

February 27, 2009

Chapter 7: Two-Sample Inference

Chapter 6 introduced hypothesis testing in the *one-sample* setting: one sample is obtained from a single population and the sample mean was compared to a hypothesized value of the mean. Practical applications more frequently involve comparing the means of two or more populations. For example:

- Compare the mean response of individuals on an experimental drug treatment to those taking a placebo.
- Compare birds living near a toxic waste site with birds living in a pristine area.

In this chapter, we introduce the hypothesis testing procedures for comparing two populations. *Analysis of Variance* (ANOVA) is the statistical tool for performing hypothesis testing to comparing more than two means and this is covered in Chapter 9.

Two-Sample t -Test for Independent Samples with Equal Variances

One of the most popular statistical testing procedures is the two sample t -test used for comparing the means of two populations. The following example will be used to illustrate the ideas:

Example. A study was conducted to examine the effect of thermal pollution from a treatment plant on Asiatic clams (*Corbicula Fluminea*) in the water. A sample of clams was collected at the intake site where there was no thermal pollution and at a discharge site where there was thermal pollution. One of the variables that was measured on the length of the clams (in cm). The goal of the study was to determine if the thermal pollution was adversely affecting the growth of the clams and leading to smaller sizes on average. (Data collected by J. Booker, 1997.)

Clam Lengths	
Intake	Discharge
7.20	7.25
7.50	7.23
6.89	6.85
6.95	7.07
6.73	6.55
7.25	7.43
7.20	7.30
6.85	6.90
7.52	7.10
7.01	6.95
6.65	7.39
7.55	6.54
7.14	6.39
7.45	6.08
7.24	6.30
7.75	6.35

6.85	7.34
6.50	6.70
6.64	7.08
7.19	7.09
7.15	7.40
7.21	6.00
7.15	6.94
7.30	5.95
6.35	

The general setting is as follows: Consider two populations to be compared in terms of a particular variable X . Let μ_1 and μ_2 denote the means of the two populations and let σ_1 and σ_2 denote the standard deviations for the two populations respectively. In the clam example, the two populations are clams at the pristine site and the polluted site. The variable X is the length of the clams.

Generally, interest lies in comparing the means of the two populations. The null hypothesis of the test is that the two population means are equal:

$$H_0 : \mu_1 - \mu_2 = 0.$$

One or two-sided alternative hypotheses can be specified depending on the nature of the problem:

$$H_a : \begin{cases} \mu_1 - \mu_2 \neq 0 & \text{Two-Sided} \\ \mu_1 - \mu_2 < 0 & \text{One Sided} \\ \mu_1 - \mu_2 > 0 & \text{One Sided} \end{cases}.$$

Clam Example. Let μ_1 denote the mean clam length at the intake site and let μ_2 denote the mean length at the discharge site. Then the null hypothesis is $H_0 : \mu_1 - \mu_2 = 0$. Because we want to determine if the thermal pollution is retarding clam growth, the appropriate alternative hypothesis will be:

$$H_a : \mu_1 - \mu_2 > 0.$$

Thus, the alternative hypothesis states that clams at the discharge site are shorter on average than clams at the intake site.

In order to test the hypotheses above, random samples are obtained from each of the populations.

Notation. The following gives the notation that will be used:

Let n_1 and n_2 denote the sample sizes obtained from the two populations. Let \bar{X}_1 and S_1 denote the sample mean and standard deviation from population 1 and \bar{X}_2 and S_2 denote the sample mean and standard deviation from population 2.

Statistics for the Clam Data.

Site	
Intake	Discharge
$n_1 = 25$	$n_2 = 24$
$\bar{x}_1 = 7.09$	$\bar{x}_2 = 6.84$
$s_1 = 0.347$	$s_2 = 0.467$

The mean length of the $n = 24$ clams at the discharge site is less than the mean length of the $n = 25$ clams at the intake site. The question of interest is whether or not this difference is statistically significant.

In order to determine if the data support or contradict the null hypothesis of equal means ($\mu_1 - \mu_2 = 0$), it is natural to examine the difference in the sample means:

$$\bar{X}_1 - \bar{X}_2.$$

If the observed difference is small, then we have evidence supporting the null hypothesis. In order to tell if the difference between sample means is small or if it is too big to be explained by chance, we need to know the sampling distribution of $\bar{X}_1 - \bar{X}_2$. Here are a couple of facts:

Fact 1. $E[\bar{X}_1 - \bar{X}_2] = \mu_1 - \mu_2$.

Fact 2. If the random samples from the two populations are independent of each other, then

$$\text{Var}(\bar{X}_1 - \bar{X}_2) = \frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}. \quad (1)$$

These two facts follow immediately from the properties of linear combinations that were covered in Chapter 4. The *central limit theorem* gives the following fact:

Fact 3. If n_1 and n_2 are both sufficiently large, then

$$(\bar{X}_1 - \bar{X}_2) \approx N(\mu_1 - \mu_2, \sigma_1^2/n_1 + \sigma_2^2/n_2).$$

If the underlying distributions for both populations are normal, then $\bar{X}_1 - \bar{X}_2$ will have an exact normal distribution.

Note that we are assuming that the two random samples are obtained independently of each other. In the clam example, this independence assumption is reasonable. However, if data is collected on individuals at two different times such as a “before and after” or data comparing a subject’s baseline measurement to an end-of-study measurement, then the two sets of measurements will be correlated and not be independent. If the two samples of data are correlated due to *repeated measures* on the same subjects, then a *paired t-test* is more appropriate – we will cover this later in this chapter.

A common assumption made when comparing two population means is that the variance of the two populations are equal: $\sigma_1^2 = \sigma_2^2$. Let σ^2 denote the common variance. If the equal variance assumption holds, then we can factor out the common variance in (1) and write the variance of the difference as:

$$\text{Var}(\bar{X}_1 - \bar{X}_2) = \sigma^2(1/n_1 + 1/n_2). \quad (2)$$

In practice, the population variances have to be estimated. If we assume the population variances are equal, then it makes sense to *pool* all the data from both samples to estimate the common variance. The pooled estimate of the variance, given by the following formula, is a weighted average of the sample variances from the two populations:

$$\text{Pooled Estimated of the Variance: } S^2 = \frac{(n_1 - 1)S_1^2 + (n_2 - 1)S_2^2}{n_1 + n_2 - 2} \quad (3)$$

The degrees of freedom for the pooled variance estimator is $n_1 + n_2 - 2$. Two degrees of freedom are lost when estimating μ_1 and μ_2 respectively.

