Exam preparation Jeff

Be able to define these terms and describe why they're important:

Describing data:

- Parameter vs. statistic:
    - The parameter is a numerical summary of the population.
    - A statistic is a numerical summary of the sample data, and our guess at the population.
- Data: Observations gathered for individuals, on various characteristics of interest.
- Data file: A data file has a separate row of data for each subject and a separate column for each characteristic. Basically, a dataset; data frame…
- Observations: They are the rows of the dataset; a single subject/unit. It is a fact/figure we collect about a given variable.
- Variable: They are the columns of the dataset; a variable is a characteristic that can vary among subjects/units in a sample or population.
- Population: The population is the total set of subjects/entities that a study observes.
- Sample: A sample is a subset of the population on which the study collects data.
- Qualitative/quantitative data:
    - Quantitative → Numeric
    - Qualitative → Categorical
- Nominal/ordinal (scale of measurement):
    - Nominal: Unordered categories (ex: sex).
    - Ordinal: Ordered categories (ex: low, middle, high income).
- Continuous/discrete (granularity):
    - Continuous: a variable is continuous if it can take an infinite continuum of possible real number values
    - Discrete: a variable is discrete if its possible values form a set of separate numbers (0, 1, 2…).
- Relationship between classifications:
    - Quantitative – interval scale
    - Categorical unordered – nominal scale
    - Categorical ordered – ordinal scale
    - Quantitative variables – discrete or continuous
    - Categorical variables – discrete
- Skew: The skew of a distribution (from a population parameter or sample statistic) describes if such distribution is asymmetrical about its mean. If a distribution has a positive skew, the tail to the right is longer (containing more observations), and the mean is larger than the median. When the skew is negative, the tail to the left is longer, and the mean is smaller than the median.
- Outliers: Formally, an observation is an outlier if it falls more than 1.5(IQR) above the upper quartile or below the lower quartile. Less formally, an observation is an outlier if it is very atypical or falls far away from where most observations cluster.

Measures of Central Tendency and Dispersion:

- These measures describe, numerically, the center of the data (a typical observation) and the variability/dispersion of the data (the spread around the center).
- Mean: it is the sum of the observations divided by the total number of observations, or the "center of gravity" of the data. Highly influenced by outliers. With dummies, the mean is the proportion of observations that equal 1.
- Median: It is the observation that falls in the middle of the ordered sample. If the sample size is even, the median is the midpoint between the two (or the average). Not affected by outliers. It is thus preferred for skewed distributions, as it better represents what is typical.
- For skewed distributions, the mean lies toward the direction of the skew (the longer tail) relative to the median.
- Deviation: it is the difference between an observation and the sample mean.
- Variance: it is the sum of deviations, divided over the sample size-1. It is approximately an average of the squared deviations.
- Standard deviation: it is the square root of the variance; the squared root of the sum of squared deviations over the sample size-1. It is approximately an average of the deviations, or a sort of typical distance of an observation from the mean.
- The reason for using n-1 has to do with statistical inference about population parameters.

Distributions:

