Advanced Techniques for Mobile Robotics

Statistical Testing

Wolfram Burgard, Cyrill Stachniss, Kai Arras, Maren Bennewitz
Statistical Testing for Evaluating Experiments

- Deals with the relationship between the **value** of data, its **variance**, and the **confidence** of a conclusion

A typical situation:
- Existing technique A
- You developed a new technique B
- Key question: Is B better than A?
Evaluating Experiments

- Define a performance measure, e.g.
  - Run-time
  - Error
  - Accuracy
  - Robustness (success rate, MTBF, ...)

- Collect data \( d \)

- Run both techniques on the data \( d \)

- How to compare the obtained results \( A(d), B(d) \)?
**1st Example**

**Scenario**
- A, B are two path planning techniques
- Score is the planning time
- Data d is a given map, start and goal pose

**Example**
- \(A(d) = 0.5\) s
- \(B(d) = 0.6\) s

What does that mean?
2nd Example

- Same scenario but four tasks

Example

- A(d) = 0.5 s, 0.4 s, 0.6 s, 0.4 s
- B(d) = 0.4 s, 0.3 s, 0.6 s, 0.5 s

What does that mean?
2\textsuperscript{nd} Example

- Same scenario but four tasks

Example

- $A(d) = 0.5 \text{ s}, 0.4 \text{ s}, 0.6 \text{ s}, 0.4 \text{ s}$
- $B(d) = 0.4 \text{ s}, 0.3 \text{ s}, 0.6 \text{ s}, 0.5 \text{ s}$

Mean of the planning time is

- $\mu_A = 1.9 \text{ s}/4 = 0.475 \text{ s}$
- $\mu_B = 1.8 \text{ s}/4 = 0.45 \text{ s}$

Is $B$ really better than $A$?
Is B better than A?

- $\mu_A = 0.475 \text{ s}, \mu_B = 0.45 \text{ s}$
- $\mu_A > \mu_B$, so B is better than A?!
- We just evaluated four tests, thus $\mu_A$ and $\mu_B$ are rough estimates only
- We saw too few data to make statements with high confidence

How can we make a confident statement that B is better than A?
Hypothesis Testing

- “Answer a yes-no question about a population and assess that the answer is wrong.” [Cohen’ 95]

- Example: To test that B is different from A, assume they are truly equal. Then, assess the probability of the obtained result. If the probability is small, reject the hypothesis.
The Null Hypothesis $H_0$

- The null hypothesis is the hypothesis that one wants to reject by analyzing data (from experiments)
- $H_0$ is the default state
- A statistical test can never proof $H_0$
- A statistical test can only reject or fail to reject $H_0$
- Example: to show that method A is better than B, use $H_0: A=B$
Typical Null Hypotheses

- Typical null and alternative hypotheses

\[ H_0 : \mu = 0 \]
\[ H_1 : \mu \neq 0 \] (two-tailed test)
\[ H_1 : \mu < 0 \] (one-tailed test)
\[ H_1 : \mu > 0 \] (one-tailed test)
\[ H_0 : \mu_1 = \mu_2 \]
\[ H_1 : \mu_1 \neq \mu_2 \] (two-tailed test)
\[ H_1 : \mu_1 < \mu_2 \] (one-tailed test)
\[ H_1 : \mu_1 > \mu_2 \] (one-tailed test)
Population and Sample

- The data we observe is often only a small fraction of the possible outcomes

- Population = set of potential measurements, values, or outcomes
- Sample = the data we observe
- Sampling distribution = distribution of possible samples given a fixed sample size
Sampling Distribution

- A sampling distribution is the distribution of a statistics calculated from all possible samples of a given size, drawn from a given population.

