

# Towards Modern Statistical Testing

Daniel Lawson University of Bristol

Lecture 02.2.2 (v2.0.0)

# Signposting

- ▶ This lecture covers three main topics:
  - ▶ Classical testing (and when its still ok to use it)
  - ▶ Modern testing (and how to use it well)
  - ▶ General Cross Validation (and why you should always do it)

# 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 blog<sup>1</sup>.
- ▶ 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.

---

<sup>1</sup><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

.	$H_0$ true	$H_0$ false
$H_0$ accepted	Correct	Type II error
$H_0$ rejected	Type I error	Correct

- Under the convention that  $H_0 = 0$  = “negative” case and  $H_1 = 1$  = “positive case”:

## Alternative notation

.	$H_0$ holds	$H_1$ holds
$H_0$ accepted	True Negative	False Negative
$H_0$ rejected	False Positive	True Positive



# 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  $\chi^2$  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

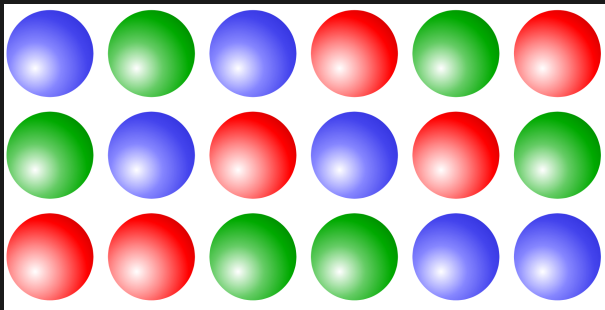
- ▶ Nonparametric tests:
  - ▶ **Mann-Whitney U or Wilcoxon rank sum** test: are two samples drawn from the same distribution? by comparing their ranks.
  - ▶ **Wilcoxon signed-rank** test - as rank sum test, for paired data.
  - ▶ **Kolmogorov-Smirnov** test - are two samples from the same distribution? 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!*

# Resampling

- ▶ The main types of resampling tests include:
  - ▶ **jackknifing**, which is analysing subsets of data to estimate (variance of) parameter estimates
  - ▶ **bootstrapping**, which is resampling with replacement, to estimate (variance of) parameter estimates
  - ▶ **permutation**, which is resampling without replacement, to test a null hypothesis
  - ▶ **cross-validation**, which is analysing subsets of data to estimate out-of-sample prediction, for model performance
- ▶ Each of these methods can be applied to a wide variety of problems, and often requires thought to use appropriately.

# Permutations

All permutations of three colors (each column is a permutation):



- ▶ Figure from Wikipedia<sup>2</sup>. There are in general  $n!$  permutations.

---

<sup>2</sup>[https://upload.wikimedia.org/wikipedia/commons/4/4c/Permutations\\_RGB.svg](https://upload.wikimedia.org/wikipedia/commons/4/4c/Permutations_RGB.svg) / 32

# Generating permutations

```
> set.seed(1)
> n = 5
> x = seq(0,20,length=n)
> x
[1] 0 5 10 15 20
> x[sample.int(n)]
[1] 5 20 15 10 0
> x[sample.int(n)]
[1] 20 15 5 10 0
```

# Use of permutations in testing

- ▶ Consider the following general class of problem:
  - ▶  $H_0$ :  $y$  is independent of  $x$ .
  - ▶  $H_1$ :  $y$  is dependent on  $x$ .
- ▶  $x$  may be continuous, categorical, etc and  $y$  may depend on a number of other things.
- ▶ A **permutation test** will:
  - ▶ resample  $x, y$  pairs **under  $H_0$** ,
  - ▶ Construct a test statistic  $T$ ,
  - ▶ Test if  $T$  extreme in the real data, compared to the permutations?

# Why permutations

- ▶ The main advantage is that the test is asymptotically correct and distribution free. We only (!) have to assume **exchangeability**.
- ▶ Exchangeability of what?
  - ▶ what would be **equal if the null hypothesis is true**, and
  - ▶ would be **different if the alternative hypothesis is true**?
- ▶ It is essential to **maintain any true correlation structure** when performing the test, otherwise the test is not correct.
- ▶ For example, if the indices were originally correlated, permutation will fail.
  - ▶ as from e.g. a time-series.



