HOW MANY DISCOVERIES ARE LOST BY IGNORING MODERN ROBUST METHODS?

A COPY OF THIS DOCUMENT CAN BE DOWNLOADED FROM THE MY WEB PAGE. GOOGLE RAND WILCOX AND FOLLOW THE LINK TO MY WEB PAGE, THEN CLICK ON WORKSHOP.

1 SOME PRELIMINARY REMARKS

HUNDREDS OF PAPERS PUBLISHED DURING THE LAST HALF CENTURY PROVIDE SUBSTANTIALLY IMPROVED METHODS FOR DEALING WITH NON-NORMALITY, OUTLIERS, HETEROSCEDASTICITY AND CURVATURE.

PRACTICAL REASONS FOR TAKING ADVANTAGE OF MORE MODERN METHODS:

• The possibility of substantially higher power relative to methods that assume normality and homoscedasticity.

• More accurate confidence intervals and better control over the probability of a Type I error.

• A deeper and more accurate sense of how groups compare and how variables are related.

The are, of course, differences in the relative merits of the methods to be described, raising a natural question: WHICH METHOD IS BEST?

Certain methods provide excellent control over the Type I error probability and are more likely to have higher power compared to other techniques that might be used. But in the end, no single method dominates and different methods provide different perspectives. This raises another question:

HOW MANY METHODS DOES IT TAKE TO UNDERSTAND DATA?

Various aspects of this issue will be discussed.
THREE MAJOR INSIGHTS WITH MANY RELATED RESULTS:

• Heavy-tailed distributions (outliers are likely) are commonly encountered and can destroy power when using means or least squares regression. (Heavy-tailed distributions refers to distributions that have tails thicker than a normal distribution. An example is given momentarily.) They can result in an inaccurate sense of what is typical and they can result in seemingly small measures of effect size that are in fact large from a graphical point of view.

• Skewed distributions can result in highly inaccurate results and low power when using any method based on means. The sample size needed to assume normality, when using means, can be very large, contrary to what was once believed. In some cases a sample size greater than 300 is required, as will be illustrated. Certain classic methods can be inaccurate under general conditions regardless of how large the sample sizes might be.

• Heteroscedasticity can be much more serious than once thought.

Here the goal is to briefly summarize the many issues and techniques that have been developed. More details can be found in


BROAD GOALS

• Understand the meaning modern of robust methods.
• Understand when and why classic routinely used methods perform poorly relative to more modern methods.
• Understand why some seemingly natural methods for dealing with non-normality, outliers and heteroscedasticity are generally unsatisfactory.
• Understand modern insights and methods relevant to curvature when dealing with regression.
• Learn how to deal with non-normality, outliers and heteroscedasticity in a theoretically sound manner. Technical details are kept to a minimum, but it is important to stress what works well and what does not.
• Elaborate on what various methods tell us and what they don’t tell us.

THE THREE BASIC COMPONENTS OF THIS TALK:

• BASICS
• METHODS FOR COMPARING GROUPS
• METHODS FOR STUDYING ASSOCIATIONS INCLUDING RECENT ADVANCES RELEVANT TO ANCOVA
What is a Measure of Location?

Consider $n$ observations, which we label $X_1,\ldots,X_n$. To indicate that the values are rewritten in ascending order, we write $X_{(1)} \leq X_{(2)} \leq \ldots \leq X_{(n)}$.

A summary of data, based on the observations $X_1, \ldots, X_n$, is called a measure of location if it satisfies three properties. First, its value must lie somewhere between the smallest and largest values observed. In symbols, the measure of location must have a value between $X_{(1)}$ and $X_{(n)}$, inclusive. Second, if all observations are multiplied by some constant $c$, then the measure of location is multiplied by $c$ as well. Third, if the constant $c$ is added to all of the observations, a measure of location is increase by $c$ as well.

Example

You measure the height, in feet, of ten women yielding the values

$$5.00 \ 5.11 \ 5.20 \ 5.20 \ 5.50 \ 5.60 \ 5.70 \ 5.80 \ 5.90 \ 6.00$$

The sample mean is $\bar{X} = 5.501$. Notice that the mean cannot be less than the smallest value and it cannot be greater than the largest value. That is, it satisfies the first criterion for being a measure of location. We could get the mean in inches by multiplying each value by 12 and recomputing the average, but it is easier to simply multiply the mean by 12 yielding 66.012. If every value is increased by 2, the mean is increased by 2 as well.

Example

The median is a measure of location. Using the data in the last example, the median is

$$(5.5+5.6)/2 = 5.55.$$ Again, if we multiply all values by 12, we see that the median becomes

$$((12(5.5) + 12(5.6))/2 = 66.6,$$ which is $12(5.55)$. If 2 is added to every value, the median is now 7.55.

Example

Consider the smallest of the $n$ values, $X_{(1)}$. It is a measure of location. In the example, if we multiply all values by 12, then in particular the smallest value, 5, becomes 60. If we add 2 to every value, in particular $X_{(1)}$ is increased by 2.
Quartiles are a measure of location.

The definition extends to the population mean, $\mu$, the median and quartiles and all quantiles. (The value $c$ is the .75 quantile or upper quartile if $P(X \leq c) = .75$.}
WHAT IS MEANT BY ROBUST METHODS?

MAINTAIN RELATIVELY HIGH POWER WHEN DEALING WITH NON-NORMAL DISTRIBUTIONS OR HETEROSEDASTICITY.

ACHIEVE ACCURATE CONFIDENCE INTERVALS.

MORE THAN JUST CONTROLLING THE TYPE I ERROR PROBABILITY. (A METHOD THAT CONTROLS THE TYPE I ERROR PROBABILITY IS SAID TO BE LEVEL ROBUST.)

ROUGHLY, SMALL CHANGES IN A DISTRIBUTION SHOULD NOT HAVE A LARGE IMPACT ON THE METHOD BEING USED.

FOR EXAMPLE, SMALL DEPARTURES FROM NORMALITY SHOULD NOT DESTROY POWER.

SMALL CHANGES IN A DISTRIBUTION SHOULD NOT HAVE A LARGE IMPACT ON THE PARAMETER BEING USED SUCH AS THE POPULATION MEAN AND VARIANCE.

ROUGHLY, WANT SUMMARIES OF DATA AND INFERENTIAL METHODS THAT PERFORM WELL OVER A BROAD RANGE OF SITUATIONS WITHOUT SPECIFYING OR ASSUMING SOME PARAMETRIC FAMILY OF DISTRIBUTIONS SUCH AS THE FAMILY OF NORMAL DISTRIBUTIONS.

REGARDING MEASURES OF LOCATION BASED ON OBSERVED DATA, A FEW VALUES SHOULD NOT BE ABLE DOMINATE RESULTING IN A MISLEADING INDICATION OF THE TYPICAL RESPONSE.


EXAMPLE

In a sexual attitude study conducted by Pedersen et al. (2002), 105 young males were asked how many sexual partners they desire over the next 30 years. The sample mean is 64.9. This is, however, a dubious indication of the typical response because 97% of the responses are less than the sample mean. This occurs because of a single extreme response: 6000. If this extreme response is removed, now the mean is 7.9. But even 7.9 is rather misleading because over 77% of the remaining observations are less than 7.9.
Breakdown Point

Unusually large or small values are called outliers. One way of quantifying the sensitivity of the sample mean to outliers is with the so-called finite sample breakdown point, which refers to the smallest proportion of observations that must be altered to destroy a measure of location or some measure of variation.

MEAN: The finite sample breakdown point of the sample mean is the smallest proportion of observations that can make it arbitrarily large or small. Said another way, the finite sample breakdown point of the sample mean is the smallest proportion of \( n \) observations that can render it meaningless. A single observation can make the sample mean arbitrarily large or small, regardless of what the other values might be, so its finite sample breakdown point is \( 1/n \).

MEDIAN: breakdown point is approximately \( .5 \). About half of the data must be changed to make the median arbitrarily large

SAMPLE VARIANCE: Its breakdown point is \( 1/n \).

EXAMPLE: In the sexual attitude study the median is 1, which is the most common response.

IS THE POPULATION VARIANCE ROBUST? NO AS WILL BE ILLUSTRATED NEXT

HEAVY-TAILED DISTRIBUTIONS: Refers to distributions that have thicker tails than the normal. Samples from such distributions are characterized by outliers.
Figure 1: For normal distributions, increasing the standard deviation from 1 to 1.5 results in a substantial change in the distribution. But when considering non-normal distributions, seemingly large differences in the variances does not necessarily mean that there is a large difference in the graphs of the distributions. The two curves shown here have an obvious similarity, yet the variances are 1 and 10.9.
Figure 2: Two probability curves having equal means and variances.
Figure 3: Two probability curves having equal means and variances.
Figure 4: The population mean can be located in the extreme portion of the tail of a distribution. That is, the mean can represent a highly atypical response.
Figure 5: In the left panel, power is .96 based on Student’s T, $\alpha = .05$. But in the left panel, power is only .28, illustrating the general principle that slight changes in the distributions being compared can have a large impact on the ability to detect true differences between the population means.
TO ILLUSTRATE ONE ASPECT REGARDING HOW THE MEAN AND MEDIAN COMPARE. FIRST CONSIDER NORMALITY.

GENERATED 30 OBSERVATIONS FROM A NORMAL DISTRIBUTION, COMPUTED THE MEAN AND MEDIAN, REPEATED THIS 10,000 TIMES. BOXPLOTS OF THESE MEANS AND MEDIANS, SHOWN IN FIGURE 6, ILLUSTRATE THE INCREASED ACCURACY OF THE MEAN OVER THE MEDIAN.

BUT NOW SUPPOSE SAMPLING IS FROM A MIXED NORMAL OR A SKEWED DISTRIBUTION FOR WHICH OUTLIERS TEND TO BE RELATIVELY COMMON. REPEATING THE SIMULATION, NOW THE MEAN PERFORMS POORLY AS INDICATED BY THE BOXPLOTS IN FIGURE 7.

ILLUSTRATIONS USING DATA FROM ACTUAL STUDIES:

A possible criticism of the non-normal distributions used to create Figure 7 is that perhaps they do not reflect realistic departures normality. Accordingly, the simulations are repeated, this time sampling (with replacement) 30 values stemming from two different studies. First, data from the Well Elderly 2 study (Clark et al., 1997; Jackson et al., 2009) were used. A portion of the study dealt with the cortisol awakening response (CAR) and its association with various measures of stress, where CAR refers to the change in cortisol measured upon awakening and again 30-60 minutes later. A boxplot of the CAR values indicates an approximately symmetric distribution with outliers. The left panel of Figure 8 shows boxplots of the resulting means and medians. Again there is less variability among the medians. The standard deviation of the means was .09 compared to .05 for the medians.

The next example is based on data from a study aimed at predicting reading ability in children. The right panel of Figure 7 shows boxplots of the means and medians for a measure of the accuracy of identifying lower case letters. Again samples of 30 were used. Once more, there is less variability among the medians. The standard deviation of the means was 4.41 compared to .52 for the medians. That is, the standard error of the mean is nearly nine times larger than the standard error of the medians.

NOTE IMPLICATIONS ABOUT REPLICATING A STUDY.
Figure 6: Boxplots of means, 20% trimmed means and medians when sampling from a normal distribution.
Figure 7: Boxplots of 10,000 means, 20% trimmed means and medians using data sampled from a mixed normal distribution.
Figure 8: Boxplots of means and medians using data from two different studies.
DEALING WITH OUTLIERS: TWO STRATEGIES THAT ARE REASONABLE, ONE THAT IS RELATIVELY INEFFECTIVE AND ANOTHER SEEMINGLY NATURAL STRATEGY THAT SHOULD NEVER BE USED.

1. TRIM. BEST-KNOWN EXAMPLE IS THE MEDIAN, BUT TYPICALLY IT TRIMS TOO MUCH.

2. REMOVE OUTLIERS AND AVERAGE THE VALUES THAT REMAIN.

3. TRANSFORM THE DATA, FOR EXAMPLE TAKE LOGARITHMS, BUT THIS APPROACH PERFORMS POORLY.

4. HIGHLY UNSATISFACTORY STRATEGY: DISCARD OUTLIERS AND APPLY STANDARD HYPOTHESIS TESTING METHODS FOR MEANS TO THE REMAINING DATA. IF, FOR EXAMPLE THE SAMPLE SIZE IS REDUCED FROM $n$ TO $m$ AFTER TRIMMING, USE A METHOD FOR MEANS THAT ASSUMES WE HAVE $m$ OBSERVATIONS. THIS RESULTS IN USING THE WRONG STANDARD ERROR. METHODS THAT REMOVE OUTLIERS CAN BE USED, BUT IT IS IMPERATIVE THAT A CORRECT ESTIMATE OF THE STANDARD BE USED, WHICH DEPENDS ON HOW OUTLIERS ARE TREATED. (DETAILS ARE GIVEN LATER.)

WHY IS IT TECHNICALLY UNSOUND TO DISCARD OUTLIERS AND APPLY STANDARD METHODS BASED ON MEANS USING THE REMAINING DATA? The reason is that the remaining data are not independent: they are correlated, which invalidates the usual derivation of the standard error. Note if $X_1$ and $X_2$ are independent, knowing that $X_1 = 6$, for example, tells nothing about the probability that $X_2$ is less than 6. But suppose we put the observations in ascending order and we are told that the smallest value is $X_{(1)} = 6$. Then we can say that the probability that the next largest value, $X_{(2)}$ cannot be less than 6. That is, $X_{(1)}$ and $X_{(2)}$ are dependent.

ALL INDICATIONS ARE THAT TYPICALLY, A 20% TRIMMED MEAN IS A GOOD CHOICE FOR GENERAL USE. MORE DETAILS WILL BE COVERED LATER IN THE COURSE. BUT THERE ARE ALWAYS EXCEPTIONS. NO SINGLE METHOD IS ALWAYS BEST. ONLY EFFECTIVE WAY TO DETERMINE WHETHER ANOTHER CHOICE MAKES A PRACTICAL DIFFERENCE IS TO TRY IT.

COMPUTING A 20% TRIMMED MEAN: Compute $g = .2n$ and round down to the nearest integer. Remove the $g$ smallest and the $g$ largest and average the remaining data.
DETECTING OUTLIERS AND RELATED MEASURES OF LOCATION

METHOD GENERALLY AGREED TO BE LEAST SATISFACTORY:

Use the sample mean $\bar{X}$ and the sample standard deviation. For example, declare the value $X$ an outlier if

$$\frac{|X - \bar{X}|}{s} > 2.$$  (1)

SUFFERS FROM MASKING. REASON: THE SAMPLE VARIANCE HAS A BREAKDOWN POINT OF ONLY $1/n$. EVEN A SINGLE OUTLIER CAN INFLATE $s$, WHICH MAKES THE LEFT SIDE OF THE LAST EQUATION SMALL.

BOXPLOT RULE (USING IDEAL FOURTHS) The boxplot is based on the interquartile range, which has a breakdown point of .25.

The R function

$$\text{outbox}(x, \text{mbox}=F),$$

MAD-MEDIAN RULE

MAD IS THE MEDIAN OF $|X_1 - M|, \ldots, |X_n - M|$. MADN is MAD/.6745 and $M$ is the usual sample median; MADN estimates $\sigma$ under normality and its breakdown point is .5.

THE MAD-MEDIAN RULE declares $X$ an outlier if

$$\frac{X - M}{MADN} > 2.24$$

MODIFIED ONE-STEP M-ESTIMATOR: REMOVE VALUES DECLARED OUTLIERS VIA THE MAD-MEDIAN RULE, AVERAGE THE REMAINING VALUES. Breakdown=.5.

The R function

$$\text{out}(x)$$

checks for outliers using the MAD-median rule.

EXAMPLE

Consider the values
The value 100,000 is not declared an outlier using the mean and standard deviation.

For the same data, using the R function outbox to detect outliers, the value 100,000 is declared an outlier. Same is true using MAD-MEDIAN RULE

EXAMPLE

Cortisol upon awakening (Well Elderly study). $n = 460$. The mean and standard deviation detect 7 outliers (using the R function outms) Boxplot detects 19 outliers (using outbox) and the MAD-MEDIAN RULE detects 34 (using the R function out). The MAD-MEDIAN RULE has a higher breakdown point than the boxplot rule.

ONE-STEP M-ESTIMATOR:

$X$ is declared an outlier if

$$\frac{|X - M|}{\text{MADN}} > 1.28.$$  \hfill (2)

If the number of small outliers ($\ell$) is not equal to the number of large outliers ($u$), the one-step M-estimator is given by

$$\bar{X}_{os} = \frac{1.26(\text{MADN})(u - \ell)}{n - u - \ell} + \frac{1}{n - u - \ell} \sum_{i=\ell+1}^{n-u} X_{(i)}.$$  \hfill (3)

Breakdown=.5.

The R function

\texttt{onestep(x)}

computes the one-step M-estimator just illustrated. And the function

\texttt{mom(x)}

computes the modified one-step M-estimator.
n=40, can assume normality when using a mean?

NO: two major insights, one of which is particularly devastating.

Early studies considered unusually light-tailed distributions. And the implicit assumption was that if the sample mean has approximately a normal distribution, the usual T test statistic will have approximately Student’s t distribution. This is not necessarily true, as will be illustrated. In practical terms, when using T, might need \( n > 300 \). Some of these illustrations are based on a lognormal distribution, which is shown in Figure 9. This distribution is light-tailed in the sense that the expected proportion of points declared outliers is relatively small. Even with \( n = 25 \), the sample mean has, approximately, a normal distribution. But look at Figure 10, which shows the distribution of Student’s T when sampling from a lognormal distribution. To get a reasonably accurate .95 confidence interval, \( n = 200 \) is required.

When sampling from a skewed, heavy-tailed distribution, Student’s T deteriorates even further as illustrated in Figure 11.

But what about real data? Do problems persist? The answer is yes as illustrated in Figure 12 using hangover data. Figure 13 shows the distribution of T when dealing with the sexual data. Removing the extreme outlier, the distribution of T is shown in Figure 14.
Figure 9: Shown is a lognormal distribution, which is skewed and relatively light-tailed, roughly meaning that the proportion of outliers found under random sampling is relatively small.
Figure 10: The left panel shows the distribution of 5000 $T$ values, with each $T$ value based on 20 observations generated from the lognormal distribution. The symmetric solid line is the distribution of $T$ under normality. The right panel is the same as the left, only now the sample size is $n = 100$. 
Figure 11: The asymmetric curve is the distribution of $T$ when sampling 20 observations from a contaminated lognormal distribution, which is heavy-tailed. The symmetric curve is the distribution of $T$ under normality.
Figure 12: The distribution of $T$ when randomly sampling from the hangover data. The symmetric curve is the distribution of $T$ under normality.
Figure 13: The distribution of $T$ when randomly sampling from the sexual attitude data.
Figure 14: The distribution of $T$ when randomly sampling from the sexual attitude data with the extreme outlier removed.
2 TRIMMED MEANS

TRIMMING REDUCES PROBLEMS ASSOCIATED WITH SKEWED DISTRIBUTIONS AND LOW POWER DUE TO OUTLIERS.

BASIC ISSUE: ESTIMATING THE SQUARED STANDARD ERROR OF THE TRIMMED MEAN

RECALL ESTIMATE OF \( \text{VAR}(\bar{X}) = \frac{s^2}{n} \)

FOR A 20% TRIMMED MEAN, WHICH HAS A BREAKDOWN POINT OF .2, ESTIMATE IS

\[
\frac{s_w^2}{6^2n},
\]

where \( s_w^2 \) is the 20% Winsorized sample variance.

WINSORIZING Trimming means to remove the smallest \( g \) and largest \( g \) values. For 20% trimming, \( g \) is \(.2n \) rounded down to the nearest integer. Winsorizing data means that rather than trim the smallest \( g \) values, set them equal to the smallest value not trimmed. Similarly, largest \( g \) values, are set equal to the largest value not trimmed. The mean of Winsorized data is called the Winsorized mean, and variance of the Winsorized data is called the squared Winsorized variance.

Simply estimating the standard error using the sample variance applied to the data left after trimming is technically unsound and can differ substantially from the estimator just described, which is motivated by theory and supported by simulations.

MORE BROADLY, SIMPLY DISCARDING OUTLIERS AND APPLYING METHODS FOR MEANS TO THE REMAINING DATA CAN RESULT IN HIGHLY INACCURATE RESULTS. STANDARD ERRORS DEPEND ON THE METHOD USED TO DETECT AND REMOVE THE INFLUENCE OF OUTLIERS.

EXAMPLE

For sexual attitude data, \( n = 105 \) males, the estimated standard error of the 20% trimmed mean is .53.

