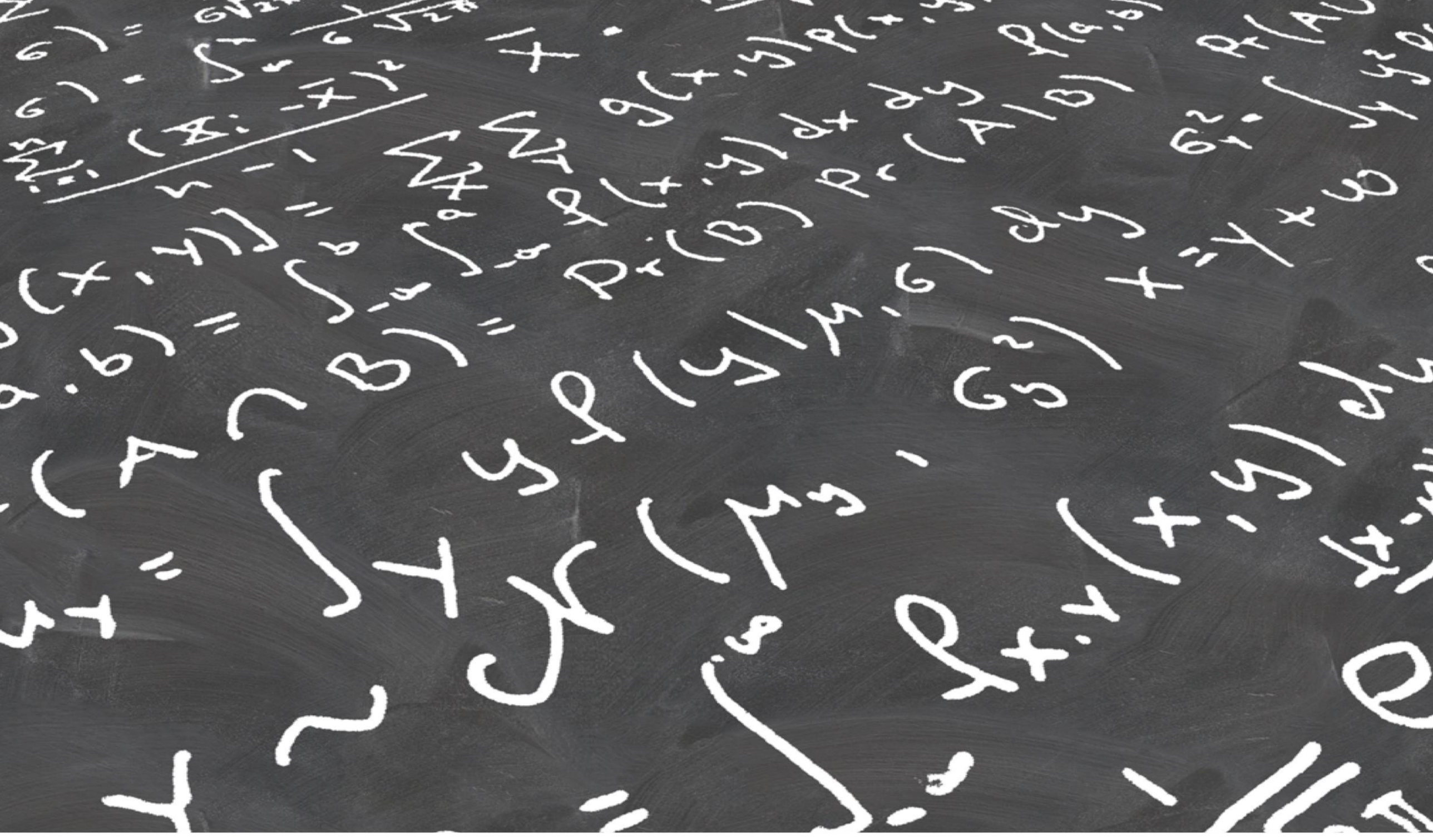


9. Introduction to Statistics

Gunvor Elisabeth Kirkelund
Lars Mandrup

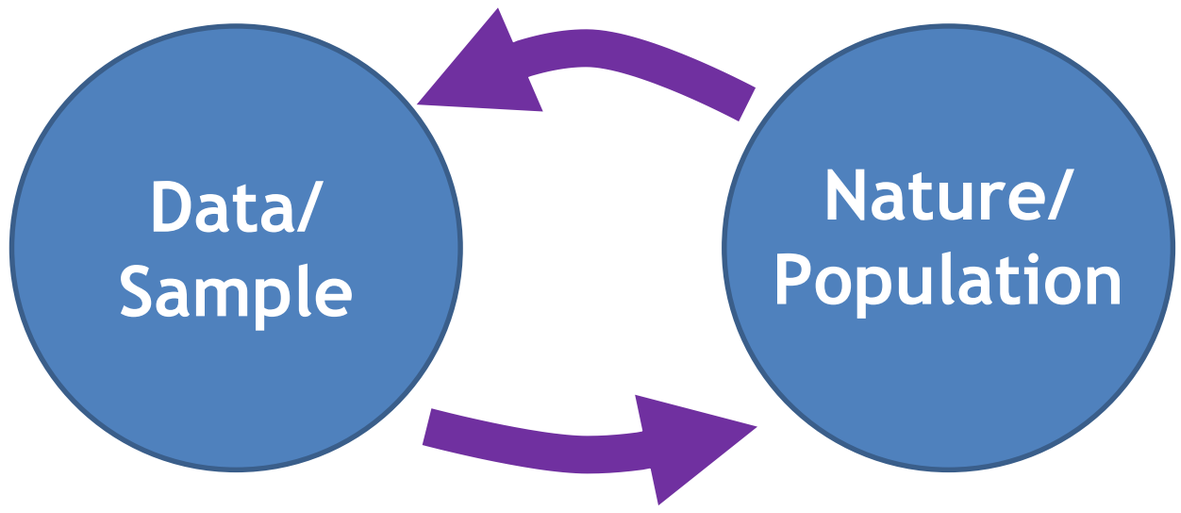
Todays Content

- 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)?

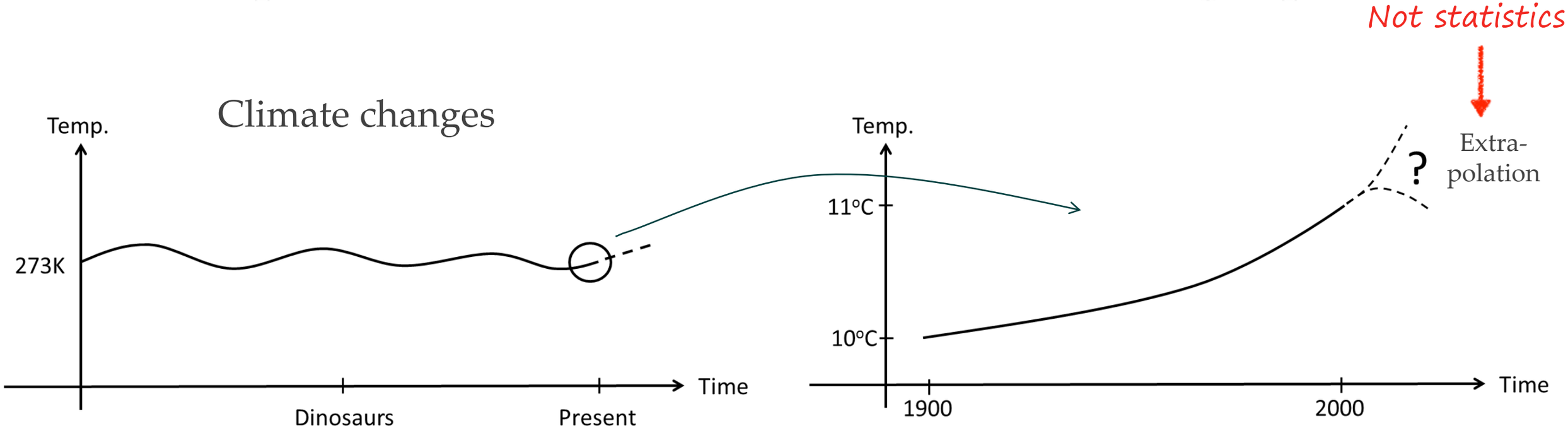
- Testing a hypotesis
- Estimating means and variances

- 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.

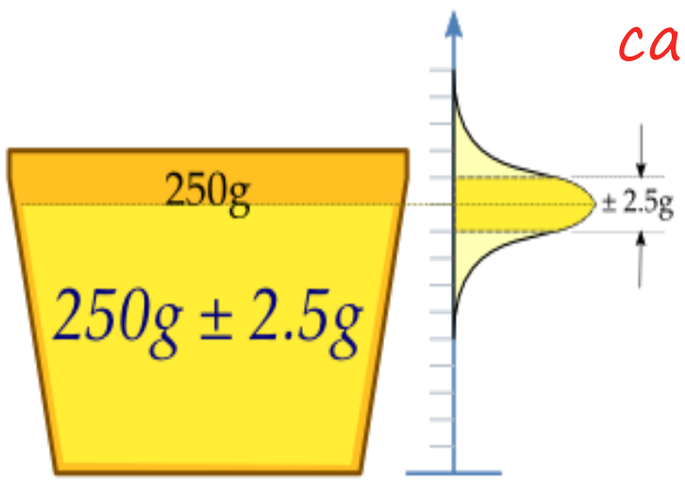


Inferential Statistics

- *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 $X_1, X_2, \ldots, X_{25}$, a random sample from X.

Population — All cups for all times

- To get an impression of the population mean (μ), we use the average (or sample mean) as an estimate:

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

^ means an estimator of the true polulation value →

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

- 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.
- 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.

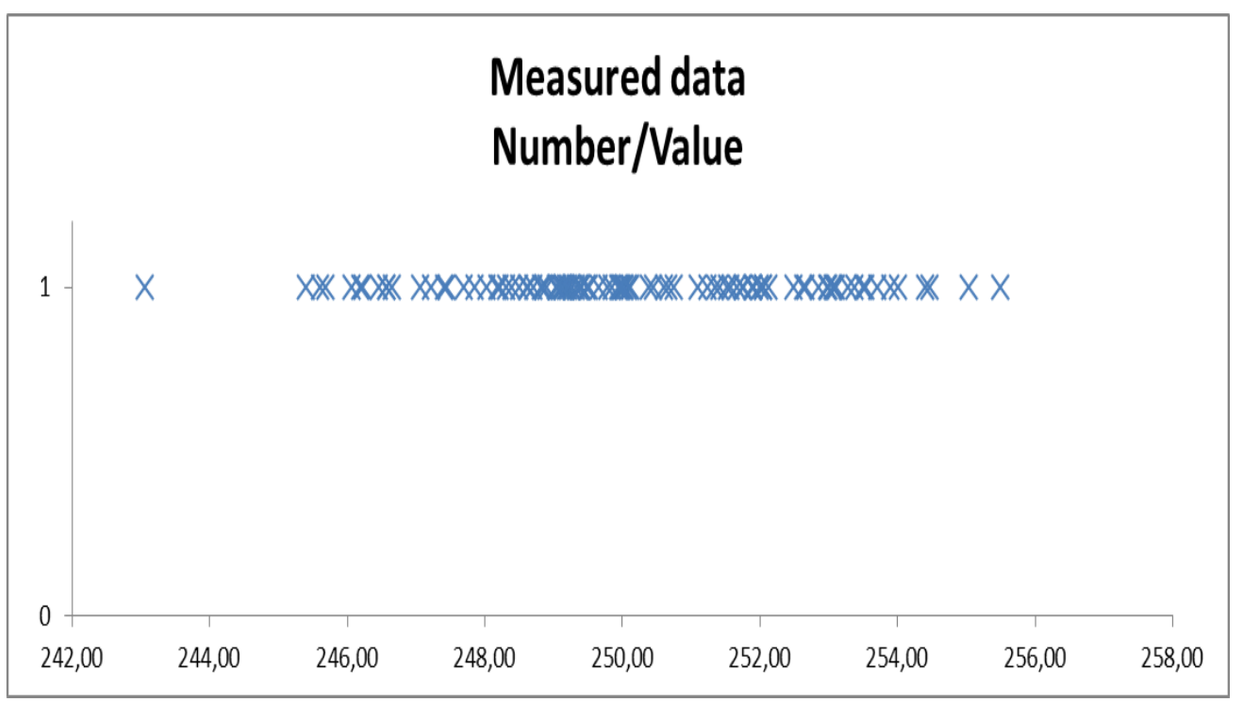
100 samples of the population – fx. in Matlab

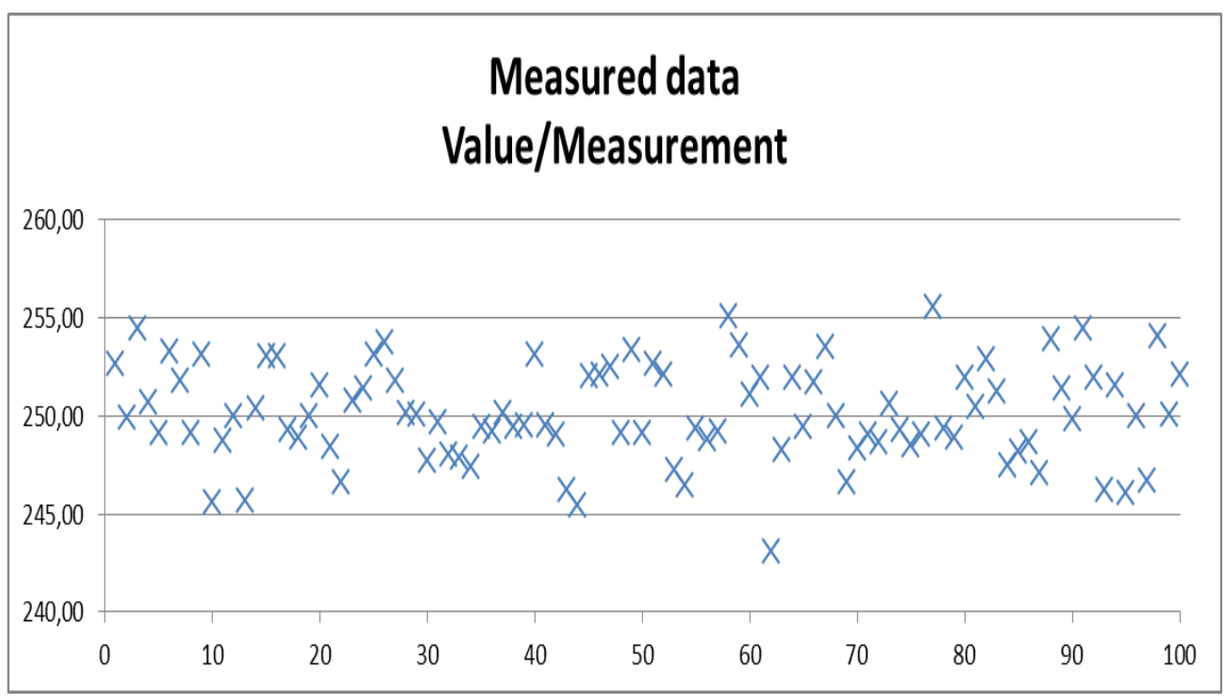
Think of the Central Limit theorem!!

Histograms

- Presentation of data

252,6	248,7	248,4	249,7	249,5	252,7	251,9	249,1	250,5	254,4
249,9	250,0	246,6	248,0	249,0	252,1	243,1	248,6	252,9	251,9
254,5	245,7	250,8	247,9	246,2	247,2	248,2	250,6	251,2	246,2
250,7	250,4	251,4	247,4	245,4	246,5	251,9	249,3	247,4	251,5
249,1	253,0	253,1	249,4	252,0	249,4	249,4	248,5	248,2	246,1
253,3	253,0	253,7	249,2	252,1	248,8	251,7	249,1	248,6	249,9
251,8	249,3	251,8	250,1	252,5	249,2	253,5	255,5	247,1	246,7
249,1	248,9	250,1	249,4	249,1	255,0	250,0	249,3	253,9	254,0
253,1	250,0	250,0	249,5	253,4	253,5	246,6	248,9	251,3	250,0
245,6	251,6	247,7	253,1	249,1	251,1	248,3	251,9	249,8	252,1

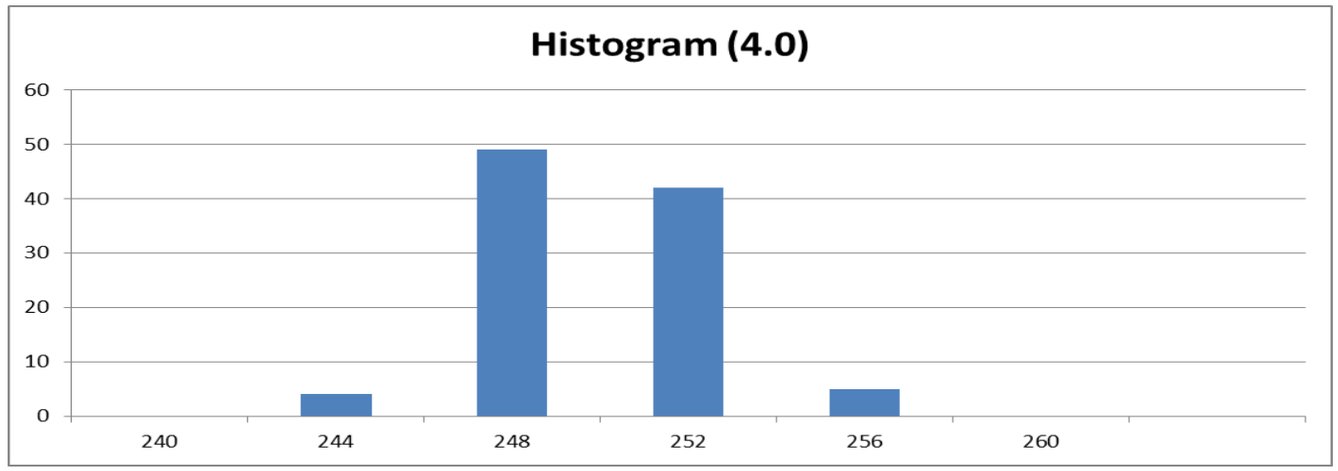


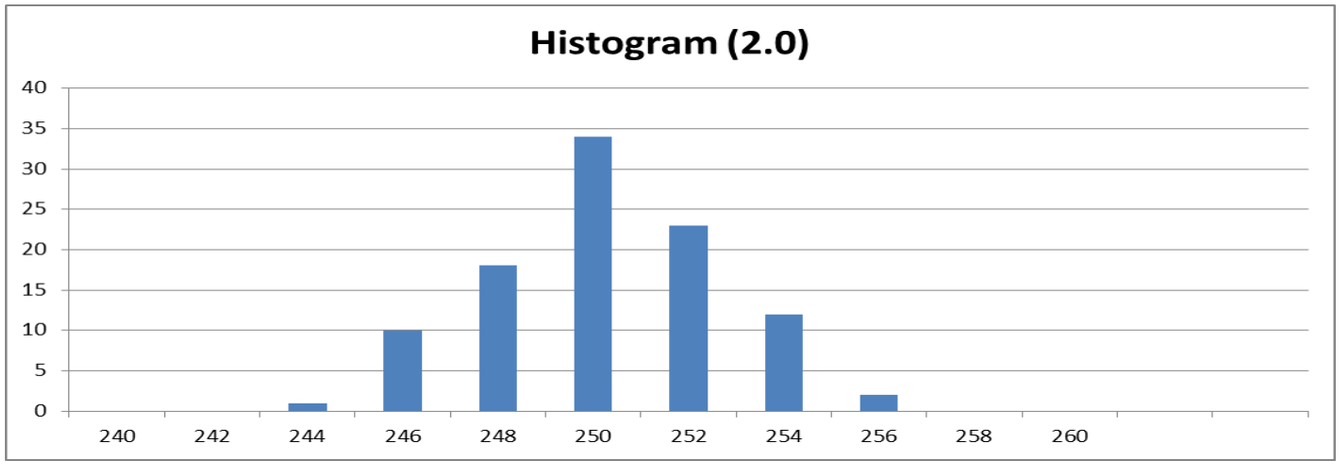


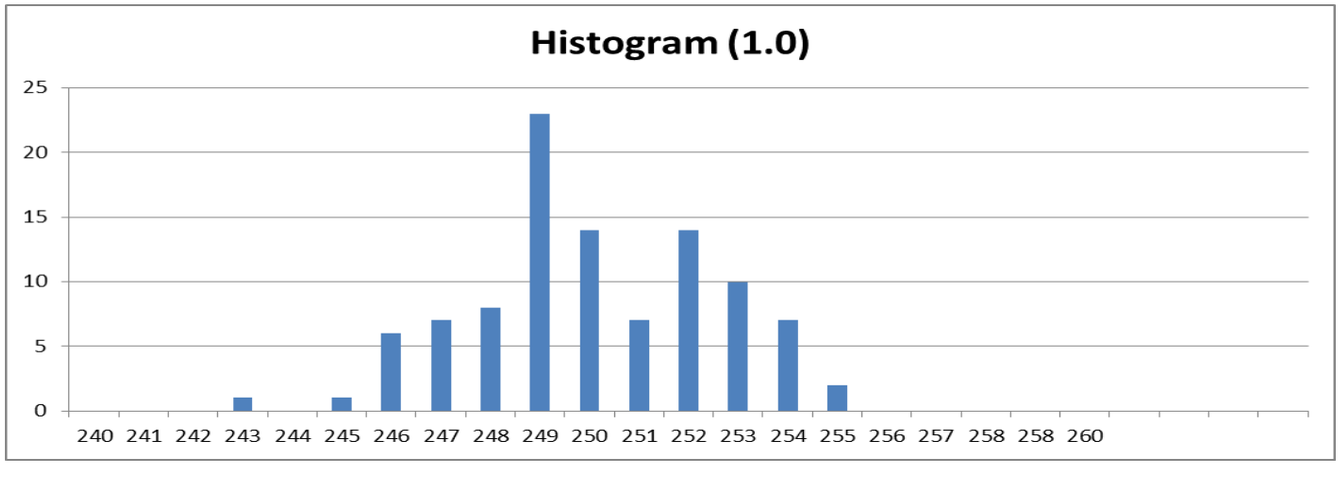
Histograms

- Grouping of data

Intervaller	Midtpunkt	Antal
239-241	240	0
241-243	242	0
243-245	244	1
245-247	246	10
247-249	248	18
249-251	250	34
251-253	252	23
253-255	254	12
255-257	256	2
257-259	258	0
259-261	260	0

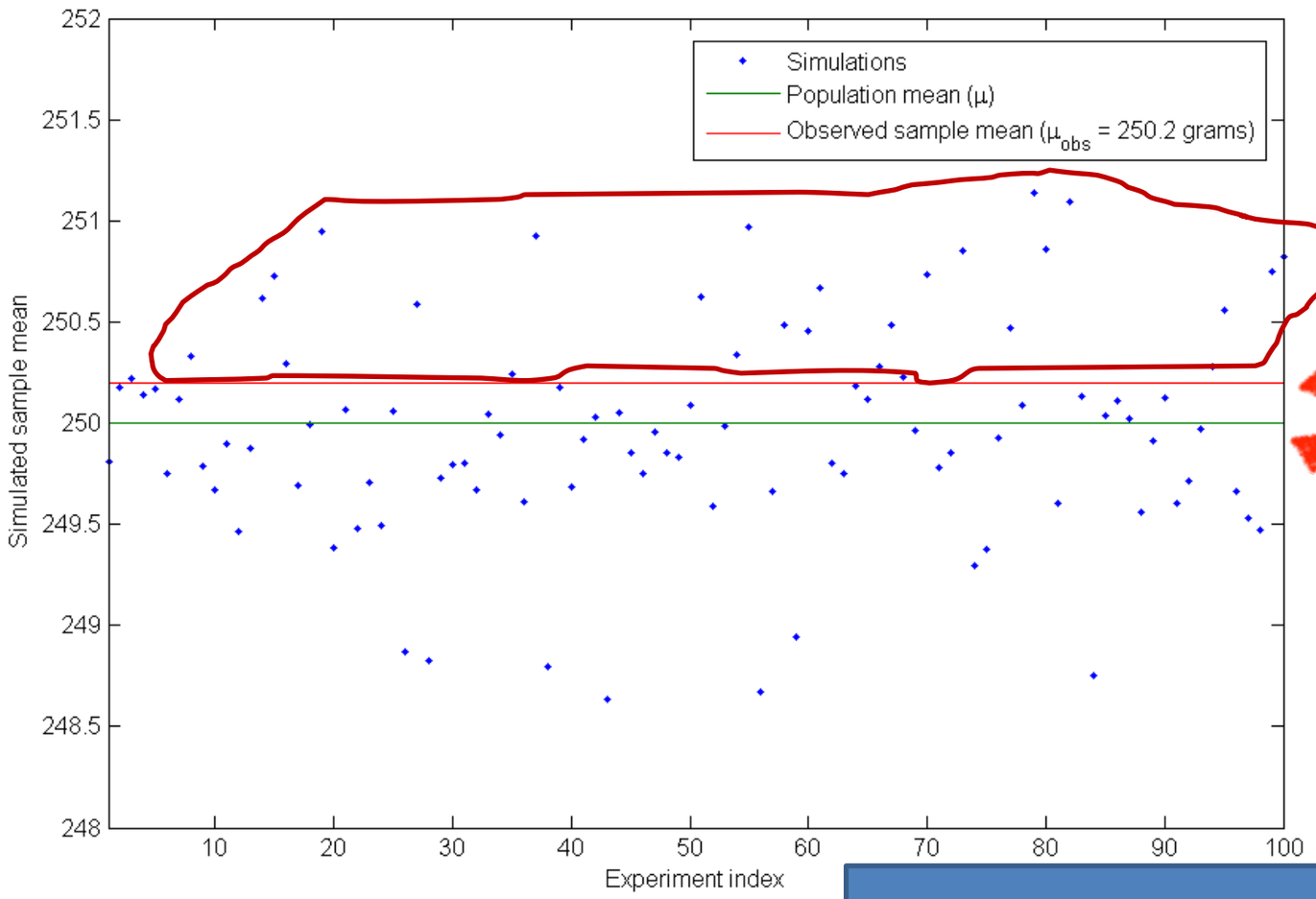






Cup Example

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



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

Observed value

Theoretical value

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 28% 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 be explained by random variation in X, and **we conclude that the machine is adequately calibrated**.

it is likely that

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 $X_1, X_2, \ldots, X_n$, 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 θ is the parameter of the PDF.

- The parameter, θ, is in general a vector and often unknown.
- If $f_X(x;\theta)$ is a normal distribution with mean μ and variance σ^2, then $\theta = [\mu, \sigma^2]$.
- The purpose of inferential statistics is to infer knowledge about the unknown parameter(s), θ.

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, θ.

- A commonly used statistic is the average or sample mean

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

Estimator

Key term

- **Definition 5 – Estimator**
  - An estimator, $\hat{\theta}(X)$, is a statistic used to estimate the unknown parameter θ of a random sample, X.
- For notational convenience, we will often write $\hat{\theta}$ instead of $\theta(X)$.
- 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, μ, 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 $f_X(x \mid \hat{\theta})$ the likelihood function is:

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

- 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 Estimate of the Mean

- Given independent observations, $x_1, x_2, \ldots, x_n$, the likelihood function given the true variance (σ^2) and some choice of mean ($\hat{\mu}$) is

Likelihood

i.i.d. data

Multiplication

$$L(\hat{\mu}) = f(x_1, x_2, \ldots, x_n | \hat{\mu}, \sigma^2) = f(x_1) \cdot f(x_2) \cdot \cdots \cdot f(x_n) = \prod_{i=1}^{n} f(x_i)$$

- where

$$f(x_i) = \frac{1}{\sqrt{2\pi\sigma^2}} \cdot e^{-(x_i - \hat{\mu})^2/\sigma^2}$$

in Gaussian data

- Inserting, we get

$$L(\hat{\mu}) = \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi\sigma^2}} \cdot e^{-(x_i - \hat{\mu})^2/\sigma^2}$$

Maximum-Likelihood Estimate of the Mean

i.i.d. Gaussian data

- We have,

$$L(\hat{\mu}) = \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi\sigma^2}} \cdot e^{-(x_i-\hat{\mu})^2/\sigma^2}$$

- Ignoring constants and taking the logarithm, we see that the optimum choice of $\hat{\mu}$ must maximize

$$-\sum_{i=1}^{n}(x_i - \hat{\mu})^2$$ *Called: the log-likelihood*

- Differentiating with respect to $\hat{\mu}$ and setting equal to zero, we get the desired result

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

Maximum-likelihood Estimator for Binomial Data

- For the binomial distribution the likelihood function is:

$$L(\hat{p}) = f_{bino}(x|\hat{p}) = \binom{n}{x} \hat{p}^x (1-\hat{p})^{n-x}$$

where $f_{bino}(x|\hat{p})$ denotes the binomial pdf, and x is the observed number of successes out of n trials.

- When maximizing, we can ignore the binomial coefficient $\binom{n}{x}$, because it does not depend on $\hat{p}$.
- Since the logarithm is a monoton function, the choise of $\hat{p}$ that maximizes $L(\hat{p})$ also maximizes $\log(L(\hat{p}))$
- So, the optimum $\hat{p}$ maximizes: $\log(\hat{p}^x(1-\hat{p})^{n-x}) = x \cdot \log(\hat{p}) + (n-x) \cdot \log(1-\hat{p})$
- Differentiate with $\hat{p}$ and setting equal to zero: $\frac{x}{\hat{p}} - \frac{n-x}{1-\hat{p}} = 0 \;\Rightarrow\; \hat{p} = \frac{x}{n}$
- Hence, the parameter estimate from the n trials, is exactly the maximum-likelihood estimate.

Unbiased Estimator

Key term

- **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, μ, of a normally distributed random variable, X.

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

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

The Sample Variance

Key term

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

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

- This is a biased estimator

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

- 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*.

Central Limit Theorem

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

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

- 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 (θ).
- Example (z-score):

expected value/true mean

sample mean

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

Standard normal distribution (μ=0 and σ^2=1)

standard deviation, sqrt(var(x))

n – number in the sample

Test Statistics – Cup example

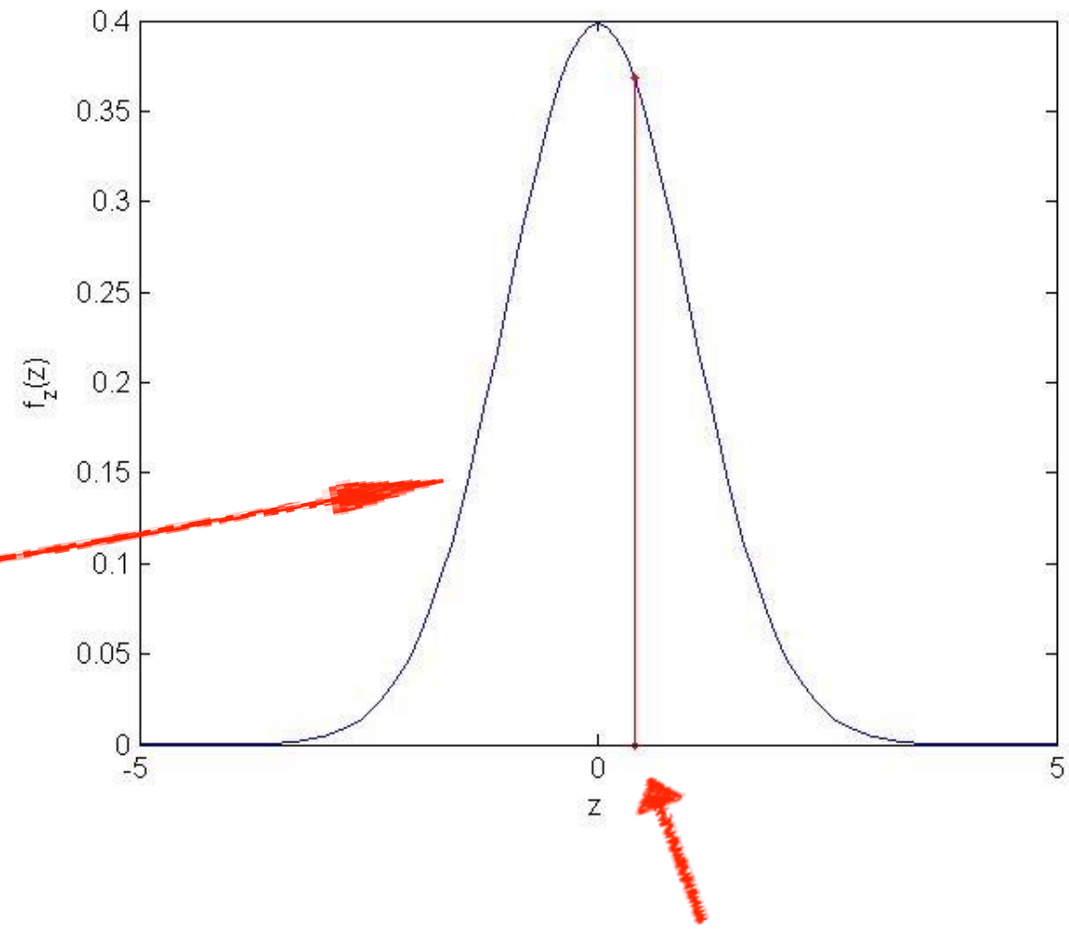
We define a test statistics z:

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

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

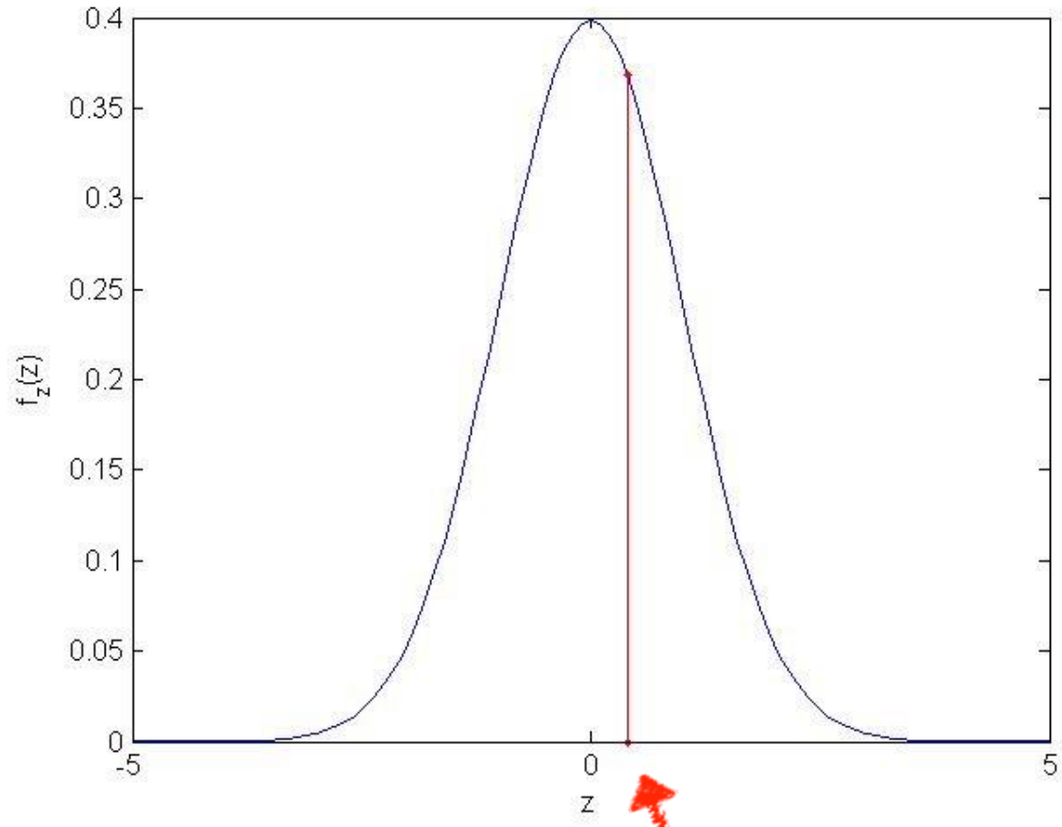
Plausible result?

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?

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?

