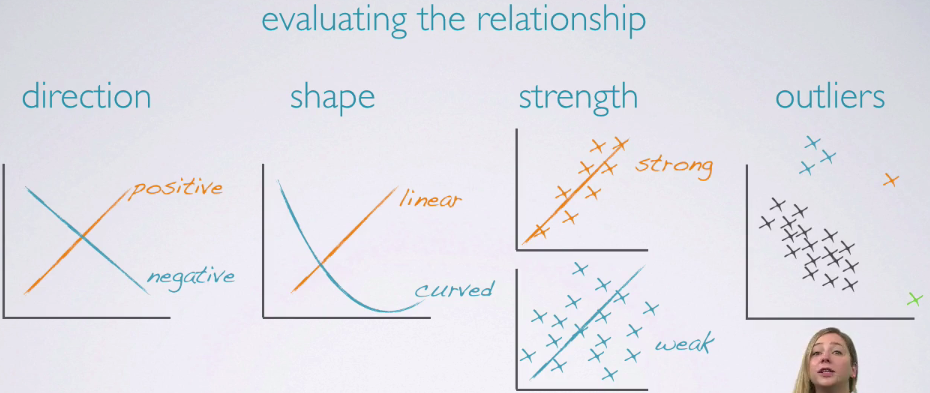
***COURSERA: STATS W/ R SPECIALIZATION***

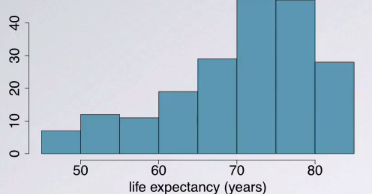
***COURSE 1 - Introduction to Probability and Data***

**WEEK 2- Exploratory Data Analysis and Introduction to Inference**

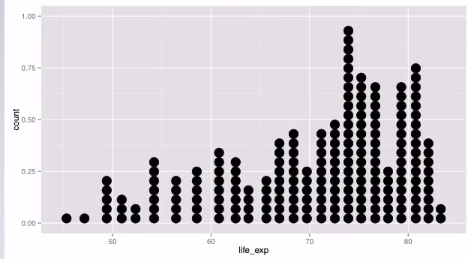
* Visualizing numerical data:
* **Scatterplots** (i.e. suspect income per person, or GDP/capita influences life expectancy, we plot life expectancy (response) as a function of GDP/capita (explanatory)
* Only correlations, not causal explanations



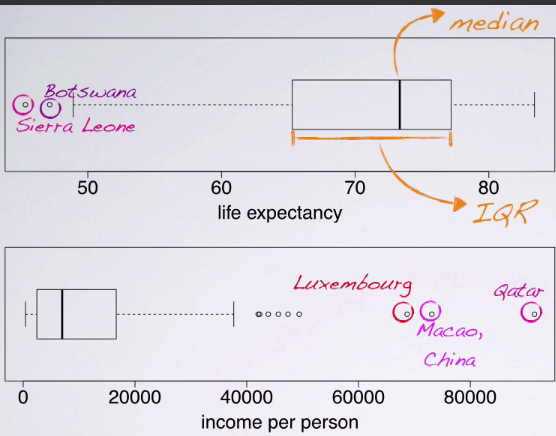
* **Outliers** may be interesting cases, so they must be handled them w/ careful consideration of the research question and of other associated variables
* **Histograms** 🡪 distributions of numerical variables = a view of **data density** (higher bar = more common data points)



