

Lecture 1

BASIC PROBABILITY

Introduction

- ▶ This section of the course consists of 11 lectures on probability and statistics.
- ▶ Probability can be thought of as a mathematical model for phenomena involving uncertainty.
- ▶ Statistics is concerned with making conclusions based on data involving uncertainty.
- ▶ In this section we will also introduce the use of MATLAB for statistical calculations.

Why probability and statistics?

Uncertainty or random variation is widespread and occurs at many levels relevant in engineering applications.

- ▶ Impurity levels in raw materials will be subject to random fluctuation;
- ▶ Inflow to a dam will vary from year to year;
- ▶ Peak power consumption will vary from day to day;
- ▶ Manufactured items will be subject to unwanted variability:
 - ▶ Proportion of defective items will vary from batch to batch;
 - ▶ Imperfections will occur on welds.
- ▶ Input voltage to an engine control unit will be subject to variation under operating conditions.

Probability Models

- ▶ Probability models allow us to make quantitative predictions for systems involving uncertainty. For example:
 - ▶ What is the water level expected for a “once in 100 year” flooding event?
 - ▶ If a quality inspector finds 5 defective items in a sample of 100, what is the probability that the batch will be rejected by a customer?
- ▶ Probability models also provide the basis for many types of simulation.
 - ▶ Stochastic simulation is a powerful computational methodology.
 - ▶ Amongst other things, it allows us to model complex systems involving multiple sources of uncertainty.

Statistics

Statistical inference is concerned with making valid conclusions from data involving uncertainty.

- ▶ Statistical estimation is often needed to estimate parameters required for probability models.
 - ▶ For example, a probability model for expected flood levels will need to be constructed using data observed over previous years.
- ▶ Statistical inference is also needed to make decisions in the face of uncertainty. For example:
 - ▶ Using experimental data to determine the optimal temperature at which to conduct a certain reaction.
 - ▶ Deciding whether to accept or reject a batch of items based on the number of defectives observed in a sample.

Matlab

- ▶ In practice, statistical calculations are performed using a statistical package.
 - ▶ Professionally written packages are generally reliable and efficient.
 - ▶ There is very little need to be able to program statistical calculations in languages such as FORTRAN or VBA.
 - ▶ Calculators can be used but tend to be limited with respect to data handling and graphics.
- ▶ The Matlab statistical toolbox provides access to many standard probability and statistical calculations.
- ▶ It is convenient for handling data.
- ▶ It is accessible in all of the CAT suites.

Elementary probability theory

Kreyszig, E. (2006) *Advanced Engineering Mathematics* 9th ed.
§24.3

- ▶ Probability theory is a part of mathematics that deals with random phenomena.
- ▶ A random phenomenon can be thought of as an *experiment* that can be repeated and whose outcome is not predictable.
- ▶ For example,
 - ▶ tossing a coin;
 - ▶ rolling a die;
 - ▶ observing arrival times of customers at a queue.

Definitions, concepts and notation

- ▶ The **sample space** S is the **set** of all possible outcomes for a given experiment.
- ▶ An **event** A is a subset $A \subseteq S$.
 - ▶ The event A is said to occur if the outcome of the experiment is an element of A .
 - ▶ The probability P is a function which assigns each event $A \subseteq S$ a number $P(A)$, and $0 \leq P(A) \leq 1$.

Frequency interpretation of probability

Consider a long sequence of trials in which a certain experiment is repeated independently and under identical conditions.

- ▶ Frequency interpretation of probability:
- ▶ the proportion of trials in which the event A occurs will *settle down* to $P(A)$ as the number of trials increases.

Note that probability is a characteristic of the actual experiment (equipment and procedure).

The axioms of probability

1. For any event A , $P(A) \geq 0$.
2. $P(S) = 1$.
3. 3.1 For disjoint events A_1 and A_2 ,

$$P(A_1 \cup A_2) = P(A_1) + P(A_2).$$

- 3.2 More generally, for a sequence of *pairwise* disjoint events A_1, A_2, A_3, \dots ,

$$P(A_1 \cup A_2 \cup A_3 \cup \dots) = P(A_1) + P(A_2) + P(A_3) + \dots$$

Recap

- ▶ $A_1 \cap A_2$ is the set of all outcomes in *both* A_1 and A_2 .
- ▶ $A_1 \cup A_2$ is the set of all outcomes in *either* A_1 and A_2 (or both).
- ▶ Disjoint means that $P(A_1 \cap A_2) = \phi$.

The complement rule I

- ▶ For any event A let A^c denote the **complementary event**.

- ▶ That is, A^c occurs $\Leftrightarrow A$ does not occur.

- ▶ The law of complementary probability states that

$$P(A^c) = 1 - P(A).$$

- ▶ This result is derived from the axioms of probability as follows:

The addition rule

Consider two events A and B .

- ▶ If $A \cap B = \phi$ then according to the third axiom of probability

$$P(A \cup B) = P(A) + P(B).$$

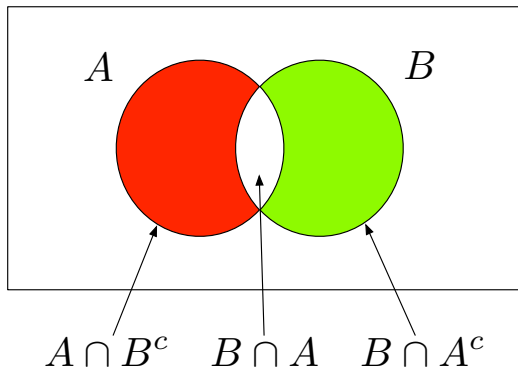
- ▶ Consider now the case of A and B not disjoint. Then

$$P(A \cup B) = P(A) + P(B) - P(A \cap B).$$

Proof of the addition rule I

Observe first that $A \cup B$ is expressible as the pairwise disjoint union.

$$A \cup B = (A \cap B^c) \cup (B \cap A^c) \cup (A \cap B)$$



Example

A single card is drawn from a well shuffled pack. What is the probability it is either a diamond or a king?

Example (continued)

We will find $P(A \cup B)$ using the formula,

$$P(A \cup B) = P(A) + P(B) - P(A \cap B).$$

Conditional Probability

- ▶ Consider two events A and B such that $P(B) > 0$.
- ▶ The **conditional probability of A given B** is defined by,

$$P(A|B) = \frac{P(A \cap B)}{P(B)}.$$

The multiplication rule

For two events A and B ,

$$P(A \cap B) = P(B) \times P(A|B)$$

and

$$P(A \cap B) = P(A) \times P(B|A).$$

Remarks:

- ▶ The multiplication rule is a simple rearrangement of the definition of conditional probability;
- ▶ The multiplication rule provides justification for multiplying probabilities on tree diagrams.

Independence

Two events A and B are said to be independent if

$$P(A \cap B) = P(A)P(B).$$

Remark

- ▶ If A and B are independent then $P(A|B) = P(A)$ and $P(B|A) = P(B)$. Therefore the occurrence of A does not affect the likelihood of B and vice versa.

Example I

A standard deck of cards is shuffled and a card is randomly selected from the deck.

1. What is the probability of a king?
2. What is the probability of a red card?
3. What is the probability of a red king?
4. Given that the card is red, what is the probability it is a king?
5. Given that the card is a king, what is the probability that it is red?
6. Are the events, red card and king independent?

The law of total probability

The multiplication rule can be used to evaluate probabilities through the law of total probability:

For events A and B such that $P(B) > 0$ and $P(B^c) > 0$:

$$P(A) = P(B)P(A|B) + P(B^c)P(A|B^c).$$

Example A manufacturer receives resistors from two supplies S_1 and S_2 such that:

- ▶ 80% of the resistors are from S_1 and 20% are from S_2 ;
- ▶ 1% of resistors from S_1 are out of spec and 2% of resistors from S_2 are out of spec.

Find the probability that a randomly chosen resistor is out of spec.

Solution

- ▶ Let A be the event that a resistor is out of spec and B be the event that a resistor was supplied by S_1 .
 - ▶ Note that B^c is the event that the resistor was supplied by S_2 .
- ▶ From the information provided:

Bayes' Theorem

The law of total probability and the multiplication can be used to calculate conditional probabilities.

Recall that

$$P(B|A) = \frac{P(B \cap A)}{P(A)}.$$

Substituting

$$P(B \cap A) = P(B)P(A|B)$$

and

$$P(A) = P(B)P(A|B) + P(B^c)P(A|B^c).$$

gives Bayes' Theorem:

$$P(B|A) = \frac{P(B)P(A|B)}{P(B)P(A|B) + P(B^c)P(A|B^c)}.$$

Example *ctd*

Continuing with the resistor example, suppose a randomly chose resistor is found to be out of spec. What is the probability it was supplied by S_1 .

Solution

- ▶ In the same notation as before, we want to find $P(B|A)$.

Lecture 2

Probability and Discrete Random Variables

Random Variables

- ▶ A **random variable (RV)**, X , is a measurement resulting in a real value determined by the random outcome of an experiment.
 - ▶ A web server may be working or not working at particular time.
 - ▶ A random variable X can be defined to have the value 1 for working and 0 for not working.
 - ▶ Calls are received at an IT helpdesk in a certain 10 minute period.
 - ▶ The number of calls, N , arriving within the 10 minute interval is a random variable. The possible values could be $0, 1, \dots$
- ▶ Random variables can be *discrete* or *continuous*.

Probability Distributions

- ▶ We use uppercase letters X, Y, Z, \dots to denote random variables, and the corresponding lower case letters x, y, z, \dots for particular values.
 - ▶ For example, $X = 5$ is the event that the random variable X takes the value 5.
 - ▶ $X = x$ is the event that the random variable X takes the value x .
 - ▶ $X \leq x$ is the event that the random variable X does not exceed the value x .
- ▶ A random variable can be described mathematically by its probability distribution.
- ▶ The specification of probability distributions is handled differently for discrete and continuous random variables.

Discrete Probability Distributions

Let X be a discrete random variable.

- ▶ The **probability (mass) function (pmf)** $p(x)$ is defined by

$$p(x) = P(X = x).$$

- ▶ A probability function p satisfies:

$$p(x) \geq 0 \text{ for all } x.$$

$$\sum_x p(x) = 1$$

where the summation is over all possible values of X .

Cumulative Distribution Functions

- ▶ The **cumulative distribution function (cdf)** F of X is defined by

$$F(x) = P(X \leq x).$$

- ▶ The cumulative distribution function F satisfies:

$$0 \leq F(x) \leq 1 \text{ for all } x,$$

$$F(x) \text{ is a non-decreasing function,}$$

$$\lim_{x \rightarrow -\infty} F(x) = 0 \text{ and } \lim_{x \rightarrow +\infty} F(x) = 1.$$

- ▶ If X is a discrete RV with pmf $p(x)$ then

$$F(x) = \sum_{y \leq x} p(y) = \sum_{y \leq x} P(X = y).$$

The Bernoulli Distribution

- ▶ The Bernoulli distribution has a single parameter

$$0 < p < 1$$

called the success probability.

- ▶ The possible values are 0 (failure) and 1 (success).
- ▶ If X has a Bernoulli distribution, its probability function is defined by

$$p(x) = p^x(1-p)^{1-x} = \begin{cases} p & \text{if } x = 1, \\ 1-p & \text{if } x = 0. \end{cases}$$

The Bernoulli Distribution *ctd*

- ▶ The Bernoulli distribution is used to code experiments with two different outcomes.
- ▶ Examples:
 - ▶ A randomly sampled component may be defective ($X = 1$) or sound ($X = 0$).
 - ▶ A single coin toss may be either a head ($X = 1$) or a tail ($X = 0$).
 - ▶ At the time of the accident, the driver wore a seat belt ($X = 1$), or did not ($X = 0$).
- ▶ Such variables are sometimes called *binary* variables or *dichotomous* variables.

The Binomial Distribution

- ▶ The binomial distribution $B(n, p)$ has two parameters, the success probability $0 \leq p \leq 1$, and a positive integer n .
- ▶ If X has a binomial distribution $B(n, p)$, its probability function is

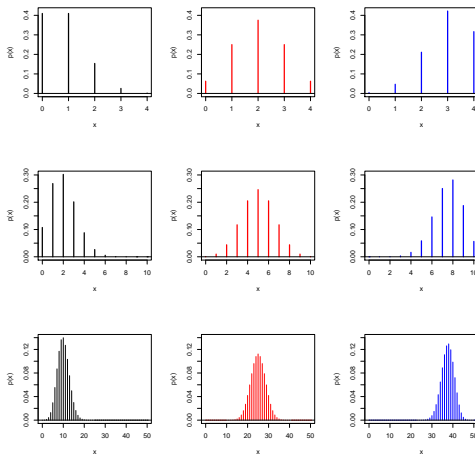
$$p(x) = \binom{n}{x} p^x (1 - p)^{n-x}.$$

- ▶ For a sequence of n independent Bernoulli trials, each with success probability p , the total number of successes, X , has the binomial distribution $B(n, p)$.
- ▶ If $n = 1$ then the binomial distribution reduces to the Bernoulli.

Binomial probabilities

Probability mass functions: $B(n, p)$ for

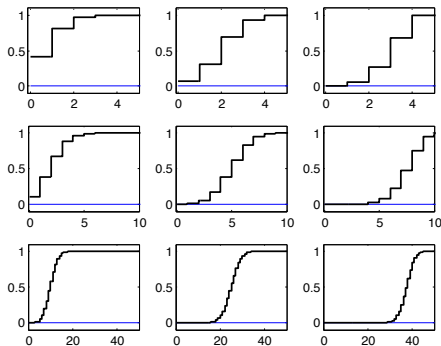
$$n = 4, 10, 50 \quad \text{and} \quad p = 0.2, 0.5, 0.75$$



The Binomial Distribution *ctd*

Cumulative distribution functions: $B(n, p)$ for

$$n = 4, 10, 50 \quad \text{and} \quad p = 0.2, 0.5, 0.75$$



The Binomial Distribution *ctd*

An outline of the derivation of the binomial distribution is as follows.

- ▶ Consider a sequence of n independent Bernoulli trials and the event $\{X = x\}$. That is, the occurrence of x successes.
- ▶ One way for this to occur is to have x successes followed by $n - x$ failures.

The probability of this outcome is

$$\overbrace{p \cdot p \cdots p}^x \overbrace{(1-p) \cdot (1-p) \cdots (1-p)}^{n-x} = p^x (1-p)^{n-x}.$$

- ▶ The x successes and $n - x$ failures could occur in many other orders.
 - ▶ There are $\binom{n}{x}$ such arrangements possible;
 - ▶ Each such arrangement occurs with probability $p^x(1 - p)^{n-x}$.
- ▶ Therefore the probability of exactly x successes is

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

Example

1% of tiles provided by a manufacturer are broken during packing and shipping. Find the probability that a randomly chosen carton of 25 tiles contains more than one broken tile.

Example *ctd* I

Solution: Let X be the number of broken tiles. Then X can be modelled as a binomial variable with $n = 25$ and $\pi = 0.01$. i.e. $X \sim B(25, 0.01)$.

$$P(X = x) = \binom{25}{x} (0.01)^x (0.99)^{25-x} \quad x = 0, 1, \dots, 25.$$

Binomial Calculations in Matlab

command	description	example
<code>binopdf(x, n, p)</code>	$B(n, p)$ pmf at x	<code>binopdf(4, 7, 0.4)</code> 0.1935
<code>binocdf(x, n, p)</code>	$B(n, p)$ cdf at x	<code>binocdf(4, 7, 0.4)</code> 0.9037
<code>binornd(n, p, m, k)</code>	generate $m \times k$ array of rvs from $B(n, p)$	<code>binornd(7, 0.4, 1, 5)</code> 4 3 2 3 3

The Binomial Distribution - Summary

- ▶ If the rv X is the number of successes in n trials, then $X \sim B(n, p)$ provided the following four conditions are satisfied:
 1. There is a fixed number n of trials (repetitions of the experiment).
 2. The n trials are independent.
 3. At each trial, we observe one of two possible outcomes, either a success or a failure.
 4. The probability p of success is the same for each trial.
- ▶ The probability function for the binomial distribution is

$$p(x) = \binom{n}{x} p^x (1-p)^{n-x} \text{ for } x = 0, 1, \dots, n.$$

The Poisson Distribution

- ▶ Consider the number of calls, X , logged by a help desk between 12:00noon and 1:00pm on a given day.
 - ▶ X is a discrete variable but the binomial model is not suitable.
 - ▶ For such variables, the Poisson distribution is often appropriate.

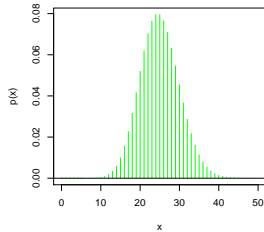
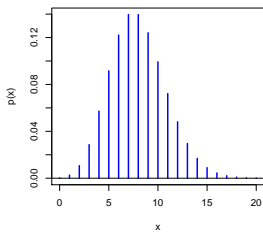
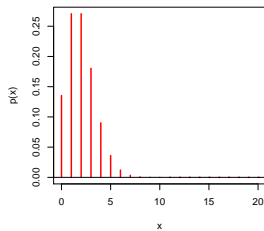
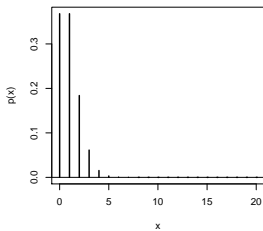
The Poisson Distribution

- ▶ The Poisson distribution $Po(\mu)$ depends on a single *mean* parameter $\mu > 0$.
 - ▶ possible values of a Poisson rv X are $0, 1, 2, 3, \dots$
 - ▶ The probability function is given by

$$p(x) = \frac{e^{-\mu} \mu^x}{x!}.$$

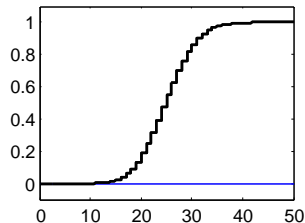
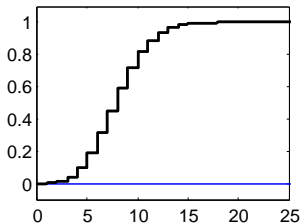
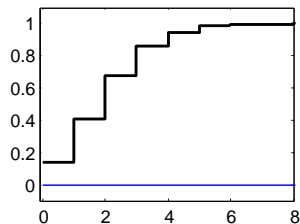
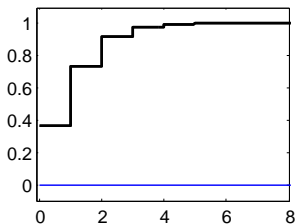
- ▶ The Poisson distribution is used to model the number of events of a given type over a fixed interval in time or space.

Poisson probabilities $Po(\mu)$ for $\mu = 1, 2, 8, 25$.



The Poisson Distribution *ctd*

Distribution functions: $Po(\mu)$ for $\mu = 1, 2, 8, 25$



Poisson Random Variables

- ▶ the number of accidents X at a certain plant in a given year;
- ▶ the number of phone calls arriving at a switchboard in a given hour;
- ▶ the number of defects in a length of rope;
- ▶ the number of particles emitted from a radioactive specimen in a given time.
- ▶ The number of defects on a sheet of zinc plated steel.

Example

Suppose the number of accidents that occur at a particular intersection in a given year has the Poisson distribution with mean 7.2. Find the probability that exactly six accidents will occur in a given year.

Poisson Calculations in Matlab

command	description	example
<code>poisspdf(x, μ)</code>	$Po(\mu)$ pmf at x	<code>poisspdf(4, 2)</code> 0.0902
<code>poisscdf(x, μ)</code>	$Po(\mu)$ cdf at x	<code>poisscdf(4, 2)</code> 0.9473
<code>poissrnd(μ, m, k)</code>	generate $m \times k$ array of rvs from $Po(\mu)$	<code>poissrnd(3, 1, 6)</code> 3 4 0 1 0 1

Continuous Random Variables

The Poisson Process

In queues, arrival times and the number of arrivals are closely related.

A *Poisson process* is a stochastic process in which *point events* occur over time, independently of each other.

