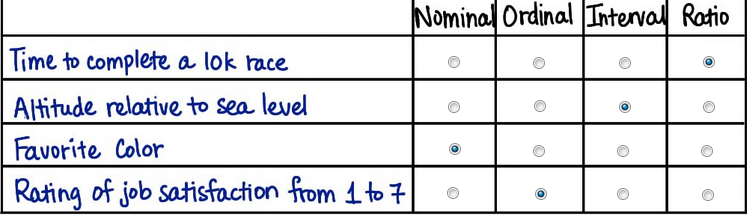
***Udacity Data Analyst Track***

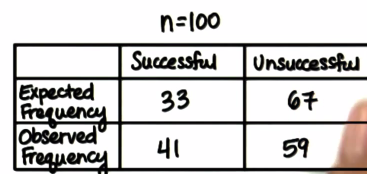
**I. Into to Inferential Stats**

11. Chi-Squared Regression

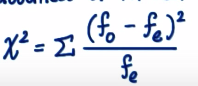
* Examples of scales of measurement
* % correct on a test 🡪 ranks w/ equal intervals + w/ an absolute 0 (cannot be < 0)
* Finishing order in a race 🡪 ranks 🡪 *no* absolute 0 b/c there is no 0 ranking + no matter how far behind someone you finish, you’re still assigned the same rank = *non-equal intervals*
* Temp in Celsius 🡪 rank w/ equal intervals, but no *absolute* 0 (can be negative)
* **Ordinal** = just a ranking w/ no interval or absolute 0 (race order, level in college (F,S,J,S), ranking of TV shows)
* Distance between 1st + 2nd is not necessarily the same as the distance between the 2nd + 3rd
* **Interval** = raking w/ equal intervals but no absolute 0 (*temps*, *years* that revolutions occurred [years go before 0 🡪 B.C.])
* **Ratio =** ranks w/ equal intervals + w/ an absolute 0 (student score, time taken to finish a maze, number of spelling errors)



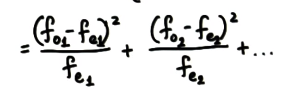
* All previous tests (hypotheses, t-test, z-test, ANOVA/F-test, Correlation, Regression) = **parametric test =** numericaltests of hypotheses that make assumptions about population parameters, mu + sig
* What if we ask a question/take a measurement taking a non-interval or non-ratio scale?
* Ex: Ask 100 people Y/N or beach or mountains?
* Can’t say “avg. fav. vacation spot is the beach”, so we use **frequencies + proportions** to describe these data
* These are **non-parametric tests** = hypothesis testing techniques that don’t require parametric info (mu + sig), such as the **chi-squared test**
* **Ex:** Want to summit a mountain. Mountain’s website says only 33% of summit attempts are successful. A professional guide says 41% of their 100 summit attempts were successful.
* We notice that in this case, there is no way to calculate a mean or SD, the data is *not* based on normal distributions, the data is still nominal (successful vs. unsuccessful), + the data is based on frequencies/proportions
* For each trip, we can write them as Y or N for a successful summit or not and count these frequencies
* From the whole population, we’d expect 33 out of 100 attempts by professional guides to be successful and 67 to not be so, while we observe 41 successful summits and 50 to not be so



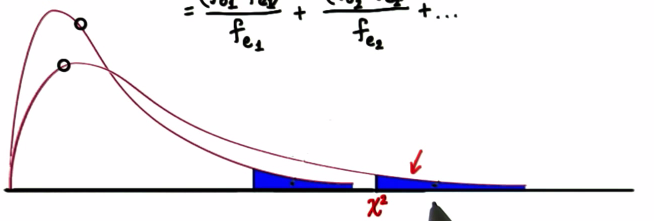
* Want to find out if this difference is significant and if we should use the guide.
* H(0) = expected frequency (guides have no effect on chances of success at a summit attempt)
* Other examples of h(0) = 50% of people prefer Coke, 50% prefer Pepsi, out of 2k people at a rodeo, 1k are male, 1k are female, 50 people out of 200 prefer rap, pop, country, or house
* ^^Not sure, so we ***assume***
* H(0) values are what we expect to find *based on the population* if there’s nothing different about each thing we’re testing (genre of music, taste of soda, chances of summiting, etc.)
* **Expected proportion \* n = expected frequencies**
* So, for the mountain example, we know what success to *expect* based on h(0) and we know what the guide company *observed* last year
* H(a) is basically just that h(0) is not true
* Now need to test how well *observed sample frequencies “fit” the population/expected proportions* via a **Chi-Squared “Goodness of Fit” Test**
* **Chi-squared =** Sum of all observed frequencies minus expected frequencies squared and divided by expected frequencies 🡪 **Sum[(f(o) – f(e)^2/f(e))]** = a ration of observed to expected frequencies

