# Information Retrieval

## WS 2012 / 2013

### Lecture 13, Wednesday February 6th, 2013
### (Hypothesis testing, statistical significance)

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

# Overview of this lecture

- Organizational
  - Your results + experiences with Ex. Sheet 12 (Ontologies)
  - The official **evaluation** of this course
- Hypothesis Testing
  - How to determine whether an observed effect is what is called statistically significant ?
  - A must in (not only) information retrieval research
  - In particular: the Z-test and Student's T-test
  - **Exercise Sheet 13:** determine the statistical significance of a simple database performance optimization (string ids → int ids)

# Experiences with ES#12 (Ontologies)

- Summary / excerpts last checked February 6, 16:00
  - Good to know / learn / refresh some SQL
  - Sadly, I'm too stupid for SQL (out-of-sleep error)
  - Took more time than expected
  - Exercise seemed far fetched … believe me, it's not
  - Jay-Z won award for Best Female Video
  - Single index increased speed, but another index (on the same table) decreased it again … interesting observation !

    Possible explanation: each table can be sorted only according to one column

    In a SPARQL-only database, you would sort according to both columns
  - Is it allowed to use notes in the exam … YES !

# Official course evaluation

- Please follow the link + instructions on the Wiki
  - We are very interested in your feedback
  - Please take your time for this

    You will get 20 wonderful points !
  - Please be honest and concrete
  - The free text comments are of particular interest for us
  - **Please complete it by Friday, February 8**

    and at the very latest by Sunday, February 10 !

# Hypothesis Testing 1/5

- Motivation
  - Typical situation in research: compare the outcome of two experiments

    In the life sciences: two studies

    In computer science: two algorithms / methods
  - **Problem:** how much of the observed difference is "real", and how much is due to random fluctuations

# Hypothesis Testing 2/5

- Example 1: Prediction of coin tosses
  - Ten predictions in a row, C = correct, W = wrong

    CCCCCCCCCC (all ten predictions are correct)
  - Do we believe in this person's ability to predict?
- Hypothesis testing answers this as follows
  - Null hypothesis $H_0$ = the person cannot predict = is just making random guesses … mathematically: $Pr(C) = ½$
  - Compute the probability of the observed (or more extreme) data assuming that $H_0$ is true

    $Pr(\text{all ten correct} \mid H_0) = 2^{-10} \le 0.001 = 0.1\%$
  - We say that we can reject $H_0$ with probability $\ge 99.9\%$

    means: it's unlikely that the great prediction was mere chance

# Hypothesis Testing 3/5

- Example 1: continuation
  - Let's assume, in a different series we get

    CCCWCCCWCC (8 correct, 2 wrong)
  - What is the probability now, that this is due to chance?
  - Note: it takes some non-trivial interpretation when formalizing "... of the observed or more extreme data"

$$\binom{10}{8} = \frac{10 \cdot 9}{1 \cdot 2}$$

$$\binom{10}{9} = \frac{10}{1}$$

$$\text{Prob}(\geq 8 \text{ correct})$$

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

$$= (45 + 10 + 1) \cdot 2^{-10} = 56 \cdot 2^{-10} > 5\%$$

# Hypothesis Testing 4/5

- General terminology
  - We start with a hypothesis H (ability to predict coin tosses)
  - Null hypothesis $H_0$ = the opposite of H (random guessing)
  - Statistical test: compute the probability p of the given or "more extreme data" assuming that $H_0$ is true
  - Typical outcome: for a given $\alpha$, say $0.05 = 5\%$

    $p \leq \alpha = 0.05$ ⇨ $H_0$ 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
  - The exact significance level p is often simply called **p-value**

# Hypothesis Testing 5/5

- Example 2: two series of measurements
  - For example, accuracies of two classification methods

    A1 : 0.87 0.88 0.87 0.90

    A2 : 0.87 0.86 0.85 0.86
  - Null hypothesis $H_0$ = the means are equal
  - Given $H_0$, what is the probability of observing A1 and A2
  - We need assumptions on the underlying prob. distribution

    Z-Test: assume normal distribution with fixed variance

    T-Test: like Z-test, but also model variance distribution
  - The T-Test is more realistic but (slightly) more complex

# Probability distributions 1/4

- General terminology
  - Continuous random variable $X$ = range is $\mathbf{R}$
  - Probability density function $\varphi(x) = \Pr(X = x)$
  - Cumulative distribution function $\Phi(x) = \Pr(X \le x)$
  - **Mean** of the distribution $\mu = \mathbf{E}\,X$
  - **Variance** of the distr. $\sigma^2 = \mathbf{E}\,(X - \mathbf{E}\,X)^2 = \mathbf{E}\,X^2 - (\mathbf{E}\,X)^2$

