

Universität Potsdam  
Institut für Informatik  
Lehrstuhl Maschinelles Lernen



# Hypothesis testing

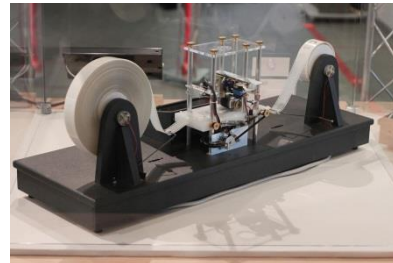
Anna Wegloop  
Niels Landwehr/Tobias Scheffer

# Why do a statistical test?

input

```
0.0549029 0.138893 0.752511 0.840297 0.524751 0.13273 0.729734 0.855847 0.101316 0.712468  
0.0636197 0.689784 0.702977 0.539049 0.078979 0.298209 0.748777 0.791585 0.948011 0.84571  
0.770593 0.528515 0.290283 0.271929 0.279516 0.177802 0.284608 0.0101261 0.844996 0.424476  
0.822005 0.685425 0.749609 0.794797 0.089714 0.282844 0.830748 0.801594 0.751484 0.0396008  
0.0818744 0.874394 0.484874 0.158244 0.444677 0.408257 0.472023 0.394385 0.833001 0.119514  
0.424143 0.700082 0.783641 0.240332 0.152441 0.288984 0.0493641 0.814848 0.229349 0.038409  
0.124311 0.644876 0.739122 0.449105 0.146004 0.444938 0.488207 0.444938 0.444938 0.00020249  
0.0293483 0.845802 0.160435 0.456124 0.150562 0.857004 0.0146457 0.129781 0.744636 0.039461  
0.132369 0.274444 0.448791 0.449704 0.224607 0.404651 0.430208 0.109469 0.12437 0.088497  
0.844622 0.602092 0.182514 0.897421 0.489414 0.89613 0.758034 0.813092 0.297254 0.220899  
0.684214 0.431232 0.408315 0.087937 0.421156 0.404651 0.284613 0.0879464 0.462075 0.724491  
0.862707 0.00449321 0.179452 0.0340682 0.339286 0.0119954 0.702457 0.451042 0.658293 0.639031  
0.375442 0.549856 0.440435 0.139129 0.708721 0.120254 0.717483 0.874444 0.0995008 0.007778  
0.712824 0.340281 0.478344 0.171463 0.922809 0.422034 0.120372 0.380051 0.42739 0.33004  
0.374132 0.389072 0.82746 0.171399 0.246414 0.324243 0.105247 0.161919 0.569251 0.743642  
0.37195 0.877723 0.353425 0.770642 0.32414 0.723518 0.943285 0.994755 0.847139 0.132033  
0.0649493 0.072223 0.4534 0.241001 0.897468 0.144244 0.959444 0.0335257 0.923009 0.682353  
0.780142 0.222001 0.407937 0.349351 0.499407 0.513144 0.413223 0.205191 0.142933 0.699597  
0.902316 0.525144 0.421232 0.151328 0.470076 0.743084 0.097742 0.745191 0.885448 0.877198  
0.479517 0.808423 0.403347 0.427878 0.154727 0.403387 0.644944 0.005332 0.333439 0.838011  
0.885382 0.131461 0.394349 0.189204 0.400788 0.004688 0.101237 0.299469 0.813315 0.873798  
0.954679 0.794509 0.6885 0.211194 0.0840302 0.147841 0.138069 0.452482 0.221254 0.789051  
0.189744 0.154995 0.287009 0.181401 0.733303 0.494742 0.077893 0.397775 0.284125 0.0240715  
0.12712 0.542043 0.778002 0.143587 0.273687 0.905927 0.954344 0.452482 0.276939 0.333303  
0.033303 0.549853 0.349718 0.274909 0.122141 0.090678 0.139078 0.080809 0.762882 0.412882  
0.11633 0.229484 0.681124 0.202124 0.579887 0.609307 0.224055 0.347391 0.244853 0.444488  
0.144444 0.150542 0.461825 0.151709 0.500734 0.463945 0.868332 0.802134 0.304847 0.80003  
0.892994 0.449718 0.600783 0.474897 0.049129 0.784944 0.0447202 0.28951 0.0346239 0.30626  
0.190364 0.300463 0.78121 0.991774 0.734949 0.471133 0.243903 0.0414329 0.404717 0.346235  
0.448008 0.649349 0.113397 0.343147 0.244989 0.101888 0.0248122 0.035777 0.0099134 0.088497
```

