



ARISTOTLE UNIVERSITY  
OF THESSALONIKI

FACULTY OF HEALTH SCIENCES - SCHOOL OF MEDICINE  
MSc Health Statistics and Data Analytics

# Hypothesis testing: Paired and two-sample t-tests Mann-Whitney U test and Wilcoxon Signed Ranks test

Kokkali Stamatia  
Research Fellow- General Practitioner



THESSALONIKI 2021-22



# Objectives

- Form statistical hypothesis to check specific research questions
- Recognize statistical tests to compare a continuous outcome between two groups
- Choose the appropriate test based on the assumptions of each test
- 
- Interpret results of the test in terms of the null and the alternative hypothesis

# Hypothesis Testing

- A **statistical hypothesis** is a research question in statistical terms.
  - Example: drug X has an effect on reducing systolic blood pressure
- A **statistical test** is the procedure which we follow in order to decide if we should reject this hypothesis or not, based on a sample.

# Hypothesis Testing

- Briefly, we need to:

1. Define the Null and Alternative Hypothesis: . The null hypothesis ( $H_0$ ) indicates that no difference exists between conditions, groups, or variables while the alternate hypothesis ( $H_a$ ) indicates a difference or association.

- Example

$H_0$ : drug X does not reduce the systolic blood pressure

$H_a$ : drug X reduces the systolic blood pressure

# Hypothesis Testing

2. Choose the significance level, represented with the Greek letter  $\alpha$ 
  - Typical values are 0.05 and 0.01
  - The significance level is the highest value of a probability value for which the null hypothesis is rejected.
  - Type I error

# Hypothesis Testing

3. Identify the appropriate test statistic and calculate the observed test statistic from the data

- Particular type of test based on characteristics of the data.
- Independent groups or not
- Normally distributed data or not

# Hypothesis Testing

4. **Compare** the probability value with the  $\alpha$  level
- The p-value is the probability of obtaining the observed results, or something more extreme, if the null hypothesis is true
  - The p-value  $< \alpha$ , this result is statistically significant. Reject the  $H_0$
  - The p-value  $> \alpha$ , this result is NOT statistically significant. We cannot reject the  $H_0$ .
  - The smaller the p-value, the greater the evidence against the null hypothesis.

# Hypothesis Testing

## 5. Interpret the results

Communicating results in a meaningful and comprehensible manner makes the research useful to others.




# Type I and Type II errors

	TRUE SITUATION	
TEST RESULT	Null hypothesis is <b>true</b>	Null hypothesis is <b>false</b>
Not significant (don't reject H0)	<b>Correct conclusion</b> Probability = $1-\alpha$	<b>Type II error</b> Probability = $\beta$ (false negative result)
Significant (reject H0)	<b>Type I error</b> Probability = $\alpha$ (false positive result)	<b>Power of the study</b> Correct conclusion Probability = $1-\beta$

- The power of the study is the probability of getting a statistically significant result with the selected sample if a true difference exists.
- The power is equal to  $1-\beta$  - the larger the power of the study, the smaller the Type II error.

# Court system and hypothesis testing




jury doesn't  
reject his  
innocence

CORRECT DECISION

INNOCENT


TEST RESULT	TRUE SITUATION	
	Null hypothesis is <u>true</u>	Null hypothesis is <u>false</u>
Not significant (don't reject H0)	Correct conclusion Probability = $1-\alpha$ 😊	<b>Type II error</b> Probability = $\beta$ (false negative result) 😞
Significant (reject H0)	<b>Type I error</b> Probability = $\alpha$ (false positive result) 😞	<b>Power of the study</b> Correct conclusion Probability = $1-\beta$ 😊



jury doesn't  
reject his  
innocence

I got away  
with it!!


GUILTY MAN



I've been  
framed!

jury rejects  
his innocence

INNOCENT



jury rejects  
his innocence

CORRECT DECISION

GUILTY MAN

## Type I error ( $\alpha$ )

- the null hypothesis is **rejected** while it is **true**
- Also called a **false positive result**
- concludes that there is an **effect when**, in reality, **there is none**.
- maximum probability is set **in advance as alpha**
- is **not affected by sample size** as it is set in advance
- **increases** with the number of tests or end points

## Type II error ( $\beta$ )

- the null hypothesis is **NOT rejected** while it is **false**
- Also called a **false negative result**
- concludes that there is **no effect when**, in reality, **there is one**.
- **depends** upon **sample size** and **alpha**
- **decreases** with **larger sample size**
- **decreases** with the number of tests or end points

