BROWN
Computer Science

CS1951A: Data Science

Lecture 13: Non-parametric testing

Lorenzo De Stefani

Spring 2022

Outline

- Limits of normal-distribution based testing
- Non-parametric testing
- Permutation testing
  - One sample
  - Two sample
- Rank-Sum test
- Signed-Rank test

Two-sample tests

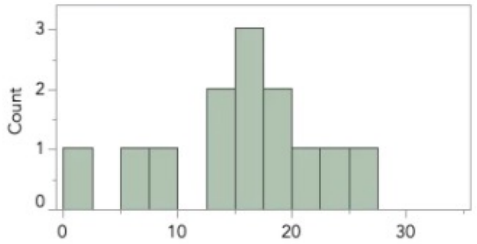
- Using a two-sample test we can compare properties of two observed populations
  - We observe a collection of samples drawn from each population
  - We assume samples are independent

Example

- Our sample data is from a group of men and women who did workouts at a gym three times a week for a year
- We measure the body fat percentages of the two populations
  - Men: 13.3, 6.0, 20.0, 8.0, 14.0, 19.0, 18.0, 25.0, 16.0, 24.0, 15.0, 1.0, 15.0
  - Women: 22.0, 16.0, 21.7, 21.0, 30.0, 26.0, 12.0, 28.0, 23.0

Group=Men

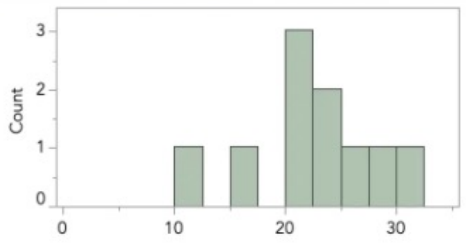
Body Fat Percentage



Summary Statistics	
Mean	14.95
Std Dev	6.84
Std Err Mean	1.90
Upper 95% Mean	19.08
Lower 95% Mean	10.81
N	13.00

Group=Women

Body Fat Percentage



Summary Statistics	
Mean	22.29
Std Dev	5.32
Std Err Mean	1.68
Upper 95% Mean	26.10
Lower 95% Mean	18.48
N	10.00

- We want to compare the means of the two populations μ_M, μ_W
  - H_a: $\mu_M \neq \mu_W$
  - H_0: $\mu_M = \mu_W$
  - $\alpha = 0.05$

Two-sample t-tests

- We can use a two-sample t-test to compare the means of two observed populations
  - We observe a collection of samples drawn from each population
  - We assume samples are independent
  - We assume that the values in the two populations are normally distributed
  - Can be used even with populations with different variance

2-sample t-test

- Compute the empirical variances of each population s_M and s_W
- Compute the pooled variance

$$s_p^2 = \frac{\left((n_M - 1)s_M^2\right) + \left((n_W - 1)s_W^2\right)}{n_M + n_W - 2}$$

- Test statistic $t = \frac{|\mu_M - \mu_W|}{s_p \times \sqrt{\frac{1}{n_M} + \frac{1}{n_W}}}$
- Degrees of freedom $df = n_M - 1 + n_W - 1$
- Compute p-value p using a table ☺
- Compare p-value with set level of confidence α

Can we always use this test?

What if I have more than two populations?

- Break down into multiple hypotheses
  - Use a MHT control procedure!
- Ad-hoc control methods
  - ANOVA
  - Tukey-Kramer test of all pairwise differences
  - Analysis of means (ANOM)
  - Dunnett's test to compare each group mean to a control mean.

Motivation

- Comparing the means of two populations is very important
- If the values of the populations are are normally distributed, we can use methods such as the t-test
- For large sample sizes, even if the populations are not normally distributed, we can invoke the central limit theorem
- However, in some cases the data are clearly NOT normally distributed, and the sample size is too small to invoke the CLT

Can we always use this test?

- What if my data is small and not nearly normally distributed?
  - If your sample sizes are very small, you might not even be able to test for normality
  - You might need to rely on your understanding of the data.
  - When you cannot safely assume normality, you can perform a nonparametric test that doesn't assume normality.

