Best walkthrough of stats meaning from initial data to regression:<http://varianceexplained.org/RData/code/code_lesson3/>

**significance level**: Area under the curve of the Null Hypothesis that you are willing to reject as coming from the Null H. The confidence level is equivalent to 1 – the alpha level. alpha level = significance level = probability of rejecting the null hypothesis when it is true = Type1 error. So, if your significance level is 0.05, the corresponding confidence level is 95%. For a significance level of 0.05, expect to obtain sample means in the critical region 5% of the time, even when the null hypothesis is true. Both the significance level and the confidence level define a distance from a limit to a mean, and they always agree. If the P value is less than alpha, the confidence interval will not contain the null hypothesis value.

Null hypothesis mean, hypothesis test representative: Hey buddy! I’ve found that you’re statistically significant because you’re more than $63.57 away from me!

Sample mean, confidence interval representative: Actually, I’m significant because you’re more than $63.57 away from me!

* Calc: critical t-value \* SEM
* Find the t-value that corresponds to the sig level you want \* (sd/sqrt(sampleSize))

**t.value,z.value**: measures the size of the difference between 2 different means (sample1-sample2, or sample-hypothesized) relative to the variation in your sample data. Put another way, The t-value is simply the calculated difference represented in units of standard error. The numerator is the signal(difference). the denominator is the noise(variation). t-value is the ratio of the signal to the noise

* Calc: t-value = (mean1-mean2)/(Variance1-Variance)
* SEM is just Variances/SD that are corrected for sample size.
* T-value(aka t-statistic) = (mean1 - mean2)/SEM
  1. SEMalone = sampleSD/sqrt(sampleSize)
  2. SEMpooled = sqrt(Var1/n1 + Var2/n2)

**p.value**: The probability of observing a particular t-value, if the Null H is true. The probability of observing a sample mean that was as many standard deviations away from the hypothesized mean as was observed, if the Null Hypothesis is true. If your p-value is smaller than your significance level, reject the Null H.

**confidence interval**: An interval around your t-value (your t-value comes from your sample mean) that is likely to contain the true population mean. If your confidence interval does not contain 0, reject the Null H. A narrow confidence interval [90 110] is more precise than a wide interval [50 150].You’ll still have a 95% Confidence Level for those values, but those values are much less precise, there’s a bigger range that the actual mean could fall intoIdeally, A 95% confidence level indicates that 19/20 samples randomly drawn from the same population will produce confidence intervals that contain the population/true parameter (eg:mean), however we generally don’t/can’t know the true population mean. So, The CI shows the range of sample means that you’d obtain 95% of the time using the sample mean as the estimate of the population mean. This means that if you repeated experiment with same size sample from same population, 95% percent of the time the mean you get from that sample will be within the 95% Confidence Interval that you generated from the current sample. Given the same data, a 95% CI will be narrower than a 99% CI.

General CI formula= mean +/- (Zvalue for the DesiredConfidence[.95]/2)(SE)

Standard Error of the Mean(SEM or SE) = sampleStandardDeviation/sqrt(sampleSize)

Concept: ME = (T/Z-value)(SE)

Concept: CI = sampleMean +/- MarginofError (ME)

**Hypothesis Test**:

Hypothesis Test(pvalue): Compares distance from HypothesizedMean to sampleMean.

CI Test: Compares distance from sampleMean to HypothesizedMean

For H-T, the distance we care about is the length of the line from the HypoMean to the SigLevel. If the sampMean is further away from the HypoMean than the SigLevel level; Reject Null. For CI-T, the dist we care about is length of line from sampMean to upper/lower bound, if HypoMean is not in that interval; Reject Null. Dist from Hypo-->sigLevel == Dist from sampMean → upper/lowerCIboundary.

The goal of inferential statistics is to discover some property or general pattern about a large group by studying a smaller group of people in the hopes that the results will generalize to the larger group.

