prepared by Professor Jenny Baglivo

2 MATH448001 Notebook 2

This notebook introduces nonparametric, permutation and bootstrap methods. Each method is important in modern applications of statistics. The notes include material from the following chapters of the Rice textbook: Chapter 3 (order statistics), Chapter 10 (summarizing data) and Chapter 11 (comparing two samples).

2.1 Order Statistics and Quantiles

This section reviews material about order statistics (from Chapter 3 of the Rice textbook), and their relationship to quantiles of continuous distributions.

2.1.1 *K*th Order Statistics and Their Distributions

Let X_1, X_2, ..., X_n be a random sample from the continuous distribution

whose PDF is $f(x)$, and whose CDF is $F(x) = P(X \leq x)$, for all real numbers x,

and let k be an integer between 1 and n: $k \in \{1, 2, \ldots, n\}$.

1. *Kth Order Statistic:* The k^{th} *order statistic*, $X_{(k)}$, is the k^{th} observation in order:

   $$X_{(k)} \text{ is the } k^{\text{th}} \text{ smallest of } X_1, X_2, \ldots, X_n.$$

2. *Sample Maximum/Minimum:* The largest observation, $X_{(n)}$, is called the *sample maximum* and the smallest observation, $X_{(1)}$, is called the *sample minimum.*

3. *Sample Median:* The *sample median* is the middle order statistic when n is odd, and the average of the two middle order statistics when n is even:

$$\textit{Sample Median} = \begin{cases} X_{\left(\frac{n+1}{2}\right)} & \text{when } n \text{ is odd} \\ \frac{1}{2}\left(X_{\left(\frac{n}{2}\right)} + X_{\left(\frac{n}{2}+1\right)}\right) & \text{when } n \text{ is even} \end{cases}$$

For example, if the following numbers were observed

18.4, 27.2, 37.0, 8.6, 2.5, 12.8, 34.9, 1.7, 23.8, 22.4, 0.1, 23.8, 25.7, 13.5

then the sample minimum is _____, the sample maximum is _____,

and the sample median is _____.

Probability distributions. Let $X_{(k)}$ be the k^{th} order statistic of a random sample of size n from a continuous distribution with PDF $f(x)$ and CDF $F(x)$, and let $k \in \{1, 2, \ldots, n\}$. Then

1. *CDF:* The cumulative distribution function of $X_{(k)}$ has the following form:

$$F_{(k)}(x) = \sum_{j=k}^{n} \binom{n}{j} (F(x))^j (1 - F(x))^{n-j}, \quad \text{for all real numbers } x.$$

2. *PDF:* The probability density function of $X_{(k)}$ has the following form:

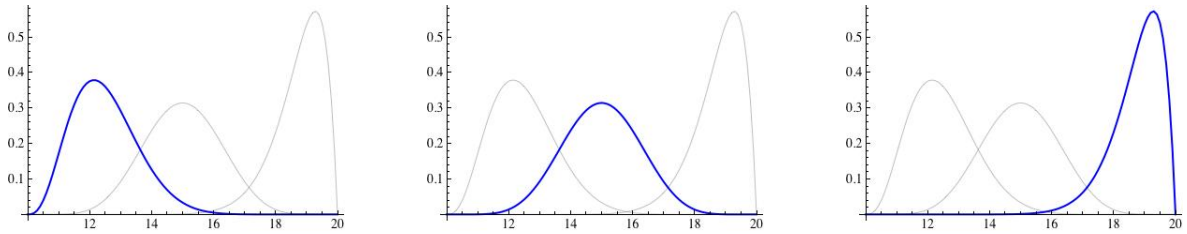
$$f_{(k)}(x) = \frac{d}{dx} F_{(k)}(x) = \binom{n}{k-1, 1, n-k} (F(x))^{k-1} f(x) (1 - F(x))^{n-k}$$

   (after simplification), whenever the derivative exists.

To demonstrate that the formula for $F_{(k)}(x) = P(X_{(k)} \leq x)$ is correct, first note that the event that "$X_{(k)} \leq x$" is equivalent to the event that "k or more of the X_i's are $\leq x$".

Now (complete the demonstration),

For example, the following plots show the density functions of the 4^{th}, 8^{th}, and 14^{th} order statistics of a random sample of size 15 from a continuous uniform distribution on $[10, 20]$:



Each distribution "focuses" on a different part of the range of the random variable.

Exercise. Let X be a continuous uniform random variable on $[a, b]$, where $a < b$ are constants. Find a simplified general formula for the density function of the k^{th} order statistic of a random sample of size n from the X distribution.

2.1.2 Approximate Mean and Variance; Probability Plots

Let $X_{(k)}$ be the k^{th} order statistic of a random sample of size n from a continuous distribution with PDF $f(x)$ and CDF $F(x)$, and assume that p^{th} quantiles exist for $p \in (0,1)$. Further, let

$$\theta = F^{-1}(p) \text{ for } p = k/(n+1).$$

Then the following theorem can be used to find approximate summaries of the $X_{(k)}$ distribution.

> ***Theorem (Summary Measures for* Kth *Order Statistic).*** Under the conditions above,
>
> $$E(X_{(k)}) \approx \theta \text{ and } Var(X_{(k)}) \approx \frac{p(1-p)}{(n+2)(f(\theta))^2} \text{ as long as } f(\theta) \neq 0.$$

> ***Corollary (Uniform Distributions).*** If X is a continuous uniform random variable, then the formulas given in the theorem are exact.

Note: Given a random sample of size n, the collection $\{E(X_{(k)}) \mid k = 1, 2, \ldots, n\}$ divide the real line into $(n+1)$ intervals, as illustrated below.

The theorem tells us that the $(n+1)$ intervals are approximately equally likely. That is, the theorem tells us that

$$P(X \in \text{The } i^{\text{th}} \text{ Interval}) \approx \frac{1}{n+1}, \text{ for } i = 1, 2, \ldots, n+1.$$

To illustrate the theorem for uniform distributions, let $[a, b] = [10, 20]$ and assume that $n = 15$. Summary measures for the 4^{th}, 8^{th}, and 14^{th} order statistics are given in the following table:

	$E(X_{(k)})$	$Var(X_{(k)})$
$k = 4$	$10 + 10\left(\frac{1}{4}\right) = 12.50$	$\frac{(1/4)(3/4)}{(17)(1/10)^2} \approx 1.103$
$k = 8$	$10 + 10\left(\frac{1}{2}\right) = 15.00$	$\frac{(1/2)(1/2)}{(17)(1/10)^2} \approx 1.471$
$k = 14$	$10 + 10\left(\frac{7}{8}\right) = 18.75$	$\frac{(7/8)(1/8)}{(17)(1/10)^2} \approx 0.643$

Probability plots. Let X be a continuous random variable, and $x_{(1)},\ x_{(2)},\ \ldots,\ x_{(n)}$ be the observed values of the order statistics of a random sample of size n from the X distribution.

A *probability plot* is a plot of pairs of the form:

$$\left(\left(\frac{k}{n+1} \right)^{\text{st}} \text{ model quantile, } x_{(k)} \right), \quad k = 1, 2, \ldots, n \ .$$

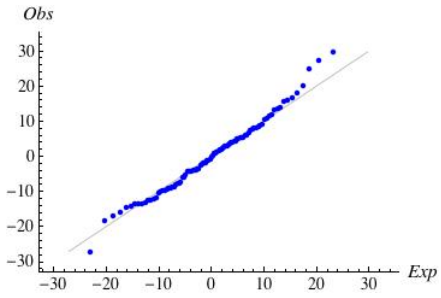
The theorem above tells us that the ordered pairs in a probability plot should lie roughly on the line $y = x$.

For example, I used the computer to generate a pseudo-random sample of size 95 from the normal distribution with mean 0 and standard deviation 10.

1. *Probability Plot:* The plot on the right is a probability plot of pairs of the form

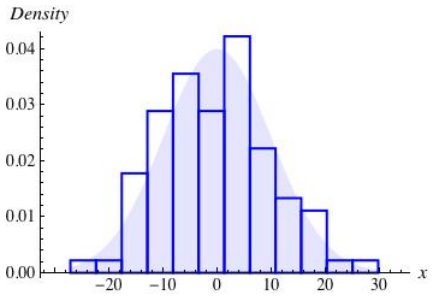
   $$\left(\left(\frac{k}{96} \right)^{\text{th}} \text{ model quantile, } x_{(k)} \right)$$

   for $k = 1, 2, \ldots, 95$. In the plot, the Observed order statistic (vertical axis) is plotted against its approximate Expected value (horizontal axis).



2. *Comparison Plot:* The plot on the right shows an empirical histogram of the same sample, superimposed on the density curve for a normal distribution with mean 0 and standard deviation 10.

   Twelve subintervals of equal length were used to construct the empirical histogram.



Footnotes: If n is large, then both plots give good graphical comparisons of model and data.

But, if n is small to moderate, then the probability plot may be a better way to compare model and data since the shape of the empirical histogram may be very different from the shape of the density function of the continuous model.

2.1.3 Exact Confidence Interval Procedure for Quantiles

Let X be a continuous random variable, and let θ be the p^{th} quantile of the X distribution, for a fixed proportion $p \in (0, 1)$.

Let $X_{(k)}$ be the k^{th} order statistic of a random sample of size n from the X distribution. Then

1. *Intervals:* The n order statistics divide the real line into $(n + 1)$ intervals:

$$(-\infty, X_{(1)}), \quad (X_{(1)}, X_{(2)}), \quad \ldots, \quad (X_{(n-1)}, X_{(n)}), \quad (X_{(n)}, \infty)$$

   (ignoring the endpoints).

2. *Binomial Probabilities:* The probability that θ lies in a given interval follows a binomial distribution with parameters n and p. Specifically,

   (a) *First Interval:* The event "$\theta \in (-\infty, X_{(1)})$" is equivalent to the event that all X_i's are greater than θ. Thus,

$$P(\theta \in (-\infty, X_{(1)})) = (1 - p)^n.$$

   (b) *Middle Intervals:* The event "$\theta \in (X_{(k)}, X_{(k+1)})$" is equivalent to the event that exactly k X_i's are less than θ. Thus,

$$P(\theta \in (X_{(k)}, X_{(k+1)})) = \binom{n}{k} p^k (1 - p)^{n-k}.$$

   (c) *Last Interval:* The event "$\theta \in (X_{(n)}, \infty)$" is equivalent to the event that all X_i's are less than θ. Thus,

$$P(\theta \in (X_{(n)}, \infty)) = p^n.$$

Let $p(k) = \binom{n}{k} p^k (1 - p)^{n-k}$, $k = 0, 1, \ldots, n$, be the binomial probabilities.

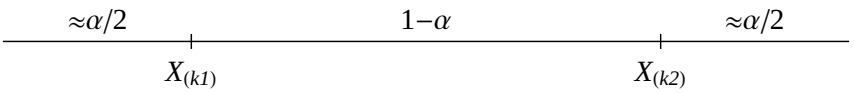
Then the following graphic illustrates the probabilities associated with each subinterval.

Further, the facts above can be used to construct a confidence interval procedure for quantiles.

Quantile Confidence Interval Theorem. Under the conditions above, if indices k_1 and k_2 are chosen so that

$$\begin{aligned} P(\theta < X_{(k_1)}) &= \textstyle\sum_{j=0}^{k_1-1} \binom{n}{j} p^j (1-p)^{n-j} = \alpha/2 \\ P(X_{(k_1)} < \theta < X_{(k_2)}) &= \textstyle\sum_{j=k_1}^{k_2-1} \binom{n}{j} p^j (1-p)^{n-j} = 1-\alpha \\ P(\theta > X_{(k_2)}) &= \textstyle\sum_{j=k_2}^{n} \binom{n}{j} p^j (1-p)^{n-j} = \alpha/2, \end{aligned}$$

then the interval $\left[X_{(k_1)}, X_{(k_2)}\right]$ is a $100(1-\alpha)\%$ confidence interval for θ.



Note that, in practice, k_1 and k_2 are chosen to make the sums in the theorem as close as possible to the values shown on the right.

Exercise (Source: Sheaffer et al, 1996). The following table shows the total yearly rainfall (in inches) for Los Angeles in the 10-year period from the beginning of 1983 to the end of 1992.

	1983	*1984*	*1985*	*1986*	*1987*	*1988*	*1989*	*1990*	*1991*	*1992*
Rainfall	34.04	8.90	8.92	18.00	9.11	11.57	4.56	6.49	15.07	22.56

Assume these data are the values of a random sample from a continuous distribution.

Construct a 90% (or as close as possible) confidence interval for the 40^{th} percentile of the rainfalls distribution. State the exact confidence level.

$p(0) = 0.0060$	
	$x_{(1)} =$ _____
$p(1) = 0.0403$	
	$x_{(2)} =$ _____
$p(2) = 0.1209$	
	$x_{(3)} =$ _____
$p(3) = 0.2150$	
	$x_{(4)} =$ _____
$p(4) = 0.2508$	
	$x_{(5)} =$ _____
$p(5) = 0.2007$	
	$x_{(6)} =$ _____
$p(6) = 0.1115$	
	$x_{(7)} =$ _____
$p(7) = 0.0425$	
	$x_{(8)} =$ _____
$p(8) = 0.0106$	
	$x_{(9)} =$ _____
$p(9) = 0.0016$	
	$x_{(10)} =$ _____
$p(10) = 0.0001$	

Example (Sample Size 65): Suppose we are interested in constructing confidence intervals for the 25^{th}, 50^{th} and 75^{th} percentiles of a continuous distribution with 95% (or slightly higher) coverage using a sample of size 65 from the distribution.

