Flowchart for statistical tests and estimation. BIEB100. Winter 2014. Prof. Scott Rifkin.

Type of data	Questions	Possible choices of statistics	Possible H0 & HA	Approach to generate a null distribution	Approach to generate a confidence interval
Single binary categorical variable	Is the sample from some hypothesized population? Is the population proportion equal to some value? Are the counts in each category what you would expect?	- sample proportion in one category: $\hat{p}$ (if you know one you can calculate the other) - sample counts in one category	H0: $\pi = \pi_0$ HA: $\pi > \neq < \pi_0$	**Simulation** Imagine (or using the computer) putting an infinite number of white and black balls into a jar so that a fraction π_0 of them are black. Then draw a sample of size *n* from the urn, calculate the fraction of black balls (or the counts if that is your chosen statistic), record this statistic on a dotplot, and put the sampled balls back in. Repeat this many times. **Theory** Conditions: at least 10 counts for each value of the categorical variable. π_0 not too close to 0 or 1. Use a normal distribution. Using a proportion as your statistic: mean: π_0 standard error: $\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}$ Using counts as your statistic: mean: (n x π_0) standard error: $\sqrt{n\hat{p}(1-\hat{p})}$ The standardized statistic is called a z-score Find the cutoff for your desired significance level from a table of normal distribution values. Pay attention to whether your HA is one-sided or two-sided. Or you can use a z-score -> p-value converter online.	**Bootstrapping** Same approach as for simulating a null distribution <u>except</u> instead of using a fraction of π_0 use a fraction of $\hat{p}$. **Theory** Conditions: at least 10 counts for each value of the categorical variable. π_0 not too close to 0 or 1. Use a normal distribution. Using a proportion as your statistic: mean: $\hat{p}$ standard error: $\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}$ Using counts as your statistic: mean: (n x $\hat{p}$) standard error: $\sqrt{n\hat{p}(1-\hat{p})}$ Find the multiplier for your desired confidence level from a table of normal distribution values. Or you can use a p-value -> z-score converter online. Remember that an X% confidence level corresponds to a (100-X)% two-sided significance level.
Single quantitative variable	Is the sample from some hypothesized population? Does the sample have approximately the same location as hypothesized? Does the sample have approximately the same spread as hypothesized? Does the sample have approximately the same shape as hypothesized?	- sample mean: $\bar{x}$ - sample median - sample standard deviation: SD or s - sample range	H0: $\mu = \mu_0$ HA: $\mu > \neq < \mu_0$ or equivalent for other statistics	**Simulation** You need to construct a population from which to sample. Unlike the case of a single categorical binary variable, here you know far less about the population. You will have to make some assumptions about what it looks like. A common situation is that you assume that your variable is normally distributed in the population with mean μ_0 and the same standard deviation as your sample. You may sometime find yourself in a situation where you want to make other assumptions about the population distribution. Repeatedly take samples of size *n* from your assumed population, calculate your statistic, and use these to construct a null distribution. **Theory** Conditions: sample size of at least 20 or the sample is known (or assumed) to come from a normally distributed population. For the mean: use a t-distribution. mean: μ_0 standard error: $\frac{SD}{\sqrt{n}}$ degrees of freedom: *n-1* The standardized statistic is called a t-statistic Find the cutoff for your desired significance level from a table of t-distribution values. Pay attention to whether your HA is one-sided or two-sided. Or you can use a t-statistic -> p-value converter online.	**Bootstrapping** Just as your sample statistic is your best estimate of the population parameter, your sample itself is your best estimate of the population. Repeatedly take a sample of size *n* <u>with replacement</u> from your sample. Calculate your sample statistic each time and use these to build a bootstrap sampling distribution for your statistic. For the percentile method (recommended), choose a desired confidence level and use the percentile method to find the boundaries of the confidence interval. Line up your bootstrap replicate statistics in order. If you have *B* bootstrap replicates (a big number) and *c* is your confidence level (e.g. 95%, 99%, etc.), then the lower bound of your confidence interval is the *B x (1-c)/2* bootstrap statistic value and the upper bound is the *B - (B x (1-c)/2)* bootstrap statistic value. For example, if *B* is 10,000 and *c* is 84% then chop off the bottom 8% (800 values) and chop off the top 8% (800 values). If your sample size is big enough and your sample statistic is a mean, you could use the standard deviation of your bootstrap distribution as an estimate of the standard error for the mean and plug that into the statistic +/- multiplier x SE formula like in the theory-based approach. **Theory** Conditions: sample size of at least 20 or the sample is known (or assumed) to come from a normally distributed population. For <u>the</u> mean: use a t-distribution. mean: $\bar{X}$ standard error: $\frac{SD}{\sqrt{n}}$ degrees of freedom: *n-1* The standardized statistic is called a t-statistic Find the multiplier for your desired confidence level from a table of t-distribution values. Or you can use a p-value -> t-statistic converter online. Remember that an X% confidence level corresponds to a (100-X)% two-sided significance level.