The test statistic for the hypothesis test is simply the standardized difference between the sample means:

$$T = \frac{\bar{X}_1 - \bar{X}_2}{S\sqrt{1/n_1 + 1/n_2}} \quad (4)$$

If we are sampling from normal populations, then the sampling distribution of T follows a Student's t -distribution on $n_1 + n_2 - 2$ degrees of freedom. Therefore, the t -distribution is our reference distribution for determining if the difference in means is small (consistent with H_0) or if the difference is too big to be explained by chance (supporting H_a).

Once the data is collected, we will have an observed value for T in (4) which is the two-sample t -test statistic:

$$t = \frac{\bar{x}_1 - \bar{x}_2}{s\sqrt{1/n_1 + 1/n_2}} \quad (5)$$

The following table summarizes the two-sample t -test procedure when testing at a significance level α .

Two-Sample t -test, Two-Tailed Alternative

$$H_0 : \mu_1 - \mu_2 = 0 \quad \text{versus} \quad H_a : \mu_1 - \mu_2 \neq 0.$$

Decision:

$$\begin{array}{ll} \text{Reject } H_0 & \text{if } t > t_{n_1+n_2-2, 1-\alpha/2} \text{ or } t < -t_{n_1+n_2-2, 1-\alpha/2}. \\ \text{Fail to Reject } H_0 & \text{otherwise.} \end{array}$$

The p -value for the two-tailed alternative is the area under the t -density to the right of $|t|$ plus the area to the left of $-|t|$, where t is the test statistic (5):

$$p\text{-value} = 2Pr(T > |t|).$$

The next table describes the testing procedure for one-tailed alternatives:

Two-Sample t -test, One-Tailed Alternative

$$H_0 : \mu_1 - \mu_2 = 0$$

$$\begin{array}{l|l} H_a : \mu_1 - \mu_2 > 0. & H_a : \mu_1 - \mu_2 < 0. \\ \text{Reject } H_0 \text{ if } t > t_{n_1+n_2-2, 1-\alpha} & \text{Reject } H_0 \text{ if } t < -t_{n_1+n_2-2, 1-\alpha} \\ p\text{-value} = Pr(T > t) & p\text{-value} = Pr(T < t). \end{array}$$

The p -value from the two-sample t -test is interpreted the same as in the case of a one-sample test: small p -values are evidence against the null hypothesis and large p -values do not provide evidence against the null hypothesis.

We can also estimate the difference in the means $\mu_1 - \mu_2$ using a confidence interval:

Confidence Interval for $\mu_1 - \mu_2$:

A $100 \times (1 - \alpha/2)$ confidence interval for $\mu_1 - \mu_2$ is given by

$$(\bar{x}_1 - \bar{x}_2) \pm t_{n_1+n_2-2, 1-\alpha/2} (s \sqrt{1/n_1 + 1/n_2}). \quad (6)$$

The meaning of *confidence* in the two-sample confidence interval for $\mu_1 - \mu_2$ is to be interpreted similar to the one-sample confidence interval: if we are computing a 95% confidence interval for $\mu_1 - \mu_2$, then in repeated sampling, roughly 95% of the intervals will contain the true difference $\mu_1 - \mu_2$.

Assumptions. The hypothesis tests and confidence interval formula just described are valid given that the following assumptions are satisfied:

1. *Independence* – Random samples from two populations and the two samples are independent of each other. This independence assumption cannot be relaxed. If the measurements in the two samples are correlated, then another inference procedure is required (such as the paired t -test).
2. *Normal Distributions* - Each sample is from a normal distribution. This assumption can be relaxed provided the underlying distributions do not deviate too strongly from normality. Always plot your data to assess the normality assumption. For small data sets, it is difficult to see the shape of distributions. However, strong deviations from normality in the underlying distributions can often be determined visually for small data sets (e.g. outlying observations). The larger the sample sizes, the more the normality assumption can be relaxed due to the central limit theorem effect.
3. *Equal Variance Assumption.* The two-sample t -test statistic in (5) and the confidence interval formula (6) both use the pooled estimate of variance which assumes that each distribution has the same variance. The t -procedure is fairly robust to deviations to the equal variance assumption, particularly if the sample sizes n_1 and n_2 are equal (or approximately equal). For this reason, it is usually recommended to plan studies with equal sample sizes from the difference populations (this holds for comparing more than two populations). For small sample sizes, the estimators S_1 and S_2 will be

quite variable and it is difficult to access the equal variance assumption. Nonetheless, a common rule of thumb is that the equal variance assumption is plausible if the larger sample standard deviation is no more than twice the smaller sample standard deviation. If the equal variance assumption is not satisfied, we can use the unequal variance procedure described below.

Two-Sample t -Test with Unequal Variances

This section describes the testing procedure for equality of means when the assumption of equal population variances in the two populations is violated. This is known as the **Behrens-Fisher** problem. Inference will still be based on the difference in the sample means: $\bar{X}_1 - \bar{X}_2$. Recall that for two independent samples, we have

$$\text{Var}(\bar{X}_1 - \bar{X}_2) = \frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}.$$

If we are not confident in assuming that $\sigma_1^2 = \sigma_2^2$, the test statistic for equality of means is still based on the standardized difference between the sample means:

$$\frac{\bar{X}_1 - \bar{X}_2}{\sqrt{S_1^2/n_1 + S_2^2/n_2}}. \quad (7)$$

The problem with the statistic in (7) is that it does not follow a t -distribution and its actual distribution is difficult to access. However, the statistic (7) can be well approximated by a t -distribution with degrees of freedom equal to the following complicated formula:

$$\text{Approximate Degrees of Freedom: } df = \frac{(s_1^2/n_1 + s_2^2/n_2)^2}{(s_1^2/n_1)^2/(n_1 - 1) + (s_2^2/n_2)^2/(n_2 - 1)}. \quad (8)$$

This approximation is known as **Satterthwaite's Method**. The testing procedure is the same as before except the original test statistic (5) is modified to be

$$t = \frac{\bar{x}_1 - \bar{x}_2}{\sqrt{s_1^2/n_1 + s_2^2/n_2}}, \quad (9)$$

and the degrees of freedom is approximated by (8). Computing (8) by hand is complicated, but statistical software packages like SAS will often do the computation for us.

