Advanced Techniques for Mobile Robotics

Statistical Testing

Wolfram Burgard, Cyrill Stachniss,

Kai Arras, Maren Bennewitz

UNI FREIBURG

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)?

1^{st} 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?

2^{nd} 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^{nd} 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

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?

- $\mu_A = 0.475$ s, $\mu_B = 0.45$ s
- $\mu_A > \mu_B$, so B is better than A?!
- We just evaluated four tests, thus μ_A and μ_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

$$H_0 \;:\; \mu = 0$$

$$H_1 \;:\; \mu \neq 0 \qquad \text{(two-tailored test)}$$

$$H_1 \;:\; \mu < 0 \qquad \text{(one-tailored test)}$$

$$H_1 \;:\; \mu > 0 \qquad \text{(one-tailored test)}$$

$$H_0 \;:\; \mu_1 = \mu_2$$

$$H_1 \;:\; \mu_1 \neq \mu_2 \qquad \text{(two-tailored test)}$$

$$H_1 \;:\; \mu_1 < \mu_2 \qquad \text{(one-tailored test)}$$

$$H_1 \;:\; \mu_1 > \mu_2 \qquad \text{(one-tailored test)}$$

Population and Sample

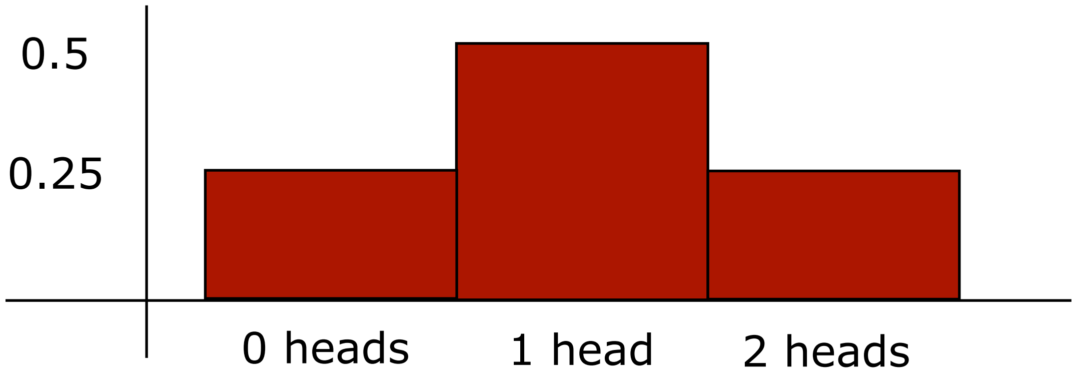
- The data we observe is often only a small fraction of the possible outcomes

