MA4413 Statistics for Computing

Lecture images/ : Normal Distribution

Kevin O'Brien

kevin.obrien@ul.ie

Dept. of Mathematics & Statistics,
University *of* Limerick

Autumn 2012

Introduction to the Normal Distribution

- Recall the experiment whereby a die was rolled 100 times, and the sum of the 100 values was recorded.
- This experiment was repeated a very large number of times (e.g. 100,000 times) in a simulation study.
- A histogram was drawn to depict the distribution of outcomes of this experiment.
- Recall that we agreed that “bell-shaped” was a good description of the histogram.

Normal Distribution

IMAGE

Normal Distribution

- The normal distribution is perhaps the most widely used distribution for a random variable.
- Normal distributions have the same general shape: the bell curve.
- They are symmetric with scores more concentrated in the middle than in the tails.
- The height of a normal distribution can be defined mathematically in terms of two fundamental parameters: the mean (μ) and the standard deviation (σ).
- A normally distributed random variable X is denoted $X \sim \mathrm{N}(\mu, \sigma^2)$ (note that we use the variance term here)
- The mean and standard deviation are vital for calculating probabilities.

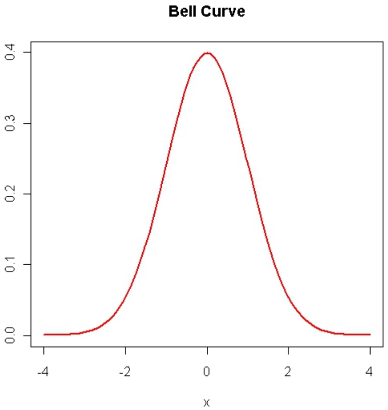
The Normal Distribution

The ***probability density function*** of the normal distribution is given as

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

Integrating this formula would allow us to compute probabilities. However, we will not use this formula, although we later discuss what a probability density function is.

Normal Distribution



Characteristics of the Normal probability distribution

1. The highest point on the normal curve is at the mean, which is also the median and mode of the distribution.
2. [VERY IMPORTANT] The normal probability curve is bell-shaped and symmetric, with the shape of the curve to the left of the mean a mirror image of the shape of the curve to the right of the mean.
3. The standard deviation determines the width of the curve. Larger values of the the standard deviation result in wider flatter curves, showing more dispersion in data.
4. The total area under the curve for the normal probability distribution is 1.

Characteristics of the Normal probability distribution

- The interval defined by the mean $\pm 1\times$ standard deviation includes approximately 68% of the observations, leaving 16% (approx) in each tail.
- The interval defined by the mean $\pm 1.96\times$ standard deviation includes approximately 95% of the observations, leaving 2.5% (approx) in each tail.
- The interval defined by the mean $\pm 2.58\times$ standard deviation includes approximately 99% of the observations, leaving 0.5% (approx) in each tail.

Remark: It is useful to know this numbers, but we will do all calculations from first principles.

The Standard Normal Distribution

- The standard normal distribution is a special case of the normal distribution with a mean $\mu = 0$ and a standard deviation $\sigma = 1$.
- We denote the standard normal random variable as Z rather than X.
- The distribution is well described in statistical tables (i.e. Murdoch Barnes Table 3)
- Rather than computing probabilities from first principles, which is very difficult, probabilities from distributions other than the Z distribution (e.g. X $\sim(\mu = 100, \sigma = 15)$) can be computed using the Z distribution, a much easier approach. (We shall demonstrate how shortly.)

Standardization formula

All normally distributed random variables have corresponding Z values, called Z-scores.

For normally distributed random variables, the z-score can be found using the ***standardization formula***;

$$z_o = \frac{x_o - \mu}{\sigma}$$

where x_o is a score from the original normal ("X") distribution, μ is the mean of the original normal distribution, and σ is the standard deviation of original normal distribution.

Therefore z_o is the z-score that corresponds to x_o.

