Hypothesis Tests
Introduction

Peter Büchel

HSLU W

SA: W05

Hypothesis Test

- Hypothesis testing: Important statistical tool to decide whether observations "fits" a certain parameter
- Assumption: True mean *not* known, but assume an "ideal" or an assumed value

Example: Bottling Plant

- A brewery orders a new bottling plant for 500 ml cans
- Bottling plant: *Never* fills *exactly* 500 ml, but only *approximately*
- Brewery: Interested that bottling plant fills as accurately as possible
- If bottling plant fills too much: Bad for brewery as it sells too much beer for given price
- Does not fill enough: Customers are dissatisfied, because they do not get enough beer for the price

- Manufacturer claims that bottling plant fills cans normally distributed with $\mu = 500$ ml and $\sigma = 1$ ml
- Brewery takes 100 samples
- Mean of these samples: 499.22 ml
- Less than 500 ml
- However: Still in range of accuracy $\mu = 500$ ml and $\sigma = 1$ ml of manufacturer of bottling plant?
- How can to check this?

Example

- Newspaper claims: Average height of adult women in Switzerland is 180 cm with a standard deviation of 10 cm
- Mean intuitively wrong, because it is much too high
- But how to check and justify this mathematically without having to rely on intuition?

Hypothesis Tests

- Aim: To introduce a standardised, reproducible procedure to decide whether mean of observations does (or not) match a certain *true* mean μ
- Following procedure: *Never* proof that, for example, a quantity does not fit observations
- By statistical means: Can only show that this quantity does not fit to the observations *with high probability*
- Newspaper: "...proven with statistics...", this is nonsense!
- Procedure: Explained with examples above

Example: Bottling Plant

- Take 10 beer cans and measure their contents:

$$498.45, \quad 500.23, \quad \ldots, \quad 499.11$$

- Assumption: Content normally distributed with $\mu_0 = 500$ and $\sigma = 1$
- Measurement data $x_1, x_2, \ldots, x_{10}$: Realisations of random variables

$$X_1, \quad X_2, \quad \ldots, \quad X_{10}$$

- Example: $x_2 = 500.23$

In General

- Measurement data $x_1, \ldots, x_n$ as realisations of:

$$X_1, \ldots, X_n \text{ i.i.d.} \sim \mathcal{N}(\mu, \sigma_X^2)$$

- Two key figures for *all* random variables X_i are (2. i in i.i.d.):

$$\mathrm{E}(X_i) = \mu \qquad \text{and} \qquad \mathrm{Var}(X_i) = \sigma_X^2$$

- Normally: Key figures *unknown*
- *Why* unknown?
  - Example: Beer can
  - To know true μ and σ_X: Would have to measure contents and standard deviation of *all* cans
  - All those already produced and all those not yet produced
  - This is impossible!

Example: Bottling Plant

- Manufacturing company claims: Machine fills cans normally distributed with $\mu = 500\,\text{ml}$ and $\sigma = 1\,\text{ml}$
- Brewery takes 10 samples
- Mean of these samples is 499.22 ml
- Less than 500 ml, but still within $\mu = 500\,\text{ml}$ and $\sigma = 1\,\text{ml}$?
- How to check this?
- Would mean be 421.54 ml: Would complain
- Where is boundary between ok and not ok?

Estimation

- Inference for μ (true μ or μ_0) from data
- Approximating $\mathrm{E}(X_i)$ and σ_X^2 (true but unknown values) by mean and variance of given data an
- Speak of an *estimate* $\widehat{\mu}$ of $\mathrm{E}(X_i)$
- Analogous: estimation $\widehat{\sigma}_X^2$ of σ_X^2
- Notation: Hat $\widehat{\ }$ denotes estimate of a quantity

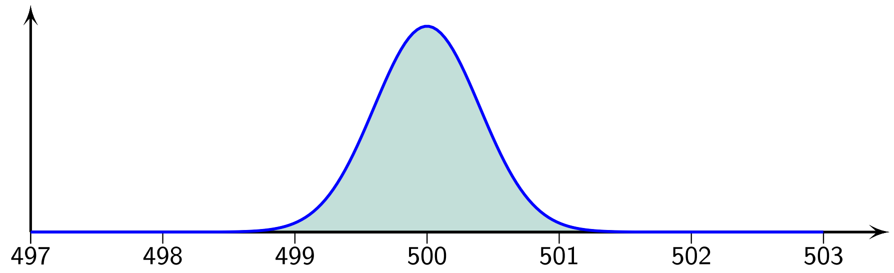
- (Point) estimates for expected value and variance are:

$$\widehat{\mu} = \frac{1}{n}\sum_{i=1}^{n} X_i, \qquad \widehat{\sigma}_X^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \overline{X}_n)^2$$

