

# 9. Introduction to Statistics

Gunvor Elisabeth Kirkelund  
Lars Mandrup

---

# Today's Content

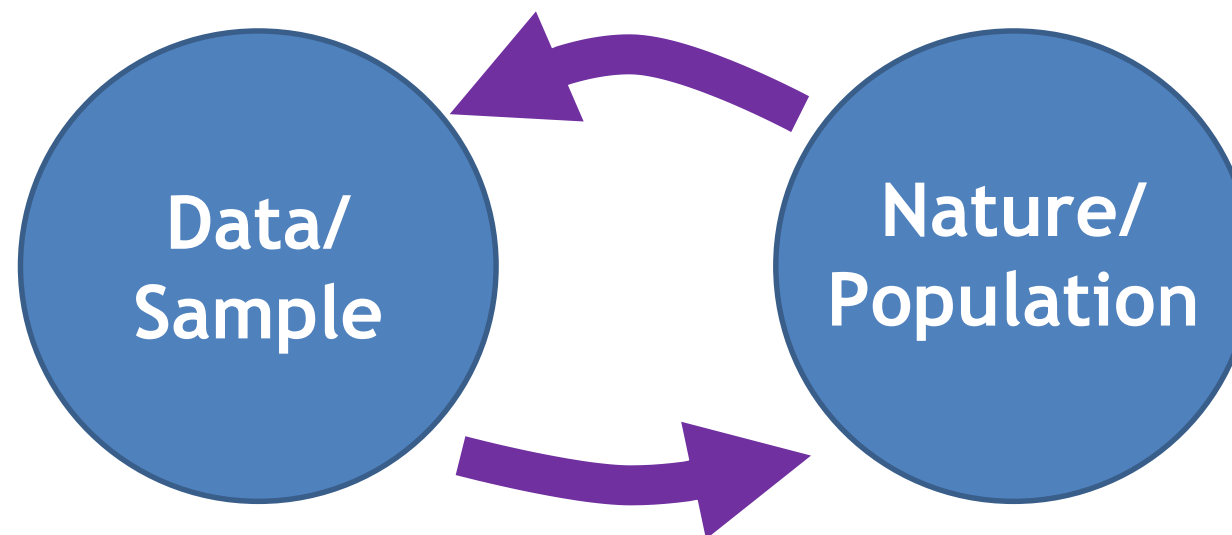
---

- ❖ Introduction to Statistics
- ❖ Estimators
- ❖ Significance Level and p-value
- ❖ Confidence Intervals
- ❖ Sample Size Determination

# Introduction to Statistics

## Probability theory

*Given the cause (population), what should the data (sample) look like?*



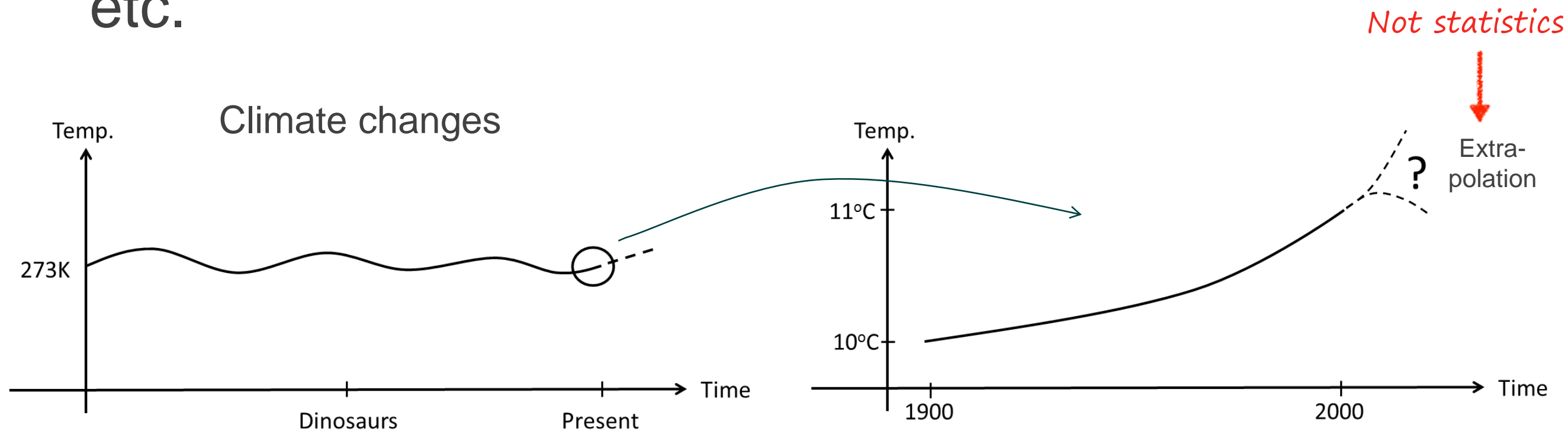
## Statistics

*Given the data (sample), what caused them (population)?*

- Testing a hypothesis
- Estimating means and variances
- If we don't know better: We assume data are normally distributed

# Descriptive Statistics

- ❖ *Descriptive statistics* summarizes data from a sample using indexes such as the mean or standard deviation.
- ❖ Descriptive statistics also includes tables, graphs, etc.





---

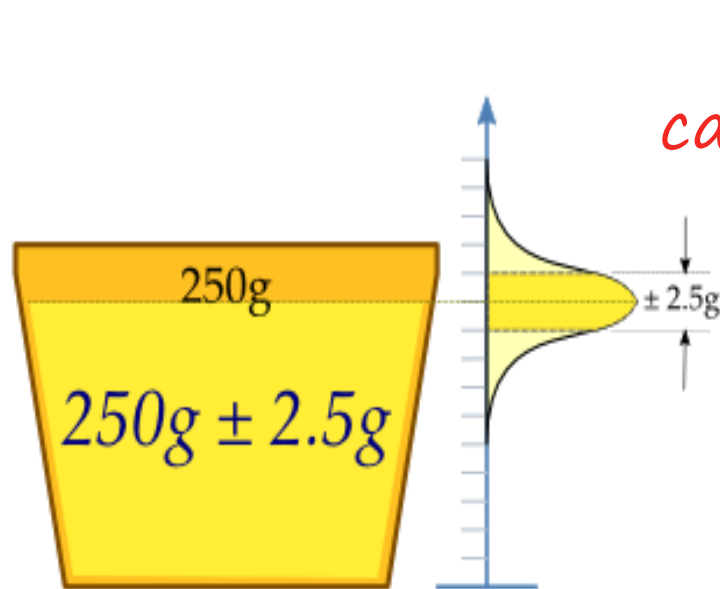
# Inferential Statistics

---

- ❖ *Inferential statistics* infers predictions about a larger population than the sample represents.
- ❖ It uses patterns in the sample data to draw inferences about the population represented, accounting for randomness.
- ❖ These inferences may take the form of:
  - ❖ answering yes/no questions about the data (*hypothesis testing*)
  - ❖ estimating numerical characteristics of the data (*estimation*)
  - ❖ describing associations within the data (*correlation*)
  - ❖ modeling relationships within the data (*regression analysis*).

# Cup Example

- ❖ A machine fills cups with a liquid, the content of the cups is 250 grams of liquid.
- ❖ The machine cannot fill with exactly 250 grams, the content added to individual cups shows some variation, and is considered a random variable,  $X$ .



*If the machine is adequately calibrated,  $X$  is normally distributed*

$$X \sim N(\mu, \sigma^2)$$

with mean  $\mu = 250$  g and  
standard deviation  $\sigma = 2.5$  g

# Cup Example

*ONE sample of the population!*

- To determine if the machine is adequately calibrated, a sample of  $n = 25$  cups of liquid is chosen at random and the cups are weighed.
- The resulting measured masses of liquid are  $X_1, X_2, \dots, X_{25}$ , a random sample from  $X$ .
- To get an impression of the population mean ( $\mu$ ), we use the average (or sample mean) as an estimate:

*Population — All cups for all times*

*Sample mean is NOT the expected value (true mean)!*


*^ means an  
estimator of the  
true population value*



$$\hat{\mu} = \frac{1}{25} \sum_{i=1}^{25} X_i = 250.2 \text{ g}$$

- Is the machine adequately calibrated?

# Cup Example

- What would we typically observe, if the machine is indeed adequately calibrated?
- Perform 100 simulations of the cup filling experiment, where in each experiment we assume that the machine is adequately calibrated.  *100 samples of the population – fx. in Matlab*
- The result of each experiment is a sample mean,  $\hat{\mu}$ , which is calculated by first drawing 25 random samples  $(X_1, X_2, \dots, X_{25})$  from a normal distribution with mean  $\mu = 250$  grams and standard deviation  $\sigma = 2.5$  grams.

