Part (C): The Normal Distribution and Data Summaries

10 The Normal (Gaussian) distribution

(Reading: 'Chance Encounters', Seber & Wild, 6.1 - 6.3)

Some historical notes:

The normal curve was developed mathematically in 1733 by DeMoivre as an approximation to the binomial distribution. His paper was not discovered until 1924 by Karl Pearson. Laplace used the normal curve in 1783 to describe the distribution of errors. Subsequently, Gauss used the normal curve to analyze astronomical data in 1809.

The normal distribution is the most used statistical distribution. The principal reasons are:

1. Normally distributed random variables arise naturally in the context of measurements such as height, weight, blood pressure, etc ...
2. Normality is important in statistical inference.

Often, if we take such a measurement over a large group of subjects, the resulting histogram has a characteristic shape.

Example:

In 1836, the chest girths of 1516 US soldiers were measured. The distribution of the measurements is shown in the following histogram.

Note:

(i) the symmetry of the histogram about the central value (35")

(ii) the 'bell' shape of the observed distribution.

If we choose a soldier at random from the 1516 they are just as likely to have chest girth $> 35''$ as $< 35''$. [due to (i)]
It is very unlikely to be $> 40''$ or $< 30''$. [due to (ii)]

The Normal distribution is often used to represent distributions such as the above.

10.1 Probability density function of the Normal distribution

If X is a normal random variable, its p.d.f. is given by

$$f_X(x) = \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{1}{2\sigma^2}(x-\mu)^2}, \qquad -\infty < x < \infty \tag{20}$$

$$-\infty < \mu < \infty, \sigma > 0.$$

The p.d.f. depends on the two basic parameters of the distribution, the mean

$$E(X) = \mu$$

and the variance

$$\text{var}(X) = \sigma^2.$$

For a normal r.v. with mean μ and variance σ^2 we use the notation

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

If we superimpose the graph of the normal p.d.f. with $\mu = 35, \sigma^2 = 4$ on the chest-girth histogram we obtain a good fit. The density looks like a 'smoother' version of the histogram.

