Stat 5102 Notes: Nonparametric Tests and Confidence Intervals

Charles J. Geyer

April 13, 2003

This handout gives a brief introduction to nonparametrics, which is what you do when you don't believe the assumptions for the stuff we've learned so far.

All frequentist procedures come in natural triples

- a *hypothesis test*,
- the *confidence interval* obtained by "inverting" the test, and
- the *point estimate* obtained by shrinking the confidence level to zero. This is called the *Hodges-Lehmann estimator* associated with the confidence interval.

A familiar example is

hypothesis test	t test
confidence interval	t confidence interval
point estimate	sample mean

Now we are going to learn about some competing triples

hypothesis test	sign test
confidence interval	associated confidence interval
point estimate	sample median

and

hypothesis test	Wilcoxon signed rank test
confidence interval	associated confidence interval
point estimate	associated Hodges-Lehmann estimator

and

1 The Sign Test and Associated Procedures

1.1 Assumptions

Suppose we have independent and identically distributed data X_1, X_2, ..., X_n from some continuous distribution. That is, our assumptions are

- independence

- identical distribution
- continuity

There are no other assumptions. In particular, we do not assume any particular shape for the distribution of the data. Any continuous distribution whatsoever will do.

The continuity assumption assures that ties are impossible. With probability one we have $X_i \neq X_j$ when $i \neq j$. The continuity assumption is necessary for exact hypothesis tests. The continuity assumption is unnecessary for the point estimate and confidence interval.

The parameter of interest is the median θ of the distribution of the X_i. Because of the continuity assumption, the median is uniquely defined. If F is the distribution function of the distribution of the X_i, then there exists exactly one point x such that $F(x) = 1/2$ and that point is the median θ.

1.2 Hypothesis Test

Define

$$Y_i = I_{(0,\infty)}(X_i - \theta).$$

In words, Y_i is the indicator of the sign of $X_i - \theta$.

The Y_i are i. i. d. Bernoulli(1/2) random variables (assuming θ is the median). Hence

$$T = \sum_{i=1}^{n} Y_i$$

is Binomial$(n, 1/2)$ under the assumption that θ is the median.

Thus a sensible test of the hypotheses

$$H_0 : \theta = \theta_0$$
$$H_1 : \theta \neq \theta_0$$

or of the corresponding one-sided hypothesis is based on the distribution of the test statistic T.

For example suppose the data are

−4.7	3.7	22.4	23.5	14.4
13.6	8.7	9.1	20.2	6.5
−7.8	10.8	15.6	10.1	−6.9

and we wish to test the null hypothesis that θ is zero (this is the usual null hypothesis when the data are differences from a paired comparison).

There are 12 positive data points so $T = 12$. Alternatively, we could consider $T = 3$ because there are 3 negative data points. It doesn't matter because the Binomial$(n, 1/2)$ null distribution of the test statistic is symmetric.

The P-value for an upper-tailed test would be

```
> 1 - pbinom(11, 15, 1 / 2)
[1] 0.01757812
```

because of the symmetry of the null distribution of the test statistic, the P-value of the two-tailed test is just twice that of a one-tailed test

```
> 2 * (1 - pbinom(11, 15, 1 / 2))
[1] 0.03515625
```

Either test says rejects the null hypothesis at the conventional 0.05 significance level and says the median appears to be not equal to zero. The evidence, however, is not strong. $P = 0.035$ is not much below $P = 0.05$. There is about one chance in 28 of getting a P-value this big even if the median is zero.

The only thing tricky here is why did we look up 11 rather than 12 when $T = 12$? Well, $P(T \geq 12) = 1 - P(T \leq 11)$ because T is discrete. Because the binomial distribution is discrete, you must be careful about discreteness. It's clear this is the right answer, because when we use the other tail of the distribution and avoid the complement rule, we get the same one-tailed P-value in a more obvious way

```
> pbinom(3, 15, 1 / 2)
[1] 0.01757813
```

and, of course, the same two-tailed P-value when we multiply by two.

