Method and design

# Lecture 1

## Population vs. sample

Population = all things of interest in our study

Sample = subset of those things



Sample statistics are estimators of population parameters.

We use statistical tests to show the uncertainty of our estimated parameters.

A sample should be:

- random

- large enough (*n*=sample size)

- representative. Sample must reflect the entire population

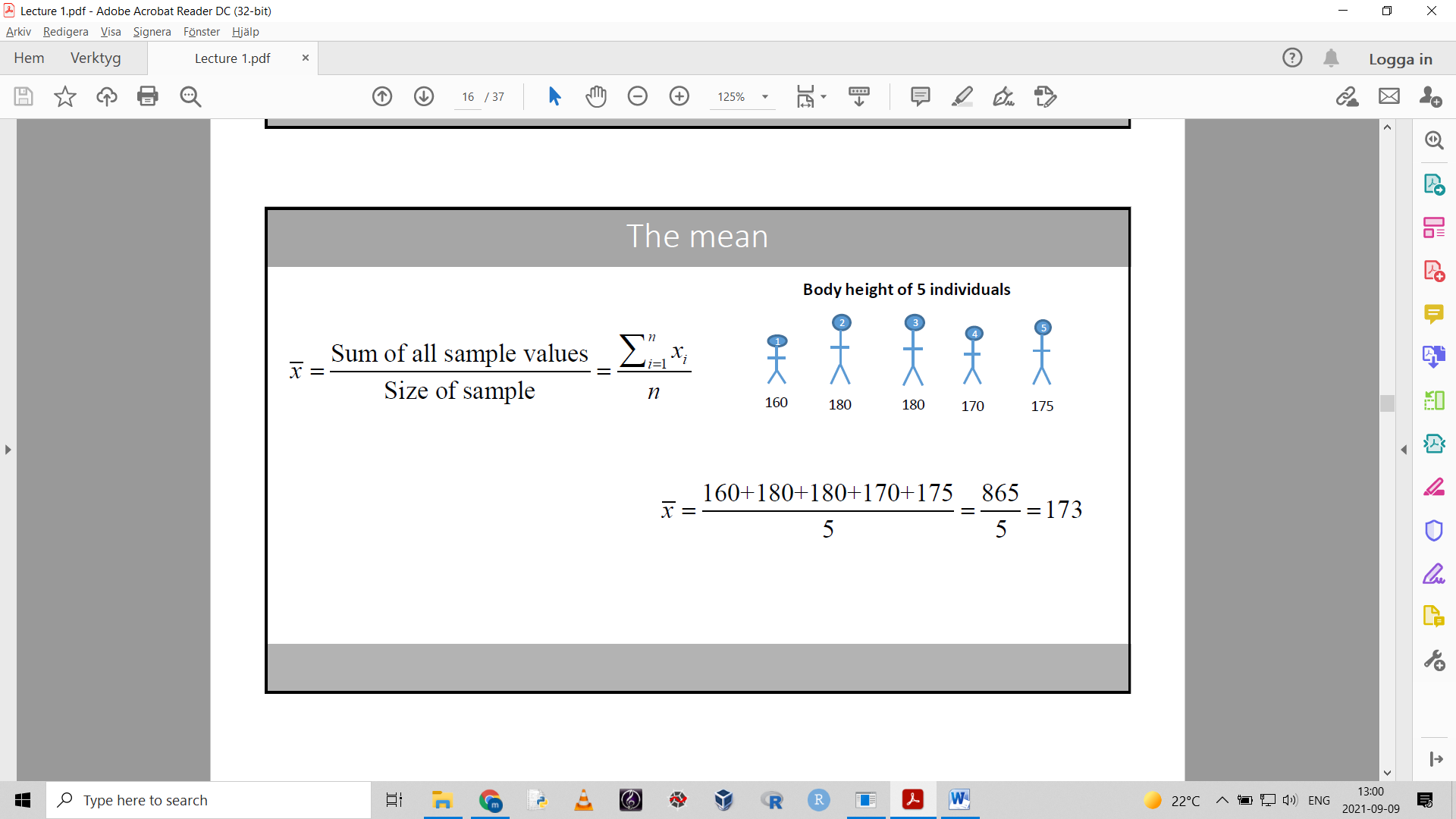
## Measures of the central tendency:

The central position of the data and some information about the spread.

### Central position:

- **Mean**

Aka average or arithmetic mean.



Used when data is symmetric

(with symmetric distribution of data, the mean and median are very close)

- **Median**

The middle value of the data set

Used when data distribution is skewed

(when data distribution is skewed, the mean and median can be very far apart)

- **Mode**

The value that appears most often in a data set.

Mainly used on categorical data

### Spread:

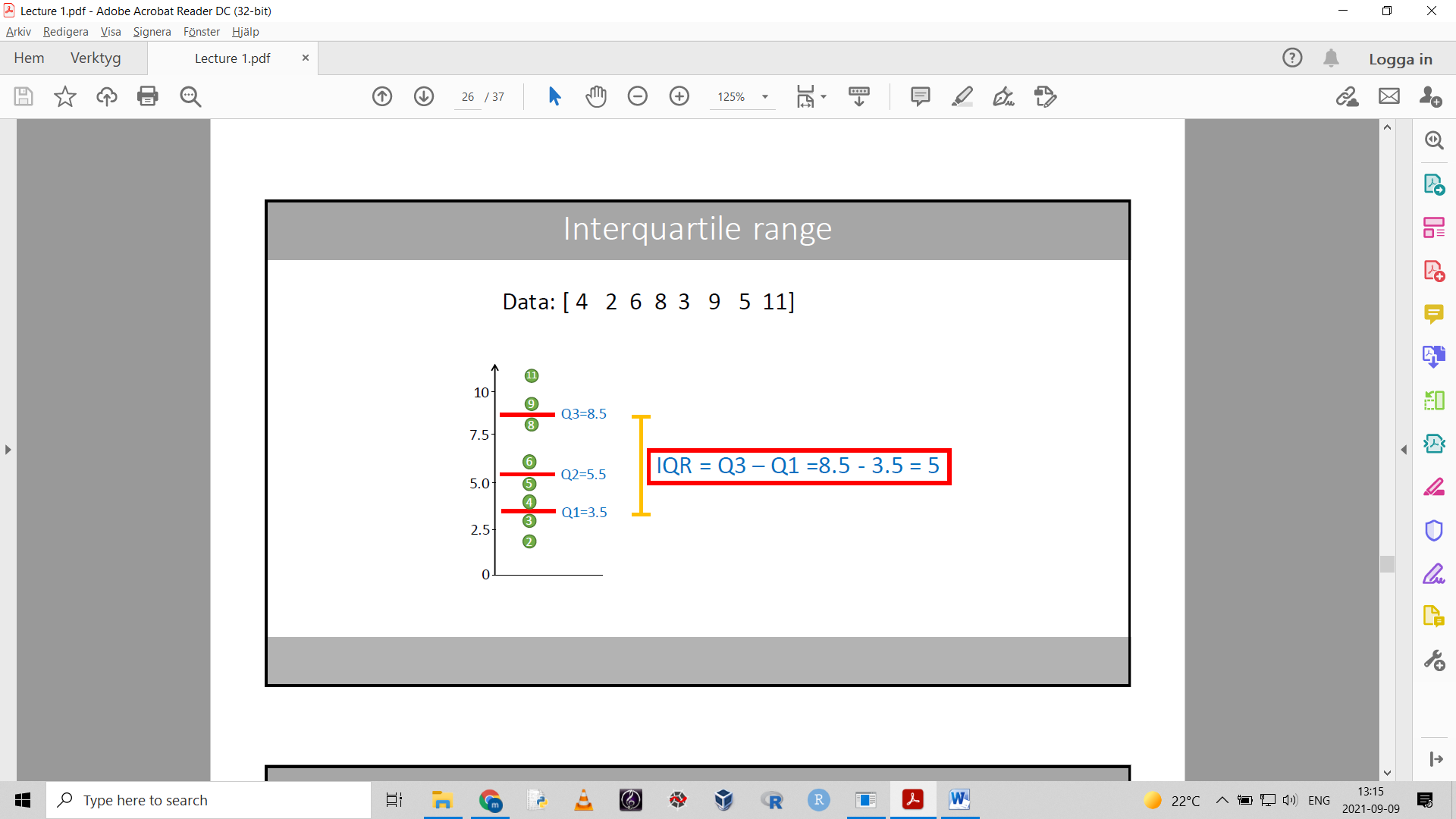
- **Range**

The difference between the largest and smallest value in a data set.

- **interquartile range IQR**

The difference between the third quartile and the first quartile.

Quartiles are values that divide ordered numbers into quarters.



Q2=median

Q1=median of the lower half of the data

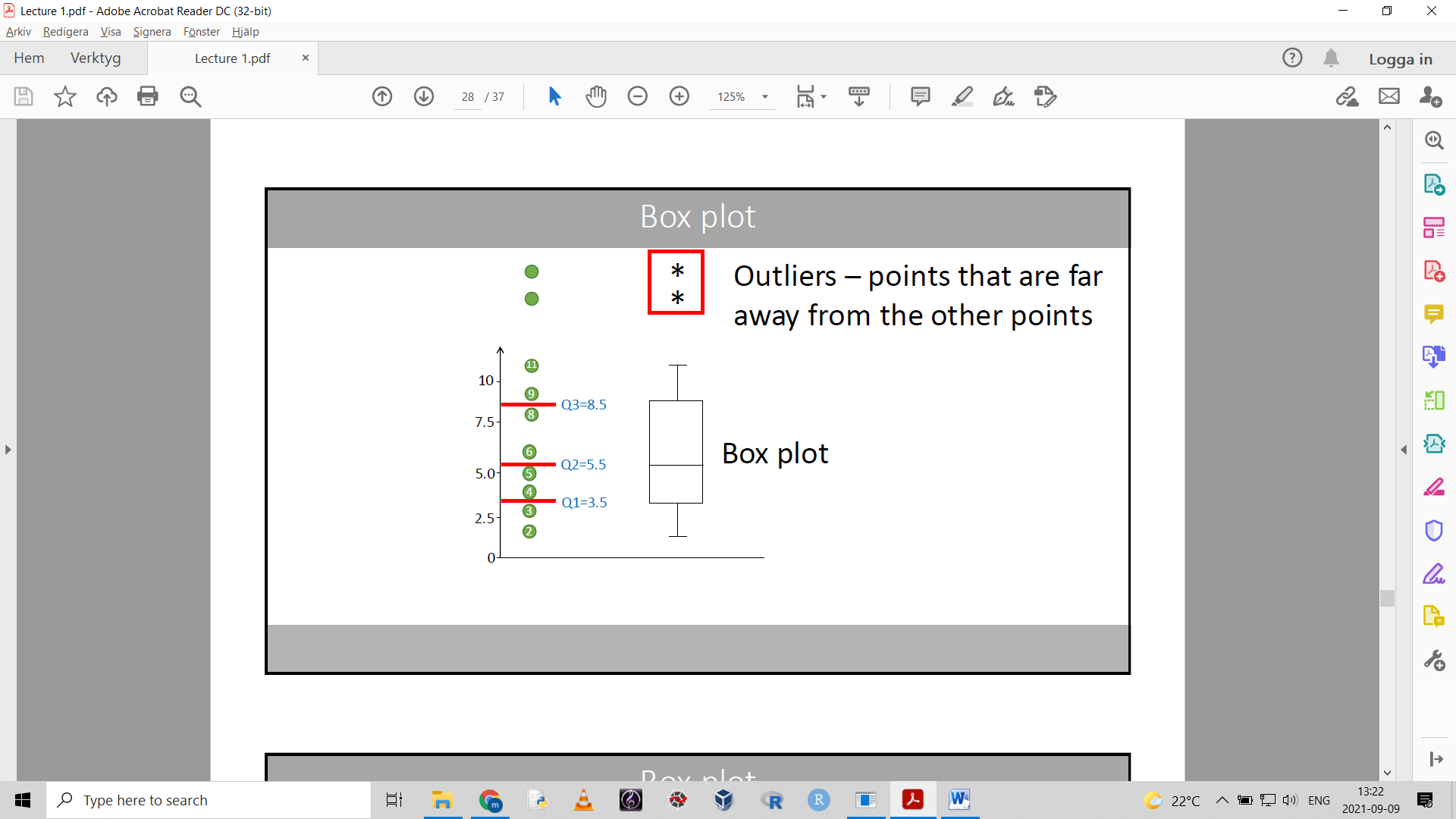
Q3=median of the upper half of the data

IQR= the range of the middle 50% of the data.

Can also be written as an interval [3.5 8.5] – 50% of the data points are located within this interval

**Box plot / box and whisker plot**

Graphical representation of the median, quartiles and range. May also show outliers.



**Quantiles and percentiles**

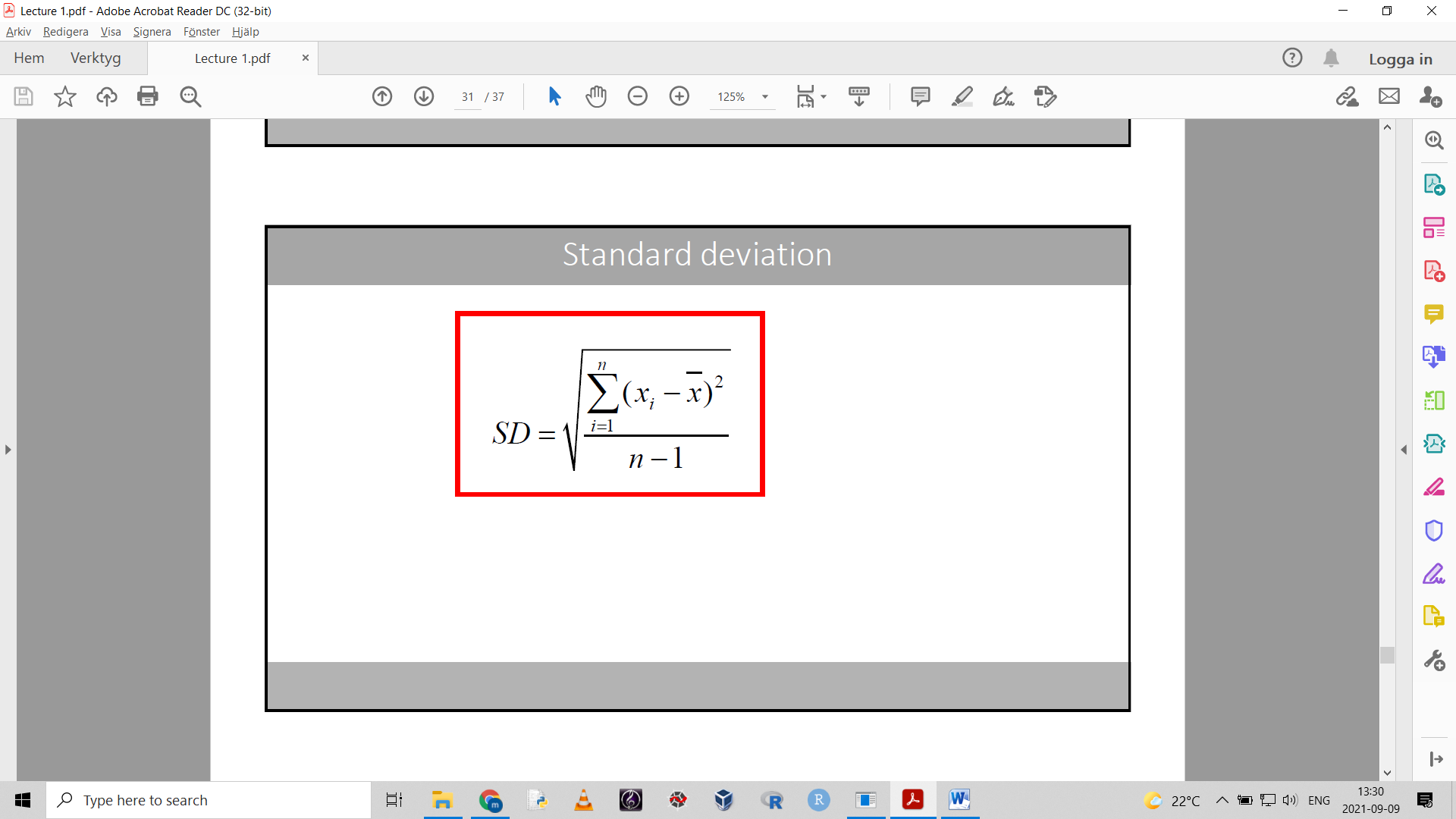
Q3 = 0.75 quantile = 75th percentile

Q2 = 0.5 quantile = 50th percentile

Q1 = 0.25 quantile = 25th percentile

You can also say for example 0.875 quantile.

- **Standard deviation or variance**



Usually presented with the mean like this: 42

One standard deviation away from the mean in both directions: [2 6]

Usually represented graphically in **bar charts**.

- the height of the bar = the mean

- the “error bar” = one standard deviation

About 68% of the data points can be found within SD. Denoted by the error bars.

- given normal distribution

The sample variance is related to the standard deviation: variance = (standard deviation)2

- no simple interpretation. Mainly used in theoretical statistics

MEAN is used with SD

MEDIAN is used with IQR

# Lecture 2

## The normal distribution

Aka Gaussian distribution – the most common distribution used in statistics.

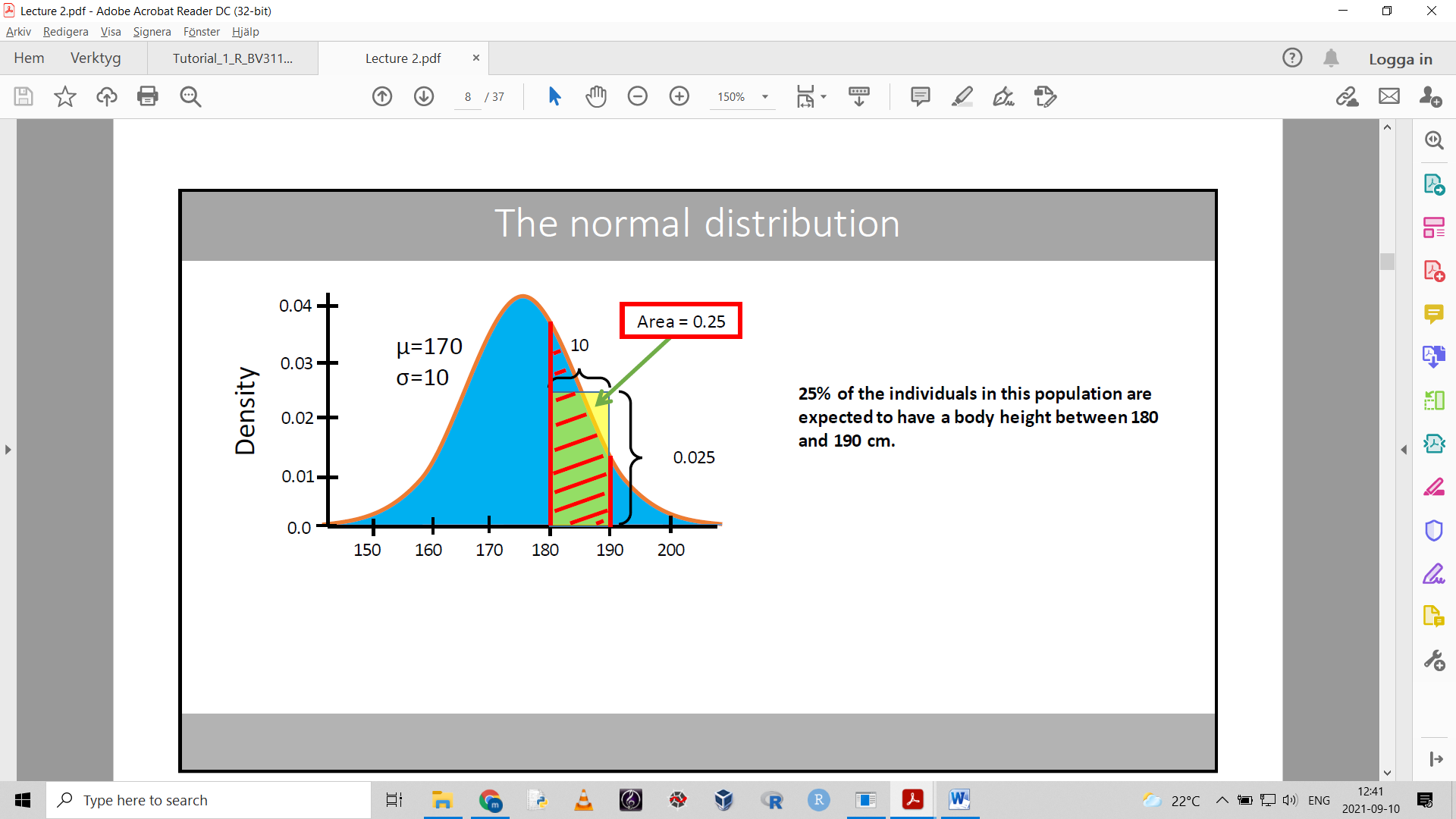
Histograms show distribution of data.

