

Statistical Testing I - Classical testing

Daniel Lawson University of Bristol

Lecture 02.2.1 (v1.0.2)

Signposting

- ▶ Last session we covered **regression**.
 - ▶ This is something of a pre-requisite for a useful analysis of **testing**.
- ▶ We'll cover testing in three sections:
 1. Classical testing (recap)
 2. Resampling methods
 3. Model selection

Intended Learning Outcomes

- ▶ ILO2 Be able to **use and apply basic machine learning** tools
- ▶ ILO3 Be able to make and report appropriate inferences from the results of applying basic tools to data

Null hypothesis test

- ▶ Given some data $\{y\}$:
 - ▶ Null Hypothesis **H0**: A statement is true about $\{y\}$.
 - ▶ Alternative Hypothesis **H1**: The statement is not true.
- ▶ We then compute a **test statistic** $T(\{y\})$ whose distribution is **computable under H0**.
 - ▶ By convention, large T is evidence against the null.
- ▶ Then compute p-value $p(T \geq T(\{y\}))$, the probability of observing a test statistic at least as large as that observed given H0 is true.
 - ▶ Example: H0: $\mathbb{E}(y) = \mu$ with $\mu = 0$. H1: $\mu \neq 0$.
 - ▶ This is **not model selection**. We favour H0 and must find evidence against it to accept H1.

Null hypothesis significance testing

- ▶ Hypothesis testing is asking: are my data consistent **with this hypothesis** when **using this measure**?
 - ▶ If you choose a silly hypothesis, testing will dutifully say “no”
 - ▶ If you use a weak measure, testing will dutifully say “yes”
 - ▶ Nothing is learned by this!
- ▶ The correct use of statistical testing is where:
 1. the **null hypothesis might plausibly be true**, or
 2. it might not be true, but you care how much **power the data has to reject the null**

When to use hypothesis testing

- ▶ Some valid use cases include:
 - ▶ To **rank hypotheses** by how much evidence there is against them
 - ▶ To obtain a **standardised scale** (0-1) for combining evidence
 - ▶ When **data are scarce**
- ▶ Also when testing plausible nulls, such as:
 - ▶ **validating simulations** with a known simulator;
 - ▶ **independence** or other non-parametric tests.
 - ▶ **broad null hypotheses**, such as testing a range of parameters.

Types of error

- ▶ The **p-value** defines the probability that H_0 is true, but is rejected.
- ▶ The **power of the test** is the probability that H_0 is false but is accepted anyway.
 - ▶ Low power situations are to be avoided: see e.g. Andrew Gelman's blog¹.
- ▶ Power is a surprisingly important problem because there are many researcher degrees of freedom.
 - ▶ so if power is low, we tend to find significant results anyway, through the (often unintentional) use of the data to choose the test.

¹<https://andrewgelman.com/2018/02/18/low-power-replication-crisis-learned-since-2004-1984-1964/>

Types of error

Error notation

.	H0 true	H0 false
H0 accepted	Correct	Type II error
H0 rejected	Type I error	Correct

Types of error

Error notation

.	H0 true	H0 false
H0 accepted	Correct	Type II error
H0 rejected	Type I error	Correct

Alternative notation

.	H0 true	H0 false
H0 accepted	True Positive	False Positive
H0 rejected	False Negative	True Negative

Example: T-test for a difference in mean

- ▶ To test if the mean of $\{x\}$ is μ_0 , we calculate the **test statistic**:

$$t = \frac{\bar{x} - \mu_0}{\frac{s}{\sqrt{n}}},$$

- ▶ where s is the standard deviation and n the sample size. Under H_0 :

$$t \sim t(t; \nu = n - 1)$$

- ▶ where ν is the degrees of freedom. (See Student's t-distribution).

Example: T-test for a difference in mean

```
## Extract TCP and UDP packet sizes  
tcpsize=conndata[conndata[,"proto"]=="tcp","orig_bytes"]  
udpsize=conndata[conndata[,"proto"]=="udp","orig_bytes"]  
ftpsize=conndata[conndata[,"service"]=="ftp","orig_bytes"]
```

Example: T-test for a difference in mean

```
## Extract TCP and UDP packet sizes
tcpsize=conndata[conndata[,"proto"]=="tcp","orig_bytes"]
udpsize=conndata[conndata[,"proto"]=="udp","orig_bytes"]
ftpsize=conndata[conndata[,"service"]=="ftp","orig_bytes"]

## Convert and omit missing data
tcpsize=as.numeric(tcpsize[tcpsize!="-"])
udpsize=as.numeric(udpsize[udpsize!="-"])
ftpsize=as.numeric(ftpsize[ftpsize!="-"])
tcpsize=tcpsize[tcpsize>0]
udpsize=udpsize[udpsize>0]
ftpsize=ftpsize[ftpsize>0]
```

