Statistics

MATH2089

UNSW SYDNEY

MATHEMATICS & STATISTICS
THE UNIVERSITY OF NEW SOUTH WALES

Semester 1, 2018 – Lecture 6

This lecture

5. Special random variables

- 5.6 Normal random variables
- 5.7 Checking if the data are normally distributed

6. Sampling distributions

- 6.1 Introduction
- 6.2 Estimation and sampling distribution

Additional reading:
Sections 1.4, 2.4, 5.5 and 5.6 (pp.228-230) in the textbook (2nd edition)
Sections 1.4, 2.4, 5.5 and 5.6 (pp.233) in the textbook (3rd edition)

Revision: the Normal distribution

A random variable is said to be normally distributed with parameters μ and σ ($\sigma > 0$), i.e.

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

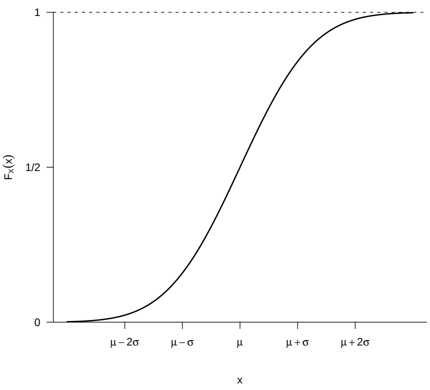
if its probability density function is given by

$$f(x) = \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{(x-\mu)^2}{2\sigma^2}} \qquad (\rightarrow S_X = \mathbb{R})$$

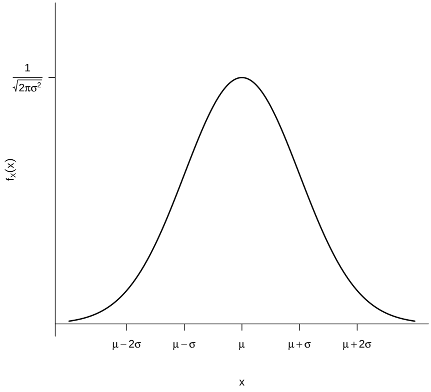
Unfortunately, no closed form exists for

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

Revision: the Normal distribution



cdf $F(x)$



pdf $f(x) = F'(x)$

Normal distribution: properties

It can be shown that, for any μ and σ,

$$\int_{S_X} f(x)\,dx = \int_{-\infty}^{+\infty} \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{(x-\mu)^2}{2\sigma^2}}\,dx = 1$$

Similarly, we can find

$$\mathbb{E}(X) = \frac{1}{\sqrt{2\pi}\sigma} \int_{-\infty}^{+\infty} x\, e^{-\frac{(x-\mu)^2}{2\sigma^2}}\,dx = \mu$$

and

$$\mathbb{V}\mathrm{ar}(X) = \frac{1}{\sqrt{2\pi}\sigma} \int_{-\infty}^{+\infty} (x-\mu)^2\, e^{-\frac{(x-\mu)^2}{2\sigma^2}}\,dx = \sigma^2$$

Mean and variance of the Normal distribution

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

$\mathbb{E}(X) = \mu$ and $\mathbb{V}\mathrm{ar}(X) = \sigma^2$ $(\rightarrow \mathrm{sd}(X) = \sigma)$

The Standard Normal distribution

The **Standard Normal distribution** is the Normal distribution with $\mu = 0$ and $\sigma = 1$. This yields

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

Usually, in this situation, the specific notation

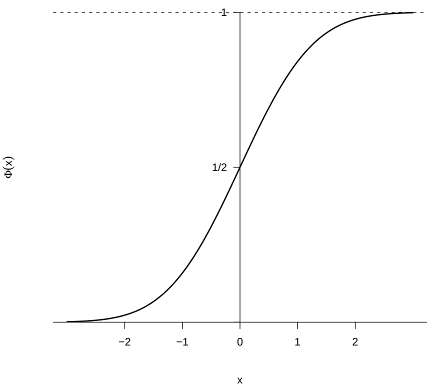
$$f(x) \doteq \phi(x) \qquad \text{and} \qquad F(x) \doteq \Phi(x)$$

is used, and a standard normal random variable is usually denoted Z.

