Faculty of Life Sciences

Course introduction

Ib Skovgaard and Claus Ekstrøm
E-mail: ims@life.ku.dk

Program

- Practical information
- Populations and samples
- Data
  - Data types
  - Visualization
  - Measures of location and variation of a distribution

Practical information

- Course material and information on Absalon
- Program
- Lectures
- Exercises
- Cases
- Material
  - The text book
  - Data sets (the `isdals` package)
  - Solution manual
  - R-guide(s)
- R

Weekly schedule (week 1-8)

- Monday 9.00-10.30 Lectures
  - 8.15-10.00 Week 1 only: Earlier lecture
- Monday 10.30-12.00 Exercises
- Tuesday 13-14 Lecture
- Tuesday 14-16 Exercises (1 hour supervised)
  - 16-17 Week 1 only: Further R-work, if needed
- Homework: week case
- Friday 8-9 Case summary
- Friday 9-10 Week summary and supplements

Introduction to statistics

Do you need statistics?

Birth sex bias of animal offspring

Are too many males born among animals in captivity?
Theories suggest male bias from strong mothers for certain species.

Two examples from births in European Zoo's.:

Kirk's dik-dik:
154 males and 96 females

Brown bear:
6 males and 12 females

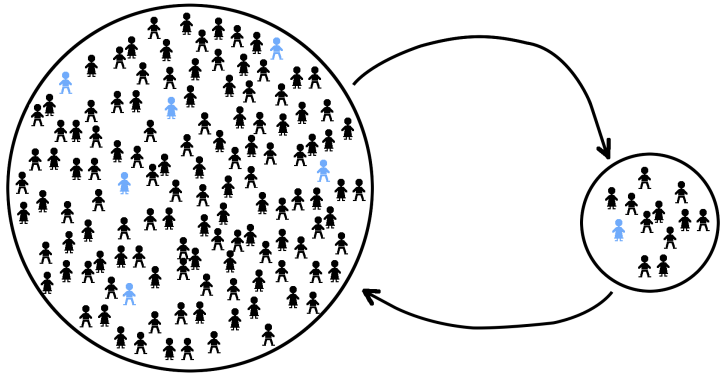
Birth sex bias II

Comparison of male birth proportion for Kirk's dik-dik in two studies:

	Males	Females
European zoo's	154	96
North American zoo's	169	134

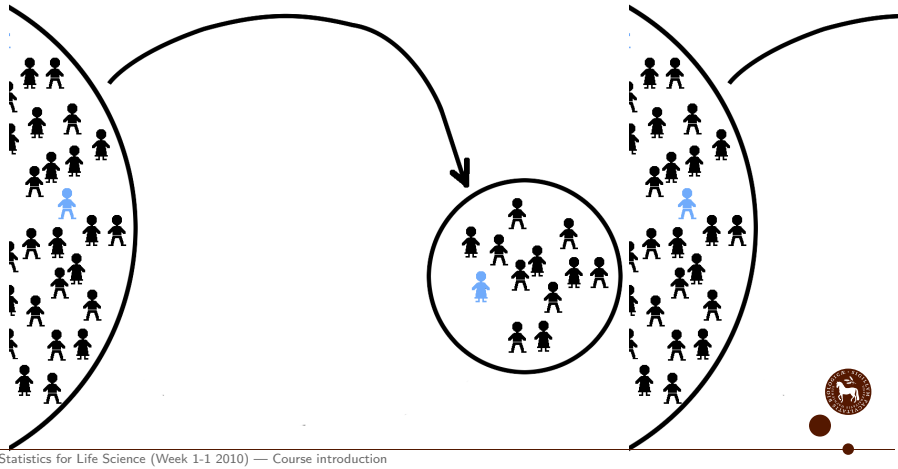
- Could the probability of male birth be the same in the two studies?
- How large is the difference, if any?

