Statistical Tests

Ingeniería y Ciencia de Datos I

true

Preliminaries

Some concepts we have to manage properly.

Descriptive Statistics

- Central tendency measures: mean, median, mode...
- Statistical dispersion measures: deviations, variance, IQR, range...
- Other measures: skewness, kurtosis...

Random variable and probability

Probability distributions

- Normal or Gaussian distribution: $N(\mu, \sigma)$.
- Binomial distribution: $B(n, p)$.
- Student t, Fisher-Snedecor F, χ^2_n.

Probability theory

- Total probability theorem: $P(A) = \sum_{S_i} P(A|S_i)P(S_i)$.
- Bayes' theorem: $P(B|A) = \dfrac{P(B)P(A|B)}{\sum_{S_i} P(A|S_i)P(S_i)}$

Statistical tests

- It is virtually impossible to work on all the units of the group (**population**) you are interested in.
- It is usual to work on a subset of units (**sample**) taken at random and make inferences from this subset.
- By chance, your random sample may not be very representative of the population. Thus, even two samples taken from two similar populations may differ greatly.
- If there is really a difference between the samples, you need to know what differences you can expect by chance, and how to deal with the variation within samples.

Statistical tests are methods to use samples to make inferences about the populations.

Many statistical tests evaluate a strict null hypothesis H_0 against an alternative hypothesis H_A.

Goal:

- Fit statistical models to data representing the hypothesis that we want to test.
- Use the probability to see when thre results are likely to have happened by chance.

Then with these two tools, we test whether our statistical models (and therefore our hypothesis) fit of the data.

We compute this idea with:

$$\text{test statistic} = \frac{\text{variance explained by the model}}{\text{variance not explained by the model}} = \frac{effect}{error}$$

A lot of test, but all of them represent the same thing: **the amount of variance explained by the model**.

- The bigger the test statistic will be, the more unlikely it is occur by chance.
- If the probability of obtaining the value of our test statistic is less than .05 then we generally accept the experiment hypothesis as true - **there is a significant effect of..** . But take care the effect would not be important.

Hypothesis

- Hypothesis - prediction from your theory.

Hypothesis:

- H_A: Alternative hypothesis (experiment hypothesis). Ej: mean dispersal distance differs between male and female butterflies.
- H_0: Null hypothesis (opposite of H_A). Ej: mean dispersal distance does NOT differ between male and female butterflies

Statistical tests calculate a test statistic from the data to find out how likely the obtained result would be under the null hypothesis.

We can not prove H_A, but we can reject H_0.

Type I and Type II errors

- We use test statistics to tell us about the true state of the world.
- We are trying to check when there is an effect in the population.
- Two possibilities: an effect in the population and no effect.
- No way to know which of these possibilities is true.
- We can compute a test statistic and their associated probability to know which of the two is more likely.

Fisher said: *we must be very sure - 95% of confidence.* But it could occur an error.

Two mistakes:

- Type I: we believe that there is a genuin effect, and in fact, there is not.
- Type II: we believe that there is not a genuin effect, and in fact, there is.

To do so, a probability distribution of the test statistic is theoretically derived assuming the null hypothesis. The probability of the test statistic from the data given that the null hypothesis is true is then found using this theoretical distribution. This probability is termed the **P-value**.

> If the test statistic calculated from the data happens to be a value that is very rare under the null hypothesis, usually occurring at a probability of less than 5%, **(P-value < 0.05), the null hypothesis is discarded** and the alternative hypothesis accepted.

Test result	H_0 true	H_A true	Decision
P-value >= 0.05	Correct	Type II error (false negative)	H_0 accepted; H_A discarded
P-value < 0.05	Type I error (false positive)	Correct	H_0 discarded; H_A accepted

Parametric tests

Many of the statistical procedures are parametric tests based on the normal distribution.

Parametric test:

- requires data from one large catalogue of distributions that statisticians have described
- data to be parametric - certain assumptions must be true
- if you use a parametric test when your data is not parametric, the results are likely to be innacurate

Assumptions of parametric tests:

- Normally distributed data
- Homogeneity of variance: samples comes from populations with the same variance
- Independence: data from different persons are independient

T-tests

> T-tests are used to determine whether the means of two groups (1, and 2) are equal to each other. Assumption: both groups are sampled from normal distributions with equal variances.
>
> H_0 the two means are equal
>
> H_a the two means are not equal

In R, t-tests are calculated using the command **t.test()**.

