Introduction to Computational BioStatistics with R: hypothesis tests

Erik Spence

SciNet HPC Consortium

17 October 2023

SciNet compute • calcul CANADA

Today's slides

Today's slides can be found here. Go to the "Introduction to Computational BioStatistics with R" page, under Lectures, "Hypothesis tests".

`https://scinet.courses/1301`

Hypothesis testing

How do I perform statistical tests on my data?

- Statistical tests are always testing *against* something.
- The thing being tested against is called the *Null Hypothesis*, $\boldsymbol{H_0}$.
- All hypothesis testing is done under the assumption that the Null Hypothesis is true.
- The non-Null Hypothesis is called the Alternate Hypothesis, $\boldsymbol{H_1}$.
- Every hypothesis test is attempting to answer the question: Should I reject the null hypothesis?

You may have heard of the null hypothesis before. What are the characteristics of the null hypothesis?

- It represents NO change from the accepted state of things.
- Whatever is 'normal', or 'default', is the null hypothesis.

Hypothesis testing, example

Step 1: make a claim: "He's dead, Jim", said Dr. McCoy to Captain Kirk. Under usual circumstances your claim will involve some test statistic (mean, variance, *etc.*).

Step 2: determine if the claim being made is the null or alternative hypothesis.

Does this statement represent a change from the normal situation?

- If yes, then it is the alternate hypothesis, $\boldsymbol{H_1}$.
- If not, then it is the null hypothesis (the normal situation is that he is not dead), $\boldsymbol{H_0}$.

Step 3: make a decision (perform a test) to determine whether the null hypothesis should be rejected or not rejected.

- Reject $\boldsymbol{H_0}$: "sufficient evidence to say the patient is dead".
- Fail to reject $\boldsymbol{H_0}$: "insufficient evidence to say patient is dead".

Note that we never accept the null hypothesis, we merely fail to reject it.

Example stolen from James Jones.

Hypothesis testing, example, continued

Based on the two possible states (dead/alive) and the two possible decisions (reject $\boldsymbol{H_0}$/fail to reject $\boldsymbol{H_0}$), there are 4 possible outcomes.

	True state of nature	
Decision	$\boldsymbol{H_0}$ True (patient is not dead)	$\boldsymbol{H_0}$ False (patient is dead)
Reject $\boldsymbol{H_0}$	Patient is not dead, Sufficient evidence of death	Patient is dead, Sufficient evidence of death
Fail to reject $\boldsymbol{H_0}$	Patient is not dead, Insufficient evidence of death	Patient is dead, Insufficient evidence of death

Or, in other words...

	True state of nature	
Decision	$\boldsymbol{H_0}$ True	$\boldsymbol{H_0}$ False
Reject $\boldsymbol{H_0}$	Dispose of a live person	Dispose of a dead person
Fail to reject $\boldsymbol{H_0}$	Try to revive a live person	Try to revive a dead person

Hypothesis testing, example, continued more

	True state of nature	
Decision	$\boldsymbol{H_0}$ True	$\boldsymbol{H_0}$ False
Reject $\boldsymbol{H_0}$	Dispose of a live person	Dispose of a dead person
Fail to reject $\boldsymbol{H_0}$	Try to revive a live person	Try to revive a dead person

These cases are so common (and well-studied) that they have been given names.

	True state of nature	
Decision	$\boldsymbol{H_0}$ True	$\boldsymbol{H_0}$ False
Reject $\boldsymbol{H_0}$	Type I error (alpha)	Correct assessment
Fail to reject $\boldsymbol{H_0}$	Correct assessment	Type II error (beta)

Type I errors are considered more serious.

How hypothesis tests work

Suppose we ask the question: does a certain antibiotic work at killing bacteria, in the lab?

- The null hypothesis is that the antibiotic does not affect the bacteria population.
- All measurements have errors, and there will be variations (randomness) in the data, so even if there is no effect the average bacteria population differences between the two measurements (before and after antibiotic application) will be non-zero.
- We only have a real difference in bacteria populations if the difference we see is *unlikely to occur by chance* if the real difference is zero.
- The probability of a difference as big as we measure occurring when the null hypothesis of no difference is true is called p-value. This is calculated with the test statistic:

$$t = \frac{\overline{x} - 0}{s/\sqrt{n}}$$

where s is the sample standard deviation, and x is the difference between the population measurements, for a given sample, before and after antibiotic application

How hypothesis tests work, continued

Due to randomness, we expect $\overline{x}$ to have an approximate Gaussian distribution, with $\mu = 0$ if the null hypothesis is true.

