DEPARTMENT OF ENGINEERING, AARHUS UNIVERSITY, DENMARK

Lecture Notes in Stochastic Modelling and Processing

Introduction to statistics

Henrik Pedersen

11/24/2014

TABLE OF CONTENTS

1 MOTIVATIONAL EXAMPLE – CUP FILLING MACHINE

A machine fills cups with a liquid, and is supposed to be adjusted so that the content of the cups is 250 grams of liquid. As the machine cannot fill every cup with exactly 250 grams, the content added to individual cups shows some variation, and is considered a random variable, X. This variation is assumed to be normally distributed (although this assumption is not necessary for the theory to work) around the true mean, μ, of 250 grams, with a standard deviation, σ, of 2.5 grams. See Figure 1.

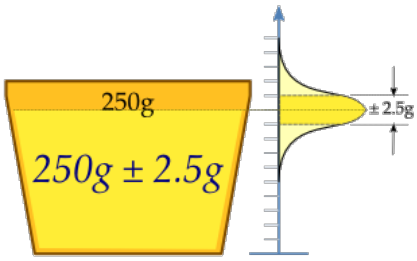


Figure 1 – When adequately calibrated the machine fills every cup with 250 grams on average. The deviation of the cup content from the average is assumed to be normally distributed with a standard deviation of 2.5 grams.

To determine if the machine is adequately calibrated, a sample of $n = 25$ cups of liquid is chosen at random and the cups are weighed. The resulting measured masses of liquid are $X_1, X_2, \dots, X_{25}$, a random sample from X.

To get an impression of the expectation or mean value, $E[X] = \mu$, it is sufficient to give an estimate. The appropriate estimator here is the *sample mean* (i.e., the average):

$$\hat{\mu} = \frac{1}{25}\sum_{i=1}^{25} X_i \qquad \textbf{(1)}$$

Let us assume that we observe a sample mean value $\hat{\mu}_{obs} = 250.2$ grams. If we take another sample of 25 cups, we could easily expect to find sample mean values like 250.4 or 251.1 grams due to random variation in X. A sample mean value of 280 grams however would be extremely rare if the mean content of the cups is in fact close to 250 grams.

The deviation between μ and $\hat{\mu}_{obs}$ is 0.2 grams. The question is; is the observed deviation due to inadequate calibration of the machine, or is it just a coincidence that results from random variation in X? (In the latter case, the machine would be adequately calibrated.).

This is a statistical question!

To answer this question, we will need to investigate what values of the sample mean we would typically observe if the machine is adequately calibrated (see Figure 2). If the observed sample mean ($\hat{\mu}_{obs} = 250.2$) turns out to be a very rare event, the 0.2 gram deviation cannot be explained random variation in X. This would suggest that the machine is not adequately calibrated. Conversely, if the observed sample mean is a common event, the deviation can be explained by random variation in X, and we could conclude that the machine is adequately calibrated.

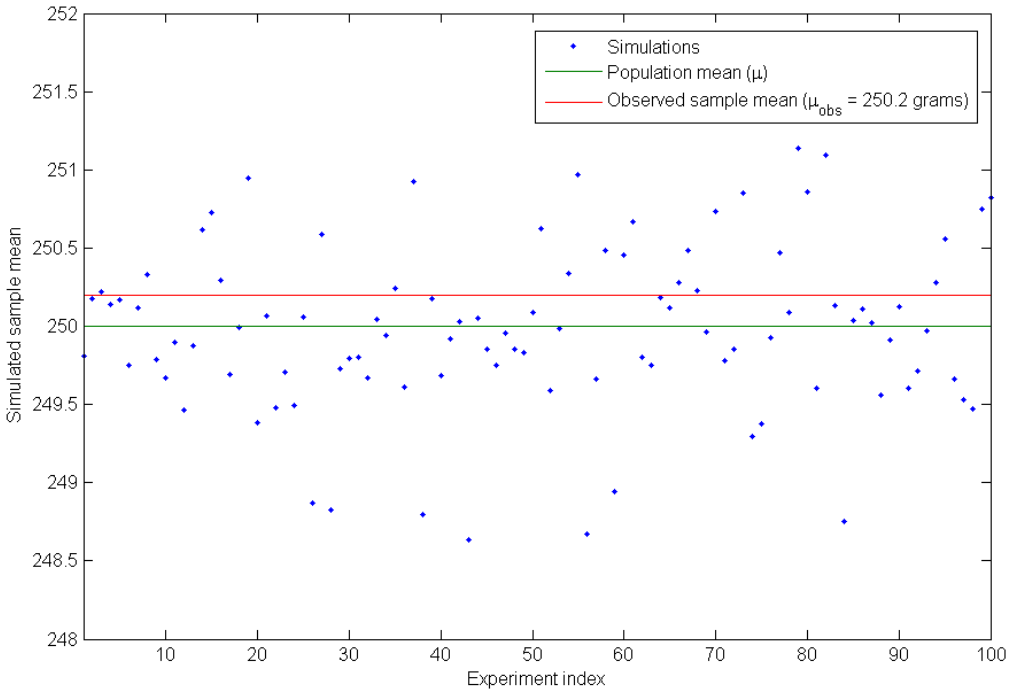


Figure 2 – 100 repeated measurements of the mean content of 25 cups, under the assumption that the true mean is 250 grams, and the true standard deviation is 2.5 grams.

To find out whether the observed sample mean ($\hat{\mu}_{obs} = 250.2$) is common or not, we could perform 100 simulations of the cup filling experiment, where in each experiment we assume that the machine is adequately calibrated. The result of each experiment is a sample mean, $\hat{\mu}$, which is calculated by first drawing 25 random samples ($X_1, X_2, \dots, X_{25}$) from a normal distribution with mean $\mu = 250$ grams and standard deviation $\sigma = 2.5$ grams, and subsequent insertion into equation (1).

The result of 100 such simulations is shown in Figure 2. Each blue dot is a simulated sample mean, which we will denote $\hat{\mu}$. The green line shows the population mean ($\mu = 250$), and the red line represents the observed sample mean ($\hat{\mu}_{obs} = 250.2$). By counting, we realize that there are 28 experiments where $\hat{\mu} \geq \hat{\mu}_{obs}$ (i.e., all blue points above the red line). A loose interpretation of this result is that – just by random variation in X – there is a 28% chance of observing a sample mean that is larger than what we have observed ($\hat{\mu}_{obs}$), even though the machine is adequately calibrated. We can write this as

$$\Pr(\hat{\mu} \geq \hat{\mu}_{obs}) = \Pr(\hat{\mu} \geq 250.2) \approx 0.28$$

This means that observing a sample mean of $\hat{\mu}_{obs} = 250.2$ or larger, when the machine is adequately calibrated, is a relatively common event. Hence, the 0.2 gram deviation can be explained by random variation in X, and we could conclude that the machine is adequately calibrated.

Figure 3 shows the Matlab code used to generate the data in Figure 2. The Matlab command, `randn`, outputs pseudo-random numbers from a standard normal distribution (i.e., with mean 0 and standard deviation 1). We need to convert these random numbers into numbers drawn from a normal distribution with mean `mu=250` and standard deviation `sigma=2.5`. Recall that if x is a normally

distributed random variable with mean μ and variance σ^2, written $x \sim N(\mu, \sigma^2)$, then the standardized random variable

$$z = \frac{x - \mu}{\sigma} \sim N(0,1) \qquad \textbf{(2)}$$

is standard normally distributed. Using `randn` in Matlab, we are given random numbers (z) that are standard normally distributed, and we want to convert these numbers into normally distributed numbers (x) with mean μ and variance σ^2. To do this, we simply isolate x in equation (2) to obtain

$$x = z \cdot \sigma + \mu$$

or in Matlab code (line 11)

```
X = randn(1,n)*sigma + mu;
```

```
 1   N        = 25;    % number of cups
 2   sigma    = 2.5    % standard deviation
 3   mu       = 250;   % mean if machine is correctly calibrated
 4   mu_hat   = [];    % estimates of the mean
 5
 6   % Do 100 repeated experiments
 7   for experiment = 1:100
 8
 9       % 25 observations drawn from a normal distribution with me
10       % mean 250 and standard deviation 2.5
11       X = randn(1,n)*sigma + mu;
12
13       % Store estimate of the mean (=average of observations)
14       mu_hat(experiment) = mean(X);
15   end
16
17   % Our observation in the example
18   mu_obs = 250.2;
19
20   % Plot the result of all 100 experiments
21   plot(1:100,mu_hat,'.',...
22        [1 100],[mu mu],...
23        [1 100],[mu_obs mu_obs])
24        legend('Simulations','Mean','Observed sample mean')
25        xlabel('Experiment index')
26        ylabel('Average content of 25 cups')
```

Figure 3 – Matlab source used to generate the data in Figure 2.

In statistics, the probability, $\Pr(\hat{\mu} \geq \hat{\mu}_{obs})$, of getting a sample mean value ($\hat{\mu}$) that is more extreme than the observed value ($\hat{\mu}_{obs}$) is called a *p-value*. The p-value is linked to a *hypothesis*. In the cup filling example, the hypothesis is that the samples $(X_1, X_2, \ldots, X_{25})$ come from a normal distribution with mean $\mu = 250$ grams and standard deviation $\sigma = 2.5$ grams. However, rather than expressing the hypothesis in terms of the samples, it is more convenient to express it in terms of a *test statistic*. The test statistic that we use here is

$$z = \frac{\hat{\mu}_{obs} - \mu}{\sigma/\sqrt{n}} = \frac{250.2 - 250}{2.5/\sqrt{25}} = 0.4$$

There are two important things to note about z: firstly, z depends on the observed sample mean ($\hat{\mu}_{obs}$), which is a random variable. Hence, z is also a random variable. Secondly, if the true population mean is indeed $\mu = 250$ grams, then z must be a sample from a standard normal distribution, which we write:

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

The reason that z must be standard normally distributed (provided that the true mean is 250 grams) follows from the central limit theorem. According to the central limit theorem, the sample mean is normally distributed with a mean that equals the population mean and a variance that equals the population variance divided by the number of samples (n). Mathematically this can be written

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

where we have implicitly assumed that the samples $X_1, X_2, \ldots, X_n$ are independent and identically distributed (i.i.d.) with mean μ and variance σ^2. We have seen in Cooper/McGillem chapter 2 that in order to standardize a random variable, we first subtract the true mean from it and then divide by its standard deviation. In the case of $\hat{\mu}_{obs}$, this is exactly

$$z = \frac{\hat{\mu}_{obs} - \mu}{\sigma/\sqrt{n}}$$

where we recall that the standard deviation is the square root of the variance, $\sqrt{\sigma^2/n} = \sigma/\sqrt{n}$. Note that if $\hat{\mu}_{obs}$ is indeed normally distributed with mean $\mu = 250$ grams and variance $\sigma^2/n = 2.5^2/25 = 0.25$ grams, then the test size z must be standard normally distributed.

Defining the random variable

$$Z = \frac{\hat{\mu} - \mu}{\sigma/\sqrt{n}}$$

the theoretical p-value (i.e., the probability of observing a sample mean that is more extreme than $\hat{\mu}_{obs}$=250.2 grams) can be calculated as

$$\begin{aligned}\Pr(\hat{\mu} \geq \hat{\mu}_{obs}) &= \Pr\left(\frac{\hat{\mu} - \mu}{\sigma/\sqrt{n}} \geq \frac{\hat{\mu}_{obs} - \mu}{\sigma/\sqrt{n}}\right) = \Pr(Z \geq z) = \Pr(Z \geq 0.4) = 1 - \Pr(Z \leq 0.4) = 1 - \Phi(0.4) \\ &= 1 - 0.6554 = 0.34\end{aligned}$$

where Φ denotes the CDF of a standard normal distribution. In order to find $\Phi(0.4)$ you can either use a lookup table or Matlab's built-in command, `normcdf`:

```
>> normcdf(0.4)

ans =

    0.6554
```

The theoretical p-value of 0.34 is slightly larger than the one we got from the simulations. On average, the simulations should yield a p-value that is equal to the theoretical value. For the record, the procedure of verifying a hypothesis is called a *hypothesis test*. In this example, because the p-value is large, we say the hypothesis is accepted. Had the p-value been small, say 0.01, we would reject the hypothesis. The level of acceptance is called the *significance level* and is often set to 0.05.

SUMMARY

In order to determine if the cup filling machine is adequately calibrated, we take a random sample of 25 cups and calculate an estimate ($\hat{\mu}_{obs}$) of the mean content of the cups. According to the central limit theorem then, if the machine is adequately calibrated, $\hat{\mu}_{obs}$ should be a random sample from a normal distribution with mean 250 and variance 0.25. This claim is called a hypothesis. The hypothesis is equivalent to saying that the test size z should be a random sample from a standard normal distribution. Here, we find that the probability of observing a test size that is more extreme than z is 0.34, which is a relatively high value. In other words, under the assumption that the hypothesis is correct (i.e., that the machine is adequately calibrated), there is a 34% chance of observing a sample mean that is larger than $\hat{\mu}_{obs}$. We thus conclude that the machine is indeed adequately calibrated.

PROBLEMS

1. In the cup filling example, suppose we had observed a sample mean of $\hat{\mu}_{obs} = 260$ grams. Would you then conclude that the machine is adequately calibrated or not? You need to redo the hypothesis test to answer this question. A p-value smaller than 0.05 suggests that the machine is not adequately calibrated. Explain in words why that is.

2. What range of values of $\hat{\mu}_{obs}$ would give rise to a p-value larger than 0.05? (i.e., what is the range of sample mean values for which we would conclude that the machine is adequately calibrated.). Note that because of the way we interpret the p-value, it only makes sense to consider $\hat{\mu}_{obs} \geq \mu = 250$.

3. It is common to look at the absolute value of the deviation from the true mean, $|\mu - \hat{\mu}_{obs}|$. In the example above this implies that we would not distinguish between $\hat{\mu}_{obs} = 250.2$ grams and $\hat{\mu}_{obs} = 249.8$, because they would both give rise to a deviation from the mean with an absolute value of 0.2. In this case, the p-value is $\Pr\big(\hat{\mu} \geq (\mu + |\mu - \hat{\mu}_{obs}|)\big) + \Pr\big(\hat{\mu} \leq (\mu - |\mu - \hat{\mu}_{obs}|)\big)$. This is called a two-sided p-value (as opposed to the above definition, which is a one-sided p-value). Calculate the two-sided p-value when $\hat{\mu}_{obs} = 250.2$.

4. Using the two-sided p-value, as defined in Problem 3, what range of values of $\hat{\mu}_{obs}$ would give rise to a p-value larger than 0.05?

2 TERMINOLOGY

2.1 POPULATION AND SAMPLE

Statistics is the study of the collection, analysis, interpretation, presentation, and organization of data. In applying statistics to, e.g., a scientific, industrial, or societal problem, it is necessary to begin with a *population* or process to be studied. Populations can be diverse topics such as "all persons living in a country" or "every atom composing a crystal". In the cup filling example of the previous chapter, the population is "all cups filled by the machine".

Definition 1 – Population

In statistics, a population is a complete set of items that share at least one property in common that is the subject of a statistical analysis.

Statistics deals with all aspects of data including the planning of data collection in terms of the design of surveys and experiments. Statisticians collect data by developing specific experiment designs and survey samples. Representative sampling assures that inferences and conclusions can safely extend from the sample to the population as a whole.

Definition 2 – Sample

A statistical sample is a subset drawn from the population to represent the population in a statistical analysis. If a sample is chosen properly, characteristics of the entire population that the sample is drawn from can be inferred from corresponding characteristics of the sample.

We will denote the sample $X_1, X_2, \dots, X_n$, where the samples are i.i.d. random variables with a given probability distribution, $f_X(x;\theta)$, explained below.

An *experimental study* involves taking measurements of the system under study, manipulating the system, and then taking additional measurements using the same procedure to determine if the manipulation has modified the values of the measurements. In contrast, an *observational study* does not involve experimental manipulation. Statistical analysis is in general more difficult in observational studies.

Two main statistical methodologies are used in data analysis: *descriptive statistics*, which summarizes data from a sample using indexes such as the mean or standard deviation, and *inferential statistics*, which infers predictions about a larger population than the sample represents. To draw meaningful conclusions about the entire population, inferential statistics is needed. It uses patterns in the sample data to draw inferences about the population represented, accounting for randomness. These inferences may take the form of: answering yes/no questions about the data (hypothesis testing), estimating numerical characteristics of the data (estimation), describing associations within the data (correlation) and modeling relationships within the data (for example, using linear regression).

Inference can extend to forecasting, prediction and estimation of unobserved values either in or associated with the population being studied; it can include extrapolation and interpolation of time series or spatial data, and can also include data mining.

2.2 STATISTICS AND ESTIMATORS

Consider a set of independent identically distributed (i.i.d.) random variables $(X_1, X_2, \dots, X_n)$ with a given probability distribution. Standard statistical inference and estimation theory defines a random sample (X) as the random column vector of these i.i.d. variables. The population being examined is described by a probability density function (PDF), $f_X(x;\theta)$, with parameter θ. In general, θ is a vector. For instance, if $f_X(x;\theta)$ is a normal distribution with mean μ and variance σ^2, then $\theta = [\mu, \sigma^2]$. Often, we are interested in inferring knowledge about one or more unknown parameters of the probability distribution. We often refer to the random sample X and its PDF as a *statistical model*.

Definition 3 – Statistical model

A random sample X and its probability density function (PDF), $f_X(x;\theta)$, where θ is the parameter of the PDF. The parameter, θ, is in general a vector and often unknown.

The purpose of inferential statistics is to infer knowledge about the unknown parameter(s), θ. For this purpose we need a summary of the observed data, called a *statistic*. A statistic is a random variable that is a function of the random sample, X, but not a function of unknown parameters. The probability distribution of the statistic, though, may have unknown parameters.

Definition 4 – Statistic

A statistic is a random variable that is a function of the random sample, X, but not a function of unknown parameters, θ.

