Chapter 6: Nonparametric procedures

If you decide that the assumptions of normality are not sufficiently met for doing a *t*-test or ANOVA, then what do you do?

One possibility is to use *nonparametric* procedures. The word *nonparametric* is in contrast to *parametric* procedures, where it is assumed that the data come from a family of distributions (such as the normal) which is *parameterized* by a small number of parameters.

For example, normal distributions are parameterized by the mean μ and variance σ^2. The t family of distributions is parameterized by the degrees of freedom, sometimes denoted ν. (It is common, but not universal, to use Greek letters for the parameters of a distribution.)

Nonparametric procedures

Nonparametric procedures make fewer assumptions about the distribution of the data than do parameteric procedures. The t-tests and ANOVA procedures are examples of parametric procedures. Probability statements such as p-values, and the width of confidence intervals based on these procedures assume a very specific family of distributions (normal distributions) for the underlying data.

Nonparametric procedures still make assumtions about the data, usually especially that each observation is independently sampled. Nonparametric procedures often make weaker assumptions than parametric procedures.

For example, some (not all) nonparametric procedures assume that the data come from a symmetric distribution, but that distribution is not assumed to be normal. If the distribution does happen to be normal, then the procudure would still be valid.

Nonparametric procedures

If the normality assumption is reasonable for t-tests and ANOVA, and if the equal variances assumption is reasonable for ANOVA, there is no need to use nonparametric procedures. However, if these assumptions seem questionable, then it is reasonable to consider nonparametric alternatives.

If nonparametric procedures are used when the assumptions of t-tests or ANOVA are met, then it is likely that any evidence against the null hypothesis (i.e., the p-value) would be weakened. Another way of saying this is that t-tests and ANOVA tend to be more *powerful* (higher probability of rejecting the null when the null is false) then nonparametric procedures when the assumptions of the procedures are met. We will explore this idea using simulation after introducing some of the methods.

Nonparametric procedures: Sign test

The sign test is a test of the hypothesis that a *median* of a population is equal to a certain value. The sign test a nonparametric alternative to the one-sample t-test.
Let η (pronounced *Ay-duh*) denote the population median. The null hypothesis can be written:

$$H_0 : \eta = \eta_0$$

If the null hypothesis is true, then approximately half of the observations should be above η_0 and half should be below η_0.

The alternative hypothesis can be based on either a two-sided or one-sided test, so we could have

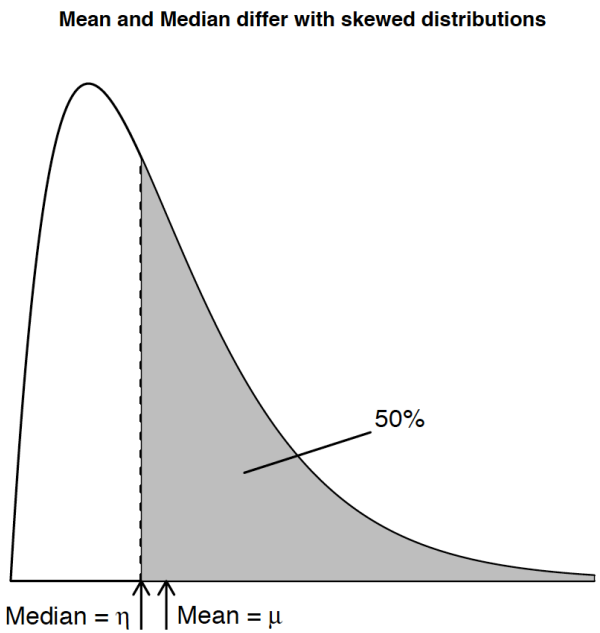
$$H_A : \eta \neq \eta_0$$
$$H_A : \eta < \eta_0, \text{ or}$$
$$H_A : \eta > \eta_0$$

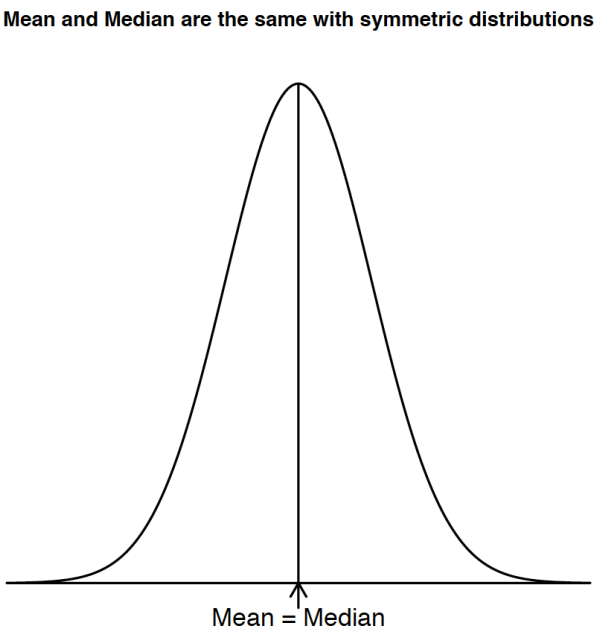
Nonparametric procedures: Sign test

If the distribution is symmetric (such as for the normal), then the population median is equal to the population mean, so the statement that the population median is a certain value is equivalent to the statment that the population mean is that value as well.

However, the test also works for distributions that are skewed, so that the population median is different from the population mean.

Nonparametric procedures: Sign test





Nonparametric procedures: Sign test

For hypothesis testing, the usual procedure is:

- to construct a test statistic based on the data (e.g., t_{obs} or F)
- determine the distribution of the test statistic under the null hypothesis
- quantify how consistent the data are with the null hypothesis (get a p-value)
- make a decision based on the test statistic or p-value

This general approach to hypothesis testing works for *many* different cases, including t-tests, ANOVA and here the sign test.

For the sign test, the test statistic is S, the number of observations larger than η_0, the hypothesized median.

Nonparametric procedures: Sign test

Once you have determined S, you need to find a distribution that S should follow under the null hypothesis. If the null hypothesis is correct, then each observation has a 50% chance to be either above or below η_0.

The procedure is similar to flipping a coin for each observation. With probability 50%, you get heads (the value is above η_0), and with probability 50%, you get tails (the value is below η_0.

The right distribution for describing this is well known in probability and is called the *binomial distirbution*. This distribution describes the probability of getting k successes in n trials, where each trial is independent and has probability p of success. For this application, $p = 1/2$.

Nonparametric procedures: Sign test

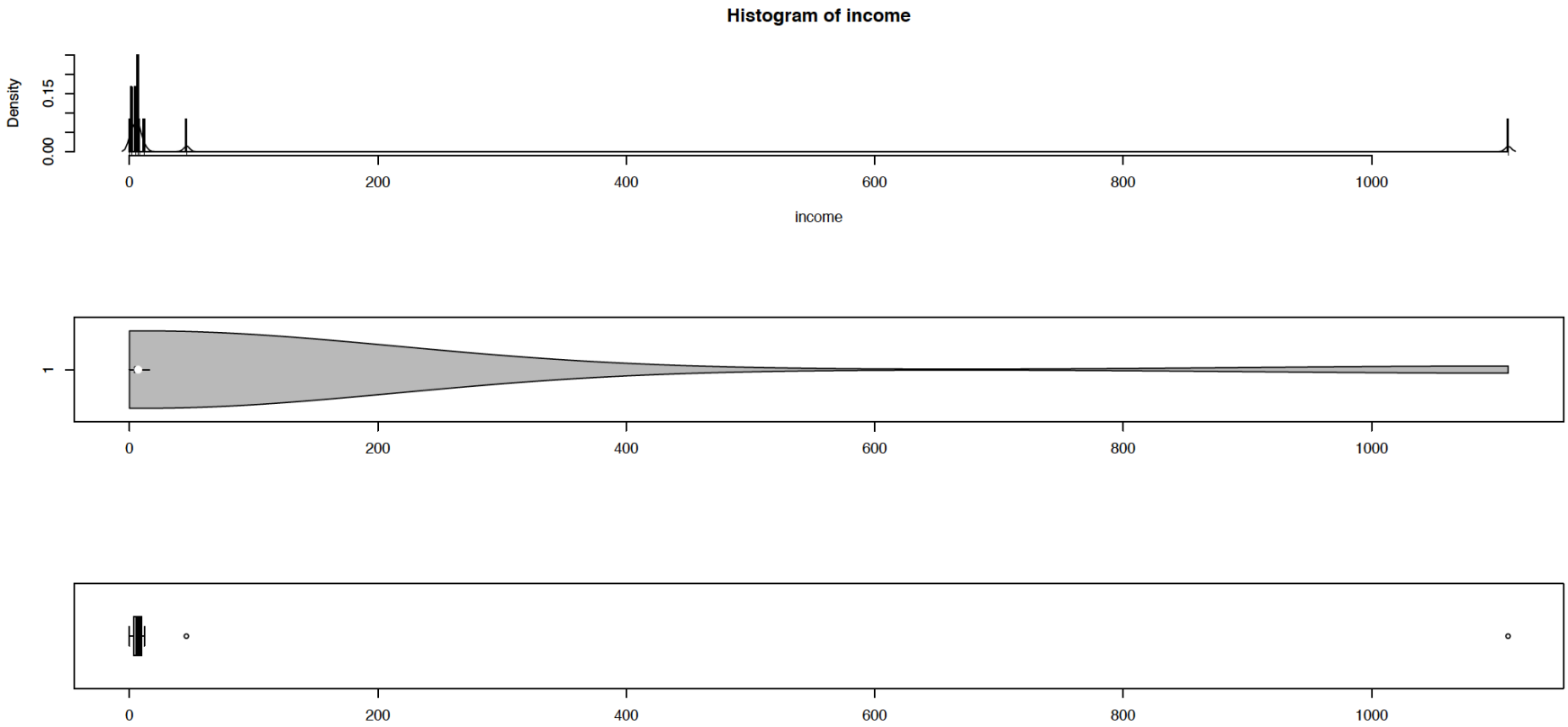
Rather than calculating the binomial probabilities yourself, you can use the function `SIGN.test()` in the `BSDA` package in R. The following is an example with an extreme outlier:

```
#### Example: Income Data
income <- c(7, 1110, 7, 5, 8, 12, 0, 5, 2, 2, 46, 7)
# sort in decreasing order
income <- sort(income, decreasing = TRUE)
income
## [1] 1110 46 12 8 7 7 7 5 5 2 2 0
summary(income)
## Min. 1st Qu. Median Mean 3rd Qu. Max.
## 0.00 4.25 7.00 100.90 9.00 1110.00
sd(income)
## [1] 318.0078
```

Nonparametric procedures: Sign test

```
par(mfrow=c(3,1))
# Histogram overlaid with kernel density curve
hist(income, freq = FALSE, breaks = 1000)
points(density(income), type = "l")
rug(income)
# violin plot
library(vioplot)
vioplot(income, horizontal=TRUE, col="gray")
# boxplot
boxplot(income, horizontal=TRUE)
```

Nonparametric procedures: Sign test



Nonparametric procedures: Sign test

Also try doing `qqnorm()` to see what the QQ-plot looks like (I'll let you do this on your own). Notice the extreme outlier.

A *t*-distribution based CI for this data is unreasonable since it includes negative values

```
income <- c(7, 1110, 7, 5, 8, 12, 0, 5, 2, 2, 46, 7)
t.test(income)

One Sample t-test

data:  income
t = 1.0993, df = 11, p-value = 0.2951
alternative hypothesis: true mean is not equal to 0
95 percent confidence interval:
 -101.1359  302.9692
sample estimates:
mean of x
 100.9167
```

Nonparametric procedures: Sign test

Instead let's try a sign test. The sign test will automatically compute the median for you.

```
library(BSDA)
SIGN.test(income)
s = 11, p-value = 0.0009766
alternative hypothesis: true median is not equal to 0
95 percent confidence interval:
  2.319091 11.574545
sample estimates:
median of x
         7
Achieved and Interpolated Confidence Intervals:
                  Conf.Level L.E.pt  U.E.pt
Lower Achieved CI     0.8540 5.0000  8.0000
Interpolated CI       0.9500 2.3191 11.5745
Upper Achieved CI     0.9614 2.0000 12.0000
```

Nonparametric procedures: Sign test

Note that the confidence interval here are for the population median, not the population mean. This is slightly different than for the t-test.

Also, because the exact distribution of incomes is not known, can only be computed with a 95% confidence level if assumptions are made about the data, which R describes as interpolation. Otherwise, the CI can include values in the data, but the confidence level will not be exactly 95%. It therefore outputs a range of CIs for you to choose from. For example, you are approximately 96% confident that the population median income is between \$2,000 and \$12,000.

Nonparametric procedures: Sign test

It might have occurred to you that since the original data had one extreme outlier, we could have analyzed the data by removing that outlier and then analyzing the remaining data using the usual t-test approach.

The advantage for this approach is that we use the more common t statistics, which can be more powerful and (often) lead to narrower confidence intervals.

For this data, even removing the observation of 1110 leads to a second outlier of 46. Potentially you could remove this outlier as well. But remember that in inferential statistics we are making inferences about a population from which we sampled. If we remove observations that are genuine (not due to typos, incorrectly copied data, etc.), what population are making inferences about? For incomes, we seem to be making inferences about the population of incomes that are not extremely high, rather than the general population of incomes, which includes some genuinely high values.