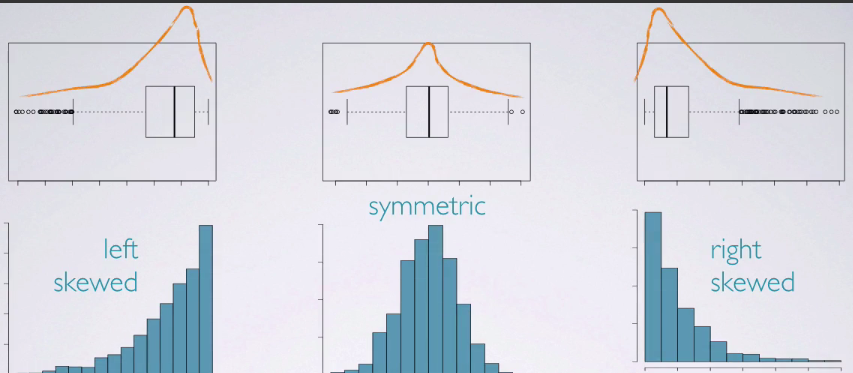
* Whether a histogram is unimodal, bimodal, multi-modal, or uniform is important
* uniform = each value has an equal chance of being represented
* Bin width/size can also alter the story a histogram tells
* Too wide = lose interesting data
* Too narrow, difficult to get overall picture of distribution
* Ideal width depends on data 🡪 play w/ it
* **Dot plot**

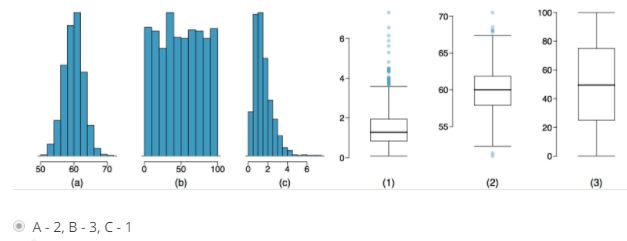


* Useful when individual data points are interesting, but can get busy w/ large sample sizes
* **Boxplots**
* Helpful to find outliers 🡪 displays median, IQR, max + min, outliers, skewness, but NOT modality (check a histogram)

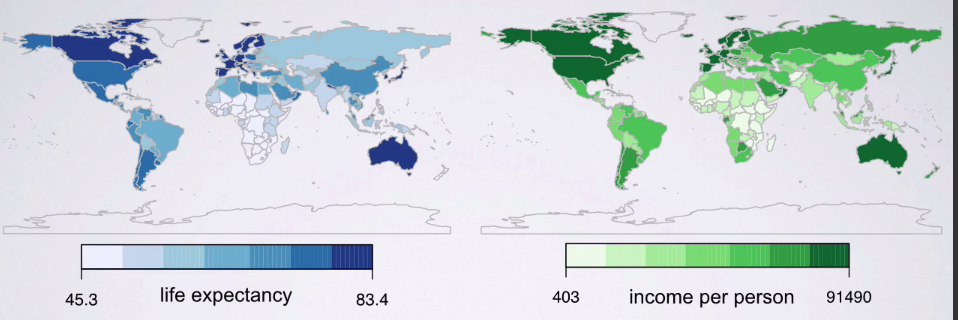


* To see skewness in boxplots, imagine its histogram 🡪 peak is around median and tails extend out to tails of boxplots



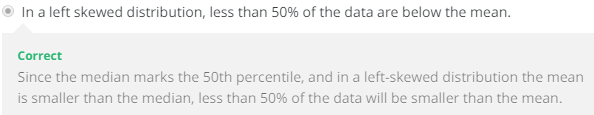


* **Intensity maps** of spacial distributions can also be helpful

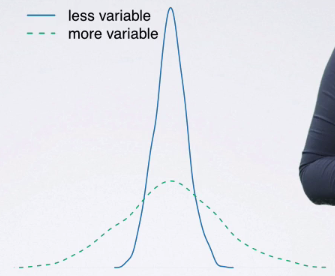


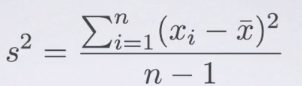
Ex: both income + life exp. are lower in Africa and higher in N. America and Europe

