Statistical Analysis

John Kloke
Department of Mathematics
Pomona College
Claremont, CA 91711
Phone: 909-621-8712
Fax: 909-607-1247
Email: john.kloke@pomona.edu

Johanna Hardin
Department of Mathematics
Pomona College
Claremont, CA 91711
Phone: 909-607-8717
Fax: 909-607-1247
Email: jo.hardin@pomona.edu

Author order: John Kloke and Johanna Hardin

SUMMARY
In this appendix we have provided an outline for methods used in analyzing molecular biology data. We have given a summary of types of data encountered and the appropriate methods to apply for the questions of interest. Statistical techniques described include the t-test, the Wilcoxon rank sum test, the Man-Whitney-Wilcoxon test, ANOVA, regression, and the Chi-Square test. For each method we have given the appropriate assumptions, the details of the test, and a complete concrete example to follow. We have also discussed related ideas such as multiple comparisons and why correlation does not imply causation.

Keywords: Data Analysis, Hypothesis Testing, Confidence Intervals

Statistical Analysis

INTRODUCTION

This appendix was written as a reference guide for data analysis of biological experiments. It should not replace a course in introductory statistics, but rather serve as a reminder of concepts learned in that course. For those interested in a deeper refresher or more derivations, we recommend the following textbooks:

Introduction to the Practice of Statistics by Moore and McCabe (2006)
Mind on Statistics by Utts and Heckard (2004)
Primer of Biostatistics by Glantz (2005)

Notice that the approaches provided below, summarized in Table 3, do not constitute a complete list of all possible statistics methods. It is possible that your data will need more sophisticated analyses, in which case you should consult a statistician or a post-introductory text book.

We have tried to be clear about which method to use with which type of research question and data. However, if you are unsure, the references above can give a more complete picture of methods and their appropriate uses. Be sure to graph the data and to check assumptions like independence carefully.

Data

Before deciding what test or method to use, you have to know a little about the data you have collected and the questions you'd like to ask. Many methods require independent data. Two data points are independent if knowledge of the first point gives no information about the value of the second point. For example, pre-test and post-test scores on the same individual are not independent. Two scores on two different individuals are independent.

We break data up three different ways:

1. Explanatory vs. Response

- Explanatory - an explanatory variable is used to explain or predict any outcome.
- Response - a response variable measures the outcome of interest.

  Example If you are interested in measuring the difference in two diets, the explanatory variable would be the diets (diet 1 vs. diet 2) and the response variable would be the number of pounds lost by each participant in the study.

2. Numeric vs Categorical

- Numeric - a numeric variable is one which takes numerical values. A numeric variable is either discrete or continuous as described below.
  - Continuous - a numeric variable which is measured on a continuous scale. (height, weight, age, time).
  - Discrete - a numeric variable which may take on only a finite number of values and usually arise from counting situations (number of offspring).
- Categorical variables - a categorical variable does not take on numeric values but rather can be placed into groups or categories such as those described below:
  - Binary or dichotomous - having only two levels like gender or on/off.
  - Multichotomous - having multiple levels like race or grade level.

3. Quantitative vs. Qualitative

- Quantitative (numerical) variables are measured on a numeric scale. These can be continuous (e.g. height) or discrete (e.g. grade in school).
- Qualitative (categorical) variables are measured as categories.

Parameters

Standard symbols will be used to represent parameters of interest are presented in Table 1.

Table 1 Standard Symbols used to Represent Parameters of Interest

Definition	Parameter
Population Mean	μ
Population Median	θ
Population Variance	σ^2
Population Standard Deviation	σ
Population Standard Error of the Mean	$\sigma/\sqrt{n}$

Population Proportion	p
Population Regression Slope	β_1
Population Correlation	ρ

Statistics

Standard symbols will be used to represent statistics of interest, as outlined in Table 2.

Table 2 Standard Symbols used to Represent Statistics of Interest

	Statistic	Definition
Sample Mean	$\bar{x}$	$\frac{1}{n}\sum_{i=1}^{n} x_i$
Sample Variance	s^2	$\frac{1}{n\text{-}1}\sum_{i=1}^{n} (x_i\text{-}\bar{x})^2$
Sample Standard Deviation	S	$\sqrt{s^2}$
Sample Standard Error of the Mean	$s/\sqrt{n}$	$\sqrt{\frac{s^2}{n}}$
Sample Proportion	$\hat{p}$	$\frac{\#\text{ of successes}}{n}$
Sample Regression Slope	b_1	see references
Sample Correlation	r	see references

Table 3 Summary of the Statistical Analysis Discussed

Variables in Dataset	Example	Method(s)	Page
one continuous response variable	Testing average blood pressure (against a known value)	One sample t-test Wilcoxon Rank Sum	7 9
	Estimating average blood pressure	CI for μ/θ	8/10
(optional: a binary pairing variable)	Testing blood pressure before and after treatment to determine if there is a difference	Paired t-test Wilcoxon Rank Sum	10
	Estimating average change in blood pressure before and after treatment	CI for μ	10
one binary explanatory variable one continuous response variable	Testing the difference in average pounds lost for two different diets.	Two sample t-test Mann-Whitney-Wilcoxon	11 14
	Estimating the difference in average pounds lost for two different diets.	CI for difference in μ/θ	11/14
One binary response variable	Estimating the cure rate of some treatment or testing against a known value	Inference for a single p	15

One binary response variable One binary explanatory variable	Testing whether gender and pet ownership are independent	Tests of proportions Chi-square test	16 24
	Estimating the differnece in proportion of men who are pet owners versus women who are pet owners	CI for difference in proportions	16
one multichotomous explanatory variable one continuous response variable	Testing the difference in average GPA for different grade levels.	ANOVA Multiple comparisons	18 19
one continuous explanatory variable one continuous response variable	Predicting weight (response) from height (explanatory)	Simple Linear Regression	20
	Correlating weight and height	Correlation	22
two binary or multichotomous variables	Testing whether race and political party are independent.	Chi-Square test	24

Inference

General Advice for Running Analyses

Typically there are two ways to analyze data: hypothesis testing and confidence intervals. Both methods respond to questions about populations using sample data. So, research questions should ask something related to a population of interest, and the analyses should be done using sample data which was collected.

Hypothesis Testing

In general, there are two hypotheses. A null hypothesis is a statement about the population that nothing interesting is going on. An alternative hypothesis is a statement about the population which would be of interest to the research community. Generally, the alternative hypothesis corresponds to the research question being asked. If a null hypothesis is rejected in favor of the alternative hypothesis, this is an indication that the research question is true. The null hypothesis is rejected if the *p*-value of the test is small enough. *p*-values are discussed below.

In general, every test of hypothesis follows the same basic steps:

1. The null and alternative hypothesis (H_0,H_1 respectively) are formed based on some research question.

   For example, if testing a chemical's ability to act as an antibiotic (i.e., kill bacteria), the null hypothesis would be "The chemical does not kill more bacteria than a control compound," while the alternative hypothesis would be "The chemical kills more bacterial than a control compound."

2. A test statistic is calculated based on a sample of representative data.

   A number of plates would be grown with either the test compound or a control compound, and the number of colonies grown over a set period of time would be counted.

3. A decision to reject or to not reject the null hypothesis is made based on how likely the data are given a true null hypothesis.

   If the p-value obtained by comparing the number of colonies appearing after incubation on plates with antibiotic versus those without is <0.05, the researcher can assume that the

null hypothesis (i.e., "The chemical does not kill significantly more bacteria than a control compound.") is false.

A *p*-value is the probability of seeing the observed sample data or data which is more extreme if the null hypothesis is true. A very small p-value, will suggest that the original assumption (the null hypothesis being true) is false. A null hypothesis is rejected if the *p*-value is less than the *level* of the test, α. Most often a level of α=0.05. Occasionally levels of α=0.10 or α=0.01 are used. The α–level represents how often you make a false positive error. You are responsible for stating the false positive error rate for your experiment. A rejected null hypothesis is often referred to as *statistical evidence* or *significant evidence* of the research question being true (i.e. H_1).

For example, suppose a coin is tossed 10 times, each time resulting in heads. You might reject the hypothesis that the coin is fair (H_0:p=0.5, where p is the true probability of heads) in favor of the hypothesis that the coin is not fair (H_1:$p\neq$0.5). Since the probability of observing results as extreme when the coin is fair (the null hypothesis is true) is small (p-value = 0.002).

