science and policy
for a healthy future

HBM4EU project

Eye-balling and descriptive statistics

Lisbeth E. Knudsen

1st HBM4EU Training School 2018

1. Overview

Introduce key terms in statistics ppt from Google

www.uky.edu/CommInfoStudies/JAT/.../Presentations/Descriptive_statistics.pptx

2. Strategy

Lecture with some interaction

REASONS for STATISTICS

aids in summarizing the results

helps us recognize underlying trends and tendencies in the data

aids in communicating the results to others

TYPES OF STATISTICS

Descriptive (which *summarize some characteristic* of a sample)

Measures of central tendency
Measures of dispersion
Measures of skewness

Inferential (which test for significant *differences* between groups and/or significant *relationships* among variables within the sample

t-ratio, chi-square, beta-value

Descriptive statistics

Descriptive statistics are used to summarize data from individual respondents, etc.

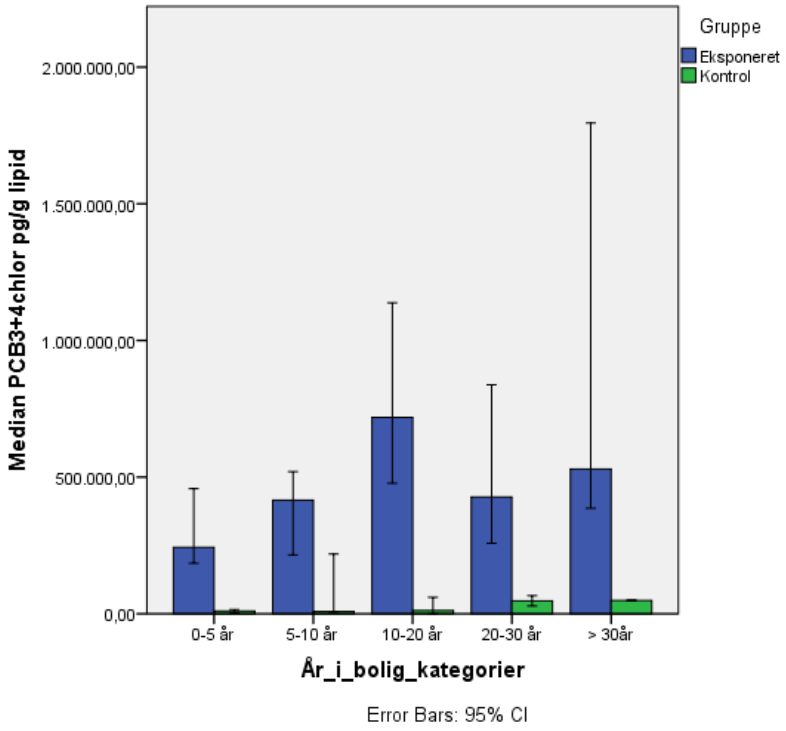
- **They help to make sense of large numbers of individual responses, to communicate the essence of those responses to others**

They focus on typical or average scores, the dispersion of scores over the available responses, and the shape of the response curve

Descriptive statistics

Example

Concentrations of lower chlorinated PCBs measured in blood, corrected for lipids and plotted versus years in dwelling