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



# We will use the central limit theorem

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

$$\lim_{N \rightarrow \infty} p(\sqrt{N}(\bar{X} - \mu)) = \mathcal{N}(0, \sigma^2)$$

# Using the CLT in an example

- A list with 5000 weights of people stored in a database
- In an experiment, we only use a sample of 16 weights. The sample mean is  $\bar{X} = 73$  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  $P(\bar{X} - b \leq \mu \leq \bar{X} + b) = 0.9$  be the probability that the population mean  $\mu$  differs  $b$  or less from a sample mean  $\bar{X}$
- Then we call the  $[\bar{X} - b, \bar{X} + b]$  90% confidence interval

# Example: confidence interval

- Derivation confidence interval

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

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

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

- Therefore

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

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  $P(\mu \leq \bar{X} + b) = 0.95$
- Because of symmetry,  $P(\bar{X} - b \leq \mu \leq \bar{X} + b) = 0.9$  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  $\{X_1, X_2, \dots, X_N\}$  be random variables
- Let  $\{x_1, x_2, \dots, x_N\}$  be their sample values
- We want to test whether our model  $p_{H_0}^M(X_1, X_2, \dots, X_N)$  for the true distribution is likely to be correct
- Define test statistic  $t = t(X_1, X_2, \dots, X_N)$  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  $t_0 = t(x_1, x_2, \dots, x_N)$
- One sided test, e.g.:  $P_{H_0}(t > t_0)$

