EN.553.211: Probability and Statistics for Life Science

Homework

James Guo

Spring 2024

Contents

This document compiles all the homeworks from `EN.553.211` Probability and Statistics for Life Science, instructed by *Dr. Taylor Jones* at *Johns Hopkins University*.
This documents includes the solution to all the assigned problems. During the homework, collaborations were made with *Annie Xian* and *Jessica Zhang*. Most of the errors pointed out by the TAs have been adjusted.

Best regard, James Guo.
May 2024.

1 Describing Data, Measures of Center and Variability

Problem 1.1. (Experimental Units). Identify the experimental units on which the following variables are measured:

(a) Gender of a student

(b) Number of errors on a midterm exam

(c) Age of a cancer patient

(d) Number of flowers on an azalea plant

(e) Color of a car entering a parking lot

Sol'n.

(a) The experimental units are $\boxed{\text{the students}}$.

(b) The experimental units are $\boxed{\text{the midterm exams}}$.

(c) The experimental units are $\boxed{\text{the cancer patients}}$.

(d) The experimental units are $\boxed{\text{the azalea plants}}$.

(e) The experimental units are $\boxed{\text{the cars entering the parking lot}}$.

⌟

Problem 1.2. (Qualitative or Quantitative?). Identify each variable as quantitative or qualitative:

(a) Amount of time it takes to assemble a simple puzzle

(b) Number of students in a first-grade classroom

(c) Rating of a newly elected politician (excellent, good, fair, poor)

(d) State in which a person lives

Sol'n.

(a) The variable is $\boxed{\text{quantitative}}$.

(b) The variable is $\boxed{\text{quantitative}}$.

(c) The variable is $\boxed{\text{qualitative}}$.

(d) The variable is $\boxed{\text{qualitative}}$.

⌟

Problem 1.3. (Discrete or Continuous?). Identify the following quantitative variables as discrete or continuous:

(a) Population in a particular area of the United States

(b) Weight of newspapers recovered for recycling on a single day

(c) Time to complete a sociology exam

(d) Number of consumers in a poll of 1000 who consider nutrition labels on food products to be important

Sol'n.

(a) The variable is $\boxed{\text{discrete}}$.

(b) The variable is $\boxed{\text{continuous}}$.

(c) The variable is $\boxed{\text{continuous}}$.

(d) The variable is $\boxed{\text{discrete}}$.

⌟

Problem 1.4. (Basic Techniques 1.10). Fifty people are grouped into four categories – A, B, C, and D – and the number of people who fall into each category is shown in the table:

Category	Frequency
A	11
B	14
C	20
D	5

(a) What is the experimental unit?

(b) What is the variable being measured? Is it qualitative or quantitative?

(c) What *proportion* of the people are in category B, C, or D?

(d) What is the *percentage* of the people who are not in category B?

Sol'n.

(a) The experimental unit is $\boxed{\text{each person of the 50 people}}$.

(b) The variable measured is $\boxed{\text{the category of each person}}$ and it is $\boxed{\text{qualitative}}$.

(c) The proportion of the people in category B, C, and D is:

$$\frac{f_B + f_C + f_D}{N} = \frac{14 + 20 + 5}{50} = \boxed{\frac{39}{50}}.$$

(d) The percentage of the people who are not in category B is:

$$\frac{N - f_B}{N} \times 100\% = \frac{50 - 14}{50} \times 100\% = \frac{36}{50} \times 100\% = \boxed{72\%}.$$

┘

Problem 1.5. (Read the Syllabus). The following can all be found in the syllabus:

(a) What are the dates and times of the three exams in this course?

(b) Where is homework submitted? How many penalty points are given for not tagging problems?

(c) Under what circumstances are late projects or homework assignments accepted?

(d) For how long are regrade requests available?

(e) What are the office hour times and locations of your instructor?

<u>Sol'n.</u>

(a) The three exams for this course will have dates and times as follows:

Exam	Date	Time
Midterm 1	February 28, 2024	7-9pm
Midterm 2	April 17, 2024	7-9pm
Final∗	May 9, 2024	2-5pm

∗ The final exam date/time was published later on the syllabus, but it can also be found on Spring 2024 Examination Schedule.

(b) The homework is submitted to $\boxed{\text{Gradescope}}$ (from Canvas). A penalty of $\boxed{\text{15 points}}$ would be given for not tagging problems.

(c) Late projects or homework will not be accepted at any circumstances, *i.e.*, under $\boxed{\text{no circumstances}}$ are they accepted.

(d) The regrade requests will be available for $\boxed{\text{1 week}}$ after the grades are published.

(e) For my instructor, Professor Jones, the office hour times and locations are as follows:

Times	Location
Mondays 1-2pm	Wyman N420
Tuesdays 10-11am	Wyman N420

⌟

Problem 1.6. (Weighted Average). Suppose we have two sets of data

(a) $\{x_1, x_2, \cdots, x_n\}$

(b) $\{y_1, y_2, \cdots, y_m\}$

note it may not necessarily be the case that $n = m$. Consider the data set $\{x_1, x_2, \cdots, x_n, y_1, y_2, \cdots, y_m\}$ and index them by z_i.

(a) Find a formula for the sample average z in terms of x and y.

(b) According to the 2010 census, the state of California had a population of $N = 37{,}253{,}956$ and a per capita income of $\bar{x} = \$41{,}893/\text{yr}$. On the other hand, the state of Idaho had a population of $M = 1{,}567{,}582$ with a per capita income of $\bar{y} = \$31{,}556/\text{yr}$. What was the 2010 per capita income of the states of California and Idaho combined, i.e. a hypothetical country comprised of CA and ID?

Sol'n.

(a) To find the average of z, we use the following method:

$$\begin{aligned}\bar{z} &= \frac{z_1 + \cdots + z_n + z_{n+1} + \cdots + z_{n+m}}{m+n} = \frac{\sum_{i=1}^{n} x_i + \sum_{i=1}^{m} y_i}{m+n} \\ &= \boxed{\frac{n\bar{x} + m\bar{y}}{m+n}},\end{aligned}$$

which implies that $\bar{z}$ is the weighted average of $\bar{x}$ and $\bar{y}$ with weight n and m, respectively.

(b) According to the proceeding problem, we have that:

$$\begin{aligned}\text{Average 2010 per capita income for two states} &= \frac{N\bar{x} + M\bar{y}}{M+N} \\ &= \frac{\$41{,}893/\text{yr} \times 37{,}253{,}956 + \$41{,}893/\text{yr} \times 1{,}567{,}582}{37{,}253{,}956 + 1{,}567{,}582} \\ &= \$\frac{1{,}560{,}679{,}978{,}708 + 65{,}670{,}712{,}726}{38{,}821{,}538}/\text{yr} \\ &= \$\frac{1{,}610{,}146{,}596{,}300}{38{,}821{,}538}/\text{yr} \\ &= \boxed{\$41{,}475.6003819323/\text{yr}} \approx \$41{,}475.60/\text{yr}.\end{aligned}$$

$\lrcorner$

Problem 1.7. (An Archaeological Find). An article in Archaeomerty involved an analysis of 26 samples of Romano-British pottery found at four different kiln sites in the United Kingdom. The samples were analyzed to determine their chemical composition. The percentage of iron oxide in each of five samples collected at the Island Thorns site was:

1.28 2.39 1.50 1.88 1.51

(a) Calculate the range.

(b) Calculate the sample variance and standard deviation.

(c) Compare the range and standard deviation. The range is approximately how many standard deviations?

Sol'n.

(a) The range of the data is:

$$\begin{aligned} \text{range} &= \max\left(\{x_i \mid 1 \le i \le 5\}\right) - \min\left(\{x_i \mid 1 \le i \le 5\}\right) \\ &= \max\left(\{1.28, 2.39, 1.50, 1.88, 1.51\}\right) - \min\left(\{1.28, 2.39, 1.50, 1.88, 1.51\}\right) \\ &= 2.39 - 1.28 = \boxed{1.11}. \end{aligned}$$

(b) Prior to these calculation, we needed to find the average of the data:

$$\text{mean} = \frac{\sum_{i=1}^{n} x_i}{n} = \frac{1.28 + 2.39 + 1.50 + 1.88 + 1.51}{5} = \frac{8.56}{5} = 1.712.$$

Thus the sample variance is then:

$$\begin{aligned} s^2 &= \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2 \\ &= \frac{1}{5-1}\left((1.28 - 1.712)^2 + (2.39 - 1.712)^2 + (1.50 - 1.712)^2 + (1.88 - 1.712)^2 + (1.51 - 1.712)^2\right) \\ &= \frac{1}{4}(0.186624 + 0.459684 + 0.044944 + 0.028224 + 0.040804) = \frac{0.76028}{4} = \boxed{0.19007}. \end{aligned}$$

The standard deviation, correspondingly, is:

$$s = \sqrt{s^2} = \sqrt{0.19007} \approx \boxed{0.4359701825}.$$

(c) The range is $\boxed{\text{larger}}$ than the standard deviation. The range is approximately $1.11 \div 0.4359701825 \approx$ $\boxed{2.5460456805}$ standard deviations.

⌟

Problem 1.8. (Basic Technique). Give a population of three numbers such that the population median and mode are both 10, and the population mean is $1,000,000$.

Sol'n. Notice that the population median is 10, with three numbers, this means that the population in the middle must be 10.
Since the mode is also 10, that means that at least two numbers must be 10.
Therefore, let the other population number be x, then the average of $1,000,000$ implies that:

$$\begin{aligned}\bar{x} = \frac{\sum_{i=1}^{3} x_i}{3} = \frac{10 + 10 + x}{3} &= 1,000,000 \\ 20 + x &= 3,000,000 \\ x &= 2,999,980.\end{aligned}$$

Therefore, the population of three numbers are:

$$\boxed{10 \qquad 10 \qquad 2,999,980}.$$

⌟

Problem 1.9. (Aaron Rodger). The number of passes completed by Aaron Rodgers during the 2010 season while playing for the Minnesota Vikings during each of his 15 regular season games is recorded below:

19	19	34	12	27	18	21	15
27	22	26	21	7	25	19	

(a) Draw a stem and leaf plot to describe the data.

(b) Calculate the mean and standard deviation of Aaron Rodgers' per game pass completions.

(c) What proportion of the measurements lie within two standard deviations of the mean?

Sol'n.

(a) A stem and leaf plot is illustrated as follows:

Stem	Leaf
0	7
1	2 5 8 9 9 9
2	1 1 2 5 6 7 7
3	4

(b) The mean could be calculated as:

$$\mu = \frac{\sum_{i=1}^{N} x_i}{N} = \frac{19 + 19 + \cdots + 25 + 19}{15} = \frac{312}{15} = \boxed{20.8}.$$

The standard deviation could be calculated as:

$$\sigma = \sqrt{\frac{\sum_{i=1}^{N}(x_i - \mu)^2}{N}} = \sqrt{\frac{(19 - 20.8)^2 + (19 - 20.8)^2 + \cdots + (25 - 20.8)^2 + (19 - 20.8)^2}{14}}$$

$$= \sqrt{\frac{190.44 + 190.44 + \cdots + 38.44 + 190.44}{14}} \approx \sqrt{44.0285710} \approx \boxed{6.635402}.$$

(c) The range of two standard deviations of the mean is:

$$[\mu - 2\sigma, \mu + 2\sigma] = [7.529194, 34.0708057],$$

which implies that 14 measurements lie within the range, so the proportion is $\boxed{\dfrac{14}{15}}$.

┘

Problem 1.10. (Descriptive statistics). Fifty students from the New York Public School system were randomly sampled and their WISC-IV scores are shown below. This sample is ordered for your convenience (in 10 columns of 5 values each).

65	84	92	100	102	106	110	116	120	134
68	84	94	100	103	107	112	116	120	134
72	85	95	101	103	108	115	116	124	136
76	85	96	101	104	108	116	118	128	136
79	91	99	102	105	110	116	119	134	137

Find $Q1$, the median, $Q3$, and the IQR and sketch a box-plot.

Sol'n. To account for the problem, we want to order the list of data first. The result, from the smallest to the largest is:

65	68	72	76	79	84	84	85	85	91
92	94	95	96	99	100	100	101	101	102
102	103	103	104	105	106	107	108	108	110
110	112	115	116	116	116	116	116	118	119
120	120	124	128	134	134	134	136	136	137

Therefore, with a total of 50 values, we have $n + 1 = 1$ in which $n_1 = 51 \div 4 = 12.75$, $n_2 = 51 \div 2 = 25.5$, and $n_3 = 51 \div 4 \times 3 = 38.25$:

$$Q1 = 0.25x_{\lfloor 12.75 \rfloor} + 0.75x_{\lceil 12.75 \rceil} = 0.25x_{12} + 0.75x_{13} = 0.25 \times 94 + 0.75 \times 95 = \boxed{94.75},$$

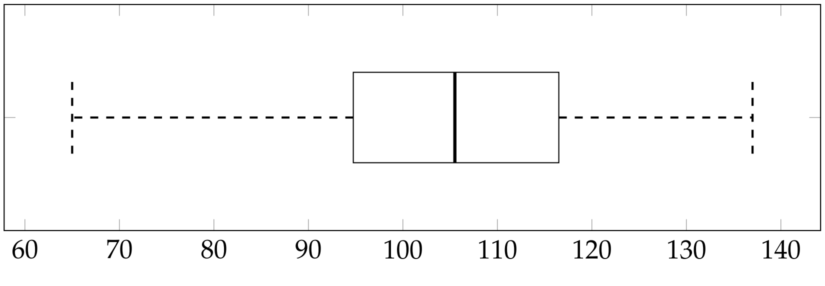
$$\tilde{x} = Q2 = 0.5x_{\lfloor 25.5 \rfloor} + 0.5x_{\lceil 25.5 \rceil} = 0.5x_{25} + 0.5x_{26} = 0.5 + 0.5 \times 106 = \boxed{105.5},$$

$$Q3 = 0.75x_{\lfloor 38.25 \rfloor} + 0.25x_{\lceil 38.25 \rceil} = 0.75x_{38} + 0.25x_{39} = 0.75 \times 116 + 0.25 \times 118 = \boxed{116.5}.$$

Correspondingly, the IQR can be calculated as:

$$IQR = Q3 - Q1 = 116.5 - 94.75 = \boxed{21.75}.$$

With these information, a box-plot can be sketched as below:



$\lrcorner$

2 Bivariate Data: Correlation, Causation, Experiments, and Observations

Problem 2.1. (Conceptual Understanding). Determine whether each of the following statements are True or False, explain your answers.

(a) Suppose the bivariate data $(x_1, y_1), \cdots, (x_n, y_n)$ has a linear relationship is negatively correlated. Then the data which arises by switching the independent and dependent variables $(y_1, x_1), \cdots, (y_n, x_n)$ will also be negatively correlated.

(b) Suppose variables x and y are positively correlated, then an increase in x causes an increase in y.

(c) Suppose that variables x and y are independent. Then best-fitting line will be approximately horizontal.

Sol'n.

(a) $\boxed{\text{True}}$. The correlation of variable $\{(x_i, y_i)\}_{i=1}^n$, denoted by $r_{x,y}$ is:

$$\begin{aligned} r_{x,y} &:= \frac{s_{xy}}{s_x \cdot s_y} \\ &= \frac{\frac{1}{n-1}\sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y})}{s_x \cdot s_y} \\ &= \frac{\frac{1}{n-1}\sum_{i=1}^n (y_i - \bar{y})(x_i - \bar{x})}{s_y \cdot s_x} \\ &= \frac{s_{yx}}{s_y \cdot s_x} \\ &= r_{y,x}. \end{aligned}$$

Hence, since the covariance are the same and the correlation coefficient are the same for $\{(x_i, y_i)\}_{i=1}^n$ and $\{(y_i, x_i)\}_{i=1}^n$, thus they are negatively correlated.

(b) $\boxed{\text{False}}$. Since the correlation implies only association, but does not imply causation, then an increase in x will not necessarily cause an increase in y. A causation would lead to correlation, but its converse is not necessarily true due to confounding variables, *i.e.*, correlation $\not\Rightarrow$ causation.

(c) $\boxed{\text{True}}$. By definition, x and y being independent implies that they are not correlated, meaning that its coefficient should be approximately 0, which implies that the slope of the best-fitting line is:

$$m = r \cdot \frac{s_y}{s_x} \approx 0,$$

implying it is approximately horizontal.

⌟

Problem 2.2. (Experiment or Observation?). For each statistical study below, identify the experimental units and both the explanatory and response variables being measured. As well, identify whether it is an observational study or experimental study. If it is an experimental study, identify the control and treatment groups.

(a) Persistent pulmonary hypertension of the newborn (PPHN) is defined as the failure of the normal circulatory transition that occurs after birth. A study was conducted to test the association between PPHN and exposure to selective serotonin-reuptake inhibitors (SSRIs) during late pregnancy. Between 1996 and 2007, data was collected on 1.6 million infants across 5 Nordic countries. Around 30,000 women had used SSRIs during pregnancy and 11,014 had been dispensed an SSRI later than gestational week 20. Exposure to SSRIs in late pregnancy was associated with an increased risk of persistent pulmonary hypertension in the newborn: 33 of 11,014 exposed infants (absolute risk 3 per 1000 liveborn infants compared with the background incidence of 1.2 per 1000).

(b) Domestic violence is a serious offense with a higher recidivism rate than other violent crimes. Researchers wished to determine whether a batterer treatment program would be effective at reducing recidivism. They conducted a study involving 376 adult males convicted of domestic violence who were randomly assigned to either a 40-hour batterer treatment program or 40 hours of community service that did not include a therapeutic component. Official records and victims' reports were examined to determine the six-month recidivism rate for both groups. The researchers found that 10% of the males assigned to the treatment program had further battering incidents, compared to 21% of the males sentenced to community service. The researcher concluded that therapeutic treatment may reduce domestic violence among convicted batterers.

(c) A randomly selected sample of 1088 high school students from 20 schools completed surveys about their lunch practices and vending machine purchases. School food policies were assessed by principal and food director surveys. The number of vending machines and their hours of operation were assessed by trained research staff. The results found that students at schools with open campus policies during lunchtime were significantly more likely to eat lunch at a fast food restaurant than students at schools with closed campus policies. Furthermore, student snack food purchases at school were significantly associated with the number of snack machines at schools and policies about the types of food that can be sold.

Sol'n.

(a) The experimental units are each of the 1.6 million (woman and) $\boxed{\text{infants across 5 Nordic countries}}$,
The explanatory variable is the $\boxed{\text{exposure to selective SSRIs}}$ during late pregnancy,
The response variable is the $\boxed{\text{risk of persistent pulmonary hypertension in the new born}}$,
This is an $\boxed{\text{observational study}}$.

(b) The experimental units are each of the 376 $\boxed{\text{adult males convicted of domestic violence}}$,
The explanatory variable is $\boxed{\text{whether one was assigned to 40-hour batterer treatment program}}$ or not (that is the community service),
The response variable is $\boxed{\text{whether the individual had further battering incidents}}$,

This is an $\boxed{\text{experimental study}}$, the control group is under $\boxed{\text{40 hours of community service}}$ and the treatment group is with $\boxed{\text{40-hour batterer treatment program}}$ that did not include a therapeutic component.

(c) The experimental units are each of the 1088 randomly selected $\boxed{\text{high school students}}$ from 20 schools,
The explanatory variables are $\boxed{\text{the open campus policies}}$ and $\boxed{\text{the number of snack machines}}$ at school (and their hours of operation),
The response variables are $\boxed{\text{place that a student goes for lunch}}$ and $\boxed{\text{number snack food purchases}}$ by students,
This is an $\boxed{\text{observational study}}$.

⌟

Problem 2.3. (Identifying Biases). Identify the type of bias which might be present in each example below.

(a) As of July 2023, the highest rated movie on IMDB is The Shawshank Redemption. A statistician uses this data to conclude that The Shawshank Redemtion is the most popular movie in America.

(b) Recall from lecture that the `SPRING_2023_Census_data` is a survey taken by 613 statistics students at the University of North Texas. A statistician uses this data to estimate that the average age of Texas residents is 19.57 years old.

(c) Heinmann, a publishing company of elementary education materials, sponsors an experimental study to test the effectiveness of cueing strategies when teaching children to read. The results of the study concluded that cueing strategies were significantly more effective at improving student's reading abilities than other methods.

<u>Sol'n.</u>

(a) First of all, people mostly give comments on movies when they extremely like it or hate it, hence the rate of the movies lacks the data on people's rating for people who moderately likes it or hate it. Hence, there could an movie that many people like it, but not to an extent that people would love it to rate it online, compared to *The Shawshank Redemtion*, which does not necessarily approve that it is the most popular. Hence, this would be a $\boxed{\text{self-selection bias}}$.
Moreover, this data only represent the rated movies up to July 2023, so the newer movies might not have sufficiently many votes on whether the general audiences like the movie. Moreover, since IMDB is based on the United Kingdom and can be accessed almost worldwide, the data does not represent the most popular movies in America, as people tend to rate movies from their country higher. These are $\boxed{\text{sampling biases}}$ or having outside data polluting the data set.

(b) This survey only contains college students, whose age is typically around 20 years old. This neglects the adults and elders in Texas which is also a large population in Texas. This is $\boxed{\text{sampling bias}}$.

(c) The sponsors being the publishing company could be favoring the use of books while reading and they could select the children deliberately to promote the company. Hence, this is $\boxed{\text{self-interest bias}}$.

⌟

Problem 2.4. (Sampling). Identify the type of sampling used in each experiment below. Explain your reasoning.

(a) A school consisting of 1000 students can be classified by hair color:

- 57% black,
- 29% brown,
- 12% blonde,
- 2% red.

The school wants to survey 200 of their students, so they randomly choose 114 students from a list of all the black haired students, 58 from all the brown haired students, 24 students from all the blonde haired students, and 2 students from all the red haired students.

(b) A fast food restaurant chain has about 5 locations in each of four cities with similar populations: Atlanta, Baltimore, Cincinnati, and Detroit. A city is randomly selected and the average revenue over all restaurant locations in that city are recorded for the week.

(c) A small retail company of 20 employees wants to select 5 of them for a performance review. The manager puts each employee's name on identical slips of paper and puts the slips of paper in a non-transparent box. The box is well shaken and five slips of paper are drawn to determine which employees will receive a performance review.

Sol'n.

(a) Notice that:

$$\begin{cases} 200 \times 57\% = 114, \\ 200 \times 29\% = 58, \\ 200 \times 12\% = 24, \\ 200 \times 2\% = 4, \end{cases}$$

which corresponds to the ratio of of each group (the hair color) represented in the school, hence this is $\boxed{\text{stratified sampling}}$.

(b) Since when the population is partitioned into groups by city and one cluster is selected randomly (indented to be SRS), then this is $\boxed{\text{cluster sampling}}$.

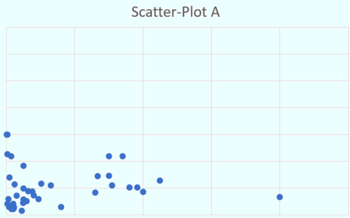
(c) By having all names on the paper, shuffled in a non-transparent box, and drawing pieces from it, the five people are selected randomly (indented to be SRS), then this is $\boxed{\text{simple random sampling}}$ (or SRS).

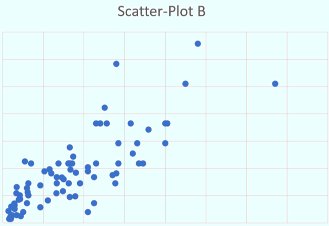
┘

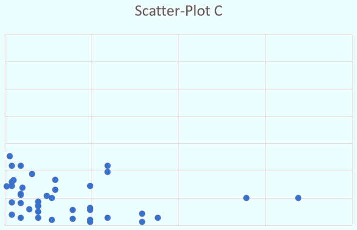
Problem 2.5. (Mammal Life Spans). In the article Evolution of Reproductive Life History in Mammals and the Associated Change of Functional Constraints, data was collected on 89 species of mammals on their average gestation time, number of offspring, weight, and other reproductive related traits to study the association these variables might have with life expectancy (in days).
Given the correlation coefficients:

- $r = 0.8105$ for average gestation period and average lifespan,
- $r = -0.4085$ for average number of offspring and average lifespan,
- $r = 0.4937$ for average weight and average lifespan,

(a) Match the scatter-plots with the corresponding variable (gestation time, number of offspring, weight) when plotted against average lifespan. Explain your choices.







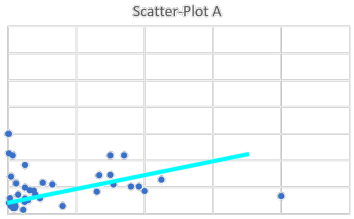
(b) Given the covariance of gestation time $\langle x \rangle$ and lifespan $\langle y \rangle$ is

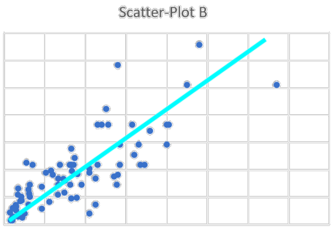
$$s_{xy} = 702713.5$$

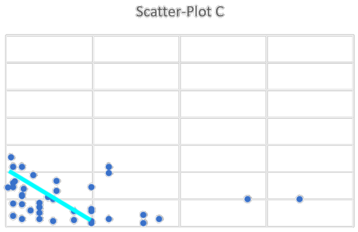
and the slope of the line of best-fit is $m = 41.284$, calculate the standard deviation of gestation time s_x.

Sol'n.

(a) A best fit line is roughly sketched below:







- The correlation coefficients of $r = 0.8105$ for *average gestation period and average lifespan* should be matched to $\boxed{\texttt{Scatter-Plot B}}$, since 0.8105 is close to 1, it would be a moderately strong positive correlation, *i.e.*, the line of the best fit should be having a positive slope.

- The correlation coefficients of $r = -0.4082$ for *average number of offspring and average lifespan* should be matched to $\boxed{\texttt{Scatter-Plot C}}$, since -0.4082 is negative, it would be a moderately weak negative correlation, *i.e.*, the line of the best fit should be roughly having a negative slope.
- The correlation coefficients of $r = 0.4937$ for *average weight and average lifespan* should be matched to $\boxed{\texttt{Scatter-Plot A}}$, since 0.4937 is positive , it would be a moderately weak negative correlation, *i.e.*, the line of the best fit should be roughly having a negative slope.

(b) Now we know that the slope of the line of best-fit is $m = 41.284$, the covariance if $s_{xy} = 702713.5$, then by the formula, we have:

$$\begin{aligned} m &= r \cdot \frac{s_y}{s_x} = \frac{s_{xy}}{s_x s_y} \cdot \frac{s_y}{s_x} \\ &= \frac{s_{xy}}{s_x^2} \\ s_x &= \sqrt{\frac{s_{xy}}{m}} = \sqrt{\frac{702713.5}{41.284}} \\ &\approx \boxed{130.4662752879}. \end{aligned}$$

⌟

Problem 2.6. (Basic Techniques). Consider the set of bivariate data

x	1	2	3	4	5	6
y	5.6	4.6	4.5	3.7	3.2	2.7

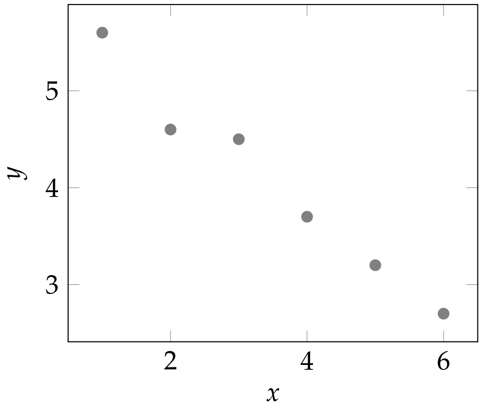
(a) Draw a scatter-plot of the data.

(b) Does there appear to be a relationship between x and y? If so, how do you describe it?

(c) Calculate the correlation coefficient r. Does the r value confirm your conclusion in part (b)?

Sol'n.

(a) The scatter-plot of the data is as follows:



(b) By looking at the scatter-plot, there seems to be a quite strong negative correlation between x and y, *i.e.*, a decrease in x *associates* a decrease in y.

(c) In calculating the formula, we first need to calculate the mean and standard deviation of x and y:

$$\bar{x} = 3.5, \qquad s_x \approx 1.870828693, \qquad \bar{y} = 4.05, \qquad s_y \approx 1.055935604.$$

Then, we shall calculate the covariance by the formula:

$$\begin{aligned} s_{xy} =& \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y}), \\ =& \frac{1}{5}\big((1-3.5)(5.6-4.05) + (2-3.5)(4.6-4.05) + (3-3.5)(4.5-4.05) \\ & + (4-3.5)(3.7-4.05) + (5-3.5)(3.2-4.05) + (6-3.5)(2.7-4.05)\big), \\ =& \frac{-3.875 - 0.825 - 0.225 - 0.175 - 1.275 - 3.375}{5} = \frac{-9.75}{5} = -1.95. \end{aligned}$$

Therefore, we can then calculate the correlation coefficient:

$$r = \frac{s_{xy}}{s_x \cdot s_y} \approx \frac{-1.95}{1.870828693 \times 1.055935604} \approx \boxed{-0.9871045542}.$$

Since r is positive and very close to 1, this implies quite a string negative correlation between x and y, same as described in part (b).

┘

Problem 2.7. (Mr. Plow's Snowplowing Business Takes a Hit). The amount of snowfall/yr. in Springfield is plotted over a 10 year period:

Year	Snow Fall (in.)	Year	Snow Fall (in.)
1	61.5	6	58.2
2	62.3	7	57.5
3	60.7	8	57.5
4	59.8	9	56.1
5	58.0	10	56.0

(a) Draw a scatter-plot to describe the changes in snowfall in Springfield over time.

(b) Calculate the correlation coefficient r given that the variance of Years $\langle x \rangle$ is $s_x^2 \approx 9.16667$, the variance of Snowfall $\langle y \rangle$ is $s_y^2 \approx 4.84933$, and the covariance is

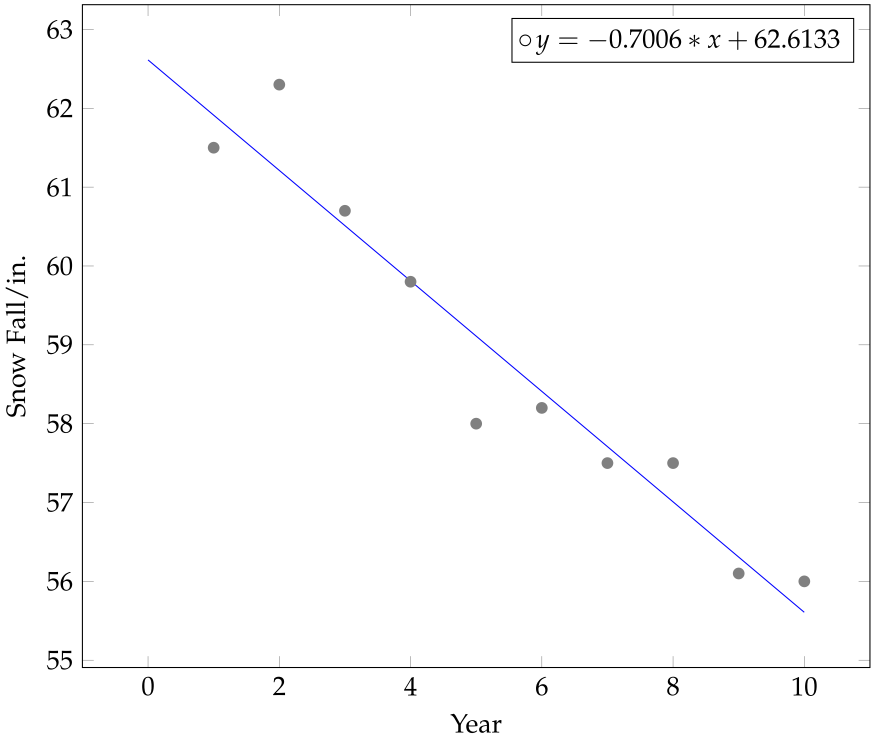
$$s_{xy} \approx -6.42222.$$

(c) Calculate the equation of the best fitting line using the result from part (b).

(d) Plot the best-fitting line on your scatter-plot from part (a).

Sol'n.

(a,d) The scatter-plot of the data and the best-fitting line is as follows:



(b) Here, we may be directly using the formula that:

$$r = \frac{s_{xy}}{s_x \cdot s_y} \approx \frac{-6.42222}{\sqrt{9.16667} \times \sqrt{4.84933}} \approx \frac{-6.42222}{3.02765090458 \times 2.2021194336} \approx \boxed{-0.9632488754}.$$

(c) The average of x and y can be calculated as:

$$\bar{x} = 5.5, \qquad \bar{y} = 58.76.$$

Therefore, the line of best fit as $mx + b$ has:

$$m = r \cdot \frac{s_y}{s_x} \approx \frac{-0.9632488754 \times 2.2021194336}{3.02765090458} \approx -0.7006055634,$$

$$b = \bar{y} - m\bar{x} \approx 58.76 - (-0.7006055634) \times 5.5 \approx 62.6133305987.$$

Therefore, the equation of the best fit line is:

$$\boxed{\hat{y} = -0.7006055634x + 62.6133305987}.$$

⌟

Problem 2.8. (Grocery Costs). The following data relates the number of household members to the amount spent on groceries per week.

x	2	2	3	4	1	5
y	\$95.75	\$110.19	\$118.33	\$150.92	\$85.86	\$180.62

(a) Find the best-fitting line for the data.

(b) Plot the points and the best-fitting line on the same graph.

(c) What would you predict a household of size six to spend on groceries per week? How about a house hold of size 9? Do you think these predictions would be very accurate? Explain.

