Lecture 5: Permutation methods and nonparametric methods

EMAT30007 Applied Statistics

Nikolai Bode & Filippo Simini

Department of Engineering Mathematics

Outline of the lecture

In this lecture you will learn:

- How to use the Kolmogorov-Smirnov test to test if a population has a given distribution.
- How to use the Kolmogorov-Smirnov test to test if two populations have the same distribution.
- How to use permutations to test if two populations have the same mean.
- How to use permutations tests for matched pairs, for a relationship between two variables, and when we need to control for certain variables.

Rationale of nonparametric methods

Nonparametric tests are general methods that don't make any assumption on the underlying distribution of the specific variable of interest.

On the contrary, the z-test and t-test are parametric tests because they are derived under the assumption that data comes from a population that is Normally distributed.

Nonparametric hypothesis testing works exactly as parametric hypothesis testing, except for the fact that the sampling distribution of the test statistic does not depend on the population distribution of the variable of interest.

In general, they are useful when the sample size is small and the underlying population distribution is unknown (not Normal).

Kolmogorov-Smirnov test

The Kolmogorov-Smirnov tests are nonparametric tests that are useful when we want to compare distributions:

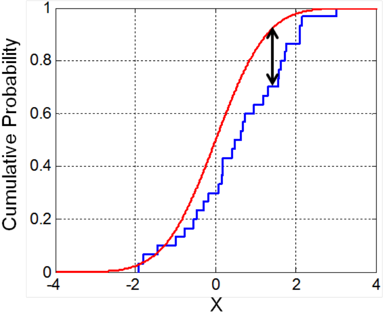
- K-S test for one sample: test if a data sample is compatible with a reference theoretical probability distribution.
- K-S test for two samples: test if two data samples come from populations with the same distribution.

The idea behind K-S tests is to measure the distance between the distributions and estimate the probability to observe a larger distance under the null hypothesis that the distributions are the same.

Kolmogorov-Smirnov test for one sample

Hypothesis: the sample comes from a given continuous distribution with CDF F.

- Test statistic:
  $D = \sup_x |F_e(x) - F(x)|$
  where F_e is the empirical CDF of the sample of size n, is called Kolmogorov–Smirnov statistic and is the supremum (greatest) distance between the empirical and theoretical CDFs.
- Null hypothesis, H_0: D is not significantly larger than zero.

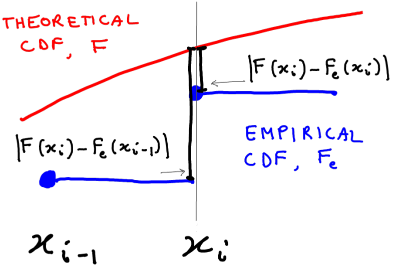


https://en.wikipedia.org/wiki/Kolmogorov%E2%80%93Smirnov_test

- For large samples, reject H_0 at significance level α if $D\sqrt{n} > K_\alpha$ where K_α is the α-quantile of the Kolmogorov distribution (Matlab code here).

Notes about the K-S test

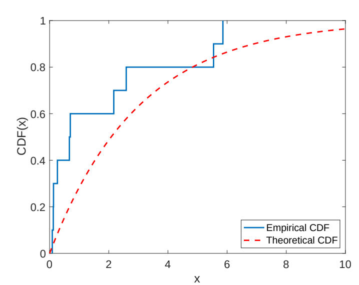
- In practice, to compute the supremum (greatest) distance between the empirical and theoretical CDFs, $D = \sup_x |F_e(x) - F(x)|$, we only have to compute the distances at the data points $x_1, ..., x_n$ between the theoretical CDF at x_i and the empirical CDF at x_i and x_{i-1}:
$$D = \max\{|F(x_i) - F_e(x_i)|, |F(x_i) - F_e(x_{i-1})| \text{ for } i = 1, ..., n\}.$$



- A more accurate condition for rejecting H_0 when the sample size n is small is $(D\sqrt{n} + \frac{1}{6\sqrt{n}} + \frac{D\sqrt{n}-1}{4n}) > K_\alpha$

Example: One-sample K-S test for an Exponential distribution

- A random sample of size 10 is drawn from an Exponential distribution with parameter $\mu = 3$: $x =$ [5.542, 0.089, 0.131, 2.168, 0.668, 5.858, 2.589, 0.264, 0.698, 0.124]
- The null hypothesis is that the sample is drawn from an Exponential distribution with parameter $\mu = 1$.