Then, the values of k_1 and k_2 and the exact confidence levels are given below:

	k_1	k_2	Confidence level
25^{th} percentile	10	24	0.956
50^{th} percentile	25	41	0.954
75^{th} percentile	42	56	0.956

Application (Hand et al., 1994): Cholesterol and triglycerides belong to the class of chemicals known as lipids (fats). As part of a study to determine the relationship between high levels of lipids and coronary artery disease, researchers measured plasma levels of cholesterol and triglycerides in milligrams per deciliter (mg/dL) in 371 men complaining of chest pain.

The cholesterol levels for a random subset of 65 men who complained of chest pain and who showed evidence of heart disease are given below:

138	140	150	151	157	159	168	169	172	174	175	176	177
184	185	186	187	190	192	193	194	196	197	199	200	201
204	209	211	212	213	214	215	216	218	220	221	223	225
226	236	238	239	243	246	247	249	251	254	256	257	258
264	265	267	269	279	280	287	294	297	299	308	313	332

Based on these data,

1. A 95.6% confidence interval for the 25^{th} percentile of the cholesterol distribution is

   [_____ , _____]

2. A 95.4% confidence interval for the 50^{th} percentile of the cholesterol distribution is

   [_____ , _____]

3. A 95.6% confidence interval for the 75^{th} percentile of the cholesterol distribution is

   [_____ , _____]

2.1.4 Sample Quantiles; Box Plots

Let X be a continuous random variable, let θ_p be the p^{th} quantile of the X distribution, and let $X_{(k)}$ be the k^{th} order statistic of a random sample of size n , for $k = 1, 2, \ldots, n$.

Sample Quantiles: Given $p \in \left[\frac{1}{n+1}, \frac{n}{n+1}\right]$, the p^{th} *sample quantile* is defined as follows:

1. if $p = \frac{k}{n+1}$ for some k, then $\widehat{\theta_p} = X_{(k)}$;
2. if $p \in \left(\frac{k}{n+1}, \frac{k+1}{n+1}\right)$ for some k, then $\widehat{\theta_p} = X_{(k)} + ((n+1)p - k)\left(X_{(k+1)} - X_{(k)}\right)$.

   With this definition, the point $(\widehat{\theta_p}, p)$ is on the piecewise linear curve connecting the successive points

$$\left(X_{(1)}, \frac{1}{n+1}\right), \left(X_{(2)}, \frac{2}{n+1}\right), \ldots, \left(X_{(n)}, \frac{n}{n+1}\right).$$

For example, suppose that the following 12 numbers were observed (and ordered):

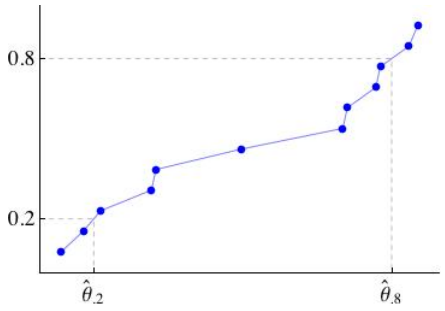
$$2.2,\ 3.2,\ 4.0,\ 6.3,\ 6.5,\ 10.4,\ 15.0,\ 15.2,\ 16.5,\ 16.7,\ 18.0,\ 18.4.$$

The plot on the right shows the pairs:

$$\left(2.2, \frac{1}{13}\right), \left(3.2, \frac{2}{13}\right), \left(4.0, \frac{3}{13}\right), \ldots$$

with successive pairs connected by line segments.

The locations of the sample 20^{th} and 80^{th} percentiles have been labeled. The values of these sample quantiles are (please complete)



Sample Quartiles: The *sample quartiles* are the estimates of the 25^{th}, 50^{th}, and 75^{th} percentiles:

$$q_1 = \widehat{\theta_{0.25}},\ q_2 = \widehat{\theta_{0.50}},\ q_3 = \widehat{\theta_{0.75}}.$$

The *sample median* is q_2, and the *sample interquartile range* is the difference $q_3 - q_1$.

Continuing with the example above, the sample median and sample interquartile range are (please complete)

Box Plots: A *box plot* is a graphical display of a data set that shows the sample median, the sample interquartile range, and the presence of possible outliers (numbers that are far from the center). Box plots were introduced by John Tukey in the 1970's.

Let q_1, q_2 and q_3 be the sample quartiles. To construct a box plot:

1. *Box:* A box is drawn from q_1 to q_3.
2. *Bar:* A bar is drawn at the sample median, q_2.
3. *Whiskers:* A whisker is drawn from q_3 to the largest observation that is less than or equal to $q_3 + 1.50(q_3 - q_1)$. Another whisker is drawn from q_1 to the smallest observation that is greater than or equal to $q_1 - 1.50(q_3 - q_1)$.
4. *Outliers:* Observations outside the interval

   $$[q_1 - 1.50(q_3 - q_1), q_3 + 1.50(q_3 - q_1)]$$

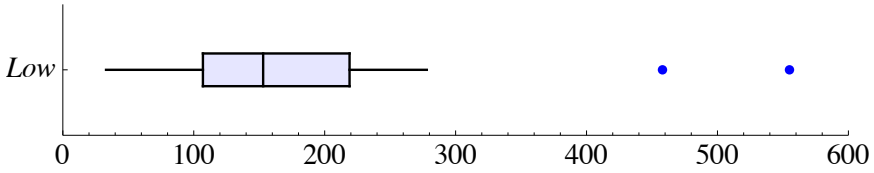
   are drawn as separate points. These observations are called the *outliers*.

Exercise (Source: Rice textbook, Chapter 10). As part of a study on the effects of an infectious disease on the lifetimes of guinea pigs, more than 400 animals were infected.

The data below are the lifetimes (in days) of 45 animals given a *low* exposure to the disease:

33	44	56	59	74	77	93	100	102	105	107	107	108	108	109
115	120	122	124	136	139	144	153	159	160	163	163	168	171	172
195	202	215	216	222	230	231	240	245	251	253	254	278	458	555

A box plot for these data is shown below:



(*a*) Find the sample quartiles and sample interquartile range.

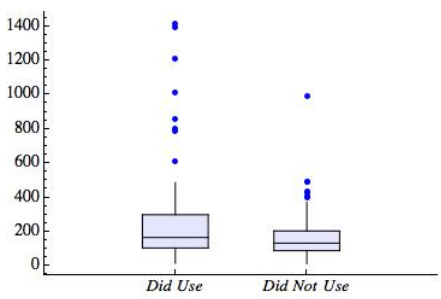
(*b*) Find the interval $[q_1 - 1.50(q_3 - q_1), q_3 + 1.50(q_3 - q_1)]$ and identify the outliers.

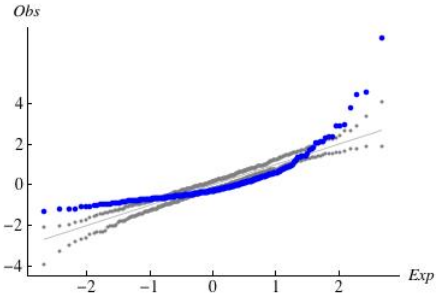
Example (Source: Stukel, 1998, FTP:lib.stat.cmu.edu/datasets/): Several studies have suggested that low plasma concentrations of beta-carotene (a precursor of vitamin A) may be associated with increased risk of certain types of cancer. As part of a study to investigate the relationship between personal characteristics (including diet) and levels of beta-carotene in the blood, measurements were made on over 300 subjects.

This example compares the plasma levels of beta-carotene in nanograms per milliliter (ng/mL) of the 108 women who regularly used vitamin supplements to the levels for the 163 women who did not use supplements regularly. Data summaries are as follows:

	Sample Size	*Mean, Standard Deviation*	*Median, Interquartile Range*
Did Use (X)	$n = 108$	$\overline{x} = 250.444$, $s_x = 255.774$	$q_2 = 166.0$, $q_3 - q_1 = 194.25$
Did Not Use (Y)	$m = 163$	$\overline{y} = 162.11$, $s_y = 115.11$	$q_2 = 133.0$, $q_3 - q_1 = 115.0$

The *left plot* below shows side-by-side box plots for the two groups, and the *right plot* is an enhanced normal probability plot of standardized residuals.





To construct the plot on the right:

1. The normal probability plot of standardized values is constructed as follows:
   (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 271 ordered standardized values (vertical axis; observed) are plotted against the $k/272^{\text{nd}}$ quantiles of the standard normal distribution (horizontal axis; expected).
2. The plot is enhanced to include the results of 100 simulations from the standard normal distribution: For each $k = 1, 2, \ldots, 271$, the minimum and maximum value of the 100 simulated k^{th} order statistics are plotted.

The side-by-side box plots are a good visual comparison of the data; the plot suggests that the distributions from which the data were drawn are different. The enhanced normal probability plot suggests that we would not use normal theory methods to analyze differences in the distributions.

2.2 Parametric & Nonparametric Methods for Two Sample Analysis

This section reviews parametric methods for comparing the means of normal distributions with equal variances, and introduces a nonparametric method for comparing general distributions. In each case, the methods use independent random samples from the distributions.

2.2.1 Parametric Methods; Review of Pooled t Methods

Statistical methods that require strong assumptions about the shapes of distributions and answer questions about parameter values are called *parametric methods.*

A useful general example of parametric methods involves comparing two normal distributions with equal variances using independent random samples.

Example: Pooled t methods. Suppose that we are interested in comparing the means of two normal distributions using independent random samples, and that we are willing to assume that the variances of the two distributions are equal. Specifically,

1. *X Sample:* Let $X_1, X_2, \ldots, X_n$ be a random sample from a normal distribution with mean μ_x and standard deviation σ.
2. *Y Sample:* Let $Y_1, Y_2, \ldots, Y_m$ be an independently chosen random sample from a normal distribution with mean μ_y and standard deviation σ.

Since the samples were chosen independently, we know that the difference in sample means is a normal random variable with the following summary measures:

$$E(\overline{X} - \overline{Y}) = \mu_x - \mu_y \text{ and } Var(\overline{X} - \overline{Y}) = \frac{\sigma^2}{n} + \frac{\sigma^2}{m} = \sigma^2 \left(\frac{1}{n} + \frac{1}{m} \right).$$

Further, we can demonstrate that the statistic

$$T = \frac{(\overline{X} - \overline{Y}) - (\mu_x - \mu_y)}{\sqrt{S_p^2 \left(\frac{1}{n} + \frac{1}{m} \right)}}, \text{ where } S_p^2 = \frac{(n-1)S_x^2 + (m-1)S_y^2}{(n+m-2)},$$

has a Student t distribution with $(n + m - 2)$ degrees of freedom.

Our strong assumptions about the distributions from which the data were drawn allow us to find the sampling distribution of a useful statistic, namely the statistic T.

Since we know the sampling distribution of T, we can determine if the means of the distributions (and hence the distributions themselves) are equal, and we can find a range of values within which we believe the true difference in means $(\mu_x - \mu_y)$ lies.

Exercise (Source: Chihara & Hesterberg, 2011): "Black spruce is a species of a slow-growing coniferous tree found across the northern part of North America. It is commonly found on wet organic soils. In a study conducted in the 1990s, a biologist interested in factors affecting the growth of the black spruce planted its seedlings on sites located in boreal peatlands in northern Manitoba, Canada."

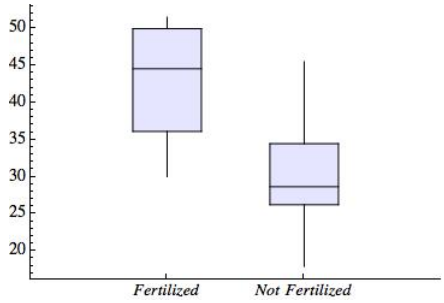
Thirty-six seedlings were randomly assigned to plots that were either fertilized or not. Heights in centimeters (cm) were measured at the beginning of the study period (at the time the seedlings were planted) and again after 5 years.

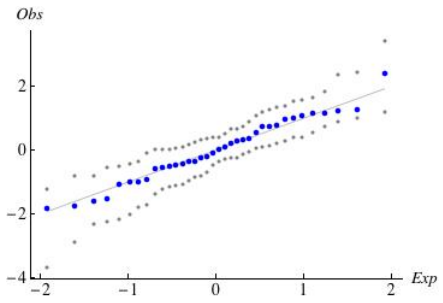
The following samples and summaries give the change in height over the 5-year period:

1. *Fertilized (X):*
   Change in height in cm for seedlings that were fertilized were: 30.0, 31.6, 35.3, 35.8, 36.2, 38.7, 39.6, 40.4, 44.4, 44.7, 45.0, 46.7, 48.0, 49.8, 50.4, 50.8, 51.0, 51.5
   (Sample summaries: $n = 18$, $\overline{x} = 42.772$, $s_x = 7.020$)

2. *Not Fertilized (Y):*
   Change in height in cm for seedlings not fertilized were: 17.9, 19.5, 22.9, 26.0, 26.3, 26.7, 27.1, 28.0, 28.3, 29.0, 29.6, 30.2, 32.0, 34.4, 34.6, 36.0, 38.0, 45.5
   (Sample summaries: $m = 18$, $\overline{y} = 29.556$, $s_y = 6.622$)

The side-by-side box plots shown on the *left* below suggest that the population means are different, and the enhanced normal probability plot of standardized residuals shown on the *right* below suggest that normal theory methods are reasonable for these data.





