Lecture 21

The Sample Total and Mean and The Central Limit Theorem

1. Statistics and Sampling Distributions

Suppose we have a random sample from some population with mean μ_X and variance σ_X^2

$$\boxed{\begin{matrix} X \\ \mu_X, \sigma_X^2 \end{matrix}} \dashrightarrow X_1, X_2, \ldots, X_n$$

and a function $w = h(x_1, x_2, \ldots, x_n)$ of n variables. Then (as we know) the combined random variable

$$W = h(X_1, X_2, \ldots, X_n)$$

is called a <u>statistic</u>.

If the population random variable X is discrete then $X_1, X_2, \ldots, X_n$ will all be discrete and since W is a combination of discrete random variables it too will be discrete.

The #64,000 question

How is W distributed?

More precisely, what is the pmf $p_W(x)$ of W.

The distribution $p_W(x)$ of W is called a "sampling distribution".

Similarly if the population random variable X is continuous we want to compute the pdf $f_W(x)$ of W (now it is continuous)

We will jump to § 5.5

The most common $h(x_1, \ldots, x_n)$ is a linear function

$$h(x_1, x_2, \ldots, x_n) = a_1 x_1 + \cdots + a_n x_n$$

where

$$W = a_1 X_1 + a_2 X_2 + \cdots + a_n X_n$$

Proposition L.C (pg 219)

Suppose $W = a_1 X_1 + \cdots + a_n X_n$.

Then

(i) $E(W) = E(a_1 X + \cdots + a_n X_n)$

$= a_1 E(X_1) + \cdots + a_n E(X_n)$

(ii) If $X_1, X_2, \ldots, X_n$ are independent then

$$V(a_1 X_1 + \cdots + a_n X_n) = a_1^2 V(X_1) + \cdots + a_n^2 V(X_n)$$

(so $V(cX) = c^2 V(X)$)

Now suppose $X_1, X_2, \ldots, X_n$ are a random sample from a population of mean μ and variance σ^2 so

$$E(X_i) = E(X) = \mu, \quad 1 \le i \le n$$

$$V(X_i) = V(X) = \sigma^2, \quad 1 \le i \le n$$

and $X_1, X_2, \ldots, X_n$ are independent.

We recall

T_0 = the sample total $= X_1 + \cdots + X_n$

$\overline{X}$ = the sample mean $= \dfrac{X_1 + \cdots + X_n}{n}$

As an immediate consequence of the previous proposition we have

Proposition M

Suppose $X_1, X_2, \ldots, X_n$ is a random sample from a population of mean μ_X and variance σ_X^2. Then

(i) $E(T_0) = n\mu_X$

(ii) $V(T_0) = n\sigma_X^2$

(iii) $E(\bar{X}) = \mu_X$

(iv) $V(\bar{X}) = \frac{\sigma_X^2}{n}$

Proof (this is important)

i) $E(T_0) = E(X_1 + \cdots + X_n)$

by the Prop.

$= E(X_1) + \cdots + E(X_n)$

why

$= \underbrace{\mu_X + \cdots + \mu_X}_{n \text{ copies}}$

$= n\mu_X$

ii) $V(T_0) = V(X_1 + \cdots + X_n)$

by the Prop

$= V(X_1) + \cdots + V(X_n)$

$= \sigma_X^2 + \cdots + \sigma_X^2$

$= n\sigma_X^2$

(iii) $E(\bar{X}) = E\left(\frac{1}{n}(X_1 + \cdots + X_n)\right)$

$= \frac{1}{n} E(X_1 + \cdots + X_n)$

by (i)

$= \frac{1}{n} (n\mu_X)$

$= \mu_X$

(iv) $V(\bar{X}) = V\left(\frac{1}{n}(X_1 + \cdots + X_n)\right)$

by the Pwp

$= \frac{1}{n^2} V(X_1 + \cdots + X_n)$

by (ii)

$= \frac{1}{n^2} (n\sigma_X^2)$

$= \frac{\sigma_X^2}{n}$

□

Remark

It is important to understand the symbols — μ_X and σ^2_X are the mean and variance of the underlying population. In fact they are called the population mean and the population variance. Given a statistic $W = h(X_1, \ldots, X_n)$ we would like to compute $E(W) = \mu_W$ and $V(W) = \sigma^2_W$ in terms of the population mean μ_X and

population variance σ_X^2.