Nonparametric procedures: Sign test

```
> shapiro.test(income)

Shapiro-Wilk normality test

data:  income
W = 0.35148, p-value = 1.718e-06

> shapiro.test(income[income<100])

data:  income[income < 100]
W = 0.59454, p-value = 2.175e-05

> shapiro.test(income[income<46])


data:  income[income < 46]
W = 0.95189, p-value = 0.6909
```

Nonparametric procedures: Sign test

To illustrate the sensitive of the t based confidence intervals (and p-values) to the outliers compare what happens to the t-tests versus signed rank tests as the extreme observations are made smaller (but still larger than other observations).

```
 income
# [1]    7 1110    7    5    8   12    0    5    2    2   46    7
income2[2] <- 110
income2[11] <- 16
income2
# [1]   7 110   7   5   8  12   0   5   2   2  16   7
income3 <- income2
income3[2] <- 17
income3
 [1]  7 17  7  5  8 12  0  5  2  2 16  7
```

Nonparametric procedures: Sign test

Sensitivity of t-test to outliers:

```
 t.test(income)$conf.int
#[1] -101.1359  302.9692
t.test(income2)$conf.int
#[1] -4.111024 34.277691
 t.test(income3)$conf.int
#[1]  3.945899 10.720768
t.test(income)$p.value
#[1] 0.295115
t.test(income2)$p.value
#[1] 0.1116271
t.test(income3)$p.value
#[1] 0.0005855308
```

Nonparametric procedures: Sign test

Robustness of sign test to outliers:

```
SIGN.test(income)$conf.int
#[1]  2.319091 11.574545
SIGN.test(income2)$conf.int
#[1]  2.319091 11.574545
SIGN.test(income3)$conf.int
#[1]  2.319091 11.574545
```

Nonparametric procedures: Sign test

We could also look at what happens if the outliers are removed. Again, the sign test is less sensitive than the t-test:

```
SIGN.test(income)
#s = 11, p-value = 0.0009766
#95 percent confidence interval: 2.319091 11.574545
SIGN.test(income[income<40])
#s = 9, p-value = 0.003906
#95 percent confidence interval: 2.000000 7.675556
t.test(income)
#t = 1.0993, df = 11, p-value = 0.2951
#95 percent confidence interval: -101.1359  302.9692
t.test(income[income<40])
t = 4.9637, df = 9, p-value = 0.0007766
#95 percent confidence interval: 2.993414 8.006586
```

Nonparametric procedures: Rank-sum test

An alternative to the sign test is the Wilcoxon signed rank test. In this nonparametric procedure, it is assumed that the underlying distirbution is symmetric, but not necessarily normal. It makes stronger assumpetions than the sign test, but not as strong as the t-test.

Here the null is that $\mu = \mu_0$ where μ is equivalently the mean or median. You compute both the signs of $X_i - \mu_0$ and the ranks of $|X_i - \mu_0|$ for each data point. By ranks, we mean that the largest deviation $|X_i - \mu_0|$ gets rank n, where n is the sample size, and the smallest deviation $|X_i - \mu_0|$ gets rank 1.

Nonparametric procedures: Rank-sum test

Example where we test $H_0 : \mu = 10$.

X_i	$X_i - 10$	sign	$\lvert X_i - 10 \rvert$	rank	rank $\times$ sign
20	10	+	10	6	6
18	8	+	8	4.5	4.5
23	13	+	13	8	8
5	−5	−	5	3	−3
14	4	+	4	2	2
8	−2	−	2	1	−1
18	8	+	8	4.5	4.5
22	12	+	12	7	7

Note that for the tied observations, these would have ranks 4 and 5, so we give them each the average of 4 and 5. Gernally, if k observations are tied for rank r, give them each rank $((r+0)+(r+1)+\cdots+(r+k-1))/k = r+(k-1)/2$.

Nonparametric procedures: Rank-sum test

The test statistic is $W =$ the sum of the positive signed ranks. For the above example

$$W = 6 + 4.5 + 8 + 2 + 4.5 + 7 = 32$$

Note that the sum of the unsigned ranks is

$$1 + 2 + \cdots + n = n(n+1)/2$$

where n is the sample size. For this example, $n = 8$, so

$$n(n+1)/2 = (8)(9)/2 = 36$$

If half of the observations are above μ_0, then you expect half of the observations to contribute to the W statistic, and the expected value of W is $(1/2) \times n(n+1)/2 = n(n+1)/4 = 18$ for this example. The question then is whether 32 is significantly different from 18. This depends on the distribution of W.

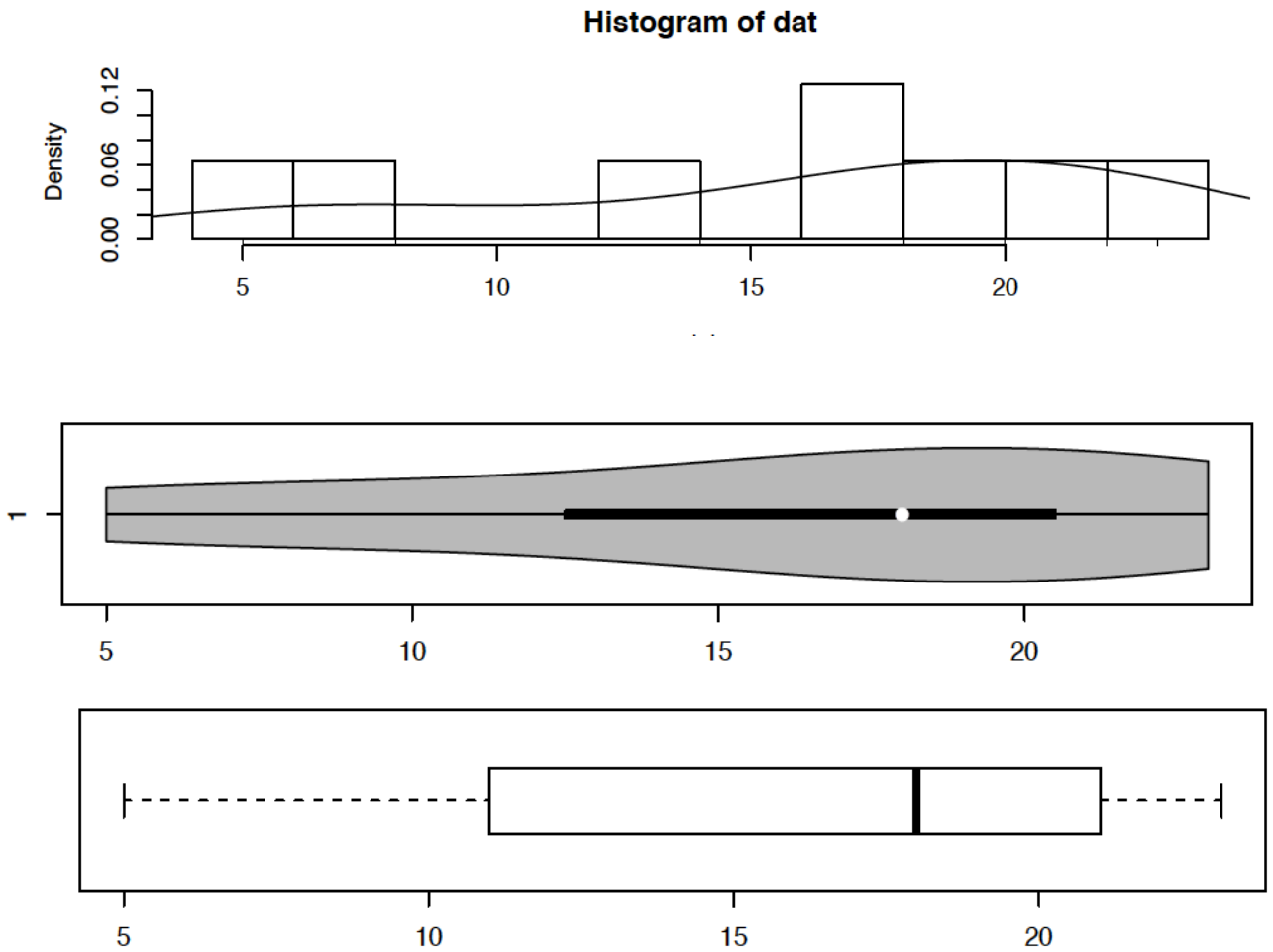
Nonparametric procedures: Rank-sum test

```
#### Example: Made-up Data
dat <- c(20, 18, 23, 5, 14, 8, 18, 22)
# sort in decreasing order
dat <- sort(dat, decreasing = TRUE)
dat
## [1] 23 22 20 18 18 14 8 5
summary(dat)
## Min. 1st Qu. Median Mean 3rd Qu. Max.
## 5.0 12.5 18.0 16.0 20.5 23.0
sd(dat)
## [1] 6.524678
```

Nonparametric procedures: Rank-sum test

```
par(mfrow=c(3,1))
# Histogram overlaid with kernel density curve
hist(dat, freq = FALSE, breaks = 10)
points(density(dat), type = "l")
rug(dat)
# violin plot
library(vioplot)
vioplot(dat, horizontal=TRUE, col="gray")
# boxplot
boxplot(dat, horizontal=TRUE)
```

Nonparametric procedures: Sign test



Nonparametric procedures: Rank-sum test

QQplot and Shapiro-Wilk test do not suggest evidence against normality. There do not appear to be outliers, the distribution is unimodal, and the there does not appear strong skew (is is slightly left-skewed). So I would be comfortable using a t-test for this data. Nevertheless, we illustrate using both the t-test and signed rank test.

Nonparametric procedures: Rank-sum test

```
t.test(dat, mu=10)
##
## One Sample t-test
##
## data: dat
## t = 2.601, df = 7, p-value = 0.03537
## alternative hypothesis: true mean is not equal to 10
## 95 percent confidence interval:
## 10.54523 21.45477
## sample estimates:
## mean of x
## 16
```

Nonparametric procedures: Rank-sum test

```
wilcox.test(dat, mu=10, conf.int=TRUE)
## Warning in wilcox.test.default(dat, mu = 10, conf.int = TRUE):
cannot compute exact p-value with ties
## Wilcoxon signed rank test with continuity correction
##
## V = 32, p-value = 0.0584
## 95 percent confidence interval:
## 9.500002 21.499942
## (pseudo)median
## 16.0056
# without continuity correction
wilcox.test(dat, mu=10, conf.int=TRUE, correct=FALSE)
## V = 32, p-value = 0.04967
## alternative hypothesis: true location is not equal to 10
## 95 percent confidence interval:
## 10.99996 21.00005
## (pseudo)median
## 16.0056
```

Nonparametric procedures: Rank-sum test

Note that the p-value is slightly different depending on whether a continuity correction is used or not. Continuity corrections are often used for discrete tests such as this one and the chi-square test (which we haven't covered), particularly when p-values are based on normal approximations.

When there are ties in the ranks or the sample size is large (50 or above), R uses normal approximations. The idea is that $W/SE(W)$ is approximately normally distributed, so this quantity acts like a z-score. The test is still considered nonparametric even when a normal approximation for the distribution of W is used.

The idea is that W is approximately normally distributed (due to the Central Limit Theorem) even if the underlying data isn't. Of course, you might argue that if W is approximately normally distributed, then so is $\overline{X}$. This is likely true as long as there are no extreme outliers.

Nonparametric procedures: Rank-sum test

You might be tempted to ask, is the correct p-value really below .05 or not? Scientifically, however, p-values of 0.0584 and 0.0497 are quite close. They indicate similar amounts of evidence against the null hypothesis.

This just happens to be a case where a slight difference in method leads to a different conclusion if you are using 0.05 as a cutoff. This is an example where people might argue that paying too much attention to p-values is a bad thing. In the end, I would prefer using the continuity correction since it is the default method and tends to lead to better performance for discrete methods.

Nonparametric procedures: Rank-sum test

If we apply the Wilcoxon test to the income data we would get the same conclusion, although the p-value is somewhat larger.

```
wilcox.test(income)
#
# Wilcoxon signed rank test with continuity correction
#
#data:  income
#V = 66, p-value = 0.003753
#alternative hypothesis: true location is not equal to 0
```

Nonparametric procedures: paired data

Just like for the t-test, you might have two paired samples (e.g., pre- vs post scores) or you might have two independent samples. For paired data, just like with the matched pairs t-test, you can analyze the differences as a single sample rather than think of it as a two-sample problem. This works for both the sign test and the Wilcoxon rank-sum test. For paired data, you will usually be interested in testing $H_0 : \eta = 0$ or $H_0 : \mu = 0$.

Simulating power

As mentioned previously, it is usually preferred to do a t-test over a nonparametric procedure if the assumptions of the t-test are satisfied. Let's do a simulation to see why.

Suppose we have a sample of size $n = 10, 20, 30, 40, 50, \ldots, 100$ from a normal distribution with mean $\mu = 1$ and variance $\sigma^2 = 3^2 = 9$. (i.e., the standard deviation is 2). We want to test $H_0 : \mu = 0$ vs $H_A : \mu \neq 0$. We can test using either a one-sample t-test or the Wilcoxon test (the sign test would also work).

Because we set up the simulation, we know that H_0 is false, so for each simulated example, a correct decision is reached if we reject the null hypothesis. For the smaller sample sizes, though, there is enough variability that we often won't be able to reject the null hypothesis because there is insufficient evidence. Power is the probability of rejecting the null hypothesis. We simulate many times for each value of n to estimate the probability of rejecting the null hypothesis for each sample size.

Simulating power

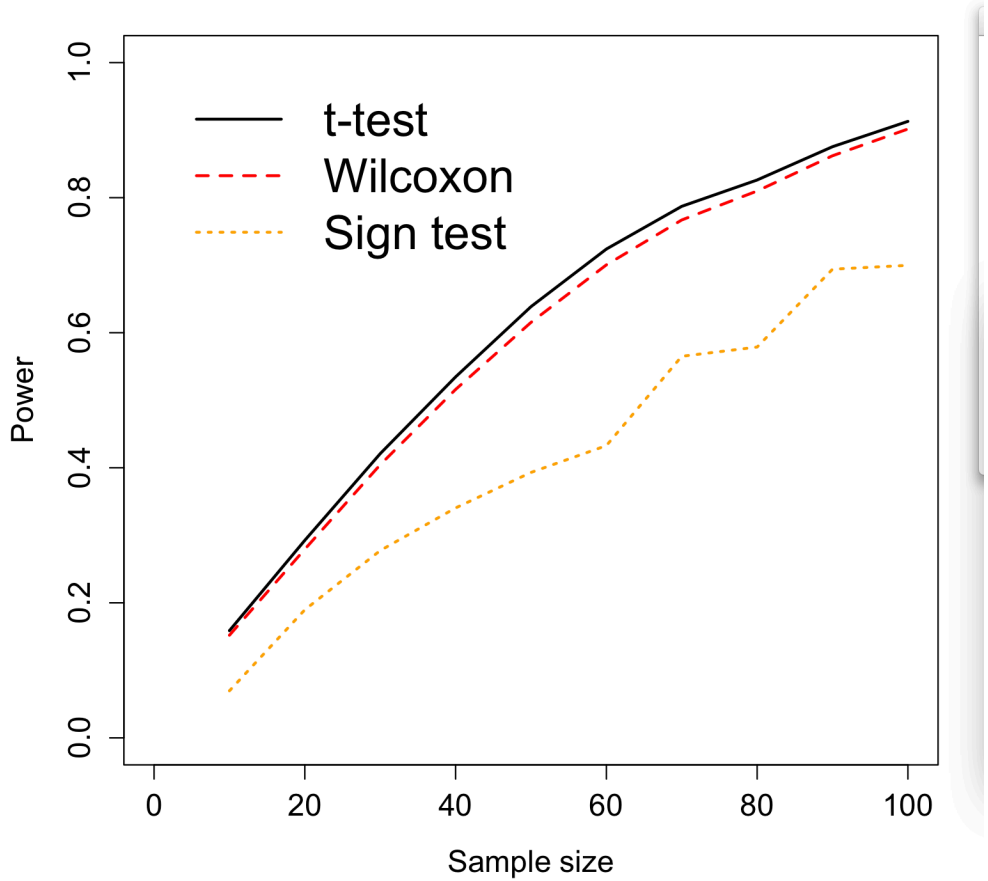
```
#code to simulate power, n=10
I <- 10000 # number of iterations
n <- 100 # sample size
decision.t <- 1:I
decision.w <- 1:I
for(i in 1:I) {
  x <- rnorm(n,1,3)
  pvalue.t <- t.test(x)$p.value
  pvalue.w <- wilcox.test(x)$p.value
  decision.t[i] <- (pvalue.t < .05) # 1=correct decision
  decision.w[i] <- (pvalue.w < .05) #1=correct decision
}
mean(decision.t) # proportion of correct decisions
#[1] 0.1557 # this is the power, about 16%
mean(decision.w)
#[1] 0.1489 # this is the power, 15%
```

Simulating power

```
#code to simulate power, n=20
I <- 10000 # number of iterations
n <- 20 # sample size
decision.t <- 1:I
decision.w <- 1:I
for(i in 1:I) {
  x <- rnorm(n,1,3)
  pvalue.t <- t.test(x)$p.value
  pvalue.w <- wilcox.test(x)$p.value
  decision.t[i] <- (pvalue.t < .05) # 1=correct decision
  decision.w[i] <- (pvalue.w < .05) #1=correct decision
}
mean(decision.t) # proportion of correct decisions
#[1] 0.2981 # this is the power, about 30%
mean(decision.w)
#[1] 0.2826 # this is the power, 28%
```

Simulating power

To plot the power as a function of the sample size, you could run the previous code for $n = 10, 20, \ldots, 100$ This gives the following plot.

