Xiao Zhehao, A0235422N

1. a) By LN3, $\hat{p}_{mle} = \frac{\sum X_i}{n} = \frac{159}{314} = 0.5064$. Since $314 \gg 30$, we can approximate $p \sim N(p, \frac{p(1-p)}{n})$ by CLT. The 90% CI is $\hat{p} \pm z_{0.05}\sqrt{\frac{p(1-p)}{n}}$. Since n is large, we can approx p by $\hat{p}_{mle}$ so the CI is $0.5064 \pm 1.645\sqrt{\frac{0.5064(0.4936)}{314}} = 0.5064 \pm 0.0464$ or $[0.4600, 0.5528]$.

b) To get $0.02 = 2 z_{0.05}\sqrt{\frac{p(1-p)}{n}}$ we need $0.01^2 = 1.645^2 \frac{0.5064(0.4936)}{n}$, $n = 1.645^2 \frac{0.5064(0.4936)}{0.01^2} \approx 6764$.

c) Since the CI is $[0.390, 0.500]$, $\hat{p} = 0.390 + \overbrace{\tfrac{1}{2}(0.5 - 0.39)}^{0.055} = 0.445$. Hence, $0.055 = z_{\alpha/2}\sqrt{\frac{0.445(0.555)}{314}}$ $z_{\alpha/2} = 0.055\sqrt{\frac{314}{0.445(0.555)}} = 1.961 \approx z_{0.025}$ so it is a $100(1-2\cdot 0.025)\% = 95\%$ CI.

2. a) Let $\bar{X}$ and $\bar{Y}$ be the sample means and our estimates for μ_X and μ_Y respectively. Hence, $E(\bar{X}) = \mu_X$, $Var(\bar{X}) = \frac{\sigma_X^2}{n}$, $E(\bar{Y}) = \mu_Y$, & $Var(\bar{Y}) = \frac{\sigma_Y^2}{m}$.
This gives $E(\bar{X}-\bar{Y}) = E(\bar{X}) - E(\bar{Y}) = \mu_X - \mu_Y$ and $Var(\bar{X}-\bar{Y}) = Var(\bar{X}) + Var(\bar{Y}) = \frac{\sigma_X^2}{n} + \frac{\sigma_Y^2}{m}$ as the 2 samples are independent. Hence, $\bar{X}-\bar{Y} \sim N(\mu_X - \mu_Y, \frac{\sigma_X^2}{n} + \frac{\sigma_Y^2}{m})$
and $Z = [(\bar{X}-\bar{Y}) - (\mu_X - \mu_Y)] / \sqrt{\frac{\sigma_X^2}{n} + \frac{\sigma_Y^2}{m}} \sim N(0,1)$. By LN5, 3.1, the $100(1-\alpha)\%$ CI is $[\bar{X}-\bar{Y} - z_{\alpha/2}\sqrt{\frac{\sigma_X^2}{n} + \frac{\sigma_Y^2}{m}}, \bar{X}-\bar{Y} + z_{\alpha/2}\sqrt{\frac{\sigma_X^2}{n} + \frac{\sigma_Y^2}{m}}]$.

b) We rewrite $m = 6000 - n$. We want $\min_n z_{\alpha/2}\sqrt{\frac{\sigma_X^2}{n} + \frac{\sigma_Y^2}{6000-n}} = \min_n 1.645\sqrt{\frac{70^2}{n} + \frac{50^2}{6000-n}}$. Hence, we solve $\frac{d}{dn}\frac{70^2}{n} + \frac{50^2}{6000-n} = 0$ as we assume the function is concave w.r.t. n and sqrt is monotonically increasing.
$\frac{d}{dn}\frac{70^2}{n} + \frac{50^2}{6000-n} = -\frac{70^2}{n^2} + \frac{50^2}{(6000-n)^2} = 0$, $\frac{50}{6000-n} = \frac{70}{n}$, $50n = 420000 - 70n$, $120n = 420000$, $n = 3500$//

3. a) Since $x_i \sim N(\mu, 121)$, $i \in [1,12]$ where x_i is the no. of chocolate chips in the i^{th} cookie, the 90% CI is $\mu \pm z_{0.05}\frac{11}{\sqrt{12}} = 41.83 \pm 1.645\frac{11}{\sqrt{12}}$ or $[36.61, 47.05]$ (*approx μ by $\bar{x}$)

b) The 95% and 99% CI are $\mu \pm z_{0.025}\frac{11}{\sqrt{12}} = 41.83 \pm 1.960\frac{11}{\sqrt{12}}$ or $[35.61, 48.05]$ and $\mu \pm z_{0.005}\frac{11}{\sqrt{12}} = 41.83 \pm 1.645\frac{11}{\sqrt{12}}$ or $[33.65, 50.01]$ respectively.

c) If σ^2 is unknown, we estimate it by $s^2 = 11.8^2$. By LN5, 3.3, the 90% CI is $\mu \pm t_{0.05}(11)\frac{s}{\sqrt{n}} = 41.83 \pm 1.796\frac{11.8}{\sqrt{12}}$ or $[35.71, 47.95]$

