Sampling Distributions and Estimation

Tom Bruning

2020-09-08

Sampling Distributions and Estimation

Sampling Variation

A sampling distribution is a distribution of all of the possible values of a sample statistic for a given sample size selected from a population.

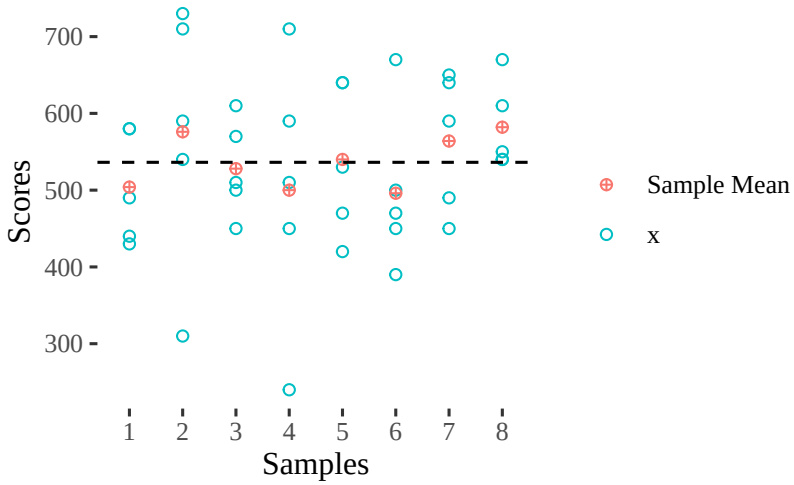
- **Sample statistic** – a random variable whose value depends on which population items are included in the random sample.
- Depending on the sample size, the sample statistic could either represent the population well or differ greatly from the population. This sampling variation can easily be illustrated.
  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

```
## `summarise()` ungrouping output (override with `.groups` argument)
```

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



The dot plot above 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)

*This is the **Most** important concept of this course.*

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 Sampling Distribution: Standard Error of the Mean

- Different samples of the same size from the same population will yield different sample means.
- A measure of the variability in the mean from sample to sample is given by the Standard Error of the Mean:

$$\sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}} \tag{4}$$

(This assumes that sampling is with replacement or sampling is without replacement from an infinite population.)

Note that the standard error of the mean decreases as the sample size increases.

Sample Mean Sampling Distribution: If the Population is Normal

If a population is normal with mean μ and standard deviation σ, the sampling distribution of is also normally distributed with:

$$\mu_{\bar{x}} = \mu \tag{5}$$

and

$$\sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}} \tag{6}$$

Z-value for Sampling Distribution of the Mean

Z-value for the sampling distribution of $\bar{X}$:

$$Z = \frac{(\bar{X} - \mu_{\bar{X}})}{\sigma_{\bar{x}}} = \frac{(\bar{X} - \mu)}{\frac{\sigma}{\sqrt{n}}} \tag{7}$$

where:
$\bar{X}$ = sample mean
μ = population mean
σ = population standard deviation
n = sample size

Determining An Interval Including A Fixed Proportion of the Sample Means

Find a symmetrically distributed interval around µ that will include 95% of the sample means when $\mu = 368$, $\sigma = 15$, and n = 25.

- Since the interval contains 95% of the sample means 5% of the sample means will be outside the interval.
- Since the interval is symmetric 2.5% will be above the upper limit and 2.5% will be below the lower limit.
- From the standardized normal table, the Z score with 2.5% (0.0250) below it is -1.96 and the Z score with 2.5% (0.0250) above it is 1.96.

Calculating the lower limit of the interval:

$$\bar{X_L} = \mu + Z(\frac{\sigma}{\sqrt{n}}) = 368 + (-1.96)(\frac{15}{\sqrt{25}}) = 362.12 \tag{8}$$

Calculating the upper limit of the interval:

$$\bar{X_U} = \mu + Z(\frac{\sigma}{\sqrt{n}}) = 368 + (1.96)(\frac{15}{\sqrt{25}}) = 373.88 \tag{9}$$

Based on samples of size 25, the sample means in 95% of all samples are between 362.12 and 373.88.

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.

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.

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.

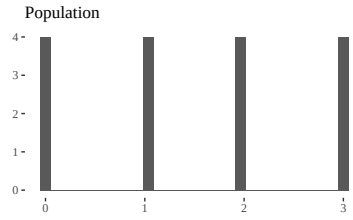


Figure 1: Uniform and Means Distribution

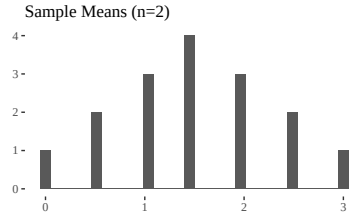


Figure 2: Uniform and Means Distribution

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:

$$\mu \pm z\frac{\sigma}{\sqrt{n}} \tag{10}$$

Example

Suppose a population has mean $\mu = 8$ and standard deviation $\sigma = 3$. Suppose a random sample of size n = 36 is selected.

What is the probability that the sample mean is between 7.8 and 8.2?

Solution:

Even if the population is not normally distributed, the central limit theorem can be used (n > 30).

so the **sampling distribution** of $\bar{X}$ is approximately normal.

with mean $\mu_{\bar{x}} = 8$ and standard deviation $\sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}} = \frac{3}{\sqrt{36}} = 0.5$

$$P(\frac{7.8-8}{\frac{3}{\sqrt{36}}} < \frac{\bar{X}-\mu}{\frac{\sigma}{\sqrt{n}}} < \frac{8.2-8}{\frac{3}{\sqrt{36}}}) \tag{11}$$

$$P(-0.4 < Z < 0.4) = 0.6554 - 0.3446 = 0.3108 \tag{12}$$

Population Proportions

π = the proportion of the population having some characteristic.

- Sample proportion (p) provides an estimate of π:

$$\frac{X}{n} = \frac{number\ of\ items\ in\ the\ sample\ that\ have\ a\ characteristic\ of\ interest}{sample\ size} \tag{13}$$

- $0 \leq p \leq 1$.
- p is approximately distributed as a normal distribution when n is large.
  (assuming sampling with replacement from a finite population or without replacement from an infinite population.)

Sampling Distribution of p is approximated by a normal distribution if:

$n\pi > 5$ *and* $n(1-\pi) > 5$.

where: $\mu_p = \pi$ and $\sigma_p = \sqrt{\frac{\pi(1-\pi)}{n}}$

where: π is the population proportation.

Z-Value for Proportions

Standardize p to a Z value with the formula:

$$Z = \frac{p-\pi}{\sigma_p} = \frac{p-\pi}{\sqrt{\frac{\pi(1-\pi)}{n}}} \tag{14}$$

Example

If the true proportion of voters who support Proposition A is $\pi = 0.4$, what is the probability that a sample of size 200 yields a sample proportion between 0.40 and 0.45?

i.e.: if $\pi = 0.4$ and n = 200, what is $P(0.40 \leq p \leq 0.45)$?
Find σ_p: $\sigma_p = \sqrt{\frac{\pi(1-\pi)}{n}} = \sqrt{\frac{0.40(1-0.40)}{200}} = 0.03464$

$$P(0.40 \leq p \leq 0.45) = P(\frac{0.40 - 0.40}{0.03464} \leq Z \leq \frac{0.45 - 0.40}{0.03464}) = P(0 \leq Z \leq 1.44) \tag{15}$$

Utilize the cumulative normal table:
$P(0 \leq Z \leq 1.44) = 0.9251–0.5000 = 0.4251$