So we solved this problem for $W = \bar{X}$ namely

$$\mu_{\bar{X}} = \mu_X$$

and $$\sigma^2_{\bar{X}} = \frac{1}{n}\sigma^2_X$$

Never confuse population quantities with sample quantities

Corollary

$\sigma_{\bar{X}}$ = the standard deviation of $\bar{X}$

$$= \frac{\sigma_X}{\sqrt{n}} = \frac{\text{population standard deviation}}{\sqrt{n}}$$

Proof

$$\sigma_{\bar{X}} = \sqrt{V(\bar{X})}$$

$$= \sqrt{\frac{\sigma_X^2}{n}}$$

$$= \frac{\sqrt{\sigma_X^2}}{\sqrt{n}} = \frac{\sigma_X}{\sqrt{n}}$$

2. Sampling from a Normal Distribution

Theorem LCN (Linear combination of normal is normal)

Suppose $X_1, X_2, \ldots, X_n$ are independent and $X_1 \sim N(\mu_1, \sigma_1^2), \ldots, X_n \sim N(\mu_n, \sigma_n^2)$.

Let $W = a_1 X_1 + \cdots + a_n X_n$. Then

$$W \sim N(a_1\mu_1 + \cdots + a_n\mu_n, a_1^2\sigma_1^2 + \cdots + a_n^2\sigma_n^2)$$

Proof At this stage we can't prove W is normal (we could if we have moment

generating functions available).

But we can compute the mean and variance of W using Proposition 2.

$$E(W) = E(a_1X_1 + \cdots + a_nX_n)$$
$$= a_1E(X_1) + \cdots + a_nE(X_n)$$
$$= a_1\mu_1 + \cdots + a_n\mu_n$$

and

$$V(W) = V(a_1X_1 + \cdots + a_nX_n)$$
$$= a_1^2V(X_1) + \cdots + a_n^2V(X_n)$$
$$= a_1^2\sigma_1^2 + \cdots + a_n^2\sigma_n^2$$

□

Now we can state the theorem we need.

Theorem N

Suppose $X_1, X_2, \ldots, X_n$ is a random sample from $N(\mu, \sigma^2)$

$$\boxed{X \sim N(\mu, \sigma^2)} \dashrightarrow X_1, X_2, \ldots, X_n$$

Then $T_0 \sim N(n\mu, n\sigma^2)$

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

Proof The hard part is that T_0 and $\bar{X}$ are normal (this is Theorem LCN)

You show the mean of $\overline{X}$ is μ using either Proposition M or Theorem LCN and the same for showing the variance of $\overline{X}$ is $\frac{\sigma^2}{n}$. □

Remark

It is very important for statistics that the sample variance

$$S^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \overline{X})^2$$

satisfies $S^2 \sim \chi^2(n-1)$.

This is one reason that the chi-squared distribution is so important.

In Theorem N we saw that if we sampled n times from a normal distribution with mean μ and variance σ^2 then

i) $T_0 \sim N(n\mu, n\sigma^2)$

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

So both T_0 and $\bar{X}$ are still normal

The Central Limit Theorem says that if we sample n times with n large enough from any distribution with mean μ and variance σ^2 then T_0 has approximately $N(n\mu, n\sigma^2)$ distribution and $\bar{X}$ has approximately $N(\mu, \sigma^2)$ distribution.

We now state the CLT.

The Central Limit Theorem

$$\boxed{X, \mu, \sigma^2} \dashrightarrow X_1, X_2, \ldots, X_n$$

$$\bar{X} \approx N(\mu, \sigma^2) \quad \text{provided } n > 30$$

Remark

This result would not be satisfactory to professional mathematicians because there is no estimate of the error involved in the approximation

However an error estimate is known - you have to take a more advanced course.

The $n > 30$ is a "rule of thumb" In this case the error will be negligible up to a large number of decimal places (but I don't know how many).