    $\sigma$ is often called the **standard deviation**
  - Recall linearity properties of $\mathbf{E}$ and var :

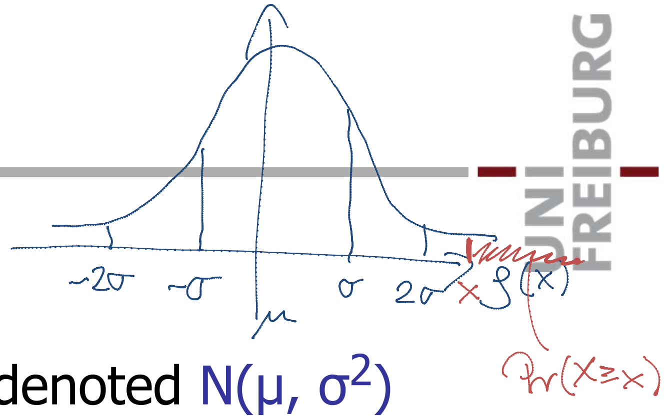
    $\mathbf{E}\,(X + Y) = \mathbf{E}\,X + \mathbf{E}\,Y$ even if $X$ and $Y$ are dependent

    $\mathrm{var}(X + Y) = \mathrm{var}(X) + \mathrm{var}(Y)$ only if $X$ and $Y$ independent

    $\mathrm{var}(a \cdot X) = a^2 \cdot \mathrm{var}(X)$ by $\mathrm{var}(X) = \mathbf{E}\,X^2 - (\mathbf{E}\,X)^2$ above

# Probability distributions 2/4

CENTRAL LIMIT THEOREM




- The normal distribution
  - There is exactly one for each $\mu$ and $\sigma$, denoted $N(\mu, \sigma^2)$
  - Density function $\varphi(x) = \exp(-(x-\mu)^2/2\sigma^2) / (\sigma \cdot \mathrm{sqrt}(2\pi))$
  - Assumed as the underlying distribution in many scenarios

    In the life sciences as well as in computer science !
  - For hypothesis testing, we need to compute, for a given $x$

    $\Pr(X \geq x) = 1 - \Pr(X \leq x)$ … the so-called $p$-value for $x$
  - That's an integral over $\varphi(x)$, no closed formulas for that
  - Either lookup in a table or use tools like **Wolfram Alpha**

    e.g. Wolfram Alpha knows $\mathrm{erf}(x)$, where $\Phi(x) = (1+\mathrm{erf}(x/\sqrt{2}))/2$
  - Lemma: if $X$ has dist $N(\mu, \sigma^2)$, then $(X-\mu) / \sigma$ has dist $N(0,1)$

# Probability distributions 3/4

- The $\chi^2$ distribution $\chi$ = small Greek letter "chi"
  - Assume $Z_1, \ldots, Z_n$ randomly picked according to $N(0, 1)$
  - Then the distribution of $Z = Z_1^2 + \ldots + Z_n^2$ is defined as:

    the $\chi^2$ distribution with $n$ degrees of freedom aka $\chi^2(n)$
  - Why is this a practically relevant distribution ?

    Consider measurements $X_1, \ldots, X_n$ , each from $N(\mu, \sigma^2)$

    Let $M = \Sigma X_i / n$ be the sample mean, $\mathbf{E}\, M = \mu$

    Let $S^2 = \Sigma (X_i - M)^2 / n$ be the sample variance, $\mathbf{E}\, S^2 = \sigma^2$

    Then $S^2 \cdot n / \sigma^2$ has a $\chi^2(n)$ distribution

    **Intuitively:** the variance of a series of measurements has a $\chi 2$ distribution (up to scaling)

# Probability distributions 4/4

sum of normal distributions is ~~aa~~ again normal distribution

independent identically distributed

- The Student's t-distribution
  - Again, let $X_1, \ldots, X_n$ be i.i.d. from $N(\mu, \sigma^2)$
  - Let $M = \Sigma X_i / n$ be the sample mean, $\mathbf{E}\, M = \mu$
  - Let $S^2 = \Sigma (X_i - M)^2 / n$ be the sample variance, $\mathbf{E}\, S^2 = \sigma^2$
  - Then $Z = (M - \mu) / \sigma \cdot \sqrt{n}$ has distribution $N(0, 1)$ … slide 11
  - And $V = S^2 \cdot n / \sigma^2$ has distribution $\chi^2(n)$ … slide 12
  - Definition: the **(Student's) t-distribution** is defined as the distribution of $T = (M - \mu) / S \cdot \sqrt{n}$ with $M$ and $S$ as above
  - Note: $Z$ and $V$ depend on $\sigma$, but $T = Z / \mathrm{sqrt}(V/n)$ does not **!**
  - Intuitively: the deviation from the mean of a series of measurements with unknown variance has a t-distribution

