Chapter 8

Two-Sample Hypothesis Tests and Confidence Intervals

Chapter Objectives

- Carry out a two-sample t test for the difference between two population means.
- Compute and interpret a two-sample t confidence interval for the difference between two population means.
- Carry out a rank sum test for the difference between two population means.
- Decide which test (the t test or the rank sum test) is more appropriate for a given set of data.

Key Takeaways

- The two-sample t test is a parametric test for the difference between two population means that requires either that the samples are from normal populations or the sample sizes are both large. We can assess normality by graphing the data. A log transformation can make right skewed data more normal prior to conducting a t test.
- The rank sum test is a nonparametric test for the difference between two population means that doesn't require a normality assumption or large sample sizes.

8.1 Introduction

We've seen how to carry out hypothesis tests using a sample from a *single* population (Chapter 7). But environmental studies often involve testing for a difference between *two* populations using samples from those populations. Here are four examples.

Example 8.1: Two-Sample Test in Control-Impact Studies

In a *control-impact* study, we might suspect that the mean contaminant level is higher at the impact site than at the control site. The hypotheses would be

$$H_0 : \mu_x = \mu_y \tag{8.1}$$
$$H_a : \mu_x > \mu_y \tag{8.2}$$

where μ_x and and μ_y are the (unknown) population mean contaminant levels at the impact and control sites, respectively. The null hypothesis says there's no difference between mean contaminant levels at the two sites, and the alternative says the impact site's mean is higher.

Example 8.2: Two-Sample Test in Before-After Studies

In a *before-after* study, we might suspect that the site became contaminated as a result of the impact event. The hypotheses would be

$$\begin{aligned} H_0 : \mu_x &= \mu_y \\ H_a : \mu_x &> \mu_y \end{aligned}$$

where μ_x and μ_y are the (unknown) population mean contaminant levels after and before the impact event, respectively. The null hypothesis says the impact event had no effect on the contaminant levels, and the alternative says it increased them.

Example 8.3: Two-Sample Test in Designed Experiments

Laboratory and field *experiments* often involve randomly assigning experimental units to treatment and control conditions and comparing their responses. To decide if there's any difference in the effects of the two conditions, we'd test

$$\begin{aligned} H_0 : \mu_x &= \mu_y \\ H_a : \mu_x &\neq \mu_y \end{aligned}$$

where μ_x and μ_y are the true (unknown) mean responses to the two conditions. The null hypothesis says the treatment has no effect on the response, and the alternative says it has an effect.

Example 8.4: Two-Sample Test in Reclamation Effectiveness Studies

Studies to evaluate the effectiveness of the reclamation (cleanup) of a contaminated site typically take one of two forms, analogous to *before-after* and *control-impact* assessment studies.

If pre-contamination data are available from the site, they're compared to data collected after the reclamation effort. Here we might test

$$\begin{aligned} H_0 : \mu_x &\leq \mu_y \quad (8.3) \\ H_a : \mu_x &> \mu_y \quad (8.4) \end{aligned}$$

where μ_x and μ_y are the site's pre-contamination and post-reclamation population mean contaminant levels, respectively. The null hypothesis says that despite the reclamation effort, the contaminant level is still as high as (or higher than) it was before the contamination event. The alternative says it's even lower than it was before the event.

If no pre-contamination data are available, data from a nearby uncontaminated control site would be used instead. In this case, μ_x and μ_y would be the population mean contaminant levels at the control and reclaimed sites, respectively.

We'll look at three two-sample hypothesis test procedures:

1. The two-sample t test
2. The pooled two-sample t test

3. The rank sum test

All three are tests for a difference between two population means $\boldsymbol{\mu_x}$ and $\boldsymbol{\mu_y}$. The first two are parametric tests, requiring either that samples are from two normal populations or the sample sizes are large. The second requires furthermore that the population standard deviations are equal. The third test is a non-parametric test, so it doesn't have the normality or large sample size requirement. Figure 8.1 below shows these three situations graphically.

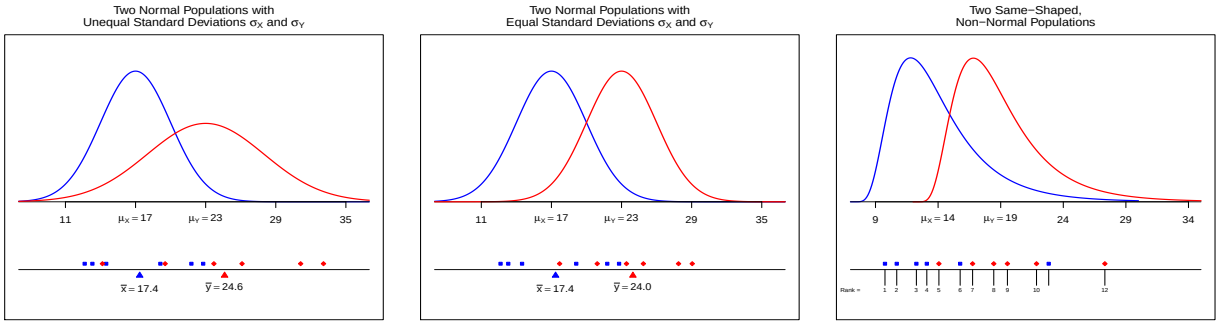


Figure 8.1: Normal populations with unequal (left) and equal (center) standard deviations. The two-sample t test could be used in either case, but the pooled t test would only be appropriate in the second case. Non-normal populations (right). Here, the rank sum test would be used. Random samples are shown as dots below their color-matched density curves.

Note: Often a set of hypotheses such as

$$\begin{aligned} H_0 &: \mu_x = \mu_y \\ H_a &: \mu_x > \mu_y \end{aligned}$$

is stated as

$$\begin{aligned} H_0 &: \mu_x - \mu_y = 0 \\ H_a &: \mu_x - \mu_y > 0 \end{aligned}$$

The difference $\boldsymbol{\mu_x - \mu_y}$ between the population means is sometimes called the ***effect size***. If μ_x and μ_y are the true (unknown) mean responses to treatment and control conditions in an experiment, the null hypothesis says that the effect size is zero, which is to say the treatment has no effect.

In addition to looking at the above three hypothesis testing procedures, we'll also look at procedures for constructing confidence intervals for an (unknown) effect size $\mu_x - \mu_y$.

8.2 Sampling Distribution of $\bar{X} - \bar{Y}$

8.2.1 Introduction and Notation

The two-sample t test is based on the means of the samples drawn from two populations. More formally, suppose $\boldsymbol{X_1, X_2, \ldots, X_{n_x}}$ and $\boldsymbol{Y_1, Y_2, \ldots, Y_{n_y}}$ are random samples from two populations. If $\boldsymbol{\bar{X}}$ and $\boldsymbol{\bar{Y}}$ are the sample means, the test is based on their difference, $\boldsymbol{\bar{X} - \bar{Y}}$, which is called the ***estimated effect size*** and is an estimate of the true effect size $\mu_x - \mu_y$. The two sample sizes, $\boldsymbol{n_x}$ and $\boldsymbol{n_y}$, don't have to be the same. Finally, we'll denote the two sample standard deviations by $\boldsymbol{S_x}$ and $\boldsymbol{S_y}$.

The next example illustrates the need to account for sampling variation in $\bar{X}$ and $\bar{Y}$ when deciding if two population means differ.

Example 8.5: Sampling Error of $\bar{X} - \bar{Y}$

A bioremediation process is to be carried out at a contaminated site by injecting nutrients into the soil to maintain a microbial community that will biodegrade the contaminants. The agency responsible for the cleanup has had success using a standard nutrient mixture, but suspects that a more expensive mixture might work better.

In an experiment to test this hypothesis, twenty plots of land are randomized to two treatment groups (ten per group), one receiving the more expensive mixture and the other the standard one. The resulting biodegradation rates (percent reduction in the contaminant) are shown below.

Biodegradation Rates

More Expensive Nutrient Mixture	Standard Nutrient Mixture
9.7	9.7
10.1	8.5
9.1	9.4
8.8	9.3
9.2	7.9
10.0	8.9
9.4	8.8
8.8	8.5
9.9	8.9
8.2	8.8

The summary statistics for the two treatment groups are

$$\begin{array}{ll} n_x = 10 & n_y = 10 \\ \bar{X} = 9.32 & \bar{Y} = 8.87 \\ S_x = 0.62 & S_y = 0.51 \end{array}$$

where X denotes the more expensive mixture and Y the standard one. The agency wants to decide if the more expensive mixture leads to higher biodegradation rates. Thus the hypotheses are

$$H_0 : \mu_x - \mu_y = 0 \tag{8.5}$$
$$H_a : \mu_x - \mu_y > 0 \tag{8.6}$$

where μ_x and μ_y are the true (unknown) population mean biodegradation rates for the more expensive and standard mixtures, respectively. The decision will be based on the *estimated effect size*

$$\bar{X} - \bar{Y} = 9.32 - 8.87 = 0.45,$$

which, because it's positive, provides some evidence that the more expensive mixture works better. But is it really better, or could the result be explained by sampling variation (chance)? A hypothesis test (Example 8.6) will help answer this question.

8.2.2 Mean and Standard Error of the Sampling Distribution of $\bar{X} - \bar{Y}$

To decide whether an observed difference between two sample means is larger than can be explained by chance, we'll need the sampling distribution that the statistic $\bar{X} - \bar{Y}$ would follow if the samples were from two populations whose means are equal. The mean and standard error of the sampling distribution are given in the following fact.

Fact 8.1 Suppose $X_1, X_2, \ldots, X_{n_x}$ and $Y_1, Y_2, \ldots, Y_{n_y}$ are two independent random samples from populations whose means are μ_x and μ_y and whose standard deviations are $\boldsymbol{\sigma_x}$ and $\boldsymbol{\sigma_y}$. Then the sampling distribution of the statistic $\bar{X} - \bar{Y}$ has mean $\boldsymbol{\mu_{\bar{X}-\bar{Y}}}$ and ***standard error*** $\boldsymbol{\sigma_{\bar{X}-\bar{Y}}}$ given by

$$\mu_{\bar{X}-\bar{Y}} \;=\; \mu_x - \mu_y \tag{8.7}$$

and

$$\sigma_{\bar{X}-\bar{Y}} \;=\; \sqrt{\frac{\sigma_x^2}{n_x} + \frac{\sigma_y^2}{n_y}}. \tag{8.8}$$

The mean $\mu_{\bar{X}-\bar{Y}}$ of the distribution is the *long-run average* value that you'd get for $\bar{X} - \bar{Y}$ if you repeatedly took samples from the two populations. Thus (8.7) says that, *on average*, the difference between the sample means will equal the difference between the population means, even though a *particular* $\bar{X} - \bar{Y}$ almost certainly won't equal $\mu_x - \mu_y$ exactly. In other words, when $\bar{X} - \bar{Y}$ is used as an *estimator* of $\mu_x - \mu_y$, it neither systematically overestimates nor systematically underestimates $\mu_x - \mu_y$. On average, it's right on target.

The discrepancy between a *particular* estimate $\bar{X} - \bar{Y}$ and the true value $\mu_x - \mu_y$ is called the ***sampling error*** of the estimate.

Sampling Error of $\bar{X} - \bar{Y}$:

$$\text{Sampling Error} \;=\; \bar{X} - \bar{Y} \;-\; (\mu_x - \mu_y)\,.$$

The sampling error measures how far off the mark a *particular* estimate of the difference between population means is.

The standard error $\sigma_{\bar{X}-\bar{Y}}$ is interpreted as the size of a *typical* sampling error (for given sample sizes n_x and n_y), and it serves as a measure of how precise $\bar{X} - \bar{Y}$ is as an estimator of $\mu_x - \mu_y$. A smaller standard error indicates a more precise estimator. The standard error will be small if either:

- The population standard deviations σ_x and σ_y are both small, or
- The sample sizes n_x and n_y are both large.

8.2.3 Normality of the Sampling Distribution of $\bar{X} - \bar{Y}$

We know (from Chapter 5) that $\bar{X}$ and $\bar{Y}$ each individually follows a normal distribution (at least approximately) when either the samples are drawn from normal populations or the sample sizes are large. The next fact tells us that the statistic $\bar{X} - \bar{Y}$ follows a normal distribution too.

Fact 8.2 Suppose $X_1, X_2, \ldots, X_{n_x}$ and $Y_1, Y_2, \ldots, Y_{n_y}$ are two independent random samples from populations whose means are μ_x and μ_y and whose standard deviations are σ_X and σ_Y. If both populations are *normal*, then

$$\bar{X} - \bar{Y} \;\sim\; \mathrm{N}\left(\mu_x - \mu_y,\; \sigma_{\bar{X}-\bar{Y}}\right),$$

where the standard error $\sigma_{\bar{X}-\bar{Y}}$ is given in Fact 8.1.

Furthermore, even if one or both populations are *non-normal*, the statistic $\bar{X} - \bar{Y}$ still follows the normal distribution *approximately* as long the sample sizes n_x and n_y are both *large*.

As a consequence, if we standardize $\bar{X} - \bar{Y}$, the resulting variable follows a standard normal distribution, that is,

$$Z \;=\; \frac{\bar{X} - \bar{Y} \;-\; (\mu_x - \mu_y)}{\sigma_{\bar{X}-\bar{Y}}} \;\sim\; \mathrm{N}(0,\,1)\,. \tag{8.9}$$

The variable Z is measured in *standard units*, so its value indicates how many standard errors the statistic $\bar{X} - \bar{Y}$ is away from $\mu_x - \mu_y$. Normality of $\bar{X} - \bar{Y}$ is a consequence of Facts 4.3 and 5.5 (Chapters 4 and 5) that linear functions and sums of normal random variables, such as $\bar{X}$ and $\bar{Y}$, are themselves normally distributed.

8.3 The Two-Sample *t* Test

The ***two-sample t test*** is a test for a difference between two population means μ_x and μ_y.

