Applied statistics (spring term 2019)

readers: Nikolai Bode and Ksenia Shalonova

14th February 2019

Contents

1 Estimating parameters

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 $\hat{p} \pm 2 \cdot \sqrt{\frac{p(1-p)}{n}}$

■ **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 $n(1 - \hat{p}) > 5$.

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.

1.1 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	$l(\lambda\|x_1,...,x_n) = n\log(n) - \lambda \sum x_i$
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.

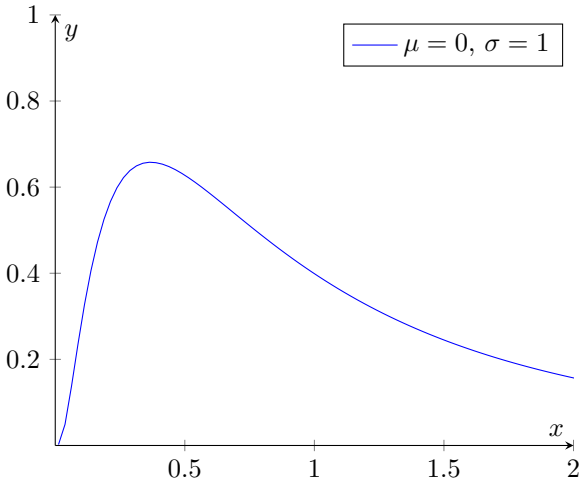
The likelihood function is maximised at a turning point of the likelihood function and could therefore be found by setting the partial derivatives of $L(\theta, \varphi)$ with respect to θ and φ to zero.

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!

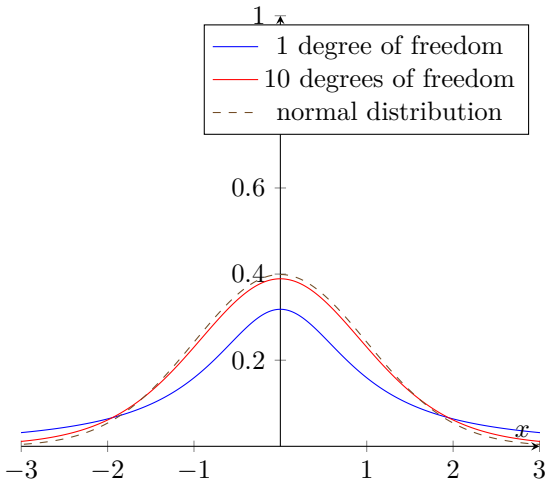
1.2 Continuous distributions

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.



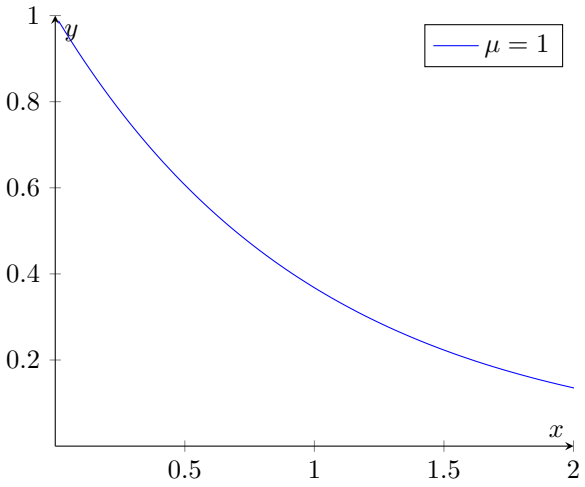
In the beta distribution the uncertain variable is a random value between 0 and positive value. The distribution is frequently used for estimating the proportions and probabilities (i.e. values between 0 and 1). The shape of the distribution is specified by two positive parameters.

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.

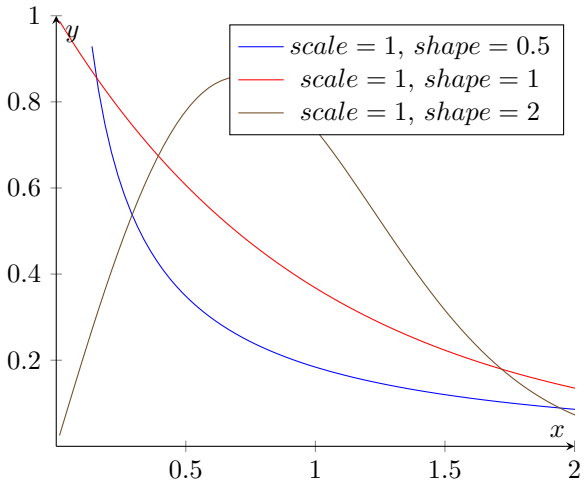


The chi-square distribution is usually used for estimating the variance in a normal distribution.

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 "'$\alpha = \ldots$"'. 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.