Non-parametric testing

- Non-parametric tests are very robust:
  - Can be used regardless of the distribution of the data
  - They do not rely on assumptions on normal distributions or the CLT
  - This is extremely useful as in practice we can hardly check the correctness of these assumptions
  - But of course, nothing is perfect: What you gain in robustness you lose in power

- The main idea is still the same we used so far:
  - "Is what I am observing a result of random noise, or representative of a statistically significant phenomenon?"
  - "Assuming the null hypothesis is true, how likely it is to observe a result as least as extreme as the one from the data?"

Permutation tests

- New idea: "Let us add some randomness to the data"
  - We then evaluate our test statistic on the scrambled data
  - We want to decide how extreme our initial evaluation (on the unscrambled data) is with respect to the distributions of the randomized ones
  - If the initial observation is still peculiar (E.g., in the 5% of the values obtainable adding randomness) then we interpret it as evidence of the fact that the observed phenomenon is not just due to chance
    - We reject the null hypothesis ☺
    - What is going to be the confidence of this decision?

- One-sample permutation test to compare means
- Two-samples permutation to compare distributions

Permutations

- The word permutation refers to the arrangement of a (multi)set of objects into some specified order.
- Each column is one possible permutation of the three colors:

- In general, given a set of size n there are $n!$ possible permutations of its elements

One-sample Permutation test

- We have n independent and identically distributed observations from an unknown population D

$$x_1, x_2, \ldots, x_n \sim D$$

- Null hypothesis is that the median θ of D is θ_0

$$H_0: \theta = \theta_0$$

- Possible alternative hypotheses:
  - $H_1: \theta > \theta_0$
  - $H_1: \theta < \theta_0$
  - $H_1: \theta \neq \theta_0$

Permutation Vector and Lemma

- The permutation vector $g = (g_1, g_2, \ldots, g_n)$ denotes
  - which observations are above θ_0 ($g_i = 1$),
  - and which are below θ_0 ($g_i = -1$)
- There are 2^n different possible g vectors (each g_i can be 1 or −1)
- If $H_0: \theta = \theta_0$ is true, then $P(x_i < \theta_0) = P(x_i > \theta_0) = 0.5$ by definition of median

Permutation Lemma:

Under $H_0: \theta = \theta_0$, the vector g has probability $\frac{1}{2^n}$ of equaling each of the 2^n different possible outcomes

Permutation Achieved Significance Level

- Test statistic computed using the data

$$T = \frac{1}{n}\sum_{i=1}^{n}(x_i - \theta_0)$$

- For each of the possible permutation vectors we compute an analogous value

$$t_j = \frac{1}{n}\sum_{i=1}^{n}|x_i - \theta_0|g_j$$

- The larger the observed absolute values of the difference, the more evidence against H_0
- The Permutation Achieved Significance (ASL) level is the probability that for a randomly chosen permutation vector $|t_j| > T$

$$ALS_{perm} = |\{t_j \; s.t. |t_j| > |T|\}|/2^n$$

Main idea

- Recall, we want to evaluate how surprising/extreme the observed value T is with respect to the t_j's
  - In particular, how extreme with respect to the distribution of the t_j's
- If the H_0 is correct, the value T should not "stand out" with respect to the other t_j's
  - Should be distributed as the t_j's
- E.g., if T lays in the bottom 5% of the values obtained by randomly permuting, the probability of that happening under the null hypothesis would be no more than 0.05

Test procedure

- One-sided upper alternative $H_1: \theta > \theta_0$

$$p_{val} = |\{t_j | t_j \geq T\}|/2^n$$

- One-sided lower alternative $H_1: \theta < \theta_0$

$$p_{val} = |\{t_j | t_j \leq T\}|/2^n$$

- Two-sided alternative $H_1: \theta \neq \theta_0$

$$p_{val} = |\{t_j | |t_j| \geq |T|\}|/2^n$$

- We can compare p_{val} with the control threshold level α
  - If $p_{val} \leq \alpha$ reject null hypothesis
  - Otherwise, fail to reject.