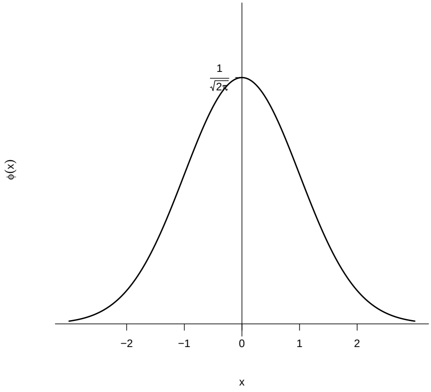
It directly follows from the preceding that

$$\mathbb{E}(Z) = 0 \qquad \text{and} \qquad \mathbb{V}\text{ar}(Z) = 1 \quad (= \text{sd}(Z))$$

The Standard Normal distribution



cdf $\Phi(x)$

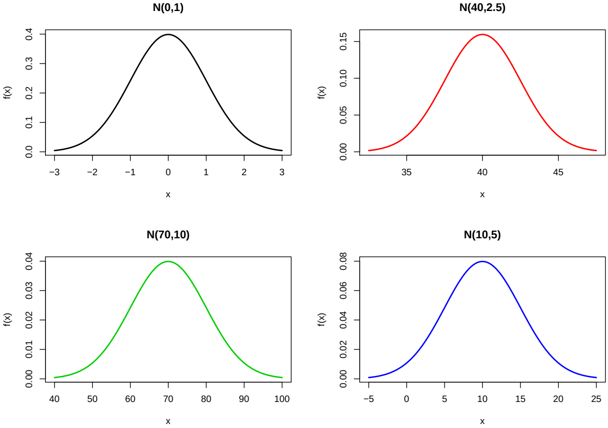


pdf $\phi(x) = \Phi'(x)$

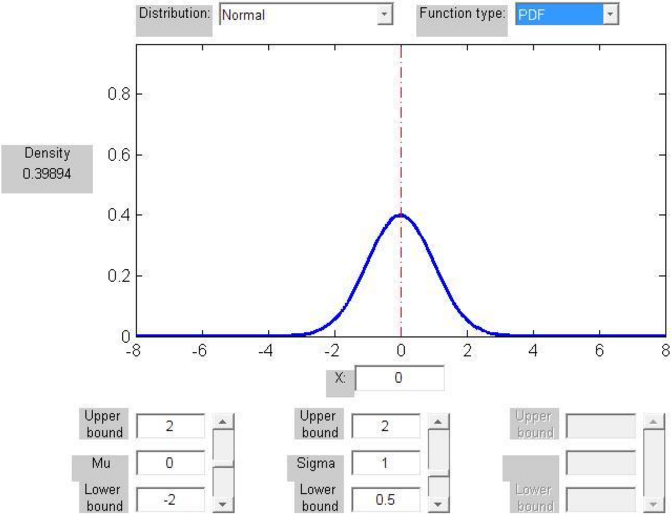
Normal distribution: properties

An important observation is that all normal probability distribution functions have the same bell shape

They only differ in where they are centred (at μ) and in their spread (quantified by σ).



In Matlab, key in `disttool` and play with the interactive probability function display tool (see Lab Week 6)



Normal distribution: standardisation

It is clear from the expression and the shape of the Normal pdf that if $X \sim \mathcal{N}(\mu, \sigma)$, then $Y = aX + b$ is normally distributed with mean $\mathbb{E}(Y) = a\mu + b$ and variance $\mathbb{V}\mathrm{ar}(Y) = a^2\sigma^2$.

The following result directly follows from the foregoing:

Property: Standardisation

If $X \sim \mathcal{N}(\mu, \sigma)$, then

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

This linear transformation is called the **standardisation** of the normal random variable X, as it transforms X into a standard normal random variable Z.

Standardisation will play a paramount role in what follows.

Normal distribution: properties

This extremely important fact allows us to deduce any required information for a given Normal distribution $\mathcal{N}(\mu, \sigma)$ from the features of the 'simple' standard normal distribution.