Imagine we use the method for the sample mean on the remaining 63 values left after trimming. That is, we compute \( s \) using these 63 values only and then compute \( s/\sqrt{63} \). This yields 0.28, which is about half of the value based on a theoretically sound technique.
The R function

```r
trimse(x, tr = .2)
```

estimates the standard error of a trimmed mean.
A CONFIDENCE INTERVAL FOR THE POPULATION TRIMMED MEAN: TUKEY-MCLAUGHLIN METHOD

\[
\left( \bar{X}_t - c \frac{s_w}{(1-2G)\sqrt{n}}, \bar{X}_t + c \frac{s_w}{(1-2G)\sqrt{n}} \right),
\]

where \( G \) is the proportion of values trimmed and \( c \) is the \( 1 - \alpha/2 \) quantile of the Student’s t distribution with \( h-1 \) degrees of freedom, where \( h = n-2G \) is the number of observations left after trimming. With 20% trimming, a \( 1-\alpha \) confidence interval is given by

\[
\left( \bar{X}_t - c \frac{s_w}{.6\sqrt{n}}, \bar{X}_t + c \frac{s_w}{.6\sqrt{n}} \right). \tag{4}
\]

The R function

\[ \text{trimci}(x, \text{tr}=.2, \text{alpha}=.05, \text{nv}=0) \]

computes a \( 1 - \alpha \) confidence interval for \( \mu_t \). The amount of trimming, indicated by the argument \( \text{tr} \), defaults to 20%. If the argument \( \text{alpha} \) is unspecified, \( \alpha = .05 \) is used. The argument \( \text{nv} \) indicates the null value when testing some hypothesis.

Figure 15 shows the distributions of \( T_t \) (left panel) and \( T \) (right panel) based on \( n = 20 \) and when sampling from the lognormal distribution in Figure 9 (a skewed, relatively light-tailed distribution). Also shown is the approximation of the distribution assuming normality. Note that \( T_t \) is better approximated by a Student’s t distribution compared to \( T \).

Figure 16 shows an estimate of the distribution of \( T \) for some cortisol data. Could the approximation be misleading? Yes, problems with \( T \) are probably being underestimated.
Figure 15: The left panel shows the actual distribution of $T_i$ and the approximation based on Student’s $t$ distribution when sampling from the lognormal distribution in Figure 9, $n = 20$. The right panel shows the actual distribution of $T$. In practical terms, using Student’s $t$ to compute confidence intervals for the 20% trimmed mean tends to be more accurate than using Student’s $t$ to compute a confidence interval for the mean.
EXAMPLE

Suppose a test of open mindedness is administered to 10 participants yielding the observations

\[5, 60, 43, 56, 32, 43, 47, 79, 39, 41.\]

\[\bar{X}_t = 44.8\] and \[\bar{X} = 44.5.\] Confidence interval for the trimmed mean (using the R function \text{trimci}) is

\[44.8 \pm 2.57 \frac{7.385}{0.6\sqrt{10}} = (34.8, 54.8).\]

In contrast, the 0.95 confidence interval for the mean is (30.7, 58.3). The ratio of the lengths of the confidence intervals is \((54.8 - 34.8)/(58.3 - 30.7) = .72.\) That is, the length of the confidence interval based on the trimmed mean is substantially shorter.

NOTE IMPLICATION REGARDING POWER.

Here is how it is done in R:

\[x = c(5, 60, 43, 56, 32, 43, 47, 79, 39, 41)\]

\[\text{trimci}(x)\]

\[\text{trimci} \text{ defaults to 20\% trimming. To compute a confidence interval based on the mean, use}\]

\[\text{trimci}(x, tr=0)\]

For 10\% trimming

\[\text{trimci}(x, tr=.1)\]

HYPOTHESIS TESTING

Test statistic is

\[T_t = \frac{0.6(\bar{X}_t - \mu_0)}{s_w/\sqrt{n}},\]  \hspace{1cm} (5)

where \(\mu_0\) is some specified value of interest.
Figure 16: The distribution of $T$ based on cortisol measures upon awakening, n=87
MEDIAN: REQUIRES SPECIAL METHODS

FOR DISTRIBUTION FREE CONFIDENCE INTERVAL, USE THE R FUNCTION

\[ \text{sint}(x, \alpha = .05) \]

The R function

\[ \text{msmedse}(x) \]

computes an estimate of the standard error of sample median, \( M \)

BUT THIS ESTIMATOR, AS WELL AS ALL OTHERS THAT HAVE BEEN PROPOSED, CAN PERFORM POORLY WHEN TIED (DUPLICATED) VALUES OCCUR.

PRACTICAL IMPLICATION: NEED ALTERNATIVE METHOD FOR COMPARING GROUPS BASED ON MEDIAN.

WE CAN DEAL WITH TIED VALUES USING A PERCENTILE BOOTSTRAP METHOD, WHICH WILL BE DESCRIBED LATER.

NOTE: STUDENT’S T CAN BE BIASED. THAT IS, PROBABILITY OF REJECTING IS NOT MINIMIZED WHEN THE NULL HYPOTHESIS IS TRUE. This is illustrated in Figure 17 when sampling from a skewed, light-tailed distribution.

COMMENTS ON M-ESTIMATOR AND MODIFIED ONE-STEP M-ESTIMATOR

STANDARD ERROR OF M-ESTIMATOR CAN BE ESTIMATED. BUT RESULTING TEST STATISTIC IS UNSATISFACTORY, IN TERMS OF TYPE I ERRORS, WHEN DEALING WITH SKEWED DISTRIBUTIONS. NEED TO USE A PERCENTILE BOOTSTRAP TO BE DESCRIBED. SAME IS TRUE WHEN USING THE MODIFIED ONE-STEP M-ESTIMATOR.
Figure 17: Power curve of Student’s T when sampling from a lognormal distribution, \( n = 20 \), \( \alpha = 0.05 \). The null hypothesis corresponds to \( \text{delta}=0 \). Ideally the power curve should be strictly increasing as \( \text{delta} \) gets large.
3  BOOTSTRAP METHODS

A COLLECTION OF NONPARAMETRIC TECHNIQUES THAT CAN BE USED TO COMPUTE CONFIDENCE INTERVALS, TEST HYPOTHESES AND SOLVE OTHER STATISTICAL PROBLEMS. THE BASIC STRATEGY IS TO RUN SIMULATIONS USING THE OBSERVED DATA, AS WILL BE ILLUSTRATED.

WHAT DOES NONPARAMETRIC MEAN? CONSIDER, FOR EXAMPLE, THE ONE-SAMPLE STUDENT’S T TEST. CAN DETERMINE A CRITICAL VALUE BY IMPOSING A PARTICULAR PARAMETRIC ASSUMPTION ABOUT THE DISTRIBUTION UNDER STUDY: NORMALITY. NORMAL DISTRIBUTIONS ARE COMPLETELY DETERMINED BY TWO UNKNOWN PARAMETERS: THE POPULATION MEAN AND POPULATION VARIANCE.

IN CONTRAST, BOOTSTRAP METHODS DO NOT ASSUME THAT THE POPULATION DISTRIBUTION IS DETERMINED BY A SPECIFIED CLASS OF DISTRIBUTIONS CHARACTERIZED BY UNKNOWN PARAMETERS SUCH AS THE POPULATION MEAN AND POPULATION VARIANCE.

NOTE: BOOTSTRAP METHODS ARE NOT DISTRIBUTION FREE. A DISTRIBUTION FREE METHOD MEANS THAT WHEN TESTING HYPOTHESES, CONTROL OVER THE TYPE I ERROR PROBABILITY CAN BE DETERMINED EXACTLY ASSUMING RANDOM SAMPLING ONLY. AN EXAMPLE OF A DISTRIBUTION FREE METHOD WILL BE GIVEN LATER. (THE SHIFT FUNCTION, TO BE DESCRIBED.) THE R FUNCTION sint COMPUTES A DISTRIBUTION FREE CONFIDENCE INTERVAL FOR THE POPULATION MEDIAN.

BOOTSTRAP METHODS ARE NOT A PANACEA FOR DEALING WITH THE MANY PRACTICAL PROBLEMS ENCOUNTERED WHEN TRYING TO UNDERSTAND DATA.

BUT THEY HAVE CONSIDERABLE PRACTICAL VALUE FOR A WIDE RANGE OF SITUATIONS.

UNDERSTANDING THE BASICS:

QUICK REVIEW OF SAMPLING DISTRIBUTIONS

If the study were replicated many times, and the sample mean computed each time, the results could be summarized as follows:
Student’s t

In a similar manner Student’s t has a sampling distribution. It is the distribution of t over many studies.

First consider *Determining Critical Values via Simulation Studies*

If the goal is to have a Type I error probability of 0.05, this can be accomplished if the distribution of

\[ T = \frac{\bar{X} - \mu_0}{s/\sqrt{n}} \]  

(6)

can be determined.

In particular, if

\[ P(T \leq t) = 0.975. \]

a 0.95 confidence interval for \( \mu \) is given by

\[ \left( \bar{X} - t \frac{s}{\sqrt{n}}, \bar{X} + t \frac{s}{\sqrt{n}} \right). \]

Conceptually, when the null hypothesis is true, \( t \) can be determined to a high degree of accuracy if a study could be repeated millions of times. If, for example, it were repeated 1000 times, put the the resulting \( T \) values in ascending order and label the results \( T_{(1)} \leq \cdots \leq T_{(1000)}. \) Then one way of estimating \( t \) is with \( T_{(975)}, \) simply because 97.5% of the \( T \) values are less than or equal to \( T_{(975)}. \)

Here is how we would use R to determine \( t \) if sampling from a lognormal distribution:
• Generate \( n \) observations from the distribution of interest. (For a lognormal distribution the R function \texttt{rlnorm} accomplishes this goal.)

• Compute the sample mean \( \bar{X} \), the sample standard deviation \( s \), and then \( T \). (The population mean of a lognormal distribution is approximately 1.65.)

• Repeat steps 1 and 2 many times. For illustrative purposes, assume they are repeated 1000 times yielding 1000 \( T \) values: \( T_1, \ldots, T_{1000} \).

• Put these values in ascending order yielding \( T_{(1)} \leq \cdots \leq T_{(1000)} \).

• An estimate of the .975 quantile of the distribution of \( T \) is \( T_{(975)} \) again because the proportion of observed \( T \) values less than or equal to \( T_{(975)} \) is 0.975. If, for example, \( T_{(975)} = 2.1 \), this suggests using \( t = 2.1 \) when computing a 0.95 confidence interval.

But this assumes \( T \) has a symmetric distribution. Alternative solution: also estimate the .025 quantile with \( T_{(25)} \). Call this \( t_\ell \) and label \( T_{(975)} t_u \).

\[
\left( \bar{X} - T_{(u)}^* \frac{s}{\sqrt{n}}, \bar{X} - T_{(\ell+1)}^* \frac{s}{\sqrt{n}} \right).
\]

CALLED AN EQUAL-TAILED CONFIDENCE INTERVAL

NOTE: It might appear that \( T_{(u)}^* \) should be used to compute the upper end of the confidence interval, not the lower end, but a little algebra shows that this is not the case.

BUT WE DON’T KNOW THE DISTRIBUTION WE ARE SAMPLING FROM.

HOWEVER, WE HAVE AN ESTIMATE IT BASED ON THE OBSERVED DATA.

SO CONDUCT A SIMULATION ON THE OBSERVED DATA TO DETERMINE DISTRIBUTION OF \( T \).

THIS IS THE STRATEGY USED BY THE BOOTSTRAP-T METHOD.

SYMMETRIC CONFIDENCE INTERVAL

Let

\[
T^* = \frac{|\bar{X}^* - \bar{X}|}{s^*/\sqrt{n}} \tag{7}
\]

and reject \( H_0 : \mu = \mu_0 \) if \( |T| \geq T_{(c)}^* \), where \( c = (1 - \alpha)B \) rounded to the nearest integer and again \( T_{(1)}^* \leq \cdots \leq T_{(B)}^* \) are the \( B \) bootstrap \( T^* \) values written in ascending order. An approximate \( 1 - \alpha \) confidence interval for \( \mu \) is now given by

\[
\bar{X} \pm T_{(c)}^* \frac{s}{\sqrt{n}}. \tag{8}
\]
Table 1: Actual Type I error probabilities for three methods based on the mean when testing at the $\alpha = .05$ level

<table>
<thead>
<tr>
<th>Method</th>
<th>Dist.</th>
<th>BT</th>
<th>SB</th>
<th>T</th>
</tr>
</thead>
<tbody>
<tr>
<td>$n = 20$</td>
<td>N</td>
<td>.054</td>
<td>.051</td>
<td>.050</td>
</tr>
<tr>
<td></td>
<td>LN</td>
<td>.078</td>
<td>.093</td>
<td>.140</td>
</tr>
<tr>
<td></td>
<td>MN</td>
<td>.100</td>
<td>.014</td>
<td>.022</td>
</tr>
<tr>
<td></td>
<td>SH</td>
<td>.198</td>
<td>.171</td>
<td>.202</td>
</tr>
<tr>
<td>$n = 100$</td>
<td>N</td>
<td>.048</td>
<td>.038</td>
<td>.050</td>
</tr>
<tr>
<td></td>
<td>LN</td>
<td>.058</td>
<td>.058</td>
<td>.083</td>
</tr>
<tr>
<td></td>
<td>MN</td>
<td>.092</td>
<td>.018</td>
<td>.041</td>
</tr>
<tr>
<td></td>
<td>SH</td>
<td>.168</td>
<td>.173</td>
<td>.190</td>
</tr>
</tbody>
</table>

N=Normal, LN=Lognormal, MN=Mixed normal, SH=Skewed, heavy-tailed, BT=Equal-tailed, bootstrap-t, SB=Symmetric bootstrap-t, T=Student’s T

4 PERCENTILE BOOTSTRAP METHOD

NOT RECOMMENDED WHEN TESTING HYPOTHESES ABOUT MEANS, BUT EASIEST TO EXPLAIN WHEN WORKING WITH THE MEAN. WORKS VERY WELL WITH A 20% TRIMMED MEAN OR MEDIAN.

GOAL: TEST

$$H_0 : \mu = \mu_0,$$

where $\mu_0$ is some hypothesized value.

STRATEGY:

ESTIMATE THE PROBABILITY THAT A BOOTSTRAP SAMPLE MEAN IS GREATER THAN $\mu_0$.

IN SYMBOLS, THE GOAL IS TO DETERMINE

$$p = P(\bar{X}^* > \mu_0),$$

the probability that the bootstrap sample mean is greater than the hypothesized value.

Here are the steps used by percentile bootstrap method when dealing with the mean:
1. Generate a bootstrap sample $X_1^*, \ldots, X_n^*$. (Randomly sample with replacement $n$ observations from $X_1, \ldots, X_n$.)

2. Compute the mean of this bootstrap sample, $\bar{X}^*$.

3. Repeat steps 1 and 2 $B$ times yielding $\bar{X}_1^*, \ldots, \bar{X}_B^*$.

4. Estimate $p = P(\bar{X}^* > \mu_0)$ with $\hat{p}$, the proportion of bootstrap sample means greater than $\mu_0$.

5. To get a .95 confidence interval, put the bootstrap sample means in ascending order. A .95 confidence interval corresponds to middle 95%.

The percentile bootstrap p-value when testing for exact equality is

$$ P = 2\min(\hat{p}, 1 - \hat{p}). $$

That is, the p-value is either $2\hat{p}$ or $2(1 - \hat{p})$, whichever is smallest.

The R function

```
onesampb(x, est = onestep, alpha = 0.05, nboot = 2000)
```

computes a confidence interval based on the one-step M-estimator, where x is an R variable containing data, alpha is $\alpha$, which defaults to 0.05, and nboot is $B$, the number of bootstrap samples to be used, which defaults to 2000. (This function contains two additional arguments, the details of which can be found in my books.) This function can be used with any measure of location via the argument est. For example,

```
onesampb(x, est = tmean)
```

would return a confidence interval based on the 20% trimmed mean.

For convenience, the R function

```
momci(x, alpha=.05, nboot=2000)
```

computes a confidence interval based on the modified one-step M-estimator and the function

```
trimpb(x, tr=.2, alpha=.05, nboot=2000)
```

uses a trimmed mean. The argument tr indicates the amount of trimming and defaults to 0.2 if not specified. Again, alpha is $\alpha$ and defaults to 0.05. It appears that $B = 500$ suffices, in terms of achieving accurate probability coverage with 20% trimming. But to be safe, $B$
(nboot) defaults to 2,000. (An argument for using $B = 2000$ can be made along the lines used by Booth and Sarker, 1998. If $B$ is relatively small, this might result in relatively low power)

**EXAMPLE**

For 15 law schools, the undergraduate GPA of entering students, in 1973, was

$3.39 \ 3.30 \ 2.81 \ 3.03 \ 3.44 \ 3.07 \ 3.00 \ 3.43 \ 3.36 \ 3.13 \ 3.12 \ 2.74 \ 2.76 \ 2.88 \ 2.96$.

The 0.95 confidence interval returned by the R function `onesampb` is $(2.95, 3.29)$. So among all law schools, it is estimated that the typical GPA of entering students is between 2.95 and 3.29. Using the MOM-estimator instead (the R function `momci`), the 0.95 confidence interval is $(2.92, 3.35)$. Using the function `trimpb`, the .95 confidence interval for the 20% trimmed mean is $(2.94, 3.25)$. Setting the argument tr=0 in `trimpb` results in a 0.95 confidence interval for the mean: $(2.98, 3.21)$, illustrating that in some cases, switching from the mean to a 20% trimmed mean makes little difference.

**EXAMPLE**

Cortisol awakening response prior to intervention (from the Well Elderly 2 study).

Percentile bootstrap, 20% trimmed, the confidence interval is $(-0.041, 0.00765)$ compared to $(-0.058, 0.030)$, the confidence interval for the mean using Student’s T, which is nearly twice as long as the confidence interval based on the 20% trimmed. The p-value based on a 20% trimmed mean, using difference scores, is 0.19. Comparing instead the marginal trimmed means (cortisol upon awakening 1 vs. cortisol 30-60 minutes later), $p=0.005$.  

40
5 BOOTSTRAP-T METHOD BASED ON TRIMMED MEANS

If the amount of trimming is 20% or more, it seems that using a percentile bootstrap method is best for general use, but with the amount of trimming close to zero, it currently seems that using a bootstrap-t method is preferable. (With 10% trimming, it is unclear whether a bootstrap-t is preferable to a percentile bootstrap method.) The computational steps of the most basic version of the bootstrap-t method are summarized here when using a trimmed mean.

Generate a bootstrap sample of size $n$ and compute the trimmed mean and Winsorized standard deviation, which we label $\bar{X}_t^*$ and $s^*_w$, respectively. Let $\gamma$ be the amount of trimming. With 20% trimming, $\gamma = 0.2$ and with 10% trimming, $\gamma = 0.1$. Next, compute

$$T_t^* = \frac{(1 - 2\gamma)(\bar{X}_t^* - \bar{X}_t)}{s^*_w / \sqrt{n}}.$$ (9)

Repeating this process $B$ times yields $B$ $T_t^*$ values. Writing these $B$ values in ascending order we get $T_{t(1)}^* \leq T_{t(2)}^* \leq \cdots \leq T_{t(B)}^*$. Letting $\ell = .025 B$, rounded to the nearest integer, and $u = B - \ell$, an estimate of the 0.025 and 0.975 quantiles of the distribution of $T_t$ is $T^*_{t(\ell+1)}$ and $T^*_{t(u)}$. The resulting 0.95 confidence interval for $\mu_t$ (the population trimmed mean) is

$$\left(\bar{X}_t - T_{t(u)}^* \frac{s_w}{(1 - 2\gamma)\sqrt{n}}, \bar{X}_t - T_{t(\ell+1)}^* \frac{s_w}{(1 - 2\gamma)\sqrt{n}}\right).$$ (10)

Hypothesis Testing.