The normal distribution has a mean and a standard deviation.

The mean μ denotes where the center of the curve is.

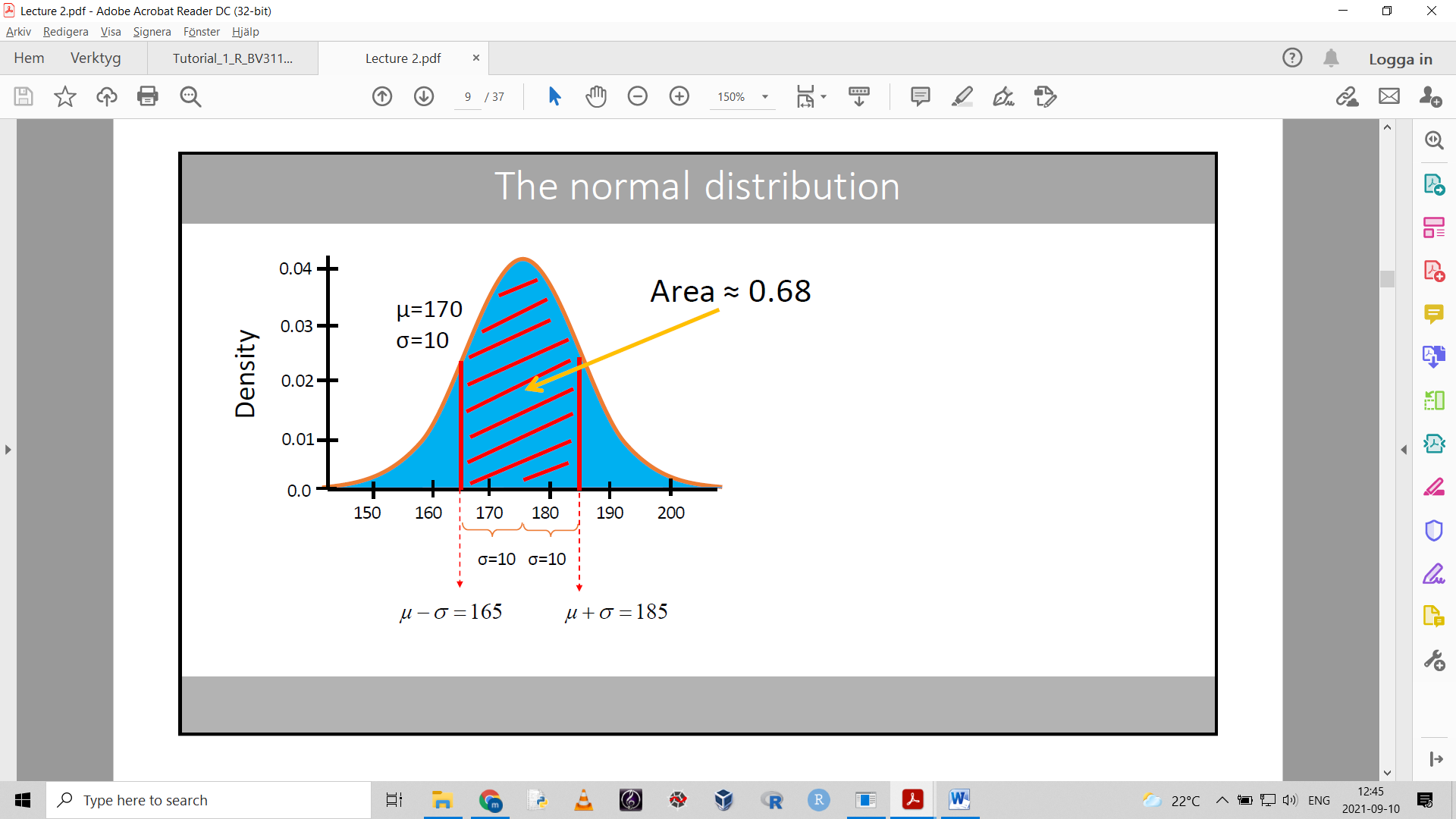
The standard deviation σ determines the width of the curve

For a normal distribution curve the area under the curve is always 1.

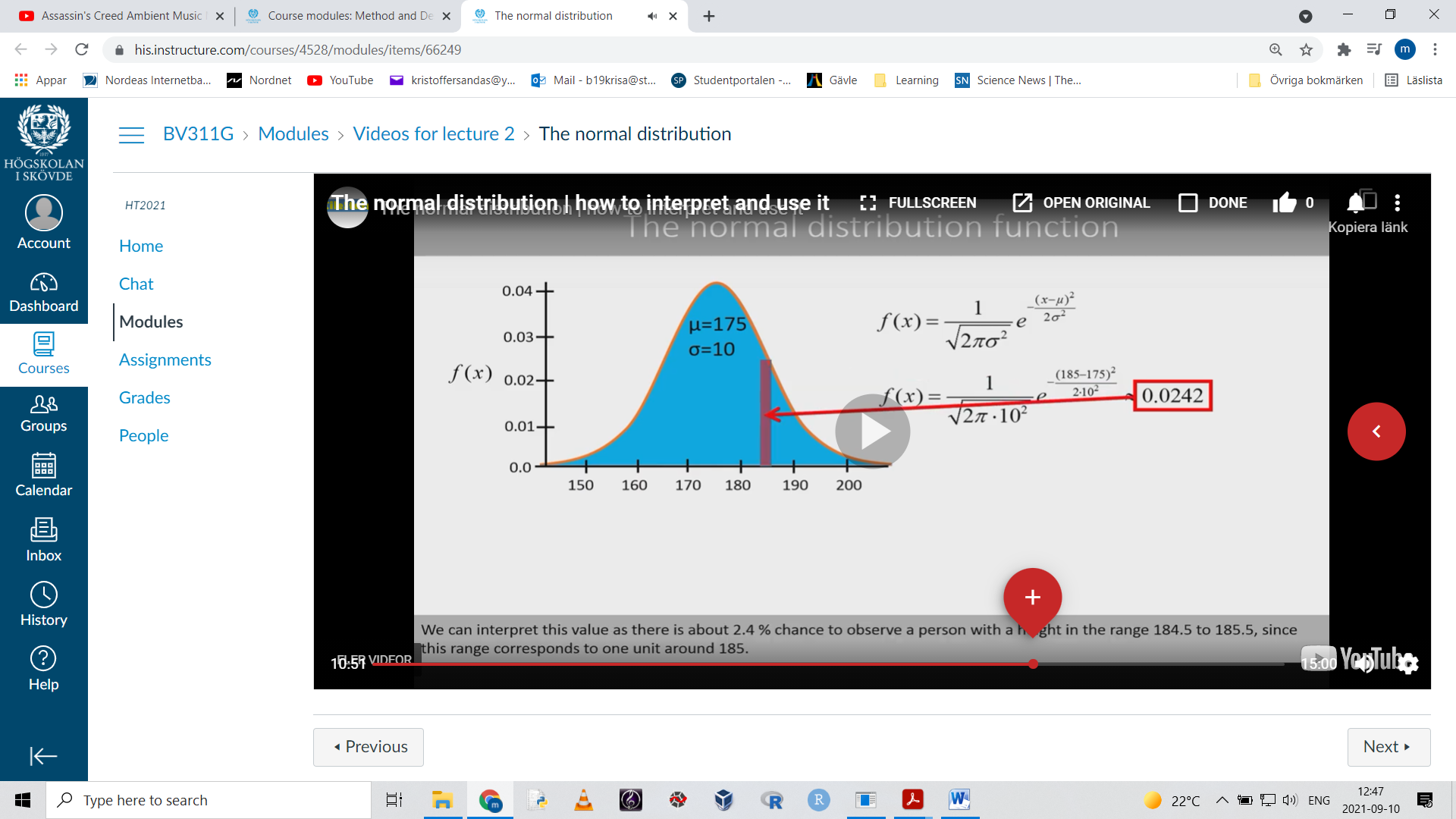


The following notation can be used: P(180<X<190)

- what is the probability of observing a value between 180-190 in this normal distribution.

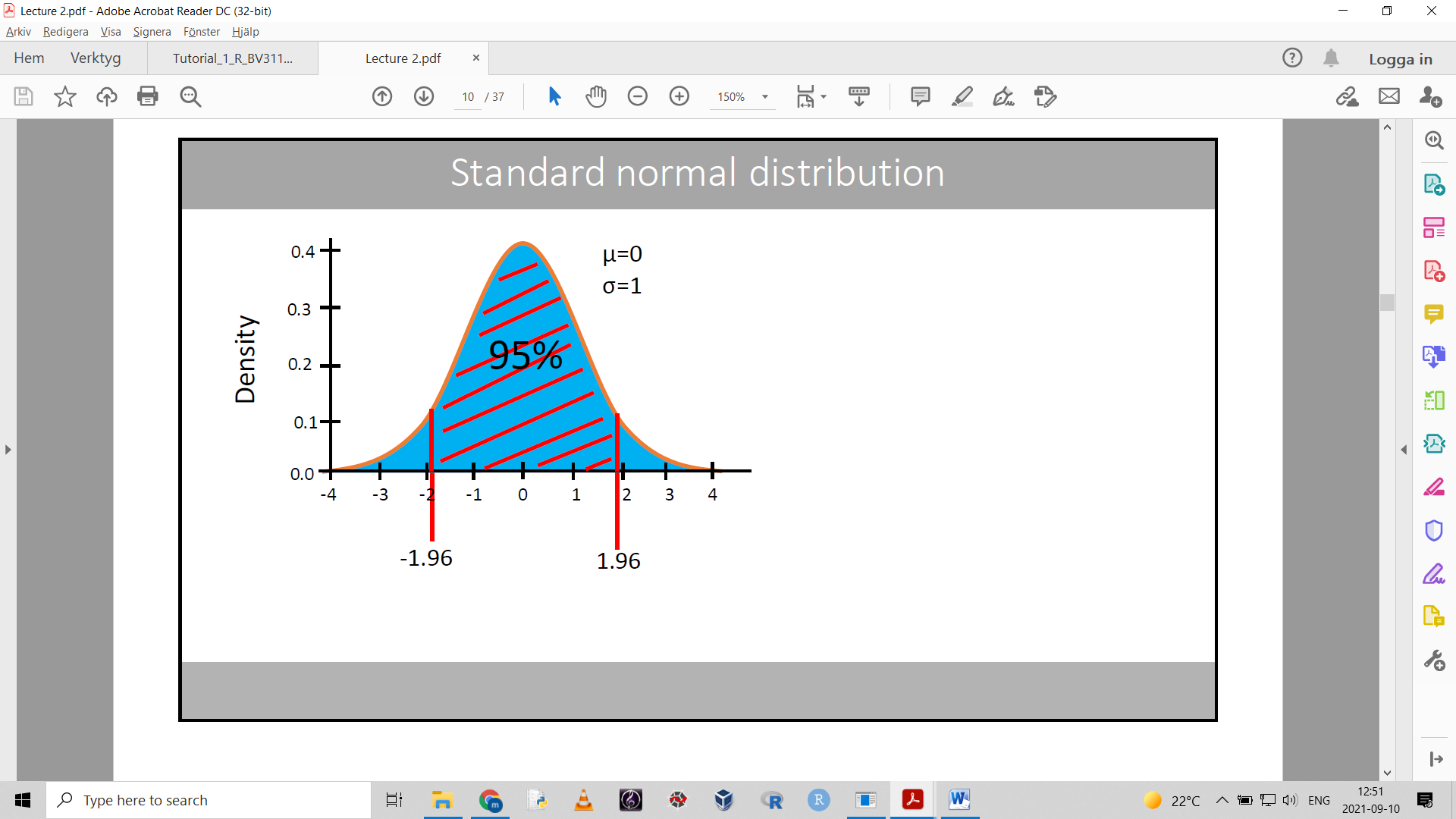


+/- 1 SD covers 68% of the area.



This is the normal curve function.

## The standard normal distribution



### Z-score

We can use a relationship between a specific normal distribution (x) and the standard normal distribution (Z).

Standardize numbers from x like so:

This gives the corresponding value on Z – or the **Z-score**.

For example, the x value 180 from our previous normal distribution:

The corresponding value for 180 on x is 0.5 on Z. The Z-score is 0.5. 0.5 standard deviations on Z.

The probability of finding a person shorter than 180 (x<180) is the same as the probability that the Z-score is smaller than 0.5 in the standard normal distribution. -> P(x < 180) = P(Z < 0.5)

The formula also works in reverse:

This lets us find a specific value on a normal distribution corresponding to a certain Z-score.

## The central limit theorem

= sample means based on large samples will be approximately normally distributed, even though the distribution we sample from is not normal.

Or

The mean values will be normally distributed even though the original distribution is non-normal.

If we increase the sample size (n), we approach normal distribution.

The more the original distribution deviates from normal, the larger sample size we need to approach normality.

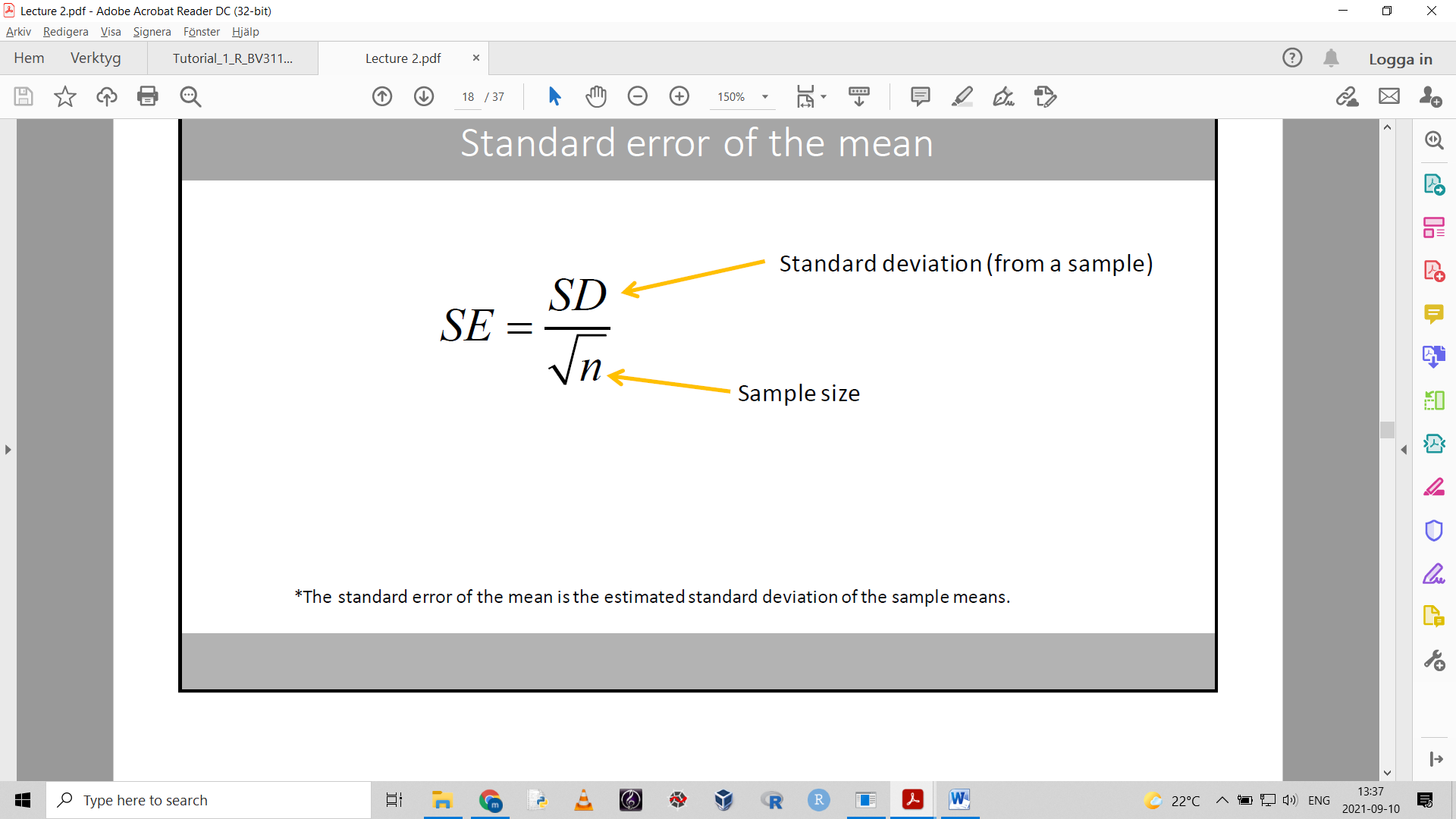
A sample size of n=30 is usually large enough to assume that the sample means are normally distributed for most underlying distributions.

It is important to know that the mean values are normally distributed because:

Most statistical tests assume normal distribution. Ex P-value, confidence interval.

(A t-statistic is the difference between two means divided by the standard error of that difference.)

## Standard error of the mean – SEM



The SEM is an estimate of the expected standard deviation of the sample means.

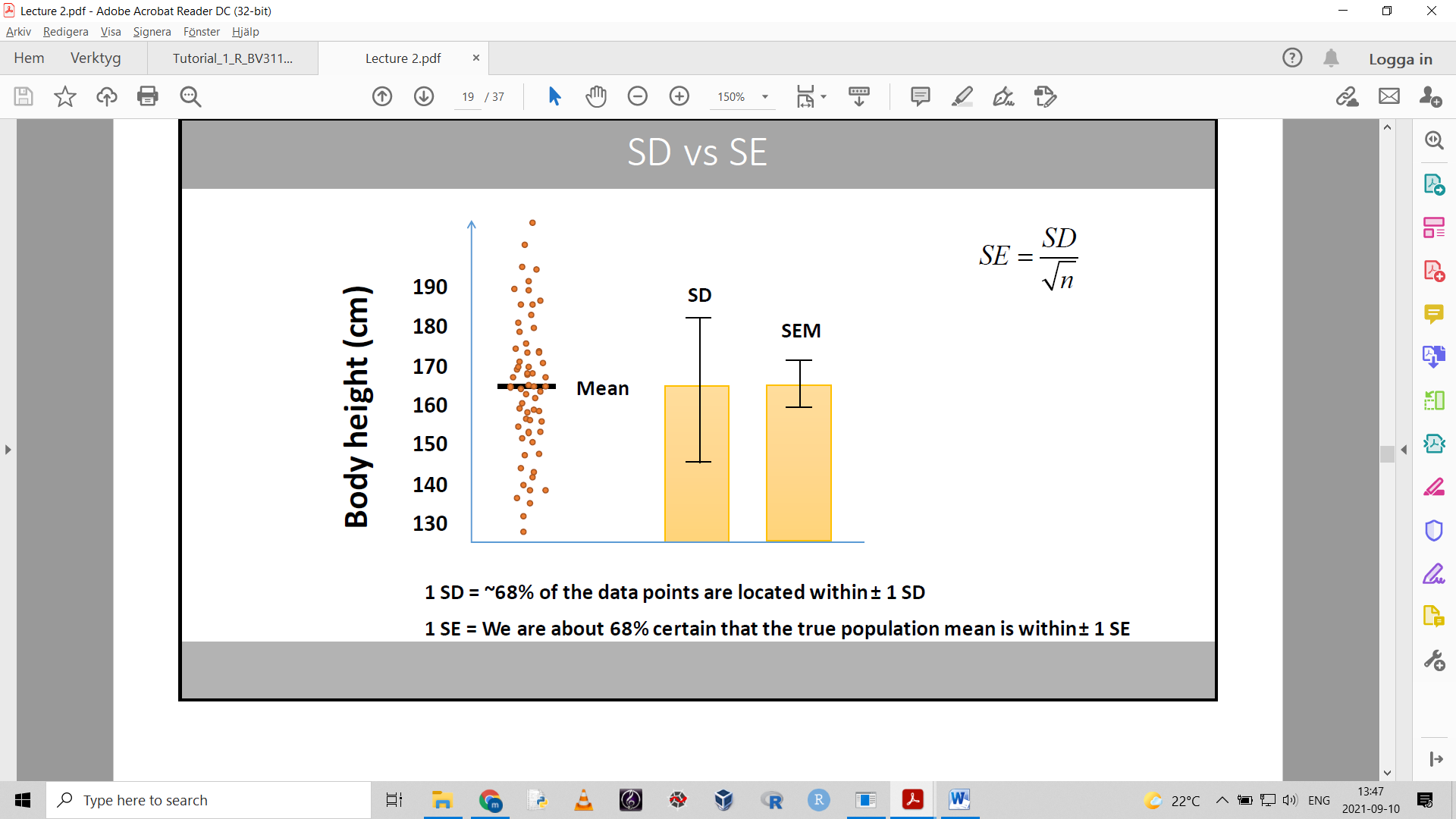
Tells us the expected spread of the sample means.