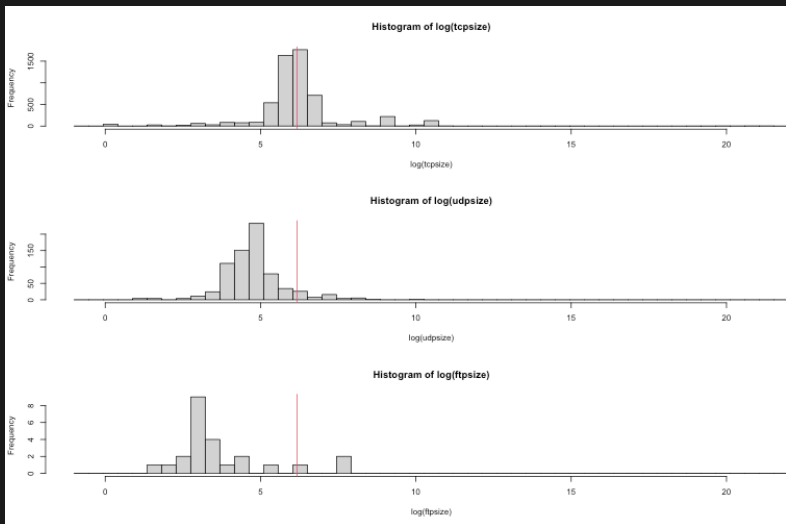
Example: T-test for a difference in mean

```
mu=mean(log(tcpsize))  
t.test(log(udpsize),mu=mu)$p.value  
t.test(log(ftpsize),mu=mu)$p.value
```

Example: T-test for a difference in mean

```
mu=mean(log(tcpsize))  
t.test(log(udpsize),mu=mu)$p.value  
t.test(log(ftpsize),mu=mu)$p.value  
  
> t.test(log(udpsize),mu=mu)$p.value  
[1] 2.733874e-182  
> t.test(log(ftpsize),mu=mu)$p.value  
[1] 8.334782e-08
```

Example: T-test for a difference in mean



t-tests

- ▶ Can be one-tailed (**H0**: $\mu \leq \mu_0$) or two-tailed (**H0**: $\mu = \mu_0$)
- ▶ Assumes:
 - ▶ independence (note: paired tests are possible) and identically distributed
 - ▶ the **data are Normal**
 - ▶ the standard deviation is either known (t is then Normal) or estimated from the data (t is then t distributed).
- ▶ Used in regression, paired tests, etc.
- ▶ *NB Incomplete notes as this is a prerequisite!*

Chi squared test

- ▶ The χ^2 test is for categorical data comparing two variables.
- ▶ **H0**: No relationship between the variables; **H1** Some relationship between them.
- ▶ The **test statistic** for N datapoints from k classes, with x_i observations of type i , with expected value $m_i = Np_i$ where p_i is the expected probabilities, is (under the null):

$$X^2 = \sum_{i=1}^k \frac{(x_i - m_i)^2}{m_i} \sim \chi^2(k - 1)$$

- ▶ This is most often used for **contingency tables** though appears elsewhere.
- ▶ See also **Fishers exact test** for small samples.
- ▶ *NB Incomplete notes as this is a prerequisite!*

Other important tests

- ▶ Nonparametric tests:
 - ▶ **Mann-Whitney U or Wilcoxon rank sum** test: are two samples are drawn from the same distribution? by comparing their ranks.
 - ▶ **Wilcoxon signed-rank** test - as rank sum test, for paired data.
 - ▶ **Kolmogorov-Smirnov** test - are two samples from the same distribution? by comparing the empirical cumulative distribution function.
- ▶ There are many online cookbooks which state exactly which circumstances each test should be used in. You should be able to use them.
- ▶ *NB Incomplete notes as this is a prerequisite!*

Statistical testing overview

- ▶ The tests we have discussed are **classic statistics**, that is, before computers (indeed pre-1950s). The most important types of test for data science are yet to come.

Reflection

- ▶ You must be able to:
 - ▶ Define and use a null hypothesis significance test,
 - ▶ Contrast classical and resampling tests, and judge appropriate uses,
 - ▶ Use statistical testing appropriately in projects.
- ▶ In Science, why does statistical testing have a bad reputation?
- ▶ Does statistical testing have a place in large-scale data science for applied domains?

Signposting

- ▶ Further reading:
 - ▶ Chapter 4 of Statistical Data Analysis by Glen Cowan
 - ▶ Statistical tests, P values, confidence intervals, and power: a guide to misinterpretations by Greenland et al
 - ▶ Andrew Gelman's blog has many examples of statistical testing failures in social science and medicine
- ▶ Next up: Resampling approaches to testing