Problem

- When 2^n is large, computing the t_j for all possible g vectors is computationally expensive (hard)
  - Exponential runtime
- Solution: use a randomized Monte Carlo approach!
  - We need an estimate/approximate of the p-value to be used in the testing

Monte Carlo approximation

- Select B permutation vectors $\boldsymbol{g}_i$ uniformly at random
- For each vector evaluate the statistic t_i
- Compute the approximate p-values

$$p_{val} = |\{t_j \in \{t_1, \ldots, t_B\} \big| |t_j| \geq |T|\}|/B$$

- How much should we sample?
  - Not a clear answer
  - For large n generally 1000, 2000 samples are used

Symmetric distribution

- A probability distribution is said to be symmetric if and only if there exist a value x_0 such that

$$f(x_0 - \delta) = f(x_0 + \delta) \text{ for all } \delta \in R$$

where *f* is the probability density function if the distribution is continuous or the probability mass function if the distribution is discrete.

- The median and the mean (if they exists) of a symmetric distribution are the same
- If we assume the population mean being symmetric, we can use the one-sample permutation test to formulate hypotheses on the population mean

Non-parametric methods to compare distributions

- We previously saw how to use the Chi-Squared test to compare distributions
- This holds based on the assumption that the the difference between observed populations converges to the Chi-Squared distribution
- Permutation tests are an alternative, non-parametric, approach to test distributions

Two-sample permutation test

- We have $N = m + n$ observations
  - $x_1, \ldots, x_m$ are iid random sample from population 1
  - $y_1, \ldots, y_n$ are iid random sample from population 2
- We want to make inferences about the difference of the populations' distribution
  - Let F_1 and F_2 denote distributions of pop. 1 and 2
  - Null hypothesis: F_1 and F_2 are the same distribution
    - $H_0 : F_1(z) = F_2(z), \forall z$
  - Alternative hypothesis is different distributions
    - $H_1 : \exists z \; s.t. F_1(z) \neq F_2(z)$

Permutation Vector and Lemma (2-Sample)

- Let $\boldsymbol{g} = (g_1, g_2, \ldots, g_N)$ be the permutation vector denoting which observation belongs to which population
  - I.e., $g_i = 1 \; if \; x_i \in population \; 1$ and $g_i = -1$ otherwise
  - $\boldsymbol{g}$ contains m X-group labels and n Y-group labels
  - g_i denotes group membership of x_i , where x_i is i-th observation for combined sample of N observations
  - There are $\binom{N}{n}$ different possible values of the permutation vector

- Permutation Lemma:

  Assuming the null-hypothesis $H_0 : F_1(z) = F_2(z), \forall z$ is true, the vector $\boldsymbol{g}$ is uniformly distributed on the $\binom{N}{n} = \frac{N!}{n!m!}$ possible values

Test statistic

- We compute the difference in the means

$$T = \frac{1}{n}\sum_{i=1}^{n} x_i - \frac{1}{m}\sum_{i=1}^{m} y_i$$

- For each of the $\binom{N}{n}$ possible permutation vectors we evaluate a similar quantity

$$t_j = \frac{1}{n}\sum_{i\ with\ label\ X} z_i - \frac{1}{m}\sum_{i\ with\ label\ Y} z_i$$

- To test H_0 we evaluate how surprising the value T is with respect to the value obtained through the random permutations (their distribution)

Test procedure

- Fix critical control/confidence level α
- We sort the computed differences t_j increasingly
  - If T falls in the middle $(1-\alpha)$ fraction of the values, then we fail to reject the null hypothesis
  - Otherwise, if T is on the tails of the observed values, we reject H_0 at significance level α
  - The difference we observed in the populations is extreme enough to give us reason to reject the null hypothesis of the two distributions being the same

Problem

- When n, m are large, computing all of the t_j for all $\binom{N}{n}$ possible random permutation vectors is computationally expensive.
- Solution: use a randomized Monte Carlo approach!