For $t > 0$, a Poisson process with rate λ is a collection of random variables

$$\{X(t) : t \geq 0\}$$

which satisfy

- ▶ $X(t)$ is the number of occurrences in $[0, t]$;
- ▶ For $a < b$, $X(b) - X(a)$ is a Poisson random variable with mean $\mu = \lambda(b - a)$.
 - ▶ That is, the number of occurrences in any interval of length t is $Po(\lambda t)$.
 - ▶ Note the units of the rate parameter λ are expected occurrences per unit time.
- ▶ The numbers of occurrences in non-overlapping intervals are independent.

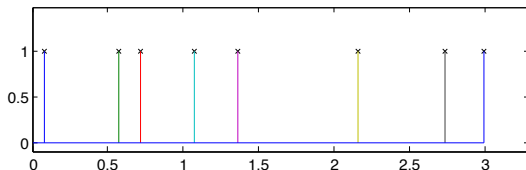
The Poisson Process *ctd*

Messages arrive randomly in a time interval $[0, t]$.

$X(t)$ be the number of messages that arrived up to t .

$T(k)$ is the arrival time of the k th message.

$$X(t) \geq k \iff T(k) \leq t$$



Example

Students queueing up for Matlab help between 12:00noon and 1:00pm on a Wednesday arrive according to a Poisson process with rate $\lambda = 0.5$ arrivals per minute.

1. What is the probability that during a given one-minute period no students arrive?
2. What is the probability that exactly four students arrive in the first five minutes?
3. If three students arrived in the first five minutes, what is the probability that there are no students in the second five minutes?

Solution

The number of students arriving in an interval of k minutes has the Poisson distribution with mean parameter $\mu = \lambda k$.

Continuous Distributions

The Bernoulli, Binomial and Poisson distributions are all *discrete* distributions in the sense that the possible values form a discrete set. In practice, many measurements are *continuous* variables in the sense that the possible values are not restricted to discrete set. For example,

- ▶ A waiting time;
- ▶ The laden mass of vehicle;
- ▶ The maximum power of an aircraft engine.

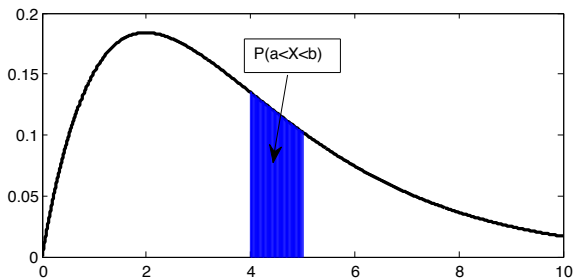
The possible values for a continuous variable are often

- ▶ A finite interval, $\{x : a < x < b\}$;
- ▶ The positive real numbers, $\{x \in \mathbb{R} : x > 0\}$;
- ▶ The set of all real numbers \mathbb{R} .

Continuous Distributions

- ▶ The probability distribution for a continuous random variable X can be specified by its *probability density function* (pdf), f .
- ▶ The probability density function represents probability by **area**.

- ▶ For any numbers $a < b$, $P(a < X \leq b) = \int_a^b f(x)dx$



The cdf

- ▶ The **(cumulative) distribution function (cdf)** F of X is defined, as previously, to be

$$F(x) = P(X \leq x).$$

- ▶ It can be computed as

$$F(x) = \int_{-\infty}^x f(t)dt.$$

- ▶ It follows that $f(x) = \frac{d}{dx}F(x)$.
- ▶ And also, $P(a < X \leq b) = F(b) - F(a)$.

Properties of the pdf

- ▶ In order to be a valid pdf, the function f must satisfy

$$f(x) \geq 0 \text{ for all } x;$$

$$\int_{-\infty}^{\infty} f(x)dx = 1.$$

- ▶ If X is a continuous random variable then for any $a \in \mathbb{R}$

$$P(X = a) = \int_a^a f(x)dx = F(a) - F(a) = 0.$$

It follows that $P(a < X \leq b) = P(a < X < b)$ etc.

The Normal Distribution

The normal distribution $N(\mu, \sigma^2)$ depends on two parameters

$$\mu \quad \text{and} \quad \sigma^2 > 0.$$

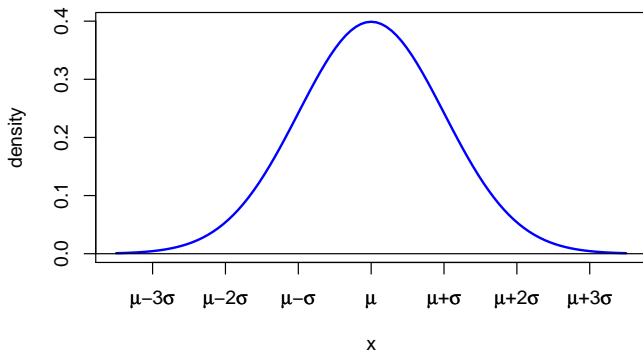
- ▶ It is also called the **Gaussian** distribution.
- ▶ The possible values of a normal rv X are $-\infty < x < \infty$
- ▶ Its pdf is given by

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

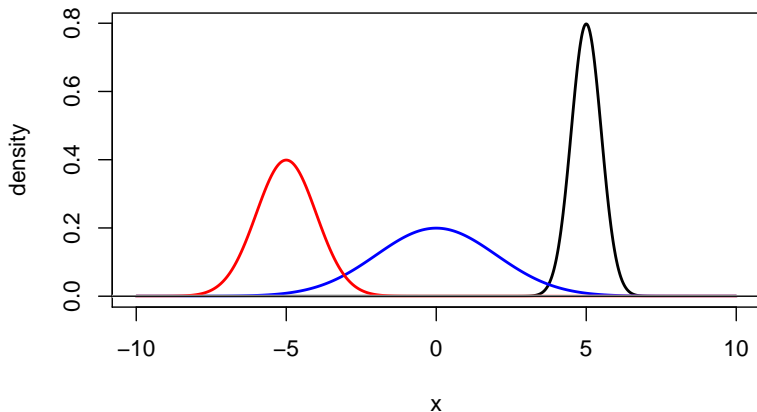
- ▶ μ is called the **mean**, and σ the **standard deviation**.

The Normal Distribution

The normal pdf is a symmetric “bell shaped” curve centred at μ and with spread determined by σ .



Normal pdfs, $N(-5, 1)$, $N(0, 2^2)$, $N(5, 0.5^2)$

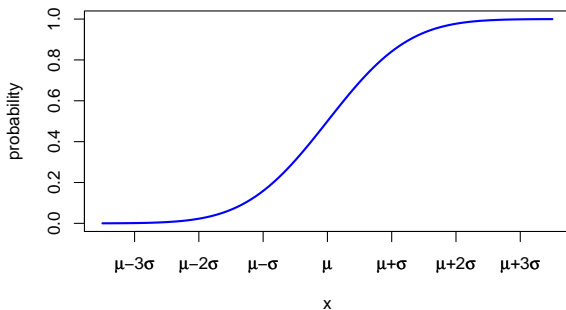


The normal cdf

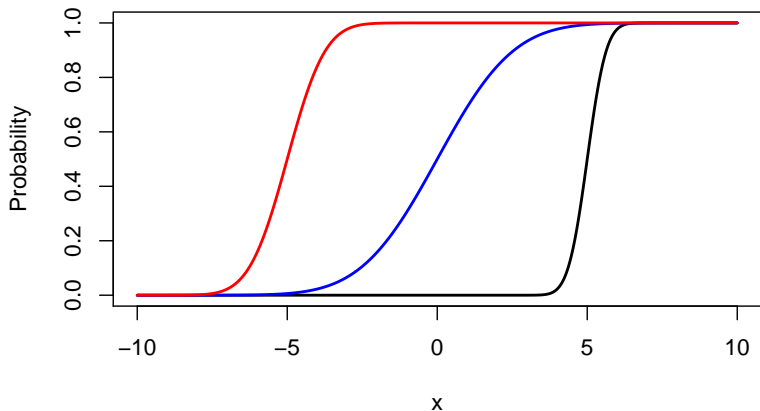
The cdf

$$F(x) = \int_{-\infty}^x \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{1}{2} \frac{(t-\mu)^2}{\sigma^2}} dt$$

cannot be simplified and is evaluated numerically.



Normal cdfs, $N(-5, 1)$, $N(0, 2^2)$, $N(5, 0.5^2)$



The 68-95-99.7 rule

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

$$P(\mu - \sigma < X < \mu + \sigma) = 0.6827 \approx 68\%$$

$$P(\mu - 2\sigma < X < \mu + 2\sigma) = 0.9545 \approx 95\%$$

$$P(\mu - 3\sigma < X < \mu + 3\sigma) = 0.9973 \approx 99.7\%$$

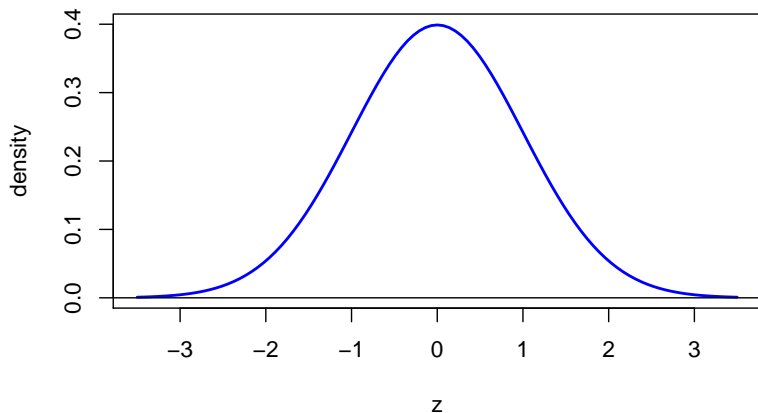
The Standard Normal Distribution

- ▶ The **standard normal** distribution, $N(0, 1)$, has $\mu = 0$ and $\sigma^2 = 1$.
- ▶ It is conventional to use
 - ▶ Z to represent a standard normal variable and
 - ▶ $\phi(z)$ and $\Phi(z)$ to denote its pdf and cdf respectively.

Hence,

$$\phi(z) = \frac{1}{\sqrt{2\pi}} e^{-z^2/2} \quad \text{and} \quad \Phi(z) = \int_{-\infty}^z \frac{1}{\sqrt{2\pi}} e^{-t^2/2} dt$$

The standard normal pdf



Standardisation

- ▶ If $X \sim N(\mu, \sigma^2)$ then the **standardised** rv

$$Z = \frac{X - \mu}{\sigma}$$

has the standard normal distribution, $N(0, 1)$.

- ▶ Conversely, if $Z \sim N(0, 1)$, then

$$X = \mu + \sigma Z \sim N(\mu, \sigma^2).$$

These facts enable the use of the standard normal probabilities for general $N(\mu, \sigma^2)$ distributions.

Matlab commands for the normal distribution

command	description	example
<code>normpdf(x, μ, σ)</code>	$N(\mu, \sigma^2)$ pdf at x	<code>normpdf(-2, -1.5, 0.5)</code> 0.4839
<code>normcdf(x, μ, σ)</code>	$N(\mu, \sigma^2)$ cdf at x	<code>normcdf(-2, -1.5, 0.5)</code> 0.1587
<code>norminv(p, μ, σ)</code>	inverse of $N(\mu, \sigma^2)$ cdf at p	<code>norminv(0.75,-1.5,0.5)</code> -1.628

Example

Arc welds on a steel frame construction have a mean tack weld length of 2.23 mm and a standard deviation of tack length of 0.63 mm. Assume lengths are normally distributed. The specification was that the length should be between 1.5 and 3.0.

1. What proportion of welds will be out of spec?
2. What modified spec could be met 99.8% of the time?
3. What should the mean be adjusted to, and the standard deviation be reduced to, for 99.8% of welds to meet the current spec with a minimum reduction in standard deviation?

Arrival and Waiting Times

- ▶ In queues, we observe the arrival time of messages, customers, *etc*
- ▶ The times are continuous random variables
- ▶ Other arrival or waiting times are
 - ▶ the arrival time of the next J1 bus;
 - ▶ the time until a light bulb stops working;
 - ▶ the survival time of a patient after a stroke.

The Exponential Distribution

The exponential distribution $Exp(\lambda)$ depends on a parameter $\lambda > 0$.

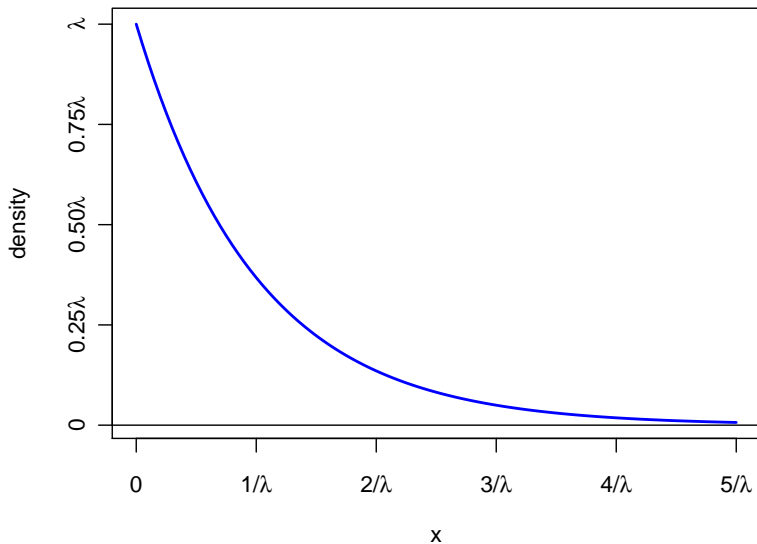
- ▶ The possible values of an exponential rv X are \mathbb{R}^+ .
- ▶ The pdf is given by

$$f(x) = \begin{cases} \lambda e^{-\lambda x} & \text{if } x \geq 0 \\ 0 & \text{if } x < 0. \end{cases}$$

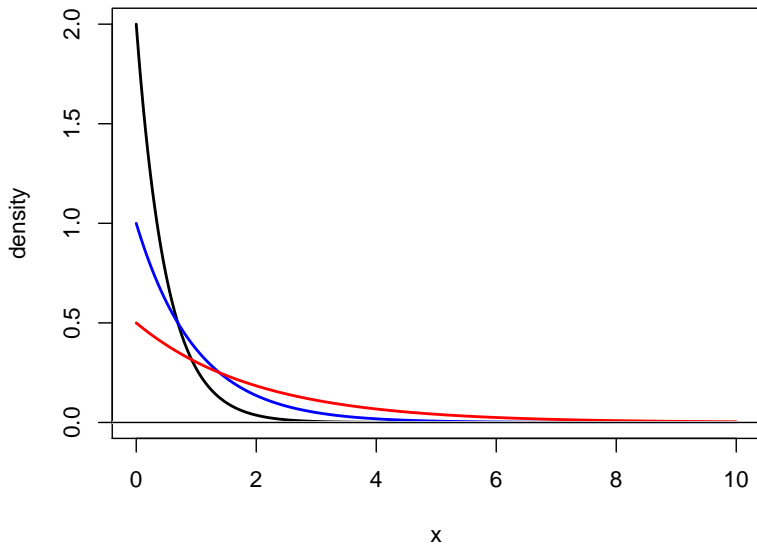
- ▶ The cdf is given by

$$F(x) = \begin{cases} 1 - e^{-\lambda x} & \text{if } x \geq 0 \\ 0 & \text{if } x < 0. \end{cases}$$

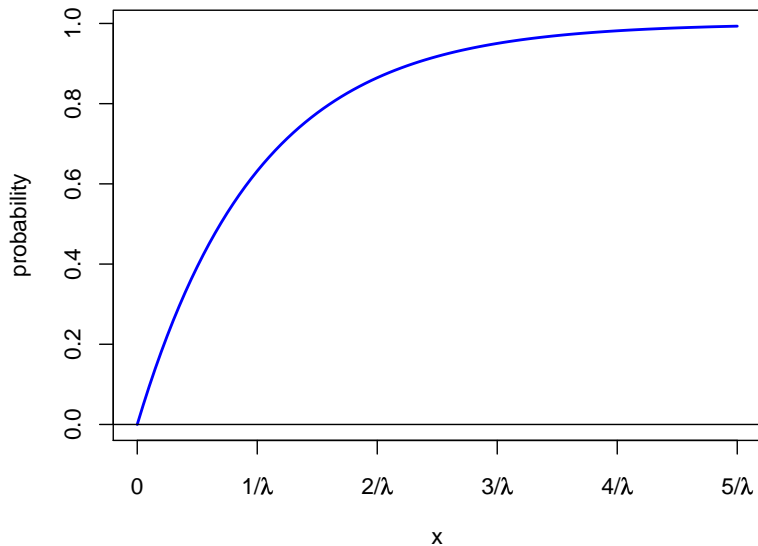
The Exponential pdf



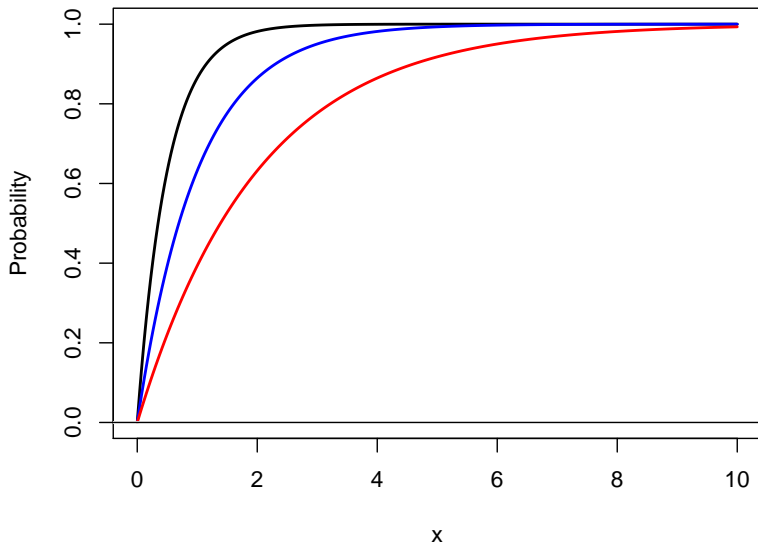
Exponential pdfs for $\lambda = 2$, $\lambda = 1$ and $\lambda = 0.5$



The Exponential cdf



Exponential cdfs for $\lambda = 2$, $\lambda = 1$ and $\lambda = 0.5$



Matlab commands for the exponential distribution

command	description	example
$\text{exppdf}(x, \mu)$	$\text{Exp}(\lambda)$ pdf at x	$f = \text{exppdf}(1.2, 2)$ 0.2744
$\text{expcdf}(x, \mu)$	$\text{Exp}(\lambda)$ cdf at x	$F = \text{expcdf}(1.2, 2)$ 0.4512
$\text{expinv}(p, \mu)$	inverse of $\text{Exp}(\lambda)$ cdf at p	$q = \text{expinv}(0.75, 2)$ 2.7726

!!Note that matlab uses μ as the parameter. This is equal to $1/\lambda$!!

Example

Suppose the lifetime, T , of a light bulb, is exponentially distributed with $\lambda = 1/1000$. Find the probability that a randomly chosen bulb will last for at least 800 hours.

Solution:

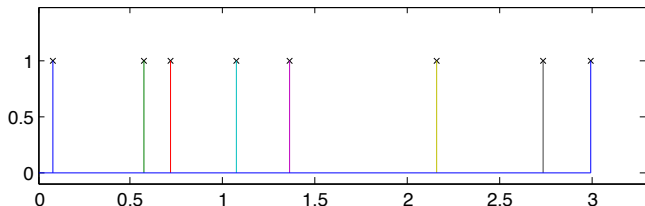
The Exponential Distribution *ctd*

The exponential distribution arises as the waiting time distribution for the Poisson process.

$X(t)$ the number of events that arrived up to time t .

$T(k)$ the time of the k th event.

$$X(t) \geq k \iff T(k) \leq t$$



The Exponential Distribution and the Poisson Process

Consider a Poisson process with rate parameter λ and let the rv T be the time until the first event occurs.

The Uniform Distribution

The uniform distribution $U(a, b)$ depends on the parameters $a < b$.