4. a) Since $\frac{X_i - \mu}{\sigma} \sim N(0,1)$, $\left(\frac{X_i-\mu}{\sigma}\right)^2 \sim \chi^2(1)$ and $\sum^n \left(\frac{X_i-\mu}{\sigma}\right)^2 = \frac{\sum(X_i-\mu)^2}{\sigma^2} \sim \chi^2(n)$ by definition of N and χ^2 distributions. Hence, $P\left(\frac{\sum(X_i-\mu)^2}{\sigma^2} > \chi^2_{\alpha/2}(n)\right) = \alpha/2$ and $P\left(\frac{\sum(X_i-\mu)^2}{\sigma^2} > \chi^2_{1-\alpha/2}(n)\right) = 1-\alpha/2$.
So $P\left(\chi^2_{1-\alpha/2}(n) \le \frac{\sum(X_i-\mu)^2}{\sigma^2} \le \chi^2_{\alpha/2}(n)\right) = 1-\alpha = P\left(\frac{1}{\chi^2_{\alpha/2}(n)} \le \frac{\sigma^2}{\sum(X_i-\mu)^2} \le \frac{1}{\chi^2_{1-\alpha/2}(n)}\right) = P\left(\frac{\sum(X_i-\mu)^2}{\chi^2_{\alpha/2}(n)} \le \sigma^2 \le \frac{\sum(X_i-\mu)^2}{\chi^2_{1-\alpha/2}(n)}\right)$. Hence, the $100(1-\alpha)\%$ CI for σ^2 is $\left[\frac{\sum(X_i-\mu)^2}{\chi^2_{\alpha/2}(n)}, \frac{\sum(X_i-\mu)^2}{\chi^2_{1-\alpha/2}(n)}\right]$.

b) We know $S^2 = \frac{1}{n-1}\sum(X_i-\bar{X})^2 \sim \chi^2(n-1)$ by definition of N distribution. Hence, $\frac{(n-1)S^2}{\sigma^2} = \frac{\sum(X_i-\bar{X})^2}{\sigma^2} \sim \chi^2(n-1)$. By replacing $\chi^2(n)$ with $\chi^2(n-1)$ and $\sum(X_i-\mu)^2$ with $\sum(X_i-\bar{X})^2$, we see that the $100(1-\alpha)\%$ CI for σ^2 is $\left[\frac{\sum(X_i-\bar{X})^2}{\chi^2_{\alpha/2}(n-1)}, \frac{\sum(X_i-\bar{X})^2}{\chi^2_{1-\alpha/2}(n-1)}\right]$.

c) To get the CI for σ we sqrt the CI of σ^2 $\sqrt{\frac{\sum(X_i-\bar{X})^2}{\chi^2_{\alpha/2}(n-1)}} = \sqrt{\frac{(n-1)S^2}{\chi^2_{\alpha/2}(n-1)}} = \sqrt{\frac{(n-1)}{\chi^2_{\alpha/2}(n-1)}}\,S$ so the CI is $\left[\sqrt{\frac{(n-1)}{\chi^2_{\alpha/2}(n-1)}}\,S, \sqrt{\frac{(n-1)}{\chi^2_{1-\alpha/2}(n-1)}}\,S\right]$.

5. a) $\bar{x} = 3055.91$, $\bar{y} = 3317.91$

b) Pooled t-Interval: $[-323.65, -200.34]$

c) Welch's t-Interval: $[-323.59, -200.40]$

ST2132 Mathematical Statistics
Assignment 2
Due date: 21 Feb 2025 23:59 on Canvas
Please submit in pdf format.

Please answer all questions. Please work on the assignment by yourself only. We follow a strict rule on academic honesty.

Generative AI policy: The use of generative AI tools (e.g., ChatGPT, Gemini, etc.) is strictly prohibited for structural or conceptual questions in this course. However, these tools may be used to assist with coding tasks, provided their use is clearly documented.

1. Suppose you flip a coin 314 times, and heads appears 159 times.
   (a) Construct an approximate 90% confidence interval for the probability that the coin comes up heads.
   (b) Approximately how many samples would you need to obtain an approximate 90% confidence interval with width 0.02, while keeping exactly 50.64% of flips appearing heads?
   (c) You give the coin to a friend, who also flips the coin 314 times, and obtains an approximate confidence interval [0.390, 0.500] instead. What confidence level did (s)he use?

2. Let $X_1, \ldots, X_n \overset{i.i.d.}{\sim} N(\mu_X, \sigma_X^2)$, $Y_1, \ldots, Y_m \overset{i.i.d.}{\sim} N(\mu_Y, \sigma_Y^2)$ and they are independent. Suppose that both σ_X^2 and σ_Y^2 are known.
   (a) Construct a two-sided $100(1-\alpha)\%$ confidence interval for the difference $\mu_X - \mu_Y$.
   (b) We want to obtain a 90% confidence interval for the difference between true average cable strengths made by Company X and by Company Y. Suppose cable strength is normally distributed for both types of cables with $\sigma_X = 70$ and $\sigma_Y = 50$. If we can make $n + m = 6000$ observations, how many of these should be on Company X cable if we want to minimize the width of the interval?

