Quantitative Applications for Data Analysis: hypothesis tests

Erik Spence

SciNet HPC Consortium

11 February 2025

Today's slides

Today's slides can be found here. Go to the "Quantitative Applications for Data Analysis" page, under Lectures, "Hypothesis Tests".

`https://scinet.courses/1376`

A review of hypothesis testing

Recall the steps we follow to perform a hypothesis test:

- 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. First, we will be examining multi-sample tests. This means tests that involve multiple samples of data (data from multiple groups), to see how their *means* compare to each other.

Questions you must ask to determine which tests are appropriate:

- 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`

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 intensional. We just don't have the time to cover the theory behind the tests.

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 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.

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.

What distribution describes my data?, 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
>
```

SciNet

What distribution describes my data?, 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 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 are slight differences between one- and two-sample tests. These differences affect how the test is set up.

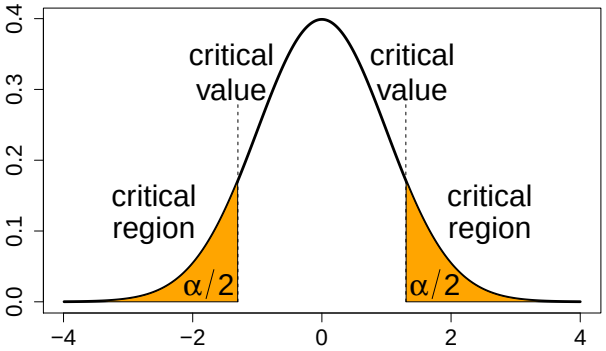
Types of test: two-tailed test

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.

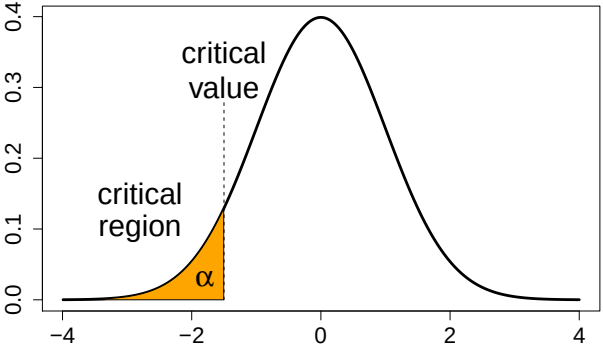


Types of test: left-tailed test

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 signficance level.

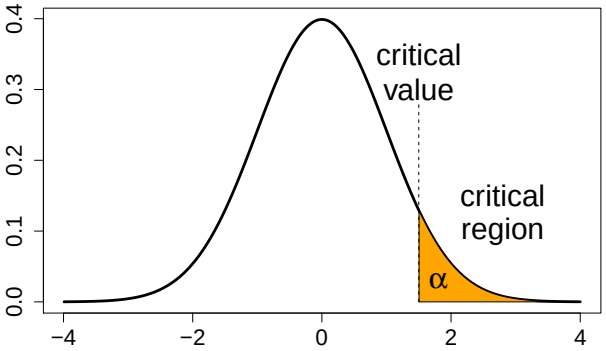


Types of test: right-tailed test

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.

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
>
```

SciNet

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 populations.

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.

SciNet

Paired tests, example

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, 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
>
```

Analysis of variance

What is analysis of variance (ANOVA)?

- ANOVA is used to examine the means of different groups in a sample.
- ANOVA, in its simplest sense, generalizes the t-test to more than 2 groups.
- ANOVA can be used if
  - the data is unpaired, numeric and Gaussian (ANOVA),
  - the data is unpaired, numeric and non-Gaussian (Kruskall-Wallis ANOVA (H test)),
  - the data is paired, numeric and non-Gaussian (Friedman's ANOVA).
- ANOVA assumes that the group variances are equal.
- ANOVA operates under the null hypothesis that all group means are the same.
- If there are many groups, and the null hypothesis is rejected, further tests must be performed to determine which groups are different from each other.

Let's look at an example.

Analysis of variance, example

Suppose that 3 drugs (A, B, X) are tested for treating ankle pain. A study with 27 volunteers is performed by randomly assigning 9 subjects to each drug, and then registering pain level.

Our question: do the drugs differ at all in their performance?

Question 1: is the data paired? No.

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

```
>
> pain <- c(4, 5, 4, 3, 2, 4, 3, 4, 4, 6, 8, 4, 5,
+       4, 6, 5, 8, 6, 6, 7, 6, 6, 7, 5, 6, 5, 5)
>
> drug <- c(rep("A", 9), rep("B", 9), rep("X", 9))
>
> treatment <- data.frame(pain, drug)
>
```

Question 3: is the data Gaussian? In this case the question is, is the data within each group Gaussian?

Let's find out.

Example stolen from `http://www.stat.columbia.edu/~martin/W2024/R3.pdf`

Analysis of variance, 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: is the data Gaussian? We may assume so.

