Estimates and Confidence intervals
Lecture: 8

Hamid Raza
Assistant Professor in Economics
raza@business.aau.dk

Aalborg University

Statistics - Statistik

Outline

Acceptance Interval:

- An acceptance interval is an interval within which a sample mean has a high probability of occurring, given that we know the population mean and variance
- If the sample mean is within that interval, then we can accept the conclusion that the random sample came from the population with the known population mean and variance.
- Assuming that we know the population mean μ and variance σ^2, then we can construct a symmetric acceptance interval:

$$\mu - z_{\alpha/2}\frac{\sigma}{\sqrt{n}}; \quad \mu + z_{\alpha/2}\frac{\sigma}{\sqrt{n}}$$

$\pm z_{\alpha/2}$ represents the critical value, and $1-\alpha$ represents the confidence level (CL)

The foundation of statistics

- Statistics is all about saying something about a population based on a sample
- E.g. we **guess** the population mean μ based on a sample
- We also say that μ is a population parameter
- Based on a specifict sample we can calculate a sample statistic such as $\bar{x}$ and s^2
- We use the sample statistic as an estimate of the population parameter, e.g., $\mu = \bar{x}$ is an estimate of of the population mean μ based on a specific sample
- When we think of the sample statistic as a random variable (something that changes every time we take a new sample) we call it an **estimator**

Properties of estimators:

- Unbiased
- Consistent

Point and interval estimates

- The estimators we have discussed so far are point estimates, meaning that it is just a single number that represents our guess of the unknown parameter value (e.g., mean and variance)
- In the world of statistics, point estimates come up with some uncertainty
- Therefore we use an **interval estimate**. This is an interval around the point estimate, in which we are confident (to a certain degree) that the population parameter is located

Confidence Interval and Confidence Level:

- A confidence interval estimator for a population parameter is a rule for determining (based on sample information) an interval, where we expect the parameter to be.
- The probability that this construction yields an interval which includes the parameter is called the **confidence level** and it is typically chosen to be 95%.
- (1-confidence level) is called the error probability (in this case $1 - 0.95 = 0.05, i.e., 5\%$). The error probability is also denoted α.

Intervals Based on the Normal Distribution

Confidence Interval for mean:

- When the population follows a normal distribution we know that:

$$\bar{X} \sim N(\mu, \sigma^2/n)$$

where $\bar{X}$ is transformed into a standard normal distribution:

$$Z = \frac{\bar{X} - \mu}{\sigma/\sqrt{n}} \sim N(0,1)$$

- We can look up in the normal distribution that:

$$1 - \alpha = P(-z_{\alpha/2} \leq Z \leq +z_{\alpha/2})$$

where $z_{\alpha/2}$ is the value from the standard normal distribution

- For a 95% confidence level it follows that:

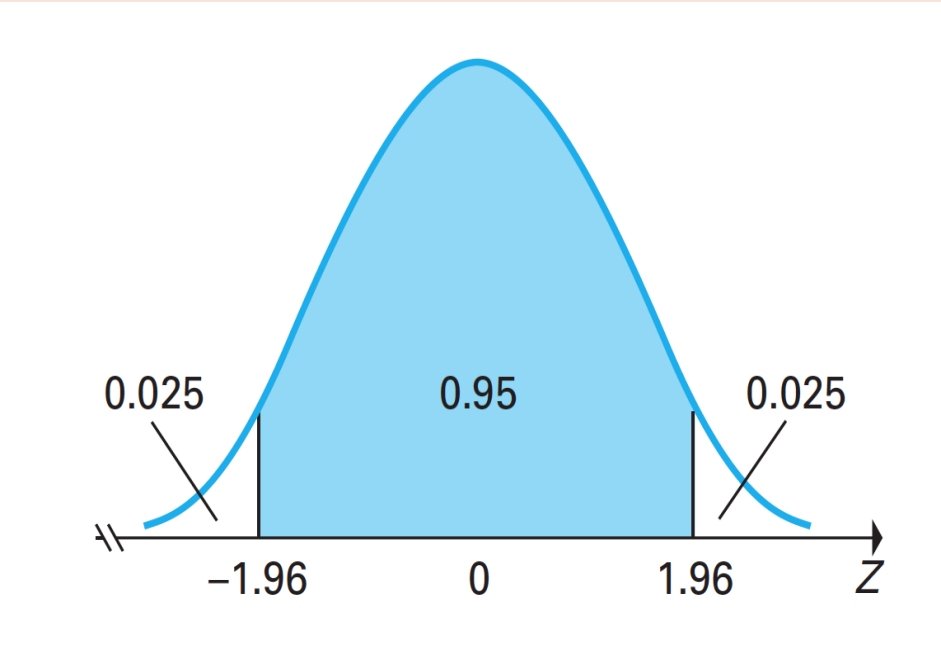
$$1 - \alpha = P(-1.96 \leq Z \leq +1.96) = 95\%$$