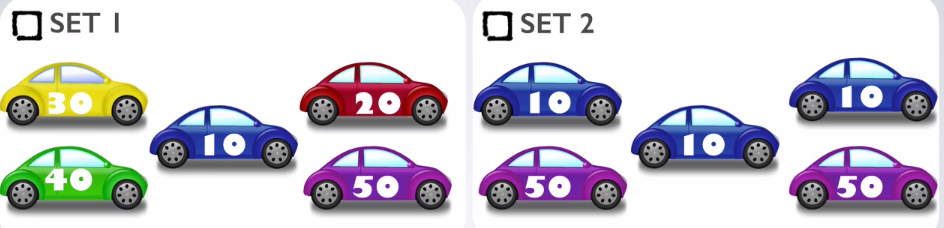
* **Sample statistics = point estimates** of **population parameters** (which are almost always unknown b/c it’s not feasible to have info on all observations in a population)
* Sample stats might not be right, but should be good estimates *if sample is good* (i.e. representative of the population)
* Note: Latin = sample, Greek = population
* W/ continuous distributions, may be very unlikely to observe the same value multiple times, so **mode** isn’t very useful
* **Mean** is pulled to the side of the tail (below **median** in left skews, above it in right skews)



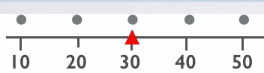
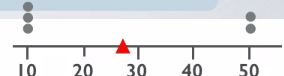
**Measures of Spread**

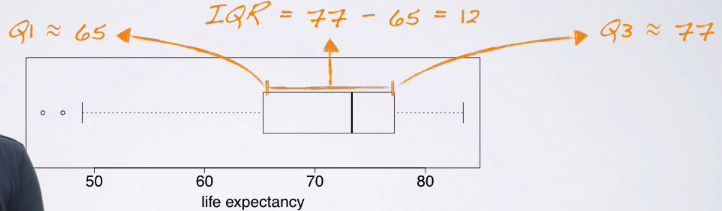
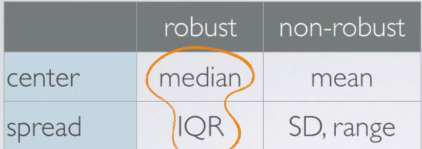
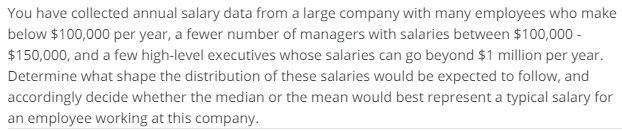
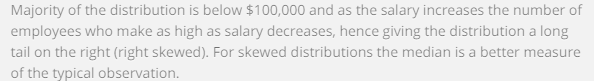


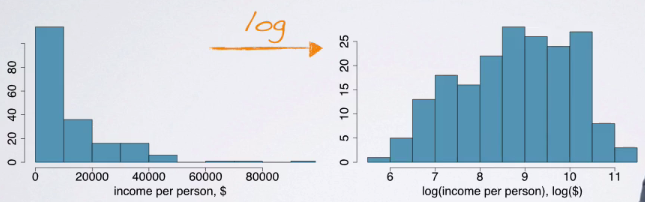
* **Range** = very easy to calculate, but not robust b/c it relies on the most extreme values of a distribution
* **Variance** = (s^2 for sample, sigma^2 for population) = average squared deviation (to account for negative differences) of each observation from the mean
* We square the deviations instead of taking absolute values to increase larger deviations more than smaller ones so that they’re weighed more heavily
* 
* **Variance** is not in units we can understand, so we take square root to get SD + a measure in units we can understand (same as data)
* **SD** = same units as data, more useful (s for sample, sigma for population)
* Variability vs. **Diversity**



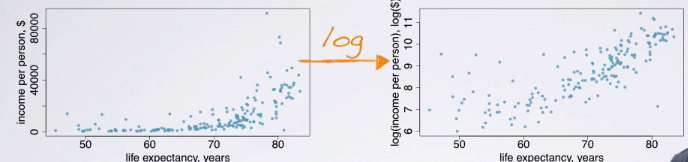
* Set 1 is more **diverse** *in color* = more different b/c more colors (each car has own color)
* Set 2 is more **variable** *in MPG* (more data at ends of distribution/away from the center)
* More observations around the center = less variable
* Avg MPG in Set 1 = 30, but has observations near the mean
* Avg. MPG in Set 2 = 26, but w/ no observations near the mean

