Applied statistics (spring term 2019)

readers: Dr Ksenia Shalonova and Dr Nikolai Bode

written by Henry Haustein

Contents

1 Estimating parameters

1.1 Confidence and tolerance intervals

In statistical analysis we want to estimate a population from a random sample. This is called interference about the parameter. Random samples are used to provide information about parameters in an underlying population distribution. Rather than estimating the full shape of the underlying distribution, we usually focus on one or two parameters.

We want the error distribution to be centered on zero. Such an estimator is called unbiased. An biased estimator tends to have negative/positive errors, i.e. it usually underestimates/overestimates the parameter that is being estimated.

We also want error distribution to be tightly concentrated on zero, i.e. to have a small spread.

A good estimator should have a small bias and small standard error. These two criteria can be combined with into single value called the estimator's mean squared error. Most estimators that we will consider are unbiased, the spread of the error distribution is most important.

Definition 1.1 (Standard error)
The standard error (SE) of an estimator $\hat{\theta}$ of a parameter θ is defined to be its standard deviation.

■ **Example 1.2**
Standard error of the mean:

- Bias (μ error) $= 0$, i.e. $E(\hat{\theta}) = \theta$
- When population standard deviation is known: SE $= \frac{\sigma}{\sqrt{n}}$
- When population standard deviation is unknown: SE $= \frac{s}{\sqrt{n}}$

Do not confuse SD (sample standard deviation) ($\rightarrow$ one sample) and SE (standard deviation of the sample mean $\bar{x}$) ($\rightarrow$ error from hypothetical samples)!

Definition 1.3 (Confidence interval for μ with known σ)
We can be $(1-\alpha)\cdot 100\%$ confident that the estimate for μ will be in the interval

$$\bar{x} - z_{\alpha/2}\frac{\sigma}{\sqrt{n}} < \mu < \bar{x} + z_{\alpha/2}\frac{\sigma}{\sqrt{n}}$$

Common exact values of $z_{\alpha/2}$ with critical values from normal distribution:

confidence	**value of $z_{\alpha/2}$**
90%	1.645
95%	1.96
99%	2.575

Definition 1.4 (Confidence interval for μ when σ is unknown)
If we simply replace σ by its sample variance the confidence level will be lower than 95%. When the sample size is large, the confidence level is close to 95% but the confidence level can be much lower if the sample size is small.

Critical value comes from the STUDENTS t distribution. The value of $t_{\alpha/2}$ depends on the sample size through the use of degrees of freedom. The confidence interval is

$$\bar{x} - t_{\alpha/2}\frac{s}{\sqrt{n}} < \mu < \bar{x} + t_{\alpha/2}\frac{s}{\sqrt{n}}$$

Consider estimation of a population mean, μ, from a random sample of size n. A confidence interval will be of the form $\bar{x} \pm t_{\alpha/2}\frac{s}{\sqrt{n}}$. If we want our estimate to be within k of μ, then we need n to be large enough so that $t_{\alpha/2}\frac{s}{\sqrt{n}} < k$. For 95% confidence interval if n is reasonably large the t-value in the inequality will be approximately 1.96: $1.96\frac{s}{\sqrt{n}} < k$ that can be re-written as $n > \left(\frac{1.96s}{k}\right)^2$. In practice, it is best to increase n a little over this value in case the sample deviation was wrongly guessed.

■ **Example 1.5**
If we expect that a particular type of measurement will have a standard deviation of about 8, and we want to estimate its mean, μ, to within 2 of its correct value with probability 0.95, the sample size should be:

$$n > \left(\frac{1.96 \cdot 8}{2}\right)^2 = 61.5$$

This suggests a sample size of at least 62. The more accurate trial-and-error method using a t-value would give a sample size of 64.

The sample proportion of successes is denoted by $\hat{p}$ and is an estimate of p. The estimation error is $\hat{p} - p$.

$$\hat{p} = \frac{\text{number of successes in sample}}{\text{sample size}}$$

A 95% confidence interval is[1] $\hat{p} \pm 2 \cdot \sqrt{\frac{p(1-p)}{n}}$

■ **Example 1.6**
In a random sample of $n = 36$ values, there were $x = 17$ successes. We estimate the population proportion $\hat{p}$ with $\hat{p} = \frac{17}{36} = 0.472$. A 95% confidence interval for $\hat{p}$ is 0.472 ± 0.166. We are therefor 95% confident that the population of successes is between 30.6% and 63.8%. A sample size of $n = 36$ is clearly too small to give a very accurate estimate.

If the sample size n is small or $\hat{p}$ is close to either 0 or 1, this normal approximation is inaccurate and the confidence level for the interval can be considerably less than 95%. Classical theory recommends to use the confidence interval for $\hat{p}$ only when $n > 30$, $n\hat{p} > 5$ and $n(1 - \hat{p}) > 5$.

[1] To get the correct result, you would have to multiply by 1.96 instead of 2. You can also use 2 for a quick calculation.

Annotation (z-value or t-value?)

- If you know the variance of the population, then you should use the z-value from normal distribution.
- If you don't know the variance of the population or the population is non-normal, then you should formally always use the t-value.
- For most non-normal population distributions, the distribution of the sample mean becomes close to normal when the sample size increases (Central Limit Theorem)
- Even for relatively small samples, the distributions are virtually the same. Therefore, it is common to approximate the t-distribution using normal distribution for sufficiently large samples (e.g. $n > 30$).

Definition 1.7 (Tolerance interval)
A $(1-\alpha)\cdot 100\%$ tolerance interval for $\gamma \cdot 100\%$ of the measurements in a normal population is given by $\bar{x} \pm Ks$ where K is a tolerance factor. Tolerance limits are the endpoints of the tolerance interval.

Do not mix up with confidence intervals! We focus on γ (a certain percentage of measurements) rather than on a population parameter.

If we knew μ and σ then the tolerance factor K is 1. Otherwise the tolerance factor depends on the level of confidence, γ and the sample size n.

■ **Example 1.8**
A corporation manufactures field rifles. To monitor the process, an inspector randomly selected 50 firing pins from the production line. The sample mean $\bar{x}$ for all observations is 0.9958 inch and standard deviation s is 0.0333. Assume that the distribution of pin lengths is normal. Find a 95% tolerance interval for 90% of the firing pin lengths.

Given $n = 50$, $\gamma = 0.9$ and $\alpha = 0.05$, work out K (you can either use a special table or MATLAB function). $K = 1.996$. The 95% tolerance interval is (0.9293, 1.0623). Approximately 95 of 100 similarly constructed tolerance intervals will contain 90% of the firing pin lengths in the population.

1.2 Maximum Likelihood Estimate

Definition 1.9 (likelihood function)
If random variables have joint probability $p(x_1, ..., x_n|\theta)$ then the function $L(\theta|x_1, ..., x_n) = p(x_1, ..., x_n|\theta)$ is called the likelihood function of θ.

The likelihood function tells the probability of getting the data that were observed if the parameter value was really θ.

Definition 1.10 (maximum likelihood estimate)
The maximum likelihood estimate of a parameter θ is the value that maximizes the likelihood function $L(\theta|x_1, ..., x_n) = p(x_1, ..., x_n|\theta)$.

In practice they maximize the logarithm of the likelihood function and solve the following equation:

$$\frac{\mathrm{d}\log L(\theta|x_1,...,x_n)}{\mathrm{d}\theta}$$

The following formula can find an approximate numerical value for the standard error of almost any maximum likelihood estimator:

$$\mathrm{SE}(\hat{\theta}) \approx \sqrt{-\frac{1}{l''(\hat{\theta})}}$$

For the 95% confidence interval we can write:

$$\hat{\theta} - 1.96 \cdot \mathrm{SE}(\hat{\theta}) < \theta < \hat{\theta} + 1.96 \cdot \mathrm{SE}(\hat{\theta})$$

For the 90% confidence interval we can write:

$$\hat{\theta} - 1.645 \cdot \mathrm{SE}(\hat{\theta}) < \theta < \hat{\theta} + 1.645 \cdot \mathrm{SE}(\hat{\theta})$$

■ **Example 1.11**

The probability density function (PDF) of exponential distribution is

$$\mathrm{PDF} = \begin{cases} \lambda e^{-\lambda x} & x \geq 0 \\ 0 & \text{otherwise} \end{cases}$$

We want to estimate the parameter λ.

Likelihood function	$L(\lambda\|x_1,...,x_n) = \lambda^n e^{-\lambda \sum x_i}$
log-likelihood function	$l(\lambda\|x_1,...,x_n) = n\log(\lambda) - \lambda \sum x_i$
MLE	$l'(\lambda\|x_1,...,x_n) = \frac{n}{\lambda} - \sum x_i \stackrel{!}{=} 0 \Rightarrow \hat{\lambda} = \frac{1}{\bar{x}}$
Standard error	$\mathrm{SE}(\hat{\lambda}) = \sqrt{-\frac{1}{l''(\hat{\lambda})}} = \frac{\hat{\lambda}}{\sqrt{n}} = \frac{1}{\sqrt{n\bar{x}}}$ where $l''(\lambda) = -\frac{n}{\lambda^2}$
95% confidence interval	$\frac{1}{\bar{x}} \pm 1.96 \cdot \frac{1}{\sqrt{n\bar{x}}}$

Lets assume that the mean time between failures of 199 air-conditioners is $\bar{x} = 90.92$ hours. The MLE for the estimated failure rate λ is $\frac{1}{\bar{x}} = 0.0110$ failure per hour.
$\Rightarrow$ 95% confidence interval for the failure rate:

$$\frac{1}{\bar{x}} \pm 1.96 \cdot \frac{1}{\sqrt{n\bar{x}}} \Rightarrow \lambda \in [0.00974, 0.01253]$$

Given a sample, we can estimate two unknown parameters in a probability distribution, for example, estimate parameters μ and σ in a normal distribution.

Definition 1.12 (likelihood function for two parameters)
If random variables have joint probability $p(x_1, ..., x_n|\theta, \varphi)$ then the function $L(\theta, \varphi|x_1, ..., x_n) = p(x_1, ..., x_n|\theta, \varphi)$ is called the likelihood function of θ and φ.

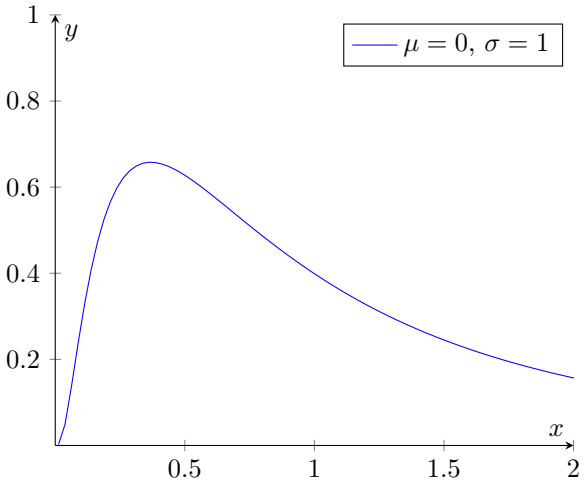
The likelihood function is maximised at a turning point of the likelihood function and could therefore be found by setting the partial derivatives of $L(\theta, \varphi)$ with respect to θ and φ to zero.

There are two important properties of the maximum likelihood estimator $\hat{\theta}$ of a parameter θ based on a random sample of size n from a distribution with a probability function $p(x_1, ..., x_n|\theta)$:

- Asymptotically unbiased: $E(\hat{\theta}) \to \theta$ when $n \to \infty$
- Asymptotically has a normal distribution: $\hat{\theta} \to$ normal distribution when $n \to \infty$ that can be used to generate confidence intervals.
- Maximum likelihood estimators have low mean squared error if the sample size is large enough. MLE can be heavily biased for small samples!

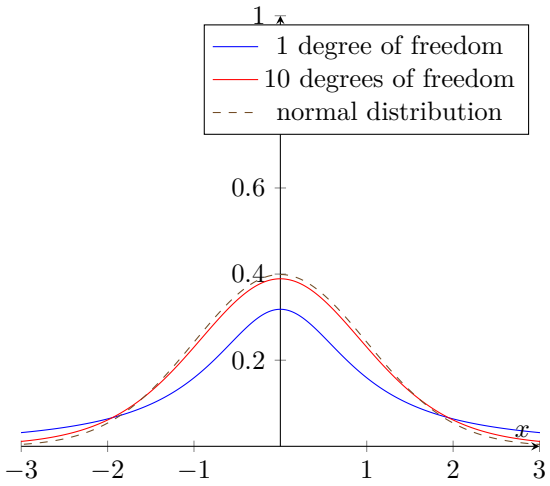
1.3 Continuous distributions

The lognormal distribution is used in situations where values are positively skewed, for example, for financial analysis of stock prices. Note that the uncertain variable can increase without limits but cannot take negative values.

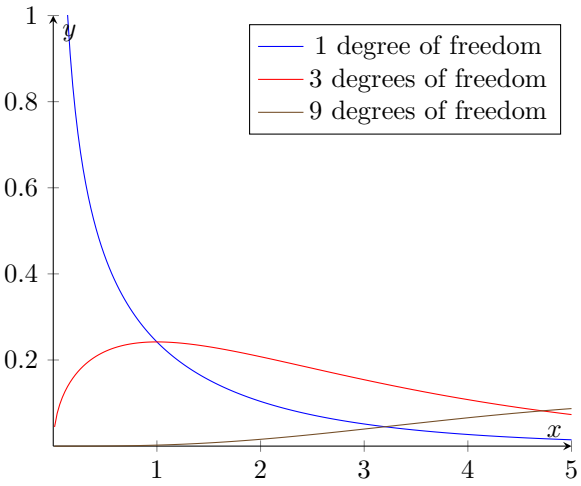


In the beta distribution the uncertain variable is a random value between 0 and positive value. The distribution is frequently used for estimating the proportions and probabilities (i.e. values between 0 and 1). The shape of the distribution is specified by two positive parameters.