- ▶ The possible values of a uniform rv X are in the interval $[a, b]$.
- ▶ The pdf is given by

$$f(x) = \begin{cases} 1/(b-a) & \text{if } x \in [a, b] \\ 0 & \text{if } x \notin [a, b]. \end{cases}$$

- ▶ The cdf is given by

$$F(x) = \begin{cases} 0 & \text{if } x < a \\ (x-a)/(b-a) & \text{if } x \in [a, b] \\ 1 & \text{if } x > b \end{cases}$$

Lecture 4

Mean and Variance

Expected Values

- ▶ Let X be a discrete random variable with a probability function p . The **expected value** or **mean** of X is

$$\mu_x = E(X) = \sum x p(x) = \sum x P(X = x)$$

where the summation is over all possible values for X .

- ▶ Let X be a continuous random variable with a pdf f . The **expected value** or **mean** of X is

$$\mu_x = E(X) = \int_{-\infty}^{\infty} x f(x) dx$$

- ▶ The symbol μ or μ_X is also used for the expected value.
- ▶ The expected value is also called the first moment.

Expected Values for Standard Distributions

1. Bernoulli distribution with parameter p : $E(X) = p$
2. Binomial distribution $B(n, p)$: $E(X) = np$
3. Poisson distribution $Po(\mu)$: $E(X) = \mu$
4. Normal distribution $N(\mu, \sigma^2)$: $E(X) = \mu$
5. Exponential distribution $Exp(\lambda)$: $E(X) = 1/\lambda$
6. Uniform distribution $U(a, b)$: $E(X) = (a + b)/2$

Variances I

Let X be a random variable with $E(X) = \mu$. The **variance** is defined by

$$\text{var}(X) = E((X - \mu)^2).$$

- ▶ For a discrete random variable, the variance is obtained as

$$\text{var}(X) = \sum_x (x - \mu)^2 p(x).$$

- ▶ For a continuous random variable, the variance is obtained as

$$\text{var}(X) = \int_{-\infty}^{\infty} (x - \mu)^2 f(x) dx.$$

Variances II

- ▶ The symbol σ^2 or σ_X^2 is often used instead of $\text{var}(X)$.
- ▶ The variance is sometimes called the second central moment.

Standard Deviation

- ▶ The **standard deviation** of a discrete or continuous rv X is

$$\sigma = \sqrt{\text{var}(X)}.$$

- ▶ The variance is a squared quantity and is a mean square error.
- ▶ The standard deviation has the same units as X .

The Bernoulli Variance

Let X be a rv from the Bernoulli distribution with success probability p , so X takes the value 1 with probability p , and the value 0 with probability $1 - p$.

The expected value of X is $E(X) = p$ and the variance can be found directly from the definition.

Means and Variances for some commonly used distributions

Distribution	Mean	Variance
Bernoulli- p	p	$p(1 - p)$
Binomial $B(n, p)$	np	$np(1 - p)$
Poisson $Po(\mu)$	μ	μ
Uniform $U(a, b)$	$\frac{a+b}{2}$	$\frac{(b-a)^2}{12}$
Normal $N(\mu, \sigma^2)$	μ	σ^2
Standard Normal $N(0, 1)$	0	1
Exponential $Exp(\lambda)$	$\frac{1}{\lambda}$	$\frac{1}{\lambda^2}$

Properties of the Mean and Variance

1. $\text{var}(X) \geq 0$ and $\text{var}(X) = 0 \Leftrightarrow X = E(X)$ with probability 1.

2.

$$\text{var}(X) = E(X^2) - [E(X)]^2$$

3. For a random variable X and real constants a, b , let $Y = a + bX$. Then

$$E(Y) = a + bE(X) \text{ and } \text{var}(Y) = b^2\text{var}(X).$$

4. Suppose X is a random variable with $E(X) = \mu$ and $\text{var}(X) = \sigma^2$. For any $k > 0$,

$$P(|X - \mu| \geq k\sigma) \leq \frac{1}{k^2}.$$

Proof of Result 2

Proof of Result 3

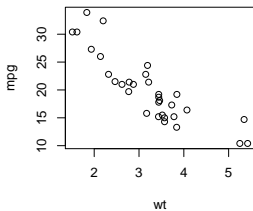
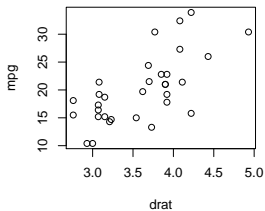
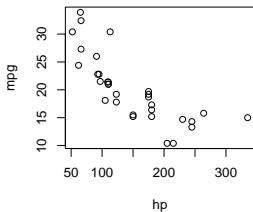
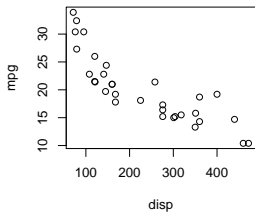
Lecture 5

Covariance, Correlation and Independence

Scatterplots of Two Variables

Many problems involve more than one variable. The scatter plot is a useful way to display variables two at a time.

Examples



The Data

`mpg` miles per gallon

`disp` displacement (cu. in.)

`hp` gross horsepower

`wt` weight (lb/1000)

Sample Mean and Sample Variance

Consider a variable X and a sample of observed values

$$x_1, x_2, \dots, x_n.$$

The **sample mean** is defined by

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

and the **sample variance** is defined by

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

The **sample standard deviation** is defined by $s = \sqrt{s^2}$.

As will be discussed later, \bar{x} is often used to estimate $\mu = E(X)$ and s^2 is often used to estimate $\sigma^2 = \text{var}(X)$.

Covariance and correlation

For bivariate data, the most commonly used measure of association is the *sample correlation coefficient*, r .

The sample correlation is usually defined via an intermediate quantity called the *sample covariance*.

For two variables X and Y , and a sample of observed values,

$$(x_1, y_1), (x_2, y_2), \dots, (x_n, y_n),$$

the sample covariance is

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

The *sample covariance* is positive if there is a *positive association* between the two variables.

- ▶ Large values of Y tend to occur with large values of X ;
- ▶ Small values of Y tend to occur with small values of X ;

It is negative if there is a *negative association* between the two variables.

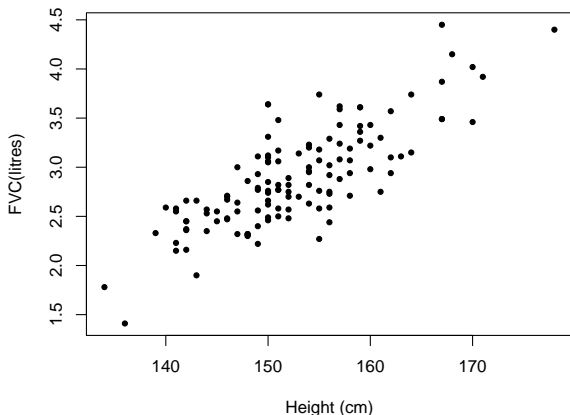
- ▶ Large values of Y tend to occur with small values of X ;
- ▶ Small values of Y tend to occur with large values of X ;

If there is no pattern in the data, the sample covariance will be approximately zero.

Example

Height in cm and FVC in litres were recorded for 127 healthy children.

The sample correlation was found to be $S_{xy} = 3.11$.

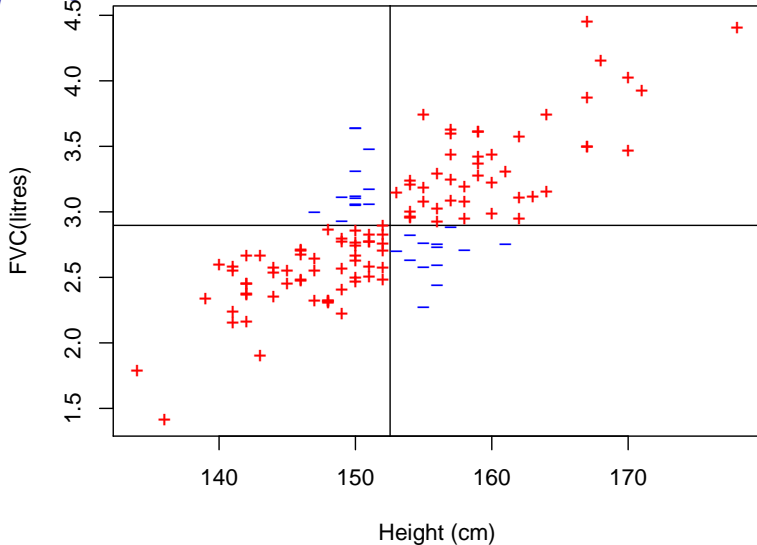


To understand how the sample covariance works, consider a pair of values (x_i, y_i) .

- ▶ If $x_i > \bar{x}$ and $y_i > \bar{y}$ then $(x_i - \bar{x})(y_i - \bar{y}) > 0$ so the contribution to the covariance will be positive.
- ▶ If $x_i < \bar{x}$ and $y_i < \bar{y}$ then $(x_i - \bar{x})(y_i - \bar{y}) > 0$ so the contribution to the covariance will be positive.
- ▶ If $x_i > \bar{x}$ and $y_i < \bar{y}$ then $(x_i - \bar{x})(y_i - \bar{y}) < 0$ so the contribution to the covariance will be negative.
- ▶ If $x_i < \bar{x}$ and $y_i > \bar{y}$ then $(x_i - \bar{x})(y_i - \bar{y}) < 0$ so the contribution to the covariance will be negative.

FVC Data: contribution to the covariance

$$s_{xy} = 3.11$$



- ▶ For scatter plots with positive association, most of the contributions will be positive. These will outweigh the negative contributions and the sample covariance will be positive.
- ▶ For scatter plots with negative association, most of the contributions will be negative. These will outweigh the positive contributions and the sample covariance will be negative.
- ▶ When there is no pattern, the positive and negative contributions will be similar in number and magnitude and will approximately cancel.

The sample correlation

The preceding logic would suggest that the larger the covariance, the stronger the association.

However, this is not true. Consider the FVC data. The sample covariance is $s_{XY} = 3.11$.

The heights are measured in cms. If we had decided to measure the heights in inches instead, every measurement would be divided by 2.54.

The sample covariance would also be divided by 2.54 and would become 1.23.

However, the association between Height and FVC would be unchanged.

The sample correlation coefficient is defined by

$$r = \frac{s_{xy}}{s_x s_y}$$

where

- ▶ s_{xy} is the sample covariance;
- ▶ s_x is the sample standard deviation of x ;
- ▶ s_y is the sample standard deviation of y ;

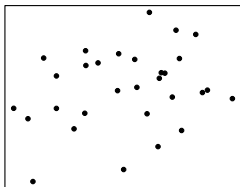
The sample correlation between Height and FVC is 0.79.

Properties of the sample correlation

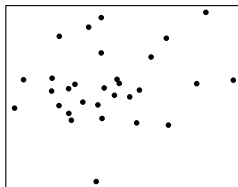
- ▶ $-1 \leq r \leq 1$;
- ▶ $r = 1$ if and only if all points lie on a straight line with positive slope;
- ▶ $r = -1$ if and only if all points lie on a straight line with negative slope;
- ▶ If the scatter plot has no pattern then $r \approx 0$.
 - ▶ Note that the converse is not true.
 - ▶ That is we may have a $r \approx 0$ but a strong *non-linear pattern may be present*.
- ▶ If all of the x (or y) values are multiplied by a positive constant, r is not changed.
- ▶ If all of the x (or y) values are multiplied by a negative constant, the sign of r is reversed.

Illustrative scatter plots and sample correlations

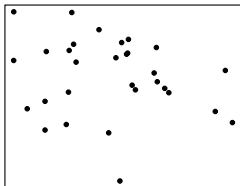
$r=0$



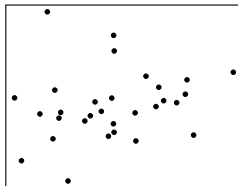
$r=0.1$



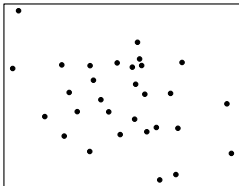
$r=-0.2$



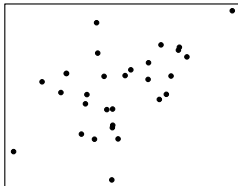
$r=0.3$



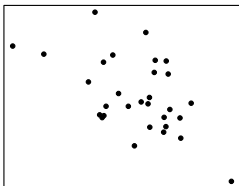
$r=-0.4$



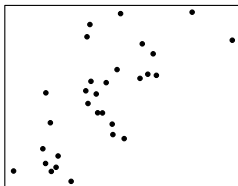
$r=0.5$



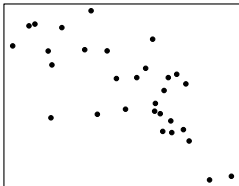
$r=-0.6$



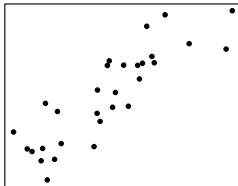
$r=0.7$



$r=-0.8$



$r=0.9$



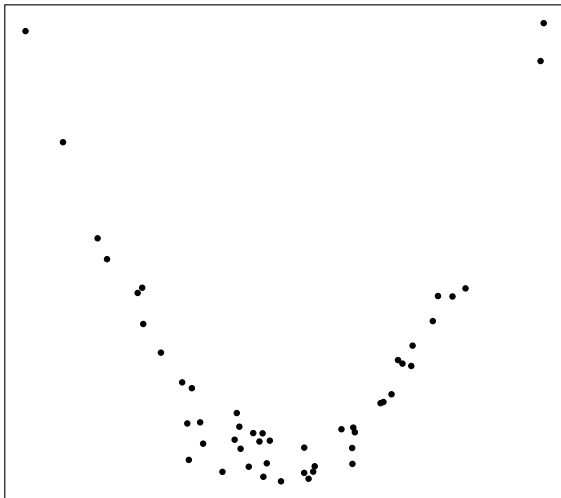
$r=-0.95$



$r=0.99$



$r=0.032$



$r=0.9$



The sample correlation coefficient is a measure of the strength of *linear association* in the data.

That is, the tendency of the points to cluster about a straight line on the scatter plot.

It is not sensitive to non-linear patterns of association.

Even when there is a high correlation, for example $r = 0.9$ we cannot say that there is strong linear association in the data.

The correct conclusion is that there is a strong linear component of association. There may also be a non-linear component.

Covariance and Correlation for Random Variables

Consider two random variables, X and Y with

$$E(X) = \mu_X, \quad E(Y) = \mu_Y, \quad \text{var}(X) = \sigma_X^2 \quad \text{and} \quad \text{var}(Y) = \sigma_Y^2.$$

The covariance and correlation for are defined, in analogy to the sample covariance and sample correlation, respectively by

$$\sigma_{XY} = \text{cov}(X, Y) = E((X - \mu_X)(Y - \mu_Y))$$

and

$$\rho = \frac{\sigma_{XY}}{\sigma_X \sigma_Y}.$$

A detailed discussion of the covariance and correlation requires the notion of bivariate distributions, the details of which are beyond the scope of this course.

It can be shown that the correlation coefficient ρ possesses properties analogous to the sample correlation r .

- ▶ $-1 \leq \rho \leq 1$;
- ▶ $\rho = 1 \Leftrightarrow Y = a + bX$ with probability 1, for some constants a and b with $b > 0$.
- ▶ $\rho = -1 \Leftrightarrow Y = a + bX$ with probability 1, for some constants a and b with $b < 0$.

Independent Random Variables

Recall that two events A and B are defined to be independent if

$$P(A \cap B) = P(A)P(B).$$

The definition can be extended to provide a definition for the independence of two random variables.

The two random variables, X and Y are said to be independent if for all real numbers x and y ,

$$P(\{X \leq x\} \cap \{Y \leq y\}) = P(X \leq x)P(Y \leq y).$$

This definition applies to any combination of discrete and continuous random variables.

Variables that arise from unrelated experiments are often assumed to be independent.

- ▶ A sample of pulverised ore is divided in two and assays for lead concentration, X and Y are performed at two different laboratories.
- ▶ The number of people X attempting a driving test at a particular centre in one day, and the number of people Y who successfully pass their test on that day.

Uncorrelated Random Variables

Random variables, X and Y are said to be uncorrelated if $\rho = 0$ or, equivalently, $\sigma_{XY} = 0$.

If the random variables X and Y are independent, then they must also be uncorrelated.

However, the converse is **not** true. When $\rho = 0$, we cannot deduce that X and Y are independent.

Linear Combinations of Random Variables

- ▶ Let X and Y be random variables with means μ_X and μ_Y , variances σ_X^2 and σ_Y^2 and covariance σ_{XY} .
- ▶ For constants a and b we define

$$W = aX + bY.$$

- ▶ W is a random variable with
 - ▶ mean

$$\mu_w = E(W) = a\mu_X + b\mu_Y,$$

- ▶ and variance

$$\begin{aligned}\sigma_w^2 &= \text{var}(W) = a^2\sigma_X^2 + b^2\sigma_Y^2 + 2ab\sigma_{XY} \\ &= a^2\sigma_X^2 + b^2\sigma_Y^2 + 2ab\rho\sigma_X\sigma_Y\end{aligned}$$

Outline derivation

Special Cases

- ▶ Let X and Y be random variables with means μ_X and μ_Y , and variances σ_X^2 and σ_Y^2 . Put

$$W = aX + bY$$

for some constants a and b .

- ▶ If X and Y are uncorrelated, then

$$\text{var}(W) = a^2\sigma_X^2 + b^2\sigma_Y^2.$$

- ▶ If X and Y are independent and normally distributed, so $X \sim N(\mu_X, \sigma_X^2)$ and $Y \sim N(\mu_Y, \sigma_Y^2)$, then

$$W \sim N(\mu_W, \sigma_W^2),$$

where $\mu_W = a\mu_X + b\mu_Y$ and $\sigma_W^2 = a^2\sigma_X^2 + b^2\sigma_Y^2$.

Example I

A certain orebody contains both silver and lead.

For a randomly chosen 1000kg section of ore, let X be the amount of lead in kilograms and Y the amount of silver in grams.

Assume that

- ▶ X has mean $\mu_X = 59\text{kg}$ and standard deviation $\sigma_X = 18\text{kg}$;
- ▶ Y has mean $\mu_Y = 850\text{g}$ and standard deviation $\sigma_Y = 270\text{g}$;
- ▶ The correlation is $\rho = 0.72$.

Example II

Suppose the value of lead is \$2.30 per kilogram and the value of silver is \$0.81 per gram.

Let V denote the total value of the lead and silver in a randomly chosen 1000kg section of ore. Find μ_V and σ_V .

Solution

Lecture 6

Sums and Averages of Random Variables

Sums of independent random variables I

Consider random variables

$$X_1, X_2, \dots, X_n$$

such that

- ▶ $E(X_i) = \mu$ for $i = 1, 2, \dots, n$;
- ▶ $\text{var}(X_i) = \sigma^2$ for $i = 1, 2, \dots, n$;
- ▶ The variables are **mutually independent**.

Sums of independent random variables II

Let

$$S = \sum_{i=1}^n X_i = X_1 + X_2 + \dots + X_n.$$

Then

$$E(S) = n\mu \text{ and } \text{var}(S) = n\sigma^2.$$

Remark This fact is a simple extension of the results for linear combinations of two variables.

The sample mean

For X_1, X_2, \dots, X_n and S as defined, let

$$\bar{X} = \frac{1}{n} \sum_{i=1}^n X_i = \frac{1}{n} S.$$

- ▶ \bar{X} is called the sample mean;
- ▶ \bar{X} should be considered as a random variable;
- ▶ What is $E(\bar{X})$ and $\text{var}(\bar{X})$?

The sample mean as an estimate for μ

Example

- ▶ Thickness in angstroms was measured for randomly chosen 23 silicon wafers after polishing.

3240	3200	3220	3210	3250	3220
3190	3190	3150	3160	3270	3180
3200	3270	3180	3300	3250	3330
3300	3280	3270	3270	3200	

- ▶ The sample mean \bar{x} was found to be 3231.7 angstroms.
- ▶ The value of \bar{x} can be considered as a summary description of these 23 numbers.
- ▶ However it can also be considered as an **estimate** of the **mean** μ for **all** wafers produced by this process.