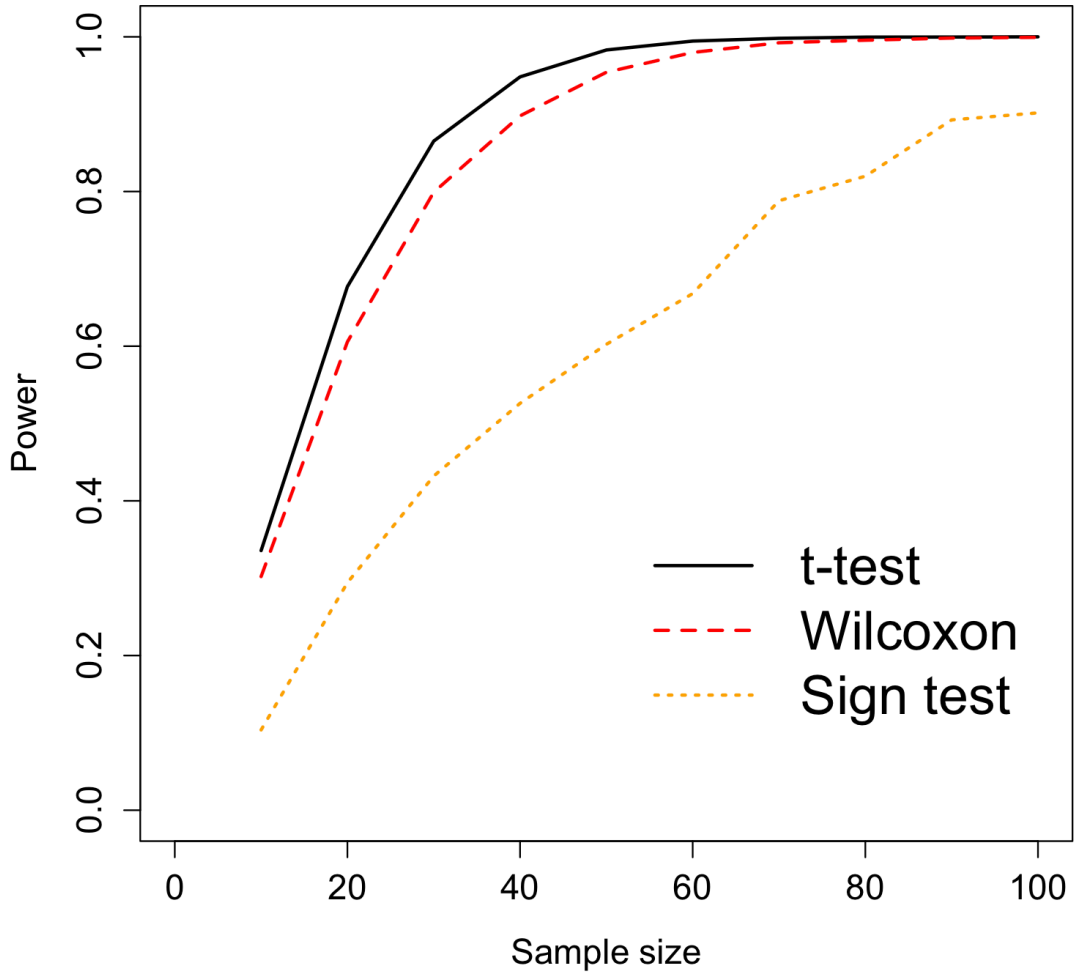


Simulating power

For each sample size, the t-test is slightly more likley to reach the correct conclusion of rejecting the null hypothesis. On the other hand, the differences are pretty small. Still, there is no reason to prefer the Wilcoxon test here.

What happens if the assumptions of the t-test are wrong? Suppose the distribution is uniform? If it is uniform from 0 to 1, then the null hypothesis of $\mu = 0$ is easy to reject from any method. Suppose the distribution is uniform from -1 to 2? Then $\mu = (-1 + 2)/2 = 0.5$, but it is harder to reject the null.

Simulating power: Unif(-1,2), $\mu = .5$, $H_0 : \mu = 0$



Simulating power: Unif(-1,2), $\mu = .5$, $H_0 : \mu = 0$

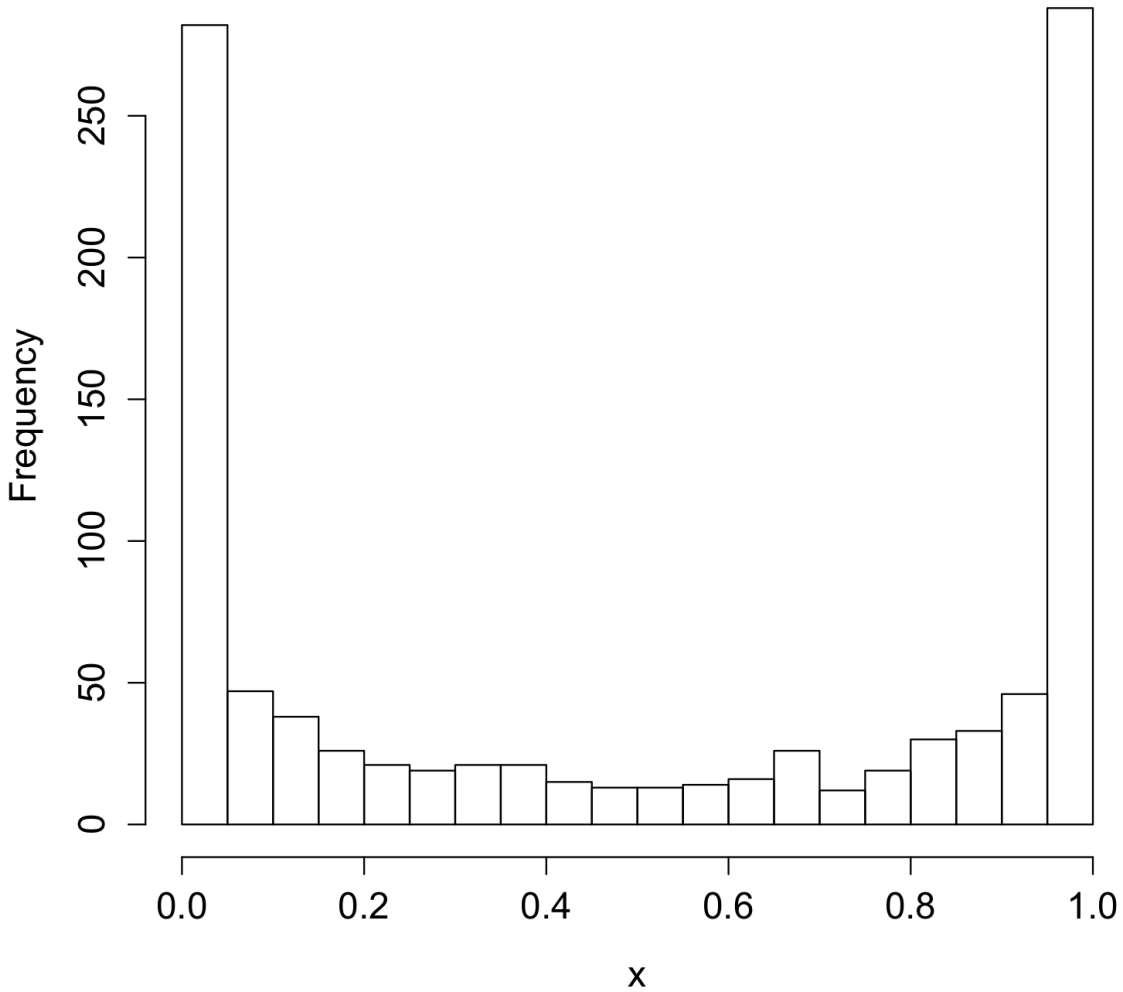
Even though the assumptions of the t-test are false (the data is not normally distributed), the t-test has better power than the Wilcoxon test.

You might note that the uniform distribution tends not to have outliers. We could also try, say a bimodal distribution. Here I'll simulate from a distribution with values between 0 and 1 that is symmetric but tends to have values close to 0 or 1, and is less likely to have values in between.

```
x <- rbeta(1000,.2,.2)
hist(x,nclass=30,xlab="x",cex.lab=1.3,cex.axis=1.3)
```

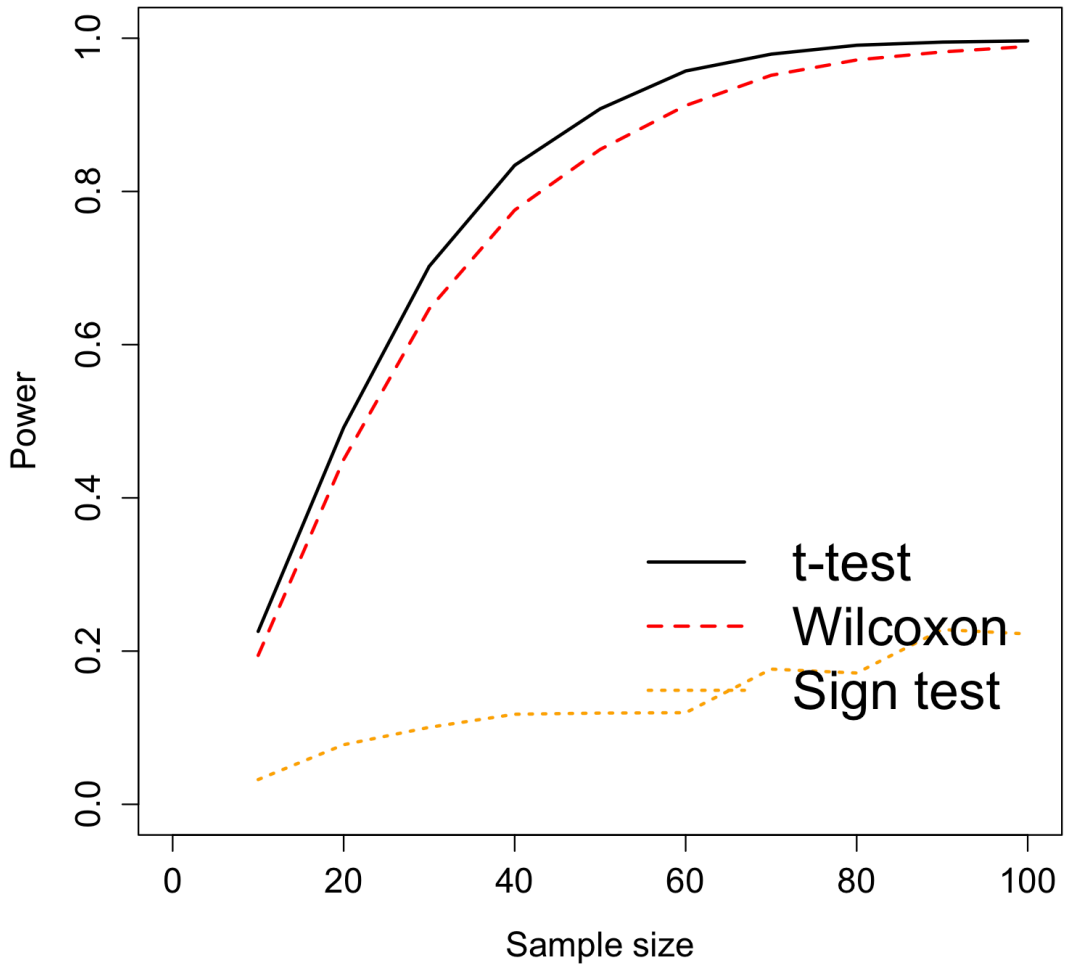
Simulating power: Beta with parameters 0.2, 0.2

```
rbeta(1000,0.2,0.2)
```