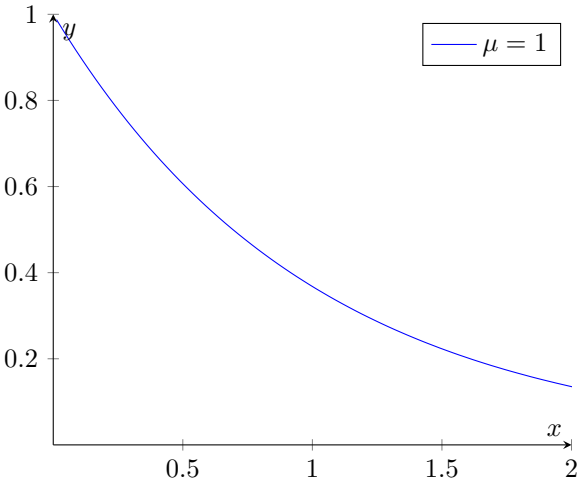
The Students t distribution is the most widely used distribution in confidence intervals and hypothesis testing. The distribution can be used to estimate the mean of a normally distributed population when the sample size is small. The t distribution comes to approximate the normal distribution as the degrees of freedom (or sample size) increases.



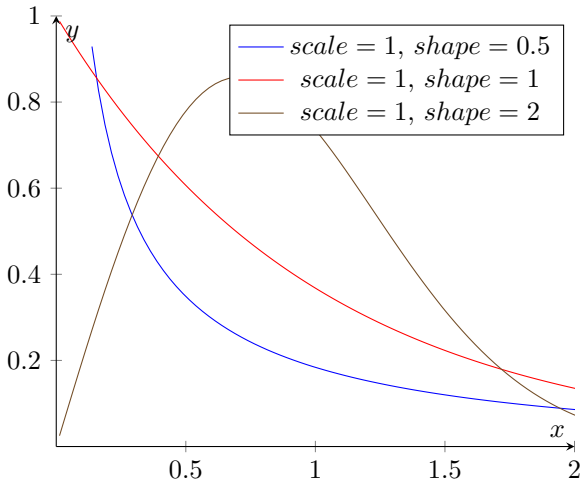
The chi-square distribution is usually used for estimating the variance in a normal distribution.



In a homogeneous Poisson process with a rate λ events per unit time, the time until the first event happens has a distribution called an exponential distribution. All exponential distributions have their highest probability density at $x = 0$ and steadily decrease as x increases.



The Weibull distribution can be used as a model for items that either deteriorate or improve over time. It's basic version has two parameters: shape and scale.



- $shape > 1$: the hazard function is increasing so the item becomes less reliable as it gets older.
- $shape < 1$: the hazard function is decreasing so the item becomes more reliable as it gets older.
- $shape = 1$: the hazard function is constant so the lifetime distribution becomes exponential.

The survival function (probability of surviving until a particular time) is $R(t) = 1 - F(t)$. The hazard rate function (failure rate) is worked out by the formula:

$$\begin{aligned} h(t) &= \frac{f(t)}{1 - F(t)} \\ &= \frac{f(t)}{R(t)} \end{aligned}$$

where $f(t)$ and $F(t)$ are PDF and CDF of the distribution.

The hazard function describes how an item ages where t affects its risk of failure. This constant hazard function in the exponential distribution corresponds to the Poisson process without memory, i.e. the chance of failing does not depend on what happened before and how long the item has already survived.

2 Hypothesis testing

There are two types of questions in statistical interference:

- **Parameter estimation:** What parameter values would be consistent with the sample data?
- **Hypothesis testing:** Are the sample data consistent with some statement about the parameters?

The Null Hypothesis H_0 often specifies a single value for the unknown parameter such as "$\alpha = \dots$". It is a default value that can be accepted as holding if there is no evidence against it. A researcher often collects data with the express hope of disapproving the null hypothesis.

If the null hypothesis is not true, we say that the alternative hypothesis H_A holds. If the data are not consistent with the null hypothesis, then we can conclude that the alternative hypothesis must be true. Either the null hypothesis or the alternative hypothesis must be true.

■ **Example 2.1**

The data show the number of operating hours between successive failures of air-conditioning equipment in ten aircrafts. The sample of 199 values is a test statistic. We can test the manufacturer's claim that the rate of failures is no more than one per 110 hours of use.

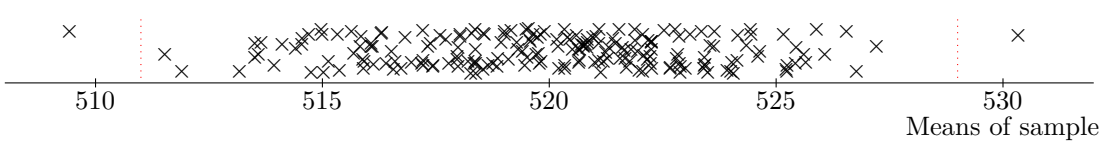
$$H_0 : \lambda \leq \frac{1}{100} \text{ (claim of a manufacturer)}$$
$$H_A : \lambda > \frac{1}{100}$$

This can be simplified:

$$H_0 : \lambda = \frac{1}{100} \text{ (claim of a manufacturer)}$$
$$H_A : \lambda > \frac{1}{100}$$

2.1 The p-value (probability value)

In an industrial process some measurement is normally distributed with standard deviation $\sigma = 10$. Its mean should be $\mu = 520$, but can differ a little bit. Samples of $n = 10$ measurements are regularly collected as part of quality control. If a sample had $\bar{x} = 529$, does the process need to be adjusted?



From the 200 simulated samples above (Monte Carlo simulation), it seems very unlikely that a sample mean of 529 would have been recorded if $\mu = 529$. There is strong evidence that the industrial process no longer has a mean of $\mu = 520$ and needs to be adjusted.

Definition 2.2 (p-value)
A p-value describes the **evidence against** H_0. A p-value is evaluated from a random sample so it has a distribution in the same way that a sample mean has a distribution.

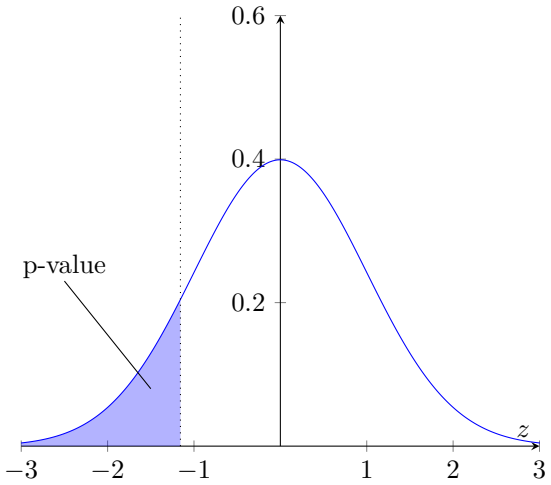
p-value	**Interpretation**
over 0.1	no evidence that H_0 does not hold
between 0.05 and 0.1	very weak evidence that H_0 does not hold
between 0.01 and 0.05	moderately strong evidence that H_0 does not hold
under 0.01	strong evidence that H_0 does not hold

■ **Example 2.3 (normal distribution with known σ, one-tailed test)**
We are given a random sample of $n = 30$ with $\bar{x} = 16.8$. Does the population have mean $\mu = 18.3$ and standard deviation $\sigma = 7.1$, or is the mean now lower than 18.3?

$$H_0 : \mu = 18.3$$
$$H_A : \mu < 18.3$$

The p-value can be evaluated using the statistical distance of 16.8 from 18.3 (a z statistic).

$$z = \frac{\bar{x} - 18.3}{\underbrace{\frac{7.1}{\sqrt{30}}}_{\text{standard error}}} = -1.157$$



$$\text{p-value} = P(z \leq -1.157) = 0.124$$

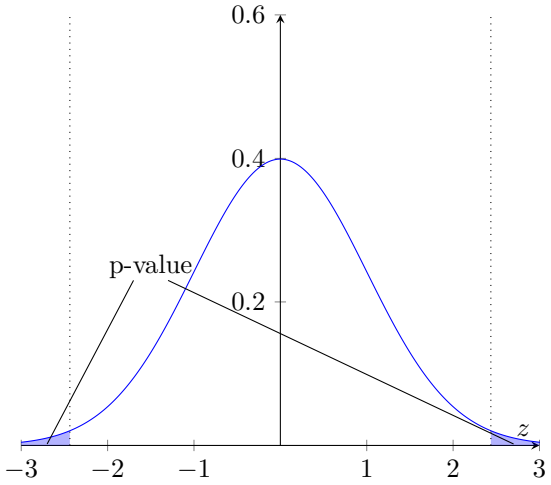
The p-value is reasonably large, meaning that a sample mean as low as 16.8 would not be unusual if $\mu = 18.3$, so there is no evidence against H_0.

Annotation
To compute the p-value you can use

$$\text{p-value} = \texttt{CDF(NormalDistribution(0,1),-1.157)}$$

■ **Example 2.4 (normal distribution with known σ, two-tailed test)**
Companies test their products to ensure that the amount of active ingredient is within some limits. However the chemical analysis is not precise and repeated measurements of the same specimen usually differ slightly. One type of analysis gives results that are normally distributed with a mean that depend on the actual product being tested and standard deviation 0.0068 grams per litre. A product is tested three times with the following concentrations of the active ingredient: 0.8403, 0.8363, 0.8447 grams per litre. are the data consistent with the target concentration of 0.85 grams per litre?

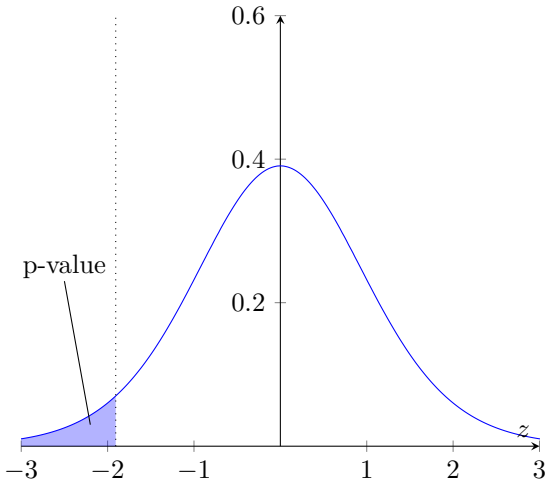
null hypothesis	H_0: $\mu = 0.85$
alternative hypothesis	H_A: $\mu \neq 0.85$
test statistic	$\bar{x} = 0.8404$, $z = \frac{0.8404-0.85}{\frac{0.0068}{\sqrt{3}}} = -2.437$, $P(z \leq -2.437) = 0.00741$
p-value	$2 \cdot 0.00741 = 0.0148$
p-value interpretation	There is moderately strong evidence that the true concentration is not 0.85.



■ **Example 2.5 (normal distribution with unknown σ, one-tailed test)**
Both cholesterol and saturated fats are often avoided by people who are trying to lose weight or reduce their blood cholesterol level. Cooking oil made from soybeans has little cholesterol and has been claimed to have only 15% saturated fat. A clinician believes that the saturated fat content is greater than 15% and randomly samples 13 bottles of soybean cooking oil for testing with the following percentage saturated fat: 15.2, 12.4, 15.4, 13.5, 15.9, 17.1, 16.9, 14.3, 19.1, 18.2, 15.5,

16.3, 20.0.

null hypothesis	H_0: $\mu = 15$
alternative hypothesis	H_A: $\mu > 15$
T-test for μ	$\bar{x} = 16.138$, $t = \frac{16.138-15}{\frac{2.154}{\sqrt{13}}} = 1.906$, $P(t \geq 1.906) = 0.040$ (t-distribution with 12 degrees of freedom)
p-value interpretation	Since this is below 0.05, we conclude that there is moderately strong evidence that the mean saturated fat content of the oils is higher than the claimed 15%.

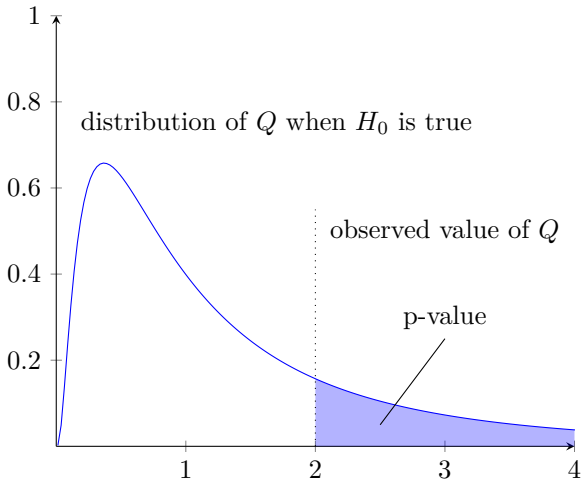


A hypothesis test is based on two competing hypotheses about the value of a parameter θ.
Null hypothesis H_0: $\theta = \theta_0$
Alternative hypothesis (one-tailed test) H_A: $\theta > \theta_0$

The hypothesis test is based on a test statistic that is some function of the data values:

$$Q = g(x_1, ..., x_n|\theta_0)$$

whose distribution is fully known when H_0 is true (i.e. when θ_0 is the true parameter value). We evaluate the test statistic to assess whether it is unusual enough to throw doubt on the null hypothesis.



Theorem 2.6
P-values close to zero throw doubt on the null hypothesis.

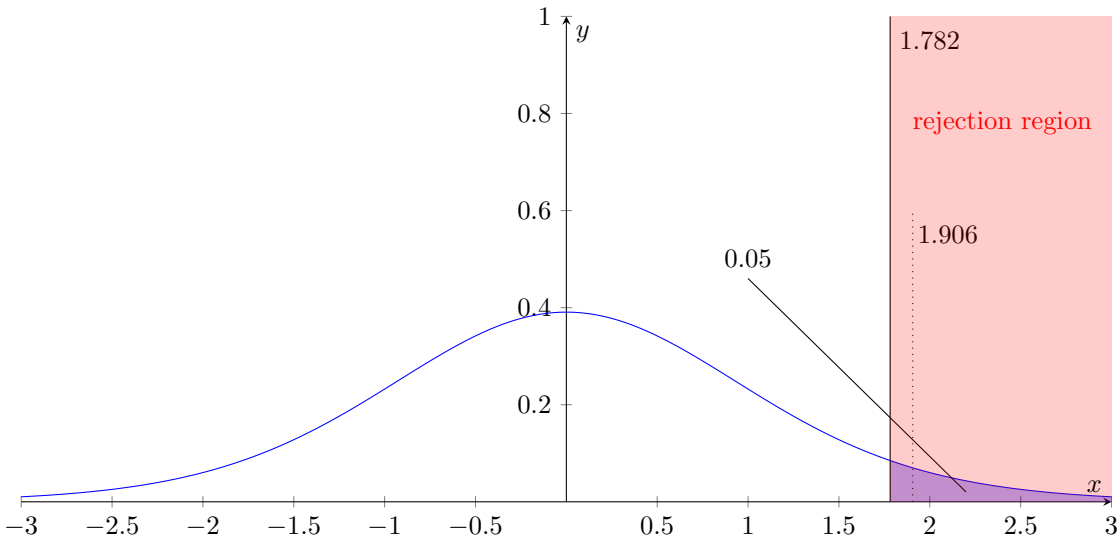
2.2 The significance level

Definition 2.7 (significance level)
The significance level is the probability of wrongly concluding that H_0 does not hold when it actually does.

