R Language Fundamentals

Data Frames

Steven Buechler

Department of Mathematics
276B Hurley Hall; 1-6233

Fall, 2007

Classical Hypothesis Testing

Review or Reading Assignment

Test of a null hypothesis against an alternative hypothesis. There are five steps, the first four of which should be done before inspecting the data.

Step 1. Declare the null hypothesis H_0 and the alternative hypothesis H_1.

In a sequence matching problem H_0 may be that two sequences are uniformly independent, in which case the probability of a match is 0.25. H_1 may be "probability of a match $= 0.35$", or "probability of a match > 0.25".

Classical Hypothesis Testing

Types of hypotheses

A hypothesis that completely specifies the parameters is called simple. If it leaves some parameter undetermined it is composite. A hypothesis is one-sided if it proposes that a parameter is $>$ some value or $<$ some value; it is two-sided if it simply says the parameter is $\neq$ some value.

Types of Error

Rejecting H_0 when it is actually true is called a Type I Error. In biomedical settings it can be considered a false positive. (Null hypothesis says "nothing is happening" but we decide "there is disease".)

Step 2. Specify an acceptable level of Type I error, α, normally 0.05 or 0.01.

This is the threshold used in deciding to reject H_0 or not. If $\alpha = 0.05$ and we determine the probability of our data assuming H_0 is 0.0001, then we reject H_0.

The Test Statistic

Step 3. Select a test statistic.

This is a quantity calculated from the data whose value leads me to reject the null hypothesis or not. For matching sequences one choice would be the number of matches. For a contingency table compute Chi-squared. Normally compute the value of the statistic from the data assuming H_0 is true.

A great deal of theory, experience and care can go into selecting the right statistic.

The Critical Value or Region

Step 4. Identify the values of the test statistic that lead to rejection of the null hypothesis.

Ensure that the test has the numerical value for type I error chosen in Step 2. For a one-sided alternative we normally find a value x_0 so that only $\alpha = 0.05$ values of the statistic are $> x_0$ (or $< x_0$ for an alternative in the other direction). For a two-sided alternative we need thresholds in both directions. We find y_0 and y_1 so that 0.025 values of the statistic are $> y_0$ and 0.025 values of the statistic are $< y_1$.

The Critical Value or Region

Example

The statistic for the number Y of matches between two sequences of nucleotides is a binomial random variable. Let n be the lengths of the two sequences (assume they are the same). Under the null hypothesis that there are only random connections between the sequences the probability of a match at any point is $p = 0.25$. We reject the null hypothesis if the observed value of Y is so large that the chance of obtaining it is < 0.05.

The Critical Value or Region

Example

There is a specific formula for the probability of Y matches in n "trials" with probability of a match $= 0.25$. We can similarly calculate the significance threshold K so that

$$Prob(Y \geq K | p = 0.25) = 0.05.$$

When $n = 100$, $Prob(Y \geq 32) = .069$ and $Prob(Y \geq 33) = .044$. Take as the significance threshold 33. Reject the null hypothesis if there are at least 33 matches.

Obtain the Data and Execute

Step 5. Obtain the data, calculate the value of the statistic assuming the null hypothesis and compare with the threshold.

P-Values

Substitute for Step 4

Once the data are obtained calculate the null hypothesis probability of obtaining the observed value of the statistic or one more extreme in the direction of a one-sided alternative. This is called the p-value. If it is $<$ the selected Type I Error threshold then we reject the null hypothesis.

P-Values

Example

Compare sequences of length 26 under the null hypothesis of only random matches; i.e., $p = 0.25$. Suppose there are 11 matches in our data. In a binomial distribution of length 26 with $p = 0.25$ the probability of ≥ 11 matches is about 0.04. So, with the Type I Error rate, α, at 0.05 we would reject the null hypothesis.

Summary of Hypothesis Testing

- Clearly state the null and alternative hypotheses before designing the experiment.
- Select an optimal test statistic. This is number calculated from the data.
- Under particular assumptions the test statistic has a well-understood distribution under the null hypothesis. Nickname: the null distribution.
- Collect the data and calculate the test statistic.
- If this value is extremely unlikely (based on α and the alternative) in the null distribution we reject the null hypothesis.

Summary of Hypothesis Testing

- Clearly state the null and alternative hypotheses before designing the experiment.
- Select an optimal test statistic. This is number calculated from the data.
- Under particular assumptions the test statistic has a well-understood distribution under the null hypothesis. Nickname: the null distribution.
- Collect the data and calculate the test statistic.
- If this value is extremely unlikely (based on α and the alternative) in the null distribution we reject the null hypothesis.

Summary of Hypothesis Testing

- Clearly state the null and alternative hypotheses before designing the experiment.
- Select an optimal test statistic. This is number calculated from the data.
- Under particular assumptions the test statistic has a well-understood distribution under the null hypothesis. Nickname: the null distribution.
- Collect the data and calculate the test statistic.
- If this value is extremely unlikely (based on α and the alternative) in the null distribution we reject the null hypothesis.

