Lecture

Data Science

Statistics Foundations

JProf. Dr. Claudia Wagner
Learning Goals

- Last Time
  - Hypothesis Testing
  - Power Analysis
  - Effect Sizes

- Today
  - How to deal with uncertainty?
  - Non parametric alternatives for hypothesis testing
  - Likelihood & MLE
CONFIDENCE INTERVALS
Uncertainty

- Estimate true population parameter from sampling distribution
- Estimate true difference in means between two population by looking at difference in means of sampling distributions
- We have estimates → i.e. we have uncertainty

How to quantify uncertainty?
  - Confidence Intervals are a statement about the percentage of future confidence intervals that will contain the true population parameter (or the true difference in populations parameters)
Interpretation CI

We take a random sample of 1000 fb user.
The mean age in the sample is 22, the 95% CI is [20,24]

1. The true mean age of fb users is between 20 and 24.
2. If we pick a fb user randomly his age will be between 20 and 24 with 95% probability.
3. There is a 95% probability that the true population mean lies between 20 and 24.
4. We are 95% confident that the true mean age of fb users lies between 20 and 24.
5. 95% of future sample means +/- 2 will contain the true mean age of fb user
A CI always contains the true parameter or not!
In the long run 95% of confidence intervals will contain the true population parameter.
CI = \bar{X} \pm MOE

MOE = SE \times z_{\alpha/2}
Example: we observe 6 heads in 10 tosses.
- What is the parameter p of my coin? 95% CI required!

- \( \hat{p} = 6/10 = 0.6 \) MLE point estimate \( \hat{p} \) of p

- MOE = critical value * SE

- \( MOE = z_{\alpha/2} \sqrt{\frac{\hat{p}(1-\hat{p})}{n}} = 1.96 \sqrt{\frac{0.6*0.4}{10}} = 0.30 \)

- CI = \( \bar{X} \) +/- MOE

- CI [0.3; 0.9]
Example: we observe 600 heads in 1000 tosses.
- What is the parameter $p$ of my coin? 95% CI required!

- $\hat{p} = \frac{600}{1000} = 0.6$ MLE point estimate $\hat{p}$ of $p$

- MOE = critical value * SE

- $MOE = z_{\frac{1-0.95}{2}} \sqrt{\frac{\hat{p}(1-\hat{p})}{n}} = 1.96 \sqrt{\frac{0.6 \times 0.4}{1000}} = 0.030$

- CI [0.57; 0.63] CI is narrower now
**Example:** we observe 900 heads in 1000 tosses.

- What is the parameter $p$ of my coin? 95% CI required!

- $\hat{p} = \frac{900}{1000} = 0.9$  
  MLE point estimate $\hat{p}$ of $p$

- MOE = critical value $\ast$ SE

- $MOE = Z_{\alpha/2} \sqrt{\frac{\hat{p}(1-\hat{p})}{n}}$

- $1.96 \sqrt{\frac{0.9 \ast 0.1}{1000}} = 0.019$

- CI [0.881; 0.919]
Example: CI for Mean

- Select 1000 random users from fb
- Mean age \( \bar{X} = 25 \)
- Standard deviation of population: 30
- Compute confidence level is 99% for true age

\[ CI = \bar{X} +/- MOE \]

\[ MOE = \frac{\sigma}{\sqrt{n}} * z_{\alpha/2} \]

- \( MOE = \frac{30}{\sqrt{1000}} * 2.58 = 2.45 \)
- \( 25 +/- 2.45 \) CI: [22.55 ; 27.45]
CI Example

- Sample of 64 students, mean commute time to campus 35km
- Std of commute time in population is 5

1. Construct 95% CI
2. What if I can only accept MOE = 0.01? What sample size would we need?

- n=64 and $\bar{X} = 35$ and $\sigma = 5$

- MOE = $SE \cdot z_{\alpha/2} = \frac{5}{\sqrt{64}} \cdot 1.96 = 1.24$

- $35 \pm 1.24$  
  CI: $[33.86, 36.24]$
A newspaper selected a random sample of 16 readers. They asked whether the paper should increase its coverage of local news. 40% of the sample want more local news.

