Experimental Methods in Computer Science

(*Metodologias Experimentais em Informática)*

Henrique Madeira

Master in Informatics Engineering
Departamento de Engenharia Informática
Faculdade de Ciências e Tecnologia da Universidade de Coimbra
2023/2024

1

Hypothesis Testing

2

Non-parametric statistical inference

3

Non-parametric hypothesis testing

- There are situations for which is not possible to apply parametric statistics:
  - Data have a ranking but no clear numerical interpretation, such as user preferences;
  - You are interested on a parameter of the population for which the distribution is unknown (medians, variances, percentiles, etc.).
- Pos: Non-parametric methods make fewer assumptions than parametric; they are distribution free
- Cons: In cases where a parametric test would be appropriate, non-parametric tests have less power.

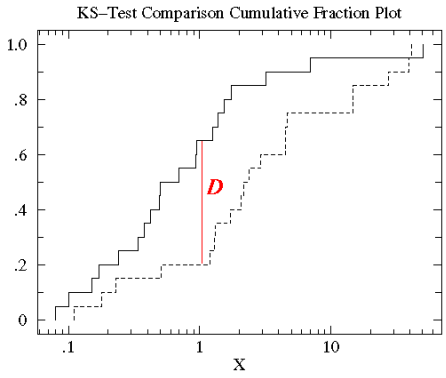
4

Kolmogorov-Smirnov test (two sample)

K-S test compares the two empirical distribution functions. The test statistic is:

$$D_{(\max)} = supx\left|F_{1,n_1}(x) - F_{2,n_2}(x)\right|$$

where F_{1,n_1} and F_{2,n_2} are the empirical distribution functions and *supx* is the supremum function. The values of the test statistic are tabulated.



In practice, K-S test assesses the significance of the **maximum divergence D** between two cumulative frequency curves. If D is larger than a **critical values** for a given **α**, then the differences between the two functions are significant.

5

Critical values for the two-sample Kolmogorov-Smirnov test (2-sided)

The table gives critical *D*-values for α = 0.05 (upper value) and α = 0.01 (lower value) for various sample sizes. The symbol * means you cannot reject H_0 regardless of observed *D*.

$n_2 \backslash n_1$	3	4	5	6	7	8	9	10	11	12
1	*	*	*	*	*	*	*	*	*	*
	*	*	*	*	*	*	*	*	*	*
2	*	*	*	*	*	16/16	18/18	20/20	22/22	24/24
	*	*	*	*	*	*	*	*	*	*
3	*	*	15/15	18/18	21/21	21/24	24/27	27/30	30/33	30/36
	*	*	*	*	*	24/24	27/27	30/30	33/33	36/36
4		16/16	20/20	20/24	24/28	28/32	28/36	30/40	33/44	36/48
		*	*	24/24	28/28	32/32	32/36	36/40	40/44	44/48
5			*	24/30	30/35	30/40	35/45	40/50	39/55	43/60
			*	30/30	35/35	35/40	40/45	45/50	45/55	50/60
6				30/36	30/42	34/48	39/54	40/60	43/66	48/72
				36/36	36/42	40/48	45/54	48/60	54/66	60/72
7					42/49	40/56	42/63	46/70	48/77	53/84
					42/49	48/56	49/63	53/70	59/77	60/84
8						48/64	46/72	48/80	53/88	60/96
						56/64	55/72	60/80	64/88	68/96
9							54/81	53/90	59/99	63/108
							63/81	70/90	70/99	75/108
10								70/100	60/110	66/120
								80/100	77/110	80/120
11									77/121	72/132
									88/121	86/132
12										96/144
										84/144

For relatively larger sample sizes, the approximate **critical value D_α** is given by the equation

$$D_\alpha = c(\alpha)\sqrt{\frac{n_1 + n_2}{n_1 n_2}}$$

The coefficient ***c(α)*** for typical values of **α** is:

α	0.10	0.05	0.025	0.01	0.005	0.001
c(α)	1.22	1.36	1.48	1.63	1.73	1.95

6

Example of K-S test

A large SW development company is analyzing the failures in SW projects in the last 3 years. The company considers that a given SW project fails if the project results in financial losses. The goal is to evaluate whether the project failures in the company branches of Coimbra and Lisbon are significantly different (95% significance) or not. If there is a difference between Coimbra and Lisbon branches, then the company will reassess the SW engineering practices used in the branch with higher failure rate.