- Sampling distribution: A sampling distribution is a probability distribution of a statistic that is obtained through repeated sampling of a specific population. If we draw samples of a population over and over, and calculate a statistic (like the mean or beta), the sampling distribution would tell me how frequent (or likely) each possible value of the statistic is.
- Normal Distribution: the normal distribution is symmetric, bell shaped, and characterized by its mean μ. The probability within any particular number of standard deviations of μ is the same for all normal distributions. This probability is roughly 0.68 within 1 sd, 0.95 within 2 and 0.997 within 3. The standard normal is a T distribution with df = infinity.
- T Distribution: the t distribution is symmetric, and bell shaped around a mean of 0. The standard deviation is a little larger than 1; its precise value depends on the degrees of freedom, and the t distribution has a slightly different spread for each value of df, and different t-scores apply for each df value.
    - For inference about the population mean, the df = n-1.
    - The larger the df value, the less spread out, and after about 30 df, the t is almost identical to the standard normal.
    - For inferences about the population slope in regression, df = n-2.
- Decision between Normal and T:
    - Según Est II, el uso de la Normal/T depende de si conocemos/no conocemos la desviación estándar (y el SE) del parámetro (en la práctica nunca lo conocemos). Según el libro y Jeffrey, el uso de la Normal/T depende del tamaño de la muestra, porque de eso depende si aplica o no el TCL (y podemos asumir normalidad de la distribución de muestreo). Para muestras más chicas (el threshold que vamos a usar es 30), usamos la T. "La sustitución del SE poblacional por el muestral introduce error extra, que puede ser grande cuando n es chica. Para arreglar este error, reemplazamos el z-score con un t-score. El intervalo de confianza se vuelve más amplio. El t-score es como un z-score, pero viene de una distrubución bell-shaped que está más esparcida que la normal estándar; la distribución T". Según el libro, "the T method makes the additional assumption of a normal population distribution". El Z score no hace esto cuando n >= 30, porque ahí sí aplica TCL.
    - Chat: The Z-score is typically used when you have a large enough sample size (usually n > 30) and either the population standard deviation (σ) is known or the population size is assumed to be very large. In such cases, the sampling distribution of the sample mean is approximately normal (follows a standard normal distribution), and you can use the Z-score to calculate the confidence interval. The T-score, on the other hand, is used when the population standard deviation (σ) is unknown, and you are working with a smaller sample size (typically n < 30). The T-distribution has heavier tails than the standard normal distribution and takes into account the additional uncertainty introduced by estimating the population standard deviation from the sample.
    - As n → ∞, the T-distribution with df → ∞ becomes a normal distribution with a mean of 0 and a standard deviation of 1, which is often referred to as the standard normal distribution. So, as n gets larger, the two distributions become very similar.

    - For the test:
        - n < 30 → Use T
        - n >= 30 → Use Z
- Degrees of Freedom:
    - DF can be thought of as the number of independent pieces of information that we have to calculate our estimate, or as the number of values that are free to vary in our dataset.
    - For three pieces of information, once we know the mean, we only need two pieces of information to know everything. The idea is that every time I make a calculation, like a mean, I'm using up one piece of information.
    - Within the T distribution, the more degrees of freedom, the more the distribution resembles the Normal.
- Standard deviation/variance: the standard deviation/variance tell us how much our data are spread around the mean. The sd allows for interpretation as the same units of our variable. Within the N/T distributions, the variance/sd are parameters of the N distrib.
- Sampling distribution: it is a probability distribution of a sample statistic, such as the sample mean or sample proportion. It specifies probabilities for the possible values of the statistic for all the possible samples.
- Central Limit Theorem:
    - For any variable, regardless of how the population behaves, the sampling distribution of the mean tends to Normal with mean = population mean, and variance = population variance / n (standard error or SE).
    - Sample sizes and cases → TLC applies for **large**, **random** samples, or:
        - If my variable X is normally distributed, the sampling distribution of the mean of X tends to Normal even with n = 1.
        - If my variable X is symmetrically distributed, the sampling distribution of the mean of X tends to Normal after n = 15.
        - If my variable X is distributed otherwise or its distribution is unknown, the sampling distribution of the mean of X tends to Normal after n = 30.
    - As n gets larger, SE gets smaller.
- Standard error (SE): It is the standard deviation of the sampling distribution of our statistic. It is equal to the sd/sqrt(n).

