

Statistics

Dr.sc. Iliriana Miftari
University of Prishtina



Master Study Program Urban Agriculture Teaching Material/ 2021

Chapter 1

A sample = a set of measurements

The sample size = the number of units
where measurements
are taken

A population = a set of hypothetical
measurements

Chapter 2

We have 2 kinds of data:

- 1) Qualitative or categorical: e.g: red, blue, green etc. tall, medium, short.
- 2) Numerical or measurement data: height, weight, shoe sizes etc.

We have 2 kinds of numerical data:

Discrete: can only take a finite or countable infinite values.

Continuous: can take all values on a continuous scale or interval. Counts are discrete, but they can be treated as continuous if they can take a wide range of values.

A variable can be qualitative, continuous or discrete.

For each unit we observe the variable.

It gets the value corresponding to what we observe on that unit.

When we have collected data, we often want to describe the dataset.

Categorical data

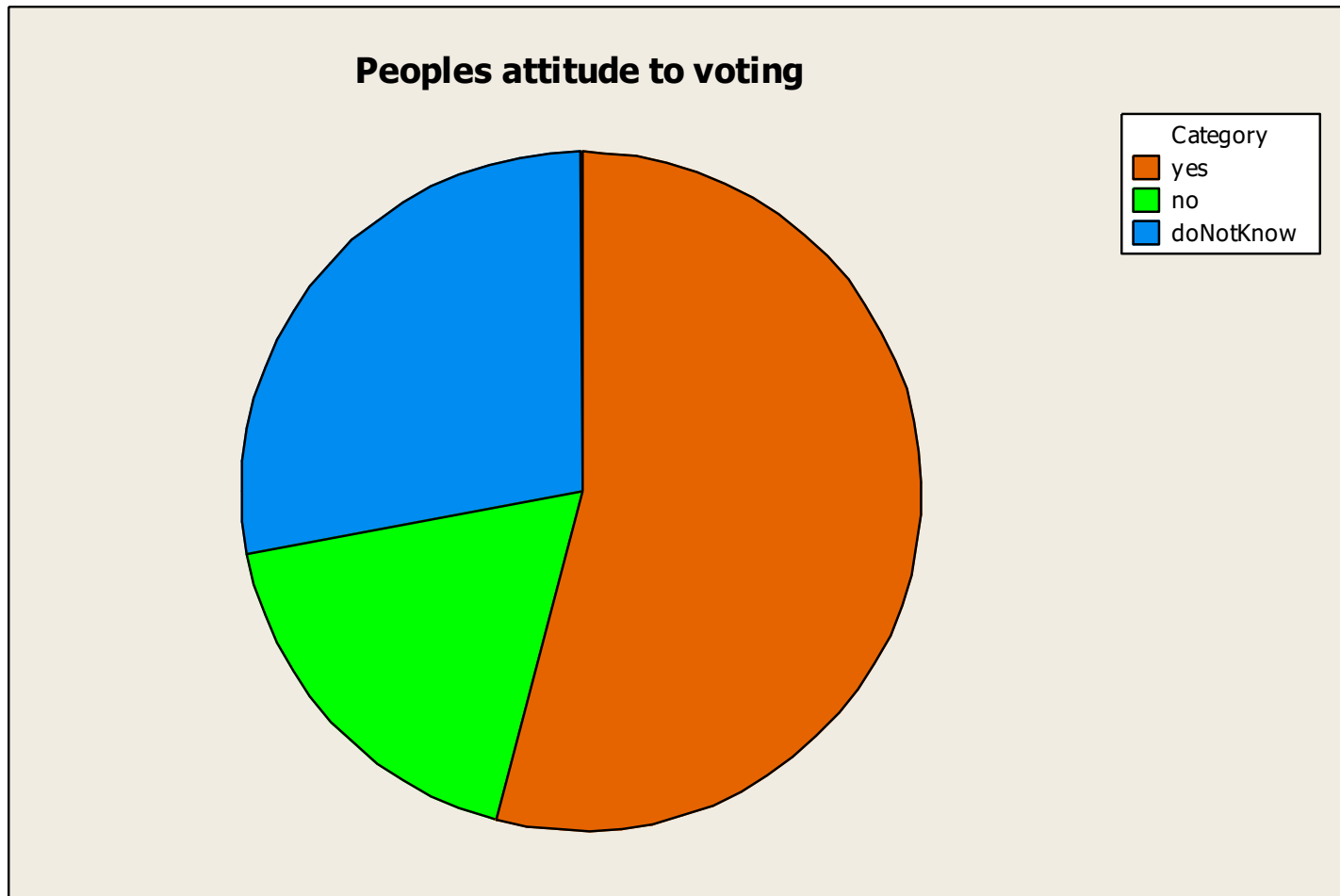
We can calculate the frequency or relative frequency for each category in the dataset. Then we can show relative frequencies in a pie chart or a bar chart.

Example: Do you want to vote?

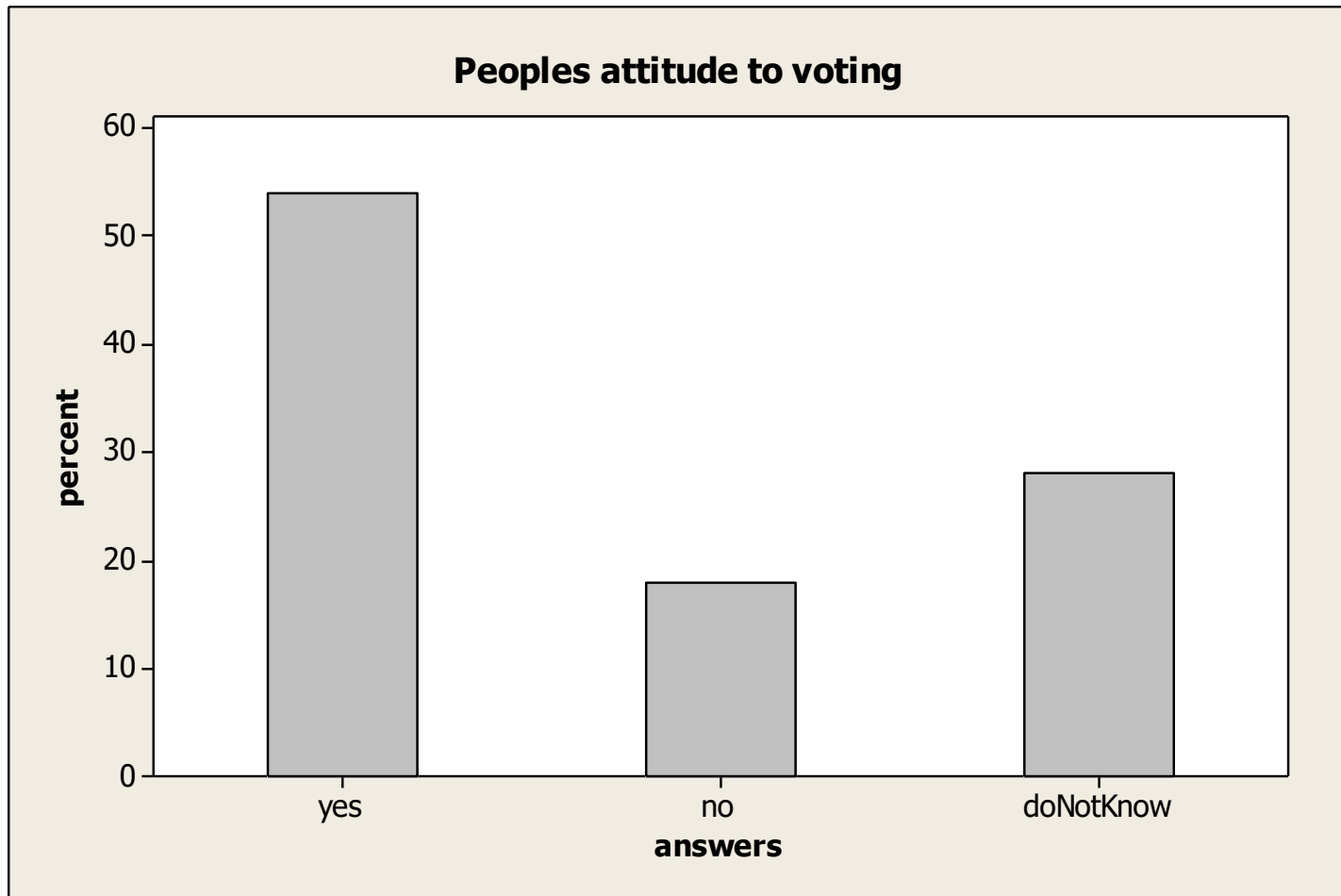
Answers:

Yes: 54% No: 18% don't know: 28%

Degrees for each category = percent*360



This bar chart shows the relative frequency of kinds of answers.



A Pareto diagram

is a bar chart, but the tallest bar comes first, then the second tallest and so on. The last bar shows the relative frequency for the category "other".

Discrete data.

Assume the distinct values observed are not too numerous. We can calculate the frequency and the relative frequency for each of the observed values. The results can be shown in a table, a line diagram or a histogram.

Example 1.

12 people picked strawberries in $\frac{1}{2}$ kg.
baskets. They picked:

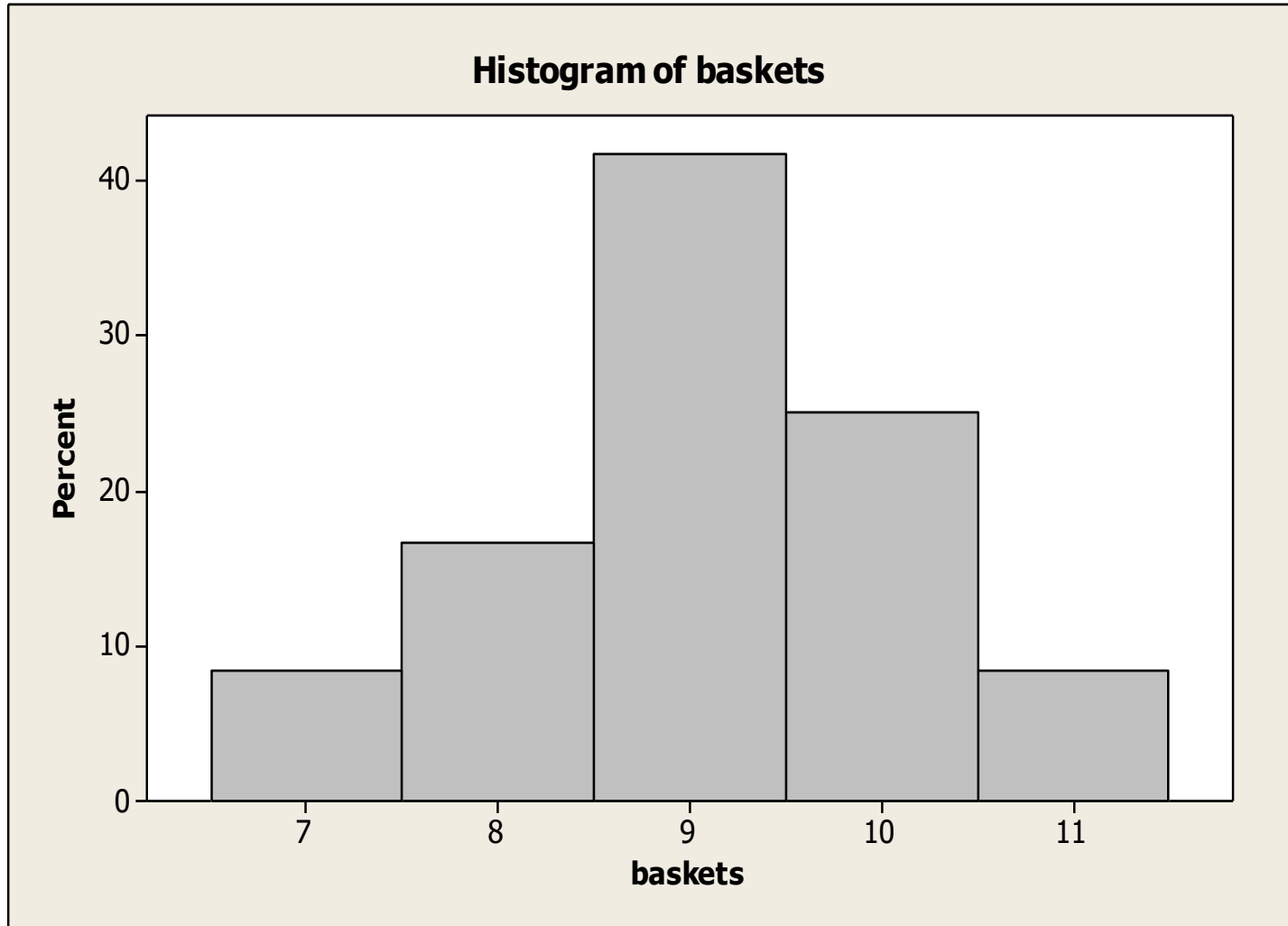
Number of baskets: 9, 9, 8, 10, 9, 11, 9, 8,
10, 10, 9, 7

Results from Minitab:

- **Results for: baskets.MTW**
-
- **Tally for Discrete Variables: baskets**

| • baskets | Count | CumCnt | Percent | CumPct |
|-----------|-------|--------|---------|--------|
| • 7 | 1 | 1 | 8,33 | 8,33 |
| • 8 | 2 | 3 | 16,67 | 25,00 |
| • 9 | 5 | 8 | 41,67 | 66,67 |
| • 10 | 3 | 11 | 25,00 | 91,67 |
| • 11 | 1 | 12 | 8,33 | 100,00 |
| • N= | 12 | | | |

Relative frequencies versus baskets:



Data on a continuous variable.

If we have 25 observations or less, we can make a dot (plot) diagram.

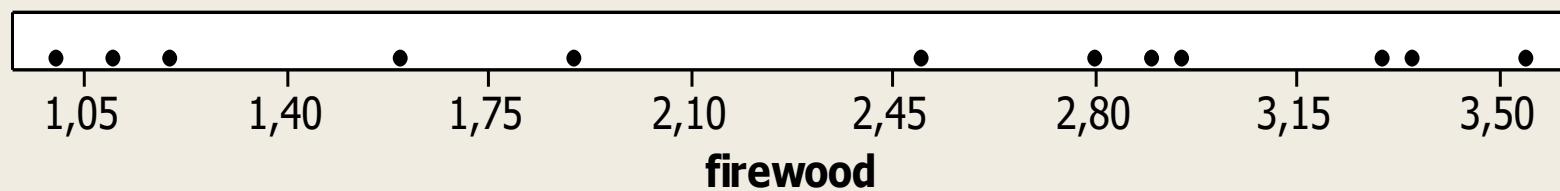
Example 2

We have 12 observations of firewood consumption in a district (in cubic meter per family per year):

| | | | | | | |
|------|------|------|------|------|------|------|
| 2.81 | 1.60 | 2.97 | 1.90 | 1.01 | | |
| 3.35 | 3.56 | 3.30 | 1.11 | 2.49 | 2.88 | 1.20 |

The values are plotted in a dot plot:

Firewood consumption in a district.



If the number of observations is more than 25, the scale will be split into intervals. Usually the intervals will have equal length, and there will be 5 – 15 intervals.

It is smart to choose intervals with boundaries which are easy to recall.

The number of observations which belong to each interval is counted. That is: class frequencies are calculated.

Relative frequencies for the intervals can be calculated.

Relative frequency =
class frequency/(total number of observations)

Make rules for what to do at boundaries.

Histogram for continuous data.

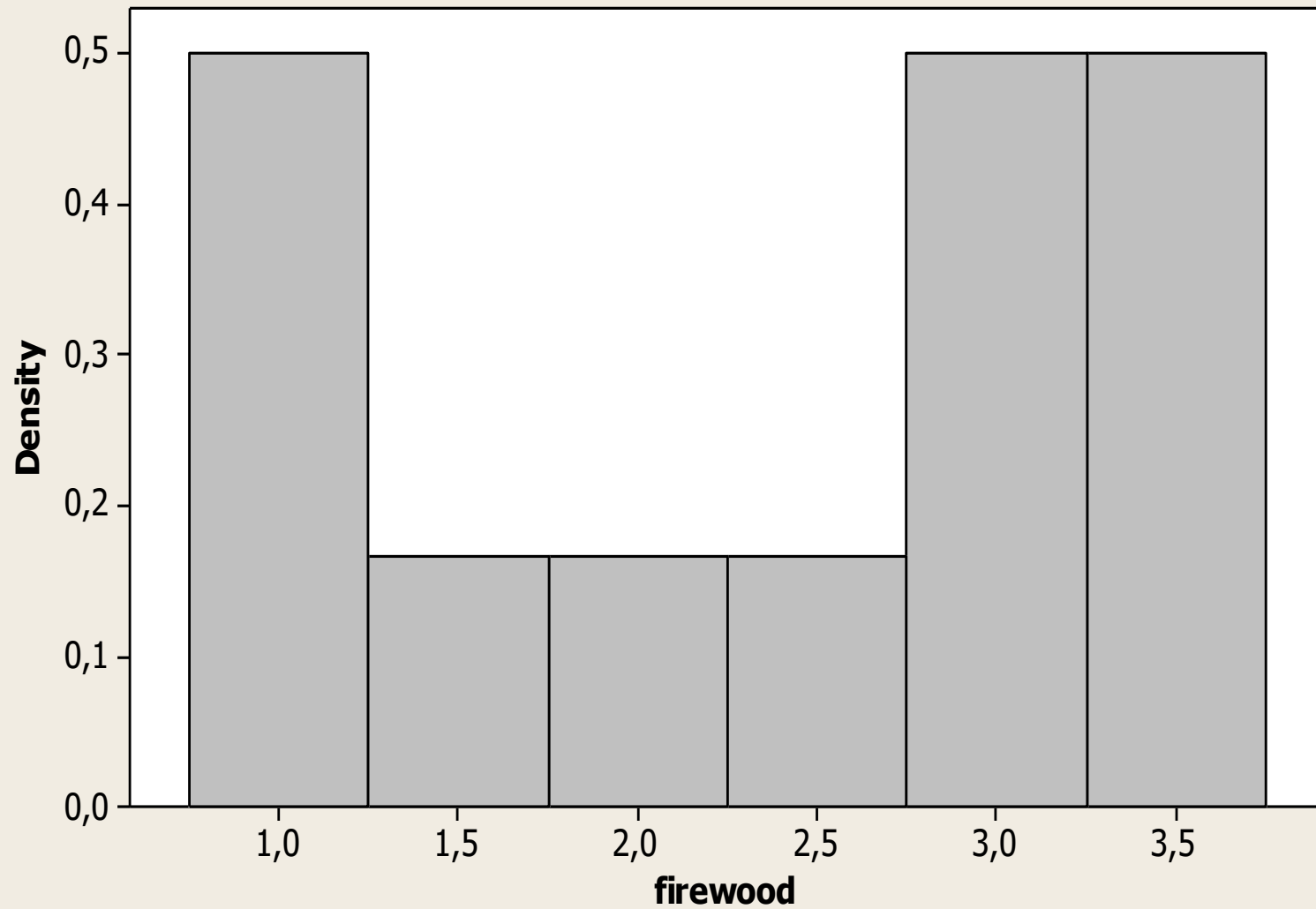
Draw vertical rectangles on each interval. The area represents the relative frequency.

The total area of the rectangles in a histogram = 1

Height= (relative frequency)/(width of interval)

The height of the rectangles must be calculated if the intervals do not have equal width. If the intervals have equal width then we can let the intervals have width 1 unit.

Histogram of firewood



Stem and leaf display.

Example 3: We have two digit numbers:

13 14 14 15 16 62 62 63 65 67

- **Stem-and-Leaf Display: observed**
- Stem-and-leaf of observed $N = 10$
- Leaf Unit = 1,0
- 5 1 34456
- 5 2
- 5 3
- 5 4
- 5 5
- 5 6 22357

Measures of center.

We have observations: X_1, X_2, \dots, X_n

Example 4.

Three newborn babies have these weights:

$X_1=2.1, X_2=3.2, X_3=3.7$

We can calculate the mean:

$$\bar{X} = \frac{1}{n} \sum_{i=1}^n X_i = \frac{1}{n} (X_1 + X_2 + \dots + X_n)$$

In our example:

$$\bar{X} = \frac{2.1 + 3.2 + 3.7}{3} = \frac{9}{3} = 3$$

To find the median: sort your observations. If n is odd, then we pick out the observation in the middle and let that observation be the median.

If n is even, then we pick out the two observations in the middle and calculate their average. This average will be the median.

We prefer the median if one or some observations are extreme and will have a large influence on the mean.

If the number of observations is large (more than 25-30) then we can calculate percentiles. We want to calculate the 100p-percentile:

Order the data from smallest to largest, we sort the data.

Calculate np . If np is an integer, say k , then calculate $(X_{(k)} + X_{(k+1)})/2$ where $X_{(k)}$ is sorted observation number k .

If np is not an integer we round it up to the next integer, say k and pick $X_{(k)}$ as the 100p-percentile

Q_1 =25th percentile=Lower (first) quartile.

Q_2 =50th percentile=Second quartile
(median)

Q_3 =75th percentile=Upper (third) quartile.

Measures of variation.

The sample variance of n observations:

$$S^2 = \frac{1}{n-1} \sum_{i=1}^n (x_i - \bar{x})^2$$

The sample standard deviation:

$$S = \sqrt{\text{Variance}} = \sqrt{\frac{1}{n-1} \sum_{i=1}^n (x_i - \bar{x})^2}$$

The sample standard deviation has the same unit as the observations.

Example 4 about newborn babies.

$$S^2 = \frac{1}{2} \left[(2.1 - 3)^2 + (3.2 - 3)^2 + (3.7 - 3)^2 \right]$$

$$= \frac{1}{2} \left[(-0.9)^2 + 0.2^2 + 0.7^2 \right]$$

$$= \frac{1}{2} \left[0.81 + 0.04 + 0.49 \right] = \frac{1}{2} \cdot 1.34 = 0.67$$

The sample standard deviation is:

$$S = \sqrt{0.67} = 0.8185$$

If our observations have a bell-shaped distribution, we expect:

68% of the observations in the dataset will lie within $\bar{x} \pm s$

95% of the observations in the dataset will lie within $\bar{x} \pm 2s$

99.7% of the observations in the dataset will lie within $\bar{x} \pm 3s$

Example 4 revisited:

$\bar{x} = 3$ and $s = 0.8185$. These results give:

$$\bar{x} \pm s \Leftrightarrow [2.181, 3.819]$$

$$\bar{x} \pm 2s \Leftrightarrow [1.363, 4.637]$$

$$\bar{x} \pm 3s \Leftrightarrow [0.544, 5.456]$$

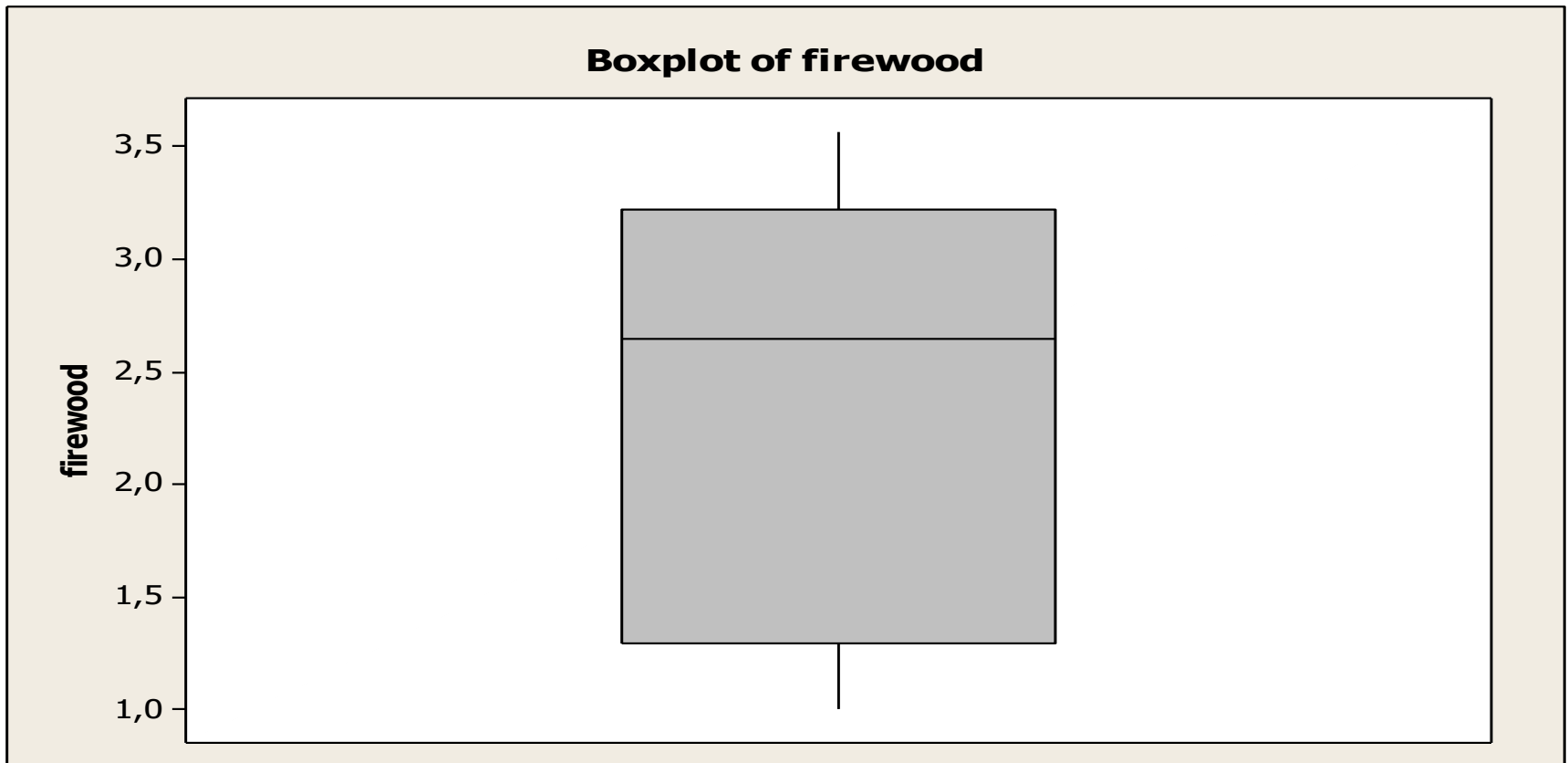
The sample range is defined as: the largest observation-the smallest observation.

