Introduction to Computational BioStatistics with R: ANOVA

Erik Spence

SciNet HPC Consortium

21 October 2021

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, "anova.pdf".

`https://scinet.courses/1182`

SciNet compute • calcul CANADA

Today's class

Today we will continue our adventures with hypothesis tests.

- Association tests.
- ANOVA.
- Power analysis.

As always, ask questions.

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.

Association tests

Is there any association between your variables? Are they correlated?

- If your data are numeric and Gaussian:
  - Pearson's r (correlation)
- If your data are numeric and not Gaussian (or unknown):
  - Spearman's (rank correlation coefficient) ρ
  - Kendall's (rank correlation coefficient) τ
- If your data are categorical:
  - If your data are 2 x 2:
    - Relative risk (risk ratio)
    - Odds ratio
  - If your data are not 2 x 2:
    - Chi-square test for trend
    - Logistic regression

We will also visit some related topics when we cover linear models.

SciNet compute • calcul CANADA

Association test, example 1

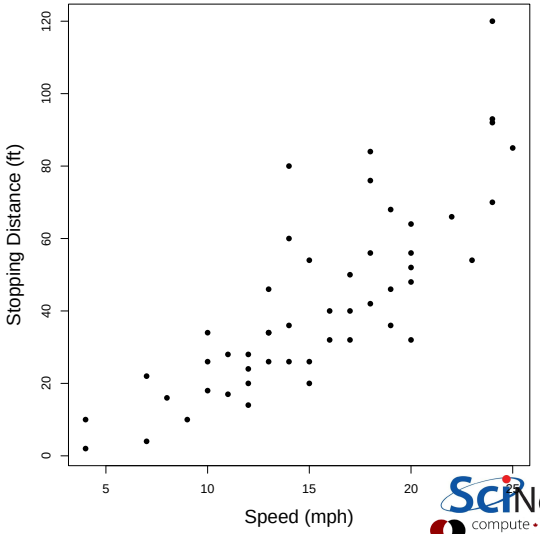
Consider the 'cars' data set, which gives the speed of 1920s cars and the distance they take to stop.

```
> plot(cars$speed, cars$dist)
```

Are the speed and distance linearly associated?

Question 1: are the data numeric or categorical? Numeric.

Question 2: are the data Gaussian? Not sure, let's find out.



Association test, example 1, 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 rejected for the distance data, but not rejected for the speed data.

Question 2: are the data Gaussian? Unlikely. The Pearson correlation test requires both sets of data to be Gaussian.

```
>
> shapiro.test(cars$speed)

     Shapiro-Wilk normality test

data:  cars$speed
W = 0.97765, p-value = 0.4576
>
> shapiro.test(cars$dist)

     Shapiro-Wilk normality test

data:  cars$dist
W = 0.95144, p-value = 0.0391
>
```

SciNet compute • calcul CANADA

Association test, example 1, continued more

Data: numeric, non-Gaussian.

Association test: Spearman's correlation test. This will check to see if there is a significant association between the two data sets.

Null hypothesis: there is no association between the two data sets.

We will use a significance level of 0.05.

Result: null hypothesis rejected!

Correlation coefficient of 0.83!

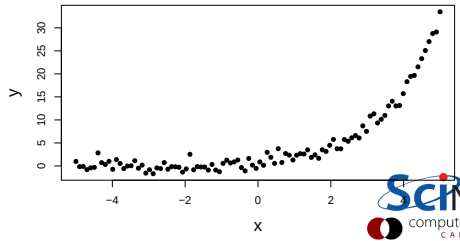
```
>
> cor.test(cars$dist, cars$speed,
+      method = 'spearman')

Spearman's rank correlation rho

data:  cars$dist and cars$speed
S = 252, p-value = 1.464e-05
alternative hypothesis:  true rho is not equal
to 0
sample estimates:
      rho
0.8303568
>
```

Pearson versus Spearman

A note about using correlation functions.

