Module C: Understanding the Data-Generating Process

Slide Deck C11:

The Central Limit Theorem

The section in which we explicitly learn about the Central Limit Theorem (CLT) and why it is so important to statistical inference. Examples are provided to illustrate the use of the CLT, allowing us to focus on the distribution of the measure of center instead of the distribution of the data.

Today's Objectives

By the end of this slidedeck, you should

1. state the Central Limit Theorem
2. state the requirements for applying it
3. state its consequences
    - with respect to the distribution of the sample mean
    - with respect to the distribution of the sample proportion

Theorem Statement

Theorem (Central Limit Theorem)

Let X be a random variable with mean μ and finite variance σ^2. Let us draw a random sample of size n from this distribution.

Then, the distribution of the sample sums converges to a Normal distribution as n gets larger. Specifically,

$$\sum_{i=1}^{n} X_i \xrightarrow{d} \mathcal{N}(n\mu,\ n\sigma^2)$$

The proof of this theorem is beyond the scope of this course. It is first proven in MATH 321: Mathematical Statistics I.

Theorem Consequences

The Central Limit Theorem (CLT) tells us the following:

- The *sum of* independent random variables is more Normal than the distribution of the variable itself, unless the variable is Normally distributed or if it has an infinite variance.
  - The Binomial is a sum of independent Bernoulli rvs
  - The Poisson is a sum of independent Poisson rvs
- Because the sample mean is just the sample sum, divided by a constant (n), the CLT tells us that *the distribution of sample means* will tend towards Normal.
- The speed of convergence depends on how closely the data distribution is to Normal. The closer, the faster.

Example 1: Uniform

Example

Let $X \sim \mathcal{U}nif(a,\ b)$. Use the Central Limit Theorem to estimate the distribution of the sum of a sample of size n.

By the CLT,

$$T \dot{\sim} \mathcal{N}(n\mu,\ n\sigma^2)$$

From our knowledge of the Uniform distribution, this means

$$T \dot{\sim} \mathcal{N}\left(n\frac{a+b}{2},\ n\frac{(b-a)^2}{12}\right)$$

Example 2: Exponential

Example

Let $X \sim \mathcal{E}xp(\lambda)$. Use the Central Limit Theorem to estimate the distribution of the sum of a sample of size n.

By the CLT,

$$T \dot{\sim} \mathcal{N}(n\mu,\ n\sigma^2)$$

From our knowledge of the Exponential distribution, this means

$$T \dot{\sim} \mathcal{N}\left(n\frac{1}{\lambda},\ n\frac{1}{\lambda^2}\right)$$

Example 3: Binomial

Example

Let $X \sim \mathcal{B}in(n,\ p)$. Use the Central Limit Theorem to approximate the distribution of X.

Note that the Binomial is just the sum of n independent Bernoulli distributions. That is, if $Y_i \sim \mathcal{B}in(1,\ p)$,

$$X = \sum_{i=1}^{n} Y_i \sim \mathcal{B}in(n,\ p)$$

Thus, by the Central Limit Theorem, we know

$$X \dot{\sim} \mathcal{N}\big(np,\ np(1-p)\big)$$

Distribution of Sample Mean

Corollary (Distribution of Sample Mean)

Let X be a random variable with mean μ and finite variance σ^2. Let us draw a random sample of size n from this distribution.

Then, the distribution of the sample mean is

$$\overline{X}_n \xrightarrow{d} \mathcal{N}\left(\mu,\ \frac{1}{n}\sigma^2\right)$$

Distribution of Sample Mean

Proof From the Central Limit Theorem, we know

$$\sum_{i=1}^{n} X_i \dot{\sim} \mathcal{N}\left(n\mu;\ n\sigma^2\right)$$

Thus,

$$\overline{X}_n = \frac{1}{n}\sum_{i=1}^{n} X_i \dot{\sim} \mathcal{N}\left(\mu;\ \frac{1}{n}\sigma^2\right)$$

Distribution of Sample Mean

Sub-Proof 1

$$\begin{aligned}
\mathbb{E}\left[\overline{X}_n\right] &= \mathbb{E}\left[\frac{1}{n}T\right] \\
&= \frac{1}{n}\mathbb{E}\left[T\right] \\
&= \frac{1}{n}\ n\mu \\
&= \mu
\end{aligned}$$

Distribution of Sample Mean

Sub-Proof 2

$$\begin{aligned}\mathbb{V}\left[\,\overline{X}_n\,\right] &= \mathbb{V}\left[\frac{1}{n}T\right] \\ &= \frac{1}{n^2}\mathbb{V}\left[\,T\,\right] \\ &= \frac{1}{n^2}\; n\sigma^2 \\ &= \frac{1}{n}\sigma^2\end{aligned}$$

Distribution of Sample Mean

Corollary (Distribution of Sample Mean)

Let X be a random variable with mean μ and finite variance σ^2. Let us draw a random sample of size n from this distribution.