8.3.1 The Two-Sample *t* Test Procedure

We'll want to test the null hypothesis

$$H_0 : \mu_x - \mu_y \;=\; 0$$

that there's no difference between the two means versus one of the three alternative hypotheses

1. $H_a : \mu_x - \mu_y > 0$ (upper-tailed test)
2. $H_a : \mu_x - \mu_y < 0$ (lower-tailed test)
3. $H_a : \mu_x - \mu_y \neq 0$ (two-tailed test)

the choice of which will depend on what we're trying to substantiate by conducting the study.

In most studies, the population standard deviations won't be known, so they'll have to be estimated by the sample standard deviations, which we'll denote by $\boldsymbol{S_x}$ and $\boldsymbol{S_y}$. The ***estimated standard error*** of the statistic $\bar{X} - \bar{Y}$, denoted $\boldsymbol{S_{\bar{X}-\bar{Y}}}$, is then

$$S_{\bar{X}-\bar{Y}} \;=\; \sqrt{\frac{S_x^2}{n_x} + \frac{S_y^2}{n_y}}$$

When we standardize $\bar{X} - \bar{Y}$ using the estimated standard error $S_{\bar{X}-\bar{Y}}$ in place of the true value $\sigma_{\bar{X}-\bar{Y}}$, the resulting variable follows a t distribution.

Fact 8.3 Suppose that $X_1, X_2, \ldots, X_{n_x}$ and $Y_1, Y_2, \ldots, Y_{n_y}$ are two independent random samples from populations whose means are μ_x and μ_y. If the populations are both normal, then the random variable

$$T \;=\; \frac{\bar{X} - \bar{Y} - (\mu_x - \mu_y)}{S_{\bar{X}-\bar{Y}}}, \qquad \text{where} \qquad S_{\bar{X}-\bar{Y}} \;=\; \sqrt{\frac{S_x^2}{n_x} + \frac{S_y^2}{n_y}}, \tag{8.10}$$

follows a t distribution with degrees of freedom given by

$$\text{df} \;=\; \frac{\left(\frac{S_x^2}{n_x}+\frac{S_y^2}{n_y}\right)^2}{\frac{(S_x^2/n_x)^2}{n_x-1}+\frac{(S_y^2/n_y)^2}{n_y-1}}, \tag{8.11}$$

which should be rounded *down* to the nearest integer.

Furthermore, even if one or both populations are *non-normal*, the variable T still follows the t distribution *approximately* as long the sample sizes n_x and n_y are both *large*.

The ***two-sample t test statistic*** is obtained by replacing $\mu_x - \mu_y$ in (8.10) by its null-hypothesized value zero.

Two-Sample t Test Statistic:

$$t \;=\; \frac{\bar{X}-\bar{Y}\;-\;0}{S_{\bar{X}-\bar{Y}}} \;=\; \frac{\bar{X}-\bar{Y}}{S_{\bar{X}-\bar{Y}}} \tag{8.12}$$

where

$$S_{\bar{X}-\bar{Y}} \;=\; \sqrt{\frac{S_x^2}{n_x}+\frac{S_y^2}{n_y}}\,.$$

If H_0 was true, and μ_x equal to μ_y, we'd expect the sample means to be approximately equal too, in which case $\bar{X}-\bar{Y}$ would be close to zero and t would be near zero too. Any discrepancy between t and zero would be due purely to chance (sampling variation). On the other hand, if H_a was true, and $\mu_x - \mu_y$ different from zero in the direction specified by H_a, we'd expect $\bar{X}-\bar{Y}$ to differ from zero in that same direction, in which case t would differ from zero in that direction too. Moreover, because the denominator of t is an estimate of the standard error of $\bar{X}-\bar{Y}$, t measures (approximately) how many standard errors $\bar{X}-\bar{Y}$ is away from zero. Therefore, we have the following.

1. *Large positive* values of t provide evidence in favor of $H_a : \mu_x - \mu_y > 0$.
2. *Large negative* values of t provide evidence in favor of $H_a : \mu_x - \mu_y < 0$.
3. *Both large positive and large negative* values of t provide evidence in favor of $H_a : \mu_x - \mu_y \neq 0$.

To decide whether an observed value of t provides statistically significant evidence to support the alternative hypothesis, we'll determine if it's among the values that would be unlikely to occur just by chance under the null hypothesis. For this, we'll need the sampling distribution that t would follow if the null was true. But because t is obtained by replacing $\mu_x - \mu_y$ in (8.10) by its null-hypothesized value zero, we have the following.

Sampling Distribution of t Under $\boldsymbol{H_0}$: Suppose that $X_1, X_2, \ldots, X_{n_x}$ and $Y_1, Y_2, \ldots, Y_{n_y}$ are two independent random samples from populations whose means are μ_x and μ_y, and that either the populations are both normal or n_x and n_y are both large. Then when

$$H_0 : \mu_x - \mu_y \;=\; 0$$

is true,

$$t \sim t(\text{df})$$

where the degrees of freedom are given by (8.11).

Values of t in the tail of the t(df) distribution, in the direction (or directions) specified by the alternative hypothesis, would be unlikely to occur by chance under the null, and would therefore support the alternative. Thus p-values (and critical values for the rejection region approach) are obtained from the corresponding tail (or tails) of the t(df) distribution, as summarized below.

Two-Sample t Test for μ_x and μ_y

Assumptions: The data $x_1, x_2, \ldots, x_{n_x}$ and $y_1, y_2, \ldots, y_{n_y}$ are independent random samples from two populations and either the populations are normal or n_x and n_y are large.

Null hypothesis: $H_0 : \mu_x - \mu_y = 0$.

Test statistic value: $t = \frac{\bar{X} - \bar{Y}}{\sqrt{S_x^2/n_x + S_y^2/n_y}}$.

Decision rule: Reject H_0 if p-value $< \alpha$ or t is in rejection region.

Alternative hypothesis	P-value = area under t distribution with d.f. given by (8.11):	Rejection region = t values such that:*
$H_a : \mu_x - \mu_y > 0$	to the right of t	$t > t_{\alpha,\text{df}}$
$H_a : \mu_x - \mu_y < 0$	to the left of t	$t < -t_{\alpha,\text{df}}$
$H_a : \mu_x - \mu_y \neq 0$	to the left of $-\lvert t\rvert$ and right of $\lvert t\rvert$	$t > t_{\alpha/2,\text{df}}$ or $t < -t_{\alpha/2,\text{df}}$

* $t_{\alpha,\text{df}}$ is the $100(1-\alpha)$th percentile of the t distribution with d.f. given by (8.11).

Note: Usually $n \geq 30$ is large enough for the t test to be valid when the sample is from a non-normal population. However, if the population is *very* skewed, an even larger sample size may be required. On the other hand, if the population distribution is fairly symmetric, and therefore closer to a normal distribution, a sample size of $n = 10$ or 15 may suffice.

8.3.2 Carrying Out the Two-Sample t Test

Here are some examples that illustrate the two-sample t test. Example 8.6 involves a one-sided alternative hypothesis and Example 8.7 a two-sided one.

Example 8.6: Two-Sample t Test

In the experiment comparing biodegradation rates for standard and more expensive nutrient mixtures (Example 8.5), the alternative hypothesis (8.6) is one-sided and the test is upper-tailed. Here are side-by-side boxplots of the data.

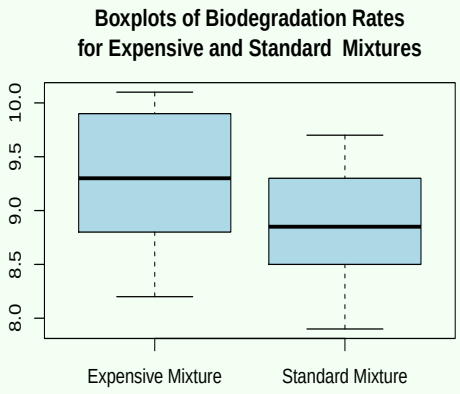


Figure 8.2: Boxplots of biodegradation rates using expensive and standard nutrient mixtures.

The boxplots suggest that the more expensive mixture leads to higher biodegradation rates than the standard one. We want to know if this observed difference is statistically significant. The normal probability plots (below) indicate that the normality assumption for is met both samples, so the use of a two-sample t test is justified.

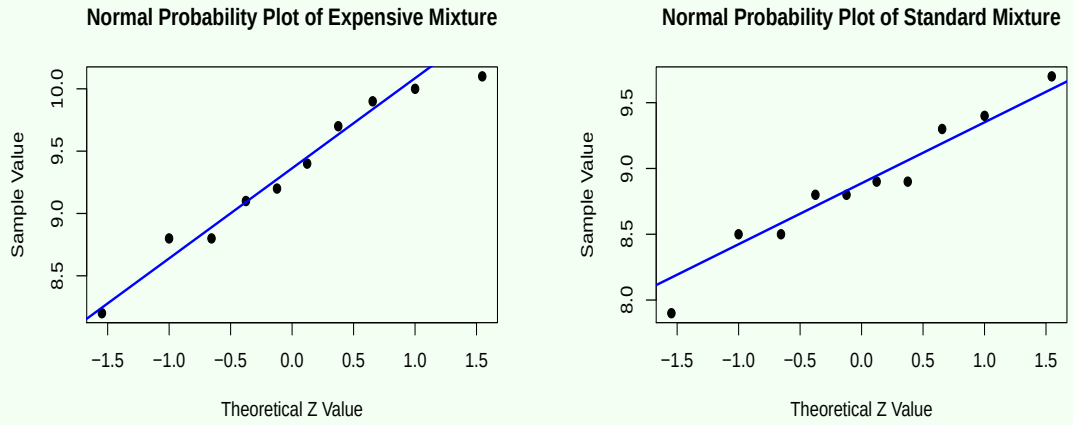


Figure 8.3: Normal probability plots of degradation rates using the more expensive (left) and standard (right) nutrient mixtures.

The estimated standard error of $\bar{X} - \bar{Y}$ (using the summary statistics from Example 8.5) is

$$S_{\bar{X}-\bar{Y}} = \sqrt{\frac{0.63^2}{10} + \frac{0.50^2}{10}} = 0.25,$$

so the observed test statistic value is

$$t = \frac{9.32 - 8.87}{0.25} = 1.80.$$

Thus the observed difference between $\bar{X}$ and $\bar{Y}$, 0.45, is about 1.80 standard errors above zero. We want to know whether this result could be explained by chance (sampling variation) if in reality there was no difference between the degradation rates for the two nutrient mixtures.

For this upper-tailed test, the p-value is the tail area to the right of 1.80 under the t distribution with degrees of freedom

$$\text{df} = \frac{\left(\frac{s_x^2}{n_x} + \frac{s_y^2}{n_y}\right)^2}{\frac{(s_x^2/n_x)^2}{n_x - 1} + \frac{(s_y^2/n_y)^2}{n_y - 1}} = \frac{\left(\frac{0.63^2}{10} + \frac{0.50^2}{10}\right)^2}{\frac{(0.63^2/10)^2}{10-1} + \frac{(0.50^2/10)^2}{10-1}} = 17.12,$$

which we round down to 17. From a table of t distribution tail areas, the p-value is found to be 0.0448. Using a level of significance $\alpha = 0.05$, we reject H_0 and conclude that the observed difference between degradation rates is statistically significant, and therefore not just due to chance. In other words, it appears that the more expensive mixture does produce higher degradation rates, on average, than the standard one.

Example 8.7: Two-Sample *t* Test

Pollutants from motor vehicles, such as particulates, heavy metals, and hydrocarbons, can accumulate on highway surfaces. These pollutants are then transported to nearby soils or bodies of water via runoff during wet weather.

A study was carried out to assess the impact of pollutants in highway runoff on the water and roadside soils in the Pear River Delta, a rapidly developing region in South China [5]. Highway runoff was sampled on 11 wet days at an urban site and 7 wet days at a rural site. For each runoff water specimen, several pollutants were measured. The table below shows the nickel (Ni) concentrations (μg/L).

Ni in Highway Runoff

Urban Site	Rural Site
20.3	14.7
23.1	21.5
37.7	10.2
20.1	6.5
30.5	7.6
33.4	12.6
11.8	20.9
18.8	
15.3	
26.7	
10.9	

Side-by-side boxplots of the data are shown below.

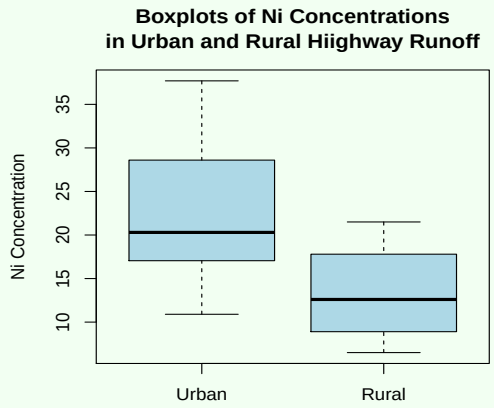


Figure 8.4: Boxplots of urban and rural nickel concentrations in highway runoff in Pear River Delta, South China.

The boxplots indicate that the Ni concentrations tend to be higher in the urban runoff. The hypothesis test will tell us whether this difference is statistically significant. Normal probability plots of the two samples are below.