- Estimators: Functions of random variables $X_1, \ldots, X_n$
- $\widehat{\mu}$ and $\widehat{\sigma}_X^2$ are themselves random variables
- In example: $\widehat{\mu} = 499.22$
- Standard deviation is ignored
- Because of CLT: Mean approximately distributed like

$$\overline{X}_{10} \sim \mathcal{N}\left(500, \frac{1^2}{10}\right)$$

- Sketch:



- If $\widehat{\mu}$ "near" at $\mu_0 = 500$: Consider $\mu_0 = 500$ as plausible
- If $\widehat{\mu}$ "far" from $\mu_0 = 500$: Consider $\mu_0 = 500$ as *not* plausible
- What does "near" or "far away" mean?

Procedure Hypothesis Test

- Assumption: Data normally distributed with $\mu = 500$ and $\sigma = 1$
- How to check if mean $\mu = 500$ is plausible?
- Basic idea: Using a observation, check whether, under assumption $\mu = 500$, mean of observations is probable or not
- To do this, select 10 observations with model

Model

10 observations are realisations of RV $X_1, X_2, \ldots, X_{10}$, where X_i is continuous RV and

$$X_1, \ldots, X_{10} \text{ i.i.d.} \sim \mathcal{N}(500, 1^2)$$

- Want to check whether *assumption* $\mu_0 = 500$ is justified
- Introduce following terms:

Null Hypothesis

$$H_0: \quad \mu = \mu_0 = 500$$

Alternative Hypothesis

$$H_A: \quad \mu \neq \mu_0 = 500 \quad (\text{or "}<\text{" or "}>\text{"})$$

- μ: True (unknown) mean of data
- μ_0: Assumed true mean of data

Remark

- Remarked earlier: $\overline{x} = \mu$ is never 500 ml
- So: Alternative hypothesis always satisfied
- In this context: $\mu \neq 500$ means that μ is "far away" from 500
- What is "far away"?

- (Estimated) mean: $\widehat{\mu} = 499.22$
- What does it mean that this mean is (un)probable?
- Probability:

$$P(\overline{X}_{10} = 499.22)$$

- Of no use, since this is 0

- Since $\widehat{\mu} < 500$ is: Consider following probability:

$$P(\overline{X}_{10} \leq 499.22)$$

- Assuming $\mu = 500$ and $\sigma = 1$, $\overline{X}_{10}$ is distributed as:

$$\overline{X}_{10} \sim \mathcal{N}\left(500, \frac{1^2}{10}\right)$$

- Test with this distribution whether assumption $\mu = 500$ is justified

Test Statistic

Distribution of test statistic T under the null hypothesis H_0:

$$T: \quad \overline{X}_{10} \sim \mathcal{N}\left(500, \frac{1^2}{10}\right)$$

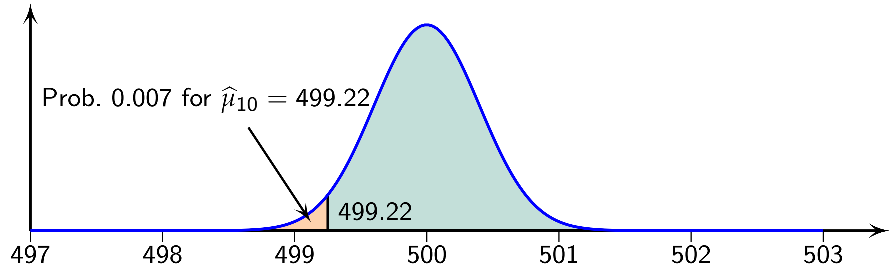
- Probability:

$$P(\overline{X}_{10} \leq 499.22) = 0.007$$

```r
pnorm(q = 499.22, mean = 500, sd = 1/sqrt(10))

[1] 0.006820578
```

