Permutation tests

Christiana Kartsonaki

christiana.kartsonaki@dph.ox.ac.uk

CGAT

February 16, 2017

Hypothesis testing

Null hypothesis: the hypothesis to be tested.

A statistical hypothesis, usually about the frequency distribution of the population of values from which the data are drawn.

e.g. are the distribution of values in two groups the same?

Often denoted by H_0 (and the alternative hypothesis H_1).

Hypothesis testing

choose a property, the test statistic such that when the null hypothesis holds, the probability distribution of the test statistic is known numerically

Hypothesis testing

choose a property, the test statistic such that when the null hypothesis holds, the probability distribution of the test statistic is known numerically

formulated such that the larger the value of the test statistic, the stronger the evidence against the null hypothesis

'small' values suggest that the data are consistent with the null hypothesis

Hypothesis testing

the value of the test statistic is calculated from the data and compared to its expected distribution under the null hypothesis

p-value (observed statistical significance level)

$p = \mathbb{P}(|T| \geq t_{\text{obs}} \mid H_0)$

where T is the test statistic and t_{obs} its observed value

Hypothesis testing

the value of the test statistic is calculated from the data and compared to its expected distribution under the null hypothesis

p-value (observed statistical significance level)

$p = \mathbb{P}(|T| \geq t_{\text{obs}} \mid H_0)$

where T is the test statistic and t_{obs} its observed value

equivalently set a significance level α and calculate the critical value c such that

$\mathbb{P}(|T| > c \mid H_0) \leq \alpha$

we reject H_0 if $|T| > c$

Parametric and non-parametric methods

can be done using parametric and non-parametric methods

Parametric and non-parametric methods

can be done using parametric and non-parametric methods

parametric methods $\rightarrow$ rely on distributional assumptions

usually more interpretable, relationship between estimation and testing

sometimes assumptions based on approximations such as the central limit theorem unrealistic, such as when sample size very small

non-parametric methods $\rightarrow$ no distributional assumptions (but in general not completely assumption-free)

usually do not yield some interpretable measure of effect/association

can be very computationally intensive

may waste some information in the data

Resampling-based methods

may be useful where either standard approximations cannot be used or where their accuracy is suspect

Permutation tests

A **permutation test** calculates the p-value as the proportion of permuted datasets which produce a test statistic at least as extreme as the one observed from the actual data.

no assumptions, but can be infeasible to calculate exactly

Permutation tests

A **permutation test** calculates the p-value as the proportion of permuted datasets which produce a test statistic at least as extreme as the one observed from the actual data.

no assumptions, but can be infeasible to calculate exactly

- calculate the value of the test statistic for the observed data
- calculate the value of the test statistic on all possible permutations of the sample
- p-value = proportion of permutations which yielded a value of the test statistic at least as extreme as the one calculated from the data

Permutation tests

estimate the sampling distribution of the test statistic

can only be used for a null hypothesis of 'no effect'

If the null hypothesis is true the shuffled data sets should look like the real data, otherwise they should look different from the actual data.

Permutation tests

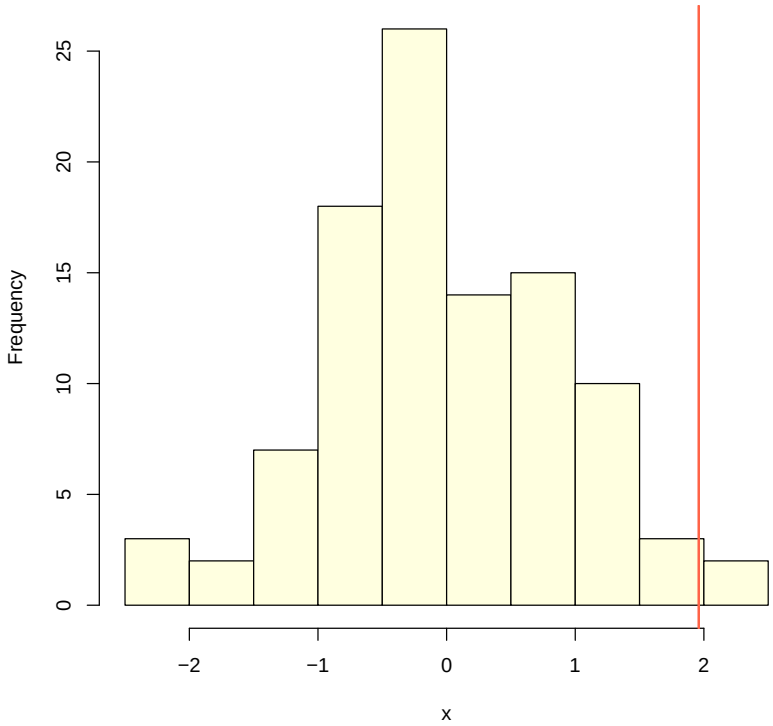

Figure: Empirical distirbution of test statistic. The red line is the value of the observed test statistic.