Estimation (and CIs):

- Point estimate: In inference, it is the single number that is the best guess for the unobserved parameter. In a nutshell, our point estimate is a sample statistic obtained from the observed sample.
    - Our point estimate is our *estimation*, resulting from using our *estimator* (the formula to calculate the parameter for the sample; for instance the formula for the sample mean).
    - No confundir con el parámetro que tengo en la formula de los estadísticos, ese solo sirve para sustituirlo con el valor que le asigno en la H0.
- Interval estimate/confidence interval:
    - A confidence interval is an interval within which the true parameter for the population is believed to fall. It consists of a point estimate (our sample statistic), +/- a margin of error (critical value at desired confidence level * SE).
    - How to interpret confidence intervals:
        - Suppose 90% CI, n = 15
        - "If we took size 15 random samples of the population over and over again, and calculated 90% Cis for all of them, then the true parameter would fall within those bounds 90% of the time." The 90% is called the confidence level.
        - "The confidence coefficient is the probability that an interval would contain the parameter with repeated sampling" (slide 71 W1).
        - NO probabilistic statements; NO "90% confidence".
        - Interpretations and conclusions of inference exercises must always be in line with assumptions and context ☺
    - Larger n, narrower intervals (the margin of error depends on the SE of the sampling distribution of the point estimator, and the SE itself depends on the sample size)
    - Lower confidence, narrower intervals (trade-off)
- Basic idea of Inference / CIs:
    - We plug in the estimated value of σ to estimate the s of the sampling distribution of the mean (or to estimate our SE, since the SE is the σ of the sample).
    - We use our mean statistic/sample mean to estimate the true mean.
    - Now we have estimated a sampling distribution N(sample mean, SE).
    - We then use our knowledge to find our CI.
- Bias:
    - An estimator is unbiased if its sampling distribution centers around the parameter / the parameter is the mean of the sampling distribution. E(µˆ) = µ with repeated sampling.
    - A biased estimator tends to under or overestimate the parameter on the average.
- Efficiency: An estimator is efficient if it has smaller standard errors than those of other estimators. SE of estimator is smaller.
- Consistency: An estimator is consistent if its variance tends to zero as n tends to infinite. And/or if its mean tends to the true parameter as n tends to infinite.
- Sufficiency: An estimator is sufficient if it is not dependent of the parameter to estimate.

Assumptions for statistical inference (these determine how and if we can make inference):

- Todos los modelos:
    - How the data is generated (randomization)
    - Sample size (and decisions based on it)
    - Population distribution (not relevant after a certain sample size)
    - Does the TCL apply? (always does after a certain sample size)
    - Type of data
    - Parameter to be estimated
    - Test statistic decision
- Regression:
    - All of the above, plus:
        - The mean of Y is related to X by the linear equation E(X) = alpha + beta(X) (linearity).
        - The error is normally distributed with mean 0 and constant variance/sd.
            - For our error:
                - Mean = 0 (this is what the model is meant to do)
                    - The expectation of our error term, given x, = 0.
                    - This means that x does not help predict e.
                - The sd of our error is the same at each value of x
            - This is equivalent to saying that yi is normally distributed, with mean of alpha + b(Xi), and variance v. Values of y at each value of x follow a normal distribution, with the same sd at each value of x (constant variance of y).
    - Independence between observations → errors are not correlated (not sure if this is additional or just another way of saying mean of errors = 0)
    - No perfect collinearity in regressors (only for multivariate) + variation in x.
    - Homoscedastic errors.

Hypothesis testing: → How unlikely is it to observe the data we observe, if H0 were true?