For instance, for the standard pdf $\phi(x)$, it can be found that

$$\begin{aligned}
\int_{-1}^{1} \phi(x)\, dx &= \mathbb{P}(-1 < Z < 1) \simeq 0.6827 \\
\int_{-2}^{2} \phi(x)\, dx &= \mathbb{P}(-2 < Z < 2) \simeq 0.9545 \\
\int_{-3}^{3} \phi(x)\, dx &= \mathbb{P}(-3 < Z < 3) \simeq 0.9973
\end{aligned}$$

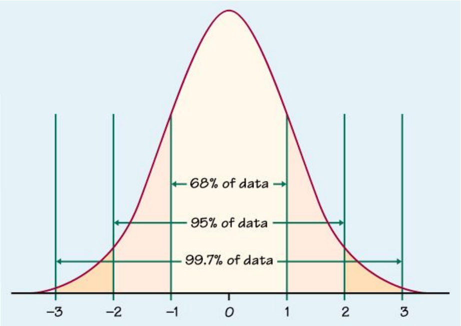
This automatically translates to the general case $X \sim \mathcal{N}(\mu, \sigma)$:

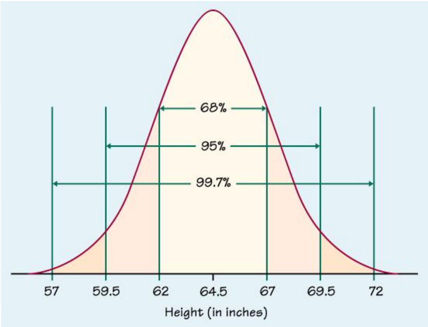
$$\begin{aligned}
\mathbb{P}(\mu - \sigma < X < \mu + \sigma) &\simeq 0.6827 \\
\mathbb{P}(\mu - 2\sigma < X < \mu + 2\sigma) &\simeq 0.9545 \\
\mathbb{P}(\mu - 3\sigma < X < \mu + 3\sigma) &\simeq 0.9973
\end{aligned}$$

This is known as the **68-95-99 rule** for normal distributions.

Normal distribution: properties

For instance, suppose we are told that women's heights in a given population follow a normal distribution with mean $\mu = 64.5$ inches and $\sigma = 2.5$ inches





$\rightarrow$ we expect 68.27 % of women to be between $\mu - \sigma = 64.5 - 2.5 = 62$ inches and $\mu + \sigma = 64.5 + 2.5 = 67$ inches tall, etc.

Normal distribution: remark

- Theoretically, the domain of variation S_X of a normally distributed random variable X is $\mathbb{R} = (-\infty, +\infty)$
- However, there is a 99.7% chance to find X between $\mu - 3\sigma$ and $\mu + 3\sigma$
- It almost impossible to find X outside that interval, and virtually impossible to find it much further away from μ

$\rightarrow$ There is in general no problem in modelling the distribution of a positive quantity with a Normal distribution, provided μ is large compared to σ

- Typical examples include weight, height, or IQ of people
- This also explains why 6σ is sometimes called the width of the normal distribution

Normal distribution: examples

Example

Suppose that $Z \sim \mathcal{N}(0,1)$. What is $\mathbb{P}(Z \leq 1.25)$?

In principle, this should be given by

$$\mathbb{P}(Z \leq 1.25) = \Phi(1.25) = \int_{-\infty}^{1.25} \frac{1}{\sqrt{2\pi}} e^{-x^2/2}\, dx$$

However, we know that this integral cannot be evaluated analytically.

$\rightarrow$ we <u>must</u> use software (command `normcdf` in Matlab) or the Standard Normal table*.
This probability is the 'area under the standard normal curve to the left of z', that is

$$\mathbb{P}(Z \leq z) = \Phi(z)$$

Matlab gives: $\mathbb{P}(Z \leq 1.25) = \Phi(1.25) = 0.8944$

* We don't use Standard Normal Table in this course.

Normal distribution: examples

Any other kind of probabilities must be written in terms of $\mathbb{P}(Z \leq z)$.

Example

Suppose that $Z \sim \mathcal{N}(0, 1)$. What is $\mathbb{P}(Z < 1.25)$? What is $\mathbb{P}(Z > 1.25)$? What is $\mathbb{P}(-0.38 \leq Z < 1.25)$?

Normal distribution: examples

Example 1.16 p.38 (textbook)

The time it takes a driver to react to the brake light on a decelerating vehicle follows a Normal distribution having parameters $\mu = 1.25$ sec and $\sigma = 0.46$ sec. In the long run, what proportion of reaction times will be between 1 and 1.75 sec? (**Hint:** $\Phi(1.09) = 0.8621, \Phi(-0.54) = 0.2946$.)

