Math 224 **Spring 2022**
Homework 6 **Drew Armstrong**

---

1. Rounding Error. Your friend has a list of ten real numbers, whose values are unknown to you: $x_1, \ldots, x_{10} \in \mathbb{R}$. Your friend rounds each number to the nearest integer and sends you the results: $X_1, X_2, \ldots, X_{10} \in \mathbb{Z}$. We will assume that $X_i = x_i + U_i$, where each U_i is a uniform random variable on the interval $[-1/2, 1/2]$.

(a) Compute $E[U_i]$ and $\text{Var}(U_i)$.
(b) Consider the sum of the rounded numbers $X = X_1 + \cdots + X_{10}$ and the sum of the unrounded numbers $x = x_1 + \cdots + x_{10}$. Prove that $E[X] = x$.
(c) Assuming that the random variables U_i are independent, use the CLT to estimate the probability that $|X - x| > 1/2$. [Hint: The CLT says that $X - x = U_1 + \cdots + U_{10}$ is approximately normal. You just need to compute $E[X - x]$ and $\text{Var}(X - x)$.]

(a): The density of U_i is given by

$$f(x) = \begin{cases} 1 & -1/2 < x < 1/2, \\ 0 & \text{otherwise.} \end{cases}$$

Hence

$$E[U_i] = \int_{-1/2}^{1/2} x \cdot 1 \, dx = 0 \quad \text{and} \quad \text{Var}(U_i) = E[U_i^2] - 0^2 = \int_{-1/2}^{1/2} x^2 \cdot 1 \, dx = 1/12.$$

(b): Since $X_i = x_i + U_i$ and since x_i is constant we have

$$E[X_i] = E[x_i + U_i] = x_i + E[U_i] = x_i + 0 = x_i.$$

Then adding these up gives

$$\begin{aligned} E[X] &= E[X_1 + X_2 + \cdots + X_{10}] \\ &= E[X_1] + E[X_2] + \cdots + E[X_{10}] \\ &= x_1 + x_2 + \cdots + x_{10} \\ &= x. \end{aligned}$$

(c): Note that $X - x = U_1 + U_2 + \cdots + U_{10}$ is a sum of independent and identically distributed random variables. Presumably 10 is a large enough number that this sum is approximately normal. The mean is[1]

$$E[X - x] = E[U_1] + E[U_2] + \cdots + E[U_{10}] = 0 + 0 + \cdots + 0 = 0,$$

and the variance is

$$\begin{aligned} \text{Var}(X - x) &= \text{Var}(U_1) + \text{Var}(U_2) + \cdots + \text{Var}(U_{10}) \\ &= 1/12 + 1/12 + \cdots + 1/12 \\ &= 10/12. \end{aligned}$$

---

[1] We can also use part (b) to get $E[X - x] = E[X] - x = x - x = 0$.

It follows that $X - x$ is approximately $N(0, 10/12)$ and hence $Z = (X - x)/\sqrt{10/12}$ is approximately $N(0, 1)$. Therefore we have

$$\begin{aligned}
P(|X - x| > 1/2) &= P(X - x < -1/2) + P(X - x > 1/2) \\
&= P\left(\frac{X - x}{\sqrt{10/12}} < \frac{-1/2}{\sqrt{10/12}}\right) + P\left(\frac{X - x}{\sqrt{10/12}} > \frac{1/2}{\sqrt{10/12}}\right) \\
&\approx P(Z < -0.55) + P(Z > 0.55) \\
&= \Phi(-0.55) + (1 - \Phi(0.55)) \\
&= (1 - \Phi(0.55)) + (1 - \Phi(0.55)) \\
&= 2\,(1 - \Phi(0.55)) \\
&= 2\,(1 - 0.7088) \\
&= 58.24\%.
\end{aligned}$$

My computer gives the exact answer 58.904%, so this approximation is pretty good.

Remark: Note that the real number x rounds to the integer X if and only if $|X - x| \le 1/2$. Therefore $P(|X - x| > 1/2)$ is the probability that x does not round to X. In other words, there is a 58.9% chance that the following two procedures yield different results:

- Round each of 10 numbers to the nearest integer and then add them.
- Add 10 numbers and then round the sum to the nearest integer.