We compare z with another value:

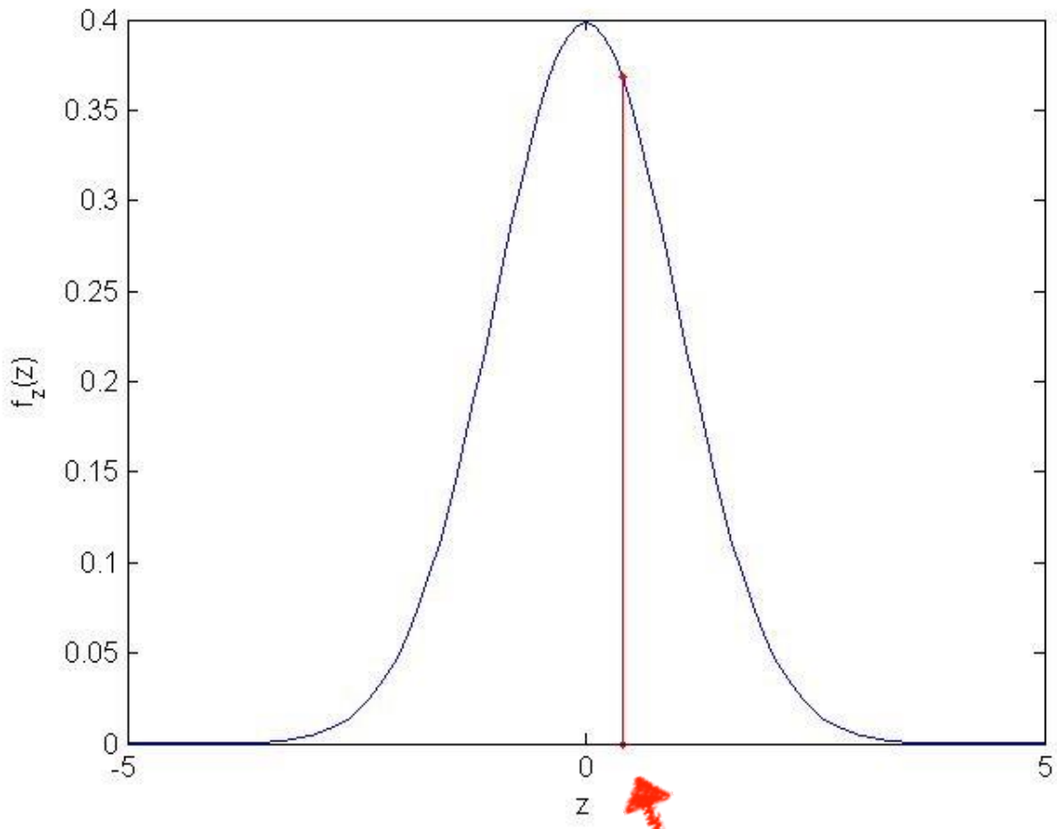
Same as asking: what is the probability of observing a test size (z) that is more extreme than 0.4?

p-value:

$$Pr(Z > z \cup Z < -z)$$
$$= \Pr(Z > z) + \Pr(Z < -z)$$
$$= 1 - \Pr(Z \leq z) + 1 - \Pr(Z \leq z)$$
$$= 2 \cdot (1 - \Pr(Z \leq z))$$
$$= 2 \cdot (1 - \Pr(Z \leq 0.4))$$
$$= 2 \cdot \big(1 - \Phi(0.4)\big) \leftarrow \texttt{normcdf(0.4)}$$
$$= 2 \cdot (1 - 0.6554)$$
$$= 0.6892$$

Compare with the α-value!

Standard normal distribution (PDF)



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

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

Significance Level

Key term

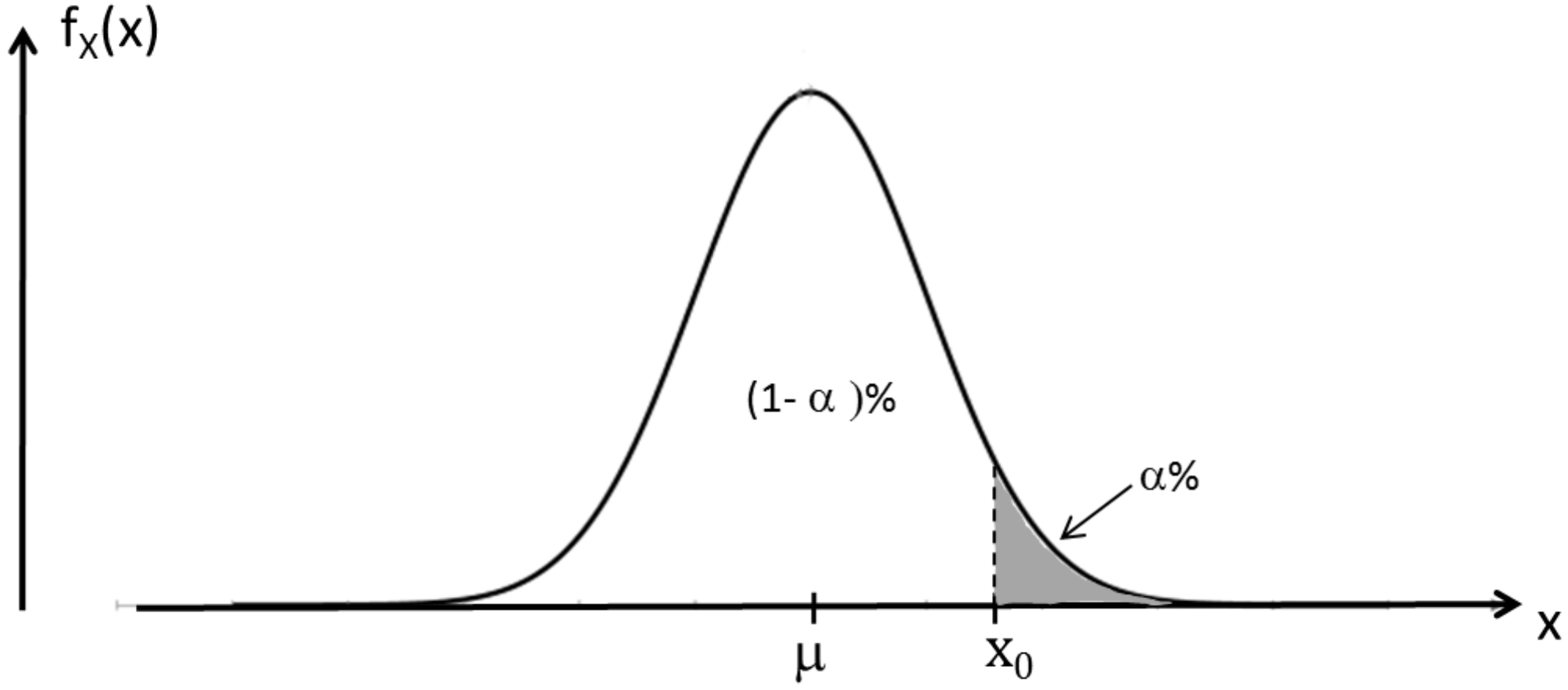
- Definition 9 – Significance level
  - The statistical significance level α is the lower limit we will accept for the probability of getting a more extreme result under the assumption of a given distribution.
  - The most common used significance level is $\alpha = 0{,}05$ (5%)

- By comparing the p-value for the test with the significance level α we can deside whether the test-result is plausible with the assumed distribution or not.

Significance Level

- If the test is <u>one-sided</u> (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:

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



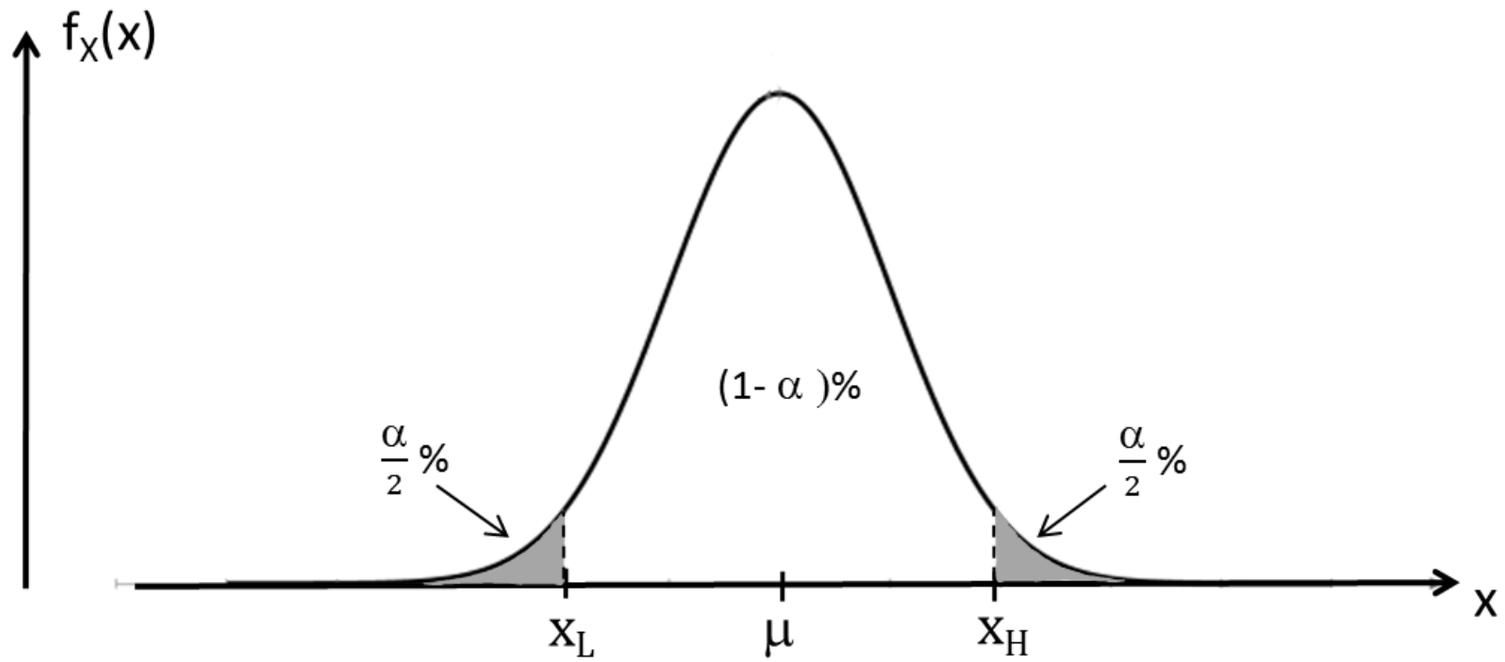
Significance Level

- If the test is <u>two-sided</u> (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:

$$\Pr(X < x_L \cup X > x_H) = \Pr(X < x_L) + \Pr(X > x_H) = \alpha$$

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

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



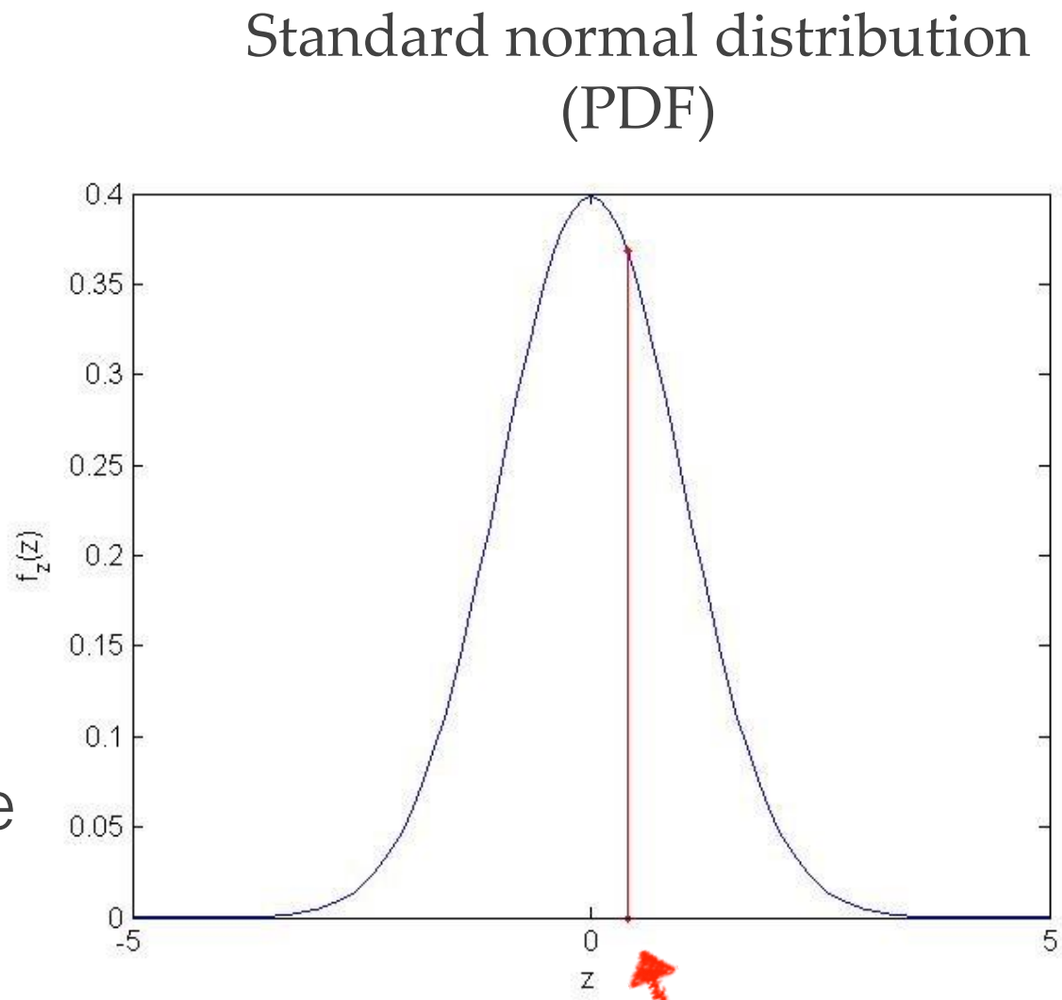
Plausible result?

Yes!

We compare the p-value with α :

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

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



- If the cup-filling machine is adequately calibrated with μ=250 g and σ=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!

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

Confidence Level

Key term

- **Definition 12 – Confidence level**
  - The confidence level is the complement of the significance level (α):

$$confidence\ level = 1 - \alpha$$

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

Confidence Interval

Key term

- **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, θ, lies within the interval is $1-\alpha$:

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

Confidence Interval

- **Example 6 – cup filling example**
  - Recall that the observed test statistic was

$$z = \frac{\bar{x} - \mu}{\sigma/\sqrt{n}} = \frac{250.2 - 250}{2.5/\sqrt{25}} = 0.4 \sim N(0,1)$$

- With a significance level of $\alpha = 0.05$, the confidence level is 1-0.05 = 0.95.
- Now, it is possible to find numbers –z and z, between which Z lies with 95% probability. So we have

$$\Pr(-z \leq Z \leq z) = 0.95$$

Confidence Interval

- The number *z* follows from the CDF function, in this case the CDF of a standard normal distribution:

$\Phi(z) = F(z)$ *for the standard normal distribution* $\mathcal{N}(0,1)$

$$\Phi(z) = \Pr(Z \leq z) = 1 - \frac{\alpha}{2} = 1 - 0.025 = 0.975$$

- We have to divide α by 2, because we are considering a two-sided test.
- The value of *z* satisfying the above relation is *z* = 1.96, i.e., $\Phi(1.96) = 0.975$.
- The way to find this *z* value is according to

$$z = \Phi^{-1}\big(\Phi(z)\big) = \Phi^{-1}(0.975) = 1.96$$

- which can be done either by an inverse table lookup in the probability table of a standard normal distribution or by using the Matlab command, norminv(0.975).

Confidence Interval

- By insertion we get

$$0.95 = \Pr(-1.96 \leq Z \leq 1.96) = \Pr\left(-1.96 \leq \frac{\bar{x} - \mu}{\sigma/\sqrt{n}} \leq 1.96\right)$$
$$= \Pr(\bar{x} - 1.96 \cdot \sigma/\sqrt{n} \leq \mu \leq \bar{x} + 1.96 \cdot \sigma/\sqrt{n})$$

- The lower endpoint of the 95% confidence interval is:

$$lower\ endpoint = \ \mu_{-} = \bar{x} - 1.96 \cdot \sigma/\sqrt{n}$$

- and the upper endpoint of the 95% confidence interval is:

$$upper\ endpoint = \mu_{+} = \ \bar{x} + 1.96 \cdot \sigma/\sqrt{n}$$

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 μ 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

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

B is the desired uncertaincy (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=σ then n≥4
If B=0.1σ then n≥385

Cup Example

- We want to estimate the average filling of the cups with a confidence (maximum uncertaincy) of ±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

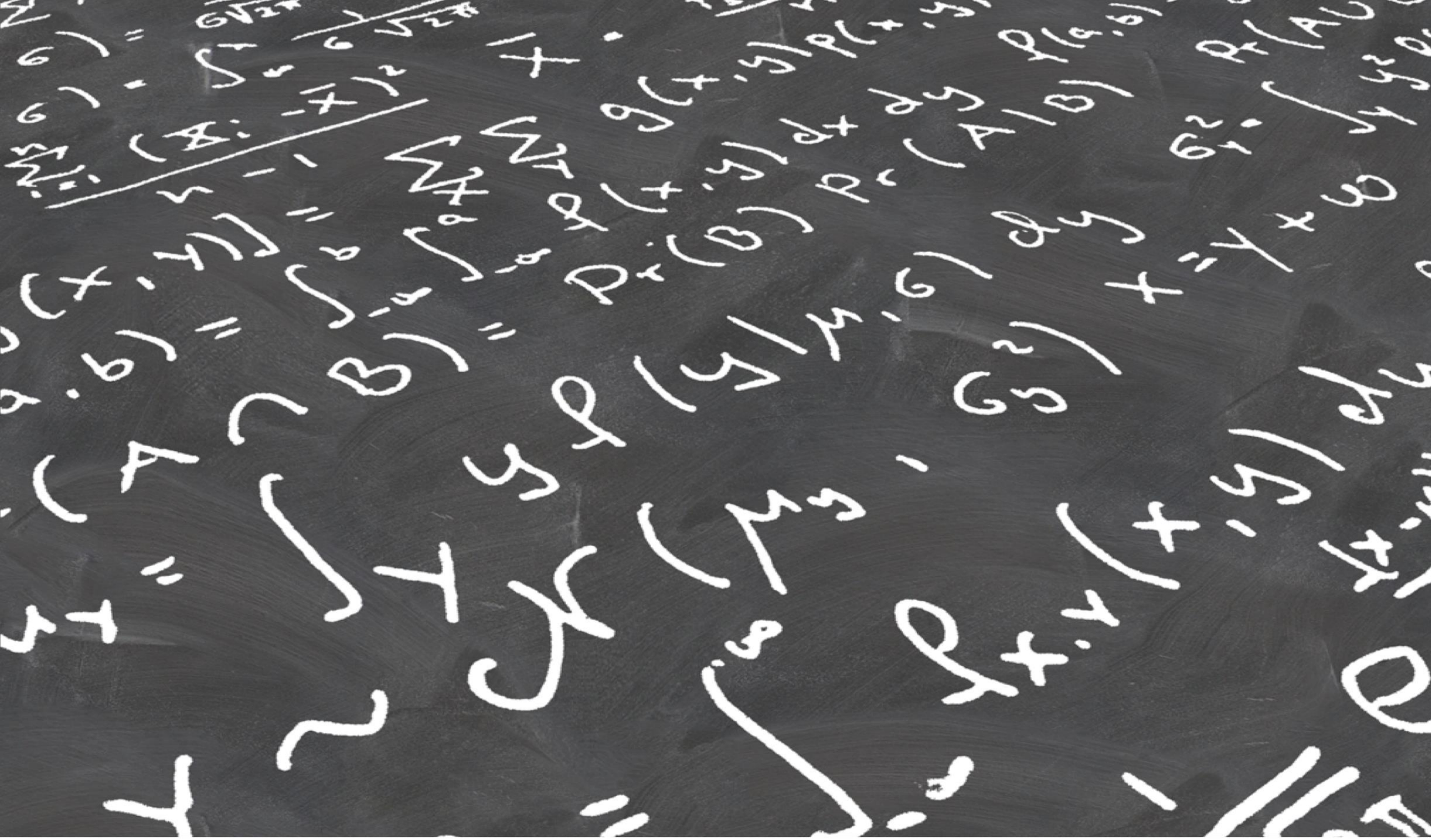
z-statistic

Sample Size

p-value

Sample

Confidence Interval



10. Hypothesis Test

Gunvor Elisabeth Kirkelund
Lars Mandrup
Slides and material provided in parts by
Henrik Pedersen

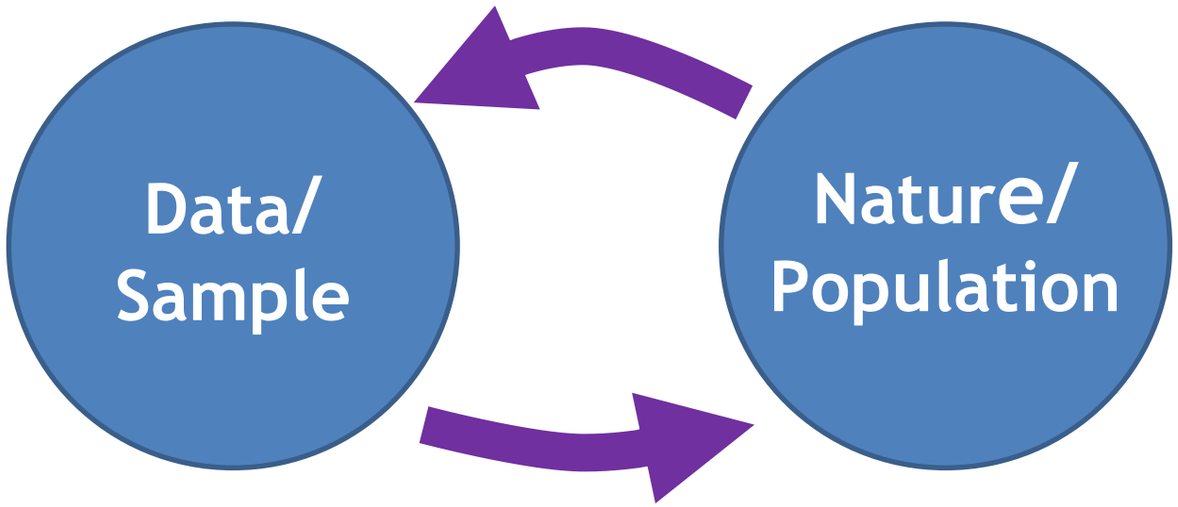
Todays Content

- Repetition from last time
- Hypothesis Test
- p-values
- Test of the mean with known variance (z-test)
- Test of the mean with unknown variance (t-test)

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 hypotesis
- Estimating means and variances

- If we don't know better: We assume data are normally distributed

Estimator

Estimator:

- An estimator $\hat{\theta}(X)$ is a statistic used to estimate the unknown parameter θ of a random sample X.
- An estimator is unbiased if $E[\hat{\theta}] = \theta$.

Unbiased estimators:

- The sample mean: $\hat{\mu} = \bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$

Degrees of freedom

- The sample variance: $\hat{s}^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2$

Degrees of freedom

Statistical Model

Statistical model:

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

- Because of the Central Limit Theorem (CLT) we often can use the normal distribution $\mathcal{N}(\mu,\sigma^2)$ with mean μ and variance σ^2 as statistical model for the sample mean $\bar{X}$

$$\bar{X} = \frac{1}{n}\sum_{i=1}^{n} X_i \sim \mathcal{N}(\mu,\sigma^2) \quad (\text{n}>30)$$

Test Statistics

Test statistics:

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

- For a sample assumed to follow the normal distribution $\mathcal{N}(\mu, \sigma^2)$ with known mean μ and variance σ^2 we can use the z-statistics (z-score):

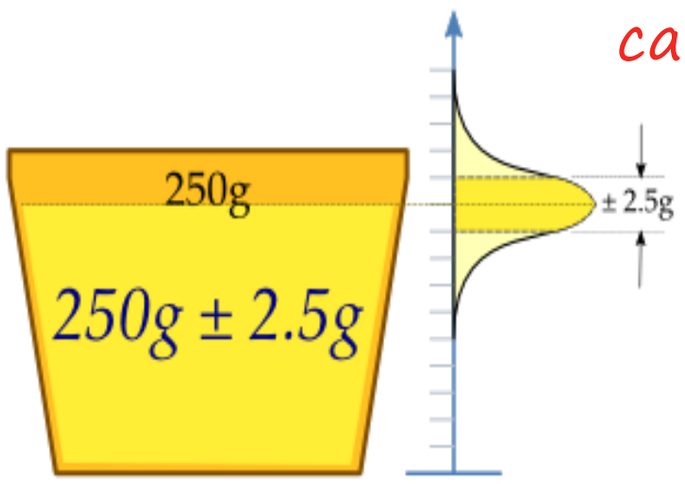
$$z = \frac{\bar{x} - \mu}{\sigma/\sqrt{n}} \sim \mathcal{N}(0,1)$$

Standard (normalized) normal distribution (μ=0 and σ^2=1)

➢ **Assignment:** Show that if $\bar{x} \sim \mathcal{N}(\mu, \sigma^2)$ then $z = \frac{\bar{x}-\mu}{\sigma/\sqrt{n}} \sim \mathcal{N}(0,1)$ (ie. having μ=0 and σ^2=1)

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 $X_1, X_2, \ldots, X_{25}$, a random sample from X.

Population — All cups for all times

- To get an impression of the population mean (μ), we use the average (or sample mean) as an estimate:

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

^ means an estimator of the true polulation value →

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

- Is the machine adequately calibrated?

Hypothesis Test

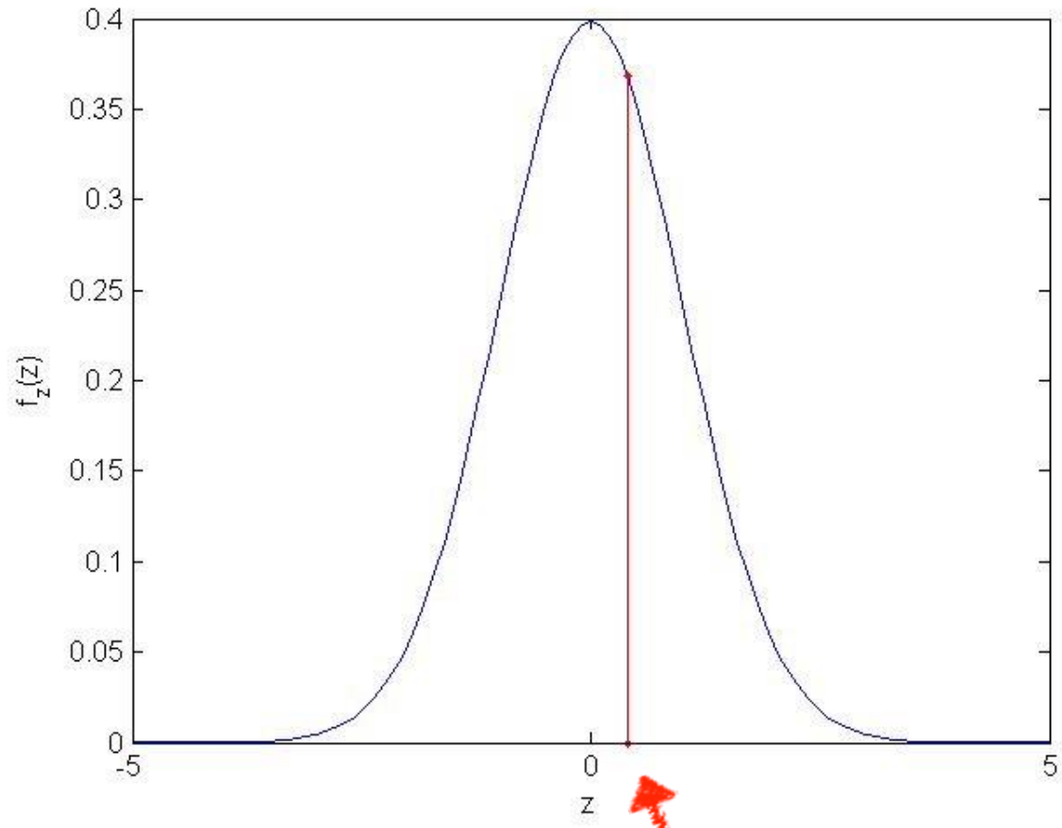
Since we know both the true mean μ and variance σ², we use the test statistics z

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

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?

Standard normal distribution (PDF)



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

Hypothesis

- Definition – Null hypothesis (H_0)
  - The statement being tested in a test of statistical significance is called the **null hypothesis**. The test of significance is designed to assess the strength of the evidence against the null hypothesis.
  - Usually, the null hypothesis is a statement of 'no effect', 'no difference' or 'no relation' between the phenomena whose relation is under investigation.
- Definition – Alternative hypothesis (H_1)
  - The statement that is hoped or expected to be true instead of the null hypothesis is the **alternative hypothesis**
  - The alternative hypothesis, as the name suggests, is the alternative to the null hypothesis: it states that there is some 'effect/difference' or some 'kind of relation'.

Important!

- One cannot "prove" a null hypothesis, one can only test how close it is to being true.
- Therefore, we never say that we *accept* the null hypothesis, but that we either ***reject it*** or ***fail to reject it***.

Hypothesis

An example of a null hypothesis:

- A certain drug may reduce the chance of having a heart attack.
- Possible null hypothesis H_0: “This drug has no effect on the chances of having a heart attack".
- An alternative hypothesis H_1: “This drug has an effect on the chances of having a heart attack”.
- The test of the hypothesis consists of giving the drug to half of the people in a study group as a controlled experiment.
- If the data show a statistically significant change in the people receiving the drug, the null hypothesis is rejected.

Hypothesis

- The term "**null hypothesis**" H_0 is a general statement or default position that there is no relationship between two measured phenomena, or no association among groups.
- Rejecting or disproving the null hypothesis is a central task in the modern practice of science; the field of statistics gives precise criteria for rejecting a null hypothesis.
- The null hypothesis H_0 is generally assumed to be true until evidence indicates otherwise.
- A null hypothesis is rejected if the observed data are significantly unlikely to have occurred if the null hypothesis were true. In this case an alternative hypothesis H_1 is accepted in its place – concluding that there are grounds for believing that there *is* a relationship between two phenomena.
- If the data are consistent with the null hypothesis, then the null hypothesis is not rejected (i.e., accepted).
- **In neither case is the null hypothesis or its alternative proven**; the null hypothesis is tested with data and a decision is made based on how **likely or unlikely** the data are. This is analogous to a criminal trial, in which the defendant is assumed to be innocent (null is not rejected) until proven guilty (null is rejected) beyond a reasonable doubt (to a statistically significant degree).

Hypothesis testing

- Hypothesis testing works by collecting a randomly selected representative sample X (data) and measuring how likely the particular set of data is, assuming the null hypothesis H_0 is true: $Pr(X|H_0)$
- The data-set is usually specified via a **test statistic** which is designed to measure the extent of apparent departure from the null hypothesis – fx. z-statistic or t-statistic.
- If the data-set of a randomly selected representative sample is very unlikely relative to the null hypothesis – i.e. only rarely (usually in less than either 5% or 1% (the significance level α)) will be observed – we rejects the null hypothesis concluding it (probably) is false.
- If the data do not contradict the null hypothesis, then only a weak conclusion can be made: namely, that the observed data set provides no strong evidence against the null hypothesis. In this case, because the null hypothesis could be true or false, it is interpreted as there is no evidence to support changing from a currently useful regime (the null hypothesis) to a different one.

Example of Hypothesis

- **Example 1 – cup filling example**
  - If the machine is adequately calibrated, the true population mean should be 250 grams. Hence, the null hypothesis is

$$H_0: \mu = 250$$

  - If we are not concerned about the direction of a possible deviation from $\mu = 250$, the alternative hypothesis is

$$H_1: \mu \neq 250$$

Significance Level

- Definition 9 – Significance level

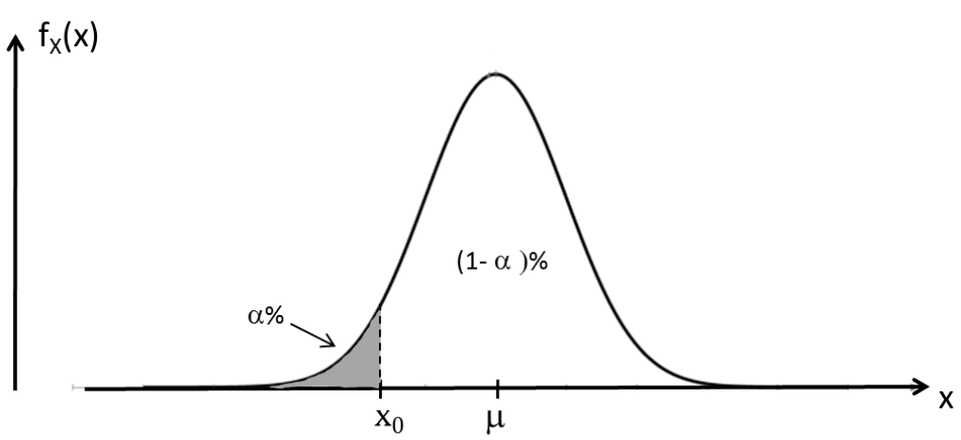
  - The statistical significance level α is the lower limit we will accept for the probability of getting a more extreme result assuming the null hypothesis H_0 is true.

  - The significance level is the cutoff level to reject the null hypothesis: If the probability of a randomly selected representative sample X under the assumption that the null hypothesis H_0 is true, is less than the significance level, we will reject the null hypothesis H_0.

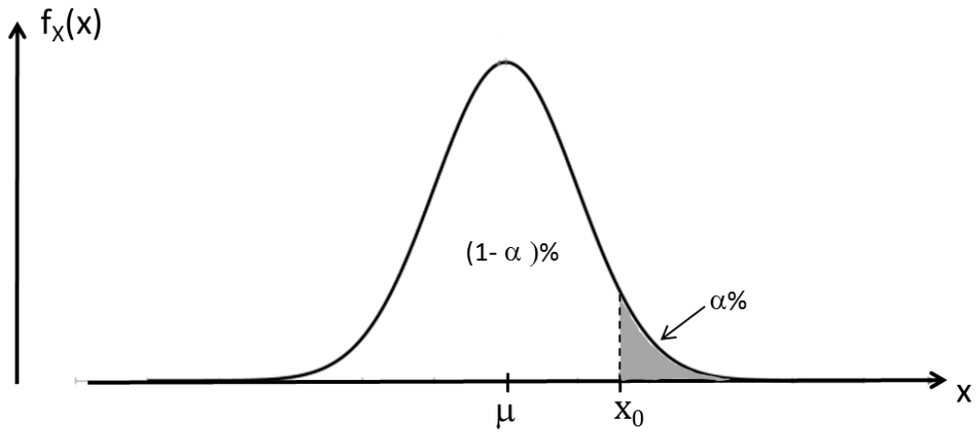
  - The most common used significance level is α = 0,05 (5%)

Significance Level

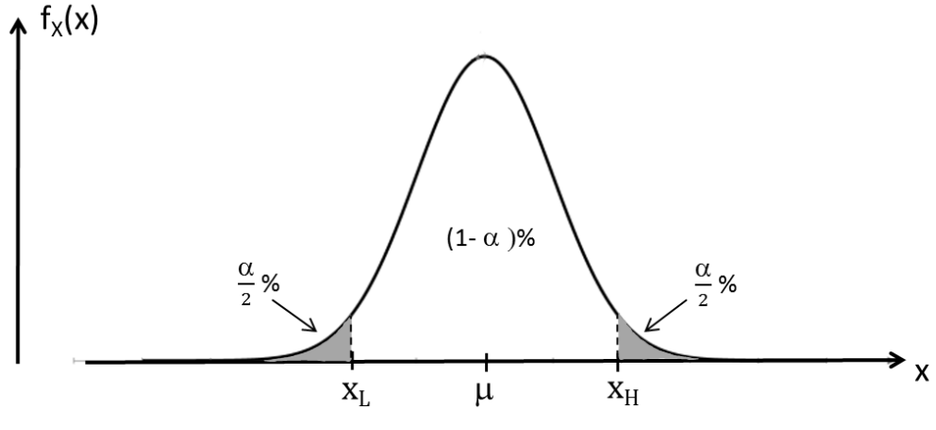
Left-tailed



Right-tailed



Two-tailed



p-value

The p value is found with matlab or in a table.

