Statistics in a nutshell

In the following, we will work along the open statistics introductory book by:

DOUGLAS S. SHAFER.

Professor of Mathematics University of North Carolina at Charlotte.

https://www.saylor.org/site/textbooks/Introductory%20Statistics.pdf

Online version:

https://saylordotorg.github.io/text_introductory-statistics/

Overview

- Basic probabilities
- Probability distributions
- Central Limit Theorem
- Hypotheses test
  (Z-,t-,f,Chi²- test)

Some definitions and terminology

Statistics (collection of methods)

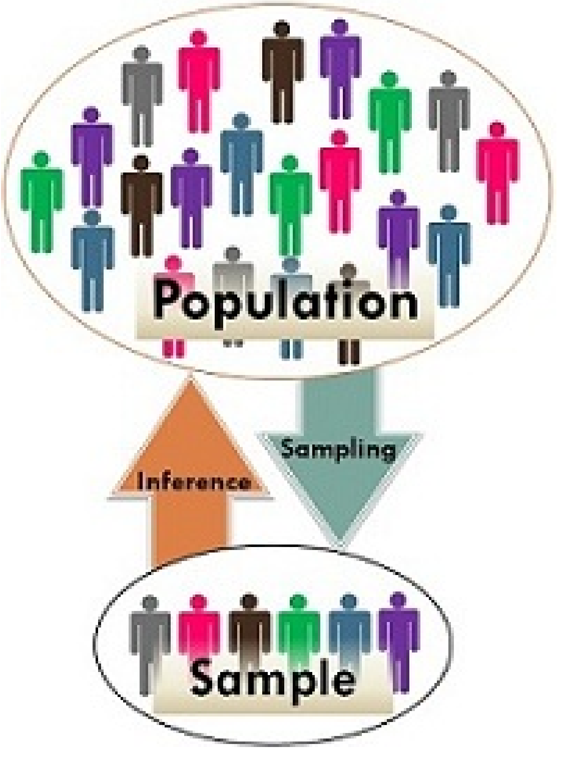
Descriptive statistics
(organizing, displaying, and describing data)

Inferential statistics
(drawing conclusions about a population
based on information contained in a sample

- A **population** is any specific collection of objects of interest
- A **sample** set of measures from the populations