Normal distribution

Many characteristics are distributed through the population in a 'normal' manner

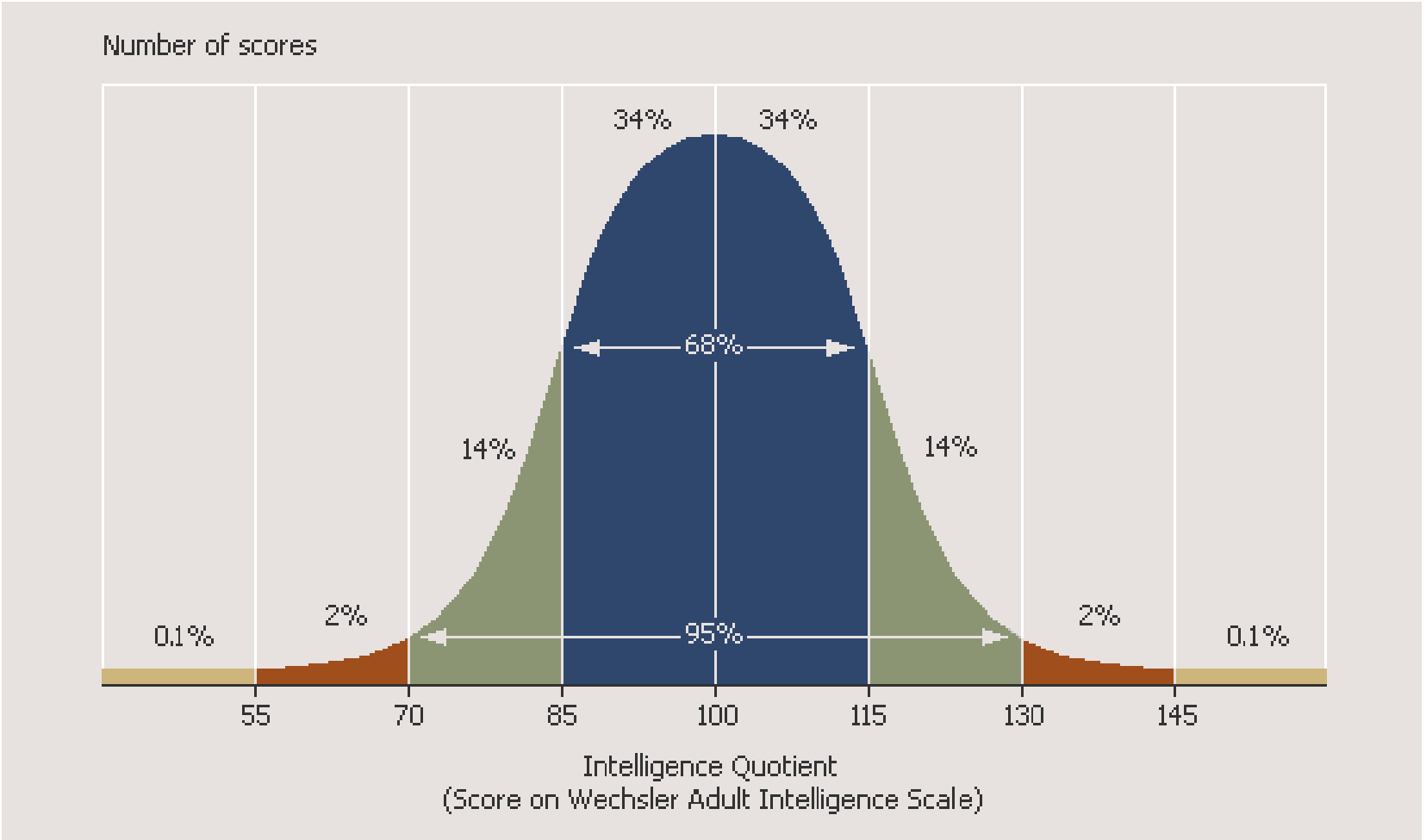
- Normal curves have well-defined statistical properties
- Parametric statistics are based on the assumption that the variables are distributed normally
  - Most commonly used statistics

This is the famous "Bell curve" where many cases fall near the middle of the distribution and few fall very high or very low

- I.Q.

Normal distribution

IQ distribution



Normal distribution

Mode (Mo): the most frequent score in a distribution

- good for nominal data

Median (Mdn): the midpoint or midscore in a distribution.

- (50% cases above/50% cases below)

– insensitive to extreme cases

--Interval or ratio

Mean

- The 'average' score—sum of all individual scores divided by the number of scores
- has a number of useful statistical properties
  - however, can be sensitive to extreme scores ("outliers")
- many statistics are based on the mean

Source : *Reasoning with Statistics*, by Frederick Williams & Peter Monge, fifth edition, Harcourt College Publishers.

Normal distribution

Some statistics look at how widely scattered over the scale the individual scores are

Groups with identical means can be more or less widely dispersed

To find out how the group is distributed, we need to know how far from or close to the mean individual scores are

Like the mean, these statistics are only meaningful for interval or ratio-level measures

Range

Distance between the highest and lowest scores in a distribution;

- sensitive to extreme scores;
- Can compensate by calculating interquartile range (distance between the 25th and 75th percentile points) which represents the range of scores for the middle half of a distribution

Usually used in combination with other measures of dispersion.

