prepared by Professor Jenny Baglivo

5 MATH4427 Notebook 5

This notebook is concerned with parametric and nonparametric methods for two sample analysis. The notes include material from Chapter 11 (comparing two samples) and Chapter 6 (distributions derived from the normal distribution) of the Rice textbook.

5.1 Two Sample Analysis: Difference in Means

In many statistical applications, interest focuses on comparing two probability distributions.

For example,

1. An education researcher might be interested in determining if the distributions of standardized test scores for students in public and private schools are equal.
2. A medical researcher might be interested in determining if the distributions of mean blood pressure levels are the same in patients on two different treatment protocols.
3. An economics researcher might be interested in determining if income distributions are the same in two different communities.

In this section, we focus on comparing two distributions by comparing their means.

5.1.1 Introduction: Notation and Model Summaries

1. *X Sample:* Let
$$X_1, X_2, \ldots, X_n$$
be a random sample from a distribution with mean μ_x and standard deviation σ_x.
2. *Y Sample:* Let
$$Y_1, Y_2, \ldots, Y_m$$
be a random sample from a distribution with mean μ_y and standard deviation σ_y.
3. *Independent Samples:* Assume that the samples were chosen independently.
Thus, the combined sample,
$$X_1,\ X_2,\ \ldots,\ X_n,\ Y_1,\ Y_2,\ \ldots,\ Y_m$$
is a list of $n + m$ mutually independent random variables, where the first n are IID (independent identically distributed) with the same distribution as X, and the last m are IID with the same distribution as Y.

Difference in means. Let $\delta = \mu_x - \mu_y$ be the difference in means of the distributions.

The difference in sample means, $\overline{X} - \overline{Y}$, is used to estimate δ. Since the samples were chosen independently, summary values are easy to compute.

Sample Summaries Theorem. Under the conditions stated above,

$$E(\overline{X} - \overline{Y}) = \delta \quad \text{and} \quad Var(\overline{X} - \overline{Y}) = \frac{\sigma_x^2}{n} + \frac{\sigma_y^2}{m}.$$

Note: This theorem can be proven using the techniques we learned in probability theory. Remember that the theorem tells us nothing about the sampling distribution of $\overline{X} - \overline{Y}$.

5.1.2 Exact Methods for Normal Distributions

If X and Y are normal random variables, then $\overline{X} - \overline{Y}$ has a normal distribution.

Exact statistical methods can be developed for analyzing the difference in means parameter for normal samples in two situations:

1. *Normal Samples, Known Variances:* Assume that σ_x and σ_y are known. Then the exactly standardized difference

   $$Z = \frac{(\overline{X} - \overline{Y}) - \delta}{\sqrt{\frac{\sigma_x^2}{n} + \frac{\sigma_y^2}{m}}} \quad \text{is a standard normal random variable.}$$

   The standard normal random variable is used in statistical inference problems.

2. *Normal Samples, Estimated Common Variance:* Assume that $\sigma_x = \sigma_y$, but the common value is not known. Then the approximately standardized difference

   $$T = \frac{(\overline{X} - \overline{Y}) - \delta}{\sqrt{S_p^2 \left(\frac{1}{n} + \frac{1}{m}\right)}} \quad \text{has a Student t distribution with } n + m - 2 \textit{ df},$$

   where S_p^2 (the *pooled estimator of the common variance*) is defined below. The Student t random variable is used in statistical inference problems.

The formula for the pooled estimator of the common variance is as follows:

$$S_p^2 = \frac{(n-1)S_x^2 + (m-1)S_y^2}{n + m - 2}$$

where S_x^2 and S_y^2 are the sample variances for the X and Y samples.

Note that if we let $\sigma = \sigma_x = \sigma_y$ be the common standard deviation, then

$$\frac{(n+m-2)S_p^2}{\sigma^2} = \frac{(n-1)S_x^2}{\sigma^2} + \frac{(m-1)S_y^2}{\sigma^2}$$

is the sum of independent chi-square random variables, and hence chi-square. The degrees of freedom of the sum is the sum of the degrees of freedom: $(n-1)+(m-1) = (n+m-2)$.

Question: S_p^2 is an unbiased estimator of the common variance. Can you see why?

Exercise. Use the definition of the Student t distribution, and your knowledge of the sampling distributions based on samples from normal distributions, to explain why the T statistic given in item 2 above has a Student t distribution.

Confidence interval procedures. The following tables give $100(1-\alpha)\%$ confidence interval procedures for the difference in means parameter, $\delta = \mu_x - \mu_y$.

1. σ_x, σ_y Known, Normal Samples:

$$\left(\overline{X} - \overline{Y}\right) \ \pm \ z(\alpha/2)\sqrt{\frac{\sigma_x^2}{n} + \frac{\sigma_y^2}{m}}$$

where $z(\alpha/2)$ is the $100(1-\alpha/2)\%$ point of the standard normal distribution.

2. $\sigma_x = \sigma_y$ Estimated, Normal Samples:

$$\left(\overline{X} - \overline{Y}\right) \ \pm \ t_{n+m-2}(\alpha/2)\sqrt{S_p^2\ \left(\frac{1}{n} + \frac{1}{m}\right)}$$

where $t_{n+m-2}(\alpha/2)$ is the $100(1-\alpha/2)\%$ point on the Student t distribution with $(n+m-2)$ df.

Hypothesis testing procedures. The following table gives size α tests of the null hypothesis that the difference in means parameter is a fixed value: $H_0 : \delta = \delta_o$.

	1. σ_x, σ_y Known, Normal Samples	*2. $\sigma_x = \sigma_y$ Estimated, Normal Samples*
Test Statistic	$Z = \dfrac{\left(\overline{X} - \overline{Y}\right) - \delta_o}{\sqrt{\frac{\sigma_x^2}{n} + \frac{\sigma_y^2}{m}}}$	$T = \dfrac{(\overline{X} - \overline{Y}) - \delta_0}{\sqrt{S_p^2\ \left(\frac{1}{n} + \frac{1}{m}\right)}}$
RR for $H_A : \delta < \delta_o$	$Z \le -z(\alpha)$	$T \le -t_{n+m-2}(\alpha)$
RR for $H_A : \delta > \delta_o$	$Z \ge z(\alpha)$	$T \ge t_{n+m-2}(\alpha)$
RR for $H_A : \delta \ne \delta_o$	$\lvert Z\rvert \ge z(\alpha/2)$	$\lvert T\rvert \ge t_{n+m-2}(\alpha/2)$

Exercise. Assume the following data are the values of independent random samples from normal distributions with common standard deviation 2.

1. *X Sample ($n = 8$, $\overline{x} = 10.1$):*

   6.07, 7.00, 9.49, 9.76, 11.19, 11.31, 12.96, 13.02

2. *Y Sample ($m = 12$, $\overline{y} = 6.83$):*

   3.86, 4.52, 5.14, 5.23, 5.33, 6.32, 7.21, 7.56, 7.94, 8.19, 9.07, 11.59

(*a*) Construct a 95% confidence interval for the difference in means, $\mu_x - \mu_y$.

(*b*) Consider testing $\mu_x - \mu_y = 4$ versus $\mu_x - \mu_y \neq 4$ using the information provided above. Would the null hypothesis be accepted or rejected at the 5% significance level? State the conclusion and report the observed significance level (p value).

Exercise (Source: Shoemaker, JSE, 1996): Body temperatures of 148 healthy adults were taken several times over two consecutive days. A total of 130 values are reported below.

