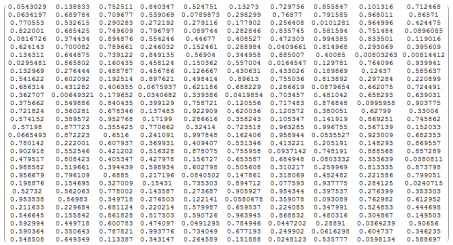
Universität Potsdam
Institut für Informatik
Lehrstuhl Maschinelles Lernen

Hypothesis testing

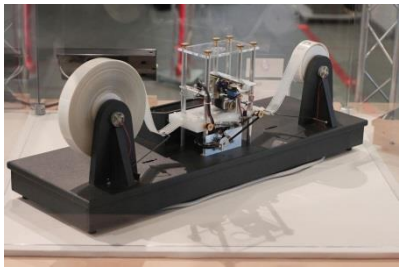
Anna Wegloop
Niels Landwehr/Tobias Scheffer

Why do a statistical test?

input



computer



output

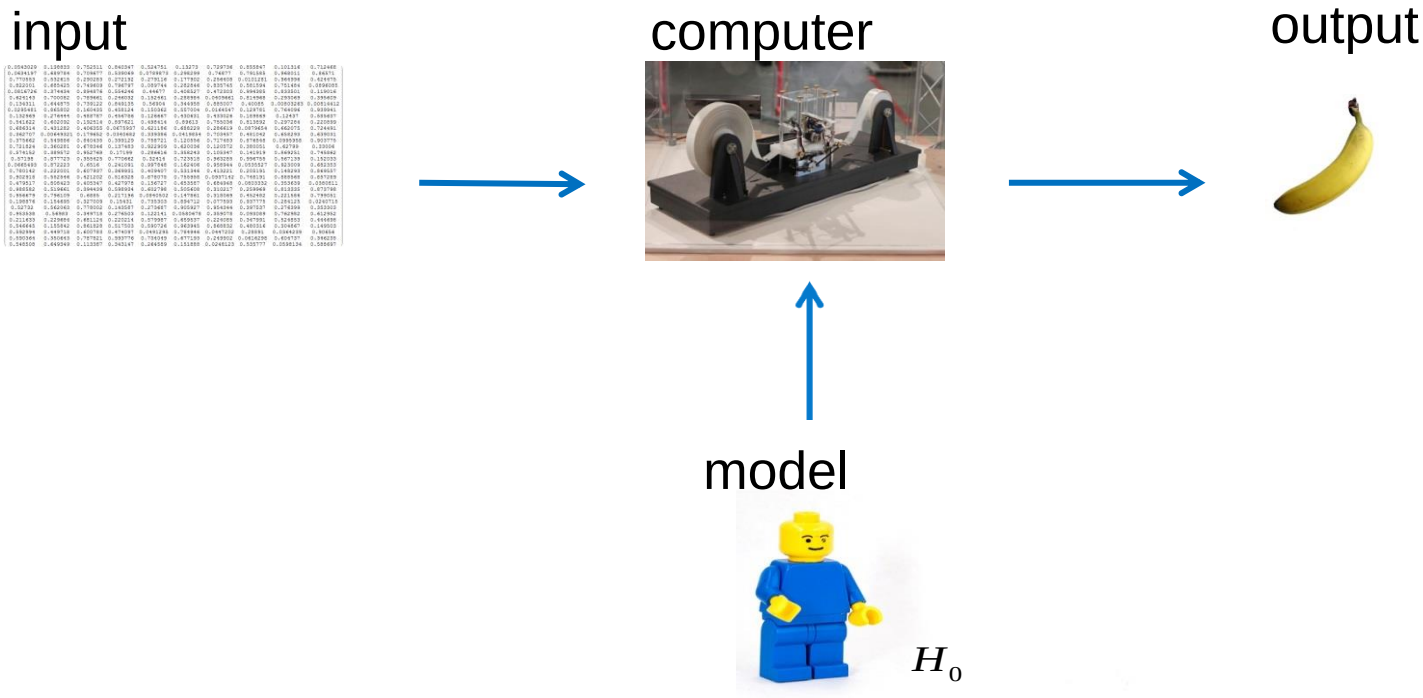
model

Outlook

- Null-hypothesis
- Some more concepts involved in hypothesis testing
  - Central limit theorem
  - Confidence interval
  - Critical value
  - P-values
  - Significance level
- One sample t-test
- Pearson´s chi-squared test
- Sign test
- Likelihood ratio test

