# The principals of statistical analysis

(Statistikk i helsefagleg forsking)

HELSTA Fall 2014

Øystein Haaland

Postdoc, Research Group for Genetic Epidemiology

Department of  Global Public Health and Primary Care

University of Bergen

# The aim of this course

- is to present the basic concepts of statistical analysis
- we will <u>not</u> go through the mathematics of statistical tests
- we will <u>not</u> show you how to do the tests on a computer
- but, we will go through some important principles, including generalization and the idea of the p-value

# Quantitative research methods

- Is based on an accumulation of data from a limited number of variables within a sample of individuals (things)

- This entails uncertainty.

- Statistical analysis gives a measure of part of this uncertainty

**Descriptive statistics** (beskrivende statistikk)

- use of numbers for describing the main features of a collection of data (sample) on individuals (or things)

**Statistical inference** (induktiv statistikk - statistiske slutninger)

- use of numbers from a sample of data for making inferences concerning some unknown aspect of a population where the sample was drawn from

# A study of length of hospitalizations at two hospitals A and B

- Length of hospitalization measured as number of bed-days
- 50 hospitalizations registered for each of the two hospitals
- How to communicate the results?

1.

Hospital A:
5, 7, 2, 1, 13, 4, 9, 83, 17, 38, 4, 12, 22,
50, 2, 1, 11, 9, 33, 7, 8, 3, 21, 9, 8, 15,
2, 11, 7, 1, 3, 2, 17, 9, 11, 6, 8, 3, 3, 14,
17, 4, 18, 9, 7, 7, 24, 11, 2, 8.

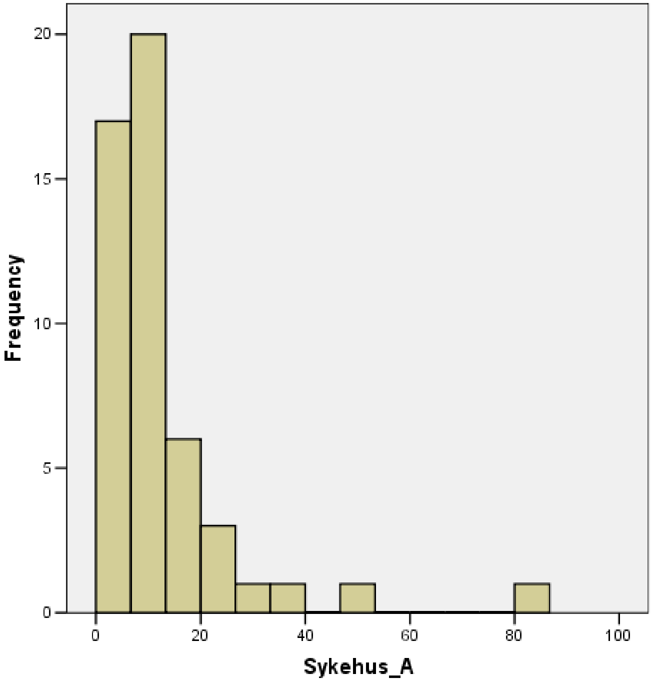
Hospital B:
9, 22, 11, 19, 17, 33, 8, 2, 17, 36, 42, 9,
17, 7, 8, 4, 40, 11, 13, 7, 19, 8, 2, 6, 28,
7, 6, 13, 14, 4, 12, 6, 5, 36, 16, 15, 19,
9, 2, 12, 9, 30, 7, 9, 12, 14, 17, 22, 7.

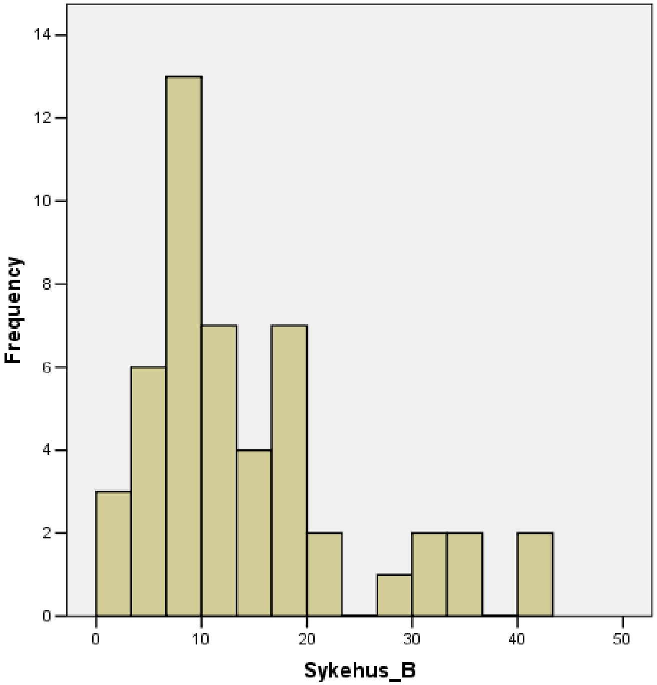
2.

At hospital A the number of bed days ranged from 1 to 83 days.

At hospital B the number of bed days ranged from 2 to 42 days.

3.


Frequency
20
15
10
5
0
0
20
40
60
80
100
Sykehus_A


Frequency
14
12
10
8
6
4
2
0
0
10
20
30
40
50
Sykehus_B

4.

Mean for hospital A:
11.96 days

Mean for hospital B:
14.24 days

5.

The empirical frequency distribution at the two hospitals could be approximated by a chi-square distribution with 9 and 11 degrees of freedom, respectively.

6.

**Statistics**

Sykehus_A

| | | |
|---|---|---|
| N | Valid | 50 |
| | Missing | 0 |
| Mean | | 11,96 |
| Std. Error of Mean | | 1,995 |
| Median | | 8,00 |
| Mode | | 2[a] |
| Std. Deviation | | 14,109 |
| Variance | | 199,060 |
| Skewness | | 3,254 |
| Std. Error of Skewness | | ,337 |
| Kurtosis | | 13,339 |
| Std. Error of Kurtosis | | ,662 |
| Range | | 82 |
| Minimum | | 1 |
| Maximum | | 83 |
| Sum | | 598 |
| Percentiles | 10 | 2,00 |
| | 20 | 3,00 |
| | 25 | 3,75 |
| | 30 | 4,30 |
| | 40 | 7,00 |
| | 50 | 8,00 |
| | 60 | 9,00 |
| | 70 | 11,70 |
| | 75 | 14,25 |
| | 80 | 17,00 |
| | 90 | 23,80 |

[a] Multiple modes exist. The smallest value is shown

**Statistics**

Sykehus_B

| | | |
|---|---|---|
| N | Valid | 49 |
| | Missing | 1 |
| Mean | | 14,24 |
| Std. Error of Mean | | 1,443 |
| Median | | 12,00 |
| Mode | | 7[a] |
| Std. Deviation | | 10,101 |
| Variance | | 102,022 |
| Skewness | | 1,263 |
| Std. Error of Skewness | | ,340 |
| Kurtosis | | ,987 |
| Std. Error of Kurtosis | | ,668 |
| Range | | 40 |
| Minimum | | 2 |
| Maximum | | 42 |
| Sum | | 698 |
| Percentiles | 10 | 4,00 |
| | 20 | 7,00 |
| | 25 | 7,00 |
| | 30 | 8,00 |
| | 40 | 9,00 |
| | 50 | 12,00 |
| | 60 | 14,00 |
| | 70 | 17,00 |
| | 75 | 18,00 |
| | 80 | 19,00 |
| | 90 | 33,00 |

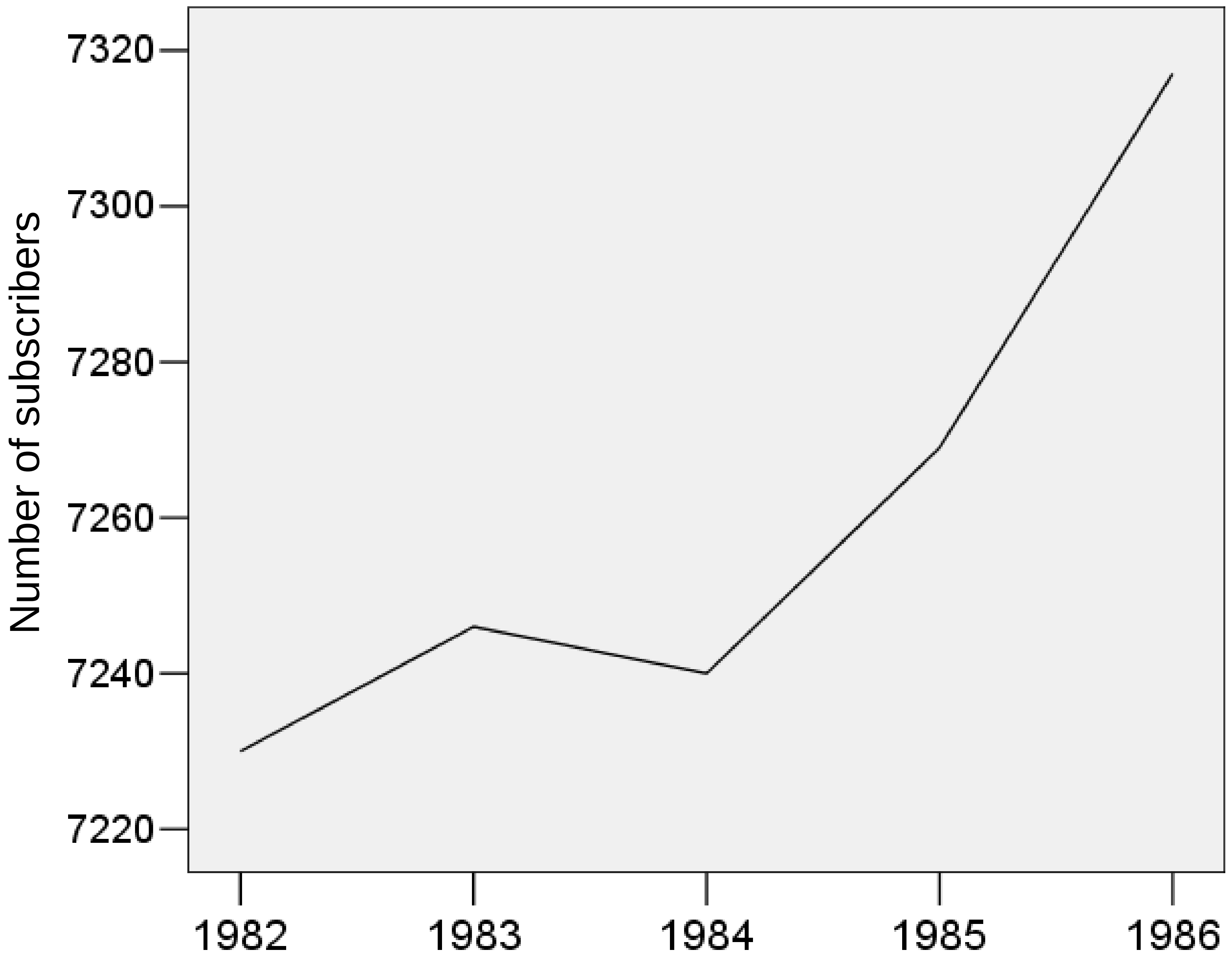
[a] Multiple modes exist. The smallest value is shown