Population and sample



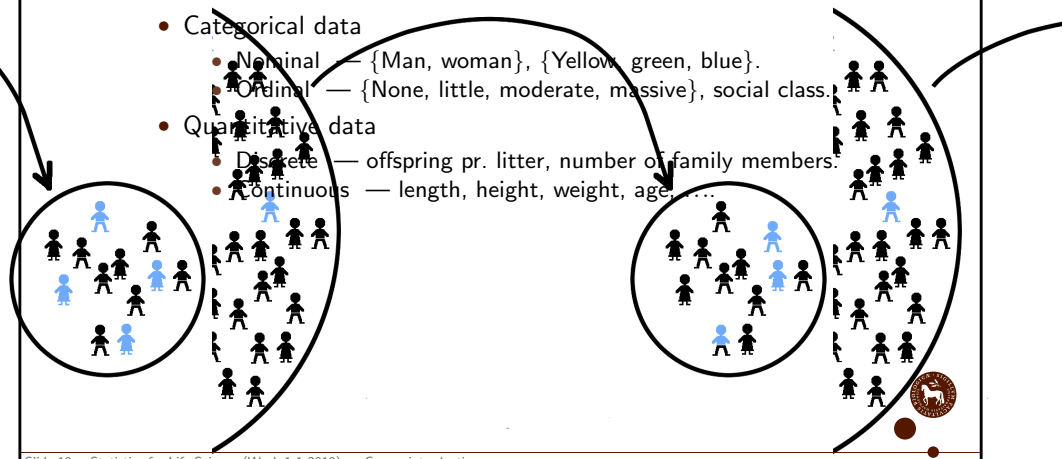
Samples

There is variation between individuals in the population. Samples therefore also vary.



Data types

- Categorical data
  - Nominal — {Man, woman}, {Yellow, green, blue}.
  - Ordinal — {None, little, moderate, massive}, social class.
- Quantitative data
  - Discrete — offspring pr. litter, number of family members.
  - Continuous — length, height, weight, age, ...



Data types — example

Lameness in cattle 72 hours after intake of sugar.

Place	Weight (kg)	Lameness score	nb of swollen joints
I	276	Mildly lame	2
I	395	Mildly lame	1
I	356	Normal	0
I	437	Lame	2
II	376	Lame	0
II	350	Moderately lame	0
II	331	Lame	1
II	331	Normal	0

Visualisation — categorical data

- Frequency = number of occurrences.
- Relative frequency = relative number of occurrences.

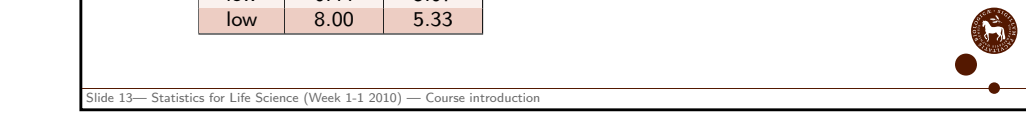
If n is the number of observations then relative frequency = $\frac{\text{frekvensen}}{n}$.

Visualization — quantitative data

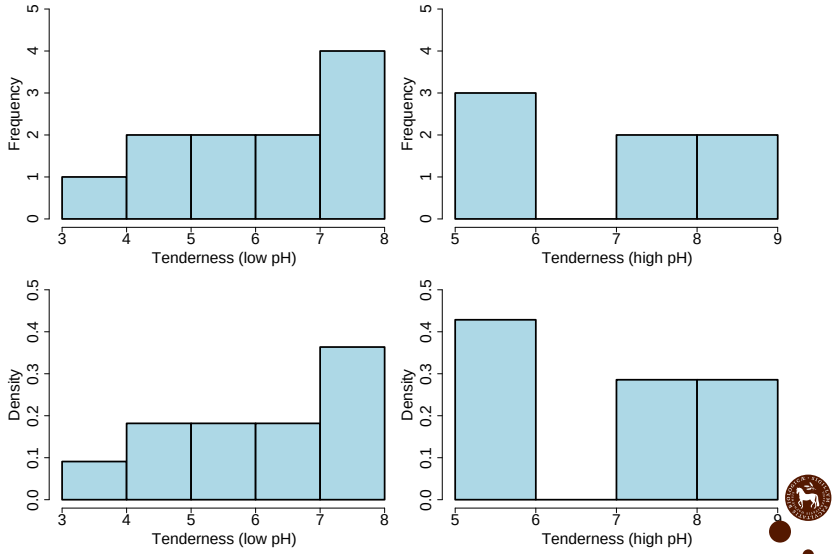
pH group	Tunnel cooling	Rapid cooling
high	8.44	8.44
high	7.11	6.00
high	6.00	5.78
high	7.56	7.67
low	7.22	5.56
high	5.11	4.56
low	3.11	3.33
high	8.67	8.00
low	7.44	7.00
low	4.33	4.89
low	6.78	6.56
low	5.56	5.67
low	7.33	6.33
low	4.22	5.67
high	5.78	7.67
low	5.78	5.56
low	6.44	5.67
low	8.00	5.33

Tenderness of pork meat by different cooling methods.