Definition of the null hypothesis

- Given a model/hypothesis: H_0
- Are the data plausible, given the model?
- If not, choose alternative model/hypothesis: H_a



We will use the central limit theorem

- Let $\{X_1, X_2, ..., X_N\}$ be i.i.d. random variables drawn from a distribution with mean μ and variance $\sigma^2 < \infty$
- Let be $\overline{X} = \frac{1}{N}\sum_{i=1}^{N} X_i$ the sample mean of a sample of size N
- Then the central limit theorem (CLT) follows:

$$\lim_{N\to\infty} p(\sqrt{N}\left(\overline{X} - \mu\right)) = \mathcal{N}\left(0, \sigma^2\right)$$

Using the CLT in an example

- A list with 5000 weights of people stored in a database
- In an experiment, we only us a sample of 16 weights. The sample mean is $\overline{X} = 73$ kg. The sample variance is $\sigma^2_{sample} = 3$ kg.
- What is the chance that the population mean differs 2kg or less from the sample mean?

Example: confidence interval

- Let $P\left(\overline{X} - b \leq \mu \leq \overline{X} + b\right) = 0.9$ be the probability that the population mean μ differs b or less from a sample mean $\overline{X}$
- Then we call the $\left[\overline{X} - b, \overline{X} + b\right]$ 90% confidence interval

Example: confidence interval

- Derivation confidence interval

$$\frac{\sqrt{N}(\bar{X}-\mu)}{\sigma} \sim \mathcal{N}(0,1)$$ [approximately, for large N, because of CLT]

- Define $\hat{\sigma}^2 = \frac{1}{N}\sigma^2_{sample}$

- Then approximately $\frac{\bar{X}-\mu}{\hat{\sigma}} \sim \mathcal{N}(0,1)$

- Therefore

$$P(\mu \le \bar{X}+b) = P(\mu-\bar{X} \le b) = P\left(\frac{\mu-\bar{X}}{\hat{\sigma}} \le \frac{b}{\hat{\sigma}}\right) \approx \Phi\left(\frac{b}{\hat{\sigma}}\right)$$

with $\Phi(x)$ the cumulative distribution function of $\mathcal{N}(1,0)$

Example: confidence interval

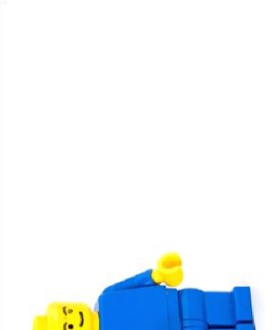
- For $b = \hat{\sigma}\ \Phi^{-1}(0.95)$, it will approximately hold that $P\left(\mu \leq \bar{X} + b\right) = 0.95$

- Because of symmetry, $P\left(\overline{X} - b \leq \mu \leq \overline{X} + b\right) = 0.9$ then approximately holds.

When to reject the null hypothesis?

- Check how likely the data are, given H_0:
  - Sample data
  - Define test statistic
  - See if sample is consistent with null hypothesis

- If very unlikely, reject H_0



The test statistic

- Let $\{X_1, X_2, ..., X_N\}$ be random variables
- Let $\{x_1, x_2, ..., x_N\}$ be their sample values
- We want to test whether our model $p^M_{H_0}(X_1, X_2, ..., X_N)$ for the true distribution is likely to be correct
- Define test statistic $t = t(X_1, X_2, ..., X_N)$ with distribution $p(t)$
- t measures some attribute of the sample
- Let $p_{H_0}(t)$ be the distribution over t under the assumption that the null-hypothesis holds
- Calculate $t_0 = t(x_1, x_2, ..., x_N)$
- One sided test, e.g.: $P_{H_0}(t > t_0)$