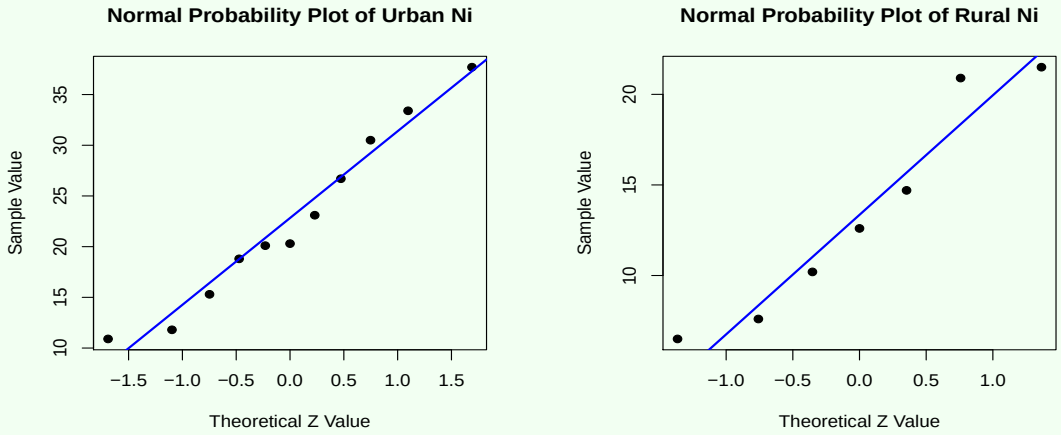


Figure 8.5: Normal probability plot of urban (left) and rural (right) nickel concentrations in highway runoff in Pear River Delta, South China.

The plots give no indications of non-normality in either sample, so the assumptions required for the two-sample t test appear to be met. The summary statistics for the two samples are

Urban	Rural
$n_x = 11$	$n_y = 7$
$\bar{X} = 22.6$	$\bar{Y} = 13.4$
$S_x = 8.7$	$S_y = 6.0$

where X denotes the urban sample and Y the rural sample.

Because the goal of the study was to decide if there's *any* difference in urban and rural Ni concentrations, the hypotheses are

$$\begin{aligned} H_0 : \mu_x - \mu_y &= 0 \\ H_a : \mu_x - \mu_y &\neq 0. \end{aligned}$$

where μ_x and μ_y are the true (unknown) population mean Ni concentrations in urban and rural runoff, respectively. The null hypothesis says there's no difference, and the alternative says there is a difference. The two sample means differ by $\bar{X} - \bar{Y} = 22.6 - 13.4 = 9.2$ μg/L, suggesting that Ni concentrations at the urban site are about 9.2 μg/L higher, on average, than those at the rural site.

The estimated standard error of $\bar{X} - \bar{Y}$ is

$$S_{\bar{X}-\bar{Y}} = \sqrt{\frac{8.7^2}{11} + \frac{6.0^2}{7}} = 3.47,$$

so the observed test statistic value is

$$t = \frac{22.6 - 13.4}{3.47} = 2.65.$$

This says the observed difference, 9.2, is about 2.65 standard errors away from zero. The p-value is the probability that we'd get a difference this far away from zero just by chance if the null hypothesis was true. It's the sum of the tail areas to the left of -2.65 and right of 2.65 under the t distribution with degrees of freedom

$$\text{df} = \frac{\left(\frac{s_x^2}{n_x} + \frac{s_y^2}{n_y}\right)^2}{\frac{(s_x^2/n_x)^2}{n_x - 1} + \frac{(s_y^2/n_y)^2}{n_y - 1}} = \frac{\left(\frac{8.7^2}{11} + \frac{6.0^2}{7}\right)^2}{\frac{(8.7^2/11)^2}{11-1} + \frac{(6.0^2/7)^2}{7-1}} = 15.81,$$

which we round *down* to 15. From a table of t distribution tail areas, the p-value is found to be $2(0.0091) = 0.0182$. Using a level of significance $\alpha = 0.05$, we reject the null hypothesis and conclude that urban and rural mean Ni concentrations differ. More precisely, it appears that they're *higher* at the urban site than at the rural one.

8.4 Two-Sample *t* Confidence Intervals

The difference between two true means, $\mu_x - \mu_y$, is sometimes called the ***effect size***. In an experiment, where μ_x and μ_y are the true mean responses to a treatment and a control, $\mu_x - \mu_y$ represents the size of the treatment effect. In an impact assessment study, where μ_x and μ_y are the true mean contaminant levels at the impact and control sites, $\mu_x - \mu_y$ represents effect of the impact event.

We can estimate the effect size by the *point estimator* $\bar{X} - \bar{Y}$, but we usually also want to include a margin of error to convey how precise the estimate is.

8.4.1 Computing and Interpreting a Two-Sample *t* Confidence Interval

A confidence interval for an effect size $\mu_x - \mu_y$ consists of an estimate and its margin of error. For the ***two-sample t confidence interval for $\mu_x - \mu_y$***, given below, we'll make the same assumptions as those required for the two-sample t test.

Two-Sample t Confidence Interval: Suppose $X_1, X_2, \ldots, X_{n_x}$ and $Y_1, Y_2, \ldots, Y_{n_y}$ are two independent random samples from populations whose means are μ_x and μ_y. Suppose also that either the populations are both normal or the sample sizes n_x and n_y are both large.

Then a **$100(1-\alpha)\%$ *two-sample t confidence interval for $\mu_x - \mu_y$*** is

$$\bar{X} - \bar{Y} \pm t_{\alpha/2,\mathrm{df}}\, S_{\bar{X}-\bar{Y}}, \tag{8.13}$$

where

$$S_{\bar{X}-\bar{Y}} \;=\; \sqrt{\frac{S_x^2}{n_x} + \frac{S_y^2}{n_y}}$$

and the degrees of freedom df is given by (8.11).

We can be $100(1-\alpha)\%$ confident that the true (unknown) effect size $\mu_x - \mu_y$ will be contained in this interval. We'll see how the confidence interval formula was derived in Subsection 8.4.3.

The *margin of error* is the "plus or minus" part of the confidence interval.

Margin of Error: For the two-sample t confidence interval (8.13), the margin of error is

$$\text{Margin of Error} \;=\; t_{\alpha/2,\mathrm{df}}\, S_{\bar{X}-\bar{Y}} \;=\; t_{\alpha/2,\mathrm{df}} \sqrt{\frac{S_x^2}{n_x} + \frac{S_y^2}{n_y}}.$$

The margin of error measures the degree of precision in the estimate $\bar{X} - \bar{Y}$ of the true effect size $\mu_x - \mu_y$. A smaller margin of error means a more precise estimate.

Example 8.8: Two-Sample t Confidence Interval

Consider again the study of Ni in urban and rural highway runoff (Example 8.7). The hypothesis test indicated that there's a difference between the Ni concentrations in urban and rural runoff. We can estimate the *size* of the difference, $\mu_x - \mu_y$, using the point estimate

$$\bar{X} - \bar{Y} \;=\; 22.6 - 13.4 \;=\; 9.2,$$

but this provides no indication of how far off the mark the estimate might be.

The estimated standard error is

$$S_{\bar{X}-\bar{Y}} \;=\; \sqrt{\frac{8.7^2}{11} + \frac{6.0^2}{7}} \;=\; 3.47,$$

so a 95% confidence interval for $\mu_x - \mu_y$ is

$$\begin{aligned} \bar{X} - \bar{Y} \pm t_{\alpha/2,\mathrm{df}}\, S_{\bar{X}-\bar{Y}} &= 22.6 - 13.4 \pm 2.13\,(3.47) \\ &= 9.2 \pm 7.39 \\ &= (1.81,\ 16.59), \end{aligned}$$

where the t critical value $t_{0.025,15} = 2.13$ was obtained from a t table and the degrees of freedom, $\mathrm{df} = 15$, was determined in Example 8.7.

The margin of error, 7.39, indicates the degree of precision in the estimate 9.2. It tells us that the estimate might be off the mark by up to 7.39 μg/L.

The confidence interval gives a range of estimates of the true difference $\mu_x - \mu_y$, and we can be 95% confident that the true difference lies somewhere between 1.81 and 16.59 μg/L. Because the entire interval lies above zero, there's convincing evidence that the true difference is greater than zero, although it might be as small as 1.81 μg/L.

8.4.2 Using Confidence Intervals to Test Two-Sided Hypotheses

In Example 8.8 the confidence interval fell entirely above zero, and we concluded that zero wasn't a plausible value for the difference $\mu_x - \mu_y$. We can use the two-sample t confidence interval, with confidence level $100(1-\alpha)\%$, to test

$$\begin{aligned} H_0 : \mu_x - \mu_y &= 0 \\ H_a : \mu_x - \mu_y &\neq 0 \end{aligned}$$

with significance level α, by invoking the decision rule

> Reject H_0 if the confidence interval doesn't contain zero
> Fail to reject H_0 if it does contain zero

and the conclusion will be the same as if we had carried out the hypothesis test. In Example 8.8, using the confidence interval method, we'd reject H_0, which is the same conclusion we came to using the two-sample t test in Example 8.7.

The argument used to confirm that the two-sample t confidence interval approach to hypothesis testing gives the same result as the two-sample t test is similar to the one given in Section 7.5 of Chapter 7 to justify the use of the one-sample t confidence interval for conducting a one-sample t test.

8.4.3 Derivation of the Two-Sample *t* Confidence Interval Formula

The derivation of the two-sample t confidence interval formula is similar to that of the one-sample z and t confidence interval formulas given in Chapter 6. From Fact 8.3,

$$1 - \alpha \;=\; P\left(-t_{\alpha/2,\mathrm{df}} \;\leq\; \frac{\bar{X} - \bar{Y} - (\mu_x - \mu_y)}{S_{\bar{X}-\bar{Y}}} \;\leq\; t_{\alpha/2,\mathrm{df}}\right).$$

After rearranging terms, this can be written as

$$1 - \alpha \;=\; P\left(\bar{X} - \bar{Y} - t_{\alpha/2,\mathrm{df}}\, S_{\bar{X}-\bar{Y}} \;\leq\; \mu_x - \mu_y \;\leq\; \bar{X} - \bar{Y} + t_{\alpha/2,\mathrm{df}}\, S_{\bar{X}-\bar{Y}}\right),$$

from which the confidence interval formula (8.13) follows.

8.5 The Pooled Two-Sample *t* Test and Confidence Interval

8.5.1 Introduction

Sometimes when comparing two population means, it's reasonable assume that the population standard deviations are equal. When σ_X and σ_Y are equal, we can carry out a version of the two-sample t test called the ***pooled two-sample t test***. The difference between this test and the one covered in the last section lies in how the standard error of $\bar{X} - \bar{Y}$ is estimated.

8.5.2 The Pooled Two-Sample t Test

When the two population standard deviations are equal, we can just write $\boldsymbol{\sigma}$ for both σ_X and σ_Y. In this case the sample standard deviations S_x and S_y both estimate the same value σ, so we can ***pool*** (combine) the two samples to get a single, more precise estimate. The ***pooled sample variance*** and ***standard deviation***, denoted $\boldsymbol{S_{\mathrm{p}}^2}$ and $\boldsymbol{S_{\mathrm{p}}}$, are defined as follows.

Pooled Sample Variance and Standard Deviation: For random samples from X and Y populations whose standard deviations are equal, the pooled sample variance is

$$S_{\mathrm{p}}^2 \;=\; \frac{(n_x-1)S_x^2 \,+\, (n_y-1)S_y^2}{n_x+n_y-2} \tag{8.14}$$

and the pooled sample standard deviation is

$$S_{\mathrm{p}} \;=\; \sqrt{\frac{(n_x-1)S_x^2 \,+\, (n_y-1)S_y^2}{n_x+n_y-2}}\,. \tag{8.15}$$

The statistic S_{p} estimates the common population standard deviation σ.

For intuition, note that when the two sample sizes are equal, say to n, it's easy to show that

$$S_{\mathrm{p}}^2 \;=\; \frac{S_x^2+S_y^2}{2}\,,$$

the average of the two sample variances. When they *aren't* equal, S_{p}^2 is a *weighted* average of the sample variances.

The *pooled two-sample t test* and *confidence interval* will use the following fact.

Fact 8.4 Suppose that $X_1, X_2, \ldots, X_{n_x}$ and $Y_1, Y_2, \ldots, Y_{n_y}$ are independent random samples from populations whose means are μ_x and μ_y and whose standard deviations are equal. Suppose also that either the populations are both normal or the sample sizes n_x and n_y are both large.

Let

$$S_{\mathrm{p},\,\bar{X}-\bar{Y}} \;=\; \sqrt{\frac{S_{\mathrm{p}}^2}{n_x}+\frac{S_{\mathrm{p}}^2}{n_y}}$$

be the estimated standard error of $\bar{X}-\bar{Y}$.

Then the random variable

$$\frac{\bar{X}-\bar{Y}-(\mu_x-\mu_y)}{S_{\mathrm{p},\,\bar{X}-\bar{Y}}}, \tag{8.16}$$

follows (at least approximately) a t distribution with n_x+n_y-2 degrees of freedom.

Now consider testing the null hypothesis

$$H_0 : \mu_x-\mu_y \;=\; 0.$$

The ***pooled two-sample t test statistic*** is obtained by replacing $\mu_x-\mu_y$ in (8.16) by its null-hypothesized value zero.

Pooled Two-Sample t Test Statistic:

$$t \;=\; \frac{\bar{X}-\bar{Y}\;-\;0}{S_{\mathrm{p},\,\bar{X}-\bar{Y}}} \;=\; \frac{\bar{X}-\bar{Y}}{S_{\mathrm{p},\,\bar{X}-\bar{Y}}}$$

where

$$S_{\mathrm{p},\,\bar{X}-\bar{Y}} \;=\; \sqrt{\frac{S_{\mathrm{p}}^2}{n_x}+\frac{S_{\mathrm{p}}^2}{n_y}}$$

From Fact 8.4, we get the sampling distribution of the pooled t test statistic under H_0.

Sampling Distribution of t Under H_0: Suppose that $X_1, X_2, \ldots, X_{n_x}$ and $Y_1, Y_2, \ldots, Y_{n_y}$ are independent random samples from populations whose means are μ_x and μ_y and whose standard deviations are equal. Suppose also that either the populations are both normal or the sample sizes n_x and n_y are both large. Then when

$$H_0 : \mu_x - \mu_y \;=\; 0$$

is true,

$$t \;\sim\; t(n_x + n_y - 2).$$

Values of t that differ from zero in the direction specified by H_a count as evidence in favor of H_a. P-values (and critical values for the rejection region approach) are obtained from the corresponding tail (or tails) of the $t(n_x + n_y - 2)$ distribution, as summarized below.