- **One-tailed test:** For example, it may be acceptable to have a 5% chance of concluding that $\theta < \theta_0$ when actually $\theta = \theta_0$. This means a significance level (tail area of the test statistic's distribution) of this test is $\alpha = 0.05$.
- **Two-tailed test:** Values at both tails of the distribution of the test statistic result in rejection of H_0, so the corresponding tail areas should each have area $\frac{\alpha}{2}$ for a test with significance level α.

■ **Example 2.8**
Cooking oil made from soybeans has little cholesterol and has been claimed to have only 15% saturated fat. A clinician believes that the saturated fat content is greater than 15% and randomly samples 13 bottles of soybean cooking oil for testing: 15.2, 12.4, 15.4, 13.5, 15.9, 17.1, 16.9, 14.3, 19.1, 18.2, 15.5, 16.3, 20.0.

Null hypothesis	H_0: $\mu = 15$
Alternative hypothesis	H_A: $\mu > 15$
A significance level of $\alpha = 0.05$ means that the clinician is willing to wrongly conclude that the saturated fat content is over 15% when it really is 15% with probability 0.05.	
t-statistic	$t = \frac{\bar{x}-15}{\frac{s}{\sqrt{13}}} = 1.906$
rejection region	$P(T > 1.782) = 0.05$ (t distribution with 12 degrees of freedom)
Conclusion	t lies in the rejection region so H_0 is rejected at the 5% significance level.



Definition 2.9 (Type 1 + 2 error)
The Type 1 error is the significance level of the test. The decision rule is usually defined to make the significance level 5% or 1%.

The Type 2 error is wrongly accepting H_0 when it is false.

Instead of the probability of a Type 2 error, it is common to use the power of a test, defined as one minus the probability of a Type 2 error. The power of a test is the probability of correctly rejecting H_0 when it is false.

		Decision	
		accept H_0	reject H_0
Truth	H_0 is true		significance level = P(Type 1 error)
	H_0 is false	P(Type 2 error)	Power = 1 - P(Type 2 error)

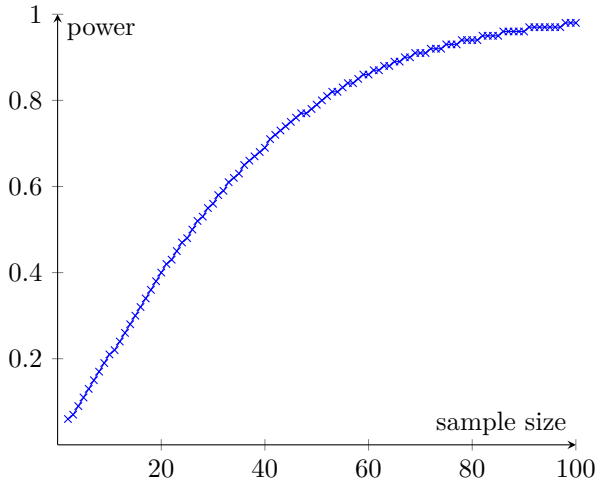
Computer software can provide the p-value for a hypothesis test at 5% or 1% significance level (Type 1 error).

It is clearly desirable to use a test whose power is as close to 1 as possible. There are three different ways to increase the power:

- **Increase the significance level:** If the critical value for the test is adjusted, increasing the probability of a Type 1 error decreases the probability of a Type 2 error and therefore increase the power.
- **Use a different decision rule:** For example, in a test about the mean of a normal population, a decision rule based on the sample median has lower power than a decision rule based on the sample mean.
- **Increase the sample size:** By increasing the amount of data on which we base our decision about whether to accept or reject H_0, the probabilities of making errors can be reduced.

When the significance level is fixed, increasing the sample size is therefore usually the only way to improve the power.

Ideally there should be a trade-off between low significance level (Type 1 error) and high power. The desired power of the test is usually 0.8. The power of a test is not a single value since the alternative hypothesis allows for a range of different parameter values. It is represented by a power function that can be graphed against the possible parameter values. MATLAB `sampsizepwr` can compute the sample size to obtain a particular power for a hypothesis test, given the parameter value of the alternative hypothesis.



There are a many number of statistical tests for assessing normality: Shapiro-Wilk test, Kolmogorov-Smirnov test, Jacque-Bera test, etc. The Shapiro-Wilk test ($n < 50$) can be used to verify whether data come from a normal distribution:
H_0: sample data are not significantly different than a normal population.
H_A: sample data are significantly different than a normal population.
P-value > 0.05 mean the data are normal
P-value < 0.05 mean the data are not normal
Monte Carlo simulations proved the efficiency of Shapiro-Wilk test. It s preferable that normality

is assessed visually as well! The KOLMOGOROV-SMIRNOV non-parametric test ($n > 50$) examines if scores are likely to follow some distribution in some population (not necessarily normal).

2.3 Likelihood ratio test

In some cases we need to perform a hypothesis test to compare two models: big "general" model (M_B) and small "simple" model (M_S) nested into the bigger model.
H_0: M_S fits the data
H_A: M_S does not fit the data and M_B should be used instead.

We need to verify if M_B fits the data significantly better.

- **Measure how well a model fits the data:** The fit of any model can be described by the maximum possible likelihood for that model:

$$L(M) = \max\{P(data|model)\}$$

  Calculate the maximum likelihood estimates for all unknown parameters and insert them into the likelihood function.

- **Work out the likelihood ratio:**

$$R = \frac{L(M_B)}{L(M_S)} \geq 1$$

  Big values of R suggests that M_S does not fit as well as M_B.

- **Work out log of likelihood ratio:**

$$\log(R) = l(M_B) - l(M_S) \geq 0$$

  Big values of R suggests that M_S does not fit as well as M_B.

■ **Example 2.10**
There are a number of defective items on a production line in 20 days that follow POISSON(λ) distribution: 1, 2, 3, 4, 2, 3, 2, 5, 5, 2, 4, 3, 5, 1, 2, 4, 0, 2, 2, 6.
M_S: the sample comes from POISSION(2)
M_B: the sample comes from POISSION(λ)

■ **Example 2.11**
Clinical records give the survival time for 30 people: 9.73,5.56, 4.28, 4.87, 1.55, 6.20, 1.08, 7.17, 28.65, 6.10, 16.16, 9.92, 2.40, 6.19. In a clinical trial of a new drug treatment 20 people had survival times of: 22.07, 12.47, 6.42, 8.15, 0.64, 20.04, 17.49, 2.22, 3.00. Is there any difference in survival times for those using the new drug?
M_S: Both examples come from the same exponential(λ) distribution.
M_B: The first sample comes from exponential(λ_1) and the second sample from exponential(λ_2).

Definition 2.12
If the data come from $L(M_S)$, and $L(M_B)$ has k more parameters than $L(M_S)$ then

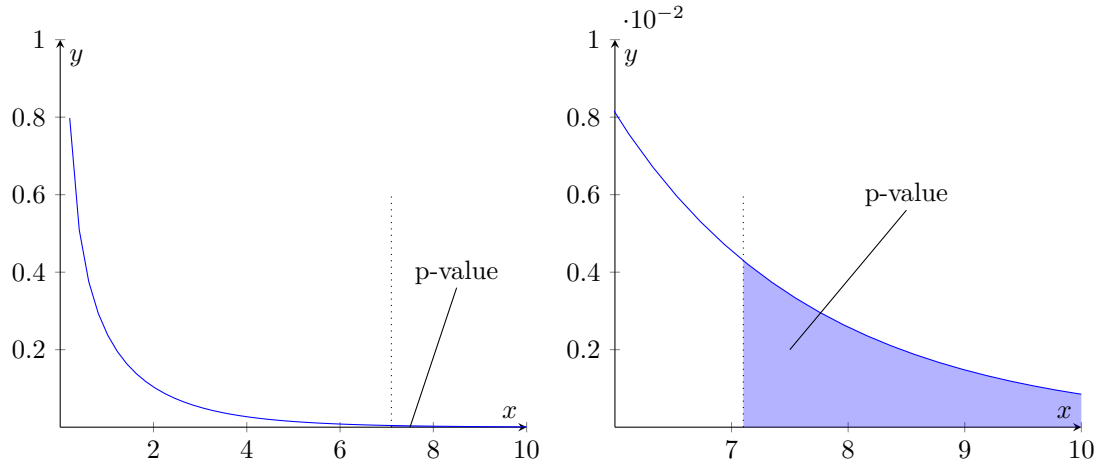
$$\begin{aligned} X^2 &= 2\log(R) \\ &= 2\big(l(M_B) - l(M_S)\big) \\ &\approx \chi^2(k \text{ degrees of freedom}) \end{aligned}$$

The main steps for the likelihood ratio test are:

1. Work out maximum likelihood estimates of all unknown parameters in M_S.
2. Work out maximum likelihood estimates of all unknown parameters in M_B.
3. Evaluate the test statistic: $\chi^2 = 2\big(l(M_B) - l(M_S)\big)$
4. The degrees of freedom for the test are the difference between the numbers of unknown parameters in two models. The p-value for the test is the upper tail probability of the $\chi^2(k$ degrees of freedom) distribution given the test statistic.
5. Interpret the p-value: small values give evidence that the null hypothesis (M_S model) does not hold.

■ **Example 2.13**
There are a number of defective items on a production line in 20 days that follow $\textsc{Poisson}(\lambda)$ distribution: 1, 2, 3, 4, 2, 3, 2, 5, 5, 2, 4, 3, 5, 1, 2, 4, 0, 2, 2, 6.

	null hypothesis	H_0: $\lambda = 2$ small model M_S
	alternative hypothesis	H_A: $\lambda \neq 2$ big model M_B
	log-likelihood for the Poisson distribution	$l(\lambda) = \left(\sum_{i=1}^{20} x_i\right) \log(\lambda) - n\lambda$
M_B	MLE for the unknown parameter	$\hat{\lambda} = \frac{\sum x_i}{n} = 2.9$
M_B	Maximum possible value for the log-likelihood	$l(M_B) = 58\log(2.9) - 20 \cdot 2.9 = 3.7532$
M_S	MLE for the unknown parameter	no unknown parameter
M_S	Maximum possible value for the log-likelihood	$l(M_S) = 58\log(2) - 20 \cdot 2 = 0.2025$
	Likelihood ratio test	$\chi^2 = 2\big(l(M_B) - l(M_S)\big) = 7.101$
	It should be compared to χ^2(1 degree of freedom) since the difference in unknown parameters is equal to 1.	
	p-value	The p-value is 0.008 (the upper tail probability above 7.101)
	Interpreting p-value	The p-value is very small and we can conclude that there is strong evidence that M_B fits the data better than M_S: $\lambda \neq 2$.



A two-sample t-test should be used to compare group means when you have independent samples. A paired t-test is needed when each sampled item in one group is associated with an item sampled from the other group.

2.4 Two-sample t-test

We can carry out a hypothesis test to verify if the two means are equal:
H_0: $\mu_1 = \mu_2$
H_A: $\mu_1 \neq \mu_2$ (The corresponding one-tailed alternative also holds.)

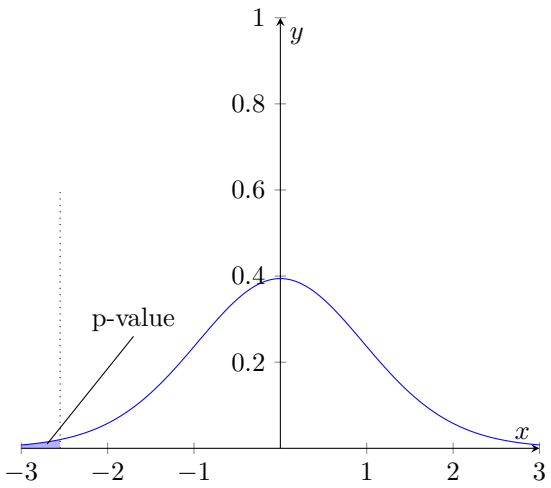
Definition 2.14
If $\bar{x_1}$ and $\bar{x_2}$ come from Normal(μ_1, σ) and Normal(μ_2, σ) with sample sizes n_1 and n_2 then

$$T = \frac{\bar{x_1} - \bar{x_2}}{\text{SE}(\bar{x_1} - \bar{x_2})} \approx t(n_1 + n_2 - 2 \text{ degrees of freedom})$$

provided $\mu_1 = \mu_2$. For relatively large sample sizes (Central Limit Theorem) we can use Z-test instead of t-test.

■ **Example 2.15**
A botanist is interested in comparing the growth response of dwarf pea stems to two different levels of the hormone indoleacetic acid (IAA). The botanist measured the growth of pea stem segments in millimetres for $0.5 \cdot 10^{-4}$ IAA level: 0.8, 1.8, 1.0, 0.1, 0.9, 1.7, 1.0, 1.4, 0.9, 1.2, 0.5 and for 10^{-4} IAA level: 1.0, 1.8, 0.8, 2.5, 1.6, 1.4, 2.6, 1.9, 1.3, 2.0, 1.1, 1.2. Test whether the larger hormone concentration results in greater growth of the pea plants.

independent samples	$n_x = 11$, $n_y = 12$
Null hypothesis	H_0: $\mu_x = \mu_y$
Alternative hypothesis	H_A: $\mu_x < \mu_y$
The pooled estimate assumes that the variance is the same in both groups	$s^2 = \frac{10s_x^2 + 11s_y^2}{21} = 0.2896$
test statistic	$t = \frac{1.027 - 1.6}{\sqrt{0.2896(\frac{1}{11} + \frac{1}{12})}} = -2.5496$
p-value for 21 degrees of freedom in t-distribution	$P(t \leq -2.5496) = 0.0093$
Interpretation	There is very strong evidence that the mean growth of the peas is higher at the higher hormone concentration.