- Hypothesis
    - Statement about a population. Usually, a prediction that a parameter describing some characteristic of a variable takes a particular numerical value or values within a certain range of values.
- Null/alternative hypotheses
    - Null: statement that the parameter takes a particular value (usually "no effect").
    - Alternative: statement that the parameter falls in some alternative range of values (usually effect).
- Test statistic (estadístico Z, estadístico T…)
    - The parameter to which the hypotheses refer has a *point estimate*, and the test statistic summarizes how far that estimate falls from the parameter value in H0. This is often expressed by the number of SEs between the estimate and the H0 value.
    - The general form of a test statistic is (observed data – what we expect if null is true)/(average variation or SE).
    - The test statistic itself is the formula to calculate this distance; to calculate the score.
    - The VALUE of the test statistic is the "score", for instance the Z score or T score.
    - And yet another thing is the critical value, which is the value of Z or T that results from a given confidence level (Z of 0.05, which is 1.96, for example).
    - Terms:
        - Parameter: Population mean.
        - Estimator: Sample mean formula.
        - Estimation / statistic: Calculated value of sample mean.
        - Test statistic: Z statistic. (Difference between observed value and what I think the true value under H0, standardized). It's called the test statistic because it also is built with sample inputs.
        - Score: Calculated value of our statistic (can be our T Statistic, Z statistic).
- P-value:
    - It is he probability that the test statistic equals the observed value or a value even more extreme in the direction predicted by Ha, if H0 were true. It is calculated by presuming that H0 is true; it answers "supposing H0 is true, how likely is it that I would observe this test statistic value, or one more in the direction of Ha?".
    - It's simply a transformation and sort of "standardization" of test statistics, because different tests use different statistics, but the p-value interpretation is "universal", it's a probability 0 to 1.
    - Ex: P-value = 0.05 means "we would observe a test-statistic this extreme or more 5% of the time (1/20 times) if H0 was true".
- Significance level (α-level)
    - It is a number such that we reject H0 if the p-value is less than or equal to it. The conventional levels are 0.05 and 0.01.
- Type I & Type II error
    - Type 1: H0 true and we rejected it. The probability of Type 1 error is α, which we set up in advance.
    - Type 2: H0 false and we did not reject it.

  - As the probability of Type 1 error decreases, that of Type 2 increases. A smaller α makes it harder to reject H0, even when it's false.
  - Also, P(Type 2) decreases as the sample size increases, and as the parameter is farther from H0.
- Link with Cis
  - "In testing H0: mu = mu0 against Ha: mu != mu0, suppose we reject H0 at the 0.05 α level. Then the 95% confidence interval for mu does not contain mu0. The 95% confidence interval level consists of those mu0 values for which we do not reject H0 at the 0.05 α level."

Steps in hypothesis testing:

1. Assumptions and decisions.
2. Hypotheses.
3. Test statistic: measure of the number of SEs that the point estimate falls from the H0 value. Comparison to critical value**: If the absolute value of the test statistic is greater than or equal to the critical value(s), reject H0** (this is connected to the decision rule with the p-value).
4. P-value: Probability that we would observe the test statistic (or an even more extreme one in the direction of Ha) that we do, given that H0 is true. Larger test statistic and smaller p-value are more evidence to reject H0. P-value = $p(Z\ or\ T\ \geq |observed\ test\ statistic|)$. We must multiply that probability by 2 in two-tailed tests (with Normal and T, which are symmetrical).
5. Conclusion: Report and interpret P-value; if p-value is smaller than α, we reject H0. This means that your sample data provides enough evidence to conclude that the alternative hypothesis (Ha) is supported.

Regression:

- Linear regression / regression analysis:
  - It is an approach to model the relationship between an input variable (X) and an output variable (Y). As the name suggests, it is linear, and thus proposes that Y is a linear function of X, illustrated by Y = alpha * beta (X).
  - The regression function describes how the mean of the response variable changes according to the value of an explanatory variable.
  - Beta and alpha are then called the regression coefficients which we estimate with least squares, which provides our sample prediction equation, which looks like ^y = a + bx.
  - At a certain value of X, y^= a + bx estimates the mean of y for all subjects in the population having that value of X.
- Y-intercept: The Y intercept (alpha) is the value of Y when X = 0. Within the linear regression model, our estimate of alpha (a) is the expected value of Y when X = 0.
- Slope: The slope (b) tells us the estimated change in Y that is associated with a one-unit increase in X. "A 1-unit increase in X corresponds to a b = something increase in the predicted number of Y".
  - More precise estimation with larger n, smaller estimated sd, and more spread in x.
- Residuals: A residual is the difference between an observed value and the predicted value of the response variable (y – y^).
- Least squares line:
  - Least squares is a method to build the regression line. It provides the line of best fit by minimizing the squared residuals from the plotted line.
  - The LS estimators are the most efficient among all unbiased. (Gauss-Markov).
- Sum of squared error / Residual sum of squares (**SSE/RSS**):
  - $\sum(y-\hat{y})^2$
  - We summarize the size of the residuals by the sum of their squared values (we squared because otherwise negative and positive would cancel themselves out). This is the Residual Sum of Squares. Same term according to book (p. 264).
  - The least squares estimates a and b are the values that provide the prediction equation y^= a + bx with the minimum (least) residual sum of squares. The prediction line is called the least squares line.
  - The SSE summarizes the variability around the prediction equation.
  - The SSE quantifies the squared distance from the Y values to the fitted line.
- Total sum of squares (**TSS**):
  - $\sum(y-\bar{y})^2$
  - The TSS summarizes the variability of the observations on y. It's quantity, divided by n-1, is the sample variance of the y-values.
  - The TSS quantifies the overall squared distance of the Y values from the overall mean of the response $\bar{Y}$
- Regression sums of squares (**RegSS**):
  - $\sum(\hat{y}-\bar{y})^2$
  - Quantifies the squared distance from the fitted line to the overall mean.
- R squared / coefficient of determination:

    - R squared = (TSS – SSE) / TSS or 1 – (SSE/TSS) or even better, RegSS/TSS→ Literally squared correlation coefficient
    - Proportion of variation in the response y that is explained by the model
    - It falls between 0 and 1, when r squared = 1, SSE = 0 (all points fall on the prediction line). Here, r =-1 or 1.
    - R squared, like the correlation, measures the strength of linear association. The closer to 1, the stronger the association, and the more effective the least squares line is compared to $\bar{y}$ at predicting y.
  - Correlation coefficient (r/R):
    - The correlation is a standardized version of the slope (of the regression line). Its value does not depend on the units of measurement of X and Y. R = (sx/sy)b.
    - R measures the strength of association, it is between-1 and 1, has the same sign as the slope… We can think of the correlation coefficient as an average of the products of standardized values of our observations.
    - Pearson correlation assumes that both x and y are approximately normally distributed.

- Y: dependent, response, **outcome**, output…
- Y: independent, **input**, explanatory…

Other notes:

- Ch. 7 A& F: Comparing 2 groups. Diferencia de medias, diferencia de proporciones.
- Ch. 8 A&F: Categorical variables & association.
    - Contingency tables / conditional distributions.
    - Chi Squared test of independence (p. 225)
    - Observed vs. expected (expected if independent)
    - Residuals and standardized residuals.
    - Connection between Chi-squared and Normal distribution for 2x2 tables.
    - We can also use difference of proportions or odds ratio.
    - Also ordinal measures of association like the numbers of concordant & discordant pairs and Gamma.
    - Chi squared good for nominal; Z test based on sample gamma good for ordinal
- Pp. 208 & 289 useful charts.
- Chi squared, correlation coefficient and gamma tests not in Stats II ITAM.
- Statistical independence: Two variables are statistically independent if the conditional distributions of the population are identical across categories.
- Standardization: subtract the mean from each observation, divide by the sd.
- Two types of prediction:
    - Predict an individual's response at any chosen value x0 the predictor variable (a prediction interval)
        - Ex: What is the predicted birth weight of an individual baby who has a gestational period 275 days?
    - Predict an average response when X = x0 (a confidence interval for the mean)
        - Ex: What is the predicted average birth weight of babies who have a gestational period of 275 days?
    - There's more variability in individual responses than in average responses, so these intervals take different forms. The point estimate/prediction is the same, what changes is the confidence level (the intervals get wider for individual responses). We do a better job in predicting a mean than an individual response.
    - We get better predictions towards the mean of x, because this is where the bulk of the observations is. Also, extrapolation will yield bad predictions.