We expect that 68% of the sample means can be found within +/- SEM

The SE decreases when n increases. If we increase sample size, the SE decreases.

The SEM is a measure of the expected spread of the sample means around the population mean.

SEM is the estimated standard deviation of the sample means.



## Confidence intervals

The mean +/- the SE

Upper limit: mean + SE

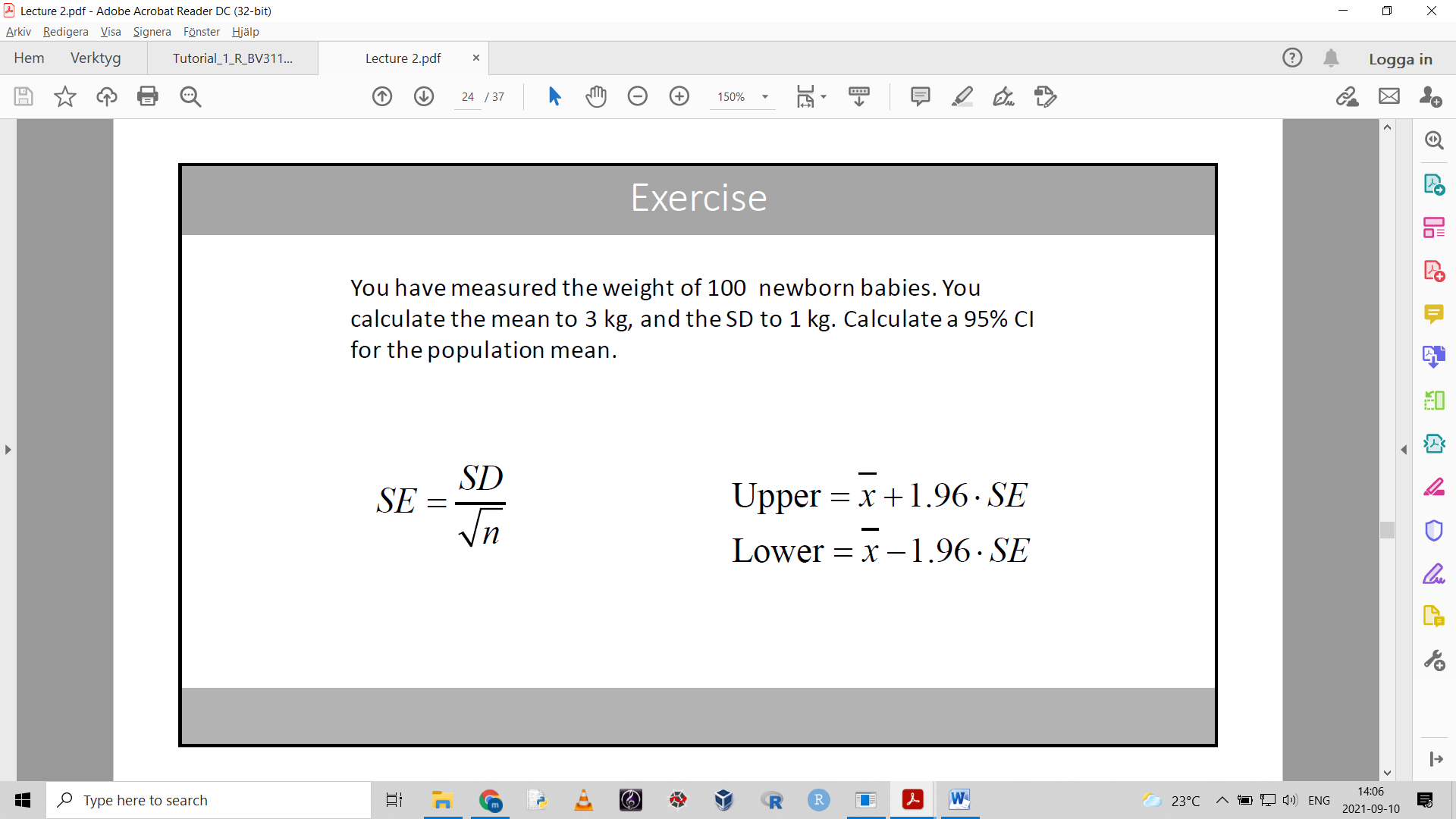
Lower limit: mean – SE

68% confidence interval: [lower upper]

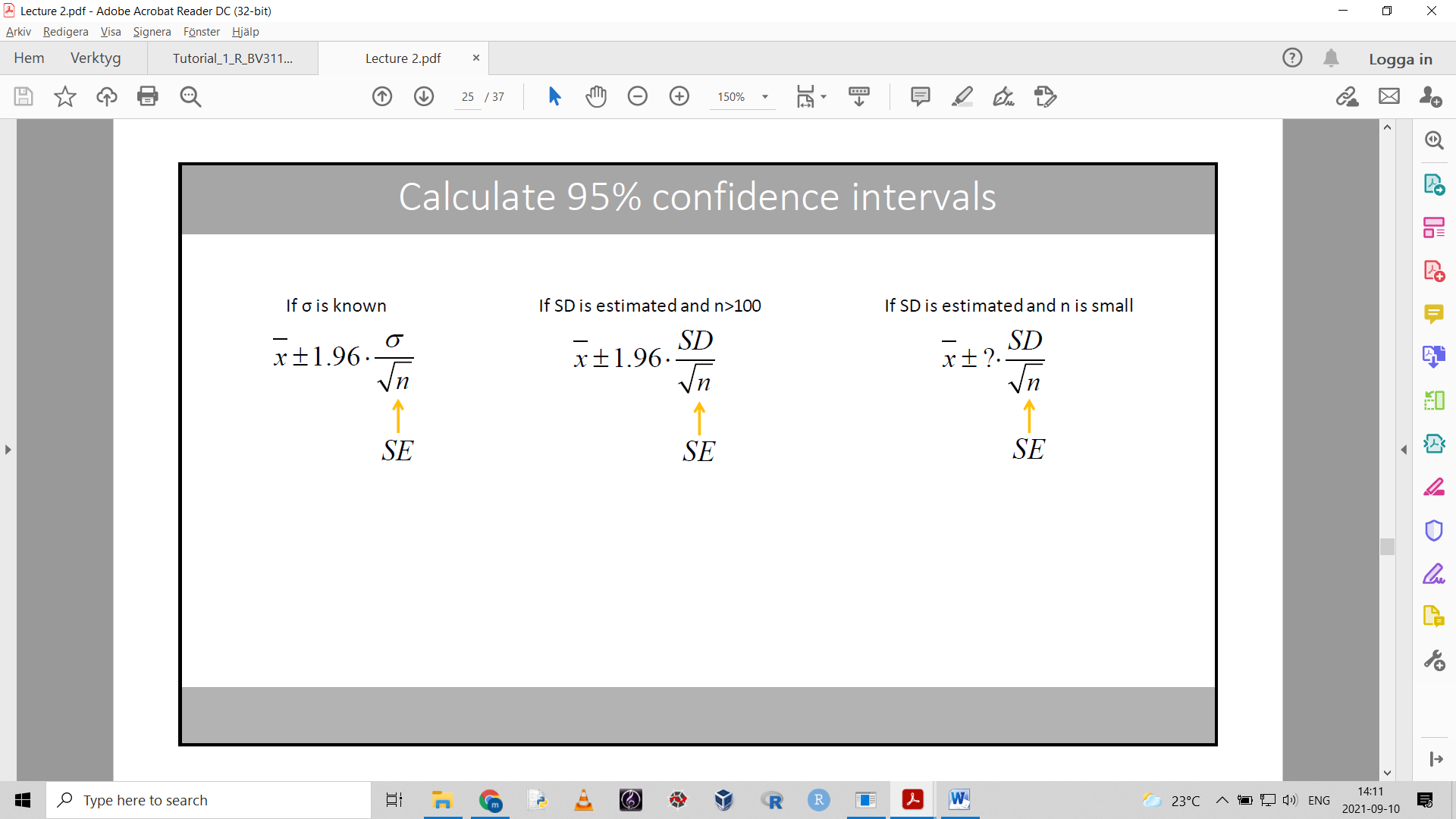
The standard error tells us how certain (in percent) we are that the true mean lies within the range.

In statistics, we want to be 95% certain.

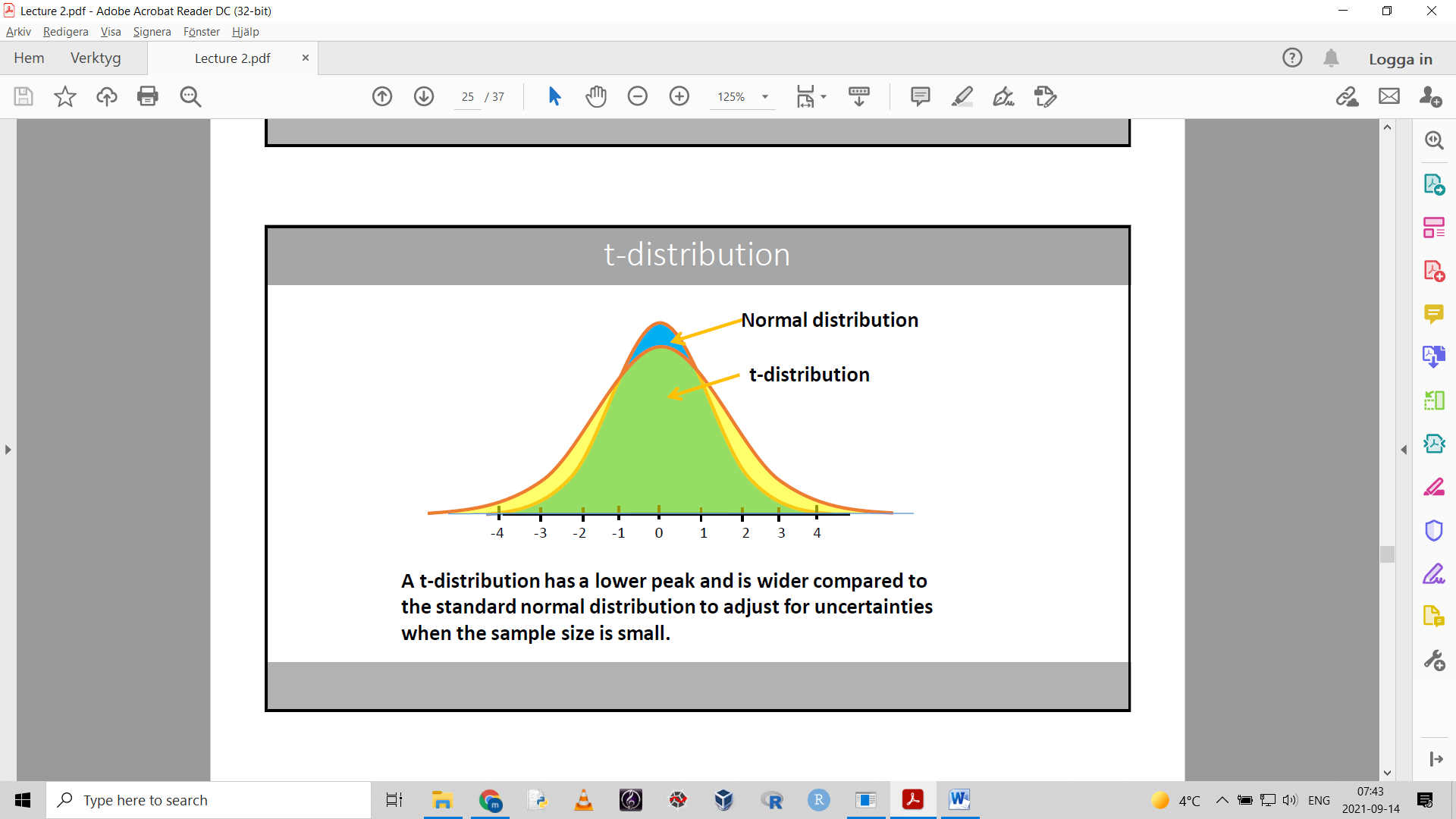
To get from 68% confidence interval to 95% we multiply the SE by 1.96



We can only use this (multipying with 1.96) if n > 100. If n is smaller, we need a t-distribution.



## T-distribution



The t-distribution curve will change shape depending on the degrees of freedom.

As the degrees of freedom increase, the t-distribution approaches the normal distribution.

At df=100 (n=101), the t-distribution is virtually the same as the normal distribution.

**Sort of extra material:**

This formula uses the symbols for t-distribution stuff. The Z means we should use a Z-score from the standard normal distribution for a certain value of alpha.

If we want for ex a 95% confidence interval, we use α = 1 - 0.95 = 0.05

Divided by 2, this gives us 0.025, which corresponds to one of the tails in the standard normal distribution. We use the Z-score for that tail (for example Z = 1.96).

If our sample size is small and we need to estimate the standard deviation:

t tells us that we should use a t-score from a t-distribution with n-1 df, for the same given value of alpha.

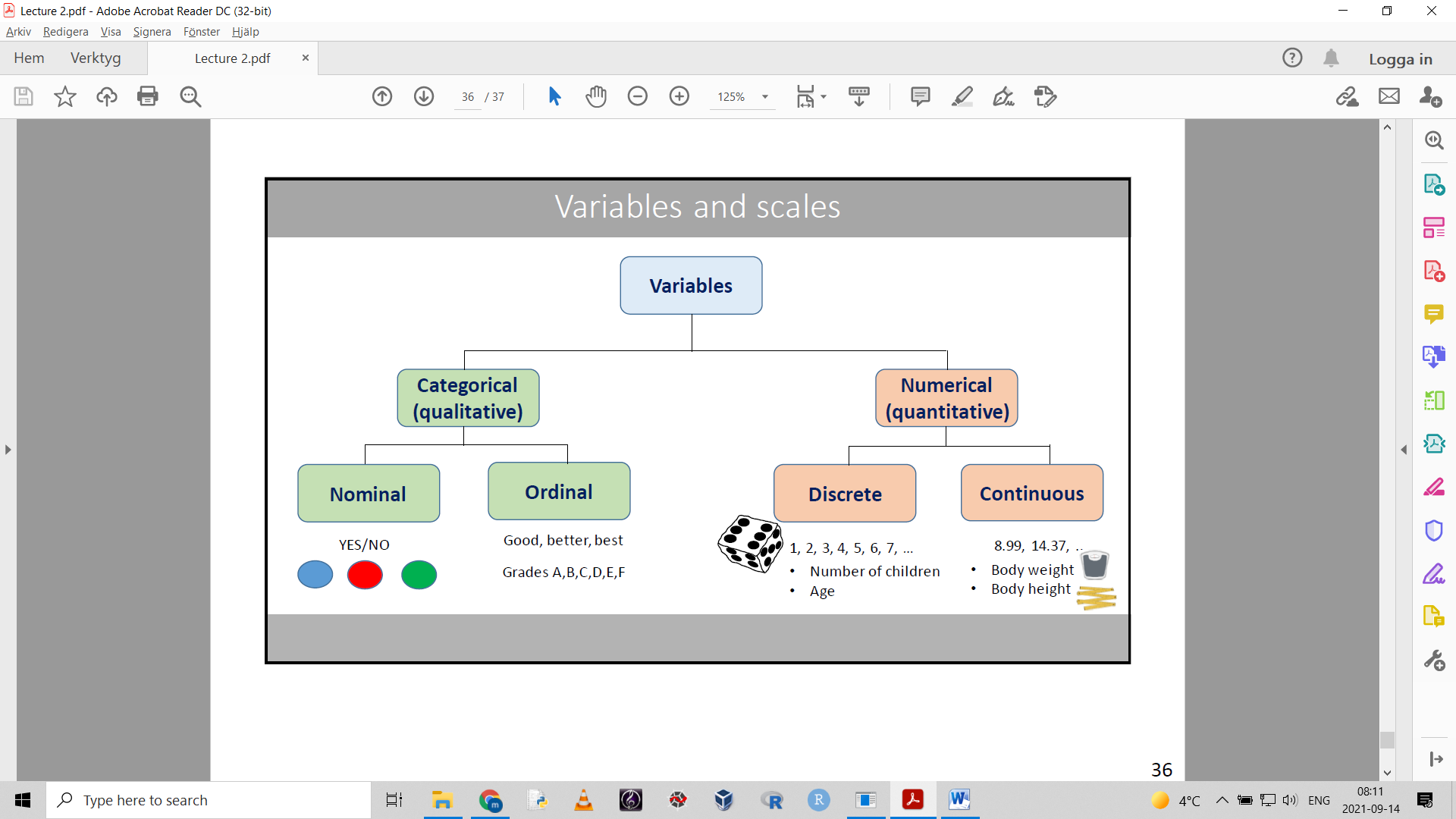
## Degrees of freedom

The number of values in the final calculation that are free to vary.

A simplified interpretation: df = n-1 when we estimate one parameter.

Generally: the df = number of observations – numbers of intermediate parameters we need to estimate

## Variables and scales

  
Nominal = cannot be ordered. Ordinal = can be ordered in a reasonable way.

A variable is something that can be measured or counted.

A parameter is a value that represents a characteristic of a population: ex mean or proportion.

We can change the scale for variables. For ex categorize a continous scale int low, med and high.

# Lecture 3

## The one sample t-test and p-values

The one sample t-test checks if a mean value from a sample is significantly different from a hypothesized or known value.