The sample interquartile range=

Third quartile-First quartile.

Box plots

We have: minimum, Q_1 , Q_2 (the median), Q_3 , maximum. It is possible to make several box plots alongside each other to compare them.



Chapter 3

Several things can be measured on the same sampling unit, e.g. gender and type of education, height and weight.

Summarization of bivariate categorical data.

If 2 traits are observed on each sampling unit, a two way frequency table can be made.

Relative frequencies for the table or relative to a marginal total can be calculated.

Example 5.

We want to compare the proportions of males and females who believe in an afterlife. These data have been collected:

counts:

| | yes | no | total |
|---------|-----|-----|-------|
| females | 435 | 147 | 582 |
| males | 375 | 134 | 509 |
| total | 810 | 281 | 1091 |

Relative frequencies can be calculated for each gender. Then we can make comparisons:

-

| | yes | no | sum |
|---------|--------|--------|-----|
| females | 0.7474 | 0.2526 | 1 |
| males | 0.7367 | 0.2633 | 1 |
| sum | 0.7424 | 0.2576 | 1 |

- It seems that the proportion of females who believe in an afterlife is similar to that of males.

Simpson's paradox

can occur if data from different sources are combined into a single table. We have 3 traits and will make two way tables for 2 of them. We can disregard the third trait, and make a two way table for the 2 traits of interest. We can also make two way tables for the 2 traits of interest for each category of the third trait. If we are getting confusing results, we will call it Simpson's paradox.

Example 6.

We ask 50 young and 50 old females and 50 young and 50 old males this question:

Do you think the cinema needs an upgrade?

The results are:

females

| | yes | no | sum |
|-------|-----|----|-----|
| young | 30 | 20 | 50 |
| old | 20 | 30 | 50 |
| sum | 50 | 50 | 100 |

males

| | yes | no | Sum |
|-------|-----|----|-----|
| young | 20 | 30 | 50 |
| old | 30 | 20 | 50 |
| sum | 50 | 50 | 100 |

Combined:

| | Yes | no | Sum |
|-------|-----|-----|-----|
| Young | 50 | 50 | 100 |
| Old | 50 | 50 | 100 |
| Sum | 100 | 100 | 200 |

The last table shows equal frequencies for young and old. The information about gender has been disregarded and we see no difference.

A designed experiment for making a comparison.

A double blind test can be conducted to eliminate the placebo effect.

People in the trial don't know whether they get real treatment or just a placebo.

We observe if the treatment has an effect or not.

Treatments can be compared statistically.

If we have one group of subjects, we have to split it into two groups by random assignment and then give the treatment to one group and placebo to the other group.

We will calculate the proportion with effect for both the treatment group and the placebo group.

Scatter diagram of bivariate measurement data.

Two numerical observations (x,y) are recorded for each sample unit.

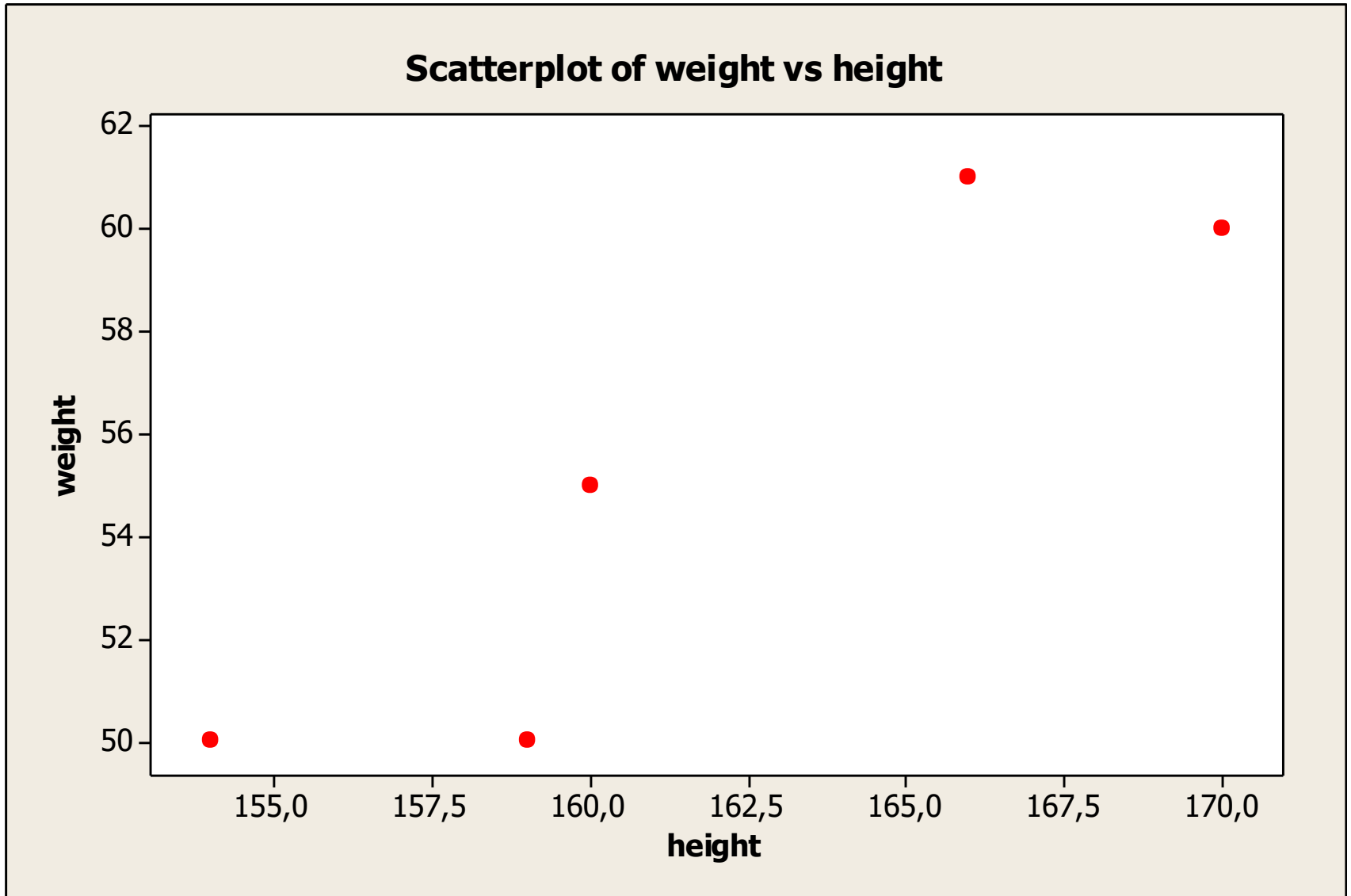
Example 7.

| | |
|-----------------------|------------|
| $X_i = \text{height}$ | } person i |
| $Y_i = \text{weight}$ | |

Observations:

- height weight
- 160 55
- 166 61
- 154 50
- 170 60
- 159 50

Plot of y versus x.



Example 8.

Pigs are observed.

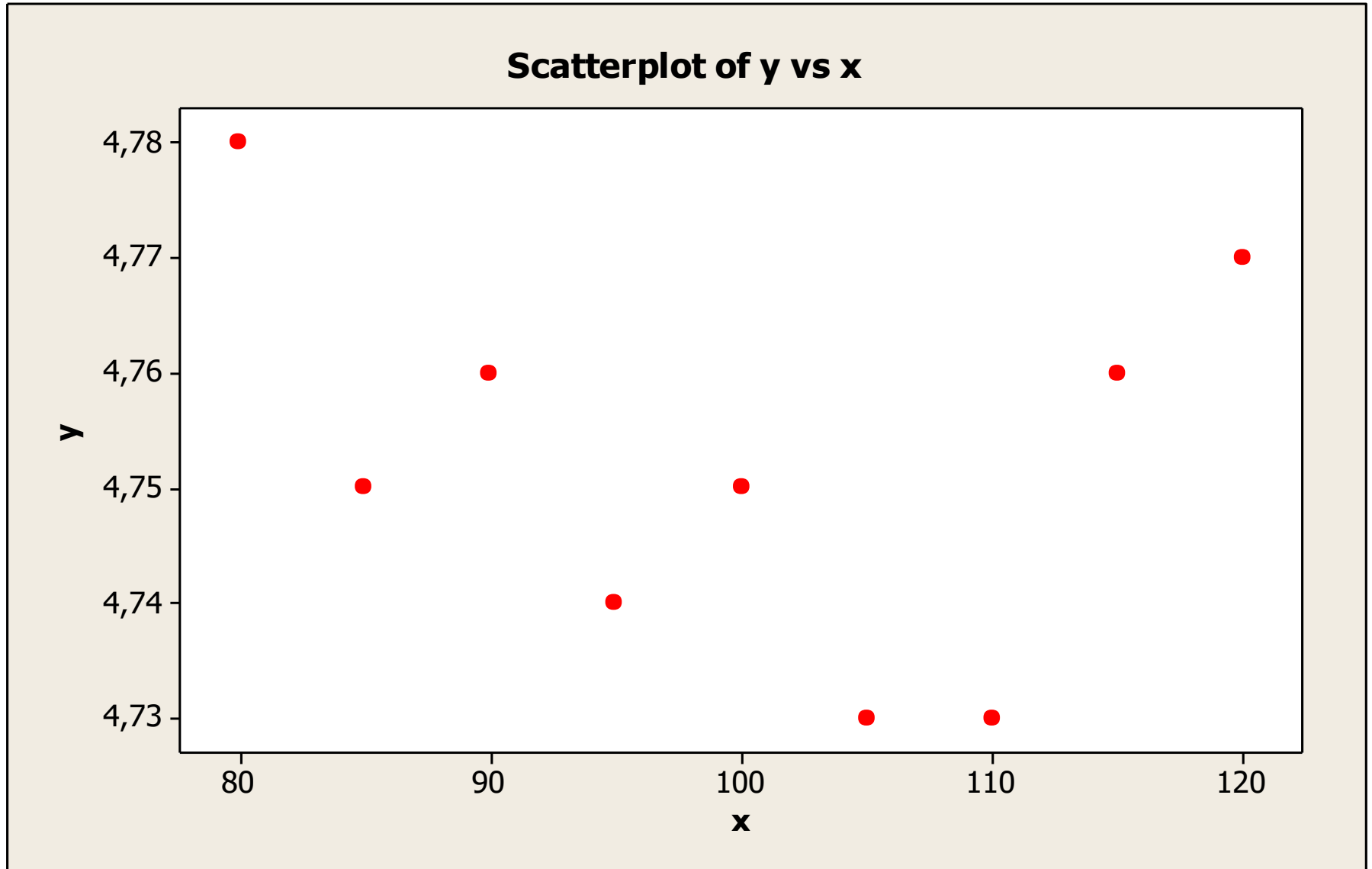
X_i =weight when slaughtered

Y_i =amount of fodder per kg. slaughter weight.

Observations:

| | | | | | | | | | |
|---|------|------|------|------|------|------|------|------|------|
| X | 80 | 85 | 90 | 95 | 100 | 105 | 110 | 115 | 120 |
| Y | 4,78 | 4,75 | 4,76 | 4,74 | 4,75 | 4,73 | 4,73 | 4,76 | 4,77 |

The relationship between x and y is not linear.



We want to assess if X and Y are related, and what kind of relationship we have. There could be no relationship between X and Y.

The correlation coefficient, a measure of linear relation.

It is given by:

$$r = \frac{\sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^n (x_i - \bar{x})^2 \sum_{i=1}^n (y_i - \bar{y})^2}}$$

$$S_{xy} = \sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y}) = \sum_{i=1}^n x_i y_i - \frac{1}{n} \left(\sum_{i=1}^n x_i \right) \left(\sum_{i=1}^n y_i \right)$$

$$S_{xx} = \sum_{i=1}^n (x_i - \bar{x})^2 = \sum_{i=1}^n x_i^2 - \frac{1}{n} \left(\sum_{i=1}^n x_i \right)^2$$

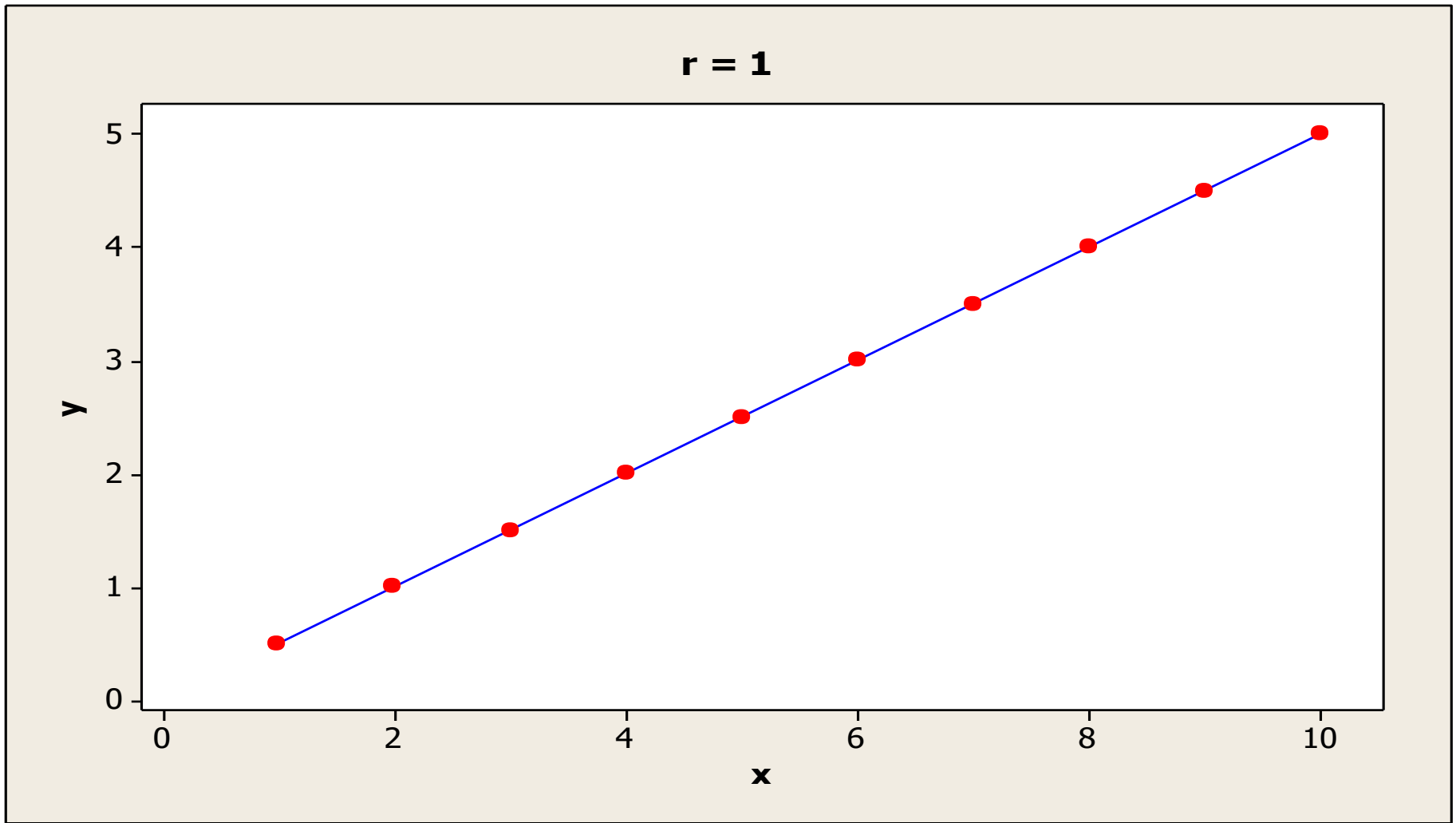
$$S_{yy} = \sum_{i=1}^n (y_i - \bar{y})^2 = \sum_{i=1}^n y_i^2 - \frac{1}{n} \left(\sum_{i=1}^n y_i \right)^2$$

$$r = \frac{S_{xy}}{\sqrt{S_{xx}S_{yy}}}$$

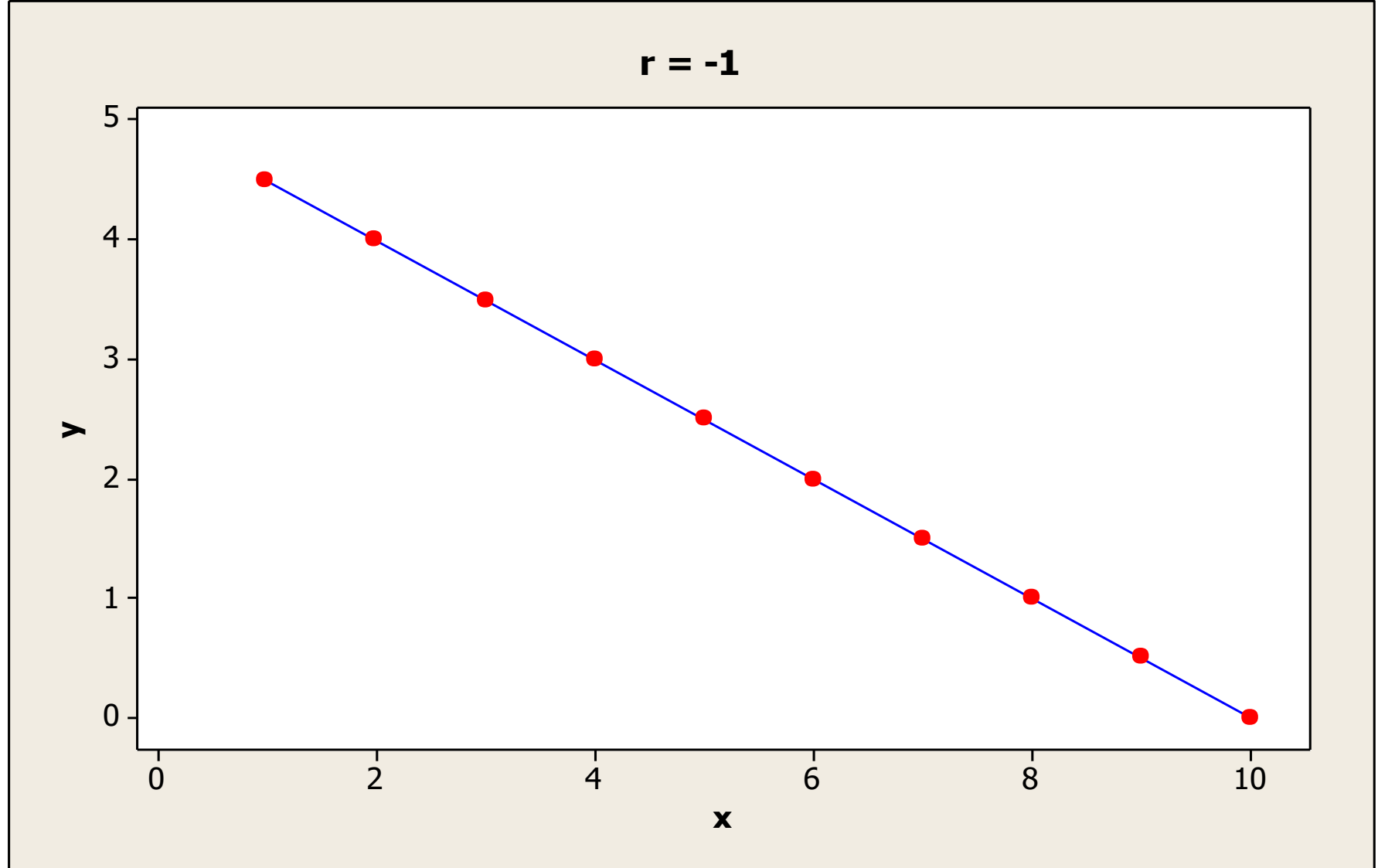
We have: $-1 \leq r \leq 1$

$r > 0$, large x and y together, small x and y together.

$r < 0$, large x and small y together, small x and large y together.



All points are on a straight line from low left to high right when $r=1$.



All points are on a straight line from high left to low right when $r = -1$.

- r shows the strength of a linear relation.
- When r is close to 1 or -1 there is a strong linear relationship between x and y .
- When r is close to 0 there is very little linear relationship between x and y .

Correlation and causation.

- Even if we have a correlation between two variables, we can't be sure that we have a cause and effect relationship.
- In some situations we can influence y by changing x . We might be able to arrange things in a most favorable way.

Prediction of one variable from another. (linear regression).

- We assume: $Y_i = \beta_0 + \beta_1 x_i + e_i$
- Y_i = the response variable
- x_i = the predictor variable
- ☐ β_0 = the intercept
- ☐ β_1 = the slope
- e_i = the error term

Example 7 revisited:

We have 5 pairs of observations:

(x_1, y_1) (x_2, y_2) (x_3, y_3) (x_4, y_4) (x_5, y_5) and fit a line to our observations.

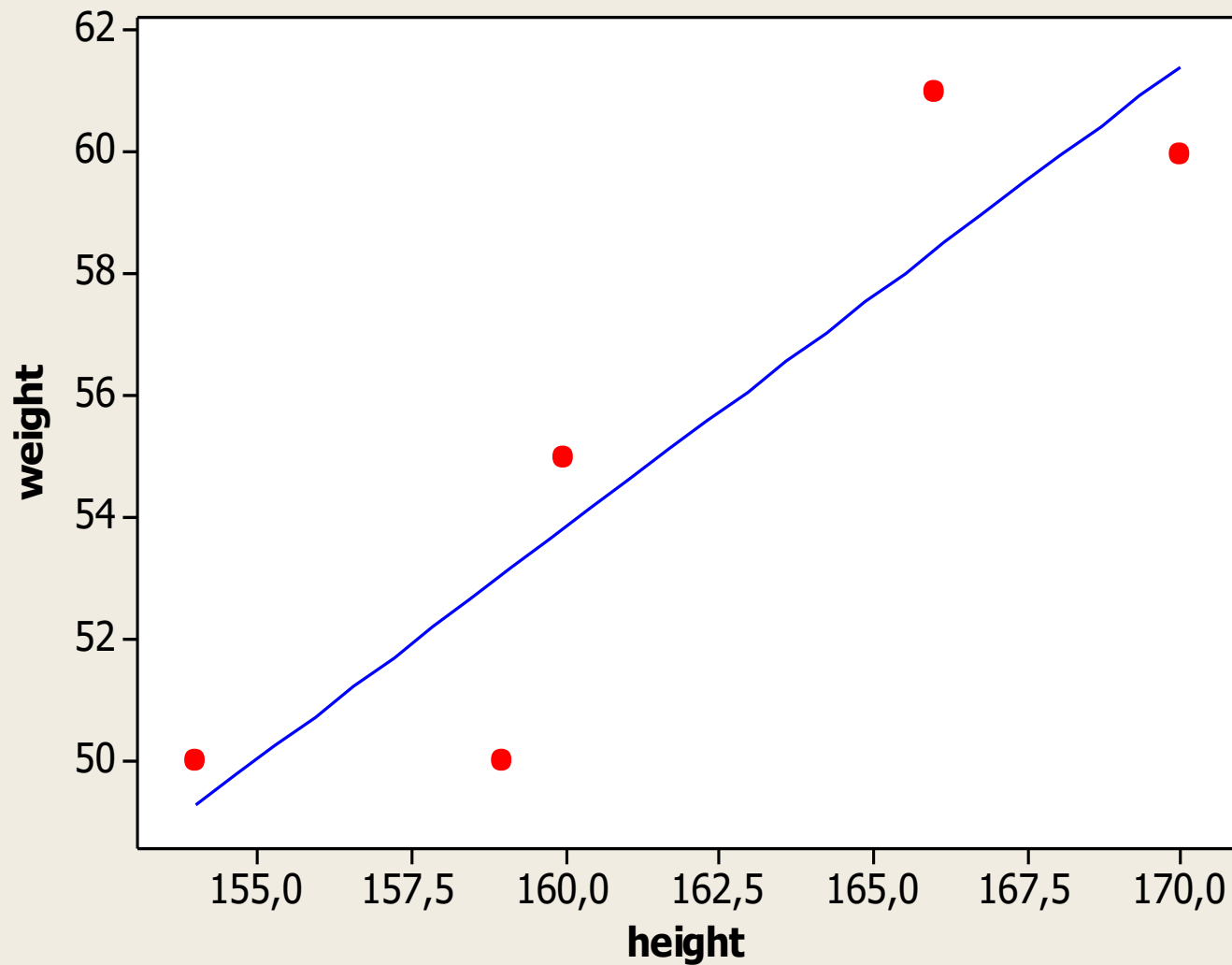
The fitted line is: $\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x$

where $\hat{\beta}_1 = \frac{S_{xy}}{S_{xx}}$ = the estimated slope

and $\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x}$ = the estimated intercept.

