Research Methods

Hypothesis Testing

Dr. Sven Magg, Prof. Dr. Stefan Wermter

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

Plan for today!

1. What is a good hypothesis?
2. The steps of a statistical test
3. What is a one- or two-tailed test?
4. Where to find a sampling distribution
5. Z-Scores and hypotheses about means

Statistical Inference

- If we have a sample drawn from a population, we can ask two kind of questions:
  1. How “good” is an estimate for a parameter of the population, drawn from this sample?
     How confident are we that the estimate is close to the real parameter value?
     Example: Guessing the average number of blonde students from a snapshot count in the Mensa
  2. When answering a yes/no question about the population using the sample, how likely is it that we are wrong?
     Example: My software A is more accurate in guessing the weather than software B

Hypotheses

***1.** a suggested explanation for a group of facts or phenomena, either accepted as a basis for further verification (working hypothesis) or accepted as likely to be true […]*
***3.** (Philosophy / Logic) an unproved theory; a conjecture*
[Collins English Dictionary]

- has to be testable and falsifiable
- follows from observation, exploratory study, or just idea

- Big question: Is my hypothesis correct?
  - What does correct mean? How well can I prove it?
  - When do I consider it verified?

Hypotheses

- Falsifiability
  - “All students are female”
    - Falsifiable by a single male student
  - “When green aliens land in Hamburg, they always step of their spaceship with their middle foot first”
    - Falsifiable in principle, but not in practice
  - “A spiritual being exists”
    - Not scientifically falsifiable/testable
- It has to be possible to think of a hypothesis stating the opposite and you can both test them in practice

Let's gamble first…

- You watch a gambler throwing a die three times and always scoring a 6
- You want to accuse him of using a manipulated die!

- What are the chances of you being right?
- You assumption is that the die is fair and under this assumption you think it's unlikely to score three 6s

- Null Hypothesis: The die is fair (H_0)
- Alternative hypothesis: The die is manipulated (H_1)

Rejecting is better

- Why is my hypothesis the "alternative" hypothesis H_1?
- You can't prove a hypothesis with statistics on a sample
- But we can estimate the likelihood that a sample was drawn from a given population!

- H_0 : Sample from this population
- H_1 : Sample from a different population

- I can statistically evaluate the probability that my sample N_h came from a given population and, if low, reject H_0
- Rejecting H_0 → Evidence for H_1

Back to gambling

- Null Hypothesis: The die is fair (H_0)
- Alternative hypothesis: The die is manipulated (H_1)

- If the die is fair, we know what the probabilities are:
  - $1/6$ to get a specific number in one go
  - $1/(6 * 6 * 6) = 1/216 = 0.0046$ to get three 6s

- You are therefore **99.54% certain** that the die was not fair
- You reject H_0 with a chance of $p = 0.0046$ to be wrong

- **P-Value:** Measure of strength of evidence against H_0

What did we do?

- We have…

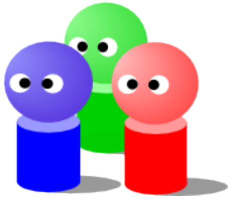
1. …stated a null hypothesis (Die is fair)
2. …observed a result (gathered a sample N_h=(6,6,6))
3. …thought about a formula to calculate chances, if H_0 is true (binomial distribution)
4. …calculated the probability p of N_h to happen, if H_0 is true
5. …decided to reject H_0 because $\boldsymbol{p}$ was too low

- Science is easy! ☺
- Is 0.46% low enough to accuse the 2m professional boxer of cheating?

Another example

- **Hypothesis**: There are more male than female students in computer science!
- **Step 1: State a Null-Hypothesis**
  - $H_0: P(male) = 0.5$

Group Task!

 2  5

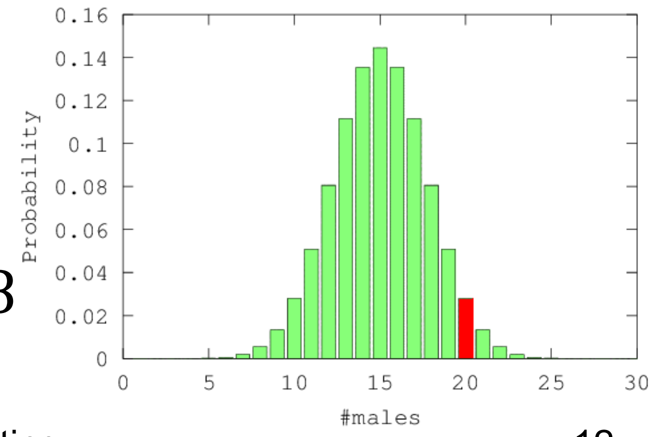
What is the difference between these two hypothesis:

1. **There are more male than female students in computer science**
2. **The ratio of male and female students is not equal in CS**

When do you reject the hypothesis?