Two binary categorical variables	Is there an association between the variables? There are 3 essential degrees of association: (1) nothing predictive or causal implied - the two variables are just associated; (2) you can predict the value of one based on the value of the other to some extent. This may not be reciprocal and may not imply anything causal; (3) varying one of the variables causes the other one to also change in value. Usually you need an experiment to definitely show this. If one variable is considered an explanatory ((2) and (3) above) (or more generally a *grouping*) variable, are the groups different? Do they come from populations with different parametric counts or proportions?	- difference between the sample proportions for in one value of the response variable in each group (A, B) (conditional proportions): $\hat{p}_A - \hat{p}_B$	H0: $\pi_A - \pi_B = 0$ HA: $\pi_A - \pi_B > = < 0$	**Simulation** Shuffling/scrambling approach. You want to break any potential association between the grouping (explanatory) and response variables. Imagine all your data in two columns. The first column has the group, the second has the response. Each row represents the data from a single observational unit. Now randomize the order of the values in one of the columns. This way the number of observational units in each group remains the same and the number of responses in each category remains the same, but any association between the two variables is scrambled. Each time you scramble calculate your statistic on the scrambled data. This generates your null distribution. **Theory** Conditions: at least 10 counts for each cell in your 2x2 table of counts. Use a normal distribution. The mean will be 0. The standard error is: $\sqrt{\hat{p}(1-\hat{p})\left(\frac{1}{n_A}+\frac{1}{n_B}\right)}$ $\hat{p}$ here is the overall proportion in the dataset, <u>not</u> a conditional proportion. You use the overall proportion because your null hypothesis is that the groups are not different (they are from the same population) and so your best estimate of the population parametric proportion is the overall proportion in your dataset. The standardized statistic is called a z-score Find the cutoff for your desired significance level from a table of normal distribution values. Pay attention to whether your HA is one-sided or two-sided. Or you can use a z-score -> p-value converter online.	**Bootstrapping** Bootstrap each group in parallel but using their own $\hat{p}$. See the instructions above for bootstrapping a single proportion. However, each time you take a bootstrap sample for the two groups, calculate your statistic: $boot\hat{p}_A - boot\hat{p}_B$. These form your bootstrap distribution. You can also bootstrap counts. **Theory** Conditions: at least 10 counts for each cell in your 2x2 table of counts. Use a normal distribution. Using a proportion as your statistic: mean: 0 standard error: $\sqrt{\left(\frac{\hat{p}_A(1-\hat{p}_A)}{n_A}+\frac{\hat{p}_B(1-\hat{p}_B)}{n_B}\right)}$ Note that this standard error is different from the hypothesis testing one in that here you use the conditional proportions while there you use the overall proportion (substitute $\hat{p}$ for $\hat{p}_A$ and $\hat{p}_B$ in the formula above and you get the formula to the right). This is because here you are estimating a difference between the two groups, however inconsequential it may be. Find the multiplier for your desired confidence level from a table of normal distribution values. Or you can use a p-value -> z-score converter online. Remember that an X% confidence level corresponds to a (100-X)% two-sided significance level.