Normal distribution: examples

Example

The actual amount of instant coffee that a filling machine puts into “4-ounce” jars may be looked upon as a random variable having a normal distribution with $\sigma = 0.04$ ounce. If only 2% of the jars are to contain less than 4 ounces, what should be the mean fill of these jars? (**Hint:** $\Phi(-2.05) = 0.02$.)

Normal distribution: quantiles

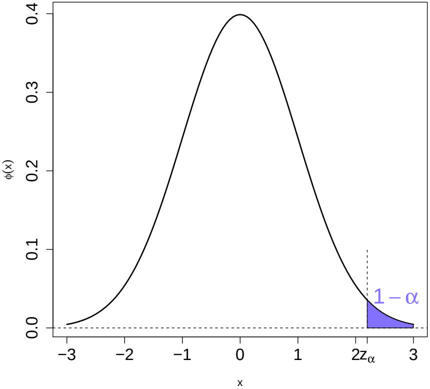
As in the previous example, we are sometimes given a probability and asked to find the corresponding value z

For instance, for any $\alpha \in (0,1)$, let z_α be such that

$$\mathbb{P}(Z > z_\alpha) = 1 - \alpha.$$

$$\text{i.e.,} \quad \mathbb{P}(Z < z_\alpha) = \alpha,$$

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

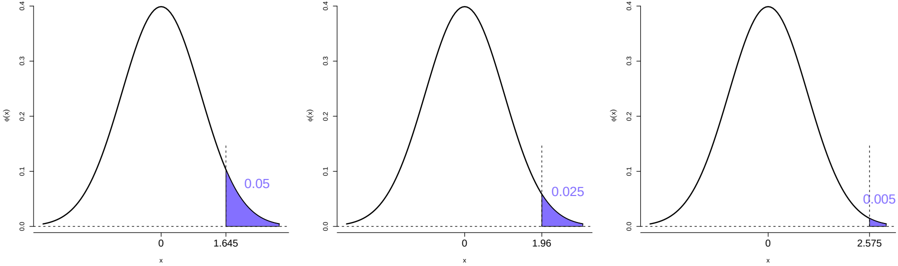


This value z_α is called the **quantile** of level α of the standard normal distribution

Normal distribution: quantiles

Some particular quantiles will be used extensively in subsequent chapters. These are the quantiles of level 0.95, 0.975 and 0.995:

$$\mathbb{P}(Z > 1.645) = 0.05, \quad \mathbb{P}(Z > 1.96) = 0.025, \quad \mathbb{P}(Z > 2.575) = 0.005$$



Note: by symmetry of the normal pdf, it is easy to see that

$$\boxed{z_{1-\alpha} = -z_\alpha}$$

$\rightarrow$ for instance, $\mathbb{P}(Z < -1.96) = 0.025$

Some further properties of the Normal distribution

We know that if $X \sim \mathcal{N}(\mu, \sigma)$, then $aX + b$ is also normally distributed, for any real values a and b.

This generalises further: if $X_1 \sim \mathcal{N}(\mu_1, \sigma_1)$ and $X_2 \sim \mathcal{N}(\mu_2, \sigma_2)$, and X_1 and X_2 are independent, then $aX_1 + bX_2$ is also normally distributed for any real values a and b.

Also, we can compute the parameters of the resulting distribution.

Property

Suppose $X_1 \sim \mathcal{N}(\mu_1, \sigma_1)$, $X_2 \sim \mathcal{N}(\mu_2, \sigma_2)$ and X_1 and X_2 are **independent**. Then, for any real values a and b,

$$aX_1 + bX_2 \sim \mathcal{N}\left(a\mu_1 + b\mu_2, \sqrt{a^2\sigma_1^2 + b^2\sigma_2^2}\right)$$

Some further properties of the Normal distribution

As a direct application of the preceding property, we have, with $X_1 \sim \mathcal{N}(\mu_1, \sigma_1)$ and $X_2 \sim \mathcal{N}(\mu_2, \sigma_2)$, X_1 and X_2 independent,

$$X_1 + X_2 \sim \mathcal{N}\left(\mu_1 + \mu_2, \sqrt{\sigma_1^2 + \sigma_2^2}\right), \quad X_1 - X_2 \sim \mathcal{N}\left(\mu_1 - \mu_2, \sqrt{\sigma_1^2 + \sigma_2^2}\right)$$

Besides, the previous property can be readily extended to an arbitrary number of independent normally distributed random variables.