Inserting Z and rearranging gives:

$$P(\bar{X} - z_{\alpha/2}\sigma/\sqrt{n} \leq \mu \leq \bar{X} + z_{\alpha/2}\sigma/\sqrt{n})$$

Intervals Based on the Normal Distribution

Confidence Interval for mean:



$CL = 1 - \alpha$	α	$\alpha/2$	$z_{\alpha/2}$
90	0.1	0.05	1.645
95	0.05	0.025	1.96
99	0.01	0.005	2.576

```
qnorm(1-0.05, mean = 0, sd=1); qnorm(1-0.025, mean = 0, sd=1); qnorm(1-0.005, mean = 0, sd=1)
```

```
## [1] 1.645
## [1] 1.96
## [1] 2.576
```

Intervals Based on the Normal Distribution

Confidence Interval Estimation for the Mean of a Population That Is Normally Distributed: Population Variance Known

Consider a random sample of n observations from a normal distribution with mean μ and variance s^2. If the sample mean is $\bar{X}$, then a **confidence interval for the population mean with known variance** is given by:

$$\bar{X} \pm z_{\alpha/2}\frac{\sigma}{\sqrt{n}}$$

or

$$\bar{X} \pm ME$$

where ME is the margin of error

The upper confidence limit (UCL) is given by:

$$\bar{X} + z_{\alpha/2}\frac{\sigma}{\sqrt{n}}$$

The lower confidence limit (LCL) is given by:

$$\bar{X} - z_{\alpha/2}\frac{\sigma}{\sqrt{n}}$$

Intervals Based on the Normal Distribution

Confidence Interval Estimation for the Mean of a Population That Is Normally Distributed: Population Variance Known

Suppose that time spent on shopping at a local mall (Sailing) are normally distributed with known population standard deviation of 20 minutes. A random sample of 64 customers has a mean time of 75 minutes.

Find the standard error, margin of error, and the upper and lower confidence limits of a 95% confidence interval for the population mean, μ.

Solution: The standard error and the margin of error are as follows:

Standard error $= \sigma/\sqrt{n} = 20/\sqrt{64} = 2.5$

$ME = z_{\alpha/2}\frac{\sigma}{\sqrt{n}} = 1.96(2.5) = 4.9$

Upper confidence limit (UCL): $\bar{X} + ME = 75 + 4.9 = 79.9$

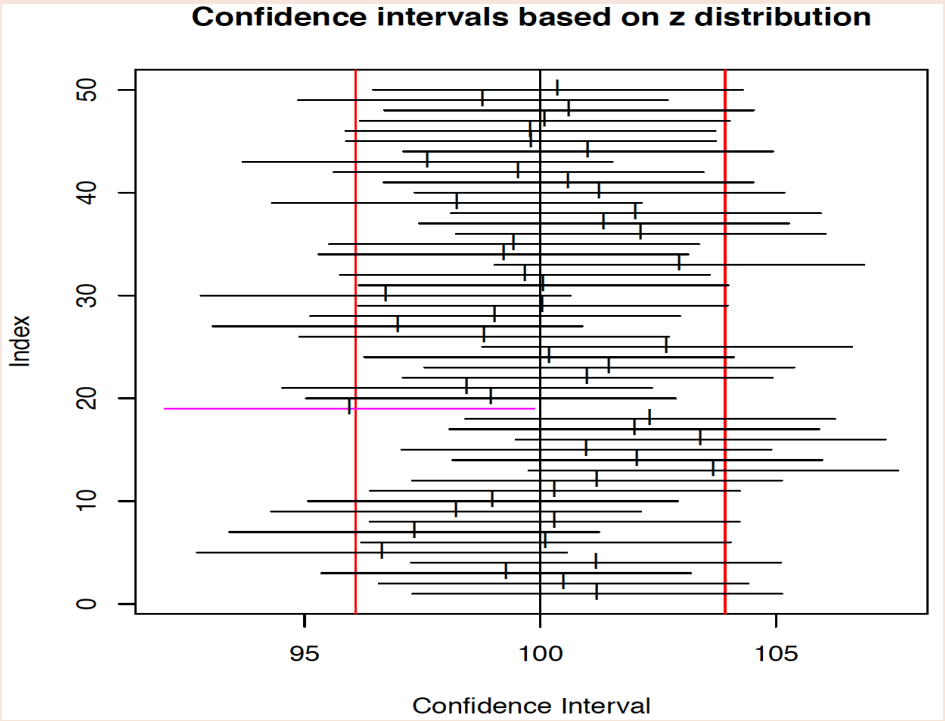
Lower confidence limit (LCL): $\bar{X} - ME = 75 - 4.9 = 70.1$