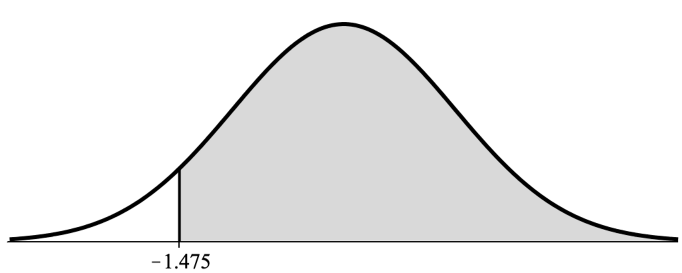
2. Tail Probabilities. Consider a standard normal variable $Z \sim N(0, 1)$. Solve for a.

(a) $P(Z > a) = 93\%$
(b) $P(Z < a) = 35\%$
(c) $P(|Z| > a) = 2\%$
(d) $P(|Z| < a) = 80\%$

(a): We have $\Phi(a) = 7\%$. Since $a < 0$ this is not in our table, so we use symmetry to write

$$\begin{aligned}
\Phi(-a) &= 1 - \Phi(a) \\
\Phi(-a) &= 93\% \\
-a &\approx 1.475 \\
a &\approx -1.475.
\end{aligned}$$

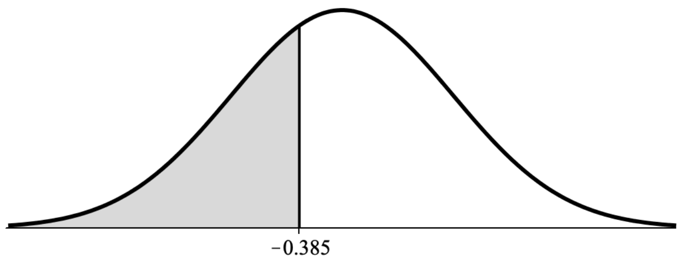
Here is a picture:



(b): We have $\Phi(a) = 35\%$. Since $a < 0$ this is not in our table, so we use symmetry:

$$\begin{aligned} \Phi(-a) &= 1 - \Phi(a) \\ \Phi(-a) &= 65\% \\ -a &\approx 0.385 \\ a &\approx -0.385. \end{aligned}$$

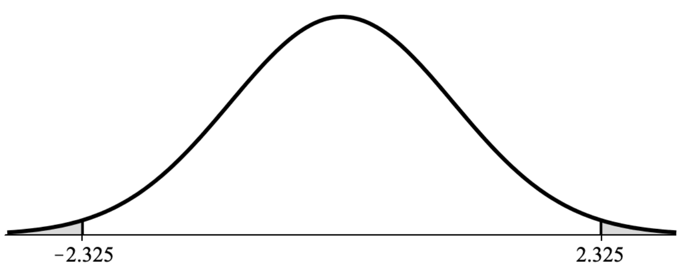
Here is a picture:



(c): We can rewrite $P(|Z| > a) = 2\%$ as

$$\begin{aligned} P(Z < -a) + P(Z > a) &= 2\% \\ \Phi(-a) + (1 - \Phi(a)) &= 2\% \\ (1 - \Phi(a)) + (1 - \Phi(a)) &= 2\% \\ 2\,(1 - \Phi(a)) &= 2\% \\ \Phi(a) &= 99\% \\ a &\approx 2.325. \end{aligned}$$

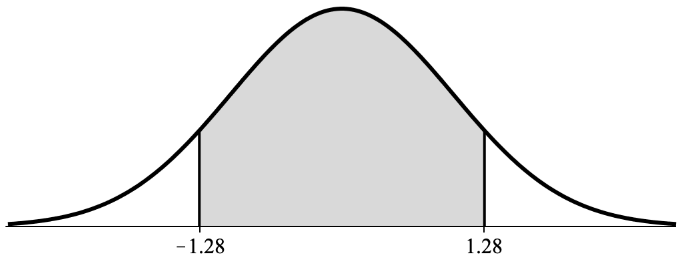
Here is a picture:



(d): We can rewrite $P(|Z| < a) = 80\%$ as

$$\begin{aligned} P(-a < Z < a) &= 80\% \\ \Phi(a) - \Phi(-a) &= 80\% \\ \Phi(a) - (1 - \Phi(a)) &= 80\% \\ 2\Phi(a) - 1 &= 80\% \\ \Phi(a) &= 90\% \\ a &\approx 1.28. \end{aligned}$$

Here is a picture:



Remark: I wrote out the solutions using algebra, but it's easier to work from a picture.