UNI FREIBURG

# Z-test

- Assumption: underlying normal distribution
  - Given two series $X1$ and $X2$ of a total of $n$ measurements
  - Let $M = M1 - M2$ be the difference of the sample means
  - Let $S^2 = S1^2 + S2^2$ be the sum of the sample variances
  - The $Z$-test assumes that $\sigma = S$ ... this is quite unrealistic
  - Hypothesis: $\mathbf{E}\,M > 0$ ($\mathbf{E}\,M < 0$ or $\mathbf{E}\,M \neq 0$ analogously)
  - Assume the null hypothesis: $\mathbf{E}\,M = 0$
  - Then $Z = (M - \mu) / \sigma \cdot \sqrt{n}$ has distribution $N(0, 1)$
  - Compute value $z$ of $Z$ for given measurements
  - The $p$-value is $\Pr(Z \geq z)$ ... estimate via table or Wolfram Alpha

    http://en.wikipedia.org/wiki/Standard_normal_table

# T-test

- Assumption: underlying $t$-distribution
  - Given two series $X1$ and $X2$ of a total of $n$ measurements
  - Let $M = M1 - M2$ be the difference of the sample means
  - Let $S^2 = S1^2 + S2^2$ be the sum of the sample variances
  - The $t$-test does not need an estimate of $\sigma$ **!**
  - Hypothesis: $\mathbf{E}\, M > 0$ ($\mathbf{E}\, M < 0$ or $\mathbf{E}\, M \neq 0$ analogously)
  - Assume the null hypothesis: $\mathbf{E}\, M = 0$
  - Then $T = M / S \cdot \sqrt{n}$ has a $t$-distribution
  - Compute value $t$ of $T$ for given measurements
  - The $p$-value is $\Pr(T \geq t)$ ... estimate via table or Wolfram Alpha

    http://en.wikipedia.org/wiki/T-distribution#Table_of_selected_values

# A working example

$S_1^2 = 0.01^2 + 0.01^2 + 0.02^2 = 6 \cdot 10^{-4}$

$S_2^2 = 0.01^2 + 0.01^2 = 2 \cdot 10^{-4}$

with $n=8$

- Both Z-test and T-Test

$n = 8$

Z has dist $N(0,1)$

T has t-distribution

(more "conservative" than $N(0,1)$)

  - Recall the accuracy measurements from slide 9

    A1 : 0.87 0.88 0.87 0.90 $\quad M_1 = 0.88$, $S_1^2 = 6 \cdot 10^{-4}$

    A2 : 0.87 0.86 0.85 0.86 $\quad M_2 = 0.86$, $S_2^2 = 2 \cdot 10^{-4}$

  - Difference M of sample means is: $M_1 - M_2 = 0.02 = 2 \cdot 10^{-2}$
  - Sum $S^2$ of sample variances is: $S_1^2 + S_2^2 = 8 \cdot 10^{-4}$
  - Value x of $M / S \cdot \sqrt{n}$ is: $2 \cdot 10^{-2} / (\sqrt{8} \cdot 10^{-2}) \cdot \sqrt{8} = 2$
  - **Z-test:** estimate for $Pr(Z \geq x)$ is: $0.0228 \approx 2\%$

    $Pr(Z \leq 2) = 0.9772$

    0228
  - **T-test:** estimate for $Pr(T \geq x)$ is: $\approx 4\%$

    $Pr(T \geq 1.86) = 0.05$

    $Pr(T \geq 2.306) = 0.025$

# Background for Exercise Sheet 13

- Let's consider a simple database optimization trick
  - Replace **string ids** in the TSV tables by **integer ids**
  - In the SQL CREATE command then say INTEGER for that column instead of TEXT

    This tells the DB engine to store the values as ints internally
  - This seems to save some time, but maybe that is simply because the integers are more compact than string
  - So repeat the same with **hexadecimal ids**
  - **Exercise:** determine the statistical significance of the performance difference between hex ids and integer ids

    try both Z-test and T-test; and try 3 and 10 measurements

# References

- Wikipedia
  - http://en.wikipedia.org/wiki/Statistical_hypothesis_testing
  - http://en.wikipedia.org/wiki/P-Value
  - http://en.wikipedia.org/wiki/Z-test
  - http://en.wikipedia.org/wiki/Student's_t-test
  - http://en.wikipedia.org/wiki/Student's_t-distribution
- Two articles by Jacob Cohen

  an American statistician and psychologist, 1923 – 1998

  The earth is round ($p < 0.05$)

  Things I have learned (so far)

  Quite entertaining + instructive !