Research Methods

Confidence Intervals and Sampling Methods

Dr. Sven Magg, Prof. Dr. Stefan Wermter

http://www.informatik.uni-hamburg.de/WTM/

Plan for today!

1. What are confidence intervals?
2. Empirical Sampling
   a) Monte-Carlo Tests
   b) Bootstrapping

Parameter Estimation

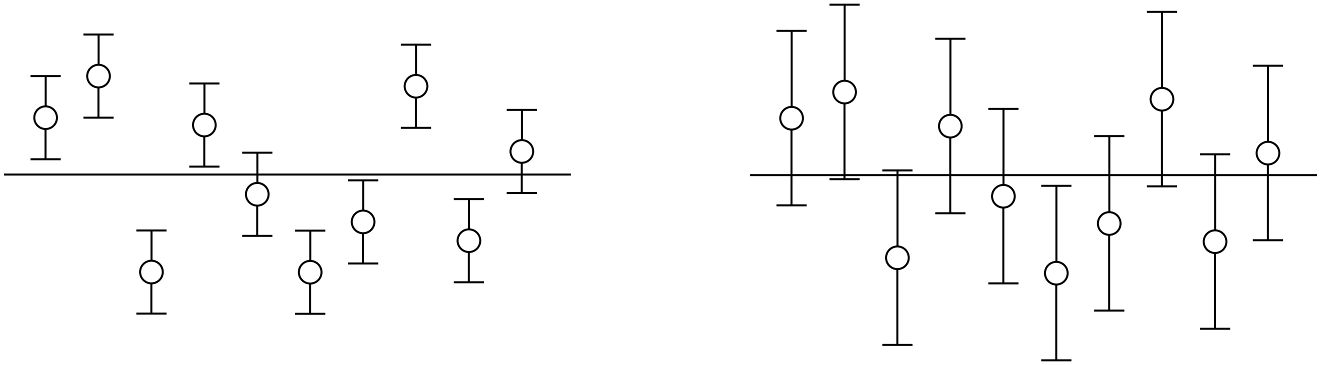
- Hypothesis Testing
  - helps to answer a yes/no question about our data
  - used to accept/reject a hypothesis in favour of another
  - helps us to estimate the error we might have made

- Parameter Estimation
  - We try to estimate a population parameter from a sample we have drawn
  - We have already learned about some estimators:

$$\hat{\mu} = \bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i \qquad \hat{\sigma}_{\bar{x}} = \frac{s}{\sqrt{N}}$$

  - How accurate are those estimates?

Confidence Intervals

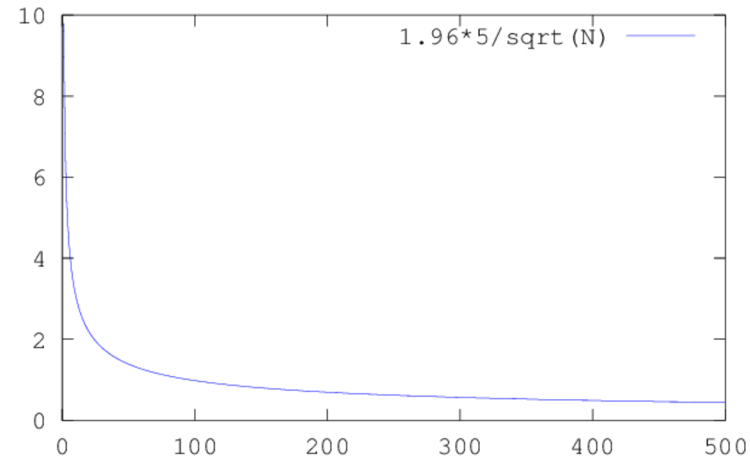


- How wide should a CI be?
- $\mu = \overline{x} \pm \varepsilon$
- How often do you expect that μ falls within ε of the sample mean?
- A 95% CI means you are 95% sure that this interval includes μ
- Remember: μ is a constant, a CI says more about $\overline{x}$

Confidence Intervals

- Example:
  - We collect samples of size N and know σ
  - $\Rightarrow$ sampling distribution of the means is normal
  - $\Rightarrow$ 95% of all sample means are in the interval given by $\mu \pm 1.96\sigma_{\bar{x}}$
  - This also means that if $\varepsilon = 1.96\sigma_{\bar{x}}$, then the interval $\bar{x} \pm 1.96\sigma_{\bar{x}}$ will contain μ in 95% of all samples we draw