3. A Bernoulli Hypothesis Test. A six-sided die has sides labeled $\{1, 2, 3, 4, 5, 6\}$. Let p be the probability of getting a 6. Before performing any experiments we will assume that $H_0 =$ "$p = 1/6$" is true. Now suppose that you roll the die 600 times and let Y be the number of times you get 6. Which values of Y would cause you to reject H_0 in favor of $H_1 =$ "$p > 1/6$" at the 99% level of confidence? That is, what is the rejection region?

We can develop the answer from scratch or we can just quote a formula.

Quote a Formula: Consider the sample proportion $\hat{p} = Y/n$. When testing $H_0 =$ "$p = p_0$" against $H_1 =$ "$p > p_0$" at confidence level $1 - \alpha$, the rejection region is

$$\hat{p} > p_0 + z_\alpha \cdot \sqrt{\frac{p_0(1-p_0)}{n}}$$

In our case we have $n = 600$, $p_0 = 1/6$ and $1 - \alpha = 99\%$ so the rejection region is

$$\begin{aligned}
\hat{p} &> \frac{1}{6} + z_{1\%} \cdot \sqrt{\frac{(1/6)(5/6)}{600}} \\
\frac{Y}{600} &> \frac{1}{6} > (2.325) \cdot \sqrt{\frac{(1/6)(5/6)}{600}} \\
\frac{Y}{600} &> 0.202 \\
Y &> 121.2.
\end{aligned}$$

Thus we will reject $H_0 =$ "$p = 1/6$" for $H_1 =$ "$p > 1/6$" when $Y \geq 122$. In other words, if we roll a die 600 times and get a 6 at least 122 times then we will declare with 99% confidence that $P(6) > 1/6$.

Develop the Formula From Scratch: Suppose that $H_0 =$ "$p = p_0$" is true. Then from the CLT we know that $\hat{p}$ is approximately $N(p_0, p_0(1-p_0)/n)$. Hence

$$\frac{\hat{p} - p_0}{\sqrt{p_0(1-p_0)/n}} \approx N(0, 1).$$

To test $H_0 =$ "$p = p_0$" against $H_1 =$ "$p > p_0$" we look for values of $\hat{p}$ that are significantly larger than p_0:

$$\begin{aligned}\alpha &= P(Z > z_\alpha) \\ &\approx P\left(\frac{\hat{p} - p_0}{\sqrt{p_0(1-p_0)/n}} > z_\alpha\right) \\ &= P\left(\hat{p} > p_0 + z_\alpha \cdot \sqrt{\frac{p_0(1-p_0)}{n}}\right).\end{aligned}$$

If such a significantly large value of $\hat{p}$ occurs then we reject H_0 in favor of H_1.

4. Confidence Intervals for a Proportion. Let p be the proportion of Americans who are left-handed. In order to estimate p, we randomly selected $n = 1000$ Americans and we found that $Y = 125$ of them are left-handed. Use this information to compute two-sided, symmetric $(1-\alpha)100\%$ confidence intervals for p when $\alpha = 5\%$, 2.5% and 1%.

We can develop the answer from scratch or just quote a formula.

Quote a Formula: The sample proportion is $\hat{p} = Y/n = 125/1000 = 12.5\%$. We have the following approximate, two-sided, symmetric $(1-\alpha)100\%$ confidence interval for p:

$$\begin{aligned}p &= \hat{p} \pm z_{\alpha/2} \cdot \sqrt{\frac{\hat{p}(1-\hat{p})}{n}} \\ p &= 12.5\% \pm z_{\alpha/2} \cdot \sqrt{\frac{(0.125)(1-0.125)}{1000}} \\ p &= 12.5\% \pm z_{\alpha/2} \cdot 0.01045825033\end{aligned}$$

Substituting $\alpha = 5\%$, 2.5% and 1% gives the following confidence intervals, respectively:

$$\begin{aligned}p &= 12.5\% \pm 2.05\%, \\ p &= 12.5\% \pm 2.34\%, \\ p &= 12.5\% \pm 2.69\%.\end{aligned}$$