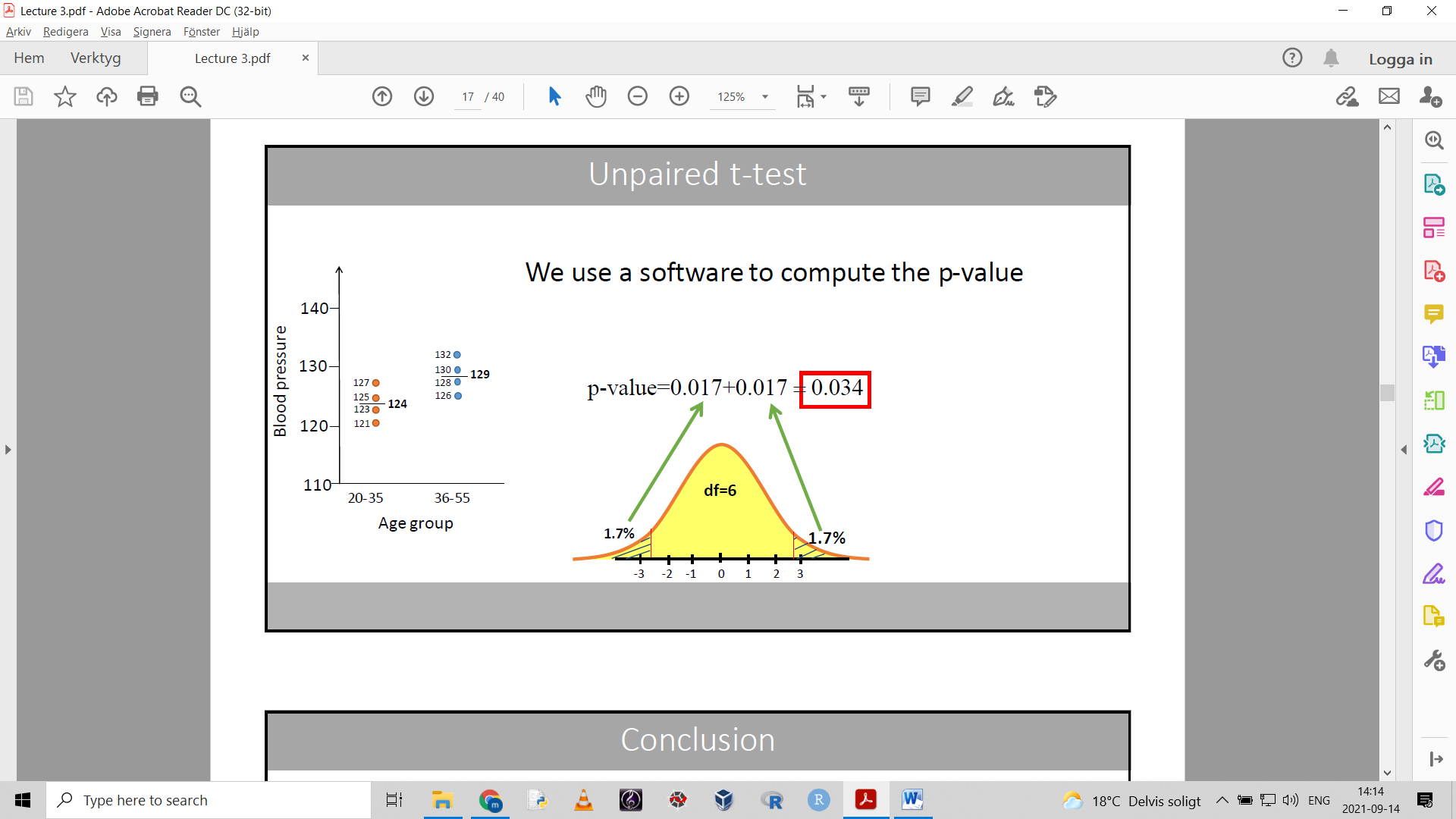
The formula for calculating the t-statistic for a one-sample t-test:

x-bar=estimated mean value

SE = standard error of the mean

μ0 = the hypothesized value (can be seen as a reference value)

The t-statistic is used on the t-distribution corresponding to the appropriate degrees of freedom.



Here for example the t-statsitic should be about 2.8

The area in the t-distribution curve that corresponds to the tails when we plugged in our t-statistic is the p-value.

the p-value is always between 0 and 1

If we assume that there is no effect, the p-value can be interpreted as the probability that the observed effect (or a more extreme effect) could be due to chance.

If p-value > 0.05: weak evidence of effect. (there is a 5% chance that the observed effect is due to chance?)

p-value 0.01-0.05: moderate evidence of effect

p-value < 0.01: strong evidence of effect

There is a significant effect only if p-value < 0.05

## The basic steps of hypothesis testing

Hypothesis = statement that can be tested true/false

In statistical hypothesis testing you test a null hypothesis against an alternative hypothesis

The null hypothesis is generally a statement against your hypothesis

Can be thought of as the idea that there is no difference, no relationship or that everything is normal

Null hypothesis (H0) = there is NO difference

Alternative hypothesis (H1) = there IS a difference, a relationship or things are not normal

Steps:

1. state the H0 and H1

2. choose a significance level (α) for the test (usually 0.05)

3. perform experiment and select appropriate statistical test for data

4. compute the p-value (computed by the statistical test)

5. compare p-value to significance level, to decide whether to reject the null hypothesis

If p-value < α -> reject H0

If we reject the H0, we say that there is a significant difference (we have evidence for H1)

You can create a one-sided or two-sided (most common) hypothesis.

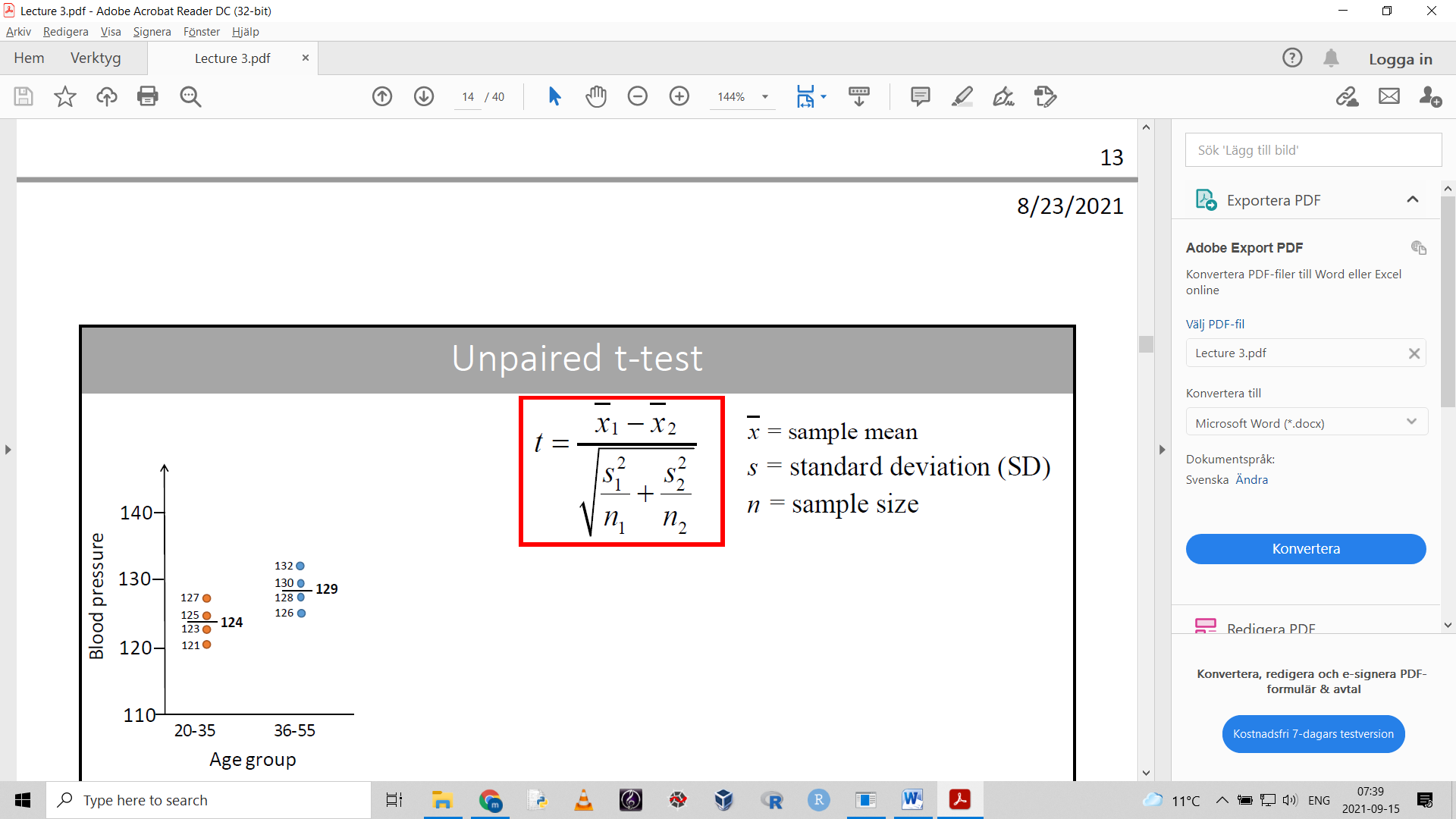
## Unpaired t-test

A t-test is a statistical test where the statistic follows a t distribution.

Unpaired t test = independent (two sample) t test = compares the means of two populations.

(If we happen to know the population variance, the corresponding Z-test is used.)

In an unpaired t test the populations should be independent.

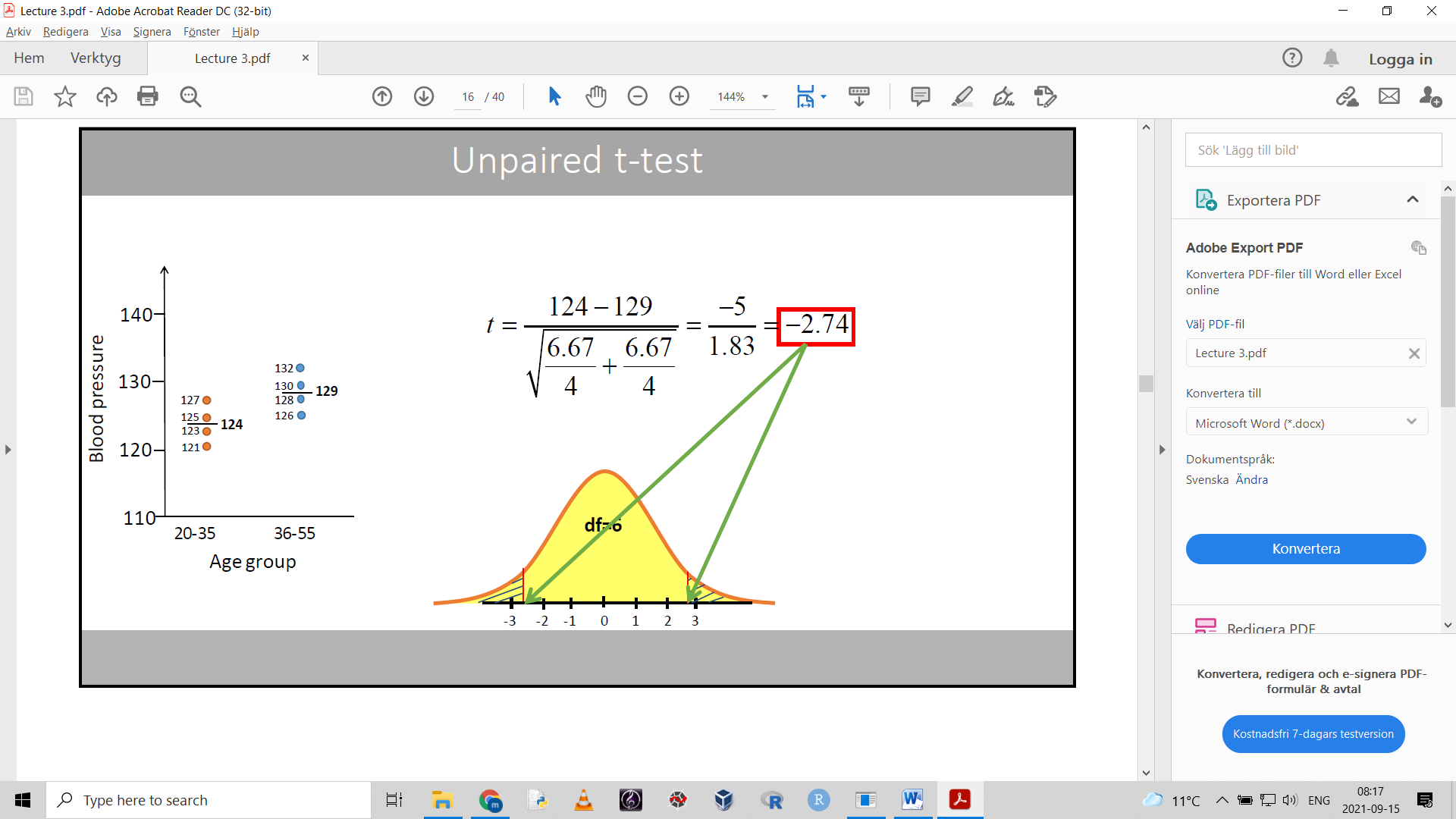


This equation only works if the two groups have the same sample size.

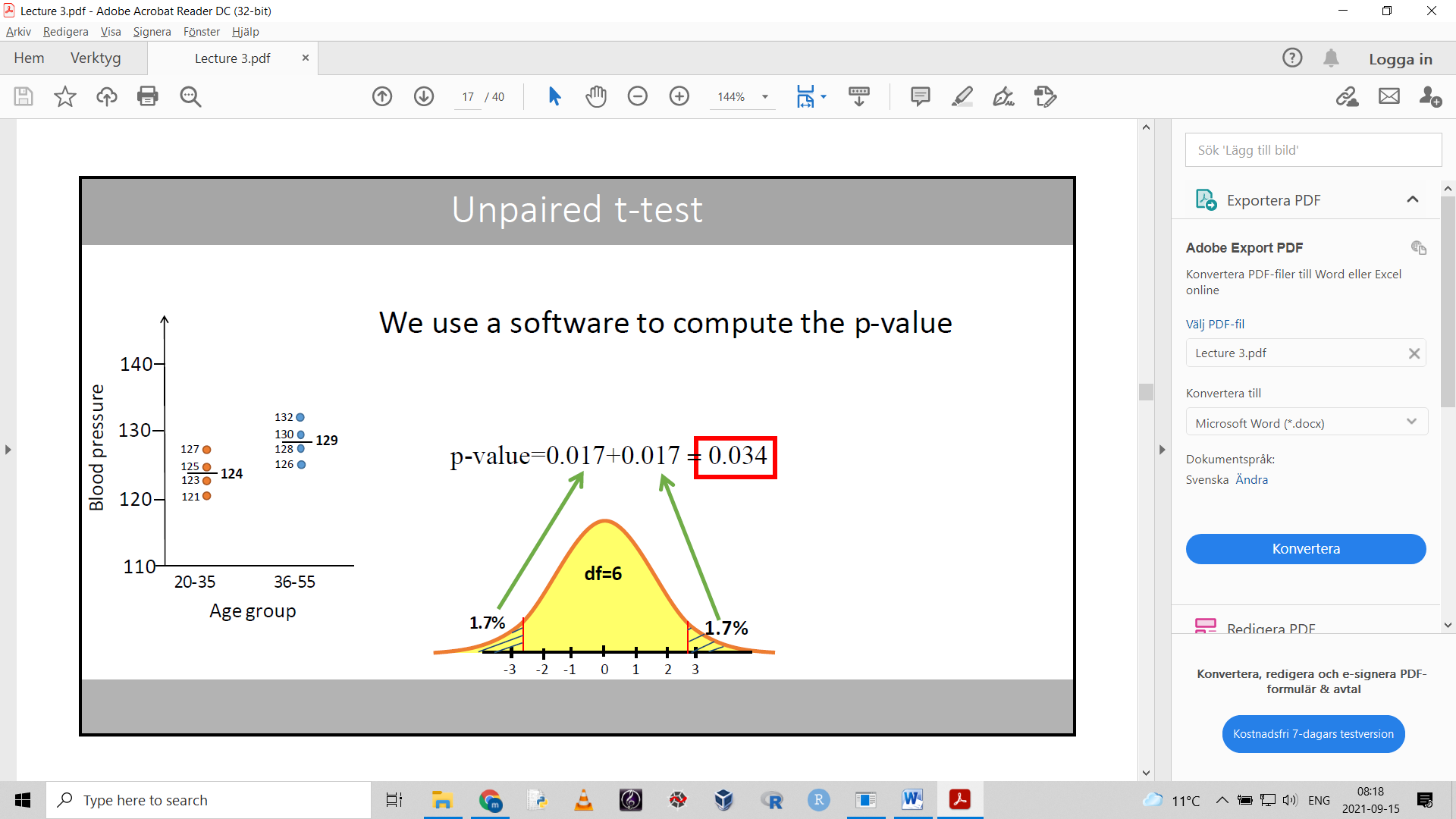
(If sample sizes differ, we put more weight on the larger one.)

We use a t-distribution with df=n-2 because we have estimated two means in the test.

Here is an example of a t-statistic of -2.47:



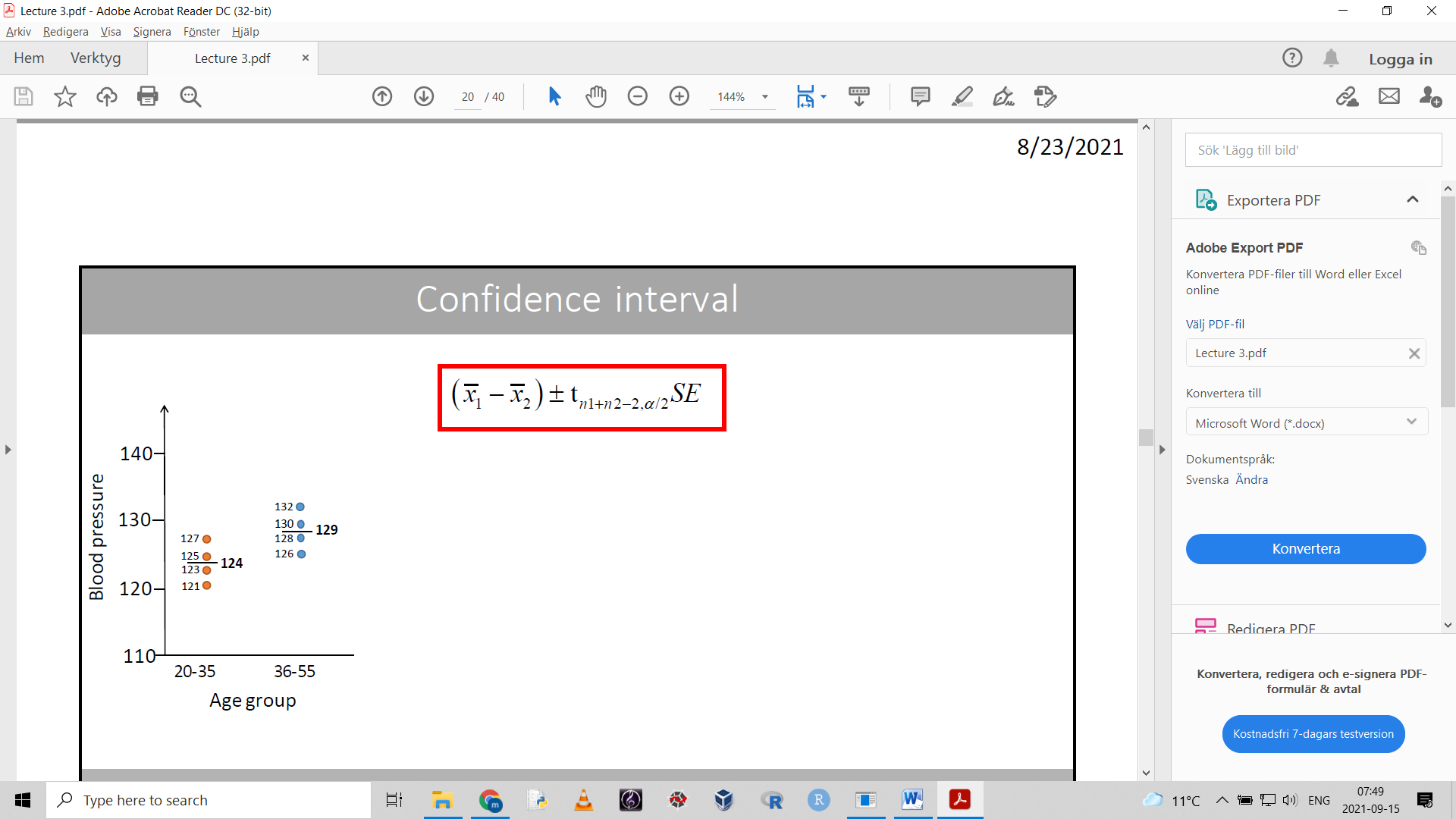
The p-value is the area of the tails.



(the area of the tails represent the H0? And the main area represents H1?)

“If we assume that there is no difference, the p-value can be interpreted as the probability that the observed difference is due to chance.”

The results of a t-test is usually reported with a 95% confidence interval.



Calculated like this. The t value is based on the df and the alpha value (usually 0.05)

If the H0 value is not within the confidence interval, we can reject H0.

The unpaired t test and confidence interval relies on normal distribution and equal variance.