- Terms with subscripts mean particular values, and are not variable names.
- The z distribution will only be a normal distribution if the original distribution (X) is normal.

The Standardized Value

- Suppose that mean $\mu = 80$ and that standard deviation $\sigma = 8$.
- What is the Z-score for $x_o = 100$?

$$z_{100} = \frac{x_0 - \mu}{\sigma} = \frac{100 - 80}{8} = \frac{20}{8} = 2.5$$

- Therefore $z_{100} = 2.5$

A Z-score always reflects the number of standard deviations above or below the mean a particular score is. Suppose the scores of a test are normally distributed with a mean of 50 and a standard deviation of 9 For instance, if a person scored a 68 on a test, then they scored 2 standard deviations above the mean.

Converting the test scores to z scores, an X value of 68 would yield:

$$Z = \frac{68 - 50}{9} = 2$$

So, a Z score of 2 means the original score was 2 standard deviations above the mean.

The Standard Normal (Z) Distribution Tables

- Importantly, probabilities relating to the z distribution are comprehensively tabulated in ***Murdoch Barnes table 3***.
- Given a value of k (with k usually between 0 and 4), the probability of a standard normal "Z" random variable being greater than (or equal to) k $P(Z \geq k)$ is given in Murdoch Barnes table 3 .
- Other statistical tables can be used, but they may tabulate probabilities in a different way.

An Important Identity

If two values z_o and x_o are related in the following way, for some values μ and σ,

$$z_0 = \frac{x_0 - \mu}{\sigma}$$

Then we can can say

$$P(X \geq x_o) = P(Z \geq z_o)$$

or alternatively

$$P(X \leq x_o) = P(Z \leq z_o)$$

This is fundamental to solving problems involving normal distributions.

Using Murdoch Barnes tables 3

- For some value z_o, between 0 and 4, the Murdoch Barnes tables set 3 tabulate $P(Z \geq z_o)$
- Ideally z_o would be specified to 2 decimal places. If it is not, round to the closest value.
- We call the third digit (i.e. the digit in the second decimal place) the "second precision".

Using Murdoch Barnes tables 3

- To compute the relevant probability we express z_o as the sum of z_o without the second precision, and the second precision.(For example $1.28 = 1.2 + 0.08$.)
- Select the row that corresponds to z_o without the second precision (e.g. 1.2).
- Select the column that corresponds to the second precision(e.g. 0.08).
- The value that contained on the intersection is $P(Z \geq z_o)$

Find $P(Z \geq 1.28)$

	…	…	0.006	0.07	0.08	0.09
…	…	…	…	…	…	…
1.0	…	…	0.1446	0.1423	0.1401	0.1379
1.1	…	…	0.1230	0.1210	0.1190	0.1170
1.2	…	…	0.1038	0.1020	0.1003	0.0985
1.3	…	…	0.0869	0.0853	0.0838	0.0823
…	…	…	…	…	…	…

Using Murdoch Barnes tables 3

- Find $P(Z \geq 0.60)$
- Find $P(Z \geq 1.64)$
- Find $P(Z \geq 1.65)$
- Estimate $P(Z \geq 1.645)$

Find $P(Z \geq 0.60)$

	0.00	0.01	0.02	0.03	...	...
...	...	...	...	...	...	...
0.4	0.3446	0.3409	0.3372	0.3336	...	...
0.5	0.3085	0.3050	0.3015	0.2981	...	...
0.6	0.2743	0.2709	0.2676	0.2643	...	...
0.7	0.2420	0.2389	0.2358	0.2327	...	...
...	...	...	...	...	...	...

