Practical Biostatistics

Department of Clinical Epidemiology, Biostatistics and Bioinformatics

AMC

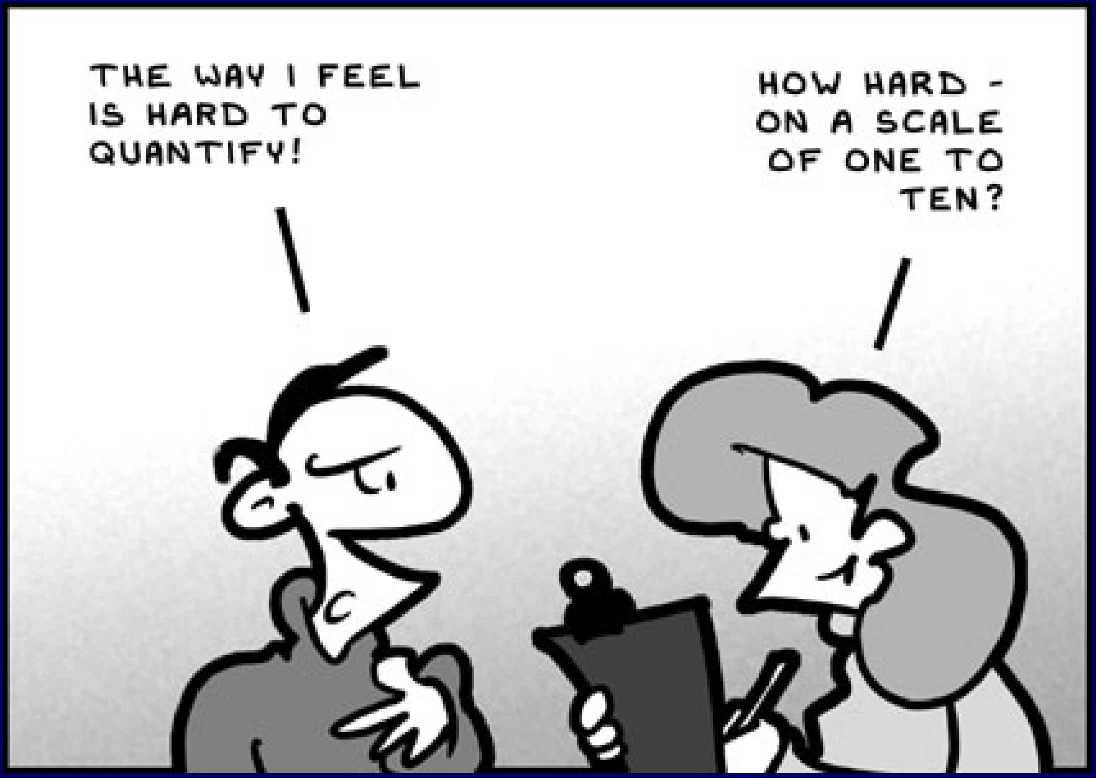
Principles of statistical testing

Review

DATA TYPE	measurement level	distribution	
numerical	*continuous*	normal	Mean: Variance, SD
		not-normal	Median: range, percentiles
	interval	normal	
		not-normal	
categorical	*binary*	balanced	
	nominal	balanced	
	ordinal	balanced	

Statistical tests (1)

Why do we use statistical tests?



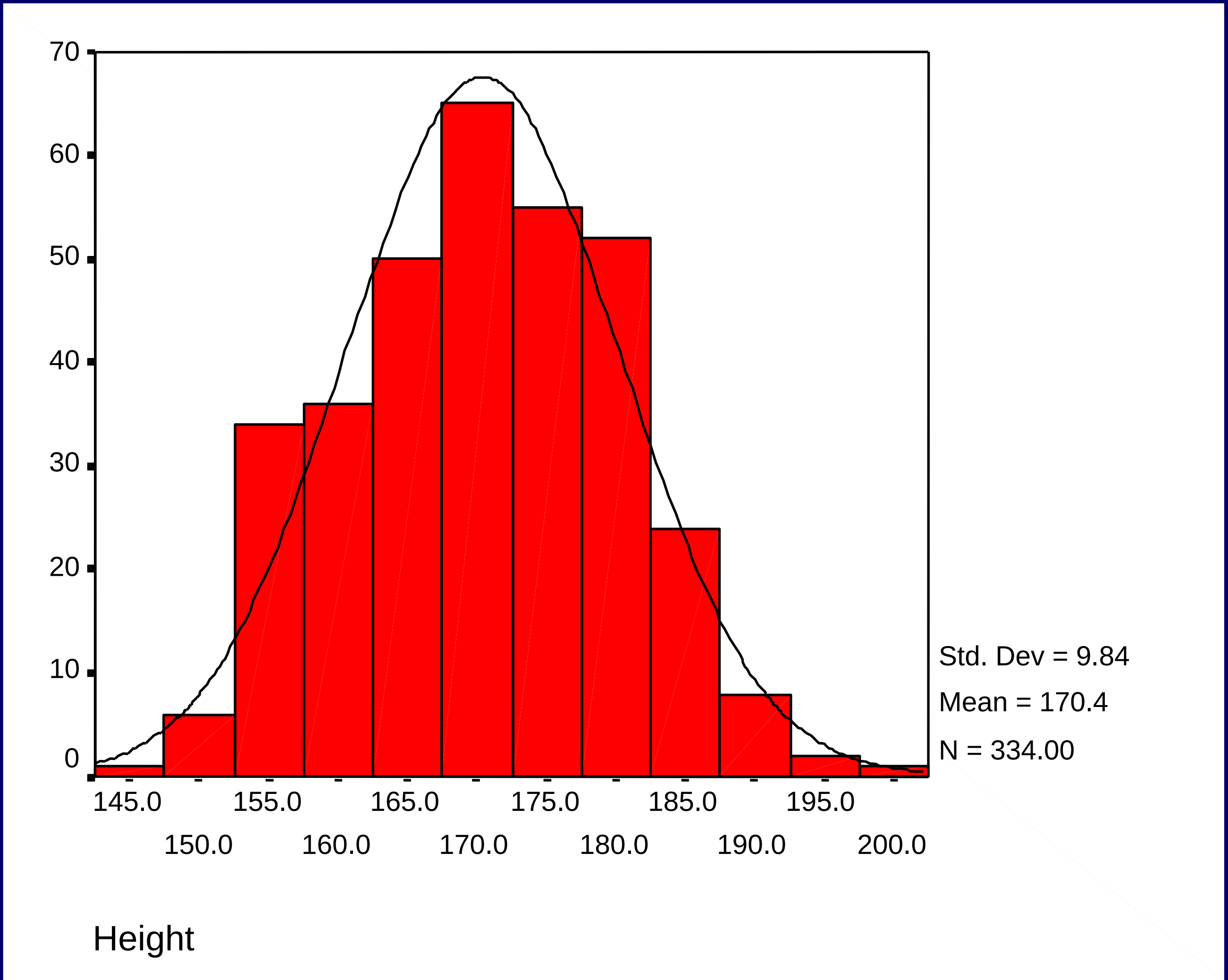
Statistical tests (2)

Statistics help provide us with information about a population, using sample information

Lecture 2

- ☒ Normal Distribution
- ☒ Sample vs. Population
- ☒ Standardizing the normal distribution
- ☒ Confidence Intervals & P-values

Normal distribution (1)



Normal distribution (2)