Monte Carlo approximation

- Randomly sample B permutation vectors $\boldsymbol{g}_i$
- For each vector evaluate the statistic t_i
- We sort the computed differences
  - If T falls in the middle $(1-\alpha)$ fraction of the B values, then we fail to reject the null hypothesis
  - Otherwise, if T is on the tails of the observed values, we reject H_0 at α significance level
    - The difference we observed in the populations is extreme enough to give us reason not reject the hypothesis of the two distributions being the same

- How much should we sample?
  - Not a clear answer
  - For large n generally 1000, 2000 samples are used

The Rank-Sum Test (Mann-Whitney U test)

- Consider the samples
  - $X_1 = \{9.0, 11.5, 11.5, 12.0, 13.0, 13.25\}$
  - $X_2 = \{9.0, 9.5, 9.5, 9.75, 10.0, 13.0\}$
- H_0: The mean of the two distributions is the same
- We start by ranking the observations according to their size relative to the whole sample
  - If there are ties, we average the ranks

measurements	9.0	9.0	9.5	9.5	9.75	10.0	11.5	11.5	12.0	13.0	13.0	13.25
ranks	1	2	3	4	5	6	7	8	9	10	11	12
modified ranks	1.5	1.5	3.5	3.5	5	6	7.5	7.5	9	10.5	10.5	12

The Rank-Sum Test (Mann-Whitney U test)

- We compute:
  - R_1 by summing the ranks of the entries of the smaller sample
  - R_2 by summing the ranks of the entries of the larger sample
- H_0: The mean of the two distributions is the same
- If the null hypothesis is true, we would expect R1 and R2 to have similar value
- The U test statistic is computed as

$$U = \min\{R_1 - \frac{n_1(n_1+1)}{2}, R_2 - \frac{n_2(n_2+1)}{2}\}$$

  - If there is a complete separation between the populations U=0
  - If the values are well interleaved, we would observe a higher values of U

The Rank-Sum Test (Mann-Whitney U test)

Critical Values of the Mann-Whitney U
(Two-Tailed Testing)

n_2	α	n_1																	
		3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20
3	.05	--	0	0	1	1	2	2	3	3	4	4	5	5	6	6	7	7	8
	.01	--	0	0	0	0	0	0	0	0	1	1	1	2	2	2	2	3	3
4	.05	--	0	1	2	3	4	4	5	6	7	8	9	10	11	11	12	13	14
	.01	--	--	0	0	0	1	1	2	2	3	3	4	5	5	6	6	7	8
5	.05	0	1	2	3	5	6	7	8	9	11	12	13	14	15	17	18	19	20
	.01	--	--	0	1	1	2	3	4	5	6	7	7	8	9	10	11	12	13
6	.05	1	2	3	5	6	8	10	11	13	14	16	17	19	21	22	24	25	27
	.01	--	0	1	2	3	4	5	6	7	9	10	11	12	13	15	16	17	18
7	.05	1	3	5	6	8	10	12	14	16	18	20	22	24	26	28	30	32	34
	.01	--	0	1	3	4	6	7	9	10	12	13	15	16	18	19	21	22	24
8	.05	2	4	6	8	10	13	15	17	19	22	24	26	29	31	34	36	38	41
	.01	--	1	2	4	6	7	9	11	13	15	17	18	20	22	24	26	28	30
9	.05	2	4	7	10	12	15	17	20	23	26	28	31	34	37	39	42	45	48
	.01	0	1	3	5	7	9	11	13	16	18	20	22	24	27	29	31	33	36

- Critical values are given for two-tailed test
- Rows and columns correspond to the sizes of the smaller and larger samples, respectively
- For every combination of row and column, there are two subrows:
  - the top gives the critical values for confidence 10% (i.e., $\alpha = 0.1$)
  - bottom the 5% ones.
- For a one-sided test at 5% use the relevant top entry.

The Rank-Sum Test (Mann-Whitney U test)

- For a given choice of n_1, n_2, α we get a value u(n_1, n_2, α)
- If our computed test statistic $U \leq$ u(n_1, n_2, α), we reject the null
- Otherwise we fail to reject the null

Larger Samples

- The table only goes up to large sample size 20
- For larger samples use normal approximation

$$z = \frac{\mathrm{U} - \mathrm{m_u}}{\sigma_U}$$

$$\mathrm{m_u} = \frac{\mathrm{n_1 n_2}}{2}$$

$$\sigma_U = \sqrt{\frac{\mathrm{n_1 n_2 (n_1 + n_2 + 1)}}{12}}$$

- Then compare with the normal table
- E.g., for two-tailed test at 0.05 reject null if $|z| > 1.96$.