Sol'n.

(a) In calculating the formula, we first need to calculate the mean and standard deviation of x and y:

$$\bar{x} \approx 2.833333333, \qquad s_x \approx 1.471960144, \qquad \bar{y} \approx 123.6116667, \qquad s_y \approx 35.79159981.$$

Then, we shall calculate the covariance by the formula:

$$\begin{aligned}s_{xy} =& \frac{1}{n-1}\sum_{i=1}^{n}(x_i-\bar{x})(y_i-\bar{y}),\\ =& \frac{1}{5}\big((2-2.8333)(95.75-123.6117)+(2-2.8333)(110.19-123.6117)+(3-2.8333)(118.33-123.6117)\\ &+(4-2.8333)(150.92-123.6117)+(1-2.8333)(85.86-123.6117)+(5-2.8333)(180.62-123.6117)\big),\\ =& \frac{23.21715461+11.18430261-0.88045939+31.86059361+69.21019161+123.51988361}{5}\\ =& \frac{258.11166666}{5} \approx 51.6223.\end{aligned}$$

Therefore, we can then calculate the correlation coefficient:

$$r = \frac{s_{xy}}{s_x \cdot s_y} \approx \frac{51.6223}{1.471960144 \times 35.79159981} \approx 0.9798513349.$$

Thereby, we can calculate the parameters for the line of best fit as $mx+b$, as:

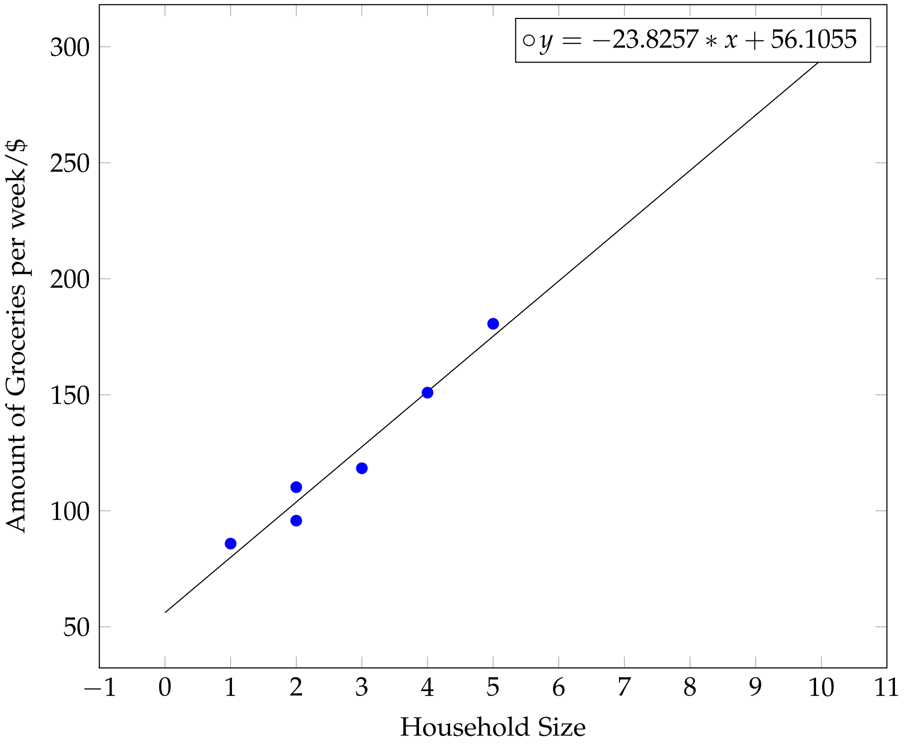
$$m = r\cdot\frac{s_y}{s_x} \approx \frac{0.9798513349 \times 35.79159981}{1.471960144} \approx 23.8256769349,$$

$$b = \bar{y} - m\bar{x} \approx 123.6116667 - 23.8256769349 \times 2.83333333 \approx 56.1055821305.$$

Therefore, the equation of the best fit line is:

$$\boxed{\hat{y} = 23.8256769349x + 56.1055821305}.$$

(b) The scatter-plot of the data and the best-fitting line is as follows:



(c) In predicting the household of size six and nine, we can use the line of the best fit:

$$\hat{y}(x = 6) = 23.8256769349 \times 6 + 56.1055821305 \approx 199.0596437399,$$

$$\hat{y}(x = 6) = 23.8256769349 \times 6 + 56.1055821305 \approx 270.5366745446,$$

which implies that for a household of size six, it should spend approximately $\boxed{\$199.06/\text{week}}$, whereas for a household of size six, it should spend approximately $\boxed{\$270.54/\text{week}}$.

These approximation could be accurate since the correlation coefficient is very close to 1, implying that there is a strong correlation.

Moreover, in general, the prediction for a household of six should be more accurate since it is closer to the x values that we have track of.

However, given that there are only 6 data points, the small sample size could not clearly represent the whole population even if it is SRS. Hence, it cannot be entirely accurate.

⌟

3 Elementary Probability

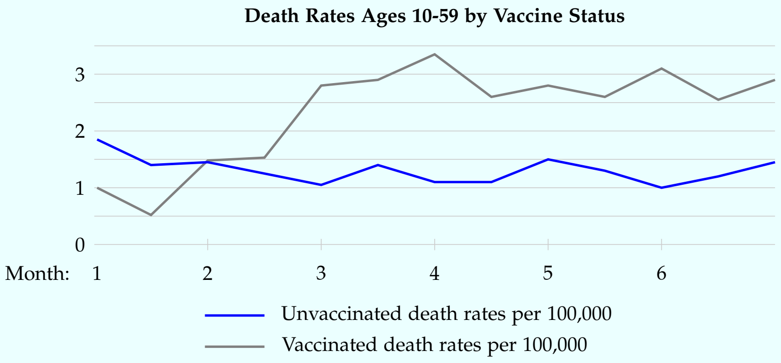
Problem 3.1. (Simpson's Paradox). This problem illustrates how combining data sets can produce counter- intuitive results.

(a) There are two hats. Hat one contains 5 red balls and 6 green balls. Hat two contains 3 red and 4 green balls. You draw a ball at random from one of the hats and if it is red, you win a prize. Which hat should you draw from?

(b) Suppose there there are two more hats, Hat three and Hat four. Hat three contains 6 red balls and 3 green. Hat four contains 9 red balls and 5 green balls. Again, you win a prize if you draw a red ball. Which hat should you draw from?

(c) Finally, suppose the contents of Hat 1 and Hat 3 are combined into a single hat (call this Hat Odd), while the contents of Hat 2 and Hat 4 are combined as well (call this Hat Even). Again, a red ball wins a prize, which Hat should you draw from?

(d) Parts (a)-(c) might seem like a meaningless trick in a meaningless game, but similar scenarios often happen in real life. The following table is based on real data collected by the Office of National Statistics:



The chart plots the death rates over time of individuals in the U.K. who received a COVID-19 vaccination (in orange) and those who did not (in blue). It appears strange, right? Indeed, this chart made its rounds in social media during Fall 2021, usually in an attempt to mislead others. Use the hat game in parts (a)-(c) as an analogy to explain what is happening with the data.

Sol'n.

(a) Suppose that the draw is completely random, then the probability for winning the prize, respectively, are:

$$P(\text{winning for hat 1}) = \frac{\#(\text{winning outcomes for hat 1})}{\#(\text{all outcomes for hat 1})} = \frac{5}{5+6} = \frac{5}{11} = \frac{35}{77},$$

$$P(\text{winning for hat 2}) = \frac{\#(\text{winning outcomes for hat 2})}{\#(\text{all outcomes for hat 2})} = \frac{3}{3+4} = \frac{3}{7} = \frac{33}{77}.$$

Note that $\frac{35}{77} > \frac{33}{77}$, thus we shall be drawing from $\boxed{\text{hat 1}}$.

(b) Here, we would continually calculating the probability for winning the prize for hat 3 and 4:

$$P(\text{winning for hat 3}) = \frac{\#(\text{winning outcomes for hat 3})}{\#(\text{all outcomes for hat 3})} = \frac{6}{6+3} = \frac{2}{3} = \frac{28}{52},$$
$$P(\text{winning for hat 4}) = \frac{\#(\text{winning outcomes for hat 4})}{\#(\text{all outcomes for hat 4})} = \frac{9}{9+5} = \frac{9}{14} = \frac{27}{52}.$$

Note that for hat 3 and 4, the probability is larger than $\frac{1}{2}$, so they have larger chances to win than hat 1 and 2. Then, note that $\frac{28}{52} < \frac{27}{52}$, we shall be drawing from $\boxed{\text{hat 3}}$.

(c) Likewise, we would continually calculating the probability for winning the prize for hat odd and even:

$$P(\text{winning for hat odd}) = \frac{\#(\text{winning outcomes for hat odd})}{\#(\text{all outcomes for hat odd})} = \frac{5+6}{5+6+6+3} = \frac{11}{20} = \frac{231}{420},$$
$$P(\text{winning for hat even}) = \frac{\#(\text{winning outcomes for hat even})}{\#(\text{all outcomes for hat even})} = \frac{3+9}{3+4+9+5} = \frac{12}{21} = \frac{240}{420}.$$

Note that $\frac{231}{420} < \frac{240}{420}$, we shall be drawing from $\boxed{\text{hat even}}$.

(d) Here if we consider getting a red ball as surviving and getting a green ball as death. Consider hat 1 and 3 as vaccinated people and consider hat 2 and 4 as unvaccinated people.

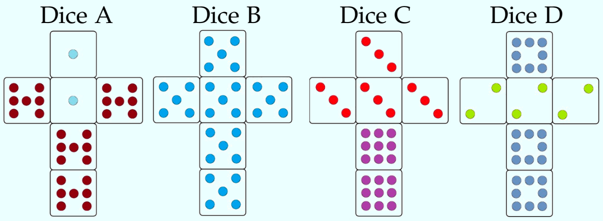
Further more, we may assume that hat 1 and 2 is a smaller population, whereas hat 3 and 4 is a larger population. Note that for parts (a) and (b), both hat 1 and 3 are more likely to survive, but when the statistics are added together, hat 2 and 4 combined has a higher survival rate.

Specifically for COVID example, the different groups could signify different population groups, there could be less children vaccinated and more middle-aged or elders being vaccinated. At the same time, the vaccinated middle-aged or elders has a higher death rate compared to children, but they could have a smaller death rater as well compared to the unvaccinated people.

Especially in the COVID context, if we divide the groups into vaccinated and unvaccinated people of younger and older ages. Since the elders receive the vaccines earlier, they might have a decreased rate of death, but it is still higher than the younger, which pulls the death rate for the whole population to higher as the elders takes up more of the vaccinated population (and consequently pulled down the death rate for unvaccinated population). In this case, a more conclusive result could be determined by having people of different age groups, especially differed by the general death rate and vaccination rate to be summarized separately.

⌟

Problem 3.2. (Non-Transitive Dice). Below are four unusually numbered dice, A, B, C, and D.



Suppose that you and a friend each pick one of the four die to roll. Whoever rolls the largest number wins.

(a) Show that the probability that dice A beats dice B is 2/3.

(b) Show that the probability that dice B beats dice C is 2/3.

(c) Show that the probability that dice C beats dice D is 2/3.

(d) Show that the probability that dice D beats dice A is 2/3.

This is essentially a probabilistic spin on the game of Rock, Paper, Scissors.

Proof. We want to show all the following cases, as desired.

(a) Note that for dice B, it can only row to 5, thus we can just consider about dice A only. Out of the six total outcomes, there is four cases that A could beat B, thus the probability is:

$$P(\text{dice A beats dice B}) = \frac{\#(\text{winning outcomes for dice A})}{\#(\text{all outcomes})} = \frac{4}{6} = \frac{2}{3}.$$

(b) Also note that we still have B which can only row to 5, thus we can just consider about dice C only. Out of the six total outcomes, there is two cases that C could beat B, thus the probability is:

$$P(\text{dice B beats dice C}) = \frac{\#(\text{winning outcomes for dice B})}{\#(\text{all outcomes})} = \frac{6-2}{6} = \frac{2}{3}.$$

(c) In the competition between C and D, we have all the possible equivalent class of outcome and occurrence as:

Result of C	Result of D	Occurrence
3	2	$4 \times 3 = 12$
3	8	$4 \times 3 = 12$
9	2	$2 \times 3 = \ \ 6$
9	8	$2 \times 3 = \ \ 6$

Therefore, out of the thirty-six outcomes, C wins twenty-four of them, thus the probability is:

$$P(\text{dice C beats dice D}) = \frac{\#(\text{winning outcomes for dice C})}{\#(\text{all outcomes})} = \frac{24}{36} = \frac{2}{3}.$$

(d) In the competition between A and D, we have all the possible equivalent class of outcome and occurrence as:

Result of A	Result of D	Occurrence
1	2	$2 \times 3 = \ \ 6$
1	8	$2 \times 3 = \ \ 6$
7	2	$4 \times 3 = 12$
7	8	$4 \times 3 = 12$

Therefore, out of the thirty-six outcomes, D wins twenty-four of them, thus the probability is:

$$P(\text{dice D beats dice A}) = \frac{\#(\text{winning outcomes for dice D})}{\#(\text{all outcomes})} = \frac{24}{36} = \frac{2}{3}.$$

□

Problem 3.3. (Board Games). Five children Amy, Brett, Chloe, Davey, and Emily are about to play the board game Candyland. The children will play until all five reach the end of the game. Since the game requires no decision making from the player, the outcome is completely determined by chance (namely, the order that the cards are shuffled). Thus each child is equally likely to finish 1^{st} through 5^{th}.

(a) How many outcomes are in the sample space? Do not list them.

(b) What is the probability that Davey finishes first or last?

(c) How many different ways can the first three finishers be recorded?

(d) What is the probability that Davey or Emily finishes among the first three?

(e) What is the probability that *either* Amy finishes first, Brett finishes second, or Chloe finishes third?

Sol'n.

(a) Since all of the five candidates are getting a place from 1 to 5, thus to total number of outcomes in the sample space is:

$$\#(S) = (5)_5 = \frac{5!}{(5-5)!} = \frac{5!}{1} = 5! = \boxed{120}.$$

(b) Since the outcome is assumed to be completely random, and since Davey cannot both first and last (these are mutually exclusive event) then the probability that Davey finishes first or last is the same as probability that Davey finishes first and probability that Davey finishes last.
The probabilities for Davey to finish first and last are:

$$P(\text{Davey finishes first}) = \frac{\#(\text{Davey finishes first})}{\#(S)} = \frac{(4)_4}{(5)_5} = \frac{4!/0!}{5!/0!} = \frac{1}{5},$$

$$P(\text{Davey finishes last}) = \frac{\#(\text{Davey finishes last})}{\#(S)} = \frac{(4)_4}{(5)_5} = \frac{4!/0!}{5!/0!} = \frac{1}{5},$$

$$P(\text{Davey finishes first or last}) = P(\text{Davey finishes first}) + P(\text{Davey finishes last}) = \frac{1}{5} + \frac{1}{5} = \boxed{\frac{2}{5}}.$$

(c) To record the first three finishers, this is the equivalent to choose the first three finishers with order from the five contestants, thus the number of ways are:

$$\#(S') = (5)_3 = \frac{5!}{(5-3)!} = \frac{5!}{2!} = 5 \times 4 \times 3 = \boxed{60}.$$

(d) Since all the probability are the same, this implies that we could look for the probability that both Davey and Emily do not finish among the first three, and the subtract it from the 1 since what we are looking for is the complementary event in the sample space S':

$$\begin{aligned}
&P(\text{Davey or Emily finishes among the first three}) \\
&= 1 - P(\text{Neither Davey nor Emily finishes among the first three}) \\
&= 1 - \frac{\#(\text{Neither Davey nor Emily finishes among the first three})}{\#(S')} \\
&= 1 - \frac{(2)_2(3)_3}{(5)_5} = 1 - \frac{2!/0! \times 3!/0!}{5!/0!} = 1 - \frac{2 \times 3 \times 2}{5 \times 4 \times 3 \times 2} = \boxed{\frac{9}{10}}.
\end{aligned}$$

(e) We still be considering all the events and the intersections. For the simplicity of notation, we denote Amy finishes first by A, Brett finishes second by B, and Chloe finishes third by C. Therefore, we have

$$\begin{aligned}
\#(A) &= (4)_4 = \frac{4!}{(4-4)!} = 4! = 24, \\
\#(B) &= (4)_4 = \frac{4!}{(4-4)!} = 5! = 24, \\
\#(C) &= (4)_4 = \frac{4!}{(4-4)!} = 5! = 24, \\
\#(A \wedge B) &= (3)_3 = \frac{3!}{(3-3)!} = 3! = 6, \\
\#(B \wedge C) &= (3)_3 = \frac{3!}{(3-3)!} = 3! = 6, \\
\#(A \wedge C) &= (3)_3 = \frac{3!}{(3-3)!} = 3! = 6, \\
\#(A \wedge B \wedge C) &= (2)_2 = \frac{2!}{(2-2)!} = 2! = 2.
\end{aligned}$$

Hence, the total number of cases where A or B or C happens is:

$$\begin{aligned}
\#(A \vee B \vee C) &= \#(A) + \#(B) + \#(C) - \#(A \wedge B) - \#(B \wedge C) - \#(A \wedge C) + \#(A \wedge B \wedge C), \\
&= 24 + 24 + 24 - 6 - 6 - 6 + 2 = 56.
\end{aligned}$$

Hence, the probability that A or B or C happens is:

$$P(A \vee B \vee C) = \frac{\#(A \vee B \vee C)}{\#(S)} = \frac{56}{120} = \boxed{\frac{7}{15}}.$$

⌟

Problem 3.4. (Cramming). A student prepares for an exam by studying a list of 10 problems. She can solve 6. For the exam, the instructor selects 5 questions at random from the list of 10. Solving three questions is sufficient for passing.

(a) What is the probability she gets a perfect score on the exam? I.e. she can solve all 5 questions.

(b) What is the minimum number of problems she should be able to solve from the list so that her probability of getting a perfect score is at least 0.9?

(c) What is the probability she passes the exam?

Sol'n.

(a) She gets a perfect score on the exam when all five questions are from the six that she knows. Assume the the selection is random, then the probability that she gets a perfect score is:

$$\begin{aligned} P(\text{Getting perfect score}) &= \frac{\#(\text{all questions selected are that she knows})}{\#(\text{selection of exams})} \\ &= \frac{\binom{6}{5}}{\binom{10}{5}} = \frac{\frac{6!}{5!(6-5)!}}{\frac{10!}{5!(10-5)!}} = \frac{6}{252} = \boxed{\frac{1}{42}}. \end{aligned}$$

(b) Here, we assume that she should be able to solve x problems to have a chance of at least 0.9 to get a perfect score, hence:

$$\frac{\binom{x}{5}}{\binom{10}{5}} \geq 0.9 \iff \frac{\frac{x!}{5!(x-5)!}}{252} \geq 0.9 \iff x(x-1)(x-2)(x-3)(x-4) \geq 27216.$$

Since a degree 5 polynomial is not solvable, while there are only 5 possible values, we shall be just calculating all these five values.

$$\begin{aligned} 5 \times (5-1) \times (5-2) \times (5-3) \times (5-4) &= 120 < 27216, \\ 6 \times (6-1) \times (6-2) \times (6-3) \times (6-4) &= 720 < 27216, \\ 7 \times (7-1) \times (7-2) \times (7-3) \times (7-4) &= 2520 < 27216, \\ 8 \times (6-1) \times (6-2) \times (6-3) \times (6-4) &= 6720 < 27216, \\ 9 \times (9-1) \times (9-2) \times (9-3) \times (9-4) &= 15120 < 27216, \\ 10 \times (10-1) \times (10-2) \times (10-3) \times (10-4) &= 30420 \geq 27216. \end{aligned}$$

Hence, she needs to be able to solve $\boxed{\text{10 questions}}$ from the list to have a probability to get a perfect score at lest 0.9.

(c) To pass the exam, then the five exam questions must contain exactly 5 or 4 or 3 questions that she knows (and they are mutually disjoint), thus, the total number of cases that she passes is:

$$\begin{aligned} \#(\text{she passes}) &= \#(\text{knows 5 questions}) + \#(\text{knows 4 questions}) + \#(\text{knows 3 questions}) \\ &= \binom{6}{5} + \binom{6}{4} \times \binom{4}{1} + \binom{6}{3} \times \binom{4}{2} \\ &= \frac{6!}{5!(6-5)!} + \frac{6!}{4!(6-4)!} \times \frac{4!}{1!(4-1)!} + \frac{6!}{3!(6-3)!} \times \frac{4!}{2!(4-2)!} \\ &= 6 + 15 \times 4 + 20 \times 6 = 186. \end{aligned}$$

Thus, the probability for her to pass the exam is:

$$\begin{aligned} P(\text{she passes}) &= \frac{\#(\text{she passes})}{\#(\text{selection of exams})} \\ &= \frac{186}{\binom{10}{5}} = \frac{186}{252} = \boxed{\frac{31}{42}}. \end{aligned}$$

┘

Problem 3.5. (Socks). A drawer contains twice as many white socks as red socks (and no other color). If two socks are drawn at random, the probability that they are both red is 1/12. How many socks are in the drawer?

Sol'n. Assume that there are x red socks, hence there are $2x$ white socks. Hence, the probability that at a random draw of two socks are red is:

$$P(\text{two draws are both red}) = \frac{x}{x+2x} \times \frac{x-1}{(x-1)+2x} = \frac{1}{12}.$$

Hence, we shall be attempting to solve the above equation of one variable, that is:

$$\begin{aligned} \frac{1}{3} \times \frac{x-1}{(x-1)+2x} &= \frac{1}{12}, \\ 4(x-1) &= 3x-1, \\ 4x-4 &= 3x-1, \\ x &= 3. \end{aligned}$$

Hence, there are 3 red socks and 6 white socks, implying that there is a total of $3+6=\boxed{9}$ socks. ⌟

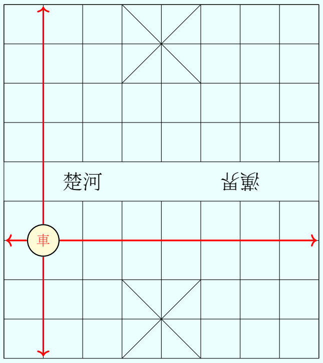
Problem 3.6. (Numbers). A positive integer is randomly picked from the set of all 5-digit positive integers whose sum of digits is 43. Assuming each number is equally likely, what is the probability that the chosen number is divisible by 11?

Sol'n. The question came up like a number theory question, but the readers should observe that the total number of 5-digit positive integers whose sum is 43 is quite limited. In fact, the number is either 4 nines and one seven or 3 nines and 2 eights, with no other choices. Hence, all of these numbers and their divisibility by 11 is:

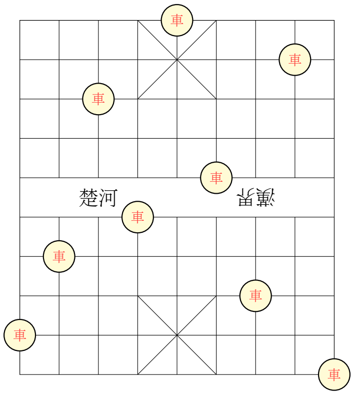
Number	Divisibility by 11
99997	No
99979	**Yes**
99799	No
97999	**Yes**
79999	No
99988	No
99898	No
98998	No
89998	No
99889	No
98989	**Yes**
89989	No
98899	No
89899	No
88999	No

By exhaustion, we have had listed all the possibility. Notice that there are a total of 15 5-digit positive integers whose sum of digits is 43, and only 3 of them is divisible by 11, hence, the probability that the chosen number is divisible by 11 is $\frac{3}{15} = \boxed{\frac{1}{5}}$. ⌟

Problem 3.7. (Xiangqi). In the game of xiangqi, two armies fight in an attempt to capture the other's king, somewhat similar to chess except the pieces play along the intersections of the squares as opposed to the interiors, as well as some pieces having different conditions on their movement. The vertical lines are called the files and the horizontal lines are called the ranks. The chariot can move as many points as they wish in either the vertical or horizontal direction, just as a rook does in chess (see figure, here chariot is written as "car"). The chariot makes captures by moving to the same intersection as the opposing piece. Suppose that 9 chariots are placed on the board at random, what is the probability that no piece will be en-prise? That is, there is no pair of chariots which share the same rank or file?



Sol'n. In observing the diagram, one would notice that there are 10 ranks and 9 files, so one possible arrangement is as follows:



Specifically, we can encode the information of each arrangement by inspecting the list of the horizontal axis of the chesses (since there are 9 files, the vertical coordinate cannot be manipulated). For example, the above arrangement can be denoted as:

$$[1,3,7,4,9,5,2,8,0].$$

Hence, out of all the possibilities, the all cases that there is no piece that will be en-prise is:

$$\#(\text{no piece that will be en-prise}) = (10)_9 = \frac{10!}{(10-9)!} = 10!.$$

Then, since there are in total $10 \times 9 = 90$ intersections, and we are choosing 9 of them to put a chess on, then the total number of arrangements is:

$$\#(\text{arrangements}) = \binom{90}{9} = \frac{90!}{9!(90-9)!} = \frac{90 \times 89 \times \cdots \times 82}{9!}.$$

Thus, the probability that no piece will be en-prise is:

$$P(\text{no piece will be en-prise}) = \frac{\#(\text{no piece will be en-prise})}{\#(\text{arrangements})} = \frac{10!}{90 \times 89 \times \cdots \times 82/9!}$$
$$= \frac{10! \times 9!}{90 \times 89 \times \cdots \times 82} = \boxed{\frac{362\,880}{70\,625\,252\,863}}.$$

┘

Problem 3.8. (Basic Techniques). Calculate the probability of the given event under the stated assumptions.

(a) If $P(A) = 0.1$ and $P(B) = 0.5$, and A and B are disjoint, calculate $P(A' \cup B)$.

(b) If $P(C) = 0.25$, $P(D) = 0.4$ and $P(C' \cap D') = 0.55$, calculate $P[(C \cap D)']$.

(c) If $P(E \cup F) = 0.5$ and $P(E \cup F') = 0.7$, calculate $P(E)$.

Sol'n. For this problem, we assume that X' denotes the complementary of the set X, always written as X^c. Moreover, for the simplicity of notation, we denote S as the sample space.

(a) Since A and B are disjoint, let $b \in B$ be arbitrary, then $b \notin A$, which implies that $b \in S \backslash A = A^c$, and hence implying that $B \subseteq A^c$, so we have $A^c \cup B = A^c$, and thus:

$$P(A^c \cup B) = P(A^c) = 1 - P(A) = 1 - 0.1 = \boxed{0.9}.$$

(b) By DeMorgan's law, we have that $(C \cap D)^c = C^c \cup D^c$, hence giving us that:

$$\begin{aligned} P[(C \cap D)^c] &= P(C^c \cup D^c) = P(C^c) + P(D^c) - P(C^c \cap D^c) \\ &= [1 - P(C)] + [1 - P(D)] - P(C^c \cap D^c) = 0.75 + 0.6 - 0.55 = \boxed{0.8}. \end{aligned}$$

(c) Notice that:

$$\begin{cases} (E \cup F^c) \cup (E \cup F) = (E \cup E) \cup (F^c \cup F) = E \cup S = S; \\ (E \cup F^c) \cap (E \cup F) = E \cup (F^c \cap F) = E \cup \emptyset = E; \end{cases}$$

thus, we then know that:

$$\begin{aligned} 1 = P(S) &= P(E \cup F) + P(E \cup F^c) - P((E \cup F^c) \cap (E \cup F)) = 0.5 + 0.7 - P(E) \\ P(E) &= 0.5 + 0.7 - 1 = \boxed{0.2}. \end{aligned}$$

┘

Problem 3.9. (Sports Viewers). A survey of a group's sports viewing habits over the last year revealed the following information:

Football	29%	Football and Baseball	14%
Baseball	28%	Baseball and Hockey	10%
Hockey	19%	Hockey and Football	12%
All Three Sports	8%		

(a) What is the percentage of the group that view none of the three sports?

(b) If a randomly selected baseball viewer of this group is chosen, calculate the probability that they also view football and hockey. Hint: The condition in the problem gives a new sample space of only baseball viewers.

Sol'n.

(a) The percentage of the group that at least one of the three sports is:

$$29\% + 28\% + 19\% - 14\% - 10\% - 12\% + 8\% = 48\%.$$

Hence, the percentage of the group that view none of the three sports is:

$$1 - 48\% = \boxed{52\%}.$$

(b) Here, we are looking for the ratio of people viewing all three sports from all the people who view baseball, that is:

$$\begin{aligned} P(\text{viewing all three sports} \mid \text{viewing baseball}) &= \frac{P(\text{viewing all three sports and baseball})}{\text{viewing baseball}} \\ &= \frac{8\%}{28\%} \times 100\% \approx \boxed{28.57142857\%}. \end{aligned}$$

⌟

4 Conditional Probability and Independence

Problem 4.1. (Mr. Smith and His Strange Kids). Mr. Smith has two children. Assume that having a boy or girl is equally likely and independent. For each question below, draw a tree diagram to represent the sample space before solving.

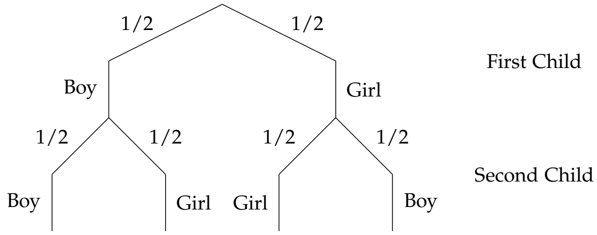
(a) Suppose Mr. Smith tells you at least one of his children is a boy, what is the probability that both of his children are boys?

(b) Suppose you go to Mr. Smith's house to visit. When you knock on the door, his two children flip a fair coin to see who has to answer the door. If the child who answers the door is a boy, what is the probability that Mr. Smith has two boys?

(c) Explain why the answers in parts (a) and (b) are different.

Sol'n.

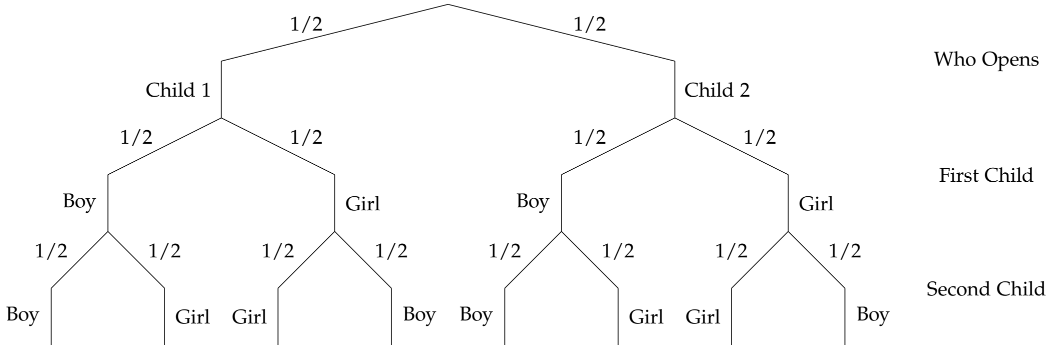
(a) First, we can illustrate the tree diagram, as follows:



Here, we denote having at least one of his children is a boy as X and having both of his children are boys as Y. Note that if both of his children are boys implies that at least one is boy, so $Y \subseteq X$. Therefore, the probability of both of his children are boys given that at least one of his children is a boy is:

$$P(Y|X) = \frac{P(X \cap Y)}{P(X)} = \frac{P(Y)}{P(X)} = \frac{\frac{1}{2} \times \frac{1}{2}}{\frac{1}{2} + \frac{1}{2} \times \frac{1}{2}} = \frac{\frac{1}{4}}{\frac{3}{4}} = \boxed{\frac{1}{3}}.$$

(b) Likewise, we can first illustrate the tree diagram, as follows:



Here, we denote having the boy answering the door after fair coin as X and having both of his children are boys as Y. Note that if a boy answers implies that at least one is boy, so $Y \subseteq X$.

Therefore, the probability of both of his children are boys given that the child answering after the fair coin flip is a boy is:

$$P(Y|X) = \frac{P(X \cap Y)}{P(X)} = \frac{P(Y)}{P(X)} = \frac{\frac{1}{2} \times \frac{1}{2} \times \frac{1}{2} + \frac{1}{2} \times \frac{1}{2} \times \frac{1}{2}}{\frac{1}{2} \times \frac{1}{2} + \frac{1}{2} \times \frac{1}{2} \times \frac{1}{2} + \frac{1}{2} \times \frac{1}{2} \times \frac{1}{2}} = \frac{\frac{1}{4}}{\frac{1}{2}} = \boxed{\frac{1}{2}}.$$

(c) The answers in (a) and (b) are different due to the amount of information given. From (a), as long as there is one boy, then a `true` response will be given, while from (b), when there is only one boy, the children answering the door being a boy has 50% chance being `true` and the other half being `false`. Since when there are exactly one boy and one girl, the two responses will vary, this assumption does not generate the same probability for both children being boys.

⌟

Problem 4.2. (Lost Luggage). You and a friend fly on an airplane, event A is that they lose your luggage, and event B is that they lose your friend's luggage. Suppose $P(A) = 0.6$ and $P(B) = 0.7$, and suppose that A and B are independent. Compute the probability of the event that they lose someone's luggage, and the probability of the event that neither luggage is lost.

