

What Statistic and Why?

Raymond R. Balise

Stanford University

Department of Health Research and Policy



SPCTRM



Stanford / Packard Center for
Translational Research in Medicine

- “Nobody ever learned statistics by reading a book about it.”

from M.J. Crawley’s statistics book

Statistical Computing: An Introduction to
Data Analysis using S-Plus

- “The hardest part of any statistical work is getting started. And one of the hardest things about getting started is choosing the right kind of statistical analysis for your data and the particular question you are trying to answer.”

The BIG Questions

- What is your response variable?
 - Is it a continuous scale, a count, a proportion, a time until an event or a category?
- What are the explanatory variables?
 - Are the explanatory variables continuous, categorical or both?
 - How many subjects do you have in each group?

A Huge Flowchart

- Once you have thought about the nature (nominal, ordinal, interval) of your variables you would want a giant flowchart to help you find the statistic that is appropriate for your task.
- The best one I have found is *Selecting Statistical Techniques for Social Science Data: A Guide For SAS Users* by Andrews, et al..
 - It is superb in that it gets you to an analysis technique and tells you the procedure to do it in SAS but it is a bit old and out of date and is missing survival analysis.

KISS

- Keep it simple stupid!
- Once you know your outcome and predictors draw a picture of what you hope to see.
 - Find a cocktail napkin or an envelope and doodle. Don't waste time with Excel or a good graphics package.
 - Come to a full and complete STOP if you can't draw a picture of the effect you expect to see.
- Your entire goal with data analysis should be to reproduce that graphic using real data.

Thoughts on Plots

- Plotting your data:
 - The response variable goes on the y axis (the ordinate).
 - The explanatory variables go on the x axis (the abscissa).
- A good plot shows you two things:
 - a point estimate of the value you expect to see and
 - a representation of the variability you expect to see.

Descriptive Plots **Before** Data

- Hand-drawn histograms, box-plots or density plots
 - Put major thought into the shape of the histogram you expect to see. Should the distribution of the outcome variables be: bell curve shaped, many scores piled up at 0 with a few big values, positive counts only?
- Hand drawn scatter plots
 - Put major thought into what happens to the variability in the outcome as your predictors change.
 - Many of the classic methods of analysis have problems if the variability changes with the outcome.

Inference

- Things change.
- With a biological system, if you measure the same thing twice, you will not get *exactly* the same score.
- In real life you expect to encounter variability but once in a while you encounter things that are so unusual that your mental model of the world changes.
- The much talked about concept of a P-value is an attempt to quantify the idea of how unusual a sample is given a basic theory about the world.
 - More on that in a bit...

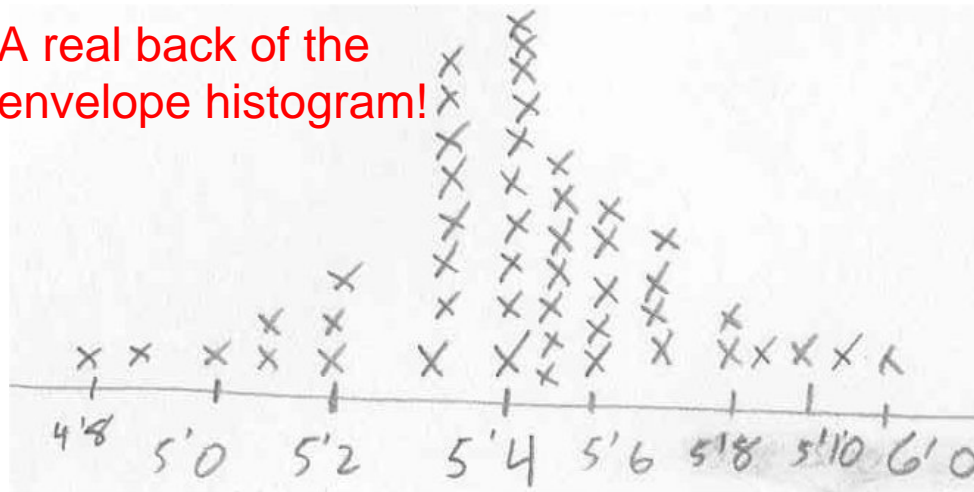
HIV and P-values

- I have a theory that the chances of acquiring the HIV virus are approximately 3.1% per event.
 - Probability of female-to-male transmission of HIV-1 in Thailand. Lancet. 1994 Jan 22;343(8891):204-7.
- If this rate is correct and someone seronegative who experienced 5 events acquires the virus, it would be unusual. Could it happen? Of course it could happen but it would be unexpected. A p-value is nothing more than an attempt to describe how unlikely it would be to see someone become seropositive given the original idea of the infection rate.

Empirical Distribution

- Sample size of 1 for height
 - Hand draw a histogram with dots or xs for each sample...

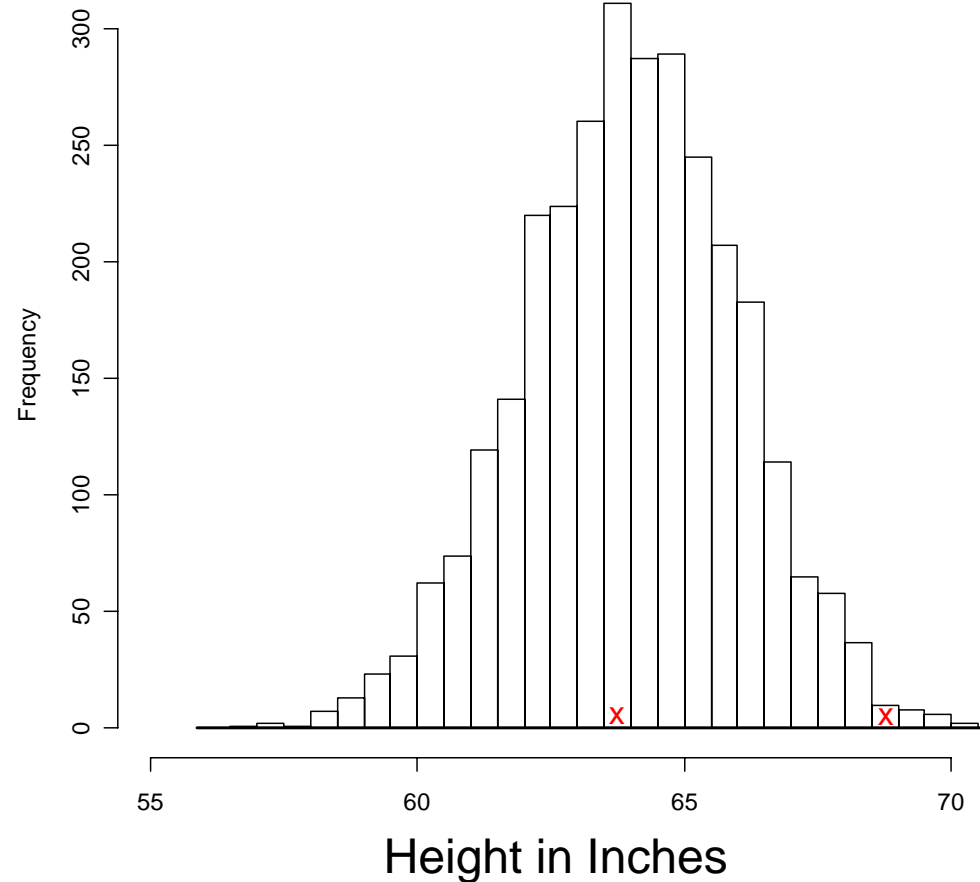
A real back of the envelope histogram!