can be used as a check of the resampled data

Example: group means

Null hypothesis: the distributions of observations from each group are the same

then the group 'labels' are irrelevant

consider a data frame with outcome and group

permute the group labels

Example: group means

Null hypothesis: the distributions of observations from each group are the same

then the group 'labels' are irrelevant

consider a data frame with outcome and group

permute the group labels

y	z
2	5
3	1
1	4
5	4

Example: group means

Null hypothesis: the distributions of observations from each group are the same

then the group 'labels' are irrelevant

consider a data frame with outcome and group

permute the group labels

y	z
2	5
3	1
1	4
5	4

value	group
2	y
3	y
1	y
5	y
5	z
1	z
4	z
4	z

Example: group means

several possible test statistics

e.g.

- means
- geometric means
- t statistic

Example: group means

several possible test statistics

e.g.

- means
- geometric means
- t statistic

re-calculate for each permutation

calculate the percentage of simulations where the simulated statistic was more extreme than the observed $\rightarrow$ p-value

Example: GWAS

can be used as an alternative to multiple-testing correction (e.g. Bonferroni correction) if the tests are thought to not be independent

- the phenotypes are randomly shuffled and all m tests are recalculated on the shuffled datasets $\rightarrow$ repeat many times to construct empirical frequency distribution
- for each permutation, the smallest p-value of the m tests is recorded
- procedure repeated may times $\rightarrow$ empirical frequency distribution of the smallest p-value
- empirical adjusted p-value $= (r+1)/(n+1)$, where n is the number of permutations and r is the number of p-values that are equal to or greater than the p-value from the actual data

distribution of the most extreme of all test statistics

Example: GWAS

set of SNPs that each have some effect on an outcome

want to test for interactions (epistasis)

Permuting the genotype data would break the links between genotype and outcome and created shuffled data with no main effects of SNPs.

Even if there are no interactions the shuffled data will look different from the real data.

Randomization tests

Statistical inference not based on a probabilistic model of the underlying data-generating process

permutation tests numerically equivalent to randomization tests but conceptually different

permutation tests $\rightarrow$ based on some symmetries induced by the probabilistic model

assumed independence and identical distributional form of the random variability

randomization tests $\rightarrow$ randomization used in allocating the treatments; no assumption about the stochastic variability of the individual units

by-product of the procedure used in design

Sign test

test of location zero

non-parametric alternative to the one-sample t-test or the t-test for paired data

pairs of data $(x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n)$

assume differences $X_i - Y_i$ are independent and identically distributed

null hypothesis: median $= 0$

test statistic: number of values greater than 0

Under the null hypothesis, positive and negative differences are equally likely, so the number of positive values follows a binomial distribution with parameters n and 0.5.

Wilcoxon signed rank test

pairs of data $(x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n)$

assume differences $X_i - Y_i$ are independent and identically distributed

null hypothesis: median $= 0$

test statistic: sum of the ranks for the differences with positive sign

large absolute values of the test statistic suggest departure from null

p-value can be calculated exactly for small samples using the permutation distribution (if there are no ties)

for large samples a normal approximation to the sampling distribution can be used

Wilcoxon signed rank test

1. Calculate paired differences
2. Calculate absolute values of differences
3. Rank the absolute values, discarding 0s
4. Multiply ranks by the sign of the difference
5. Calculate the rank sum of the positive ranks

for small sample sizes the rank sum has an exact distribution under the null

Wilcoxon rank sum (Mann–Whitney) test

A permutation test on the ranks rather than the observations themselves.

non-parametric alternative to the two-sample t-test

location shifts between two independent samples

If the samples are of size n_1 and n_2 respectively, then the test statistic is the sum of the ranks of the observations from the first sample minus $n_1(n_1 + 1)/2$.

For small sample sizes the rank sum has an exact distribution under the null.

