review

- big picture of applied stats: see 36200 image idk
- we have statistics ($\overline{x}, \hat{p}, \ldots$) and standard error ($\text{SE}_{\overline{x}}, \text{SE}_{\hat{p}}, \ldots$)
- population: literally everyone, hard to measure
- sample: subset of population
- parameter: perfect summary (e.g. mean height)
- statistic: measurable summary (e.g. mean height of sample)
- stderr of stat: typical variation due to random sampling.
  - diff error formulae for each stat.
  - this course: simply calc with software
- inference: give estimate and measure of how far off it might be
  - if statistic is random and sampling distribution known, we have probabilistic inference; can get p-value or margin or err

1 variable EDA

- categorical
  - bar graph
  - percent summaries
- quantitative
  - histogram
  - center: $\overline{x}$, median
  - spread: stddev, IQR, range
  - five number summary/box plot

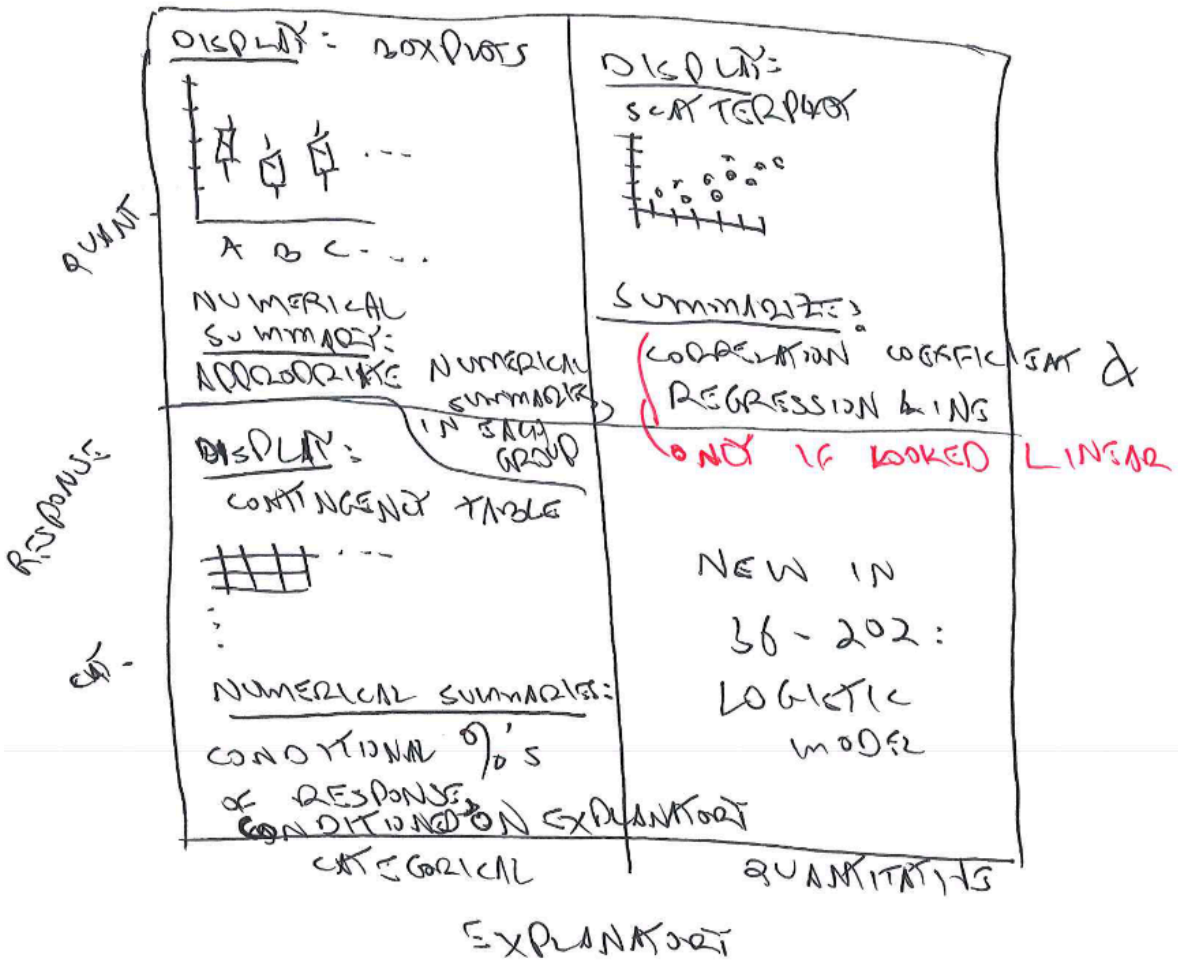
1 variable transformations

- need normal distributions?
- $x^{\frac{1}{n}}$, $\log(x + c)$ so everything is > 1.
- the above's inverses
- quantile plots (qqplot) can help us diagnose if normal enough (look for straight line)

2 variable EDA

- explanatory x axis $\rightarrow$ response y axis

Review of 2 Variable EDA (graphs and summaries to explore bivariate relationships)
[Reference: prerequisite course]



1 variable inference

- statistics ($\overline{x}, S_x, \ldots$) predicts parameters ($\mu, \sigma, \ldots$)
- components:
  - point estimation: estimate via single number calculated
  - interval estimation: give plausible interview and how plausible
  - significance testing about hypotheses: assess evidence for/against claim about
- 95% confidence interval for μ is $\overline{X} \pm 2 \cdot \text{SE}_{\overline{X}}$
  - (works for arbitrary parameter/statistic estimate)
  - any sample Standard Error SE is $\frac{S}{\sqrt{n}}$ with sample stddev S (but remember, we just use software)
  - technically, 2 should be t_{crit} which varies with n, but it approximates to 2 for 95% confidence when large n
- hypotheses testing
  - H_0 vs H_A
  - "p value is compared to significance level. we do (not) reject the null hypothesis. we do (not) have sufficient evidence that ..."
  - remember: p finds boolean evidence of difference from norm, not magitude of difference

