Applied Probability II

Section 5: The Central Limit Theorem

Professor Caroline Brophy

Semester 2, 2020-21

Applied Probability II. Section 5: The Central Limit Theorem

Section 5.1: Statement of the Central Limit Theorem (CLT)

Introduction

In Section 4, we saw the following results.

When $X_1, ...X_n \sim N(\mu, \sigma^2)$,

$$\frac{\bar{X} - \mu}{\sigma/\sqrt{n}} \sim N(0, 1)$$

$$\frac{\bar{X} - \mu}{S/\sqrt{n}} \sim T(n-1)$$

$$\frac{(n-1)S^2}{\sigma^2} \sim \chi^2(n-1)$$

We used these to construct confidence intervals and hypothesis tests for μ and σ^2.

However, if the X_is depart from the normal distribution, how can we carry out hypothesis tests and construct confidence intervals for the population mean μ?

The Central Limit Theorem is a very general (and important) results describing the properties of $\bar{X}$ in general.

Statement of the Central Limit Theorem (CLT)

Theorem (without proof)

Let $X_1, X_2, ..., X_n$ be independent and identically distributed random variables with mean μ and variance σ^2, i.e., $E[X_i] = \mu$ and $Var(X_i) = \sigma^2$. Then, any linear combination of the X_i random variables, follows a normal distribution.

Let that sink in for a moment...

In particular, we have not assumed anything about the shape of the distribution of the X_i, only that they are IID with mean μ and variance σ^2.

We will now look at some particular cases of the CLT.

Case 1

Let $X_1, ..., X_n$ be independent random variables.

Let $E[X_i] = \mu$ and $Var(X_i) = \sigma^2$.

Consider $\sum_{i=1}^{n} X_i$. This is a linear combination of the X_is. Therefore, by the CLT,

$$\sum_{i=1}^{n} X_i \underset{approx}{\sim} N(n\mu, n\sigma^2)$$

Where do the mean and variance come from? We can see that

$$E[\sum_{i=1}^{n} X_i] = \sum_{i=1}^{n} E[X_i] = \sum_{i=1}^{n} \mu = n\mu$$

and

$$Var(\sum_{i=1}^{n} X_i) = \sum_{i=1}^{n} Var(X_i) = \sum_{i=1}^{n} \sigma^2 = n\sigma^2$$

(with covariances equal to 0).

Case 2

Let $X_1, ..., X_n$ be independent random variables.

Let $E[X_i] = \mu$ and $Var(X_i) = \sigma^2$.

Consider $\bar{X} = \frac{\sum_{i=1}^{n} X_i}{n}$. This is a linear combination of the X_is. Therefore, by the CLT,

$$\bar{X} \underset{approx}{\sim} N(\mu, \frac{\sigma^2}{n})$$

Where do the mean and variance come from? We can see that

$$E[\frac{\sum_{i=1}^{n} X_i}{n}] = \frac{1}{n}\sum_{i=1}^{n} E[X_i] = \frac{1}{n}\sum_{i=1}^{n} \mu = \frac{n\mu}{n} = \mu$$

and

$$Var(\frac{\sum_{i=1}^{n} X_i}{n}) = \frac{1}{n^2}\sum_{i=1}^{n} Var(X_i) = \frac{1}{n^2}\sum_{i=1}^{n} \sigma^2 = \frac{n\sigma^2}{n^2} = \frac{\sigma^2}{n}$$

(with covariances equal to 0).

The value of the CLT

The Central Limit Theorem is a powerful result.

Case 2 on the previous slide is one of the most common uses of the CLT.

It also tells us that, if $X_1, ..., X_n$ are independent random variables with $E[X_i] = \mu$ and $Var(X_i) = \sigma^2$, then the random variable

$$\frac{\bar{X} - \mu}{\sigma/\sqrt{n}} \underset{approx}{\sim} N(0, 1)$$

The larger the n, the closer the approximation (asymptotic result).

How large is large? Typically, as a rule of thumb, we say that the approximation works well for samples ≥ 30.