Wilcoxon rank sum (Mann–Whitney) test

under H_0 the two groups are exchangeable

therefore any allocation of the ranks between the two groups is equally likely

two samples of size n_1 and n_2

1. Rank the observations $1, \ldots, n_1 + n_2$
2. Permute the ranks (if there are ties, the rank of the tied observations is the average of the ranks of the tied observations)
3. Take the first n_1 and assign them to group 1 and the remaining n_2 to group 2
4. Calculate the test statistic
5. Repeat 1–4
6. p-value = proportion of times the test statistic is more extreme than the observed value

Fisher's exact test

contingency table

		Y		
		0	1	Row total
X	0	a	b	$a+b$
	1	c	d	$c+d$
Column total		$a+c$	$b+d$	$a+b+c+d(=n)$

Fisher's exact test

contingency table

		Y		
		0	1	Row total
X	0	a	b	$a+b$
	1	c	d	$c+d$
Column total		$a+c$	$b+d$	$a+b+c+d(=n)$

under a null hypothesis of independence of rows and columns – hypergeometric distribution of the numbers in the cells of the table (conditionally on margin totals)

$$p = \frac{\binom{a+b}{a}\binom{c+d}{c}}{\binom{n}{a+c}} = \frac{(a+b)!(c+d)!(a+c)!(b+d)!}{a!b!c!d!n!}$$

Jackknife

Jacknife a resampling-based method

each observation is deleted in turn and an estimate is calculated based on the remaining $n - 1$ of them

this set of estimates is then used for estimating quantities like bias and variance

useful for quantities that may not be unbiased or have known variance

Jackknife

set of data $x_1, \ldots, x_n$

estimate a parameter θ

$\hat{\theta}$: estimate based on the full data set

$\hat{\theta}_{-i}$: estimate of θ obtained by deleting observation i

$\bar{\theta} = \frac{1}{n} \sum_{i=1}^{n} \hat{\theta}_{-i}$

Jackknife

Jackknife estimate of bias

$$(n-1)(\bar{\theta}-\hat{\theta})$$

Jackknife estimate of standard error

$$\left\{\frac{n-1}{n}\sum_{i=1}^{n}(\hat{\theta}_i-\bar{\theta})^2\right\}^{1/2}$$

Bootstrap

Bootstrap (Efron and Tibshirani, 1993; Davison and Hinkley, 1997)

using repeated sampling with replacement from the data to approximate the sampling distribution of a parameter

distribution-free method

many variations

confidence intervals, tests, ...

Bootstrap

$\hat{\theta} = T(\mathbf{x})$ symmetric function of the sample (does not depend on the sample order)

take m samples from $\mathbf{x}$ with replacement and calculate $\hat{\theta}^*$ for these samples

the new samples consist of an integer number of copies of each of the original data points and so will have ties

Bootstrap

$\hat{\theta} = T(\mathbf{x})$ symmetric function of the sample (does not depend on the sample order)

take m samples from $\mathbf{x}$ with replacement and calculate $\hat{\theta}^*$ for these samples

the new samples consist of an integer number of copies of each of the original data points and so will have ties

assess the variability of $\hat{\theta}$ about the unknown true θ by the variability of $\hat{\theta}^*$ about $\hat{\theta}$

bias of $\hat{\theta}$: mean of $\hat{\theta}^* - \hat{\theta}$

commonly used when the distribution of θ cannot be found analytically

In R

```
?sample
```

```
library(boot)
library(bootstrap)
```

When using simulation-based procedures, always make them reproducible by setting the 'seed' for the random number generation.

```
set.seed(5)
```

Discussion

- permutation tests useful when parametric tests unavailable or assumptions implausible
- some assumptions still implied, e.g. independence
- usually give similar results as corresponding parametric tests
- no simple direct relationship with estimation
- more resistant to extreme observations than parametric tests

References

- Cox, D. R. (2006). *Principles of Statistical Inference.* Cambridge University Press.
- Cox, D. R. (2016). Statistical significance tests. *Diagnostic Histopathology* **22** (7), 243–245.
- Davison, A. C. and Hinkley, D. V. (1997). *Bootstrap Methods and Their Application.* Cambridge University Press.
- Efron, B. and Tibshirani, R. J. (1994). *An Introduction to the Bootstrap.* Chapman and Hall / CRC.
- Venables, W. N. and Ripley, B. D. (2002). *Modern Applied Statistics with S.* Springer. Section 5.7