Then, the distribution of the sample mean is

$$\overline{X}_n \xrightarrow{d} \mathcal{N}\left(\mu, \frac{1}{n}\sigma^2\right)$$

Example 1: A First Look

Example

I draw a sample of size $n = 14$ from a population with mean $\mathbb{E}\left[\, X \,\right] = 126$ and variance $\mathbb{V}\left[\, X \,\right] = 42$. What is the approximate distribution of the sample means?

Solution. From the Mean Corollary to the CLT, the approximate distribution of the sample mean is

$$\overline{X} \mathrel{\dot\sim} \mathcal{N}(\mu_{\bar{x}} = 126,\ \sigma^2_{\bar{x}} = \frac{1}{14}\ 42) = \mathcal{N}(126,\ 3)$$

Example 2: Heights

Example

I have been told that the average adult height for males in the United States has mean $\mu = 69$ inches and standard deviation $\sigma = 3$ inches. What is the probability of having the mean of a **sample of size 2** being less than 65 inches?

Solution. Here, we are asked to calculate

$$\mathbb{P}\left[\ \overline{X} < 65\ \right]$$

To calculate this, we need to determine the distribution of $\overline{X}$. From the CLT, this is

$$\overline{X} \mathrel{\dot\sim} \mathcal{N}\left(69,\ \frac{1}{2}\ 3^2\right) = \mathcal{N}\left(\mu_{\bar{x}} = 69,\ \sigma^2_{\bar{x}} = 4.5\right)$$

Example 2: Heights

And so, since $\overline{X} \dot{\sim} \mathcal{N}\left(69,\ \frac{3^2}{2}\right)$,

$$\begin{aligned}\mathbb{P}\left[\ \overline{X} < 65\ \right] &= \texttt{pnorm(65, m=69, s=sqrt(4.5))} \\ &\approx 0.0297\end{aligned}$$

Thus, the probability of observing this event, given our assumptions are correct, is quite small. So, either I did not observe this event *or* my assumptions are unlikely to be true.

Example 3: More Heights

Example

I have been told that the average adult height for males in the United States has mean $\mu = 69$ inches and standard deviation $\sigma = 3$ inches. What is the probability of having the mean of a **sample of size 10** being less than 65 inches?

Solution. Here, we are asked to calculate

$$\mathbb{P}\left[\ \overline{X} < 65\ \right]$$

To calculate this, we need to determine the distribution of $\overline{X}$. From the CLT, this is

$$\overline{X} \dot{\sim} \mathcal{N}\left(69,\ \frac{1}{10}\ 3^2\right) = \mathcal{N}\left(69,\ 0.9\right)$$

Example 3: More Heights

And so, since $\overline{X} \dot{\sim} \mathcal{N}(69,\ 0.9)$,

$$\begin{aligned}\mathbb{P}\left[\ \overline{X} < 65\ \right] &\approx \texttt{pnorm(65, m=69,s=sqrt(0.9))} \\ &= 1.24133 \times 10^{-5} \\ &= 0.0000124\end{aligned}$$

Thus, the probability of observing this event, given our assumptions are correct, is very close to zero. So, either I did not observe this event *or* my assumptions are not correct.

Example 4: Crime

Example

The 2000 violent crime rate for the 50 states (+DC) are given in the data file `crime`. What is a 95% central confidence interval for the mean violent crime rate?

Solution. Here, we are asked to calculate the 2.5^{th} and 97.5^{th} quantiles (percentiles) of the sample means drawn from the 2000 violent crime rates. Note that $n = 51$ here.

One way of estimating this confidence interval is to apply the corollary to the Central Limit Theorem. From the data, we have a mean of 441.55 and a standard deviation of 241.45. The approximate sampling distribution will be

$$\overline{X} \dot{\sim} \mathcal{N}\left(441.55,\ \frac{1}{51}\ 241.45^2\right)$$

Example 4: Crime

Thus, we have a distribution of the sample means

$$\overline{X} \dot{\sim} \mathcal{N}\left(441.55,\ \frac{1}{51}241.45^2\right)$$

We know that the endpoints of a 95% confidence interval will be at the 0.025 and 0.975 quantiles of this distribution:

```
qnorm(c(0.025,0.975), m=441.55, s=241.45/sqrt(51))
```

We are 95% confident that the population mean is between 375 and 508 violent crimes per 100,000 people.