Summary of Hypothesis Testing

- Clearly state the null and alternative hypotheses before designing the experiment.
- Select an optimal test statistic. This is number calculated from the data.
- Under particular assumptions the test statistic has a well-understood distribution under the null hypothesis. Nickname: the null distribution.
- Collect the data and calculate the test statistic.
- If this value is extremely unlikely (based on α and the alternative) in the null distribution we reject the null hypothesis.

Summary of Hypothesis Testing

- Clearly state the null and alternative hypotheses before designing the experiment.
- Select an optimal test statistic. This is number calculated from the data.
- Under particular assumptions the test statistic has a well-understood distribution under the null hypothesis. Nickname: the null distribution.
- Collect the data and calculate the test statistic.
- If this value is extremely unlikely (based on α and the alternative) in the null distribution we reject the null hypothesis.

Outline

Mean of a Normal

Suppose we are given a normally distributed random variable of unkown mean μ but known variance σ^2. In one test the null hypothesis is that the mean is μ_0 and the one-sided alternative is "the mean is $> \mu_0$". Set the type I error as $\alpha = 0.05$. In the experiment we sample n values, $X_1, \ldots, X_n$, of the random variable. The chosen test statistic is the average $\bar{X} = (X_1 + \cdots + X_n)/n$.

Mean of a Normal

Normal; unknown mean, known variance

This is a random variable itself that takes different values for different samples. The theory of sums of random variable implies that $\bar{X}$ is normally distributed with mean μ and variance σ^2/n.

Z-scores

Standardization

Normally distributed random variables are often standardized. If Y is normally distributed with mean m and variance s^2, then $(Y - m)/s$ has a standard normal distribution. This is the Z-score. It measures the number of standard deviations from the mean.

So, if q_{95} is the .95 quantile of the standard normal, the .95 quantile of Y is $m + q_{95}s$.

Calculate Theshold

For one-sided alternative: $\mu > \mu_0$

The .95 quantile of the standard normal is

```
> qnorm(0.95)
```

```
[1] 1.645
```

The .95 quantile of the null distribution is then $t_0 = \mu_0 + 1.645\sigma/\sqrt{n}$. Thus, we reject the null hypothesis if $\bar{X} > t_0$.

Two-sided Alternative

With a two-sided alternative; i.e., that $\mu \neq \mu_0$, we must set thresholds for the alternatives $\mu > \mu_0$ and $\mu < \mu_0$. With a type I error of 0.05 we use the extreme thresholds of the .025 quantile and the .975 quantile.

```
> qnorm(0.025)

[1] -1.96
```

The thresholds are $\mu_0 - 1.96\sigma/\sqrt{n}$ and $\mu_0 + 1.96\sigma/\sqrt{n}$. Rule of thumb: 2 standard deviations from the mean is extreme.

Averages for Non-normal Distributions

Suppose that X is a random variable with mean μ and variance σ^2, which may not be normal. If $X_0, \ldots, X_n$ are indpendent samples from X, then for n sufficiently large, $\bar{X}$ approaches a normal distribution with mean μ and variance σ^2/n. This is by the Central Limit Theorem. For X of any distribution we can test the hypothesis $\mu = \mu_0$ with enough samples.

Outline

Compare Gene Expression Levels

between two cell types

Problem Given a particular gene we want to know if it is expressed differently in two different cell types. That is, is the gene differentially expressed in the two cell types.

Biological and technical variations require that we use numerous replicates of each cell type, taking the mean as the expession level of the cell type.

Compare Means of Two Sample Groups

Strategy Measure the expression levels of m cells of one type and n cells of the second type, and test the null hypothesis that the means are equal.

Assume the measurements are $X_{11}, \ldots, X_{1m}$ for the first cell type and $X_{21}, \ldots, X_{2n}$ for the second cell type.

$\bar{X}_1 = (X_{11} + \cdots + X_{1m})/m$, $\bar{X}_2 = (X_{21} + \cdots + X_{2n})/n$, are the two means.

Assumption about Distributions

of gene expression

Assumption The gene expression levels in the first cell type are normally distributed with mean μ_1 and variance σ^2, and in the second they are normally distributed with mean μ_2 and the same variance σ^2.

Not totally unreasonable when the replicates are true replicates and variance is small.

Select Test Statistic

using the assumption

With these assumptions we use the two-sample t test (with equal variance) calculated as follows.

$$t_0 = \frac{(\bar{X}_1 - \bar{X}_2)\sqrt{mn}}{S\sqrt{m+n}},$$

where S is defined from

$$S^2 = \frac{\sum_{i=1}^{m}(X_{1i} - \bar{X}_1)^2 + \sum_{i=1}^{n}(X_{2i} - \bar{X}_2)^2}{m+n-2}.$$

The Distribution of t_0

it's t

Under the null hypothesis and the assumptions of the case, t_0 calculated from the data as above follows a t distribution with $m + n - 2$ degrees of freedom. This distribution is used to set the threshold for rejecting the null hypothesis.