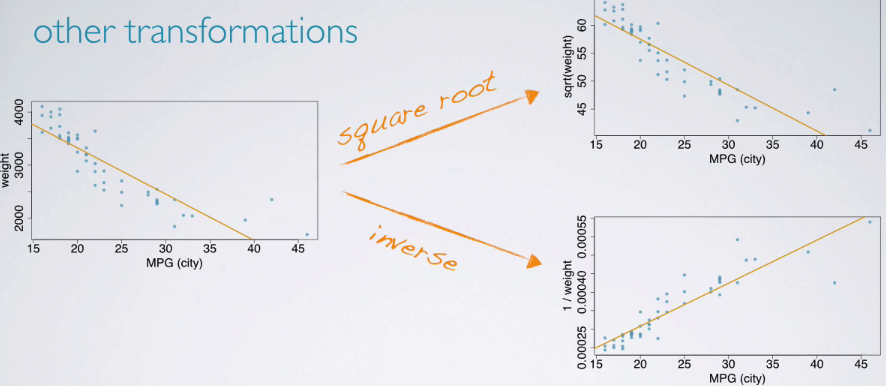
* **IQR** = middle 50% of data = Q3 – Q1 = 75th percentile – 25th percentile
* 
* “middle 50% of people in population have life expectancies between 65 and 77 years old”
* Value of IQR is not that useful on its own, but is when comparing distributions
* More reliable than range b/c it DOES NOT rely on endpoints
* **Robust statistics** = measures on which extreme data points have no effects
* Median, NOT mean (arithmetic average vs . middle data point (Q2/50th percentile) where endpoints are irrelevant to its calculation
* IQR therefore, which is based on the median, is more robust than SD (uses the mean) + range (relies solely on most extreme values)
* 
* Use robust statistics for *skewed distributions*/those w/ extreme values
* 
* 
* **Data Transformations** = useful tricks for making certain types of data easier to model
* **Transformation =** rescaling data using a function (sometimes used w/ skewed data so it’s easier to model so outliers are not as extreme)
* Most common = **(natural) log transformation** 🡪 used w/ much of the data cluster to 0 *relative to very large value in the data set + all observations are positive*



* **log transformation** can be applied to a variable in a scatterplot to make a relationship more linear + hence easier to model w/ simple methods



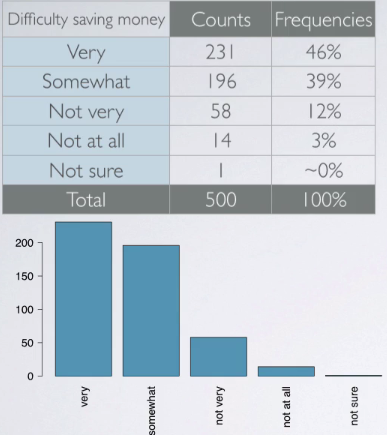
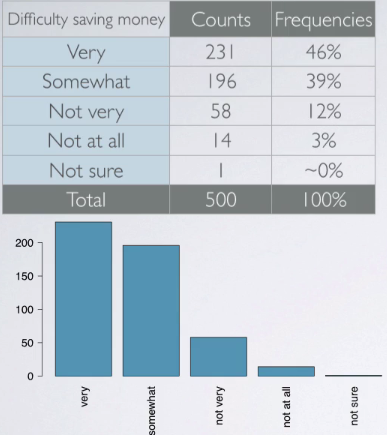
* relationship remains positive, but is now more linear + easier to model
* **Square root transform** or **inverse transformation** for cars weights vs. city MPG, with a negative non-linear relationship



