Kind of Questions to Ask Yourself about Data and How to Approach Answering Them

UVA Research Data Services: StatLab
Caitlin I. Steiner

February 4, 2015

Contents

1 How Many Variables are of Interest?

1.1 How Many Response Variables are of Interest?

While the answer to this question is usually one response variable for most statistical hypothesis tests, it is still an important question to ask. There are numerous statistical analysis methods that deal with two or more response variables. To answer this question you need to determine what is the main variable(s) of interest that you want to draw a conclusion about.

1.2 How Many Explanatory Variables are of Interest?

Furthermore, sometimes when analyzing data you want to test the effect of the response variable(s) based on other variables denoting characteristics of the dataset, known as explanatory variables. If there are no explanatory variables of interest, then your data should lead to either a one-sample or two-sample analysis. But, if there is at least one explanatory variable of interest, then you are trying to determine the relationship between the explanatory variable(s) and the response variable(s). Hence, you are testing the variability of your response based on characteristics of the subjects in the data.

1.3 Not Sure? Do Some Preliminary Analysis and Background Research

Before doing any statistical analysis on a dataset, one should always calculate some descriptive statistics about the variables. This not only allows one to notice outliers and possible problems about the data but provides quantitative information about the variables in a manageable form.
Descriptive statistics include but are not limited to the following:

Univariate Analysis Analyzing one variable at a time

the distribution a summary of the frequency/count of individual values or ranges for a variable

- frequency table
- bar graph or histogram
- stem-and-leaf plot

the central tendency an estimate of the "center" of a distribution of values

- mean
- median
- mode

the dispersion the spread of the values about the central tendency

- standard deviation/variance
- quartiles (minimum, 1st quartile, median, 3rd quartile, maximum)
- boxplot

Multivariate Analysis Analyzing the effect one or more variables has on a response

- scatterplot (2D or 3D)
- correlation
- covariance
- cross-tabulations and contingency tables
- conditional distributions

Ultimately, the choice of what variables one should use is dependent on the dataset and one's background knowledge of the material. If it does not logically make sense to use a specific variable, then it is usually best to leave that out of the analysis. Also keep in mind that sometimes transformations of variables are required and this should also be considered.

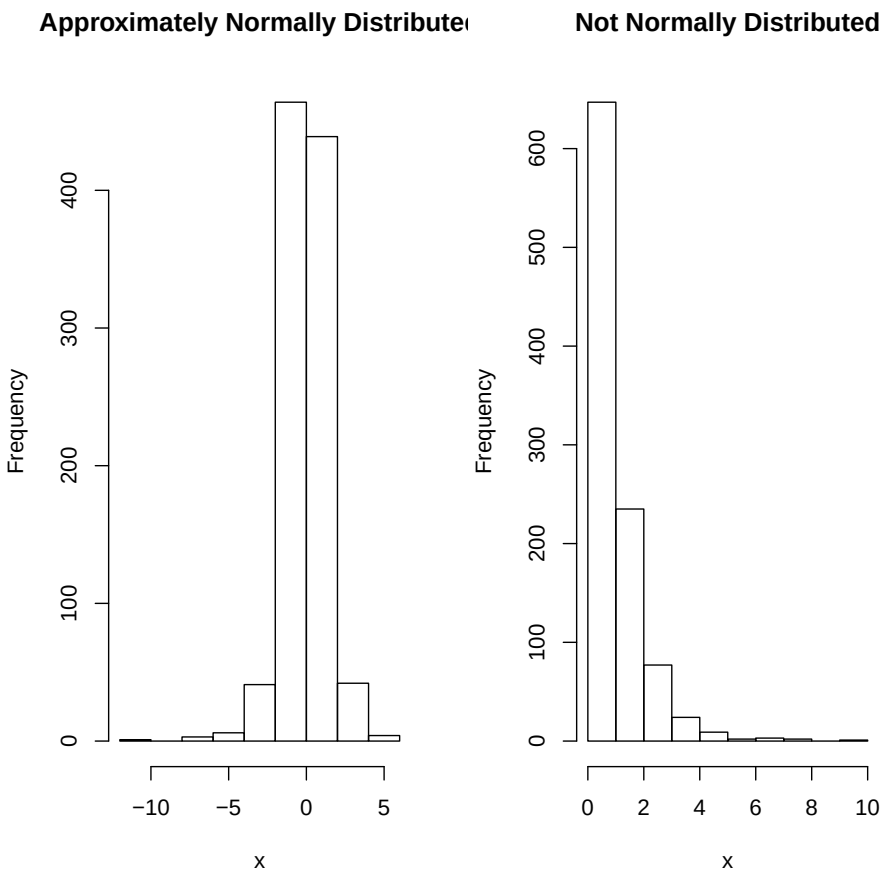
2 What is the Underlying Distribution(s) of the Variable(s) of Interest?