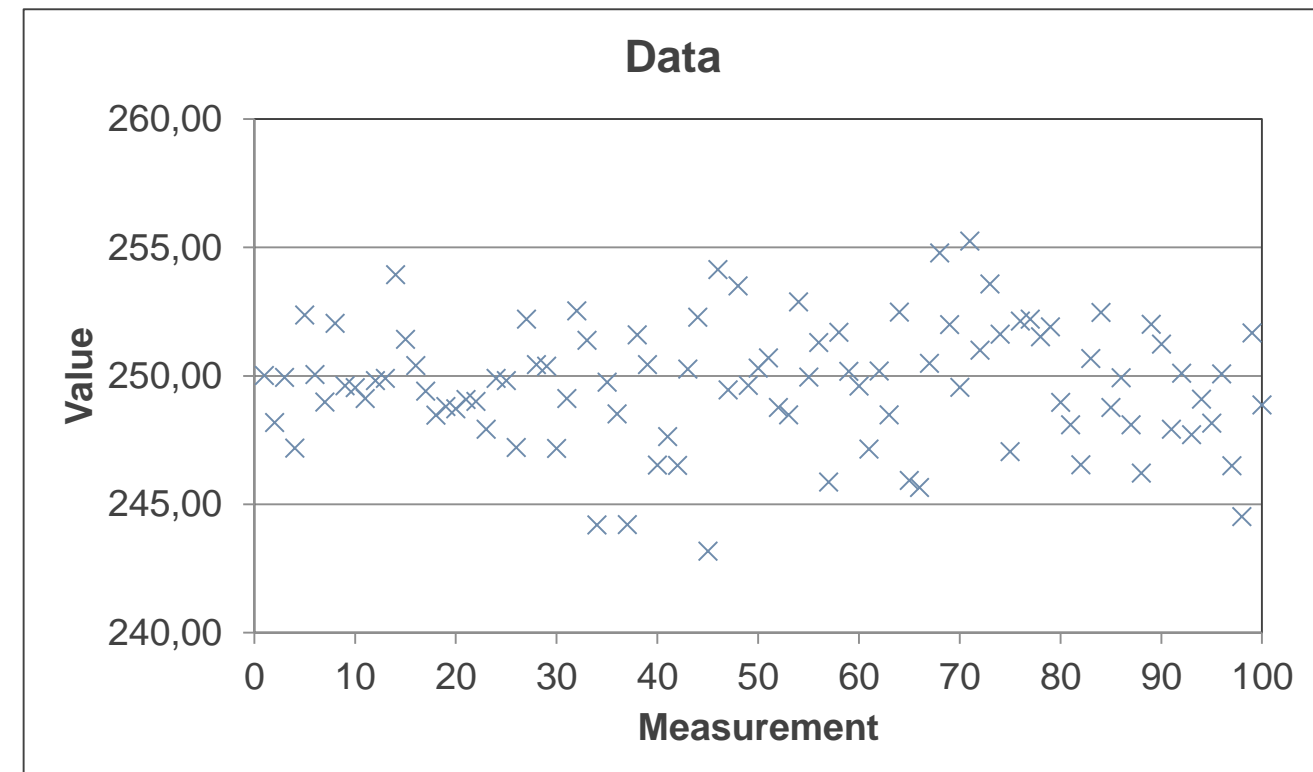
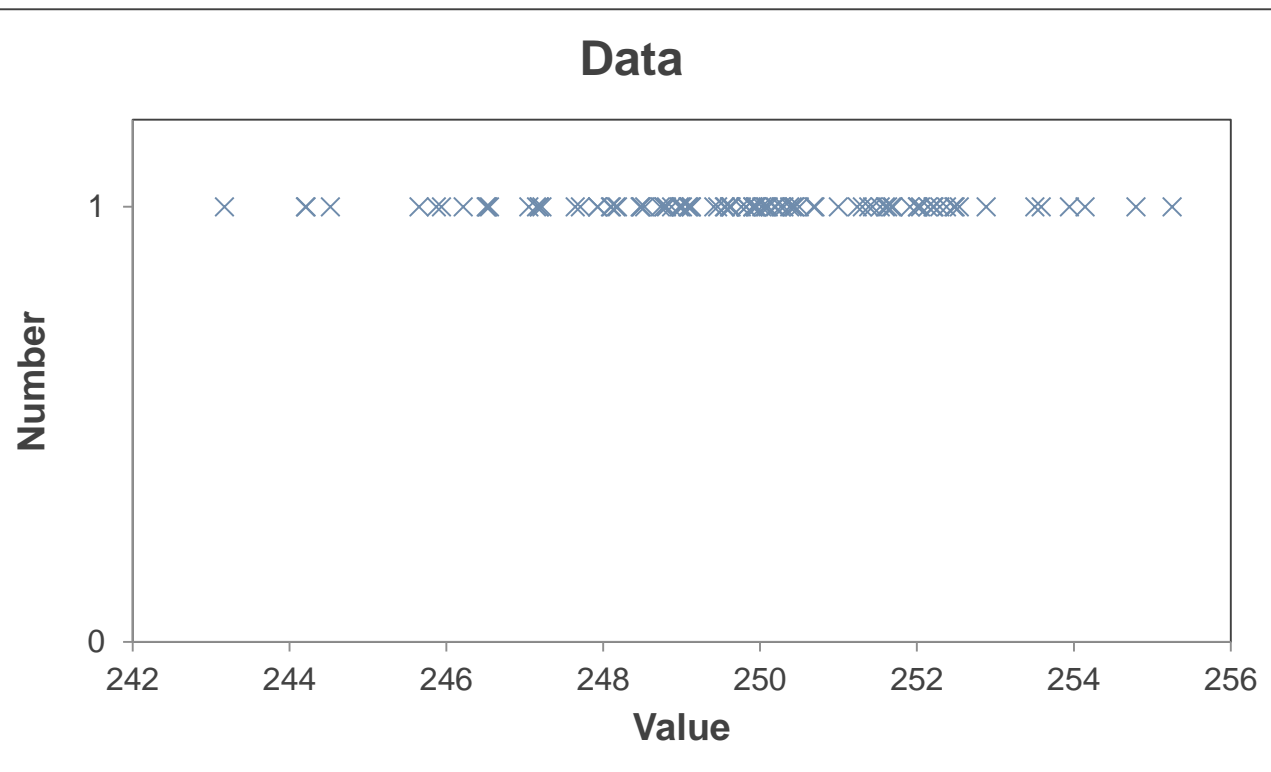
*Think of the Central Limit theorem!!*



# Histograms

- Presentation of data

250,02	249,12	249,08	249,13	247,64	250,70	247,15	255,25	248,09	247,93
248,18	249,82	249,01	252,54	246,52	248,77	250,20	251,00	246,55	250,10
249,95	249,91	247,93	251,39	250,27	248,48	248,49	253,59	250,69	247,71
247,19	253,95	249,91	244,20	252,29	252,88	252,49	251,63	252,47	249,10
252,38	251,43	249,82	249,76	243,17	249,96	245,94	247,05	248,78	248,16
250,06	250,41	247,22	248,53	254,14	251,30	245,65	252,14	249,94	250,08
248,99	249,41	252,21	244,21	249,46	245,87	250,50	252,22	248,11	246,51
252,05	248,47	250,44	251,60	253,50	251,70	254,79	251,53	246,22	244,52
249,62	248,81	250,38	250,44	249,63	250,18	252,01	251,91	252,02	251,68
249,54	248,73	247,18	246,52	250,31	249,61	249,56	248,97	251,24	248,87



