AARHUS UNIVERSITY

Methods 2 - 2

Chris Mathys

SOLIDUM PETIT IN PROFUNDIS · UNIVERSITAS ARHUSIENSIS ·

BSc Programme in Cognitive Science

Spring 2023

Weighted averages

Stratum, j	Label	Population, N_j	Average age, $\bar{y}_j$
1	United States	310 million	36.8
2	Mexico	112 million	26.7
3	Canada	34 million	40.7

Figure 3.1 *Populations and average ages of countries in North America. (Data from CIA World Factbook 2010.) The average age of all North Americans is a* weighted average *of the average ages within each country.*

Weighted averages

$$\begin{aligned}\text{average age} &= \frac{310\,000\,000 * 36.8 + 112\,000\,000 * 26.7 + 34\,000\,000 * 40.7}{310\,000\,000 + 112\,000\,000 + 34\,000\,000} \\ &= 34.6.\end{aligned}$$

$$\begin{aligned}\text{average age} &= \frac{310\,000\,000}{456\,000\,000} * 36.8 + \frac{112\,000\,000}{456\,000\,000} * 26.7 + \frac{34\,000\,000}{456\,000\,000} * 40.7 \\ &= 0.6798 * 36.8 + 0.2456 * 26.7 + 0.0746 * 40.7 \\ &= 34.6.\end{aligned}$$

$$\text{weighted average} = \frac{\sum_j N_j \bar{y}_j}{\sum_j N_j},$$

Vectors and matrices

$$\hat{y} = 46.3 + 3.0x,$$

$$\hat{y} = \hat{a} + \hat{b}x.$$

The expressions $\hat{a}$ and $\hat{b}$ denote estimates—the coefficients 46.3 and 3.0 were obtained by fitting a line to past data—and $\hat{y}$ denotes a predicted value. In this case, we would use y to represent an actual election result, and $\hat{y}$ is the prediction from the model. Here we are focusing on the linear prediction, and so we work with $\hat{y}$.

We can define x as the vector that comprises these three cases, that is $x = (-1, 0, 3)$.

We can put these three predictions together:

$$\begin{aligned}
\hat{y}_1 &= 43.3 = 46.3 + 3.0 * (-1),\\
\hat{y}_2 &= 46.3 = 46.3 + 3.0 * 0,\\
\hat{y}_3 &= 55.3 = 46.3 + 3.0 * 3,
\end{aligned}$$

Vectors and matrices

We can put these three predictions together:

$$
\begin{aligned}
\hat{y}_1 &= 43.3 = 46.3 + 3.0 * (-1),\\
\hat{y}_2 &= 46.3 = 46.3 + 3.0 * 0,\\
\hat{y}_3 &= 55.3 = 46.3 + 3.0 * 3,
\end{aligned}
$$

which can be written as vectors:

$$
\hat{y} = \begin{pmatrix} 43.3 \\ 46.3 \\ 55.3 \end{pmatrix} = \begin{pmatrix} 46.3 + 3.0 * (-1) \\ 46.3 + 3.0 * 0 \\ 46.3 + 3.0 * 3 \end{pmatrix},
$$

or, in matrix form,

$$
\hat{y} = \begin{pmatrix} 43.3 \\ 46.3 \\ 55.3 \end{pmatrix} = \begin{pmatrix} 1 & -1 \\ 1 & 0 \\ 1 & 3 \end{pmatrix} \begin{pmatrix} 46.3 \\ 3.0 \end{pmatrix},
$$

or, more abstractly,

$$
\hat{y} = X\hat{\beta}.
$$

Vectors and matrices (this and more: video 2a)

At the simplest level:

- *Vector:* list of numbers

$$\begin{pmatrix} 43.3 \\ 46.3 \\ 55.3 \end{pmatrix}$$

- *Matrix:* rectangular array of numbers

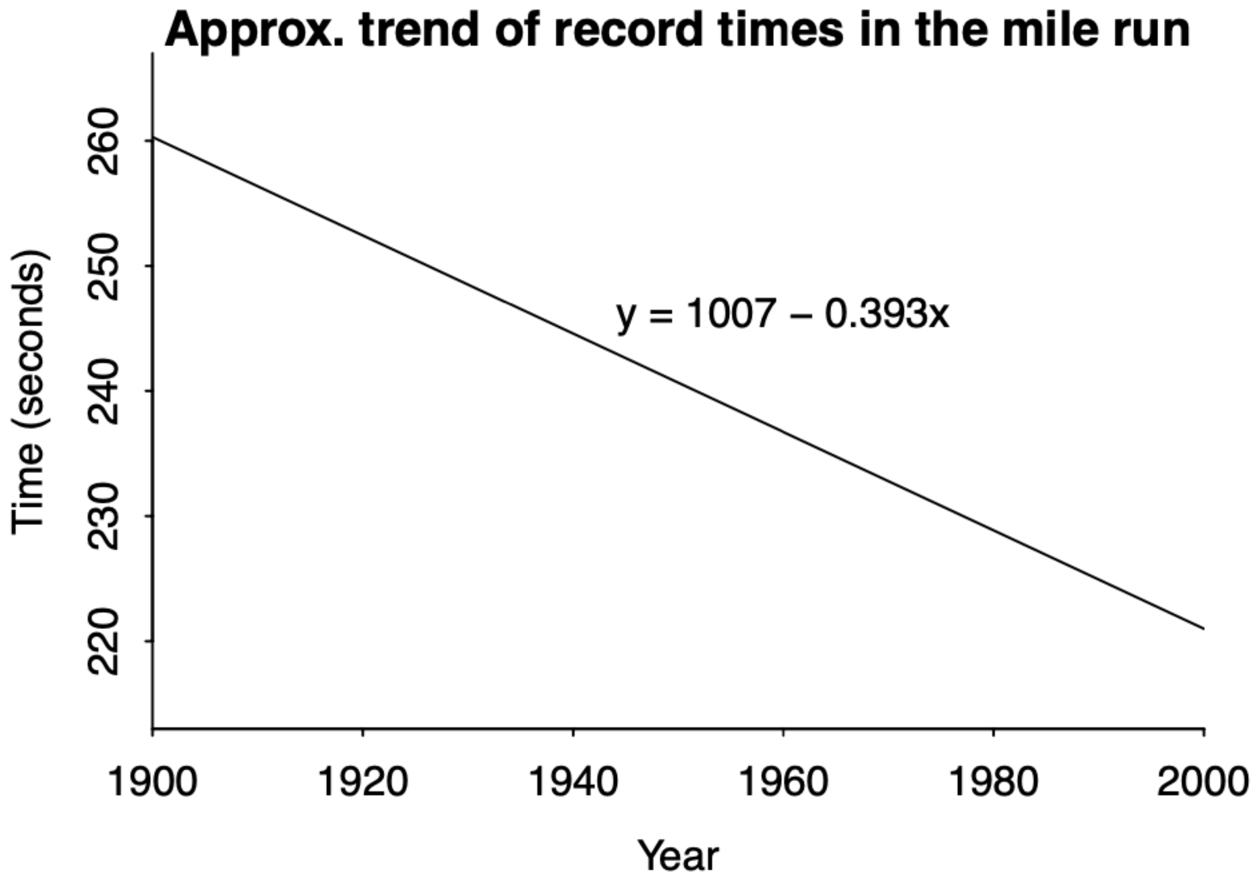
$$\begin{pmatrix} 1 & -1 \\ 1 & 0 \\ 1 & 3 \end{pmatrix}$$

Multiplying a vector by a matrix

$$\hat{y} = X\hat{\beta}.$$

$$\hat{y} = \begin{pmatrix} 43.3 \\ 46.3 \\ 55.3 \end{pmatrix} = \begin{pmatrix} 1 & -1 \\ 1 & 0 \\ 1 & 3 \end{pmatrix} \begin{pmatrix} 46.3 \\ 3.0 \end{pmatrix}$$

Graphing a line



- Invent your own example (2 points in a plane) and calculate the corresponding equation!

Exponential and power-law relations

The formula $\log y = a + bx$ represents

exponential growth (if $b > 0$)

or decline (if $b < 0$):

$y = Ae^{bx}$, where $A = e^a$.