The sample mean as an estimate for μ

The justification for this use of the sample mean follows from the properties of \bar{X} . Namely

$$E(\bar{X}) = \mu \text{ and } \text{var}(\bar{X}) = \frac{\sigma^2}{n}.$$

- ▶ We think of an observed value \bar{x} (e.g. 3231.7) as a realization of the random variable \bar{X} .
- ▶ The fact that $E(\bar{X}) = \mu$ means that the probability distribution of \bar{X} is centred at μ .
- ▶ The fact that $\text{var}(\bar{X}) = \sigma^2/n$ justifies the common intuition that larger samples are more reliable.

Contexts I

- ▶ μ could be a **process mean** as in the silicon wafer example.
 - ▶ The data arise by sampling products produced by the process.
- ▶ μ could be the mean of a finite population.
 - ▶ For example, the population of students at the University of Adelaide.
 - ▶ The data arise by collecting a random sample of students and recording the variable of interest.
- ▶ μ could be an unknown constant.
 - ▶ For example, the physicist Simon Newcomb made 66 determinations of the speed of light.

Contexts II

- ▶ A simple model for measurement error is that

$$X_i = \mu + \mathcal{E}_i \text{ where } E(\mathcal{E}_i) = 0 \text{ and } \text{var}(\mathcal{E}_i) = \sigma^2.$$

- ▶ In this case μ would represent the exact speed of light and X_i would be a single measurement.

Sums and averages of normal variables

Suppose now that

$$X_1, X_2, \dots, X_n$$

are independent random variables such that

$$X_i \sim N(\mu, \sigma^2).$$

If $S = \sum_{i=1}^n X_i$ then

$$S \sim N(n\mu, n\sigma^2).$$

If $\bar{X} = \frac{1}{n} \sum_{i=1}^n X_i$ then

$$\bar{X} \sim N\left(\mu, \frac{\sigma^2}{n}\right).$$

Example

A helicopter operating in Bass Strait can carry ten passengers. Suppose passenger weights, X are normally distributed $X \sim N(80, 14^2)$. Find the probability that:

1. The weight of a randomly selected passenger exceeds 90kg;
2. The total weight of 10 randomly selected passengers exceeds 900kg;
3. The average weight of 10 randomly selected passengers exceeds 90kg.

The central limit theorem

Suppose X_1, X_2, \dots, X_n are independent random variables with

$$E(X) = \mu \text{ and } \text{var}(X) = \sigma^2$$

but are not assumed to normally distributed.

Let

$$S = \sum_{i=1}^n X_i \text{ and } \bar{X} = \frac{1}{n} \sum_{i=1}^n X_i.$$

Then for large n , the approximate distributions apply:

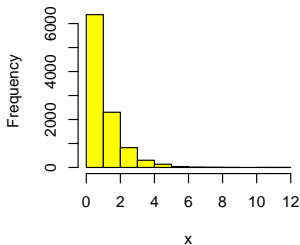
$$S \approx N(n\mu, n\sigma^2)$$

and

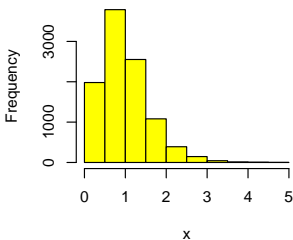
$$\bar{X} \approx N(\mu, \sigma^2/n)$$

Example: sample means of exponential variables

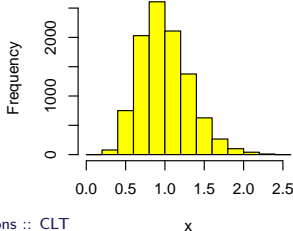
n=1



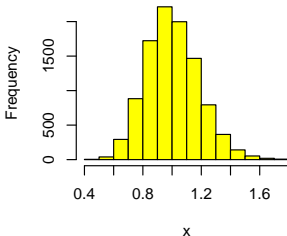
n=3



n=10



n=30



Example I

Suppose the weight, W , of a randomly chosen bag of cement is uniformly distributed $W \sim U(19.8, 20.5)$. Find an approximate value for the probability that the total weight of 50 randomly chosen bags is no more than 1005kg.

Estimation and confidence I

Consider data

$$x_1, x_2, \dots, x_n$$

assumed to be observations of independent random variables

$$X_1, X_2, \dots, X_n$$

such that $X_i \sim N(\mu, \sigma^2)$ for $i = 1, 2, \dots, n$.

It is reasonable to use \bar{x} as an estimate of μ .

A key point is that we can use random variable theory to quantify the accuracy of the estimate.

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

Estimation and confidence II

$$\begin{aligned} & P(\mu - 1.96\sigma/\sqrt{n} < \bar{X} < \mu + 1.96\sigma/\sqrt{n}) \\ = & P\left(-1.96 < \frac{\bar{X} - \mu}{\sigma/\sqrt{n}} < 1.96\right) \\ = & P(-1.96 < Z < 1.96) \\ = & 0.95 \end{aligned}$$

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

In other words, there is a 95% chance that $|\bar{X} - \mu| < 1.96\sigma/\sqrt{n}$.

Estimation and confidence III

Equivalently, there is a 95% chance that the random interval

$$(\bar{X} - 1.96\sigma/\sqrt{n}, \bar{X} + 1.96\sigma/\sqrt{n})$$

will contain the target quantity μ .

The 95% confidence interval for μ

In practice, we can calculate the interval

$$(\bar{x} - 1.96\sigma/\sqrt{n}, \bar{x} + 1.96\sigma/\sqrt{n})$$

using the observed value of the sample mean \bar{x} provided σ is known.

This interval is called a **95% confidence interval for μ** .

Example (continued)

- ▶ Recall in the silicon wafer example, we found the $\bar{x} = 3231.7$ from a random sample of $n = 23$ wafers.
- ▶ Suppose now that the thickness of wafers produced by the process is normally distributed and that the standard deviation is known to be $\sigma = 50$ angstroms.
- ▶ The 95% confidence interval for μ is thus

$$(3231.7 - 1.96 \times 50/\sqrt{23}, 3231.7 + 1.96 \times 50/\sqrt{23})$$

which gives $(3211.3, 3252.1)$

- ▶ In other words, we can be **95% confident** that the mean thickness of wafers produced by the process lies between 3211.3 and 3252.1 angstroms.

Remarks

- ▶ Although confidence is not, strictly speaking, the same thing as probability, the practical interpretation is similar.

Someone who always assumes that the 95% confidence interval contains μ will, in the long run, be correct for 95% of intervals.

- ▶ The quantity σ/\sqrt{n} is called the **standard error (SE)** (of \bar{x}).
- ▶ The form of the confidence interval is

$$\text{Estimate} \pm z^* \text{SE}$$

where

- ▶ \bar{x} is the estimate;
- ▶ $z^* = 1.96$;
- ▶ $\text{SE} = \sigma/\sqrt{n}$.

Level of confidence

- ▶ The value $z^* = 1.96$ was chosen to give 95% confidence.
 - ▶ The key property is $P(-1.96 < Z < 1.96) = 0.95$ for $Z \sim N(0, 1)$.
 - ▶ We can calculate this value using `norminv(0.975,0,1)`.
- ▶ We can also obtain different levels of confidence by changing z^* .
 - ▶ For example, $z^* = \text{norminv}(0.95,0,1) = 1.645$ would give a 90% confidence interval

$$\bar{x} \pm 1.645\sigma/\sqrt{n}.$$

- ▶ In general:
 - ▶ Larger $z^* \Rightarrow$ greater confidence but also wider interval;
 - ▶ Smaller $z^* \Rightarrow$ narrower interval but also lower confidence;

Confidence intervals for μ with σ unknown

- ▶ The preceding confidence interval for μ requires that σ is known.
- ▶ In most practical circumstances σ will not be known.
- ▶ In this case we can estimate σ by the sample standard deviation to obtain an estimated standard error

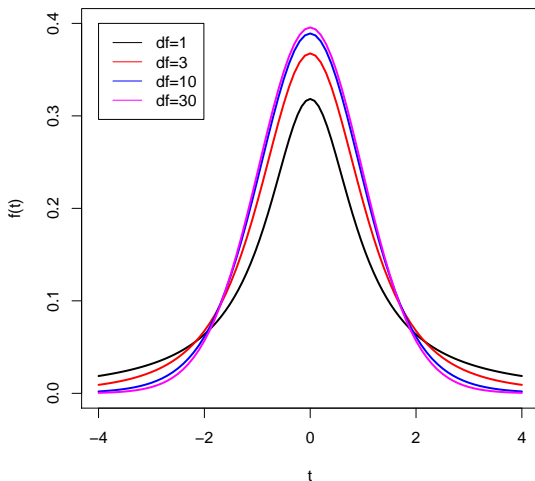
$$SE = \frac{s}{\sqrt{n}}.$$

- ▶ The value $z^* = 1.96$ does not result in 95% confidence when the estimated SE is used.
- ▶ Instead, the value t^* obtained from a different distribution, called the “t-distribution with $n-1$ ” degrees of freedom is used.
 - ▶ That is, we find t^* for which

$$P(T \leq t^*) = 0.975$$

The t-distribution

The t-distribution is a family of continuous distributions depending on a parameter k called the degrees of freedom.



- ▶ If X_1, X_2, \dots, X_n are independent $N(\mu, \sigma^2)$ random variables, it can be shown (mathematically) that

$$T = \frac{\bar{X} - \mu}{s/\sqrt{n}} \sim t_{n-1}.$$

- ▶ In comparison, for σ known, we have

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

- ▶ Matlab Commands

command	description	example
$\text{tpdf}(x, k)$	t_k pdf at x	$f = \text{tpdf}(0, 3)=0.3676$
$\text{tcdf}(x, k)$	t_k cdf at x	$F = \text{tcdf}(2, 3)=0.9303$
$\text{tinv}(p, x)$	inverse of t_k cdf at p	$q = \text{tinv}(0.75, 3)=0.7649$

Example continued

- ▶ In the silicon wafer example, the sample mean thickness is $\bar{x} = 3231.7$ angstroms and the sample standard deviation is $s = 48.77$.
- ▶ The (estimated) standard error is

$$SE = \frac{48.77}{\sqrt{23}} = 10.17.$$

- ▶ The degrees of freedom for the t-distribution are $n - 1 = 22$ and

$$t^* = \text{tinv}(0.975, 22) = 2.0739.$$

- ▶ The 95% confidence interval is $3231.7 \pm 2.0739 \times 10.17$ or, equivalently

$$(3210.649, 3252.829)$$

Confidence Intervals and Hypothesis Tests

Summary of confidence intervals so far... I

- ▶ Confidence intervals are used to allow for the error that arises when a sample mean is used to estimate the corresponding mean of a distribution.
- ▶ The general form for the confidence intervals we consider is

$$\text{Estimate} \pm k \text{ SE}$$

where k is a constant chosen to give the required level of confidence.

- ▶ For independent $N(\mu, \sigma^2)$ observations x_1, x_2, \dots, x_n with σ^2 known, we use

$$\bar{x} \pm z^* \sigma / \sqrt{n}$$

where z^* is obtained from the $N(0, 1)$ distribution.

Summary of confidence intervals so far... II

- ▶ For independent $N(\mu, \sigma^2)$ observations x_1, x_2, \dots, x_n with σ^2 not known, we use

$$\bar{x} \pm t^* s / \sqrt{n}$$

where t^* is obtained from the t_{n-1} distribution.

Hypothesis tests

- ▶ Consider independent $N(\mu, \sigma^2)$ observations x_1, x_2, \dots, x_n .
- ▶ Sometimes it is of interest to ask whether the mean μ takes a particular value.

Example

- ▶ A water-ethanol distillation column produces a mean yield of 93% when it is operating correctly.
- ▶ A random sample of 8 recent batches showed yields of

0.90, 0.93, 0.95, 0.86, 0.90, 0.87, 0.93, 0.92

- ▶ Based on this data, we want to determine whether the column is operating correctly.
- ▶ If we assume the 8 observations to be independent $N(\mu, \sigma^2)$ observations, the question becomes:

Is $\mu = 0.93$?

The null and alternative hypotheses I

- ▶ Consider independent $N(\mu, \sigma^2)$ observations x_1, x_2, \dots, x_n .
- ▶ We consider the **null hypothesis**,

$$H_0 : \mu = \mu_0$$

where

- ▶ μ is the unknown mean parameter;
- ▶ μ_0 is the particular value of interest to us;
- ▶ For example, for the ethanol column we can pose the question of interest as the hypothesis

$$H_0 : \mu = 0.93.$$

The null and alternative hypotheses II

- ▶ The **alternative hypothesis** is, in general, the statement that the null is not true. In this case,

$$H_A : \mu \neq \mu_0.$$

- ▶ It is also possible to consider “one-sided” hypotheses such as $H_0 : \mu \leq \mu_0$ vs $H_A : \mu > \mu_0$ but this is beyond the scope of our course.

The hypothesis test I

- ▶ A hypothesis test is a procedure that uses the data x_1, x_2, \dots, x_n to reach one of the two conclusions:
 - Accept H_0 The null hypothesis is a plausible model for the observed data;
 - Reject H_0 The null hypothesis is not a plausible model for the observed data.
- ▶ For $H_0 : \mu = \mu_0$, recall that \bar{x} is our estimate of μ .
 - ▶ The basic idea is that we should accept H_0 if \bar{x} is close to μ_0 and reject otherwise.
 - ▶ To define what we mean by “close to μ_0 ” recall that the **precision** of the estimate \bar{x} is measured by the **standard error**

The hypothesis test II

- ▶ Therefore, “close” should be measured in terms of the standard error.

The test statistic I

- ▶ Consider independent $N(\mu, \sigma^2)$ observations x_1, x_2, \dots, x_n and the hypotheses $H_0 : \mu = \mu_0$ vs $H_A : \mu \neq \mu_0$.
- ▶ For σ^2 not known, the test of H_0 is based on the **test statistic**

$$t = \frac{\bar{x} - \mu_0}{s/\sqrt{n}}.$$

- ▶ The generic form of this test statistic is

$$\text{Test Statistic} = \frac{\text{Estimate} - \text{Null Value}}{\text{Standard Error}}$$

- ▶ The decision rule is of the form
 - ▶ Reject H_0 for large values of $|t|$;

The test statistic II

- ▶ Accept H_0 otherwise.
- ▶ To determine how large $|t|$ should be for us to reject H_0 we calculate a quantity called the P-value.

The P-value I

- ▶ Consider the test of $H_0 : \mu = \mu_0$ vs $H_A : \mu \neq \mu_0$ using the test statistic

$$t = \frac{\bar{x} - \mu_0}{s/\sqrt{n}}.$$

- ▶ If H_0 is true, then the observed value of t can be considered as an observation from the t_{n-1} distribution.
- ▶ Let $T \sim t_{n-1}$ denote a random variable from the t_{n-1} distribution.
- ▶ The P-value is

$$\text{P-value} = P(|T| \geq |t|) = 2P(T \geq |t|)$$

where $|t|$ is the observed value of the test statistic.

The P-value II

- ▶ To test H_0 at the **5% level of significance**

Reject H_0 for P-value ≤ 0.05 ;

Accept H_0 for P-value > 0.05 ;

Example: The ethanol data I

- ▶ The data are

0.90, 0.93, 0.95, 0.86, 0.90, 0.87, 0.93, 0.92

and the null hypothesis is

$$H_0 : \mu = 0.93.$$

- ▶ The sample mean and sample standard deviation for these data are

$$\bar{x} = 0.9075 \text{ and } s = 0.03105.$$

- ▶ The standard error for \bar{x} is

$$s/\sqrt{n} = 0.03105/\sqrt{8} = 0.01098.$$

Example: The ethanol data II

- ▶ The test statistic is

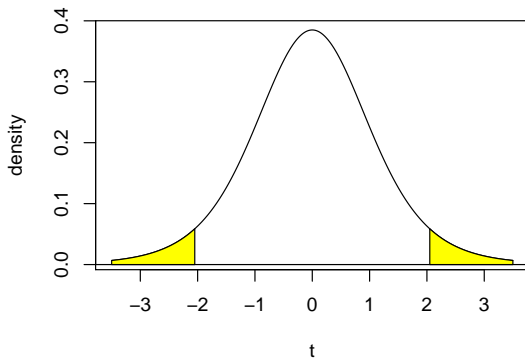
$$t = \frac{\bar{x} - \mu_0}{s/\sqrt{n}} = -2.0494.$$

Example (continued) I

- ▶ The degrees of freedom are $n - 1 = 7$.
- ▶ The P-value is

$$\begin{aligned}\text{P-value} &= P(T \geq |t|) = 2P(T \geq 2.0494) \\ &= 2*(1-\text{tcdf}(2.0494, 7)) = 0.0796.\end{aligned}$$

Example (continued) II



- ▶ Since $P\text{-value} > 0.05$, H_0 is accepted.

The logic of accept and reject I

- ▶ Large values of $|t|$ represent evidence against H_0 .
- ▶ The P-value is the probability we would observe a value of $|t|$ as large (or larger) than that actually observed if H_0 were true.
 - ▶ A small P-value means that the observed data would be unlikely to have happened if H_0 were true.
This is the reason for rejecting.
 - ▶ A larger P-value means that the observed data could easily have happened if H_0 were true.
This means there is no reason to reject so we accept H_0 .
- ▶ The widely accepted convention to reject H_0 for P-value ≤ 0.05 and accept H_0 for P-value > 0.05 .

The logic of accept and reject II

- ▶ More generally, we could fix a different threshold $0 < \alpha < 1$ with the rule reject H_0 for P-value $\leq \alpha$ and accept H_0 for P-value $> \alpha$.
- ▶ α is called the significance level of the test and is usually expressed as a percentage.
 - ▶ It also useful to present the P-value as a measure of the **strength of evidence**

Remarks about test and P-values I

- ▶ Rejecting H_0 means we have strong evidence that it is not true.
- ▶ Accepting H_0 means that we don't have strong evidence that it is not true.
 - ▶ This is not the same thing as having strong evidence that H_0 is true;
 - ▶ Even if the P-value is close to 1!
- ▶ Small P-values mean we have strong evidence that H_0 is not true.
 - ▶ In this case, the observed discrepancy is said to be **statistically significant**.

Remarks about test and P-values II

- ▶ Whether or not a statistically significant effect is of practical importance is a separate question.
- ▶ Statistical significance \neq practical importance.

Hypothesis tests and confidence intervals I

- ▶ For the situations considered in this course, there is a close relationship between hypothesis tests and confidence intervals.
- ▶ Let (ℓ, u) be the 95% confidence interval for μ and consider a test of $H_0 : \mu = \mu_0$.
- ▶ The following statements are equivalent.
 - ▶ H_0 is accepted at the 5% level of significance;
 - ▶ The null value μ_0 is contained in the 95% confidence interval. That is, $\ell < \mu_0 < u$.
- ▶ **Example** For the alcohol distillation example, recall that $\bar{x} = 0.9075$, $SE = 0.01098$ and from Matlab $t^* = \text{tinv}(0.975, 7) = 2.3646$.

Hypothesis tests and confidence intervals II

- ▶ The 95% confidence interval for μ is $\bar{x} \pm t^* \text{SE}$ which gives (0.8815, 0.9335).
- ▶ The null value $\mu_0 = 0.93$ is contained in the interval.
- ▶ When the test was performed, H_0 was accepted.

Hypothesis tests and confidence intervals in Matlab I

Matlab provides a function called `ttest` that performs hypothesis tests and confidence interval calculations for a single normal sample.

```
>> x=[0.90, 0.93, 0.95, 0.86, 0.90, 0.87, 0.93, 0.92]
x =
    0.9000    0.9300    0.9500    0.8600    0.9000    0.8700
>> [h,p,ci,stats]=ttest(x,0.93)
h =
    0

p =
    0.0796

ci =
    0.8815    0.9335

stats =
    tstat: -2.0494
        df: 7
        sd: 0.0311
>>
```

Matlab

- ▶ The default significance level is 5% and the default confidence level is 95%.
- ▶ `h` is an indicator:

$$h = \begin{cases} 0 & \text{if } H_0 \text{ is accepted} \\ 1 & \text{if } H_0 \text{ is rejected} \end{cases}$$

- ▶ `p` is the P-value.
- ▶ `ci` is vector containing the endpoints of the 95% confidence interval for μ .
- ▶ `stats` is vector containing, the t-statistic, the degrees of freedom and the sample standard deviation of the data.