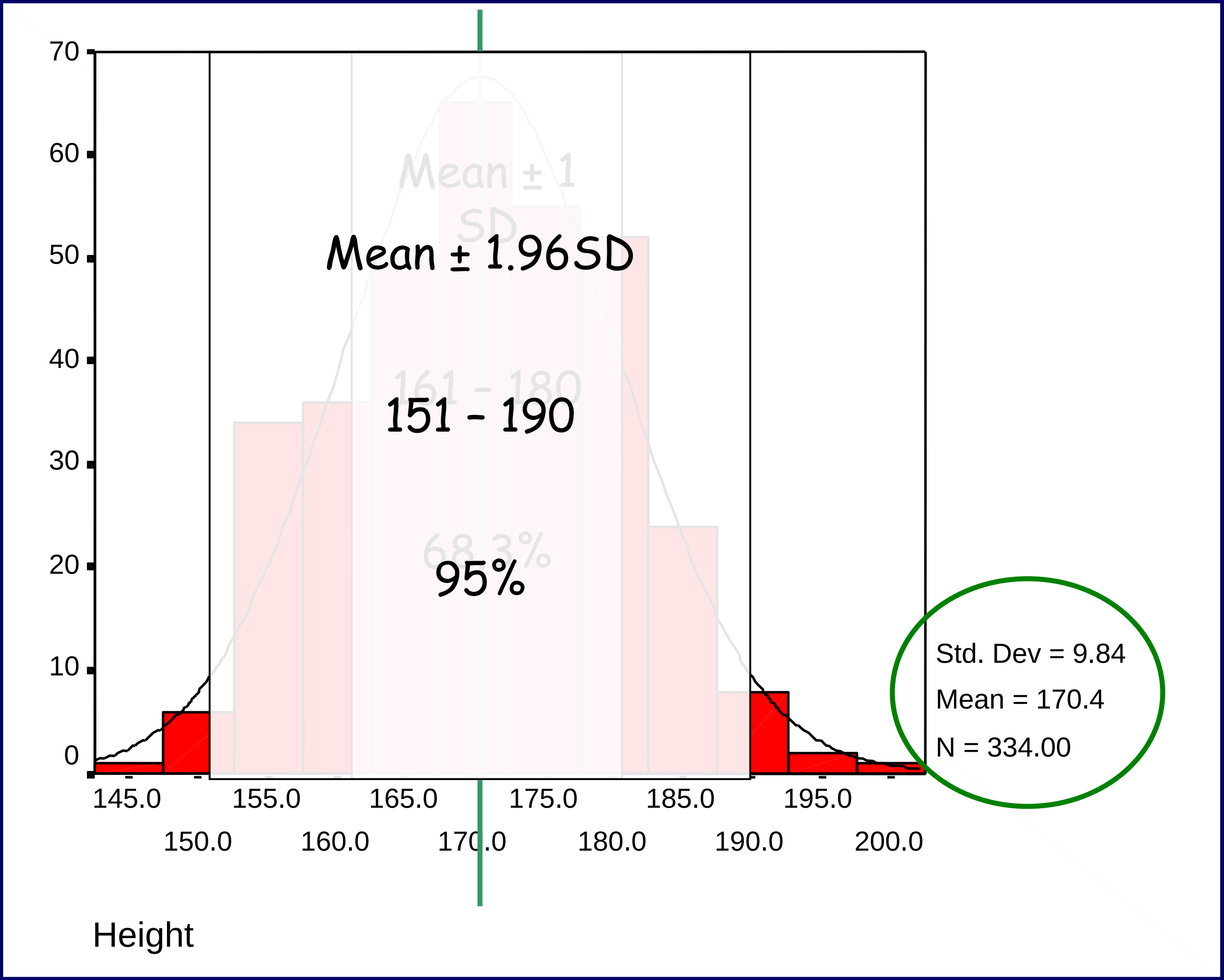
- Bell-shaped
- One top
- Symmetrical around its mean
- Mean and median are equal

Petrie chapter 7: theoretical distributions: the normal distribution

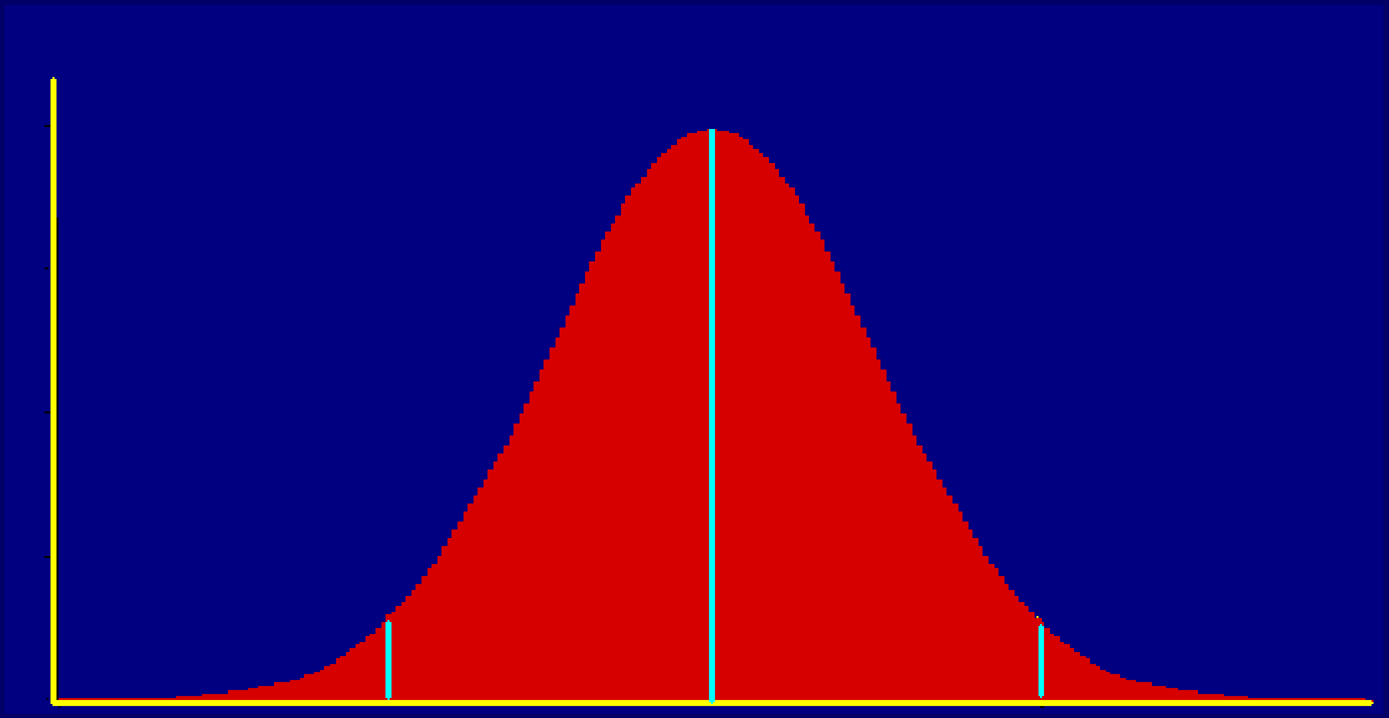
Normal distribution (3)

- mean ± 1 SD
  - 68,3% of all scores
- mean ± 1.96 SD
  - 95% of all scores
- mean ± 3 SD
  - 99,7% of all scores

Normal distribution



Normal Distribution



Probability that a Normally distributed random variable x, with mean μ, and SD σ, lies between Mean $\pm$ 1.96sd is 95%