Normal distribution

Variance (S^2)

- Average of squared distances of individual points from the mean
  - sample variance
- High variance means that most scores are far away from the mean. Low variance indicates that most scores cluster tightly about the mean.
- The amount that one score differs from the mean is called its deviation score (deviate)
- The sum of all deviation scores in a sample is called the *sum of squares*

Standard Deviation (SD)

- A summary statistic of how much scores vary from the mean Square root of the Variance
- expressed in the original units of measurement
- Represents the average amount of dispersion in a sample
- Used in a number of inferential statistics

Normal distribution

Measures look at how lopsided distributions are—how far from the ideal of the normal curve they are

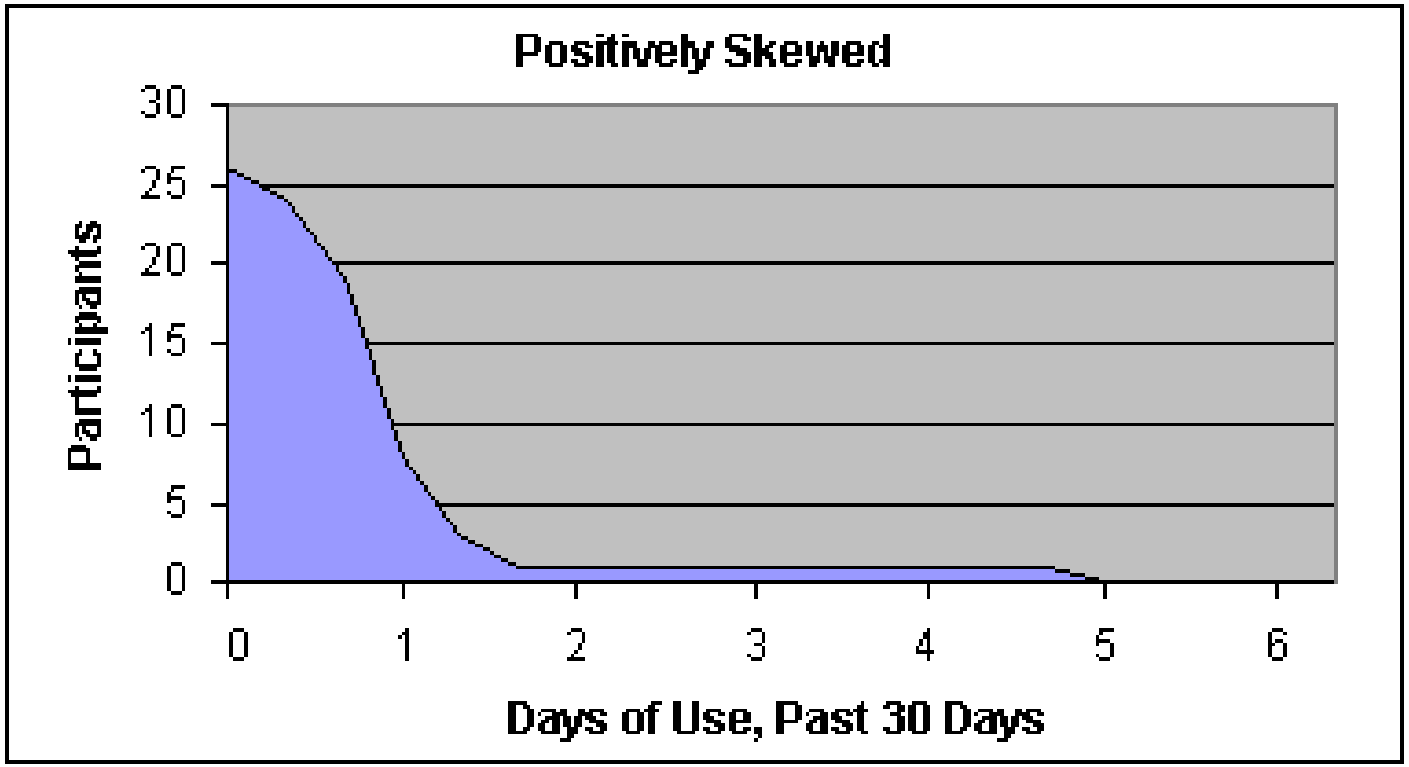
When the median and the mean are different, the distribution is skewed. The greater the difference, the greater the skew.