- What would be weird is to get an extremely big or small value.
- If the true average female height at Stanford is 5'4", I could get a sample of 5'9".

Theoretical Density

- If I were to get a very large sample with very fine measurements, the x's would blur together and we would see some very short and very tall people but most would be piled up around the hypothetical mean of 5'4".



Sampling

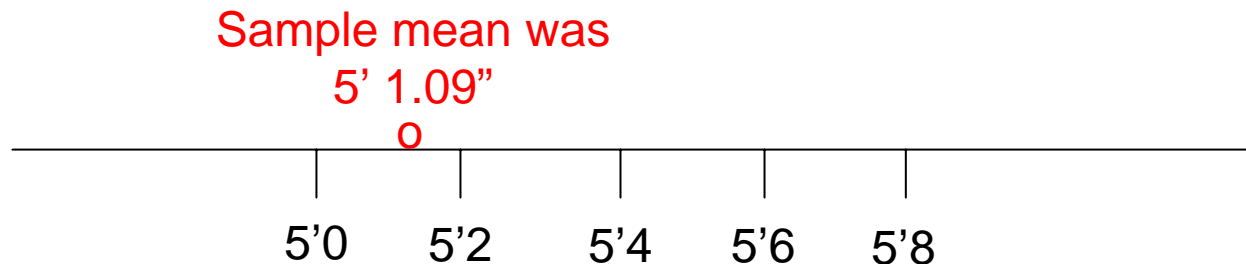
- You are not likely to be able to get such a large sample. Rather, you will collect a sample, get the mean of the sample and look to see if it is very different from the expected value.
- The same idea comes into play if the observed mean is “too weird”, causing you to reject the hypothesis of a mean of 5'4". What constitutes “too weird” depends on the distribution of samples.

Bigger Samples....

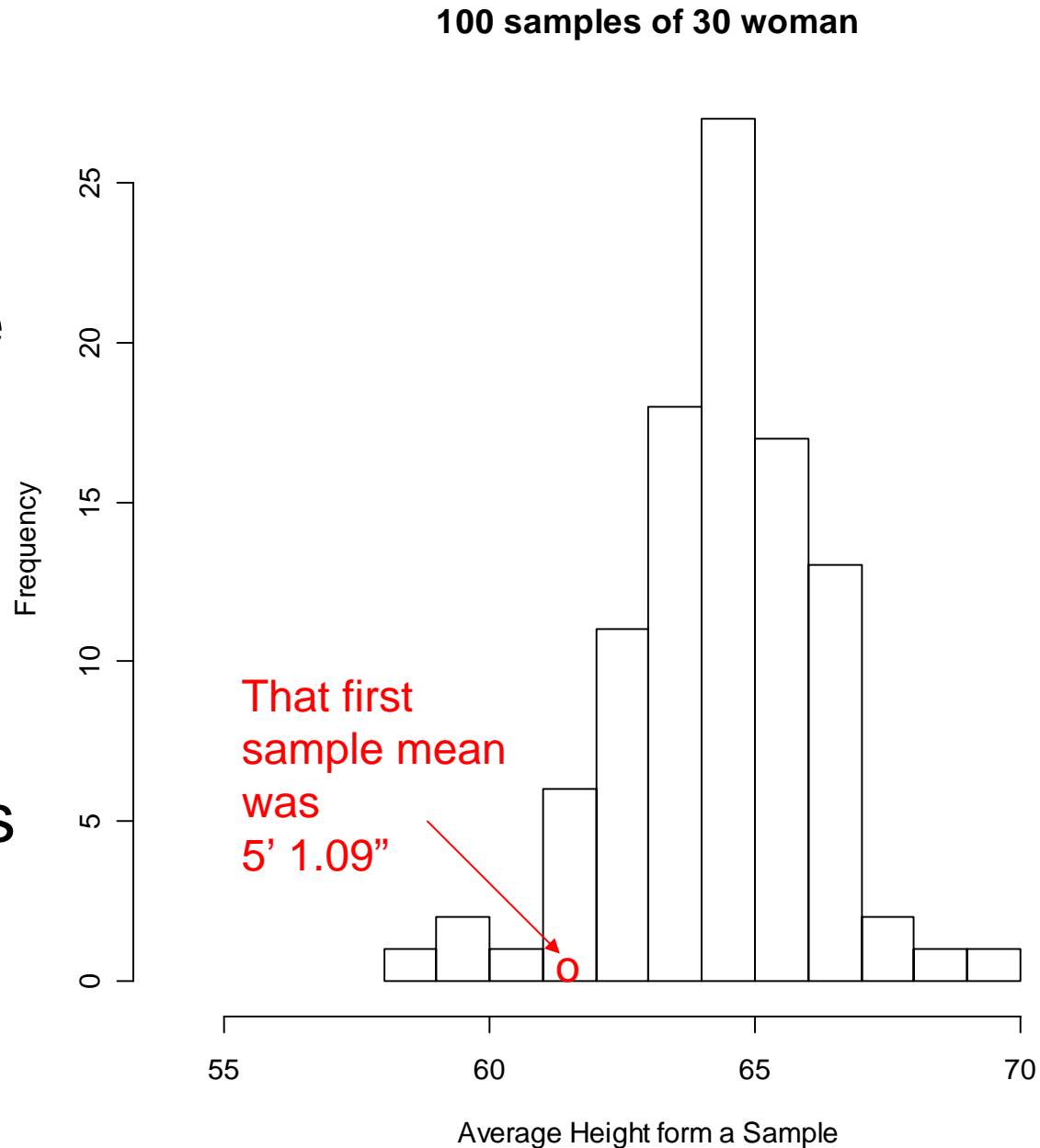
- If I use a sample size of 30 women, I don't expect to get a lot of folks who are 5'9". If I randomly pull 30 women, I expect most of the sample to be clustered around 5'4." So a typical sample mean will be somewhere around 5'4". Could I get a mean of 30 women that is 5'11"? Yes, it would be bizarre but I could. The concept of a p-value is a measure of just how bizarre it would be.