Lecture 2

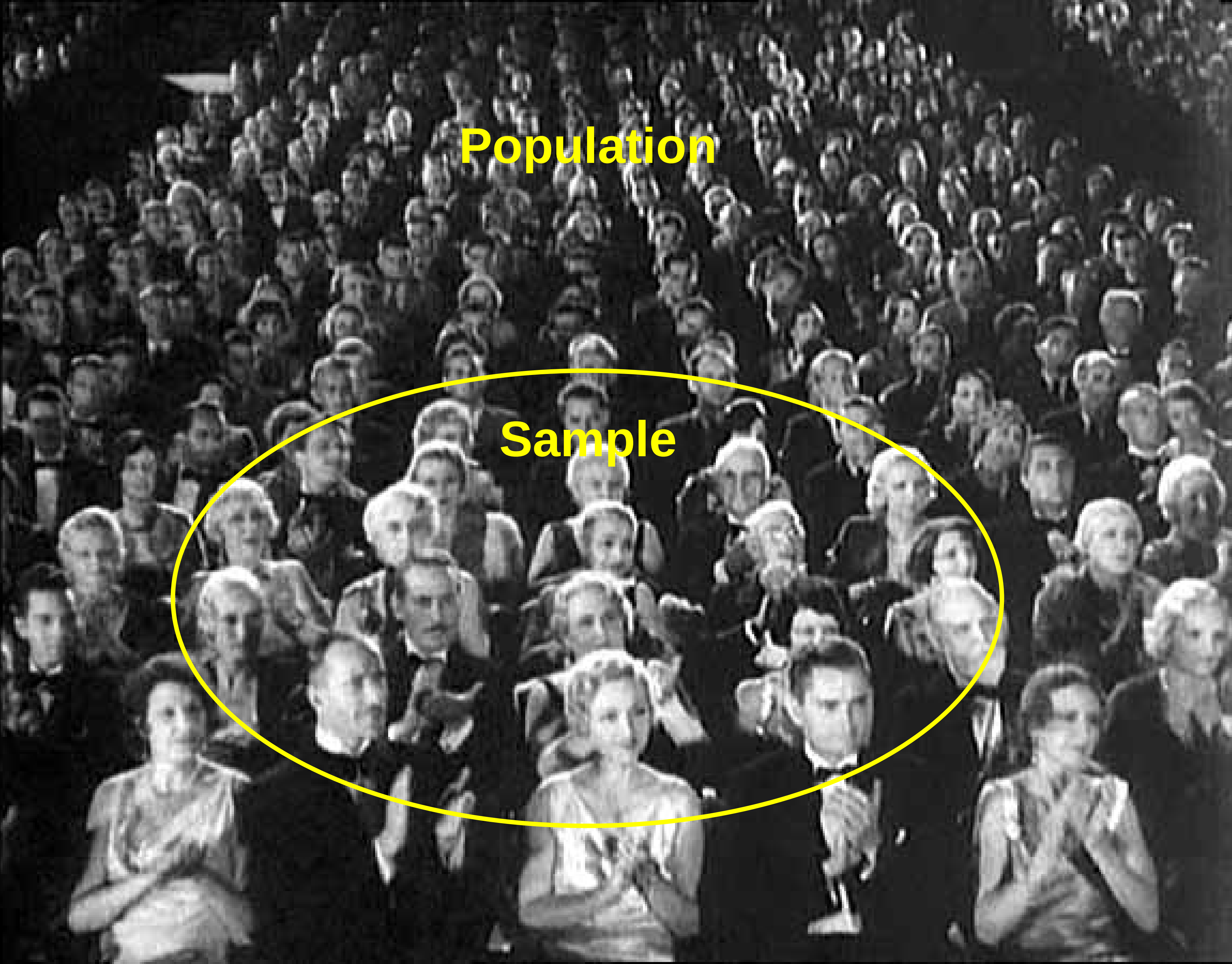
- ☒ Normal Distribution
- ☒ Sample vs. Population ←
- ☒ Standardizing the normal distribution
- ☒ Confidence Intervals & P-values

Sample vs. population (1)

- **Population** = an entire group of individuals in whom we are interested

- **Sample** = a representative subset of the population in whom we are interested

What is a representative sample?


Population
Sample

Sample vs. Population (2)

- ☒ Based on a sample of individuals, we estimate the value for the population

- ☒ Make a statement for a larger group based on a limited number of observations

Petrie chapter 10: sampling and sampling distributions

Sample vs population (3)

- ☒ **A Sample** is used to obtain knowledge about characteristics of a **Population**
- ☒ **A Sample statistic** is a value calculated from the observations from a sample
- ☒ **An inferential statistic:** is an estimate of the characteristics of the **population**

Estimate statistics (1)

- In Lecture 1 in a Normally distributed sample, we learnt:
  - Means
  - Standard deviations

- Lecture 2: we learn how well our sample data represents our population

- In second, third, fourth samples the sample statistics may differ

Estimate statistics (2)

We use inferential statistics to determine how well our sample statistic (eg. mean, SD) represents our population

Estimate Statistic (3)

- means -

- ☒ The mean in a sample can be taken as a ‘point estimate’ of the mean in the population

- ☒ How good is my estimation?
  - Depending on the spread in the sample
  - Depending on the size of the sample
  - Fluctuation between samples (sampling error)

Estimate statistics (4)

- ☒ The **Standard Error** (SE) is the standard deviation (SD) of the sampling distribution of a estimate statistic

- ☒ Spread of the means of various samples round the population mean is S**tandard Error (SE)**

Estimate Statistics (5)

- Standard error -

Calculating the Standard error (SE):

- ☒ Take repeated samples of the same size from a population
- ☒ Calculate the mean of each sample
- ☒ Mean of these sample means, estimates the mean of the population
- ☒ Sampling means follow a Normal distribution

Standard error (1)

Estimate standard error (mean):

$$SE(mean) = SD/\sqrt{N}$$

SE example

- serum albumin -

- ☒ N = 216
- ☒ Mean of the sample = 34,5 g/l
- ☒ SD = 5,8 g/l
  - – Sample mean = estimate of population mean