10.2 Important properties

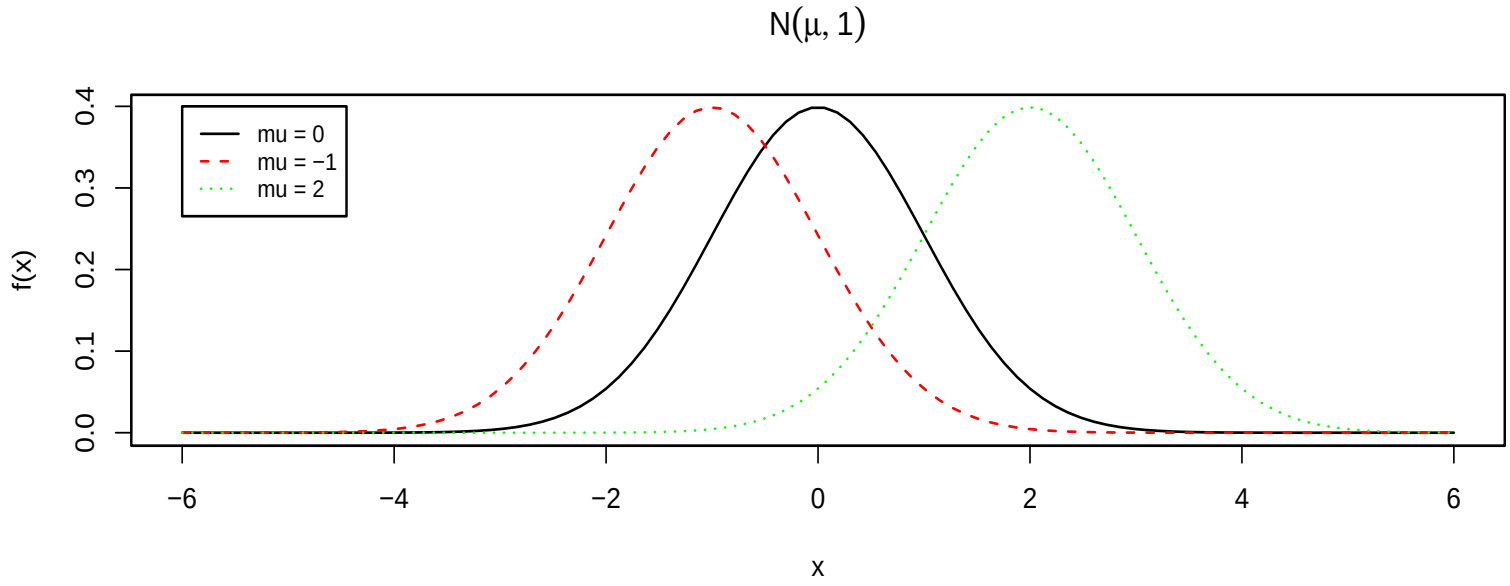
1. The shape of the distribution

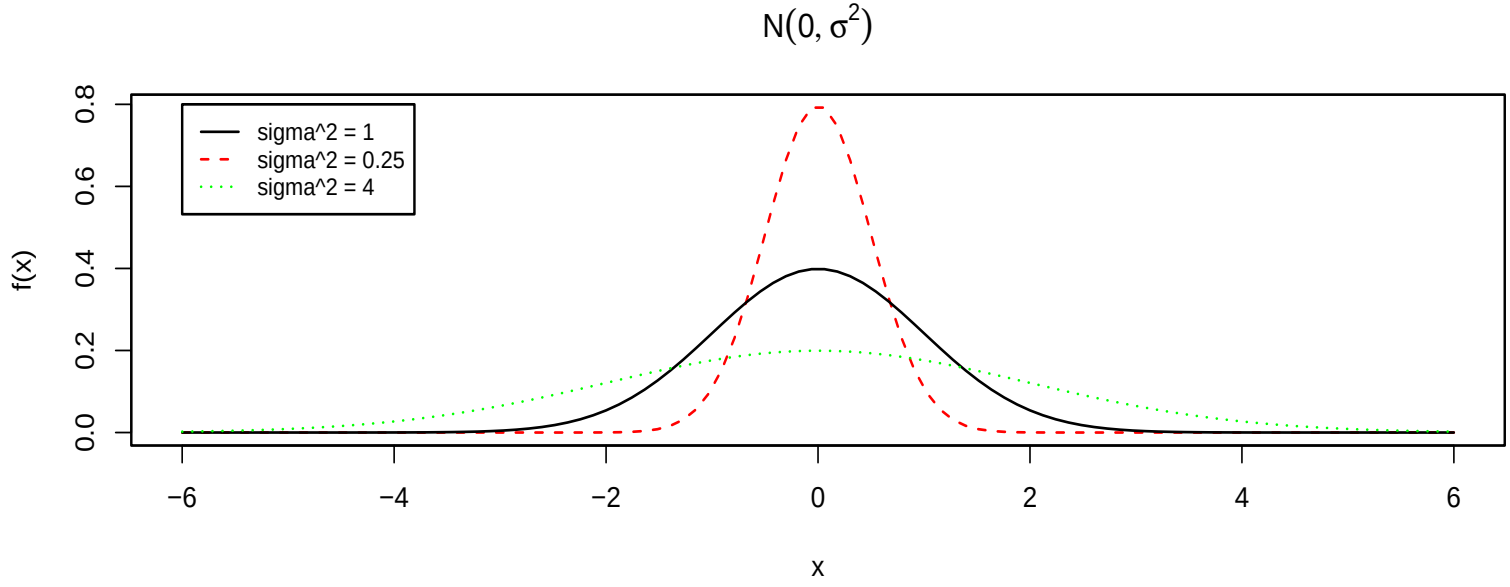
All normal densities have a 'bell-shaped' graph. They differ only in location (μ) and spread (σ).

The graph of the $N(\mu, \sigma^2)$ density is symmetric about the mean μ.

If σ is large the distribution ('bell-shape') will be wide.

If σ is small the distribution will be narrow.