Assumptions

- ▶ The t-test and confidence interval for a single normal mean begin with the assumption that

$$x_1, x_2, \dots, x_n$$

are independent $N(\mu, \sigma^2)$ observations.

- ▶ That is, the data are observations of independent $N(\mu, \sigma^2)$ random variables

$$X_1, X_2, \dots, X_n.$$

- ▶ This involves two assumptions
 - ▶ The observations are independent;
 - ▶ The observations are normally distributed.

Independence I

- ▶ Independence is the more important of the two assumptions but it is also the most difficult to check for.
- ▶ Sometimes we can tell from the circumstances that independence is not feasible.
 - ▶ For example, 20 samples of a certain polymer product were collected and the viscosity was measured for each sample.
 - ▶ If the 20 samples had been collected from 4 different batches of the product with 5 samples per batch the independence assumption would be violated.
 - ▶ We would expect measurements from the same batch to be more similar than measurements from different batches.

Independence II

- ▶ Sometimes we can detect violations of independence in the data.
 - ▶ For example, suppose the data x_1, x_2, \dots, x_n were recorded serially.
 - ▶ That is, x_1 was recorded first, x_2 was recorded second and so on.
 - ▶ In such cases we can check for dependence by plotting X against time.
 - ▶ Or by correlating x_1, x_2, \dots, x_{n-1} with x_2, x_3, \dots, x_n .

Independence III

- ▶ If we just obtain set of numbers

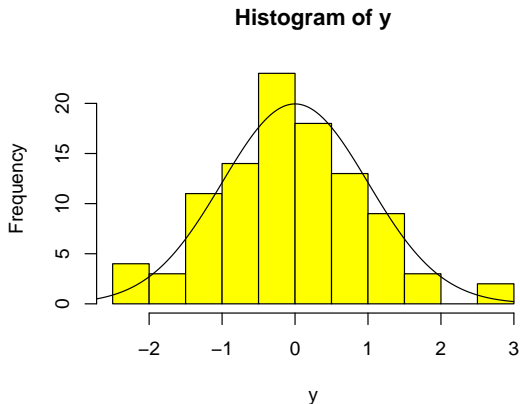
$$x_1, x_2, \dots, x_n$$

without knowing the context it is impossible to determine whether or not independence is a reasonable assumption.

- ▶ In practice, it is important to get as much information about how the data were collected as possible.
- ▶ If you suspect that there are problems with the assumption of independence, then you may need to get advice from a statistician about how to analyse the data.

Normality I

- ▶ The simplest way to check whether the data are normally distributed is to examine a histogram of the data.



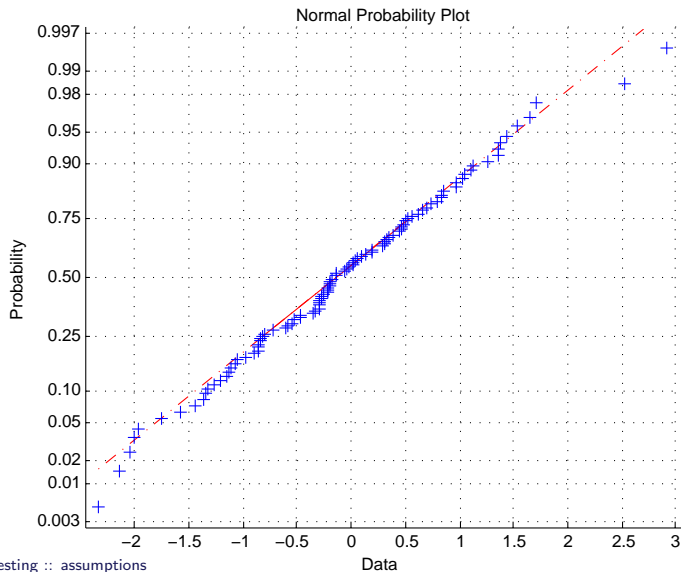
Normality II

- ▶ If the histogram appears “roughly normal” then the assumption is reasonable.

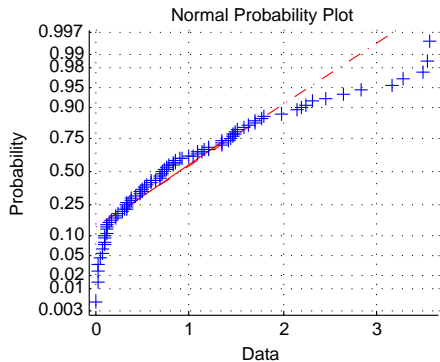
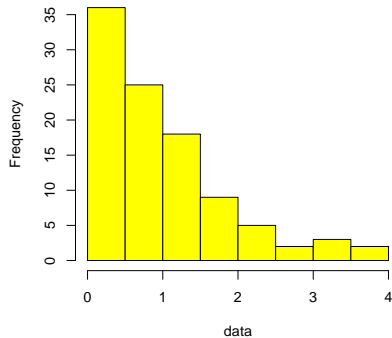
The normal probability plot

- ▶ It is easier to check for normality using a specialised plot called the normal probability plot.
- ▶ The criterion is that the points should roughly follow a straight line.
 - ▶ When the data are normal the line is generally straightest near the middle of the plot.
 - ▶ It will also be “more wiggly” at the ends.
 - ▶ For larger samples, less wiggle is acceptable.
 - ▶ It takes some experience to distinguish between acceptable variability and serious departures from normality.
- ▶ Normal probability plots can be obtained in Matlab using the `normplot` command.

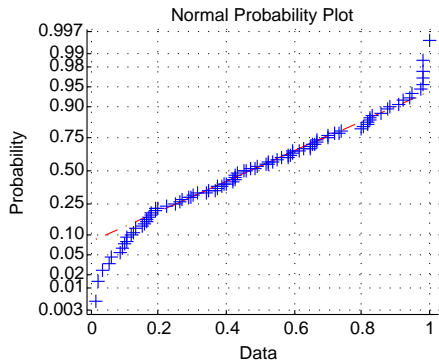
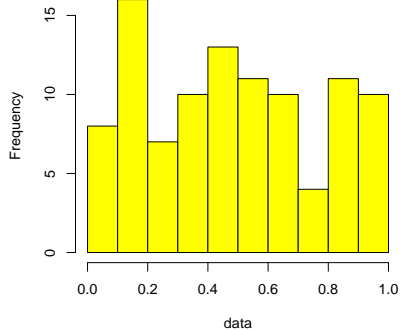
A normal probability plot for a normal sample of size 100



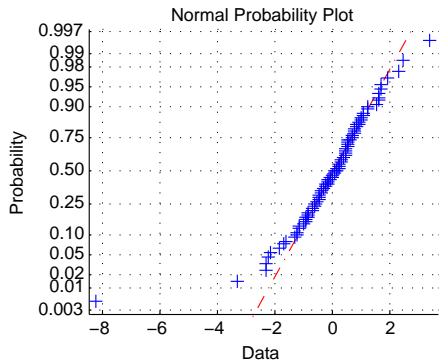
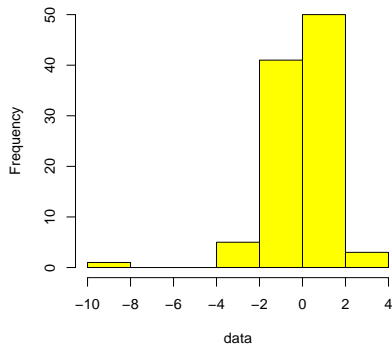
Skewed data



Uniform Data



Data with outlier



Assumption checking

- ▶ Assumption checking is a key step.
- ▶ Independence is the most important assumption but also the hardest to check.
- ▶ Use a normal probability plot to check for normality
 - ▶ Difficult to determine if normality is reasonable for very small samples ($n < 25$).
 - ▶ Skewness or outliers can cause serious problems with the validity of the t-test and confidence interval.
 - ▶ Uniform type violations of normality tend to be less problematic.

Summary of confidence interval calculations I

1. Identify the parameter of interest.
 - ▶ μ .
2. Identify and check the assumptions using the data.
 - ▶ x_1, x_2, \dots, x_n are independent normal observations.
 - ▶ Use a normal probability plot to assess normality.
3. Choose the confidence level.
 - ▶ Almost always use 95% confidence.
4. 4.1 Enter the data into Matlab and calculate the interval using the `ttest` command.
4.2 Or, for hand calculations:
 - ▶ Calculate \bar{x} , s and $SE = s/\sqrt{n}$.

Summary of confidence interval calculations II

- ▶ Identify the degrees of freedom $n - 1$ and hence find t^* from the t_{n-1} distribution.
- ▶ The confidence interval is

$$(\bar{x} - t^* s / \sqrt{n}, \bar{x} + t^* s / \sqrt{n})$$

Summary of hypothesis test calculations I

1. Identify the parameter of interest.
 - ▶ μ .
2. Identify the null and alternative hypotheses.
 - ▶ $H_0 : \mu = \mu_0$ vs $H_A : \mu \neq \mu_0$.
3. Choose the level of significance.
 - ▶ Almost always $\alpha = 0.05$.
4. Identify and check the assumptions using the data.
 - ▶ x_1, x_2, \dots, x_n are independent normal observations.
 - ▶ Use a normal probability plot to assess normality.
5. 5.1 Enter the data into Matlab and perform the test using the `ttest` command.

Summary of hypothesis test calculations II

5.2 Or, for hand calculations:

- ▶ Calculate \bar{x} , s and $SE = s/\sqrt{n}$ and the test statistic

$$t = \frac{\bar{x} - \mu_0}{s/\sqrt{n}}.$$

- ▶ Identify the degrees of freedom $n - 1$ and find the P-value from the t_{n-1} distribution.
- ▶ Reject H_0 for P-value $\leq \alpha$ otherwise accept H_0 .

Approximate hypothesis and confidence intervals I

- ▶ When the data are not normally distributed, the t-test and confidence intervals can still be used as an approximate procedure provided the sample size is large.
- ▶ This can be justified using the central limit theorem.
- ▶ The mathematical basis for both the t-test and confidence interval is the fact that for independent random variables X_1, X_2, \dots, X_n with $X_i \sim N(\mu, \sigma^2)$ for $i = 1, 2, \dots, n$,

$$\frac{\bar{X} - \mu}{s/\sqrt{n}} \sim t_{n-1}.$$

- ▶ Suppose now that X_1, X_2, \dots, X_n are independent (not necessarily normal) random variables with $E(X_i) = \mu$ and $\text{var}(X_i) = \sigma^2$ for $i = 1, 2, \dots, n$.

Approximate hypothesis and confidence intervals II

- ▶ The central limit theorem implies the approximate result:

$$\frac{\bar{X} - \mu}{\sigma/\sqrt{n}} \approx N(0, 1) \text{ for sufficiently large } n.$$

- ▶ It follows that for large samples, an approximate confidence interval for μ is

$$\bar{x} \pm z^* s / \sqrt{n}$$

and a test of $H_0 : \mu = \mu_0$ can be obtained from the test statistic

$$t = \frac{\bar{x} - \mu_0}{s/\sqrt{n}}$$

by calculating an approximate p-value using the $N(0, 1)$ distribution.

Approximate hypothesis and confidence intervals III

- ▶ Unfortunately there is no simple rule to decide how large the sample needs to be in order for the approximation to be good.
 - ▶ For non-normal distributions without pronounced skewness or outliers, the approximation will be good for relatively small sample sizes. For example $n = 50$ or $n = 100$.
 - ▶ For highly skewed distributions, the sample size may need to be very large before the approximation is useful.

Tests and confidence intervals for proportions

- ▶ In some applications, the parameter of interest is a proportion (or probability) rather than a mean.
- ▶ For example, in a random sample of 214 bricks, 18 were found to not be first grade (non-conforming).
 - ▶ In this case, the parameter of interest would be the **population proportion** p of non-conforming bricks.
- ▶ In general, suppose a random sample of n objects is chosen and that x are found to have an certain attribute of interest.
 - ▶ It is assumed that $X \sim B(n, p)$.
 - ▶ The sample proportion $\hat{p} = x/n$ is the estimate for p .
 - ▶ The standard error of the estimate is

$$SE = \sqrt{p(1-p)/n}.$$

Confidence interval for p

- ▶ An approximate confidence interval for p is given by

$$\hat{p} \pm z^* \sqrt{\hat{p}(1 - \hat{p})/n}$$

- ▶ The approximation is accurate provided $np \geq 10$ and $n(1 - p) \geq 10$.
- ▶ Example (continued): Find a 95% confidence interval for the (population) proportion of non-conforming bricks.
- ▶ Solution:
 - ▶ Since $x = 18$ and $n = 214$, $\hat{p} = 18/214 = 0.08411$.
 - ▶ The estimate standard error is $SE = \sqrt{\hat{p}(1 - \hat{p})/n} = 0.01897$.

Confidence interval for p II

- ▶ For 95% confidence, we use
 $z^* = \text{norminv}(0.975) = 1.96$.
- ▶ The confidence interval is (0.0469, 0.1213).
- ▶ In other words, we can be 95% confident that the proportion of non-conforming bricks lies between 4.69% and 12.13%.

Hypothesis test for p . I

- ▶ Consider the hypotheses $H_0 : p = p_0$ vs $H_A : p \neq p_0$.
- ▶ The test statistic is

$$z = \frac{\hat{p} - p_0}{\sqrt{p_0(1 - p_0)/n}}.$$

- ▶ An approximate P-value is $P(|Z| \geq |z|) = 2P(Z \geq |z|)$ for a standard normal random variable Z .
- ▶ Example (continued): Test, at the 5% level of significance, the hypothesis that the proportion of non-conforming bricks is 3%.
- ▶ Solution:
 - ▶ The null hypothesis is $H_0 : p = 0.03$.

Hypothesis test for p . II

- ▶ The test statistic is $z = \frac{\hat{p}-0.03}{\sqrt{0.03 \times (1-0.03)/214}} = 4.64$.
- ▶ The approximate P-value is

$$\text{P-value} = 2 * (1 - \text{normcdf}(4.64)) = 3.48 \times 10^{-6}$$

- ▶ Since P-value $\ll 0.05$ H_0 , is strongly rejected.
- ▶ We conclude there is strong evidence that $p > 0.03$.

Lecture 8

Confidence Intervals and Hypothesis Tests

Summary of confidence intervals and hypothesis tests

- ▶ We have considered various cases where confidence intervals and hypothesis tests can be applied.
- ▶ Confidence intervals have been of the form

$$\text{Estimate} \pm k \text{ SE}$$

- ▶ The constant k is chosen from the **reference distribution** to give the required level of confidence.
- ▶ The test statistics have been of the form

$$\frac{\text{Estimate} - \text{Null Value}}{\text{SE}}$$

- ▶ The P-value is calculated from the reference distribution.

The normal mean

Data x_1, x_2, \dots, x_n .

Assumptions Data are independent $N(\mu, \sigma^2)$ observations.

Parameter The normal mean μ .

Hypotheses $H_0 : \mu = \mu_0$ vs $H_A : \mu \neq \mu_0$.

Estimate The sample mean \bar{x} .

Standard Error

- ▶ $SE = \sigma / \sqrt{n}$ if σ^2 known.
- ▶ $SE = s / \sqrt{n}$ if σ^2 not known.

Reference Distribution

- ▶ $N(0, 1)$ if σ^2 known.
- ▶ t_{n-1} if σ^2 not known.

The binomial proportion

Data x, n .

Assumptions x is the number of successes from n independent trials, $X \sim B(n, p)$.

Parameter The success probability p .

Hypotheses $H_0 : p = p_0$ vs $H_A : p \neq p_0$.

Estimate The sample proportion $\hat{p} = x/n$.

Standard Error

- ▶ $SE = \sqrt{\hat{p}(1 - \hat{p})/n}$ for the confidence interval.
- ▶ $SE = \sqrt{p_0(1 - p_0)/n}$ for the hypothesis test.

Reference Distribution $N(0, 1)$.

Comparison of two means

- ▶ In many situations it is of interest to be able to compare the means for two populations.
- ▶ **Examples**
 - ▶ Up to a certain point, paint viscosity affects the coating thickness. In an experiment to determine whether they have reached this limit, engineers prepared samples at two levels of viscosity and recorded the coating thickness.
 - ▶ In an experiment to determine whether the time to drill through rock was different for wet and dry drilling, the time taken was recorded for 12 specimens using wet drilling and 12 specimens using dry drilling.
 - ▶ The fracture load in a 4 point bending test was measured for 30 specimens of each of two different high-strength microalloyed structural steels.

Two independent samples I

- ▶ Consider two samples:

Group 1 $x_{11}, x_{12}, \dots, x_{1n_1}$ assumed to be independent $N(\mu_1, \sigma_1^2)$ observations;

Group 2 $x_{21}, x_{22}, \dots, x_{2n_2}$ assumed to be independent $N(\mu_2, \sigma_2^2)$ observations;

- ▶ Assume also that observations in different groups are independent of each other.
- ▶ In this case the parameter of interest is the difference

$$\mu_1 - \mu_2.$$

Two independent samples II

- ▶ The estimate is

$$\bar{x}_1 - \bar{x}_2$$

where

$$\bar{x}_1 = \frac{1}{n_1} \sum_{j=1}^{n_1} x_{1j} \text{ and } \bar{x}_2 = \frac{1}{n_2} \sum_{j=1}^{n_2} x_{2j}.$$

The standard error

- ▶ The standard error of the estimate is

$$\text{SE}(\bar{x}_1 - \bar{x}_2) = \sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}$$

where

$$s_1^2 = \frac{1}{n_1 - 1} \sum_{j=1}^{n_1} (x_{1j} - \bar{x}_1)^2 \text{ and } s_2^2 = \frac{1}{n_2 - 1} \sum_{j=1}^{n_2} (x_{2j} - \bar{x}_2)^2$$

The reference distribution

- ▶ For this kind of t-test, a t-distribution can be used as an approximate reference distribution. The degrees of freedom ν can be obtained by two different methods:

Hand calculation Use $\nu = \min(n_1 - 1, n_2 - 1)$.

Computer packages A more complicated approximation resulting in fractional degrees of freedom is used.

$$\nu = \left(\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2} \right)^2 \bigg/ \left(\frac{s_1^4}{n_1^2(n_1 - 1)} + \frac{s_2^4}{n_2^2(n_2 - 1)} \right)$$

- ▶ **Note** Another variant of the two-sample t-test called the pooled t-test can be applied when it is assumed that $\sigma_1^2 = \sigma_2^2$. We do not consider that test in this course.

Assumptions

- ▶ There are two independence assumptions.
 - ▶ Observations within each sample are independent.
 - ▶ Observations from the two different samples are independent.
- ▶ It is also assumed that the data have normal distributions separately in the two groups.
 - ▶ Can check this assumption by constructing separate normal probability plots for the two groups.

Example: Steel breaking stress

Alloy A					
24700	31300	29400	31500	28700	31900
28400	27200	32800	30200	30200	32700
30900	42000	22800	36000	28000	27700
30500	28500	40400	31300	32600	34700
33100	23200	24300	35500	25300	24700

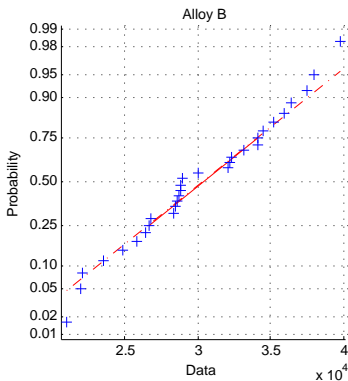
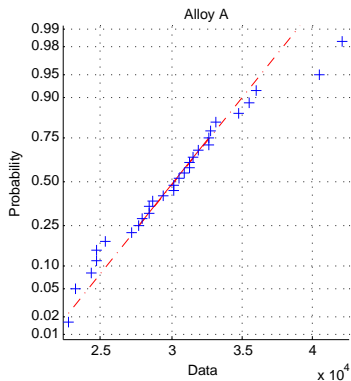
Alloy B					
34510	28730	36380	32060	25810	21900
28450	37510	29960	39730	21010	28880
35150	28750	35970	37960	28840	26400
24770	26580	34110	26750	23540	32300
34140	33140	28330	32220	22040	28610

Problems

Let μ_1 denote the mean breaking force for specimens of Alloy A and let μ_2 be the mean breaking force for specimens of Alloy B.