Intervals Based on the Normal Distribution

What does Confidence interval really mean?

- The confidence level of the interval implies that in the long run, 95% of intervals found will contain the true value of the population mean
- In other words, if we repeat the process for a large number of independent random samples, 95% of the times the confidence intervals will contain the true value of the population mean.
- **Note:** in a single sample case, it is not known whether our interval is one of the good 95% or bad 5% without knowing μ
- This idea can be graphically illustrated

Intervals Based on the Normal Distribution



Fifty (50) samples of size twenty five (25) were generated from a norm(mean = 100; sd = 10) distribution, and each sample was used to find a 95% confidence interval for the population mean. The 50 confidence intervals are represented above by horizontal lines, and the respective sample means are denoted by vertical slashes. Confidence intervals that "cover" the true mean value of 100 are plotted in black; those that fail to cover are plotted in a lighter color. In the plot we see that only one (1) of the simulated intervals out of the 50 failed to cover $\mu = 100$, which is a success rate of 98%. If the number of generated samples were to increase from 50 to 500 to 50000, . . . , then we would expect our success rate to approach the exact value of 95%.

Outline

Student's t Distribution

In the preceding section confidence intervals for the mean of a normal population when the population variance was known were derived

But a major problem is that population variance is not known so we cannot use the formula

In the case where the population standard deviation is unknown, the result ($Z = \frac{\bar{X}-\mu}{\sigma/\sqrt{n}}$) cannot be used directly

Student's t Distribution

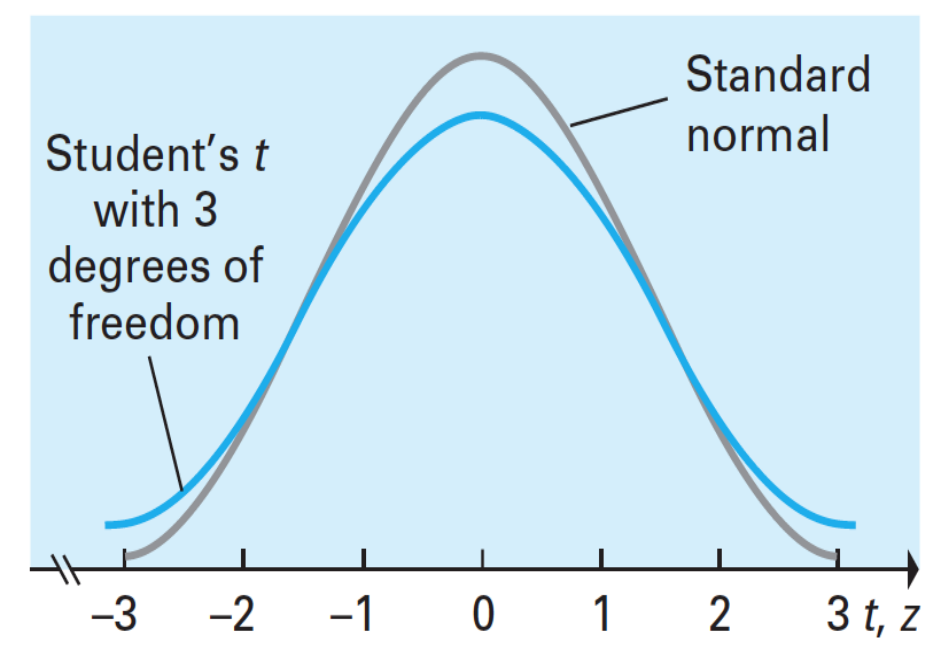
In this case, σ is replaced by the sample standard deviation (s):

$$t = \frac{\bar{X} - \mu}{s/\sqrt{n}}$$

This random variable follows a member of a family of distributions called **Student's t**