# Two sided and one sided tests

- Tests can be either single-sided

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

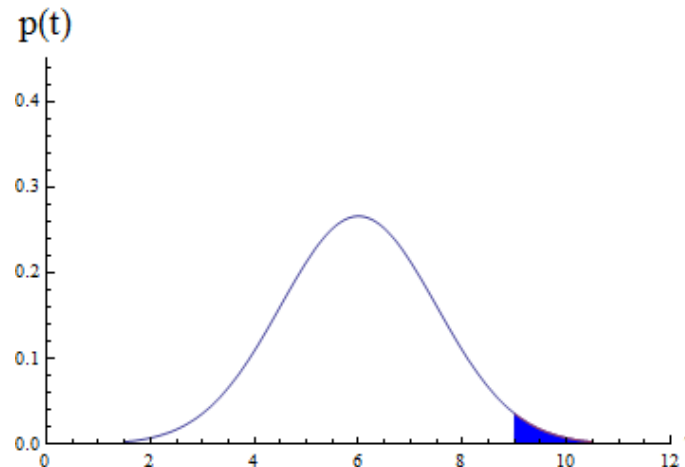
- Or two-sided

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

- Where  $\theta, \theta_0$  are attributes of the random variables

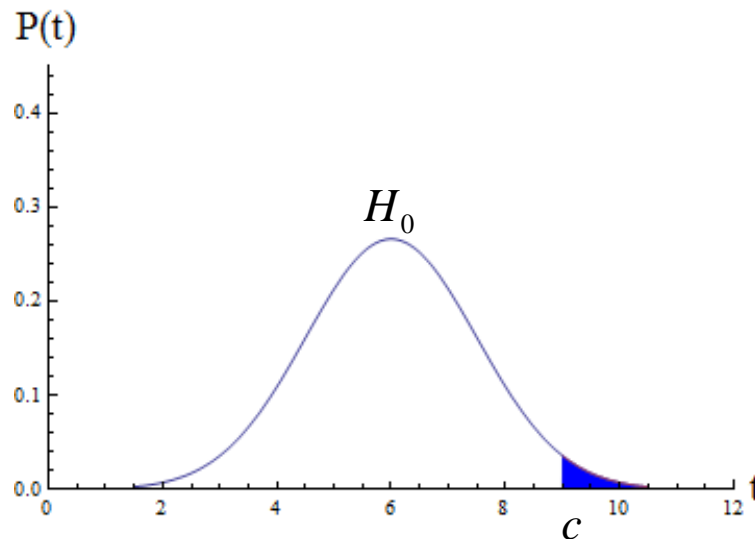
# Testing procedure

- 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  $\alpha$  (corresponding to the critical value  $c$ ) if  $P_{H_0}(t > t_0) < \alpha$

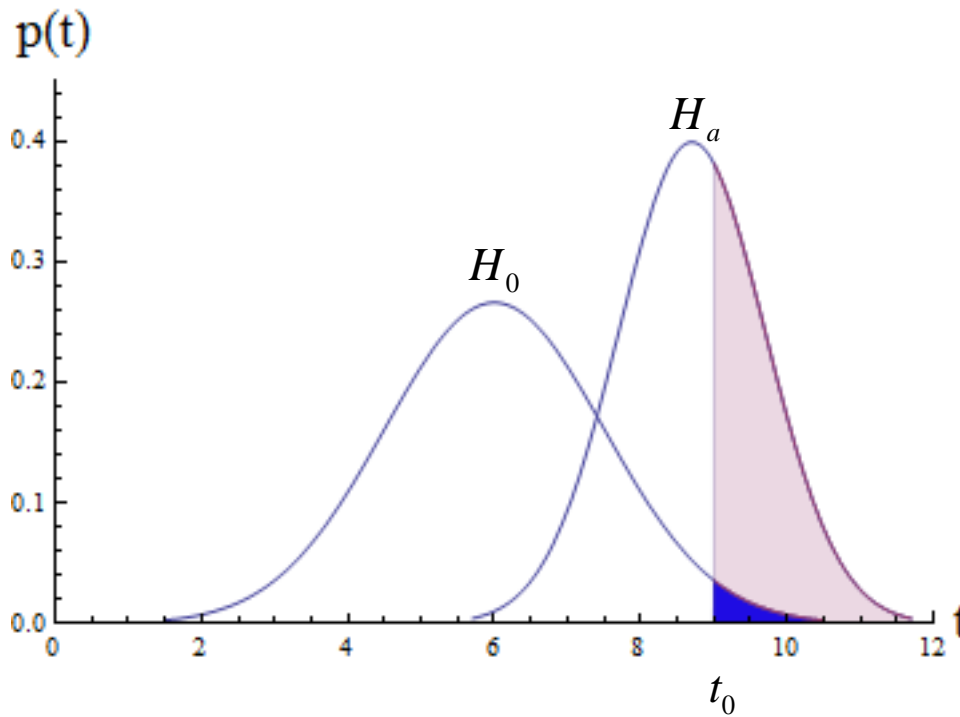


# Critical value

- Area under the distribution  $\int_{-\infty}^c p_{H_0}(t) dt = CDF(c) = 1 - \alpha$
- 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  $\{x_1, x_2, \dots, x_N\}$  of weights
- $H_0$ : The population mean  $\mu_0$  is 80 kg

# Assumptions of the (one sample) t-test

- The data generating variables  $X_1, X_2, \dots, X_N$  are independent of each other
- Means of the random variables are normally distributed (assumption is often satisfied due to CLT).