A commonly used statistic is the average or sample mean

$$\bar{x} = \frac{1}{n}\sum_{i=1}^{n} X_i \tag{3}$$

Consider now the unknown parameter, θ: an estimator is a statistic used to estimate θ. Commonly used estimators include the sample mean and the unbiased sample variance.

Definition 5 – Estimator

An estimator, $\hat{\theta}(X)$, is a statistic used to estimate the unknown parameter θ of a random sample, X. For notational convenience, we will often write $\hat{\theta}$ instead of $\hat{\theta}(X)$.

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

$$\hat{\mu} = \bar{x} = \frac{1}{n}\sum_{i=1}^{n} X_i \tag{4}$$

which is an estimator of the (true) population mean, μ, of a normally distributed random variable, X.

A random variable that is a function of the random sample and of the unknown parameter, but whose probability distribution does not depend on the unknown parameter is called a pivotal quantity or pivot. Pivotal quantities are commonly used for normalization to allow data from different data sets to be compared. Widely used pivots include the z-score defined in equation (2), the chi square statistic, and Student's t-value.

Between two estimators of a given parameter, the one with lower mean squared error is said to be more efficient (it will become clear later what this means exactly). Furthermore, an estimator is said to be unbiased if it's expected value is equal to the true value of the unknown parameter being estimated.

Definition 6 – Unbiased estimator

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

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

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

This still leaves the question of how to obtain estimators in a given situation and carry out the computation. In these lecture notes, we will use the so-called *maximum likelihood method*.

2.3 NULL HYPOTHESIS AND ALTERNATIVE HYPOTHESIS

Inferential statistics often involves testing a *hypothesis*, which is an expectation about how a particular process or phenomenon works. When a possible correlation or similar relation between phenomena is investigated, such as whether a proposed remedy is effective in treating a disease, the hypothesis that a relation exists cannot be examined the same way one might examine a proposed new law of nature. In such an investigation, if the tested remedy shows no effect in a few cases, these do not necessarily falsify the hypothesis. Instead, statistical tests are used to determine how likely it is that the overall

effect would be observed if the hypothesized relation does not exist. If that likelihood is sufficiently small (e.g., less than 5%), the existence of a relation may be assumed. Otherwise, any observed effect may be due to pure chance.

Two types of hypotheses are usually compared. These are called the *null hypothesis* and the *alternative hypothesis*.

Definition 7 – Null hypothesis (H_0)

The null hypothesis is the hypothesis that states that there is no relation between the phenomena whose relation is under investigation, or at least not of the form given by the alternative hypothesis.

Definition 8 – Alternative hypothesis (H_1)

The alternative hypothesis, as the name suggests, is the alternative to the null hypothesis: it states that there is some kind of relation.

The alternative hypothesis is simply a hypothesis that contradicts the null hypothesis. The alternative hypothesis may take several forms, depending on the nature of the hypothesized relation; in particular, it can be two-sided (for example: there is some effect, in a yet unknown direction) or one-sided (the direction of the hypothesized relation, positive or negative, is fixed in advance).

One cannot "prove" a null hypothesis, one can test how close it is to being true. Therefore, we never say that we *accept* the null hypothesis, but that we *fail to reject it*. Also, rejecting the null hypothesis does not automatically prove the alternative hypothesis.

Hypotheses are often expressed in terms of the unknown parameter, θ of X's distribution. A common example, which we saw in chapter 1, concerns the population mean (μ) of normally distributed data.

Example 1 – Hypothesis for the cup filling machine example

In the example of chapter 1, the null hypothesis (H_0) is that the cup filling machine is adequately calibrated. The alternative hypothesis (H_1) is that the machine is not adequately calibrated.

If the machine is indeed adequately calibrated, the true population mean should be 250 grams. Hence, the null hypothesis is

$$H_0: \mu = 250$$

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

$$H_1: \mu \neq 250$$

Based on our observation ($\hat{\mu}_{obs} = \bar{x} = 250.2$ grams) we are either going to either reject or fail to reject H_0 (see below).

2.4 TYPE 1 AND TYPE 2 ERRORS

Working from a null hypothesis two basic forms of error are recognized:

- Type I errors where the null hypothesis is falsely rejected giving a "false positive".
- Type II errors where the null hypothesis fails to be rejected and an actual difference between populations is missed giving a "false negative".

Example 2 – Type I error (cup filling machine example)

Suppose the cup filling described in chapter 1 is adequately calibrated, and suppose we observe a sample mean of $\hat{\mu}_{obs} = 260$ *grams.* Then we obtain a p-value of 0 (see Problem 1), implying that we incorrectly reject the null hypothesis (H_0) that the cup filling machine is adequately calibrated. This is an example of a type I error.

Example 3 – Type II error (cup filling machine example)

Suppose the cup filling machine is not adequately calibrated, and suppose we observe a sample mean of $\hat{\mu}_{obs} = 250.2$ *grams.* Then, as we have seen, we obtain a p-value of 0.34, implying that we incorrectly fail to reject H_0. This is an example of a type II error.

2.5 SIGNIFICANCE AND STATISTICAL TESTS

Statistics rarely give a simple Yes/No type answer to the question under analysis. Interpretation often comes down to the level of statistical significance applied to the numbers and often refers to the probability of a value accurately rejecting the null hypothesis (this value is referred to as the *p-value*).

Definition 9 – Significance level

The statistical significance level, denoted α, is the low probability of obtaining at least as extreme results given that the null hypothesis is true. We often use a probability of one in twenty ($\alpha = 0.05$) as a convenient cutoff level to reject the null hypothesis.

To determine if a result is statistically significant, a researcher would have to calculate a p-value, which is the probability of observing an effect given that the null hypothesis is true. The p-value is formally defined as the probability, under the assumption of the null hypothesis H_0, of obtaining a result equal to or more extreme than what was actually observed. Depending on how we look at it, the "more extreme than what was actually observed" can either mean $\{X \geq x\}$ (right-tailed event) or $\{X \leq x\}$ (left-tailed event) or two times the "smaller" of $\{X \geq x\}$ and $\{X \leq x\}$ (two-tailed event), where *x* denotes the observed quantity (for instance the sample mean, $\bar{x}$). Thus the p-value is given by

Definition 10 – p-value

The p-value is the smallest significance level that allows the test to reject the null hypothesis. If x denotes the observed quantity, the p-value is given by

$$Pr(X \geq x) \text{ for a right-tailed event,}$$

$$Pr(X \leq x) \text{ for a left-tailed event,}$$

$$2 \cdot min\{Pr(X \geq x), Pr(X \leq x)\} \text{ for a two-tailed event.}$$

The smaller the p-value, the larger the significance because it tells the investigator that the hypothesis under consideration may not adequately explain the observation. The null hypothesis H_0 is rejected if any of these probabilities is less than or equal to the chosen significance level, α. Unlike the p-value, the α level is not derived from any observational data nor does it depend on the underlying hypothesis; the value of α is instead determined based on the consensus of the research community that the investigator is working in. This is logically equivalent to saying that the p-value is the probability, assuming the null hypothesis is true, of observing a result at least as extreme as the test statistic. Therefore the smaller the p-value, the lower the probability of committing type I error. If the α level is 0.05, then the conditional probability of a type I error, given that the null hypothesis is true, is 5%. Then a statistically significant result is one in which the observed p-value is less than 5%, which is formally written as $p < 0.05$.

Figure 4 shows a graphical interpretation of the p-value for a normally distributed random variable. Given the observed data point (red dot), the p-value of a right-tailed event is the probability of observing a value that is larger than the observed value, $\Pr(X \geq x)$. The black curve is the PDF of the random variable, and the p-value corresponds to the green shaded region.

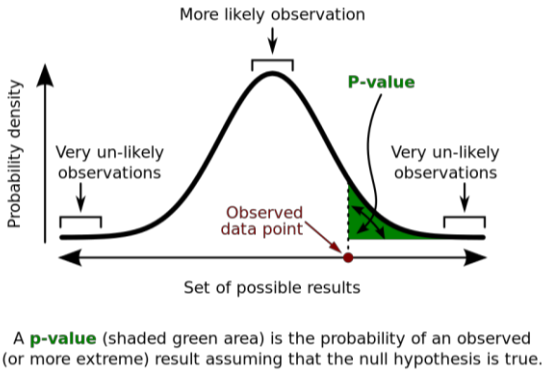


Figure 4 – Illustration of the p-value for a right-tailed event.

Example 4 – Hypothesis test for the cup filling machine example

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

$$H_0: \mu = 250$$

We observe the sample mean $\bar{x} = 250.2$ grams, which also turns to be the maximum likelihood estimator of the population mean. Hence,

$$\hat{\mu} = \bar{x}$$

According to the central limit theorem, the sample mean, $\bar{X}$, is normally distributed with a mean that is equal to the population mean (μ) and a variance that is scaled by $1/n$, where n denotes the number of samples. Hence,

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

where the variance is $\sigma^2/n = 2.5^2/25 = 6.25/25 = 0.25$.

Denoting by $F_{\bar{X}}\left(\bar{x}; (\mu, \sigma^2/n)\right)$ the CDF of the normal distribution of $\bar{X}$, the two-tailed p-value is

$$2 \cdot min\{Pr(\bar{X} \geq \bar{x}), Pr(\bar{X} \leq \bar{x})\}$$

$$= 2 \cdot min\{1 - Pr(\bar{X} \leq \bar{x}), Pr(\bar{X} \leq \bar{x})\}$$

$$= 2 \cdot min\left\{1 - F_{\bar{X}}\left(\bar{x}; \left(\mu, \frac{\sigma^2}{n}\right)\right), F_{\bar{X}}\left(\bar{x}; \left(\mu, \frac{\sigma^2}{n}\right)\right)\right\}$$

$$= 2 \cdot min\left\{1 - F_{\bar{X}}\left(250.2; (250, 0.25)\right), F_{\bar{X}}\left(250.2; (250., 0.25)\right)\right\}$$

$$= 2 \cdot min\{1 - 0.6554, 0.6554\}$$

$$= 2 \cdot min\{0.3446, 0.6554\} = 0.6892$$

where I have used the Matlab command `normcdf(250.2,250,sqrt(0.25))` to calculate $F_{\bar{X}}(250.2; (250,0.25))$.

Since the p-value is not smaller than the significance level, $\alpha = 0.05$, we fail to reject H_0 and conclude that the machine is adequately calibrated.

The distribution of the statistic used in the above example ($\bar{X}$) depends on the parameters of the distribution governing the random variable *X*, from which we have observed the samples. Hence, the sample mean ($\bar{X}$) is not a pivotal quantity – it is instead referred to as a *descriptive statistic*. A hypothesis test is typically specified in terms of a *test statistic*, which is a pivotal quantity. In general, a test statistic is selected or defined in such a way as to quantify, within observed data, behaviors that would distinguish the null hypothesis from the alternative hypothesis, where such an alternative is prescribed, or that would characterize the null hypothesis if there is no explicitly stated alternative hypothesis.

An important property of a test statistic is that its sampling distribution under the null hypothesis must be calculable, either exactly or approximately, which allows p-values to be calculated. A test statistic shares some of the same qualities of a descriptive statistic, and many statistics can be used as both test statistics and descriptive statistics. However, a test statistic is specifically intended for use in statistical testing, whereas the main quality of a descriptive statistic is that it is easily interpretable. Some informative descriptive statistics, such as the sample range, do not make good test statistics since it is difficult to determine their sampling distribution.

Definition 11 – Test statistic

A random variable that summarizes a data-set by reducing the data to one value that can be used to perform the hypothesis test. The probability distribution of a test statistic, as opposed to descriptive statistics, does not depend on the unknown parameter (θ).

One commonly used test statistic is the z-score. For a normally distributed random variable, *X*, with mean μ and variance σ^2 (i.e., $X \sim N(\mu, \sigma^2)$), the z-score is

$$z = \frac{x - \mu}{\sigma} \sim N(0,1) \qquad \textbf{(5)}$$

where *x* is an observation from *X*'s distribution. Note that the z-score is standard normally distributed (i.e., with mean 0 and variance 1).

An important special case of the z-score concerns the sample mean, $\bar{x}$, as defined in equation (3). If the samples $(X_1, X_2, \dots, X_n)$ being averaged are independent and identically distributed (i.i.d.) random variables with mean μ and variance σ^2, then according to the central limit theorem, we have

$$z = \frac{\bar{x} - \mu}{\sigma/\sqrt{n}} \sim N(0,1) \tag{6}$$

Note that the samples being averaged are not required to be normally distributed; they can have any distribution, as long as the samples are i.i.d..

In the cup filling machine example, we can use the z-score as test statistic.

Example 5 – Test statistic for the cup filling machine example

As we saw in Example 4, we could in principle use the sample mean statistic ($\bar{x}$) in order to test the null hypothesis. However, it is more common and more convenient to use the z-score instead. The z-score in this example is as defined in equation (6),

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

where $\mu = 250$ is the population mean, $\sigma = 2.5$ is the population standard deviation, and $n = 25$ denotes the number of samples.

Inserting into the equation above, we get the following z-score:

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

If the null hypothesis holds, the z-score is guaranteed to be standard normally distributed. Now, in order to calculate the p-value, we simply need to calculate

$$2 \cdot min\{Pr(Z \geq z), Pr(Z \leq z)\}$$

where the normalized random variable, Z, is standard normally distributed, $Z \sim N(0,1)$. The CDF of a standard normal distribution is denoted Φ. Hence the p-value is

$$2 \cdot min\{Pr(Z \geq z), Pr(Z \leq z)\}$$

$$= 2 \cdot min\{2 - Pr(Z \leq z), Pr(Z \leq z)\}$$

$$= 2 \cdot min\{1 - \Phi(z), \Phi(z)\}$$

$$= 2 \cdot min\{1 - \Phi(0.4), \Phi(0.4)\}$$

$$= 2 \cdot min\{1 - 0.6554, 0.6554\} = 0.6892$$

2.7 CONFIDENCE INTERVALS (ESTIMATION)

A difference that is highly statistically significant can still be of no practical significance, but it is possible to properly formulate tests in account for this. One response involves going beyond reporting only the significance level to include the p-value when reporting whether a hypothesis is rejected or accepted. The p-value, however, does not indicate the size or importance of the observed effect and can also seem to exaggerate the importance of minor differences in large studies. A better and increasingly common approach is to report *confidence intervals*. Although these are produced from the same calculations as those of hypothesis tests or p-values, they describe both the size of the effect and the uncertainty surrounding it.

In the cup filling machine example, we found out that observing a sample mean of 250.2 grams results in a p-value that is larger than our significance level of 0.05. For sure, a sample mean of 250 grams also yield a p-value larger than 0.05. So there are many possible values of the sample mean that would be consistent with the null hypothesis, and hence good estimates of the population mean. Confidence intervals consist of a range of values (interval) that act as good estimates of the unknown population parameter. It is an observed interval (i.e. it is calculated from the observations), in principle different from sample to sample, that frequently includes the parameter of interest if the experiment is repeated. How frequently the observed interval contains the parameter is determined by the *confidence level*. More specifically, the meaning of the term "confidence level" is that, if confidence intervals are constructed across many separate data analyses of repeated (and possibly different) experiments, the proportion of such intervals that contain the true value of the parameter will match the confidence level; this is guaranteed by the reasoning underlying the construction of confidence intervals.

If a corresponding hypothesis test is performed, the confidence level is the complement of the level of significance, i.e., a 95% confidence interval reflects a significance level of 0.05. The confidence interval contains the parameter values that, when tested, should not be rejected with the same sample.

Definition 12 – Confidence level

The confidence level is the complement of the significance level (α):

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

The definition of confidence intervals is best understood from an example. So let us have another look at the cup filling machine example.

Example 6 – Confidence interval for the cup filling machine example

Recall that the observed test statistic was

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

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

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

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

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

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

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

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

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

In other words, the lower endpoint of the 95% confidence interval is:

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

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

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

With the values in the above example, the confidence interval is:

$$0.95 = \Pr\left(250.2 - 1.96 \cdot \frac{2.5}{\sqrt{25}} \leq \mu \leq 250.2 + 1.96 \cdot \frac{2.5}{\sqrt{25}}\right)$$

$$= \Pr(250.2 - 1.96 \cdot 0.5 \leq \mu \leq 250.2 + 1.96 \cdot 0.5)$$

$$= \Pr(250.2 - 0.98 \leq \mu \leq 250.2 + 0.98) = \Pr(249.22 \leq \mu \leq 251.18)$$

In other words, the 95% confidence interval is between the lower endpoint 249.22 grams and the upper endpoint 251.18 grams. As the desired value 250 of μ is within the resulted confidence interval, there is no reason to believe the machine is wrongly calibrated. This does not mean there is 0.95 probability that the true value of parameter μ is in the interval obtained by using the currently computed value of the sample mean,

$$[249.22;\ 251.18]$$

Instead, every time the measurements are repeated, there will be another value for the sample mean $\bar{x}$. In 95% of the cases μ will be between the endpoints calculated from this mean, but in 5% of the cases it will not be. This is illustrated in Figure 5.

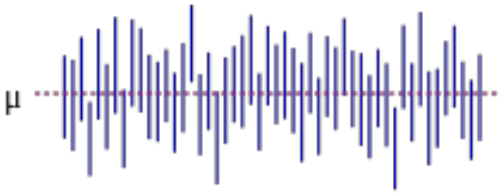


Figure 5 – The blue vertical line segments represent 50 realizations of a confidence interval for the population mean μ, represented as a red horizontal dashed line; note that some confidence intervals do not contain the population mean, as expected.

We are now ready to give or more formal definition of a confidence interval.