Fitted Line Plot

$$\text{weight} = -67,80 + 0,7602 \text{ height}$$



| | |
|-----------|---------|
| S | 2,59382 |
| R-Sq | 81,8% |
| R-Sq(adj) | 75,7% |

$\hat{\beta}_0$ and $\hat{\beta}_1$ are determined by the method of least squares. That is to minimize:

$$D = \sum_{i=1}^n (y_i - \beta_0 - \beta_1 x_i)^2$$

with respect to the two unknown parameters. The best fitting line is:

$$\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 \mathbf{x}$$

Now we can predict y_i if we know x_i :

$$\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_i$$

The value of x can be one of the values in the dataset, or another hypothetical value. But the hypothetical value should not be far away from all x -values in the dataset.

Chapter 4. Probability.

An experiment is the process of observing a phenomenon that has variation in its outcomes.

E.g: success or failure, rain, sun, storm (weather), head or tail (tossing a coin) etc.

Definition:

- The sample space = all possible distinct outcomes of the experiment.
- An event = a subspace of the sample space, the outcome of an experiment.
- We want to find the probability or chance that an event will occur.
- In some situations we can calculate probabilities by reasoning.

- In other situations we can give probabilities by our experience. These kind of probabilities will probably not be accurate.
- We can also try to give probabilities after conducting an experiment. To rely on these results we should have a lot of observations.
- E.g: A weather forecast can be very uncertain.

- Tossing a coin: 50% chance of getting a head, 50% chance of getting a tail.
- Probability: $P(\text{"head"}) = 0.5$ $P(\text{"tail"}) = 0.5$
- We have always:

$$0 \leq P(\text{event}) \leq 1$$

$$P(\text{sample space}) = 1$$

$$P(\text{the null event}) = 0 = P(\emptyset)$$

- Let the sample space be $S = \{e_1, e_2, \dots, e_k\}$
- e_1, e_2, \dots, e_k are elements in S .
- Let A be an event.

$$P(A) = \sum_{\text{all } e \text{ in } A} P(e)$$

If all elements in S have the same probability of occurring, then:

$$P(e_i) = \frac{1}{k}$$

$$P(A) = \frac{m}{k} = \frac{\text{number_of_elements_in_}A}{\text{number_of_elements_in_}S}$$

In a trial the event A can occur or not. We can do the trial many times, say n trials. Assume A occurs in m of them. We will estimate $P(A)$:

$$P(A) \approx \frac{m}{n}$$

Let A be an event. A or \bar{A} (= the complement of A)

$$1 = P(S) = P(A \cup \bar{A}) = P(A) + P(\bar{A})$$

$$P(A) = 1 - P(\bar{A})$$

If the events A and B are mutually exclusive then: $A \cap B = \emptyset$ and

$$P(A \cup B) = P(A) + P(B)$$

If A and B are not mutually exclusive then:

$$P(A \cup B) = P(A) + P(B) - P(A \cap B)$$

Conditional probability and independence.

- Assume that we know that the event B has occurred. What is now the probability that A will occur?
- We write this: $P(A|B)$ = the conditional probability of A given B.
- In some situations we will have that:
 $P(A|B) = P(A)$
- That is: The probability that A will occur is the same whether B occurs or if we do not know if B occurs.

- Then A and B are independent. We have:

$$P(A | B) = \frac{P(A \cap B)}{P(B)}$$

- or

$$P(A \cap B) = P(A | B)P(B)$$

- If A and B are independent then:

$$P(A \cap B) = P(A)P(B)$$

= the multiplication law of probability.

- If

$$P(A \cap B) = P(A)P(B)$$

- then A and B are independent.

Example 9

Let C = "colorblind" ,

M = "male" and F = "female"

We know: $P(C|M)=0.08$ and $P(C|F)=0.02$

Calculate the probability of being colorblind in a population with 52% males.

Solution

$$P(C | M) = \frac{P(C \cap M)}{P(M)}$$

gives:

$$P(C \cap M) = P(C | M)P(M) = 0.08 \cdot 0.52 = 0.0416$$

$$P(C \cap F) = P(C | F)P(F) = 0.02 \cdot 0.48 = 0.0096$$

$$P(C) = P(C \cap M) + P(C \cap F) = 0.0416 + 0.0096 = \underline{\underline{0.0512}}$$

Chapter 5. Probability distributions.

A random variable X associates a numerical value with each outcome of an experiment.

A random variable is discrete if it has a finite number of values or countable infinite number of values.

A random variable is continuous if it can take all values on a continuous scale or interval.

Example 10:

We toss a coin three times. Let

X = the number of heads.

X can take the values 0, 1, 2, 3.

Example 11:

We count

X = the number of traffic accidents at an intersection.

X can take the values 0, 1, 2,...infinite.

- The probability distribution of a discrete random variable X is a list of the distinct numerical values of X along with their associated probabilities.
- The probabilities could be given by a formula.

Example 10 revisited (about tossing a coin).

$$X=0 \Leftrightarrow \{TTT\} \quad P(X=0) = \frac{1}{2} \cdot \frac{1}{2} \cdot \frac{1}{2} = \frac{1}{8}$$

$$X=1 \Leftrightarrow \{HTT, THT, TTH\} \quad P(X=1) = \frac{1}{8} + \frac{1}{8} + \frac{1}{8} = \frac{3}{8}$$

$$X=2 \Leftrightarrow \{HHT, HTH, THH\} \quad P(X=2) = \frac{3}{8}$$

$$X=3 \Leftrightarrow \{HHH\} \quad P(X=3) = \frac{1}{8}$$

$$P(X = 0) = \frac{1}{8}$$

$$P(X = 1) = \frac{3}{8}$$

$$P(X = 2) = \frac{3}{8}$$

$$P(X = 3) = \frac{1}{8}$$

$$\sum_{i=0}^3 P(X = i) = P(X = 0) + P(X = 1) + P(X = 2) + P(X = 3) = 1$$

Form of a discrete probability distribution.

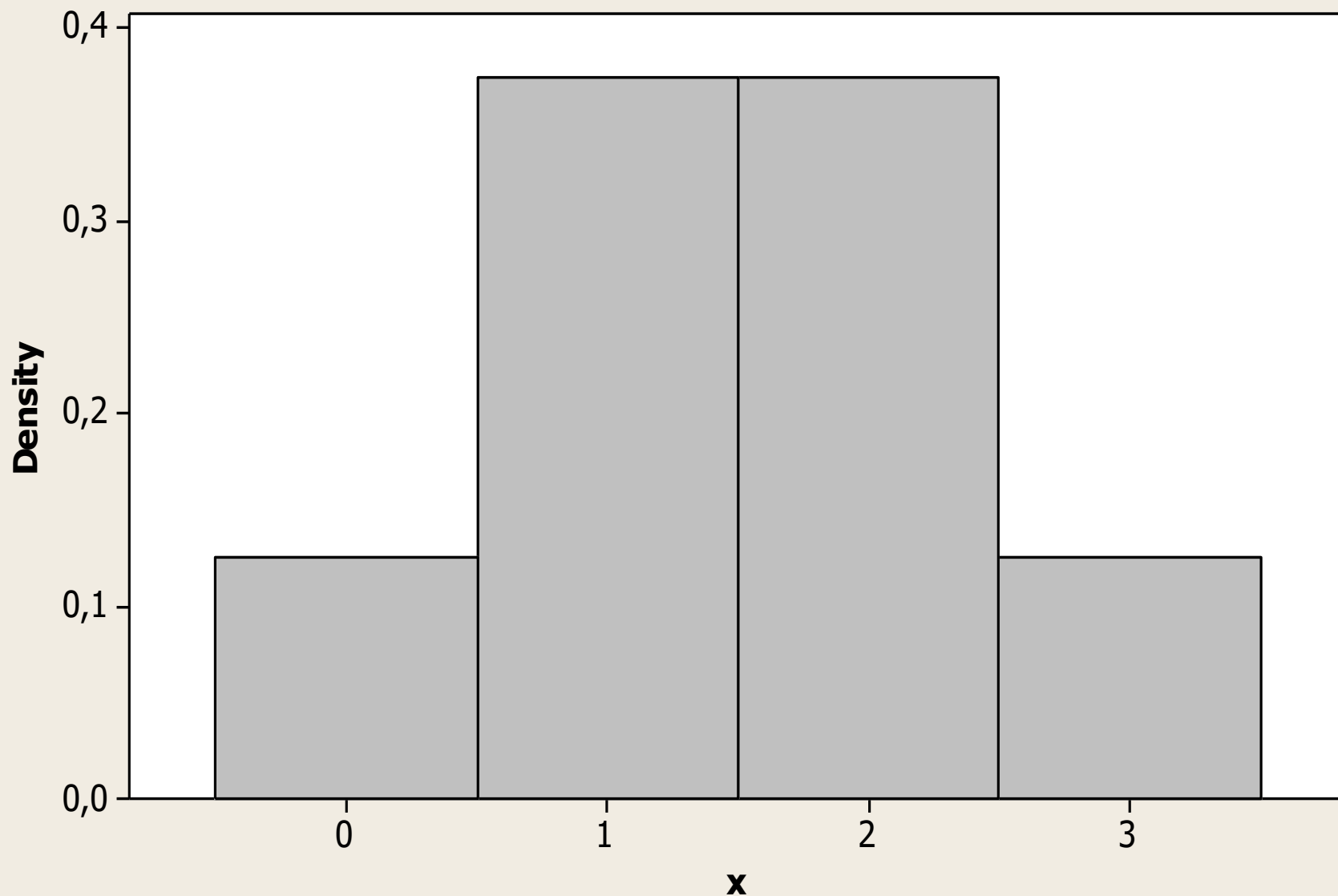
The values X can take | Probability

| | |
|-------|----------|
| x_1 | $f(x_1)$ |
| x_2 | $f(x_2)$ |
| · | · |
| x_k | $f(x_k)$ |

The probability distribution must satisfy these conditions: $0 \leq f(x_i) \leq 1$ for each x_i , $i=1, \dots, k$

$$\sum_{i=1}^k f(x_i) = 1$$

The probability histogram of X =the number of heads in 3 tosses of a coin.



- A probability distribution can be given by experience, a trial or by reasoning.
- In some situations we do not know the probability distribution of X .
- We can conduct a trial and then estimate the probability distribution of X .

- We will calculate the relative frequencies for the values X can take and use the results as the probability distribution of X .
- If the trial is vast then the relative frequencies will be close to the probabilities.
- If the trial is small, the results will be very uncertain.

Example 12. (Look at page 179)

- Let X = the number of magazines to which a college senior subscribes.

| subscriptions | Frequency | Relative frequency |
|---------------|-----------|--------------------|
| 0 | 61 | $61/400 = 0.15$ |
| 1 | 153 | $153/400 = 0.38$ |
| 2 | 106 | $106/400 = 0.27$ |
| 3 | 56 | $56/400 = 0.14$ |
| 4 | 24 | $24/400 = 0.06$ |
| total | 400 | 1 |

- We use the relative frequencies in the table as the probability distribution of X .
- If we want to find $P(X \geq 2)$:
- $P(X \geq 2) = P(X=2) + P(X=3) + P(X=4) = 0.27 + 0.14 + 0.06 = \underline{0.47}$
- $P(X \leq 1) = P(X=0) + P(X=1) = 0.15 + 0.38 = \underline{0.53}$
- We have: $P(X \leq 1) = 1 - P(X \geq 2) = 1 - 0.47 = \underline{0.53}$

Expectation (mean) and standard deviation of a probability distribution.

- The expectation of a probability distribution = the center of a probability distribution.
- It is given by the formula:

$$\square \mu = E(X) = \sum x_i f(x_i) = \sum x_i P(X = x_i)$$

i can have values 1,...,n or from 1 to infinite.

Example 10 revisited.

- X = the number of heads in 3 tosses of a coin.

$$\mu = E(X) = \sum_{i=0}^3 i \cdot P(X = i)$$

$$= 0 \cdot P(X = 0) + 1 \cdot P(X = 1) + 2 \cdot P(X = 2) + 3 \cdot P(X = 3)$$

$$= 0 \cdot \frac{1}{8} + 1 \cdot \frac{3}{8} + 2 \cdot \frac{3}{8} + 3 \cdot \frac{1}{8} = \underline{\underline{1.5}}$$

The balance point of the distribution is 1.5

- The standard deviation of a probability distribution = a measure for the spread of a probability distribution.
- The standard deviation of $X = \text{sd}(X) = \sigma$.

$$= \sqrt{\text{Var}(X)} = \sqrt{\sum (x_i - \mu)^2 P(X = x_i)}$$

$\sigma^2 = \text{Var}(X)$ = the population variance.

Example 10 about 3 tosses of a coin.

- X = the number of heads in 3 tosses.

$$\sigma^2 = \sum_{i=0}^3 (i - \mu)^2 P(X = i)$$

$$= (0 - 1.5)^2 \cdot \frac{1}{8} + (1 - 1.5)^2 \cdot \frac{3}{8} + (2 - 1.5)^2 \cdot \frac{3}{8} + (3 - 1.5)^2 \cdot \frac{1}{8}$$

$$= 0.75$$

$$\sigma = \sqrt{\text{Var}(X)} = \sqrt{0.75} = \underline{\underline{0.875}}$$

- We have an alternative formula for the calculation of $\text{Var}(X)$:

$$\sigma^2 = \sum_{i=1}^n x_i^2 P(X = x_i) - \mu^2$$

- **Example 10:**

- $\text{Var}(X) = 0^2 \cdot \frac{1}{8} + 1^2 \cdot \frac{3}{8} + 2^2 \cdot \frac{3}{8} + 3^2 \cdot \frac{1}{8} - 1.5^2 = \underline{\underline{0.75}}$

Bernoulli trials.

- A trial which can have only two outcomes is called a Bernoulli trial.
- The outcome can be “success” or “failure”.
- If we conduct n such trials, and they are independent and $P(\text{“success”})=p$ is the same for all trials, we have a binomial situation.
- We can observe X = the number of successes in n trials.
- X has a binomial distribution.

- X could be:
- The number of heads in n tosses of a coin. Here $p = P(\text{"head"}) = 0.5$ if the coin is fair.
- Assume $p = P(\text{"a random tree has beetles"}) = 0.30$ in a large forest, and that the trees have beetles or not independently. Then $X =$ the number of trees with beetles out of n randomly picked trees.

Small populations.

- Assume we have a population with two categories of elements, e.g. B and \overline{B}
- The population consists of 15 items and 5 are in category B.
- We draw 2 items, one at the time.
- We have: $P(B_1) = \frac{5}{15}$ for the first drawing.
- For the next drawing we have: $P(B_2|B_1) = \frac{4}{14}$ so we do not have a binomial situation.

- If the sampling is with replacement, we have a binomial situation.
- If we have a large population, say at least 100 and we only draw a small sample, say less than 10% of the population, we will have: $P(B_2|B_1) \approx P(B_1)$ and so on. It is approximately a binomial situation.

The Binomial Distribution.

- $P(X=x) = \binom{n}{x} p^x (1-p)^{n-x}$ for $x=0,1,..n$
- Where: $\binom{n}{x} = \frac{n!}{x!(n-x)!}$
- **Example 10:** Let X = the number of heads in 3 tosses of a coin.
- $P(X=0) = \binom{3}{0} \left(\frac{1}{2}\right)^0 \left(1 - \frac{1}{2}\right)^{3-0} = \frac{3!}{0!3!} \cdot \left(\frac{1}{2}\right)^3 = \frac{1}{8}$

- $P(X=1) = \binom{3}{1} \left(\frac{1}{2}\right)^1 \left(1 - \frac{1}{2}\right)^{3-1} = \frac{3!}{1!2!} \cdot \left(\frac{1}{2}\right) \left(\frac{1}{2}\right)^2 = \frac{3}{8}$

- $P(X=2) = \binom{3}{2} \left(\frac{1}{2}\right)^2 \left(1 - \frac{1}{2}\right)^{3-2} = \frac{3!}{2!1!} \cdot \left(\frac{1}{2}\right)^2 \left(\frac{1}{2}\right) = \frac{3}{8}$

- $P(X=3) = \binom{3}{3} \left(\frac{1}{2}\right)^3 \left(1 - \frac{1}{2}\right)^{3-0} = \frac{3!}{3!0!} \cdot \left(\frac{1}{2}\right)^3 = \frac{1}{8}$

- If we know p in our formula, we can calculate probabilities and make a figure of the distribution.

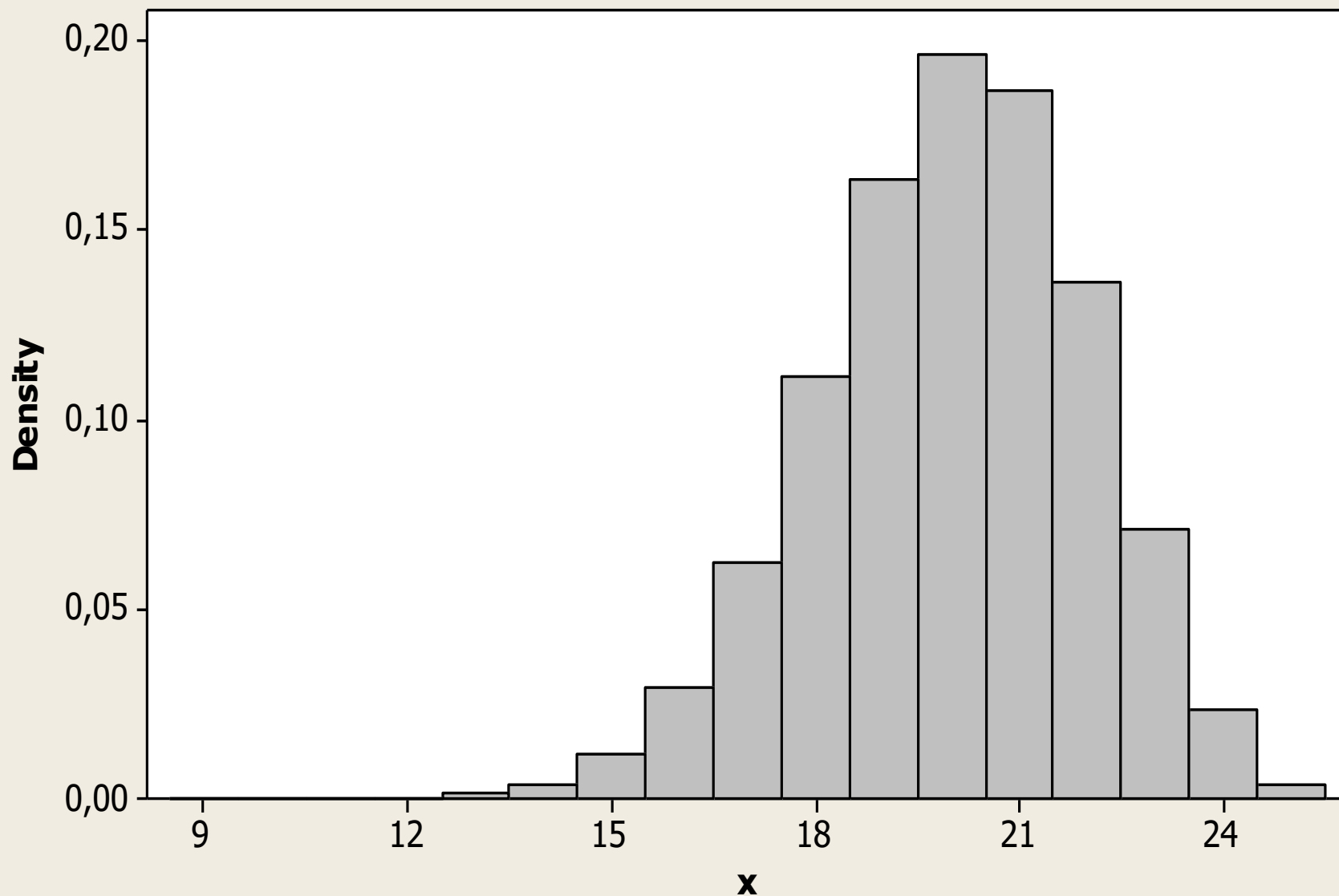
- A probability model is an assumed form of the probability distribution that describes the chance behaviour for a random variable X .
- Cumulative Binomial probabilities are tabulated in table 2 in the text book (page 617) when $n < 26$ and $p = \{0.05, 0.1, 0.2, 0.3, 0.4, 0.5, 0.6, 0.7, 0.8, 0.9, 0.95\}$

Example 13.

- We sow 10 seeds. On the envelope it says: The probability of germination is 80%. The seeds are a bit old, so we don't know if this is still true.
- After a while we observe 6 germs out of 10 seeds.
- What is the probability of getting 6 or less germs out of 10 seeds if the probability of germination = $p = 0.80$?

- $P(X \leq 6) = B_{10}(6, 0.80) = \underline{0.121}$
- This is not very likely, but it can happen.
- We now sow 25 seeds and get 16 germs.
- We find: $P(X \leq 16) = B_{25}(16, 0.8) = \underline{0.047}$
- It is unlikely to get 16 or less germs out of 25 seeds if $p = 0.80$.
- This could indicate that $p < 0.80$

Probability histogram for the binomial distribution when $n=25$ and $p=0.80$



The Cumulative Binomial Distribution.

- $P(X \leq c) = B_n(c, p) =$

$$\sum_{x=0}^c \binom{n}{x} p^x (1-p)^{n-x}$$

= the area of all the bars to the left of c , the bar for c inclusive.

- We will also find the probability of getting 16 or more germs out of 25 seeds when $p = 0.80$:
- $P(X \geq 16) = 1 - P(X \leq 15) = 1 - B_{25}(15, 0.80) = 1 - 0.017 = \underline{0.983}$
- What is the probability of finding between 14 and 17 germs if $n=25$ and $p=0.80$?
- $P(14 \leq X \leq 17) = B_{25}(17, 0.80) - B_{25}(13, 0.80) = 0.109 - 0.002 = \underline{0.107}$

The Mean and Standard deviation of the Binomial Distribution.