The t Distribution

The probability density function for the t distribution is `dt`, the cumulative distribution function is `pt` and the quantile function is `qt`. Each of these has a parameter `df` for degrees of freedom.

```
> qt(0.95, df = 5)
```

```
[1] 2.015
```

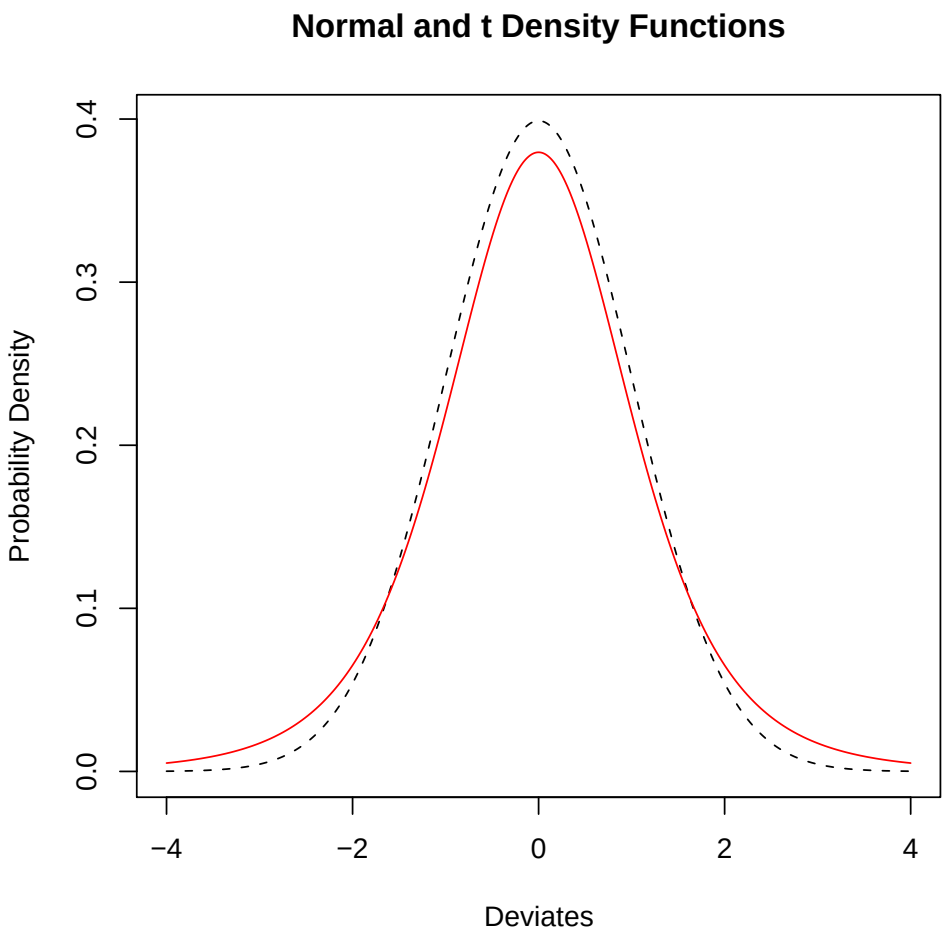
The t Distribution

Density plot compared to normal

How does a t distribution compare to a normal distribution?

```
> xvs <- seq(-4, 4, 0.01)
> plot(xvs, dnorm(xvs), type = "l", lty = 2,
+      ylab = "Probability Density", xlab = "Deviates",
+      main = "Normal and t Density Functions")
> lines(xvs, dt(xvs, df = 5), col = "red")
```

Normal vs. t Density



The t has fatter tails.