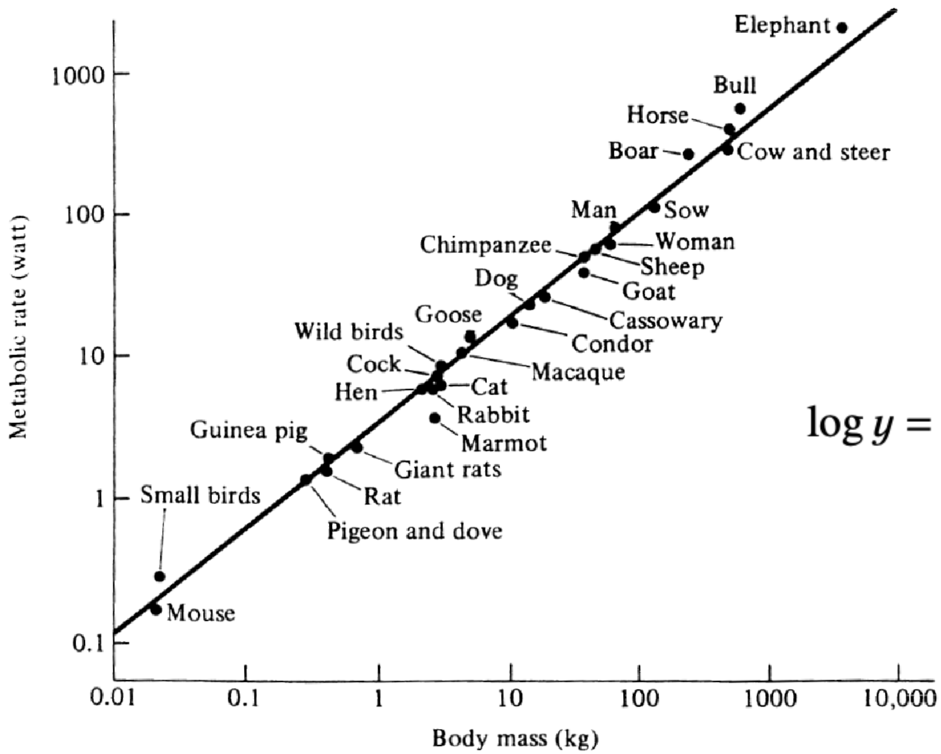
The formula $\log y = a + b \log x$ represents

power-law growth (if $b > 0$)

or decline (if $b < 0$):

$y = Ax^b$, where $A = e^a$.

Exponential and power-law relations



$$\log y = 1.4 + 0.74 \log x.$$

Figure 3.4 *Log metabolic rate vs. log body mass of animals, from Schmidt-Nielsen (1984). These data illustrate the log-log transformation. The fitted line has a slope of 0.74. See also Figure 3.5.*

- Invent your own examples for exponential and power law relations and calculate the corresponding equation!

Probability distributions

Probability distributions represent the unmodeled aspects of reality—the *error term* ϵ in the expression $y = a + bx + \epsilon$—and it is randomness that greases the wheels of inference.

A probability distribution corresponds to an urn with a potentially infinite number of balls inside. When a ball is drawn at random, the "random variable" is what is written on this ball.

A probability distribution of a random variable z takes on some range of values (the numbers written on the balls in the urn). The *mean* of this distribution is the average of all these numbers or, equivalently, the value that would be obtained on average from a random sample from the distribution. The mean is also called the expectation or expected value and is written as $\mathrm{E}(z)$ or μ_z.

The *variance* of the distribution of z is $\mathrm{E}((z - \mu_z)^2)$

The *standard deviation* is the square root of the variance.