Often when analyzing data we either know what kind of distribution a variable comes from based on previous research or we are able to make assumptions about the distribution based on preliminary analysis or testing the variable.
Depending on the variable type (qualitative/categorical or quantitative/continuous), different assumptions about the distribution can be made, which leads to different hypothesis tests. Hypothesis tests are generally classified as one of the three sets:

1. Parametric Tests
2. Nonparametric Tests
3. Categorical Tests

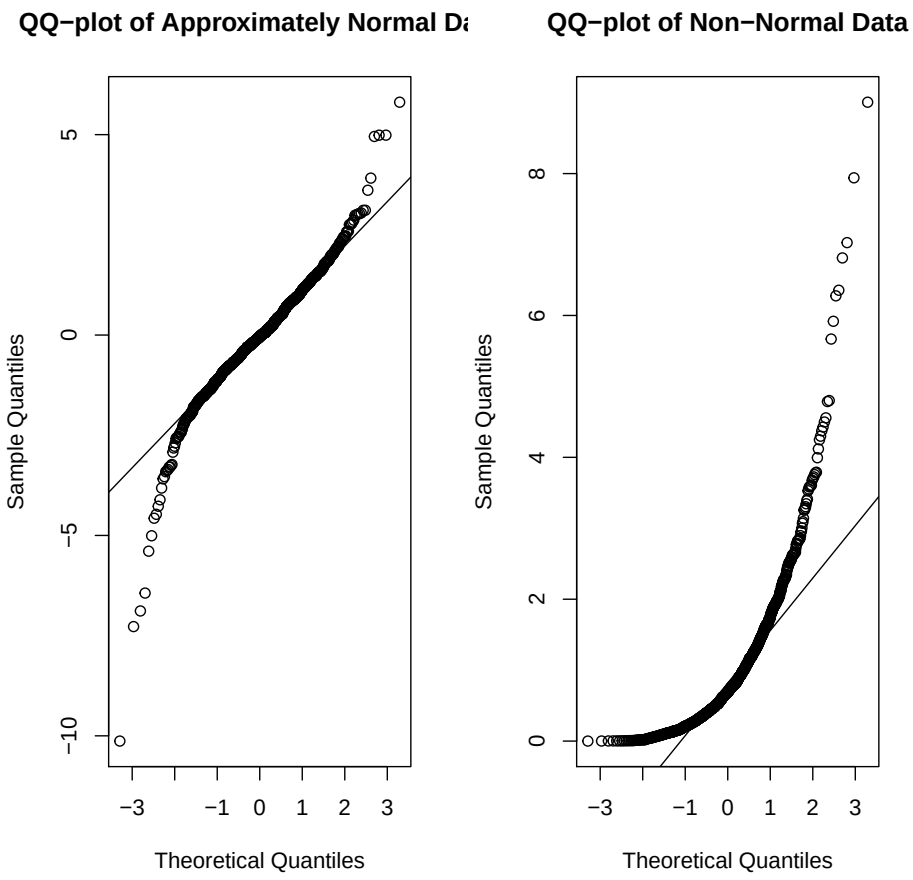
2.1 Parametric Tests: Underlying Distribution is Normal

Parametric tests requires normality of the variable(s). Hence, for a variable to be considered approximately normal a histogram of that variable should have an overall shape that looks symmetric about a peak-value. In other words, it should look bell-shaped.



Several ways of testing normality of a variable exist ranging from making an opinion about a visual representation of the variable to implementing a hypothesis test. Which method one uses depends on how liberal or conservative the researcher wants to be about the assumptions of a dataset.

A simple preliminary analysis of normality can be performed by creating a histogram of the variable for if the histogram does not look normal then there is no point in wasting time testing normality, unless you want to statistically show that a variable is not normal. The most common method of determining normality is to analyze the Quantile-Quantile plot (QQ-plot) of a variable. A QQ-plot is a graphical method of comparing a variable's probability distribution against a normal distribution by plotting their quantiles against each other. If the two distributions being comparied are similar, then the points on the QQ-plot will approximately lie on a straight line. Ergo, if the points in the QQ-plot result in a straight line, then one can conclude that the variable is approximately normal.