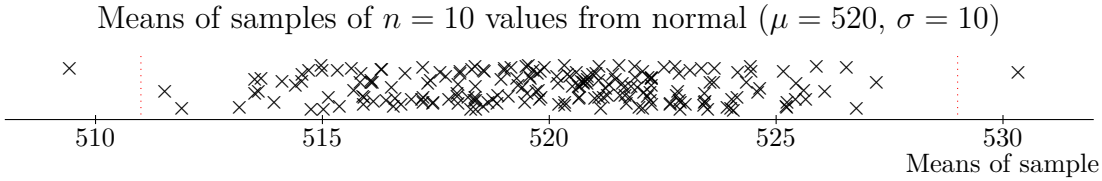
$$H_0 : \lambda \leq \frac{1}{100} \text{ (claim of a manufacturer)}$$
$$H_A : \lambda > \frac{1}{100}$$

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.

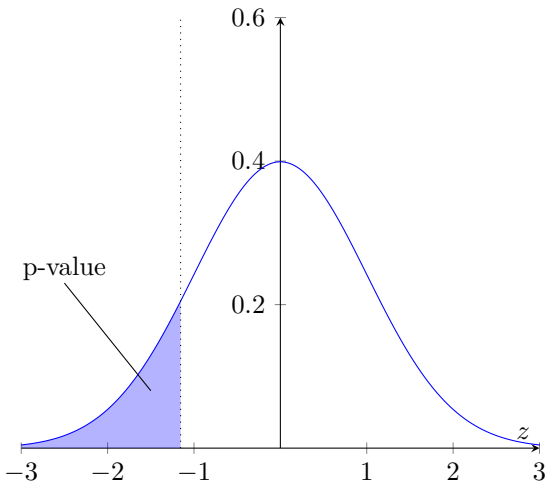
p-value	**Interpretation**
over 0.1	no evidence that H_0 does not hold
between 0.05 and 0.1	very weak evidence that H_0 does not hold
between 0.01 and 0.05	moderately strong evidence that H_0 does not hold
under 0.01	strong evidence that H_0 does not hold

■ **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).

$$z = \frac{\bar{x} - 18.3}{\underbrace{\frac{7.1}{\sqrt{30}}}_{\text{standard error}}} = -1.155$$



$$\text{p-value} = P(z \leq -1.155) = 0.124$$

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.

■ **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.

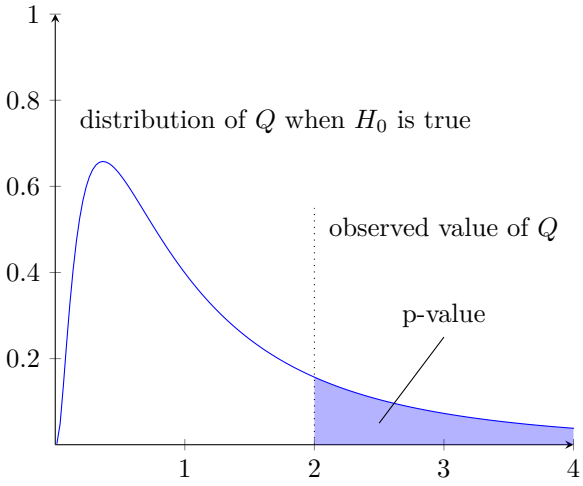
null hypothesis	H_0: $\mu = 15$
alternative hypothesis	H_A: $\mu > 15$
T-test for μ	$\bar{x} = 16.138$, $z = \frac{16.138-15}{\frac{2.154}{\sqrt{13}}} = 1.906$, $P(t \geq 1.906) = 0.040$ (t-distribution with 12 degrees of freedom)
p-value interpretation	Since this is below 0.05, we conclude that there 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	$P(T > 1.782) = 0.05$
Conclusion	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)

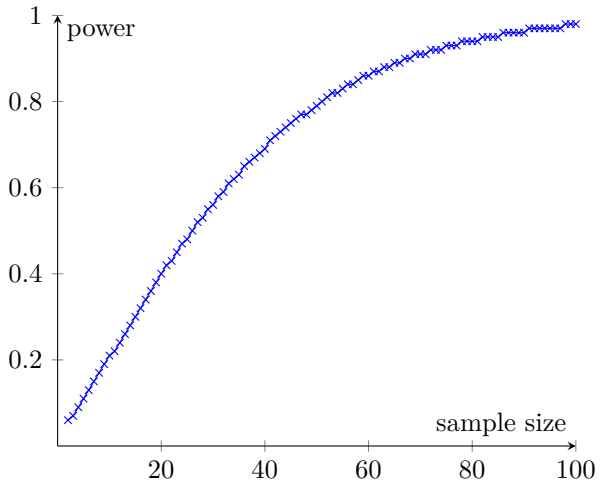
Computer software can provide the p-value for a hypothesis test at 5% or 1% significance level (Type 1 error).

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).