As for testing $H_0 : \mu_t = \mu_0$, compute

$$T_t = \frac{(1 - 2\gamma)(\bar{X}_t - \mu_0)}{s_w / \sqrt{n}}$$

and reject if

$$T_t \leq T_{t(\ell+1)}^*,$$

or if

$$T_t \geq T_{t(u)}^*.$$ 

The symmetric bootstrap-t method can be used as well when testing a two-sided hypothesis. Now we use

$$T_t^* = \frac{|(1 - 2\gamma)(\bar{X}_t^* - \bar{X}_t)|}{s^*_w / \sqrt{n}}.$$ (11)
Table 2: Actual Type I error probabilities using 20% trimmed means, $\alpha = .05$

<table>
<thead>
<tr>
<th>Method</th>
<th>Dist.</th>
<th>BT $n=20$</th>
<th>SB $n=20$</th>
<th>P $n=20$</th>
<th>TM $n=20$</th>
</tr>
</thead>
<tbody>
<tr>
<td>N</td>
<td>.067</td>
<td>.052</td>
<td>.063</td>
<td>.042</td>
<td></td>
</tr>
<tr>
<td>LN</td>
<td>.049</td>
<td>.050</td>
<td>.066</td>
<td>.068</td>
<td></td>
</tr>
<tr>
<td>MN</td>
<td>.022</td>
<td>.019</td>
<td>.053</td>
<td>.015</td>
<td></td>
</tr>
<tr>
<td>SH</td>
<td>.014</td>
<td>.018</td>
<td>.066</td>
<td>.020</td>
<td></td>
</tr>
</tbody>
</table>

N=Normal, LN=Lognormal, MN=Mixed normal, SH=Skewed, heavy-tailed, BT=Equal-tailed, bootstrap-t, SB=Symmetric bootstrap-t, P=Percentile bootstrap, TM=Tukey-McLaughlin

and reject $H_0$ if $|T_t| > T^*_t(c)$, where $c = (1 - \alpha)B$ rounded to the nearest integer. An approximate $1 - \alpha$ confidence interval for $\mu_t$ is

$$\bar{X}_t \pm T^*_t(c) \frac{s_w}{\sqrt{n}}.$$  \hspace{1cm} (12)

The R function

```
trimcibt(x, tr = 0.2, alpha = 0.05, nboot = 599, side = T)
```

computes a bootstrap-t confidence interval for a trimmed mean. The argument side indicates whether an equal-tailed or a symmetric confidence interval is to be computed. The default is side=T resulting in a symmetric confidence interval. Using side=F means that an equal-tailed confidence interval will be computed. The argument tr indicates the amount of trimming, which defaults to 20%. So to compute a confidence interval for the mean, set tr=0.

**EXAMPLE**

Data on the desired number of sexual partners over next 30 years among 105 college males. The Tukey-McLaughlin .95 confidence interval for the 20% is (1.62, 3.75). Using the R function trimcibt with side=F yields an equal-tailed .95 confidence interval of (1.28, 3.61). With side=T it is (1.51, 3.61). Using the percentile bootstrap method, the R function trimpb returns (1.86, 3.95). So in this particular case, the lengths of the confidence intervals do not
vary that much among the methods used, but the intervals are centered around different values, which might affect any conclusions made. If trimcibt is used to compute a 0.95 confidence for the mean (by setting the argument tr=0), the result is \((-2.46, 4704.59)\), which differs drastically for the confidence interval for a 20% trimmed mean.

In summary, all indications are that the percentile bootstrap is more stable (with at least 20% trimming) than the bootstrap-t method. That is, the actual Type I error probability tends to be closer to the nominal level. And it has the added advantage of more power, at least in some situations, compared to any other method we might choose.

However, there are situations where the bootstrap-t method outperforms the percentile method. And there are additional situations where the percentile bootstrap is best. So both methods are important to know.
6 COMPARING TWO INDEPENDENT GROUPS

FOUR GENERAL APPROACHES WHEN COMPARING TWO GROUPS:

1. Compare measures of location, such as the mean or median.
2. Compare measures of variation.
3. Focus on the probability that a randomly sampled observation from the first group is smaller than a randomly sampled observation from second group.
4. Simultaneously compare all of the quantiles to get a global sense of where the distributions differ and by how much. For example, low scoring participants in group 1 might be very similar to low scoring participants in group 2, but for high scoring participants, the reverse might be true.

CONCERNS ABOUT STUDENT’S T

EXAMPLE

Consider again the two normal distributions shown in the left panel of Figure 5. Both have variances one and the means differ by 1. As previously noted, using Student’s T with $\alpha = .05$, power is 0.96 with $n_1 = n_2 = 25$. But now look at the two contaminated normal distributions shown in the right panel of Figure 5. The difference in the means is the same as in the left panel and the plot of the distributions has an obvious similarity to the distributions in the left panel. But for the right panel, power is only 0.28. One reason power is low is that when sampling from a heavy-tailed distribution, the actual probability of a Type I error can be substantially lower than the specified $\alpha$ value. For example, if you use Student’s T with $\alpha = 0.05$, the actual probability of a Type I error can drop below .01. If an adjustment could be made so that the actual probability of a Type I error is indeed .05, power would be better, but it would still be low relative to alternative methods that are less sensitive to outliers. The reason is that in the right panel, the variances are 10.9, compared to 1 for the distributions in left panel. Said another way, outliers are more likely to occur when sampling from the distributions in the left panel which can inflate the standard deviations. And even when outliers are not a concern, having unequal variances or even different degrees of skewness can result in relatively poor power as well.
Finally, we note that when using Student’s T, even a single outlier in only one group can result in a rather undesirable property. The following example illustrates the problem.

**EXAMPLE**

Consider the following values.

\[
\begin{align*}
\text{Group 1:} & \quad 4 \ 5 \ 6 \ 7 \ 8 \ 9 \ 10 \ 11 \ 12 \ 13 \\
\text{Group 2:} & \quad 1 \ 2 \ 3 \ 4 \ 5 \ 6 \ 7 \ 8 \ 9 \ 10
\end{align*}
\]

The corresponding sample means are \( \bar{X}_1 = 8.5 \) and \( \bar{X}_2 = 5.5 \) and \( T = 2.22 \). With \( \alpha = 0.05 \), the critical value is \( t = 2.1 \), so Student’s T would reject the hypothesis of equal means and conclude that the first group has a larger population mean than the second (because the first group has the larger sample mean). Now, if we increase the largest observation in the first group from 13 to 23, the sample mean increases to \( \bar{X}_1 = 9.5 \). So the difference between \( \bar{X}_1 \) and \( \bar{X}_2 \) has increased from 3 to 4 and this would seem to suggest that we have stronger evidence that the population means differ and in fact the first group has the larger population mean. However, increasing the largest observation in the first group also inflates the corresponding sample variance, \( s_1^2 \). In particular, \( s_1^2 \) increases from 9.17 to 29.17. The result is that \( T \) decreases to \( T = 2.04 \) and we no longer reject. That is, increasing the largest observation has more of an effect on the sample variance than the sample mean in the sense that now we are no longer able to conclude that the population means differ. Increasing the largest observation in the first group to 33, the sample mean increases to 10.5, the difference between the two sample means increases to 5 and now \( T = 1.79 \). So again we do not reject and in fact our test statistic is getting smaller! This illustration provides another perspective on how outliers can mask differences between population means.

**Yuen’s Method for Trimmed Means**

Yuen (1974) derived a method for comparing the population trimmed means of two independent groups that reduces to Welch’s method for means when there is no trimming. Trimming 20% is generally a good choice, but there are exceptions.

Focusing on 20% trimming, let \( g_j = \lfloor .2n_j \rfloor \), where again \( n_j \) is the sample size associated with the \( j \)th group \( (j = 1, 2) \) and \( \lfloor .2n_j \rfloor \) is the value of \( .2n_j \) rounded down to the nearest integer. (With 10% trimming, now \( g_j = \lfloor .1n_j \rfloor \).) Let \( h_j = n_j - 2g_j \). That is, \( h_j \) is the number of observations left in the \( j \)th group after trimming. Let

\[
d_j = \frac{(n_j - 1)s_{wj}^2}{h_j(h_j - 1)},
\]

(13)

where \( s_{wj}^2 \) is the Winsorized variance for the \( j \)th group. (The amount of Winsorizing is always the same as the amount of trimming. So for a 20% trimmed mean, Winsorize 20%
as well.) Yuen’s test statistic is

\[ T_y = \frac{\bar{X}_{t1} - \bar{X}_{t2}}{\sqrt{d_1 + d_2}}. \]  (14)

The degrees of freedom are

\[ \hat{\nu}_y = \frac{(d_1 + d_2)^2}{\frac{d_1}{h_1-1} + \frac{d_2}{h_2-1}}. \]

**Hypothesis Testing.** The hypothesis of equal trimmed means, \( H_0 : \mu_{t1} = \mu_{t2} \), is rejected if

\[ |T_y| \geq t. \]

As before, \( t \) is the \( 1 - \alpha/2 \) quantile of Student’s t distribution with \( \hat{\nu}_y \) degrees of freedom.

The improvement in power, achieving accurate confidence intervals, and controlling Type I error probabilities, can be substantial when using Yuen’s method with 20% trimming rather than Welch’s test.

The R function

\[ \text{yuen}(x, y, \text{alpha}=.05, \text{tr}=.2) \]

**EXAMPLE.**

In a study of sexual attitudes, different from the one previously discussed, 1327 males and 2282 females were asked how many sexual partners they desired over the next 30 years. We can compare the means with the R function yuen by setting tr=0. (This results in using Welch’s test.) The 0.95 confidence interval for the difference between the means is \((-1491.087, 4823.244)\) and the p-value is 0.30. Given the large sample sizes, a tempting conclusion might be that power is adequate and that the groups do not differ. However, the 0.95 confidence interval for the difference between the 20% trimmed means is \((0.408, 2.109)\) with a p-value less than 0.001. In terms of Tukey’s three-decision rule, decide that males have the larger population 20% trimmed mean.

The function

\[ z=\text{fac2list}(x, g) \]
fac2list can be used to separate data into groups, where \( x \) is the R variable (often the column of a matrix or data frame) containing the data to be analyzed and \( g \) indicates the level of the corresponding value stored in \( x \).

The function

\[
z = \text{split}(x, g)
\]

can be used as well.

**EXAMPLE.**

Plasma retinol data are available from

http://lib.stat.cmu.edu/datasets/Plasma_Retinol.

Retinol is vitamin A, and plasma retinol appears to be related to the utilization of vitamin A in rats. For future reference, the variable names in the plasma retinol data set are:

1. AGE: Age (years)
2. SEX: Sex (1=Male, 2=Female).
3. SMOKSTAT: Smoking status (1=Never, 2=Former, 3=Current Smoker)
4. QUETELET: Quetelet (weight/(height^2))
5. VITUSE: Vitamin Use (1=Yes, fairly often, 2=Yes, not often, 3=No)
6. CALORIES: Number of calories consumed per day.
7. FAT: Grams of fat consumed per day.
8. FIBER: Grams of fiber consumed per day.
9. ALCOHOL: Number of alcoholic drinks consumed per week.
10. CHOLESTEROL: Cholesterol consumed (mg per day).
11. BETADIET: Dietary beta-carotene consumed (mcg per day).
12. RETDIET: Dietary retinol consumed (mcg per day).
13. BETAPLASMA: Plasma beta-carotene (ng/ml)
14. RETPLASMA: Plasma Retinol (ng/ml)

The first few lines of the data set look like this:

64 2 2 21.48380 1 1298.8 57.0 6.3 0.0 170.3 1945 890 200 915
76 2 1 23.87631 1 1032.5 50.1 15.8 0.0 75.8 2653 451 124 727
38 2 2 20.01080 2 2372.3 83.6 19.1 14.1 257.9 6321 660 328 721
40 2 2 25.14062 3 2449.5 97.5 26.5 0.5 332.6 1061 864 153 615
We compare males and females based on the daily consumption of cholesterol using Yuen’s test. Assuming the data are stored in the R variable plasma, column 2 indicates sex (male=1, female=2) and column 10 indicates the daily consumption of cholesterol. So the R command

\[
z = \text{fac2list}(\text{plasma[,10]}, \text{plasma[,2]})
\]

separates the data into groups corresponding to males and females and stores the results in z in list mode. Because male is indicated by the value 1, which is less than the value used to indicate female, z[[1]] will contain the data for the males and z[[2]] contains the data for the females. The R command

\[
yuen(z[[1]],z[[2]])
\]

will compare males to females using 20% trimmed means.
COMPARING MEDIANS

Best overall method, especially when dealing tied (duplicated) values: percentile bootstrap.

With no tied values, a non-bootstrap method can be used based on McKean-Schrader estimate of standard error works well.

NOTE: UNDER GENERAL CONDITIONS, ALL RANK-BASED METHODS PERFORM POORLY GIVEN THE GOAL OF COMPARING MEDIANS.

Point Worth Stressing

Although the median belongs to the class of trimmed means, special methods are required for comparing groups based on medians. Put another way, it might seem that Yuen’s method could be used to compare medians simply by setting the amount of trimming to 0.5. (That is, when using the R function `yuen`, set `tr=0.5`.) When the amount of trimming is 0.2, Yuen’s method performs reasonably well in terms of Type I errors and accurate confidence intervals, but as the amount of trimming gets close to 0.5, the method breaks down and should not be used. (The method used to estimate the standard error performs poorly.)

If there are no tied (duplicated) values in either group, an approach that currently seems to have practical value is as follows. Let $M_1$ and $M_2$ be the sample medians corresponding to groups 1 and 2, respectively, and let $S_1^2$ and $S_2^2$ be the corresponding McKean-Schrader estimate of the squared standard errors, which was described in Section 4.9.2. Then an approximate $1 - \alpha$ confidence interval for the difference between the population medians is

$$(M_1 - M_2) \pm c\sqrt{S_1^2 + S_2^2}$$

where $c$ is the $1 - \alpha/2$ quantile of a standard normal distribution.

The R function

```r
msmed(x,y,alpha=.05)
```

compares medians using the McKean-Schrader estimate of the standard error.

7 Percentile Bootstrap Methods for Comparing Measures of Location

Let $\bar{X}_1^*$ and $\bar{X}_2^*$ be the bootstrap means corresponding to groups 1 and 2, respectively. Reject if

$$P(\bar{X}_1^* > \bar{X}_2^*)$$
is relatively close to 0 or 1. (A p-value can be computed as indicated momentarily.)

A 1 − α confidence interval for \( \mu_1 - \mu_2 \) is computed as follows. Let

\[
D^* = \bar{X}_1^* - \bar{X}_2^*
\]

be the difference between the bootstrap means. Now suppose we repeat this process \( B \) times yielding \( D_1^*, \ldots, D_B^* \). (The software written for my books uses \( B = 2000 \).) The middle 95% of these values, after putting them in ascending order, yields a .95 confidence interval for the difference between the population means. In symbols, put the values \( D_1^*, \ldots, D_B^* \) in ascending order yielding \( D_{(1)}^* \leq \cdots \leq D_{(B)}^* \). Then an approximate \( 1 - \alpha \) confidence interval for the difference between the population means, \( \mu_1 - \mu_2 \), is

\[
(D_{(\ell+1)}, D_{(u)})
\]

where as usual \( \ell = \alpha B/2 \), rounded to the nearest integer, and \( u = B - \ell \). So for a 0.95 confidence interval, \( \ell = 0.025 B \).

**Computing a p-value**

A p-value can be computed and is based on the probability that a bootstrap mean from the first group is greater than a bootstrap mean from the second. In symbols, a p-value can be computed if we can determine

\[
p^* = P(\bar{X}_1^* > \bar{X}_2^*).
\]

The value of \( p^* \) reflects the degree of separation between the two (bootstrap) sampling distributions. If the means based on the observed data are identical, meaning that \( \bar{X}_1 = \bar{X}_2 \), then \( p^* \) will have a value approximately equal to 0.5. (Under normality, \( p^* \) is exactly equal to 0.5.) Moreover, the larger the difference between the sample means \( \bar{X}_1 \) and \( \bar{X}_2 \), the closer \( p^* \) will be to 0 or 1. If \( \bar{X}_1 \) is substantially larger than \( \bar{X}_2 \), \( p^* \) will be close to 1, and if \( \bar{X}_1 \) is substantially smaller than \( \bar{X}_2 \), \( p^* \) will be close to 0. (Hall, 1988a, provides relevant theoretical details and results in Hall 1988b are readily extended to trimmed means.) Theoretical results not covered here suggest the following decision rule when the goal is to have a Type I error probability \( \alpha \): Reject the hypothesis of equal means if \( p^* \) is less than or equal to \( \alpha/2 \), or greater than or equal to equal to \( 1 - \alpha/2 \). Said another way, if we let

\[
p^*_m = \min(p^*, 1 - p^*),
\]

meaning that \( p^*_m \) is equal to \( p^* \) or \( 1 - p^* \), whichever is smaller, then reject if

\[
p^*_m \leq \frac{\alpha}{2}.
\]

50
We do not know \( p^* \), but it can be estimated by generating many bootstrap samples and computing the proportion of times a bootstrap mean from the first group is greater than a bootstrap mean from the second. That is, if \( A \) represents the number of values among \( D_1^*, \ldots, D_B^* \) that are greater than zero, then we estimate \( p^* \) with

\[
\hat{p}^* = \frac{A}{B}. 
\]  

(18)

Finally, reject the hypothesis of equal population means if \( \hat{p}^* \) is less than or equal to \( \alpha/2 \) or greater than or equal to \( 1 - \alpha/2 \). Or setting

\[
\hat{p}_m^* = \min(\hat{p}^*, 1 - \hat{p}^*),
\]

reject if

\[
\hat{p}_m^* \leq \frac{\alpha}{2}. 
\]  

(19)

The p-value is

\[
2\hat{p}_m^* 
\]  

(Liu & Singh, 1997).

When working with the usual sample medians, a slight extension of the method must be used when there are tied values. Let \( M_1^* \) and \( M_2^* \) be the bootstrap sample medians. Now, rather than using \( p^* \) as just defined, use

\[
p^* = P(M_1^* > M_2^*) + .5P(M_1^* = M_2^*). 
\]

In words, \( p^* \) is the probability that the bootstrap sample median from the first group is larger than the bootstrap sample median from the second, plus half the probability that they are equal. So among \( B \) bootstrap samples from each group, if \( A \) is the number of times \( M_1^* > M_2^* \), and \( C \) is the number of times \( M_1^* = M_2^* \), the estimate of \( p^* \) is

\[
\hat{p}^* = \frac{A}{B} + .5\frac{C}{B}. 
\]

As before, the p-value is

\[
2\min(\hat{p}^*, 1 - \hat{p}^*). 
\]

**EXAMPLE**

As a simple illustration, imagine that we generate bootstrap samples from each group and compute the difference between the bootstrap sample means. Further imagine that this process is repeated 10 times (so \( B = 10 \) is being used) resulting in the following \( D^* \) values:
There is one positive difference, so \( \hat{p}^* = 1/10 \). The smaller of the two numbers \( \hat{p}^* \) and \( 1 - \hat{p}^* \) is 0.1. Consequently, the p-value is \( 2(0.1) = 0.2 \).

**THE SAME STRATEGY IS USED WITH OTHER MEASURES OF LOCATION.**

**WITH SKEWED DISTRIBUTIONS, WHEN COMPARING M-ESTIMATORS, THIS IS THE BEST APPROACH TO HYPOTHESIS TESTING.**

**WORKS WELL WHEN USING 20% TRIMMED MEANS**

**COMPARING MEDIANS, WORKS WELL GENERALLY AND THE ONLY KNOWN METHOD THAT WORKS WELL WHEN THERE ARE TIED VALUES.**

The R function

\[
\text{medpb2}(x, y, \alpha = 0.05, \text{nboot} = 2000, \text{SEED} = \text{T})
\]

tests the hypothesis of equal medians using the percentile bootstrap method just described. The function also returns a \( 1 - \alpha \) confidence interval for the difference between the population medians.

The R function

\[
\text{trimpb2}(x, y, \text{tr} = 0.2, \alpha = 0.05, \text{nboot} = 2000)
\]

compares trimmed means, including medians as a special case, using the percentile bootstrap method just described. Here \( x \) is any R variable containing the data for group 1 and \( y \) contains the data for group 2. The amount of trimming, \( \text{tr} \), defaults to 20\%, \( \alpha \) defaults to 0.05, and \( \text{nboot} (B) \) defaults to 2000. This function returns a p-value plus a \( 1 - \alpha \) confidence interval for the difference between the trimmed means.

The R function

\[
\text{pb2gen}(x, y, \alpha = 0.05, \text{nboot} = 2000, \text{est} = \text{onestep}, \ldots)
\]

can be used to compare groups using any measure of location in conjunction with the percentile bootstrap method. Again, \( x \) and \( y \) are any R variables containing data and \( \text{nboot} \) is \( B \), the number of bootstrap samples to be used. By default, \( B = 2000 \). The argument \( \text{est} \) indicates which measure of location is to be employed. It can be any R function that
computes a measure of location and defaults to the function \texttt{onestep}, which is the one-step M estimator.

\begin{example}

In an unpublished study by Dana (1990), the general goal was to investigate issues related to self-awareness and self-evaluation. In one portion of the study, he recorded the times individuals could keep an apparatus in contact with a specified target. Storing the data for group 1 in the R variable \texttt{G1}, and storing the data for group 2 in \texttt{G2}, the command

\[ \text{yuenbt(G1, G2)} \]

returns a 0.95 confidence interval of \((-312.5, 16.46)\). This interval contains zero so we would not reject. In terms of Tukey’s three-decision rule, we would make no decision about which group has the larger 20% trimmed mean. If we increase the number of bootstrap samples (\(B\)) by setting the argument \texttt{nboot} to 999, now the confidence interval is \((-305.7, 10.7)\). We still do not reject, but increasing \(B\) alters the confidence interval slightly. In contrast, comparing medians, the 0.95 confidence interval is \((-441.4, -28.6)\), so we reject, the only point being that even among robust estimators, the choice of method can alter the conclusions reached.

