The Normal Distribution

Contents

.

Normal distribution

Continuous probability distributions

With continuous variables, our concern is then to associate probabilities with outcomes of certain ranges (or intervals), not single values. To facilitate the assignment of probabilities in this way, a continuous distribution is used, the most common being *the normal distribution.*

Similarly, strictly discrete variables which are *effectively* continuous should be treated as continuous. An example would be observing the number of heads in 10 tosses of coin compared to 100 tosses. In the first case it may be useful to calculate the probability of exactly 7 heads in 10 tosses; but it would not be useful to calculate the probability of exactly 70 heads in 100 tosses, since the probability would be so small. In the latter case we would use a range, such as, $P(70 \leq X \leq 75)$. That is, what is the chance we could have 70 to 75 heads inclusive in 100 tosses. Variables such as time, size and weight follow continuous probability distributions and so we have to work with ranges of values, rather than specific individual values.

Properties of the normal distribution

The normal distribution is one of the most important distributions in statistics. There are two reasons for this. First, many real-life examples and naturally occurring phenomena exhibit variation in line with the normal distribution, and many processes (such as output from a production line) have a variation that can be approximated by a normal distribution. Second, the normal is one of the distributions that form the basis of sampling theory and statistical inference.

Using *Z* scores

An important feature of the normal distribution is that we work in terms of the '*number of standard deviations above and/or below the mean'*. We use *Z* to denote the number of standard deviations in which we are interested.

The *Z* score can help us examine every normal variable in terms of the number of standard deviations a value is above or below the mean. Thus, try to remember that in your reading from now on:

Z **means nothing more than the 'number of standard deviations (σ) a particular value (*X*) is away from the mean (μ)'.**

The transformation is no more than a scale transformation of the normal variable *X* into the *Z* scale. Hence, instead of measuring the distance of *X* from the mean (μ) as

$$X - \mu$$

which has the units of *X*, we measure this distance as units of standard deviations (σ) from the mean, that is,

$Z = (X - \mu)/\sigma$

which has no units.

The two important types of calculations we carry out with the normal distribution are:

- finding a probability (area), given an *X* value
- finding an *X* value, given a probability.

These calculations can be done via either computer or statistical tables (see tutorial for instructions on how to use both).

APPLICATION: AUTOMATED PROCESS

Suppose we have an automated production process where the weight of output can be approximated by a normal distribution with mean or $\mu = 80$ g, and standard deviation or $\sigma = 4$ g. We wish to calculate what per cent of output will weigh more than 86g.

To solve it 'by hand' we would transform the problem to the *Z* scale:

$$Z = (X - \mu)/\sigma = (86 - 80)/4 = 1.50$$

$$P(X > 86) = P(z > 1.50)$$

From the normal table (see CloudDeakin), Z = 1.50 gives a value of 0.9332. However this corresponds to the probability of:

$$P(z > 1.50) = 0.9332$$

Our desired probability is obtained by:

$$P(z > 1.50) =1 - P(z < 1.50),\ 1 - 0.9332 = 0.0668$$

We have determined that 6.68% of output will exceed 86g in weight.

EXERCISE

Weekly demand for a particular brand of a perishable product is assumed to be approximately normal, with a mean of 150 and a standard deviation of 8. Management orders the product at the beginning of the week and throws out leftover stock at the end of the week.

(a) What would you recommend as the minimum and maximum order quantities for this product? Explain.

(b) The ordering policy has been to order 160 units of the brand. What is the probability that some of the product will need to be thrown out at the end of the week?

(c) If two consecutive weeks are examined, what is the joint probability that there will be wastage in both weeks? (Hint: Use your probability rules.) This is optional; however, if you attempt it, it is difficult.

EXERCISE

Suppose that the diameters of steel rods manufactured by a company are normally distributed with mean of 2 cm and a standard deviation of 0.02 cm. Output is not acceptable if the diameter is below 1.95 cm. In a production run what percentage of output will be defective?

EXERCISE

Complete these questions by hand and verify the results by computer.

1 Suppose you have an automated process that fills soft drink bottles. The process is approximately normally distributed and meant to fill with a mean of 750 ml and a standard deviation of 9 ml.