3 Bootstrapping

4 Linear models (Simple linear regression)

5 Linear models (Multiple linear regression)

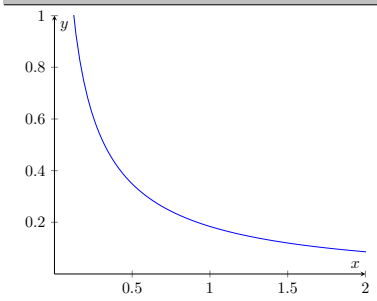
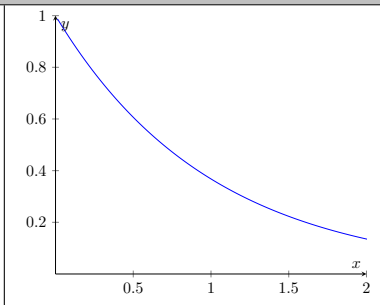
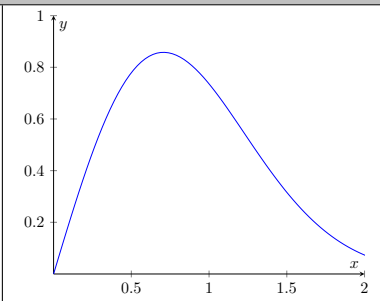
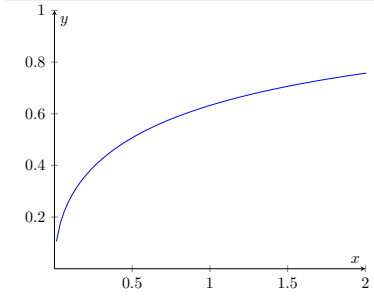
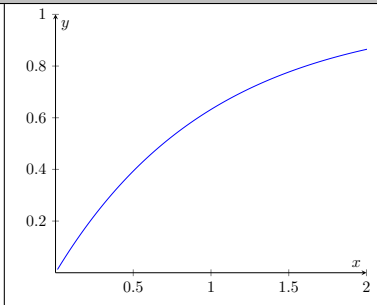
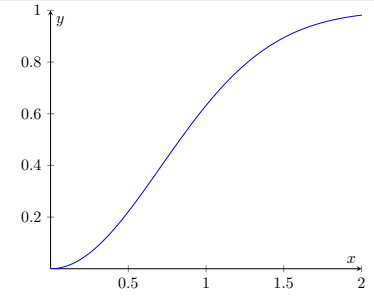
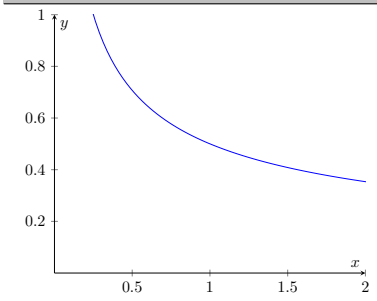
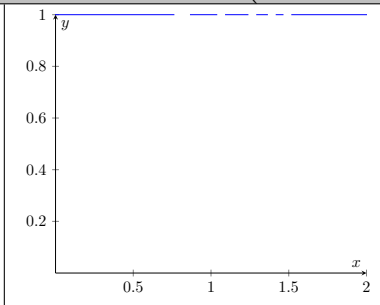
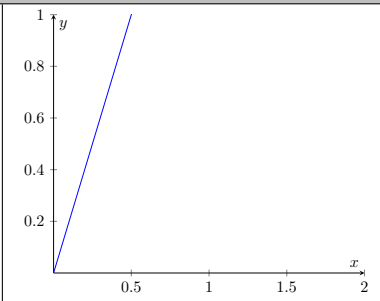
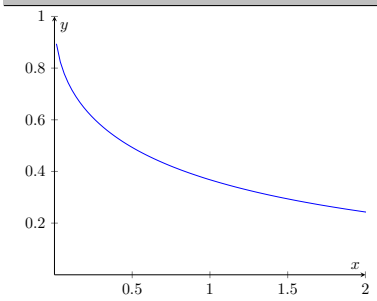
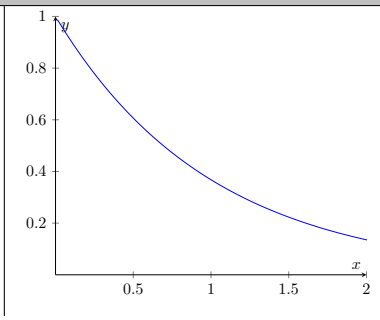
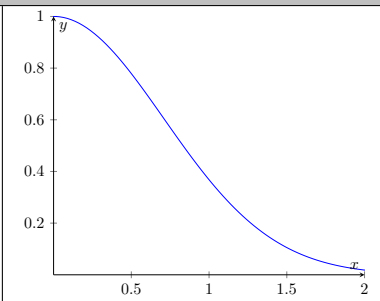
6 Model building

7 Experimental design and ANOVA

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