Definition 13 – Confidence interval

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

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

2.8 Relation between p-value and Confidence interval

Suppose the observed data has a probability distribution with unknown parameter, θ. Let $pval(x; \theta_0)$ denote the p-value obtained with the null hypothesis H_0: $\theta = \theta_0$. If $pval(x; \theta_0) > \alpha$, then the observed data do not contradict the null hypothesis. Typically, there will be many choices of θ_0, which satisfy $pval(x; \theta_0) > \alpha$. Now, define the set of all such parameters: $\{\theta | pval(x; \theta) > \alpha\}$. This set is an interval of parameters, all of which satisfy the null hypothesis. This is exactly the $1 - \alpha$ confidence interval.

The practical implication of this observation is that it is enough to calculate the 95% confidence interval in order to do a hypothesis test. If the hypothesized parameter, θ_0, is included in the 95% confidence interval

$$Pr(\theta_- \leq \theta_0 \leq \theta_+) = 0.95$$

then θ_0 must satisfy the null hypothesis, H_0: $\theta = \theta_0$.

SUMMARY

In inferential statistics, we take samples from a given population and attempt to infer knowledge about the entire population based on the samples. Usually the knowledge that we wish to infer concerns the probability distribution of the population. To this end, we need to make assumptions about the distribution before we can infer knowledge about its (unknown) parameters. In hypothesis testing, we state an explicit claim about the unknown parameter in the form of a null hypothesis and an alternative hypothesis. We then calculate a test size to obtain a p-value. The p-value is formally defined as the probability, under the assumption of the null hypothesis, of obtaining a result equal to or more extreme than what was actually observed. If the p-value is smaller than a chosen significance level (typically 0.05), we reject the claim stated by the null hypothesis. Otherwise we say that we fail to reject the null hypothesis (or equivalently that we reject the alternative hypothesis). When more detailed knowledge about an unknown parameter is required, we calculate its confidence interval. The confidence interval is always centered around the observed value. The correct interpretation of a 95% confidence interval is: every time the measurements are repeated, there will be another value for the observation. In 95% of the cases the true value of the parameter will be between the endpoints calculated from the observation, but in 5% of the cases it will not be.

PROBLEMS

1. Daily demand for widgets (manufactured devices) is normally distributed with a mean of 100 and a standard deviation of 15.
    a. What is the probability that the demand in a day will exceed 125?
    b. What is the probability that the demand will be less than 75? Less than 70?
    c. How many widgets should be stocked to ensure with 95% probability all demands will be met?

2. Let $X_1, X_2, \ldots, X_{10}$ be a random sample from a normally distributed population with mean 50 and standard deviation 4. Let $\bar{X}$ be the sample mean.
    a. Find the theoretical expected value (mean) and standard deviation of $\bar{X}$.
    b. In Matlab, generate 10 samples from a normally distributed population with mean 50 and standard deviation 4 and calculate the sample mean, $\bar{X}$.
    c. Repeat the experiment 100 times and use the commands `mean` and `std` to estimate the expected value and standard deviation of $\bar{X}$. How do the experimental values compare with the theoretical values?

3. Let $X_1, X_2, \ldots, X_{10}$ be a random sample from a normal population with unknown mean μ and standard deviation 4. Let $\bar{X}$ be the sample mean, and suppose we observe $\bar{x} = 48$.

a. Test H_0: $\mu = 45$ versus H_1: $\mu \neq 45$ at the 5% level.
b. Test H_0: $\mu \leq 45$ versus H_1: $\mu > 45$ at the 5% level.

4. In a random sample of 40 students graduated from Aarhus University in 2010, the average starting salary is DKK 350,000 and the standard deviation is known to be DKK 60,000. How significant is the evidence that the population mean is greater than DKK 300,000? Give a p-value.

5. Let $X_1, X_2, \dots, X_{10}$ be a random sample from a normal population with unknown mean μ and standard deviation 4. Let $\bar{X}$ be the sample mean, and suppose we observe $\bar{x} = 48$. Find a 95% confidence interval for the population mean, μ.

6. In a random sample of 40 students graduated from Aarhus University in 2010, the average starting salary is DKK 350,000 and the standard deviation is known to be DKK 60,000. Find a 95% confidence interval for the population mean, μ.

7. For the cup filling example, calculate the sample mean from 100 simulations of the data (similar to the Matlab code in Figure 3). For each of the 100 sample means, calculate the 95% confidence interval. How many of the intervals include the true population mean ($\mu = 250$)?

3 THE BINOMIAL DISTRIBUTION

3.1 MENDELS PEA PLANT EXPERIMENT

Gregor Johann Mendel (1822-1884) was a German-speaking scientist who gained posthumous fame as the founder of the modern science of genetics. Though farmers had known for centuries that crossbreeding of animals and plants could favor certain desirable traits, Mendel's pea plant experiments conducted between 1856 and 1863 established many of the rules of heredity (the passing of traits to offspring from their parents), now referred to as the laws of Mendelian inheritance.

Mendel worked with seven characteristics of pea plants: plant height, pod shape and color, seed shape and color, and flower position and color. With seed color, he showed that when a yellow pea and a green pea were bred together their offspring plant was always yellow. However, in the next generation of plants, the green peas reappeared at a ratio of 1:3. To explain this phenomenon, Mendel coined the terms “recessive” and “dominant” in reference to certain traits. (In the preceding example, green peas are recessive and yellow peas are dominant.) He published his work in 1866, demonstrating the actions of invisible “factors”—now called genes—in providing for visible traits in predictable ways.

Denote by ‘A’ the dominant (green) allele and denote by ‘a’ the recessive (yellow) allele. Then Mendel’s scientific hypothesis was that the three genotypes ‘AA’, ‘Aa’, ‘aA’ should produce a green pea, whereas the genotype ‘aa’ should produce a yellow pea. To investigate his hypothesis, Mendel performed 580 experiments, where in each experiment two parent plants of genotype ‘Aa’ were crossed over to produce a new offspring. According to his hypothesis, this should result in equally many pea plants of each genotype, and hence looking at the colors of the 580 offspring plants, he should observe

$$\Pr(yellow\ plant) = \frac{1}{4}$$

$$\Pr(green\ plant) = \frac{3}{4}$$

In an idealized experiment, Mendel should observe $580/4 = 145$ yellow plants and 435 green plants. In the actual experiment, he observed 152 yellow plants and 428 green plants. Could this deviation be explained by random variation? Or is Mendel’s hypothesis incorrect?

To answer this question we need statistics. In this case, however, the data are not normally distributed. The correct distribution to use here is the binomial distribution.

3.2 THE BINOMIAL PROBABILITY DENSITY FUNCTION

In probability theory and statistics, the Bernoulli distribution, named after Swiss scientist Jacob Bernoulli, is the probability distribution of a random variable which takes value 1 with success probability p and value 0 with failure probability $q = 1 - p$. It can be used, for example, to represent the toss of a coin, where "1" is defined to mean "heads" and "0" is defined to mean "tails" (or vice versa).

Let B be a Bernoulli distributed random variable with probability parameter p. We write this as

$$B \sim bernoulli(p)$$

This formally means that $B = \{0,1\}$ and

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

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

Now, let $B_1, B_2, \ldots, B_n$ be independent random variables, where

$$B_i \sim bernoulli(p)$$

then the number of successes

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

is a binomially distributed random variable with parameters *n* and *p*. We write this as

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

In words, the binomial distribution with parameters *n* and *p* is the discrete probability distribution of the number of successes in a sequence of *n* independent yes/no experiments, each of which yields success with probability *p*. A Bernoulli experiment or Bernoulli trial is a special case; when n = 1, the binomial distribution is a Bernoulli distribution. The binomial distribution is the basis for the popular binomial test of statistical significance.

The binomial distribution is frequently used to model the number of successes in a sample of size *n* drawn with replacement from a population of size *N*. If the sampling is carried out without replacement, the draws are not independent and so the resulting distribution is a hypergeometric distribution, not a binomial one. However, for *N* much larger than *n*, the binomial distribution is a good approximation, and widely used.

Recall from probability theory that the probability of getting *x* successes in *n* trials is

$$\Pr(X = x) = \frac{n!}{x!\,(n-x)!} p^k (1-p)^{n-x} = \binom{n}{x} p^x (1-p)^{n-\mathrm{x}}$$

Example 7 – Statistical model and hypothesis for Mendel's experiment

Returning to Mendel's experiment, we can denote by *X* the number of yellow plants. Then *X* is a binomially distributed random variable with *n* = 580 and with *p* unknown:

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

The null hypothesis concerns the unknown parameter, *p*, and states that

$$H_0 : p = 1/4$$

---

with the alternative hypothesis

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

The PDFs and CDFs of three different binomial distributions are shown in Figure 6. Figure 7 shows the PDF underlying Mendel's experiment (n=580, p=1/4) along with the observed value, x=152. The corresponding Matlab code is displayed in Figure 8. Just by looking at the plot in Figure 7, it seems plausible that x=152 is a sample from the PDF, because x lies close to the center of the distribution. In other words, it is very likely that due to random variation in X, we will often observe a value that is more extreme than 152. This means that when we are going to calculate the p-value below, we expect that we fail to reject the null hypothesis (i.e., that p=1/4). Before we calculate the p-value, notice that the PDF in Figure 7 is strikingly similar to a Gaussian or normal distribution. This is no coincidence; it follows from the central limit theorem. As we shall see below, we can often – but not always – approximate the PDF of a binomial distribution with a normal distribution.

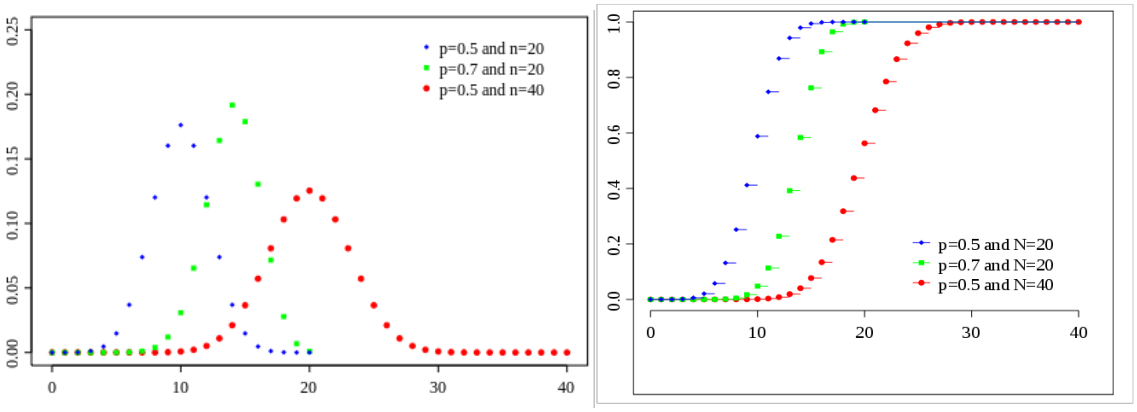


Figure 6 – Examples of PDFs and CDFs of the binomial distribution. Left: PDFs. Right: CDFs.

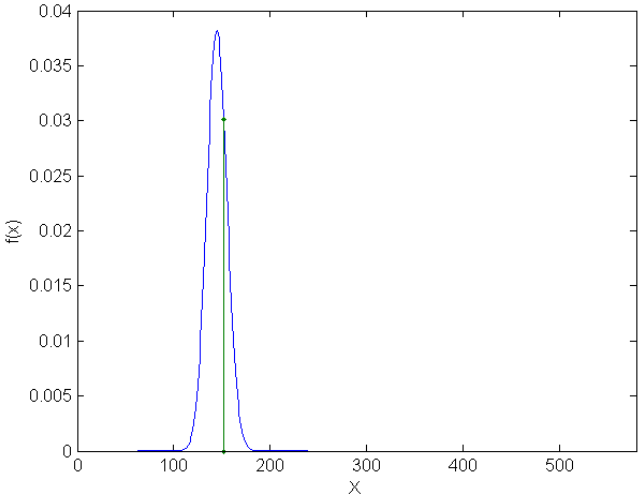


Figure 7 – PDF for the binomial distribution used in Mendel's experiment (n=50, p=1/4). The observed value (x=152) is marked in green.

```
n       = 580;  % Number of trials
p       = 0.25; % Probability of success
x       = 0:n;  % x-values for plotting the PDF
fx      = binopdf(x,n,p); % PDF
xobs    = 152;  % Observed x (=number of successes)
plot(x,fx,...
     [xobs xobs],[0 binopdf(xobs,n,p)],'.-')
axis([0 580 0 0.04])
xlabel('X')
ylabel('f(x)')
```

Figure 8 – Matlab source used to generate the data in Figure 7.

Let us first see how to calculate the p-value exactly (i.e., without using a normal approximation).

Example 8 – Exact p-value for Mendel's experiment

Recall that the statistical model is

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

the null hypothesis is

$$H_0: p = 1/4$$

and the observed number of successes is x = 152. The p-value is the probability of observing a value of the random variable X that is more extreme than 152. By extreme we mean with respect to the value of X that we would observe in an idealized experiment, given that the null hypothesis is true. This value is $p \cdot n = \frac{1}{4} \cdot 580 = 145$. Hence, for a two-tailed test, we need to consider the events $\{X \geq 152\}$ and $\{X \leq 138\}$, both of which deviate by 7 from the theoretical value of 145. The p-value is

$$2 \cdot min\{Pr(X \geq x), Pr(X \leq x)\}$$

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

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

$$= 2 \cdot min\{1 - F_{bino}(152; n = 580, p = 1/4), F_{bino}(138; n = 580, p = 1/4)\}$$

$$= 2 \cdot min\{1 - 0.7652, 0.2682\}$$

$$= 2 \cdot min\{0.2348, 0.2682\} = 0.4697$$

---

where $F_{bino}(x; n, p)$ denotes the CDF of a binomial distribution with parameters *n* and *p.* Since p>0.05, we fail to reject the null hypothesis and the data support Mendel's hypothesis that crossing pea plants of genotype 'Aa' produces equally many pea plants of each genotype.

3.3 WORKING WITH THE BINOMIAL DISTRIBUTION IN MATLAB

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

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

you can use the following Matlab commands:

$\Pr(X = x)$ = `binopdf(x,n,p)`

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

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

3.4 NORMAL APPROXIMATION TO THE BINOMIAL DISTRIBUTION

This section covers the normal approximation of the binomial distribution, ultimately leading to a simpler calculation of p-values and confidence intervals for binomially distributed data.

We are going to need mathematical expressions to calculate the mean and standard deviation of a binomially distributed random variable. First, let us look at a Bernoulli distributed random variable with parameter *p*

$$B \sim bernoulli(p)$$

According to the formulas derived in chapter 2 in Cooper/McGillem, the mean of *B* is

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

and the variance is

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

Now, define the binomially distributed random variable

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

where $B_i \sim bernoulli(p)$, and recall that the $B_i's$ are independent. Then the mean of *X* is

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

where we have used that $E[X+Y] = E[X] + E[Y]$. The variance of X is

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

where we have used that $Var(X+Y) = Var(X) + Var(Y)$ if X and Y are independent.

Now define the standardized random variable

$$Z = \frac{X - E[X]}{\sqrt{Var(X)}} = \frac{X - np}{\sqrt{np(1-p)}} \qquad \textbf{(7)}$$

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

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

This fact follows from the central limit theorem; in the limit as $n \to \infty$, the sum of i.i.d. variables B_i is normal distributed with a mean that is n times similar to the mean of B_i and a variance that is also scaled by n:

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

Standardizing the random variable X as in equation (7), we must have $Z \sim N(0,1)$.

Example 9 – Approximate p-value for Mendel's experiment

Standardizing the observation, x=152, we get

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

and the two-tailed p-value is

$$2 \cdot min\{Pr(Z \geq z), Pr(Z \leq z)\}$$

$$= 2 \cdot min\{1 - Pr(Z \leq 0.6712), Pr(Z \leq 0.6712)\}$$

$$= 2 \cdot min\{1 - \Phi(0.6712), \Phi(0.6712)\}$$

$$= 2 \cdot min\{1 - 0.7490, 0.7490\}$$

$$= 2 \cdot min\{0.2510, 0.7490\} = 0.5021$$

which is slightly larger than the exact p-value calculated in Example 8, but leads to the same result: failure to reject the null hypothesis.

Note that since the normal distribution is symmetric, we can simply calculate the two-tailed p-value as

$$pval = 2 \cdot |1 - \Phi(|z|)| \quad \textbf{(8)}$$

where $|\cdot|$ denotes the numerical value.

3.5 Estimation of the probability parameter in binomially distributed data

In general, the probability parameter *p* is unknown and has to be estimated from observed data. Given the observation *x* = 'number of successes" in *n* trials, the estimator is

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

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

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

The variance of the estimate is

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

Notice that the variance of the estimate decreases with $1/n$. This is a very nice property, because it means that the uncertainty of the estimate decreases as we add more data points.

In statistics, maximum-likelihood estimation (MLE) is a method of estimating the parameters of a statistical model. When applied to a data set and given a statistical model, maximum-likelihood estimation provides estimates for the model's parameters. In general, for a fixed set of data and underlying statistical model, the method of maximum likelihood selects the set of values of the model parameters that maximizes the likelihood function. Intuitively, this maximizes the "agreement" of the selected model with the observed data, and for discrete random variables it indeed maximizes the probability of the observed data under the resulting distribution. Maximum-likelihood estimation gives a unified approach to estimation, which is well-defined in the case of the normal distribution and many other problems.

For the binomial distribution the likelihood function is

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

where $f_{bino}(x|\hat{p})$ denotes the binomial PDF, and *x* is the observed number of successes out of *n* trials. It is important to note that the likelihood function is a function of the parameter estimate, $\hat{p}$, not *x*;

$L(\hat{p})$ is the probability of observing x successes, if the true parameter is $\hat{p}$. The optimum (MLE) estimate of the parameter is the one that maximizes $L(\hat{p})$. Looking at the definition, we see that we can ignore the binomial coefficient $\binom{n}{x}$, because it does not depend on $\hat{p}$. Also, since the logarithm is a monotonic function, the choice of $\hat{p}$ that maximizes $L(\hat{p})$ also maximizes $\log\big(L(\hat{p})\big)$. Accordingly, the optimum $\hat{p}$ maximizes

$$\log(\hat{p}^x(1-\hat{p})^{n-x}) = x \cdot \log(\hat{p}) + (n-x) \cdot \log(1-\hat{p})$$

To find the $\hat{p}$ that maximizes this function, we differentiate and set equal to zero

$$\frac{x}{\hat{p}} - \frac{n-x}{1-\hat{p}} = 0$$

Isolating $\hat{p}$, we get

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

Hence, the parameter estimate that we used above is exactly the maximum-likelihood estimate. In general we prefer estimators that are unbiased, who's variance decrease with the number of samples (n), and that maximize the likelihood function.