Sol'n. Since we assume that the two events are independent, this implies that:

$$P(A \cap B) = P(A) \times P(B) = 0.6 \times 0.7 = 0.42.$$

Therefore, the probability that they lose someone's luggage is:

$$P(A \cup B) = P(A) + P(B) - P(A \cap B) = 0.6 + 0.7 - 0.42 = \boxed{0.88}.$$

Hence, the probability that neither luggages is lost is the probability of the converse event as of above, that is:

$$P(A^c \cap B^c) = P[(A \cup B)^c] = 1 - P(A \cup B) = 1 - 0.88 = \boxed{0.12}.$$

⌟

Problem 4.3. (Independent Dice). You roll two 6-sided fair dice, one after the other. Consider the events:

- $A =$ the first die shows a number strictly greater than the second die;
- $B =$ the second die shows a 3 or a 4;
- $C =$ the first die shows a 1.

Note that there are three possible pairs of events: (A, B), (A, C), and (B, C).

(a) Compute $P(A)$, $P(B)$, and $P(C)$.

(b) Which pairs of events are independent?

(c) Which of the above three pairs of events are mutually exclusive?

Sol'n.

(a) For simplicity, we denote (x, y) as the first dice rolls to x and second rolls to y, therefore, we can list all possibilities in the following table:

(1,1)	(1,2)	(1,3)	(1,4)	(1,5)	(1,6)
(2,1)	(2,2)	(2,3)	(2,4)	(2,5)	(2,6)
(3,1)	**(3,2)**	(3,3)	(3,4)	(3,5)	(3,6)
(4,1)	**(4,2)**	**(4,3)**	(4,4)	(4,5)	(4,6)
(5,1)	**(5,2)**	**(5,3)**	**(5,4)**	(5,5)	(5,6)
(6,1)	**(6,2)**	**(6,3)**	**(6,4)**	**(6,5)**	(6,6)

Note that cases satisfying A is **bold**, the cases satisfying B is in blue, and the cases satisfying C is marked by shaded color.

Hence, we can just count the occurrence of each event given that the total sample space has 36 events, so we have:

$$P(A) = \frac{\#(A)}{\#(S)} = \frac{15}{36} = \boxed{\frac{5}{12}},$$

$$P(B) = \frac{\#(B)}{\#(S)} = \frac{12}{36} = \boxed{\frac{1}{3}},$$

$$P(C) = \frac{\#(C)}{\#(S)} = \frac{6}{36} = \boxed{\frac{1}{6}}.$$

(b) Note that we can calculate the probability of intersection of events, as follows:

$$P(A \cap B) = \frac{\#(A \cap B)}{\#(S)} = \frac{5}{36},$$

$$P(B \cap C) = \frac{\#(B \cap C)}{\#(S)} = \frac{2}{36} = \frac{1}{18},$$

$$P(A \cap C) = \frac{\#(A \cap C)}{\#(S)} = \frac{0}{36} = 0.$$

Note that:

$$\begin{aligned} P(A) \times P(B) &= \frac{5}{12} \times \frac{1}{3} = \frac{5}{36} = P(A \cap B), \\ P(B) \times P(C) &= \frac{1}{3} \times \frac{1}{6} = \frac{1}{18} = P(B \cap C), \\ P(A) \times P(C) &= \frac{5}{12} \times \frac{1}{6} = \frac{5}{72} \neq P(A \cap C). \end{aligned}$$

From the above calculation, we know $\boxed{A \text{ and } B \text{ are independent}}$ and $\boxed{B \text{ and } C \text{ are independent}}$. For A, C, since A happening implies that the first draw is 1, which implies that it is not strictly larger than any number of the dice, thus C is impossible, hence A, C are not independent by definition.

(c) Note that from the above table, the calculation, and logical deduction, event A and C will not happen at the same time, thus $\boxed{A \text{ and } C \text{ are mutually exclusive}}$.

⌟

Problem 4.4. (Jury Duty). 26.8% of adult Baltimore residents have a college degree. If 12 Baltimore residents are selected at random to sit on a jury in a criminal case, what is...

(a) the probability that at least one of them has a college degree?

(b) the probability that *exactly* one of them has a college degree?

Explain any assumptions you made in order to solve this problem.

Sol'n.

(a) Here, we want to calculate the probability of the complementary event, *i.e.*, no one of them has a college degree, that is:

$$P(\text{no one has a college degree}) = (1 - 26.8\%)^{12} \approx 0.02366644208.$$

Note that the above calculation assumes that there are sufficiently large population in Baltimore such that the selection of one people with or without college degree has negligible affect on the probability of the next trial. Note that we have concluded in class that the probability could be arbitrarily close to the original when the population number is very large.

Thus, the probability that at least one has a college degree is:

$$\begin{aligned} P(\text{at least one has a college degree}) &= 1 - P(\text{no one has a college degree}) \\ &= 1 - (1 - 26.8\%)^{12} \approx \boxed{0.9763335579}. \end{aligned}$$

(b) In calculating probability that *exactly* one of them has a college degree, we keep the same assumption from the proceeding part.

Therefore, since the person with the college degree could be the any of them and there are 12 positions (or $\binom{12}{1} = 12$), therefore, the probability is:

$$P(\text{exactly one of them has a college degree}) = 12 \times 26.8\% \times (1 - 26.8\%)^{11} = \boxed{0.1039771554}.$$

⌟

Problem 4.5. (Bases are Loaded and Kasey's At-Bat). At the family wiffle-ball game it's uncle Kasey's turn to bat. He hits a single with probability 0.35, a double with probability 0.25, a triple with probability 0.1, and get's out with probability 0.3.
Once on base, his probability of scoring after a single is 0.2, after a double is 0.3, and after a triple is 0.4. What is the probability that he scores this turn?

Sol'n. The probability that he scores is:

$$\begin{aligned} P(\text{he score}) &= P(\text{hits a single}) \times P(\text{scoring w/ single}) + P(\text{hits a double}) \times P(\text{scoring w/ double}) \\ &\quad + P(\text{hits a triple}) \times P(\text{scoring w/ triple}) \\ &= 0.35 \times 0.2 + 0.25 \times 0.3 + 0.1 \times 0.4 \\ &= 0.07 + 0.075 + 0.04 = \boxed{0.185}. \end{aligned}$$

┘

Problem 4.6. (Ballistic Statistics). Matching fired bullets found at a crime scene to the gun they were fired from remains a popular tool for law enforcement to identify perpetrators. Thus, many studies have been done to measure this accuracy.

In one such study, Accuracy of comparison decisions by forensic firearms examiners, many trials were run where an unknown specimen was sent to firearm examiners along with known comparison specimens. In each trial, the examiner either

- Identified (ID) the unknown specimen, claiming it to be fired from the same gun.
- Eliminated the unknown specimen, claiming it to not be fired from the same gun.
- Could neither eliminate nor ID the unknown specimen (Inconclusive).

The data is summarized in the table below. Each cell in the table counts the number of observations of an examiner's decision under the true situation. E.g. there were 961 trials where an examiner correctly eliminates the unknown specimen as being fired from the same gun when they are not a match.

Examiner Decision:	ID	Inconclusive	Elimination
Match:	1076	288	41
Not a Match:	20	1861	961

Use this data to estimate the probability of

(a) a false-positive (The event that an examiner incorrectly identifies the unknown specimen as being fired from the same gun, even though it was not a match), and

(b) a false-negative (the event that an examiner incorrectly eliminates, even though it was a match).

Sol'n.

(a) First a false positive is when it is identified given that it is not a match. In total, the number of not a match is:

$$20 + 1861 + 961 = 2\,842.$$

Hence, the probability for false-positive is:

$$P(\text{false-positive}) = P(\text{ID}|\text{Not a match}) = \frac{\#(\text{ID and not a match})}{\#(\text{Not a match})} = \boxed{\frac{20}{2\,842}} \approx 0.007037297678.$$

(b) For the other case, it is eliminated given that it is a match. In total, the number of being a match is:

$$1076 + 288 + 41 = 1\,405.$$

Hence, the probability for false-negative is:

$$P(\text{false-negative}) = P(\text{Elimination}|\text{Match}) = \frac{\#(\text{Elimination and match})}{\#(\text{Match})} = \boxed{\frac{41}{1\,405}} \approx 0.02918149466.$$

┘

Problem 4.7. (Marbles in an Urn). There are 3 urns with the containing colored marbles:

Urn	**#of black marbles**	**# of white marbles**
1	2	1
2	3	2
3	4	3

We play the following game:

- I pick an urn at random (each urn equally likely, and you do not get to see which I picked).
- I draw a marble from the urn and reveal it to you.
- If you guess the correct urn from whence the marble came, you win a prize. Say, a high-five or something.

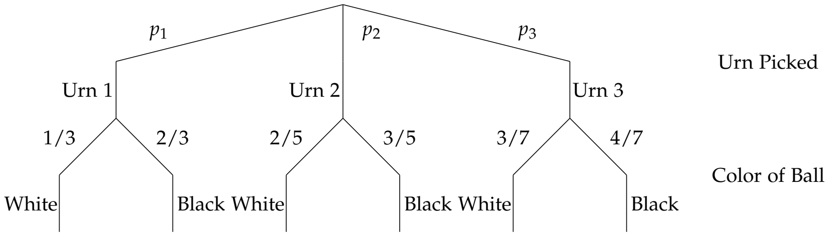
(a) Suppose we play the game and I reveal to you a black marble. Which urn should you choose to maximize your probability of winning? What is that probability?

(b) Suppose now that I *don't* pick the urns with equal probability. Instead, I pick them with probabilities p_1, p_2, p_3 where $0 \leq p_i \leq 1$ represents the probability that I pick Urn i and $p_1 + p_2 + p_3 = 1$. Each time we play the game you will know whether you won or lost, but in the case that you lost, you do not get to know what the correct urn was. If you're allowed to play the game as many times as you'd like (forever, even) then would you be able to figure out the values of p_1, p_2 and p_3? Explain your strategy.

(c) Finally, suppose you've figured out that I'm picking the Urns with probabilities $p_1 = 18/59$, $p_2 = 20/59$, and $p_3 = 21/59$. We play the game one more time and I reveal to you a Black marble. Which urn should you choose to maximize your probability of winning? What is that probability?

Sol'n.

(a) For simplicity, a tree-diagram is drawn for the general case:



Therefore, the probability for winning is the probability that the urn is i ($i \in \{1,2,3\}$) given that the ball picked is black. Here, we denote X_i as the event of urn i is chosen, and we denote Y as the event where a black ball is picked, then:

$$P(X_1|Y) = \frac{P(X_1 \cap Y)}{P(Y)} = \frac{\frac{1}{3} \times \frac{2}{3}}{\frac{1}{3} \times \frac{4}{7} + \frac{1}{3} \times \frac{3}{5} + \frac{1}{3} \times \frac{2}{3}} = \frac{2/3}{4/7 + 3/5 + 2/3} = \frac{70}{193}.$$

$$P(X_2|Y) = \frac{P(X_2 \cap Y)}{P(Y)} = \frac{\frac{1}{3} \times \frac{3}{5}}{\frac{1}{3} \times \frac{4}{7} + \frac{1}{3} \times \frac{3}{5} + \frac{1}{3} \times \frac{2}{3}} = \frac{3/5}{4/7 + 3/5 + 2/3} = \frac{63}{193}.$$

$$P(X_3|Y) = \frac{P(X_1 \cap Y)}{P(Y)} = \frac{\frac{1}{3} \times \frac{4}{7}}{\frac{1}{3} \times \frac{4}{7} + \frac{1}{3} \times \frac{3}{5} + \frac{1}{3} \times \frac{2}{3}} = \frac{4/7}{4/7 + 3/5 + 2/3} = \frac{60}{193}.$$

We shall pick $\boxed{\text{Urn 1}}$ since it has the highest probability for winning, which is $\boxed{\frac{70}{193}}$.

(b) Here, we know that given the three probabilities of picking a urn is constant, and assume that the process is completely random, we know that the probability to win by guessing each urn is, respectively:

$$\begin{cases} P(\text{Urn 1 wins}) &= \dfrac{\frac{2}{3}p_1}{\frac{4}{7}p_3 + \frac{3}{5}p_2 + \frac{2}{3}p_1}, \\ P(\text{Urn 2 wins}) &= \dfrac{\frac{3}{5}p_2}{\frac{4}{7}p_3 + \frac{3}{5}p_2 + \frac{2}{3}p_1}, \\ P(\text{Urn 3 wins}) &= \dfrac{\frac{4}{7}p_3}{\frac{4}{7}p_3 + \frac{3}{5}p_2 + \frac{2}{3}p_1}. \end{cases} \tag{$*$}$$

Note that assuming randomness, as we test for countably (infinite) many times, we can get a good *approximation* of the winning probability for always selecting a urn.

Note that the times needing to test each urn is countable, hence their union is countable. We are still able to enumerate the finite union of countable testings. *In fact, a very easy way is to guess urn 1, then urn 2, and then urn 3* every time when a black ball is picked out, and the result of winning or not should be recorded. When a white ball is drawn out, we can just guess anything as we do not care about the result. In fact, there should be ways to let the probability converging *faster*, but since this question *only* asks for the existence of a strategy, the optimization is, hence, not discussed.

Within a large number of testings, we can eventually get the three probabilities with a good approximation, then $(*)$ becomes a system of three equations with three variables, which is:

$$\begin{cases} \left(\frac{4}{7}p_3 + \frac{3}{5}p_2 + \frac{2}{3}p_1\right) P(\text{Urn 1 wins}) &= \frac{2}{3}p_1, \\ \left(\frac{4}{7}p_3 + \frac{3}{5}p_2 + \frac{2}{3}p_1\right) P(\text{Urn 2 wins}) &= \frac{3}{5}p_2, \\ \left(\frac{4}{7}p_3 + \frac{3}{5}p_2 + \frac{2}{3}p_1\right) P(\text{Urn 3 wins}) &= \frac{4}{7}p_3. \end{cases}$$

By the subtraction of the right hand side of each case, the coefficients should still be linearly independent, and with three functions and three unknown variables, the system linear equations implies the existence of three unique solutions. Hence, by this strategy, we should be able to get a good approximation of each probability of p_1, p_2, and p_3.

Remark: Note that this can only be a approximation, *i.e.*, the results would converge at countably (infinite) trails. This implies that for any finitely many steps, the probability could still variate from the actual value. Hence, the values of p_1, p_2, and p_3 can be approximated to arbitrary close to the actual value, but getting the exact value is not possible for any finitely many steps (*and it is impossible to test for infinitely many times too*).

(c) Here, using our formula from $(*)$, we have:

$$\begin{aligned} P(\text{Urn 1 wins}) &= \frac{\frac{2}{3}\times\frac{18}{59}}{\frac{4}{7}\times\frac{21}{59}+\frac{3}{5}\times\frac{20}{59}+\frac{2}{3}\times\frac{18}{59}} = \frac{1}{3}, \\ P(\text{Urn 2 wins}) &= \frac{\frac{3}{5}\times\frac{20}{59}}{\frac{4}{7}\times\frac{21}{59}+\frac{3}{5}\times\frac{20}{59}+\frac{2}{3}\times\frac{18}{59}} = \frac{1}{3}, \\ P(\text{Urn 3 wins}) &= \frac{\frac{4}{7}\times\frac{21}{59}}{\frac{4}{7}\times\frac{21}{59}+\frac{3}{5}\times\frac{20}{59}+\frac{2}{3}\times\frac{18}{59}} = \frac{1}{3}. \end{aligned}$$

In this case, picking $\boxed{\text{any urn is the same}}$, with a probability wining being $\boxed{\frac{1}{3}}$.

$\lrcorner$

Problem 4.8. (Diagnostic Testing). Assume that a rare disease is present in 0.05% of the population. For each individual of the population, the disease is either present (D) or not (D'). There is a test for this disease with **sensitivity** (i.e. the probability of a positive $(+)$ test result, conditioned on the disease actually being present in the individual) is given by

$$P(+|D) = 0.99.$$

On the other hand, the **specificity** (i.e. the probability of a negative $(-)$ test result, conditioned on the disease not being present in the individual) is

$$P(-|D') = 0.95.$$

Calculate the probability that someone has the disease given that they've tested positive.

Sol'n. Here, we assume that the population has size 1 000 000 and perfect distribution. Therefore, there will be 500 people who have the disease.

Assume that every individual is tested, for the 500 people who have the disease, $500 \times 0.99 = 495$ people will be tested positive.

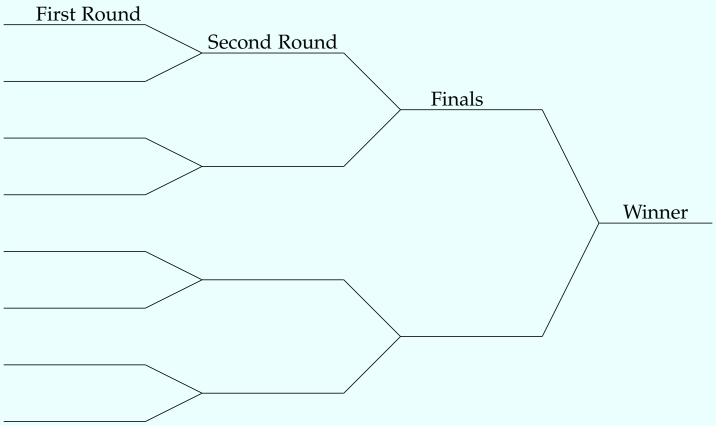
For the 999 500 people who do not have the disease, $999\ 500 \times (1 - 0.95) = 49\ 975$ will be tested positive.

Hence, if an individual is tested positive, the probability that they has the disease is:

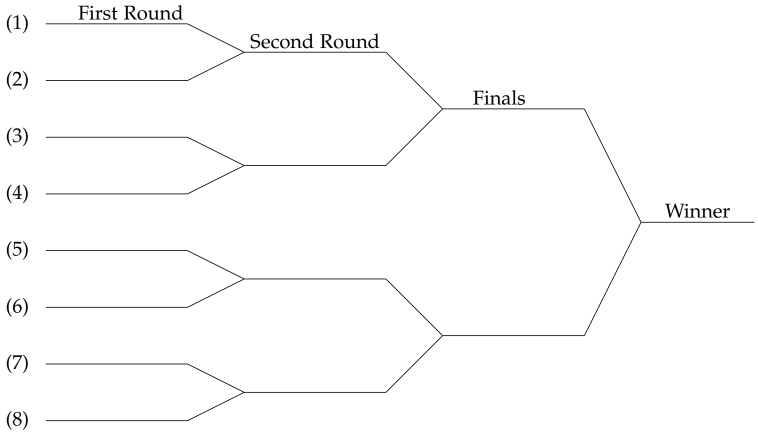
$$\frac{495}{49\ 975 + 495} = \boxed{\frac{495}{50\ 470}} \approx 0.9807806618\%,$$

which is a quite small probability. ⌟

Problem 4.9. (Knights of the Round Table). King Arthur holds a jousting tournament among 8 knights. Each joust is in pairs, and the initial parings of the first round are random. Each knight is evenly matched and there are no ties. Among the 8 knights are the twins Balin and Balan, what is the probability that the twins meet in the tournament?



Sol'n. First, we want to enumerate the player with an order, as of follows:



Then, we want to get some conclusions for the general results:

- If the two players are both in $\{1,2\}$, $\{3,4\}$, $\{5,6\}$, or $\{7,8\}$, they are guaranteed to meet each other (in which they meet at first round).
- Otherwise, if the two players are in $\{1,2,3,4\}$ or $\{5,6,7,8\}$, they have a probability of $1/2 \times 1/2 = 1/4$ to meet each other (in which they meet at second round).
- Otherwise, they have a probability of $1/2 \times 1/2 \times 1/2 \times 1/2 = 1/16$ to meet each other (in which they meet at finals).

Thus, we shall then consider if Balin and Balan might meet in which round.

- For players are both in $\{1,2\}$, $\{3,4\}$, $\{5,6\}$, or $\{7,8\}$, the probability that this happens is $1/7$.
- Otherwise, for players are both in $\{1,2,3,4\}$ or $\{5,6,7,8\}$, the probability that this happens is $2/7$.

- Otherwise, the probability of the remaining happens is 4/7.

Therefore, the probability for Balin and Balan to meet is:

$$1 \times \frac{1}{7} + \frac{1}{4} \times \frac{2}{7} + \frac{1}{16} \times \frac{4}{7} = \frac{4+2+1}{28} = \frac{7}{28} = \boxed{\frac{1}{4}}.$$

⌟

5 Discrete Random Variables

Problem 5.1. (Biased Dice). Your friend gives you a loaded 6-sided dice. This dice lands on 6 twice as often as it lands on 4, and lands on 4 twice as often as it lands on 2. The dice lands on 5 with the same frequency as 6, 3 with the same frequency as 4, and 1 with the same frequency as 2. Let X be the number that the dice lands on when rolled a single time.

(a) Find the pmf of X.

(b) Find the expectation of X.

Sol'n.

(a) Here we can define the probability mass function as: $f : \mathbb{Z} \to [0, 1]$.
Note that the dice has six sides, enumerated from 1 to 6, only the sides of 1 to 6 can have a non-zero probability. Here, we set the probability of landing on 2 as x, then we have landing on 1 with probability same as x, we have probability of landing on 3 and 4 as $2x$, and on 5 and 6 as $2(2x) = 4x$. Since all these probability must add up to 1, then we have:

$$x + x + 2x + 2x + 4x + 4x = 1 \Longrightarrow 14x = 1 \Longrightarrow x = \frac{1}{14}.$$

Hence, our probability mass function can be defined as:

$$f(x) = P(X = x) = \boxed{\begin{cases} \dfrac{1}{14}, & x = 1 \text{ or } 2; \\ \dfrac{1}{7}, & x = 3 \text{ or } 4; \\ \dfrac{2}{7}, & x = 5 \text{ or } 6; \\ 0, & \text{otherwise.} \end{cases}}$$

(b) Then, we can calculate the expectation of X as:

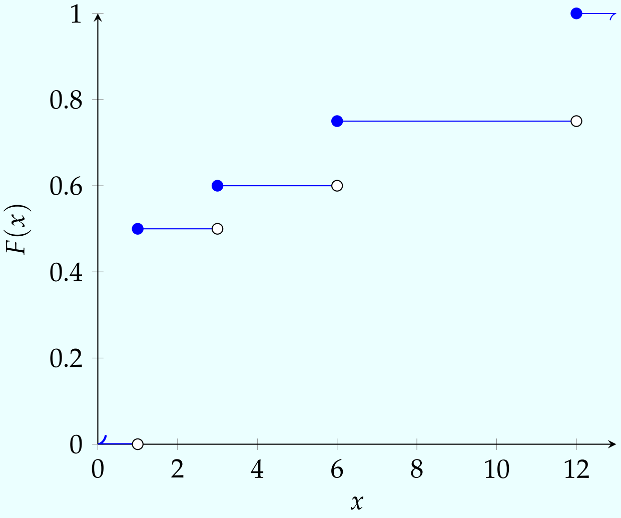
$$\begin{aligned} \mathbb{E}(X) = \mu_X &= \sum_{x \in \text{range}(X)} x \cdot f(x) \\ &= \frac{1}{14} \times 1 + \frac{1}{14} \times 2 + \frac{1}{7} \times 3 + \frac{1}{7} \times 4 + \frac{2}{7} \times 5 + \frac{2}{7} \times 6 + 0 \\ &= \frac{1 + 2 + 6 + 8 + 20 + 24}{14} \\ &= \boxed{\frac{61}{14}} \approx 4.3571428571. \end{aligned}$$

⌟

Problem 5.2. (Payment Options). A technology company offers 4 different payment plans for their subscription services:

payment due every month, payment due every 3 months,
payment due every 6 months, payment due every year.

Let X be the number of months between payments of a random customer of this technology company. E.g. $X = 12$ for a customer who pays every year, and $X = 1$ for a customer who pays every month. Below is the graph of the cdf, $F(x) = P(X \leq x)$:



(a) What is the range of X?

(b) What percentage of customers have less than 9 months between payments? What percentage have more than 5 months between payments?

(c) What proportion of customers pay every month?

(d) For each x in the range of X, evaluate the pmf $f(x) = P(X = x)$.

(e) Calculate $\mathbb{E}(X)$.

Sol'n.

(a) Note that the range of X includes is the number of months between payments, and there are 4 plans with x being 1, 3, 6, and 12. Hence, the range is:

$$\text{range}(X) = \boxed{\{1, 3, 6, 12\}}.$$

(b) First, since there is not an option of paying every 9 month, so the probability of paying every 9 month is $P(X = 9) = 0$, thus:

$$P(X < 9) = P(X \leq 9) = F(9) = 0.75 = 0.75 \times 100\% = \boxed{75\%}.$$

Likewise, since there is also not an option of paying every 5 month, so the probability of paying every 5 month is $P(X = 5) = 0$, thus:

$$P(X \leq 5) = F(5) = 0.6 = 0.6 \times 100\% = 60\%.$$

Hence, the percentage of having more than 5 months between payments is:

$$P(X > 5) = 1 - 60\% = \boxed{40\%}.$$

(c) Since paying by every month is the minimal month allowed method, thus we have $P(X < 1) = 0$, thus we have:

$$P(X = 1) = P(X \leq 1) = F(1) = 0.5 = \boxed{\frac{1}{2}}.$$

(d) For each $x \in \text{range}(X)$, we have:

- For $x = 1$, we have:
$$f(1) = P(X = 1) = \boxed{0.5}.$$
- For $x = 3$, we have:
$$f(3) = P(X = 3) = F(3) - F(1) = 0.6 - 0.5 = \boxed{0.1}.$$
- For $x = 6$, we have:
$$f(6) = P(X = 6) = F(6) - F(3) = 0.75 - 0.6 = \boxed{0.15}.$$
- For $x = 12$, we have:
$$f(12) = P(X = 12) = F(12) - F(6) = 1 - 0.75 = \boxed{0.25}.$$

In general, we have:

$$f(x) = \begin{cases} 0.5 & x = 1, \\ 0.1 & x = 3, \\ 0.15 & x = 6, \\ 0.25 & x = 12, \\ 0 & \text{otherwise.} \end{cases}$$

(e) Then, we can calculate the expectation, as follows:

$$\begin{aligned} \mathbb{E}(X) = \mu_X &= \sum_{x \in \text{range}(X)} x \cdot f(x) \\ &= 0.5 \times 1 + 0.1 \times 3 + 0.15 \times 6 + 0.25 \times 12 \\ &= \boxed{4.7}. \end{aligned}$$

⌟

Problem 5.3. (A Fair Game). You play a game where you roll a fair dice, then flip a coin the same number of times as shown on the dice. You win \$1 each time the coin lands heads, and nothing if the coin lands tails. What is the expectation of your total winnings in this game? *Hint:* Let Y be the number that the dice lands on, and X be the number of heads flipped, can you figure the probability that $X = x$ given that $Y = y$? Then think of all the ways you could win x dollars...

Sol'n. Here, as X denotes the total number of heads rolled, it is equivalent to the total number of money earned. Note that $\text{range}(X) = \{0, 1, 2, 3, 4, 5, 6\}$ since one can win at most \$6 when flipping 6 heads. Technically, we shall be considering the probability mass function evaluated at each $x \in \text{range}(X)$, but since $x = 0$ has no impact on the expectation, we do not need to calculate that.

For simplicity of notation, we define the following distribution on a random variable of Z:

$$Z_n \sim \text{Binomial}\left(n, \frac{1}{2}\right),$$

which corresponding to getting how many heads for flipping the coin n times.

Thus, the consideration of all cases are:

- $X = 1$, this case can be achieve by having exactly one head when the dice rolls to any number, thus its probability is:

$$\begin{aligned} P(X = 1) &= \sum_{i=1}^{6} P(Y = i) \times P(Z_i = 1) \\ &= \frac{1}{6} \times \binom{1}{1} \times \left(\frac{1}{2}\right)^1 + \frac{1}{6} \times \binom{2}{1} \times \left(\frac{1}{2}\right)^1 \times \left(\frac{1}{2}\right)^1 + \frac{1}{6} \times \binom{3}{1} \times \left(\frac{1}{2}\right)^1 \times \left(\frac{1}{2}\right)^2 \\ &\quad + \frac{1}{6} \times \binom{4}{1} \times \left(\frac{1}{2}\right)^1 \times \left(\frac{1}{2}\right)^3 + \frac{1}{6} \times \binom{5}{1} \times \left(\frac{1}{2}\right)^1 \times \left(\frac{1}{2}\right)^4 + \frac{1}{6} \times \binom{6}{1} \times \left(\frac{1}{2}\right)^1 \times \left(\frac{1}{2}\right)^5 \\ &= \frac{1}{6}\left(\frac{1}{2} + \frac{1}{2} + \frac{3}{8} + \frac{1}{4} + \frac{5}{32} + \frac{3}{32}\right) \\ &= \frac{1}{6} \times \frac{15}{8} = \frac{5}{16}. \end{aligned}$$

- $X = 2$, this case can be achieve by having exactly two head when the dice rolls to a number greater than 1, thus its probability is:

$$\begin{aligned} P(X = 2) &= \sum_{i=2}^{6} P(Y = i) \times P(Z_i = 2) \\ &= \frac{1}{6} \times \binom{2}{2} \times \left(\frac{1}{2}\right)^2 + \frac{1}{6} \times \binom{3}{2} \times \left(\frac{1}{2}\right)^2 \times \left(\frac{1}{2}\right)^1 + \frac{1}{6} \times \binom{4}{2} \times \left(\frac{1}{2}\right)^2 \times \left(\frac{1}{2}\right)^2 \\ &\quad + \frac{1}{6} \times \binom{5}{2} \times \left(\frac{1}{2}\right)^2 \times \left(\frac{1}{2}\right)^3 + \frac{1}{6} \times \binom{6}{2} \times \left(\frac{1}{2}\right)^2 \times \left(\frac{1}{2}\right)^4 \\ &= \frac{1}{6}\left(\frac{1}{4} + \frac{3}{8} + \frac{3}{8} + \frac{5}{16} + \frac{15}{64}\right) \\ &= \frac{1}{6} \times \frac{99}{64} = \frac{33}{128}. \end{aligned}$$

- $X = 3$, this case can be achieve by having exactly three head when the dice rolls to a number greater than 2, thus its probability is:

$$\begin{aligned}
P(X=3) &= \sum_{i=3}^{6} P(Y=i) \times P(Z_i=3) \\
&= \frac{1}{6} \times \binom{3}{3} \times \left(\frac{1}{2}\right)^3 + \frac{1}{6} \times \binom{4}{3} \times \left(\frac{1}{2}\right)^3 \times \left(\frac{1}{2}\right)^1 + \frac{1}{6} \times \binom{5}{3} \times \left(\frac{1}{2}\right)^3 \times \left(\frac{1}{2}\right)^2 \\
&\quad + \frac{1}{6} \times \binom{6}{3} \times \left(\frac{1}{2}\right)^3 \times \left(\frac{1}{2}\right)^3 \\
&= \frac{1}{6}\left(\frac{1}{8} + \frac{1}{4} + \frac{5}{16} + \frac{5}{16}\right) \\
&= \frac{1}{6} \times 1 = \frac{1}{6}.
\end{aligned}$$

- $X = 4$, this case can be achieve by having exactly four head when the dice rolls to a number greater than 3, thus its probability is:

$$\begin{aligned}
P(X=4) &= \sum_{i=4}^{6} P(Y=i) \times P(Z_i=4) \\
&= \frac{1}{6} \times \binom{4}{4} \times \left(\frac{1}{2}\right)^4 + \frac{1}{6} \times \binom{5}{4} \times \left(\frac{1}{2}\right)^4 \times \left(\frac{1}{2}\right)^1 + \frac{1}{6} \times \binom{6}{4} \times \left(\frac{1}{2}\right)^4 \times \left(\frac{1}{2}\right)^2 \\
&= \frac{1}{6}\left(\frac{1}{16} + \frac{5}{32} + \frac{15}{64}\right) \\
&= \frac{1}{6} \times \frac{29}{64} = \frac{29}{384}.
\end{aligned}$$

- $X = 5$, this case can be achieve by having exactly five head when the dice rolls to a number greater than 4, thus its probability is:

$$\begin{aligned}
P(X=5) &= \sum_{i=5}^{6} P(Y=i) \times P(Z_i=5) \\
&= \frac{1}{6} \times \binom{5}{5} \times \left(\frac{1}{2}\right)^5 + \frac{1}{6} \times \binom{6}{5} \times \left(\frac{1}{2}\right)^5 \times \left(\frac{1}{2}\right)^1 \\
&= \frac{1}{6}\left(\frac{1}{32} + \frac{3}{32}\right) \\
&= \frac{1}{6} \times \frac{1}{8} = \frac{1}{48}.
\end{aligned}$$

- $X = 6$, this case can be achieved only if the dice rolls to 6 and there are six heads, thus its probability is:

$$P(X=6) = \frac{1}{6} \times \left(\frac{1}{2}\right)^6 = \frac{1}{6} \times \frac{1}{64} = \frac{1}{384}.$$

Then, with all the probability mass functions, we are ready to calculate the expectation on the winning, which is:

$$\begin{aligned}
\mathbb{E}(X) = \mu_X &= \sum_{x \in \text{range}(X)} x \cdot f(x) \\
&= 1 \times \frac{5}{16} + 2 \times \frac{33}{128} + 3 \times \frac{1}{6} + 4 \times \frac{29}{384} + 5 \times \frac{1}{48} + 6 \times \frac{1}{384} \\
&= \boxed{\frac{7}{4}}
\end{aligned}$$

Remark. Note that this results corresponds to our intuitions. When rolling a dice, the expectation shall be 3.5 and out of all the rollings, flipping a choice shall give you half of the total number of flips, which is $3.5 \div 2 = 1.75$, which aligns with our case-wise discussions. ⌟

Problem 5.4. (Bull's-EyeBull's-Eye). A player throws darts at a target. On each trial, independently of the other trials, she hits the bull's-eye with probability 0.05. Just how many times should she throw so that her probability of hitting the bull's-eye (at least once) is at least 0.5?

