

# Social Robotics

Albert-Ludwigs-Universität Freiburg



Bernhard Nebel, Felix Lindner, Thorsten Engesser,  
Barbara Kuhnert, Laura Wächter  
WS 2017/18

## Non-parametric Tests

Nebel, Lindner, Engesser, Kuhnert, Wächter – Social Robotics

2 / 25

## Overview

- Wilcoxon signed-rank test
- Wilcoxon rank-sum test (aka Mann-Whitney test)
- Kruskal-Wallis test

Nebel, Lindner, Engesser, Kuhnert, Wächter – Social Robotics

3 / 25

## Ranks

- Ranks are natural numbers starting with 1, which get assigned to scores sorted in increasing order.
- Ranks can be assigned to any data which is at least ordinal.
- Ranks are robust against outliers (because ranks are used instead of the actual data).

### Example

- Data: 0, 7, 3; Rank: 1, 3, 2
- Data: -100, 99, 98; Rank: 1, 3, 2
- Data: d, a, b; Rank: 3, 1, 2

Nebel, Lindner, Engesser, Kuhnert, Wächter – Social Robotics

4 / 25

- In case of ties, the average rank is assigned to the whole group of scores that constitutes the tie.

### Example

- Data: 1, 6, 4, 4, 2, 2, 2
- Rank: 1, 7, 5.5, 5.5, 3, 3, 3

- Likert scales are a popular means of measurement.
- Likert scales in most cases have no interval-scale reading.

### Example

Five participants are asked to rate their belief in the possibility that humans will one day be the slaves of robots before and after they have watched a Sci-Fi movie. As a measurement instrument, a 3-Point Likert-Scale "never ever!" (1), "maybe" (2), "yes, sure!" (3) was used.

- Before: 1, 2, 2, 3, 3; After: 2, 3, 3, 3, 1
- Difference: -1, -1, -1, 0, +2
- Differences without 0: -1, -1, -1, +2
- Ranks: 2, 2, 2, 4

## Wilcoxon Signed-Rank Test: Example Continued

### Example

Five participants are asked to rate their belief in the possibility that humans will one day be the slaves of robots before and after they have watched a Sci-Fi movie. As a measurement instrument, a 3-Point Likert-Scale "never ever!" (1), "maybe" (2), "yes, sure!" (3) was used.

- Before: 1, 2, 2, 3, 3; After: 2, 3, 3, 3, 1
- Difference: -1, -1, -1, 0, +2; Without 0: -1, -1, -1, +2
- Ranks: 2, 2, 2, 4
- Let  $V = \sum_i^n Z_i R_i$  be the sum of the positive ranks ( $Z_i = 1$  if difference  $i$  is positive, and  $Z_i = 0$  else).
- In the example  $V = 4$ . Well, so what?

## Towards the Null Hypothesis

- Imagine two paired samples and consider their rank differences.
- Consider  $V = \sum_i^n Z_i R_i$ . What could happen?
  - 1 Case  $V = 0$ : All the rank differences are negative.
  - 2 Case  $V = \sum_i^n R_i = \frac{n(n+1)}{2}$ : All rank differences are positive.
  - 3 Else:  $V$  ranges between 0 and  $\frac{n(n+1)}{2}$ .
- If the groups do not differ ( $H_0$ ), then 50% of the differences should be below 0 and 50% above. This is like saying that the median of the difference is 0. And in that case,  $V$  should be close to  $\frac{\frac{n(n+1)}{2}}{2} = \frac{n(n+1)}{4}$ .
- Hence, we will test  $H_0 : \text{Mdn} = 0$  against its alternatives, and we will do that by using  $V$ .

- The nice thing about  $V$  is that (for  $n > 25$ ) its distribution is well approximated by a normal distribution  $\mathcal{N}(\mu_V, \sigma_V)$  with
  - $\mu_V = \frac{n(n+1)}{4}$
  - $\sigma_V = \sqrt{\frac{n(n+1)(2n+1)}{24}}$

- The nice thing about  $V$  is that (for  $n > 25$ ) its distribution under  $H_0$  is well approximated by a normal distribution  $\mathcal{N}(\mu_V, \sigma_V)$  with
  - $\mu_V = \frac{n(n+1)}{4}$
  - $\sigma_V = \sqrt{\frac{n(n+1)(2n+1)}{24}}$
- **Proof (Mean):** We already came to this conclusion earlier on Slide 8.