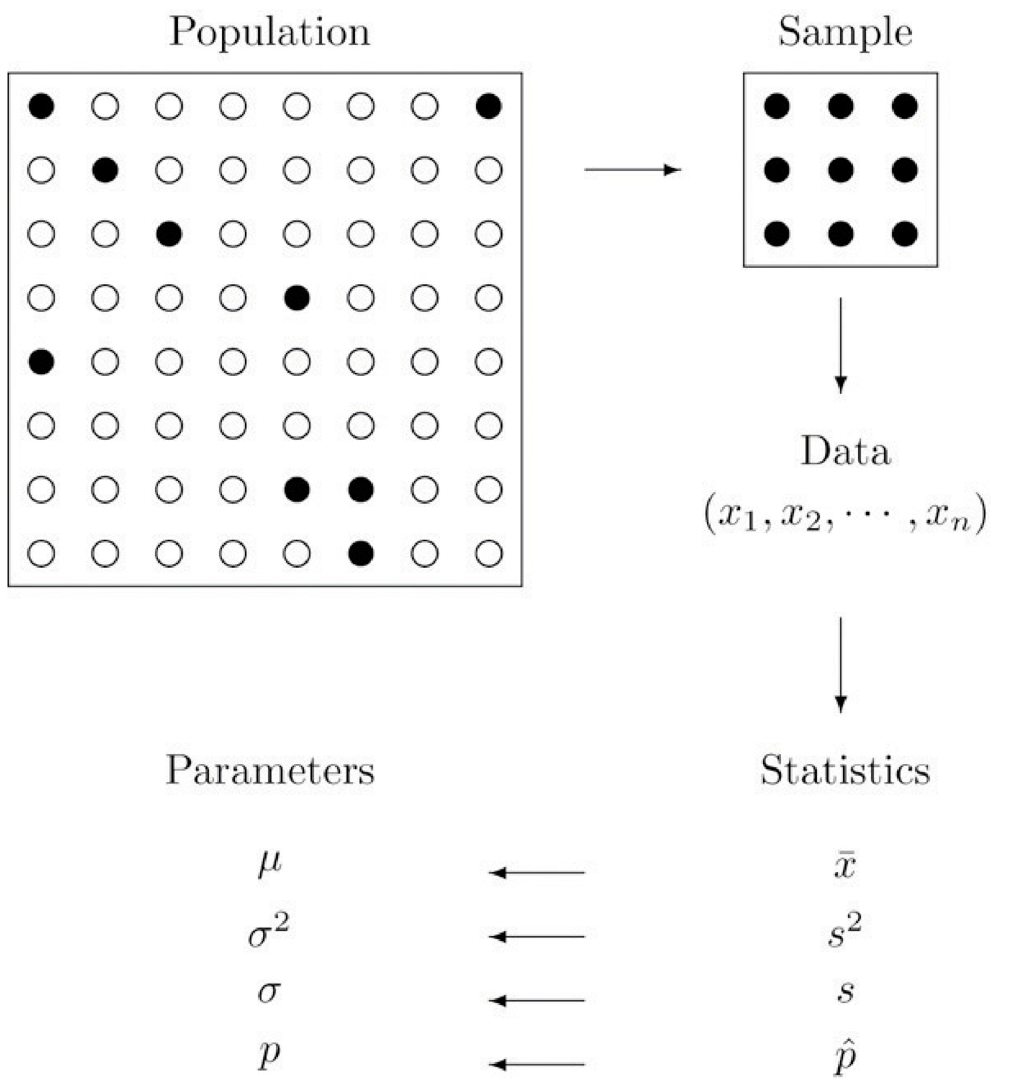
Population vs sample

- We take a random sample from the population.
- Using statistics we can learn about the population.
- Qualitative data are measurements for which there is no natural numerical scale, but which consist of attributes, labels, or other nonnumerical characteristics.
- Quantitative data are numerical measurements that arise from a natural numerical scale.



Inference from statistics

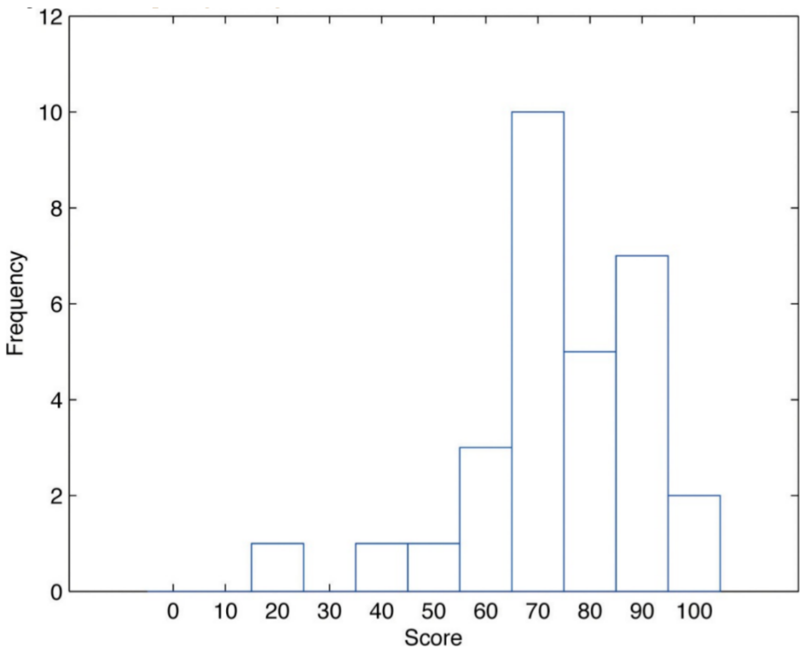
- A statistic is a number computed from the sample data.
- A parameter is a number that summarizes an aspect of the population.

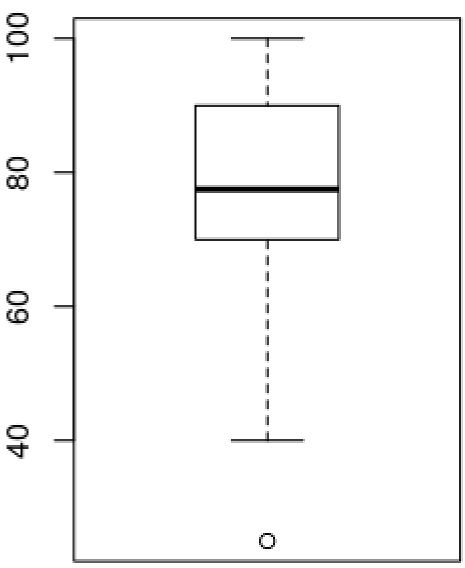


Descriptive statistics

- We received a sample of students class scores:

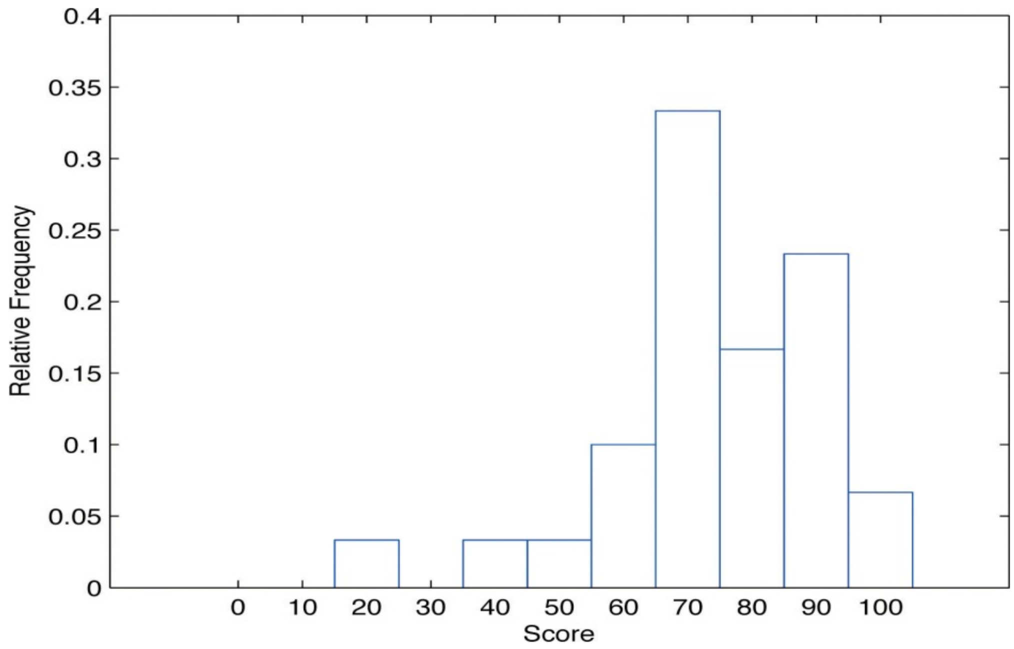
  scores={86, 80, 25, 77, 73, 76, 100, 90, 69, 93, 90, 83, 70, 73, 73, 70, 90, 83, 71, 95, 40, 58, 68, 69, 100, 78, 87, 97, 92, 74}





Relative frequency and sample size

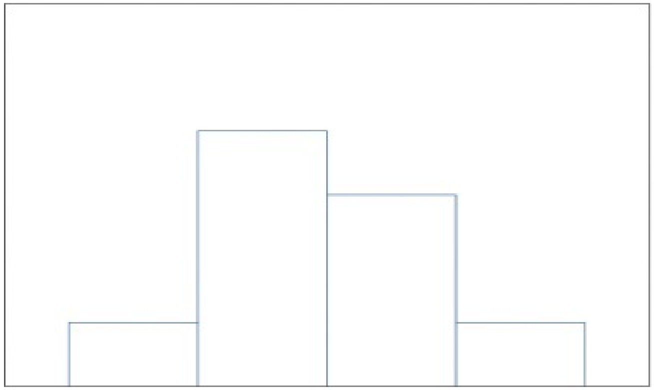
- By dividing the number of each score by the sample size, we obtain their relative frequencies.



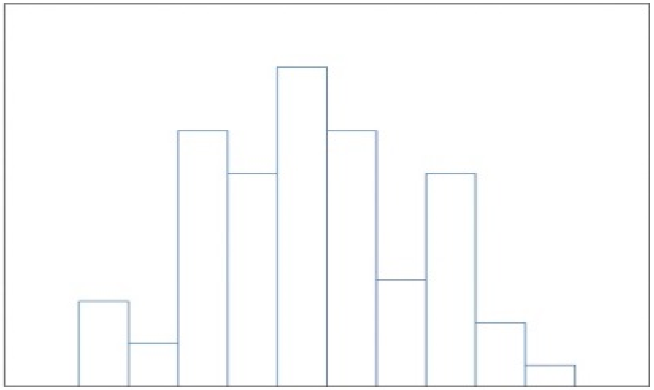
- The sum over all relative scores is 1.

How is the relative histogram effected by the sample size?

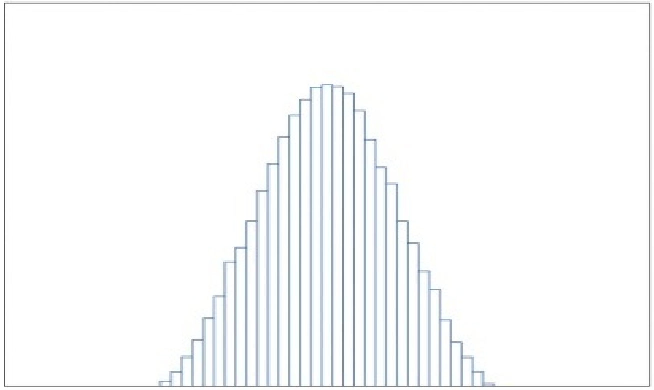
- With an increasing sample size, more cases are observed and the histogram becomes smoother.