The question becomes, is there a difference between 2 populations & if there is, is the difference due to random chance? To answer these questions, we use a Hypothesis Test (of some form). A hypothesis test allows us quantify the probability that our sample mean is unusual/unlikely, given that the Null Hypothesis is true.

We make two hypotheses before we start our study, the null hypothesis (samples come from same population), and the alternative hypothesis (samples come from different populations).

At a certain basic level, all inferential statistics procedures are the same: they seek

to determine if the observed (sample) characteristics are sufficiently deviant from

the Null Hypothesis to justify rejecting it. At the end of the day, based on the data analysis (evidence) you can either Reject the Null or not Reject the Null. Notice the language, this is different from how most people use stats. These are tests of falsifiability, the only assertions you make are about what is false (w/in a certain degree of error), never what is true.

Null H and Alt H

1. H0: The Null H. The status quo. The thing being introduced has NO effect. The drug doesn’t work. Mean response w/the drug is same as mean response w/o the drug. Everyone likes all things equally. Men& women do/achieve this thing equally
2. H1: Alternative H. Status quo has changed. Thing being introduced HAS effect. The mean response w/the drug is DIFFERENT than mean response w/o the drug. People like coke more than Pepsi. There’s a difference between men&women here.

I see a difference between my sample mean and the expected mean (or between the means of my two samples), how could this just be due to Random Chance?

Sampling error is the difference between a sample and the entire population.Thanks to sampling error, it’s entirely possible that we could sample men’s height in the US and get a sample mean of 1.7m, but that the population mean could still be our Null H of 2.0m (previous mean we want to check to see if it’s changed). Or, to put it another way, if we repeated the experiment, it’s possible that the second sample mean could be close to 2.0. A hypothesis test helps assess the likelihood of this possibility! A hypothesis test allows us quantify the probability that our sample mean is unusual/unlikely, given that the Null Hypothesis is true.

Judges, Juries, and Errors.

With H-testing, like w/ a jury, there are 2 types of error you can make. The Null H is the man on trial.

**Type I Error:** Rejecting a True Null. Convicting an innocent man. Concluding drug was effective when it really wasn’t

**Type II Error:** Failing to reject a False Null. Letting a guilty man go free. Concluding drug was not effective when it was.

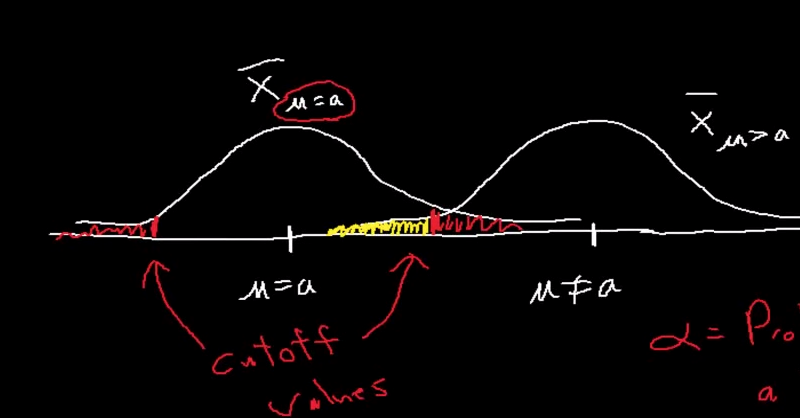
**The Power** of a Hypothesis Test is the ability of the test to correctly reject a False Null Hypothesis (of NOT commiting a Type2 Error).

The significance level, also denoted as alpha or α, is the probability of rejecting the null hypothesis when it is true (Type1 Error). For example, a significance level of 0.05 indicates a 5% risk of concluding that a difference exists when there is no actual difference.