Two sided and one sided tests

- Tests can be either single-sided

$$H_0 : \theta \leq \theta_0 \text{ versus some } H_a : \theta > \theta_0$$

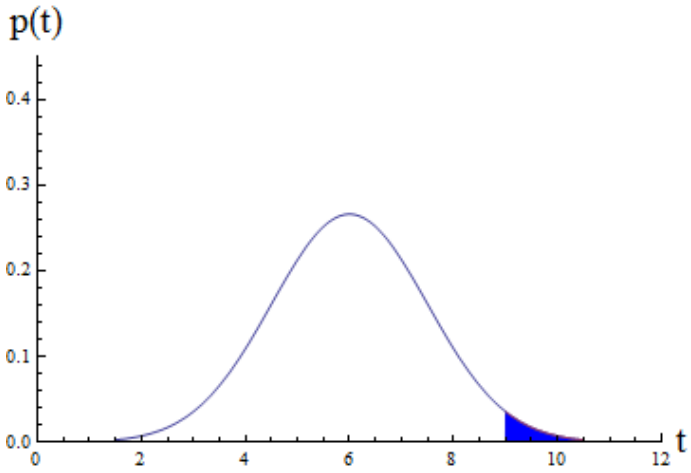
- Or two-sided

$$H_0 : \theta = \theta_0 \text{ versus some } H_a : \theta \neq \theta_0$$

- Where θ, θ_0 are attributes of the random variables

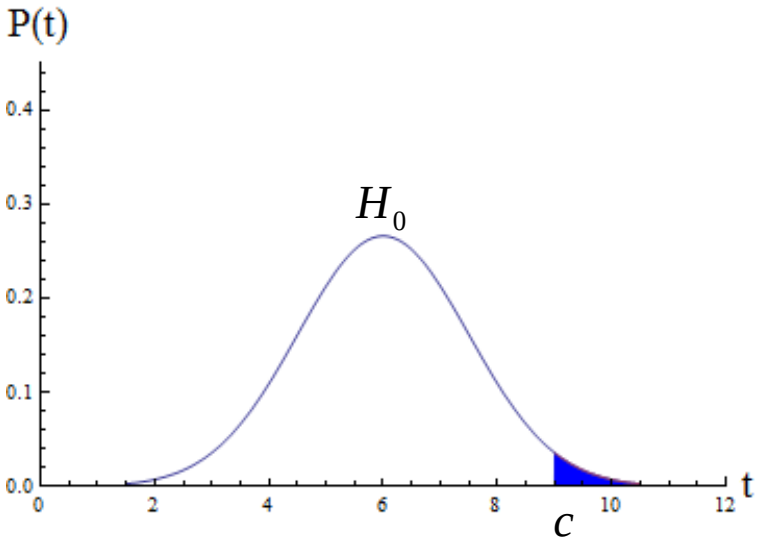
Testing procedure

- Define test statistic t
- Calculate value t_0 of test statistic for sample
- Calculate p-value: $P_{H_0}(t > t_0)$
- Reject H_0 with predefined significance level α (corresponding to the critical value c) if $P_{H_0}(t > t_0) < \alpha$

