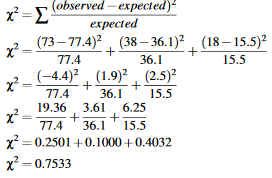
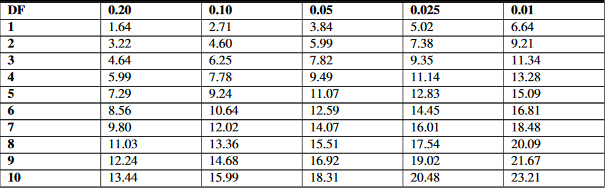
**Chi-Squared**

* The type of test to choose depends on the data available + what question we are trying to answer.
* We analyze simple descriptive statistics (mean, median, mode, SD to give us an idea of the distribution + to remove outliers, if necessary.
* We calculate probabilities to determine the likelihood of something happening.
* Finally, we use regression analysis to examine the relationship between 2+ continuous variables.
* But suppose you wanted to evaluate a recent statistic stating iOS + Android represents 32% + 51% of active smart phones.
* You would like to know if the statistic actually reflects the distribution of phones among your friends
* How could you evaluate the data you collect to see if it supports this hypothesis?
* The primary difference of a **chi-square test** is that chi square tests are for used for categorical data.
* Can be used to estimate how closely the distribution of a categorical variable matches an expected distribution (**goodness-of-fit test**), or to estimate whether 2 categorical variables are independent of one another (**test of independence**).
* The chi square test of independence is a natural extension of contingency tables to examine whether or not 2 variables appeared to be independent of each other.
* A **goodness of fit test** is concerned w/ the distribution of ONE categorical variable. 

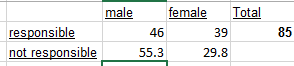


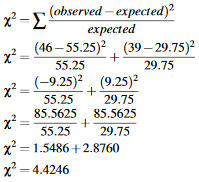


* **Observed** = actual count values in each category
* **Expected** = the predicted/expected counts in each category if the null were true
* Conducting a Chi-Square test is much like conducting a Z-test or T-test.
* We will follow the same basic series of steps + compare a calculated value to a value on a distribution table to evaluate the probability of getting the results we have if null is true, just as we did with the Z and T tests.
* Now, we will use something called the chi-square distribution table.
* For the goodness-of-fit chi-square test, the dF = # of levels in the categorical variable – 1
* Assumptions of the Chi-Square test
* Random sample.
* Independent observations for the sample (1 observation per subject).
* No expected counts less than 5.
* Notice that the last 2 assumptions are concerned w/ *expected* counts, not the raw observed counts.
* Ex: The American Pet Products Association conducted a survey in 2011 + determined that 60% of dog owners have only 1 dog, 28% have 2 dogs, and 12% have 3+.
* Supposing you conduct your own survey + have collected the data below, determine whether your data supports the results of the APPA study w/ significance level of 0.05.
* Data: Out of 129 dog owners, 73 had one dog and 38 had two dogs.
* 1) Clearly state the null and alternative hypotheses.
* H(0) = The proportion of dog owners w/ 1, 2 or 4 dogs is 0.60, 0.28 and 0.12 respectively.
* H(a) = The proportion of dog owners w/ 1, 2 or 4 dogs does NOT match the proposed model
* 2) ID an appropriate test and significance level.
* A Chi-Square goodness of fit test is appropriate because we are examining the *distribution of a single categorical* variable.
* In the absence of a stated significance level in the problem, assume the default 0.05.
* 3) Analyze sample data.
* Create a table to organize data + compare the observed data to the expected data, where expected data is the expected % from APPA multiplied by total observations
* 
* To calculate chi-square d, first sum the squared difference between each observed + expected value divided by the expected value
* 
* Now that we have our chi-square statistic, we need to compare it to the chi-square value for the significance level 0.05
* dF = # of observed category values - 1.
* In this case, there are 3 category values: 1 dog, 2 dogs, and 3+ dogs.