We shall now apply the two-sample procedures to the clam data.

Clam Example continued. Recall that we wish to test if the mean clam length at the discharge site is less than the mean clam length at the intake site. Histograms of the clam length data, generated using SAS analyst, at the two sites are shown in Figure 1. The length distributions at both sites look skewed to the left. Recall that the two-sample t -test assumes the underlying distributions are approximately normal, but that the test is robust to departures from this assumption, particularly if the sample sizes are nearly equal. Later we will introduce a nonparametric test that does not require the normality assumption.

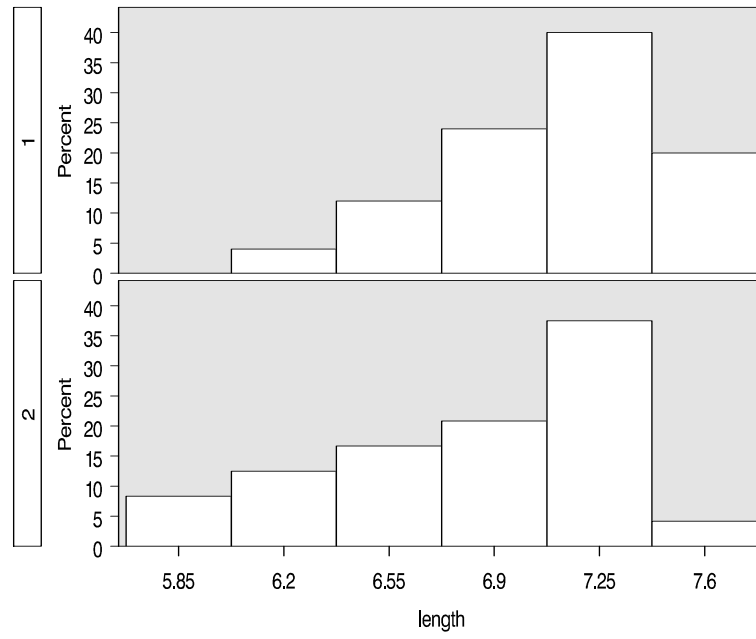


Figure 1: Histograms of clam lengths at the Intake (1) and Discharge (2) sites

Because the sample sizes are $n_1 = 25$ and $n_2 = 24$, the two-sample t -test assuming equal variances will use the t -distribution on $n_1 + n_2 - 2 = 25 + 24 - 2 = 47$ degrees of freedom. The sample standard deviations (see SAS output below) are $s_1 = 0.3466$ and $s_2 = 0.4670$ respectively. The pooled estimate of the sample variance is

$$s^2 = \frac{(n_1 - 1)s_1^2 + (n_2 - 1)s_2^2}{n_1 + n_2 - 2} = \frac{(25 - 1)0.3466^2 + (24 - 1)0.4670^2}{25 + 24 - 2} = 0.16355,$$

and the pooled estimate of the standard deviation is

$$s = \sqrt{s^2} = \sqrt{0.16355} = 0.40441.$$

Note that the pooled estimate of the standard deviation will always lie between the sample standard deviations for the two populations. The t -test statistic (5), assuming equal variances, is given by

$$t = \frac{\bar{x}_1 - \bar{x}_2}{s\sqrt{1/n_1 + 1/n_2}} = \frac{7.0888 - 6.8408}{0.4041\sqrt{1/25 + 1/24}} = 2.12.$$

These computations can be performed by SAS using PROC TTEST as the following SAS program illustrates:

```

/*****
Clam Data: study on the effect of thermal pollution from a
treatment plant on Asiatic clams (Corbicula Fluminea).
Data:
Column 1 Site: 1 = Intake site, 2 = Discharge site
Column 2 Length (cm)
(Data collected by J. Booker, 1997.)
*****/
options ls=80;
data clams;
infile 'c:\stt630\notes\clams.dat';
input site length;
run;
proc ttest;
class site;
var length;
run;

```

The “CLASS” statement from PROC TTEST indicates the classification variable which in this context is the independent variable. The CLASS statement is required for PROC TTEST. The variable specified by the CLASS statement can only take two values for the two different populations. In this example the CLASS variable is SITE which takes the values 1 and 2 for the intake and discharge sites respectively. The “var length” statement tells SAS to perform the t -test using the dependent variable length (length of the clam depends on the site where the clam is from).

The output from PROC TTEST is shown below:

The TTEST Procedure

Statistics

Variable	site	N	Lower CL		Upper CL	Lower CL	
			Mean	Mean	Mean	Std Dev	Std Dev
length	1	25	6.9457	7.0888	7.2319	0.2706	0.3466
length	2	24	6.6436	6.8408	7.038	0.3629	0.467
length	Diff (1-2)		0.0123	0.248	0.4836	0.3413	0.4099

Statistics

Variable	site	Upper CL			
		Std Dev	Std Err	Minimum	Maximum
length	1	0.4821	0.0693	6.35	7.75
length	2	0.6551	0.0953	5.95	7.43
length	Diff (1-2)	0.5135	0.1172		

T-Tests

Variable	Method	Variances	DF	t Value	Pr > t
length	Pooled	Equal	47	2.12	0.0396
length	Satterthwaite	Unequal	42.4	2.10	0.0414

Equality of Variances

Variable	Method	Num DF	Den DF	F Value	Pr > F
length	Folded F	23	24	1.82	0.1540

Note that PROC TTEST automatically computes 95% confidence intervals for both means as well as the difference in means using (6). The t -test statistic from the SAS output (assuming equal variances) is $t = 2.12$. SAS automatically produces a *two-tailed* p -value. In this example, the two-tailed p -value is 0.0396. However, recall that we are performing a one-tailed test. Therefore, since the mean difference is positive which is consistent with the alternative hypothesis, the actual p -value is half the two-tailed p -value. Thus, the p -value for the one-sided test is

$$p = 0.0396/2 = 0.0198$$

which provides strong evidence against the null hypothesis.

If we were performing this test using a significance level α , then we would reject the null hypothesis if the p -value is less than α since the p -value is the smallest significance level at which we would reject H_0 .