Annotation (pooled variance)

In statistics, pooled variance (also known as combined, composite, or overall variance) is a method for estimating variance of several different populations when the mean of each population may be different, but one may assume that the variance of each population is the same. The numerical estimate resulting from the use of this method is also called the pooled variance.

Under the assumption of equal population variances, the pooled sample variance provides a higher precision estimate of variance than the individual sample variances. This higher precision can lead to increased statistical power when used in statistical tests that compare the populations, such as the t-test.

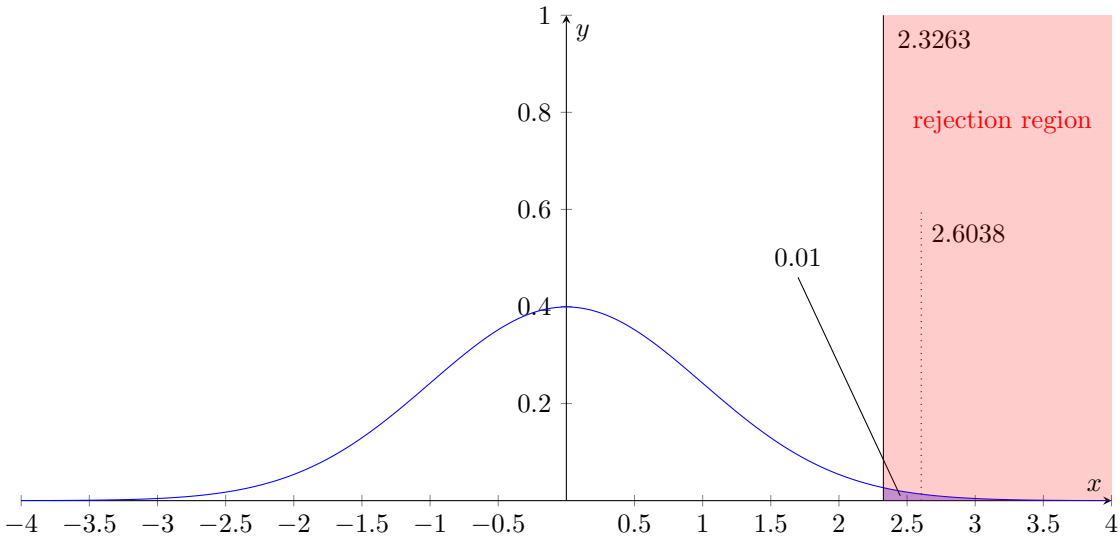
$$s^2 = \frac{\sum_{i=1}^{k}(n_i - 1)s_i^2}{\sum_{i=1}^{k}(n_i - 1)}$$

Adapted from https://en.wikipedia.org/wiki/Pooled_variance.

■ **Example 2.16 (from MATLAB session)**

When you sing in to your Facebook account, you are granted access to more than 1 million relying party (RP) websites. RP websites were categorized as server-flow or client-flow websites. Of the 40 server-flow sites studied, 20 were found to be vulnerable to impersonation attacks. Of the 54 client-flow sites, 41 were found to be vulnerable to impersonation attacks. Do these results indicate that a client-flow website is more likely to be vulnerable to an attack than a server-flow website? Test using $\alpha = 0.01$.

Null hypothesis	H_0: $p_{server} = p_{client} \Rightarrow \frac{20}{40} = \frac{41}{54}$
Alternative hypothesis	H_A: $p_{server} < p_{client}$
pooled sample proportion	$p = \frac{40 \cdot \frac{20}{40} + 54 \cdot \frac{41}{54}}{40+54} = 0.6489$
test statistic	$z = \frac{p_{client} - p_{server}}{\sqrt{0.6489(\frac{1}{40} + \frac{1}{54})}} = 2.6038$
rejection region for $\alpha = 0.01$	`norminv(0.01)` $= 2.3268$
Interpretation	z lies in the rejection region so H_0 is rejected.



2.5 Paired t-test

Testing whether two paired measurements X and Y have equal means is done in terms of the difference $D = Y - X$. The hypothesis

H_0: $\mu_x = \mu_y$

H_A: $\mu_x \neq \mu_y$

can be re-written as

H_0: $\mu_d = 0$

H_A: $\mu_d \neq 0$.

This can reduce the paired data set to a univariate data set of differences. The hypothesis can be assigned using t-test:

$$t = \frac{\bar{d} - 0}{\frac{s_d}{\sqrt{n}}}$$

Z-test can be used for relatively large sample sizes.

■ **Example 2.17**

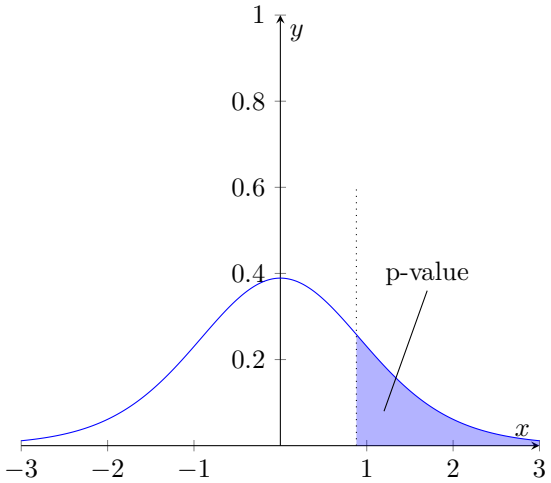
A researcher studying congenital heard disease wants to compare the development of cyanotic children with normal children. Among the measurement of interest is the age at which the children

speak their first word.

pair of siblings	cyanotic sibling	normal sibling	difference
1	11.8	9.8	2.0
2	20.8	16.5	4.3
3	14.5	14.5	0.0
4	9.5	15.2	-5.7
5	13.5	11.8	1.7
6	22.6	12.2	10.4
7	11.1	15.2	-4.1
8	14.9	15.6	-0.7
9	16.5	17.2	-0.7
10	16.5	10.5	6.0

The researcher wants to test whether cyanotic children speak their first word later on average than children without the disease.

Null hypothesis	H_0: $\mu_d = 0$
Alternative hypothesis	H_A: $\mu_d > 0$
test statistic	$t = \frac{\bar{d}-0}{\frac{s_d}{\sqrt{n}}} = 0.8802$
Interpretation	The p-value is well above zero (0.1997), so there is no evidence that the cyanotic children learn to speak later.



■ **Example 2.18**

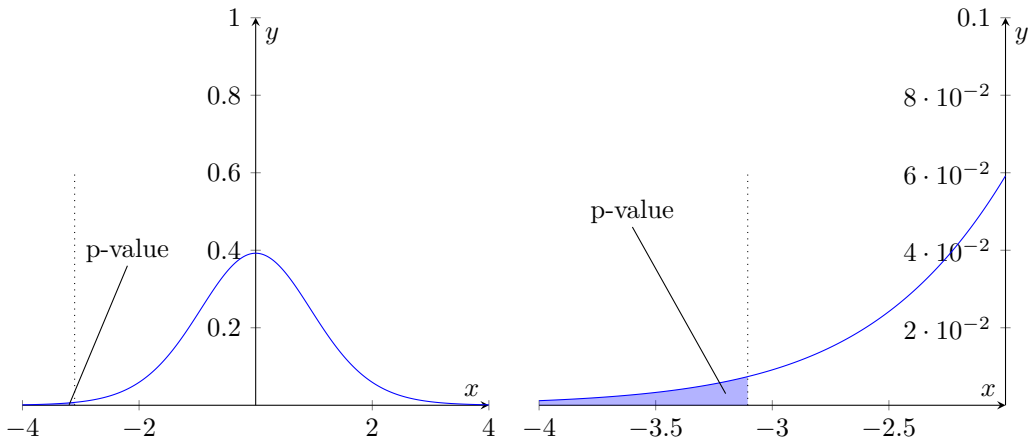
The blood pressure of 15 college-aged woman was measured before starting to take the pill and

after 6 months of use.

	blood pressure	
subject	**before pill**	**after pill**
1	70	68
2	80	72
3	72	62
4	76	70
5	76	58
6	76	66
7	72	68
8	78	52
9	82	64
10	64	72
11	74	74
12	92	60
13	74	74
14	68	72
15	84	74

A two-tailed test is used as the pill might either increase or decrease blood pressure.

Null hypothesis	H_0: $\mu_d = 0$
Alternative hypothesis	H_A: $\mu_d \neq 0$
test statistic	$t = \frac{\bar{d}-0}{\frac{s_d}{\sqrt{n}}} = -3.1054$
Interpretation	The p-value (0.0072) is very small that gives strong evidence that the blood pressure has changed. The negative t-value suggests that the blood pressure has decreased.


1
y
0.8
0.6
0.4
0.2
p-value
x
−4
−2
2
4
0.1
y
$8 \cdot 10^{-2}$
$6 \cdot 10^{-2}$
$4 \cdot 10^{-2}$
$2 \cdot 10^{-2}$
p-value
x
−4
−3.5
−3
−2.5

3 Bootstrapping

3.1 A word of warning

The limitation of the bootstrap is the assumption that the distribution of the data represented by one sample is an accurate estimate of the population distribution. If the sample does not reflect the population distribution, then the random sampling performed in the bootstrap procedure may add another level of sampling error, resulting in inaccurate statistical estimations.

It is important to get quality data that accurately reflects the population being sampled. The smaller the original sample, the less likely it is to accurately represent the entire population.

We use bootstrap if

- we have a small but representative random sample or
- we have a not-normal distribution or aren't sure about it.

3.2 Why to use it?

In normal population the mean μ is the parameter that is most often estimated. But other parameters are possible too:

- standard deviation
- interquartile range (upper quartile - lower quartile)
- median
- other percentiles (e.g. upper quartile)

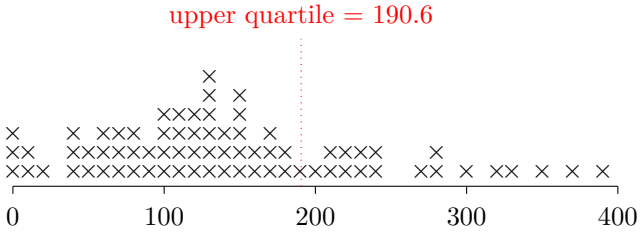
These parameters can be estimated using the corresponding summary statistic from a random sample, but the error distribution may be difficult to obtain theoretically.

Resampling techniques are normally used to estimate parameters and confidence intervals from sample data when parametric test assumptions are not met or for small samples from non-normal distributions.

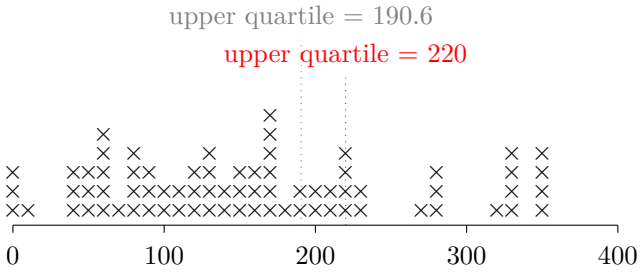
- non-parametric bootstrap
- parametric bootstrap
- Jackknife
- permutation tests

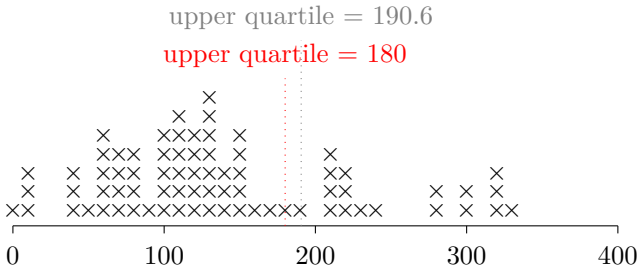
Non-parametric bootstrap means that only a random sample is known and no prior knowledge on the population density function.

■ **Example 3.1**
Monthly rainfall in Dodoma, Tanzania has a skew distribution in some months. The distribution of a sample is provided below.



If a normal distribution does not seem a reasonable model, an alternative is to treat the actual sample as the "population" for the simulation and take random samples with replacement from this sample. Such samples are called bootstrap samples. A simulation with these bootstrap samples can again show the error distribution and provide approximate values for the bias and standard error.





3.3 Bootstrap distribution

The standard error of a statistic is the standard deviation of the sample statistic. The standard error can be calculated as the standard deviation of the sampling distribution.

Bootstrap sample is a random sample taken with replacement from the original sample, of the same size as the original sample. A bootstrap statistic is the statistic computed on a bootstrap sample. A bootstrap distribution is the distribution of many bootstrap statistics. The standard error of a statistic can be estimated using the standard deviation of the bootstrap distribution.

Let $\hat{\theta}$ a statistic calculated from a sample ($\hat{\theta} = \bar{x}$). We draw r observations with replacement to create a bootstrap sample and calculate the statistic $\hat{\theta}^*$ for this sample.

- **bootstrap standard error:** the sample standard deviation of the bootstrap distribution:

$$\mathrm{SE}_b = \sqrt{\frac{\sum(\hat{\theta}_b^* - \bar{\theta}^*)^2}{B-1}}$$

where B is the number of bootstrap replications (usually $B > 10000$)

- **bootstrap bias:** $\bar{\theta}^* - \hat{\theta}$
- **bootstrap confidence intervals:** bootstrap percentile interval, t confidence interval with bootstrap standard error, bootstrap t-interval, etc.