Theoretical Distribution N of 30

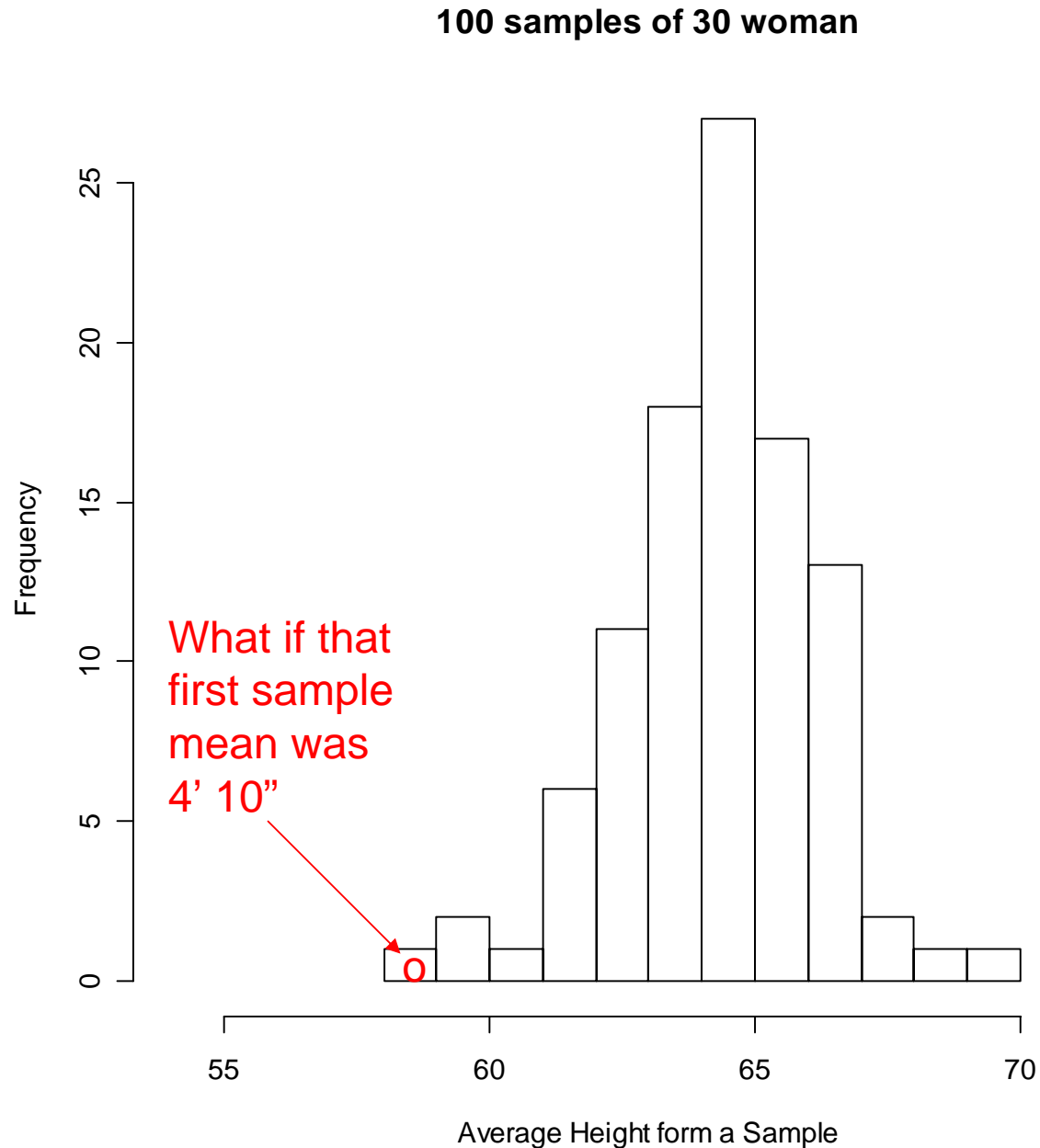
- Instead of plotting individual scores, I can plot the mean value for the sample. Is the value compatible with the original hypothesis?



- Just like you could compare the sample of size 1 with the theoretical distribution, you can compare the mean vs. the theoretical distribution of means for sample size 30.



- If you expected the data to look something like this and you saw a value way off to the side, you would think that your original model of the world was wrong.
- You would reject the null hypothesis of the mean being 5' 4".