Develop the Formula From Scratch: We know from the CLT that $\hat{p}$ is approximately $N(p, p(1-p)/n)$, and hence

$$\frac{\hat{p} - p}{\sqrt{p(1-p)/n}} \approx N(0,1).$$

Then we use some algebra to obtain

$$\begin{aligned}1 - \alpha &= P(-z_{\alpha/2} < Z < z_{\alpha/2}) \\ &\approx P\left(-z_{\alpha/2} < \frac{\hat{p} - p}{\sqrt{p(1-p)/n}} < z_{\alpha/2}\right) \\ &\vdots \\ &= P\left(\hat{p} - z_{\alpha/2} \cdot \sqrt{\frac{p(1-p)}{n}} < p < \hat{p} - z_{\alpha/2} \cdot \sqrt{\frac{p(1-p)}{n}}\right)\end{aligned}$$

Since the error bounds involve the unknown p we replace it by $\hat{p}$ in the expression $\sqrt{p(1-p)/n}$. This is mathematically irresponsible but hopefully it's not too bad for large n. A computer would use a more accurate formula that is harder to derive and harder to memorize. If the sample is drawn from a population of size N then we get even more accuracy by using the variance of a hypergeometric distribution

$$\text{Var}(\hat{p}) = \frac{p(1-p)}{n} \cdot \frac{N-n}{N-1}.$$

In our case, $N = 329.5$ million and $n = 1000$, so $(N-n)/(N-1) = 0.9999969681$.

5. Sample Variance. Consider an iid sample $X_1, \ldots, X_n$ with unknown mean μ and unknown variance σ^2. In order to estimate σ^2 we define the *sample variance* as follows:

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

where $\overline{X} = (\sum_{i=1}^{n} X_i)/n$ is the usual sample mean.

(a) Show that $\sum_{i=1}^{n} \left(X_i - \overline{X}\right)^2 = \left(\sum_{i=1}^{n} X_i^2\right) - n\overline{X}^2$. [Hint: $n\overline{X} = \sum_{i=1}^{n} X_i$.]

(b) Show that $E[X_i^2] = \mu^2 + \sigma^2$ and $E[\overline{X}^2] = \mu^2 + \sigma^2/n$. [Hint: By definition we have $E[X_i] = \mu$ and $\text{Var}(X_i) = \sigma^2$, which implies that $E[\overline{X}] = \mu$ and $\text{Var}(\overline{X}) = \sigma^2/n$.]

(c) Combine (a) and (b) to show that $E[S^2] = \sigma^2$. This is why the definition of S^2 has $n-1$ in the denominator instead of n.

(a): First we note that

$$(X_i - \overline{X})^2 = X_i^2 - 2\overline{X}X_i + \overline{X}^2.$$

Then we sum over i to obtain

$$\begin{aligned}
\sum_{i=1}^{n}(X_i - \overline{X})^2 &= \sum_{i=1}^{n}(X_i^2 - 2\overline{X}X_i + \overline{X}^2) \\
&= \sum_{i=1}^{n} X_i^2 - 2\overline{X}\sum_{i=1}^{n} X_i + n\overline{X}^2 \\
&= \sum_{i=1}^{n} X_i^2 - 2n\overline{X}^2 + n\overline{X}^2 \\
&= \sum_{i=1}^{n} X_i^2 - n\overline{X}^2.
\end{aligned}$$

(b): For any random variable Y we have $\text{Var}(Y) = E[Y^2] - E[Y]^2$ and hence $E[Y^2] = E[Y]^2 + \text{Var}(Y)$. Since $E[X_i] = \mu$ and $\text{Var}(X_i) = \sigma^2$ we get

$$E[X_i^2] = E[X_i]^2 + \text{Var}(X_i) = \mu^2 + \sigma^2.$$

And since $E[\overline{X}] = \mu$ and $\text{Var}(\overline{X}) = \sigma^2/n$ we get

$$E[\overline{X}^2] = E[\overline{X}]^2 + \text{Var}(\overline{X}) = \mu^2 + \sigma^2/n.$$

(c): Combining (a) and (b) gives

$$
\begin{aligned}
E[S^2] &= E\left[\frac{1}{n-1}\sum_{i=1}(X_i - \overline{X})^2\right] \\
&= \frac{1}{n-1}\cdot E\left[\sum_{i=1}^{n}(X_i - \overline{X})^2\right] \\
&= \frac{1}{n-1}\cdot E\left[\sum_{i=1}^{n} X_i^2 - n\overline{X}^2\right] \\
&= \frac{1}{n-1}\cdot \left(\sum_{i=1}^{n} E[X_i^2] - n\cdot E[\overline{X}^2]\right) \\
&= \frac{1}{n-1}\cdot \left(n\cdot(\cancel{\mu^2}+\sigma^2) - n\cdot(\cancel{\mu^2}+\sigma^2/n)\right) \\
&= \frac{1}{n-1}\cdot \left(n\sigma^2 - \sigma^2\right) \\
&= \frac{1}{n-1}\cdot (n-1)\sigma^2 = \sigma^2.
\end{aligned}
$$