1.3 Confidence Interval

In order to "invert" the test to obtain a confidence interval, we need to consider tests of all possible null hypotheses. If the null hypothesis is some θ that is not zero, then the test is based on the signs of the $Y_i = X_i - \theta$.

In theory, we need to consider an infinite collection of null hypotheses, a different test for each real number θ. In practice we only need to consider $n + 1$ null hypotheses, because the sign of $X_i - \theta$ changes only when θ "goes past" X_i. Thus the n data points divide the real line into $n + 1$ intervals (2 semi-infinite exterior intervals and $n - 1$ interior intervals).

The result of the test goes from "reject" to "accept" or vice versa as the θ specified by the null hypothesis goes past one of the data points, thus the endpoints of the confidence interval (the endpoints of the set of θ that are not rejected) are data points. By symmetry (the symmetry of the binomial null distribution of the test statistic) the endpoints will be the same number in from each end when the data values are put in *sorted order*

−7.8	−6.9	−4.7	3.7	6.5
8.7	9.1	10.1	10.8	13.6
14.4	15.6	20.2	22.4	23.5

This means there are only a finite set of possible confidence intervals

$$
\begin{aligned}
(-7.8, 23.5) \\
(-6.9, 22.4) \\
(-4.7, 20.2) \\
(3.7, 15.6)
\end{aligned}
$$

the ends, one in from each end, two in, three in, and so forth.

The only thing left to do is figure out the confidence levels that go with each such interval.

The widest possible interval $(-7.8, 23.5)$ fails to cover θ if and only if θ is not within the range of the data. The corresponding test statistic for the sign test is $T = 0$ or $T = n$ (all the $X_i - \theta$ have the same sign if all the X_i are on the same side of θ). The probability of this happening under the null hypothesis is

$$P(T = 0) + P(T = n) = 2P(T = 0).$$

The next widest interval fails to cover θ if and only if θ is not within the range of the data except for the two extremes. The corresponding test statistic for the sign test is $T \le 1$ or $T \ge n - 1$. The probability of this happening under the null hypothesis is $2P(T \le 1)$.

The general pattern should now be clear. The confidence interval whose endpoints are the data points k in from each end in sorted order has confidence level $1 - 2P(T \le k)$. For sample size 15 the possible confidence levels are

```
> 1 - 2 * pbinom(k, 15, 1 / 2)
[1] 0.99994 0.99902 0.99261 0.96484 0.88153 0.69824 0.39276
```

No one is much interested in confidence intervals that have 99.9% coverage or 69% or 39%. That means the only interesting intervals are

2 in from each end	99.2% confidence interval	$(-4.7, 20.2)$
3 in from each end	96.5% confidence interval	$(3.7, 15.6)$
4 in from each end	88.2% confidence interval	$(6.5, 14.4)$

You can't have a 95% confidence interval. You must pick one of these. On the other hand, these confidence intervals are

- *exact*. They have the exact stated confidence level.
- *nonparametric*. They have the stated confidence level under no assumptions other than that the data are i. i. d.

1.4 Point Estimate

When the confidence level shrinks to zero, the interval shrinks to the middle value in sorted order (or the two middle values when n is even). This is the *sample median*, which by convention is defined to be the middle data value in sorted order when n is odd and the average of the two middle data values in sorted order when n is even.

This general notion of the point estimator derived by sending the confidence level to zero is called a *Hodges-Lehmann* estimator. The Hodges-Lehmann estimator associated with the sign test is the sample median. For our example data, the median is 10.1.

1.5 Summary

Thus the sign test gives us a complete theory about the median. The sample median is the natural point estimator of the median. The confidence interval

dual to the sign test is the natural confidence interval. And the sign test is the natural way to test hypotheses about the median.

This triple of procedures is a complete competitor to the triple based on the mean (sample mean and t test and confidence interval). There's no reason you should only know about the latter. A knowledgeable statistician knows about both and uses whichever is appropriate.