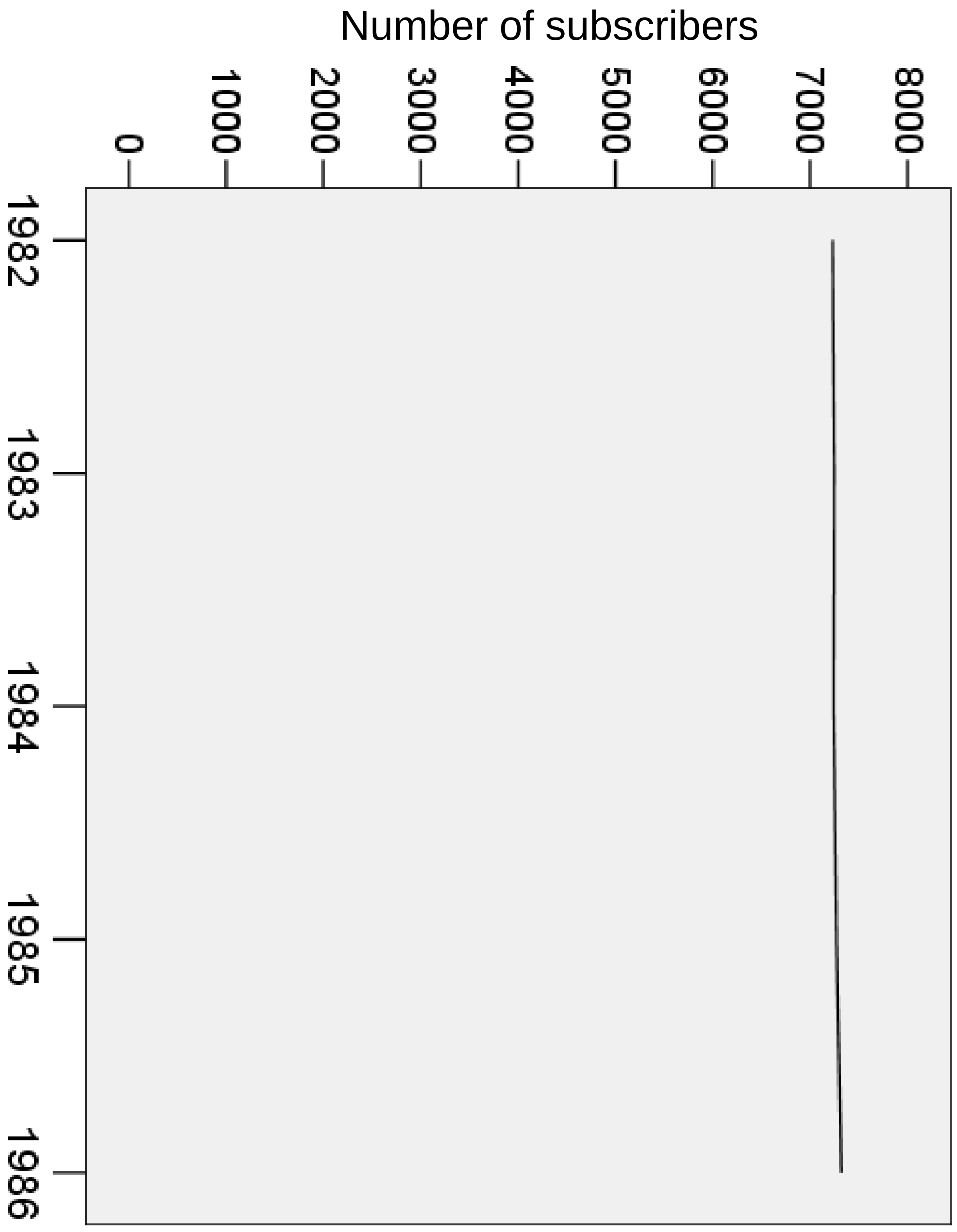
# Statistical inference

*“At least one-third of the patients at hospital B have more that 25 bed days”*

*“The patients at hospital B are staying longer in the hospital than at hospital A”*

Requires representativity of the groups being investigated


Number of subscribers
7220
7240
7260
7280
7300
7320
1982
1983
1984
1985
1986


Number of subscribers
0
1000
2000
3000
4000
5000
6000
7000
8000
1982
1983
1984
1985
1986

# Empirical study

- Population
  - the group of individuals or thing one wants to study
- Sample
  - the part of the population where real data actually are collected
  - needs to be representative for the population
  - in principle should be drawn at random from the population

**Study of total population** (totalundersøkelse) – the complete population is being examined

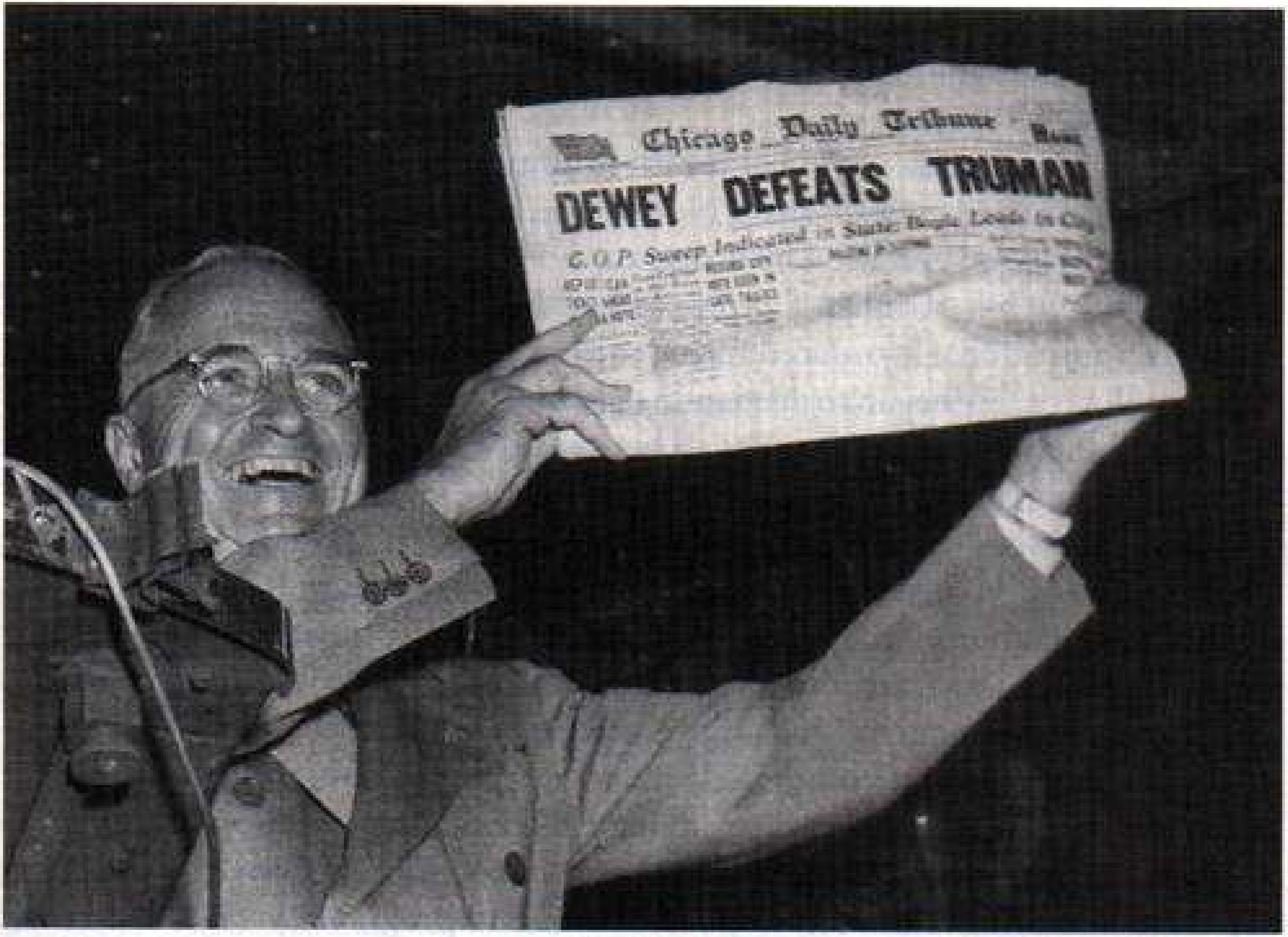
## A random sample

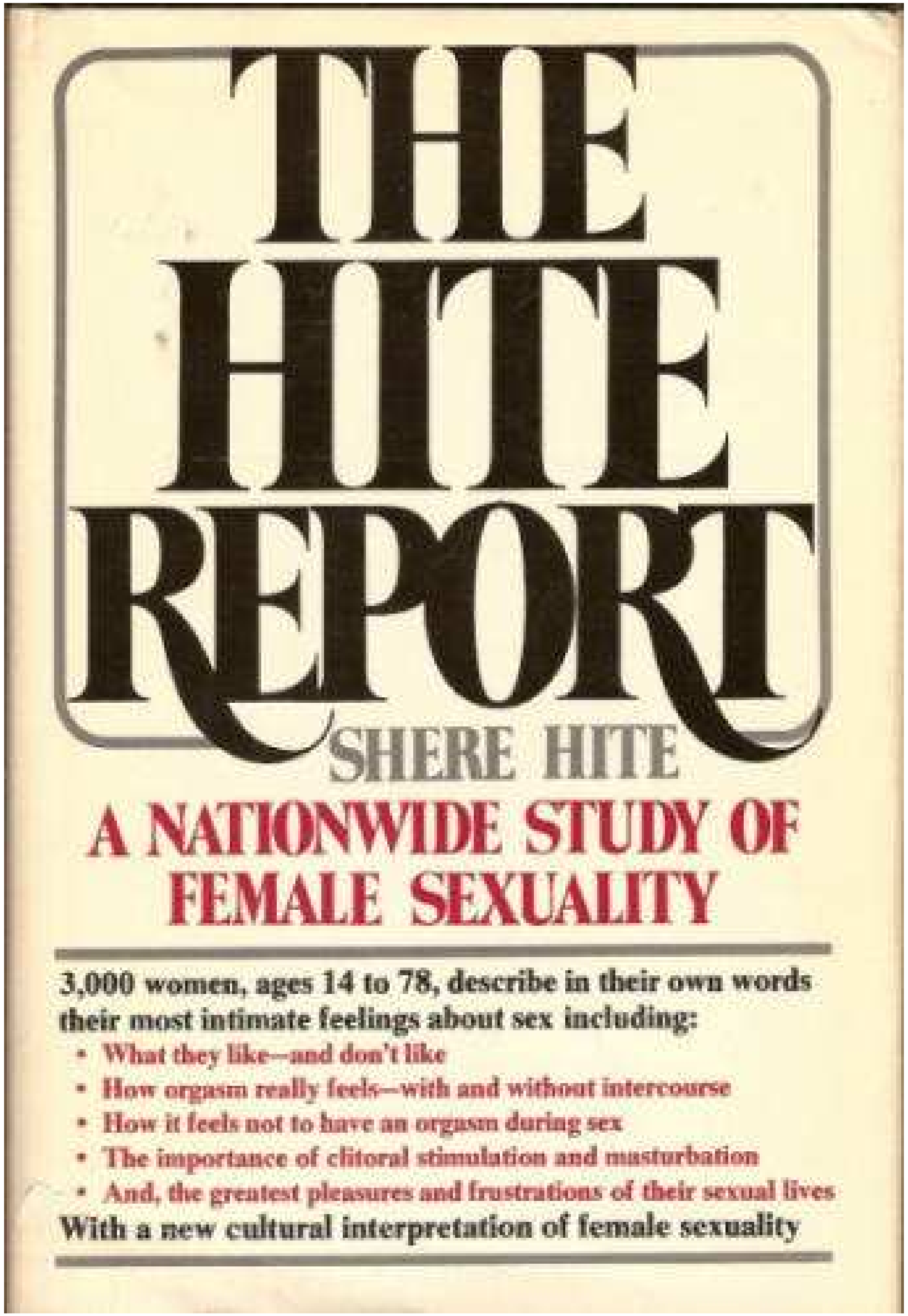
A sample is a random sample if all individuals in the population have had the same probability of being part of the sample.