- Example: Toss a coin twice
Sampling Distribution

- Sampling distributions are rather theoretical entities
- Distributions of all possible samples are likely to be large or infinite
- Very few closed form solutions only
- However, one can compute empirical sampling distributions based on a set of samples
Central Limit Theorem

- The sampling distribution of the mean of samples of size $N$ approaches a normal distribution as $N$ increases.
- If the samples are drawn from a population with mean $\mu$ and standard deviation $\sigma$, then the mean of the sampling distribution is $\mu$ with standard deviation $\sigma/N^{0.5}$.
- These statements hold irrespectively of the shape of the population distribution from which the samples are drawn.
p(one sample)

\[ \mu = 0 \]
\[ \sigma = 1.45 \]

[Illustration of the central limit theorem, Wikipedia]
p(average of two samples)

[Illustration of the central limit theorem, Wikipedia]
p(average of three samples)

[Illustration of the central limit theorem, Wikipedia]
p(average of four samples)

[Illustration of the central limit theorem, Wikipedia]
Standard Error of the Mean

- Standard deviation of the sampling distribution of the mean is often called **standard error (of the mean)**, SE.

- Central limit theorem: \( \lim_{N \to \infty} \frac{\bar{x}}{\sqrt{N}} = \mu \)

- The standard error represents the uncertainty about the mean and is given by \( \sigma_{\bar{x}} = \sigma / \sqrt{N} \ ( = SE) \)
The Normal Distribution

Probability of Cases in portions of the curve

Standard Deviations From The Mean

Cumulative %

Z Scores
Z Score

- Z score indicates how many standard deviations an observation $x$ is above or below the mean

\[ Z = \frac{x - \mu}{\sigma} \]

- Z table provides the probability for this event
  - $Z < 3 : p = 99.9\%$
  - $Z < 0 : p = 50\%$
  - $Z < -1 : p = 15.9\%$
  - $-2 < Z < -2 : p = \sim 95\%$
One Sample Z-Test

- One sample location test
- Given a $\mu$ and $\sigma$ of a population
- Test if a sample (from the population) has a significantly different mean than the population
- Sample of size N
- Compute the Z score $Z = \frac{\bar{x} - \mu}{SE}$
- Look up the Z score in a Z table to obtain the probability that the sample
Z-Test Example

- Scores of all German students in a test
- In Germany: \( \mu = 100, \sigma = 12 \)
- A sample of 55 students in Freiburg obtained an average score of 96
- Null hypothesis: Students from Freiburg are as good as the average German?

\[
SE = \frac{\sigma}{\sqrt{N}} = \frac{12}{\sqrt{55}} \approx 1.62
\]

\[
Z = \frac{\bar{x} - \mu}{SE} = \frac{96 - 100}{1.62} = -2.47
\]

- Z-table: the probability of observing a value below -2.47 is approximately 0.68%
- Reject the null hypothesis
**Z-Test: Assumptions**

- Independently generated samples
- Mean and variance of the population distribution are known
- Sampling distribution approx. normal (population distributions normal or large N)
- The sample set is sufficiently large (N>~30)

**Comments**

- Often, $\sigma$ can be approximated using the variance in the sample set
- In practice, the size of the sample set is often too small for the Z-Test
When N is Small: t-Test

Relax and have a Guinness! 😊

- Test to cheaply monitor the quality of stout at Guinness brewery (~1908)

William Sealy Gosset
When N is Small: t-Test

- Variant of the Z-Test for N<30
- Instead of the Normal distribution, it uses the t-distribution
- The t-distribution is the sampling distribution for the mean for small N under the assumption that the population is normally distributed
- t-distribution is similar to a normal distribution but has bigger tails
t-Distribution

- The t-distribution depends on $N$
- For large $N$, it approaches a normal
One Sample t-Test

- t-value is similar to the Z value
  \[ t = \frac{\bar{x} - \mu}{\frac{s}{\sqrt{N}}} = \frac{\bar{x} - \mu}{s/\sqrt{N}} \]