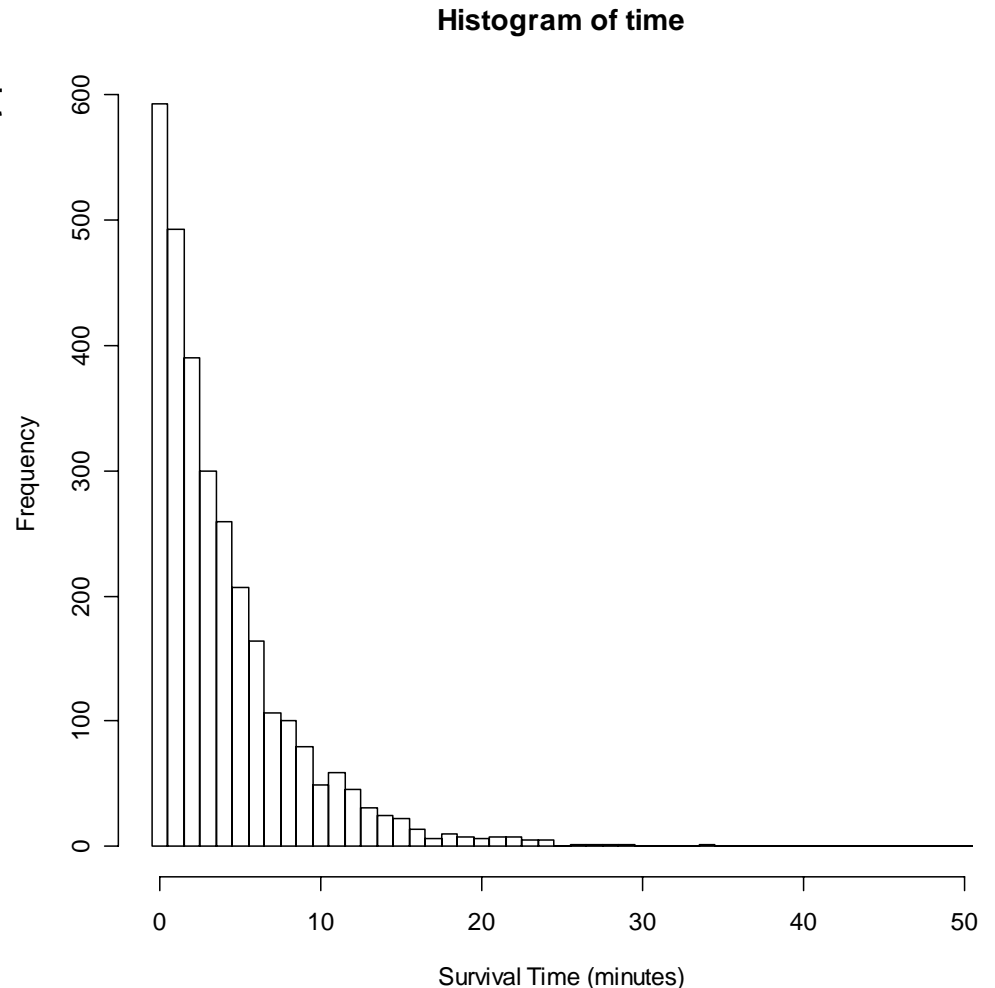


One-sample T-test

- Testing a sample against a hypothetical mean is usually done with a one-sample t-test. In the past, you would standardize scores (subtract values from the mean and divide by the standard deviation) and look up values in tables to see how incompatible your observed data was with the original (null hypothesis). Happily the math for that “unlikelihood” value is stored away in functions in the major analysis packages.

Expected Value for Not Normal

- At 11:33 AM 6/11/2007, aMadScientist wrote:
“I think I'd make an excellent Brain Surgeon! Hehehe No really...”
- I built a brain transplantation simulator... What is the expected time of death for someone undergoing a brain transplant?
- I think it would look nothing like normal and more like an exponential or a negative binomial function, like the histogram to the right.



Sample of 30 with Brain Transplants

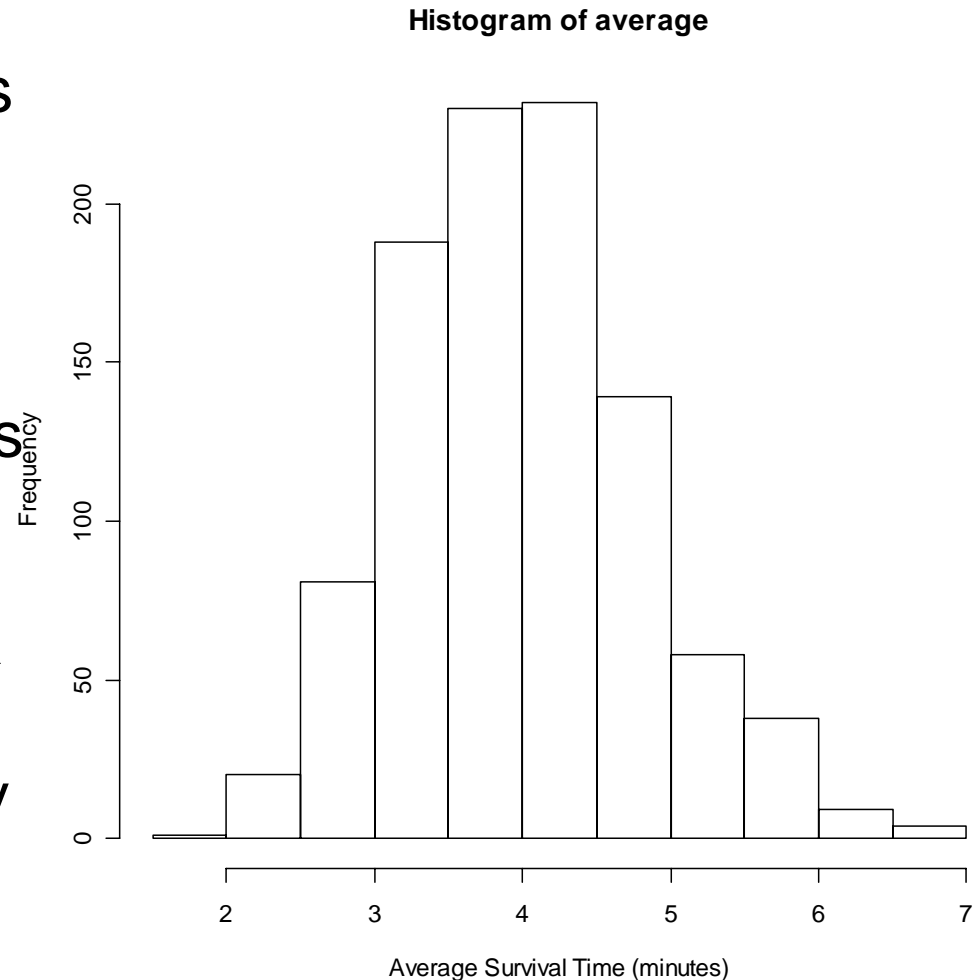
- What is the expected value for 30 people undergoing a brain transplant?
- You don't expect to get many people living a long time, so to have a sample with 30 people show up with a mean survival of an hour, it is exceptionally odd.
- You need to know what is the distribution of the average survival for a negative binomial distribution with sample size of 30.
- Happily there are functions already written to describe something this weird.