- **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



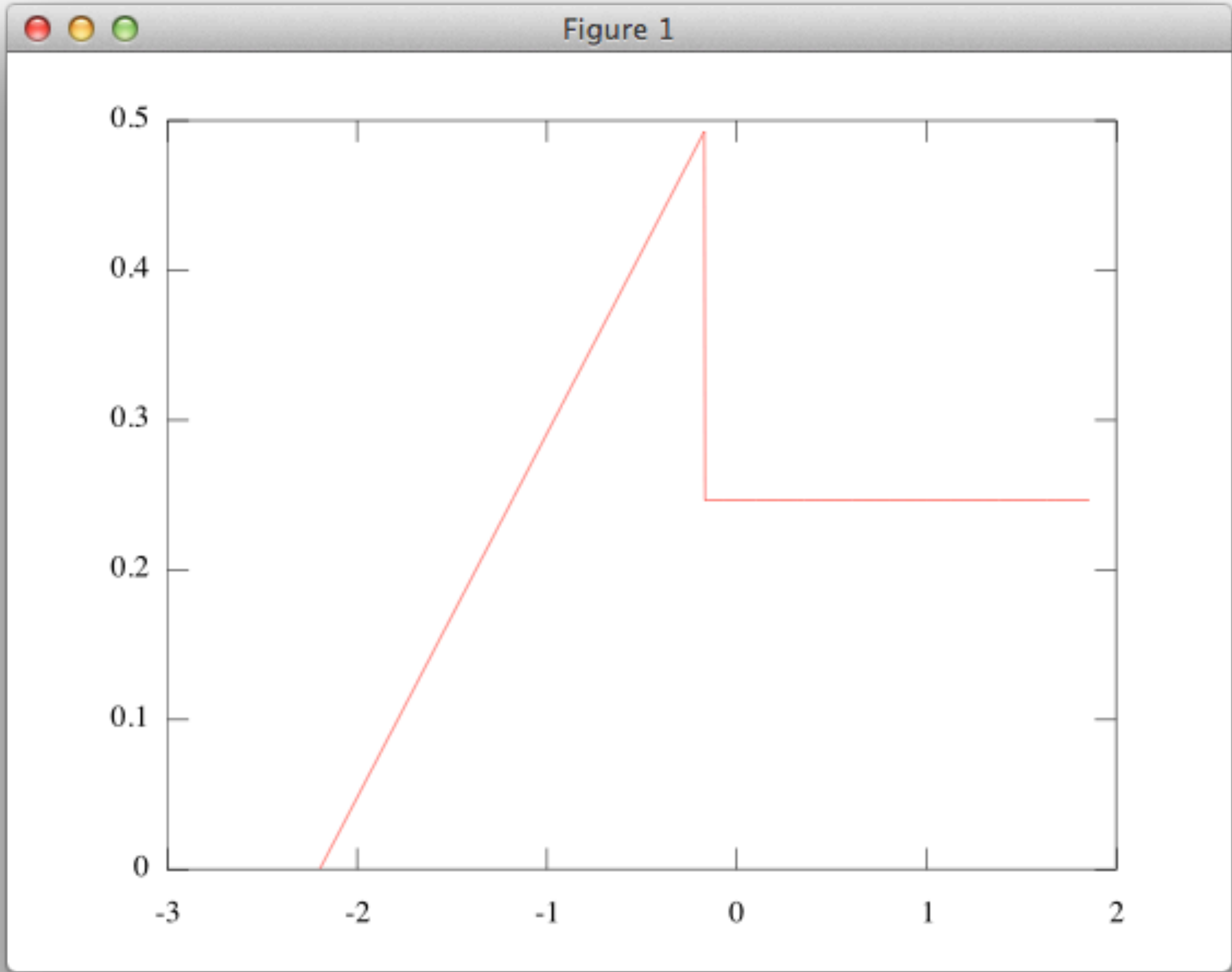
Sampling Distribution

- 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 μ and standard deviation σ, then the mean of the sampling distribution is μ 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)