- The nice thing about  $V$  is that (for  $n > 25$ ) its distribution is well approximated by a normal distribution  $\mathcal{N}(\mu_V, \sigma_V)$  with
  - $\mu_V = \frac{n(n+1)}{4}$
  - $\sigma_V = \sqrt{\frac{n(n+1)(2n+1)}{24}}$
- **Proof (Variance)**
  - First, we define  $V' = \sum_i^n V'_i$  with
 
$$V'_i = \begin{cases} 0 & \text{with probability } 0.5 \\ i & \text{with probability } 0.5 \end{cases}$$
  - ( $V'$  has the same distribution as  $V$ , because, for every rank, it either belongs to the sum of  $V$  or not with probability 0.5.)
  - $\text{Var}(V) = \text{Var}(V') = \sum_i^n \text{Var}(V'_i)$  (independence of  $V'_i$ ).
  - $\text{Var}(V'_i) = E(V_i'^2) - E(V_i')^2 = (0^2 \cdot \frac{1}{2} + i^2 \cdot \frac{1}{2}) - (\frac{1}{2}i)^2 = \frac{i^2}{4}$
  - $\text{Var}(V) = \sum_i^n \text{Var}(V_i) = \sum_i^n \frac{i^2}{4} = \frac{n(n+1)(2n+1)}{24}$ .

## Example

Five participants are asked to rate their belief in the possibility that humans will one day be the slaves of robots before and after they have watched a Sci-Fi movie. As a measurement instrument, a 3-Point Likert-Scale "never ever!" (1), "maybe" (2), "yes, sure!" (3) was used.

- Before: 1, 2, 2, 3, 3; After: 2, 3, 3, 3, 1
- Difference: -1, -1, -1, 0, +2; Without 0: -1, -1, -1, +2
- Ranks: 2, 2, 2, 4
- $V = 4$ ,  $\mu_V = 4(4+1)/4 = 5$ ,  $\sigma_V = \sqrt{4(4+1)(2 \times 4 + 1)/24}$
- $z = \frac{V - \mu_V}{\sigma_V} = (4 - 5)/2.74 = -0.365$
- $p = P(z \leq -0.365) + 1 - P(z \leq 0.365) = 0.715$

## Comparison to Paired t-Test

### Example: t-Test

Five participants are asked to rate their belief in the possibility that humans will one day be the slaves of robots before and after they have watched a Sci-Fi movie. As a measurement instrument, a 3-Point Likert-Scale "never ever!" (1), "maybe" (2), "yes, sure!" (3) was used.

- Before: 1, 2, 2, 3, 3; After: 2, 3, 3, 3, 1
- Difference: -1, -1, -1, 0, +2
- $\bar{D} = 0.20$ ,  $s_D = 1.30$ ,  $n = 5$
- $t = \sqrt{5} \times 0.20 / 1.30 = 0.344$ ,  $df = 4$
- $p = 0.748$

## Wilcoxon Rank-Sum Test: Motivation

### Example

Five participants are asked to rate their belief in the possibility that humans will one day be the slaves of robots after they have watched the Sci-Fi movie M1, and five participants rate their belief after watching Sci-Fi movie M2. As a measurement instrument, a 3-Point Likert-Scale "never ever!" (1), "maybe" (2), "yes, sure!" (3) was used.

- M1: 1, 1, 2, 2, 2; M2: 2, 3, 3, 3, 2
- $H_0$ : The two groups are equal.
- Reject  $H_0$  or not?

## Wilcoxon Rank-Sum Test: General Setting

- First, all scores are ranked together.
- First group's rank sum:  $R_1 = \sum_{i=1}^{n_1} r_{1,i}$
- Second group's rank sum:  $R_2 = \sum_{i=1}^{n_2} r_{2,i}$
- First group's W:  $W_1 = R_1 - \sum_{i=1}^{n_1} i = R_1 - \frac{n_1(n_1+1)}{2}$
- Second group's W:  $W_2 = R_2 - \sum_{i=1}^{n_2} i = R_2 - \frac{n_2(n_2+1)}{2}$
- $W_1 + W_2 = R_1 - \frac{n_1(n_1+1)}{2} + R_2 - \frac{n_2(n_2+1)}{2} = n_1 n_2$
- **Note:** The Wilcoxon Rank-Sum Test is also known as Mann-Whitney U-Test, and W is also called U. There are various ways of defining W (resp. U), which are all equal! R uses the statistics W the way shown above.

## Wilcoxon Rank-Sum Test: Distribution of W

- For larger samples ( $n_1 > 10, n_2 > 10$ ),  $W \sim \mathcal{N}(\mu_W, \sigma_W)$ :
  - $\mu_W = \frac{n_1 n_2}{2}$
  - $\sigma_W = \sqrt{\frac{n_1 n_2 (n_1 + n_2 + 1)}{12}}$
  - Also see simulation in `lecture11.Rmd` in the git repository slides.
- Again, we can calculate z-values to decide whether or not W is extreme, i.e., whether or not to reject  $H_0$ .