- ☒ SE = 5,8 / √ 216 = 0,4 g/l

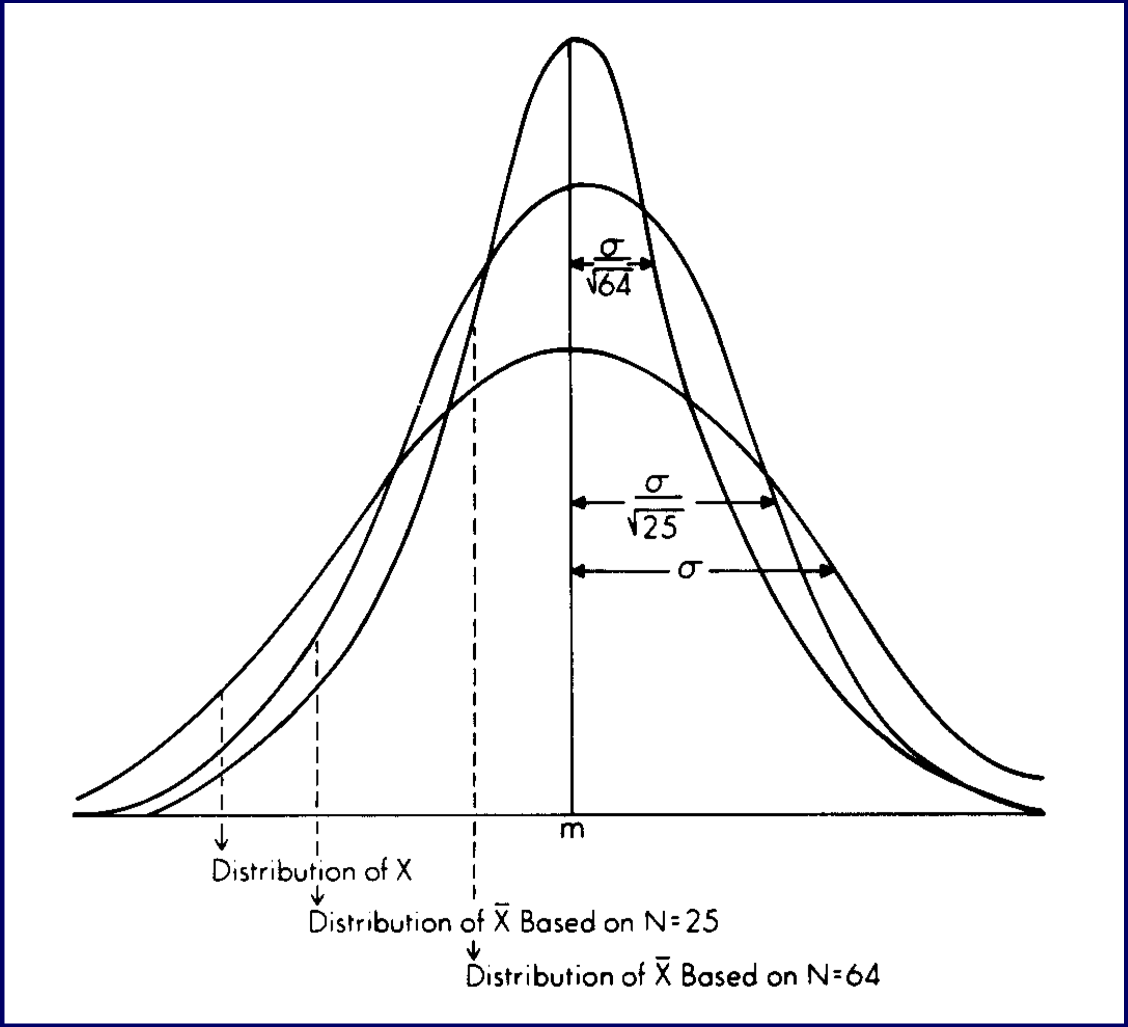
SD and SE

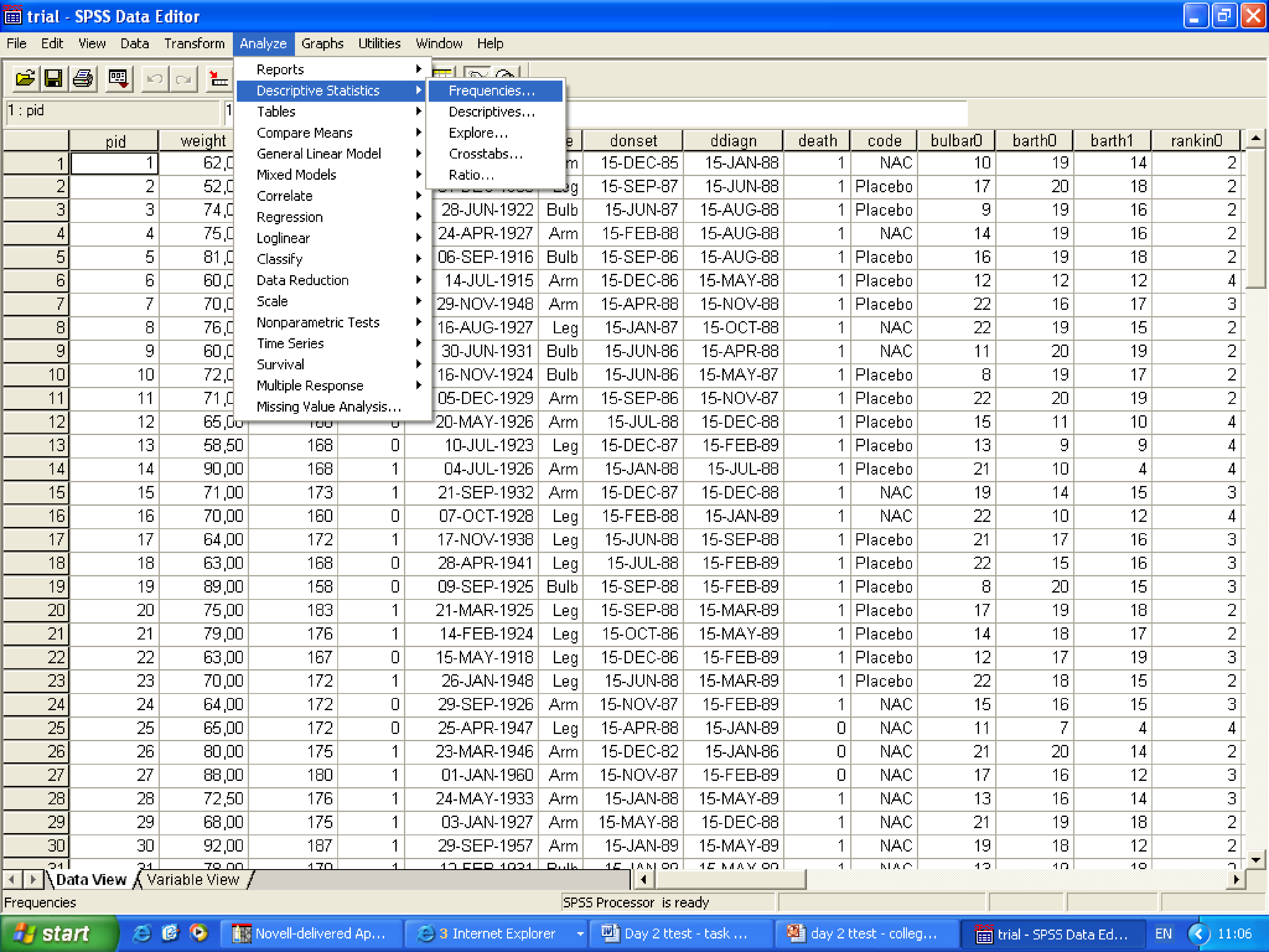
- ☒ **SD:** the variation in the data values of a sample, and should be quoted if you wish to show the spread in the sample data

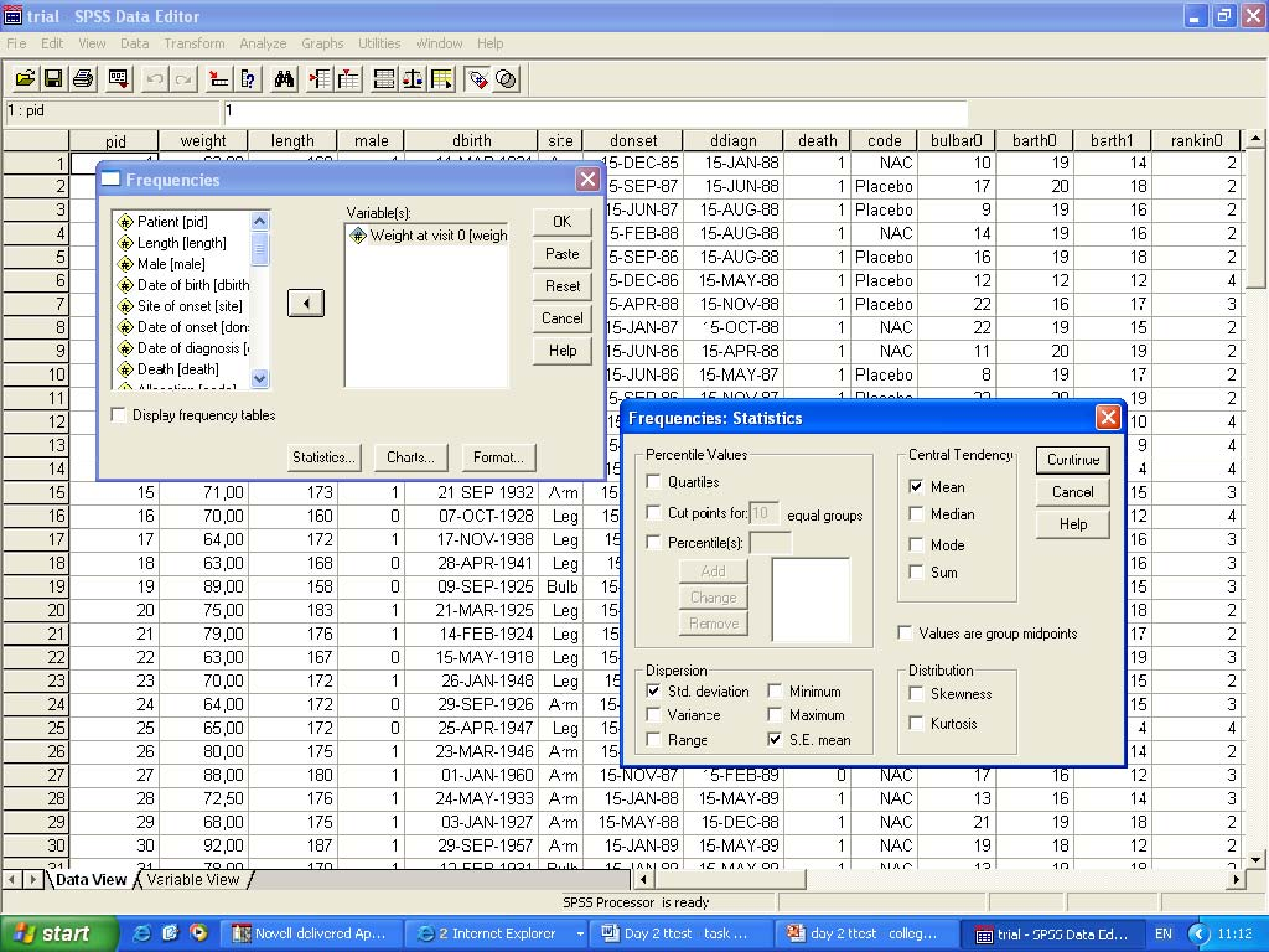
- ☒ **SE:** precision of the sample mean (as estimate of the population mean) and should be quoted if you are interested in the prediction of the sample mean representing the population mean