Sol'n. Here, we could be considering the complementary event, which is the probability that the player never hits the bull's-eye for all n trials, this probability is:

$$P(\text{hitting none for } n \text{ trials}) = (1-0.05)^n = 0.95^n.$$

Note that for having 0.5 probability of hitting the bull's-eye at least once implies 0.5 probability of never hitting. Hence, we want to find the minimum n such that $0.95^n \leq 0.5$. We can solve the equality case as:

$$0.95^n = 0.5 \Longrightarrow n = \log_{0.95}(0.5) \approx 13.5134073340.$$

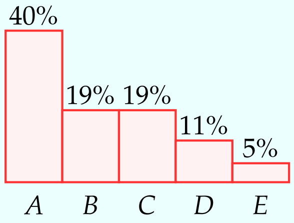
Hence, she needs to throw $\boxed{14}$ times to have a probability of hitting the bull's-eye (at least once) being at least 0.5. ⌟

Problem 5.5. (Recognizing Distributions). Identify the approximate distribution of the each random variable X in the experiments below as either Binomial(n, p), Poisson(λ), or HyperGeometric(N, K, n), and give the numerical parameters in case (E.g. don't just say Binomial, write Binomial$(10, 0.5)$ if that's what the distribution is). Explain any assumptions you make about the experiment in order to approximate the distributions.

(a) In 2022, the average SAT math score was 521. Suppose a random sample of 100 high school students who took the SAT in 2022 is collected. Let X be the number of students in the sample who scored above average on the math portion.

(b) Each seedling of a particular species of tree has a zone of resource depletion around it. If this zone overlaps with that of other seedlings in the vicinity, it can be detrimental to the growth of both. When seeds are randomly dispersed over a wide area, the number of neighbors that a given seedling may have is a random variable X with mean $4/\text{m}^2$.

(c) In a very large community, different drivers may one of five differing opinion on what road fixes should be made to improve commutes. They are A = *Improved road conditions*, B = *Better traffic signals*, C = *More highway lanes*, D = *Reduced construction delays*, and E = *Improved signage*. The percentages of the population having such opinions are collected in the histogram below.



Two different samples of the population are collected, the first is a sample of size 15, and Y is the number of individuals in the sample who think *better traffic signals* are the best way to improve roads. The second sample is of size 20, and Z is the number of individuals in the sample who think *more highway lanes* are the best way to improve roads. We are interested in the distribution of $X = Y + Z$.

(d) 12 refrigerators of a certain type has been returned to a distributor because of the presence of a high-pitched oscillating noise. Suppose that 5 of these 12 have defective compressors and the other 7 have less serious problems. If they are examined in random order, let X be the number among the first 6 examined that have a defective compressor.

Sol'n.

(a) Here, we may assume that there are sufficiently many students so that picking a student out does not impact the probability of next students having a math score above average. Here, we denote $S_i \sim$ Bernoulli(p), where p is the probability that a random student scores above average on SAT math.

Therefore, we can consider X as:

$$X = \sum_{i=1}^{100} S_i = S_1 + S_2 + \cdots + S_{100},$$

in which S_i are assumed to be independent since the population size is large enough, hence have:

$$X \sim \boxed{\text{Binomial}(100, p)},$$

where p is the probability (or proportion) of student have SAT score above average.

Remark. If we assume that the SAT Math score is close to a normal distribution (which is fair since there are large enough number of test candidates), the probability of scoring above average should be (approximately) 0.5, which makes $X \sim \boxed{\text{Binomial}(100, 0.5)}$.

(b) Note that we were given the average is $4/\text{ m}^2$. Here, we assume that the location of any given seed is independent with the location of the other seed, and the number shall follow a Poisson Distribution, that is $X \sim \text{Poisson}(\lambda)$. Since $\mathbb{E}(X) = 4 = \lambda$, then we have:

$$X \sim \boxed{\text{Poisson}(4)}.$$

(c) Here, we still assume that the sample size is big enough so that the samples are independent, so we can conclude Y and Z as:

$$Y = \text{Binomial}(15, 0.19),$$
$$Z = \text{Binomial}(20, 0.19).$$

Here, we can denote $S_i \sim \text{Bernoulli}(0.19)$ as a distribution for an individual, then we have:

$$X = Y + Z \sim \underbrace{S_1 + \cdots + S_{15}}_{\text{components of } Y} + \underbrace{S_{16} + \cdots + S_{35}}_{\text{components of } Z} \sim \boxed{\text{Binomial}(35, 0.19)}.$$

(d) Since the sample size is very much limited, the probability cannot remain unchanged after picks. Here, we assume that the order is random. Hence, we consider having a defective compressor as *success*, hence, the distribution of *success* is:

$$X \sim \boxed{\text{Hyp. Geo.}(12, 5, 6)}.$$

⌟

Problem 5.6. (Plant Density). Refer back to **Problem 5.5** (b):

(a) What is the probability that a given seedling has no neighbors within 1 m^2?

(b) What is the probability that a seedling has at most 3 neighbors per m^2?

(c) What is the probability that a seedling has 5 or more neighbors per m^2?

Sol'n. First of all, from **Problem 5.5** (b), we have the distribution as:

$$X \sim \text{Poisson}(4).$$

(a) For no neighbors, this is:

$$P(X=0) = \exp(-\lambda) \times \frac{\lambda^k}{k!} = \exp(-4) \times \frac{4^0}{0!} = \boxed{\exp(-4)} \approx 0.01831563889.$$

(b) For at most 3 neighbors, this is:

$$\begin{aligned} P(X \leq 3) &= P(X=3) + P(X=2) + P(X=1) + P(X=0) \\ &= \exp(-4) \times \frac{4^3}{3!} + \exp(-4) \times \frac{4^2}{2!} + \exp(-4) \times \frac{4^1}{1!} + \exp(-4) \times \frac{4^0}{0!} \\ &= \exp(-4) \times \left(\frac{32}{3} + 8 + 4 + 1\right) \\ &= \boxed{\frac{71}{3}\exp(-4)} \approx 0.4334701204. \end{aligned}$$

(c) For having 5 or more neighbors, we can first find the probability of having strictly less than 5 neighbors, which is:

$$\begin{aligned} P(X<5) = P(X \leq 4) &= P(X \leq 3) + P(X=4) \\ &= \frac{71}{3}\exp(-4) + \exp(-4) \times \frac{4^4}{4!} \\ &= \frac{71}{3}\exp(-4) + \exp(-4) \times \frac{32}{3} \\ &= \frac{103}{3}\exp(-4). \end{aligned}$$

Then, as complementary event, we have:

$$P(X \geq 5) = 1 - P(X<5) = \boxed{1 - \frac{103}{3}\exp(-4)} \approx 0.3711630648.$$

┘

Problem 5.7. (Road Improvements). Refer back to **Problem 5.5** (c):

(a) What is $P(X \leq 4)$?

(b) What is the largest c for which $P(X \leq c) \leq 1/2$?

(c) What is the probability that 2 individuals from the first sample think that *better traffic signals* and 3 individuals from the second sample think that *more highway lanes* is the best way to improve roads?

Sol'n. Note that from **Problem 5.5** (c), we have the distribution as:

$$X \sim \text{Binomial}(35, 0.19).$$

(a) The probability for $X \leq 4$ is:

$$\begin{aligned} P(X \leq 4) &= P(X=0) + P(X=1) + P(X=2) + P(X=3) + P(X=4) \\ &= \binom{35}{0}(0.19)^0(1-0.19)^{35} + \binom{35}{1}(0.19)^1(1-0.19)^{34} + \binom{35}{2}(0.19)^2(1-0.19)^{33} \\ &\quad + \binom{35}{3}(0.19)^3(1-0.19)^{32} + \binom{35}{4}(0.19)^4(1-0.19)^{31} \\ &= 0.81^{35} + 35 \times 0.19 \times 0.81^{34} + 595 \times 0.19^2 \times 0.81^{33} + 6545 \times 0.19^3 \times 0.81^{32} \\ &\quad + 52360 \times 0.19^4 \times 0.81^{31} \\ &\approx \boxed{0.1785353750}. \end{aligned}$$

(b) Although we have not proven the relationship between the expectation (mean) and median of the distribution, it could be solid guess to check the cumulative mass function around the expectation. By the formula, we have:

$$\mathbb{E}(X) = n \cdot p = 35 \times 0.19 = 6.65.$$

Observe that:

$$\begin{aligned} P(X \leq 6) &= \sum_{i=0}^{6} P(X = i) \approx 0.4923965041, \\ P(X \leq 7) &= \sum_{i=0}^{7} P(X = i) \approx 0.6570295639. \end{aligned}$$

Since the cumulative mass function is monotonically increasing, this implies that the largest c for which $P(X \leq c) < 1/2$ is $\boxed{6}$.

(c) Here, we need the specific distribution again:

$$\begin{aligned} Y &= \text{Binomial}(15, 0.19), \\ Z &= \text{Binomial}(20, 0.19). \end{aligned}$$

Thus, the probability of such event happening is:

$$\begin{aligned} P(Y=2) \times P(Z=3) &= \binom{15}{2}(0.19)^2(1-0.19)^{13} \times \binom{20}{3}(0.19)^3(1-0.19)^{17} \\ &= 105 \times 0.19^2 \times 0.81^{13} \times 1140 \times 0.19^3 \times 0.81^{17} \\ &\approx \boxed{0.05326141762}. \end{aligned}$$

┘

Problem 5.8. (Is Your Refrigerator Running?). Refer back to **Problem 5.5** (d):

(a) What is the probability that exactly 2 of the first 6 refrigerators examined have defective compressors.

(b) What is the probability that at least half of the first 6 refrigerators examined have defective compressors.

(c) What is the expected number of the first 6 refrigerators examined which do *not* have defective compressors.

Sol'n. Recall from **Problem 5.5** (d), we have the distribution as:

$$X \sim \text{Hyp. Geo.}(12, 5, 6).$$

(a) For exactly 2 having defective compressors, the probability is:

$$\begin{aligned} P(X=2) &= \frac{\binom{5}{2} \times \binom{12-5}{6-2}}{\binom{12}{6}} = \frac{\binom{5}{2} \times \binom{7}{4}}{\binom{12}{6}} \\ &= \frac{10 \times 35}{924} = \boxed{\frac{25}{66}} \approx 0.3787878788. \end{aligned}$$

(b) We can calculate for the event, which is 3, 4, or 5 of the first 6 refrigerators have defective compressors (since there are no more than 5 that have defective compressors), which is:

$$\begin{aligned} P(X \geq 3) &= P(X=3) + P(X=4) + P(X=5) \\ &= \frac{\binom{5}{3} \times \binom{12-5}{6-3}}{\binom{12}{6}} + \frac{\binom{5}{4} \times \binom{12-5}{6-4}}{\binom{12}{6}} + \frac{\binom{5}{5} \times \binom{12-5}{6-5}}{\binom{12}{6}} \\ &= \frac{10 \times \binom{7}{3} + 5 \times \binom{7}{2} + 1 \times \binom{7}{1}}{\binom{12}{6}} \\ &= \frac{350 + 105 + 7}{924} = \boxed{\frac{1}{2}} = 0.5. \end{aligned}$$

(c) We still consider the complementary. The expectation of X is:

$$\mathbb{E}(X) = \mu_X = 6 \times \frac{5}{12} = 2.5.$$

Hence, for the first 6 refrigerators, we expect 2.5 machines that have defective compressors, hence we expect $6 - 2.5 = \boxed{3.5}$ refrigerators which do not have defective compressors.

┘

Problem 5.9. (Golden Ticket). Augustus Gloop loves chocolate. His local candy store currently stocks 1000 chocolate bars, and he buys 10 of them. Out of those 1000 chocolate bars, there are 5 which contain a golden ticket!

(a) What is the probability that at least one of Augustus' chocolate bars will contain a prize?

(b) Estimate the probability in (a) using a Binomial distribution.

(c) Estimate the probability in (b) using a Poisson distribution.

(d) Briefly explain why the answers in (a), (b), and (c) are all so close together.

Sol'n. For this question, we denote X as the number of tickets that Augustus gets from the purchases.

(a) We consider the complementary event in this case. The probability that none of the chocolate has a golden ticket is:

$$P(X=0) = \frac{995}{1000} \times \frac{994}{999} \times \frac{993}{998} \times \cdots \times \frac{986}{991} = \frac{435\,841\,667\,261}{458\,349\,513\,900}.$$

Hence, its complementary, which is getting at least one prize is:

$$P(X \geq 1) = 1 - \frac{986}{991} = \frac{435\,841\,667\,261}{458\,349\,513\,900} = \boxed{\frac{22\,507\,846\,639}{458\,349\,513\,900}} \approx 0.0491062954.$$

(b) If we are to approximate by binomial distribution, we have:

$$X \sim \text{Binomial}(10, 0.005).$$

Likewise, we calculate the probability of the complementary event, which is:

$$\begin{aligned} P(X=0) &= \binom{10}{0} \times (0.005)^0 \times (0.995)^{10} \\ &= 1 \times 0.995^{995} = 0.995^{995}. \end{aligned}$$

Hence, its complementary, which is getting at least one prize is:

$$P(X \geq 1) = 1 - 0.995^{995} \approx \boxed{0.0488898695}.$$

(c) If we are to approximate by Poisson distribution, we have:

$$X \sim \text{Poisson}(0.005 \times 10) = X \sim \text{Poisson}(0.05).$$

Likewise, we calculate the probability of the complementary event, which is:

$$P(X=0) = \exp(-0.05) \times \frac{0.05^0}{0!} = \exp(-0.05).$$

Hence, its complementary, which is getting at least one prize is:

$$P(X \geq 1) = 1 - \exp(-0.05) \approx \boxed{0.0487705755}.$$

(d) Notice that these numbers are pretty close. First of all, as n is very large, we know that the effects of replacements is small. We can roughly consider the events as independent since:

$$\frac{986}{991} \approx 0.9949545913 \sim 0.995 = \frac{995}{1000}.$$

Hence, this implies that the events are roughly independent (or at least is safe to approximate that they are independent), so results of (a) and (b) should be similar. Likewise, we have concluded that as $n \to \infty$, we have the Binomial distribution converging to Poisson distribution, hence, the results of (b) and (c) should be close.

Therefore, both (a), (b), and (c) should be giving relatively close results.

⌟

Problem 5.10. (Emails). Let X be the number of emails a University professor receives in a week during business hours (8am-5pm, Mon. through Fri.).

(a) What would be a reasonable distribution for X? Explain any assumptions you need to make.

(b) Suppose that last month, the professor received and average of 6 emails per day. Use this information along with part (a) to estimate $\mathbb{E}(X)$ and explain any assumptions you make in your reasoning.

(c) Use part (b) to estimate the probability that the professor receives at least 40 emails this week. You may leave your answer as an expression which can be plugged into a calculator.

Sol'n.

(a) In general the distribution should be approximately:

$$X \sim \boxed{\text{Poisson}(\lambda)},$$

where λ is the number of emails that a University professor receives every week during business hours.

This approximation is made by assuming the the professor randomly receives emails from each students and this does not vary much by the weeks.

Moreover, we would consider receiving an email as an independent event and the rate of receiving should be approximately constant.

(b) Then, considering the average is 6 per day, this is approximately $6 \times 5 = 30$ per week, which implies that our distribution is now:

$$X \sim \text{Poisson}(30).$$

Therefore, the expectation of the number of emails is:

$$\mathbb{E}(X) = \lambda = \boxed{30}.$$

By reaching here, we still make the same assumptions in the preceding part, *i.e.*, the model follows along with the Poisson distribution where the professor gets approximate the same number of emails each week (which is approximately the constant rate).

(c) The probability of the professor receives ar least 40 email is:

$$\begin{aligned} P(X \geq 40) &= 1 - P(X < 40) \\ &= 1 - \sum_{i=0}^{39} P(X = i) \\ &= \boxed{1 - \sum_{i=0}^{39} \exp(-30) \times \frac{30^i}{i!}} \\ &\approx 0.0462530373. \end{aligned}$$

┘

6 The Normal Distribution & CLT

Problem 6.1. (LA Weather). The average daily high temperature in Los Angeles is $\mu = 77^\circ F$ with a standard deviation of $\sigma = 5^\circ F$. Suppose that the temperatures in June closely follow a normal distribution; that is, if we let X be the temperature of a randomly chosen day in June, we have that $X \sim \mathcal{N}(\mu, \sigma)$.

(a) What is the probability of observing an $83^\circ F$ temperature or higher during a randomly chosen day in June?

(b) What is instead the probability of observing a temperature between $70^\circ F$ and $80^\circ F$?

(c) How cold are the coldest 10% of the days during June in Los Angeles?

(d) Let Y be temperature of a randomly chosen day in June measured in Celsius degrees ($^\circ C$) instead of Fahrenheit ($^\circ F$) degrees: that is, $Y = (X - 32) \times 5/9$. What is the probability distribution of Y?

(e) Verify that $P(Y \geq 28.33) = P(X \geq 83)$. Are you surprised? Explain.

(f) Estimate the IQR of the temperature (in $^\circ C$) in June in Los Angeles.

Sol'n.

(a) Here, we can convert the probability to the standard normal distribution:

$$\begin{aligned} P(X \geq 83^\circ F) &= 1 - P(X < 83^\circ F) \\ &= 1 - P\left(Z < \frac{83^\circ F - 77^\circ F}{5^\circ F}\right) \\ &= 1 - P(Z < 1.2) \\ &= 1 - \Phi(1.2) \approx 1 - 0.88493 = \boxed{0.11507}. \end{aligned}$$

(b) Here, we can also convert the results to the standard normal distribution:

$$\begin{aligned} P(70^\circ F < X < 80^\circ F) &= P\left(\frac{70^\circ F - 77^\circ F}{5^\circ F} < Z < \frac{80^\circ F - 77^\circ F}{5^\circ F}\right) \\ &= P(-1.4 < Z < 0.6) \\ &= \Phi(0.6) - \Phi(-1.4) \approx 0.72575 - 0.08076 = \boxed{0.64499}. \end{aligned}$$

(c) Note that the coldest 10% of temperature x must satisfy that:

$$P(X < x) = 0.1,$$

which is equivalent in claiming that:

$$P\left(Z < \frac{x - 77^\circ F}{5^\circ F}\right) = 0.1,$$

which implies that:

$$\frac{x - 77^\circ F}{5^\circ F} = \Phi^{-1}(0.1) \approx -1.28 \Longrightarrow x = -1.28 \times 5^\circ F + 77^\circ F = \boxed{70.6^\circ F}.$$

(d) First, by the linearity of normal distribution, we know that distribution of Y is also normal.

Then, we want to calculate the parameters:

$$\begin{cases} \mu_Y = (\mu_X - 32) \times 5/9 = (77 - 32) \times 5/9 = 25; \\ \sigma_Y = |5/9| \times \sigma_X = 5 \times 5/9 = 25/9. \end{cases}$$

Hence, the distribution of Y is $Y \sim \boxed{\mathcal{N}(25, 25/9)}$.

(e) First, we want to calculate the probability for $Y \geq 28.33$, which is:

$$\begin{aligned} P(Y \geq 28.33) &= P\left(\frac{(X-32) \times 5}{9} \geq 28.33\right) \\ &= P(X \geq 28.33 \times 9/5 + 32) \\ &= P(X \geq 82.994) \\ &= P\left(Z \geq \frac{82.994 - 77}{5}\right) \\ &= P(Z \geq 1.1988) \\ &= 1 - \Phi(1.1988) \approx 1 - 0.88493 = \boxed{0.11507}. \end{aligned}$$

Note that the values corresponds (approximately). Since we have approximately $28.33°C \approx 83°F$ by conversion, so the probability of temperature falling above it should be roughly the same.

(f) To estimate the IQR, we need to know the lower and upper quarter, say q_1 and q_3, respectively, which is:

$q_1 = \Phi^{-1}(0.25) \times 5°F + 77°F \approx -0.67 \times 5°F + 77°F = 73.65°F = (73.65 - 32) \times 5/9°C \approx 23.13889°C,$

$q_3 = \Phi^{-1}(0.75) \times 5°F + 77°F \approx 0.67 \times 5°F + 77°F = 80.35°F = (80.35 - 32) \times 5/9°C = 26.86111°C.$

Therefore, we have the IQR as:

$$IQR = q_3 - q_1 = 26.86111°C - 23.13889°C = \boxed{3.72222°C}.$$

┘

Problem 6.2. (Basic Technique). Let $Z \sim \mathcal{N}(0,1)$.

(a) Find c such that $P(-c < Z \leq c) = 0.86$.

(b) Find a and b if $P(a < Z < b) = 0.3$ and $P(Z > b) = 0.1$.

Sol'n.

(a) Note that by symmetry, we have that:

$$P(0 < Z < c) = \frac{P(-c < Z < c)}{2} 0.86 \div 2 = 0.43,$$
$$P(Z < c) = P(Z \leq 0) + P(0 < Z < c) = 0.5 + 0.43 = 0.93.$$

Thus, by consulting with the standard normal table, we have:

$$c = \Phi^{-1}(0.93) \approx \boxed{1.48}.$$

(b) Here, we can solve them separately.

$$P(Z < b) = 1 - P(Z > b) = 1 - 0.1 = 0.9,$$
$$P(Z < a) = 1 - P(Z > a) = 1 - (P(a < Z < b) + P(Z > b)) = 1 - (0.3 + 0.1) = 1 - 0.4 = 0.6.$$

Thus, by consulting with the standard normal table, we have:

$$a = \Phi^{-1}(0.6) \approx \boxed{0.25},$$
$$b = \Phi^{-1}(0.9) \approx \boxed{1.28}.$$

⌟

Problem 6.3. (Normal distribution). The article "Reliability of domestic waste biofilm reactors" in J. of Envir. Eng. (1995) 785-790 suggests that substrate concentration (mg/cm^3) of influent to a reactor is normally distributed with $\mu = 0.30$ and $\sigma = 0.06$.

(a) What is the probability that the concentration exceeds 0.25?

(b) What is the probability that the concentration is at most 0.10?

(c) Give a number x such that the probability is exactly .05 that the concentration is higher than x.

Sol'n.

(a) First off, let X denote the substrate concentration of influent to a reactor, so it follows that:

$$X \sim \mathcal{N}(0.30, 0.06).$$

Thus, the probability that the concentration exceeds 0.25 is:

$$\begin{aligned} P(X > 0.25) &= P\left(Z > \frac{0.25 - 0.30}{0.06}\right) \approx P(Z > -0.83333) \\ &= 1 - \Phi(-0.83333) = 1 - 0.20327 = \boxed{0.79673}. \end{aligned}$$

(b) Then, for the probability that the concentration being at most 0.10 is:

$$\begin{aligned} P(X \leq 0.10) &= P\left(Z < \frac{0.10 - 0.30}{0.06}\right) \approx P(Z < -3.33333) \\ &= \Phi(-3.33333) \approx \boxed{0.00043}. \end{aligned}$$

(c) Eventually for the x such that $P(X > x) = 0.05$, we have:

$$P\left(Z > \frac{x - 0.30}{0.06}\right) = 0.05,$$

by consulting with the standard normal table, we have:

$$\frac{x - 0.30}{0.06} = \Phi^{-1}(1 - 0.05) = \Phi^{-1}(0.95) \approx 1.64 \Longrightarrow x \approx 1.64 \times 0.06 + 0.30 = \boxed{0.3984(\text{mg}/\text{cm}^3)}.$$

┘

Problem 6.4. (Coke or Pepsi). The amount of soft drink in a bottle is a Normal random variable. Suppose that in 7% of the bottles containing this soft drink there are less than 15.5 ounces, and in 10% of them there are more than 16.3 ounces. What are the mean and standard deviation of the amount of soft drink in a randomly selected bottle?

Sol'n. Here, let X denote the amount of soft drink in a bottle, and let its distribution be:

$$X \sim \mathcal{N}(\mu, \sigma).$$

From the problem, we know the following information:

$$\begin{cases} P(X < 15.5) = 0.07; \\ P(X > 16.3) = 0.10. \end{cases} \implies \begin{cases} P\left(Z < \dfrac{15.5 - \mu}{\sigma}\right) = 0.07; \\ P\left(Z < \dfrac{16.3 - \mu}{\sigma}\right) = 0.90. \end{cases}$$

Then, we can use the standard normal table to observe that:

$$\begin{cases} \dfrac{15.5 - \mu}{\sigma} = \Phi^{-1}(0.07) = -1.48; \\ \dfrac{16.3 - \mu}{\sigma} = \Phi^{-1}(0.90) = 1.28. \end{cases} \implies \begin{cases} 15.5 - \mu = -1.48\sigma; \\ 16.3 - \mu = 1.28\sigma. \end{cases}$$

Therefore, by subtracting the two equations, we have:

$$16.3 - 15.5 = 1.28\sigma - (-1.48\sigma) \implies 0.8 = 2.76\sigma \implies \sigma = \boxed{\frac{20}{69}} \approx 0.28986(\text{ounces}).$$

Then, we want to find the mean, which is:

$$16.3 - \mu = 1.28 \times \frac{20}{69} \implies \mu = 16.3 - 1.28 \times \frac{20}{69} \implies \mu = 16.3 - 1.28 \times \frac{20}{69} = \boxed{\frac{10991}{690}} \approx 15.92899(\text{ounces}).$$

┘

Problem 6.5. (Lefties). Suppose the probability that a randomly selected person is left-handed is 0.10. In a class of 250 students just how many left-handed seats should we have to be 95% sure that no left-handed person goes without a seat?

Sol'n. Note that we can model each individual person $\{X_i\}_{i=1}^{25}$ as:

$$X_i \sim \text{Bernoulli}(0.10).$$

Therefore, we can model the total 250 students as $Y := \sum_{i=1}^{250} X_i$, in which each one of them is independent. Thus, we have:

$$Y \sim \text{Binomial}(250, 0.10).$$

Hence to achieve a 95% sure that no left-handed person goes without a seat, we set the number of left-handed seats needed as n, and we want to achieve that:

$$P(Y < n) \geq 0.95.$$

Note that we are not introduced to a generate CDF formula for binomial distribution, we shall be approximating Y by a normal distribution, say $\tilde{Y} \sim \mathcal{N}(250 \times 0.10, \sqrt{250 \times 0.1 \times 0.9})$, which is $\tilde{Y} \sim \mathcal{N}(25, \sqrt{22.5})$. Hence, now want to have (with the correction) that:

$$P(\tilde{Y} < n + 0.5) \geq 0.95.$$

By consulting with the standard normal table, we have that:

$$\begin{aligned} & P(\tilde{Y} < n + 0.5) = P\left(Z < \frac{n + 0.5 - 25}{\sqrt{22.5}}\right) = 0.95 \\ \Longrightarrow & \frac{n + 0.5 - 25}{\sqrt{22.5}} = \Phi^{-1}(0.95) \\ \Longrightarrow & n = \left\lceil \Phi^{-1}(0.95) \times \sqrt{22.5} + 25 - 0.5 \right\rceil \approx \left\lceil 1.64 \times \sqrt{22.5} + 25 - 0.5 \right\rceil \approx \lceil 32.27920 \rceil = \boxed{33(\text{seats})}. \end{aligned}$$

⌟

Problem 6.6. (Grant Proposals). The Associate Vice Provost of a university must read many research proposals from grant applicants. The amount of time the Provost spends reading any given research proposal is $\mathcal{N}(\mu, \sigma)$ distributed and independent of the time spent on any other proposals. It is known that the Provost must read 49 proposals to ensure that the sample mean is within 3 minutes of μ with probability 0.9. Estimate the standard deviation σ.

Sol'n. Note that we can have the time of reading each proposal as $\{X_i\}_{i=1}^{49}$ as:

$$X_i \sim \mathcal{N}(\mu, \sigma),$$

thus the mean time to read all 49 proposals, can be modeled by $Y := \frac{1}{49}\sum_{i=1}^{49} X_i$, that is:

$$Y \sim \mathcal{N}\left(\mu, \frac{\sigma}{\sqrt{49}}\right) = \mathcal{N}\left(\mu, \frac{\sigma}{7}\right).$$

Give that $P(\mu - 3 < Y < \mu + 3) = 0.9$, we know that:

$$P\left(\frac{\mu - 3 - \mu}{\sigma/7} < Z < \frac{\mu + 3 - \mu}{\sigma/7}\right) = P\left(\frac{-21}{\sigma} < Z < \frac{21}{\sigma}\right) = 0.9,$$

which then, by symmetry, implies that:

$$P\left(0 < Z < \frac{21}{\sigma}\right) = 0.45 \Longrightarrow P\left(Z < \frac{21}{\sigma}\right) = 0.95.$$

Hence, we can then use the table of standard normal distribution to have:

$$\frac{21}{\sigma} = \Phi^{-1}(0.95) \Longrightarrow \sigma = \frac{21}{\Phi^{-1}(0.95)} \approx \frac{21}{1.64} \approx \boxed{12.80488}.$$

⌟

Problem 6.7. (Man-o-War). The Portuguese Man-o-War is a siphonore, (an oceanic creature similar to a jellyfish). The tentacle length of a randomly selected Man-o-War from any region is normally distributed with standard deviation 6ft. However, The average length may differ from region to region.
In the Gulf of Mexico, 5% of all Man-o-War have tentacles over 39.84ft.
Off the coast of Australia, 12.75% of all Man-o-War have tentacles over 39.84ft.
On average, How much longer are the tentacles of a Man-o-War from Australia than a Man-o-War from the Gulf of Mexico?

Sol'n. Here, we denote X_M modeling the length of a randomly selected Man-o-War from Gulf of Mexico and denote X_A modeling the length of a randomly selected Man-o-War from coast of Australia, hence, their distributions are:

$$\begin{cases} X_M \sim \mathcal{N}(\mu_M, 6); \\ X_A \sim \mathcal{N}(\mu_A, 6). \end{cases}$$

Therefore, we can now reverse engineer the average by the given information. For Gulf of Mexico, we have:

$$\begin{aligned} P(X_M > 39.84) = 0.05 &\Longrightarrow P(X_M < 39.84) = 1 - 0.05 = 0.95 \Longrightarrow P\left(Z < \frac{39.84 - \mu_X}{6}\right) = 0.95 \\ &\Longrightarrow \frac{39.84 - \mu_X}{6} = \Phi^{-1}(0.95) \Longrightarrow \mu_X = 39.84 - 6\Phi^{-1}(0.95) \\ P(X_A > 39.84) = 0.1275 &\Longrightarrow P(X_M < 39.84) = 1 - 0.1275 = 0.8725 \Longrightarrow P\left(Z < \frac{39.84 - \mu_A}{6}\right) = 0.8725 \\ &\Longrightarrow \frac{39.84 - \mu_A}{6} = \Phi^{-1}(0.8725) \Longrightarrow \mu_A = 39.84 - 6\Phi^{-1}(0.8725). \end{aligned}$$

Therefore, the Man-o-War from Australia has an longer length of:

$$\begin{aligned} \mu_A - \mu_X &= 39.84 - 6\Phi^{-1}(0.8725) - 39.84 + 6\Phi^{-1}(0.95) \\ &= 6\left(\Phi^{-1}(0.95) - \Phi^{-1}(0.8725)\right) \approx 6(1.64 - 1.14) = \boxed{3(\text{ft})}. \end{aligned}$$

┘

Problem 6.8. (Genetic Defects). Data collected over a long period of time show that a particular genetic defect occurs in 1 every 1000 children. The records of a medical clinic show 60 with the defect in a total of 50,000 examined. If the 50,000 children were a random sample from the population of children represented by past records, what is the probability of observing a value of N (of children with the genetic defect) equal to 60 or less? Approximate the distribution of N and solve the problem by:

(a) Using the correction for continuity, and

(b) not using the correction for continuity. You will note that the difference is not negligible, despite n being very large! That is because the standard deviation of N is very small, relative to n.

Sol'n.

(a) Note that for the 50 000 children, they can each be independently modeled by $\{X_i\}_{i=1}^{50\,000}$ with each being modeled as:

$$X_i \sim \text{Bernoulli}(1/1\,000).$$

Thus, the information of the total population would be $Y := \sum_{i=1}^{50\,000} X_i$, hence it will be:

$$Y \sim \text{Binomial}(50\,000, 0.001).$$

Since we do not have a general formula for CDF of binomial, we can approximate Y by a normal distribution, $\tilde{Y}$:

$$\tilde{Y} \sim \mathcal{N}(50\,000 \times 0.001, \sqrt{50\,000 \times 0.001 \times 0.999}) = \mathcal{N}(50, \sqrt{49.95}).$$

Thus with the correction for continuity, we have:

$$\begin{aligned} P(Y \leq 60) &\approx P(\tilde{Y} \leq 60 + 0.5) = P(Y < 60.5) = P\left(Z < \frac{60.5 - 50}{\sqrt{49.95}}\right) \\ &= \Phi\left(\frac{60.5 - 50}{\sqrt{49.95}}\right) \approx \Phi(1.48567) \approx \boxed{0.93189}. \end{aligned}$$

(b) When not accounting for the correction for continuity, we have:

$$\begin{aligned} P(Y \leq 60) &\approx P(\tilde{Y} \leq 60) = P(Y < 60) = P\left(Z < \frac{60 - 50}{\sqrt{49.95}}\right) \\ &= \Phi\left(\frac{60 - 50}{\sqrt{49.95}}\right) \approx \Phi(1.41492) \approx \boxed{0.92073}. \end{aligned}$$

Rmk. Notice that the two values are very close, this is because that the 0.5 differences divided by the standard deviation is $0.5 \div \sqrt{49.95} \approx 0.07075$, so this would be negligible especially when we are not very close to the center (mean). Moreover, we have:

$$\mathbb{P}(60 < \tilde{Y} < 60.5) = \int_{[60,60.5]}^{\mathcal{L}} f(x)dx = \int_{[60,60.5]}^{\mathcal{L}} \frac{1}{\sqrt{2\pi} \cdot \sigma} \exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right) dx.$$

Note that $m([60, 60.5]) = 0.5$ and we have $\sigma \to \infty$ when $n \to \infty$ (or becoming arbitrarily large), and thus the integrand would approach 0 on a finite support. Thus, this probability mass would also tend to 0, meaning that it will become less important with larger n value.