Distributions that trail away to the left are negatively skewed and those that trail away to the right are positively skewed

If the skewness is extreme, the researcher should either transform the data to make them better resemble a normal curve or else use a different set of statistics—nonparametric statistics—to carry out the analysis

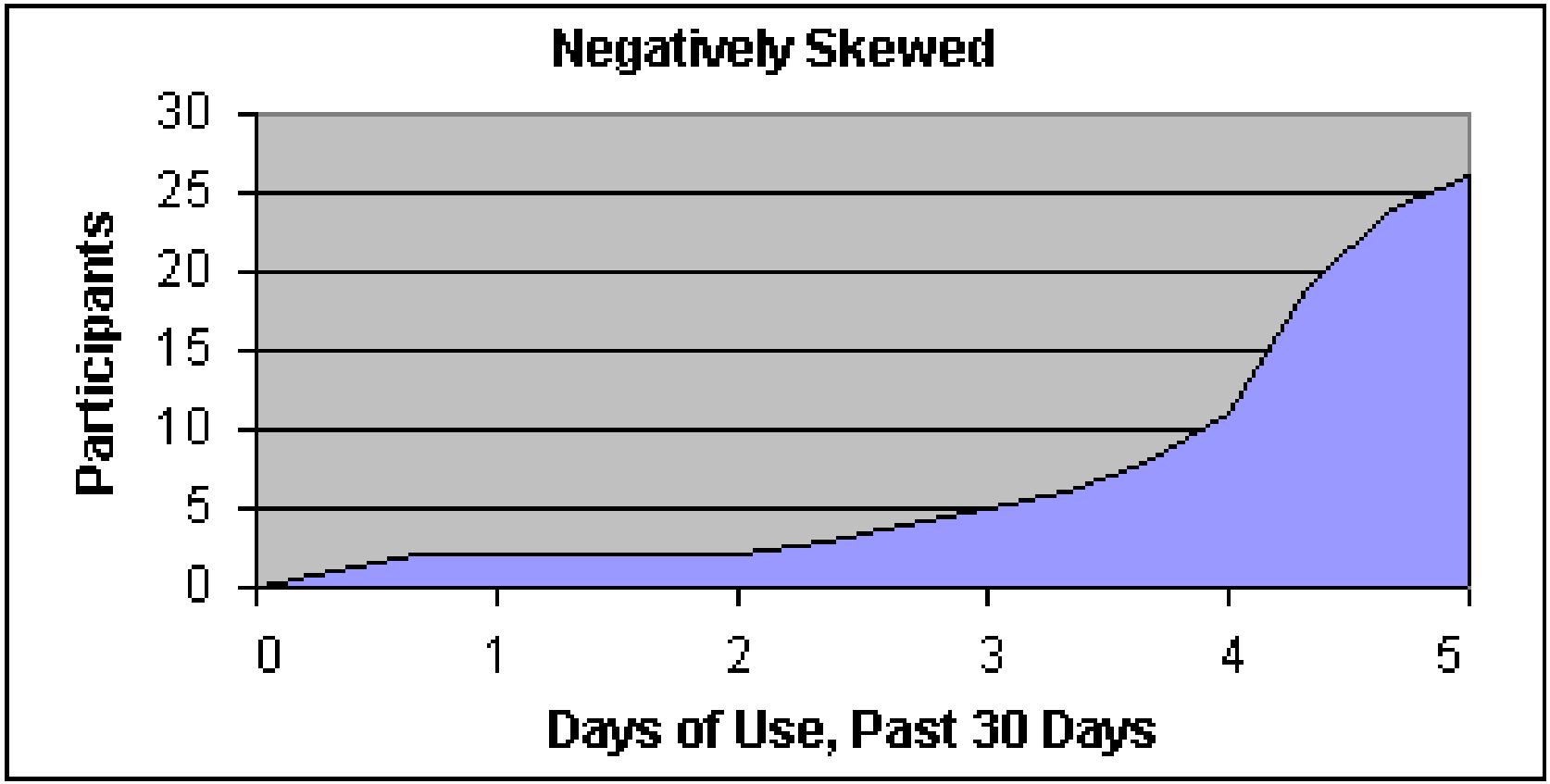
Normal distribution

Positive skewness

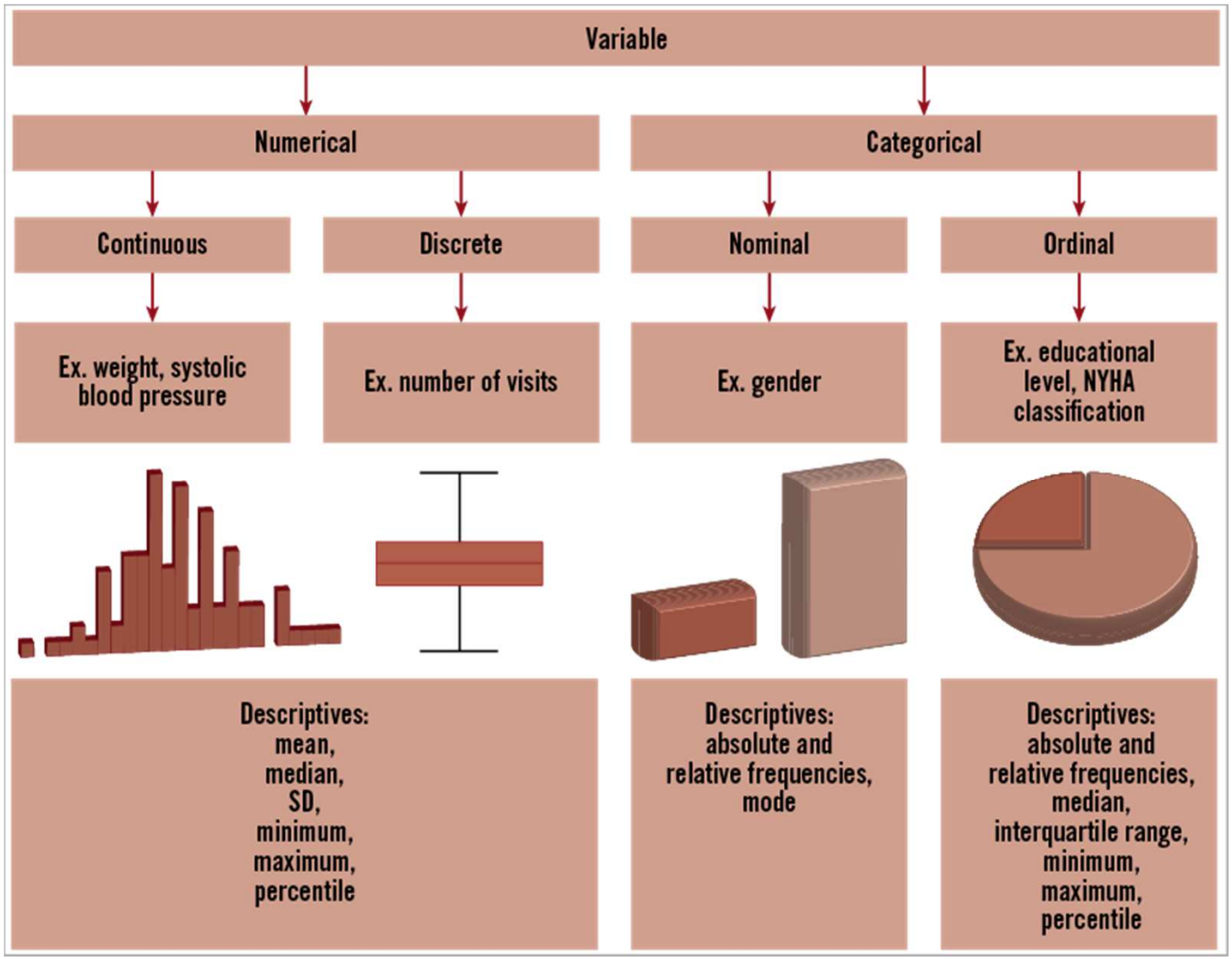


Normal distribution

Negatively skewness

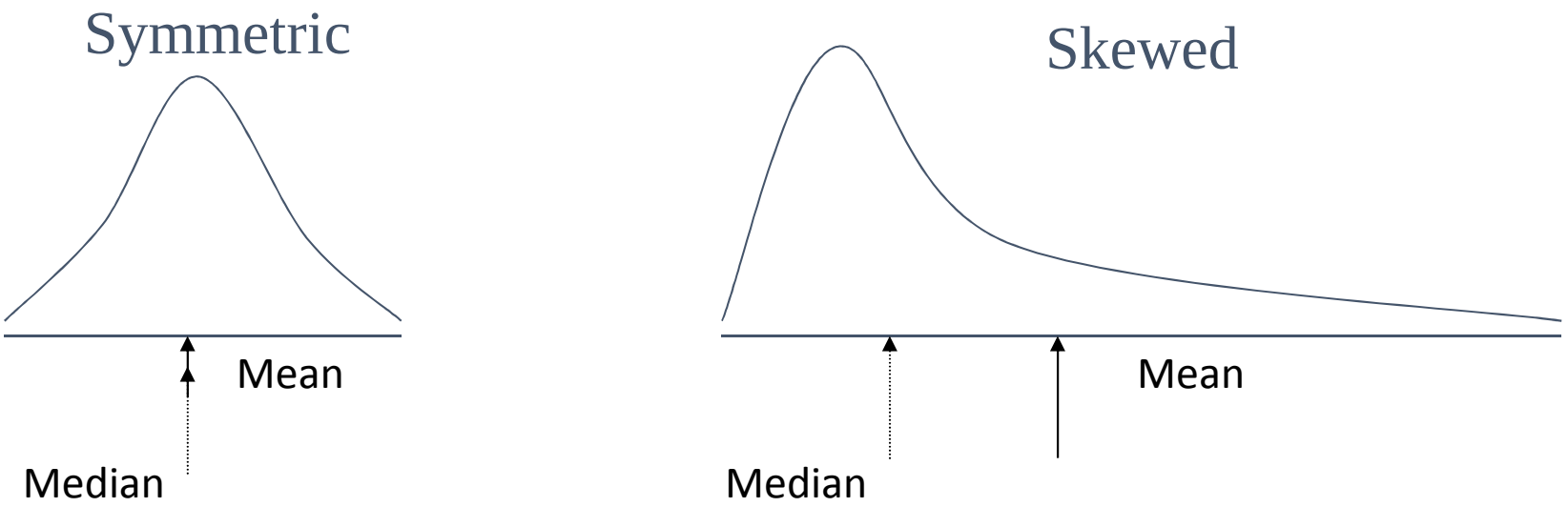


Variable types and suitable statistical measures for descriptive presentation.



Median