Assume these data are the values of independent random samples from normal distributions with equal variances.

(a) Find the pooled estimate of the common variance.

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

2.2.2 Nonparametric Methods; Stochastically Larger/Smaller

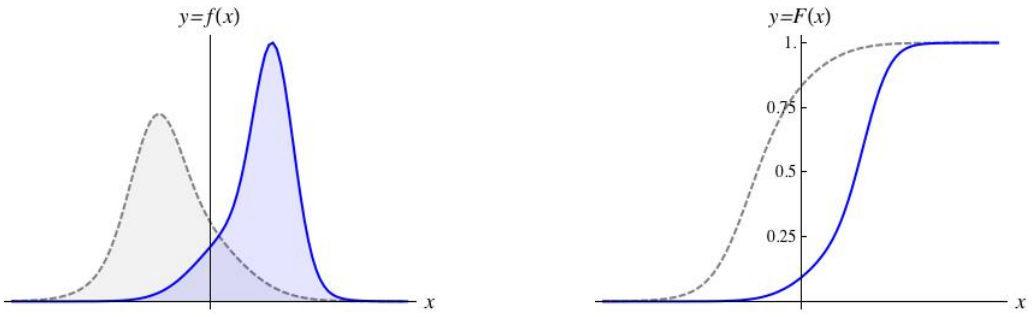
Parametric methods require strong assumptions about the shapes of the distributions from which the data were drawn. By contrast, *nonparametric methods* (also known as *distribution-free methods*) make mild assumptions, such as, "the distributions are continuous" or "the continuous distributions are symmetric around their centers."

Stochastically larger/smaller: 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 "≥") for at least one x.

To illustrate the definition of stochastically larger/smaller, consider the following plots of the density functions (*left* plot), and the cumulative distribution functions (*right* plot) of two 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

$$F_V(x) \leq F_W(x) \text{ for all } x,$$

with strict inequality for at least one x.

2.2.3 Wilcoxon Rank Sum Statistic, Distribution and Methods

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, 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 and 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 _____

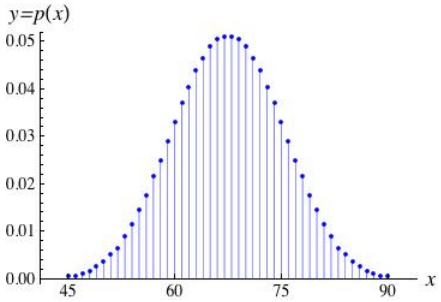
The observed value of R_2 is _____

Distribution of R_1. 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.

> ***Theorem (Rank Sum Distribution).*** 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.)

<u>*Note:*</u> The distribution theorem can be proven using combinatorics. Under the null hypothesis of equality of the X and Y distributions, each subset of size n from the collection of 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 R_1 for all values in the range of the random variable.



The range of the R_1 distribution is the integers between 45 and 90. Further,

- The mean of the R_1 distribution is _____.

- The variance of the R_1 distribution is _____.

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.)

Computer programs for rank sum tests usually report the observed significance level (p value). It is up to the user to interpret the p value.

Continuing with the example above, where $n = 9$ and $m = 5$, the table below gives probabilities and cumulative probabilities for the R_1 distribution for values of x in the range.

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			

Using this table,

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 85, 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 were 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, and the percentage of iron retained was calculated."

Results for the second concentration (1.2 millimolar) are reported below.

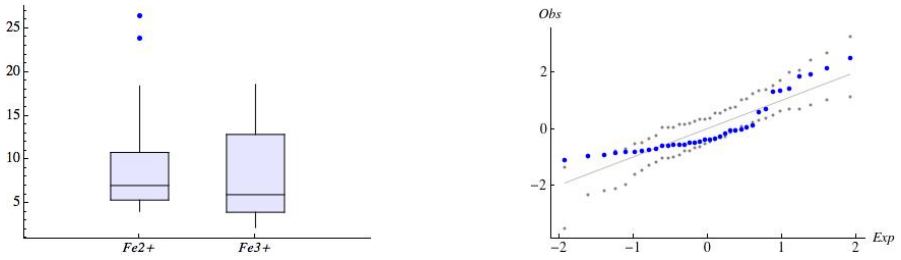
1. *X sample:* Percent retention for the 18 mice given Fe2+.

4.04, 4.16, 4.42, 4.93, 5.49, 5.77, 5.86, 6.28, 6.97, 7.06, 7.78, 9.23, 9.34, 9.91, 13.46, 18.4, 23.89, 26.39.

2. *Y sample:* Percent retention for the 18 mice given Fe3+.

2.20, 2.93, 3.08, 3.49, 4.11, 4.95, 5.16, 5.54, 5.68, 6.25, 7.25, 7.90, 8.85, 11.96, 15.54, 15.89, 18.30, 18.59.

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.



These plots suggest that the X and Y distributions are *not* approximately normal.

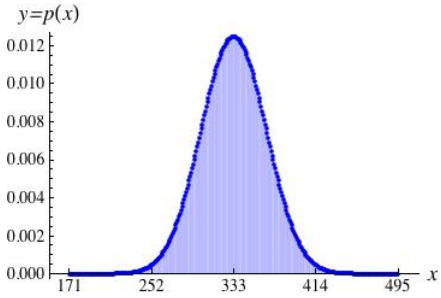
The equality of the X and Y distributions will be tested using the Wilcoxon rank sum test, a two-sided alternative, and 5% significance level.

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 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 when some observations are tied, 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:

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

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 at 0^oC in calories per gram of mass."

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

   79.97, 79.98, 80., 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 _____

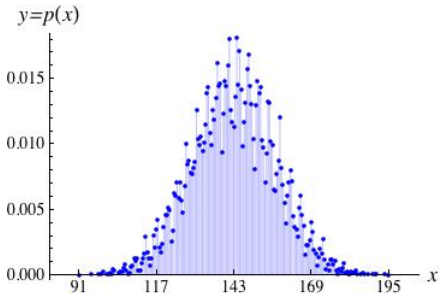
The equality of the X and Y distributions will be tested using the Wilcoxon rank sum test, a two-sided alternative, and 5% significance level.

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),



2.3 Parametric & Nonparametric Methods for Paired Samples Analysis

This section considers parametric and nonparametric methods for paired samples

$$\{\{x_1, y_1\}, \{x_2, y_2\}, \ldots, \{x_n, y_n\}\},$$

or corresponding lists of differences $\{d_1, d_2, \ldots, d_n\} = \{x_1 - y_1, x_2 - y_2, \ldots, x_n - y_n\}$. Paired samples are assumed to be the values of a random sample from a joint (X, Y) distribution.

Experimental settings. Paired samples arise in many experimental settings. For example, the observed data could be the result of a

1. *Before-and-After Experiment:*

   For each of n individuals, the x value is a measurement made before a treatment begins, and the y value is the corresponding measurement after a fixed treatment period.

2. *Randomized Pairs Experiment:*

   For each of n pairs of individuals, where members of each pair are matched on important factors (such as age, sex, severity of disease), one member of the pair is randomly assigned to treatment 1 and the other to treatment 2. After a fixed period of time, measurements are taken on each individual.

Paired samples are used to reduce variability. Researchers use paired samples designs in order to reduce the variability of the results. In paired designs, individuals within each pair are expected to respond similarly to treatment, while individuals in different pairs are expected to respond differently to treatment.

To understand this idea better, let X be a random variable with mean μ_x and standard deviation σ_x, let Y be a random variable with mean μ_y and standard deviation σ_y, and suppose that $\rho = Corr(X, Y) > 0$. Then $E(X - Y) = \mu_x - \mu_y$, and

$$Var(X - Y) = \sigma_x^2 + \sigma_y^2 - 2\rho\sigma_x\sigma_y$$

is smaller than the variance of the difference if X and Y were chosen independently.

Exercise. Use properties of expectation to demonstrate that the variance formula is correct.

2.3.1 Paired t Methods

Assume that $D = X - Y$ is a normal random variable. Since

$$E(D) = E(X) - E(Y),$$

the differences data from a paired samples experiment can be used to answer questions about the difference in means of the X and Y distributions. Further, we can use one sample t methods to analyze the data.

Note: In this setting, one sample t methods are called *paired t methods.*

Exercise (FTP: lib.stat.cmu.edu/DASL/). The table below gives the proportion of women in the labor force in 1972 (x) and 1968 (y), and the difference in proportions ($d = x - y$), for women living in 19 U.S. cities.

City	x	y	d	City	x	y	d
Baltimore	0.57	0.49	0.08	Minneapolis/St.Paul	0.59	0.50	0.09
Boston	0.60	0.45	0.15	Newark	0.53	0.54	−0.01
Buffalo	0.64	0.58	0.06	New York	0.45	0.42	0.03
Chicago	0.52	0.52	0.00	Patterson	0.57	0.56	0.01
Cinncinati	0.53	0.51	0.02	Philadelphia	0.45	0.45	0.00
Connecticut	0.55	0.54	0.01	Pittsburg	0.49	0.34	0.15
Dallas	0.64	0.63	0.01	San Francisco	0.55	0.55	0.00
Detroit	0.46	0.43	0.03	St. Louis	0.35	0.45	−0.10
Houston	0.50	0.49	0.01	Washington,D.C.	0.52	0.42	0.10
Los Angeles	0.50	0.50	0.00				

(*Summaries:* $n = 19$, $\overline{d} = 0.03368$, $s_d^2 = 0.05974^2$.)

Assume the differences data are the values of a random sample from a normal distribution. Construct and interpret a 95% confidence interval for the difference in means of the proportions of women in the labor force in 1972 and 1968, respectively.

Important footnote to this exercise: Consider data summaries by year:

$$\text{1972 labor force participation:} \quad n = 19,\ \overline{x} = 0.527,\ s_x^2 = 0.071^2$$

$$\text{1968 labor force participation:} \quad n = 19,\ \overline{y} = 0.493,\ s_y^2 = 0.068^2$$

If these numbers had been based on independent random samples, then a 95% confidence interval for the difference in means based on pooled t methods is $[-0.012, 0.079]$.

Since the lists of numbers for the two years are quite close (in fact, the observed correlation between the x and y samples is about 0.63), there would be insufficient evidence to conclude that the mean changed in the four-year period.

2.3.2 Wilcoxon Signed Ranks Statistic, Distribution and Methods

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 in the paired samples setting.

Given a random sample of size n from the joint (X, Y) distribution, and the corresponding list of differences of the form

$$D_i = X_i - Y_i, \quad \text{for} \quad i = 1, 2, \ldots, n,$$

Wilcoxon *signed ranks statistics* for positive differences (W_+) and for negative differences (W_-) are computed as follows:

1. Sort the list of absolute differences.
2. Replace each observed absolute difference by its rank (or position) in the sorted list.
3. Let W_+ equal the sum of the ranks for positive differences and let W_- equal the sum of the ranks for negative differences.

<u>*Note*</u> that, since

$$W_+ + W_- = 1 + 2 + \cdots + n = \frac{n(n+1)}{2},$$

tests based on W_+ are equivalent to tests based on W_-. We will use the W_+ statistic.

For example, if $n = 10$ and the differences data are as follows:

$$-3.54,\ -3.05,\ -0.68,\ 0.65,\ 1.66,\ 2.16,\ 2.75,\ 3.23,\ 4.24,\ 5.15$$

then the ordered list of absolute differences, and corresponding ranks are as follows:

0.65	0.68	1.66	2.16	2.75	3.05	3.23	3.54	4.24	5.15
1	2	3	4	5	6	7	8	9	10

The observed value of W_+ is _____

The observed value of W_- is _____

Distribution of W_+. The following theorem gives us information about the sampling distribution of W_+ under the null hypothesis that the X and Y distributions are equal.

> ***Theorem (W_+ Statistic Distribution).*** Let X and Y be continuous distributions, and W_+ be the Wilcoxon signed ranks statistic for positive differences based on a random sample of size n from the joint (X, Y) distribution. If the X and Y distributions are equal, then
>
> 1. The range of W_+ is $0, 1, 2, \ldots, n(n+1)/2$.
> 2. $E(W_+) = n(n+1)/4$ and $Var(W_+) = n(n+1)(2n+1)/24$.
> 3. The distribution of W_+ is symmetric around its mean. In particular,
>
> $$P(W_+ = x) = P\left(W_+ = \frac{n(n+1)}{2} - x\right) \quad \text{for all } x.$$
>
> 4. If n is large, then the distribution of W_+ is approximately normal.

<u>*Note:*</u> The proof of the distribution theorem uses combinatorics:

1. Under the null hypothesis, the distribution of $D = X - Y$ is symmetric around 0. Thus,

$$P(D_i > 0) = P(D_i < 0) = \frac{1}{2} \qquad i = 1, 2, \ldots, n$$

   where $D_i = X_i - Y_i$.

2. Since the pairs were chosen independently, the 2^n events of the form

$$\begin{Bmatrix} D_1 > 0 \\ D_1 < 0 \end{Bmatrix} \text{ and } \begin{Bmatrix} D_2 > 0 \\ D_2 < 0 \end{Bmatrix} \text{ and } \cdots \text{ and } \begin{Bmatrix} D_n > 0 \\ D_n < 0 \end{Bmatrix}$$

   are equally likely, and each choice of signs for the observed differences is equally likely.

3. Thus, each subset of ranks associated with positive differences is equally likely.

Finding p values. Large values of W_+ support the alternative that the X distribution is stochastically larger than the Y distribution, while small values support the alternative that Y is stochastically larger than X. Thus, the computation of p values in the paired samples setting follows the same pattern we used in the two sample setting with the R_1 statistic.

