LDI Seminar Series in Biostatistics: Lecture 1

Kevin McGregor

September 18, 2017

About me

- 3rd year PhD candidate in biostatistics in the department of Epidemiology, Biostatistics, and Occupational Health at McGill.
- Supervised by Drs. Celia Greenwood (LDI Centre for Clinical Epidemiology) and Aurélie Labbe (HEC Montréal).
- Specialize in statistical methodology for genetic and genomic data (microbiome, DNA methylation, RNA-seq)
  - Involved in genomic studies in autoimmune disease, asthma, colorectal cancer, and anorexia nervosa.

Intro to seminar series

- Seminar series to give insight into proper implementation of statistics in biomedical research.
- Meant to be introductory. Today's lecture will be quite basic for someone who has taken statistics before.
- Will focus on biomedical applications... strong emphasis on proper implementation and interpretation. We won't go into the math too much.

Software

- Will give examples of doing analysis in SPSS (available to all McGill students)
  - Open-source version called PSPP (https://www.gnu.org/software/pspp/) available to those who do not have access to SPSS.
- Analyses covered will all likely be available in any statistical software.
- Dataset and instructions for SPSS tutorial will be available online for practice.

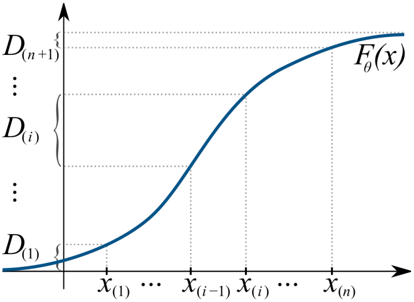
Schedule

- Sept 18: Intro to (bio)statistics
- Sept 25: Hypothesis testing (illustrated through t-tests)
- Oct 2: Simple and Multiple Linear Regression
- Oct 9: Thanksgiving (No lecture this week)
- Oct 16: One-way and Factorial ANOVA
- Oct 23: Putting it all Together

Intro to Biostatistics

Why do we need (bio)statistics?

- We often have questions that we want to answer through quantitative means.
- These days there is often a plethora of data available. Need to effectively summarize data to understand what the data show.
- Need a way to report results in a responsible manner.



https://commons.wikimedia.org/wiki/File:Spacings.svg

Population parameters

- Population parameter: quantity that describes the distribution of some variable of interest in a population.
- We don't know the true values of population parameters! But we can use statistics to try to get good estimates.
- Population parameters often represented as Greek letters, e.g.
  - μ usually represents the average (arithmetic mean) of a variable of interest in the population.
  - σ usually represents the standard deviation of a variable of interest in the population.

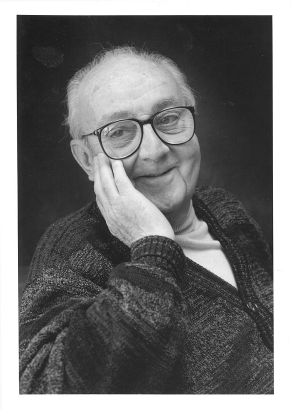
Sample statistics

- In experiments and observational studies, we can usually only take measurements from a subset of the population. This subset is called a *sample*, and all the observed variables (qualitative or quantitative) in this sample form the *data*.
- Conclusions drawn about the sample can only be generalized to the entire population if the sample adequately resembles the population.
- Quantities derived from the data are referred to as *statistics*, e.g.
  - The average (arithmetic mean) of a variable over all individuals in the sample
  - The standard deviation of a variable over all individuals in the sample
  - The maximum/minimum values of a variable in the sample

Population vs. sample

- A statistic is used to get an estimate of the true value of a population parameter.
  - The proportion of a random sample of Montrealers who plan on voting for Candidate A for mayor in the next election vs. proportion who actually vote for the candidate.
  - Slope of best fit line between age and blood pressure in cross sectional sample vs. average increase in blood pressure per additional year of age.
- The estimated value obtained from a statistic is almost certainly wrong!
  - But we don't care... we just need it to be close enough to the true parameter to be informative.

All models are wrong!