- If X has a binomial distribution with n and p , then
- $E(X) = np$
- $\text{Var}(X) = np(1-p)$
- The standard deviation of X is:

$$\sigma = \sqrt{\text{Var}(X)} = \sqrt{np(1-p)}$$

- If $n = 25$ and $p = 0.80$ then:

- $E(X) = 25 \cdot 0.80 = \underline{\underline{20}}$

- $\text{Var}(X) = 25 \cdot 0.80 \cdot 0.20 = \underline{4}$

$$\sigma = \sqrt{\text{Var}(X)} = \text{sd}(X) = \sqrt{4} = \underline{\underline{2}}$$

Chapter 6. Continuous distributions

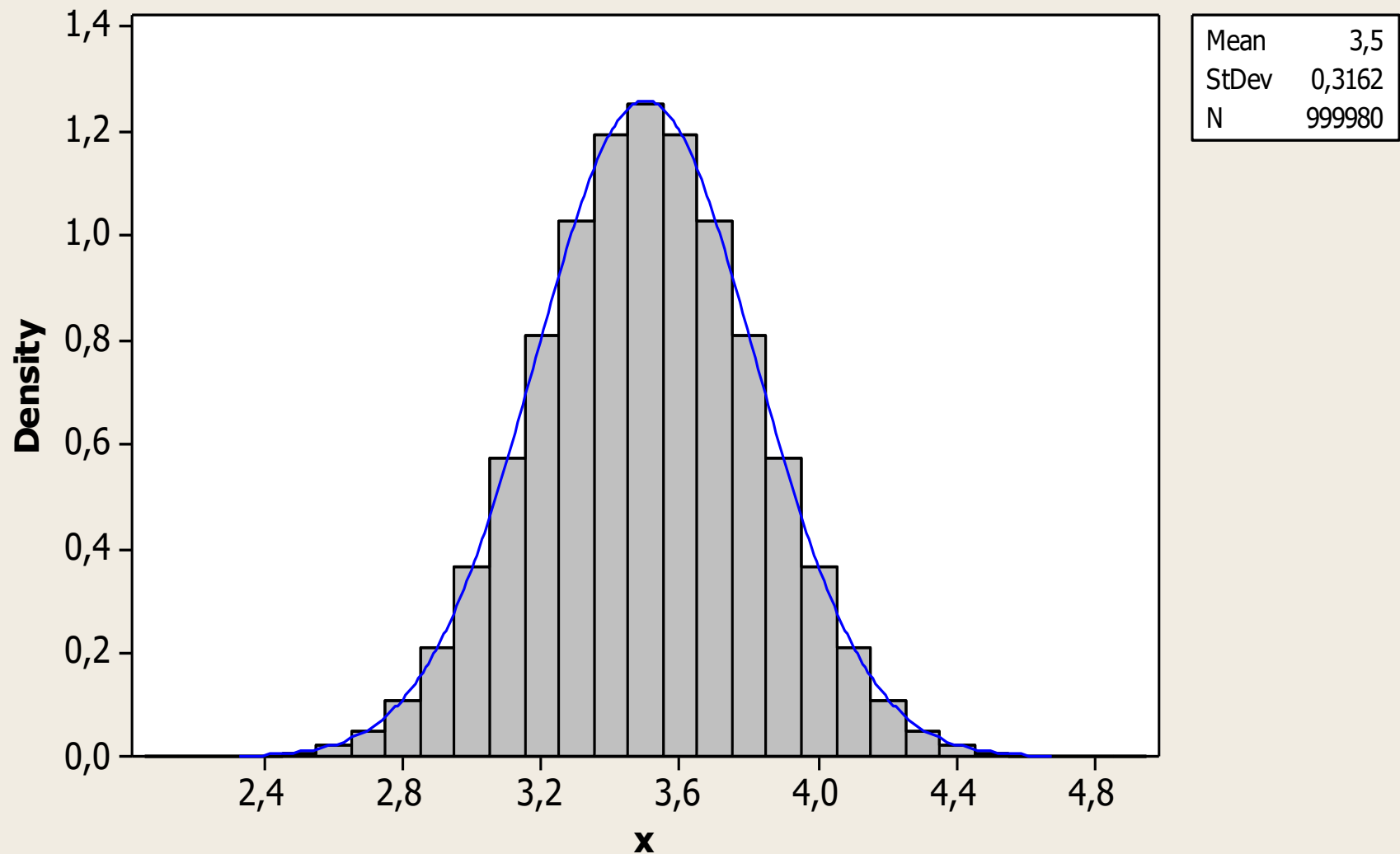
We measure:

X = the weight of a newborn baby.

We can have registrations for 100 – 500 – 1000 babies. Then we can make a histogram. The histogram can be approximated to a smooth curve.

Histogram of the weights of newborn babies.

Normal

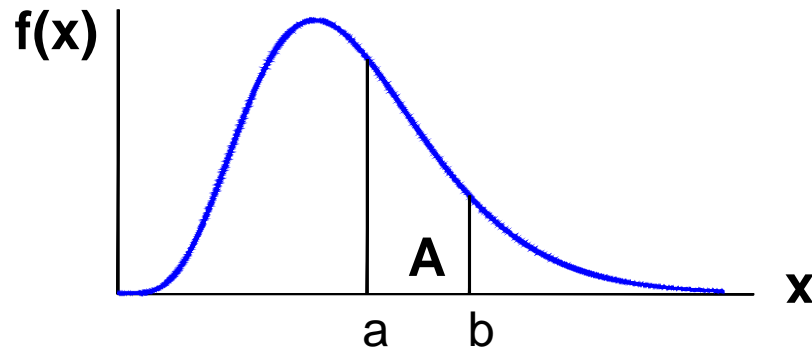


- We will call this smooth curve $f(x)$.
- $f(x)$ is the probability density function for X , a probability distribution for X .
- It has the properties:
- 1) The total area under the probability density curve = 1. That is:

$$\int f(x)dx = 1$$

- 2) $P(a \leq X \leq b)$ = the area under the probability density curve between a and b =

$$\int_a^b f(x) dx = A$$



- 3) $f(x) \geq 0$ for all x .
- We have: $P(X = a) = 0$
- We must think of probabilities for X as areas under the probability density curve.

- Some important distributions have tables of areas.
- The standard normal distribution is tabulated for the area to the left of a point z (look at page 624-625 in the textbook.)
- We have: $P(a \leq X \leq b) = P(X \leq b) - P(X \leq a)$
- We don't have to worry about whether $P(X \leq a)$ or $P(X < a)$, these two probabilities are the same.

- Some probability distributions are tabulated for the area to the right of a point b , e.g:
- The t-distribution
- The Chi-square distribution
- The F – distribution
- The normal and t distribution are symmetric
- The Chi-square and F distribution are not symmetric.

- A continuous random variable X has a mean $E(X)$, a variance $\text{Var}(X)$ and a standard deviation $\sqrt{\text{Var}(X)}$
- $\mu = E(X)$ is the balance point of the probability mass. If we have a symmetric distribution, then $E(X)$ is the point of symmetry = μ is the same as the median.
- The median Q_2 is: $P(X \leq Q_2) = 0.5 = P(X \geq Q_2)$

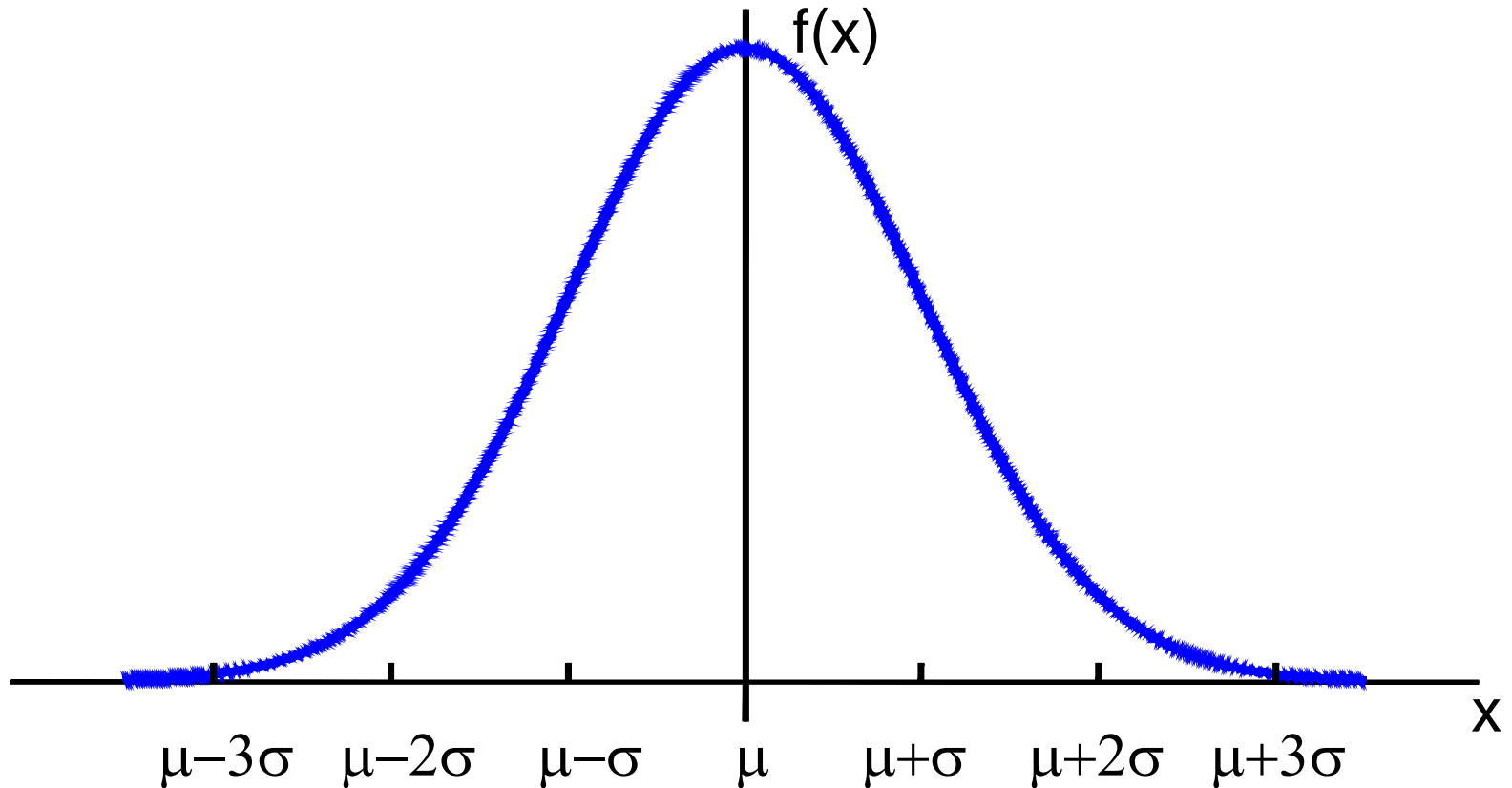
- If the distribution is not symmetric, then the median and $E(X)$ are usually different.
- We have: If Q_1 is the first quartile (25th percentile) then: $P(X \leq Q_1) = 0.25$
- If Q_3 is the third quartile then:
 $P(X \leq Q_3) = 0.75$

- A random variable can be standardized:

$$Z = \frac{X - E(X)}{\sqrt{\text{Var}(X)}}$$

- Then $E(Z) = 0$ and $\text{Var}(Z) = 1$

The Normal distribution.



X has a normal distribution with $E(X) = \mu$ and $\text{Var}(X) = \sigma^2$

The distribution of x can be given by this formula:

$$f(x) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(x-\mu)^2}{2\sigma^2}},$$

$$-\infty < x < \infty$$

$$-\infty < \mu < \infty \quad \text{og} \quad \sigma^2 > 0.$$

- We have:
- $P(\mu - \sigma \leq X \leq \mu + \sigma) = 0.6826$
- $P(\mu - 2\sigma \leq X \leq \mu + 2\sigma) = 0.9544$
- $P(\mu - 3\sigma \leq X \leq \mu + 3\sigma) = 0.9974$
- If X has a normal distribution with $E(X) = \mu$ and $\text{Var}(X) = \sigma^2$ then we write:
- $X \sim N(\mu, \sigma)$
- If $X \sim N(0, 1)$ then X has a standard normal distribution.

Probability calculations with the standard normal distribution.

- Assume $Z \sim N(0,1)$
- $P(Z < 0.83) = \underline{0.7967}$, look at page 625.
- $P(Z < -1.03) = \underline{0.1515}$
- $P(Z > 1.03) = 1 - P(Z \leq 1.03) = 1 - 0.8485 = \underline{0.1515}$

Use of the table in the opposite direction.

- $P(Z \leq a) = 0.20$ gives $a \approx -0.84$
- $P(Z \geq b) = 0.125 \iff P(Z \leq b) = 1 - 0.125 = 0.875$ and $b \approx 1.15$
- $P(-a \leq Z \leq a) = 0.95 = P(Z \leq a) - P(Z \leq -a) = P(Z \leq a) - (1 - P(Z \geq a)) = 2P(Z \leq a) - 1 \iff$
- $2P(Z \leq a) = 1.95 \iff P(Z \leq a) = 0.975$ gives
- $a = 1.96$

Probability calculations with normal distributions.

- If $X \sim N(\mu, \sigma)$ then

$$Z = \frac{X - E(X)}{\sqrt{\text{Var}(X)}} = \frac{X - \mu}{\sigma} \sim N(0, 1)$$

$$P(a \leq X \leq b) = P\left(\frac{a - \mu}{\sigma} \leq Z \leq \frac{b - \mu}{\sigma}\right)$$

Example 14.

- Let X = the height of people in a population (given in cm)
- Assume $X \sim N(170, 10)$
- $P(155 \leq X \leq 180) = P\left(\frac{155-170}{10} \leq Z \leq \frac{180-170}{10}\right)$

$$= P(-1.5 \leq Z \leq 1) = P(Z \leq 1) - P(Z \leq -1.5) = 0.8413 - 0.0668 = \underline{0.7745}$$

The Normal approximation to the Binomial.

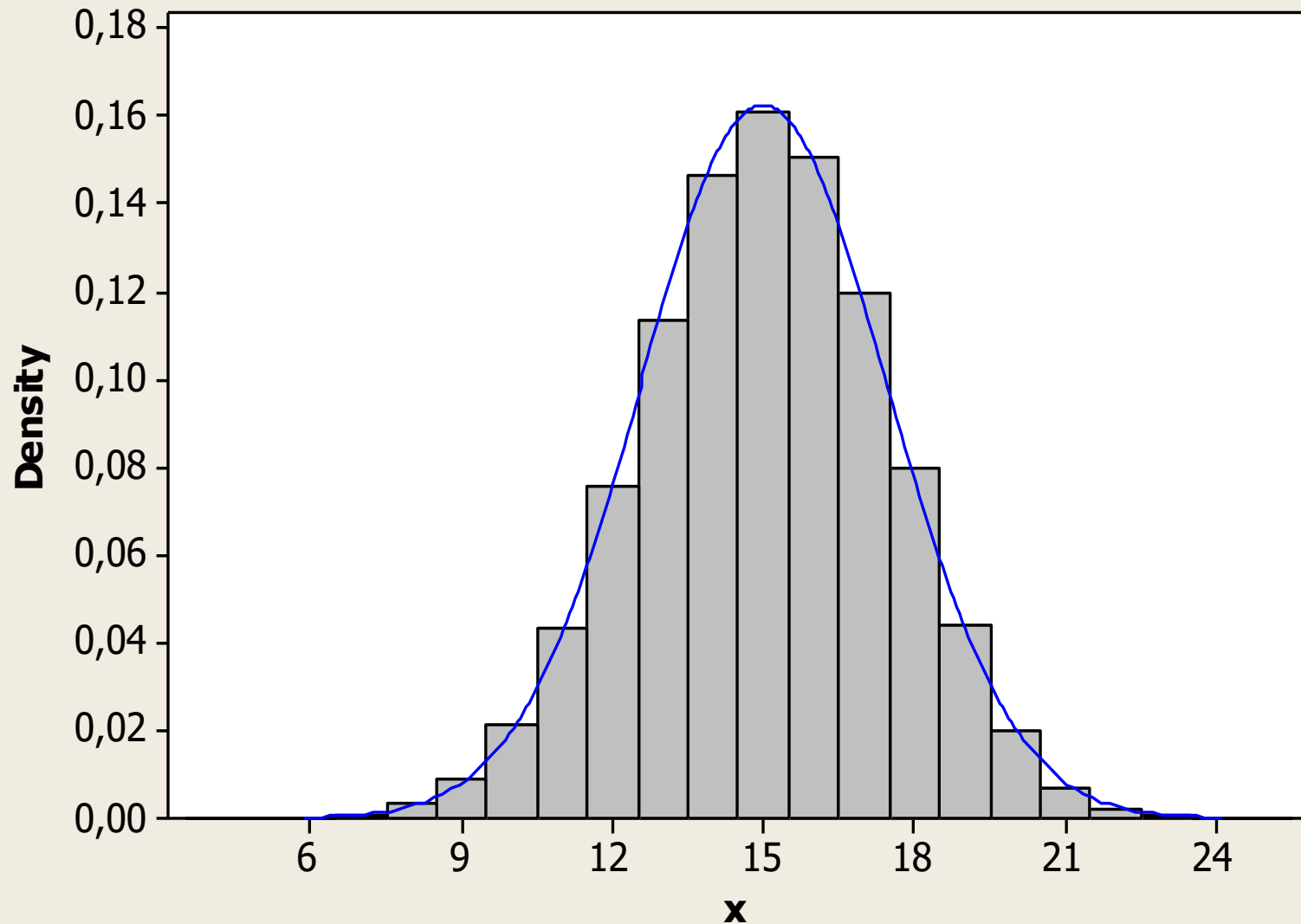
- Assume X has a binomial distribution, $n > 20$ and p is not too near 0 or 1.
- Then the binomial distribution can be approximated by a normal distribution.

- Let $n = 25$ and $p = 0.6$
- $E(X) = np = 25 \cdot 0.6 = 15$

$$\sqrt{\text{Var}(X)} = \sqrt{np(1-p)} = \sqrt{25 \cdot 0.6(1-0.6)} = 2.45$$

- X is approximately $N(15, 2.45)$

Probability histogram for the binomial distribution when $n=25$ and $p=0.6$



ean
Dev

- $P(X=12) = B_{25}(12,0.6) - B_{25}(11,0.6) = 0.154 - 0.078 = \underline{0.076}$

- $P(X=12) = P(11.5 < X < 12.5) =$

$$P\left(\frac{11.5 - 15}{2.45} < \frac{X - 15}{2.45} < \frac{12.5 - 15}{2.45}\right)$$

$$\approx P\left(\frac{-3.5}{2.45} < Z < \frac{-2.5}{2.45}\right) = P(-1.43 < Z < -1.02)$$

$$= P(Z < -1.02) - P(Z < -1.43) = 0.1539 - 0.0764 = \underline{0.0775}$$

- If X has a binomial distribution and $n > 20$ and p is not too near 0 or 1 then:

- $Z = \frac{X - np}{\sqrt{np(1-p)}}$ is approximately $N(0,1)$

- We have: $P(X \leq x) = P\left(\frac{X - np}{\sqrt{np(1-p)}} \leq \frac{x - np}{\sqrt{np(1-p)}}\right)$
 $\approx P\left(Z \leq \frac{x + 0.5 - np}{\sqrt{np(1-p)}}\right)$

The correction constant is 0.5

- If n is very large, we don't need the continuity correction 0.5.
- To approximate the binomial distribution to the normal distribution, we should have:
 $np \geq 5$ and $n(1-p) \geq 5$ (the textbook suggests 15)
- In our example: $n=25$ and $p=0.6$, $np=15$ and $n(1-p)=25(1-0.6) = 10$

Chapter 7. Sampling Distributions

The probability distribution of a statistic is called its sampling distribution.

A statistic can be a sampling mean.

A statistic can be a sample standard deviation.

- We assume: Our hypothetical observations have a distribution $f(x)$. All observations have: $E(X) = \mu$ and $\text{Var}(X) = \sigma^2$.
- Let \bar{X} = the sample mean.
- n = the sample size.
- It can be shown: $E(\bar{X}) = \mu$
- If the observations in the sample are independent, then:
- $\text{Var}(\bar{X}) = \frac{\sigma^2}{n}$ and $\sqrt{\text{Var}(\bar{X})} = \text{sd}(\bar{X}) = \frac{\sigma}{\sqrt{n}}$

We have:

- Samples: Sample mean
- Sample 1: observations \bar{X}_1
- Sample 2: observations \bar{X}_2
- Sample k: observations \bar{X}_k
- The sample means have a distribution.

- If the population has a normal distribution, then the sample mean \bar{X} has a normal distribution
- That is: X_1, X_2, \dots, X_n are independent and $N(\mu, \sigma)$ then \bar{X} is

$$N(\mu, \frac{\sigma}{\sqrt{n}})$$

The central limit theorem.

- If the observations follow the same distribution with mean μ and standard deviation σ and n is large ($n > 30$), then \bar{X} has an approximately normal distribution with
- $E(\bar{X}) = \mu$ and
$$\text{Var}(\bar{X}) = \frac{\sigma^2}{n}$$

- Then: $Z = \frac{\bar{X} - \mu}{\frac{\sigma}{\sqrt{n}}}$ is approximately $N(0,1)$
- **Example 15:** assume we have 36 observations from a distribution with $\mu = 25$ and $\sigma = 12$.
- Calculate $P(26 \leq \bar{X} \leq 29)$

$$P\left(\frac{26-25}{\frac{12}{\sqrt{36}}} \leq \frac{\bar{X}-25}{\frac{12}{\sqrt{36}}} \leq \frac{29-25}{\frac{12}{\sqrt{36}}}\right) \approx P\left(\frac{1}{2} \leq Z \leq 2\right) = 0.9772 - 0.6915 = 0.2857$$

Chapter 8. Drawing inferences from large samples.

Statistical inference deals with drawing conclusions about population parameters from an analysis of the sample data.

If we have a dataset drawn from a population, we can calculate descriptive statistics from it.

- We can calculate:
- The sample mean = \bar{x}
- The sample standard deviation =

$$s = \sqrt{\frac{1}{n-1} \sum_{i=1}^n (x_i - \bar{x})^2}$$

- The median = Q_2
- First quartile = Q_1
- Third quartile = Q_3

- If we want to find out something more about the distribution of the population, we might want to find $\mu = E(X)$.
- We could want to:
 - 1) Estimate μ (point estimation)
 - 2) Find an interval for μ (confidence interval)
 - 3) Test if μ has a specific value (testing a hypothesis)

- A statistic intended for estimating a parameter is called a point estimator, or simply an estimator.
- The standard deviation of an estimator is called its standard error: SE
- If we want to estimate μ then we usually will choose $\hat{\mu} = \bar{X}$. \bar{X} will be a point estimate of μ . This estimate of μ might be uncertain.

- $SE(\bar{X}) = \frac{\sigma}{\sqrt{n}}$ is the standard deviation of \bar{X}
- If we know σ , this will tell us about the variability in the distribution of the estimator $\hat{\mu} = \bar{X}$
- Usually we don't know σ . If we have a large sample, that is: n is large, then we use $\frac{S}{\sqrt{n}}$ as the standard deviation of \bar{X}

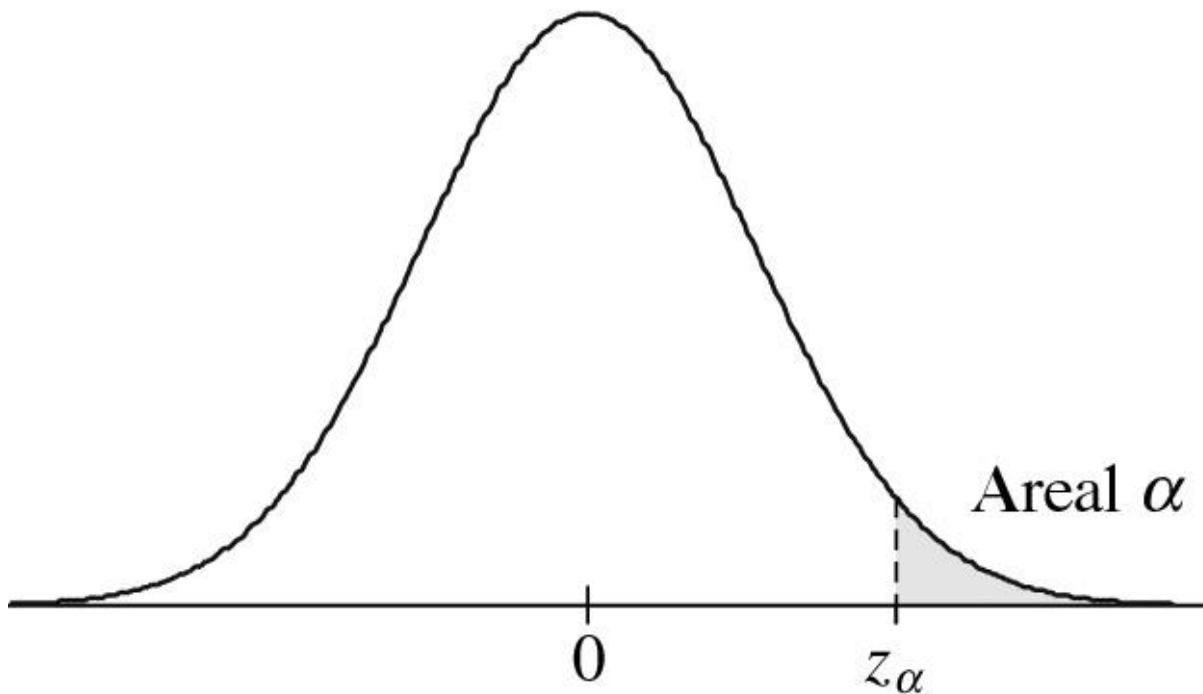
Strictly speaking, $\frac{S}{\sqrt{n}}$ is an estimate of $SE(\bar{X})$.