Handling tied differences. Tied absolute differences in the paired samples setting are handled similarly to the two sample setting.

For example, suppose the differences data from above are altered as follows:

$$-3.54,\ -3.05,\ -0.65,\ 0.65,\ 1.66,\ 2.16,\ 2.75,\ 3.23,\ 4.24,\ 5.15$$

Since the two smallest absolute differences are equal, the average of their ranks (that is, the *midrank* of the two observations) is assigned to each. This leads to the following table of absolute differences and corresponding midranks:

0.65	0.65	1.66	2.16	2.75	3.05	3.23	3.54	4.24	5.15
1.5	1.5	3	4	5	6	7	8	9	10

The sampling distribution of W_+ is now obtained by considering all $2^{10} = 1024$ subsets chosen from the set of slips

$$\boxed{1.5}\ \boxed{1.5}\ \boxed{3}\ \boxed{4}\ \boxed{5}\ \boxed{6}\ \boxed{7}\ \boxed{8}\ \boxed{9}\ \boxed{10}$$

Interpreting zero differences. Since $+0 = -0 = 0$, a difference of zero will effectively reduce the sample size since the midranks associated with the zero differences can never be assigned to either W_+ or W_-.

For example, consider again the labor force participation example from page 26. The following table lists the differences, ordered by their absolute values, and midranks for the absolute differences:

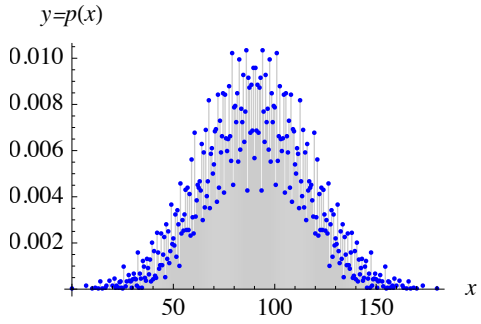
	d_i	$\lvert d_i\rvert$	Midrank		d_i	$\lvert d_i\rvert$	Midrank		d_i	$\lvert d_i\rvert$	Midrank
1	0.00	0.00	2.5	8	+0.01	0.01	7.0	15	+0.09	0.09	15.0
2	0.00	0.00	2.5	9	−0.01	0.01	7.0	16	−0.10	0.10	16.5
3	0.00	0.00	2.5	10	+0.02	0.02	10.0	17	+0.10	0.10	16.5
4	0.00	0.00	2.5	11	+0.03	0.03	11.5	18	+0.15	0.15	18.5
5	+0.01	0.01	7.0	12	+0.03	0.03	11.5	19	+0.15	0.15	18.5
6	+0.01	0.01	7.0	13	+0.06	0.06	13.0				
7	+0.01	0.01	7.0	14	+0.08	0.08	14.0				

The observed value of W_+ is _____

The observed value of W_- is _____

To determine if the labor force participation data suggest a difference in participation of women in the years 1968 and 1972, we conduct a Wilcoxon signed ranks test, using a two-sided alternative and the 5% significance level.

The plot below shows the exact distribution of W_+ under the null hypothesis. The exact distribution is based on all $2^{15} = 32,768$ choices of signs for the nonzero differences.



Summary measures are as follows:

$$E(W_+) = 90.0 \quad \text{and} \quad Var(W_+) = 607.125.$$

The observed significance level is

$$p \text{ Value} = 2P(W_+ \geq \_\_\_\_\_) \approx 0.0049.$$

Thus (please complete),

2.3.3 Paired Samples Sign Test

The sign test, introduced by Arbuthnot in the 1700's, is a simple test for differences data that is less sensitive than the Wilcoxon signed rank test, but easy to explain and apply.

Let S_+ be the number of positive signs among the list of n differences. Under the null hypothesis that the X and Y distributions are equal, S_+ is a binomial random variable with success probability $p = 1/2$, and with number of trials equal to the number of nonzero differences.

Large values of S_+ support the alternative that positive differences are more likely than negative differences; small values support the opposite alternative.

Exercise. Find the observed significance level for a two sided test of the null hypothesis of equality of the labor force participation distributions for years 1968 and 1972 using the labor force participation data from page 26. Are the results significant at the 5% significance level?

2.4 Random Samples, Simple Random Samples, Sampling Models

The methods discussed in Sections 2.2 (two sample analysis) and 2.3 (paired samples analysis) assume that observed data are the values of random samples from continuous distributions.

This section describes situations when the methods can be applied to measurements made on individuals drawn from finite populations.

Simple random sample. A *simple random sample* of size n from a population of size N is a subset of n individuals chosen in such a way that each choice of subset is equally likely. If the population size is large enough, then the resulting measurements can be treated as if they were the values a random sample.

For example, the labor force participation study illustrates the use of a simple random sample of cities from the rather large population of cities in the United States.

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

It is interesting to note that sampling can be done in many different ways in the two sample setting, for example, including the following:

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.

Randomization model for two sample analysis. Consider again the last research scenario in the two sample setting (comparing cholesterol treatments), where chance was applied

- In choosing the initial sample of $n + m$ individuals and
- In randomly assigning individuals to one of two treatments.

If the initial sample is *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* for two sample analysis.

Under the randomization model, chance enters into the experiment only through the assignment of subjects to treatments (only the second application of chance). 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 in this setting.

Footnote: Statistical methods that can be applied under both population and randomization models are important in modern applications. We will study these methods further in the next sections, and we will return to them when we study multiple sample analysis and least squares analysis.

2.5 Permutation Analysis

In many statistical applications, the null and alternative hypotheses of interest can be paraphrased in the following simple terms:

H_O: Any patterns appearing in the data are due to chance alone.

H_A: There is a tendency for a certain type of pattern to appear.

Permutation methods allow researchers to determine whether or not to accept a null hypothesis of randomness and, in some cases, to construct confidence intervals for unknown parameters.

The methods are applicable in many settings since they require few mathematical assumptions. Analyses can be based on knowledge of the distributions from which the data were drawn, or upon permuting ranks, or upon permuting the observations themselves.

General example for two sample analysis. Assume that X and Y are continuous random variables and consider testing the null hypothesis that the X and Y distributions are equal, versus the alternative that one distribution is stochastically larger than the other, using independent random samples of sizes n and m, respectively. Let

$$Z_1 < Z_2 < \cdots < Z_{n+m}$$

be the ordered combined list of $n + m$ observations.

- Under the null hypothesis of equality of distributions, each ordering of the combined list of $n + m$ observations is equally likely. Thus, an observed ordering can be thought of as being randomly chosen from $(n + m)!$ equally likely possibilities.
- Under the alternative hypothesis that one distribution is stochastically larger than the other, the values in one sample will tend to be larger than those in the other. Thus, the observed patterns of interest are that X_i's tend to be larger than Y_j's, or that X_i's tend to be smaller than Y_j's.

2.5.1 Permutation Statistics, Distributions and Methods

To conduct a permutation test you need to:

1. Choose a statistic, T, that measures the effect of interest.
2. Construct the sampling distribution that T would have if the effect were *not* present. (This is accomplished by computing the value of T for each reordering of the data.)
3. Find the observed significance level (or p value) based on the sampling distribution from the second step.

The sampling distribution from the second step is known as the *permutation distribution* of the statistic T, and the p value is called a *permutation p value*.

Permutation tests are conditional tests. It is important note that the sampling distribution of T is computed conditional on the observations. The distribution of T must be recomputed each time a new study is conducted.

Exact versus approximate permutation tests. If the number of reorderings of the data is not very large, then an exact permutation p value can be computed, and we say that we have conducted an *exact permutation test*.

Otherwise, the computer can be used to approximate the sampling distribution of T by computing its value for a fixed number of random reorderings of the data. The approximate sampling distribution can then be used to estimate the p value, and we say that we have conducted an *approximate permutation test* or that we have conducted the test using *Monte Carlo methods*.

In typical applications, Monte Carlo methods are needed.

Footnotes:

1. Nonparametric tests are examples of permutation tests where the ranks of observations (rather than the observations themselves) are reordered in all possible ways.
2. Monte Carlo analyses are used in many disciplines to estimate quantities of interest. In permutation testing, the quantity of interest is the permutation p value.
3. When we use Monte Carlo methods to estimate the permutation p value, we will include a confidence interval as well; see examples on pages 39 and 46.

2.5.2 Two Sample Analyses: Difference in Means Tests

A natural starting point is the two sample setting. Let

$$\{x_1, x_2, \ldots, x_n\} \text{ and } \{y_1, y_2, \ldots, y_m\}$$

be the observed samples (lists of numbers with repetitions).

Difference in means statistic. The difference in means statistic D for the two samples is defined as follows:

$$D = \text{Mean of } x \text{ sample} - \text{Mean of } y \text{ sample.}$$

Tests based on D are appropriate in the following situations:

1. *Population model.* The observed data are the values of independent random samples from distributions differing in mean only. The null hypothesis is that the distributions are equal; equivalently, that $\mu_x = \mu_y$. Alternatives of interest are that the mean of one distribution is larger than the mean of the other.
2. *Randomization model.* The data are measurements taken on $n+m$ individuals in distinguishable groups of sizes n and m. The null hypothesis is that the observed difference in means is due to chance alone. Alternatives of interest are that values in one group tend to be larger (but not more variable) than values in the other group.

Permutation distribution. The sampling distribution of the difference in means statistic under the null hypothesis is obtained by randomly partitioning the collection of $n+m$ observations into subsets of sizes n and m, respectively, and calculating the difference in sample means each time. If there are ties in the data, then sampling is done using an urn model, where we imagine that each observation is written on a slip of paper and the collection of slips is partitioned into subsets of sizes n and m.

Summary measures for the permutation distribution of D are given in the following theorem.

> ***Theorem (D Statistic Distribution).*** Conditional on the observed values in the two samples, the permutation distribution of D has the following summary measures:
>
> $$E(D) = 0 \quad \text{and} \quad Var(D) = \frac{n+m}{n\ m\ (n+m-1)} \sum_{i=1}^{n+m} (z_i - \overline{z})^2$$
>
> where $z_1, z_2, \ldots, z_{n+m}$ is the combined list (with repetitions) of the $n+m$ observations, and $\overline{z}$ is the mean of the $n+m$ observations.

Example (Body Fat Study). As part of a study on body fat in men, researchers considered the relationship between body fat and age.

The following table gives body fat measurements (in percent of total weight) for 8 men in the thirties (the X sample) and 10 men in their fifties (the Y sample) who participated in the study, along with certain summaries.

	Values:	*Size:*	*Mean:*
X Sample:	9.4, 22.9, 14.6, 7.9, 0.7, 19.2, 16.9, 5.6	$n = 8$	$\overline{x} = 12.15$
Y Sample:	18.7, 19.5, 15.0, 17.4, 22.6, 26.0, 31.5, 8.8, 8.5, 10.6	$m = 10$	$\overline{y} = 17.86$

For the data in this example, the sampling distribution of the difference in means statistic has the following summary measures:

$$E(D) = 0 \quad \text{and} \quad Var(D) = 13.8617.$$

The observed difference in means is $d_{\text{obs}} =$ _____.

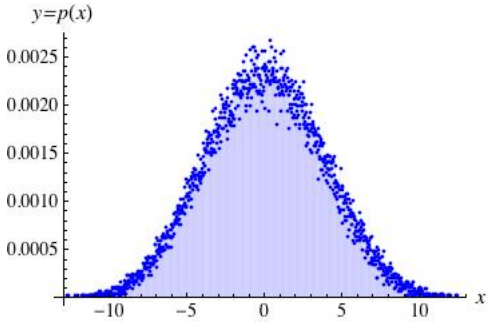
To determine if the observed difference in means is due to chance alone, a permutation analysis will be conducted using a two sided alternative and 5% significance level.

There are $\binom{18}{8} = 43,758$ partitions to consider.

The plot on the right shows the distribution. The observed significance level is

$$p \text{ Value} = P(|D| \geq |d_{\text{obs}}|) = 0.128754.$$

Thus (please complete),



2.5.3 Two Sample Analyses: Smirnov Test

In the 1930's, Smirnov proposed a two sample test based on a comparison of empirical distribution function. The test is appropriate in the following situations:

1. *Population model.* The observed data are the values of independent random samples. The null hypothesis is that the distributions from which the data were drawn are equal versus the general alternative that the distributions are not equal.
2. *Randomization model.* The data are measurements taken on $n+m$ individuals in distinguishable groups of sizes n and m. The null hypothesis is that observed differences in the empirical CDFs are due to chance alone. Alternatives of interest are that the samples differ in some way.

As in the previous section, we consider the analysis of two samples

$$\{x_1, x_2, \ldots, x_n\} \quad \text{and} \quad \{y_1, y_2, \ldots, y_m\},$$

where the samples are lists of numbers with repetitions.

Empirical cumulative distribution function (ECDF). We first define the empirical distribution function of a single sample.

Let $\{x_1, x_2, \ldots, x_n\}$ be the sample of interest. Then

$$ECDF(x) = \frac{\#(x_i\text{'s} \leq x)}{n} \qquad \text{for all real numbers } x.$$

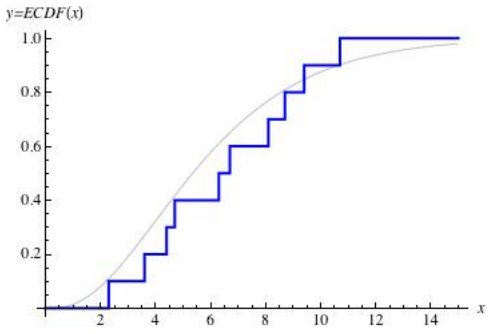
The graph of the $ECDF$ is a step function.

