Sampling Distributions and Estimation

Tom Bruning

2018-01-24

Sampling Distributions and Estimation

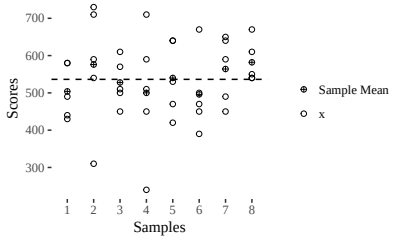
Sampling Variation

- **Sample statistic** – a random variable whose value depends on which population items are included in the random sample.
- Depending on the sample size, the sample statistic could either represent the population well or differ greatly from the population. This sampling variation can easily be illustrated.
  Consider eight random samples of size n = 5 from a large population of GMAT scores for MBA applicants.

Table 1: Random sample from GMAT Scores Population

1	2	3	4	5	6	7	8
490	310	500	450	420	450	490	670
580	590	450	590	640	670	450	610
440	730	510	710	470	390	590	550
580	710	570	240	530	500	640	540
430	540	610	510	640	470	650	540

Sample	Mean
1	504
2	576
3	528
4	500
5	540
6	496
7	564
8	582



The dot plot on the right shows that the sample means have much less variation than the individual sample items. The mean of the population is the dotted line which equals 520.28.

8.2 Estimators and Sampling Distributions

Some Terminology

- **Estimator** – a statistic derived from a sample to infer the value of a population parameter.
- **Estimate** – the value of the estimator in a particular sample.
- Population parameters are usually represented by Greek letters and the corresponding statistic by Roman letters.

The sample mean ($\bar{x}$) is the estimator for the population mean (μ).

The sample proportion (p) is the estimator for the population proportion (π).

The sample standard deviation (s) is the estimator for the population standard deviation (σ)

Sampling error is the difference between an estimate and the corresponding population parameter. For example, if we use the sample mean as an estimate for the population mean.

$$Sampling\ Error = \bar{x} - \mu \tag{1}$$

Bias is the difference between the expected value of the estimator and the true parameter.

$$Bias = E(\bar{X}) - \mu \tag{2}$$

An estimator is unbiased if its expected value is the parameter being estimated. The sample mean is an unbiased estimator of the population mean since:

$$Bias = E(\bar{X}) - \mu. \tag{3}$$

On average, an unbiased estimator neither overstates nor understates the true parameter.

Sample Mean and the Central Limit Theorem

The Central Limit Theorem for a Mean - If a random sample of size n is drawn from a population with mean μ and standard deviation σ, the distribution of the sample mean $\bar{X}$ approaches a normal distribution with mean μ and standard deviation $\sigma_{\bar{x}} = \sigma/\sqrt{n}$ as the sample size increases.

The Central Limit Theorem is a powerful result that allows us to approximate the shape of the sampling distribution of the sample mean even when we don't know what the population looks like.

- If the population is exactly normal, then the sample mean follows a normal distribution.
- As the sample size n increases, the distribution of sample means narrows in on the population mean μ.

- If the sample is large enough, the sample means will have approximately a normal distribution even if your population is not normal.

Applying The Central Limit Theorem

The Central Limit Theorem permits us to define an **interval** within which the sample means are expected to fall. As long as the sample size n is large enough, we can use the normal distribution regardless of the population shape (or any n if the population is normal to begin with).

- Expected range of Sample means:

$$\mu \pm z\frac{\sigma}{\sqrt{n}} \tag{4}$$

Illustration: All Possible Samples from a Uniform Population

- Consider a discrete uniform population consisting of the integers {0, 1, 2, 3}.

Table 3: All possible samples n=2

a	b	c	d
(0,0)	(1,0)	(2,0)	(3,0)
(0,1)	(1,1)	(2,1)	(3,1)
(0,2)	(1,2)	(2,2)	(3,2)
(0,3)	(1,3)	(2,3)	(3,3)

Table 4: Means of all possible samples n=2

a	b	c	d
0.0	0.5	1.0	1.5
0.5	1.0	1.5	2.0
1.0	1.5	2.0	2.5
1.5	2.0	2.5	3.0

The population parameters are: $\mu = 1.5$, $\sigma = 1.118$.

As you can see in the two graphs to the right, the top one is the histogram of the data above. The x-axis is the first item in each of the cells above, and the y-axis is the count of each of the corresponding cell. So, there are 4 combinations with 0 as the first number, 4 with the 1 as the first number, and so on. This is a uniform distribution.

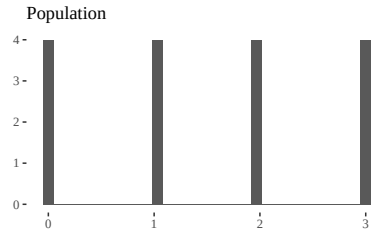


Figure 1: Uniform and Means Distribution

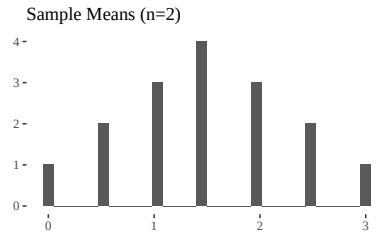


Figure 2: Uniform and Means Distribution

The bottom graph is a histogram of the means of each of the cells above. So there is one cell with a mean 0, 2 with a mean of 1.5, 3 with a mean of 1, etc. The means of the uniform distribution is a normal distribution with the mean of 1.5.

What is a Confidence Interval?

A sample mean $\bar{x}$ calculated from a random sample $x_1, x_2, x_3, ...x_n$ is a **point estimate** of the unknown population mean μ. Because samples vary, we need to indicate our uncertainty about the true value of μ. Based on our knowledge of the sampling distribution of $\bar{X}$, we can create an **interval estimate** for μ. We construct a **confidence interval** for the unknown mean μ by adding and subtracting a **margin of error** from $\bar{x}$, the mean of our random sample. The **confidence level** for this interval is expressed as a percentage such as 90, 95, or 99 percent.

The confidence interval for a mean μ with known σ is:

$$\bar{x} \pm z_{\alpha/2}\frac{\sigma}{\sqrt{n}} \tag{5}$$

Choosing a Confidence Level

- A higher confidence level leads to a wider confidence interval.
- Greater confidence implies loss of precision (i.e. greater margin of error).
  - 95% confidence is most often used.

Confidence Level: 90, $z_{.05} = 1.645$
Confidence Level: 95, $z_{.025} = 1.960$
Confidence Level: 98, $z_{.01} = 2.326$
Confidence Level: 99, $z_{.005} = 2.576$

Confidence Interval for a Mean (μ) with known (σ)

Interpretation

- A confidence interval either does or does not contain μ.
- The confidence level quantifies the risk.
- Out of 100 confidence intervals, approximately 95% may contain μ, while approximately 5% might not contain μ when constructing 95% confidence intervals.
  When Can We Assume Normality?
- If σ is known and the population is normal, then we can safely use the formula to compute the confidence interval.

- If σ is known and we do not know whether the population is normal, a common rule of thumb is that $n \geq 30$ is sufficient to use the formula as long as the distribution is approximately symmetric with no outliers.
- Larger n may be needed to assume normality if you are sampling from a strongly skewed population or one with outliers.

Confidence Interval Width

Confidence interval width reflects

- the sample size,
- the confidence level and
- the standard deviation.
  To obtain a narrower interval and more precision
- increase the sample size or
- lower the confidence level (e.g., from 90% to 80% confidence).