WELCOME TO:

STAT.1030:
Tilastotieteen perusteet
Introduction to Statistics

Bernd Pape
University of Vaasa
Department of Mathematics and Statistics

TERVETULOA!

www.uwasa.fi/∼bepa/TilPer.html

Contents

Midterm exam: chapters 1–4
Final exam: chapters 5–8

0. A survival kit of mathematical tools

Exponentiation (Potenssifunktiot)

Exponentiation with base (kantaluku) $a \neq 0$ and positive integer exponent (eksponentti) n is defined as an n times repeated multiplication of a:

$$a^n := \underbrace{a \times \cdots \times a}_{n}.$$

For negative integer exponents we define:

$$a^{-n} := \frac{1}{a^n}, \quad a \neq 0,\ n = 1, 2, 3 \ldots$$

Furthermore:

$$a^0 := 1 \text{ for } a \neq 0$$

and

$$0^n := 0 \text{ for all } n.$$

Basic Identities:

$$a^{m+n} = a^m \cdot a^n,$$
$$(a^m)^n = a^{m \cdot n},$$
$$(a \cdot b)^n = a^n \cdot b^n.$$

Roots (Juuret)

A root r of a number x is any number which yields x when repeatedly multiplied by itself:

$$\underbrace{r \times r \times \cdots \times r}_{n} = x.$$

In terms of exponentiation, r is a root of x if

$$r^n = x.$$

The number n is called the degree (aste) of the root, and a root of degree n is called the n'th root (n's juuri):

$$r = \sqrt[n]{x}.$$

We define fractional powers (murtoeksponentit) as the n'th root of the base:

$$x^{1/n} := \sqrt[n]{x}.$$

The basic identities from the previous page remain valid also for fractional powers.

Logarithms (Logaritmit)

The logarithm of a number x to a given base a is the power or exponent to which the base must be raised in order to produce the number. For example, the logarithm of 1 000 to the base 10 is 3, because 3 is how many 10's you must multiply to get 1000. Formally:

$$a^y = x \quad \Leftrightarrow: \quad y = \log_a x.$$

In this course we will mainly consider the natural logarithm (luonnollinen logaritmi) with base $a = e = 2.7182\ldots$ (Neperin luku) and then drop the subscript a.

Note: You can find the logarithm to any arbitrary base by taking the ratio of logarithms to some other arbitrary base which is on your calculator, for example:

$$\log_2 5 = \frac{\log_{10} 5}{\log_{10} 2} = \frac{\log_e 5}{\log_e 2} \approx 2.32.$$

Basic Identities:

$$\log(x\,y) = \log x + \log y,$$
$$\log\,(x^n) = n \log x.$$

The Summation Sign (Summamerkki)

The symbol $\sum$ is used to denote the sum of the terms that follow it, taken over the range given above and below the symbol:

$$\sum_{i=1}^{n} a_i = a_1 + a_2 + \cdots + a_n.$$

Examples:

$$\sum_{i=1}^{4} x_i = x_1 + x_2 + x_3 + x_4,$$

$$\sum_{i=1}^{3} i^2 = 1^2 + 2^2 + 3^2 = 14.$$

The representation of a sum is not unique:

$$a_4 + a_5 + a_6 = \sum_{i=4}^{6} a_i = \sum_{i=1}^{3} a_{i+3} = \sum_{i=0}^{2} a_{6-i}.$$

Sums and Differences:

$$\sum_{i=1}^{n}(a_i \pm b_i) = \sum_{i=1}^{n} a_i \pm \sum_{i=1}^{n} b_i.$$

Example:

$$\sum_{i=1}^{3}(a_i \pm b_i) = (a_1 \pm b_1) + (a_2 \pm b_2) + (a_3 \pm b_3)$$

$$=(a_1 + a_2 + a_3) \pm (b_1 + b_2 + b_3) = \sum_{i=1}^{3} a_i \pm \sum_{i=1}^{3} b_i.$$

Constant Factors:

$$\sum_{i=1}^{n} ca_i = c\sum_{i=1}^{n} a_i.$$

Example:

$$\sum_{i=1}^{2} ca_i = ca_1 + ca_2 = c(a_1 + a_2) = c\sum_{i=1}^{2} a_i.$$

Constants:

$$\sum_{i=1}^{n} c = nc.$$

Example:

$$\sum_{i=1}^{4} c = c + c + c + c = 4c.$$

Common exam mistake: In general

$$\sum_{i=1}^{n} a_i b_i \neq \left(\sum_{i=1}^{n} a_i \right) \left(\sum_{i=1}^{n} b_i \right).$$