- Pearson correlation:
  - tests for a linear relationship. If the data is correlated, but in a non-linear way, Pearson's won't work correctly.
  - strongly sensitive to outliers.
  - data needs to be normally distributed.
- Spearman correlation:
  - tests for a monotonic relationship.
  - robust to outliers.
  - data does not need to be normally distributed; data can be ordinal.
- Kendall correlation:
  - same assumptions as Spearman's correlation. Less commonly sed.

```
> x <- seq(-5, 5, length = 100)
> y <- 2**x + rnorm(100)
> cor.test(x,y, method = 'pearson')
⋮
0.7761637
> cor.test(x,y, method='spearman')
⋮
0.8631263
```



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.

SciNet compute • calcul CANADA

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.

SciNet compute • calcul CANADA

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 means of 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.

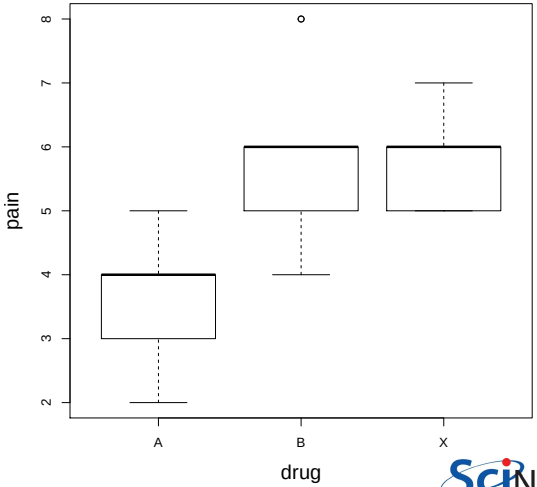
The problem now, as mentioned before, is that we don't know which particular groups are different from which.

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.

SciNet compute • calcul CANADA

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.

SciNet compute • calcul CANADA

Power analysis

The "power" of a binary hypothesis test is the probability that the test will correctly reject the null hypothesis when the alternative hypothesis is true. The "statistical power" ranges from 0 to 1. As the power increases the probability of making a Type II error (beta) decreases.

$$\textbf{power} = P(\textbf{reject } \mathbf{H_0} \,|\mathbf{H_1} \textbf{ is true})$$

Statistical power can also be thought of as the probability of detecting a specific effect when that effect does indeed exist.

Power analysis relates 4 quantities (if you have the other 3, you can calculate the fourth):

- Sample size: n, the number of data points.
- "Effect size": h,
- Significance: usually 0.05,
- Statistical power: usually 0.8.

Effect size?

Just because you've found a statistically significant result doesn't mean that the result itself is important.

- For example, if you have large sample sizes, it can be easy to get a statistically significant difference between means of populations.
- Recall that statistical significance just indicates that it's unlikely that the calculated differences occur randomly.
- But is the "size of the result" important?
- Effect size is a measure of the "amount" $\boldsymbol{H_0}$ is false, or the "amount" two variables relate to each other.
- Effect size is measured in standard deviations, for continuous variables.

Effect size is an important consideration when dealing with the practical application of statistically significant results.

SciNet compute • calcul CANADA

Effect size, continued

The value of effect size that we use in our power calculation depends upon

- the type of test being run,
- how big an effect we desire.

The "cohen.ES" function can be used to return a representative value, based on the test and the relative effect size.

The "ES.h" function is also used to calculate effect sizes for proportion tests.

Test	small	medium	large
tests for proportion (p)	0.2	0.5	0.8
tests for means (t)	0.2	0.5	0.8
chi-squared tests (chisq)	0.1	0.3	0.5
correlation tests (r)	0.1	0.3	0.5
ANOVA (anov)	0.1	0.25	0.4
general linear model (f2)	0.02	0.15	0.35

```
> library(pwr)
>
> cohen.ES(test = "r", size = "medium")
Conventional effect size from Cohen (1982)
test = r
size = medium
effect.size = 0.3
>
```

Power analysis, continued

There are several R packages out there that will calculate statistical power for you. One good one is the "pwr" package.