Wilcoxon Signed-Rank Test for Paired Data

- Used to study the distribution of the difference of paired observations
  - Let N be the sample size, thus $2N$ data points in pairs
  - $(x_{i,1}, x_{i,2})$ denote the measurements
- Assumptions:
  - Data are paired $(x_{i,1}, x_{i,2})$ and come from the same population
  - Each pair is chosen uniformly and independently at random
  - The data are measured on at least an interval scale when, as is usual, within-pair *differences* are calculated to perform the test

Testing procedure

- H_0: difference between the pairs follows a symmetric distribution around zero
- H_1: difference between the pairs does not follow a symmetric distribution around zero.

1. Calculate $d_i = |x_{2,i} - x_{1,i}|$ and $sgn = (x_{2,i} - x_{1,i})$
2. Exclude pairs for which $d_i = 0$, and let N_r be the adjusted sample size
3. Rank the pairs according to increasing values for the d_i's
   - Smallest d_i has rank 1
   - Let R_i be the rank associated with d_i
   - Ties are split by averaging the ranks

Testing procedure

4. Calculate the test statistic

$$W = \sum_{i=1}^{N_r} sgn(x_{2,i} - x_{1,i}) R_i$$

Under the null hypothesis W follows the "W-distribution" with mean 0 and variance

$$\frac{N_r(N_r+1)(2N_r+1)}{6}$$

5. Given a critical value α, we can obtain a corresponding value W_{α,N_r} using the opportune table

6. If $|W| > W_{\alpha,N_r}$ reject H_0

Example

- We consider a comparative study between different methods of preparing breasts for breastfeeding.
- Each mother treated one breast, leaving the other untreated.
- The following data gives the difference in the level of discomfort (1 to 4) between treated and untreated breast for a particular treatment.
- There are 19 measurements overall.

-0.525, 0.172, -0.577, 0.200, 0.040, -0.143, 0.043, 0.010, 0.000, -0.522, 0.007, -0.122, -0.040, 0.000, -0.100, 0.050, -0.575, 0.031, -0.060.

Example

- We rank the observations by absolute value after dropping the zero values

Diff	0.007	0.010	0.031	0.040	-0.040	0.043	0.050	-0.060	-0.100
Rank	1	2	3	4.5	4.5	6	7	8	9
Diff	-0.122	-0.143	0.172	0.200	-0.522	-0.525	-0.575	-0.577	
Rank	10	11	12	13	14	15	16	17	

- We thus have $W = 48.5$

Example

n	P = 0.10	P = 0.05
5	2	-
6	2	0
7	3	2
8	5	3
9	8	5
10	10	8
11	14	10
12	17	13
13	21	17
14	26	21
15	30	25
16	36	29
17	41	34
18	47	40
19	53	46
20	60	52
21	67	58
22	75	65
23	83	73
24	91	81
25	100	89

- Since we dropped two values our sample size is 19-2=17
- Looking at the corresponding row we find the critical value of 34 at the 5% level
- To reject we would have to observe $T \leq 34$.
- We fail to reject the null

Sample sizes

- For $N_r < 20$ it is necessary to use the exact distribution (i.e. the table)
- As N_r increases the sampling distribution of W converges to a normal distribution
  - z-score $z=\frac{W}{\sigma_W}$, where $\sigma_W = \sqrt{\frac{N_r(N_r+1)(2N_r+1)}{6}}$
  - Given a critical confidence threshold α, we obtain the corresponding critical value z_α using opportune table
  - Reject H_0 if $|z| > z_\alpha$

How Would the t-test Do?

- Paired t-test:
  - For every pair $(x_{1,i}, x_{2,i})$ compute the difference $d_i = (x_{2,i} - x_{1,i})$
  - Just run the "standard" t-test on the d_i values!
- We would have $\bar{d} = -0.11$, $s_d = 0.25$

$$t = \sqrt{n}\frac{\bar{d}}{s_d} = -1.95$$

- Given $\alpha = 0.05$, we can obtain the corresponding threshold using the table
  - For two-tailed test the threshold would be 2.10
  - Since 2.10>-1.95 we would fail to reject H_0

Is the t-test Justified?

- Does the data look like it comes from a normal distribution? Let's look at the histogram.