What is the 99% confidence interval for the proportion of readers who would like more coverage of local news?

- Sample proportion ($\hat{p} = 0.40$)

- MOE = critical value * SE
- critical value $\rightarrow$ alpha = 1% $\rightarrow$ alpha/2 = 0.5%
- Use t-distribution with 15 degree of freedom $\rightarrow$ critical value is 2.9467
CI for proportions

\[
SE = \sqrt{\frac{\hat{p} \times (1 - \hat{p})}{n}} = \sqrt{\frac{0.6 \times 0.4}{16}} = 0.1225
\]

MOE = critical value * SE = 2.9467*0.1225 = 0.361

99% confidence interval for \(\hat{p}\) =0.40:

[0.039, 0.791]
NON-PARAMETRIC TESTS
- **Parametric tests**
  - Make assumptions about parameter of sampling distribution
  - For large sample sizes sampling distribution of many statistics will be normal

- **Nonparametric tests**
  - Are often based on ranks rather than the raw data.
  - Often use the median rather than the mean
  - Tend to be more robust to outliers
  - But are less powerful
  - Nonparametric tests are distribution-free but not assumption free!
Example: t-test comparing two sets of measurements

- Sample 1: 1 2 3 4 5 6 7 8 9 10
- Sample 2: 7 8 9 10 11 12 13 14 15 16 17 18 19 20
- Sample averages: 5.5 and 13.5
- T-test $p = 0.000019$

- Sample 2: 7 8 9 10 11 12 13 14 15 16 17 18 19 20 200
- Sample averages: 5.5 and 25.9
- T-test $p = 0.12$

$$t = \frac{\bar{x}_1 - \bar{x}_2 - \Delta}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}}$$
Overview

- One sample mean test
  - Parametric: one sample t-test or z-test
  - Non-Parametric: sign test for median

- Two dependent samples (paired test)
  - Parametric test: paired t-test or z-test
  - Non-Parametric test: Wilcoxon signed rank test

- Two independent samples (unpaired test)
  - Parametric: two sample unpaired t-test or z-test
  - Non-Parametric: Mann Withney U test
One sample Test (Sign Test)

- Test if the median \( m \) takes on a particular value \( m_0 \)
  - \( H_0: m = m_0 \)
  - \( H_A: m > m_0 \) or \( m < m_0 \) or \( m \neq m_0 \)

- Number of positive and negative differences between \( X_i \) and \( m_0 \) follows a binomial distribution with \( p = 0.5 \) under \( H_0: m = m_0 \):
  \[
  N_- \sim b\left(n, \frac{1}{2}\right) \quad \text{and} \quad N_+ \sim b\left(n, \frac{1}{2}\right)
  \]

https://onlinecourses.science.psu.edu/stat414/node/318
Example

- Sample data (daily study times in minutes):
  - 88, 70, 66, 55, 52, 50, 45, 43, 40, 39

- Someone claims median study time was 60 minutes.

- We believe that the true median study time was shorter

- $H_0: m = m_0$
- $H_A: m < m_0$
We observe 3 positive signs out of 10

How surprising is this observation under H0: $N_+ \sim \text{Binom}(n, 0.5)$?

The probability of observing $k$ heads in 10 tosses of a fair coin, with $p(\text{heads}) = 0.5$

$$P(X \leq 3) = 0.1719$$
Non Parametric 2-sample Tests

- Is the weight of men and women significantly different?