George Box

All models are wrong but some are useful!
-George Box 1978

What should we be concerned about?

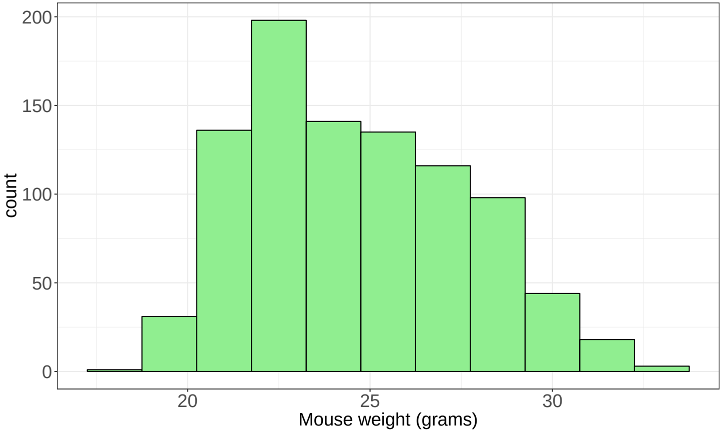
- Though we can accept error in our estimates, we don't want to systematically over- or under-estimate things (bias).
- Statistical testing often relies on unverifiable assumptions!
  - Sometimes have diagnostic tools
  - Can use external knowledge
  - Oftentimes assumptions are simply ignored!
- Have to make sure that the test we're performing actually answers the question of interest.

Investigating a single variable

Frequency distribution

- Categorical data (discrete data)... can often just use a table to summarize counts or percentages of each category
  - E.g. Sex, disease case vs. control, race, neighbourhood
  - Ordinal variables (categorical with natural ordering): self-reported health (poor, good, excellent), level of education (high school, university, grad school)
- Continuous data
  - Age, height, weight, cell count per unit volume of blood
  - Distribution of data can be shown in a histogram

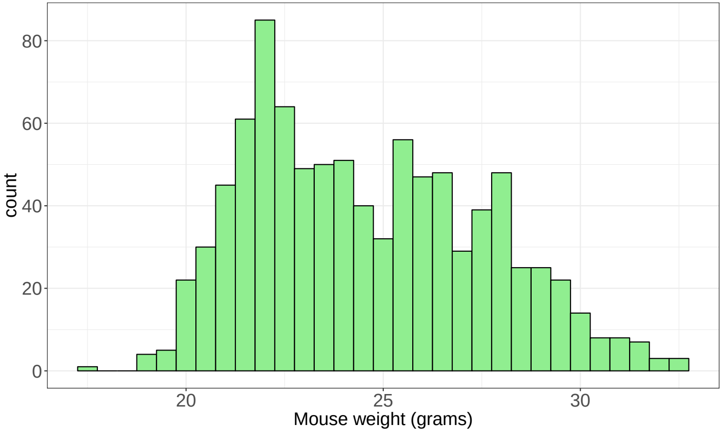
Histogram



Distribution of mouse body weight

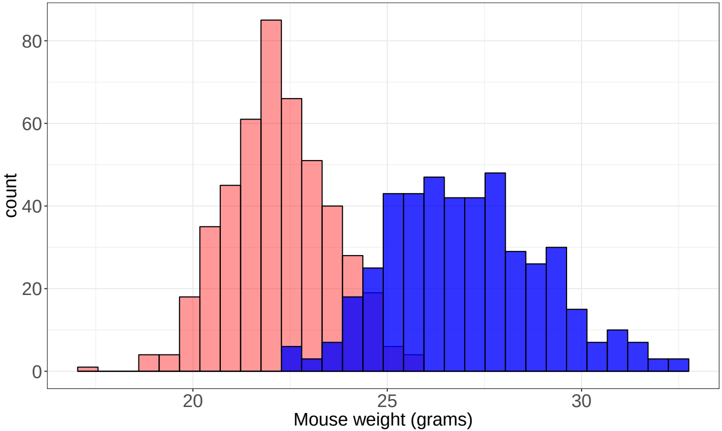
International Mouse Phenotyping Consortium data (http://www.mousephenotype.org/)

Histogram (finer bin width)



Finer bin width gives more detail

Histogram (subgroups)



What two subgroups are identified here?