Confidence interval for the population mean μ .

- We want to construct an interval of values which is likely to contain the true value of μ .
- We assume: X_1, X_2, \dots, X_n are independent and $N(\mu, \sigma)$ where σ is known.
- Before sampling, we want to find an interval which covers μ with the probability $1 - \alpha$. α is a small number.

α can be 0.01, 0.05, 0.10 or some other small number.

The standard normal distribution:



| α | z_α |
|----------|------------|
| 0.100 | 1.282 |
| 0.050 | 1.645 |
| 0.025 | 1.960 |
| 0.010 | 2.326 |
| 0.005 | 2.576 |
| 0.001 | 3.090 |

- We have: $\bar{X} \sim N(\mu, \frac{\sigma}{\sqrt{n}})$

$$1 - \alpha = P \left(-z_{\frac{\alpha}{2}} \leq \frac{\bar{X} - \mu}{\frac{\sigma}{\sqrt{n}}} \leq z_{\frac{\alpha}{2}} \right)$$

$$= P \left(\bar{X} - z_{\frac{\alpha}{2}} \cdot \frac{\sigma}{\sqrt{n}} \leq \mu \leq \bar{X} + z_{\frac{\alpha}{2}} \cdot \frac{\sigma}{\sqrt{n}} \right)$$

- A $100(1-\alpha)\%$ confidence interval for μ is given by:

$$\left[\bar{X} - z_{\frac{\alpha}{2}} \cdot \frac{\sigma}{\sqrt{n}}, \bar{X} + z_{\frac{\alpha}{2}} \cdot \frac{\sigma}{\sqrt{n}} \right]$$

- Before we take the sample, this is a random interval.
- When we have collected a sample we can calculate a 90%, a 95% or a 99% confidence interval for μ .

Example 16.

- We have 25 observations.
- They are independent $N(\mu, \sigma)$ where σ is known to be 8. $\bar{X} = 42.7$, and we will construct a 90% confidence interval for μ .

$$42.7 \pm 1.645 \cdot \frac{8}{\sqrt{25}} = 42.7 \pm 2.632$$

- $42.7 - 2.632 = 40.068 \approx 40.07$
- $42.7 + 2.632 = 45.332 \approx 45.33$

- $[40.07, 45.33]$ is a 90% confidence interval for μ .
- The method is so that the chance is 90% that this interval will cover μ .
- In most cases we don't know σ . But if we have a large sample, we have:
- A $100(1-\alpha)\%$ confidence interval for μ is given by:

$$\left[\bar{X} - z_{\frac{\alpha}{2}} \cdot \frac{S}{\sqrt{n}}, \bar{X} + z_{\frac{\alpha}{2}} \cdot \frac{S}{\sqrt{n}} \right]$$

- In the formula, S is the sample standard deviation.
- When we have a large sample, we know that \bar{X} is approximately

$$N(\mu, \frac{\sigma}{\sqrt{n}})$$

even if the observations in the sample don't have a normal distribution.

- We also know that S is a very good estimator of σ .

The error margin.

- The error margin is one half of the length of the confidence interval.
- We want to decide the sample size so that our confidence interval will have length $2d$.

That is:

$$z_{\frac{\alpha}{2}} \frac{\sigma}{\sqrt{n}} = d$$

- This gives: $n = \left(\frac{z_{\frac{\alpha}{2}} \cdot \sigma}{d} \right)^2$

Example 17.

- Assume $\sigma = 8$ and $\alpha = 0.1$, $z_{\frac{\alpha}{2}} = z_{0.05} = 1.645$
- $d = 1$ (the length of the interval is 2)

$$\left(\frac{1.645 \cdot 8}{1} \right)^2 = 173.19$$

- We have to choose $n = 174$ to get the length 2 of the 90% confidence interval for μ .
- This is the same as specifying the 90% error margin to 1.

- We can calculate 90%, 95% and 99% confidence intervals for μ after collecting a sample. The level of confidence is then 0.90, 0.95 or 0.99 respectively.
- The interval with level of confidence 0.90 will be shorter than the interval with level of confidence 0.95.
- If the level of confidence increases, then the length of the interval also increases.

Definition of a confidence interval for a parameter.

- An interval (L,U) is a $100(1-\alpha)\%$ confidence interval for a parameter if
- $P(L \leq \text{parameter} \leq U) = 1 - \alpha$
- and the endpoints L and U are computable from the sample.

Testing hypotheses about a population mean μ .

- We have two hypotheses called H_0 and H_1 .
- H_0 = the null hypothesis.
- H_1 = the alternative hypothesis.
- We want to say that H_1 is correct, but in many cases we will not be able to do so.
- Then we must say that we accept H_0 .

- In most cases we will state that μ has a value μ_0 in our null hypothesis. The alternative hypothesis can be:

$$\mu \neq \mu_0, \mu < \mu_0 \text{ or } \mu > \mu_0$$

- The first: $\mu \neq \mu_0$ is a two sided alternative.

$\mu < \mu_0$ or $\mu > \mu_0$ are one sided alternatives.

- If we want to examine if there has been a change, we will state in H_0 that no change has appeared.

- In H_1 it will be stated that there has been a change.
- If someone claims that things are in a certain way, and we doubt this and want to say: no it is not, we will put our claim in H_1 .
- H_0 will contain the assertion we don't believe in.
- We must decide whether we believe in H_0 or H_1 .

- From a trial we will make a decision rule for what to do.
- If we want to test $H_0: \mu = \mu_0$ against $H_1: \mu < \mu_0$ we will reject H_0 if $\bar{X} \leq c = \text{a constant}$.
- If we have: $\bar{X} > c$ we will retain H_0 .
- We will determine c such that there is just a small probability α of rejecting H_0 if H_0 is true. α is the level of significance for the test

- We have: $\alpha = P(\bar{X} \leq c \mid H_0 \text{ is true}) =$

$$P\left(\frac{\bar{X} - \mu_0}{\frac{\sigma}{\sqrt{n}}} \leq \frac{c - \mu_0}{\frac{\sigma}{\sqrt{n}}} \mid H_0 \text{ is true}\right) = P\left(Z \leq \frac{c - \mu_0}{\frac{\sigma}{\sqrt{n}}} \mid H_0 \text{ is true}\right)$$

$$\frac{c - \mu_0}{\frac{\sigma}{\sqrt{n}}} = -z_\alpha \Leftrightarrow c = \mu_0 - z_\alpha \cdot \frac{\sigma}{\sqrt{n}}$$

If σ is unknown and $n > 30$ we will replace σ by s in the formulas.

Example 18.

- Let X = the time it takes for a customer to get cash out of a teller machine.
- We know from previous experience that $\mu = E(X) = 270$ seconds and $\sigma = 24$ seconds.
- A vendor suggests that an upgrade of the machine will make $\mu < 270$.
- We will examine if he is right.

- We test $H_0: \mu = 270$ against $H_1: \mu < 270$ at a 5% level of significance.
- We do a trial with an upgraded machine. 38 observations are taken and we find: $\bar{X} = 261$
- We assume that $\sigma = 24$ as before, but μ may have changed.
- We find: $c = 270 - 1.645 \cdot \frac{24}{\sqrt{38}} = 263.60$
- We have: $\bar{X} < 263.6$ and we reject H_0 .
- We will recommend an upgrade.

- \bar{X} is called a test statistic.

α = the level of significance =
 $P(\text{rejecting } H_0 | H_0 \text{ is true}) = P(\text{doing type I error})$

- We have chosen α very small. Usually we choose $\alpha = 0.01, 0.05$ or 0.10
- We want a small chance of doing type I error.

- In hypotheses testing we also can do type II error, that is: retaining H_0 if H_1 is true.
- We have: $\beta = P(\text{doing type II error}) = P(\text{retaining } H_0 | H_1 \text{ is true})$.
- This probability depends on the value of μ , and μ can take a value in a region. We don't have a specific value to put into the formula. But, unfortunately β can be large. We have no control on β .

- If we can reject H_0 we know the chance of doing an error (of type I). But if we retain H_0 we do not know the chance of doing an error (of type II).
- If we retain H_0 , it is possible that H_0 is true. But if H_0 is not true, the sample size is too small to be able to assert H_1 .
- We can try to increase the sample size if possible, if we believe in H_1 .

P-value.

- P-value = the probability that the test statistic takes a value equal to or more extreme than the observed value calculated under H_0 .
- In our example 18: $\bar{X} = 261$
- We find: $P(\bar{X} \leq 261 | H_0 \text{ is true})$

$$\bullet = P\left(\frac{\bar{X} - 270}{\frac{24}{\sqrt{38}}} \leq \frac{261 - 270}{\frac{24}{\sqrt{38}}} \mid H_0 \text{ is true} \right) = P(Z \leq -2.31)$$

$$= 0.0104 = \text{p-value}$$

- If the p-value is smaller than α then we will reject H_0 .
- \bar{X} is a test statistic, and
- $Z = \frac{\bar{X} - \mu_0}{\frac{\sigma}{\sqrt{n}}}$ is a test statistic if we know σ .
- If we don't know σ , we will use
- $Z = \frac{\bar{X} - \mu_0}{\frac{s}{\sqrt{n}}}$ as a test statistic.

- If $n > 30$ and H_0 is true, $Z \sim N(0,1)$ approximately
- Usually we will calculate Z and compare the value to a value in the standard normal probability table.
- At the α level of significance we can test:
 - $H_0: \mu = \mu_0$ against $H_1: \mu < \mu_0$
 - Reject H_0 if $Z \leq -z_\alpha$

- or: $H_0: \mu = \mu_0$ against $H_1: \mu > \mu_0$
- Reject H_0 if $Z \geq z_\alpha$
- or: $H_0: \mu = \mu_0$ against $H_1: \mu \neq \mu_0$
- Reject H_0 if $|Z| \geq z_{\frac{\alpha}{2}}$
- Assume we test the last hypotheses.
- We calculate $|Z| = t$ from a trial. We now find:
- $$\text{P-value} = P(|Z| \geq t | \mu = \mu_0) = P(Z \leq -t | \mu = \mu_0) + P(Z \geq t | \mu = \mu_0) = 2P(Z \leq -t | \mu = \mu_0)$$

Example 18 revisited.

- Assume we are testing $H_0: \mu=270$ against $H_1: \mu \neq 270$ and found $\bar{X} = 261$ and $\sigma = 24$

$$|Z| = \left| \frac{261 - 270}{\frac{24}{\sqrt{38}}} \right| = 2.31$$

- P-value = $P(|Z| \geq 2.31 | H_0 \text{ is true}) =$
- $2P(Z \leq -2.31 | H_0 \text{ is true}) = 2 \cdot 0.0104 = \underline{\underline{0.0208}}$

- We have: $p\text{-value} = 0.0208 < 0.05$ and we reject H_0 with the level of significance $= 0.05$.
- We can have a 0.03 level of significance and still reject H_0 .
- We can say: The chance of a false conclusion is less than 3%.
- If we can reject H_0 at the level of significance α then the probability of drawing the wrong conclusion is at most α .

- We call this test: a large sample normal test or a Z-test.
- Steps for testing hypotheses:
 - 1) Formulate H_0 and H_1 .
 - 2) State the test statistic and calculate it from the data.
 - 3) Determine the rejection region.
 - 4) Draw a conclusion. Tell the probability of a false conclusion if you can.

Inferences about a population proportion.

Assume $X \sim \text{bin}(n, p)$, it has a binomial distribution. We have:
 $\hat{p} = \frac{X}{n}$ is an estimator of p . \hat{p} is a statistic.

$$E(\hat{p}) = \frac{E(X)}{n} = \frac{np}{n} = p$$

$$\text{Var}(\hat{p}) = \frac{\text{Var}(X)}{n^2} = \frac{np(1-p)}{n^2} = \frac{p(1-p)}{n}$$

$$SE(\hat{p}) = \sqrt{\text{Var}(\hat{p})} = \sqrt{\frac{p(1-p)}{n}}$$

If n is large:
$$Z = \frac{\hat{p} - p}{\sqrt{\frac{p(1-p)}{n}}} = \frac{X - np}{\sqrt{np(1-p)}}$$

is approximately $N(0,1)$.

We can estimate

$$SE(\hat{p}) \text{ by } \sqrt{\frac{\hat{p}(1-\hat{p})}{n}}$$

A large sample confidence interval for p.

- For large n, a $100(1-\alpha)\%$ confidence interval for p is given by:

$$\left[\hat{p} - z_{\frac{\alpha}{2}} \sqrt{\frac{\hat{p}(1-\hat{p})}{n}}, \hat{p} + z_{\frac{\alpha}{2}} \sqrt{\frac{\hat{p}(1-\hat{p})}{n}} \right]$$

Example 19:

- We have $n = 100$ and $X = 80$. $\hat{p} = 0.80$

$$z_{0.025} \sqrt{\frac{0.8 \cdot 0.2}{100}} = 1.96 \cdot 0.04 = 0.0784$$

- $[0.80 - 0.0784, 0.80 + 0.0784]$
- $[0.7216, 0.8784]$ is a 95% large sample confidence interval for p .

Determining the sample size.

- An approximate $100(1-\alpha)\%$ error margin is:

$$z_{\frac{\alpha}{2}} \sqrt{\frac{p(1-p)}{n}} \leq z_{\frac{\alpha}{2}} \sqrt{\frac{0.5 \cdot 0.5}{n}} = z_{\frac{\alpha}{2}} \frac{0.5}{\sqrt{n}} = \frac{z_{\frac{\alpha}{2}}}{2\sqrt{n}}$$

- If we want the $100(1-\alpha)\%$ error margin to be d , then we choose n so that:

$$d = \frac{z_{\frac{\alpha}{2}}}{2\sqrt{n}} \quad \text{which gives: } n = \frac{1}{4} \left(\frac{z_{\frac{\alpha}{2}}}{d} \right)^2$$

- If we want a 95% error margin to be 0.08:
(think about the seeds):
- $d = 0.08 \quad z_{0.025} = 1.96$
- $n = \frac{1}{4} \left(\frac{1.96}{0.08} \right)^2 = 150.0625$
- We must choose $n = 151$ to be sure that the 95% error margin is not more than 0.08.
Then we also can be sure that our confidence interval will be no longer than 0.16.

- When we construct a confidence interval for p , we might get:
- The lower limit < 0 or the upper limit > 1 .
- We then truncate the interval so it will not be outside the interval $[0,1]$

Large sample tests about p.

- If n is large: $\hat{p} = \frac{X}{n} \sim N(p, \sqrt{\frac{p(1-p)}{n}})$ approximately
- We will test $H_0: p = p_0$ against $H_1: p \neq p_0$.
- $Z = \frac{\hat{p} - p_0}{\sqrt{\frac{p_0(1-p_0)}{n}}} \sim N(0,1)$ approximately if H_0

is true. We reject H_0 at the α level of significance if $|Z| \geq z_{\frac{\alpha}{2}}$

- We will test:
- $H_0: p = p_0$ against $H_1: p > p_0$.
- We reject H_0 at the α level of significance if $Z \geq z_\alpha$.

- We will test:
- $H_0: p = p_0$ against $H_1: p < p_0$.
- We reject H_0 at the α level of significance if $Z \leq -z_\alpha$.

Example 19 revisited.

- $n = 100$ and $X = 80$
- We want to test $H_0: p = 0.90$ against $H_1: p \neq 0.90$.

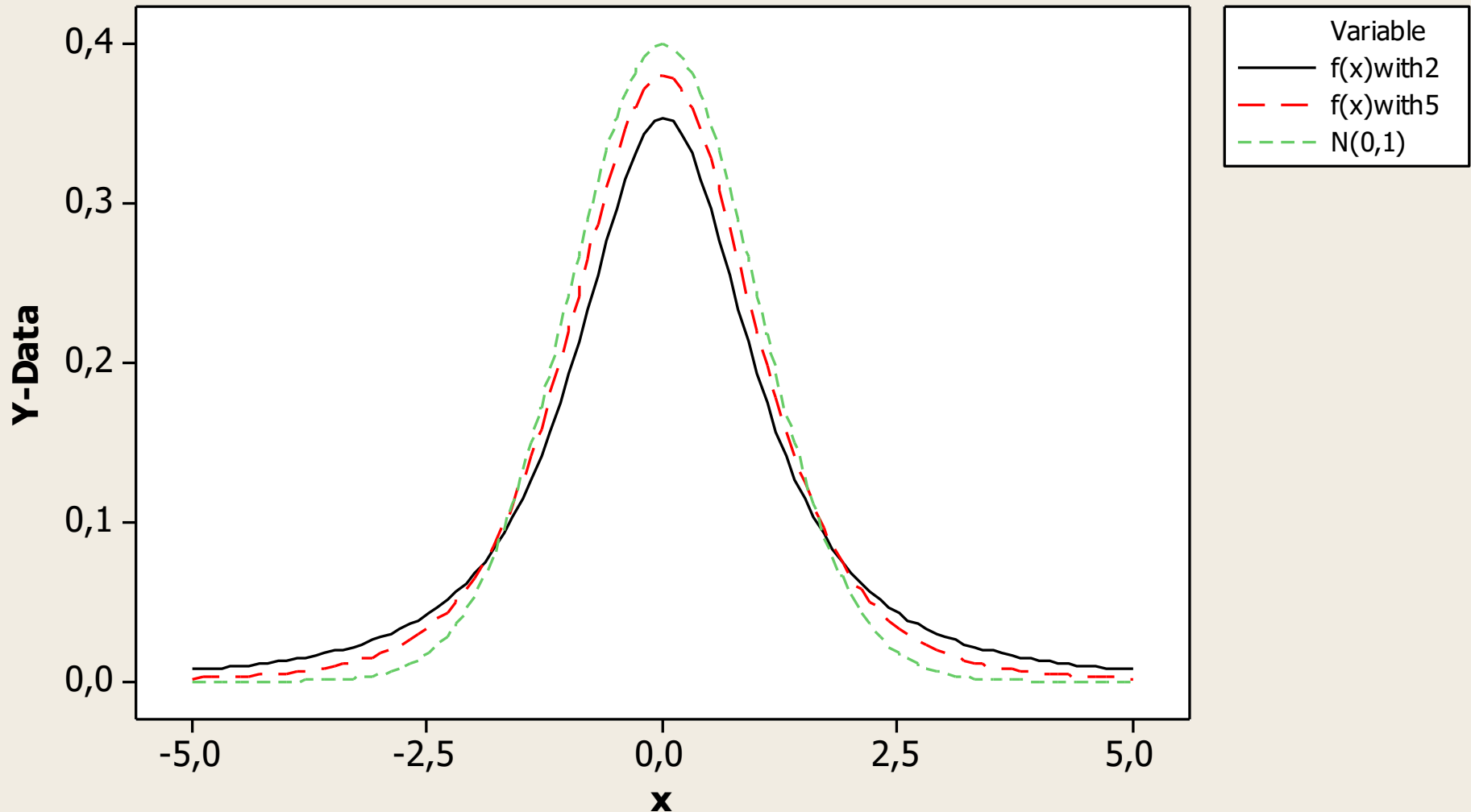
$$Z = \frac{0.8 - 0.9}{\sqrt{\frac{0.9(1 - 0.9)}{100}}} = -3.33$$

- $|Z| = 3.33 \geq z_{0.025} = 1.96$ and we reject H_0 at a 5% level of significance.

- A 95% confidence interval for p can be used to test $H_0: p = 0.90$ against $H_1: p \neq 0.90$.
- We found the interval $[0.7216, 0.8784]$.
- 0.90 is not in the interval, so we can reject H_0 at a 5% level of significance.

Chapter 9. Students t-distribution.

**t-distributions with 2 or 5 degrees of freedom and
the standard normal distribution.**



- The Students t-distribution is symmetric around 0.
- We have to know the degrees of freedom.
- There is one distribution for each value of the degrees of freedom: 1, 2, , infinite.
- The Students t-distribution with infinite degrees of freedom is the same as $N(0,1)$.

- $T = \frac{\bar{X} - \mu}{\frac{S}{\sqrt{n}}}$ has a Students t-distribution with
- $n-1$ degrees of freedom if X_1, X_2, \dots, X_n are independent and $N(\mu, \sigma)$ with σ unknown.
- A $100(1-\alpha)\%$ confidence interval for $\mu = E(X)$ is:

$$\left[\bar{X} - t_{\frac{\alpha}{2}, n-1} \cdot \frac{S}{\sqrt{n}}, \bar{X} + t_{\frac{\alpha}{2}, n-1} \cdot \frac{S}{\sqrt{n}} \right]$$

Example 20.

- Let X = the weight of a pack of butter.
- We observe 8 packs of butter and assume: the weights are independent and $N(\mu, \sigma)$.
- From our sample we calculate $\bar{X} = 503\text{g}$ and $S = 11.82\text{g}$.
- We want to calculate a 95% confidence interval for μ .

We find in the t-table: $t_{0.025,7} = 2.365$

$$\text{Lower limit} = \bar{X} - t_{0.025,7} \cdot \frac{S}{\sqrt{n}} = 503 - \frac{2.365 \cdot 11.82}{\sqrt{8}}$$

$$= 503 - 9.88 = \underline{493.12}$$

$$\text{Upper limit} = \bar{X} + t_{0.025,7} \cdot \frac{S}{\sqrt{n}} = 503 + \frac{2.365 \cdot 11.82}{\sqrt{8}}$$

$$= 503 + 9.88 = \underline{512.88}$$

[493.12 , 512.88] is a 95% confidence interval for μ .

- We can now test with level of significance α :
- $H_0: \mu = \mu_0$ against $H_1: \mu \neq \mu_0$.
- Calculate $t = \frac{\bar{X} - \mu_0}{\frac{s}{\sqrt{n}}}$
- Reject H_0 if $|t| \geq t_{\frac{\alpha}{2}, n-1}$

- Test: $H_0: \mu = \mu_0$ against $H_1: \mu > \mu_0$.
- Reject H_0 if $t \geq t_{\alpha, n-1}$
- Test: $H_0: \mu = \mu_0$ against $H_1: \mu < \mu_0$.
- Reject H_0 if $t \leq -t_{\alpha, n-1}$
- We assume that X_1, X_2, \dots, X_n are independent and $N(\mu, \sigma)$ with σ unknown.

- If the observations are far from normal, we can increase n , or conduct a nonparametric test.
- **Example 20** revisited: There are 8 packs of butter.
- We test: $H_0: \mu = 500$ against $H_1: \mu \neq 500$

$$|t| = \left| \frac{\bar{x} - \mu_0}{\frac{s}{\sqrt{n}}} \right| = \left| \frac{503 - 500}{\frac{11.82}{\sqrt{8}}} \right| = 0.72 < 2.365 = t_{0.025, 7}$$

- We retain H_0 at a 5% level of significance.

Chapter 10. Comparing two treatments.

There are two populations. We draw a random sample from each.

Sample 1 (from population 1): X_1, X_2, \dots, X_{n_1}

Sample 2 (from population 2): Y_1, Y_2, \dots, Y_{n_2}

We can calculate: \bar{X} $S_1 = \sqrt{\frac{1}{n_1 - 1} \sum_{i=1}^{n_1} (X_i - \bar{X})^2}$

$$\bar{Y} \quad S_2 = \sqrt{\frac{1}{n_2 - 1} \sum_{j=1}^{n_2} (Y_j - \bar{Y})^2}$$

- We assume: X_1, X_2, \dots, X_{n_1} are $N(\mu_1, \sigma)$
- Y_1, Y_2, \dots, Y_{n_2} are $N(\mu_2, \sigma)$
- All observations are independent.

□ μ_1 is estimated by $\hat{\mu}_1 = \bar{X}$ and μ_2 by $\hat{\mu}_2 = \bar{Y}$

- The common σ is estimated by:

$$S_p = S_{\text{pooled}} = \sqrt{\frac{\sum_{i=1}^{n_1} (X_i - \bar{X})^2 + \sum_{j=1}^{n_2} (Y_j - \bar{Y})^2}{n_1 + n_2 - 2}}$$

- If $n_1 = n_2$ then: $S_{\text{pooled}} = \sqrt{\frac{S_1^2 + S_2^2}{2}}$
- We can assume equal σ for the two populations if
$$\frac{1}{2} \leq \frac{S_1}{S_2} \leq 2$$
- It can be shown: $E(\bar{X}) - E(\bar{Y}) = \mu_1 - \mu_2$

- $T = \frac{\bar{X} - \bar{Y} - (\mu_1 - \mu_2)}{S_{\text{pooled}} \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}}$ has a t-distribution

- with $n_1 + n_2 - 2$ degrees of freedom if our assumptions are valid.