N(μ, 1)
mu = 0
mu = −1
mu = 2
f(x)
0.0
0.1
0.2
0.3
0.4
−6
−4
−2
0
2
4
6
x


N(0, σ²)
sigma^2 = 1
sigma^2 = 0.25
sigma^2 = 4
f(x)
0.0
0.2
0.4
0.6
0.8
−6
−4
−2
0
2
4
6
x

2. The standard normal distribution

The distribution $N(0,1)$ is referred to as the standard Normal distribution. Its p.d.f. is given as

$$f_Z(z) = \frac{1}{\sqrt{2\pi}}\, e^{-\frac{z^2}{2}}, \qquad -\infty < z < \infty. \tag{21}$$

The distribution function (CDF) of a $N(0,1)$ r.v. is usually denoted by $\Phi(\cdot)$, i.e.

$$\Phi(z) = P(Z \leq z) = \int_{-\infty}^{z} \frac{1}{\sqrt{2\pi}}\, e^{-\frac{t^2}{2}}\, dt \tag{22}$$

and its values are given in statistical tables.

$N(0,1)$ **table values** (e.g. *NCST*, p.34)

z	$\Phi(z) = P(Z \leq z)$
0.00	0.5000
1.00	0.8413
1.65	0.9505
1.96	0.9750
2.33	0.9901
3.30	0.9995

Tables also give values of the **inverse** $N(0,1)$. In *NCST* (p.35) these are the values z such that

$$1 - \Phi(z) = p \Rightarrow P(Z > z) = p.$$

p	z
0.10	1.2816
0.05	1.6449
0.025	1.9600
0.01	2.3263
0.005	2.5758

The $N(0,1)$ distribution is symmetric around zero, and therefore

- $f(-a) = f(a)$
- $\Phi(0) = 0.5$
- $\Phi(-a) = P(Z \leq -a) = P(Z \geq a) = 1 - \Phi(a)$
- $P(-a \leq Z \leq a) = \Phi(a) - \Phi(-a) = 2\Phi(a) - 1$

e.g.

$$\begin{aligned} P(X \leq -1.96) &= \Phi(-1.96) = 1 - \Phi(1.96) \\ &= 1 - 0.975 = 0.025. \end{aligned}$$

3. Fundamental result

If $X \sim N(\mu, \sigma^2)$, then the r.v.

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

is a $N(0, 1)$ r.v.

Proof ...

[In general, if $X \sim N(\mu, \sigma^2)$, then $aX + b \sim N(a\mu + b, a^2\sigma^2)$.]

Then:

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

We can use this result to reduce all probability calculations for any normal r.v. to a calculation for the standard normal distribution.

4. Fundamental result 2

If X_1, X_2 are independent r.v$_s$ with

$$X_1 \sim N(\mu_1, \sigma_1^2), \quad X_2 \sim N(\mu_2, \sigma_2^2),$$

then for the r.v. $Y = aX_1 + bX_2$ we have:

$$Y \sim N(a\mu_1 + b\mu_2, a^2\sigma_1^2 + b^2\sigma_2^2).$$

Example

The nominal volume of a can of a particular brand of lager beer is 440 *ml.*

According to the manufacturer, the true volumes are normally distributed with $\mu = 442$ *ml and* $\sigma = 2.5$ *ml.*

What is the probability that a randomly selected can contains less than 438 *ml?*

Solution ...

Example

Suppose that (in a group of people) weights of men are $M \sim N(68, 4)$ *and of women* $W \sim N(65, 1)$ *(both in kg). Select a man and a woman independently at random.*

Find the probability that the weight of the woman is greater than that of the man.

Solution ...

10.3 Normal approximation to the binomial distribution

[Reading: Freund, paragraph 6.6]

If the r.v. $X \sim \text{bin}(n, p)$, then

$$X \overset{approx}{\sim} N(np, np(1-p))$$

as $n \to \infty$.

This implies that

$$Z = \frac{X - np}{\sqrt{np(1-p)}}$$

approaches the $N(0, 1)$ distribution as $n \to \infty$.

In practice we can use this approximation when both np and $np(1-p)$ are greater than 5.

Continuity correction

Remember that the binomial is a discrete distribution. We must allow for the change from a discrete to a continuous r.v.