Descriptive statistics

Distributions can be summarized through descriptive statistics. Most basic descriptive statistics fall into one of these four categories:

- Frequency
  - Counts or proportions of different categorical variables
  - How many observations fall into the different possible categories.
- Central Tendency
  - Mean, median, mode
  - Give an idea of the “centre” of the data, or a “typical” value.
- Position
  - Maximum, minimum, percentile
  - Give an idea of where the data fall in relation to one another.
- Variation or Dispersion
  - Variance, standard deviation
  - Give an idea of how spread out the data are.

Descriptive statistics

Distributions can be summarized through descriptive statistics. Most basic descriptive statistics fall into one of these four categories:

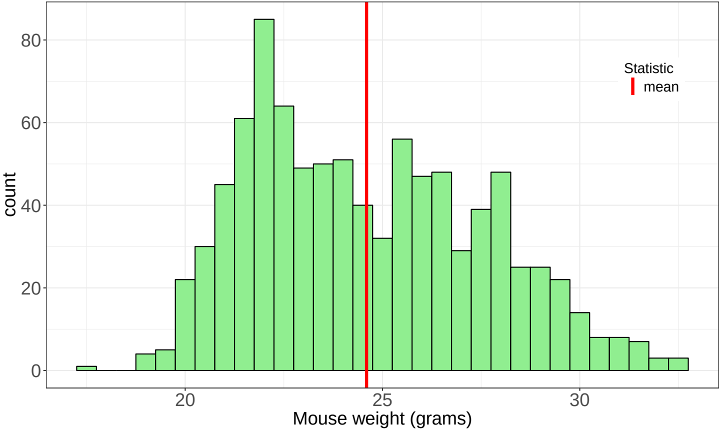
- Frequency
  - Counts or proportions of different categorical variables
  - How many observations fall into the different possible categories.
- Central Tendency
  - Mean, median, mode
  - Give an idea of the "centre" of the data, or a "typical" value.
- Position
  - Maximum, minimum, percentile
  - Give an idea of where the data fall in relation to one another.
- Variation or Dispersion
  - Variance, standard deviation
  - Give an idea of how spread out the data are.

Arithmetic mean

- By far the most common measure of central tendency is the *arithmetic mean* (aka just the mean, or average).
- Assume a sample size of n. Let $x = (x_1, x_2, \ldots, x_n)$ represent the observed values of each of the n individuals.
- The sample mean, represented as $\overline{x}$, is defined as:

$$\begin{aligned} \overline{x} &= \frac{x_1 + x_2 + \cdots + x_n}{n} \\ &= \frac{1}{n}\sum_{i=1}^{n} x_i \end{aligned}$$

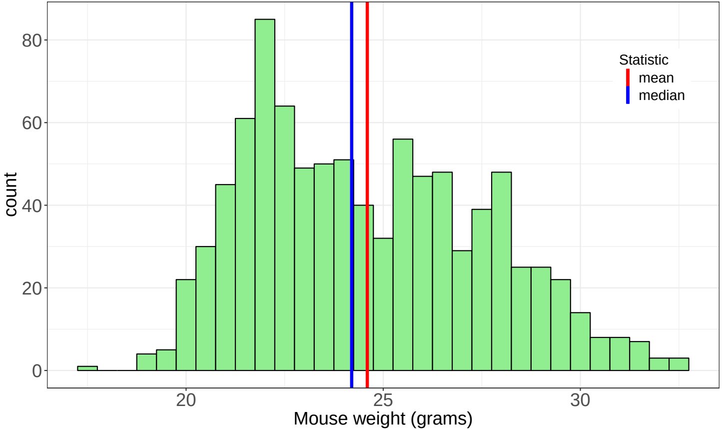
Mouse weight mean



Median

- The *median* is a value that separates the sample into an upper and lower half.
- If there is an odd number of observations, just order the observations and take the middle value. i.e. if the data is:
  - 1, 23, 34, 56, 57, 71, 86
  - The median is 56
- If there is an even number of observations, then order the observations and take the average of the two middle observations. E.g.
  - 1, 23, 34, 56, 57, 71, 86, 92
  - The median is $\frac{56+57}{2} = 56.5$