**alpha** = pr(Type1 error) = 1- signficance level = pr(rejecting a true null H)

**beta** = pr(Type2 error) = pr(Failing to reject False null H). Beta is the region where the distribution from the Alt H overlaps with the distribution of the Null H, that is closer to the Null mean than the signficance level. A value in this region could come from either distribution. If you only considered the Null H, taking a value from this region would not provide evidence to reject the Null H. However, since/if the Alt H distribution reaches into this space it is possible that the value you picked actually belonged to the Alt H. Hence, based on that value, you would choose not to reject the Null H, even though the value came from the Alt H. So beta depends on alpha and it also requires a sampleMean which requires a sample Size.

<https://www.youtube.com/watch?v=eoYYFVVcMlM>



Legend of image: Null H distribution on Left. Alt H distribution on right. Red shows significance and type1 error, yellow shows beta and type2 error.

So power, beta, alpha, & n are linked. If you know 3 of them, you can calculate the 4th

Why do we care about power?

Power analysis can be used to calculate the minimum sample size required so that one can be reasonably likely to detect an effect of a given size. Power analysis can also be used to calculate the minimum effect size that is likely to be detected in a study using a given sample size. In addition, the concept of power is used to make comparisons between different statistical testing procedures: for example, between a parametric and a nonparametric test of the same hypothesis.

The power of a hypothesis test is affected by three factors.

1. Sample size (n). Other things being equal, the greater the sample size, the greater the power of the test.
2. Significance level (α). The higher the significance level, the higher the power of the test. If you increase the significance level, you reduce the region of acceptance. As a result, you are more likely to reject the null hypothesis. This means you are less likely to accept the null hypothesis when it is false; i.e., less likely to make a Type II error. Hence, the power of the test is increased.
3. Effect Size. the greater the effect size, the greater the power of the test.

The "alternative" value of the parameter being tested. The greater the difference between the "alternative" value of a parameter and the value specified in the null hypothesis, the greater the power of the test.

An Alternative value is just an actual value for the Alternative Hypothesis. Could be set before or after the experiment?

Effect size = Alt value - Hypothesized(Null H)value

Interpretation of Power

power (sometimes referred to as π). π=0.80 is a common standard for adequacy. This convention implies a 4:1 trade off between β-risk and α-risk. (β is the probability of a Type II error; α is the probability of a Type I error, 0.2 and 0.05 are conventional values for β and α).

However, there will be times when this 4-to-1 weighting is inappropriate. In medicine, for example, tests are often designed in such a way that no false negatives (Type II errors) will be produced. But this inevitably raises the risk of obtaining a false positive (a Type I error). The rationale is that it is better to tell a healthy patient "we may have found something - let's test further", than to tell a diseased patient "all is well".[1]

POWER IN R

n = sample size

delta = Effect Size = difference between sampleMean and HypothesizedMean

\*\*\*Make sure delta is positive, negative delta destroys test. You can just switch order of sM and hM

sd = standard deviation

alternative = one sided or two sided hypothesis test

type = c("two.sample", "one.sample", "paired")

sig.level=significance level (Type I error probability) (default = .05)

Find the power if you know n (assuming samMean was .01, and H-mean was 0)

power.t.test(n = 100, delta = .01-0, sd = .04, type = "one.sample", alt = "one.sided")$power

Find the n necessary to produce a particular power

power.t.test(power = .9, delta = .01-0, sd = .04, type = "one.sample", alt = "one.sided")$n

A NOTE ON SAMPLING

Law of Large Numbers(LLN)

The sample mean will approach the population mean, which is the same thing as the expected value E(X), as the number of samples/trials (n) goes to infinity.

Central Limit Theorem

<https://www.youtube.com/watch?v=JNm3M9cqWyc&list=PL4C863861E3B2E380&index=25>

You can start with ANY distribution (could be continuous or discrete) that has a well defined mean and variance and end up w/a normal distribution if you:

Take a sample size(n) from the distribution(population) and compute the mean of that sample. If you do repeated sampling (eg 10,000 samples) of your population (drawing n values each time) and you plot the frequency that each sample mean appears, you will end up w/a normal distribution.