Example 4: Crime (Bootstrapping)

We could also estimate the confidence interval from the data. This process is called "bootstrapping," and here is the code:

```
mn=numeric()
for(i in 1:1e4) {
  x = sample(vcrime00, replace=TRUE)
  mn[i]=mean(x)
}
quantile(mn, c(0.025,0.975))
```

This gives a 95% confidence interval of 380 to 510.

Question: Why the difference between the two confidence intervals?

Distribution of Sample Proportion

Corollary (Distribution of Sample Proportion)

Let $X \sim \mathcal{B}in\,(n,\ p)$ be a random sample of size n from Bernoulli random variables.

Then, the distribution of the sample proportion is

$$P = \frac{1}{n}X \xrightarrow{d} \mathcal{N}\left(p,\ \frac{p(1-p)}{n}\right)$$

Distribution of Sample Proportion

Proof From the Central Limit Theorem, we know

$$X \mathrel{\dot\sim} \mathcal{N}\left(np,\ np(1-p)\right)$$

Thus,

$$\mathbb{E}\left[P\right] = \mathbb{E}\left[\frac{X}{n}\right] = \frac{\mathbb{E}\left[X\right]}{n} = \frac{np}{n} = p$$
$$\mathbb{V}\left[P\right] = \mathbb{V}\left[\frac{X}{n}\right] = \frac{\mathbb{V}\left[X\right]}{n^2} = \frac{np(1-p)}{n^2} = \frac{p(1-p)}{n}$$

and...

$$P \mathrel{\dot\sim} \mathcal{N}\left(p,\ \frac{p(1-p)}{n}\right)$$

Distribution of Sample Proportion

Corollary (Distribution of Sample Proportion)

Let $X \sim \mathcal{B}in\,(n,\ p)$ be a random sample of size n from Bernoulli random variables.

Then, the distribution of the sample proportion is

$$P = \frac{1}{n} X \xrightarrow{d} \mathcal{N}\left(p,\ \frac{p(1-p)}{n}\right)$$

Example 1: Poverty

Example

According to the US Census, 18% of Americans are below the poverty line. If I randomly sample $n = 10$ people from the United States, what is the probability that more than 20% of them are below the poverty line?

Solution. We are asked to calculate $\mathbb{P}[\ P > 0.20\ ]$. Thus, since the probability statement deals with P, we need to know the distribution of P.

From the corollary, we know that the approximate distribution of the sample proportion is

$$P \mathrel{\dot\sim} \mathcal{N}\left(p,\ \frac{p(1-p)}{n}\right) = \mathcal{N}\left(0.18,\ \frac{0.18(0.82)}{10}\right) = \mathcal{N}\left(0.18,\ 0.01476\right)$$

Example 1: Poverty

Solution (cont.). We have $P \dot{\sim} \mathcal{N}(0.18,\ 0.001476)$. Thus,

$$\begin{aligned}\mathbb{P}[\ P > 0.20\ ] &\approx 1 - \mathbb{P}[\ P \leq 0.20\ ] \\ &= \texttt{1 - pnorm(0.20, m=0.18, s=sqrt(0.01476))} \\ &= 0.4346\end{aligned}$$

This is not a small value. Thus, it should not shock me if more than 20% of my (rather small) sample is below the poverty line.

Example 2: More Poverty

Example

According to the US Census, 18% of Americans are below the poverty line. If I randomly sample $n = 100$ people from the United States, what is the probability that more than 20% of them are below the poverty line?

Solution. We are asked to calculate $\mathbb{P}[\ P > 0.20\ ]$. Thus, since the probability statement deals with P, we need to know the distribution of P.

From the corollary, we know that the approximate distribution of the sample proportion is

$$P \dot{\sim} \mathcal{N}\left(0.18,\ \frac{0.18(0.82)}{100}\right) = \mathcal{N}(0.18,\ 0.001476)$$

Example 2: More Poverty

Solution (cont.). We have $P \dot{\sim} \mathcal{N}(0.18,\ 0.001476)$. Thus,

$$\begin{aligned}\mathbb{P}[\ P > 0.20\ ] &\approx 1 - \mathbb{P}[\ P \leq 0.20\ ] \\ &= \texttt{1 - pnorm(0.20, m=0.18, s=sqrt(0.001476))} \\ &= 0.30133\end{aligned}$$

This is *also* not a small value. Thus, it should not shock me if more than 20% of my larger sample is below the poverty line.