## Wilcoxon Rank-Sum Test: Example Continued

### Example

- M1: 1, 1, 2, 2, 2
- M2: 2, 3, 3, 3, 2
- All Scores: 1, 1, 2, 2, 2, 2, 3, 3, 3, 2
- Ranks: 1.5, 1.5, 5, 5, 5, 5, 9, 9, 9, 5
- $R_1 = 18$ ,  $W = 18 - 15 = 3$
- $z = \frac{\frac{3 - (5 \times 5)}{2}}{\sqrt{\frac{5 \times 5(5+1)}{12}}} = -2.298$
- $p = P(z \leq -2.298) + 1 - P(z \leq 2.298) = 0.022$

## Wilcoxon Rank-Sum Test vs. t-Test

### Example

- M1: 1, 1, 2, 2, 2
- M2: 2, 3, 3, 3, 2
- $\bar{X}_1 = 1.6$ ,  $\bar{X}_2 = 2.6$ ,  $s_1^2 = 0.3$ ,  $s_2^2 = 0.3$ ,  $n = 5$ ,  $df = 8$
- $t = \sqrt{n} \times \frac{\bar{X}_1 - \bar{X}_2}{\sqrt{s_1^2 + s_2^2}} = \sqrt{5} \times \frac{1.6 - 2.6}{\sqrt{0.3 + 0.3}} = -2.887$
- $p = P(t \leq -2.887) + 1 - P(t \leq 2.887) = 0.020$
- For a simulation comparing Wilcoxon and t-Test see `lecture11.Rmd` in the git repository.

## Kruskal-Wallis Test: Setting

- Also for rank-based methods, there is an analog to ANOVA that can cope with more than two groups: [Kruskal-Wallis Test](#). As for ANOVA,  $H_0$  reads “There is no difference between the groups”.
- First, the scores of all groups are ranked together (like for Wilcoxon Rank-Sum Test).
- The test statistics is called H:
  - $H = (N - 1) \frac{\sum_i^p n_i (\bar{r}_i - \bar{r})^2}{\sum_i^p \sum_{j=1}^{n_i} (r_{ij} - \bar{r})^2}$ , with  $N = \sum_i^p n_i$ ,  $\bar{r}_i = \frac{\sum_{j=1}^{n_i} r_{ij}}{n_i}$ ,  $\bar{r} = \frac{N+1}{2}$
  - H can be simplified to  $H = \frac{12}{N(N+1)} \sum_i^p n_i \bar{r}_i^2 - 3(N+1)$
- $H \sim \chi_{p-1}^2$ , with  $p$  being the number of groups.

## Kruskal-Wallis Test: Example

### Example

- M1: 1, 1, 2, 2, 2; Ranks: 2.5, 2.5, 8.5, 8.5, 8.5
- M2: 2, 3, 3, 3, 2; Ranks: 8.5, 14, 14, 14, 8.5
- M3: 1, 2, 2, 1, 2; Ranks: 2.5, 8.5, 8.5, 2.5, 8.5
- $\bar{r}_1 = 6.1$ ,  $\bar{r}_2 = 11.8$ ,  $\bar{r}_3 = 6.1$ ,  $N = 15$ ,  $\bar{r} = (15+1)/2 = 8$
- $H = \frac{12}{15 \times 16} \times 5(37.21 + 139.24 + 37.21) - 3 \times 16 = 5.41$
- $p = 1 - P(\chi^2 \leq 5.41) = 0.067$
- R will report different values, see next slide to learn why.

## Ties call for Corrections

If there are long ties (i.e., a lot of scores getting the same rank), the variance of the statistics become smaller and thus some corrections have to be applied.

- The V-statistics's standard deviation becomes:

- $\sigma_V = \sqrt{\frac{n(n+1)(2n+1)}{24} - \sum_i^k \frac{t_i^3 - t_i}{48}}$  (cf., slide 9)

- The W-statistics's standard deviation becomes:

- $\sigma_W = \sqrt{\frac{n_1 n_2}{12} \left( (n_1 + n_2 + 1) - \sum_i^k \frac{t_i^3 - t_i}{(n_1 + n_2)(n_1 + n_2 - 1)} \right)}$  (cf., slide 16)

- And the H-statistics can be corrected by dividing H by the term  $corr = 1 - \frac{\sum_i^k (t_i^3 - t_i)}{N^3 - N}$

- In the example:  $corr = 1 - \frac{(4^3 - 4) + (8^3 - 8) + (3^3 - 3)}{(15^3 - 15)}$

- The corrected H value then is  $H_{corr} = 6.56$

- Because all this is rather tedious, you are allowed to skip these corrections in your assignments (also in the exam).

## Current State of our Toolkit

- Categorical Scale
  - $\chi^2$ -statistics ( $\chi^2$ -distributed)
- Interval Scale
  - Variance known: z-statistics (normally distributed)
  - Variance unknown (but equal): t-statistics (Student's t distribution), F-statistics (F-distributed)
- Ordinal Scale
  - W-, V-statistics (both normally distributed), H-statistics ( $\chi^2$ -distributed)

## What comes next

- We started out defining four types of hypotheses

- 1 Directional difference hypotheses
- 2 Undirectional difference hypotheses
- 3 Directional relationship hypotheses
- 4 Undirectional relationship hypotheses

- We can so far only deal with (1) and (2). This is going to be fixed during the next statistics block starting from January 16th. The timeline:

- **Next:** We will have a mock exam on January 9th.
  - **NextNext:** We will have another Reading Session on January 11.
  - **NextNextNext:** Correlation and Regression for testing relationship hypotheses from January 16th.
  - **New Exam Date: February 21th!**

## Merry Christmas!



# Sketches

Intentionally left blank :-)