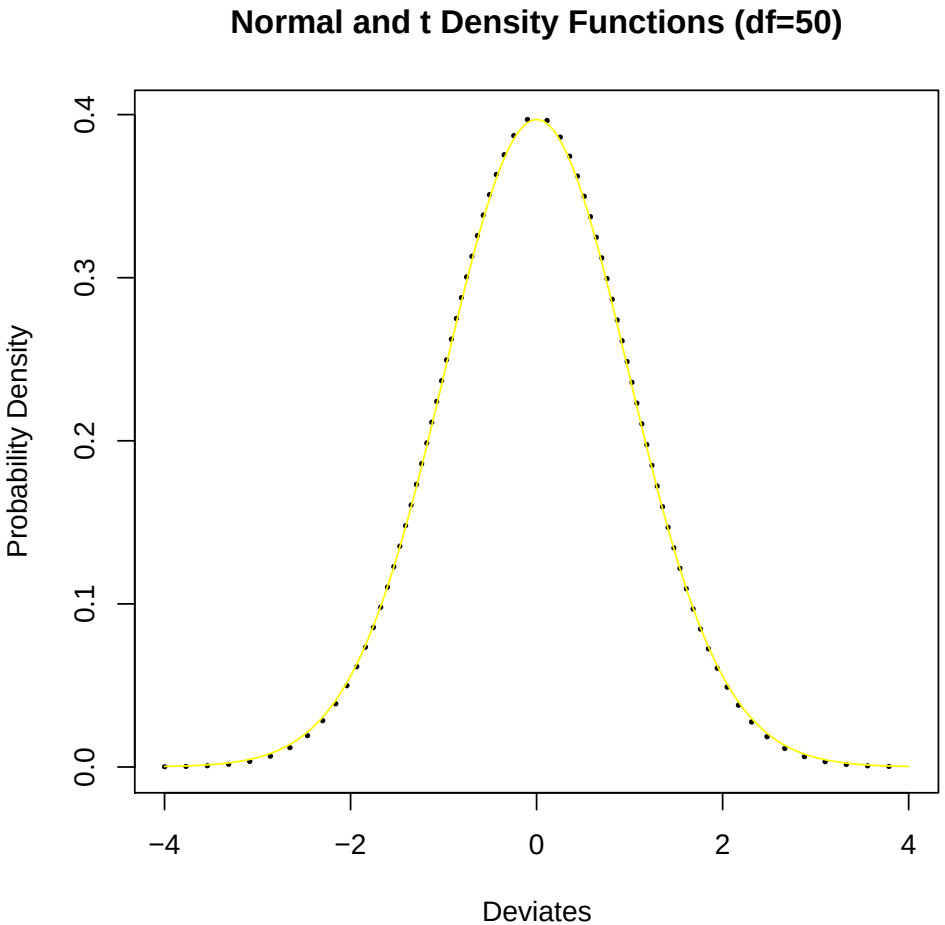
Normal vs. t Density

Effect of df

```
> plot(xvs, dnorm(xvs), type = "l", lty = 3,
+      lwd = 3, ylab = "Probability Density",
+      xlab = "Deviates", main = "Normal and t Density Funct
> lines(xvs, dt(xvs, df = 50), col = "yellow")
```

Normal vs. t Density

Effect of df



Effect of Equal Variance

assumption in this case

It may not be reasonable to assume the variances in the two cell types are the same. There is an alternative statistic, calculated with a different formula than t_0, the Welch's two-sample t test with unequal variance. This also follows a t distribution (with a complicated calculation of degrees of freedom).

More robust is a non-parametric Mann-Whitney test, which needs no assumptions on the distribution of the two sample groups, except that they have the same shape.

t Tests in *R*

R has a simple function `t.test(...)` for carrying out a t test. It has numerous parameters for setting options, like equal variance or unequal variance.

Given sample vectors `x1, x2`, both from normally distributed random variables, the format of a t test is

```
result <- t.test( x1, x2, ... )
```

where ... are optional parameters. See the help on t.test.

t Test Example

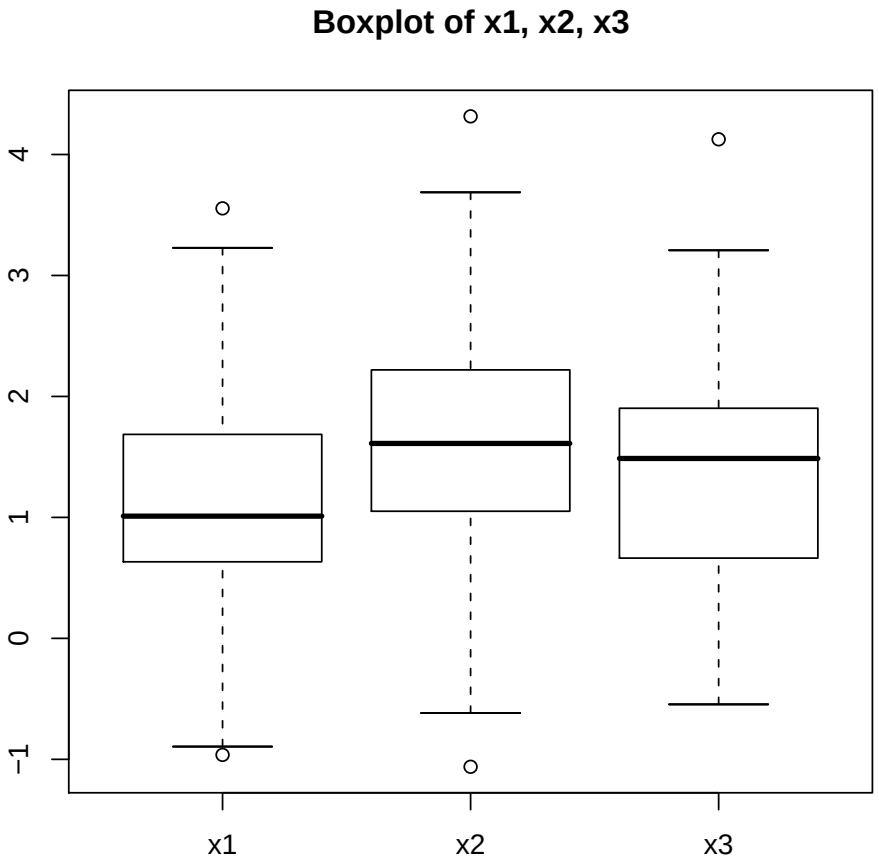
equal variance

We have variables `x1`, `x2`, `x3` supplied as experimental data and we want to compare the means. We can use `summary` to get some feel for the numbers.

```
> summary(x1)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
 -0.963   0.639   1.010   1.070   1.670   3.560
> summary(x2)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
  -1.06    1.05    1.61    1.61    2.21    4.32
> summary(x3)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
 -0.546   0.666   1.490   1.330   1.880   4.130
```