- Definition 10 – p-value
  - The p-value is the probability of getting a result equal to or more extreme than the observed test-sample $\boldsymbol{X}$ under the assumption of a null hypothesis H_0:

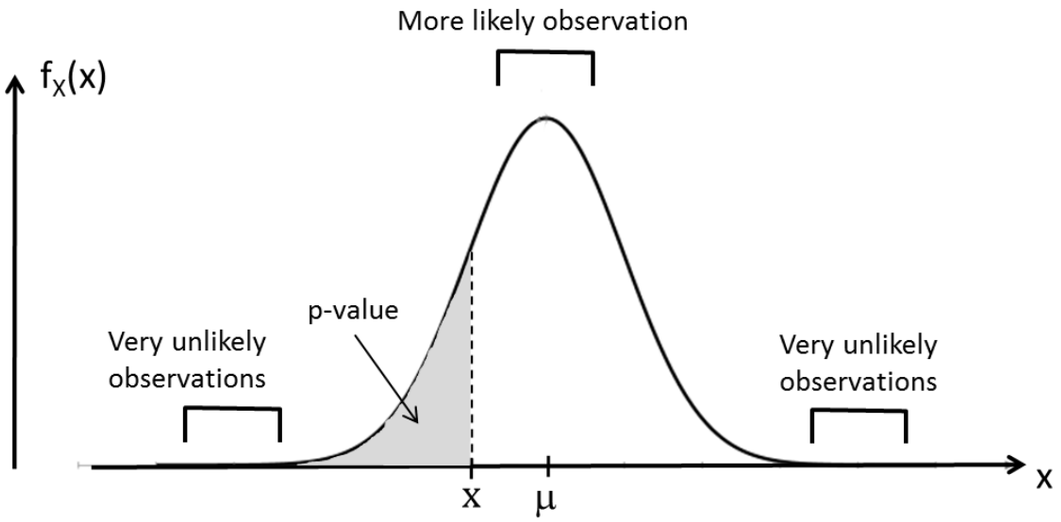
$$\boldsymbol{p - value = Pr(Worse\ result\ than\ X | H_0)}$$

- If $\boldsymbol{x}$ denotes the observed quantity, the p-value is:
  - $\boldsymbol{Pr(X \geq x | H_0)}$ for a right-tailed event
  - $\boldsymbol{Pr(X \leq x | H_0)}$ for a left-tailed event
  - $2 \cdot min\{\boldsymbol{Pr(X \leq x | H_0), Pr(X \geq x | H_0)}\}$ for a two-tailed event

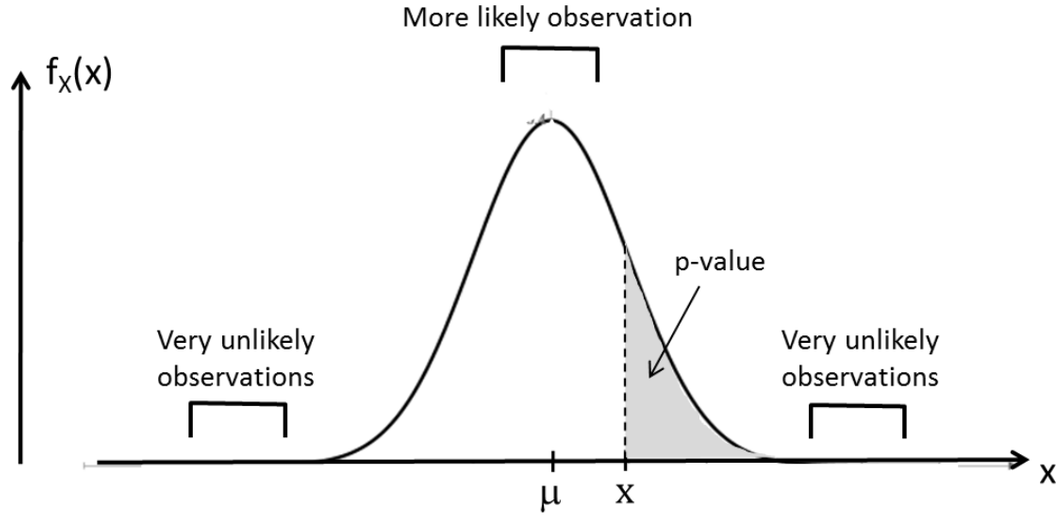
- By comparing the p-value for the test with the significance level α we can deside whether the null hypothesis H_0 should be rejected ($p < \alpha$) or not ($p > \alpha$).

p-value

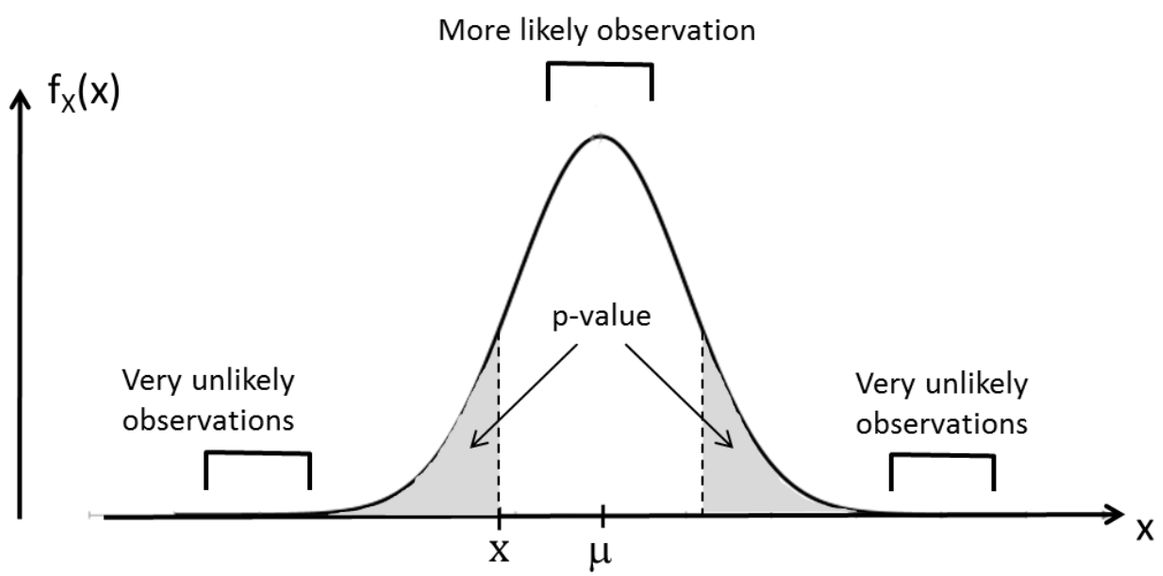
Left-tailed $\boldsymbol{p = Pr(X \leq x|H_0)}$



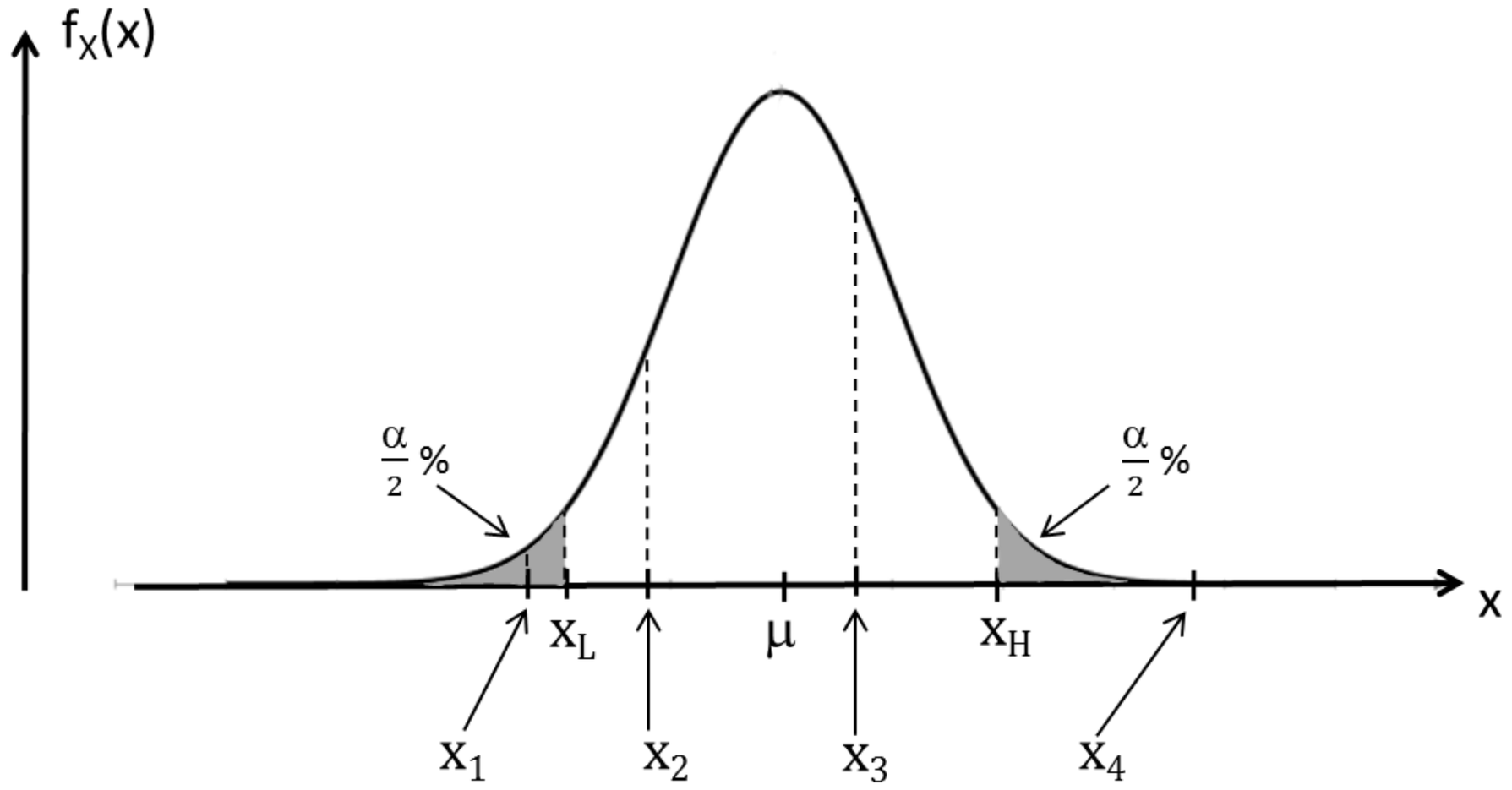
Right-tailed $\boldsymbol{p = Pr(X \geq x|H_0)}$



Two-tailed $p = 2 \cdot min\{\boldsymbol{Pr(X \leq x|H_0)}, \boldsymbol{Pr(X \geq x|H_0)}\}$



Hypothesis testing



Observations (test-samples): x_1 and x_4: Reject H_0
x_2 and x_3: Failed to reject H_0

Cup Example

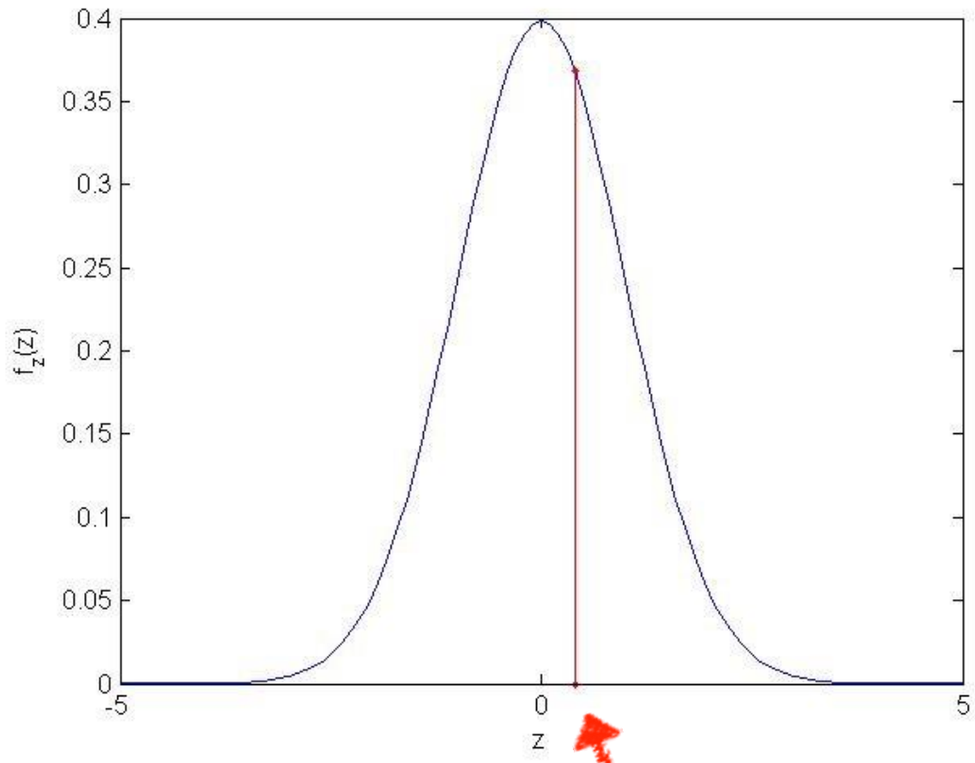
Here, we fail to reject the null hypothesis (H_0: $\mu = 250$), because the p-value is larger than $\alpha = 0.05$.

p-value:

$$Pr(Z > z \cup Z < -z)$$
$$= \Pr(Z > z) + \Pr(Z < -z)$$
$$= 1 - \Pr(Z \leq z) + 1 - \Pr(Z \leq z)$$
$$= 2 \cdot (1 - \Pr(Z \leq z))$$
$$= 2 \cdot (1 - \Pr(Z \leq 0.4))$$
$$= 2 \cdot \big(1 - \Phi(0.4)\big) \leftarrow \texttt{normcdf(0.4)}$$
$$= 2 \cdot (1 - 0.6554)$$
$$= 0.6892$$

Compare with $\alpha = 0.05$

Standard normal distribution (PDF)



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

Tests and Types of Errors

- There will sometimes be wrong conclusions in hypothesis testing
- We can classify the errors in hypothesis testing as:

Table of error types		Null hypothesis H_0	
		True	False
Hypothesis test result	Reject	Type I Error (False positive)	Correct inference (True positive)
	Fail to reject	Correct inference (True negative)	Type II Error (False negative)

- The Type I Error rate is the significance level α
- Decreasing the Type I Error rate (α) will increase the Type II Error rate

Also called the Gaussian Distribution

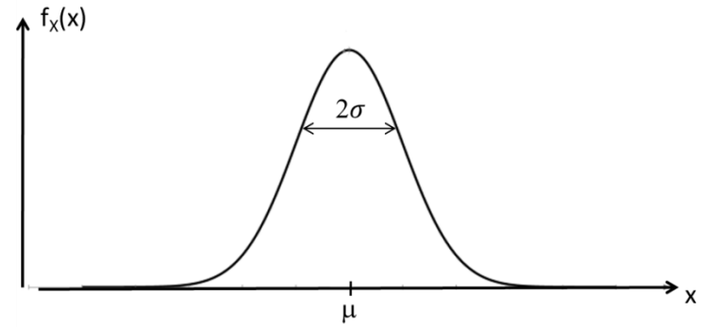
The Normal Distribution

- Let X be a normally distributed random variable with mean μ and variance σ^2:

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

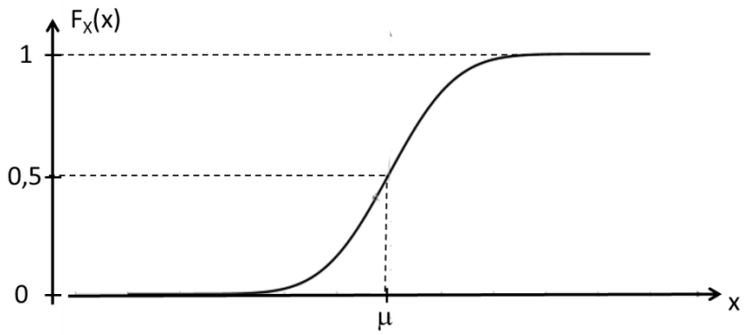
- Probability density function (pdf):

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



- There is no closed expression for the cdf.
  We use a lookup table for the standardized x:

$$z = \frac{x-\mu}{\sigma} \sim \mathcal{N}(0,1)$$



- The cdf is found as:

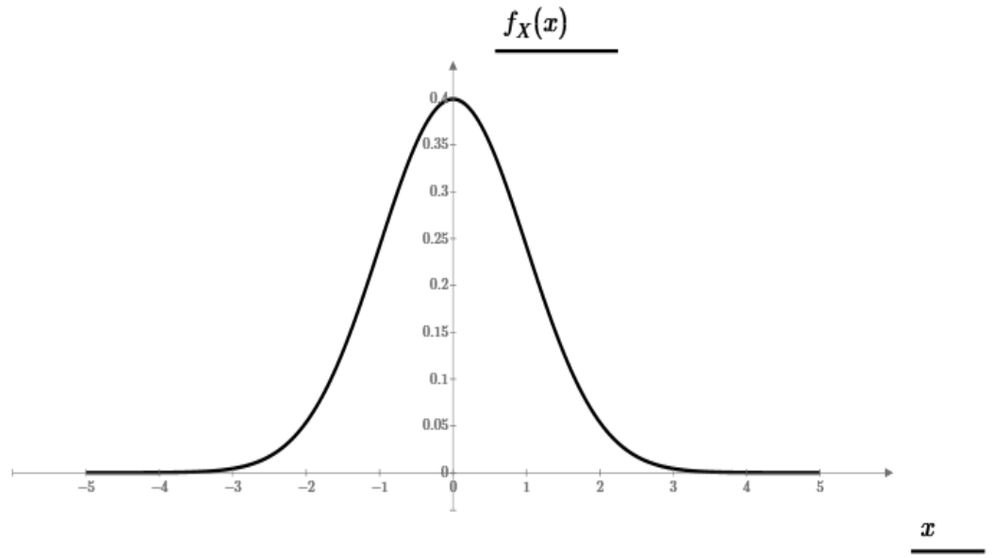
$$F_X(x) = Pr(X \leq x) = Pr\left(\frac{X-\mu}{\sigma} \leq \frac{x-\mu}{\sigma}\right) = \Phi(z) = 1 - Q(z)$$

- $\Phi(z)$ is the cdf of the standardized normal random variable Z used in most tabels
- $Q(z) = 1 - \Phi(z) = 1 - F_X(x) = Pr(X \geq x)$ is the tabelled quantity in app. D in ”Random Signals”

The Normal Distribution

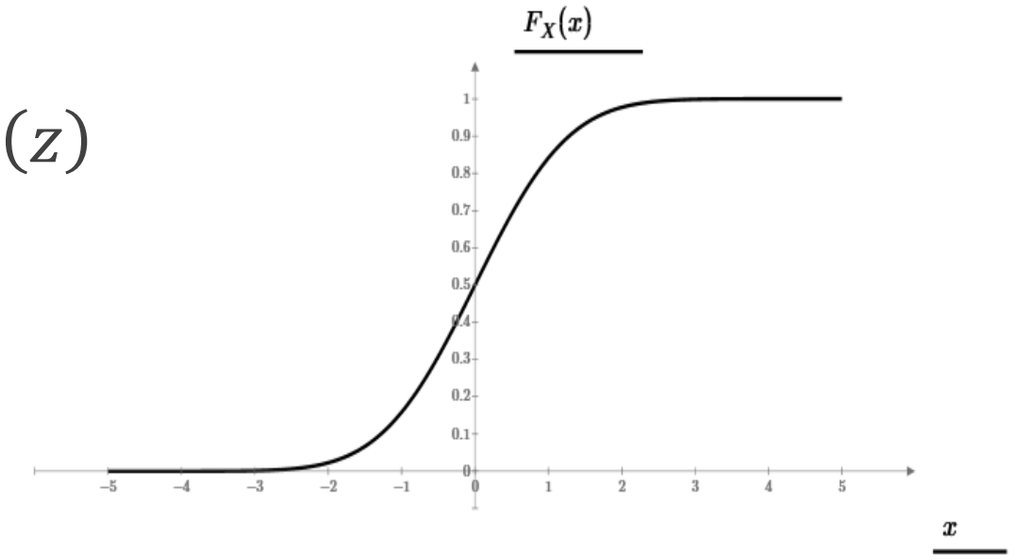
- Note that due to symmetry of the pdf:

$$Pr(Z \leq -z) = Pr(Z \geq z) = 1 - Pr(Z \leq z)$$



- Which imply that:

$$\Phi(-z) = 1 - \Phi(z) \quad \text{and} \quad \mathrm{Q}(-z) = 1 - \mathrm{Q}(z)$$



- Also due to symmetry:

$$Pr(Z \leq 0) = \Phi(0) = Q(0) = \frac{1}{2}$$

The Normal Distribution in Matlab

- Calculating probabilities of a normally distributed random variable

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

- In Matlab: $f_X(x) =$ `normpdf(x,μ,σ)` $= \sigma \cdot$`normpdf(x)`$+\mu$ *density function*

$$F_X(x) = \Pr(X \leq x) =$$ `normcdf(x,μ,σ)` *cumulative function*

- where μ is the mean and σ is the standard deviation ($\sigma = \sqrt{\sigma^2}$).

- If X is standard normally distributed $\sim \mathcal{N}(0,1)$ (ie. μ=0, σ=1), you can skip the arguments μ and σ:

`normpdf(x)= normpdf(x,0,1)`

`normcdf(x)= normcdf(x,0,1)= Φ(x)`

Simplified Calculation of the p-value from a z-Statistic

- Test size: $z \sim \mathcal{N}(0,1)$
- p-value (two-tailed event): $pval = 2 \cdot (1 - \Phi(|z|))$
  - If z is negative:

$$\begin{aligned} pval &= 2 \cdot min\{\boldsymbol{Pr}(\boldsymbol{Z} \leq \boldsymbol{z}), \boldsymbol{Pr}(\boldsymbol{Z} \geq \boldsymbol{z})\} \\ &= 2 \cdot \boldsymbol{Pr}(\boldsymbol{Z} \leq \boldsymbol{z}) \\ &= 2 \cdot \Phi(z) = 2 \cdot \big(1 - \Phi(-z)\big) = 2 \cdot (1 - \Phi(|z|)) \end{aligned}$$

  - If z is positive:

$$\begin{aligned} pval &= 2 \cdot min\{\boldsymbol{Pr}(\boldsymbol{Z} \leq \boldsymbol{z}), \boldsymbol{Pr}(\boldsymbol{Z} \geq \boldsymbol{z})\} \\ &= 2 \cdot \boldsymbol{Pr}(\boldsymbol{Z} \geq \boldsymbol{z}) = 2 \cdot (1 - \boldsymbol{Pr}(\boldsymbol{Z} \leq \boldsymbol{z})) \\ &= 2 \cdot \big(1 - \Phi(z)\big) = 2 \cdot (1 - \Phi(|z|)) \end{aligned}$$

Confidence Interval

- Conficence Interval
  - The $1-\alpha$ confidence interval is an interval $[\theta_{-};\theta_{+}]$ such that the probability that the true value of the unknown parameter θ lies within the interval is $1-\alpha$:

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

- For a two-tailed event with significance level $\alpha = 0{,}05$ the 95% confidence interval for the mean $[\mu_{-};\mu_{+}]$ is given by:

$$Pr(\mu_{-} \leq \mu \leq \mu_{+}) = 1-\alpha = 0{,}95$$

- where the interval endpoints are:

$$\Phi^{-1}(0{,}975)$$

$$\mu_{-} = \bar{x} - 1{,}96\frac{\sigma}{\sqrt{n}} \quad \text{and} \quad \mu_{+} = \bar{x} + 1{,}96\frac{\sigma}{\sqrt{n}}$$

Confidence Interval

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

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

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

$$= [250.2 - 0.98;\ 250.2 + 0.98] = [249.22;\ 251.18]$$

ie. μ=250 can't be rejected

TEST CATALOG FOR THE MEAN (KNOWN VARIANCE)

- **Statistical model:**
- $X_1, X_2, \ldots, X_n$ are i.i.d. samples of a random variable *X* with mean μ and variance σ^2.
- Parameter estimate:

$$\hat{\mu} = \bar{x} = \frac{1}{n}\sum_{i=1}^{n} X_i \sim N(\mu, \sigma^2/n)$$

- Where the observation is $\bar{x}$ = 'the average of *n* samples drawn from *X*'s distribution'.
- NOTE: The statistical model is only true if *n* is sufficiently large ($n \geq 30$) or if the samples are drawn from a normal population with mean μ and variance σ^2.

- **Hypothesis test (two-tailed):**
- $H_0: \mu = \mu_0$
- $H_1: \mu \neq \mu_0$
- Test size: $z = \frac{\bar{x} - \mu_0}{\sigma/\sqrt{n}} \sim N(0,1)$
- Approximate p-value: $2 \cdot |1 - \Phi(|z|)|$

- **95% confidence interval:**
- $\mu_{\_} = \bar{x} - 1.96 \cdot \sigma/\sqrt{n}$
- $\mu_{+} = \bar{x} + 1.96 \cdot \sigma/\sqrt{n}$

t-Score

When the variance is unknown!

- Now, consider the usual *z* statistic

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

- The equivalent statistic, when replacing the standard deviation (σ) with the empirical standard deviation (s), is called a *t*-score

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

Called the student t-distribution

- The *t*-score is *not* normally distributed; it is *t*-distributed with $v = n - 1$ degrees of freedom, which we write

$$t = \frac{\bar{x} - \mu}{s/\sqrt{n}} \sim t(n-1)$$

n-1: degrees of freedom

Student's t-distribution

- Students t-distribution: t(ν)=t(n-1):

  - pdf: $f_X(x) = \frac{\Gamma(\frac{\nu+1}{2})}{\sqrt{\nu\pi}\,\Gamma(\frac{\nu}{2})}(1+\frac{x^2}{\nu})^{-\frac{\nu+1}{2}}$

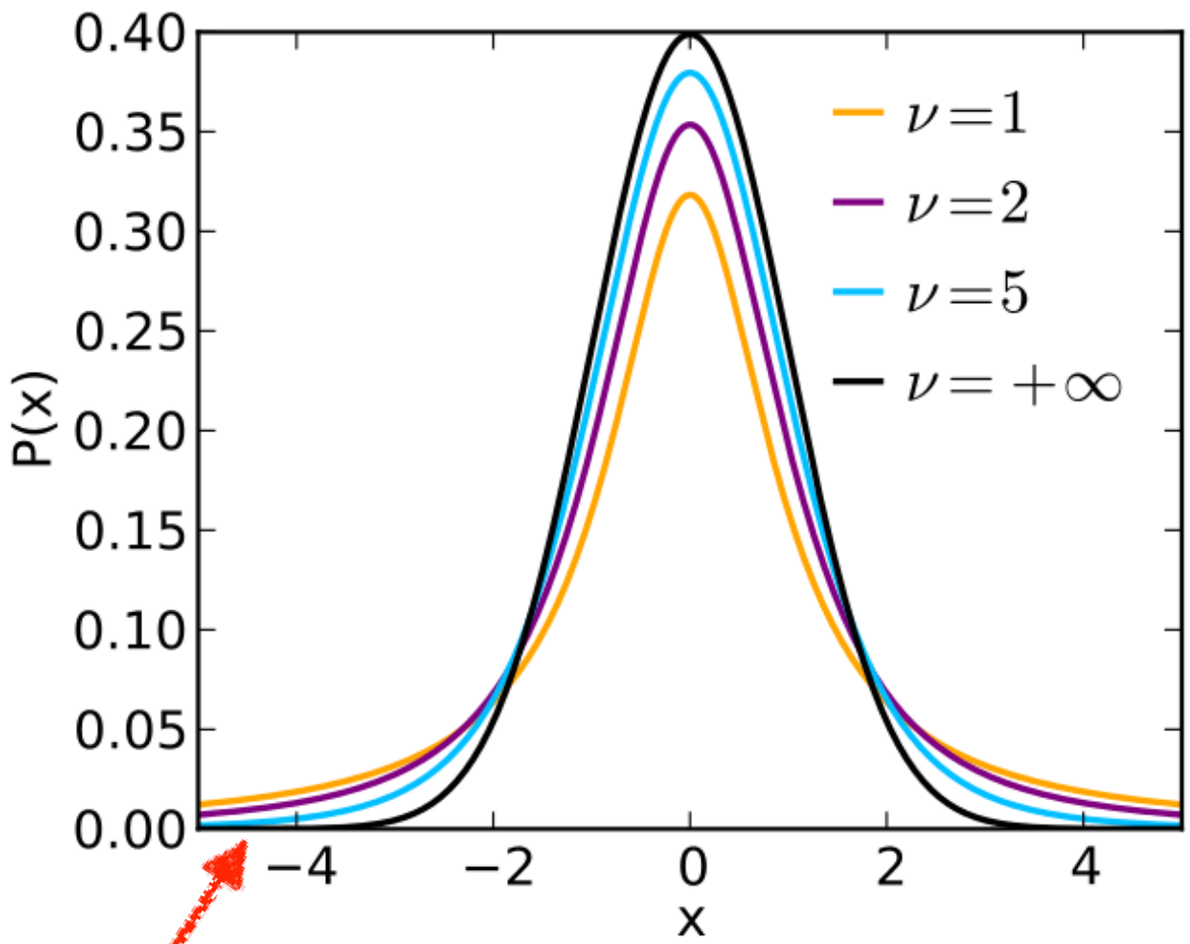
    where the gamma-function:

$$\Gamma(n) = \int_0^\infty y^{n-1}e^{-y}dy$$

- Even/symmetric: $f_X(x) = f_X(-x)$
- Mean: $\mu_t = 0$ for $\nu>1$
- Variance: $\sigma_t^2 = \frac{\nu}{\nu-2}$ for $\nu>2$
- $n \to \infty$: $t(n-1) \sim \mathcal{N}(0,1)$

ν: Degrees of freedom

n: Number of samples



t(ν) is heavy (large tail) for small ν

When the variance is unknown!

Find the Mean Using the t-score

- To test the null hypothesis H_0: $\mu = \mu_0$ using the *t* statistic instead of the *z* statistic, the p-value is

$$pval = 2 \cdot \left(1 - t_{cdf}(|t|, n-1)\right)$$

- where $t_{cdf}(t, n-1) = \Pr(T \leq t)$ denotes the CDF of a t distribution with n-1 degrees of freedom.
- The 95% confidence interval for the mean is

$$\bar{x} \pm t_0 \cdot s/\sqrt{n}$$

- where t_0 is chosen such that

Depends on n-1 (degrees of freedom)

$$\Pr(T \leq t_0) = 1 - \frac{\alpha}{2} = 1 - \frac{0.05}{2} = 0.975$$

The t-Distribution in Matlab

- Calculating probabilities of a t-distributed random variable

$$T \sim t(n-1)$$

- $f(t)$ = `tpdf(t,n-1)`
- $\Pr(T \leq t)$ = `tcdf(t,n-1)`

- where n is the number of samples.
- Given a probability $1 - \alpha/2$, what is the corresponding value t_0, such that $\Pr(T \leq t_0) = 1 - \alpha/2$?

two-tailed event

- `t0 = tinv(1-alpha/2,n-1)`

The Mean for a Population with Unknown Variance

```
x = [ 247.7092
      249.7320
      248.4911
      245.7529
      248.9114
      251.7742
      247.7648
      253.8474
      245.8562
      251.6590
      249.5829
      250.1789
      251.9603
      244.4238
      251.0956
      248.1759
      249.1428
      246.5550
      248.4083
      248.9627
      249.1712
      252.4946
      249.7412
      253.1700
    253.7220 ]
```

Data (n=25)

Then the sample mean ($\bar{x}$) is

```
>> mean(x)

ans =

   249.5313
```

and the (unbiased) estimate of the variance is

```
>> var(x)

ans =

    6.2900
```

corresponding to an empirical standard deviation of $\sqrt{6.29} = 2.508$.

The Mean for a Population with Unknown Variance

- Recall that in the cup filling machine example, the null hypothesis is

$$H_0: \mu = 250$$

- Test size

$$t = \frac{\bar{x} - \mu}{s/\sqrt{n}} = \frac{249.53 - 250}{2.51/\sqrt{25}} = -0.9363 \sim t(n-1)$$

- P-value

$$pval = 2 \cdot \left(1 - t_{cdf}(|t|, n-1)\right) = 2 \cdot \left(1 - t_{cdf}(0.9363, 25-1)\right)$$
$$= 2 \cdot (1 - 0.8208) = 0.3584 \quad > 0.05$$

- and we fail to reject the null hypothesis.

The Mean for a Population with Unknown Variance

- The 95% confidence interval for the mean is $\bar{x} \pm t_0 \cdot s/\sqrt{n}$, so the endpoints are

Lower bound:

$$\mu_- = \bar{x} - t_0 \cdot \frac{s}{\sqrt{n}} = 249.53 - 2.0639 \cdot \frac{2.51}{\sqrt{25}} = 248.49$$

Upper bound:

$$\mu_+ = \bar{x} + t_0 \cdot \frac{s}{\sqrt{n}} = 249.53 + 2.0639 \cdot \frac{2.51}{\sqrt{25}} = 250.57$$

- where t0 = tinv(1-alpha/2,n-1) = tinv(0.975,24)= 2.0639

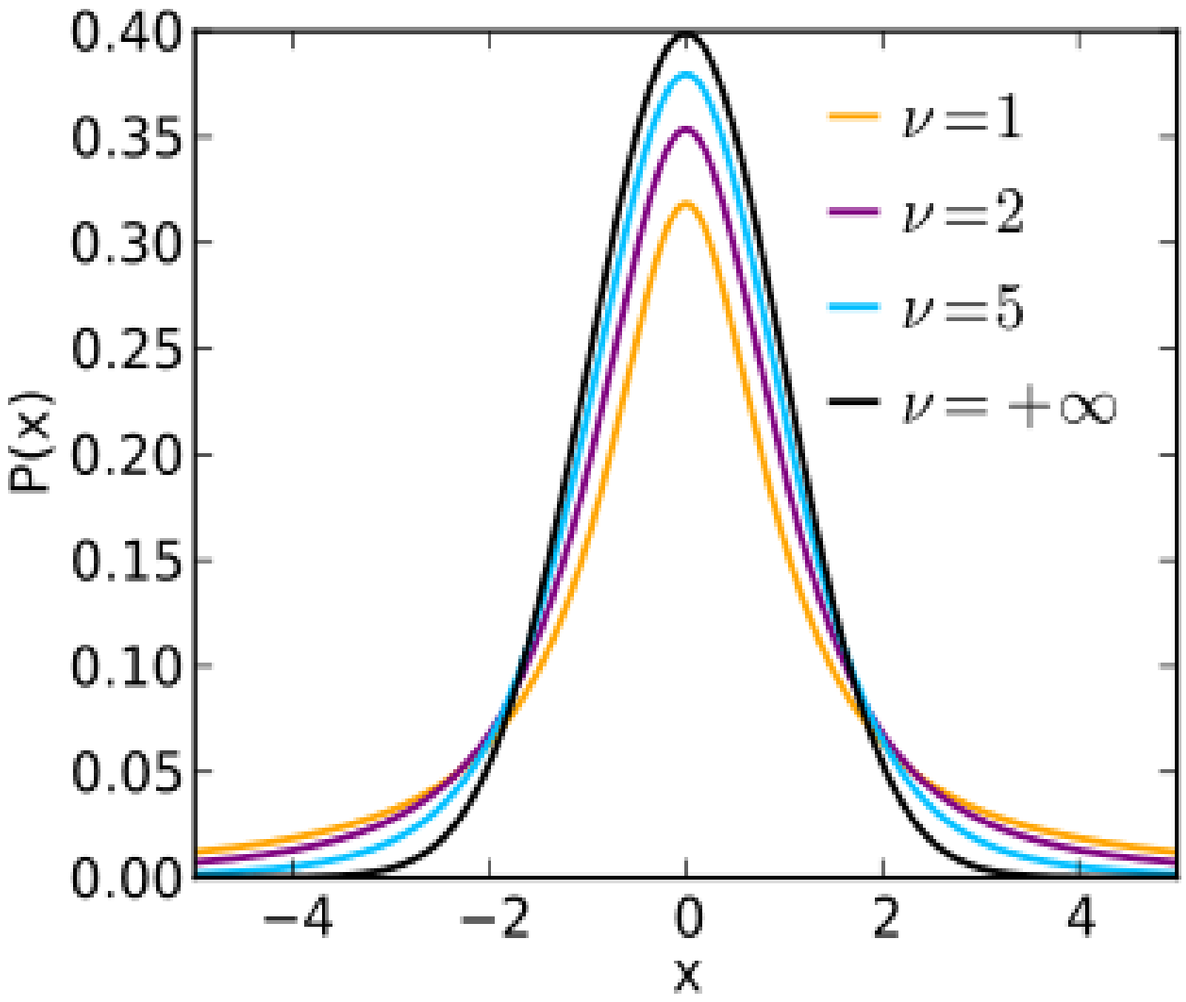
Convergence of the t-dist. towards a Std. Norm. dist.