Classical t-test

The classical t-test considers that the variance of the two groups are equivalent. It is defined:

$$t = \frac{m_1 - m_2}{\sqrt{\frac{S^2}{n_1} + \frac{S^2}{n_2}}}$$

where m_i is the mean of the group i, n_i is the size of the group i and S^2 is an estimator of the pooled variance of the two groups.

$$S = \frac{\sum(x - m_1)^2 + \sum(x - m_2)^2}{n_1 + n_2 - 2}$$

Welch t-statistic

It considers that the variance of the two groups are different and/or unequal sample sizes

In this case S_i is the standard deviation of the two group i.

T-tests in R

```
# independent 2-group t-test
t.test(y ~ x) # where y is numeric and x is a binary factor
# independent 2-group t-test
t.test(y1, y2) # where y1 and y2 are numeric
# paired t-test
t.test(y1, y2, paired = TRUE) # where y1 & y2 are numeric
# one sample t-test
t.test(y, mu = 3)
# mu is a number indicating the true value of the mean
```

Example

Data simulated from a normal distribution

```
# rnorm is a function() that draws random numbers from a normal distribution.
x <- rnorm(10)
y <- rnorm(10)
t.test(x, y)
##
##  Welch Two Sample t-test
##
## data:  x and y
## t = 0.53116, df = 16.441, p-value = 0.6024
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -0.7364004  1.2302359
## sample estimates:
##   mean of x   mean of y
## -0.02522335 -0.27214112
```

Example

Given the following list of heights of persons, let us check whether the average human height is significantly different from 1.77m.

```r
height <- c(1.43, 1.75, 1.85, 1.74, 1.65,
            1.83, 1.91, 1.52, 1.92, 1.83)
t.test(height, mu = 1.77)
##
##  One Sample t-test
##
## data:  height
## t = -0.52046, df = 9, p-value = 0.6153
## alternative hypothesis: true mean is not equal to 1.77
## 95 percent confidence interval:
##  1.625646 1.860354
## sample estimates:
## mean of x
##     1.743
```

Therefore, we cannot reject the hypothesis that average human height is significantly different from 1.77m.

Example

Dispersal distance in male and female butterflies.
- H_0: male butterflies dispersal is similar from female butterflies. - H_A: male butterflies dispersal is different from female butterflies.

```r
distance <- c(3, 5, 5, 4, 5, 3, 1, 2, 2, 3)
sex <- c("male", "male", "male", "male", "male",
         "female", "female", "female", "female",
         "female")
```

Assume equal variances and perform a two-tailed test.

```r
t.test(distance ~ sex, var.equal = TRUE)
##
##  Two Sample t-test
##
## data:  distance by sex
## t = -4.0166, df = 8, p-value = 0.003859
## alternative hypothesis: true difference in means between group female and group male is not equal to
## 95 percent confidence interval:
##  -3.4630505 -0.9369495
## sample estimates:
## mean in group female   mean in group male
##                  2.2                  4.4
```

Thus, male butterflies dispersal is significantly different from female butterflies.

We can also specify a one-sided alternative hypothesis by adding the argument alternative= “less” or alternative = “greater” depending on which tail is to be tested

```r
t.test(distance ~ sex,
       var.equal = TRUE,
       alternative = "greater")
##
```

```
##  Two Sample t-test
##
## data:  distance by sex
## t = -4.0166, df = 8, p-value = 0.9981
## alternative hypothesis: true difference in means between group female and group male is greater than
## 95 percent confidence interval:
##  -3.218516       Inf
## sample estimates:
## mean in group female   mean in group male
##                  2.2                  4.4
```

The results of these tests state that female dispersal distance is not significantly greater than male dispersal distance.

Paired sample t-test

The sleep of students is affected by an exam. You ask 6 students how long they sleep the night before an exam and the night after an exam.

- H_0: There is a significant difference in means.
- H_A: There is not a significant difference in means.

```
sleep.before <- c(4, 2, 7, 4, 3, 2)
sleep.after <- c(5, 1, 3, 6, 2, 1)
# Note the _paired = TRUE_ argument!!
t.test(sleep.before, sleep.after, paired = TRUE)
##
##  Paired t-test
##
## data:  sleep.before and sleep.after
## t = 0.79057, df = 5, p-value = 0.465
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -1.501038  2.834372
## sample estimates:
## mean of the differences
##               0.6666667
```

Thus the data does not support an effect of exams on students sleeping time. **There is not a statistically significant difference in the means.**

Independent samples

```
t.test(y1, y2, paired = FALSE)
#By default, the variances are unequal
t.test(y1, y2, paired = FALSE, var.equal = TRUE)
```

Correlation analysis

To assess the relationship between two variables.