Simulating power: Beta with parameters 0.2, 0.2

Two sided test of $H_0 : \mu = 0.3$ or $H_0 : \eta = 0.3$



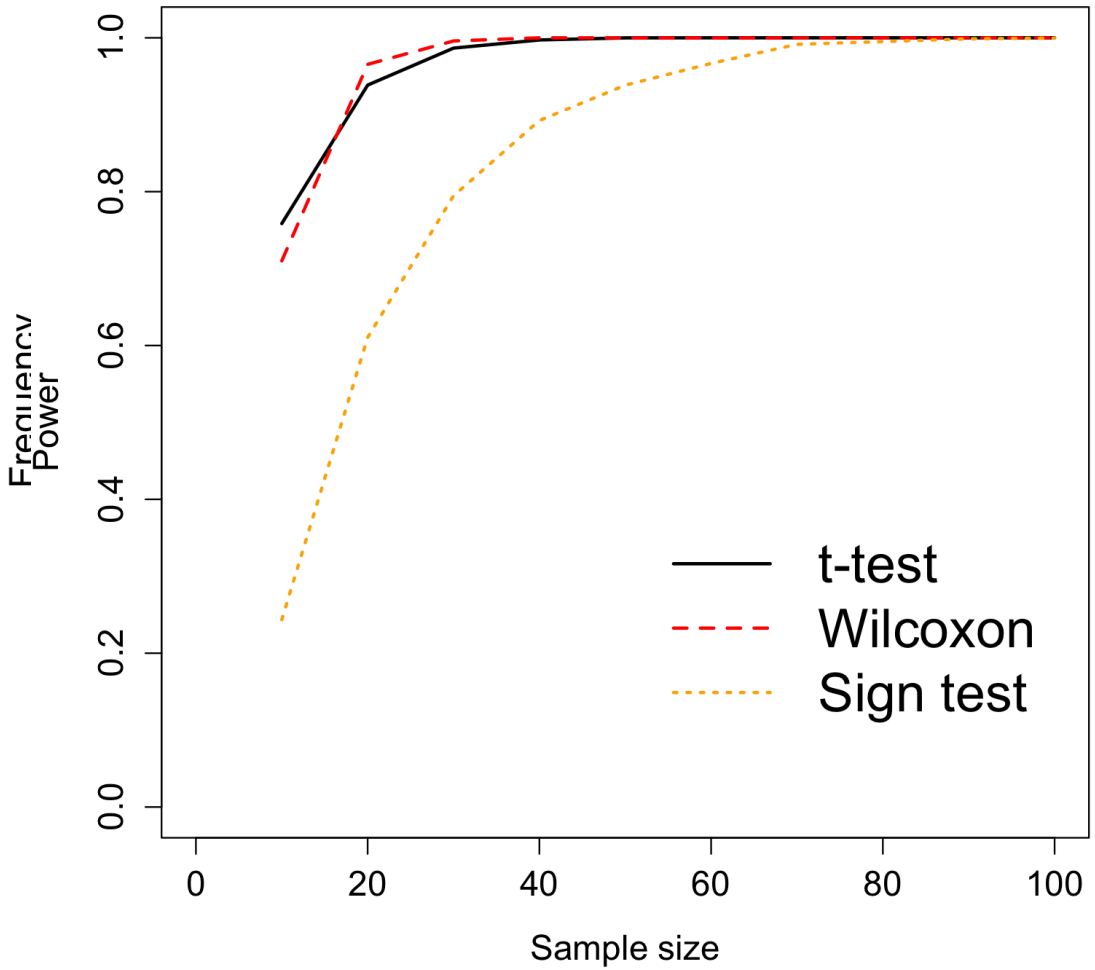
Simulating power

What happens if we simulate from a distribution that violates the assumptions of both the t-test and Wilcoxon test? We'll try with an exponential distribution with mean 1 to test $H_0 : \mu = 2$ versus $H_0 : \mu \neq 2$. For the sign test, I'm using $H_0 : \eta = 1.38$, which is twice the value of the population median of 0.69.

The assumptions of both tests are violated because the distribution isn't symmetric.

Simulating power: Exponential

Two sided test of $H_0 : \mu = 0.3$ or $H_0 : \eta = 0.3$



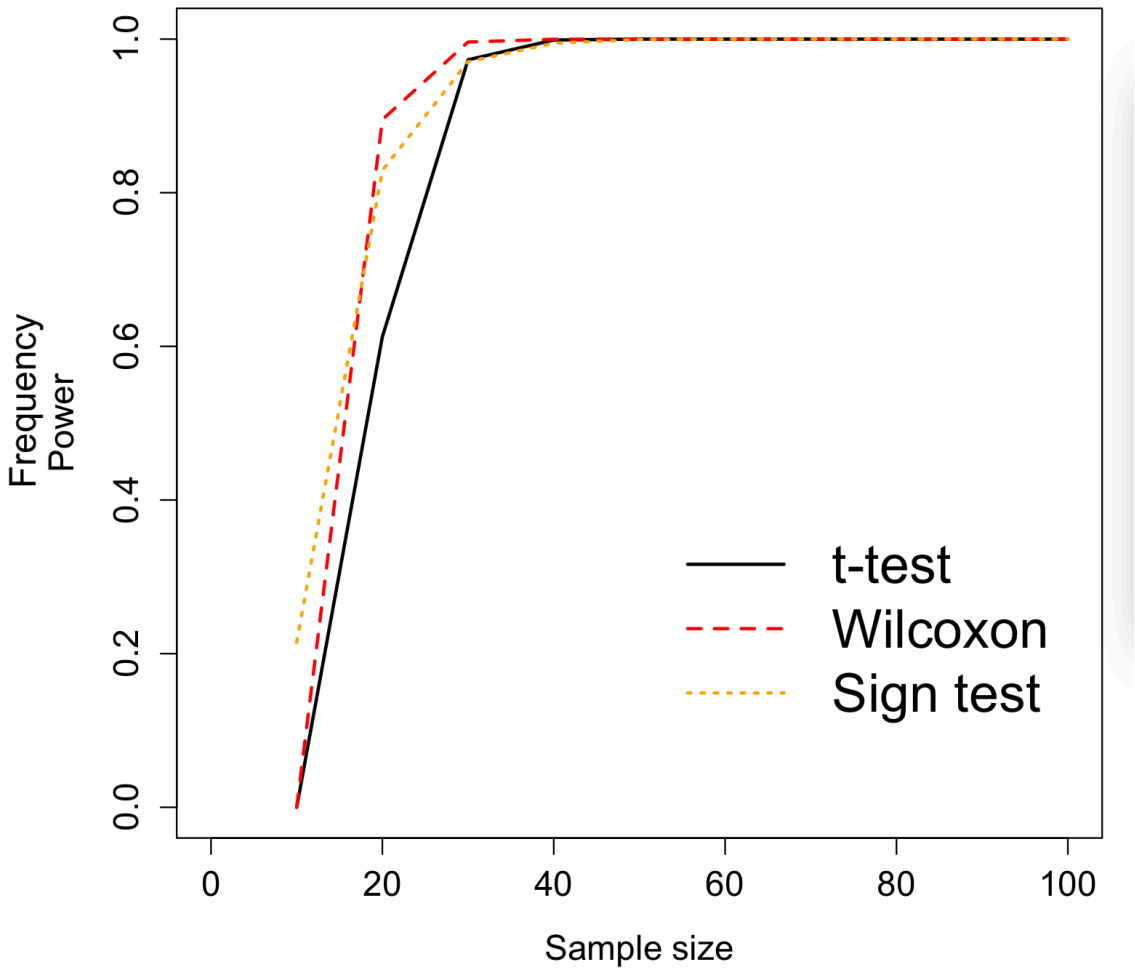
Simulating power: extreme outliers

Here we'll do an example where the distribution is standard normal for all observations except one, and an outlier is added with a value of 5. For a standard normal, this is a fairly extreme outlier. Here we simulate $n-1$ standard normals, and add an additional observation with value 5. We test $H_0 : \mu = 1$ and $H_0 : \eta = 1$. For $n = 10$, this looks like this:

```
x <- c(rnorm(9),5)
```

Simulating power: extreme outliers

The t-test is not as good here....



Type I error

When the assumptions of a test are violated, you might also be concerned with type I error: the probability of falsely rejected the null hypothesis when it is true. Here we'll do the example of the exponential with mean 1, and test $H_0 : \mu = 1$ and $H_0 : \eta = 0.69$. For these cases, the null hypothesis is true but the background assumptions of symmetric (or normal) distributions are incorrect for the Wilcoxon and t-tests.

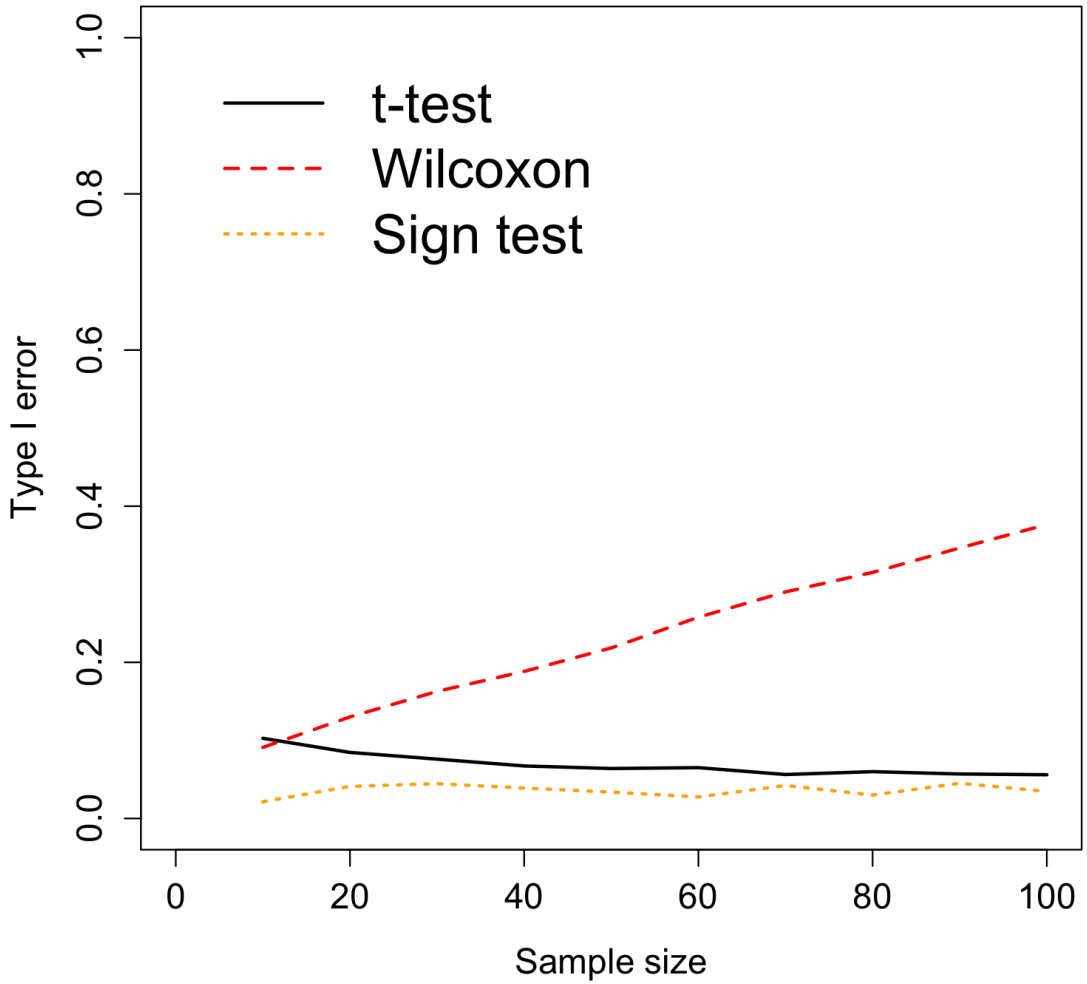
From the plot on the next slide, we see the Wilcoxon test is doing something bad. As the sample size increases, it's type I error rate is increasing. For large samples, it is increasingly likely to reject the null hypothesis even when the null hypothesis is true! This is because the p-value is based on assuming both the hypothesized value of $\mu = 1$ and the symmetry of the distribution. The low p-values (causing the incorrect decisions) are due to the lack of symmetry, not due to the null hypothesis being wrong.

Type I error

Both the t-test and the Willcoxon test are based on assuming symmetry in the population distribution. However, the t-test is not very sensitive to this assumption (as long as there are no extreme outliers). The Wilcoxon test is not very sensitive to extreme outliers if the rest of the distribution is symmetric, but it is senstive to the distribution being symmetric.

Simulating Type I error

Something bad with Wilcoxon—error increasing with sample size.



Type I error

The type I error for both Wilcoxon and the t-test is about 0.10 here when testing at level α. That means that if the null hypothesis is true, both tests are rejecting twice as often as they are supposed to due to the violations of the assumptions. When using either test on skewed data, you should keep this in mind. A *p*-value of say 0.04, might be too low compared to what it should be if the assumptions of the test had been met. As the sample size increases, the Central Limit Theorem means that the sampling distribution of the mean is closer to normal, and the α level for the *t*-test is getting closer to 0.05 (the proportion of false rejections was 0.056 for $n = 100$). Unfortunately, large sample sizes are not helping the Wilcoxon test.

Newcombe's Data

Experiments of historical importance were performed beginning in the eighteenth century to determine physical constants, such as the mean density of the earth, the distance from the earth to the sun, and the velocity of light. An interesting series of experiments to determine the velocity of light was begun in 1875. The first method used, and reused with refinements several times thereafter, was the rotating mirror method3.