3.4 Bootstrap methods

Name	calculate	repeat	get distribution	confidence interval
Bootstrap percentile CI or EFRON method	$\hat{\theta}_b^*$	B times	$\left\{\hat{\theta}_b^*\right\}_{b=1}^{B}$	$[q_{\alpha/2}, q_{1-\alpha/2}]$
Bootstrap CI - bootstrap t	$\frac{\hat{\theta}_b^* - \hat{\theta}}{\mathrm{SE}(\hat{\theta}_b^*)}$	B times	$\left\{\frac{\hat{\theta}_b^* - \hat{\theta}}{\mathrm{SE}(\hat{\theta}_b^*)}\right\}_{b=1}^{B}$	$[\hat{\theta} - \mathrm{SE}(\hat{\theta}) \cdot q_{1-\alpha/2}, \hat{\theta} - \mathrm{SE}(\hat{\theta}) \cdot q_{\alpha/2}]$
Bootstrap CI symmetric t-percentile	$\frac{\hat{\theta}_b^* - \hat{\theta}}{\mathrm{SE}(\hat{\theta}_b^*)}$	B times	$\left\{\frac{\hat{\theta}_b^* - \hat{\theta}}{\mathrm{SE}(\hat{\theta}_b^*)}\right\}_{b=1}^{B}$	$[\hat{\theta} - \mathrm{SE}(\hat{\theta}) \cdot q_{1-\alpha}, \hat{\theta} + \mathrm{SE}(\hat{\theta}) \cdot q_{1-\alpha}]$
Bootstrap CI HALL method	$\hat{\theta}_b^* - \hat{\theta}$	B times	$\left\{\hat{\theta}_b^* - \hat{\theta}\right\}_{b=1}^{B}$	$[\hat{\theta} - q_{1-\alpha/2}, \hat{\theta} - q_{\alpha/2}]$

Bootstrap using t CI - not recommended

$$\hat{\theta} \pm t_{\alpha/2} \cdot \mathrm{SE}_b$$

Bootstrap standard error is the sample standard deviation of the bootstrap distribution

$$\mathrm{SE}_b = \sqrt{\frac{\sum(\hat{\theta}_b^* - \bar{\theta}^*)^2}{B-1}}$$

where B is the number of bootstrap replications (usually $B > 10000$). The bootstrap bias is $\bar{\theta}^* - \hat{\theta}$. It can be useful when the standard error is difficult to derive. It has a poor performance when distributions are highly skewed.

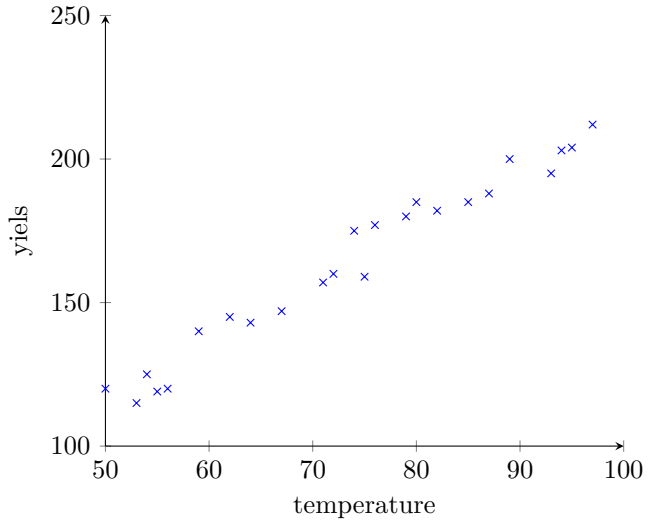
Bootstrap percentile CI or Elfron method
For a 90% confidence interval keep the middle 90%, leaving 5% in each tail and 5% in the head. The 90% confidence interval boundaries would be the 5th percentile and the 95th percentile. In case we have 10000 bootstrap replications: $\theta_1^* \leq \theta_2^* \leq \cdots \leq \theta_{10000}^*$ the 90% confidence interval is $[\theta_{500}^*, \theta_{9500}^*]$.

- **Advantages:** A very intuitive and easy to implement method. Can also outperform some other bootstrap CI methods for skewed distributions.
- **Disadvantages:** Can be too narrow for small samples.

4 Linear models (Simple linear regression)

■ **Example 4.1**

Suppose a chemical reaction produces higher yields of a product, the higher the ambient temperature



Simple linear regression is what we can use to investigate if a relationship between two variables exists when we don't know about the underlying process. We model a linear relationship between two variables and

(a) quantify this pattern before

(b) testing if we can believe it

4.1 Structure of simple linear regression models

Definition 4.2 (simple linear regression)

For n observed data pairs $\{(x_i, Y_i) \mid i = 1, ..., n\}$ simple linear regression assumes we have the relationship

$$Y_i = \beta_0 + \beta_1 x_i + \varepsilon_i$$

Y is called the respose variable and x the explainatory variable. β_0 and β_1 are model parameters. ε_i are error terms (spread around the regression line) and crucially, we assume $\varepsilon_i \overset{i.i.d}{\sim} Normal(0, \sigma)$ in simple linear regression.

Definition 4.3 (alternative representations for simple linear regression models)

(a) algebraic notation:

$$\begin{aligned} Y_1 &= \beta_0 + \beta_1 x_1 + \varepsilon_1 \\ Y_2 &= \beta_0 + \beta_1 x_2 + \varepsilon_2 \\ &\vdots \\ Y_n &= \beta_0 + \beta_1 x_n + \varepsilon_n \end{aligned}$$

(b) Matrix notation:

$$Y = \beta X + \varepsilon$$

Because of the error term ε, the response Y is a random variable. So, $Y_i = E(Y_i) + \varepsilon_i = \beta_0 + \beta_1 x_i + \varepsilon_i$ and thus $E(Y_i) = \beta_0 + \beta_1 x_i$. This leads to the following notation

(c) Random variable notation:

$$\begin{aligned} Y_i &\overset{i.i.d}{\sim} Normal(\beta_0 + \beta_1 x_i, \sigma) \\ Y &\sim Normal(X\beta, \mathbb{1}\sigma) \end{aligned}$$

Since the mean of the response is a linear function of the explanatory variables, there are often called simple linear models.

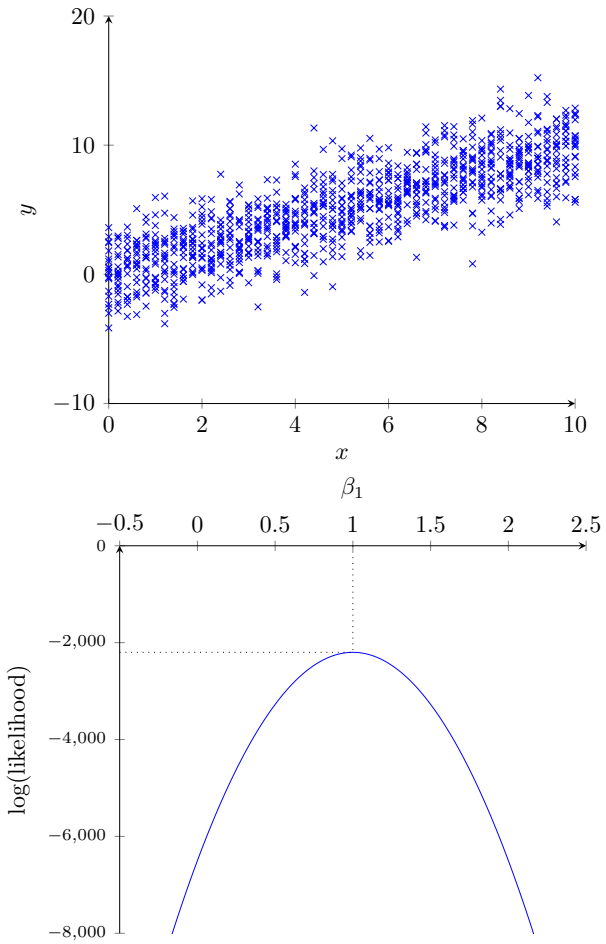
Given data, we want to estimate the parameters of the model. In this course, we always use Maximum Likelihood Estimation (MLE). That means, we find parameter values that maximise the likelihood of our model. The likelihood for an simple linear model is

$$L(\beta \mid X) = \prod_{i=1}^{n} f_{Normal}(Y_i, \mu = \beta_0 + \beta_1 x_i, \sigma)$$

where f_{Normal} is the PDF for the normal distribution with mean $\mu = \beta_0 + \beta_1 x_i$ and standard deviation σ evaluated at Y_i. For linear models, it has been shown that MLE is equivalent to Ordinary Least Squares (OLS). For linear models exact equations for these parameter estimates exist

$$\begin{aligned} \hat{\beta} &= (X^T X)^{-1} X^T Y \\ \hat{\sigma}^2 &= \frac{1}{n-2} (Y - X\hat{\beta})^T (Y - X\hat{\beta}) \end{aligned}$$

Fitted values for the response are: $\hat{Y} = X\hat{\beta}$



In $Y_i = \beta_0 + \beta_1 x_i + \varepsilon_i$, assuming we know $\beta_0 = 0$and σ, we can see that the likelihood is maximal for $\beta_1 = 1$. Seeming plausible given the data on the left.

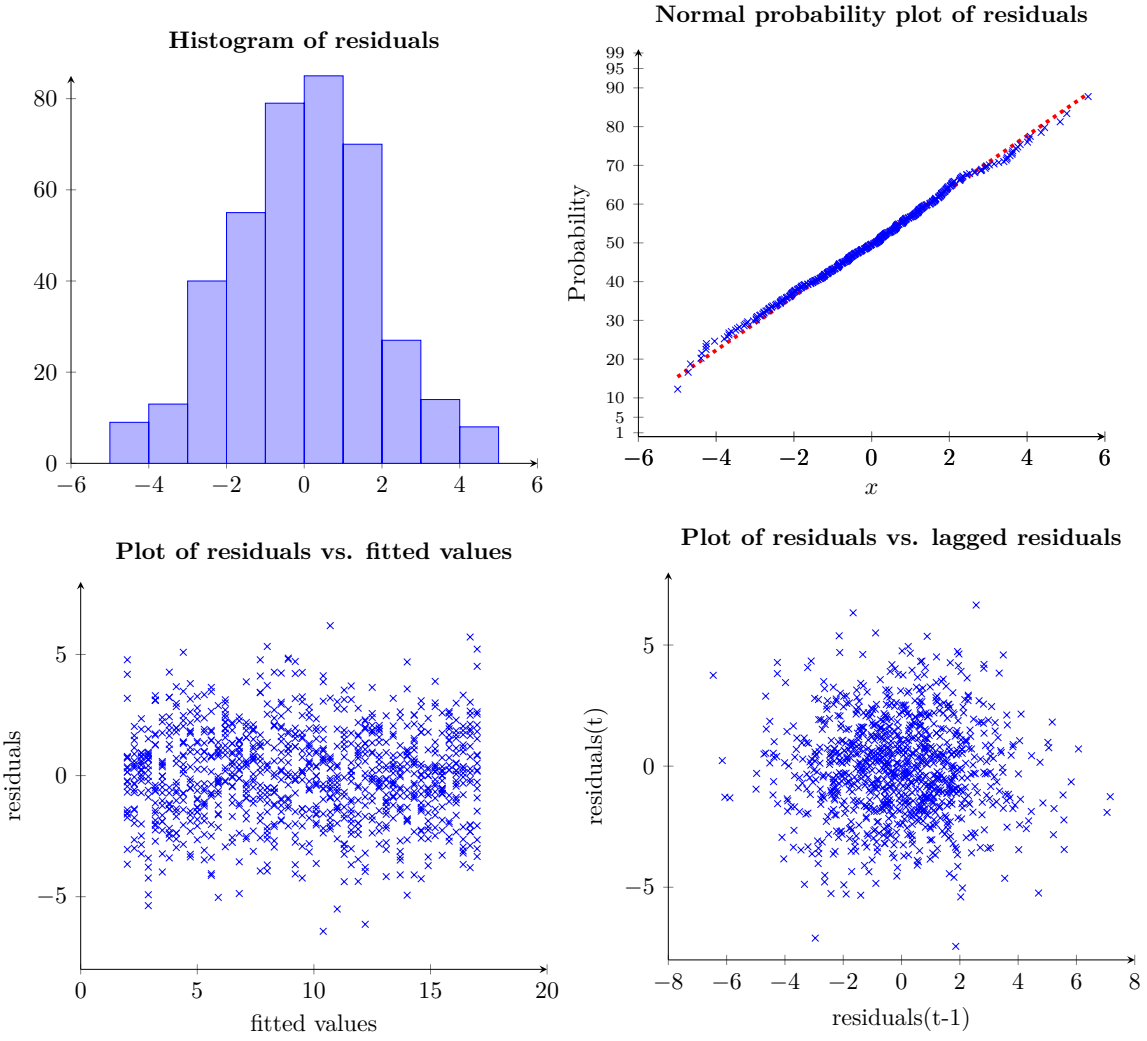
4.2 Assumptions of simple linear models

All statistical models make assumptions. Linear models are no exception. They assume, most importantly:

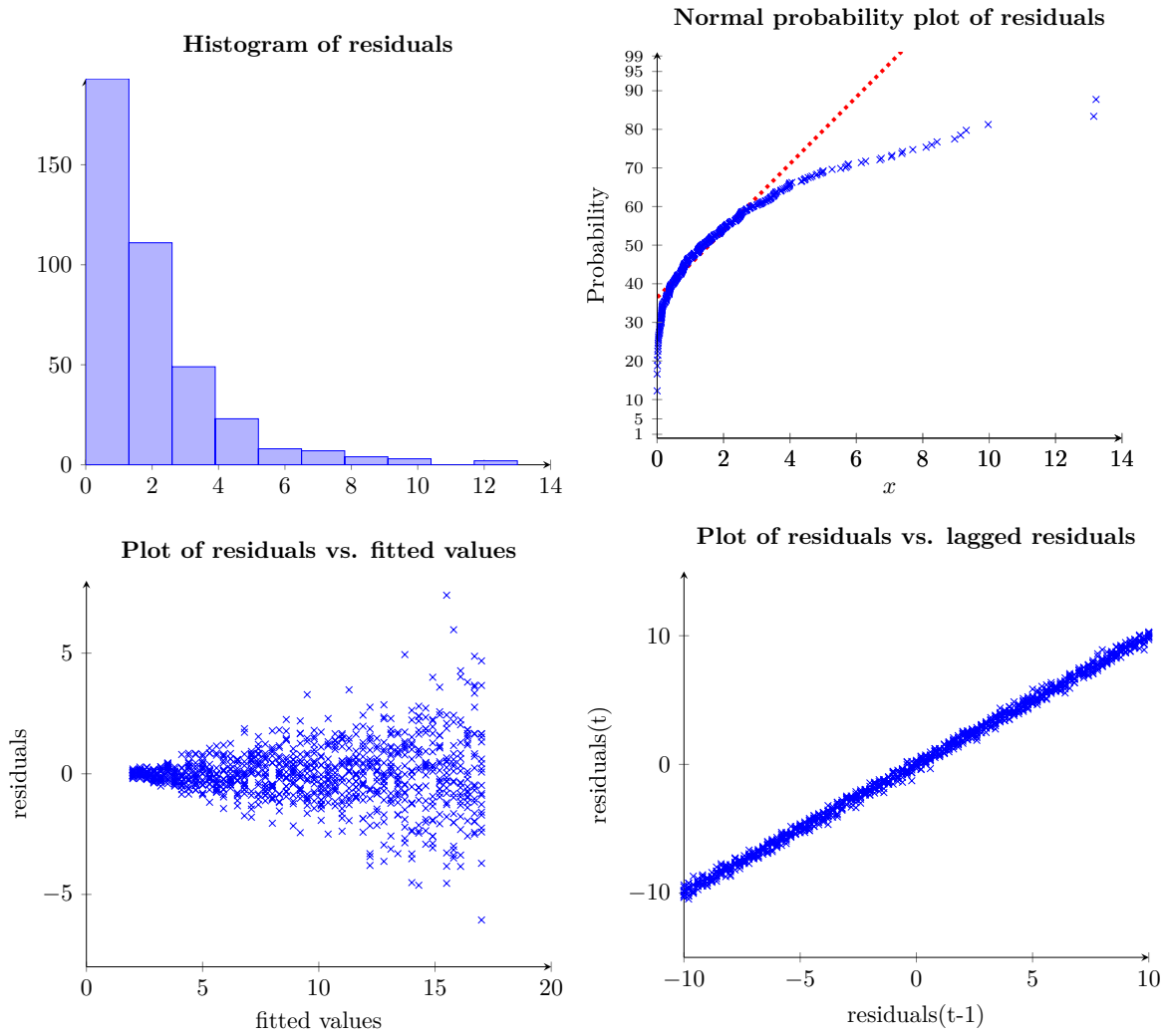
- **Linearity:** response variable is a linear combination of the explanatory variables
- **Normality:** errors follow a normal distribution
- **Constant error variance (homoscedasticity):** variance of the response variable (or errors) does not depend on the value of the explanatory variables. If this assumption is invalid it's called heteroscedasticity.
- **Independence:** the errors are uncorrelated (ideally statistically independent). This means that the response variable observations are conditionally independent. Need this for the product in the likelihood function.
- **Weak exogeneity:** the explanatory variables can be treated as fixed values, rather than random variables.

It is important to check that the most important assumptions hold (approximately) when fitting linear models to data.

To check if the model assumptions hold, we look at the errors, or residuals: $\varepsilon = Y - X\hat{\beta}$. Hypothesis testing on residuals is possible, but we will focus on residual plots for model checking. Perfect residual plots look like this:



The following residual plots gone wrong. Note that these plots are not all from the same data.



4.3 Hypothesis testing on simple linear model parameters

How to test if the x_is contribute information for the prediction of Y in $Y_i = \beta_0 + \beta_1 x_i + \varepsilon_i$? One way of doing this is to test the hypothesis that Y does not change as the explanatory x changes. In other words, we test the hypothesis

H_0: $\beta_1 = 0$

H_A: $\beta_1 \neq 0$

Fortunately, 'math boffins' have found that $\hat{\beta}_1$ follows a normal distribution with mean β_1 and standard error

$$\sigma_{\hat{\beta}_1} = \frac{\sigma}{\sqrt{SS_{xx}}} \approx \frac{s}{\sqrt{SS_{xx}}}$$

where $SS_{xx} = \sum(x_i - \bar{x})^2$ and $s^2 = \frac{\sum(Y_i - \hat{Y}_i)^2}{n-2}$. Since σ is usually unknown, we need to use a STUDENT's t-test on

$$t = \frac{\hat{\beta}_1 - \text{hypothesised value}}{\frac{s}{\sqrt{SS_{xx}}}}$$

with degrees of freedom based on the number of data points and model parameters. In practice, software will does this for us!

As an alternative for interference on parameter estimates, we can also compute confidence intervals. The $(1-\alpha)100\%$ confidence interval for the gradient β_1 is

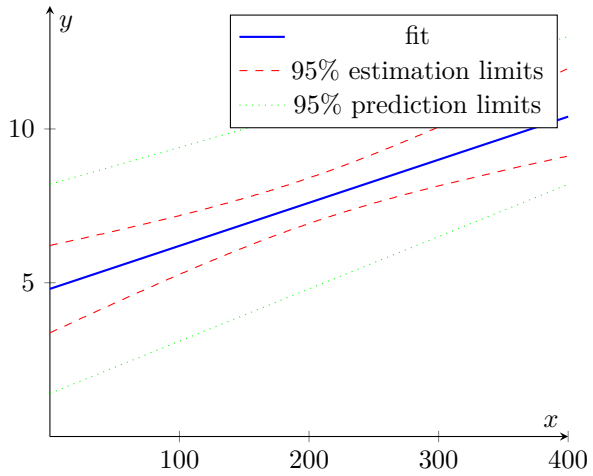
$$\hat{\beta}_1 \pm t_{\alpha/2} s_{\hat{\beta}_1}$$

where $s_{\hat{\beta}_1} = \frac{s}{\sqrt{SS_{xx}}}$ and $t_{\alpha/2}$ is based on $n-2$ degrees of freedom. $t_{\alpha/2}$ is obtained from the STUNDENT's t-distribution in the usual way. Some statistical software provides these values by default, but often (e.g. in MATLAB), only the standard errors for parameter estimates are provided.

4.4 Estimation and prediction for simple linear models

Definition 4.4 (estimation)
Estimation: estimate mean value of Y over many data points

Definition 4.5 (prediction)
Prediction: predict Y for a particular value of x. This leads to higher error bounds (add error in mean to variation around mean).



Be careful with predictions far away from mean of explanatory variable or outside of region covered by data.

It's good practise to always check the fit visually[2]. Looking at residual plots is also useful.

Definition 4.6 (Coefficient of Determination)
The Coefficient of Determination, R^2, is defined as

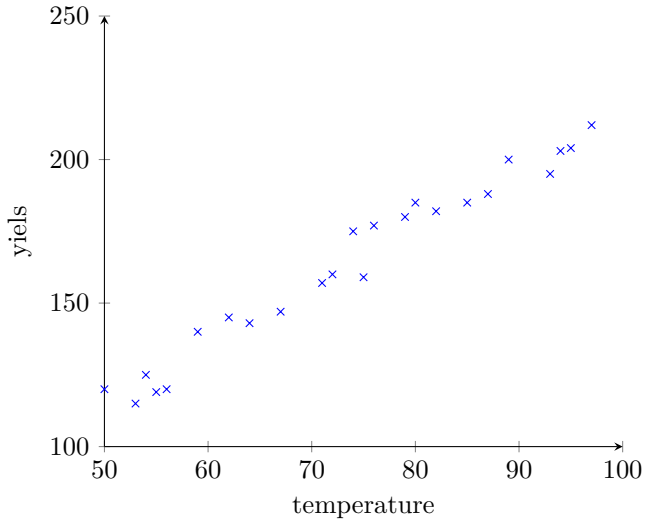
$$R^2 = \frac{SS_{yy} - SSE}{SS_{yy}}$$

where $SS_{yy} = \sum(Y_i - \bar{Y})^2$ and $SSE = \sum(Y_i - \hat{Y}_i)^2$. R^2 can be interpreted as the proportion of the variance in the response variable that is explained by (or attributed to) the explanatory variable. There are other measures, similar to R^2 and there are a few issues making it problematic for assessing goodness of fit. We'll revisit this later in the course.

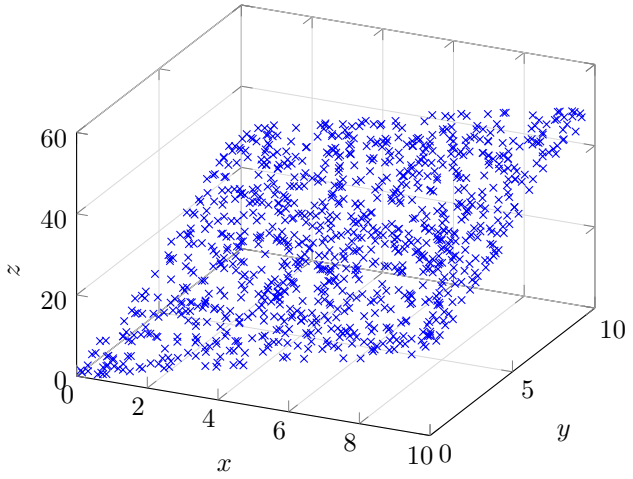
[2]If you don't believe that have a look at https://en.wikipedia.org/wiki/Anscombe%27s_quartet

5 Linear models (Multiple linear regression)

Last lecture we've looked at simple linear regression ($Y_i = \beta_0 + \beta_1 x_i + \varepsilon_i$)



This lecture we'll look at multiple linear regression (more than one predictor)



5.1 Structure of multiple linear regression models

Good news: simple linear models are a special case of multiple linear regression and with minor extensions everything we've looked at so far still applies. For p predictors and data tuples $\{(x_{1i}, x_{2i}, ..., x_{pi}, Y_i) \mid i = 1, ..., n\}$, we assume the relationship

$$Y_i = \beta_0 + \sum_{j=1}^{p} \beta_j x_{ji} + \varepsilon_i$$

As before, we assume $\varepsilon_i \overset{i.i.d}{\sim} Normal(0, \sigma)$.

The matrix notation, $Y = X\beta + \varepsilon$, now becomes very convenient:

$$Y = \begin{pmatrix} Y_1 \\ \vdots \\ Y_n \end{pmatrix} \quad X = \begin{pmatrix} 1 & x_{11} & \dots & x_{p1} \\ \vdots & \vdots & \vdots & \vdots \\ 1 & x_{1n} & \dots & x_{pn} \end{pmatrix} \quad \beta = \begin{pmatrix} \beta_0 \\ \vdots \\ \beta_p \end{pmatrix} \quad \varepsilon = \begin{pmatrix} \varepsilon_1 \\ \vdots \\ \varepsilon_n \end{pmatrix}$$

As before, the random variable notation is

$$Y_i \overset{i.i.d}{\sim} Normal\left(\beta_0 + \sum_{j=1}^{p} \beta_j x_{ji}, \sigma\right)$$
$$Y \sim Normal(X\beta, \mathbb{1}\sigma)$$

Model fitting using Maximum Likelihood Estimation (MLE) proceeds in the same way as for simple linear models seen before and is equivalent to Ordinary Least Squares (OLS) fitting. The likelihood function is given by

$$L(\beta \mid X) = \prod_{i=1}^{n} f_{Normal}\left(Y_i, \mu = \beta_0 + \sum_{j=1}^{p} \beta_j x_{ji}, \sigma\right)$$

where f_{Normal} is the PDF for the normal distribution evaluated at Y_i. The exact equations for parameter MLEs are

$$\begin{aligned} \hat{\beta} &= (X^T X)^{-1} X^T Y \\ \hat{\sigma}^2 &= \frac{1}{n-(p+1)} (Y - X\hat{\beta})^T (Y - X\hat{\beta}) \end{aligned} \tag{1}$$

Fitted values for the response are: $\hat{Y} = X\hat{\beta}$. Parameter estimates for one explanatory variable describe the relationship between this variable and the response variable when all other explanatory variables are held fixed.

5.2 Assumptions of linear models

The assumptions for multiple linear regression models are the same as for simple linear models:

- **Linearity:** response variable is a linear combination of the explanatory variables
- **Normality:** errors follow a normal distribution
- **Homoscedasticity:** variance of the response variable (or errors) is constant.
- **Independence:** the errors are uncorrelated (ideally statistically independent).
- **Weak exogeneity:** the explanatory variables can be treated as fixed values, rather than random variables.

However, there is one important addition:

- **Lack of perfect multicollinearity:** if two explanatory variables are perfectly correlated, we can not solve the equation for parameter estimates (eq. (1)). Some (but not perfect!) correlation between explanatory variables my be permissible.

Model checking works in the same way as for simple linear models, using residuals. Additionally, may want to plot each explanatory variables versus residuals.

5.3 Hypothesis testing on linear model parameters

Using a conceptionally similar approach as for simple linear models, we can test hypothesis about the βs and construct confidence intervals for them. Importantly, the test statistics depend on the entries of the matrix

$$(X^T X)^{-1} = \begin{pmatrix} c_{00} & c_{01} & \dots & c_{0p} \\ \vdots & \vdots & \vdots & \vdots \\ c_{p0} & c_{p1} & \dots & c_{pp} \end{pmatrix}$$

as $\sigma_{\hat{\beta}_j} = \hat{\sigma} \cdot \sqrt{c_{jj}}$ $(j = 1, ..., p)$. Off-diagonal entries determine the covariance of estimates and are important for the variance in prediction. Generally, σ needs to be estimated, so in practise we use Student's t-distribution and the test statistic

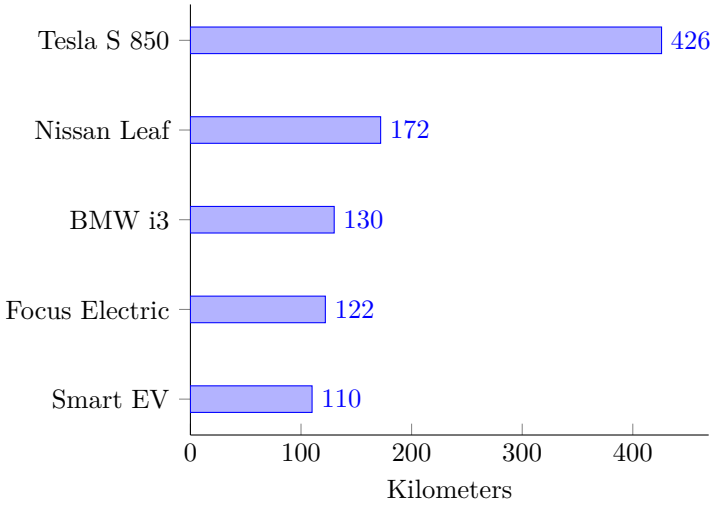
$$t = \frac{\hat{\beta}_j - \text{hypothesised value}}{\hat{\sigma}\sqrt{c_{jj}}}$$

Many details omitted - software does the work for us!