1. Check the assumption of normality separately for each group.
2. Using Matlab, test the hypothesis $H_0 : \mu_1 - \mu_2 = 0$ at the 5% level of significance and also obtain a 95% confidence interval for $\mu_1 - \mu_2$.
3. Use Matlab to calculate the sample means and sample variances for the two groups.
4. Use the sample means and variances to obtain a confidence interval and also to perform the hypothesis test.

Normal Probability Plots



Matlab Analysis

```
>> [h,p,ci,stats]=ttest2(AlloyA,AlloyB,[],[],'unequal')  
h =  
    0  
  
p =  
    0.8737  
  
ci =  
    1.0e+03 *  
  
    -2.2963  
     2.6943  
  
stats =  
    tstat: 0.1597  
         df: 57.4817  
         sd: [2x1 double]  
  
>>
```

Means and variances in Matlab

```
>> x1_bar=mean(AlloyA)  
x1_bar = 30350
```

```
>> x2_bar=mean(AlloyB)  
x2_bar = 30151
```

```
>> s1_squared=var(AlloyA)  
s1_squared = 2.1087e+07
```

```
>> s2_squared=var(AlloyB)  
s2_squared = 2.5513e+07
```

```
>> n1=length(AlloyA)  
n1 = 30
```

```
>> n2=length(AlloyB)  
n2 = 30
```

Confidence interval in Matlab

```
>> SE=sqrt(s1_squared/n1+s2_squared/n2)
```

```
SE =
```

```
1.2463e+03
```

```
>> df=min(n1,n2)-1
```

```
df =
```

```
29
```

```
>> k=tinvt(0.975,df)
```

```
k =
```

```
2.0452
```

```
>> lower=x1_bar-x2_bar-k*SE
```

```
lower =
```

```
-2.3500e+03
```

```
>> upper=x1_bar-x2_bar+k*SE
```

```
upper =
```

```
2.7480e+03
```

Hypothesis test in Matlab

```
>> t=(x1_bar-x2_bar)/SE  
t =  
    0.1597
```

```
>> pval=2*(1-tcdf(t,df))  
pval =  
    0.8742  
>>
```

Conclusion: There is no evidence of a difference between the means strengths of these two types of alloys.

Remark: The differences between our calculations and those obtained from `ttest2` arise because of the different degrees of freedom. We used 29 df and formula used in `ttest2` gives 57.48df.

Paired Comparisons

- ▶ Another commonly used experimental design for the comparison of two means is the matched pairs design.
- ▶ In this type of experiment, n items are considered and two measurements are made on each item.
- ▶ **Example** In an experiment to compare two different methods of measuring the wall thickness of cylinder heads, two measurements were made on each of 18 cylinder heads, one using ultrasound and one by sectioning. The purpose of the experiment was to determine whether there is a systematic difference between the two methods of measurement.

Cylinder Head Measurements

Cylinder Head	1	2	3	4	5	6
Ultrasound Section	0.223	0.193	0.218	0.201	0.231	0.204
	0.224	0.207	0.216	0.204	0.230	0.203
Cylinder Head	7	8	9	10	11	12
Ultrasound Section	0.228	0.223	0.215	0.223	0.237	0.226
	0.222	0.225	0.224	0.223	0.226	0.232
Cylinder Head	13	14	15	16	17	18
Ultrasound Section	0.214	0.213	0.233	0.224	0.217	0.210
	0.217	0.217	0.237	0.224	0.219	0.192

Analysis of paired data

- ▶ Suppose two variables X_1 and X_2 are recorded on each of n items.
- ▶ The data can be represented as

$$(x_{11}, x_{12}), (x_{21}, x_{22}), \dots, (x_{n1}, x_{n2})$$

- ▶ The data cannot be treated as two independent samples because observations made on the same item could be expected to be correlated.
 - ▶ This violates the assumption that observations from different groups are independent.
- ▶ The strategy for paired data is to calculate the differences

$$d_i = x_{i1} - x_{i2}$$

and apply the one-sample t-procedures to those data.

Analysis of paired data

- ▶ The data are

$$d_1, d_2, \dots, d_n.$$

- ▶ The assumption is that the values are independent $N(\delta, \sigma^2)$ observations.
- ▶ The parameter of interest is $\delta = E(X_1 - X_2)$.
- ▶ The estimate is

$$\bar{d} = \frac{1}{n} \sum_{i=1}^n d_i.$$

- ▶ The standard error is

$$\text{SE} = s_d / \sqrt{n} \text{ where } s_d = \sqrt{\frac{1}{n-1} \sum_{i=1}^n (d_i - \bar{d})^2}.$$

- ▶ The reference distribution is t_{n-1} .

Example - Cylinder Head Data I

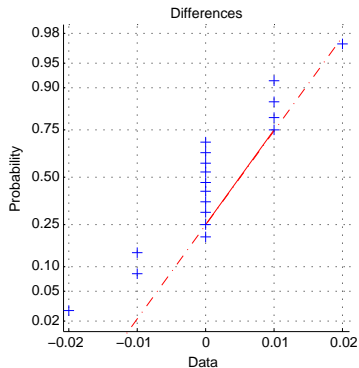
Perform a suitable hypothesis test and obtain a confidence interval to determine whether there are systematic differences between the section and ultrasound measurements.

- ▶ Because this is paired data, we work with the differences

$$\text{Difference} = \text{Section} - \text{Ultrasound}$$

- ▶ The first consider the normal probability plot of the differences.

Example - Cylinder Head Data II



Matlab analysis based on the differences

```
>> Difference=Section-Ultrasound;  
>> normplot(Difference)  
>> [h,p,ci,stats]=ttest(Difference)
```

```
h =  
    0
```

```
p =  
    0.6073
```

```
ci =  
   -0.0034  
    0.0056
```

```
stats =  
    tstat: 0.5236  
       df: 17  
       sd: 0.0090
```

Matlab analysis using built-in paired data feature

```
>> [h,p,ci,stats]=ttest(Section,Ultrasound)
```

```
h =
```

```
0
```

```
p =
```

```
0.6073
```

```
ci =
```

```
-0.0034
```

```
0.0056
```

```
stats =
```

```
tstat: 0.5236
```

```
df: 17
```

```
sd: 0.0090
```

```
>>
```

Conclusions

- ▶ The pattern apparent in the normal probability plot results from the discreteness in the recorded data.
 - ▶ It is strictly speaking not normally distributed but this violation would not impact greatly on the validity of the procedure.
- ▶ The large P-value (0.6073) leads us to accept $H_0 : \delta = 0$ so we conclude that there is no evidence of a systematic difference between the two methods.
- ▶ The 95% confidence interval $(-0.0034, 0.0056)$ is relatively narrow compared to the range of the data.
- ▶ Note that although there are no negative indications, this analysis does not prove that the two methods are equivalent.

Paired comparisons or two independent samples?

- ▶ In both situations the parameter of interest is a difference of two means.
- ▶ The correct analysis is very different for these two designs so it is important not to get them confused.
- ▶ Two independent samples generally arise from $n_1 + n_2$ separate items with a single measurement on each item.
- ▶ Paired samples most often arise from n items with two measurements on each item.

Comparison of two proportions

- ▶ Sometimes it is of interest to compare two binomial probabilities.
- ▶ For example, in an experiment to investigate the effect of Gamma irradiation on foodstuffs:
 - ▶ Out of 180 garlic bulbs that were irradiated, 153 were suitable for sale after 240 days in storage.
 - ▶ Out of 180 untreated garlic bulbs, 119 were suitable for sale after 240 days in storage.
- ▶ In general, consider two independent binomial observations, $X_1 \sim B(n_1, p_1)$ and $X_2 \sim B(n_2, p_2)$.
- ▶ The parameter of interest is the difference

$$p_1 - p_2.$$

Estimation

- ▶ The estimate is

$$\hat{p}_1 - \hat{p}_2 \quad \text{where} \quad \hat{p}_1 = \frac{x_1}{n_1} \quad \text{and} \quad \hat{p}_2 = \frac{x_2}{n_2}.$$

- ▶ The standard error is

$$SE = \sqrt{\frac{p_1(1-p_1)}{n_1} + \frac{p_2(1-p_2)}{n_2}}.$$

- ▶ For confidence intervals we use

$$SE = \sqrt{\frac{\hat{p}_1(1-\hat{p}_1)}{n_1} + \frac{\hat{p}_2(1-\hat{p}_2)}{n_2}}.$$

- ▶ To test $H_0 : p_1 - p_2 = 0$ we use

$$SE = \sqrt{\hat{p}(1-\hat{p}) \left(\frac{1}{n_1} + \frac{1}{n_2} \right)} \quad \text{where} \quad \hat{p} = \frac{x_1 + x_2}{n_1 + n_2}.$$

Example

- ▶ Test, at the 5% level of significance, the hypothesis that the proportion irradiating the garlic bulbs does not affect the proportion that will be saleable after 240 days in storage.
- ▶ Obtain a 95% confidence interval for the difference in the proportions.

```
>> x1=153; n1=180; p1=x1/n1  
p1 =  
    0.8500
```

```
>> x2=119; n2=180; p2=x2/n2  
p2 =  
    0.6611
```

```
>> diff=p1-p2  
diff =  
    0.1889
```

Matlab hypothesis test calculations

```
>> p=(x1+x2)/(n1+n2)
```

```
p =  
    0.7556
```

```
>> se_test=sqrt(p*(1-p)*(1/n1+1/n2))
```

```
se_test =  
    0.0453
```

```
>> z=diff/se_test
```

```
z =  
    4.1697
```

```
>> p_val=2*(1-normcdf(z))
```

```
pval =  
    3.0500e-05
```

Conclusion: $H_0 : p_1 = p_2$ is strongly rejected so we conclude that Gamma irradiation improves the storage life of garlic.

Matlab confidence interval calculations

```
>> k=norminv(0.975)
k =
    1.9600

>> se_ci=sqrt(p1*(1-p1)/n1+p2*(1-p2)/n2)
se_ci =
    0.0442

>> ci=[diff-k*se_ci,diff+k*se_ci]
ci =
    0.1023    0.2755

>>
```

Conclusion: We can claim with 95% confidence that Gamma irradiation increases the percentage of saleable garlic bulbs by between 10.23% and 27.55%.

Summary of confidence intervals and hypothesis tests

- ▶ We have considered various cases where confidence intervals and hypothesis tests can be applied.
- ▶ Confidence intervals have been of the form

$$\text{Estimate} \pm k \text{ SE}$$

- ▶ The constant k is chosen from the reference distribution to give the required level of confidence.
- ▶ The test statistics have been of the form

$$\frac{\text{Estimate} - \text{Null Value}}{\text{SE}}$$

- ▶ The P-value is calculated from the reference distribution.

Summary Table

Context	Parameter	Estimate	SE	Reference Dist
Normal Mean σ^2 known	μ	\bar{x}	σ/\sqrt{n}	$N(0, 1)$
Normal Mean σ^2 not known	μ	\bar{x}	s/\sqrt{n}	t_{n-1}
2 Independent Samples	$\mu_1 - \mu_2$	$\bar{x}_1 - \bar{x}_2$	$\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}$	t_ν $\nu = \min(n_1, n_2) - 1$
Paired comparison	δ	\bar{d}	s_d/\sqrt{n}	t_{n-1}
Binomial CI	p	$\hat{p} = x/n$	$\sqrt{\hat{p}(1-\hat{p})/n}$	$N(0, 1)$
Binomial Test	p	$\hat{p} = x/n$	$\sqrt{p_0(1-\hat{p}_0)/n}$	$N(0, 1)$
2 Proportions CI	$p_1 - p_2$	$\frac{x_1}{n_1} - \frac{x_2}{n_2}$	$\sqrt{\frac{\hat{p}_1(1-\hat{p}_1)}{n_1} + \frac{\hat{p}_2(1-\hat{p}_2)}{n_2}}$	$N(0, 1)$
2 Proportions Test	$p_1 - p_2$	$\frac{x_1}{n_1} - \frac{x_2}{n_2}$	$\sqrt{\hat{p}(1-\hat{p})\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}$	$N(0, 1)$

Lecture 9

Linear Regression

Linear Regression

Consider data that arises when two variables X and Y are recorded on each of n subjects.

In what follows we will assume that X is a *predictor variable*, (also called *stimulus* or *independent variable*) and Y is the response (also called *dependent variable*).

Regression analysis is used to quantify the effect of the predictor variable X on the response Y .

The data can be represented as

$$(x_1, y_1), (x_2, y_2), \dots, (x_n, y_n).$$

Examples I

- ▶ A municipal incinerator generates electricity using the energy released when rubbish is burnt. Deliveries of rubbish vary in water content and the operator would like to estimate the available energy from the measured water content.

Predictor variable: X is the water content as a percentage by weight;

Response variable: Y is the energy density in kCal/kg.

- ▶ In an electroplating process, it is important to control the thickness of the coating. The goal here is to understand the relationship between coating thickness and coating time.

Predictor variable: X is the coating time in seconds;

Examples II

Response variable: Y is the thickness in mm.

- ▶ The available chlorine in a product decays with time. A manufacturer wishes to determine the shelf life of the product.

Predictor variable: X is the time since manufacture in weeks;

Response variable: Y is the available chlorine.

The linear regression model

The linear regression model postulates that

$$y_i = \beta_0 + \beta_1 x_i + e_i$$

for $i = 1, 2, 3, \dots, n$, where the errors

$$e_1, e_2, \dots, e_n$$

are assumed to be *independent* $N(0, \sigma^2)$ realizations.

The regression coefficients

When the regression model fits the data, it means that the patterns in the data set can be described by three numbers.

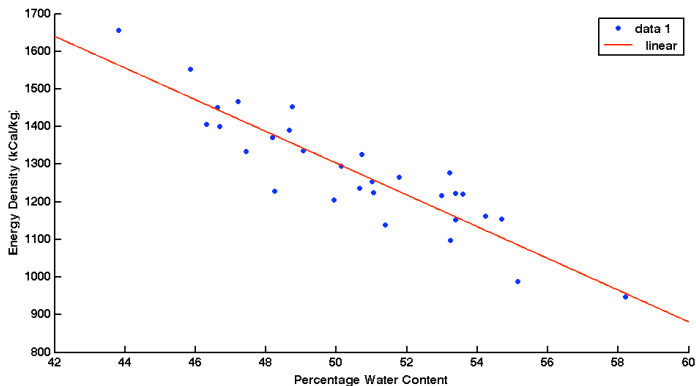
Namely the parameters β_0 , β_1 and σ .

For the incinerator data, these parameters are estimated to be

$$\hat{\beta}_0 = 3,412.2, \quad \hat{\beta}_1 = -42.18 \quad \text{and} \quad s_e = 67.73.$$

Note that s_e is used to estimate σ and is called the *residual standard deviation*.

Incinerator data with the line $y = 3,412.2 - 42.18x$



The regression coefficients I

The coefficients β_0 and β_1 are called the regression coefficients.

- ▶ β_1 is the slope of the regression line.
 - ▶ The estimated value $\hat{\beta}_1 = -42.18$ in the present example means that for each increase of 1% in water content, the energy density reduces, on average, by 42.18 kCal per kg.
- ▶ β_0 is the intercept of the regression line.
 - ▶ In principle, it would be the average energy density for water content 0.

The regression coefficients II

- ▶ In the present data, this is a moot point because it appears that water content for rubbish varies between roughly 40% and 60%.
- ▶ In general β_0 only has a useful interpretation when $X = 0$ is a realistic value.

The residual standard deviation I

- ▶ The parameter σ is called the residual standard deviation.
- ▶ Recall the regression model states that

$$y_i = \beta_0 + \beta_1 x_i + e_i.$$

- ▶ Equivalently,

$$e_i = y_i - (\beta_0 + \beta_1 x_i)$$

so the magnitude of the e_i determines the **vertical** spread of points about the regression line.

- ▶ Since the regression model assumes $e_i \sim N(0, \sigma^2)$, it follows that:
 - ▶ If $\sigma = 0$, then all points would lie exactly on the regression line;

The residual standard deviation II

- ▶ For $\sigma > 0$, about 68% of the y_i should lie in the range $\beta_0 + \beta_1 x_i \pm \sigma$;
- ▶ For $\sigma > 0$, about 95% of the y_i should lie in the range $\beta_0 + \beta_1 x_i \pm 2\sigma$;
- ▶ For $\sigma > 0$, about 99.7% of the y_i should lie in the range $\beta_0 + \beta_1 x_i \pm 3\sigma$;

Regression calculations I

Estimation

In practice, β_0 , β_1 and σ are not known and must be estimated from data. We will:

- ▶ Explain the mathematical principles used for estimation;
- ▶ Demonstrate how the calculations are performed in Matlab.

Diagnostics

The regression model is a set of assumptions that may or may not apply to given data. We will:

- ▶ Describe regression diagnostics that allow us to check whether the assumptions are reasonable;

Regression calculations II

- ▶ Demonstrate that regression calculations can produce misleading conclusions when the assumptions are violated.
- ▶ Show how diagnostics are produced in Matlab.

Inference

The estimated regression coefficients are **statistical estimates** of the underlying **model parameters** in the same way that a sample mean \bar{x} can be used as an estimate of a population mean μ . We will:

- ▶ Introduce standard errors and show how they can be used to obtain confidence intervals and test hypotheses for the regression parameters;
- ▶ Demonstrate how the calculations are performed in Matlab.

Regression calculations III

Prediction

Prediction is an important application of regression. We will:

- ▶ Define a prediction problem for regression;
- ▶ Introduce confidence intervals and prediction intervals;
- ▶ Demonstrate how the calculations are performed in Matlab.

Estimation

The regression parameters β_0, β_1 are usually estimated using the method of least squares.

That is, $\hat{\beta}_0, \hat{\beta}_1$ are chosen jointly to minimize

$$Q(\beta_0, \beta_1) = \sum_{i=1}^n (y_i - (\beta_0 + \beta_1 x_i))^2.$$

The solutions are

$$\hat{\beta}_1 = \frac{S_{xy}}{S_{xx}} \text{ and } \hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x}$$

where

$$S_{xy} = \sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y}) \text{ and } S_{xx} = \sum_{i=1}^n (x_i - \bar{x})^2.$$

The **residual** for the i th observation is defined by

$$\hat{e}_i = y_i - (\hat{\beta}_0 + \hat{\beta}_1 x_i).$$

The residual variance σ^2 is estimated by the residual mean squared error,

$$s_e^2 = \frac{1}{n-2} \sum_{i=1}^n (y_i - (\hat{\beta}_0 + \hat{\beta}_1 x_i))^2 = \frac{1}{n-2} \sum_{i=1}^n \hat{e}_i^2.$$

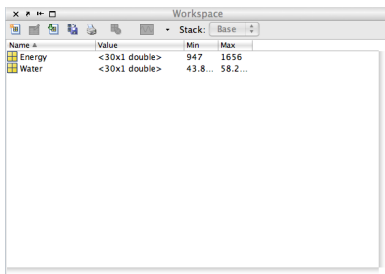
The residual standard deviation is estimated by

$$s_e = \sqrt{s_e^2}.$$

Regression Calculations in Matlab

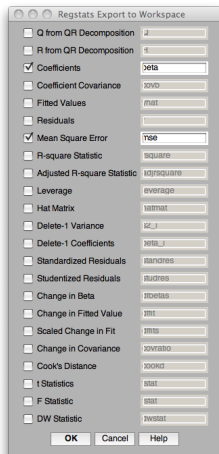
One way to perform basic regression calculations in Matlab is using the `regstat` function. The simplest usage is `regstat(Y,X)` where `Y` is the response and `X` is the predictor.

We illustrate with the incinerator data. Assume that the data have been entered into Matlab.