3.6 Approximate confidence interval for binomially distributed data

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

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

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

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

where

$$Z = \frac{X - np}{\sqrt{np(1-p)}} \sim N(0,1)$$

is the standardized random variable defined in equation (7). The choice of limits -1.96 and 1.96 is explained in Example 6. Inserting, we get

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

Isolating p in the inequality, one obtains

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

or

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

where

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

and

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

Example 10 – Approximate 95% confidence interval for Mendel's experiment

The parameter estimate is

$$\hat{p}=\frac{x}{n}=\frac{152}{580}=0.2621$$

Since $np=145>5$ and $n(1-p)=435>5$, we can use the normal approximation of the confidence interval:

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

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

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

$$= 0.2993$$

Note that since the hypothesized parameter (p=1/4) lies within the 95% confidence interval, we accept the null hypothesis.

In general, the limits of the $1-\alpha$ confidence interval for p are

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

and

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

where

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

We have to divide α by 2 for a two-tailed test/confidence interval. Also recall that in Matlab code

```
u = norminv(1-alpha/2)
```

3.7 Applications of the binomial distribution

The importance of the binomial distribution is that it has very wide application. This is because at its heart is a binary situation: one with two possible outcomes. Many random phenomena worth studying have two outcomes. Most notably, this occurs when we examine a sample from a large population of 'units' for the presence of a characteristic; each unit either has the characteristic or it doesn't. The generic term 'unit' is used precisely because the situation is so general. The population is often people, in which case a unit is a person; but a unit might be a school, an insect, a bank loan, a company, a DNA sequence, or any of a number of other possibilities.

For the practical part of this course, we will only consider two-tailed hypothesis test/confidence intervals. Also, you will only be presented with problems/assignments, where the normal approximation to the binomial distribution holds. Thus, all you need to perform statistical analysis using the binomial distribution is listed in the test catalog below.

Example 11 – Girl and boy births

In a given region of Denmark, 231 children were born in 2005, of which 108 were girls and 123 were boys. We wish to find out whether girl and boy births are equally likely.

Statistical model:

x = number of girls born = 108

$X \sim binomial(n, p)$, where n = 231

Hypothesis test:

$$H_0: p = 1/2$$

$$H_1: p \neq 1/2$$

Test size:

$$z = \frac{x - np_0}{\sqrt{np_0(1-p_0)}} = \frac{108 - 231 \cdot 1/2}{\sqrt{231 \cdot 1/2 \cdot (1-1/2}} = \frac{108 - 231 \cdot 1/2}{\sqrt{231 \cdot 1/2 \cdot (1-1/2)}} = -0.9869 \sim N(0,1)$$

Approximate p-value: $2 \cdot |1 - \Phi(|z|)| = 2 \cdot |1 - \Phi(0.9869)| = 0.3237$

Since p>0.05, we fail to reject the null hypothesis and conclude that the data suggest that girl and boy births are equally likely.

95% confidence interval:

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

$$= \frac{1}{231 + 1.96^2}\left[108 + \frac{1.96^2}{2} - 1.96\sqrt{\frac{108(231-108)}{231} + \frac{1.96^2}{4}}\right]$$

$$= 0.4042$$

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

$$= \frac{1}{231 + 1.96^2}\left[108 + \frac{1.96^2}{2} + 1.96\sqrt{\frac{108(231-108)}{231} + \frac{1.96^2}{4}}\right]$$

$$= 0.5319$$

SUMMARY

We have now covered the binomial distribution, which can be used to infer knowledge about data with two possible outcomes. All assignments in this course make use of the normal approximation to the binomial distribution, meaning that all you have to know in order to calculate p-values and confidence intervals is the test catalog below.

PROBLEMS

1. A company believes that roughly 35% of consumers rate its brand first in quality. In a survey, 100 consumers are asked to rate the company's brand 'first quality' or 'not first quality'. According to the company's believe,
    a. What is the probability that more than 40 customers will rate the brand first quality?
    b. What is the probability that less than 30 customers will rate the brand first quality?
    c. What is the probability that exactly 45 customers will rate the brand first quality?

2. Let $X \sim binomial(n = 10, p)$.
    a. Find the mean and standard deviation of X for p = 0.1, 0.5, and 0.9.
    b. In Matlab, plot the PDF of X for p = 0.1, 0.5, and 0.9.
    c. Can you figure out a way to generate random samples from $X \sim binomial(10, p)$ in Matlab?

3. A company believes that roughly 35% of consumers rate its brand first in quality. In a survey, 30 out of 100 consumers rate the company's brand first quality.
    a. Write down a statistical model describing the survey.
    b. Write down a null hypothesis and an alternative hypothesis for the company's believe.
    c. Test your null hypothesis using the exact p-value (see **Example 8 – Exact p-value for Mendel's experiment**).
    d. Test your null hypothesis using the normal approximation to the p-value (see **Example 9 – Approximate p-value for Mendel's experiment**).

4. In a survey, x=15 out of n=50 consumers rate the company's brand first quality. The statistical model is $x \sim binomial(50, p)$.
    a. Find the maximum-likelihood estimate of the parameter, p.
    b. Find the 95% confidence interval of the parameter, p.

TEST CATALOG FOR THE BINOMIAL DISTRIBUTION

Statistical model:

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

Hypothesis test (two-tailed):

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

95% confidence interval:

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

4 THE POISSON DISTRIBUTION

In probability theory and statistics, the Poisson distribution is a discrete probability distribution that expresses the probability of a given number of events occurring in a fixed interval of time if these events occur with a known average rate and independently of the time since the last event. The Poisson distribution can also be used for the number of events in other specified intervals such as distance, area or volume.

For instance, an individual keeping track of the amount of mail they receive each day may notice that they receive an average number of 4 letters per day. As it is reasonable to assume that receiving one piece of mail will not affect the arrival times of future pieces of mail – that pieces of mail from a wide range of sources arrive independently of one another – the number of pieces of mail received per day would obey a Poisson distribution. Other examples might include: the number of phone calls received by a call center per hour, the number of decay events per second from a radioactive source, or the number of taxis passing a particular street corner per hour.

4.1 THE POISSON PROBABILITY DENSITY FUNCTION

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

Denoting by γ the known average rate of the arrivals/event, we have

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

That is, the probability of observing an event in the i'th interval is proportional to the length (Δt) of the interval. Furthermore we assume that the observations $B_1, B_2, \dots, B_N$ are independent. Then, the probability of observing $X = x$ events over the entire period of duration $t = N \cdot \Delta t$ is binomially distributed:

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

Now, observe that

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

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

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

This is the PDF of the Poisson distribution. The PDFs and CDFs of three different binomial distributions are shown in Figure 9. Just like we observed for the binomial distribution, as we move towards higher values of the parameter, $\lambda = \gamma/t$, the PDF of a Poisson distributions converges towards a Gaussian.

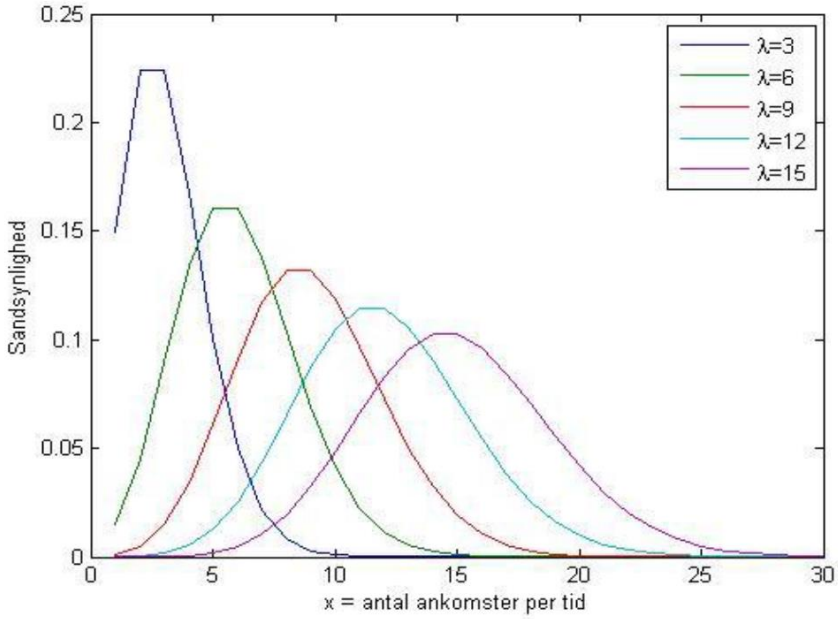


Figure 9 – Examples of PDFs of the Poisson distribution.

The parameter of a Poisson can be expressed as $\gamma = t \cdot \lambda$, where $\lambda = \gamma/t$ should be interpreted as the average rate of arrivals/events.

4.2 Working with the Poisson distribution in Matlab

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

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

you can use the following Matlab commands:

$\Pr(X = x)$ = `poisspdf(x,lambda)`

$\Pr(X \leq x)$ = `poisscdf(x,lambda)`

x must be an integer value and can in general be a vector or array. Note that Matlab expects the parameter $\lambda = \gamma/t$ as input.

4.3 Normal approximation to the Poisson distribution

We shall skip the exact p-value calculation for the Poisson distribution and instead focus on the normal approximation. As for the binomial distribution, we are going to need mathematical expressions to calculate the mean and standard deviation of a Poisson distributed random variable.

Defining the Poisson distributed random variable

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

it can be shown that

$$E[X] = \gamma = t \cdot \lambda$$

and

$$Var(X) = \gamma = t \cdot \lambda$$

Now define the standardized random variable

$$Z = \frac{X - E[X]}{\sqrt{Var(X)}} = \frac{X - \gamma}{\sqrt{\gamma}} = \frac{X - t \cdot \lambda}{\sqrt{t \cdot \lambda}} \tag{9}$$

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

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

This fact follows from the central limit theorem.

Example 12 – Approximate p-value

A store claims that they have an average rate of $\lambda = \gamma/t = 150$ customers per hour. Suppose we observe $x = 280$ customers over a period of two hours.

The store's claim can be formulated using the following null hypothesis:

$$H_0: \gamma = \lambda = 150$$

because $t = 1 \rightarrow \gamma = \lambda$ when we state the hypothesis. The observed quantity (x = 280 customers in $t = 2$ hours) corresponds to an estimated average rate of $\hat{\lambda} = x/t = 280/2 = 140$ customers per hour.

Standardizing the observation, we get the test size

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

According to equation equation (9), the two-tailed p-value is

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

and we conclude that we fail to reject the null hypothesis. Hence, the observation x = 28o customers in two hours does not contract the store's claim that they have an average 150 customers per hour.

4.4 Estimation of the average rate parameter in Poisson distributed data

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

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

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

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

4.5 Approximate confidence interval for Poisson distributed data

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

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

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

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

where

$$Z = \frac{X - t \cdot \lambda}{\sqrt{t \cdot \lambda}} \sim N(0,1)$$

is the standardized random variable defined in equation (9). The observation time, t, is assumed known. Inserting, we get

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

Isolating γ in the inequality, one obtains

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

or

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

where

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

and

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

Example 13 – Approximate 95% confidence interval for the store example

The parameter estimate is

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

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

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

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

Note that since the hypothesized parameter (λ =150) lies within the 95% confidence interval, we accept the null hypothesis.

In general, the limits of the $1 - \alpha$ confidence interval for λ are

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

and

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

Where we recall that $u = \Phi^{-1}\left(1 - \frac{\alpha}{2}\right)$.

4.6 APPLICATIONS OF THE POISSON DISTRIBUTION

Here is another example to demonstrate the use of the Poisson distribution.

Example 14 – Geiger–Marsden experiment

The Geiger–Marsden experiments (also called the Rutherford gold foil experiment) were a landmark series of experiments by which scientists discovered that every atom contains a nucleus where its positive charge and most of its mass is concentrated. They deduced this by measuring how an alpha particle beam is scattered when it strikes a thin metal foil. The experiments were performed between 1908 and 1913 by Hans Geiger and Ernest Marsden under the direction of Ernest Rutherford at the Physical Laboratories of the University of Manchester.

In one of their experiments, Geiger and Marsden detected $x = 11571$ alpha particles (using a Geiger counter) over a time period of $t = 187776$ seconds.

For illustration purposes only, let us assume that we hypothesize,$\lambda = 0.060$.

Statistical model:

x = number of alpha particles detected = 11571

$X \sim poisson(t \cdot \lambda)$, where t = 187776 seconds

Hypothesis test:

$$H_0: \lambda = 0.060$$

$$H_1: \lambda \neq 0.060$$

Test size:

$$z = \frac{x - t \cdot \lambda}{\sqrt{t \cdot \lambda}} = \frac{11571 - 187776 \cdot 0.06}{\sqrt{187776 \cdot 0.06}} = 2.8682 \sim N(0,1)$$

Approximate p-value:

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

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

95% confidence interval:

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

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

and

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

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

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

SUMMARY

In this chapter we have covered the Poisson distribution, which can be used to infer knowledge about processes where arrivals/events occur randomly over time. All assignments in this course make use of the normal approximation to the Poisson distribution, meaning that all you have to know in order to calculate p-values and confidence intervals is the test catalog below.

PROBLEMS

1. For a given traffic light in a city, the local traffic management team claims that 10 cars pass the traffic light per hour. Over a period of 4 hours, the management team counts the number of cars passing the traffic light.
    a. What is the probability that more than 45 cars will pass the traffic light?
    b. What is the probability that less than 38 cars will pass the traffic light?
    c. What is the probability that exactly 45 cars will pass the traffic light?

2. Let $X \sim poisson(\lambda)$.
    a. Find the mean and standard deviation of X for λ = 2, 5, and 10.
    b. In Matlab, plot the PDF of X for λ = 2, 5, and 10.
    c. Can you figure out a way to generate random samples from $X \sim poisson(\lambda)$ in Matlab?

3. For a given traffic light in a city, the local traffic management team claims that 10 cars pass the traffic light per hour. Over a period of 4 hours, the management team counts 45 cars passing the traffic light.
    a. Write down a statistical model describing the experiment.
    b. Write down a null hypothesis and an alternative hypothesis for the management team's claim.
    c. Test your null hypothesis using the normal approximation to the p-value.

4. Over a period of 2 hours, the management team counts 25 cars passing the traffic light. The statistical model is $x \sim poisson(\lambda)$.
    a. Find the maximum-likelihood estimate of the parameter, λ.
    b. Find the 95% confidence interval of the parameter, λ.

TEST CATALOG FOR THE POISSON DISTRIBUTION

Statistical model:

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

Hypothesis test (two-tailed):

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

95% confidence interval:

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

5 NORMALLY DISTRIBUTED DATA

In probability theory, the normal (or Gaussian) distribution is a very commonly occurring continuous probability distribution – a function that tells the probability that any real observation will fall between any two real limits or real numbers, as the curve approaches zero on either side. Normal distributions are extremely important in statistics and are often used in the natural and social sciences for real-valued random variables whose distributions are not known.

The normal distribution is immensely useful because of the central limit theorem, which states that, under mild conditions, the mean of many random variables independently drawn from the same distribution is distributed approximately normally, irrespective of the form of the original distribution: physical quantities that are expected to be the sum of many independent processes (such as measurement errors) often have a distribution very close to the normal. Moreover, many results and methods (such as propagation of uncertainty and least squares parameter fitting) can be derived analytically in explicit form when the relevant variables are normally distributed.

5.1 THE NORMAL PROBABILITY DENSITY FUNCTION

Let X be a normally distributed random variable with mean μ and variance σ^2, which we write

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

Then the PDF is given by

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

As we learned earlier in the course, there is no closed expression for the CDF of a normal distribution. In order to calculate the probability $\Pr(X \leq x)$, we first standardize x:

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

Then

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

where $\Phi(z)$ denotes the CDF of a standard normally distributed random variable, Z. Also note that since the normal PDF is symmetric, we have

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

implying that

$$\Phi(-z) = 1 - \Phi(z)$$

Another important observation that follows from the symmetry of the PDF is that

$$\Pr(Z \leq 0) = \Phi(0) = 1/2$$

Example 15 – Simplified calculation of the p-value from a z-statistic

Suppose we wish to calculate the two-tailed p-value based on a test statistic, $z \sim N(0,1)$. Then we have claimed earlier that

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

Let us now verify this claim. First note that according to the complement rule, we have

$$pval = 2 \cdot min\{Pr(Z \geq z), Pr(Z \leq z)\} = 2 \cdot min\{1 - Pr(Z \leq z), Pr(Z \leq z)\}$$

If z is negative, we get

$$pval = 2 \cdot min\{1 - \Phi(z), \Phi(z)\} = 2 \cdot min\{\Phi(|z|), 1 - \Phi(|z|)\}$$

where the last equality follows from the symmetry of the PDF. Now, since $|z|$ is zero or positive, we must have $\Phi(|z|) \geq 1/2$. Accordingly, the minimum of $\Phi(|z|)$ and $1 - \Phi(|z|)$ is the latter, and we conclude that if z is negative, then

$$pval = 2 \cdot \big(1 - \Phi(|z|)\big)$$

If z is positive

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

using the same argument as above. This concludes the verification of our claim.