┘

Problem 6.9. (Coins in a Jar). A collector has two very large jars of pennies. In Jar 1, the mean weight of a penny is $\mu_1 = 1.08$ grams with a standard deviation of $\sigma_1 = 0.1$ grams. In Jar 2, the mean weight of a penny is $\mu_2 = 1.10$ grams with a standard deviation of $\sigma_2 = 0.05$ grams. She picks one of the jars at random, and from that jar draws 25 pennies (again chosen at random). If her sample mean is $\bar{x} > 1.11$ grams, what is the probability that she chose Jar 2?

Sol'n. Note that we shall calculate the distribution of the jars.

- For Jar 1 and 2, with each individual pennies having the same distribution and independent, we can let mean of pick of 25 pennies as:
$$X_1 \sim \mathcal{N}(1.08, 0.1/\sqrt{25}) = \mathcal{N}(1.08, 0.02),$$
$$X_2 \sim \mathcal{N}(1.10, 0.05/\sqrt{25}) = \mathcal{N}(1.10, 0.01).$$
- Considering this, we can give the probability that the sample mean is larger than 1.11 for each Jar if they are selected.
$$P(X_1 > 1.11) = P\left(Z > \frac{1.11 - 1.08}{0.02}\right) = P(Z > 1.5) = 1 - \Phi(1.5) \approx 1 - 0.93319 = 0.06681,$$
$$P(X_2 > 1.11) = P\left(Z > \frac{1.11 - 1.10}{0.01}\right) = P(Z > 1) = 1 - \Phi(1) \approx 1 - 0.84134 = 0.15866.$$

Then, we want to use the conditional probability formula to get:
$$\begin{aligned} P(\text{Jar 2 is chosen}|\bar{x} > 1.11) &= \frac{P(\text{Jar 2 is chosen and } \bar{x} > 1.11)}{P(\text{Jar 1 is chosen and } \bar{x} > 1.11) + P(\text{Jar 2 is chosen and } \bar{x} > 1.11)} \\ &\approx \frac{1/2 \times 0.15866}{1/2 \times 0.06681 + 1/2 \times 0.15866} = \frac{0.15866}{0.06681 + 0.15866} \approx \boxed{0.70369}. \end{aligned}$$

⌟

Problem 6.10. (Yes, We Have No Bananas). The normal daily human potassium requirement is in the range of 2,000 to 6,000 milligrams (mg), with larger amounts required during hot summer weather. The amount of potassium in food varies, but bananas are often associated with high potassium, with approximately 422 mg in a medium-sized banana. Suppose that the distribution of potassium in a banana is Normally distributed, with mean equal to $\mu = 422$ mg and standard deviation $\sigma = 13$ mg per banana. You eat $n = 3$ bananas per day, and T is the total number of milligrams of potassium you receive from them.

(a) Find the mean and standard deviation of T.

(b) Find the probability that your total daily intake of potassium from the 3 bananas will exceed 1300 mg.

Sol'n.

(a) Here, we may have X_i for $i = 1, 2, 3$ to model the potassium amount in each banana, which is a normal distribution and is independent, so:

$$X_i \sim \mathcal{N}(422, 13),$$

Thus with T denoting the total number, we have:

$$T = X_1 + X_2 + X_3,$$

and hence:

$$T \sim \mathcal{N}(422 \times 3, 13 \times \sqrt{3}) = \mathcal{N}(1266, 13\sqrt{3}).$$

Thus, the mean is $\boxed{1266}$ mg and the standard deviation is $\boxed{13\sqrt{3}}$ mg.

(b) Here, we want to calculate the probability as follows:

$$\begin{aligned} P(T > 1300) &= P\left(Z > \frac{1300 - 1266}{13\sqrt{3}}\right) \approx P(Z > 1.50999) \\ &= 1 - \Phi(1.50999) \approx 1 - 0.93448 = \boxed{0.06552}. \end{aligned}$$

┘

Problem 6.11. (Buy Low, Sell High). There are two stocks, A and B, which you are considering purchasing. Each stock is currently priced at \$100. The price of stock A one month from now is modeled by a $\mathcal{N}(103, 8)$ random variable. On the other hand, the price of stock B one month from now is modeled by a $\mathcal{N}(97, 20)$ random variable.

(a) Which stock is more likely to be worth at least \$120 one month from now?

(b) Which stock is more likely to be worth less than it is today, one month from now?

Sol'n.

(a) Note that for the two stocks, we can consider \$120 in their z-scores:

$$\begin{cases} z_A(120) = \dfrac{120 - 103}{8} = 2.125; \\ z_B(120) = \dfrac{120 - 97}{20} = 1.15. \end{cases}$$

Note that $2.125 > 1.15$ so there is a higher probability that $\boxed{\text{stock 2}}$ will worth at lease \$120 as it differs from the mean with less number of standard deviations.

(b) Note that we have the above as normal distributions, denote the two stocks as X_1 and X_2 in one month. Note that the probability of less than their respective average should be the same, then:

$$P(X_1 < 100) < P(X_1 < 103) = P(X_2 < 97) < P(X_2 < 100).$$

by the above inequalities, there is a higher probability that $\boxed{\text{stock 2}}$ will worth less than it is today.

⌟

Problem 6.12. (College Admissions). The ideal size of a first-year class at a particular college is 150 students. The college, knowing from past experience that, on the average, only 30 percent of those accepted for admission will actually attend, uses a policy of approving the applications of 450 students. Let N be the number of first-year students who will attend this college.

(a) What is the probability distribution of N, and what are its expectation and standard deviation?

(b) Using a Normal approximation (with continuity correction), compute the probability that more than 150 first-year students attend this college.

Sol'n.

(a) For each of the 450 students, we can model them individually and independently as in $\{X_i\}_{i=1}^{450}$ being:

$$X_i \sim \text{Bernoulli}(0.30).$$

Hence, for collecting their sum, we have $N := \sum_{i=1}^{450} X_i$, hence we have:

$$X \sim \boxed{\text{Binomial}(450, 0.30)}.$$

Here, the expectation is $\mathbb{E}(X) = 450 \times 0.30 = \boxed{135}$ and the variance is $\text{Var}(X) = 450 \times 0.3 \times 0.7 = 94.5$, then the standard deviation is $\boxed{9.72111}$.

(b) Then we want to use a normal approximation, that is:

$$\tilde{N} \sim \mathcal{N}(135, 9.72111).$$

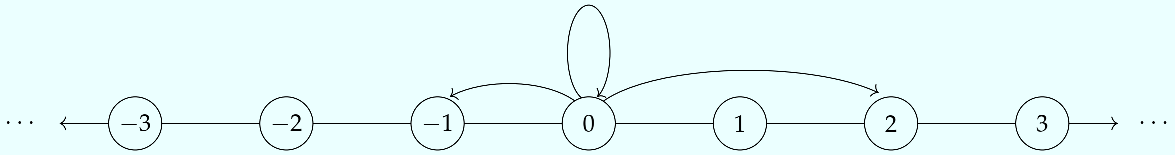
With such the probability is:

$$\begin{aligned} P(N > 150) = P(N \geq 151) &\approx P(\tilde{N} > 151 - 0.5) = P(\tilde{N} > 150.5) \\ &= P\left(Z > \frac{150.5 - 135}{9.72111}\right) \approx P(Z > 1.594468) \\ &= 1 - \Phi(1.594468) \approx 1 - 0.9441 = \boxed{0.0559}. \end{aligned}$$

┘

7 More CLT, Sampling Distributions, & Confidence Intervals

Problem 7.1. (Random Walks). A *random walk* is a model that is often used in the sciences, e.g. in statistical mechanics to describe the random motion of a particle that is subject to different forces (but also in epidemiology to model the diffusion of viruses and bacteria, in computer science for network modeling, et cetera). See the simple random walk shown in the figure below:



$S_0 = 0$ is the initial position of the particle, and let S_n be the position of the particle at times $n = 0, 1, 2, 3, \cdots$ The position S_n for $n \geq 1$ can be thought of as a sum of random displacements: $S_n = X_1 + X_2 + \cdots + X_n$. Here, the range of each X_i is $\{-1, 0, 2\}$, assume each X_i is independent from all the others, and that for each i, $P(X_i = -1) = P(X_i = 0) = P(X_i = 2) = 1/3$. (So note that there is a bit of a "drift" to the right).

(a) What is the pmf of S_2?

(b) Estimate $P(S_{45000} \leq 14750)$.

Sol'n.

(a) We want to first investigate the probability mass function for S_1, which is:

$$P(S_1 = x) = \begin{cases} 1/3, & x = -1; \\ 1/3, & x = 0; \\ 1/3, & x = 2. \end{cases}$$

Then for accounting that range is $\{-2, -1, 0, 1, 2, 4\}$, we have the probability mass function as:

$$f(x) = P(S_2 = x) = \boxed{\begin{cases} 1/9, & x = -2; \\ 1/9 + 1/9 = 2/9, & x = -1; \\ 1/9, & x = 0; \\ 1/9 + 1/9 = 2/9, & x = 1; \\ 1/9 + 1/9 = 2/9, & x = 2; \\ 1/9, & x = 4. \end{cases}}$$

(b) Since we are having a very large, *i.e.*, $n \gg 30$, and since each X_i is independent, we may use the Central limit theorem to approximate $S_{45\,000} = \sum_{i=1}^{45\,000} X_i$ by the normal distribution.

Give that for each X_i, we have the average and standard deviation, respectively, be:

$$\begin{aligned}\mu_{X_i} &= \frac{1}{3} \times (-1+0+2) = \frac{1}{3}, \\ \sigma_{X_i}^2 &= \frac{1}{3} \times \left((-1-1/3)^2 + (0-1/3)^2 + (2-1/3)^2\right) = \frac{14}{9}, \\ \sigma_{X_i} &= \sqrt{\frac{14}{9}} = \frac{\sqrt{14}}{3}.\end{aligned}$$

Hence, we can consider the mean and standard deviation of $S_{45\ 000}$ as:

$$\begin{aligned}\mu_{S_{45\ 000}} &= 45\ 000 \times \frac{1}{3} = 15\ 000, \\ \sigma_{S_{45\ 000}} &= \frac{\sqrt{14}}{3} \times \sqrt{45\ 000} = 100\sqrt{7}.\end{aligned}$$

Therefore, we approximate the distribution by:

$$S_{45\ 000} \sim \mathcal{N}(15\ 000, 100\sqrt{7}),$$

then the probability can be estimated as:

$$\begin{aligned}P(S_{45\ 000} \leq 14750) &\approx P\left(Z < \frac{14750-15000}{100\sqrt{7}}\right) = \Phi\left(-\frac{5}{2\sqrt{7}}\right) \\ &\approx \Phi(-0.944911) \approx \boxed{0.17361}.\end{aligned}$$

┘

Problem 7.2. (Digit Distribution). Let D_i be a randomly chosen digit between 0 and 9, let X_i be the last digit of D_i^2, (e.g. if $D_i = 9$, then $D_i^2 = 81$ and $X_i = 1$). Let $\overline{X_n} = \frac{1}{n}(X_1 + \cdots + X_n)$ be the average of such last digits obtained from independent and randomly selected digits $D_1, \cdots, D_n$.

(a) What do you think $\overline{X_n}$ will be for large n?

(b) Find approximately the least value n such that your prediction of $\overline{X_n}$ is correct to within 0.01 with probability at least 0.99.

(c) Which can be more accurately predicted, $\overline{D_n} = \frac{1}{n}(D_1 + \cdots + D_n)$, or $\overline{X_n}$ for large value of n?

(d) If you had to predict the first digit of $\overline{X_{100}}$, what would you choose to maximize your chance of being right? What is that chance?

Sol'n.

(a) Note that if we exhaust all square of integers, we have:

$$0^2 = 0,\ 1^2 = 1,\ 2^2 = 4,\ 3^2 = 9,\ 4^2 = 16,\ 5^2 = 25,\ 6^2 = 36,\ 7^2 = 49,\ 8^2 = 64,\ 9^2 = 81.$$

Therefore, we can conclude a probability mass function for each X_i, assuming that D_i is chosen randomly:

$$P(X_i = x) = \begin{cases} 0.1, & x = 0; \\ 0.2, & x = 1; \\ 0.2, & x = 4; \\ 0.1, & x = 5; \\ 0.2, & x = 6; \\ 0.2, & x = 9; \\ 0, & \text{otherwise.} \end{cases}$$

Here, we want to find the mean (or expectation) and standard deviation for X_i to facilitate the approximation, that:

$$\mu_{X_i} = \mathbb{E}(X_i) = \sum_{x \in \text{range } X_i} x \cdot P(X_i = x) = 0.1 \times 0 + 0.2 \times 1 + 0.2 \times 4 + 0.1 \times 5 + 0.2 \times 6 + 0.2 \times 9 = 4.5,$$

$$\begin{aligned} \sigma_{X_i}^2 = \text{Var}(X_i) &= \sum_{x \in \text{range } X_i} (x - 4.5)^2 \cdot P(X_i = x) \\ &= 0.1 \times (0 - 4.5)^2 + 0.2 \times (1 - 4.5)^2 + 0.2 \times (4 - 4.5)^2 + 0.1 \times (5 - 4.5)^2 \\ &\quad + 0.2 \times (6 - 4.5)^2 + 0.2 \times (9 - 4.5)^2 \\ &= 9.05, \end{aligned}$$

$$\sigma_{X_i} = \sqrt{9.05}.$$

Since we assume that n is arbitrarily large, we assume that $x \gg 30$, then by the Central Limit Theorem, we have $\overline{X_n}$ approximately a normal distribution, with the mean and standard deviation being:

$$\mu_{\overline{X_n}} = \mu_{X_i} = 4.5,$$
$$\sigma_{\overline{X_n}} = \sqrt{9.05}/\sqrt{n},$$

leading to the distribution that:

$$\overline{X_n} \sim \boxed{\mathcal{N}(4.5, \sqrt{9.05/n})}.$$

Rmk. Note that if $x \to \infty$, we would have $\bar{X}_n \to \boxed{4.5}$.

(b) To translate the problem, we are looking for the least n such that:

$$P(\mu_{\overline{X_i}} - 0.01 \leq \overline{X_n} \leq \mu_{\overline{X_i}} + 0.01) \geq 0.99,$$

which can be translates into:

$$P(-0.01/\sqrt{9.05/n} \leq Z \leq 0.01/\sqrt{9.05/n}) \geq 0.99 \implies P(Z < -0.01\sqrt{n}/\sqrt{9.05}) \geq 0.005.$$

By consulting the standard normal distribution table, we have:

$$-\frac{0.01 \times \sqrt{n}}{\sqrt{9.05}} \leq \Phi^{-1}(0.005) \approx -2.58 \implies \sqrt{n} \gtrapprox \sqrt{9.05} \times 2.58 \times 100,$$
$$\implies n \gtrapprox 9.05 \times (2.58 \times 100)^2 = 602\,404.2$$

Thus we want n to be at least $\boxed{602\,405}$ for the prediction to hold.

(c) Assume that the prediction is the expectation, if we discuss the mean and standard deviation for each D_i, we have:

$$\mu_{D_i} = \frac{0 + 1 + \cdots + 9}{10} = 4.5,$$
$$\sigma^2_{D_i} = \frac{4.5^2 \times 2 + 3.5^2 \times 2 + \cdots + 0.5^2 \times 2}{10} = 8.25,$$
$$\sigma_{D_i} = \sqrt{8.25}.$$

Hence, we have the standard deviation of $\overline{D_n}$ as:

$$\sigma_{\overline{D_n}} = \frac{\sqrt{8.25}}{\sqrt{n}}.$$

Note that:

$$\sigma_{\overline{D_n}} = \frac{\sqrt{8.25}}{\sqrt{n}} < \frac{\sqrt{9.05}}{\sqrt{n}} = \sigma_{\overline{X_n}},$$

this implies that for the same n such that $n \gg 30$, we can approximate both distributions as normal distribution, but since $\overline{D_n}$ has a smaller standard deviation, a higher proportion of the probability mass would be around the center (mean) compared to $\overline{X_i}$, hence $\boxed{\overline{D_i} \text{ will be slightly more accurate}}$ compared to $\overline{X_i}$.

(d) Given that $n = 100 \gg 30$, its distribution is approximately:

$$\overline{X_{100}} \sim \mathcal{N}(4.5, \sqrt{0.0905}).$$

We want to first consider the interval around 4.5, which is $[4, 5)$, its probability is:

$$\begin{aligned} P(4 \leq \overline{X_{100}} < 5) &= P\left(-0.5/\sqrt{0.0905} < Z < 0.5/\sqrt{0.0905}\right) \\ &= [\Phi(0.5/\sqrt{0.0905}) - 0.5] \times 2 \approx [\Phi(1.662056) - 0.5] \times 2 \\ &\approx (0.95154 - 0.5) \times 2 = 0.903499, \end{aligned}$$

Note that this probability is greater than 0.5, so it is most likely to happen, hence we shall guess the first digit is $\boxed{4}$ with a winning probability being approximately $\boxed{0.903499}$.

$\lrcorner$

Problem 7.3. (Achilles Injuries). Frequent participation in strenuous athletic activities could put individuals at risk for Achilles tendinopathy (AT), an inflammation and thickening of the Achilles tendon. A study in The American Journal of Sports Medicine looked at the diameter (in mm) of the affected and non-affected tendons for patients who participated in strenuous sports activities. Suppose that the Achilles tendon diameter in the general population have a mean 5.97mm with a standard deviation of 1.95mm.

(a) What is the probability that a randomly selected sample of 31 patients would produce an average diameter of 7.5mm or less for the non-affected tendon?

(b) When the diameters of the affected tendon were measured for a sample of 31 patients, the average diameter was 9.80. If it *were* the case that the average tendon diameter in the population of patients with AT is no different than the average diameter of the non-affected tendons (5.97mm), then what would be the probability of observing an average diameter of 9.8mm or higher?

(c) What conclusion might you draw from the results of part (b)?

Sol'n.

(a) Suppose that for the general population, the proportion of people having Achilles tendinopathy is negligible, or *i.e.*, everyone has nonaffected tendons.
Thus we each individual of the 31 sample $\{X_i\}_{i=1}^{31}$ is independent, and thus their average, by the Central Limit Theorem, is:

$$\overline{X} = \frac{1}{31}\sum_{i=1}^{31} X_i,$$

which can be modeled by:

$$\overline{X} \sim \mathcal{N}(5.97, 1.95/\sqrt{31}).$$

Thus, the probability of the diameter being less than 7.5mm is:

$$P(\overline{X} < 7.5) = P(Z < (7.5 - 5.97)/(1.95/\sqrt{31})) = \Phi(1.53\sqrt{31}/1.95) \approx \Phi(4.368554) > 0.99997.$$

The z-score is out of range for the standard normal table, so we know that this probability of the average less than 7.5mm is $\boxed{\text{above } 0.99997}$, and by a calculator, the probability is approximately 0.999994.

(b) Now we assume that the diameters is the same distribution for the affected tendon, we can reuse the distribution of $\overline{X}$ being the average of the 31 samples, which is:

$$\overline{X} \sim \mathcal{N}(5.97, 1.95/\sqrt{31}).$$

Note that now, if we calculate the probability that this happens, it is:

$$P(\overline{X} > 9.8) = P(Z > (9.8 - 5.97)/(1.95/\sqrt{31})) \approx 1 - \Phi(10.9357) \approx \boxed{0}.$$

(c) From part two, even if the average are approximately the same, we should have the probability of greater than the average approximately 0.5, but however that probability is ~ 0, thus this is a contradiction. Thus, our assumption that the average tendon diameter in the population of patients with AT is no different than the average diameter of the non-affected tendons, must be false. Hence $\boxed{\text{population of patients with AT is different from the average diameter of the non-affected tendons}}$.

Rmk. Note that conceptually, we are almost certain that for the non-affected population, the achilles is less than 7.5mm, but we are almost certain that for the affected population that the achilles is more than 9.80mm, which means that they have to be $\boxed{\text{different}}$.

⌟

Problem 7.4. (Cheating in an experiment). Suppose that $(X_1, \cdots, X_n)$ is a sample form a population with population mean μ and population variance σ^2: consider $X_1, \cdots, X_n$ as i.i.d. (independent and identically distributed) random variables with distribution given by the population distribution (so they each have mean μ and variance σ^2). Define the sample mean as usual, i.e. $\overline{X} = \frac{1}{n}(X_1 + \cdots + X_n)$.

(a) Compute $\mathbb{E}(\overline{X})$, $\text{Var}(\overline{X})$, and $\text{SD}(\overline{X})$, expressing them in terms of μ, σ^2, and n.

(b) You work in a lab, and your boss asks you to run an experiment 6 times so to collect the sample $X_1, \cdots, X_6$ (again, suppose these are i.i.d., each with mean μ and variance σ^2). You are feeling lazy and you run the experiment only 4 times, and you "reuse" the result X_4 for the two experiments that you did not run (the 5th and 6th trials). So, instead of computing

$$\overline{X} = \frac{X_1 + X_2 + X_3 + X_4 + X_5 + X_6}{6},$$

you "cheat" and compute instead:

$$\overline{X_C} = \frac{X_1 + X_2 + X_3 + 3X_4}{6},$$

Compare $\mathbb{E}(\overline{X})$ and $\mathbb{E}(\overline{X_c})$ (which may make you think you can get away with it!), and also compute $\text{Var}(\overline{X})$, $\text{Var}(\overline{X_c})$ and their ratio $\text{Var}(\overline{X_c})/\text{Var}(\overline{X})$. How does this show the risks of your behavior?

Sol'n.

(a) Given such, since each measure is distributed independently and identically, thus we have:

$$\mathbb{E}(\overline{X}) = \frac{1}{n}\sum_{i=1}^{n}\mathbb{E}(X_i) = n \cdot \frac{\mu}{n} = \boxed{\mu},$$

$$\text{Var}(\overline{X}) = \frac{1}{n^2}\sum_{i=1}^{n}\text{Var}(X_i) = \frac{1}{n^2} \cdot n \cdot \sigma^2 = \boxed{\frac{\sigma^2}{n}},$$

$$\text{SD}(\overline{X}) = \sqrt{\frac{\sigma^2}{n}} = \boxed{\frac{\sigma}{\sqrt{n}}}.$$

(b) For the expectation, we have them respectively be:

$$\mathbb{E}(\overline{X}) = \frac{1}{6}\sum_{i=1}^{6}\mathbb{E}(X_i) = \frac{1}{6} \times 6 \times \mu = \boxed{\mu},$$

$$\mathbb{E}(\overline{X_c}) = \frac{1}{6}[\mathbb{E}(X_1) + \mathbb{E}(X_2) + \mathbb{E}(X_3) + 3\mathbb{E}(X_4)] = \frac{1}{6} \times (\mu + \mu + \mu + 3\mu) = \boxed{\mu}.$$

Hence, the mean are same. However, concerning variance, since each event is independent, we have:

$$\text{Var}(\overline{X}) = \frac{1}{6^2}\sum_{i=1}^{6}\text{Var}(X_i) = \frac{1}{36} \times 6 \times \sigma^2 = \boxed{\frac{\sigma^2}{6}},$$

$$\begin{aligned}\text{Var}(\overline{X_c}) &= \frac{1}{6^2}[\text{Var}(X_1) + \text{Var}(X_2) + \text{Var}(X_3) + \text{Var}(3X_4)] \\ &= \frac{1}{36} \times (\sigma^2 + \sigma^2 + \sigma^2 + 9\sigma^2) = \boxed{\frac{\sigma^2}{3}}.\end{aligned}$$

Note that for the variance, issue comes in, we have the ratio being:

$$\frac{\text{Var}(\overline{X_c})}{\text{Var}(\overline{X})} = \frac{\sigma^2/3}{\sigma^2/6} = \boxed{2},$$

meaning that the variance of the cheated version would be larger than the standard procedure. This implies that $\boxed{\text{the calculation would be more likely to deviate from the actual mean}}$. If the boss knows approximately what the mean should be around, it is more likely that you deviate from it more and get caught. On the other hand, this is an unethical behalf and would pollute data set.

⌟

Problem 7.5. (Sampling from a small population). Consider the population $(2, 4, 6, 8)$, of population size $N = 4$.

(a) Compute the population mean μ and population variance σ^2.

(b) We sample the population without replacement. In how many ways can we draw an (ordered) sample (X_1, X_2) of size $n = 2$? In other words: in how many ordered arrangements (X_1, X_2) of the elements of the population are there? List all the possible (ordered) samples of size 2 (note that you are building the sample space S).

(c) For each of such samples (X_1, X_2), compute the corresponding sample mean $\overline{X} = \frac{X_1+X_2}{2}$. What is range$(\overline{X})$?

(d) If sampling is performed via Simple Random Sampling (SRS), then the samples listed in part (b) are all equally likely: under this assumption, find the distribution of $\overline{X}$.

(e) Compute $\mathbb{E}(\overline{X})$, and verify that $\mathbb{E}(\overline{X}) = \mu$, as predicted by the theory.

(f) Finally, compute $\text{Var}(\overline{X})$ and verify that $\text{Var}(\overline{X}) = \frac{\sigma^2}{n}\frac{N-n}{N-1}$

Note: The factor $\frac{N-n}{N-1}$ is called finite population correction. We must use it because the population size N is not very large with respect to the sample size n. For this reason, X_1 and X_2 cannot be treated as independent random variables and the formula for $\text{Var}(X)$ that you derived in Problem 7.4(a) does not apply.

Sol'n.

(a) The population mean and variance, respectively, are:

$$\mu = \frac{2+4+6+8}{4} = \frac{20}{4} = \boxed{5},$$

$$\sigma^2 = \frac{(2-5)^2 + (4-5)^2 + (6-5)^2 + (8-5)^2}{4} = \boxed{5}.$$

(b) Since we are choosing two elements with order, the number of choices are:

$$\#(S) = |S| = (4)_2 = \frac{4!}{(4-2)!} = 4 \times 3 = \boxed{12}.$$

If we want to list them all, they are:

$$S = \{\boxed{(2,4), (2,6), (2,8), (4,2), (4,6), (4,8), (6,2), (6,4), (6,8), (8,2), (8,4), (8,6)}\}.$$

(c) Here, if we do the calculation of averaging and make it multi-set (preserving order compared to the previous set), we have:

$$\mathcal{X} = \{3, 4, 5, 3, 5, 6, 4, 5, 7, 5, 6, 7\},$$

which since we are considering the range set, we have:

$$\text{range}(X) = X = \boxed{\{3, 4, 5, 6, 7\}}, \qquad \text{and} \qquad \#(\text{range}(X)) = 4.$$

(d) The distribution of $\overline{X}$ can be expressed by the probability mass function by the occurrence of each element in the multi-set $\mathcal{X}$:

$$P(\overline{X} = x) = \boxed{\begin{cases} 2/12 = 1/6, & x = 3, \\ 2/12 = 1/6, & x = 4, \\ 4/12 = 1/3, & x = 5, \\ 2/12 = 1/6, & x = 6, \\ 2/12 = 1/6, & x = 7. \end{cases}}$$

(e) To compute $\mathbb{E}(\overline{X})$, we have:

$$\mathbb{E}(\overline{X}) = \sum_{x \in \text{range}(X)} x \cdot P(\overline{X} = x) = 3 \times 1/6 + 4 \times 1/6 + 5 \times 1/3 + 6 \times 1/6 + 7 \times 1/6 = \boxed{5},$$

and this value corresponds to $\mu = 5$.

(f) To compute $\text{Var}(\overline{X})$, we have:

$$\begin{aligned} \text{Var}(\overline{X}) &= \sum_{x \in \text{range}(X)} (x-5)^2 \cdot P(\overline{X} = x) \\ &= (3-5)^2 \times 1/6 + (4-5)^2 \times 1/6 + (5-5)^2 \times 1/3 + (6-5)^2 \times 1/6 + (7-5)^2 \times 1/6 \\ &= \frac{4+1+1+4}{6} = \boxed{\frac{5}{3}}. \end{aligned}$$

With the given formula, we compute that:

$$\frac{\sigma^2}{n} \cdot \frac{N-n}{N-1} = \frac{5}{2} \cdot \frac{4-2}{4-1} = \frac{5 \cdot 2}{2 \cdot 3} = \frac{5}{3},$$

which corresponds to the our calculate above.

⌟

Problem 7.6. (Blood Pressure). Systolic blood pressure was measured in a sample of 100 healthy women between the ages of 25 and 29 years. The sample mean pressure was $\overline{X} = 120$mm Hg (millimeters of mercury), and the sample standard deviation was $S = 10$mm Hg. Find the 95% and 99% confidence intervals for the true mean μ.

Sol'n. Note that we want to first calculate the standard error:

$$\text{SE}(\overline{X}) = \frac{\sigma}{\sqrt{n}} \simeq \frac{s}{\sqrt{n}} = \frac{10}{\sqrt{100}} = 1.$$

For the 95% confidence interval, we have $\alpha = 0.05$, then:

$$z_{0.025} \approx 1.96,$$

then we have the confidence interval being:

$$[120 - 1.96 \times 1, 120 + 1.96 \times 1] = \boxed{[118.04, 121.96]}.$$

For the 99% confidence interval, we have $\alpha = 0.01$, then:

$$z_{0.005} \approx 2.58,$$

then we have the confidence interval being:

$$[120 - 2.58 \times 1, 120 + 2.58 \times 1] = \boxed{[117.42, 122.58]}.$$

┘

Problem 7.7. (Battery Powered). The capacities (in ampere-hours) of 258 batteries were measured. The average was 181.5 and the standard deviation was 15.1. Suppose that it's desired to have a 98% confidence interval whose margin of error (or half-width) is only 1 ampere-hour. Find the number of ADDITIONAL batteries that have to be tested. (In other words, not counting the 258 batteries that have already been tested) to obtain this margin of error.

Sol'n. For this question, since the standard deviation of the battery capacity is unknown, we will estimate the sample standard deviation as:

$$s = 15.1,$$

First, we want to find the z for $\alpha = 0.02$, that is:

$$z_{0.01} \approx 2.33$$

Therefore, our (half-width) margin of error is:

$$\begin{aligned} \text{ME} = z_{0.01} \times \frac{s}{\sqrt{n}} &\Longrightarrow 1 \approx 3.09 \times \frac{15.1}{\sqrt{n}} \\ &\Longrightarrow \sqrt{n} \approx 2.33 \times 15.1 \approx 35.183 \\ &\Longrightarrow n \approx (35.183)^2 \approx 1\,237.843489, \end{aligned}$$

hence, we need approximately $1\,238 - 258 = \boxed{980}$ more batteries. ⌟

Problem 7.8. (Education Level). Suppose 50 different survey organizations visit eastern Tennessee to estimate the average number of years of schooling completed among adults age 25 and over. Each organization surveys 400 people and reports a 90% confidence interval.

(a) Of these 50 intervals, how many of these intervals would you expect to contain the true population average?

(b) Suppose one of these organizations in Bristol TN found that 363 of the surveyed adults were high school graduates. What is the 90% confidence interval the organization would report for the proportion of the population which are high school graduates?

Sol'n.

(a) By definition, the confidence interval of 90% implies that the probability that the interval contains the average.
Hence, the expected number of intervals containing the true population average is:

$$\mathbb{E}(\text{Number of intervals containing the true population average}) = 90\% \times 50 = \boxed{45}.$$

(b) First, we want to find the mean and standard deviation of this sample, we have:

$$\begin{aligned} \hat{p} &= \frac{1}{400} \times 363 = \frac{363}{400} \approx 0.9075, \\ \hat{p}(1-\hat{p}) &= \frac{363}{400} \cdot \frac{400-363}{400} = \frac{13\,431}{160\,000}, \\ \sqrt{\hat{p}(1-\hat{p})} &= \frac{\sqrt{13\,431}}{400}. \end{aligned}$$

Following that, for the z value, we have calculated $\alpha = 0.1$, so:

$$z_{0.05} \approx 1.64.$$

Then, the margin of error is:

$$\text{ME} \approx 1.64 \times \frac{\sqrt{13\,431}/400}{\sqrt{400}} = 0.0237579,$$

thus, the proportion interval of 90% confidence is:

$$[0.9075 - 0.0237579, 0.9075 + 0.0237579] = \boxed{[0.8837421, 0.9312579]}.$$

⌟

Problem 7.9. (Noise and Stress). To compare the effect of stress in the form of noise on the ability to preform a simple task, 70 subjects were divided into two groups. The first group of 30 subjects acted as the control, while the second group of 40 were the experimental group. Although each subject completed the task, the experimental group subjects performed the task while loud music was played. The time to finish the task was recorded for each subject and the following summary was obtained:

	Control	**Experimental**
n	30	40
$\bar{x}$	15 min	23 min
s	4 min	10 min

(a) Find a 99% confidence interval for the difference in mean completion times from these two groups.

(b) Based on part (a), do you think there is sufficient evidence to indicate a difference in the average time to completion for the two groups? Explain.

Sol'n.