Example

Let X_1, X_2, X_3 represent the times necessary to perform three successive repair tasks at a certain service facility. Suppose they are independent normal random variables with expected values μ_1, μ_2 and μ_3 and variances σ_1^2, σ_2^2 and σ_3^2, respectively. What can be said about the distribution of $X_1 + X_2 + X_3$?

From the previous property, we can conclude that

$$X_1 + X_2 + X_3 \sim \mathcal{N}\left(\mu_1 + \mu_2 + \mu_3, \sqrt{\sigma_1^2 + \sigma_2^2 + \sigma_3^2}\right)$$

Some further properties of the Normal distribution

Example (ctd.)

If $\mu_1 = 40$ min, $\mu_2 = 50$ min and $\mu_3 = 60$ min, and $\sigma_1^2 = 10$ min^2, $\sigma_2^2 = 12$ min^2 and $\sigma_3^2 = 14$ min^2, what is the probability that the full task would take less than 160 min? (**Hint:** $\Phi(1.67) = 0.9525$.)

Checking if the data are normally distributed

Fact

Many of the statistical techniques presented in the coming chapters are based on an assumption that the distribution of the random variable of interest is (almost) normal.

$\rightarrow$ In many instances, we will need to check whether a data set appears to be generated by a normally distributed random variable

How do we do that ?

Although they involve an element of subjective judgement, graphical procedures are the most helpful for detecting serious departures from normality.

Some of the visual displays we have used earlier, such as the density histogram, can provide a first insight about the form of the underlying distribution.

Density histograms to check for normality

Think of a density histogram as a piecewise constant function $h_n(x)$, where n is the number of observations in the data set

$\rightarrow$then, if the r.v. X having generated the data has density f on a support S_X, it can be shown that, under some regularity assumptions, for any $x \in S_X$,

$$h_n(x) \rightarrow f(x)$$

as $n \rightarrow \infty$ (and the number of classes $\rightarrow \infty$)

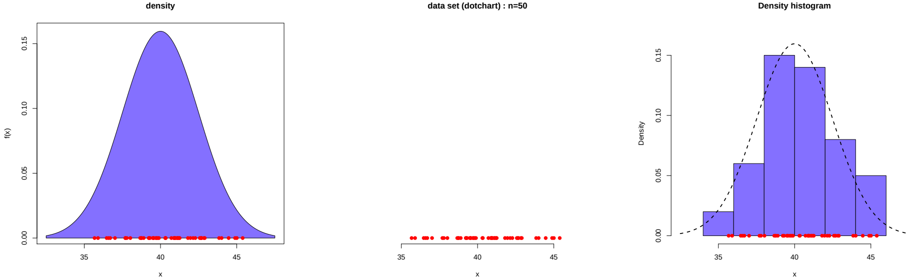
(the convergence "$h_n(x) \rightarrow f(x)$" has to be understood in a particular probabilistic sense, but details are beyond the scope of this course)

Concretely, **the larger the number of observations, the more similar the density histogram and the 'true' (unknown) density f are**

$\rightarrow$ look at the histogram and decide whether it looks enough like the symmetric 'bell-shaped' normal curve or not

Density histograms to check for normality

Suppose we have a data set of size $n = 50$, drawn from a normal distribution

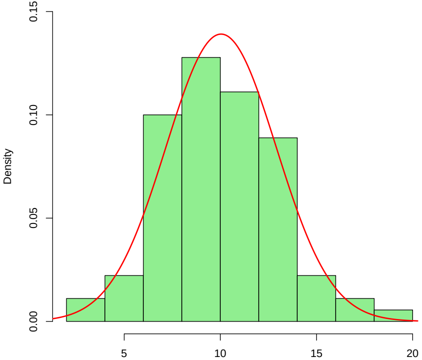


$\rightarrow$ the density histogram 'looks like' the bell-shaped curve

Also, as both $f(x)$ and the density histogram are scaled such that the blue-ish purple areas are 1, they are easily superimposed and compared.

Density histograms to check for normality

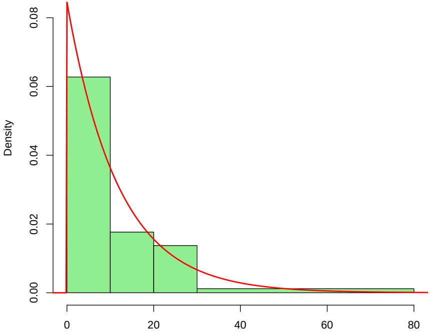
Look again at the (density) histogram on Slide 24, Lecture 2



→ the histogram is symmetric and bell-shaped, without outliers

→ the normality assumption is reasonable

Look again at the density histogram on Slide 33, Lecture 2



→ clear lack of symmetry (skewed to the right)

→ serious departure from normality (Exponential?)