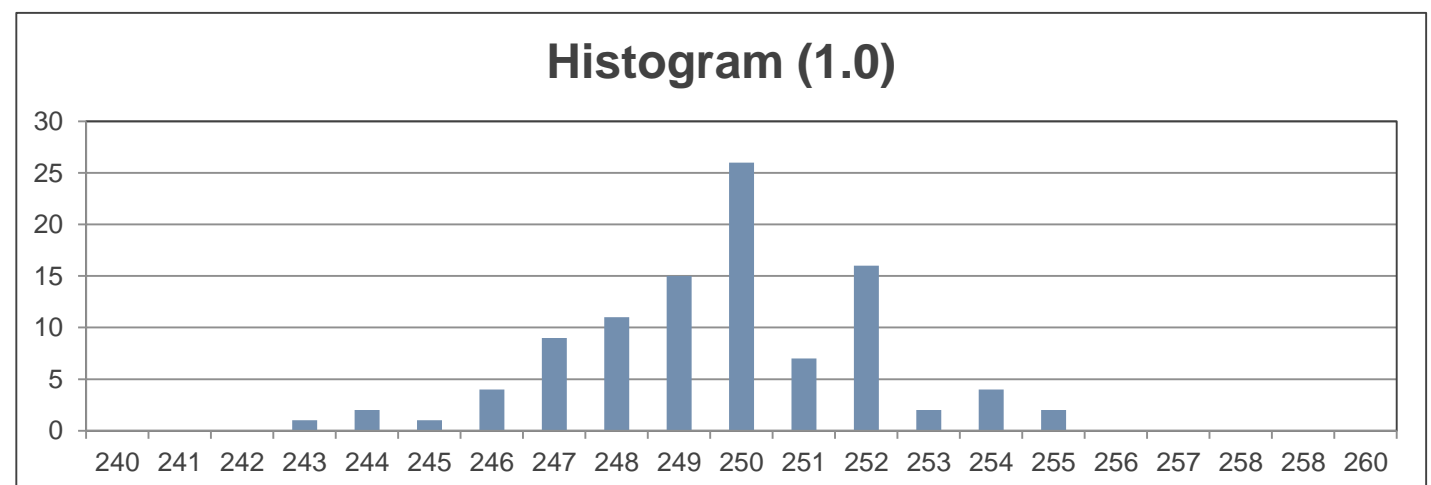
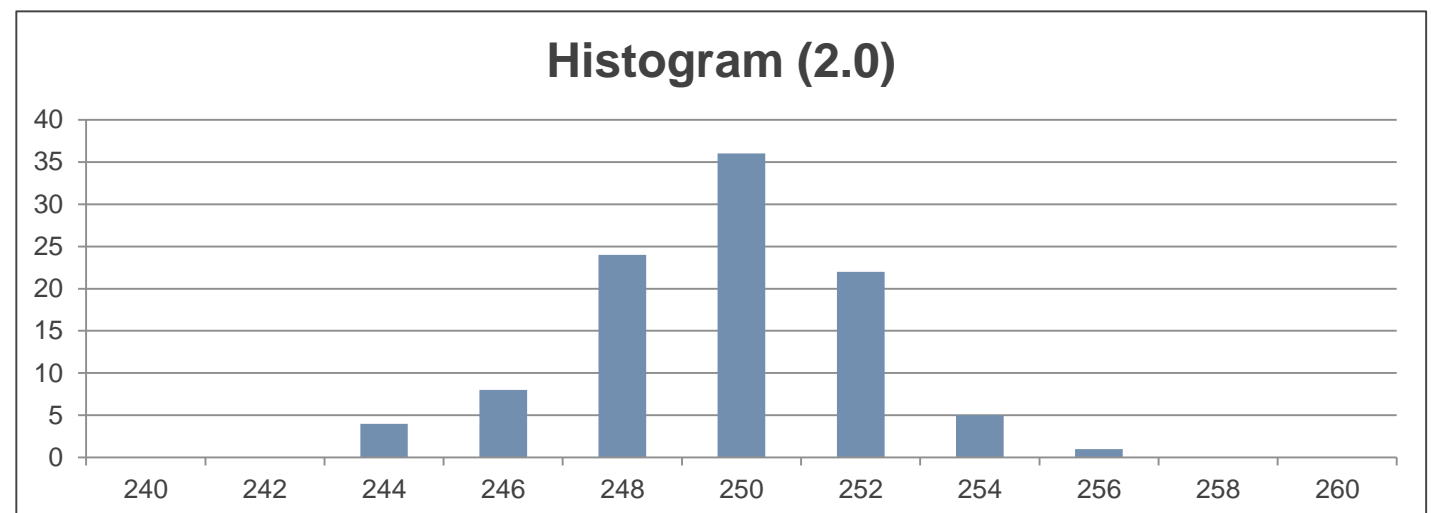
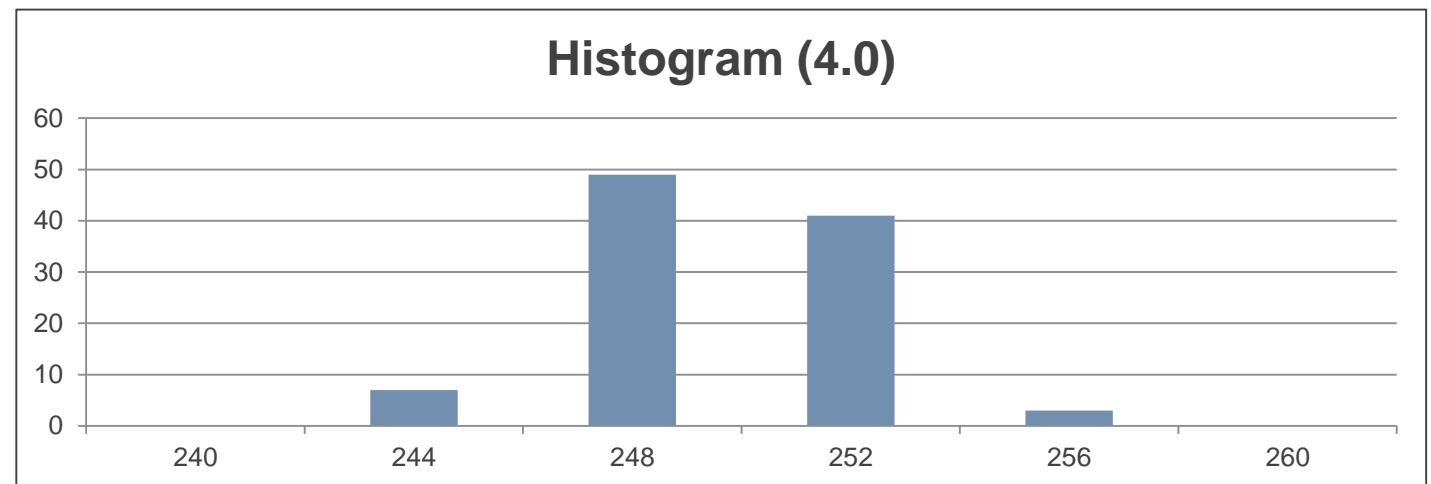
# Histograms

- Grouping of data

Interval	Center	Number
239-241	240	0
241-243	242	0
243-245	244	4
245-247	246	8
247-249	248	24
249-251	250	36
251-253	252	22
253-255	254	5
255-257	256	1
257-259	258	0
259-261	260	0

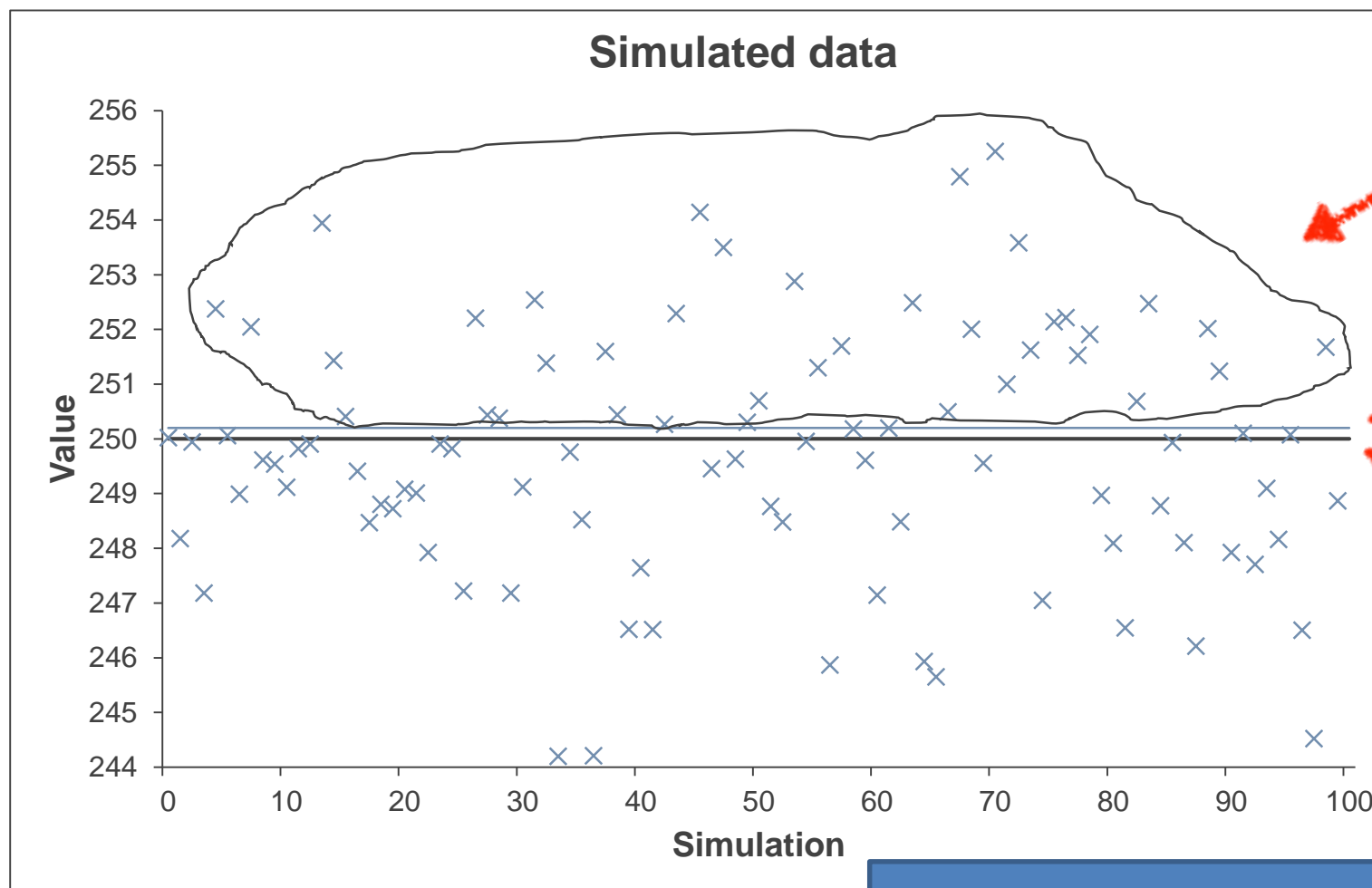
## Matlab:

- `hist(x)` → 10 intervaller/søjler
- `hist(x,n)` → n intervaller/søjler



# Cup Example

- ❖ What would we **typically** observe, if the machine is adequately calibrated?



38 of the 100 simulated values are more extreme than our observation.

Observed value: 250.2

Theoretical value: 250

**Conclusion:** If the machine is adequately calibrated, then due to random variation, we will often observe a sample mean (or average) that is more extreme than the actually observed value of 250.2 g.

# Cup Example

- A loose interpretation of this result is that – just by random variation in  $X$  – there is a 38% chance of observing a sample mean that is larger than what we have observed (250.2 g.), even though the machine is adequately calibrated.
- This means that observing a sample mean of 250.2 g. or larger, when the machine is adequately calibrated, is a relatively common event.
- Hence, the 0.2 gram deviation from the hypothesized mean ( $\mu = 250$  g.) can – maybe – be explained by random variation in  $X$ , and **we conclude that it may be likely that the machine is adequately calibrated.**

*How to make a  
objective conclusion?*

---

# Population

---

*Key term*

- **Definition 1 - Population**

- In statistics, a population is a complete set of items that share at least one property in common that is the subject of a statistical analysis.
- Populations can be diverse topics such as “all persons living in a country” or “every atom composing a crystal”.
- In the cup filling example, the population is “all cups filled by the machine”.



---

# Sample

---

*Key term*

- **Definition 2 - Sample**

- A statistical sample is a subset drawn from the population to represent the population in a statistical analysis.
- If a sample is chosen properly, characteristics of the entire population that the sample is drawn from can be inferred from corresponding characteristics of the sample.
- ❖ • We will denote the sample  $X_1, X_2, \dots, X_n$ , where the samples are iid random variables with a given probability distribution.

---

# Statistical Model

---

*Key term*

- **Definition 3 – Statistical model**

- A random sample  $X$  and its PDF,  $f_X(x; \theta)$ , where  $\theta$  is the parameter of the PDF.

- The parameter,  $\theta$ , is in general a vector and often unknown.



- If  $f_X(x; \theta)$  is a normal distribution with mean  $\mu$  and variance  $\sigma^2$ , then  $\theta = [\mu, \sigma^2]$ .

- The purpose of inferential statistics is to infer knowledge about the unknown parameter(s),  $\theta$ .

---

# Statistic

---

Key term

- **Definition 4 – Statistic**

- A *statistic* is a random variable that is a function of the random sample,  $X$ , but not a function of unknown parameters,  $\theta$ .
- A commonly used statistic is the average or sample mean

$$\bar{x} = \frac{1}{n} \sum_{i=1}^n X_i$$

# Estimator

- **Definition 5 – Estimator**

- An estimator,  $\hat{\theta}(X)$ , is a statistic used to estimate the unknown parameter  $\theta$  of a random sample,  $X$ .

- For notational convenience, we will often write  $\hat{\theta}$  instead of  $\hat{\theta}(X)$ .

- We have already seen an example of an estimator,  
❖ namely the sample mean

$$\hat{\mu} = \bar{x} = \frac{1}{n} \sum_{i=1}^n X_i$$

- which is an estimator of the (true) population mean,  $\mu$ , of a normally distributed random variable,  $X$ .

---

# Maximum-likelihood Estimator

---

- Maximum-likelihood estimation (MLE) is a method of estimation the parameters of a statistical model.
- For a given distribution with the parameter  $\hat{\theta}$  and the pdf  $f_X(x | \hat{\theta})$  the likelihood function is:

$$L(\hat{\theta}) = f_X(x | \hat{\theta})$$

- Important:
  - The likelihood function  $L(\hat{\theta})$  is a function of the parameter estimate  $\hat{\theta}$  and not  $x$
  - The likelihood function  $L(\hat{\theta})$  is the probability density of observing  $x$  if the true parameter is  $\hat{\theta}$
- **The optimum (MLE) estimate of the parameter is the one that maximizes  $L(\hat{\theta})$ .**



# Maximum-likelihood Estimators

## Mean-estimator:

- Independent observations  $x_1, x_2, \dots, x_n$  of a stochastic variable  $X$

Maximum-likelihood Estimator of the mean  $\longrightarrow$

$$\hat{\mu} = \frac{1}{n} \sum_{i=1}^n x_i = \bar{x}$$

$\longleftarrow$  Average of observations

## Success rate-estimator:

- $x$  = number of successes in  $n$  independent Bernoulli trials

Maximum-likelihood Estimator of the success rate  $\longrightarrow$

$$\hat{p} = \frac{x}{n}$$

$\longleftarrow$  Average of observed success rate

## Event rate-estimator:

- $x$  = number of observed independent events in time  $t$

Maximum-likelihood Estimator of the event rate  $\longrightarrow$

$$\hat{\lambda} = \frac{x}{t}$$

$\longleftarrow$  Average of observed event rate

# Unbiased Estimator<sup>Key term</sup>

- **Definition 6 – Unbiased estimator**

- An estimator,  $\hat{\theta}$ , is unbiased if

$$E[\hat{\theta}] = \theta$$

- i.e., if it's expected value is equal to the true value of the unknown parameter being estimated.

- The sample mean

$$\hat{\mu} = \bar{x} = \frac{1}{n} \sum_{i=1}^n X_i$$

- is an unbiased estimator of the population mean,  $\mu$ , of a normally distributed random variable,  $X$ .

- The sample succesrate  $\hat{p} = \frac{x}{n}$  is also an unbiased estimator:

$$E[\hat{p}] = E\left[\frac{x}{n}\right] = \frac{1}{n} \cdot E[x] = \frac{1}{n} \cdot np = p$$

# The Sample Variance Key term

- Let  $X_1, X_2, \dots, X_n$  be i.i.d. samples of a random variable  $X$  with mean  $\mu$  and variance  $\sigma^2$ . Then the maximum-likelihood estimate of the variance is

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

- This is a biased estimator

$$E[\hat{\sigma}^2] = \sigma^2 - \frac{1}{n} \sigma^2 \neq \sigma^2$$

- The unbiased estimate of the variance is

*n-1 degrees of freedom*

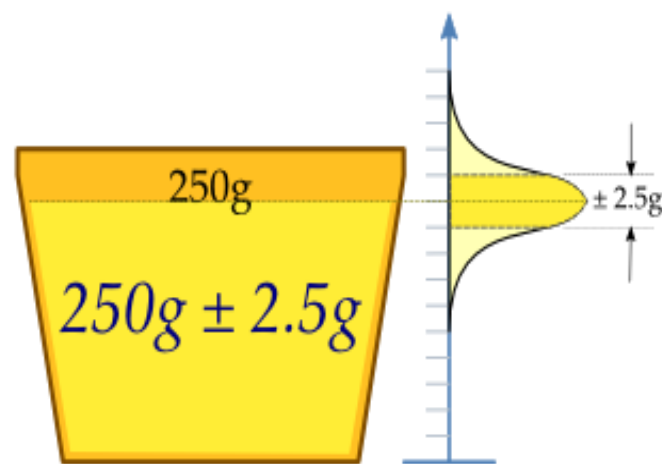
$$s^2 = \frac{1}{n-1} \sum_{i=1}^n (x_i - \bar{x})^2$$

*Matlab*

- We refer to this unbiased estimator as the *sample variance* or *empirical variance*.

