Information Retrieval

WS 2013 / 2014

Lecture 13, Tuesday February 4th, 2014
(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 ?
  - This is a **must** when the observed effect is small, and the variation is large
  - Specifically today: R(andomization)-Test, Z-test, T-test
  - **Exercise Sheet 13:** determine the statistical significance of the difference between BM25 and TF-IDF on five queries

Experiences with ES#12 (Ontologies)

- Summary / excerpts last checked February 4, 15:30
  - Liked the introduction to Ontologies & SPARQL
  - Too much effort for parsing the SPARQL query (in C++)

    Sorry, I forgot to say that Python would have been ok for this exercise too, since efficiency was not the issue
  - Constant values were not explained … settled on Forum

Official course evaluation

- Instructions
  - You received an email from EvaSys Admin on Thursday, January 30 with a link to an online evaluation form
  - 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 most interesting for us

    Please complete by Sunday, February 9

    The evaluation is centralized this time, and will be closed after that date, and there is nothing we can do about that

Hypothesis Testing   1/5

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

    E.g. in the **life sciences**: health status for two groups of people, one taking a particular medication and one not

    E.g. in **computer science**: the performance of two systems, using different algorithms or different parameter settings
  - The outcome of the experiments will be different

    But even carrying out the same experiment twice will give different results because of random fluctuations

    Key question: how to tell a "real" difference between the two experiments from mere random fluctuation

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 data, under the assumption that the null hypothesis H_0 is true

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

    Thus very unlikely that this great prediction was mere chance

Hypothesis Testing 3/5

$\binom{n}{k} = \frac{n!}{k!\,(n-k)!}$

- Example 1: continuation

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

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

  – Let's assume, in a different series we get

    CCCWCCCWCC (8 correct, 2 wrong)

  – What is now the probability that this is due to chance?

    Note: we should **not** ask for the probability of **exactly 8** correct guesses to happen; it makes more sense to ask for the prob. of **8 or more** correct guesses to happen

$$\Pr(\geq 8 \text{ correct} \mid H_0) = \binom{10}{8} \cdot 2^{-10} + \binom{10}{9} \cdot 2^{-10} + 2^{-10}$$

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

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

UNI FREIBURG

Hypothesis Testing 4/5

- General terminology
  - We start with a hypothesis H   e.g. ability to predict coin tosses
  - Null hypothesis H_0 = the opposite of H   e.g. random guessing
  - **Statistical test:** compute the probability p of the given or more extreme data assuming that H_0 is true

    This probability p is called the **p-value**

    If p is small enough, one says something like:

    The outcome of the experiment is **statistically significant** (for the hypothesis) with significance level p

    In the life sciences, people are usually happy with $p < 0.05$ or $p < 0.01$

Hypothesis Testing 5/5

- Example 2: two dice with unknown distribution
  - Two dice A and B, four rolls each

    A : 1 , 3 , 3 , 5

    B : 6 , 6 , 4 , 4

    H = the two dice have a different distribution
  - Null hypothesis H_0 = the two dice A and B are identical
  - Given H_0, what is the probability of observing A and B
  - We will look at three well-known statistical tests

    R-Test: simple + makes no probabilistic assumptions

    Z-Test: assume normal distribution with fixed variance

    T-Test: like Z-test, but also model variance distribution

R(andomization)-Test 1/3

- One of the simplest statistical tests
  - Assume we have two series of measurements, A and B
  - Null hypothesis = no difference between A and B
  - Then we can assume that the measurements come from one experiment + assignment to either A or B is arbitrary
  - The R-Test considers all 2^n possible assignments of the n measurements to either A or B
  - For each assignment, compute the difference $\Delta\mu$ of the means, and see if it is $\geq$ the $\Delta\mu$ on the observed data

    The fraction of assignments for which this is the case is the p-value according to the R-Test