- The confidence interval obtained with t statistic is wider than the one obtained with the z statistic.
- This results from the fact that we do not know the true standard deviation; we have to use the estimate s instead of the true value σ.
- As a result, we always have $t_0 \geq 1.96$ for a significance level of $\alpha = 0.05$, which on average leads to a wider confidence interval for the t statistic.

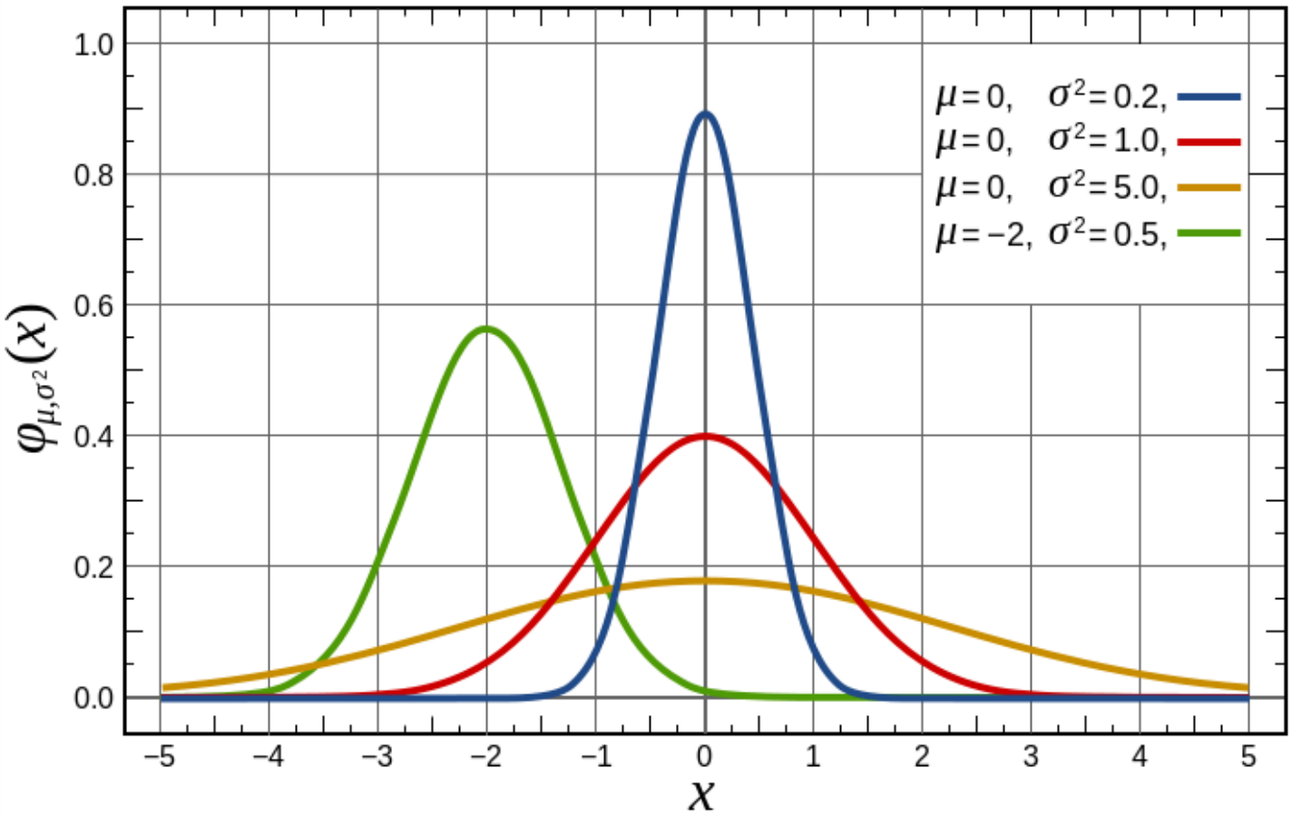
n	2	3	5	10	30	∞
$\boldsymbol{t_0}$	12.71	4.30	2.78	2.26	2.05	1.96

Convergence of the t-dist. towards a Std. Norm. dist.

density of the t-distribution



density of the normal distribution



Checking for Normality in Sampled Data (Q-Q plots)

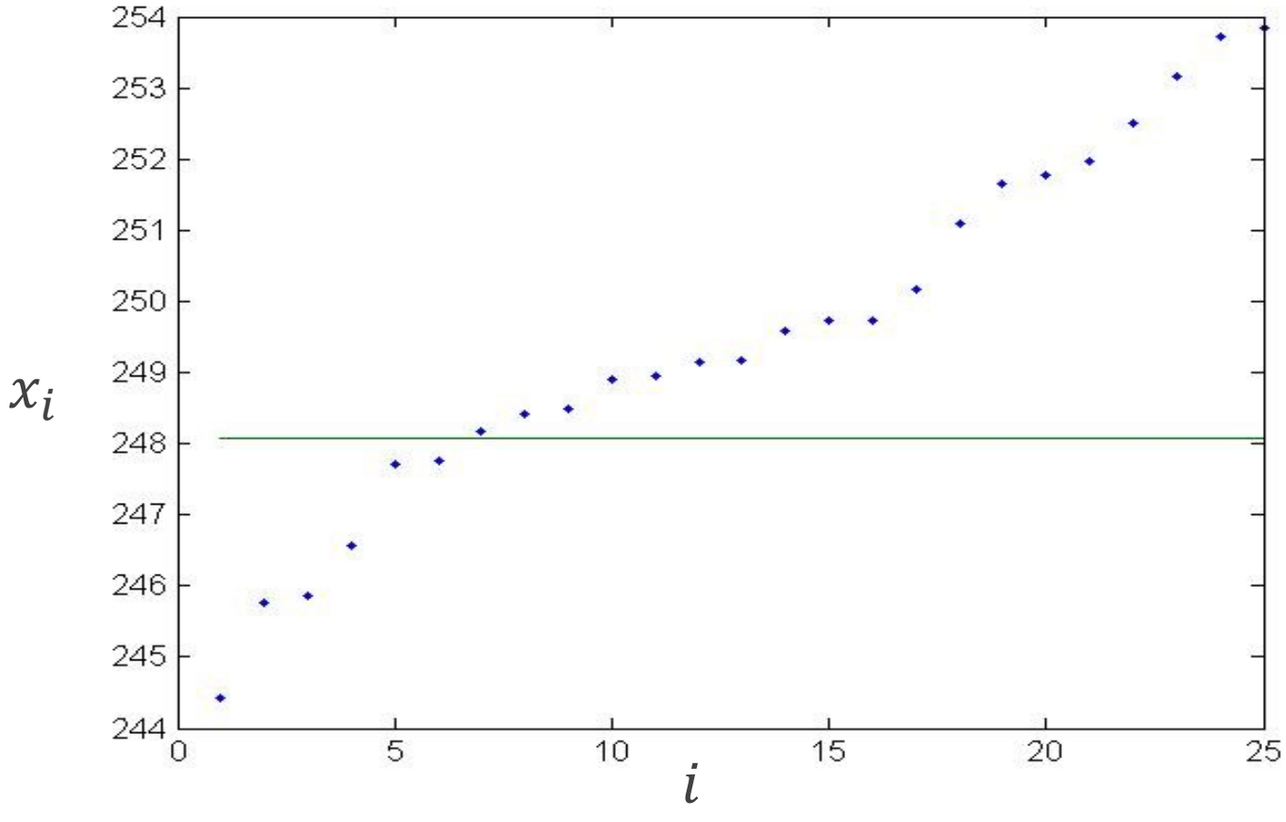
- We can quite safely use the central limit theorem (CLT) to make inference about the mean of any population (i.e., distribution), provided that the sample size is sufficiently large (say $n \geq 30$).
- However, if *n* is small the CLT does not hold anymore.
- In this case, statistical inference based on either the *z*-score or *t*-score only works, if the sampled data $x_1, x_2, \ldots, x_n$ are themselves normally distributed.
- **Hence, we need a method to check whether the data are normally distributed.**

Quantiles *(Fraktiler)*

- The 25% quantile of the previous data

```
q25 = quantile(x,0.25)
q25 = 248.0731
```

$$\Pr(X \leq q25) = \Phi(z25) = 0.25$$
$$\Downarrow$$
$$z25 = \frac{q25 - \mu}{\sigma/\sqrt{n}} = \Phi^{-1}(0.25) = -0.675$$



Sorted data values with the estimated 25% percentile = 248.07.

Roughly 25% of the data should lie below this value.

Q-Q plot

- The quantiles of standard normally distributed data with *n* samples are roughly such that

$$x_{[i]} \leftrightarrow \Phi^{-1}\left(\frac{i-0.5}{n}\right) = z_{[i]} = \frac{x_{[i]} - \mu}{\sigma}$$

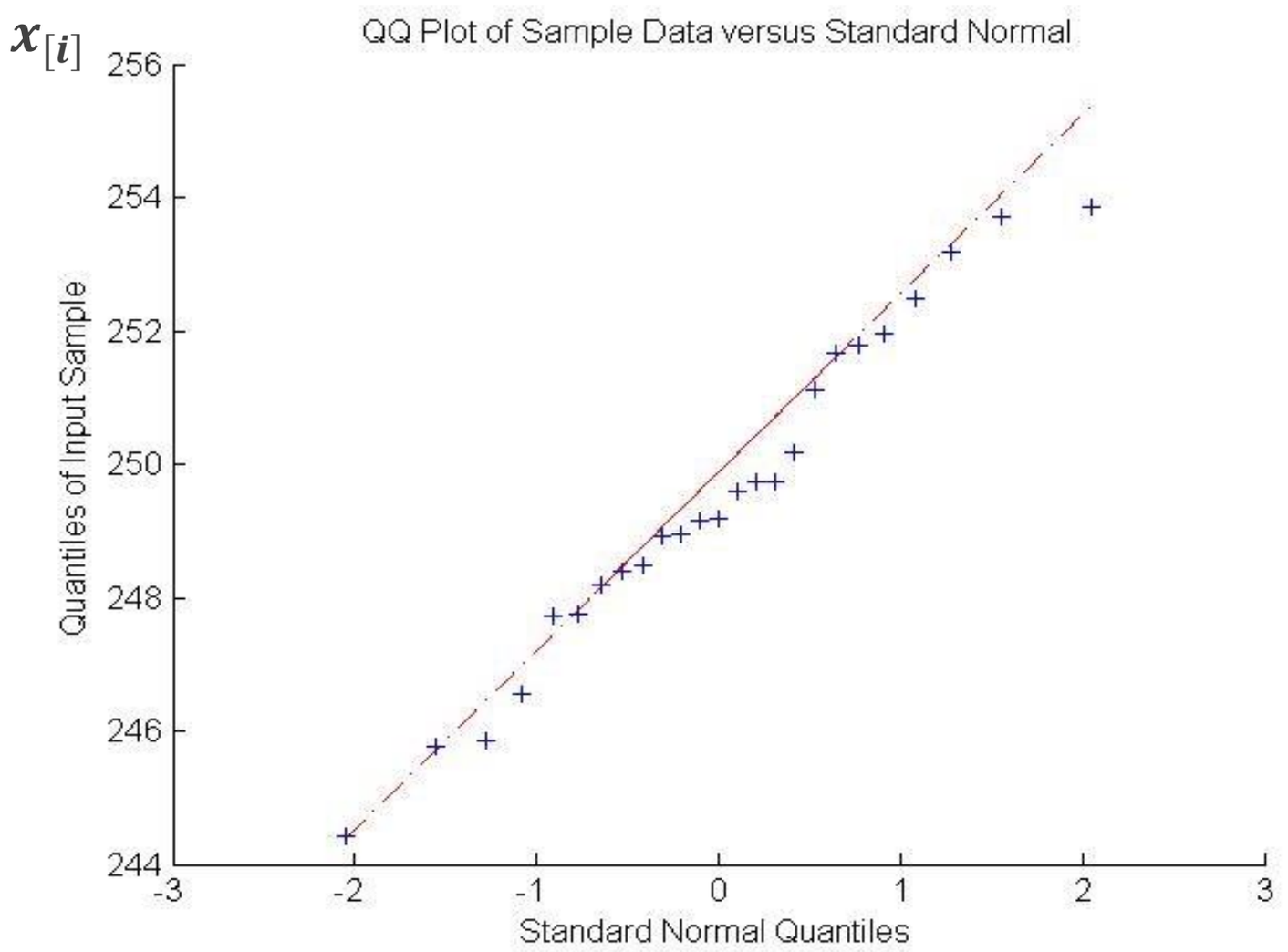
$\sim[0;1]$

$]-\infty;\infty[$

- where $x_{[i]}$ denotes the *i*'th sample after sorting the samples $x_1, x_2, \ldots, x_n$ in ascending order.
- If the data are consistent with a sample from a normal distribution, then plotting $x_{[i]}$ vs. $\Phi^{-1}\left(\frac{i-0.5}{n}\right)$ should result in a straight line.
- **This is the Q–Q plot.**

$$x_{[i]} = \sigma \cdot \Phi_{[i]}^{-1} + \mu$$

Example



$$\Phi^{-1}\left(\frac{i-0.5}{n}\right)$$

Q-Q plot of the data from slide 33. The data points lie roughly on a straight line, and we conclude that the data are in fact normally distributed.

The Cavendish Experiment

- The Cavendish experiment, performed in 1797–98 by British scientist Henry Cavendish, was the first experiment to measure the force of gravity between masses in the laboratory and the first to yield accurate values for the gravitational constant.
- The value generally accepted today is 5.517.
- Cavendish's measurements were

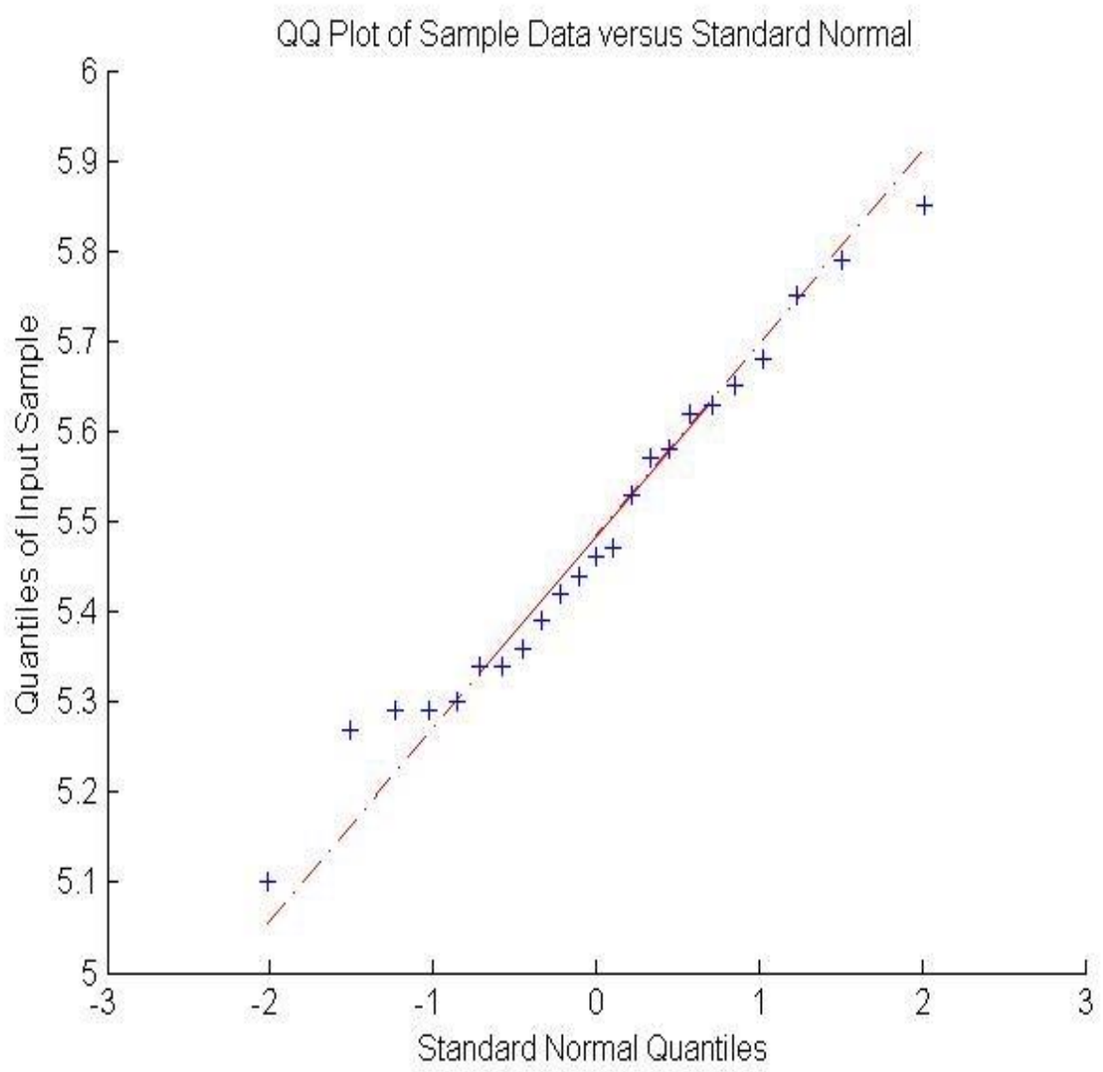
```
x = [ 5.36 5.29 5.58 5.65 5.57 5.53 5.62 5.29 ...
      5.44 5.34 5.79 5.10 5.27 5.39 5.42 5.47 ...
      5.63 5.34 5.46 5.30 5.75 5.68 5.85 ];
```

The Cavendish Experiment

- Check for normality:

  `qqplot(x)`

- The Q-Q plot results in a straight line, and hence we can conclude that the data are normally distributed.



The Cavendish Experiment

- The sample mean (and hence Cavendish's estimate of the gravitational constant) is

```
>> mean(x)
    5.4835
```

- with an empirical variance of

```
>> var(x)
    0.0363
```

- The null hypothesis needed to test if Cavendish's estimate corresponds to the accepted value today is

$$H_0\colon \mu = 5.517$$

The Cavendish Experiment

- Since we observe a sample mean of $\bar{x} = 5.4835$ and an empirical standard deviation of $s = \sqrt{0.0363} = 0.1904$, the test size with n = 23 is

$$t = \frac{\bar{x} - \mu}{s/\sqrt{n}} = \frac{5.4835 - 5.517}{0.1904/\sqrt{23}} = -0.8438 \sim t(n-1)$$

- P-value

$$pval = 2 \cdot \left(1 - t_{cdf}(|t|, n-1)\right) = 2 \cdot \left(1 - t_{cdf}(0.8438, 23-1)\right)$$
$$= 2 \cdot (1 - 0.7961) = 0.4078 \quad > 0.05$$

- and we fail to reject the null hypothesis.
- In other words, we conclude that Cavendish's estimate of earth's gravitational constant corresponds to the accepted value today.

The Cavendish Experiment

- The 95% confidence interval for the mean is $\bar{x} \pm t_0 \cdot s/\sqrt{n}$. We have

```
t0 = tinv(1-0.05/2,n-1)= tinv(0.975,23-1)= 2.0739
```

- so the endpoints of the confidence interval are

Lower bound:

$$\mu_{\_} = \bar{x} - t_0 \cdot \frac{s}{\sqrt{n}} = 5.4835 - 2.0739 \cdot \frac{0.1904}{\sqrt{23}} = 5.4012$$

Upper bound:

$$\mu_{+} = \bar{x} + t_0 \cdot \frac{s}{\sqrt{n}} = 5.4835 + 2.0739 \cdot \frac{0.1904}{\sqrt{23}} = 5.5658$$

TEST CATALOG FOR THE MEAN (UNKNOWN VARIANCE)

- **Statistical model:**
- $X_1, X_2, \dots, X_n$ are i.i.d. samples of a random variable *X* with mean μ and variance σ^2.
- Parameter estimates:

$$\hat{\mu} = \bar{x} = \frac{1}{n}\sum_{i=1}^{n} X_i \sim N(\mu, \sigma^2/n)$$

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

- Where the observation is $\bar{x}$ = 'the average of *n* samples drawn from *X*'s distribution'.
- NOTE: The statistical model is only true if *n* is sufficiently large ($n \geq 30$) or if the samples are drawn from a normal population with mean μ and variance σ^2.

- **Hypothesis test (two-tailed):**
- $H_0: \mu = \mu_0$
- $H_1: \mu \neq \mu_0$
- Test size: $t = \frac{\bar{x}-\mu_0}{s/\sqrt{n}} \sim t(n-1)$
- Approximate p-value: $2 \cdot \left|1 - t_{cdf}(|t|)\right|$

- **95% confidence interval:**
- $\mu_- = \bar{x} - t_0 \cdot s/\sqrt{n}$
- $\mu_+ = \bar{x} + t_0 \cdot s/\sqrt{n}$

- where t0 = tinv(1-0.05/2,n-1)

Words and Concepts to Know

Heavy

Quantiles

Left-tailed

Null hypothesis

Test catalog

Reject

Q-Q plot

Alternative hypothesis

t-score

Fail to reject

Right-tailed

Students t-distribution

Two-tailed

Hypothesis test

Degrees of freedom

11. Chi-Square tests, the Binomial and Poision Distributions

Gunvor Elisabeth Kirkelund
Lars Mandrup
Slides and material provided in parts by
Henrik Pedersen

Todays Content

- Repetition from last time
- Chi-Square Test
- The Binomial Distribution
  - Approximation to the Normal distribution
- The Poisson Distribution
  - Approximation to the Normal distribution

Hypothesis

- Definition – Null hypothesis (H_0)
  - The statement being tested in a test of statistical significance is called the **null hypothesis**. The test of significance is designed to assess the strength of the evidence against the null hypothesis.
  - Usually, the null hypothesis is a statement of 'no effect', 'no difference' or 'no relation' between the phenomena whose relation is under investigation.
- Definition – Alternative hypothesis (H_1)
  - The statement that is hoped or expected to be true instead of the null hypothesis is the **alternative hypothesis**
  - The alternative hypothesis, as the name suggests, is the alternative to the null hypothesis: it states that there is some 'effect/difference' or some 'kind of relation'.

Important!

- One cannot "prove" a null hypothesis, one can only test how close it is to being true.
- Therefore, we never say that we *accept* the null hypothesis, but that we either ***reject it*** or ***fail to reject it***.

Test Statistics, p-value, significance level and confidence interval

- **Test statistics:**
  - A random variable that summarized a data-set by reducing the data to one value that can be used to perform the hypothesis test.

  - Known μ and σ^2: $z = \frac{\bar{x}-\mu}{\sigma/\sqrt{n}} \sim \mathcal{N}(0,1)$ z-statistic

  - Known μ and unknown σ^2: $t = \frac{\bar{x}-\mu}{s/\sqrt{n}} \sim t(n-1)$ Students t

- **p-value:** $p - value = Pr(Worse\ result\ than\ X|H_0)$

- **Significance level α:** Limit for the p-value to reject the NULL hypothesis.
  Typical we use **α = 0,05 = 5%.**

- **Confidence interval:** $[\theta_-; \theta_+]$ such that $Pr(\theta_- \leq \theta \leq \theta_+) = 1 - \alpha$
  Typical the 95% confidence interval.

TEST CATALOG FOR THE MEAN (KNOWN VARIANCE)

- **Statistical model:**
- $X_1, X_2, \ldots, X_n$ are i.i.d. samples of a random variable *X* with mean μ and variance σ^2.
- Parameter estimate:

$$\hat{\mu} = \bar{x} = \frac{1}{n}\sum_{i=1}^{n} X_i \sim N(\mu, \sigma^2/n)$$

- Where the observation is $\bar{x}$ = 'the average of *n* samples drawn from *X*'s distribution'.
- NOTE: The statistical model is only true if *n* is sufficiently large ($n \geq 30$) or if the samples are drawn from a normal population with mean μ and variance σ^2.

- **Hypothesis test (two-tailed):**
- $H_0: \mu = \mu_0$
- $H_1: \mu \neq \mu_0$
- Test size: $z = \frac{\bar{x}-\mu_0}{\sigma/\sqrt{n}} \sim N(0,1)$
- Approximate p-value: $2 \cdot |1 - \Phi(|z|)|$

- **95% confidence interval:**
- $\mu_{\_} = \bar{x} - 1.96 \cdot \sigma/\sqrt{n}$
- $\mu_{+} = \bar{x} + 1.96 \cdot \sigma/\sqrt{n}$

TEST CATALOG FOR THE MEAN (UNKNOWN VARIANCE)

- **Statistical model:**
- $X_1, X_2, \dots, X_n$ are i.i.d. samples of a random variable *X* with mean μ and variance σ^2.
- Parameter estimates:

$$\hat{\mu} = \bar{x} = \frac{1}{n}\sum_{i=1}^{n} X_i \sim N(\mu, \sigma^2/n)$$

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

- Where the observation is $\bar{x}$ = 'the average of *n* samples drawn from *X*'s distribution'.
- NOTE: The statistical model is only true if *n* is sufficiently large ($n \geq 30$) or if the samples are drawn from a normal population with mean μ and variance σ^2.

- **Hypothesis test (two-tailed):**
- $H_0: \mu = \mu_0$
- $H_1: \mu \neq \mu_0$
- Test size: $t = \frac{\bar{x}-\mu_0}{s/\sqrt{n}} \sim t(n-1)$
- Approximate p-value: $2 \cdot \left|1 - t_{cdf}(|t|)\right|$

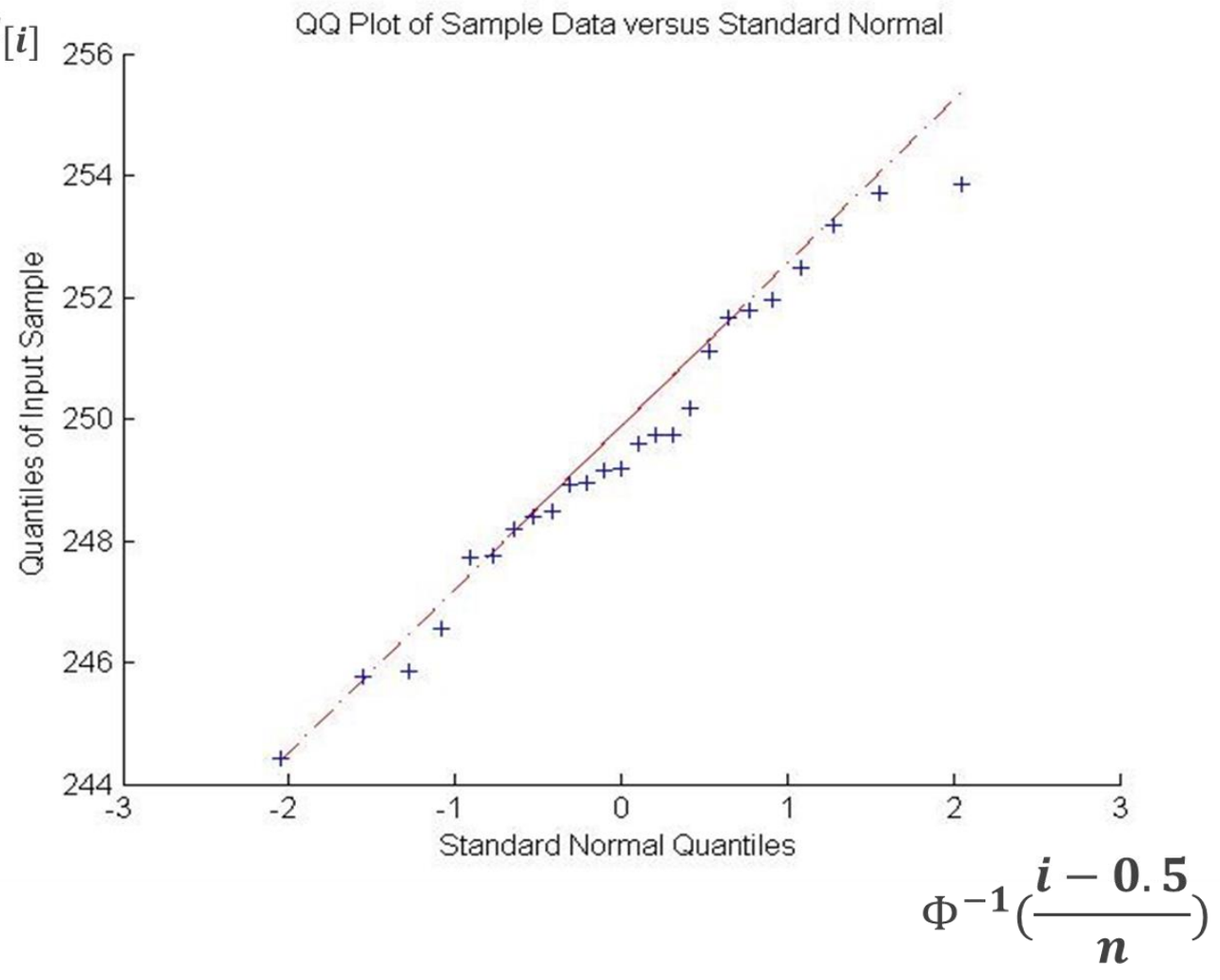
- **95% confidence interval:**
- $\mu_- = \bar{x} - t_0 \cdot s/\sqrt{n}$
- $\mu_+ = \bar{x} + t_0 \cdot s/\sqrt{n}$

- where t0 = tinv(1-0.05/2,n-1)

Q-Q plot

- Q-Q plot – a method to check whether the data are normally distributed

- Sort the samples in ascending order:
  $x_1, x_2, \dots, x_n$
- Plot $x_{[i]}$ vs. $\Phi^{-1}(\frac{i-0.5}{n})$
- If the data are consistent with a sample from a normal distribution it should result in a straight line.



Chi-Square Distribution

- $\hat{\mu}$ → Gaussian Distributed $(\sum X_i)$ (CLT)
- $\hat{\sigma}^2$ → Not Gaussian Distributed $\left(\sum X_i^2\right) \rightarrow \chi^2$-distributed

- If we have a set of i.i.d. data $X_1, X_2, \ldots, X_n$ distributed according to:

$$X_i \sim \mathcal{N}(0,1)$$

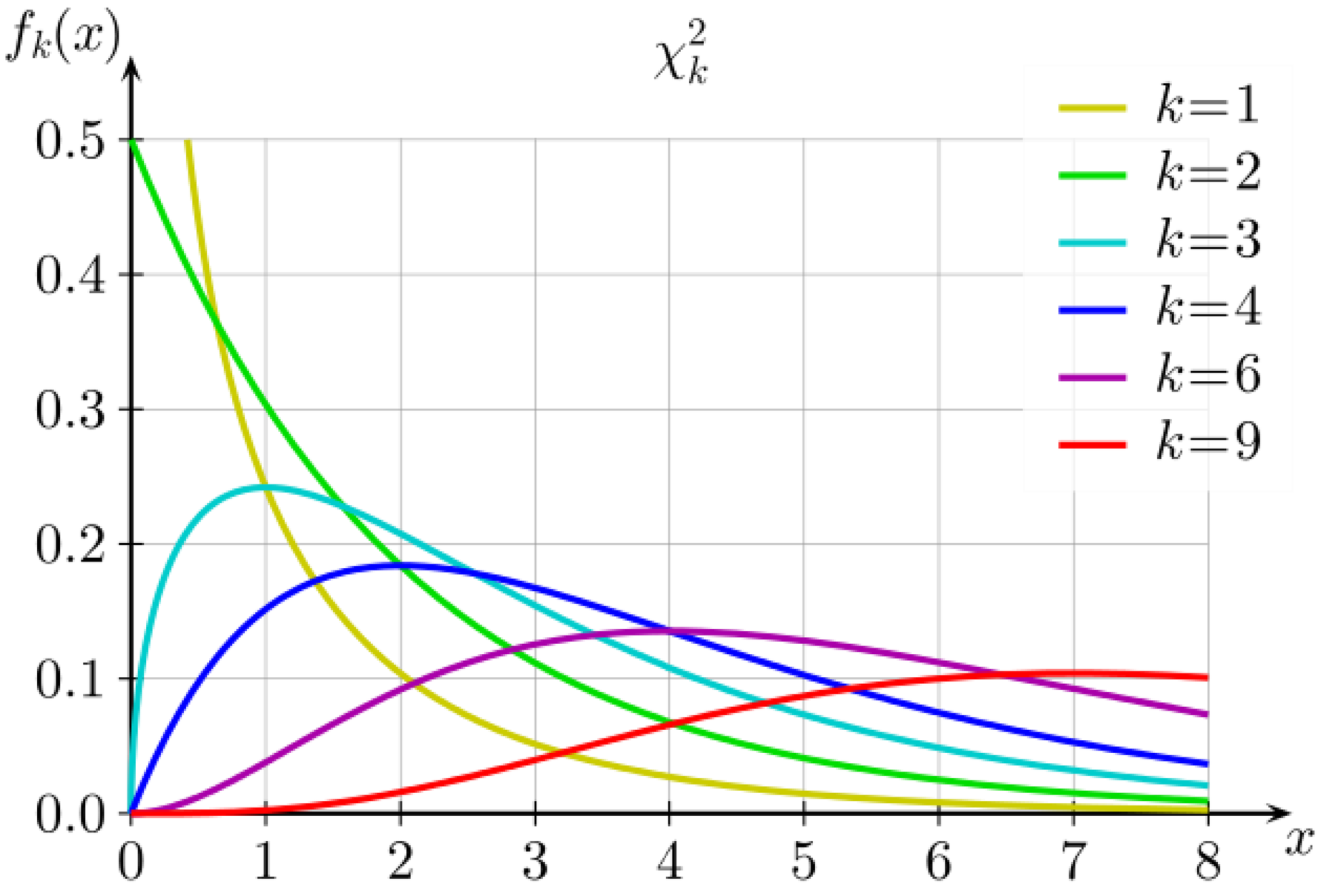
OBS: Standard (normalized) normal distibition

- Then have that: $\mathrm{Q} = \sum_{i=1}^{n} X_i^2$
  is χ_k^2 distributed – k is the degree of freedom.

n-1

- Pdf: $f_{\chi_k^2}(x) = \frac{1}{2^{k/2} \cdot \Gamma(k/2)} \cdot x^{k/2-1} \cdot e^{-x/2}$ (Matlab: chi2pdf(x,k))

Chi-Square Distribution



$-\infty \geq \mathcal{N}(0,1) \geq \infty$

$\chi^2_k \geq 0$

Chi-Square Test for Independence

- We can compute the sample variance for an experiment with n observations:

$$X^2 = \sum_{i=1}^{n} \frac{(observed - expected)^2}{Expected}$$

- if X^2 is large, then the data is a poor fit, and thus the NULL hypothesis is rejected.

Observed = Expected

k=4, α=0.05: H_0 rejected if $X^2>9.4$
(see tabel 3 in ''Random Signals'')

- Fx: How well fits data with an expected function (curve)?