5.2 Working with the normal distribution in Matlab

To calculate probabilities of a normally distributed random variable

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

you can use the following Matlab commands:

$f(\mathrm{x})$ = `normpdf(x,mu,sigma)`

$\Pr(X \leq x)$ = `normcdf(x,mu,sigma)`

where `mu` is the mean (μ) and `sigma` the standard deviation ($\sqrt{\sigma^2} = \sigma$). If X is standard normally distributed (μ=1, σ=1), you can skip the arguments `mu` and `sigma`.

5.3 ESTIMATING THE MEAN OF NORMALLY DISTRIBUTED DATA

Given a sample $(X_1, X_2, \dots, X_n)$ of i.i.d. and normally distributed random variables with mean μ and variance σ^2, the maximum-likelihood estimator of the mean $(\hat{\mu})$ is the sample mean $(\bar{x})$

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

This is an unbiased estimator, because

$$E[\hat{\mu}] = E\left[\frac{1}{n}\sum_{i=1}^{n} X_i\right] = \frac{1}{n}\cdot E\left[\sum_{i=1}^{n} X_i\right] = \frac{1}{n}\cdot \sum_{i=1}^{n} E[X_i] = \frac{1}{n}\cdot \sum_{i=1}^{n} \mu = \mu$$

The variance of the estimate decreases with $1/n$

$$Var(\hat{\mu}) = Var\left(\frac{1}{n}\sum_{i=1}^{n} X_i\right) = \frac{1}{n^2}\cdot Var\left(\sum_{i=1}^{n} X_i\right) = \frac{1}{n^2}\cdot \sum_{i=1}^{n} Var(X_i) = \frac{1}{n^2}\cdot \sum_{i=1}^{n} \sigma^2 = \frac{1}{n}\cdot \sigma^2$$

To show that the above estimator is in fact the maximum-likelihood estimator, let $x_1, x_2, \dots, x_n$ be independent observations from a normal distribution with mean μ and variance σ^2. Then, the likelihood function (i.e., the probability of observing $x_1, x_2, \dots, x_n$) given the true variance (σ^2) and some choice of mean $(\hat{\mu})$ is given by

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

where

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

Here, we have used that

$$f(X, Y) = f_X(X)\cdot f_Y(Y)$$

when *X* and *Y* are independent random variables. Inserting, we get

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

The maximum-likelihood estimate is the $\hat{\mu}$ that maximizes this function. Ignoring constants and taking the logarithm, we see that the optimum choice of $\hat{\mu}$ must maximize

$$-\sum_{i=1}^{n} (x_i - \hat{\mu})^2$$

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

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

5.4 THE CENTRAL LIMIT THEOREM

In probability theory, the central limit theorem (CLT) states that, given certain conditions, the arithmetic mean of a sufficiently large number of samples of i.i.d. random variables, each with a well-defined expected value (μ) and well-defined variance (σ^2), will be approximately normally distributed, *regardless of the underlying distribution*. That is, suppose that a sample is obtained containing a large number of observations, each observation being randomly generated in a way that does not depend on the values of the other observations, and that the arithmetic average of the observed values is computed. If this procedure is performed many times, the central limit theorem says that the computed values of the average will be distributed according to the normal distribution.

The central limit theorem has a number of variants. The variant that we are going to use is the following.

Definition 14 – The central limit theorem

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

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

Note that X can have any distribution, i.e., it is *not* required to be normally distributed.

Although the exact number is subject of debate, it is common practice to require that the number of samples (n) should be 30 or larger in order to apply the CLT:

$$n \geq 30$$

We have already seen an example. Using the CLT and the simplified p-value computation stated in Example 15, the hypothesis test for the mean of the population becomes very simple (see Example 16 below). The complete test catalog for testing a mean of samples drawn independently from an arbitrary distribution is given at the end of the chapter – Test catalog for the mean (known variance). Notice that in order to use this test catalog, the true population variance (σ^2) must be given, and *preferably* the number of samples should be 30 or larger.

Example 16 – Hypothesis test for the mean in the cup filling example

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

$$H_0: \mu = 250$$

We observe the sample mean $\bar{x} = 250.2$ grams. According to the CLT, we have

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

We can test the null hypothesis using the z-score

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

The p-value is

$$pval = 2 \cdot \big(1 - \Phi(|z|)\big) = 2 \cdot \big(1 - \Phi(0.4)\big) = 2 \cdot (1 - 0.6554) = 0.6892$$

5.5 SAMPLE SIZE DETERMINATION

Larger sample sizes generally lead to increased precision when estimating unknown parameters. For example, if we wish to know the effect of a medical treatment, we would generally have a more accurate estimate of this effect if we sampled and examined 200 rather than 100 patients. However, if sufficient statistical power can be obtained using just 100 patients, we prefer that instead of 200 patients because it reduces the economic costs of the experiment.

Sample sizes are judged based on the quality of the resulting estimates. For example, if a mean is being estimated, one may wish to have the 95% confidence interval be less than 0.1 units wide. Recall that the 95% confidence interval of the mean estimator ($\hat{\mu}$) is

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

(See Test catalog for the mean (known variance) or Example 6 – Confidence interval for the cup filling machine example.)

The general form of the 95% confidence interval is

$$\bar{x} \pm B$$

Suppose that you – as a scientist or engineer – wish to make an estimate of the mean of a population, where the 95% confidence interval is less than B units wide. Then, assuming that the population standard deviation is known, we require that

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

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

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

Let us see an example of sample size calculation.

Example 17– Sample size calculation

A large manufacturing firm is interested in estimating the average distance traveled to work by its employees. Past studies of this type indicate that the standard deviation of the distances should be in the neighborhood of 2 km. How many employees should be samples if the estimate is to be within 0.1 km of the true average, with 95% confidence?

The resulting interval is to be of the form $\bar{X} \pm 0.1$. Thus, $B = 0.1$. It follows that

$$n \geq \left(\frac{1.96 \cdot \sigma}{B}\right)^2 = \left(\frac{1.96 \cdot 2}{0.1}\right)^2 \approx 1537$$

Thus, at least 1537 employees should be sampled to achieve the desired result.

5.6 STUDENTS T-DISTRIBUTION

In all of the examples we have seen so far, we have assumed that we know the true variance or standard deviation of the population. In practice, however, this is rarely the case. In probability and statistics, Student's t-distribution (or simply the t-distribution) is any member of a family of continuous probability distributions that arise when estimating the mean of a normally distributed population in situations where the sample size is small and population standard deviation is unknown. Whereas a normal distribution describes a full population, t-distributions describe samples drawn from a full population; accordingly, the t-distribution for each sample size is different, and the larger the sample, the more the distribution resembles a normal distribution.

If we take a sample of n observations from a normal distribution, then the t-distribution with $v = n-1$ degrees of freedom can be defined as the distribution of the location of the true mean, relative to the sample mean and divided by the sample standard deviation, after multiplying by the normalizing term $\sqrt{N}$. In this way, the t-distribution can be used to estimate how likely it is that the true mean lies in any given range. The t-distribution is symmetric and bell-shaped, like the normal distribution, but has heavier tails, meaning that it is more prone to producing values that fall far from its mean.

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

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

where $\bar{x}$ is the average of the samples. This estimator is biased, because it can be shown that

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

i.e., the expected value of the variance estimate, $\hat{\sigma}^2$, is *not* identical to the true variance, σ^2. The unbiased estimate of the variance, which we denote s^2, is

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

We refer to this unbiased estimator as the *sample variance* or *empirical variance*. It is unbiased because

$$E[s^2] = \sigma^2$$

Now, consider the usual *z* statistic

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

This statistic contains only one random variable, $\bar{x}$. The equivalent statistic, when replacing the standard deviation (σ) with the empirical standard deviation (s) is called a *t*-score

$$t = \frac{\bar{x} - \mu}{\mathrm{s}/\sqrt{n}} \; or \; t = \frac{\bar{x} - \mu}{\sqrt{s^2/\mathrm{n}}}$$

The *t*-score is a function of two random variables, $\bar{x}$ and *s*. The *t*-score is *not* normally distributed; it is *t* distributed with $v = n - 1$ degrees of freedom, which we write

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

Now, suppose we wish to test the null hypothesis H_0: $\mu = \mu_0$ using the t statistic instead of the z statistic. Then the p-value is given by

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

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

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

where t_0 is chosen such that

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

where $\alpha = 0.05$ is the significance level.

5.7 WORKING WITH THE T-DISTRIBUTION IN MATLAB

To calculate probabilities of a t-distributed random variable

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

you can use the following Matlab commands:

$$f(t) = \texttt{tpdf(t,n-1)}$$

$$\Pr(T \leq t) = \texttt{tcdf(t,n-1)}$$

where `n` is the number of samples. We are going to need to be able to calculate the inverse of the CDF below. That is, given a probability $1 - \alpha/2$, what is the corresponding value t_0, such that $\Pr(T \leq t_0) = 1 - \alpha/2$?

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

5.7 INFERENCE ON THE MEAN FOR A POPULATION WITH UNKNOWN VARIANCE

Let us do inference on the population mean in the cup filling machine using the *t* statistic instead of the *z* statistic. Hence, we are assuming that we do not know the true variance of the population. The example is clearer if we actually have sampled data, so let the 25 sampled weights of the cups be

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

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

```
>> mean(x)

ans =

  249.5313
```

and the (unbiased) estimate of the variance is

```
>> var(x)

ans =

    6.2900
```

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

Example 18 – Hypothesis test and 95% confidence interval for the mean in the cup filling example (unknown variance)

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

$$H_0: \mu = 250$$

We observe a sample mean of $\bar{x} = 249.53$ grams and an empirical standard deviation of $s = 2.51$.

The test size is

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

Then the p-value is

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

and we fail to reject the null hypothesis. The 95% confidence interval for the mean is $\bar{x} \pm t_0 \cdot s/\sqrt{n}$, so the endpoints are

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

and

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

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

Note that the results obtained above are not directly comparable to the results obtained in chapter 2, because the observations are different ($\bar{x}$=249.53 grams vs $\bar{x}$=250.2 grams). For the sake of argument, suppose that we had instead observed $\bar{x}$=250.2 grams and $s = 2.5$ grams, making the results comparable. Then the endpoints of the 95% confidence interval using the *t* statistic would be

$$\mu_- = \bar{x} - t_0 \cdot \frac{s}{\sqrt{n}} = 250.2 - 2.0639 \cdot \frac{2.5}{\sqrt{25}} = 249.17$$

and

$$\mu_+ = \bar{x} + t_0 \cdot \frac{s}{\sqrt{n}} = 250.2 + 2.0639 \cdot \frac{2.5}{\sqrt{25}} = 251.23$$

These values differ slightly from the values obtained with the *z* statistic in **Example 6 – Confidence interval for the cup filling machine example**, where we obtained

$$\mu_- = \bar{x} - 1.96 \cdot \frac{\sigma}{\sqrt{n}} = 250.2 - 1.96 \cdot \frac{2.5}{\sqrt{25}} = 249.22$$

and

$$\mu_+ = \bar{x} + 1.96 \cdot \frac{\sigma}{\sqrt{n}} = 250.2 + 1.96 \cdot \frac{2.5}{\sqrt{25}} = 251.18$$

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

Recalling that t_0 depends on the number of samples (n), we see that the width of the confidence interval obtained with the *t* statistic also depends on the number of samples. The table below shows the value of t_0 (which is proportional to the width of the confidence interval!) for $\alpha = 0.05$ and increasing *n*.

n	2	3	5	10	30	∞
t_0	12.71	4.30	2.78	2.26	2.05	1.96

The values in the table have been obtained using the Matlab command

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

for n = 2, 3, 5, 10, and 30. We observe that t_0 approaches 1.96 as $n \to \infty$. This is because that as $n \to \infty$, the uncertainty in the estimate (s) of the true standard deviation (σ) goes towards zero. So, in the limit $n \to \infty$, the t-distribution converges towards a standard normal distribution. This is illustrated in Figure 10.

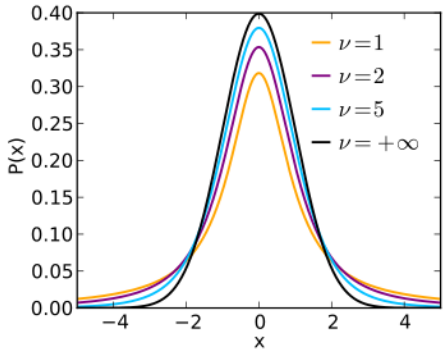


Figure 10 – The PDF of the t-distribution for varying number of degrees of freedom, v = n-1. As n goes towards infinity, the t-distribution converges towards a standard normal distribution.

5.8 CHECKING FOR NORMALITY IN SAMPLED DATA (Q-Q PLOTS)

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

In statistics, a Q–Q plot ("Q" stands for quantile) is a probability plot, which is a graphical method for comparing two probability distributions by plotting their quantiles against each other. The normal probability plot is a special case of the Q–Q probability plot for a normal distribution. This is the only type of Q-Q plot that we will consider in this course. The Q-Q plot is used to identify substantive departures from normality. In a Q-Q plot, the sorted data are plotted vs. values selected to make the resulting image look close to a straight line if the data are approximately normally distributed. Deviations from a straight line suggest departures from normality. The plotting can be performed using Matlab's built-in command, qqplot.

Quantiles are values taken at regular intervals from the inverse of the CDF of a random variable. Dividing ordered data into q essentially equal-sized data subsets is the motivation for q-quantiles; the quantiles are the data values marking the boundaries between consecutive subsets. For 25% quantile given the data on page 51 is

```
q25 = quantile(x,0.25)

q25 =

  248.0731
```

The interpretation of this result is that roughly 25% of the data points should lie below 248.07. Figure 11 shows the sorted data along with the 25% percentile. Six of the sampled values lie below 248.07, which roughly corresponds to 25% of n=25.

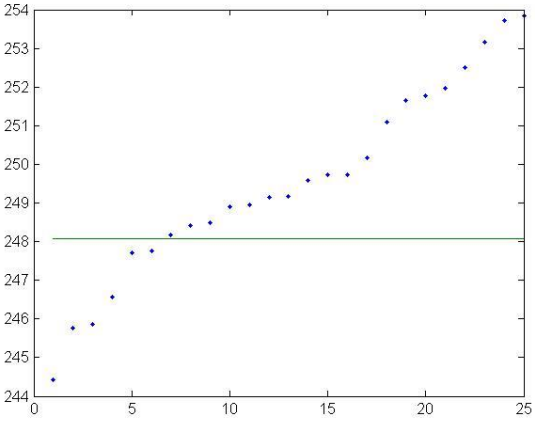


Figure 11 – Sorted data values from page 51 along with the estimated 25% percentile = 248.07. Roughly 25% of the data should lie below this value.

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

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

where $x_{[i]}$ denotes the *i*'th sample after sorting the samples $x_1, x_2, \ldots, x_n$ in ascending order, and the double-arrow means "corresponds to". If the data are consistent with a sample from a normal distribution, then plotting $x_{[i]}$ vs. $\Phi^{-1}\left(\frac{i-0.5}{n}\right)$ should result in a straight line. This is the Q–Q plot. As a reference, a straight line can be fit to the points. The further the points vary from this line, the greater the indication of departure from normality. If the sample has mean 0, standard deviation 1 then a line through 0 with slope 1 could be used. Figure 12 shows the Q-Q plot of the data from page 51. The corresponding Matlab command is

```
qqplot(x)
```

Since the sampled data lie roughly on a straight line, we conclude that the data are normally distributed.

With more points, random deviations from a line will be less pronounced. Normal plots are often used with as few as 7 points, e.g., with plotting the effects in a saturated model from a 2-level fractional factorial experiment. With fewer points, it becomes harder to distinguish between random variability and a substantive deviation from normality.

5.9 Application of the t-distribution

The t-distribution can be used to make inference about the mean of any population with unknown variance. If the sample size is large ($n \geq 30$) it is safe to proceed using the *t*-score to calculate p-values and confidence intervals. If, on the hand, the number of samples is small, you need to check for normally first by making a Q-Q plot. In the theoretical case, where the true population variance is given, you can base your statistical inference on the *z*-score.

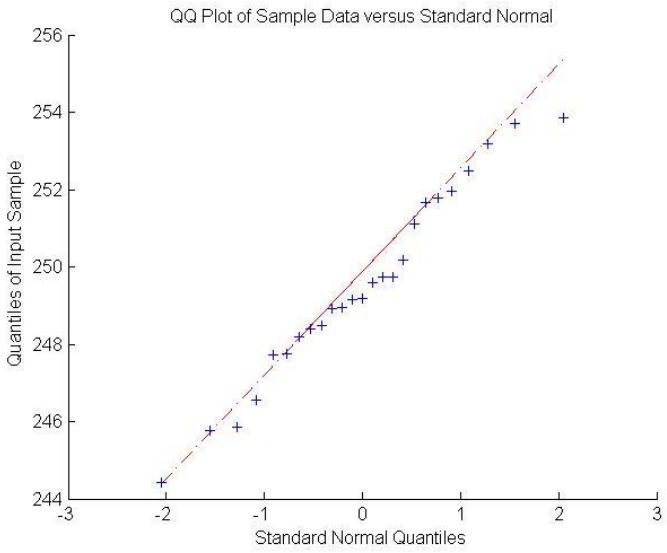


Figure 12 – Q-Q plot of the data from page 51. The data points lie roughly on a straight line, at we conclude that the data are in fact normally distributed.