Regardless of the size of the sample(n) that is drawn each time, the mean of the resulting frequency plot will always be the mean of the original distribution (population).

If you increase n (ie instead of drawing 4 at a time, you draw 20 at time) the standard deviation of the resulting frequency plot will get tighter

**Types of Hypothesis Tests:**

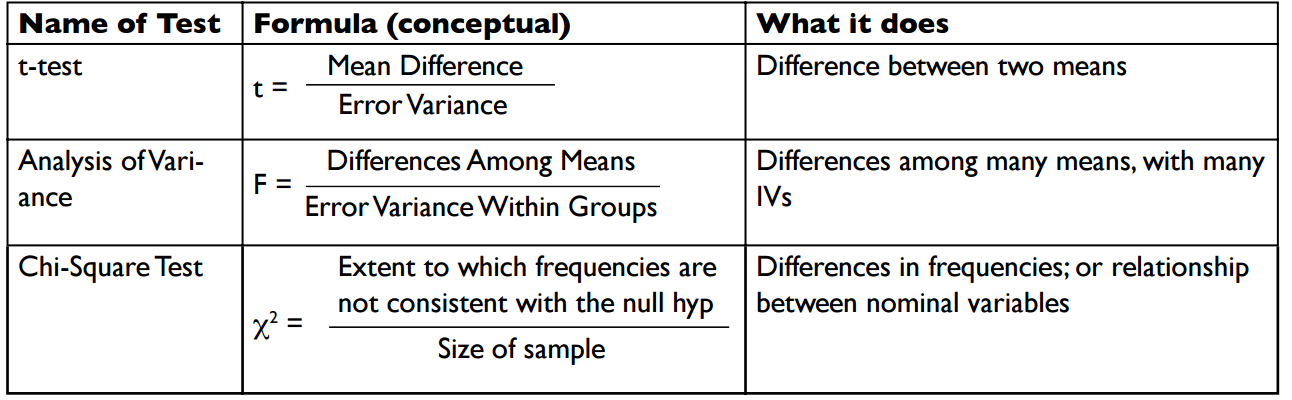
What Type of stat analysis should I use, chart?: <http://www.ats.ucla.edu/stat/mult_pkg/whatstat/>

What Type of stat analysis should I use, Descriptions: <http://www.ats.ucla.edu/stat/spss/whatstat/whatstat.htm>

Simple explainations w/example of different types of t-tests (1sample,2sample,paired) in R:

<http://www.stat.columbia.edu/~martin/W2024/R2.pdf>

Doing t-tests in R:<http://www.cookbook-r.com/Statistical_analysis/t-test/>



T-Test

<http://blog.minitab.com/blog/statistics-and-quality-data-analysis/what-is-a-t-test-and-why-is-it-like-telling-a-kid-to-clean-up-that-mess-in-the-kitchen>

The t-value measures the size of the difference between the means relative to the variation in your sample data. Put another way, T is simply the calculated difference represented in units of standard error. The greater the magnitude of T (it can be either positive or negative), the greater the evidence against the null hypothesis. The closer T is to 0, the more likely there isn't a significant difference.

As the above formula shows, the t-value simply compares the strength of the signal (the difference) to the amount of noise (the variation) in the data.

If the signal is weak relative to the noise, the (absolute) size of the t-value will be smaller. So the difference is not likely to be statistically significant. However, if the signal is strong relative to the noise, the (absolute) size of the t-value will be larger.