The package comes with many useful functions for calculating statistical power:

- pwr.t.test: one-sample, two-sample and paired t-tests.
- pwr.t2n.test: two-sample t-test of unequal sample sizes.
- pwr.p.test: proportion test.
- pwr.anova.test: one-way ANOVA.
- pwr.r.test: correlation test.
- pwr.chisq.test: chi-squared test.

These functions work by leaving out the argument for the part of the calculation you are interested in.

Power analysis, example

Suppose we suspect that we have an unfair coin, which lands on heads 75% of the time, instead of the expected 50%.

We decide to run an experiment to test whether or not the coin is fair. We will flip the coin many times, and count the number of heads.

We will then perform a one-sample proportion test to see if the proportion of heads is significantly different from 50%.

The null hypothesis is that it is a fair coin, and that we will get a head 50% of the time. The alternative hypothesis is that the coin is unfair, and that we will get heads more than 50% of the time.

Problem: how many times must we flip the coin to decide that the coin is unfair?

Example stolen from `https://cran.r-project.org/web/packages/pwr/vignettes/pwr-vignette.html`

Power analysis, example, continued

We will use the pwr.p.test function, which calculates the power for a proportion test. The test will calculate whichever argument is not supplied to the function.

The function takes the arguments:

- n: the sample size.
- h: effect size,
- sig.level: significance level,
- power: statistical power,
- alternative: alternative hypothesis type.

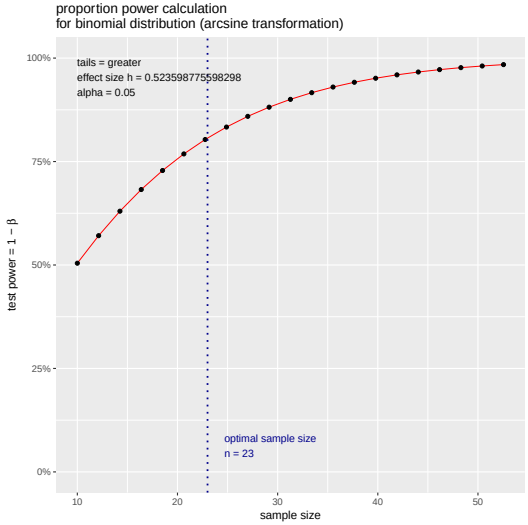
```
>
> coin.power <- pwr.p.test(power = 0.8,
+     h = ES.h(p1 = 0.75, p2 = 0.5),
+     sig.level = 0.05, alternative = "greater")
>
> coin.power
proportion power calculation for binomial
distribution (arcsine transformation)
h = 0.5235988
n = 22.55126
sig.level = 0.05
power = 0.8
alternative = greater
>
```

n = 23!

Power analysis, example, continued more

```
>
> plot(coin.power)
>
```

The output of pwr.p.test includes the ability to plot the power as a function of sample size.



Power analysis, example 2

Suppose we want to check the differences in means between two groups. We will use the two-sample t-test. What is the power of the test if

- there are 30 individuals in each group,
- the significance is 0.05,
- we have an "medium" effect size?

We will use the pwr.t.test function.

```
> cohen.ES(test = "t", size = "medium")
Conventional effect size from Cohen (1982)

test = t
size = medium
effect.size = 0.5
```

```
>
```

```
> pwr.t.test(n = 30, d = 0.5, sig.level = 0.05)
Two-sample t test power calculation

n = 30
d = 0.5
sig.level = 0.05
power = 0.4778965
alternative = two.sided

NOTE: n is number in *each* group
```

```
>
```

Summary

We've now covered many of the categories of tests which are out there.

- Association tests examine whether quantities are correlated with each other.
- Analysis of variance (ANOVA) is used to analyse different groups simultaneously.
- If the ANOVA test is significant, post hoc tests must be performed to determine which groups are different from which.
- Power analyses are performed, without data, to determine the statistical power of a test.
- Power analyses are often done during the design stage of a study, to determine how many subjects are needed.

SciNet compute • calcul CANADA