When evaluating a QQ-plot it is important to note that if a majority of the points (mainly the middle 50%) lie on a line while the tails bow out, one can still assume that the variable is approximately normal if there is a large sample size and inference about the mean is desired. This is due to the Central Limit Theorem that states, for any given population with a mean and variance, as long as the sample size is large enough (usually greater than or equal to 30) then the sampling distribution of the mean is approximately normal.

Besides using graphical analysis to determine if a variable is approximately normal, one can also use a hypothesis test to test normality. Normality tests include:

- D'Agostino's K-squared test
- Jarque-Bera test
- Anderson-Darling test
- Cramer-von Mises criterion
- Shapiro-Wilk test
- Pearson's chi-squared test
- Shapiro-Fancia test

2.2 Nonparametric Tests: Underlying Distribution is Not Normal

Nonparametric tests reduce the assumptions required by parametric tests by analyzing the ranks rather than the raw values of a variable. The main difference is that normality of the variable(s) is not essential for nonparametric tests. Therefore when the sample size too small to assume the sampling distribution is approximately normal, the variable is skewed or has a high kurtosis (unusually peaked), or any other parametric assumption is violated, a nonparametric test should be used instead. In essence, for any given parametric hypothesis test there is an equivalent nonparametric test.

2.3 Categorical Tests: Underlying Distribution is Binomial

Categorical tests are similar to nonparametric tests in that they do not require the assumption of normality, but they deal with discrete/qualitative variable(s) instead of continuous/quantitative variable(s). There are three types of categorical variables:

Nominal variables with two or more categories, but do not have an intrinsic order

- e.g. Type of Property (houses, condos, and co-ops)

Dichotomous nominal variables with only two categories

- e.g. Gender (male or female)

Ordinal variables with two or more categories, but do have an intrinsic order or rank

- e.g. Survey Question (Not very much, Neutral, Yes a lot)

Since categorical data is recorded as the frequency of occurrence (or count) of a particular level, one generally creates a contingency table for analysis. A contingency table shows the counts of how many times each of the levels actually happened in a particular sample.

	Blue eyes	Brown eyes
Fair hair	38	11
Dark hair	14	51

3 Review of Hypothesis Testing

A hypothesis test uses a sample to test hypotheses about the population from which the sample is drawn. This helps you make decisions or draw conclusions about the population.

3.1 Hypotheses

In hypothesis tests we formulate a null and alternative hypothesis about the unknown population parameter of some random variable's distribution from which sample(s) are drawn.

- H_0: null hypothesis
  - Hypothesis about the population from which sample or samples are drawn
  - Usually a hypothesis about the value of an unknown parameter
  - Hypothesis you want to "give the benefit of the doubt" to
  - Commonly written as an equal inequality
    $H_0 : \mu = \mu_0$ or $H_0 : \mu_1 = \mu_2$
- H_1: alternative hypothesis
  - Hypothesis you wish to establish
  - Hypothesis that will be accepted if there is enough evidence to reject the null hypothesis
  - Can be directional (one-tailed) or non-directional (two-tailed)
    $H_0 : \mu < 0$ or $H_0 : \mu > 0$ or $H_0 : \mu \neq 0$

We conclude in favor of H_0 or H_1, but not both. All inconclusive data is resolved in favor of H_0 (failing to reject the null hypothesis).This is not saying that the H_0 is true but that we don't have enough evidence to say our result didn't just happen by chance. Thus a non-significant result tells us is that the difference between what we are testing and what we know is not big enough to be anything other than a chance finding.