If the *p*-value is greater than α=0.05 we are unable to say which of the hypotheses is true. A correct interpretation of a large *p*-value is: "The data do not provide evidence to reject the null hypothesis." A p-value of 0.05 is equivalent to saying that *if the null hypothesis is true,* data like those collected would happen 5% of the time.

Confidence Intervals

Sometimes an estimate of a population parameter is desired instead of testing a particular claim about the population. For example, interest might be in estimating the average blood pressure of women taking Hormone Replacement Therapy (HRT). When estimating a population parameter, using a confidence interval is the appropriate method of analysis. A confidence interval is a set (an interval, in fact) of values which serves as an estimate of a population parameter.

Typically, 90%, 95%, or 99% confidence intervals are used. Consider a 95% confidence interval for some population characteristic, the population mean μ, say. A mathematical derivation shows that out of all possible samples of size *n*, 95% of the intervals will contain the true population value.

The correct interpretation of a 95% confidence interval is, "I am 95% **confident** that the true population parameter lies within the endpoints of the interval." Or, "I am 95% **confident** that the true average blood pressure of women on HRT is between the bounds I have calculated." It would be incorrect to say "There is a 95% **probability** that the true blood pressure of women on HRT lies within the bounds I have calculated." Once the interval has been calculated, there is no more probability associated with the result.

Errors in Inference

As alluded to above, there is no way to know for sure which hypothesis is in fact true. One only has evidence to suggest which is true. At times, because of random variability, the data suggest a hypothesis which is in fact false. When this occurs, it said than an *error* has occurred. There are two possible types of errors which can be made when conducting statistical analysis.

The first type of error, a *type I error* happens when the null hypothesis is incorrectly rejected. Type I errors happen, on average, at a rate of α, usually 0.05. So, 5% of the time when the null hypothesis is in fact true, the data will indicate that the null hypothesis should be rejected.

A *type II error* happens when a false null hypothesis is not rejected. The rate of type II errors cannot be directly controlled and depends on what the actual (unknown) population values are. Because the rate of type II errors is unknown, the statement of "fail to reject the null hypothesis" is preferred to "accept the null hypothesis".

Note that the errors in inference are not the same as measurement variability. A larger sample size will reduce the type II error. Type I error is fixed (typically at 0.05.)

Graphics

A graphical analysis of the data is a very important part of the data analysis process. In addition to giving a pictorial representation of the data, graphics allow the analyst to check any assumptions for the desired statistical method/procedure. Several useful graphical tools exist for this particular purpose. We will primarily use boxplots and scatterplots, but histograms and quantile-quantile plots (not described here, see references in introduction for information on histograms and quantile-quantile plots) are also extremely useful tools.

A boxplot is a representation of a univariate (one variable) set of data. The middle line represents the median (or 50% point), and the outer edges of the box represent the 25% point (lower quartile) and the 75% point (upper quartile.) The whiskers extend to the minimum and maximum values within a certain threshold. If the minimum (or maximum) value is outside of a threshold, the minimum (or maximum) value will be represented by a point. See figures 1-6.

Figure 1 represents the data of birth weight of children whose mothers smoked. 50% of the mothers delivered babies who weighed less than 108.5 oz; 25% of mothers delivered babies who weighed less than 99.5 oz; 75% of mothers delivered babies who weighed less than 121 oz. The smallest baby in the dataset weighed 74oz, and the largest weighed 147oz.

A scatterplot is a representation of a bivariate (two variables) set of data. The x-axis represents the value of the first variable and the y-axis represents the value of the second variable. Each individual is given by a dot in the x-y plane. See figures 7-8.

Figure 7 represents the speeds of 47 independent lizards at 20℃ and 35℃. We can see that there is a positive relationship between the two variables. However, because of the natural variability, we cannot predict perfectly the speed at one temperature from the speed at the other temperature.

Sample Sizes

For each of the methods we have given general sample size requirements. These should be used as a guide and not absolute law. In certain cases larger sample sizes may be required for the methods to be valid. Also, in rare cases, smaller sample sizes may be sufficient. As a general rule, when it comes to sample sizes, think bigger is better. As mentioned in the section on errors, a larger sample size will reduce the type II error. That is, if the true state of nature is "significant

differences", the data are more likely to demonstrate significant differences with a larger sample size. The guidelines for sample sizes are given within each methodological section.

Note, *n* represents the number of independent measurements needed or taken. If there is measurement error, multiple measurements may need to be taken on each independent individual. Doing so will create two sources of variability: within variability and between variability. Analysis of such data is called repeated measures analysis and is outside the scope of this appendix. However, taking the average (or median) of a few repeated measurements will usually provide reasonable data with which to work.

Parametric versus Nonparametric

Parametric methods assume an underlying distribution (usually normality) of the data; nonparametric methods do not assume a known structure of the data. For the majority of topics which we discuss, we have included both parametric and nonparametric methods. In general the nonparametric approaches require fewer assumptions than their parametric counterparts. The parametric approaches are exact when the underlying population is normal, otherwise the inference is approximate. The approximation gets closer as the sample size increases. Though, exact inference for the nonparametric methods exist regardless of sample size, we have chosen to only discuss the approximate versions. As with the parametric approaches, the approximations for nonparametric approaches decline as the sample size increases. Most software packages allow the user to chose between exact and approximate inference for the nonparametric methods. For a thorough treatment of nonparametric methods, see Hollander & Wolfe (1999).

Statistical Software

There are many statistical software packages, in the following table we list only the ones which we are most familiar with. The analyses in this appendix were done using R.

Table 4 Statistical Software

Name	www	Interface	Free or Open Source
Arc	www.stat.umn.edu/arc/software.html	GUI and Command line	Yes
JMP	www.jmp.com/	GUI	No
Minitab	www.minitab.com/	GUI or Command line	No
R	www.r-project.org/	Command line	Yes
SAS	www.sas.com/	Command line	No
S-PLUS	www.insightful.com/products/splus/default.asp	GUI or Command line	No
SPSS	http://www.spss.com/	GUI and Command line	No

THE ONE SAMPLE LOCATION PROBLEM

The one sample location problem is concerned with inference on the center (mean μ or median θ) of a single population. For example, one may want to estimate the average blood pressure of some population or to test if (on average) a certain drug contains the correct amount of active ingredient.

Parametric Procedures

Assumptions

- The response variable is measured on a numeric scale.
- For symmetric populations $n \geq 15$ otherwise $n \geq 30$. Symmetric populations are those that have the same type of distribution of values on the left and right side of the center. For example, income is typically not symmetric because of the extremely large values. Height is typically symmetric because there is the same spread of heights above the mean as below the mean.

Inference

Both the interval estimate and the hypothesis test are based on a t-test which gives a range of plausible values for the mean of the population (interval estimation) or tests a value of interest (μ_0) for the mean of the population (hypothesis testing.)

Interval estimation

A $(1-\alpha)*100\%$ confidence interval for μ is

$$\bar{x} \pm t_{1-\alpha/2,n-1} \frac{s}{\sqrt{n}}.$$

Equation 1

Hypothesis Testing

The *t*-statistic is

$$t^* = \frac{\bar{x}-\mu_0}{\frac{s}{\sqrt{n}}}.$$

Equation 2

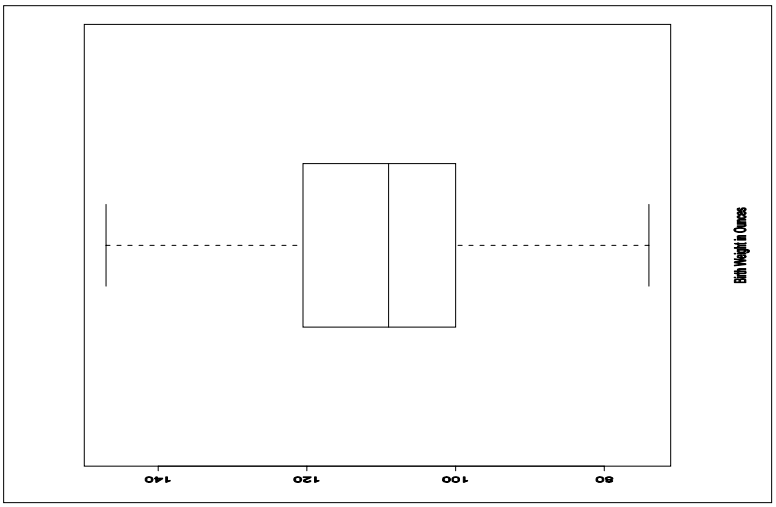
Hypothesis are shown in Table 5

Table 5 Decision rules for the one sample t-test

Hypotheses	Decision Rule	*p*-value
$H_0:\mu=\mu_0$ vs $H_1:\mu\neq\mu_0$	Reject H_0 if $\lvert t^*\rvert > t_{1-\alpha/2,n-1}$	$2*P(t_{n-1}>\lvert t^*\rvert)$
$H_0:\mu\geq\mu_0$ vs $H_1:\mu<\mu_0$	Reject H_0 if $t^* < -t_{1-\alpha,n-1}$	$P(t_{n-1}<t^*)$
$H_0:\mu\leq\mu_0$ vs $H_1:\mu>\mu_0$	Reject H_0 if $t^* > t_{1-\alpha,n-1}$	$P(t_{n-1}>t^*)$

Example

Suppose that we are interested in estimating the mean birth weight of children whose mother admitted to smoking during pregnancy. Suppose a sample of size $n=47$ is taken resulting in a sample mean of $\bar{x}=109.5$ ounces and a sample standard deviation of $s=15.8$ ounces. The actual data are given in the example for the nonparametric test in the next subsection, a plot of the data is in Figure 1.