Pooled Two-Sample t Test for μ_x and μ_y

Assumptions: The data $x_1, x_2, \ldots, x_{n_x}$ and $y_1, y_2, \ldots, y_{n_y}$ are independent random samples from populations whose standard deviations are equal, and either the two populations are *normal* or n_x and n_y are large.

Null hypothesis: $H_0 : \mu_x - \mu_y = 0$.

Test statistic value: $t = \frac{\bar{X}-\bar{Y}}{\sqrt{s_{\mathrm{p}}^2/n_x + S_{\mathrm{p}}^2/n_y}}$.

Decision rule: Reject H_0 if p-value $< \alpha$ or t is in rejection region.

Alternative hypothesis	P-value = area under t-distribution with $n_x + n_y - 2$ d.f.:	Rejection region = t values such that:*
$H_a : \mu_x - \mu_y > 0$	to the right of t	$t > t_{\alpha, n_x+n_y-2}$
$H_a : \mu_x - \mu_y < 0$	to the left of t	$t < -t_{\alpha, n_x+n_y-2}$
$H_a : \mu_x - \mu_y \neq 0$	to the left of $-\lvert t\rvert$ and right of $\lvert t\rvert$	$t > t_{\alpha/2, n_x+n_y-2}$ or $t < -t_{\alpha/2, n_x+n_y-2}$

* t_{α, n_x+n_y-2} is the $100(1-\alpha)$th percentile of the t distribution with $n_x + n_y - 2$ d.f.

In practice, we'll rarely know whether two population standard deviations are equal, but we can gauge

the plausibility of this assumption using the two sample standard deviations. It's reasonable to assume that the population standard deviations are equal as long as the larger sample standard deviation isn't more than twice as large as the smaller one.

Example 8.9: Pooled Two-Sample t Test

Storm water runoff from roads contains heavy metals, both dissolved in the water and bound to particulates. Water treatment detention ponds remove some of the particulate-bound metals by allowing them to settle, but small particles and dissolved metals pass through the ponds and must then be filtered out using specially designed filtration systems.

A laboratory experiment was carried out to compare the efficiencies of different filter systems for removing metals from storm water [4]. The filters consisted of different mixtures of natural materials. We'll compare the results for two mixtures. The first, Filter A, consisted of 80% natural opoka (a calcium silicate rock) and 20% zeolite (a mineral). The second, Filter B, consisted of 50% burned opoka and 50% zeolite.

Fifteen columns were filled with the Filter A mixture and fifteen with the Filter B mixture. Then storm water was poured through each column. The response variable is the percent reduction in chromium (Cr). Here are the summary statistics.

Percent Reduction in Cr

Filter A	Filter B
$n_x = 15$	$n_y = 15$
$\bar{X} = 50.0$	$\bar{Y} = 39.0$
$S_x = 25.0$	$S_y = 24.0$

We want to know if there's any difference in the efficiencies of the two mixtures. Thus the hypotheses are

$$\begin{aligned} H_0 : \mu_x - \mu_y &= 0 \\ H_a : \mu_x - \mu_y &\neq 0 \end{aligned}$$

where μ_x and μ_y are the true mean percent reductions using Filters A and B, respectively.

Suppose that histograms and normal probability plots indicate that the normality assumption is met for both samples. Also, because the larger of the two sample standard deviations is less than twice as large as the smaller, it's reasonable to assume the population standard deviations are equal. Thus a pooled two-sample t test is appropriate.

The pooled estimate of the common population standard deviation is

$$\begin{aligned} S_{\mathrm{p}} &= \sqrt{\frac{(15-1)25.0^2 + (15-1)24.0^2}{15+15-2}} \\ &= 24.51, \end{aligned}$$

so the estimated standard error of $\bar{X} - \bar{Y}$ is

$$S_{\mathrm{p},\,\bar{X}-\bar{Y}} = \sqrt{\frac{24.51^2}{15} + \frac{24.51^2}{15}} = 8.95$$

and the observed test statistic value is

$$t \;=\; \frac{50.0 - 39.0}{8.95} \;=\; 1.23.$$

From a t distribution table, using $n_x + n_y - 2 = 28$ degrees of freedom, the p-value is $2(0.1145) = 0.2290$. Using a level of significance $\alpha = 0.05$, we fail to reject H_0, so we conclude that there's no statistically significant evidence for any difference between the efficiencies of the two filter mixtures.

8.5.3 The Pooled Two-Sample t Confidence Interval

If the assumptions for the pooled two-sample t test are met, then we can estimate an effect size by the following ***pooled two-sample t confidence interval for $\mu_x - \mu_y$***.

Pooled Two-Sample t Confidence Interval: Suppose that $X_1, X_2, \ldots, X_{n_x}$ and $Y_1, Y_2, \ldots, Y_{n_y}$ are independent random samples from populations whose means are μ_x and μ_y and whose standard deviations are equal. Suppose also that either the populations are both normal or the sample sizes n_x and n_y are both large.

Then a **$100(1-\alpha)\%$ *pooled two-sample t confidence interval for $\mu_x - \mu_y$*** is

$$\bar{X} - \bar{Y} \;\pm\; t_{\alpha/2, n_x+n_y-2}\, S_{\mathrm{p},\,\bar{X}-\bar{Y}}\,, \tag{8.17}$$

where

$$S_{\mathrm{p},\,\bar{X}-\bar{Y}} \;=\; \sqrt{\frac{S_{\mathrm{p}}^2}{n_x} + \frac{S_{\mathrm{p}}^2}{n_y}}\,.$$

We can be $100(1-\alpha)\%$ confident that the true effect size $\mu_x - \mu_y$ will be contained in this interval. As always, the *margin of error* is the "plus or minus" part.

Margin of Error: For the pooled two-sample t confidence interval (8.17), the margin of error is

$$\text{Margin of Error} \;=\; t_{\alpha/2, n_x+n_y-2}\, S_{\mathrm{p},\,\bar{X}-\bar{Y}} \;=\; t_{\alpha/2, n_x+n_y-2} \sqrt{\frac{S_{\mathrm{p}}^2}{n_x} + \frac{S_{\mathrm{p}}^2}{n_y}}\,.$$

Example 8.10: Pooled Two-Sample t Confidence Interval

Returning to the experiment to compare the efficiencies of two filter mixtures for removing chromium from storm water (Example 8.9), a 95% pooled two sample t confidence interval for the effect size $\mu_x - \mu_y$ is

$$\begin{aligned} \bar{X} - \bar{Y} \pm t_{\alpha/2, n_x+n_y-2}\, S_{\mathrm{p},\,\bar{X}-\bar{Y}} &= 50.0 - 39.0 \pm 2.05\,(8.95) \\ &= 11.0 \pm 18.35 \\ &= (-7.35,\; 29.35). \end{aligned}$$

where $\bar{X}$, $\bar{Y}$, and $S_{\mathrm{p},\,\bar{X}-\bar{Y}}$ are from Example 8.9 and $t_{0.025,28} = 2.05$ is from a table of t distribution critical values, using $n_x + n_y - 2 = 28$ degrees of freedom.

Notice that the interval contains zero, which is consistent with our failure to reject H_0 in Example 8.9.

8.6 Dealing With Non-Normal Data: Transformations and Nonparametric Procedures

The two-sample and pooled two-sample t tests are both are both *parametric* tests because they rely on an assumption of normality of the populations. If this assumption isn't met (and n_x and n_y aren't large), there are two options.

1. **Transform the data to normality**: It's sometimes possible to transform both samples, for example by taking their logs or using another transformation in the Ladder of Powers, so that the transformed values are more normally distributed, and then carry out the test on the transformed data.
2. **Carry out a nonparametric test**: We can carry out a *nonparametric* test that doesn't rely on an assumption of normality. The *rank-sum test* described in the next section is a nonparametric alternative to the two-sample and pooled t tests.

8.7 The Rank Sum Test

8.7.1 Introduction

The ***rank sum test*** (also called the ***Wilcoxon rank sum test***), like the two-sample t test, is a test for the difference between two population means μ_x and μ_y. But unlike the t test, the rank sum test doesn't require the normality assumption, so it's a *nonparametric* alternative to the t test.

The rank sum test is equivalent to another test called the ***Mann-Whitney test***. The two tests use different test statistics but their p-values will be the same. Some statistical software packages will carry out a Mann-Whitney test but not a the rank sum test.

8.7.2 The Rank Sum Test Procedure

We assume only that we have two independent random samples $X_1, X_2, \ldots, X_{n_x}$ and $Y_1, Y_2, \ldots, Y_{n_y}$ from *any* two continuous populations (not necessarily normal) whose distributions differ, if at all, by their means μ_x and μ_y but not their shapes. Figure 8.1 (right plot) shows two distributions that differ by their means but not their shapes.

We'll test the null hypothesis

$$H_0 : \mu_x - \mu_y \ = \ 0$$

that the population means are the same, versus one of the three alternatives

1. $H_a : \mu_x - \mu_y > 0$ (upper-tailed test)
2. $H_a : \mu_x - \mu_y < 0$ (lower-tailed test)
3. $H_a : \mu_x - \mu_y \neq 0$ (two-tailed test)

Comment: Because the X and Y populations are assumed to have the same shape, if they also have the same means ("centers"), they're identical. Thus the null hypothesis could be stated as

$$H_0 : \text{The two populations are identical.}$$

Likewise, if μ_x and μ_y are different, the populations will differ only by their "centers", or locations along the horizontal axis (right plot of Fig. 8.1). Thus, because the population median is another measure of center, we could also state the null hypothesis in terms of the *medians* $\tilde{\mu}_x$ and $\tilde{\mu}_y$ as

$$H_0 : \tilde{\mu}_x - \tilde{\mu}_y \ = \ 0,$$

in which case we'd state H_a in terms of the medians too.

The rank sum test is carried out by *combining* the observations from the two samples together, *sorting* these from smallest to largest, and then ***ranking*** them by recording their *positions* in the sorted list. The test statistic is then the *sum* of the *ranks* of the sample whose size is smaller. If the two sample sizes are the same, the ranks of either sample can be used to compute the test statistic.

Example 8.11: Rank Sum Test Statistic

Waste from mining activities can impact water quality in nearby rivers and streams by altering their levels of suspended solids, metals, and acidity. In a study of the impact of weathering waste rock from an inactive gold mine on water quality in the nearby North Fork Humboldt River in Nevada, several variables were measured upstream and downstream of the former mine [3]. The table below shows the arsenic (As) concentrations (μg/L) made upstream on $n_x = 6$ days and downstream on $n_y = 7$ days.

As in Water

Upstream	Downstream
5	10
4	9
6	8
10	7
12	10
9	16
	10

Side-by-side boxplots of the two samples are below.

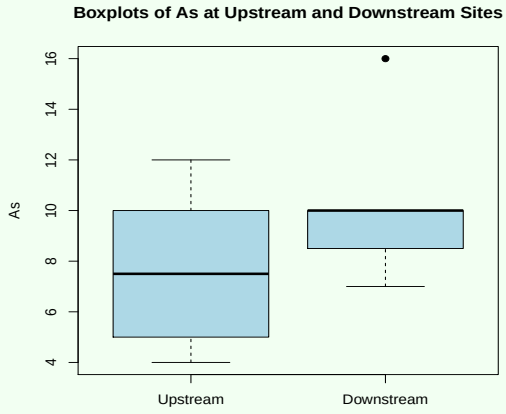


Figure 8.6: Side-by-side boxplots of arsenic concentrations upstream and downstream of the inactive gold mine.

The boxplots show that there's substantial overlap between the two samples, but that that the arsenic levels downstream of the gold mine might be slightly higher. We want to know if the difference is statistically significant. Histograms and normal probability plots of the data are below.

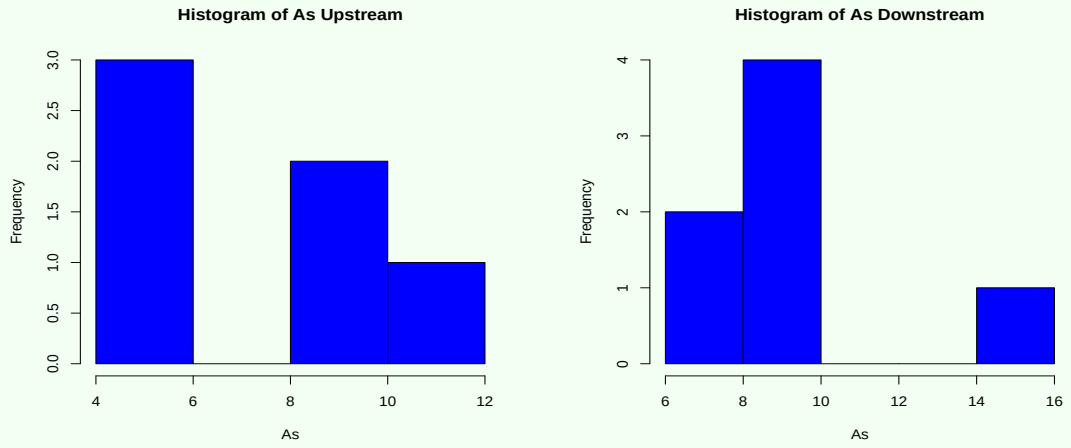


Figure 8.7: Histograms of arsenic upstream (left) and downstream (right) of the inactive gold mine.