## Some main types of test (1)

x1	x2	x3	y1	y2
4	12	-3	2	-24

- Permutation of **indices**:

x2	y1	x3	y2	x1
4	12	-3	2	-24

## Some main types of test (1)

x1	x2	x3	y1	y2
4	12	-3	2	-24

- Permutation of **indices**:

x2	y1	x3	y2	x1
4	12	-3	2	-24

- Permutation of **signs**, retaining magnitudes:

x1	x2	x3	y1	y2
4	-12	3	-2	24

## Some main types of test (2)

x1	x2	x3	y1	y2
4	12	-3	2	-24

- ▶ Permutation of **group** labels:

x1	y1	y2	x2	x3
4	12	-3	2	-24

## Some main types of test (2)

x1	x2	x3	y1	y2
4	12	-3	2	-24

- ▶ Permutation of **group** labels:

x1	y1	y2	x2	x3
4	12	-3	2	-24

- ▶ Permutation **within group** labels:

x1	x2	x3	y1	y2
12	-3	4	-24	2

# Monte-Carlo testing

- ▶ There are in general  $n!$  permutations. This is typically too many for  $n > 20$ .
- ▶ We instead choose  $N$  **random permutations** from all the possible ones.
- ▶ Monte-Carlo testing is an important subject in its own right.
- ▶ Its often possible to place guarantees on the  $p$ -value from very few samples.

# Monte-Carlo test

- ▶ To conduct a Monte-Carlo test, we construct  $N$  random datasets and add our real dataset.
- ▶ We then ask, is the **real dataset an outlier** with respect to the random datasets?
- ▶ Specifically, the p-value for a test  $T$  applied to  $X$  (where large values are considered strange) is:

$$\frac{\text{Rank}(T(X); T(\{x_i\}))}{N + 1}$$

- ▶ where Rank simply counts the number of cases as large or larger.

## Heuristics for how many permutations to use

- ▶ The **smallest possible p-value** with  $N$  permutations is  $1/(N + 1)$ . So 999 permutations gives a minimum of 0.001.
- ▶ The **variance** around a chosen threshold, say  $p = 0.05$ , is determined by the sampling distribution of the Binomial:

$$\text{sd}(p) = \text{sd}(\text{Bin}(N, p)) = \sqrt{\frac{p(1 - p)}{n}}$$

- ▶  $p$  is of course the true unknown probability, not the observed one.
- ▶ But variance is an increasing function of  $p$  (for  $p < 0.5$ )
- ▶ A heuristic rule is: to be 95% confident that  $p \leq t$  we need the empirical p-value to be less than  $t - 1.96\text{sd}(p = t)$
- ▶ For  $N = 999$  and  $t = 0.05$ ,  $\text{sd}(p = t) = 0.0135$  and therefore  $p < 0.036$
- ▶ A similar calculation shows  $N = 999$  wouldn't be enough to be sure we were less than 0.005.
- ▶ This is conservative... only if the distribution is Normal...(!)  
**Plot the distribution of  $T$ !**

# Permutation testing summary

- ▶ **Distributional assumptions** are often invalid (regular tests)
- ▶ **Exchangeability assumptions** are often plausible (permutation tests)
- ▶ It is possible to get misleading inference if the assumptions of a test don't hold
- ▶ Permutation tests are really important for generating **plausible null hypotheses**



# Model Selection

- ▶ Imagine that we have run two different inference procedures (models) on our data.
- ▶ We want to decide which of these gives the **best** description of the data.
  - ▶ (For the moment we will pretend we want to know which one is **right**...)
- ▶ Model selection formalises how to make this assessment.