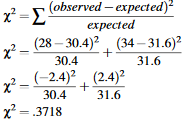


* Using the table, we find that the critical value for a 0.05 significance level w/ dF = 3 = 5.99
* That means that 95 times out of 100, a survey that agrees w/ a sample will have a chi-squared value of 5.99 or less.
* If our chi-square value is *greater than 5.995*, the measurements we took only occur 5 or fewer times out of 100 = null hypothesis is incorrect.
* Our chi-square statistic is only **0.7533**, so we will NOT reject the null.
* 4) Interpret the results.
* Since our chi-square statistic was less than the critical value, we do not reject the null + can say that our survey data *does support the data from the APPA.*
* Ex: Rachel told Eric the reason her car insurance is less expensive is that female drivers get in fewer accidents than male drivers.
* Specifically, she says male drivers are held responsible in 65% of accidents involving drivers under 23. Eric does some research of his own + discovers 46 out of 85 accidents he investigates involve male drivers, so does his data support or refute Rachel’s hypothesis?
* 1) Clearly state the null and alternative hypotheses.
* H(0) = Males are responsible for 65% of accidents + females are responsible for 35%.
* H(a) = The data do not match the proposed model.
* 2) ID an appropriate test and significance level.
* Again, a Chi-Square goodness of fit test is appropriate. In the absence of a stated significance level in the, assume the default 0.05.
* 3) Analyze sample data.



* 
* Now that we have our chi-square statistic, we need to compare it to the chi-square critical value for 0.05 w/ 2 – 1 = 1 degree of freedom, since we have 2 categories.
* Using the table, we find the critical value to be 3.84, which indicates that only 0.05, or 5%, of values would be as high as 3.84.
* If our χ2 of our data is *greater than 3.84*, then *fewer than 5 times out of 100 would we expect to get that result if the null is true*.
* 4) Interpret your results.
* Our calculated data value of χ2 is greater than the 0.05 significance level critical value of 3.84, so we REJECT the null = the data Eric observed does NOT support the distribution that Rachel claimed.
* Ex: The online car magazine “Camaro5.com” claims that, given a choice between a Ford Mustang or Chevy Camaro, 51% of readers will choose a Camaro. Ellen is a Mustang lover and decides to do some research. If Ellen collects the data below, does her data support the magazine’s claim?
* Data: Mustang owners: 28, Camaro owners: 34
* 1) ID the null and alternative hypotheses
* H(0) - 51% of Camaro5 users prefer the Camaro to the Mustang and 49% prefer the Mustang to the Camaro
* H(a) – Readers do not have this proposed preference
* 2) ID an appropriate test and significance level.
* A Chi-Square Goodness of Fit test is appropriate. In the absence of a stated significance level in the problem, we assume the default 0.05.
* 3) Analyze sample data.





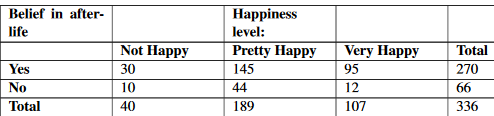
* The chi-square critical value for dF = 1 and a significance level of 0.05 is 3.8414
* 4) Interpret your results.
* Our calculated data value of χ2 is significantly less than the 0.05 significance level critical value of so we FAIL to reject the null hypothesis.
* This means that, unfortunately for Ellen, her research did NOT allow her to deny the claim that Camaros are more popular.
* Features of the Goodness-of-Fit Test
* Goodness-of-fit test is used to determine *patterns of distinct categorical variables*.
* The test requires that data be obtained through a random sample.
* dF associated with a particular chi-square test = number of categories minus one.
* There are many situations that use the goodness-of-fit test, including surveys, taste tests, + analysis of behaviors.
* Interestingly, goodness-of-fit tests are also used in casinos to determine if there is cheating in games of chance, such as cards or dice.
* Ex: If a certain card or number on a die shows up more than expected (high observed frequency compared to expected), officials use goodness-of-fit to determine the likelihood that the player may be cheating or that the game may not be fair.