1. *X Sample:* 65 temperatures (in degrees Fahrenheit) for women

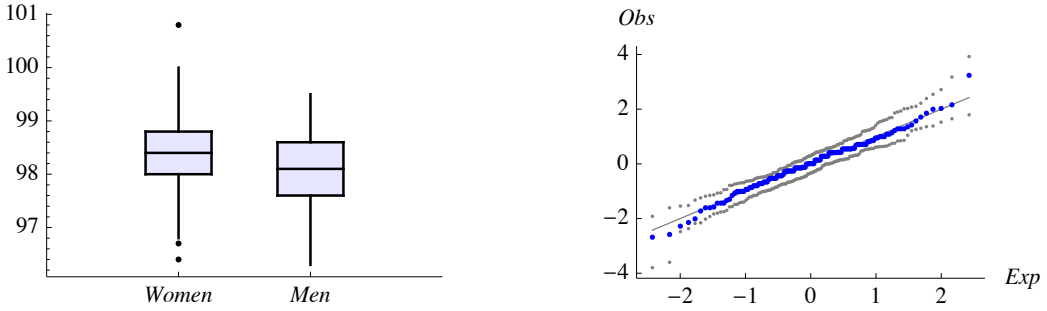
96.4	96.7	96.8	97.2	97.2	97.4	97.6	97.7	97.7	97.8	97.8	97.8	97.9
97.9	97.9	98.0	98.0	98.0	98.0	98.0	98.1	98.2	98.2	98.2	98.2	98.2
98.2	98.3	98.3	98.3	98.4	98.4	98.4	98.4	98.4	98.5	98.6	98.6	98.6
98.6	98.7	98.7	98.7	98.7	98.7	98.7	98.8	98.8	98.8	98.8	98.8	98.8
98.8	98.9	99.0	99.0	99.1	99.1	99.2	99.2	99.3	99.4	99.9	100.0	100.8

Sample summaries: $n = 65$, $\overline{x} = 98.3938$, $s_x = 0.7435$

2. *Y Sample:* 65 temperatures (in degrees Fahrenheit) for men

96.3	96.7	96.9	97.0	97.1	97.1	97.1	97.2	97.3	97.4	97.4	97.4	97.4
97.5	97.5	97.6	97.6	97.6	97.7	97.8	97.8	97.8	97.8	97.9	97.9	98.0
98.0	98.0	98.0	98.0	98.0	98.1	98.1	98.2	98.2	98.2	98.2	98.3	98.3
98.4	98.4	98.4	98.4	98.5	98.5	98.6	98.6	98.6	98.6	98.6	98.6	98.7
98.7	98.8	98.8	98.8	98.9	99.0	99.0	99.0	99.1	99.2	99.3	99.4	99.5

Sample summaries: $m = 65$, $\overline{y} = 98.1046$, $s_y = 0.6988$



1. *Left Plot:* Side-by-side box plots of the two samples are shown on the left. The sample distributions are approximately symmetric.

2. *Right Plot:* A normal probability plot of standardized temperatures is shown on the right, where

   (a) each x value is replaced by $(x - \overline{x})/s_x$;

   (b) each y value is replaced by $(y - \overline{y})/s_y$; and

   (c) the 130 ordered standardized values (vertical axis; observed) are plotted against the $k/131^{\text{st}}$ quantiles of the standard normal distribution (horizontal axis; expected).

   The normal probability plot has been enhanced to include the results of 100 simulations from the standard normal distribution: For each $k = 1, 2, \ldots, 130$, the minimum and maximum value of the 100 simulated k^{th} order statistics are plotted.

Assume these data are the values of independent random samples from normal distributions with a common variance.

(*a*) Test the null hypothesis that $\mu_x = \mu_y$ versus the alternative hypothesis that $\mu_x \neq \mu_y$ at the 5% level. Clearly state your conclusion.

(*b*) Construct and interpret a 95% confidence interval for the difference in means, $\mu_x - \mu_y$.

(*c*) Comment on the results of parts (*a*) and (*b*).

Exercise (Source: Larsen & Marx, 1986): Electroencephalograms are records showing fluctuations of electrical activity in the brain. Among the several different kinds of brain waves produced, the dominant ones are usually *alpha* waves. These have a characteristic frequency of anywhere from 8 to 13 cycles per second.

As part of a study to determine if sensory deprivation over an extended period of time has any effect on alpha-wave pattern, 20 male inmates in a Canadian prison were randomly split into two equal-sized groups. Members of one group (*control group*) were allowed to remain in their cells, while members of the other group (*treated group*) were placed in solitary confinement. After seven days, alpha-wave frequencies were measured in all 20 men:

1. *X Sample:* Average number of cycles per second for members of the control group:

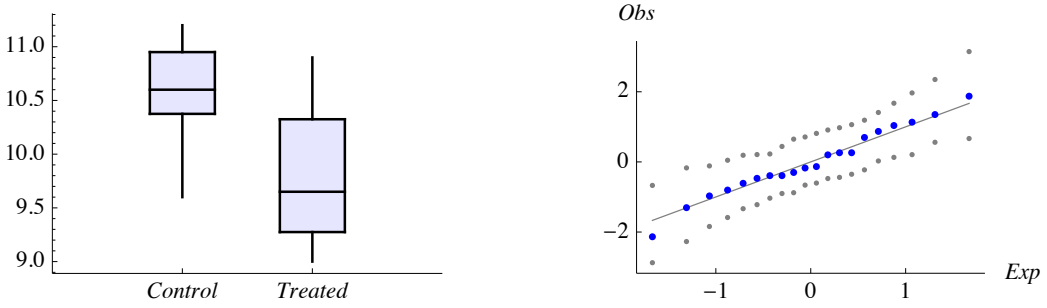
   9.6, 10.3, 10.4, 10.4, 10.5, 10.7, 10.7, 10.9, 11.1, 11.2

   Sample summaries: $n = 10$, $\overline{x} = 10.58$, $s_x = 0.4590$

2. *Y Sample:* Average number of cycles per second for members of the treated group:

   9.0, 9.2, 9.3, 9.5, 9.6, 9.7, 9.9, 10.3, 10.4, 10.9

   Sample summaries: $m = 10$, $\overline{y} = 9.78$, $s_y = 0.5978$



1. *Left Plot:* Side-by-side box plots, shown on the left, suggest that population means for non-confined and solitary-confined prisoners are different.

2. *Right Plot:* Enhanced normal probability plot of standardized averages suggests that normal theory methods are reasonable (although the plot tells us nothing about whether the assumption of a common variance is reasonable).

Note that use of a two-sample method in this example is justified under what is known as a "randomization model" for inference. By randomizing, the researchers have assured that the inmates were as alike as possible, except for the "treatment."

Assume these data are the values of independent random samples from normal distributions with a common variance.

(*a*) Test the null hypothesis that $\mu_x = \mu_y$ versus the alternative hypothesis that $\mu_x \neq \mu_y$ at the 5% level. Clearly state your conclusion.

(*b*) Construct and interpret a 95% confidence interval for the difference in means, $\mu_x - \mu_y$.

(*c*) Comment on the results of parts (*a*) and (*b*).

5.1.3 Approximate Methods for Normal Distributions

Our third method for normal samples handles the case where X and Y have distinct variances that must be estimated from the data. In this case, an exact sampling distribution that can be used for statistical inference questions is not possible. Instead, a clever method developed by B. Welch in the 1940's allows us to analyze the difference of means parameter:

3. *Normal Samples, Estimated Distinct Variances:* Assume that $\sigma_x \neq \sigma_y$ and that both are unknown. Then the approximately standardized difference

$$T \;=\; \frac{(\overline{X}-\overline{Y})-\delta}{\sqrt{\frac{S_x^2}{n}+\frac{S_y^2}{m}}} \quad \text{has an approximate Student t distribution,}$$

with degrees of freedom as follows:

$$df \;=\; \frac{\left((S_x^2/n)+(S_y^2/m)\right)^2}{\left((S_x^2/n)^2/n+(S_y^2/m)^2/m\right)} \;-\; 2.$$

The approximate Student t distribution is used in statistical inference problems.

Note: The value of *df* is often called the "effective" degrees of freedom. The effective degrees of freedom always lies between the degrees of freedom for analyzing one sample (using the smaller sample size) and the degrees of freedom for analyzing two samples drawn from distributions with an estimated common variance:

$$\min(n,m)-1 \;\leq\; df \;\leq\; n+m-2.$$

It is interesting to note that Welch's initial formula for the degrees of freedom did not include the "-2" term at the end; the extra term was added in later refinements of the method.