Box and Whisker Plots

```
> boxplot(x1, x2, x3, names = c("x1", "x2",
+     "x3"), main = "Boxplot of x1, x2, x3")
```



Check Hypotheses for t Test

for x1, x2

Are the variances equal?

```
> var(x1)
```

```
[1] 0.8949
```

```
> var(x2)
```

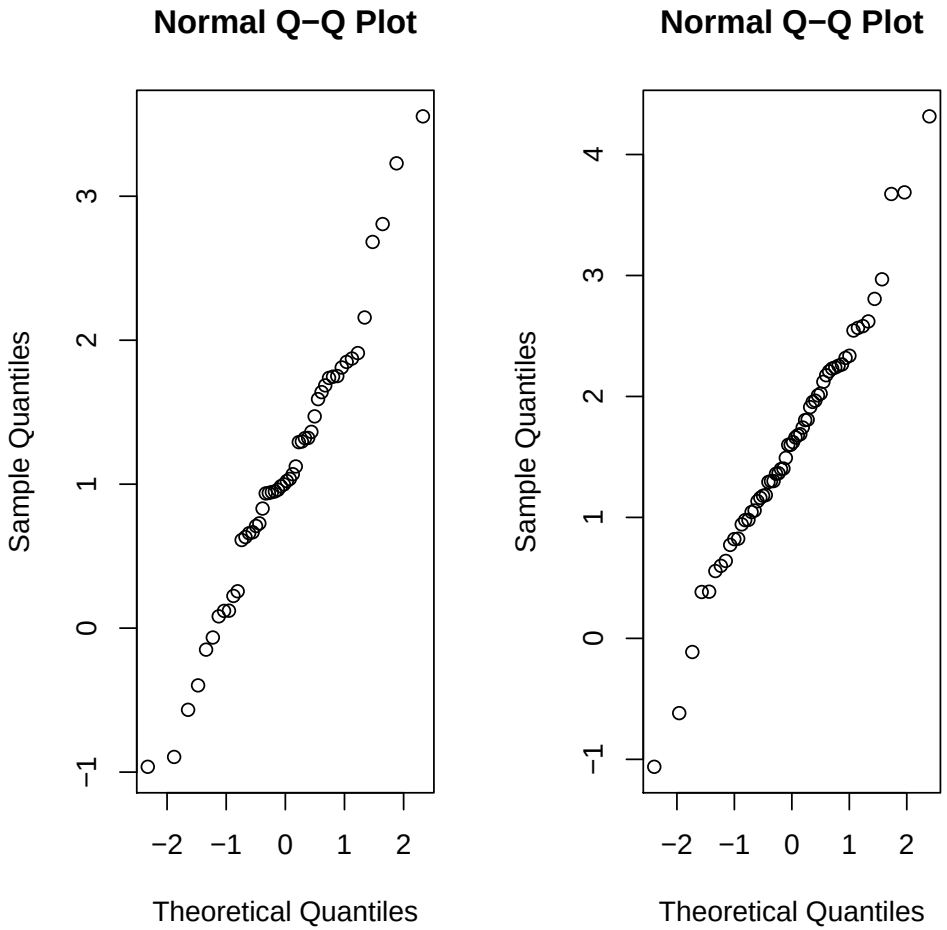
```
[1] 0.9076
```

Check that `x1, x2` are approximately normally distributed.

```
> par(mfrow = c(1, 2))
> qqnorm(x1)
> qqnorm(x2)
> par(mfrow = c(1, 1))
```

Q-Q Normal Plots of Samples

both in one figure



Execute t Test

on x1, x2

Null hypothesis: `mean(x1)=mean(x2)`
Alternative: `mean(x1)`$\neq$`mean(x2)` (two-sided)
Type I Error: 0.05

```
> tx1x2 <- t.test(x1, x2, var.equal = TRUE)
```

Normally, with the result of a test, or fitting a statistical model, the results can be obtained just by typing the object or maybe `summary(object)`. For a t.test, just print.

Output of the t Test

on x1, x2

Result of the test:

```
> tx1x2

        Two Sample t-test

data:  x1 and x2
t = -2.966, df = 108, p-value = 0.003711
alternative hypothesis: true difference in means is not equ
95 percent confidence interval:
 -0.8999 -0.1790
sample estimates:
mean of x mean of y
    1.073     1.613
```

What Kind of Object is Returned?

Interrogate the object as follows:

```
> class(tx1x2)

[1] "htest"

> names(tx1x2)

[1] "statistic"   "parameter"   "p.value"
[4] "conf.int"    "estimate"    "null.value"
[7] "alternative" "method"      "data.name"
```

Often objects are coded like lists so the components carry different aspects of the analysis. These parameters are used to prepare the "report" seen above.

Extracting Individual Components

```
> tx1x2$statistic

     t
-2.966

> tx1x2$parameter

 df
108

> tx1x2$p.value

[1] 0.003711
```

What is p.value for Two-sided Test?