   (a) The actual capacity of the bottles is 775 ml. Can you predict the probability that a selected bottle will be filled to overflow?

   (b) If a selected bottle contained only 720 ml, what conclusion would you draw?

   (c) If the selected bottle contained 730 ml…?

   (d) If the selected bottle contained 740 ml…?

2 You offer a warranty on a type of television set picture tube. The lifetime of these picture tubes is distributed approximately normally with a mean of 48 months and a standard deviation of 12 months.

   (a) What is the probability that a picture tube will last more than 72 months?

   (b) What % of tubes will last more than 36 months if the warranty period is set for 3 years?

   (c) If the firm wanted to offer a warranty period that would cover only 1% of the least lasting picture tubes, what warranty period should it offer?

In this next section we look at sampling and some of the pitfalls that can occur when collecting samples of data from a population. As you will see there are good and bad ways to collect data through samples.

Sampling issues

We take samples rather than a census because a sample (of just 400 respondents) can save a great deal of time and money compared to a census (of say 10,000 employees in a company or 10,000,000 voters in a country), and in circumstances where items are destroyed (like finding out the average life of a new type of light globe, or new style of car tyre), a census does not make sense.

Perfect accuracy is not always, and perhaps rarely, justified:

- If a new product is expected to earn you $500,000 over five years, you can’t justify spending $1,000,000 to interview every Australian (census) to see if they would buy the product. You need a cost-effective data collection procedure.

- If you need to make a decision in three weeks about accepting a new contract, but need some data to help make that decision, you can't wait three months for a census of all people/items to be carried out, processed and analysed. You need a timely data collection procedure.
- If you need to tell consumers the 'average life of a new make of tyre', a census means you can be perfectly honest to the consumer, but you would have nothing to sell (all tyres destroyed).

Most of the inferential statistics techniques depend on simple random sampling being used when selecting the samples taken from the population. Statistical/inference—using confidence intervals and hypothesis tests—is valid *only for random samples.* Thus, you need to have a clear understanding *of random sampling* before proceeding.

Sampling distributions

When we have random sample data we use sample statistics (that is, the summary measures from a sample) such as the sample mean and sample proportion as estimators of population parameters (summary measures pertaining to a population) such as the population mean and population proportion. We therefore need to understand how sample statistics relate to their population counterparts. The sampling distribution is a concept which explains how a sample statistics can vary in relation to a population parameter.

Sampling distribution of the sample mean, $\overline{X}$

Consider a repeated sampling scenario where many different samples of the same size are taken from a population. If the sample mean from each sample is calculated, it is not likely to be the same value in each sample, and it is not likely to be the same as μ. For example, if you have 20 students and know their ages, take out one student aged 19 and replace with a student aged 23 and the mean moves.

Comparing the sample mean $\overline{X}$ to the population mean μ we can make the following observations:

- The sample mean $\overline{X}$ could be above or below μ (or equal to it, but how likely is that!)
- The minimum $\overline{X}$ you could get cannot be less than the minimum individual *X*, and it will be greater than this minimum value unless the sample contains *X* values all equal to the minimum (but how likely is that!)
- The maximum $\overline{X}$ you could get cannot be greater than the maximum individual *X*, and it will be less than this maximum value unless the sample contains *X* values all equal to the maximum (but how likely is that!)
- The more the values of *X* are representative of the population (that is, covering more of the population's range of values) the more *central* will be $\overline{X}$ (i.e. the closer $\overline{X}$ will be to the true central value, μ)

- The bigger the sample you take, the more likely the mean of the sample will be even more central and closer to μ (as you are getting closer to taking a census).

The repeated sampling exercise above gives us a data set of sample means. From a result known as the ***Central Limit Theorem*** it turns out that the average of all the sample means is identical to the population mean, that is:

$$\mu_{\bar{X}} = \mu$$

Read this as 'the mean of all possible sample means is equal to the population mean'. The standard deviation of the sample means distribution, *termed the standard error*, is given as:

$$\sigma_{\bar{\mathrm{X}}} = \frac{\sigma}{\sqrt{n}}$$

This equation shows us that the standard deviation varies according to the sample size. Larger samples yielding a smaller standard deviation (as *n* is in the denominator).