# General considerations

- ▶ To make Cross-Validation work, we need to be able to define our inference goal cleanly. Some scenarios:
  - ▶ **Same source, single datapoint**: Within a single datastream, how well can we predict the **next** point?
  - ▶ **Same source, segment of data**: Within a single datastream, how well could we predict everything that happens within an hour?
  - ▶ **New but understood source**: We have multiple datastreams, each of which might be different but all are generated by a similar process. How well can we predict a new such datasource?
  - ▶ **Unexpected source**: We have many classes of datastream. How well can we predict what would happen on a new class of datastream?

# Problems with LOOCV

- ▶ We might worry that leaving out one datapoint at a time isn't enough:
  - ▶ **Cost**. It is straightforward to apply LOOCV to an arbitrary loss function, including a Likelihood. However, it can be costly.
  - ▶ **Quality**. LOOCV estimates of out-of-sample loss has high variance because each test datapoint using  $n - 2$  of the **same training datapoints**. . .
    - ▶ Empirically, we often choose a different model on different data generated under the same distribution!
  - ▶ **Correlation**. Any correlation breaks LOOCV.

# K-fold CV

- ▶ Naive **k-fold CV** addresses the first issue by creating a **bias-variance tradeoff**: we introduce a bias (towards simpler models) but also significantly reduce the variance of the MSE estimation.
- ▶ More complicated sampling in k-fold settings can also address correlation.
- ▶ **Split** the data into  $k$  “folds”  $f(i)$ , that is, **random non-overlapping samples** of the data of size  $n/k$ . Then:
- ▶ **For each fold  $i$ :**
  - ▶ Call  $X^{-(f(i))}$  the “training” dataset and  $X^{(f(i))}$  the “test” dataset
  - ▶ Learn parameters  $\hat{\theta}_i$  with data  $X^{-(f(i))}$
  - ▶ Evaluate  $l_i = \text{Loss}(X^{(f(i))} | \hat{\theta}_i)$
- ▶ And report  $\frac{1}{n} \sum_{i=1}^k l_i$

# How many folds?

- ▶  $k$ -fold CV loses a fraction of the data, whereas LOOCV only loses a constant.
- ▶ This means that (under the assumption that the **true model is not in the model space**)  $k$ -fold CV will choose a **simpler model** with less predictive power than was possible.
- ▶ However, smaller  $k$  can make the inference more consistent across different data.
- ▶ For **small data**, LOOCV is recommended. For **larger data**,  $k = 10$  is often chosen:
  - ▶ **cost**.  $k$  defines the minimum number of times you need to run the models. If you can afford to run a model once, you can probably afford 10 times.
  - ▶ **practicality**. If you had only 10% more data you might expect to get the same performance as LOOCV. We frequently lose this amount of data to quality control, etc.

# Handling correlation

- ▶ **Correlation** structures can be handled in k-fold CV by **careful sampling**:
  - ▶ a-priori there is a correlation in time or space expected. we can therefore **remove windows**.
  - ▶ the data have some associated covariate, which can be removed en-masse.
  - ▶ empirical correlation structures can be used to select a point  $i$  and all points correlated with it above some **correlation threshold**.
- ▶ Some of these can be used in other contexts. Examples include:
  - ▶ **block bootstrap**
  - ▶ Using a different definition of a “datapoint” in a leave-one-out context, for example: datapoints are countries instead of countries at timepoints

# Reflection

- ▶ You should understand how 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?
- ▶ When are sampling approaches to testing appropriate?
- ▶ What do they test?
- ▶ What are the main ways to implement them?
- ▶ What problems can resampling tests solve? Where are they still difficult to apply?

# Further reading

- ▶ Classical Testing
  - ▶ 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
- ▶ Modern Testing
  - ▶ Cosma Shalizi's Modern Regression Lectures (Lectures 26,28)
  - ▶ Cross Validation and Bootstrap Aggregating on Wikipedia
  - ▶ Chapters 18.7 of The Elements of Statistical Learning: Data Mining, Inference, and Prediction (Friedman, Hastie and Tibshirani).
- ▶ Cross Validation
  - ▶ Chapters 2.3 and 7.10 of The Elements of Statistical Learning: Data Mining, Inference, and Prediction (Friedman, Hastie and Tibshirani).
  - ▶ Cosma Shalizi's Modern Regression Lectures (Lectures 20, 26)