- A small sample $100(1-\alpha)\%$ confidence interval for $\mu_1 - \mu_2$ is:

$$\left[\bar{X} - \bar{Y} - t_{\frac{\alpha}{2}, (n_1+n_2-2)} S_p \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}, \bar{X} - \bar{Y} + t_{\frac{\alpha}{2}, (n_1+n_2-2)} S_p \sqrt{\frac{1}{n_1} + \frac{1}{n_2}} \right]$$

- **Example 21:** Out of 21 pigs of the same age and race we select 10 randomly. They are given diet A. The other 11 pigs are given diet B. After a certain period of time we observe the increase in weights in kilos.
- The observations from diet A are: X_1, X_2, \dots, X_{10} . We assume they are: $N(\mu_1, \sigma)$.
- The observations from diet B are: Y_1, Y_2, \dots, Y_{11} . We assume they are: $N(\mu_2, \sigma)$.

- All observations are assumed to be independent.
- The observations are:
- | Row | dietA | dietB |
|-----|-------|-------|
| 1 | 11,8 | 15,4 |
| 2 | 15,4 | 10,0 |
| 3 | 11,3 | 17,7 |
| 4 | 13,6 | 14,0 |
| 5 | 12,2 | 13,6 |
| 6 | 10,9 | 15,9 |
| 7 | 7,7 | 12,2 |
| 8 | 8,6 | 11,3 |
| 9 | 14,5 | 13,2 |
| 10 | 10,9 | 8,6 |
| 11 | | 13,6 |

- We want to calculate a 95% confidence interval for $\mu_1 - \mu_2$: From the observations we calculate:

$$\bar{x} = 11.69 \quad \bar{y} = 13.23$$

$$s_1 = 2.414 \quad s_2 = 2.633$$

$$s_p = \sqrt{\frac{9 \cdot 2.414^2 + 10 \cdot 2.633^2}{19}} = 2.53$$

- In the t-table we find: $t_{0.025,19} = 2.093$

- Lower limit =

$$11.69 - 13.23 - 2.093 \cdot 2.53 \sqrt{\frac{1}{10} + \frac{1}{11}} = -3.85$$

- Upper limit =

$$11.69 - 13.23 + 2.093 \cdot 2.53 \sqrt{\frac{1}{10} + \frac{1}{11}} = 0.77$$

- $[-3.85, 0.77]$ is a 95% confidence interval for $\mu_1 - \mu_2$.

- The confidence interval can be used to test $H_0: \mu_1 - \mu_2 = \delta_0$ against $H_1: \mu_1 - \mu_2 \neq \delta_0$ at the α level of significance.
- We reject H_0 if δ_0 is outside the interval.
- In example 21 we test $H_0: \mu_1 - \mu_2 = 0$ against $H_1: \mu_1 - \mu_2 \neq 0$.
- We retain H_0 at a 5% level of significance because 0 is inside the interval.

- If we want to test
- $H_0: \mu_1 - \mu_2 = \delta_0$ against $H_1: \mu_1 - \mu_2 \neq \delta_0$ with the α level of significance and we don't have a $100(1-\alpha)\%$ confidence interval for $\mu_1 - \mu_2$

- we calculate:

$$t = \frac{\bar{x} - \bar{y} - \delta_0}{S_{\text{pooled}} \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}}$$

- We reject H_0 if $|t| \geq t_{\frac{\alpha}{2}, n_1 + n_2 - 2}$ = the upper $\frac{\alpha}{2}$

-percentile in the t-distribution with $n_1 + n_2 - 2$ degrees of freedom.

- To test: $H_0: \mu_1 - \mu_2 = \delta_0$ against $H_1: \mu_1 - \mu_2 > \delta_0$ with the α level of significance :
- Reject H_0 if $t \geq t_{\alpha, n_1 + n_2 - 2}$
- To test: $H_0: \mu_1 - \mu_2 = \delta_0$ against $H_1: \mu_1 - \mu_2 < \delta_0$ with the α level of significance :
- Reject H_0 if $t \leq -t_{\alpha, n_1 + n_2 - 2}$

Example 21 revisited.

- We test: $H_0: \mu_1 - \mu_2 = 0$ against $H_1: \mu_1 - \mu_2 < 0$ with confidence level 0.05.

$$t = \frac{11.69 - 13.23}{2.53 \sqrt{\frac{1}{10} + \frac{1}{11}}} = -1.39$$

- $t_{0.05, 19} = 1.729$
- $t > -1.729$ and we retain H_0 .

Two large samples.

- If we have two large samples, say n_1 and $n_2 > 30$, we can replace
- $S_p \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}$ and $t_{\frac{\alpha}{2}, n_1 + n_2 - 2}$ in the formula for the confidence interval by $\sqrt{\frac{S_1^2}{n_1} + \frac{S_2^2}{n_2}}$ and $z_{\frac{\alpha}{2}}$ respectively.

- In testing hypotheses, we will calculate

$$Z = \frac{\bar{x} - \bar{y} - \delta_0}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}}$$

- and compare to z in the standard normal distribution.
- But if we have n_1 and n_2 both >30 I guess we will rather use Minitab.
- In Minitab, the two populations can have different standard deviations.

Matched pairs comparisons.

- We want to compare two kinds of medical treatments, treatment A and B, and we have n pairs of identical twins.
- For each pair, we decide at random who is going to get treatment A and who is going to get treatment B.
- Let x_i = observed weight loss for person getting treatment A in pair number i .
- y_i = observed weight loss for person getting treatment B in pair number i .

- For each pair, we calculate $d_i = x_i - y_i$.
- Before we observe, we have the random variables D_1, \dots, D_n where $D_i = X_i - Y_i$
- We assume D_1, \dots, D_n are independent and $N(\delta, \sigma_D)$
- Now we have a 1-sample situation where the differences are treated as the observations. If we want to calculate a $100(1-\alpha)\%$ confidence interval for δ or test hypotheses about δ , we put the differences into the formulas and procedures for one population mean.

- This means:
- If the number of pairs is 30 or less, we will calculate a t interval or perform a t-test for δ .
- If the number of pairs is more than 30, we will calculate a z interval or perform a z-test for δ .
- We will often use matched pairs if we are growing something at different places.
- If we want to compare two kinds of seeds and we look at crop yields, we can arrange a trial like this:

| | |
|--------|--------|
| Kind A | Kind B |
|--------|--------|

Place 1

| | |
|--------|--------|
| Kind B | Kind A |
|--------|--------|

Place 2

| | |
|--------|--------|
| Kind B | Kind A |
|--------|--------|

Place 3

| | |
|--------|--------|
| Kind A | Kind B |
|--------|--------|

Place 4

This is an experiment in blocks.

- In every block it should be decided at random if kind A should be grown to the left or to the right.
- Climate and soil quality may be different from block to block.
- When we analyze d_1, \dots, d_n where $d_i = x_i - y_i$ we have eluded effects from climate and soil quality, and it will be easier to compare expected crop yields.

Example 22. Growing strawberries.

- We will compare expected crop yields for two sorts of strawberry plants. Both sorts are grown at 12 different locations. The fields are of the same size and there are approximately the same number of plants of each sort at each location.
- Let X_i = mean crop yield per plant in gram, sort 1, location i .
- Y_i = mean crop yield per plant in gram, sort 2, location i .

- The observations are:

| • | Row | location | x | y | d |
|---|-----|----------|-------|-------|-------|
| • | 1 | 1 | 330,5 | 289,8 | 40,7 |
| • | 2 | 2 | 299,0 | 300,0 | -1,0 |
| • | 3 | 3 | 334,5 | 310,2 | 24,3 |
| • | 4 | 4 | 307,7 | 317,0 | -9,3 |
| • | 5 | 5 | 351,0 | 340,6 | 10,4 |
| • | 6 | 6 | 318,3 | 323,4 | -5,1 |
| • | 7 | 7 | 338,1 | 312,0 | 26,1 |
| • | 8 | 8 | 321,2 | 304,0 | 17,2 |
| • | 9 | 9 | 344,0 | 316,8 | 27,2 |
| • | 10 | 10 | 301,4 | 322,7 | -21,3 |
| • | 11 | 11 | 331,6 | 311,1 | 20,5 |
| • | 12 | 12 | 348,3 | 323,9 | 24,4 |

- Calculations give: $\bar{d} = 12.8417$

$$s_D = \sqrt{\frac{1}{n-1} \sum_{i=1}^n (d_i - \bar{d})^2} = 18.28$$

- We want to calculate a 95% confidence interval for $\delta = E(D_i)$
- Lower limit =

$$\bar{d} - t_{\frac{\alpha}{2}, n-1} \frac{s_D}{\sqrt{n}} = 12.8417 - 2.201 \cdot \frac{18.28}{\sqrt{12}} = 1.23$$

- Upper limit =

$$\bar{d} + t_{\frac{\alpha}{2}, n-1} \frac{s_D}{\sqrt{n}} = 12.8417 + 2.201 \cdot \frac{18.28}{\sqrt{12}} = 24.46$$

- $[1.23, 24.46]$ is a 95% confidence interval for δ .
- The interval can be used for testing
- $H_0: \delta = 0$ against $H_1: \delta \neq 0$
- 0 is outside the interval, and we reject H_0 at a 5% level of significance.
- We test $H_0: \delta = 0$ against $H_1: \delta > 0$

$$t = \frac{\bar{d}}{\frac{s_D}{\sqrt{n}}} = \frac{12.8417}{\frac{18.28}{\sqrt{12}}} = 2.4335 > 1.796 = t_{0.05,11}$$

- We reject H_0 at a 5% level of significance.
- We state that sort 1 has a greater mean crop yield than sort 2.

10.6 Comparing two population proportions.

We have two populations. A random sample of size n_1 is taken from the first population and a random sample of size n_2 is taken from the second population.

For each sample unit we look for an event A .

$P(A) = p_1$ in the first population and

$P(A) = p_2$ in the second population.

Each sample unit is assumed to have A or not independently of the other units.

- Let X = the number of units having A in the first sample.
- Let Y = the number of units having A in the second sample.
- We estimate p_1 by $\hat{p}_1 = \frac{X}{n_1}$ and p_2 by $\hat{p}_2 = \frac{Y}{n_2}$
-
- It can be shown:

$$SE(\hat{p}_1 - \hat{p}_2) = \sqrt{\frac{p_1(1-p_1)}{n_1} + \frac{p_2(1-p_2)}{n_2}}$$

- A large sample $100(1-\alpha)\%$ confidence interval for $p_1 - p_2$ is:

$$\left[\hat{p}_1 - \hat{p}_2 - z_{\frac{\alpha}{2}} \sqrt{\frac{\hat{p}_1(1-\hat{p}_1)}{n_1} + \frac{\hat{p}_2(1-\hat{p}_2)}{n_2}}, \hat{p}_1 - \hat{p}_2 + z_{\frac{\alpha}{2}} \sqrt{\frac{\hat{p}_1(1-\hat{p}_1)}{n_1} + \frac{\hat{p}_2(1-\hat{p}_2)}{n_2}} \right]$$

Example 23. There are two sorts of seed. Let $p=P(\text{"germination"})$.

We sow 37 seeds of sort 1 and observe

X = the number of germs. $X=30$, $n_1=37$.

We sow 32 seeds of sort 2 and observe

Y = the number of germs. $Y=29$, $n_2 = 32$.

- We will construct a large sample 95% confidence interval for $p_1 - p_2$:

$$\hat{p}_1 = \frac{30}{37} = 0.811 \quad \hat{p}_2 = \frac{29}{32} = 0.906$$

$$\hat{p}_1 - \hat{p}_2 = -0.095$$

$$\sqrt{\frac{\hat{p}_1(1-\hat{p}_1)}{n_1} + \frac{\hat{p}_2(1-\hat{p}_2)}{n_2}} = \sqrt{\frac{0.811 \cdot 0.189}{37} + \frac{0.906 \cdot 0.094}{32}} = 0.082$$

- Lower limit = $\hat{p}_1 - \hat{p}_2 - z_{0.025} \sqrt{\frac{\hat{p}_1(1-\hat{p}_1)}{n_1} + \frac{\hat{p}_2(1-\hat{p}_2)}{n_2}}$
 $= -0.095 - 1.96 \cdot 0.082 \approx -0.256$
- Upper limit = $\hat{p}_1 - \hat{p}_2 + z_{0.025} \sqrt{\frac{\hat{p}_1(1-\hat{p}_1)}{n_1} + \frac{\hat{p}_2(1-\hat{p}_2)}{n_2}}$
 $= -0.095 + 1.96 \cdot 0.082 \approx 0.066$
- $[-0.256, 0.066]$ is a 95% large sample confidence interval for $p_1 - p_2$.

- We can test: $H_0: p_1 - p_2 = 0$ against $H_1: p_1 - p_2 \neq 0$

- Calculate:
$$z = \frac{\hat{p}_1 - \hat{p}_2}{\sqrt{\hat{p}(1 - \hat{p})\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}}$$

- where
$$\hat{p} = \frac{X + Y}{n_1 + n_2}$$

- We reject H_0 if $|z| \geq z_{\frac{\alpha}{2}}$

- Test $H_0: p_1 - p_2 = 0$ against $H_1: p_1 - p_2 < 0$
 - Reject H_0 with the α level of significance if $z \leq -z_\alpha$.
-
- Test $H_0: p_1 - p_2 = 0$ against $H_1: p_1 - p_2 > 0$
 - Reject H_0 with the α level of significance if $z \geq z_\alpha$.
-
- **Example 23 revisited:** We test:
 - $H_0: p_1 - p_2 = 0$ against $H_1: p_1 - p_2 \neq 0$

$$\hat{p} = \frac{30 + 29}{37 + 32} = 0.855$$

$$z = \frac{0.811 - 0.906}{\sqrt{0.855(1 - 0.855)\left(\frac{1}{37} + \frac{1}{32}\right)}} = -1.118$$

$$z_{0.025} = 1.96 > |z| \text{ and we retain } H_0.$$

We can also test: $H_0: p_1 - p_2 = 0$ against $H_1: p_1 - p_2 < 0$

$Z_{0.05} = 1.645$ $-1.645 < -1.118$ and H_0 is retained.

Chapter 11. Regression analysis 1.

Simple linear regression.

- We observe X and Y from n units. The purpose for doing this could be:
- We want to find a general connection between X and Y or we want to predict Y from X .
- In regression analysis $X=x$ is considered as known, while Y is a random variable and it depends on x .
- Y is called the response variable
- X is called the predictor variable.

Example 24.

- People in forestry need to estimate the amount of timber in a given area of a forest.
- To examine the relationship between diameter (x) and volume (y) of an eucalyptus tree we have these observations from 8 trees:

- Row diameter volume

- 1 0,34 0,67

- 2 0,38 0,71

- 3 0,41 0,78

- 4 0,44 0,83

- 5 0,46 0,87

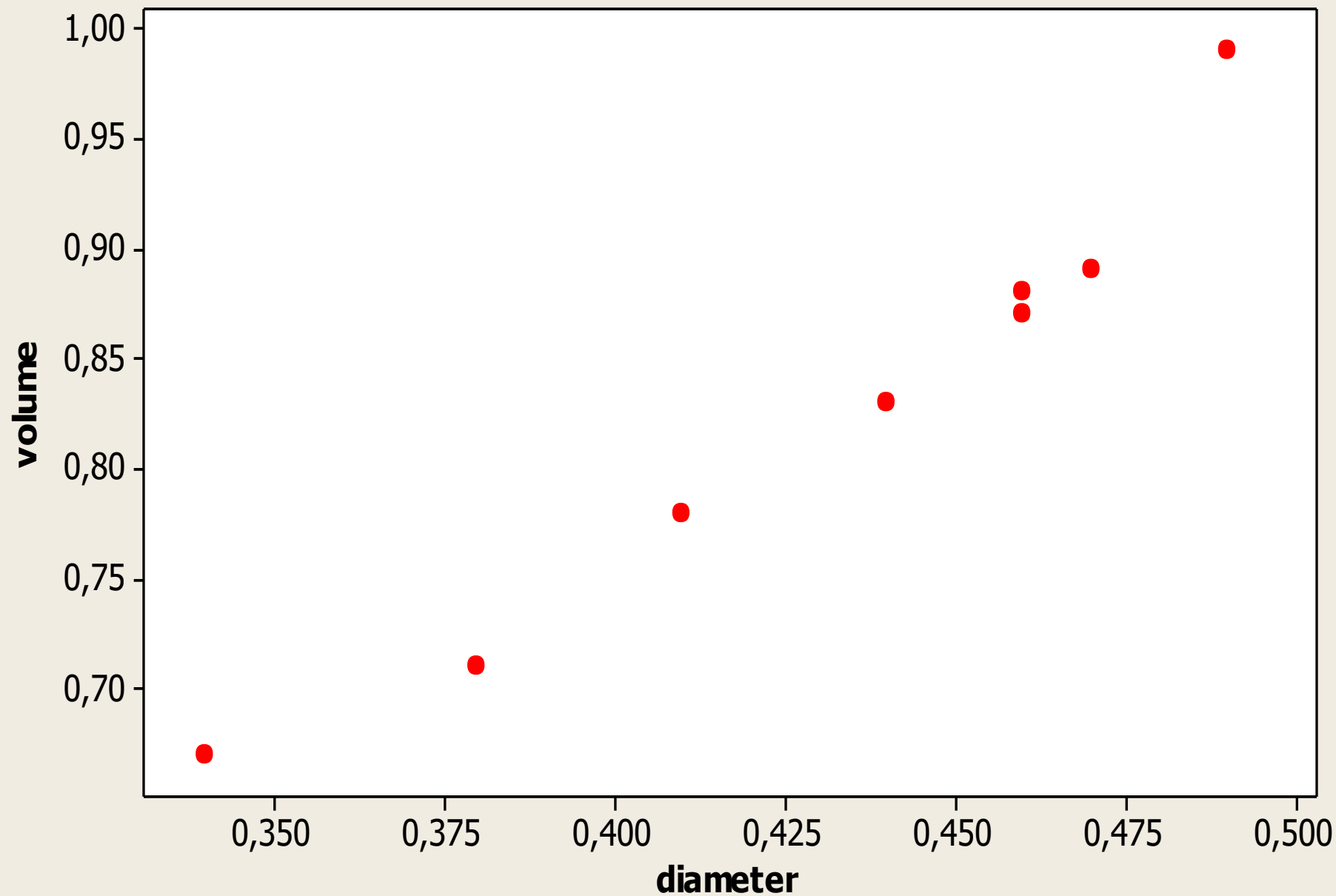
- 6 0,47 0,89

- 7 0,46 0,88

- 8 0,49 0,99

- Diameters are give in meters, volumes in cubic meters.

Volumes versus diameters for eucalyptus trees.



- If we find a plot like this, we will assume the model: $Y_i = \beta_0 + \beta_1 x_i + e_i$
- This is the model of linear regression.
- The random errors e_1, e_2, \dots, e_n are assumed to be independent $N(0, \sigma)$.
- Then Y_1, Y_2, \dots, Y_n are independent.
- Given $X=x$, Y_i is $N(\beta_0 + \beta_1 x_i, \sigma)$

- The unknown parameters β_0 and β_1 are estimated by the least square estimators $\hat{\beta}_0$ and $\hat{\beta}_1$ respectively.
- The estimates for the eucalyptus trees in example 24 are: $\hat{\beta}_0 = -0.03$, $\hat{\beta}_1 = 2$
- For eucalyptus trees with diameter 0.4 m. we will estimate the mean volume $\mu_0 = E(Y_0)$ as:

$$\hat{\mu}_0 = \hat{\beta}_0 + \hat{\beta}_1 x_0 = -0.03 + 2 \cdot 0.4 = 0.77$$

- If we have one eucalyptus tree with diameter 0.4 m. we will predict the volume of the tree by:

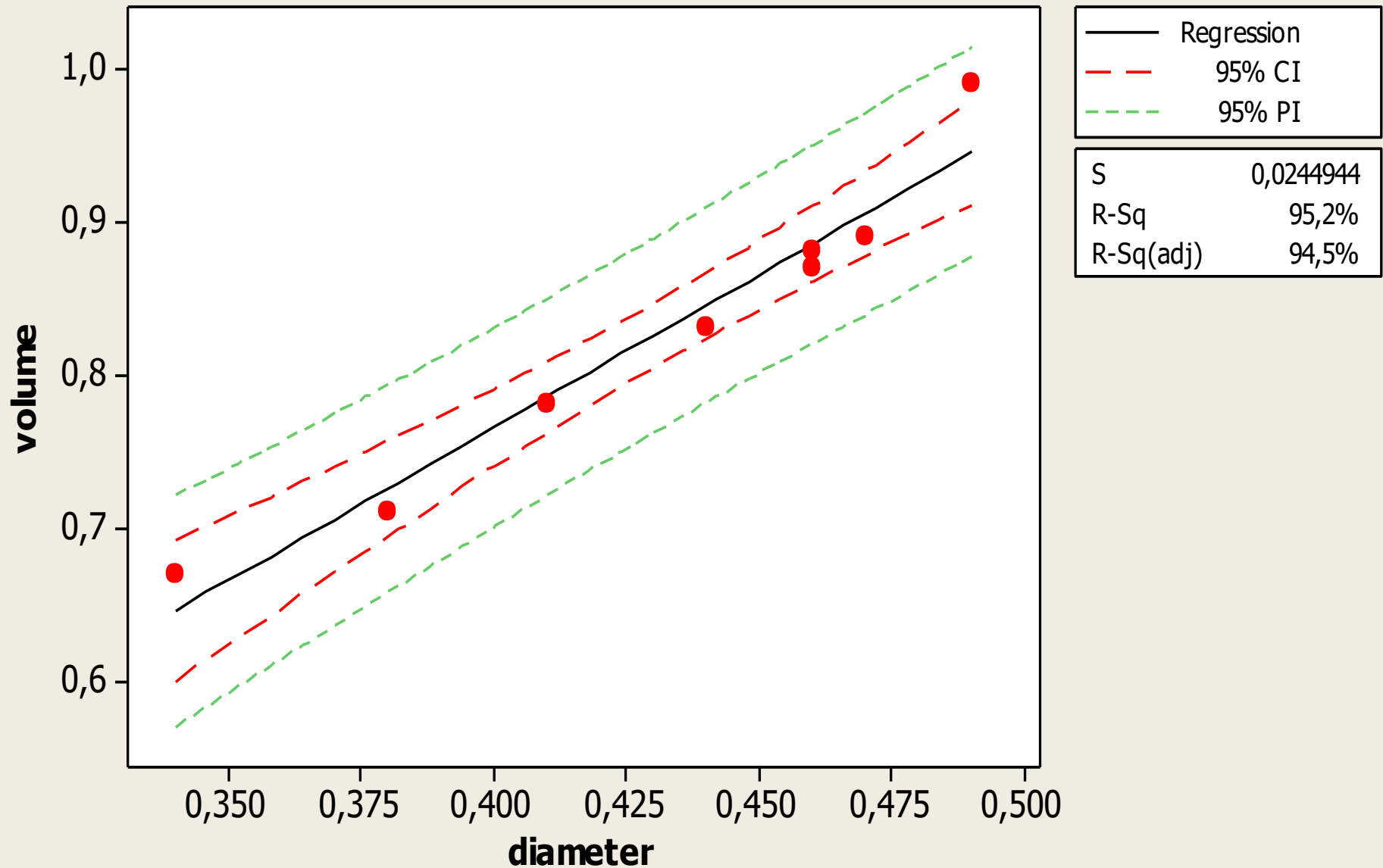
$$\hat{Y}_0 = \hat{\beta}_0 + \hat{\beta}_1 x_0 = -0.03 + 2 \cdot 0.4 = 0.77$$

- We can predict the volumes of all trees in the dataset. The points $(x_1, \hat{y}_1), \dots, (x_8, \hat{y}_8)$ are on the fitted line. \hat{y}_i is called a fitted value.
- The residual of observation i is: $\hat{e}_i = y_i - \hat{y}_i$

- Interpretation of $\hat{\beta}_0$: if $x=0$ then we predict Y to be $\hat{\beta}_0$
- Interpretation of $\hat{\beta}_1$: if x increases by 1 then the predicted Y increases by $\hat{\beta}_1$
- In example 24: $\hat{\beta}_0 = -0.03$ which means that the predicted volume of a tree with diameter 0 is -0.03. This does not give sense, because we have not observed Y when $x=0$.
- $\hat{\beta}_1 = 2$ which means that if the diameter of a tree increases by 1m then the predicted volume of the tree increases by 2m^3 .

Fitted line plot for 8 eucalyptus trees.

$$\text{volume} = -0,03381 + 1,997 \text{ diameter}$$



□ σ^2 is estimated by: $s^2 = \frac{\sum_{i=1}^n \hat{e}_i^2}{n-2} = \frac{\text{SSE}}{n-2}$

- SSE is the sum of squares for errors.