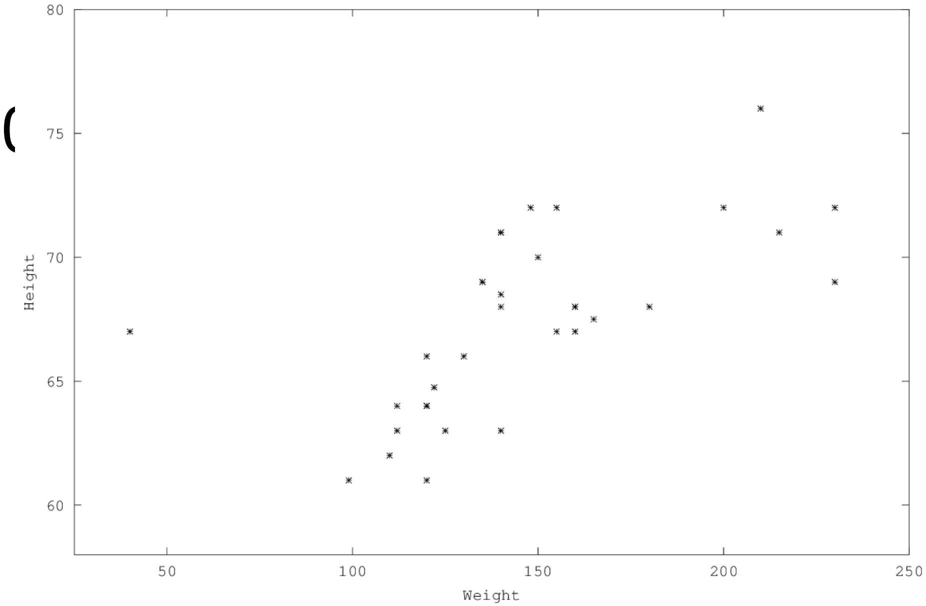
- CI is dependent on standard error and therefore N



CI Example

- Height and weight of 33 students
- Correlation Coefficient:

$r_{XY} = ($



$$z(r_{XY}) = 0.5 ln \frac{1 + 0.64}{1 - 0.64} = 0.759$$

$$\hat{\sigma}_{z(r_{XY})} = \frac{1}{\sqrt{n-3}} = \frac{1}{\sqrt{30}} = 0.183$$

$H_0(\rho_{XY} = 0)$: $Z = 4.16$

$$z(\rho) = z(r_{XY}) \pm 1.96\hat{\sigma}_{z(r_{XY})} = 0.759 \pm 0.358 = (0.40, 1.12)$$

$\Rightarrow$ 95% Confidence interval: $0.38 \leq \rho \leq 0.81$

What if we don't know σ?