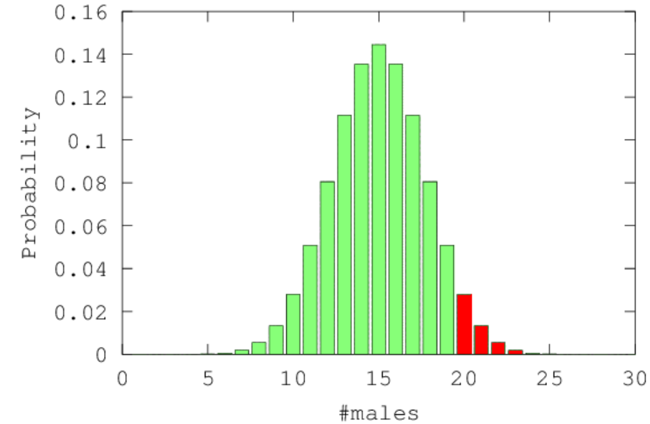
Another example

- **Hypothesis**: There are more male than female students in computer science!
- **Step 1: State a Null-Hypothesis**
  - $H_0: P(male) = 0.5$
- **Step 2: Gather a sample statistic** (run an experiment)
  - 30 students in the Mensa: $N_h = 20$ (male students)
- **Step 3: Find a sampling distribution $\boldsymbol{N_h}$, if $\boldsymbol{H_0}$ is true**
  - From $H_0: P(male) = P(female) = 0.5$
  - Binomial Distribution
- **Step 4: Calculate p value using $\boldsymbol{N_h}$**
- **Step 5: Decide**: Reject with $p = 0.028$



When to reject?

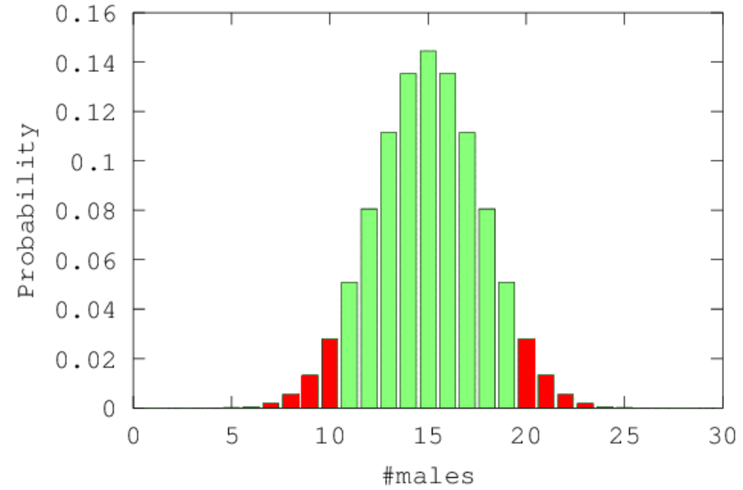
- We rejected H_0 because having 20 males in a sample of 30 is very unlikely
- How about 21? Or 10?
- We would reject if we see 20 or more!



- If H_0 would be rejected for several results, the p-Value of the combined result is the sum of individual p-Values:
- $P_{oneTailed} = P(20) + \dots + P(30) = 0.049!$
- **One-Tailed Test**: We reject for all values greater (smaller) than a given cut-off point (left/right tail)
- Used for directional hypotheses (e.g. "*greater than*")

When to reject

- 2-Tailed Test
  - Reject H_0 if observed value is greater or lower than one of two cut-off points
  - $P_{twoTailed} = 0.099$
  - Typical use:
    Reject H_0 when the observed value differs more than a given maximum from the mean



- Typical values to determine cut-off points:
  $p \leq 0.05$ or $p \leq 0.01$
  - Level of significance

Adjust procedure

- **Step 1-3 as before**
- **Step 4a: Set α**
  - Decide on a maximum acceptable probability α of incorrectly rejecting H_0
- **Step 4b: Find cut-off points**
  - Use sampling distribution N_h to find critical values c^+ and c^-
  - Set c^+ and c^- such that $\mathrm{P}(N_h \geq c^+) + P(N_h \leq c^-) \leq \alpha$
- **Step 5: Decide**
  - If $(N_h \geq c^+)\ or\ (N_h \leq c^-)$, reject H_0

- In the example: Set $\alpha = 0.05 \Rightarrow c^+ = 21$ and $c^- = 9$
  - $P(N_h \leq 9) + P(N_h \geq 21) = 0.043 \leq \alpha$

Sampling Distributions

- One problem still remains:
  Where do I find sampling distributions for H_0?