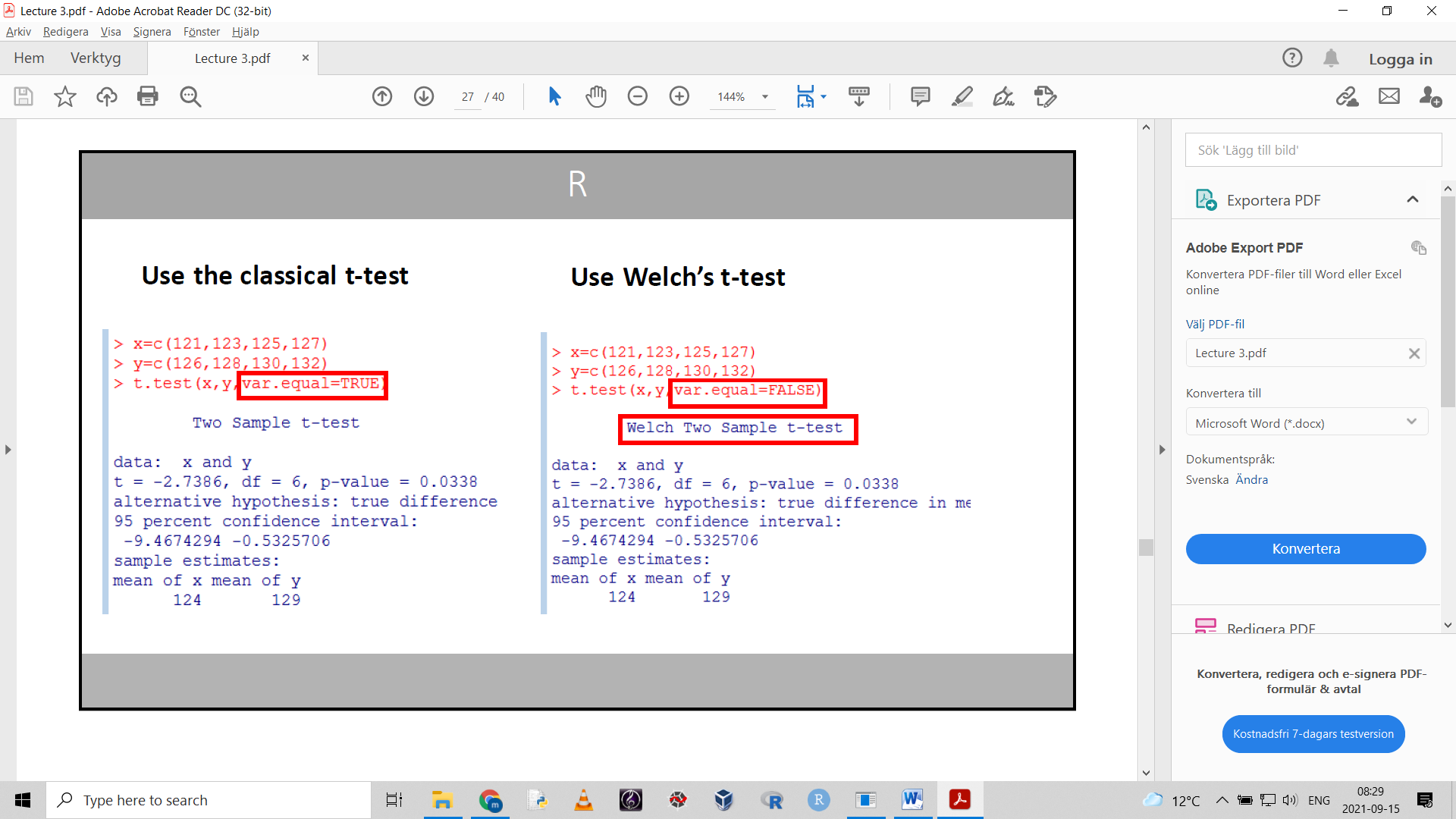
Normality can be tested with shapiro-wilks test.

If the sample size is > 30, the sample means will be normally distributed even if the underlying data is skewed.

If the sample size is small but we know that the underlying values are normally distributed (ex height of humans) we can use the t-test and assume normality.

Calculate and compare variances of the two samples. Levenes test can do this (H0=equal variance).

If the populations don’t have equal variances, we can use Welch’s t-test which can handle unequal variances.



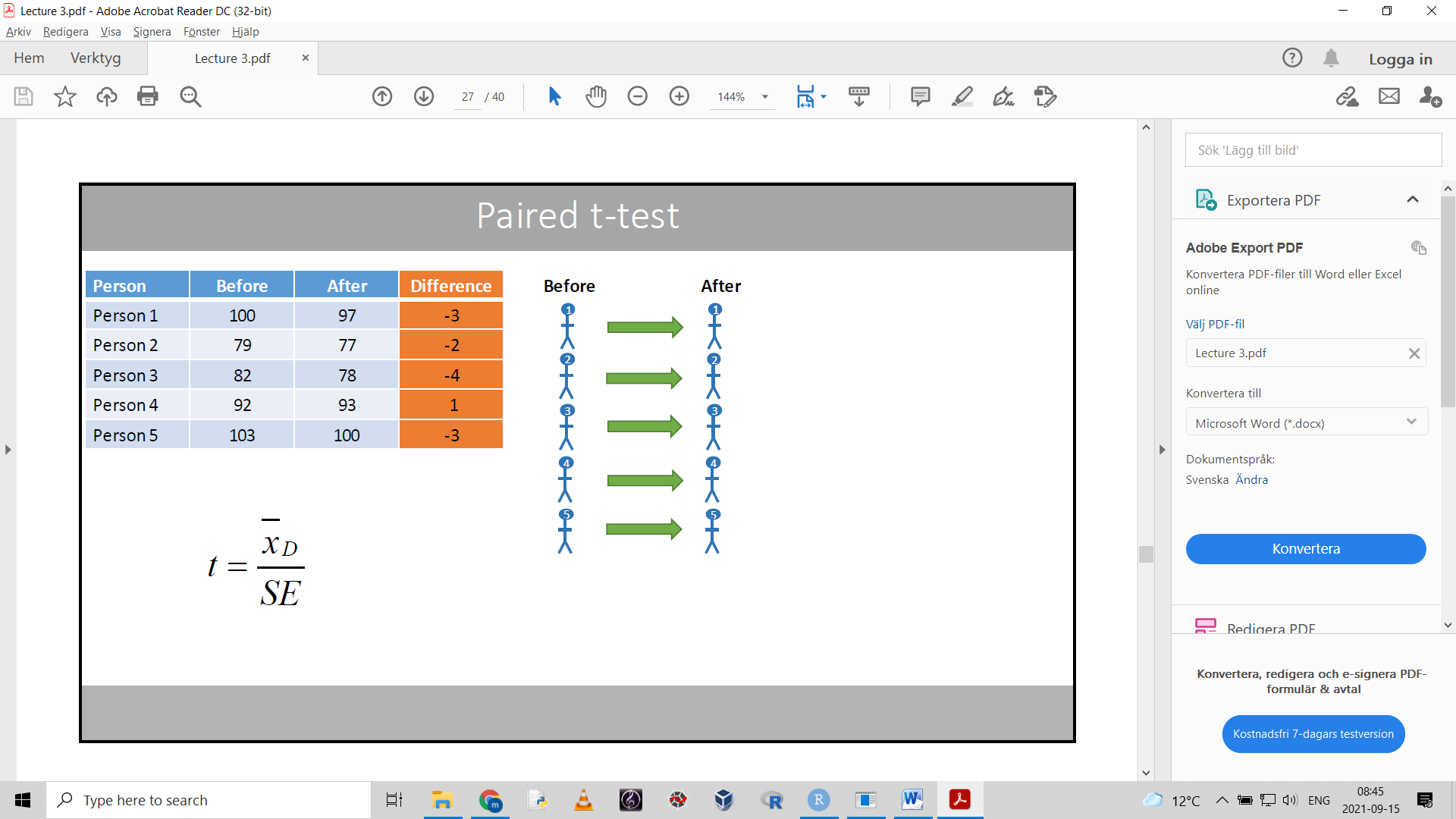
## Paired t test

Aka dependend sample t test.

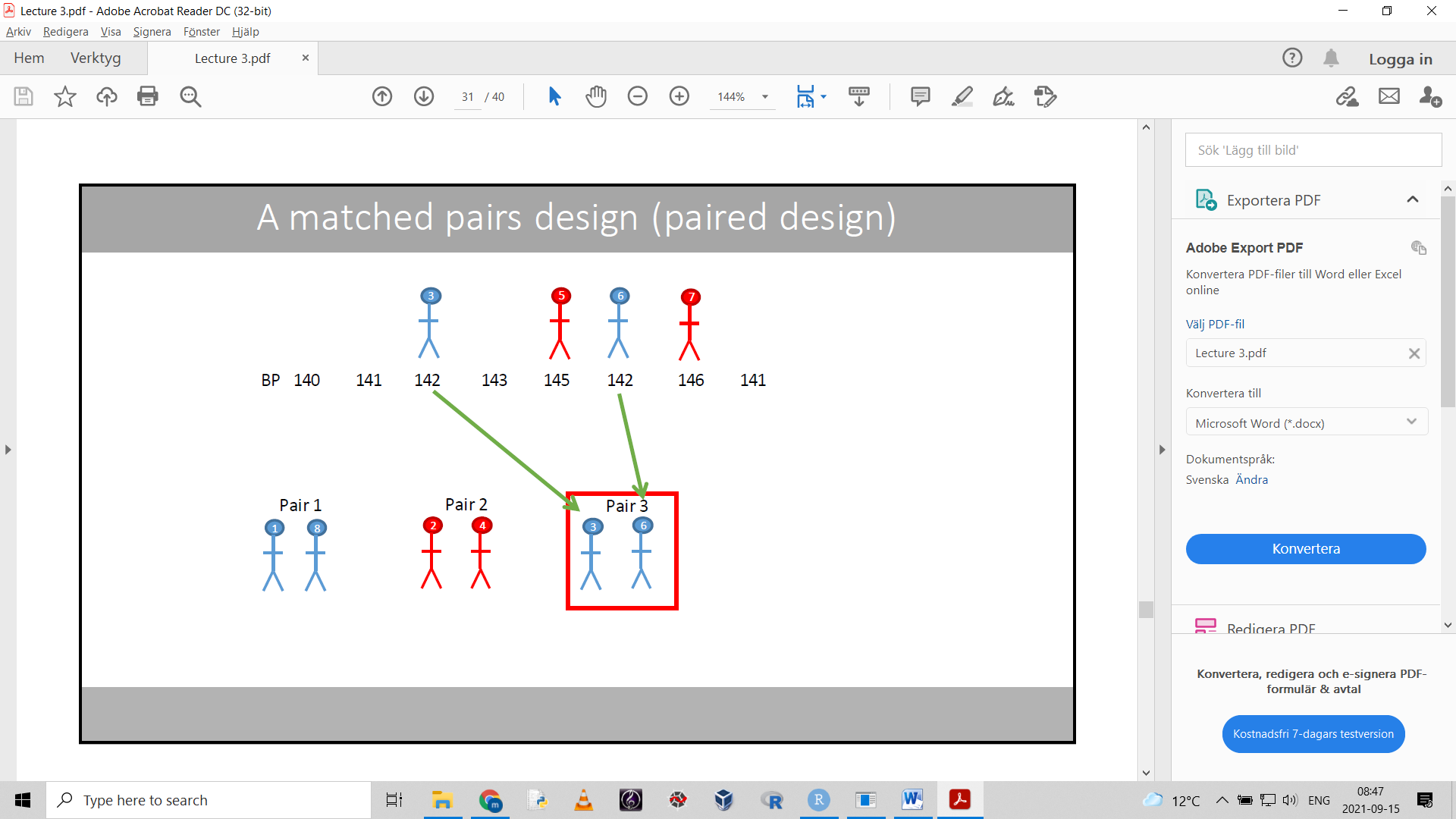
Used to determine if the mean difference from pairs of observations is different from zero.

Compares for example observations on paired individuals.

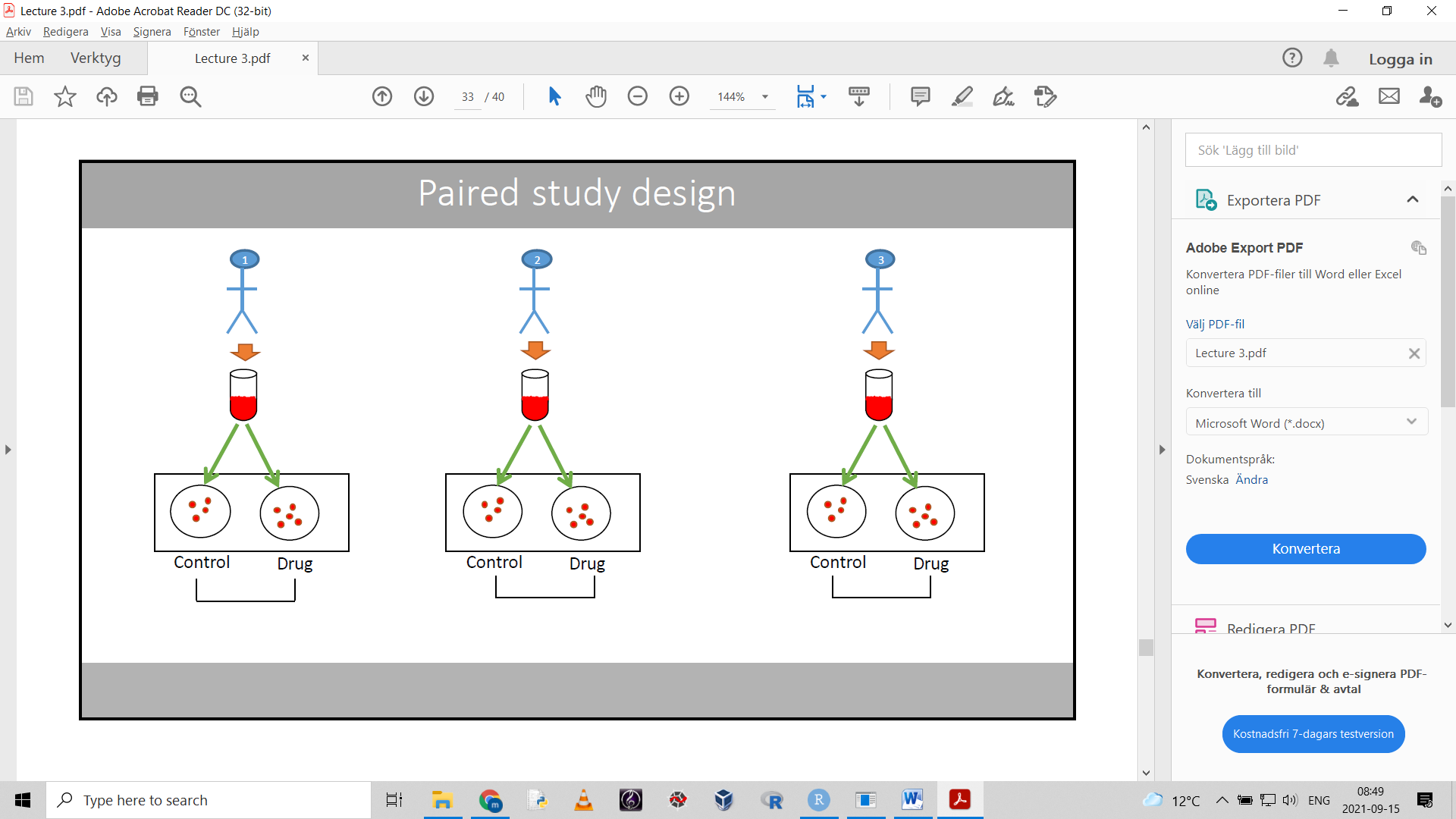
Example: before and after studies



Example: matched pair design



Paired study design:



Heres how its done:

Calculate the mean and the sd of the differences.

The null hypothesis states that the population mean difference between the paired values = 0

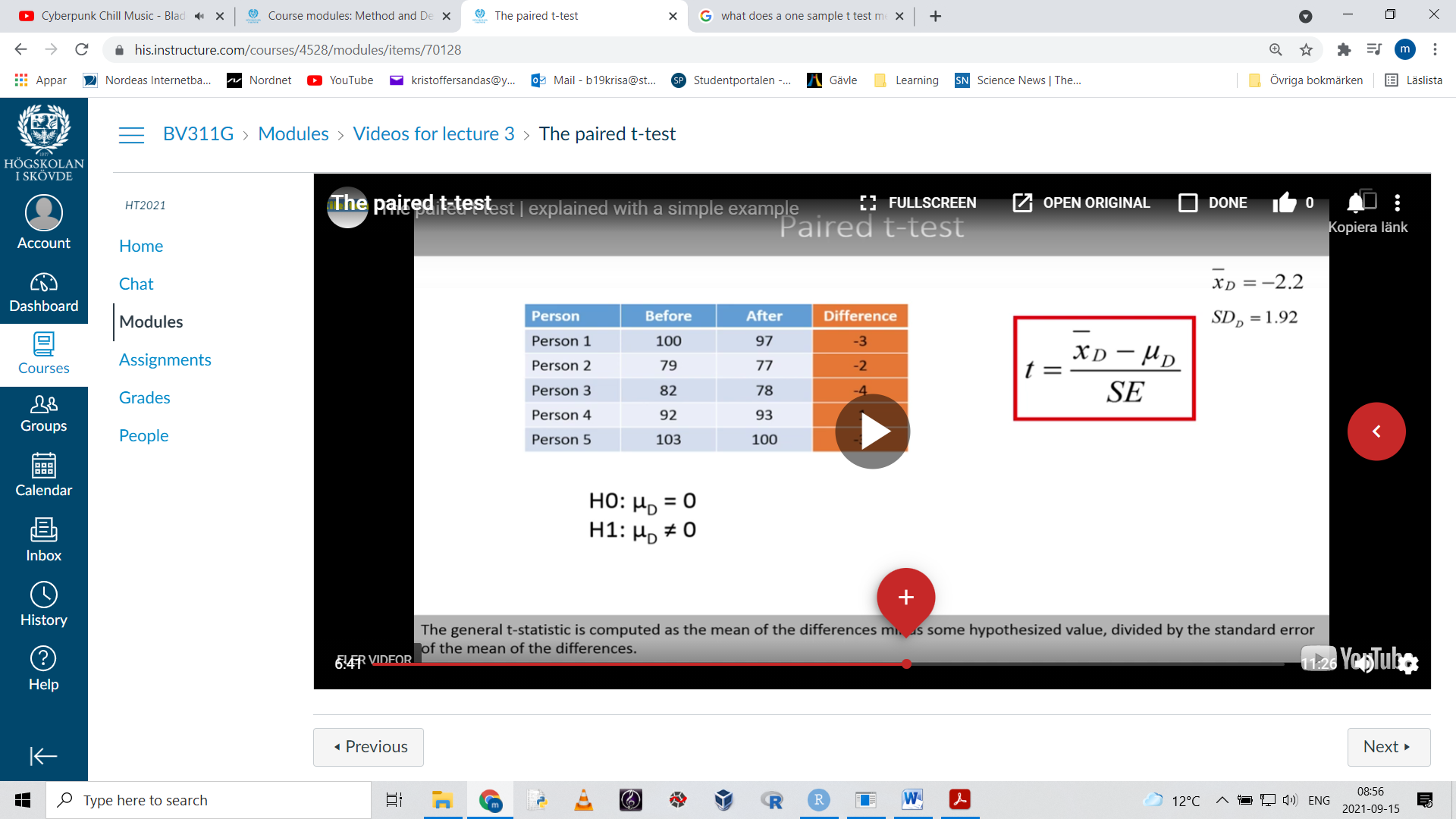
H0: μD = 0

The alternative hypothesis (for a two sided test) states that the population mean difference is not = 0