Average Survival Time...

- Given my model of the world, the average survival time of 30 people is going to be around 5 minutes. Could I get an average that is an hour? Yes, but it would be really unexpected! What is truly weird is the shape of distribution of the means.
- It looks...

...normal

- You can use the same logic to make inferences from lots of different distributions (if your sample size is big enough). Things break down if the distribution is very skewed and in this case if the data can not be well described with a mean.
- This pattern of normality in the sample means is sometimes called the central limit theorem...



Cute Little Thingy

- The Central Limit Theorem (CLT) is such a big deal because it allows you to make inferences about what is weird (unexpected value) given a null hypothesis even when your continuous outcome is not normally distributed.

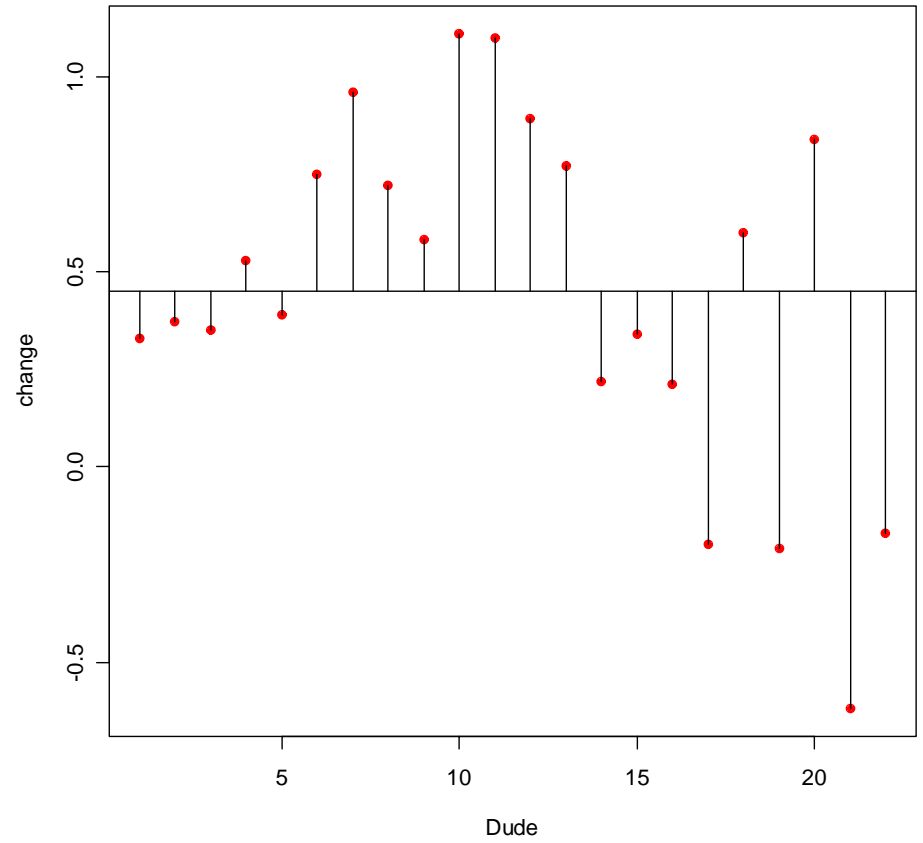
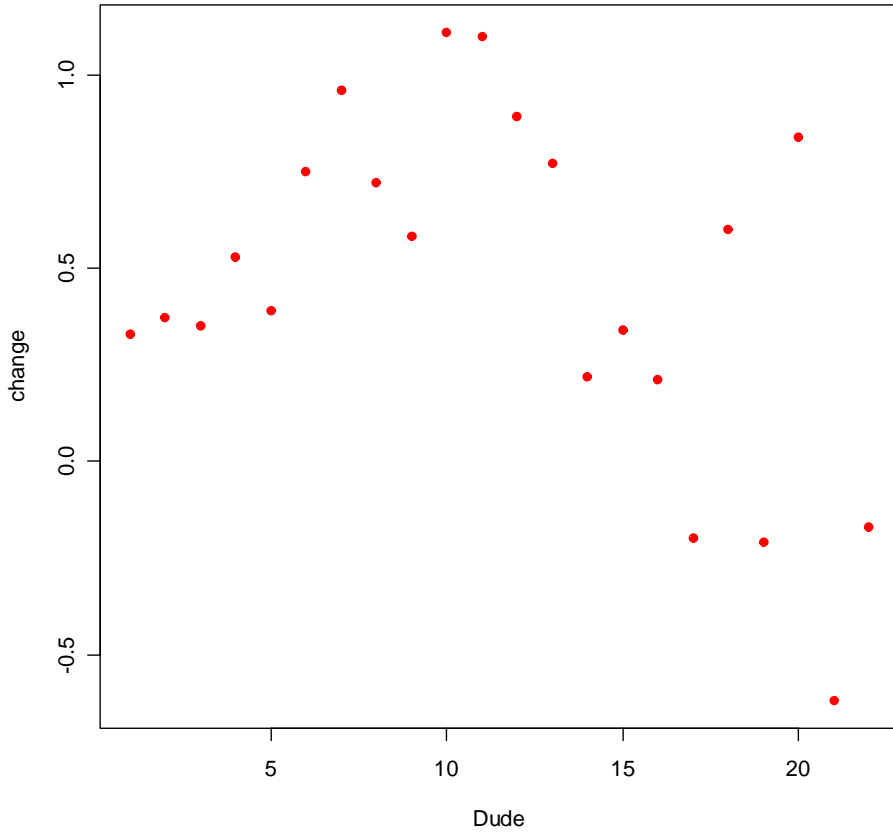
Classic Tests of Means

- If you took a one quarter class in statistics you will have heard of t-tests, ANOVA and (ordinary least squares) regression.
 - These tests attempt to quantify the relationship between a continuous outcome and a two level categorical predictor (t-test), a multilevel categorical predictor (ANOVA) or continuous predictor (regression).
- The height example began by saying I think women are 5'4" (the null hypothesis) and tested to see if the data was incompatible with this idea. These techniques begin by saying there is no relationship between the continuous outcome and the predictors and you then test to see if the data are incompatible with the null hypothesis.