Calculate/execute these concepts:

Measures of central tendency:

- Mean
  - Sum all values; divide by sample size.
  - $\bar{x} = \frac{1}{n}\sum X_I$
- Median
  - Order and choose middle, or average of middle for even n.

Measures of dispersion:

- Variance
  - $s^2 = \frac{\sum(x_i-\bar{x})^2}{n-1}$
- Standard deviation
  - $s = \sqrt{\frac{\sum(x_i-\bar{x})^2}{n-1}}$
- Standard error
  - It is the standard deviation of sample statistics.
  - For example, the "standard error of the mean" refers to the standard deviation of the distribution of sample means taken from a population. The smaller the standard error, the more representative the sample will be of the overall population.
  - $SE = \frac{s \; or \; \sigma}{\sqrt{n}}$

Distributions and probability:

- Define a sampling distribution of a mean
  - If CLT applies (n>30), the sample mean is normally distributed, centered around the true mean (the mean of the sampling distribution is the true mean) and with standard deviation equal to the population sd/n.
  - Otherwise, the sample mean follows a t distribution with n-1 df.
- Calculate a Z-score
  - A Z score is the result of substituting in my test statistic (Z statistic, in this case) formula. The Z statistic for the mean is the difference between my point estimate and my hypothesized population mean, over the standard error (sd/sqrt(n)).
- Use the Z-score to find p-value
  - The P-Value is the probability of observing my Z score or an even more extreme one, given that H0 is true $p(Z \; \geq |observed\ test\ statistic|)$.

Regression

- Interpret y-intercept and slope for a linear function
    - y-intercept: value of Y when X=0
- Write a prediction equation
    - $y = a + bx + e$
- Calculate Sum of Squared Errors (RSS)
    - $\sum(y - \hat{y})^2$
- Calculate Total Sum of Squares (TSS)
    - $\sum(y - \bar{y})^2$s
- Calculate regression sum of squares (RegSS)
    - $\sum(\hat{y} - \bar{y})^2$
- Keep in mind: $TSS = RegSS + RSS$
- Interpret a scatter plot:
    - Form/pattern
    - Direction
    - Strength
    - Outliers
- Construct a CI around β
    - Our CI in this case is $\boldsymbol{\hat{\beta}}$ +/- t(alpha/2) * se$\boldsymbol{\hat{\beta}}$
- Conduct a hypothesis test for β
    - Out test statistic for β is $t = \frac{\hat{\beta} - \beta}{SE_{\hat{\beta}}}$, with df = n-2
- Calculate standard error for β
    - The standard error for β (the slope) is (quite ugly):
        - $SE_{\hat{\beta}} = \frac{\hat{\sigma}}{\sqrt{\sum(Xi - \bar{x})^2}}$
- Calculate Sx and Sy
    - $s_x = \sqrt{\frac{\sum(x_i - \bar{x})^2}{n-1}}$
    - $s_y = \sqrt{\frac{\sum(y_i - \bar{y})^2}{n-1}}$
- Calculate r and interpret its meaning
    - $r = \frac{s_{xy}}{s_x s_y}$
    - $s_{xy} = \frac{\sum(x_i - \bar{x})(y_i - \bar{y})}{n-1}$
- Calculate r2/R2
    - $r^2 = (\frac{s_{xy}}{s_x s_y})^2$
- Interpret a regression analysis table
    - Know how to interpret intercept, slope, r and r squared…
- Interpret regression coefficients
    - α: intercept; expected value of Y when X = 0.
    - β: slope; expected change in outcome for 1-unit increase in input.

Practice problem set:

A) The following table contains the GDP per capita (in thousands of international dollars) for four European countries.