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

# Test statistic of the t-test (single sample)

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

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

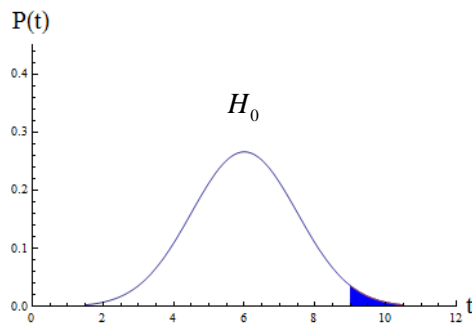
- Test statistic:

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

# Example: one sample t-test

- Data set  $\{x_1, x_2, \dots, x_N\}$  of weights
- Is the population mean  $\mu_0$  80 kg?

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



# 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  $v = (N-1) \frac{\sigma_{sample}^2}{\sigma^2}$  and  $u = \sqrt{N} \frac{(\bar{X} - \mu_0)}{\sigma}$   
and  $\sigma^2$  the population variance

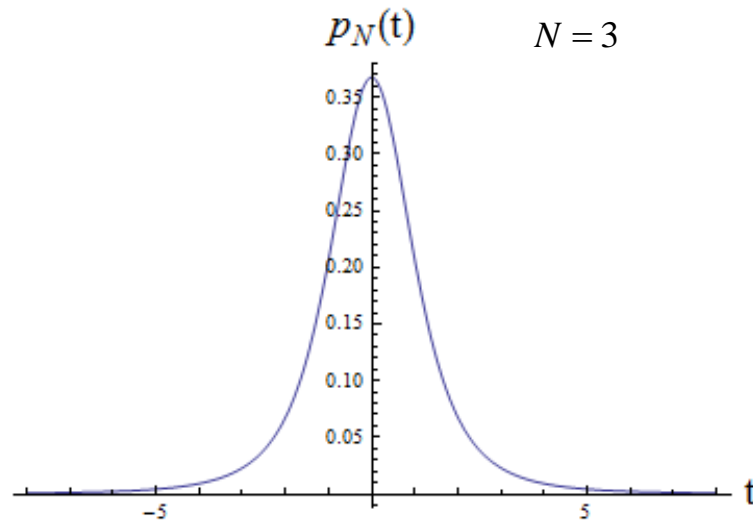
- If we assume (for this proof ) that  $X_i$  normally distributed, then:

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

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

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

# The t-distribution



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

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

# Pearson's $\chi^2$ -Test

- Let  $\{X_1, \dots, X_N\}$  i.i.d., from a multinomial distribution  
 $X_i = (X_i^1, \dots, X_i^k)$  with  $X_i^j \in \{0, 1\}$  and  $\|X_i\|_1 = 1$   
and expectation value  $\mu = (\mu^1, \dots, \mu^k)$
- We want to test  $H_0 : \mu = \mu_0$  vs.  $H_a : \mu \neq \mu_0$
- Test statistic  $t$  obeys a chi-squared distribution

$$t = \sum_{j=1}^k \frac{(\bar{X}^j - \mu_0^j)^2}{\mu_0^j} \quad \bar{X}^j = \frac{1}{N} \sum_{i=1}^N X_i^j$$

# Sign test

- Test whether the distributions of two random variables  $X, Y$  have zero difference median
$$H_0 : P(X > Y) = \frac{1}{2}$$
- Collect pairs  $\{(X_1, Y_1), (X_2, Y_2), \dots, (X_N, Y_N)\}$
- Discard pairs where  $X_i = Y_i$  with  $i \in \{1, 2, \dots, N\}$
- 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:  $p_{H_0}(t) = \text{Binom}(M, \frac{1}{2})$



# 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: 
$$t = -2 \log \left( \frac{f(x_1, x_2, \dots, x_N | \theta_a)}{f(x_1, x_2, \dots, x_N | \theta_0)} \right)$$
- 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:

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

- With  $\sigma$  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  $\{X_1, X_2, \dots, X_N\}$  i.i.d., means normally distributed
- Chi-squared test: assume  $\{X_1, X_2, \dots, X_N\}$  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