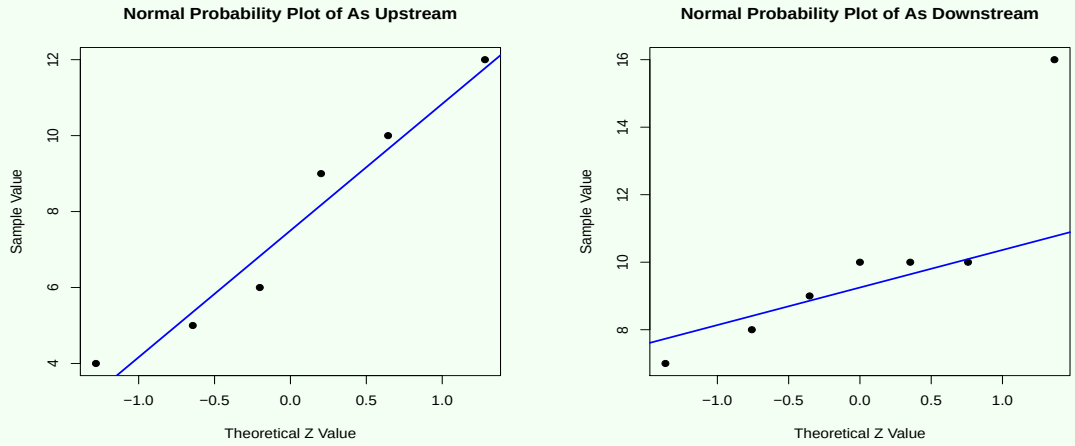


Figure 8.8: Normal probability plots of arsenic upstream (left) and downstream (right) of the inactive gold mine.

There are subtle hints from the plots that arsenic concentrations follow right skewed distributions at both sites, and so because the sample sizes are small, a two-sample t test isn't appropriate. Instead, we'll carry out a rank sum test.

To compute the test statistic, we first combine the two samples, keeping track of which observations were made upstream and which were made downstream. Then we *rank* the observations in the combined sample from smallest (rank = 1) to largest (rank = $n_x + n_y = 13$). If two or more observations are tied, they're each assigned the *average* of the ranks they would've been assigned had they not been tied.

Letting X denote the upstream sample and Y the downstream one, the combined, sorted, and ranked

observations are:

Sample	***X***	***X***	***X***	*Y*	*Y*	***X***	*Y*	***X***	*Y*
Observation	**4**	**5**	**6**	7	8	**9**	9	**10**	10
Rank	**1**	**2**	**3**	4	5	**6.5**	6.5	**9.5**	9.5

	Y	*Y*	***X***	*Y*
(cont'd)	10	10	**12**	16
	9.5	9.5	**12**	13

Notice that the sixth and seventh observations are tied, so they're given the average of the ranks 6 and 7. Likewise, the eighth through eleventh are tied, so they're given the average of the ranks 8, 9, 10, and 11.

The test statistic, denoted $\boldsymbol{W_{rs}}$, is the sum of the ranks for the upstream site (whose sample size is the smaller of the two).

$$\begin{aligned} W_{rs} &= \text{Sum of ranks of } X \text{ sample} \\ &= 1 + 2 + 3 + 6.5 + 9.5 + 12 \\ &= 34. \end{aligned}$$

We'll finish the hypothesis test in Example 8.13.

Here's the general procedure for computing the ***rank sum test statistic***, denoted $\boldsymbol{W_{rs}}$.

Rank Sum Test Statistic:

1. Label the observations in the sample whose size is *smaller* by $X_1, X_2, \ldots, X_{n_x}$ and those in the sample whose size is *larger* by $Y_1, Y_2, \ldots, Y_{n_y}$. If the sample sizes are the same, either sample may be labeled $X_1, X_2, \ldots, X_{n_x}$ and the other $Y_1, Y_2, \ldots, Y_{n_y}$.
2. *Combine* the two samples, *sort* the observations, and *rank* them from smallest to largest. If two or more observations are tied, assign to each of them the *average* of the ranks they would've been assigned if they hadn't been tied.
3. Sum the *ranks* of the X_i's. This gives the test statistic:

$$W_{rs} = \text{Sum of ranks of } X_i\text{'s.}$$

In Example 8.11, the upstream and downstream arsenic concentrations were somewhat "intermingled" after the two samples were combined and sorted, meaning there was a lot of overlap between the samples and suggesting that the gold mine wasn't affecting the stream's arsenic levels. In the next example, the two samples are completely "segregated" after combining and sorting them, suggesting that the mine is adversely impacting the water quality.

Example 8.12: Rank Sum Test

Another variable measured in the study of water quality upstream and downstream of the inactive gold mine described in Example 8.11 was sodium (Na). The table below shows the data (in mg/L).

Na in Water

Upstream	Downstream
2.17	4.23
2.22	4.43
0.85	3.80
2.03	3.80
1.54	3.66
2.05	2.58
	3.48

If we combine the two samples, order them from smallest to largest, and rank them, we end up with the following.

Sample	***X***	***X***	***X***	***X***	***X***	***X***	*Y*	*Y*	*Y*
Observation	**0.85**	**1.54**	**2.03**	**2.05**	**2.17**	**2.22**	2.58	3.48	3.66
Rank	**1**	**2**	**3**	**4**	**5**	**6**	7	8	9

	Y	*Y*	*Y*	*Y*
(cont'd)	3.80	3.80	4.23	4.43
	10.5	10.5	12	13

where an X corresponds to an upstream observation and a Y to a downstream one. Notice that the upstream Na concentrations are *all lower* than the downstream ones. The rank sum test statistic is

$$\begin{aligned} W_{rs} &= \text{Sum of ranks of } X \text{ sample} \\ &= 1+2+3+4+5+6 \\ &= 21, \end{aligned}$$

which is smaller than the value we got in Example 8.11. We'll find out in Example 8.14 if this value provides statistically significant evidence that the gold mine is affecting the Na levels.

In Example 8.12, the upstream sodium concentrations fell entirely at the lower end of the combined, sorted sample, suggesting that the mine is adversely impacting the stream's water quality. It resulted in a small value of W_{rs}. In general, if the X's in the combined, sorted sample, all lie near the lower end, it suggests that μ_x is less than μ_y and results in a small value of W_{rs}. On the other hand, if they lie near the upper end, it suggests that μ_x is greater than μ_y and leads to a large W_{rs} value.

We'll see that when the two samples are evenly "intermingled" after combining and sorting them, W_{rs} will be about equal to $n_x(N+1)/2$, where $\boldsymbol{N}$ denotes the ***overall (combined) sample size*** (that is, $N = n_x + n_y$). Thus

1. *Large* values of W_{rs} (larger than $n_x(N+1)/2$) provide evidence in favor of $H_a : \mu_x - \mu_y > 0$.
2. *Small* values of W_{rs} (smaller than $n_x(N+1)/2$) provide evidence in favor of $H_a : \mu_x - \mu_y < 0$.
3. Both *large and small* values of W_{rs} (larger or smaller than $n_x(N+1)/2$) provide evidence in favor of $H_a : \mu_x - \mu_y \neq 0$.

To decide whether an observed value of W_{rs} provides statistically significant evidence in support of the alternative hypothesis, we'll need to know its sampling distribution under the null hypothesis.

Sampling Distribution of W_{rs} Under H_0: Suppose $X_1, X_2, \ldots, X_{n_x}$ and $Y_1, Y_2, \ldots, Y_{n_y}$ are independent random samples from *any* two continuous populations whose distributions differ, if at all, by their means μ_x and μ_y but not their shapes. Then when

$$H_0 : \mu_x - \mu_y \;=\; 0$$

is true, W_{rs} follows a discrete probability distribution called the ***Wilcoxon rank sum distribution***, which has two parameters, $\boldsymbol{n_x}$ and $\boldsymbol{n_y}$. We write this as

$$W_{rs} \;\sim\; \text{Wilcoxon}(n_x,\, n_y).$$

The *Wilcoxon rank sum distribution* is symmetric and approximately bell-shaped. The distribution is shown below for sample sizes $n_x = 3$ and $n_y = 4$.

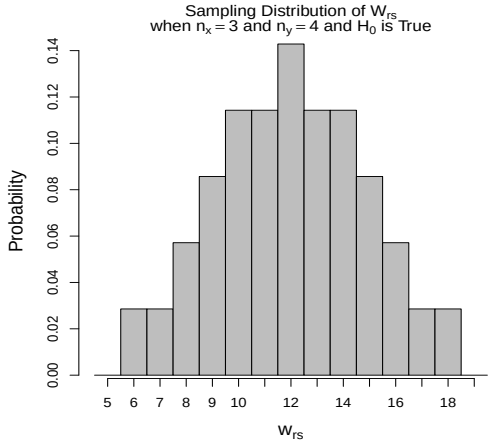


Figure 8.9: Sampling distribution of the rank sum test statistic W_{rs} when $n_x = 3$ and $n_y = 4$ and H_0 is true.

The mean and standard error (standard deviation) of the Wilcoxon distribution, denoted $\boldsymbol{\mu_{w_{rs}}}$ and $\boldsymbol{\sigma_{w_{rs}}}$, are given by the following.

Mean and Standard Error of the Sampling Distribution of W_{rs}: The mean $\boldsymbol{\mu_{w_{rs}}}$ and standard error $\boldsymbol{\sigma_{w_{rs}}}$ of the Wilcoxon rank sum distribution are

$$\mu_{w_{rs}} \;=\; \frac{n_x(N+1)}{2} \tag{8.18}$$

and

$$\sigma_{w_{rs}} \;=\; \sqrt{\frac{n_x n_y (N+1)}{12}}\,. \tag{8.19}$$

Where $N = n_x + n_y$ is the overall sample size (when the two samples are combined).

The mean $\mu_{w_{rs}}$ is the value we'd expect to get for W_{rs}, on average, when the two samples are drawn from populations whose means are equal (that is, when H_0 is true). Thus if the null hypothesis was true, we'd expect our test statistic to be roughly equal to $n_x(N+1)/2$. Values of W_{rs} in the extreme tail of the distribution (in the direction specified by H_a) provide evidence against H_0, so the rejection region is

comprised of W_{rs} values in the the extreme $100\alpha\%$ of the distribution, and the p-value is the tail probability outward from the observed W_{rs} value.

Details about how the sampling distribution of W_{rs} was determined will be given in Subsection 8.7.5. For now, p-values (tail probabilities) and critical values (for the rejection region approach) will be found in a Wilcoxon rank sum distribution table.

The rank sum test procedure is summarized in the table below.

Rank Sum Test for μ_x and μ_y

Assumptions: The data $x_1, x_2, \ldots, x_{n_x}$ and $y_1, y_2, \ldots, y_{n_y}$, with $n_x \leq n_y$, are independent random samples from two continuous populations whose distributions differ, if at all, by their means μ_x and μ_y but not their shapes.

Null hypothesis: $H_0 : \mu_x - \mu_y = 0$.

Test statistic value: w_{rs} = Sum of the ranks of x's in the combined, sorted set of x's and y's.

Decision rule: Reject H_0 if p-value $< \alpha$ or w_{rs} is in rejection region.

Alternative hypothesis	P-value = tail probability of the W_{rs} distribution under H_0:*	Rejection region = w_{rs} values such that:**
$H_a : \mu_x - \mu_y > 0$	to the right of (and including) w_{rs}	$w_{rs} \geq w_{\alpha,n_x,n_y}$
$H_a : \mu_x - \mu_y < 0$	to the left of (and including) w_{rs}	$w_{rs} \leq w'_{\alpha,n_x,n_y}$
$H_a : \mu_x - \mu_y \neq 0$	$2 \cdot$ (the smaller of the tail probabilities to the right of (and including) w_{rs} and to the left of (and including) w_{rs})	$w_{rs} \leq w'_{\alpha/2,n_x,n_y}$ or $w_{rs} \geq w_{\alpha/2,n_x,n_y}$

* Tail probabilities of the W_{rs} distribution under H_0, for given sample sizes n_x and n_y, are found in the Wilcoxon rank sum distribution table under the column p for different observed values of W_{rs} (denoted w in the right tail and w' in the left tail).

** For a given level of significance α and sample sizes n_x and n_y, the upper tail critical value w_{α,n_x,n_y} is obtained from the Wilcoxon rank sum distribution table by locating the smallest w for which the tail area p to its right is less than α. Likewise, w'_{α,n_x,n_y} is obtained by locating the largest w' for which the tail area p to its left less than α.

Note: If your observed W_{rs} value doesn't appear in the Wilcoxon distribution table (even though your sample sizes are represented in the table), there are three possibilities:

1. Largest $w' < W_{rs} <$ Smallest w. In this case W_{rs} is near enough to the middle of the distribution that you should fail to reject H_0.
2. $W_{rs} <$ Smallest w'. In this case you should reject H_0 if you're doing a lower-tailed test or a two-tailed test, and fail to reject H_0 if you're doing an upper tailed test.
3. $W_{rs} >$ Largest w. In this case you should reject H_0 if you're doing an upper-tailed test or a two-tailed test, and fail to reject H_0 if you're doing a lower tailed test.

8.7.3 Carrying Out the Rank Sum Test

We've seen how to compute the rank sum test statistic. Now we'll now turn to a few examples showing how to carry out the rest of test.

Example 8.13: Rank Sum Test

In Example 8.11, we had $n_x = 6$ arsenic measurements made upstream and $n_y = 7$ downstream of the former gold mine. We're interested in deciding if the true mean As concentration is higher downstream than upstream. Thus the hypotheses are

$$\begin{aligned} H_0 : \mu_x - \mu_y &= 0 \\ H_a : \mu_x - \mu_y &< 0 \end{aligned}$$

where μ_x and μ_y are the true mean upstream and downstream concentrations, respectively. Notice that H_0 says the mine has *no effect* on the river's As concentration, and H_a says it *does* have an effect.

From Example 8.11, the observed test statistic value is $W_{rs} = 34$. The p-value is the probability that we'd get a test statistic this small just by chance if the gold mine had no effect.