1 categorical (with 2 or more categories) and 1 quantitative variable	Is there an association between the variables? There are 3 essential degrees of association: (1) nothing predictive or causal implied - the two variables are just associated; (2) you can predict the value of one based on the value of the other to some extent. This may not be reciprocal and may not imply anything causal; (3) varying one of the variables causes the other one to also change in value. Usually you need an experiment to definitely show this. Consider the categorical variable the grouping variable. Are the groups different? Do they come from populations with different parameters?	Choose your statistic to summarize the groups. See possibilities for the single quantitative variable case above. You will need to construct a statistic comparing the individual group statistics. For two groups, you might investigate whether the group influences location by using the difference in means or the difference in medians as your statistic. If you are interested in comparing the spread of the data in each group you might use the ratio of the variances of the data in each group. For two groups or more: - MAD: mean absolute difference. Take the mean distance between group means. (Alternatively this could be medians if it makes more sense for the data). - F statistic. Effectively this compares the variation between groups (between group means) to the variation within groups (the variation of each datapoint from its group mean). It is a ratio of a variance-like term for group means to a variance-like term summarizing within group variation. An F statistic is: $\frac{\text{Mean Square Groups}}{\text{Mean Square Within}}$ Mean Square Groups is almost the variance of the group means. The only difference is that in a variance, you subtract off the mean of your individual items. If the group sizes are unequal, then the grand mean (of all the data together) will not be the same as the mean of the group means (try it!). To calculate the MS_{groups} you subtract the grand mean from each group mean, square that difference, add them up across groups, and divide by *g - 1*, where *g* is the number of groups. To get Mean Square Within, you take each individual datapoint, subtract off the mean of its group, square that difference, and add those all up. This is SS_{within}. Then divide by *n - g* to get MS_{within}.	For comparing the location of two groups: H0: $\mu_A - \mu_B = 0$ HA: $\mu_A - \mu_B > \neq < 0$ or you might have chosen a different statistic like median. For comparing the location of two or more groups: H0: $\mu_A = \mu_B = ... = \mu_Z$ HA: At least one of the group means is different. In practice, this formulation of the null and alternative hypotheses is difficult to test directly. So they have to be rephrased in terms of the statistic used (we have learned MAD and F). H0: MAD = 0 HA: MAD > 0 H0: F = 1 HA: F > 1 Note that both of these are one-sided. MAD cannot be negative. If F < 1 then this means that groups are very similar to each other, but we aren't interested in whether they are overly similar, only if there is evidence that there is at least one difference.	**Simulation** Shuffling/scrambling approach. You want to break any potential association between the grouping (explanatory) and response variables. Imagine all your data in two columns. The first column has the group, the second has the response. Each row represents the data from a single observational unit. Now randomize the order of the values in one of the columns. This way thenumber of observational units in each group remains the same and the specific values in the dataset remain the same, but any association between the two variables is scrambled. Each time you scramble calculate your statistic on the scrambled data. This generates your null distribution. The MAD and F statistics are overall statistics - they summarize variation in all of the groups together. Let's say you reject your null hypothesis that all the group means are equal. You may then want to know which group is different. You can use a "post-hoc" test to test whether pairs of groups are different from each other. The most famous one is Tukey's Honestly Significant Difference test. You can read about it here: http://web.mst.edu/~psyworld/tukeysteps.htm or just search for it on Wikipedia. For the simulation version, do the scrambling approach as described above, except you want to construct a null distribution comprised of the maximal differences between group means. In one iteration you would scramble your data, calculate the group means, compute the absolute difference between all pairs of group means, and then find the biggest absolute difference. This is your statistic and is one dot on your null distribution dotplot. Repeat this many times. Then look at the differences between means from your actual data and compare them to this distribution. You do this in a particular order. Order your means from smallest to largest. Say there are 4 groups. Then compare 4 vs. 1, 4 vs. 2, 4 vs. 3, 3 vs. 1, 3 vs. 2, 2 vs. 1. Use the following logic: If 4 vs. 2 are not significantly different, then neither will 4 vs. 3 or 3 vs. 2 (remember they are ordered). So don't bother doing those. However, this doesn't tell you about 3 vs. 1 or 2 vs. 1 so you'd want to test those. Your null distribution is the biggest differences between groups that you would expect by chance and so you are asking, "is the difference between group A and group B even bigger than the biggest difference I'd expect by chance." Remember that the simulation approach doesn't confine you to using the mean as your summary statistic. You could choose to use the median or other statistic that makes sense for your question and data. **Theory** Conditions: sample size of at least 20 for each group or the samples are known (or assumed) to come from a normally