2. If the recorded values for a variable form a symmetric distribution, the median and mean are identical.
3. In skewed data, the mean lies further toward the skew than the median.



Interquartile Range

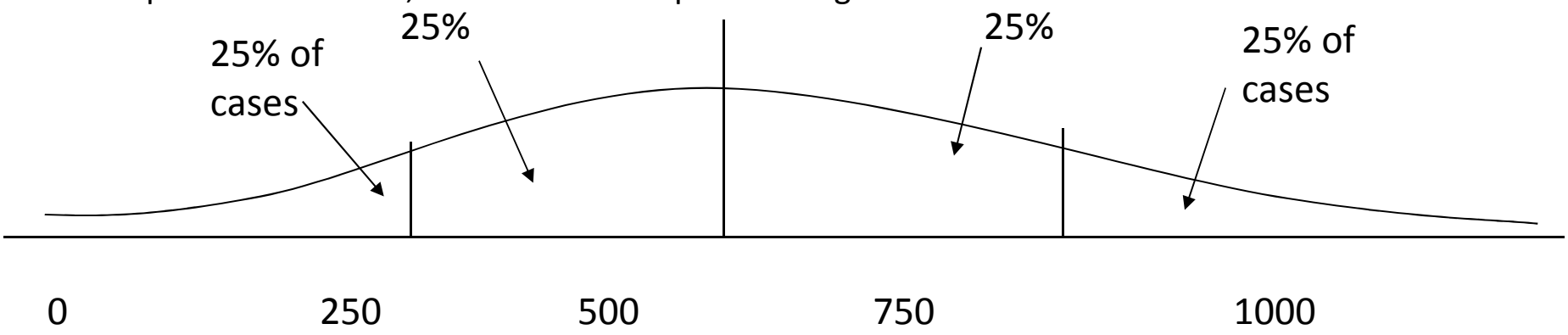
A quartile is the value that marks one of the divisions that breaks a series of values into four equal parts.