A birth weight of 116 ounces is considered typical. The research question we wish to answer is: does smoking during pregnancy result in lower birth weights, on average?
That is we wish to test the hypotheses

$$H_0: \mu \geq 116 \text{ vs } H_1: \mu < 116.$$

Equation 3

Using the equation for a *t*-test statistic from above we see that the test statistic we observe for these data is

$$t^* = \frac{109.5-116}{15.8/\sqrt{47}} = -2.82.$$

Equation 4

Since $t^* = -2.82 < -1.68 = -t_{0.95,46}$ we conclude that there is significant evidence to reject the null hypothesis. There is significant evidence that smoking is linked with lower than typical birth weight. (The t-value of -1.68 was computed from a t-table with 46 degrees of freedom and 0.95 probability to the left. Once you have a t-table, from software or a textbook, use degrees of freedom = n-1. The degrees of freedom simply point to the correct table.)

A 95% confidence interval for the true mean birth weight of children whose mothers admitted to smoking during pregnancy is $109.5 \pm 2.01 * 15.8/\sqrt{47}$ or (104.87 ounces, 114.13 ounces). Note that 2.01 is the t-value from a t-table with 46 degrees of freedom at 97.5% (which gives 2.5% error on each side of the confidence interval.) We are 95% confident that the true mean birth weight for all children born to mothers who admitted to smoking during pregnancy is between about 104.87 and 114.13 ounces.

Nonparametric

There are several nonparametric approaches to the one sample location problem. The sign test is a very general test which requires almost no assumptions. We will take a large sample Wilcoxon approach which assumes symmetry of the underlying population. See, for example, Hollander & Wolfe (1999) for Wilcoxon small sample methods as well as additional methods based on the sign test.

Assumptions

- Data are measured on a numeric scale.

- The data are fairly symmetric.
- $n \geq 20$

Hypothesis Testing

The Wilcoxon signed rank test statistic is defined as

$$T^{+} = \sum_{i=1}^{n} R(|x_i - \mu_0|) I(x_i - \mu_0).$$

Equation 5

Where $R(|x_i - \mu_0|)$ denotes the rank of $|x_i - \mu_0|$ among $|x_1 - \mu_0| ... |x_n - \mu_0|$ and $I(x_i - \mu_0)$ is the indicator function which takes value 1 when $x_i - \mu > 0$ and 0 otherwise.

For large samples we may use the test statistic

$$z^{*} = \frac{T^{+} - n(n+1)/4}{\sqrt{n(n+1)(2n+1)/24}}$$

Equation 6

With hypotheses as given in Table 6.

Table 6 Decision Rules for the Wilcoxon signed rank test

Hypotheses	Decision Rule	*p*-value
$H_0: \theta = \theta_0$ vs $H_1: \theta \neq \theta_0$	Reject H_0 if $\lvert z^{*} \rvert > z_{1-\alpha/2}$	$2 * P(Z > \lvert z^{*} \rvert)$
$H_0: \theta \geq \theta_0$ vs $H_1: \theta < \theta_0$	Reject H_0 if $z^{*} < -z_{1-\alpha}$	$P(Z < z^{*})$
$H_0: \theta \leq \theta_0$ vs $H_1: \theta > \theta_0$	Reject H_0 if $z^{*} > z_{1-\alpha}$	$P(Z > z^{*})$

Interval Estimation

An interval estimate for the population median, based on the Wilcoxon signed rank, is the interval of values such that the null hypothesis is not rejected. See, for example, Hollander & Wolfe (1999).

Example

We use the same example as we did for the parametric analysis. The sorted data of sample brith weights are presented below.

```
  74  81  87  88  88  89  89  92  98  98  99 100 100 100 102 104
104 104 106
 106 107 107 108 109 109 111 111 113 113 113 115 117 117 119 120
121 121 121
 123 123 124 125 130 132 141 142 147
```

Judging from the boxplot (see figure 1), the assumption of symmetry seems quite valid. That is, there are just as many points above as below the median; also the distance from the 25% to the minimum value is about the same as the distance from the 75% to the maximum value. The plot tells us that the assumptions for the statistical method are not violated. The value of the Wilcoxon signed rank test statistic for these data is $T^{+}=303$. The large sample standardized test statistic is

$$z^* = \frac{303\text{-}47*48/4}{\sqrt{47*48*95/24}} = -2.76.$$

Equation 7

Using a $-z_{0.95}=-1.96$ as our critical value we see that we can reject the null hypothesis and conclude that smoking is related to lower birth weight. Also, *p*-value = 0.0029.

The software package R reports (104.5 ounces, 114.0 ounces) as a interval estimate of the true median birth weight of children whose mothers admitted to smoking based on the Wilcoxon signed rank test.

Paired Designs

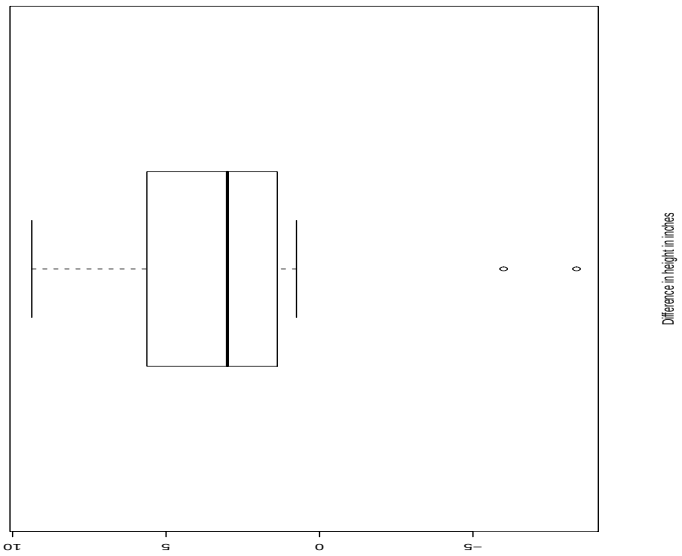
A paired design, sometimes referred to as matched pairs, occurs when two repeated measurements are taken of the same individual or experimental unit or when measurements are taken on pairs of subjects or experimental units which are matched somehow. The parametric analysis of such data is often called a paired t-test. For other methods (including the nonparametric analysis), there is often a "paired" option in the statistical software. For two examples, measurements of siblings on two different treatments are matched; a baseline measurement on some individual and then a measurement on the same individual taken after some treatment is also matched.

A paired data analysis is a one sample analysis performed on the differences of the response variables. The null and alternative hypotheses almost always address the question of whether the pairs are different; that is, do they have a difference of zero.

Analysis

These data where taken from Hettmansperger & McKean (1998). The data consist of 15 pairs of heights in inches of cross-fertilized and self-fertilized plants, each pair grown in the same plot (Figure 2). The data are given below.

```
cross:
 23.500 12.000 21.000 22.000 19.125 21.550 22.125 20.375 18.250
21.625 23.250 21.000 22.125 23.000 12.000
self:
  17.375 20.375 20.000 20.000 18.375 18.625 18.625 15.250 16.500
18.000 16.250 18.000 12.750 15.500 18.000
differences:
  6.125 -8.375  1.000  2.000  0.750  2.925  3.500  5.125  1.750
3.625 7.000  3.000  9.375  7.500 -6.000
```



Parametric analysis

The output from the software package R is given below.

```
        One Sample t-test

data:  diff
t = 2.1506, df = 14, p-value = 0.04946
alternative hypothesis: true mean is not equal to 0
95 percent confidence interval:
 0.007114427 5.232885573
```

As we can see the *t*-test is only marginally significant at the α=0.05 level (*p*-value = 0.04946).