# Hypothesis Testing

1. From the research question, determine the appropriate null hypothesis,  $H_0$ , and the alternative,  $H_1$ .
2. Set the level of significance,  $\alpha$  ( $\alpha=0.05$ )
3. Identify the appropriate test statistic and check the assumptions
4. Decide whether or not the result is statistically significant.
  - The ***p-value***  $< 0.05$ , this result ***is statistically significant***.  
**Reject the  $H_0$**
  - The ***p-value***  $> 0.05$ , this result ***is NOT statistically significant***. **We cannot reject the  $H_0$** .
5. Interpret the results

# Two Independent samples test

1. State the null and alternative hypothesis
  - $H_0: \mu_1 = \mu_2$  or  $\mu_1 - \mu_2 = 0$   $H_1: \mu_1 \neq \mu_2$  or  $\mu_1 - \mu_2 \neq 0$
2. Check for normality in the continuous variable within the two groups
  - If normal ( $P > 0.05$ ) in both groups then t-test
    - Check for equality of variances (Levene's test)
      - If  $P > 0.05$  then unpaired t-test with pooled variance
      - If  $P < 0.05$  then unpaired t-test with unpooled variance (Welch's test)
  - If not normal ( $P < 0.05$ ) in at least one group then Mann-Whitney U test
    - $H_0: md_1 = md_2$   $H_1: md_1 \neq md_2$

# Paired test

1. State the null and alternative hypothesis
  - $H_0: \delta = 0$      $H_1: \delta \neq 0$
2. Calculate the difference  $d_i$  and check for normality
  - If normal ( $P > 0.05$ ) then Paired t-test
  - If not normal ( $P < 0.05$ ) the Signed-Ranks Wilcoxon test

# What test statistic to use

- Test selection depends on:
  - Type of variable whose values we want to test (i.e. continuous, categorical, ordinal)
  - The number of observations (e.g. number of individuals)
  - The existence of correlation between observations (e.g. do we measure blood pressure before and after giving the patient anti-hypertension pills?)
  - How many groups are we comparing?

# Continuous outcome

Example: Weight, Height, levels of Hematocrit, levels of INR

	Two groups	
	Independent (e.g males vs.females, treated vs.untreated)	Correlated (e.g. before vs. after treatment)
Parametric	<b>t-test:</b> compares means between two independent groups	<b>Paired t-test:</b> compares means between two related groups (e.g., the same subjects before and after)
Non-parametric	<b>Mann-Whitney U test</b>	<b>Wilcoxon Signed Ranks test</b>



# The independent samples T-test

- Is the difference in means that we observe between two groups more than we'd expect to see based on chance alone?
- Hypothesis Testing
  - $H_0: \mu_1 = \mu_2$  or  $\mu_1 - \mu_2 = 0$
  - $H_1: \mu_1 \neq \mu_2$  or  $\mu_1 - \mu_2 \neq 0$

# Assumptions of the independent samples t-test

- **Independence:** We need two independent groups, i.e. male/female, treated/untreated.
- **Normality:** The observations within each group should be approximately normally distributed (and be measured on the continuous scale).
- **Assumption of Homogeneity of Variances**
  - Levene's test
    - If  $p > 0.05$  then unpaired t-test with pooled variance
    - If  $p < 0.05$  then unpaired t-test with unpooled variance (Welch's test)

## Example

- It is well known that women have lower hemoglobin levels compared to men.
- As an example, we will test this hypothesis using a sample of 40 individuals from the transfusion dataset
- HB levels (g/dL)
  - **Males (N=20):** 14.7 13.3 15.0 15.0 15.2 13.9 12.2 13.7 13.1 15.3 12.0 14.3 14.3 12.4 15.4 14.7 13.9 13.8 14.8 14.1
  - **Females (N=20):** 13.4 15.4 12.9 12.7 14.0 12.3 15.2 14.0 13.3 13.9 12.3 13.6 13.5 13.7 14.6 15.8 13.9 13.6 12.4 14.4

# Data Summary

	n	Sample Mean	Sample Standard Deviation
Group 1: women	20	13.74g/dl	0.99
Group 2: men	20	14.05g/dl	1.03

# Independent samples t-test

## 1. Define your hypotheses (null, alternative)

$H_0$ : mean Hb difference between men and women is 0 ( $\mu_{\text{men}} = \mu_{\text{women}}$ ).

$H_a$ : mean Hb difference between men and women is not 0. ( $\mu_{\text{men}} \neq \mu_{\text{women}}$  [two-sided]).