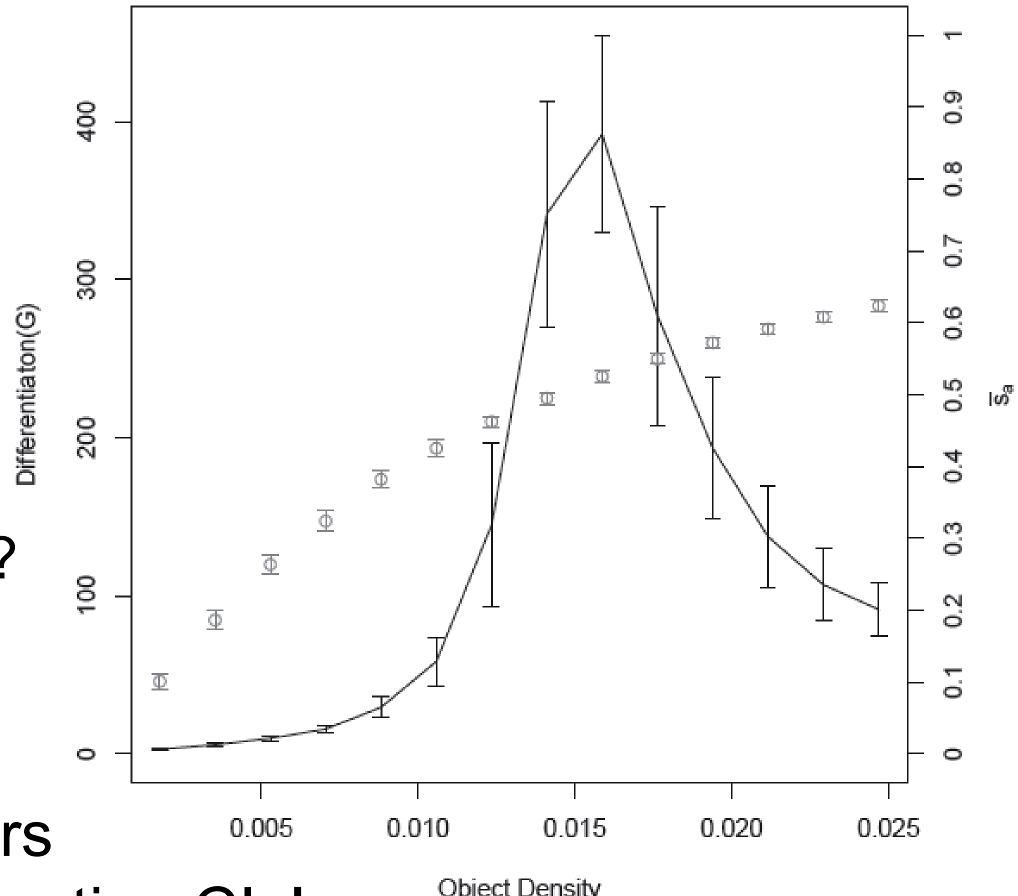
- If σ is not known, we estimate the sample error!
  $\hat{\sigma}_{\bar{x}} = \frac{s}{\sqrt{N}}$
- **The problem now**: The sampling distribution of the mean is NOT normally distributed but rather t-distributed
- What do we do?
- We look up the critical values in a t-distribution table!
- Since a 95% CI is symmetric, we have to look up $t_{97.5}$ (or $t_{.025}$)
- Then calculate $\bar{x} \pm t_{97.5}\sigma_{\bar{x}}$
- $t_{97.5}$ will be different for different degrees of freedom

df	**1**	**2**	**5**	**20**	**60**
$\boldsymbol{t_{97.5}}$	12.71	4.303	2.571	2.086	2.000

How to use CIs?

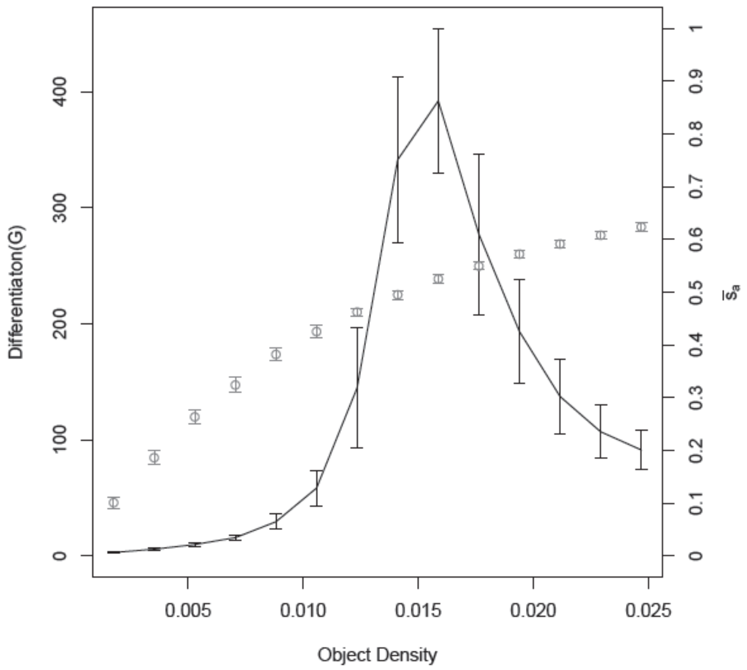
- Show them in graphs as error bars!

- What can we see?
- Different sizes of CI
  - Due to different s?
  - Due to different number of samples?
  - Due to boundary effects?



- CIs are useful as error bars but be careful when interpreting CIs!

CI and Hypothesis Testing



- It is tempting to test hypotheses with confidence intervals
- **Confidence Interval**: Estimated from standard error of individual means
- **Hypothesis Tests**, e.g. t-Test Uses standard error of the difference between the means
- **Poor mans t-Test:** If CIs of two means do not overlap $\Rightarrow$ Two-sample t-Test will show they are different
- If they overlap: Don't use them for hypothesis tests!