- The t-value has to be compared to the values available in a t-table
- A t-table shows also a degree of freedom (DoF) which is closely related to the sample size (here: DoF=N-1)
<table>
<thead>
<tr>
<th>Degree of Freedom</th>
<th>Confidence Level 75%</th>
<th>80%</th>
<th>85%</th>
<th>90%</th>
<th>95%</th>
<th>97.5%</th>
<th>99%</th>
<th>99.5%</th>
<th>99.75%</th>
<th>99.9%</th>
<th>99.95%</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>1.000</td>
<td>1.376</td>
<td>1.963</td>
<td>3.078</td>
<td>6.314</td>
<td>12.71</td>
<td>31.82</td>
<td>63.66</td>
<td>127.3</td>
<td>318.3</td>
<td>636.6</td>
</tr>
<tr>
<td>2</td>
<td>0.816</td>
<td>1.061</td>
<td>1.386</td>
<td>1.886</td>
<td>2.920</td>
<td>4.303</td>
<td>6.966</td>
<td>9.295</td>
<td>14.09</td>
<td>22.33</td>
<td>31.60</td>
</tr>
<tr>
<td>3</td>
<td>0.765</td>
<td>0.978</td>
<td>1.250</td>
<td>1.638</td>
<td>2.353</td>
<td>3.182</td>
<td>4.541</td>
<td>5.841</td>
<td>7.453</td>
<td>10.21</td>
<td>12.92</td>
</tr>
<tr>
<td>4</td>
<td>0.741</td>
<td>0.941</td>
<td>1.190</td>
<td>1.533</td>
<td>2.132</td>
<td>2.776</td>
<td>3.747</td>
<td>4.604</td>
<td>5.596</td>
<td>7.173</td>
<td>8.610</td>
</tr>
<tr>
<td>5</td>
<td>0.727</td>
<td>0.920</td>
<td>1.156</td>
<td>1.476</td>
<td>2.015</td>
<td>2.571</td>
<td>3.365</td>
<td>4.032</td>
<td>4.773</td>
<td>5.693</td>
<td>6.869</td>
</tr>
<tr>
<td>6</td>
<td>0.718</td>
<td>0.906</td>
<td>1.134</td>
<td>1.440</td>
<td>1.943</td>
<td>2.447</td>
<td>3.143</td>
<td>3.707</td>
<td>4.317</td>
<td>5.208</td>
<td>5.959</td>
</tr>
<tr>
<td>7</td>
<td>0.711</td>
<td>0.896</td>
<td>1.119</td>
<td>1.415</td>
<td>1.895</td>
<td>2.365</td>
<td>2.998</td>
<td>3.499</td>
<td>4.029</td>
<td>4.785</td>
<td>5.408</td>
</tr>
<tr>
<td>8</td>
<td>0.706</td>
<td>0.889</td>
<td>1.108</td>
<td>1.397</td>
<td>1.860</td>
<td>2.306</td>
<td>2.896</td>
<td>3.355</td>
<td>3.833</td>
<td>4.501</td>
<td>5.041</td>
</tr>
<tr>
<td>9</td>
<td>0.703</td>
<td>0.883</td>
<td>1.100</td>
<td>1.383</td>
<td>1.833</td>
<td>2.262</td>
<td>2.821</td>
<td>3.250</td>
<td>3.690</td>
<td>4.297</td>
<td>4.781</td>