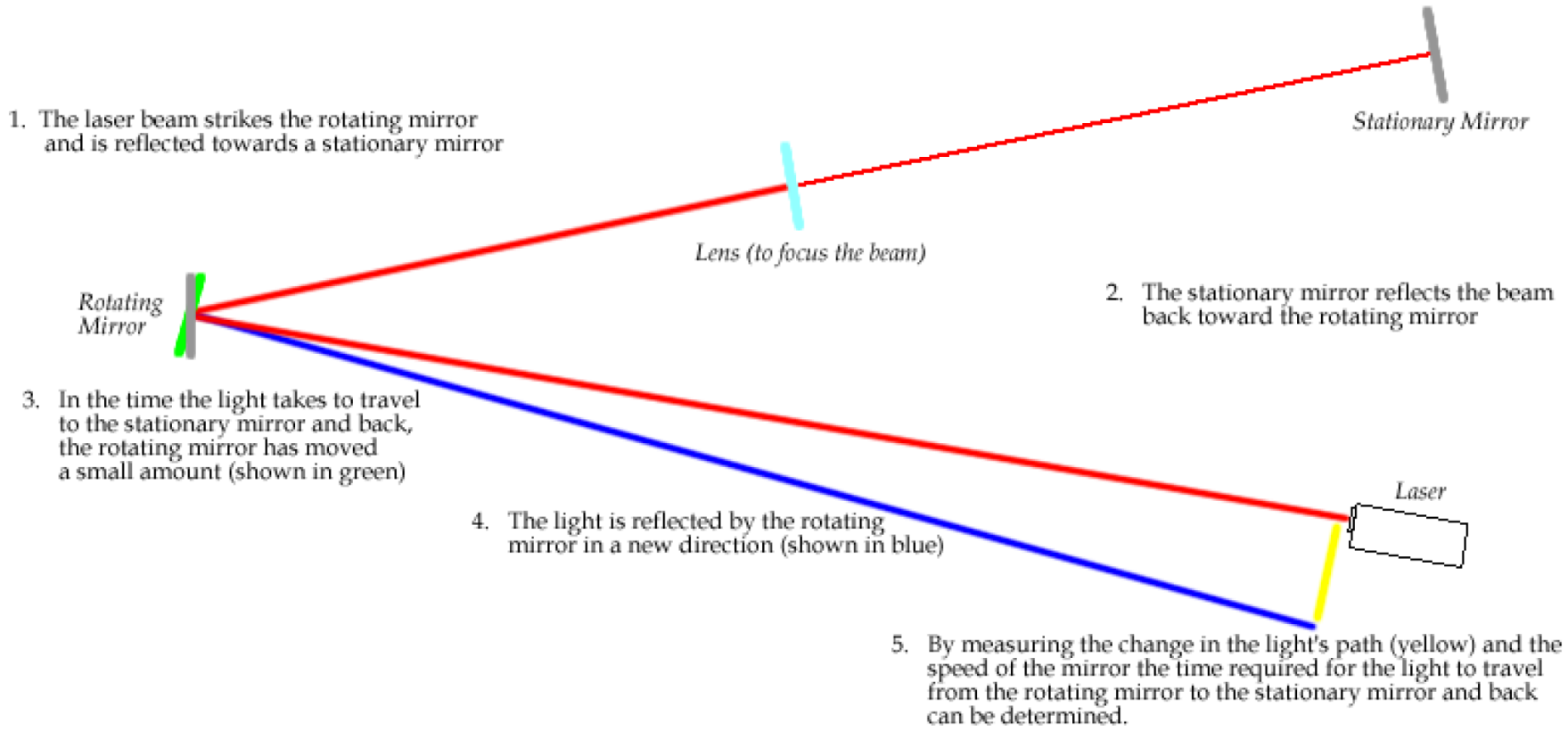
In this method a beam of light is reflected on a rapidly rotating mirror to a fixed mirror at a carefully measured distance from the source. The returning light is re-reflected from the rotating mirror at a different angle, because the mirror has turned slightly during the passage of the corresponding light pulses. From the speed of rotation of the mirror and from careful measurements of the angular difference between the outward-bound and returning light beams, the passage time of light can be calculated for the given distance.

Newcombe's Data

After averaging several calculations and applying various corrections, the experimenter can combine mean passage time and distance for a determination of the velocity of light. Simon Newcombe, a distinguished American scientist, used this method during the year 1882 to generate the passage time measurements given below, in microseconds. The travel path for this experiment was 3721 meters in length, extending from Ft. Meyer, on the west bank of the Potomac River in Washington, D.C., to a fixed mirror at the base of the Washington Monument.

The problem is to determine a 95% CI for the "true" passage time, which is taken to be the typical time (mean or median) of the population of measurements that were or could have been taken by this experiment.

Newcombe's Data

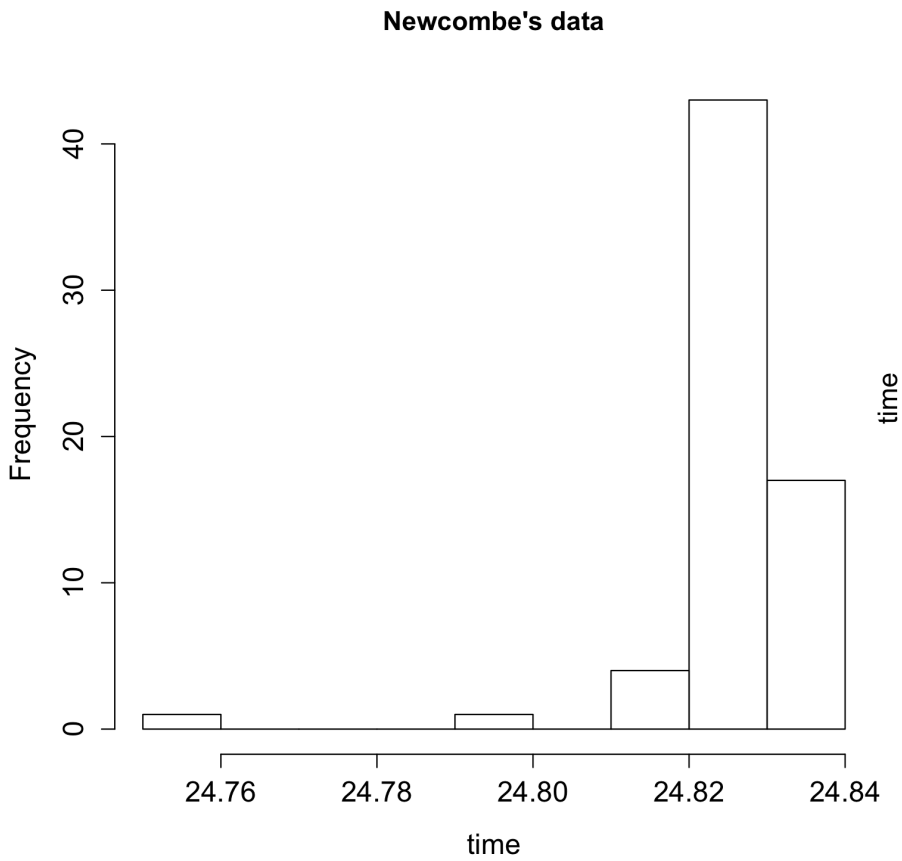


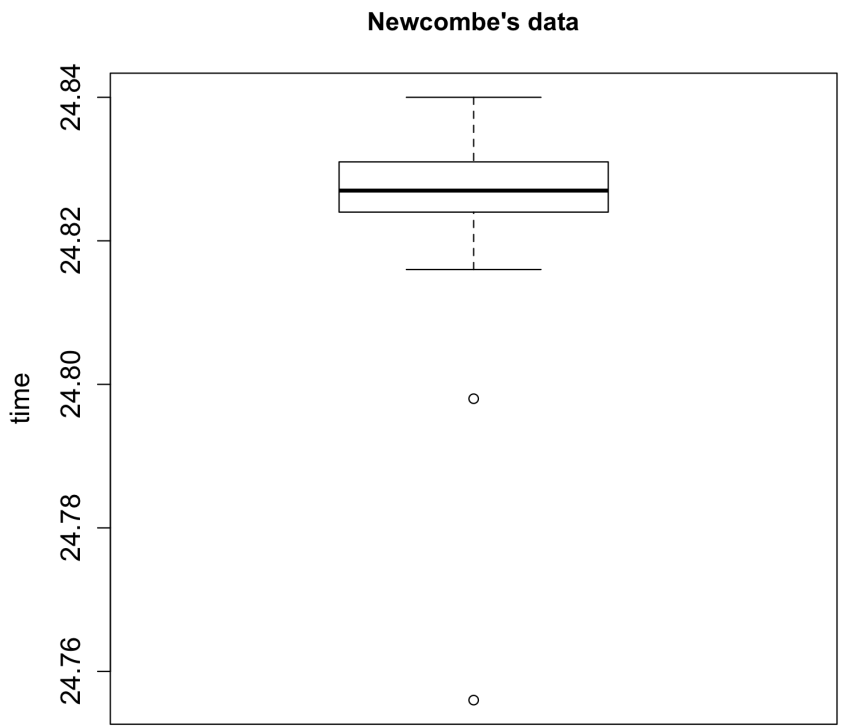
Newcombe's Data

```
#### Example: Newcombe's Data
time <- c(24.828, 24.833, 24.834, 24.826, 24.824, 24.756
, 24.827, 24.840, 24.829, 24.816, 24.798, 24.822
, 24.824, 24.825, 24.823, 24.821, 24.830, 24.829
, 24.831, 24.824, 24.836, 24.819, 24.820, 24.832
, 24.836, 24.825, 24.828, 24.828, 24.821, 24.829
, 24.837, 24.828, 24.830, 24.825, 24.826, 24.832
, 24.836, 24.830, 24.836, 24.826, 24.822, 24.823
, 24.827, 24.828, 24.831, 24.827, 24.827, 24.827
, 24.826, 24.826, 24.832, 24.833, 24.832, 24.824
, 24.839, 24.824, 24.832, 24.828, 24.825, 24.825
, 24.829, 24.828, 24.816, 24.827, 24.829, 24.823)
```

Newcombe's Data

Plotting the data shows that it is left-skewed with two outliers.





Newcombe's Data

Here it might be interesting to compare the width of the confidence intervals for different methods. The Wilcoxon confidence interval is half the width of the t-based confidence interval.

```
t.test(time)$conf
#[1] 24.82357 24.82885
#attr(,"conf.level")
#[1] 0.95
t.test(time)$conf[2] - t.test(time)$conf[1]
#[1] 0.005283061
wilcox.test(time,conf.int=T)
# 24.82604 24.82853
wilcox.test(time,conf.int=T)$conf[2]-
wilcox.test(time,conf.int=T)$conf[1]
#[1] 0.002487969
```

Nonparametric tests for two independent samples

A nonparametric alternative to the two-sample t-test is the Mann-Whitney or Wilcoxon-Mann-Whitney (WMW) test.

The test assumes that two distributions have the same shapes and spreads (e.g., they should have the same standard deviations), but they are not assumed to be symmetric. The null hypothesis is often stated as that the population medians are equal, $H_0 : \eta_1 = \eta_2$. If the distributions are symmetric, then it is testing that the means are equal as well, $H_0 : \mu_1 = \mu_2$. More generally, it can be thought of as testing whether the two populations have the same distribution. The two samples are pooled and then ranked, where the ranking is similar to the Wilcoxon one-sample test, so that a rank of 1 is used for the smalles observation, and a rank of $n_1 + n_2$ is used for the largest observation.

WMW test

The test statistic will be computed in R, but it is based on the sum of the ranks in one of the samples adjusted by the sample size. The idea is that if the two samples come from the same distirbution and had the same sample size, then their sum of ranks should be about the same. This has to be adjusted for potentially having different sample sizes.

The WMW can also be used for ordinal but non-numeric data, for example, data where there are ordered categories but not measurement scale data. Ordinal data includes Likert scale data where people indicate that they strongly disagree, disagree, are neutral, agree, or strongly disagree.

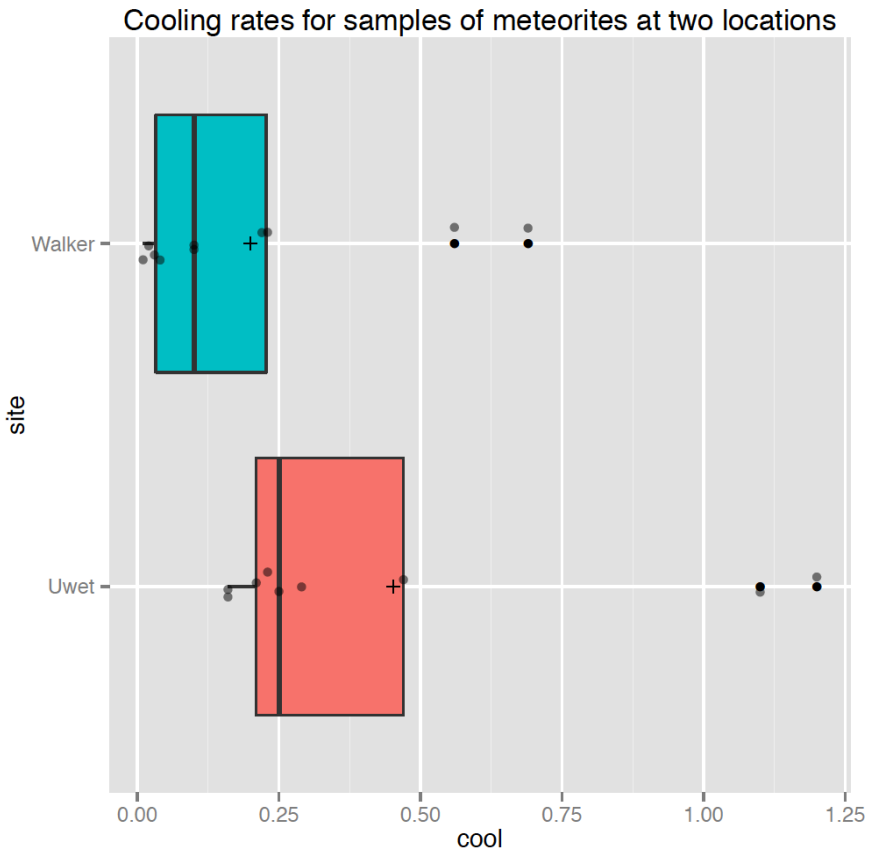
WMW test: example

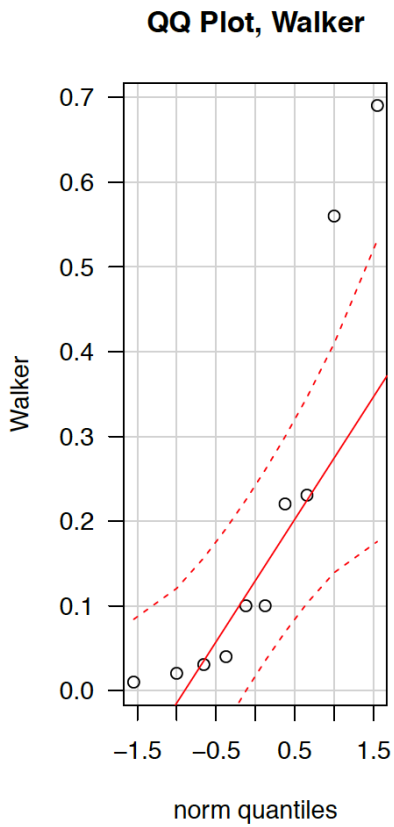
Here an example is from comparing cooling rate (degrees per million years) for meteorite fragments from two locations: Uwet (Cross River, Nigeria), and Walker County in Alabama (US).