The goal of correlation analysis is to determine how related two variables are. This differs from regression analysis, which seeks to determine a line of best fit from the relationship and assumes that predictor variables is directly affecting (causing) the outcomes of the response variable.

Remember: **Correlation is not causation!**

Pearson Correlation

This test seeks to determine the level of relatedness between two variables using a score that runs from -1 (perfect negative correlation) to 1 (perfect positive correlation). A value of zero indicates no correlation.

It is advisable to test the assumption of normality before running this test.

```
cor.test(iris$Sepal.Length,
         iris$Petal.Length)
##
##  Pearson's product-moment correlation
##
## data:  iris$Sepal.Length and iris$Petal.Length
## t = 21.646, df = 148, p-value < 2.2e-16
## alternative hypothesis: true correlation is not equal to 0
## 95 percent confidence interval:
##  0.8270363 0.9055080
## sample estimates:
##       cor
## 0.8717538
```

This data shows a highly-significant (P-value $<$ 2.2e-16) and strongly positive (0.87) correlation between these two variables. Note that in this case, the P-value is used to reject the null hypothesis that the true correlation is equal to zero.

Spearman Correlation

Spearman's ρ (rho) determines the level of correlation of two variables ranging from -1 to 1. The difference between the two measures is that Spearman uses the *rank-order* of the data rather than the raw values.

```
cor.test(iris$Sepal.Length,
         iris$Petal.Length,
         method = "spearman")
## Warning in cor.test.default(iris$Sepal.Length, iris$Petal.Length, method =
## "spearman"): Cannot compute exact p-value with ties
##
##  Spearman's rank correlation rho
##
## data:  iris$Sepal.Length and iris$Petal.Length
## S = 66429, p-value < 2.2e-16
```

```
## alternative hypothesis: true rho is not equal to 0
## sample estimates:
##       rho
## 0.8818981
```

produced similar, but not identical, results compared with Pearson's R

Cross-tabulation and the χ^2 test

Basic contingency tables would have two categorical variables. In many cases we may wish to test whether the two grouping variables are independent or there is any dependence bewteen the two categorical variables.

Two random variables x and y are called independent if the probability distribution of one variable is not affected by the presence of another.

One of the most common ways to analyze contingency tables is with the χ^2-test (Chi-square test). The χ^2 tests work by first calculating the difference between expected and observed values:

$$\sum \frac{(Observed - Expected)^2}{Expected}$$

In this case, expected values are those that would be obtained in the ideal case where the two variables are independent. If variables are A and B, the independence of A and B occurs if $P(A = a \cap B = b) = P(A = a) \cdot P(B = b)$ for all possible values a and b of the variables.

We propose the following hypothesis:

- H_0: The two variables are independent.
- H_1: The two variables are related.

Example

See Effectiveness of a Drug Treatment

Example

```
flyeyes <- read.csv("3.9.csv", header = T, stringsAsFactors = TRUE)
# Classical method with base R
tab <- table(flyeyes$Eyecolor, flyeyes$Group)
tab
##
##          A  B
##   Red   34 41
##   White 16  9
chisq.test(tab)
##
##  Pearson's Chi-squared test with Yates' continuity correction
##
## data:  tab
## X-squared = 1.92, df = 1, p-value = 0.1659
```

```r
# An alternative
library(lsr)
associationTest(~ Eyecolor + Group, data = flyeyes)
##
##      Chi-square test of categorical association
##
## Variables:   Eyecolor, Group
##
## Hypotheses:
##    null:        variables are independent of one another
##    alternative: some contingency exists between variables
##
## Observed contingency table:
##         Group
## Eyecolor  A  B
##    Red   34 41
##    White 16  9
##
## Expected contingency table under the null hypothesis:
##         Group
## Eyecolor    A    B
##    Red   37.5 37.5
##    White 12.5 12.5
##
## Test results:
##    X-squared statistic:  1.92
##    degrees of freedom:  1
##    p-value:  0.166
##
## Other information:
##    estimated effect size (Cramer's v):  0.139
##    Yates' continuity correction has been applied
```

The p-value indicates that these variables are independent.

The greater the difference between current and expected values, the greater the Chi-Sq value.

References

- Statistical Inference via Data Science
- Learning Statistics with R
- Answering questions with data
- Data Science Live Book