# Cup Example

- A machine fills cups with a liquid, the content of the cups is 250 g. of liquid.
- The machine cannot fill with exactly 250 g. The content added to individual cups shows some variation, and is considered a random variable  $X$ .



- Statistical model:  $X \sim \mathcal{N}(250, 2.5)$
- Statistics:  $\bar{X} = \frac{1}{n} \sum_{i=1}^n X_i$
- Test sample:  $X_1, X_2, \dots, X_{25}$
- Estimator:  $\hat{\mu} = \bar{X} = \frac{1}{25} \sum_{i=1}^{25} X_i = 250.2 \text{ g.}$

➤ Is the machine adequately calibrated?

# Central Limit Theorem

- Let  $X_1, X_2, \dots, X_n$  be i.i.d. samples of a random variable  $X$  with mean  $\mu$  and variance  $\sigma^2$ .
- Then, as  $n \rightarrow \infty$ , the sample mean ( $\bar{X}$ ) becomes normally distributed with a mean that is equal to the population mean ( $\mu$ ) and a variance that is scaled by  $1/n$ :

*Dependent of  $\mu$*       *Independent of  $\mu$*

*Sample mean*       $\bar{X} \sim N(\mu, \sigma^2/n) \rightarrow Z = \frac{\bar{X} - \mu}{\sigma/\sqrt{n}} \sim \mathcal{N}(0,1)$

- Note that  $X$  can have any distribution, i.e., it is *not* required to be normally distributed.
- Although the exact number is subject of debate, it is common practice to require that the number of samples ( $n$ ) should be 30 or larger in order to apply the CLT:

$$n \geq 30$$

*...most important number in statistics*



# Test Statistics

*Key term*

*We can define which test statistics we use – z is an option.*

- **Definition 11 – Test statistic**

- A random variable that summarizes a data-set by reducing the data to one value that can be used to perform the hypothesis test.

- The probability distribution of a test statistic, as opposed to descriptive statistics, does not depend on the unknown parameter ( $\theta$ ).  
❖

- Example (z-score):

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

*sample mean* →  $\bar{x}$

*expected value/true mean* →  $\mu$

*standard deviation, sqrt(var(x))* →  $\sigma$

*n – number in the sample* →  $n$

*Standard normal distribution ( $\mu=0$  and  $\sigma^2=1$ )* →  $N(0,1)$

# Test Statistics – Cup example

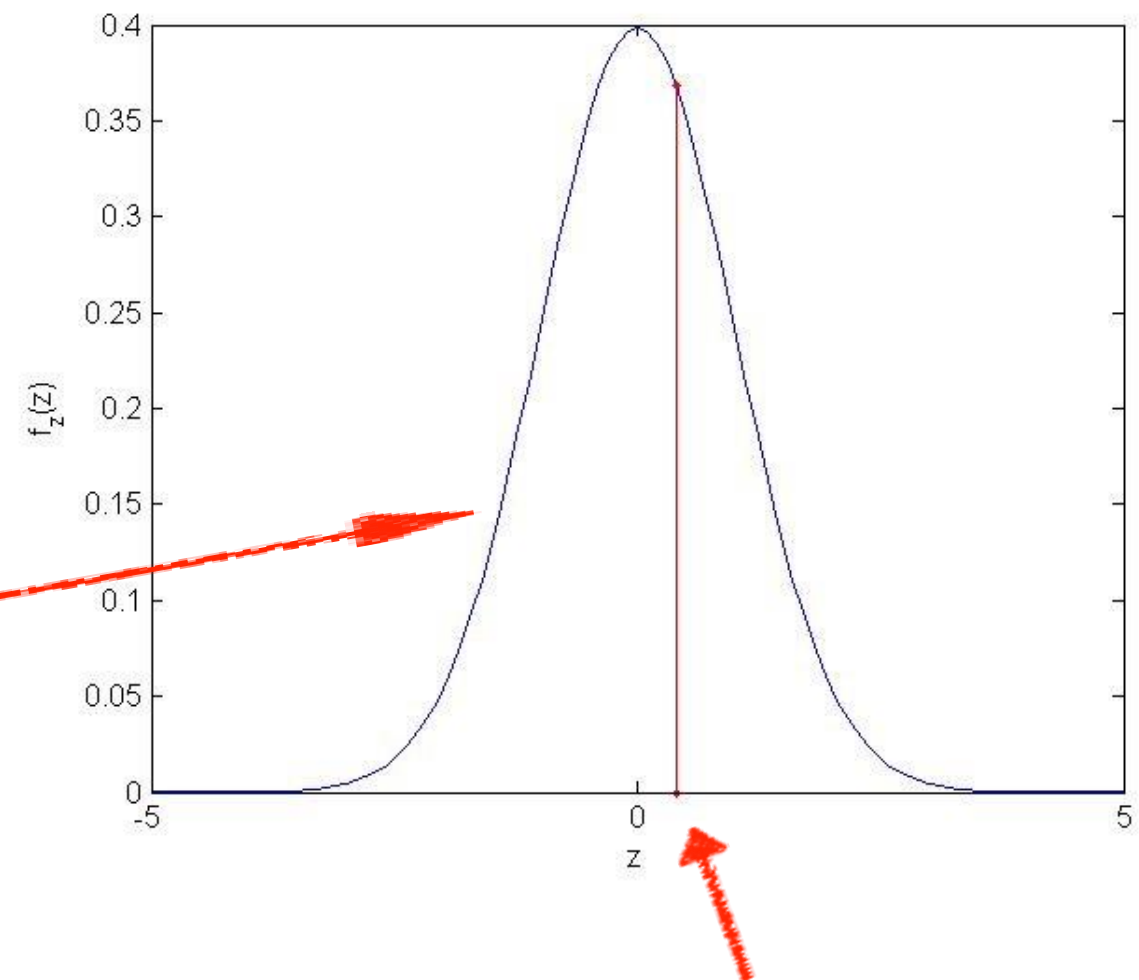
*We define a test statistics  $z$ :*

$$z = \frac{\bar{x} - \mu}{\sigma / \sqrt{n}}$$

*Statistical model:*

$$f_Z(z) = \frac{1}{\sqrt{2\pi}} \cdot e^{-\frac{z^2}{2}} \sim \mathcal{N}(0,1)$$

Standard normal distribution  
(PDF)



*Test statistics:*  $z = \frac{\bar{x} - \mu}{\sigma / \sqrt{n}} = \frac{250.2 - 250}{2.5 / \sqrt{25}} = 0.4$

# Plausible result?

*Let's make a test for z*

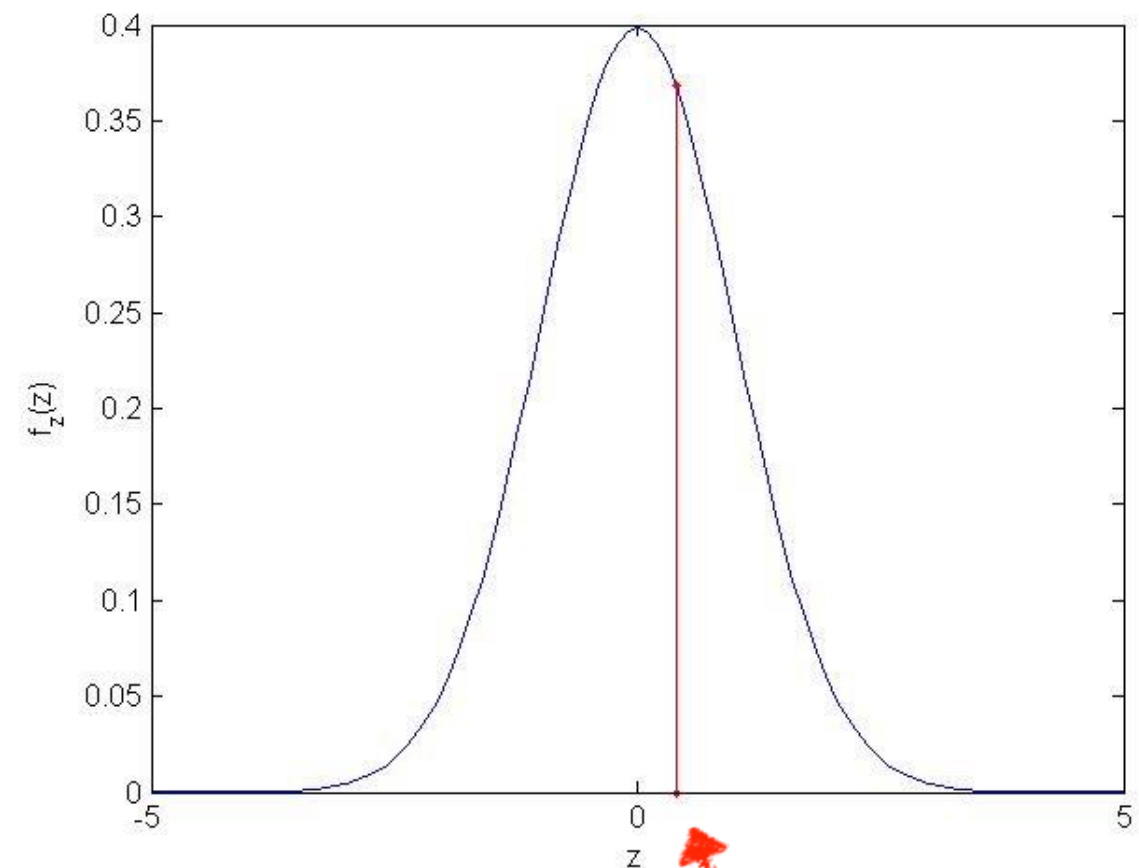
Does it seem plausible that  $z=0.4$  is an observation drawn from a standard normal distribution?