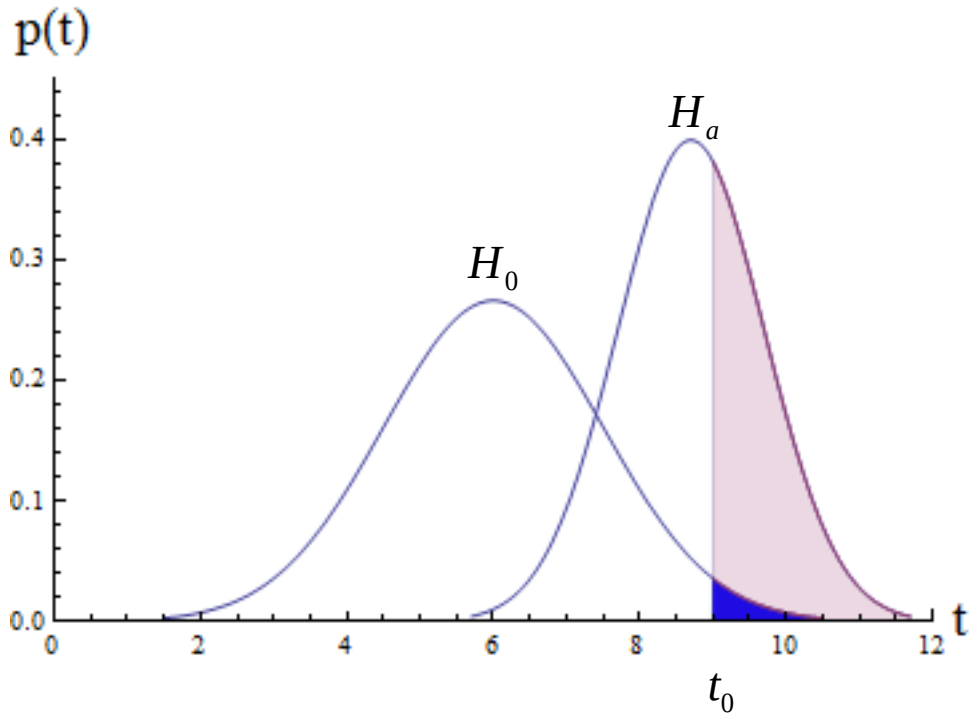


Critical value

- Area under the distribution $\int_{-\infty}^{c} p_{H_0}(\mathrm{t})\,\mathrm{dt} = CDF(\mathrm{c}) = 1-\alpha$
- Critical value: $c = CDF^{-1}(1-\alpha)$
- Reject H_0 if $t_0 > c$. This is the same as: $P(t > t_0) < \alpha$



Comparing hypotheses



Examples of statistical tests

- One sample t-test
- Pearsons chi-squared test
- Sign test
- Likelihood ratio test

Example: one sample t-test

- Sample $\{x_1, x_2, ..., x_N\}$ of weights
- H_0 :The population mean μ_0 is 80 kg

Assumptions of the (one sample) t-test

- The data generating variables $X_1, X_2, ..., X_N$ are independent of each other

- Means of the random variables are normally distributed (assumption is often satisfied due to CLT).

$$p(\overline{X}) = \mathcal{N}\left(\overline{X}|\mu,\sigma^2\right)$$

Test statistic of the t-test (single sample)

- Corrected estimate of the standard deviation of a sample of size N.

$$\sigma_{sample} = \sqrt{\frac{\sum_{i=1}^{N}(X_i)^2 - \left(\frac{1}{N}\sum_{i=1}^{N} X_i\right)^2}{N-1}}$$

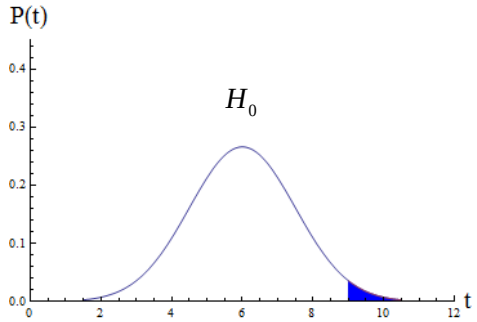
- Test statistic:

$$t = \frac{\frac{1}{\sqrt{N}}\sum_{i=1}^{N} X_i - \mu_0}{\sigma_{sample}} = \frac{\sqrt{N}(\bar{X} - \mu_0)}{\sigma_{sample}}$$

Example: one sample t-test

- Data set $\{x_1, x_2, \ldots, x_N\}$ of weights
- Is the population mean μ_0 80 kg?

$$t = \frac{\sqrt{N}(\bar{X} - \mu_0)}{\sigma_{sample}}$$



Derive the t-distribution from assumptions

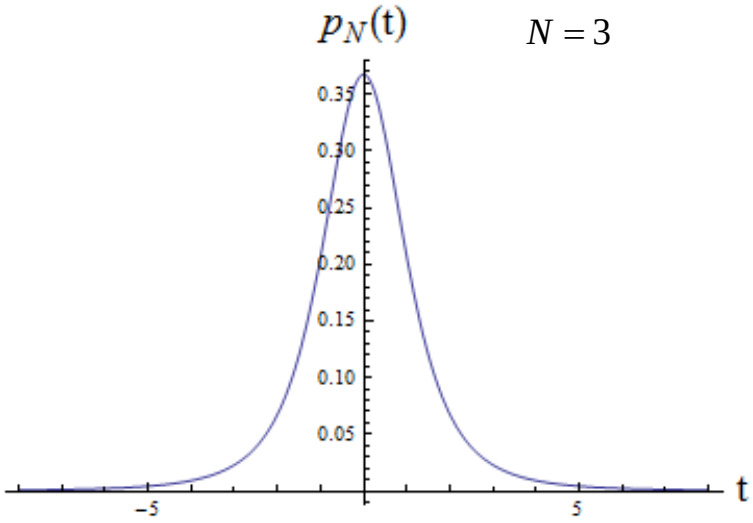
- Test statistic $t = \frac{\sqrt{N}(\bar{X} - \mu_0)}{\sigma_{sample}} = \frac{\sqrt{N-1}u}{\sqrt{v}}$
- With $v = (\mathrm{N}-1)\frac{\sigma_{sample}^2}{\sigma^2}$ and $u = \sqrt{N}\frac{(\overline{X} - \mu_0)}{\sigma}$ and σ^2 the population variance
- If we assume (for this proof) that X_i normally distributed, then:

$$p_{H_0}(\mathrm{u}) = \frac{1}{2\pi} e^{-\frac{1}{2}u^2}$$ normal distribution

$$p_N(v) = \frac{1}{2^{\frac{N}{2}}\Gamma\left(\frac{N}{2}\right)} v^{\frac{N}{2}-1} e^{-\frac{v}{2}}$$ chi-squared distribution