**If your t-test results don't achieve statistical significance, it could be for any of the following reasons:**

* The difference (signal) isn't large enough. Nothing you can do about that, assuming that your study is properly designed and you've collected a representative sample. This is meaningful insignificance.
* The variation (noise) is too great. This is why it's important to remove or account for extraneous sources of variation when you plan your analysis. For example, you could use a control chart to identify and eliminate sources of special-cause variation from your process before you collect data for a t-test on the process mean.
* The sample is too small. Remember the effect of variation is lessened by sample size. That means for a given difference and a given amount of variation, a larger sample is more likely to achieve statistical significance. This effect also explains why an extremely large sample can produce statistically significant results even when a difference is very small and has no practical consequence

T-TESTS in R

Doing t-tests in R:<http://www.cookbook-r.com/Statistical_analysis/t-test/>

Doing t.tests w/o only mean, sd,n

The tsum.test() is for doing a t-test using only the SUMMARY data. tSUMMARY.test :)

library(BSDA)

#2-sample test

tsum.test(mean.x=.1, s.x=.01, n.x=5, mean.y=.136, s.y=.02, n.y=7)

#1-sample test

tsum.test(mean.x = 1100,s.x=30, n.x=9)

BINOMIAL AND POISSON TESTS:

Doing binomial tests: For when people fall into 1 of 2 categories (like something, don’t like it : have a trait, don’t have it, etc)

binom.test() : Performs an exact test of a simple null hypothesis about the probability of success in a Bernoulli experiment.

#3 out of 4 people chose coke over pepsi. Do people like coke more(coke greater?)

binom.test(3,4,alternative = "greater")

Poisson tests: For binomial tests w/large n. OR for comparing against a known/acceptable/target RATE

x: number successes

T:number trials

r: hypothesized rate. This could be a standard or benchmark. IE 1/10 failures is acceptable.

alt: two-sided(from r), greater(is sample rate >than r), less(than r)

poisson.test(x=10,T=1787,r=1/100,alternative="less")

**CHI-SQUARE TEST**

Types of Data: There are basically two types of random variables and they yield two types of data: numerical and categorical. A chi square(x^2) statistic is used to investigate whether distributions of categorical variables differ from one another.

\*There are 2 types chi-square tests:

* Goodness of Fit (are observed values significantly different from expected values)
* Test of Independence (Are 2 observed variables independent of each other. Independence is the Null H, the Alt H is that the 2 variables are related/dependent)
* NOTE, THESE ARE THE SAME ACTUAL TEST. They’re only different conceptually. In both cases, you’re comparing variables from 2 (or more columns). When testing for goodness of fit, you’re columns are [expected, observed] and you’ll have 1 row for each variable (white, asian, etc).

A chi-square goodness of fit test allows us to test whether the observed proportions for a categorical variable differ from hypothesized proportions. For example, let's suppose that we believe that the general population consists of 10% Hispanic, 10% Asian, 10% African American and 70% White folks. We want to test whether the observed proportions from our sample differ significantly from these hypothesized proportions.

Independence example:

let's see if there is a relationship between the type of school attended (schtyp, public or private) and students' gender (female). (The Null would be that females would equally choose public and private)

Basically, you get a chi-square-VALUE, which plays a similar role to a t-value, and can be used w/degrees of freedom. Chi-square df = (#columns -1)(#rows-1)

Tutorial on using chi.sq() tests in R:<http://courses.statistics.com/software/R/Rchisq.htm>

Here’s another tutorial w/practice problems: <http://ww2.coastal.edu/kingw/statistics/R-tutorials/independ.html>

Some great examples of laying out your tables & calc chi-square by hand:

<http://math.hws.edu/javamath/ryan/ChiSquare.html>

**MULTIPLE COMPARISONS: ANOVA (Analysis of Variance)**

ANOVA is used to compare the means of more than two samples.

When we have only two samples we can use the t-test to compare the means of the samples but it might become unreliable in case of more than two samples. If we only compare two means, then the t-test (independent samples) will give the same results as the ANOVA.

You get an F-value (F-statistic) instead of a T-value

Awesome overview of use of ANOVA in R including how to VISUALIZE the results:<http://www.statmethods.net/stats/anova.html>

I wrote a script walking through exactly how to do this. anova.HowTo.R in ~/datascience.