We can assume that F and M have relatively similar standard deviations/variances, so make a “pooled” estimate of variance.

# Results reporting

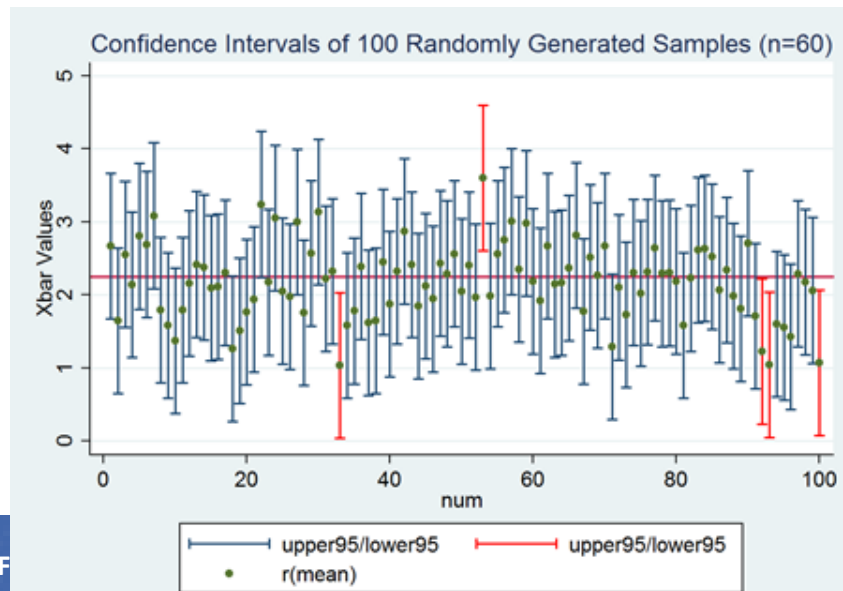
	Females N=31	Males N=31	p
Heamoglobin, Hb (g/dl) Mean (SD)	13.74 (0.99)	14.05(1.03)	0.338
Difference Mean (95% CI)	0.31 (-0.34,0.96)		

95% CI includes 0, so difference not significant at  $\alpha=0.05$

- An independent samples t-test was used to compare the heamoglobin levels between men and women.
- Results show that there is no difference in the heamoglobin levels between men and women ( $p=0.338$ )

# Confidence intervals (CI)

- is a type of interval estimate of an effect of interest
- 95% CI if the experiment was repeated 100 times under the same conditions on the same population then 95 of these CIs would be expected to capture the true effect.



The confidence intervals of the 100 randomly generated samples (sample size = 60). Ninety-five of them covering the population mean (the horizontal red line  $\mu=2.25$ )

# Non-parametric tests

- ***t-tests*** require your outcome variable to be normally distributed (or close enough), for small samples.
- **Non-parametric or distribution free tests** are based on **RANKS** instead of means and standard deviations (=“population *parameters*”).



# Mann-Whitney U test

- This is the nonparametric equivalent of the Independent samples t-test
- The Mann-Whitney U test is more broadly used to interpret whether there are differences in the *distributions* of two groups or differences in the *medians* of two groups.
- In R the test is called Wilcoxon Mann-Whitney test

# Mann-Whitney U test

- It is much more robust against outliers and heavy tail distributions.
- **Null Hypothesis:** No difference in scores of the two groups (i.e. the sum of ranks for group 1 is no different than the sum of ranks for group 2).
  - $H_0: md_1 = md_2$
- **Alternative Hypothesis:** There is a difference between the scores of the two groups (i.e. the sum of ranks for group 1 is significantly different from the sum of ranks for group 2).
  - $H_1: md_1 \neq md_2$

# Mann Whitney U Test

- To compute the Mann Whitney U:

- Rank the scores in both groups (together) from highest to lowest.
- Sum the ranks of the scores for each group.
- The sum of ranks for each group are used to make the statistical comparison.

Income	Rank	No Income	Rank
25	12	27	10
32	5	19	17
36	3	16	20
40	1	33	4
22	14	30	7
37	2	17	19
20	16	21	15
18	18	23	13
31	6	26	11
29	8	28	9
	85		125

# Calculation of Mann-Whitney U test with R

- The null hypothesis states that there is no difference in the scores of the populations from which the samples were drawn.
- The Mann Whitney  $U$  is sensitive to both the central tendency of the scores and the distribution of the scores.
- R results:

```
> wilcox.test(group1,group2)
```

```
Wilcoxon rank sum test
```

```
data: x and y
```

```
W = 70, p-value = 0.1431
```

	Income	No income	p
Population Scores (median(IQR))	30 (14.75)	24.5 (10)	0.143

There was no difference in the scores between those with income and those with no income ( $U=70$ ,  $p=0.143$ ).