In practice

- $X = k$ (discrete) is equivalent to $k - \frac{1}{2} < X < k + \frac{1}{2}$ (continuous)
- $X > k$ (discrete) is equivalent to $X > k + \frac{1}{2}$ (continuous)
- $X \geq k$ (discrete) is equivalent to $X > k - \frac{1}{2}$ (continuous)

e.g. If $X \sim \text{Bin}(16, 0.5)$

$$P(X = 6) \approx P(5.5 < Y < 6.5)$$

where $Y \sim N(16 \times 0.5, 16 \times 0.5 \times 0.5)$, i.e.

$$\begin{aligned} P(X = 6) &\approx P(5.5 < Y < 6.5) \\ &= P\left(\frac{5.5 - 8}{2} < Z < \frac{6.5 - 8}{2}\right) \qquad \text{where } Z \sim N(0, 1) \\ &= P(-1.25 < Z < -0.75) = 0.2266 - 0.1056 \qquad (\text{Tables, p.34}) \\ &= 0.1210. \end{aligned}$$

Notice that exact probability is $P(X = 6) = 0.1222$ (using binomial p.m.f.)

Descriptive statistics

- Organisation and presentation of data
  - tables, frequency distributions
  - visual displays: bar charts, histograms, stem-and-leaf plots etc.
- Summary of data
  - measures of location: median, mean, mode etc.
  - measures of spread: range, standard deviation etc.
  - description of shape: e.g. skewness
  - outliers (extreme values), transformations
- Overall summaries
  - e.g. 5-point summaries, boxplots

11 Summarising continuous data: Graphical summaries

This should be the first step to a good statistical analysis.

Graphical summaries help identify the main characteristics of the data, e.g. whether our data look consistent with a Normal distribution.

All graphs should:

- be reasonably accurate
- be on the correct (appropriate) scale
- have good annotation (title,labels on axes, units etc.)

11.1 Dot plots

These are useful for small to medium sample sizes. The data values are plotted as dots along a continuous horizontal axis (in ascending order).

Dot plots are good for showing clusters of points, gaps (where there are no observations) and atypical observations or outliers.

Example

Consider the data:

$\{210, 210, 212, 214, 218, 220, 222, 223, 223, 225, 227, 238, 239\}$.

The dot plot looks like:

11.2 Stem-and-leaf plots

These retain all the information in the data. They are constructed by splitting each observation into two parts:

a	b
preceding digit(s)	last digit
(stem)	(leaf)

Procedure:

i) Identify min and max values for the stem and write down all stems in this range vertically.

ii) For each observation enter the leaf beside the appropriate stem.

We can lengthen the plot (i.e. have more stems) by splitting each stem over two lines. We can also shorten it by appropriate rounding.

Example

Consider the data from previous example:

$\{210, 210, 212, 214, 218, 220, 222, 223, 223, 225, 227, 238, 239\}$.

The stem-and-leaf plot looks like:

11.3 Histograms

Also very useful for representing continuous data graphically.

To form a histogram from the observations $x_1, x_2, \ldots, x_n$ we simply:

i) Identify min and max observations.

ii) Split the range of the values into equal intervals (bins) (e.g. between 5 and 15 bins).

iii) Count the observations falling in each bin.

iv) Draw a rectangle above each interval with height equal to the number of observations in that interval.

Notice that:

- Alternatively we can draw rectangles with height equal to the proportion of observations falling in each bin.
- The intervals (in step ii. above) can have unequal length (perhaps at the extremes of the distribution).
  In that case the area of the rectangles (not their height) should be equal (or proportional) to the bin frequencies.

Histograms and stem-and-leaf plots are very useful for checking whether the data look Normal.

Remember that if this is the case the histogram should look (roughly)

- symmetric
- 'bell-shaped'

Watch out for obvious deviations from this pattern, e.g.

- More than one mode (peak)

  This pattern can occur when a measurement is taken on a population consisting of two distinct sub-populations (e.g. males - females)

- Lack of symmetry (skewness)

12 Summarising continuous data: Numerical summaries

Numerical summaries provide useful and informative measures of the main features of the data.