This implies by setting $a_i = b_i$, that generally

$$\sum_{i=1}^{n} a_i^2 \neq \left(\sum_{i=1}^{n} a_i \right)^2.$$

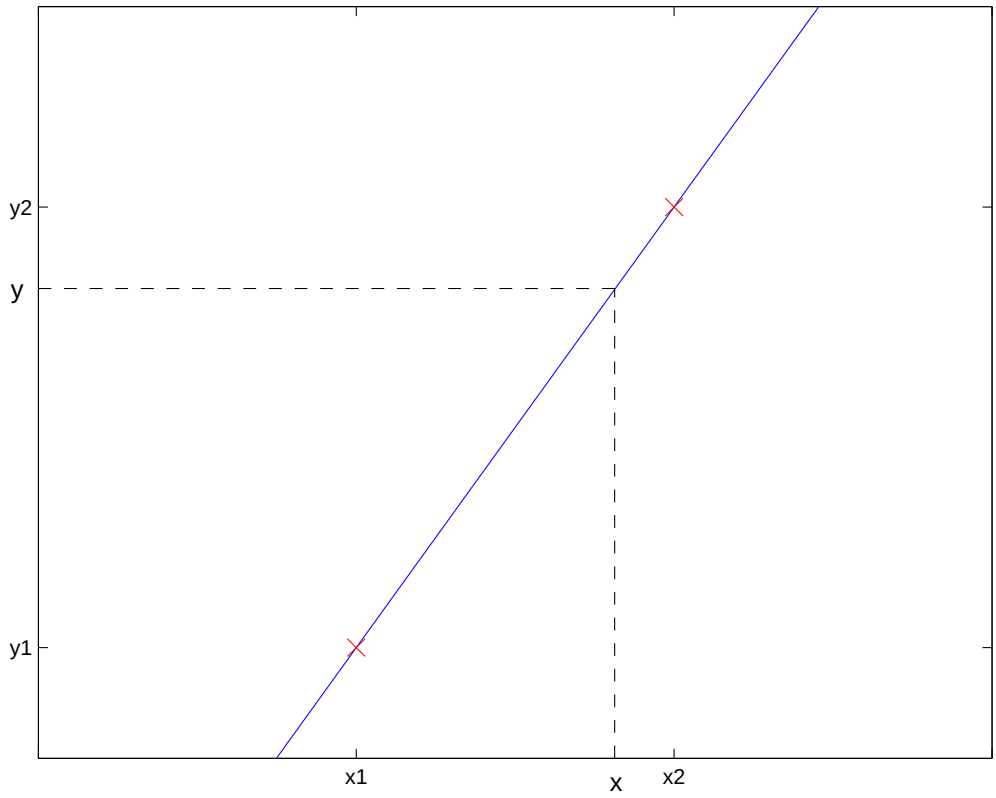
Example:

$$\sum_{i=1}^{2} a_i^2 = a_1^2 + a_2^2, \qquad \text{but}$$

$$\begin{aligned} \left(\sum_{i=1}^{2} a_i \right)^2 &= (a_1 + a_2)(a_1 + a_2) \\ &= a_1^2 + 2a_1 a_2 + a_2^2, \end{aligned}$$

which coincides with the expression above only in the special case that $a_1 a_2 = 0$, that is, only when at least one of the summands a_1 or a_2 is zero.

Linear Interpolation



Suppose you got two data points (x_1, y_1) and (x_2, y_2) and you would like to get an estimate for the ordinate of a point with a given x-coordinate somewhere between x_1 and x_2. Suppose furthermore you got reasons to believe that all the points between (x_1, y_1) and (x_2, y_2) lie on a straight line. Then you may obtain the missing y-coordinate as follows.

The as yet unknown ordinate y may be written as

$$y = y_1 + \Delta y,$$

where $\Delta y = y - y_1$ denotes the distance on the y-axis corresponding to an increase of $\Delta x = x - x_1$ on the x-axis.

The ratio of these distances is called the slope (jyrkkyys) of the line, that is,

$$\text{slope} = \frac{\Delta y}{\Delta x} \quad \Rightarrow \quad \Delta y = \text{slope} \cdot \Delta x.$$

Therefore,

$$y = y_1 + \text{slope} \cdot \Delta x.$$

Now the slope of a line is constant, which means that we might just as well have calculated it as the ratio of the distance between y_1 and y_2 to the distance between x_1 and x_2:

$$\text{slope} = \frac{y_2 - y_1}{x_2 - x_1}.$$

This yields the linear interpolation formula:

$$y = y_1 + \frac{y_2 - y_1}{x_2 - x_1} \cdot (x - x_1).$$

Differentiation (Differentiaalilaskentaa)

The slope of a line tells by how much the value of $y = f(x)$ changes (Δy), when we change the value of x by $\Delta x = 1$ unit.