SAS also automatically produces the Satterthwaite t -test statistic that does not assume equal variances (9), the corresponding approximate degrees of freedom from (8), and the p -value. Note that Satterthwaite's

approximation gives a t -test statistic of 2.10 which is very close in value to the original test statistic. The approximate degrees of freedom is less (42.2 compared to 47), but the p -value of $0.0414/2 = 0.0207$ yields essentially the same result as the equal variance t -test (i.e. reject H_0).

In conclusion: we reject H_0 with $p = 0.0207$ and conclude that the clams at the discharge site have shorter lengths on average than clams at the intake site.

Based on the evidence we have available, we cannot necessarily claim that the thermal pollution at the discharge site is causing the clams to have a shorter length on average. The thermal pollution may very well be the cause of the statistically significant difference, but because this is an observational study, there could be other factors that were not controlled for that could influence the length of clams at the two sites.

More General Hypotheses. Up to this point we have considered the null hypothesis that the two means are equal: $H_0 : \mu_1 - \mu_2 = 0$. It is possible and sometimes of interest to perform a test that specifies that the means differ by some quantity δ other than zero:

$$H_0 : \mu_1 - \mu_2 = \delta_0$$

versus an alternative that the difference is not equal to δ_0 (Two-sided), or greater (less) than δ_0 (one-sided). The two-sample t -test is easily modified to handle this by using the test statistic

$$t = \frac{(\bar{x}_1 - \bar{x}_2) - \delta_0}{s_p \sqrt{1/n_1 + 1/n_2}}$$

Confidence Interval. If the original goal of the study had been to estimate the difference in the mean lengths using a confidence interval, then, from the SAS output, we could state the following:

With 95% confidence, we estimate that clams at the intake site are 0.0123 to 0.4836 centimeters longer on average than clams at the discharge site. Note that zero is not in this confidence interval indicating that a difference of zero is not a plausible value for the mean difference.

Sample Size and Power When Comparing Two Means

As for one-sample hypothesis tests and confidence intervals, it is important to plan experiments when comparing two population means in terms of power and sample size. The same principals that held in the one-sample setting hold in the two-sample setting:

- The larger the sample sizes n_1 and n_2 , the higher the power $1 - \beta$ of the test.
- As the difference between means $\mu_1 - \mu_2$ grows bigger, the power of the two-sample t -test increases for fixed sample sizes.
- Smaller population variances σ_1^2 and σ_2^2 lead to higher power.
- Increasing the significance level (probability of committing a type I error) will increase the power for a fixed sample size.

In order to find a required sample size for testing the null hypothesis $H_0 : \mu_1 - \mu_2 = 0$, one typically needs to specify the significance level α and the desired power $1 - \beta$ as well as an estimate of the population standard deviations σ_1, σ_2 . In addition, the desired difference between means $\Delta = |\mu_1 - \mu_2|$ that one would like to be able to detect from the test. Generally sample size and power computations are performed assuming equal sample sizes obtained from both populations. If the population standard deviations are known, then the required sample size for a two-sided alternative is given by:

$$n = \frac{(\sigma_1^2 + \sigma_2^2)(z_{1-\alpha/2} + z_{1-\beta})^2}{\Delta^2}. \quad (10)$$

The n in this formula is the sample size to obtain from *both* populations so that the test has power $1 - \beta$ to detect a difference between the means of at least $\Delta = |\mu_1 - \mu_2|$ when testing at a significance level α . For a one-sided test, formula (10) can be modified by replacing $z_{1-\alpha/2}$ by $z_{1-\alpha}$. There is a modification of formula (10) if unequal sample sizes are desired or required (not given here).

Turning (10) around, we can solve for the power of a two-sided test given the sample sizes $n = n_1 = n_2$:

$$\text{Power} = \Phi\left(-z_{1-\alpha} + \frac{\sqrt{n}\Delta}{\sqrt{\sigma_1^2 + \sigma_2^2}}\right). \quad (11)$$

Of course, in practice, the standard deviations are not known and estimates will be needed to apply formula (10). However, in this case, percentiles from the t -distribution are required and one typically needs to iterate the sample size formula to get an answer. As in the one-sample case, statistical software packages can perform these calculations for us.

In the clam example above, we can use SAS to perform a retrospective power analysis. Given the observed statistics from the data, SAS's analyst will do power computations for specified sample sizes. Once in Analyst, Click on:

Statistics \rightarrow Hypothesis Tests \rightarrow Two-sample t-test for Means,

then click on TESTS and POWER ANALYSIS. From the clam example, we obtain the following output:

Power Analysis

Alpha	Sample Size	Observed Power
-----	----	-----
0.050	49	0.670

Alpha	Sample Size	Hypothetical Power
-----	----	-----
0.050	10	0.221
0.050	20	0.366
0.050	30	0.488
0.050	40	0.592
0.050	50	0.678
0.050	60	0.749

0.050	70	0.805
0.050	80	0.850
0.050	90	0.885
0.050	100	0.913

As the sample size increases, the power increases. These computations were done using a significance level α . What would happen to the power results if $\alpha = 0.01$ were used instead?

Nonparametric Tests.

The t -testing procedure described above rests on the assumption that independent samples are obtained from two *normal* populations. The t -test is known as a **parametric** statistical test because they assume a parametric form of the underlying distributions. For modest violations of the normality assumption, the t -test will still be approximately valid. However, if the normality assumption is not reasonable, then *non-parametric* tests can be used which do not require that the underlying distribution belongs to a particular parametric family (such as the normal distribution).

Before describing nonparametric tests, we first describe different data types:

Definition. Cardinal Data are data that are on a scale where it is meaningful to measure distances between different data values.

Variables like weights, heights, volumes etc. are on a cardinal scale.

Definition. Interval Scale data is cardinal data where the zero point is arbitrary.

The best example of interval scale data is temperature. The zero temperature differs depending on the scale (e.g. Kelvin, Celsius, Fahrenheit).

Definition. Ratio Scale data is cardinal data with a fixed zero point.

A variable like weight or volume is on a ratio scale. For data on a ratio scale, it is meaningful to measure ratios of values (e.g. “this weighs twice as much as that”).

Another data type that occurs very frequently in practice is *ordinal data*:

Definition. Ordinal Data are data that can be ordered but do not correspond to *cardinalities* of sets.