(*actually:* A sample is a random sample if all possible combination of samples have the same probability of being the final sample)

**A stratified random sample** is achieved by drawing random samples in subgroups of the population and then add these to one sample.

- gender, age, area of residence (polls)


Chicago Daily Tribune
DEWEY DEFEATS TRUMAN
G.O.P. Sweep Indicated in State; Boyle Leads in City



# THE HITE REPORT

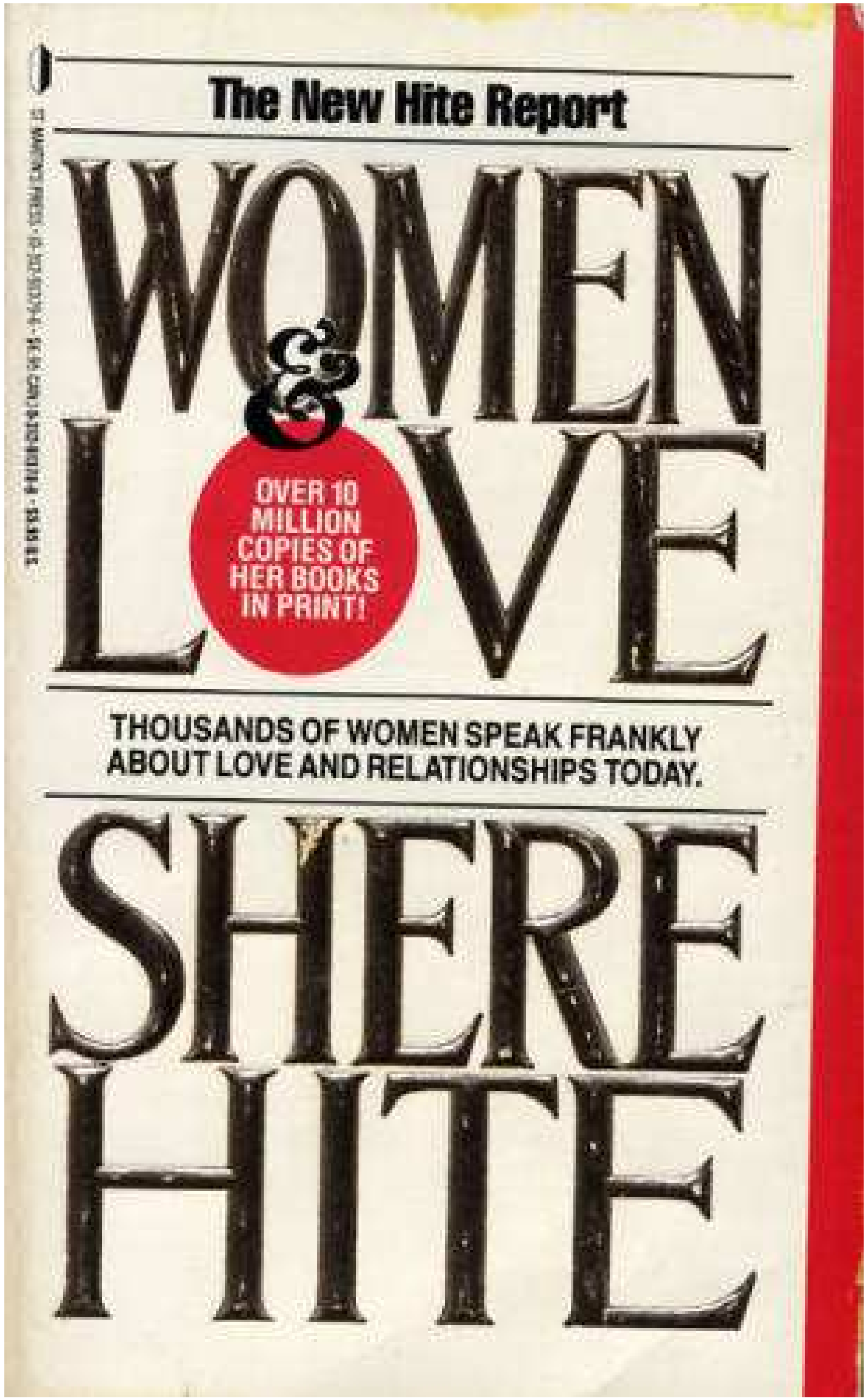
SHERE HITE

## A NATIONWIDE STUDY OF FEMALE SEXUALITY

3,000 women, ages 14 to 78, describe in their own words their most intimate feelings about sex including:

- What they like--and don't like
- How orgasm really feels—with and without intercourse
- How it feels not to have an orgasm during sex
- The importance of clitoral stimulation and masturbation
- And, the greatest pleasures and frustrations of their sexual lives

With a new cultural interpretation of female sexuality



The New Hite Report

# WOMEN & LOVE

OVER 10 MILLION COPIES OF HER BOOKS IN PRINT!

THOUSANDS OF WOMEN SPEAK FRANKLY ABOUT LOVE AND RELATIONSHIPS TODAY.

SHERE HITE

ST. MARTIN'S PRESS · 0-312-91379-6 · $6.95 CAN. | 0-312-91378-8 · $5.95 U.S.

**WHY IS SHERE HITE'S NEW REPORT
THE MOST PROVOCATIVE
BESTSELLER TODAY?**

**BECAUSE OF THE STARTLING HONESTY
WITH WHICH THOUSANDS OF WOMEN
REVEAL THEIR INNERMOST FEELINGS
ABOUT LOVE AND RELATIONSHIPS.**

91378
0 44903 00595 3

ISBN 0-312-91378-8

# **Variable** – what is measured in a study

- **Nominal (categorical) variables**
  - Type of treatment (new / old - type A / type B / type C)
  - Gender, Area of residence
  - Social group, Occupation

- **Ordinal variables** (may be ordered)
  - Dose of treatment
  - The amount of information given to a patient
  - Age, height, weight, number of children in a family

- **Observations**
  - The specific values of the variable being measured (i.e. man, woman; 24 years of age, 29 years of age)

# Ordinal variables

- Ordinal scale
  - Observations may be ordered (given a rank)
    - Freezing, liquid, boiling temperatures
- Interval scale
  - as above + same distance between each level
    - Degrees Celcius
- Ratio scale
  - as above + existence of a natural zero
    - Degrees Kelvin

# Translation of a concept to a measurable variable (operationalization)

*Examples:*

- Quality of life
- Hope
- Fatigue
- Mobility of a joint
- Pain

*Involves evaluation of:*

- Validity
- Reliability

- ***Dependent variable*** – the variable that can be considered being “influenced” by other variables

- ***Independent variable*** – the variable that is “influencing” the dependent variable

# Dependent – independent variable

Two problems:

- *Are there systematic geographical differences in how the social security offices organise the child welfare work?*
- *Do nurses working for the city spend more time on preventive health measures than physicians working for the city?*

What would you define as dependent and independent variable for these two approaches?

# Why calculate descriptive statistics?

- To describe the distribution of the observations in a data set as well as possible
  - a) to find errors in the data
  - b) for simply describing the sample
- To check whether an assumption for doing some specific statistical analyses is fulfilled, i.e.
  - a) whether a variable has a normal distribution
  - b) whether the variance of the a variable is similar for two groups being compared
- Simply to explore the data in search of new and unexpected findings not related to a previous formal hypothesis (explorative data analysis, hypothesis generating analysis)