Example 19

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

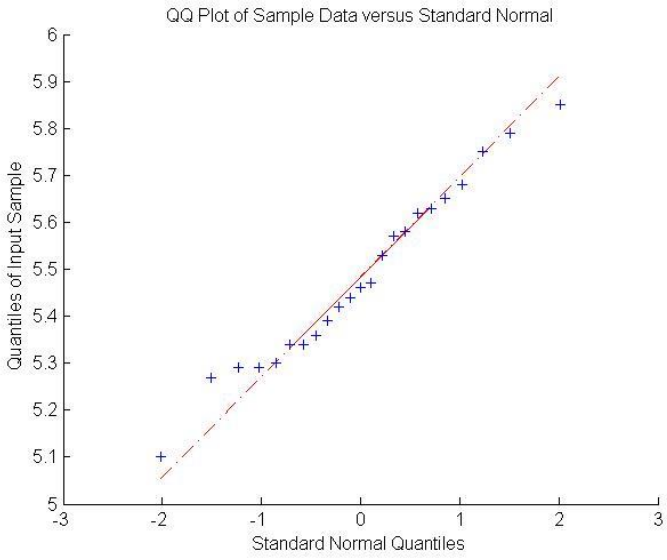
Cavendish's measurements were

```
x = [ 5.36 5.29 5.58 5.65 5.57 5.53 5.62 5.29 ...

5.44 5.34 5.79 5.10 5.27 5.39 5.42 5.47 ...

5.63 5.34 5.46 5.30 5.75 5.68 5.85 ];
```

Check for normality:

```
qqplot(x)
```

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



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

```
>> mean(x)

     5.4835
```

with an empirical variance of

```
var(x)

     0.0363
```

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

$$H_0: \mu = 5.517$$

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

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

Then the p-value is

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

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

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

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

so the endpoints of the confidence interval are

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

and

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

5.10 Where does the t-distribution come from?

Without going into too much detail, let us just have a brief look at where the t-distribution comes from. Let $X_1, X_2, \ldots, X_n$ be i.i.d. samples of a standard normally distributed random variable *X.* Then, the distribution of the sum-of-squares is

$$\sum_{i=1}^{n} X_i^2 \sim \chi^2(n-1)$$

where $\chi^2(n-1)$ is called "chi-square" (in Danish, "Ki-i-anden") distribution with $(n-1)$ degrees of freedom. Now, if $\bar{X} \sim N(\mu, \sigma^2/n)$ is a sample mean drawn from n i.i.d. random variables, then it can be shown that

$$V = \frac{(n-1)s^2}{\sigma^2} \sim \chi^2(n-1)$$

where s^2 is the empirical variance and σ^2 the true population variance. This distribution can be used to estimate a 95% confidence for the variance in normally distributed data (however, this is beyond the scope of this course). Now, let

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

Then it can be shown that the *t* statistic is

$$T = \frac{Z}{\sqrt{V/(n-1)}} = \frac{\bar{X} - \mu}{s/\sqrt{n}}$$

and has probability distribution $\frac{N(0,1)}{\sqrt{\chi^2(n-1)/(n-1)}}$, depending only on n. This is exactly the t-distribution.

The χ^2 distribution is one of the most widely used probability distributions in inferential statistics, e.g., in hypothesis testing or in construction of confidence intervals. The chi-squared distribution is used in the common chi-squared tests for goodness of fit of an observed distribution to a theoretical one, the independence of two criteria of classification of qualitative data, and in confidence interval estimation for a population standard deviation of a normal distribution from a sample standard deviation. Many other statistical tests also use this distribution.

SUMMARY

In this chapter we have learned how to make inference about the mean of a population. If the sample size is sufficiently large, the central limit theorem states that the sample mean is normally distributed, and we can use the *t*-score to make inference about the population mean. If the sample size is small, we are only allowed to make inferences about the population mean if the samples are normally distributed. We use the Q-Q plot to check for normality. In the rare case where the true population variance is given (such as in the cup filling example), we can use the *z*-score instead of the *t*-score. Furthermore, we have seen how to calculate the required sample size if we want the 95% confidence interval to be less than, say, 0.1 units wide.

PROBLEMS

1. Let $X_1, X_2, \ldots, X_{10}$ be a random sample from a normal population with unknown mean and variance. Let $\bar{X}$ be the sample mean, and suppose we observe $\bar{x} = 48$. Also, let S^2 be the empirical variance, and suppose we observe $s^2 = 16$.
    a. Test H_0: $\mu = 50$ versus H_1: $\mu \neq 50$ at the 5% level.
    b. Find a 95% confidence interval for the population mean, μ.

2. In the cup filling example, suppose the true population standard deviation is 2.5 grams. How many cups do we need to sample to make the 95% confidence interval less than 0.5 grams wide?

3. You are given the sample

```
x = [ 54.0748
      56.6827
      54.7552
      44.1039
      50.2046
      53.6727
      63.9488
      50.5385
      50.0734
      52.0398]
```

   from some population.
    a. Verify that the samples are normally distributed.
    b. Give an estimate of the mean and variance of the population.
    c. Find a 95% confidence interval for the population mean, μ.

TEST CATALOG FOR THE MEAN (KNOWN VARIANCE)

Statistical model:

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

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

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

Hypothesis test (two-tailed):

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

95% confidence interval:

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

TEST CATALOG FOR THE MEAN (UNKNOWN VARIANCE)

Statistical model:

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

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

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

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

Hypothesis test (two-tailed):

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

95% confidence interval:

- $\mu_- = \bar{x} - t_0 \cdot s/\sqrt{n}$
- $\mu_+ = \bar{x} + t_0 \cdot s/\sqrt{n}$

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

6 COMPARING TWO POPULATION MEANS

In this chapter, we are going to develop tests to determine if two population means are equal. A common application is to test if a new process or treatment is superior to a current process or treatment.

There are several variations on this test.

The data may either be paired or not paired. By paired, we mean that there is a one-to-one correspondence between the values in the two samples. That is, if $X_1, X_2, \dots, X_n$ and $Y_1, Y_2, \dots, Y_n$ are the two samples, then X_i corresponds to Y_i. For paired samples, the difference $X_i - Y_i$ is usually calculated. For unpaired samples, the sample sizes for the two samples may or may not be equal. The formulas for paired data are somewhat simpler than the formulas for unpaired data.

The variances of the two samples may be assumed to be equal or unequal. Equal variances yields somewhat simpler formulas, and this is the only case that we will consider in this course.

In some applications, you may want to adopt a new process or treatment only if it exceeds the current treatment by some threshold. In this case, we can state the null hypothesis in the form that the difference between the two populations means is equal to some constant $\mu_1 - \mu_2 = d_0$, where the constant is the desired threshold.

6.1 TWO-SAMPLE Z-TEST FOR UNPAIRED DATA (KNOWN VARIANCE)

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

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

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

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

Note that in general, the two samples are of different size (i.e., $n_1 \neq n_2$). The statistical question of interest is whether the two population means, denoted μ_1 and μ_2, are equal. This is the same as asking whether

$$\mu_1 - \mu_2 = 0$$

From the central limit theorem, we know that the estimates of the two population means are

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

and

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

The estimate of the difference between the two population means is

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

where the distribution of $\bar{x}_1 - \bar{x}_2$ follows from the fact that the PDF of the sum of two independent random variables is the convolution of their respective PDFs, and the convolution of two Gaussians is another Gaussian.

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

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

The null hypothesis for testing whether the two population means are equal is

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

with the alternative hypothesis

$$H_1: \mu_1 - \mu_2 \neq 0$$

(This corresponds to a two-tailed test). Under the null hypothesis we must have

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

Since we have implicitly assumed that we know the true variance, we can test the null hypothesis by considering whether the standardized test size is a sample from a standard normal distribution:

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

Example 20 – Hypothesis test for comparing two population means (with known and identical variances)

Suppose we observe sample means $\bar{x}_1 = 3$ and $\bar{x}_2 = 4$ from two normally distributed populations with standard deviation 1. Furthermore, let us assume that we have taken ten samples from population 1 ($n_1 = 10$) and twenty samples from population two ($n_2 = 20$).

We wish to test whether the two samples are consistent with the hypothesis that the two populations have identical means. Hence, the null hypothesis is

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

The z-score is

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

Inserting into the equation of the p-value, we get

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

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

As we have seen in chapters 2-5, it may be desirable to specify a 95% confidence interval for the difference between the two population means. For this purpose we need an estimate of the difference between the two means. Denoting the true difference

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

the estimator is

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

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

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

Now, since

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

The standardization of the random variable, $\hat{\delta} = \bar{x}_1 - \bar{x}_2$, becomes

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

Using the same trick as in the previous chapters, namely

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

We can isolate the true difference between the population means (δ) to get

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

Hence, the endpoints of the 95% confidence interval are

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

and

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

Example 21 – 95% confidence interval for the difference between two population means (with known and identical variances)

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

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

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

and

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

The interpretation of this result is that all choices of δ within this interval would satisfy the null hypothesis

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

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

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

6.2 TWO-SAMPLE T-TEST FOR UNPAIRED DATA (UNKNOWN VARIANCE)

In the case, where the true variance is unknown we have to estimate it. The effect of this is that we must use the t-statistic instead of the z-statistic. If the two population distributions can be assumed to have the same variance – and, therefore, the same standard deviation – the empirical variances of the two populations can be pooled together, each weighted by the number of cases in each sample. The formula for the pooled variance estimator is

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

where

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

and

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

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

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

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

Example 22– Hypothesis test and confidence interval for comparing two means with unknown and identical variances

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

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

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

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

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

With the null hypothesis, $H_0: \delta = \mu_1 - \mu_2 = 0$, the t-score becomes

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

Inserting into the equation of the p-value, we get

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

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

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

$$\delta_- = (\bar{x}_1 - \bar{x}_2) - t_0 \cdot s\sqrt{\frac{1}{n_1} + \frac{1}{n_2}} = (3 - 4) - 2.0484 \cdot 1.0938 \cdot \sqrt{\frac{1}{10} + \frac{1}{20}}$$
$$= -1.8678$$

$$\delta_+ = (\bar{x}_1 - \bar{x}_2) + t_0 \cdot s\sqrt{\frac{1}{n_1} + \frac{1}{n_2}} = (3 - 4) + 2.0484 \cdot 1.0938 \cdot \sqrt{\frac{1}{10} + \frac{1}{20}}$$
$$= -0.1322$$

where `t0 = tinv(0.975,n1+n2-2) = tinv(0.975,30-2) = 2.0484.`

Since $\delta = 0$ is not included in the 95% confidence interval, we reject the null hypothesis stating that the population means are equal.

We will conclude this section with another example.

Example 23 - Does right- or left-handedness affect how fast people type?

Random samples of students from a typing class are given a typing speed test (words per minute), and the results are compared.

Group	-Handed	n	$\bar{x}$	s
1	Right	16	55.8	5.7
2	Left	9	59.3	4.3

Because you are looking for a difference between the groups in either direction (right-handed faster than left, or vice versa), this is a two-tailed test.

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

$$H_1: \mu_1 - \mu_2 \neq 0$$

We first calculate the pooled variance

$$s^2 = \frac{1}{n_1 + n_2 - 2}\left((n_1 - 1)s_1^2 + (n_2 - 1)s_2^2\right)$$
$$= \frac{1}{16 + 9 - 2}\left((16 - 1) \cdot 5.7^2 + (9 - 1) \cdot 4.3^2\right) = 27.62$$

Next, we calculate the t-score

$$t = \frac{(\bar{x}_1 - \bar{x}_2)}{s\sqrt{1/n_1 + 1/n_2}} = \frac{55.8 - 59.3}{\sqrt{27.62} \cdot \sqrt{1/16 + 1/9}} = -1.598 \sim t(n_1 + n_2 - 2)$$

Inserting into the equation of the p-value, we get

$$2 \cdot \left(1 - t_{cdf}(|t|, n_1 + n_2 - 2)\right) = 2 \cdot \left(1 - t_{cdf}(1.598{,}25 - 2)\right)$$
$$= 2 \cdot (1 - 0.9382) = 0.1237$$

Since p>0.05 the null hypothesis of equal population means cannot be rejected. There is no evidence that right- or left -handedness has any effect on typing speed.

6.3 PAIRED DIFFERENCE TEST

Here we will look at a paired t-test, which compares one set of measurements with a second set from the same sample. It is often used to compare "before" and "after" scores in experiments to determine whether significant change has occurred. The data are paired. By paired, we mean that there is a one-to-one correspondence between the values in the two samples. That is, if $X_{11}, X_{12}, \dots, X_{1n}$ and $X_{21}, X_{22}, \dots, X_{2n}$ are the two samples, then X_{1i} corresponds to X_{2i}.

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

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

where δ is the true (unknown) difference between X_1 and X_2. The estimate of δ is the average difference between X_1 and X_2

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

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

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

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

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

Denoting by s_d the corresponding unbiased estimate of the standard deviation, then under the null hypothesis, $H_0: \delta = \delta_0$, the standardized test size is

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

Example 24 – Does fertilizer have an effect on corn growth?

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

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

The null hypothesis is

$$H_0: \delta = 6$$

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

As shown in the calculations below, the average difference in stalk height before and after is

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

$$45.3 - 35.5 = 9.8$$
$$36.0 - 31.7 = 4.3$$
$$38.6 - 31.2 = 7.4$$
$$44.7 - 36.3 = 8.4$$
$$31.4 - 22.8 = 8.6$$
$$33.5 - 28.0 = 5.5$$
$$28.8 - 24.6 = 4.2$$
$$35.8 - 26.1 = 9.7$$
$$42.9 - 34.5 = 8.4$$
$$35.0 - 27.7 = 7.3$$
$$\frac{73.6}{10} = 7.36$$

The estimated variance is

$$s_d^2 = \frac{1}{n-1}\sum_{i=1}^{n}\left(d_i - \bar{d}\right)^2 = \frac{(9.8 - 7.36)^2 + (4.3 - 7.36)^2 + \cdots + (7.3 - 7.36)^2}{10 - 1}$$
$$= 4.216$$

and the estimated standard deviation becomes $s_d = \sqrt{4.216} = 2.05$. The test size is

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

Inserting into the equation of the p-value, we get

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

Since p>0.05 the null hypothesis cannot be rejected. Hence, the test does not provide evidence that the fertilizer caused the corn to grow more or less than if it had not been fertilized.

Note: A one-tailed test would have resulted the opposite conclusion (p<0.05).

Recall that the estimator of δ is

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

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

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

and

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

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

Example 25 – Which sowing machine is the better one?

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

Field	*Machine 1*	*Machine 2*	*Difference*
1	8.0	5.6	2.4
2	8.4	7.4	1.0
3	8.0	7.3	0.7
4	6.4	6.4	0.0
5	8.6	7.5	1.1
6	7.7	6.1	1.6
7	7.7	6.6	1.1
8	5.6	6.0	−0.4
9	5.6	5.5	0.1
10	6.2	5.5	0.7

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

$$H_0: \delta = 0$$

and the alternative hypothesis states that there is a difference

$$H_1: \delta \neq 0$$

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

The data expressed in Matlab code are

```
data = [
    8.0 5.6
    8.4 7.4
    8.0 7.3
    6.4 6.4
    8.6 7.5
    7.7 6.1
    7.7 6.6
    5.6 6.0
    5.6 5.5
    6.2 5.5 ];
x1 = data(:,1);
x2 = data(:,2);
d = x1-x2;
n = length(d);
```

where `x1` is the yield of machine 1, and `x2` is the yield of machine 2. The average difference ($\bar{d}$) between `x1` and `x2` is

```
d_bar = mean(d)
d_bar =
    0.8300
```

The estimated variance (s_d^2) is

```
sd2 = var(d)
```

```
sd2 =

    0.6668
```

and the estimated standard deviation (s_d) is

```
sd = sqrt(sd2)

sd =

    0.8166
```

The test size

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

in Matlab code is

```
t = d_bar/(sd*sqrt(1/n))

t =

    3.2143
```

and the resulting p-value is

```
pval = 2*(1-tcdf(abs(t),n-1))

pval =

    0.0106
```

Since p<0.05, we reject the null hypothesis that there is no difference between the yield of the two machines. Since the average difference $\bar{d}$ is positive, we conclude that the data suggest that machine 1 outperforms machine 2.

The endpoints of the 95% confidence interval δ for are

```
t0 = tinv(0.975,n-1)

d_minus = d_bar - t0*sd*sqrt(1/n)

d_plus = d_bar + t0*sd*sqrt(1/n)

t0 =

    2.2622
```

```
d_minus =

    0.2459

d_plus =

    1.4141
```

That is,

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

and

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

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

6.4 PAIRED VS. UNPAIRED TEST

In **Example 25 – Which sowing machine is the better one?**, consider what would happen if we performed an *unpaired* comparison between the two machines.

***Example 26* – Which sowing machine is the better one? (Unpaired test.)**

The sample means are $\bar{x}_1 = 7.22$ and $\bar{x}_2 = 6.39$, and the empirical variances are $s_1^2 = 1.32$ and $s_2^2 = 0.62$. With $n_1 = n_2 = 10$, and assuming equal variances, the pooled variance is

$$s^2 = \frac{1}{10 + 10 - 2}\big((10 - 1) \cdot 1.13 + (10 - 1) \cdot 1.44\big) = 0.97$$

and the empirical standard deviation is $\sqrt{0.97} = 0.99$.

With the null hypothesis, $H_0: \mu_1 - \mu_2 = 0$, the t-score becomes

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

Inserting into the equation of the p-value, we get

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

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

Matlab code:

```
x1_bar = mean(x1)

x2_bar = mean(x2)

s21 = var(x1)

s22 = var(x2)

n1 = length(x1)

n2 = length(x2)

s2 = 1/(n1+n2-2)*((n1-1)*s21+(n2-1)*s22)

s = sqrt(s2)

t = (x1_bar-x2_bar)/(s*sqrt(1/n1+1/n2))

pval = 2*(1-tcdf(abs(t),n1+n2-2))
```

We see that the conclusions drawn in **Example 25 – Which sowing machine is the better one?** and **Example 26 – *Which sowing machine is the better one? (Unpaired test.)*** are contradictory. That is, the paired test suggests a difference between the two sowing machines, whereas the unpaired test doesn't. The explanation is that in the unpaired test, the difference between the two population means is due to a combination of machine effects and field effects. By pairing the data, such that we look at the difference between the two machines in similar fields, the field-effects are removed, and we can detect a difference between the machines.

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

The bottom line is; use paired data, whenever possible.