For example, if $n = 10$ and the numbers are

$$2.3, 3.6, 4.4, 4.7, 6.3, 6.7, 8.1, 8.7, 9.4, 10.7,$$

then the step curve shown on the right is the empirical CDF of the numbers.

Vertical lines have been added to the plot for emphasis, as well as a smooth curve that suggests a possible model for these data.



Note that the steps in the plot are all of equal height since the data have distinct values.

Smirnov statistic and permutation distribution. Let $ECDF_1$ and $ECDF_2$ be the empirical CDFs of the x and y samples, respectively.

The *Smirnov statistic*, S, is the maximum absolute difference in the $ECDF$s:

$$S \;=\; \max_x \Big| ECDF_1(x) - ECDF_2(x) \Big|.$$

The permutation distribution of S is obtained by randomly partitioning the collection of $n+m$ observations into subsets of sizes n and m, respectively, and calculating the maximum absolute difference in resulting ECDFs each time. If there are ties in the data, then sampling is done using an urn model, where we imagine that each observation is written on a slip of paper, and the collection of slips is randomly partitioned into subsets of sizes n and m, respectively.

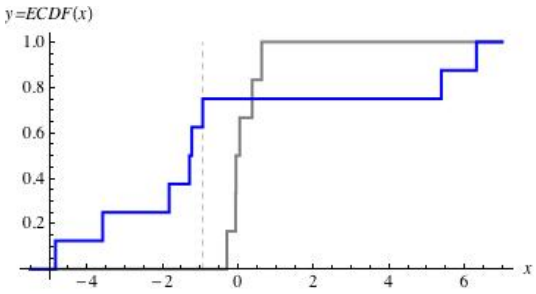
Values of the Smirnov statistic S lie in the interval $[0, 1]$. Observed values near 0 support the null hypothesis; large observed values support the alternative.

For example, consider testing the null hypothesis of randomness at the 5% significance level using samples of sizes 6 and 8, and using Smirnov's two sample statistic, S.

Assume the observed samples are listed below, and the empirical CDFs are shown on the right.

x sample ($n = 6$, light plot):
$-0.29, -0.06, -0.05, 0.05, 0.38, 0.63$

y sample ($m = 8$, dark plot):
$-4.83, -3.58, -1.82, -1.28, -1.22, -0.93, 5.39, 6.32$



(1): To find the maximum absolute difference in the ECDFs, we just need to consider absolute differences at points in the combined list of 14 observations. (Please complete the table.)

x	$\text{ECDF}_1(x)$	$\text{ECDF}_2(x)$	|DIFF|	x	$\text{ECDF}_1(x)$	$\text{ECDF}_2(x)$	|DIFF|
-4.83	0	1/8	____/24	-0.06	2/6	6/8	____/24
-3.85	0	2/8	____/24	-0.05	3/6	6/8	____/24
-1.82	0	3/8	____/24	0.05	4/6	6/8	____/24
-1.28	0	4/8	____/24	0.38	5/6	6/8	____/24
-1.22	0	5/8	____/24	0.63	1	6/8	____/24
-0.93	0	6/8	____/24	5.39	1	7/8	____/24
-0.29	1/6	6/8	____/24	6.32	1	1	____/24

The observed value of Smirnov's statistic is $s_{\text{obs}} =$ _____.

(2): The sampling distribution of S is summarized in the table below.

3/24	*4/24*	*5/24*	*6/24*	*7/24*	*8/24*	*9/24*
0.00033	0.00500	0.04262	0.09890	0.06926	0.14652	0.17716

10/24	*11/24*	*12/24*	*13/24*	*14/24*	*15/24*	*16/24*
0.08392	0.07459	0.10922	0.05328	0.04662	0.03197	0.01798

17/24	*18/24*	*20/24*	*21/24*	*1*
0.01998	0.01332	0.00466	0.00400	0.00067

The observed significance level is $P(S \geq s_{\text{obs}}) =$ _____.

(3): Thus, (please complete)

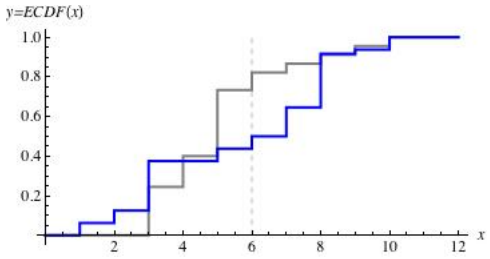
Example (Consumer Preferences Study). As part of a study on consumer preferences for sweet foods, 45 Australian consumers and 48 Japanese consumers were asked to rate a particular brand of chocolate on a ten-point scale, where

- A score of 10 indicates the consumer liked the sweetness, while
- A score of 1 indicates a consumer did not like the sweetness at all.

Results are summarized in the following table:

	Scores:										*Sample Size:*	*Sample Mean:*	*Sample Median:*
	1	*2*	*3*	*4*	*5*	*6*	*7*	*8*	*9*	*10*			
Australian:	0	0	11	7	15	4	2	2	2	2	45	5.067	5.0
Japanese:	3	3	12	0	3	3	7	13	1	3	48	5.625	6.5

ECDFs for the Australian (light plot) and Japanese (dark plot) samples are shown at the right.



The observed value of the Smirnov statistic is 0.32 (87/270), occurring when the score is 6.

Consider testing the null hypothesis of randomness using the 5% significance level.

In a Monte Carlo analysis using 2000 random partitions (including the observed partition of the 93 scores), 0.55% (11/2000) of S values were greater than or equal to the observed value. Further, a 99% confidence interval for the true p value is $[0.00216, 0.01136]$.

Thus (please complete),

2.5.4 Two Sample Analyses: Comparison of Tests

We have seen several approaches to the analysis of two samples. A useful comparison of these methods is as follows:

1. *Pooled t versus difference in means tests.*

   In situations where both the pooled t test of the null hypothesis of equality of means and the difference in means tests are appropriate, the pooled t test is preferred. However, it is interesting to note that the tests give similar results.

2. *Difference in means versus rank sum tests.*

   In situations where both the difference in means and rank sum tests are appropriate:

   - If the samples are highly skewed or have extreme outliers, then the rank sum test is preferred to the difference in means test.
   - Otherwise, the difference in means test is preferred.

   In practice, the rank sum test is used almost exclusively since it is easy to implement.

3. *General alternatives.*

   The Smirnov test is preferred in situations where the alternative hypothesis is that the distributions differ in some way.

2.5.5 Paired Samples Analyses: Fisher Symmetry Test

We next consider another approach to the analysis of paired samples

$$\{\{x_1, y_1\}, \{x_2, y_2\}, \ldots, \{x_n, y_n\}\},$$

or their corresponding lists of differences $\{d_1, d_2, \ldots, d_n\} = \{x_1 - y_1, x_2 - y_2, \ldots, x_n - y_n\}$.

Fisher symmetry statistic. R.A. Fisher, who pioneered the use of permutation methods, proposed using the sum of differences statistic, $S = \sum_{i=1}^{n} d_i$.

Tests based on the sum of differences statistics are appropriate in the following situations:

1. *Population model.* The paired data are the values of a random sample from a bivariate continuous distribution. The difference $D = X - Y$ satisfies a shift model with $\Delta = E(D) = E(X - Y)$. The null hypothesis is that the mean is zero. Alternatives of interest are that the mean is positive or negative.

2. *Randomization model.* The paired data are measurements taken on n individuals (or n pairs of individuals). The null hypothesis is that the signs (positive or negative) of observed differences are due to chance alone. Alternatives of interest are that the observed differences tend to be positive or negative.

Permutation distribution. The sampling distribution of the sum of differences statistic under the null hypothesis of randomness is obtained by computing the sum for each assignment of signs to the observed differences.

Summary measures for the permutation distribution of S are given in the following theorem.

> ***Theorem (Sum of Differences Distribution).*** Conditional on the observed differences, the permutation distribution of the sum of differences statistic, S, has the following summary measures
>
> $$E(S) = 0 \quad \text{and} \quad Var(S) = \sum_{i=1}^{n} d_i^2,$$
>
> where the d_i's are the observed differences. If n is large enough, then the S distribution is approximately normal.

For example, if $n = 8$ and the observed differences are -19, -6, -3, -2, 1, 2, 15, 29, then a total of $2^8 = 256$ sums of the form

$$\pm 19 \pm 6 \pm 3 \pm 2 \pm 1 \pm 2 \pm 15 \pm 29$$

would be computed, where either $+$ or $-$ is chosen in each summand.

The resulting distribution has mean 0, variance 1481 and standard deviation 38.48.

The sum of differences statistic S takes both positive and negative values. Large values support the alternative hypothesis that differences tend to be positive, while small values suppose the alternative that differences tend to be negative.

Example (Calcium in Animal Feed). An experiment was conducted to compare two methods of measuring the percentage of calcium in animal feed. The standard method is accurate but time consuming; the new method is faster.

The question of interest is whether the results are equivalent.

The following data are differences ($d = x - y$) between the percentage measured using the standard method (x) and the percentage measured using the new method (y) for 25 samples:

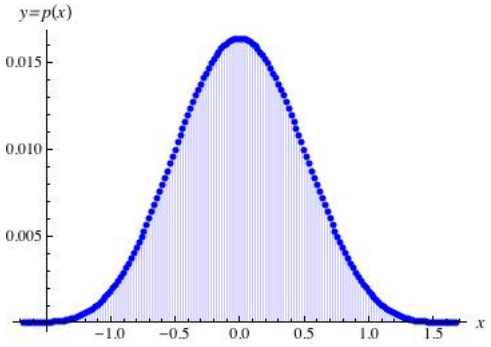
$$-0.25, -0.20, -0.19, -0.14, -0.12, -0.08, -0.07, -0.06, -0.06, -0.05, -0.05, -0.03,$$

$$-0.02, -0.01, 0.01, 0.01, 0.02, 0.02, 0.02, 0.02, 0.02, 0.03, 0.04, 0.06, 0.10.$$

For these data, $\overline{d} = -0.0392$ and $s_d^2 = 0.086^2$.

To determine if the observed differences between measurements made using the standard method and the new method are due to chance alone, a two sided test will be conducted at the 5% significance level.

The exact distribution of the S statistic is shown on the right. The mean is 0, the variance is 21.58, and the standard deviation is 4.65.



The observed value of S is -0.98, and the observed significance level is

$$p \text{ Value} = P(|S| \geq |-0.98|) = 0.032142.$$

Thus (please complete),

2.5.6 Paired Samples Analyses: Comparison of Tests

We have seen several approaches to the analysis of paired samples. A useful comparison of these methods is as follows:

1. *Paired t versus Fisher symmetry tests.*

   In situations where both the paired t test of the null hypothesis that the mean is zero and the Fisher symmetry test are appropriate, the paired t test is preferred. However, it is interesting to note that the tests give similar results.

2. *Fisher symmetry versus signed ranks tests.*

   In situations where both the Fisher symmetry and signed ranks tests are appropriate:

   - if the differences data are highly skewed or have extreme outliers, then the signed rank test is preferred;
   - otherwise, the Fisher symmetry test is preferred.

   In practice, the signed ranks test is used almost exclusively since it is easy to implement.

2.5.7 Correlation Analyses: Sample Correlation Test

This section considers permutation methods for analyzing lists of pairs of numbers

$$\{\{x_1, y_1\}, \{x_2, y_2\}, \ldots, \{x_n, y_n\}\}$$

in one of the following situations:

1. *Population model.* The paired data are the values of a random sample from a bivariate continuous distribution. The null hypothesis of interest is that X and Y are independent, versus alternatives that there is a positive or negative association between the variables.

2. *Randomization model.* The paired data are measurements of two characteristics in each of n individuals. The null hypothesis of interest is that there is no relationship between the characteristics. Alternatives of interest are that the characteristics are positively or negatively associated.

Sampling under population model. Consider sampling under the population model and, for convenience, index the observations so that the event

$$X_1 < X_2 < \cdots < X_n \quad \text{is observed.}$$

If X and Y are independent (the null hypothesis of interest), then the observed ordering of the Y_i's is one of $n!$ equally likely choices.

This fact forms the basis of the permutation methods below.

Sample correlation statistic. The sample correlation statistic,

$$R = \frac{\sum_{i=1}^{n}(x_i - \overline{x})(y_i - \overline{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \overline{x})^2 \ \sum_{i=1}^{n}(y_i - \overline{y})^2}},$$

can be used to test the null hypothesis of randomness.

Values of R lie in the interval $[-1, 1]$. Positive values favor the alternative that the characteristics under study are positively associated, and negative values favor the alternative that the characteristics under study are negatively associated.

<u>*Note*</u> that under the population model, R is an estimate of the correlation coefficient

$$\rho = Corr(X, Y) = \frac{\sigma_{xy}}{\sigma_x \sigma_y} = \frac{Cov(X, Y)}{\sqrt{Var(X)Var(Y)}} = \frac{E((X - \mu_x)(Y - \mu_y))}{\sqrt{E((X - \mu_x)^2)E((Y - \mu_y)^2)}},$$

and tests based on R are tests of the null hypothesis that X and Y are uncorrelated.

Thus, tests based on R are not tests of independence.

