Search Engines

WS 2009 / 2010

Lecture 14, Thursday February 11th, 2010

(Statistical Hypothesis Testing)

Prof. Dr. Hannah Bast
Chair of Algorithms and Data Structures
Department of Computer Science
University of Freiburg

Overview of Today´s Lecture

- Learn what it means
  - that a difference between two results is statistically significant
  - for example
    - the running times of two programs
    - or their space consumption
    - or the precision of two search engines
    - or anything ...
  - also learn about the pitfalls of statistical tests
    - the interpretation of the results is everything!

Hypothesis Testing — Example 1

- A guy says he can predict
  - whether it will SNOW tomorrow or not
  - let´s observe the truth of his prediction on 10 days

    TTTTTTTTTT (he is right on all 10 days)
  - do we believe in this guy´s ability to predict?
- Here is how hypothesis testing answers this question
  - null hypothesis H_0 = he can´t predict
    - mathematically: $Pr(T) = ½$
  - now compute the probability of the observed (or more extreme) data assuming that the null hypothesis is true
    - $Pr(\text{all ten correct} \mid H_0) = 2^{-10} \leq 0.001 = 0.1\%$
  - we can reject the null hypothesis with probability $\geq 99.9\%$

Hypothesis Testing — Example 1

- Let´s now assume we observed

  TTTFTTTTFT (he is right on 8 of the 10 days)

  – do we still believe in this guy´s ability to predict?

  – note that it takes some non-trivial interpretation when precisely formulating the event of "the observed or more extreme data"

$$Pr(\text{right} \geq 8 \text{ times} \mid H_0)$$

$$= \binom{10}{8} \cdot 2^{-10} + \binom{10}{9} \cdot 2^{-10} + \binom{10}{10} \cdot 2^{-10}$$

$$= (45 + 10 + 1) \cdot 2^{-10}$$

$$= 56 \cdot 2^{-10} \approx 0.056 = 5.6\%$$

$$\frac{10 \cdot 9}{2 \cdot 1}$$

Hypothesis Testing — General

- Formulation
  - a hypothesis H (e.g. the guy can predict SNOW)
  - the null hypothesis H_0 = the opposite of H
- Test

  room for interpretation here!

  - Compute the probability p of the given or more extreme data assuming that the null hypothesis is true
- Outcome
  - $p \leq \alpha = 0.05$ ⇨ H_0 rejected with significance level 5%

    one says: the observed data is statistically significant for H
  - $p > \alpha = 0.05$ ⇨ H_0 cannot be rejected

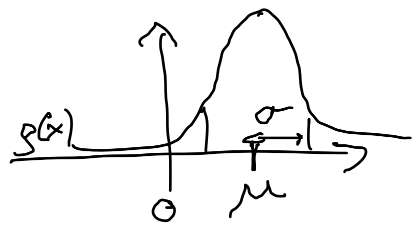
    one says: the observed data is not statistically significant for H

Hypothesis Testing — Example 2

- Difference between two means
  - like the difference between the averaged running times of two programs ... or their space consumption, or precision, or ...
  - is the difference random or statistically significant?

    D1: 8.6 4.3 3.2 5.1   Ø = 5.3

    D2: 2.1 4.2 7.6 3.2 2.9   Ø = 4.0
  - null hypothesis H_0 = the means are equal , namely 4.6
  - what is the probability of observing D1 and D2 given H_0 ?
  - we need to make some assumptions so that we can compute this probability
    - data has a normal distribution (see next slide)
    - the variance is the same for both D1 and D2



The normal distribution

- First recall some facts about random variables

$$E(X+Y) = E(X) + E(Y)$$

even if they are not independent

$$\mathrm{var}(X) = E((X - E(X))^2)$$
$$= E(X^2 - 2 \cdot X \cdot E(X) - E(X)^2) = E(X^2) - (EX)^2$$

$$\mathrm{var}(X+Y) = \mathrm{var}(X) + \mathrm{var}(Y)$$

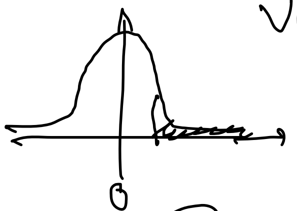
if X and Y are independent

$$\mathrm{var}(\alpha X) = \alpha^2 \cdot \mathrm{var}(X)$$

- And about the normal distribution

$N(\mu, \sigma^2)$ mean μ, variance σ^2

The sum of independent normal random variables is again normal.