Nonparametric analysis

Again, we include the output from the software package R.

```
        Wilcoxon signed rank test

data:  diff
V = 96, p-value = 0.04089
alternative hypothesis: true mu is not equal to 0
95 percent confidence interval:
 0.4999872 5.2125081
```

The results for the Wilcoxon analysis are similarly borderline (*p*-value = 0.04089).

THE TWO INDEPENDENT SAMPLES LOCATION PROBLEM

The two sample location problem is concerned with inference on the change in means (parametric approach) or the change in medians (nonparametric approach) between two treatment or experimental groups. For example, one might be interested in testing if the LDL cholesterol levels of a group of treated quail are lower than a group of untreated quail. Or one might be interested in estimating how much taller, on average, adult males are than adult females. Note if a

population is symmetric then the population mean is equal to the population median, which is the situation represented in the plots below (see figure 3) and discussed in the next paragraph.

The boxplots below (see figure 3) represent two typical situations for the two sample location problem. The plot on the left represents a case where either the MWW test for medians or the pooled t-test for means is appropriate. The plot on the right represents a case where the unpooled t-test is appropriate.

[*Figure 3 here]

Data

Suppose we are interested in comparing two populations, *X* and *Y*. We take a sample of observations from each of the two groups or under the two experimental conditions. Let $x_1,...x_m$ be a random sample of size *m* from population *X*. Let $y_1,...y_n$ be a random sample of size *n* from population *Y*.

Parametric

There are two types of t-tests for the two sample problem, the pooled and unpooled. The pooled requires the distributions differ only in location but are similar in terms of variability. The unpooled requires the shapes of the two populations be the same but does not require the variability of the two populations to be the same. When the assumptions of the pooled analysis are met, the pooled analysis results in a more efficient (that is, fewer samples needed) analysis than the unpooled counterpart.

In both cases μ_X represents the true mean of population *X*, and μ_Y represents the true mean of population *Y*. Further $\bar{x}$ and s_x denote the sample mean and sample standard deviation of the *m* observations $x_1,...,x_m$. Likewise $\bar{y}$ and s_y denote the sample mean and sample standard deviation of *n* observations $y_1,...,y_n$.

Unpooled inference

Assumptions

- The response variable is numeric.
- The two distributions have the same shape.
- Two independent samples are drawn from the two distributions.
- If the distributions are symmetric, you should have at least $n>15$ in each group. If the distributions are not symmetric, you should have at least $n>30$ in each group.

Interval estimation

A $(1-\alpha)*100\%$ confidence interval for the difference in the population means $(\mu_X-\mu_Y)$ is

$$\bar{x}-\bar{y} \pm t_{1-\alpha/2,df}\sqrt{\frac{s_x^2}{m}+\frac{s_y^2}{n}}$$

Equation 8

Most software packages use a *Satterthwaite* approximation to determine the degrees of freedom, *df*, leads to a conservative analysis. Degrees of freedom simple point to the correct table to us. Here, we can also use df=minimum (n – 1, m-1) as an approximation.

Hypothesis testing

The unpooled *t*-statistic is

$$t^* = \frac{\bar{x} - \bar{y}}{\sqrt{\frac{s_x^2}{m} + \frac{s_y^2}{n}}}$$

Equation 9

With hypotheses shown in Table 7.

Table 7 Decision rules for the two independent sample t-test with unpooled variance

Hypotheses	Decision Rule	*p*-value
H_0:μ_x=μ_y vs H_1:μ_x≠μ_y	Reject H_0 if $\lvert t^*\rvert > t_{1-\alpha/2,df}$	$2*P(t_{df}>\lvert t^*\rvert)$
H_0:μ_x≥μ_y vs H_1:μ_x<μ_y	Reject H_0 if $t^* < -t_{1-\alpha,df}$	$P(t_{df}<t^*)$
H_0:μ_x≤μ_y vs H_1:μ_x>μ_y	Reject H_0 if $t^* > t_{1-\alpha,df}$	$P(t_{df}>t^*)$

Where *df* is as described above. Sometimes the difference in the two population means is called the *shift* and is denoted by $\Delta=\mu_X-\mu_Y$.

Example

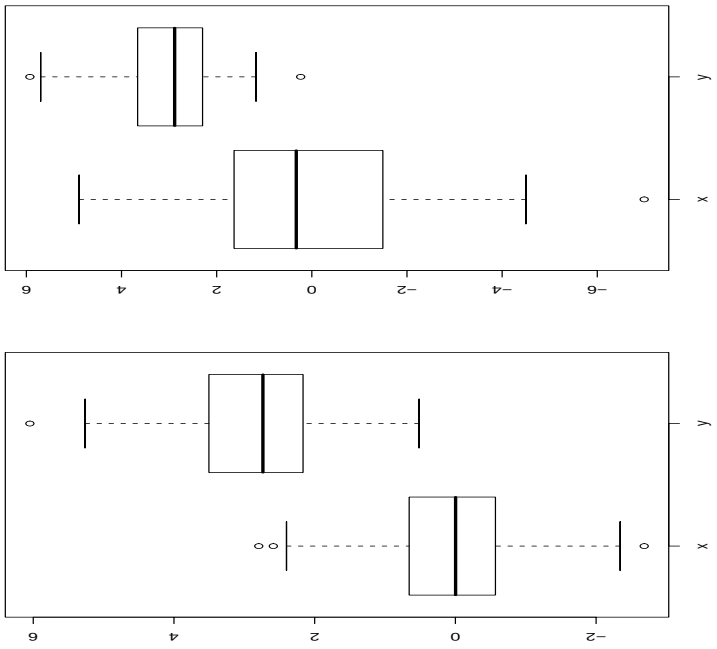
Suppose we are interested in comparing the cholesterol levels of two groups of quail, one group who receives a treatment designed to lower cholesterol and one which has not. A random sample of n=30 quail were selected for control, and a sample of m=20 quail were selected for treatment. The data are presented below with comparison boxplots shown in (Fig. 4). These indicate that the assumption of a common variance is not valid so we will perform an unpooled analysis.

```
control
 44 51 50 52 41 69 56 67 45 37 40 44 46 55 50 60 65 53 46 38 58
58 48 29 65 62 59 45 61 56
treated
 50 49 46 58 50 59 58 48 44 41 49 47 49 48 45 48 49 51 52 52
```

A 95% confidence interval for the mean difference in the two populations μ_x-μ_y (control - treatment) is (-2.14 mg/dL, 6.18 mg/dL). To test the hypothesis that the treatment is effective in lowering cholesterol we test

$$H_0:\mu_x \le \mu_y \text{ vs } H_1:\mu_x > \mu_y$$

The *t*-test statistic is t^*=0.977, which is not significant (*p*-value=0.1670). So, based on these data we do not have significant evidence to conclude that the treatment is effective in lower cholesterol.



Pooled inference

Assumptions

- The response variable is numeric.
- The two distributions have the same shape and variance.
- Two independent samples are drawn from the two distributions.
- If the distributions are symmetric, you should have at least $n>15$ in each group. If the distributions are not symmetric, you should have at least $n>30$ in each group.

Interval estimation

An interval estimate for the difference in the population means (μ_x-μ_y) is

$$\bar{x}\text{-}\bar{y} \pm t_{1-\alpha/2,m+n-2} s_p \sqrt{\frac{1}{m}+\frac{1}{n}}$$

Equation 10

where

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

Equation 11

is a pooled estimate of the population standard deviation (σ) of the two groups.

Hypothesis testing

The pooled *t*-statistic is

$$t^* = \frac{\bar{x}\text{-}\bar{y}}{s_p \sqrt{\frac{1}{m}+\frac{1}{n}}}$$

Equation 12

where s_p is as described above for the interval estimate.