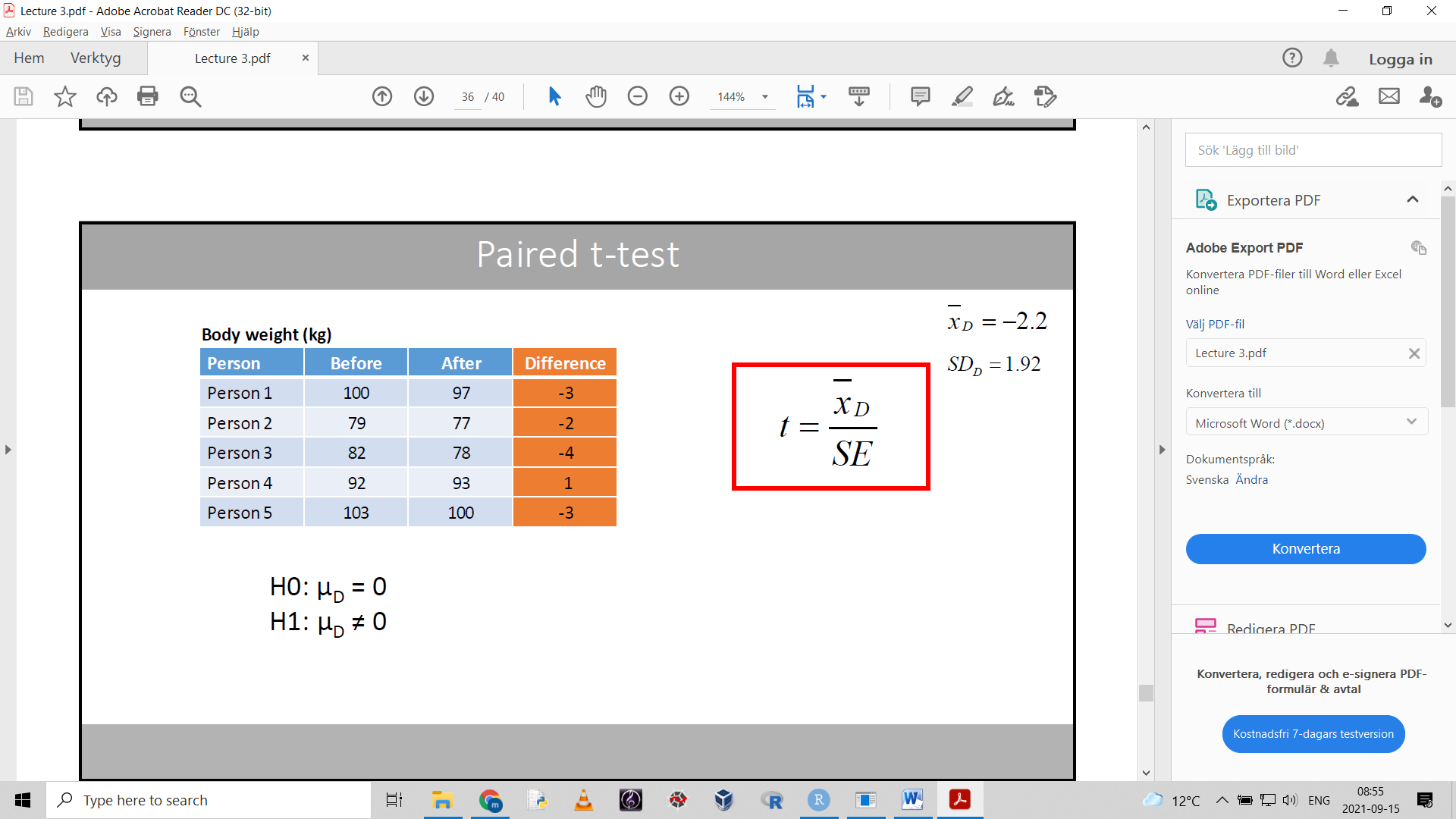
H1: μD  0

We now used a paired t test to see if the observed stuff in not just due to chance.

Formula for the t-statistic:



The mean of the differences minus the hypothesised value, in this case 0, gives:

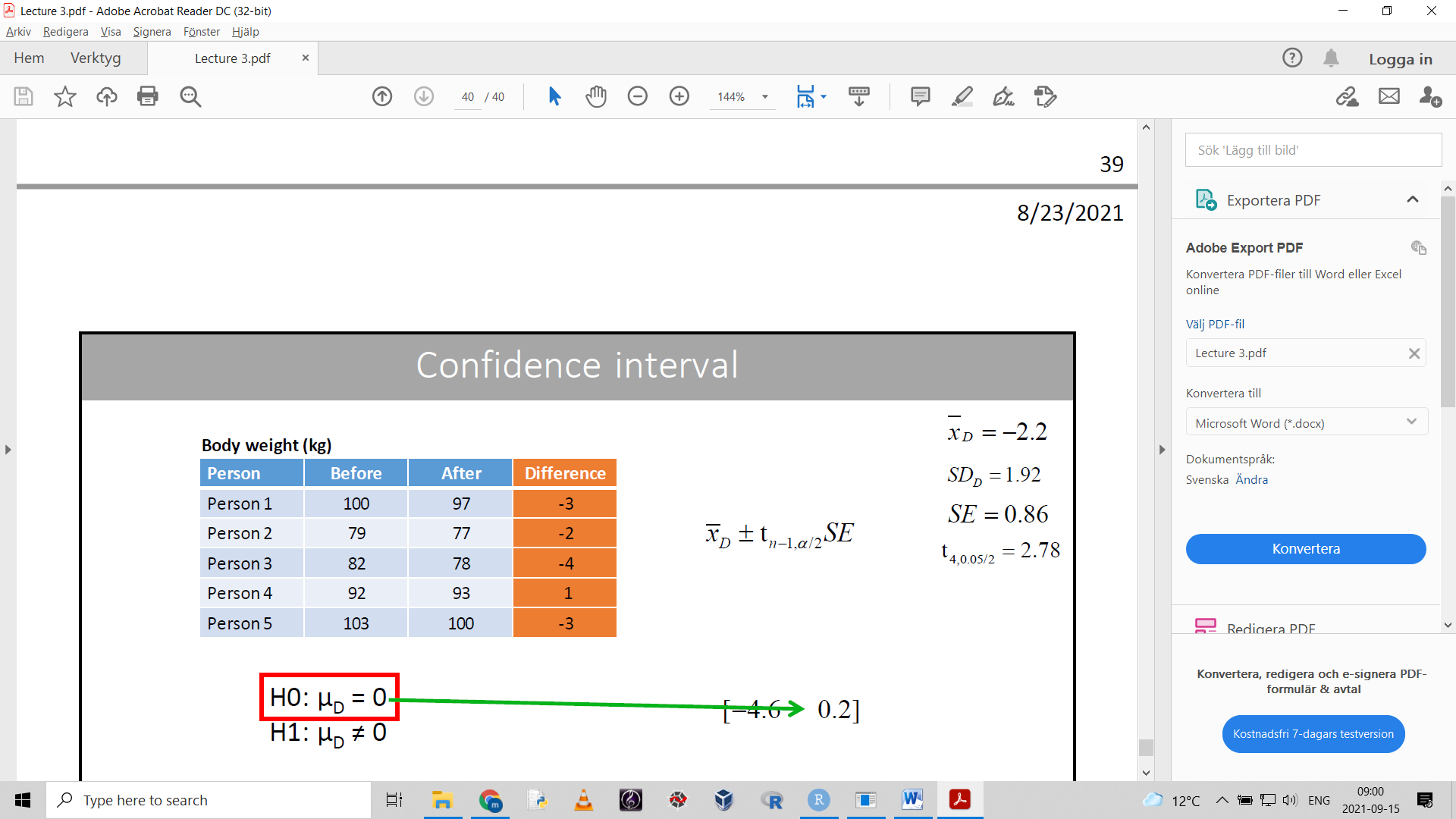


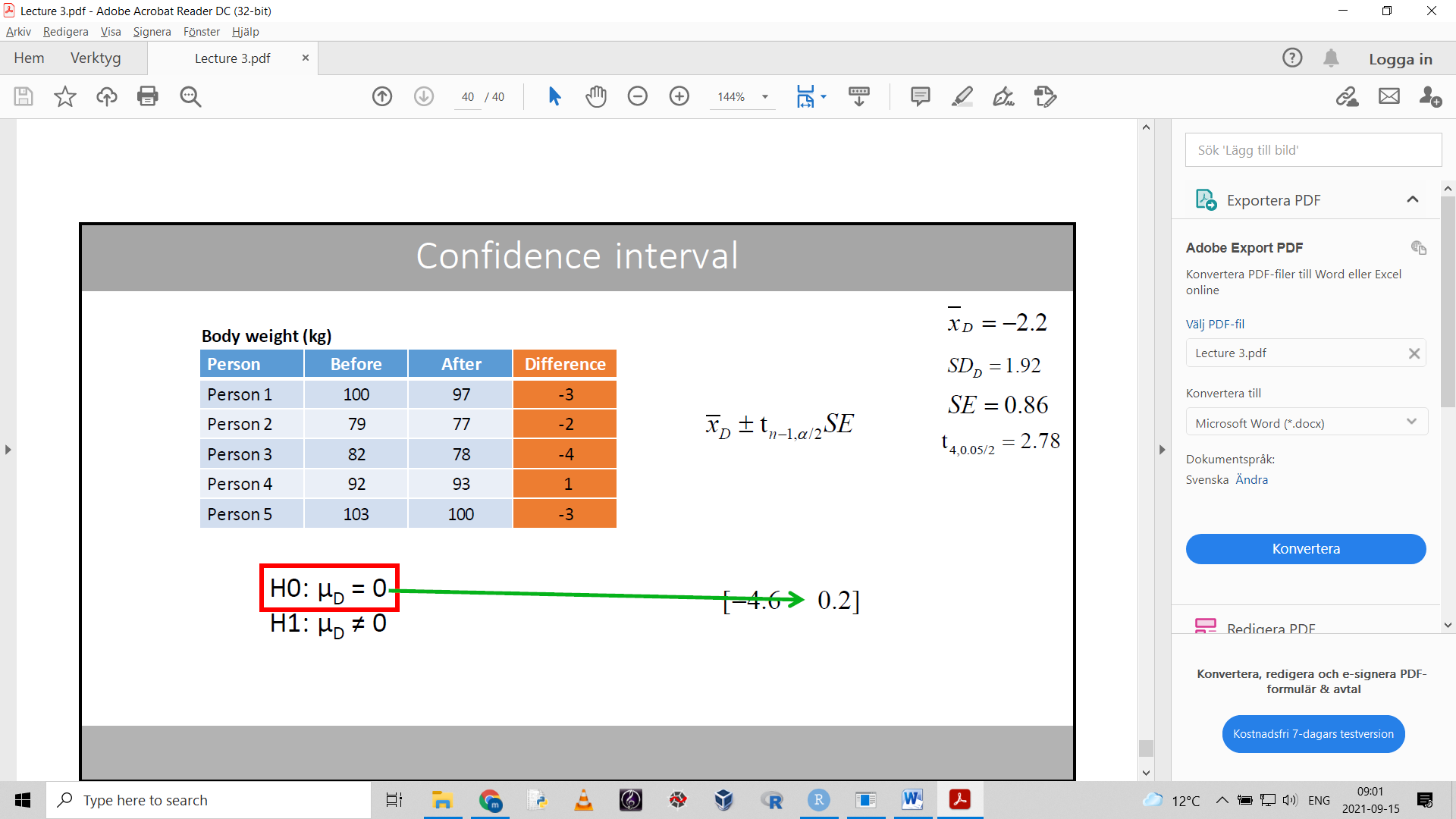
The t statistic of a paired t-test is the mean of the differences divided by the

SE=standard error of the mean differences

-> A p-value is obtained

Confidence interval:





If the H0 value is included in the confidence interval, that means its plausible.

Assumptions for the paired t test

- normal distribution (with small sample sizes)

If the differences are not normally distributed, we should use a non-parametric test.

The paired t-test is just a one-sample t-test based on the differences between the pairs.

# Lecture 4

## ANOVA

Analysis of variance

One-way means it involves only one factor

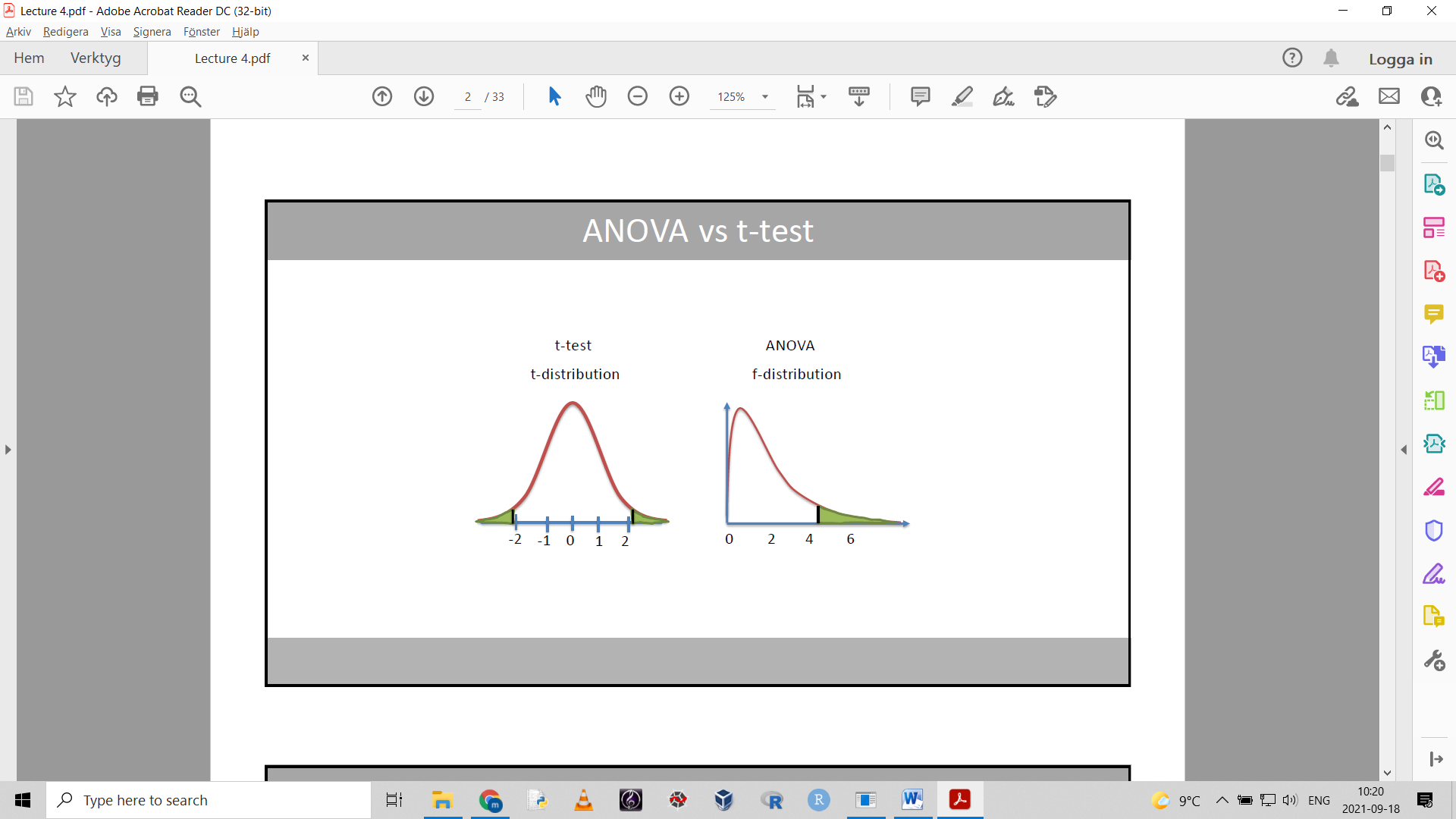
Anova can be used to compare the means of 3 or more groups.

- the independent variable (the factor) should have a categorical scale.

**Anova and t test**

(if you use anova to compare the means of two groups, you will get the same result as a t-test)

The f-statistic computed by anova is the square of the t-statistic: F=t2



ANOVA can only test a two sided hypotesis since the f-distribution only has positive values(t-test can test one or two sided).

If we use a t-test to make comparisons between three or more groups, we increase the chance of making a type 1 error = rejecting a true null hypothesis.

We have a 5% error chance for every t-test. These chances compound when we do many tests.

ADVANTAGES OF ANOVA WHEN COMPARING 3 OR MORE GROUPS:

- ANOVA compares all means in one go and maintains a single 0.05 risk of type 1 error

- a lot easier to run and interpret one test

**How anova works:**

H0: the population means are equal for all groups

H1: not all popolation means are equal

Example: H0: μ1 = μ2 = μ3

H1: at least one mean differ from the others

Assumptions for anova:

- the dependent variable should be normally distributed within each group:

The observations within each group should be normally distributed

- homogeneity of variance = the variance of each group should be the same (test by levenes test)

- observations are independent

**The F-ratio**

Anova calculates the f-ratio/f-statistic:

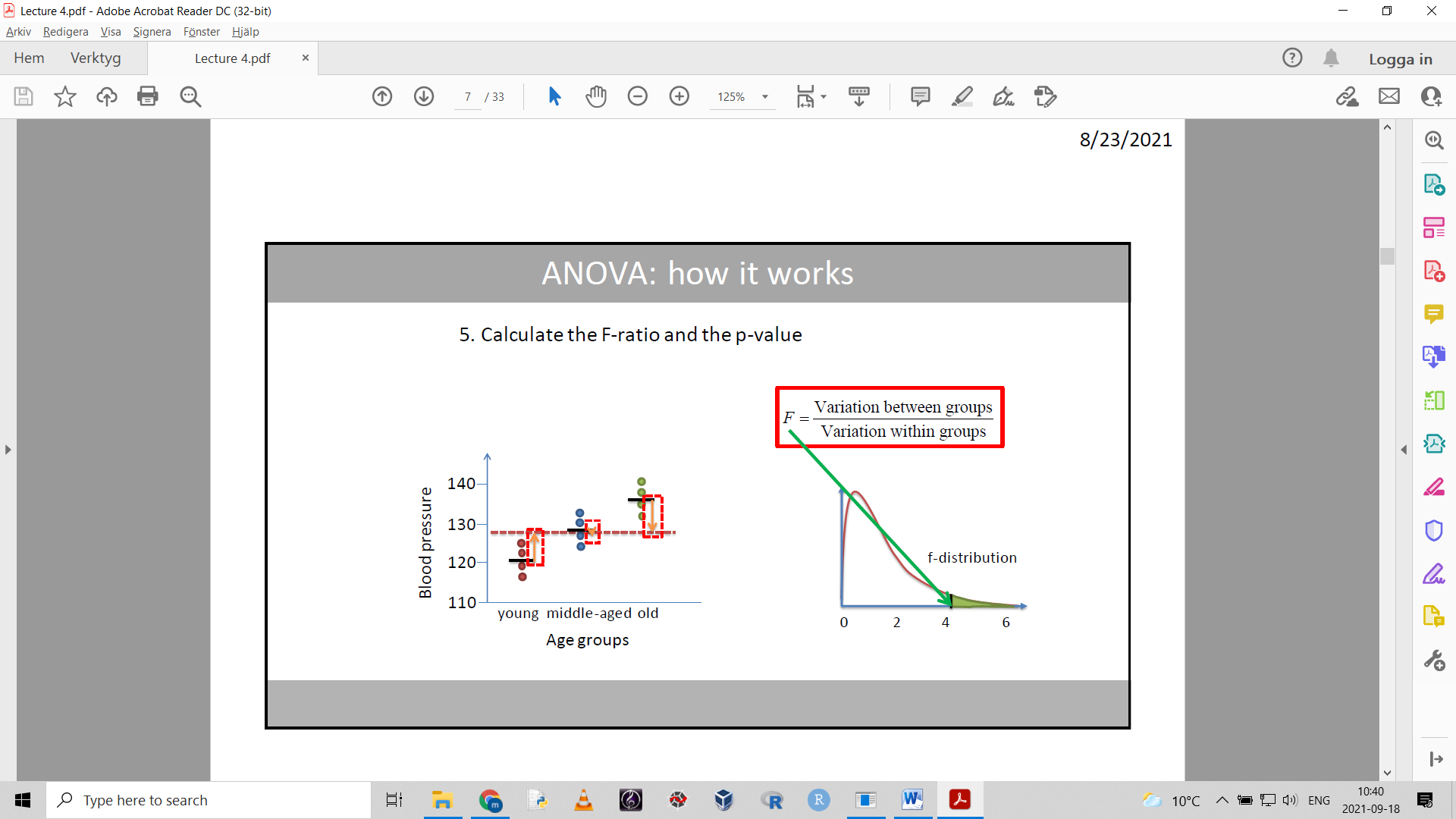
F =

1. calculate the means within the groups

2. calculate the grand mean = the mean of all observations

3. calculate the pooled variation within the groups (the distance between data points and their individual group mean)