3. Michael owns a bakery. The number of chocolate chips that he adds to his cookies is normally distributed with some mean μ and known variance $\sigma^2 = 121$. A customer buys a dozen of these cookies, and obtains the sample

   $$31,\ 41,\ 59,\ 26,\ 53,\ 59, 47,\ 43,\ 23,\ 34,\ 42,\ 44$$

   for which the sample mean is $\bar{x} = 41.83$ and the sample variance is $s^2 \approx 11.8^2$.
   (a) Compute a 90% confidence interval for μ.
   (b) Compute 95% and 99% confidence intervals for μ.
   (c) Repeat part (a), assuming that the customer has no idea what σ^2 is.

4. Let $X_1, \ldots, X_n \overset{i.i.d.}{\sim} N(\mu, \sigma^2)$. In this question, we will construct confidence interval for σ^2. Let $C \sim \chi^2(r)$, and as usual for any $\alpha \in [0, 1]$ we denote $\chi^2_\alpha(r)$ to be

   $$P(C > \chi^2_\alpha(r)) = \alpha.$$

   (a) Suppose that μ is known. By considering the distribution of

   $$\sum_{i=1}^{n} \left(\frac{X_i - \mu}{\sigma} \right)^2,$$

prove that

$$\left[\frac{\sum_{i=1}^{n}(X_i-\mu)^2}{\chi^2_{\alpha/2}(n)}, \frac{\sum_{i=1}^{n}(X_i-\mu)^2}{\chi^2_{1-\alpha/2}(n)}\right]$$

is a $100(1-\alpha)\%$ confidence interval for σ^2.

(b) Suppose that μ is unknown. By considering the distribution of

$$\sum_{i=1}^{n}\left(\frac{X_i-\overline{X}}{\sigma}\right)^2 = \frac{(n-1)S^2}{\sigma^2},$$

prove that

$$\left[\frac{\sum_{i=1}^{n}(X_i-\overline{X})^2}{\chi^2_{\alpha/2}(n-1)}, \frac{\sum_{i=1}^{n}(X_i-\overline{X})^2}{\chi^2_{1-\alpha/2}(n-1)}\right]$$

is a $100(1-\alpha)\%$ confidence interval for σ^2.

(c) In the same setting as part (b), that is, suppose that μ is unknown. Construct a $100(1-\alpha)\%$ confidence interval for σ.

5. (Traffic volume) Michael lives in Minneapolis. He is interested in the traffic volume of Minneapolis. Download the dataset "traffic.csv" on Canvas. There are 9 columns and 48,204 data points. We will only use two columns for this question: "weather_main" and "traffic_volume". "weather_main" is a short description of the weather, and "traffic_volume" is the hourly traffic volume. Similar to Assignment 1, use your favourite computing language and attach your code at the end of your answer.

Note: the dataset in this question is a real-world dataset from UC Irvine Machine Learning depository. The link is here: CLICK.

(a) Michael is interested in doing a comparison between the hourly traffic volume when the weather is "Clear" verus the hourly traffic volume when the weather is "Rain". He has the feeling that there is less traffic when the weather is "Rain", and would like to derive some statistical insights from the data to support his idea. What is the sample mean of the traffic volume when the weather is "Clear", say $\overline{x}$? What is the sample mean of the traffic volume when the weather is "Rain", say $\overline{y}$?

(b) Assume the hourly traffic volume when the weather is "Rain" are $Y_1, \ldots, Y_m \overset{i.i.d.}{\sim} N(\mu_Y, \sigma^2)$, and the hourly traffic volume when the weather is "Clear" are $X_1, \ldots, X_n \overset{i.i.d.}{\sim} N(\mu_X, \sigma^2)$, and X_i and Y_i are independent. These assumptions are not realistic but still provide us a model to work on. Compute the 95% two-sample pooled t-interval for the difference $\mu_X - \mu_Y$. As n and m are large in this dataset, please use $z_{0.025}$ to replace $t_{0.025}(n+m-2)$ when you compute the interval.

(c) Assume the hourly traffic volume when the weather is "Rain" are $Y_1, \ldots, Y_m \overset{i.i.d.}{\sim} N(\mu_Y, \sigma_Y^2)$, and the hourly traffic volume when the weather is "Clear" are $X_1, \ldots, X_n \overset{i.i.d.}{\sim} N(\mu_X, \sigma_X^2)$, and X_i and Y_i are independent and $\sigma_X^2 \neq \sigma_Y^2$. Compute the 95% Welch's t-interval for the difference $\mu_X - \mu_Y$. As n and m are large in this dataset, please use $z_{0.025}$ to replace the t distribution quantiles when you compute the interval.