Midterm 1 - Fall 2025

ECON 4753 — University of Arkansas

1. Say you have a sample of observations of some variable. You want to summarize the variable to a stakeholder.

   (a) [10 points] Describe two ways you can help someone understand the distribution of a variable and what function you might use in `R` to do this.

   Answer: E.g. `summary` to present summary statistics of the variable; `hist` to plot the full distribution of the variable visually.

   (b) [10 points] In your own words, describe what the concept of the sampling distribution of a statistic is. Why is it helpful to know about the sample distribution of a statistic?

   Answer: A sampling distribution represents the notion of 'repeated sampling', where you grab many different samples from the population of the sample size and calculate the statistic for each sample. The distribution of estimates is the sample distribution.

   (c) [5 points] Typically, we report 95% confidence intervals. Give an example where someone might want to use a higher level of confidence (e.g 99% or 99.9%)

   Answer: Higher levels of confidence make us wrong about the population mean less often. This is important when the cost of being wrong is high (e.g. trying to predict natural disasters).

2. This question is based on our review of statistics. Say you observe a sample of workers from a firm with sample size $n = 100$. You observe their wages w_i and want to estimate the average wage at the firm. You estimate the following statistics in your sample: $\bar{w} = 17.53$ and $\text{var}\left(w\right) = 4.2$.

   (a) [5 points] Given this information what is the (approximate) sample distribution of the sample mean?

   Answer: $\bar{w} \sim \mathcal{N}(\mu, 4.2/100)$. Note, it should not be $\mathcal{N}(17.53, 4.2/100)$.

   (b) [10 points] Form a 95% confidence interval for your sample mean. Interpret this in words.

   Answer: $17.53 \pm 1.96 * \sqrt{(}4.2/100) = (17.13, 17.93)$.

   With 95% confidence, the population mean falls between 17.13 and 17.93.

(c) [5 points] Another student claims the average worker earns \$17. Using your confidence interval, would you reject this null with a 5% significance level?

Answer: We can reject the null that the true average is \$17 because it does not fall within the 95% confidence interval.

As an alternative answer, we could calculate the t-statistic as $\frac{17.53-17}{\sqrt{4.2/100}} = 2.58$. Since this is larger than the critical value of 1.96, we reject the null with 5% level of significance.

Below is the result of a two regressions using data on nutrition information on Starbucks' food items.

3. First, we will look at this regression of the number of calories in the food item on indicators for each food type with 'bakery' beign the omitted group.

```
OLS estimation, Dep. Var.: calories
Observations: 73
Standard-errors: IID
                      Estimate Std. Error   t value   Pr(>|t|)
(Intercept)          368.78049    12.9273 28.527289  < 2.2e-16 ***
type::bistro box       8.71951    31.9934  0.272541 7.8603e-01
type::hot breakfast  -43.78049    31.9934 -1.368422 1.7568e-01
type::petite        -191.00271    30.4699 -6.268568 2.8685e-08 ***
type::sandwich        26.93380    33.8516  0.795644 4.2901e-01
```

(a) [5 points] What is the average amount of calories for food items in the 'petite' type?

Answer: $368.78 + -191.00 = 177.78$.

(b) [5 points] Which food type has the largest number of calories on average?

Answer: Sandwiches have the largest number of calories on average.

(c) [10 points] What is the difference in average amount of calories for 'bistro box' foods relative to sandwiches? How would you modify this regression to test if the difference is statistically significant?

Answer: $(368.78 + 8.72) - (368.78 + 26.93) = -18.21$.

To test for significance, we could set sandwiches as the omitted category.

4. Second, we regress the number of calories in the item on the amount of fat in each item (in grams).

```
OLS estimation, Dep. Var.: calories
Observations: 73
Standard-errors: Heteroskedasticity-robust
            Estimate Std. Error  t value   Pr(>|t|)
(Intercept) 183.2400   20.98025  8.73393 7.2892e-13 ***
fat          11.2768    1.09564 10.29240 1.0109e-15 ***
```

(a) [10 points] Interpret the coefficient on ‘`fat`’ in words. Comment on its statistical significance.

Answer: For every additional gram of fat, we predict the food will have on average 11.27 extra calories. This estimate is statisticallly significant at the 5% level.

(b) [5 points] Predict the number of calories in a food item with 14g of fat.

Answer: $183.2400 + 11.2768 * 14 = 341.11$ calories.

(c) [10 points] Construct a 95% confidence interval around the slope coefficient. What is the smallest slope that you can not reject with a 5% level of significance?

Answer: $11.2768 \pm 1.96 * 1.09564 = (9.13, 13.42)$. The smallest slope we can not reject with a 5% level of significance is 9.13.

Finally, we add the number of carbs in that item as a second explanatory variable.

```
OLS estimation, Dep. Var.: calories
Observations: 73
Standard-errors: Heteroskedasticity-robust
            Estimate Std. Error  t value  Pr(>|t|)
(Intercept) 39.10401  13.709974  2.85223 0.0057027 **
fat          9.98090   0.476424 20.94964 < 2.2e-16 ***
carb         3.62562   0.202739 17.88319 < 2.2e-16 ***
```

(d) [10 points] Why do you think the slope coefficient on ‘`fat`’ went down after controlling for ‘`carb`’?

Answer: Since `fat` and `carb` are likely positively correlated in foods, the coefficient on `fat` was previously getting ‘credit’ for the differences in `carb`.

Any answer that comments on the two being correlated will get points.