From the Wilcoxon distribution table, after locating our sample sizes $n_x = 6$ and $n_y = 7$, we find that $W_{rs} = 34$ isn't in the table. It lies between the largest w' and smallest w values. This tells us that the p-value is greater than 0.117, the p-value corresponding to the largest w' value in the table.

Using a level of significance $\alpha = 0.05$, therefore, we fail to reject H_0. There's no statistically significant evidence that the mine is affecting the downstream As concentrations.

Example 8.14: Rank Sum Test

In Example 8.12, we're interested in deciding if the mean sodium concentration is higher downstream than upstream. The hypotheses are

$$\begin{aligned} H_0 : \mu_x - \mu_y &= 0 \\ H_a : \mu_x - \mu_y &< 0 \end{aligned}$$

where μ_x and μ_y are the true mean upstream and downstream Na concentrations, respectively.

From Example 8.12, the test statistic is $W_{rs} = 21$.

From the Wilcoxon distribution table, using $n_x = 6$ and $n_y = 7$, we find that $W_{rs} = 21$ isn't in the table. It lies below the smallest w' value. This tells us that the p-value is less than 0.007, the p-value corresponding to the smallest w' value in the table.

Thus using $\alpha = 0.05$, we reject H_0. There's statistically significant evidence that Na concentrations are higher downstream of the mine than upstream.

8.7.4 What if the Population Distributions Don't Have the Same Shape?

The rank sum test, as described above, relies on an assumption that the two population distributions have the same shape. It turns out, though, that the test *can* still be carried out *without* assuming that they have the same shape. However, in this case it's not testing for a difference between the population means μ_x and μ_y. Rather, it's testing hypotheses that can be stated in words as:

H_0 : The two population distributions are identical
H_a : One distribution has values that are systematically larger

Here's more formally what's meant by "systematically larger" in the context of arsenic observations made upstream and downstream of the former gold mine. Each time we measure the As upstream, it's a random variable X, and each time we measure it downstream it's a random variable Y. There's a probability that X will be greater than 8 μg/L, $P(X > 8)$, and a probability that Y will be greater than 8 μg/L, $P(Y > 8)$. If the downstream As distribution has values that are "systematically larger" than those of the upstream As distribution, then we'd have

$$P(Y > 8) \ > \ P(X > 8).$$

The alternative hypothesis says that this inequality holds not just for 8 but for *any* As concentration value we specify. In other words, it says that the downstream As distribution has more probability to the right of whatever As concentration value we specify.

8.7.5 Some Comments on the Sampling Distribution of W_{rs}

To fully understand the rank sum test, it's useful to know how the sampling distribution of the test statistic under the null hypothesis is determined. We'll consider the case when the sample sizes are $n_x = 2$ and $n_y = 5$ and the X and Y populations have the same shape.

When the null hypothesis is true, the population means *and* shapes are the same, so the populations are *identical*. In this case, the two samples (combined) are just seven randomly selected observations all from the same population, so the ranks of the two X's are equally likely to take any two values from one to seven.

In the table below, each of the possible sets of two values for the X ranks is shown (assuming no ties), along with the resulting W_{rs} value. Note that in some cases, more than one set of X ranks lead to the same W_{rs} value. Also shown is the frequency of each W_{rs} value.

Ranks of X's	Value of W_{rs}	Frequency of W_{rs}
1, 2	3	1
1, 3	4	1
1, 4; 2, 3	5	2
1, 5; 2, 4	6	2
1, 6; 2, 5; 3, 4	7	3
1, 7; 2, 6; 3, 5	8	3
2, 7; 3, 6; 4, 5	9	3
3, 7; 4, 6	10	2
4, 7; 5, 6	11	2
5, 7	12	1
6, 7	13	1
		21

There are a total of 21 possible sets X ranks, all of which are equally likely when H_0 true, but W_{rs} can only take the values $3, 4, \ldots, 13$. Thus the sampling distribution of W_{rs} is:

Value of w_{rs}	3	4	5	6	7	8	9	10	11	12	13
$p(w_{rs})$	$\frac{1}{21}$	$\frac{1}{21}$	$\frac{2}{21}$	$\frac{2}{21}$	$\frac{3}{21}$	$\frac{3}{21}$	$\frac{3}{21}$	$\frac{2}{21}$	$\frac{2}{21}$	$\frac{1}{21}$	$\frac{1}{21}$

which is shown as a probability histogram below.

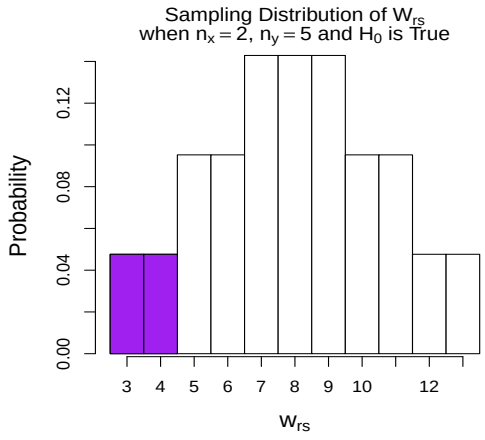


Figure 8.10: Sampling distribution of the rank sum test statistic W_{rs} when $n_x = 2$ and $n_y = 5$ and H_0 is true.

If the observed test statistic value was $W_{rs} = 4$, the p-value for a lower tailed test would be the shaded probability in Fig. 8.10, which, from the probability distribution table above, is $1/21 + 1/21 = 0.0952$.

Properties of the Sampling Distribution of W_{rs}: The following properties of W_{rs} may provide insight into interpreting its value. Recall that N denotes the total (combined) sample size, that is, $N = n_x + n_y$.

1. W_{rs} takes its smallest possible value when the X's are all smaller than the smallest Y. In this case the X's are ranked $1, 2, \ldots, n_x$, and W_{rs} is equal to $1 + 2 + \cdots + n_x = n_x(n_x + 1)/2$.
2. W_{rs} takes its largest possible value when the X's are all larger than the largest Y. In this case the X's are ranked $(n_y+1), (n_y+2), \ldots, N$, and W_{rs} is equal to $(n_y+1)+(n_y+2)+(n_y+3)+\cdots+(n_y+n_x) = n_x n_y + n_x(n_x + 1)/2$.
3. Recall that when H_0 is true, the mean of the sampling distribution of W_{rs} is

$$\mu_{w_{rs}} = \frac{n_x(N+1)}{2}.$$

   This makes sense if we note that when H_0 is true, we expect the X's and Y's to be evenly "intermingled" in the combined, sorted sample. In this case, the *average* of the X ranks should be about equal to the average of *all* the ranks $(1, 2, \ldots, N)$, which is $(N+1)/2$, and so the *sum* of the X ranks should be about equal to n_x times its average, or $n_x(N + 1)/2$.

Comment: Properties 1 - 3 above make use of the fact that for any positive integer n,

$$1 + 2 + 3 + \cdots + n = \frac{n(n+1)}{2}. \tag{8.20}$$

To see why, consider the case in which $n = 100$. Rearranging the numbers $1, 2, \ldots, 100$ so that 1 is paired with 100, 2 is paired with 99, and so on gives

$$\begin{aligned} 1 + 2 + \cdots + 100 &= (1 + 100) + (2 + 99) + \cdots + (50 + 51) \\ &= 101 + 101 + \cdots + 101 \\ &= 50(101) \\ &= \frac{100(100 + 1)}{2}, \end{aligned}$$

which is (8.20) with $n = 100$.

8.7.6 Large Sample Version of the Rank Sum Test

Notice from Figs. 8.9 and 8.10 that the sampling distribution of W_{rs} when H_0 is true is symmetric and roughly bell-shaped. The plots below show that the distribution becomes more and more bell-shaped as the sample sizes n_x and n_y get bigger.

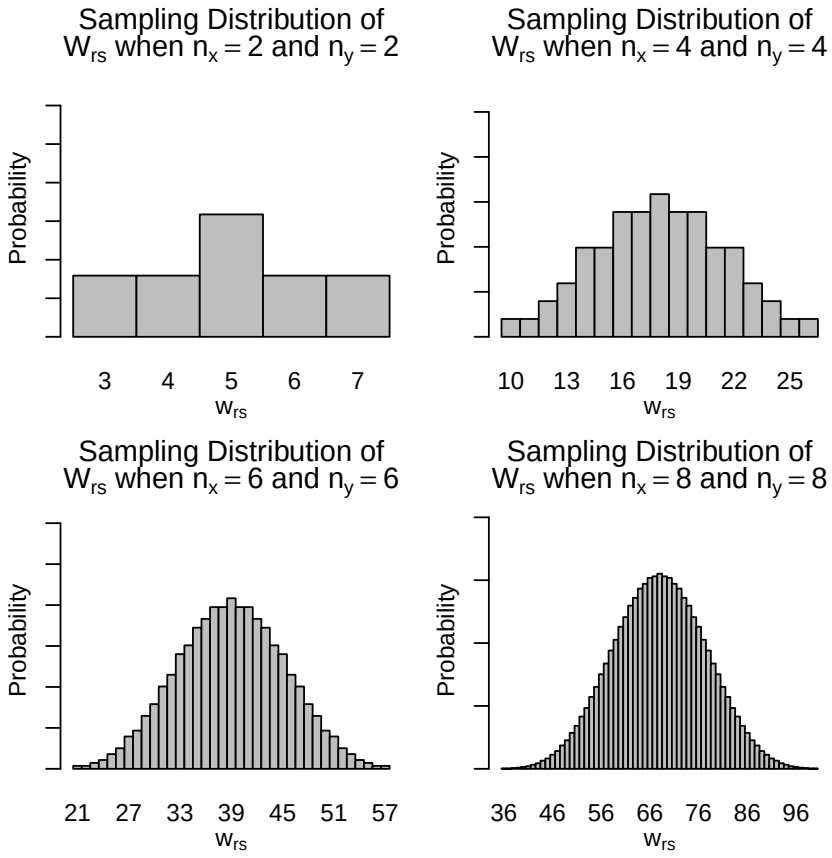


Figure 8.11: Sampling distribution of the rank sum test statistic W_{rs} for various sample sizes n_x and n_y when H_0 is true.

It turns out that as n_x and n_y increase, the distribution gets closer and closer to a normal distribution, as stated in the following fact.

Fact 8.5 Suppose $X_1, X_2, \ldots, X_{n_x}$ and $Y_1, Y_2, \ldots, Y_{n_y}$ are two independent random samples from continuous populations whose means are μ_x and μ_y. Then if n_x and n_y are large ($n_x \geq 10$ and $n_y \geq 10$ is sufficient) and $H_0 : \mu_x - \mu_y = 0$ is true,

$$W_{rs} \sim \mathrm{N}(\mu_{w_{rs}}, \sigma_{w_{rs}})$$

approximately, where the mean $\mu_{w_{rs}}$ and standard error $\sigma_{w_{rs}}$ are given by (8.18) and (8.19).

It follows that if we standardize W_{rs}, the resulting random variable Z follows a standard normal

random distribution, that is,

$$Z = \frac{W_{rs} - \mu_{w_{rs}}}{\sigma_{w_{rs}}} \sim \text{N}(0, 1)$$

approximately.

When n_x and n_y are large, the appropriate test statistic for carrying out the rank sum is the ***large sample rank sum test statistic***, denoted $\boldsymbol{Z_{rs}}$, given by the following.

Large Sample Rank Sum Test Statistic:

$$Z_{rs} = \frac{W_{rs} - \mu_{w_{rs}}}{\sigma_{w_{rs}}}$$

where $\mu_{w_{rs}}$ and $\sigma_{w_{rs}}$ are given by (8.18) and (8.19).

P-values (and critical values for the rejection region approach) are obtained from the tails of the N(0, 1) distribution in the direction specified by H_a.

Comment: As was the case with the large sample version of the sign test of Chapter 7, most statistical software packages use a slightly more accurate ***continuity corrected*** version of Z_{rs} when computing p-values for the large sample version of the rank sum test. The continuity correction accounts for the fact that a continuous distribution (the standard normal) is approximating a discrete one (the true distribution of Z_{rs}). Details about the continuity correction can be found in many statistics textbooks, including [9].

8.8 Which Test Should Be Used, the Two-Sample *t* Test, the Pooled *t* Test, or the Rank Sum Test?

8.8.1 Deciding Between the Pooled *t* Test and the Two-Sample *t* Test

Both the two-sample t test and the pooled t test require an assumption that the samples came from normally distributed populations, but the pooled t test requires an additional assumption that the population standard deviations are equal. In practice, we almost never know for sure whether this additional assumption is valid, and the rule of thumb given in Subsection 8.5.2 that the larger sample standard deviation be no more than twice as large as the smaller one makes no guarantees.

If we mistakenly use the pooled t procedure even though σ_X and σ_Y aren't equal, the *power* for detecting a difference between the population means will be diminished compared to that of a two-sample t test. It's generally recommended, therefore, that we play it safe and *use the two-sample t procedure if there's any doubt as to whether σ_X and σ_Y are the same.*

Comment: Even when the population standard deviations *are* equal, there's usually no harm done by using the two-sample t procedure instead of the pooled t procedure. Although the pooled t procedure is *slightly* more powerful under these circumstances, the gain in power is small relative to the loss of power that would result if it turned out that the population standard deviations weren't in fact equal.

8.8.2 Deciding Between the Two-Sample *t* Test and the Rank Sum Test

If the normality assumption required for the two-sample (or pooled) t test isn't met (and n_x and n_y are small), and we don't want to transform the data to normality, we have little choice but to use the rank sum test.

But if the normality assumption *is* met, we have a choice of whether to use the t test or the rank sum test. It can be shown that the t test is *more powerful* than the rank sum test when the normality

assumption is met. In other words, when you carry out a study, you're more likely to detect a difference or effect, if there is one, using the t test. Therefore *when you have a choice, the t test should be used.*

The intuitive explanation for why the t test is more powerful is that it uses the complete information that's contained in the data (that is, it uses their actual numerical values), whereas the rank sum test only uses information about where they're ranked in the combined sample. Thus the rank sum test ignores some of the information that's contained in the data, resulting in a loss of power.