The median is a quartile and divides the cases in half.

25th percentile is a quartile that divides the first ¼ of cases from the latter ¾.

75th percentile is a quartile that divides the first ¾ of cases from the latter ¼.

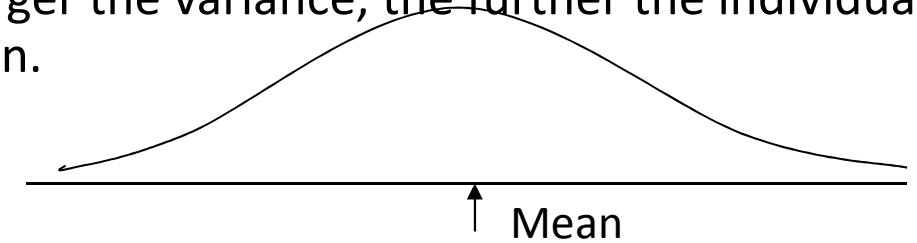
The interquartile range is the distance or range between the 25th percentile and the 75th percentile. Below, what is the interquartile range?



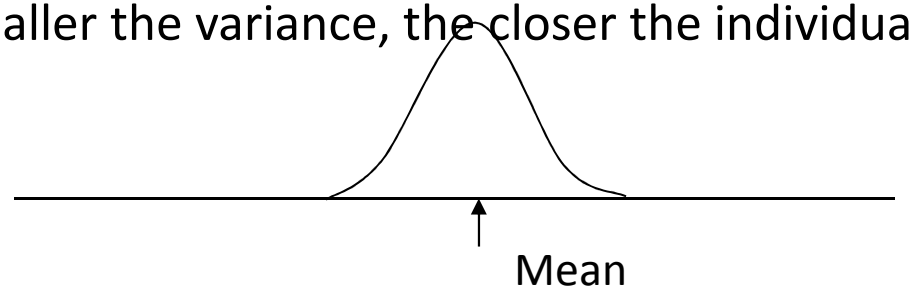
Variance

A measure of the spread of the recorded values on a variable. A measure of dispersion.

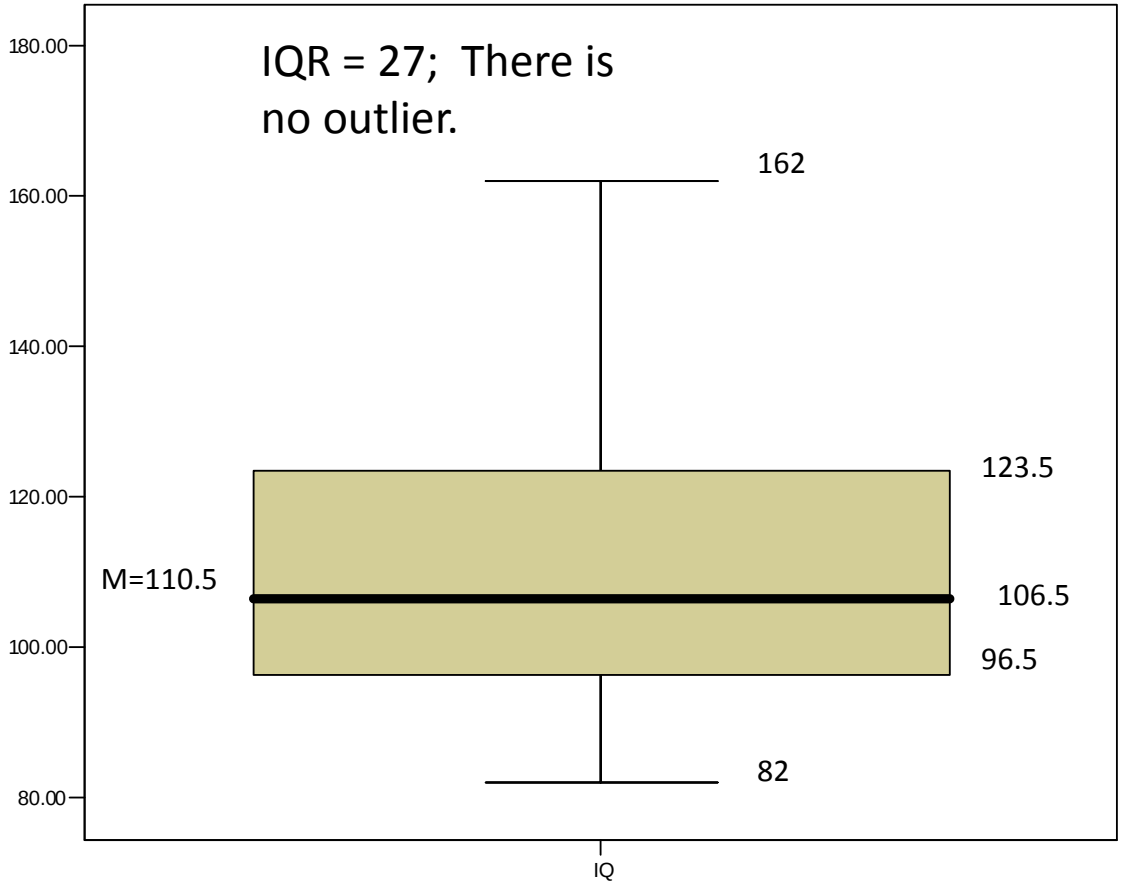
The larger the variance, the further the individual cases are from the mean.