12.1 Measures of location

1. The sample mean

If $x_1, x_2, \ldots, x_n$ constitute a random sample, then the sample mean is given by

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

2. The sample median

Sort the data in ascending order:

$$x_{(1)} \leq x_{(2)} \leq \ldots \leq x_{(n)}$$

The sample median is simply the value that splits the data into two halves, i.e. the observation with rank $\frac{n+1}{2}$ (ie. lying in position $\frac{n+1}{2}$). We can write

$$\text{sample median} = x_{\left(\frac{n+1}{2}\right)}$$

If n is odd we have

$$x_{(1)} \leq \ldots \leq x_{\left(\frac{n+1}{2}\right)} \leq \ldots \leq x_{(n)}$$

Notice however that if n is even, $\frac{n+1}{2}$ is not an integer and we have

$$x_{(1)} \leq \ldots \leq x_{\left(\frac{n}{2}\right)} \leq x_{\left(\frac{n}{2}+1\right)} \leq \ldots \leq x_{(n)}.$$

In this case the sample median is given as

$$\text{sample median} = \frac{1}{2}\left\{x_{\left(\frac{n}{2}\right)} + x_{\left(\frac{n}{2}+1\right)}\right\}$$

Example

Consider the two data sets $\{9, 12, 12, 13, 17, 20, 25\}$ *and* $\{9, 12, 12, 13, 17, 20, 75\}$. *Find the sample mean and sample variance.*

Solution ...

The median is more robust than the (arithmetic) mean, i.e. is not affected by extreme values in the data set.

Thus the median is more appropriate for skewed data, whereas for symmetric data mean $\approx$ median.

- Positive skewness (tail to the right): mean $>$ median
- Negative skewness (tail to the left): mean $<$ median.

3. The sample mode

The sample mode is given by the most frequently observed value in the data.

In the previous example, the mode was equal to 12.

12.2 Measures of spread (dispersion, variation)

1. The sample variance

Again, if the observations are $x_1, x_2, \ldots, x_n$, the sample variance is given by

$$\begin{aligned} s^2 &= \frac{1}{(n-1)} \sum_{i=1}^{n} (x_i - \bar{x})^2 \\ &= \frac{1}{(n-1)} \left\{ \sum_{i=1}^{n} x_i^2 - \frac{1}{n} \left(\sum_{i=1}^{n} x_i \right)^2 \right\}. \end{aligned}$$

The last expression is usually the easiest way to calculate s^2.

The sample standard deviation is defined as

$$s = \sqrt{\text{sample variance}}.$$

2. The interquartile range (IQR)

We first define the quartiles:

- First quartile: $Q_1 = x_{\left(\frac{n+1}{4}\right)}$

  i.e. it is the value that exceeds exactly the 25% of the observations.

- Third quartile: $Q_3 = x_{\left(\frac{3(n+1)}{4}\right)}$

  i.e. it is the value that exceeds exactly the 75% of the observations.

Use of appropriate interpolation may be needed for the calculation of the quartiles.

Notice that the median can be viewed as the second quartile (Q_2).

The interquartile range is given by

$$\text{IQR} = Q_3 - Q_1$$

3. The five-point summary

Again, sort the data in ascending order:

$$x_{(1)} \leq x_{(2)} \leq \ldots \leq x_{(n)}$$

The five-point summary is given by

min value	1st quartile	median	3rd quartile	max value
$x_{(1)}$	Q_1	Q_2	Q_3	$x_{(n)}$

Example

Find the 5-point summary of the data:

$$1 \quad 2 \quad 2 \quad 5 \quad 8 \quad 8 \quad 11 \quad 12 \quad 14$$

$$1 \quad 2 \quad 2 \quad 5 \quad 8 \quad 8 \quad 11 \quad 12 \quad 14$$

Solution

Median $= x_{\left(\frac{n+1}{2}\right)} = x_{\left(\frac{9+1}{2}\right)} = x_{(5)} = 8.$

$Q_1 = x_{\left(\frac{n+1}{4}\right)} = x_{\left(\frac{9+1}{4}\right)} = x_{(2.5)} = 2.$

$Q_3 = x_{\left(\frac{3(n+1)}{4}\right)} = x_{\left(\frac{3(9+1)}{4}\right)} = x_{(7.5)} = 11.5.$

Therefore we have:

$$\min = 1 \qquad Q_1 = 2 \qquad \text{median} = 8 \qquad Q_3 = 11.5 \qquad \max = 14$$

Also: IQR $= Q3 - Q1 = 9.5.$

12.3 Rules for normal data

Finally we give two approximate rules which appear to conform to a normal distribution (unimodal, symmetric).

Approximately:

- 68% of the data lie between

$$\bar{x} - s \text{ and } \bar{x} + s$$

- 95% of the data lie between

$$\bar{x} - 2s \text{ and } \bar{x} + 2s$$