Belgium	Germany	France	Luxembourg
38	38	35	90

1. Find the variance.

- mean $= \frac{38(3) - 3 + 90}{4} = \frac{114 - 3 + 90}{4} = \frac{111 + 90}{4} = \frac{201}{4} = 50.25$
- var $= \frac{(50.25 - 38)^2 2 + (50.25 - 35)^2 + (50.25 - 90)^2}{4}$

$= \frac{(-12.25)^2(2) + (-15.25)^2 + (39.75)^2}{4} = \ldots = 515.68$

2. Find the standard deviation.

- sd $= \sqrt{var} = \sqrt{515.68} \approx 22.72$

3. Would you say that one of these observations is an outlier?

- we can't really calculate the IQ range $(Q_3 - Q_1)$, and outlier would be a value
  - ↳ $Q_1 - 1.5 \cdot IQR$
  - ↳ $Q_3 + 1.5 \cdot IQR$
- however, the lux. obs is an outlier....

B) The ”Freshman Fifteen” is an expression that commonly refers to an amount (some- what arbitrarily set at fifteen pounds) of weight often gained during a student’s first year at university. You decide to test whether this expression holds true for Trinity College students. You randomly select 16 second years and gather data on how much weight (in pounds) they gained the previous year. The mean change of your data is 14.5 lbs. and the sample standard deviation is 0.8 lbs.

1. Identify the population for this study.

- The population is the whole body of TCD students.

- $\bar{x} = 14.5$ lb
- $\bar{sd} = 0.8$ lb.

2. Describe the sample distribution for this study.

- It consists of 16 randomly selected second-year TCD students who provided data on weight gain.

3. Describe the sampling distribution for this study as precisely as possible.

- The sampling dist. represents the dist. of sample means (mean weight gain) for random samples of 16 sec-year TCD students. It describes how the sample mean weight gain varies accross different random samples of the same size for the population of TCD students.

4. Calculate the point estimate and a 95% confidence interval for the population mean. Explain what your confidence interval means.

- Point estim : 14.5 lbs. = sample mean.
- 95 % CI :
  - ↳ $z_{95} = 1.96$ (for large n)
  - ↳ $t_{15} = 2.13$ (since $n < 30$)
  - ↳ $SE = \frac{sd}{\sqrt{n}} = \frac{0.8}{\sqrt{16}} = \frac{0.8}{4} = 0.2$
  - ↳ CI = point est ± score · SE
  - ↳ $CI_z = 14.5 \pm 1.96 \cdot 0.2 \Rightarrow (14.108, 14.892)$
  - ↳ $CI_t = 14.5 \pm 2.13 \cdot 0.2 \Rightarrow (14.07, 14.92)$
- this means "if we were to draw random <u>n = 16</u> samples of TCD students over & over again, and calculated <u>95%</u> CIs for all, the <u>true pop. mean</u> weight gain would fall within those CIs <u>95% of the time.</u>"
- Mamadas.
- Also, upper bound = either 14.89 or 14.92 ; 15 16 is not contained in our CIs.

D) Way back when, Apple claimed that iPhone 5 is "the biggest thing to happen to iPhone since the iPhone." Among other improvements, the iPhone 5 claimed improved battery life over the old versions. For example, the standby time has been improved to 225 hours (a 25-hour improvement over iPhone 4S). To test this claim, you collect a sample of battery longevity from 100 randomly selected owners of the iPhone 5. Among these 100 owners, you find that the battery life in the new iPhone 5 is 217 hours with a standard deviation of 40 hours. Test the research hypothesis that the batteries in the new iPhone 5 differ from the 225 hours claimed by Apple. Use a 0.05 level of significance.