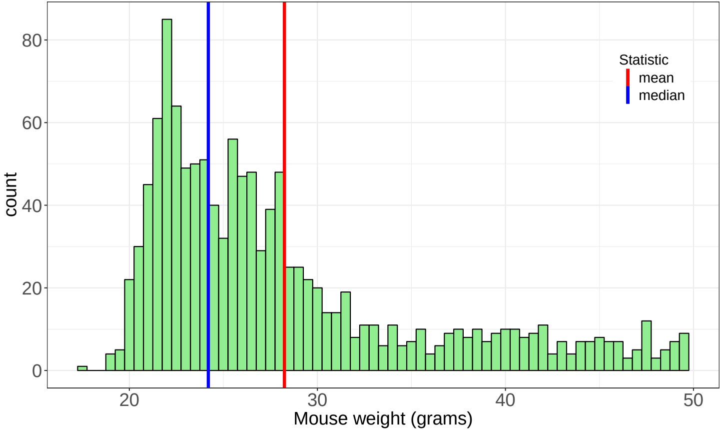
Mouse weight median



Median and outliers

- The median is less sensitive to outliers (uncharacteristically high or low values) than the mean!
- Data: 1, 23, 34, 56, 57, 71, 86
  - Mean $= \overline{x} = \frac{1+23+34+56+57+71+86}{7} = 46.86$
  - Median $= 56$
- Data with outliers: 1, 23, 34, 56, 57, **109**, **172**
  - Mean $= \overline{x} = \frac{1+23+34+56+57+109+172}{7} = 64.57$
  - Median is still 56

Mouse weight skewed



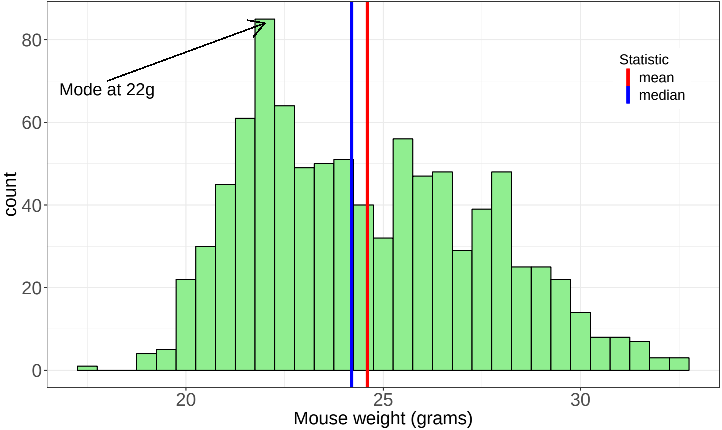
Mode

- A somewhat less common measure of centrality is called the *mode*. This is the most common value in the data.
- Data: 2, 2, 13, 20, 20, 20, 20, 25, 25, 30, 30

Value	Frequency
2	2
13	1
20	4
25	2
30	2

- Mode $= 20$, since it appears more than any other value.
- Becomes a little bit trickier when dealing with continuous data as there are often no repeated values. Can infer mode from the maximum point on the histogram.

Mouse weight mode

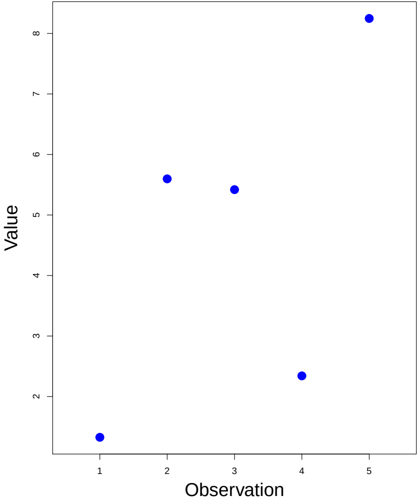


Measures of variation

- Measures of central tendency are only part of the story
- A distribution with mean at 25 and with all observations between 15 and 40 is very different from a distribution with mean at 25 and all observations between 0 and 100.
- Also need measures of how spread out the observations are
    - Will be critical when performing statistical testing procedures