\end{example}

\section{Bootstrap-t Methods for Comparing Measures of Location}

When testing the hypothesis of equal population means, Welch’s test statistic is

\[ W = \frac{(\bar{X}_1 - \bar{X}_2)}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}} \]

The probability of a Type I error can be controlled exactly if the distribution of \(W\) (over many studies) can be determined when the null hypothesis of equal means is true. Welch’s strategy was to approximate the distribution of \(W\) with a Student’s \(t\) distribution and the degrees of freedom estimated based on the sample variances and sample sizes. The bootstrap-t strategy is to use bootstrap samples instead to estimate the distribution of \(W\) when the null hypothesis is true.

An outline of the method is as follows. Generate a bootstrap sample of size \(n_1\) from the first group and label the resulting sample mean and standard deviation \(\bar{X}_1^*\) and \(s_1^*\),
respectively. Do the same for the second group and label the bootstrap sample mean and standard deviation $\bar{X}_2^*$ and $s_2^*$. Let

$$W^* = \frac{(\bar{X}_1^* - \bar{X}_2^*) - (\bar{X}_1 - \bar{X}_2)}{\sqrt{\frac{(s_1^*)^2}{n_1} + \frac{(s_2^*)^2}{n_2}}}.$$  \hspace{1cm} (20)

Repeat this process $B$ times yielding $B$ $W^*$ values: $W_1^*, \ldots, W_B^*$. Next, put these $B$ values in ascending order, which we label $W^*_1 \leq \cdots \leq W^*_B$. Let $\ell = \alpha B / 2$, rounded to the nearest integer, and $u = B - \ell$. Then an approximate $1 - \alpha$ confidence interval for the difference between the means ($\mu_1 - \mu_2$) is

$$\left( (\bar{X}_1 - \bar{X}_2) - W^*_u \sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}, \frac{s_1^2}{n_1} + \frac{s_2^2}{n_2} \right), \frac{s_1^2}{n_1} + \frac{s_2^2}{n_2} \right), \frac{s_1^2}{n_1} + \frac{s_2^2}{n_2} \right),$$  \hspace{1cm} (21)

**METHOD READILY EXTENDED TO TRIMMED MEANS**

The R function

```
yuenbt(x, y, tr = 0.2, alpha = 0.05, nboot = 599, side=F)
```

uses a bootstrap-t method to compare trimmed means. The arguments are the same as those used by trimpb2 plus an additional argument labeled side, which indicates whether a symmetric or equal-tailed confidence interval will be used. Side defaults to F for false meaning that the equal-tailed confidence interval will be computed. Setting side equal to T yields the symmetric confidence interval.

**EXAMPLE**

**Hangover Symptoms.** The next example again illustrates that the measure of central tendency that is used can make a practical difference. Even if the focus is on a single robust measure of central tendency, the method used to compute a confidence interval or p value can matter. The data are from a study dealing with the effects of consuming alcohol on hangover symptoms. Group 1 was a control group and measures reflect hangover symptoms after consuming a specific amount of alcohol in a laboratory setting. Group 2 consisted of sons of alcoholic fathers. The sample size for both groups is 20. Comparing means, the estimated difference is 4.5, CI [-1.63, 10.73], $p = .14$. Figure 18 shows boxplots of the data. As is evident, the data are skewed with outliers. Using 20% trimmed means (R function yuenv2) yields an estimated difference of 3.7, $p = .076$, CI [-0.456, 7.788]. Note that the length of the confidence intervals differ substantially; the ratio of the lengths is .67.
Figure 18: Boxplots of hangover symptoms.
An important point here is what happens when using bootstrap methods for comparing trimmed means (and means as a special case). In some situations they offer better control over the Type I error probability (e.g., Wilcox, 2012a, b). Using a bootstrap-t technique on the hangover data (via the R function yuenbt with 20% trimming) yields CI [0.388, 11.552]. So in contrast to the non-bootstrap methods, the bootstrap-t method provides stronger evidence that the typical level of hangover symptoms for the control group is greater than the typical level for sons of alcoholic fathers, as the interval now entirely consists of positive values. Using a percentile bootstrap method with a 20% trimmed mean (and the R function trimpb2), p = .0475 and CI [0.083, 8.333] suggesting again that typical hangover symptoms are higher for the control group. The confidence interval is narrower when using the percentile bootstrap method compared to the bootstrap-t. Comparing medians (with the R function pb2gen and the argument est=hd) gives similar results: p = .038 and CI [0.095, 8.498].

TUKEY’S THREE DECISION RULE:

Before continuing, it might help to comment on the common goal of testing for exactly equality. Tukey (1991) argued that this is nonsensical because surely the null hypothesis is not true at some decimal place. Jones and Tukey (2000) suggest dealing with this issue using Tukey’s three decision rule. If the null hypothesis is rejected, make a decision about which group has the larger measure of location. If not rejected, make no decision.

EXAMPLE

In an unpublished study by Dana (1990), the general goal was to investigate issues related to self-awareness and self-evaluation. In one portion of the study, he recorded the times individuals could keep an apparatus in contact with a specified target. Storing the data for group 1 in the R variable G1, and storing the data for group 2 in G2, the command

\[
yuenbt(G1, G2)
\]

returns a 0.95 confidence interval of (−312.5, 16.46) based on 20% trimmed means. This interval contains zero so we would not reject. In terms of Tukey’s three-decision rule, we would make no decision about which group has the larger 20% trimmed mean. If we increase the number of bootstrap samples (\( B \)) by setting the argument nboot to 999, now the confidence interval is (−305.7, 10.7). We still do not reject, but increasing \( B \) alters the confidence interval slightly. In contrast, comparing medians using the McKean–Schrader estimate of the standard error (via the R function mimed) the 0.95 confidence interval is (−460.93, −9.07), so we reject, the only point being that even among robust estimators, the choice of method can alter the conclusions reached.
9  Rank-Based and Nonparametric Methods

Let $p$ be the probability that a randomly sampled observation from the first group is less than a randomly sampled observation from the second. (Momentarily, it is assumed that tied values never occur.) If the groups do not differ, then in particular it should be the case that

$$ H_0 : p = 0.5 $$

is true. The quantity $p$ has been called a *probabilistic measure of effect size*, the *probability of concordance*, the *measure of stochastic superiority* and the *common language measure of effect size*.

WILCOXON–MANN-WHITNEY IS BASED ON AN ESTIMATE OF $p$, but it does not test this hypothesis. It uses the wrong standard error when distributions differ. Rather it tests the hypothesis that distributions are identical. And generally it does not test the hypothesis of equal medians.

Currently, a method derived by Cliff seems to perform relatively well. (Another method that performs well was derived by Brunner and Munzel.) It deals effectively with both heteroscedasticity and tied values, unlike the classic Wilcoxon-Mann-Whitney test. The R function

```
cid(x,y,alpha=.05,plotit=F),
```

performs the calculations. Twelve methods were compared by Ruscio and Mullen (2012) who found that if the total sample size, $n_1 + n_2$, is greater than or equal to 60, a bootstrap method not covered here tends to provide a more accurate confidence interval than Cliff’s method.

Let

$$ p_1 = P(X_{i1} > X_{i2}), $$
$$ p_2 = P(X_{i1} = X_{i2}), $$

and

$$ p_3 = P(X_{i1} < X_{i2}). $$

The function also reports a confidence interval for $P = p_2 + 0.5p_2$, which is labeled $ci.p$. The estimate of $P$ is labeled $phat$. To get a p-value, use the function

```
cidv2(x,y,plotit=F).
```

When the argument $plotit=T$, these functions plot the distribution of $D$, where $D$ is the difference between a randomly sampled observation from the first group, minus a randomly
sampled observation from the second group. $D$ will have a symmetric distribution around zero when the distributions are identical. The plot provides perspective on the extent to which this is the case.
COMPARING ALL QUANTILES SIMULTANEOUSLY

Roughly, when comparing medians, the goal is to compare the central values of the two distributions. But an additional issue is how low scoring individuals in the first group compare to low scoring individuals in the second. And in a similar manner, how do relatively high scores within each group compare? A way of addressing this issue is to compare the 0.25 quantiles of both groups as well as the 0.75 quantiles. Or to get a more detailed sense of how the distributions differ, all of the quantiles might be compared. There is a method for comparing all quantiles in a manner that controls the probability of a Type I error exactly assuming random sampling only. The method was derived by Doksum and Sievers (1976) and is based on an extension of the Kolmogorov-Smirnov method. Complete computational details are not provided, but a function that applies the method is supplied and illustrated next.

The R function

\[
\text{sbond}(x,y, \text{flag} = F, \text{plotit} = T, \text{xlab} = "x (First Group)", \text{ylab} = "Delta")
\]

computes confidence intervals for the difference between the quantiles using the data stored in the R variables x and y. Moreover, it plots the estimated differences as a function of the estimated quantiles associated with the first group, the first group being the data stored in the first argument, x. This difference between the quantiles, viewed as a function of the quantiles of the first group, is called a shift function. To avoid the plot, set the argument plotit=F.

EXAMPLE

In a study by J. Victoroff et al. (2008), 52 14-year-old refugee boys in Gaza were classified into one of two groups according to whether a family member had been wounded or killed by an Israeli. One issue was how these two groups compare based on a measures of depression. In particular, among boys with relatively high depression, does having a family member killed or wounded have more of an impact than among boys with relatively low measures of depression? Here is a portion of the output from sband:

<table>
<thead>
<tr>
<th>qhat</th>
<th>lower</th>
<th>upper</th>
</tr>
</thead>
<tbody>
<tr>
<td>[1,]</td>
<td>0.03448276</td>
<td>NA 18</td>
</tr>
<tr>
<td>[2,]</td>
<td>0.06896552</td>
<td>NA 15</td>
</tr>
<tr>
<td>[3,]</td>
<td>0.10344828</td>
<td>NA 15</td>
</tr>
<tr>
<td>[4,]</td>
<td>0.13793103</td>
<td>NA 15</td>
</tr>
<tr>
<td>[5,]</td>
<td>0.17241379</td>
<td>NA 16</td>
</tr>
<tr>
<td>[6,]</td>
<td>0.20689655</td>
<td>NA 16</td>
</tr>
</tbody>
</table>
The column headed by qhat indicates the quantile being compared. The first value listed is 0.03448276, meaning that a confidence interval for the difference between the 0.03448276 quantiles is given in the next two columns. In the column headed by lower, entries NA indicate $-\infty$. In the column headed by upper, NA indicates $\infty$. So the first row of the output says that the confidence interval for the difference between the 0.03448276 quantiles is $(-\infty, 18)$. This interval contains 0, so you would fail to conclude that the quantiles differ. The function also returns a value labeled numsig, which indicates how many differences among all of the estimated quantiles are significant. That is, the confidence interval does not contain 0.

Now look at row 20 of the output. This says that when comparing the 0.68965517 quantiles, the confidence interval is $(2, \infty)$. This interval does not contain 0, so reject. Looking at rows 21–26, we again reject. So no difference between the groups is found when looking at the lower quantiles, but a difference is found from the 0.69 to 0.90 quantiles. Roughly, the results indicate than among boys who have had a family member wounded or killed, the effect is more pronounced among boys with high depression scores. Moreover, the probability
that all of these confidence intervals simultaneously contain the true differences is approximately 0.95. If it is desired to compute the exact probability, this can be done by setting the argument flag=T. If flag=T is used in the example, the output labeled pc (probability coverage) has the value 0.96762 meaning that all 29 confidence intervals contain the true differences with probability 0.96762. Said another way, the probability of making at least one Type I error among the 29 quantiles being compared is $1 - 0.96762 = 0.03238$. Figure 19 shows a plot of the shift function, where the y-axis indicates the estimated difference between the quantiles.
Figure 19: A plot of the shift function based on the Gaza data. The plot indicates that among boys with low measures of depression, there is little difference between the two groups. But as we move toward subpopulations of boys with high depression, the difference between the two groups increases.
COMPARING INDIVIDUAL QUANTILES

Better power when dealing with the tails of distributions. For comparing independent groups, use the R function. It is based in part of the Harrell–Davis quantile estimator:

\[ \sum w_i X_{(i)}, \]

where the weights \( w_i \) can be chosen to estimate the quantile (percentile) of interest. So unlike the usual median, it uses all the data to estimate the population median. That is, use a weighted sum of the ordered data.

\[ \text{qcomhd}(x,y,q=c(1,.25,.5,.75,.9),\text{nboot}=2000,\text{plotit}=\text{TRUE},\text{SEED}=\text{TRUE},\text{xlab}="\text{Group 1}"\text{,ylab}="\text{Est.1-Est.2}"\text{,alpha}=.05) \]

For dependent groups, use

\[ \text{Dqcomhd}(x,y,q=c(1:9)/10,\text{nboot}=1000,\text{plotit}=\text{TRUE},\text{SEED}=\text{TRUE},\text{xlab}="\text{Group 1}"\text{,ylab}="\text{Est.1-Est.2}"\text{,na.rm}=\text{TRUE},\text{alpha}=.05) \]

EXAMPLE

COMPARE CONTROL GROUP TO EXPERIMENTAL GROUP THAT RECEIVED INTERVENTION BASED ON DEPRESSIVE SYMPTOMS.

Near the mode, there seems to be little difference between the groups. Comparing the medians and 20% trimmed means, no significant difference is found. But the right tails of the distributions appear to differ. See Figure 20. Comparing the .8, .85 and .9 quantiles (i.e., the 80th, 85th and 90th percentiles of the intervention and control groups), the corresponding p-values are .061, .006 and .012, respectively, and the corresponding estimated differences are 3.9, 4.8 and 7.0. The main point is that, had only differences between the groups in central tendency been examined, potentially important differences would have been missed.

EXAMPLE

Sexual attitude study. Instead of comparing the groups using a trimmed mean, it is possible to probe differences using other measures, such as the quartiles. Comparing the upper quartiles (using the R function qcomhd with the argument \( q=\cdot75 \)), the estimated difference between the upper quartiles is 6.77-2.81=3.96, CI [1.31, 7.44], \( p < .001 \). So the results indicate that males tend to give more extreme responses than females. (For details about the method being used see Wilcox et al., 2013. For comparing dependent groups, see Wilcox & Erceg-Hurn, 2012.)
Figure 20: Distribution of CES-D before intervention (solid line) and after intervention
10 Measuring Effect Size

Generally, how might we measure or characterize the difference between two groups? Some possibilities are:

- Compute a confidence interval for the difference between some measure of location.
- Use a standardized difference or some related method that measures the difference between the means relative to the standard deviations.
- Use a generalization of Pearson’s correlation (based on explanatory power).
- Compare and plot the differences between all of the quantiles.
- Plot the distributions. (For example, use the R function g2plot or use boxplots.)
- Estimate the probability that a randomly sampled observation from the first group is larger than a randomly sampled observation from the second.

A commonly used approach assumes the two groups have a common variance, which we label $\sigma^2$. That is, $\sigma_1^2 = \sigma_2^2 = \sigma^2$ is assumed. Then the so-called standardized difference between the groups is

$$\delta = \frac{\mu_1 - \mu_2}{\sigma}. \quad (23)$$

Cohen (1977) suggests that as a general guide, when dealing with normal distributions, $\delta = 0.2$, 0.5 and 0.8 correspond to small, medium and large effect sizes, respectively. An estimate of $\delta$ is

$$d = \frac{\bar{X}_1 - \bar{X}_2}{s_p}, \quad (24)$$

which is often called Cohen’s $d$, where $s_p$ is the pooled standard deviation.

TWO SERIOUS CONCERNS

1. The measure of effect size $\delta$ can be seriously affected by nonnormality. The left panel of Figure 21, which is the same as in Figure 5, shows two normal distributions where the difference between the means is 1 ($\mu_1 - \mu_2 = 1$) and both standard deviations are one. So $\delta = 1$,

which is often viewed as being relatively large. Now look at the right panel of Figure 21 As is evident, the difference between the two distributions appears to be very similar to the difference shown in the left panel, so according to Cohen we again have a large effect.
Figure 21: In the left panel, $\delta = 1$. In the right panel, $\delta = .3$, illustrating that a slight departure from normality can lower $\delta$ substantially.

size. However, in the right panel, $\delta = 0.3$ because these two distributions are mixed normals with variances 10.9. This illustrates the general principle that arbitrarily small departures from normality can render the magnitude of $\delta$ meaningless. In practical terms, if we rely exclusively on $\delta$ to judge whether there is a substantial difference between two groups, situations will arise where we will grossly underestimate the degree to which groups differ, particularly when outliers occur.

2. Assumes homoscedasticity.

Here is another concern about $\delta$ when trying to characterize how groups differ. Look again at Figures 2 and 3. The distributions have equal means and equal variances but they differ in an obvious way that might have practical importance. Although the difference between measures of location provide a useful measure of effect size, we might need additional ways
of gaining perspective on the extent to which groups differ such as the shift function, already
described.

DEALING WITH NONNORMALITY

One way of dealing with nonnormality is to use a generalization of $\delta$ based on 20% trimmed means and Winsorized variances, where the Winsorized variances are rescaled so that under normality they estimate the variance (Algina, Keselman & Penfield, 2005). With 20% trimming, this means that the Winsorized variance is divided by 0.4121. That is, under normality, $s_w^2/0.4121$ estimates $\sigma^2$. If we assume the groups have equal Winsorized variances, $\delta$ becomes

$$\delta_t = 0.642 \frac{\bar{X}_{t1} - \bar{X}_{t2}}{s_w},$$

where

$$S_w^2 = \frac{(n_1 - 1)s_{w1}^2 + (n_2 - 1)s_{w2}^2}{n_1 + n_2 - 2}$$

is the pooled Winsorized variance. Under normality, and when the variances are equal, $\delta = \delta_t$. If the Winsorized variances are not equal, Algina et al. suggest using both

$$\delta_{t1} = 0.642 \frac{\bar{X}_{t1} - \bar{X}_{t2}}{s_{w1}},$$

and

$$\delta_{t2} = 0.642 \frac{\bar{X}_{t1} - \bar{X}_{t2}}{s_{w2}}.$$

10.1 A Heteroscedastic and Robust Measure of Effect Size

Momentarily imagine that an equal number of observations is sampled from each group and let $\sigma_{pool}^2$ be the (population) variance corresponding to the pooled observations. That is, we are pooling together two variables that might have different variances resulting in a variable that has variance $\sigma_{pool}^2$. Then a heteroscedastic, explanatory measure of effect size, $x_i$, is

$$\xi = \sqrt{\frac{\sigma_\mu^2}{\sigma_{pool}^2}},$$

where $\sigma_\mu^2$ is the variation among the population means.

To add perspective, it is noted that in the context of least squares regression, the approach leading to $\xi^2$ results in Pearson’s squared correlation coefficient, the coefficient of determination. (Kulinskaya and Staudte, 2006, studied another approach that is somewhat
related to $\xi^2$.) Also, as previously noted, Cohen suggested that under normality and homoscedasticity, $\delta = 0.2$, 0.5 and 0.8 correspond to small, medium and large effect sizes, respectively. The corresponding values of $\xi$ are approximately, 0.10, 0.30 and 0.50. Note that for Pearson’s correlation, it is sometimes suggested that .1, .3 and .5 are small, medium and large. To the extent this seems reasonable, the same can be said about $\xi$.