Confidence interval procedures. The following table gives an approximate $100(1-\alpha)\%$ confidence interval procedure for the difference in means parameter, $\delta = \mu_x - \mu_y$.

3. $\sigma_x \neq \sigma_y$ Estimated, Normal Samples:

$$(\overline{X}-\overline{Y}) \;\pm\; t_{df}(\alpha/2)\sqrt{\frac{S_x^2}{n}+\frac{S_y^2}{m}}$$

where $t_{df}(\alpha/2)$ is the $100(1-\alpha/2)\%$ point on the Student t distribution with *df* degrees of freedom.

Hypothesis testing procedures. The following table gives approximate size α tests of the null hypothesis that the difference in means parameter is a fixed value: $H_0 : \delta = \delta_o$.

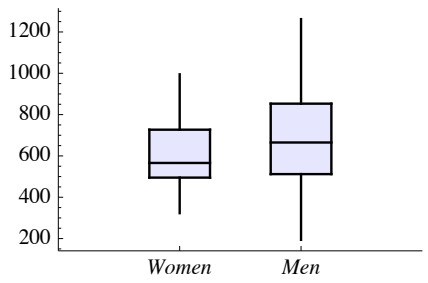
3. $\sigma_x \neq \sigma_y$ Estimated, Normal Samples

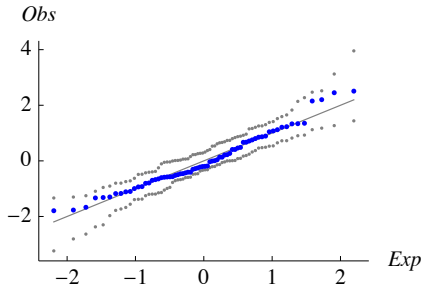
Test Statistic	$T = \dfrac{(\overline{X} - \overline{Y}) - \delta_o}{\sqrt{\frac{S_x^2}{n} + \frac{S_y^2}{m}}}$
RR for $H_A : \delta < \delta_o$	$T \leq -t_{df}(\alpha)$
RR for $H_A : \delta > \delta_o$	$T \geq t_{df}(\alpha)$
RR for $H_A : \delta \neq \delta_o$	$\lvert T \rvert \geq t_{df}(\alpha/2)$

Exercise (Source: Stukel, 1998, FTP:lib.stat.cmu.edu/datasets/): Several studies have suggested that low levels of plasma retinol (vitamin A) are associated with increased risk of certain types of cancer. As part of a study to investigate the relationship between personal characteristics and cancer incidence, data were gathered on 315 subjects. Study subjects were patients who had an elective surgical procedure to biopsy or remove a lesion that was found to be non-cancerous. This exercise compares mean plasma levels of retinol in nanograms per milliliter (ng/ml) for 35 women and 35 men who participated in the study. Numerical and graphical summaries are as follows:

Women: $n = 35,\ \overline{x} = 600.943,\ s_x = 157.103$

Men: $m = 35,\ \overline{y} = 673.457,\ s_y = 267.370$





Let X and Y be the plasma retinol levels for women and men, respectively, with non-cancerous lesions. Assume that X and Y have normal distributions and that the information above summarizes the results of independent random samples from these distributions.

(*a*) Use Welch's formula to find df for the approximate Student t distribution.

.

(*b*) Construct and interpret an approximate 95% confidence for the difference in means parameter, $\delta = \mu_x - \mu_y$.

5.1.4 Recap; Pooled t Methods, Welch t Methods

In the previous sections, we discussed three methods for analyzing the difference in means parameter when samples are drawn independently from normal distributions:

1. *Known Variances:* When both variances are known, we can use the standard normal distribution to find cutoffs.
2. *Estimated Common Variance:* When we can assume that the distributions have a common (but unknown) variance, we can use S_p^2 to estimate the common variance and the Student t distribution with $(n + m - 2)$ degrees of freedom to find cutoffs.
3. *Estimated Distinct Variances:* When we cannot assume that the unknown variances are equal, we can use the Student t distribution with degrees of freedom given by Welch's formula to find cutoffs.

Analyses that use the pooled estimate of the common variance are often called *pooled t methods*; those using Welch's formula to find degrees of freedom are often called *Welch t methods*.

5.2 Two Sample Analysis: Ratio of Variances

Assume that X and Y are normal random variables.

This section develops methods for answering statistical questions about the ratio of variances parameter, $r = \sigma_x^2/\sigma_y^2$, for normal distributions.

The ratio of sample variances, S_x^2/S_y^2, is used to estimate r. The sampling distribution of the ratio of sample variances is related to the *f ratio distribution*, which is introduced first.

5.2.1 F Ratio Distribution

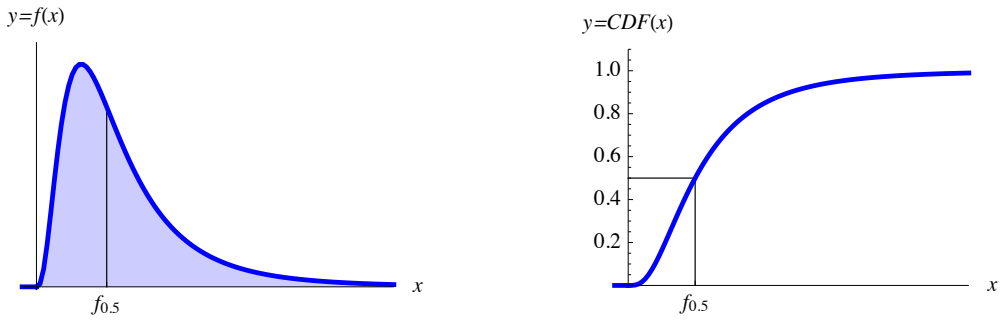
Let U and V be independent chi-square random variables with n_1 and n_2 degrees of freedom, respectively. Then

$$F \;=\; \frac{U/n_1}{V/n_2}$$

is said to be an *f ratio random variable*, or to have an *f ratio distribution*, with n_1 and n_2 degrees of freedom (*df*). The PDF of F is as follows:

$$f(x) = \frac{\Gamma((n_1+n_2)/2)}{\Gamma(n_1/2)\Gamma(n_2/2)} \left(\frac{n_1}{n_2}\right)^{n_1/2} \left(\frac{n_2}{n_2+n_1 x}\right)^{(n_1+n_2)/2} \quad \text{when } x > 0, \text{ and 0 otherwise.}$$

Typical forms for the PDF and CDF of F are shown below.



The location of the median, $f_{0.5}$, has been labeled in each plot.

Notes:

1. *"Fisher" Ratio Distribution:* The f in "f ratio distribution" is for R.A. Fisher, who pioneered its use in analyzing the results of comparative studies (that is, in analyzing the results of studies comparing two or more samples).

2. *Shape:* Both parameters govern shape and scale.

   (a) If $n_2 > 2$, then $E(F) = \dfrac{n_2}{n_2 - 2}$.

   (b) If $n_2 > 4$, then $Var(F) = \dfrac{2n_2^2(n_1 + n_2 - 2)}{n_1(n_2 - 4)(n_2 - 2)^2}$.

   In all other cases, the summaries are indeterminate. Note that $E(F) \to 1$ as $n_2 \to \infty$.

3. *Reciprocal:* If F has an f ratio distribution with n_1 and n_2 degrees of freedom, then the reciprocal of F has an f ratio distribution with n_2 and n_1 degrees of freedom.

4. *Quantiles:* The notation f_p is used to denote the pth quantile ($100p$th percentile) of the f ratio distribution. The Rice textbook includes tables for $p = 0.90$ (page A10), $p = 0.95$ (page A11), $p = 0.975$ (page A12), and $p = 0.99$ (page A13).