Hypotheses are shown in Table 8

Table 8 Decision rules for the two independent sample t-test with pooled variance

Hypotheses	Decision Rule	*p*-value
H_0:$\mu_x=\mu_y$ vs H_1:$\mu_x\neq\mu_y$	Reject H_0 if $\lvert t^*\rvert>t_{1-\alpha/2,m+n-2}$	$2*P(t_{m+n-2}>\lvert t^*\rvert)$
H_0:$\mu_y\geq\mu_z$ vs H_1:$\mu_y<\mu_z$	Reject H_0 if $t^*<-t_{1-\alpha,m+n-2}$	$P(t_{m+n-2}<t^*)$
H_0:$\mu_y\leq\mu_z$ vs H_1:$\mu_y>\mu_z$	Reject H_0 if $t^*>t_{1-\alpha,m+n-2}$	$P(t_{m+n-2}>t^*)$

Example

We wish to test the hypothesis that men are taller than women. Suppose that the heights, in inches, of a sample of m=20 women are

```
68 60 61 60 60 59 63 59 63 65 70 61 67 66 61 65 68 63 64 67
```

Suppose that the heights, in inches, of a sample of n=22 men are

```
68 65 66 68 72 70 73 72 71 70 69 71 66 72 69 73 70 72 65 69 70 72
```

Let μ_x denote the true mean height for women and μ_y denote the true mean height for men. The comparison boxplots (Fig. 4) for these data indicate that the assumption of equal variances is appropriate. The hypotheses are

$$H_0:\mu_x\geq\mu_y \text{ vs } H_1:\mu_x<\mu_y.$$

Equation 13

From software we obtain the *t*-test statistic, t^*=-6.8054, since the value observed is so extreme we may say that the *p*-value is approximately zero and conclude that indeed men are taller than women. The confidence interval for μ_x-μ_y reported by the software is (-8.02 inches, -4.35 inches). This says that, on average, women are between about 4 and a half and 8 inches shorter than men with 95% confidence.

Nonparametric

Though the Mann-Whitney-Wilcoxon (MWW) test is general in that it tests if one population is larger than another, it is often used to test for differences in the medians of the two populations in which case the populations are assumed to have the same shape. For example, one might want to know which bacterial strain, *X* or *Y* has a longer lifespan. The distributions of the two random variables might be quite different in shape, etc. A rejected null hypothesis says that *Y* tends to outlive *X*. We will discuss only the inference on medians.

Inference

Assumptions

- The two samples are independent.
- The response variable is numeric.
- The two population differ possibly only in location.
- The sample sizes are sufficiently large: $m\geq10$ and $n\geq10$.

Test of hypothesis

The MWW test statistic is

$$T=\sum_{i=1}^{n} R(x_i)$$

Equation 14

where $R(x_i)$ denotes the rank of x_i among the combined sample $x_1,...,x_n,y_1,...,y_m$. The large sample standardized test statistic is

$$z^*=\frac{T-m(n+m+1)/2}{\sqrt{nm(n+m+1)/12}}$$

Equation 15

Hypotheses are shown in Table 9

Table 9 Decision rules for the Mann-Whitney-Wilcoxon test

Hypotheses	Decision Rule	*p*-value
$H_0:\theta_x=\theta_y$ vs $H_1:\theta_x\neq\theta_y$	Reject H_0 if $\|z^*\|>z_{1-\alpha/2}$	$2*P(Z>\|z^*\|)$
$H_0:\theta_x\geq\theta_y$ vs $H_1:\theta_x<\theta_y$	Reject H_0 if $z^*<-z_{1-\alpha}$	$P(Z<z^*)$
$H_0:\theta_x\leq\theta_y$ vs $H_1:\theta_x>\theta_y$	Reject H_0 if $z^*>z_{1-\alpha}$	$P(Z>z^*)$

Interval estimation

Similar to the one sample location problem, an interval estimate based on the MWW test is the set of values for which the null hypothesis is not rejected.

Example

Using the height data from pooled inference example (page 13), comparison boxplots (Fig. 5) show that the assumption of a *shift model* is appropriate. The appropriate hypotheses to answer if men have a median height which is greater than the median height of all women are

$$H_0:\theta_x\geq\theta_y \text{ vs } H_1:\theta_x<\theta_y.$$

Equation 16

The value of the rank sum statistic, adjusted for ties, is T=244 and the standardized version is z^*=-4.684. Since the *p*-value = 1.406 is small (less than α=0.05) the null hypothesis is rejected. There is significant evidence that men are taller then women.

A 95% confidence interval for the shift $\theta_x-\theta_y$ is (-9 inches, -4 inches).

THE ONE SAMPLE PROPORTION PROBLEM

The one sample proportion problem is concerned with inference on one population proportion or the probability of success of some process. For example, we might be interested in the success rate of a new treatment for some disease.

Data

Let x be the number of successes out of n. Then $\hat{p}=\frac{x}{n}$ is a point estimate of p.

Assumptions

- The observations are independent.
- The sample size n is large enough so that $n*p_0>5$ and $n(1-p_0)>5$ for testing.

• The sample size n is large enough so that $n*\hat{p}>5$ and $n(1-\hat{p})>5$ for interval estimation.

Interval estimation

A (1-α)*100% confidence interval for p, the population proportion, is

$$\hat{p}\pm z_{1-\alpha/2}\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}.$$

Equation 17

Hypothesis testing

The test statistic is

$$z^*=\frac{\hat{p}-p_0}{\sqrt{\frac{p_0(1-p_0)}{n}}}$$

Equation 18

Hypotheses are shown in Table 10

Table 10 Decision rules for the one sample test of a proportion

Hypotheses	Decision Rule	p-value
$H_0:p=p_0$ vs $H_1:p\neq p_0$	Reject H_0 if $\lvert z^*\rvert>z_{1-\alpha/2}$	$2*P(Z>\lvert z^*\rvert)$
$H_0:p\geq p_0$ vs $H_1:p<p_0$	Reject H_0 if $z^*<-z_{1-\alpha}$	$P(Z<z^*)$
$H_0:p\leq p_0$ vs $H_1:p>p_0$	Reject H_0 if $z^*>z_{1-\alpha}$	$P(Z>z^*)$

Example

It is well known that the cure rate for the standard treatment of a certain disease is 0.45. A new treatment has been developed and we wish to test if the new treatment has a higher cure rate. The hypotheses of interest are

$$H_0:p\leq 0.45 \text{ vs } H_1:p>0.45.$$

Equation 19

Out of a sample of n=99 patients with the disease, x=46 of them were cured so that $\hat{p}=\frac{46}{99}=0.4646$ is a point estimate of the true cure rate. An 95% confidence interval estimate is $0.4646\pm1.96\sqrt{0.4646*0.5354/99}=0.4646\pm0.0982$ or (0.3665,0.5629). That is the true cure rate for the new drug is between about 37% and 56% with 95% confidence. The test statistic is

$$z^*=\frac{0.4646-0.45}{\sqrt{\frac{0.4646*0.5354}{99}}}=0.2913.$$

Equation 20

The p-value of the test is p-value=0.3854, which indicates that there is not significant evidence to say that the new disease is significantly better at curing the disease than the current standard treatment.

The researcher may choose to do only an interval estimate or a hypothesis or both. The determination should be made based on the research question of interest. If interest is in getting an estimate of the rate, an interval estimate should be created. If interest is in testing a particular plausible value, a hypothesis test should be done. If the problem is new to the literature, both methods might be applied.

THE TWO INDEPENDENT SAMPLES PROPORTION PROBLEM

The two sample proportion problem is concerned with inference on the difference between two population proportions or the difference in the success rates of two treatments. For example if treatment 1 has success rate p_1 and treatment 2 has success rate p_2 (both unknown, both needing to be estimated from the data), we may want to test if treatment 2 has a higher success rate than treatment 1. We will provide information on the degree of difference of success rate.

Data

Let x_1 be the number of successes out of n_1 for treatment 1. Let x_2 be the number of successes out of n_2 for treatment 2. Then $\hat{p}_1=\frac{x_1}{n_1}$ and $\hat{p}_2=\frac{x_2}{n_2}$ are the sample proportions.

Assumptions

- The two samples are independent.
- The sample sizes n_1 and n_2 are large enough so that $n_i*\hat{p}_i>5$ and $n_i(1-\hat{p}_i)>5$.

Interval estimation

An interval estimate for the difference in the two populations proportions (p_1-p_2) is

$$\hat{p}_1-\hat{p}_2\pm z_{1-\alpha/2}\sqrt{\frac{\hat{p}_1(1-\hat{p}_1)}{n_1}+\frac{\hat{p}_2(1-\hat{p}_2)}{n_2}}.$$

Equation 21

Hypothesis testing

The test statistic is

$$z^*=\frac{\hat{p}_1-\hat{p}_2}{\sqrt{\hat{p}(1-\hat{p})\left(\frac{1}{n_1}+\frac{1}{n_2}\right)}}$$

Equation 22

where

$$\hat{p}=\frac{x_1+x_2}{n_1+n_2}.$$

Equation 23

Hypotheses are shown in Table 11.