Estimation of $\xi^2$ is straightforward when there are equal sample sizes. First estimate $\sigma^2_\mu$ by replacing $\mu_1$ and $\mu_2$ with the corresponding sample means. Next, pool all $2n$ values and compute the sample variance, say $s^2_{\text{pool}}$, which estimates $\sigma^2_{\text{pool}}$. (Note that $s^2_{\text{pool}}$ and $s^2_p$ are not the same. The latter is based on a weighted average of the individual sample variances.) But when there are unequal sample sizes, this estimation method can be shown to be unsatisfactory. To deal with this, suppose the sample sizes are $n_1 < n_2$ for groups 1 and 2, respectively. If we randomly sample (without replacement) $n_1$ observations from the second group, we get a satisfactory estimate of $\xi^2$. To use all of the data in the second group, we repeat this process many times yielding a series of estimates for $\xi^2$, which are then averaged to get a final estimate, which we label $\hat{\xi}^2$. And the estimate of $\xi$ is just

$$\hat{\xi} = \sqrt{\hat{\xi}^2}.$$

**THIS MEASURE OF EFFECT SIZE IS NOT ROBUST**

**BUT A ROBUST VERSION IS READILY OBTAINED BY REPLACING THE MEANS AND VARIANCES WITH SOME ROBUST MEASURE OF LOCATION AND SCATTER.**

**Robust Variations**

The measure of effect size $\xi$ can be made more robust. Again replace the means with trimmed means, and replace the pooled sample variance with the Winsorized variance that has been rescaled to estimate the variance under normality. Henceforth, when using the measure of effect size $\xi$, the version based on 20% trimmed means and the 20% Winsorized variances will be assumed unless stated otherwise.

**EXAMPLE**

In the right panel of Figure 21, $\xi = 0.56$ (based on 20% trimmed means) indicating a
large effect size. This is in contrast to $\delta$, which indicates a relatively small effect size.

**EXAMPLE**

For data comparing hangover symptoms of sons of alcoholics versus a control group, we reject the hypothesis of equal 20% trimmed means. If we use a standardized difference between the two groups based on the means and the standard deviation of the first group we get $\hat{\delta}_1 = 0.4$. Using the standard deviation of the second group yields $\hat{\delta}_2 = 0.6$. So taken together, and assuming normality, these results suggest a medium effect size. The estimate of $\xi$ is 0.44 suggesting that this robust, heteroscedastic measure of effect size is fairly large.

**EXAMPLE**

Replication of the sexual attitude data. The sample sizes are $n_1 = 1327$ and $n_2 = 2282$, Welch’s test returns a p-value of 0.30, but Yuen’s test has a p-value less than 0.001. Cohen’s effect size, $d$, is less than 0.0001. In contrast, $\hat{\delta}_i = 0.48$, suggesting a medium effect size and $\hat{\xi} = 0.47$, suggesting a large effect size.

The R function

$$\texttt{yuenv2}(x,y)$$

is exactly like the R function $\texttt{yuen}$, only it also reports the effect size $\xi$. The R function

$$\texttt{akp.effec}\texttt{t}(x,y,\texttt{EQVAR}=\texttt{T},\texttt{tr}=0.2)$$

computes the measure of effect size $\delta_i$, which defaults to using a 20% trimmed mean. If the argument $\texttt{EQVAR}=\texttt{F}$, the function returns both $\delta_1$ and $\delta_2$. Setting $\texttt{tr}=0$, $\texttt{akp.effec}\texttt{t}$ returns Cohen’s $d$ when $\texttt{EQVAR}=\texttt{T}$.

11 Comparing Correlations and Least Squares Regression Slopes

The goal is to test

$$H_0 : \rho_1 = \rho_2,$$

the hypothesis that the two groups have equal population correlation coefficients.
DO NOT USE FISHER’S R-TO-Z TRANSFORMATION. CAN PERFORM POORLY UNDER NON-NORMALITY.

Currently, one of the more effective approaches is to use a (modified) percentile bootstrap method.

When comparing slopes, based on least squares, a wild bootstrap method can be used as well as a non-bootstrap method based in part of the HC4 estimate of the standard, which deals well with heteroscedasticity.

But a seemingly better approach is the so-called HC4 method. It uses a correct estimate of the standard error when there is heteroscedasticity.

IGNORING HETEROSCEDASTICITY CAN RESULT IN USING THE WRONG STANDARD ERROR, INVALIDATING THE RESULTS.

The R function

\[
\text{olsJ2}(x1, y1, x2, y2, xout = \text{FALSE}, outfun = \text{outpro}, plotit = \text{TRUE}, xlab = 'X', ylab = 'Y', \text{ISO} = \text{FALSE}, \ldots)
\]

can be used test the hypothesis that two independent groups have identical least squares regression lines. Setting the argument ISO=TRUE, the slopes are compared.

\[
\text{twohc4cor}(x1, y1, x2, y2)
\]

**EXAMPLE**

In an unpublished study by L. Doi, there was interest in whether a measure of orthographic ability (\(Y\)) is associated with a measure of sound blending (\(X\)). Here we consider whether an auditory analysis variable (\(Z\)) modifies the association between \(X\) and \(Y\). One way of approaching this issue is to partition the pairs of points \((X, Y)\) according to whether \(Z\) is \(\leq\) or \(>\) 14, and then enter the resulting pairs of points into the R function \text{twopcor}. The 0.95 confidence interval for \(\rho_1 - \rho_2\), the difference between the correlations, is \((-0.64, 0.14)\). This interval contains zero so we would not reject the hypothesis of equal correlations. If we compare regression slopes instead, the 0.95 confidence interval is \((-0.55, 0.18)\) and again we fail to reject. It is stressed, however, that this analysis does not establish that the association does not differ for the two groups under study. A concern is that power might be low when attention is restricted to Pearson’s correlation or least squares regression. Methods covered later indicate that the measure of auditory analysis does modify the association between orthographic ability and sound blending.
12 Making Decisions About Which Method To Use

Numerous methods have been described for comparing two independent groups. How does one choose which method to use? There is no agreed upon strategy for addressing this issue, but a few comments might help.

First, and perhaps most obvious, consider what you want to know. If, for example, there is explicit interest in knowing something about the probability that an observation from the first group is smaller than an observation from the second, use Cliff’s method or the Brunner-Munzel technique. Despite the many problems with methods for comparing means, it might be that there is explicit interest in the means, as opposed to other measures of location, for reasons discussed in Chapter 2. If this is case, the R function `yuenbt` seems to be a relatively good choice for general use (setting the argument `tr=0`) with the understanding that all methods for means can result in relatively low poor, inaccurate confidence intervals, and unsatisfactory control over the probability of a Type I error. But if any method based on means rejects, it seems reasonable to conclude that the distributions differ in some manner. A possible argument for using other methods for comparing means is that they might detect differences in the distributions that are missed by other techniques (such as differences in skewness). If, for example, boxplots indicate that the groups have no outliers and a similar amount of skewness, a reasonable speculation is that using other measures of location will make little or no difference. But the only way to be sure is to actually compare groups with another measure of location.

If the goal is to maximize power, a study by Wu (2002), where data from various dissertations were reanalyzed, indicates that comparing groups with a 20% trimmed mean is likely to be best, but the only certainty is that exceptions occur. Generally, the method that has the highest power depends on how the groups differ, which is unknown. With sufficiently heavy tails, methods based on the median might have more power. To complicate matters, situations are encountered where a rank-based method has the most power. Keep in mind that for skewed distributions, comparing means is not the same as comparing trimmed means or medians.

In terms of controlling the probability of a Type I error and getting accurate confidence intervals, all indications are that when comparing identical distributions, it makes little difference which method is used. If, however, the groups differ in some manner, the choice of method might make a substantial difference in terms of both power and accurate probability coverage. If the goal is to make inferences about a measure of location, without being sensitive to heteroscedasticity and skewness, methods based on a 20% trimmed mean seem best for general use. The bootstrap method for medians also performs well based on this criterion. Generally it is better to use a method that allows heteroscedasticity rather than a method that assumes homoscedasticity. Bootstrap methods appear to have an advantage over non-bootstrap methods when sample sizes are small. Just how large the sample sizes
must be so that non-bootstrap methods generally work as well as bootstrap methods is unknown. Also, keep in mind that testing assumptions, in order to justify a method based on means, is satisfactory only if the method for testing assumptions has sufficiently high power. Determining whether this is the case is difficult at best.

An issue of some importance is whether it is sufficient to use a single method for summarizing how groups differ. Different methods provide different perspectives, as was illustrated. So one strategy might be to focus on a single method for deciding whether groups differ and then use other methods, such as rank-based techniques or a shift function, to get a deeper understanding of how and where the groups differ.

A criticism of performing multiple tests is that as the number of tests performed increases, the more likely it is that at least one test will reject even when the groups do not differ. If we perform three tests at the 0.05 level, the probability of at least one Type I error is greater than 0.05. (Chapter 13 discusses this issue in more detail.) However, if several tests reject at the 0.05 level, say, and if each test controls the probability of a Type I error, then it is reasonable to conclude that the groups differ and that the probability of at least one Type I error, among all the tests performed, does not exceed 0.05. (This conclusion is justified based on sequentially rejective methods in Chapter 13.) Also, if a method fails to reject, it seems unwise to not consider whether other methods find a difference. At a minimum, using alternative methods might provide an indication of where to look for differences when conducting future investigations. Moreover, different methods provide alternative perspectives on how groups differ.

A final suggestion is to take advantage of the various plots that were described. They can be invaluable for understanding how groups differ.

13 COMPARING TWO DEPENDENT GROUPS

There are several general ways two dependent groups might be compared:

1. Compare measures of location, such as the mean or median.
2. Compare measures of variation.
3. Focus on the probability that a randomly sampled observation from the first group is less than a randomly sampled observation from second group.
4. Use a rank-based method to test the hypothesis that the marginal distributions are identical.
5. Simultaneously compare all of the quantiles to get a global sense of where the distributions differ and by how much.
All of the methods in for comparing independent groups have analogs for the case of two dependent groups.

13.1 When Does the Paired T Test Perform Well?

The good news about the paired T test is that if the observations in the first group (the $X_{i1}$ values) have the same population distribution as the observations in the second group, so in particular they have equal means, variances and the same amount of skewness, generally Type I error probabilities substantially higher than the nominal level can be avoided. The reason is that for this special case, the difference scores (the $D_i$ values) have a symmetric (population) distribution in which case methods based on means perform reasonably well in terms of avoiding Type I error probabilities substantially higher than the nominal level. However, practical problems can arise when the two groups (or the two dependent variables) differ in some manner. Again, arbitrarily small departures from normality can destroy power, even when comparing groups having symmetric distributions. And if groups differ in terms of skewness, the difference scores (the $D_i$ values) have a skewed distribution, in which case the paired T test can be severely biased, meaning that power can actually decrease as the difference between the means gets large. Yet another problem is poor probability coverage when computing a confidence interval for the difference between the means. If the goal is to test the hypothesis that the two variables under study have identical distributions, the paired T test is satisfactory in terms of Type I errors. But if we reject, there is doubt as to whether this is primarily due to the difference between the means or some other way the distributions differ. With a large enough sample size, these concerns become negligible, but just how large the sample size must be is difficult to determine.

14 Comparing Robust Measure of Location

Again, one way of dealing with the risk of relatively low power when comparing means is to use 20% trimmed means, medians or M-estimators instead. A possible appeal of methods for comparing 20% trimmed means is that by design, they perform about as well as the paired T test for means when the normality assumption is true. But in fairness, if a distribution is sufficiently light-tailed, comparing means might have a power advantage. And if distributions are sufficiently heavy-tailed, medians or M-estimators might provide more power. As usual, if distributions are skewed, comparing means is not the same as comparing a 20% trimmed mean or other robust measures of location. And despite any concerns about the mean, there are situations where it is more appropriate than a 20% trimmed mean. (See Chapter 2 of Wilcox, 2012.) But to get good power under a broad range of conditions, as well as improved control over the Type I error probability, using a 20% trimmed mean has appeal.
For convenience, the focus is on trimmed means, but the methods about to be described extend to medians and M-estimators as will be made clear. If two dependent groups have identical distributions, then the population trimmed mean of the difference scores is equal to the difference between the individual trimmed means, which is zero. In symbols, $\mu_{tD} = \mu_{t1} - \mu_{t2} = 0$. However, if the distributions differ, in general it will be the case that $\mu_{tD} \neq \mu_{t1} - \mu_{t2}$.

In practical terms, computing a confidence interval for $\mu_{tD}$ is not necessarily the same as computing a confidence interval for $\mu_{t1} - \mu_{t2}$. And the same is true when using medians or an M-estimator. So an issue is whether one should test

$$H_0 : \mu_{tD} = 0,$$  \hspace{1cm} (25)

or

$$H_0 : \mu_{t1} = \mu_{t2}.$$  \hspace{1cm} (26)

The latter approach is called comparing the marginal trimmed means. Put another way, the distributions associated with each of the dependent groups, ignoring the other group, are called marginal distributions. The goal is to compare the trimmed means of the marginal distributions.

To elaborate on the difference between these two hypotheses in more concrete terms, consider a study aimed at assessing feelings of life satisfaction based on randomly sampled pairs of brothers and sisters. If we test Equation (25), the goal in effect is to assess how the typical brother compares to his sister. Testing Equation (26) instead, the goal is to compare how the typical female compares to the typical male.

In terms of controlling the Type I error probability, currently it seems that there are only slight differences between the two approaches. A rough rule is that using difference scores, meaning that the goal is to test Equation (9.3), Type I errors tend to be a bit higher compared to testing (9.4). For example, under normality, if the pairs of observations have correlation $\rho = .4$, testing (25) at the 0.05 level (using the R function trimci), the actual Type I error probability is approximately 0.059 compared to 0.044 when testing (26) (using the R function yuend described momentarily). With $\rho = 0$, the actual Type I error probabilities are approximately 0.059 and 0.047, respectively. So in this particular case, comparing marginal trimmed means results in a Type I error probability that is closer to the nominal 0.05 level. But if the distributions are lognormal, the Type I error probabilities are approximately 0.043 and 0.033, suggesting that using difference scores might have more power. In terms of maximizing power, the optimal choice depends on how the groups differ, which of course is unknown. If forced to choose one approach over the other, with goal of maximizing power, using difference scores seems to have an advantage. But perhaps situations arise where, from a substantive point of view, the choice between testing Equation (25) and (26) is relevant.
15 Missing Values

All indications are that from a robust point of view, something other than imputation should be used. If values are missing at random, use a percentile bootstrap method, which uses all of the available data.

The R function

\[ \text{rmmismcp}(x, y, \text{est} = \text{tmean}) \]

test the hypothesis \( H_0: \mu_{t1} = \mu_{t2} \) using this approach. With 20% trimming or more, it appears to be one of the better methods for general use when there are missing values. By default, a 20% trimmed mean is used, but other measures of location can be used via the argument est. For example, \( \text{rmmismcp}(x, y, \text{est} = \text{onestep}) \) would compare the groups with a one-step M-estimator. The function also computes a confidence interval.

The R function

\[ \text{rm2miss}(x, y, \text{tr} = 0) \]

also tests \( H_0: \mu_{t1} = \mu_{t2} \) assuming any missing values occur at random. It uses a bootstrap-t method in conjunction with an extension of the method derived by Lin and Stivers (1974), and by default it tests the hypothesis of equal means. It appears to be one of the better methods when the amount of trimming is small.

16 Comparing All Quantiles

Again, it might be of interest to compare low scoring individuals in one group to low scoring participants in another. This might be done by comparing the lower quartiles. Simultaneously, one might want to compare high scoring participants using the upper quartiles. This can done using an analog of the shift function for dependent groups.

The R function

\[ \text{lband}(x, y = \text{NA}, \alpha = 0.05, \text{plotit} = \text{T}) \]

compares all quantiles of two dependent groups and plots the shift function if the argument plotit=T. If the argument y is not specified, it is assumed that the argument x is a matrix with two columns.
Again, for individual quantiles, use

\[
Dqcomhd(x, y, q=c(1:9)/10, nboot=1000, plotit=TRUE, SEED=TRUE, xlab="Group 1", ylab="Est.1-Est.2", na.rm=TRUE, alpha=.05)
\]

It uses the Harrell-Davis estimator. Power can be much better when dealing with the tails. Figure 22 illustrates the plot created by this function.

**EXAMPLE**

A study was performed where EEG measures of convicted murderers were compared to a control group. Measures for both groups were taken at four sites in the brain. For illustrative purposes, the first two sites are compared.

The R function `lband` finds differences at the 0.14, 0.5 and 0.57 quantiles. That is, the 0.14 quantile for the first site differs significantly different from the 0.14 quantile of the second site, and the same is true for the 0.5 and 0.57 quantiles.

## 17 One-Way and Higher ANOVA Designs

Violating assumptions is even more serious compared to the one- and two-sample situations.

All of the methods for comparing two groups in a robust manner can be extended to higher-way designs.

Test the assumptions of normality and homoscedasticity? Not supported based on published studies. Such tests often do not have enough power to detect violations of assumptions that have practical importance.

R functions for dealing with two-way, three-way designs, including repeated measures designs are described in my 2012 books. New editions of these books are in progress.

## 18 Multiple Comparisons

\( J \) independent groups.
Figure 22: Example using the R function Dqcomhd. Decrease in depression is indicated by the y-axis. The x-axis indicates the level of depression prior to intervention.
Dunnett’s T3 method is readily extended to trimmed means. Briefly, for each pair of groups, apply Yuen’s method and control FWE use a critical employed by T3, with the degrees of freedom adjusted based on the amount of trimming.

The R function
\[
\text{lincon}(x, \text{con}=0, \text{tr}=0.2, \text{alpha}=0.05),
\]
compares all pairs of groups using the extension of the T3 method to trimmed means. The argument \(x\) is assumed to be a matrix with \(J\) columns or to have list mode. The argument \(\text{tr}\) controls the amount of trimming and defaults to 20%.

### 18.1 Percentile Bootstrap Methods for Comparing Trimmed Means, Medians and M-estimators

The only known method for comparing medians that remains reasonably accurate when tied values occur is based on a percentile bootstrap method. Also, when there are no tied values, there seems to be no compelling reason to use a non-bootstrap method.

When comparing M-estimators or MOM, the same method used to compare medians appears to be best for general use when the sample sizes are large. That is, for each group, an M-estimator is computed rather than a median, otherwise the computations are the same. But with small sample sizes, roughly meaning that all sample sizes are less than or equal to 80, the method is too conservative in terms of Type I errors: the actual level can drop well below the nominal level. An adjustment is available that appears to be reasonably effective.

The R function
\[
\text{medpb}(x, \text{alpha}=.05, \text{nboot}=\text{NA}, \text{grp}=\text{NA}, \text{est} = \text{median}, \text{con}=0, \text{bhop}=\text{F}),
\]
performs all pairwise comparisons of \(J\) independent groups using medians. The argument \(\text{nboot}\) determines how many bootstrap samples will be used. By default, \(\text{nboot}=\text{NA}\), meaning that the function will choose how many bootstrap samples will be used depending how many groups are being compared.

The R function
\[
\text{tmcppb}(x, \text{alpha}=.05, \text{nboot}=\text{NA}, \text{grp}=\text{NA}, \text{est} = \text{tmean}, \text{con}=0, \text{bhop}=\text{F}, \ldots),
\]
is exactly the same as \text{mdepb}, only it defaults to using a 20% trimmed mean.

The R function
mcppb20(x)

compares all pairs of groups via 20% trimmed means using the method in Wilcox (2001d). It is limited to $\alpha = 0.05$ and can be used with 20% trimmed means only.

The R function

\[
\text{pbmcp}(x, \text{alpha} = 0.05, \text{nboot} = \text{NA}, \text{grp} = \text{NA}, \text{est} = \text{onestep}, \text{con} = 0, \text{bhop} = \text{F})
\]

can be used to compare measures of location indicated by the argument \text{est}, which defaults to a one-step M-estimator. If the largest sample size is less than 80 and \text{bhop} = \text{F}, method SR is used to control FWE. But if the largest sample size exceeds 80, Hochberg’s method is used, still assuming that \text{bhop} = \text{F}. If \text{bhop} = \text{T}, the Benjamini - Hochberg method is used. With small sample sizes, this function appears to be best when comparing groups based on an M-estimator or MOM. If the goal is to compare groups using MOM, set the argument \text{est} = \text{mom}.

The R function

\[
\text{linconb}(x, \text{con} = 0, \text{tr} = 0.2, \text{alpha} = 0.05),
\]

compares all pairs of groups using the bootstrap-t method just described.

TWO-WAY AND THREE-WAY DESIGNS CAN BE HANDLES AS WELL, INCLUDING WITHIN SUBJECTS (REPEATED MEASURES) DESIGNS.

19 SOME MULTIVARIATE METHODS

19.1 Detecting Outliers

Usual Mahalanobis distance: Suffers from masking. One of the worst possible methods.

Need a method that uses a robust measure of location and scatter. Numerous methods have been proposed. Two that perform relatively well are a projection method and what is called the minimum generalized variance method.