   The $p = 0.10\,, 0.05,\ 0.025,\ 0.01$ quantiles can be computed using reciprocals. Specifically,

$$f_p \text{ on } n_1,\, n_2 \ df \ = \ \frac{1}{f_{1-p} \text{ on } n_2,\, n_1 \ df}$$

To illustrate the use of the tables in the textbook, suppose that we are interested in determining the 5% point of the f ratio distribution with 10 and 8 degrees of freedom.

Using the relationship above, we compute

$$f_{0.05} \text{ on } 10{,}8 \ df \ = \ \frac{1}{f_{0.95} \text{ on } 8{,}10 \ df} \ = \ \frac{1}{3.35} = 0.30.$$

5.2.2 Sampling Distribution of Ratio of Sample Variances

Let X be a normal random variable with mean μ_x and standard deviation σ_x, and let Y be a normal random variable with mean μ_y and standard deviation σ_y.

The following theorem tells us about the sampling distribution of the ratio of sample variances when samples are chosen independently from the X and Y distributions.

> ***Theorem (Sampling Distribution).*** Let S_x^2 and S_y^2 be the sample variances of independent random samples of sizes n and m, respectively, from the X and Y distributions. Then
>
> $$F = \frac{S_x^2/S_y^2}{\sigma_x^2/\sigma_y^2}$$
>
> has an f ratio distribution with $(n-1)$ and $(m-1)$ degrees of freedom, where the numerator is the ratio of sample variances and the denominator is the ratio of model variances.

To demonstrate that the conclusion of the theorem is correct, first note that

1. $U = \frac{(n-1)}{\sigma_x^2} S_x^2$ has a chi-square distribution with $(n-1)$ *df*, and
2. $V = \frac{(m-1)}{\sigma_y^2} S_y^2$ has a chi-square distribution with $(m-1)$ *df*.

Now (please complete the demonstration),

5.2.3 Exact Methods for Normal Distributions

Let X and Y be normal random variables. Under the conditions of the last section, the following tables give exact confidence interval and hypothesis test methods for the ratio of variances parameter, $r = \sigma_x^2/\sigma_y^2$.

1. $100(1-\alpha)\%$ *CI for* $r = \sigma_x^2/\sigma_y^2$*, when* μ_x *and* μ_y *are estimated:*

$$\left[\frac{S_x^2/S_y^2}{f_{n-1,m-1}(\alpha/2)},\ \frac{S_x^2/S_y^2}{f_{n-1,m-1}(1-\alpha/2)}\right]$$

where $f_{n-1,m-1}(p)$ is the $100(1-p)\%$ point of the f ratio distribution with $(n-1)$ and $(m-1)$ *df.*

2. $100\alpha\%$ *tests of* $H_0 : r = r_o$*, when* μ_x *and* μ_y *are estimated:*

Test Statistic:	$F = \dfrac{S_x^2/S_y^2}{r_o}$
RR for $H_A : r < r_o$*:*	$F \le f_{n-1,m-1}(1-\alpha)$
RR for $H_A : r > r_o$*:*	$F \ge f_{n-1,m-1}(\alpha)$
RR for $H_A : r \ne r_o$*:*	$F \le f_{n-1,m-1}(1-\alpha/2)$ or $F \ge f_{n-1,m-1}(\alpha/2)$

First-step analyses. F ratio methods are often used as a first step in an analysis of the difference in means.

Exercise, continued. For example, in the alpha waves exercise (beginning on page 11), a confidence interval for the difference in means was constructed under the assumption that the population variances were equal.

To demonstrate that this assumption is justified, we test

$$\frac{\sigma_x^2}{\sigma_y^2} = 1 \quad \text{versus} \quad \frac{\sigma_x^2}{\sigma_y^2} \ne 1$$

at the 5% significance level. The rejection region for the test is

$$F \le f_{9,9}(0.975) = \frac{1}{4.03} = 0.25 \quad \text{or} \quad F \ge f_{9,9}(0.025) = 4.03$$

and the observed value of the test statistic is $s_x^2/s_y^2 = 0.5897$.

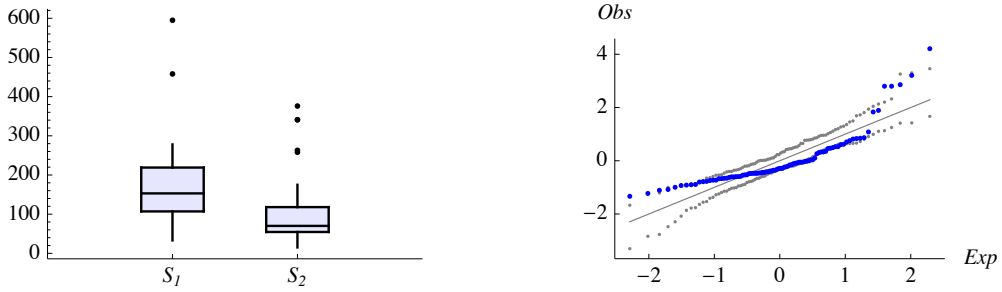
Since the observed value of the test statistic is in the acceptance region, the hypothesis of equal variances is accepted; there is insufficient evidence to conclude otherwise.

5.3 Transformations to Normality

Methods based on sampling from normal distributions are popular and easy to apply.

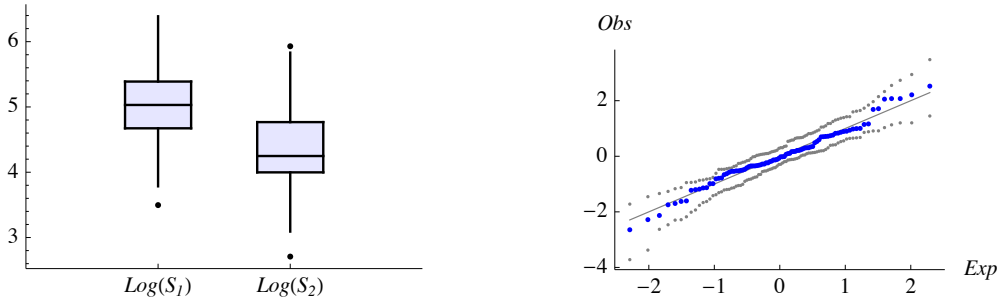
For this reason, researchers often transform their data to achieve approximate normality, and then use normal theory methods on the transformed scale.

For example, let S_1 and S_2 be samples of size 45 each drawn from skewed positive distributions. The *left* plot shows side-by-side box plots and the *right* plot shows an enhanced normal probability plot of combined standardized values.



Notice that the boxes are asymmetric, there are large outliers, and the normal probability plot has a pronounced bend.

Now, let $\log(S_1)$ and $\log(S_2)$ be the log-transformed samples.



The log-transformed samples are more symmetric and closer to being normally distributed. The plots above suggest that normal theory methods could now be used.

Footnotes: Although the use of transformations is attractive, there are many drawbacks. For example, it may be difficult to find an appropriate transformation, or it may be difficult to interpret the results back on the original scale.

For this reason, we will study methods applicable to broad ranges of distributions in the remaining sections of this notebook.

5.4 Nonparametric Methods for Two Sample Analysis

This section focuses on broadly-applicable statistical methods for two sample analyses.

We begin by contrasting parametric and nonparametric methods:

1. *Parametric Methods:* Statistical methods that require strong assumptions about the shapes of the distributions from which the data are drawn (for example, assuming the data are drawn from a normal distribution), and ask questions about the parameter(s) of the distribution.
2. *Nonparametric Methods:* Statistical methods that make mild assumptions about the distributions from which the data are drawn, such as "the distributions are continuous" or "the continuous distributions are symmetric around their centers."