```
> shapiro.test(pain[drug == "A"])

     Shapiro-Wilk normality test

data:  pain[drug == "A"]
W = 0.87282, p-value = 0.1318
```

```
> shapiro.test(pain[drug == "B"])

     Shapiro-Wilk normality test

data:  pain[drug == "B"]
W = 0.88654, p-value = 0.1838
```

```
> shapiro.test(pain[drug == "X"])

     Shapiro-Wilk normality test

data:  pain[drug == "X"]
W = 0.83798, p-value = 0.05485
```

Homogeneity of variance, an aside

The ANOVA assumes that the variances of the data within groups are the same (homoscedasticity). How is this assumption checked? Shockingly, "there's a test for that":

- F-test of equality of variances: tests if two normal populations have the same variance.
- Hartley's test: assumes the data are normal, and that each groups has the same number of entries.
- Levene's test: useful for data with more than one group.
- Brown-Forsythe test: also used for data with more than one group.
- Barlett's test, and others.

These tests use the null hypothesis that the variances are equal.

Analysis of variance, example, continued more

Question 4: Is the data homoscedastic? Don't know, let's find out.

Our null hypothesis: each group of the data have the same variance. Let us use Levene's test to check.

We will use the standard significance level of 0.05.

Result: null hypothesis not rejected!

```
>
> library(car)
>
> leveneTest(pain ~ drug, data = treatment)
Levene's Test for Homogeneity of Variance (center =
median)
         Df   F value   Pr(>F)
  group   2    1.6667     0.21
         24
>
```

Question 4: is the data homoscedastic? We may assume so.

Analysis of variance, example, continued some more

Data: unpaired, numeric, Gaussian, homoscedastic, 3 groups.

Test: Analysis of variance (ANOVA).

Null hypothesis: no significant difference between the three groups.

We will use a significance level of 0.05.

```
>
> anova.result <- aov(pain ~ drug, data = treatment)
>
> summary(anova.result)
             Df  Sum Sq  Mean Sq  F value     Pr(>F)
drug          2   28.22   14.111    11.91   0.000256   ***
Residuals    24   28.44    1.185
>
```

Result: the null hypothesis is rejected. There IS a difference between these groups.

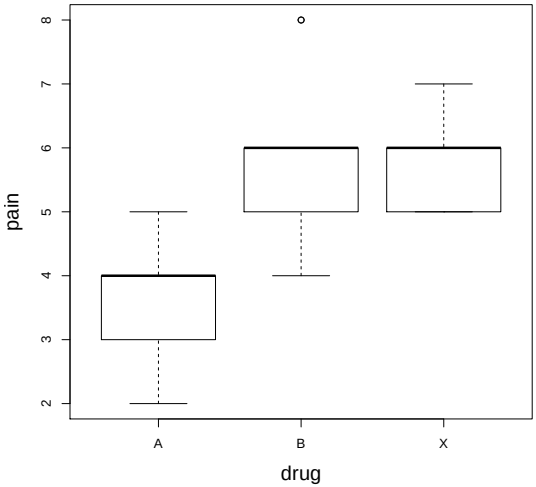
The problem now, as mentioned before, is that we don't know which particular groups are different from which.

Analysis of variance, example, visual result

```
>
> plot(pain ~ drug, data = treatment)
>
```

One quick and easy way to get some intuition as to which groups are doing what is to just make a quick bar plot of the data.

This is not rigorous, of course, but helps with understanding what the data is doing.



ANOVA post hoc tests

If the p-value of a one-way ANOVA is significant we know that some of the group means are different, but we don't know which ones. To assess this we perform post hoc tests.

There are two tests which are usually run:

- Tukey HSD test (Honest Significant Differences)
- Pairwise t-tests with averaged variances.

Let's use these to check our one-way ANOVA.

ANOVA post hoc tests, continued

The p values of the pairwise t-test end up accumulating "family error". The p.adjust = "bonferroni" argument indicates which algorithm to use to try to fix this error.

The result is a table of p-values for the group-to-group comparisons.

There is a statistically significant difference between A and B and A and X.

```
>
> pairwise.t.test(pain, drug, p.adjust = 'bonferroni')
Pairwise comparisons using t tests with pooled SD

data:  pain and drug

     A         B
  B  0.00119   -
  X  0.00068   1.00000
P value adjustment method:  bonferroni
>
```

Once again, the null hypothesis is that there are no differences between the means.

ANOVA post hoc tests, continued more

Using the Tukey HSD test we get similar group-to-group p-values.

Note that applying Tukey HSD, or paired t-tests, before the ANOVA's null hypothesis has been rejected increases the possibility of incorrectly rejecting the null hypothesis.

```
>
> TukeyHSD(anova.result, conf.level = 0.95)
Tukey multiple comparisons of means
95% family-wise confidence level

Fit:  aov(formula = pain   drug, data = treatment)

$drug
            diff          lwr        upr       p adj
B-A    2.1111111    0.8295028   3.392719   0.0011107
X-A    2.2222222    0.9406139   3.503831   0.0006453
X-B    0.1111111   -1.1704972   1.392719   0.9745173
>
```

ANOVA-like tests

There are other ANOVA-like tests out there.

- Analysis of Covariance (ANCOVA): whereas ANOVA determines differences in group means, ANCOVA determines differences in adjusted means (adjusting for a covariate, a "confounding variable", a third variable which may be affecting the result).
- Multivariate analysis of variance (MANOVA): similar to ANOVA, but with multiple dependent variables.
- Multivariate analysis of covariance (MANCOVA): like ANCOVA, but now with multiple dependent variables.

These tests lie outside of the test decision tree from this class, since they are either multivariate or controlling for a third variable.

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`