Quantile plots

Density histograms are easy to use; however, they are usually not really reliable indicators of the distributional form unless the number of observations is very large.

$\rightarrow$ another special graph, called a **normal quantile plot**, is more effective in detecting departure from normality

The plot essentially compares the data ordered from smallest to largest with what to expect to get for the smallest to largest in a sample if the theoretical distribution from which the data have come is normal

$\rightarrow$if the data were effectively selected from the normal distribution, the two sets of values should be reasonably close to one another

Note: a quantile plot is also sometimes called **qq-plot**

($\rightarrow$ command in Matlab: `qqplot`)

Quantile plots

Procedure for building a quantile plot:

- observations $\{x_1, x_2, \dots, x_n\}$
- ordered observations: $x_{(1)} \leq x_{(2)} \leq \dots \leq x_{(n)}$
- cumulative probabilities $\alpha_i = \frac{i-0.5}{n}$, for all $i = 1, \dots, n$
- standard normal quantiles of level α_i: for all $i = 1, \dots, n$, z_{α_i} chosen such that $\mathbb{P}(Z \leq z_{\alpha_i}) = \alpha_i$, where $Z \sim \mathcal{N}(0, 1)$
- Quantile plot: plot the n pairs $(x_{(i)}, z_{\alpha_i})$

If the sample comes from the Standard Normal distribution, $x_{(i)} \simeq z_{\alpha_i}$ and the points would fall close to a 45°straight line passing by (0,0).

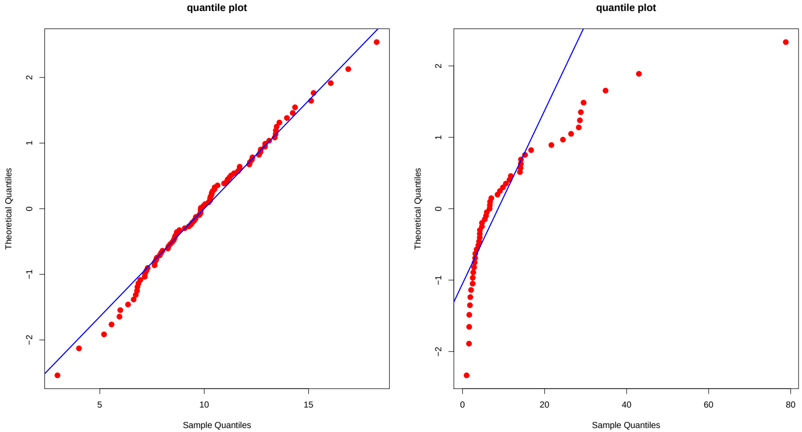
If the sample comes from some other normal distribution, the points would still fall around a straight line, as there is a linear relationship between the quantiles of $\mathcal{N}(\mu, \sigma)$ and the standard normal quantiles.

Fact

If the sample comes from some normal distribution, the points should follow (at least approximately) a straight line.

Quantile plots: examples

The figure below displays quantile plots for the previous two examples



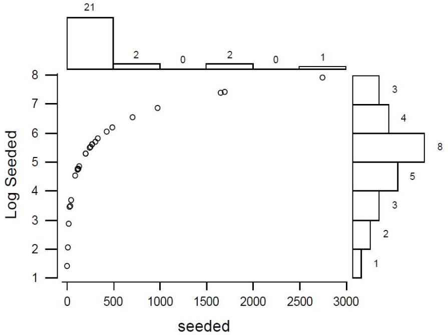
$\rightarrow$ the normality assumption appears acceptable for the first data set, not at all for the second

Transforming observations

When the density histogram and the qq-plot indicate that the assumption of a normal distribution is invalid, transformations of the data can often improve the agreement with normality (see lab classes 2 & 4).

Scientists regularly express their observations in natural logs.

Let's look again at the seeded clouds rainfall data.