3.2 Terminology

Test Statistic is a function of the sample data

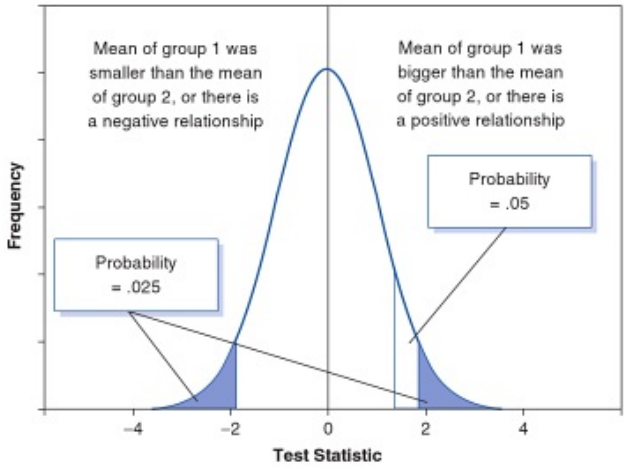
$$\text{test statistic} = \frac{(\text{what testing}) - (\text{mean of what testing})}{(\text{standard error of what testing})}$$

It varies depending on the type of hypothesis test being used and is a statistic calculated from the sample values, which has a known distribution under the null hypothesis. Also there is no direct interpretation of a test statistic value.

Rejection (Critical) Region is the range of values of the test statistic that would tend to make us doubt H_0 and conclude in favor of H_1.

Significance level (α) is the probability, assuming the truth of H_0, that the test statistic is in the rejection region. It measures how convincing the evidence is when we reject H_0 and are in favor of H_1. One can view it as the cut-off at which the null hypothesis is rejected.
Note that the significance level should be determined before beginning the test

- default level is $\alpha = 0.05$
- $\alpha = 0.01$ or $\alpha = 0.1$ are also popular



P-value is the probability calculated assuming H_0 is true (aka how likely H_0 is correct).

- If p-value$\leq \alpha$, H_0 is rejected and one can conclude that the data is significant
- If p-value$> \alpha$, we fail to reject H_0. This does not mean H_1 is true, just means we conclude in favor of H_1.

3.3 Test Procedure Overview

1. Identify parameter of interest
   - it could be the population mean, proportion, correlation, etc.
2. State null and alterative hypothesis about that parameter
3. Choose an appropriate test statistic, T, and substitute parameter's null value and other known parameter values but leave sample related variables as unknowns.
4. State the rejection region, R, for the selected significant level
5. Collect the data (sample) and compute value of the test statistic by filling in sample-related variable values
   - If $T \in R$, we reject H_0 and are in favor of H_1: This is strong evidence that H_1 is true and H_0 is false
   - If $T \notin R$, we fail to reject H_0: Either H_1 is true or the evidence is inconclusive
6. Calculate corresponding p-value for test statistic
7. Draw conclusions

4 Are the Variances Homogeneous?

Another common assumption besides that of the underlying distribution is whether the variances are equal. In order to accurately analyze the varibles against one another, it is important to know if they have similar spreads. While one could try and analyze a visual representation of a boxplot, it is best to perform a hypothesis test for equal variances.

4.1 Two Sample F-test for equal variances

Summary

A two sample test that analyzes whether the variances of the measurement variable are different between two groups.

Assumptions

- There are two samples from two independent populations (can be different sizes)
- The two samples are independent
- Both populations are normally distributed (even if the sample sizes are large)
- Both populations variances, σ_1^2 and σ_2^2, are unknown.

Procedure

1. Hypotheses

$$H_0 : \sigma_1^2 = \sigma_2^2 \quad \text{vs.} \quad H_1 : \sigma_1^2 \neq \sigma_2^2 \text{ or } \sigma_1^2 > \sigma_2^2 \text{ or } \sigma_1^2 < \sigma_2^2$$

2. Test Statistics

$$f = \frac{s_1^2}{s_2^2} \sim F_{n_1-1, n_2-1}$$

where

- s_x^2 is the sample variance of population 1
- s_y^2 is the sample variance of poppulation 2

follows a F distribution with $n-1$ numerator degrees of freedom and $m-1$ denominator degrees of freedom.

3. Find the p-value. Note the F distribution is not symmetric so the probability of each tail must be caluclated separately and that the order of the degrees of freedom matters.

- If $H_1 : \sigma_1^2 > \sigma_2^2$, then $p - value = P(F \geq f)$
- If $H_1 : \sigma_1^2 < \sigma_2^2$, then $p - value = P(F \leq f)$
- If $H_1 : \sigma_1^2 \neq \sigma_2^2$, then $p - value = 2 * P(F \geq f)$

5 Quick Reference - One and Two Sample Tests

5.1 One Sample Z Test for Population Mean

Summary

A one sample test that analyzes whether the mean of the measurement variable is different from the theoretical expectation of what the mean should be when the population variance is known.