Note: Nonparametric methods are also known as *distribution-free methods.*

5.4.1 Stochastically Larger and Stochastically Smaller Random Variables

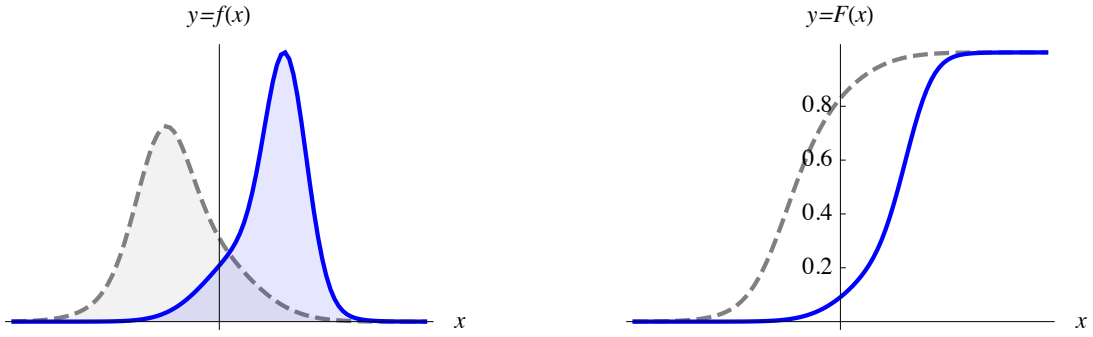
Let V and W be continuous random variables.

V is *stochastically larger* than W (corresponding, W is *stochastically smaller* than V) if

$$P(V \geq x) \;\geq\; P(W \geq x) \quad \text{for all real numbers } x,$$

with strict inequality (that is, where ">" replaces "$\geq$") for at least one x.

To illustrate the definition of stochastically larger/smaller, consider the following plots of density functions (*left* plot), and cumulative distribution functions (*right* plot), of two continuous random variables: V (solid curve) and W (dashed curve).



V is stochastically larger than W (correspondingly, W is stochastically smaller than V).

Note: If V is stochastically larger than W, then their CDFs satisfy the inequality $$F_V(x) \leq F_W(x) \quad \text{for all } x,$$ with strict inequality for at least one x.

5.4.2 Wilcoxon Rank Sum Statistic

In the 1940's, Wilcoxon developed a nonparametric method for testing the null hypothesis that two continuous distributions are equal versus the alternative hypothesis that one distribution is stochastically larger than the other.

Given independent random samples,

$$X_1, X_2, \ldots, X_n, \quad \text{and} \quad Y_1, Y_2, \ldots, Y_m,$$

from the X and Y distributions, Wilcoxon *rank sum statistics* for the X sample (R_1) and for the Y sample (R_2) are computed as follows:

1. Pool and sort the $n + m$ observations.
2. Replace each observation by its *rank* (or position) in the sorted list.
3. Let R_1 equal the sum of the ranks for observations in the X sample, and R_2 equal the sum of the ranks for observations in the Y sample.

Note that since

$$R_1 + R_2 = 1 + 2 + \cdots + (n + m) = \frac{(n + m)(n + m + 1)}{2},$$

tests based on R_1 are equivalent to tests based on R_2. We will use the R_1 statistic.

For example,

1. If $n = 4$, $m = 6$ and the data are as follows

   $$1.1,\ 2.5,\ 3.2,\ 4.1 \qquad \text{and} \qquad 2.8,\ 3.6,\ 4.0,\ 5.2,\ 5.8,\ 7.2$$

   then the sorted combined list of $n + m = 10$ observations is

   $$1.1,\ 2.5,\ 2.8,\ 3.2,\ 3.6,\ 4.0,\ 4.1,\ 5.2,\ 5.8,\ 7.2.$$

   The observed value of R_1 is _____

   The observed value of R_2 is _____

2. If $n = 9$, $m = 5$ and the data are as follows

   $$12.8,\ 15.6,\ 15.7,\ 17.3,\ 18.5,\ 22.9,\ 27.5,\ 29.7,\ 35.1 \quad \text{and} \quad 8.2,\ 12.6,\ 16.7,\ 21.6,\ 32.4$$

   then the sorted combined list of $n + m = 14$ observations is

   $$8.2,\ 12.6,\ 12.8,\ 15.6,\ 15.7,\ 16.7,\ 17.3,\ 18.5,\ 21.6,\ 22.9,\ 27.5,\ 29.7,\ 32.4,\ 35.1.$$

   The observed value of R_1 is _____

   The observed value of R_2 is _____

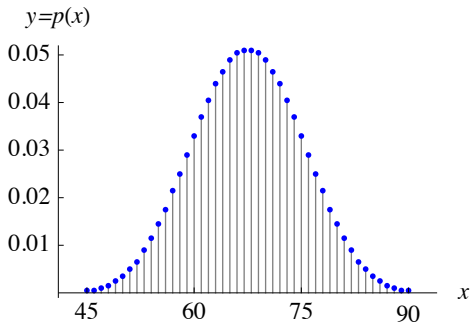
5.4.3 Wilcoxon Rank Sum Distribution and Methods

The following theorem gives us information about the distribution of the Wilcoxon rank sum statistic for the X sample under the null hypothesis that the X and Y distributions are equal.

> ***Rank Sum Distribution Theorem.*** Let X and Y be continuous distributions, and R_1 be the Wilcoxon rank sum statistic for the X sample based on independent random samples of sizes n and m, respectively, from the X and Y distributions. If the distributions of X and Y are equal, then
>
> 1. The range of R_1 is $\frac{n(n+1)}{2}$, $\frac{n(n+1)}{2}+1$, ..., $nm+\frac{n(n+1)}{2}$.
> 2. $E(R_1)=\frac{n(n+m+1)}{2}$ and $Var(R_1)=\frac{nm(n+m+1)}{12}$.
> 3. The distribution of R_1 is symmetric around its mean. In particular,
>
> $$P(R_1=x) \;=\; P(R_1=n(n+m+1)-x)\ .$$
>
> 4. If n and m are large, then the distribution of R_1 is approximately normal. (If both are greater than 20, then the approximation is reasonably good.)

The proof of the distribution theorem uses combinatorics since, under the null hypothesis of equality of distributions, each subset of size n from ranks $\{1,2,\ldots,n+m\}$ is equally likely.

To illustrate the distribution theorem, let $n=9$ and $m=5$. The plot below shows the PDF of the R_1 statistic for all values in the range of the random variable.



- The range of R_1 is the integers between _____ and _____.
- The mean of the R_1 distribution is _____.
- The variance of the R_1 distribution is _____.

Exercise. Let $n = 2$ and $m = 4$.

(*a*) List all $\binom{6}{2} = 15$ subsets of size 2 from $\{1, 2, 3, 4, 5, 6\}$.

(*b*) Use your answer to part (*a*) to completely specify the PDF of R_1.

(*c*) Use your answer to part (*b*) to find the mean and variance of R_1.

Finding p values. Let r_{obs} be the observed value of R_1 for a given set of data. Then observed significance levels (p values) are obtained as follows:

Alternative Hypothesis	*P Value*
X is stochastically larger than Y	$P(R_1 \geq r_{\text{obs}})$
X is stochastically smaller than Y	$P(R_1 \leq r_{\text{obs}})$
One random variable is stochastically larger or smaller than the other	If $r_{\text{obs}} > E(R_1)$, then $2P(R_1 \geq r_{\text{obs}})$, and if $r_{\text{obs}} < E(R_1)$, then $2P(R_1 \leq r_{\text{obs}})$. (If $r_{\text{obs}} = E(R_1)$, then the p value is 1.)