# Descriptive statistics – main types

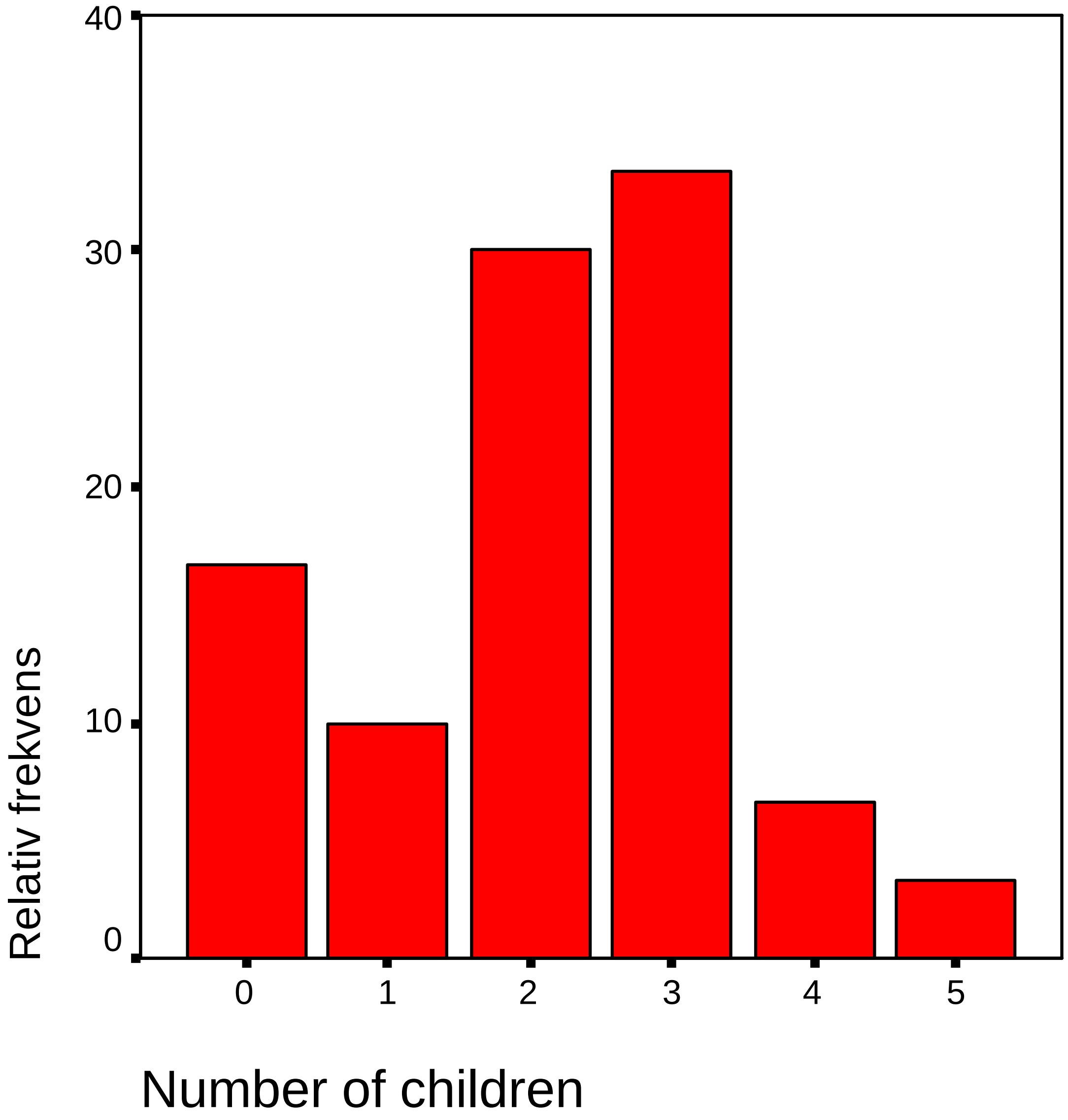
- Frequency tables
- Graphical presentation
- Summary statistics

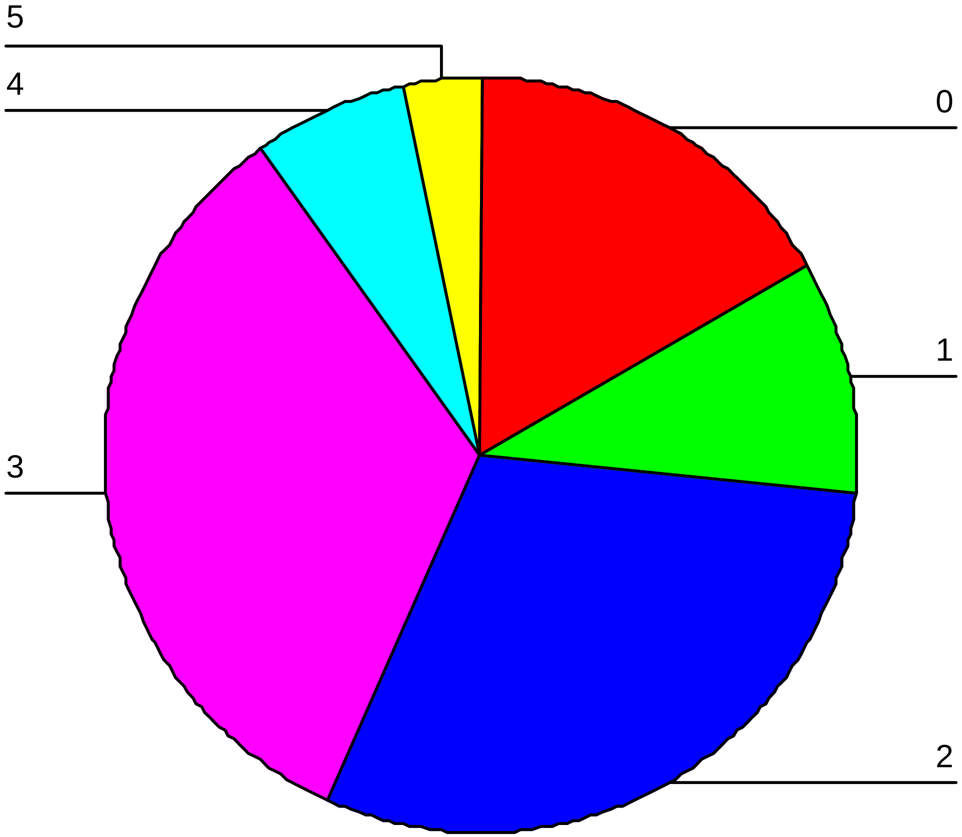
# Frequency table

Number of children of 30 randomly chosen individuals aged 40-45 years in Hordaland, Norway

3, 0, 2, 0, 2, 1, 2, 3, 0, 3, 5, 4, 2, 3, 2, 2, 3, 3, 3, 4, 2, 0, 3, 1, 1, 2, 0, 3, 3, 2.

|  | Absolutt frekvens | Relativ frekvens | Kumulativ frekvens |
|---|---|---|---|
| 0 | 5 | 0.17 | 0.17 |
| 1 | 3 | 0.10 | 0.27 |
| 2 | 9 | 0.30 | 0.57 |
| 3 | 10 | 0.33 | 0.90 |
| 4 | 2 | 0.07 | 0.97 |
| 5 | 1 | 0.03 | 1.00 |
| Totalt | 30 | 1.00 |  |


40
30
20
10
0
Relativ frekvens
0
1
2
3
4
5
Number of children


5
4
0
1
3
2

# Descriptive statistics

*Summary statistics*

- Measure of central tendency
- Measure of variation

# Measures of central tendency

- Mean
  - Sum of all observations divided by the number of observations
  - Strongly influenced by extreme values
- Median
  - The middle value in the distribution arranged according to size
- Mode
  - The value that is most frequently observed
  - Little used, but is an intuitive measure

# Height of Norwegian women in cm

- 166
- 170
- 162
- 158
- 173
- 166
- 163
- 176
- 151
- 155

Mean = (166+170+ … +155) / 10 = 1640/10 = 164

Median: 151, 155, 158, 162, 163, 166, 166, 170, 173, 176

**164,5**

Mode : 166

# Problem – measure of central tendency

A new drug for headache is tested and one wish to estimate the duration before the drug is taking effect.

15 individuals with headache is given the drug at a recommended dose and they are asked to report when the drug is taking effect.

These are the results in minutes from intake of tablet to effect:

21, 15, 2, 17, 14, 85, 19, 16, 340, 15, 12, 8, 45, 11, 19

*Based on these data give an estimated time before effect is reached for this drug.*

# Measures of variability

- **Range**
  - The distance between the lowest and highest value
  - Strongly influenced by extreme values

- **Standard deviation**
  - Gives the ”mean variation” around the mean

- **Interquartile range**
  - The distance between the upper and lower quartile

# Standard deviation

- Two samples A and B
- A: 8  9 10 10 10 11 12
- B: 4  4  6  10 14 16 16

# Standard deviation

- Two samples A and B

- A: 8 9 10 10 10 11 12, mean=10
  sd=1.3
- B: 4 4 6 10 14 16 16, mean=10
  sd=5.4

# Height of Norwegian women in cm

- 166
- 170
- 162
- 158
- 173
- 166
- 163
- 176
- 151
- 155

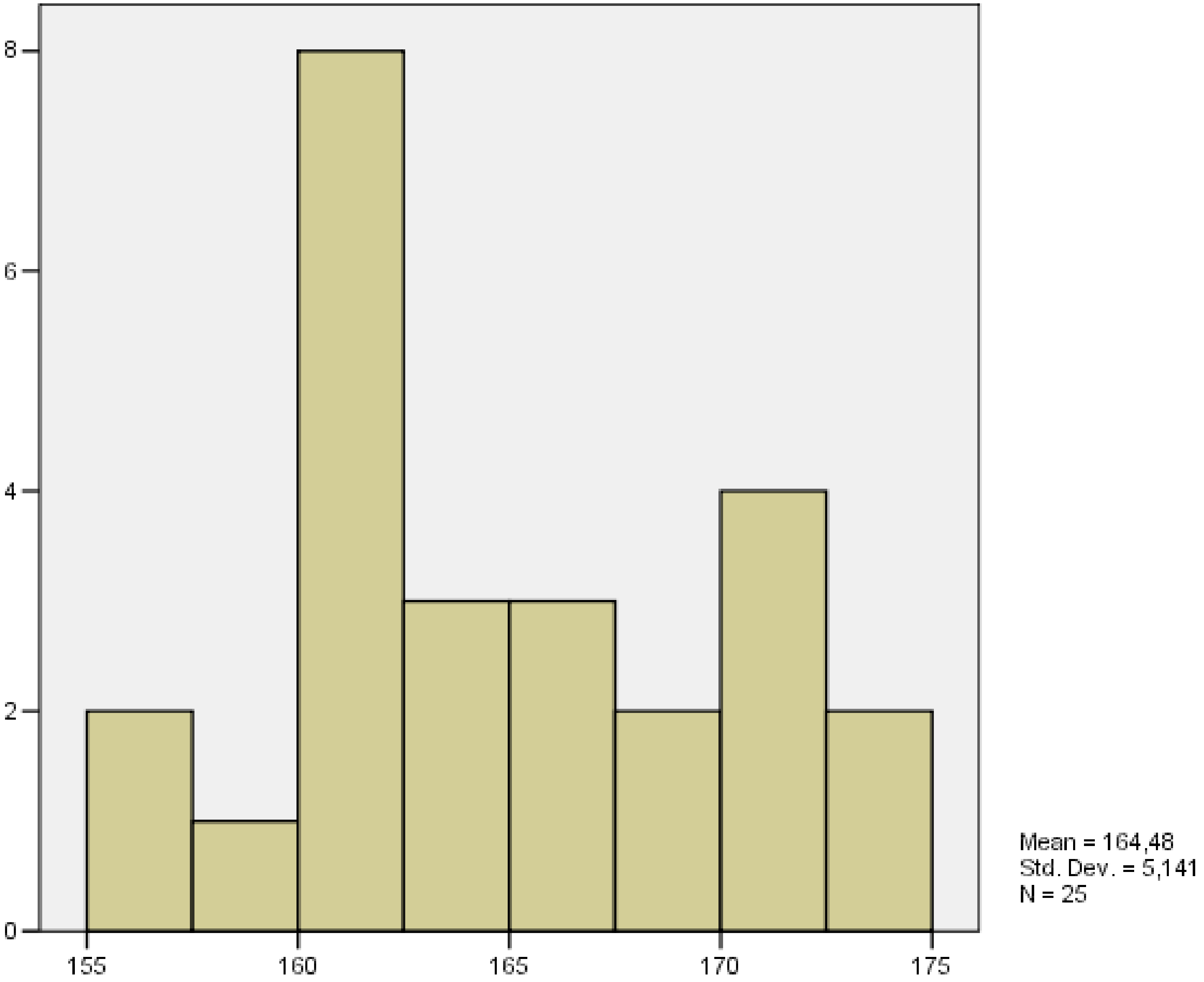
Range: 176 - 151 = 25

Standard deviation:

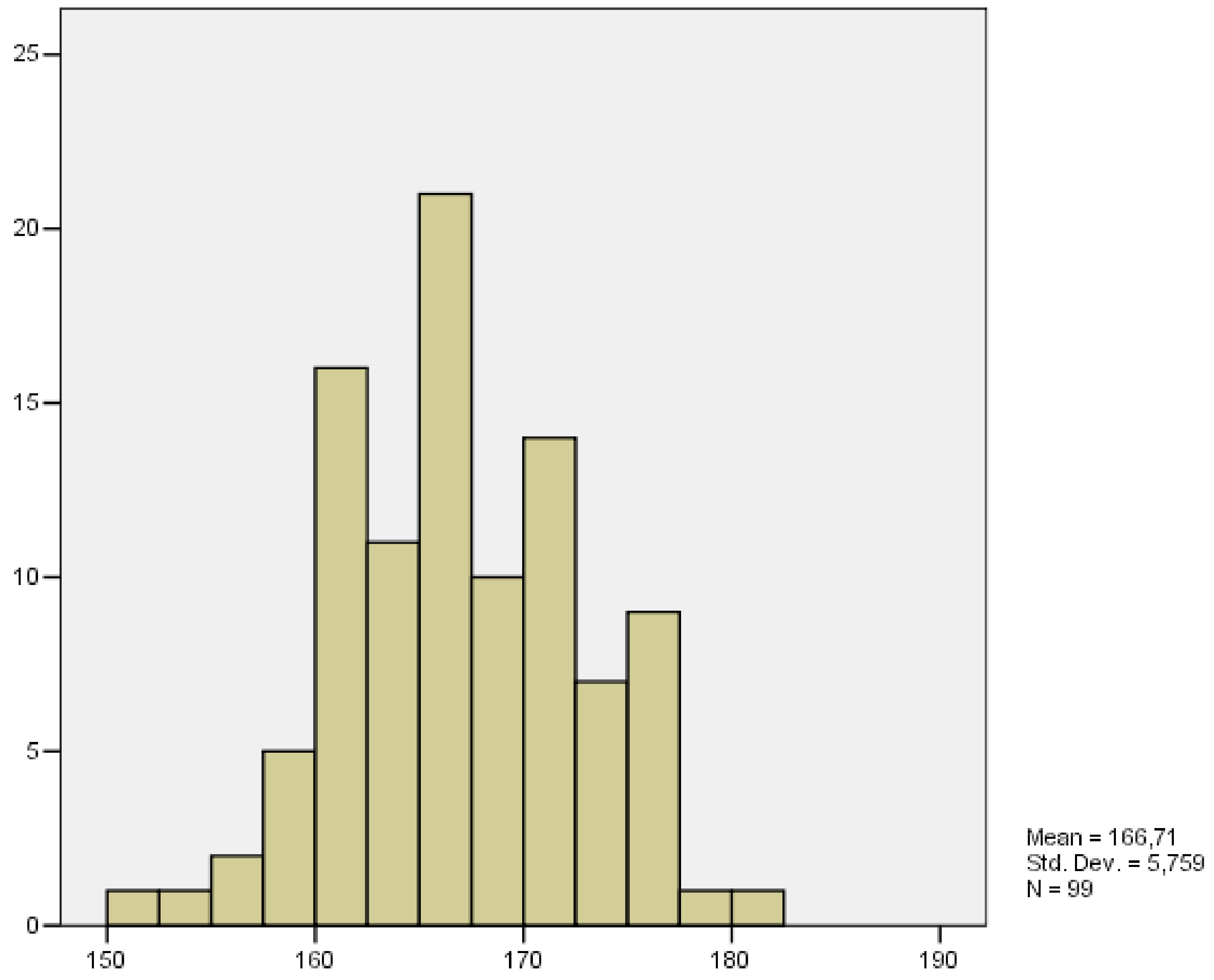
$(166 – 164)^2 + (170 – 164)^2 + … + (155 – 164)^2$

$2^2 + 6^2 + … + 9^2 = 560$

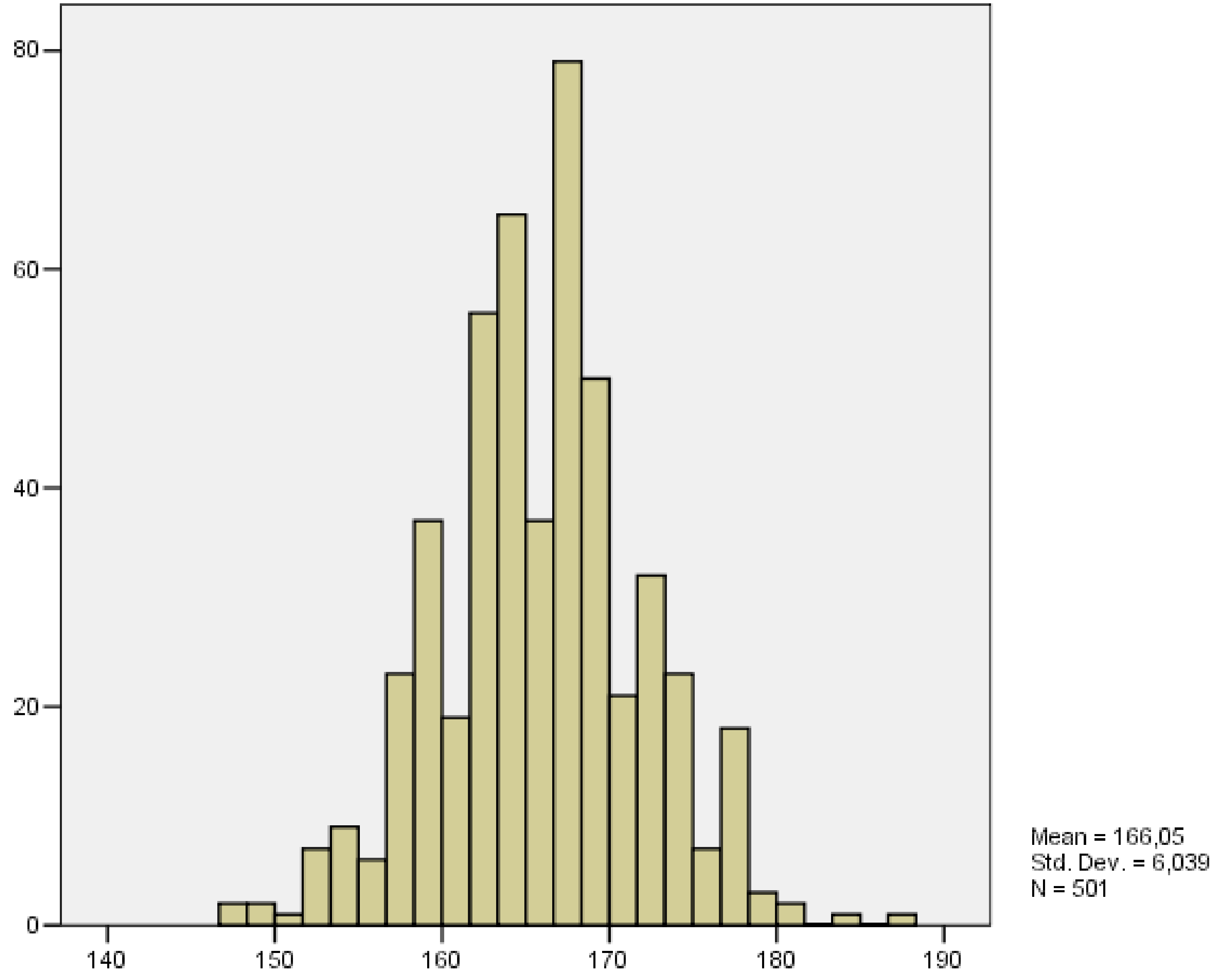
The square root of 560/9 = 7.9



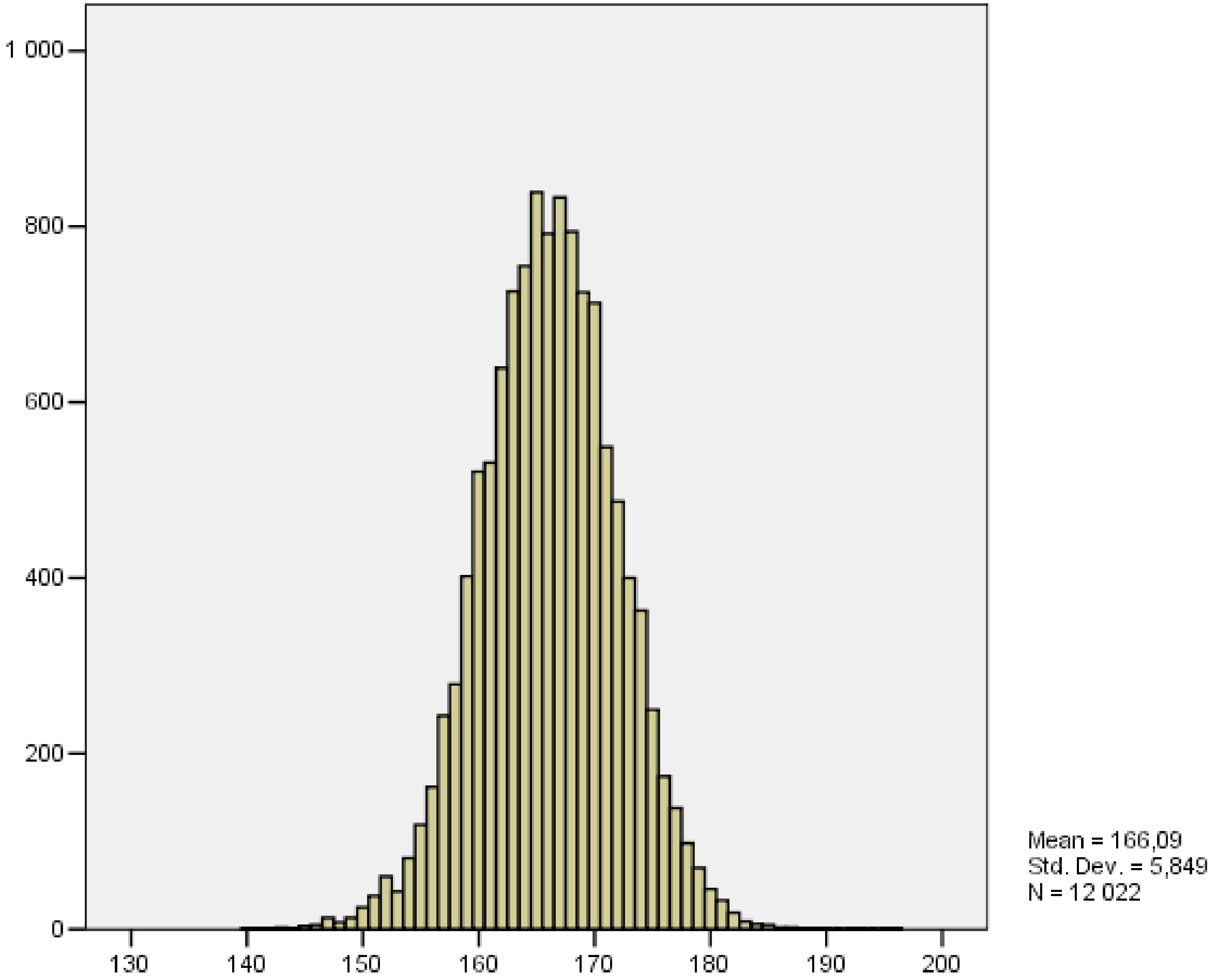
Height of women in Hordaland aged 40-45 years