- No Systematic Rank Difference

- Systematic Rank Difference

- Works for ordinal data!
### Test Group Difference (unpaired)

<table>
<thead>
<tr>
<th>Men</th>
<th>Women</th>
</tr>
</thead>
<tbody>
<tr>
<td>88.5</td>
<td>66</td>
</tr>
<tr>
<td>96</td>
<td>55</td>
</tr>
<tr>
<td>62</td>
<td>60</td>
</tr>
<tr>
<td>71.5</td>
<td>62</td>
</tr>
<tr>
<td>84</td>
<td>80</td>
</tr>
<tr>
<td>70.5</td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Gender</th>
<th>Score</th>
<th>Rank</th>
</tr>
</thead>
<tbody>
<tr>
<td>W</td>
<td>55</td>
<td>1</td>
</tr>
<tr>
<td>W</td>
<td>60</td>
<td>2</td>
</tr>
<tr>
<td>W</td>
<td>62</td>
<td>3</td>
</tr>
<tr>
<td>M</td>
<td>62</td>
<td>4</td>
</tr>
<tr>
<td>W</td>
<td>66</td>
<td>5</td>
</tr>
<tr>
<td>W</td>
<td>70.5</td>
<td>6</td>
</tr>
<tr>
<td>M</td>
<td>71.5</td>
<td>7</td>
</tr>
<tr>
<td>W</td>
<td>80</td>
<td>8</td>
</tr>
<tr>
<td>M</td>
<td>84</td>
<td>9</td>
</tr>
<tr>
<td>M</td>
<td>88.5</td>
<td>10</td>
</tr>
<tr>
<td>M</td>
<td>96</td>
<td>11</td>
</tr>
</tbody>
</table>

Median: 70.5
### Example: Mann-Whitney U test

<table>
<thead>
<tr>
<th></th>
<th></th>
<th>Rank</th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>W</td>
<td>55</td>
<td>1</td>
<td></td>
<td></td>
</tr>
<tr>
<td>W</td>
<td>60</td>
<td>2</td>
<td></td>
<td></td>
</tr>
<tr>
<td>W</td>
<td>62</td>
<td>3</td>
<td></td>
<td></td>
</tr>
<tr>
<td>M</td>
<td>63</td>
<td>4</td>
<td></td>
<td></td>
</tr>
<tr>
<td>W</td>
<td>66</td>
<td>5</td>
<td></td>
<td></td>
</tr>
<tr>
<td>W</td>
<td>70.5</td>
<td>6</td>
<td></td>
<td></td>
</tr>
<tr>
<td>M</td>
<td>71.5</td>
<td>7</td>
<td></td>
<td></td>
</tr>
<tr>
<td>W</td>
<td>80</td>
<td>8</td>
<td></td>
<td></td>
</tr>
<tr>
<td>M</td>
<td>84</td>
<td>9</td>
<td></td>
<td></td>
</tr>
<tr>
<td>M</td>
<td>88.5</td>
<td>10</td>
<td></td>
<td></td>
</tr>
<tr>
<td>M</td>
<td>96</td>
<td>11</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

#### Median

- Men: 88.5
- Women: 66

#### Mean

<table>
<thead>
<tr>
<th></th>
<th>Mean Rank</th>
<th>Rank Sum</th>
</tr>
</thead>
<tbody>
<tr>
<td>Men</td>
<td>8.2</td>
<td>41</td>
</tr>
<tr>
<td>Women</td>
<td>4.2</td>
<td>25</td>
</tr>
</tbody>
</table>

- Median: 70.5
- How many data points in one group have higher rank than data points in the other group?

- Degree of overlap of ranks between 2 groups
Degree of overlap of rank between 2 groups

High Rank

Women

U=0

Low Rank

Men
Example: Mann-Whitney U test

<table>
<thead>
<tr>
<th></th>
<th>Mean Rank</th>
<th>Rank Sum</th>
</tr>
</thead>
<tbody>
<tr>
<td>Men</td>
<td>8.2</td>
<td>41</td>
</tr>
<tr>
<td>Women</td>
<td>4.2</td>
<td>25</td>
</tr>
</tbody>
</table>