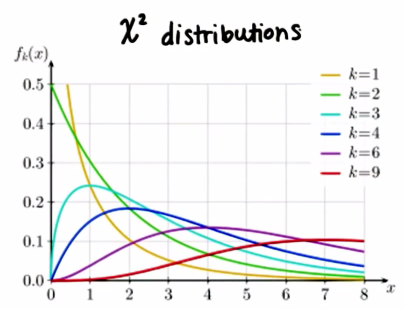


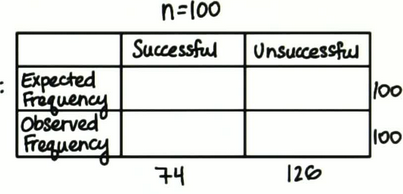
* Chi-squared is *never* negative (*frequencies are counts so are always positive*), so therefore a chi-square test is always one-directional in the positive direction (1-tail on the right w/ a critical value)
* *As we add more categories to our frequency table, the more likely our chi-squared value will be larger*, since we have to add up the frequency proportion for *each* category

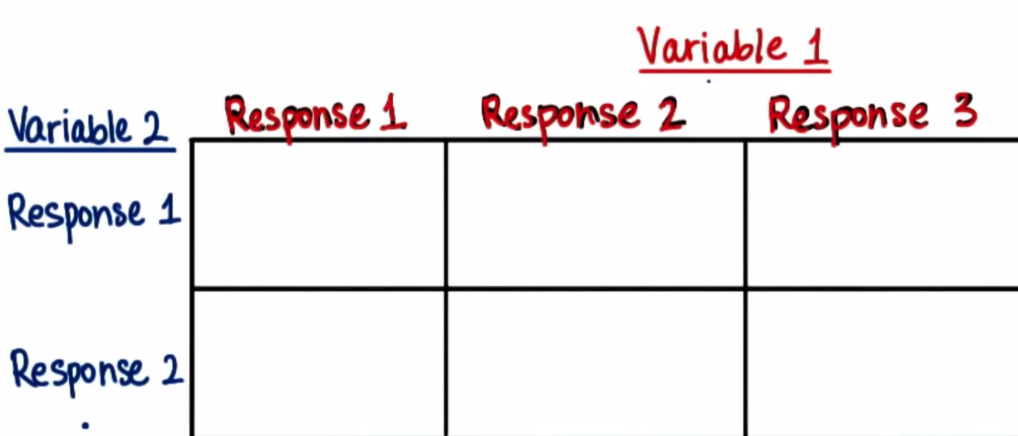


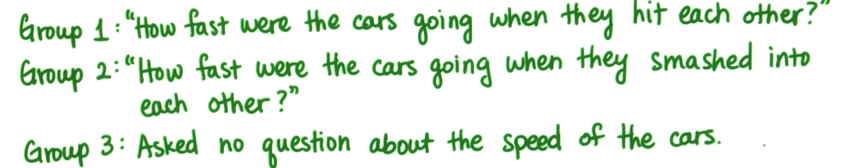
* More categories = more dF = larger chi-squared, but we don’t want to reject h(0) just b/c there’s a lot of categories
* Therefore, w/ more categories, we need a higher critical value in order to reject h(0)



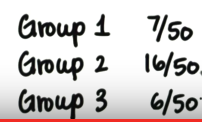
* Shorter distribution 🡪 critical value further to right = distribution has more categories = more Df
* All chi-squared distributions are positively-skewed w/ skewness decreasing as # of categories (+ therefore dF increases)
* 
* See the red line has the highest dF = 9 + least skewness
* So as categories (+ therefore dF) increases, skewness decreases = chi-squared distribution better approximates normal distribution (but it never becomes perfectly normal)