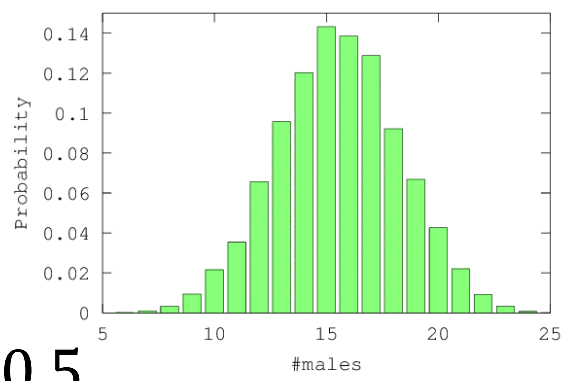
Sampling distributions

- Different ways to get sampling distributions
  - Exact distributions
    - Derived analytically/mathematically
  - Estimated distributions
    - Central Limit Theorem (CLT)
    - Z-distribution (standard normal distribution)
    - t-Distributions
    - Fisher’s z-distribution
- How to get sampling distributions in case there is no analytical distribution or good estimate

$\Rightarrow$ Computer-intensive methods through simulation of the sampling process

Monte-Carlo Tests

- We often know the population distribution from which we draw samples, but **not** the sampling distribution
  - e.g. a hypothetic population as defined by H_0
- Since we know the parameters, we can simulate the process
- **Simple example:**
  H_0: Same number of men and women at the Informatikum



- Samples of 30 students with $P(male) = 0.5$
- Select 10000 pseudo-samples of 30 students and record number of male students
- **Empirical sampling**

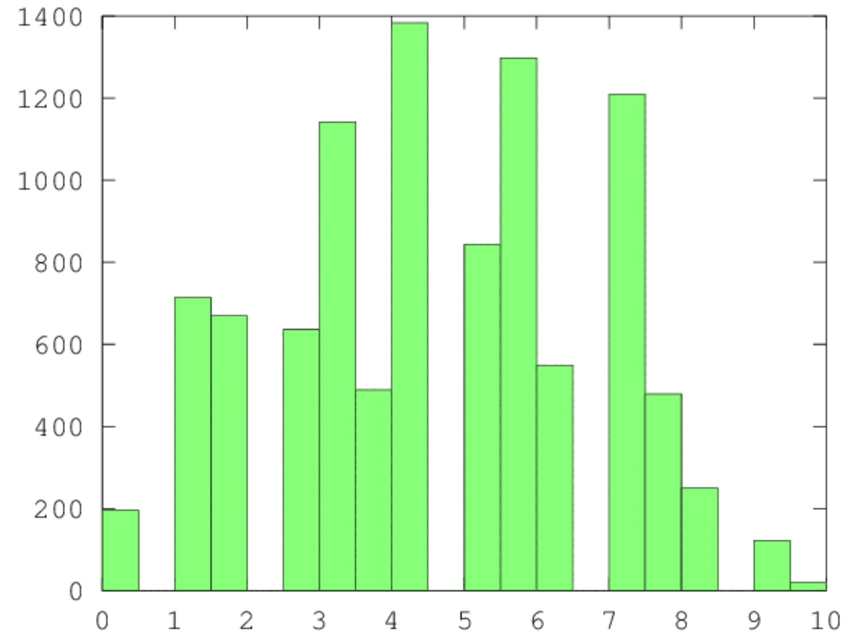
Monte-Carlo Tests

- General approach

1. Determine population parameters and test statistic T to use
2. For i = 1 to K
   a. Draw pseudo-sample of size N from the population
   b. Calculate test statistic ${T_i}^*$ for pseudo-sample
   c. Record the frequencies
3. Use the distribution of T^* to determine probability of original sample under H_0

Realistic Example

- Agent in a grid-world with 4 directions of movement
- We measure Euclidian distance travelled after 10 steps
- We hypothesise it has a 25% chance to turn left and 25% to turn right in each step
- What is the probability of a measure of 8 or more?
- How does the sampling distribution look?
- Frequency histogram for 10000 sample runs
- $P(x \geq 8) = 0.0392$