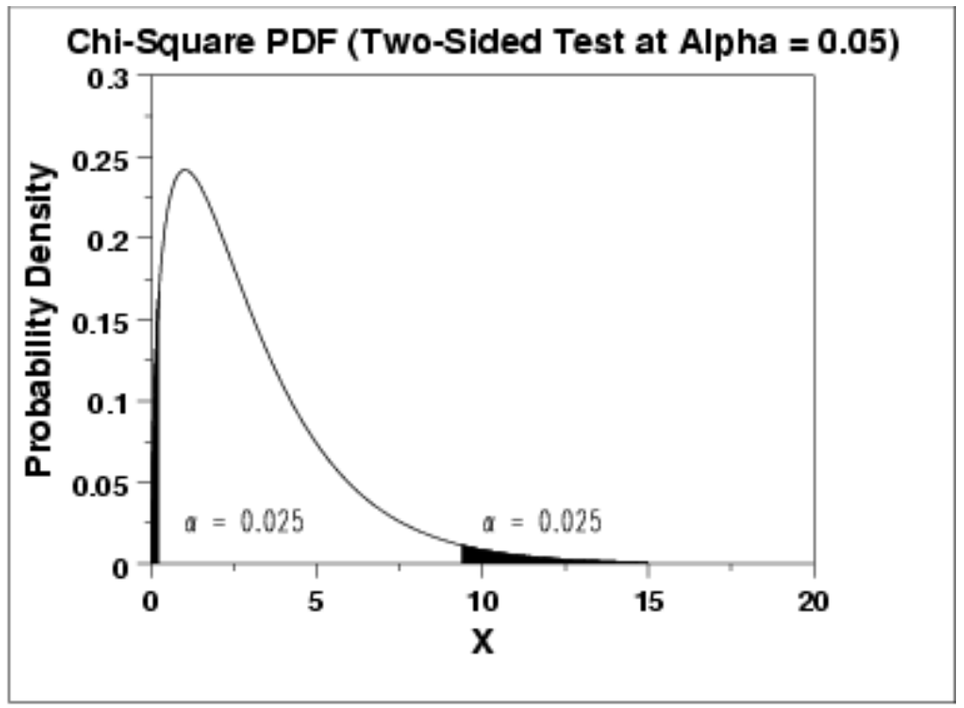
Chi-Square Test for Variance

- Hypothesis: $H_0 : \sigma^2 = \sigma_0^2$

  $H_1 : \sigma^2 \neq \sigma_0^2$

- Test statistics: $T = (N-1) \cdot \frac{s^2}{\sigma_0^2}$

- N = sample size (should be large).



From "Engineering Statistics Handbook"

- For a two tailed test we fail to reject NULL if:

$$\chi^2_{N-1,\alpha/2} < T < \chi^2_{N-1,1-\alpha/2}$$

- where $\chi^2_{N-1,*}$ are the lower and upper critical values of the Chi-Square distribution with N-1 degrees of freedom

Bernoulli Trial

- Two possible outcomes

$$B = \{0,1\}$$

- Probabilities

$$\Pr(B = 1) = p \qquad \text{(success)}$$
$$\Pr(B = 0) = 1 - p \qquad \text{(failure)}$$

- Notation

$$B \sim bernoulli(p)$$

The Binomial Distribution

- Let $B_1, B_2, \ldots, B_n$ be independent random variables, where

$$B_i \sim bernoulli(p)$$

- Then the number of successes

$$X = \sum_{i=1}^{n} B_i$$

- is a binomially distributed random variable with parameters *n* and *p*.

- Notation

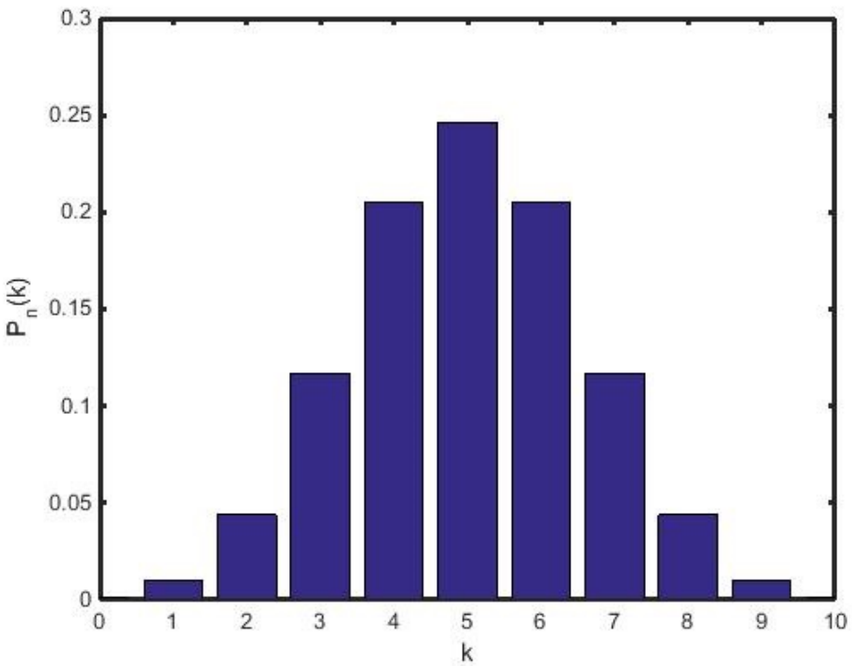
$$X \sim binomial(n, p)$$

The Binomial Distribution

- We have n repeated trials.
- Each trial has two possible outcomes
  - **Success** — probability p
  - **Failure** — probability 1-p
- We write the mass function as:

$$f(k|n,p) = \frac{n!}{k!\,(n-k)!}p^k(1-p)^{n-k}$$

$$= \binom{n}{k} p^k (1-p)^{n-k}$$



Mendel's Experiment

- Pea seed color is controlled by two alleles
  - 'A' the dominant (green)
  - 'a' the recessive (yellow)
- Genotypes
  - 'AA', 'Aa', 'aA' should produce a green pea.
  - 'aa' should produce a yellow pea.
- Mendel's hypothesis
  - Crossing 'Aa' genotypes should result in equally many pea plants of each genotype.
- More formally
  - $\Pr(yellow\ plant) = \frac{1}{4}$
  - $\Pr(green\ plant) = \frac{3}{4}$

Mendel's Experiment

- Mendel looked at the colors of 580 offspring plants.
- Result
  - 152 yellow plants
  - 428 green plants

Two possible outcomes → Bernoulli Trial

- In an idealized experiment
  - 145 yellow plants (580/4)
  - 435 green plants (580*3/4)
- Could the deviation be explained by random variation?
- Or is Mendel's hypothesis incorrect?

Mendel's Experiment

- We denote by X the number of yellow plants.
- Then the statistical model is

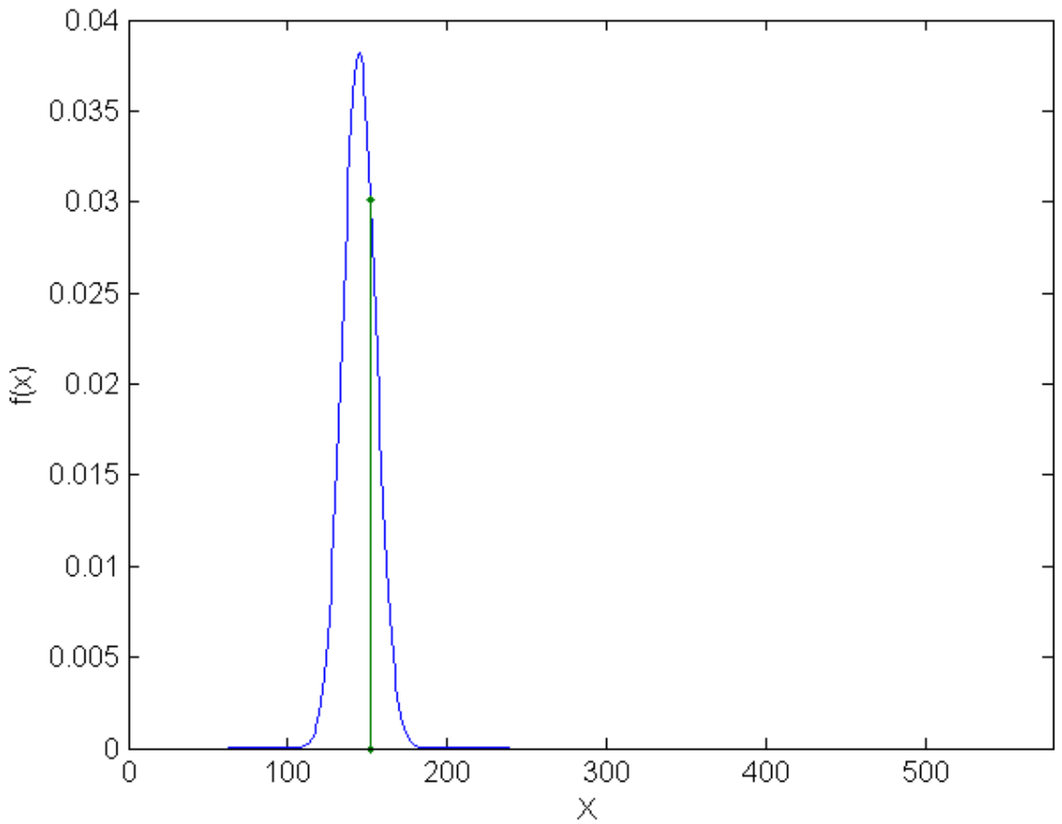
$$X \sim binomial(n = 580, p)$$

- The hypothesis concerns the unknown parameter, p

$$H_0: p = 1/4$$
$$H_1: p \neq 1/4$$

Mendel's Experiment

- The pdf for the underlying Mendel's experiment (n=580, p=1/4) along with the observed value, x=152.




```
n      = 580;  % Number of trials
p      = 0.25; % Probability of success
x      = 0:n;  % x-values for plotting the PDF
fx     = binopdf(x,n,p); % PDF
xobs   = 152;  % Observed x (=number of
                                  successes)

plot(x,fx,…
     …[xobs xobs],[0 binopdf(xobs,n,p)],'.-')
axis([0 580 0 0.04])
xlabel('X')

ylabel('f(x)')
```

p-value for Mendel's Experiment

- The p-value is the probability of observing a value of the random variable X that is more extreme than 152.
- By extreme we mean with respect to the value of X that we would observe in an idealized experiment, given that the null hypothesis is true:

$$p \cdot n = \frac{1}{4} \cdot 580 = 145$$

- Hence, for a two-tailed test, we need to consider the events $\{X \geq 152\}$ and $\{X \leq 138\}$, both of which deviate by 7 from the theoretical value of 145.

p-value for Mendel's Experiment

- Calculation of p-value:

$$pval = 2 \cdot min\{Pr(X \geq x_{max}), Pr(X \leq x_{min})\}$$

$$= 2 \cdot min\{Pr(X \geq 152), Pr(X \leq 138)\}$$

$$= 2 \cdot min\{1 - Pr(X < 152), Pr(X \leq 138)\}$$

$$= 2 \cdot min\left\{1 - F_{bino}\left(151; n = 580, p = \frac{1}{4}\right), F_{bino}\left(138; n = 580, p = \frac{1}{4}\right)\right\}$$

$$= 2 \cdot min\{1 - 0.7350, 0.2682\} = 2 \cdot min\{0.2650, 0.2682\}$$

$$= 0.5300 > 0.05 = \alpha$$

- where $F_{bino}(x;n,p)$ denotes the cdf of a binomial distribution with parameters n and p.
- Since pval>α=0.05, we fail to reject the null hypothesis (Mendel's hypothesis)

Binomial Distribution in Matlab

- Calculating the probabilities $\Pr(X = x)$ and $\Pr(X \leq x)$ of a binomially distributed random variable

$$X \sim binomial(n, p)$$

- $\Pr(X = x)$ = binopdf(x,n,p)
- $\Pr(X \leq x)$ = binocdf(x,n,p)

- x must be an integer value and can in general be a vector or array.

Normal Approximation to the Binomial Distribution

- First, consider

$$B \sim bernoulli(p)$$

- Mean

$$E[B] = \sum_{b=\{0,1\}} b \cdot \Pr(B = b) = 0 \cdot \Pr(B = 0) + 1 \cdot \Pr(B = 1)$$

❖
$$= 0 \cdot (1 - \mathrm{p}) + 1 \cdot \mathrm{p} = \mathrm{p}$$

- Variance

$$Var(B) = \sum_{b=\{0,1\}} (b - p)^2 \cdot \Pr(B = b)$$
$$= (0 - p)^2 \cdot \Pr(B = 0) + (1 - p)^2 \cdot \Pr(B = 1)$$
$$= p^2(1 - p) + (1 - p)^2 p = p(1 - p)$$

Normal Approximation to the Binomial Distribution

- Now, define

$$X = \sum_{i=1}^{n} B_i$$

- where $B_i \sim bernoulli(p)$, and $B_i's$ are independent.
- Mean

$$E[X] = E\left[\sum_{i=1}^{n} B_i\right] = \sum_{i=1}^{n} E[B_i] = \sum_{i=1}^{n} p = np$$

- Variance

$$Var(X) = Var\left(\sum_{i=1}^{n} B_i\right) = \sum_{i=1}^{n} Var(B_i) = \sum_{i=1}^{n} p(1-p) = np(1-p)$$

Normal Approximation to the Binomial Distribution

- Now define the standardized random variable

$$Z = \frac{X - E[X]}{\sqrt{Var(X)}} = \frac{X - np}{\sqrt{np(1-p)}}, \text{ where } X \sim binomial(n, p)$$

approximately

- Then if $np > 5$ and $n(1-p) > 5$, Z is standard normally distributed

$$Z \sim N(0,1)$$

- This fact follows from the central limit theorem:

$$\boldsymbol{X = \sum_{i=1}^{n} B_i \sim N(n \cdot E[B], n \cdot Var(B)) = N(np, np(1-p))}$$

Approximate p-value for Mendel's Experiment

- Standardizing the observation, x=152, we get

$$z = \frac{x - np}{\sqrt{np(1-p)}} = \frac{152 - 580 \cdot 1/4}{\sqrt{580 \cdot 1/4 \cdot (1 - 1/4)}} = \frac{152 - 145}{\sqrt{145 \cdot 3/4}} = 0.6712$$

- and the two-tailed p-value is

$$\begin{aligned} pval &= 2 \cdot min\{Pr(Z \geq z), Pr(Z \leq z)\} \\ &= 2 \cdot min\{Pr(Z \geq 0.6712), Pr(Z \leq 0.6712)\} \\ &= 2 \cdot min\{1 - Pr(z < 0.6712), Pr(X \leq 0.6712)\} \\ &= 2 \cdot min\{1 - \Phi(0.6712), \Phi(0.6712)\} \\ &= 2 \cdot min\{1 - 0.7490, 0.7490\} = 2 \cdot min\{0.2510, 0.7490\} \\ &= 0.5021 > 0.05 = \alpha \end{aligned}$$

- which is slightly smaller than the exact p-value calculated earlier, but leads to the same result: failure to reject the null hypothesis.

➢ OBS: If $p \approx \alpha$ – the normal distribution approximation should not be used

Estimation of p in Binomially Distributed Data

- The estimator of *p* is

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

- Unbiased

$$E[\hat{p}] = E[x/n] = \frac{1}{n}E[x] = \frac{1}{n}np = p$$

- Variance

$$Var(\hat{p}) = Var\left(\frac{x}{n}\right) = \frac{1}{n^2}Var(x) = \frac{1}{n^2}np(1-p) = \frac{1}{n}p(1-p)$$

- Notice that the variance of the estimate decreases with $1/n$.

Approximate 95% Confidence Interval

- To find the 95% confidence interval for the parameter p, we must find the limits p_- and p_+, such that the true parameter p lies in the interval $[p_-; p_+]$ with probability 0.95:

$$Pr(p_- \leq p \leq p_+) = 0.95$$

- Assuming that we can use the normal approximation, this condition is equivalent to:

$$Pr(-1.96 \leq z \leq 1.96) = 0.95$$

- where z is the standarized random variable defined earlier: $z = \frac{x-np}{\sqrt{np(1-p)}}$

- Inserting we get:

$$Pr\left(-1.96 \leq \frac{x-np}{\sqrt{np(1-p)}} \leq 1.96\right) = 0.95 \quad \Rightarrow$$

$$Pr\left(\frac{1}{n+1.96^2}\left[x+\frac{1.96^2}{2}-1.96\sqrt{\frac{x(n-x)}{n}+\frac{1.96^2}{4}}\right] \leq p \leq \frac{1}{n+1.96^2}\left[x+\frac{1.96^2}{2}+1.96\sqrt{\frac{x(n-x)}{n}+\frac{1.96^2}{4}}\right]\right) = 0.95$$

Approximate 95% Confidence Interval

- And therefore we get for the 95% confidence interval $[p_-; p_+]$:

$$p_- = \frac{1}{n + 1.96^2}\left[x + \frac{1.96^2}{2} - 1.96\sqrt{\frac{x(n-x)}{n} + \frac{1.96^2}{4}}\right]$$

$$p_+ = \frac{1}{n + 1.96^2}\left[x + \frac{1.96^2}{2} + 1.96\sqrt{\frac{x(n-x)}{n} + \frac{1.96^2}{4}}\right]$$

- In generel, the limits of the 1-α confidence interval for p are:

$$p_- = \frac{1}{n + u^2}\left[x + \frac{u^2}{2} - u\sqrt{\frac{x(n-x)}{n} + \frac{u^2}{4}}\right]$$

$$p_+ = \frac{1}{n + u^2}\left[x + \frac{u^2}{2} + u\sqrt{\frac{x(n-x)}{n} + \frac{u^2}{4}}\right]$$

- where $u = \Phi^{-1}(1 - \frac{\alpha}{2})$

Estimation of p and 95% Confidence Interval Mendel's Experiment

- Mendel's Experiment:
  - Number of trials (plants): n = 580
  - Number of successsses (yellow): x = 152

  - Estimated parameter: $\hat{p} = \frac{x}{n} = \frac{152}{580} = 0.2621$
  - Estimated variance: $\mathrm{Var}(\hat{p}) = \frac{p(1-p)}{n} = \frac{0.2621 \cdot 0.7379}{580} = 0.000333$

  - 95% confidence interval:

$$p_{-} = \frac{1}{580 + 1.96^2}\left[152 + \frac{1.96^2}{2} - 1.96\sqrt{\frac{152(580-152)}{580} + \frac{1.96^2}{4}}\right] = 0.2279$$

$$p_{+} = \frac{1}{580 + 1.96^2}\left[152 + \frac{1.96^2}{2} + 1.96\sqrt{\frac{152(580-152)}{580} + \frac{1.96^2}{4}}\right] = 0.2993$$

- Since $p = 0.25$ lies within the 95% confidence interval, the null hypothesis can't be rejected.

Test catalog for the Binomial Distribution

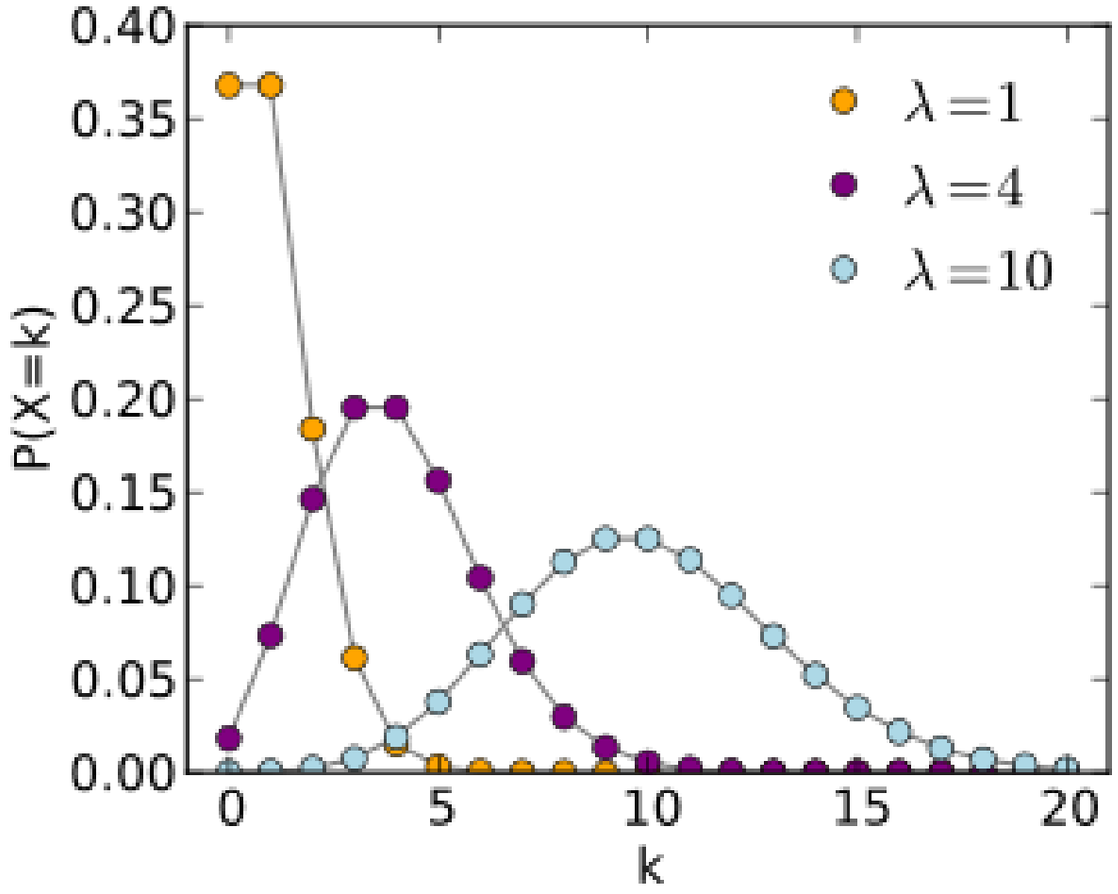
- **Statistical model:**
- $X \sim binomial(n, p)$
- Parameter estimate: $\hat{p} = x/n$
- Where the observation is *x* = 'number of successes out of *n* trials'

- **Hypothesis test (two-tailed):**
- $H_0: p = p_0$
- $H_1: p \neq p_0$
- Test size: $z = \frac{x - np_0}{\sqrt{np_0(1-p_0)}} \sim N(0,1)$
- Approximate p-value: $2 \cdot |1 - \Phi(|z|)|$

- **95% confidence interval:**
- $p_- = \frac{1}{n+1.96^2}\left[x + \frac{1.96^2}{2} - 1.96\sqrt{\frac{x(n-x)}{n} + \frac{1.96^2}{4}}\right]$
- $p_+ = \frac{1}{n+1.96^2}\left[x + \frac{1.96^2}{2} + 1.96\sqrt{\frac{x(n-x)}{n} + \frac{1.96^2}{4}}\right]$

The Poisson Distribution

- The Poisson distribution is a discrete probability distribution.
- The probability of a given number of events k occurring in a fixed interval of time, when:
  - these events occur with a known average rate λ
  - the events are independently of the time since the last event



The Poisson Distribution

- Assume that we have a time axis that is divided into *N* intervals of length Δt.
- For each interval there is one Bernoulli distributed random variable, denoted B_i for the *i*'th interval, denoting the number of arrivals/events in that interval.
- Recalling that B_i={0,1}, there can be either 0 or 1 arrival in each interval.
- Denoting by λ the known average rate of the arrivals/event, we have

$$B_i \sim bernoulli(\lambda \cdot \Delta t)$$

Δt so small that $\lambda \cdot \Delta t < 1$

- That is, the probability of observing an event in the *i*'th interval is proportional to the length (Δt) of the interval.

The Poisson Distribution

- We assume that the observations $B_1, B_2, \ldots, B_N$ are independent.
- Then, the probability of observing $X = x$ events over the entire period of duration $t = N \cdot \Delta t$ is binomially distributed:

$$X \sim binomial(N, \lambda \cdot \Delta t)$$

- Observe that

$$N \cdot (\lambda \cdot \Delta t) = constant = \frac{t}{\Delta t} \cdot (\lambda \cdot \Delta t) = t \cdot \lambda = \gamma$$

- In the limit, as $N \to \infty$ (or $\Delta t \to 0$), it can be shown that

$$\Pr(X = x) = \frac{(\lambda t)^x}{x!} e^{-\lambda t} = \frac{(\gamma)^x}{x!} e^{-\gamma}$$

Poisson Distribution in Matlab

- Calculating the probabilities $\Pr(X = x)$ and $\Pr(X \leq x)$ of a Poisson distributed random variable

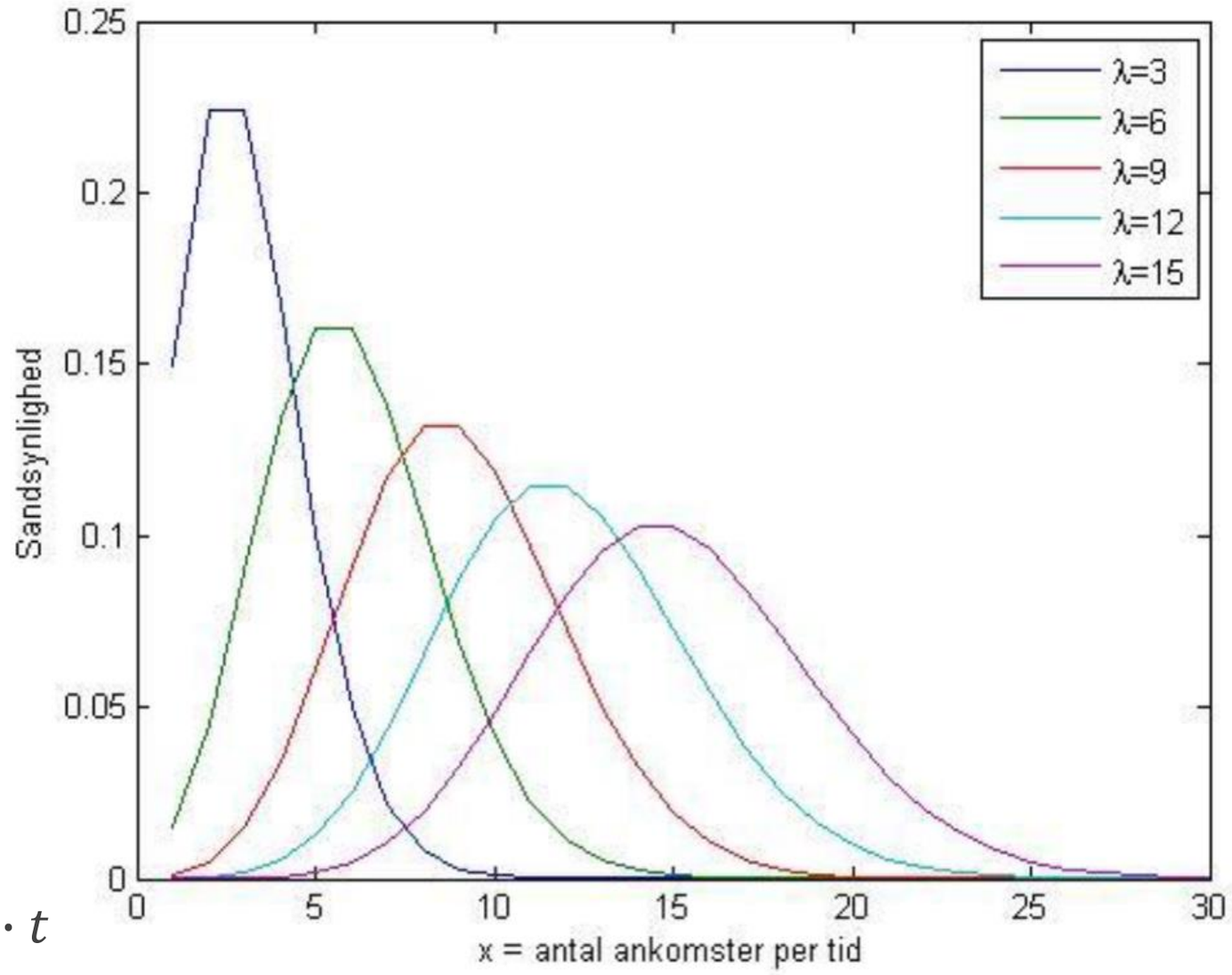
$$X \sim poisson(t \cdot \lambda = \gamma)$$

- $\Pr(X = x) = poisspdf(x, \gamma)$
- $\Pr(X \leq x) = poisscdf(x, \gamma)$

OBS:

x hændelser i tiden t

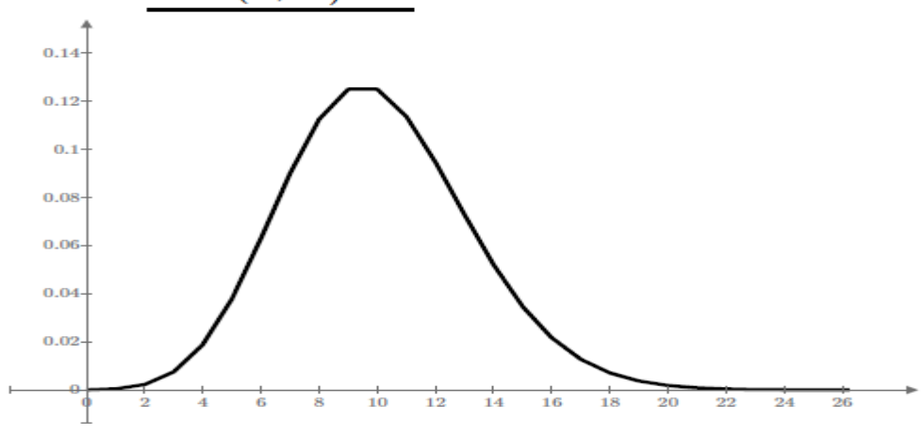
$\gamma = \lambda \cdot t$



The Poisson Distribution

- A custom-handling system receive in average 2 orders pr. minute: $\lambda = 2/min$

- What is the probability that within the next 5 minutes the system should handle x orders:
  - ➢ $Pr(X = x) = \frac{\gamma^x}{x!} e^{-\gamma} = \frac{10^x}{x!} e^{-10}$ $(\gamma = \lambda \cdot t = 10)$
    $Pois(x, 10)$
  - 0 orders? ➢ $Pr(0) = 0.000045$
  - 8 orders? ➢ $Pr(8) = 0.113$
  - 10 orders? ➢ $Pr(10) = 0.125$
  - 15 orders? ➢ $Pr(15) = 0.035$



- If 99% of all orders should be handled within 5 minutes, how many orders shall the system be designed to handle:
  - $\Pr(X > x) < 0.01 \Rightarrow x \geq poissinv(0.99,10) = 18$ → poisscdf(x, γ)=0.99

The Poisson Distribution

- A store claim to have 150 customers pr. hour. $(\lambda = 150/time)$
- What is the probability that within the next 2 hours the store will have x customers:
  - $Pr(X = x) = \frac{\gamma^x}{x!} e^{-\gamma} = \frac{300^x}{x!} e^{-300}$ $(\gamma = \lambda \cdot t = 300)$

Very large

Very small

- 200 customers? $Pr(200) = 3.01 \cdot 10^{-20}$
- 300 customers? $Pr(300) = 2.30 \cdot 10^{-2}$
- 400 customers? $Pr(400) = 5.67 \cdot 10^{-9}$
- 500 customers? $Pr(500) = 1.53 \cdot 10^{-26}$

Very small numbers

The Poisson Distribution

- A store claim to have 150 customers pr. hour ($\lambda = 150/time$).

- What is the probability that within the next 2 hours the store will have:
  - $Pr(X \leq x) = poisscdf(x, \gamma)$ $\quad (\gamma = \lambda \cdot t = 300)$

- <250 customers? ➢ $poisscdf(250{,}300) = 0.002$
- 275-325 customers? ➢ $poisscdf(325{,}300) - poisscdf(275{,}300) = 0.851$
- >325 customers? ➢ $1 - poisscdf(325{,}300) = 0.072$

Normal Approximation to the Poisson Distribution

- Defining the Poisson distributed random variable

$$X \sim poisson(t \cdot \lambda = \gamma)$$

- it can be shown that

$$E[X] = t \cdot \lambda = \gamma$$
$$Var(X) = t \cdot \lambda = \gamma$$

- Now define the standardized random variable

$$Z = \frac{X - E[X]}{\sqrt{Var(X)}} = \frac{X - t \cdot \lambda}{\sqrt{t \cdot \lambda}} = \frac{X - \gamma}{\sqrt{\gamma}}$$

- Then, if $t \cdot \lambda = \gamma > 5$, Z is approximately standard normally distributed

$$Z \sim N(0,1)$$

Approximate p-value

- A store claims: average rate of 150 customers pr. hour.
  $\lambda = \gamma/t = 150$.

- Observe: $x = 280$ customers for 2 hours.

- Formulation of null hypothesis:

$$H_0: \quad \lambda = 150$$

- Standardising the observation, test statistics:

$$z = \frac{x - t \cdot \lambda}{\sqrt{t \cdot \lambda}} = \frac{280 - 2 \cdot 150}{\sqrt{t \cdot 150}} = -1.1547$$

- Two-tailed p-value:

$$2 \cdot |1 - \Phi(|z|) = 2 \cdot |1 - \Phi(1.1547)| = 2 \cdot |1 - 0.8759| = 0.2482$$

- We Fail to reject til null hypothesis.

Estimation of the Average Rate Parameter

- In general, the average rate parameter $\lambda = \gamma/\mathrm{t}$ is unknown and has to be estimated from observed data.
- Given the observation *x* = 'number of arrivals/events" over a time period of duration *t*, the maximum-likelihood estimator is

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

- This is an unbiased estimator, because the expected value of $\hat{\lambda}$ is the true parameter

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

Approximate 95% Confidence Interval

- To find the 95% confidence interval for the parameter λ, we must find the limits λ_- and λ_+, such that the true parameter λ lies in the interval $[\lambda_-; \lambda_+]$ with probability 0.95:

$$Pr(\lambda_- \leq \lambda \leq \lambda_+) = 0.95$$

- Assuming that we can use the normal approximation, this condition is equivalent to:

$$Pr(-1.96 \leq z \leq 1.96) = 0.95$$

- where z is the standarized random variable defined earlier: $z = \frac{x - t \cdot \lambda}{\sqrt{t \cdot \lambda}}$

- Inserting we get:

$$Pr\left(-1.96 \leq \frac{x - t \cdot \lambda}{\sqrt{t \cdot \lambda}} \leq 1.96\right) = 0.95 \quad \Rightarrow$$

$$P\,r\left(\frac{1}{t}\left[x + \frac{1.96^2}{2} - 1.96\sqrt{x + \frac{1.96^2}{4}}\right] \leq \lambda \leq \frac{1}{t}\left[x + \frac{1.96^2}{2} + 1.96\sqrt{x + \frac{1.96^2}{4}}\right]\right) = 0.95$$

Approximate 95% Confidence Interval

- And therefore we get for the 95% confidence interval $[\lambda_-; \lambda_+]$:

$$\lambda_- = \frac{1}{t}\left[x + \frac{1.96^2}{2} - 1.96\sqrt{x + \frac{1.96^2}{4}}\right]$$

$$\lambda_+ = \frac{1}{t}\left[x + \frac{1.96^2}{2} + 1.96\sqrt{x + \frac{1.96^2}{4}}\right]$$

- In generel, the limits of the 1-α confidence interval for p are:

$$\lambda_- = \frac{1}{t}\left[x + \frac{u^2}{2} - u\sqrt{x + \frac{u^2}{4}}\right]$$

$$\lambda_+ = \frac{1}{t}\left[x + \frac{u^2}{2} + u\sqrt{x + \frac{u^2}{4}}\right]$$

- where $u = \Phi^{-1}(1 - \frac{\alpha}{2})$

Approximate 95% Confidence Interval

- The parameter estimate is

$$\hat{\lambda} = \frac{x}{t} = \frac{280}{2} = 140$$

- Since $t \cdot \lambda = 2 \cdot 150 = 300 > 5$ we can use the normal approximation of the confidencde interval:

$$\lambda_{-} = \frac{1}{t} \cdot \left[x + \frac{1.96^2}{2} - 1.96 \cdot \sqrt{x + \frac{1.96^2}{4}} \right] = \frac{1}{2} \cdot \left[280 + \frac{1.96^2}{2} - 1.96 \cdot \sqrt{280 + \frac{1.96^2}{4}} \right] = 124.5$$

$$\lambda_{+} = \frac{1}{t} \cdot \left[x + \frac{1.96^2}{2} + 1.96 \cdot \sqrt{x + \frac{1.96^2}{4}} \right] = \frac{1}{2} \cdot \left[280 + \frac{1.96^2}{2} + 1.96 \cdot \sqrt{280 + \frac{1.96^2}{4}} \right] = 157.4$$

- Note that since the hypothesized parameter ($\lambda = 150$) lies within til 95% confidence interval, we accept the NULL hypothesis.