Assumptions

- The population is normally distributed or the sample size is large enough that the mean is normally distributed (aka $n \geq 30$).
- The population standard devitaion, σ, is known.

Procedure

1. Hypotheses
$$H_0 : \mu = \mu_0 \quad \text{vs.} \quad H_1 : \mu \neq \mu_0 \text{ or } \mu > \mu_0 \text{ or } \mu < \mu_0$$
2. Test Statistics
$$z = \frac{\bar{x} - \mu_0}{\sigma/\sqrt{n}} \sim N(0, 1)$$
where
   - $\bar{x}$ is the sample mean
   - μ_0 is the hypothesized true population mean
   - σ is the population standard deviation
   - n is the sample size

   follows a Standard Normal Distribution.
3. Find the p-value
   - If $H_1 : \mu > \mu_0$, then $p - value = P(Z \geq z)$
   - If $H_1 : \mu < \mu_0$, then $p - value = P(Z \leq z)$
   - If $H_1 : \mu \neq \mu_0$, then $p - value = 2 * P(Z \geq |z|)$

5.2 One Sample T Test for Population Mean

Summary

A one sample test that analyzes whether the mean of the measurement variable is different from the theoretical expectation of what the mean should be when the population variance is unknown.

Assumptions

- The population is normally distributed or the sample size is large enough that the mean is normally distributed.
- Do not use this test if there are outliers or the population is very skewed (Skewness can be ignored if $n \geq 40$).
- The population standard devitaion, σ, is unknown. The sample standard deviation, s, will be used in calculatiosn instead

Procedure

1. Hypotheses

$$H_0 : \mu = \mu_0 \quad H_1 : \mu \neq \mu_0 \text{ or } \mu > \mu_0 \text{ or } \mu < \mu_0$$

2. Test Statistics

$$t = \frac{\bar{x} - \mu_0}{s/\sqrt{n}} \sim T_{n-1}$$

   where

   - $\bar{x}$ is the sample mean
   - μ_0 is the hypothesized true population mean
   - s is the sample standard deviation
   - n is the sample size

   follows a T Distribution with $n - 1$ degrees of freedom

3. Find the p-value

   - If $H_1 : \mu > \mu_0$, then $p - value = P(T \geq t)$
   - If $H_1 : \mu < \mu_0$, then $p - value = P(T \leq t)$
   - If $H_1 : \mu \neq \mu_0$, then $p - value = 2 * P(T \geq |t|)$

5.3 One Sample Chi-Square Test for the Population Variance

Summary

A one sample test that analyzes whether the variance of the measurement variable is different from the theoretical variance.

Assumptions

- The population is normally distributed.
- The population variance, σ^2, is unknown

Procedure

1. Hypotheses

$$H_0 : \sigma^2 = \sigma_0^2 \quad \text{vs.} \quad H_1 : \sigma^2 \neq \sigma_0^2 \text{ or } \sigma^2 > \sigma_0^2 \text{ or } \sigma^2 < \sigma_0^2$$

2. Test Statistics

$$X^2 = \frac{(n-1)s^2}{\sigma_0^2} \sim \chi^2_{n-1}$$

where

- n is sample size
- s^2 is the sample variance
- σ_0^2 is the hypothesized population variance

follows a chi-squared distribution with $n-1$ degrees of freedom.

3. Find the p-value. Note the χ^2 distribution is not symmetric so the probability of each tail must be caluclated separeately.

- If $H_1 : \sigma^2 > \sigma_0^2$, then $p - value = P(\chi^2 \geq X^2)$
- If $H_1 : \sigma^2 < \sigma_0^2$, then $p - value = P(\chi^2 \leq X^2)$
- If $H_1 : \sigma^2 \neq \sigma_0^2$, then $p - value = 2 * P(\chi^2 \leq X^2)$ if X^2 is less than the median or $p - value = 2 * P(\chi^2 \geq X^2)$ if X^2 is greater than the median
  Note - The median is the number a, such that $P(\chi^2 \geq a) = 0.5$

5.4 One Sample Binomial Test (Z Test for Population Proportion)

Summary

A one sample test that analyzes whether the proportion of the successes is different from the theoretical proportion.

Assumptions

- Data comes from a binomial experiment.
- The sample size is large (number of expected successes $n * p_0 \geq 10$ and number of expected failures $n(1 - p_0) \geq 10$).
- The population is at least 20 times the size of the sample.