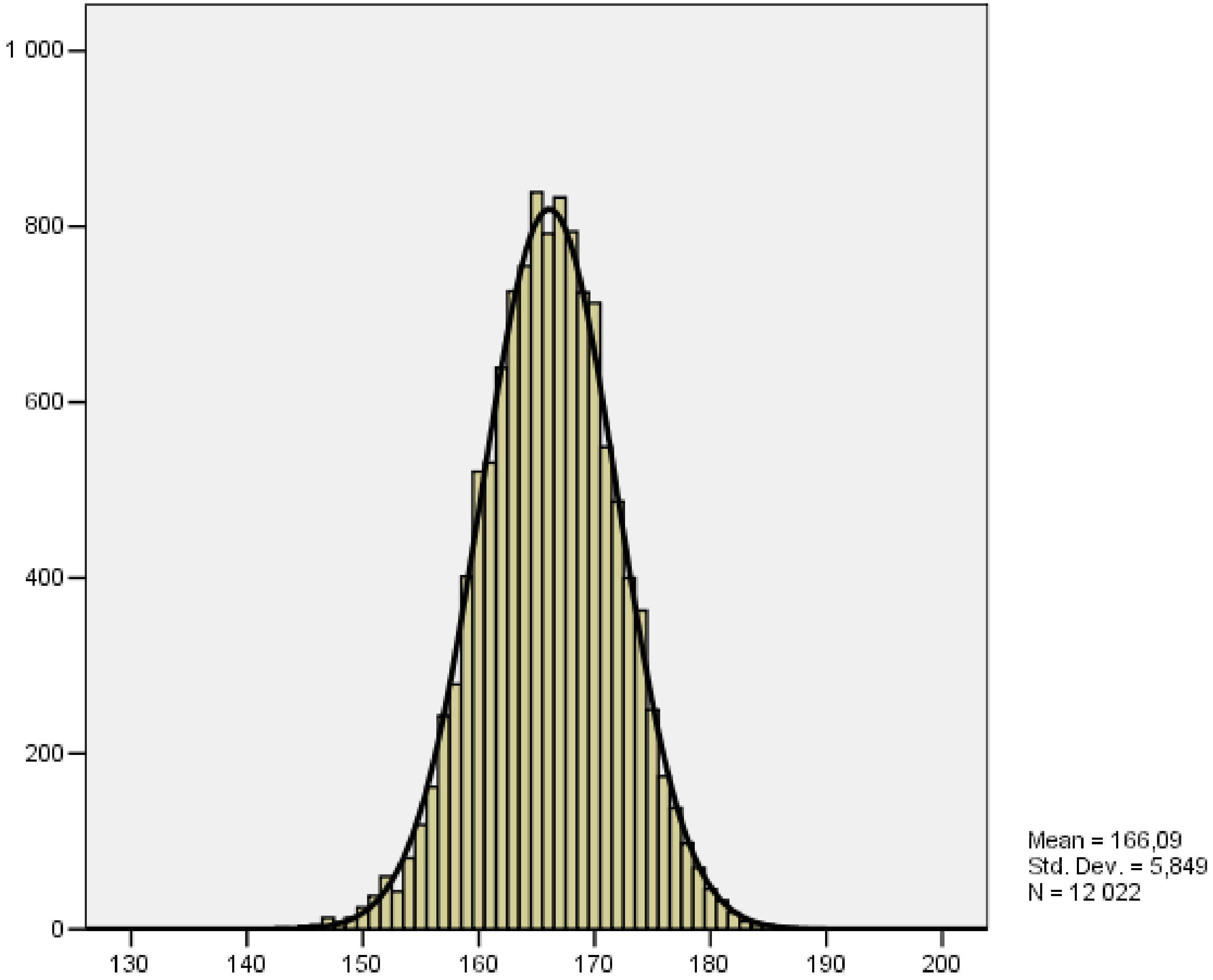
Height of women in Hordaland aged 40-45 years



Height of women in Hordaland aged 40-45 years



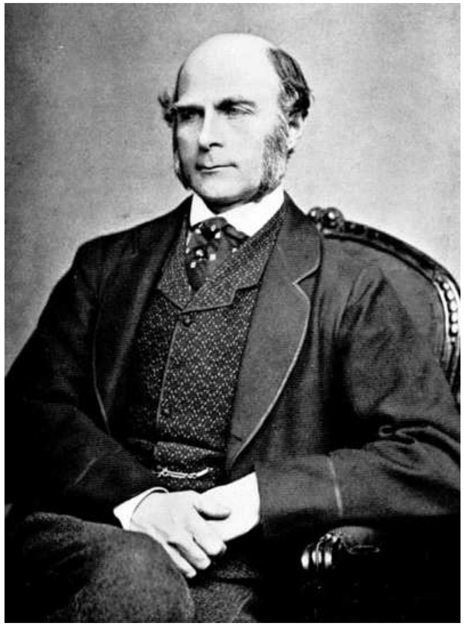
Height of women in Hordaland aged 40-45 years



Height of women in Hordaland aged 40-45 years

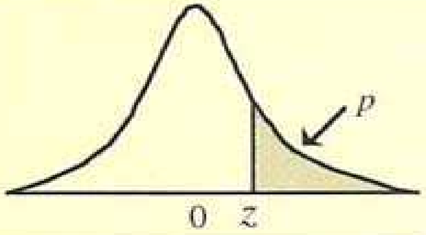
***I know of scarce anything so apt to impress the imagination as the wonderful form of cosmic order expressed by the "Law of Frequency of Error"***

Francis Galton 1822-1911



# Normal distribution

- Many biological phenomena are normally distributed when measured.
- The normal distribution is defined by a mathematical formula which is uniquely determined by the mean and the standard deviation.
- For a given mean and standard deviation it is possible to accurately calculate the percentage of observations that falls between to values.



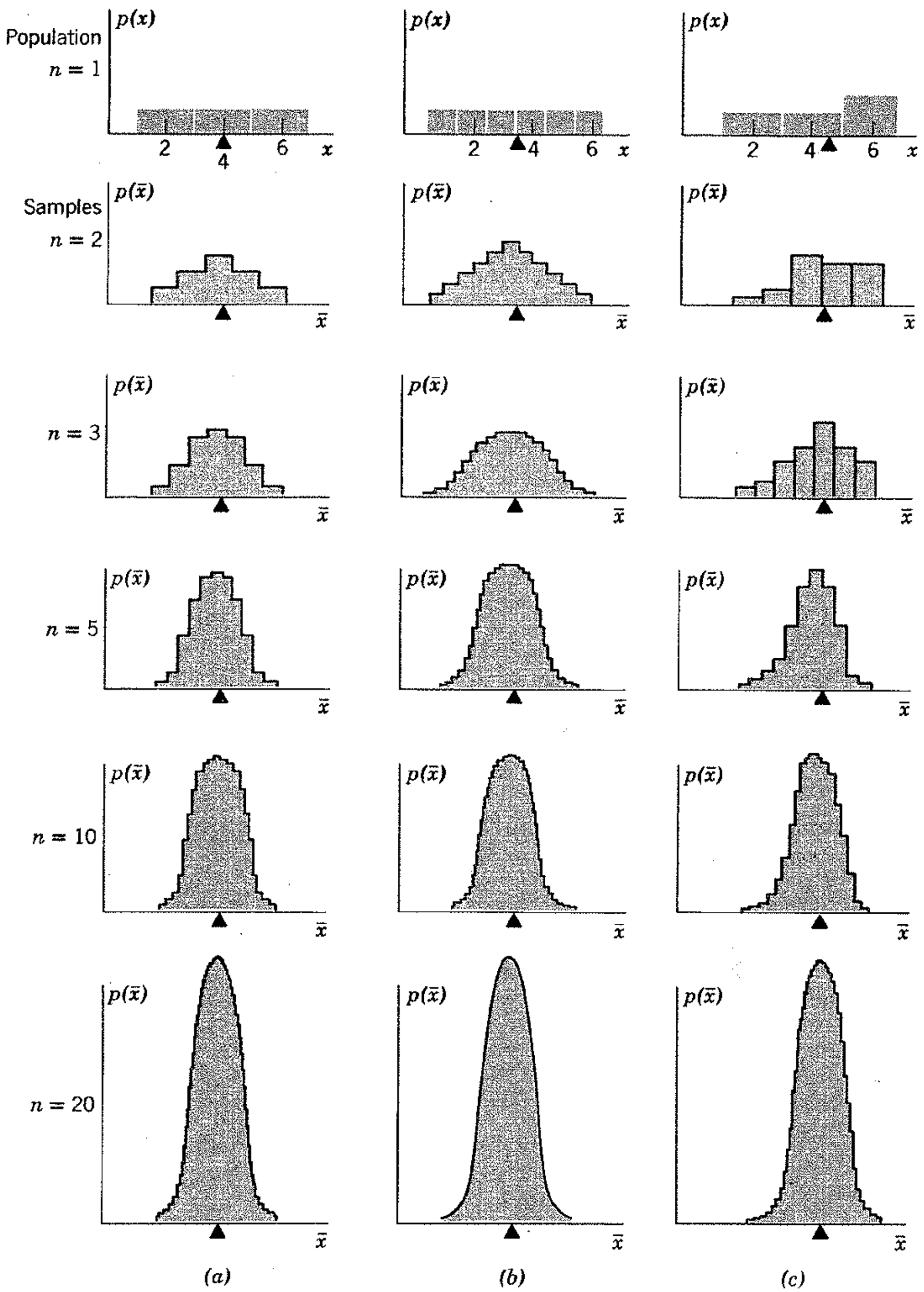
| | Second decimal place of *z* | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| *z* | .00 | .01 | .02 | .03 | .04 | .05 | .06 | .07 | .08 | .09 |
| 0.0 | .5000 | .4960 | .4920 | .4880 | .4840 | .4801 | .4761 | .4721 | .4681 | .4641 |
| 0.1 | .4602 | .4562 | .4522 | .4483 | .4443 | .4404 | .4364 | .4325 | .4286 | .4247 |
| 0.2 | .4207 | .4168 | .4129 | .4090 | .4052 | .4013 | .3974 | .3936 | .3897 | .3859 |
| 0.3 | .3821 | .3783 | .3745 | .3707 | .3669 | .3632 | .3594 | .3557 | .3520 | .3483 |
| 0.4 | .3446 | .3409 | .3372 | .3336 | .3300 | .3264 | .3228 | .3192 | .3156 | .3121 |
| 0.5 | .3085 | .3050 | .3015 | .2981 | .2946 | .2912 | .2877 | .2843 | .2810 | .2776 |
| 0.6 | .2743 | .2709 | .2676 | .2643 | .2611 | .2578 | .2546 | .2514 | .2483 | .2451 |
| 0.7 | .2420 | .2389 | .2358 | .2327 | .2297 | .2266 | .2236 | .2206 | .2177 | .2148 |
| 0.8 | .2119 | .2090 | .2061 | .2033 | .2005 | .1977 | .1949 | .1922 | .1894 | .1867 |
| 0.9 | .1841 | .1814 | .1788 | .1762 | .1736 | .1711 | .1685 | .1660 | .1635 | .1611 |
| 1.0 | .1587 | .1562 | .1539 | .1515 | .1492 | .1469 | .1446 | .1423 | .1401 | .1379 |
| 1.1 | .1357 | .1335 | .1314 | .1292 | .1271 | .1251 | .1230 | .1210 | .1190 | .1170 |
| 1.2 | .1151 | .1131 | .1112 | .1093 | .1075 | .1056 | .1038 | .1020 | .1003 | .0985 |
| 1.3 | .0968 | .0951 | .0934 | .0918 | .0901 | .0885 | .0869 | .0853 | .0838 | .0823 |
| 1.4 | .0808 | .0793 | .0778 | .0764 | .0749 | .0735 | .0721 | .0708 | .0694 | .0681 |
| 1.5 | .0668 | .0655 | .0643 | .0630 | .0618 | .0606 | .0594 | .0582 | .0571 | .0559 |
| 1.6 | .0548 | .0537 | .0526 | .0516 | .0505 | .0495 | .0485 | .0475 | .0465 | .0455 |
| 1.7 | .0446 | .0436 | .0427 | .0418 | .0409 | .0401 | .0392 | .0384 | .0375 | .0367 |
| 1.8 | .0359 | .0351 | .0344 | .0336 | .0329 | .0322 | .0314 | .0307 | .0301 | .0294 |
| 1.9 | .0287 | .0281 | .0274 | .0268 | .0262 | .0256 | .0250 | .0244 | .0239 | .0233 |
| 2.0 | .0228 | .0222 | .0217 | .0212 | .0207 | .0202 | .0197 | .0192 | .0188 | .0183 |
| 2.1 | .0179 | .0174 | .0170 | .0166 | .0162 | .0158 | .0154 | .0150 | .0146 | .0143 |
| 2.2 | .0139 | .0136 | .0132 | .0129 | .0125 | .0122 | .0119 | .0116 | .0113 | .0110 |
| 2.3 | .0107 | .0104 | .0102 | .0099 | .0096 | .0094 | .0091 | .0089 | .0087 | .0084 |
| 2.4 | .0082 | .0080 | .0078 | .0075 | .0073 | .0071 | .0069 | .0068 | .0066 | .0064 |
| 2.5 | .0062 | .0060 | .0059 | .0057 | .0055 | .0054 | .0052 | .0051 | .0049 | .0048 |
| 2.6 | .0047 | .0045 | .0044 | .0043 | .0041 | .0040 | .0039 | .0038 | .0037 | .0036 |
| 2.7 | .0035 | .0034 | .0033 | .0032 | .0031 | .0030 | .0029 | .0028 | .0027 | .0026 |