If f is nonlinear, then the slope is no longer constant, but we may sometimes be able to assign a variable slope to f as the slope of a straight line which is tangent to f at $y = f(x)$. This slope, if it exists, is in general itself a function of x and called the derivative of f with respect to x (f:n derivaatta x:n suhteen). Notation:

$$\frac{dy}{dx} \quad \text{or} \quad \frac{df}{dx}(x) \quad \text{or} \quad \frac{d}{dx}f(x) \quad \text{or} \quad f'(x).$$

Basic Derivatives:

$$f(x) = ax^n \qquad \Rightarrow \qquad f'(x) = anx^{n-1}.$$

In particular: $(ax)' = a$ and $a' = 0$.

$$f(x) = e^x \qquad \Rightarrow \qquad f'(x) = e^x.$$

Differentiation Rules (Derivoimissääntöjä)

Constant Factor: $[a \cdot f(x)]' = a \cdot f'(x)$
Example:

$$f(x) = ae^x \quad \Rightarrow \quad f'(x) = ae^x.$$

Sum Rule: $[f(x) + g(x)]' = f'(x) + g'(x)$
Example:

$$f(x) = x^2 + x \quad \Rightarrow \quad f'(x) = 2x + 1.$$

Product Rule: $[f(x)g(x)]' = f'(x)g(x) + f(x)g'(x)$
Example:

$$[x(x+1)]' = 1 \cdot (x+1) + x \cdot 1 = 2x + 1.$$

Chain Rule: If $f(x) = h(g(x))$, then

$$f'(x) = h'(g(x)) \cdot g'(x).$$

Example:

$$\left(e^{-ax}\right)' = e^{-ax} \cdot (-a) = -ae^{-ax}.$$

Extrema and Derivatives (Ääriarvoja)

x is called a (local) maximum/minimum of f, if $f(x)$ is the largest/smallest value which f attains within a neighbourhood of x. It is called an extremum, if x is either a maximum or a minimum.

If f is differentiable in x (that is, $f'(x)$ exists), then the definition of extrema implies that the tangent to f in x must be parallel to the x-axis, such that its slope is zero. In other words, if f is differentiable in x and x is an extremum, then $f'(x) = 0$.

Note: $f'(x) = 0$ is only a necessary condition (välttämätön ehto) for extrema of differentiable functions, that is, it doesn't guarantee us that x is indeed an extremum of f. It just tells us that x cannot be an extremum if f is differentiable in x and $f(x) \neq 0$. It also doesn't tell us whether x is a maximum or a minimum. Sufficient conditions for extrema and decision criteria whether an extremum is a minimum or a maximum can be found by considering higher derivatives of f.

Partial Derivatives (Osittaisderivaatta)

Let $z = f(x, y)$. Taking the derivative of f with respect to x while keeping y constant is called the partial derivative of f with repect to x. Notation:

$$\frac{\partial z}{\partial x} \quad \text{or} \quad \frac{\partial f}{\partial x} \quad \text{or} \quad \frac{\partial}{\partial x} f(x, y) \quad \text{or} \quad f_x.$$

Example:

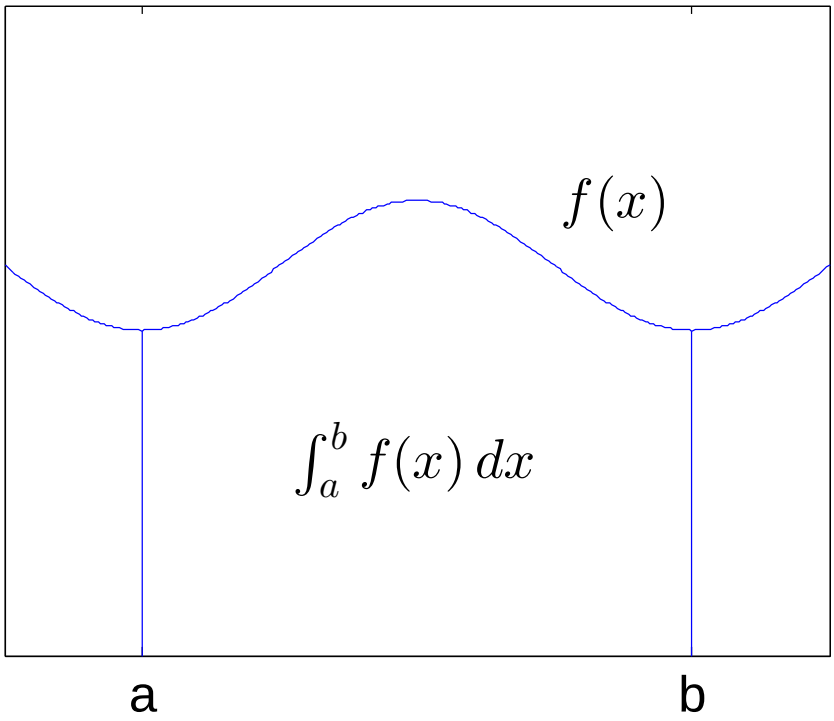
$$f(x, y) = 2x^3 y \quad \Rightarrow \quad \frac{\partial f}{\partial x} = 6x^2 y, \ \frac{\partial f}{\partial y} = 2x^3.$$