distributed population. For multiple groups, the standard deviations of the groups should be roughly equal: (the largest should be less than twice the smallest) For two groups: For the mean: use a t-distribution. mean: 0 standard error: $\sqrt{\left(\frac{SD_A^2}{n_A}+\frac{SD_B^2}{n_B}\right)}$ degrees of freedom: $n_A\text{-}1 + n_B\text{-}1$ The standardized statistic is called a t-statistic Find the cutoff for your desired significance level from a table of t-distribution values. Pay attention to whether your HA is one-sided or two-sided. Or you can use a t-statistic -> p-value converter online. For two or more groups: For the mean: use an F statistic and distribution. numerator degrees of freedom: *g - 1* denominator degrees of freedom: *n - g* Compare the F statistic to the cutoff value from the F distribution for your desired significance level. Like the t-distribution, the F-distribution is actually a family of distributions, and the exact shape depends on the degrees of freedom for the numerator and denominator. Look up the cutoff value in a table or use an F-statistic <-> pvalue converter online. See the web links above in the simulation section for the theory-based ways to determine which means are different.	**Bootstrapping** For two groups: Bootstrap each group in parallel but keeping the data from each group separate. See the instructions above for bootstrapping a single quantitative variable. However, each time you take a bootstrap sample for the two groups, calculate your statistic on the bootstrapped sample groups. These form your bootstrap distribution. For more than two groups: This is more complicated. It depends on what you really want to estimate. Often you don't really want a confidence interval for MAD or F. Rather you want confidence intervals for the differences between pairs of populations. You can do this by bootstrapping in the same way as for two groups above. Keep in mind, though, that you are doing multiple comparisons, so you have similar kinds of inaccuracies as you do when doing multiple hypothesis tests. If you have evidence from post-hoc tests that some groups are different while others may have come from the same population, you should take this into account in your confidence interval estimation. If you have no evidence that the groups are different, then pool the data together before bootstrapping. Note that you can estimate confidence intervals for the population statistics alone instead of the differences between them or for the (population mean - grand mean) or whatever makes sense for your question. **Theory** Conditions: sample size of at least 20 for each group or the samples are known (or assumed) to come from a normally distributed population. For two groups: For <u>the</u> mean: use a t-distribution. mean: $\bar{X}_A - \bar{X}_B$ standard error: $\sqrt{\left(\frac{SD_A^2}{n_A}+\frac{SD_B^2}{n_B}\right)}$ degrees of freedom: $n_A\text{-}1 + n_B\text{-}1$ The standardized statistic is called a t-statistic Find the multiplier for your desired confidence level from a table of t-distribution values. Or you can use a p-value -> t-statistic converter online. Remember that an X% confidence level corresponds to a (100-X)% two-sided significance level. For two or more groups: This is a bit more complicated. The F statistic isn't really what you want to estimate. Instead, For the mean: use an F statistic and distribution. numerator degrees of freedom: *g - 1* denominator degrees of freedom: *n - g* Compare the F statistic to the cutoff value from the F distribution for your desired significance level. Like the t-distribution, the F-distribution is actually a family of distributions, and the exact shape depends on the degrees of freedom for the numerator and denominator. Look up the cutoff value in a table or use an F-statistic <-> pvalue converter online.
2 quantitative variables	Is there an association between the variables? There are 3 essential degrees of association: (1) nothing predictive or causal implied - the two variables are just associated; (2) you can predict the value of one based on the value of the other to some extent. This may not be reciprocal and may not imply anything causal; (3) varying one of the variables causes the other one to also change in value. Usually you need an experiment to definitely show this. For correlation the two variables are on equal footing. This statistic/analysis is used primarily for (1) above, although the statistic is related to the statistic used for regression in (2) and (3). Does variation in one variable predict or cause variation in the other variable?	For correlation: The usual statisic is *r*, the correlation coefficient which has a minimal value of -1 (perfectly inversely correlated) to maximal value of 1 (perfectly correlated). 0 means not correlated at all. For regression: The slope of the regression line, *b*, is the key summary of how the value of one variable depends on the value of the other. One important thing to know is: $b = r\frac{SD_Y}{SD_X}$ If there were a perfect correlation (*r=1*) then the data would all fall on a straight line. Then the slope of the line would be the standard deviation of the Y values divided by the standard deviation of the X values. You can also use an approach similar to that for multiple groups/quantitative response above for regression - you can think of each different value of your x variable as being the basis of a group. In that situation, you might want to use the F statistic, but we will not cover that in class. In the theory-based approach, you use a z_r, t-, or F-statistic for the correlation and a t- or F-statistic for regression.	