Mean Rank

<table>
<thead>
<tr>
<th>Rank</th>
<th>Mean</th>
</tr>
</thead>
<tbody>
<tr>
<td>W</td>
<td>55</td>
</tr>
<tr>
<td>W</td>
<td>60</td>
</tr>
<tr>
<td>W</td>
<td>62</td>
</tr>
<tr>
<td>M</td>
<td>63</td>
</tr>
<tr>
<td>W</td>
<td>66</td>
</tr>
<tr>
<td>W</td>
<td>70.5</td>
</tr>
<tr>
<td>M</td>
<td>71.5</td>
</tr>
<tr>
<td>W</td>
<td>80</td>
</tr>
<tr>
<td>M</td>
<td>84</td>
</tr>
<tr>
<td>M</td>
<td>88.5</td>
</tr>
<tr>
<td>M</td>
<td>96</td>
</tr>
</tbody>
</table>

Max rank sum for n1 items

\[ U_1 = R_1 - \frac{n_1(n_1 + 1)}{2} = 4 \]

\[ U_2 = R_2 - \frac{n_2(n_2 + 1)}{2} = 26 \]

\[ U_1 + U_2 = n1 \times n2 \]
We would need to observe $U \leq 3$ in order to reject $H_0$ (no difference between the 2 groups) at significance level of 5%.
U-Statistic

- **U-Statistic**: the **minimal overlap** we expect to see if both groups come from the same distribution and have size $N_1$ and $N_2$.

- If $U_{obs} \leq U_{critical}(n_1,n_2) \rightarrow$ reject $H_0$

- **Concordance Probability (Effect Size)**

  \[ C = \frac{\bar{R} - \frac{n_1+1}{2}}{n_2} \]

  For Women: \[ (4.2 - 3.5) / 5 = 0.14 \]

  The probability that a randomly chosen female has a value greater than a randomly chosen male is 0.14
Example: Mann-Whitney U test

<table>
<thead>
<tr>
<th></th>
<th>Rank Sum</th>
</tr>
</thead>
<tbody>
<tr>
<td>Men</td>
<td>30</td>
</tr>
<tr>
<td>Women</td>
<td>36</td>
</tr>
</tbody>
</table>

\[
U_1 = R_1 - \frac{n_1(n_1 + 1)}{2} = 15
\]

\[
U_2 = R_2 - \frac{n_2(n_2 + 1)}{2} = 15
\]

\[
U_1 + U_2 = n_1 * n_2
\]
Mann Whitney U Test (Rank Sum Test)

- Test the null hypothesis that two samples come from the same population

<table>
<thead>
<tr>
<th>Rating</th>
<th>Rating</th>
</tr>
</thead>
<tbody>
<tr>
<td>3</td>
<td>9</td>
</tr>
<tr>
<td>4</td>
<td>7</td>
</tr>
<tr>
<td>2</td>
<td>5</td>
</tr>
<tr>
<td>6</td>
<td>10</td>
</tr>
<tr>
<td>2</td>
<td>6</td>
</tr>
<tr>
<td>5</td>
<td>8</td>
</tr>
<tr>
<td>3.5</td>
<td>7.5</td>
</tr>
</tbody>
</table>
Mann Whitney U Test

- Create combined rank and sum rankings per group

<table>
<thead>
<tr>
<th>Rating</th>
<th>Rank</th>
</tr>
</thead>
<tbody>
<tr>
<td>3</td>
<td>3</td>
</tr>
<tr>
<td>4</td>
<td>4</td>
</tr>
<tr>
<td>2</td>
<td>1.5</td>
</tr>
<tr>
<td>6</td>
<td>7.5</td>
</tr>
<tr>
<td>2</td>
<td>1.5</td>
</tr>
<tr>
<td>5</td>
<td>5.5</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Rating</th>
<th>Rank</th>
</tr>
</thead>
<tbody>
<tr>
<td>9</td>
<td>11</td>
</tr>
<tr>
<td>7</td>
<td>9</td>
</tr>
<tr>
<td>5</td>
<td>5.5</td>
</tr>
<tr>
<td>10</td>
<td>12</td>
</tr>
<tr>
<td>6</td>
<td>7.5</td>
</tr>
<tr>
<td>8</td>
<td>10</td>
</tr>
</tbody>
</table>
### Mann Whitney U Test (Rank Sum Test)