Probability distributions

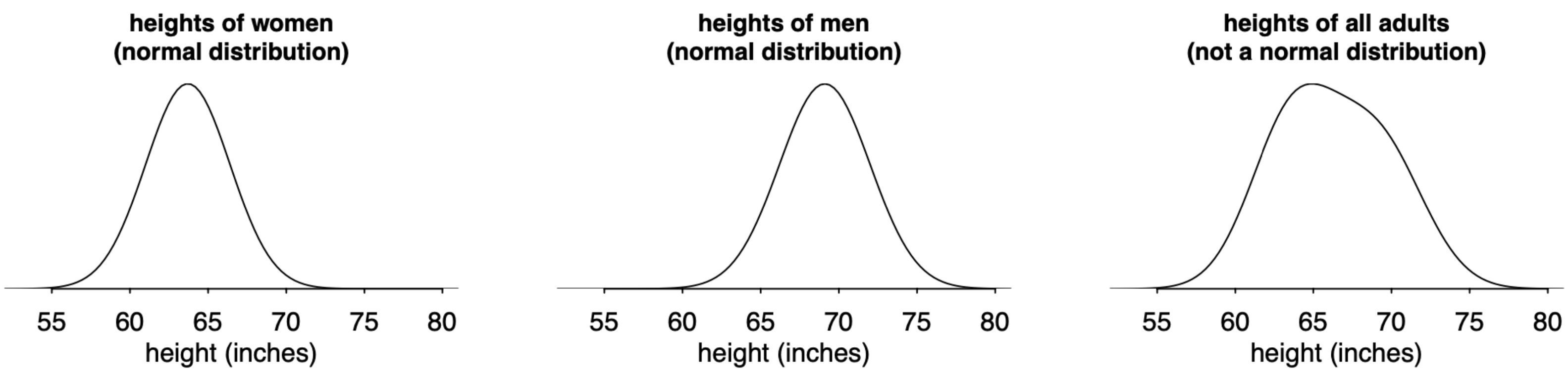


Figure 3.6 *(a) Heights of women, which approximately follow a normal distribution, as predicted from the Central Limit Theorem. The distribution has mean 63.7 and standard deviation 2.7, so about 68% of women have heights in the range* 63.7 ± 2.7. *(b) Heights of men, approximately following a normal distribution with mean 69.1 and standard deviation 2.9. (c) Heights of all adults in the United States, which have the form of a mixture of two normal distributions, one for each sex.*

The Gaussian ("normal") distribution

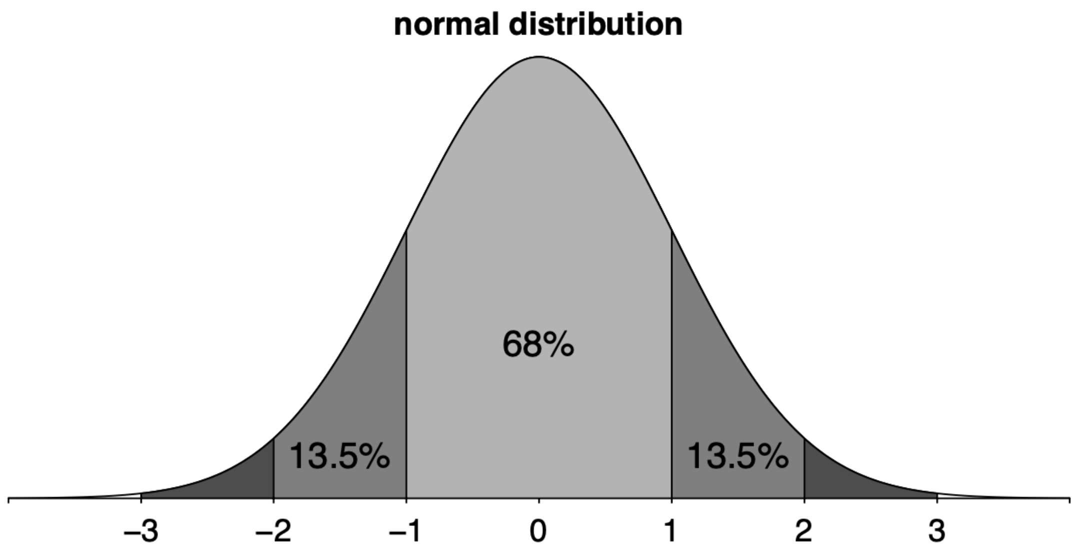


Figure 3.7 *Approximately 50% of the mass of the normal distribution falls within 0.67 standard deviations from the mean, 68% of the mass falls within 1 standard deviation from the mean, 95% within 2 standard deviations of the mean, and 99.7% within 3 standard deviations.*

Correlation

If two random variables u and v have mean μ_u, μ_v and standard deviations σ_u, σ_v, then their *correlation* is defined as $\rho_{uv} = \mathrm{E}((u - \mu_u)(v - \mu_v))/(\sigma_u \sigma_v)$. It can be shown mathematically that the correlation must be in the range $[-1, 1]$, attaining the extremes only when u and v are linear functions of each other.