For correlation: H0: $\rho = 0$ HA: $\rho > \neq < 0$ (note that 0 means no correlation. It is possible to test whether the correlation coefficient equals some hypothesized null value with a modification of the test statistic. See the theory-based section to the right. Or use CIs.) For regression: H0: $\beta = \beta_0$ HA: $\beta > \neq < \beta_0$ Often you will want to test whether the slope is $\beta_0 = 0$, but you may have prior information or a previous result that suggests a non-zero null value to test. If you use a theory-based approach, you will use either a z-statistic or a t-statistic so the hypotheses get translated into: H0: Z = 0 HA: Z > ≠ < 0 and H0: t = 0 HA: t > ≠ < 0 Compare this to a t-distribution with *n-2* degrees of freedom. If $\rho_0 \neq 0$, then first transform *r* into what is confusingly often called z. This is not yet a z-score but is the precursor to one: $z = 0.5\ x\ ln\left(\frac{1+r}{1-r}\right)$ where *ln* is the natural log. Transform ρ_0 using this formula too into a null parameter called ζ_0 (that's a Greek zeta). Then compute the standardized z-score in the normal way using the standard error of z: $\sqrt{\frac{1}{n-3}}$ like most aspects of the theory-based approach, this estimate of SE_z improves as n gets bigger. final standardized z-statistic: $\frac{z-\zeta_0}{SE_z}$ compare this statistic to a normal distribution. For regression: Conditions: for each value of X there is a normal distribution of Y values in the population and all these normal distributions have the same variance. The actual relationship between X and Y is linear. The X values are known without error or at least the error (e.g. measurement error) is much smaller than the error in measuring Y. The theory approach works alright even if some of these don't hold. standard error for the slope: $\frac{SD_Y}{SD_X}\sqrt{\frac{1-r^2}{n-2}}$ Note that this is just the standard error for the correlation coefficient multiplied by the slope if the correlation were perfect. Then construct the standardized statistic in the usual way. Compare it to a t-distribution with *n-2* degrees of freedom.	**Simulation** **(much easier than theory)** Shuffling/scrambling approach. You want to break any potential association between the two variables. Imagine all your data in two columns. Each column has the values for one of the variables. Each row represents the data from a single observational unit. Now randomize the order the values in one of the columns. This way he number of observational units remains the same and the values in the dataset remain the same, but any association between the two variables is scrambled. Each time you scramble calculate your statistic on the scrambled data. This generates your null distribution. While the simulation approaches are just great in general, the one for correlation is especially clean and simple. Because this shuffling breaks any associations between the variables, the null is no association for both correlation and regression (ρ=0 or β=0). There are cases where you might want to test whether your sample slope or correlation is different from a specified non-zero value. In these cases either use the theory-based approach below or construct a confidence interval and see if the hypothesized null parameter falls within it. **Theory** For correlation: Conditions: both variables come from a normally distributed population and the Y values at each X are normally distributed, and the X values at each Y are normally distributed in the population. If $\rho_0 = 0$, then standardize *r* by dividing by its standard error: $\sqrt{\frac{1-r^2}{n-2}}$ Compare this to a t-distribution with *n-2* degrees of freedom. If $\rho_0 \neq 0$, then first transform *r* into what is confusingly often called z. This is not yet a z-score but is the precursor to one: $z = 0.5\ x\ ln\left(\frac{1+r}{1-r}\right)$ where *ln* is the natural log. Transform ρ_0 using this formula too into a null parameter called ζ_0 (that's a Greek zeta). Then compute the standardized z-score in the normal way using the standard error of z: $\sqrt{\frac{1}{n-3}}$ like most aspects of the theory-based approach, this estimate of SE_z improves as n gets bigger. final standardized z-statistic: $\frac{z-\zeta_0}{SE_z}$ compare this statistic to a normal distribution. For regression: Conditions: for each value of X there is a normal distribution of Y values in the population and all these normal distributions have the same variance. The actual relationship between X and Y is linear. The X values are known without error or at least the error (e.g. measurement error) is much smaller than the error in measuring Y. The theory approach works alright even if some of these don't hold. standard error for the slope: $\frac{SD_Y}{SD_X}\sqrt{\frac{1-r^2}{n-2}}$ Note that this is just the standard error for the correlation coefficient multiplied by the slope if the correlation were perfect. Then construct the standardized statistic in the usual way. Compare it to a t-distribution with *n-2* degrees of freedom.	**Bootstrapping** **(much easier and more versatile than theory)** For both correlation and regression: Sample observational units with replacement. These are your bootstrap samples. Calculate the correlation coefficient or any feature of the regression line including the slope (which is the mean response), the intercept, and a confidence interval for an individual response. These bootstrap statistics form your bootstrap sampling distribution. **Theory** Conditions: see the conditions for hypothesis testing. For correlation and regression: See instructions for finding the standard errors in hypothesis testing. Then use the usual way of constructing theory based confidence intervals: statistic +/- multiplier x SE with multiplier from the t or normal distributions. See the hypothesis testing instructions. Note that for regression the standard error in the left column is for the slope which gives the mean response. To construct a confidence interval for the intercept or a "prediction interval" for an individual response for a specific value of X, you would need to use a different SE.