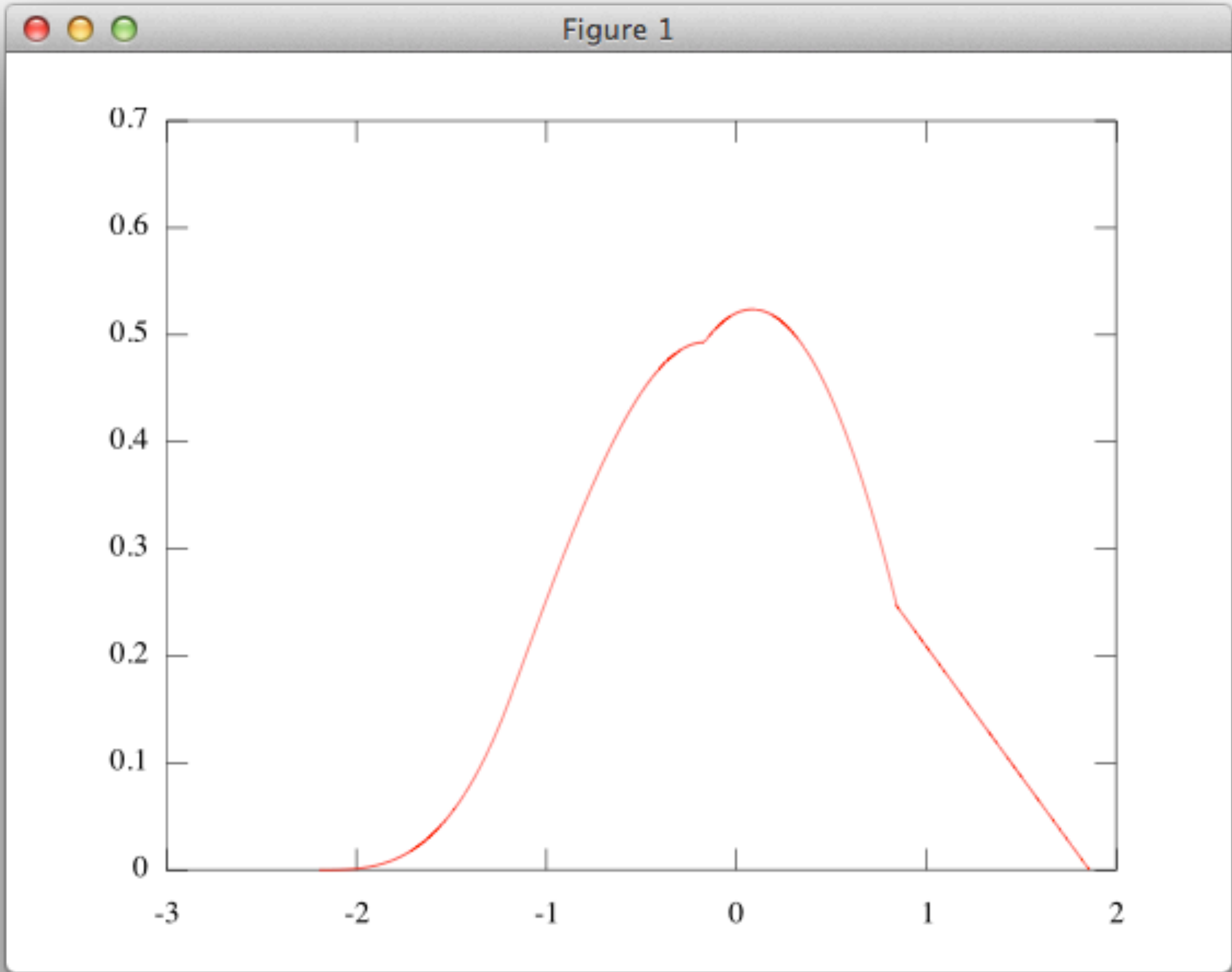


$\mu = 0$

$\sigma = 1.45$

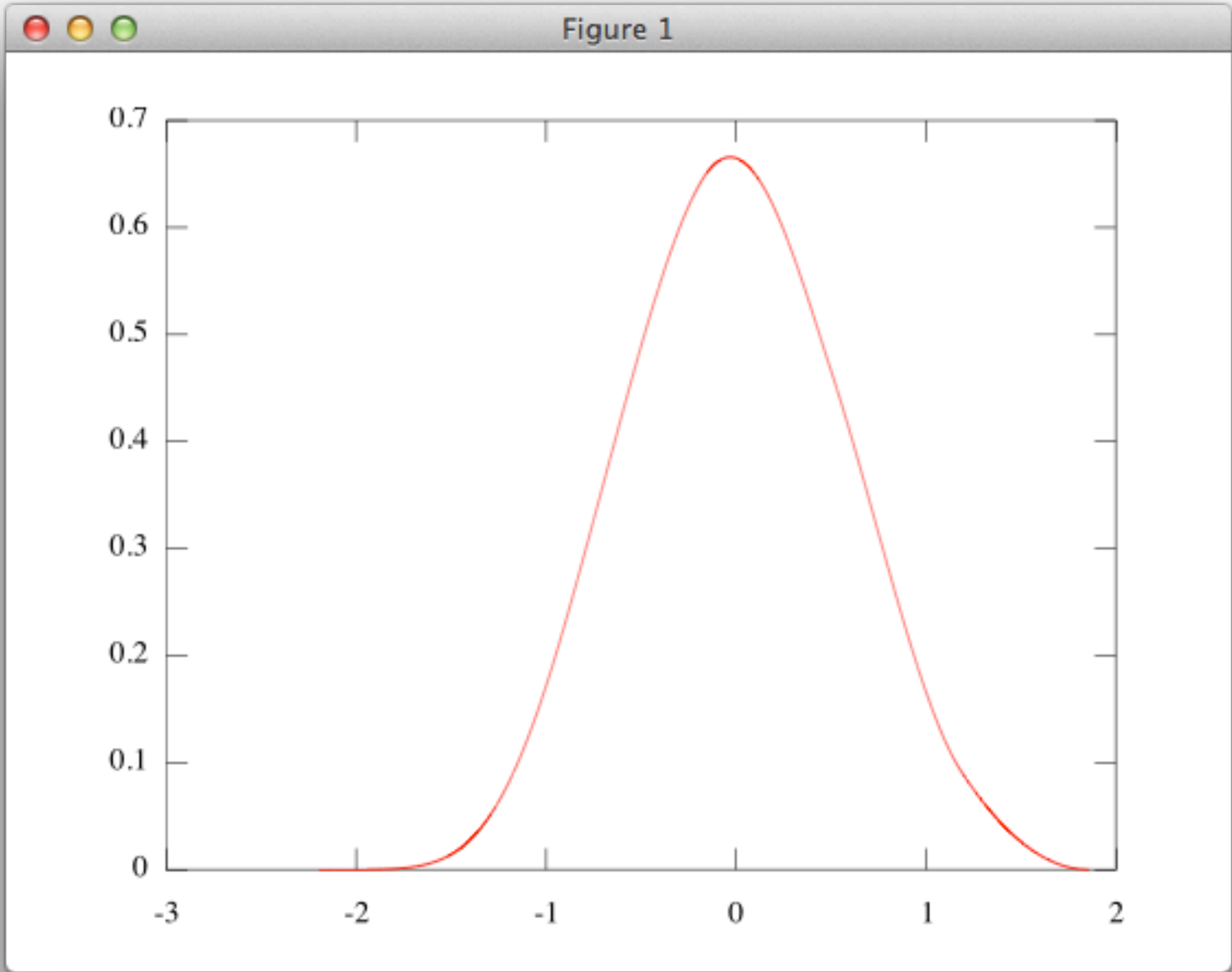
[Illustration of the central limit theorem, Wikipedia]

p(average of two samples)