Permutation distribution. The permutation distribution of the sample correlation statistic is obtained by computing the sample correlation for each matching of a permutation of the y values to the ordered x values.

The following theorem gives information about the resulting distribution.

> ***Theorem (R Statistic Distribution).*** Conditional on the observed pairs, the permutation distribution of R has the following summary measures:
>
> $$E(R) = 0 \ \text{ and } \ Var(R) = \frac{1}{n-1}.$$
>
> If n is large enough, then the R distribution is approximately normal.

For example, consider testing the null hypothesis that X and Y are independent versus the alternative that X and Y are *positively associated* at the 5% significance level, using a random sample of size $n = 4$ from the joint (X, Y) distribution.

Assume the following data are the values of a random sample from the joint distribution:

$$\{\{-1.3, -0.2\}, \{-1.2, -1.1\}, \{0.6, 0.7\}, \{0.8, -0.8\}\}$$

The sampling distribution of R is obtained by computing its value for each of $4! = 24$ matchings of y values to ordered x values, as shown in the table below.

r	Permutation of y's	r	Permutation of y's
-0.90	$\{\ 0.7, -0.2, -0.8, -1.1\}$	0.14	$\{-0.2, -0.8,\ \ 0.7, -1.1\}$
-0.88	$\{\ 0.7, -0.2, -1.1, -0.8\}$	0.16	$\{-0.8, -0.2,\ \ 0.7, -1.1\}$
-0.86	$\{-0.2,\ \ 0.7, -0.8, -1.1\}$	0.27	$\{-0.2, -0.8, -1.1,\ \ 0.7\}$
-0.84	$\{-0.2,\ \ 0.7, -1.1, -0.8\}$	0.30	$\{-0.8, -0.2, -1.1,\ \ 0.7\}$
-0.50	$\{\ 0.7, -0.8, -0.2, -1.1\}$	0.36	$\{-0.2, -1.1,\ \ 0.7, -0.8\}$
-0.44	$\{-0.8,\ \ 0.7, -0.2, -1.1\}$	0.40	$\{-1.1, -0.2,\ \ 0.7, -0.8\}$
-0.43	$\{\ 0.7, -0.8, -1.1, -0.2\}$	0.47	$\{-0.2, -1.1, -0.8,\ \ 0.7\}$
-0.37	$\{-0.8,\ \ 0.7, -1.1, -0.2\}$	0.51	$\{-1.1, -0.2, -0.8,\ \ 0.7\}$
-0.27	$\{\ 0.7, -1.1, -0.2, -0.8\}$	0.83	$\{-0.8, -1.1,\ \ 0.7, -0.2\}$
-0.23	$\{\ 0.7, -1.1, -0.8, -0.2\}$	0.84	$\{-1.1, -0.8,\ \ 0.7, -0.2\}$
-0.21	$\{-1.1,\ \ 0.7, -0.2, -0.8\}$	0.90	$\{-0.8, -1.1, -0.2,\ \ 0.7\}$
-0.16	$\{-1.1,\ \ 0.7, -0.8, -0.2\}$	0.91	$\{-1.1, -0.8, -0.2,\ \ 0.7\}$

The observed value of the test statistic is $r_{\text{obs}} =$ _____.

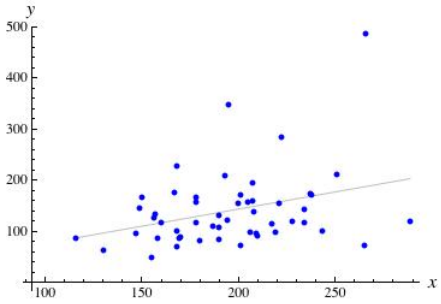
The observed significance level is _____.

Thus (please complete),

Example (Hand et al., Chapman & Hall, 1994). Cholesterol and triglycerides belong to the class of chemicals known as lipids (fats). As part of a study to determine the relationship between high levels of lipids and coronary artery disease, researchers measured plasma levels of cholesterol and triglycerides in milligrams per deciliter (mg/dL) in 371 men complaining of chest pain.

This example considers the relationship between the cholesterol (x) and triglycerides (y) measurements for the 51 men with no evidence of disease.

The plot on the right shows the cholesterol-triglycerides pairs, superimposed on a least squares fit line.



Sample summaries are

$$\overline{x} = 195.275,\ s_x^2 = 36.11^2;\ \overline{y} = 140.353,\ s_y^2 = 74.29^2.$$

The sample correlation is 0.325.

To determine if the observed association is due to chance alone, a permutation test will be conducted using a two sided alternative and 5% significance level.

In a Monte Carlo analysis using 5000 random permutations (including the observed permutation of the y values), 1.96% (98/5000) of $|R|$ values were greater than or equal to (the absolute value of) the observed correlation.

Further, a 99% confidence interval for the permutation p value was $[0.0149, 0.0252]$.

Thus (please complete),

2.5.8 Correlation Analyses: Rank Correlation Test

In the early 1900's, Spearman proposed a test based on the ranks of the x and y values. Spearman's *rank correlation statistic*, R_s, is computed as follows:

1. Replace each x by its rank (or midrank) in the ordered x values.
2. Replace each y by its rank (or midrank) in the ordered y values.
3. Let R_s equal the sample correlation of the paired ranks.

For example, if $n = 6$ and the list of paired data is

$$\{\{10.42, 13.18\}, \{11.43, 14.03\}, \{11.79, 13.24\}, \{13.17, 12.03\}, \{13.4, 11.75\}, \{13.53, 11.83\}\}$$

then the list of paired ranks is:

$$\{\{1, \_\_\_\_\_\}, \{2, \_\_\_\_\_\}, \{3, \_\_\_\_\_\}, \{4, \_\_\_\_\_\}, \{5, \_\_\_\_\_\}, \{6, \_\_\_\_\_\}\}$$

and the observed value of Spearman's statistic is -0.771429.

Permutation distribution and test. The distribution of R_s is computed as follows: for each matching of y ranks to ordered x ranks, the rank correlation value is computed.

Permutation tests using R_s are conducted in the same way as permutation tests using R. Unless there are many ties in the data or n is very small, the large sample normal approximation to the R_s distribution can be used to estimate p values.

Exercise. Consider again the cholesterol-triglycerides example from page 46. The observed value of Spearman's statistic is 0.288. Use the normal approximation to the sampling distribution of R_s to find the observed significance level for a two sided test of the null hypothesis that the observed association is due to chance alone.

2.5.9 Correlation Analyses: Comparison of Tests, Effect of Outliers

Permutation tests based on R and R_s are valid in the same situations. The following is a list of useful facts:

1. *Skewed data or extreme outliers:*

   If the data are highly skewed or have extreme outliers, then the rank correlation test is preferred; otherwise, the sample correlation statistic should be used.

2. *Invariance to transformation:*

   The rank correlation statistic is unchanged if either the x values or the y values are transformed using an increasing function (such as square or square-root); the sample correlation statistic is not invariant to such transformations.

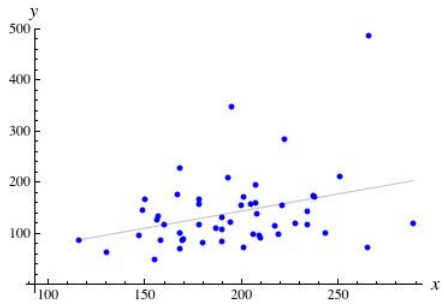
3. *Large sample methods:*

   The large sample normal approximation to the sampling distribution of the rank correlation statistic is generally good when there are 30 or more observations.

   When working with the sample correlation statistic, it is best to use Monte Carlo methods to approximate the p value even when the sample size is very large.

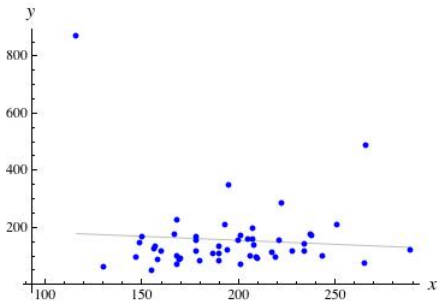
Effect of outliers. To illustrate the effect of outliers, consider changing the y-coordinate of the pair with the smallest x value in the cholesterol-triglycerides data from page 46.

Original Data: $(x, y) = (116, 87)$



$r = 0.325$, $r_s = 0.288$

Changed Data: $(x, y) = (116, 870)$



$r = -0.081$, $r_s = 0.186$

1. *Left Plot:* The left plot shows the original data set, along with the values of the sample correlation and rank correlation for these data. Both numerical and graphical summaries suggest a positive association between cholesterol and triglycerides.

2. *Right Plot:* The right plot shows the result of changing the y-coordinate from a number within the original range of triglyceride values to a number far above the original range. The observed value of the sample correlation and the graph now suggest a very weak negative association, while the rank correlation remains positive.

2.6 Bootstrap Analysis

In many statistical applications, interest focuses on estimating a quantity using a random sample from a probability distribution, the distribution from which the data were drawn is not known exactly, and the sampling distribution of the statistic used to estimate the quantity is not known (either exactly or approximately).

Bootstrap methods allow researchers to make approximate probability calculations in these situations by using the computer to simulate the original experiment many times.

2.6.1 Bootstrap Resampling: Estimated Model, Observed Distribution

Let θ be a parameter of interest, T a statistic used to estimate θ from sample data, and t_{obs} the observed value of T. Assume the sample data are the values of a random sample, or of independent random samples.

For example, in a one sample analysis of the mean of a probability distribution, we would let

$$\theta = E(X) = \mu \quad \text{and} \quad T = T(X_1, X_2, \ldots, X_n) = \overline{X},$$

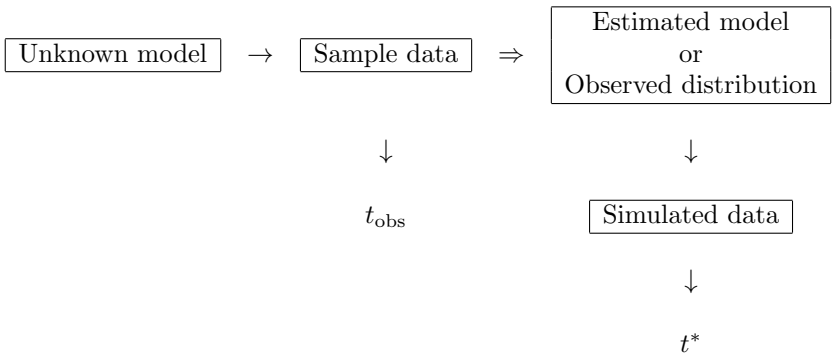
where $X_1, X_2, \ldots, X_n$ is a random sample of size n from that distribution.

Similarly, in a two sample analysis of the difference in population means, we would let

$$\theta = E(X) - E(Y) = \mu_x - \mu_y \quad \text{and} \quad T = T(X_1, \ldots, X_n, Y_1, \ldots, Y_m) = \overline{X} - \overline{Y},$$

where $X_1, \ldots, X_n$ and $Y_1, \ldots, Y_m$ are independently chosen random samples from their respective distributions.

Bootstrap resampling. In the resampling step of a bootstrap analysis, a model estimated using the sample data or the observed distribution of the sample data is used to produce simulated data and simulated values of T, say t^*, as illustrated in the right column below.



This process is repeated a large number of times, say B, to produce an approximate sample from the sampling distribution of T:

$$t_1^*,\ t_2^*,\ t_3^*,\ \ldots,\ t_B^*.$$

The bootstrap resampled values are then studied.

Parametric versus nonparametric bootstrap. *Parametric bootstrap analyses* assume that the distributions of interest have given functional forms; estimated models are used to produce resampled values. By contrast, *nonparametric bootstrap analyses* make no assumptions about the forms of the distributions and use observed distributions of sample data to produce resampled values. The following examples illustrate each approach.

Parametric bootstrap example. To illustrate the simulation process when working with an estimated model, we consider a study where the data can be modeled using a gamma distribution and where our interest focuses on estimating the sampling distribution of the method of moments estimator of α (the shape parameter).

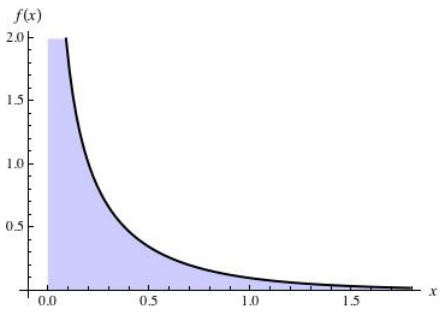
Example (Source: Rice textbook, Chapter 8). In a study of the natural variability of rainfall, the rainfall of summer storms was measured at a series of rain gauges in southern Illinois in the years 1960 to 1964. There were 227 storms in the five-year period.

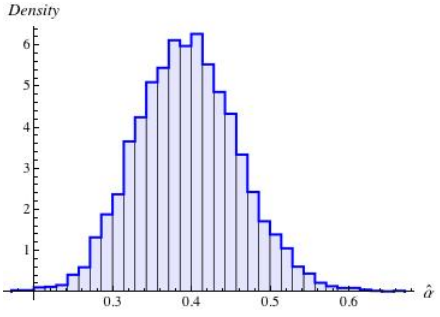
Let X be the rainfall (in inches) of a summer storm in Illinois, and assume that the observed rainfalls for the years 1960 to 1964 are the values of a random sample from the X distribution.