For example, let $n = 9$ and $m = 5$.

x	$P(R_1 = x)$	$P(R_1 \leq x)$	x	$P(R_1 = x)$	$P(R_1 \leq x)$	x	$P(R_1 = x)$	$P(R_1 \leq x)$
45	0.0005	0.0005	61	0.0370	0.2188	77	0.0250	0.9051
46	0.0005	0.0010	62	0.0405	0.2592	78	0.0215	0.9266
47	0.0010	0.0020	63	0.0440	0.3032	79	0.0175	0.9441
48	0.0015	0.0035	64	0.0465	0.3497	80	0.0145	0.9585
49	0.0025	0.0060	65	0.0490	0.3986	81	0.0115	0.9700
50	0.0035	0.0095	66	0.0504	0.4491	82	0.0090	0.9790
51	0.0050	0.0145	67	0.0509	0.5000	83	0.0065	0.9855
52	0.0065	0.0210	68	0.0509	0.5509	84	0.0050	0.9905
53	0.0090	0.0300	69	0.0504	0.6014	85	0.0035	0.9940
54	0.0115	0.0415	70	0.0490	0.6503	86	0.0025	0.9965
55	0.0145	0.0559	71	0.0465	0.6968	87	0.0015	0.9980
56	0.0175	0.0734	72	0.0440	0.7408	88	0.0010	0.9990
57	0.0215	0.0949	73	0.0405	0.7812	89	0.0005	0.9995
58	0.0250	0.1199	74	0.0370	0.8182	90	0.0005	1.0000
59	0.0290	0.1489	75	0.0330	0.8511			
60	0.0330	0.1818	76	0.0290	0.8801			

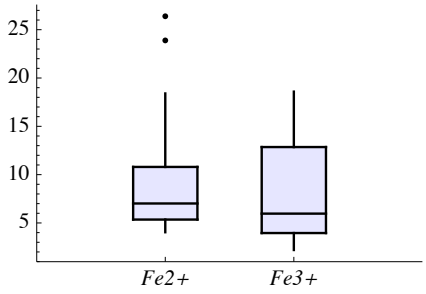
1. If the alternative hypothesis is "X is stochastically smaller than Y" and the observed value of R_1 is 60, then the observed significance level is

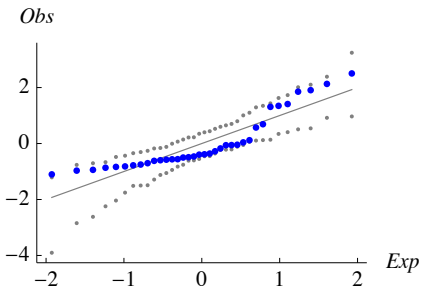
2. If the alternative hypothesis is "X is stochastically larger than Y" and the observed value of R_1 is 74, then the observed significance level is

3. If the alternative hypothesis is "One random variable is stochastically larger than the other" and the observed value of R_1 is 48, then the observed significance level is

Example (Source: Rice textbook, Chapter 11). "An experiment was performed to determine whether two forms of iron (Fe2+ and Fe3+) are retained differently. (If one form of iron was retained especially well, it would be the better dietary supplement.) The investigators divided 108 mice randomly into 6 groups of 18 each; three groups were given Fe2+ in three different concentrations, 10.2, 1.2, and 0.3 millimolar, and three groups were given Fe3+ at the same concentrations. The mice were given the iron orally; the iron was radioactively labeled so that a counter could be used to measure the initial amount given. At a later time, another count was taken for each mouse, and the percentage of iron retained was calculated."

This example considers results for the second concentration (1.2 millimolar). The *left* plot below shows side-by-side box plots of the percent retention for each group, and the *right* plot is an enhanced normal probability plot of the 36 standardized values.





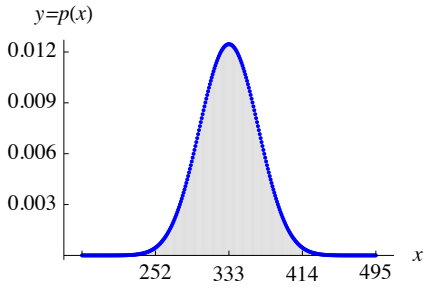
Let X and Y represent percentages of iron retained at 1.2 millimolar concentration by mice given Fe2+ and Fe3+, respectively, and assume that the data collected by the researchers are the values of independent random samples from the X and Y distributions.

Since the plots suggest that the distributions are *not* approximately normal, the equality of the X and Y distributions will be tested using the Wilcoxon rank sum test. A two-sided alternative will be used, and the 5% significance level.

The sampling distribution of R_1 has range

$$\mathcal{R} = \{171, 172, \ldots, 495\},$$

and is centered at 333.



The observed value of R_1 for these data is 362, and the p value is

$$2P(R_1 \geq 362) \approx 0.372.$$

Thus (please complete),

Handling equal observations. Continuous data are often rounded to a fixed number of decimal places, causing two or more observations to be equal.

1. *Tied Observations:* Equal observations are said to be *tied* at a given value.
2. *Midranks:* If two or more observations are tied at a given value, then their average rank (or *midrank*) is used to compute the rank sum statistic. For example, if the two smallest observations are equal, they would each be assigned rank $(1 + 2)/2 = 1.5$.
3. *Sampling Distribution:* To obtain the sampling distribution of R_1, we use a simple urn model: Imagine writing the $n + m$ midranks on separate slips of paper and placing the slips in an urn. After thoroughly mixing the urn, choose a subset of size n and compute the sum of the values on the chosen slips.

   If each choice of subset is equally likely, then the resulting probability distribution is the distribution of R_1 for the given collection of midranks.

For example, suppose that $n = 4$, $m = 6$ and the first row of the following table represents the sorted combined sample of 10 observations.

Combined Sample:	4.1	5.5	8.7	11.1	11.1	12.6	12.6	12.6	18.3	19.7
List of Midranks:	1	2	3	4.5	4.5	7	7	7	9	10

Then the second row of the table represents the numbers used to construct the distribution of R_1 under the null hypothesis that the X and Y distributions are equal.

The sampling distribution of R_1 is obtained by considering all $\binom{10}{4} = 210$ subsets of size 4 chosen from the following set of slips:

$$\boxed{1}\ \boxed{2}\ \boxed{3}\ \boxed{4.5}\ \boxed{4.5}\ \boxed{7}\ \boxed{7}\ \boxed{7}\ \boxed{9}\ \boxed{10}$$

Several random partitions of the slips are given in the following table, along with the value of R_1 in each case (please complete):

Midranks in First Sample:	*Midranks in Second Sample:*	*Value of R_1:*
2, 4.5, 7, 7	1, 3, 4.5, 7, 9, 10	_____
3, 4.5, 4.5, 7	1, 2, 7, 7, 9, 10	_____
3, 7, 9, 10	1, 2, 4.5, 4.5, 7, 7	_____
4.5, 7, 7, 9	1, 2, 3, 4.5, 7, 10	_____
...	...	...

Under the null hypothesis, each choice of subset for the first sample is equally likely.

Example (Source: Rice textbook, Chapter 11). "Two methods, A and B, were used in a determination of the latent heat of fusion of ice (Natrella 1963). The investigators wished to find out by how much the methods differed. The following table gives the change in total heat from ice at -0.72^oC to water 0^oC in calories per gram of mass."

1. *X sample:* 13 observations (calories/gram) using Method A:

   79.97 79.98 80.00 80.02 80.02 80.02 80.03 80.03 80.03 80.04 80.04 80.04 80.05

2. *Y sample:* 8 observations (calories/gram) using Method B:

   79.94 79.95 79.97 79.97 79.97 79.98 80.02 80.03

The following table shows the ordered values and corresponding midranks:

	Observation:	*Midrank:*		*Observation:*	*Midrank:*		*Observation:*	*Midrank:*
1	79.94	1.0	8	79.98	7.5	15	80.03	15.5
2	79.95	2.0	9	80.00	9.0	16	80.03	15.5
3	79.97	4.5	10	80.02	11.5	17	80.03	15.5
4	79.97	4.5	11	80.02	11.5	18	80.04	19.0
5	79.97	4.5	12	80.02	11.5	19	80.04	19.0
6	79.97	4.5	13	80.02	11.5	20	80.04	19.0
7	79.98	7.5	14	80.03	15.5	21	80.05	21.0

The observed value of R_1 is _____

The observed value of R_2 is _____

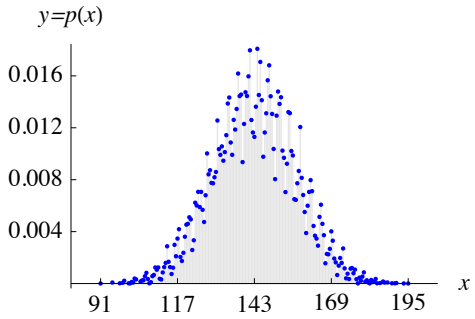
Let X and Y represent measurements taken with methods A and B, respectively, and assume that the data above are the values of independent random samples from the X and Y distributions. The equality of the distributions will be tested using the Wilcoxon rank sum test, a two-sided alternative, and 5% significance level.