Transforming observations

Apart from the $\log$, other transformations may be useful:

$$\frac{-1}{x}, \quad \sqrt{x}, \quad \sqrt[4]{x}, \quad x^2, \quad x^3$$

If the observations are positive and the distribution has a long tail on the right, then concave transformations like $\log(x)$ or $\sqrt{x}$ put the large values down farther than they pull the central or small values
$\rightarrow$ observations 'more symmetric'

Convex transformations work the other way

If the transformed observations are approximately normal (check with a quantile plot), it is usually advantageous to use the normality of this new scale to perform any statistical analysis.

6. Sampling distributions

Statistical Inference: Introduction

In this chapter, we introduce the last main topic for this course: **statistical inference** (that will keep us busy until the end).

Recall (Lecture 1) the general problem that is addressed:

- statistical methods are used to draw conclusions and make decisions about a population of interest
- however, for some reasons, we have no access to the whole population and we must do with observations on a subset of the population only. That subset is called the sample
- if the sample is effectively representative of the population, what we observe on the sample can be generalised to the population as a whole, at least to some extent ...
- ... taking chance factors properly into account

$\rightarrow$ what we have learned about descriptive statistics, probability and random variables in the previous chapters, will play important roles here

Statistical Inference: Introduction

Populations are often described by the distribution of their values
→ for instance, it is quite common practice to refer to a 'normal population', when the variable of interest is thought to be normally distributed

In statistical inference, we focus on drawing conclusions about one parameter describing the population.

In engineering, the parameters we are mainly interested in are

- the **mean** μ of the population
- the **variance** σ^2 (or standard deviation σ) of the population
- the **proportion** π of individuals in the population that belong to a class of interest
- the **difference in means of two sub-populations,** $\mu_1 - \mu_2$
- the **difference in two sub-population proportions,** $\pi_1 - \pi_2$

These parameters are unknown (otherwise, no need to make inferences about them) → the first part of the process is thus to estimate the unknown parameters.

Random sampling

The importance of **random sampling** has been emphasised in Lecture 1:

- to assure that a sample is representative of the population from which it is obtained, and
- to provide a framework for the application of probability theory to problems of sampling

As we said, the assumption of random sampling is very important: if the sample is not random and is based on judgement or flawed in some other way, then statistical methods will not work properly and will lead to incorrect decisions.

Therefore, we should now properly define a random sample.

Random sampling

Before a sample of size n is selected at random from the population, the observations are modelled as random variables $X_1, X_2, \dots, X_n$.

Definition

The set of observations $X_1, X_2, \dots, X_n$ constitutes a random sample if

1. the X_i's are independent random variables, and
2. every X_i has the same probability distribution

This is often abbreviated to i.i.d., for 'independent and identically distributed'
$\rightarrow$ it is common to talk about an i.i.d. sample

We also apply the terms 'random sample' to the set of observed values

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

of the random variables, but this should not cause any confusion.

Note: as usual, the lower case distinguishes the realisation of a random sample (the actual data) from the upper case, which represents the random variables before they are observed

Statistic, estimator and sampling distribution

- A numerical measure calculated from the sample is called a **statistic**
- Denote the unknown parameter of interest θ (so this can be μ, σ^2, or any other parameter of interest)
- The only information we have to estimate that parameter θ is the information contained in the sample

$\rightarrow$ an **estimator** of θ must be a statistic, i.e. a function of the sample

$$\hat{\Theta} = h(X_1, X_2, \dots, X_n)$$

- Note that an estimator is a random variable, as it is a function of random variables $\rightarrow$ it must have a distribution
- That distribution is called a **sampling distribution**, it generally depends on the population distribution and the sample size
- After the sample has been selected, $\hat{\Theta}$ takes on a particular value $\hat{\theta} = h(x_1, x_2, \dots, x_n)$, called the **estimate** of θ

Estimation: some remarks

Remark: the hat notation conventionally distinguishes the sample-based quantities (estimator $\hat{\Theta}$ or estimate $\hat{\theta}$) from the 'true' population parameter (θ). Also, as usual, capital letters denote the random variables, like $\hat{\Theta}$, whereas lower-case letters are for particular numerical values, like $\hat{\theta}$.