Same as asking: what is the probability of observing a test size ( $z$ ) that is more extreme than 0.4?

**p-value:**

$$\begin{aligned} &Pr(Z > z \cup Z < -z) \\ &= Pr(Z > z) + Pr(Z < -z) \\ &= 2 \cdot Pr(Z > z) \\ &= 2 \cdot (1 - Pr(Z \leq z)) \\ &= 2 \cdot (1 - Pr(Z \leq 0.4)) \\ &= 2 \cdot (1 - \Phi(0.4)) \leftarrow \text{normcdf}(0.4) \\ &= 2 \cdot (1 - 0.6554) \\ &= 0.6892 \end{aligned}$$

Standard normal distribution  
(PDF)



*Test statistics:*  $z = \frac{\bar{x} - \mu}{\sigma/\sqrt{n}} = \frac{250.2 - 250}{2.5/\sqrt{25}} = 0.4$

*Compare with the  $\alpha$ -value!*

---

# p-value

---

Key term

- **Definition 10 – p-value**

- The p-value is the probability of getting a result equal to or more extreme than the observed test-result under the assumption of a given distribution:

$$Pr(\text{Worse test-result than } \hat{X} \mid f_X(\theta))$$

- One-sided:  $p - value = \Pr(X \geq x_{test}) = \Pr(Z \geq z_{test})$  (right-sided)  
 $p - value = \Pr(X \leq x_{test}) = \Pr(Z \leq z_{test})$  (left-sided)
- Two-sided:  $p - value = \Pr(X \geq \mu + \delta) + \Pr(X \leq \mu - \delta); \quad \delta = |\mu - x_{test}|$   
 $= \Pr(Z \geq |z_{test}|) + \Pr(X \leq -|z_{test}|)$

---

# Significance Level *Key term*

---

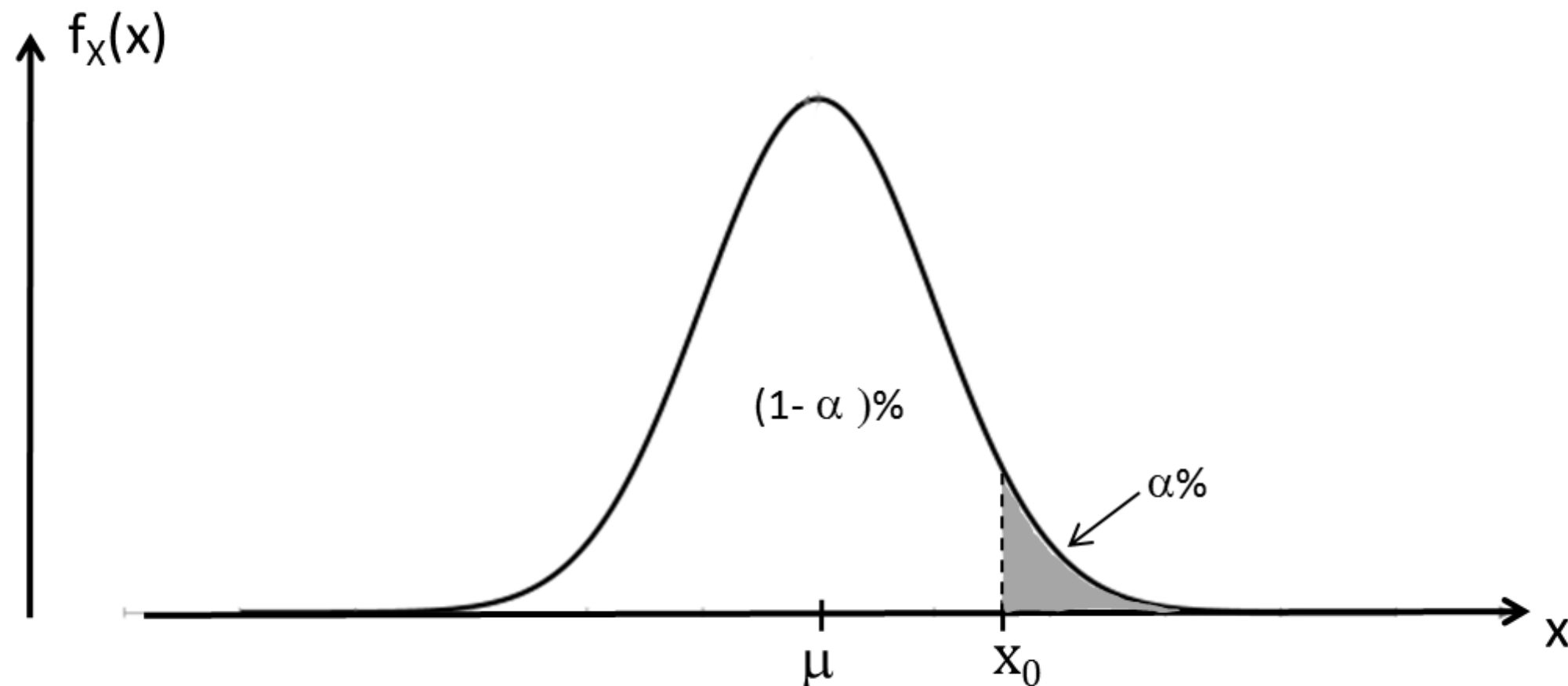
- **Definition 9 – Significance level  $\alpha$** 
  - The statistical significance level  $\alpha$  is the lower limit we will accept for the probability of getting a more extreme result under the assumption of a given distribution.
  - The most common used significance level is  $\alpha = 0,05$  (5%)
- By comparing the p-value for the test with the significance level  $\alpha$  we can decide whether the test-result is plausible with the assumed distribution or not.



# Significance Level

- If the test is one-sided (there is only a limit of how much we will allow the result to deviate from the expected value to one side (larger or lower)), this means that:

$$\Pr(X \geq x_0) = 1 - \Pr(X < x_0) = 1 - F_X(x_0) = \alpha$$



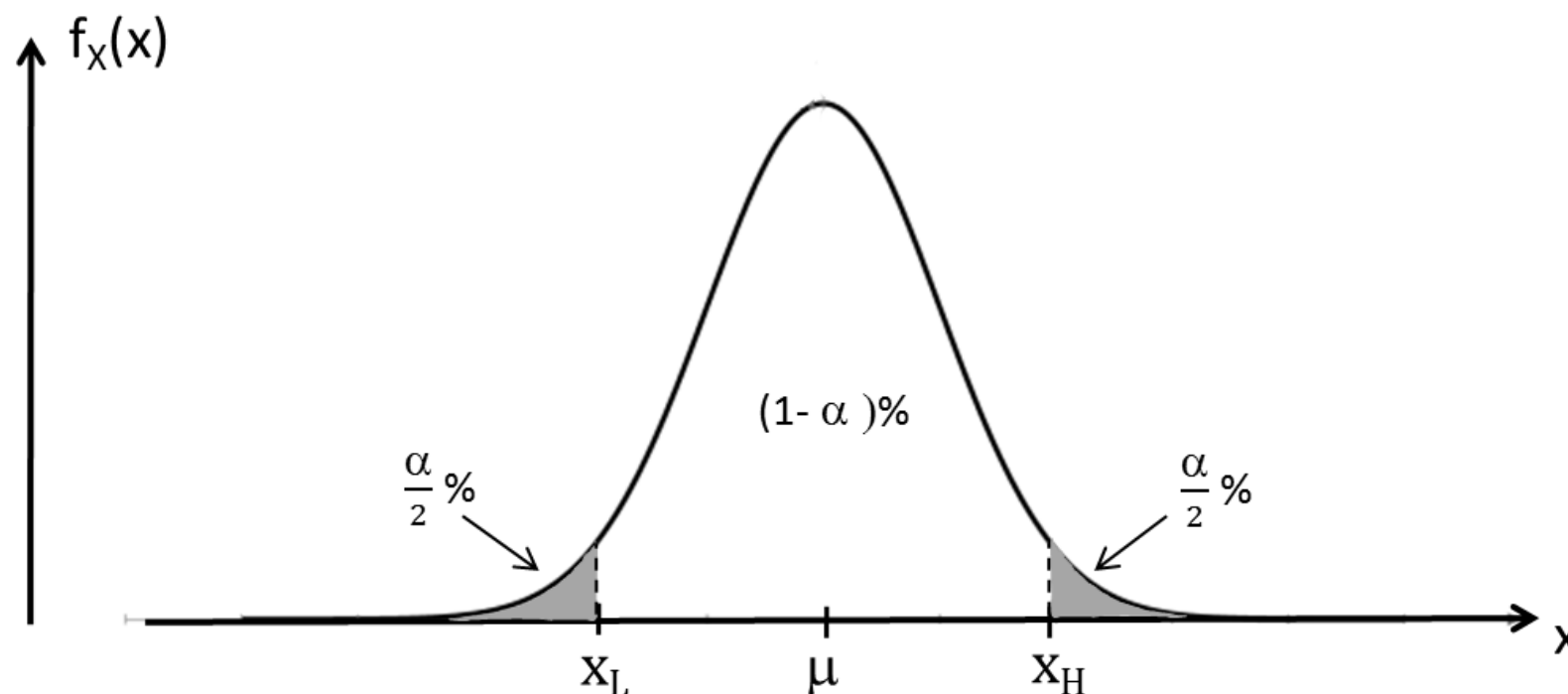
# Significance Level

- If the test is two-sided (there are limits of how much we will allow the result to deviate from the expected value to both sides (larger and lower)), this means that:

$$\Pr(X \leq x_L \cup X \geq x_H) = \Pr(X \leq x_L) + \Pr(X \geq x_H) = \alpha$$

If the distribution is symmetric  $|\mu_X - x_L| = |\mu_X - x_H|$  and  $\Pr(X \leq x_L) = \Pr(X \geq x_H)$  so:

$$\Pr(X \leq x_L) = \Pr(X \geq x_H) = 1 - \Pr(X \leq x_H) = \frac{\alpha}{2}$$



# Plausible result?

Yes!

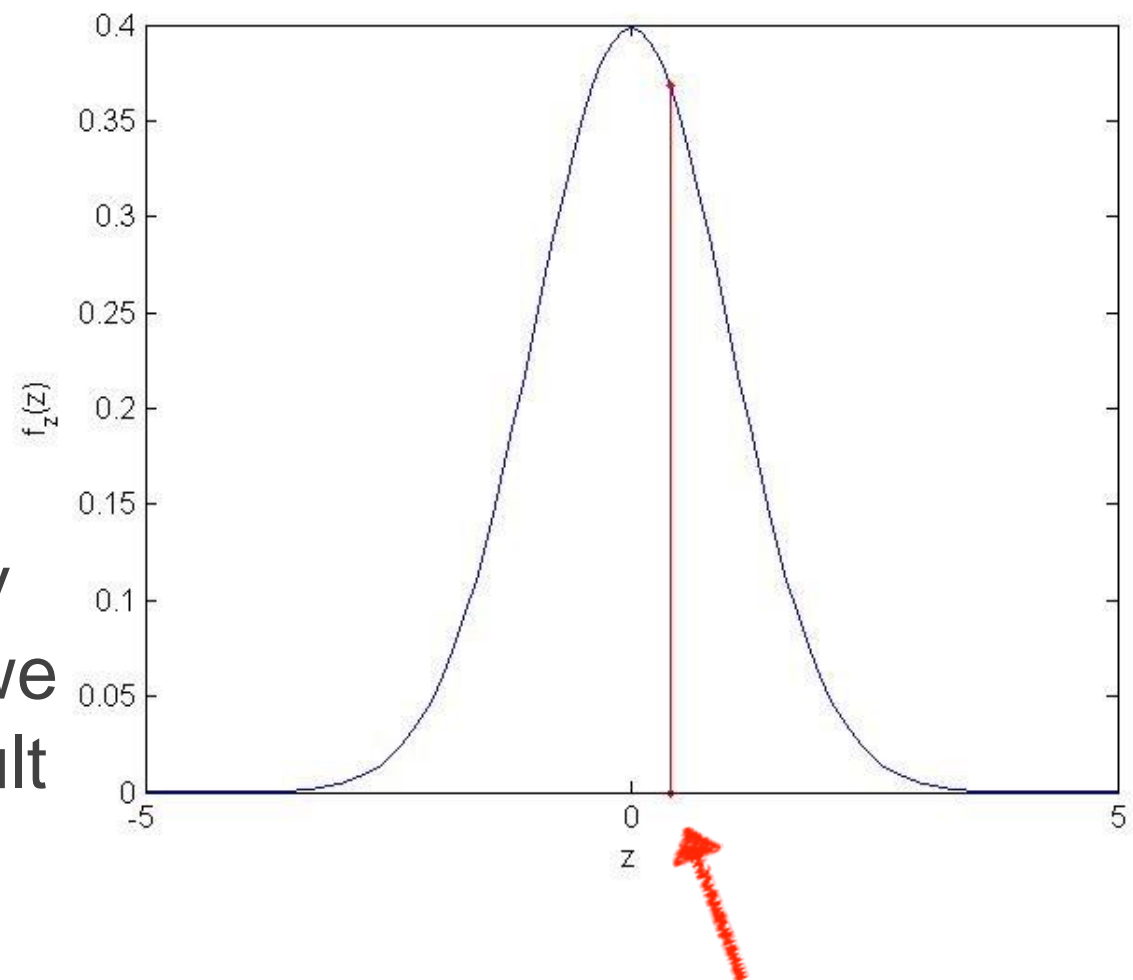
We compare the  $p$ -value with  $\alpha$ :

Same as asking: what is the probability of observing a test size ( $z$ ) that is more extreme than 0.4?

$$p = \Pr(Z \leq -0.4 \cup Z \geq 0.4) = 0.6892 > 0.05 = \alpha$$

- If the cup-filling machine is adequately calibrated with  $\mu=250$  g and  $\sigma=2.5$  g, we will in 68.9% of the time get a test-result more extreme than the observed one.
- So the test result is very plausible:  
→ We can't reject that the cup-filling machine is adequately calibrated!

Standard normal distribution (PDF)



$$z = \frac{\bar{x} - \mu}{\sigma/\sqrt{n}} = \frac{250.2 - 250}{2.5/\sqrt{25}} = 0.4$$

---

# Confidence Level *Key term*

---

- **Definition 12 – Confidence level**

- The confidence level is the complement of the significance level ( $\alpha$ ):



$$\text{confidence level} = 1 - \alpha$$

*Now, how confident are we in the estimate of the mean?*

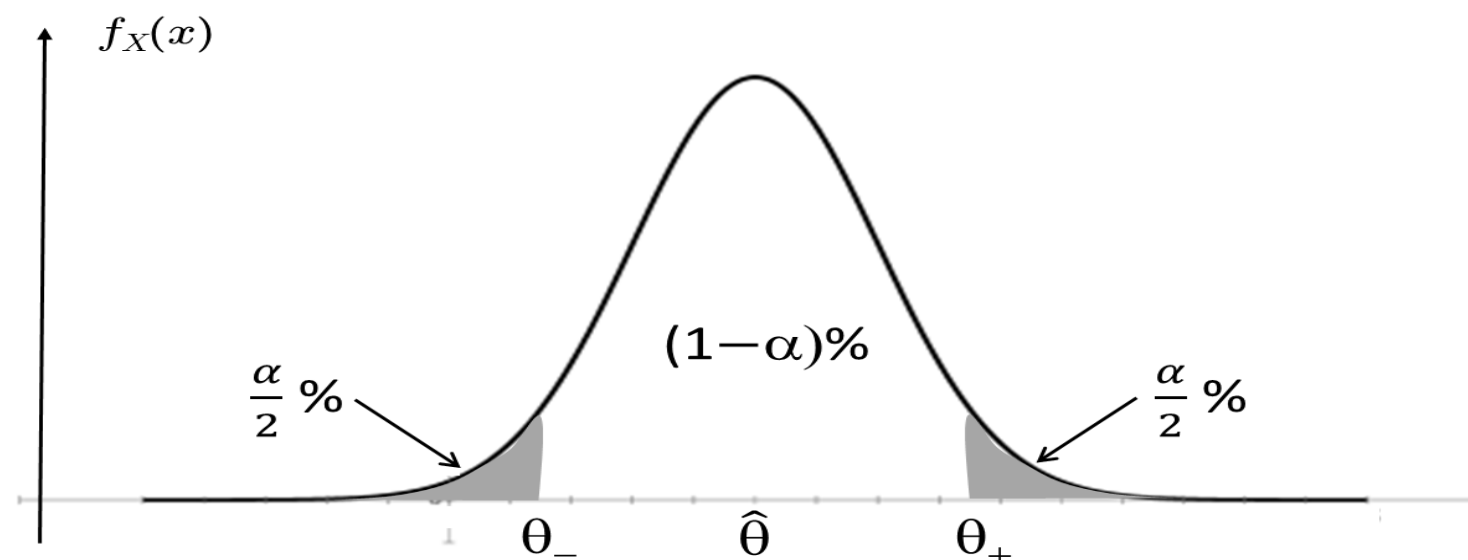
*Significance level:  $\alpha = 0,05 = 5\% \rightarrow 0,95 = 95\%$  confidence level*