5.4 Model selection

With the more general Linear Model formulation, we can create many different models. How to choose which model to use?

■ **Example 5.1**
Consider the range of electric vehicles.



Depends on battery size, driving style or ambient temperature? Or all factors?

The right model depends on the purpose: Find all relevant factors? Prediction of range on a given day?

There is not one correct approach to model selection. What makes a good model depends on what it is designed for. Consequently, there are may different tools and techniques. Broadly, there are three types of model selection approaches:

- Hypothesis tests on model parameters within models (include/exclude parameters).
- Measures that describe the quality or goodness of fit of models.
- Hypothesis tests comparing the fit of entire models (often related to measures from the previous point).

Fundamentally, model selection is at the heart of scientific inquiry. Statistical techniques offer one quantitative and rigorous approach. We'll go through a few of the most common statistical techniques.

5.4.1 Hypothesis tests on model parameters

We've encountered these already (test H_0: $\beta_j = 0$). Could determine our model by fitting a full model that includes all conceivable explanatory variables first and than removing the explanatory variables for which we can't reject H_0.

Problems:

- Multiple comparisons: conducting many hypothesis tests means that some may be rejected/accepted by chance (there are ways of dealing with this, e.g. Bonferroni correction).
- The model fit changes every time we remove an explanatory variable.

While these tests are useful, we may want other approaches that allow us to test global hypotheses about models.

5.4.2 R^2 and adjusted R^2

Last lecture we looked at the coefficient of correlation, R^2 (proportion of variance in response explained by explanatory variables). Could use this to distinguish between models: the closer to 1 it is, the better the model.

Problem: R^2 always increases when we include more explanatory variables.

> **Proposition 5.2 (Occam's Razor by William of Occam, 1287-1347)**
> Entities are not to be multiplied without necessity (Latin: Non sunt multiplicanda entia sine necessitate.) Basically: prefer simpler models.

Solution: Adjusted R^2, R^2_a: unlike R^2, this takes into account the sample size n and the number of model parameters.

Warning: I caution against relying on R^2_a or R^2. They can only be usefully applied in particular circumstances.

5.4.3 F-test on linear models

Suppose we want to test a more general, global hypothesis about a linear model:
H_0: $\beta_1 = \beta_2 = \cdots = \beta_p = 0$
H_A: At least one of the $\beta_j \neq 0$
i.e. we compare our specified model to a constant model that only has an intercept β_0. This is called F-test or Analysis of Variance (Global). It uses the F-distribution and the test statistic can be expressed in terms of R^2 (we skip the details).

Warnings:

- The test is very specific (we'll see a more general test in a moment). It asks if the improvement in model fit is larger than expected under H_0.
- It makes several assumptions (e.g. normally distributed errors), so check model assumptions hold.
- It doesn't tell us whether our model is correct.

5.4.4 Quality measures based on the likelihood

Techniques so far rely on the variance in the response explained by models. Recall the analogy for MLE of maximising $P(data \mid parameters)$. Could use the maximum likelihood of models, $\hat{L}$, to compare them: highest $\hat{L}$ indicates best model.

Warning: the likelihood faces the same problem as R^2 - it will always improve when more parameters are added to the model.

Solution: penalty for parameters in measures for relative model quality.

- **Akaike Information Criterion:**

$$\text{AIC} = 2k - 2\ln(\hat{L})$$

  where k is the number of model parameters, including the intercept, but typically excluding the error variance (as this is included in all models). Model with smallest AIC is best.

- **Bayesian Information Criterion:**

$$\text{BIC} = \ln(n)k - 2\ln(\hat{L})$$

  Same idea, but penalty for parameters is stronger (based on different assumptions).

Warning: candidate models must be fitted to the same response data.

5.4.5 Likelihood-ratio test for nested models

The Likelihood-ratio test is a much more general version of the F-test. Consider a linear model, M_1 with parameters $\beta_{M_1} = \{\beta_1, \beta_2, ..., \beta_p\}$. The test considers a restriction M_0 of M_1 where e.g. $\beta_{M_0} = \{\beta_1, \beta_2, ..., \beta_q, 0, ..., 0\}$ with $q < p$. We say M_0 is nested in M_1. We test the hypothesis:
H_0: $\beta = \beta_{M_0}$
H_A: $\beta = \beta_{M_1}$

using the test statistic

$$D = 2\ln\left(\frac{\widehat{L_{M_1}}}{\widehat{L_{M_0}}}\right) = 2\Big(\ln\left(\widehat{L_{M_1}}\right) - \ln\left(\widehat{L_{M_0}}\right)\Big)$$

Under some conditions and assuming H_0, D is asymptotically distributed as $D \sim \chi^2_{p-q}$. This is a flexible and very useful test (e.g. tests on individual parameters). Also works when error distribution is not normal.

5.5 Automated or standardised model selection strategies

Model selection is time consuming. So people have tried to come up with standardised or even automated procedures, e.g. stepwise regression. Many software packages, including MATLAB, implement such procedures. There are many flavours, but the basic idea is:

1. Identify response Y and all potentially important explanatory variables $x_1, ..., x_p$.
2. Automatically work through models defined by all possible combinations of the x_js, starting with simpler models.
3. At each step use standard hypothesis tests (or other measures) to assess if additional explanatory variables improve model.
4. Continue until some stopping criterion is reached.

Warning: control of the process is relinquished to software that can be very intricate. Procedure conducts many statistical tests (multiple comparisons!). Because of these and other issues, many statisticians recommend not to use this approach.

use	**technique**
Hypothesis tests for single parameters	t-test, Likelihood-ratio test
Hypothesis tests for several parameters	F-test, Likelihood-ratio test
Measures for relative quality of models	R^2, R^2_a, AIC, BIC

Warning: none of these techniques tell you anything about the correctness of a model. So checking model assumptions separately is important!

Remember, none of these techniques are perfect. Choose your approach depending on what you are trying to achieve with your models. Within this in mind, prediction intervals and residual plots could be also used in model selection.

6 Model building

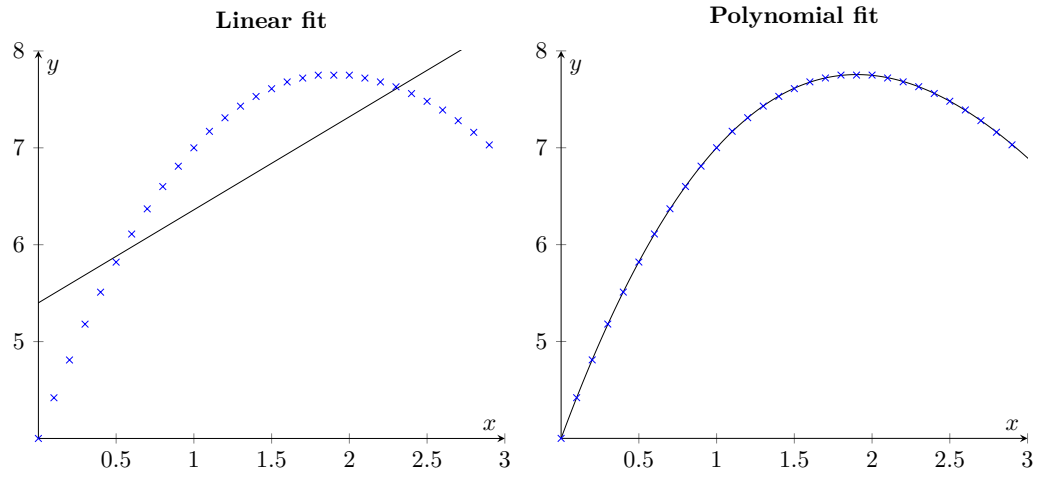
So far, we have looked at linear models of the form

$$Y_i = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \cdots + \beta_p x_{pi} + \varepsilon_i$$

In this lecture, we extend this concept:

- We look at how we can use different types of predictors to capture a wide range of relationships (e.g. capture non-linear relationship with linear model).
- We'll think about how to formulate models, depending on their use.
- We'll consider some common pitfalls in use of linear models.

■ **Example 6.1**
Consider energy usage in houses, based on their size.



A simple linear models might tell us that a predictor is important, but it's no good prediction. Model building is a process that includes:

- **Formulating a model:** model structure, we'll look at different types of predictors today.
- **Model fitting:** last two lectures
- **Model evaluation:** Check model assumptions hold, avoid common pitfalls

Before formulating a statistical model, it is good practice to explore the data. Look at scatter plots of predictors against the response and predictors against each other. Can help to develop an intuition for model structure. Look at distributions of response and predictors (e.g. look for outliers, can you capture the distribution with your model?) Think about what the model will be used for (e.g. prediction, or simply to find relevant explanatory variables).

6.1 Types of predictors

6.1.1 qualitative vs quantitative

So far, we have looked at quantitative predictors (numerical variables), e.g. temperature, energy usage, waiting time before computer processes data, ... But we can have qualitative predictors as well (categorical variables), e.g. type of engine, type fuel used, type of processor used, ... These are included in models using dummy variables.

■ **Example 6.2**
Consider the performance Y_i of diesel engines for three different fuel types A, B and C. We use the model $Y_i = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \varepsilon_i$, where x_{1i} and x_{2i} are dummy variables such that

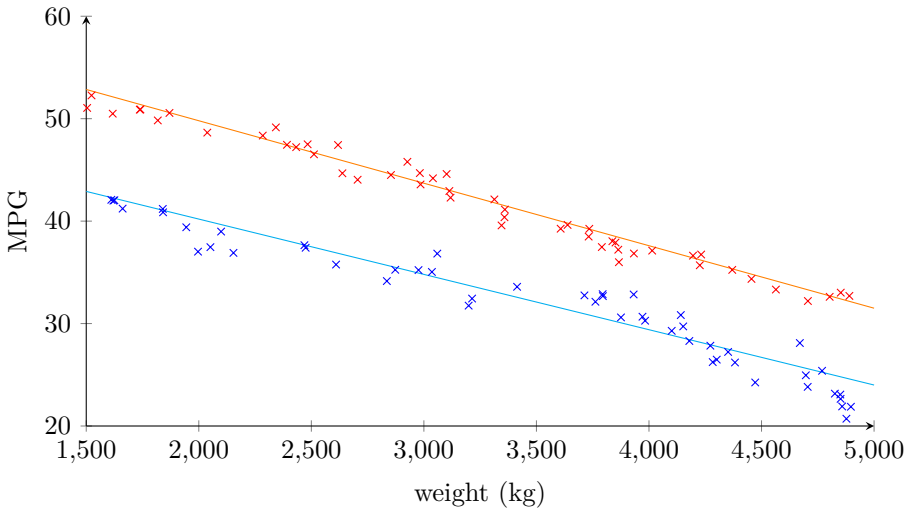
$$x_{1i} = \begin{cases} 1 & \text{if fuel B is used} \\ 0 & \text{if not} \end{cases}$$

$$x_{2i} = \begin{cases} 1 & \text{if fuel C is used} \\ 0 & \text{if not} \end{cases}$$

The performance for fuel type A is captured in β_0.

Qualitative and quantitative predictors can be combined in models.

■ **Example 6.3**
Consider fuel efficiency of car models from the 1950s (type 1) and from the 1960s (type 2), depending on their weight. For car models $i = 1, ..., n$, we might consider the model $Y_i = \beta_0 + \beta_1 x_i + \beta_2 w_i + \varepsilon_i$, where Y_i is the efficiency in miles per gallon (MPG), x_i is a dummy variable for the car type and w_i is the weight of a car model. We might find:



6.1.2 interaction terms

So far, we have assumed that the effects of all explanatory variables are additive, e.g. as is $Y_i = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \varepsilon_i$. What if the relationship between Y_i and x_{1i} depends on the value of x_{2i}?

Then we need to consider interaction terms in our model, e.g. $Y_i = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \beta_3 x_{1i} x_{2i} + \varepsilon_i$. Interpretation of model parameters:

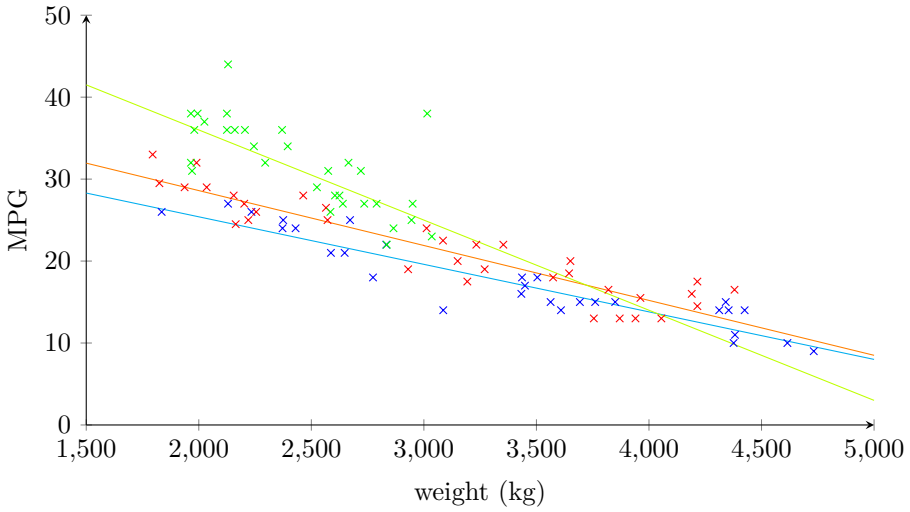
- $(\beta_1 + \beta_3 x_2)$ represents the change in Y for every unit increase in x_1, holding x_2 fixed.
- $(\beta_2 + \beta_3 x_1)$ represents the change in Y for every unit increase in x_2, holding x_1 fixed.
- In model checking: if interaction term is important, then the interacting explanatory variable must be important and t-tests on them are meaningless.

This is still a linear model - the effects captured by parameters are additive.