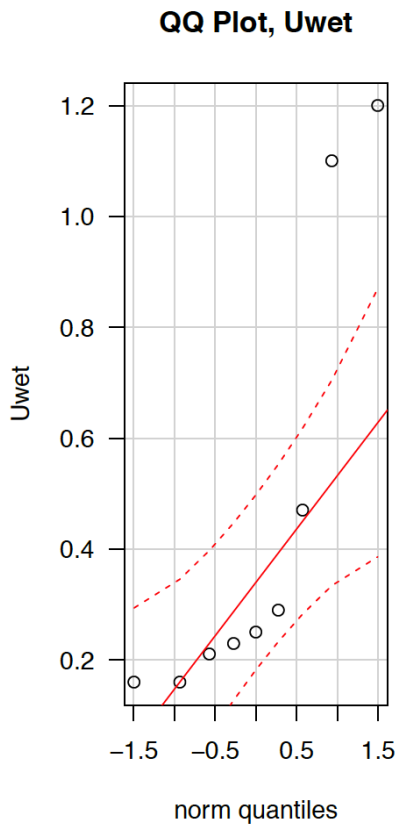
```
Uwet <- c(.21, .25, .16, .23, .47, 1.20, .29, 1.10, .16)
Walker <- c(.69, .23, .10, .03, .56, .10, .01, .02, .04, .22)
```

WMW test: example

Something bad with Wilcoxon—error increasing with sample size.







WMW test: example

This example has a small sample size and a rank-based nonparametric test, so we might expect the WMW test to do well compared to a t-test. Unlike a simulation though, we don't really know whether the null is true.

```
wilcox.test(Uwet,Walker,conf.int=T))
#95 percent confidence interval:
# 0.0000449737 0.4499654518

#W = 69.5, p-value = 0.04974

t.test(Uwet,Walker)

#t = 1.6242, df = 12.652, p-value = 0.129
#95 percent confidence interval:
# -0.08420858  0.58865302
```

WMW test: example

One can get very similar results from the WMW by doing a two-sample t-test on the ranks. I.e., replace the data measurement values by their ranks and then perform the t-test using the equal variance assumption. This could be done as follows:

```
dat <- c(Uwet,Walker)
group <- c(rep("a",length(Uwet)),rep("b",length(Walker)))
a <- rank(dat)
a
[1]  9.0 13.0  7.5 11.5 15.0 19.0 14.0 18.0  7.5 17.0 11.5  5.5  3
[16]  1.0  2.0  4.0 10.0
t.test(a ~ group,var.equal=T)
t = 2.2082, df = 17, p-value = 0.04125
95 percent confidence interval:
  0.2304938 10.1139507
```

Nonparametric alternative to ANOVA: Kruskal-Wallis

A nonparametric alternative to ANOVA is the Kruskal-Wallis test (KW). This generalizes the WMW test much the same way that ANOVA generlizes the two-sample equal-variance t-test.

Kruskal-Wallis tests the hypothesis that all populations have the same median assuming that they have equal spreads. The alternative hypothesis is that at least two population medians are different. The null can be written as

$$H_0 : \eta_1 = \eta_2 = \cdots = \eta_k$$

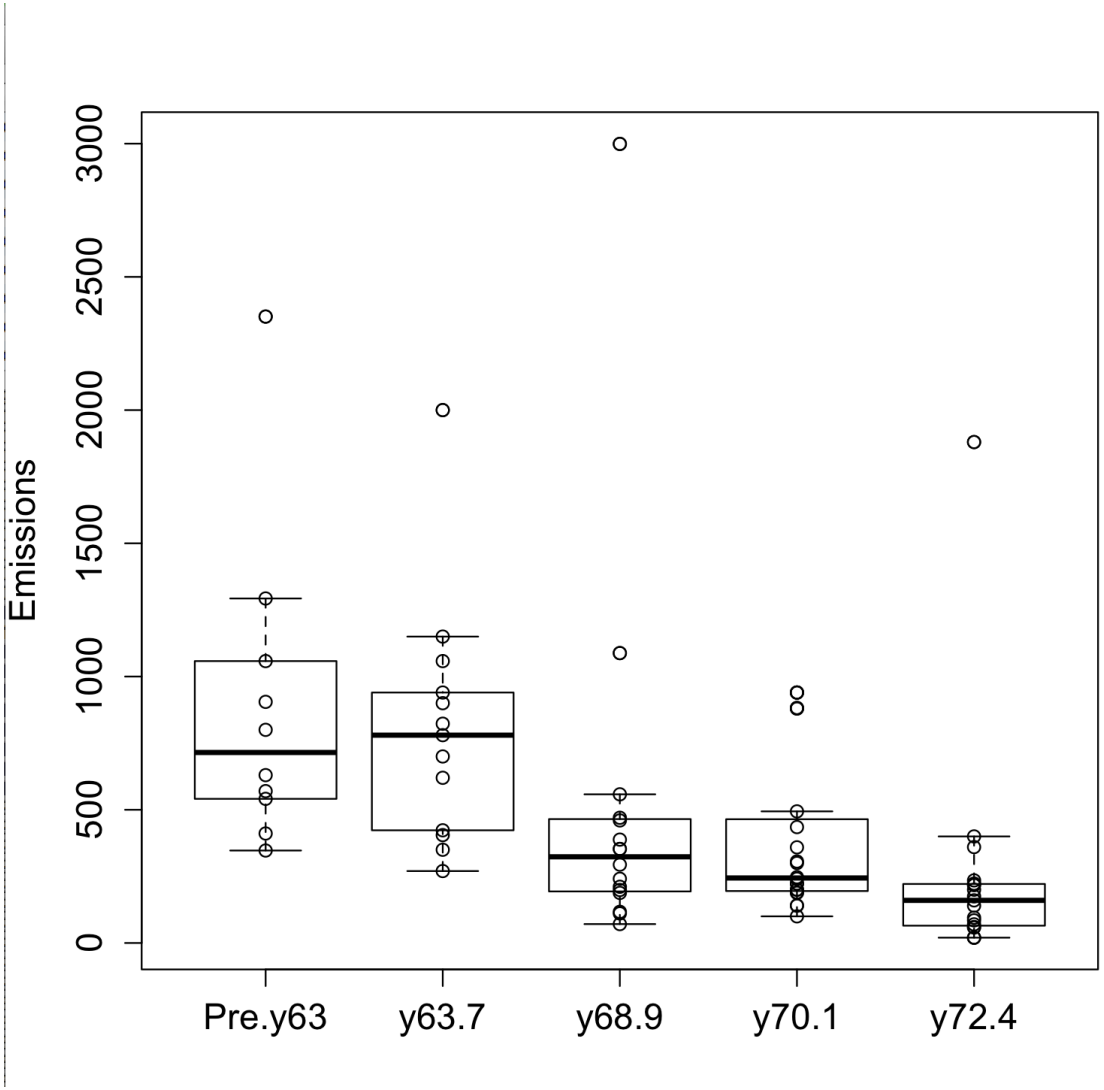
when there are k groups.

The idea for the KW test is similar to the WMW: pool all the data and rank them ignoring group membership, averaging the ranks in the case of ties. Then each group should have similar distributions of ranks if the null is true.

```
#### Example: Hydrocarbon (HC) Emissions Data (from 45 yrs ago!)
emis <- read.table(text="
Pre-y63 y63-7 y68-9 y70-1 y72-4
2351 620 1088 141 140
1293 940 388 359 160
541 350 111 247 20
1058 700 558 940 20
411 1150 294 882 223
570 2000 211 494 60
800 823 460 306 20
630 1058 470 200 95
905 423 353 100 360
347 900 71 300 70
NA 405 241 223 220
NA 780 2999 190 400
NA 270 199 140 217
NA NA 188 880 58
NA NA 353 200 235
NA NA 117 223 1880
NA NA NA 188 200
NA NA NA 435 175
NA NA NA 940 85
NA NA NA 241 NA
", header=TRUE)
```

```
library(reshape2)
# convert to long format
emis.long <- melt(emis,
variable.name = "year",
value.name = "hc",
na.rm = TRUE
)
# No id variables; using all as measure variables
attach(emis.long)
boxplot(hc ~ year,cex.axis=1.3,cex.lab=1.3,ylab="Emissions")
points(hc ~ yr)
```

KW test: example



KW test: example

```
> by(emis.long$hc,emis.long$year,median)
emis.long$year: Pre.y63
[1] 715
emis.long$year: y63.7
[1] 780
emis.long$year: y68.9
[1] 323.5
emis.long$year: y70.1
[1] 244
emis.long$year: y72.4
[1] 160
```

KW test: example

```
> by(emis.long$hc,emis.long$year,sd)
emis.long$year: Pre.y63
[1] 591.5673
emis.long$year: y63.7
[1] 454.9285
emis.long$year: y68.9
[1] 707.8026
emis.long$year: y70.1
[1] 287.8864
emis.long$year: y72.4
[1] 410.7866
```

KW test: example

The different groups seem to have different spreads, which violates the assumptions of both ANOVA and the KW test. We'll see how both methods perform with this data. The techinical name for different spreads is "heteroscedasticity".
For some discussion on the KW test and why it is used and sometimes misused when assumptions of ANOVA are violated, see the webpage

`http://influentialpoints.com/Training/Multiple_comparison_tests_after_ANOVA_use_and_misuse.htm`

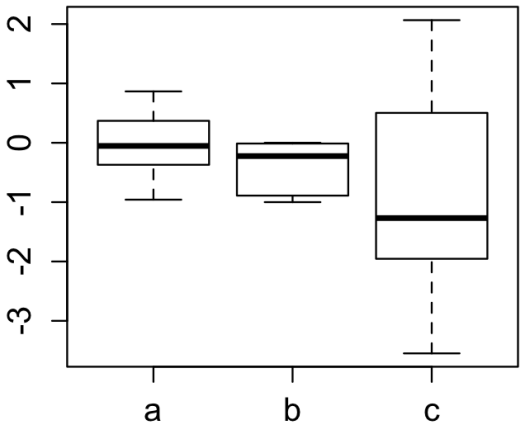
KW test: example

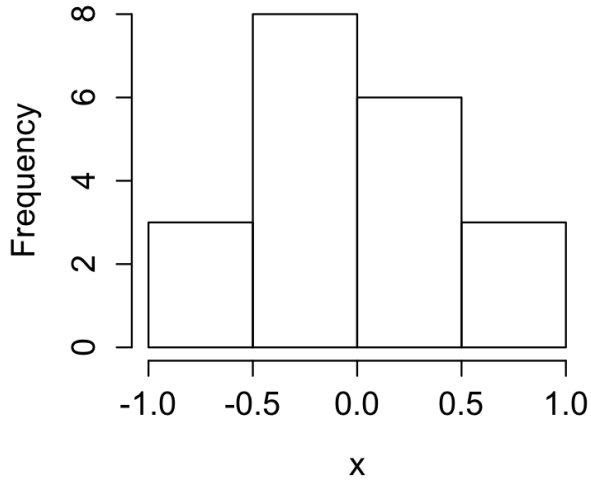
```
 kruskal.test(emis.long$hc ~ factor(emis.long$year))
#
# Kruskal-Wallis rank sum test
#
#data:  emis.long$hc by factor(emis.long$year)
#Kruskal-Wallis chi-squared = 31.808, df = 4, p-value = 2.093e-06
a <- aov(emis.long$hc ~ factor(emis.long$year))
summary(a)
   #                        Df    Sum Sq Mean Sq F value  Pr(>F)
#factor(emis.long$year)   4   4226834 1056709   4.343 0.00331 **
#Residuals                73 17759968  243287
```

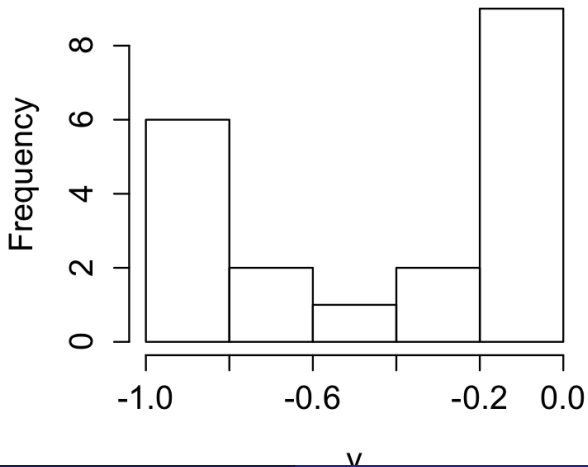
Type I error: different populations and heteroscedasticity

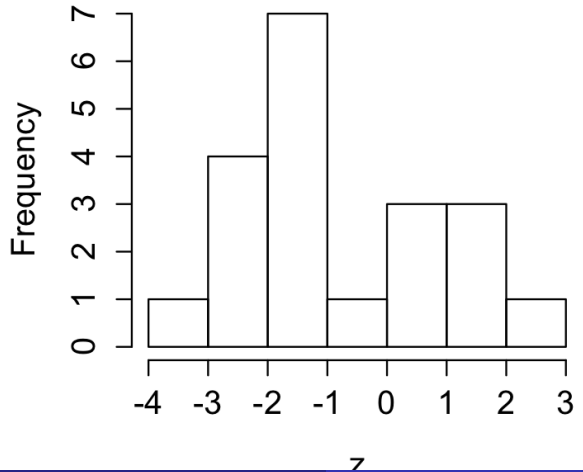
```
I <- 10000
myp.1 <- 1:I
myp.2 <- 1:I
for(i in 1:I) {
  x <- runif(20,-1,1)
  y <- rbeta(20,.2,.2)-0.5
  z <- rnorm(20,0,2)
  dat <- c(x,y,z)
  group <- c(rep("a",20),rep("b",20),rep("c",20))
  a <- aov(dat ~ factor(group))
  #this next line was tricky
  myp.1[i] <- summary(a)[[1]][["Pr(>F)"]][1]
  a <- kruskal.test(dat ~ factor(group))
  myp.2[i] <- a$p.value
  myp.2
}
print(c(mean(myp.1<.05),mean(myp.2<.05)))
#[1] 0.0769 0.0755 # similar type I errors, ideally <= 0.05
```