6. A Small Sample. The label weight of a Cadbury Creme Egg is 1.2oz. In order to test this you weighed 10 eggs and obtained the following values (in ounces):

1.12	1.01	1.04	1.10	1.00	1.04	1.28	1.17	1.19	1.24

Let X represent the underlying distribution with unknown mean $\mu = E[X]$. For simplicity we assume that X is normal.

(a) Compute the sample mean $\overline{X}$ and the sample variance S^2.
(b) Look up the t-tail probabilities $t_{5\%}(9)$ and $t_{2.5\%}(9)$.
(c) Test the hypothesis $H_0 =$ "$\mu = 1.2$" against the one-sided alternative $H_1 =$ "$\mu < 1.2$" at the 5% level of significance.
(d) Compute a two-sided symmetric 95% confidence interval for the unknown μ.

(a): My computer gives

$$\overline{X} = 1.119 \quad \text{and} \quad S^2 = 0.009677.$$

(b): The t-table says

$$t_{5\%}(9) = 1.833 \quad \text{and} \quad t_{2.5\%}(9) = 2.262.$$

(c): We just quote a formula. When testing $H_0 =$ "$\mu = \mu_0$" against $H_1 =$ "$\mu < \mu_0$". The rejection region is

$$\overline{X} < \mu_0 - t_\alpha(n-1)\cdot\sqrt{\frac{S^2}{n}}.$$

In our case we have $n = 10$, $\mu_0 = 1.2$ and $\alpha = 5\%$. Plugging in the values of $\overline{X}$, S^2 and $t_{5\%}(9)$ from parts (a) and (b) gives

$$
\begin{aligned}
1.119 &< 1.2 - 1.833\cdot\sqrt{\frac{0.009677}{10}} \\
1.119 &< 1.143.
\end{aligned}
$$

Since this inequality is true, we reject "$\mu = 1.2$" in favor of "$\mu < 1.2$".

Remark: Outside the classroom, you would just plug your data into a computer and press a button. Here's what my computer says:

```
> X := [1.12, 1.01, 1.04, 1.10, 1.0, 1.04, 1.28, 1.17, 1.19, 1.24];
```

$$X := [1.12, 1.01, 1.04, 1.10, 1.0, 1.04, 1.28, 1.17, 1.19, 1.24] \qquad (3)$$

```
> Statistics[OneSampleTTest](X,1.2,confidence=0.95,alternative='lowertail',summarize=embed):
```

Standard T-Test on One Sample

Null Hypothesis:	Sample drawn from population with mean greater than 1.2
Alternative Hypothesis:	Sample drawn from population with mean less than 1.2

Sample Size	Sample Mean	Sample Standard Deviation	Distribution	Computed Statistic	Computed p-value	Confidence Interval
10.	1.11900	0.0983700	*StudentT*(9)	−2.60389	0.0142778	Float(- ∞) ..1.17602

Result:	**Rejected**: This statistical test provides evidence that the null hypothesis is false.

The "p-value" is the smallest value of α such that H_0 would be rejected. It is more useful to quote this than just saying that H_0 was rejected at $\alpha = 5\%$.

(d): We just quote a formula. We have the following two-sided, symmetric $(1 - \alpha)100\%$ confidence interval for μ:

$$\mu = \overline{X} \pm t_{\alpha/2}(n-1) \cdot \sqrt{\frac{S^2}{n}}$$

$$\mu = 1.119 \pm t_{\alpha/2}(9) \cdot \sqrt{\frac{0.009677}{10}}.$$

In our case we have $\alpha/2 = 2.5\%$ and $t_{2.5\%}(9) = 2.262$. Plugging this in gives

$$\mu = 1.119 \pm 0.0704,$$

or

$$1.049 < \mu < 1.1894.$$

In words: We are 95% confident that the true value of μ is between 1.049 and 1.1894.

How to Develop the Formulas From Scratch: We start with Gosset's definition of the random variable T_{n-1}, which says that

$$\frac{\overline{X} - \mu}{\sqrt{S^2/n}} \sim T_{n-1}.$$

Then we use algebra as in Problems 3 and 4.