The central limit theorem explains that different conditions apply when sampling from normal and non-normal situations. In summary:

- if you are taking a sample from a *normally distributed population*, the sampling distribution of $\bar{X}$ will be normal for *n* of any size.
- if you are taking a sample from a *non-normally distributed population*, the sampling distribution of $\bar{X}$ may not be normal for small samples, and we use *n* ≥ 30 as a general cut-off for when *n* is large enough to use the normal distribution for the sampling distribution.

In the typical situation it is highly unlikely that we will know that the population is normal. In most cases, we will have little knowledge of the precise distribution of the population from which we are sampling. However, provided that the sample size is large we can determine that the sampling distribution will be *approximately normal.* (Large is determined by the symmetry of the population distribution. For a reasonably symmetric 'bell-shaped' population, *n* = 6 to 10 may be large enough, while for highly skewed distributions a sample size of *n* > 50 may be needed for a good approximation to a normal distribution. But as a conservative rule, usually *n* ≥ 30 is suggested.)

If a sampling distribution is not normal, we have no framework from which to make probability statements about sampling error.

In summary, provided the right conditions apply:

- for a given sample size, say *n* = 100, the sample mean $\bar{X}$ will fall either side of, or equal to, the population mean, μ
- the term 'sampling distribution' of the sample mean, $\bar{X}$, is the term used to describe this probability distribution

- the mean of the sampling distribution is μ. That is, while $\overline{X}$ could fall above or below μ, on average a sample mean will equal μ, the mean of all the individual values in a population
- the standard deviation of the sampling distribution for sample size is called the standard error and can be conveniently calculated using the formula:

$$\sigma_{\overline{X}} = \frac{\sigma}{\sqrt{n}}$$

- the sampling distribution follows (or approximately follows) the normal distribution if the population distribution from which the sample is taken is normal or if the sample size is large enough > 30.
- even if we don't know μ, provided we can calculate or obtain an acceptable estimate of the standard error, $\sigma_{\overline{X}}$, we can make probability judgments about the mean, $\overline{X}$, of a particular sample
- we use the term standard error for $\sigma_{\overline{X}}$ since it is a measure of the 'average' or 'standard' error that could be made in using $\overline{X}$ as a point estimate of μ.

Calculating probabilities for the sample mean, $\overline{X}$

In order to calculate probabilities for the sample mean we need to know the distribution of the sample mean. As the preceding section suggests, if the sampling distribution is normal or approximately normal, then the *Z* transform can be used to calculate areas under the curve:

$$Z = \frac{\overline{X} - \mu_{\overline{X}}}{\sigma_{\overline{X}}}$$

You need to calculate the standard error:

$$\sigma_{\overline{X}} = \frac{\sigma}{\sqrt{n}}$$

You then follow the same techniques covered earlier in this topic for calculating probabilities from a normal distribution.

APPLICATION: PRODUCTION TASK

Suppose a bank claims that the time taken to complete a transaction task by phone is normally distributed with $\mu_X = 25$ seconds and $\sigma_X = 4$ seconds. To test this claim, you take a random sample of n = 9 individuals who have used the phone system. The average time taken to complete the task by the 9 individuals was 28 seconds. Is this a likely result for the assumed distribution? Does this result fit in with what the bank is claiming?

We can do some probability calculations as follows. You will notice that since we are now in a *sampling* situation, we do not use the standard deviation (which analyses how individual values could occur), but calculate the *standard error*, (i.e. how a sample mean from a sample of size *n* = 9 could occur).

Since the population is assumed to be normal then we assume:

$\mu = \mu_X = 25$

$\sigma_{\overline{X}} = \frac{\sigma}{\sqrt{n}} = 4/3 = 1.33$ seconds.

What is the probability that $\overline{X}$ could be equal to or exceed 28? We use the *Z* transformation:

$z = (28 - 25)/1.33 = 2.25$

Using software or from the standardised normal probability table the probability equals 0.0122, that is, just over 1% or 1 chance in 100.

This probability surely raises some questions, in the same way as some of our exercises and applications in topic 5. Is the population average time really 25 seconds or something greater? Is our sample representative of the population?