# Confidence Interval <sup>Key term</sup>

- **Definition 13 – Confidence interval**

- The  $1 - \alpha$  confidence interval is an interval  $[\theta_-; \theta_+]$  such that the probability that the true value of the unknown parameter,  $\theta$ , lies within the interval is  $1 - \alpha$ :

❖ 
$$\Pr(\theta_- \leq \theta \leq \theta_+) = 1 - \alpha$$



# Confidence Interval

## Standard Normal distribution:

- Significance level:  $\alpha = 0.05 \rightarrow$  Confidence level =  $1 - 0.05 = 0.95 = 95\%$
- We should find numbers  $-z$  and  $z$ , between which  $Z$  lies with 95% probability:

$$\Pr(-z < Z < z) = 0.95 \Rightarrow \Pr(Z < z) = F(z) = \Phi(z) = 1 - \frac{\alpha}{2} = 0.975$$

$$\Rightarrow z_+ = -z_- = \Phi^{-1}(0.975) = 1.96$$

*Norminv(0.975) in Matlab*

The **Confidence Interval for  $X$**  can now be found:

$$\bullet \quad 0.95 = \Pr(-1.96 < Z < 1.96) = \Pr(-1.96 < \frac{\bar{X} - \mu}{\sigma/\sqrt{n}} < 1.96)$$

$$= \Pr\left(\bar{x} - 1.96 \frac{\sigma}{\sqrt{n}} < \mu < \bar{x} + 1.96 \frac{\sigma}{\sqrt{n}}\right)$$

*Lower endpoint  $\mu_-$*

*Upper endpoint  $\mu_+$*



# Confidence Interval

- In the cup-filling example, the 95% confidence interval is

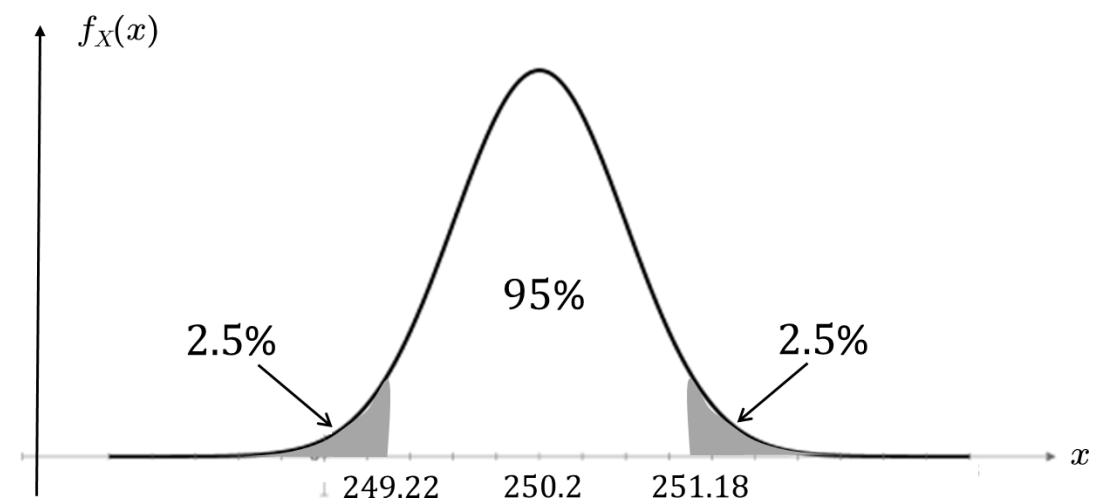
$$[\mu_-; \mu_+] = \left[ \bar{x} - 1.96 \cdot \frac{\sigma}{\sqrt{n}}; \bar{x} + 1.96 \cdot \frac{\sigma}{\sqrt{n}} \right]$$

❖

$$= \left[ 250.2 - 1.96 \cdot \frac{2.5}{\sqrt{25}}; 250.2 + 1.96 \cdot \frac{2.5}{\sqrt{25}} \right]$$

$$= [250.2 - 0.98; 250.2 + 0.98]$$

$$= [249.22; 251.18]$$



---

# Confidence Interval

---

- Correct interpretation:
  - Every time the experiment is repeated, there will be a new estimate of the mean (i.e., we will observe another value for the sample mean  $\bar{x}$ ).
  - ❖ – This changes the endpoints of the 95% confidence interval.
  - In 95% of the cases, the true mean  $\mu$  will lie between the endpoints calculated from  $\bar{x}$ , but in 5% of the cases it will not.

---

# Sample Size Determination

---

- ❖ Larger sample sizes generally lead to **increased precision** when estimating unknown parameters.
- ❖ **For example**, if we wish to know the effect of a medical treatment, we would generally have a more accurate estimate of this effect if we sampled and examined 200 rather than 100 patients.
- ❖ However, if sufficient statistical power can be obtained using just 100 patients, we prefer that instead of 200 patients because it reduces the **economic costs of the experiment**.
- ❖ **Sample sizes are judged based on the quality of the resulting estimates.**
- ❖ **For example:** If estimating a mean, one may wish to have the 95% confidence interval be less than 0.1 units wide.

# Sample Size Determination

*How big a sample size do you need?*

- Recall that the 95% confidence interval of the mean estimator ( $\hat{\mu}$ ) is

$$\bar{x} \pm 1.96 \cdot \sigma / \sqrt{n}$$

- The general form of the 95% confidence interval is  $\bar{x} \pm B$
- Suppose we wish to make an estimate of the mean of a population, where the 95% confidence interval is less than  $B$  units wide.
- Then, assuming that the population standard deviation is known, we require that

$$1.96 \cdot \sigma / \sqrt{n} \leq B$$

*B is the desired uncertainty (95% confidence) on  $\hat{\mu}$*

- Isolating the sample size,  $n$ , in this equation results in

$$n \geq \left( \frac{1.96 \cdot \sigma}{B} \right)^2$$

*If  $B = \sigma$  then  $n \geq 4$   
If  $B = 0.1\sigma$  then  $n \geq 385$*

# Cup Example

- We want to estimate the average filling of the cups with a confidence (maximum uncertainty) of  $\pm B$  g.
- How many cups are needed to be tested (size of test-sample)?
- $B = 2g \Rightarrow n \geq \left(\frac{1.96 \cdot \sigma}{B}\right)^2 = \left(\frac{1.96 \cdot 2.5}{2}\right)^2 = 6$
- $B = 1g \Rightarrow n \geq \left(\frac{1.96 \cdot \sigma}{B}\right)^2 = \left(\frac{1.96 \cdot 2.5}{1}\right)^2 = 24$
- $B = 0,5g \Rightarrow n \geq \left(\frac{1.96 \cdot \sigma}{B}\right)^2 = \left(\frac{1.96 \cdot 2.5}{0.5}\right)^2 = 96$
- $B = 0,1g \Rightarrow n \geq \left(\frac{1.96 \cdot \sigma}{B}\right)^2 = \left(\frac{1.96 \cdot 2.5}{0.1}\right)^2 = 2401$

---

# Words and Concepts to Know

---

Descriptive statistics	Statistic	Estimator
		Biased/Unbiased
Population	Central Limit Theorem	
		Sample variance
Two-sided	Significance Level	
		Inferential statistics
Statistical model	Test statistics	
		Confidence Level
Maximum-likelihood	True mean	
		Sample mean
	One-sided	
Standard Normal Distribution		z-statistic
Sample Size	p-value	Confidence Interval
	Sample	