Sample size and SE

The larger the sample size, the better the prediction of the mean of the population




trial - SPSS Data Editor
File Edit View Data Transform Analyze Graphs Utilities Window Help
Reports
Descriptive Statistics
Tables
Compare Means
General Linear Model
Mixed Models
Correlate
Regression
Loglinear
Classify
Data Reduction
Scale
Nonparametric Tests
Time Series
Survival
Multiple Response
Missing Value Analysis...
Frequencies...
Descriptives...
Explore...
Crosstabs...
Ratio...
1 : pid
pid weight donset ddiagn death code bulbar0 barth0 barth1 rankin0
1 1 62,0 m 15-DEC-85 15-JAN-88 1 NAC 10 19 14 2
2 2 52,0 eg 15-SEP-87 15-JUN-88 1 Placebo 17 20 18 2
3 3 74,0 28-JUN-1922 Bulb 15-JUN-87 15-AUG-88 1 Placebo 9 19 16 2
4 4 75,0 24-APR-1927 Arm 15-FEB-88 15-AUG-88 1 NAC 14 19 16 2
5 5 81,0 06-SEP-1916 Bulb 15-SEP-86 15-AUG-88 1 Placebo 16 19 18 2
6 6 60,0 14-JUL-1915 Arm 15-DEC-86 15-MAY-88 1 Placebo 12 12 12 4
7 7 70,0 29-NOV-1948 Arm 15-APR-88 15-NOV-88 1 Placebo 22 16 17 3
8 8 76,0 16-AUG-1927 Leg 15-JAN-87 15-OCT-88 1 NAC 22 19 15 2
9 9 60,0 30-JUN-1931 Bulb 15-JUN-86 15-APR-88 1 NAC 11 20 19 2
10 10 72,0 16-NOV-1924 Bulb 15-JUN-86 15-MAY-87 1 Placebo 8 19 17 2
11 11 71,0 05-DEC-1929 Arm 15-SEP-86 15-NOV-87 1 Placebo 22 20 19 2
12 12 65,00 20-MAY-1926 Arm 15-JUL-88 15-DEC-88 1 Placebo 15 11 10 4
13 13 58,50 168 0 10-JUL-1923 Leg 15-DEC-87 15-FEB-89 1 Placebo 13 9 9 4
14 14 90,00 168 1 04-JUL-1926 Arm 15-JAN-88 15-JUL-88 1 Placebo 21 10 4 4
15 15 71,00 173 1 21-SEP-1932 Arm 15-DEC-87 15-DEC-88 1 NAC 19 14 15 3
16 16 70,00 160 0 07-OCT-1928 Leg 15-FEB-88 15-JAN-89 1 NAC 22 10 12 4
17 17 64,00 172 1 17-NOV-1938 Leg 15-JUN-88 15-SEP-88 1 Placebo 21 17 16 3
18 18 63,00 168 0 28-APR-1941 Leg 15-JUL-88 15-FEB-89 1 Placebo 22 15 16 3
19 19 89,00 158 0 09-SEP-1925 Bulb 15-SEP-88 15-FEB-89 1 Placebo 8 20 15 3
20 20 75,00 183 1 21-MAR-1925 Leg 15-SEP-88 15-MAR-89 1 Placebo 17 19 18 2
21 21 79,00 176 1 14-FEB-1924 Leg 15-OCT-86 15-MAY-89 1 Placebo 14 18 17 2
22 22 63,00 167 0 15-MAY-1918 Leg 15-DEC-86 15-FEB-89 1 Placebo 12 17 19 3
23 23 70,00 172 1 26-JAN-1948 Leg 15-JUN-88 15-MAR-89 1 Placebo 22 18 15 2
24 24 64,00 172 0 29-SEP-1926 Arm 15-NOV-87 15-FEB-89 1 NAC 15 16 15 3
25 25 65,00 172 0 25-APR-1947 Leg 15-APR-88 15-JAN-89 0 NAC 11 7 4 4
26 26 80,00 175 1 23-MAR-1946 Arm 15-DEC-82 15-JAN-86 0 NAC 21 20 14 2
27 27 88,00 180 1 01-JAN-1960 Arm 15-NOV-87 15-FEB-89 0 NAC 17 16 12 3
28 28 72,50 176 1 24-MAY-1933 Arm 15-JAN-88 15-MAY-89 1 NAC 13 16 14 3
29 29 68,00 175 1 03-JAN-1927 Arm 15-MAY-88 15-DEC-88 1 NAC 21 19 18 2
30 30 92,00 187 1 29-SEP-1957 Arm 15-JAN-89 15-MAY-89 1 NAC 19 18 12 2
Data View
Variable View
Frequencies
SPSS Processor is ready
start
Novell-delivered Ap...
3 Internet Explorer
Day 2 ttest - task ...
day 2 ttest - colleg...
trial - SPSS Data Ed...
EN
11:06


trial - SPSS Data Editor
File Edit View Data Transform Analyze Graphs Utilities Window Help
1 : pid
pid weight length male dbirth site donset ddiagn death code bulbar0 barth0 barth1 rankin0
Frequencies
Patient [pid]
Length [length]
Male [male]
Date of birth [dbirth
Site of onset [site]
Date of onset [don
Date of diagnosis [
Death [death]
Variable(s):
Weight at visit 0 [weigh
OK
Paste
Reset
Cancel
Help
Display frequency tables
Statistics...
Charts...
Format...
Frequencies: Statistics
Percentile Values
Quartiles
Cut points for: 10 equal groups
Percentile(s):
Add
Change
Remove
Central Tendency
Mean
Median
Mode
Sum
Continue
Cancel
Help
Values are group midpoints
Dispersion
Std. deviation
Minimum
Variance
Maximum
Range
S.E. mean
Distribution
Skewness
Kurtosis
Data View
Variable View
SPSS Processor is ready
start
Novell-delivered Ap...
2 Internet Explorer
Day 2 ttest - task ...
day 2 ttest - colleg...
trial - SPSS Data Ed...
EN
11:12

Standard error (2)
- important concept-

- ☒ SE can be used for two statistical calculations
  - ☒ calculation of the confidence interval for the mean (CI)
  - ☒ calculation of the 'critical value' required in testing (p-value)

Standard Error (3):

NB: The SE can also be used for ordinal, nominal and binary data, but is not expressed in means and SD, but in proportions.