R(andomization)-Test 2/3

- Application to our dice example

  A : 1 , 3 , 3 , 5 $\quad \mu_A = (1+3+3+5)/4 = 3$

  B : 6 , 6 , 4 , 4 $\quad \mu_B = (6+6+4+4)/4 = 5$

  $\Rightarrow \underline{\Delta\mu = 2}$

  – Here are some of the 2^8 possible assignments of these 8 measurements to either A or B and the respective Δμ

    Note: we ignore the two assignments, where all measurements are assigned all to A or all to B, because we can't compute a meaningful mean difference then

	1	3	3	5	6	6	4	4	
≥ 2	A	A	A	A	B	B	B	B	$\rightarrow \Delta\mu = \frac{1+3+3+5}{4} - \frac{6+6+4+4}{4} = 2$
< 2	A	A	A	A	A	A	B	B	$\rightarrow \Delta\mu = \frac{1+3+3+5+6+6}{6} - \frac{4+4}{2} = 0$
< 2	A	B	A	B	A	A	B	A	$\rightarrow \Delta\mu = \frac{1+3+6+6+4}{5} - \frac{3+5+4}{3} = \frac{20}{5} - \frac{12}{3} = 4-4 = 0$
...	...								

R(andomization)-Test 3/3

- Continuation of the example
  - Let's write a program together to iterate over all $2^8 - 2$ assignments and compute the p-value as explained

    For 46 out of 254 assignments, the $|\Delta\mu|$ is 2 or more (in either direction)

    $\Rightarrow$ p-value $= 46/254 = 18.1\%$ = really right!
  - **Note:** for a small number n of measurements, we can easily try out (on a computer) all $2^n - 2$ assignments

    But for larger n, this quickly becomes infeasible

    For $n = 30$ we already have $2^{30} \approx 1$ billion assignments

    Then we can take a (large enough) random sample of assignments and compute the fraction for those

Z-Test and T-Test 1/9

- Assumptions
  - The Z-Test and the T-Test both assume an underlying probability distribution
  - Z-Test: underlying **normal distribution**
  - T-Test: underlying **t-distribution**
  - Then, for our setting, the p-value is $Pr(M \geq \Delta\mu)$, where:

    M is a random variable, modelling the difference of the means with the assumed probability distribution

    $\Delta\mu$ is the value of M on the observed measurements

    As a preparation, let us recap (on the next slides) some foundations from probability theory …

Z-Test and T-Test 2/9

- General terminology
  - Continuous random variable X = range is $\mathbf{R}$
  - Cumulative distribution function $\Phi(x) = \Pr(X \le x)$

    In particular: $\lim_{x\to\infty} \Phi(x) = 1$
  - **Mean** of the distribution $\mu = \mathbf{E}\,X$
  - **Variance** of the distr. $\sigma^2 = \mathbf{E}\,(X - \mathbf{E}\,X)^2 = \mathbf{E}\,X^2 - (\mathbf{E}\,X)^2$

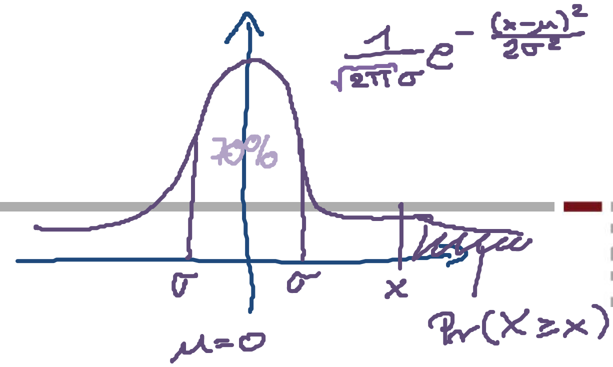
    The sqrt σ of the variance is known as standard deviation
  - Basic linearity properties of $\mathbf{E}$ and $\mathbf{var}$:

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

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

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

Z-Test and T-Test 3a/9



- The **normal distribution**
  - Assumed as the underlying distribution in many scenarios

    In the life sciences as well as in computer science
  - Two parameters: the mean μ and the variance σ^2

    The corresponding distribution is denoted by $N(\mu, \sigma^2)$
  - We will need to compute $Pr(X \geq x)$ where X has normal dist.