</tr>
<tr>
<td>10</td>
<td>0.700</td>
<td>0.879</td>
<td>1.093</td>
<td>1.372</td>
<td>1.812</td>
<td>2.228</td>
<td>2.764</td>
<td>3.169</td>
<td>3.591</td>
<td>4.144</td>
<td>4.587</td>
</tr>
<tr>
<td>11</td>
<td>0.697</td>
<td>0.876</td>
<td>1.088</td>
<td>1.363</td>
<td>1.796</td>
<td>2.201</td>
<td>2.718</td>
<td>3.106</td>
<td>3.497</td>
<td>4.025</td>
<td>4.437</td>
</tr>
<tr>
<td>12</td>
<td>0.695</td>
<td>0.873</td>
<td>1.083</td>
<td>1.356</td>
<td>1.782</td>
<td>2.179</td>
<td>2.681</td>
<td>3.055</td>
<td>3.428</td>
<td>3.930</td>
<td>4.318</td>
</tr>
<tr>
<td>13</td>
<td>0.694</td>
<td>0.870</td>
<td>1.079</td>
<td>1.350</td>
<td>1.771</td>
<td>2.160</td>
<td>2.650</td>
<td>3.012</td>
<td>3.372</td>
<td>3.852</td>
<td>4.221</td>
</tr>
<tr>
<td>14</td>
<td>0.692</td>
<td>0.868</td>
<td>1.076</td>
<td>1.345</td>
<td>1.761</td>
<td>2.145</td>
<td>2.624</td>
<td>2.977</td>
<td>3.326</td>
<td>3.787</td>
<td>4.140</td>
</tr>
<tr>
<td>15</td>
<td>0.691</td>
<td>0.866</td>
<td>1.074</td>
<td>1.341</td>
<td>1.753</td>
<td>2.131</td>
<td>2.602</td>
<td>2.947</td>
<td>3.296</td>
<td>3.733</td>
<td>4.073</td>
</tr>
<tr>
<td>16</td>
<td>0.690</td>
<td>0.865</td>
<td>1.071</td>
<td>1.337</td>
<td>1.745</td>
<td>2.120</td>
<td>2.583</td>
<td>2.921</td>
<td>3.252</td>
<td>3.686</td>
<td>4.015</td>
</tr>
<tr>
<td>17</td>
<td>0.689</td>
<td>0.863</td>
<td>1.069</td>
<td>1.333</td>
<td>1.740</td>
<td>2.110</td>
<td>2.567</td>
<td>2.898</td>
<td>3.222</td>
<td>3.646</td>
<td>3.965</td>
</tr>
<tr>
<td>18</td>
<td>0.688</td>
<td>0.862</td>
<td>1.067</td>
<td>1.330</td>
<td>1.734</td>
<td>2.101</td>
<td>2.552</td>
<td>2.878</td>
<td>3.197</td>
<td>3.610</td>
<td>3.922</td>
</tr>
<tr>
<td>19</td>
<td>0.688</td>
<td>0.861</td>
<td>1.066</td>
<td>1.328</td>
<td>1.729</td>
<td>2.093</td>
<td>2.539</td>
<td>2.861</td>
<td>3.174</td>
<td>3.579</td>
<td>3.883</td>
</tr>
<tr>
<td>20</td>
<td>0.687</td>
<td>0.860</td>
<td>1.064</td>
<td>1.325</td>
<td>1.725</td>
<td>2.086</td>
<td>2.528</td>
<td>2.845</td>
<td>3.153</td>
<td>3.562</td>
<td>3.850</td>
</tr>
</tbody>
</table>
### t-Table 2/2