Lecture 2

- ☒ Normal Distribution
- ☒ Sample vs. Population
- ☒ Standardizing the normal distribution
- ☒ Confidence Intervals & P-values

SD and SE

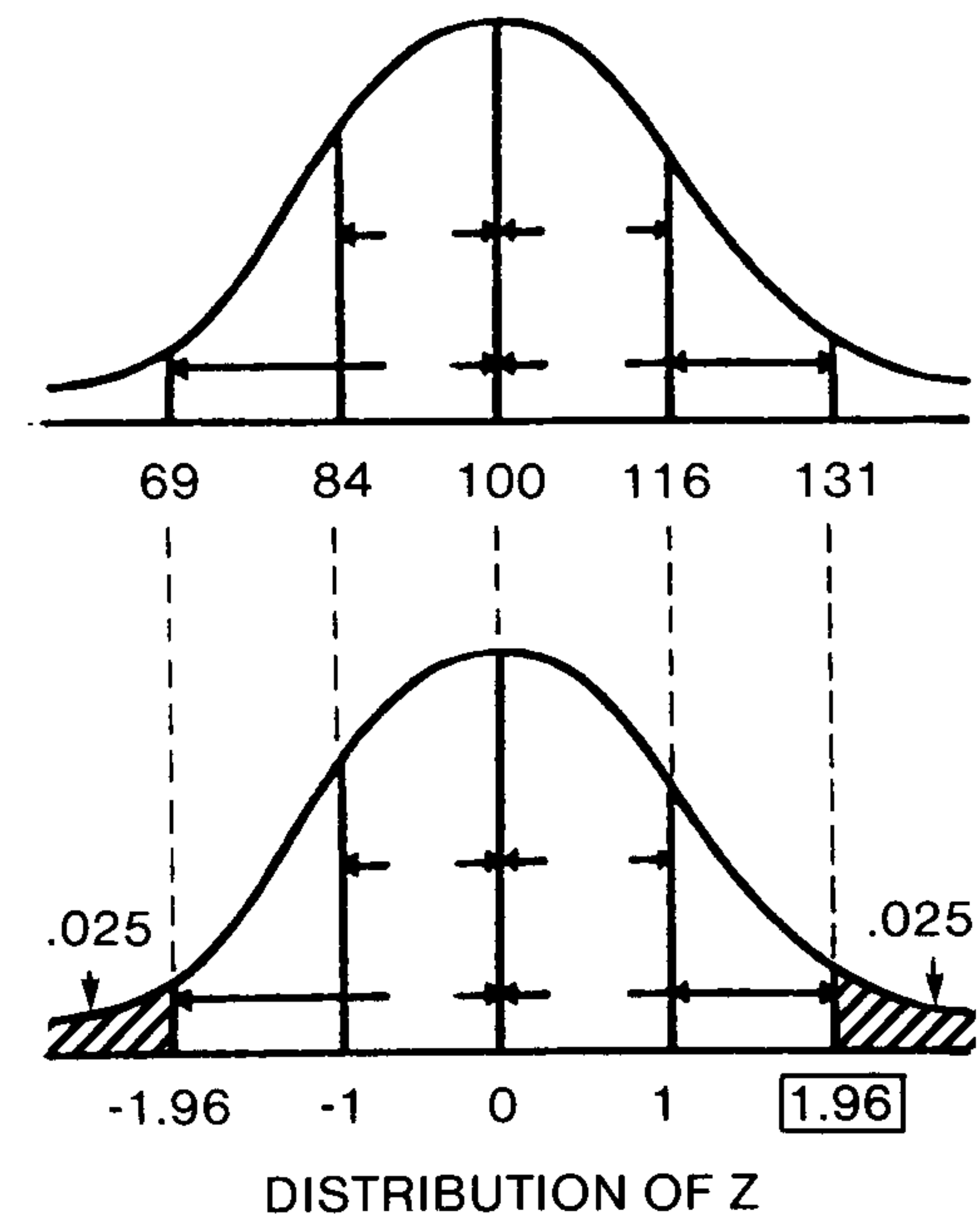
- ☒ **Mean and SD:** calculated mean of sample and spread in the data values of a sample

- ☒ **Mean and SE:** estimated mean of population and precision of this estimation sample mean (as estimate of the population mean)

- ☒ Both follow normal distribution

Standardize Normal distribution
- z-distribution -

- There are many different normal distributions
- But only one z-distribution where:
  - mean = 0
  - SD = 1
- Every Normal distribution can be transposed into a z-distribution



Normally distributed weight data

Z-distribution

Standardized Normal distribution
- z-distribution -

- z-score on a *sample* level indicates the number of SD's a score is above or below the *sample* mean

- z-score on *population* level indicates the number of SE's a sample mean is above or below the *population* mean

How?

- standard scores on *sample* level-

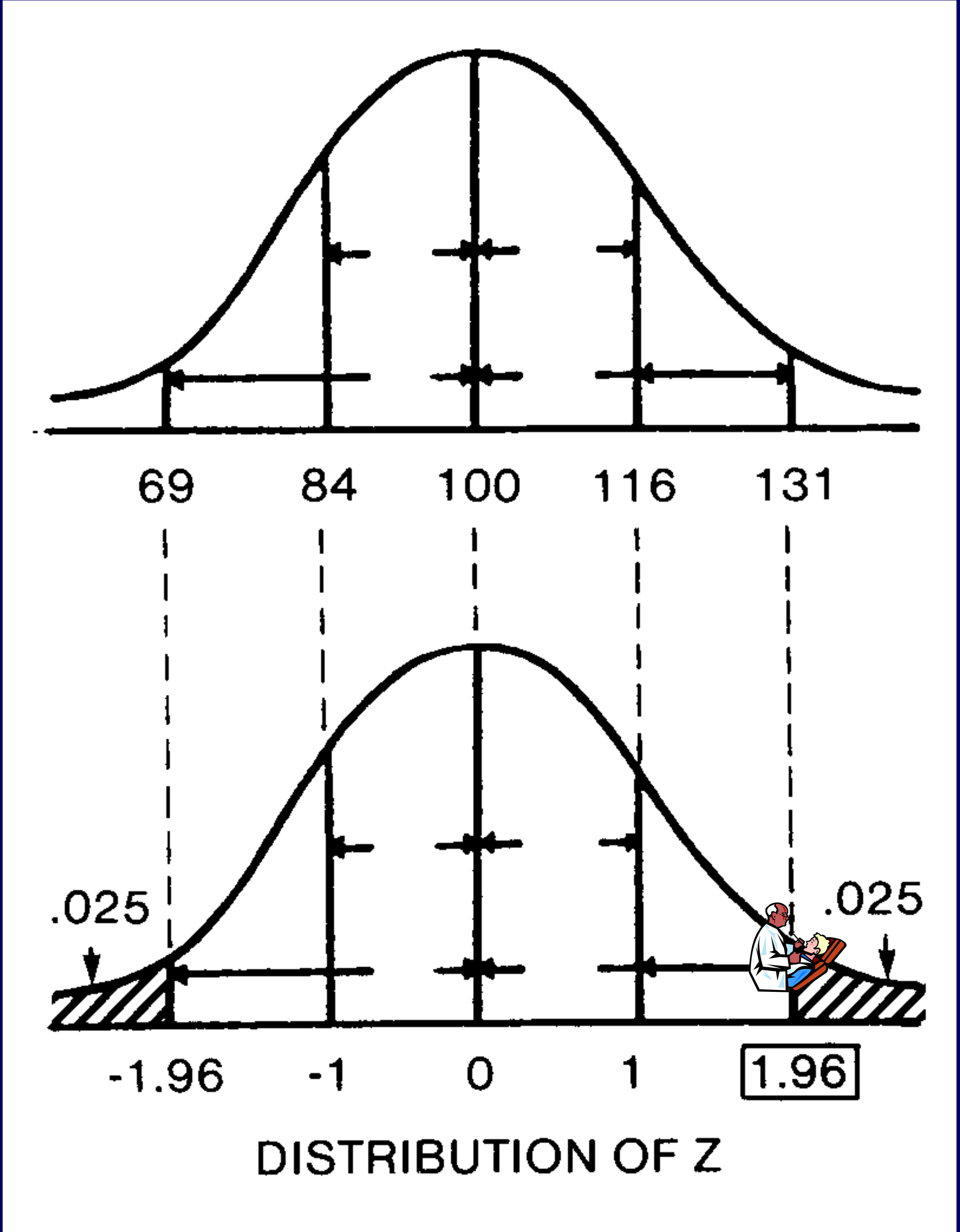
- Difference between *an individual score* and the *sample mean*
  - mean = 100; SD = 15,8; N = 25
  - Patients' score is 131 (difference = 31 points)
- z-score of the patient = 31 / 15,8 = 1,96
- score of this patient is 1,96 SD above the sample mean

why?