Procedure

1. Hypotheses

$$H_0 : p = p_0 \quad \text{vs.} \quad H_1 : p \neq p_0 \text{ or } p > p_0 \text{ or } p < p_0$$

2. Test Statistics

$$z = \frac{\hat{p} - p_0}{\sqrt{\frac{p_0(1-p_0)}{n}}} \sim N(0, 1)$$

where

- $\hat{p}$ is the sample proportion
- p_0 is the hypothesized true population proportion
- n is the sample size

follows a Standard Normal Distribution.

3. Find the p-value

- If $H_1 : \mu > \mu_0$, then $p - value = P(Z \geq z)$
- If $H_1 : \mu < \mu_0$, then $p - value = P(Z \leq z)$
- If $H_1 : \mu \neq \mu_0$, then $p - value = 2 * P(Z \geq |z|)$

5.5 Two Sample Z Test for Difference of the Population Mean

Summary

A two variable test that analyzes whether the means of the measurement variable are different in the two groups (nominal variable) when the population variance of both groups is known.

Assumptions

- There are two samples from two populations. (The samples can be different sizes.)
- The two samples are independent
- Both population are normally distributed or the sample size is large enough that the mean is normally distributed.
- Both population standard devitaions, σ_x and σ_y, are known.

Procedure

1. Hypotheses

$$H_0 : \mu_x - \mu_y = D \quad \text{vs.} \quad H_1 : \mu_x - \mu_y \neq D \text{ or } \mu_x - \mu_y > D \text{ or } \mu_x - \mu_y < D$$

2. Test Statistics

$$z = \frac{(\bar{x} - \bar{y}) - D}{\sqrt{\frac{\sigma_x^2}{n} + \frac{\sigma_y^2}{m}}} \sim N(0, 1)$$

   where

   - $\bar{x} - \bar{y}$ is the difference in sample mean
   - D is the hypothesized difference in population
   - n is the sample 1 size
   - m is the sample 2 size
   - σ_x^2 is the population 1 variance
   - σ_y^2 is the population 2 variance

   follows a Standard Normal Distribution.

3. Find the p-value

   - If $H_1 : \mu_x - \mu_y > D$, then $p - value = P(Z \geq z)$
   - If $H_1 : \mu_x - \mu_y < D$, then $p - value = P(Z \leq z)$
   - If $H_1 : \mu_x - \mu_y \neq D$, then $p - value = 2 * P(Z \geq |z|)$

5.6 Two Sample T Test for Difference of the Population Mean (Equal Variances)

Summary

A two sample test that analyzes whether the means of the measurement variable are different in the two groups (nominal variable) when the population variance of both groups is unknown but assumed to be equal.

Assumptions

- There are two samples from two populations. (The samples can be different sizes.)
- The two samples are independent
- Both population are normally distributed or the sample size is large enough that the mean is normally distributed.
- Both population standard devitaions, σ_x and σ_y, are unknown, but are assumed to be equal.

Procedure

1. Hypotheses

$$H_0 : \mu_x - \mu_y = D \quad \text{vs.} \quad H_1 : \mu_x - \mu_y \neq D \text{ or } \mu_x - \mu_y > D \text{ or } \mu_x - \mu_y < D$$

2. Test Statistics

$$t = \frac{(\bar{x} - \bar{y}) - D}{\sqrt{s_p^2 \left(\frac{1}{n} + \frac{1}{m}\right)}} \sim T_{n+m-2}$$

$$s_p^2 = \frac{(n-1)s_x^2 + (m-1)s_y^2}{n+m-2}$$

where

- $\bar{x} - \bar{y}$ is the difference in sample mean
- D is the hypothesized difference in population
- s_p^2 is the pooled variance estimate
- n is the sample size of sample 1
- m is the sample size of sample 2

follows a T Distribution with $n + m - 2$ degrees of freedom

3. Find the p-value

- If $H_1 : \mu_x - \mu_y > D$, then $p - value = P(T \geq t)$
- If $H_1 : \mu_x - \mu_y < D$, then $p - value = P(T \leq t)$
- If $H_1 : \mu_x - \mu_y \neq D$, then $p - value = 2 * P(T \geq |t|)$

5.7 Two Sample T Test for Difference of the Population Mean (Unequal Variances)

Summary

A two sample test that analyzes whether the means of the measurement variable are different in the two groups (nominal variable) when the population variance of both groups is unknown but assumed to be unequal.