<table>
<thead>
<tr>
<th></th>
<th>0.687</th>
<th>0.860</th>
<th>1.064</th>
<th>1.325</th>
<th>1.725</th>
<th>2.066</th>
<th>2.528</th>
<th>2.845</th>
<th>3.153</th>
<th>3.552</th>
<th>3.850</th>
</tr>
</thead>
<tbody>
<tr>
<td>20</td>
<td>0.686</td>
<td>0.859</td>
<td>1.063</td>
<td>1.323</td>
<td>1.721</td>
<td>2.060</td>
<td>2.518</td>
<td>2.831</td>
<td>3.135</td>
<td>3.527</td>
<td>3.819</td>
</tr>
<tr>
<td>21</td>
<td>0.686</td>
<td>0.859</td>
<td>1.061</td>
<td>1.321</td>
<td>1.717</td>
<td>2.074</td>
<td>2.508</td>
<td>2.819</td>
<td>3.119</td>
<td>3.506</td>
<td>3.792</td>
</tr>
<tr>
<td>22</td>
<td>0.685</td>
<td>0.858</td>
<td>1.060</td>
<td>1.319</td>
<td>1.714</td>
<td>2.069</td>
<td>2.500</td>
<td>2.807</td>
<td>3.104</td>
<td>3.485</td>
<td>3.767</td>
</tr>
<tr>
<td>23</td>
<td>0.685</td>
<td>0.857</td>
<td>1.059</td>
<td>1.318</td>
<td>1.711</td>
<td>2.064</td>
<td>2.492</td>
<td>2.797</td>
<td>3.091</td>
<td>3.467</td>
<td>3.745</td>
</tr>
<tr>
<td>24</td>
<td>0.684</td>
<td>0.856</td>
<td>1.058</td>
<td>1.316</td>
<td>1.708</td>
<td>2.060</td>
<td>2.485</td>
<td>2.787</td>
<td>3.078</td>
<td>3.450</td>
<td>3.725</td>
</tr>
<tr>
<td>25</td>
<td>0.684</td>
<td>0.856</td>
<td>1.056</td>
<td>1.314</td>
<td>1.706</td>
<td>2.056</td>
<td>2.479</td>
<td>2.779</td>
<td>3.067</td>
<td>3.435</td>
<td>3.707</td>
</tr>
<tr>
<td>26</td>
<td>0.684</td>
<td>0.855</td>
<td>1.055</td>
<td>1.311</td>
<td>1.704</td>
<td>2.052</td>
<td>2.473</td>
<td>2.771</td>
<td>3.057</td>
<td>3.421</td>
<td>3.690</td>
</tr>
<tr>
<td>27</td>
<td>0.683</td>
<td>0.855</td>
<td>1.054</td>
<td>1.311</td>
<td>1.701</td>
<td>2.048</td>
<td>2.467</td>
<td>2.763</td>
<td>3.047</td>
<td>3.408</td>
<td>3.674</td>
</tr>
<tr>
<td>28</td>
<td>0.683</td>
<td>0.854</td>
<td>1.053</td>
<td>1.310</td>
<td>1.699</td>
<td>2.045</td>
<td>2.462</td>
<td>2.756</td>
<td>3.038</td>
<td>3.396</td>
<td>3.659</td>
</tr>
<tr>
<td>29</td>
<td>0.683</td>
<td>0.854</td>
<td>1.052</td>
<td>1.310</td>
<td>1.697</td>
<td>2.042</td>
<td>2.457</td>
<td>2.750</td>
<td>3.030</td>
<td>3.385</td>
<td>3.646</td>
</tr>
<tr>
<td>30</td>
<td>0.681</td>
<td>0.851</td>
<td>1.050</td>
<td>1.303</td>
<td>1.684</td>
<td>2.021</td>
<td>2.423</td>
<td>2.704</td>
<td>2.971</td>
<td>3.307</td>
<td>3.551</td>
</tr>
<tr>
<td>40</td>
<td>0.679</td>
<td>0.849</td>
<td>1.047</td>
<td>1.299</td>
<td>1.676</td>
<td>2.009</td>
<td>2.403</td>
<td>2.678</td>
<td>2.937</td>
<td>3.261</td>
<td>3.496</td>
</tr>
<tr>
<td>60</td>
<td>0.679</td>
<td>0.848</td>
<td>1.045</td>
<td>1.296</td>
<td>1.671</td>
<td>2.000</td>
<td>2.390</td>
<td>2.660</td>
<td>2.915</td>
<td>3.232</td>
<td>3.460</td>
</tr>
<tr>
<td>80</td>
<td>0.678</td>
<td>0.846</td>
<td>1.043</td>
<td>1.293</td>
<td>1.664</td>
<td>1.990</td>
<td>2.374</td>
<td>2.639</td>
<td>2.887</td>
<td>3.195</td>
<td>3.416</td>
</tr>
<tr>
<td>100</td>
<td>0.677</td>
<td>0.845</td>
<td>1.042</td>
<td>1.290</td>
<td>1.660</td>
<td>1.984</td>
<td>2.364</td>
<td>2.625</td>
<td>2.871</td>
<td>3.174</td>
<td>3.390</td>
</tr>
<tr>
<td>120</td>
<td>0.677</td>
<td>0.845</td>
<td>1.041</td>
<td>1.289</td>
<td>1.658</td>
<td>1.980</td>
<td>2.358</td>
<td>2.617</td>
<td>2.860</td>
<td>3.160</td>
<td>3.373</td>
</tr>
<tr>
<td>∞</td>
<td>0.674</td>
<td>0.842</td>
<td>1.036</td>
<td>1.282</td>
<td>1.645</td>
<td>1.960</td>
<td>2.326</td>
<td>2.576</td>
<td>2.807</td>
<td>3.090</td>
<td>3.291</td>
</tr>
</tbody>
</table>