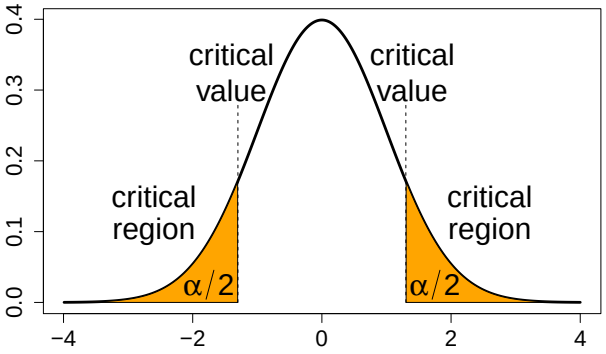
- Using our data, we calculate t.
- This tells us how many standard deviations away from 0 our data is.
- Using this, we can determine the probability that this value of t, or greater, would occur if the null hypothesis is true.
- This is our p-value.
- This is equivalent to determining the probability of committing a Type I error (incorrectly rejecting the null hypothesis).

For those familiar with these sorts of tests, this is an example of a paired two-sample t-test.

Test significance

The significance level (alpha) is used to determine if the test statistic is far enough out in the tails to reject the null hypothesis.

- Commonly used values of significance are 0.05 and 0.01.
- Most tests return a "p-value". This is the probability of committing a Type I error, given the input data.
- If the p-value is less than the significance, then the null hypothesis can be rejected.



You must decide what significance you are using before you run the test!

How to perform a hypothesis test

Ok, so you want to perform a hypothesis test on your data. What are the steps involved?

- Write the claim, and determine whether it is the null or alternate hypothesis.
- Choose the level of significance (α).
- Perform the test.
- Reject or fail to reject the null hypothesis.
- Write a conclusion.

We've already discussed determining the type of hypothesis from the claim, and the significance level.

Which test should I run?

Choosing which tests to run on your data is a problem everyone faces.

Questions you must ask to determine which tests are appropriate:

- Are your data of different groups (multi-sample tests) or not (single-sample tests)?
- If your data are in groups, are your groups paired or unpaired (are the groups related to/dependent on each other?
- If your data are numeric, do the data follow a known distribution (if they do, use "parametric" tests, otherwise "non-parametric" tests)?
- If your data are categorical, how many groups are there?

We will review some of the most-commonly used tests, what they are used for, and when they apply.

With material stolen from `https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3116565`

SciNet compute • calcul CANADA

Warning!

This class is very black-box in its approach. What does this mean?

- The approach will be very cook-book.
- We will not be going into the statistical theory; this is not a course in statistics.
- Instead we will merely cover the when, what and why of the tests.

This oversight is intentional. We just don't have the time to cover the theory behind the tests.

SciNet compute • calcul CANADA

One-sample tests

If your data only consists of a single "group" of data, and there's no separate group you're testing against, it's likely you should run a "one-sample" test.

- These tests involve a single sample of data.
- As such, the null hypothesis we are testing against must now be some hypothetical quantity or quality.
- Examples of such hypothetical quantities and qualities include:
  - The sample's mean (one-sample t-test (Gaussian) or Wilcoxon Signed-Rank (non-Gaussian)).
  - The sample's proportion (proportion test).
  - The sample's distribution (Shapiro-Wilk normality test).

Note that these tests are all numeric.

SciNet compute • calcul CANADA

Is my data Gaussian?

Suppose you've got some data, and you're curious as to the distribution which generates it. Maybe it's Gaussian?

Tests exist to determine whether a set of data is likely from a given distribution. Tests for the normal distribution include:

- Shapiro-Wilk (shapiro.test)
- Anderson-Darling (ad.test)
- Lilliefors (lillie.test)
- Pearson Chi-square (pearson.test)

These tests usually take the null hypothesis to be the case that the data IS normally distributed. Nonetheless, always be sure to read the documentation to confirm what the null hypothesis is. Otherwise, you won't know the p-value is referring to.

SciNet compute • calcul CANADA

Is my data Gaussian?, continued

Suppose that we are examining the 'trees' data set.

The null hypothesis is that the data IS normally distributed.

The p-value from two different tests suggest that the null hypothesis cannot be rejected.