Example 3: Much More Poverty

Example

According to the US Census, 18% of Americans are below the poverty line. If I randomly sample $n = 1000$ people from the United States, what is the probability that more than 20% of them are below the poverty line?

Solution. We are asked to calculate $\mathbb{P}[\ P > 0.20\ ]$. Thus, since the probability statement deals with P, we need to know the distribution of P.

From the corollary, we know that the approximate distribution of the sample proportion is

$$P \dot{\sim} \mathcal{N}\left(0.18,\ \frac{0.18(0.82)}{1000}\right) = \mathcal{N}(0.18,\ 0.0001476)$$

Example 3: Much More Poverty

Solution (cont.). We have $P \dot{\sim} \mathcal{N}(0.18,\ 0.0001476)$. Thus,

$$\begin{aligned}\mathbb{P}[\ P > 0.20\ ] &\approx 1 - \mathbb{P}[\ P \leq 0.20\ ] \\ &= \texttt{1 - pnorm(0.20, m=0.18, s=sqrt(0.0001476))} \\ &= 0.0499\end{aligned}$$

Is *this* a small value? If we decide it is, then I need to question whether my sample was representative of the population *or* whether my assumptions about poverty are incorrect.

On the other hand, if we decide it is not particularly small, then observing more than 20% below the poverty line in my sample is a reasonable result of our sample.

Example 4: Much, Much More Poverty

Example

According to the US Census, 18% of Americans are below the poverty line. If I randomly sample $n = 10,000$ people from the United States, what is the probability that more than 20% of them are below the poverty line?

Solution. We are asked to calculate $\mathbb{P}[\ P > 0.20\ ]$. Thus, since the probability statement deals with P, we need to know the distribution of P.

From the corollary, we know that the approximate distribution of the sample proportion is

$$P \dot{\sim} \mathcal{N}\left(0.18,\ \frac{0.18(0.82)}{10000}\right) = \mathcal{N}(0.18,\ 0.00001476)$$

Example 4: Much, Much More Poverty

Solution (cont.). We have $P \dot{\sim} \mathcal{N}(0.18,\ 0.00001476)$. Thus,

$$\begin{aligned}\mathbb{P}[\ P > 0.20\ ] &\approx 1 - \mathbb{P}[\ P \leq 0.20\ ]\\ &= \texttt{1 - pnorm(0.20, m=0.18, s=sqrt(0.00001476))}\\ &= 0.000\,000\,096\,585\end{aligned}$$

Without question, this is a small value. Thus, I need to question whether my sample was representative of the population *and* whether my assumptions about poverty are incorrect.

Today's Summary

Now that we have concluded this lecture, you should be able to

1. state the Central Limit Theorem
2. state the requirements for applying it
3. state its consequences
    - with respect to the distribution of the sample mean
    - with respect to the distribution of the sample proportion

Today's R Functions

In this slide deck, we covered three R functions. This is in addition to ones we have already experienced and ones we *will* experience in the future:

- `pbinom(x, size,prob)` is the CDF for the Binomial, $\mathbb{P}[\ X \leq x\ ]$
- `pnorm(x, m,s)` is the CDF for the Normal $= F(x) = \mathbb{P}[\ X \leq x\ ]$

Supplemental Activities

The following are supplements for the topics covered today.

- SCA 7a is for the distribution of the sample mean.
- SCA 7c is for bootstrapping, a general method for estimating confidence intervals.

Note that you can access all Statistical Computing Activities here:
`https://www.kvasaheim.com/courses/stat200/sca/`

In addition to the SCA, **Laboratory Activity C** is helpful for learning how to handle some continuous distributions (including the Normal distribution). The lab actually illustrates the Central Limit Theorem, which is central to why the Normal can be used to approximate the Binomial.
`https://www.kvasaheim.com/courses/stat200/labs/`

Supplemental Readings

The following are some readings that may be of interest to you in terms of understanding continuous distributions, including the Exponential:

- Hawkes Learning: Chapter 7
- Intro to Modern Statistics: Section 16.1
- R for Starters: Appendix C

- Wikipedia: Binomial Distribution
  Normal Distribution
  Central Limit Theorem
  Sampling Distributions