For gamma models, the method of moments estimators of α (shape) and β (scale) are

$$\widehat{\alpha} = \frac{(\overline{X})^2}{(1/n)\sum_{i=1}^{n}(X_i - \overline{X})^2} \quad \text{and} \quad \widehat{\beta} = \frac{(1/n)\sum_{i=1}^{n}(X_i - \overline{X})^2}{\overline{X}}.$$

For the rainfalls data, the estimates are $\widehat{\alpha} = 0.378$ and $\widehat{\beta} = 0.594$.





1. *Left plot:* The left plot shows the PDF of the estimated gamma distribution. That is, the left plot shows the PDF of the gamma distribution with shape parameter 0.378 and scale parameter 0.594.

2. *Right plot:* The right plot is a histogram of 5000 randomly resampled estimates of α, each based on a pseudo-random sample of size 227 drawn from the estimated gamma model. The 5000 randomly resampled estimates have mean 0.395 and standard deviation 0.065.

Nonparametric bootstrap example. To illustrate the simulation process when working with observed distributions of sample data, we consider a study comparing telephone repair services provided by Verizon to its own customers and to customers of its local competitors.

Example (Source: Hesterberg, 2003). As the primary telephone company in the eastern U.S., Verizon is required by law to allow local telephone companies to use their cables, and to provide repair service to customers of the local companies. They are also required to provide evidence that the average repair time for customers of other companies is the same as for Verizon customers.

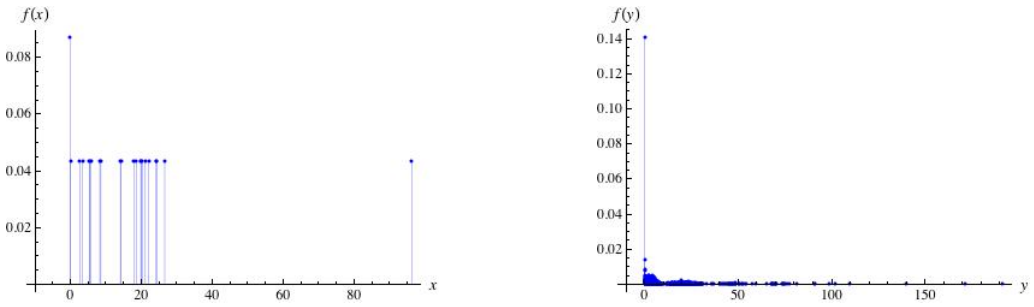
The following are summaries for repair times in hours for local (x) and Verizon (y) customers:

Local Customers:	$n = 23$, $\overline{x} = 16.51$, $s_x^2 = 19.50^2$
Verizon Customers:	$m = 1664$, $\overline{y} = 8.41$, $s_y^2 = 14.69^2$

The observed difference in means is 8.1 hours.

Let X be the repair time in hours for the customer of a local company in the eastern U.S., and let Y be the repair time for a Verizon customer. Assume that the information above summarizes the values of independent random samples from the X and Y distributions.

The observed distributions of repair times for local and Verizon customers are shown below:



1. *Left Plot:* Repair times for local customers varied from 0 (repair was handled by flipping a switch at the office) to 96.32 hours. The observed distribution for local customers is the discrete distribution with PDF

$$f(x) = \frac{\#(x_i\text{'s equal to } x)}{23} \quad \text{for all } x.$$

   There were 22 distinct repair times. Two customers had a repair time of 0 hours.

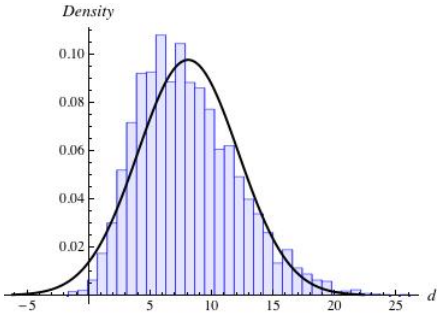
2. *Right Plot:* Repair times for Verizon customers varied from 0 (repair was handled by flipping a switch at the office) to 191.6 hours. The observed distribution for Verizon customers is the discrete distribution with PDF

$$f(y) = \frac{\#(y_j\text{'s equal to } y)}{1664} \quad \text{for all } y.$$

   There were 729 distinct repair times, and there were a total of 234 zero times.

The plot on the right below gives a histogram of 5000 randomly resampled mean differences, where each difference was obtained by

1. Computing the mean of a random resample of size 23 from the local customers distribution,
2. Computing the mean of a random resample of size 1664 from the Verizon customers distribution, and
3. Recording the difference in means.



The normal distribution with mean 8.1 and standard deviation $\sqrt{\frac{19.50^2}{23} + \frac{14.69^2}{1664}} \approx 4.08$ is superimposed on the histogram.

The simulated $\overline{X} - \overline{Y}$ values do not follow the large sample approximate distribution; this suggests that statistical analyses should not be based on normal theory approximations.

Footnote on sampling from observed distributions. A useful way to think about sampling from the observed distribution is as follows: Imagine writing the n observations on n slips of paper, and placing the slips in an urn. The following experiment is repeated n times:

"Thoroughly mix the urn, choose a slip, record the value, return the slip to the urn."

Thus, each replicate data set is the result of sampling *with* replacement n times from the original list of n observations.

If the observed distribution represents the true distribution well, then it is likely that the simulated values of T will represent the sampling distribution of T well.

2.6.2 Bootstrap Estimates of Bias and Standard Error

Let θ be a parameter of interest and T a statistic used to estimate θ from sample data.

Recall that the

1. *Bias* of the estimator T is the difference between its expected value and the parameter we are interested in estimating: $BIAS(T) = E(T) - \theta$.
2. *Standard error (SE)* of T is another name for its standard deviation: $SE(T) = SD(T)$.

Bootstrap methods can be used to estimate these quantities.

Specifically, let t_{obs} be the observed value of T for the sample data,

$$t_1^*,\ t_2^*,\ t_3^*,\ \ldots,\ t_B^* \text{ be a list of } B \text{ bootstrap resampled values,}$$

and $\overline{t^*} = \frac{1}{B}\sum_{i=1}^{B} t_i^*$ be mean of the resampled values.

Then, the bootstrap estimates of bias and standard error are as follows:

$$bias = \overline{t^*} - t_{\text{obs}} \quad \text{and} \quad se = \sqrt{\frac{1}{B-1}\sum_{i=1}^{B}(t_i^* - \overline{t^*})^2}.$$

Thus, the bootstrap estimate of bias is obtained by

$$\text{replacing } E(T) \text{ by } \overline{t^*} \text{ and } \theta \text{ by } t_{\text{obs}} \text{ in the definition of the } BIAS(T),$$

and the bootstrap estimate of standard error is obtained by using the sample standard deviation of the B resampled values.

For example, we can use the 5000 bootstrap resampled values of the MOM estimator of the shape parameter in the Illinois rainfall example from page 50 to estimate the bias and standard error of the MOM estimator when 227 storms are reported:

$$bias = \overline{\alpha^*} - \widehat{\alpha} = 0.395 - 0.378 = 0.017 \quad \text{and} \quad se = sd(\{\alpha_i^*\}) = 0.065.$$

2.6.3 Bootstrap Error Distribution

The *error distribution* of an estimator T is the distribution of $T - \theta$.

The error distribution has mean equal to the bias of T and standard deviation equal to the standard error of T since

$$E(T-\theta) = E(T) - \theta = BIAS(T) \text{ and } SD(T-\theta) = SD(T) = SE(T).$$

Bootstrap methods can be used to estimate the error distribution. Specifically, shifted bootstrap resampled values (where t_{obs} replaces the unknown θ),

$$t_1^* - t_{\text{obs}},\ t_2^* - t_{\text{obs}},\ t_3^* - t_{\text{obs}},\ \ldots,\ t_B^* - t_{\text{obs}},$$

are used to construct an estimated shifted distribution.

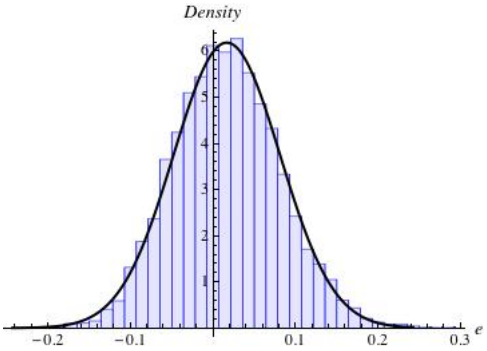
Why should we care? Practitioners often use the bootstrap error distribution as a way to help visualize the results of a bootstrap analysis. For example, if the center of the shifted distribution is close enough to zero, then the practitioner may conclude that the bias of a particular estimator is negligible when data sets are of the size as the observed data.

In addition, since sampling distributions for many commonly used estimators are approximately normal when sample sizes are large enough, practitioners will often enhance the visualization by superimposing a normal density with mean equal to the bootstrap bias and with standard deviation equal to the bootstrap estimate of the standard error.

Consider again the 5000 bootstrap resampled values of the MOM estimator of the shape parameter in the Illinois rainfall example from page 50.

The graph to the right shows

- the estimated error distribution with
- the normal density mean 0.017 and standard deviation 0.065 superimposed.



The estimated bias is a fairly large proportion of the estimated standard error,

$$0.017/0.065 \approx 0.262,$$

and the center of the normal approximation is visibly different from zero. This suggests that the MOM estimator is a biased estimator of α for samples of size 227.

Lastly, the computations suggest that the sampling distribution of the MOM estimator is fairly symmetric for samples of size 227.

Note on an approximate 5% rule for bias:

In order to determine if simulation results suggest that an estimator is biased, we first compute absolute bias as a percentage of standard error. If the result is greater than 5%, then the simulation results suggest that the estimator is biased.

2.6.4 Implementations, Sources of Errors

Bootstrap methods are used in situations where the sampling distribution of T is not known exactly or even approximately. Monte Carlo sampling from the observed distribution or an estimated model is used to produce B replicate values.

Sources of errors. There are two sources of error in bootstrap analyses:

1. The error in using observed distributions or estimated models instead of the true probability distributions of interest, and
2. The error in using a fixed number of replicate data sets to approximate the sampling distribution of T and to estimate its summary measures.

If the sample size n is large, the sample data approximate the distributions of interest well, and the resampling scheme does not rely strongly on a small subset of the observed data, then the results of bootstrap analyses are generally good.

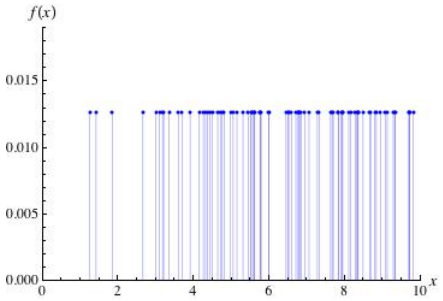
The following example illustrates a situation where the estimated T distribution is not close to its true distribution.

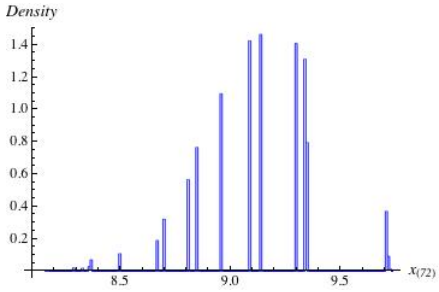
Example. Assume the following data are the values of a random sample of size 79 from a continuous distribution,

1.26	1.43	1.86	2.67	3.01	3.11	3.19	3.22	3.36	3.61	3.71	3.92	4.16	4.27	4.33
4.38	4.45	4.51	4.65	4.72	4.77	4.82	4.99	5.06	5.16	5.32	5.46	5.53	5.54	5.60
5.61	5.63	5.77	5.78	5.80	6.00	6.01	6.46	6.52	6.53	6.60	6.72	6.78	6.79	6.83
6.86	6.95	7.07	7.29	7.34	7.64	7.68	7.71	7.84	7.85	7.92	7.95	7.96	8.12	8.17
8.29	8.33	8.36	8.37	8.50	8.67	8.70	8.81	8.85	8.96	9.09	9.14	9.30	9.34	9.35
9.71	9.72	9.73	9.83											

and assume that we are interested in using the 72$^{\text{nd}}$ order statistic to estimate the 90$^{\text{th}}$ percentile of the distribution.

The observed value of the 72$^{\text{nd}}$ order statistic is _____.





1. *Left Plot:* The left plot is the observed distribution of the sample data.

2. *Right Plot:* The right plot is a histogram of 5000 sample 72$^{\text{nd}}$ order statistics, where each sample order statistic is based on a pseudo-random sample of size 79 from the observed distribution. The mean is 9.13 and the standard deviation is 0.26.

   Note that 94.9% (4746/5000) of the simulated values were equal to one of the following 10 numbers:

   $$8.70,\ 8.81,\ 8.85,\ 8.96,\ 9.09,\ 9.14,\ 9.30,\ 9.34,\ 9.35,\ 9.71.$$

   Thus, the histogram does not approximate a continuous curve very well.

Although 9.13 and 0.26 are reasonable estimates of $E(X_{(72)})$ and $SD(X_{(72)})$, respectively, the shape of the distribution on the right is not close to the shape of the distribution of an order statistic from a continuous distribution.

Note that increasing the number of resamples would not help.

2.6.5 Bootstrap Confidence Intervals

The bootstrap was introduced by Stanford University Professor Brad Efron in the late 1970's as a quick-and-dirty way to help researchers gain information about the sampling distributions of statistics whose exact or approximate forms were unknown. Researchers were soon interested in learning if the bootstrap could be used for more than descriptive purposes.

This section considers three approaches to finding approximate confidence intervals for θ.