KW test with different populations









Analyzing the log of the data

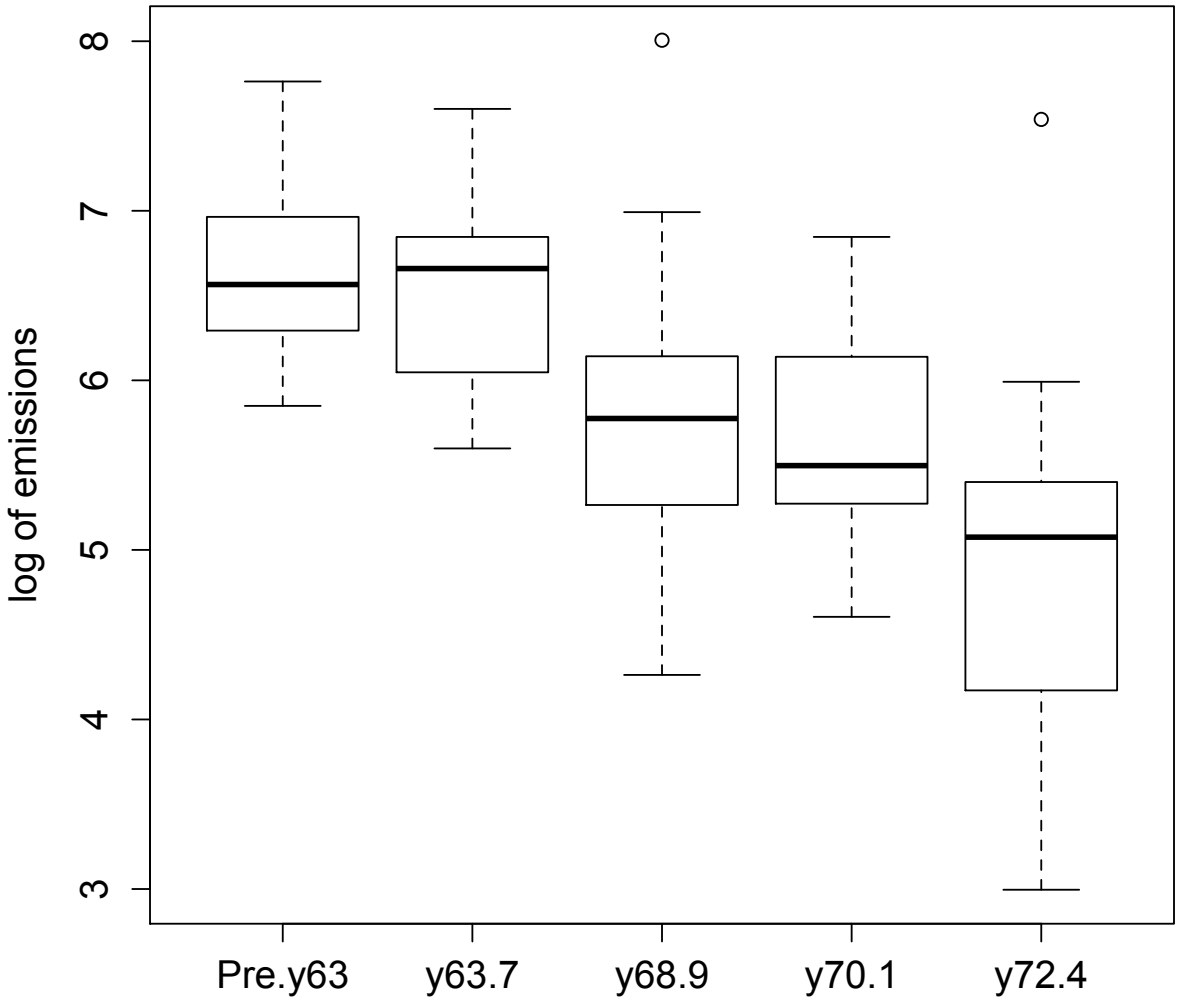
For highly skewed data, another possibility is to analyze the data on a transformed scale, such as the log scale. This often reduces the number or severity of the outliers and can make the distributions more nearly symmetric. Note that Bartlett's test suggests considerably closer variances under the log transform than under the original scale, although the p-value is still under 0.05.

```
boxplot(log(emis.long$hc) ~ emis.long$year)
bartlett.test(log(emis.long$hc),emis.long$year)
#
#data:  log(emis.long$hc) and emis.long$year
#Bartlett's K-squared = 10.879, df = 4, p-value = 0.02795

bartlett.test(emis.long$hc,emis.long$year)

#data:  emis.long$hc and emis.long$year
#Bartlett's K-squared = 14.451, df = 4, p-value = 0.005986
```

Analyzing the log of the data



Analyzing the log of the data

Note that taking the log of the pooled data doesn't change the ranking of the data, so that performing a KW test on the log of the data gives identical results to using the original data. This suggests that the KW test might have some robustness to heteroscedasticity since the log-transformed data have more reasonably close spread.

On the other hand, for the example using three different distributions (uniform, beta, and normal), the t-test and KW test had similarly inflated type I error rates.

We'll encounter other approaches to analyzing transformed data later on.

Multiple comparisons

With the ANOVA, if the null hyptohesis is rejected, you often follow up with pairwise comparisons, adjusting for multiple comparisons using Fisher's Least Significant Differences (LSD or FSD), Bonferroni, or Tukey's Honest Differences.

For Kruskal-Wallis, the same approach can be used, usually using the WMW tests for the follow up pairwise comparisons.

Multiple comparisons

We'll illustrate with an example from treating Hodgkin's disease:

Hodgkin's Disease Study Plasma. bradykininogen levels were measured in normal subjects, in patients with active Hodgkin's disease, and in patients with inactive Hodgkin's disease. The globulin bradykininogen is the precursor substance for bradykinin, which is thought to be a chemical mediator of inflammation. The data (in micrograms of bradykininogen per milliliter of plasma) are displayed below. The three samples are denoted by **nc** for normal controls, **ahd** for active Hodgkin's disease patients, and **ihd** for inactive Hodgkin's disease patients. The medical investigators wanted to know if the three samples differed in their bradykininogen levels.

Multiple comparisons

```
#### Example: Hodgkins Disease Study
nc <- c(5.37,5.8,4.7,5.7,3.4,8.6,7.48,5.77,7.15,6.49,4.09,5.94,
6.38,9.24,5.68,4.53,6.51,7.0,6.2,7.04,4.82,6.73,5.26)
ahd <- c(3.96,3.04,5.28,3.4,4.1,3.61,6.16,3.22,7.48,3.87,4.27
,4.05,2.40,5.81,4.29,2.77,4.40)
ihd <- c(5,37,10.6,5.02,14.3,9.9,4.27,5.75,5.03,5.74,7.85,6.82,
7.9,8.36,5.72,6.0,4.75,5.83,7.3,7.52,5.32,6.05,5.68,7.57,5.68,
8.91,5.39,4.40,7.13)
hd <- c(nc,ahd,ihd)
group <- c(rep(``nc'',length(nc)),rep(``ahd'',length(ahd)),
rep(``ihd'',length(ihd)))
hd.long <- as.data.frame(cbind(hd,group))
#the next line is tricky, to convert a factor to numeric
hd.long$hd <- as.numeric(levels(hd.long$hd))[hd.long$hd]
```

Note that only using `cbind()` creates a matrix object. Using `as.data.frame()` converts the matrix to a data frame. The `as.numeric()` command converts from string to numeric for the numbers.

Multiple comparisons

```
by(hd.long$hd, hd.long$group, summary)
#hd.long$group: ahd
#    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
#    3.00    7.00   12.00   16.76   17.00   54.00
#------------------------------------------------------------
#hd.long$group: ihd
#   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
# 1.00   23.00   33.00   35.52   53.00   63.00
#------------------------------------------------------------
#hd.long$group: nc
#    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
#    7.00   26.50   39.00   37.13   48.00   62.00
```

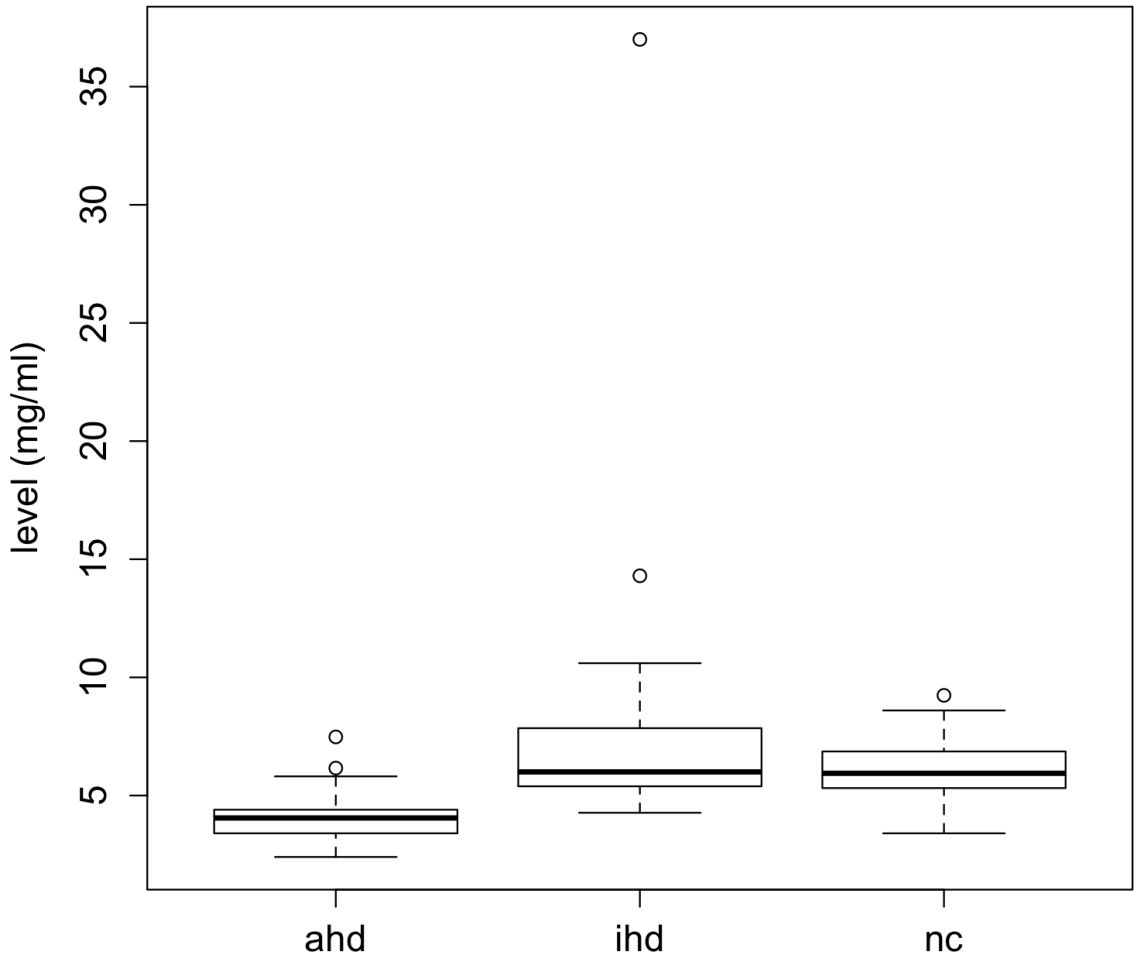
Multiple comparisons

Fancy code to extract several summary statistics at the same time...

```
by(hd.long$hd, hd.long$group, function(X) { c(IQR(X), sd(X),
length(X)) } )
#hd.long$group: ahd
#[1] 10.00000 14.65661 17.00000
#------------------------------------------------------------
#hd.long$group: ihd
#[1] 30.00000 17.99726 29.00000
#------------------------------------------------------------
#hd.long$group: nc
#[1] 21.50000 15.14567 23.00000
```

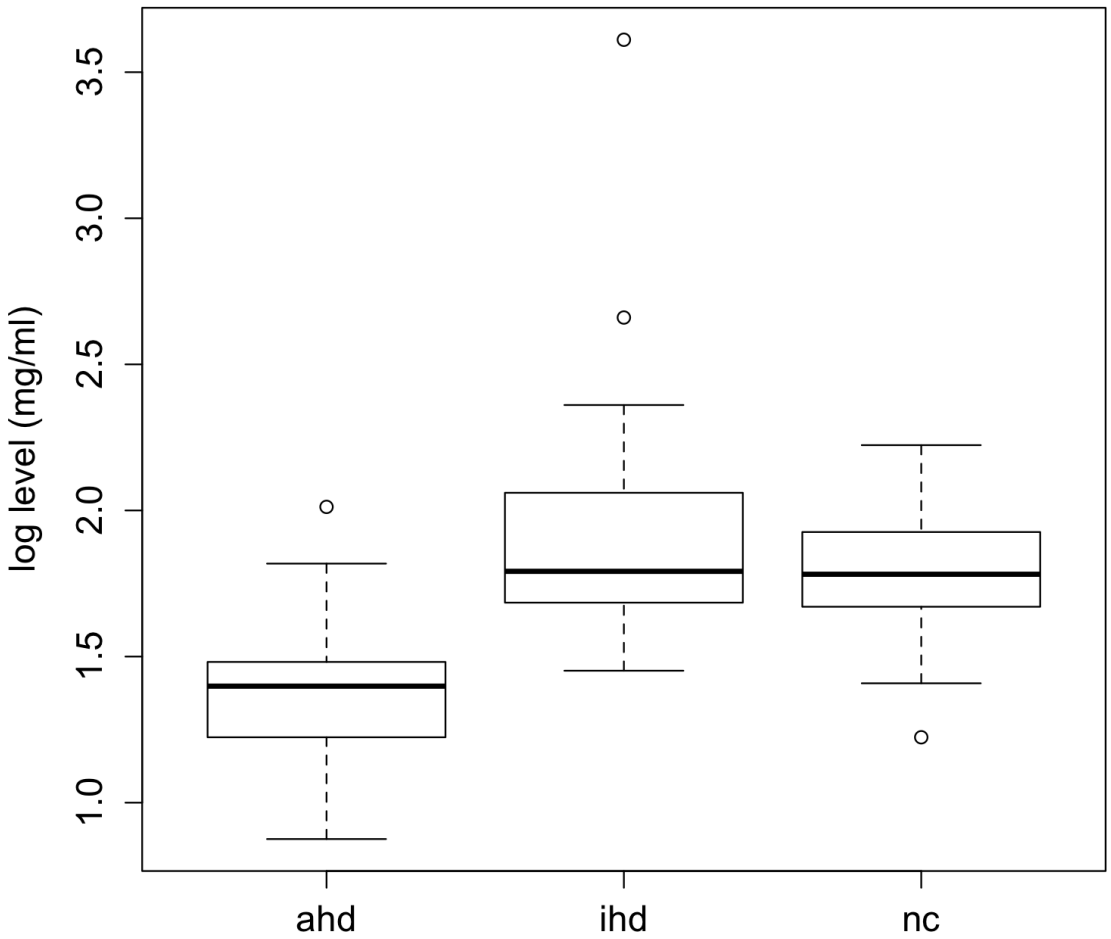
Boxplots for Hodgkin's data

There are some extreme outliers.