The Hamilton Depression scale (HAM-D) records items using a rating 0 = absent, 1 = doubtful to mild, 2 = mild to moderate, 3 = moderate to severe, 4 = very severe. Ordinal data are clearly non-normal and using a two-sample t -test may not be valid with such data. A nonparametric test may be more appropriate with ordinal data.

Definition. Nominal Data are data that record values in different categories (e.g. married, single, divorced) with no meaningful ordering between categories. Categorical data analysis tools are needed to analyze this type of data.

The Wilcoxon Rank-Sum Test

There are several nonparametric tests for testing if two distributions are equal. We shall describe one of the best known tests: the Wilcoxon Rank-Sum Test. If the normality assumption of the two-sample t -test does not seem plausible, then the Wilcoxon rank-sum test can be used which does not require the normality assumption.

The basic idea of the Wilcoxon rank-sum test is to replace the raw data in the two samples by their relative ranks and then compare the rank values of the two groups. Formally, the Wilcoxon rank-sum test is testing equality of the population *medians*.

Here is how the test is carried out:

1. Combine all the data from the two samples and order the values from lowest to highest.
2. Assign ranks to the values: $1, 2, 3, \dots, n_1 + n_2$ (if ties exist, assign the average rank to the tied observations).

In SAS, for the clam data, we can obtain the ranks using the following code:

```
proc sort; by length;
run;
proc print;
run;
```

The first several lines from the proc print command are:

Obs	site	length
1	2	5.95
2	2	6.00
3	2	6.08
4	2	6.30
5	1	6.35
6	2	6.35
7	2	6.39
8	1	6.50
9	2	6.54
10	2	6.55
11	1	6.64
12	1	6.65
13	2	6.70
14	1	6.73
15	1	6.85

The first column (Obs) gives the ranks of the pooled data. Note that the first several ranks are all from site 2, the discharge site. To perform the test, one computes the rank-sums R_1 and R_2 for each group

where R_1 is simply the sum of the ranks from group 1. One can show that under H_0 , the average rank-sum for the combined sample is

$$\frac{1 + n_1 + n_2}{2},$$

and therefore, under H_0 ,

$$E[R_1] = n_1 \left(\frac{1 + n_1 + n_2}{2} \right).$$

Using combinatoric counting techniques, one can show that under H_0 ,

$$\text{var}(R_1) = n_1 n_2 (n_1 + n_2 + 1) / 12.$$

The Wilcoxon rank-sum test statistic is computed by comparing the observed rank-sum to the expected rank-sum (under H_0):

$$\text{Wilcoxon Rank-Sum Test Statistic: } T = \frac{|R_1 - \frac{n_1(n_1+n_2+1)}{2}| - 1/2}{\sqrt{(n_1 n_2 / 12)(n_1 + n_2 + 1)}}.$$

If there are ties among the ranks, then this formula is modified slightly.

The exact distribution of T can be worked out again by combinatoric counting techniques. However, the distribution of the test statistic T will follow an approximate standard normal distribution when H_0 is true for large sample sizes due the central limit theorem effect.

Note that the T test statistic is defined in terms of R_1 only. Using R_2 and n_2 instead would give exactly the same result.

Using the normal approximation, one rejects H_0 at significance level α if

$$T > z_{1-\alpha/2}.$$

For the normal approximation to be approximately valid, sample sizes from each population should be at least 10.

In SAS, the Wilcoxon rank-sum test can be carried out using PROC NPAR1WAY. The following SAS code demonstrates its use for the clam data:

```
data clams;
infile 'clams.dat';
input site length;
run;
proc npar1way;
var length;
class site;
run;
```

PROC NPAR1WAY gives output for several different testing procedures but we shall focus only on the Wilcoxon rank-sum output:

The NPAR1WAY Procedure

Wilcoxon Scores (Rank Sums) for Variable length
Classified by Variable site

site	N	Sum of Scores	Expected Under H0	Std Dev Under H0	Mean Score
1	25	710.0	625.0	49.987243	28.400000
2	24	515.0	600.0	49.987243	21.458333

Average scores were used for ties.

Wilcoxon Two-Sample Test

Statistic 515.0000

Normal Approximation

Z -1.6904

One-Sided Pr < Z 0.0455

Two-Sided Pr > |Z| 0.0909

t Approximation

One-Sided Pr < Z 0.0487

Two-Sided Pr > |Z| 0.0974

Z includes a continuity correction of 0.5.

The one-sided p -value using the normal approximation is $p = 0.0455$ which again shows evidence that distribution of clam lengths at the discharge site is lower than at the intake site. Note that the Wilcoxon p -value is higher than it was for the two-sample t -test.

The Wilcoxon rank-sum test is sometimes referred to as the *Mann-Whitney U test*: both tests yield equivalent results.

Paired t -Test

A clinical trial was run to test the medication *captopril* for the treatment of high blood pressure. The question arises: how do we design the study? One possibility is to enroll subjects into the study and randomize the subjects into two treatment arms: captopril and a placebo. Once the experiment is over, a two-sample t -test could be used to analyze the data to determine if the group receiving captopril had significantly lower blood pressure at the end of the study.

Can you think of another way to conduct this trial? The problem with the two-sample t -test procedure is that there is likely to be a lot of patient-to-patient variability in blood pressures. If the effect of the drug is small relative to this *between* subject variability, then the statistical test may not be able to detect the drug's effect.

We are not interested in the patient-to-patient variability in blood pressures. One way to factor out this variability is to use each subject as their own control. That is, measure each patient's blood pressure before taking the drug and after taking the drug and record the difference. Next, perform the statistical analysis on the differences (before – after).

Below is a table showing data from the captopril data (MacGregor et al 1979). The first column is the systolic blood pressure (in mm Hg) before taking the drug and the second column is the blood pressure after taking the drug. The last column is the difference (Before – After).

Before	After	Difference
210	201	9
169	165	4
187	166	21
160	157	3
167	147	20
176	145	31
185	168	17
206	180	26
173	147	26
146	136	10
174	151	23
201	168	33
198	179	19
148	129	19
154	131	23

Table 1: Systolic Blood Pressure Before and After Medication

Figure 2 shows a scatterplot of the systolic blood pressures before and after taking the drug for the 15 subjects. Note that there is quite a bit of variability between subjects. Pairing the subjects by taking before and after readings on each factors out all of the between subject variability evident in Figure 2. Note that subject 1 has a relatively high blood pressure before and after (compared to subject 2 say). Therefore, the before and after blood pressure readings appear to be positively correlated – that is, if a subject's before blood pressure is higher than average, then the subject's after blood pressure reading will tend to be above average as well.