- Exact distributions
  - We already have used the binomial distribution
  - For sample size N and r the probability for each single event: $P(N_h = k) = \frac{N!}{k!(N-k)!} r^k (1-r)^{N-i}$
  - Many other probability distributions: e.g. Normal, Log, t (student), Geometric, Exponential, Poisson, $\chi^2, \ldots$
  - All statistics packages include functions to calculate probabilities given a distribution (probability density “pdf” or cumulative density “cdf”)

Estimating Distributions

- Use (large) sample collected for H_0 as estimate

- Computer intensive methods
- Simulate the sampling process to derive distribution
  - Monte-Carlo Sampling
  - Bootstrap Methods
  - Randomisation Tests

- Sampling distribution of the mean
  - Central limit theorem!
  - If my individuals are means of samples of size N
    $\Rightarrow$ Normal distribution with $\bar{x} = \mu$ and standard error $\sigma_{\bar{x}} = \frac{\sigma}{\sqrt{N}}$

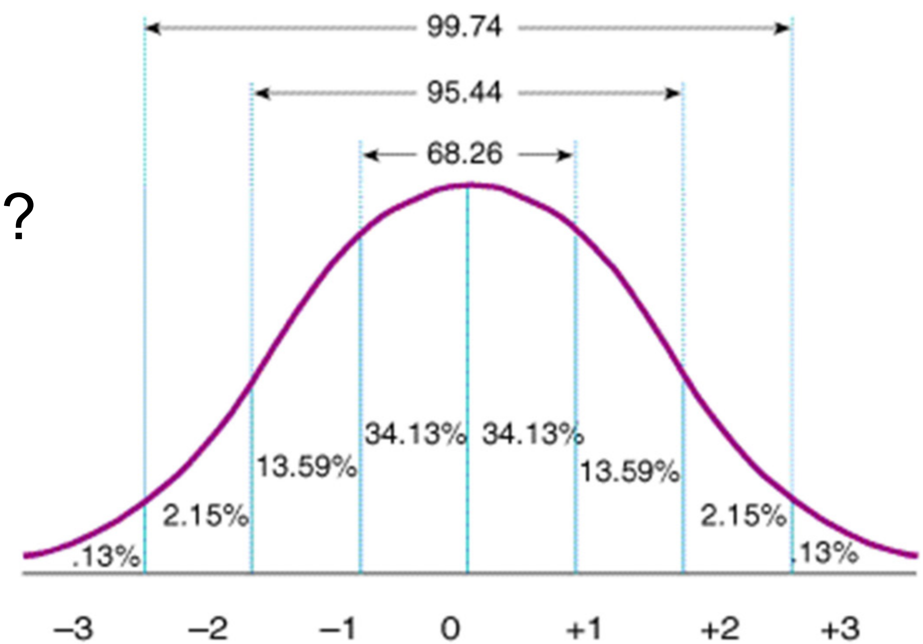
Hypotheses about means

- Example:
  - Individuals are means of 25 test scores
  - Old system: $\mu = 1.0$, $\sigma = 0.948$
  - New system test run: $\bar{x}$ = 2.8
  - Is this performance evidence for an improvement?
  - H_0: Old and new system are performing equally
  - Due to CLT: $\sigma_{\bar{x}} = \frac{\sigma}{\sqrt{N}} = \frac{0.948}{\sqrt{25}} = 0.19$
  - Mean of sampling distribution is 1.0, so $\bar{x}$ is 1.8 units above μ