- standard scores on *sample* level-

- Using a z-score, we can calculate the probability of a specific individual score related to the sample mean
- Probability of z => 1,96 or higher = 2,5%
- remark:
  - z <= -1,96 is 2,5% as well

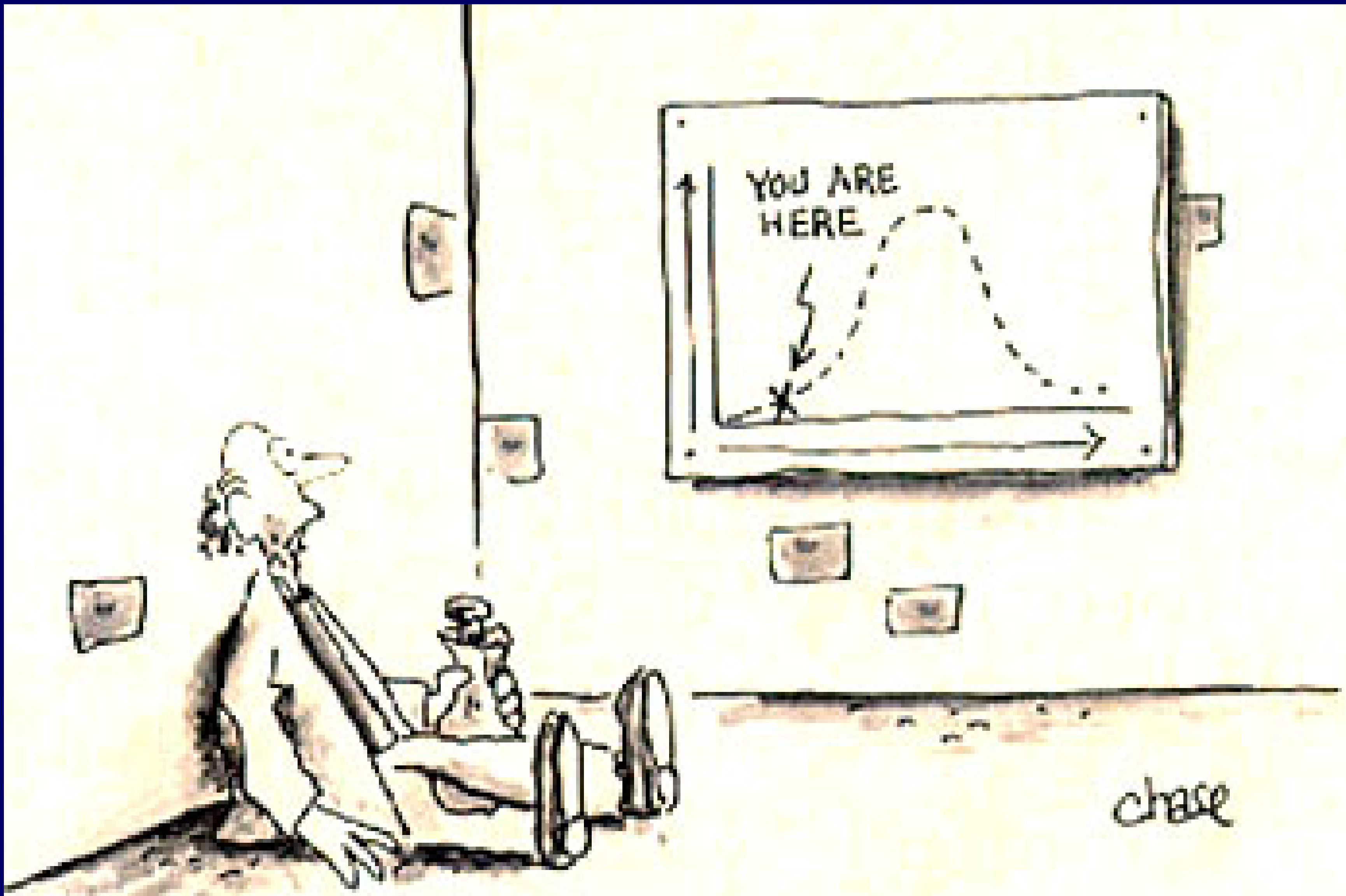
Sample

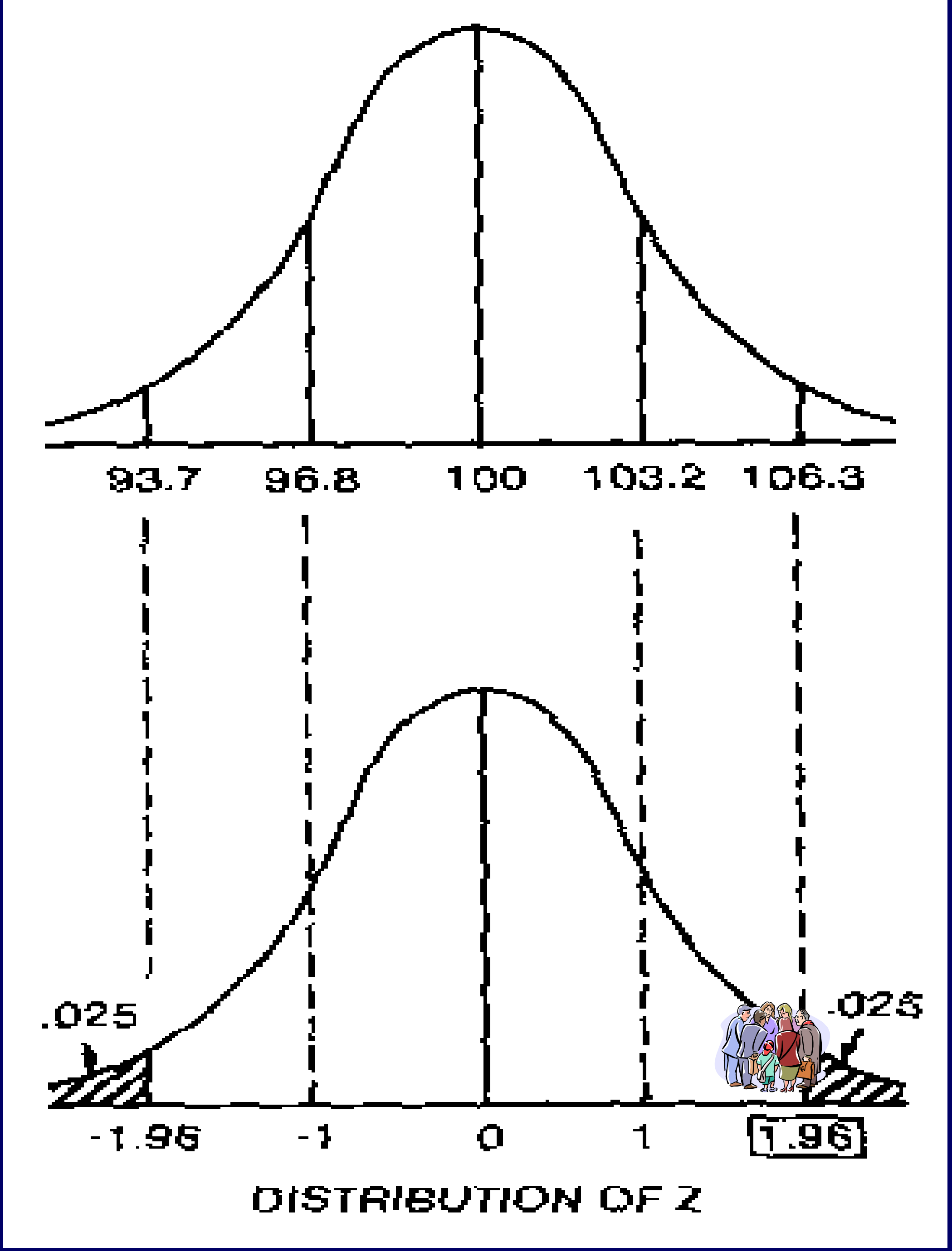


How?