SUMMARY

The comparison of two population means is an important application within statistics. Population means can be compared using (unpaired) two-sample z-tests or t-tests, depending on whether the true population variances are known or not. In both cases, the methods provided in this chapter require that the underlying distributions of the two populations being compared have equal variance. A potential problem with unpaired tests is that a difference between two population means can be caused by other effects than the one you are interested in, or alternatively an actual difference between the populations can be masked by other effects. This problem can be addressed with paired tests, where confounding effects are eliminated by pairing the data appropriately. The typical use of paired tests is to compare an effect-measure before and after some intervention. The comparison is made on a per-sample level, such that confounding effects are removed for all samples.

PROBLEMS

1. A study wishes to compare the body temperature of men and women. The table below shows the average temperature ($\bar{x}$) measured in each group, along with the empirical standard deviation (s).

Gender	$\boldsymbol{n}$	$\boldsymbol{\bar{x}}$	$\boldsymbol{s}$
Male	65	36.725	0.699
Female	65	36.886	0.743

   a. Write down the null hypothesis that the mean body temperature of men and women is the same.
   b. Test your null hypothesis vs. the alternative hypothesis that the mean body temperature of men and women is not the same. Use a significance level of 0.05.
   c. Compute a 95% confidence interval for the difference (δ) between the population means.

2. An experiment is conducted to determine whether intensive tutoring (covering a great deal of material in a fixed amount of time) is more effective than paced tutoring (covering less material in the same amount of time). Two randomly chosen groups are tutored separately and then administered proficiency tests. See the results in the table below. Let μ_1 represent the population mean for the intensive tutoring group and μ_2 represent the population mean for the

paced tutoring group. Test the null hypothesis $H_0: \mu_1 - \mu_2 = 0$ vs. the alternative hypothesis, $H_1: \mu_1 - \mu_2 \neq 0$.

Group	Method	n	$\bar{x}$	s
1	Intensive	12	46.31	6.44
2	Paced	10	42.79	7.52

3. Using the data in the table below, estimate a 95% confidence interval for the difference (δ) between the number of raisins per box in two brands of breakfast cereal.

Brand	n	$\bar{x}$	s
A	6	102.1	12.3
B	9	93.6	7.52

4. The table below shows scores on a vocabulary test before using a new study guide and tests on a similar vocabulary test after using the study guide for a random sample of 10 students. Let δ denote the true effect of using the study guide.
    a. Test null hypothesis, $H_0: \delta = 0$, against the alternative hypothesis $H_1: \delta \neq 0$.
    b. Compute a 95% confidence interval for the effect, δ.

Person	1	2	3	4	5	6	7	8	9	10
Before	88	92	85	80	83	84	86	78	81	95
After	89	90	87	84	84	86	86	84	83	92

5. In a text reading study, 10 participants were given 2 minutes to read a simple text and a complex text, each text containing 10 words. They were informed that after 10 minutes they would be asked to recall the words of both texts. The same participants were given both texts. Half of the participants were given the simple text first and half were given the complex text first. The null hypothesis expressed that there would be no difference in the recall of words when participants looked at a simple text and a complex text. Test the null hypothesis given the data below.

Participants	Condition 1 (Simple texts)	Condition 2 (Complex texts)	Differences
1	9	3	6
2	4	3	1
3	7	5	2
4	5	6	-1
5	10	4	6
6	7	3	4
7	9	7	2
8	4	4	0
9	8	1	7
10	5	6	-1
Means	6.8	4.2	2.6

TEST CATALOG FOR COMPARING TWO MEANS (KNOWN VARIANCE)

Statistical model:

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

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

- Where the observation is $\bar{x}_1 - \bar{x}_2$ = 'the difference between two sample means'.

Hypothesis test (two-tailed):

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

95% confidence interval:

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

TEST CATALOG FOR COMPARING TWO MEANS (UNKNOWN VARIANCE)

Statistical model:

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

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

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

- Where the observation is $\bar{x}_1 - \bar{x}_2$ = 'the difference between two sample means'.

Hypothesis test (two-tailed):

- $H_0: \mu_1 = \mu_2$
- $H_1: \mu_1 \neq \mu_2$
- Test size: $t = \frac{(\bar{x}_1 - \bar{x}_2)}{s\sqrt{1/n_1 + 1/n_2}} = \sim t(n_1 + n_2 - 2)$
- Approximate p-value: $2 \cdot \left(1 - t_{cdf}(|t|, n_1 + n_2 - 2)\right)$

95% confidence interval:

- $\delta_- = (\bar{x}_1 - \bar{x}_2) - t_0 \cdot s\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}$
- $\delta_+ = (\bar{x}_1 - \bar{x}_2) + t_0 \cdot s\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}$

where `t0 = tinv(1-0.05/2,n1+n2-2)`

TEST CATALOG FOR PAIRED DATA

Statistical model:

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

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

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

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

Hypothesis test (two-tailed):

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

95% confidence interval:

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

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

7 SIMPLE LINEAR REGRESSION

In statistics, simple linear regression is the least squares estimator of a linear regression model with a single explanatory variable. In other words, simple linear regression fits a straight line through the set of n points in such a way that makes the sum of squared residuals of the model (that is, vertical distances between the points of the data set and the fitted line) as small as possible.

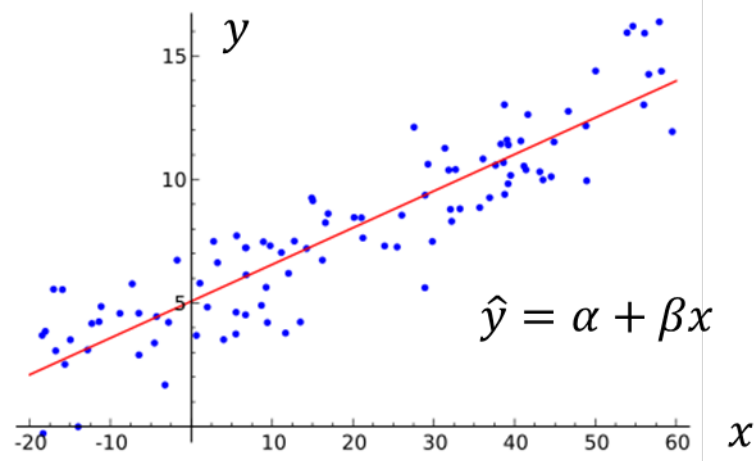


Figure 13 – Simple linear regression fits a straight line through the data points.

The adjective *simple* refers to the fact that this regression is one of the simplest in statistics. The slope of the fitted line is equal to the correlation between y and x corrected by the ratio of standard deviations of these variables. The intercept of the fitted line is such that it passes through the center of mass $(\bar{x}, \bar{y})$ of the data points.

7.1 STATISTICAL MODEL

In linear regression, the data come in pairs

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

where x is the independent variable and y is the dependent (or response) variable. Only the response variable is random. Specifically, we assume that y is normally distributed with a mean that depends linearly on x and constant variance (i.e., a variance that does not depend on x). We write this

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

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

The statistical model is illustrated graphically in Figure 14. For a given choice of x-value (x_i), the corresponding y-sample (y_i) is normally distributed with a mean value of $\alpha + \beta x_i$ and constant variance, σ^2. The figure shows the PDFs (blue curves) and samples (blue dots) from the PDFs for five different choices of x_i. The goal of linear regression is – given data pairs (x_i, y_i) – to determine the choice of slope (β) and intercept (α) that minimizes the sum of squared residuals of the model. The residual of the i'th sample (y_i) for a given choice of α and β is denoted ε_i and is given by

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

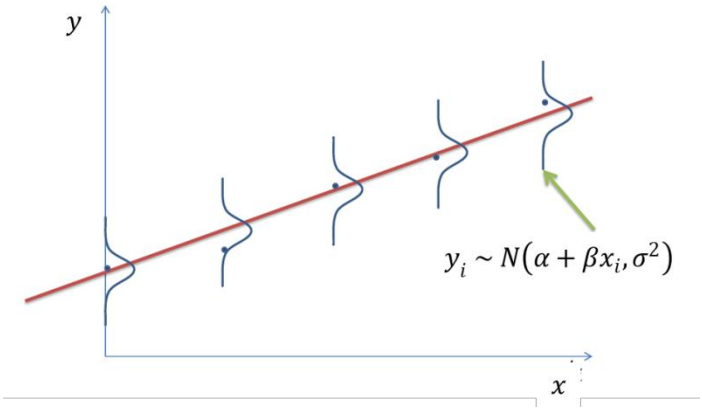


Figure 14 – Illustration of the statistical model of linear regression. Samples from *y*'s distribution are normally distributed with a mean that depends linearly on *x*, but a variance that is fixed for all *x*. The red line represents the regression line. The blue curves represent Gaussian PDFs, and the blut dots represent samples from the PDFs.

where y_i is the sampled *y*-value, and $\alpha + \beta x_i$ is the predicted *y*-value. Accordingly, the sum of squared residuals is

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

The parameter estimates that minimize *R* are

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

and

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

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

Recalling that $y_i \sim N(\alpha + \beta x_i, \sigma^2)$, we see that we also need an estimate of the variance σ^2. The unbiased estimator is

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

Statistical inference in linear regression concerns the parameter estimates: $\hat{\alpha}$, $\hat{\beta}$, and s_r^2. For instance, we can state and test hypothesis about the slope and/or intercept, or we can calculate their 95% confidence intervals.

7.2 Maximum-likelihood parameter estimation

The estimates of the slope and intercept given above are in fact the maximum-likelihood estimates. To see this, let us analyze the likelihood function (i.e., the probability of observing the data (x_i, y_i), given

some choice of parameters. Assuming that the measurements are independent, the likelihood function is

$$L(\hat{\alpha},\hat{\beta}) = f(y_1, y_2, \dots, y_n|\hat{\alpha},\hat{\beta},\sigma^2, x_1, x_2, \dots, x_n) = \prod_{i=1}^{n} f(y_i|\hat{\alpha},\hat{\beta},\sigma^2, x_i)$$

where $f(y_i|\hat{\alpha},\hat{\beta},\sigma^2, x_i)$ is a Gaussian PDF with mean $\hat{\alpha} + \hat{\beta}x_i$, variance σ^2, and observation y_i

$$f(y_i|\hat{\alpha},\hat{\beta},\sigma^2, x_i) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\left(y_i-(\hat{\alpha}+\hat{\beta}x_i)\right)^2/2\sigma^2}$$

Inserting and rearranging, we get

$$L(\hat{\alpha},\hat{\beta}) = \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\left(y_i-(\hat{\alpha}+\hat{\beta}x_i)\right)^2/2\sigma^2} = \left(\frac{1}{\sqrt{2\pi\sigma^2}}\right)^n \cdot \prod_{i=1}^{n} e^{-\left(y_i-(\hat{\alpha}+\hat{\beta}x_i)\right)^2/2\sigma^2}$$
$$= \left(\frac{1}{\sqrt{2\pi\sigma^2}}\right)^n \cdot e^{-\sum_{i=1}^{n}\left(y_i-(\hat{\alpha}+\hat{\beta}x_i)\right)^2/2\sigma^2} = \left(\frac{1}{\sqrt{2\pi\sigma^2}}\right)^n \cdot e^{-R(\hat{\alpha},\hat{\beta})/2\sigma^2}$$

where $R(\hat{\alpha},\hat{\beta})$ is the sum of squared residuals defined above. To get rid of the product, we have used the familiar relation, $e^a \cdot e^b = e^{a+b}$. It follows from the above equation that maximizing the likelihood function with respect to $\hat{\alpha}$ and $\hat{\beta}$ is the equivalent to minimizing $R(\hat{\alpha},\hat{\beta})$. Differentiating $R(\hat{\alpha},\hat{\beta})$ and setting to zero gives the maximum-likelihood estimators of choice of α and β.

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

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

It follows that

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

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

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

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

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

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

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

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

It follows that

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

Thus, we have shown that the maximum-likelihood estimates of the slope and intercept are

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

and

$$\hat{\alpha} = \bar{y} - \hat{\beta} \cdot \bar{x} \quad \text{(intercept)}$$

Let us see an example.

Example 27 – Hubble's low

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

```
Distance = [ 0.032 0.034 0.214 0.263 0.275 0.275 0.450 ...
             0.500 0.500 0.630 0.800 0.900 0.900 0.900 ...
             0.900 1.000 1.100 1.100 1.400 1.700 2.000 ...
             2.000 2.000 2.000 ];
Speed = [ 170 290 -130 -70 -185 -220 200 290 ...
          270 200 300 -30 650 150 500 920 ...
          450 500 500 960 500 850 800 1090 ];
```

Choosing x = 'Distance' and y = 'Speed',

```
x = Distance;

y = Speed;
```

we wish to estimate the regression slope and intercept. The estimated slope is

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

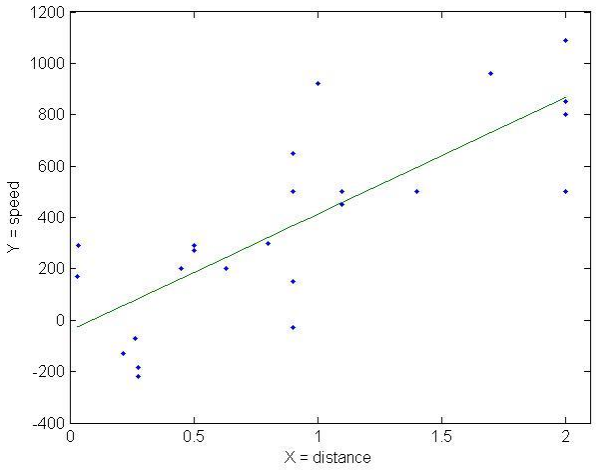
```
beta = sum((x-mean(x)).*(y-mean(y)))/sum((x-mean(x)).^2)
```

and the estimated intercept is

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

```
alpha = mean(y) - beta*mean(x)
```

Plotting the data along with the fitted straight line, we see that there is indeed an approximate linear relationship between distance and speed.



Matlab code for plotting:

```
plot(x,y,'.',...
     x,alpha+beta*x)
xlabel('X = distance')
ylabel('Y = speed')
axis([0 2.1 -400 1200])
```

7.3 STATISTICAL INFERENCE ON THE REGRESSION SLOPE

When considering the regression slope (β) the typical question of interest is whether the slope deviates significantly from zero; this would suggest that there is some correlation between *x* and *y*. In general, the null hypothesis that we wish to test takes the following form

$$H_0: \beta = \beta_0$$

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

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

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

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

where

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

$$\hat{\alpha} = \bar{y} - \hat{\beta} \cdot \bar{x} \quad \text{(intercept)},$$

and

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

The p-value is

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

and the 95% confidence interval for the slope is

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

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

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

Example 28 – Statistical inference on the regression slope (Hubble's low)

Using the data provided in **Example 27 – Hubble's low**, let us test whether the regression slope deviates significantly from zero. The null hypothesis is

$$H_0: \beta = 0$$

with the alternative hypothesis

$$H_1: \beta \neq 0$$

Parameter estimates:

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

```
beta = sum((x-mean(x)).*(y-mean(y)))/sum((x-mean(x)).^2)
```

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

```
alpha = mean(y) - beta*mean(x)
```

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

```
n = length(x);
```

```
sr2 = 1/(n-2)*sum((y-(alpha+beta*x)).^2)
```

$$s_r = \sqrt{54247} = 232.91$$

Test size:

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

```
sr=sqrt(sr2)
```

```
t = (beta-0)/(sr*sqrt(1/sum((x-mean(x)).^2)))
```

p-value:

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

```
2*(1-tcdf(abs(t),n-2))
```

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

95% confidence interval:

$$t_0 = 2.0739$$

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

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

```
t0 = tinv(0.975,n-2)
```

```
beta_minus = beta - t0*sr*sqrt(1/sum((x-mean(x)).^2))
```

```
beta_plus = beta + t0*sr*sqrt(1/sum((x-mean(x)).^2))
```

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

7.4 STATISTICAL INFERENCE ON THE REGRESSION INTERCEPT

When considering the regression intercept (α) the typical question of interest is whether the intercept deviates significantly from zero. In general, the null hypothesis that we wish to test takes the following form

$$H_0: \alpha = \alpha_0$$

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

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

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

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

where $\hat{\beta}$, $\hat{\alpha}$, and s_r^2 are given as above.

The p-value is

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

and the 95% confidence interval for the intercept is

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

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

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

Example 29 – Statistical inference on the regression intercept (Hubble's low)

Using the data provided in **Example 27 – Hubble's low**, let us test whether the regression intercept deviates significantly from zero. The null hypothesis is

$$H_0: \alpha = 0$$

with the alternative hypothesis

$$H_1: \alpha \neq 0$$

Parameter estimates are the same as above:

$$\hat{\beta} = 454.1584$$

$$\hat{\alpha} = -40.7836$$

$$s_r^2 = 54247$$

$$s_r = 232.91$$

Test size:

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

```
t = alpha/(sr*sqrt(1/n+mean(x)^2/sum((x-mean(x)).^2)))
```

p-value:

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

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

95% confidence interval:

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

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

```
alph_minus = alpha-sr*sqrt(1/n+mean(x)^2/sum((x-mean(x)).^2))
```

```
alph_plus = alpha+sr*sqrt(1/n+mean(x)^2/sum((x-mean(x)).^2))
```

Since $\alpha = 0$ is included in the 95% confidence interval, we fail to reject the null hypothesis.

7.5 CHECKING FOR NORMALITY

Recalling that the statistical model underlying linear regression is

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

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

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

Hence, a good way to check whether the assumption of linearity between x and y holds is to first fit the linear model and subsequently check that the residuals are normally distributed. We have already seen a method for checking whether sampled data are normally distributed, namely the Q-Q plot. Thus, one way to check that the statistical model holds is to make a Q-Q plot of the residuals (ϵ_i) and check that they lie approximately on a straight line.