Theoretically, after setting α in a two-sided test we find regions at extremes in the negative and positive directions that each contain $\alpha/2$ of the values. Do we reject the null hypothesis if the `p.value` is $< \alpha$ or $< \alpha/2$?

The p.value is supposedly the quantile value of the test statistic applied to the data. Compute the quantile of the value for our specific example.

What is p.value for Two-sided Test?

```
> pt(tx1x2$statistic, df = 108)

        t
0.001856

> tx1x2$p.value

[1] 0.003711
```

The quantile is half the p.value. Specifying that the test is two-sided caused *R* to adjust the p.value so that we reject the null if the p.value is $< \alpha$.

A Failed t Test

```
> t.test(x1, x3, var.equal = TRUE)

        Two Sample t-test

data:  x1 and x3
t = -1.405, df = 108, p-value = 0.1630
alternative hypothesis: true difference in means is not equ
95 percent confidence interval:
 -0.6075  0.1036
sample estimates:
mean of x mean of y
    1.073     1.325
```

Can't conclude that the mean of `x1` is different from the mean of `x2`.

t Test with Unequal Variances

Welch's Two-Sample t

Given two samples, normally distributed with unknown mean and unknown variance, test the null hypothesis that they have the same mean.

Another version of the t test handles this more general case. Calculate a quantity t_1 much like t_0 from before. Calculate a *pseudo-* degrees of freedom d_1 from a complicated formula. Under the null hypothesis t_1 satisfies a t distribution with d_1 degrees of freedom.

Welch's t Test in *R*

It is trivial to perform this test in *R*; it is the default option of `t.test`.

```
> t2x1x2 <- t.test(x1, x2)
> t2x1x2

        Welch Two Sample t-test

data:  x1 and x2
t = -2.968, df = 104.7, p-value = 0.003713
alternative hypothesis: true difference in means is not equ
95 percent confidence interval:
 -0.8998 -0.1791
sample estimates:
mean of x mean of y
    1.073     1.613
```

Compare the two t's

This test did slightly worse than under the equal variance assumption.

```
> t2x1x2$p.value
```

```
[1] 0.003713
```

```
> tx1x2$p.value
```

```
[1] 0.003711
```

It is harder to "pass" a test with fewer restrictions on the samples.

Outline

Two samples, Arbitrary Distribution

Given: Two samples with a common distribution, which may not be normal. Test the null hypothesis that they have the same mean.

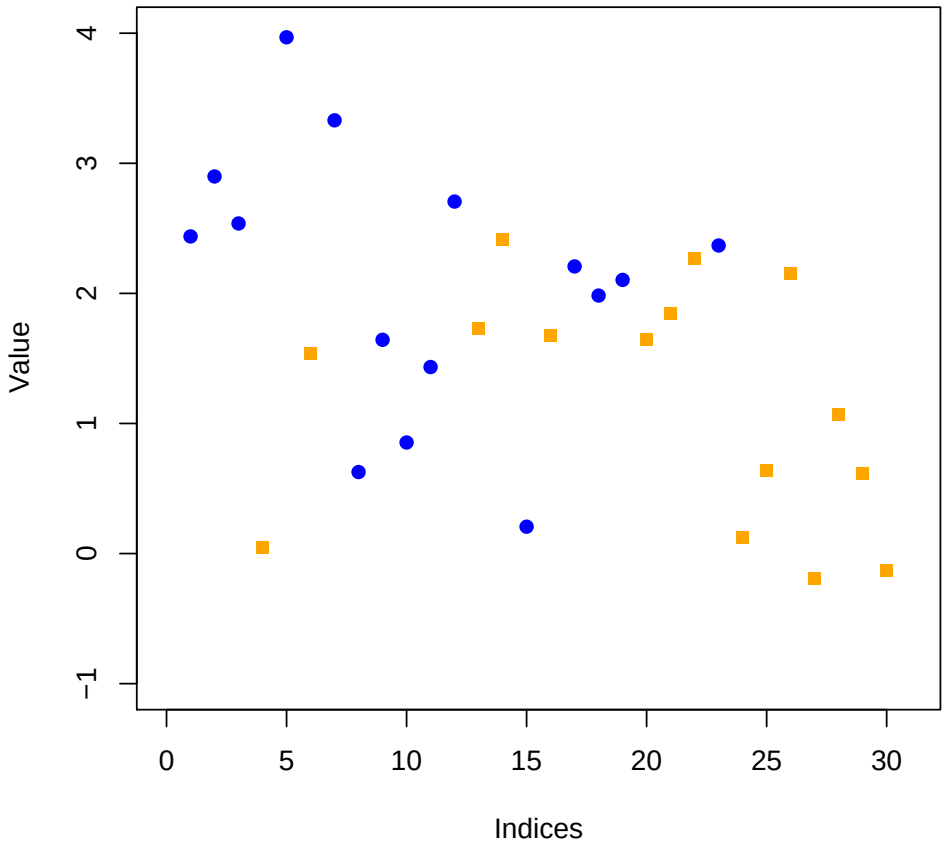
Such methods are called nonparametric or (more accurately) distribution-free. Such tests are conservative in that we make no assumptions about the distribution (except that they're the same in both samples). It is also conservative in that it is difficult to reject the null hypothesis.