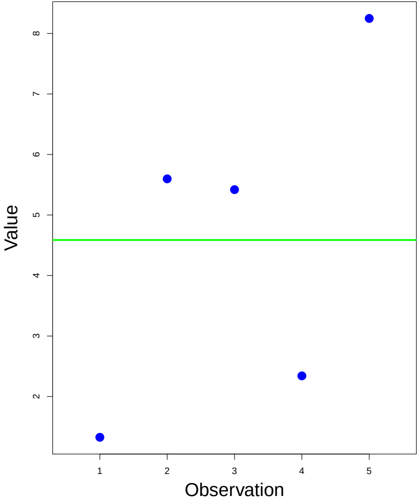
Simple dataset

Simple dataset with 5 observations: 1.33, 5.60, 5.42, 2.34, and 8.25



Points with mean

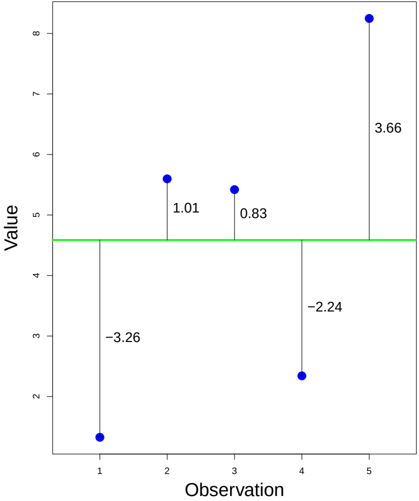
Can add a
line to represent the mean:
$\overline{x} = \frac{1.33+5.60+5.42+2.34+8.25}{5} = 4.588$



Points with residuals

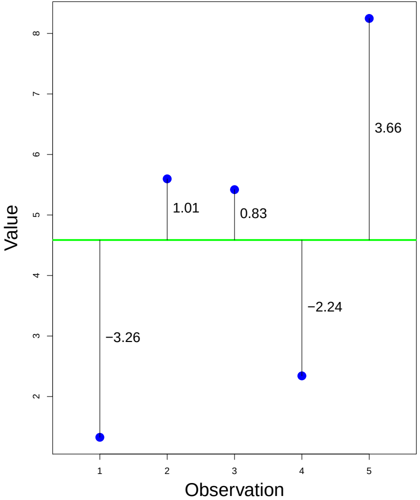
- Can calculate the difference between the mean and each of the observed points (called *residuals*)
- Very simple statistical model:

  $outcome = mean + error$



Quantifying variation

- How to measure the amount of variation?
- Taking the mean of the residuals gives: $\frac{-3.26+1.01+0.83-2.24+3.66}{5} = 0$
- In fact, the mean of the residuals will always be *exactly* zero!

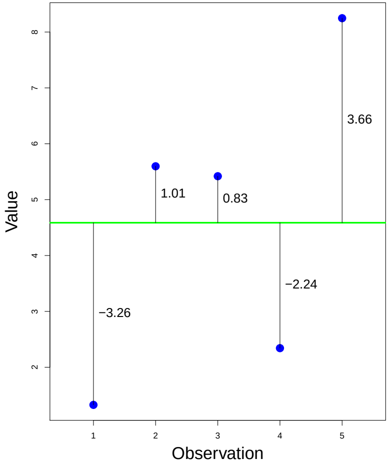


Quantifying variation

- Could ignore whether residuals are positive/negative
- Taking the mean of the absolute residuals gives:

$$\frac{|-3.26| + |1.01| + |0.83| + |-2.24| + |3.66|}{5} = \frac{3.26 + 1.01 + 0.83 + 2.24 + 3.66}{5} = 2.2$$

- Intuitive, but this turns out to be mathematically inconvenient.