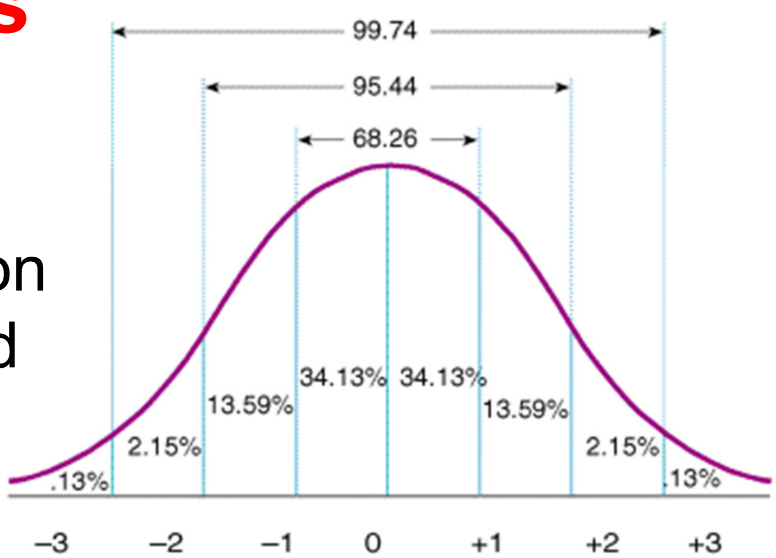
- Is 1.8 within the usual variability of the population?

Z-Scores

- We can express this 1.8 value in terms of standard deviations:
- $Z = \frac{(\bar{x} - \mu)}{\sigma_{\bar{x}}} = \frac{1.8}{0.19} = 9.47$
- The sample result is 9.47 standard deviations above the expectation!
- Standard score or Z-Score
- So what does Z=9.47 mean?
  - $\bar{x} > \mu: 50\%$
  - $\bar{x} > \mu + \sigma: 16\%$
  - $\bar{x} > \mu + 2\sigma: 2.3\%$
  - $\bar{x} > \mu + 1.645\sigma: 5\%$



Z-Scores



- We have transformed our sample distribution into another distribution centered around 0.0 and standard deviation 1.0: Z-distribution
- Standard normal distribution
- With this distribution, we can define cut-off points for different levels of significance:
- One-Tailed:
  - $p \leq 0.05 : Z \geq 1.645$ and $p \leq 0.01 : Z \geq 2.33$
- Two-Tailed
  - $p \leq 0.05 : Z \geq 1.96 \; or \; Z \leq -1.96$
  - $p \leq 0.01 : Z \geq 2.58 \; or \; Z \leq -2.58$

Z-Test

- The Z-Test does 3 things:
  - Estimates the sampling distribution of the mean
  - Transform this distribution into a standard normal distribution
  - Express sample mean $\bar{x}$ as Z standard deviations from μ
- Finding critical values
  - $\bar{x}_{crit} = \mu \pm 1.96\sigma_{\bar{x}}$
  - For a two-tailed test, $\mu = 2.5, \sigma = 1.2, N = 27$:
    $\bar{x}_{crit} = 2.5 \pm 1.96\left(\frac{1.2}{\sqrt{27}}\right) = \pm 2.9$
  - So for values larger or smaller than 2.9 we reject H_0 with $p \leq 0.05$ (5% significance level)

What have we learned?

1. Good hypotheses have to be falsifiable in practice
2. We define H_0 and H_1 and try to reject H_0
3. When we reject, there is a chance that we are wrong
4. P-Values are a measure of the probability to falsely reject H_0
5. We can either use a calculated P-Value directly as the strength of evidence against H_0 or set bounds and reject H_0 if the p-value is within the rejection regions
6. We have to find sampling distributions to make decisions!
7. For means we can use a Z-Test

The perfect Apple Experiment

- **1. Round**
  - Review the measures for the dependent variables (in turns)
    - Discuss especially validity and reliability
    - Result: A rated list of measures
  - Do the same for the independent variables

The perfect Apple Experiment

- **2. Round**
  - Discuss the procedures to reduce the effects of the following factors:
    - The knife used for cutting
    - Pre-contamination of the apple(s)
  - Combine the procedures into one (or two) detailed procedure addressing this problem.
    What concepts are you using and why?

Procedure

- Cutting the apple
  - Take a clean kitchen knife
  - Cut apple in half, label the halves A and B
  - Rotate plate by 90 degrees
  - Cut both halves into 3 equal slices without changing the position
  - Label the top half from left to right C, L, H
  - Label the lower half from right to left C, L, H
  - Rotate back by 90 degrees and cut both slices vertically into 6 pieces.
  - Add label U to the outside slices, L to the inside
  - Results are lables like ACU, etc for each apple