The R function

\[
\text{outpro}(m, gval = \text{NA}, \text{center} = \text{NA}, \text{plotit} = \text{T}, \text{op} = \text{T}, \text{MM} = \text{F}, \text{cop} = 3, \text{xlab} = ‘VAR 1’, \text{ylab} = ‘VAR 2’)
\]
checks for outliers using the projection method described in the previous section. By default, the method uses the boxplot rule on each projection. To use the MAD-median rule, set the argument **MM=T**. The argument cop determines how the center of the data cloud is computed. By default, the marginal medians are used. The options cop=2, 3, 4, 5, 6, 7 correspond to MCD, marginal medians, MVE, TBS, RMBA and a measure of location called the *spatial (L1) median*, respectively. These choices can make a difference in the results, but from a practical point of view it is unclear which method is best for routine use. An alternative choice for the center of the data can be specified with the argument center. (The argument op has to do with how the data are plotted; see Wilcox, 2012, for more information.) (The argument gval corresponds to $c$ to a critical value used to declare a point an outlier. The function chooses a value for gval if one is not specified.)

Filzmoser, Maronna and Werner (2008) noted that under normality, if the number of variables is large, the proportion of points declared outliers by the better-known outlier detection methods can be relatively high. This concern applies to all the methods covered here with the projection method seemingly best at avoiding this problem. But with more than nine variables ($p > 9$), it breaks down as well. Currently, the best way of dealing with this problem is to use the projection method, but rather than use the boxplot rule or the MAD-median rule (as described in Chapter 2 of Wilcox, 2011), determine an adjustment so that the proportion of points declared outliers is small, say 5%, when sampling from a normal distribution.

The R function

```r
outproad(m,gval = NA, center = NA, plotit = T, op = T, MM = F, cop = 3, xlab = "VAR 1", ylab = "VAR 2",rate = 0.05)
```

is the same as `outpro`, only the decision rule is adjusted so that the expected proportion of points declared outliers, under normality, is approximately equal to the value indicated by the argument `rate`, which defaults to 0.05. Use this function if the number of variables is greater than 9.

The R functions

```r
outproMC(m, gval, center = NA, plotit = T, op = T, MM = F, cop = 3, xlab = "VAR 1", ylab = "VAR 2")
```

and

```r
outproadMC(m, center = NA, plotit = T, op = T, MM = F, cop = 3, xlab = "VAR 1", ylab = "VAR 2",rate = 0.05)
```
are the same as outpro and outproad, respectively, but they take advantage of a multi-core processor if one is available. (They require the R package multicore.)

When working with three dimensional data, the R function

\[
\text{out3d}(x, \text{outfun}=\text{outpro}, \text{xlab}="\text{Var 1}" , \text{ylab}="\text{Var 2}" , \text{zlab}="\text{Var 3}" , \\
\text{reg.plane}=\text{F}, \text{regfun}=\text{tsreg})
\]

will plot the data and indicate outliers with an *}. If the argument \text{reg.plane}=\text{T}, this function also plots the regression plane based on the regression estimator indicated by the argument \text{regfun}, assuming that \text{regfun} returns the intercept and slope estimates in \text{regfun$coef}. When plotting the regression surface, the function assumes the outcome variable is stored in the third column of the argument \text{x}. And the two predictors are stored in the first two columns. To plot the least squares regression surface, use \text{regfun=lsift}. By default, the function uses the Theil–Sen regression estimator. (The function out3d requires the R package scatterplot3d, which can be installed as described in Chapter 1 of Wilcox, 2012.)

**EXAMPLE**

Predictors of reading ability study. Figure 23 shows the points flagged as outliers plus the regression plane based on the Theil–Sen estimator. The R command is

\[
\text{out3d(read[,c(3,4,8)], \text{reg.plane}=\text{T}, \text{regfun}=\text{tsreg}, \text{xout}=FALSE, \text{xlab}="\text{SBT1}" , \text{ylab}="\text{RAN1T1}" , \\
\text{zlab}="\text{WWISST2}" )}
\]

Figure 24 is the same, only the regression plane estimated ignoring the outliers. The R command is now

\[
\text{out3d(read[,c(3,4,8)], \text{reg.plane}=\text{T}, \text{regfun}=\text{tsreg}, \text{xout}=\text{TRUE}, \text{xlab}="\text{SBT1}" , \text{ylab}="\text{RAN1T1}" , \\
\text{zlab}="\text{WWISST2}" )}
\]

the two regression planes differ substantially.

20 REGRESSION AND CORRELATION

Recall from basic principles

\[
\hat{\sigma}^2 = \frac{\sum (X_i - \bar{X})^2}{\sum (X_i - \bar{X})^2}
\]

estimates standard error of \(b_1\). Note that outliers among independent variable reduce the standard error, but outliers among the dependent variable increase it.
Figure 23: A scatterplot of reading data based on the R function out3d.
Figure 24: A scatterplot of reading data based on the R function `out3d`, only now outliers among the independent variables are ignored.
Figure 25: An example of homoscedasticity. The conditional variance of $Y$, given $X$, does not change with $X$.

Review the notion of homoscedasticity and heteroscedasticity. Figure 25 illustrates homoscedasticity and Figure 26 illustrates heteroscedasticity.

WHEN USING LEAST SQUARES REGRESSION, SUGGEST DEALING WITH HETEROSEDASTICITY USING THE HC4 ESTIMATOR.

IN PRACTICAL TERMS, USE ONE OF THE FOLLOWING R FUNCTIONS:

The R function

```r
olshc4(x, y, alpha=.05, xout=F, outfun=out)
```

computes $1 - \alpha$ confidence intervals and p-values for each of the individual parameters. By
Figure 26: An example of heteroscedasticity. The conditional variance of $Y$, given $X$, changes with $X$. 
default, 0.95 confidence intervals are returned. Setting the argument alpha equal to 0.1, for example, will result in 0.9 confidence intervals.

The function

\[ \text{hc4test}(x, y, xout=F, outfun=out) \]

tests the hypothesis that all of the slopes are equal to zero. Note that both functions include the option of removing leverage points via the arguments \( xout \) and \( outfun \).

21 Problems with Least Squares

**EXAMPLE**

Figure 28 shows a scatterplot of data dealing with the association between surface temperature of stars and their light intensity. The solid line is the least squares regression line using all of the data. The dashed line is the least squares regression line ignoring leverage points.

In regression, any outlier among the \( X \) values is called a *leverage point*. There are two kinds: good and bad.

**BAD LEVERAGE POINTS**

Roughly, bad leverage points are outliers that can result in a misleading summary of how the bulk of the points are associated. That is, bad leverage points are not consistent with the association among most of the data.

**NOTE:** Even when no outliers are detected among the data for the independent variable, unusual points can have a substantial impact on the least squares estimator as illustrated in Figure 27.

**EXAMPLE**

Consider data based on measures taken from a lake. There are two predictors. When both predictors are included in the model, the p-value, when testing \( H_0: \ beta_1 = 0 \), is 0.56.
Figure 27: The two points marked by the square in the lower right corner have a substantial impact on the least squares regression line. Ignoring these two points, the least squares regression is given by the solid line. Including them, the least squares regression line is given by the dotted line. Moreover, none of the $X$ or $Y$ values are declared outliers using the boxplot rule or MAD-median rule.
Figure 28: The solid line is the least squares regression line using all of the star data. Ignoring outliers among the $X$ values, the least squares regression line is now given by the dotted line.
Figure 29: The left panel shows a scatterplot of the lake data. The bad leverage points, located in the lower right portion of the scatterplot, have a tremendous influence on the least squares estimate of the slope resulting in missing any indication of an association when using least squares regression. The right panel shows a scatterplot of the reading data. Now the bad leverage points mask a negative association among the bulk of the points.
Ignoring the second predictor, the p-value is 0.72 illustrating that p-values are a function in part of which predictors are entered into the model. If we perform the analysis again with both predictors, but with outliers among the predictors removed, now the p-value when testing $H_0: \beta_1 = 0$ is 0.0009, illustrating that a few bad leverage points can have a large impact when testing hypotheses. See the left panel of Figure 29.

EXAMPLE

Consider again the reading study (conducted by L. Doi) dealing with predictors of reading ability. One portion of her study used a measure of digit naming speed (RAN1T) to predict the ability to identify words (WWISST2). Using all of the data, the slope of the least squares regression line is nearly equal to zero, which is the nearly horizontal line in the right panel of Figure 30. When testing $H_0: \beta_1 = 0$, the p-value is .76. But the scatterplot clearly reveals that the six largest X values are outliers. When these outliers are removed, the dotted line shows the resulting least squares regression line and now the p-value is 0.002. See the right panel of Figure 29.

GOOD LEVERAGE POINTS

Leverage points can distort the association among most of the data, but not all leverage points are bad. A crude description of a good leverage point is a point that is reasonably consistent with the regression line associated with most of the data. Figure 30 illustrates the basic idea.

BEWARE OF DISCARDING OUTLIERS AMONG THE Y VALUES (THE DEPENDENT VARIABLE)

THIS RESULTS IN USING AN INVALID ESTIMATE OF THE STANDARD ERROR

DO NOT ASSUME THE REGRESSION LINE IS STRAIGHT

EXAMPLE

In a study by C. Chen and F. Manis, one general goal was to investigate the extent
Figure 30: Shown are both good and bad leverage points. Good leverage points do not mask the true association among the bulk of the points and they have the practical advantage of resulting in shorter confidence intervals.
Asian immigrants to the United States learn how to pronounce English words and sounds. Figure 31 shows a scatterplot of age versus an individual’s score on a standardized test of pronunciation. As indicated, there appears to be curvature as well as heteroscedasticity. (The middle regression line in Figure 31 shows the estimated median score on the pronunciation test, given an individual’s age, and was created with the function qsmcobs. (See my book on robust methods for more details.) The lower and upper regression lines show the predicted 0.25 and 0.75 quantiles, respectively.)

BUT: qsmcobs can produce wavy regression lines that are spurious. That is, it can indicate more curvature than what actually exists.

Can reduce this problem by switching to the R function:

\[
\text{qhdsm}(x,y,\text{qval}=5,\text{q=NULL},\text{pr}=\text{FALSE},\text{out}=\text{FALSE},\text{outfun}=\text{outpro},\text{plotit}=\text{TRUE}, \text{xlab}='X',\text{ylab}='Y',\text{zlab}='Z',\text{pyhat}=\text{FALSE},\text{fr}=\text{NULL},\text{LP}=\text{TRUE},\text{theta}=50,\text{phi}=25,\text{ticktype}=\text{'simple'},\text{nmin}=0,\text{scale}=\text{FALSE},\text{pr.qhd}=\text{TRUE},...)
\]

It uses a combination of the running interval smooth, the Harrell–Davis estimator and LOESS.

**PEARSON’S CORRELATION**

The following features of data influence the magnitude of Pearson’s correlation:

- The slope of the line around which points are clustered.
- The magnitude of the residuals.
- Outliers
- Restricting the range of the \(X\) values, which can cause \(r\) to go up or down.
- Curvature.

Regardless of how large the sample size might be, Pearson’s correlation can miss a strong association among the bulk of the participants. The left panel of Figure 32 shows a bivariate normal distribution with \(\rho = .8\). In the right panel, \(\rho = .2\). But look at Figure 33. The correlation appears to be high, but \(\rho = .2\)

The R function

\[ \text{pcorhc4}(x,y) \]
Figure 31: Shown are age and an individual’s score on a pronunciation test. The lower, middle and upper regression lines indicate the predicted 0.25 quantile, the median, and the 0.75 quantile, respectively. Note that the regression lines appear to be fairly horizontal among young individuals, but as age increases, scores on the test tend to decrease, and the variation among the scores appears to increase.
Figure 32: When both $X$ and $Y$ are normal, increasing $\rho$ from .2 to .8 has a noticeable effect on the bivariate distribution of $X$ and $Y$. 
Figure 33: Two bivariate distributions can appear to be very similar yet have substantially different correlations. Shown is a bivariate distribution with $\rho = .2$, but the graph is very similar to the left panel in the previous figure where $\rho = .8$. 
has been provided to compute a confidence interval for ρ using the HC4 method, meaning that it deals with heteroscedasticity, and a p-value is returned as well.

22 Robust Regression and Measures of Association

22.1 The Theil-Sen Estimator

Momentarily consider a single predictor and imagine that we have n pairs of values:

\[(X_1, Y_1), \ldots, (X_n, Y_n)\].

Consider any two pairs of points for which \(X_i > X_j\). The slope corresponding to the two points \((X_i, Y_i)\) and \((X_j, Y_j)\) is

\[b_{ij} = \frac{Y_i - Y_j}{X_i - X_j}.\] (27)

The Theil (1950) and Sen (1964) estimate of \(\beta_1\) is the median of all the slopes represented by \(b_{ij}\), which is labeled \(b_{1ts}\). The intercept is estimated with

\[b_{0ts} = M_y - b_{1ts}M_x,\]

where \(M_y\) and \(M_x\) are the sample medians corresponding to the \(Y\) and \(X\) values, respectively.

The R function

\[\text{tsreg}(x, y, \text{xout}=F, \text{outfun}=\text{out}, \text{iter}=10, \text{varfun}=\text{pbvar}, \text{corfun}=\text{pbcor}, \ldots)\]

computes the Theil–Sen estimator. If the argument \(\text{xout}=T\), leverage points are removed based on the outlier detection method specified by the argument \(\text{outfun}\).

A recent modification gives better results (e.g., higher power) when there are tied (duplicated) values among the dependent variable. Use the R function:

\[\text{tshdreg}(x, y, \text{HD}=\text{TRUE}, \text{xout}=\text{FALSE}, \text{outfun}=\text{out}, \text{iter}=10, \text{varfun}=\text{pbvar}, \text{corfun}=\text{pbcor}, \text{plotit}=\text{FALSE}, \text{tol}=1\text{e-04}, \text{RES}=\text{FALSE}, \ldots)\]

22.2 M-Estimators

M-estimators of location can be extended to regression. Many variations have been proposed. Currently, a so-called MM-estimator is popular among some statisticians. The Coakley-Hettmansperger estimator also has excellent theoretical properties.
But a concern is contamination bias. Non-bootstrap methods for making inferences about the MM-estimator are available.

But generally, hypothesis testing methods based on M-estimators, that use a non-bootstrap method in conjunction with an estimate of the standard errors, perform poorly when dealing with skewed distributions. Whether this is a practical problem for the MM-estimator has not been investigated.

The R function

\[
\text{chreg}(x,y,\text{bend}=1.345).
\]

computes the Coakley-Hettmansperger regression estimator. As will all regression functions, the argument x can be a vector or an \( n \) by \( p \) matrix of predictor values. The argument bend is the value of \( K \) used in Huber’s \( \Psi \).

The R function

\[
\text{MMreg}(x,y)
\]

computes Yohai’s MM-estimator, assuming that the R package robustbase has been installed.

**EXAMPLE.** If the star data in Figure 28 are stored in the R variables starx and stary, the command

\[
\text{chreg(starx,stary)}
\]

returns an estimate of the slope equal to 4.0. The R function MMreg estimates the slope to be 2.25, the only point being that the choice of estimator can make a practical difference.

### 22.3 Other Regression Estimators

Quantile regression: an extension of least absolute value regression.

The R

\[
\text{rqfit}(x,y,\text{qval}=0.5,\text{xout}=\text{F},\text{outfun}=\text{out},\text{res}=\text{F}),
\]


performs the calculations; calls the function \texttt{rq}. (Both \texttt{rq} and \texttt{rqfit} assume that you have installed the R package \texttt{quantreg}. To install this package, start R and use the command \texttt{install.packages("quantreg")}.) One advantage of \texttt{rqfit} over the built-in function \texttt{rq} is that, by setting the argument \texttt{xout=T}, it will remove any leverage points found by the function indicated by the argument \texttt{outfun}, which defaults to the function \texttt{out}.

The R function

\begin{verbatim}
proc {mdepreg(x,y)}
end proc
\end{verbatim}

computes the deepest regression line. It is based on the goal of a regression line giving the median of \( Y \) given \( X \) in a manner that protects against bad leverage points.

Other methods are least trimmed squares, S-estimators, skipped estimators, plus several others listed in Wilcox (2005). Skipped estimators remove outliers, using an appropriate multivariate outlier detection method (such as a projection method or the MGV method), and apply some regression estimator to the data that remain. Least squares is not a good choice. Better is the Theil - Sen estimator. The R functions

\begin{verbatim}
opreg
end proc
\end{verbatim}

and

\begin{verbatim}
mgvreg
end proc
\end{verbatim}

apply skipped estimators. They perform well when there is heteroscedasticity.

22.4 Dealing with Curvature: Smoothers

Roughly, given the goal of predicting \( Y \), given \( X = x \), look at \( X_i \) values close to \( x \) and use the corresponding \( Y_i \) values to determine the predicted value of \( Y \). Do this for all \( X_i \) \((i=1,\ldots,n)\) yielding \( \hat{Y}_i \). Many variations have been proposed.

The R function

\begin{verbatim}
lpplot(x,y,span=.75,pyhat=F,eout=F,xout=F,outfun=out,plotit=T, expand=.5,low.span=2/3,varfun=pbvar, scale=F,xlab="X",ylab="Y",zlab="",theta=50,phi=25),
\end{verbatim}

plots Cleveland’s smoother automatically and provides a variety of options that can be useful. For example, it will remove all outliers if \texttt{eout=T}. To remove leverage points only, use \texttt{xout=T}. If \texttt{pyhat=T}, the function returns \( \hat{Y}_i \), the estimate of \( Y \) given that \( X = X_i \). To
suppress the plot, use \texttt{plotit=F}. The argument \texttt{low.span} is the value of \( p \), the span, when using a single predictor. The arguments \texttt{xlab} and \texttt{ylab} can be used to label the x-axis and y-axis, respectively. The arguments \texttt{theta} and \texttt{phi} can be used to rotate three dimensional plots. The argument \texttt{scale} is important when plotting a three-dimensional plot. If there is no association, the default \texttt{scale=F} typically yields a good plot. But when there is an association, \texttt{scale=T} is usually more satisfactory. (The argument \texttt{varfun} has to do with measuring the strength of the association using an approach to be explained.)

\textit{EXAMPLE}

For a diabetes study, the goal was to understand the association between the age of children at diagnosis and their C-peptide levels. As noted there, the hypothesis of a zero slope is rejected with the R function \texttt{hc4test}; the p-value is 0.034. Student’s T test of a zero slope has a p-value of 0.008. So a temptation might be to conclude that as age increases, C-peptide levels increase as well. But look at Figure 34, which shows Cleveland’s smooth. Note that for children up to about the age of 7, there seems to be a positive association. But after the age of 7, it seems that there is little or no association at all.

\textit{EXAMPLE}

The Well Elderly 2 study by Clark et al. (2012) was generally aimed at assessing the impact of an intervention program designed to enhance the physical and emotional well being of older adults. A portion of the study dealt with the association between the cortisol awakening response and a measure of depressive symptoms, which is labeled CESD. Cortisol is a hormone produced by the adrenal cortex. The cortisol awakening response (CAR) refers to the change in cortisol upon awakening and measured again 30-60 minutes later. Here the focus is on measures taken after six months of intervention. The sample size is \( n = 328 \). Using least squares regression, which is the most commonly used method for detecting an association, no association is found. The p-value is 0.218. Using least squares regression in conjunction with a method that allows heteroscedasticity (via the R function \texttt{olshc4}), the p-value is 0.168. Using the Theil–Sen estimator (via the R function \texttt{regci}), the p-value is 0.531. Eliminating leverage points and again using the Theil–Sen estimator, the p-value is 0.242. Given the reasonably large sample size, these results might suggest that there is little or no association.

However, look at Figure 35, which shows a plot of the regression line using the running interval smoother. (The data are stored on the author’s web page in the file A3B3C.txt, which can be downloaded as described in Section 1.5.) Assuming the data have been read into the R variable \texttt{A3B3C}, here are the R commands that were used:
Figure 34: A smooth created by the R function \texttt{lplot} using the diabetes data. Note that there seems to be a positive association up to about the age of 7, after which there is little no association.
dif = A3B3C$cort1 - A3B3C$cort2
rplot(dif, A3B3C$CESD, xout=TRUE, xlab='CAR', ylab='CESD').

The R variable A3B3C$cort1 contains the cortisol measures upon awakening and A3B3C$cort2 contains the cortisol measures 30-60 minutes after awakening. So the first R command stores the CAR values in the R variable dif. The argument xout=TRUE indicates that observations associated with leverage points were removed. Notice that when the CAR is negative (cortisol increases after awakening), the regression line appears to be nearly horizontal. But when the CAR is positive, there appears to be a positive association with CESD (depressive symptoms). Focusing on only the data for which the CAR is positive, the p-value based on the Theil–Sen estimator is 0.038. For CAR negative, no association is found; the p-value is 0.631.