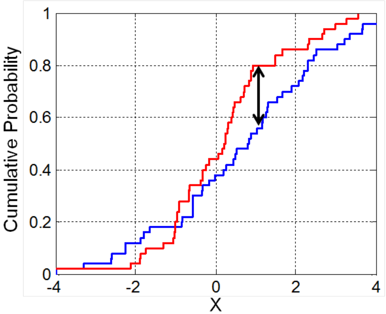
- The K-S test statistic is:
  $D = \max\{|F(x_i) - F_e(x_i)|, |F(x_i) - F_e(x_{i-1})|$ for $i = 1, ..., n\} = 0.39$
- The p-value is $p = 1 - F_K(D\sqrt{n}) = 0.09$, where F_K is the CDF of the Kolmogorov distribution. The p-value is the probability to observe a bigger value of the test statistic D if the null hypothesis is true.
- A more accurate p-value is computed using the formula
  $p = 1 - F_K(D\sqrt{n} + \frac{1}{6\sqrt{n}} + \frac{D\sqrt{n}-1}{4n}) = 0.068$

Kolmogorov-Smirnov test for two samples

Hypothesis: the samples come from the same probability distribution.

- Two-sample K-S test statistic: $D = \sup_x |F_1(x) - F_2(x)|$ where F_1 and F_2 are the empirical CDFs of the two samples of sizes n and m, respectively.
- Null hypothesis, H_0: D is not significantly larger than zero.
- For large samples, reject H_0 at significance level α if

$$D\sqrt{\frac{nm}{n+m}} > \sqrt{-\frac{1}{2}\ln\frac{\alpha}{2}} = K_\alpha$$



https://en.wikipedia.org/wiki/Kolmogorov%E2%80%93Smirnov_test

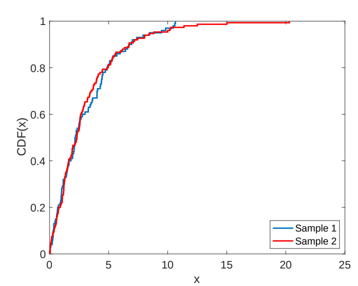
The two-sample K-S test is reliable only for large samples.

Example: Two-sample K-S test

- Two random samples x1 and x2 of sizes 100 and 150 are drawn from an Exponential distribution with parameter $\mu = 3$.
- The null hypothesis is that the samples have the same distribution.
- We compute the K-S test statistic and the p-value using Matlab's function `kstest2`:

```
[h,p,D] = kstest2(x1, x2)
```

  We get $p = 0.78$ and $D = 0.08$.



- The p-value can be computed with the formula $p = 1 - F_K(D\sqrt{\frac{nm}{n+m}}) = 0.80$. This formula for the two-sample K-S p-value is less accurate than Matlab's formula, but the difference between the two tends to zero for large sample sizes.

Resampling and permutation tests

Permutation or resampling tests work exactly as parametric and nonparametric tests except for the fact that the sampling distribution of the test statistic is empirically obained by resampling in a manner that is consistent with the null hypothesis, without relying on theoretical formulae.

A permutation test consists in *scrambling, shuffling or sampling without replacement* from the original dataset in a way that would not produce any difference in the distribution of the test statistic if the null hypothesis is true.

To carry out a permutation test:

1. Compute the test statistic for the original data.
2. Perform a permutation test sampling *without replacement* from the data in a way that is consistent with the null hypothesis.
3. Estimate the sampling distribution constructing the permutation distribution of the test statistic for a large number of resamples.
4. Find the p-value of the original test statistic on the permutation distribution.