2 The Wilcoxon Signed Rank Test and Associated Procedures

One way to think about what the sign test does is to replace the data X_i by "signs" Y_i (actually indicator variables for signs). We start with numbers and throw away the sizes of the numbers leaving only the signs. This allows us to get an exact nonparametric tests.

But perhaps that throws away too much. Can we throw away less information and still get a nonparametric test? Yes.

An idea that enables us to do this (invented in the 1940's by Wilcoxon) is to replace the data by their *signed ranks*. Each data point is replaced by the integer giving its position in the sorted order by absolute value but keeping its actual sign.

For the example we previously used for the sign test, the data now sorted in order of absolute value are

3.7	−4.7	6.5	−6.9	−7.8
8.7	9.1	10.1	10.8	13.6
14.4	15.6	20.2	22.4	23.5

and the corresponding signed ranks are

1	−2	3	−4	−5
6	7	8	9	10
11	12	13	14	15

Because we are throwing away less information, keeping some information about size as well as information about sign, we can expect to get more powerful tests, more accurate estimation, and so forth. And generally we do.

2.1 Assumptions

Suppose we have independent and identically distributed data X_1, X_2, ..., X_n from some symmetric continuous distribution. That is, our assumptions are

- independence
- identical distribution
- continuity
- symmetry

There are no other assumptions. In particular, we do not assume any particular shape for the distribution of the data. Any symmetric continuous distribution whatsoever will do.

Note that the assumptions are exactly the same as for the sign test with the addition of *symmetry*. Note also that the assumptions are exactly the same as for the t test except that for the t we must strengthen the assumption of *symmetry* to an assumption of *normality*. Thus the assumptions that are different for the three tests are

sign test	any continuous distribution
signed rank test	any symmetric continuous distribution
t test	any normal distribution

The other assumptions (i. i. d.) are the same for all three tests (and their associated confidence intervals and Hodges-Lehmann estimators). Clearly the sign test has the weakest assumptions (assumes the least), the t test has the strongest assumptions (assumes the most), and the Wilcoxon signed rank test is in the middle.

The continuity assumption assures that tied ranks are impossible. With probability one we have $|X_i| \neq |X_j|$ when $i \neq j$. In order to have exact tests, this assumption is necessary. The continuity assumption is unnecessary for the point estimate and confidence interval.

The parameter of interest is the center of symmetry θ of the distribution of the X_i. We know that the center of symmetry is also the median and the mean (if first moments exist). Thus this parameter is sometimes also referred to as the median or mean.

How restrictive is an assumption of symmetry? Fairly restrictive. Many data distributions are obviously skewed. For such distributions the Wilcoxon test is bogus, but the sign test is valid.

The symmetry assumption is sometimes justified by the following argument. When the data consist of pairs (X_i, Y_i), the so-called *paired comparison* situation, we know the standard trick (already used with the t test) is to form the differences $Z_i = X_i - Y_i$ and use the one sample procedure on the Z_i. If we assume X_i and Y_i are independent and assume that $X_i + \theta$ and Y_i are equal in distribution (that is, the distribution of the Y_i is the same as the distribution of the X_i except shifted by the distance θ), then we can conclude that the distribution of the Z_i is symmetric about θ. However, these assumptions about (X_i, Y_i) pairs are neither necessary nor sufficient for this conclusion about the Z_i. Thus this argument should be taken with a grain of salt.

2.2 Hypothesis Test

Define

$$Y_i = (X_i - \theta). \tag{2.1}$$

and let R_i denote the signed ranks corresponding to the Y_i.

Under the null hypothesis that the θ subtracted off in (2.1) is the true parameter value, and under the rest of the assumptions made in the preceding section, the distribution of the R_i is known. The absolute values of the ranks are just the numbers from 1 to n. (Ties are impossible because of the continuity assumption.) The sign of any R_i is equally likely to be plus or minus

because of the symmetry assumption. Thus any function of the R_i has a known distribution under the null hypothesis.