8.9 Problems

8.1 To assess the impact of a wastewater treatment plant's effluent discharge into the Febros River in northwest Portugal, gudgeon (*Gobio gobio*) and mullet (*Mugil cephalus*) fish were sampled upstream and downstream of the plant and their weights (g) and lengths (cm) measured [11].

The water was warmer downstream than upstream, possibly due to the effluent discharge, and fish are more active and have higher metabolic activity in warmer water, which can lead to growth. Also, the nutrient load was higher downstream, and this can lead to more available food for the fish. For these reasons, the researchers hypothesized that the fish would be larger downstream than upstream. The summary statistics are below.

	Gudgeon		**Mullet**	
	Upstream	Downstream	Upstream	Downstream
Weight	$n_x = 23$ $\bar{X} = 12.12$ $S_x = 2.57$	$n_y = 22$ $\bar{Y} = 16.93$ $S_y = 4.17$	$n_x = 7$ $\bar{X} = 21.39$ $S_x = 2.00$	$n_y = 7$ $\bar{Y} = 42.93$ $S_y = 27.88$
Length	$n_x = 23$ $\bar{X} = 10.83$ $S_x = 0.69$	$n_y = 22$ $\bar{Y} = 11.80$ $S_y = 0.87$	$n_x = 7$ $\bar{X} = 14.11$ $S_x = 0.35$	$n_y = 7$ $\bar{Y} = 16.00$ $S_y = 3.04$

a) Carry out a two-sample t test to decide if the true mean gudgeon weight is greater downstream than upstream. Use a level of significance $\alpha = 0.05$.

b) Carry out a two-sample t test to decide if the true mean gudgeon length is greater downstream than upstream. Use a level of significance $\alpha = 0.05$.

8.2 Refer to Problem 8.1, in which weights and lengths of gudgeon and mullet fish were measured upstream and downstream of a wastewater treatment plant.

a) Carry out a two-sample t test to decide if the true mean mullet weight is greater downstream than upstream. Use a level of significance $\alpha = 0.05$.

b) Carry out a two-sample t test to decide if the true mean mullet length is greater downstream than upstream. Use a level of significance $\alpha = 0.05$.

8.3 Refer to Problem 8.1, in which weights and lengths of gudgeon and mullet fish were measured upstream and downstream of a wastewater treatment plant.

a) Compute and interpret a 95% two-sample t confidence interval for the true difference in the mean upstream and downstream gudgeon weights.

b) Compute and interpret a 95% two-sample t confidence interval for the true difference in the mean upstream and downstream gudgeon lengths.

8.4 Refer to Problem 8.1, in which weights and lengths of gudgeon and mullet fish were measured upstream and downstream of a wastewater treatment plant.

a) Compute and interpret a 95% two-sample t confidence interval for the true difference in the mean upstream and downstream mullet weights.

b) Compute and interpret a 95% two-sample t confidence interval for the true difference in the mean upstream and downstream mullet lengths.

8.5 Problems 7.9 and 7.11 in Chapter 7 described a U.S. Geological Survey study of the use of biosolid fertilizer in the Metrogro Farm area east of Denver, Colorado. One question of interest was whether the use of this fertilizer would have any effect on crop quality. To answer this question, several crop quality variables were measured in wheat grown on six fields, two treated with biosolids and four untreated. With the exception of fertilizer use, the six 20-acre fields were farmed in the same manner. The table on the left below shows the cadmium (Cd) concentrations (mg/kg) in the wheat. The table on the right shows their summary statistics.

Cd in Wheat

Control Fields	Fertilized Fields
0.05	0.13
0.14	0.03
0.03	
0.02	

Summary Statistics

Control Fields	Fertilized Fields
$n_x = 4$	$n_y = 2$
$\bar{X} = 0.06$	$\bar{Y} = 0.08$
$S_x = 0.055$	$S_y = 0.071$

a) Carry out a two-sample t test to decide if the mean Cd concentration is statistically significantly higher on fertilized fields than on control fields. Use a level of significance $\alpha = 0.05$.

b) The sample sizes are both small. What assumptions about the data should be met in order for the t test results to be valid?

8.6 A before-after study was carried out to assess the effect of a dam on various water quality characteristics in the Kelkit Stream, Turkey [10]. The variables were measured at a water quality monitoring station located downstream of the dam 60 times during the four years before the dam was built and again 144 times during the 11 years after the dam began operating. One of the variables measured in the water was chlorides (Cl, in mg/L). Here are the summary statistics for this variable.

Cl in Kelkit Stream

Pre-Dam	Post-Dam
$n_x = 60$	$n_y = 144$
$\bar{X} = 12.7$	$\bar{Y} = 11.4$
$S_x = 3.47$	$S_y = 3.22$

a) Carry out a two-sample t test to decide if there's convincing evidence for a decrease in the mean Cl concentration? Use a level of significance $\alpha = 0.05$.

b) Give the value of a point estimate of the unknown true change (effect size) $\mu_x - \mu_y$ in the mean Cl concentration.

c) Compute and interpret a 95% two-sample t confidence interval for $\mu_x - \mu_y$.

8.7 Because mercury (Hg) can accumulate in the feathers of sea birds, concentrations in their feathers are sometimes used as indicators of Hg pollution. A study was carried out to investigate the change over time in Hg pollution on the German North Sea coast by analyzing Hg concentrations in feathers of herring

gulls and common terns from German museum collections dating back to the 1880's [13].

During World War II (1939 - 1945), Hg was used in the manufacture of ammunition, mines, and bombs, and during that time the manufacturers were unlikely to have been concerned with pollution control. To decide if this resulted in an increase in Hg pollution, concentrations of Hg in feathers collected before 1940 were compared to those in feathers collected after 1940. The summary statistics are below.

Mercury in Feathers

	Pre-1940	Post-1940
Herring Gulls	$n_x = 27$	$n_y = 113$
	$\bar{X} = 4.56$	$\bar{Y} = 7.91$
	$S_x = 1.97$	$S_y = 3.86$
Common Tern	$n_x = 5$	$n_y = 37$
	$\bar{X} = 0.98$	$\bar{Y} = 3.47$
	$S_x = 0.85$	$S_y = 2.47$

The authors of the cited study found that the normality assumptions required for two-sample t tests and confidence intervals were met.

a) Carry out a two-sample t test to decide if the true mean Hg concentration in herring gull feathers increased from before 1940 to after 1940. Use a level of significance $\alpha = 0.05$.

b) Compute and interpret a 95% two-sample t confidence interval to estimate the increase in the true mean Hg concentration in herring gull feathers from before 1940 to after 1940.

c) Carry out a two-sample t test to decide if the true mean Hg concentration in common tern feathers increased from before 1940 to after 1940. Use a level of significance $\alpha = 0.05$.

d) Compute and interpret a 95% two-sample t confidence interval to estimate the increase in the true mean Hg concentration in common tern from before 1940 to after 1940.

8.8 Problem 6.3 in Chapter 6 described a study in which PCBs (polychlorinated biphenyls) were measured in the blubber of 8 female dolphins that were nulliparous (had never given birth) and another 17 that were parous (had given birth) from Sarasota Bay, Florida. The summary statistics (ppm) are below.

PCBs in Dolphins

Nulliparous	Parous
$n_x = 8$	$n_y = 17$
$\bar{X} = 27.7$	$\bar{Y} = 6.8$
$S_x = 10.67$	$S_y = 5.45$

Previous studies have found that for male dolphins, accumulation of PCBs continues throughout the lifetime, but for females, concentrations tend to decline with reproductive activity through transfer across the placenta and via lactation.

We want to know if the true mean PCB concentration is higher in nulliparous dolphins than in parous ones, and if so, by how much.

a) Carry out a two-sample t test to decide if the observed mean PCB concentration is statistically significantly higher in nulliparous dolphins than parous ones. Use a level of significance $\alpha = 0.05$.

b) Give the value of the point estimate of the amount by which the true mean PCB concentration for nulliparous dolphins exceeds that for parous ones.

c) Compute and interpret a 95% two-sample t confidence interval for the amount by which the true mean PCB concentration for nulliparous dolphins exceeds that for parous ones.

8.9 Perfluorooctane sulfonate (PFOS) is a fluorochemical compound used in a variety of applications such as lubricants, fire retardants, and pesticides.

The table below shows PFOS concentrations (ng/L) measured at $n_x = 18$ sites along the Tennessee River near a fluorochemical manufacturing facility in Decatur, Alabama and $n_y = 21$ sites downstream [7].

PFOS in Water

Upstream	Downstream
27.8	54.1
28.9	37.3
28.8	30.3
25.8	74.8
36.9	96.4
16.8	98.0
27.4	107.0
31.0	136.0
26.9	140.0
22.3	106.0
21.8	134.0
21.4	106.0
18.4	112.0
31.6	144.0
51.9	92.3
52.6	110.0
37.1	105.0
39.4	119.0
	133.0
	127.0
	119.0

a) Make side-by-side boxplots of the upstream and downstream PFOS concentrations.

b) Make normal probability plots of the upstream and downstream PFOS concentrations.

c) Based on the normal probability plots of part b, does it appear that the normality assumptions required for a two-sample t test are met?

d) Carry out a two-sample t test to decide if there's statistically significant evidence that the true mean PFOS concentration is higher downstream of the fluorochemical facility than upstream. Use level of significance $\alpha = 0.05$.

e) Compute and interpret a 95% two-sample t confidence interval for the difference in true mean PFOS concentrations upstream and downstream of the facility.

8.10 Problem 3.20 in Chapter 3 described a study of persistent contaminants such as DDTs, PCBs, HCHs in eggs from nine randomly selected Little Egret nests in Mai Po Village and nine randomly selected Black-Crowned Night Heron nests in the A Chau egretry, both near Hong Kong. Concentrations of PCB, DDT, and HCH contaminants (ng/g) were measured in each of the sampled eggs. The table below shows the data.

PCBs		DDTs		HCHs	
Egret	Heron	Egret	Heron	Egret	Heron
1700	530	1600	1200	21	28
1000	140	810	1100	14	8
800	110	980	820	19	17
970	600	2200	440	20	30
1600	160	1600	340	17	20
1000	85	1400	350	18	21
270	170	560	680	14	14
370	150	670	280	18	29
970	120	690	210	12	21

One objective of the study was to determine whether there are any statistically significant differences in the contaminant levels in the two types of birds eggs.

a) Make side-by-side boxplots of the PCB concentrations for the two types of birds.

b) Carry out a two-sample t test to decide if there is a statistically significant difference in PCB concentrations. Use a level of significance $\alpha = 0.05$.

c) Make side-by-side boxplots of the DDT concentrations for the two types of birds.

d) Carry out a two-sample t test to decide if there is a statistically significant difference in DDT concentrations. Use a level of significance $\alpha = 0.05$.

e) Make side-by-side boxplots of the HCH concentrations for the two types of birds.

f) Carry out a two-sample t test to decide if there is a statistically significant difference in HCH concentrations. Use a level of significance $\alpha = 0.05$.

8.11 A BACI study design was used to assess the impact of logging on the physical properties of nearby lakes [1]. Several variables were measured on each of nine lakes before and after their watersheds were logged, and on each of 13 control lakes whose watersheds remained unlogged over the same time period. The control lakes had areas, depths, and catchment characteristics that were similar to those of the lakes whose watersheds had been logged.

The table below shows the before-logging and after-logging dissolved organic carbon (DOC) concentrations (μg/L) in each of the lakes along with the amounts by which their DOC concentrations changed.

DOC for Lakes in Unlogged Watersheds			
Lake	Before	After	Change
AB220	4.1	4.4	0.3
AB35	5.4	5.6	0.2
CSL2	5.7	6.3	0.6
CSL5	10.5	10.6	0.1
DA4	8.5	9.0	0.5
DA9	9.7	10.4	0.7
DF4	9.2	9.4	0.2
K2	8.4	9.2	0.8
N35	4.4	4.1	-0.3
N43	9.5	9.4	-0.1
N55	5.0	5.2	0.2
N70	5.8	5.5	-0.3
N89	3.7	4.1	0.4

DOC for Lakes in Logged Watersheds			
Lake	Before	After	Change
DF2	7.2	8.3	1.1
DF5	8.0	8.1	0.1
DF7	11.0	12.5	1.5
DF9	12.7	14.9	2.2
K1	6.5	6.7	0.2
K3	6.1	6.6	0.5
K4	7.3	8.2	0.9
K8	9.9	11.1	1.2
P109	6.1	7.0	0.9

a) Make side-by-side boxplots comparing the changes in DOC for lakes in logged and unlogged watersheds.

b) Make normal probability plots of the changes in DOC for lakes in logged and unlogged watersheds.

c) Based on the plots in part *b*, does the assumption of normality required for the two-sample t test appear to be met?

d) Carry out a two-sample t test to decide if there's statistically significant evidence that the true mean change in DOC in logged-watershed lakes is greater than the mean change in unlogged-watershed lakes. Use a level of significance $\alpha = 0.05$.

8.12 An experiment was carried out to investigate the effects of global warming on soil moisture and temperature in Gunnison County, Colorado [8]. Ten 30 m^2 plots in a subalpine meadow in the Rocky Mountains were used as experimental units. Overhead infrared radiators were placed over five of the plots to create a downward heat flux at a constant rate of 15 W/m^2 for a period of two years. The remaining five plots received no artificial warming and served as controls. The flux of 15 W/m^2 represents a typical estimate of the additional flux resulting from a doubling of atmospheric carbon dioxide.

The table below gives the soil temperatures for the years 1991 and 1992 and their changes over that time period. A negative change means the temperature decreased.