Student's t Distribution

Student's t Distribution (cont)



- The shape of the Student's t distribution is similar to that of the standard normal distribution
- Both distributions have mean 0, and the probability density functions of both are symmetric about their means
- The density function of the Student's t distribution has a wider dispersion (reflected in a larger variance) than the standard normal distribution
- For large degrees of freedom, the two distributions are virtually identical, i.e, student's t dist. converges to $N(1, 0)$

Intervals Based on Student's t distribution

Confidence Interval for mean:

- We now consider:

$$t = \frac{\bar{X} - \mu}{s/\sqrt{n}} \sim N(0,1)$$

- Like before, we can write it as:

$$(1-\alpha) = P(\bar{X} - t_{v,\alpha/2} \leq t \leq \bar{X} + t_{v,\alpha/2})$$

- Insert the value of 't' and re-arranging gives:

$$P(\bar{X} - t_{v,\alpha/2}s/\sqrt{n} \leq \mu \leq \bar{X} + t_{v,\alpha/2}s/\sqrt{n})$$

v represents the degrees of freedom, e.g., for 95% confidence interval with 10 degrees of freedom:

```
qt(1-0.025, df = 10)
```

```
## [1] 2.228
```

Intervals Based on the Student's t Distribution

Confidence Intervals for the Mean

Intervals Based on the Student's t Distribution

Confidence Intervals for the Mean

If the sample mean and standard deviation are, respectively, $\bar{X}$ and s, then the degrees of freedom is $v = n - 1$. The confidence limit is given by:

Intervals Based on the Student's t Distribution

Confidence Intervals for the Mean

If the sample mean and standard deviation are, respectively, $\bar{X}$ and s, then the degrees of freedom is $v = n - 1$. The confidence limit is given by:

$$\bar{X} \pm t_{v,\alpha/2} \frac{s}{\sqrt{n}}$$

Intervals Based on the Student's t Distribution

Confidence Intervals for the Mean

If the sample mean and standard deviation are, respectively, $\bar{X}$ and s, then the degrees of freedom is $v = n - 1$. The confidence limit is given by:

$$\bar{X} \pm t_{v,\alpha/2} \frac{s}{\sqrt{n}}$$

or

$$\bar{X} \pm ME$$

Intervals Based on the Student's t Distribution

Confidence Intervals for the Mean

If the sample mean and standard deviation are, respectively, $\bar{X}$ and s, then the degrees of freedom is $v = n - 1$. The confidence limit is given by:

$$\bar{X} \pm t_{v,\alpha/2} \frac{s}{\sqrt{n}}$$

or

$$\bar{X} \pm ME$$

where ME is the margin of error

Intervals Based on the Student's t Distribution

Confidence Intervals for the Mean

If the sample mean and standard deviation are, respectively, $\bar{X}$ and s, then the degrees of freedom is $v = n - 1$. The confidence limit is given by:

$$\bar{X} \pm t_{v,\alpha/2} \frac{s}{\sqrt{n}}$$

or

$$\bar{X} \pm ME$$

where ME is the margin of error

Intervals Based on the Student's t Distribution

Exercise in R:

The dataset (ryder.csv) contains information on individuals. Calculate the confidence interval of the mean of height of individuals in this dataset.

```
data = read.csv("~/Dropbox/Teaching/Statistics/Lecture 8/data/ryder.csv", header = T)
x = as.numeric(data$hoejde) # we tell R this is a numerical data
x = na.omit(x) # We leave out individuals for whom data is missing
```

```
n <- length(x) # calculates the no. of observations (n)
m <- mean(x) # calculates the sample mean
s <- sd(x) # calculates the sample standard deviation
se <- s/sqrt(n) # calculates the standard error
tscore <- qt(1-.025, df = n-1) # calculates the t score
ME <- tscore*se  # calculates the margin of error (ME)
m + c(-ME,ME) # Calculates the lower and upper confidence limits of the mean of x


## [1] 172.3 173.1
```

R can directly calculate this for us:

```
t.test(x)
```

```
##
##  One Sample t-test
##
## data:  x
## t = 919, df = 2626, p-value <2e-16
## alternative hypothesis: true mean is not equal to 0
## 95 percent confidence interval:
##  172.3 173.1
## sample estimates:
```

Outline

Confidence Intervals for Population Proportion

Confidence Intervals for Population Proportion:

We consider a population, where the distribution of a given characteristic is P to have this characteristic and $1 - P$ not to have it (e.g., male or female)

Let $\hat{p}$ denote the observed proportion of "successes" in a random sample of n observations from a population with a proportion of successes P

- The standard error of $\hat{p}$: $\sigma_{\hat{p}} = \sqrt{\frac{P(1-P)}{n}}$

We do not know P but we insert the estimate and get the

- **estimated standard error** of $\hat{p}$: $\sigma_{\hat{p}} = \sqrt{\frac{\hat{p}(1-\hat{p})}{n}}$
- Like before, we have:

$$\hat{p} \pm z_{\alpha/2}\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}$$

$$\hat{p} \pm ME$$

Confidence Intervals for Population Proportion

Exercise in R:

The dataset (ryder.csv) contains information on individuals. Calculate the confidence interval of the population proportions (for females in this case).

```
data = read.csv("~/Dropbox/Teaching/Statistics/Lecture 8/data/ryder.csv", header = T)
```

```
y = table(data$koen)
y    # we tell R this is categorical variable
```

```
##
## Kvinde   Mand
##   1471   1225
```

```
y/sum(y)          # check the proportion of males and females in the data
```

```
##
## Kvinde   Mand
## 0.5456 0.4544
```

Population proportion of females (F) : p

$$\hat{p} = \frac{Females}{Females + Males} = \frac{1471}{1471 + 1225} = 0.54$$

Confidence Intervals for Population Proportion

Exercise in R (cont):

R automatically calculates the confidence interval for the proportion of females when we do a so-called hypothesis test (discussed later):

```
prop.test(y, correct = FALSE)
```

```
##
##  1-sample proportions test without continuity correction
##
## data:  y, null probability 0.5
## X-squared = 22, df = 1, p-value = 2e-06
## alternative hypothesis: true p is not equal to 0.5
## 95 percent confidence interval:
##  0.5268 0.5643
## sample estimates:
##      p
## 0.5456
```

Outline

Confidence interval for variance

Confidence interval for variance:

Suppose a random sample of n observations from a normally distributed population with variance σ^2 and sample variance s^2 is taken. The random variable follows follows a chi-square distribution with $(n-1)$ degrees of freedom:

$$\chi^2_{n-1} = \frac{(n-1)}{\sigma^2} s^2$$

To find confidence intervals for the population variance:

$$1 - \alpha = P(\chi^2_{n-1,1-\alpha/2} < \chi^2_{n-1} < \chi^2_{n-1,\alpha/2})$$

Insert the value of χ^2_{n-1} and re-arranging gives:

$$\frac{(n-1)}{\chi^2_{n-1,\alpha/2}} s^2 < \sigma^2 < \frac{(n-1)}{\chi^2_{n-1,1-\alpha/2}} s^2$$

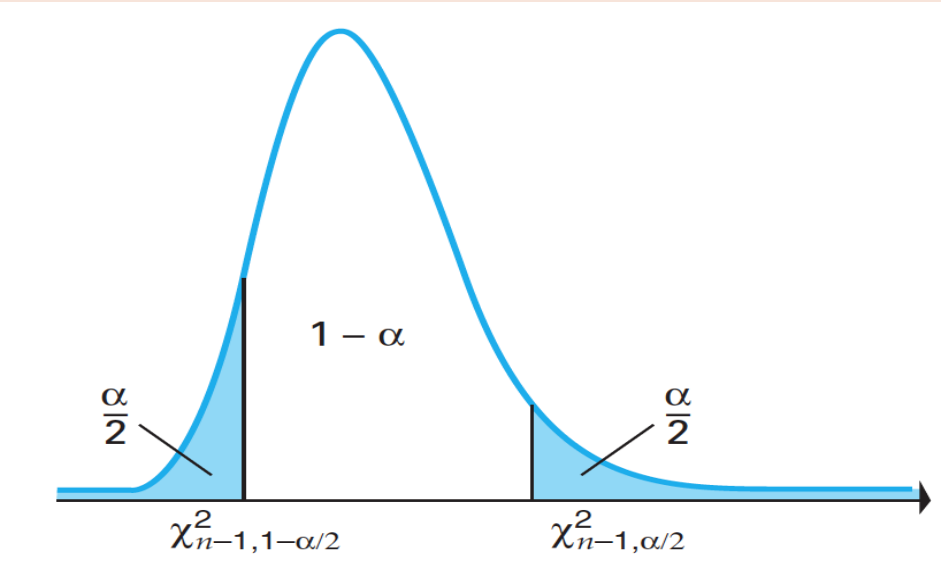
LCL is $\frac{(n-1)}{\chi^2_{n-1,\alpha/2}} s^2$

UCL is $\frac{(n-1)}{\chi^2_{n-1,1-\alpha/2}} s^2$

Note: the general assumption is that the parent pop is normally dist.