If we let (X_i, Y_i) denote the systolic blood pressure before (X_i) and after (Y_i), then we can define the difference as $D_i = X_i - Y_i$. The mean difference can be denoted by

$$\mu_d = E[D_i] = E[X_i - Y_i].$$

The null hypothesis is that the drug will not help: $H_0 : \mu_d = 0$. The alternative hypothesis is that the drug will lower blood pressure, making the mean difference positive:

$$H_a : \mu_d > 0.$$

To test this hypothesis, we can compute the average difference from the observed differences (right-most column in the above table):

$$\bar{d} = (d_1 + d_2 + \cdots + d_n)/n.$$

Letting s_d denote the standard deviations of the differences, we have

$$s_d = \frac{1}{n-1} \sum_{i=1}^n (d_i - \bar{d})^2.$$

The test statistic then is the standardized difference between \bar{d} and zero:

$$\text{Paired Difference Test Statistic: } t = \frac{\bar{d}}{s_d/\sqrt{n}}.$$

If the differences follow an approximate normal distribution, then this t -test statistic follows a normal distribution on $n - 1$ degrees of freedom where n is the number of differences. We can follow the same testing procedure as in the one-sample t -test which is summarized below:

Two-Sided Test: If $H_a : \mu_d \neq 0$ then reject H_0 if

$$t > t_{n-1, 1-\alpha/2} \text{ or } t < -t_{n-1, 1-\alpha/2}.$$

One-Sided Test – Lower Tail: If $H_a : \mu_d < 0$ then reject H_0 if $t < -t_{n-1, 1-\alpha}$.

One-Sided Test – Upper Tail: If $H_a : \mu_d > 0$ then reject H_0 if $t > t_{n-1, 1-\alpha}$.

If we want to test if the drug captopril lowers systolic blood pressure, then our null and alternative hypotheses are: $H_0 : \mu_d = 0$ versus $H_a : \mu_d > 0$. From the $n = 15$ differences in Table 1, we find that $\bar{d} = 18.93$ and $s_d = 9.027$. The t -test statistic is

$$t = \frac{\bar{d}}{s_d/\sqrt{n}} = \frac{18.93}{9.027/\sqrt{15}} = 8.213.$$

From SAS, the one-tailed p -value was found to be $p < 0.0001$ which indicates a highly significant difference. Thus, we have very strong evidence ($p < 0.0001$) that captopril lowers systolic blood pressure.

Confidence Interval. If an estimate of the mean difference is desired, we could compute a $100 \times (1 - \alpha)$ confidence interval for the mean difference μ_d as

$$\text{Confidence Interval for Paired Difference: } \bar{d} \pm t_{n-1, 1-\alpha/2}(s_d/\sqrt{n}).$$

In the blood pressure example, the degrees of freedom is $n - 1 = 15 - 1 = 14$. A 95% confidence interval mean difference is

$$18.93 \pm 2.145(9.027/\sqrt{15})$$

which gives an interval of (13.93, 23.93). With 95% confidence we estimate that blood pressure is lowered by 13.93 to 23.93 mm Hg on average when taking captopril.

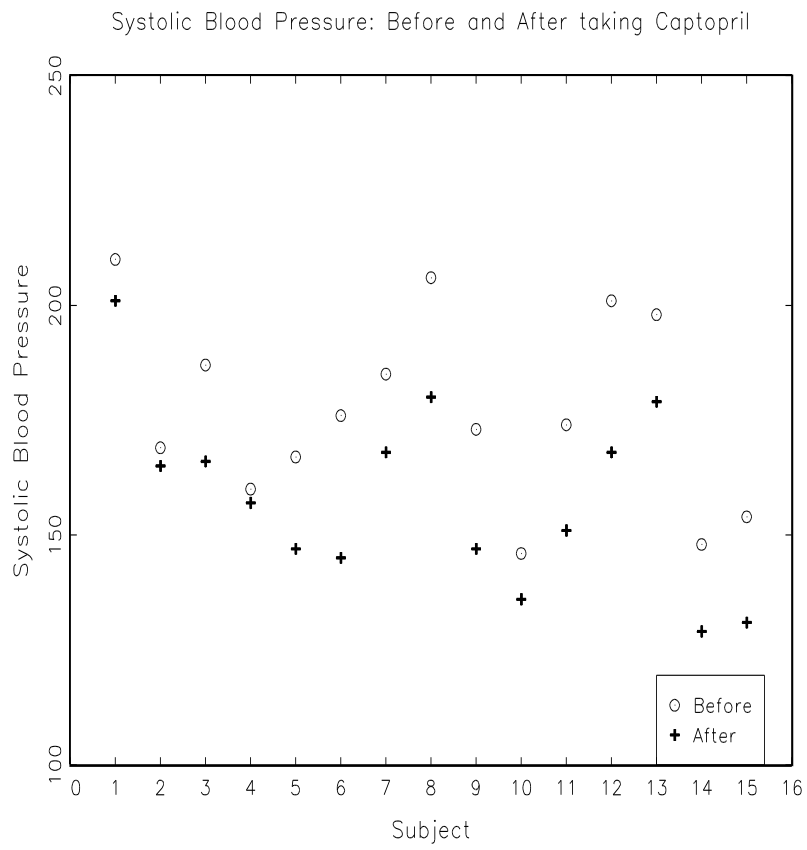


Figure 2:

Note that we cannot use the two-sample t -test in the blood pressure example because the before and after readings will be correlated – they are not independent. The use of the paired t -test in this example makes advantageous use of the correlation between before and after readings. Suppose X and Y represent random measurements of interest (e.g. blood pressure on and off the drug) and we want to compare the difference in their means $\mu_x - \mu_y = E[X - Y]$. If X and Y are independent, then the variance of the difference is

$$\text{var}(X - Y) = \sigma_x^2 + \sigma_y^2. \quad (12)$$

Recall that in our t -test statistics, the denominator is the standard error of the numerator. If we can make the test statistic larger, then the test can be more powerful since we generally reject the null hypothesis for large values of the test statistic. The t -test statistic becomes larger when the denominator becomes smaller. If X and Y are positively correlated, then

$$\text{var}(X - Y) = \sigma_x^2 + \sigma_y^2 - 2\text{cov}(X, Y), \quad (13)$$

which will be smaller than (12). In other words, the positive covariance allows the denominator of the test statistic to be come smaller and hence makes the test more powerful. Figure 3 shows a scatterplot of the before and after blood pressure readings from the preceding example. This plot shows a very strong positive correlation between the before and after blood pressure measurements.

