# Central Limit Theorem


DASC 512

# Overview

- Population distribution
- Sampling distribution
- Central Limit Theorem (CLT)
- Example application
- Simulations

# Population distribution

So far, we have mostly been talking about the population distribution

- Each case is a single observation
- Observational unit is one member of the population

We want to make inferences on the parameters of the population distribution

# Sampling distribution

The sampling distribution is the distribution of point estimates of the population parameters

- Each case is a point estimate calculated from the sample
- Observational unit is one sample
- The sampling distribution is not observed unless the experiment is repeated

Every sample will have some variability in point estimates, but it will be less than that of the population

# Binomial distribution

Example: the binomial distribution is the sampling distribution for the Bernoulli distribution! Let $X \sim \text{Bernoulli}$.

- Each case is a 1 or 0 value

$$x_i \in \{0,1\}$$

- Parameter of interest: proportion

$$p = E(X) = \mu$$

- Measured statistic: sample proportion

$$\hat{p} = \bar{x} = \frac{\sum_{i=1}^{n} x_i}{n}$$

# Sampling distribution

Discussing beer, we said that if the population is normally distributed…

- Small sample means follow a Student t-distribution
- Large sample means approximate a Normal distribution

Put mathematically, if $X_i \sim N(\mu, \sigma), i = 1,2, \ldots, n$ and $n \geq 30$ then

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

# Standard Error

The standard error about the mean is the standard deviation of the sampling distribution of sample means.

Given population standard deviation $\sigma$, the standard error is

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

We can estimate this using a sample standard deviation

$$s_{\bar{X}} = \frac{s}{\sqrt{n}} = \sqrt{\left(\sum_{i=1}^{n} \frac{(x_i - \bar{x})^2}{n(n-1)}\right)}$$

# Standard Error

$$s_{\bar{X}} = \frac{s}{\sqrt{n}} = \sqrt{\left(\sum_{i=1}^{n} \frac{(x_i - \bar{x})^2}{n(n-1)}\right)}$$

Note that the following hold from this formula:

- Lower population variance → Lower sampling distribution variance
- Larger sample size → Lower sampling distribution variance

# Central Limit Theorem

Let $X_1, X_2, \ldots, X_n$ be a random sample from a distribution with mean $\mu$ and variance $\sigma^2$. If the sample size $n$ is sufficiently large, then

- The sample mean $\bar{X}$ follows an approximate normal distribution with mean $\mu_{\bar{X}} = \mu$ and variance $\sigma^2_{\bar{X}} = \frac{\sigma^2}{n}$.

That is, the sampling distribution of the sample mean is approximately normally distributed **regardless of the distribution of the underlying random sample**. This includes discrete and continuous distributions.

- Exception: distributions for which mean/variance do not exist.

# “Sufficiently Large”???

The meaning of “sufficiently large” depends on the median, mean, and mode. If the distribution of $X_i$ is…

- Symmetric, unimodal, and continuous
  Sample size as small as 5-10 is sufficient
- Skewed, multimodal, or discrete
  Sample size of 25-30 is typically sufficient
- Extremely skewed
  Even larger samples may be required (rarely more than 100)

For binomial distribution, $np \geq 10$ and $n(1-p) \geq 10$.

# Central Limit Theorem

Proof of CLT requires statistical tools we do not have for this class

Let’s look at an example problem, then do some simulations in Python

# Grocery Store Checkout

At a particular grocery store, the manager suspects that customers are taking longer to check out than they used to. In the past, the average checkout time of customers was 2 minutes.

The manager observes a random sample of 36 customers and finds that they took an average of 3.2 minutes to check out. Does this prove her claim?

Assume that check out times follow an exponential distribution.

# The Distribution

Assuming an exponential distribution, we know that the mean is $\frac{1}{\lambda}$ and the variance is $\frac{1}{\lambda^2}$. Given that $\mu = 2$, that gives us $\lambda = \frac{1}{2}$ and $\sigma^2 = 4$.

What does this tell us we can assume about the sampling distribution?

# The Distribution

Assuming an exponential distribution, we know that the mean is $\frac{1}{\lambda}$ and the variance is $\frac{1}{\lambda^2}$. Given that $\mu = 2$, that gives us $\lambda = \frac{1}{2}$ and $\sigma^2 = 4$.

Based on the Central Limit Theorem, we expect the sampling distribution of the mean to be:

$$\bar{X} \sim N\left(\mu = 2, \sigma^2 = \frac{4}{36}\right) = N\left(\mu = 2, \sigma^2 = \frac{1}{9}\right)$$

# Probability

So how likely was an observation at least as extreme as the one observed?

# Probability

So how likely was an observation at least as extreme as the one observed?

$$P\left(\bar{X} > 3.2 \middle| \bar{X} \sim N\left(2, \frac{1}{9}\right)\right) = P\left(Z > \frac{3.2 - 2}{\sqrt{1/9}}\right) = P(Z > 3.6)$$

What do you conclude? Are check out times longer than they used to be?

# Probability

So how likely was an observation at least as extreme as the one observed?

$$P\left(\bar{X} > 3.2 \middle| \bar{X} \sim N\left(2, \frac{1}{9}\right)\right) = P\left(Z > \frac{3.2 - 2}{\sqrt{1/9}}\right) = P(Z > 3.6)$$

If the population mean were really still 2, this observation would be expected to occur once in 6,250 observations – not very likely! The checkout time has probably changed for some reason.

This application is called hypothesis testing and is coming soon.

# And now to examine CLT in action…

# Recap

- Population distribution
- Sampling distribution
- Central Limit Theorem (CLT)
- Example application
- Simulations