So the Central Limit Theorem says that for the purposes of sampling if $n > 30$ then the sample mean behaves as if the sample were drawn from a NORMAL population with the same mean and variation of the actual population

Example 5.27

A certain consumer organization reports the number of major defects for each new automobile that it tests. Suppose that the number of such defects for a certain model is a random variable with mean 3.2 and standard deviation 2.4. Among 100 randomly selected cars of this model what is the probability that the <u>average</u> number of defects exceeds 4.

Solution

Let X_i = # of defects for the i-th car

$$\boxed{\begin{matrix} X \\ \mu = 3.2, \sigma = 2.4 \end{matrix}} \dashrightarrow X_1, X_2, \ldots, X_{100}$$

$n = 100 > 30$ so we can use the CLT

$$\bar{X} = \frac{X_1 + X_2 + \cdots + X_{100}}{100}$$

so

$\bar{X}$ = average number of defects

So we want

$$P(\bar{X} > 4)$$

Now $E(\bar{X}) = \mu = 3.2$

$$V(\bar{X}) = \frac{\sigma^2}{n} = \frac{(2.4)^2}{100}$$

Let $Y \sim N\left(3.2, \frac{(2.4)^2}{100}\right)$

$\sigma_Y = \frac{2.4}{10} = .24$

By the CLT $\bar{X} \approx Y$ so

$$P(\bar{X} \geq 4) \approx P(Y \geq 4)$$

don't use correction for continuity

$$= P\left(\frac{Y - 3.2}{.24} \geq \frac{4 - 3.2}{.24}\right)$$

$$= P\left(Z \geq \frac{.8}{.24}\right) \quad 3.33$$

$$\approx 1 - \Phi(3.33) = 1 - .9996$$

$$= .0004$$

How the Central Limit Theorem Gets Used More Often

The CLT is much more useful than one would expect. That is because many well-known distributions can be realized as sample totals of a sample drawn from another distribution.

I will state this as

General Principle

Suppose a random variable W can be realized as a sample total $W = T_0 = X_1 + \cdots + X_n$ from some X and $n > 30$.

Then W is approximately normal

Examples (This isn't obvious)

1. $W \sim Bin(n, p)$ with n large
2. $W \sim Gamma(\alpha, \beta)$ with α large
3. $W \sim Poisson(\lambda)$ with λ large

We will do the example of $W \sim Bin(n, p)$ and recover (more or less) the normal approximation to the binomial as

CLT $\Rightarrow$ normal approx to binomial

The point is

Theorem (sum of binomials is binomial)

Suppose X and Y are independent, $X \sim Bin(m, p)$ and $Y \sim Bin(n, p)$. Then

$$W = X + Y \sim Bin(m+n, p)$$

"**Proof**"

For simplicity we will assume $p = \frac{1}{2}$.

Suppose Fred tosses a fair coin m times and then Jack tosses a fair coin n times.

Let X = # of heads Fred observes

Y = # of heads Jack observes

So $X \sim Bin(m, \frac{1}{2})$ and $Y \sim Bin(n, \frac{1}{2})$

What is $X+Y$?

Forget who was doing the tossing, $X+Y$ is just the total number of heads in $m+n$ tosses of a fair coin so

$X+Y \sim Bin(m+n, \frac{1}{2})$.

□

Now suppose we have

$\boxed{X \sim Bin(1,p)} \dashrightarrow X_1, \ldots, X_n$

Then $X_i \sim Bin(1,p)$ i $1 \le i \le n$

so $T_0 = X_1 + X_2 + \cdots + X_n \sim Bin(n,p)$

Now if $n > 30$ we know

T_0 is approximately normal so

if $W \sim Bin(n,p)$ and $n > 30$

then $W \approx$ normal

$E(W) = np$ and $V(W) = npq$ AND

$W \approx N(np, npq)$

So we get the normal approximation to the binomial (with $n > 30$ replacing $np \geq 10$ and $nq \geq 10$)

Remark

If $p = \frac{1}{2}$ then the second conditions give $n > 20$ — so better than CLT but

if $p = \frac{1}{5}$ then the second conditions give $n > 50$

— so worse than the CLT