EXAMPLE

The next example involves examining the relationship between quality of life and anxiety among 47 patients diagnosed with depressive and anxiety disorders. The data come from a study by McEvoy et al. (2014). Quality of Life (QOL) was measured using the Quality of Life Enjoyment and Satisfaction Questionnaire Short Form (Endicott, Nee, Harrison & Blumenthal, 1993), with scores ranging from 0 to 100, larger scores indicating greater quality of life. Anxiety was measured using the Beck Anxiety Inventory (Beck, Epstein, Brown, & Steer, 1988). Scores can range from 0 to 63, with higher scores indicative of higher levels of anxiety. The hypothesis under investigation was that as anxiety increases, quality of life should decrease in a linear fashion. The data are stored in the file anxiety.csv. Correlation coefficients suggest that there is a strong linear relationship between the two variables. Pearson’s correlation is -.53 and the robust skipped correlation is -.59. (The skipped correlation can be computed with the R function scor or the matlab function described by Pernet et al. 2013.) Classical (least squares) and robust regression estimators also suggest that there is a linear relationship between the two variables. Based on the classical (least squares) regression estimator (using the R function olshc4, which allows heteroscedasticity), the slope is estimated to be -.70, CI [-1.06, -.34]. Using the modified Theil–Sen estimator, the estimated slope is similar, -.84, CI [-1.32, -.41]. Confidence intervals were computed with the R function regci and the argument regfun=tshdreg.) A scatterpot of the data and the estimate of the regression line are shown Figure 36.

Next, we fit a regression line using lplot, which is the curved line in Figure 37. This suggests that there is a strong linear relationship between quality of life and anxiety when anxiety scores are between 0 and 20, but for anxiety scores above 20 there appears to be little or no association.

This finding can be probed further by splitting the data into two groups, according to
Figure 35: Shown is the smooth created by the R function `rplot`. The solid line reflects an estimate of the typical CESD measure (depressive symptoms) given a value for the CAR (the cortisol awakening response). Notice that there appears to be a distinct bend close to where the CAR is zero.
Figure 36: The straight line is the estimate of the regression line using a modification of the Theil–Sen estimator.

whether anxiety scores are above or below 20, and computing regression slopes separately for each group. If the regression line is truly straight, then the slopes should be similar. If the regression line is curved, they should differ. The estimated slope (using the R function regci) is -2.15, CI [-3.17, -1.46] for anxiety scores under 20 and the strength of association is estimated to be .63 (using the R function tshdreg). For scores above 20 the estimated slope is -.06 and now the strength of the association is .04. (When using least squares regression and the usual sample variance, the strength of the association used here reduces to the absolute value of Pearson’s correlation.) There is a significant difference between the two slopes (using the R function reg2ci and the argument est=tshdreg). The estimated difference is -2.2 CI [-3.65, -.97], which provides additional evidence of a curvilinear association.
Figure 37: The LOESS regression line
EXAMPLE

Consider again the Well Elderly study, only now the goal is to predict a measure of meaningful activities, labeled MAPA, using two independent variables: the CAR (the cortisol awakening response) and CESD (a measure of depressive symptoms). Using least squares regression, the slope associated with the CAR is not significant using the R function `olshc4`; the p-value is 0.787 (with leverage points removed). So when taking CESD into account, the usual regression model does not indicate any association between the CAR and MAPA. Also, a smooth indicates a fairly straight, horizontal regression line between MAPA and the CAR (ignoring CESD), and no association is found using least squares or the Theil–Sen estimator when ignoring CESD. Moreover, no association between MAPA and the CAR is found when focusing on only the CAR values greater than zero. That is, unlike depressive symptoms, no association is found between the CAR and MAPA when cortisol decreases shortly after awakening. (The same is true when CAR is negative.) In summary, all of the analyses just described find no association between MAPA and the CAR.

However, look at Figure 38, which shows the regression surface for predicting MAPA given a value for CAR and CESD. Notice the distinct bend close to where CESD is 16. The nature of the association appears to depend crucially on whether depressive symptoms are high or low. Applying least squares regression again, but now using only the data for which the CAR is less than 16, both slopes for predicting MAPA are significant when testing at the 0.05 level. For the slope associated with CAR the p-value is 0.022 and for the slope associated with CESD the p-value is less than 0.001. Here are the R commands that were used to compute these p-values, again assuming the data have been read into the R variable `A3B3C`:

```r
dif = A3B3C$cort1 - A3B3C$cort2
flag2 = A3B3C$CESD < 16
olshc4(cbind(dif[flag2], A3B3C$CESD[flag2]), A3B3C$MAPAGLOB[flag2], xout = T).
```

It is noted, though, that if CESD is ignored, again no association between the CAR and MAPA is found when testing at the 0.05 level; the p-value is 0.09. So this example touches on two points: curvature can mask an association when it is assumed that a regression surface is a plane and detecting whether an independent variable is associated with a dependent variable can depend on the other independent variables included in the model.

RUNNING INTERVAL SMOOTHER:

The running interval smoother is a good choice when dealing with a robust measure of location. It can be applied via the R function

105
Figure 38: Shown is the smooth created by the R function \texttt{lplot} using the CAR and CESD to predict the typical MAPA score. Notice the distinct bend close to where CESD is equal to 16. Focusing on only those participants who have a CESD score less than 16, an association is found between CAR and MAPA, in contrast to an analysis where any possibility of curvature is ignored.
rplot( (x, y, est = tmean, scat = TRUE, fr = NA, plotit = TRUE, pyhat = FALSE, efr = 0.5, theta = 50, phi = 25, scale = TRUE, expand = 0.5, SEED = TRUE, varfun = pbvar, outfun = outpro, mmin = 0, xout = FALSE, out = FALSE, eout = FALSE, xlab = "X", ylab = "Y", zscale = FALSE, zlab = "", pr = TRUE, duplicate = "error", ticktype = "simple", LP = TRUE, ...)  

NOTE: rplot has been updated since the publication of my 2012 book. LP=TRUE: first get the usual smooth then smooth again using LOESS  

Also, ticktype = "simple" is the default setting meaning that for 3D plots, no values on the axes are printed. ticktype = "detail", values will be printed.  

The argument est determines the measure of location that is used and defaults to a 20% trimmed mean. The argument fr corresponds to the span \((f)\) and defaults to 0.8. The function returns the \(\hat{Y}_i\) values if the argument pyhat is set to TRUE. By default, a scatterplot of the points is created with a plot of the smooth. To avoid the scatterplot, set scat=F. The function  

\[
\text{rplotsm}(x,y,\text{est=tmean,fr=1,plotit=T,pyhat=F,}\text{nboot=40,}
\text{ atr=0,nmin=0,outfun=out,eout=F,xlab="X",ylab="Y",scat=T,}
\text{ SEED=T,expand=.5,scale=F, varfun=pbvar,pr=T,...})
\]

estimates the regression line with the running-interval smoother in conjunction with bootstrap bagging.  

The R function  

\[
\text{kerreg}(x, y, \text{pyhat = F, pts = NA, plotit = T, theta = 50, phi = 25, expand = 0.5, scale = F, zscale = F, eout = F, xout = F, outfun = out, np = 100, xlab = "X", ylab = "Y", zlab="", varfun = pbvar, e.pow = T})
\]

computes a kernel smoother. If \(pts\) is a matrix having \(p\) columns, and \(pyhat=T\), the function estimates \(Y\) for each row in \(pts\).  

The R function  

\[
\text{runpd}(x,y, \text{est=tmean, fr=0.8, plotit=T, pyhat=F, nmin = 0, theta=50, phi=25, expand=0.5, scale=F, xlab="X", ylab="Y", MC=F, ...})
\]

computes the running-interval smoother based on projection distances. By default it estimate the 20% trimmed mean of \(Y\) given values for \(X_1, \ldots, X_p\). With \(MC=T\), a multi-core processor will be used if one is available.
The function

\[ qsmcobs(x, y, sqval = 0.5, xlab = "X", ylab = "Y", FIT = T, pc = ".", plotit = T) \]

performs COBS. The quantile to be estimated corresponds to the argument \( qval \). Unlike the other smoothers in this section, this particular smoother is limited to a single predictor.

**MORE R functions based on smoothers:**

- rplotCI: For one predictor, same as rplot but includes a 1-alpha confidence band.
- rplotpbCI: Like rplotCI, only use percentile bootstrap.

The function

\[ qhdsm(x, y, qval = 0.5, q = NULL, pr = FALSE, xout = FALSE, outfun = outpro, plotit = TRUE, xlab = "X", ylab = "Y", zlab = "Z", pyhat = FALSE, fr = NULL, LP = TRUE, theta = 50, phi = 25, ticktype = "simple", nmin = 0, scale = FALSE, pr.qhd = TRUE, ...) \]

does a smooth for one or more quantiles using the running interval smoother in conjunction with Harrell–Davis estimator. Seems that it is generally better than qsmcobs.

### 22.5 Comparing the Slopes of Two Independent Groups Based on Robust Estimator

Consider two independent groups and imagine that the goal is to test

\[ H_0 : \beta_{11} = \beta_{12}, \]  

(28)

the hypothesis that the two groups have equal slopes. Use percentile bootstrap method.

The R function

\[ \text{reg2ci}(x1, y1, x2, y2, \text{regfun}=\text{tsreg}, \text{nboot}=599, \alpha=0.05, \text{plotit}=T) \]

compares the slopes of two groups. The data for group 1 are stored in \( x1 \) and \( y1 \), and for group 2 they are stored in \( x2 \) and \( y2 \). As usual, \( \text{nboot} \) is \( B \), the number of bootstrap samples, \( \text{regfun} \) indicates which regression estimator is to be used and defaults to the Theil-Sen estimator, and \( \text{plotit}=T \) creates a plot of the bootstrap estimates.

To provide some visual sense of how the regression lines differ, and to provide an informal check on whether both regression lines are reasonably straight, the R function
runmean2g(x1, y1, x2, y2, fr=0.8, est=tmean, xlab="X", ylab="Y", ...)

has been supplied. It creates a scatterplot for both groups (with a + used to indicate points that correspond to the second group) and it plots an estimate of the regression lines for both groups using the running-interval smoother. By default, it estimates the 20% trimmed mean of \( Y \), given \( X \). But some other measure of location can be used via the argument \texttt{est}. The smooth for the first group is indicated by a solid line, and a dashed line is used for the other.

The \texttt{R} function

\[
\texttt{l2plot(x1, y1, x2, y2, span=2/3, xlab = "X", ylab = "Y")}
\]

also plots smoothers for each group, only it uses \texttt{LOESS} to estimate the regression lines.

\textit{EXAMPLE}

\section*{23 Robust Measures of Association}

The \texttt{R} Function

\[
\texttt{wincor(x,y, tr=.2)}
\]

computes the Winsorized correlation and tests the hypothesis \( H_0: \rho_w = 0 \) using the Student’s \texttt{T} test. The amount of Winsorization is controlled by the argument \texttt{tr} and defaults to 0.2.

Winsorized correlation does not take into account the overall structure of the data when dealing with outliers. The skipped correlation does. It uses the projection method to remove outliers and then computes a correlation using remaining data.

The \texttt{R} Function

\[
\texttt{scor(x,y, plotit = T, xlab = "VAR 1", ylab = "VAR 2", MC = F)}
\]

computes the skipped correlation. It tests the hypothesis of a zero correlation, but using a method that does not deal with heteroscedasticity. A percentile bootstrap that does deal with heteroscedasticity can be applied with the \texttt{R} function

\[
\texttt{scorci(x,y, nboot = 1000, SEED = TRUE, plotit = TRUE)}
\]
The R Function

\[ \text{corb}(x,y, \text{corfun} = \text{pbcor}, n\text{boot} = 599, \ldots) \]

can be used to test hypotheses and compute confidence intervals for any robust correlation. By default, the function uses the percentage bend correlation, which is not covered here. (See Wilcox, 2012, Section 9.3.1 for details.) It is an M-type correlation.

## 23.1 Measuring the Strength of an Association Based on a Robust Fit

Goal: Given a fit to the data based on a robust regression estimator or smoother, measure the strength of the association. Use a simple generalizations of the notion of explanatory power, which was studied in a general context by Doksum and Samarov (1995). Taking \( \tilde{Y} \) to be the predicted value of \( Y \) based on any regression estimator or smoother, let \( \tau^2(Y) \) be any measure of variation. Then a robust analog of *explanatory power* is

\[
\eta^2 = \frac{\tau^2(\tilde{Y})}{\tau^2(Y)}.
\]  

(29)

The *explanatory strength of association* is the (positive) square root of explanatory power, \( \eta \).

To put \( \eta^2 \) in perspective, if \( \tilde{Y} \) is the predicted value of \( Y \) based on the least squares regression line, and \( \tau^2 \) is the usual variance, then \( \eta^2 \) reduces to the squared multiple correlation coefficient. In the case of a single predictor, still using least squares regression, \( \eta \) is just Pearson’s correlation, \( \rho \), assuming that the sign of \( \eta \) is taken to be the sign of the slope of the least squares regression line.

Choice of smoother matters in terms of getting an accurate estimate of explanatory power. Using the R function lplot performs relatively well.

## 23.2 Tests for Linearity

The R function

\[ \text{ltest}(x,y, \text{regfun}=\text{tsreg}, n\text{boot}=500, \alpha=.05) \]

tests the hypothesis that a regression surface is a plane. It uses a wild bootstrap method.
The R function

\[
\text{lintestMC}(x, y, \text{regfun}=\text{tsreg}, \text{nboot}=500, \text{alpha}=.05)
\]

is the same as \text{lintest}, only it takes advantage of a multi-core processor, if one is available, with the goal of reducing execution time.

### 24 Moderator Analysis

Standard approach uses least squares assuming that

\[
Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_1X_2 + e.
\]  

(30)

Example: predictors of reading ability.

```r
model=lm(y~x[,1]*x[,2])
summary.aov(model)
```

<table>
<thead>
<tr>
<th></th>
<th>Df</th>
<th>Sum Sq</th>
<th>Mean Sq</th>
<th>F value</th>
<th>Pr(&gt;F)</th>
</tr>
</thead>
<tbody>
<tr>
<td>x[, 1]</td>
<td>1</td>
<td>6997.9</td>
<td>6997.9</td>
<td>30.0057</td>
<td>5.224e-07 ***</td>
</tr>
<tr>
<td>x[, 2]</td>
<td>1</td>
<td>3247.4</td>
<td>3247.4</td>
<td>13.9240</td>
<td>0.0003623 ***</td>
</tr>
<tr>
<td>x[, 1]:x[, 2]</td>
<td>1</td>
<td>86.9</td>
<td>86.9</td>
<td>0.3724</td>
<td>0.5434918</td>
</tr>
<tr>
<td>Residuals</td>
<td>77</td>
<td>17957.9</td>
<td>233.2</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

To deal with heteroscedasticity, still assuming normality, the R function \text{olshc4} (or the R function \text{olswbtest}) can be used. Now the appropriate R commands are

```r
xx=cbind(x,x[,1]*x[,2])
olshc4(xx,y)
```

To deal with both heteroscedasticity and nonnormality, and simultaneously remove leverage points, one possibility is

```r
xx=cbind(x,x[,1]*x[,2])
regci(xx,y,xout=T,outfun=outpro)
```
Now the p-value for $H_0: \beta_3 = 0$ is 0.023

A more flexible approach for establishing that there is an interaction is to test

$$H_0: Y = \beta_0 + f_1(X_1) + f_2(X_2) + e,$$

the hypothesis that for some unknown functions $f_1$ and $f_2$, a generalized additive model fits the data, versus the alternative hypothesis