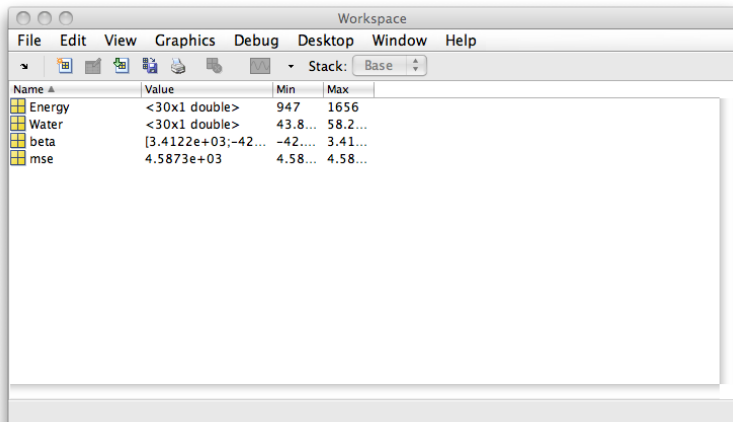
Matlab Calculations (continued)

1. The first step is to enter the regstats(Energy,Water) in the command window.
2. The tick the Coefficients and Mean Square Error check boxes in the Regstats dialog box and click OK.



Matlab Calculations (continued)

3. Two new items, `beta` and `mse` have been created and appear in the workspace.



Matlab Calculations (continued)

4. These items can be examined in the command window.

```
>> beta
beta =
    3.4122e+03
   -4.2182e+01
>> mse
mse =
    4.5873e+03
>> sqrt(mse)
ans =
    6.7729e+01
>>
```

Regression Diagnostics I

Most regression diagnostics are based on the **residuals** defined by

$$\hat{e}_i = y_i - (\hat{\beta}_0 + \hat{\beta}_1 x_i).$$

The logic of considering residuals is that if the linear regression model

$$y_i = \beta_0 + \beta_1 x_i + e_i$$

is correct and $\hat{\beta}_0 \approx \beta_0$ and $\hat{\beta}_1 \approx \beta_1$ then we should have

$$\hat{e}_i \approx e_i.$$

If the linear regression model is correct, then

$$\hat{e}_1, \hat{e}_2, \dots, \hat{e}_n$$

Regression Diagnostics II

should appear *roughly* like a sample of n independent $N(0, \sigma^2)$ observations.

The residual plot is the plot of the residuals \hat{e}_i vs fitted values \hat{y}_i defined by

$$\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_i.$$

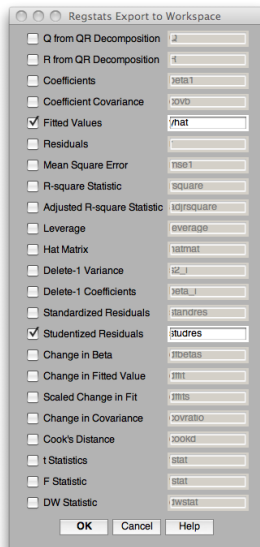
- ▶ In simple linear regression, plotting \hat{e}_i vs \hat{y}_i is equivalent to just plotting \hat{e}_i vs x_i .
 - ▶ The reason for considering the fitted values is that the definition extends to multiple regression.
- ▶ Many texts suggest using a refinement of the ordinary residuals called the *studentized* residuals.

Regression Diagnostics III

- ▶ Studentized residuals are more appropriate for identifying problems with non-constant variance;
- ▶ Studentized residuals provide a more sensitive tool for identifying outliers.
- ▶ The fitted values and studentized residuals can be obtained in Matlab from `regstats` and then plotted.

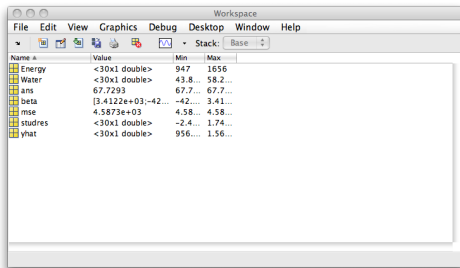
Diagnostics (continued)

1. In the regstats dialog box, tick the Fitted Values and Studentized Residuals check boxes.



Diagnostics (continued)

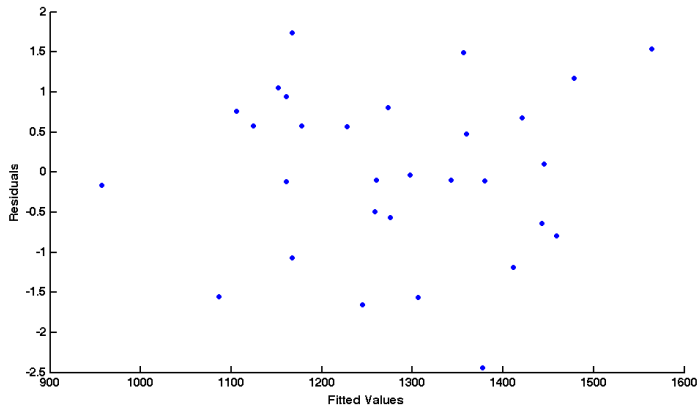
2. Two new items, `studres` and `yhat` will appear in the workspace.



3. The residual plot is obtained using the `scatter` command.

```
>> scatter(yhat,studres)
>> xlabel('Fitted Values'); ylabel('Residuals');
```

Example: residual plot for the incinerator data



Example (continued)

For the incinerator example, the residuals appear to be distributed randomly with respect to the fitted values.

- ▶ There is no evidence of curvature in the residual plot.
- ▶ There is also no evidence that the *residual variance* is not constant.

Therefore we conclude that the regression assumption:

$$y_i = \beta_0 + \beta_1 x_i + e_i$$

with

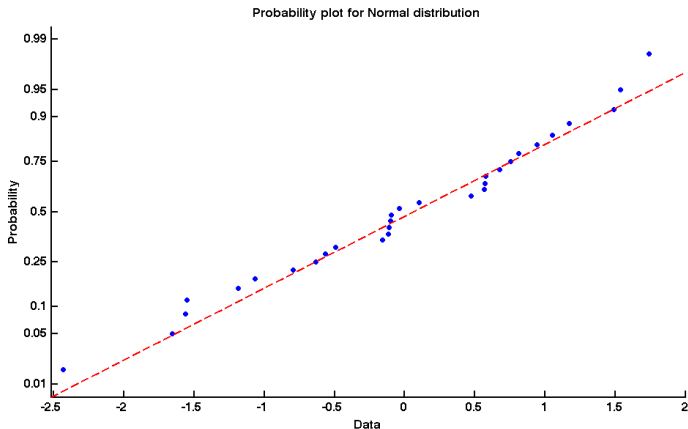
$$E(e_i) = 0 \quad \text{and} \quad \text{var}(e_i) = \sigma^2$$

for $i = 1, 2, \dots, n$, appears reasonable.

Finally the assumption of normality can be checked by considering a normal probability plot of the residuals.

```
>> normplot(studres)
```

Example (continued)



Since the points closely follow a straight line, we conclude that the assumption of normality is also reasonable.

Lecture 10

Linear Regression II

Regression Diagnostics I

Recall the residual plot was defined to be the scatter plot of the residuals vs fitted values for the linear regression model.

Notes:

- ▶ We use the studentized residuals;
- ▶ For simple linear regression plotting the fitted values is equivalent to plotting X -values.

When the linear regression model

$$y_i = \beta_0 + \beta_1 x_i + e_i \quad \text{with} \quad e_i \sim N(0, \sigma^2) \quad \text{independently for} \quad i = 1,$$

is correct, the vertical coordinates of the residual plot should be randomly scattered.

Regression Diagnostics II

This behaviour was exemplified with the incinerator example.

In what follows, we illustrate some common violations of the linear regression model.

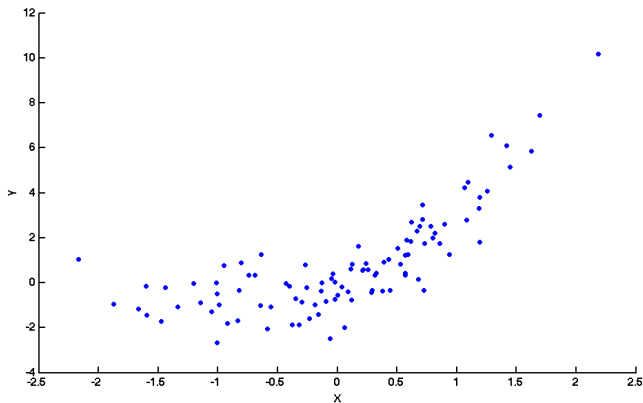
Non-linearity I

According to the linear regression model the data should be clustered about the line $\mu_Y = \beta_0 + \beta_1 x$.

If the relationship between Y and X is not linear then it may happen that the data are clustered about a curve rather than a line.

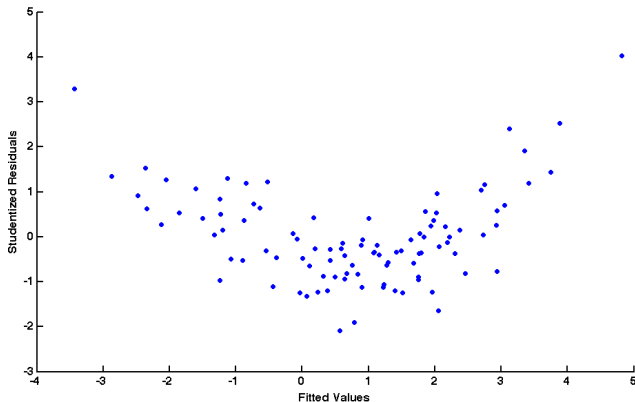
Shown below is an artificial data set illustrating curvature.

Non-linearity II



Residual plot showing curvature

The residual plot for the same data is shown below.



Non-constant variance

The linear regression model also includes a constant variance assumption.

In particular, the assumption that $e_i \sim N(0, \sigma^2)$ for all i means that the variance is the same for all observations.

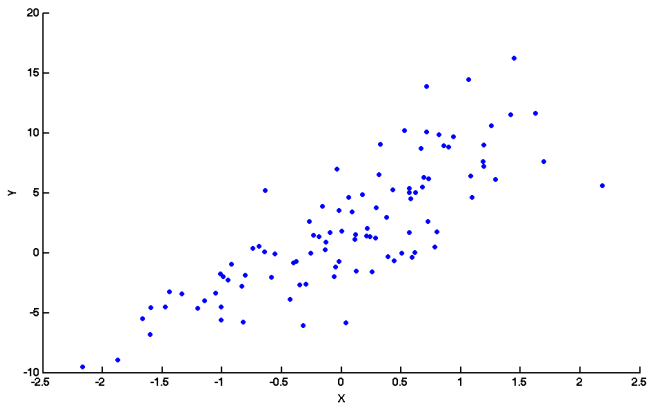
This condition is called **homoscedasticity**.

It means that the spread of points about the regression line should be roughly constant.

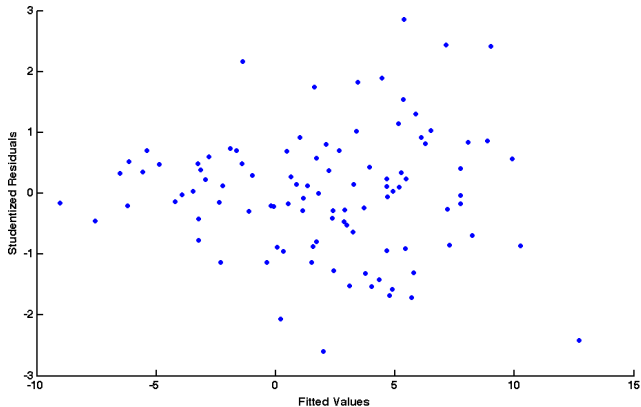
If the variance is not constant, then **heteroscedasticity** is said to occur.

A frequently occurring form of heteroscedasticity is when the variance increases with the fitted value.

Artificial data showing increasing variance

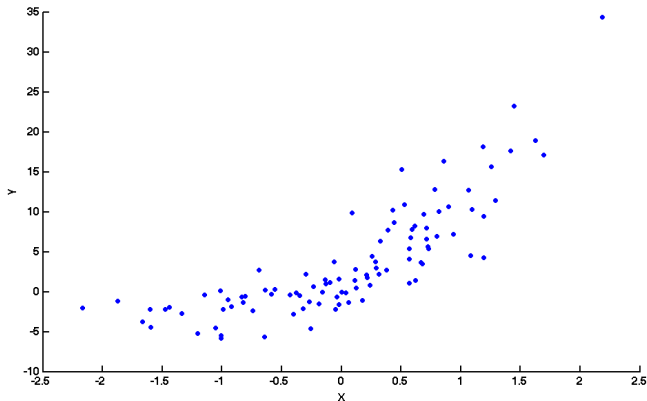


Residual plot for increasing variance data

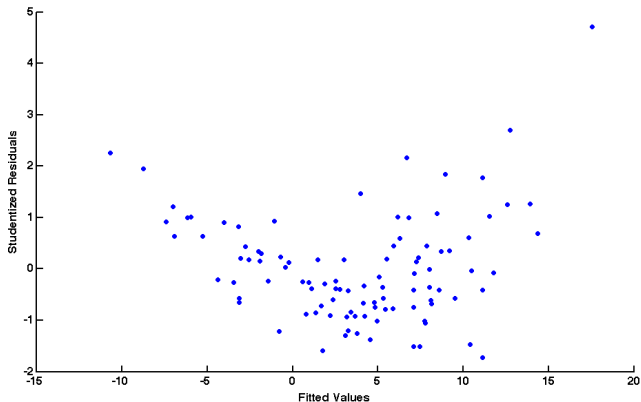


Artificial data with both curvature and increasing variance

When the linear regression assumptions are violated it can happen that the data displays both curvature and increasing variance.

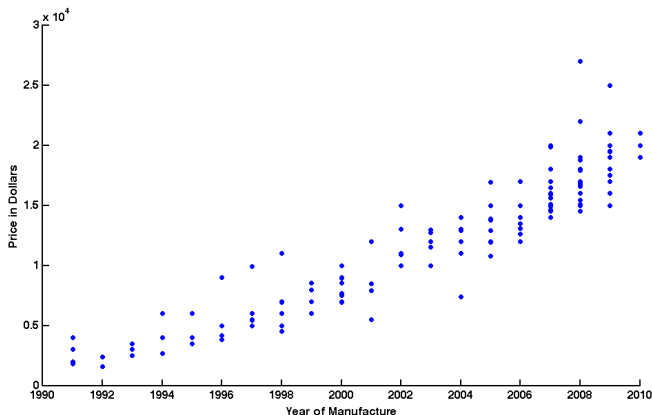


Residual plot for curvature and increasing variance data



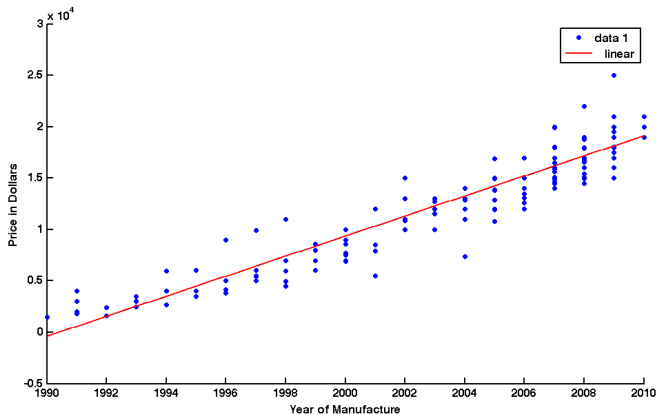
Used car price example

The year of manufacture and price in \$ for 190 second-hand toyota corollas advertised in SA on 19 Feb 2011 on www.carsguide.com.au.



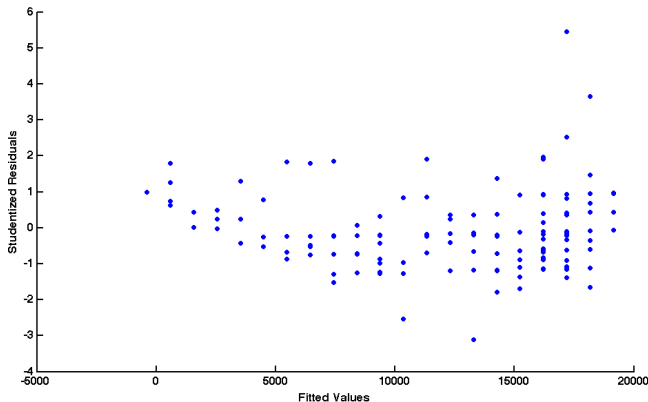
Example (continued)

At first sight, it might appear that the linear regression model is a reasonable approximation.



Example (continued)

However, the residual plot shows clear evidence of curvature and increasing variance.



Example (continued)

Suppose the linear regression model is fitted to these data.

```
>> beta
```

```
beta =
```

```
-1.9424e+06
```

```
9.7591e+02
```

The estimated regression equation is

$$\text{Price} = -1942443 + 975.9 \times \text{Year}.$$

The predicted average price for cars manufactured in 1990 is

$$-1942443 + 975.9 \times 1990 = -383.35$$

which is clearly an absurd conclusion.

Regression diagnostics summary

The assumptions of the linear regression model can be listed as:

Linearity: The linear form $E(Y) = \beta_0 + \beta_1 x$ is appropriate or, equivalently, $E(e_i) = 0$;

Constant variance: $\text{var}(e_i) = \sigma^2$ for all $i = 1, 2, \dots, n$;

Independence: e_1, e_2, \dots, e_n are statistically independent;

Normality: $e_i \sim N(0, \sigma^2)$ for all $i = 1, 2, \dots, n$.

Of these assumptions, independence is the most difficult to check. If the data are recorded serially in time, then a partial check for independence can be obtained by plotting the residuals against time.

Summary (continued)

It is recommended that you check the assumptions of the regression model as far as possible before using the model to draw conclusions.

1. Use the residual plot to check the assumptions of
 - ▶ linearity,
 - ▶ constant variance.
2. If applicable, plot the residuals vs time to check for possible dependence in the data.
3. If the assumptions of linearity, constant variance and independence appear reasonable, use a normal quantile plot to check for normality.

Inference for regression coefficients

The least squares estimates $\hat{\beta}_0$ and $\hat{\beta}_1$ are estimates of the “true” intercept and slope parameters in the regression model

$$y_i = \beta_0 + \beta_1 x_i + e_i.$$

This is analogous to considering the sample mean \bar{x} as an estimate of the underlying population mean μ .

The key statistical concepts:

- ▶ Standard error;
- ▶ Confidence interval;
- ▶ Hypothesis tests;

that were introduced for means can also be applied to regression coefficients.

Standard errors for the regression coefficients I

- ▶ The standard error of the estimated slope coefficient is

$$\text{SE}(\hat{\beta}_1) = \frac{\sigma}{\sqrt{S_{xx}}} \quad \text{where} \quad S_{xx} = \sum_{i=1}^n (x_i - \bar{x})^2.$$

- ▶ In practice, σ is not known so we use the estimated standard error,

$$\frac{s_e}{\sqrt{S_{xx}}}.$$

- ▶ The standard error of the estimated intercept coefficient is

$$\text{SE}(\hat{\beta}_0) = \sigma \sqrt{\frac{1}{n} + \frac{\bar{x}^2}{S_{xx}}}.$$

Standard errors for the regression coefficients II

- ▶ In practice, σ is not known so we use the estimated standard error,

$$s_e \sqrt{\frac{1}{n} + \frac{\bar{x}^2}{S_{xx}}}.$$

Key results

Confidence intervals and hypothesis tests for the regression coefficients are derived from the following results.

Suppose the random variables Y_1, \dots, Y_n satisfy the linear regression model

$$Y_i = \beta_0 + \beta_1 x_i + \mathcal{E}_i$$

with $\mathcal{E}_i \sim N(0, \sigma^2)$ independently.