# Probability

The probability for an event (A) varies between 0 and 1 and is 0 if the event cannot happen, and 1 if it surely has to happen.

$P(A) = \lim_{n \to \infty} \#A/n$

«The probability of A is the number of times A occurs in a sample of size n, where n is extremely large (infinity).»


Population
$n = 1$
Samples
$n = 2$
$n = 3$
$n = 5$
$n = 10$
$n = 20$
$p(x)$
$p(\bar{x})$
2
4
6
$x$
$\bar{x}$
(a)
(b)
(c)

# Standard error

## Standard deviation of the mean (SEM – *se*$_m$)

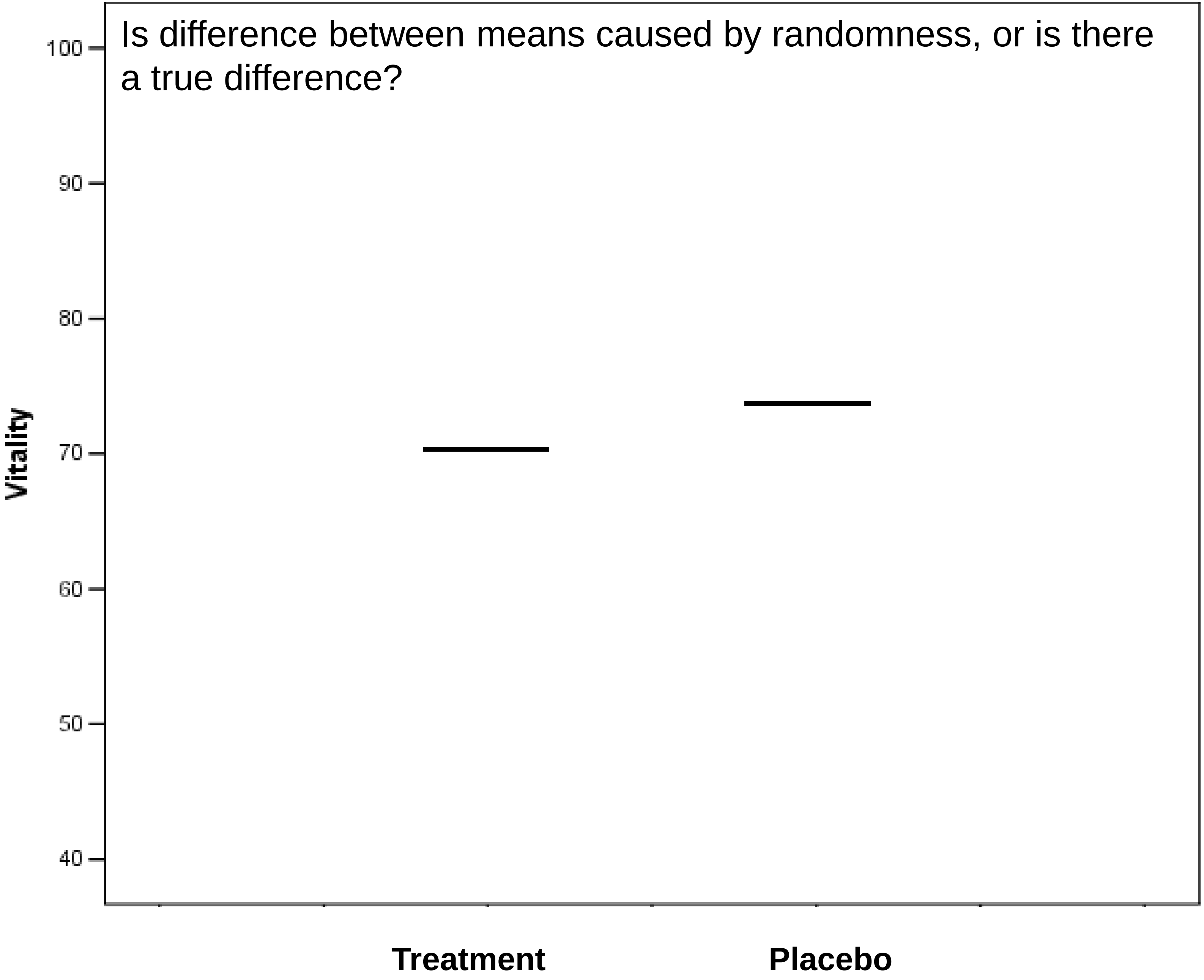
Standard deviation of a normal distribution of means

Standard error = $\frac{sd}{\sqrt{n}}$

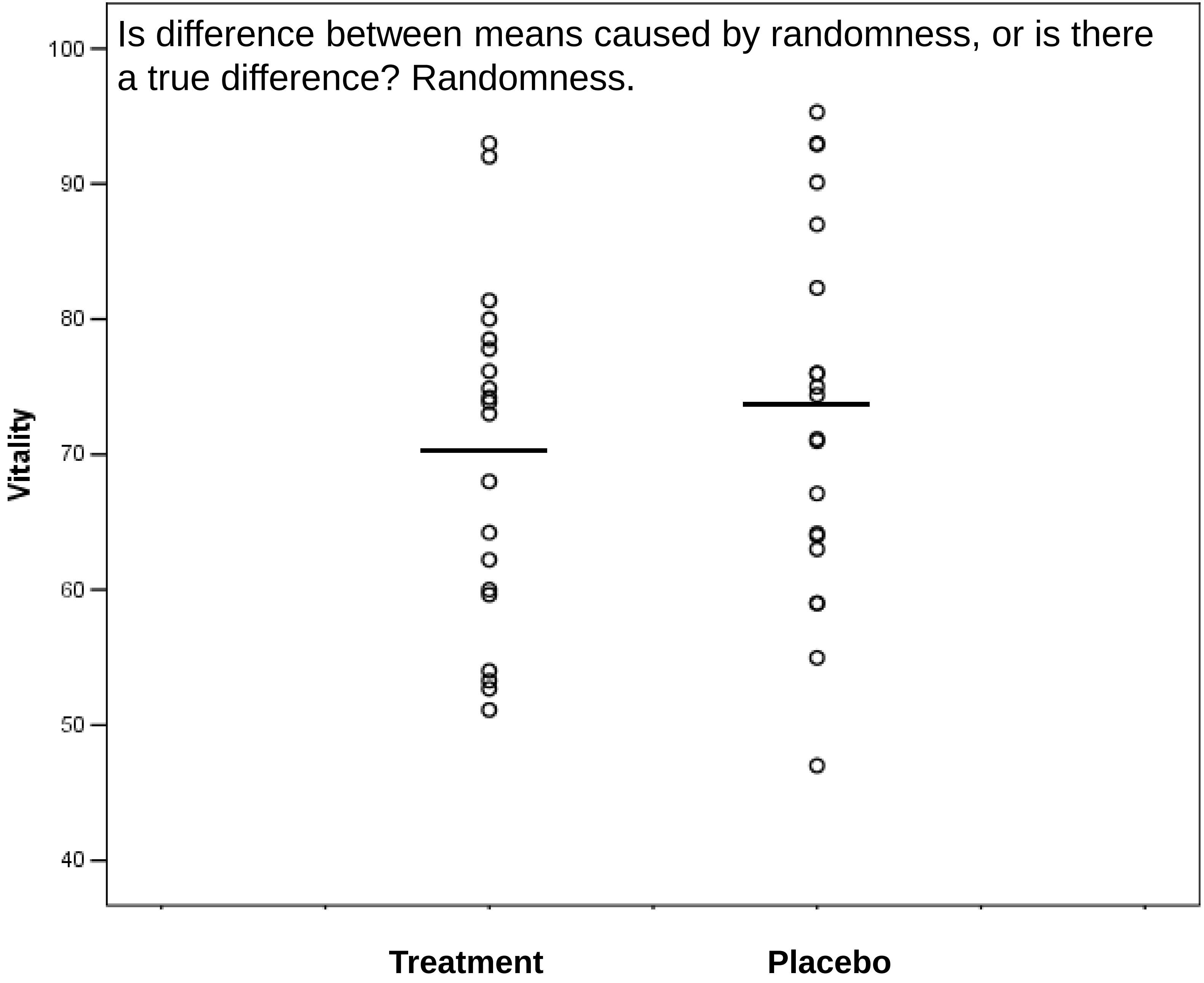
Based on the variation (sd) and sample size (n)