This opens the door for us to do tests of hypothesis and construct approximate confidence intervals for population parameters for data that comes from distributions other than the normal distribution.

Common misunderstanding of the CLT

Please do not make this mistake!

If $X_1, ..., X_n$ are independent random variables, and $E[X_i] = \mu$ and $Var(X_i) = \sigma^2$. Then the CLT tells us that, as n increases, $\bar{X}$ tends to a normal distribution. This does not mean that the $X_1, ..., X_n$ values are changing distribution as n increases!

For example, if $X_1, ..., X_n$ are independent and exponentially distributed, as n increases, $\bar{X}$ tends to a normal distribution. But the $X_1, ..., X_n$'s remain exponentially distributed, no matter the n.

Section 5.2: Application of the CLT - finding probabilities

Example

Suppose scores from a national test have a distribution with mean 50 and standard deviation 10.

Suppose 35 results are selected at random from the test scores. What is the probability that their average test result is greater than 55?

- Let the random variable X_i denote an individual test score result.
- We know the $E[X_i] = \mu = 50$ and $Var(X_i) = \sigma^2 = 100$, but we don't know the shape of the distribution (and cannot make the assumption that the X_i are normally distributed).
- Let $\bar{X} = \frac{\sum_{i=1}^{35} X_i}{35}$. The CLT tells us that $\bar{X}$ is approximately normally distributed, or that

$$\bar{X} \underset{approx}{\sim} N(\mu, \frac{\sigma^2}{n}) = N(50, \frac{100}{35})$$

We can use this to work out an approximate probability.

$$\begin{aligned} P(\bar{X} > 55) &= P(\frac{\bar{X} - 50}{\sqrt{100/35}} > \frac{55 - 50}{\sqrt{100/35}}) \\ &= P(Z > 2.96) = 1 - 0.9985 \\ &= 0.0015 \end{aligned}$$

Section 5.3: Application of the CLT - confidence intervals

Example

A study was conducted where a random sample of 500 people were surveyed and asked if they supported a particular measure being imposed by government policy.

- Let p be the true proportion of people in the (large) population that support the policy.
- Let X be a random variable representing the number in a randomly selected sample of 500 people that will answer yes. What distribution does X have?
- $X \sim \text{Binomial}(n, p)$.

Suppose that 315 people of those surveyed support the policy. Then we can estimate the true population proportion p, using

$$\hat{p} = \frac{315}{500} = 0.63$$

But how reliable is this estimate?

Example contd.

We will use the CLT to construct a confidence interval for p.

- We have that $X \sim \text{Binomial}(n = 500, p)$. We have an estimate for p, but do not know the true population value.
- Remember that for the Binomial distribution, $\text{E}[X]=np$ and $\text{Var}(X)=np(1-p)$. Aside challenge (recap from Applied Probability I): prove these in your own time!
- We can write $X = X_1 + ... + X_{500}$, where $X_i = 1$ if the person answers yes, and 0 if the person answer no. Each $X_i \sim \text{Binomial}(n = 1, p)$, or Bernoulli(p).
- This means that X is a sum of independent random variables $X_1, ..., X_{500}$ and each X_i has $\text{E}[X_i]=p$ and $\text{Var}(X_i)=p(1-p)$.
- Then, by the CLT, since n is sufficiently large,

$$X \underset{approx}{\sim} N(np, np(1-p))$$

$$\frac{X - np}{\sqrt{np(1-p)}} \underset{approx}{\sim} N(0, 1)$$

And, we can then get:

$$\frac{X/n - p}{\sqrt{\frac{p(1-p)}{n}}} \underset{approx}{\sim} N(0, 1)$$

Example contd.

To construct a confidence interval for p (the population parameter), use

$$SE(\hat{p}) = \sqrt{\frac{\hat{p}(1-\hat{p})}{n}}$$

This gives us the (approximate) confidence interval

$$\hat{p} \pm z_{\alpha/2}\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}$$

Back to our example, the 95% confidence interval for p, the true proportion of people in the population that support the government policy, is:

$$0.63 \pm 1.96 \times \sqrt{\frac{0.63(1-0.63)}{500}}$$

$$0.63 \pm 1.96 \times 0.021592$$

$$= (0.588, 0.672)$$

We are 95% confident that the true population proportion of people that support the measure lies in this interval.