The smaller the variance, the closer the individual scores are to the mean.



Box-Plots


IQR = 27; There is no outlier.
180.00
160.00
140.00
120.00
100.00
80.00
162
123.5
M=110.5
106.5
96.5
82
IQ

Six persistent research misconceptions

1) There is a hierarchy of study designs; randomized trials provide the greatest validity, followed by cohort studies, with case-control studies being least reliable.
2) 2) An essential element for valid generalization is that the study subjects constitute a representative sample of a target population.
3) If a term that denotes the product of two factors in a regression model is not statistically significant, then there is no biologic interaction between those factors.
4) When categorizing a continuous variable, a reasonable scheme for choosing category cut-points is to use percentile-defined boundaries, such as quartiles or quintiles of the distribution.
5) One should always report P values or confidence intervals that have been adjusted for multiple comparisons.
6) Significance testing is useful and important for the interpretation of data.

ASA Statement on Statistical Significance and P-values

1. P-values can indicate how incompatible the data are with a specified statistical model.

2. P-values do not measure the probability that the studiedhypothesis is true, or the probability that the data were produced by random chance alone.

3. Scientific conclusions and business or policy decisions should not be based only on whether a *p*-value passes a specific threshold.

4. Proper inference requires full reporting and Transparency

5. A *p*-value, or statistical significance, does notmeasure the size of an effect or the importance of a result.

6. By itself, a *p*-value does not provide a goodmeasure of evidence regarding a model or hypothesis.

Conclusion: Good statistical practice, as an essential component of good scientific practice, emphasizes principles of good study design and conduct, a variety of numerical and graphical summaries of data, understanding of the phenomenon under study, interpretation of results in context, complete reporting and proper logical and quantitative understanding of what data summaries mean. No single index should substitute for scientific reasoning

Important references

- Ronald L. Wasserstein & Nicole A. Lazar (2016) The ASA's Statement on p-Values: Context, Process, and Purpose, The American Statistician, 70:2, 129-133, DOI: 10.1080/00031305.2016.1154108
- Hoenig, J.M., andHeisey,D.M. (2001), "The Abuse of Power: The Pervasive Fallacy of Power Calculations for Data Analysis," *The American Statistician*, 55, 19–24.
- *www.sjsu.edu/people/james.lee/courses/102/s1/asDescriptive_Statistics2.ppt*

UNIVERSITY OF COPENHAGEN

Contacts

Lisbeth E. Knudsen, Professor
University of Copenhagen
Faculty of Health and Medical Sciences,
Dept of Public Health, Environmental Epidemiology Group
liek@sund.ku.dk
https://cms.ku.dk/sund-sites/ifsv-sites/ifsv-inst/

Speaker's information

Lisbeth E. Knudsen.,MSc, PhD professor in animal free toxicology has worked with ethics and HBM since 1987 in designing, performing and reporting field studies. Appointed scientific member of the Regional ethics committee. The chair of the institutional ethics committee. Leader of Task 1.5 in HBM4EU: Ethics and data protection and partner in the Task 2.5 Training. National Hub contact point of Denmark

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 733032.