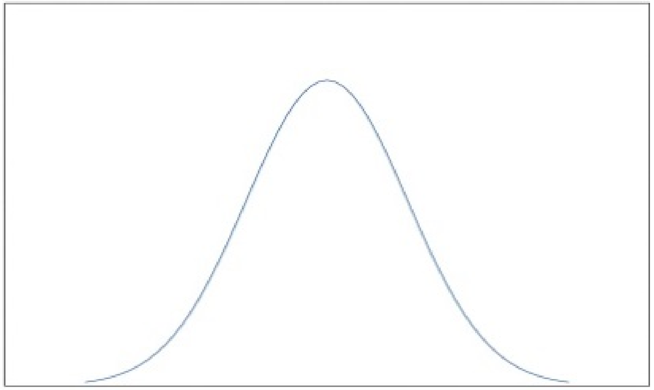
(a) Small Sample



(b) Medium Sample



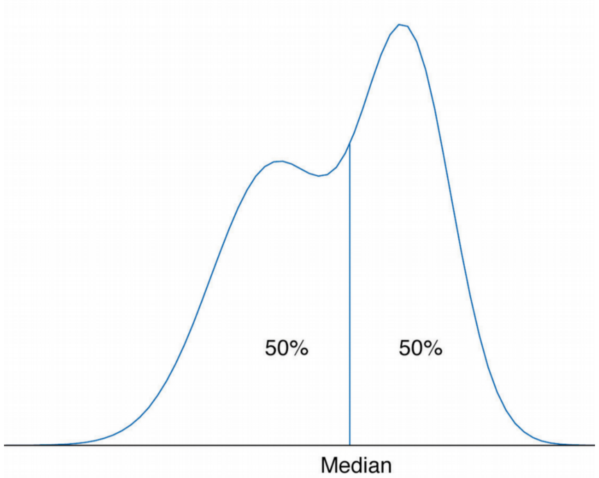
(c) Large Sample



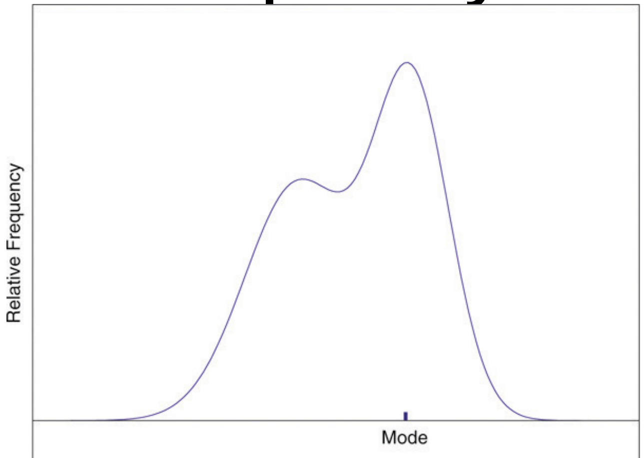
(d) Very Large Sample

Measures of central location

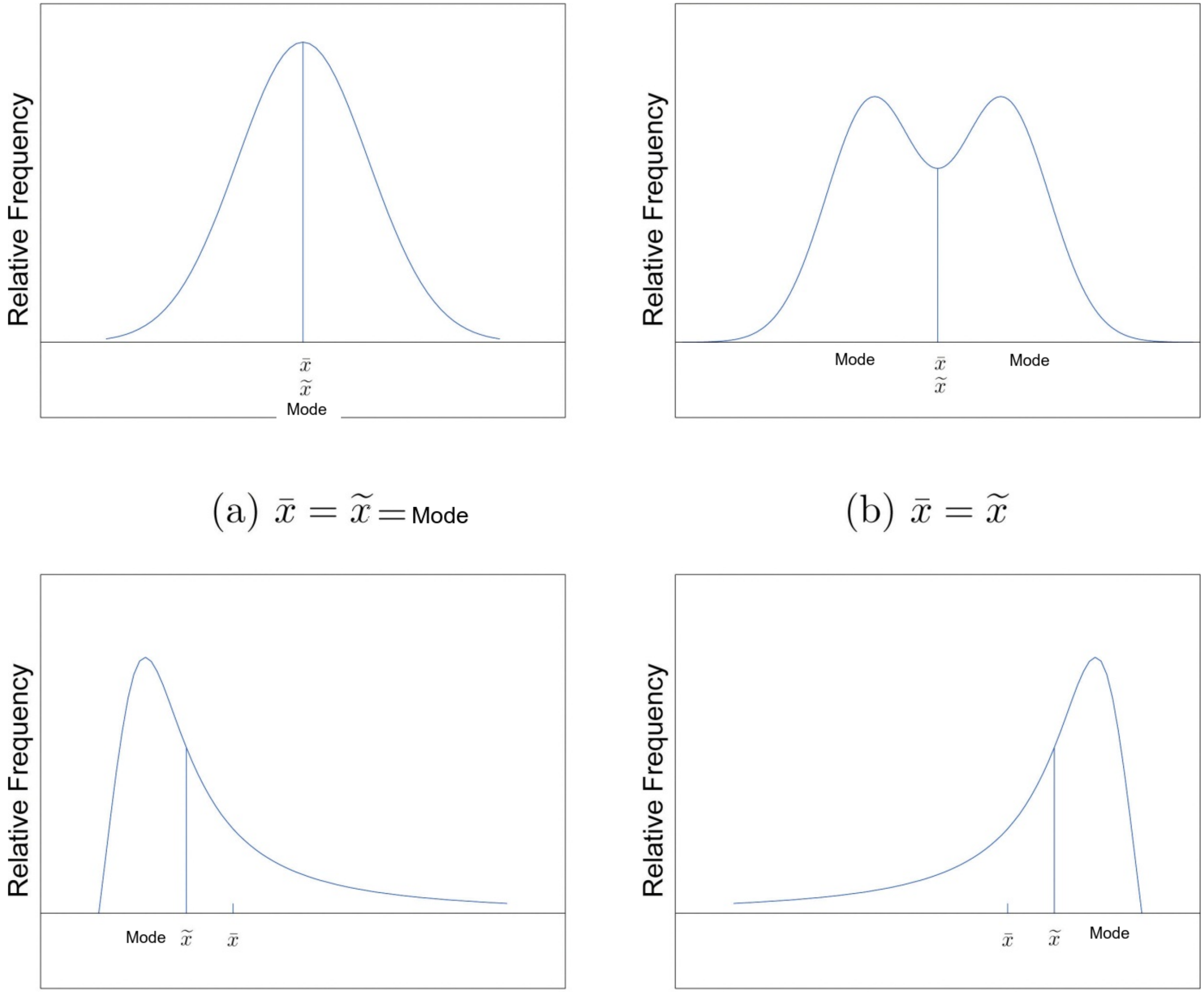
- What is the center location, though?
- Sample mean : $\overline{x} = \frac{\Sigma x}{n}$
- Median: $\tilde{x} = \begin{cases} x_{\frac{n+1}{2}} & n \text{ ungerade} \\ \frac{1}{2}\left(x_{\frac{n}{2}} + x_{\frac{n}{2}+1}\right) & n \text{ gerade.} \end{cases}$