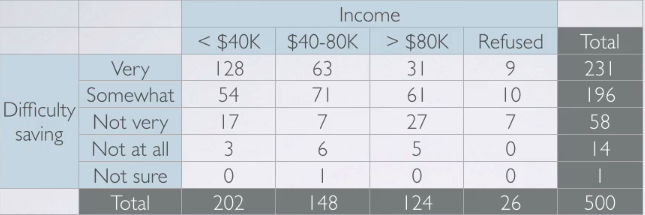
* Difficult to tell which transformation works better or if either give something better than the original data
* Transformations can be useful even if they complicate the interpretations a bit (log of income isn’t easy to evaluate)
* Goals of Transforms
* See data differently
* Reduce skew to assist in modeling
* Straighten a non-linear relationship to assist in modeling

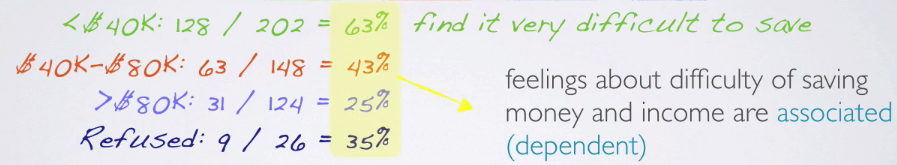
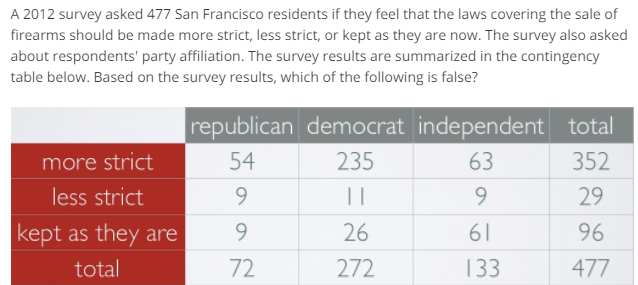
**Exploring Categorical Variables**

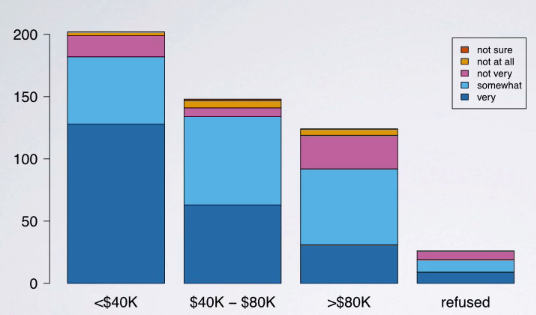
* Can visualize this kind of data in a **frequency table, relative frequency table, bar plot**, or relative **frequency bar plot** (same shape as raw bar plot but w/ frequency on y-axis instead), where the raw values tell us something about the data

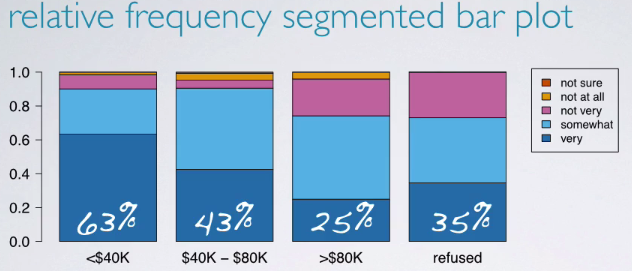
* Bar plots = categorical distribution w/ x-axis of naturally unordered classes
* Histogram = numerical distribution w/ a naturally ordered number line x-axis
* Pie charts are not that informative + can be bad w/ many levels of a class/factor
* **Contingency Table** 🡪 combine 2 categorical variables (income level bin, difficulty saving)