Mann-Whitney

The method developed here has a couple of equivalent names: Mann-Whitney test or Wilcoxon rank sum test.The term Mann-Whitney seems most common in biostatistics.

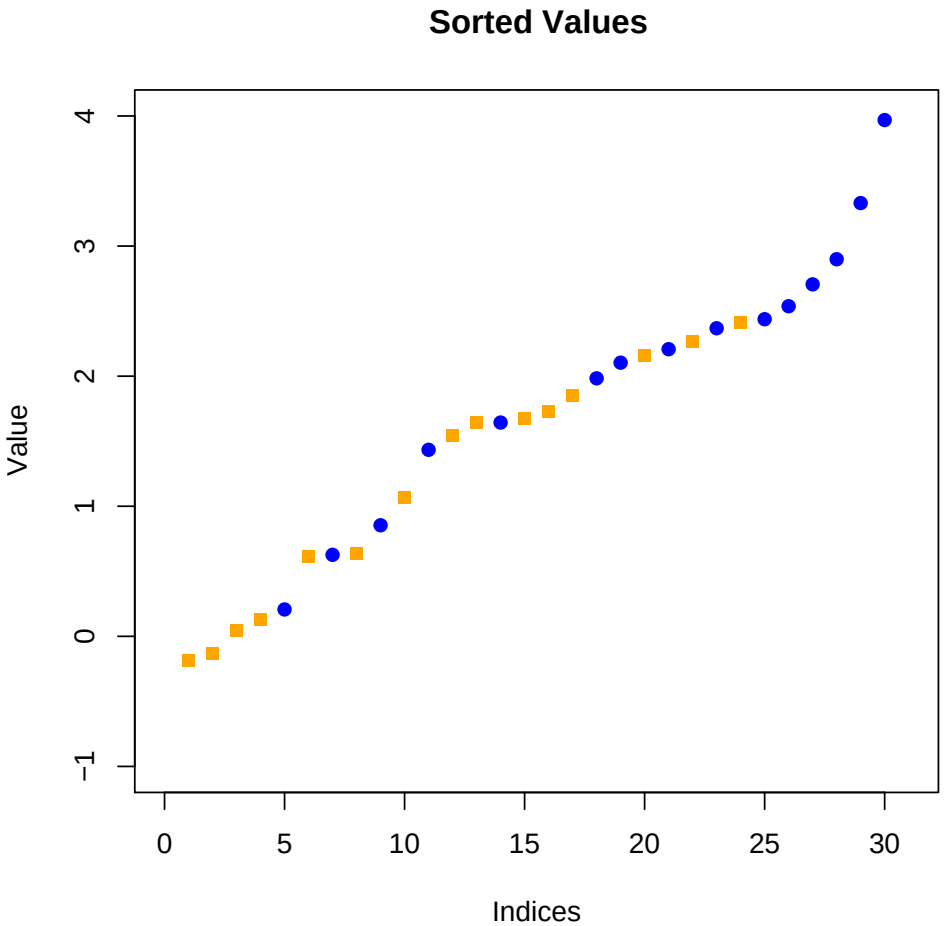
First some point plots to illustrate what's happening here.

Two Samples



How do we decide if one mean is larger than the other?

Sort and Rank



Without calculating means, differences from means, errors, etc., we can study relative sizes.

