What are Statistical Tests?

- Numerical data analysis
- Provides Between-group comparison
    - 2+ populations, intervention+control
- Is there a difference? How much of a difference? How much data supports our conclusion? What can we predict?
    - Depends on which test you run
    - Measures if there is a difference (not how much of a difference)

How do we apply them?

- Generate hypotheses
- Test them with the appropriate test
- Usually you will have a null and alternative
    - Null hypothesis: no difference between groups/no trend
    - Alternative: Significant results big difference in mean etc
- There are general questions/hypotheses, there are specific hypotheses
    - Dependending on statistical method

One sample test

- Comparing sample to expect mean
    - Sample of heart rates to expected population average heart rates
- Population standard deviation known
    - No: One sample t test
    - Yes:
        - n>30 - One sample z test
        - n<30 - one sample t test

Two Independent Sample Test

- Comparing two sample means with no relationships
    - Eg, test scores between people in different countries
- Population standard deviation known
    - No = two sample t-test
    - Yes
        - n>30 = two sample z test
        - n<30 = two sample t test

Two Paired Sample Tests

- Comparing two sample means with relationships
    - E.g, cholesterol rates before and after taking a drug
- Population standard deviation known
    - No - two paired t-test
    - Yes
        - n>30 = two paired z test
        - n<30 = two paired t test

More than two group comparison

- Multiple categories/groups
- ANOVA test
- One independent variable with 2 subgroups (levels) and one dependent variable = one way ANOVA
    - example: Brand of car vs life satisfaction
- More than two independent variables with one dependent variable = two way ANOVA
    - Example: brand of car, location of house, marital status VS life satisfaction

Probability Distribution Video:

- The chance of everything happening
- Distribution - ways of measuring patterns of occurrence in a dataset
- Links to a histogram - ways of visualizing probability of something happening (relative to other values)
- Equations model probability and are helpful in trying to approximate and understand how data might behave
    - even though real-world data are much messier and often less straightforward
- Two of the most basic distributions
    - Bernoulli - when there is a single trial with two outcomes
        - E.g, be like flipping a coin once and each of the two possible outcomes has some probability of occurring
        - Coin - 50/50
        - P (success)= 1- P(failure)
    - Binomial - Multiple bernoulli samples strung together
- Geometric - Continuation of binomials
    - Measures the number of failures you would need before success
    - Stringing together bernoulli trials
- Main one: normal distribution (Gaussian distribution
    - Sample mean is always at the peak of the bell curve
    - You can find 95 percent of the samples will have some values between one standard deviation (+ and - a standard deviation)
        - With two on each side, you will cover 99.7% of the sample
    - Used when there are continuous random variables
    - You can see whether the distribution pattern fits a histogram
    - Can be seen in larger data sets
- Poisson: Number of successful events over a time period
- Exponential: Time between the events rather than the actual number of events

Terminology

- Significance level: The probability of rejecting a null hypothesis when it is true
    - You don’t pick up on a valid/positive result
- Distribution: Representation of probabilities of where a value will lie

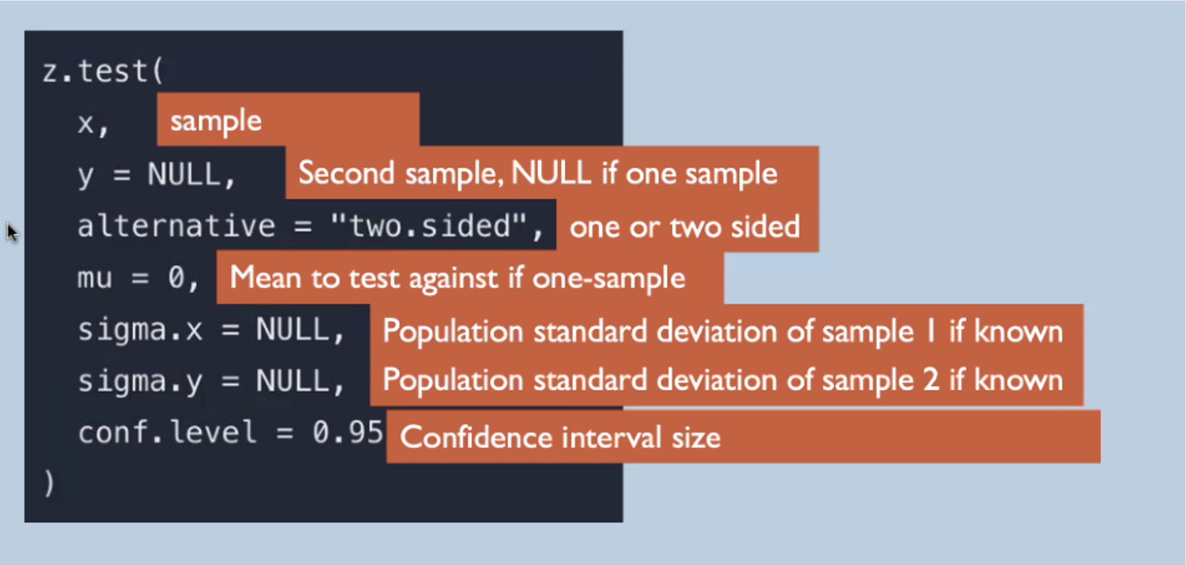
- P value: Probability of obtaining a result at least as extreme as the observed result, assuming the null hypothesis is correct
    - Probability of getting that significant result by chance alone