Methods 1 and 2. The first two methods are not true confidence interval procedures since they do not guarantee approximate accuracy and are not transformation preserving.

1. *Normal Approximation Interval:* In many applications, the distribution of T is approximately normal when n is large, although the mean and standard deviation are unknown. The idea of the first method is to construct an approximate confidence interval using cutoffs from the standard normal distribution.

   Specifically, let t_{obs} be the observed value of T for the sample data, and let *bias* be the bootstrap estimate of bias and *se* be the bootstrap estimate of standard error based on B bootstrap resamples.

   Then the $100(1-\alpha)\%$ *normal approximation interval* for θ has the following form:

   $$(t_{\text{obs}} - bias) \ \pm \ z(\alpha/2)se$$

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

   [Note that the center of the interval is the "bias corrected" estimate of θ.]

2. *Basic Bootstrap Interval:* The second method uses estimated quantiles from the bootstrap error distribution as cutoffs for a confidence interval procedure and relies on the following idea:

   Let e_p be the p^{th} quantile of the error distribution. Since

   $$1-\alpha = P(e_{\alpha/2} \le T-\theta \le e_{1-\alpha/2}) = P(T - e_{1-\alpha/2} \le \theta \le T - e_{\alpha/2}),$$

   a $100(1-\alpha)\%$ confidence interval for θ is $\left[T - e_{1-\alpha/2}, T - e_{\alpha/2}\right]$.

   To apply the method, let

   - t_{obs} be the observed value of T for the sample data and
   - e^*_p be the sample p^{th} quantile of the estimated error distribution.

   Then the $100(1-\alpha)\%$ *basic bootstrap interval* for θ has the following form

   $$\left[t_{\text{obs}} - e^*_{1-\alpha/2}, \ t_{\text{obs}} - e^*_{\alpha/2}\right].$$

Note that if the distribution of T truly is approximately normal, then the first two methods will produce intervals that are close.

For example, 95% intervals for the shape parameter α based on the simulation results for the rainfall data study from page 50 are as follows:

95% Normal Approximation Interval	*95% Basic Bootstrap Interval*
$[0.234, 0.487]$	$[0.229, 0.481]$

The intervals are indeed close in this case.

Method 3: Efron's improved bootstrap method.[1] As mentioned above, the first two methods for constructing intervals are not true confidence interval procedures, but rather descriptive summaries of the simulations.

Efron's goals in developing a new method were to produce intervals that were approximately accurate and that were transformation preservation, defined as follows:

1. *Approximate Accuracy:*

   The interval $[L, U]$ is a $100(1-\alpha)\%$ confidence interval for θ if

   $$P(\theta < L) = P(\theta > U) = \tfrac{\alpha}{2} \text{ and } P(L \leq \theta \leq U) = 1 - \alpha.$$

   A method for computing intervals is said to be *approximately accurate* when

   $$P(\theta < L) \approx \tfrac{\alpha}{2} \text{ and } P(\theta > U) \approx \tfrac{\alpha}{2} \text{ and } P(L \leq \theta \leq U) \approx 1 - \alpha.$$

2. *Transformation-Preserving:*

   Confidence interval procedures are transformation-preserving.

   That is, if $[L, U]$ is a $100(1-\alpha)\%$ confidence interval for θ and g is a monotone real-valued function, then

   (a) if g is increasing, then $[g(L), g(U)]$ is a $100(1-\alpha)\%$ CI for $g(\theta)$;

   (b) if g is decreasing, then $[g(U), g(L)]$ is a $100(1-\alpha)\%$ CI for $g(\theta)$.

[1] Known as the *bias-corrected adjusted percentile method* (or BC_a method). Source: Efron & Tibshirani 1993.

Beginning with the following assumptions,

> There exists an increasing real-valued function g so that the error distribution on the transformed scale, $g(T) - g(\theta)$, is approximately normally distributed with
>
> $$\text{mean } -z_o \text{ and standard deviation } 1 + ag(\theta),$$
>
> where a and z_o and constants.
>
> 1. the constant z_o reflects the bias of $g(T)$ as an estimator of $g(\theta)$, and
> 2. the constant a describes how the standard deviation of the transformed estimator changes as the parameter changes.

Efron showed that

1. Estimates of the two constants can be obtained from bootstrapped values of T.
2. After transforming back to the original scale (i.e. by using g^{-1}), a $100(1-\alpha)\%$ bootstrap interval of the form
$$\left[t^*_{p_1}, t^*_{p_2}\right]$$
has the properties of being approximately accurate and transformation-preserving, where
   (a) t^*_p is the sample p^{th} quantiles of the bootstrapped values of T, and
   (b) the quantiles p_1 and p_2 are calculated as follows:
$$p_1 = \Phi\left(\widehat{z_o} + \frac{\widehat{z_o} - z(\alpha/2)}{1 - \widehat{a}(\widehat{z_o} - z(\alpha/2))}\right) \quad \text{and} \quad p_2 = \Phi\left(\widehat{z_o} + \frac{\widehat{z_o} + z(\alpha/2)}{1 - \widehat{a}(\widehat{z_o} + z(\alpha/2))}\right),$$
   where $z(\alpha/2)$ is the $100(1-\alpha/2)\%$ point of the standard normal distribution.

   (Thus, you never need to find the transformation g. But, you do need to estimate the two constants from the bootstrapped values.)

Efron's improved method will be illustrated in the examples in the following sections.

2.6.6 Bootstrap Application: Trimmed Mean Analysis

Let X be a continuous random variable with density function $f(x)$, and let α be a proportion in the interval $0 < \alpha < \frac{1}{2}$.

Trimmed mean. The $100\alpha\%$ *trimmed mean* of the X distribution is the expected value of the middle $100(1-2\alpha)\%$ of the distribution:

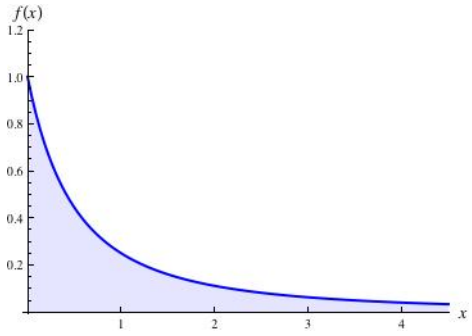
$$100\alpha\% \text{ Trimmed Mean} = \frac{1}{1-2\alpha} \int_{x_\alpha}^{x_{1-\alpha}} x\, f(x)dx$$

where x_p is the p^{th} quantile of the X distribution.

For example, let X be the continuous random variable with density function

$$f(x) = \frac{1}{(1+x)^2} \quad \text{when } x > 0,$$

and $f(x) = 0$ otherwise.



This distribution has an indeterminate mean, but well-defined quantiles.

We can show that

1. A general formula for the p^{th} quantile of the X distribution is $x_p = \frac{p}{1-p}$. Further, the median of the distribution is 1 and the interquartile range is 8/3.
2. A general formula for the $100\alpha\%$ trimmed mean is

$$100\alpha\% \text{ Trimmed Mean} = \frac{1}{1-2\alpha} \ln\left(\frac{1-\alpha}{\alpha}\right) - 1.$$

In particular, the 20% trimmed mean of the distribution is approximately 1.3105.

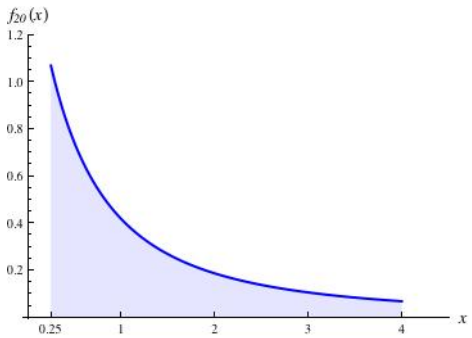
That is, the mean of the continuous distribution with density

$$f_{20}(x) = \frac{1}{(0.60)(1+x)^2}$$

when x is in the restricted interval

$$(x_{.20}, x_{.80}) = (0.25, 4.00),$$

and $f_{20}(x) = 0$ otherwise, is about 1.3105.



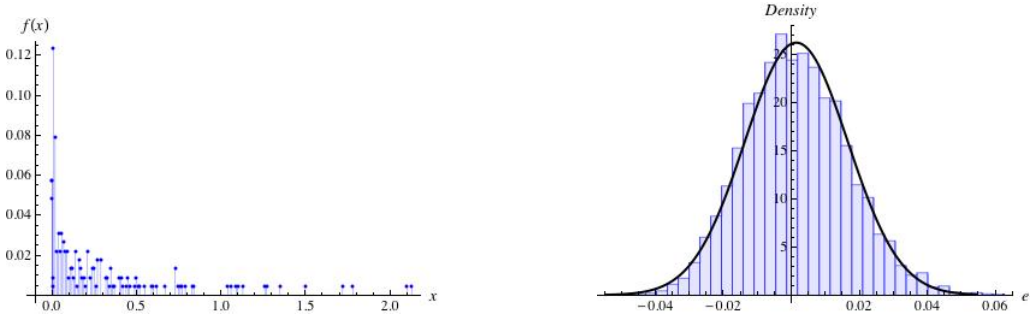
Sample trimmed mean. The *sample* $100\alpha\%$ *trimmed mean* is the sample mean of the middle $100(1-2\alpha)\%$ of the observed data.

For example, if $n = 10$ and the ordered observed data are

$$2.2,\ 6.3,\ 8.6,\ 8.8,\ 11.8,\ 13.6,\ 17.4,\ 27.8,\ 29.4,\ 58.4,$$

then the sample 20% trimmed mean is _____.

Example (Source: Rice textbook, Chapter 8). This example uses nonparametric bootstrap methods to analyze the 20% trimmed mean of the distribution of Illinois rainfalls using the rainfall data introduced on page 50.



1. *Left Plot:* The left plot shows the observed distribution of the 227 observations. For these data, the sample mean is 0.224 inches, the sample median is 0.07 inches, and the sample 20% trimmed mean is 0.106 inches.

2. *Right Plot:* The right plot shows an estimated error distribution, based on 5000 random resamples from the observed distribution; for each resample, a 20% trimmed mean was computed.

   For this simulation, the bootstrap estimate of bias is 0.0015 and the bootstrap estimate of standard error is 0.0152. Bias is about 9.82% of standard error. Further, 95% summary intervals are as follows:

95% Normal Approximation Interval	*95% Basic Bootstrap Interval*
$[0.0747, 0.1343]$	$[0.0729, 0.1323]$

A 95% improved bootstrap interval, based 5000 random resamples, is $[0.0759, 0.1393]$.

Comments (please complete):

2.6.7 Bootstrap Application: Ratio of IQRs Analysis

We can compare the spread of two distributions by comparing their interquartile ranges (IQRs).

Let X and Y be continuous random variables with interquartile ranges IQR_x and IQR_y, respectively. Let

$$\theta = \text{IQR}_x/\text{IQR}_y,$$

be the ratio of the interquartile ranges (IQRs) of the X and Y distributions, and T be the ratio of sample IQRs based on independent random samples from the X and Y distributions.

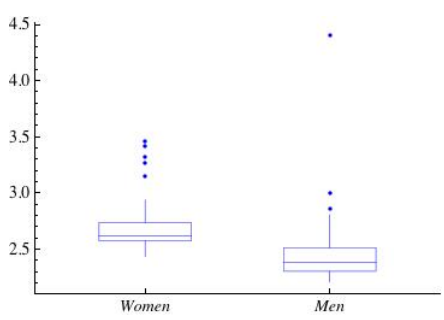
Example (Source: www.olympic.org). This example uses nonparametric bootstrap methods to analyze the ratio of interquartile ranges for the women and men who participated in the 1996 Olympic marathon competition held in Atlanta, GA.

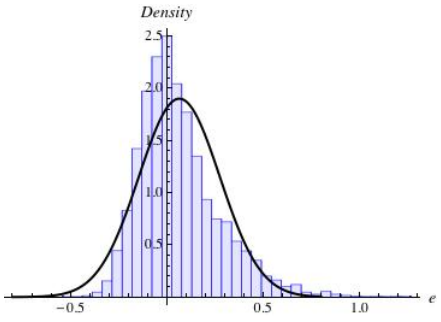
Finishing times in hours for the 65 women and 111 men who completed the competition were downloaded from the official Olympics website. Data summaries are as follows:

	Sample Size:	*Sample Median:*	*Sample IQR:*
Women:	65	2.621 hr	0.163 hr
Men:	111	2.384 hr	0.209 hr

The observed ratio of IQRs is 0.7793.

Let X and Y be the finishing times (in hours) for women and men, respectively, competing in circumstances similar to the 1996 games. Assume the observed finishing times are the values of independent random samples from these distributions.





1. *Left Plot:* The left plot shows side-by-side box plots of the observed times.

   Values in the combined sample range from 2.21 hours to 4.41 hours. Each sample distribution has several outliers.

2. *Right Plot:* The right plot shows an estimate of the error distribution for ratio of IQRs based on 5000 random resamples; each resampled ratio is based on separate resamples from the observed distributions of the women and men, respectively.

   For this simulation, the bootstrap estimate of bias is 0.0653 and the bootstrap estimate of standard error is 0.2099. Bias is about 31.12% of standard error. Further, 95% summary intervals are as follows:

95% Normal Approximation Interval	*95% Basic Bootstrap Interval*
$[0.3026, 1.1254]$	$[0.2068, 1.0305]$

A 95% improved bootstrap interval, based 5000 random resamples, is $[0.5086, 1.2415]$.

Comments (please complete):