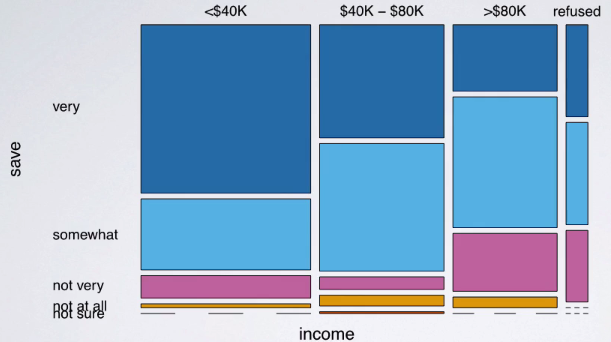
* Ex: want to compare whether income + perception of difficult saving are related, can’t compare raw counts since sample sizes for each income level are different
* instead, compare the distribution of 1 variable **conditional on the other**
* 
* Can assume that these are related b/c wide variability in perception of difficulty among income bins (*suggested* by data, not proved)
* 
* 
* Also, opinions on gun laws + party affiliations appear to be independent (all around same %, regardless of party)
* **Segmented bar plots =** useful for visualizing 2 categorical variables as 1 variable being **conditional** on the other 🡺 distribution of response variable conditional on levels of the explanatory



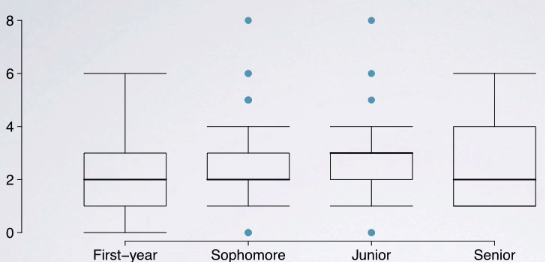
* Heights of bars = responses in various categories, segmented by color in another category
* Notice these are raw counts, not relative frequencies
* To explore the relationship between these variables, need a viz of relative frequency segmented bar plot



* Or via a **mosaic plot** of perception of saving difficulty conditional on income level + also shows **marginal distribution** of income (width of bars) as well

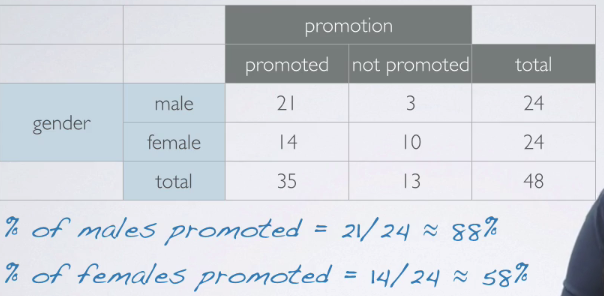


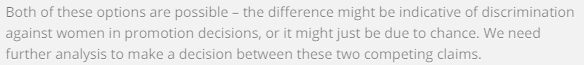
* More people make < $40k in this sample (wider bar), and 63% think its hard to save money (tall dark blue segment)
* Can see that the length of difficult perception colors varies by income level, indicating a difference of opinion among income bins/groups, suggesting a relationship between the 2 variables
* **Side-by-side box plots** are useful for relationships between a categorical + numerical variable
* Used when comparing numerical variable distribution across levels of a categorical variables (say # of clubs college students are involved in by class year)

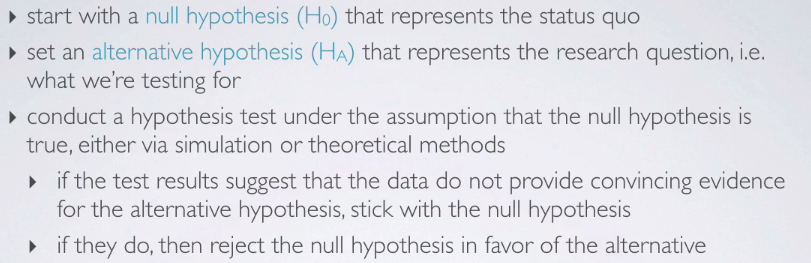


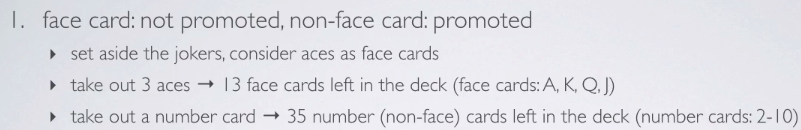
* See that, regardless of grade, students belong to roughly same # of clubs w/ more variability in seniors and freshman (larger IQRs/wider boxes) + some sophomore and juniors belong to unusually low and high numbers of clubs
* Since distributions across class years are pretty similar, its suggests grade level + # of clubs joined are not related (is independent of class)