Project type	Coimbra Branch	Lisbon Branch
Tiny	5	9
Small	8	12
Medium	7	8
Large	4	14
Very large	7	5
Huge	4	8
Total	**35**	**56**

7

Example: Kolmogorov-Smirnov test
Step 1 - State the hypothesis be tested

- **H_0: $Distrib_1 = Distrib_2$**
  The two cumulative frequency distributions are similar
  (i.e., the project failure rate are the same in the two branches of the SW development company)

- **H_1: $Distrib_1 \neq Distrib_2$**
  The two cumulative frequency distributions are different
  (i.e., the project failure rate are different; the company branch with higher percentage of project failure rates are worst)

8

Example: Kolmogorov-Smirnov test
Step 2 - Select the criteria for a decision

- For a significance of 95% → α = 0.05
- The K-S critical value for relatively large values of n_1 and n_2 is (approximately) given by the formula:

$$D_{\alpha} = c(\alpha)\sqrt{\frac{n_1 + n_2}{n_1 n_2}} = 1.36\sqrt{\frac{35 + 56}{35 \times 56}} = 0.29$$

Coefficient ***c(α)*** for **α** = 0.05

Critical value
(If the difference between the two cumulative distributions is higher than this → reject H_0)

9

Example: Kolmogorov-Smirnov test
Step 3 - Compute the test statistic

The test statistic is:

$$D_{(\max)} = supx\left|F_{1,n_1}(x) - F_{2,n_2}(x)\right|$$

This is the maximum difference in absolute value

Projects type	Coimbra Branch	Lisbon Branch	Cum% Coimbra	Cum% Lisbon	Diff abs. value
Tiny	5	9	0,143	0,161	0,018
Small	8	12	0,371	0,375	0,004
Medium	7	8	0,571	0,518	0,054
Large	4	14	0,686	0,768	0,082
Very large	7	5	0,886	0,857	0,029
Huge	4	8	1,000	1,000	0,000
Total	**35**	**56**			

Note the classes of project types are ordered

10

Example: Kolmogorov-Smirnov test
Step 4 - Make a decision

Critical value;

$D_{\alpha\,(critic.)} = 0.29$

Test statistic:

$D_{(max)} = 0.082$

$D_{(max)}$ is lower than the critical value $D_{\alpha\,(critic.)}$

Conclusion: We cannot reject H_0

We cannot say that the distributions of failed SW projects in the two branches of the company are different, with 95% of confidence.

11

Test normality assumption

- Normality of the data is required for t-test. You may need to test whether your data follows a normal distribution.
- Graphical methods for testing normality include the histogram and qq-plot.
- Non-parametric tests such as Kolmogorov-Smirnov and Shapiro-Walk can also be used.

12

Bootstrapping

- Estimates the precision of sample statistics by drawing randomly with replacement from a set of data points. It is used to estimate the standard deviation and computing CI.
- Procedure:

1. Resample the data with replacement, such that the size of the resample is equal to the size of the original data set.
2. Compute the statistic from the resample from the first step.
3. Repeat step 1 and 2 many times (1000 to 10000) to get a more precise estimate of the bootstrap distribution of the statistic.

13

Bootstrapping

- The bootstrap distribution approximates the sampling distribution of the statistic. This statistic can be other than the mean (median, variance, IQR, correlation, etc.)
- Bootstrap distributions usually have approximately the same shape and spread as the sampling distribution but are not centered at the statistic (from the original sample).
- It is also known as a *computer-intensive statistical method*.

14

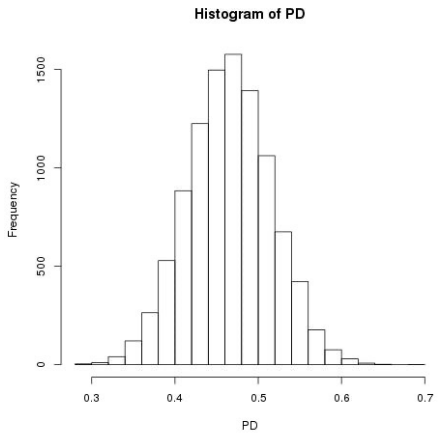
Bootstrapping: Example

- Approximate the distribution of sample proportion for the number of heads found in 100 coin flips.
- Generate a sample of size 100 ($x_1, x_2, \ldots, x_{100}$)
- Randomly resample the observations with replacement and compute the proportion of heads for each sample.
- Resample several times (Monte Carlo) to obtain an empirical bootstrap distribution of the sample mean.