(a) Here, we assume the difference is how much more time experimental takes than the control group. For the 99% confidence, we have $\alpha = 0.01$:

$$z_{0.005} \approx 2.58,$$

therefore, we have the mean and the standard error being:

$$\bar{x}_1 - \bar{x}_2 = 23 - 15 = 8 \text{ (min)},$$

$$\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}} = \sqrt{\frac{16}{30} + \frac{100}{40}} = \sqrt{\frac{91}{30}} \approx 1.741647.$$

then we have the confidence interval being:

$$[8 - 2.58 \times 1.741647, 8 + 2.58 \times 1.741647] = \boxed{[3.506551, 12.493449]}.$$

(b) Note that for with 99% of certainty, we know that it takes a longer time for the experimental group to complete the experiment, *i.e.*, the difference is positive, so the different is more than 99% likely to be positive. Hence, we are quite certain that $\boxed{\text{noise make people complete the task relative slower}}$.

┘

Problem 7.10. (Comparing Training Methods). A track coach wants to compare the effectiveness of two training regiments for the middle distance runners on her team. She creates the following experiment:

- All runners run an 800 meter time trial at the beginning of the the season.
- The runners are randomly placed into two equally sized groups: The first group receives training regiment A throughout the season, the second group receives training regiment B throughout the season.
- The runners run a 800 meter race at the district track meet that the end of the season. The difference between the preseason time and this time are recorded. *E.g.* If you ran a 2:08 at the start of the season, then improved to a 1:59, your difference would be 9 sec. On the other hand, if you ran a slower 2:10, your difference would be -2 sec.

The track coach expects the the time differences for any runner from either group to be no worse than -5 sec. but no better than 15 sec. For the estimate of the difference in mean between the two groups to be correct within 2.5 seconds with a probability equal to 0.95, how many runners must be included in each training group?

Sol'n. For this question, since we are not given a standard deviation, we estimate it as one fourth of the population range, which is:

$$\sigma = 20 \div 4 = 5.$$

Hence, as we want to ensure that the margin of error is less than 2.5 sec with 95% confidence, then we have $\alpha = 0.05$, then:

$$z_{0.025} \approx 1.96,$$

then we have the margin of error being:

$$1.96 \times \sqrt{\frac{\sigma^2}{n} + \frac{\sigma^2}{n}} \leq 2.5,$$

then we have:

$$\sqrt{n} \geq 1.96 \times \sqrt{50} \div 2.5 \implies n \geq 30.7328$$

so if we have at least $\boxed{31}$ students in each group, we can be guaranteed that the difference between the two groups to be correct within 2.5 seconds with a probability equal to 0.95. ⌟

8 Large Sample Hypothesis Testing

Problem 8.1. (Working Backwards). You are given the hypotheses: $H_0 : \mu = 60$, and $H_a : \mu < 60$. We know that the sample standard deviation is $s = 8$ and the sample size is $n = 200$. For what sample mean would the p-value be equal to 0.05?

Sol'n. Initially, we have $\mu_0 = 60$ with standard deviation of $s = 8$, and with $n = 200 \gg 30$, we form the null distribution of the test statistics:

$$\frac{\overline{X} - 60}{8/\sqrt{200}} \sim \mathcal{N}(0, 1).$$

Here, we let the sample mean be $\bar{x}$, we compute the test statistics:

$$z^* = \frac{\bar{x} - 60}{8/\sqrt{200}}.$$

Note that we have a (left) one-side test, we have p-value as:

$$p\text{-value} = P(Z \leq -z^*) = \Phi\left(-\frac{\bar{x} - 60}{8/\sqrt{200}}\right) = 0.05,$$

which gives us that:

$$-\frac{\bar{x} - 60}{8/\sqrt{200}} = \Phi^{-1}(0.05) \Longrightarrow \bar{x} = 60 + \frac{8 \times \Phi^{-1}(0.05)}{\sqrt{200}} \approx 60 + \frac{8 \times (-1.644854)}{14.142136} \approx \boxed{59.069530}.$$

┘

Problem 8.2. (Conceptual Understanding). True or False: Justify each answer.

(a) When conducting a two-tailed test of $H_0 : \mu_0 = \mu$ against $H_a : \mu_0 \neq \mu$ with significance level $\alpha = 0.05$, we can reject H_0 if either $\Phi(z^*) < 0.025$ or $\Phi(-z^*) < 0.025$.

(b) A 95% confidence interval for μ is calculated. In testing $H_0 : \mu = \mu_0$ against $H_a : \mu \neq \mu_0$, we have p-value > 0.05 if μ_0 is any of the numbers inside the confidence interval.

Sol'n.

(a) True.
First, since this is a two-side test, we have p-value as:

$$p\text{-value} = P(Z \leq -|z^*|) + P(Z > |z^*|) = 2\Phi(-|z^*|).$$

Then, we want to analyze the condition, by either $\Phi(z^*) < 0.025$ or $\Phi(-z^*) < 0.025$, by symmetry, we know that:

- If $z^* < 0$, we have:

$$\Phi(-|z^*|) = \Phi(z^*) < \Phi(-z^*),$$

so we have p-value being strictly less than $2 \times 0.025 = 0.05$, which is bounded above by $\alpha = 0.05$, so we can reject H_0.

- If $z^* \geq 0$, we have:

$$\Phi(-|z^*|) = \Phi(-z^*) \leq \Phi(z^*),$$

so we have p-value being strictly less than $2 \times 0.025 = 0.05$, which is bounded above by $\alpha = 0.05$, so we can, still, reject H_0.

Since H_0 is rejected in both cases, the statement is justified.

(b) True.
Note that we can prove the statement by proving its contraposition. Suppose we have p-value < 0.05, this implies that we reject H_0, thus $z^* \in R_0$, which implies that:

$$\left| \frac{\bar{x} - \mu_0}{\sigma\sqrt{n}} \right| > -\Phi(0.025) \Longrightarrow |\bar{x} - \mu_0| > z_{0.025}\sigma\sqrt{n},$$

so their distance is more than the confidence interval, so we know that $\bar{x}$ is not in the confidence interval of μ_0, *i.e.*, μ_0 and $\bar{x}$ is separated by more than the margin of error (their distance is strictly larger than the margin of error), hence μ_0 is not in the confidence interval of $\bar{x}$.
Thus, by contrapositive, we have μ_0 is any of the numbers inside the confidence interval implying that p-value > 0.05.
Rmk. Of course, we can construct the argument through a direct proof, which involves almost the same argument but reversing the arrows and inequalities. Moreover, one could note that the converse also holds, leading to an equivalence relationship.

┘

Problem 8.3. (Plant Genetics). A peony with red petals was crossed with another plant having streaky petals. A geneticist states that 75% of the offspring resulting from this cross will have red petals. To test this claim, 100 seeds from this cross were collected and germinated, and 58 plants had red petals.

(a) What hypothesis should you use to test the geneticist's claim?

(b) Calculate the test statistic and it's p-value to evaluate the statistical significance of the results with $\alpha = 0.01$.

Sol'n.

(a) To test the geneticist's claim we have:
H_0: The proportion of offspring resulting from this cross will have red petals is 75%, or $\boxed{p = 0.75}$.
H_a: The proportion of offspring resulting from this cross will have red petals is not 75%, or $\boxed{p \neq 0.75}$.

(b) Note that that we have $np_0 = 75 \gg 5$ and $n(1-p_0) = 25 \gg 5$, so we can form the null distribution for the test statistics:

$$Z = \frac{\hat{p} - 0.75}{\sqrt{0.75 \times (1-0.75)/100}} = \frac{\hat{p} - 0.75}{\sqrt{3}/40} \sim \mathcal{N}(0,1),$$

Therefore, we compute the test statistic, as:

$$z^* = \frac{0.58 - 0.75}{\sqrt{3}/40} = -\frac{6.8}{\sqrt{3}}.$$

Notice that we have a two-side test, we then find the p-value as:

$$p\text{-value} = P(Z \leq -|z^*|) + P(Z \geq |z^*|) = 2\Phi\left(-\frac{6.8}{\sqrt{3}}\right) \approx 0.000086 \leq 0.01 = \alpha,$$

thus we reject H_0.
Hence, we reject H_0 at significance level $\alpha = 0.01$, and we have evidence to conclude that the geneticist's claim is not right, *i.e.*, $\boxed{\text{have evidence to conclude this cross will have red petals is not 75\%}}$.

┘

Problem 8.4. (Body Temperature). In the research article What's Normal? Temperature, Gender, and Heart Rate, a random sample of 130 human body temperatures had a mean of 98.25°F and a standard deviation of 0.73°F.

(a) Does the data indicate that the average human body temperature is lower than the often reported 98.6°F? Use $\alpha = 0.05$.

(b) Calculate the power of the test if the true average human body temperature is 98.3°F.

(c) In 1868, the physician Carl Wunderlich claimed to have recorded one million temperatures in the course of his research and found an average of 98.6°F. What are your thought about this result in conjunction with your conclusion from part (a)?

Sol'n.

(a) Here, we construct a hypothesis test, with the hypothesis:
H_0: The average human body temperature is the same as reported, or $\mu = 98.6$°F.
H_a: The average human body temperature is lower than reported, or $\mu < 98.6$°F.
Then, since the sample size is $130 \gg 30$, we construct our null distribution of the test statistics as:

$$\frac{\overline{X} - 98.6}{0.73/\sqrt{130}} \sim \mathcal{N}(0,1).$$

Following that, we calculate the test statistics:

$$z^* = \frac{98.25 - 98.6}{0.73/\sqrt{130}} = -\frac{35\sqrt{130}}{73},$$

Moving forward, we calculate the p-value, which is:

$$p\text{-value} = P(Z < z^*) = \Phi\left(-\frac{35\sqrt{130}}{73}\right) \approx 2.299325 \times 10^{-8} \leq 0.05 = \alpha,$$

hence we reject H_0.
Therefore, we reject H_0 at significance level $\alpha = 0.05$, and we have evidence to conclude that the $\boxed{\text{average human body temperature is lower than often reported 98.6°F}}$.

(b) The probability of a Type II error is:

$$\begin{aligned}\beta(98.3, 0.05) &= P\left(\frac{\overline{X} - 98.6}{0.73/\sqrt{130}} > -1.64485 \mid \overline{X} \sim \mathcal{N}(98.3, 0.73/\sqrt{130})\right) \\ &= P(\overline{X} > 98.494688 \mid \overline{X} \sim \mathcal{N}(98.3, 0.73/\sqrt{130})) \\ &\approx 1 - \Phi\left(\frac{98.494688 - 98.3}{0.73/\sqrt{130}}\right) \\ &\approx 1 - 0.998820,\end{aligned}$$

hence, the power of the test is:

$$1 - \beta \approx 1 - (1 - 0.998820) = \boxed{0.998820}.$$

(c) From the proceeding part, we know that: $\boxed{\text{a decrease in average human body temperature}}$ is highly significant.

┘

Problem 8.5. (The Digits of π). The ratio of a circle's circumference to its diameter is the number $\pi = 3.14159\cdots$ In 1761 it was proven that π is irrational and, as such, its decimal expansion is nonterminating and nonrepeating. This does not, however, imply anything about the distribution of its digits. For example, we suspect that the proportion of 0's appearing in the decimal expansion of π is $p = 1/10$, but we don't know for sure. To test this, a random sample of 100 digits of π was taken and found that there were 9 zeros in the sample.

(a) Use this information to test whether p, the proportion of zeros appearing in the decimal expansion of π is not significantly different than $1/10$. Use $\alpha = 0.05$.

(b) Does your conclusion prove or disprove that $p = 1/10$?

<u>Sol'n.</u>

(a) For this question, we construct a hypothesis test, with the following hypothesis:
H_0: The proportion of zeros appearing in the decimal expansion of π is the same as suspected, or $p = 1/10$.
H_a: The proportion of zeros appearing in the decimal expansion of π is different from suspected, or $p \neq 1/10$.
Then, since we have $np_0 = 10 > 5$ and $n(1-p_0) = 90 \gg 5$, we construct our null distribution of the test statistics as:

$$Z = \frac{\hat{p} - 1/10}{\sqrt{1/10 \times 9/10/100}} = \frac{p - 1/10}{3/100} \sim \mathcal{N}(0,1).$$

Following that, we calculate the test statistics:

$$z^* = \frac{9/100 - 10/100}{3/100} = -\frac{1}{3}.$$

Moving forward, we calculate the p-value for the two-sided test, which is:

$$p\text{-value} = P(Z < -|z^*|) + P(Z > |z^*|) = 2\Phi\left(-\frac{1}{3}\right) \approx 0.738883 > 0.05 = \alpha,$$

hence we retain (or fail to reject) H_0.
Therefore, we fail to reject H_0 at significance level $\alpha = 0.05$, and we have $\boxed{\text{no evidence to conclude}}$ that the decimal expansion of π is different from suspected.

(b) The conclusion $\boxed{\text{fails to disprove}}$ that $p = 1/10$, however, it is not does not necessarily prove that $p = 1/10$.

┘

Problem 8.6. (Batting Averages; Revisited). Recall in Excel project 1 we compared the distribution of batting averages of American and Nation League hitters. In said project, we found the following information:

American League	**National League**
$n_A = 207$	$n_N = 207$
$\bar{x}_A = 0.2731$	$\bar{x}_N = 0.2742$
$s_A = 0.0312$	$s_N = 0.0339$

Is this enough evidence to suggest that there is a difference in batting averages between the leagues? Use $\alpha = 0.05$.

Sol'n. In tackling this problem, we want to form a hypothesis test for difference between population means. In doing so, we first establish the null and alternative hypotheses:
H_0 : The means of the American and National League are the same, or $\mu_1 - \mu_2 = 0$.
H_a : The means of the American and National League are different, or $\mu - \mu_2 \neq 0$.
Note that with the sample size $n_A = n_N = 207 \gg 30$, we may construct our null distribution, as:

$$Z = \frac{\overline{X_1} - \overline{X_2} - 0}{\sqrt{\frac{0.0312^2}{207} + \frac{0.0339^2}{207}}} \approx \frac{\overline{X_1} - \overline{X_2}}{\sqrt{0.002123/207}} \sim \mathcal{N}(0,1).$$

Following that, we calculate the test statistics, which is:

$$z^* \approx \frac{0.2731 - 0.2742}{\sqrt{0.002123/207}} = -\frac{0.0011}{\sqrt{0.002123/207}}.$$

Following that, we find the p-value as two-sided test, which is:

$$p\text{-value} = P(Z < -|z^*|) + P(Z > |z^*|) = 2\Phi\left(-\frac{0.0011}{\sqrt{0.002123/207}}\right) \approx 0.731215 > 0.05 = \alpha,$$

hence we retain (or fail to reject) H_0.
Therefore, we fail to reject H_0 at significant level $\alpha = 0.05$, and we have $\boxed{\text{no sufficient evidence to conclude}}$ that there is a difference in batting averages between the leagues. ⌟

Problem 8.7. (An Aspirin a Day?). In an 1988 article[1] researchers conducted a study to test the hypothesis that 325 mg of aspirin taken every other day decreases the mortality from cardiovascular disease. 22071 physicians participating in the study were randomized into one of two groups:

- $n_1 = 11037$ physicians who took buffered aspirin every other day, and
- $n_2 = 11034$ physicians who took aspirin placebo every other day.

Neither the participants, nor the investigators responsible for following them knew which group they were in. The results of this study, we'll call Study 1, are summarized below[2]:

STUDY 1

	Aspirin	Placebo
Myocardial Infraction	104	189
Fatal	5	18
Nonfatal	99	171

In each of the following, use the standard $\alpha = 0.05$.

(a) Test whether Study 1 does in fact indicate that the rate of heart attacks for physicians taking aspirin is significantly lower than the rate of those on the placebo.

(b) Test whether Study 1 indicates that the mortality rate of heart attacks is lower for physicians taking aspirin than those on the placebo.

[1] J.B. Greenhouse, and S.W. Greenhouse, An Aspirin a Day...? Chance: New directions for Statistics and Computing 1, no. 4 (1988):24-31

[2] Myocardial Infractions are colloquially called "heart attacks".

Sol'n.

(a) For this part, we want to test using a hypothesis test. First, we construct the null and alternative hypothesis for Study 1, as follows:
H_0 : The rate of heart attacks for physicians taking aspirin and placebo are the same in study 1, or $p_1 - p_2 = 0$.
H_a : The rate of heart attacks for physicians taking aspirin are lower than the rate of those taking placebo in study 1, or $p_1 - p_2 < 0$.
Suppose the population proportions are unknown, but since $n_1\hat{p}_1 = 104 \gg 5$, $n_1(1-\hat{p}_1) = 11\,037 - 104 = 10\,933 \gg 5$, $n_2\hat{p}_2 = 189 \gg 5$, and $n_2(1-\hat{p}_2) = 11\,034 - 189 = 10\,845 \gg 5$, we can construct the null distribution for the test statistic as:

$$Z = \frac{\hat{p}_1 - \hat{p}_2}{\sqrt{\hat{p}(1-\hat{p})(1/11\,037 + 1/11\,034)}} \sim \mathcal{N}(0,1),$$

where we have the pooled sample proportion as:

$$\hat{p} = \frac{104 + 189}{11\,037 + 11\,034} = \frac{293}{22\,071}.$$

Hence, we now have the z-score as:

$$z^* = \frac{104/11\,037 - 189/11\,034}{\sqrt{293/22\,071 \times 21\,778/22\,071 \times (1/11\,037 + 1/11\,034)}} \approx -5.001388.$$

Therefore, we have the p-value as:

$$p\text{-value} = P(Z < z^*) \approx \Phi(-5.001388) \approx 2.850456 \times 10^{-7} \leq 0.05 = \alpha,$$

hence we reject H_0.

Therefore, we reject H_0 at significant level $\alpha = 0.05$, and we have $\boxed{\text{sufficient evidence to conclude}}$ that the rate of heart attacks for physicians taking aspirin is significantly lower than the rate of those on placebo in study 1.

(b) For this part, we cannot obtain a good hypothesis statistical test, but we will attempt to make a hypothesis. First, we construct the null and alternative hypothesis for Study 1, as follows:

H_0 : The mortality rate of heart attacks for physicians taking aspirin and placebo are the same in study 1, or $p_1 - p_2 = 0$.

H_a : The mortality rate of heart attacks for physicians taking aspirin is lower than the ones taking placebo in study 1, or $p_1 - p_2 < 0$.

Suppose the population proportions are unknown, but note that $n_1\hat{p}_1 = 5 \not> 5$, $n_1(1-\hat{p}_1) = 99 \gg 5$, $n_2\hat{p}_2 = 18 > 5$, and $n_2(1-\hat{p}_2) = 171 \gg 5$.

Notice that for $n_1\hat{p}_1$, we fail to have this greater than 5. Since this is not vary by far, so attempt to construct the null distribution for the test statistic as:

$$Z = \frac{\hat{p}_1 - \hat{p}_2}{\sqrt{\hat{p}(1-\hat{p})(1/104 + 1/189)}} \sim \mathcal{N}(0,1),$$

where we have the pooled sample proportion as:

$$\hat{p} = \frac{5+18}{104+189} = \frac{23}{293}.$$

Hence, we now have the z-score as:

$$z^* = \frac{5/104 - 18/189}{\sqrt{23/293 \times 270/293 \times (1/104 + 1/189)}} \approx -1.436217.$$

Therefore, we have the p-value as:

$$p\text{-value} = P(Z < z^*) \approx \Phi(-1.436217.) \approx 0.075470 > 0.05 = \alpha,$$

hence we retain (fail to reject) H_0 with the assumption of Central Limit Theorem holds.

Therefore, we retain H_0 at significant level $\alpha = 0.05$, and we have $\boxed{\text{no sufficient evidence to conclude}}$ that the mortality rate of heart attacks is lower for physicians taking aspirin is significantly compared to the rate of those on placebo in study 1 if we assume CLT.

Rmk. Note that we did have $5 \not< 5$, in which we have not reached a large enough population for CLT to hold. However, since:

- it is not varied by much, and
- we have not learned the variants of Hypothesis test,

we would consider the above result as a reasonable and compensable result overall.

┘

However, on January 30, 1988, a headline in the *New York Times* read “Value of daily aspirin disputed in British study of heart attacks.” In the study that the *Times* referenced, 5139 physicians were randomly allocated into one of two groups:

- $m_1 = 3429$ doctors who took 500 mg of aspirin daily, and
- $m_2 = 1710$ doctors who were instructed to avoid aspirin.

The results of this study, which we’ll call Study 2 are summarized below:

STUDY 2

	Aspirin	Control
Myocardial Infraction	169	88
Fatal	89	47
Nonfatal	80	41

(c) Repeat part (a) using the data from Study 2, in which one group took daily aspirin and the other took none.

(d) Can you think of any possible reasons why these two studies, which were similar in some aspects, produced such different conclusions?

Sol’n.

(c) For this part, we want to test using a hypothesis test. First, we construct the null and alternative hypothesis for Study 2, as follows:
H_0 : The rate of heart attacks for physicians taking aspirin and placebo are the same in study 2, or $p_1 - p_2 = 0$.
H_a : The rate of heart attacks for physicians taking aspirin are lower than the rate of those taking placebo in study 1, or $p_1 - p_2 < 0$.
Suppose the population proportions are unknown, but since $n_1\hat{p}_1 = 169 \gg 5$, $n_1(1-\hat{p}_1) = 3\,249 - 169 = 3\,260 \gg 5$, $n_2\hat{p}_2 = 88 \gg 5$, and $n_2(1-\hat{p}_2) = 1\,710 - 88 = 1\,622 \gg 5$, we can construct the null distribution for the test statistic as:

$$Z = \frac{\hat{p}_1 - \hat{p}_2}{\sqrt{\hat{p}(1-\hat{p})(1/3\,429 + 1/1\,710)}} \sim \mathcal{N}(0,1),$$

where we have the pooled sample proportion as:

$$\hat{p} = \frac{169+88}{3\,429+1\,710} = \frac{258}{5\,139}.$$

Hence, we now have the z-score as:

$$z^* = \frac{169/3\,429 - 88/1\,710}{\sqrt{258/5\,139 \times 4\,882/5\,139 \times (1/3\,429 + 1/1\,710)}} \approx -0.336641.$$

Therefore, we have the p-value as:

$$p\text{-value} = P(Z < z^*) \approx \Phi(-0.336641) \approx 0.368194 > 0.05 = \alpha,$$

hence we retain (or fail to reject) H_0.
Therefore, we retain H_0 at significant level $\alpha = 0.05$, and we have $\boxed{\text{no sufficient evidence to conclude}}$ that the rate of heart attacks for physicians taking aspirin is significantly lower than the rate of those

on placebo in study 2.

(d) First, the first experiment is conducted under double blind experiment but the second is not double blind so this might impact the mental status for the controlled group.
Also, we note that study 1 is conducted on a much larger sample size, hence its significance is larger, so it is more likely to reject H_0.

⌟

Problem 8.8. From 2017 to 2018, a large sample of prospective law students who sat for the Law School Admissions Test (LSAT) was taken in by the Law School Admissions Council. In the table below are the average scores by major. You are interested in determining whether LSAT scores of STEM majors differ from those of Philosophy majors on average.

STEM

Major	Applicants	Score
Mathematics	293	161.6
Biology	1101	154.9
Chemistry	246	155.5
Environmental Sciences	420	155.7
Electrical Engineering	177	158.1
Mechanical Engineering	197	157.3

Philo.

Major	Applicants	Score
Philosophy	2238	157.2

(a) Find the sample average LSAT score for all STEM majors. *Hint*: It is not simply the average of the Score column.

(b) Since the standard deviations were not provided, We will estimate the standard deviation of LSAT scores with $s = 15$ for both STEM and Philosophy majors. This is because the range of LSAT scores is $180 - 120 = 60$, and $60/4 = 15$. Does the data suggest there is a significant difference in the average LSAT score for STEM and Philosophy majors? Use $\alpha = 0.1$.

Sol'n.

(a) The sample average for the STEM major is a weighted average, which is:

$$\bar{x}_S = \frac{293 \times 161.6 + 1\,101 \times 154.9 + \cdots + 197 \times 157.3}{293 + 1\,101 + \cdots + 197} \approx \boxed{156.332169}.$$

(b) Now, we want to test using hypothesis test, our test hypotheses are:

H_0 : The average LSAT of all STEM and Philosophy majors are the same, or $\mu_S - \mu_P = 0$.

H_a : The average LSAT of all STEM and Philosophy majors are different, or $\mu_S - \mu_P \neq 0$.

Note that with the sample size $n_S = 2\,434 \gg 30$, and $n_N = 2\,238 \gg 30$, we may construct our null distribution, as:

$$Z = \frac{\overline{X_1} - \overline{X_2} - 0}{\sqrt{\frac{15^2}{2\,434} + \frac{15^2}{2\,238}}} \approx \frac{\overline{X_1} - \overline{X_2}}{\sqrt{87\,600/453\,941}} \sim \mathcal{N}(0,1).$$

Following that, we calculate the test statistics, which is:

$$z^* \approx \frac{156.332169 - 157.2}{\sqrt{87\,600/453\,941}} \approx -1.975526.$$

Following that, we find the p-value as two-sided test, which is:

$$p\text{-value} = P(Z < -|z^*|) + P(Z > |z^*|) \approx 2\Phi\left(-1.975526\right) \approx 0.048208 \leq 0.1 = \alpha,$$

hence we reject H_0.

Therefore, we reject H_0 at significant level $\alpha = 0.1$, and we have $\boxed{\text{sufficient evidence to conclude}}$ that there is a difference in the LSAT score for STEM and Philosophy majors.

⌟

9 Small Sample Inference

Problem 9.1. (Conceptual Understanding). A student wants to know if the pass rate of a particular professor is below 70%. To investigate, they set up the hypotheses $H_0 : p = 0.7$ against $H_a : p < 0.7$. They ask 10 randomly chosen students who took the professor's class before whether they passed or not. Of those 10, half of them said they passed. Since their sample is small, the student computes a test t-statistic of:

$$t^* = \frac{\hat{p} - p_0}{\sqrt{\dfrac{p_0(1-p_0)}{n}}} = \frac{0.5 - 0.7}{\sqrt{\dfrac{0.7 \cdot 0.3}{10}}} \approx -2.0702.$$

With $df = 9$, this produces a p-value of about $0.0342 < 0.05 = \alpha$, which leads the student to reject H_0. However, the student has made an egregious error with this test, explain.

Sol'n. Note that there is not a t-test for proportions. The assumption for the t-test is that when the population distribution is normal.

Note that here, p is the probability that it passes, while there are 10 random students. However, the probability that n out of 10 students passed, we have:

$$\binom{10}{n} p^n (1-p)^{10-n},$$

which is a binomial distribution. As the proportion does not follow a normal distribution, so the conditions for a small sample inference test cannot be met. Hence, the above argument is invalid. ⌟

Problem 9.2. (Lobster). In a study of the *Thenus orientalis* lobster and their infestation of barnacles, researchers measured the lengths (in mm) of the carapace (the hard upper shell) of 10 randomly selected lobsters caught in the seas near Singapore. The measurements are recorded below:

$$78 \quad 66 \quad 65 \quad 63 \quad 60 \quad 60 \quad 58 \quad 56 \quad 52 \quad 50.$$

(a) Find the 99% confidence interval for the mean carapace length of the *T. orientalis* lobsters.

(b) Find the 95% confidence interval...

(c) Find the 90% confidence interval...

Sol'n. First, we assume that the measurement for lobsters are normal and independent. In prior for any specific question, to facilitate the questions, we calculate the follows about the statistics:

- The sample mean and standard deviation, we have:

$$\bar{x} = \frac{78+66+65+\cdots+50}{10} = 60.8,$$

$$s = \sqrt{\frac{(78-\bar{x})^2+(66-\bar{x})^2+\cdots+(50-\bar{x})^2}{9}} \approx 7.969386.$$

- Then, since there are 10 samples, so we have $df = 9$, the confidence interval with $(1-\alpha)\times 100\%$ confidence is:

$$CI_{(1-\alpha)\times 100\%} = \left[60.8 - t_{9,\alpha/2} \times \frac{7.969386}{\sqrt{10}}, 60.8 + t_{9,\alpha/2} \times \frac{7.969386}{\sqrt{10}}\right].$$

Now, we discuss the confidence interval for each α

(a) For 99% confidence interval, we have $\alpha/2 = 0.005$, by consulting with the table:

$$z_{9,0.005} = 3.250,$$

so we have the confidence interval:

$$CI_{99\%} = \left[60.8 - 3.250 \times \frac{7.969386}{\sqrt{10}}, 60.8 + 3.250 \times \frac{7.969386}{\sqrt{10}}\right] = \boxed{[52.609541, 68.990459]}.$$

(b) For 95% confidence interval, we have $\alpha/2 = 0.025$, by consulting with the table:

$$z_{9,0.025} = 2.262,$$

so we have the confidence interval:

$$CI_{99\%} = \left[60.8 - 2.262 \times \frac{7.969386}{\sqrt{10}}, 60.8 + 2.262 \times \frac{7.969386}{\sqrt{10}}\right] = \boxed{[55.099441, 66.500559]}.$$

(c) For 90% confidence interval, we have $\alpha/2 = 0.05$, by consulting with the table:

$$z_{9,0.05} = 1.833,$$

so we have the confidence interval:

$$CI_{99\%} = \left[60.8 - 1.833 \times \frac{7.969386}{\sqrt{10}}, 60.8 + 1.833 \times \frac{7.969386}{\sqrt{10}}\right] = \boxed{[56.180581, 65.419419]}.$$

┘

Problem 9.3. (Cholesterol). A doctor is interested in determining whether the average serum total cholesterol level (in mg/dL) of adults in their region is different than the national average of 191 mg/dL as reported by the Center for Disease Control and Prevention. They take a random sample of 15 of their patients and records the results:

$$\begin{array}{ccccc} 273 & 304 & 178 & 225 & 227 \\ 247 & 212 & 254 & 139 & 282 \\ 261 & 169 & 221 & 222 & 229. \end{array}$$

Does the data suggest that the average cholesterol level in the doctor's region is different than that of the national average? Use $\alpha = 0.01$.

Sol'n. Since we have a small sample, we would want to create a hypothesis t-test. First, we form our null and alternative hypothesis:

- H_0: $\mu = 191$ mg/dL, or the average cholesterol level is not different from the national average.
- H_a: $\mu \neq 191$ mg/dL, or the average cholesterol level is different from the national average.

As the sample size is small $15 < 30$ and we assume that the population is approximately normal. With $df = 14$, we then construct the distribution:

$$T_{14} = \frac{\overline{X} - 191}{s/\sqrt{15}} \sim \mathcal{T}_{14}.$$

Now, we calculate the sample statistics, *i.e.*, the mean and standard deviation:

$$\bar{x} = \frac{273 + 304 + \cdots + 229}{15} \approx 229.533333,$$

$$s = \sqrt{\frac{(273-\bar{x})^2 + (304-\bar{x})^2 + \cdots + (229-\bar{x})^2}{14}} \approx 43.965679.$$

Then, the test statistics is:

$$t_{14} \approx \frac{229.533333 - 191}{43.965679/\sqrt{15}} \approx 3.394442.$$

Given $\alpha = 0.01$, our rejection region is:

$$R_0 = (-\infty, -t_{14,0.005}] \sqcup [t_{14,0.005}, +\infty) = (-\infty, -2.977] \sqcup [2.977, +\infty).$$

Notice that $t_{14} \approx 3.394442 \in R_0$, we can reject H_0.

Thus, we reject H_0 at significance level $\alpha = 0.01$, and we $\boxed{\text{have evidence to conclude}}$ that the average serum total cholesterol level differs from 191 mg/dL, the national average. ⌟

Problem 9.4. (Ads and Sales). Suppose you own a small company, STATS. CO. You hope to increase your sales by running online advertisements on YouTube. To justify the cost of advertising, you require a daily average sales of over \$10,000. To test this, you run your ads for one week (7 days), and finds that the daily average sales for that week was \$10,972 with a variance of $\$^2$810,000. Use hypothesis testing at significance level $\alpha = 0.05$ to decide whether you should continue running online ads for STATS CO.

Sol'n. Since we have a small sample, we would want to create a hypothesis t-test. First, we form our null and alternative hypothesis:

- H_0: $\mu = 10\ 000$, or daily average sale is \$ 10 000.
- H_a: $\mu > 10\ 000x$, or daily average sale is above \$ 10 000.

As the sample size is small $7 < 30$ and we assume that the population is approximately normal. With $df = 6$, we then construct the distribution:

$$T_6 = \frac{\overline{X} - 10\ 000}{s/\sqrt{7}} \sim \mathcal{T}_6.$$

Now, we calculate the sample statistics, *i.e.*, the standard deviation:

$$s = \sqrt{810\ 000} = 900.$$

Then, the test statistics is:

$$t_6 = \frac{10\ 972 - 10\ 000}{900/\sqrt{7}} \approx 2.857411.$$

Given $\alpha = 0.05$ (single tailed), our rejection region is:

$$R_0 = [t_{6,0.05}, +\infty) = [1.943, +\infty).$$

Notice that $t_6 \approx 2.857411 \in R_0$, we can reject H_0.

Thus, we reject H_0 at significance level $\alpha = 0.05$, and we have evidence to conclude that the daily average sale is above \$10 000, so with $\alpha = 0.05$, we $\boxed{\text{should continue running online ads for STATS CO}}$. ⌟

Problem 9.5. (Virtual Reality in the Classroom). In an article about the impact of VR on visualizing partial derivatives, the researchers randomly sorted students into two groups:

- The Control Group: These students received classroom instruction from a professor on partial derivatives and were given a quiz at the end of the lecture.
- The Treatment Group: These students did not receive classroom instruction, but instead, went to the Virtual Reality Laboratory which allowed students to visualize and manipulate the surfaces and contour maps in 3D. Afterwards, these students took the same quiz that was given to the control group.