The usual test statistic is the sum of the positive ranks

$$T^{(+)} = \sum_{i=1}^{n} R_i I_{(0,\infty)}(R_i) \tag{2.2}$$

equivalent test statistics linearly related to (2.2) so they produce the same P-values are the sum of the negative ranks

$$T^{(-)} = -\sum_{i=1}^{n} R_i I_{(-\infty,0)}(R_i) \tag{2.3}$$

or the sum of the of the signed ranks

$$T = T^{(+)} - T^{(-)}$$

these statistics are related because the sum of $T^{(+)}$ and $T^{(-)}$ is just the sum of the numbers from 1 to n

$$T^{(+)} + T^{(-)} = \frac{n(n+1)}{2}$$

Hence

$$T^{(-)} = \frac{n(n+1)}{2} - T^{(+)}$$

and

$$T = 2T^{(+)} - \frac{n(n+1)}{2}$$

The distributions under the null hypothesis of $T^{(+)}$ and $T^{(-)}$ are the same, symmetric about $n(n+1)/4$ and ranging from 0 to $n(n+1)/2$. The distribution of T is symmetric about zero and ranges from $-n(n+1)/2$ to $+n(n+1)/2$.

The distribution under the null hypothesis of $T^{(+)}$ or $T^{(-)}$ is not a brand name distribution (at least not one in the textbook or the handout). It is officially called the *null distribution of the Wilcoxon signed rank test statistic* and is calculated by the R function `psignrank`.

The whole test is done by an R function `wilcox.test`. If the data for the example we have been using all along are in a vector `x`, then

```
> wilcox.test(x)

        Wilcoxon signed rank test

data:  x
V = 109, p-value = 0.003357
alternative hypothesis: true mu is not equal to 0
```

does a two-tailed Wilcoxon signed rank test of the null hypothesis $H_0 : \theta = 0$ (just as the printout says).

2.3 Confidence Interval

From the description of the Wilcoxon signed rank test it is not clear how to invert the test to get a confidence interval. Fortunately, there is an equivalent description of the Wilcoxon test for which the inversion is very analogous to the inversion of the sign test.

For data X_1, ..., X_n the $n(n+1)/2$ numbers

$$\frac{X_i + X_j}{2}, \qquad i \le j$$

are called the *Walsh averages*. Let $m = n(n+1)/2$, and let $W_{(1)}$, ..., $W_{(m)}$ denote the Walsh averages in sorted order.

Then it is a (non-obvious) fact that the Wilcoxon signed rank test statistic $T^{(+)}$ is equal to the number of positive Walsh averages. Thus an argument exactly analogous to the argument about inverting the sign test says that possible confidence intervals associated with the signed rank test have the form $(W_{(k+1)}, W_{(m-k)})$, that is, they have endpoints obtained by counting k in from each end of the Walsh averages in sorted order.

For the data in our running example the Walsh averages are

−7.80	−7.35	−6.90	−6.25	−5.80	−4.70	−2.05	−1.60
−0.65	−0.50	−0.20	0.45	0.65	0.90	0.90	1.10
1.15	1.50	1.60	1.95	2.00	2.20	2.70	2.90
3.05	3.30	3.35	3.70	3.75	3.90	4.35	4.45
4.85	5.10	5.45	6.20	6.20	6.40	6.50	6.65
6.90	7.25	7.30	7.60	7.75	7.75	7.80	7.85
8.30	8.30	8.65	8.65	8.70	8.85	8.90	9.05
9.10	9.40	9.40	9.60	9.65	9.75	9.95	10.05
10.10	10.45	10.45	10.80	11.05	11.15	11.35	11.55
11.75	11.85	11.95	12.15	12.20	12.25	12.35	12.60
12.85	13.05	13.20	13.35	13.60	13.60	14.00	14.40
14.45	14.45	14.60	14.65	15.00	15.00	15.15	15.50
15.55	15.60	15.75	16.10	16.25	16.30	16.60	16.80
16.90	17.15	17.30	17.90	18.00	18.40	18.55	18.95
19.00	19.55	20.20	21.30	21.85	22.40	22.95	23.50

The possible confidence intervals are

$$(-7.80, 23.50)$$
$$(-7.35, 22.95)$$
$$(-6.90, 22.40)$$
$$(-6.25, 21.85)$$

and so forth.