http://en.wikipedia.org/wiki/T_distribution
One Sample t-Test: Example

- The average price of a car in city is $12k
- Five cars park in front of a house with an average price of $20,270 and standard deviation of $5,811
- Null hypothesis (H_0): the cars are not more expensive than in the rest of the city

\[
t = \frac{\bar{x} - \mu}{s/\sqrt{N}} = \frac{20270 - 12000}{5811/\sqrt{5}} = 3.18
\]

- DoF=4 (for the one sample t-Test: sample size -1)
- Set confidence level to 95% (5% error probability)
- Since \( t = 3.18 > 2.132 \) (see t-table) reject H_0
- The cars are significantly more expensive (with 5% error probability)
One Sample t-Test: Assumptions

- Independently generated samples
- The population distribution is Gaussian (otherwise the t-distribution is not the correct choice)
- Mean is known

Comments

- The t-Test is quite robust under non-Gaussian distributions
- Often a 95% or 99% confidence (=5% or 1% significance) level is used
- t-Test is one of the most frequently used tests in science
Two Sample t-Test

- Often, one wants to compare the means of two samples to see if both are drawn from populations with equal means
- Example: Compare two estimation procedures (operating on potentially different data sets)
Typical Hypotheses

- Typical null and alternative hypotheses
  \[ H_0 : \mu_1 = \mu_2 \]
  \[ H_1 : \mu_1 \neq \mu_2 \] (two-tailed test)
  \[ H_1 : \mu_1 < \mu_2 \] (one-tailed test)
  \[ H_1 : \mu_1 > \mu_2 \] (one-tailed test)

- Logic of the test is similar as before
- Slightly different statistics
Pooled Variance (1)

- One sample t-Test

\[ \hat{\sigma}_x = \sqrt{\frac{s^2}{N}} = \sqrt{\frac{\sum(x_i - \bar{x})^2}{(N-1)N}} = \sqrt{\frac{SS}{N \times DoF}} \]

- For the two sample t-Test, we have two variances.

- The pooled, estimated variance of the sampling distribution of the difference of means is:

\[ \hat{\sigma}_{pooled}^2 = \frac{SS_1 + SS_2}{df_1 + df_2} = \frac{(N_1-1)s_1^2 + (N_2-1)s_2^2}{N_1 + N_2 - 2} \]
Pooled Variance (2)

- Which leads to the pooled, estimated SE of the sampling distribution of the difference of means

\[ \hat{\sigma}_{\bar{x}_1 - \bar{x}_2} = \sqrt{\hat{\sigma}_{\text{pooled}}^2 \left( \frac{1}{N_1} + \frac{1}{N_2} \right)} \]

- We are interested in the differences, thus the t-statistics turns into

\[ t_{\bar{x}_1 - \bar{x}_2} = \frac{\bar{x}_1 - \bar{x}_2}{\hat{\sigma}_{\bar{x}_1 - \bar{x}_2}} \]
Two Sample t-Test Example

- Two planning algorithms A and B
- Evaluate A and B, each in 25 randomly generated scenarios \((N_A = N_B = 25)\)
- \(H_0 : \mu_A = \mu_B \leftrightarrow \mu_A - \mu_B = 0\)
- \(H_1 : \mu_A \neq \mu_B \leftrightarrow \mu_A - \mu_B \neq 0\)
- \(\bar{x}_A = 127, s_A = 33; \bar{x}_B = 131, s_B = 28\)
- \(\sigma^2_{pooled} = 936.5; \hat{\sigma}_{\bar{x}_A-\bar{x}_B} = 8.65\)
- \(t_{\bar{x}_1-\bar{x}_2} = (\bar{x}_A - \bar{x}_B)/(\hat{\sigma}_{\bar{x}_A-\bar{x}_B}) = -0.46\)
- DoF is \(N_A + N_B - 2 = 48\)
- We cannot reject \(H_0\) since \(|t| < 2.01\)
Paired Sample t-Test

- Observation: The smaller the variance, the easier it is to show a significant difference.
- Design the experiments to directly measure the performance boost of a technique by considering differences.
- Test if the mean of \((A(d) - B(d))\) is significantly different from zero.