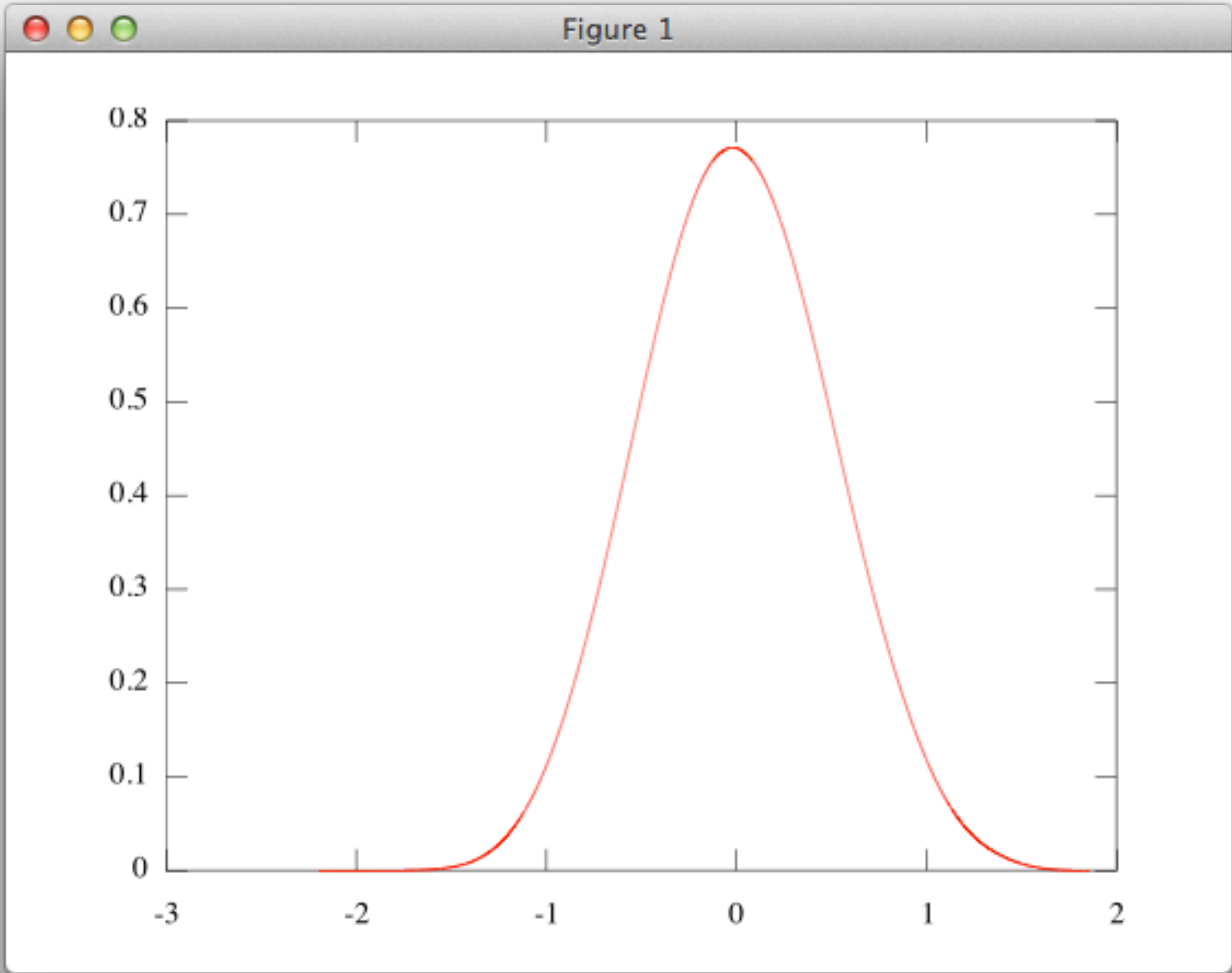
[Illustration of the central limit theorem, Wikipedia]

p(average of three samples)



[Illustration of the central limit theorem, Wikipedia]

p(average of four samples)