The corresponding confidence levels are given by

```
> n <- length(x)
> m <- n * (n + 1) / 2
> k <- 1:(m / 2)
> conf.lev <- 1 - 2 * psignrank(k, n)
```

```
> names(conf.lev) <- k
> round(conf.lev[0.80 < conf.lev & conf.lev < 0.995], 4)
    13     14     15     16     17     18     19     20     21
0.9946 0.9933 0.9916 0.9897 0.9875 0.9849 0.9819 0.9785 0.9744
    22     23     24     25     26     27     28     29     30
0.9698 0.9647 0.9587 0.9521 0.9446 0.9363 0.9270 0.9167 0.9054
    31     32     33     34     35     36
0.8930 0.8795 0.8646 0.8486 0.8312 0.8124
```

To get an actual confidence level of 95.21% we use `k = 25`

```
> w <- outer(x, x, "+") / 2
> w <- w[lower.tri(w, diag = TRUE)]
> w <- sort(w)
> sum(w > 0)
[1] 109
>
> k <- 25
> n <- length(x)
> m <- n * (n + 1) / 2
> c(w[k + 1], w[m - k])
[1]  3.30 15.15
> 1 - 2 * psignrank(k, n)
[1] 0.9520874
```

So that's it. The 95.2% confidence interval associated with the Wilcoxon signed rank test is $(3.30, 15.15)$.

This is all a lot of work, but fortunately there is a function in R that does it all for you.

```
> wilcox.test(x, conf.int = TRUE)

        Wilcoxon signed rank test

data:  x
V = 109, p-value = 0.003357
alternative hypothesis: true mu is not equal to 0
95 percent confidence interval:
  3.30 15.15
sample estimates:
(pseudo)median
         9.625
```

The only grumble about this function is that it doesn't give the actual confidence level, but the level requested. This matters less for the Wilcoxon than for the sign test because the discreteness is less coarse.

2.4 Point Estimate

As the confidence level goes to zero the interval shrinks to the *median of the Walsh averages*, so that is the Hodges-Lehmann estimator of the center of symmetry associated with the Wilcoxon signed rank test.

With the code above calculating the Walsh averages already executed

```
> median(w)
[1] 9.625
```

calculates the Hodges-Lehmann estimator.

The `wilcox.test` function also calculates this point estimate, which it calls the `(pseudo)median`.

The reason for that terminology is that the center of symmetry is the median *assuming the distribution is symmetric*.

When the distribution of the X_i is not symmetric, the median of the Walsh averages does not estimate the median but the median of the distribution of a typical Walsh average $(X_i + X_j)/2$ where X_i and X_j are independent and identically distributed. In order to take care of that case, someone coined the term *pseudomedian* of a distribution F to mean the median of the distribution of $(X_1 + X_2)/2$ where X_1 and X_2 are independent random variables with distribution F.

But having a name (pseudomedian) doesn't make it an interesting parameter. If you don't believe the distribution is symmetric, you probably don't want to use this estimator.

2.5 Summary

The Wilcoxon signed rank test also induces a triple of procedures, with confidence intervals and point estimate based on the Walsh averages.

For comparison we give all three triples

type	level (%)	interval	P-value	estimator
sign	96.5	(3.70, 15.60)	0.0352	10.1
Wilcoxon	95.2	(3.30, 15.15)	0.0034	9.625
Student t	95	(3.80, 14.76)	0.0027	9.28

3 Robustness

So why would one want to use one of these "triples" rather than another? There are three issues: one, *assumptions*, already covered, another the subject of this section, another the subject of the following section.