- It can be shown that: $\hat{\beta}_0 \sim N(\beta_0, \text{SE}(\hat{\beta}_0))$

$$\hat{\beta}_1 \sim N(\beta_1, \text{SE}(\hat{\beta}_1)) \text{ and } \hat{\mu}_0 \sim N(\mu_0, \text{SE}(\hat{\mu}_0))$$

It can be shown that: $\text{SE}(\hat{\beta}_0) = \sigma \sqrt{\frac{1}{n} + \frac{\bar{x}^2}{\sum_{i=1}^n (x_i - \bar{x})^2}}$

$$SE(\hat{\beta}_1) = \frac{\sigma}{\sqrt{\sum_{i=1}^n (x_i - \bar{x})^2}}$$

$$SE(\hat{\mu}_0) = \sigma \sqrt{\frac{1}{n} + \frac{(x_0 - \bar{x})^2}{\sum_{i=1}^n (x_i - \bar{x})^2}}$$

- These standard deviations are estimated by replacing σ by

$$s = \sqrt{\frac{SSE}{n-2}}$$

- These estimated standard deviations will be denoted as: $S_{\hat{\beta}_0}$ $S_{\hat{\beta}_1}$ $S_{\hat{\mu}_0}$
- Now we can calculate $100(1-\alpha)\%$ confidence intervals(the t-distribution has $n-2$ degrees of freedom):

- For β_0 :
$$\left[\hat{\beta}_0 - t_{\frac{\alpha}{2}, n-2} S_{\hat{\beta}_0}, \hat{\beta}_0 + t_{\frac{\alpha}{2}, n-2} S_{\hat{\beta}_0} \right]$$

- For β_1 :
$$\left[\hat{\beta}_1 - t_{\frac{\alpha}{2}, n-2} S_{\hat{\beta}_1}, \hat{\beta}_1 + t_{\frac{\alpha}{2}, n-2} S_{\hat{\beta}_1} \right]$$

- For μ_0 : $\left[\hat{\mu}_0 - t_{\frac{\alpha}{2}, n-2} S_{\hat{\mu}_0}, \hat{\mu}_0 + t_{\frac{\alpha}{2}, n-2} S_{\hat{\mu}_0} \right]$
 - A $100(1-\alpha)\%$ prediction interval for Y_0 is: $\left[\hat{Y}_0 - t_{\frac{\alpha}{2}, n-2} S_{\hat{Y}_0}, \hat{Y}_0 + t_{\frac{\alpha}{2}, n-2} S_{\hat{Y}_0} \right]$
 - Here $S_{\hat{Y}_0} = S \sqrt{1 + \frac{1}{n} + \frac{(x_0 - \bar{x})^2}{\sum_{i=1}^n (x_i - \bar{x})^2}}$
- is the estimated standard error of \hat{Y}_0

- The prediction interval for Y_0 when $X=x_0$ will always be wider than the confidence interval for μ_0 when $X=x_0$.
- Both intervals will be more narrow when $x_0 = \bar{X}$ than when x_0 has any other value.

Interpretation of a $100(1-\alpha)\%$ prediction interval for Y_0 :

The probability that Y_0 will be inside the interval is $1-\alpha$.

- Interpretation of a $100(1-\alpha)\%$ confidence interval for μ_0 :
- We are $100(1-\alpha)\%$ confident that the interval will cover μ_0 .
- If some x-values in the dataset are close to 0 we can perform a test for the intercept β_0 . It is based on the test-statistic:

$$t_{\hat{\beta}_0} = \frac{\hat{\beta}_0 - b_0}{S_{\hat{\beta}_0}}$$

Here b_0 is the value which is specified in H_0 .

- Usually b_0 is 0. This means that we are testing if the line does not go through the origin.
- A test for $H_0: \beta_0 = b_0$ against $H_1: \beta_0 \neq b_0$:
- Reject H_0 if $\left| t_{\hat{\beta}_0} \right| \geq t_{\frac{\alpha}{2}, n-2}$
- A test for $H_0: \beta_0 = b_0$ against $H_1: \beta_0 > b_0$:
- Reject H_0 if $t_{\hat{\beta}_0} \geq t_{\alpha, n-2}$
- A test for $H_0: \beta_0 = b_0$ against $H_1: \beta_0 < b_0$:
- Reject H_0 if $t_{\hat{\beta}_0} \leq -t_{\alpha, n-2}$

- A test for the slope β_1 is based on the test-statistic:

$$t_{\hat{\beta}_1} = \frac{\hat{\beta}_1 - b_1}{S_{\hat{\beta}_1}}$$

- Here b_1 is the value which is specified in H_0 .
- A test for $H_0: \beta_1 = b_1$ against $H_1: \beta_1 \neq b_1$:
- Reject H_0 if $\left| t_{\hat{\beta}_1} \right| \geq t_{\frac{\alpha}{2}, n-2}$

- A test for $H_0: \beta_1 = b_1$ against $H_1: \beta_1 > b_1$:
- Reject H_0 if $t_{\hat{\beta}_1} \geq t_{\alpha, n-2}$
- A test for $H_0: \beta_1 = b_1$ against $H_1: \beta_1 < b_1$:
- Reject H_0 if $t_{\hat{\beta}_1} \leq -t_{\alpha, n-2}$

The strength of a linear relation.

The linear relation is strong if SSE is small.

It can be shown:

$$SSE = (1 - r^2) \sum_{i=1}^n (y_i - \bar{y})^2$$

- Here r is the sample correlation coefficient.
- If r^2 is close to 1 then SSE will be small.
- r^2 comes out of Minitab as R-sq. It is given in %.
- If r^2 is close to 1 we have a strong linear relationship between X and Y .
- If r^2 is close to 0 we don't have a linear relationship between X and Y .

- We have:
$$r^2 = R^2 = 1 - \frac{\sum_{i=1}^n (y_i - \hat{y}_i)^2}{\sum_{i=1}^n (y_i - \bar{y})^2}$$

- It is the fraction of the variability in Y which is explained by the fitted linear regression model.

- Minitab also calculates R-sq(adj):

$$R^2 - \text{adj} = 1 - \frac{\frac{1}{n-2} \sum_{i=1}^n (y_i - \hat{y}_i)^2}{\frac{1}{n-1} \sum_{i=1}^n (y_i - \bar{y})^2}$$

- We evaluate the fit of the model by R^2 and a plot of the residuals against the fitted values.
- $0 \leq R^2 \leq 1$ always.
- If R^2 is large (≥ 0.7) and we don't see a systematic pattern in the plot, we have a good model.
- The model has a poor fit if $R^2 < 0.3$

Example 24 revisited.

Regression Analysis: volume versus diameter

The regression equation is

$$\text{volume} = -0,0338 + 2,00 \text{ diameter}$$

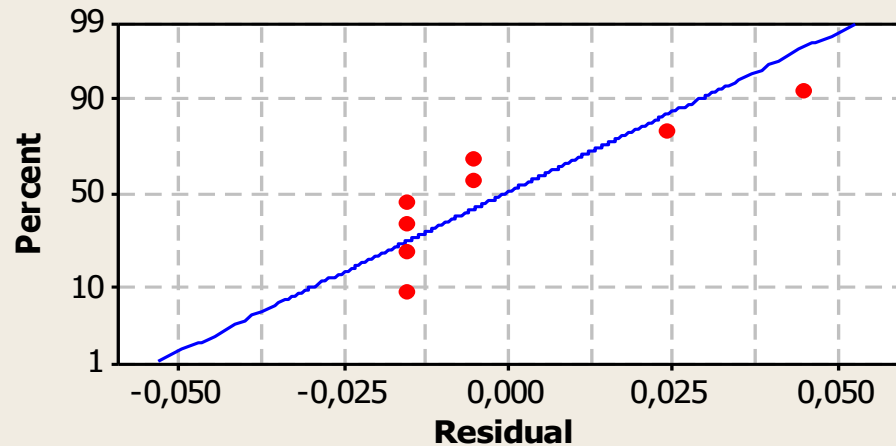
| Predictor | Coef | SE Coef | T | P |
|-----------|----------|---------|-------|-------|
| Constant | -0,03381 | 0,07902 | -0,43 | 0,684 |
| diameter | 1,9972 | 0,1821 | 10,97 | 0,000 |

$$S = 0,0244944 \quad R\text{-Sq} = 95,2\% \quad R\text{-Sq}(\text{adj}) = 94,5\%$$

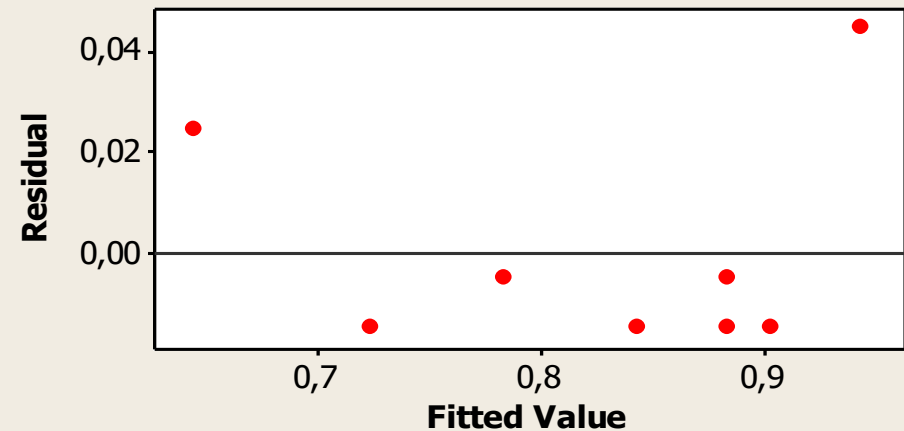
Residual plots for 8 eucalyptus trees.

Residual Plots for volume

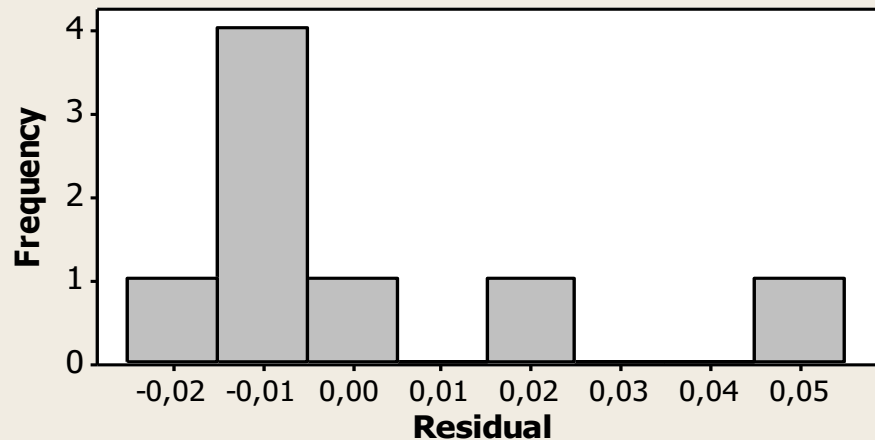
Normal Probability Plot of the Residuals



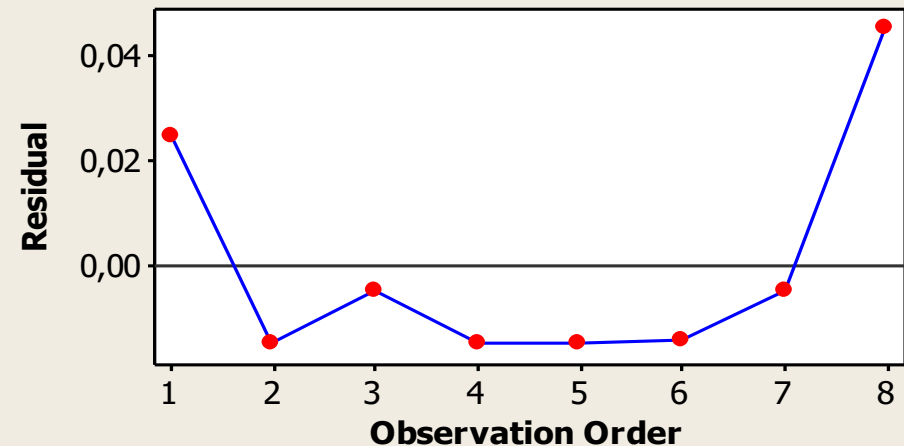
Residuals Versus the Fitted Values



Histogram of the Residuals



Residuals Versus the Order of the Data



Chapter 12. Regression Analysis 2.

Multiple linear regression.

The response variable can depend on more than one predictor variable.

- **Example 25:** For location i we observe:
- Y_i = the yield of a crop of carrots in kilo per 1000m².
- X_{i1} = the average temperature per day during the growth season.
- X_{i2} = the average rainfall per day during the growth season.
- X_{i3} = The average radiation from the sun per day during the growth season.
- X_{i4} = The number of days from April 30. until sowing.
- X_{i5} = The number of days from sowing to harvesting.

- There are observations from 12 places, $i=1, \dots, 12$.
- A multiple linear regression model can be:
- $Y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \beta_3 x_{i3} + \beta_4 x_{i4} + \beta_5 x_{i5} + e_i$
- $i=1, \dots, 12$.
- The error terms are assumed to be independent and $N(0, \sigma)$.
- The unknown parameters are: $\beta_0, \beta_1, \beta_2, \beta_3, \beta_4, \beta_5$ and σ .

- Minitab estimates these unknown parameters by the method of least squares.
- In example 25, that is to minimize

$$Q = \sum_{i=1}^{12} (y_i - \beta_0 - \beta_1 x_{i1} - \beta_2 x_{i2} - \beta_3 x_{i3} - \beta_4 x_{i4} - \beta_5 x_{i5})^2$$

- With respect to $\beta_0, \beta_1, \beta_2, \beta_3, \beta_4, \beta_5$.
- The values of the parameters that minimize Q are:

$$\hat{\beta}_0, \hat{\beta}_1, \hat{\beta}_2, \hat{\beta}_3, \hat{\beta}_4, \hat{\beta}_5$$

- A fitted value for unit no i in the dataset is:

$$\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_{i1} + \hat{\beta}_2 x_{i2} + \hat{\beta}_3 x_{i3} + \hat{\beta}_4 x_{i4} + \hat{\beta}_5 x_{i5}$$

- An estimator of σ^2 is: $S^2 = \frac{1}{n - k - 1} \sum_{i=1}^n (y_i - \hat{y}_i)^2$
- Here k = the number of predictor variables in the model and n=the number of observed units in the dataset which is used to fit the model. (k=5 and n=12 in example 25)

- In regression analysis: n must be greater than k.
- If $n \leq k$, some other analysis can be performed, e.g. PCR or PLS.
- In regression analysis the coefficient of determination is important.

- It is:
$$R^2 = 1 - \frac{\sum_{i=1}^n (y_i - \hat{y}_i)^2}{\sum_{i=1}^n (y_i - \bar{y})^2}$$

- $0 \leq R^2 \leq 1$ always.
- If $R^2 < 0.3$ the fit of the model is poor.
- If $0.3 \leq R^2 < 0.7$, the fit of the model is fair
- If $R^2 \geq 0.7$, the fit of the model can be very good.
- But we also have to examine the residual plot.

- A residual plot is a plot of the residuals versus the fitted values.
- If this plot shows a systematic pattern, it indicates that the error terms are not random, they are systematic. Then we might be able to find a better model.
- If you find a pattern, you can plot the residuals versus each of the predictor variables in the model, one predictor at the time.

- Minitab also calculates R^2 -adjusted.

- It is:
$$R^2 - \text{adj} = 1 - \frac{\frac{1}{n - k - 1} \sum_{i=1}^n (y_i - \hat{y}_i)^2}{\frac{1}{n - 1} \sum_{i=1}^n (y_i - \bar{y})^2}$$

- R-square adjusted compares two estimates of $\text{Var}(Y)$, $\hat{\sigma}_{\text{reg}}^2$ using the regression model, and $\hat{\sigma}^2$ using the one sample model.

- We can test if one or some predictor variables can be removed from the model.
- Our aim is to simplify the model and still have a large R^2 .
- First we can test if at least one predictor variable is significant:
 - $H_0: \beta_1 = \beta_2 = \dots = \beta_k = 0$ against
 - $H_1: \text{at least one } \beta_i \neq 0, i=1, \dots, k.$

- Reject H_0 at the level of significance α if the p-value in the analysis of variance table $\leq \alpha$.
- If H_0 is rejected, we can go on testing one slope at the time.
- Start with the parameter with the highest p-value.
- In example 25, this is β_1 . Then we test $H_0: \beta_1=0$ against $H_1: \beta_1 \neq 0$, assume $\beta_2, \beta_3, \beta_4$ and β_5 are not equal to 0.

- If the p-value is $\leq \alpha$, all predictor variables are significant, and the residuals should be inspected.
- If the p-value is $> \alpha$, H_0 is retained. Remove the corresponding predictor variable from the model. Fit a model with the remaining predictor variables. Repeat the test-procedure until all p-values in the suggested model are $\leq \alpha$.
- This method is called backward elimination.

- The level of significance should not be too small.
- Often $\alpha = 0.05$ is O.K.
- Minitab lists the unusual observations.
- If you get a long list of unusual observations, something must be wrong, and you should try another model.

- When just one parameter is considered in the null hypothesis, a t-test with $n-k-1$ degrees of freedom is performed. Here k = the number of parameters which are $\neq 0$ in H_1 .
- If we find a good or fair model, it can be used for prediction of the response for a unit where the predictor variables have been observed.

- The model can also be used to assess the relationship between the response and a predictor variable.
- Interpretation of an estimated slope:
- If the corresponding predictor variable is increased by 1 unit, and the other predictor variables are kept constant, then the predicted response will increase by the value of the estimated slope.
- If the estimated slope is negative, we will say the predictor variable will decrease instead of increase.

- Interpretation of the estimated intercept:
- It is the predicted response when all the predictor variables in the model are 0. Often this interpretation does not give sense because there is no unit in the dataset with all predictor variables equal to 0.
- If we can control a predictor variable, we can arrange a situation in a most favorable way.

Chapter 14. Analysis of variance.

Comparison of several treatments, the completely randomized design.

Independent random samples from two populations is a special case of this. We want to compare the mean of more than two populations.

Example 26.

- We want to compare 4 varieties of grain (4 populations) with respect to yield.
- We draw one sample from each population, the sample sizes could be: $n_1=5$, $n_2=4$, $n_3=4$ and $n_4=5$ respectively.
- Let Y_{ij} = the yield of grain, variety i , observation j . $j=1, 2, 3, 4$, $j = 1, \dots, n_i$.

- We assume: all observations are independent $Y_{ij} \sim N(\mu_i, \sigma)$ $i=1, 2, \dots, k$
- Alternatively: $Y_{ij} = \mu + \alpha_i + e_{ij}$ $i=1, \dots, k, j=1, \dots, n_i$
- e_{ij} = the error term for observation ij .
- We assume: $e_{ij} \sim N(0, \sigma)$ and all error-terms are independent.

□ μ = The grand mean.

□ α_i = effect of treatment i and: $\sum_{i=1}^k \alpha_i = 0$

- There are 5 unknown parameters here (k+1)
- They are: $\mu_1, \mu_2, \mu_3, \mu_4$, and σ .
- They are estimated by:

$$\hat{\mu}_1 = \bar{y}_1. \quad \hat{\mu}_2 = \bar{y}_2. \quad \hat{\mu}_3 = \bar{y}_3. \quad \hat{\mu}_4 = \bar{y}_4.$$

- (the sample means).

- σ can be estimated by the sample standard deviation of sample i:

$$S_i = \sqrt{\frac{1}{n_i - 1} \sum_{j=1}^{n_i} (y_{ij} - \bar{y}_{i.})^2}$$

- Let N = the total number of observation = $\sum_{i=1}^k n_i$
- and k = the number of populations we compare, then the best unbiased estimator of σ is:

$$\hat{\sigma} = S_p = \sqrt{\frac{(n_1 - 1)s_1^2 + (n_2 - 1)s_2^2 + \dots + (n_k - 1)s_k^2}{N - k}}$$

$$= \sqrt{\frac{\sum_{i=1}^k \sum_{j=1}^{n_i} (y_{ij} - \bar{y}_{i\cdot})^2}{N - k}} = \sqrt{\frac{\text{SSE}}{N - k}}$$

- SSE = sums of squares for errors.
- SST = sum of squares for treatment =

$$\sum_{i=1}^k n_i (\bar{y}_{i.} - \bar{y}_{..})^2$$

- It can be shown that $SStotal = \sum_{i=1}^k \sum_{j=1}^{n_i} (y_{ij} - \bar{y}_{..})^2$

 $= SST + SSE$

- The effect we consider can be effect of something else than treatments. There could be effects of days, seasons, age groups, and so on.
- We can test the hypotheses:
- H_0 : The treatments have equal means.
- That is: $H_0: \mu_1 = \mu_2 = \dots = \mu_k$
- The alternative hypothesis is:
- H_1 : At least two treatments have different means

- If H_0 is true then $SST/(k-1)$ will estimate σ^2 .
- If H_0 is not true, $SST/(k-1)$ will estimate something greater than σ^2 .
- $S_p^2 = \frac{SSE}{N - k}$ will always estimate σ^2 , no matter if H_0 is true or false.

- We reject H_0 at the α level of significance if:

$$F = \frac{SST/(k-1)}{SSE/(N-k)} \geq F_{\alpha, k-1, (N-k)}$$

= the upper α -point of the F-distribution with degrees of freedom = df = (k-1, N-k).

If we reject H_0 , we will state that at least two population means are different.

- To examine which population means are different, we construct confidence intervals for $\mu_{i1} - \mu_{i2}$. Often we will choose the level of significance smaller than α for these intervals.
- A $100(1-\alpha)\%$ confidence interval for $\mu_1 - \mu_2$ is:

$$\left[\bar{y}_{1.} - \bar{y}_{2.} - t_{\frac{\alpha}{2}, (N-k)} S_p \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}, \bar{y}_{1.} - \bar{y}_{2.} + t_{\frac{\alpha}{2}, (N-k)} S_p \sqrt{\frac{1}{n_1} + \frac{1}{n_2}} \right]$$

- The interval can be used to test
- $H_0: \mu_1 - \mu_2 = 0$ against $H_1: \mu_1 - \mu_2 \neq 0$
- Retain H_0 if 0 is inside the interval.
- Reject H_0 if 0 is outside the interval.
- Minitab can calculate confidence intervals for the mean difference between two population means, e.g. Fisher intervals.

- Minitab calculates the simultaneous confidence level which often is much smaller than $1-\alpha$. It is the probability that all intervals cover the true population differences they are calculated for.
- An alternative hypothesis can be one-sided:
- $H_0: \mu_1 - \mu_2 = 0$ against $H_1: \mu_1 - \mu_2 < 0$

- Reject H_0 with significance level α if:

$$t = \frac{\bar{y}_{1.} - \bar{y}_{2.}}{Sp \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}} \leq -t_{\alpha, (N-k)}$$

- or test: $H_0: \mu_1 - \mu_2 = 0$ against $H_1: \mu_1 - \mu_2 > 0$
- Reject H_0 with significance level α if:

$$t = \frac{\bar{y}_{1.} - \bar{y}_{2.}}{Sp \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}} \geq t_{\alpha, (N-k)}$$

- To examine if the error terms are random and have a normal distribution, we consider fitted values and residuals:
- A fitted value is: $\hat{y}_{ij} = \hat{\mu}_i = \bar{y}_i$.
- A residual is: $\hat{e}_{ij} = y_{ij} - \hat{y}_{ij} = y_{ij} - \bar{y}_i$.
- The residuals can be plotted against the corresponding fitted values or the population number.
- The model can also be assessed by dot plots or box-plots.

Two way analysis of variance.

Randomized block experiments for comparing treatments.

We want to compare k treatments. There are b blocks. Each block is split into k equal pieces.

One block:

| | | | | |
|--|--|--|--|--|
| | | | | |
|--|--|--|--|--|

 5 treatments

- We try all the treatments in one block, and decide at random which unit we shall give which treatment.
- If there are k treatments and b blocks, we get kb observations: $Y_{11}, Y_{12}, \dots, Y_{1b}, Y_{21}, Y_{22}, \dots, Y_{2b}, \dots, Y_{k1}, Y_{k2}, \dots, Y_{kb}$
- Assume: $Y_{ij} = \mu + \alpha_i + \beta_j + e_{ij}$
 $i=1, \dots, k, j=1, \dots, b$

- e_{ij} = the error term for observation ij .
- We assume: $e_{ij} \sim N(0, \sigma)$ and all error-terms are independent.

□ μ = The grand mean.

□ α_i = effect of treatment i

□ β_j = effect of block j

$$\sum_{i=1}^k \alpha_i = 0 \quad \sum_{j=1}^b \beta_j = 0$$

- A block effect can be: One large field which is split into k small fields. We have b large fields.

- A block effect can also be an effect of litter if there are b litters, and each litter have at least k pigs.
- The α -effects can be effects of something else than treatments. There could be effects of days, seasons, age groups, and so on.
- We can test the hypotheses:
- H_0 : The treatment effects are equal
- H_1 : At least two treatment effects are different.