The R_1 statistic takes whole-number and half-number values between 91 and 195, and is centered at 143.

The observed value of R_1 is _____, and the observed significance level is

$$2P(R_1 \geq \_\_\_\_\_) \approx 0.005.$$

Thus (please complete),



5.4.4 Mann-Whitney U Statistic

In the 1940's, Mann & Whitney developed a nonparametric two-sample test for the null hypothesis that two continuous distributions are equal versus the alternative hypothesis that one distribution is stochastically larger than the other.

Given independent random samples,

$$X_1, X_2, \ldots, X_n, \quad \text{and} \quad Y_1, Y_2, \ldots, Y_m,$$

from the X and Y distributions, Mann-Whitney *U statistics* for the X sample (U_1) and for the Y sample (U_2) are defined as follows:

1. U_1 *Statistic:* The U_1 statistic equals the number of times an X observation is greater than a Y observation:

$$U_1 = \#(X_i > Y_j) = \sum_{i=1}^{n} \sum_{j=1}^{m} I(X_i > Y_j),$$

   where $I(X_i > Y_j) = 1$ if the inequality is true, and 0 otherwise.

2. U_2 *Statistic:* The U_2 statistic equals the number of times a Y observation is greater than an X observation:

$$U_2 = \#(Y_j > X_i) = \sum_{j=1}^{m} \sum_{i=1}^{n} I(Y_j > X_i),$$

   where $I(Y_j > X_i) = 1$ if the inequality is true, and 0 otherwise.

Note: If all $n + m$ observations are distinct, then $U_1 + U_2 =$ _____.

For example,

1. If $n = 4$, $m = 6$, and the data are as follows:

   1.1, 2.5, 3.2, 4.1 and 2.8, 3.6, 4.0, 5.2, 5.8, 7.2,

   then the sorted combined list of 10 observations (with the x-values underlined) is:

   <u>1.1</u>, <u>2.5</u>, 2.8, <u>3.2</u>, 3.6, 4.0, <u>4.1</u>, 5.2, 5.8, 7.2.

   The observed value of U_1 is _____

   The observed value of U_2 is _____

2. If $n = 5$, $m = 7$, and the data are as follows,

$$4.9,\ 7.3,\ 9.2,\ 11.0,\ 17.3 \qquad \text{and} \qquad 0.5,\ 0.7,\ 1.5,\ 2.7,\ 5.6,\ 8.7,\ 13.4,$$

then the sorted combined list of 12 observations (with the x-values underlined) is

$$0.5,\ 0.7,\ 1.5,\ 2.7,\ \underline{4.9},\ 5.6,\ \underline{7.3},\ 8.7,\ \underline{9.2},\ \underline{11.0},\ 13.4,\ \underline{17.3}.$$

The observed value of U_1 is _____

The observed value of U_2 is _____

5.4.5 Mann-Whitney U Distribution and Methods

The following theorem gives us information about the Mann-Whitney U statistic for the X sample under the null hypothesis that the X and Y distributions are equal, and relates the Mann-Whitney and Wilcoxon statistics.

U Statistic Distribution. Let X and Y be continuous distributions, and let U_1 and R_1 be the Mann-Whitney and Wilcoxon statistics for the X sample based on independent random samples of sizes n and m from the X and Y distributions. If the distributions of X and Y are equal, then

1. $U_1 = R_1 - \frac{n(n+1)}{2}$.
2. The range of U_1 is 0, 1, 2, ..., nm.
3. $E(U_1) = \frac{nm}{2}$ and $Var(U_1) = \frac{nm(n+m+1)}{12}$.
4. The distribution of U_1 is symmetric around its mean. In particular,

$$P(U_1 = x) \ = \ P(U_1 = nm - x)\ .$$

5. If n and m are large, then the distribution of U_1 is approximately normal. (If both are greater than 20, then the approximation is reasonably good.)

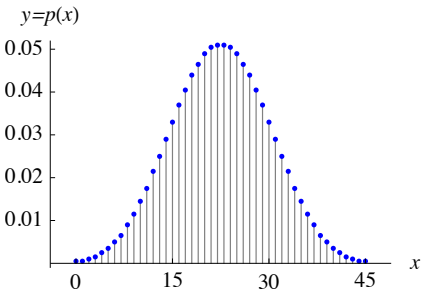
For example, if $n = 9$ and $m = 5$, then

- the range of U_1 is the integers between 0 and _____.
- the mean of the U_1 distribution is _____, and
- the variance of the U_1 distribution is _____.

The sampling distribution of U_1 is obtained by considering all

$$\binom{14}{9} = 2002$$

assignments of 9 observations to the first sample, with the remaining observations being assigned to the second sample. If the X and Y distributions are equal, then each assignment is equally likely.



It is instructive to prove the first part of the sampling distribution theorem, assuming that each observation can be written with infinite precision.

Let r_i be the rank of i^{th} order statistic of the X sample, $X_{(i)}$, for $i = 1, 2, \ldots, n$, in the ordered combined sample of $n + m$ observations.

Now (please complete the proof),

5.4.6 Shift Model; Hodges-Lehmann (HL) Estimation

The continuous random variables X and Y are said to satisfy a *shift model* if

$$X - \Delta \text{ and } Y \text{ have the same distribution,}$$

where Δ (the *shift parameter*) is the difference in medians: $\Delta = \text{Median}(X) - \text{Median}(Y)$.

Suppose that X and Y satisfy a shift model with shift parameter Δ. Then

1. *Same shape:* The distributions of X and Y must have the same shape.
2. *Stochastically larger/smaller:* We can use the shift parameter when answering questions about whether one distribution is stochastically larger than the other

   The following table summarizes the conclusions:

Value of Δ	Comparison of Distributions
$\Delta = 0$	X and Y have the same distribution
$\Delta > 0$	X is stochastically larger than Y
$\Delta < 0$	X is stochastically smaller than Y

Note: In many experimental settings, the goal is to compare a treatment group (where individuals are administered a treatment of interest) to a control group (where individuals are administered the standard treatment or a placebo).

If the effect of the treatment under study is additive, then a shift model holds and the shift parameter is referred to as the *treatment effect*.

Estimating the shift parameter. In the 1960's, Hodges & Lehmann developed a method to estimate the shift parameter in a shift model.

Given independent random samples of sizes n and m from the X and Y distributions:

1. *Walsh Differences:* The following list of nm differences,

   $$X_i - Y_j, \quad \text{for} \quad i = 1, 2, \ldots, n; \; j = 1, 2, \ldots, m,$$

   are called the *Walsh differences*. They are the *elementary estimates* of Δ.

   Note that the Walsh differences are a list of nm dependent random variables.

2. *Hodges-Lehmann Estimator:* The Hodges-Lehmann (HL) estimator of Δ is the median of the list of nm Walsh differences.

   Note that HL estimator of Δ is not necessarily equal to the difference of sample medians of the separate X and Y samples.