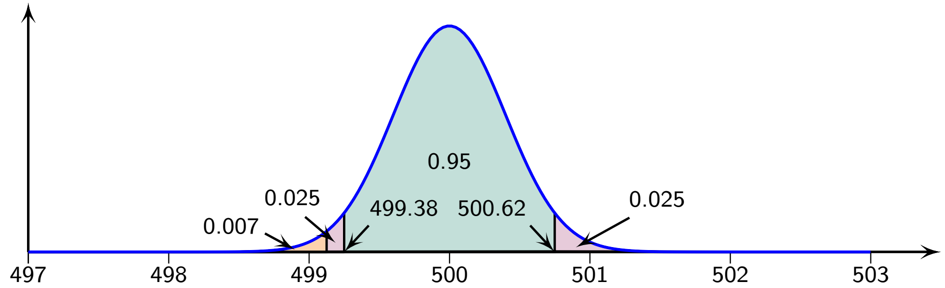
- Sketch:



- This value is small: About 0.7 %
- But is it *too* small?
- *Convention*: It has proven practical to set this limit of what is too small and what is not at 2.5 %
- According to convention:

$$P(\overline{X}_{10} \leq 499.22) < 0.025$$

- Sketch:

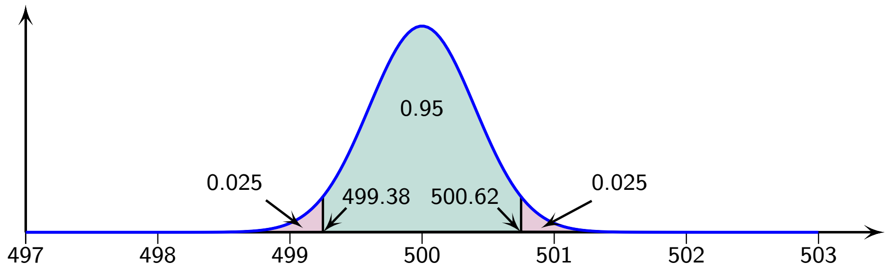


- Consider mean $\widehat{\mu} = 499.22$ as *too unlikely* to fit (true) mean $\mu = 500$
- *So assume that given mean of* $\mu_0 = 500$ *is not plausible!*
- Say: "We reject null hypothesis!"

Graphical Representation

- Divide normal distribution curve into three parts:

- Figure:



  - Symmetric part around mean $\mu = 500$: Amounts to 0.95
  - Both parts left and right must add up to 0.05
  - So for each part 0.025

- Notion:

Significance Level α

- Significance level α, indicates how high a risk one is willing to take of making a wrong decision
- For most tests: α value of 0.05 or 0.01
- Use here:

$$\alpha = 0.05$$

- Significance level sets red area in Figure before

- Red area: *Rejection range*

- Boundary rejection range: 0.025- and 0.975-quantiles:

```
qnorm(p = c(0.025, 0.975), mean = 500, sd = 1/sqrt(10))

[1] 499.3802 500.6198
```

- If observed mean lies in red area of Figure, null hypothesis $\mu_0 = 80$ is rejected

- This area is called:

Rejection Range

$$K = (-\infty, 499.38] \cup [500.62, \infty)$$

- Assume that a mean of observations in rejection range is so improbable that the correctness of $\mu = 500$ must be doubted

- Use measurements to check whether or not its mean is in rejection range

- Make so-called:

Test Decision

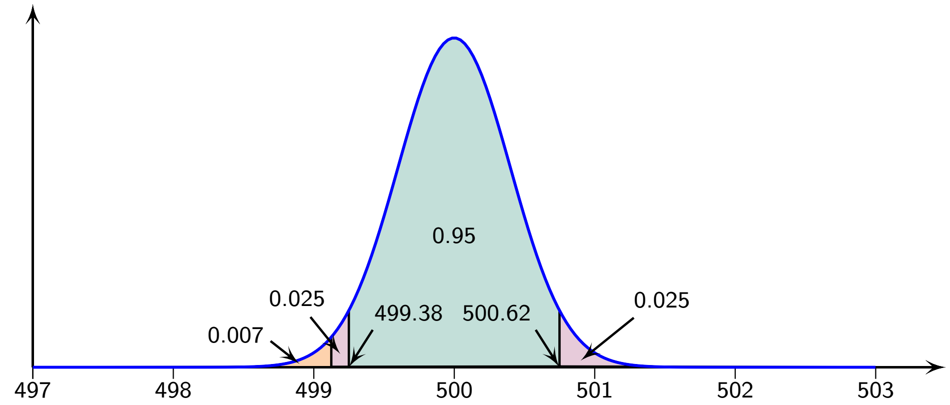
In example:

$$\overline{X}_{10} = 499.22 \notin K$$

- This value is not in the rejection range

- Do not reject null hypothesis

- Sketch:



Remarks

- Why divide rejection range to left *and* right if already known that observed mean is less than $\mu_0 = 500$?
- Well: *Not* known before measurement
- Observed mean could also have been larger than $\mu_0 = 500$
- In this case: Speak of a *two-sided test*
- There are also *one-sided tests* (see later)
- Best practice: Decide on test *before* analysing data

- Made an *assumption* that total rejection range should be 5 % (significance level 5 %)
- This assumption has proved to be practical, but it is also possible to choose 1 %, which is done from time to time

Larger Measurement Series

- Want to check whether scope of dataset affects test decision
- Choose observations of different lengths n, which all have observed mean value $\widehat{\mu} = 499.22$

- Determine for all measurement series:

$$P(\overline{X}_n \leq 499.22) \qquad \text{with} \qquad \overline{X}_n \sim \mathcal{N}\left(500, \frac{1^2}{n}\right)$$

- If this value is greater than 0.025, then null hypothesis is not rejected, otherwise it is rejected

- For $n = 2$:

$$P(\overline{X}_2 \leq 499.22) = 0.079 > 0.025$$

```
pnorm(q = 499.22, mean = 500, sd = 1/sqrt(2))

[1] 0.1349948
```

- Null hypothesis is not rejected at significance level 5 %

- For $n = 4$:

$$P(\overline{X}_4 \leq 499.22) = 0.053 > 0.025$$

```
pnorm(q = 499.22, mean = 500, sd = 1/sqrt(4))

[1] 0.05937994
```

- Null hypothesis is not rejected

- For $n = 6$:

$$P(\overline{X}_6 \leq 499.22) = 0.028 > 0.025$$

```
pnorm(q = 499.22, mean = 500, sd = 1/sqrt(6))

[1] 0.02802787
```

- Null hypothesis is (barely) not rejected for $n = 6$

- And finally, for $n = 8$:

$$P(\overline{X}_8 \leq 499.22) = 0.014 < 0.025$$

```
pnorm(q = 499.22, mean = 500, sd = 1/sqrt(8))

[1] 0.01368594
```

- Null hypothesis rejected for $n = 8$

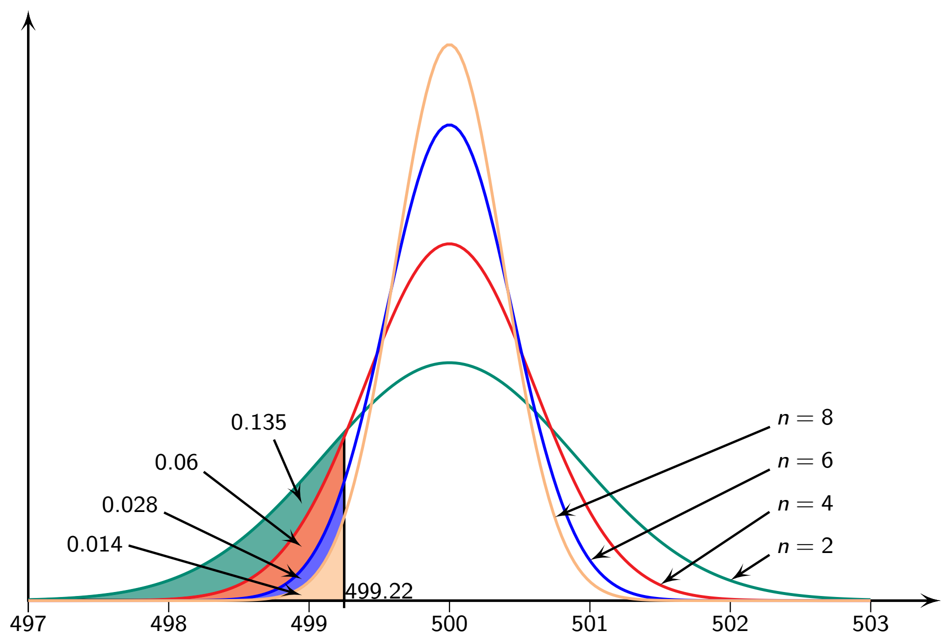
- With increasing n following value becomes smaller and smaller:

$$P(\overline{X}_n \leq 499.22)$$

- Reason: Standard deviation becomes smaller with larger n

- Normal distribution curves become narrower (Figure next slide)