$$H_1: Y = \beta_0 + f_1(X_1) + f_2(X_2) + f_3(X_1, X_2) + e.$$

The R function

```
adtest(x, y, nboot=100, alpha=0.05, xout=F, outfun=out, SEED=T, ...)```

tests this hypothesis.

Could have substantially more power, as well as substantially less power, versus the robust method just illustrated. (If the regression model used by the robust method is correct, the expectation is that it might have a substantial power advantage compared to using adtest.)

**EXAMPLE**

A portion of a study conducted by Shelley Tom and David Schwartz dealt with the association between a Totagg score and two predictors: grade point average (GPA) and a measure of academic engagement. The Totagg score was a sum of peer nomination items that were based on an inventory that included descriptors focusing on adolescents’ behaviors and social standing. (The peer nomination items were obtained by giving children a roster sheet and asking them to nominate a certain amount of peers who fit particular behavioral descriptors.) The sample size is $n = 336$. Assuming that the model given by Equation (30) is true, the hypothesis of no interaction ($H_0: \beta_3 = 0$) is not rejected using the least squares estimator. The p-value returned by the R function olswbtest is .6. (And the p-value returned by olsbc4 is 0.64.) But look at the left panel of Figure 39, which shows the plot of the regression surface assuming Equation (30) is true. (This plot was created with the R function ols.plot.inter.) And compare this to the right panel, which is an estimate of the regression surface using LOESS (created by the R function lplot). This suggests that using the usual interaction model is unsatisfactory for the situation at hand. The R function adtest returns a p-value less than .01 indicating that an interaction exists.

24.1 R Functions kercon, runsm2g, regi, ols.plot.inter and reg.plot.inter

The R function
Figure 39: The left panel shows the plot created by `ols.plot.inter`, which assumes that an interaction can be modeled with $Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_1 X_2 + \epsilon$ and where the least squares estimate of the parameters is used. The right panel shows an approximation of the regression surface based on the R function `lplot`
kercon(x,y,cval=NA,eout=F,xout=F, outfun=out,xlab="X",ylab="Y"),

creates the first of the plots mentioned in the previous section. The argument x is assumed to be a matrix with two columns. By default, three plots are created: a smooth of Y and X₁, given that X₂ is equal to its lower quartile, its median, and its upper quartile. Different choices for the X₂ values can be specified via the argument cval.

The R function

runsm2g(x1,y,x2, val = median(x2), est = tmean, sm = F, xlab = "X", ylab = "Y", ...)

splits the data in x1 and y into two groups based on the value in the argument val and the data stored in the argument x2. By default, a median split is used. The function returns a smooth for both groups. If there is no interaction, these two smooths should be approximately parallel. The smooths are based on the goal of estimating the trimmed mean of the outcome variable. But other measures of location can be used via the argument est.

The R function

regi(x,y,z,pt=median(z),est=onestep,regfun=tsreg, testit=F,...)

creates two smooths in a manner similar to the function runsm2g. In fact this function simply calls the function runsm2g, only it uses by default a one-step M-estimator rather than a 20% trimmed mean. An advantage of this function is that it also has an option for replacing the smoother with a robust regression estimator, via the argument regfun, which by default is the Theil - Sen estimator. This is done by setting the argument testit=T, in which case the function also tests the hypothesis that the two regression lines have equal slopes. Rejecting the hypothesis of equal slopes indicates an interaction.

The R function

ols.plot.inter(x, y, pyhat = F, eout = F, xout = F, outfun = out, plotit = T, expand = 0.5, scale = F, xlab = "X", ylab = "Y", zlab = ", theta = 50, phi = 25, family = "gaussian", duplicate = "error")

plots the regression surface assuming that the commonly used interaction model

\[ Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_1 X_2 + e \]

is true and that the least squares estimate of the parameters is used.

The R function
reg.plot.inter(x, y, regfun=tsreg, pyhat = F, cout = F, outfun = out, plotit = T, expand = 0.5, scale = F, xlab = "X", ylab = "Y", zlab = "", theta = 50, phi = 25, family = "gaussian", duplicate = "error")

is exactly like the function ols.plot.inter, only it can be used with any regression estimator that returns the residuals in $\text{residuals}$. By default, the Theil - Sen estimator is used.

## 25 Mediation Analysis

Brief comments about *mediation analysis*. (For a book dedicated to this topic, see MacKinnon, 2008.) Mediation analysis is similar to a moderator analysis in the sense that the goal is to understand how the association between two variables is related to a third (mediating) variable. In the parlance of researchers working on this problem, an *indirect effect*, also known as a *mediation effect*, refers to a situation where two variables of interest are associated via a third variable. For example, stress and obesity are believed to be associated through cortisol secretion (Rosmond et al., 1998). The strategy behind a mediation analysis is to assume that the three variables of interest satisfy three linear models. The first is that two primary variables of interest $x$ and $y$ (e.g., stress and obesity) are related via the usual linear model

$$
y = \beta_{01} + \beta_{11}x + \epsilon_1. \quad (32)
$$

The second assumption is that the mediating variable (cortisol in the example), which here is labeled $x_m$, is related to $x$ via

$$
x_m = \beta_{02} + \beta_{12}x + \epsilon_2. \quad (33)
$$

And finally, it is assumed that

$$
y = \beta_{03} + \beta_{13}x + \beta_{23}x_m + \epsilon_3. \quad (34)
$$

Roughly, if $\beta_{13} = 0$, this is said to constitute full mediation (Judd and Kenny, 1981a, 1981b). If the strength of the association between of $x$ and $y$ is reduced when the mediator is included, in the sense that $\beta_{13} < \beta_{11}$, there is said to be partial mediation.

Various strategies have been proposed for assessing whether $x_m$ mediates the association between $y$ and $x$. One is to focus on $\beta_{11} - \beta_{13}$. Another is to focus on the product $\beta_{12}\beta_{23}$, which has been called the *mediated effect* or *indirect effect*. One way this latter approach arises is that under certain circumstances (normality and when using the least squares estimator) the total effect represented by the slope in Eq. (32) satisfies $\beta_{11} = \beta_{12}\beta_{23} + \beta_{13}$ (MacKinnon et al., 1995). But even without this assumption, a common goal is testing

$$
H_0 : \beta_{12}\beta_{23} = 0. \quad (35)
$$
Under normality and homoscedasticity, a bootstrap method for testing this hypothesis, using the least squares estimator, has been found to perform reasonably well in simulations. But under nonnormality, or when there is heteroscedasticity, this is no longer the case (Ng, 2009a). Replacing the least squares estimator with the Theil-Sen estimator, Ng (2009a) found that a percentile bootstrap method performs well in simulations when $\beta_{12} = \beta_{23} = 0$. But otherwise, control over the probability of a Type I error can be unsatisfactory in some situations. Consequently, when the goal is to test Eq. (35), a better approach appears to be to test the two hypotheses $H_0 : \beta_{12} = 0$ and $H_0 : \beta_{23} = 0$. If both hypotheses are rejected, also reject the hypothesis given by Eq. (35). So, for example, the R function regci might be used and the probability of at least one Type I error could be controlled using Hochberg’s method.

Recently, Zu and Yuan (2010) derived an approach to testing Eq. (35) based on a Huber-type M-estimator. Their method is based on an appropriate estimate of the standard error, which is used in conjunction with a bootstrap method to control the probability of a type I error. Currently, it is unknown how well the method performs when dealing with heteroscedasticity. (They have supplied an R function for applying their method.) Other recent results relevant to the effect of nonnormality are reported by Biesanz et al. (2010).

For an approach to mediation analysis based on a least absolute value regression estimator, see Yuan and MacKinnon (2014). Their method allows heteroscedasticity. They used a non-bootstrap method that is based on an asymptotic expression for the standard error. Their simulation results are limited to symmetric distributions.

## 26 ANCOVA

One general goal is test

$$H_0 : m_1(X) = m_2(X), \text{ for all } X$$

That is, the regression lines do not differ in any manner.

Another goal is determining where the regression lines differ and by how much.

Methods are now available that allow both types of heteroscedasticity, non-parallel regression lines and outliers. There are even substantially improved methods dealing with curvature in a very flexible manner. The practical importance of the latter methods cannot be stressed enough.
26.1 R Functions ancsm, Qancsm, ancova, ancpb, ancbbpb and ancboot

The R function

\[
\text{ancsm}(x_1, y_1, x_2, y_2, \text{nboot}=200, \text{plotit}=\text{T}, \text{sm}=\text{F}, \text{tr}=0.2)
\]

tests the hypothesis that two independent groups have identical regression lines. That is, it performs a global nonparametric ANCOVA method aimed at testing (36) based on the running interval smoother in conjunction with trimmed means. The arguments \(x_1\) and \(y_1\) contain the data for group 1, and the group 2 data are assumed to be stored in the next two arguments, \(x_2\) and \(y_2\). The R function

\[
\text{Qancsm}(x_1, y_1, x_2, y_2, \text{nboot} = 200, \text{qval} = 0.5, \text{xlab} = "X", \text{ylab} = "Y", \text{plotit} = \text{T})
\]

is another global method. The only difference from \text{ancsm} is that it is based on a quantile regression smoother. By default, the smoother is aimed at estimating the (conditional) median of \(Y\), given \(X\). Other quantiles can be used via the argument \text{qval}, but for quantiles other than the median, it is unknown how well the method controls the probability of a Type I error.

The R function

\[
\text{ancova}(x_1,y_1,x_2,y_2,\text{fr1}=1,\text{fr2}=1,\text{tr}=0.2,\text{alpha}=0.05,\text{plotit}=\text{T},\text{pts} = \text{NA})
\]

performs a local nonparametric ANCOVA analysis based on the running interval smoother, which defaults to estimating the 20% trimmed of \(Y\) given \(X\). The arguments \(x_1\), \(y_1\), \(x_2\), \(y_2\), \(\text{tr}\) and \(\text{alpha}\) have their usual meanings. The arguments \(\text{fr1}\) and \(\text{fr2}\) are the spans used for groups one and two respectively. The argument \(\text{pts}\) can be used to specify the \(X\) values at which the two groups are to be compared. For example, \(\text{pts}=12\) will result in comparing the trimmed mean for group 1 (based on the \(y_1\) values) to the trimmed mean of group 2 given that \(X = 12\). If there is no trimming, the null hypothesis is \(H_0: E(Y_1|X = 12) = E(Y_2|X = 12)\), where \(Y_1\) and \(Y_2\) are the outcome variables of interest corresponding to the two groups. Using \(\text{pts}=c(22,36)\) will result in testing two hypotheses. The first is \(H_0: m_1(22) = m_2(22)\) and the second is \(H_0: m_1(36) = m_2(36)\). If no values for \(\text{pts}\) are specified, then the function picks five \(X\) values and performs the appropriate tests. The values that it picks are reported in the output as illustrated below. Generally, this function controls FWE using the method in Wilcox (2012, section 13.1.8). If \(\text{plotit} = \text{T}\) is used, the function also creates a scatterplot and smooth for both groups with a + and a dashed line indicating the points and the smooth, respectively, for group 2.

The function

\[
117
\]
is like the R function ancova only a percentile bootstrap method is used to test hypotheses and by default the measure of location is the one-step M-estimator. Now FWE is controlled as described in Section 13.1.11. In essence, the function creates groups based on the values in \texttt{pts}, in conjunction with the strategy behind the smooth, it creates the appropriate set of linear contrasts, and then it calls the function \texttt{pbmcp}.

The function

\begin{verbatim}
ancboot(x1,y1,x2,y2,fr1=1,fr2=1,tr=0.2,nboot=599,pts=NA,plotit = T)
\end{verbatim}

compares trimmed means using a bootstrap-t method.

The function

\begin{verbatim}
ancbbpb(x1,y1,x2,y2, fr1 = 1, fr2 = 1, nboot = 200, pts = NA, plotit = T, SEED = T, alpha = 0.05)
\end{verbatim}

is the same as \texttt{ancova}, only a bootstrap bagging method is used to estimate the regression line. Roughly, \textit{bootstrap bagging} means that $B$ bootstrap samples are taken, for each bootstrap sample a smoother is applied, and the final estimate of the regression line is based on the average of the $B$ bootstrap estimates. (See Wilcox, 2009b, for more details.) The practical advantage is that it might provide higher power than \texttt{ancova}. A negative feature is that execution time can be high.

\textit{EXAMPLE}

Do males and females differ in terms of depressive symptoms when a measure of interpersonal support (PEOP) is used as a covariate? Classic ANCOVA: no. Assuming straight regression lines and using a robust regression estimator: no. But when using a smoother that deals with curvature in a flexible method: yes. For PEOP relatively low, females tend to have higher depressive symptoms. Figure 40 shows the plot created by the R function ancova.

The ANCOVA methods described in this section are illustrated with the Pygmalion data. The goal is to compare posttest scores for the two groups taking into account the pretest scores. If the data for the experimental group are stored in the R matrix \texttt{pyge}, with the pretest scores in column 1, and the data for the control group are stored in \texttt{pygc}, the command

\begin{verbatim}
ancpb(x1,y1,x2,y2,est=onestep,pts=NA,fr1=1,fr2=1,
nboot=599,plotit=T,...)
\end{verbatim}
Figure 40: A comparison of males (solid line) and females (dotted line) based on a measure of depressive symptoms using a measure of personal support as a covariate.
ancsm(pyge[,1],pyge[,2],pygc[,1],pygc[,2])

returns p-value of 0.025, indicating that the regression lines differ for some values of the covariate.

ancova(pyge[,1],pyge[,2],pygc[,1],pygc[,2])

returns

<table>
<thead>
<tr>
<th>X</th>
<th>n1</th>
<th>n2</th>
<th>DIF</th>
<th>TEST</th>
<th>se</th>
<th>ci.low</th>
<th>ci.hi</th>
</tr>
</thead>
<tbody>
<tr>
<td>72</td>
<td>12</td>
<td>63</td>
<td>13.39</td>
<td>1.85</td>
<td>7.24</td>
<td>-9.02</td>
<td>35.79</td>
</tr>
<tr>
<td>82</td>
<td>16</td>
<td>68</td>
<td>14.79</td>
<td>1.93</td>
<td>7.68</td>
<td>-8.21</td>
<td>37.80</td>
</tr>
<tr>
<td>101</td>
<td>14</td>
<td>59</td>
<td>22.43</td>
<td>1.43</td>
<td>15.67</td>
<td>-26.24</td>
<td>71.11</td>
</tr>
<tr>
<td>111</td>
<td>12</td>
<td>47</td>
<td>23.79</td>
<td>1.32</td>
<td>17.99</td>
<td>-35.64</td>
<td>83.22</td>
</tr>
<tr>
<td>114</td>
<td>12</td>
<td>43</td>
<td>21.50</td>
<td>1.19</td>
<td>18.02</td>
<td>-37.83</td>
<td>81.03</td>
</tr>
</tbody>
</table>

The first column headed by X says that posttest scores are being compared given that pretest scores (X) have the values 72, 82, 101, 111 and 114. The sample sizes used to make the comparisons are given in the next two columns. For example, when X = 72, there are twelve observations being used from the experimental group and sixty-three from the control. That is, there are twelve pretest scores in the experimental group and sixty-three values in the control group that are close to X = 72. The column headed by DIF contains the estimated difference between the trimmed means. For example, the estimated difference between the posttest scores, given that X = 72, is 13.39. The last two columns indicate the ends of the confidence intervals. These confidence intervals are designed so that FWE is approximately \( \alpha \). The critical value is also reported and is 3.33 for the situation here. All of the confidence intervals contain zero, none of the tests statistics exceeds the critical value, so we fail to detect any differences between posttest scores taking into account the pretest scores of these individuals.

If we apply the function \texttt{ancpb} with the argument \texttt{est=mom}, a portion of the output is

\[
\begin{array}{cccccc}
\text{X} & \text{n1} & \text{n2} & \text{DIF} & \text{TEST} & \text{se} \\
72 & 12 & 63 & 13.39 & 1.85 & 7.24 \\
82 & 16 & 68 & 14.79 & 1.93 & 7.68 \\
101 & 14 & 59 & 22.43 & 1.43 & 15.67 \\
111 & 12 & 47 & 23.79 & 1.32 & 17.99 \\
114 & 12 & 43 & 21.50 & 1.19 & 18.02 \\
\end{array}
\]
Again we fail to find any differences. However, using the function `ancbbpb`, which uses bootstrap bagging, the results are

```
X  n1  n2  DIF  ci.low  ci.hi  p.value
[1,]  72  12  63  12.03672 -2.7430218  22.37841  0.16528926
[2,]  82  16  68  16.24183  0.8489948  24.69427  0.03007519
[3,] 101  14  59  28.32713  3.5099184  48.23010  0.01398601
[4,] 111  12  47  31.94660  8.9667976  70.64249  0.01273885
[5,] 114  12  43  34.23546  7.4661331  71.24803  0.01149425
```

So now we reject at the 0.05 level for four of the five designs points. The R function `ancbbpb` might provide more power than `ancova`, but this comes at the cost of higher execution time.

**SOME RECENT ADDITIONS TO THE R PACKAGE**

`ancGLOB`: like `ancova`, allows curvature, but tests the global hypothesis that $M(Y|X = x)$ is the same for two independent groups based on a collection of $x$ values.

```
ancGLOB( x1, y1, x2, y2, xout = FALSE, outfun = outpro, est = tmean, p.crit = NULL,
         iter = 500, alpha = 0.05, pr = TRUE, nboot = 500, SEED = TRUE, MC = FALSE, CR = FALSE,
         nmin = 12, pts = NULL, fr1 = 1, fr2 = 1, plotit = TRUE, xlab = "X", ylab = "Y", LP = TRUE, cpp = FALSE, ...)
```

For dependent groups, use

```
DancGLOB( x1 = NULL, y1 = NULL, x2 = NULL, y2 = NULL, xy = NULL, fr1 = 1, fr2 = 1, p.crit = NULL, est = tmean, alpha = 0.05, plotit = TRUE, xlab = "X", ylab = "Y",
         pts = NULL, qvals = c(0.25, 0.5, 0.75), sm = FALSE, xout = FALSE, eout = FALSE,
         outfun = out, DIF = FALSE, LP = TRUE, nboot = 500, SEED = TRUE, iter = 2000, MC = TRUE, cpp = TRUE, nmin = 12, q = 0.5, ...)
```

**ADVANTAGE: POSSIBLY A SUBSTANTIAL INCREASE IN POWER VS INDIVIDUAL TESTS PERFORMED AT THE P COVARIATE VALUES**

**DISADVANTAGE: DON’T KNOW AT WHICH POINTS THERE IS A SIG. DIFFERENCE.**
ancovaWMW like ancova, only use Cliff modification of the Wilcoxon-Mann-Whitney test.

FOR TWO OR MORE GROUPS, FIT A PARAMETRIC REGRESSION LINE FOR THE DATA IN EACH GROUP. THEN

\[
\text{reg1wayISO}(x, y, \text{regfun}=\text{tsreg}, nboot=100, \text{SEED}=\text{TRUE}, \text{xout}=\text{FALSE}, \text{outfun}=\text{outpro}, \\
\text{STAND}=\text{TRUE}, \text{AD}=\text{FALSE}, \text{alpha}=.05, \text{pr}=\text{TRUE}, \ldots)
\]

tests equality of slopes only. That is, perform a global test and do not include the hypothesis that the intercepts are equal. The data for each group are assumed to be stored in list mode. That is, x and y are assumed to have list mode having length J equal to the number of groups. For example, x[[1]] and y[[1]] contain the data for group 1.

The function

\[
\text{reg1way}(x, y, \text{regfun}=\text{tsreg}, nboot=100, \text{SEED}=\text{TRUE}, \text{xout}=\text{FALSE}, \text{outfun}=\text{outpro}, \\
\text{STAND}=\text{TRUE}, \text{AD}=\text{FALSE}, \text{alpha}=.05, \text{pr}=\text{TRUE}, \ldots)
\]

Tests hypothesis that for two or more independent groups, all regression parameters (the intercepts and slopes) are equal. By default the Theil–Sen estimator is used.

\text{reg1wayMC} and \text{reg1wayISOMC} use a multicore processor.

\text{regGmcp} does all pairwise comparisons where the null for each pair is that all parameters are equal.

FOR OLS, use \text{ols1way}; \text{ols1wayISO} compares slopes only.

\text{ancGpar}(x_1, y_1, x_2, y_2): \text{Supplied for convenience. Same as \text{reg1way} only it handles two groups only and data are not stored in list mode as assumed by \text{reg1way}}

\text{ancGparMC} uses multicore processor.

For OLS, use \text{olsJ2}

\text{ancpar}: \text{For two independent groups, compare Y hat at specified design points. One or more covariates is allowed. It fits a robust parametric linear model. By default, uses Theil–Sen. Covariate values are chosen as done by the function ancova if none are specified. Basically, a robust analog of the Johnson-Neyman method. This function uses the functions ancts and anctsmcp.}
26.2 Multiple Covariates

There are various ways multiple covariates might be handled. Momentarily focus on the \( i \)th value of the covariate in the first group, \( x_{i1} \). Then it is a simple matter to determine the set of observed points close to \( x_{i1} \). The same can be done for the second group, in which case we proceed as done by the R function `ancova`.

The R function

\[
\text{ancovamp}(x1,y1,x2,y2,fr1=1,fr2=1,tr=.2,alpha=.05,pts=NA)
\]

compares two groups based on trimmed means and takes into account multiple covariates.

The R function

\[
\text{ancovampG}(x1,y1,x2,y2,fr1=1,fr2=1,tr=.2,alpha=.05,pts=NULL,DH=FALSE)
\]

might provide substantially more power by setting the argument \( \text{DH} = \text{TRUE} \).

The function

\[
\text{ancmppb}(x1,y1,x2,y2,fr1=1,fr2=1,tr=.2,alpha=.05,pts=NA,est=tmean,
\quad nboot=NA,bhop=F,...)
\]

is like `ancovamp`, only a percentile bootstrap method is used and any measure of location can be employed.

Several new and improved ANCOVA will appear in the 4th edition of my book: *Introduction to Robust Estimation and Hypothesis Testing*

26.3 ANCOVA with two covariates when there is curvature

ANCOVA with two covariates can be performed in a manner that allows curvature. (Recent advances are covered in the 4th edition of my book on robust methods, due out around Jan. 1, 2017.

Data from the Well Elderly 2 study (Clark et al., 2011; Jackson et al., 2009) are used to illustrate an ANCOVA method for two covariates when there is curvature. A general goal in the Well Elderly 2 study was to assess the efficacy of an intervention strategy aimed at improving the physical and emotional health of older adults. A portion of the study was
aimed at understanding the impact of intervention on a measure of self-perceived physical health and mental well-being, which was based on the RAND 36-item (SF36) Health Survey (Hays, 1993; McHorney et al., 1993). Higher scores reflect greater perceived health and well-being. There were two covariates. The first is a measure of depressive symptoms based on the Center for Epidemiologic Studies Depressive Scale (CESD). The CESD (Radloff, 1977) is sensitive to change in depressive status over time and has been successfully used to assess ethnically diverse older people (Lewinsohn et al., 1988; Foley et al., 2002). Higher scores indicate a higher level of depressive symptoms. The other covariate was the cortisol awakening response (CAR), which is defined as the change in cortisol concentration that occurs during the first hour after waking from sleep. Extant studies (e.g., Clow et al., 2004; Steptoe, 2007; Chida & Steptoe, 2009) indicate that measures of stress are associated with the CAR. (The CAR is taken to be the cortisol level upon awakening minus the level of cortisol after the participants were awake for about an hour.) The sample size for the control group was 187 and the sample size for the group that received intervention was 228. Figures 41 shows the regression surface for predicting the typical difference in SF36 scores for a control groups versus an experimental group.

27 Some Final Comments on How To Proceed

Rigid rules about how to gain multiple perspectives seem unwise. Technology keeps changing and improving, and substantive issues might dictate to some extent which perspectives are relatively useful. However, some rough guidelines can be offered:

1. Whenever possible, plot the data. Error bars are popular, they provide some useful information, but they reflect a rather narrow feature of the data. Boxplots, estimates of the distributions, shift function, and smoothers can be invaluable.

2. Be careful about using the mean to the exclusion of all other measures of central tendency. From a substantive point of view, situations might be encountered where the mean is the most meaningful measure or there might be a reason to focus on a particular measure such as the median. As illustrated here, different measures of central tendency can broaden our understanding of data.

3. Consider standardized measures of effect size other than Cohen’s d. If they paint a different picture regarding how groups compare, the reasons for the differences need to be understood.

4. Be aware that the method used to detect outliers can make a practical difference. This is particularly true when dealing with multivariate data.

5. When dealing with regression, consider measures of association beyond the obvious three choices: Pearson, Kendall’s tau and Spearman’s rho. Skipped correlations, as well as measures of association based on some robust regression estimator, are among the possibilities.
Figure 41: Regression surface predicting the typical difference in SF36 scores as a function of the CAR and CESD.
6. When dealing with regression, consider robust estimators including methods that deal with curvature in a flexible manner. For more details regarding the relative merits of the estimators that might be used, see Wilcox (2012b, Chapters 10–11).

7. Routinely use methods that allow heteroscedasticity. Homoscedastic methods are satisfactory in terms of testing the hypothesis that groups have identical distributions or that variables are independent. But when trying to understand how groups differ and how variables are related, there is the concern that typically, homoscedastic methods use the wrong standard error, which might have a substantial impact when testing hypotheses or computing confidence intervals. Heteroscedasticity can now be accommodated when using any of the usual ANOVA designs as well as all of the robust regression estimators that have been derived.

8. Be cautious with models. As Huber (2011, p. 31) puts it: “Insight is gained by thinking in models, but reliance on models can prevent insight.” An example is the model where groups have normal distributions and differ only in terms of their means. This model leads naturally to Cohen’s d, but it can prevent insight as was illustrated.