An extremum of a differentiable function f at the point (x, y) requires in analogy to the one-dimensional case:

$$\frac{\partial}{\partial x} f(x, y) = 0 \qquad \text{and} \qquad \frac{\partial}{\partial y} f(x, y) = 0.$$

These are again just necessary conditions for extrema and do not allow us to distinguish between minima and maxima. If there is only one point (x_0, y_0) for which $f_x = f_y = 0$, then $f(x_0, y_0)$ is a minimum/maximum if $f(x_0, y_0) \lesseqgtr f(x, y)$ for all 4 pairs $x, y \to \pm\infty$.

The definite Integral



Let f be a continuous nonnegative function of a real variable x and $a \leq b$. Then the region of the plane bounded by the graph of f, the x-axis, and the vertical lines $x = a$ and $x = b$ is called the definite integral of f from a to b (f:n määrätty integraali a:sta b:hen) $\int_a^b f(x)\,dx$.

The value of a definite integral can be determined to any desired precision by approximating the area under the graph as the sum of the surfaces of rectangles bounded by the x-axis and f.

The calculation rules below become evident by considering the geometric interpretation of the definite integral:

$$\int_a^a f(x)\,dx = 0,$$

$$\int_a^b f(x)\,dx + \int_b^c f(x)\,dx = \int_a^c f(x)\,dx \text{ for } a \leq b \leq c,$$

$$\int_a^b c \cdot f(x)\,dx = c \cdot \int_a^b f(x)\,dx,$$

$$\int_a^b f(x) \pm g(x)\,dx = \int_a^b f(x)\,dx \pm \int_a^b g(x)\,dx.$$

F is called an antiderivative (integraalifunktio) of f if $F'(x) = f(x)$. If an antiderivative F of f is known, then we may determine the integral of f from

$$\int_a^b f(x)\,dx = [F(x)]_a^b := F(b) - F(a).$$

Example:

$$\begin{aligned}
\int_{-1}^{1} |x^3|\,dx &= \int_{-1}^{0} |x^3|\,dx + \int_{0}^{1} |x^3|\,dx \\
&= \int_{-1}^{0} -x^3\,dx + \int_{0}^{1} x^3\,dx \\
&= -\frac{1}{4}\int_{-1}^{0} 4x^3\,dx + \frac{1}{4}\int_{0}^{1} 4x^3\,dx \\
&= -\frac{1}{4}\left[x^4\right]_{-1}^{0} + \frac{1}{4}\left[x^4\right]_{0}^{1} \\
&= -\frac{1}{4}\left[0^4 - (-1)^4\right] + \frac{1}{4}\left[1^4 - 0^4\right] \\
&= -\frac{1}{4}\cdot(-1) + \frac{1}{4}\cdot 1 = \frac{1}{2}
\end{aligned}$$

Note that the technique above works only if we can guess an antiderivative of f. If this is not the case (as it often is in statistics), then we need to approximate the area under the curve by rectangles (or other clever shapes). This is done in software packages like Mathematica and matlab and integral tables such as those in Lehtonen/Niemi.

1. Introduction

Statistics is the science of collecting, organizing, analyzing, and interpreting data.

Why should one study statistics?

Because most interesting questions cannot cannot be answered without processing data.

Consider the following hypothetical example presented in a book by Andy Field*: Andy is a psychologist and suspects that most contestants in Big Brother TV shows have a narcissistic personality disorder. Suppose that 75% of all contestants have the disorder, while its level of occurence in the general population is only about 1%. Andy comes up with 2 theories (research hypotheses) explaining his finding.

Theory 1: People with narcissistic personality disorder are more likely to audition for Big Brother than those without.

Theory 2: The producers of Big Brother are more likely to select people with narcissistic personality disorder than people without.

*Discovering Statistics Using SPSS, Sage, 2009

Suppose that Andy applies a personality test to all people turning up for the audition, keeping track of who gets selected for the show, with the following result:

	No Disorder	Disorder	Total
Selected	3	9	12
Rejected	6805	845	7650
Total	6808	854	7662

Theory 1 is supported by the data, because 11%(= 854/7662·100) of the applicants were diagnosed with the disorder, which is much higher than the usual 1%.

Theory 2 is also supported by the data, because 75% of the contestants selected have the disorder, while from the pool of applicants we would have expected a rate of only 11%.