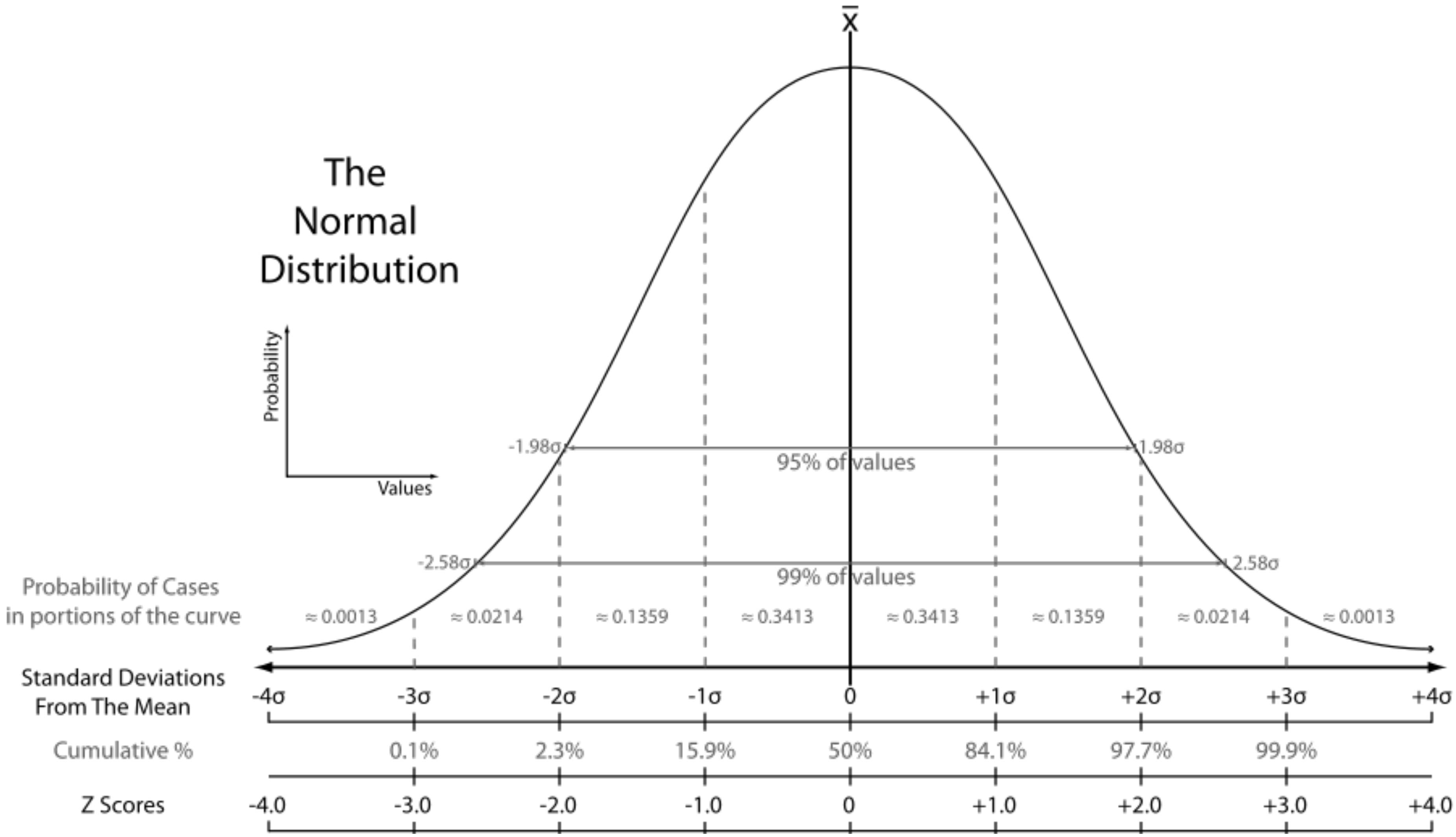


[Illustration of the central limit theorem, Wikipedia]

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: $\lim_{N \to \infty} \bar{x} = \mu$
- The standard error represents the uncertainty about the mean and is given by $\sigma_{\bar{x}} = \sigma / \sqrt{N} \ (= SE)$

The Normal Distribution



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=~95%

One Sample Z-Test

- One sample location test
- Given a μ and σ 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: μ=100, σ=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?
- $SE = \sigma/\sqrt{N} = 12/\sqrt{55} \simeq 1.62$
- $Z = \frac{\bar{x}-\mu}{SE} = \frac{96-100}{1.62} = -2.47$
- 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, σ 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! ☺