The paired t -test can be used in other types of applications besides “before and after” type experiments. For example, we could test two different medications on each subject. In this case we would once again be using the individual subjects to test both medications which would factor out sources of variation between subjects. However, in such a study, a placebo may also be needed to determine a baseline for comparing the two medications. For instance, if the responses are similar for both medications, can we say if either drug is working or not? Another very important consideration is the randomization of the order in which subjects receive treatments. If all subjects get drug A first followed by drug B, then one may not be able to tell how much of any observed difference is due to the drug or due to the order in which the drugs are taken.

The paired t -test can also be used by pairing observational units (patients, plants, sites, etc) that are similar to each other. Again, this can help factor out unwanted sources of variability. For instance, in the blood pressure example, we could pair patients who have similar weights if it is thought that variability in blood pressures is due in part to varying weights of people. Pairing individuals on the basis of body weight could help to factor out this source of variability and allow a more focused analysis of changes in blood pressure due to the drug only. In such examples, the experimenter has a choice ahead of time as to use a paired t -test or a two-sample t -test. The two-sample t -test could be used if we ignore the pairing based on body weights. If n subjects take medication A and n take medication B, then the degrees of freedom for the two-sample t -test is $n + n - 2$. However, if subjects are paired and a paired t -test is used, then there will be only n differences and $n - 1$ degrees of freedom for the paired t -test. Thus, pairing leads to a loss of degrees of freedom. However, this loss of degrees of freedom can be compensated for if it is thought that pairing will factor out a great deal of unwanted variability. The decision to pair or not should be made before analyzing the data.

1 Problems

1. A study on sperm quality for two groups of men: those exposed to organophosphate pesticides ($n_1 = 32$) on their jobs and those not exposed ($n_2 = 43$). Two variables, sperm concentration and

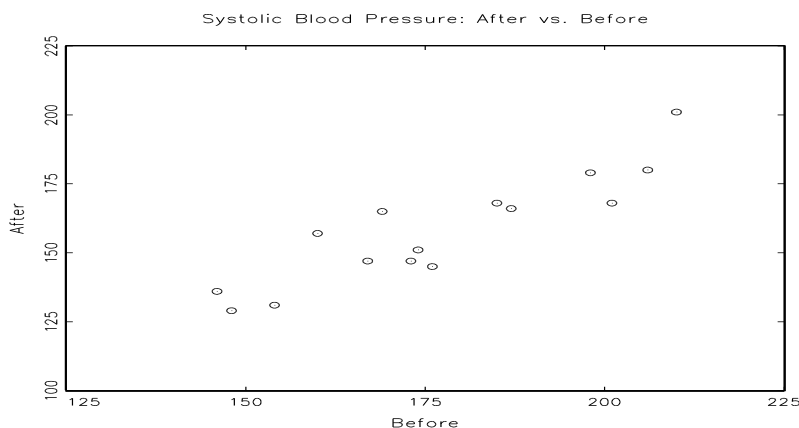


Figure 3:

motility, were measured. Two-sample t -tests were conducted for each variable to test if the exposed men had lower concentrations and lower motility on average than the non-exposed population. The p -value for the test on sperm concentration was $p_1 = 0.0076$ and the p -value for the test on motility yielded $p_2 = 0.13$. Use this information to answer the following questions.

- a) Can the researchers claim that the exposed men have lower sperm concentrations on average than non-exposed men? Why or why not?
 - b) A researcher, noting the p -value of $p_2 = 0.13$ for the two-sample t -test on motility, states that “on average, the exposed men and unexposed men have the same sperm motility.” Do you agree with this statement? Why or why not?
 - c) What are the degrees of freedom associated with the t -tests if we assume equal variances for the two populations of men?
 - d) If the sperm concentration data was heavily skewed to the left for the two groups of men, would you feel comfortable with the t -test results? Can you recommend a different statistical test?
2. A study was conducted to determine if the cholesterol lowering drug lipitor performed better than its competitor pravachol. The decrease in cholesterol levels after taking the drugs for a year were recorded for fifteen subjects taking lipitor and fifteen subjects taking pravachol. The data is summarized in the following table:

Group	n	Sample Mean	Standard Deviation
Lipitor	15	48.5	23.35
Pravachol	15	41.6	26.67

Perform a two-sample t -test using a level of significance $\alpha = 0.05$ to determine if the average decrease in cholesterol using lipitor is more than the average decrease in cholesterol using pravachol. To do this problem, do the following parts:

- a) Define appropriate parameters and state H_0 and H_a .