Assumptions

- There are two samples from two populations. (The samples can be different sizes.)
- The two samples are independent
- Both population are normally distributed or the sample size is large enough that the mean is normally distributed.
- Both population standard devitaions, σ_x and σ_y, are unknown, but are assumed to be not equal.

Procedure

1. Hypotheses

$$H_0 : \mu_x - \mu_y = D \quad \text{vs.} \quad H_1 : \mu_x - \mu_y \neq D \text{ or } \mu_x - \mu_y > D \text{ or } \mu_x - \mu_y < D$$

2. Test Statistics

$$t = \frac{(\bar{x} - \bar{y}) - D}{\sqrt{\frac{s_x^2}{n} + \frac{s_y^2}{m}}} \sim T_{d.f.} \quad d.f. = \frac{\left(\frac{s_x^2}{n} + \frac{s_y^2}{m}\right)}{\frac{s_x^4}{n^2(n-1)} + \frac{s_y^4}{m^2(m-1)}}$$

where

- $\bar{x} - \bar{y}$ is the difference in sample mean
- D is the hypothesized difference in population
- n is the sample 1 size
- m is the sample 2 size
- s_x^2 is the sample 1 variance
- s_y^2 is the sample 2 variance

follows a T Distribution with $d.f.$ degrees of freedom

3. Find the p-value

- If $H_1 : \mu_x - \mu_y > D$, then $p - value = P(T \geq t)$
- If $H_1 : \mu_x - \mu_y < D$, then $p - value = P(T \leq t)$
- If $H_1 : \mu_x - \mu_y \neq D$, then $p - value = 2 * P(T \geq |t|)$

5.8 Paired Sample T Test for Population Mean of Pairwise Differences

Summary

A two variable test that analyzes whether the mean difference in the pairs (two dependent groups) is different from zero when the population variance of both groups is unknown.

Assumptions

- There are two samples that are the same size.
- The two samples are dependent.
- Both population are normally distributed or the sample size is large enough that the mean is normally distributed.
- The standard devaition of the population's difference is unknown.

Procedure

1. Hypotheses

$$H_0 : \mu_1 - \mu_2 = D \quad \text{vs.} \quad H_1 : \mu_1 - \mu_2 \neq D \text{ or } \mu_1 - \mu_2 > D \text{ or } \mu_1 - \mu_2 < D$$

2. Test Statistics

$$t = \frac{\bar{d} - D}{s_d/\sqrt{n}} \sim T_{n-1}$$

where

- $\bar{d}$ is the sample mean of pairwise differences
- D is the hypothesized mean of pairwise difference
- s_d is the sample standard deviation of pairwise differences
- n is the sample size

follows a T Distribution with $n - 1$ degrees of freedom

3. Find the p-value

- If $H_1 : \mu_1 - \mu_2 > D$, then $p - value = P(T \geq t)$
- If $H_1 : \mu_1 - \mu_2 < D$, then $p - value = P(T \leq t)$
- If $H_1 : \mu_1 - \mu_2 \neq D$, then $p - value = 2 * P(T \geq |t|)$

5.9 Two Sample F Test for Population Variance

Summary

A two sample test that analyzes whether the variances of the measurement variable are different in the two groups.

Assumptions

- There are two samples from two populations (can be different sizes)
- The two samples are independent
- Both populations are normally distributed (even if the sample sizes are large)
- Both populations variances, σ_X^2 and σ_y^2, are unknown.

Procedure

1. Hypotheses

$$H_0 : \sigma_x^2 = \sigma_y^2 \quad \text{vs.} \quad H_1 : \sigma_x^2 \neq \sigma_y^2 \text{ or } \sigma_x^2 > \sigma_y^2 \text{ or } \sigma_x^2 < \sigma_y^2$$

2. Test Statistics

$$f = \frac{s_x^2}{s_y^2} \sim F_{n-1,m-1}$$

where

- s_x^2 is the sample variance of population 1
- s_y^2 is the sample variance of poppulation 2

follows a F distribution with $n-1$ numerator degrees of freedom and $m-1$ denominator degrees of freedom.

3. Find the p-value. Note the F distribution is not symmetric so the probability of each tail must be caluclated separeately and that the order of the degrees of freedom matters.