Statistical Model Primer

- statistical models are often of form: quantity = expectation + error
- in 1 variable, eg: $Y_i = \mu + \varepsilon_i$ where μ is the prediction and ε_i is the error at i.
  - we also specify the distribution and mean + stddev of the errors
- in 2 variables, eg: for some X axis value, $Y_i = \mu_{Y|X} + \varepsilon_i$
  - we also specify the shape, center, spread of the distribution of errors

Simple Linear Regression

- our model idea is $Y_i = \beta_0 + \beta_1 X + \varepsilon_i$ where we assume the errors are
  - independent, mean 0, constant stddev/spread (for required for least squares)
  - are normal (required for inference)
  - (can be denoted iid, $N(\mu = 0, \text{variance} = \sigma^2)$)
- our **sample** regression equation is $\hat{y} = b_0 + b_1 X$
- notice that we have three parameters: β_0, β_1, σ
  - they are estimated by b_0, b_1 (when using least squares), and $\hat{\sigma}$: what R calls "Residual standard error"
- to apply the model:
  1. **state** the model
     - eg: "we use the SLR model. $\text{vision distance} = \beta_0 + \beta_1 \cdot \text{age} + \varepsilon_i$ where errors are independent, mean 0, constant stddev, normal.
  2. **validate** the data works for the model
     - linearity: visual inspection
     - errors are:
       - independent: residual plot. residuals "patternlessly" above and below 0 line.
       - mean 0: residual plot. reasonably centered around 0.
       - constant stddev: residual plot. reasonably constant spread, scanning left to right
     - if there are problems, consider diff model/transformations
  3. **estimate** the parameters
     - use software to find $b_0, b_1, \hat{\sigma}$
  4. **inference**: is data probably showing a relationship between X and Y?
     - t test for $\beta_1 =$ or $\neq 0$
  5. **measure strength** of model with R^2 (if not chance)
     - R^2 is the percent of variability in Y that can be attibuted to the linear relationship with X
     - "Multiple R-squared" in R. NOT "Adjusted"
  6. **predict** of Y from X (for individual with X or all people with X)
     - the equation predicts the point estimate of Y given X
     - get prediction vs confidence interval via R for probable values of Y for individual or all at X