1. Random samp - ; quantitative data ...

   $\alpha = 0.05$

2. Ho :

   Ha : $\mu \neq 225$

3. test statistic :

   - $z = \frac{\bar{x} - \mu}{\sigma / \sqrt{n}}$
   - $t = \frac{\bar{x} - \mu}{s / \sqrt{n}} = \frac{217 - 225}{40 / \sqrt{100}} = -8/4 = -2$
   - critical value $t_{0.05,99} = 1.66$ & -1.66

     $(z_{0.05} = 1.64)$

     calculator
   - $t < 1.66$ & $t < -1.66$ ∴ t stat falls into rejection region.

4. P-value < 0.05

   calculator

5. Reject Ho ....

D) Imagine you are interested in the different patterns of support for the Spanish government among citizens of Catalan population. You decide to conduct a survey asking people "Do you have confidence in the national government?" Possible answers include Yes or No. Your were able to poll 243 Catalans. Of these 243 respondents, 86 said, "Yes."

1. Provide a point estimate for the percent of Catalans that have confidence in the government.

- $\hat{p} = \frac{86}{243} \cdot 100 = 35.4\% =$ point estimate

2. Identify the sampling distribution of this study. Be precise.

- the sampling distribution for this study represents the distrib. of sample proportions $\hat{p}$ (the proportion of catalans who have confidence in gov.) for random samples of $n = 243$ catalans.

$\nearrow \ \alpha = 1 - 0.92 = 0.08 \ ; \ \alpha/2 = 0.04$

3. Construct a 92% confidence interval of the percent of Catalans that have confidence in the government.

$\circ \quad CI = \text{sample prop.} \pm (z_{0.04}) \cdot SE$

$\hookrightarrow \quad z_{0.04} = 1.75$

$\hookrightarrow \quad SE = \sqrt{\frac{\hat{p}\hat{q}}{n}} = \sqrt{\frac{0.354 \cdot 0.646}{243}} = \sqrt{\frac{0.22}{243}} = \sqrt{0.00094}$

$$\hookrightarrow \quad CI = 0.354 \pm 1.75 (\sqrt{0.00094})$$
$$= 0.354 \pm 0.0536$$
$$= (0.3004 \ , \ 0.4076)$$

4. Test the theory that less than 40% of ~~Greeks~~ Catalans support the government using a 0.05 significance level. $\longrightarrow \ \alpha = 0.05$

1. Random, categorical ...

2. $H_0 : p \geq 0.4$

   $H_a : p < 0.4$

3. $$z^* = \frac{\hat{p} - p}{\sqrt{\frac{pq}{n}}} = \frac{0.354 - 0.4}{\sqrt{\frac{0.22}{243}}} = \frac{-0.046}{0.0306} = -1.50$$

   $z_{0.05} = -1.6$

   $|-1.5| < |-1.6| \quad \therefore \quad \underline{\text{NO}}$ rejection

4. p-value = 0.0668

5. p-value

$\alpha = 0.08$

$\alpha/2 = 0.04$

E) Suppose a random sample is taken of 200 rat-hunting dogs in New York City. The mean number of rats killed by a dog is 19, with a standard deviation of 2. Construct and interpret a 92% confidence interval for the mean number of rats killed.

- $\bar{x} = 19$
- $sd = 2$
- $n = 200$
- $$\begin{aligned} CI &= \bar{x} \pm z_{\alpha/2} \cdot SE \\ &= \bar{x} \pm z_{\alpha/2} \cdot \frac{sd}{\sqrt{n}} \\ &= 19 \pm z_{0.04} \cdot \frac{2}{\sqrt{200}} \\ &= 19 \pm 1.75 \cdot \frac{2}{\sqrt{200}} \\ &= 19 \pm 1.75 \cdot 0.1414 \\ &= 19 \pm 0.2472 \\ &= \boxed{(18.75,\ 19.24)} \end{aligned}$$

- "If we were to draw n = 200, random samples of rat-hunting dogs in NYC over & over again, and constructing 92% CIs for all of them, the true pop. mean would be contained in those CIs 92% of the time".