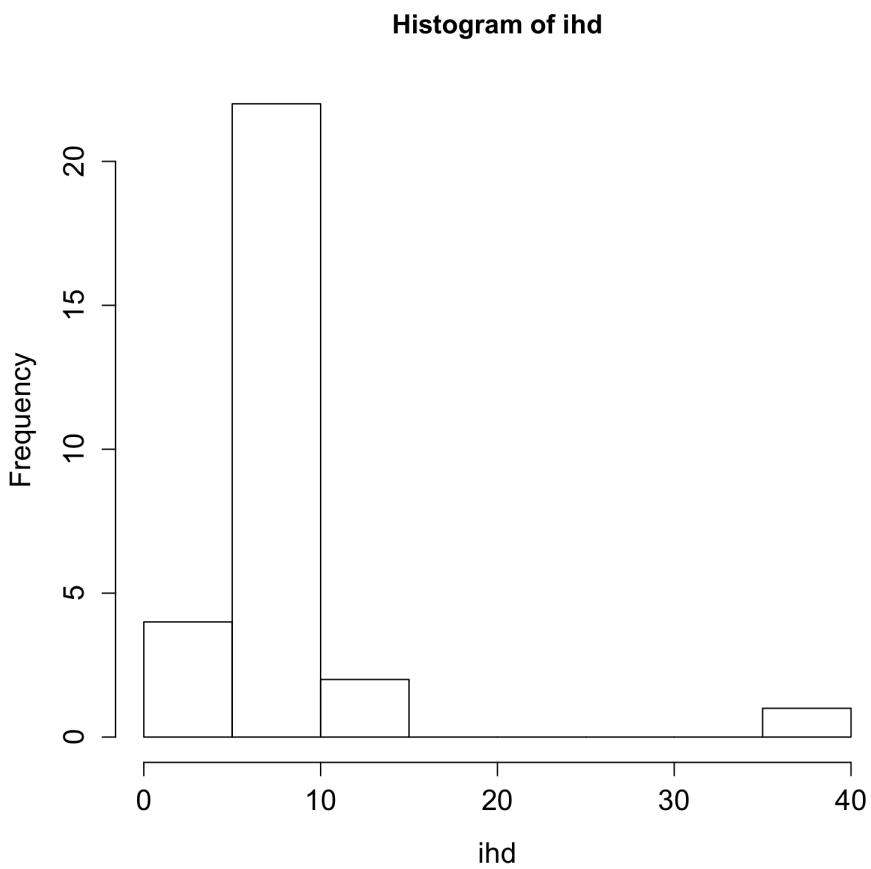


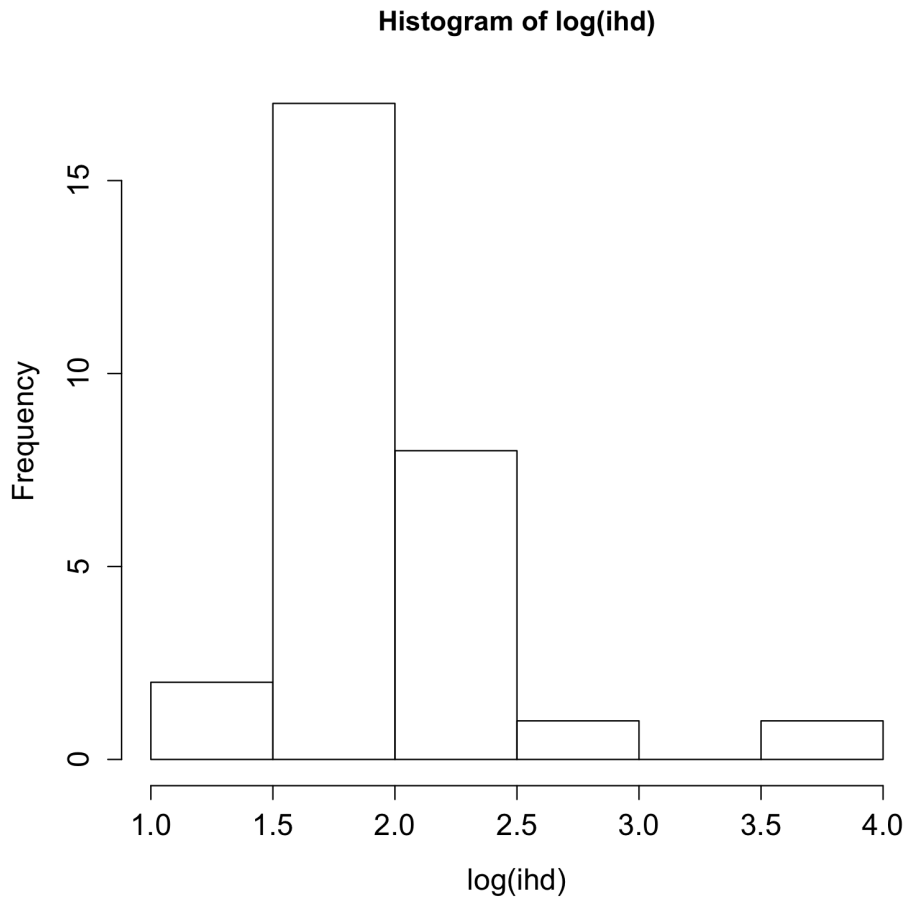
Boxplots for Hodgkin's data (log scale)

Plotting the log of the data looks more reasonable...



Hodgkin's data: ihd group only





Hodgkin's data

Based on the plots of the data, it might be reasonable to use either the KW test or to use ANOVA on the log-transformed data. The null hypothesis is that the population medians are equal:

$$H_0 : \eta_{nc} = \eta_{ahd} = \eta_{ihd}$$

```
fit.h <- kruskal.test(hd ~ group, data = hd.long)
summary(fit.h)
fit.h
Kruskal-Wallis chi-squared = 21.025, df = 2, p-value = 2.719e-05
```

Hodgkin's data

To do multiple comparisons, you can use, for example, Bonferroni comparisons and do the Wilcoxon test pairwise for each pair of groups. There are three pairs: (nc,ahd), (nc,ihd), (ahd,ihd). This suggests using two-sample tests with $\alpha = .05/3 = 0.0167$.

```
wilcox.test(nc,ihd)
#W = 279, p-value = 0.3197
wilcox.test(nc,ahd)
#W = 329, p-value = 0.0002735
wilcox.test(ihd,ahd)
#W = 436, p-value = 1.696e-05
## The results imply the following grouping
##  ahd  nc  ihd
##  ------   ---------
```

Hodgkin's data

If we use ANOVA to analyze the data instead, it is interesting to compare analyzing the log data versus the original data. Both would reject the null hypothesis, but using the log-transformed data, the evidence against the null appears stronger.

```
summary(aov(hd~group,data=hd.long))
#             Df Sum Sq Mean Sq F value Pr(>F)
#group         2  139.9   69.95   4.279 0.0179 *
#Residuals    66 1078.9   16.35

summary(aov(log(hd)~group,data=hd.long))
             Df Sum Sq Mean Sq F value   Pr(>F)
group         2  3.020  1.5101   13.11 1.61e-05 ***
Residuals    66  7.604  0.1152
```

Hodgkin's data

Doing pairwise comparisons with Bonferroni adjustments gives slightly different results using the original data or the transformed data. The transformed data gives the same grouping as the KW test, and given that this analysis satisfied the assumptions better of equal variances, I would prefer using the transformed data over the untransformed data with ANOVA.

```
pairwise.t.test(hd,group, data = hd.long,p.adjust.method="bonf")
#    ahd   ihd
#ihd 0.015 -
#nc  0.478 0.385
## ahd nc ihd
## ------
##     -------
pairwise.t.test(log(hd),group, data = hd.long,
p.adjust.method="bonf")
    ahd     ihd
ihd 9.4e-06 -
nc  0.0028  0.3441
```

Nonparametric methods: permutation tests

A set of nonparametric methods with a very different approach from rank-based methods are permutation tests.

These methods are useful for alternatives to t-tests and ANOVA as well as some other procedures. The idea for two-sample t-tests and ANOVA is that if each group has the same distribution, then you should get roughly the same test statistic if all of the groups are the same, regardless of which group each observation belongs to.

The idea is to recompute the test statistic under different assignments of labels. The distribution of the test statistic is not assumed to follow a standard, named distribution. Instead, the distribution of the test statistic is computed under different permuations of the group labels.

Nonparametric methods: permutation tests

Once the distribution of the test statistic is determined, you can decide whether the observed test statistic is sufficiently unusual compared to the distribution to count as evidence the null hypothesis.

To illustrate with an example, suppose the data are:

data	1 3 6 5 2 7	t_{obs}
group	a a a a b b	-0.274

Then a permutation is

data	1 3 6 5 2 7	t_{obs}
group	b a a b a a	0.645

Here we've permuted the labels and kept the sample sizes the same for each group.

Nonparametric methods: permutation tests

For a permutation test, you repeat this permutation of the labels many times. Each time compute a statistic such as the t-statistic. Keeping track of the computed t-statistic gives you a distribution of the t-statistic that might differ from an actual t-distribution.

For a small example like this, you could in principle enumerate all possible permutations of the labels. Typically, there are too many to enumerate them all, so you just generate a large number of permutations randomly to estimate the distribution.

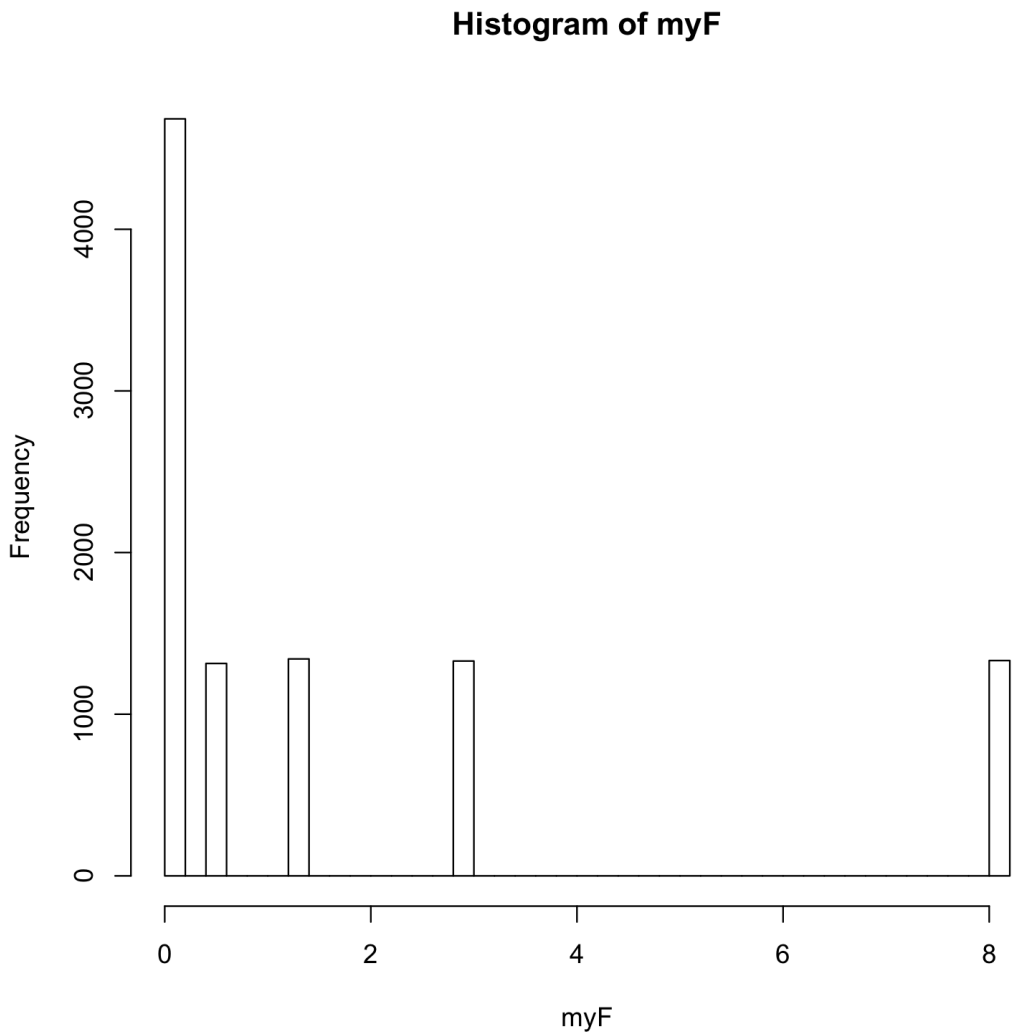
Nonparametric methods: permutation tests

Theory for permutation tests was developed in the 1930s by Fisher and Pitman. However, it is computationally intensive since you have to randomize the labels and recompute quantities. Here is some code. I'm using the `anova()` command so that you can generalize to ANOVA with with 3 or more groups if desired. The idea is the same.

Nonparametric methods: permutation tests

```
x <- c(1,3,6,5,2,7)
group <- c("a","a","a","a","b","b")
a <- aov(x ~ group)
Fobs <- summary(a)[[1]][[4]][1] #tricky to extract F-stat.
I <- 1000
myF <- 1:I
for(i in 1:I) {
  mygroup <- sample(group)
  # do ANOVA but randomize group membership
  temp <- aov(x ~ mygroup)
  myF[i] <- summary(temp)[[1]][[4]][1]
}
hist(myF,nclass=30)
Fobs
#[1] 0.1100917
mean(myF>=Fobs)
# 0.7339  # this is the simulated p-value
```

Nonparametric methods: permutation tests



Nonparametric methods: permutation tests

The simulated p-value is the proportion of F-tests from the permuated labels that values at least as large as the oberved value of 0.1100917.

You can also use a package (of course!) to do the permutation test. One package is the `coin` package. It uses the function `oneway_test()`, which has similar syntax to the `aov()` command and gives a similar p-value.

```
oneway_test(x ~ factor(group))
#
# Asymptotic Two-Sample Fisher-Pitman Permutation Test
#
#data:  x by factor(group) (a, b)
#Z = -0.36596, p-value = 0.7144
#alternative hypothesis: true mu is not equal to 0
```

Nonparametric methods: permutation test on the Hodgkin data

```
x <- hd.long$hd
group <- hd.long$group
a <- aov(x ~ group)
Fobs <- summary(a)[[1]][[4]][1] # returns 4.279
I <- 10000
myF <- 1:I
for(i in 1:I) {
  mygroup <- sample(group)
  # do ANOVA but randomize group membership
  temp <- aov(x ~ mygroup)
  # extract F-statistic, this was a little tricky
  myF[i] <- summary(temp)[[1]][[4]][1]
}
hist(myF,nclass=30)
mean(myF>=Fobs)
#[1] 0.0019
```

Permutation tests on Hodgkin data