UNIVERSITY OF LONDON

Probability and Statistics: To p, or not to p?

Module Leader: Dr James Abdey

4.5 Sampling distribution of the sample mean

Like any distribution, we care about a sampling distribution's mean and variance. Together, we can assess how 'good' an estimator is.

First, consider the mean. We seek an estimator which does not mislead us *systematically*. So the 'average' (mean) value of an estimator, over all possible samples, should be equal to the population parameter itself.

Returning to our example:

$\bar{x}$	Frequency	Product
3.5	1	3.5
4.5	1	4.5
5.0	3	15.0
5.5	2	11.0
6.0	1	6.0
6.5	3	19.5
7.0	1	7.0
7.5	1	7.5
8.0	2	16.0
Total	15	90.0

Hence the mean of this sampling distribution is $90/15 = 6 = \mu$.

An important difference between a sampling distribution and other distributions is that the values in a sampling distribution are summary measures of whole samples (i.e. statistics, or estimators) rather than individual observations.

Formally, the mean of a sampling distribution is called the **expected value** of the estimator, denoted by $\mathrm{E}(\cdot)$.

Hence the expected value of the sample mean is $\mathrm{E}(\bar{X})$.

An **unbiased estimator** has its expected value equal to the parameter being estimated. For our example, $\mathrm{E}(\bar{X}) = 6 = \mu$.

Fortunately, the sample mean $\bar{X}$ is *always* an unbiased estimator of μ in simple random sampling, regardless of the:

- sample size, n
- distribution of the (parent) population.

This is a good illustration of a population parameter (here, μ) being estimated by its sample counterpart (here, $\bar{X}$).

The unbiasedness of an estimator is clearly desirable. However, we also need to take into account the *dispersion* of the estimator's sampling distribution. Ideally, the possible values of the estimator should not vary much around the true parameter value. So, we seek an estimator with a small variance.

Recall the variance is defined to be the *mean of the squared deviations about the mean* of the distribution. In the case of sampling distributions, it is referred to as the **sampling variance**.

Returning to our example:

$\bar{x}$	$\bar{x} - \mu$	$(\bar{x} - \mu)^2$	Frequency	Product
3.5	-2.5	6.25	1	6.25
4.5	-1.5	2.25	1	2.25
5.0	-1.0	1.00	3	3.00
5.5	-0.5	0.25	2	0.50
6.0	0.0	0.00	1	0.00
6.5	0.5	0.25	3	1.75
7.0	1.0	1.00	1	1.00
7.5	1.5	2.25	1	2.25
8.0	2.0	4.00	2	8.00
		Total	15	24.00

Hence the sampling variance is $24/15 = 1.6$.

The population itself has a variance, the population variance, σ^2.

x	$x - \mu$	$(x - \mu)^2$	Frequency	Product
3	-3	9	1	9
6	0	0	1	0
4	-2	4	1	4
9	3	9	1	9
7	1	1	2	2

Hence the population variance is $\sigma^2 = 24/6 = 4$.

We now consider the relationship between σ^2 and the sampling variance. Intuitively, a larger σ^2 should lead to a larger sampling variance. For population size N and sample size n, we note the following result when sampling without replacement:

$$\boldsymbol{\mathrm{Var}(\bar{X}) = \frac{N-n}{N-1} \times \frac{\sigma^2}{n}.}$$

So, for our example, we get:

$$\mathrm{Var}(\bar{X}) = \frac{6-2}{6-1} \times \frac{4}{2} = 1.6.$$

We use the term **standard error** to refer to the standard deviation of the sampling distribution, so:

$$\boldsymbol{\mathrm{S.E.}(\bar{X}) = \sqrt{\mathrm{Var}(\bar{X})} = \sqrt{\frac{N-n}{N-1} \times \frac{\sigma^2}{n}} = \sigma_{\bar{X}}.}$$

Some implications are the following.

- As the sample size n increases, the sampling variance decreases, i.e. the **precision** increases.[1]
- Provided the **sampling fraction**, n/N, is small, the term:

$$\boldsymbol{\frac{N-n}{N-1} \approx 1}$$

  so can be ignored. Therefore, the precision depends effectively on n only.