Geiger–Marsden Experiment

- In one of their experiments, Geiger and Marsden detected $x = 11571$ alpha particles (using a Geiger counter) over a time period of $t = 187776$ seconds. 52t 9min 36sek
- For illustration purposes only, let us assume that we hypothesize $\lambda = 0.060$.
- **Statistical model:**
- x = number of alpha particles detected = 11571
- $X \sim poisson(t \cdot \lambda)$, where t = 187776 seconds

Parameter estimate: $\hat{\lambda} = \frac{11571}{187776} = 0.06162$

Expected number of alpha particles: $\gamma = t \cdot \lambda = 11266$

Geiger–Marsden Experiment

Hypothesis test:

- H_0: $\lambda = 0.060$
- H_1: $\lambda \neq 0.060$

- Test size:

$$z = \frac{x - t \cdot \gamma}{\sqrt{t \cdot \gamma}} = \frac{11571 - 187776 \cdot 0.060}{\sqrt{187776 \cdot 0.060}} = 2.8682 \sim \mathcal{N}(0,1)$$

- Approximative p-value:

$$p = 2 \cdot |1 - \Phi(|z|)| = 2 \cdot |1 - \Phi(2.8682)| = 2 \cdot |1 - 0.9979| = 0.0041$$

- Since $p < 0.05$ we reject the null hypothesis and conclude that it is very unlikely that the true parameter is $\lambda = 0.060$.

Geiger–Marsden Experiment

- **95% confidence interval:**

$$\lambda_{-} = \frac{1}{t}\left[x + \frac{1.96^2}{2} - 1.96\sqrt{x + \frac{1.96^2}{4}}\right] = \frac{1}{187776}\left[11571 + \frac{1.96^2}{2} - 1.96\sqrt{11571 + \frac{1.96^2}{4}}\right] = 0.0605$$

$$\lambda_{+} = \frac{1}{t}\left[x + \frac{1.96^2}{2} + 1.96\sqrt{x + \frac{1.96^2}{4}}\right] = \frac{1}{187776}\left[11571 + \frac{1.96^2}{2} + 1.96\sqrt{11571 + \frac{1.96^2}{4}}\right] = 0.0628$$

- Note that since the hypothesized parameter (λ = 0.06) does not lie within the 95% confidence interval, we reject the null hypothesis.

Test Catalog for the Poisson Distribution

- **Statistical model:**
- $X \sim poisson(\lambda \cdot t)$
- Parameter estimate: $\hat{\lambda} = x/t$
- Where the observation is *x* = 'number of arrivals/events observed over a period of time *t*'

- **Hypothesis test (two-tailed):**
- $H_0: \lambda = \lambda_0$
- $H_1: \lambda \neq \lambda_0$
- Test size: $z = \frac{x - \lambda \cdot t}{\sqrt{\lambda \cdot t}} \sim N(0,1)$
- Approximate p-value: $2 \cdot |1 - \Phi(|z|)|$

- **95% confidence interval:**
- $\lambda_- = \frac{1}{t}\left[x + \frac{1.96^2}{2} - 1.96\sqrt{x + \frac{1.96^2}{4}}\right]$
- $\lambda_+ = \frac{1}{t}\left[x + \frac{1.96^2}{2} + 1.96\sqrt{x + \frac{1.96^2}{4}}\right]$

Words and Concepts to Know

Chi-Square Distribution

Bernoulli Trial

Normal approximation

Average rate

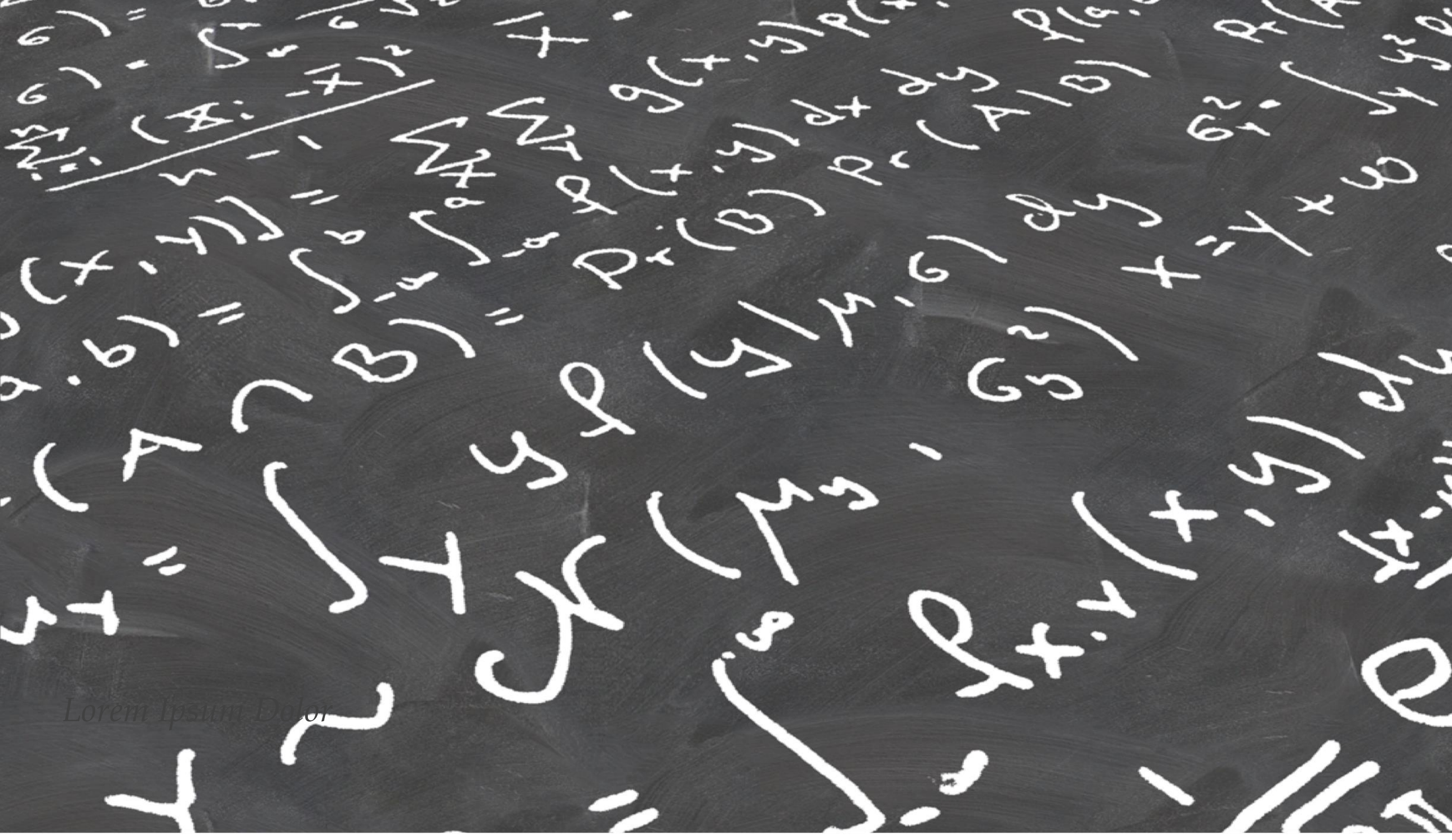
χ_k^2

Poisson Distribution

Binominal Distribution

Critical values

Chi-Square Test



Lorem Ipsum Dolor

12. Comparison of the Mean of Two Sample Sets

Gunvor Elisabeth Kirkelund
Lars Mandrup
Slides and material provided in parts by
Henrik Pedersen

Todays Content

- Repetition from last time
- Comparison the mean of two populations
  - With known variance
  - With unknown variance
- Paired and unpaired data.
- Scientific experimental test

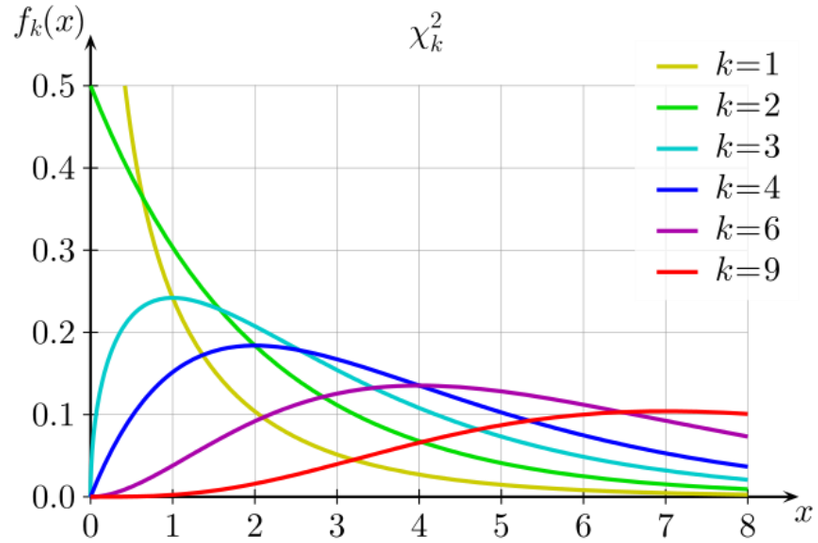
Chi-Square Distribution

- If we have a set of i.i.d. data $X_1, X_2, \ldots, X_n$ distributed according to:

$$X_i \sim \mathcal{N}(0,1)$$

- Then we have that: $\mathrm{Q} = \sum_{i=1}^{n} X_i^2$

  is χ_k^2 distributed with k degrees of freedom.



- χ^2 -test for independence: $X^2 = \sum_{i=1}^{n} \frac{(observed - expected)^2}{Expected}$ How well does the observed data fits the expected values

- χ^2 -test for variance: Test statistics $T = (N-1) \cdot \frac{s^2}{\sigma_0^2}$ Hypothesis test of the variance σ_0^2

The Binomial Distribution

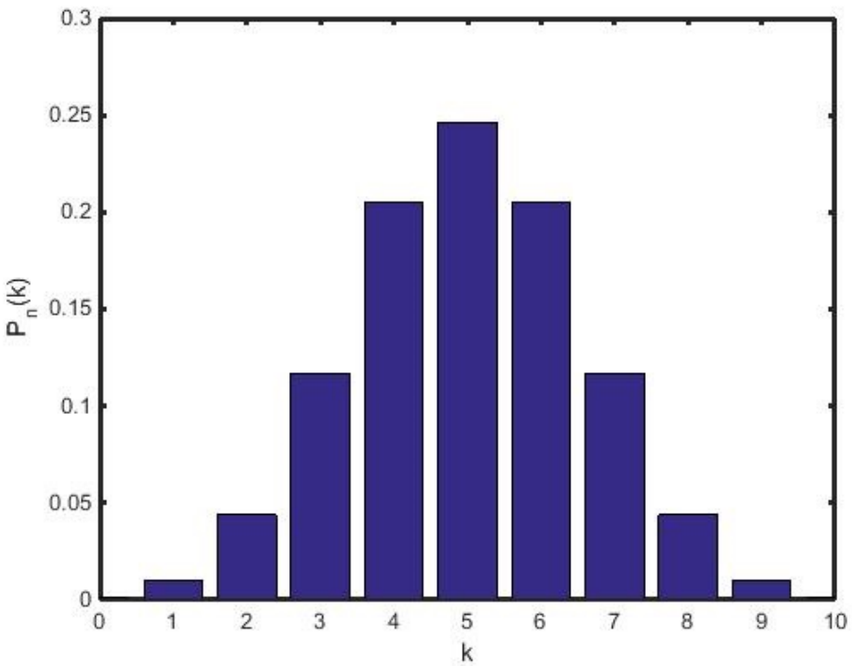
- We have n repeated trials.
- Each trial has two possible outcomes *Bernoulli Event*
  - **Success** — probability p
  - **Failure** — probability 1-p
- We write the mass function as:

  X =Number of successes in n trials

$$Pr(X = k) = f(k|n,p)$$

$$f(k|n,p) = \frac{n!}{k!\,(n-k)!}p^k(1-p)^{n-k}$$

$$= \binom{n}{k}p^k(1-p)^{n-k}$$



Test catalog for the Binomial Distribution

- **Statistical model:**
- $X \sim binomial(n, p)$
- Parameter estimate: $\hat{p} = x/n$
- Where the observation is *x* = 'number of successes out of *n* trials'

- **Hypothesis test (two-tailed):**
- $H_0: p = p_0$
- $H_1: p \neq p_0$
- Test size: $z = \frac{x - np_0}{\sqrt{np_0(1-p_0)}} \sim N(0,1)$
- Approximate p-value: $2 \cdot |1 - \Phi(|z|)|$

- **95% confidence interval:**

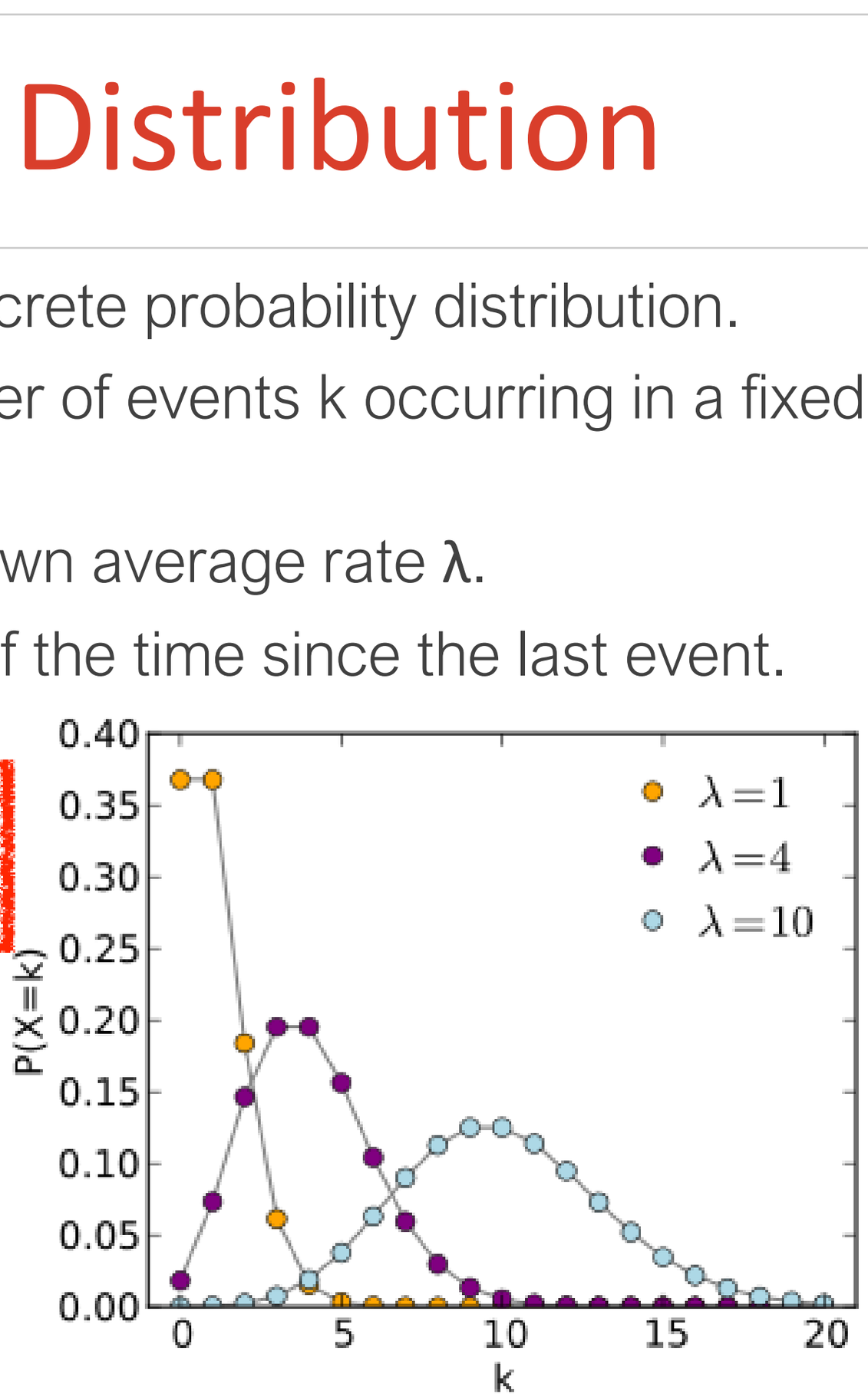
- $p_- = \frac{1}{n+1.96^2}\left[x + \frac{1.96^2}{2} - 1.96\sqrt{\frac{x(n-x)}{n} + \frac{1.96^2}{4}}\right]$
- $p_+ = \frac{1}{n+1.96^2}\left[x + \frac{1.96^2}{2} + 1.96\sqrt{\frac{x(n-x)}{n} + \frac{1.96^2}{4}}\right]$

The Poisson Distribution

- The Poisson distribution is a discrete probability distribution.
- The probability of a given number of events k occurring in a fixed interval of time t.
- If these events occur with a known average rate λ.
- And events are independently of the time since the last event.

- $$Pr(X = k) = \frac{(t \cdot \lambda)^k}{k!} e^{-t \cdot \lambda} = \frac{\gamma^k}{k!} e^{-\gamma}$$

- where $\gamma = t \cdot \lambda$ are the expected number of events in time interval t.



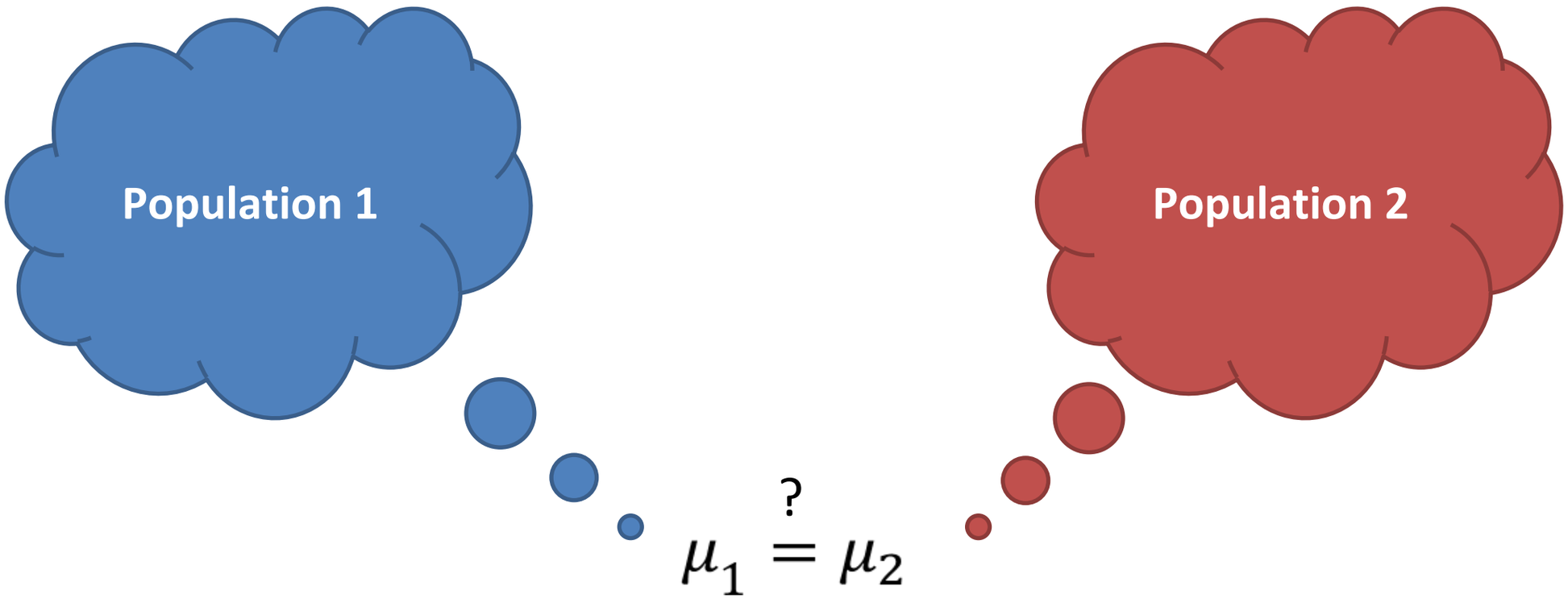
Test Catalog for the Poisson Distribution

- **Statistical model:**
- $X \sim poisson(\lambda \cdot t)$
- Parameter estimate: $\hat{\lambda} = x/t$
- Where the observation is *x* = 'number of arrivals/events observed over a period of time *t*'

- **Hypothesis test (two-tailed):**
- $H_0: \lambda = \lambda_0$
- $H_1: \lambda \neq \lambda_0$
- Test size: $z = \frac{x - \lambda \cdot t}{\sqrt{\lambda \cdot t}} \sim N(0,1)$
- Approximate p-value: $2 \cdot |1 - \Phi(|z|)|$

- **95% confidence interval:**
- $\lambda_- = \frac{1}{t}\left[x + \frac{1.96^2}{2} - 1.96\sqrt{x + \frac{1.96^2}{4}}\right]$
- $\lambda_+ = \frac{1}{t}\left[x + \frac{1.96^2}{2} + 1.96\sqrt{x + \frac{1.96^2}{4}}\right]$

Comparing two population means



- Fx. The height of people from Funen (μ_1) and Jutland (μ_2)

Statistical Model

- Suppose we have two populations with samples $X_{11}, X_{12}, \dots, X_{1n_1}$ drawn from a normally distributed population

$$X_{1i} \sim N\left(\mu_1, \sigma_1^2\right), i = 1,2, \dots, n_1$$

- and samples $X_{21}, X_{22}, \dots, X_{2n_2}$ drawn from a second normally distributed population

$$X_{2i} \sim N\left(\mu_2, \sigma_2^2\right)\ i = 1,2, \dots, n_2$$

The exact models not important

- Statistical question of interest:

$$\mu_1 = \mu_2?$$

- Note that in general, the two samples are of different size (i.e., $n_1 \neq n_2$).

Parameter Estimate

- This is the same as asking whether $\mu_1 - \mu_2 = 0$?
- From the central limit theorem, we know that the estimates of the two population means are

$$\hat{\mu}_1 = \bar{x}_1 = \frac{1}{n_1}\sum_{i=1}^{n_1} X_{1i} \sim N(\mu_1, \sigma_1^2/n_1)$$

$$\hat{\mu}_2 = \bar{x}_2 = \frac{1}{n_2}\sum_{i=1}^{n_2} X_{2i} \sim N(\mu_2, \sigma_2^2/n_2)$$

- The estimate of the difference between the two population means is

$$\bar{x}_1 - \bar{x}_2 \sim N\left(\mu_1 - \mu_2, \sigma_1^2/n_1 + \sigma_2^2/n_2\right)$$

- (the sum of two Gaussian PDFs is another Gaussian).

$E(aX + bY) = a(E(X) + bE(Y);$ $\quad Var(aX + bY) = a^2 Var(X) + b^2 Var(Y)$

Test Size

Known and identical variances!

- Here we will assume that the variances of the two populations are equal

$$\sigma^2 = \sigma_1^2 = \sigma_2^2$$

Assumption

- Under the null hypothesis H_0: $\boldsymbol{\delta = \mu_1 - \mu_2 = 0}$, we must have

$$\hat{\delta} = \bar{x}_1 - \bar{x}_2 \sim N\left(0, \sigma^2/n_1 + \sigma^2/n_2\right)$$

- Standardizing, we get

$$z = \frac{(\bar{x}_1 - \bar{x}_2) - E[\bar{x}_1 - \bar{x}_2]}{\sqrt{Var(\bar{x}_1 - \bar{x}_2)}} = \frac{(\bar{x}_1 - \bar{x}_2) - 0}{\sqrt{\sigma^2/n_1 + \sigma^2/n_2}} = \frac{\bar{x}_1 - \bar{x}_2}{\sigma\sqrt{1/n_1 + 1/n_2}} \sim N(0,1)$$

$$Var(aX + bY) = a^2 Var(X) + b^2 Var(Y)$$

Hypothesis Test for Comparing Two Population Means

with known and identical variances

- Suppose we observe sample means $\bar{x}_1 = 3$ and $\bar{x}_2 = 4$ from two normally distributed populations with standard deviation 1.
- Furthermore, let us assume that $n_1 = 10$ and $n_2 = 20$.
- Null hypothesis:

$$H_0: \mu_1 - \mu_2 = 0$$

Null Hypothesis

- Z-score

$$z = \frac{(\bar{x}_1 - \bar{x}_2)}{\sigma\sqrt{1/n_1 + 1/n_2}} = \frac{(\bar{x}_1 - \bar{x}_2)}{\sigma\sqrt{1/n_1 + 1/n_2}} = \frac{3 - 4}{1 \cdot \sqrt{1/10 + 1/20}} = -2.582$$

- P-value

$$2 \cdot \big(1 - \Phi(|z|)\big) = 2 \cdot \big(1 - \Phi(2.582)\big) = 2 \cdot (1 - 0.9951) = 0.0098$$

- Since p<0.05, we reject the null hypothesis and conclude that the two populations do not have identical mean values.

for the difference between two population means

95% Confidence Interval

Known and identical variances!

- True difference

$$\delta = \mu_1 - \mu_2$$

Statistical parameter in question

- Estimator

$$\hat{\delta} = \bar{x}_1 - \bar{x}_2 \sim N\left(\mu_1 - \mu_2, \sigma_1^2/n_1 + \sigma_2^2/n_2\right)$$

- To find the 95% confidence interval of δ, we need to find limits δ_- and δ_+ such that

$$\Pr(\delta_- \leq \delta \leq \delta_+) = 0.95$$

- Standardizing, we get

$$z = \frac{(\bar{x}_1 - \bar{x}_2) - \delta}{\sigma\sqrt{1/n_1 + 1/n_2}} \sim N(0,1)$$

for the difference between two population means

95% Confidence Interval

Known and identical variances!

- Same trick as in the previous chapters

$$0.95 = \Pr(-1.96 \leq Z \leq 1.96) = \Pr\left(-1.96 \leq \frac{(\bar{x}_1 - \bar{x}_2) - \delta}{\sigma\sqrt{1/n_1 + 1/n_2}} \leq 1.96\right)$$

- Isolating, we get

$$0.95 = \Pr\Big((\bar{x}_1 - \bar{x}_2) - 1.96 \cdot \sigma\sqrt{1/n_1 + 1/n_2} \leq \delta \leq (\bar{x}_1 - \bar{x}_2) + 1.96 \cdot \sigma\sqrt{1/n_1 + 1/n_2}\Big)$$

- Hence, the endpoints of the 95% confidence interval are

$$\delta_- = (\bar{x}_1 - \bar{x}_2) - 1.96 \cdot \sigma\sqrt{1/n_1 + 1/n_2}$$
$$\delta_+ = (\bar{x}_1 - \bar{x}_2) + 1.96 \cdot \sigma\sqrt{1/n_1 + 1/n_2}$$

for the difference between two population means

95% Confidence Interval

Known and identical variances!

- We observe sample means $\bar{x}_1 = 3$ and $\bar{x}_2 = 4$ from two normally distributed populations with standard deviation 1, $n_1 = 10$, and $n_2 = 20$.

- The endpoints of the 95% confidence interval for the true difference between the population means (δ), are

Lower bound:

$$\delta_- = (\bar{x}_1 - \bar{x}_2) - 1.96 \cdot \sigma \sqrt{\frac{1}{n_1} + \frac{1}{n_2}} = (3 - 4) - 1.96 \cdot 1 \cdot \sqrt{\frac{1}{10} + \frac{1}{20}} = -1.7591$$

Upper bound:

$$\delta_+ = (\bar{x}_1 - \bar{x}_2) + 1.96 \cdot \sigma \sqrt{\frac{1}{n_1} + \frac{1}{n_2}} = (3 - 4) + 1.96 \cdot 1 \cdot \sqrt{\frac{1}{10} + \frac{1}{20}} = -0.2409$$

- Since $\delta = 0$ is not included in the 95% confidence interval, we reject the null hypothesis stating that the population means are equal, i.e., $H_0: \mu_1 - \mu_2 = 0$

Test catalog for Comparing Two Means (known variance)

Statistical model:

- $X_{1i} \sim N\left(\mu_1, \sigma_1^2\right), i = 1,2, \dots, n_1$ and $X_{2i} \sim N\left(\mu_2, \sigma_2^2\right)\ i = 1,2, \dots, n_2$
- Parameter estimate: $\hat{\delta} = \bar{x}_1 - \bar{x}_2 \sim N\left(\mu_1 - \mu_2, \sigma_1^2/n_1 + \sigma_2^2/n_2\right)$
- Where the observation is $\bar{x}_1 - \bar{x}_2$ = 'the difference between two sample means'.

Hypothesis test (two-tailed):

- $H_0: \mu_1 = \mu_2$
- $H_1: \mu_1 \neq \mu_2$
- Test size: $z = \frac{(\bar{x_1} - \bar{x_2})}{\sigma\sqrt{1/n_1 + 1/n_2}} \sim N(0,1)$
- Approximate p-value: $2 \cdot |1 - \Phi(|z|)|$

95% confidence interval:

- $\delta_- = (\bar{x}_1 - \bar{x}_2) - 1.96 \cdot \sigma\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}$
- $\delta_+ = (\bar{x}_1 - \bar{x}_2) + 1.96 \cdot \sigma\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}$

also called pooled variance

Sample Variance

Unknown variances!

- In the case, where the true variance is unknown we have to estimate it.
- The formula for the pooled variance estimator is

Pooled variance

$$s^2 = \frac{1}{n_1 + n_2 - 2}\left((n_1 - 1)s_1^2 + (n_2 - 1)s_2^2\right)$$

weighted mean of s_1 and s_2

- where

Sample variance for 1: $s_1^2 = \frac{1}{n_1 - 1}\sum_{i=1}^{n_1}(x_{1i} - \bar{x}_1)^2$

Sample variance for 2: $s_2^2 = \frac{1}{n_2 - 1}\sum_{i=1}^{n_2}(x_{2i} - \bar{x}_2)^2$

Test Size

Unknown variances!

- The effect of using the empirical variance instead of the true variance is that we have to use the t-score instead of the z-score

$$t = \frac{\bar{x}_1 - \bar{x}_2 - \delta}{s\sqrt{1/n_1 + 1/n_2}} \sim t(n_1 + n_2 - 2)$$

- where s is the empirical standard deviation. *Pooled variance*
- Notice that the number of degrees of freedom for the t distribution is $n_1 + n_2 - 2$.

Hypothesis Test and Confidence Interval for Comparison of two Means with Unknown and Identical Variances

- Like before, suppose we observe sample means $\bar{x}_1 = 3$ and $\bar{x}_2 = 4$ from two normally distributed populations with empirical variances

$$s_1^2 = 1.4 \text{ and } s_2^2 = 1.1$$

- where $n_1 = 10$ and $n_2 = 20$. Assuming equal variances, the pooled estimate of the variance is

$$s^2 = \frac{1}{n_1 + n_2 - 2}\left((n_1 - 1)s_1^2 + (n_2 - 1)s_2^2\right)$$
$$= \frac{1}{10 + 20 - 2}\left((10 - 1) \cdot 1.4 + (20 - 1) \cdot 1.1\right) = 1.1964$$

- and the empirical standard deviation is $\sqrt{1.1964} = 1.0938$.

Hypothesis Test and Confidence Interval for Comparison of two Means with Unknown and Identical Variances

- Null hypothesis:

$$H_0: \delta = \mu_1 - \mu_2 = 0$$

- t-score

$$t = \frac{(\bar{x}_1 - \bar{x}_2) - \delta}{s\sqrt{1/n_1 + 1/n_2}} = \frac{3 - 4}{1.0938 \cdot \sqrt{1/10 + 1/20}}$$
$$= -2.3606 \sim t(n_1 + n_2 - 2)$$

degrees of freedom

- P-value

$$2 \cdot \left(1 - t_{cdf}(|t|, n_1 + n_2 - 2)\right) = 2 \cdot \left(1 - t_{cdf}(2.3606, 30 - 2)\right)$$
$$= 2 \cdot (1 - 0.9873) = 0.0254$$

- Since p<0.05, we reject the null hypothesis and conclude that the two populations do not have identical mean values.

Hypothesis Test and Confidence Interval for Comparison of two Means with Unknown and Identical Variances

- The endpoints of the 95% confidence interval for the true difference between the population means (δ), are

Lower bound:

$$\delta_- = (\bar{x}_1 - \bar{x}_2) - t_0 \cdot s\sqrt{\frac{1}{n_1} + \frac{1}{n_2}} = (3-4) - 2.0484 \cdot 1.0938 \cdot \sqrt{\frac{1}{10} + \frac{1}{20}}$$

$$= -1.8678$$

Upper bound:

$$\delta_+ = (\bar{x}_1 - \bar{x}_2) + t_0 \cdot s\sqrt{\frac{1}{n_1} + \frac{1}{n_2}} = (3-4) + 2.0484 \cdot 1.0938 \cdot \sqrt{\frac{1}{10} + \frac{1}{20}}$$

$$= -0.1322$$

- where t0 = tinv(0.975,n1+n2-2) = tinv(0.975,30-2) = 2.0484.
- Since $\delta = 0$ is not included in the 95% confidence interval, we reject the null hypothesis stating that the population means are equal.

Test Catalog for Comparing Two Means (unknown variance)

Statistical model:

- $X_{1i} \sim N\left(\mu_1, \sigma_1^2\right), i = 1,2,\dots,n_1$ and $X_{2i} \sim N\left(\mu_2, \sigma_2^2\right)\ i = 1,2,\dots,n_2$
- Parameter estimate:

$$\hat{\delta} = \bar{x}_1 - \bar{x}_2 \sim N\left(\mu_1 - \mu_2, \frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}\right)$$

$$s^2 = \frac{1}{n_1+n_2-2}\left((n_1-1)s_1^2 + (n_2-1)s_2^2\right)$$

- Where the observation is $\dot{x_1} - \dot{x_2}$ = 'the difference between two sample means'.

Hypothesis test (two-tailed):

- $H_0: \mu_1 = \mu_2$
- $H_1: \mu_1 \neq \mu_2$
- Test size: $t = \frac{(\dot{x_1}-\dot{x_2})}{s\sqrt{1/n_1+1/n_2}} = \sim t(n_1+n_2-2)$
- Approximate p-value: $2 \cdot \left(1 - t_{cdf}(|t|, n_1+n_2-2)\right)$

95% confidence interval:

- $\delta_- = (\dot{x_1} - \dot{x_2}) - t_0 \cdot s\sqrt{\frac{1}{n_1}+\frac{1}{n_2}}$
- $\delta_+ = (\dot{x_1} - \dot{x_2}) + t_0 \cdot s\sqrt{\frac{1}{n_1}+\frac{1}{n_2}}$
- where t0 = tinv(1-0.05/2,n1+n2-2)

OBS:
t-test (compared with Z-test)

- *Less knowledge*
- *Larger uncertaincy*
- *Confidence interval larger*
- *More difficult to reject H_0*

Paired vs. Unpaired Tests

- The best way to look at the effect of, say, a medical treatment is to measure some physiological parameter *in the same patient* before and after treatment.
- Consider the alternative; one group of patients gets the treatment, the other group doesn't.
- Comparing the mean of the physiological parameter between the two groups could be both due to the treatment *and* differences between the patient groups.
- If a difference was detected, there would be no way of telling whether that difference was due to the treatment or differences between the patient groups.

Paired Difference Test

Each point in one data-set correspond to a point in the other

- Paired difference tests are often used to compare "before" and "after" scores in experiments to determine whether significant change has occurred.
- The data are paired.
  - If $X_{11}, X_{12}, \ldots, X_{1n}$ and $X_{21}, X_{22}, \ldots, X_{2n}$ are the two samples, then X_{1i} corresponds to X_{2i}.