ANOVA and T-tests

- A smart person who we will call “I” thinks that a new drug can treat neuropathic pain. Contrast with this with the idea that there is no clinically important effect of the drug. That is, there is no difference in pain reduction between drug and placebo. The *nothing going on* (no difference between drug and placebo) concept is the *null hypothesis*. If there is no difference between the two groups, you could use a common mean to describe both. Of course there would be variability (which a statistician would call error, error variance or noise) between each person and the mean.

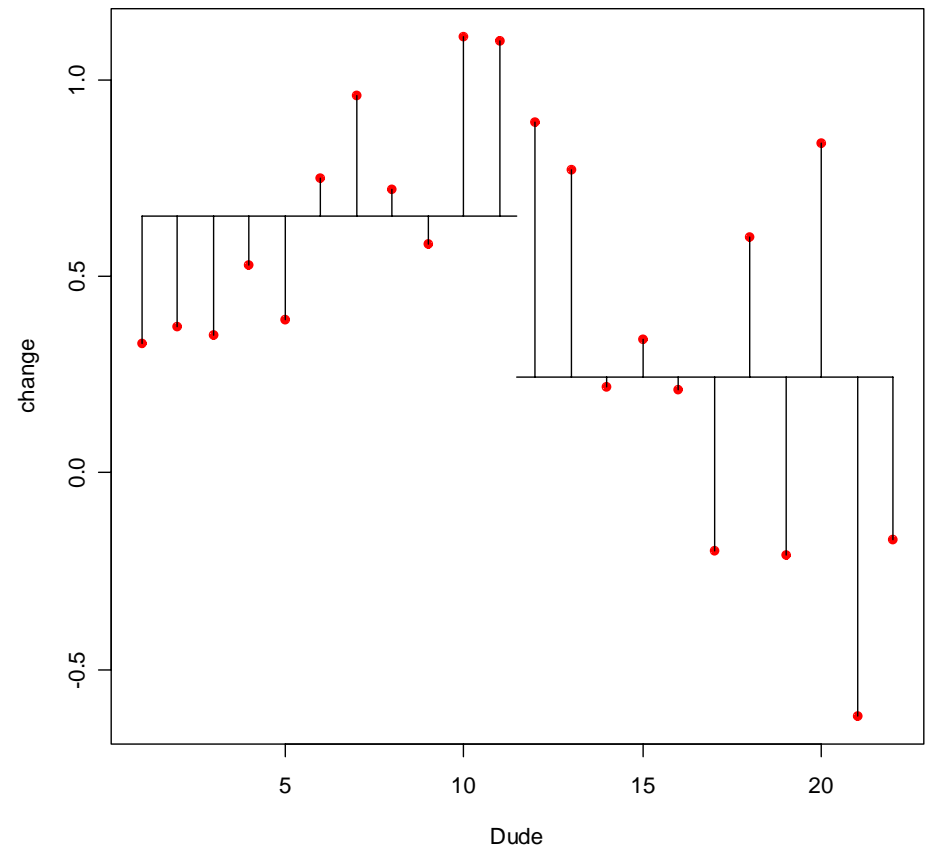
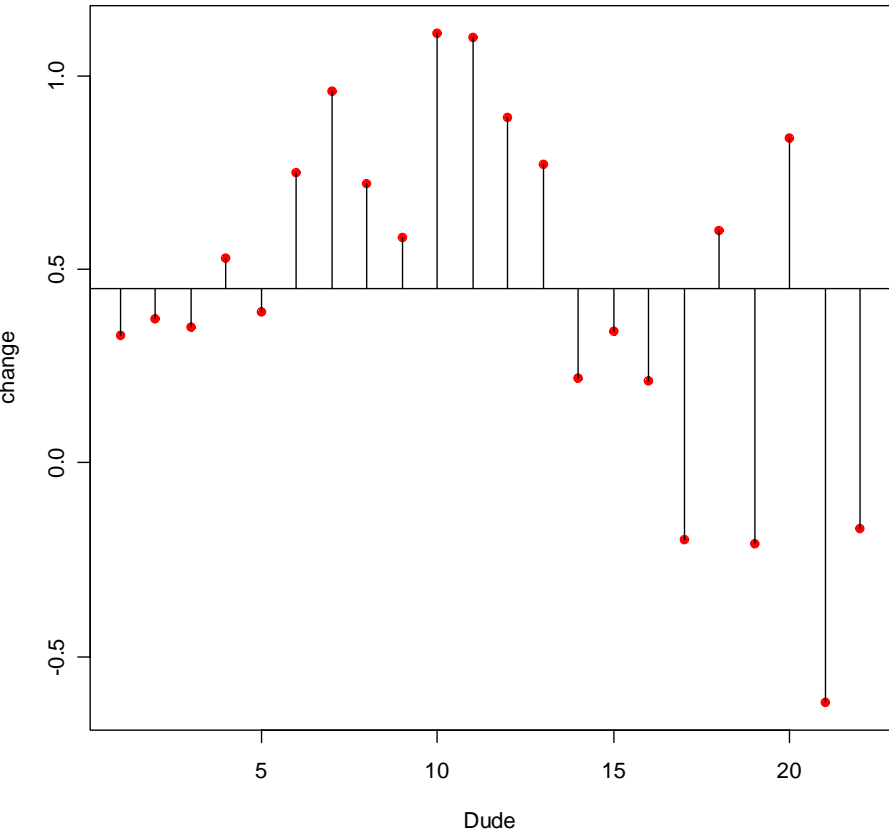
Error in Using a Single Mean



Two Groups

- If there were differences in pain between the two groups you would want to use two different means (one for the placebo and one for the drug) to describe the data. There would be errors between the drugged mean and each drugged person and there would be errors between the undrugged mean and each undrugged person. If there is a significant improvement in using two means instead of one, you should see it as different means and shorter error whiskers.

Error Reduction Using 2 Means



Errors Using Two Means

- The ratio of the overall mean error and the errors when you use the different means is what goes into the famous (t or F) statistics in t-tests or ANOVAs. One way to describe the length of the those lines is to use variance. Variance is (sort of) the average length of the whiskers between the mean line and the data points. Therefore the analysis is called **AN**alysis **O**f **V**ariance.