**U1 = 23 - \(\frac{6 \times 7}{2}\) = 2**  
**U2 = 55 - \(\frac{6 \times 7}{2}\) = 34**

<table>
<thead>
<tr>
<th>Rating</th>
<th>Rank</th>
<th>Rating</th>
<th>Rank</th>
</tr>
</thead>
<tbody>
<tr>
<td>3</td>
<td>3</td>
<td>9</td>
<td>11</td>
</tr>
<tr>
<td>4</td>
<td>4</td>
<td>7</td>
<td>9</td>
</tr>
<tr>
<td>2</td>
<td>1.5</td>
<td>5</td>
<td>5.5</td>
</tr>
<tr>
<td>6</td>
<td>7.5</td>
<td>10</td>
<td>12</td>
</tr>
<tr>
<td>2</td>
<td>1.5</td>
<td>6</td>
<td>7.5</td>
</tr>
<tr>
<td>5</td>
<td>5.5</td>
<td>8</td>
<td>10</td>
</tr>
</tbody>
</table>

\[ \sum 23 \quad \sum 55 \]
Use a table of critical U values for the Mann-Whitney U Test
For two-tailed test. 5% significance level.

<table>
<thead>
<tr>
<th>N2</th>
<th>5</th>
<th>6</th>
<th>7</th>
<th>8</th>
<th>9</th>
<th>10</th>
<th>11</th>
<th>12</th>
<th>13</th>
<th>14</th>
<th>15</th>
<th>16</th>
<th>17</th>
<th>18</th>
<th>19</th>
<th>20</th>
</tr>
</thead>
<tbody>
<tr>
<td>2</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>2</td>
<td>2</td>
<td>2</td>
<td>2</td>
<td>2</td>
<td>2</td>
<td>2</td>
</tr>
<tr>
<td>3</td>
<td>0</td>
<td>1</td>
<td>1</td>
<td>2</td>
<td>2</td>
<td>3</td>
<td>3</td>
<td>4</td>
<td>4</td>
<td>4</td>
<td>5</td>
<td>5</td>
<td>6</td>
<td>6</td>
<td>7</td>
<td>7</td>
</tr>
<tr>
<td>4</td>
<td>0</td>
<td>1</td>
<td>2</td>
<td>3</td>
<td>4</td>
<td>4</td>
<td>4</td>
<td>5</td>
<td>6</td>
<td>7</td>
<td>9</td>
<td>9</td>
<td>10</td>
<td>11</td>
<td>11</td>
<td>12</td>
</tr>
<tr>
<td>5</td>
<td>2</td>
<td>3</td>
<td>5</td>
<td>6</td>
<td>7</td>
<td>8</td>
<td>8</td>
<td>9</td>
<td>11</td>
<td>12</td>
<td>13</td>
<td>14</td>
<td>15</td>
<td>17</td>
<td>17</td>
<td>18</td>
</tr>
<tr>
<td>5</td>
<td>6</td>
<td>8</td>
<td>10</td>
<td>11</td>
<td>13</td>
<td>14</td>
<td>14</td>
<td>16</td>
<td>17</td>
<td>19</td>
<td>21</td>
<td>22</td>
<td>24</td>
<td>25</td>
<td>27</td>
<td>27</td>
</tr>
<tr>
<td>7</td>
<td>8</td>
<td>10</td>
<td>12</td>
<td>14</td>
<td>16</td>
<td>18</td>
<td>20</td>
<td>22</td>
<td>24</td>
<td>26</td>
<td>28</td>
<td>30</td>
<td>32</td>
<td>34</td>
<td>34</td>
<td>34</td>
</tr>
<tr>
<td>8</td>
<td>13</td>
<td>15</td>
<td>17</td>
<td>19</td>
<td>22</td>
<td>24</td>
<td>26</td>
<td>29</td>
<td>31</td>
<td>34</td>
<td>36</td>
<td>38</td>
<td>41</td>
<td>41</td>
<td>41</td>
<td>41</td>
</tr>
<tr>
<td>9</td>
<td>17</td>
<td>20</td>
<td>23</td>
<td>26</td>
<td>28</td>
<td>31</td>
<td>34</td>
<td>37</td>
<td>39</td>
<td>42</td>
<td>45</td>
<td>48</td>
<td>50</td>
<td>50</td>
<td>50</td>
<td>50</td>
</tr>
<tr>
<td>10</td>
<td>23</td>
<td>26</td>
<td>29</td>
<td>32</td>
<td>35</td>
<td>38</td>
<td>42</td>
<td>45</td>
<td>48</td>
<td>52</td>
<td>55</td>
<td>58</td>
<td>61</td>
<td>61</td>
<td>61</td>
<td>61</td>
</tr>
<tr>
<td>11</td>
<td>30</td>
<td>33</td>
<td>37</td>
<td>40</td>
<td>44</td>
<td>47</td>
<td>51</td>
<td>55</td>
<td>58</td>
<td>62</td>
<td>66</td>
<td>70</td>
<td>73</td>
<td>73</td>
<td>73</td>
<td>73</td>
</tr>
<tr>
<td>12</td>
<td>37</td>
<td>41</td>
<td>45</td>
<td>49</td>
<td>53</td>
<td>57</td>
<td>61</td>
<td>65</td>
<td>69</td>
<td>73</td>
<td>77</td>
<td>81</td>
<td>85</td>
<td>85</td>
<td>85</td>
<td>85</td>
</tr>
<tr>
<td>13</td>
<td>45</td>
<td>50</td>
<td>54</td>
<td>59</td>
<td>63</td>
<td>67</td>
<td>72</td>
<td>76</td>
<td>80</td>
<td>84</td>
<td>88</td>
<td>92</td>
<td>96</td>
<td>96</td>
<td>96</td>
<td>96</td>
</tr>
<tr>
<td>14</td>
<td>55</td>
<td>59</td>
<td>64</td>
<td>69</td>
<td>74</td>
<td>78</td>
<td>83</td>
<td>87</td>
<td>92</td>
<td>97</td>
<td>101</td>
<td>105</td>
<td>110</td>
<td>110</td>
<td>110</td>
<td>110</td>
</tr>
</tbody>
</table>