- Mode: is the most frequently occurring value



Difference in mean,median,mode



(a) $\bar{x} = \widetilde{x} =$ Mode

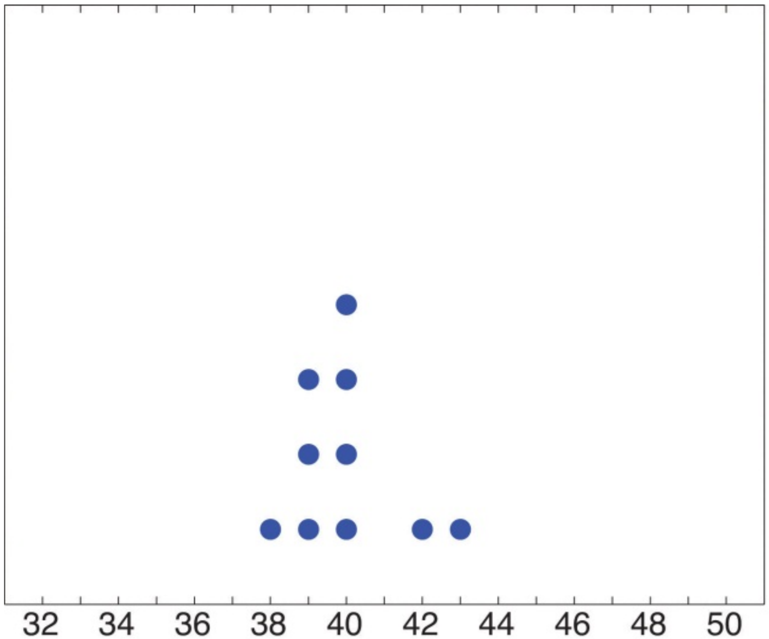
(b) $\bar{x} = \widetilde{x}$

(c) $\bar{x} > \widetilde{x} >$ Mode

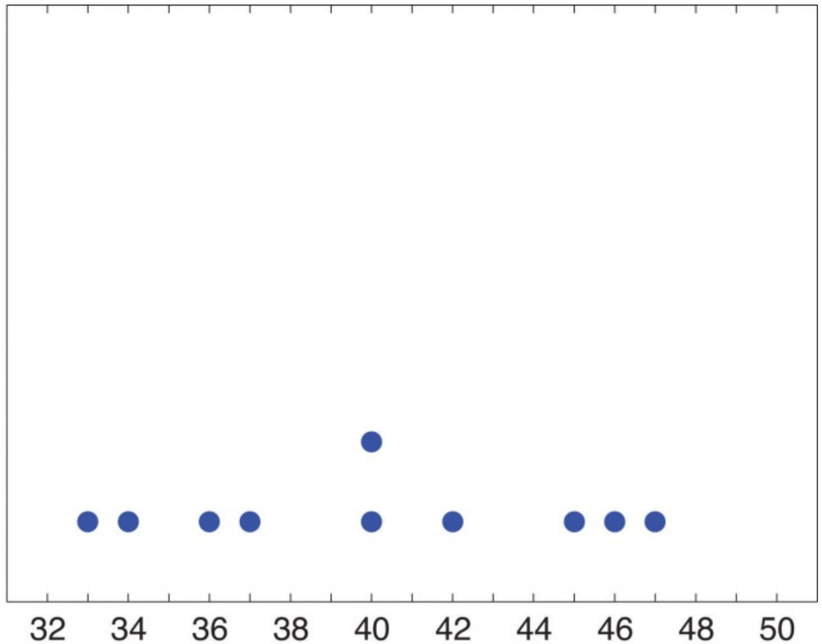
(d) $\bar{x} < \widetilde{x} <$ Mode

Measure of variability

- Set 1 and 2 both have the same mean, median, and mode of 40.