Section 5.4: Application of the CLT - hypothesis testing

Example dataset: Lateral bias

- There is a well observed phenomenon that human mothers tend to nurse their babies on the left hand side.
- One belief is that the left field of vision causes flow of information in the right hemisphere of the brain (which controls many aspects of social behaviour).

- Is there a preference for the left for other mammals?
- This has been studied in two phylogenetically distant species of mammal: Pacific Walrus and Indian Flying Fox. (Source: Giljov, Karenina and Malashichev, 2018, Biology Letters, 14, 20170707.)

- Want to test whether there is a preference for the left side.
- Think of this like tossing a (possibly biased) coin in each instance, where p (probability of heads) is the probability of the baby being on the left.
- We wish to carry out the hypothesis test

$$\mathrm{H}_0 : p = 0.5, \text{ versus } \mathrm{H}_A : p > 0.5$$

Pacific Walrus and Indian Flying Fox

What was observed?

Whether or not there was a preference for the left was observed for the following scenarios

1. floating on side of mother for Pacific Walrus

and

2. hanging at side of mother for Indian Flying Fox.

Is there evidence of a preference for the left?

Setting up the hypothesis test

In each experiment we are assuming to carry out n independent Bernoulli trials $X_1, ..., X_n$ each with outcome 0 (no lateral bias) or 1 (lateral bias).

We have: $P(X_i = 1) = p, P(X_i = 0) = 1 - p$.

Aside challenge: Prove that the $E[X_i] = p$ and that $Var(X_i) = p(1 - p)$.

Let $X = \sum_{i=1}^{n} X_i$.

We have, $E[X/n] = p$ and $\text{Var}(X/n) = \frac{p(1-p)}{n}$.

With collected data, $\hat{p} = \frac{x}{n}$.

From the CLT,

$$\frac{X/n - p}{\sqrt{\frac{p(1-p)}{n}}} \underset{approx}{\sim} N(0, 1)$$

If we assume a null hypothesis of $H_0 : p = p_0$, then this becomes:

$$\frac{X/n - p_0}{\sqrt{\frac{p_0(1-p_0)}{n}}} \underset{approx}{\sim} N(0, 1)$$

Performing the hypothesis test - Pacific Walrus

For the Pacific Walrus data, $n = 44$.

There were 36 on the left, giving $\sum x_i = 36$.

H_0: $p = p_0$ versus $H_A : p > p_0$.

(One-sided because there was *a priori* belief that there is left lateral bias.)

$\hat{p} = 36/44 = 0.8182$.

Assume that the H_0 is true. If H_0 really is true, then

$$z_{obs} = \frac{0.8182 - 0.5}{\sqrt{\frac{0.5(1-0.5)}{44}}} = 4.221$$

is a realisation from a N(0, 1).

Using $\alpha = 0.05$, we reject the test if $z_{obs} > 1.645$ (where $P(Z > 1.645) = 0.05$). We reject the H_0 and conclude that there is evidence of left preference in Pacifc Walrus.

Performing the hypothesis test - Indian Flying Fox

For the Indian Flying Fox data, $n = 59$.

There were 43 on the left, giving $\sum x_i = 43$.

H_0: $p = p_0$ versus $H_A : p > p_0$.

(One-sided because there was *a priori* belief that there is left lateral bias.)

$\hat{p} = 43/59 = 0.7288$.

Assume that the H_0 is true. If H_0 really is true, then

$$z_{obs} = \frac{0.7288 - 0.5}{\sqrt{\frac{0.5(1-0.5)}{59}}} = 3.515$$

is a realisation from a N(0, 1).

Using $\alpha = 0.05$, we reject the test if $z_{obs} > 1.645$ (where $P(Z > 1.645) = 0.05$). We reject the H_0 and conclude that there is evidence of left preference in Indian Flying Fox.