Confidence interval for variance

Confidence interval for variance:



Confidence interval for variance

Confidence interval for variance: Example

Assume we have a sample of 20 car dealers in Denmark, and we measure the number of weekly car sales. The standard deviation of the car sales is 3. Calculate the confidence interval of variance at 90%, 95%, and 99%?

- **SOLUTION:** We have the formula: $\frac{(n-1)}{\chi^2_{n-1,\alpha/2}} s^2 < \sigma^2 < \frac{(n-1)}{\chi^2_{n-1,1-\alpha/2}} s^2$
- We have all the values but we need to check the value (in blue) from chi-square dist. table (here I will use R for only 95% confidence interval:)
- First, I calculate the (critical) value for lower tail $\chi^2_{n-1,\alpha/2}$ at 95% with 19 degrees of freedom $(n-1)$.

```
qchisq(0.025, df=19)
```

```
## [1] 8.907
```

- Now, I calculate the (critical) value for upper tail $\chi^2_{n-1,1-\alpha/2}$ at 95% with 19 degrees of freedom $(n-1)$.

```
qchisq(1-0.025, df=19)
```

```
## [1] 32.85
```

Confidence interval for variance

Confidence interval for variance: Example (cont)

- **SOLUTION:** We can now calcualte the 95% confidence intervals for variance:
- Calculate the lower confidence limit (LCL)

```
n=20 # size of the sample
s2=9 # variance of the sample
(n-1)*s2/32.85


## [1] 5.205
```

- Calculate the upper confidence limit (UCL)

```
(n-1)*s2/8.907


## [1] 19.2
```

The variance 9 in this example has a 95% confidence intervals of 5.205 and 19.2

Outline

Sample size determination

Sample Size for the Mean of a Normally Distributed Population with Known Population Variance

When the population variance σ^2 was known, we earlier had:

$$\bar{X} \pm z_{\alpha/2}\frac{\sigma}{\sqrt{n}}$$

where

$$ME = z_{\alpha/2}\frac{\sigma}{\sqrt{n}} \tag{1}$$

Suppose, we want to fix the margin of error, ME (say ME=5), in advance. And we want a $100(1-\alpha)\%$ confidence interval for μ.

The question is: how big does n have to be?

- From basic algebra (re-arrangin equation 1) it follows that:

$$n = \frac{z_{\alpha/2}^2\sigma^2}{ME^2}$$

Sample size determination

Sample Size for Population Proportion

Earlier we saw that for a random sample of n observations, a $100(1-\alpha)\%$ confidence interval for the population proportion P is:

$$\hat{p} \pm z_{\alpha/2}\sqrt{\frac{\hat{p}(1-\hat{p})}{\hat{p}}}$$

where

$$ME = z_{\alpha/2}\sqrt{\frac{\hat{p}(1-\hat{p})}{\hat{p}}} \tag{2}$$

- This result cannot be used directly to determine the sample size necessary to obtain a confidence interval of some specific width since it involves the sample proportion

Sample size determination

Sample Size for Population Proportion (cont)

- Whatever the outcome, $\hat{p}(1-\hat{p})$ cannot be bigger than 0.25 (i.e, when the sample proportion is 0.5)
- Thus, the **largest possible value for the margin of error, ME**, is given by the following:

$$n = \frac{0.25(z_{\alpha/2})^2}{(ME)^2}$$

Outline

Exercise

1. Discussion question: In Chi-square case, you noticed that we calculated the upper tail and lower tail values. Why did we not do that in the normal dist and student's t distribution
2. Assume a sample of size (n=29) with mean 20 and variance of the population is known to be 500? Calculate the 95% confidence interval of mean
3. Assume we do not have the population variance, and the sample variance is 450. Calculate the 95% confidence interval of mean.
4. Calculate the 99% confidence interval of the variance
5. Calculate the 90% confidence interval of the variance of height in the example
6. Plot the distribution of heights. What kind of distribution does it look like?