In this framework, the **random variables** $\hat{\beta}_0$ and $\hat{\beta}_1$ satisfy

$$\frac{\hat{\beta}_1 - \beta_1}{s_e / \sqrt{S_{xx}}} \sim t_{n-2}$$

and

$$\frac{\hat{\beta}_0 - \beta_0}{s_e \sqrt{1/n + \bar{x}^2 / S_{xx}}} \sim t_{n-2}.$$

Confidence intervals for $\hat{\beta}_1$

For $0 < \alpha < 1$, a $100(1 - \alpha)\%$ confidence interval for the slope is given by

$$\hat{\beta}_1 \pm t^* \frac{s_e}{\sqrt{S_{xx}}}$$

where t^* is chosen from the t_{n-2} reference distribution.

In practice, we usually consider 95% confidence intervals. These confidence intervals are generally used to describe the uncertainty associated with the estimate of the slope.

Example

In the incinerator example, we previously found $\hat{\beta}_1 = -42.18$.

- ▶ From Matlab, the estimated standard error is $SE(\hat{\beta}_1) = 3.814$.
- ▶ Since there were $n = 30$ observations, the residual degrees of freedom is $n - 2 = 28$.
- ▶ From Matlab, $t^* = \text{tinv}(0.975, 28) = 2.0484$, for a 95% confidence interval.

The 95% confidence interval is

$$-42.18 \pm 2.0484 \times 3.814$$

or equivalently,

$$(-49.99, -34.37).$$

Example (continued)

Previously estimated the reduction in energy density associated with a 1% increase in water content to be 42.18 kCal per kg.

Using the confidence interval, we can now claim with 95% confidence that the reduction in energy density associated with a 1% increase in water content lies between 34.37 kCal per kg and 49.99 kCal per kg.

Confidence intervals for β_0

Confidence intervals for β_0 can also be produced:

$$\hat{\beta}_0 \pm t^* s_e \sqrt{\frac{1}{n} + \frac{\bar{x}^2}{S_{xx}}}.$$

However, in most practical situations, the slope β_1 rather than the intercept β_0 is of interest.

Tests of $H_0 : \beta_1 = 0$

In the context of the linear regression model,

$$y_i = \beta_0 + \beta_1 x_i + e_i \quad \text{where} \quad e_i \sim N(0, \sigma^2),$$

it is relevant to consider the hypothesis

$$H_0 : \beta_1 = 0.$$

- ▶ If H_0 is true, then Y and X are unrelated.
- ▶ If H_0 is false, then Y and X are related.

Hence if H_0 is **rejected** then the data can be said to provide strong evidence of a relationship between X and Y .

Hypothesis tests (continued)

To test $H_0 : \beta_1 = 0$:

Test statistic:

$$t = \frac{\hat{\beta}_1}{s_e / \sqrt{S_{xx}}}.$$

Degrees of freedom: $n - 2$.

Reference distribution: t_{n-2} .

P-value: The P-value is defined by $2P(T \geq |t|)$ for a random variable $T \sim t_{n-2}$.

Decision rule:

- ▶ Reject H_0 for P-value $\leq \alpha$;
- ▶ Accept H_0 for P-value $> \alpha$.

Example

To test $H_0 : \beta_1 = 0$ in the incinerator example, the test statistic is

$$t = \frac{\hat{\beta}_1}{\text{SE}(\hat{\beta}_1)} = \frac{-42.18}{3.814} = -11.06.$$

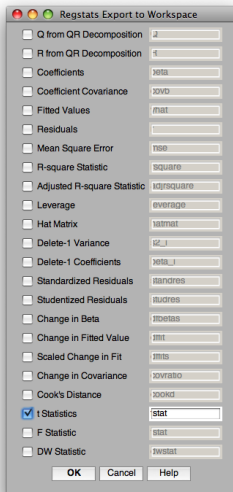
The P-value is

$$P - \text{value} = 2 * (1 - \text{tcdf}(11.06, 28)) = 9.9 \times 10^{-12}.$$

In this case the conclusion is that the data provide overwhelming evidence of a relationship between water content and energy density.

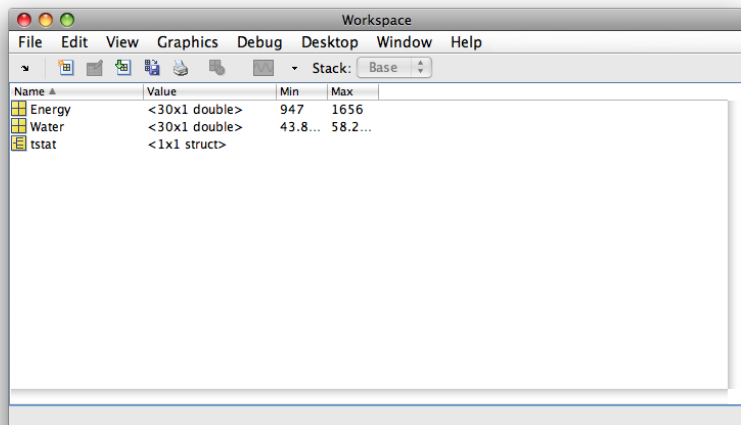
Regression inference in Matlab

To obtain confidence intervals and perform hypothesis tests for the regression coefficients, the t Statistics check box is ticked in the regstats dialog box.



Matlab (continued)

A new data structure called tstat is created.



Matlab (continued) I

The structure has named components:

beta: The estimated regression coefficients, $\hat{\beta}_0$ and $\hat{\beta}_1$.

se: The standard errors.

t: The t-statistics

pval: The p-values.

dfe: The residual degrees of freedom, $n - 2$.

Matlab (continued) II

```
>> tstat
```

```
tstat =
```

```
beta: [2x1 double]  
se: [2x1 double]  
t: [2x1 double]  
pval: [2x1 double]  
dfe: 28
```

Matlab (continued) I

The components may be extracted using "."

For example, `tstat.beta` extracts the vector of regression coefficients.

```
>> tstat.beta
```

```
ans =
```

```
3.4122e+03  
-4.2182e+01
```

A useful way to display the results is to arrange them in a matrix.

Matlab (continued) II

```
>> [tstat.beta,tstat.se,tstat.t,tstat.pval]
```

```
ans =
```

```
    3.4122e+03    1.9304e+02    1.7676e+01    1.0083e-16  
   -4.2182e+01    3.8135e+00   -1.1061e+01    9.9221e-12
```

Matlab (continued) I

The ordering within this matrix is shown below.

Coefficient	Estimate	SE	t-statistic	P-value
Intercept (β_0)	3.4122e+03	1.9304e+02	1.7676e+01	1.0083e-16
Slope (β_1)	-4.2182e+01	3.8135e+00	-1.1061e+01	9.9221e-12

The components of `tstat` can also be used to construct confidence intervals for both β_0 and β_1 .

```
>> lower=tstat.beta-tinv(0.975,tstat.dfe)*tstat.se;  
>> upper=tstat.beta+tinv(0.975,tstat.dfe)*tstat.se;  
>> [lower,upper]
```

Matlab (continued) II

```
ans =
```

```
    3.0168e+03    3.8076e+03  
   -4.9994e+01   -3.4371e+01
```

```
>>
```

Lecture 11

Regression III

Prediction

An important application of regression is prediction.

Suppose a new observation of Y is to be made and we already know the value of X .

To distinguish the new value from the observed data

$$(x_1, y_1), (x_2, y_2), \dots, (x_n, y_n)$$

we use the notation (x_0, Y_0) .

x_0 : The known value of the predictor variable X for the new observation;

Y_0 : The yet to be observed value of the response Y that we would like to predict.

Example

In the incinerator example, suppose a shipment of garbage is received and the water content is determined to be 58%.

The problem at hand is to obtain a prediction for the energy density of the shipment using that information.

In this case $x_0 = 58$ and Y_0 is the energy density for the shipment.

Assumptions for prediction

As previously, we assume the linear regression model

$$y_i = \beta_0 + \beta_1 x_i + e_i$$

where e_1, e_2, \dots, e_n are independent realisation with $e_i \sim N(0, \sigma^2)$.

Assumptions (continued)

In addition, Y_0 is assumed to have the normal distribution,

$$Y_0 \sim N(\beta_0 + \beta_1 x_0, \sigma^2)$$

independently of

$$(x_1, y_1), (x_2, y_2), \dots, (x_n, y_n).$$

This assumption can be expressed equivalently as

$$Y_0 = \beta + \beta_1 x_0 + \mathcal{E}_0$$

where $\mathcal{E}_0 \sim N(0, \sigma^2)$.

In other words, it is assumed the pair (x_0, Y_0) will be an observation from the same regression model that generated the observed data.

Confidence intervals

The parameter $\mu_{Y|x_0} = \beta_0 + \beta_1 x_0$ may be considered as a population mean.

In particular, $\mu_{Y|x_0}$ is the mean of the response Y in the sub-population defined by the condition $X = x_0$.

Example

In the incinerator example, we consider $x_0 = 58$.

- ▶ The quantity $\beta_0 + \beta_1 \times 58$ corresponds to the (population) mean energy density for the sub-population of shipments that have water content 58%.

Confidence intervals (continued)

In previous contexts, population means have been estimated using sample means.

In the regression context, the data needed to calculate a sample mean are not available.

Instead, the regression equation can be used to estimate $\mu_{Y|x_0}$ indirectly by calculating

$$\hat{y}_0 = \hat{\beta}_0 + \hat{\beta}_1 x_0.$$

The standard error can be shown to be

$$\text{SE}(\hat{y}_0) = s_e \sqrt{\frac{1}{n} + \frac{(x_0 - \bar{x})^2}{S_{xx}}}$$

and a $100(1 - \alpha)\%$ confidence interval is

$$\hat{y}_0 \pm t^* \text{SE}(\hat{y}_0)$$

Example

Taking $x_0 = 58$ in the incinerator example, and recalling that $\hat{\beta}_0 = 3,412.2$ and $\hat{\beta}_1 = -42.18$, the estimate for $\mu_{Y|X=58}$ is

$$\hat{y}_0 = 3,412.2 - 42.18 \times 58 = 965.6.$$

To construct the 95% confidence interval, we obtain

$$t^* = 2.0484, \quad s_e = 67.73, \quad \bar{x} = 50.52, \quad \text{and} \quad S_{xx} = 315.43.$$

The standard error is

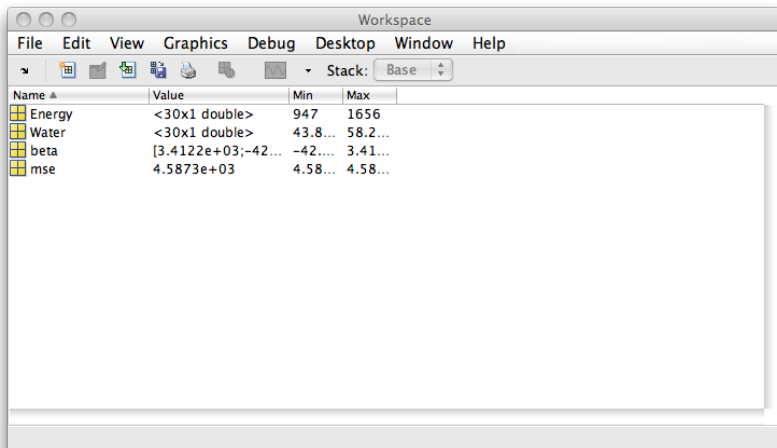
$$\text{SE}(\hat{y}_0) = 67.73 \sqrt{\frac{1}{30} + \frac{(50.52 - 58)^2}{315.43}} = 31.0993.$$

The confidence interval is thus

$$(901.93, 1029.33)$$

Confidence intervals in Matlab

The preceding calculation can be performed easily in Matlab. In what follows, assume the energy data has been loaded and `beta` and `mse` have been generated from the `regstats` command.



The screenshot shows the Matlab Workspace window with the following table of variables:

Name	Value	Min	Max
Energy	<30x1 double>	947	1656
Water	<30x1 double>	43.8...	58.2...
beta	[3.4122e+03;-42...	-42....	3.41...
mse	4.5873e+03	4.58...	4.58...

Matlab (continued)

```
>> n=30;
>> x0=58;
>> yhat=beta(1)+beta(2)*x0;
>> k=tinv(0.975,n-2);
>> xbar=mean(Water);
>> Sxx=sum((Water-xbar).^2);
>> SE=sqrt(mse)*sqrt(1/n+(x0-xbar)^2/Sxx);
>> [yhat,SE]
ans =
    965.6294    31.0993

>> [yhat-k*SE,yhat+k*SE]
ans =
    1.0e+03 *

    0.9019    1.0293
```

Prediction intervals I

The estimate $\hat{y}_0 = \hat{\beta}_0 + \hat{\beta}_1 x_0$ could also be used to predict the value of the (yet to be observed) Y_0 .

This type of prediction is sometimes called a **point prediction** because it is just a single number.

Because $Y_0 \sim N(\beta_0 + \beta_1 x_0, \sigma^2)$ has a continuous distribution, it is extremely unlikely that the realized value will be **exactly** equal to the point prediction.

It is more useful to consider a **prediction interval** for Y_0 .

That is, an interval within which Y_0 will fall with pre-specified probability.

Prediction intervals II

If $Y_0 \sim N(\beta_0 + \beta_1 x_0, \sigma^2)$, then

$$P(\beta_0 + \beta_1 x_0 - z^* \sigma < Y_0 < \beta_0 + \beta_1 x_0 + z^* \sigma) = 1 - \alpha.$$

for z^* such that $P(-z^* < Z < z^*) = 1 - \alpha$.

Prediction intervals (continued)

If the parameters β_0 , β_1 , σ were known, then

$$\beta_0 + \beta_1 x_0 \pm z^* \sigma,$$

would be a $100(1 - \alpha)\%$ prediction interval for Y_0 .

In practice, the parameters are not known and the $100(1 - \alpha)\%$ prediction interval for Y_0 is given by

$$\hat{\beta}_0 + \hat{\beta}_1 x_0 \pm t^* s_e \sqrt{1 + \frac{1}{n} + \frac{(x_0 - \bar{x})^2}{S_{xx}}}$$

where t^* is chosen from the t_{n-2} reference distribution. This interval is derived by:

- ▶ Using $\hat{\beta}_0$, $\hat{\beta}_1$, s_e in place of β_0 , β_1 , σ ;
- ▶ Making allowance for errors in those estimates;
- ▶ The detailed derivation is beyond the scope of this course.

Example I

To obtain a 95% prediction interval for the energy density of a single shipment of garbage with water content 58%, the calculations are similar to those for the confidence interval.

The only difference is in the term

$$\sqrt{\mathbf{1} + \frac{1}{n} + \frac{(x_0 - \bar{x})^2}{S_{xx}}}$$

As previously, $x_0 = 58$ so that $\hat{y}_0 = 965.6$ and

$$t^* = 2.0484, \quad s_e = 67.73, \quad \bar{x} = 50.52, \quad \text{and} \quad S_{xx} = 315.43.$$

Example II

The multiplier term is

$$\sqrt{1 + \frac{1}{n} + \frac{(x_0 - \bar{x})^2}{S_{xx}}} = \sqrt{1 + \frac{1}{30} + \frac{(50.52 - 58)^2}{315.43}} = 1.1004$$

and the prediction interval is

$$(812.97, 1118.29).$$

Prediction intervals in Matlab

```
>> %Same commands as for confidence interval
>> n=30; x0=58; yhat=beta(1)+beta(2)*x0;
>> k=tinv(0.975,n-2); xbar=mean(Water);
>> Sxx=sum((Water-xbar).^2);
>> %The only difference is the standard error term
>> SE_Predict=sqrt(mse)*sqrt(1+1/n+(x0-xbar)^2/Sxx);
>> [yhat,SE_Predict]
ans =

    9.6563e+02    7.4528e+01

>> [yhat-k*SE_Predict,yhat+k*SE_Predict]
ans =

    8.1297e+02    1.1183e+03
```

Comparison I

For the incinerator example, we have considered shipments with water content 58% and calculated:

Confidence interval: (901.93, 1029.33)

Prediction interval: (812.97, 1118.29)

The meanings of these intervals are very different.

- ▶ The confidence interval describes the accuracy of our best estimate of the parameter $\mu_{Y|X=58}$. That is, for the mean energy density for the sub-population of all shipments with water content 58%.
- ▶ The prediction interval specifies, for a single shipment with 58% water content, a range of values for the energy density that has probability 95%.

Comparison II

- ▶ The prediction interval is always wider than the confidence interval.
- ▶ Confusing the two intervals can be a serious error.

Extrapolation

When using regression for prediction, it is important to be aware of the X values in the data set.

Generally, it is advisable to make predictions only for values of x_0 within the observed range

$$x_1, x_2, \dots, x_n.$$

Making predictions beyond the observed range of the data is sometimes called **extrapolation**.

The problem with extrapolation is that large errors may occur if the model breaks down outside the observed range.

If extrapolation is used there is no way to determine whether the predictions are reliable.

The coefficient of determination R^2 I

The coefficient of determination for simple linear regression is defined by

$$R^2 = \frac{\sum_{i=1}^n (\hat{y}_i - \bar{y})^2}{\sum_{i=1}^n (y_i - \bar{y})^2}.$$

Recall that $\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_i$ for $i = 1, 2, \dots, n$.

The value of R^2 can be obtained from Matlab by ticking the R-Square Statistic checkbox in regstats.

The coefficient of determination R^2 II

It can be shown mathematically in the case of simple linear regression that $R^2 = r^2$. That is, the square of the sample correlation coefficient r defined in lecture 5.

For the incinerator example, $R^2 = 0.8138 = 81.38\%$ (from Matlab).

R^2 (continued) I

Before using the regression model to obtain predictions, it is good practice to check that it is appropriate.

- ▶ The residuals vs fitted values plot and the normal quantile plot of the residuals should be checked to determine whether the regression assumptions are reasonable.
- ▶ A hypothesis test can be used to determine whether there is a significant relationship between Y and X .

When a regression model is found to be appropriate for prediction, R^2 can be used as a measure of the **predictive strength** of the model.

- ▶ $0 < R^2 < 1$.

R^2 (continued) II

- ▶ Small values of R^2 indicate low predictive strength.
- ▶ Large values of R^2 indicate high predictive strength.
- ▶ An R^2 of 100% would mean that X can predict Y without error.

Predictive strength

Consider regression data

$$(x_1, y_1), (x_2, y_2), \dots, (x_n, y_n)$$

A $100(1 - \alpha)\%$ prediction interval for Y_0 is

$$\hat{y}_0 \pm t^* s_e \sqrt{1 + \frac{1}{n} + \frac{(x_0 - \bar{x})^2}{S_{xx}}}$$

Provided n is fairly large and if $x_0 - \bar{x}$ is not too large, then

$$\sqrt{1 + \frac{1}{n} + \frac{(x_0 - \bar{x})^2}{S_{xx}}} \approx 1.$$

In this case the prediction interval is roughly

$$\hat{y}_0 \pm t^* s_e$$

Predictive strength (continued)

Suppose now, that we have to make a prediction for Y_0 **without knowledge of** x_0 . That is, not using regression.

Our best guess would be \bar{y} and the $100(1 - \alpha)\%$ prediction limits would be

$$\bar{y} \pm t^* s_y$$

where

$$s_y^2 = \frac{1}{n-1} \sum_{i=1}^n (y_i - \bar{y})^2.$$

The regression based prediction interval will therefore be narrower by a factor of roughly s_e/s_y .

- ▶ $s_e/s_y \approx 1$ would imply that the prediction interval obtained using regression is just as wide as that obtained without regression.

Predictive strength and R^2 I

The interpretation of R^2 as a measure of predictive strength arises from the equality

$$\frac{s_e}{s_y} = \sqrt{\frac{n-1}{n-2}(1-R^2)} \approx \sqrt{1-R^2}.$$

- ▶ If R^2 is large then the ratio will be small, so the regression based prediction will be significantly narrower.
 - ▶ For the incinerator example, $R^2 = 0.837$ and $\sqrt{1-R^2} = 0.432$.
 - ▶ Therefore knowing the water content of a shipment significantly reduces the uncertainty about its energy content.

Predictive strength and R^2 II

- ▶ On the other hand, for a small value such as $R^2 = 0.25$, the width would be reduced by a factor of $\sqrt{1 - 0.25} = 0.867$.
 - ▶ In many practical contexts, this improvement in accuracy would be considered too small to justify the use of regression.