5 represents the smallest value we could expect to obtain for U if H0 was true and n1=n2=6

2 < 5 → Difference between ratings is significant with p < 0.05
Effect Size

\[ C = \frac{\bar{R} - \frac{n_1+1}{2}}{n_2} \]

- Mean Rank green book = 55/6
- \( n_2 = 6 \), \( n_2 = 6 \)
- \( C = 0.94 \)
- Probability that green book is higher ranked than red book is 0.94

| 1.5 | 1.5 | 3 | 4 | 5.5 | 5.5 | 7.5 | 7.5 | 9 | 10 | 11 | 12 |
Summary

- U-Table tells us what overlap we should expect to see if we sample from the same distribution → which value would be considered as too extreme

- Alternatives? Assume you never heard about U-statistic. What could you do?
Alternative: Simulations

- Discard the group levels and permute the data
- Compute difference in rank sum between the 2 groups after permutation
- Repeat doing that → what difference in the rank sum do you expect to observe by chance?
- How extreme is the difference you observe?
Test Group Difference (paired)

Two sample Wilcoxon signed rank test

Example:
- Ten costumers get 2 products and need to rate them
- We want to test if costumers prefer product A over B.
  - $H_0: m_A - m_B > 0$
  - $H_A: m_A - m_B \leq 0$ (left tail test)

- Number of costumers preferred B: 7 (- sign)
- Number of costumers preferred A: 2 (+ sign)
- Number of ties: 1

- $P(8 \text{ or } 9 \text{ heads in } 9 \text{ flips of a fair coin}) = 0.0195$
Kendall rank Tau-A

- Measures the similarity of two variables when their values are ranked → can be used to determine effect size of a Wilcoxon signed rank

\[ \tau = \frac{\text{num concordant pairs} - \text{num discordant pairs}}{\frac{1}{2}n(n-1)} \]