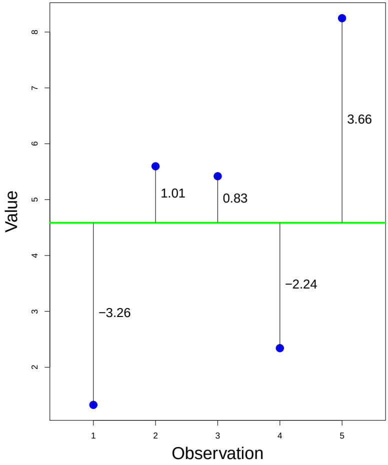


Quantifying variation

- Could try to square the residuals, then take the mean:

$$\frac{(-3.26)^2 + (1.01)^2 + (0.83)^2 + (-2.24)^2 + (3.66)^2}{5} = \frac{10.63 + 1.02 + 0.69 + 5.02 + 14.40}{5} = 6.15$$

- Less intuitive, but is very convenient to work with mathematically.



Sample variance

This leads to a statistic called the *sample variance*, denoted by s^2, which measures the average squared distance between each point and the mean:

$$s^2 = \frac{(x_1 - \overline{x})^2 + \cdots + (x_n - \overline{x})^2}{n-1}$$
$$= \frac{1}{n-1} \sum_{i=1}^{n} (x_i - \overline{x})^2.$$

Note that the denominator is $n-1$, not n. (Though for large n it would not make much of a difference)

Sample variance

- Variance is a very important quantity in statistics.
- A small value of sample variance means the observations are very concentrated around the mean and a large value means they're very spread out.
- But we squared the residuals so it's hard to relate the magnitude of the variance to the original observations.

Standard deviation

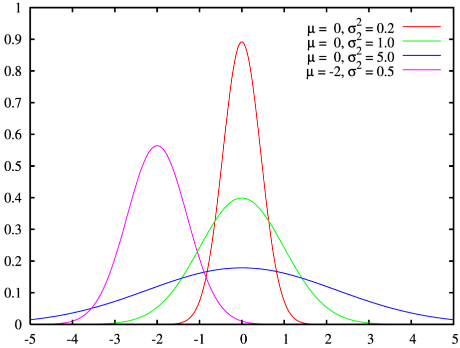
- Though variance is used often in calculations, it is much more common to report its square root which is called the *standard deviation*, denoted by $s = \sqrt{s^2}$.

$$
\begin{aligned}
s &= \sqrt{\frac{(x_1 - \overline{x})^2 + \cdots + (x_n - \overline{x})^2}{n-1}} \\
&= \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(x_i - \overline{x})^2}
\end{aligned}
$$

- The standard deviation can be interpreted in terms of the units of measurement of the data $x_1, \ldots, x_n$.

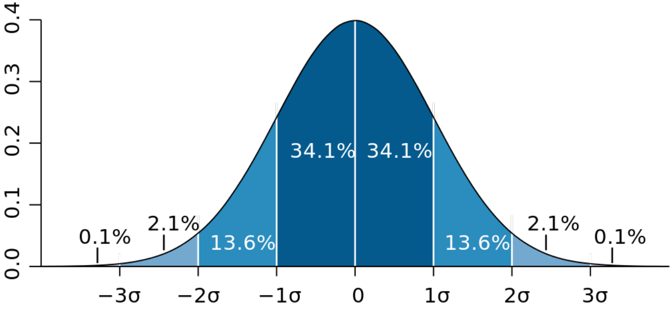
Normal distribution

- The standard deviation (or variance) appears in the most famous statistical distribution: the *normal distribution*.
- Each normal distribution is characterized by a mean μ and a standard deviation σ (variance σ^2). **These are population quantities**.



Normal distribution

- The mean, median, and mode are all equal to μ.
- Over 68% of the data fall within one standard deviation of the mean.
- About 95% of the data fall within two standard deviations of the mean.