$$N(\mu_1, \sigma_1^2) + N(\mu_2, \sigma_2^2) = N(\mu_1 + \mu_2, \sigma_1^2 + \sigma_2^2)$$

$$\Pr(N(0,1) \geq x) = \tfrac{1}{2}\left(1 - \mathrm{erf}\left(\frac{x}{\sqrt{2}}\right)\right)$$

$$\mathrm{erf}(x) = \frac{2}{\sqrt{\pi}} \int_0^x e^{-t^2}\, dt$$

$\mu_1 = 5.3$ [average of D_1] $\quad m_1 = 4$

$\mu_2 = 4.0$ (average of D_2) $\quad m_2 = 5$

$\mu = 4.6$ (overall average)

$\sigma^2 = 4.2$ (overall variance)

$(*)\quad N(0, m_1 \cdot \sigma^2) = \underline{\sqrt{m_1}\,\sigma} \cdot N(0,1)$

$$\Pr\left(N(m_1 \cdot \mu, m_1 \cdot \sigma^2) \geq m_1 \cdot \mu_1\right)$$

$$= \Pr\left(N(0, m_1 \cdot \sigma^2) \geq m_1 \cdot (\mu_1 - \mu)\right)$$

$$\stackrel{(*)}{=} \Pr\left(N(0,1) \geq \underbrace{\sqrt{m_1} \cdot \frac{\mu_1 - \mu}{\sigma}}_{= 0.683130\ldots \;=: x_1}\right)$$

$$= \tfrac{1}{2}\left(1 - \operatorname{erf}\left(\tfrac{x_1}{\sqrt{2}}\right)\right) = 0.247262 \approx 24.7\%$$

$$\Pr\left(N(m_2 \cdot \mu, m_2 \cdot \sigma^2) \leq m_2 \cdot \mu_2\right) = \ldots$$

Hypothesis Testing — BEWARE

- What one would ideally like:
  - given this data, what is the probability that my hypothesis if true?
  - formally: Pr(H | D)
- What one gets from hypothesis testing
  - given that my hypothesis is true, what is the probability of this (or more extreme) data
  - formally: Pr(D | H)
  - but Pr(D | H) could be low for other reasons than the hypothesis!!
- So is hypothesis testing useful at all?
  - OK: challenge theory by attempting to reject it
  - NO: confirm theory by rejecting corresponding null hypothesis

$Pr(H \mid D)$ versus $Pr(D \mid H)$

2% of the population have some sickness S

There is a test $Pr(\text{YES} \mid S) = 0.95 = 95\%$

$Pr(\text{No} \mid \neg S) = 0.97 = 97\%$

$Pr(S) = 2\% = 0.02$

$$Pr(S \mid \text{YES}) = Pr(\text{YES} \mid S) \cdot \frac{Pr(S)}{Pr(\text{YES})}$$

$$= \frac{0.95 \cdot 0.02}{Pr(\text{YES} \mid S) \cdot Pr(S) + Pr(\text{YES} \mid \neg S) \cdot Pr(\neg S)}$$

$$= \frac{0.95 \cdot 0.02}{0.95 \cdot 0.02 + 0.03 \cdot 0.98}$$

$$= 0.4$$

$Pr(\neg S \mid \text{YES}) = 0.6 = 60\%$

NORMAL	SICK		
949	1	950	NO
30	20	50	YES
979	21	1000	

References

- Read the wonderful articles by Jacob Cohen
  - Things I have learned (so far)
    American Psychologist, 45(12):1304–1312, 1990
  - The earth is round ($p < .05$)
    American Psychologist 49(12):997–1003, 1994
- Wikipedia articles
  - http://en.wikipedia.org/wiki/Statistical_hypothesis_testing

... and don´t forget the EXAM

on Friday, March 12, 14:00 – 17:00 h in HS 026

(oral exams scheduled on the same day)