Suppose that the quiz grades (out of 10) of 8 students from the treatment group and 12 students from the control group are randomly select to be evaluated. A summary of the data is given below:

	Treatment	**Control**
Mean	$\overline{x_{\text{tr.}}} = 7.808$	$\overline{x_{\text{co.}}} = 8.904$
Std. Dev.	$s_{\text{tr.}} = 1.3153$	$s_{\text{co.}} = 1.2246$

(a) Find a 95% confidence interval for the true difference in mean quiz scores $\mu_{\text{co.}} - \mu_{\text{tr.}}$.

(b) Use part (a) to test whether there is enough evidence to conclude that there is a significant ($\alpha = 0.05$) difference in the means, do not evaluate the test statistic t^*. Hint: Look back at Problem 8.2 (Conceptual Understanding) Part b.

Sol'n.

(a) Here, we have that the true difference can be modeled by $\overline{X_{\text{co.}}} - \overline{X_{\text{tr.}}}$.
Since the sample size is small $8 < 30$ and we assume that the population is approximately normal. With $df = 14$, we then construct the distribution:

$$T_{14} = \frac{\overline{X_{\text{co.}}} - \overline{X_{\text{tr.}}}}{s\sqrt{\frac{1}{8} + \frac{1}{12}}} \sim \mathcal{T}_{14}.$$

Now, we calculate the sample statistics, *i.e.*, mean and the standard deviation:

$$\bar{x} = \overline{x_{\text{co.}}} - \overline{x_{\text{tr.}}} = 8.904 - 7.808 = 1.096$$

$$s^2 = \frac{7s_{\text{tr.}}^2 + 11s_{\text{co.}}^2}{18} = \frac{1}{18}(7 \times 1.3153^2 + 11 \times 1.2246^2) \approx 1.589233,$$

$$s \approx 1.260647.$$

Then, the test distribution is:

$$T_{18} \approx \frac{\overline{X_{\text{co.}}} - \overline{X_{\text{tr.}}}}{1.260647 \cdot \sqrt{5/24}} \sim \mathcal{T}_{18}.$$

Therefore, with 95% confidence interval, we have:

$$t_{18,0.025} = 2.101,$$

so we have the confidence interval being:

$$CI_{95\%} \approx [1.096 - 2.101 \times 0.575404, 1.096 + 2.101 \times 0.630324] = \boxed{[-0.112924, 2.304924]}.$$

(b) Then, note that $0 \in CI_{95\%}$, this implies that we have p-value > 0.05, so we retain (or fail to reject) H_0. In particular, we would have the null and alternative hypothesis as:

- H_0: $\mu_{\text{co.}} - \mu_{\text{tr.}} = 0$, or the treatment and control group do not have different results.
- H_a: $\mu_{\text{co.}} - \mu_{\text{tr.}} \neq 0$, or the treatment and control group have different results.

Thus, we retain H_0 at significant level of $\alpha = 0.05$, and we $\boxed{\text{do not have enough evidence to conclude}}$ that there is a significant difference between the mean of the treatment and control group.

⌟

Problem 9.6. (Cable Strength). A manufacturing company will be purchasing a large order of towing cables soon. There are two options for cables, type A and type B, the company would like to purchase the strongest cables. Testing is expensive, so they can only conduct 9 trails to test the breaking strength of each type of cable. The data is collected below:

Breaking Strength in Pounds

Type A	35615	25732	32325	37370	29724	41445	35481	43239	36797
Type B	37740	35094	29135	39520	26845	29917	33136	38322	34952

Based on these trials, the company believes $\mu_A > \mu_B$ (where μ_A and μ_B are the average breaking strengths of Type A and Type B cables respectively).

(a) State the hypothesis H_0 and H_a and calculate the test t-statistic.

(b) Calculate the degrees of freedom. What assumption did you need to make? Is it reasonable? Explain.

(c) Give an upper and lower bound on the p-value using the table posted in Canvas.

(d) Based on parts (a) – (c), should retain or reject H_0? Explain what this means in plain English. (Assume a significance level of $\alpha = 0.05$.)

Sol'n. Since we have a small sample, we would want to create a hypothesis t-test.
Initially, we want to find the sample statistics, *i.e.*, the means and samples (since they have the same sample size, we have):

$$\overline{x_A} \approx 35\,303.111111, \qquad \overline{x_B} \approx 33\,851.222222, \qquad \overline{x_A} - \overline{x_B} \approx 1\,451.888888$$

$$s_A \approx 5\,465.066089, \qquad s_B \approx 4\,434.541205 \qquad s = \sqrt{\frac{s_A^2 + s_B^2}{2}} \approx 4\,976.550163.$$

(a) First, we form our null and alternative hypothesis:

- H_0: $\boxed{\mu_A - \mu_B = 0 \text{ lbs.}}$, or the average breaking strength for the two types are the same.
- H_a: $\boxed{\mu_A - \mu_B > 0 \text{ lbs.}}$, or the average breaking strength for Type A is larger than Type B.

(b) As the sample size is small $9 < 30$, in order to do the t-test, we must have made the following assumptions:

- The $\boxed{\text{populations are approximately normal}}$, *i.e.*, their differences is approximately normal;
- The $\boxed{\text{two populations have the same standard deviation}}$, or $\sigma_1 = \sigma_2$.

These assumptions are reasonable as:

- For the cables to be manufactured, they should follow normal distribution;
- We have the ratio of the variances as $s_1^2/s_2^2 \approx 1.518775 < 3$ and $s_2^2/s_1^2 \approx 0.658425 < 3$, so the assumption is reasonable.

Then, we have the degree of freedom as:

$$df = 9 + 9 - 2 = \boxed{16}.$$

(c) Here, we then construct the distribution:

$$T_{16} \approx \frac{\overline{X_A} - \overline{X_B}}{4\,976.550163 \cdot \sqrt{\dfrac{1}{9} + \dfrac{1}{9}}} \approx \frac{\overline{X_A} - \overline{X_B}}{2\,345.968245} \sim \mathcal{T}_{16}.$$

Note that if we attempt to calculate the probability, we have:

$$t^* \approx \frac{1\,451.888888}{2\,345.968245} \approx 0.618887,$$

then, by observing the t distribution table, we have:

$$t_{16,0.5} = 0 \le t^* \approx 0.618887 \le t_{16,0.25} = 0.69,$$

so we know that:

$$0.25 \le P(T_{16} \ge t^*) \le 0.5.$$

(d) Note that $P(T_{16} \ge t^*) \ge 0.25 > 0.05 = \alpha$, so we retain (or fail to reject) H_0. This implies that we $\boxed{\text{do not have enough evidence to conclude}}$ that Type A cable has a larger breaking strength than table B in pounds, *i.e.*, we do not have sufficient evidence to conclude $\mu_A > \mu_B$.

┘

Problem 9.7. (Auto Insurance). In the table below are the annual premiums for a male, licensed 6 to 8 years diving a Honda Accord with less than 15000 miles and has no accidents or violations across different California cities:

City	GEICO	21st Century
Long Beach	$2780	$2352
Pomona	$2411	$2462
San Bernadino	$2261	$2284
Moreno Valley	$2263	$2520

(a) Why would you expect these pairs of observations to be dependent?

(b) Does the data provide sufficient evidence that there is a difference in the average annual premiums between GEICO and 21st Century? Test using $\alpha = 0.01$.

(c) Find a 99% confidence interval for the difference in average annual premiums for GEICO and 21st Centrury insurance.

Sol'n.

(a) There are dependence between the annual premiums for each city. Notice that we are having the $\boxed{\text{same research subjects}}$, they are dependent on each other. Specifically, given different factors about the $\boxed{\text{environment of respective cities}}$, such as the weathering, non-accidental damages, or even income of a city. Hence it is unreasonable to compare across the cities.

(b) Here, we first calculate the differences:

City	GEICO	21st Century	*Difference*
Long Beach	$2 780	$2 352	$ 428
Pomona	$2 411	$2 462	$ − 51
San Bernadino	$2 261	$2 284	$ − 23
Moreno Valley	$2 263	$2 520	$ − 257

Correspondingly, the mean and standard deviation are:

$$\bar{d} = 24.25, \qquad s_d \approx 288.681341.$$

Then, we form the null and alternative hypothesis:

- H_0: $\mu_d = 0$, or the average annual premiums between GEICO and 21st Century are the same.
- H_a: $\mu_d \neq 0$, or the average annual premiums between GEICO and 21st Century are different.

Then, we form the distribution:

$$T_3 \approx \frac{D}{288.681341/\sqrt{4}} = \frac{D}{144.340671} \sim \mathcal{T}_3.$$

Thereby, the test statistic is:

$$t_3 = \frac{24.25}{144.340671} \approx 0.168005,$$

and with $df = 3$, we have the rejection region as:

$$R_0 = (-\infty, -t_{3,0.005}] \sqcup [t_{3,0.005}, +\infty) = (-\infty, -5.841] \sqcup [5.841, \infty).$$

Notice that $t_3 \notin R_0$, so we retain (or fail to reject) H_0.

This implies that we $\boxed{\text{do not have enough evidence to conclude}}$ that there is a difference in the average annual premiums between GEICO and 21st Century, *i.e.*, we do not have sufficient evidence to conclude $d \neq 0$.

(c) Here, the below difference is the subtraction of cost for 21st Century from GEICO. Given $\alpha/2 = 0.005$, we have $t_{3,0.005} = 5.841$, so our confidence interval is:

$$IC_{99\%} \approx \left[24.25 - 5.841 \times \frac{288.691341}{\sqrt{4}}, 24.25 + 5.841 \times \frac{288.691341}{\sqrt{4}}\right] \approx \boxed{[-818.873061, 857.373061]}.$$

⌟

Problem 9.8. (Runners and Cyclists). Creatine phosphokinase (CPK), a measure of muscle damage, was determined for each of 10 competitive runners and 10 competitive cyclists in an article published in the American Journal of Sports Medicine. For each athlete, pressures were measured from rest, and again after 15 minutes of 80% VO2 maximum exercise. The data summary – CPK values in units/liter – is as follows:

Runners

	Mean	Std. Dev.
Before Exercise	255.63	115.48
After Exercise	284.75	132.64
Difference	29.13	21.01

Cyclists

	Mean	Std. Dev.
Before Exercise	173.8	60.69
After Exercise	177.1	64.53
Difference	3.3	6.85

(a) For this particular study, explain briefly why a paired-difference test would be necessary to determine whether there is significance difference in the before and after exercise levels of CPK.

(b) Does the data suggest that there is significant difference in the before and after exercise levels of CPK in runners? Use $\alpha = 0.05$.

(c) Repeat the same analysis in part (b), but for cyclists.

(d) Suppose you are a doctor, and your patient suffers from chronic anterior compartment syndrome (an exercise-induced muscle and nerve condition that causes pain, swelling and sometimes disability in the affected muscles of the legs or arms). What might you recommend to your patient based on your results in parts (b) and (c)?

Sol'n.

(a) First, since our research is researching on the $\boxed{\text{same sets of subjects}}$ with similarities, it is important to incur a paired difference check. Note that the runners and cyclists exhibits very different training procedures and adapts different CPK changes from exercises. Here, we must compare the pair differences as $\boxed{\text{they are dependent}}$.

(b) Since we have a small sample, we would want to create a hypothesis t-test. First, we form our null and alternative hypothesis:

- H_0: $\mu_B - \mu_A = 0$, or there are no differences in CPK values before and after exercise for runners.
- H_a: $\mu_B - \mu_A \neq 0$, or there are differences in CPK values before and after exercise for runners.

As the sample size is small $10 < 30$ and we assume that the population is approximately normal. With $df = 18$, we then construct the distribution:

$$T_9 = \frac{\overline{X_B} - \overline{X_A}}{21.01/\sqrt{10}} \sim \mathcal{T}_9.$$

Then, the test statistics is:

$$t_9 = \frac{29.13}{21.01/\sqrt{10}} \approx 4.384443.$$

Given $\alpha = 0.05$ (single tailed), our rejection region is:

$$R_0 = (-\infty, -t_{9,0.025}] \sqcup [t_{9,0.025}, +\infty) = (-\infty, -2.262] \sqcup [2.262, +\infty).$$

Notice that $t_9 \approx 4.384443 \in R_0$, we can reject H_0.
Thus, we reject H_0 at significance level $\alpha = 0.05$, and we $\boxed{\text{have evidence to conclude}}$ that there is significant difference in the before and after levels of CPK for runners.

(c) Since we have a small sample, we would want to create a hypothesis t-test. First, we form our null and alternative hypothesis:

- H_0: $\mu_B - \mu_A = 0$, or there are no differences in CPK values before and after exercise for cyclists.
- H_a: $\mu_B - \mu_A \neq 0$, or there are differences in CPK values before and after exercise for cyclists.

As the sample size is small $10 < 30$ and we assume that the population is approximately normal. With $df = 18$, we then construct the distribution:

$$T_9 = \frac{\overline{X_B} - \overline{X_A}}{6.85/\sqrt{10}} \sim \mathcal{T}_9.$$

Then, the test statistics is:

$$t_9 = \frac{3.3}{6.85/\sqrt{10}} \approx 1.523433.$$

Given $\alpha = 0.05$ (double tailed), our rejection region is:

$$R_0 = (-\infty, -t_{9,0.025}] \sqcup [t_{9,0.025}, +\infty) = (-\infty, -2.262] \sqcup [2.262, +\infty).$$

Notice that $t_9 \approx 1.523433 \notin R_0$, we retain (or fail to reject) H_0.
Thus, we retain H_0 at significance level $\alpha = 0.05$, and we $\boxed{\text{do not have evidence to conclude}}$ that there is significant difference in the before and after levels of CPK for cyclists.

(d) Based on part (b) and (c), we have sufficient evidence to conclude that there is a significant difference in level and by observation, this is an increase, guaranteed with 90% confidence, so we would suggest the patient to not do intensive running exercises. For cycling, we with 90% confidence, we do not have sufficient evidence that it causes increase in CPK, so we do not enough support to suggest the patient to not cycle.
Rmk. Technically, we also do not have evidence to conclude that cycling has no impacts. However, we could suggest them to do cycling (over running) if they really want to exercise, but this is unsupported, but at least has smaller impact than running.

┘

10 Inference on Population Variance

Problem 10.1. (Conceptual Understanding). Determine whether each of the indicated statements are True or False. Explain your answers.

(a) A researcher tests the hypothesis $H_0 : \sigma^2 = \sigma_0^2$ against $H_a : \sigma^2 \neq \sigma_0^2$ at significance level 0.05 and finds sufficient evidence to reject H_0. (T/F): Then the researcher would have reached the same conclusion (i.e. reject H_0) at any significance level $\alpha > 0.05$.

(b) A researcher finds that their hypothesized variance σ_0^2 is not contained in a 95% confidence interval they calculated for the true variance σ^2. (T/F): Then for all $\alpha < 0.05$, σ_0^2 will not be contained in the $(1 - \alpha) \times 100\%$ confidence interval.

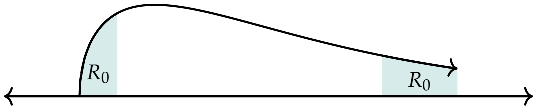
Sol'n.

(a) $\boxed{\text{True}}$. Being rejected implies that the p-value bounded above by 0.05. *Although the p-value cannot be obtained by the table, the value still exists*, thus, we immediately obtain the relationship:

$$p\text{-value} \leq 0.05 < \alpha, \ \forall \alpha > 0.05,$$

hence we still reject H_0.

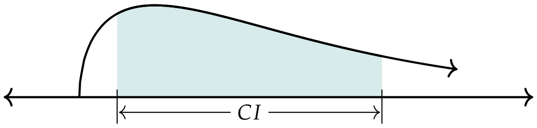
On the other hand, we can also consider about the rejection region (shaded region):



When we have a larger α value, it will has a larger rejection region, which includes the current shaded area, so the research would still reject H_0.

(b) $\boxed{\text{False}}$. The easiest approach is to consider $\alpha \to 0^+$, this implies that we have to be more certain that σ_0^2 is contained, and when $\alpha = 0$, then the confident interval must have contained all non-negative value, hence including σ_0^2.

Formally, as the confidence interval can be interpreted as the region in the middle, shaded:



When we have a smaller α value, it will has a larger confidence interval, so not in the current shaded confidence interval does not imply that it will not appear in the larger intervals.

⌟

Problem 10.2. (Instrument Precision). An engineer creates an instrument to quickly measure the height in ft. of tall buildings. A sample of four readings on the same building yielded:

$$353 \qquad 351 \qquad 351 \qquad 355.$$

(a) Test the null hypothesis that $\sigma = 0.7$ against $\sigma > 0.7$ at significance level $\alpha = 0.05$.

(b) Find a 90% confidence interval for the true variance of the instrument readings.

Sol'n.

(a) Here we want to first calculate the sample variance of the given 4 readings:

$$\bar{x} = \frac{705}{2}, \qquad s^2 = \frac{11}{3}.$$

Since we know that the standard deviation is non-negative, we create the hypotheses being:

- H_0: $\sigma^2 = 0.49$, or equivalently $\sigma = 0.7$,
- H_a: $\sigma^2 > 0.49$, or equivalently $\sigma > 0.7$.

Here, we assume that the measurement of the height of the building is approximately normal, so we can construct the χ-square distribution, thus:

$$\chi^2 = \frac{(n-1)s^2}{\sigma_0^2} \sim \chi^2_{n-1}.$$

Given such, our test statistics is:

$$\chi^2 = \frac{(4-1) \times 11/3}{0.49} \approx 22.448980.$$

With a degree of freedom being $df = 3$ and $\alpha = 0.05$, the rejection region for the single tailed test is:

$$R_0 = [\chi^2_{0.05,df=3}, +\infty) \approx [7.815, +\infty).$$

Notice that $\chi^2 \approx 22.448980 \in R_0$, so we reject H_0.
Therefore, at $\alpha = 0.05$, we reject that $\sigma = 0.7$, hence we $\boxed{\text{have sufficient evidence to conclude}}$ that $\sigma > 0.07$.

(b) To find the confidence interval with $\alpha = 0.1$ and $df = 3$, we have the critical values being:

$$\chi_{0.05,df=3} \approx 7.815, \qquad \chi_{0.95,df=3} \approx 0.352,$$

giving us the confidence interval of:

$$CI_{90\%} = \left[\frac{3 \times 11/3}{\chi_{0.05,df=3}}, \frac{3 \times 11/3}{\chi_{0.95,df=3}}\right] \approx \left[\frac{11}{7.815}, \frac{11}{0.352}\right] \approx \boxed{[1.407550, 31.25]}.$$

⌟

Problem 10.3. (Light Bulbs). A manufacturer of industrial light bulbs wants to control the variability in the lifetimes of their bulbs so that σ is less than 150 hr. A sample of 20 bulbs tests produced the following lengths of life in hours:

2100	2302	1951	2067	2415	1883	2101	2146	2278	2019
1924	2183	2077	2392	2286	2290	1946	2161	2253	1827.

Does the sample indicate that the manufacturer is achieving their goal? Use $\alpha = 0.05$.

<u>Sol'n.</u> Here, we need to first obtain the statistics on the 20 bulbs test:

$$\bar{x} = 2130.05, \qquad s^2 \approx 29\,075.734211.$$

Since we know that the standard deviation is non-negative, we create the hypotheses being:

- H_0: $\sigma^2 = 22\,500$, or equivalently $\sigma = 150$,
- H_a: $\sigma^2 < 22\,500$, or equivalently $\sigma < 150$.

Here, we assume that the measurement of the lifetimes of the bulbs is approximately normal, so we can construct the χ-square distribution, thus:

$$\chi^2 = \frac{(n-1)s^2}{\sigma_0^2} \sim \chi^2_{n-1}.$$

Given such, our test statistics is:

$$\chi^2 \approx \frac{(20-1) \times 29\,075.734211}{22\,500} \approx 24.552842.$$

With a degree of freedom being $df = 19$ and $\alpha = 0.05$, the rejection region for the single tailed test is:

$$R_0 = (-\infty, \chi^2_{0.95, df=19}] \approx (-\infty, 10.117].$$

Notice that $\chi^2 \approx 24.552842 \notin R_0$, so we retain H_0.
Therefore, at $\alpha = 0.05$, we fail to reject that $\sigma = 150$, hence we $\boxed{\text{do not have sufficient evidence to conclude}}$ that $\sigma < 150$. ⌟

Problem 10.4. (SAT Scores). The SAT subject exams in chemistry and physics for two groups of 15 students each electing take theses tests are given below:

Chemistry	**Physics**
$\bar{x} = 664$	$\bar{x} = 658$
$s = 114$	$s = 103$

To use the two sample t-test with a pooled estimate of σ^2, you must assume that the two population variances are equal. Test this assumption using the F-test for equality of variances.

Sol'n. Note that an α value is not given for this problem, so we use $\alpha = 0.05$ by default.
First, we want to form our hypothesis test with the hypotheses:

- H_0: $\frac{\sigma_C^2}{\sigma_P^2} = 1$, in which they have the same variance,
- H_a: $\frac{\sigma_C^2}{\sigma_P^2} \neq 1$, in which they have different variance.

If we assume that the two populations are approximately normal, and under H_0, we have that $\sigma_C^2 = \sigma_P^2$, so we form our null distributions as:

$$F = \frac{S_C^2}{S_P^2} \sim F_{n_C-1,n_P-1},$$
$$F^{-1} = \frac{S_P^2}{S_C^2} \sim F_{n_P-1,n_C-1}.$$

so we have the test statistics as, which is:

$$F^* = \frac{s_C^2}{s_P^2} = \frac{114^2}{103^2} \approx 1.224998,$$
$$F^{-1^*} = \frac{s_P^2}{s_C^2} = \frac{103^2}{114^2} \approx 0.816328.$$

Here, we need to take care of the two respectively rejection regions, which is:

$$\text{For } F: \quad R_0 = [F_{0.025,df_1=14,df_2=14}, +\infty) = [2.8621, +\infty),$$
$$\text{For } F^{-1}: \quad \widetilde{R_0} = [F_{0.025,df_1=14,df_2=14}, +\infty) = [2.8621, +\infty).$$

Note that both test statistics do not belong to their respective rejection region, so we retain H_0 at $\alpha = 0.05$. Therefore, at $\alpha = 0.05$, we fail to reject $\frac{\sigma_C^2}{\sigma_P^2} = 1$, so we $\boxed{\text{do not have sufficient evidence to conclude}}$ that the variances of the two populations are different. ⌟

Problem 10.5. (Further χ^2 Applications). There are several other instances where the chi-square distribution appears. One of those, which we'll see later in the semester, is called the "goodness-of-fit" test, this exercise will help us understand why the formulas in that test are the way that they are. Suppose you have an experiment which can be modeled by a binomial RV $X \sim \text{Binomial}(n, p)$.

(a) Show that:

$$\frac{(X-np)^2}{np} + \frac{\big((n-X)-nq\big)^2}{nq} = \frac{(X-np)^2}{npq}.$$

Hint: You could multiply everything out on each side, compare them, and remark that they're the same. However, you should look for a simpler way with the fact that $p + q = 1$.

Proof. Here, we start from the LHS with:

$$\begin{aligned}
\frac{(X-np)^2}{np} + \frac{\big((n-X)-nq\big)^2}{nq} &= \frac{X^2 - 2npX + n^2p^2}{np} + \frac{n^2 - 2nX + X^2 - 2n^2q + 2nqX + n^2q^2}{nq} \\
&= \frac{qX^2 - 2npqX + n^2p^2q + n^2p - 2npX + pX^2 - 2n^2pq + 2npqX + n^2pq^2}{npq} \\
&= \frac{(p+q)X^2 - 2npX + n^2p(pq + 1 - 2q + q^2)}{npq} \\
&= \frac{X^2 - 2npX + n^2p\big(pq + (1-q)^2\big)}{npq} \\
&= \frac{X^2 - 2npX + np\big(pq + p^2\big)}{npq} = \frac{X^2 - 2npX + np\big(p(q+p)\big)}{npq} \\
&= \frac{X^2 - 2npX + np^2}{npq} = \frac{(X-np)^2}{npq},
\end{aligned}$$

as desired. □

Once you run the experiment and observe the number of successful outcomes, X, you have a sample variance of $S^2 = (X - np)^2$. Since the true variance of X is $\sigma^2 = npq$, and the Binomial(n, p) distribution approximates a $\mathcal{N}(np, \sqrt{npq})$ distribution by CLT, we have that the expression in (a) has approximately a χ^2 distribution with $df = 1$.

(b) Suppose your friend gives you a coin you suspect is biased towards heads. You flip the coin 100 times and see that 61 of those times the coin landed heads. Assuming the coin was fair, use a χ^2 distribution with 1 degree of freedom to estimate the probability of observing at least 61 heads. Compare this with the actual probability of observing at least 61 heads, assuming the coin is fair, as given by the Binomial$(100, 0.5)$ distribution. The Excel functions `CHISQ.DIST` and `BINOM.DIST` may be used here.

Sol'n. Here, assume that $n = 100$ and $p = q = 0.5$, we have:

$$\frac{(X-np)^2}{npq} = \frac{(X-50)^2}{25} \sim \chi^2_{df=1},$$

which, consequently, gives us that:

$$\begin{aligned} P(X \geq 61) &= P\left(\frac{(X-50)^2}{25} \geq \frac{121}{25} = 4.84\right) = P(\chi^2_{df=1} \geq 4.84) \\ &= 1 - \texttt{CHISQ.DIST}(4.84, 1, \texttt{TRUE}) \approx 1 - 0.972193 = \boxed{0.027807}. \end{aligned}$$

Then, we also calculate this probability directly with Binomial Distribution, giving us that:

$$P(X \geq 61) = 1 - \texttt{BINOM.DIST}(61, 100, 0.5, \texttt{TRUE}) \approx 1 - 0.989511 = \boxed{0.010489}.$$

┘

Problem 10.6. (QB Pass Completions). A college football coach is looking to two recruit one of two quarterbacks, Aaron or Bret. Not only does the coach want a quarterback with a high passing percentage, but they need to be consistent as well. That is, the variability in number of passes completed per game should be small. Below are the number of passes completed for Aaron and Bret during their senior high school seasons:

Aaron		**Brett**	
19	18	16	18
19	21	19	20
34	15	17	22
12	27	17	21
27	22	30	23

(a) Does the data indicate that there is a difference in the variability in the number of passes completed for the two quarterbacks? Use $\alpha = 0.05$.

(b) Test whether there is significant ($\alpha = 0.05$) difference in the average number of passes completed between the two quarterbacks. Explain any assumptions you made in your test and why they would be appropriate.

Sol'n.

(a) First, we calculate the test statistics of the two players:

$$\overline{x_A} = 21.4, \qquad s_A^2 = 41.6, \qquad \overline{x_B} = 20.3, \qquad s_B^2 = 16.9.$$

Here, we want to form our hypothesis test with the following hypotheses:

- H_0: $\frac{\sigma_A^2}{\sigma_B^2} = 1$, in which they have the same variance,
- H_a: $\frac{\sigma_A^2}{\sigma_B^2} \neq 1$, in which they have different variance.

If we assume that the two populations are approximately normal, and under H_0, we have that $\sigma_A^2 = \sigma_B^2$, so we form our null distributions as:

$$F = \frac{S_A^2}{S_B^2} \sim F_{n_A-1,n_B-1},$$
$$F^{-1} = \frac{S_B^2}{S_A^2} \sim F_{n_B-1,n_A-1}.$$

so we have the test statistics as, which is:

$$F^* = \frac{s_A^2}{s_B^2} = \frac{41.6}{16.9} \approx 2.461538,$$
$$F^{-1*} = \frac{s_B^2}{s_A^2} = \frac{16.9}{41.6} = 0.40625.$$

Here, we need to take care of the two respectively rejection regions, which is:

$$\text{For } F: \qquad R_0 = [F_{0.025,df_1=9,df_2=9}, +\infty) = [4.1020, +\infty),$$
$$\text{For } F^{-1}: \qquad \widetilde{R_0} = [F_{0.025,df_1=9,df_2=9}, +\infty) = [4.1020, +\infty).$$

Note that both test statistics do not belong to their respective rejection region, so we retain H_0 at $\alpha = 0.05$.
Therefore, at $\alpha = 0.05$, we fail to reject $\frac{\sigma_A^2}{\sigma_B^2} = 1$, so we $\boxed{\text{do not have sufficient evidence to conclude}}$ that there is a difference in the variability in the number of passes completed for the two quaterbacks.

(b) In doing so, we are forming another t-test on the difference between population means. This is because that $n_A = n_B = 10 \ll 30$. Here, in order to do so, we have to make the following assumptions:

- The populations of the number of passes completed per game are approximately normal, *i.e.*, their differences is approximately normal, which is in our assumption;
- The two populations have the same variance, or $\sigma_1^2 = \sigma_2^2$, which is retained assumption by the previous part.

Meanwhile, we form our hypotheses as:

- H_0: $\mu_A - \mu_B = 0$, or the average number of passes per game are the same,
- H_a: $\mu_A - \mu_B \neq 0$, or the average number of passes per game are different.

Then, since the population size are the same, we calculate the pooled standard deviation:

$$s = \sqrt{\frac{41.6 + 16.9}{2}} = \sqrt{29.25} \approx 5.408327.$$

Therefore, we construct the null distribution as:

$$T_{18} = \frac{\overline{X_A} - \overline{X_B}}{s \cdot \sqrt{1/10 + 1/10}} \approx \frac{\overline{X_A} - \overline{X_B}}{2.418677} \sim \mathcal{T}_{18}.$$

Then, we calculate the test statistics as:

$$t^* \approx \frac{21.4 - 20.3}{2.418677} \approx 0.454794.$$

Here, we note that the rejection region is:

$$R_0 = (-\infty, -t_{0.025,df=18}] \sqcup (t_{0.025,df=18}, +\infty) = (-\infty, -2.101] \sqcup [2.101, +\infty),$$

and notice that $t^* \notin R_0$, which implies that we retain H_0.
Therefore, at $\alpha = 0.05$, we fail to reject that $\mu_A = \mu_B$, so we $\boxed{\text{do not have significant evidence}}$ to conclude that the average passing number is different between Aaron and Brett.

⌟

Problem 10.7. (Battery Life). A manufacture of lithium batteries used in digital cameras suspects that one of the production lines is producing batteries with a wide variation in length of life. To test this theory, a randomly selected sample of 15 batteries from the suspected line was compared with randomly selected sample of 12 batteries from a line that was judged to be in control.
The length of time in hours until depletion to $0.85V$ with a 5-Ohm load was measured for both samples:

Suspect Line	**Control Line**
$\bar{x} = 9.40$	$\bar{x} = 9.25$
$s = 0.25$	$s = 0.12$

(a) Does the data provide sufficient evidence to indicate that batteries produced by the suspect line have larger variance in lifetimes than those produced by the line assumed to be in control? Test using $\alpha = 0.05$.

(b) Approximate the p-value for the test and interpret its meaning. Here, approximate means–is it greater than 0.1, between 0.05 and 0.1, between 0.025 and 0.05, etc. How do you know? Do not use a calculator.

Sol'n.

(a) Here, we want to form a hypothesis test on the test with the following hypothesis:

- H_0: $\frac{\sigma_S^2}{\sigma_C^2} = 1$, in which they have the same variance,
- H_a: $\frac{\sigma_S^2}{\sigma_C^2} > 1$, in which the suspect line has a higher variance.

If we assume that the two populations are approximately normal, and under H_0, we have that $\sigma_S^2 = \sigma_C^2$, so we form our null distribution (as this is right tail, we test on F) as:

$$F = \frac{S_S^2}{S_C^2} \sim F_{n_S-1,n_C-1},$$

so we have the test statistics as, which is:

$$F^* = \frac{s_S^2}{s_C^2} = \frac{0.25^2}{0.12^2} = \frac{625}{144} = 4.340278.$$

Here, we consider the rejection region for F, which is:

$$R_0 = [F_{0.05,df_1=14,df_2=11}, +\infty) = [2.7386, +\infty).$$

Note that F^* is in the rejection region, so we reject H_0 at $\alpha = 0.05$.
Therefore, at $\alpha = 0.05$, we reject $\frac{\sigma_S^2}{\sigma_C^2} = 1$, so we $\boxed{\text{have sufficient evidence to conclude}}$ that the suspect line has a higher variance compared to the control line.

(b) Observing from the table that:

$$F_{0.01,df_1=14,df_2=11} = 3.8640 \leq F^* \leq 4.5085 = F_{0.005,df_1=14,df_2=11},$$

which implies that:

$$\boxed{0.005 \leq p\text{-value} \leq 0.01},$$

which falls into the category of p-value ≤ 0.01, implying that H_a is highly significant.

⌟

11 ANOVA + Linear Regression

Problem 11.1. (Breakfast of Champions). In an experiment to determine the effect of nutrition on the attention spans of elementary school students, a group of 15 students were randomly assigned to each of three meal plans: no breakfast, light breakfast, and full breakfast. Their attention spans (in minutes) were recorded during a morning reading period and are shown in the table below:

No Breakfast	**Light Breakfast**	**Full Breakfast**
8	14	10
7	16	12
9	12	16
13	17	15
10	11	12

Construct an ANOVA table for this data and determine whether there is evidence to suggest a difference in average attention spans for at least one of the treatments.