The word "robustness" is purported to be a technical term of statistics, but, unfortunately, it has been used in many different ways by many different researchers. Some of the ways are quite precise, others are extremely vague. Without a nearby precise definition of the term, "robust" means little more than "good" (in the opinion of the person using the term).

DeGroot and Schervish have a section (their Section 9.7) titled "Robust Estimation" but nowhere do they say what they mean by robustness more specific than

> An estimator that performs well for several different types of distributions, even though it may not be the best for any particular type of distribution, is called a *robust estimator* [their italics].

3.1 Breakdown Point

This section introduces one very precise robustness concept.

The *finite sample breakdown point* of an estimator

$$\hat{\theta}_n = g(x_1, \ldots, x_n) \tag{3.1}$$

is m/n where m is the maximum number of the data points x_{i_1}, ..., x_{i_m} that can be changed arbitrarily while the rest remain fixed and the estimator (3.1) remain bounded.

The *asymptotic breakdown point* is the limit of the finite sample breakdown point as n goes to infinity. Usually, the finite sample breakdown point is complicated and the asymptotic breakdown point is simpler. Hence when we say "breakdown point" without qualification we mean the asymptotic breakdown point.

That's the precise technical definition. The sloppy motivating idea is the breakdown point is the fraction of "junk data" the estimator tolerates. We imagine that some fraction of the data are "junk" also called "outliers" or "gross errors" that may have any values at all. We don't want the estimator to be affected too much by the junk, and that's what the breakdown point concept makes precise.

Example 3.1 (Breakdown Point of the Sample Mean).
The sample mean is

$$\bar{x}_n = \frac{x_1 + \ldots + x_n}{n} \tag{3.2}$$

if we leave x_2, ..., x_n fixed and let $x_1 \to \infty$, then $\bar{x}_n \to \infty$ as well. Hence the finite sample breakdown point is $0/n = 0$. The m in the definition is zero, that is, *no* data points can be arbitrarily changed while the estimator (3.2) remains bounded.

Thus the finite sample and asymptotic breakdown points of the sample mean are both zero. The sample mean tolerates *no* junk data. It should be used only with "perfect" data.

Example 3.2 (Breakdown Point of the Sample Median).
Here we need to distinguish two cases. For n odd, the sample median is the middle value in sorted order. Then the finite sample breakdown point is $(n - 1)/(2n)$ because so long as the majority of the data points, that is, $(n + 1)/2$ of them, remain fixed, the other $(n - 1)/2$ can be arbitrarily changed and the median will always remain somewhere in the range of the fixed data points. For n even, the sample median is the average of the two middle values in sorted order. Then the finite sample breakdown point is $(n/2 - 1)/n$ by a similar argument.

In either case the asymptotic breakdown point is $n/2$. (See why asymptotic breakdown point is the simpler concept?)

The sample median tolerates up to 50% junk data. It is about as robust as an estimator can be (when robustness is measured by breakdown point).

Example 3.3 (Breakdown Point of the Sample Trimmed Mean).
DeGroot and Schervish define the average of the sample with the k smallest and k largest observations the *kth level trimmed mean*. This is their eccentric

terminology. Everyone else calls it a $100(k/n)\%$ trimmed mean, that is, the trimming fraction k/n is expressed as a percent and the word "level" is not used. Note that a 10% trimmed mean "trims" 10% of the data at the low end and 10% of the data at the high end (20% all together).

It is fairly easy to see that the asymptotic breakdown point of an $100\alpha\%$ trimmed mean is just α. The breakdown point is the trimming fraction.

Thus trimmed means can have any desired breakdown point between 0 and 1 / 2. Note that the ordinary sample mean is a 0% trimmed mean and has breakdown point that is its trimming fraction. Also note that the sample median is a 50% trimmed mean, more precisely $50(\frac{n-2}{n})\%$ but the $\frac{n-2}{n}$ becomes irrelevant as n goes to infinity, and has breakdown point that is its trimming fraction.