Note that many of the tests for normality are found in the 'nortest' library.

```
>
> shapiro.test(trees$Height)

     Shapiro-Wilk normality test

data:  trees$Height
W = 0.96545, p-value = 0.4034
>
> library(nortest)
>
> ad.test(trees$Height)

     Anderson-Darling normality test

data:  trees$Height
A = 0.35926, p-value = 0.4282
>
```

Is my data Gaussian?, continued more

If you're not sure what the null hypothesis is, you can always test it by testing the test.

If you go down this route, be sure to run the tests many times, to make sure you don't accidentally stumble upon a case that incorrectly rejects the null hypothesis.

In this case the null hypothesis that the data is normally distributed, when the data is drawn from the uniform distribution, can be confidently rejected.

```
>
> shapiro.test(rnorm(1000))

     Shapiro-Wilk normality test

data:  rnorm(1000)
W = 0.99882, p-value = 0.767
>
> ad.test(runif(1000))

     Anderson-Darling normality test

data:  runif(1000)
A = 12.326, p-value < 2.2e-16
>
```

One-sample test, example

An outbreak of Salmonella-related illness was attributed to ice cream produced at a certain facility. Levels of Salmonella were measured in 9 randomly sampled batches.

Is there evidence that the mean Salmonella concentration is greater than 0.3 MPN/g?

Question 1: is the data Gaussian? Not sure, let's find out.

```
>
> ice.cream <- c(0.593, 0.142, 0.329, 0.691,
+       0.231, 0.793, 0.519, 0.392, 0.418)
>
```

One-sample test, example, continued

Our first null hypothesis: the data are normally distributed. Let us use the Shapiro-Wilk test to check.

We will use the standard significance level of 0.05.

Result: null hypothesis not rejected!

Question 1: We may assume so.

```
>
> shapiro.test(ice.cream)

     Shapiro-Wilk normality test

data:  ice.cream
W = 0.98271, p-value = 0.9767
>
```

One-sample test, example, continued more

Data: numeric, Gaussian.

Test: one-sample t-test. This will test to see if mean of the data (μ) is greater than 0.3.

Our null hypothesis: the mean is equal to 0.3 ($\mu = \mathbf{0.3}$).
Alternative hypothesis: the mean is greater than 0.3 ($\mu > \mathbf{0.3}$).

We will use a significance level of 0.05.

Before we do the test, however, it's important to understand that there can be slight differences between one- and two-sample tests. These differences sometimes affect how the test is set up.

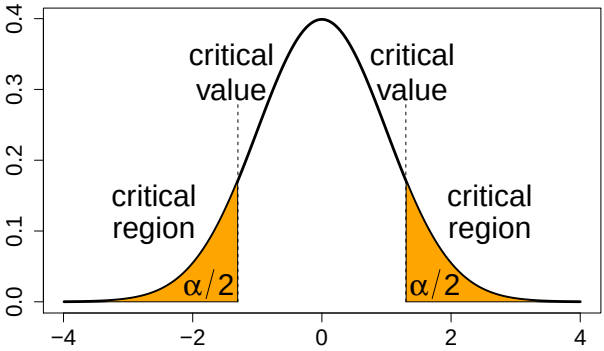
Types of test: two-tailed test

The standard type of test is the "two-tailed" test.

In this case, the null hypothesis is rejected if the test statistic is either

- greater than the upper critical value,
- lower than the lower critical value.

This is the default test, if no side is specified.

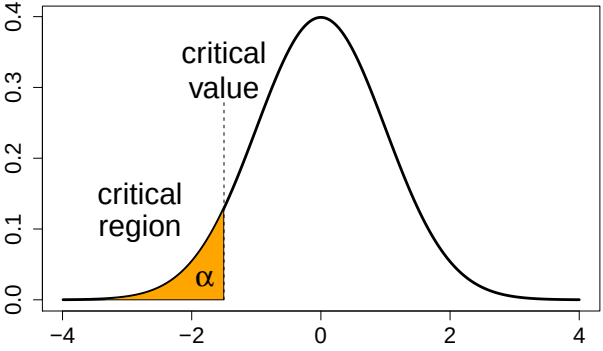


Types of test: left-tailed test

There are several types of test, determined by how you might accidentally commit a Type I error (incorrectly reject the null hypothesis).

For the left-tailed test, we reject the null hypothesis if the test statistic is less than the critical value.

More typically, as with previous tests, we use the p-value and the prechosen significance level.