- Perfect agreement between the two rankings → coef=1
- Perfect disagreement between the two rankings → coef=-1
- If X and Y are independent → coef=0
Kendall rank Tau-A

- **Example:**

<table>
<thead>
<tr>
<th>Product</th>
<th>Rating G1</th>
<th>Rating G2</th>
</tr>
</thead>
<tbody>
<tr>
<td>s1</td>
<td>4</td>
<td>1</td>
</tr>
<tr>
<td>s2</td>
<td>5</td>
<td>2</td>
</tr>
<tr>
<td>s3</td>
<td>3</td>
<td>4</td>
</tr>
</tbody>
</table>

- **Concordant pairs:**
  - (s1, s2)

- **Discordant pairs:**
  - (s2, s3), (s1, s4)

\[
\tau = \frac{\text{num concordant pairs} - \text{num discordant pairs}}{1/2 n(n - 1)}
\]

\[
\tau = \frac{1-2}{1/2^3(3-1)} = -1/3
\]
Kendall Tau-B

- Correct for ties

\[
\tau_B = \frac{(\text{number of concordant pairs}) - (\text{number of discordant pairs})}{\sqrt{N_1} \times \sqrt{N_2}}
\]

N1: number of pairs not tied in group 1
N2: number of pairs not tied in group 2

<table>
<thead>
<tr>
<th></th>
<th>Rating G1</th>
<th>Rating G2</th>
</tr>
</thead>
<tbody>
<tr>
<td>Product s1</td>
<td>5</td>
<td>1</td>
</tr>
<tr>
<td>Product s2</td>
<td>5</td>
<td>2</td>
</tr>
<tr>
<td>Product s3</td>
<td>3</td>
<td>4</td>
</tr>
</tbody>
</table>

Discordant pairs: (s2, s3), (s1, s4)

\[ N_1=2 \text{ and } N_2=3 \]

\[ \tau = \frac{-2}{\sqrt{2*3}} = -0.81 \]
PARAMETER ESTIMATION
Population mean is called parameter.

Sample mean is called sample statistic.

Find „good“ estimator.
- The **average of X** in a sample of size N is a sample statistic. In each sample it is different.

- The **expected value of X** in the population is a parameter. It's always the same.

- The **sample statistic** can be a good estimator of the parameter if N is large enough (**Law of Large Numbers**).
How to find the parameter?

- We use the **likelihood**
- The likelihood gives the **function of a parameter given the observed data**
  - How likely is the data given different values of the parameter?

For discrete probability distributions the likelihood ranges between 0 and 1.

For continuous distributions it goes from 0 to infinity. Likelihood can then only be interpreted in relative terms!
Likelihood

Likelihood: 8 out of 10

θ
Likelihood Ratios

- Which parameter is more likely?

\[
\frac{L(H_1|D)}{L(H_2|D)}
\]

P=0.8 is almost 7 times more likely than p=0.5
- Likelihood ratios expose the relative evidence of $H_0$ compared to $H_1$

- BUT: both hypothesis can be unlikely
Example

- We conduct 3 studies
- In 2 we find a significant effect. The power of the test is 80% and our alpha level is 0.05

- How likely is it that there was a true effect?

- If there was a true effect:
  - $0.8 \times 0.8 \times 0.2 = 0.128$

- If there was no true effect:
  - $0.05 \times 0.05 \times 0.8 = 0.0024$
Maximum Likelihood Estimation

Given data $D = (X_1, X_2, \ldots, X_n)$
Assume a set of distributions $\{P_\theta: \theta \in \Theta\}$
Assume $X_1, X_2, \ldots, X_n$ is a iid sample from $P_\theta$ for one specific $\theta$

Goal: estimate the true $\theta$ that $D$ comes from

$$
\theta_{MLE} = \underset{\theta \in \Theta}{\arg \max} P(D|\theta)
$$

$$
P(D|\theta) = \prod_{i=1}^{n} P(x_i|\theta)
$$