Ratio of Errors

- This example was done with two groups but behind the scenes an ANOVA is comparing the amount of error variance from using a single mean vs. the amount of variance using different means for each subgroup and comparing these ratios to an F distribution to quantify the weirdness associated with seeing an effect this big by chance alone.

P Value for this Example

- Could it be that the amount of pain is the same between the two groups but you just happen to get more reduced pain in the drug group? Of course, the p value gives a guestimate of how likely this is. Probability $< .03$ in this case.

Core assumptions for t-test and ANOVA

- You saw graphically that variance matters A LOT for these analyses. If you measure the same person repeatedly (or otherwise use related samples), your variance comparisons will be systematically biased.
- Therefore, one of the core assumptions for comparing means is that you have independent samples.
- Another is that the errors are normally distributed (a plot the length of the whiskers as a histogram).

Is that all?

- The tests I have talked about were to compare means. You have seen that with the CLT you can made inferences even when the data are not normally distributed.
 - The CLT works reasonably well for large samples. “What is a large sample” is a great deal of work to compute. What happens when you have **smaller samples** ($N < 30$)? How do you judge what is weird when the distribution of the sample statistics is totally unknown?
 - What about if you don't want to compare means?

Other Measures of Central Tendency

- There are some distributions that are not well described by a mean. Plenty of peptides are measured as zero then with a smattering of high values and some outlandishly large values.
 - You may want to use a median to compare the groups.
 - If you have a single smallish sample of non-normally distributed data (with independent samples) to be compared vs. a population value you can use Wilcoxon's signed rank test.
 - If you have two smallish samples of non-normally distributed data (with independent samples) and you want to compare the central tendencies, you can use Wilcoxon's rank sum test.

Non-parametric

- In cases where you do not want to trust the CLT you can fall back on non-parametric tests.
- These are also known as rank statistics.
- In the case of comparing two groups you sort the data from lowest to highest and you then look to see if the lowest score is in group A or B, if the 2nd lowest is in group A or B etc. You then do statistics on the ranks. For example A was ranked the 1st 2nd 4th lowest scores and group B was 3rd 5th and 6th lowest scores so compare the means of (1, 2, 4) vs (3, 5, 6). There are many variants on these rankings for dealing with things like ties but the idea is to use the rank orders.

Paired Data

- Sometimes you have paired data and you want to know if the averages are the same. For example you may have right eye and left eye acuity and you want to see if the means are the same.
- Acknowledging that the data are paired increases your chances of seeing difference if it is truly there.
- This is typically done with a paired t-test.

Ordinary Least Squares Regression

- In theory, mad scientists should be happier if people were giving them more money so they have the best technology at their disposal. Alternatively some of them would be so completely out of their minds that they did not care about the amount of money being spent on their “science”.
- There are two hypotheses here. One is that there is no meaningful relationship between money and happiness. The other is that money increases (or decreases) happiness in mad scientists.

Regression is Like ANOVA

- You can think about ordinary least squares regression just like ANOVA. You do a scatter plot of the data and put a single horizontal line through the data and measure the distance between the points in the scatter plot that flat line. Then put an optimal line through the data and measure the amount of variability.

What is an optimal line?

- Maximum likelihood... is a VERY big deal. Your data are the truth (after you remove the pregnant men). You can fit any number of different models (lines) to describe the data. You do **not** fit data to a model. The semantics are very important.
- You want to fit a line (model) that makes the data as likely as possible.

Maximum Likelihood Plots for Scatter Plots

- If you have only 2 variables it is easy to visualize the relationship with a scatter plot. Draw the outcome on the y-axis and the predictor on the x-axis and then draw a line that gets as close as possible to the data. Measure the total difference between the line and the points (using a ruler going up and down) for each and every point. Scoot the line and repeat and get the total amount of difference. Scoot the line to a new spot and repeat. With a simple scatter plot you don't have to do the iterative re-measuring. In other cases you do.
- If you were to plot the amount of total differences at some point it would be minimized. That would be the maximum likelihood estimate for the model and it would tell you the optimal line.

Closed Form Solutions

- In some cases your software can jump to the maximum likely solution using some simple simultaneous equations with algebra and/or calculus. In other cases you have to iterate to find the solution and it may not even find a solution.

Beyond the Simple Stuff

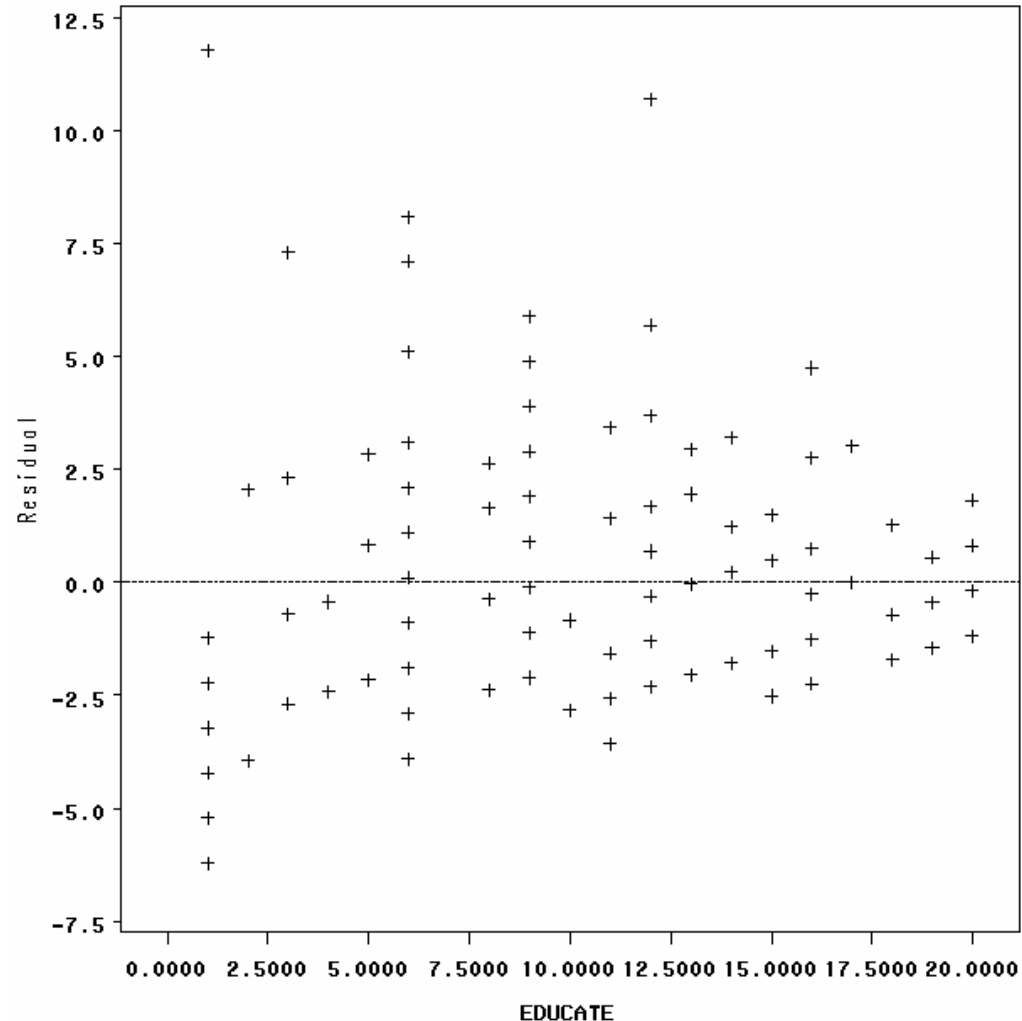
- To this point I have talked about models where the outcome is continuous variable where the range is not bound at (a lower or) an upper limit. There are other types of analyses where this would be unreasonable.
- What about a binary 1 = yes 0 = no outcome for having a disease?
- What about the count of zoonotic mites cases in a geographic region?