Comparing two variables

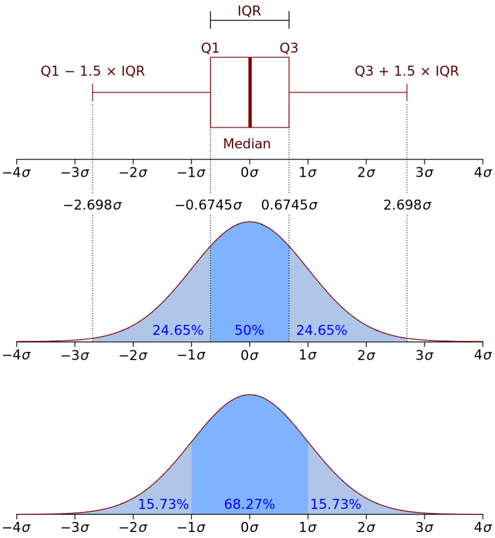
Comparing two variables

- Everything thus far dealt with only a single variable.
- Most research questions are concerned with comparing how the distribution of one variable is affected by another variable.
- A lot of statistical tests exist to study the relationships between variables (will get into some of these later on).
- For now, will focus on graphical comparisons.

Boxplot

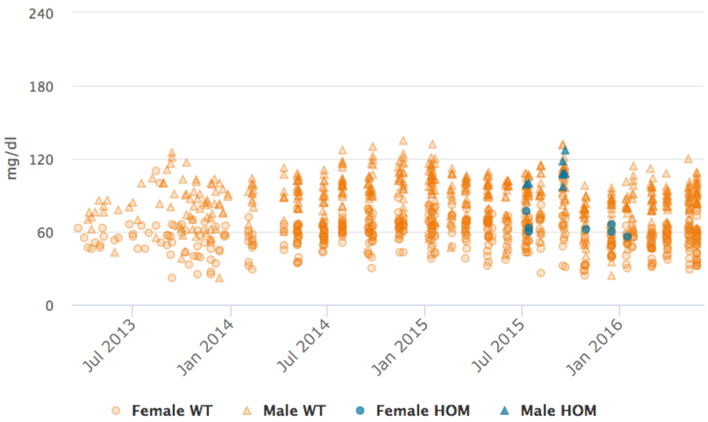
- *Box-and-whisker plot*: A simple type of plot to compare the distributions of a continuous variable with respect to the levels of one or more categorical variable. (Also known as just a *boxplot*)
- The different elements of the boxplot show how the data points are spread out and whether there are any outliers.

Boxplot



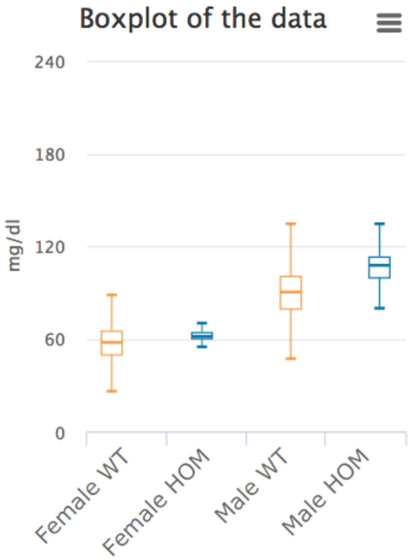
https://commons.wikimedia.org/wiki/User:Chen-Pan_Liao

HDL-cholesterol (IMPC)



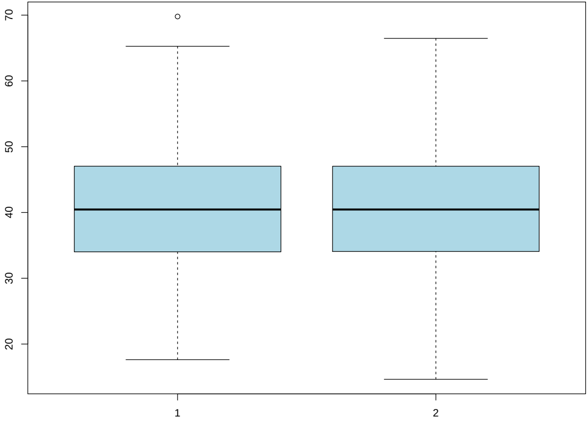
HDL-cholesterol of mice with: IL13 wild type vs. IL13 knocked out