# Central limit theorem

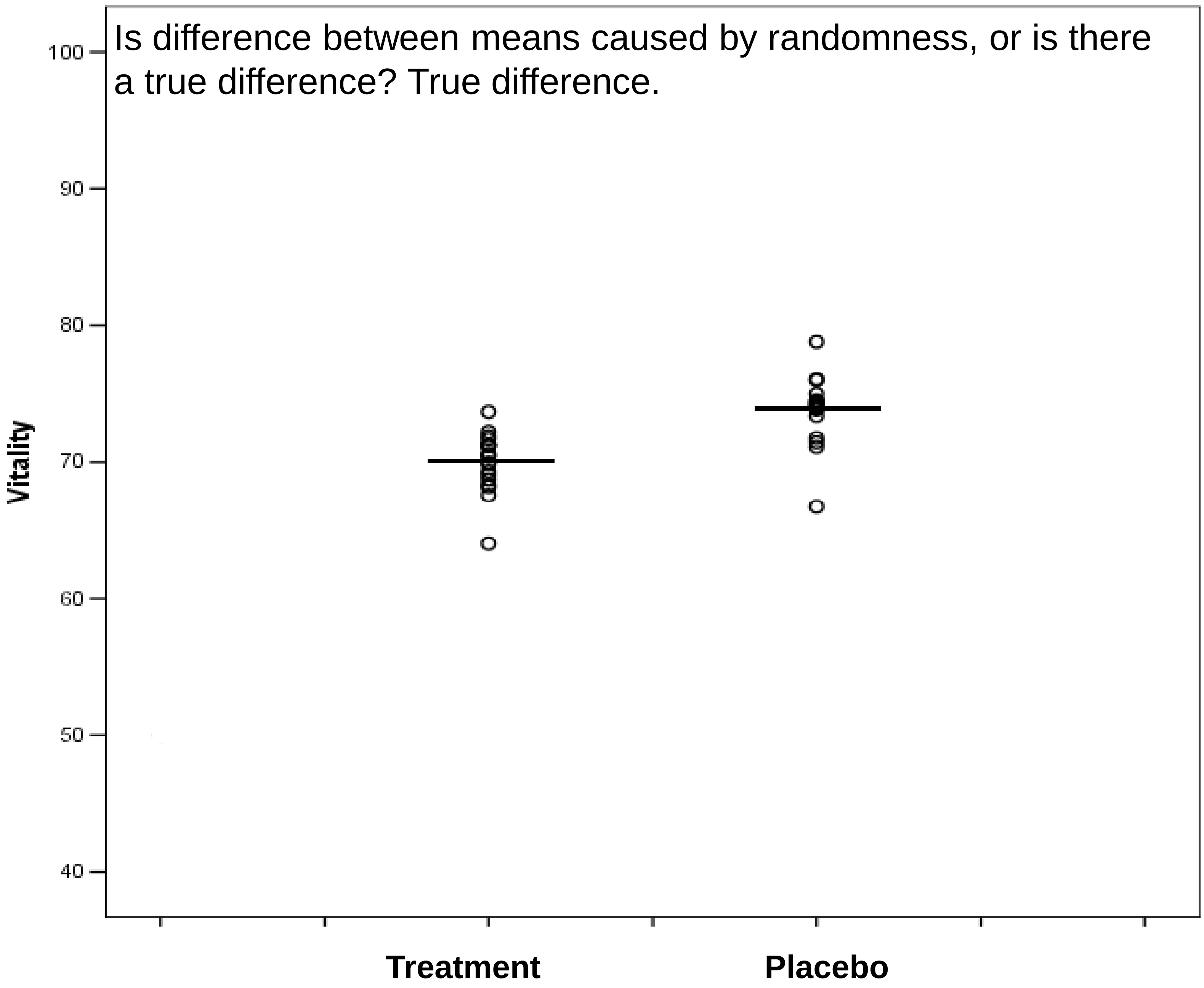
If you draw a series of n samples from a population, then the distribution of the means of these samples will approximate the normal distribution as n increases (independent of the distribution of the original population) with the same mean as the original population and with a standard deviation equals to the standard deviation from the original population divided by the square root of n (standard error).



Is difference between means caused by randomness, or is there a true difference?



Is difference between means caused by randomness, or is there a true difference? Randomness.



Is difference between means caused by randomness, or is there a true difference? True difference.

# Example of a statistical test I

- Problem: To what extent will a congenital heart malformation (non-cyanotic) influence the motor development of a child?
- A study is performed where 18 children with congenital heart malformation are followed for observation of when the children first were able to walk.
  The mean age in months was 14.1

# Congenital heart malformation and motor development – test II

- From large studies of normal children it has been shown that the mean age of children at their first steps alone is 13 months with a standard deviation of 1.75 months.

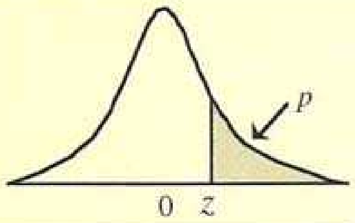
- Based on the problem in question a null-hypothesis ($H_0$) is defined:
  - Children with congenital heart malformation has the same mean age when they learn to walk
  - $H_0$: µ = 13 months

# Congenital heart malformation and motor development – test III

- Assume a normal distribution of the mean and assume that the $H_0$ is true, then:

- Calculate the probability that a random sample of 18 children has a mean “as far away from” 13 as 14.1. This is the p-value.

- If the p-value is small (less than 0.05), this means that what we have observed is unlikely to observe. Self-contradiction. Reject the underlying assumption.



|  | Second decimal place of *z* |  |  |  |  |  |  |  |  |  |
|---|---|---|---|---|---|---|---|---|---|---|
| *z* | .00 | .01 | .02 | .03 | .04 | .05 | .06 | .07 | .08 | .09 |
| 0.0 | .5000 | .4960 | .4920 | .4880 | .4840 | .4801 | .4761 | .4721 | .4681 | .4641 |
| 0.1 | .4602 | .4562 | .4522 | .4483 | .4443 | .4404 | .4364 | .4325 | .4286 | .4247 |
| 0.2 | .4207 | .4168 | .4129 | .4090 | .4052 | .4013 | .3974 | .3936 | .3897 | .3859 |
| 0.3 | .3821 | .3783 | .3745 | .3707 | .3669 | .3632 | .3594 | .3557 | .3520 | .3483 |
| 0.4 | .3446 | .3409 | .3372 | .3336 | .3300 | .3264 | .3228 | .3192 | .3156 | .3121 |
| 0.5 | .3085 | .3050 | .3015 | .2981 | .2946 | .2912 | .2877 | .2843 | .2810 | .2776 |
| 0.6 | .2743 | .2709 | .2676 | .2643 | .2611 | .2578 | .2546 | .2514 | .2483 | .2451 |
| 0.7 | .2420 | .2389 | .2358 | .2327 | .2297 | .2266 | .2236 | .2206 | .2177 | .2148 |
| 0.8 | .2119 | .2090 | .2061 | .2033 | .2005 | .1977 | .1949 | .1922 | .1894 | .1867 |
| 0.9 | .1841 | .1814 | .1788 | .1762 | .1736 | .1711 | .1685 | .1660 | .1635 | .1611 |
| 1.0 | .1587 | .1562 | .1539 | .1515 | .1492 | .1469 | .1446 | .1423 | .1401 | .1379 |
| 1.1 | .1357 | .1335 | .1314 | .1292 | .1271 | .1251 | .1230 | .1210 | .1190 | .1170 |
| 1.2 | .1151 | .1131 | .1112 | .1093 | .1075 | .1056 | .1038 | .1020 | .1003 | .0985 |
| 1.3 | .0968 | .0951 | .0934 | .0918 | .0901 | .0885 | .0869 | .0853 | .0838 | .0823 |
| 1.4 | .0808 | .0793 | .0778 | .0764 | .0749 | .0735 | .0721 | .0708 | .0694 | .0681 |
| 1.5 | .0668 | .0655 | .0643 | .0630 | .0618 | .0606 | .0594 | .0582 | .0571 | .0559 |
| 1.6 | .0548 | .0537 | .0526 | .0516 | .0505 | .0495 | .0485 | .0475 | .0465 | .0455 |
| 1.7 | .0446 | .0436 | .0427 | .0418 | .0409 | .0401 | .0392 | .0384 | .0375 | .0367 |
| 1.8 | .0359 | .0351 | .0344 | .0336 | .0329 | .0322 | .0314 | .0307 | .0301 | .0294 |
| 1.9 | .0287 | .0281 | .0274 | .0268 | .0262 | .0256 | .0250 | .0244 | .0239 | .0233 |
| 2.0 | .0228 | .0222 | .0217 | .0212 | .0207 | .0202 | .0197 | .0192 | .0188 | .0183 |
| 2.1 | .0179 | .0174 | .0170 | .0166 | .0162 | .0158 | .0154 | .0150 | .0146 | .0143 |
| 2.2 | .0139 | .0136 | .0132 | .0129 | .0125 | .0122 | .0119 | .0116 | .0113 | .0110 |
| 2.3 | .0107 | .0104 | .0102 | .0099 | .0096 | .0094 | .0091 | .0089 | .0087 | .0084 |
| 2.4 | .0082 | .0080 | .0078 | .0075 | .0073 | .0071 | .0069 | .0068 | .0066 | .0064 |
| 2.5 | .0062 | .0060 | .0059 | .0057 | .0055 | .0054 | .0052 | .0051 | .0049 | .0048 |
| 2.6 | .0047 | .0045 | .0044 | .0043 | .0041 | .0040 | .0039 | .0038 | .0037 | .0036 |
| 2.7 | .0035 | .0034 | .0033 | .0032 | .0031 | .0030 | .0029 | .0028 | .0027 | .0026 |