- b) Compute the pooled estimate of the standard deviation s_p .
- c) What are the degrees of freedom for the test statistic?
- d) What is the critical region?
- e) What is the value of the test statistic?
- f) In plain English, what is the conclusion of the test? Write one or two sentences.
3. Ten moderately trained male runners underwent an experiment where they ran 5 kilometers in the morning (low temperature) and 5 kilometers at midday (higher temperature) and their pulse rates were measured after each run. The purpose of the study was to determine if running during higher temperatures lead to an increased pulse rate. Which of the following is an appropriate statistical technique for answering the research question?
- a) Two-sample t -test
- b) Binomial test
- c) Wilcoxin Rank Sum test
- d) Paired t -test
4. The p -value from the test in the previous problem (#3) was found to be $p = 0.034$. Which of the following is a correct interpretation of the p -value?
- a) Accept the null hypothesis H_0 and conclude that the mean pulse rate is the same in the morning and at midday.
- b) The probability the null hypothesis is true is 0.034 which is low, so we reject the null hypothesis.
- c) Reject the null hypothesis and conclude that the mean pulse rate is higher after running at midday than in the morning.
- d) Fail to reject the null hypothesis because the probability of a type I error is 0.034.
5. A study comparing mercury levels in bass fish at two sites in the San Francisco Bay is being planned. A two-sample t -test using a significance level $\alpha = 0.05$ will be used and power of 80% achieved to detect a difference in mean mercury levels of $0.1mg/g$. It is assumed that standard deviations at each site are equal $\sigma := \sigma_1 = \sigma_2 = 0.06$. The required sample size at each site was found to be $n_1 = n_2 = 6$. What would happen to the required sample size in each of the following circumstances (circle one answer for each part)
- a) The power is increased to 90% from 80%. The required sample size would:
Increase Decrease Stay the Same Not enough information
- b) The standard deviation is $\sigma = 0.04$ instead of 0.06 . The required sample size would:
Increase Decrease Stay the Same Not enough information
- c) The detectable difference was decreased from 0.1 to $0.05mg/g$. The required sample size would:
Increase Decrease Stay the Same Not enough information
- d) The significance level is decreased from 0.05 to 0.01. The required sample size would:
Increase Decrease Stay the Same Not enough information
- e) Instead of bass fish, the study will compare mercury levels in jacksmelt fish. The required sample size would:
Increase Decrease Stay the Same Not enough information

6. In the data set “nortemp.dat” (see appendix from Chapter 2), data on pulse rates and body temperatures are provided for a sample of $n = 65$ healthy men and $n = 65$ health women. Do men and women have the same body temperature on average? Analyze this data set using SAS and do the following parts to answer this question.
- In order to test if body temperatures for men and women differ on average, set up the appropriate null and alternative hypothesis. Define the parameters of interest and state H_0 and H_a in terms of these parameters.
 - Suppose we test the hypothesis in part (a) using a significance level $\alpha = 0.05$. Assuming the variances for men and women are equal, determine the rejection region for this test. Sketch the t -density and shade the rejection region.
 - Run SAS to get the means and standard deviations for body temperatures of men and women in the sample and use these values to manually compute the t -test statistic assuming the variances for both groups are equal.
 - Run PROC TTEST in SAS to confirm you obtained the correct t -test statistic.
 - Does the t -test statistic fall in the rejection region from part (b)? State the conclusion of your test in the *context of this problem*.
 - What are the results of the test if we do not assume equal variances for men and women?
 - Compute a 95% confidence interval for the difference in mean temperatures between men and women. Write a sentence interpreting this estimate of the difference.
 - Re-do part (g) using a 99% confidence interval. Is this interval narrower or wider than the 95% confidence interval? Is zero in this interval? Comment on this.
7. Re-do the previous problem for the pulse rates of men and women.
8. Selective serotonin reuptake inhibitors (SSRIs) are used in treating depression. A study was conducted to investigate how depression is related to how well serotonin binds to serotonin transporters in the brain. In particular, it is hypothesized that depressed individuals will have lower binding potentials than non-depressed individuals. In order to test this hypothesis, binding potentials were estimated for depressed ($n_1 = 12$) and normal control ($n_2 = 12$) subjects from positron emission tomography (PET) scans of the brains. This study focuses the hippocampus region of interest (ROI). Use the SAS program 'binding.sas' in the Appendix to do this problem. For this problem, type a short (1-2 page) report that contains the following:
- Short introductory paragraph explaining the purpose of the study.
 - Sample statistics (means, standard deviations, etc.).
 - Results of the t -test.
 - A paragraph giving the conclusion of the t -test in the context of this study.
9. Refer to the previous problem on binding potentials study. Suppose a new study to compare depressed and normal control subjects is being planned. How many subjects in each group would be needed if the researchers want to detect a difference of 2 units in mean binding potential between the depressed and normal controls using a significance level $\alpha = 0.05$ and having power of at least 80%? (You can use the Russ Lenth webpage for this problem or any other statistical software – be sure to state how you found your solution though.)

10. A study was done to determine whether or not fire can be used as a viable management tool to increase the amount of forage available to deer. The question of interest is if fire changes calcium levels present in the soil? The experiment was conducted on 12 plots and calcium was measured on soil samples before and after the burn. Data are in units of kilograms per plot. (Source: Probability and Statistics for Engineers and Scientist by Walpole et al, 7th edition.)
- Define the parameters of interest here and set up the appropriate null and alternative hypotheses.
 - What sort of statistical test should be performed for this data?
 - Perform a t -test by running the SAS program below. Write a one-paragraph report explaining the results (give the sample statistics, t -test statistic, p -value, and conclusion in the context of the problem).
 - Compute and interpret a 95% confidence interval for the difference in mean calcium levels before and after burning the field.

```

*****/
options ls=76;
data fire;
input pre post;
diff=pre-post;
datalines;
50 9
50 18
82 45
64 18
82 18
73 9
77 32
54 9
23 18
45 9
36 9
54 9
;
run;
proc means;
run;

```

References

MacGregor, Markandu, Roulston, and Jones (1979), "Essential hypertension: effect of an oral inhibitor of angiotensin-converting enzyme," *British Medical Journal*, **2**, 1106-1109.

2 Appendix

Binding.sas

```

/*****
Binding potentials measured from Positron Emission Tomography (PET)
scans on depressed and normal control subjects in the hippocampus
region of the brain. The binding potential is a measure how well
serotonin binds to serotonin transporters. It is believed that
depressed subjects will have a lower binding potential.

Column 1: binding potential
Column 2: 0=depressed, 1=normal control
*****/
options ls=76 nodate;
proc format;      * this statement labels the output;
    value gpfmt 1='Normal Control' 0='Depressed';
run;
data binding;
input bp group;
format group gpfmt.;
datalines;
 6.31  0.00
 1.56  0.00
 6.23  0.00
 9.05  0.00
 5.78  0.00
 4.42  0.00
 4.59  0.00
 3.64  0.00
 8.11  0.00
10.49  0.00
 4.75  0.00
 3.47  0.00
11.21  1.00
 7.13  1.00
11.55  1.00
 8.85  1.00
 7.09  1.00
 5.25  1.00
13.85  1.00
 9.03  1.00
 4.95  1.00
 6.19  1.00
 7.16  1.00
 5.92  1.00
;

```

```
proc ttest;  
  class group;  
proc univariate plot normal;  
  var bp;  
  by group;  
run;
```