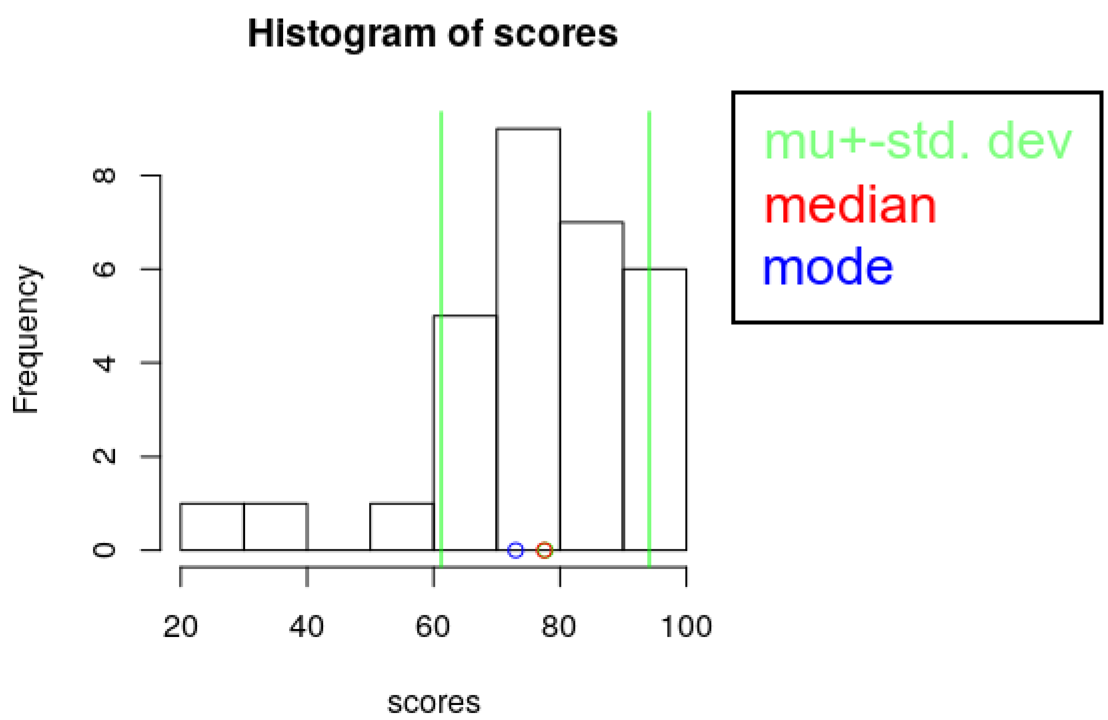
(a) Set I



(b) Set II

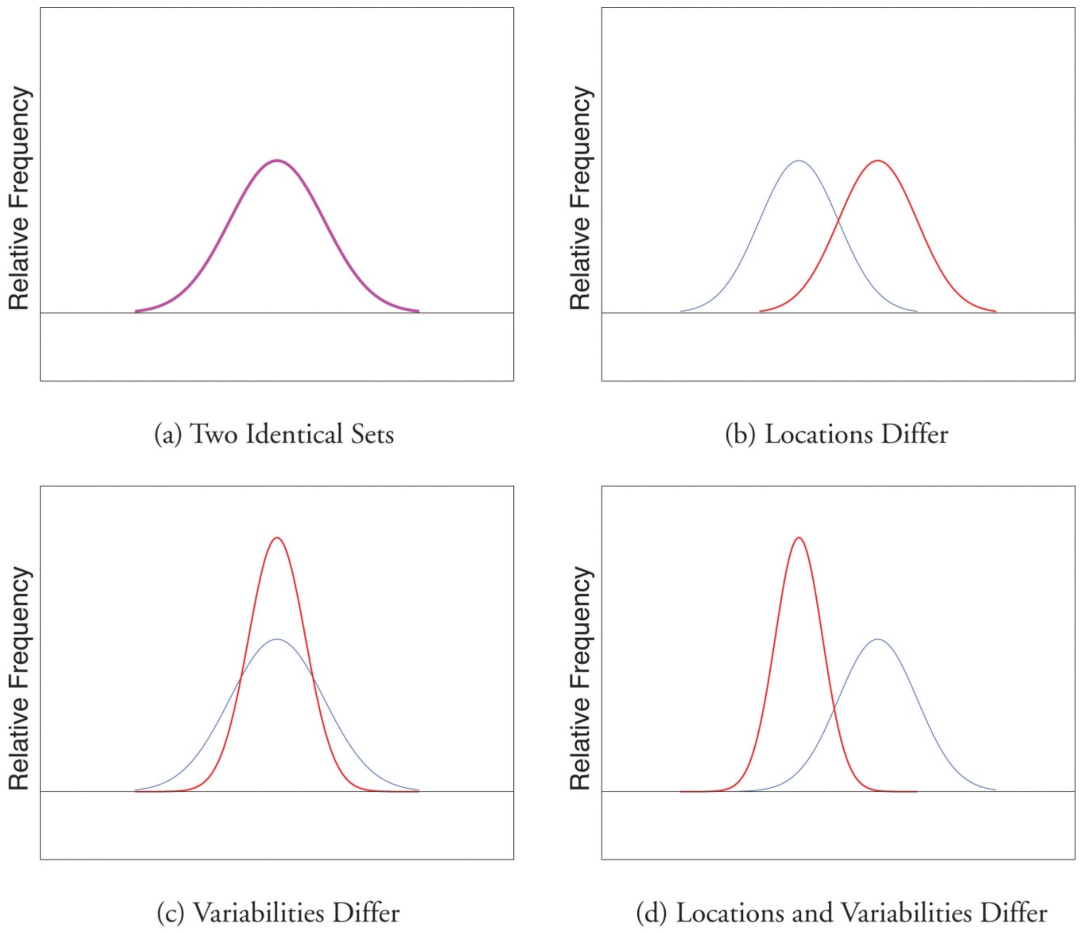
Variance and standard deviation

- Variance of a sample: $s^2 = \frac{\Sigma(x-\bar{x})^2}{n-1}$
- Sample standard deviation: $s = \sqrt{\frac{\Sigma(x-\bar{x})^2}{n-1}}$



Comparison of difference in center and variance of samples

- Figures show difference in samples.
- Statistic often compares different samples.



Percentiles and Quartiles

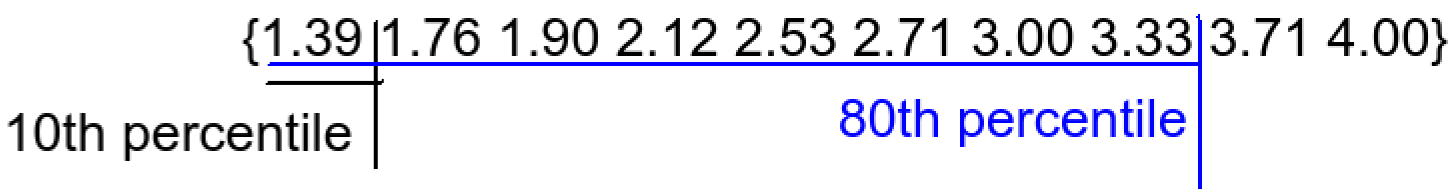
- How well is your exam score compare compared to other students if you made a 70 but the average score was 85? You did relatively poorly.
- If you made a 70, but the average score was only 55 then you did relatively well.
- Therefore, we wish to attach to each observed value a number that measures its relative position.

Pth percentile

- Given a value x in a sample, the percentile is the percentage of data less or equal than x.
- Given the data sample:

  {1.39 1.76 1.90 2.12 2.53 2.71 3.00 3.33 3.71 4.00}

  - What percentile are 1.39 and 3.33 ?



The P th percentile cuts the data set in two, so that approximately P % of the data lie below it and (100−P) % of the data lie above it.