Normal Distribution

- Critical value: cut off value in a distribution, calculated depending on significance level
    - Use value table or calculator for sig level of 0.05 the critical value for a 2 sample test is 1.96
    - Below this result may be due to chance alone
- Confidence Interval (CI) = a range of values (interval) in which you would expect the true population mean to lie
    - Large confidence interval = significant result
- Quantile = % of values below a certain value

Tests in R

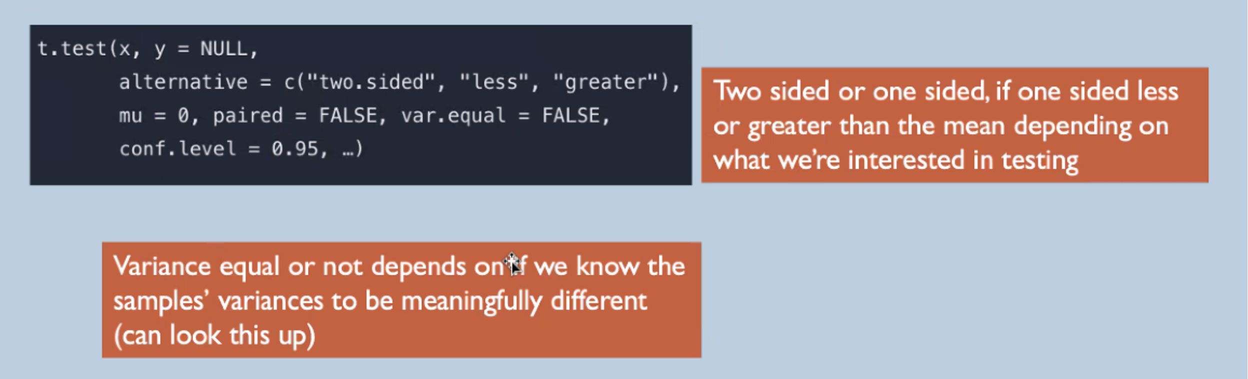
Z Test Command



```
z.test(
  x,
  y = NULL,
  alternative = "two.sided",
  mu = 0,
  sigma.x = NULL,
  sigma.y = NULL,
  conf.level = 0.95
)
```

sample
Second sample, NULL if one sample
one or two sided
Mean to test against if one-sample
Population standard deviation of sample 1 if known
Population standard deviation of sample 2 if known
Confidence interval size

T Test Command



```
t.test(x, y = NULL,
       alternative = c("two.sided", "less", "greater"),
       mu = 0, paired = FALSE, var.equal = FALSE,
       conf.level = 0.95, …)
```

Two sided or one sided, if one sided less or greater than the mean depending on what we're interested in testing

Variance equal or not depends on if we know the samples' variances to be meaningfully different (can look this up)

Exploratory Data Analysis Tools

Why?

- Understanding data
- Checking assumptions of tests
- Looking for interesting outliers
    - Don't just remove outliers - must have a compelling reason to do so

Box plots

- Useful when looking at numerical data
- Line in middle: Median Value
    - Better representation of data than mean when there are outliers
- Representation of the quartiles

Histogram

- Representing frequency of values (y axis is always frequency)

QQ Plots (Quantile Quantile)

- Scatter plot with quantiles of the data instead of raw numbers
- If curved at the top tail = it's a normal distribution

Scatter Plots

- Useful when you are comparing two variables
    - Helpful in identifying the relationship between them