For example, if $n = 5$ and $m = 7$, and the data are as follows,

$$4.9,\ 7.3,\ 9.2,\ 11.0,\ 17.3 \qquad \text{and} \qquad 0.5,\ 0.7,\ 1.5,\ 2.7,\ 5.6,\ 8.7,\ 13.4,$$

then the following 5-by-7 table gives the 35 Walsh differences:

	0.5	0.7	1.5	2.7	5.6	8.7	13.4
4.9	4.4	4.2	3.4	2.2	–0.7	–3.8	–8.5
7.3	6.8	6.6	5.8	4.6	1.7	–1.4	–6.1
9.2	8.7	8.5	7.7	6.5	3.6	0.5	–4.2
11.0	10.5	10.3	9.5	8.3	5.4	2.3	–2.4
17.3	16.8	16.6	15.8	14.6	11.7	8.6	3.9

The HL estimate of Δ is

5.4.7 Exact Confidence Interval Procedure for Shift Parameter

Let X and Y be continuous distributions satisfying a shift model with shift parameter Δ, and

$$D_{(1)} < D_{(2)} < \cdots < D_{(nm)}$$

be the ordered Walsh differences based on independent random samples of sizes n and m. Then

1. *Intervals:* The nm Walsh differences divide the real line into $(nm + 1)$ intervals:

   $$(-\infty, D_{(1)}),\ (D_{(1)}, D_{(2)}),\ \ \ldots,\ (D_{(nm-1)}, D_{(nm)}),\ (D_{(nm)}, \infty),$$

   where the endpoints are ignored.

2. *Mann-Whitney Probabilities:* The probability that Δ lies in a given interval follows the distribution of the Mann-Whitney statistic for the X sample.

   Specifically, if we let $D_{(0)} = -\infty$ and $D_{(nm+1)} = \infty$ for convenience, then

   $$P(D_{(k)} < \Delta < D_{(k+1)}) = P(U_1 = k) = p(k), \quad \text{for} \quad k = 0, 1, 2, \ldots, nm,$$

   as illustrated below.

Note that the main ideas needed to use Mann-Whitney probabilities are the following:

(a) If X and Y satisfy a shift model model with shift parameter Δ, then the samples

$$\begin{array}{ll} \text{Sample 1:} & X_1 - \Delta, X_2 - \Delta, \ldots, X_n - \Delta \\ \text{Sample 2:} & Y_1, Y_2, \ldots, Y_m \end{array}$$

are independent random samples from the same distribution. Thus, the distribution of

$$U_1 = \#(X_i - \Delta > Y_j) = \#(X_i - Y_j > \Delta)$$

can be tabulated assuming that each assignment of n observations to the first sample, with the remaining m observations being assigned to the second sample, are equally likely.

(b) Under the assumptions of part (a), the U_1 distribution is symmetric around $\frac{nm}{2}$.

Note, also, that these facts can be used to prove the following theorem.

Shift Parameter Confidence Interval Theorem. Under the assumptions above, if k is chosen so that the null probability

$$P(U_1 \leq k-1) \;=\; \frac{\alpha}{2}\;,$$

then $\left[D_{(k)}, D_{(nm-k+1)}\right]$ is a $100(1-\alpha)\%$ confidence interval for Δ.

Exercise. Assume that $n = 4$, $m = 6$, and the data are as follows:

$$4.1,\ 12.5,\ 12.9,\ 13.9 \qquad \text{and} \qquad 10.6,\ 12.2,\ 15.5,\ 16.7,\ 17.0,\ 20.6.$$

Assume these data are the values of independent random samples from continuous distributions satisfying a shift model, with $\Delta = \text{Median}(X) - \text{Median}(Y)$.

(a) Find the HL estimate of Δ.

	10.6	12.2	15.5	16.7	17.0	20.6
4.1	–6.5	–8.1	–11.4	–12.6	–12.9	–16.5
12.5	1.9	0.3	–3.0	–4.2	–4.5	–8.1
12.9	2.3	0.7	–2.6	–3.8	–4.1	–7.7
13.9	3.3	1.7	–1.6	–2.8	–3.1	–6.7

(*b*) Find a 90% (as close as possible) confidence interval for Δ. State the exact confidence level.

$p(0) = 0.0048$	$d_{(1)} =$ _____
$p(1) = 0.0048$	$d_{(2)} =$ _____
$p(2) = 0.0095$	$d_{(3)} =$ _____
$p(3) = 0.0143$	$d_{(4)} =$ _____
$p(4) = 0.0238$	$d_{(5)} =$ _____
$p(5) = 0.0286$	$d_{(6)} =$ _____
$p(6) = 0.0429$	$d_{(7)} =$ _____
$p(7) = 0.0476$	$d_{(8)} =$ _____
	$d_{(17)} =$ _____
$p(17) = 0.0476$	$d_{(18)} =$ _____
$p(18) = 0.0429$	$d_{(19)} =$ _____
$p(19) = 0.0286$	$d_{(20)} =$ _____
$p(20) = 0.0238$	$d_{(21)} =$ _____
$p(21) = 0.0143$	$d_{(22)} =$ _____
$p(22) = 0.0095$	$d_{(23)} =$ _____
$p(23) = 0.0048$	$d_{(24)} =$ _____
$p(24) = 0.0048$	

5.5 Sampling Models

The methods of this chapter assume that the measurements under study are the values of independent random samples from continuous distributions.

In most applications, simple random samples of individuals are drawn from finite populations and measurements are made on these individuals. If population sizes are large enough, then the resulting measurements can be treated as if they were the values of independent random samples. (Recall that a *simple random sample* of size n from the population of size N is a subset of n individuals chosen in such a way that each choice of subset is equally likely.)

5.5.1 Population Model

If simple random samples are drawn from sufficiently large populations of individuals, then sampling is said to be done under a *population model*. Under a population model, measurements can be treated as if they were the values of independent random samples.

When comparing two distributions, sampling can be done in many different ways, including:

1. *Sampling from Separate Subpopulations:* Individuals can be sampled from separate subpopulations. For example, a researcher interested in comparing achievement test scores of girls and boys in the fifth grade might sample separately from the subpopulations of fifth-grade girls and fifth-grade boys.

2. *Sampling from a Total Population Followed by Splitting:* Individuals can be sampled from a total population and then separated. For example, the researcher interested in comparing achievement scores might sample from the total population of fifth graders, and then split the sample into subsamples of girls and boys.

3. *Sampling from a Total Population Followed by Randomization:* Individuals can be sampled from a total population and then *randomized* to one of two treatments. For example, a medical researcher interested in determining if a new treatment to reduce serum cholesterol levels is more effective than the standard treatment in a population of women with very high levels of cholesterol might do the following:

   (a) Choose a simple random sample of $n + m$ subjects from the population of women with very high levels of serum cholesterol.
   (b) Partition the $n + m$ subjects into distinguishable subsets (or groups) of sizes n and m.
   (c) Administer the standard treatment to each subject in the first group for a fixed period of time, and the new treatment to each subject in the second group for the same fixed period of time.

   By randomly assigning subjects to treatment groups, the effect is as if sampling was done from two subpopulations: the subpopulation of women with high cholesterol who have been treated with the standard treatment for a fixed period of time, and the subpopulation of women with high cholesterol who have been treated with the new treatment for a fixed period of time. Note that, by design, the subpopulations differ in treatment only.

5.5.2 Randomization Model

The following is a common research scenario:

> A researcher is interested in comparing two treatments, and has $n+m$ subjects willing to participate in a study. The researcher randomly assigns n subjects to receive the first treatment; the remaining m subjects will receive the second treatment.

Treatments could be competing drugs for reducing cholesterol (as above), or competing methods for teaching multivariable calculus.

If the $n + m$ subjects are *not* a simple random sample from the study population, but the assignment of subjects to treatments is one of $\binom{n+m}{n}$ equally likely assignments, then sampling is said to be done under a *randomization model*.

Under a randomization model for the comparison of treatments, chance enters into the experiment only through the assignment of subjects to treatments. The results of experiments conducted under a randomization model cannot be generalized to a larger population of interest, but may still be of interest to researchers.

The Wilcoxon rank sum test is an example of a method that can be used to analyze data sampled under either the population model or the randomization model.