William Sealy Gosset

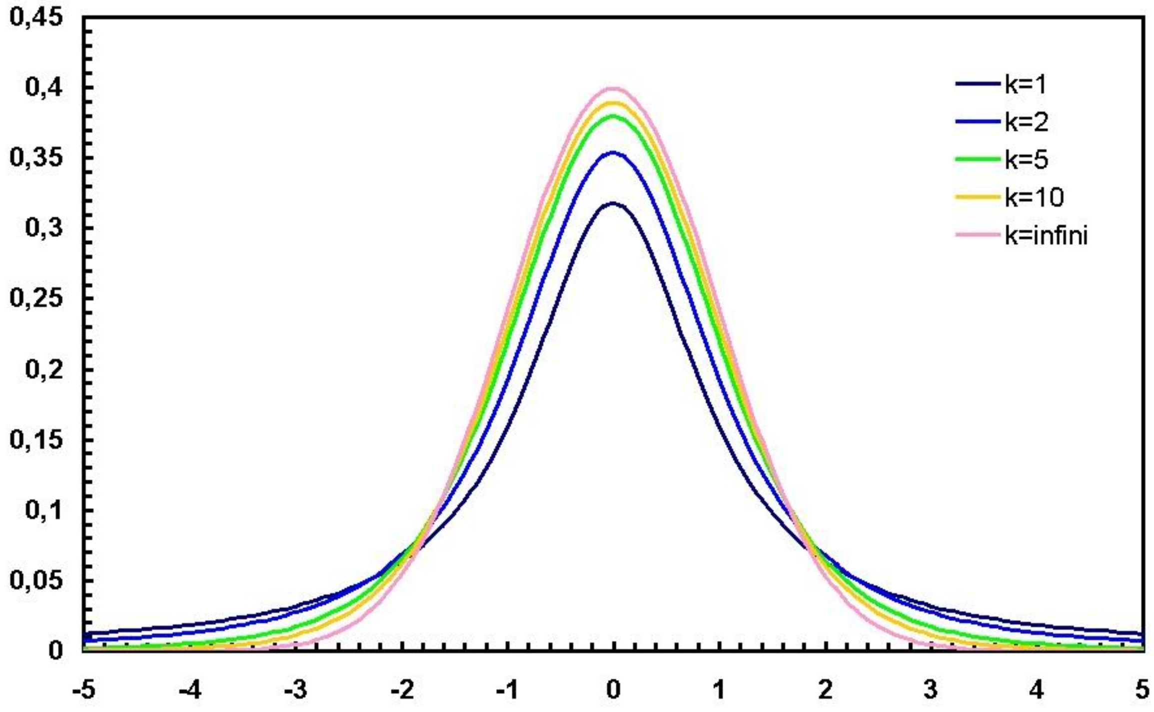
- Test to cheaply monitor the quality of stout at Guinness brewery (~1908)

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

$$t = \frac{\bar{x}-\mu}{\hat{\sigma}_{\bar{x}}} = \frac{\bar{x}-\mu}{s/\sqrt{N}}$$

std. dev estimated form the sample

sample size

- 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)

t-Table 1/2

confidence level

degree of freedom

One Sided	75%	80%	85%	90%	95%	97.5%	99%	99.5%	99.75%	99.9%	99.95%
Two Sided	**50%**	**60%**	**70%**	**80%**	**90%**	**95%**	**98%**	**99%**	**99.5%**	**99.8%**	**99.9%**
1	1.000	1.376	1.963	3.078	6.314	12.71	31.82	63.66	127.3	318.3	636.6
2	0.816	1.061	1.386	1.886	2.920	4.303	6.965	9.925	14.09	22.33	31.60
3	0.765	0.978	1.250	1.638	2.353	3.182	4.541	5.841	7.453	10.21	12.92
4	0.741	0.941	1.190	1.533	2.132	2.776	3.747	4.604	5.598	7.173	8.610
5	0.727	0.920	1.156	1.476	2.015	2.571	3.365	4.032	4.773	5.893	6.869
6	0.718	0.906	1.134	1.440	1.943	2.447	3.143	3.707	4.317	5.208	5.959
7	0.711	0.896	1.119	1.415	1.895	2.365	2.998	3.499	4.029	4.785	5.408
8	0.706	0.889	1.108	1.397	1.860	2.306	2.896	3.355	3.833	4.501	5.041
9	0.703	0.883	1.100	1.383	1.833	2.262	2.821	3.250	3.690	4.297	4.781
10	0.700	0.879	1.093	1.372	1.812	2.228	2.764	3.169	3.581	4.144	4.587
11	0.697	0.876	1.088	1.363	1.796	2.201	2.718	3.106	3.497	4.025	4.437
12	0.695	0.873	1.083	1.356	1.782	2.179	2.681	3.055	3.428	3.930	4.318
13	0.694	0.870	1.079	1.350	1.771	2.160	2.650	3.012	3.372	3.852	4.221
14	0.692	0.868	1.076	1.345	1.761	2.145	2.624	2.977	3.326	3.787	4.140
15	0.691	0.866	1.074	1.341	1.753	2.131	2.602	2.947	3.286	3.733	4.073
16	0.690	0.865	1.071	1.337	1.746	2.120	2.583	2.921	3.252	3.686	4.015
17	0.689	0.863	1.069	1.333	1.740	2.110	2.567	2.898	3.222	3.646	3.965
18	0.688	0.862	1.067	1.330	1.734	2.101	2.552	2.878	3.197	3.610	3.922
19	0.688	0.861	1.066	1.328	1.729	2.093	2.539	2.861	3.174	3.579	3.883
20	0.687	0.860	1.064	1.325	1.725	2.086	2.528	2.845	3.153	3.552	3.850