- How do we compare the results from the two methods?



Visualization — quantitative data



Median, range and quartiles

The median is the "middle observation" among the ranked data. If the number of observations is even, take the mean of the two middle observations:

The range is defined the largest minus den smallest observation:

The quartiles (denoted Q_1, Q_2, Q_3) split the data into four parts such that the smallest 25%, 50%, 75% of the observations are below the 1st, 2nd, and 3rd quartile, respectively.

The inter-quartile range (IQR) is the range of the middle 50% of the observations, that is,

$$\text{IQR} = Q_3 - Q_1$$

Boxplots

A boxplot illustrates a distribution by plotting the 5 measures: minimum, Q_1, median, Q_3 and maximum.

In a modified boxplot minimum and maximum is replaced by the smallest and the largest observed value contained in the interval

$$[Q_1 - 1.5 \cdot \text{IQR}; Q_3 + 1.5 \cdot \text{IQR}].$$

Observations outside this interval are plotted as points.

Mean and standard deviation

The sample standard deviation (sample SD) is defined as:

$$s = \sqrt{\frac{\sum_{i=1}^{n}(y_i - \bar{y})^2}{n-1}}.$$

The sample variance is the square of the sample standard deviation.

There is also a population standard deviation (SD) defined similarly, but for the entire (possibly infinite) population.

Note that the standard deviation as well as the mean are measured in the units of the observations.

Infobox 1.1

Linear transformation of mean and standard deviation

Let $\bar{y}$ and s_y be the sample mean and sample standard deviation from observations $y_1, \ldots, y_n$ and let $z_i = c \cdot y_i + b$ be a linear transformation of the y's with constants b and c. Then $\bar{z} = c \cdot \bar{y} + b$ and $s_z = |c| \cdot s_y$.

This means that if we add a constant to the data, or multiply our data by a factor, then the mean and SD change in a natural way.

Median or mean?

Median

- The median splits the data in two equally large parts.
- Can also be used for ordinal data.
- Not so sensitive to outliers.

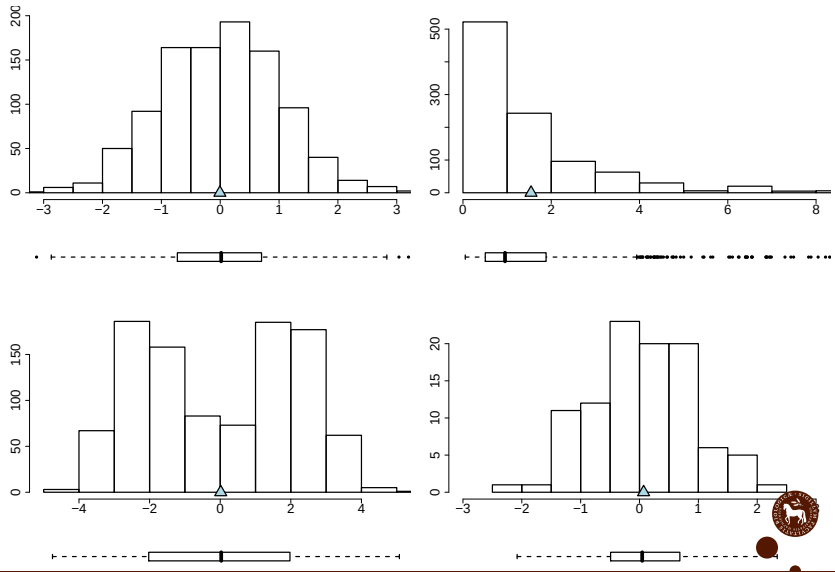
Mean

- The mean uses the values of the data, not just their order.
- Not well suited for highly skewed (non-symmetrical) distributions.
- Sensitive to outliers.

Standard deviation

- The standard deviation (SD) is a typical distance from the mean.
- More precisely SD^2 equals half of the mean squared difference between any two observations in the sample.

Four examples



Summary of main points

- Populations / samples and inference.
- Data types.
- Visualisation.
- Median, mean, standard deviation and IQR. Their definition and their interpretation.