Permutation test for equal means of two populations

A streaming media service wants to test if the new version of its website is more engaging than the old one. Two groups of random users are selected:

- the minutes spent on the old version of the webiste by the users in group 1 are: $11, 9, 16, 15, 34$.
- the minutes spent on the new version of the webiste by the users in group 2 are: $14, 7, 21, 12, 42$.

Permutation test:

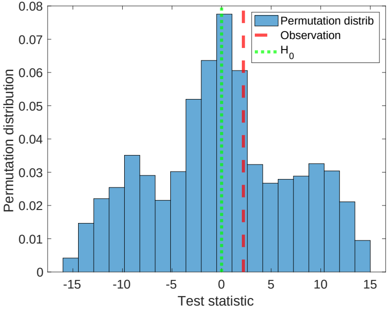
- The null hypothesis is that the average time spent on the old and new versions is the same. The test statistic is the difference between the average time of group 2 minus the average time of group 1.
- If the null hypothesis is true, the difference between the times of the two groups is zero, and any deviation from zero is due to chance resulting from the random assignment of users to the two groups.
- A permutation test compatible with this null hypothesis consists in reassigning users to the two groups randomly.

Permutation test for equal means of two populations

resamples	group 1	group 2	test statistic
#1	34, 16, 12, 9, 7	11, 14, 42, 21, 15	5.0
#2	9, 14, 21, 42, 34	16, 7, 12, 11, 15	-11.8
#3	21, 15, 12, 11, 42	9, 14, 34, 7, 16	-4.2
#4	7, 9, 42, 11, 21	12, 34, 15, 14, 16	0.2

Permutation distribution after resampling 10000 times

Right-tailed p-value = 0.37
It is the fraction of resamples with test statistic (mean difference) higher than the observed value 2.2.



Permutation test for matched pairs

A streaming media service wants to test if the new version of its website is more engaging than the old one. A group of 5 random users is selected and for each user the minutes spent on the old and new versions of the website have been recorded:

- The null hypothesis is that the average time spent by each user on the old and new versions is the same. The test statistic is the average difference between the time spent on the new and old versions.
- If the null hypothesis is true, the difference between the times is zero and any difference is due to chance.
- A permutation test compatible with this null hypothesis consists in randomly shuffling each user's old and new times.

	old	new
user 1	11	14
user 2	9	7
user 3	16	21
user 4	15	12
user 5	34	42
test statistic: 2.2		

Permutation test for matched pairs

#1	old	new
user 1	11	14
user 2	7	9
user 3	16	21
user 4	15	12
user 5	34	42
test statistic: 3		

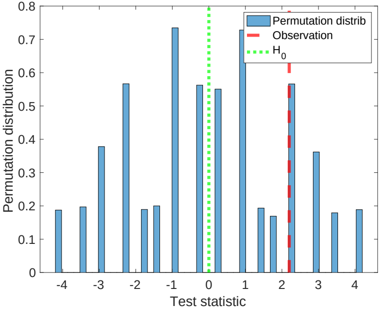
#2	old	new
user 1	14	11
user 2	9	7
user 3	21	16
user 4	12	15
user 5	42	34
test statistic: -3		

#3	old	new
user 1	14	11
user 2	7	9
user 3	21	16
user 4	15	12
user 5	34	42
test statistic: -0.2		

. . .

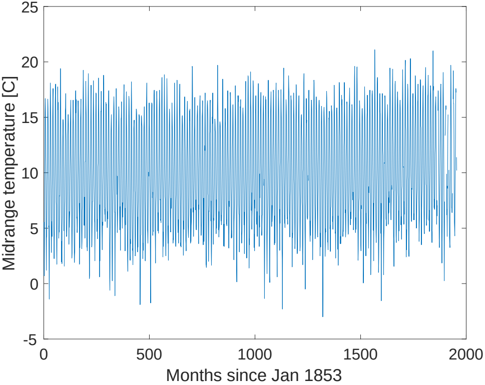
Permutation distribution after resampling 10000 times

Right-tailed p-value = 0.21
It is the fraction of resamples with test statistic higher than the observed value 2.2.