Examples

- Two estimation procedures operating on the same data set.
- Blood values of patients before and after a treatment.
Two Sample t-Test vs. Paired Sample t-Test

- **Two sample test:** Test if the differences of the means differs from zero
- **Paired sample test:** Test if the means computed over the individual differences is differ from zero

\[ H_0 : \mu_\delta = 0 ; \quad H_1 : \mu_\delta \neq 0 \]

\[ t_\delta = \frac{\bar{x}_\delta - \mu_\delta}{\hat{\sigma}_\delta} = \frac{\bar{x}_\delta}{\hat{\sigma}_\delta} \]

\[ \hat{\sigma}_\delta = \frac{s_\delta}{\sqrt{N_\delta}} \]
Paired Sample t-Test

- **Paired sample test:** Test if the means computed over the individual differences is differ from zero (or a constant $\mu_\delta$)

- Hypotheses $H_0 : \mu_\delta = 0$ ; $H_1 : \mu_\delta \neq 0$

- Test statistic

$$t_\delta = \frac{\bar{x}_\delta - \mu_\delta}{\hat{\sigma}_\delta} = \frac{\bar{x}_\delta}{\hat{\sigma}_\delta}$$

$$\hat{\sigma}_\delta = \frac{s_\delta}{\sqrt{N_\delta}}$$

- $DoF = N_\delta - 1$ (number of pairs -1)

- Use t-values as in the One sample test

- Whenever possible, use the paired sample t-Test since is minimized the variance
Confidence Intervals

- For a normal with known $\mu$ and $\sigma$, 95% of the samples fall within $\mu \pm 1.96\sigma$
- Thus, we can state that $\bar{x} \pm 1.96\sigma_{\bar{x}}$ contains the mean (for large N) with 95% probability
- Correct statement: “I am 95% sure that the $1.96\sigma_{\bar{x}}$ interval around $\bar{x}$ contains the mean.”
Confidence Intervals for Small N

- In case N is small, we need to use the t distribution to come up with the correct intervals

\[ \bar{x} \pm 1.96\sigma_{\bar{x}} \quad \rightarrow \quad \bar{x} \pm t'\hat{\sigma}_{\bar{x}} \]

- value from the t table for 95% confidence and corresponding DoF

- \( t' \) is bigger than 1.96, depending on the DoF and thus the sample size N
Visualizing Confidence Intervals

- Non-overlapping confidence intervals indicate a significant difference
- Overlapping intervals indicate nothing
Overlapping Confidence Intervals and Significance

- Consider two samples (with large N)
- The means are significantly different when:
  \[ ||x_1 - x_2|| > 1.96 \sqrt{SE_1^2 + SE_2^2} \]
- There is no overlap between CI when:
  \[ ||x_1 - x_2|| > 1.96(SE_1 + SE_2) \]
- Note that \( \sqrt{SE_1^2 + SE_2^2} < SE_1 + SE_2 \), so we have

\[
\begin{array}{c}
1.96\sqrt{SE_1^2 + SE_2^2} \\
\text{No significance}
\end{array}
\]

\[
\begin{array}{c}
1.96(SE_1 + SE_2) \\
\text{Significance}
\end{array}
\]

\[
\begin{array}{c}
\text{CIs overlap} \\
\text{CIs do not overlap}
\end{array}
\]
What Happens for Large N?

- The larger the sample size, the easier it is to show differences...
- ... but for large sample sizes, we can show any statistical significant difference no matter how small it is
- A statistically significant difference does not tell anything about if the difference is meaningful!
- See concept of “informativeness”
Conclusion

- To support the claim that A is better than B, use statistical tests
- t-Test is the most frequently used test
- Prefer the paired t-Test over the two sample t-Test (if applicable)
- Sometimes it is nice to visualize results with confidence intervals.
  - Non-overlapping CI imply significance
  - Overlapping CI imply nothing
- For large N, differences may by statistically significant but practically meaningless!
Further Reading

- Cohen’95: Empirical Methods for AI (highly recommended)

- Wikipedia offers rather articles as well on this topic