t-Table 2/2

20	0.687	0.860	1.064	1.325	1.725	2.086	2.528	2.845	3.153	3.552	3.850
21	0.686	0.859	1.063	1.323	1.721	2.080	2.518	2.831	3.135	3.527	3.819
22	0.686	0.858	1.061	1.321	1.717	2.074	2.508	2.819	3.119	3.505	3.792
23	0.685	0.858	1.060	1.319	1.714	2.069	2.500	2.807	3.104	3.485	3.767
24	0.685	0.857	1.059	1.318	1.711	2.064	2.492	2.797	3.091	3.467	3.745
25	0.684	0.856	1.058	1.316	1.708	2.060	2.485	2.787	3.078	3.450	3.725
26	0.684	0.856	1.058	1.315	1.706	2.056	2.479	2.779	3.067	3.435	3.707
27	0.684	0.855	1.057	1.314	1.703	2.052	2.473	2.771	3.057	3.421	3.690
28	0.683	0.855	1.056	1.313	1.701	2.048	2.467	2.763	3.047	3.408	3.674
29	0.683	0.854	1.055	1.311	1.699	2.045	2.462	2.756	3.038	3.396	3.659
30	0.683	0.854	1.055	1.310	1.697	2.042	2.457	2.750	3.030	3.385	3.646
40	0.681	0.851	1.050	1.303	1.684	2.021	2.423	2.704	2.971	3.307	3.551
50	0.679	0.849	1.047	1.299	1.676	2.009	2.403	2.678	2.937	3.261	3.496
60	0.679	0.848	1.045	1.296	1.671	2.000	2.390	2.660	2.915	3.232	3.460
80	0.678	0.846	1.043	1.292	1.664	1.990	2.374	2.639	2.887	3.195	3.416
100	0.677	0.845	1.042	1.290	1.660	1.984	2.364	2.626	2.871	3.174	3.390
120	0.677	0.845	1.041	1.289	1.658	1.980	2.358	2.617	2.860	3.160	3.373
∞	0.674	0.842	1.036	1.282	1.645	1.960	2.326	2.576	2.807	3.090	3.291

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 expansive (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

$$
\begin{aligned}
H_0 &: \quad \mu_1 = \mu_2 \\
H_1 &: \quad \mu_1 \neq \mu_2 \quad \text{(two-tailored test)} \\
H_1 &: \quad \mu_1 < \mu_2 \quad \text{(one-tailored test)} \\
H_1 &: \quad \mu_1 > \mu_2 \quad \text{(one-tailored test)}
\end{aligned}
$$

- Logic of the test is similar as before
- Slightly different statistics

Pooled Variance (1)

- One sample t-Test

"sum of squares"

$$\hat{\sigma}_{\bar{x}} = \sqrt{s^2/N} = \sqrt{\frac{\sum(x_i - \bar{x})^2}{(N-1)N}} = \sqrt{\frac{SS}{N \times DoF}}$$

degree of freedom

- For the two sample t-Test, we have two variances.
- The pooled, estimated variance of the sampling distribution of the difference of means is:

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

Pooled Variance (2)

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

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

- 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 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

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

$$t_\delta = \frac{\bar{x}_\delta - \mu_\delta}{\hat{\sigma}_\delta} = \frac{\bar{x}_\delta}{\hat{\sigma}_\delta} \qquad \hat{\sigma}_\delta = \frac{s_\delta}{\sqrt{N_\delta}}$$

Paired Sample t-Test

- **Paired sample test:** Test if the means computed over the individual differences is differ from zero (or a constant μ_δ)
- Hypotheses $H_0 : \mu_\delta = 0$; $H_1 : \mu_\delta \neq 0$
- Test statistic

$$t_\delta = \frac{\bar{x}_\delta - \mu_\delta}{\hat{\sigma}_\delta} = \frac{\bar{x}_\delta}{\hat{\sigma}_\delta} \qquad \hat{\sigma}_\delta = \frac{s_\delta}{\sqrt{N_\delta}}$$

- $DoF = N_\delta - 1$ (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 μ and σ, 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