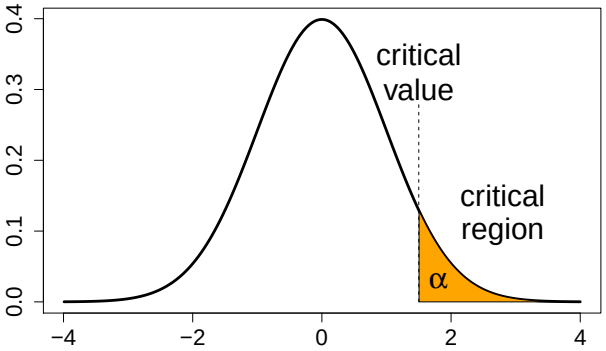


Types of test: right-tailed test

The other "one-tailed" or "one-sided" test is the right-tailed test.

For the right-tailed test, we reject the null hypothesis if the test statistic is greater than the critical value.

Again, the p-value and significance level are used to interpret the test.



One-sample test, example, continued even more

Data: numeric, Gaussian.

Test: one-sample t-test.

$\boldsymbol{H_0}$: mean of the data equal to 0.3 ($\boldsymbol{\mu = 0.3}$).
$\boldsymbol{H_1}$: the mean is greater than 0.3 ($\boldsymbol{\mu > 0.3}$).

Note that we must specify the type of test and the mean we are testing against. The "alternative = 'greater'" argument indicates that this is a right-tailed test.

We will use a significance level of 0.05.

Result: null hypothesis rejected!

```
>
> t.test(ice.cream, mu = 0.3,
+     alternative = "greater")

One Sample t-test

data:  ice.cream
t = 2.2051, df = 8, p-value = 0.02927
alternative hypothesis:  true mean is
greater than 0.3
95 percent confidence interval:
0.3245133 Inf
sample estimates:
mean of x
0.4564444
>
```

Unpaired tests

Does your data consist of groups which are not paired? Unpaired means the groups are unrelated to, or are not influenced by, the others.

Which tests should I run?

- If your data are numeric and Gaussian:
  - 2 groups: unpaired t-test.
  - >2 groups: Analysis of variance (ANOVA) or F test.
- If your data are numeric and not Gaussian (or unknown):
  - 2 groups: Mann-Whitney U test.
  - 2 groups: Wilcoxon's rank sum test.
  - >2 groups: Kruskall-Wallis H test (Kruskall-Wallis ANOVA).
- If your data are categorical:
  - 2 or more groups: Chi-squared test.
  - 2 groups: Fisher's exact test.

These tests all examine the means of the two groups.

Unpaired test, example

The birth weight of newborns was collected at a Massachusetts hospital in 1986. Some mothers were smokers, others weren't. Did the smoking status of the mother affect the birth weights of the babies?

Question 1: are the data paired? No.

Question 2: are the data numeric or categorical? Numeric.

Question 3: are the data Gaussian? Not sure, let's find out.

```
>
> library(MASS)
>
> smoking <- birthwt$bwt[birthwt$smoke == 1]
>
> non.smoking <- birthwt$bwt[birthwt$smoke == 0]
>
> length(smoking)
[1] 74
>
> length(non.smoking)
[1] 115
>
```

Unpaired tests, example, continued

Our null hypothesis: the data are normally distributed. Let us use the Shapiro-Wilk test to check.

We will use the standard significance level of 0.05.

Result: null hypothesis not rejected!

Question 3: are the data Gaussian? We may assume so.

```
>
> shapiro.test(smoking)

     Shapiro-Wilk normality test

data:  smoking
W = 0.98296, p-value = 0.4195
>
> shapiro.test(non.smoking)

     Shapiro-Wilk normality test

data:  non.smoking
W = 0.98694, p-value = 0.3337
>
```

Unpaired tests, example, continued more

Data: unpaired, numeric, Gaussian.

Test: t-test. This will check to see if there is a significant difference in the two population means.

Null hypothesis: there is no difference between the two populations.

We will use a significance level of 0.05.

Result: null hypothesis rejected!

```
>
> t.test(smoking, non.smoking)

Welch Two Sample t-test

data:  smoking and non.smoking
t = -2.7299, df = 170.1, p-value = 0.007003
alternative hypothesis:  true difference in
means is not equal to 0
95 percent confidence interval:
-488.97860 -78.57486
sample estimates:
mean of x mean of y
2771.919 3055.696
>
```

Paired data tests

Does your data consist of groups which are paired?