15

Bootstrapping: Example

```
A =
sample(c(0,1),100,replace=TRUE)
PD = c()
for (i in 1:10000){
            S =
sample(A,replace=TRUE)
            p = sum(S)/length(S)
            PD = c(PD,p)
}
hist(PD, freq=NULL,breaks=20)
```



16

Randomization tests

- Statistical test in which the distribution of the test statistic under H_0 is obtained by calculating all possible values of the test statistic under rearrangements of the labels on the observed data points. Is very often used as a two-sample statistical test.
- Procedure: Similar to bootstrap but the resample is constructed without replacement.
- Assumption: the labels are exchangeable under H_0. Sufficient condition for exchangeability: variables are i.i.d.
- See Section 5.3. of Cohen's book.

17

Example - Hypothesis testing using randomization tests (two independent samples)

Assume you are the database administrator of a big information system. The database has just been installed and you are trying two tuning configurations: Conf. **A** and Conf. **B**.

You use a given SQL package to test the execution time for each configuration.

After running several times the SQL package in both configurations you want to take a decision.

Question: **what is the best configuration?**

Conf. A exec. time	Conf. B exec. time
74	69
66	71
88	80
68	88
79	64
68	65
87	74
79	76
78	89
72	68
86	67
85	72
86	

$\mu_1 = 78.15$ $\mu_2 = 73.58$
$s_1 = 7.94$ $s_2 = 8.33$
$n = 13$ $n = 12$

18

Example - Hypothesis testing using randomization tests (two independent samples)

- Randomly reassign the observations between the two groups and compute the mean.
- Resample and compute the mean several times (Monte Carlo) in order to obtain an empirical distribution of the sample mean.
- Compute the *p-value* for a given significance value based on the empirical distribution of the sample mean.

19

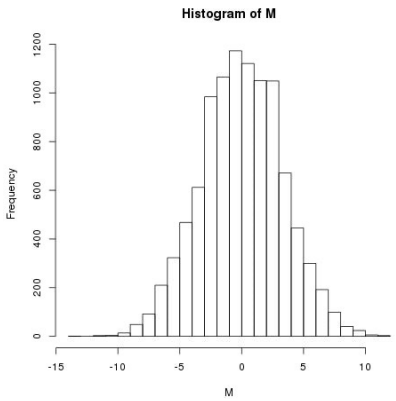
Example - Hypothesis testing using randomization tests (two independent samples)

```
A = c(74,66,88,68,79,68,87,79,78,72,86,85,86)
B = c(69,71,80,88,64,65,74,76,89,68,67,72)

dif = mean(A)-mean(B)

C = c(A,B)
M = c()
pvalue = 0
for (i in 1:10000){
            P = sample(C,replace=FALSE)
            PA = P[1:length(A)]
            PB = P[(length(A)+1):length(P)]
            pdif = mean(PA) - mean(PB)
            M = c(M,pdif)
            if (pdif >= dif)
                        pvalue = pvalue + 1
}

hist(M, breaks=20)
print(pvalue/10000)
```



P-value = 8.89% (it may change slightly if you run again)

20

Mann-Withney test

- This test assumes only an ordinal level of measurement, since it is based on the ranking of scores. It tests whether the two samples come from the same population (equivalent to the two-sample unpaired t-test).

1. Rank the set of n_1 and n_2 scores from lowest (rank 1) to highest.
2. Let R_1 be the sum of ranks of the smallest sample (that of size n_1)
3. Test statistics is $U = n_1 n_2 + \frac{n_1(n_1+1)}{2} - R_1$

- The values for U are tabulated for n<20. It approximates the normal distribution for larger sizes.

Henrique Madeira, DEI-FCTUC, 2018-2023

21

Wilcoxon signed-ranks test

- This test assumes an interval level of measurement. It tests whether the two paired samples come from the same population (equivalent to the two-sample matched-pair t-test).

1. Compute the differences between each pair and rank them.
2. Each rank is given the sign of the difference it corresponds to.
3. Sum the positive ranks and sum the negative ranks. The test statistic is the smallest sum.

- The values are tabulated for n<20. It approximates the normal distribution for larger sizes.

Henrique Madeira, DEI-FCTUC, 2018-2023

22

Sign test

- This test takes only into account the sign of the differences between pairs (equivalent to the two-sample matched-pair t-test and less powerful than Wilcoxon signed-ranks). It is applied when there is no interval information.

1. Compute the sign of the differences between each pair and ignore those that have no difference (reduce sample size accordingly)
2. The test statistic is the number of pairs with less frequent sign.

- The values are tabulated for n<25. It approximates the normal distribution for larger sizes (since it is related to the binomial distribution).

23