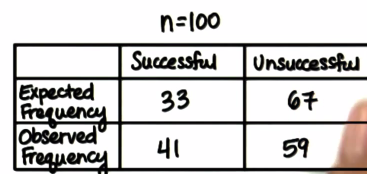
***Udacity Data Analyst Track***

**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))
* 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)
* 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 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**