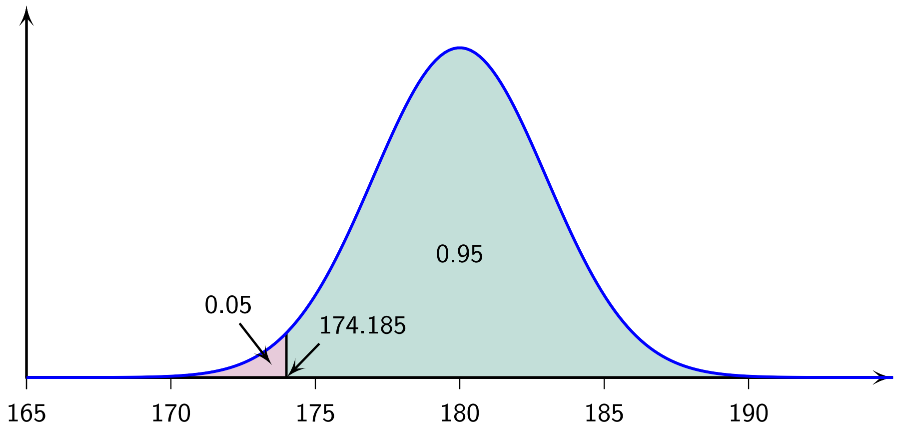
- Sketch:



Example: Body Height Women

- Newspaper claims: Average height of adult women in Switzerland at 180 cm with a standard deviation 10 cm
- Assumption: Mean too large
- Two-sided test makes little sense, because "known" that this mean is too large
- I.e.: True value is most likely to be lower
- Consideration similar to before, but do not divide rejection range to both sides, but only to left, as expected true mean to be lower than $\mu_0 = 180$ (Figure next slide)
- Make a *one-sided* test

- Sketch:



- *Model*:

$$X_1, \ldots, X_n \text{ i.i.d.} \qquad X_i \sim \mathcal{N}(180, 10^2)$$

- Assumption: The true mean is 180 cm
- *Null hypothesis:*

$$H_0: \qquad \mu = \mu_0 = 180$$

- *Alternative hypothesis*:

$$H_A: \qquad \mu < 180$$

- Investigate n people and test whether:

$$P(\overline{X}_n < \overline{x}_n) < 0.05$$

- Rejection range here is therefore one-sided to the left (left-tailed)

- Figure before: Rejection range for $n = 8$ drawn in pink
- *Test statistic under the null hypothesis H_0*:

$$\overline{X}_8 \sim \mathcal{N}\left(180, \frac{10^2}{8}\right)$$

- *Significance level*:

$$\alpha = 0.05$$

- Boundary of rejection range:

```
qnorm(p = 0.05, mean = 180, sd = 10/sqrt(8))

[1] 174.1846
```

- *Rejection range* (see Figure before):

$$K = (-\infty, 174.1846)$$

- Rejection range much too large, since hardly any heights of adult women under 40 cm are to be expected
- Working with *model*: Makes sense only in a certain area
- Now randomly select eight adult women, measure their height and determine mean, which is 171.54 cm
- *Test decision:* Observed mean in rejection range and thus null hypothesis *reject* that the true $\mu = 180$ holds
- This mean of randomly selected eight women still seems relatively high, but it is enough to make one doubt the assumption $\mu_0 = 180$