What we do know is that if the population average is 25 and the population standard deviation is 4, there is slightly greater than 1 chance in 100 that the average time taken to complete the task for a sample of 9 elements is 28 seconds. How do you react if something occurs yet it is only meant to have a chance of 1 in 100 of occurring? Some decision makers would conclude that this is an 'unlikely' occurrence and conclude that it is more likely, based on the sample result that μ is not 25 but that the population average is greater than 25 seconds (that is, $\mu > 25$). Some decision makers—who have different sensitivity to probabilities—might adopt a different attitude and conclude that a chance of 1 in 100 is bearable, and stay with the original claim that $\mu = 25$ and $\sigma = 4$.

This example illustrates one way in which statisticians use probabilities to draw inferences. We will take up that type of thinking in detail when we study hypothesis testing and confidence intervals which use probabilities in another (but related) way.

APPLICATION: AVERAGE SPEND

Last financial year, a detailed study of customers spending patterns proved that the average (mean) spending of shoppers in our store was $245 with a standard deviation of $60. Spending was quite severely skewed to the right. We wish to carry out some analysis on this year's customers to see if there has been any change in spending patterns since last year.

(a) Initially, it was proposed to take a random sample of 16 of this year's customers. Your advice was:

'Given that spending was severely skewed to the right, a sample of size *n* = 16 is too small for us to confidently invoke the Central Limit Theorem (CLT); therefore, we recommend a sample of size 30 or more.'

(b) Eventually, we took a random sample of 100 of this year's customers and found their average expenditure was $230. What does this result tell us about average expenditure for this financial year compared to last financial year?

With a sample size of 100 shoppers we can invoke the CLT and the problem would be solved as follows:

Z = (230– 245)/(60/sqrt(100)) = -15/(60/10) = -15/6 = -2.50

$P(Z \leq -2.5) = 0.0062$

Interpretation of the result suggests that if the population mean is $245 and the standard deviation is $60, it is unlikely (0.62%) that a sample of 100 customers will have an average spending $230 or lower.

Given that a sample of 100 customers was taken and the sample mean was less than $230, what conclusions can be made?

There are really two alternative conclusions:

1 Our assumption about the population mean spending (or standard deviation or both) may be wrong. We may be over-estimating the average spending of customers. Our sample result of $230 is more consistent with a population mean below $245 and closer to $230.

2 The mean spending is as assumed (that is, the same as last year), and the reason that the sample mean spent was less than $230 was due to the sampling process. The sample generated was (due to bad luck) not representative of the population. (However, our preceding calculations determined that the chance of this was only 0.62% - less than a 1% chance.)

It is up to the analyst to decide. But in general we would adopt alternative 1, and conclude (on the basis of the probabilities, 0.62%) that mean expenditure for all customers for this year is less than $245. Our sample result is consistent with μ < $245. We will examine this issue more closely in next topics.

EXERCISE

The time taken to complete a unit of output for an automated process is normally distributed with mean equal to 45 seconds and standard deviation equal to 4 seconds. Suppose a sample of 16 units is taken and the completion time measured.

(a) What is the probability that a single unit of output will take longer than 47 seconds to be completed?

(b) What is the probability that the average of the sample of 16 units will exceed 47 seconds?

Summary

We make use of our knowledge about the normal distribution and show how probability is an essential aid to decision making in a sampling situation. In particular, we introduce the important concept of the sampling distribution of a sample statistic and show how, in conjunction with the normal distribution, we are able to predict and manage error from sampling.

The Central Limit Theorem assures us that, under the right circumstances, the sampling distribution for the sample mean and for the sample proportion will be (approximately) normally distributed. This means that probabilities for sample outcomes can be easily determined from the normal distribution, and inferences about population parameters are systematically and objectively based. The sampling distribution properties depend on the application of random sampling.

Further resources

Black, K 2008, *Business statistics for contemporary decision making*, 5th edn, Wiley, NJ.

Anderson, DR, Sweeney, DJ & Williams, TA 2008, *Statistics for business and economics*, 10th edn, South-Western Thomson Learning, Cincinnati, Mass.

Selvanathan, A, Selvanathan, S, Keller, G & Warrack, B 2006, *Australian business statistics*, 4th edn, Nelson Thomson Learning, Melbourne.