Thus sample trimmed means form a continuum with the sample mean and the sample median at the ends.

Example 3.4 (Breakdown Point of the Hodges-Lehmann Estimator associated with the Wilcoxon Signed Rank Test).
Each Walsh average involves two data points. If we arbitrarily change m data points and leave $n - m$ fixed, then we leave fixed $(n - m)(n - m + 1)/2$ Walsh averages (and mess up the rest). The median of the Walsh averages will stay bounded so long as the fixed Walsh averages are a majority, that is, so long as

$$\frac{(n-m)(n-m+1)}{2} > \frac{1}{2} \cdot \frac{n(n+1)}{2}$$

If the asymptotic breakdown point is α then $m \approx n\alpha$ for large n and the equation above becomes

$$\frac{n^2(1-\alpha)^2}{2} \gtrsim \frac{1}{2} \cdot \frac{n^2}{2}$$

So $1 - \alpha = \sqrt{1/2}$ and $\alpha = 1 - \sqrt{1/2} = 0.2928932$.

So this Hodges-Lehmann estimator (median of the Walsh averages) is intermediate in robustness (as measured by breakdown point) between the sample mean and sample median and has about the same robustness as a 30% trimmed mean.

4 Asymptotic Relative Efficiency

So if the sample mean is so bad (breakdown point zero, useful only for perfect data), why does anyone ever use it?

The answer is efficiency. Generally, the more robust an estimator is, the less efficient it is. There is an unavoidable trade-off between robustness and efficiency.

We say an estimator $\hat{\theta}_n$ of a parameter θ is consistent and asymptotically normal if

$$\sqrt{n}(\hat{\theta}_n - \theta) \xrightarrow{\mathcal{D}} \text{Normal}(0, \sigma^2)$$

for some constant σ^2 which is called the *asymptotic variance* of the estimator and is not necessarily related to the population variance (all of the estimators discussed in this handout have this property). The *asymptotic relative efficiency* (ARE) of two such estimators is the ratio of their asymptotic variances.

The reason why ratio of variances (as opposed to, say, ratio of standard deviations) is interesting is that variances are proportional to costs in the following sense. Suppose two estimators have asymptotic variances σ_1^2 and σ_2^2, and suppose we want to have equal precision in estimation, say we want asymptotic 95% confidence intervals with half-width ϵ, that is,

$$\epsilon = 1.96 \frac{\sigma_1}{\sqrt{n_1}} = 1.96 \frac{\sigma_2}{\sqrt{n_2}}$$

where n_1 and n_2 are the sample sizes for the two estimators. But this implies

$$\frac{n_1}{n_2} = \frac{\sigma_1^2}{\sigma_2^2} = \text{ARE}$$

If costs are proportional to sample size (not always true, but a reasonable general assumption), then ARE is proportional to costs.

When we say the ARE of two estimators is 2.0, that means one estimator is twice as efficient as the other, but doesn't make clear which is better. That is unavoidable because which estimator we call "estimator 1" and which "estimator 2" is arbitrary. Hence it is good practice to also say which is better.

Although ARE is the proper measure of efficiency, it is very complicated because ARE depends on the true unknown distribution of the data. In order to see this we need to know more about asymptotics. The asymptotic variance of some of the estimators discussed in this handout are rather complicated. We will just state the results without proof. Consider a distribution symmetric about zero with probability density function f and cumulative density function F such that f is everywhere continuous and $f(0) > 0$.