Stalk	1	2	3	4	5	6	7	8	9	10
Before height	35.5	31.7	31.2	36.3	22.8	28.0	24.6	26.1	34.5	27.7
After height	45.3	36.0	38.6	44.7	31.4	33.5	28.8	35.8	42.9	35.0

Corn stalk height before and after using a fertilizer.

sample size n1=n2

Paired Test

Statistical Model

- For paired samples, we look at the difference $d_i = X_{1i} - X_{2i}$ and make the assumption that

$$d_i \sim N(\delta, \sigma^2), i = 1,2, \dots, n$$

- where δ is the true (unknown) difference between X_1 and X_2.
- Estimator

Maximum Likelihood:

$$\hat{\delta} = \bar{d} = \frac{1}{n}\sum_{i=1}^{n}(X_{1i} - X_{2i})$$

- Since the individual terms in the sum are normally distributed, it follows from the central limit theorem that

$$\bar{d} \sim N(\delta, \sigma^2/n)$$

Test Size

- In general, we cannot assume that we know the true variance (σ^2), so we will have to estimate it.
- We will use the unbiased estimate of the variance,

Sample variance:

$$s_d^2 = \frac{1}{n-1}\sum_{i=1}^{n}(d_i - \bar{d})^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_{1i} - X_{2i} - \bar{d})^2$$

- Under the null hypothesis, $H_0: \delta = \delta_0$, the standardized test size is

$$t = \frac{\bar{d} - \delta_0}{s_d/\sqrt{\mathrm{n}}} \sim t(n-1)$$

Distributed according to the students t-distribution, because the variance is unknown

Does Fertiliser have an Effect on Corn Growth?

- A farmer decides to try out a new fertiliser on a test plot containing 10 stalks of corn.
- Before applying the fertiliser, he measures the height of each stalk.
- Two weeks later, he measures the stalks again, being careful to match each stalk's new height to its previous one.
- The stalks would have grown an average of 6 inches during that time even without the fertiliser.
- Did the fertiliser change the expected mean?

Stalk	1	2	3	4	5	6	7	8	9	10
Before height	35.5	31.7	31.2	36.3	22.8	28.0	24.6	26.1	34.5	27.7
After height	45.3	36.0	38.6	44.7	31.4	33.5	28.8	35.8	42.9	35.0
d_i	9.8	4.3	7.4	8.4	8.6	5.5	4.2	9.7	8.4	7.3

Does Fertiliser have an Effect on Corn Growth?

- The null hypothesis is

$$H_0: \delta = 6$$ *Mean difference in height*

- If the fertilizer is thought to increase corn growth, the correct alternative hypothesis is $H_1: \delta > 6$, which would result in a one-tailed p-value.
- However, for simplicity let us just choose the two-tailed test, $H_1: \delta \neq 6$.

It would be possible to include a second alternative hypothesis, saying δ<6

Does Fertiliser have an Effect on Corn Growth?

- The average difference in stalk height before and after is

ML estimator $\bar{d} = \frac{1}{n}\sum_{i=1}^{n}(X_{1i} - X_{2i}) = 7.36.$

- The estimated variance is

Unbiased ML estimator $s_d^2 = \frac{1}{n-1}\sum_{i=1}^{n}(d_i - \bar{d})^2 = 4.216$

- and the estimated standard deviation becomes $s_d = \sqrt{4.216} = 2.05$.

Does Fertiliser have an Effect on Corn Growth?

- Test size

$$t = \frac{\bar{d} - \delta}{s_d/\sqrt{n}} = \frac{7.36 - 6}{2.05/\sqrt{10}} = 2.09 \sim t(10 - 1)$$

- P-value

$$2 \cdot \left(1 - t_{cdf}(|t|, n - 1)\right) = 2 \cdot \left(1 - t_{cdf}(2.09, 10 - 1)\right) = 2 \cdot (1 - 0.9669)$$
$$= 0.0662$$

The p-value is very small, so it is not likely, but we cannot reject the null-hypothesis.

- Since p>0.05 the null hypothesis cannot be rejected. Hence, the test does not provide evidence that the fertilizer caused the corn to grow more or less than if it had not been fertilized.
- Note: A one-tailed test would have resulted in the opposite conclusion (p<0.05).

Because the 5% is placed in one tail instead of dividing it between two.

Paired Test

95% Confidence Interval

- Recall that the estimator of δ is

ML estimator $$\hat{\delta} = \bar{d} = \frac{1}{n}\sum_{i=1}^{n}(X_{1i} - X_{2i}) \sim N(\delta, \sigma^2/n)$$

- After standardizing it is straight forward to show that the endpoints of the 95% confidence interval are

$$\delta_{-} = \bar{d} - t_0 \cdot s_d/\sqrt{n} \quad = 5.89$$

$$\delta_{+} = \bar{d} + t_0 \cdot s_d/\sqrt{n} \quad = 8.83$$

- where t_0 = *tinv*(0.975,*n*-1).

Which Sowing Machine is the Better One?

- In an agricultural research study from 1934, two sowing machines were compared in terms of the yield after harvesting.
- A total of twenty fields of equal size were sowed; ten fields with machine 1 and ten fields with machine 2.
- The fields were paired such that the two fields in a pair were neighbours.
- One field in a pair was sowed with machine 1 and the other field was sowed with machine 2.
- By pairing and sowing the fields in this way, potential field-effects could be removed.
- Hence, any differences in yield between two paired fields could be attributed to a difference between the machines (not a difference between the fields).

Which Sowing Machine is the Better One? (Paired test)

Field	*Machine 1*	*Machine 2*	*Difference* $d_i = x_{1i} - x_{2i}$	$(d_i - \bar{x})^2$
1	8.0	5.6	2.4	2,46
2	8.4	7.4	1.0	0,03
3	8.0	7.3	0.7	0,02
4	6.4	6.4	0.0	0,69
5	8.6	7.5	1.1	0,07
6	7.7	6.1	1.6	0,59
7	7.7	6.6	1.1	0,07
8	5.6	6.0	−0.4	1,51
9	5.6	5.5	0.1	0,53
10	6.2	5.5	0.7	0,02
$\bar{x}$	7.22	6.39	0.83	
s^2	1.33	0.62	0.67	

- Paired test, the fields are assumed to be the same.

Which Sowing Machine is the Better One? (Paired test)

Hypothesis testing

- The null hypothesis in this experiment states that there is no difference between the two machines:

$$H_0: \delta = 0$$

- and the alternative hypothesis states that there is a difference

$$H_1: \delta \neq 0$$

- If the data suggest that there is indeed a difference between the two machines, the p-value should be smaller than, say, 0.05.

Which Sowing Machine is the Better One? (Paired test)

- Test size

Variance unknown, thus t distribution

$$t = \frac{\bar{d} - \delta}{s_d/\sqrt{n}} = \frac{0.83 - 0}{0.8166/\sqrt{10}} \sim t(10-1)$$

t = 3.214

- P-value

Remember p-value is a probability.

$$2 \cdot \left(1 - t_{cdf}(|t|, n-1)\right) = 0.0106$$

- Since p<0.05, we reject the null hypothesis that there is no difference between the yield of the two machines.

- The average difference $\bar{d}$ is positive, and we conclude that the data suggest that machine 1 outperforms machine 2.

A positive difference can only be concluded on, after the hypothesis testing.

Which Sowing Machine is the Better One? (Paired test)

- The endpoints of the 95% confidence interval δ for are

$$\delta_{-} = \overline{d} - t_0 \cdot \frac{s_d}{\sqrt{n}} = 0.2459$$

$$\delta_{+} = \overline{d} + t_0 \cdot \frac{s_d}{\sqrt{n}} = 1.4141$$

- Since $\delta = 0$ is not included in the 95% confidence interval, we reject the null hypothesis.

Paired vs. Unpaired Test

Consider what would happen if we performed an *unpaired* comparison between the two sowing machines.

Which Sowing Machine is the Better One? (Unpaired test)

- The sample means are $\bar{x}_1 = 7.22$ and $\bar{x}_2 = 6.39$, and the empirical variances are $\bar{s}_1^2 = 1.33$ and $\bar{s}_2^2 = 0.62$.

- With $n_1 = n_2 = 10$, the pooled variance is:

Pooled sample variance

$$s^2 = \frac{1}{10 + 10 - 2}\big((10 - 1) \cdot 1.33 + (10 - 1) \cdot 0.62\big) = 0.97$$

- And the empirical standard deviation is: $s = \sqrt{0.97} = 0.99$

Which Sowing Machine is the Better One? (Unpaired test)

- Null hypothesis: $H_0: \mu_1 - \mu_2 = 0$
- t-score

$$t = \frac{(\bar{x}_1 - \bar{x}_2)}{s\sqrt{1/n_1 + 1/n_2}} = \frac{7.22 - 6.39}{0.99 \cdot \sqrt{1/10 + 1/10}} = 1.88 \sim t(n_1 + n_2 - 2)$$

unknown variance.

- P-value

$$2 \cdot \left(1 - t_{cdf}(|t|, n_1 + n_2 - 2)\right) = 2 \cdot \left(1 - t_{cdf}(1.88\,, 20 - 2)\right)$$
$$= 2 \cdot (1 - 0.9619) = 0.076$$

p-value is still small.

- Since p>0.05, we fail to reject the null hypothesis and conclude that the two machines are <u>not</u> different!

Which Sowing Machine is the Better One? (Unpaired test)

- We see that the conclusions drawn with the paired test and the unpaired test are contradictory.
  - The paired test suggests a difference between the two sowing machines.
  - The unpaired test does not.
- Explanation:
  - In the unpaired test, the difference between the two population means is due to a combination of machine effects and field effects.
  - By pairing the data, such that we look at the difference between the two machines in similar fields, the field-effects are removed, and we can detect a difference between the machines.

Variance for estimate decreases

Paired vs. Unpaired test

- The best way to look at the effect of a medical treatment is to measure some physiological parameter *in the same patient* before and after treatment.
- Consider the alternative; one group of patients gets the treatment, another group does not.
- Comparing the mean of the physiological parameter between the two groups could be both due to the treatment *and* differences between the patient groups.
- If a difference was detected, there would be no way of telling whether that difference was due to the treatment or differences between the patient groups.

Paired vs. Unpaired test

- **Unpaired:**
  - No one-to-one correspondance between X_1 and X_2 data
  - Sample size n_1 and n_2 could be different
  - Many different factors could influence the result
  - Larger uncertaincy
  - More difficult to reject the H_0 hypothesis

- **Paired:**
  - A one-to-one correspondance between X_1 and X_2 data
  - Sample size n_1 and n_2 equal
  - Elimination of factors not related to the test
  - Reducing uncertaincy
  - Easier to reject the H_0 hypothesis

Test Catalog for Paired Data

Statistical model:

- $d_i = X_{1i} - X_{2i}$, where $d_i \sim N(\delta, \sigma^2), i = 1,2, \dots, n$
- Parameter estimate:

$$\hat{\delta} = \overline{d} = \frac{1}{n}\sum_{i=1}^{n} X_{1i} - X_{2i}$$

$$s_d^2 = \frac{1}{n-1}\sum_{i=1}^{n}(d_i - \overline{d})^2$$

- Where the observation is $\overline{d}$ = 'the average of the differences between paired samples'.

Hypothesis test (two-tailed):

- $H_0: \delta = \delta_0$
- $H_1: \delta \neq \delta_0$
- Test size: $t = \frac{\overline{d} - \delta_0}{s_d/\sqrt{\mathrm{n}}} = \sim t(\mathrm{n}-1)$
- Approximate p-value: $2 \cdot \left(1 - t_{cdf}(|t|, \mathrm{n}-1)\right)$

95% confidence interval:

- $\delta_- = \overline{d} - t_0 \cdot \frac{s_d}{\sqrt{n}}$
- $\delta_+ = \overline{d} + t_0 \cdot \frac{s_d}{\sqrt{n}}$

- where t0 = tinv(1-0.05/2,n-1)

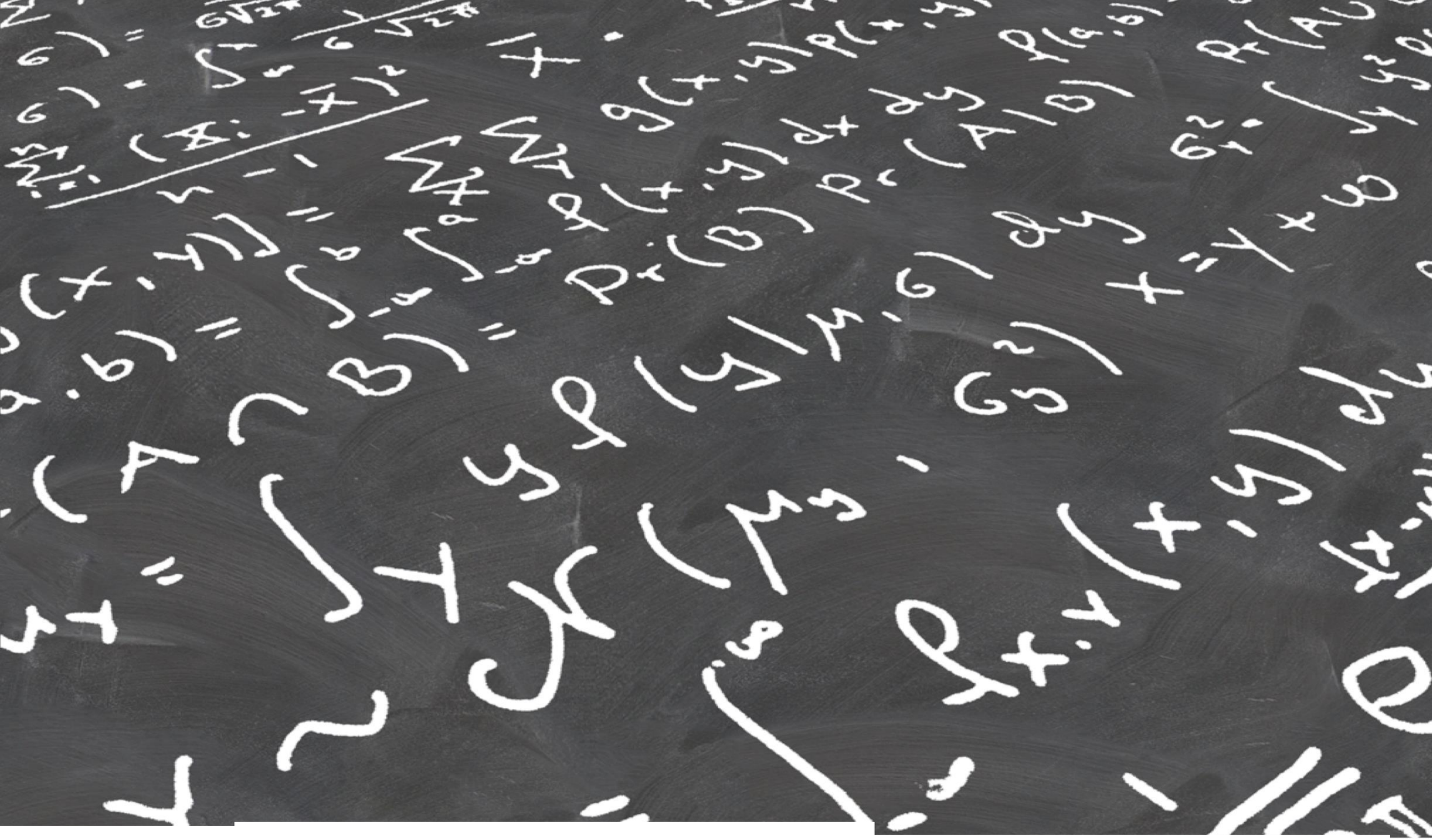
Words and Concepts to Know

Pooled variance

Paired test

Unpaired test

Comparing two population means

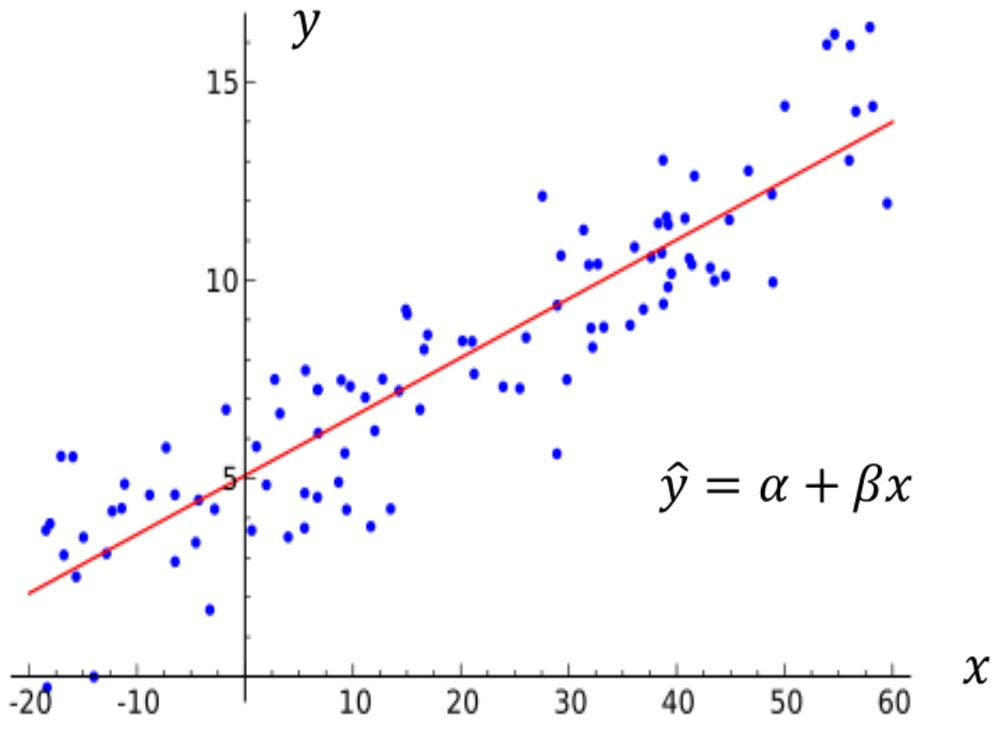


13. Linear Regression Models

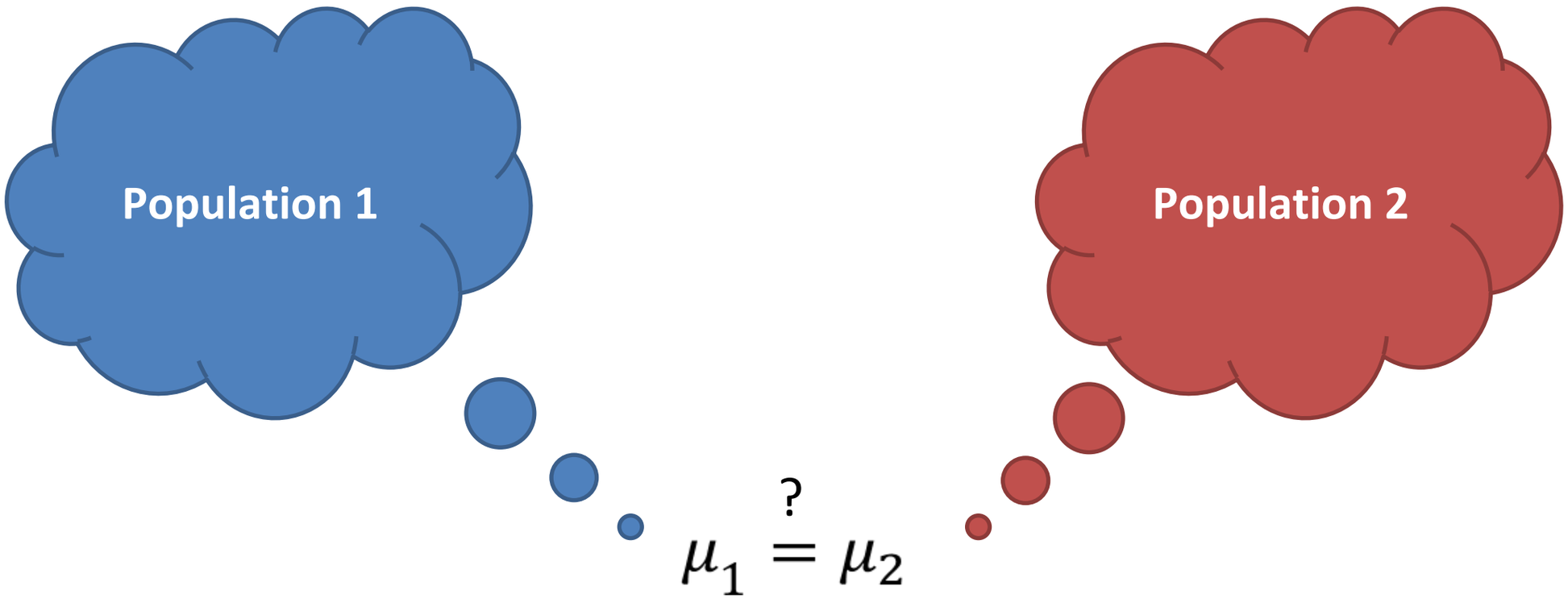
Gunvor Elisabeth Kirkelund
Lars Mandrup
Slides and material provided in parts by
Henrik Pedersen

Todays Content

- Repetition from last time
- Linear regression
- RANSAC



Comparing two population means



- Fx. The height of people from Funen (μ_1) and Jutland (μ_2)

Test catalog for Comparing Two Means (known variance)

Statistical model:

- $X_{1i} \sim N\left(\mu_1, \sigma_1^2\right), i = 1,2, \dots, n_1$ and $X_{2i} \sim N\left(\mu_2, \sigma_2^2\right)\ i = 1,2, \dots, n_2$
- Parameter estimate: $\hat{\delta} = \bar{x}_1 - \bar{x}_2 \sim N\left(\mu_1 - \mu_2, \sigma_1^2/n_1 + \sigma_2^2/n_2\right)$
- Where the observation is $\bar{x}_1 - \bar{x}_2$ = 'the difference between two sample means'.

Hypothesis test (two-tailed):

- $H_0: \mu_1 = \mu_2$
- $H_1: \mu_1 \neq \mu_2$
- Test size: $z = \frac{(\bar{x}_1 - \bar{x}_2)}{\sigma\sqrt{1/n_1 + 1/n_2}} \sim N(0,1)$
- Approximate p-value: $2 \cdot |1 - \Phi(|z|)|$

95% confidence interval:

- $\delta_- = (\bar{x}_1 - \bar{x}_2) - 1.96 \cdot \sigma\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}$
- $\delta_+ = (\bar{x}_1 - \bar{x}_2) + 1.96 \cdot \sigma\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}$

Test Catalog for Comparing Two Means (unknown variance)

Statistical model:

- $X_{1i} \sim N\left(\mu_1, \sigma_1^2\right), i = 1,2,\dots,n_1$ and $X_{2i} \sim N\left(\mu_2, \sigma_2^2\right)\ i = 1,2,\dots,n_2$
- Parameter estimate:

$$\hat{\delta} = \bar{x}_1 - \bar{x}_2 \sim N\left(\mu_1 - \mu_2, \frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}\right)$$

$$s^2 = \frac{1}{n_1+n_2-2}\left((n_1-1)s_1^2 + (n_2-1)s_2^2\right)$$

- Where the observation is $\dot{x_1} - \dot{x_2}$ = 'the difference between two sample means'.

Hypothesis test (two-tailed):

- $H_0: \mu_1 = \mu_2$
- $H_1: \mu_1 \neq \mu_2$
- Test size: $t = \frac{(\dot{x_1}-\dot{x_2})}{s\sqrt{1/n_1+1/n_2}} = \sim t(n_1+n_2-2)$
- Approximate p-value: $2 \cdot \left(1 - t_{cdf}(|t|, n_1+n_2-2)\right)$

95% confidence interval:

- $\delta_- = (\dot{x_1} - \dot{x_2}) - t_0 \cdot s\sqrt{\frac{1}{n_1}+\frac{1}{n_2}}$
- $\delta_+ = (\dot{x_1} - \dot{x_2}) + t_0 \cdot s\sqrt{\frac{1}{n_1}+\frac{1}{n_2}}$
- where t0 = tinv(1-0.05/2,n1+n2-2)

OBS:
t-test (compared with Z-test)

- *Less knowledge*
- *Larger uncertaincy*
- *Confidence interval larger*
- *More difficult to reject H_0*

Test Catalog for Paired Data

Statistical model:

- $d_i = X_{1i} - X_{2i}$, where $d_i \sim N(\delta, \sigma^2), i = 1,2,\dots,n$
- Parameter estimate:

$$\hat{\delta} = \overline{d} = \frac{1}{n}\sum_{i=1}^{n} X_{1i} - X_{2i}$$

$$s_d^2 = \frac{1}{n-1}\sum_{i=1}^{n}(d_i - \overline{d})^2$$

- Where the observation is $\overline{d}$ = 'the average of the differences between paired samples'.

Hypothesis test (two-tailed):

- $H_0: \delta = \delta_0$
- $H_1: \delta \neq \delta_0$
- Test size: $t = \frac{\overline{d} - \delta_0}{s_d/\sqrt{\mathrm{n}}} = \sim t(\mathrm{n}-1)$
- Approximate p-value: $2 \cdot \left(1 - t_{cdf}(|t|, \mathrm{n}-1)\right)$

95% confidence interval:

- $\delta_- = \overline{d} - t_0 \cdot \frac{s_d}{\sqrt{n}}$
- $\delta_+ = \overline{d} + t_0 \cdot \frac{s_d}{\sqrt{n}}$

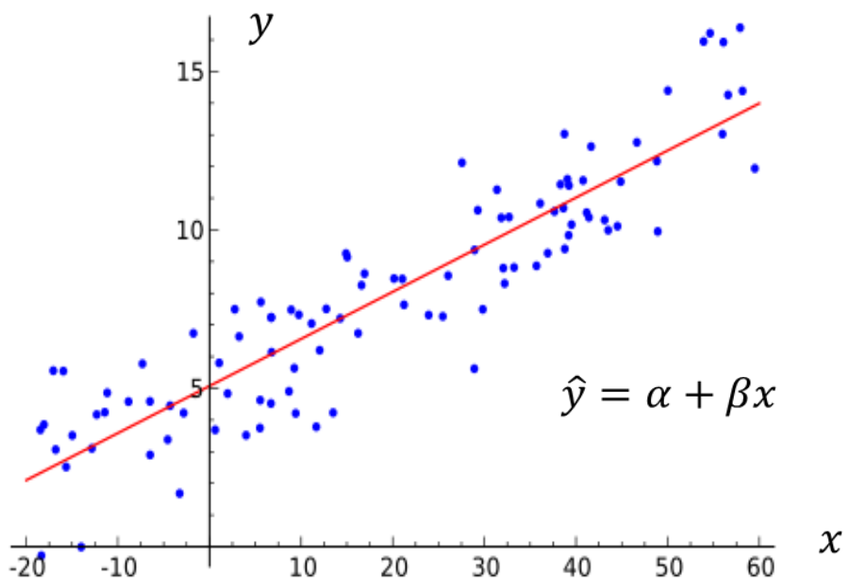
- where t0 = tinv(1-0.05/2,n-1)

Paired test (vs. unpaired test):

- *A one-to-one correspondance between X_1 and X_2 data*
- *Sample size n_1 and n_2 equal*
- *Elimination of factors not related to the test*
- *Reducing uncertaincy*
- *Easier to reject the H_o hopothesis*

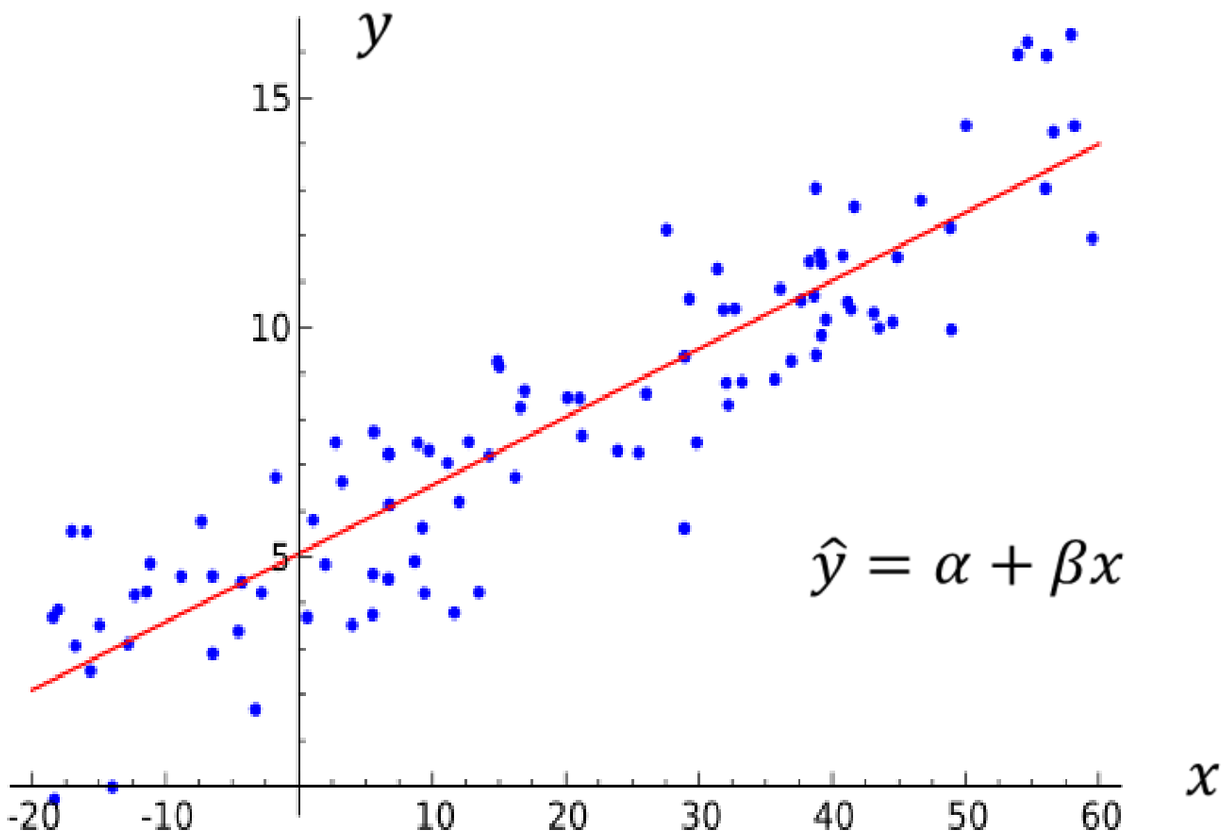
Linear Models – When/Why?

- Can be used when the mean changes over time.
- The variance should not change over time
- The mean is connected to time linearly!
- The simplest model – more advanced models not necesarily the best → start simpel (linear regression)



Linear Regression

- Fits a straight line through the set of n points (x_i, y_i)
- Make the sum of squared residuals ($\epsilon^2 = \sum_{i=1}^{n}(y_i - \hat{y}_i)^2 \sim \chi^2$ - the vertical distances between til points of data and the fitted line) of the model as small as possible



Statistical Model

- In linear regression, the data come in pairs

$$(x_i, y_i), \qquad \text{for } i = 1,2, \dots, n$$

fx. time t

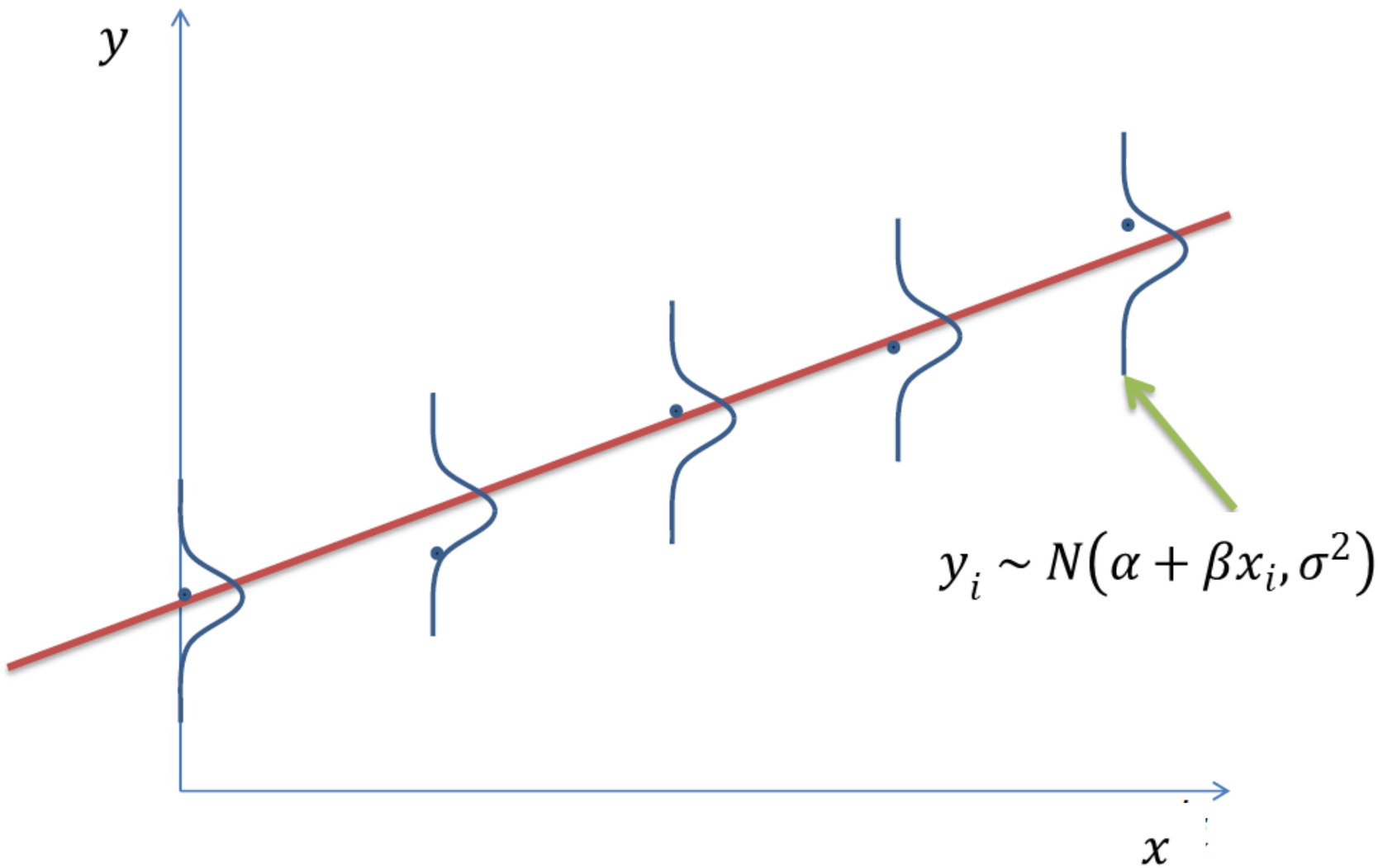
- where x is the independent variable and y is the dependent (or response) variable.
- Statistical model

$$y_i \sim N(\alpha + \beta x_i, \sigma^2)$$

- where β is the slope of the straight line and α it's intercept with the y-axis.

OBS! In ''Random Signals'': $\alpha \rightarrow b_0$ and $\beta \rightarrow b_1$

Statistical Model



Residual

- A residual is the difference between the measured and predicted data.
- The residual of the *i*'th sample (y_i) for a given choice of α and β is denoted ε_i and is given by

$$\epsilon_i = y_i - (\alpha + \beta x_i), \qquad \text{for } i = 1,2,\dots,n$$

measured

estimated ($\hat{y}_i$)

Empirical Variance

- Recall that $y_i \sim \mathcal{N}(\alpha + \beta x_i, \sigma^2)$
- The unbiased estimator of the variance is

$$s_r^2 = \frac{1}{n-2}\sum_{i=1}^{n}(y_i - (\hat{\alpha} + \hat{\beta}x_i))^2 = \frac{1}{n-2}\sum_{i=1}^{n}{\epsilon_i}^2$$

Two constrains $(\hat{\alpha}, \hat{\beta})$ → ÷ two degrees of freedom

- Statistical interence in linear regression concerns the parameter estimates: $\hat{\alpha}, \hat{\beta}$ and s_r^2.

Model Fitting

- The goal of linear regression is to determine the choice of slope (β) and intercept (α) that minimizes the sum of squared residuals of the model.

$$R(\alpha, \beta) = \sum_{i=1}^{n} (y_i - (\alpha + \beta x_i))^2 = \sum_{i=1}^{n} (y_i - \alpha - \beta x_i)^2 ,$$

$\epsilon_i^{\,2}$

- The parameter estimates that minimize *R* are

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

$$\hat{\alpha} = \bar{y} - \hat{\beta} \cdot \bar{x}$$

- where $\bar{x}$ is the average of $x_1, x_2, \ldots, x_n$ and $\bar{y}$ is the average of $y_1, y_2, \ldots, y_n$.