    Beware: there is no closed formula for this

    In the ancient past, lookup tables were used

    Nowadays, just use a tool like **Wolfram Alpha** and type

    "Pr(X >= 2.3) for standard normal distribution"

    "Pr(X >= 2.3) for t-distribution with 8 degrees of freedom"

Z-Test and T-Test 3b/9

standard
normal
distribution

- Properties of the normal distribution
  - **Property 1:** If X has distribution $N(\mu, \sigma^2)$, then $(X - \mu) / \sigma$ has distribution $N(0, 1)$

    Every normal distr. can be reduced to N(0, 1) by scaling
  - **Property 2:** If X_1 has distribution $N(\mu_1, \sigma_1^2)$ and X_2 has distribution $N(\mu_2, \sigma_2^2)$, and X_1 and X_2 are independent then $X_1 + X_2$ has distribution $N(\mu_1 + \mu_2, \sigma_1^2 + \sigma_2^2)$

    The sum of normal random variables is again normal
  - **Property 3:** Let $X_1,\ldots,X_n$ be n **i.i.d.** (independent identically distributed) random variables, each with mean μ and variance σ^2 . Then $(X_1 + \ldots + X_n) / n$ converges to $N(\mu, \sigma^2)$ as $n \rightarrow \infty$

    Property 3 is also known as the Central Limit Theorem

Z-Test and T-Test 4/9

Zas distribution
$N(0,1)$

- The χ^2 **distribution** χ = small Greek letter "chi"
  - Let $Z_1, \ldots, Z_n$ be i.i.d. from $N(0, 1)$
  - Then the distribution of $Z = Z_1^2 + \ldots + Z_n^2$ is defined as: the χ^2 distribution with n degrees of freedom aka $\chi^2(n)$
  - Why this is a practically relevant distribution:

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

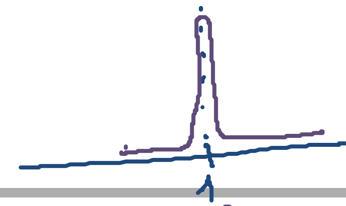
    Let $M = \Sigma X_i / n$ be the estimated mean, $\mathbf{E}\, M = \mu$

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

    $EV = n$ Then $S^2 \cdot n / \sigma^2 = \Sigma ((X_i - M) / \sigma)^2$ has a $\chi^2(n)$ distribution

    Intuitively: the variance of a series of measurements has a χ2 distribution (up to scaling)

Z-Test and T-Test   5/9



V/n for large n

- The **Student's t-distribution**
  - Let us define it by how we pick a random X from it, in comparison to the standard normal distribution:

    Standard Normal distribution (μ = 0, σ = 1):

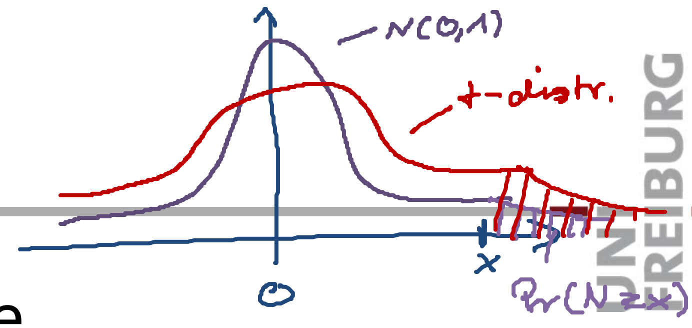
    Pick X from N(0, 1)

    T-distribution with n degrees of freedom: E V/n = 1

    First pick V from $\chi^2(n)$, then pick X from N(0, V / n)
  - Note that **E** V = n (slide 17) and that for $n \to \infty$ we have $V / n \to 1$ and the two distributions become the same

    Actually, there is a marked difference between the two distributions only for small n, say n ≤ 50

