- The basicz of statisticz
- Awright let's goooooooo
- Satistics: The art and science of ***LEARNING FROM DATAAAAAAAAA***
- (a bit more review I did on paper but eh I'll go back to the TeXey fun timez)
- The point of statistical inference: Make a best guess, since you can't work with the entire population
- Our unknown is $\hat{\theta}$
- Point estimation: We have $\bar{x}$, and the standard error $SE(x)$
- Or interval estimation: Take a sample out of the population. Calculate $\bar{x}$. Validate that μ is what we think it is
- The variables we care about might be either categorical or quantitative
- Categorical: It places an individual into a category (ie gender, zip code). Numbers may be assigned
- Quantitative/numerical: Take certain values and do mathey stuff with them. Ex height, weight, GPA, whatever
- The key is that it makes sense to average. Average GPA makes sense, average zip code doesn't
- Quantitative may be discrete or continuous
- We care about response variables and explanatory variables tooooo
- Response variable: The dependent variable, which we are modelling/predicting. Usually y
- Explanatory variable: Independent vars. Used to explain the response. They affect y in some way, and they're usually $\hat{\mathbf{x}}$ (vector because maybe more than one x)
- Very common one: $y = \beta_0 + \beta_1 + \epsilon$, typical linear regression
- Okay back to flower boiz. Are the vars (number of red petals, distance from road, height) categorical or quantitative?
- Ans: Closeness to road is categorical, others are qualitative
- Number of petals is discrete, the height is continuous because we don't know for sure it will be an integer value
- Explanatory variables: height, distance from highway. Response variables is # of red petals

- A taaaable!
- One quantitative variable $\rightarrow$ simple linear regression, and multiple linear regression if you have a bunch
- ANOVA is uesful with categorical variables. Not sure what ANOVA is
- If you have both categorical and quantitative you can do ANCOVA but not in this class, and logistic regression is good if you want to find categorical variables for the response
- TIIIIME FOOOOR CONFIDENCE IIIINTERVAAALS
- Confidence interval: An interval within which we can be confident the parameter lies
- For estimating number of red petals, parameter of interest is population mean of number of red petals, μ
- Our entire population is 200 flowers. Our mean is $\mu = 3.955$. Sample size 20 taken and plot 'em, $\bar{x} = 4.35$
- How to estimate μ from $\bar{x}$?
- We got ourselves a funkeh histograaaam (two in fact)
- The funkeh histogram is unfortunately skewed, and not symmetric, so not normal. Laame. Ditto for sample
- So instead of one sample take like a billion (actually 10K)
- Get all the sample means, all the $\bar{x}$s up in this hizouse
- Do it 10K times, result is the mean of all the $\bar{x}$s is the mean of the means and that is 3.959, and standard deviation across all those is .51
- OH HEY! If you take a bunch of samples the average of them is μ. So $E(\bar{x}) = \mu$ and so it's unbiased!
- We also get standard deviation $SD(\bar{x})$ or $\sqrt{V(\bar{x})}$
- And even if you start with two skew boys, if you plot the values for all the $\bar{x}$s it's normal! CENTRAL LIMIT THEOREEEEEM
- Unfortunately we can't really take 10000 samples. We can only take one. So now what?
- Well we can at least guess based on that sample. Parameter is θ, statistic is $\hat{\theta}$
- Estimate of variability of $\hat{\theta}$ is $SE(\theta)$. In this case that's $\frac{\sigma}{\sqrt{n}}$
- So: Start with the pivotal quantity. What's the pivotal quantity?

- Pivotal quantity: A variable whose distribution is free from any parameters
- Ex: If $P \sim N(0,1)$ it depends on no parameters or statistics
- So look at the quantiles of pivotal quantity K, construct the confidence interval
- Confidence interval is: $\hat{\theta} \pm K_{1-\frac{\alpha}{2}} SE(\hat{\theta})$
- Ex: 20 flowers from 200. 95% confidence interval. $[3.11, 5.59]$.
- So θ = mean number of red petals in 200 flowers. Given as 3.955 (but we usually won't know this for sure)
- What about $\hat{\theta}$? Mean number of petals in sample of 20. It's our first sample and that sample has $\mu = 4.35$
- What is α? 95% confident so $1 - \alpha = .95 \Rightarrow \alpha = .05$
- What's K? It's $N(0,1)$ (we could use a student's t but let's just go for a normal for now)
- What is $SE(\hat{\theta})$? We don't actually know
- But we know $\hat{\theta} + Z_{\frac{\alpha}{2}} SE(\hat{\theta}) = 5.59 \Rightarrow 4.35 + 1.96 \times SE(\hat{\theta}) = 5.59$
- $\Rightarrow SE(\hat{\theta}) = .6327$
- How the heck do we get our K value? We are bounding it on both the lower and upper bound, so $1 - \frac{\alpha}{2} = .975$
- So figure out the value of Z so that the CDF of the normal distribution is equal to .975. That happens to be 1.96
- And now for the hypothesis test!
- Alternative hypothesis: We wanna prove this
- Null hypothesis: The thing we hope is not true
- Say we want to prove that the mean number of red petals is different for flowers near or far from highway
- μ_{far} =: mean number of red petals when distance = far
- μ_{near} =: Similar but near
- H_1, the alternative guy we wanna prove: $\mu_{far} \neq \mu_{near}$
- H_0, the null boi: $\mu_{far} = \mu_{near}$
- Equivalent to: $H_0 =: \mu_{far} - \mu_{near} = 0$

- $H_1 : \mu_{far} - \mu_{near} \neq 0$
- So how do we do this? We get the test statistic assume the null guy is true, and if the CDF reports that the chance of that happening is super rare ie $< 5\%$ chance then you can reject the null dude
- Type 1 Error: A false positive. Reject the null when it is true. Probability of this is α

$$P(\text{ Reject null}|\text{null is true}) = \alpha$$

- Type 2: Fail to reject the null when it actually isn't true

$$P(\text{Fail to reject null}|\text{null is false}) =: \beta$$

- β is related to the power of the test and I don't really know what that means but I guess I don't have to cuz I'm not a grad lol
- P-value: Assume null hypothesis is true. Construct a normal distribution with this in mind, convert to standard normal or something like it, and use the CDF to see how likely that actually was. Then reject null if it's super unlikely (less than α where you decide on α before you collect the data)
- If the P-value you get is more than α you cannot draw any conclusions
- In any statistical problem or analysis, you usually pick $\alpha = .05$. Why? I dunno. It's up to you to decide it
- What is the relationship between the two? I guess one is you figure out the
- We're comparing two guys here: so $\theta = \mu_{far} - \mu_{near}$
- But we don't know the μ guys! So we'll use sample guys, $\hat{\theta} = \bar{x}_{far} - \bar{x}_{near}$ (divide your sample into the far dudes and the near dudes)
- Now calculate the mean number of red guys who like frappes and you get what you want
- Hypothesized value: Hypothesize they're the same so $\theta_0 = 0$
- Construct the normal and compute the CDF, $p = 8.76 \times 10^{-8}$. That's real small, so reject null dude and accept alternative dude
- But oh no, maybe we had a type 1 error! Not type 2 because type 2 means we fail to reject the null, and we did reject the null so it must be type 1 (if any error happened at all)
- (There is an error $\cap$ H_0 was rejected) $\Rightarrow$ (the error was type 1)

- Interpret the p-value here: The chance that the near flowers and far flowers have the same number of petals and we just ended up with a weird sample, is very very low
- Relationship between confidence interval and hyp test: If the hypothesized value is within CI, you fail to reject, but if hypothesized value is outside of CI, you reject it