Derivation of the Intercept Parameter

- Partial derivative w.r.t. α and setting to zero:

$$\frac{\partial R(\alpha,\beta)}{\partial \alpha} = \frac{\partial}{\partial \alpha}\sum_{i=1}^{n}(y_i - \alpha - \beta x_i)^2 = -2\sum_{i=1}^{n}(y_i - \alpha - \beta x_i) = 0$$

- It follows that:

$$2n\alpha = 2\sum_{i=1}^{n} y_i - 2\beta\sum_{i=1}^{n} x_i = 2n\bar{y} - 2\beta n\bar{x}$$

$$\Downarrow$$

$$\hat{\alpha} = \bar{y} - \hat{\beta}\bar{x}$$

Derivation of the Slope Parameter

- Partial derivative w.r.t. β and setting to zero:

$$\frac{\partial R(\alpha,\beta)}{\partial \beta} = \frac{\partial}{\partial \beta}\sum_{i=1}^{n}(y_i - \alpha - \beta x_i)^2 = -2\sum_{i=1}^{n} x_i(y_i - \alpha - \beta x_i)$$
$$= -2\sum_{i=1}^{n}\left(x_i y_i - \alpha x_i - \beta x_i^2\right) = -2\sum_{i=1}^{n} x_i y_i + 2\alpha\sum_{i=1}^{n} x_i + 2\beta\sum_{i=1}^{n} x_i^2 = 0$$

- Inserting the result $\alpha = \bar{y} - \beta\bar{x}$ we get:

$$-2\sum_{i=1}^{n} x_i y_i + 2(\bar{y} - \beta\bar{x})\sum_{i=1}^{n} x_i + 2\beta\sum_{i=1}^{n} x_i^2$$
$$= -2\sum_{i=1}^{n} x_i(y_i - \bar{y}) + 2\beta\sum_{i=1}^{n} x_i(x_i - \bar{x})$$
$$= -2\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y}) + 2\beta\sum_{i=1}^{n}(x_i - \bar{x})^2 = 0$$

- It follows that:

$$\hat{\beta} = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{n}(x_i - \bar{x})^2} = \frac{\sum_{i=1}^{n}(x_i y_i - \bar{x}\,\bar{y})}{\sum_{i=1}^{n}(x_i - \bar{x})^2} = \frac{s_{xy}^2}{s_x^2}$$

- Where: $s_{xy}^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i y_i - \bar{x}\,\bar{y})$ and $s_x^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2$

Sample covariance

Sample variance

Example - Hubble's Law

- Hubble's law is the name for the observation in physical cosmology that objects observed in deep space are found to have a relative velocity away from the Earth that is approximately proportional to their distance from the Earth: $v = H \cdot x$
- Edwin Hubble's original measurements for 24 distant galaxies were (in Matlab notation)

```
Distance = [ 0.032 0.034 0.214 0.263 0.275 0.275 ...
             0.450 0.500 0.500 0.630 0.800 0.900 ...
             0.900 0.900 0.900 1.000 1.100 1.100 ...
             1.400 1.700 2.000 2.000 2.000 2.000 ];

Speed = [ 170 290 -130 -70 -185 -220 200  290 ...
          270 200  300 -30  650  150 500  920 ...
          450 500  500 960  500  850 800 1090 ];
```

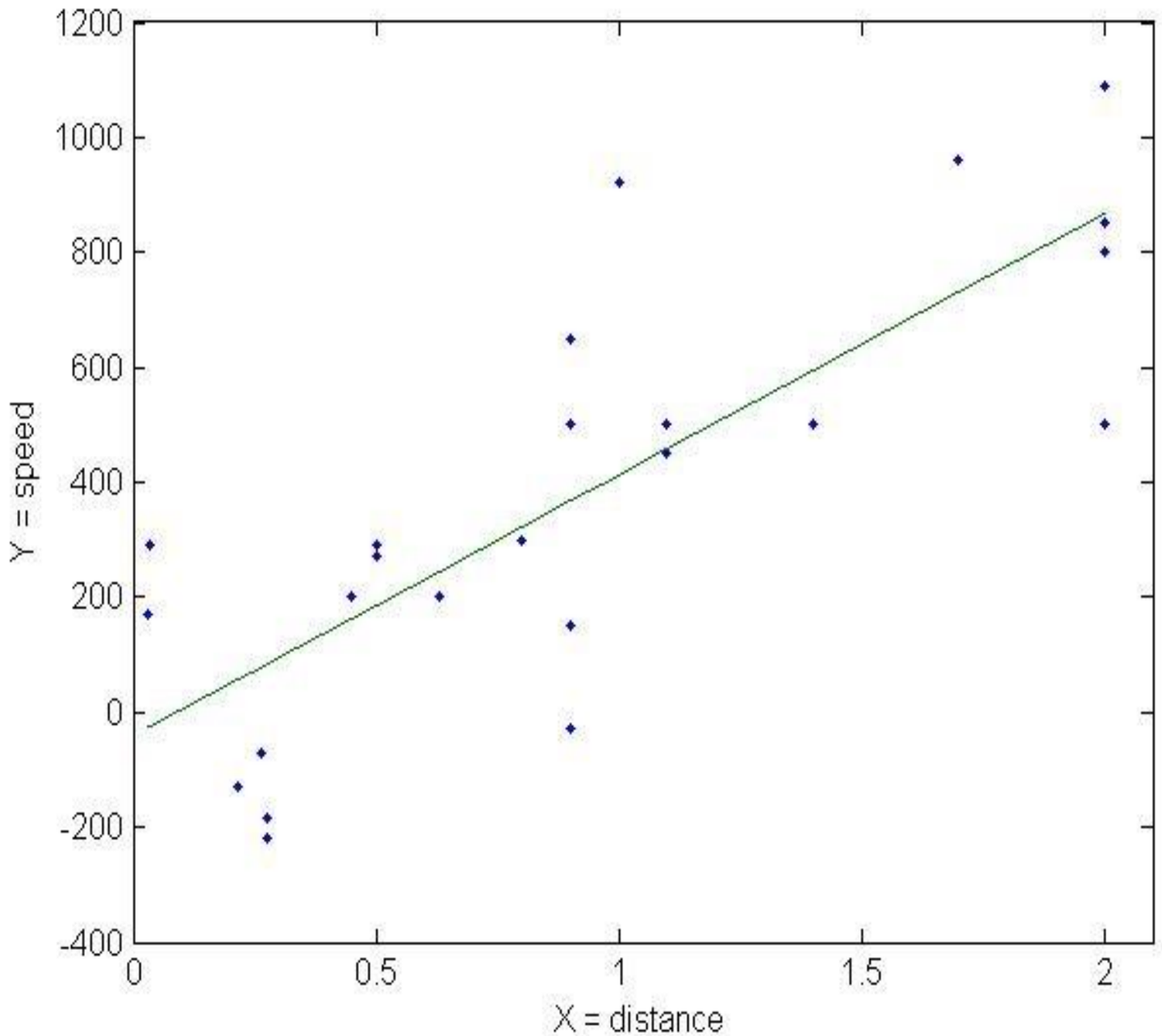
Example - Hubble's Law

- Choosing
  - x = Distance;
  - y = Speed;
- Slope estimate

$$\widehat{\beta} = \frac{\sum_{i=1}^{n}(x_i - \bar{x}) \cdot (y_i - \bar{y})}{\sum_{i=1}^{n}(x_i - \bar{x})^2} = 454.1584$$

- Intercept estimate

$$\widehat{\alpha} = \bar{y} - \widehat{\beta} \cdot \bar{x} = -40.7836$$



Statistical Inference on the Regression Slope

- In general, the null hypothesis about the slope that we wish to test takes the following form

$$H_0: \beta = \beta_0$$

- It can be shown that the estimator of the slope is normally distributed with mean β and variance

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

- where σ^2 is the variance used in the statistical model, $y_i \sim N(\alpha + \beta x_i, \sigma^2)$.
- Using the estimated variance, s_r^2, instead of the population variance, the appropriate test statistic for $\widehat{\beta}$ is

$$t = \frac{\widehat{\beta} - \beta_0}{s_r \sqrt{1/\sum_{i=1}^{n}(x_i - \bar{x})^2}} \sim t(n-2)$$

- The p-value is

$$2 \cdot \left(1 - t_{cdf}(|t|, n-2)\right)$$

Example - Hubble's Law

- Let us test whether the regression slope deviates significantly from zero.
- Null hypothesis

$$H_0: \beta = 0$$

- Parameter estimates:

$$\hat{\beta} = 454.1584 \quad \text{and} \quad \hat{\alpha} = -40.7836$$

- Empirical variance/s.d.

$$s_r^2 = \frac{1}{n-2}\sum_{i=1}^{n}\left(y_i - \left(\hat{\alpha} + \hat{\beta}x_i\right)\right)^2 = 54247 \text{ and } s_r = \sqrt{54247} = 232.91$$

- Test size:

$$t = \frac{\hat{\beta} - 0}{s_r\sqrt{1/\sum_{i=1}^{n}(x_i - \bar{x})^2}} = 6.0364$$

- p-value:

$$2 \cdot \left(1 - t_{cdf}(|t|, n-2)\right) \approx 0$$

- Since p<0.05, we reject the null hypothesis that $\beta = 0$. In other words, the data suggest that the regression slope deviates significantly from zero.

Statistical Inference on the Regression Slope

- The 95% confidence interval for the slope is

$$\beta_{-} = \hat{\beta} - t_0 \cdot \frac{s_r}{\sqrt{\sum_{i=1}^{n}(x_i - \overline{x})^2}} = \hat{\beta} - t_0 \cdot \frac{s_r}{s_x} \cdot \frac{1}{\sqrt{n-1}}$$

$$\beta_{+} = \hat{\beta} + t_0 \cdot \frac{s_r}{\sqrt{\sum_{i=1}^{n}(x_i - \overline{x})^2}} = \hat{\beta} + t_0 \cdot \frac{s_r}{s_x} \cdot \frac{1}{\sqrt{n-1}}$$

- where

$$t_0 = tinv\left(1 - \frac{0.05}{2}, n-2\right) = tinv(0.975, n-2)$$

Example - Hubble's law

- 95% confidence interval:

$$t_0 = tinv(0.975,22) = 2.0739$$

$$\beta_- = \hat{\beta} - t_0 \cdot \frac{s_r}{\sqrt{\sum_{i=1}^{n}(x_i - \overline{x})^2}} = 298.12$$

$$\beta_+ = \hat{\beta} + t_0 \cdot \frac{s_r}{\sqrt{\sum_{i=1}^{n}(x_i - \overline{x})^2}} = 610.19$$

- The NULL hypothesis H_0: $\beta = 0$ is not within the 95% confidence interval, so we reject the NULL hypothesis

Statistical Inference on the Regression Intercept

- In general, the null hypothesis that we wish to test takes the following form

$$H_0: \alpha = \alpha_0$$

- It can be shown that the estimator of the intercept is normally distributed with mean α and variance

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

- where σ^2 is the variance used in the statistical model, $y_i \sim N\left(\alpha + \beta x_i, \sigma^2\right)$.
- The appropriate test statistic for $\widehat{\alpha}$ is

$$t = \frac{\widehat{\alpha} - \alpha_0}{s_r \sqrt{\frac{1}{n} + \frac{\bar{x}^2}{\sum_{i=1}^{n} (x_i - \bar{x})^2}}} \sim t(n-2)$$

- The p-value is

$$2 \cdot \left(1 - t_{cdf}(|t|, n-2)\right)$$

Example - Hubble's Law

- Let us test whether the regression intercept deviates significantly from zero.
- Null hypothesis

$$H_0\colon\ \alpha = 0$$

- Parameter estimates are the same as above:
- Test size:

$$t = \frac{\hat{\alpha} - 0}{s_r \sqrt{\frac{1}{n} + \frac{\bar{x}^2}{\sum_{i=1}^{n}(x_i - \bar{x})^2}}} = -0.4888$$

- p-value:

$$2 \cdot \left(1 - t_{cdf}\left(|t|, n-2\right)\right) = 0.6298$$

- Since p>0.05, we fail to reject the null hypothesis that $\alpha = 0$. In other words, the data suggest that the regression intercept does not deviate significantly from zero.

Statistical Inference on the Regression Intercept

- The 95% confidence interval for the intercept α is:

$$\alpha_{-} = \hat{\alpha} - t_0 \cdot s_r \cdot \sqrt{\frac{1}{n} + \frac{\overline{x}^2}{\sum_{i=1}^{n}(x_i - \overline{x})^2}}$$

$$\alpha_{+} = \hat{\alpha} + t_0 \cdot s_r \cdot \sqrt{\frac{1}{n} + \frac{\overline{x}^2}{\sum_{i=1}^{n}(x_i - \overline{x})^2}}$$

- where

$$t_0 = tinv\left(1 - \frac{0.05}{2}, n - 2\right) = tinv(0.975, n - 2)$$

Example - Hubble's law

- 95% confidence interval:

$$\alpha_- = \hat{\alpha} - t_0 \cdot s_r \sqrt{\frac{1}{n} + \frac{\bar{x}^2}{\sum_{i=1}^{n}(x_i - \bar{x})^2}} = -124.2$$

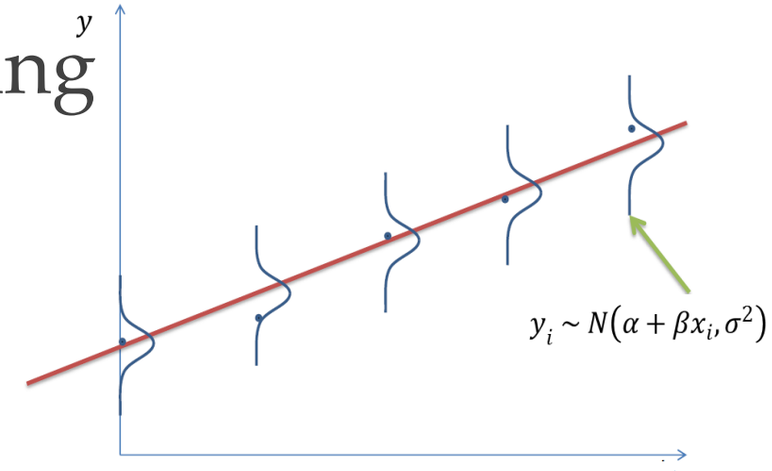
$$\alpha_+ = \hat{\alpha} + t_0 \cdot s_r \sqrt{\frac{1}{n} + \frac{\bar{x}^2}{\sum_{i=1}^{n}(x_i - \bar{x})^2}} = 42.6$$

- The NULL hypothesis H_0: $\alpha = 0$ is within the 95% confidence interval, so we fail to reject the NULL hypothesis.

Checking for Normality

- Recalling that the statistical model underlying linear regression is

$$y_i \sim \mathcal{N}(\alpha + \beta x_i, \sigma^2)$$



- the residual of the i'th sample should be normally distributed with zero mean and variance σ^2

$$\epsilon_i = y_i - (\alpha + \beta x_i) \sim \mathcal{N}(0, \sigma^2)$$

- Hence, a good way to check whether the assumption of linearity between x and y holds is to first fit the linear model and subsequently check that the residuals ϵ_i are normally distributed using a Q-Q plot.

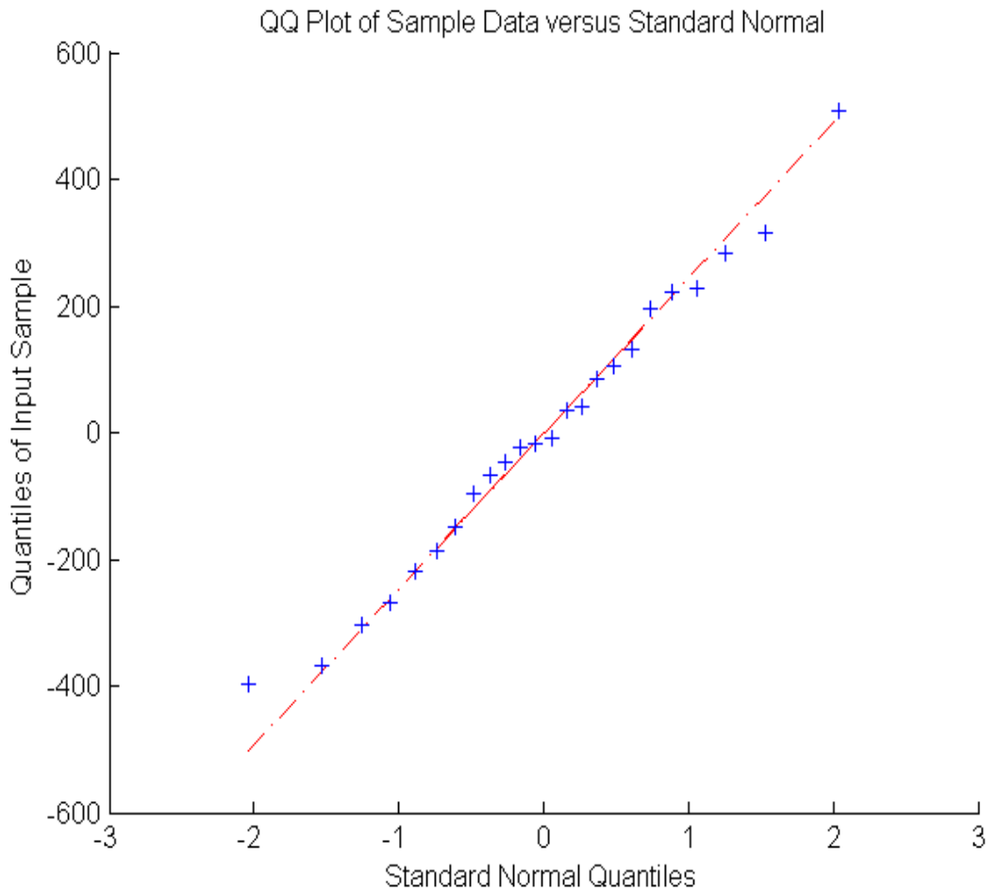
Checking for Normality Using Q-Q plot (Hubble's law)

- The residuals in Hubble's law example are

```
res=y-alpha-beta*x
```

- The resulting Q-Q plot

```
qqplot(res)
```



- shows that the residuals are approximately normally distributed, because the data points lie approximately on a straight line.

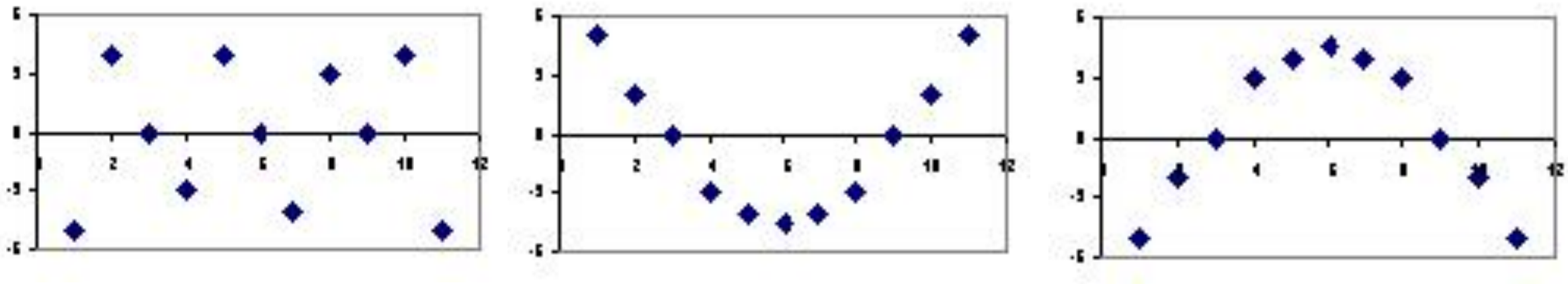
- Hence, it is safe to use simple linear regression to find the relation between the Speed and Distance of galaxies.

Residual Plots

- Another way to check the normality assumption is to make a so-called *residual plot*.
- A residual plot is a graph that shows the residuals on the vertical axis and the independent variable (***x***) on the horizontal axis.
- If the points in a residual plot are randomly dispersed around the horizontal axis, a linear regression model is appropriate for the data; otherwise, a non-linear model is more appropriate.

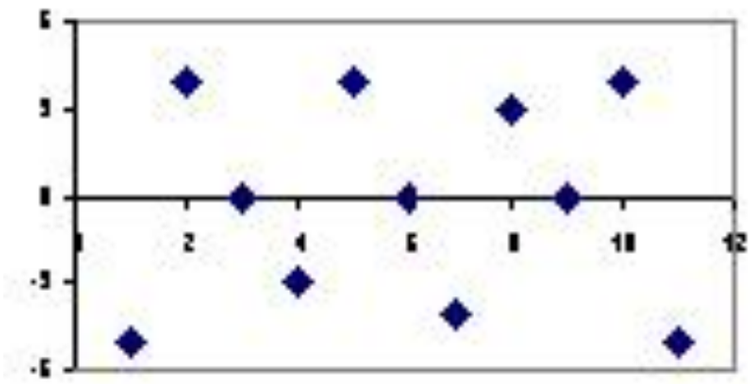
Residual Plots

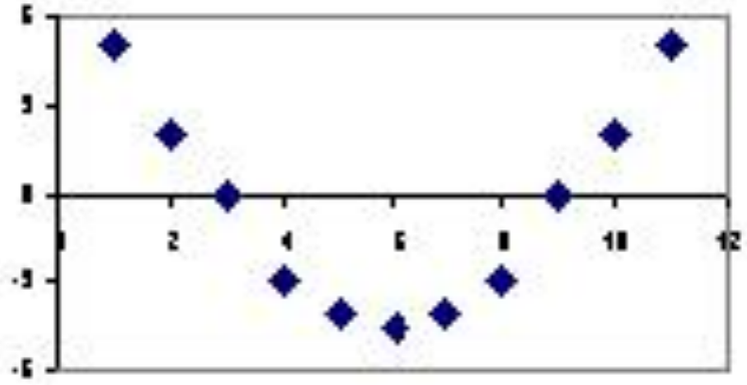
- Below, the residual plots show three typical patterns.
- The first plot shows a random pattern, indicating a good fit for a linear model.
- The other plot patterns are non-random (U-shaped and inverted U), suggesting a better fit for a non-linear model.

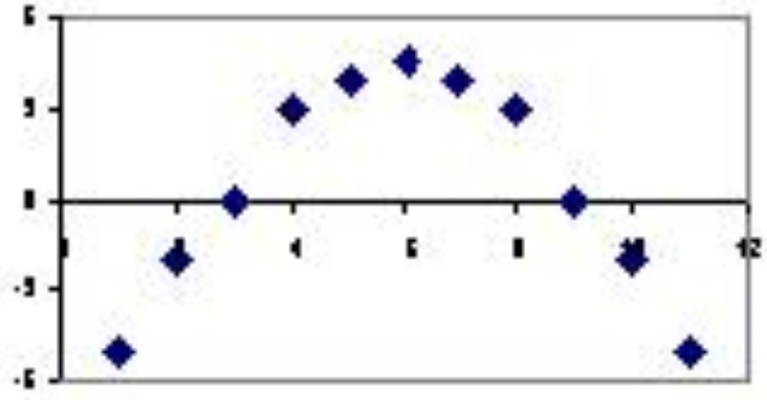


Residual Plots

- Formally, you must check the following two conditions:
  - The value of the residuals $\epsilon_i = y_i - (\alpha + \beta x_i)$ must not depend on x_i, but should lie randomly distributed around zero.
  - The variance of the residuals must not depend on x_i either.







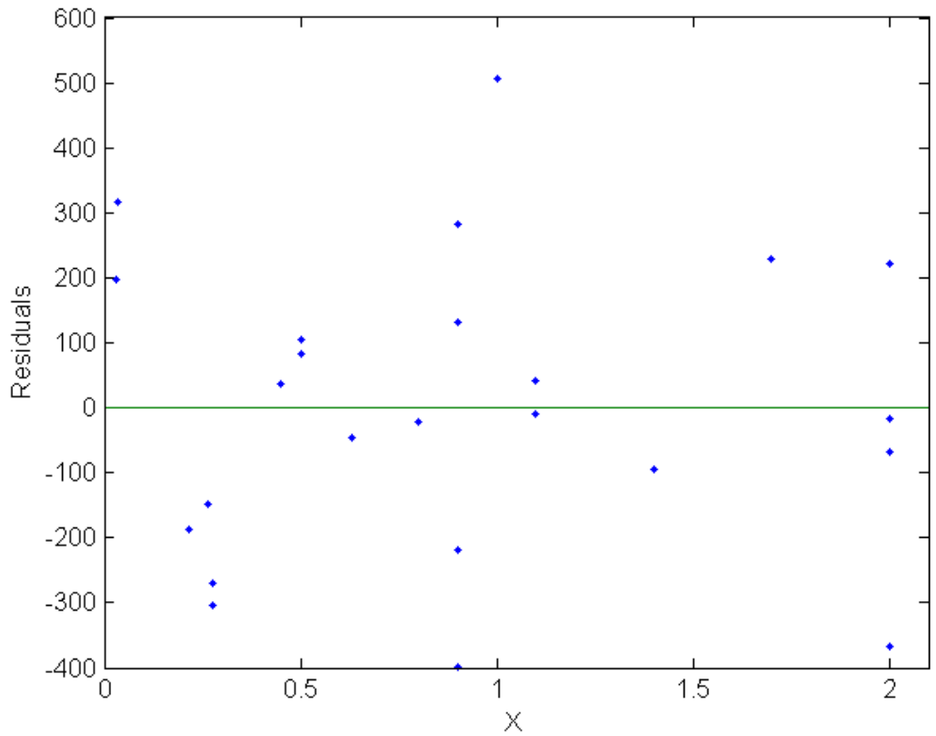
Checking for Normality Using Residual Plot (Hubble's law)

- The residuals in Hubble's law example are

```
res=y-alpha-beta*x
```

- The resulting residual plot

```
plot(x,res,'.',...
          [0 2.1],[0 0])
axis([0 2.1 -400 600])
xlabel('X')
ylabel('Residuals')
```

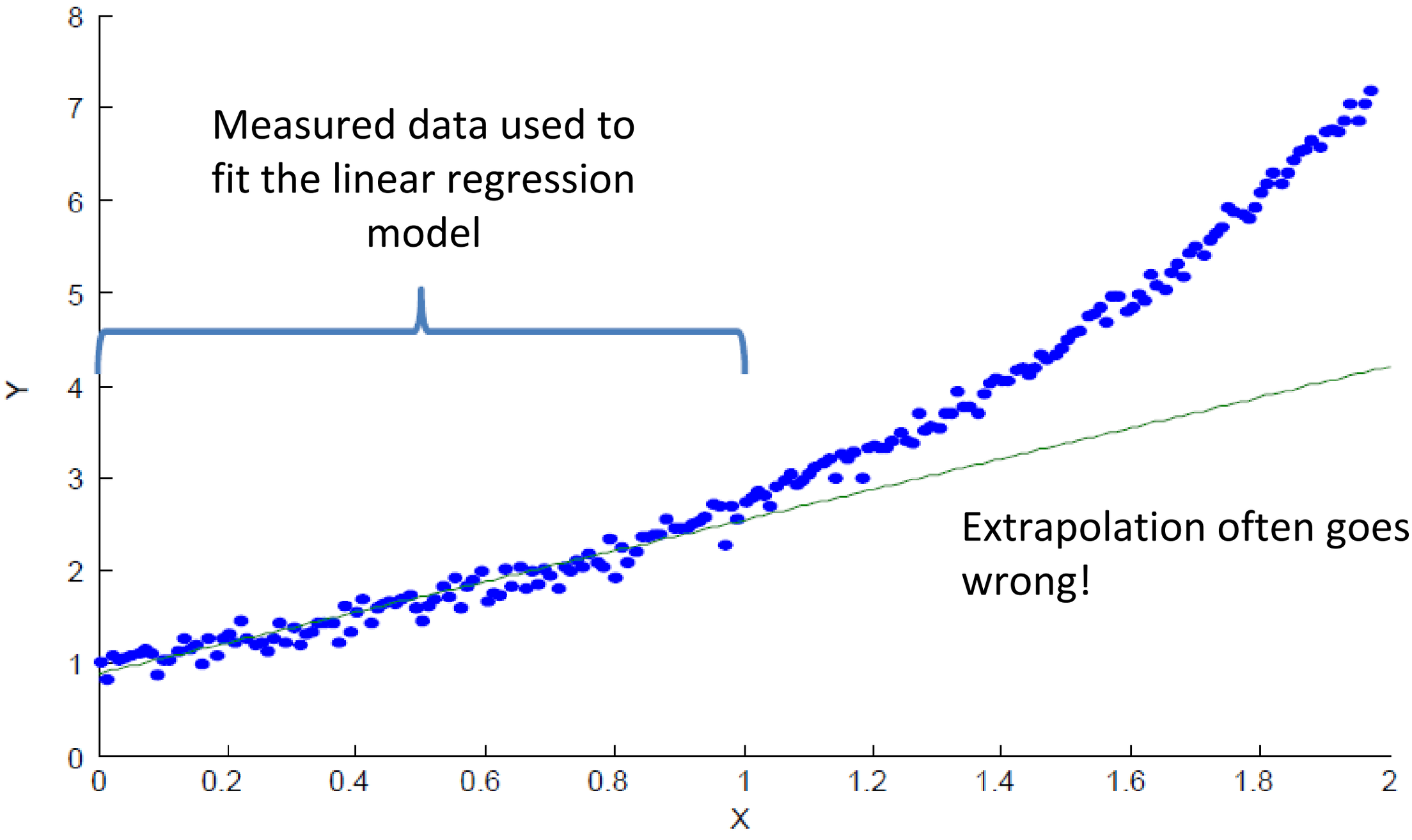


- Shows that the residuals are randomly distributed around zero and do not depend on x.
- Also, it appears that the variance of the residuals is independent of x.

Usage of Linear Regression

- Linear regression is often used for prediction.
- Suppose, for instance, that the relationship between daily energy consumption of a power plant and the outside temperature is linear.
- Then, given the temperature of tomorrow (from a weather forecast), we can give an estimate of tomorrows energy consumption of the power plant based on a linear model.
- When you use a regression equation, do not use values for the independent variable that are outside the range of values used to create the equation.
- That is called **extrapolation**, and it can produce unreasonable estimates.

Extrapolation



Sample Correlation Coefficient

- If we wish to quantify the strength of a linear relation, we can use the sample correlation coefficient

$$r = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2} \cdot \sqrt{\sum_{i=1}^{n}(y_i - \bar{y})^2}} = \frac{{s_{xy}}^2}{s_x \cdot s_y} = \frac{Cov(x,y)}{\sqrt{Var(x) \cdot Var(y)}}$$

- where s_x and s_y are the empitical standard deviations of x and y.
- As we saw in ”Random Signals”, chap. 2.3.3, the correlaltions coefficient (ϱ) takes on values from -1 to 1.
- It can be shown that the estimate of the regression slope is lineary related to the correlation coefficient:

$$\hat{\beta} = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{n}(x_i - \bar{x})^2} = \frac{{s_{xy}}^2}{{s_x}^2} = r\frac{s_x}{s_y}$$

Large β → large r → strong correlation

Coefficient of Determination

- In simple linear regression, the *coefficient of determination*

$$R^2 = r^2,$$

- indicates how well the data fit the linear model.
- The coefficient of determination ranges from 0 to 1 with value close to 1 suggesting a strong linear relationship, and values close to 0 suggesting no linear relationship.
- The coefficient of determination in the example with Hubble's law is $R^2 = 0.6235$.
- To calculate the sample correlation coefficient between *x* and *y* in Matlab, use the command `corr2(x,y)`.

Outliers

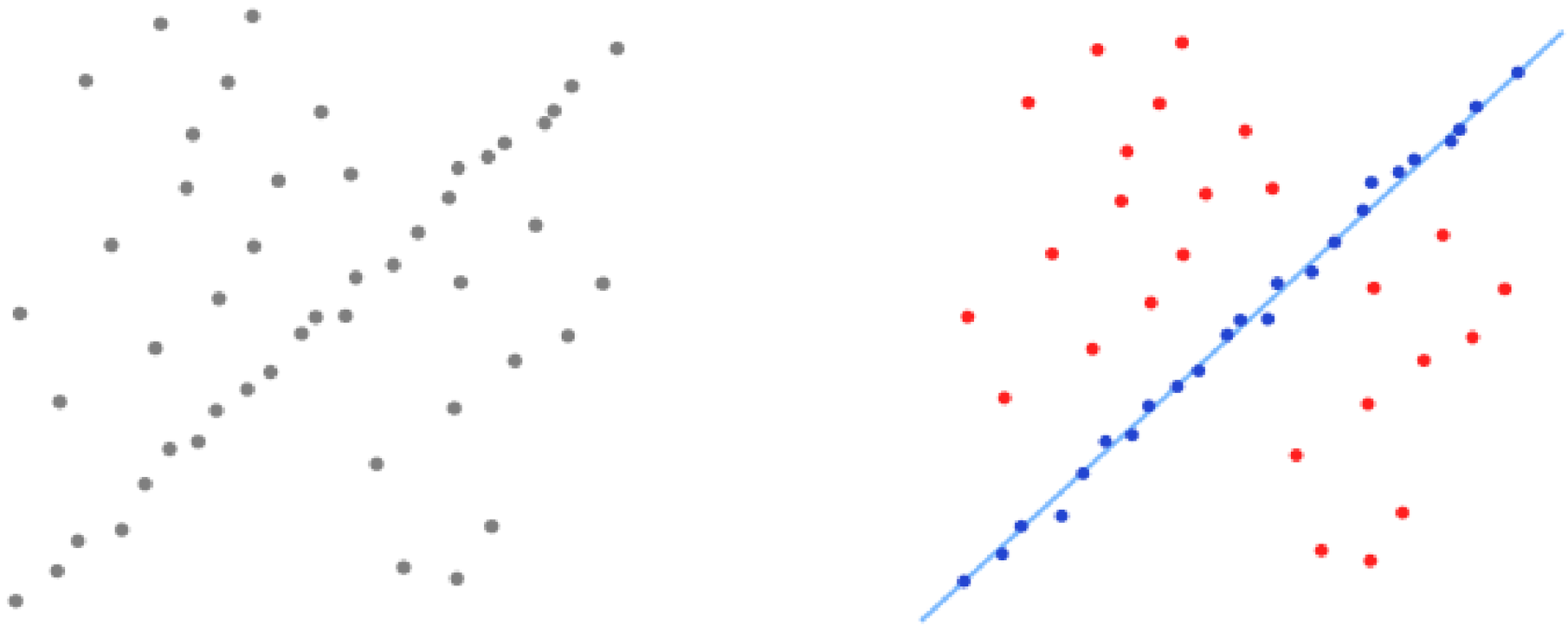
- Outliers are data points that are separated from the rest of the data and potentially influential for the regression analysis.
- Outliers can have a dramatic on the sample correlation coefficient (and therefore the slope).
- Recalling the definition of the sample correlation coefficient,

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

- an outlier is a point (x_i, y_i), such that either $(x_i - \bar{x})$ or $(y_i - \bar{y})$, or both, is large.
- The extent of influence of any point can be judged in part by computing the correlation coefficient with and without that point.

RANSAC

Random Sample Consensus



RANSAC is an iterative method to estimate parameters of a mathematical model from a set of observed data which contains outliers. It is a non-deterministic algorithm in the sense that it produces a reasonable result only with a certain probability, with this probability increasing as more iterations are allowed.

RANSAC

- Step 1: Select a subset of the data.
- Step 2: Find the best fitted model to that subset.
- Step 3: Determine the dataset of the data that fits with the model (inliers).
- Step 4: Repeat set 1-3, and if the new model has more inliers than the previous one, replaced the model with the new.
- Step 5: After a number of iterations, reject all datapoints that are not inliers. This is the outliers.
- Step 6: Re-estimate the model based on all the inliers.

➢ i.e. Find the model (linear regression) that gives the largest number of inliers (smallest number of outliers)

Linear Models – When/Why?

- Can be used when the mean changes over time.
- The variance should not change over time
- The mean is connected to time linearly!

- Outliers must not be omitted from a conclusion
  - Fx. may new medication damage indiviual patients (allergy)
- Outliers can – with justification – be omitted from a linear regression

Words and Concepts to Know

Linear Regression

Random Sample Consensus

Linear Model

Slope parameter

Sample Correlation Coefficient

Regression Intercept

Intercept parameter

Extrapolation

Predicted data

Outliers

Model Fitting

Slope parameter

Residual

Empirical Variance

Residual plot

Inliers

RANSAC

Measured data

Coefficient of Determinaton