Control Plot
Soil Temperatures

Plot	1991	1992	Change
1	13.35	12.25	-1.10
2	12.39	11.65	-0.74
3	14.06	12.54	-1.52
4	12.82	11.89	-0.93
5	12.44	11.78	-0.66

Warmed Plot
Soil Temperatures

Plot	1991	1992	Change
6	13.46	12.52	-0.94
7	13.36	12.30	-1.06
8	12.62	11.75	-0.87
9	13.18	12.58	-0.60
10	12.42	11.88	-0.54

a) Make side by side boxplots of the changes in soil temperature for the control and warmed plots.

b) Make normal probability plots of the changes for the control and warmed plots.

c) Based on the normal probability plots of part *b*, does the assumption of normality required for the two-sample t test appear to be met?

d) Carry out a two-sample t test to decide if the mean change in temperature was statistically significantly greater on the warmed plots than on the control plots. Use a level of significance $\alpha = 0.05$.

8.13 A BACI study design was used to assess the effect on benthic (bottom dwelling) macroinvertibrate populations of remediation measures carried out to improve water and sediment quality in the Orielton Lagoon, Australia [2]. The remediation measures included improving the drainage of the lagoon and diverting suburban wastewater away from it.

Macroinvertebrates were identified in grab samples from the lagoon bottom in 1999 (before the remediation) at six sites, and again at the *same* six sites in 2005 (after the remediation). They were also identified in both 1999 and 2005 in grab samples from six control sites in the nearby Pittwater and Carlton River estuaries, neither of which underwent the remediation measures.

The table below shows the total taxa in the grab samples and the changes from 1999 to 2005. A negative change means the total taxa decreased.

Total Taxa at the Impact Sites

Site	1999	2005	Change
O1	16	17	1
O2	11	8	-3
O3	7	13	6
O4	5	5	0
O5	6	13	7
O6	6	10	4

Total Taxa at the Control Sites

Site	1999	2005	Change
P1	22	25	3
P2	20	20	0
P3	7	21	14
C1	15	26	11
C2	10	28	18
C3	18	19	1

a) Make side-by-side boxplots of the changes in total taxa for the control and impact sites.

b) Carry out a two-sample t test to decide if the mean change in total taxa at the impact sites was statistically significantly less than the mean change at the control sites. Use level of significance $\alpha = 0.05$.

8.14 Refer to the impact assessment study described in Problem 8.13. The table below shows the total number of individual macroinvertibrates in the grab samples and the changes from 1999 to 2005.

Total Macroinvertibrates at the Impact Sites

Site	1999	2005	Change
O1	639	247	-392
O2	337	51	-286
O3	1048	190	-858
O4	127	60	-67
O5	291	118	-173
O6	732	52	-680

Total Macroinvertibrates at the Control Sites

Site	1999	2005	Change
P1	77	315	238
P2	91	128	37
P3	61	68	7
C1	58	113	55
C2	157	212	55
C3	442	104	-338

a) Make side-by-side boxplots of the changes in total macroinvertibrate individuals for the control and impact sites.

b) Carry out a two-sample t test to decide if the mean change in total macroinvertibrate individuals at the impact sites was statistically significantly less than the mean change at the control sites. Use level of significance $\alpha = 0.05$.

8.15 Refer to the impact assessment study described in Problem 8.13. The table below shows the total number of individual amphipods in the grab samples and the changes from 1999 to 2005.

Total Amphipods at the Impact Sites

Site	1999	2005	Change
O1	573	88	-485
O2	313	5	-308
O3	1038	141	-897
O4	120	0	-120
O5	258	34	-224
O6	719	4	-715

Total Amphipods at the Control Sites

Site	1999	2005	Change
P1	15	73	58
P2	0	32	32
P3	2	18	16
C1	22	5	-17
C2	27	64	37
C3	177	16	-161

a) Make side-by-side boxplots of the changes in total amphipod individuals for the control and impact sites.

b) Carry out a two-sample t test to decide if the mean change in total amphipod individuals at the impact sites was statistically significantly less than the mean change at the control sites. Use level of significance $\alpha = 0.05$.

8.16 Refer to the study of the impact of weathering waste rock from an inactive gold mine on water quality in the nearby North Fork Humboldt River in Nevada described in Examples 8.11 - 8.14. The table below shows iron (Fe) concentrations (mg/L) measured upstream of the former gold mine on on six days and downstream on seven days.

Fe in Water

Upstream	Downstream
0.06	0.34
0.65	0.43
0.18	0.34
0.91	0.25
0.58	0.34
0.09	0.46
	0.45

The researchers had reason to believe that the normality assumption required for a two-sample t test wasn't met.

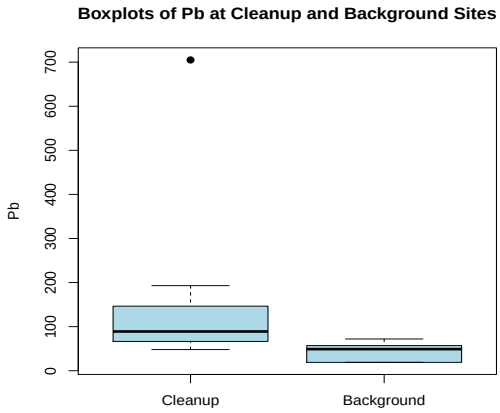
Carry out a rank sum test to decide if there's statistically significant evidence that the true mean Fe concentration is higher downstream than upstream. Use a level of significance $\alpha = 0.05$.

8.17 The data below are lead (Pb) measurements (mg/kg) made in soil at $n_x = 7$ locations within a reclaimed cleanup area of a U.S. Environmental Protection Agency Superfund site and $n_y = 7$ locations at an unpolluted background reference site [6].

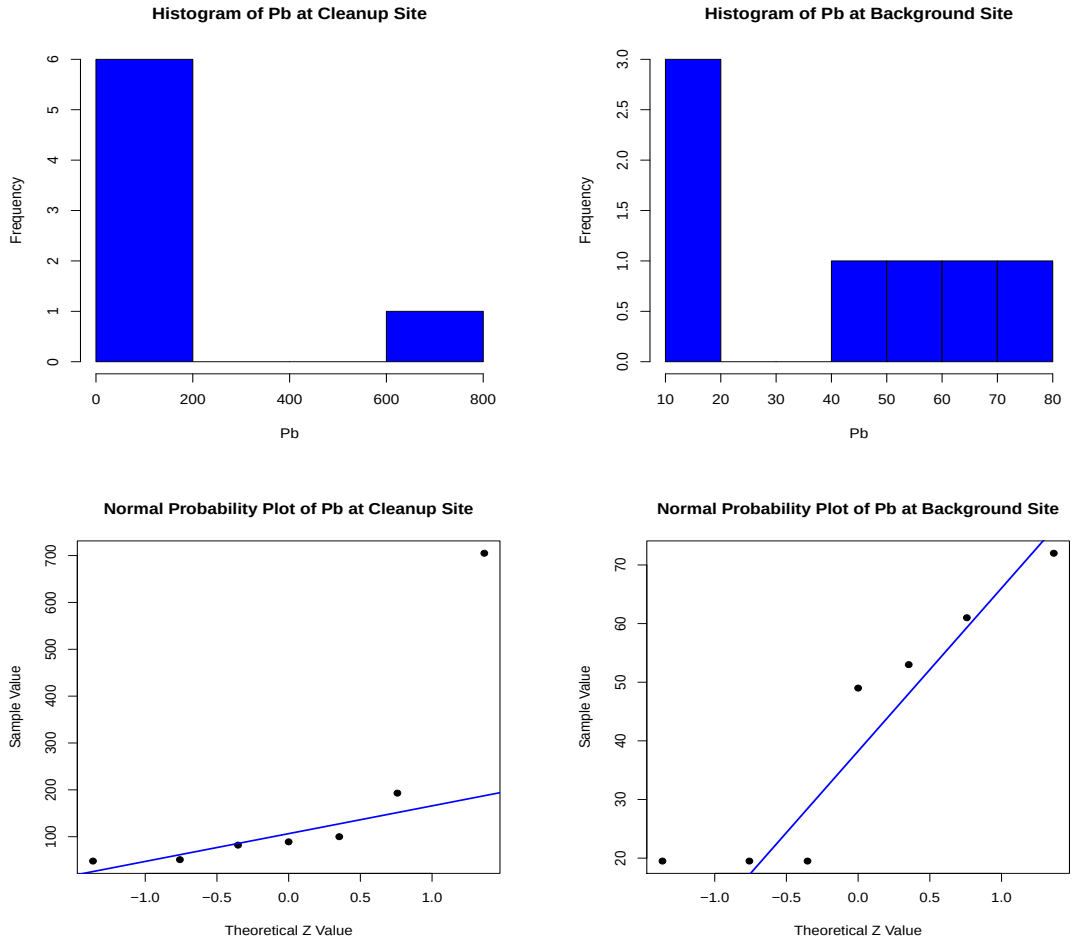
Lead in Soil

Reclaimed Cleanup Site	Background Reference Site
48	49
100	53
193	19
51	72
82	19
705	19
89	61

These data are part of a larger data set that was used to assess whether the reclaimed site met background-based standards. We'll use them to decide if the Pb levels are still higher at the reclaimed site than at the background site. Side-by-side boxplots of the two samples are below.



The boxplots suggest that the Pb levels are still higher at the reclaimed site. We want to know if the difference is statistically significant. Histograms and normal probability plots of the data are below.



The plots indicate that Pb concentrations follow right skewed distribution at both sites, and so because the sample sizes are small, a two-sample t test isn't appropriate. Instead, we'll carry out a rank sum test.

Carry out a rank sum test to decide if the data provide statistically significant evidence that the (unknown) true mean Pb concentration is still higher at the reclaimed site than at the background site. Use level of

significance $\alpha = 0.05$.

8.18 Water pollution by toxic metals from industrial processes is a concern for the health of people and ecosystems in developing countries. A study was carried out to investigate the impact of industrial processes in northern Pakistan on concentrations of metals in drainage water [12].

Metal concentrations (mg/L) were measured in water effluent from three industrial areas whose industries include electronics, manufacturing, food, chemical plants, pharmaceuticals, textiles, and steel. The metals were also measured in water effluent from five control sites along the nearby Warsak Canal, which has no prominent industries around it.

The lead (Pb) and arsenic (As) concentrations are below.

Pb in Water	
Industrial Sites	Control Sites
0.309	0.096
0.350	0.120
0.275	0.210
	0.256
	0.280

As in Water	
Industrial Sites	Control Sites
0.643	0.269
0.475	0.046
0.942	0.369
	0.420
	0.380

a) Carry out a rank sum test to decide if there's statistically significant evidence that the true mean Pb concentration is higher in industrial areas than in control areas. Use a level of significance $\alpha = 0.05$.

b) Carry out a rank sum test to decide if there's statistically significant evidence that the true mean As concentration is higher in industrial areas than in control areas. Use a level of significance $\alpha = 0.05$.

Bibliography

[1] A. Bertolo and P. Magnan. Logging-induced variations in dissolved organic carbon affect yellow perch (*Perca flavescens*) recruitment in Canadian Shield lakes. *Canadian Journal of Fisheries and Aquatic Sciences*, 64:181 – 186, 2007.

[2] P. Davies, L. Cook, and T. Sloane. Orielton Lagoon: Changes in the benthic macroinvertebrate community between 1999 and 2005, report to Sorell Council. March 2006.

[3] S. Earman and R. L. Hershey. Water quality impacts from waste rock at a Carlin-type gold mine, Elko County, Nevada. *Environmental Geology*, 45:1043 – 1053, 2004.

[4] Carina Farm. Metal sorption to natural filter substrates for storm water treatment-column studies. *The Science of the Total Environment*, 298:17 – 24, 2002.

[5] H. Gan, M. Zhuo, D. Li, and Y. Zhou. Quality characterization and impact assessment of highway runoff in urban and rural area of Guangzhou, China. *Environmental Monitoring and Assessment*. DOI 10.1007/s10661-007-9856-2.

[6] R. O. Gilbert and J. C. Simpson. Statistical methods for evaluating the attainment of cleanup standards, volume 3: Reference-based standards for soils and solid media. Technical Report US EPA Report PNL-7409, U.S. Environmental Protection Agency, Dec 1992. National Technical Information Services.

[7] K. J. Hansen et al. Quantitative characterization of trace levels of PFOS and PFOA in the Tennessee River. *Environmental Science and Technology*, 36(8):1681 – 1685, 2002.

[8] J. Harte, M. Torn, F. Chang, B. Feifarek, A. Kinzig, R. Shaw, and K. Shen. Global warming and soil microclimate: Results from a meadow-warming experiment. *Ecological Applications*, 5(1):132 – 150, 1995.

[9] D.R. Helsel. *Nondetects and Data Analysis, Statistics for Censored Environmental Data*. John Wiley and Sons, Inc., 2005.

[10] Ahmet Kurunc, Kadri Yurekli, and Cengiz Okman. Effects of Kilickaya Dam on concentration and load values of water quality constituents in Kelkit Stream in Turkey. *Journal of Hydrology*, 317:17 – 30, 2006.

[11] Ana Lúcia Pinto, Simone Varandas, Ana Maria Coimbra, João Carrola, and António Fontaínhas-Fernandes. Mullet and gudgeon liver histopathology and macroinvertebrate indexes and metrics upstream and downstream from a wastewater treatment plant (Febros River – Portugal). *Environmental Monitoring and Assessment*, 169:569 – 585, 2010.

[12] W. Rehman, A. Zeb, N. Noor, and M. Nawaz. Heavy metal pollution assessment in various industries of Pakistan. *Environmental Geology*, 55(2):353 – 358, July 2008.

[13] D. R. Thompson, P. H. Becker, and R. W. Furness. Long-term changes in mercury concentrations in herring gulls *Larus argentatus* and common terns *Sterna hirundo* from the German North Sea coast. *Journal of Applied Ecology*, 30(2):316 – 320, 1993.