HDL-cholesterol boxplot



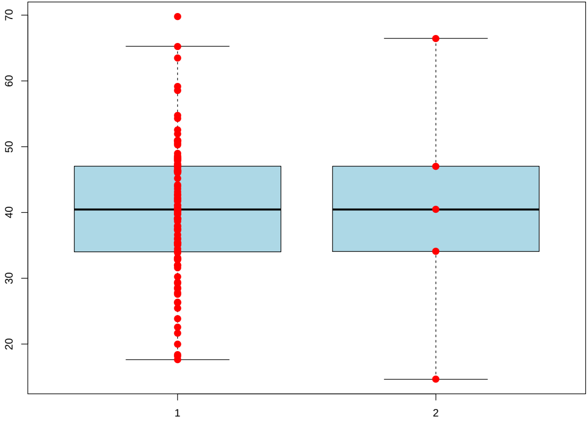
- Boxplot makes the comparison between wild type and knockout mice much easier.
- Can get a sense of variation of HDL-cholesterol within groups.

Similar boxplots



Two seemingly identical boxplots

Similar boxplots

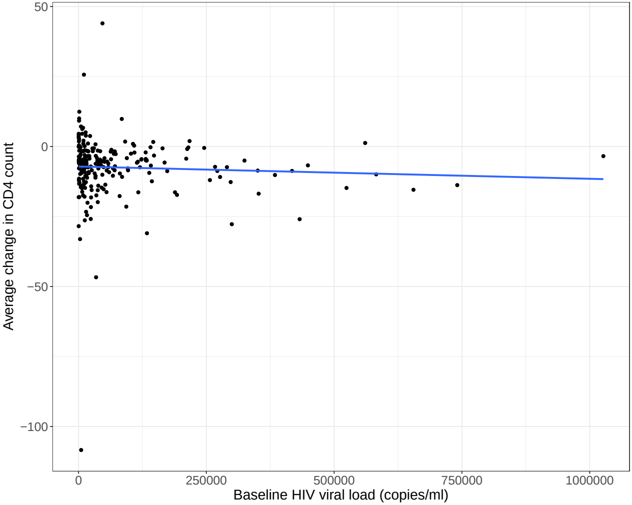


Superimposing the actual data points shows a different story!

Scatterplot

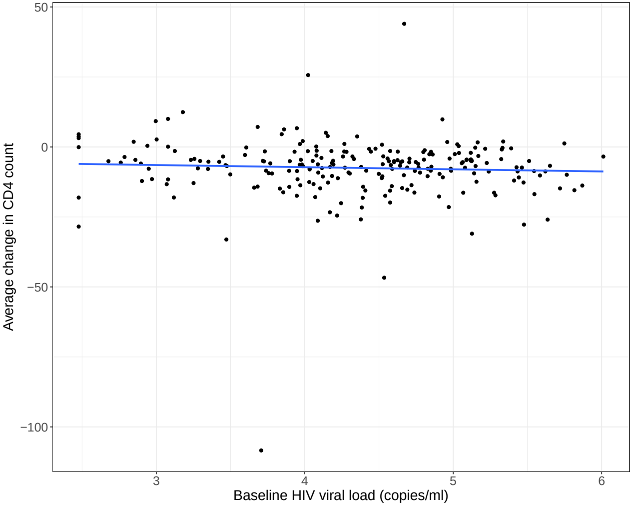
- A *scatterplot* can be used to study the relationship between two continuous variables.
- Very commonly used to illustrate linear regression (finding a "best fit" line for the data).

Scatterplot



Multicenter AIDS Cohort Data

Scatterplot (log-transformed)



$\log_{10}$-transformed viral load makes it easier to look for a trend

Take-home message

- Remember that all quantities calculated are subject to error, and potentially bias.
- Think about the population being studied, and ask yourself whether your sample is representative of that population.
- Always do a thorough data exploration (plots, descriptive statistics, etc.) before doing and formal statistical testing. You never know what you might find!
- Know how to properly interpret things.

Thank you! - Merci!

Questions?