## The paired t-test (Dependent samples t-test)

- Tests whether the mean difference between two sets of observations is zero.
- Hypothesis Testing:
  - $H_0: \mu_d = 0$
  - $H_1: \mu_d \neq 0$

## Assumptions of the Paired $t$ -test

- The mean difference should be measured on a continuous scale.
- The mean difference consists of two related groups.
- The differences between pairs are approximately normally distributed.
- **Note:** The paired  $t$ -test does *not* assume that observations within each group are normal, only that the differences are normal.

## Example

- Assume we had data on WBC before and after transfusion
- We would not apply a two-sample t-test to answer this question, since this assumes independent observation
- After-Before yields a single sample of differences.
  - WBC\_b: 8647 7524 7887 8230 7192 7218 8069 7960 7759 7926  
5467 7782 7182 7094 5927 7603 8437 7141 7576 7184 6676  
7395 6491 6504 7664 7743 6961 8573 7578 6681 5268 8047  
7298 6808 6011 7179 7086 7850 7726 6604
  - WBC\_a: 6772 7773 7621 9550 5645 8688 7884 8816 8262 6705  
9715 7855 5511 8305 7281 8126 8555 8154 8877 7579 6994  
8026 9365 7976 8525 9743 7718 8248 8787 7475 6858 9845  
9367 8606 7109 6989 7484 7047 9666 9749



# Data Summary

## Mean Change (WBC)

	N	Sample Mean Change $\bar{\delta}$	Sample Standard Deviation $sd_{\bar{\delta}}$
Group 1: Change (cells/ml)	80	782.57	1267.71

# Do the transfusion affected the WBC levels significantly?

- R results:

$t = -3.9043, df = 39, p\text{-value}=0.0004$   
95% CI:  $(-1188.007, -377.143)$

- So, reject the  $H_0$  at  $\alpha=0.05$  significance level

# Results reporting

	Before transfusion	After transfusion	p
WBC (cells/ml) (mean(SD))	7298.70 (777.66)	8081.27 (1090.08)	<0.001
Mean Difference (95% CI)	-782.57 (-1188.01, -377.14)		

95% CI does not include 0, so difference significant at  $\alpha=0.05$

- A paired samples t-test was used to compare WBC levels before and after transfusion.
- Results WBC levels before transfusion were significantly lower compared to the WBC levels after transfusion (MD: -782.57; 95%CI(-1188.01, -377.14;  $p<0.001$ ).

# Alternative tests when normality is violated: Non-parametric tests

# Wilcoxon signed Ranks test

1. The nonparametric analog of the two-sample case with dependent samples
2. The null hypothesis states that there is no difference on an identified variable before and after treatment or between two matched groups.
3. The test statistic for the Wilcoxon test is  $T$ .
4. The sampling distribution is the  $T$  distribution.

# Hypothesis Testing

- **Null Hypothesis:** There is no difference in scores before and after an intervention (i.e. the sums of the positive and negative ranks will be similar).
  - $H_0: md = 0$
- **Non-Directional Research Hypothesis:** There is a difference in scores before and after an intervention (i.e. the sums of the positive and negative ranks will be different).
  - $H_1: md \neq 0$

# Wilcoxon Test

- To compute the Wilcoxon  $T$ :

- Determine the differences between scores.
- Rank the absolute values of the differences.
- Place the appropriate sign with the rank (each rank retains the positive or negative value of its corresponding difference)
- $T$  = the sum of the ranks with the less frequent sign
- $T = -4$

Pretest	Posttest	Difference	Rank
36	21	15	11
23	24	-1	-1
48	36	12	10
54	30	24	12
40	32	8	7
32	35	-3	-3
50	43	7	6
44	40	4	4
36	30	6	5
29	27	2	2
33	22	11	9
45	36	9	8

# Wilcoxon Test with R

```
> wilcox.test(x,y, paired=TRUE)
```

Wilcoxon signed rank test

data: x and y

V = 74, p-value = 0.003418

	Before Intervention	After Intervention	p
Scores (median(IQR))	38 (15.00)	31 (11.25)	0.003

There is significant difference in the scores before and after the intervention (V=74, p=0.003). Scores before the intervention were significantly higher (median(IQR)= 38 (15.00)) compared to scores after the intervention (31(11.25)).