- As in the one way procedure, these sums of squares need to be calculated:

- $$SS_{\text{total}} = \sum_{i=1}^k \sum_{j=1}^b (y_{ij} - \bar{y}_{..})^2$$

- $$SS_{\text{treatment}} = SST = b \sum_{i=1}^k (\bar{y}_{i.} - \bar{y}_{..})^2$$

- In addition, sum of squares for blocks is:

$$SS_b = k \sum_{j=1}^b (\bar{y}_{.j} - \bar{y}_{..})^2$$

- And sum of squares for errors =

$$SSE = SS_{\text{total}} - SST - SS_b$$

| Source | SS | DF |
|-----------|---------|------------------|
| Treatment | SST | $k - 1$ |
| Block | SSb | $b - 1$ |
| Error | SSE | $(k - 1)(b - 1)$ |
| Total | SStotal | $kb - 1$ |

- We reject H_0 if:

$$F = \frac{SST/(k-1)}{SSE/(k-1)(b-1)} \geq F_{\alpha, k-1, (k-1)(b-1)}$$

- The level of significance of the test is α .
- We can also test:
- H_0 : All block effects are equal
- H_1 : at least 2 block effects are different.

- Reject H_0 if:

$$F = \frac{SSb/(b-1)}{SSE/(k-1)(b-1)} \geq F_{\alpha, b-1, (k-1)(b-1)}$$

- The level of significance of the test is α .
- Parameter estimates are:

$$\hat{\mu} = \bar{y}_{..} \quad \hat{\alpha}_i = \bar{y}_{i.} - \bar{y}_{..} \quad \hat{\beta}_j = \bar{y}_{.j} - \bar{y}_{..}$$

- A fitted value is: $\hat{y}_{ij} = \hat{\mu} + \hat{\alpha}_i + \hat{\beta}_j = \bar{y}_{i\cdot} + \bar{y}_{\cdot j} - \bar{y}_{..}$
- A residual is: $\hat{e}_{ij} = y_{ij} - \hat{y}_{ij} = y_{ij} - \bar{y}_{i\cdot} - \bar{y}_{\cdot j} + \bar{y}_{..}$
- The residuals can be plotted against the corresponding fitted values and against each effect.
- An estimator for σ is:

$$S = \hat{\sigma} = \sqrt{\frac{\text{SSE}}{(k-1)(b-1)}} = \sqrt{\frac{\sum_{i=1}^k \sum_{j=1}^b \hat{e}_{ij}^2}{(k-1)(b-1)}}$$

- We can construct a $100(1-\alpha)\%$ confidence interval for $\alpha_1 - \alpha_2$:

$$\left[\bar{y}_{1.} - \bar{y}_{2.} - t_{\frac{\alpha}{2}, (k-1)(b-1)} S \sqrt{\frac{2}{b}}, \bar{y}_{1.} - \bar{y}_{2.} + t_{\frac{\alpha}{2}, (k-1)(b-1)} S \sqrt{\frac{2}{b}} \right]$$

- The interval can be used to test
- $H_0: \alpha_1 - \alpha_2 = 0$ against $H_1: \alpha_1 - \alpha_2 \neq 0$
- Retain H_0 if 0 is inside the interval.
- Reject H_0 if 0 is outside the interval.

- The test has significance level α .
- We can construct a $100(1-\alpha)\%$ confidence interval for $\beta_1 - \beta_2$:

$$\left[\bar{y}_{.1} - \bar{y}_{.2} - t_{\frac{\alpha}{2}, (k-1)(b-1)} S \sqrt{\frac{2}{k}}, \bar{y}_{.1} - \bar{y}_{.2} + t_{\frac{\alpha}{2}, (k-1)(b-1)} S \sqrt{\frac{2}{k}} \right]$$

- The interval can be used to test
- $H_0: \beta_1 - \beta_2 = 0$ against $H_1: \beta_1 - \beta_2 \neq 0$

An alternative hypothesis can be one-sided:

$H_0: \alpha_1 - \alpha_2 = 0$ against $H_1: \alpha_1 - \alpha_2 < 0$

Reject H_0 at the α level of significance if:

$$t = \frac{\bar{y}_{1\cdot} - \bar{y}_{2\cdot}}{S\sqrt{\frac{2}{b}}} \leq -t_{\alpha, (k-1)(b-1)}$$

- Or test: $H_0: \alpha_1 - \alpha_2 = 0$ against $H_1: \alpha_1 - \alpha_2 > 0$
- Reject H_0 at the α level of significance if:

$$t = \frac{\bar{y}_{1\cdot} - \bar{y}_{2\cdot}}{S\sqrt{\frac{2}{b}}} \geq t_{\alpha, (k-1)(b-1)}$$

Chapter 13. Analysis of Categorical Data.

Chi-square tests.

We have the events A_1, A_2, \dots, A_r . They are mutually exclusive and

$$A_1 \cup A_2 \cup \dots \cup A_r = S$$

We also have the events B_1, B_2, \dots, B_c . They are mutually exclusive and

$$B_1 \cup B_2 \cup \dots \cup B_c = S$$

Example 27.

- A random sample of 219 alligators were captured in Florida. They were examined with respect to gender and primary food choice.
- Let B_1 =male, B_2 =female.
- Let A_1 =fish, A_2 =invertebrate, A_3 =reptile, A_4 =bird, A_5 =other.
- We want to test if gender and primary food choice are dependent.

- In general: We must have a random sample. Each unit in the sample must have exactly one of the events A_1, A_2, \dots, A_r and exactly one of the events B_1, B_2, \dots, B_c .
- We observe X_{ij} = the number of units having
- $A_i \cap B_j$ $i=1, 2, \dots, r$ and $j = 1, 2, \dots, c$.
- Let $p_{ij} = P(A_i \cap B_j)$
- $p_{i\cdot} = P(A_i)$ and $p_{\cdot j} = P(B_j)$ $\sum_{i=1}^r p_{i\cdot} = 1$ $\sum_{j=1}^c p_{\cdot j} = 1$

Observations:

| | B ₁ | B ₂ | B _c | total |
|----------------|-----------------|-----------------|-----------------|-----------------|
| A ₁ | X ₁₁ | X ₁₂ | | X _{1.} |
| A ₂ | X ₂₁ | X ₂₂ | | X _{2.} |
| | | | | |
| A _r | X _{r1} | X _{2r} | | X _{r.} |
| | X _{.1} | X _{.2} | X _{.c} | n |

- A test of independence:
- $H_0: P(A_i \cap B_j) = P(A_i) \cdot P(B_j)$
- for every i and j .
- Against H_1 : At least one
$$P(A_i \cap B_j) \neq P(A_i) \cdot P(B_j)$$

- We reject H_0 at the α level of significance if:

$$\chi^2 = \sum_{i=1}^r \sum_{j=1}^c \frac{\left(X_{ij} - \frac{X_{i.} X_{.j}}{n} \right)^2}{\frac{X_{i.} X_{.j}}{n}} \geq \chi_{\alpha, (c-1)(r-1)}^2$$

= the upper α -point of the chi-square distribution.

Example 27 revisited.

- We test H_0 : gender and food choice are independent.
- H_1 : gender and food choice are dependent.
- Minitab calculates the p-value = 0.719 and we retain H_0 at a 5% level of significance because the p-value > 0.05 .
- We can not state that gender and primary food choice are dependent.

A test of homogeneity.

- We still have a random sample of n units. On each unit we just observe if one of the events B_1, \dots, B_c has occurred.
- In advance we have decided the number of units we have in each of the categories A_1, \dots, A_r .
- Then $x_{1.}, \dots, x_{r.}$ will be fixed.

- Let $p_{1j} = P(B_j)$ in the first row.
- Let $p_{2j} = P(B_j)$ in the second row.
- Let $p_{rj} = P(B_j)$ in the last row.
- We test: $H_0: p_{1j} = p_{2j} = \dots = p_{rj}$ against
- H_1 : at least 2 probabilities in the same column are different.

- The test statistic and the test criteria are the same for this test as for a test of independence, but the hypotheses are different.
- Here we can compare districts, diets and so on.
- The test statistic has approximately a chi-square distribution with $(r-1)(c-1)$ degrees of freedom if H_0 is true and the expected counts are at least 5.

Example 28.

- Three sorts of seeds are compared with respect to germination. We sow 37 seeds of sort 1, 32 seeds of sort 2 and 30 seeds of sort 3.
- The results are:

| | germs | no germs | total |
|--------|-------|----------|-------|
| Sort 1 | 30 | 7 | 37 |
| Sort 2 | 29 | 3 | 32 |
| Sort 3 | 20 | 10 | 30 |
| Total | 79 | 20 | 99 |

- Let $p_{11}=P(\text{germ})$ for sort 1,
 - Let $p_{21}=P(\text{germ})$ for sort 2 and
 - Let $p_{31}=P(\text{germ})$ for sort 3.
-
- We test $H_0: p_{11}=p_{21}=p_{31}$ against
 - $H_1: p_{11} \neq p_{21}, \text{ or } p_{11} \neq p_{31} \text{ or } p_{21} \neq p_{31}$
-
- The p-value for the test of homogeneity is:
 $0.062 > 0.05$ and we retain H_0 at a 5% level of significance.

- If we want to compare the probability of germination for just 2 sorts of seed, we will perform a z-test for comparing 2 proportions.
- If a cell has an expected count less than 5, Minitab will give a warning.

Comparing 2 treatments,
some examples.

Two samples or matched pairs?

Example1: comparing the mean content of vitamin C in orange and grapefruit juice.

- At a factory, orange juice and grapefruit juice are produced. It should be tested if the orange juice contains less vitamin C than the grapefruit juice. A random sample of 5 packages of each kind of juice are taken and the content of vitamin C in mg. per liter is observed. The results are:

- | • Package | Orange | grapefruit |
|-----------|--------|------------|
| • 1 | 340 | 350 |
| • 2 | 350 | 340 |
| • 3 | 290 | 290 |
| • 4 | 280 | 290 |
| • 5 | 270 | 300 |
- What can be assumed about the observations?
 - Is this a 2 sample situation, or is it matched pairs?
 - What hypotheses will we test?

Example2: comparing mean prices of fresh and frozen salmon sold on the export market.

- In week 11 to 14 (2008), these registrations are taken on the price of 1 kilo salmon (given in Norwegian kroner):

| • Week | fresh | frozen |
|--------|-------|--------|
| • 11 | 26,96 | 26,81 |
| • 12 | 28,03 | 27,18 |
| • 13 | 27,21 | 25,85 |
| • 14 | 26,41 | 25,81 |

- We want to compare the mean price for fresh and frozen salmon sold on the export market. Are the mean prices different?
- What can be assumed about the observations?
- Is this a 2 sample situation, or is it matched pairs?
- What hypotheses will we test?

Global warming.

- **Summary**
- Autumn temperatures are at a record 5° C above normal, due to the major loss of sea ice in recent years which allows more solar heating of the ocean. Winter and springtime temperatures remain relatively warm over the entire Arctic, in contrast to the 20th century and consistent with an emerging global warming influence.
- The year 2007 was the warmest on record for the Arctic, continuing a general, Arctic-wide warming trend that began in the mid-1960s (Fig. A1).

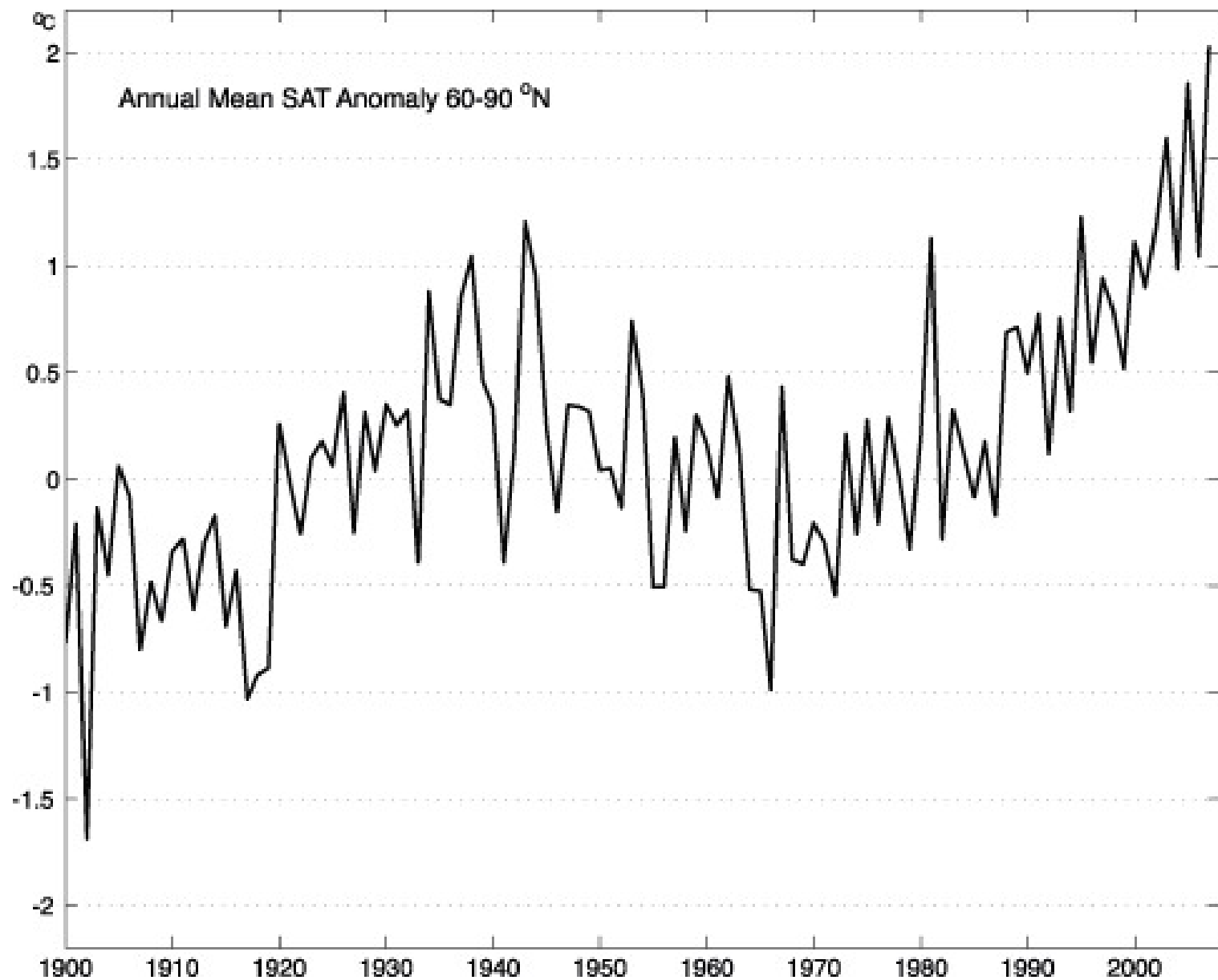
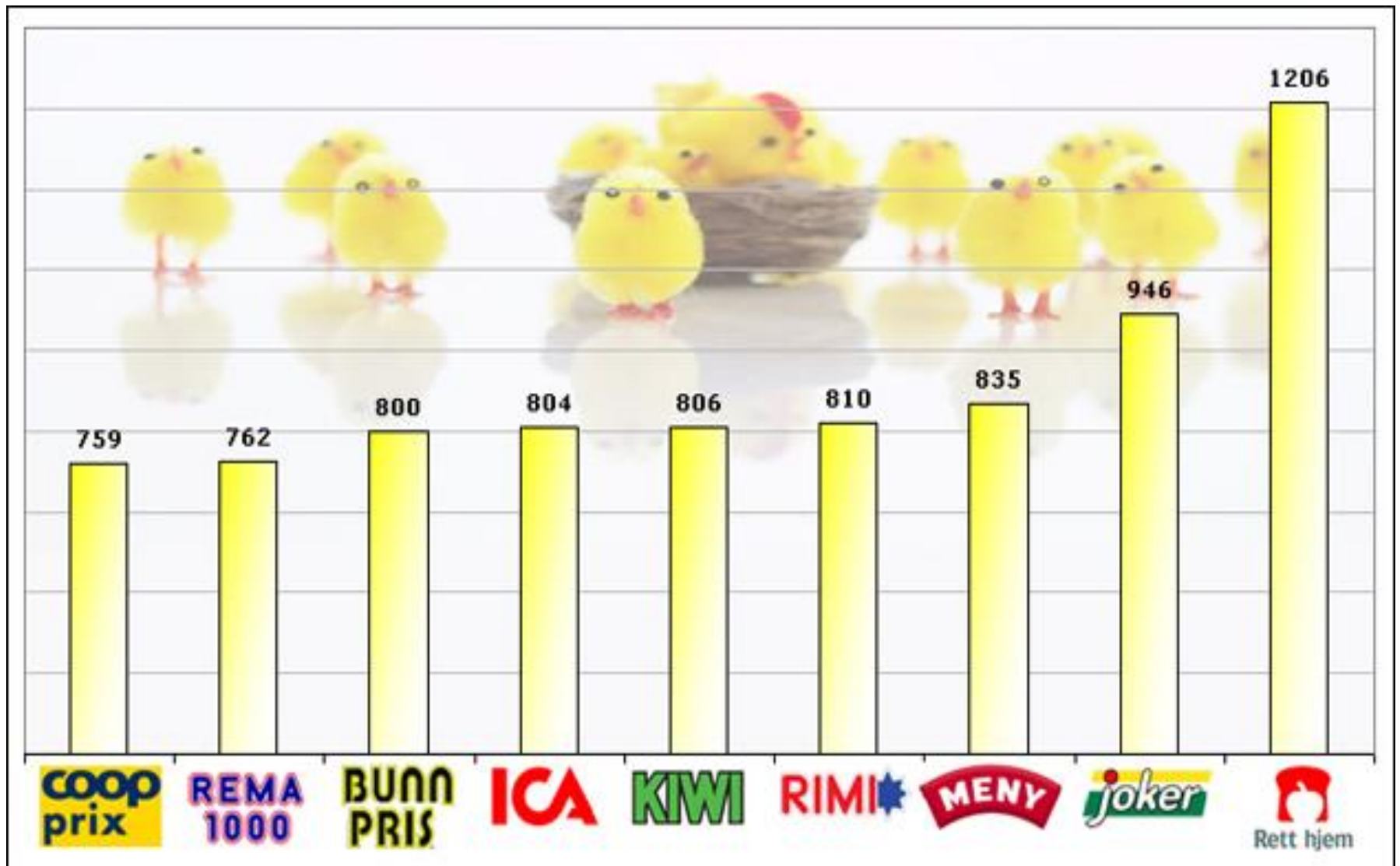


Figure A1. Arctic-wide annual averaged surface air temperature anomalies (60°–90°N) based on land stations north of 60°N relative to the 1961–90 mean.

Yellow prices at Easter.

**We saved 447 kroner by choosing
the cheapest supermarket.**

- An inspection of the expenses when buying typical Easter groceries has been performed.
- Here you can see the results from 8 supermarket chains and 1 Internet store.



The bar-chart shows the total price for selected items in a typical Easter-shopping basket. Coop Prix and Rema 1000 are cheapest, while Joker and Rett Hjem are most expensive.

- If we ignore The Internet store Rett Hjem from the comparison, and concentrate on the supermarkets, Joker is the most expensive. Our shopping basket costed 186 kroner more at Joker than at Coop Prix. Many items were most expensive at Joker, and 130 kroner more expensive than the average.
- Here are prices for selected items in the different supermarkets:



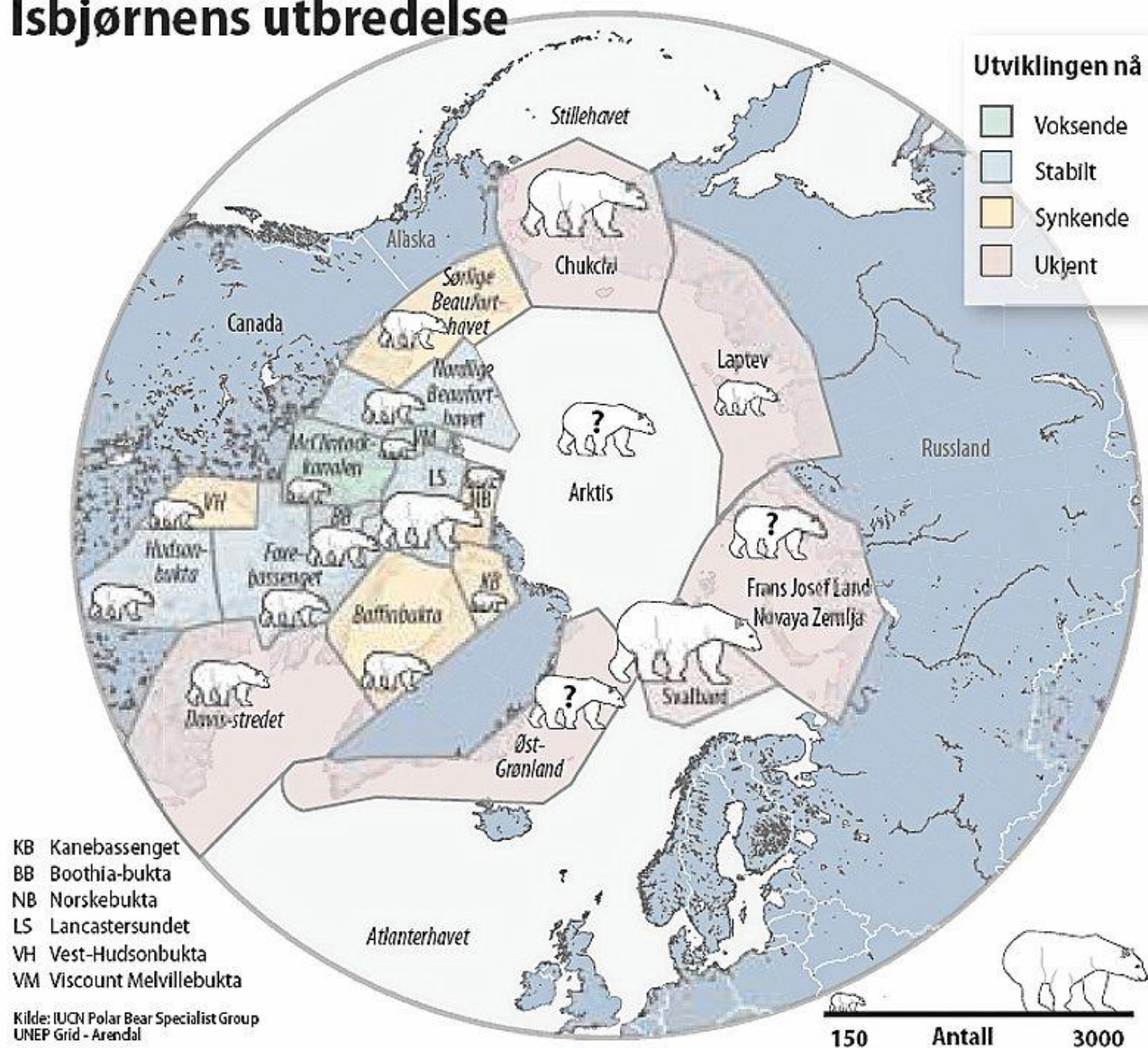
| | | | |
|------------------|-------|-------|-------|
| Rett hjem | 20,22 | 32,48 | 19,90 |
| Ica Nær | 15,90 | 16,90 | 10,14 |
| Meny | 15,50 | 24,90 | 16,90 |
| Joker | 17,90 | 26,90 | 20,50 |
| Coop Prix | 14,50 | 24,00 | 14,50 |
| Rema 1000 | 15,00 | 23,90 | 13,00 |
| Rimi | 13,50 | 23,50 | 15,00 |
| Bunnpris | 15,90 | 19,90 | 7,90 |
| Kiwi | 14,40 | 24,90 | 13,90 |

- If we want to compare 2 stores with respect to mean prices, should we perform a 2 sample t-test or matched pairs?
- If we want to compare all stores with respect to mean prices, should we perform a one way or a two way Anova?

Polar bears are forced towards the north

The polar bears in the Artic Zone are not threatened by fast extinction. But they have become fewer, and increased melting of the ice forces them to the north and west.

Isbjørnens utbredelse



We want to estimate the size of a population

- First we take a random sample of m units (animals, could be polar bears). The units are marked and then released. We let them mix with the population, and then we take a new random sample of size n . The number of marked units in this sample is X . Then we have: $\frac{m}{N} \approx \frac{X}{n}$ and N is the population size.
- This gives: $N \approx \frac{mn}{X}$

- We take a random sample of size $m=400$, could be polar bears and mark them. Then the animals are released and they mix with the population. A new random sample of size 500 is taken, and $X= 100$ is observed.
- We estimate the population size:

$$N \approx \frac{400 \cdot 500}{100} = 2000$$