Table 11 Decision rules for the two independent sample test of proportions

Hypotheses	Decision Rule	*p*-value
$H_0:p_1=p_2$ vs $H_1:p_1\neq p_2$	Reject H_0 if $\lvert z^*\rvert>z_{1-\alpha/2}$	$2*P(Z>\lvert z^*\rvert)$
$H_0:p_1\geq p_2$ vs $H_1:p_1<p_2$	Reject H_0 if $z^*<-z_{1-\alpha}$	$P(Z<z^*)$
$H_0:p_1\leq p_2$ vs $H_1:p_1>p_2$	Reject H_0 if $z^*>z_{1-\alpha}$	$P(Z>z^*)$

Examples

Suppose we are interested in knowing if more patients respond to a new treatment (treatment 1) than a standard treatment (treatment 2). That is, we are interested in comparing the response rates of the two treatments, p_1 and p_2. In a sample of n_1=100 subjects on treatment 1 there were x_1=33 responders. In a sample of n_2=110 subjects on treatment 2 there were x_1=20 responders. The point estimates for the response rates are $\hat{p}_1=\frac{33}{100}=0.330$ and $\hat{p}_2=\frac{20}{110}=0.182$. A 95% confidence interval is

$$0.330-0.182 \pm 1.96\sqrt{\frac{0.330*0.670}{100}+\frac{0.182*0.818}{110}}$$

Equation 24

or (0.031.0.265). With 95% confidence, treatment 1 has a response rate with is between 3% and 26.5%, higher than the response rate of treatment 2.

Since we are interested in knowing if treatment 1 has a higher response rate than treatment 2 we perform a lower tail test (that is, we test whether treatment 1 has a higher response rate than treatment 2, instead of simply testing that the response rates are different):

$$H_0: p_1 \leq p_2 \text{ vs } H_1: p_1 > p_2$$

Equation 25

The test statistic we observe is

$$z^*=\frac{0.330-0.182}{\sqrt{0.252*0.748\left(\frac{1}{100}+\frac{1}{110}\right)}}=2.477$$

Equation 26

since

$$\hat{p}=\frac{33+20}{100+110}=0.252.$$

Equation 27

Since $z^*=2.477>1.645=z_{0.95}$ we would reject the null hypothesis and conclude that treatment 2 is probably doing better. Also p-value=0.0066<0.05=α.

ANOVA

Analysis of Variance (ANOVA) is used for testing for a difference in the means of three or more groups. For example, one might want to know if the mean blood pressure is the same or different across several ethnic groups. One could also use ANOVA to test for differences in blood pressure for three different treatment groups. The main assumption is that the variability across the groups is the same. If this assumptions is not met, then pairwise *t*-tests are appropriate, though a correction for multiple comparisons should be made (see below). For a thorough treatment of ANOVA and related topics, see Kutner, et. al. (2005).

Inference

Assumptions

- The response variable is numeric.
- The explanatory variable is categorical.

• Variability across groups is the same ($\sigma_1=...=\sigma_k=\sigma$).

• If the distribution is symmetric, you should have at least $n>15$ in each group. If the distribution is not symmetric, you should have at least $n>30$ in each group.

Hypothesis testing

$$H_0:\mu_1=\mu_2=...=\mu_k$$

Equation 28

$$H_1:\mu_i\neq\mu_j \text{ for some } i\neq j$$

The test statistic for this Analysis of Variance test of means is beyond the scope of this appendix. Computer software will provide you with a test statistic and a p-value. Reject H_0 if the p-value is less than the specified level of significance (usually $\alpha=0.05$ or 0.01.)

Confidence intervals

If, after running the hypothesis test, you want to find a confidence interval for a difference in means, refer to the section on Two Sample Location Problem, Parametric, Pooled Inference. However, instead of using the pooled estimate of variance (s_p^2), use the mean squared error, MSE, from the ANOVA output. Additionally, the degrees of freedom for the t-multiplier will be the degrees of freedom in the residual (or error) row from the ANOVA table.

Example

Suppose we want to compare the average life of three strains of yeast. The first strain (C) is a control strain; the second strain (D) has a transcription factor gene deleted; and the third strain (A) has an extra copy of the same gene added. We want to know whether modifying the transcription factor gene changes the average lifespan (in generations) of the strain of yeast. We collect 80 data points, summary statistics are shown in Table 12.

Table 12 Average Life Span of Three Strains of Yeast

	Control	Deletion	Addition
sample size	30	27	23
Average lifespan	15.89	13.85	16.98
st. dev.	2.84	2.85	2.65

As seen in the data table as well as the boxplots (Fig. 6), the data are consistent with the assumptions (numeric, symmetric, constant variance.)

[*Figure 6 near here]

Using statistical software, the ANOVA table shown in Table 13 is obtained

Table 13 ANOVA Table Obtained from Yeast Life Span Data

	df	Sum Sq	Mean Sq	F test stat	p-value
Groups	2	128.7	64.35	8.259	0.0005631
Residual	77	599.95	7.79		

The ANOVA table gives a small p-value which leads to rejection of the null hypothesis that the average lifespans are the same across all three groups. However, it is not clear which of the 3 groups are different. Confidence intervals for each of the pairwise differences of population

means are calculated using the formula from shown in Equation 29. An interval estimate for the difference in two population means (μ_1-μ_2) is

$$\bar{x}_1\text{-}\bar{x}_2 \pm t_{\alpha^*/2,df}\sqrt{\text{MSE}}\sqrt{1/n_1+1/n_2}$$

Equation 29

Here, MSE = 7.79 and df=77. Confidence intervals for the three groups at the α^*=0.05/3=0.017 level (98.3% confidence) are created; see the multiple comparisons section below. The multiplier is $t_{.0083,77}$=2.45 .

Table 14 Estimates of pairwise differences of mean lifespan for different strains of yeast

Parameter	Estimate	confidence interval
μ_C-μ_D	2.045	(0.233 generations, 3.857 generations)
μ_C-μ_A	-1.087	(-2.980 generations, 0.806 generations)
μ_A-μ_D	3.132	(1.194 generations, 5.070 generations)

Because neither of the confidence intervals for the deleted strain overlap zero, the average lifespan for the deleted strain is significantly different from either the wildtype or the addition strain with a family-wise confidence rate of 95%. However the wildtype is not significantly different from the addition strain.

Multiple Comparisons

When computing one level α hypothesis test (in any setting, not just ANOVA), the probability of rejecting H_0 when H_0 is really true is set at α. Similarly, when finding a confidence interval, the probability of missing the true parameter is also α (or one minus the level of confidence.) Notice that in 100 hypothesis tests where nothing interesting is going on (no alternative hypothesis is true), on average five true null hypotheses will be rejected when using α=0.05. Similarly, out of 100 95% confidence intervals, on average five of them will not overlap the true parameter of interest. This problem of multiple comparisons happens in any situation where there is more than one hypothesis test or confidence interval.

When finding confidence intervals for differences of means after running an ANOVA test, multiple comparisons should be considered. If we have 3 experimental treatment, we might want to find 3 confidence intervals for differences of means: 1 vs. 2, 1 vs. 3, and 2 vs. 3. There are many ways to adjust for multiple comparisons; the Bonferroni correction is discussed here. Note that the Bonferroni adjustment is quite conservative.

Instead of controlling the type I error rate (rejecting a true null hypothesis) for each test, the Bonferroni adjustment controls the familywise error rate. The familywise error rate is the probability of rejecting one or more true hypotheses when doing numerous hypothesis tests. (For confidence intervals, a familywise error rate would be the probability of at least one confidence interval failing to capture the true parameter of interest.)

When running k hypothesis tests, the level of significance should be adjusted to $\alpha^* = \alpha/k$. Then, α^* is used as the significance level for each individual test, and α is the conservative familywise error rate. For example, if 12 hypothesis tests are to be performed, and a familywise error rate of

no more than 0.05 is desired, each test should be performed at level α=0.05/12=0.00417. That is, for each of the *k* hypothesis tests, reject the null hypothesis the *p*-value is less than 0.05/12 = 0.00417. In order to control the familywise error rate for the 12 confidence intervals, create 99.583% confidence intervals.

Nonparametric

An extension of the Mann-Whitney-Wilcoxon test (p. 14) is the Kruskal-Wallis test. Kruskal-Wallis is a test for medians.

Assumptions

- The *k* samples are independent.
- The response variable is numeric.
- The *k* populations differ only in location.

$$H_0:\theta_1=\theta_2=...=\theta_k \text{ vs } H_A:\theta_i\neq\theta_j \text{ for some } i\neq j$$

Equation 30

Example

Continuing the yeast example (p.18) and using software to obtain the p-value = 0.0011. As with the parametric approach, the null hypothesis is rejected.