- Pairing usually takes the form of repeated measurements of the same subjects.
- Pairing can also apply for different subjects that are connected to each other some how (twins, siblings, parent-child).

If your data are paired, numeric and Gaussian:

- 2 groups: paired t-test (compares if two groups are from the same distribution).
- >2 groups: repeated measures analysis of variance (ANOVA).

If your data are paired, numeric and not Gaussian (or you don't know):

- 2 groups: Wilcoxon signed-ranks test (like the t-test, but the distributions must be symmetric).
- >2 groups: Friedman's ANOVA.

If your data are paired and categorical:

- 2 groups: McNemar's test.
- >2 groups: Cochran's Q test.

Paired tests, example 1

Some mice received treatment X. Did the treatment X have any impact on the weight of the mice?

The weights of the mice were measured before and after treatment.

Question 1: are the data paired? Yes.

Question 2: are the data numeric or categorical? Numeric.

Question 3: are the data Gaussian? Not sure, let's find out.

```
>
> before <- c(200.1, 190.9, 192.7, 213,
+   241.4, 196.9, 172.2, 185.5, 205.2, 193.7)
>
> after <- c(392.9, 393.2, 345.1, 393, 434,
+   427.9, 422, 383.9, 392.3, 352.2)
>
> my.data <- data.frame(weight = c(before, after)
+   group = rep(c("before", "after"), each = 10))
>
```

Example stolen from `http://www.sthda.com/english/wiki/paired-samples-t-test-in-r`

Paired tests, example 1, continued

Our first null hypothesis: the data are normally distributed. Let us use the Shapiro-Wilk test to check.

We will use the standard significance level of 0.05.

Result: null hypothesis not rejected!

Question 3: are the data Gaussian? We may assume so.

```
>
> shapiro.test(before)

     Shapiro-Wilk normality test

data:  before
W = 0.90938, p-value = 0.2768
>
> shapiro.test(after)

     Shapiro-Wilk normality test

data:  after
W = 0.91121, p-value = 0.2894
>
```

SciNet compute • calcul CANADA

Paired tests, example 1, continued more

Data: paired, numeric, Gaussian.

Test: paired t-test. This will test to see if there is a significant difference between the means in the two populations after treatment.

Our null hypothesis: the data are the same after treatment as before (no difference in means).

We will use a significance level of 0.05.

Result: null hypothesis rejected!

Conclusion: "We conclude that the weights of the mice after treatment are significantly different than the weights before treatment."

```
>
> t.test(before, after, paired = TRUE)

Paired t-test

data:  before and after
t = -20.883, df = 9, p-value = 6.2e-09
alternative hypothesis:  true difference
in means is not equal to 0
95 percent confidence interval:
-215.5581 -173.4219
sample estimates:
mean of the differences
-194.49
>
```

Paired tests, example 2

Children were surveyed at ages 12 and 14. They were asked if they had had a severe cold in the previous 12 months.

Based on these data, was there a significant increase in the number of severe colds?

Question 1: are the data paired? Yes.

Question 2: are the data numeric or categorical? Categorical.

Question 3: how many groups? 2.

```
>
> my.data <- matrix(c(212, 256, 144, 707), nrow = 2,
+   dimnames = list(
+      "Colds at age 12" = c("Yes", "No"),
+      "Colds at age 14" = c("Yes", "No")))
>
> my.data
                Colds at age 14
Colds at age 12 Yes No
            Yes 212 144
            No  256 707
>
```

This type of data can be organized in a 2 x 2 table.

Paired tests, example 2, continued

Data: paired, categorical, 2 groups.

Test: McNemar's test.

Null hypothesis: no significant change in the frequency of severe colds.

We will use a significance level of 0.05.

Result: the null hypothesis is rejected.

```
>
> mcnemar.test(my.data, correct = FALSE)

McNemar's Chi-squared test

data:  my.data
McNemar's chi-squared = 31.36, df = 1,
p-value = 2.144e-08
>
```

Other types of test

There are many other categories of tests which are out there. The one for your data problem may not have been covered here. Here are other classes of data analysis tests and techniques.

- Association tests.
- Graphical methods.
- Power analysis.
- Survival analysis techniques.
- Time-series analysis techniques.

If you're not sure which test to try, visit this site:

`https://stats.idre.ucla.edu/other/mult-pkg/whatstat`

SciNet compute • calcul CANADA