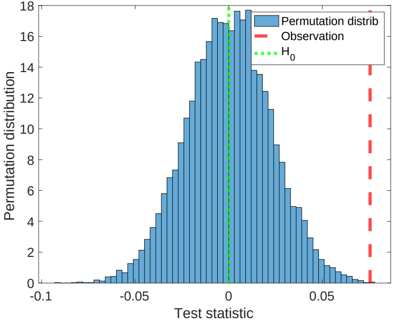
Permutation test for a relationship

We want to establish if temperature has been increasing over the years. To this end, we consider the monthly time series of midrange temperatures, that is the arithmetic mean of the maximum and minimum temperature in each month, recorded in Oxford since 1853.



Permutation test for a relationship

- The null hypothesis is that temperature and time (months since January 1853) are not correlated.
- The test statistic is the Pearson correlation coefficient between the two variables.
- If the null hypothesis is true, the correlation coefficient is zero and any difference is due to chance.
- A permutation test compatible with this null hypothesis consists in randomly shuffling one of the two variables, for example the time.



Permutation distribution after resampling 10000 times

Right-tailed p-value = 0.0002

Constrained permutations

We want to test the hypothesis that male and female students are equally likely to pass the summer exam. Last year's data is reported in the table below:

	Male	Female	totals
Pass	18	17	35
Fail	12	3	15
totals	30	20	50

- The null hypothesis is that the probability to pass the exam does not depend on the student's gender, that is, the proportion of female students failing the exam is equal the proportion of male students failing the exam.
- The test statistic is the number of female students failing the exam, given that we fix the totals of male and female students and the totals of pass and fail.
- A permutation test compatible with this null hypothesis consists in randomly reassigning the exam outcomes (35 pass and 15 fail) to students, irrespective of their gender.

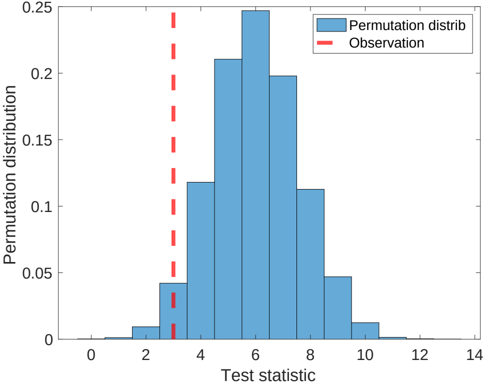
Constrained permutations

	Male	Female	totals
Pass	22	13	35
Fail	8	7	15
totals	30	20	50

	Male	Female	totals
Pass	16	19	35
Fail	14	1	15
totals	30	20	50

	Male	Female	totals
Pass	15	16	35
Fail	11	4	15
totals	30	20	50

· · ·



Permutation distribution with 10000 resamples

Left-tailed p-value 0.058

When Can We Use Permutation Tests?

We can use a permutation test only when we can see how to resample in a way that is consistent with the null hypothesis.

We applied permutation tests for the following types of problems:

- Two-sample problems when the null hypothesis says that the two populations are identical. We may wish to compare population means, proportions, standard deviations, or other statistics.
- Matched pairs designs when the null hypothesis says that there are only random differences within pairs.
- Relationships (e.g. correlation) between two quantitative variables when the null hypothesis says that the variables are not related.

Permutation tests can't be used for testing hypotheses about a single population or comparing populations that differ even under the null hypothesis.

Pros and cons of permutation tests

- Permutation tests work even when populations are not normally distributed.
- The tradeoff is that they require the two populations to have identical distributions when the null hypothesis is true – not only the same means, but also the same spreads and shapes. We need this to be able to move observations randomly between groups.
- They allow greater flexibility in the choice of test statistic.
- Constrained permutations allow to control for certain variables.