Limiting the Outcome Variable

- The math gets ugly quickly but you have a couple options for doing things like preventing negative counts. You could, for example, take a log of the counts and model that.
- There are other important issues with variability with count data. In theory you want to have similar amounts of variability in your outcome regardless of if your predictor is high or low (aka homoscedasticity).

Heteroscedasticity

- Usually you see the variability in the outcome increase with the predictor but not always.
- Heteroscedasticity is bad.



GLM

- There is a class of models called Generalized Linear Models that allow you to predict outcomes where you know there are restrictions on the possible values. For example, if you need to predict non-zero counts or a probability of being diseased (between 0 and 1).
- GLMs allow you to make predictions that are limited like this and have variances that behave well.
- I will say more about this in an upcoming talk called “Predicting the Future” but you should know the most common GLMs.

Common flavors of GLMs

- Ordinary least squares
- Logistic regression
- Poisson regression

Categorical Data Analysis

- I have discussed most of the classic analysis techniques but I didn't talk about contingency tables (categorical outcomes with categorical predictors).
- There are tests to look for an association between variables,
 - Chi-square tests
- ... and there are measures of the strength of the association between variables.
 - Relative risks
 - Odds Ratios

The Chi-Square Test

- The classic chi-square for contingency tables figures out what is an unexpected pattern given the marginal and the observed values. In English... do some of the cells have more observations than expected?

| | Yes Happy | No Happy | |
|------------|-----------|----------|-----|
| Yes Cookie | 50 | 0 | 50 |
| No Cookie | 0 | 50 | 50 |
| | 50 | 50 | 100 |

Small Expected Frequency

- In this case we expected 25 people in each cell if there was no association between cookies and happiness. You need to be careful if the expected frequency is small (say less than 5). When this happens, you need to use a different statistic (a Fisher's exact test) because the shape of the theoretical distribution no longer matches reality.

Chi-square for Association

- The P value you get from a chi-square just tells you if there are more people (things) in a cell than expected under the null hypothesis of no association. Who cares?! It is only a measure of association.
- You usually are more interested in specific patterns. You want to know things like whether the odds of getting a disease are related to exposure to a risk factor.

Odds Ratios for Strength of the Relation

- You really care about things like the odds of being happy if exposed to cookies vs. the odds of being happy if unexposed to cookies. In this case the odds are $(48/2)/(2/48)$. That is, the odds of being happy are more than 500x greater if exposed to cookies.

| | Yes Happy | No Happy | |
|------------|-----------|----------|-----|
| Yes Cookie | 48 | 2 | 50 |
| No Cookie | 2 | 48 | 50 |
| | 50 | 50 | 100 |

Summary - With Large Samples

- Outcome continuous
 - Predictor Continuous (Regression)
 - Predictor Categorical (t-test, ANOVA)
 - Predictors Continuous and Categorical (ANCOVA)
- Outcome Categorical
 - Predictor Categorical (Chi-square Odds Ratios)
 - Predictor Continuous (Logistic Regression)
- Be careful of repeated measures.

Other Important Ones...

- Small sample (not normal) outcome is continuous with a categorical predictor (Wilcoxon, Kruskal-Wallis)
- Comparing two proportions (Binomial test)
- Comparing two paired proportions (McNemar's test)
- Groups of contingency tables (Cochran-Mantel-Haenszel)

Survival Analyses

- Time until an event (Log-Rank Tests)
- What predictors effect time until an event (Cox Proportional Hazards)

Stop In and Say Hi!

<http://clinicaltrials.stanford.edu>

The screenshot shows a web browser window with the URL <http://clinicaltrials.stanford.edu/>. The page header includes navigation links for Stanford University, Stanford Hospital & Clinics, Lucile Packard Children's Hospital, VA Palo Alto, and Santa Clara Valley Medical. The main header features the Stanford School of Medicine logo and the text "Stanford/Packard Center for Translational Research in Medicine (SPCTRM)". A search bar is present with a "SEARCH" button and options for "This Site Only" and "All Stanford Medicine Sites". The breadcrumb trail reads "School of Medicine Home » Dean's Office » SPCTRM".

SPCTRM
Stanford / Packard Center for Translational Research in Medicine

CLINICAL RESEARCH INFORMATION

- Home
- SPCTRM Clinic

Facilitating Translational Research at Stanford

SPCTRM serves faculty and clinical research personnel at Stanford: in the School of Medicine, Stanford University Medical Center, Lucile Packard Children's Hospital (LPCH) and Veterans Affairs Palo Alto Health Care System (VAPAHCS).

BIostatistician

Needing assistance with the biostatistical portion of your grant proposal? Find out more about SPCTRM's [Biostatistical Consultation Services](#).