■ **Example 6.4**
Fuel efficiency Y_i of car models from 1970, '76 and '82 depending on their weight w_i. Consider interaction between year built and weight.

$$Y_i = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \beta_3 w_i + \beta_4 x_{1i} w_i + \beta_5 x_{2i} w_i + \varepsilon_i$$

This model has main effects for year and weight and interaction terms. Model fitting:



To test if interactions are important, could use the Likelihood-ratio test (H_0: $\beta_4 = \beta_5 = 0$).

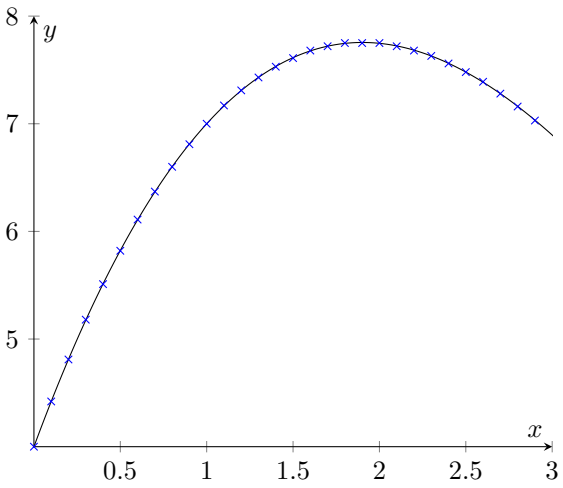
6.1.3 polynomials of predictors

How to deal with non-linear relationships between variables?

Can account for this in models using polynomials of predictors, e.g. $Y_i = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{1i}^2 + \varepsilon_i$, where β_0 is the intercept, β_1 is a shift parameter and β_2 is the rate and direction of curvature. Higher-order polynomials are also possible.

This is still a linear model - the effects captured by parameters are additive.

■ **Example 6.5**
Consider data on energy usage in houses depending on their size and second-order model $Y_i = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{1i}^2 + \varepsilon_i$.



Interpreting model parameters:

- β_0 can only be interpreted directly if range of data includes $x = 0$.
- β_1 no longer represents a slope and can't be interpreted in isolation.
- The sign of β_2 indicates the direction of the curvature (concave upward or downward).

Warning: polynomials of predictors lead to correlations between predictors by design.

6.1.4 data transformations

Data transformations can help to address problems in model fit (e.g. when relationship is not linear, but residuals are normally distributed). There are many data transformations and there are two examples below, but in general I'd recommend to be cautious about this.

- Log-transform predictors, response or both
- Code predictors, so that their range is similar. E.g. for temperature T, code this as $x = \frac{T-100}{50}$. Can reduce computational rounding errors in model fit and can help to address problems with multicollinearity in polynomial regression models.

Warning: Data transformations do make residuals more normal and statistical tests performed on transformed data are not necessarily relevant for the original data.[3]

6.2 Pitfalls

Often there is not one correct way to build a model, especially for large data sets with many predictors.

[3] Feng et al. (2014) *Shanghai Arch Psychiatry.* 26: 105-109

Proposition 6.6 (George Box, 1919-2013)
Essentially, all models are wrong, but some are useful.

Whether a model is useful or not often depends on how it is used. Things to consider are: prediction or data analysis, exploratory or for decision-making. There are many wrong ways of doing things. Common pitfalls are listed in the following.

- **Multicollinearity** arises when predictors are correlated. This can be data-based or structural when new predictors are created from existing ones (e.g. in polynomial regression). This causes problems: parameter estimates can't be interpreted sensibly and statistical tests on them are meaningless. However, prediction from models within the range covered by the data is not affected by multicollinearity.
- **Model assumptions are violated.** See previous lectures for model assumptions (e.g. normality and independence of errors).
- **Extrapolation beyond the scope of the model.** Trends identified in data by a model do not necessarily hold beyond the range of the data[4].
- **Excluding important predictors.** Can lead to models that contain misleading associations between variables. Avoid by data exploration and considering background information on data.
- **Parameter interpretation.** A common misconception is that parameter estimates always measure the effect of a predictor on the response independent from other predictors (e.g. not the case in models with interactions). Another misinterpretation is that significant p-values for a parameter indicate a cause-and-effect relationship. Unless we control for all other effects, they do not.
- **Overfitting.** Recall Occam's Razor (Proposition 5.2). An extreme case of overfitting is to use as many model parameters as there are data points. In less extreme cases, including too many predictors makes interpretation difficult.
- **Power and sample size.** Small data sets can lead to poorly fitted models with large standard errors for parameter estimates. The more data, the better. General guidelines, such as number of data points per predictor, are not possible, as they depend on the context (e.g. effect size, variability in the data).

[4] https://xkcd.com/605/

7 Experimental design and ANOVA

Recap: statistical analysis with linear models. Selecting the right model is a key challenge. Ideally, data collection and model building go hand in hand. We can analyse observational data (data observed in natural setting) or experimental data (we control the explanatory variables). The former finds correlations, the latter can establish causal links.

7.1 Designing an experiment

Proposition 7.1
The study of experimental design originated with R. A. Fischer's work in the UK in the 1900s.

In an experiment, we collect data in a structured way. We decide on:

- what to measure - **response**
- what to measure the response on - **experimental unit**
- the independent variables whose effect we study - **factors**
- the combinations of factors we want to study - **treatments**
- how many data points to collect - **sample size**
- how to assign treatments to experimental units

■ **Example 7.2**
Compare the range of three electric car models. Range is the response, car model is a factor with three levels. With only one factor, there are no further treatments. Experimental units are cars and we can decide how many of each model we want to test.

The more data we collect, the more certain we can be about trends we find. However, collecting data is often expensive. Measurement errors, environmental variation or other effects lead to noise in data. Noise-reducing experimental design can be used to counter this. Once we have identified factors we want to investigate, we have to think carefully about which treatment to test, to get the most out of our experiment (volume-increasing design).

7.1.1 Noise-reducing design

Noise-reducing designs assign treatments to experimental units in such a way that extraneous noise is reduced. The simplest approach: completely randomised design - treatments assigned randomly to experimental units.

■ **Example 7.3**
Length of time to assemble a watch using three different methods A, B and C. Select 15 workers and assign them randomly to A, B or C.

But: assembly times could vary substantially between workers. This could skew our findings. To avoid this, we could get 5 workers to each use A, B or C in turn (randomised block design).

In randomised block design, we compare p treatments by b blocks. Each block contains p relatively homogeneous (or identical) experimental units. The p treatments are assigned randomly to experimental units in each block (one experimental unit assigned per treatment).

These experimental designs can be captured in linear models to investigate differences in the mean response across treatments. For the watch example:

- Mode for completely randomised design: $Y_i = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \varepsilon_i$,

$$x_{1i} = \begin{cases} 1 & \text{if worker } i \text{ uses method A} \\ 0 & \text{if not} \end{cases}$$

$$x_{2i} = \begin{cases} 1 & \text{if worker } i \text{ uses method B} \\ 0 & \text{if not} \end{cases}$$

  ... selected method C as the base level.

- Model for randomised block design:

$$Y_i = \beta_0 + \underbrace{\beta_1 x_{1i} + \beta_2 x_{2i}}_{\text{treatment effects}} + \underbrace{\beta_3 x_{3i} + \beta_4 x_{4i} + \beta_5 x_{5i} + \beta_6 x_{6i}}_{\text{block effects}} + \varepsilon_i$$

  ... x_{3i} up to x_{6i} are dummy variables for which worker assembles.

We can use the usual methods to test hypothesis on our data (e.g. t-test for individual parameters, F-test, Likelihood-ratio test for nested models).

7.1.2 Volume-increasing design

Volume-increasing designs combine factors in experiments into treatments that are maximally informative.

■ **Example 7.4**

An electricity company wants to measure customer satisfaction for two levels of peak time price increase, x_1, and two different peak period lengths, x_2 (2 levels). How should the levels of factors x_1 and x_2 be combined into treatments?

- Option 1: keep one factor fixed and vary the other. This is consistent with block designs. However, it misses interactions between factors.
- Option 2: consider all possible combinations of factor levels (complete factorial design). For this example, we call it a 2×2 factorial design.

Warning: if many factors are tested, complete factorial designs require a lot of treatments.

Complete factorial designs can be captured in linear models with interaction terms. For this example the model is

$$Y_i = \beta_0 + \underbrace{\beta_1 x_{1i} + \beta_2 x_{2i}}_{\text{main effects}} + \underbrace{\beta_3 x_{1i} x_{2i}}_{\text{interaction}} + \varepsilon_i$$

... x_{1i} and x_{2i} are dummy variables for peak time price increase levels and peak period length levels,

respectively. The number of parameters is the same as the number of treatments. This is always the case for complete factorial designs. Thus, we need replicate measurements for each treatment.

Aside: this also works for quantitative predictors.

7.2 Selecting the sample size

Deciding how many data points to collect is important: On the one hand, the more data we have, the more certain we can be about observed trends (e.g. standard errors for parameter estimates in lecture 6). On the other hand, collecting data is expensive, so we only want to collect what's necessary.

Power analysis allows us to determine the sample size required to detect an effect of a given size with a given degree of confidence. In general, the smaller the effect and the more confident we want to be, the more replicates we need.

7.3 Introduction to ANOVA

7.3.1 One-way ANOVA

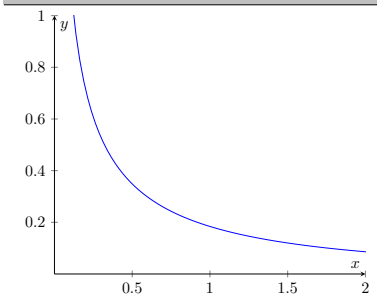
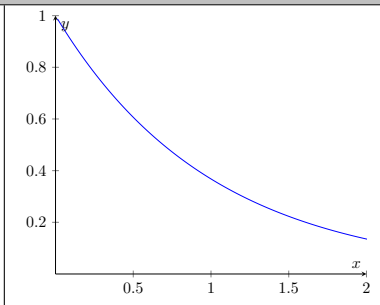
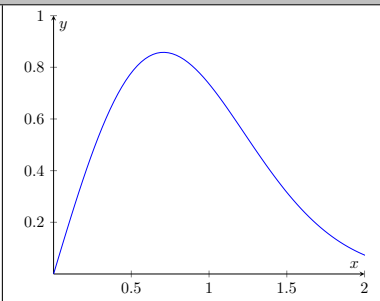
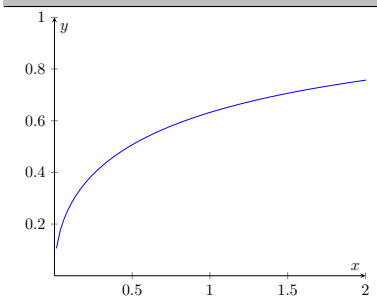
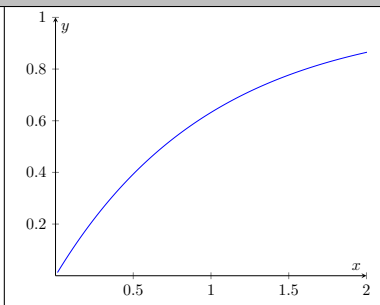
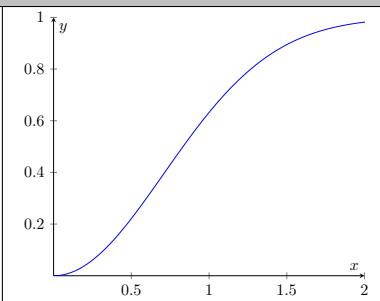
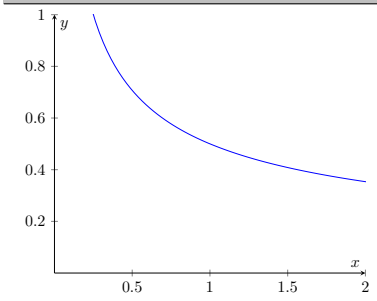
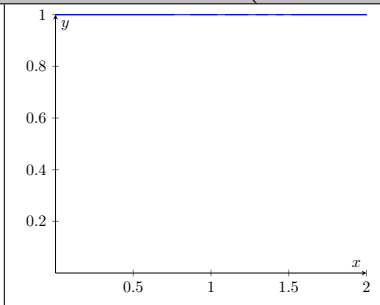
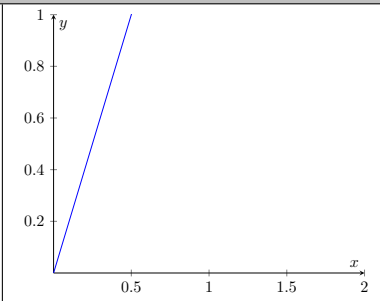
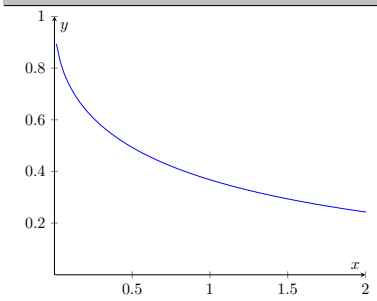
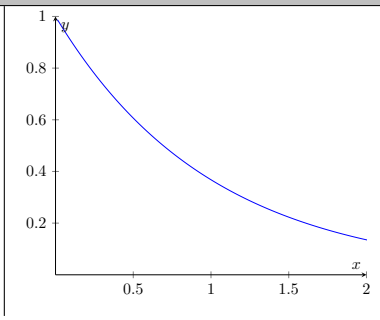
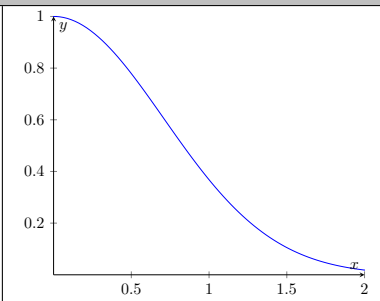
7.3.2 Two-way ANOVA

7.4 Observational data - sampling

8 Generalised linear models

9 Appendix

9.1 Weibulls Distribution - Graphs

$shape = 0.5$	$shape = 1$	$shape = 2$
Weibull PDF ($scale = 1$)		
		
Weibull CDF ($scale = 1$)		
		
Hazard function $\left(h = \frac{\text{PDF}}{1-\text{CDF}}\right)$		
		
Survival function ($R = 1 - \text{CDF}$)		
		

Index