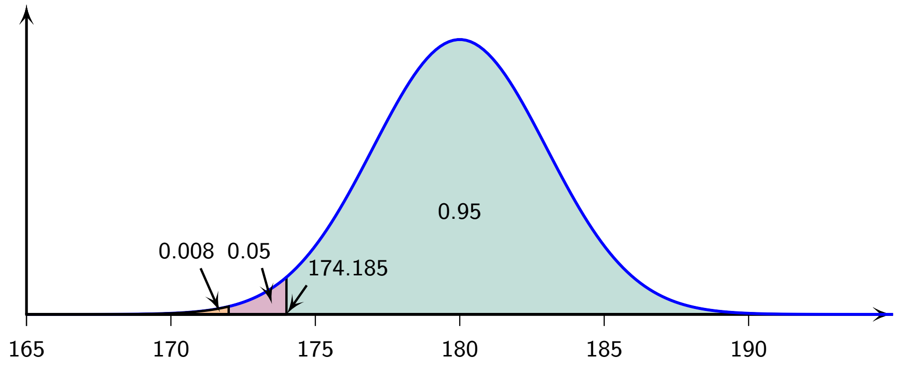
- $P(\overline{X}_6 < 171.54)$ (see Figure below)

$$P(\overline{X}_6 < 171.54) = 0.008\,359\,052$$

```
pnorm(q = 171.54, mean = 180, sd = 10/sqrt(8))

[1] 0.008359052
```

- Sketch:



- This value called *p*-value: Certainty with which test decision is made
- If null hypothesis is rejected: Very small *p*-value (close to 0) indicates that null hypothesis is rejected with more certainty than if it is close to significance level (here $\alpha = 0.05$)

p-Value

- *p*-value 0.008 is a value between 0 and 1
- Indicates how well *null hypothesis* and *data* fit together:
  - 0: does not fit at all
  - 1: fits very well
- More precisely: *p*-value is possibility of obtaining result obtained or a more extreme result under null hypothesis

- p-value thus indicates how extreme result is: The smaller p-value, the more result argues *against* null hypothesis

- Values smaller than a predetermined limit, such as 5 %, 1 % or 0.1 % are reason to reject the null hypothesis

p-value

p-value is probability of observing an event under null hypothesis that is at least as extreme (in direction of alternative) as currently observed event

- Test decision using p-value

p-value and Statistical Test

- One can directly make test decision from p-value: If p-value is smaller than significance level, one discards H_0, otherwise not
- Compared to simple test decision, p-value contains more information: Can see directly how strongly null hypothesis is rejected
- For a given significance level α (e.g. $\alpha = 0.05$), definition of the p-value applies to a one-sided test:
  - ★ Reject H_0 if p-value $\leq \alpha$
  - ★ Do not reject H_0 if p-value $> \alpha$

- Computer packages: Test decision only indirectly with p-value

- In addition to this decision rule: p-value quantifies how significant an alternative is (i.e. how much evidence there is for rejecting H_0)

- Sometimes linguistic formulas or symbols are given instead of p-values:

$$
\begin{aligned}
p\text{-value} &\approx 0.05 : && \text{weakly significant} \\
p\text{-value} &\approx 0.01 : && \text{significant} \\
p\text{-value} &\approx 0.001 : && \text{strongly significant} \\
p\text{-value} &\leq 10^{-4} : && \text{highly significant}
\end{aligned}
$$

p-Value for Two-Sided Test

- Have defined p-value for one-sided tests

- But what is p-value for two-sided tests?

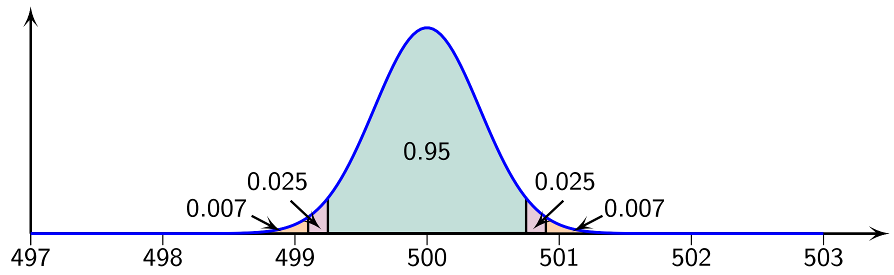
- Example from earlier:

$$P(\overline{X}_6 \leq 499.22) = 0.007$$

- Less than 0.025

- *Could* consider this to be p-value, but do not

- Sketch:



- However, since significance level is $\alpha = 0.05$, probability above is converted to 5 %, i.e. doubled:

$$p\text{-value} = 2 \cdot P(\overline{X}_6 \leq 499.22) = 0.014$$

- Then compare p-value with significance level

- Computer software returns p-value *always* at significance level