Sol'n. For simplicity of notation, we enumerate *No Breakfast* as 1, *Light Breakfast* as 2, and *Full Breakfast* as 3. Here, we want to first calculate the following:

- The mean and variance for each sample:

$$\begin{aligned} \overline{x_1} &= 9.4, & s_1^2 &= 5.3, \\ \overline{x_2} &= 14, & s_2^2 &= 6.5, \\ \overline{x_3} &= 13, & s_3^2 &= 6. \end{aligned}$$

- Then, we calculate the overall sample mean as:

$$\overline{x} = \frac{1}{15}\sum_{i=1}^{3} n_i \overline{x_i} \approx 12.133333.$$

- Now, we have the sum of squares between treatments as:

$$SST = \sum_{i=1}^{3} n_i(\overline{x_i} - \overline{x})^2 \approx 58.533333.$$

- Following that, we calculate the sum of squares with treatments due to error:

$$SSE = \sum_{i=1}^{3}(n_i - 1)s_i^2 = 71.2.$$

- Hence, that leads to the total sum of squares, as:

$$\text{Totoal } SS = SST + SSE \approx 129.733333.$$

Then, we have sufficient information to fill in the ANOVA table:

Source	df	SS	MS	F
Treatment	$k-1=2$	58.533333	$\frac{SST}{k-1} = 29.266667$	$\frac{MST}{MSE} = 4.932584$
Error	$N-k=12$	71.2	$\frac{SSE}{N-k} = 5.933333$	
Total	$N-1=14$	129.733333		

Notice that we can construct a *Hypothesis test for one-way ANOVA table*, which has hypotheses as:

- H_0: There is no difference in mean of response for different treatments, *i.e.*, $\mu_1 = \mu_2 = \mu_3$;
- H_a: There is a difference in mean of response for different treatments, *i.e.*, $\mu_i \neq \mu_j$ for some $i \neq j$.

Here, we suppose that the populations are approximate normal and all populations have a common variance, we have our statistics with the following distribution:

$$F \sim F_{df_1=2, df_2=12},$$

and suppose $\alpha = 0.05$, we have:

$$F \approx 4.932584 \geq 3.8853 \approx F_{df_1=2, df_2=12, \alpha=0.05},$$

so we reject the null hypothesis. Hence, at the level of $\alpha = 0.05$, we $\boxed{\text{have sufficient evidence}}$ to conclude that there is a difference in average attention spans for at least one of the treatments. ⌟

Problem 11.2. (Trees in Fields). A local FFA club wants to study the fruit yield of three different types of apple tree *Malus pumila*, *Malus domestica 'Fuji'*, and *Malus domestica 'Gala'*. They suspect there might also be variation in the yield based on what field they are planted in, but that there will be no interaction between the species of tree and the field. To measure this, four different fields are partitioned into thirds and 10 randomly selected trees from each species of apple tree are planted in one of the partitions in each field. The average number apples produced by each tree is recorded over a 4 year period and the data sum of squares are calculated in recorded in the ANOVA table:

Source	df	SS	MS	F
Tree Species	?	285.5	?	?
Field	?	822.92	?	?
Error	?	1083.83	?	
Total	?	2192.25		

(a) Fill in the missing values.

(b) Is there evidence to suggest that there is a difference in average yield between species of tree? Use $\alpha = 0.05$.

(c) Is there evidence to suggest that there is a difference in average yield between the different fields?

Sol'n.

(a) First, we fill in the table correspondingly:

Source	df	SS	MS	F
Tree Species	$k-1=\boxed{2}$	285.5	$\frac{SST}{k-1}=\boxed{142.75}$	$\frac{MST}{MSE}\approx\boxed{0.790253}$
Field	$b-1=\boxed{3}$	822.92	$\frac{SSB}{b-1}\approx\boxed{274.306667}$	$\frac{MSB}{MSE}\approx\boxed{1.518541}$
Error	$(b-1)(k-1)=\boxed{6}$	1083.83	$\frac{SSE}{(b-1)(k-1)}\approx\boxed{180.638333}$	
Total	$bk-1=\boxed{11}$	2192.25		

(b) Here, we can construct a *Hypothesis test for simple block design ANOVA table*, which has hypotheses as:

- H_0: There is no difference in mean of response for different treatments (tree species), *i.e.*, $\mu_{T_1} = \mu_{T_2} = \mu_{T_3}$;
- H_a: There is a difference in mean of response for different treatments (tree species), *i.e.*, $\mu_{T_i} \neq \mu_{T_j}$ for some $i \neq j$.

Here, we suppose that the populations are approximate normal and all populations have a common variance, we have our statistics with the following distribution:

$$F \sim F_{df_1=2, df_2=6},$$

and with $\alpha = 0.05$, we have:

$$F \approx 0.790253 < 5.1433 \approx F_{df_1=2, df_2=6, \alpha=0.05},$$

so we retain the null hypothesis. Hence, at the level of $\alpha = 0.05$, we $\boxed{\text{do not have sufficient evidence}}$ to conclude that there is a difference in average yield between species of trees.

(c) Again, we can construct a *Hypothesis test for simple block design ANOVA table*, which has hypotheses as:

- H_0: There is no difference in mean of response for different blocks (fields), *i.e.*, $\mu_{B_1} = \mu_{B_2} = \mu_{B_3} = \mu_{B_4}$;
- H_a: There is a difference in mean of response for different blocks (fields), *i.e.*, $\mu_{B_i} \neq \mu_{B_j}$ for some $i \neq j$.

Here, we suppose that the populations are approximate normal and all populations have a common variance, we have our statistics with the following distribution:

$$F \sim F_{df_1=3, df_2=6},$$

and suppose $\alpha = 0.05$, we have:

$$F \approx 1.518541 < 4.7571 \approx F_{df_1=3, df_2=6, \alpha=0.05},$$

so we retain the null hypothesis. Hence, at the level of $\alpha = 0.05$, we $\boxed{\text{do not have sufficient evidence}}$ to conclude that there is a difference in average yield between the different fields.

┘

Problem 11.3. (Whitefly). The whitefly, which causes defoliation of shrubs and trees and a reduction in salable crop yeilds, has emerged as a pest in southern California. In a study to determine factors that affect the life cycle of the whitefly, an experiment was conducted in which caged whiteflies were placed on two different types of plants (Cotton or Cucumber) at three different temperatures (70°F, 77°F, and 82°F). The observation of interest was the total number of eggs laid by the whiteflies under one of the six possible treatment combinations. The experiment was replicated the same number of times for each combination of treatments. The following ANOVA table was created based on the data collected in the experiment:

Source	df	SS	MS	F
Plant Type	1	1512.3	1512.3	12.293
Temperature	2	487.4667	243.733	1.981
Interaction	2	111.2	55.6	0.452
Error	24	2952.4	123.017	
Total	29	5063.367		

(a) What type of experimental design has been used?

(b) How many observations were collected from each combination of treatments?

(c) Does the data provide sufficient ($\alpha = 0.05$) evidence that there is a difference in the average number of eggs laid by Whiteflies between cotton and cucumber plants?

(d) Does the data provide sufficient evidence that there is a difference in the average number of eggs laid by Whiteflies between different temperatures?

(e) Does the data provide sufficient evidence that there is an interaction between the type of plant and temperature on the average number of eggs laid by Whiteflies?

Sol'n.

(a) The experimental design is $\boxed{\text{Two-Way Classification study}}$ with 2 × 3 factorial experiment.

(b) The total number of observations is:

$$df_{\text{total}} + 1 = 29 + 1 = \boxed{30}.$$

(c) Here, we can construct a *Hypothesis test for Two-Way classification design ANOVA table*, which has hypotheses as:

- H_0: There is no difference in mean of response for the factor of plant types, *i.e.*, $\mu_{A_1} = \mu_{A_2}$;
- H_a: There is a difference in mean of response for different plant types, *i.e.*, $\mu_{A_1} \neq \mu_{A_2}$.

Here, we suppose that the populations are approximate normal and all populations have a common variance, we have our statistics with the following distribution:

$$F \sim F_{df_1=1, df_2=24},$$

and with $\alpha = 0.05$, we have:

$$F = 12.293 \geq 4.2597 \approx F_{df_1=1, df_2=24, \alpha=0.05},$$

so we reject the null hypothesis. Hence, at the level of $\alpha = 0.05$, we $\boxed{\text{have sufficient evidence}}$ to conclude that there is a difference in average number of eggs laid by Whitefiles between cotton and cucumber plants.

(d) Again, we can construct a *Hypothesis test for Two-Way classification design ANOVA table*, which has hypotheses as:

- H_0: There is no difference in mean of response for the factor of temperatures, *i.e.*, $\mu_{B_1} = \mu_{B_2} = \mu_{B_3}$;
- H_a: There is a difference in mean of response for different temperatures, *i.e.*, $\mu_{B_i} \neq \mu_{B_j}$ for some $i \neq j$.

Here, we suppose that the populations are approximate normal and all populations have a common variance, we have our statistics with the following distribution:

$$F \sim F_{df_1=2, df_2=24},$$

and still suppose $\alpha = 0.05$, we have:

$$F = 1.982 < 3.4028 \approx F_{df_1=2, df_2=24, \alpha=0.05},$$

so we retain the null hypothesis. Hence, at the level of $\alpha = 0.05$, we $\boxed{\text{do not have sufficient evidence}}$ to conclude that there is a difference in average number of eggs laid by Whitefiles between different temperatures.

(e) Once again, we can construct a *Hypothesis test for Two-Way classification design ANOVA table*, which has hypotheses as:

- H_0: There is no interaction between the plant type and temperature, *i.e.*, $\gamma_{ij} = 0$ for all i, j;
- H_a: There is interaction between the plant type and temperature, *i.e.*, $\gamma_{ij} \neq 0$ for some i, j.

Here, we suppose that the populations are approximate normal and all populations have a common variance, we have our statistics with the following distribution:

$$F \sim F_{df_1=2, df_2=24},$$

and still suppose $\alpha = 0.05$, we have:

$$F = 0.452 < 3.4028 \approx F_{df_1=2, df_2=24, \alpha=0.05},$$

so we retain the null hypothesis. Hence, at the level of $\alpha = 0.05$, we $\boxed{\text{do not have sufficient evidence}}$ to conclude that there is an interaction between the type of plant and temperature on the average number of eggs laid by Whiteflies.

┘

Problem 11.4. (I Hate Mondays). In a study conducted by researchers at RCSI University of Medicine and Health Sciences in Ireland, data of 10,528 patients admitted to hospital between 2013 and 2018 with the most serious type of heart attack were analyzed. The researchers found that the odds of suffering a heart attack of this kind were about 13% higher on a Monday as compared with the other days of the week. Suppose you wish to replicate their study to verify the claim. You survey 200 working people who had recently had heart attacks and recorded the day on which their heart attack occurred:

Day:	Mon.	Tue.	Wed.	Thur.	Fri.	Sat.	Sun.
Count:	36	27	26	32	26	29	24

Does the data present sufficient ($\alpha = 0.05$) evidence to indicate that there is a difference in the incident of heart attacks depending on the day of the week?

Sol'n. Here, we perform a *Goodness-of-Fit test*. For the simplicity of notation, we enumerate the day of the week from Monday to Sunday as 1 to 7, so we propose the following hypotheses:

- H_0: There is no difference in the incident of heart attacks depending on the day of the week, *i.e.*, $p_1 = p_2 = \cdots = p_7 = 1/7$, which is $E_i = 200/7$ for all $i = 1, 2, \cdots, 7$;
- H_a: There are difference in the incident of hear attacks depending on the day of the week, *i.e.*, there is at least one i such that $p_i \neq 1/7$.

Then, we want to calculate the Person's test statistic as:

$$X^2 = \sum_{i=1}^{7} \frac{(O_i - 200/7)^2}{200/7} = 3.63.$$

Here, with $df = 7 - 1 = 6$, we have it corresponding to the χ^2-distribution as:

$$X^2 \sim \chi^2_{df=6}.$$

Hence, with $\alpha = 0.05$, our test statistics is:

$$X^2 = 3.63 < 12.592 = \chi_{df=6,\alpha=0.05},$$

so we retain the null hypothesis. Hence, at the level of $\alpha = 0.05$, we $\boxed{\text{do note have sufficient evidence}}$ to conclude that there is a difference in the incident of heart attacks depending on the day of the week. ⌟

Problem 11.5. (Anxious Babies). In a study conducted to determine the influence of infant-mother attachment pattern on the development of peer interaction of toddlers, 24 eighteen-month-old children were classified base on their attachment security (Secure, Avoidant, Ambivalent) according to the strange situation procedure. Once the children turned 2, they were each observed during a play session with a control group of children, and the total number of interactions initiated by the children were counted and classified as either Positive or Agonistic. Does the data present sufficient ($\alpha = 0.05$) evidence that the type of interactions a child initiates varies with their attachment level?

Attachment Level:	Secure	Avoidant	Ambivalent
Positive Initiations:	16	11	15
Agonistic Initiations:	4	4	6

Sol'n. Here, we perform a *Test of Independence for Contingency Tables*. For the simplicity of notation, we enumerate Secure, Avoidant, and Ambivalent as 1, 2, and 3, and the positive and antagonistic initiations as 1 and 2, so we propose the following hypotheses:

- H_0: The two classification methods are independent;
- H_a: The two classification methods are dependent.

Prior to testing the test statistics, we sum up each row and column:

Attachment Level:	Secure	Avoidant	Ambivalent	*Sum of Row*
Positive Initiations:	16	11	15	42
Agonistic Initiations:	4	4	6	14
Sum of Column	20	15	21	56

Then, we can calculate the respective E_{ij} as:

$$E_{1\,1} = \frac{r_1c_1}{n} = \frac{42 \times 20}{56} = 15, \qquad E_{1\,2} = \frac{r_1c_2}{n} = \frac{42 \times 15}{56} = 11.25, \qquad E_{1\,3} = \frac{r_1c_3}{n} = \frac{42 \times 21}{56} = 15.75,$$
$$E_{2\,1} = \frac{r_2c_1}{n} = \frac{14 \times 20}{56} = 5, \qquad E_{2\,2} = \frac{r_2c_2}{n} = \frac{14 \times 15}{56} = 3.75, \qquad E_{2\,3} = \frac{r_2c_3}{n} = \frac{14 \times 21}{56} = 5.25.$$

Eventually, we calculate the Pearson's test statistic:

$$\begin{aligned} X^2 &= \sum_{i=1}^{3}\sum_{j=1}^{2} \frac{(O_{ij} - E_{ij})^2}{E_{ij}} \\ &= \frac{(16-15)^2}{15} + \frac{(11-11.25)^2}{11.25} + \frac{(15-15.75)^2}{15.75} + \frac{(4-5)^2}{5} + \frac{(4-3.75)^2}{3.75} + \frac{(6-5.25)^2}{5.25} \\ &\approx 0.431746. \end{aligned}$$

Here, with $df = (3-1)(2-1) = 2$, we have it corresponding to the χ^2-distribution as:

$$X^2 \sim \chi^2_{df=1,\alpha=0.05}.$$

Hence, with $\alpha = 0.05$, our test statistics is:

$$X^2 = 0.431746 < 3.841 = \chi_{df=1,\alpha=0.05},$$

so we retain the null hypothesis. Hence, at the level of $\alpha = 0.05$, we $\boxed{\text{do note have sufficient evidence}}$ to conclude that two classification methods are dependent, *i.e.*, we do not have sufficient evidence to conclude that the type of interactions a child initiates varies with their attachment level. ⌟

Problem 11.6. (Tennis Racquets). A student is looking to purchase a new tennis racquet. Since racquets can have widely varying characteristics, she decides to regress the price (Y, in USD) onto the weight (X_1, in ounces) and stiffness (X_2 measured by a Babolat diagnostic machine at her local shop with a scale from 0 (most flexible) to 100 (most stiff)). Her findings are presented below:

Racquet	**Price**	**Weight**	**Stiffness**
Head Titanium Ti.S6	99	8	75
Wilson Triad XP5	150	10.3	46
Dunlop CS 10.0	149	9.49	76
ProKennex Heritage C98	219	12.1	57
Volkl V-Feel 9	150	10.93	74
Prince Phantom Pro 100	129	11.4	54
Wilson XP1	190	16	73
Head MicroGEL Radical OS	213	10.41	56
Tecnifibre Tflash 300 CES	199	10.6	72
Wilson Clash 100L	179	10.4	54

Calculating the sum of squares from her regression analysis she finds:

$$RSS = 9633.65, \qquad \text{and} \qquad \text{Total } SS = 13482.10.$$

(a) Calculate the R^2 value.

(b) Complete the ANOVA table.

(c) Based on the R^2 value in part (a), and the F-statistic in part (b), do you think the P-values of the coefficients $\hat{\beta_1}$ and $\hat{\beta_2}$ in the LS regression line will be significant or not?

(d) Suppose the LS line produced the following coefficients and standard errors:

	Coefficient	Standard Error
Weight (X_1)	$\hat{\beta_1} = 9.1153$	5.9406
Stiffness (X_2)	$\hat{\beta_2} = -0.7403$	1.0958

Find a 95% confidence interval for β_1 and β_2. Do your confidence intervals support your thoughts in part (c)? Explain.

Sol'n.

(a) The R^2 statistic is given by:

$$R^2 = \frac{\text{Total } SS - RSS}{\text{Total } SS} = \frac{13482.10 - 9633.65}{13482.10} \approx \boxed{0.285449}.$$

(b) Here, we construct the ANOVA table as:

Source	df	SS	MS	F
Regression	2	3 848.45	$\frac{3\,848.45}{2} = 1\,924.225$	$\frac{1\,924.225}{1\,376.235714} \approx 1.398180$
Error	$10 - 1 - 2 = 7$	9 633.65	$\frac{9\,633.65}{7} \approx 1\,376.235714$	
Total	$10 - 1 = 9$	13 482.10		

(c) Note that from previous parts, we have:

$$R^2 \approx 0.285449 \ \text{ and } \ F \approx 1.398180.$$

Notice that R^2 is quite small, so our model does not do a good job at predicting the price of the racquets.

On the other hand, we observe that the F-statistics satisfies that:

$$F = 1.398180 < 3.2574 = F_{df_1=2, df_2=7, \alpha=0.05},$$

so we conclude that the P-value is greater than 0.1, meaning that the coefficients will be not significant.

(d) In obtaining the confidence interval, we find the critical value of t with $df = 10 - 2 = 8$ and $\alpha/2 = 0.025$, so we have:

$$t_{df=8, \alpha/2=0.025} = 2.306.$$

Hence, for β_1, the confidence interval is:

$$CI_{95\%}(\beta_1) = [9.1153 - 2.306 \times 5.9406, 9.1153 + 2.306 \times 5.9406] \approx \boxed{[-4.583724, 22.814324]}.$$

Likewise, for β_2, the confidence interval is:

$$CI_{95\%}(\beta_2) = [-0.7403 - 2.306 \times 1.0958, -0.7403 + 2.306 \times 1.0958] \approx \boxed{[-3.267215, 1.786615]}.$$

Note that for the confidence interval for 95%, 0 is contained in both the interval for β_1 and β_2, so this aligns with part (c) that the coefficients are not significant.

┘

Problem 11.7. (Mr. Plow's Snow Plow Business; Revisited). Back in Homework 2 we looked at the average snowfall in inches in Springfield over a 10 year period:

Year	Snow Fall (in.)	Year	Snow Fall (in.)
1	61.5	6	58.2
2	62.3	7	57.5
3	60.7	8	57.5
4	59.8	9	56.1
5	58.0	10	56.0

In that problem we calculated the LS regression line:

$$\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x = 62.21 - 0.7x.$$

(a) How should you interpret the intercept and slope of the line $\hat{y}$?

(b) Calculate the residual standard error.

(c) How should you interpret the number you found in part (b)?

(d) Use $\alpha = 0.05$ to test $H_0 : \beta_1 = 0$. If you rejected H_0, then how do you explain why $\hat{\beta}_1 = -0.7$ is pretty close to zero?

Sol'n.

(a) Here the intercept is $\hat{\beta}_0 = 62.21$, which can be interpreted as the average snowfall in inches is 62.21 in year 0.
The slope is $\hat{\beta}_1 = -0.7$, which implies that the average snowfall in inches decreases by 0.7 each year.

(b) Here, we want to first calculate the prediction for each year in terms of snow fall in inches, which is:

Year	Snow Fall (in.)	*Prediction (in.)*	Year	Snow Fall (in.)	*Prediction (in.)*
1	61.5	61.51	6	58.2	58.01
2	62.3	60.81	7	57.5	57.31
3	60.7	60.11	8	57.5	56.61
4	59.8	59.41	9	56.1	55.91
5	58.0	58.71	10	56.0	55.21

Now, we calculate RSS by summing the square of the differences:

$$RSS = (61.5 - 61.51)^2 + \cdots + (56.0 - 55.21)^2 = \boxed{4.749}$$

(c) Note that $RSS = 4.749$ is pretty small, so $\boxed{\text{we have a good regression line}}$.

(d) Here, we form a hypothesis test on β_1, so we have the hypotheses as:

- H_0: There is no relationship between year and snow fall in inches, *i.e.*, $\beta_1 = 0$;
- H_a: There is some relationship between year and snow fall in inches, *i.e.*, $\beta_1 \neq 0$.

Since we do not know σ^2, we estimate it using the sample variance of the residuals, which is:

$$s_\varepsilon^2 = \frac{1}{10-2} \cdot RSS = 0.593625.$$

Then, we compute the standard error of $\hat{\beta_1}$, with $\sigma^2 \approx s_\varepsilon^2$, as:

$$SE(\hat{\beta_1}) = \sqrt{\frac{s_\varepsilon^2}{(10-1)s_x^2}} \approx \sqrt{\frac{0.593625}{9 \times 9.166667}} \approx 0.084826.$$

Now, we have the null distribution for the test statistic as:

$$t_8 = \frac{\hat{\beta_1}}{SE(\hat{\beta_1})} \sim \mathcal{T}_8,$$

in which the test statistic for the given sample is:

$$t_8 \approx \frac{-0.7}{0.007195} \approx -8.252184.$$

Note that for the critical value, we have:

$$t_{df=8,\alpha/2=0.025} = 2.306,$$

and note that:

$$t_8 \approx -8.252184 < -2.306 = -t_{df=8,\alpha/2=0.025},$$

we reject H_0.

Hence, we have sufficient evidence to conclude that there is some relationship between year and snow fall in inches.

Then, consider that $\hat{\beta_1} = -0.7$ is close to zero, but given that its standard error $SE(\hat{\beta_1}) \approx 0.084826$, which is very small, it is reasonable to say that -0.7 is not very closed to zero compared to the standard error.

$\lrcorner$

Problem 11.8. (Easier Than It Looks). You've collected the data $(x_1, y_1), \cdots, (x_n, y_n)$ for a simple linear regression model. Suppose you find both $\overline{y} = 0$ and $\overline{x} = 0$ (or at least very nearly 0). In this case the formula for the LS line reduces to:

$$\hat{y} = \hat{\beta_1} x.$$

(understand why before continuing). Show that for each observation i, the predicted value $\hat{y_i}$ is a linear combination of the observed response values $y_1, \cdots, y_n$, *i.e.*, you need to find constants $c_1, c_2, \cdots, c_n$ for which $\hat{y_i} = c_1 y_1 + c_2 y_2 + \cdots + c_n y_n$.

Proof. The conclusion of this problem follows immediately from the following cases:

- First, suppose that $y_i = 0$ for some $i \in \{1, 2, \cdots, n\}$, then we have:

  $$\hat{y_k} = \frac{\hat{y_k}}{y_i} y_i,$$

  where $\frac{\hat{y_k}}{y_i}$ is a the constant C_i with all other coefficients being 0, so this case holds trivially.

- Otherwise, suppose that $y_1 = y_2 = \cdots = y_n = 0$, then we have $\hat{y_k} = 0$, so it is trivially the linear combination with all entries being 0.

Hence, any $\hat{y_k}$ can be immediately written as a linear combination of y_i's. □

Of course, the above proof does not explain anything as any non-zero element in this field "generates" every element in this field. However, diligent readers should wonder *if there is a more "elegant" way to find the coefficients*, say using only the given numbers, which leads to the following proposition.

Proposition. In particular, we can express each c_i as the basic operators, *i.e.*, $+$, $-$, $\times$, and $\div$, of x_i's for $i \in \{1, 2, \cdots, n\}$.

A more elegant proof to the Proposition. Suppose that we have $\hat{y} = \hat{\beta_1} x$, for each $\hat{y_k}$ with $k \in \{1, 2, \cdots, n\}$, we can express it as:

$$\begin{aligned} \hat{y_k} &= \hat{\beta_1} x_k \\ &= \frac{\sum_{i=1}^{n} (y_i - \overline{y})(x_i - \overline{x})}{\sum_{i=1}^{n} (x_i - \overline{x})^2} \cdot x_k \\ &= \frac{\sum_{i=1}^{n} y_i x_i}{\sum_{i=1}^{n} x_i^2} \cdot x_k \\ &= \sum_{j=1}^{n} \frac{x_k}{\sum_{j=1}^{n} x_i^2} \cdot y_j \end{aligned}$$

Thus, we can define each c_j as the coefficient for y_j as:

$$c_j = \frac{x_k}{\sum_{j=1}^{n} x_i^2},$$

as desired. □

Problem 11.9. (Birds in Trees). A student is interested in the number of blue jays (λ) nesting in the average oak tree in Wyman park. The student suspects that the number of birds in a randomly selected tree will depend on the age (X, in years) of the tree, since an older tree will likely be taller and offer more resources. Thus, they initially believe the true population regression line to be:

$$\lambda = \beta_0 + \beta_1 X + \varepsilon.$$

The student surveys 15 randomly selected oak trees and records the number of blue jays nesting in them:

(x_i, λ_i) = (age in years, number of blue jays)				
(24, 4)	(28, 1)	(35, 3)	(52, 5)	(52, 6)
(53, 6)	(64, 9)	(75, 10)	(79, 13)	(82, 12)
(95, 17)	(97, 17)	(99, 20)	(105, 23)	(108, 26)

(a) For each row in the above table, calculate the mean and variance of the λ entries. That is, find the mean and variance of the number of birds nesting in the 5 youngest trees, the 5 middle oldest trees, and the 5 oldest trees.

(b) Find the least squares regression line for the data given that the variance of age is 799.4095 and the covariance between age and number of birds is 207.5667, then use this to draw a residual plot.

(c) You'll note that your plot in part (b) appears strange, which may lead us to suspect the true linear relationship between λ and X is not linear.

In fact, since we are counting the number of birds within a particular space (in our case, an oak tree) it might be reasonable to suspect that that λ given the age of the tree may be modeled by a Poisson random variable. This assumption is further supported by your answer in part (a), since the mean and variance of a Poisson random variable are equal, we would expect to see the variance of the number birds grow with the mean.

With this in mind, let's update our assumption on the population line:

$$\log(\lambda) = \beta_0 + \beta_1 X + \varepsilon.$$

Calculate the least squares regression line using the data $\big(x_i, \log(\lambda_i)\big)$, then use this to graph a residual plot. You should notice a much more natural looking residual plot this time.

Sol'n.

(a) As instructed, we calculate the mean and variances:

(x_i, λ_i) = (age in years, number of blue jays)					*mean # of birds*	*variance of # of birds*
(24, 4)	(28, 1)	(35, 3)	(52, 5)	(52, 6)	3.8	3.7
(53, 6)	(64, 9)	(75, 10)	(79, 13)	(82, 12)	10	7.5
(95, 17)	(97, 17)	(99, 20)	(105, 23)	(108, 26)	20.6	15.3

(b) Here, note that we have the slope of the linear regression as:

$$\hat{\beta_1} = \frac{s_{xy}}{s_x \cdot s_y} \cdot \frac{s_y}{s_x} = \frac{s_{xy}}{s_x^2} = \frac{207.5667}{799.4095} \approx 0.259650.$$

Note that the averages are:

$$\overline{x} \approx 69.866667, \quad \text{and} \quad \overline{y} \approx 11.466667.$$

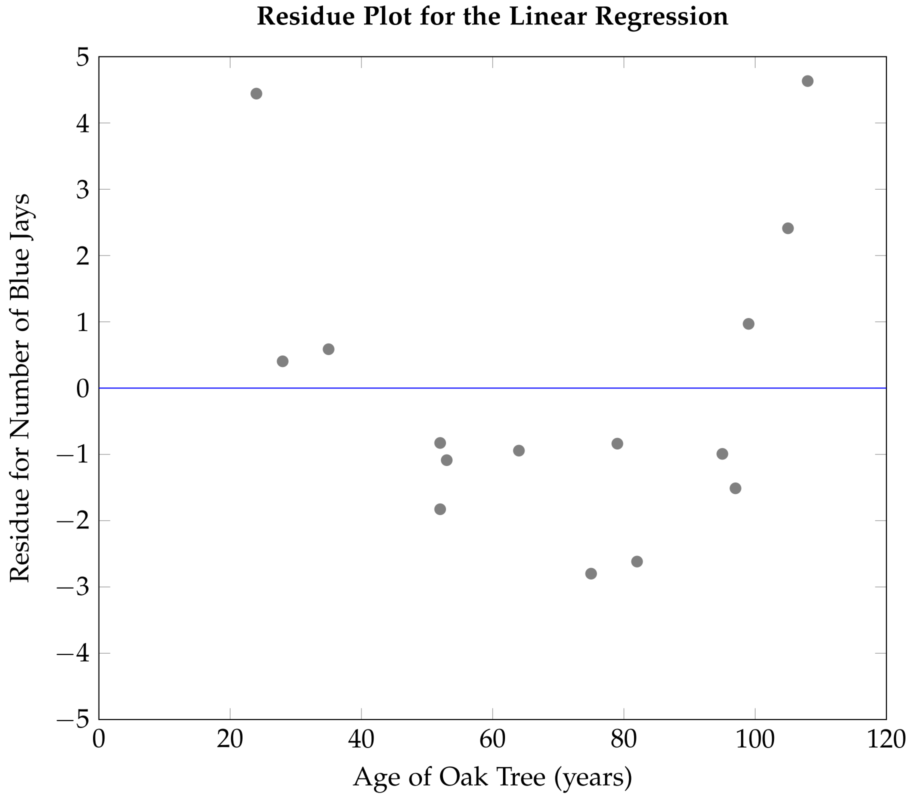
Then, we know that:

$$\hat{\beta_0} = \overline{y} - \hat{\beta_1} \cdot \overline{x} \approx 11.466667 - 0.259650 \times 69.866667 \approx -6.674213.$$

Thus, the line of the best fit is:

$$\hat{y} = \boxed{0.259650x - 6.674213}.$$

(c) The residue plot is as follows:



(d) Then, we construct the data with the natural log:

$(x_i, \log(\lambda_i))$				
(24, 1.386294)	(28, 0)	(35, 1.098612)	(52, 1.609438)	(52, 1.791759)
(53, 1.791759)	(64, 2.197225)	(75, 2.302585)	(79, 2.564949)	(82, 2.484907)
(95, 2.833213)	(97, 2.833213)	(99, 2.995732)	(105, 3.135494)	(108, 3.258097)

Here, one can calculate the covariance as:

$$s_{x\ (\log(y))} \approx 23.527736,$$

and with the same variance on x, we have:

$$\hat{\beta_1} = \frac{s_{x\ (\log(y))}}{s_x^2} \approx \frac{23.527736}{799.4095} \approx 0.029431.$$

By calculation, we observe that the average of $\log(\lambda_i)$ is:

$$\overline{\log(\lambda_i)} \approx 2.152219.$$

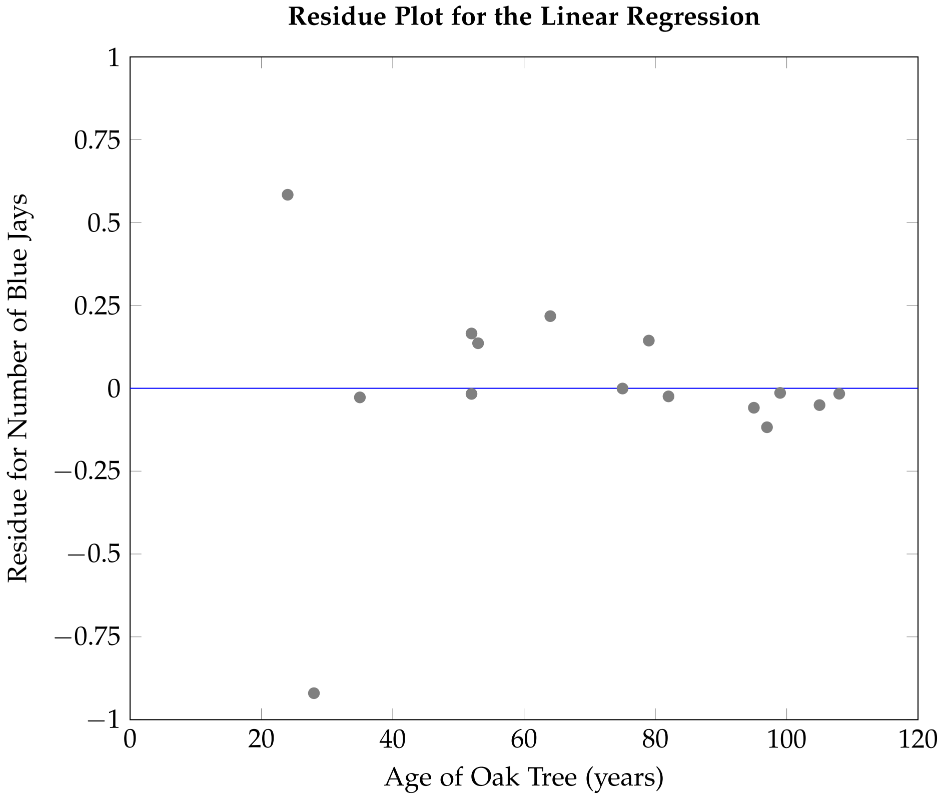
Hence, this leading us to the constant as:

$$\hat{\beta_0} = \overline{\log(\lambda_i)} - \hat{\beta_1} \cdot \overline{x} \approx 2.152219 - 0.029431 \times 69.866667 \approx 0.095973.$$

Now, the least square regression line is:

$$\hat{y} = \boxed{0.029431x + 0.095973}.$$

Now, the residue plot lies as follows:



One can notice that the residue plot is more "natural", *i.e.*, we the residue is smaller and there is not an obvious pattern here.

⌟