Returning to our example, the larger the sample, the less variability there will be between samples.

$\bar{x}$	$n=2$	$n=4$
3.50	1	–
4.50	1	–
5.00	3	2
5.25	–	1
5.50	2	1
5.75	–	3
6.00	1	1
6.25	–	2
6.50	–	3
6.75	–	1
7.00	1	–
7.25	–	1
7.50	1	–
8.00	2	–

[1] Although greater precision is desirable, data collection costs will rise with n. Remember why we sample in the first place!

We can see that there is a striking improvement in the precision of the estimator, because the variability has decreased considerably.

The range of possible $\bar{x}$ values goes from 3.5 to 8.0 down to 5.0 to 7.25. The sampling variance is reduced from 1.6 to 0.4.

The factor $(N-n)/(N-1)$ decreases steadily as $n \to N$. When $n = 1$ the factor equals 1, and when $n = N$ it equals 0.

When sampling *without replacement*, increasing n must increase precision since less of the population is left out. In much practical sampling N is *very* large (for example, several million), while n is comparably small (at most 1,000, say).

Therefore, in such cases the factor $(N-n)/(N-1)$ is close to 1, hence:

$$\mathrm{Var}(\bar{X}) = \frac{N-n}{N-1} \times \frac{\sigma^2}{n} \approx \frac{\sigma^2}{n} = \frac{\mathrm{Var}(X)}{n}$$

for small n/N. When N is large, it is the sample size n which is important in determining precision, *not* the sampling fraction.

Example

Consider two populations: $N_1 = 3$ million and $N_2 = 200$ million, both with the same variance, σ^2. We sample $n_1 = n_2 = 1000$ from each population, then:

$$\sigma^2_{\bar{X}_1} = \frac{N_1 - n_1}{N_1 - 1} \times \frac{\sigma^2}{n_1} = (0.999667) \times \frac{\sigma^2}{1000}$$

and:

$$\sigma^2_{\bar{X}_2} = \frac{N_2 - n_2}{N_2 - 1} \times \frac{\sigma^2}{n_2} = (0.999995) \times \frac{\sigma^2}{1000}.$$

So $\sigma^2_{\bar{X}_1} \approx \sigma^2_{\bar{X}_2}$, despite N_1 being much less than N_2.

Sampling from the normal distribution

The mean and variance of $\bar{X}$ are $\mathrm{E}(X)$ and $\mathrm{Var}(X)/n$, respectively, for a random sample of size n from *any* population distribution of X. What about the form of the sampling distribution of $\bar{X}$?

This depends on the distribution of X, and is not generally known. However, when the distribution of X is *normal*, the sampling distribution of $\bar{X}$ is also normal.

Suppose that $\{X_1, \ldots, X_n\}$ is a random sample from a normal distribution with mean μ and variance σ^2. Therefore:

$$\bar{X} \sim N\left(\mu, \frac{\sigma^2}{n}\right).$$

So we note $\mathrm{E}(\bar{X}) = \mathrm{E}(X) = \mu$.

- In an individual sample, $\bar{x}$ is not usually equal to μ, the expected value of the population.
- However, *over repeated samples* the values of $\bar{X}$ are centred at μ.

We also note $\text{Var}(\bar{X}) = \text{Var}(X)/n = \sigma^2/n$, and so the standard error is $\sigma/\sqrt{n}$.

The variation of values of $\bar{X}$ in different samples (the sampling variance) is large when the population variance of X is large. More interestingly, the sampling variance gets smaller when the sample size n increases.

In other words, when n is large the distribution of $\bar{X}$ is more tightly concentrated around μ than when n is small.

Example

Suppose $X \sim N(5,\, 1)$, then:

$$\bar{X} \sim N\left(5,\, \frac{1}{n}\right).$$

The figure below shows the sampling distribution of $\bar{X}$ for $n = 5$, $n = 20$ and $n = 100$. Note how all three sampling distributions are centred on 5, since:

$$\text{E}(\bar{X}) = \text{E}(X) = \mu = 5$$

while the sampling variance decreases as n increases since:

$$\text{Var}(\bar{X}) = \frac{\text{Var}(X)}{n} = \frac{\sigma^2}{n} = \frac{1}{n}.$$