Find $P(Z \geq 1.64)$ and $P(Z \geq 1.65)$

	$\ldots$	$\ldots$	0.04	0.05	0.06	0.07
$\ldots$	$\ldots$	$\ldots$	$\ldots$	$\ldots$	$\ldots$	$\ldots$
1.5	$\ldots$	0.0630	0.0618	0.0606	0.0594	$\ldots$
1.6	$\ldots$	0.0516	0.0505	0.0495	0.0485	$\ldots$
1.7	$\ldots$	0.0418	0.0409	0.0401	0.0392	$\ldots$
$\ldots$	$\ldots$	$\ldots$	$\ldots$	$\ldots$	$\ldots$	$\ldots$

Using Murdoch Barnes tables 3

- $P(Z \geq 1.64) = 0.505$
- $P(Z \geq 1.65) = 0.495$

- $P(Z \geq 1.645)$ is approximately the average value of $P(Z \geq 1.64)$ and $P(Z \geq 1.65)$.
- $P(Z \geq 1.645)$ = (0.0495 + 0.0505)/2 = 0.0500. (i.e. 5%)

Exact Probability

Remarks: This is for continuous distributions only.

- The probability that a continuous random variable will take an exact value is infinitely small. We will usually treat it as if it was zero.
- When we write probabilities for continuous random variables in mathematical notation, we often retain the equality component (i.e. the ”...or equal to..”).
  For example, we would write expressions $P(X \leq 2)$ or $P(X \geq 5)$.
- Because the probability of an exact value is almost zero, these two expression are equivalent to $P(X < 2)$ or $P(X > 5)$.
- The complement of $P(X \geq k)$ can be written as $P(X \leq k)$.

Complement and Symmetry Rules

Any normal distribution problem can be solved with some combination of the following rules.

1. Complement rule

- Common to all continuous random variables

$$P(Z \geq k) = 1 - P(Z \leq k)$$

Similarly

$$P(X \geq k) = 1 - P(X \leq k)$$

$$P(Z \leq 1.28) = 1 - P(Z \geq 1.28) = 1 - 0.1003 = 0.8997$$

Complement and Symmetry Rules

2. Symmetry rule

- This rule is based on the property of symmetry mentioned previously.
- Only the probabilities corresponding to values between 0 and 4 are tabulated in Murdoch Barnes.
- If we have a negative value of k, we can use the symmetry rule.

$$P(Z \leq -k) = P(Z \geq k)$$

by extension, we can say

$$P(Z \geq -k) = P(Z \leq k)$$

Example

Find $P(Z \geq -1.28)$ **Solution**

- Using the symmetry rule

$$P(Z \geq -1.28) = P(Z \leq 1.28)$$

- Using the complement rule

$$P(Z \geq -1.28) = 1 - P(Z \geq 1.28)$$

$$P(Z \geq -1.28) = 1 - 0.1003 = 0.8997$$

Find the probability of a “z” random variable being between -1.8 and 1.96?
i.e. Compute $P(-1.8 \leq Z \leq 1.96)$
Solution

- Consider the complement event of being in this interval: a combination of being too low or too high.
- The probability of being too low for this interval is $P(Z \leq -1.80) = 0.0359$ (check)
- The probability of being too high for this interval is $P(Z \geq 1.96) = 0.0250$ (check)
- Therefore the probability of being **outside** the interval is 0.0359 + 0.0250 = 0.0609.
- Therefore the probability of being **inside** the interval is 1- 0.0609 = 0.9391 $P(-1.8 \leq Z \leq 1.96) = 0.9391$

Solutions

Let x be the normal random variable describing waiting times
$P(X \geq 15) = ?$

First , we find the z-value that corresponds to x = 15 (remember $\mu = 10$ and $\sigma = 3$)

$$z_o = \frac{x_o - \mu}{\sigma} = \frac{15 - 10}{3} = 1.666$$

- We will use $z_o = 1.67$
- Therefore we can say $P(X \geq 15) = P(Z \geq 1.67)$
- The Murdoch Barnes tables are tabulated to give $P(Z \geq z_o)$ for some value z_o .
- We can evaluate $P(Z \geq 1.67)$ as 0.0475.
- Necessarily $P(X \geq 15) = 0.0475$.