Knowing the correlation gives information about linear combinations of u and v. Their sum $u + v$ has mean $\mu_u + \mu_v$ and standard deviation $\sqrt{\sigma_u^2 + \sigma_v^2 + 2\rho\sigma_u\sigma_v}$. More generally, the weighted sum $au + bv$ has mean $a\mu_u + b\mu_v$, and its standard deviation is $\sqrt{a^2\sigma_u^2 + b^2\sigma_v^2 + 2ab\rho\sigma_u\sigma_v}$. From this we can derive, for example, that $u - v$ has mean $\mu_u - \mu_v$ and standard deviation $\sqrt{\sigma_u^2 + \sigma_v^2 - 2\rho\sigma_u\sigma_v}$.

The log-Gaussian distribution

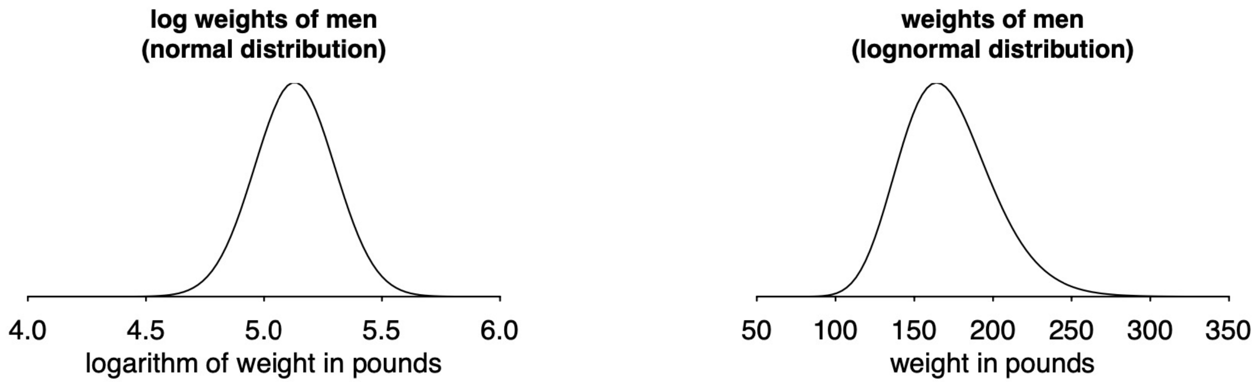


Figure 3.8 *Weights of men (which approximately follow a lognormal distribution, as predicted from the Central Limit Theorem from combining many small multiplicative factors), plotted on the logarithmic and original scales.*

Potential outcomes of a treatment

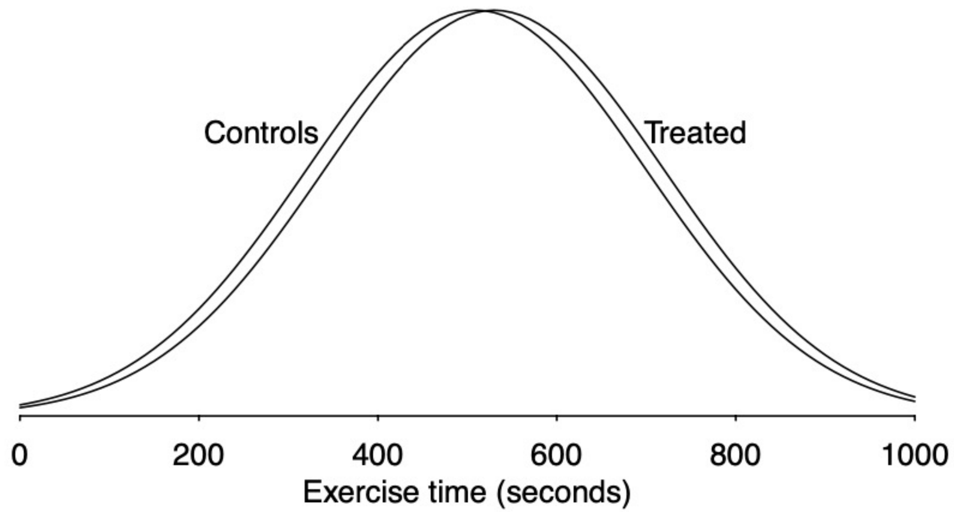


Figure 3.9 *Distributions of potential outcomes for patients given placebo or heart stents, using a normal approximation and assuming a treatment effect in which stents improve exercise time by 20 seconds, a shift which corresponds to taking a patient from the 50*th *to the 54*th *percentile of the distribution under the placebo.*