Quartiles

- The three percentiles that cut the data into fourths are called the quartiles

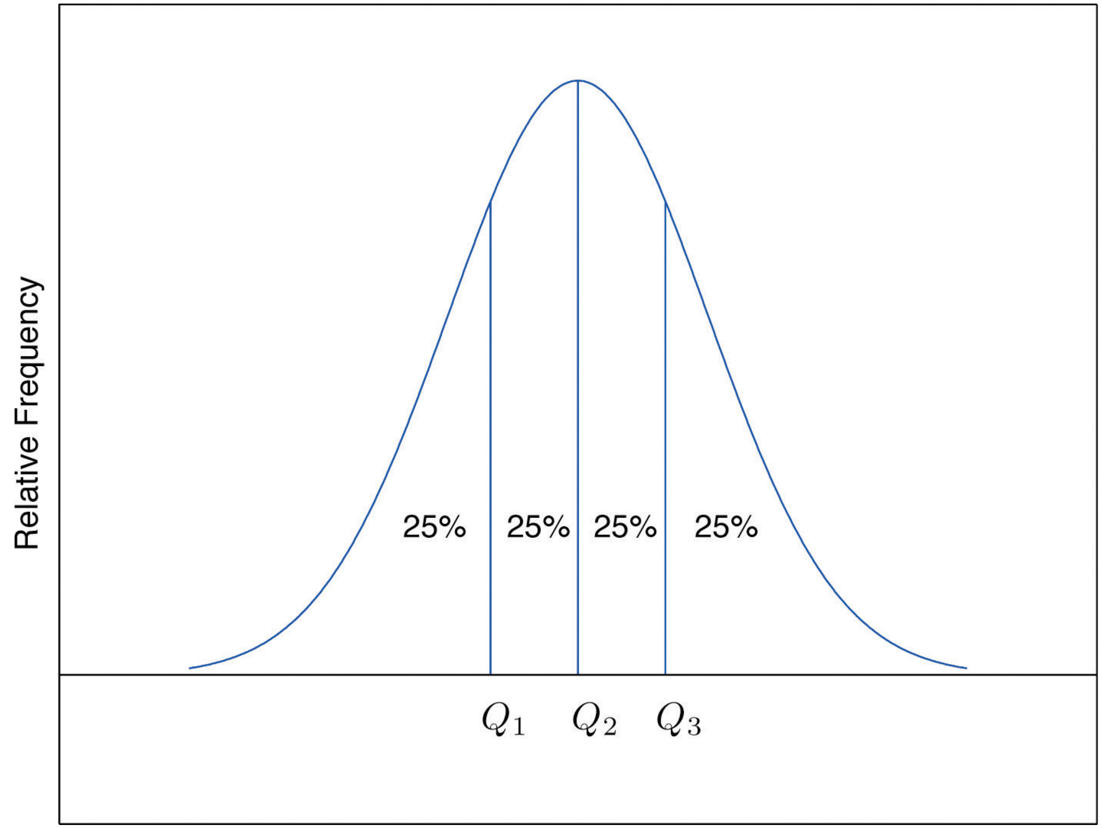
The second quartile Q_2 of the data set is its median.
It define two subsets:

1. the lower set: all observations that are strictly less than Q_2 ;

2. the upper set: all observations that are strictly greater than Q_2 .

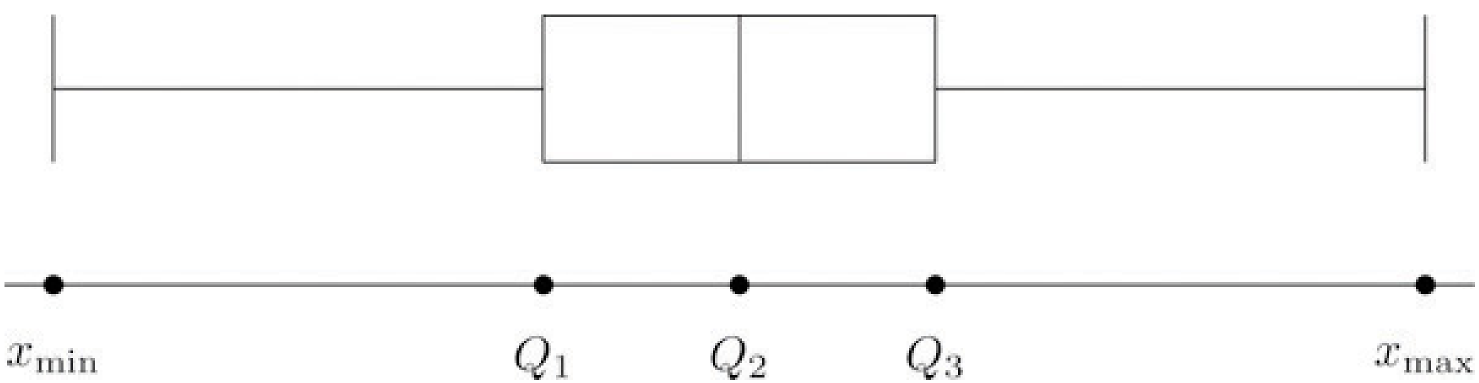
The first quartile Q_1 of the data set is the median of the lower set

The third quartile Q_3 of the data set is the median of the upper set.



Boxplot

- In addition to the three quartiles, the two extreme values, the minimum x_{min} and the maximum x_{max} are useful in describing the data.
- The five-number summary: $\{x_{min}, Q_1, Q_2, Q_3, x_{max}\}$ is used to construct a box plot



z-score

- Another way to locate a particular observation x in a data set is to compute its distance from the mean in units of standard deviation.
- Z-score: $z = \frac{x - \overline{x}}{s}$
- The z-score indicates how many standard deviations an individual observation x is from the mean of the data set.