4. estimate the variation between groups (distance between grand mean and group means)

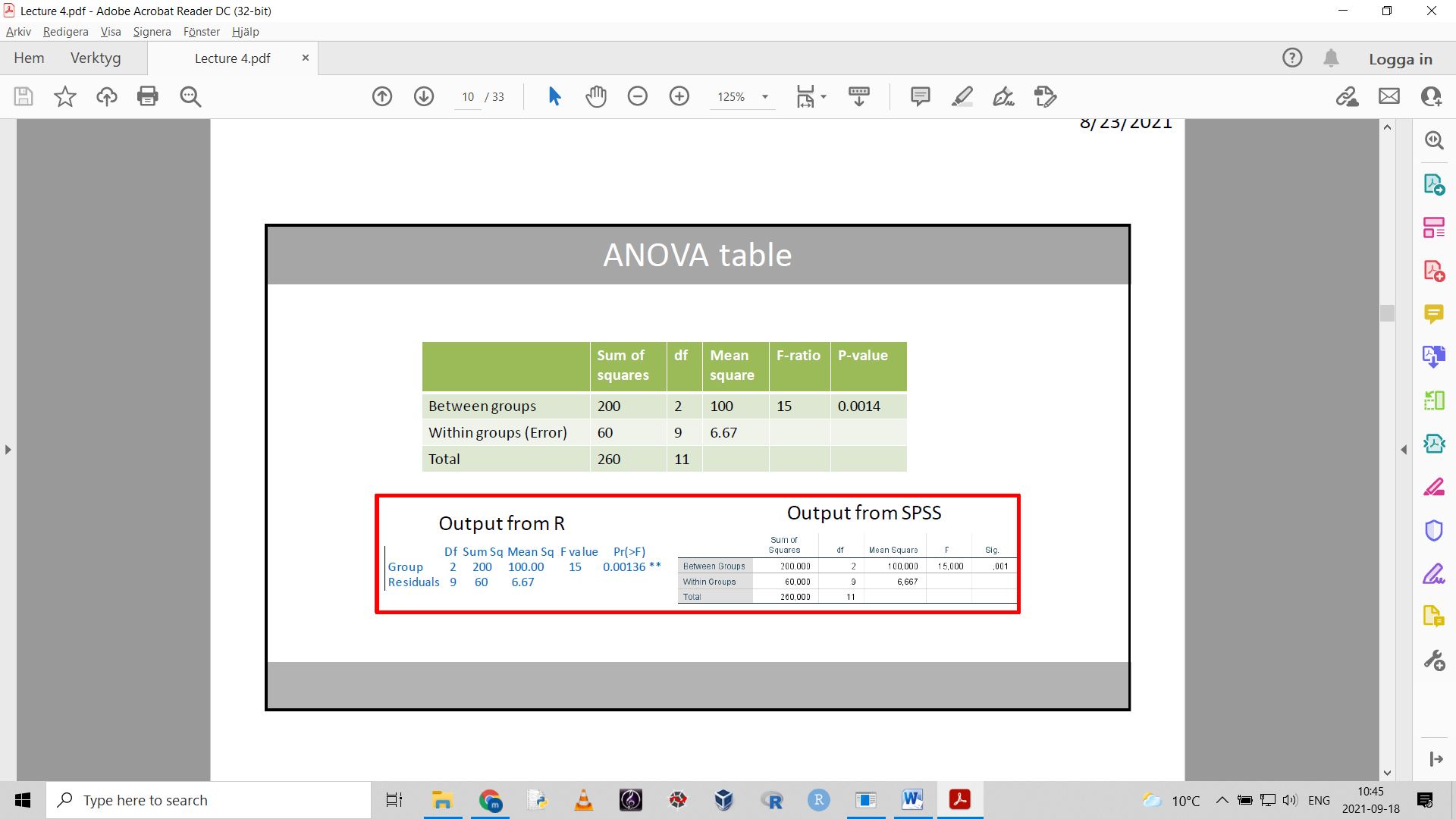


We will reject the H0 of anova only if variation between groups is larger than variation within groups.

If we reject the H0, it means there is a difference between at least one of the means.

- we can do a post-hoc test to see which one

Ex by doing 3 separate t-tests



## One way ANOVA calculations

The sum of squared (SS) distances:

The sum of squares within the groups: Error Sum of Squares or SSE.

This is done by summing the squares of the differences in all the goups to their respecitve mean.

You should end up with one number.

The sum of squares between groups: SSB

The sum of the sample size of each group multipiled by the squared difference between the corresponding group mean and the grand mean. (grand mean is subtracted from group means)

The total sum of squares: SST = SSB + SSE

SST is also the sum of the squared differences between each data point and the grand mean

The df between groups is the number of groups minus one or k-1

The df within groups is the total number of data points minus the number of groups or N-k

The df for the SST is the total number of data points minus one or N-1

Mean square = the sum of squares (SS) divided by corresponding df

MSB = mean square between groups

MSE = mean square of error (within groups)

F-ratio = MSB/MSE

The f-distribution to be used corresponds to the degrees of freedom of the MSB and MSE.

Use this f-distribution to find the p value.

**POST-HOC TESTS**

Fisher’s Least Significant Difference Test (LSD)

Basically a set of pairwise t-tests.

When we have more than 3 groups, we should use a post-hoc test that corrects for multiple comparisons.

Ex Bonferroni method or Tukey’s test

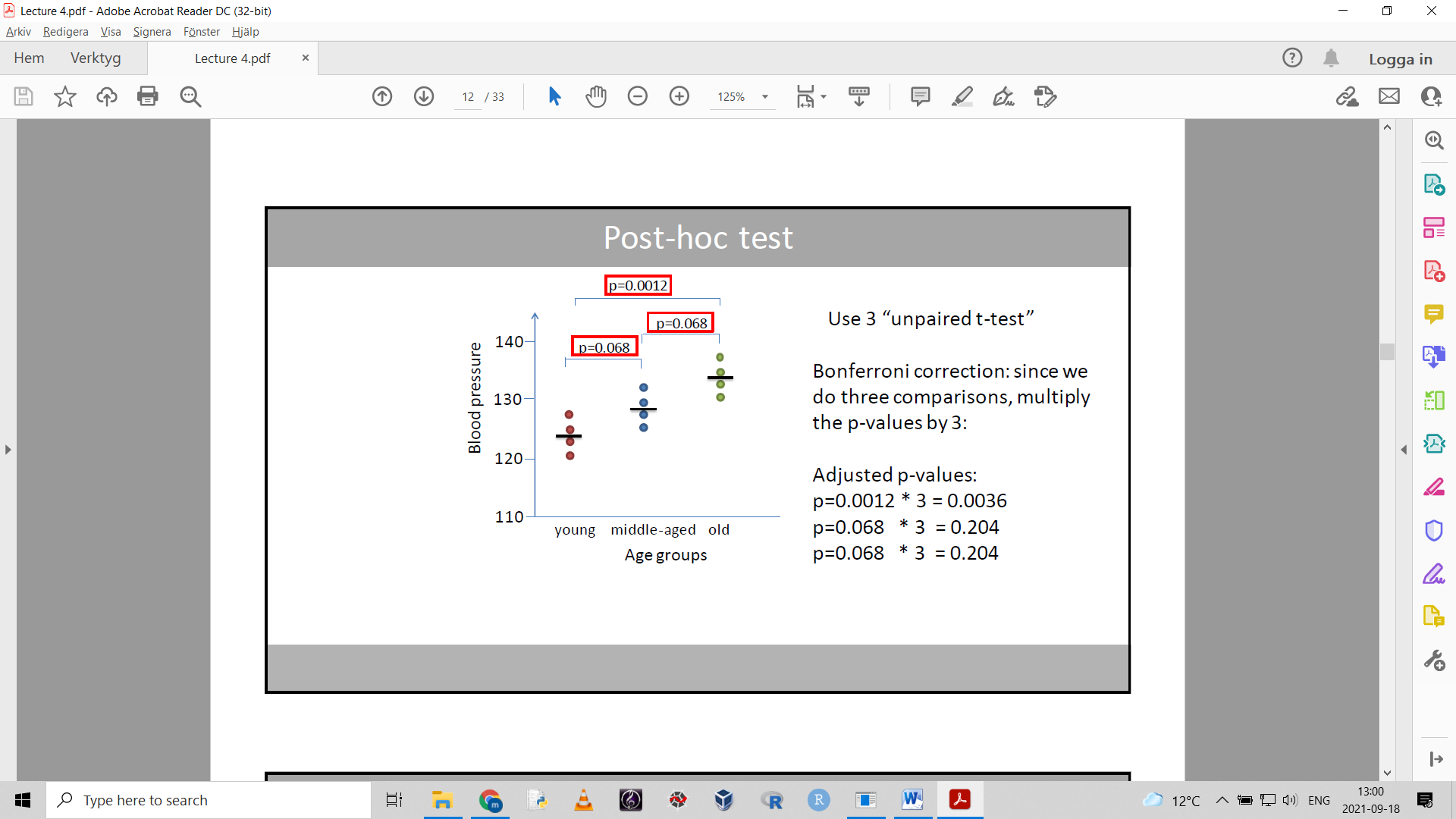
Post-hoc tests have been created to deal with the problems when we make several comparisons.

They are commonly applied after a significant ANOVA result.

Most post-hoc tests keep the family-wise error rate (FWER) at 0.05 even though we make several comparisons.

Bonferroni method is the simplest method to adjust for multiple comparisons.

Multiply the p-values by the number of tests we do.



## Repeated measures ANOVA

Can be seen as a version of the paired t-test when we have more than 2 measurements.

In repeated measures anova:

F =

The error variability is the variation within the groups, when the variation between subjects have been removed.

If we remove the difference between the subjects in a group from the sum of squares within the group, we’re left with only the variance from measurement to measurment.

The df of the variance in error is the df of the total within group variance minus the df for the subjects.

H0: all population means of the groups are equal

H1: not all means are equal



**Assumptions for repeated measures anova**

- Sphericity: the variances of the differences between the groups should be equal.

Can be tested with Mauchly’s test of sphericity

H0: sphericity

H1: no sphericity

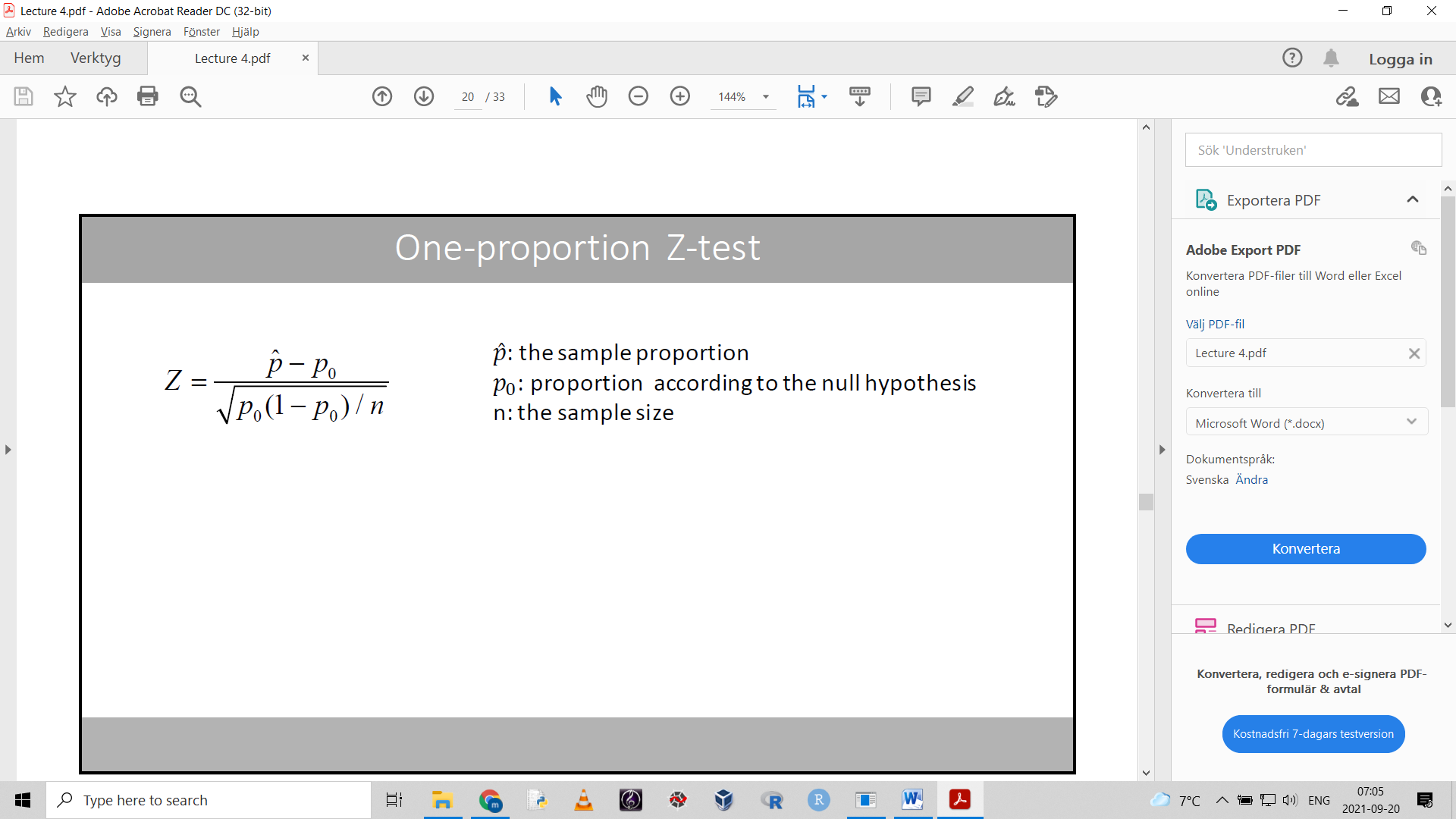
## One proportion Z-test

Used to analyze a proportion from just one sample.

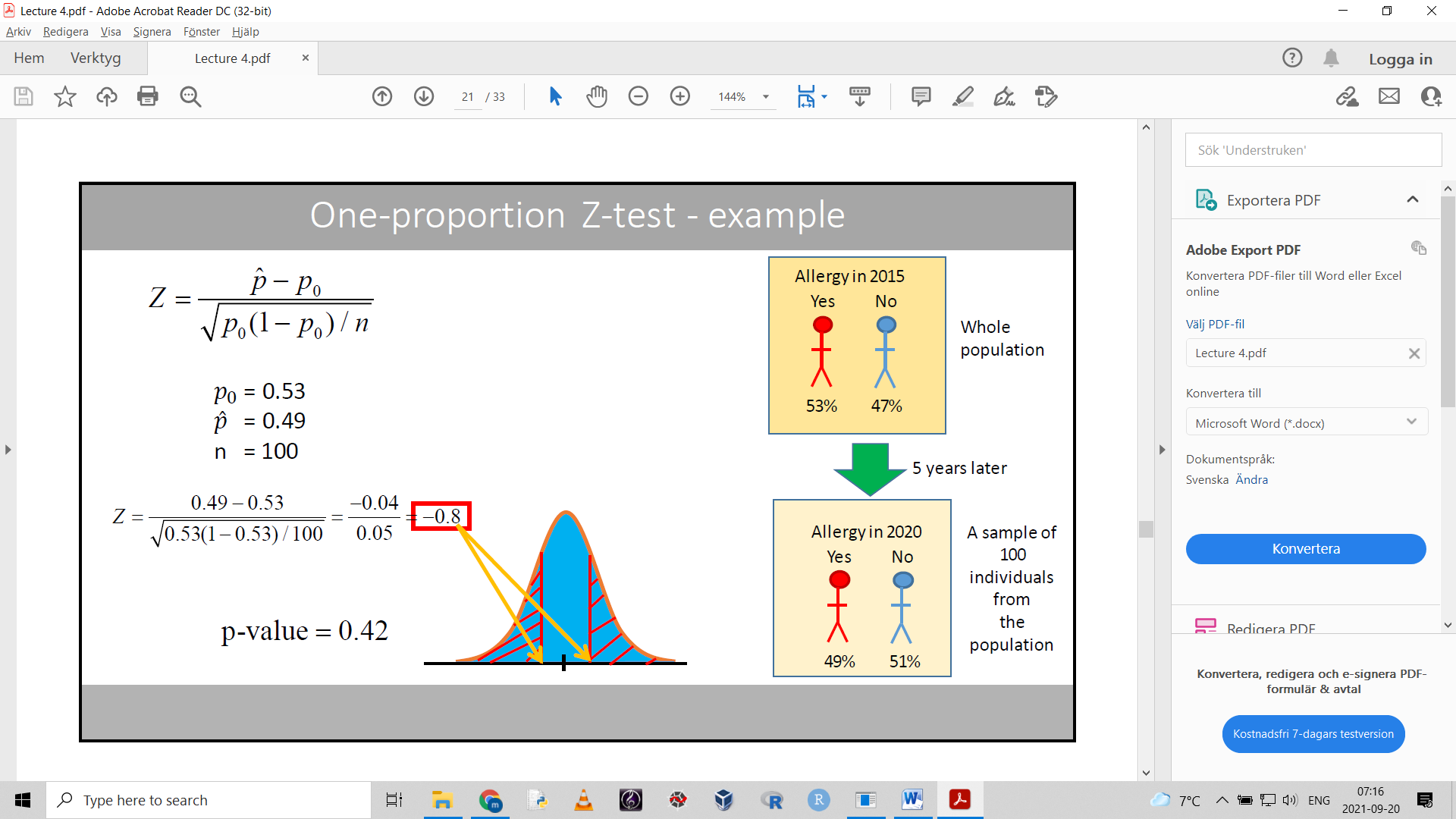
Chi-square goodness of fit can also be used.

If we want to analyze the difference between two sample proportions

Use a two-proportion Z-test or an appropriate chi-square test



The one-proportion Z-test tests if there is a differnece between a proportion from a sample and a proportion according to the null hypothesis.