$\rightarrow$ two notable exceptions are:

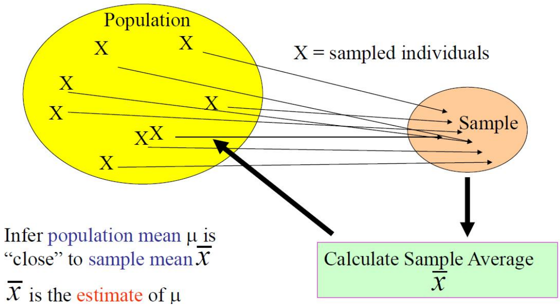
- the sample mean, usually denoted $\bar{X}$; its observed value, calculated once we have observed a sample $x_1, x_2, \ldots, x_n$, is denoted $\bar{x}$
- the sample standard deviation (variance), usually denoted S (S^2); its observed value, calculated once we have observed a sample $x_1, x_2, \ldots, x_n$, is denoted s (s^2)

An example: estimating μ in a normal population

Suppose that the random variable X of interest is normally distributed, with unknown mean μ and *known* standard deviation σ.

From a random sample, a natural estimator for μ is the sample mean $\bar{X} = \frac{1}{n}\sum_{i=1}^{n} X_i$, whose observed value is $\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$ (= estimate).

As the sample is random, it should be representative of the whole population, and the population mean μ should be "close" to the observed sample mean $\bar{x}$.



An example: estimating μ in a normal population

What does that mean, μ should be "close" to $\bar{x}$?

$\rightarrow$ the sampling distribution will answer this question

We know that each X_i in the sample follows the $\mathcal{N}(\mu, \sigma)$ distribution, and they are independent (i.i.d. sample).

Then, because linear combinations of independent normal r.v. remain normally distributed, we conclude that

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

has a normal distribution with expectation

$$\mathbb{E}\left(\bar{X}\right) = \mathbb{E}\left(\frac{1}{n}\sum_{i=1}^{n} X_i\right) = \frac{1}{n}\sum_{i=1}^{n}\mathbb{E}(X_i) \overset{\text{i.d.}}{=} \frac{1}{n}\sum_{i=1}^{n}\mu = \mu$$

and variance

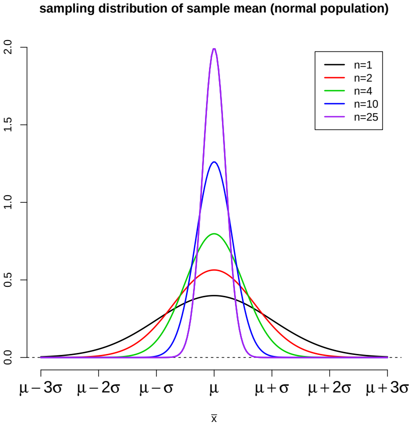
$$\mathbb{V}\text{ar}\left(\bar{X}\right) = \mathbb{V}\text{ar}\left(\frac{1}{n}\sum_{i=1}^{n} X_i\right) \overset{\text{i.}}{=} \frac{1}{n^2}\sum_{i=1}^{n}\mathbb{V}\text{ar}(X_i) \overset{\text{i.d.}}{=} \frac{1}{n^2}\sum_{i=1}^{n}\sigma^2 = \frac{\sigma^2}{n}$$

An example: estimating μ in a normal population

$\rightarrow$ the sampling distribution of $\bar{X}$ is

$$\bar{X} \sim \mathcal{N}\left(\mu, \frac{\sigma}{\sqrt{n}}\right)$$

if the population is normal



$\bar{X}$ is a normal random variable, centred about the 'true' population mean μ, and with spread becoming more and more reduced as the sample size increases

$\rightarrow$ **the larger the sample, the more accurate the estimation is**

Objectives

Now you should be able to:

- Calculate probabilities, determine mean, variance and standard deviation for normal distributions ☐
- Standardise normal random variables, and understand why this is useful ☐
- Use the cdf of a standard normal distribution to determine probabilities of interest ☐
- Explain the general concepts of estimating the parameters of a population, in particular the difference between estimator and estimate, and the role played by the sampling distribution of an estimator ☐
- Illustrate those concepts with the particular case of the estimation of the mean in a normal population

Recommended exercises

$\rightarrow$ Q31, Q33 p.41, Q35, Q37 p.42, Q73 p.58, Q27 p.78, Q47 p.93, Q65 p.97, Q51, Q53 p.237 (2nd edition)

$\rightarrow$ Q31 p.44, Q33, Q35, Q37 p.45, Q75 p.60, Q28 p.79, Q66 p.100, Q53, Q55 p.241 (3rd edition)