The asymptotic variance of the **sample mean** is the population variance

$$\sigma_{\text{mean}}^2 = \int_{-\infty}^{\infty} x^2 f(x) \, dx.$$

The asymptotic variance of the **sample median** is

$$\sigma_{\text{median}}^2 = \frac{1}{4 f(\theta)^2}.$$

The asymptotic variance of the **Hodges-Lehmann estimator associated with the Wilcoxon signed rank test** is

$$\sigma_{\text{H-L}}^2 = \frac{1}{12 \left[\int_{-\infty}^{\infty} f(x)^2 \, dx \right]^2}.$$

The asymptotic variance of the $100\alpha\%$ **trimmed mean** is

$$\sigma_\alpha^2 = \frac{2}{(1 - 2\alpha)^2} \left[\int_0^{F^{-1}(1-\alpha)} x^2 f(x) \, dx + \alpha F^{-1}(1 - \alpha)^2 \right]$$

The asymptotic variance of the **maximum likelihood estimator** is inverse Fisher information

$$\sigma_{\text{MLE}}^2 = \frac{1}{\displaystyle\int_{-\infty}^{\infty} \left(\frac{d \log f(x)}{dx} \right)^2 f(x) \, dx}$$

We throw the latter in because it is the best possible estimator. The MLE has no special robustness or nonparametric properties.

A sequence of distributions that go from very heavy to very light tails is the family of Student t distributions, that has the Cauchy distribution at one end (t distribution with one degree of freedom) and the normal distribution at the other (limit of t distributions as the degrees of freedom goes to infinity). For these distributions we have the following ARE for the estimators discussed. The table gives the ARE of each estimator relative to the MLE for the location family generated by the distribution, which is the best estimator. Of course, the MLE is different for each distribution.

d. f.	0%	5%	10%	20%	30%	40%	50%	H-L
1	0.000	0.228	0.419	0.696	0.844	0.876	0.811	0.608
2	0.000	0.676	0.815	0.943	0.967	0.924	0.833	0.867
3	0.500	0.849	0.932	0.987	0.968	0.905	0.811	0.950
4	0.700	0.922	0.973	0.991	0.953	0.882	0.788	0.981
5	0.800	0.956	0.989	0.986	0.938	0.863	0.769	0.993
6	0.857	0.974	0.995	0.978	0.924	0.847	0.753	0.998
7	0.893	0.984	0.996	0.970	0.912	0.834	0.741	0.999
8	0.917	0.990	0.995	0.963	0.902	0.823	0.731	0.999
9	0.933	0.993	0.994	0.957	0.893	0.814	0.723	0.998
10	0.945	0.995	0.992	0.951	0.886	0.806	0.716	0.996
15	0.975	0.996	0.983	0.931	0.861	0.781	0.693	0.988
20	0.986	0.993	0.975	0.919	0.848	0.767	0.680	0.983
25	0.991	0.991	0.970	0.911	0.839	0.759	0.672	0.978
30	0.994	0.989	0.966	0.905	0.833	0.753	0.666	0.975
40	0.996	0.986	0.961	0.898	0.825	0.745	0.659	0.971
50	0.998	0.984	0.958	0.893	0.820	0.740	0.655	0.968
100	0.999	0.980	0.951	0.884	0.809	0.730	0.646	0.962
∞	1.000	0.974	0.943	0.874	0.799	0.720	0.637	0.955

The column labeled 0% is the sample mean (the 0% trimmed mean). Note that it is not the best estimator unless the degrees of freedom are at least 25 (actually it isn't even the best estimator there, the 5% trimmed mean being slightly better though the ARE are the same when rounded to three significant figures). For any of the heavier tailed distributions (the rows of the table above 25 degrees of freedom), some other estimator is better.

Of course, the fact that all of the numbers in the table are less than 1.000 (except in the lower left corner) means none of these estimators are as good as the MLE. But the MLE is a parametric estimator, it cannot be calculated without knowing the likelihood, and the likelihood depends on the parametric model.

The column labeled 50% is the sample median (the 50% trimmed mean). Note that it is never the best estimator (for any t distribution). Some trimmed mean or the Hodges-Lehmann estimator (the column labeled H-L) is always better. In fact, the 30% trimmed mean is better than the 50% trimmed mean for every t distribution.

The Moral of the Story: What estimator is best depends on the true population distribution, which is unknown. So you never know which estimator is best.