Example 30 – Checking for normality using Q-Q plot (Hubble's low)

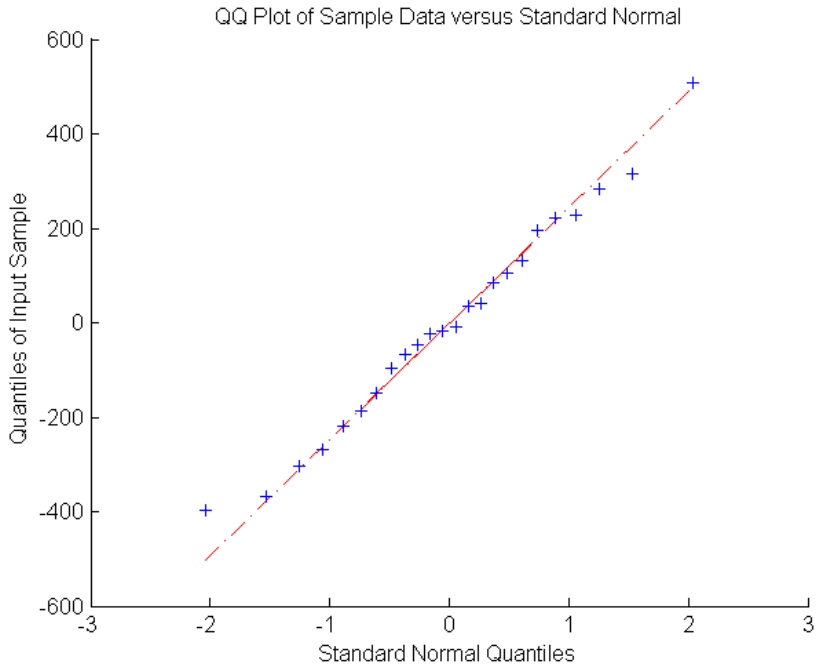
The residuals in Hubble's low example are

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

The resulting Q-Q plot

```
qqplot(res)
```

shows that the residuals are approximately normally distributed, because the data points lie approximately on a straight line. Hence, it is safe to use simple linear regression to find the relation between the Speed and Distance of galaxies.



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

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

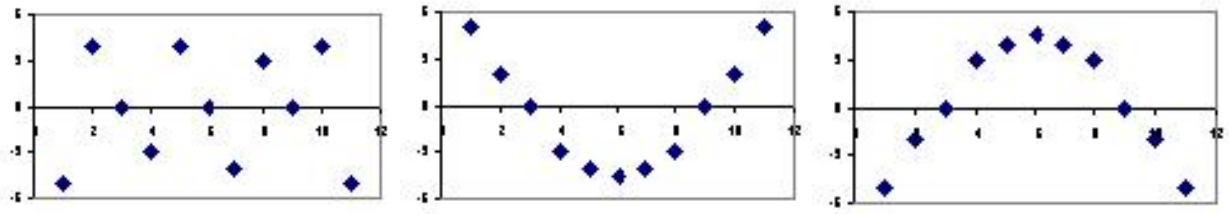


Figure 15 – Examples of residual plots. The first plot shows a random pattern, indicating a good fit for a linear model. The other plot patterns are non-random, suggesting a better fit for a non-linear model.

Formally, you must check the following two conditions:

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

Example 31– Checking for normality using residual plot (Hubble's low)

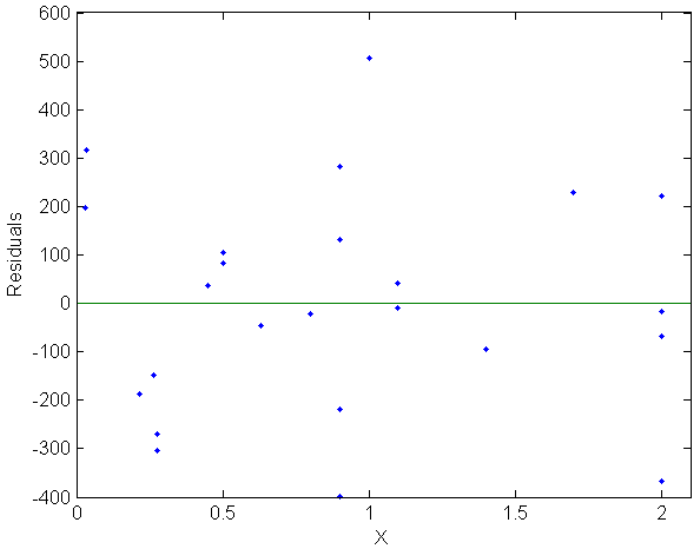
Again, the residuals in Hubble's low example are

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

The resulting residual plot

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

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



7.6 USAGE OF LINEAR REGRESSION

PREDICTION AND EXTRAPOLATION

Linear regression is often used for prediction. Suppose, for instance, that the relationship between daily energy consumption of a power plant and the outside temperature is linear. Then, given the temperature of tomorrow (from a weather forecast), we can give an estimate of tomorrows energy consumption of the power plant based on a linear model. However, make sure never to predict values outside the range of *x*-values that was used to fit the linear model (this is called extrapolation). Figure 16 shows an example were extrapolation goes terribly wrong.

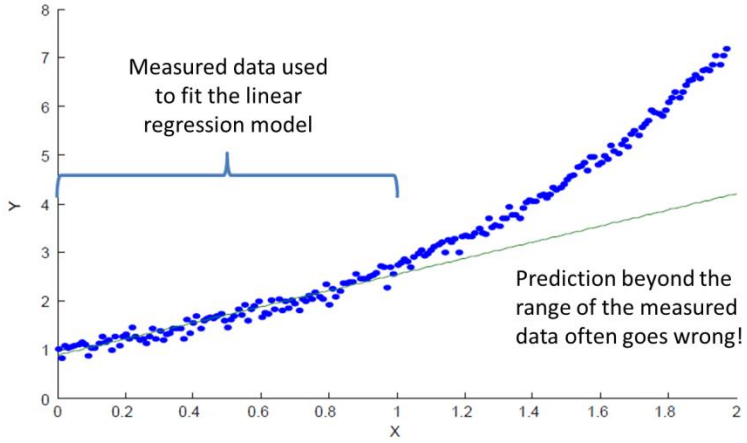


Figure 16 – Prediction outside the range of the measured data (extrapolation) is in general prohibited.

COEFFICIENT OF DETERMINATION

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

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

where s_x and s_y are the empirical standard deviations of x and y. As we saw in Cooper/McGillem chapter 3, the correlation coefficient takes on values from -1 to 1. It can be shown that the estimate of the regression slope is linearly related to the correlation coefficient:

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

Let us verify

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

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

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

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

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

$$= \hat{\beta}$$

In simple linear regression, the *coefficient of determination*

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

pronounced R squared, is a number that indicates how well the data fit the linear model. As can be seen, R^2 is simply the square of the sample correlation coefficient between the outcomes and their predicted values. The coefficient of determination ranges from 0 to 1 with value close to 1 suggesting a strong linear relationship, and values close to 0 suggesting no linear relationship (an R^2 of 1 indicates that the regression line perfectly fits the data).

The coefficient of determination in the example with Hubble's low is $R^2 = 0.62$. A coefficient of determination equal to 0.62 indicates that about 62% of the variation in galaxy speed (the dependent variable) can be explained by the relationship to galaxy distance (the independent variable). To calculate the sample correlation coefficient between *x* and *y* in Matlab, use the command `corr2(x,y)`.

Note that R^2 does not indicate whether:

- the independent variable (x) is a cause of the changes in the response variable (y);
- the correct regression was used;
- there are enough data points to make a solid conclusion.

TRANSFORMATIONS TO ACHIEVE LINEARITY

When a residual plot reveals a data set to be nonlinear, it is often possible to "transform" the raw data to make it more linear. This allows us to use linear regression techniques more effectively with nonlinear data. A nonlinear transformation changes (increases or decreases) linear relationships between variables and, thus, changes the correlation between variables. Examples of a nonlinear transformation of variable *y* would be taking the square root of *y*, the logarithm of *y*, or the reciprocal of *y*. In general, it may be necessary to transform the independent variable, the dependent variable, or both.

OUTLIERS

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

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

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

Figure 17 – Example of a data set with many outliers hiding the linear relationship.

Outliers can be handled either by excluding them from the regression analysis or by using regression techniques that are less sensitive to outliers. One example of such a technique is the “RANdom Sample Consensus” (RANSAC). RANSAC is an iterative method to estimate parameters of a mathematical model from a set of observed data which contains outliers. It is a non-deterministic algorithm in the sense that it produces a reasonable result only with a certain probability, with this probability increasing as more iterations are allowed.

A basic assumption is that the data consists of "inliers", i.e., data whose distribution can be explained by some set of model parameters, though may be subject to noise, and "outliers" which are data that do not fit the model. The outliers can come, e.g., from extreme values of the noise or from erroneous measurements or incorrect hypotheses about the interpretation of data. RANSAC also assumes that, given a (usually small) set of inliers, there exists a procedure which can estimate the parameters of a model that optimally explains or fits this data.

The input to the RANSAC algorithm is a set of observed data values, a way of fitting some kind of model to the observations, and some confidence parameters. RANSAC achieves its goal by repeating the following steps:

1. Select a random subset of the original data. Call this subset the hypothetical inliers.
2. A model is fitted to the set of hypothetical inliers.
3. All other data are then tested against the fitted model. Those points that fit the estimated model well, according to some model-specific loss function, are considered as part of the consensus set.
4. The estimated model is reasonably good if sufficiently many points have been classified as part of the consensus set.
5. Afterwards, the model may be improved by reestimating it using all members of the consensus set.

Figure 18 shows the result of applying linear regression with RANSAC to the data shown in Figure 17.

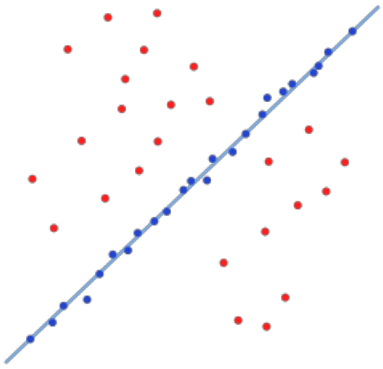


Figure 18 – Result of applying RANSAC to make linear regression more robust towards outliers.

PROBLEMS

1. When deciding whether to admit an applicant, colleges take lots of factors, such as grades, sports, activities, leadership positions, awards, teacher recommendations, and test scores, into consideration. Using SAT scores as a basis of whether to admit a student or not has created some controversy. Among other things, people question whether the SATs are fair and whether they predict college performance. This study examines the SAT and GPA information of 105

students who graduated from a state university with a B.Sc. in computer science. Using the grades and test scores from high school, can you predict a student's college grades?

Question to answer: Are the high school and college GPAs related?

Copy-paste the data on pages 100-101 below into your Matlab editor, and define the variables

```
x = data(:,1); % High school GPA
y = data(:,4); % University GPA
```

a. Draw a scatterplot comparing the students' high school GPAs (x) to their overall university GPAs (y). What does the relationship appear to be? (Hint: you can use the `scatter` command in Matlab).
b. What is the sample correlation coefficient between high school GPA (x) and overall university GPA (y)?
c. Find the regression line for predicting the overall university GPA (y) from the high school GPA (x).
d. What is the slope?
e. What is the y-intercept?
f. Does the assumption of a linear relationship between x and y seem to hold? (Use Q-Q plot or residual plot to answer this question).
g. If someone had a 2.2 GPA in high school, what is the best estimate of his or her college GPA?
h. If someone had a 4.0 GPA in high school, what is the best estimate of his or her college GPA?
i. Test the null hypothesis that the slope is zero, $H_0: \beta = 0$.
j. Compute the 95% confidence interval for the intercept (α).

2. Given the following data

```
x = [ 1 2 3 4 5 6 7 8 9 ]
y = [ 2 1 6 14 15 30 40 74 75 ]
```

a. Make a scatterplot of y vs. x
b. Find the regression line for predicting y from x.
c. Make a residual plot. Does the assumption of linearity seem to hold? (the answer is no!).
d. Transform y by taking the square root:

```
y = sqrt([ 2 1 6 14 15 30 40 74 75 ])
```

e. Find the regression line for predicting sqrt(y) from x.
f. Make a residual plot. Does the assumption of linearity seem to hold?

```
% DATA
% high_GPA math_SAT verb_SAT comp_GPA univ_GPA
data = [ 3.45 643 589 3.76 3.52
2.78 558 512 2.87 2.91
2.52 583 503 2.54 2.4
3.67 685 602 3.83 3.47
3.24 592 538 3.29 3.47
2.1 562 486 2.64 2.37
2.82 573 548 2.86 2.4
2.36 559 536 2.03 2.24
2.42 552 583 2.81 3.02
3.51 617 591 3.41 3.32
3.48 684 649 3.61 3.59
2.14 568 592 2.48 2.54
2.59 604 582 3.21 3.19
3.46 619 624 3.52 3.71
3.51 642 619 3.41 3.58
3.68 683 642 3.52 3.4
3.91 703 684 3.84 3.73
3.72 712 652 3.64 3.49
2.15 564 501 2.14 2.25
2.48 557 549 2.21 2.37
3.09 591 584 3.17 3.29
2.71 599 562 3.01 3.19
2.46 607 619 3.17 3.28
3.32 619 558 3.01 3.37
3.61 700 721 3.72 3.61
3.82 718 732 3.78 3.81
2.64 580 538 2.51 2.4
2.19 562 507 2.1 2.21
3.34 683 648 3.21 3.58
3.48 717 724 3.68 3.51
3.56 701 714 3.48 3.62
3.81 691 684 3.71 3.6
3.92 714 706 3.81 3.65
4 689 673 3.84 3.76
2.52 554 507 2.09 2.27
2.71 564 543 2.17 2.35
3.15 668 604 2.98 3.17
3.22 691 662 3.28 3.47
2.29 573 591 2.74 3
2.03 568 517 2.19 2.74
3.14 607 624 3.28 3.37
3.52 651 683 3.68 3.54
2.91 604 583 3.17 3.28
2.83 560 542 3.17 3.39
2.65 604 617 3.31 3.28
2.41 574 548 3.07 3.19
2.54 564 500 2.38 2.52
2.66 607 528 2.94 3.08
3.21 619 573 2.84 3.01
3.34 647 608 3.17 3.42
3.68 651 683 3.72 3.6
2.84 571 543 2.17 2.4
2.74 583 510 2.42 2.83
2.71 554 538 2.49 2.38
2.24 568 519 3.38 3.21
2.48 574 602 2.07 2.24
3.14 605 619 3.22 3.4
```

```
2.83 591 584 2.71 3.07
3.44 642 608 3.31 3.52
2.89 608 573 3.28 3.47
2.67 574 538 3.19 3.08
3.24 643 607 3.24 3.38
3.29 608 649 3.53 3.41
3.87 709 688 3.72 3.64
3.94 691 645 3.98 3.71
3.42 667 583 3.09 3.01
3.52 656 609 3.42 3.37
2.24 554 542 2.07 2.34
3.29 692 563 3.17 3.29
3.41 684 672 3.51 3.4
3.56 717 649 3.49 3.38
3.61 712 708 3.51 3.28
3.28 641 608 3.4 3.31
3.21 675 632 3.38 3.42
3.48 692 698 3.54 3.39
3.62 684 609 3.48 3.51
2.92 564 591 3.09 3.17
2.81 554 509 3.14 3.2
3.11 685 694 3.28 3.41
3.28 671 609 3.41 3.29
2.7 571 503 3.02 3.17
2.62 582 591 2.97 3.12
3.72 621 589 4 3.71
3.42 651 642 3.34 3.5
3.51 673 681 3.28 3.34
3.28 651 640 3.32 3.48
3.42 672 607 3.51 3.44
3.9 591 587 3.68 3.59
3.12 582 612 3.07 3.28
2.83 609 555 2.78 3
2.09 554 480 3.68 3.42
3.17 612 590 3.3 3.41
3.28 628 580 3.34 3.49
3.02 567 602 3.17 3.28
3.42 619 623 3.07 3.17
3.06 691 683 3.19 3.24
2.76 564 549 2.15 2.34
3.19 650 684 3.11 3.28
2.23 551 554 2.17 2.29
2.48 568 541 2.14 2.08
3.76 605 590 3.74 3.64
3.49 692 683 3.27 3.42
3.07 680 692 3.19 3.25
2.19 617 503 2.98 2.76
3.46 516 528 3.28 3.41 ];
```

TEST CATALOG FOR THE SLOPE IN SIMPLE LINEAR REGRESSION

Statistical model:

- $y_i \sim N(\alpha + \beta x_i, \sigma^2)$ for $i = 1,2, \dots, n$ are independent samples.
- Parameter estimates:

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

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

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

Hypothesis test (two-tailed):

- $H_0: \beta = \beta_0$
- $H_1: \beta \neq \beta_0$
- Test size: $t = \frac{\hat{\beta} - \beta_0}{s_r\sqrt{1/\sum_{i=1}^{n}(x_i - \bar{x})^2}} \sim t(n-2)$
- Approximate p-value: $2 \cdot \left(1 - t_{cdf}(|t|, n-2)\right)$

95% confidence interval:

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

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

TEST CATALOG FOR THE INTERCEPT IN SIMPLE LINEAR REGRESSION

Statistical model:

- $y_i \sim N(\alpha + \beta x_i, \sigma^2)$ for $i = 1,2, \dots, n$ are independent samples.
- Parameter estimates:

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

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

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

Hypothesis test (two-tailed):

- $H_0: \alpha = \alpha_0$
- $H_1: \alpha \neq \alpha_0$
- Test size: $t = \frac{\hat{\alpha} - \alpha_0}{s_r\sqrt{\frac{1}{n} + \frac{\bar{x}^2}{\sum_{i=1}^{n}(x_i - \bar{x})^2}}} \sim t(n-2)$
- Approximate p-value: $2 \cdot \left(1 - t_{cdf}(|t|, n-2)\right)$

95% confidence interval:

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

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