- standard scores on *population* level -

- Difference between *sample mean* and *population mean*
  - Population mean = 100
  - Sample mean= 106.3; SD= 16; N= 25
  - SE = 16 / √ 25 = 3,2
- z-score (sample mean) = 6,3 / 3,2 = 1,96
- Sample mean is 1,96 SE above the population mean


YOU ARE HERE
chase



population

Lecture 2

- ☒ Normal Distribution
- ☒ Sample vs. Population
- ☒ Standardizing the normal distribution
- ☒ Confidence Intervals & P-values

95% confidence interval of population mean
- what -

- ☒ `Interval estimate’: range of values around point estimate

- ☒ 95% confident that the true population mean lies within these limits

95% confidence interval
- how -

- ☒ Sample mean is estimate of population mean
- ☒ SE = precision of the sample mean
- ☒ Mean ± 1.96 SE → 95% of sample means

= 95% Confidence Interval (95% CI)

95% confidence interval
- example -

- ☒ Cohort (N = 600) cirrhosis patients
- ☒ Mean = 34,46 g/l; SE = 0,397 g/l
- ☒ 95% CI = mean ± **1,96** * SE

  – 95% CI population mean =
  – 34,46 ± 1,96 * 0,397
  – 33,68 tot 35,24 g/l

Size of confidence interval

- ☒ 90%, 95%, or 99% CI
  - more confidence that range includes population value (high number) means wider intervals
- ☒ SD (dispersion in the sample)
  - studies with much sample variation have wider intervals (less clinical information)
- ☒ N (sample size)
  - studies with large sample sizes have smaller intervals (more clinical information)

95% confidence interval

- estimation -

- ☒ Cohort (N = 600) cirrhosis patients
- ☒ Mean = 34,46 g/l; SE = 0,397 g/l
- ☒ 95% CI = 33,68 tot 35,24 g/l
- ☒ 5% probability population mean is outside 95% CI

- ☒ Normal population: < 33 g/l

- ☒ Probability that value of cirrhosis patients is like that of the normal population < 5% ($P < 0.05$)

Testing with 95% CI

- estimation of the treatment interval -

- ☒ Calculate 95% confidence interval
- ☒ Define the null hypothesis
  - H_0: no difference between groups
  - $H_{\underline{1}}$: groups are different
- ☒ If the null hypothesis value is within 95% CI → accept H_0
- ☒ If the null hypothesis value is outside 95% CI → reject H_0, in favour of H_1: there is a difference

Interpreted the p-value

- ☒ $p < 0,05 \Rightarrow$ reject H_0
- ☒ $p \geq 0,05 \Rightarrow$ accept H_0

- ☒ $p < 0,05 \Rightarrow$ statistically significant
- ☒ $p \geq 0,05 \Rightarrow$ not statistically significant

p-value (1)

Is the probability of obtaining our results, or something more extreme, **assuming** that the null hypothesis is true!

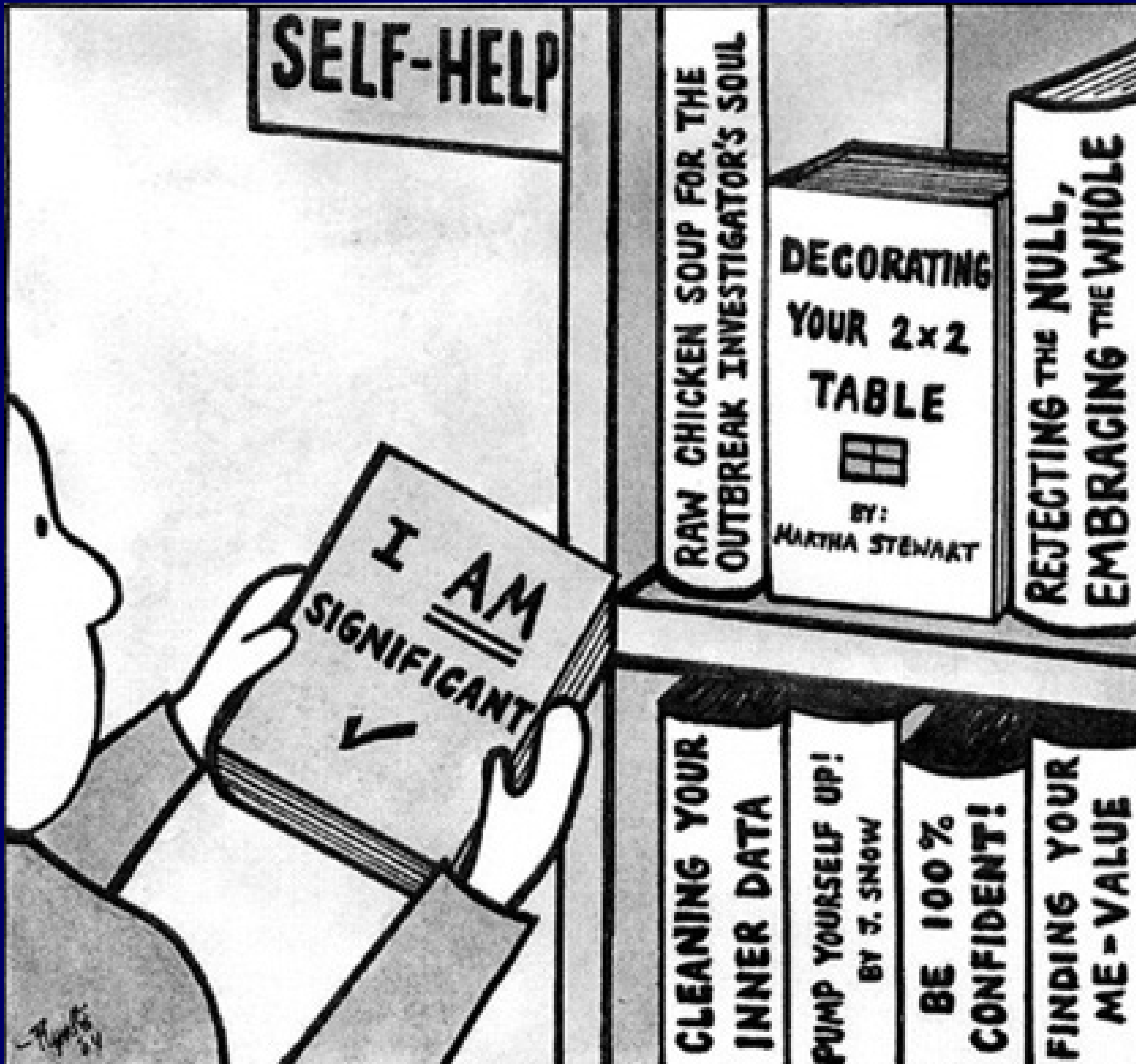
p-value (2)

When the p-value is less than 0.05 (or 0.01) there is only a small probability that the result would have occurred when the null hypothesis was true

→ In that case, we *reject the null hypothesis* (H_0) in favour of the alternative hypothesis (H_1)

Is it, or is it not significant?

- ☒ Usually not a good question
  - How large is the treatment effect?
  - Reader wants to know the uncertainty around a treatment effect, they need concrete information


SELF-HELP
RAW CHICKEN SOUP FOR THE OUTBREAK INVESTIGATOR'S SOUL
DECORATING YOUR 2x2 TABLE
BY: MARTHA STEWART
REJECTING THE NULL, EMBRACING THE WHOLE
I AM SIGNIFICANT
CLEANING YOUR INNER DATA
PUMP YOURSELF UP!
BY J. SNOW
BE 100% CONFIDENT!
FINDING YOUR ME-VALUE

Summary

- ☒ Sample vs. population
- ☒ SD vs. SE(mean)
- ☒ Point estimates (population mean) vs. interval estimates (95% CI of the mean)
- ☒ Hypothesis testing using:
  - p-values (reject or accept H_0)
  - CI (more clinical information)