REGRESSION

Regression analysis is a statistical technique designed to fit a straight line through a cloud of points. Let the response variable be called *y* and the explanatory variable be called *x*, then the regression analysis will find b_o (the y-intercept) and b_1 (the slope) such that

$$y=b_o+b_1x$$

Equation 31

is the best fit line between *x* and *y*. For data, we have a random sample of *n* pairs of observations: $(x_1,y_1),(x_2,y_2),...,(x_n,y_n)$. For a thorough treatment of regression analysis, see Kutner, et. al. (2005).

Inference

Assumptions

- *Y* must be a continuous and numerical variable.
- *X* must be a numerical variable.
- *Y* should be normally distributed around the regression line.
- The variability around the regression line should be relatively constant for all values of *X*.
- The *Y* values should all be independent observations.
- A sample of $n>25$ pairs of observations is needed.

In most situations, the hypotheses of interest are whether or not *Y* changes linearly with *X* (i.e., whether *X* and *Y* are correlated.) If the slope of the regression line is close to zero, we say that an increase in *X* is not statistically associated with any linear change in *Y*, or *X* and *Y* are not significantly correlated.

Let β_1 be the slope of the regression on the population of *X*s and *Y*s.

Interval estimation

A (1-α)*100% confidence interval for β_1 is

$$b_1 \pm t_{(1-\alpha/2,n-2)} s(b_1)$$

Equation 32

where b_1 and $s(b_1)$ are computed using statistical software (formulae can also be found in the introductory statistics texts listed above.)

Hypothesis testing

One-sided and two-sided tests concerning the populations slope, β_1, are constructing based on the test statistic

$$t^* = \frac{b_1}{s(b_1)}$$

Equation 33

Table 15 contains the decision rules for the three possible cases, with the probability of making a type I error controlled at α.

Table 15 Decision rules for tests of the slope for linear regression

Hypotheses	Decision Rule	*p*-value
H_0:β_1=0 vs H_1:$\beta_1 \neq 0$	Reject H_0 if $\lvert t^* \rvert > t_{1-\alpha/2,n-2}$	$2P(t_{n-2} > \lvert t^* \rvert)$
H_0:$\beta_1 \geq 0$ vs H_1:$\beta_1 < 0$	Reject H_0 if $t^* < -t_{1-\alpha,n-2}$	$P(t_{n-2} < t^*)$
H_0:$\beta_1 \leq 0$ vs H_1:$\beta_1 > 0$	Reject H_0 if $t^* > t_{1-\alpha,n-2}$	$P(t_{n-2} > t^*)$

Example

Consider a situation investigating the linear relationship between lizard speed and outside temperature. In particular, interest is in determining whether a change in lizard speed at 20°C (the explanatory variable, *X*) is linearly associated with lizard speed at 35°C (the response variable, *Y*). Forty seven data points are collected (Fig. 7), the first 12 points (in m/s) are shown in Table 16.

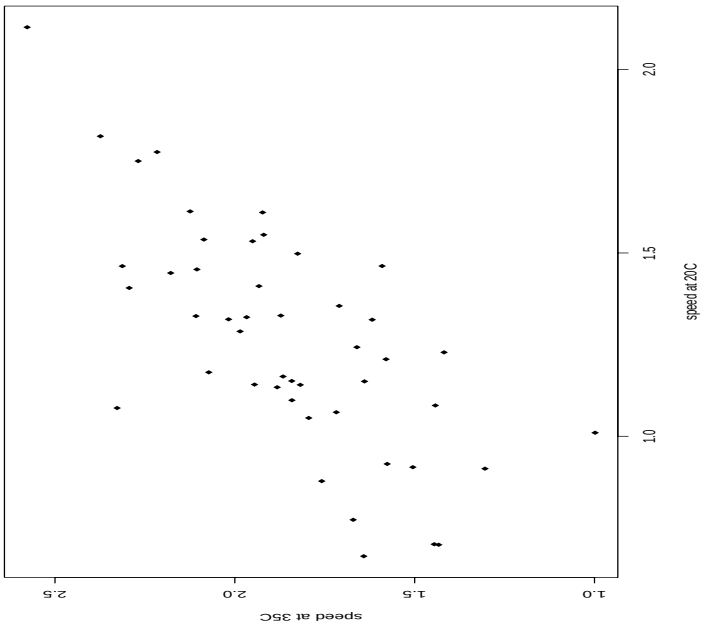


Table 16 First Twelve Data Points Collected Correlating Lizard Speed and Outside Temperature

Lizard #	1	2	3	4	5	6	7	8	9	10	11	12
Speed at 20° C	0.96	1.05	1.18	1.32	1.37	1.59	1.39	1.23	1.74	1.13	1.06	1.30
Speed at 35° C	1.85	1.31	1.73	1.42	2.18	1.79	1.85	1.92	2.33	1.78	1.48	2.03

Analysis

Checking assumptions: the points are reasonably spread out around a line with constant variance across the explanatory variable. The relevant test is:

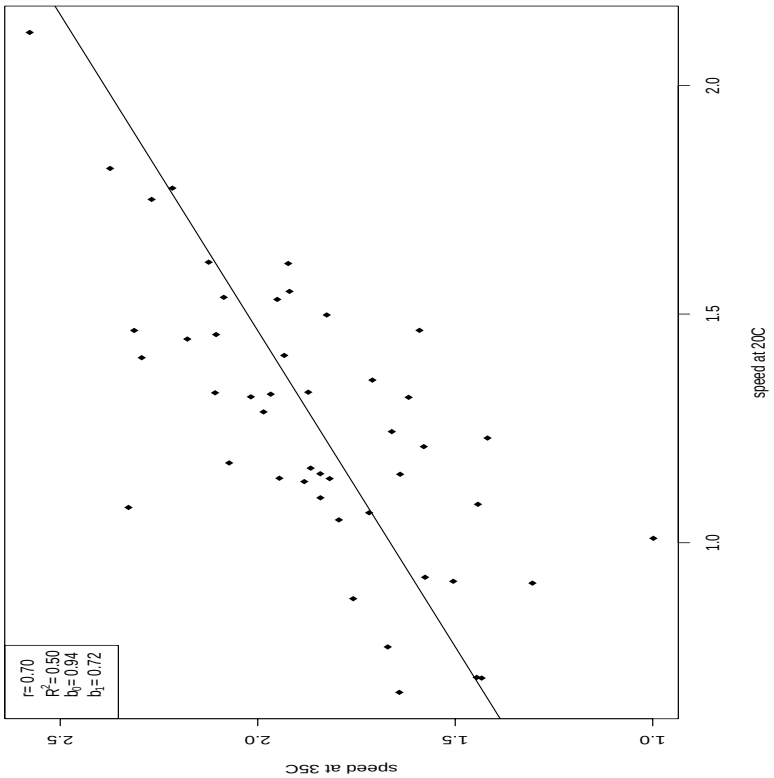
H_0:$\beta_1 \leq 0$ no linear relationship or negative linear relationship

Equation 34

H_1:$\beta > 0$ positive linear relationship

Equation 35

From the software, $b_1 = 0.723$ and $s(b_1) = 0.109$ which give $t^* = 6.660$. The critical value is $t_{1-\alpha,n-2} = t_{.95,45} = 1.68$, so the null hypothesis is rejected, there is significant evidence to conclude a positive linear relationship between speed at 20°C and 35°C. Additionally, the associated p-value is 0.0000033 which says that if the two speeds were not linearly associated, we would only see data like we got 0.00033% of the time, which is very unusual. The p-value is smaller than 0.05, again, confirming our decision to reject H_0. Figure 8 shows the resulting regression line.



The appropriate 95% CI for β_1 is:

$$b_1 \pm t_{0.975,10} * s(b_1)$$

Equation 36

$$0.723 \quad \pm \quad 2.01 * 0.109$$

$$(0.504 \text{ m/s}, 0.942 \text{ m/s})$$

We are 95% confident that for every increase in 1 m/s in speed at 20˚C the speed at 35˚C increases by between 0.504 m/s and 0.942 m/s.

Note that regression analysis should only be done for x-axis values that are within the constraints of the data. It would not make sense to predict the running speed of a lizard who runs 10m/s or 0m/s as neither of those explanatory values are realistic. Similarly, it is not wise to predict values outside the x-data because one is never certain that the same linear model holds for all possible speeds.