**Test of Independence**

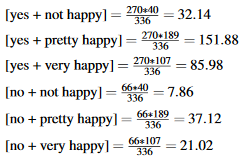
* Chi-square can be used to both estimate how closely an observed distribution matches an expected distribution (goodness-of-fit test) AND to estimate whether 2 random variables are independent of one another (**test-of-independence**).
* The chi-square test of independence is often used in social science research to determine if factors are independent of each other
* Ex: determine relationships between voting patterns + race, income + gender, behavior + education
* In general, when running a test of independence, we ask, “*Is Variable X independent of Variable Y?*”
* It is important to note that this test does not test HOW the variables are related, just simply whether or not they ARE independent of one another.
* This chi-square test can be used to determine if observed data indicates that 2 variables are dependent in much the same way that the chi-square test can be used to determine goodness of fit.
* Just as w/ goodness of fit, we calculate expected values, calculate a chi-square statistic, + compare it to the appropriate chi-square value from a reference to see if we should reject H(0) = the variables are not related
* H(0): There is NO association between the 2 categorical variables
* H(A): There IS an association (the 2 variables are NOT independent)
* In fact, the only major difference in process between a goodness of fit test and a test of independence is *how we calculate the expected values*



* C = observed column total for the cell, R = observed row total for the cell, n = total number of samples. (see Example A for clarification of the use of the formula)
* dF = (rows – 1)\*(cols – 1)
* **Contingency tables** can help us frame our hypotheses and solve problems.
* Often, we use contingency tables to list the variables + observational patterns that will help us to run a chi-square test
* Could use a contingency table to record answers to phone surveys or observed behavioral patterns.
* Assumptions of the Chi-Square Independence Test
* Random sample.
* Independent observations for the sample (one observation per subject).
* All expected counts greater than one.
* No more than 20% of cells with an expected count less than five.
* Ex: 336 individuals identified both their happiness level and their belief in an afterlife.



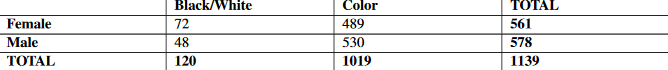
* Let’s calculate the expected values in each cell of the table:

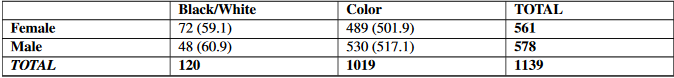


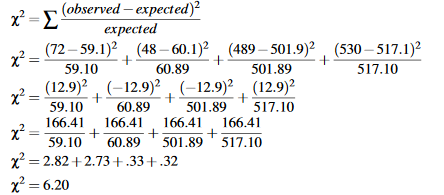
* We’ll now use the data from the above contingency table, and the expected values, to statistically test if happiness is associated with a belief in the afterlife.
* Hypothesis Step 1: Clearly state the Null and Alternative Hypothesis.
* H(0): There is NO association between happiness and the belief in the afterlife
* H(a): There IS an association (the 2 variables are *not independent*)
* Hypothesis Step 2: ID the appropriate significance level and confirm the test assumptions.
* Use the standard significance level of 0.05 + we were told the survey was random, + we’ll assume that a single subject is not living in 2 places at once.
* From our expected counts that we calculated, we can confirm that the last 2 assumptions have been met.
* Hypothesis Step 3 : Analyze the data.
* Using the chi-square statistic:



* We already know we have 2 dF so we find a critical value of 5.99.
* Hypothesis Step 4: Interpret your results.
* We find our calculated chi-square value is greater than our critical value, so we REJECT the Null + conclude that happiness is NOT independent of a belief in the afterlife = ARE related
* Ex: Rachel claims girls take more black + white and color photographs than boys, but Jack (who is a photographer) is skeptical. If Jack collects the following data, would it be correct to say that he should reject Rachel’s claim that gender affects tendency to take photographs?



* The question here is whether gender affects tendency to take more photographs, or, in other words, are gender and photograph-taking tendency dependent?
* To run a chi-squared test, we need to know the expected value for each of the four cells containing observations.
* 
* Now we can calculate our χ2 statistic



* To see if our χ2 statistic is greater/less than the critical value at the default significance level of 0.05, we need the dF = (rows – 1)\*(cols – 1) = (2-1)\*(2-1) = 1
* So the critical value for 0.05 with dF = 1 is 3.8414.
* Finally, we compare our calculated χ2 of 6.2 to the critical value of 3.8414 + determine that since our calculated value is greater than the critical value, we can REJECT h(0) = reject the independence of the variables
* ***The observed data indicates that there is a gender bias on picture-taking tendency.***