Sampling Distribution

Manpreet S Katari

Sampling Distribution

We have been discussing the distribution of the population, however most of the time we are looking at a specific sample.

If we were to **sample** many times from the **sample population**, our samples would create a distribution of their own.

The Expected value of a sampling distribution will be the Expected value of the population.

The Variance of the sampling distribution is the standard deviation squared over n

.

$$E(\bar{X}) = \mu, \quad \text{and} \quad \text{Var}(\bar{X}) = \frac{\sigma^2}{n}.$$

Standard error

The **standard deviation** of the **sampling distribution** is called the **standard error**

Since standard deviation is square root of variance, the standard error is standard deviation of sampling distribution divided by square root of n

$$\sigma/\sqrt{n}.$$

Central Limit Theorem

There are two main statements:

- If the parent distribution is normal with mean μ and variance σ^2 then sample distribution is normal with mean μ and variance σ^2 over n
- Regardless of what the original distribution is, as n gets large, the sample distribution will become normal.
- A sample size of 30 is usually considered as cutoff.

Other distributions

- The sampling distribution of the population variance is a χ^2 distribution.
- The ratio of two variance from two population is an F distribution.
- A t-distribution is created using a t-statistic which is the difference between the sample mean and population mean divided by the the standard error.
- These will all serve important when we start working with null hypothesis in the next chapter.

$$F^* = \frac{S_1^2}{S_2^2} \qquad t^* = \frac{\sqrt{n}(\bar{X} - \mu_0)}{S}$$

Confidence Intervals

Confidence intervals provide a range where if the population was to be sampled, it's mean will be present X% of the time. The most common values used for X is 95%.

Confidence intervals do **NOT** represent the probability that the mean is within the range.

$$\bar{X} \pm z_{1-(\alpha/2)} \times \frac{\sigma}{\sqrt{n}}$$

EXAMPLE 5.7

An alpine vegetation study using 16 samples at alpine late snowbank sites found that the mean cover of the grass *Agrostis variabilis* was 14.6%. Assume that we know $\sigma = 4$, and calculate a 95% confidence interval for μ.

Famously, the z-quantile for the probability of $1 - (\alpha/2) = 0.975$ equals approximately 1.96.

Resampling Distributions

To understand the sample distribution, it is often helpful to sample from the current sample to reduce bias

Bootstrapping - sample n times , with replacement, from a sample of size n

- Nonparametric methods assume to weights.
- Parametric methods influence which values get selected more often.

Values from bootstrap method can also be used to determine confidence intervals.

Jackknifing - calculate the metrics while remove one value at a time.