Z-Test and T-Test 6/9



- More intuition about the difference
  - By also considering the variance as a random variable, the t-distribution is **less concentrated** around its mean than the corresponding normal distribution
  - Here is an example which provides some intuition

    Experiment 1: pick X uniformly from [-10, 10]

    Experiment 2: first pick V uniformly from [5, 15], then pick X uniformly from [-V, V]

    Now extreme values (< -10 or > 10) become more likely, and values around the mean become less likely

    Note that the mean remains zero in Experiment 2

Z-Test and T-Test 7/9

- The **Z-Test** assumption: underlying **normal distribution**
  - Given two series X_1 and X_2 of a total of n measurements
  - Let $M = M_1 - M_2$ be the difference of the means of X_1 and X_2
  - Let $\sigma^2 = \Sigma (X_i - \mu)^2 / n$ be the estimated variance, $\mu = \Sigma_i X_i / n$
  - **Null hypothesis:** M has distribution $N(0, 4\sigma^2 / n)$
  - Then $Z = \sqrt{n} \cdot M / (2\sigma)$ has distribution $\mathbf{N(0, 1)}$
  - The p-value of the Z-Test is then $Pr(M \geq \Delta\mu) = Pr(Z \geq x)$, where $x = \sqrt{n} \cdot \Delta\mu / (2\sigma)$ and $\Delta\mu$ is the observed value of M

    Estimate via Wolfram Alpha (see slide 15) or via lookup table:

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

Z-Test and T-Test 8/9

- The **T-Test** assumption: underlying **t-distribution**
  - Given two series X_1 and X_2 of a total of n measurements
  - Let $M = M_1 - M_2$ be the difference of the means of X_1 and X_2
  - Let $\sigma^2 = \Sigma (X_i - \mu)^2 / n$ be the estimated variance, $\mu = \Sigma_i X_i / n$
  - **Null hypothesis:** M has distribution $N(0, V \cdot 4\sigma^2 / n^2)$, where V has distribution $\chi^2(n)$ with n deg. of freedom … see slide 17
  - Then $T = \sqrt{n} \cdot M / (2\sigma)$ has **t-distrib.** with n deg. of freedom
  - The p-value of the T-Test is then $\Pr(M \geq \Delta\mu) = \Pr(T \geq x)$, where $x = \sqrt{n} \cdot \Delta\mu / (2\sigma)$ and $\Delta\mu$ is the observed value of M

    Estimate via Wolfram Alpha (see slide 15) or via lookup table:

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

Z-Test and T-Test 9/9

- Back to our rolling dice example $\mu = \frac{\mu_A + \mu_B}{2} = 4$
  - Recall our two series of dice rolls

    A : 1 , 3 , 3 , 5 $\mu_A = \frac{1+3+3+5}{4} = 3$

    B : 6 , 6 , 4 , 4 $\mu_B = \frac{6+6+4+4}{4} = 5$

  - Difference of means $\Delta\mu$ is: 2
  - Estimated variance σ^2 is: $\frac{(1-4)^2 + (3-4)^2 + (3-4)^2 + (5-4)^2 + (6-4)^2 \cdot 2 + (4-4)^2 \cdot 2}{8}$

    $= (9 + 1 + 1 + 1 + 4 + 4) / 8 = 20/8 = 2.5$

  - Value x of $\sqrt{n} \cdot \Delta\mu / (2\sigma)$ is: $= \sqrt{8} \cdot 2 / (2\sqrt{2.5}) \approx 1.789$
  - **Z-test:** p-value $\Pr(Z \geq x)$ is: $\approx 0.0368 = 3.68\%$
  - **T-test:** p-value $\Pr(T \geq x)$ is: $\approx 0.0520 = 5.20\%$

    For "two-sided" p-values, simply multiply by 2

References

- Further reading

  Smucker, Allan, Carterette: A Comparison of Statistical Significance Tests for IR Evaluation, CIKM 2007

  http://ciir-publications.cs.umass.edu/getpdf.php?id=744

- 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