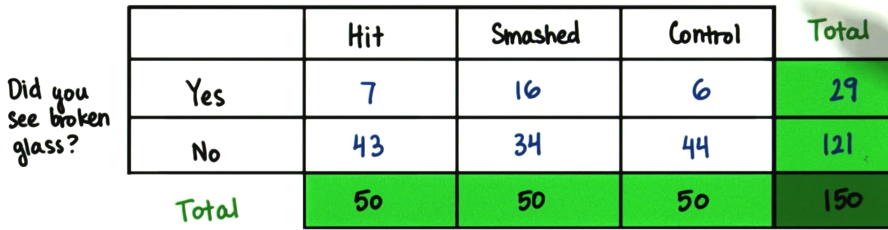


* dF here = 1 🡪 can only choose value for 1 box, and the rest have to automatically add up to those marginal totals
* So **dF = (number of categories – 1)\*(number of response – 1) = (2-1)\*(2-1) =** 1\*1 = 1
* Chi-squared Goodness of Fit test measure how well observed values match expected values for a certain variable
* Can also help us determine if 2 variables are independent 🡪 **Chi-Squared Test For Independence**
* 
* Here, instead of just having expected + observed values, we also look at the # of people who answered response 1 for both variable 1 and variable 2, who answered response 2 for variable 1 and response 1 for variable 2, and so on
* Ex: Does the wording of a question influence how well people remember details? n = 150 students of University of Washington watched 1 minute clip of a car accident + split into 3 groups of 50

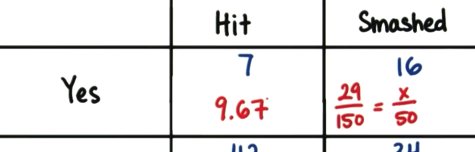


* After 1 week, all students were asked if they saw any broken glass (there was none)

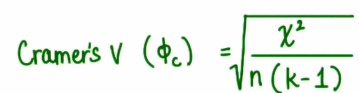




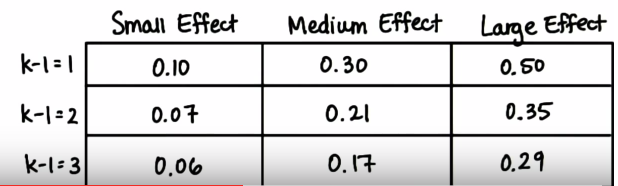
* Goal: To see if there’s independence between the 3 groups (hit, smash, control)
* By **independent**, we mean there’s *no consistent predictable relationship between the variables*
* H(0): Response for the glass question is independent of the wording in the question
* To show independence, we have to compare observed to expected
* Calculating expected frequencies can be tricky b/c they’re based on the # of points in each of the groups (3), but also the marginal totals for the responses (2)



* Basically take the marginal total and divide by the # of groups = expected value for each group for each response 🡪 **column total \* row total / grand total**
* But testing for significance is only part of the problem, then we should measure *the strength of the relationship between variables* = measure of **effect size**
* For this, we use **Cramer’s V 🡪 phi(c) = Sqrt(Chi/(n\*(k-1) where k = smaller of the # of rows or cols**



* How to interpret Cramer’s V as k increases:



* But we want to use this just as a labels to aid future power analysis, rather than fast + hard descriptors
* Assumptions + Restrictions for Chi-squared
* Avoid dependent observations, only use independence
* In the car crash example, data would be invalid if someone contribute data to more than 1 cell (asked 2 out of the questions)
* Avoid small expected frequencies
* In general, have a large # of participants/data points
* Chi-squared is just based off of a sample of observations, + we have corresponding expected values
* If using chi-squared to make assumptions about the population a sample is from, the total # of participants should be at least 20
* Conservative rule of thumb = each expected cell frequency should be at least 5
* So chi-squared values can check to see how well observed values fit expected values for categorical data
* In the case of **goodness of fit**, expected values are what we guess for h(0)
* determining if a set of categorical data came from a claimed discrete distribution or not.
* h(0) = they did, h(a) = they didn't.
* Answers if frequencies observed for the categorical variable consistent w/ theory
* Expands the 1-proportion z-test 🡪 1+ categories.
* But for **test of independence,** expected values are based on marginal totals
* determining whether 2 categorical variables are associated w/ one another in the population (race + smoking, education level + political affiliation)