**Intro to Inference**



* Can see a noticeable difference in proportions/% of each gender being promoted
* 
* 
* 
* 
* 2 **competing claims**:
* Promotion + gender are independent + difference in proportions is due to chance = **null**
* There are dependent + the observed difference is not due to chance = **alternative**
* **Hypotheses testing** is like court trials 🡪 innocent until proven guilty (null), alternative provides evidence to show null is guilty (false)
* Trying to figure out if data are possible (likely) if null were true
* If data were likely to occur under null, we retain/fail to reject null (say “not guilty”, but do NOT say “innocent” 🡪 have no way to be sure + don’t have enough evidence to “convict”
* If data were NOT likely (low probability (p-value) of occurring when null is true), reject null 🡪 have reasonable doubt null is true in favor of hypotheses
* Burden of proof is on the “unusual” claim/alternative, while null is the “ordinary” claim

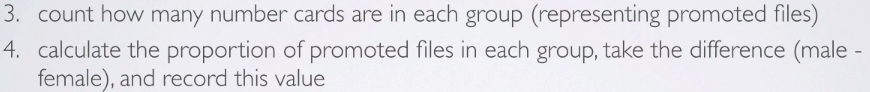


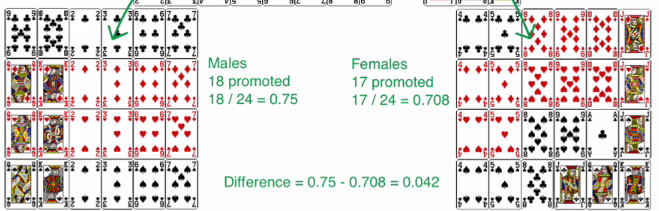


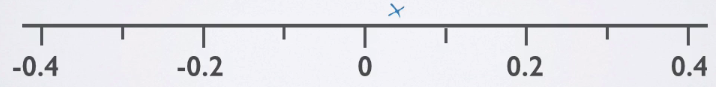
* Take out 4 cards to represent out promotion dataset 🡪 face cards = promoted, non-face = not promoted

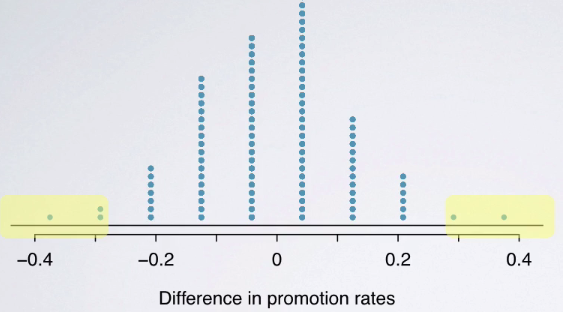


* Shuffling leaves this up to chance



* Would expect the difference between simulated proportions of males and females to be 0 if left due to chance b/c we randomly shuffled so we expect equal/near equal proportions of promoted + non-promoted in each group = 0 difference
* 
* Note this difference before running the simulation many, many times to build a distribution





* If the results from our simulations look like the observed data, we decide the difference between proportions of promotions by gender is due to chance + they are independent
* If not, it was unlikely to have happened by chance but was due to an effect of gender (discrimination) + that there’s evidence promotion + gender are dependent
* Can see our simulation distribution is centered at 0 (the null = difference is = 0 b/c there’s no bias in promotions)
* It’s also *very* rare to get a difference of ~0.3 like we observed in our data
* Our conclusion = These data show convincing evidence of an association between gender + promotion decisions made by male bank supervisors.