- If $H_1 : \sigma_x^2 > \sigma_y^2$, then $p - value = P(F \geq f)$
- If $H_1 : \sigma_x^2 < \sigma_y^2$, then $p - value = P(F \leq f)$
- If $H_1 : \sigma_x^2 \neq \sigma_y^2$, then $p - value = 2 * P(F \geq f)$

5.10 Two Sample Binomial Test (Z Test for Difference of the Population Mean)

Summary

A two sample test that analyzes whether the proportions are different in the two groups.

Assumptions

- The data comes from a binomial experiment.
- Both sample sizes are large (number of succes and number of failures is at least 5 for both samples)

Procedure

1. Hypotheses

$$H_0 : p_x - p_y = D \quad \text{vs.} \quad H_1 : p_x \neq p_y \text{ or } p_x \geq p_y \text{ or } p_x \leq p_y$$

2. Test Statistics

$$z = \frac{\hat{p}_x - \hat{p}_y}{\sqrt{\hat{p}(1 - \hat{p}\left(\frac{1}{n} + \frac{1}{m}\right))}} \sim N(0, 1)$$

   where

   - $\hat{p}_x$ is sample 1 proportion
   - $\hat{p}_y$ is sample 2 proportion
   - $\hat{p} = \frac{X+Y}{n+m}$ is the overall proportion of successes
   - n is the sample 1 size
   - m is the sample 2 size

   follows a Standard Normal Distribution.

3. Find the p-value

   - If $H_1 : p_x > p_y$, then $p - value = P(Z \geq z)$
   - If $H_1 : p_x < p_y$, then $p - value = P(Z \leq z)$
   - If $H_1 : p_x \neq p_y$, then $p - value = 2 * P(Z \geq |z|)$

5.11 Fisher's Exact Test

Summary

A small sample, two nominal variable test that analyzes whether the proportions of one variable are different depending on the value of the other variable.

Assumptions

- The individual observations are independent
- The row and column totals are fixed, or "conditioned'

Procedure

1. Hypotheses

$$H_0 : \theta = 1 (\text{independence}) \quad \text{vs.} \quad H_1 : \theta < 1 \text{ or } \theta \neq 1 \text{ or } \theta > 1$$

where θ is the odds ratio.
Interpretation of the null hypothesis is that the proportions at one variable are the same for different values of the second variable or relative proportions of one variable are independent of the second variable.

2. Find the p-value

- The p-value is the sum of hypergeometric probabilities for outcomes at least as favorable to the alternative hypothesis as the observed outcomes

$$P(n_{11}) = \frac{\binom{n_{1+}}{n_{11}}\binom{n_{2+}}{n_{+1}-n_{11}}}{\binom{n}{n_{+1}}}$$

where

	Col 1	Col 2	Row Total
Row1	n_{11}	n_{12}	n_{1+}
Row2	n_{21}	n_{22}	n_{2+}
Column Total	n_{+1}	n_{+2}	n

5.12 McNemar's Test

Summary

A statistical test used on a dichotomous contingency table with matched pairs of subjects to determine whether the marginal frequencies are equal.

Assumptions

- The sample data has been randomly selected.
- The sample data consists of matched pairs.
- There is one categorical dependent variable with two categories (i.e.,a dichotomous variable) and one categorical independent variable with two related groups.
- The two groups of your dependent variable must be mutually exclusive, in other words no groups can overlap.
- The frequencies are big enough such that $b + c \geq 10$.

Procedure

1. Hypotheses

   $$H_0 : p_b = p_c \quad \text{vs.} \quad H_1 : p_b \neq p_c$$

   Interpretation of the null hypothesis is that the two marginal probabilities for each outcome are the same, i.e. $p_a + p_b = P_a + p_c$ and $p_c + p_d = p_b + p_d$.

2. Test Statistics

   $$X^2 = \frac{(|b - c| - 1)^2}{b + c} \sim \chi_1^2$$

   where follows a chi-squared distribution with 1 degree of freedom

	Test 2 Positive	Test 2 Negative	Row Total
Test 1 Positive	a	b	a+b
Test 1 Negative	c	d	c+d
Column Total	a+c	b+d	n

3. Find the p-value

   - $p - value = P(\chi^2 \geq X^2)$

6 Works Cited

"The R Book" by Michael J. Crawley
"Discovering Statistics Using R by Andy Field, Jeremy Miles, and Zoe Field
"An Introduction to Categorical Data Analysis" by Alan Agresti