CORRELATION (r)

Related to linear regression are ideas of correlation. Correlation (typically denoted by "r") measures the degree of linear association between two continuous variables. If two variables have a perfect positive linear relationship (e.g., miles and kilometers), they have a correlation of one (1). If two variables have a perfect negative linear relationship (e.g., number right and number wrong on an exam), they have a correlation of negative one (-1). If two variables are not correlated (e.g., age and beak length for adult chickens), they have a correlation of zero (0).

Inference

Although formulae exist for calculating the correlation between two continuous variables, we will use statistical software for calculations. The value of r gives an idea of how far the data

points fall from a line. As mentioned above, the sign of r (either positive or negative) gives an indication of the relationship (either positive or negative) between the two variables. However, r^2 is often reported and interpreted as the proportion of variability explained by the regression model. Interest is usually in testing whether the population correlation (ρ) is significantly different from zero.

Hypothesis testing

The test for correlation at zero is the same test for the population slope at zero. Therefore, the summary of the correlation test is shown in Table 17.

Table 17 Decision rules for tests of the population correlation

Hypotheses	Decision Rule	*p*-value
H_0:ρ=0 vs H_1:$\rho\neq 0$	Reject H_0 if $\lvert t^*\rvert > t_{1-\alpha/2,n-2}$	$2P(t_{n-2}>\lvert t^*\rvert)$
H_0:$\rho\geq 0$ vs H_1:$\rho<0$	Reject H_0 if $t^*<-t_{1-\alpha,n-2}$	$P(t_{n-2}<t^*)$
H_0:$\rho\leq 0$ vs H_1:$\rho>0$	Reject H_0 if $t^*>t_{1-\alpha,n-2}$	$P(t_{n-2}>t^*)$

An estimate of the population correlation using the sample correlation, *r* is obtained using statistical software.

Example

Continuing with the lizard example (p.21)above, correlation is tested using b_1 the estimated slope. Note that because it was concluded that the population slope was significantly larger than zero we may conclude that the correlation between running speed at 20°C and 35°C is significantly positive. The estimated value is r=0.704. Note that here r^2 = 0.496 which says that 49.6% of the variability in the running speeds at 35°C can be explained by the information given by the running speeds at 20°C.

Correlation does not imply causation

It is important to remember that when measuring the correlation between two variables, a significant correlation does not mean that one variable causes the other variable to increase or decrease. As an example, notice that ice-cream sales and boating accidents tend to be strongly positively correlated. However, no one would believe that eating ice-cream causes boating accidents. Both increase when the weather is warm. Be very careful with your conclusions when reporting correlations.

CHI SQUARE (χ^2)

Chi Square analysis is a technique used to compare two categorical variables. Usually, the question of interest is whether or not the two variables are independent. That is, does information about one variable provide information about the second variable (and indicate a dependent relationship.) The variables can be binary or have multiple levels.

Assumptions

- Independent random samples from two or more populations, with each subject classified according to one categorical variable (the other categorical variable represents the population from which the subject came.)

or

- A single simple random sample where each subject is classified into each of two categorical variables.

additionally

• All expected cell counts must be at least 1, and no more than 20% of the counts can be less than 5.

The data will be organized in a table such as that shown in Table 18 (here we have M rows and N columns).

Table 18 Example of the structure for a data set which has N categories for the first variables and M categories for the second variable

	Variable 1			
Variable 2	Group 1	Group 2	...	Group N
Group A				
Group B				
...				
Group M				

Inference

In general, the null hypothesis (H_0) is that there is no relationship between the two categorical variables. When the null hypothesis is true, the expected data values are easily calculated. That is, to investigate whether smoking and gender are related, start by figuring out what proportion of smokers are female (and male) if the two variables are not related. That is, we'd expect about half of the smokers to be male and half to be female if, in fact, the two variables were not related.

Expected counts

The expected count represents the number of subjects expected in each cell if in fact the null hypothesis is true (that is, if there is no relationship between the two variables.)

$$\text{expected count} = \frac{\text{row total x column total}}{\text{table total}}$$

Equation 37

Hypothesis testing

With the Chi Square test, there is not an obvious population parameter that is being estimated (unlike the one sample t-test of μ where inference is about the population mean.) Here, the χ^2 test statistic combines the expected and observed counts. The test statistic will allow us to make conclusions about the null and alternative hypotheses:

$$\chi^{2*} = \sum \cdot$$

Equation 38

(observed count - expected count)^2 / expected count

Only Two Categories per Variable

A chi square statistic is distributed according to the chi square distribution. There are chi square tables in the backs the introductory reference books, or use statistical software. The chi square table is sorted by degrees of freedom. For the chi square test of independence, the degrees of freedom are based on the number of rows and columns you have in the data table:

degrees of freedom = df=(number of rows-1)(number of columns-1)

Equation 39

Note that the Chi Square Test is always one-sided in the sense of rejecting the null hypothesis and calculating a p-value. Notice that if the null hypothesis is not true the expected counts will be much different from the observed counts. It does not matter in which direction the null hypothesis is wrong, any direction will produce a large test statistic. Therefore, always reject the null hypothesis when the test statistic is big. The decision rule for a chi square test are given in table 19.

Table 19 Decision rules for a Chi-square test

Hypotheses	Decision Rule	p-value
H_0: there is no relationship between the two categorical variables	Reject H_0 if $\chi^{2*} > \chi^2_{1-\alpha,df}$	$P(\chi^2_{df} > \chi^{2*})$

Example

Suppose interest is in determining whether or not there is a relationship between type of bird egg and ability of the egg to withstand an overnight freeze. Data on 100 eggs are collected and tabulated as shown in Table 20.

Table 20 Relationship Between Type of Bird Egg and Ability to Withstand Overnight Freezing

	Egg Status		
Egg Type	Okay	Cracked	Broken
Robin	9	8	4
Wren	11	12	7
Sparrow	5	10	9
Cuckoo	9	11	5

Table 21 shows the expected values computed assuming the null hypothesis of no relationship across variables is true. Note that the expected values do not seem particularly different from the observed values.

Table 21 Expected number of eggs in each category assuming no relationship between the two variables (i.e. assuming H_0 is true)

	Egg Status		
Egg Type	Okay	Cracked	Broken
Robin	7.14	8.61	5.25
Wren	10.20	12.30	7.50
Sparrow	8.16	9.84	6.00
Cuckoo	8.50	10.25	6.25

The associated test statistic is $\chi^{2*}=3.99$ with a p-value = 0.678. The p-value is large and says that if the null hypothesis is true, we would see data at least as extreme as ours about 67.8% of the time. The large p-value leads to failure to reject the null hypothesis. There is not significant evidence that egg type and egg status are related.

A special case occurs when there are only two levels (or categories) for one of the variables. For example, suppose a cracked egg is considered to be broken (that is, we've collapsed the cracked and broken data into one column.) . The data then become that shown in Table 21.

Table 22 Relationship between type of bird egg and ability to withstand overnight freezing with cracked and broken collapsed

	Egg Status	
Egg Type	Okay	Broken
Robin	9	12

Wren	11	19
Sparrow	5	19
Cuckoo	9	16

The resulting test statistic is $\chi^{2*}=2.73$ with a p-value = 0.4355. The null hypothesis of no relationship is still unable to be rejected. However, the hypotheses this test are stated and interpreted slightly differently (though the computations are the same as when there are more than two rows or columns.) Note that the hypotheses here can be stated as:

$$H_0: p_R = p_W = p_S = p_C$$

Equation 40

H_1: at least one of the proportions is different

Where p_R is the true proportion of Robin eggs; p_W is the true proportion of Wren eggs; p_S is the true proportion of Sparrow eggs; and p_C is the true proportion of Cuckoo eggs broken after an overnight freeze. The Chi Square Test of independence is actually testing whether or not the true proportion of broken eggs is the same across the four types of eggs.

LITERATURE CITED

Glantz, S. A. (2005) *Primer of Biostatistics, 6th Edition* New York: McGraw-Hill.

Hettmansperger, T. P. and McKean, J. W. (1998), *Robust Nonparametric Statistical Methods*, London: Arnold.

Hollander, M. and Wolfe, D. A. (1999), *Nonparametric Statistical Methods, 2nd Edition*, New York: John Wiley and Sons.

Kutner, M. H., Nachtsheim, C. J., Neter, J., and Li, W. (2005), *Applied Linear Statistical Models*, *5th Edition*, New York: McGraw-Hill.

Moore, D. S. and McCabe, G. P. (2006), *Introduction to the Practice of Statistics*, *5th Edition*, New York: W. H. Freeman and Company.

Utts, J. M. and Heckard, R. F. (2004) *Mind on Statistics, 2nd Edition* Belmont, CA: Brooks/Cole.