Monte-Carlo Sampling

- Advantages
  - Straightforward and usually simple to calculate
  - Cheap for most computer science problems
  - Can be used for any statistic

- Disadvantages
  - We have to know the population parameters to know where to draw samples from
  - Often the population distribution is not known

Bootstrapping

- Let's assume
  - We have a sample S of a reasonable size N
  - We don't have the population parameters
- We can perform Monte-Carlo Sampling on the sample
  - Treat the sample as the population
  - Run Monte-Carlo Sampling with replacement

1. For i = 1 to K
   a. Select a sample ${S_i}^*$ of size N from S with replacement
   b. Calculate and record pseudo-statistic for ${S_i}^*$
2. Determine and use probability distribution of S^*

Boostrapping

- We now have an estimated sampling distribution S^* based on S itself
- We have to replace, because:
  - Obviously, without we only draw S k times
  - We assume a population that comprises all elements of S in the same proportions
- Beware:
  - We now have a sampling distribution for our sample S, NOT a null hypothesis
  - Example: H_0: $\mu = 50$, H_1: $\mu \neq 50$, $\bar{x} = 43$
  - The mean of the sampling distribution is 43, not 50!

Bootstrapping for H_0

- Shift Method
  - We assume that S^* and the distribution under H_0 have the same shape, but different means
  - We can “shift” all values in S^* by the difference of $\mu - \bar{x}$
  - In our example: Shift all values in S^* by $50 - 43 = 7$
- Normal Approximation Method
  - We assume that $\bar{x} - \mu$ is normally distributed
  - Then $Z = \frac{(\bar{x}-\mu)}{\sigma_{\bar{x}}}$, but we don't know $\sigma_{\bar{x}}$
  - We bootstrap the standard deviation from S^* and run a Z-Test

Bootstrap Example

- Sample S: (23, 42, 67, 53, 43, 60, 45, 32, 41, 24)
- $\bar{x} = 43,\ s = 4.54$
- Gather $S_i{}^*$ for $H_0\colon\ \mu = 50, H_1\colon\ \mu \neq 50$

	1	2	3	4	5	6	7	8	9	10	$\bar{x}^*$
$S_1{}^*$	32	32	45	45	23	67	53	67	41	53	45.8
$S_2{}^*$	43	24	42	23	23	43	23	60	32	41	35.4
$S_3{}^*$	60	41	41	60	67	67	24	32	45	42	47.9
$S_4{}^*$	45	23	32	67	43	67	43	41	42	43	44.6
$S_5{}^*$	23	53	53	45	32	53	60	42	45	45	45.1

- Run 5000 times: $\overline{\boldsymbol{x}}_{S^*} = 43.045,\ \sigma_{\overline{\boldsymbol{x}}_{S^*}} = 4.33$

Bootstrap Example

- Shift method:
  - Sort S^* to get critical values at positions $\mathrm{K} * \alpha/2$ and $K * (1 - \alpha/2)$.
  - For $\alpha = .05$: $c^- = 41.66, c^+ = 58.56$
  - Find p value by counting values below 43 and above 57: 432 $\Rightarrow p = 0.106$
  - We can't reject H_0 at the $\alpha = .05$ level
- Normal Approximation Method
  - $Z = \frac{(43.045 - 50)}{4.33} = -1.6051$
  - 2-Tailed boundary: ± 1.96, therefore not reject
  - $-1.96 = \frac{(c^- - 50)}{4.33} \Rightarrow c^- = 41.507$

Bootstrap Example

- t-Test:
  - $t = \frac{(43-50)}{s/\sqrt{N}} = \frac{7}{4.54} = 1.54$
  - $t_{.025} = 2.262$ for 9 degrees of freedom
  - $t_{.1}$ < 1.54 < $t_{.05}$

- All critical values close together
- Not surprising when we deal with means, i.e. a sampling distribution that can be considered normal
- In this case both shifting and normal approximation can be used

What have we learned?

1. A confidence interval gives us the interval in which we can expect μ with a given probability
2. CIs can be used as error bars around sample means
3. They become smaller with increasing N
4. If we know the population parameters, we can get a sampling distribution empirically by Monte-Carlo Sampling
5. We can derive a sampling distribution directly from a sample by bootstrapping, generating pseudo-samples by drawing from the original with replacement
6. Beware that a bootstrapped S^* is not S_{H_0}, we have to use the shift or normal approximation method