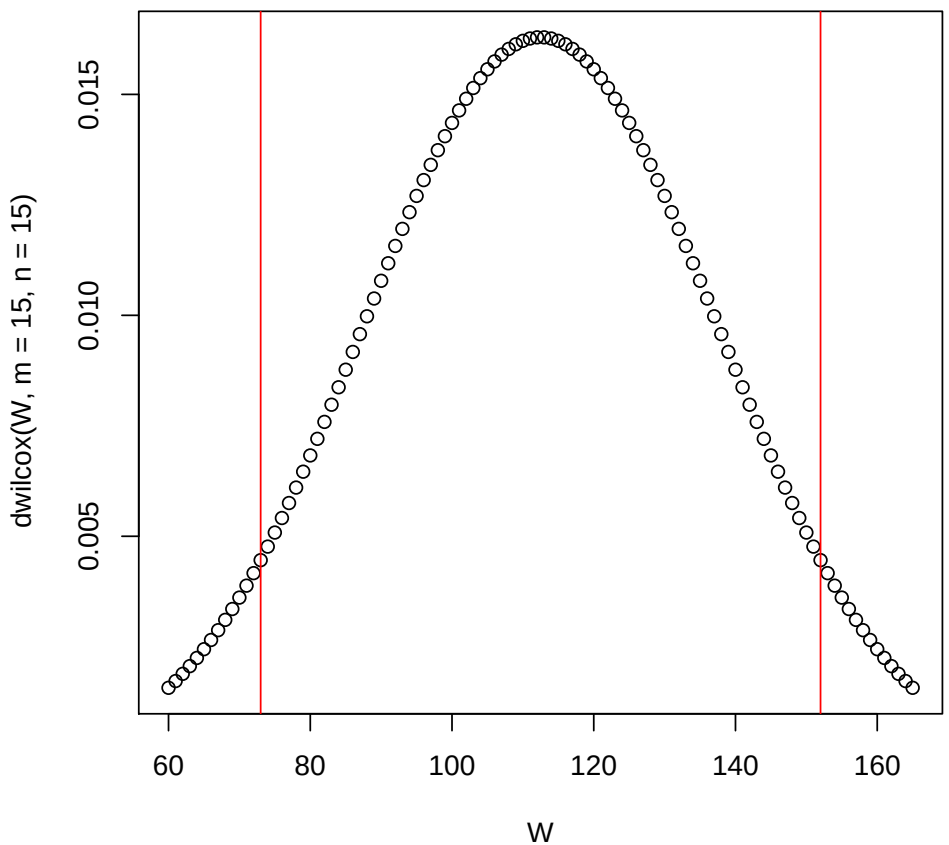
Wilcoxon Rank Sum Statistic

Suppose the first sample group contains m samples and the second n samples.

- Rank order the samples in both groups taken together with the smallest value ranked 1 and the largest ranked $m + n$.
- Add up the ranks of the samples in the first group and call it W.
- Under the null hypothesis W follows a Wilcoxon distribution with parameters for m and n.

Wilcoxon Distribution

Close to Normal



Wilcoxon Test in *R*

Mann-Whitney

Just like with the t test there is a function that performs the Wilcoxon rank sum test in *R*. The format is

```
Wres <- wilcox.test(sampleA, sampleB)
```

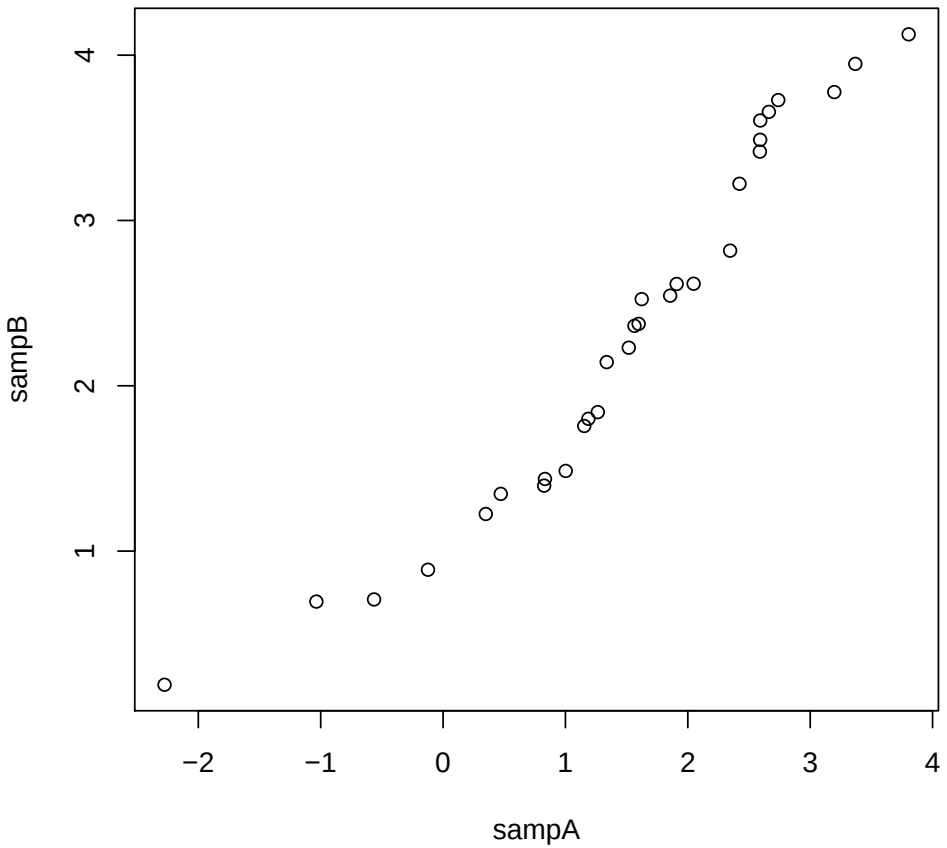
The output, `Wres` is an htest object, just as with a t test.

Example of Wilcoxon Test

We're given two samples, `sampA`, `sampB`, and want to test if the means are the same. Each has 15 points. In this case they aren't far from normal, but let's use a rank sum test for illustration.

Check Distributions

```
> qqplot(sampA, sampB)
```



Not great but OK for samples with few points.

Execute the Wilcoxon Test

```
> wTest <- wilcox.test(sampA, sampB)
> wTest

        Wilcoxon rank sum test

data:  sampA and sampB
W = 297, p-value = 0.02339
alternative hypothesis: true location shift is not equal to
```

Comparing Means in General

The Wilcoxon test is not entirely distribution-free. It assumes the two samples have roughly the same distribution except possibly a difference of means. A more general test to compare the means of two samples that has no assumptions about the distributions can be executed using a bootstrap approximation of the underlying distribution.