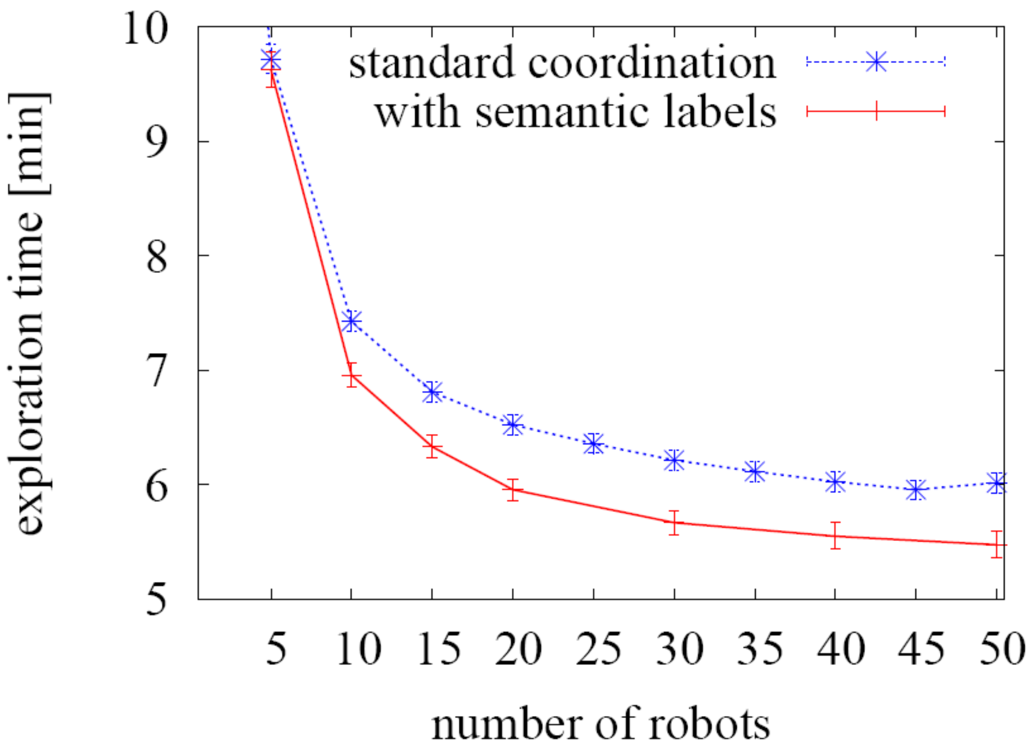
$$\bar{x} \pm 1.96\sigma_{\bar{x}} \quad \Rightarrow \quad \bar{x} \pm t'\widehat{\sigma}_{\bar{x}}$$

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

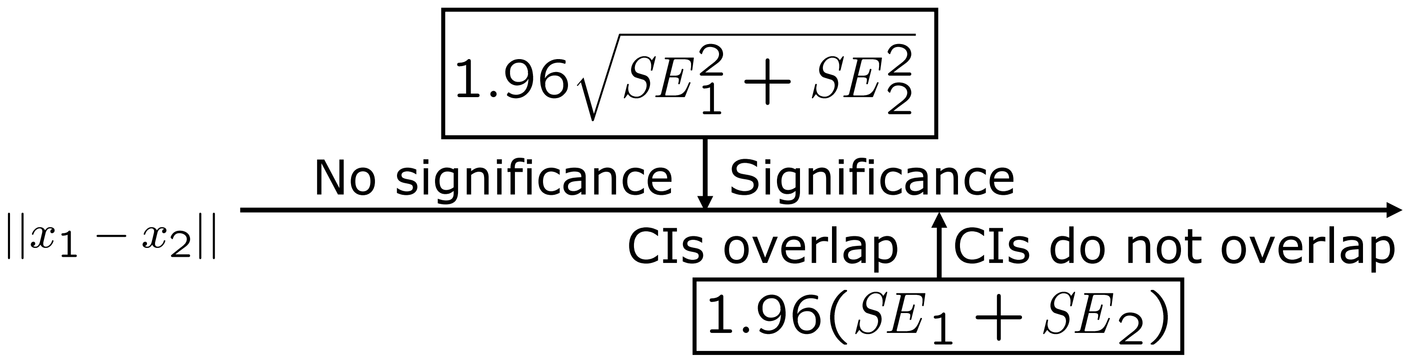
- Consider two samples (with large N)
- The means are significantly different when:

$$||x_1 - x_2|| > 1.96\sqrt{SE_1^2 + SE_2^2}$$

- 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



What Happens for Large N?

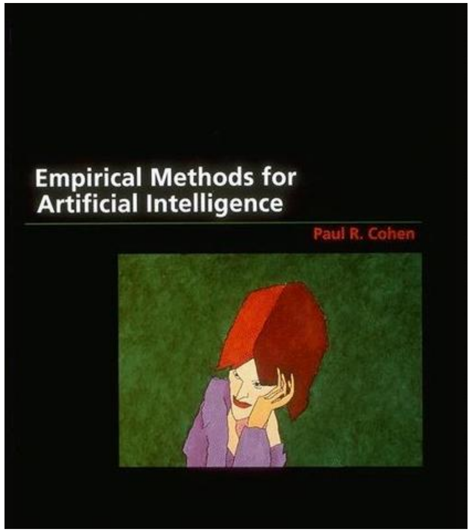
- The larger the sample size, the easier it is to show differences…
- … but for large sample sizes, we can show any statistical significant difference **no matter how small it is**
- A statistically significant difference does **not tell anything about if the difference is meaningful!**
- See concept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