**Assumptions for the one-proportion Z-test**

- independent individuals randomly selected from population

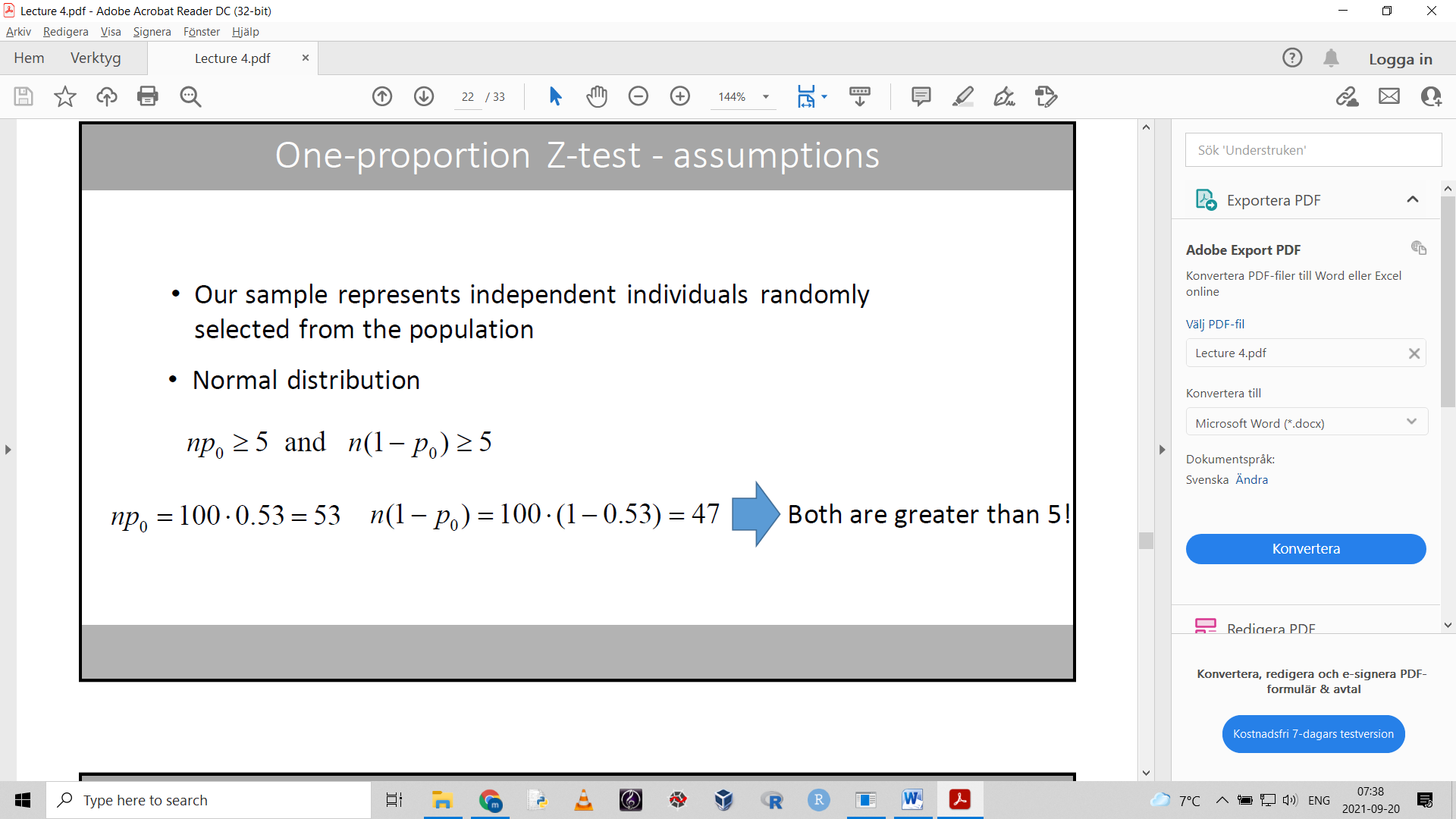
- normal distribution (the Z-statistic follows a normal distribution under the null hypothesis)

The H0 actually has a Bernoulli distribution (only two possible outcomes).

The central limit theorem plays in here.

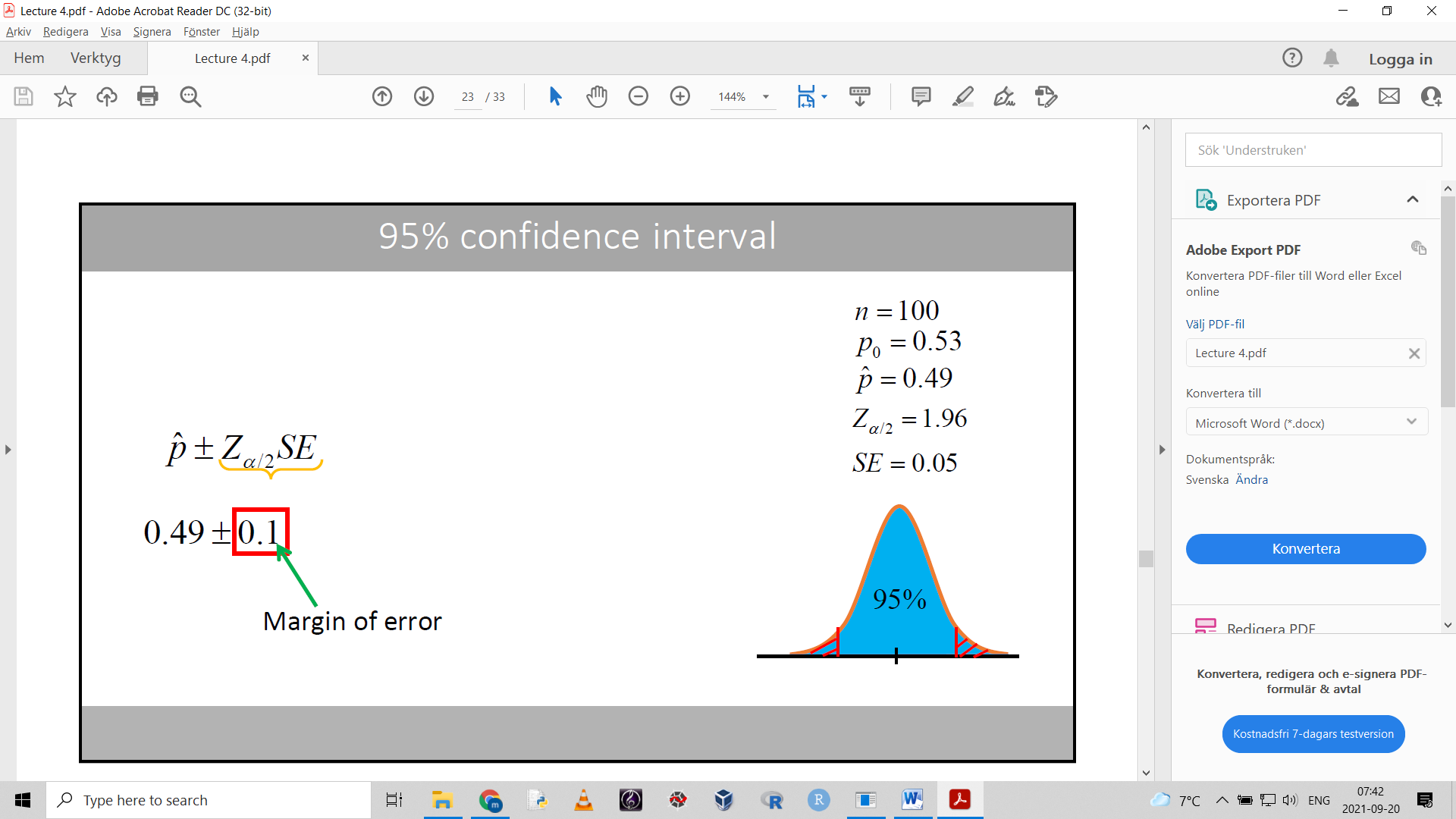
The more uneven the proportion is, the larger sample size you need to approach normal distribution of sample proportion means.

We use this to be able to assume that the sample proportions are normally distributed.



Meaning both proportions multiplied by sample size must be equal to or larger than 5.

Calculate the proportion confidence interval as such:



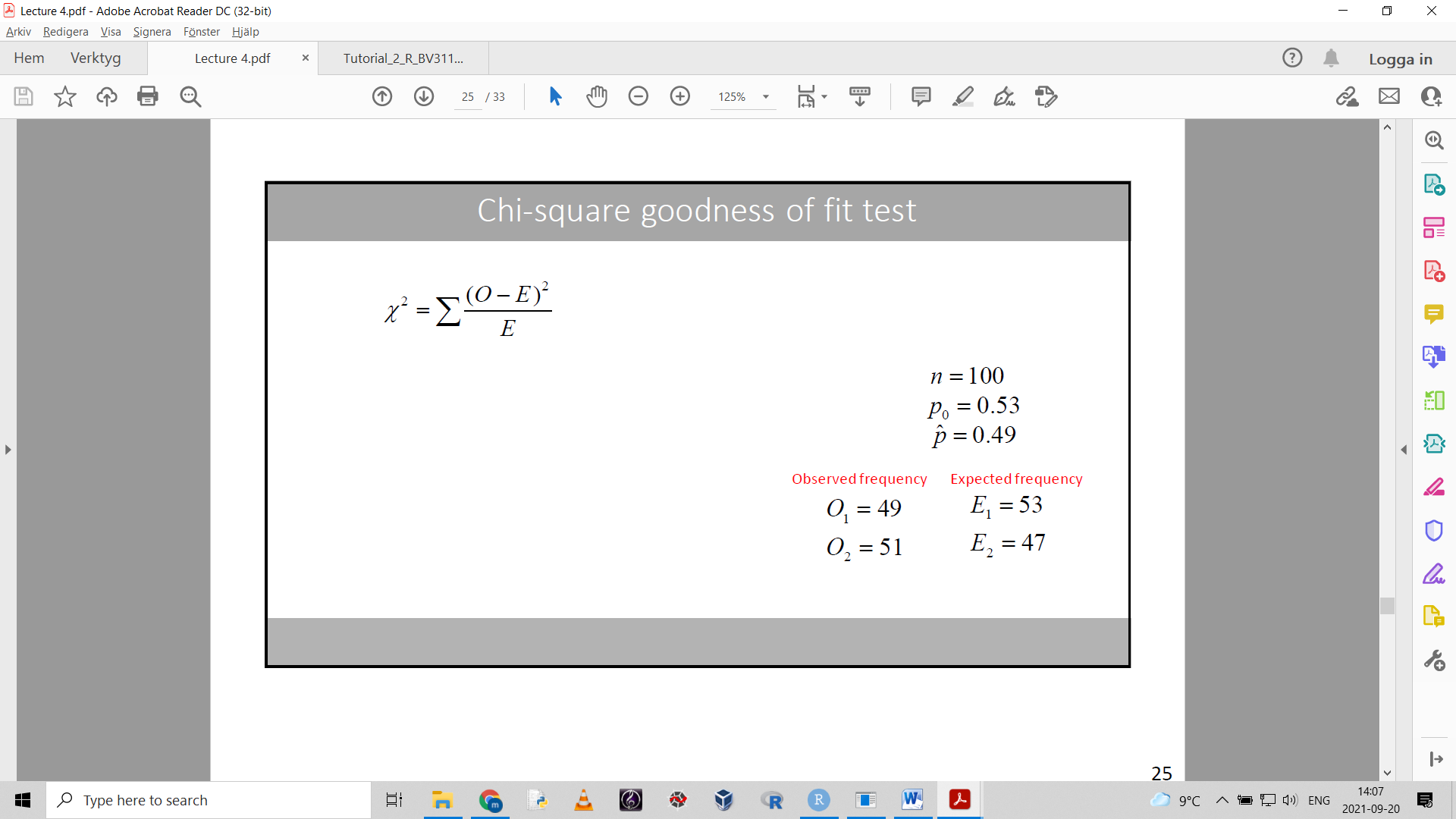
This is called the margin of error: represents how far from our estimate we think the true population proportion is.

## Chi square goodness of fit

If we analyze a proportion of one sample, we can use the one-sample proportion test or chi square goodness of fit.

There are different kinds of chi square tests.

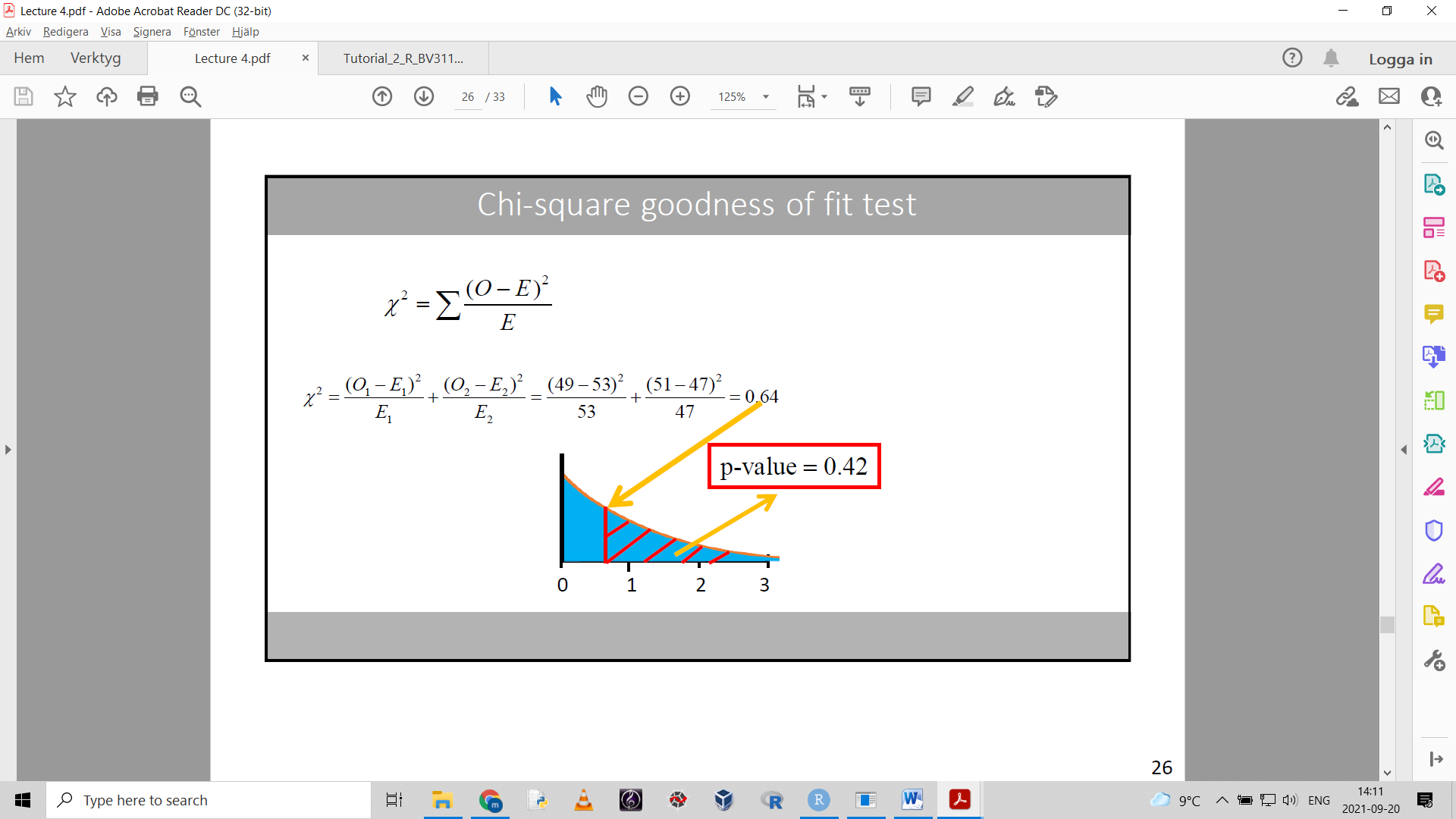
The so called goodness of fit test if the observed frequency distribution differs from a known/hypothesized distribution.



The goodness of fit test uses frequencies, not proportions.

Df = k-1 (k is the number of levels [the categorical variable])

We use a chi-square distribution corresponding to the df:



**Assumptions for the chi-square goodness of fit test**

- independent observations

- the expected frequencies in each category must be equal to or greater than 5.

**Differences between chi square goodness of fit and one proportion z-test**

- the Z statistic squared is the chi-square statistic

- the chi-square statistic is always positive, which means we can only do a one-tailed test.

A chi square goodness of fit test can be used when we have 3 or more categories.

(one proportion z-test cannot)

**Some different shi-square tests:**

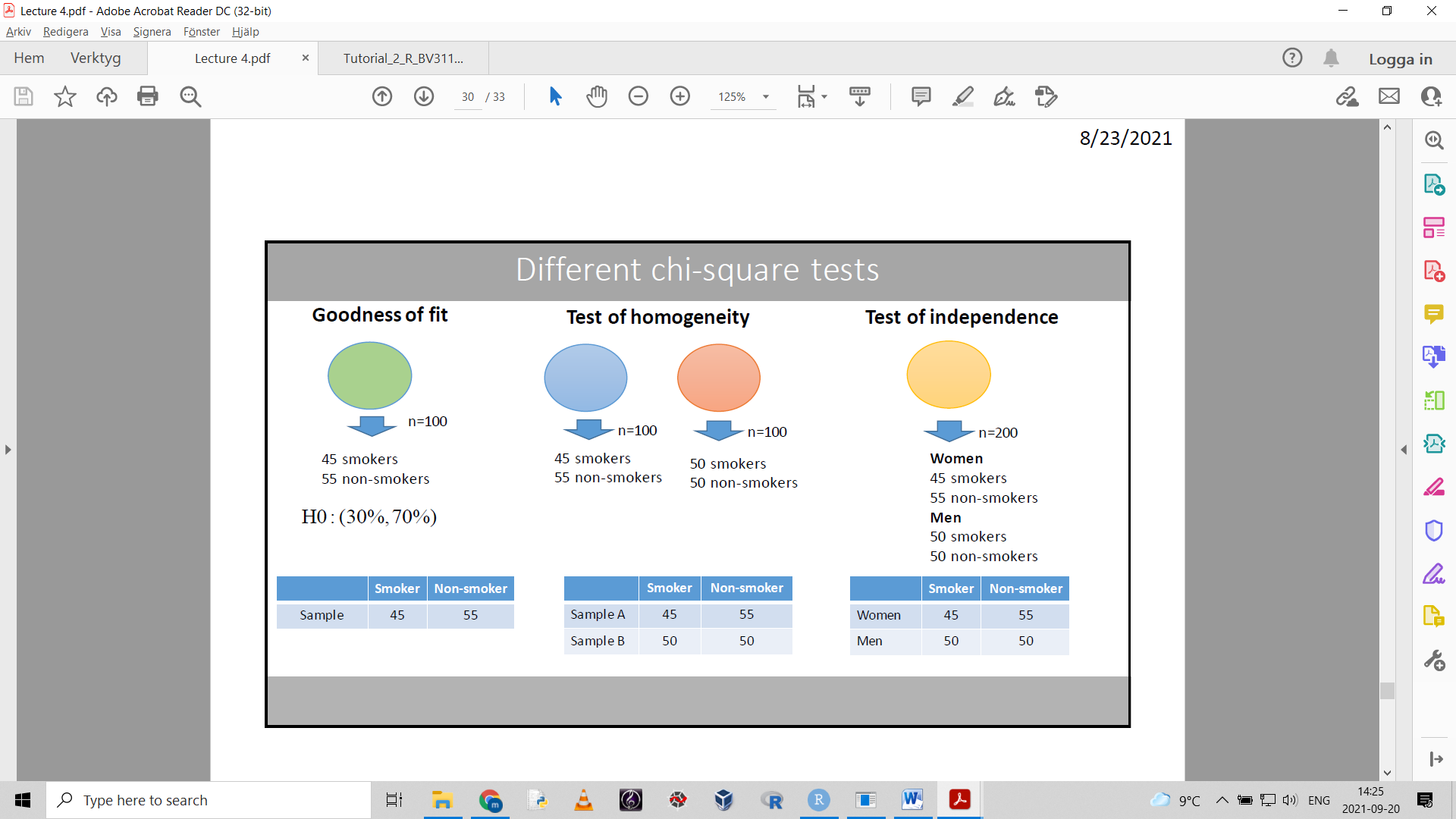
- goodness of fit

- test of homogeneity

- test of independence

## Tests for comparing two-sample proportions

A contingency table or crosstab:



Displays frequencies from two samples (from two different populations).

Can also display totals, percentages, sample sizes

**A two-proportion z-test can be used**

The hypotheses state:

H0: there is no difference in proportion between the populations

H1: there is a diffference

**Chi-square test of homogeneity**

Analyzes two sample proportions

The observed frequencies are compared with the expected frequencies from the null hypothesis . these can be calculated using some bothersom formula

**Assumption for both these tests:**

- all expected frequencies should be at least 5.

If one of the expected freqencies are less than 5, we can use an exact test such as Fishers exact test

The chi square test of homogeneity can analyze data from more than two groups. The two proportion Z-test cannot.

## Chi square test of independence

Tests if there is a relationship between two categorical variables on a nominal scale.

**The Pearson’s chi-square tests:**

- goodness of fit

- test of homogeneity

- test of independence

The goodness of fit is used to test if the observed freqency for a categorical variable is consistent with a known/hypothesized distribution.

The chi square test of homogeneity is similar, only that we have two or more samples we compare.

The test of independence tests two or more (categorical) variables from only one sample.

“Confounder = a variable that is associated with both the exposure and the outcome.”

## McNemar test

The corresponding chi-square for paired data.

- used on categorical data with only two categories, where a treatment is paired with a control group, or when we have matched pairs

We’ll see if we’ll ever get into this.

# Lecture 5