- Then $p_{H_0,N}(\mathrm{t}) = \int \delta\left(\mathrm{t} - \frac{u}{\sqrt{v/\mathrm{N}}}\right) \mathrm{p}(\mathrm{u})\,\mathrm{p}_N(\mathrm{v})\,\mathrm{dudv}$ is the t-distribution

The t-distribution



$$p_{H_0,N}(t) = \frac{\Gamma\left(\frac{N+1}{2}\right)}{\sqrt{N\pi}\,\Gamma\left(\frac{N}{2}\right)}\left(1+\frac{t^2}{N}\right)^{-\frac{N+1}{2}}$$

$$\lim_{N\to\infty} p_N(t) = \mathcal{N}(0,1)$$

Pearson`s χ^2-Test

- Let $\{X_1,\ldots,X_N\}$ i.i.d., from a multinomial distribution $X_i=\left(X_i^1,\ldots,X_i^k\right)$ with $X_i^j\in\{0,1\}$ and $\|X_i\|_1=1$ and expectation value $\mu=\left(\mu^1,\ldots,\mu^k\right)$

- We want to test $H_0:\mu=\mu_0$ vs. $H_a:\mu\neq\mu_0$

- Test statistic t obeys a chi-squared distribution

$$t=\sum_{j=1}^{k}\frac{\left(\overline{X}^j-{\mu_0}^j\right)^2}{{\mu_0}^j} \qquad \overline{X}^j=\frac{1}{N}\sum_{i=1}^{N}X_i^j$$

Sign test

- Test whether the distributions of two random variables X, Y have zero difference median
  $H_0 : P(\mathrm{X} > \mathrm{Y}) = \frac{1}{2}$
- Collect pairs $\{(\mathrm{X}_1, \mathrm{Y}_1), (\mathrm{X}_2, \mathrm{Y}_2), ..., (\mathrm{X}_N, \mathrm{Y}_N)\}$
- Discard pairs where $\mathrm{X}_i = \mathrm{Y}_i$ with $i \in \{1, 2, ..., \mathrm{N}\}$
- We keep $M \leq N$ pairs
- Let t be the number of pairs for which $X_i - Y_i > 0$
- Assuming H_0, it follows that: $p_{H_0}(t) = Binom(\mathrm{M}, \frac{1}{2})$

Likelihood ratio test

- Compare likelihood f of data given model H_0 with likelihood of data given model H_a
- H_0 is a special case of H_a
- Test statistic: $t = -2\log\left(\frac{f(\mathrm{x}_1, \mathrm{x}_2, ..., \mathrm{x}_N \mid \theta_a)}{f(\mathrm{x}_1, \mathrm{x}_2, ..., \mathrm{x}_N \mid \theta_0)}\right)$
- Under certain conditions, the test statistic approaches a chi-squared distribution when the sample size approaches infinity

Wald test

- Test $H_0 : \theta = \theta_0$ vs. $H_1 : \theta \neq \theta_0$
- Test statistic $t = \sqrt{N}\,\frac{\hat{\theta} - \theta_0}{\sigma}$
- When the null-hypothesis is true, then:

$$p_{H_0}(t) \simeq \mathcal{N}(0,1)$$

- With σ the estimator of the standard deviation of $\hat{\theta}$

Summary

- Use statistical test to decide whether the null hypothesis is likely given a sample: define significance level, calculate p-value
- One sample t-test: assume $\{X_1, X_2, ..., X_N\}$ i.i.d., means normally distributed
- Chi-squared test: assume $\{X_1, X_2, ..., X_N\}$ i.i.d., normally distributed
- Two sample sign test: assume pairwise samples i.i.d
- Likelihood ratio test: assume test statistic is chi-squared distributed
- Wald test: assume likelihood estimator normally distributed