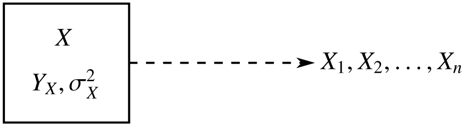
Lecture 21 : The Sample Total and Mean and The Central Limit Theorem

1. Statistics and Sampling Distributions

Suppose we have a random sample from some population with mean μ_X and variance σ_X^2.
In the next diagram Y_X should by μ_X.



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 *statistic.*

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 $h(x_1, \dots, x_n)$ is a linear function

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

where

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

Proposition L (page 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* $X_1, X_2, \dots, X_n$ *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 $V(cX) = c^2 V(X)$*)*

Proposition L (Cont.)

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

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

and $X_1, X_2, \ldots, X_n$ *are independent.*
We recall

$$T_0 = \text{the sample total } = X_1 + \cdots + X_n$$
$$\overline{X} = \text{the sample mean } = \frac{X_1 + \cdots + X_n}{n}$$

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) $E(T_0) = n\mu_{X_2}$

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

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

(iv) $V(\overline{X}) = \dfrac{\sigma_X^2}{n}$

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

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

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

$= n\sigma_X^2$

Proof (Cont.)

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

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

$=$ by (i)

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

$= \mu_X$

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

by the Prop.

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

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_X^2 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 $V(W) = \sigma_W^2$ in terms of the population mean μ_X and the population variance σ_X^2.

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

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

and

$$\sigma_{\overline{X}}^2 = \frac{1}{n}\sigma_X^2$$

Never confuse population quantities with sample quantities.

Corollary

$$\sigma_{\overline{X}} = \textit{the standard deviation of } \overline{X}$$
$$= \frac{\sigma_X}{\sqrt{n}} = \frac{\textit{population standard deviation}}{\sqrt{n}}$$

Proof.

$$\sigma_{\overline{X}} = \sqrt{V(\overline{X})}$$
$$= \sqrt{\frac{\sigma_X^2}{n}}$$
$$= \frac{\sqrt{\sigma_X^2}}{\sqrt{n}} = \frac{\sigma_X}{\sqrt{n}}$$

□

Sampling from a Normal Distribution

Theorem LCN (Linear combination of normal is normal)

Suppose $X_1, X_2, \ldots, X_n$ *are independent and*

$$X_1 \sim N(\mu, \sigma_1^2), \ldots, X_n \sim N(\mu_n, \sigma_n^2).$$

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*

Proof (Cont.)

generating functions available).
But we can compute the mean and variance of W using Proposition L.

$$\begin{aligned} E(W) &= E(a_1 X_1 + \cdots + a_n X_n) \\ &= a_1 E(X_1) + \cdots + a_n E(X_n) \\ &= a_1 \mu_1 + \cdots + a_n \mu_n \end{aligned}$$

and

$$\begin{aligned} V(W) &= V(a_1 X_1 + \cdots + a_n X_n) \\ &= a_1^2 V(X_1) + \cdots + a_n^2 V(X_n) \\ &= a_1^2 \sigma_1^2 + \cdots + a_n^2 \sigma_n^2 \end{aligned}$$

□

Now we can state the theorem we need.

Theorem N

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

Then

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

and

$$\overline{X} \sim N\left(\mu, \frac{\sigma^2}{n}\right)$$

Proof

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

Proof (Cont.)

You show the mean of $\overline{X}$ is μ using either Proposition M or Theorem 10 and the same for showing the variance of $\overline{X}$ is $\dfrac{\sigma^2}{n}$. □

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.

3. The Central Limit Theorem (§5.4)

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) $\overline{X} \sim N\left(\mu, \frac{\sigma^2}{n}\right)$

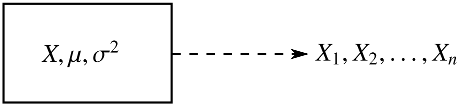
So both T_0 and $\overline{X}$ are still normal

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 $\overline{X}$ has approximately $N(\mu, \sigma^2)$ distribution.

We now state the CLT.

The Central Limit Theorem

In the figure σ^2 should be $\frac{\sigma^2}{n}$



$\overline{X} \approx N(\mu, \frac{\sigma^2}{n})$ 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 neglible up to a large number of decimal places (but I don't know how many).

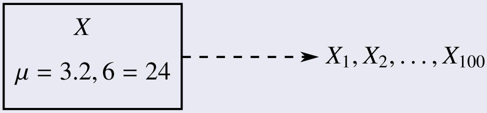
So the Central Limit Theorem says that for the purposes of sampling if $n > 30$ then the sample mean *behaves as if the sample were drawn from a NORMAL population with the same mean and variance equal to the variance of the actual population divided by* n.

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 *average* number of defects exceeds 4.

Solution

Let $X_i = \sharp$ *of defects for the i-th car.*
In the following figure the equation $6 = 24$ *should be* $\sigma = 24$.



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

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

So

$$\overline{X} = \textit{average number of defects}$$

So we want

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

Solution (Cont.)

Now

$$E(\overline{X}) = \mu = 3.2$$

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

Let Y *be a normal random with the same mean and variance as* $\overline{X}$ *so* $\mu_Y = 3.2$ *and* $\sigma_Y^2 = \frac{(2.4)^2}{100}$ *and so*

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

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

By the CLT $\overline{X} \approx Y$ *so*

don't use correction for continuity

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

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

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

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

$$= .0004$$

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

1. $W \sim \mathrm{Bin}(n, p)$ *with n large.*
2. $W \sim$ *Gamma*(α, β) *with* α *large.*
3. $W \sim$ *Poisson*(λ) *with* λ *large.*

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

$$\text{CLT} \Rightarrow \text{normal approx to binomial.}$$

The point is

Theorem (sum of binomials is binomial)

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

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

Proof

For simplicity we will assume $p = \frac{1}{2}$.
Suppose Fred tosses a fair coin m times and Jack tosses a fair coin n times.

Proof (Cont.)

Let

$$X = \sharp \text{ of head Fred observes}$$
$$Y = \sharp \text{ of heads Jack observes}$$

So

$$X \sim \mathrm{Bin}\left(m, \frac{1}{2}\right) \quad \text{and} \quad Y \sim \mathrm{Bin}\left(n, \frac{1}{2}\right)$$

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

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

□

Now suppose we have

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

Then $X_i \sim \text{Bin}(1, p)$, $1 \leq i \leq n$,

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

Now if $n > 30$ we know T_0 is approximately normal so if $W \sim \text{Bin}(n, p)$ and $n > 30$ the $W \approx$ normal

$$E(W) = np \quad \text{and} \quad V(W) = npq \qquad \text{AND}$$

$$W \sim N(np, npq)$$

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 gives $n > 20$.

- so better then CLT but if $p = \frac{1}{5}$ then the second conditions gives $n > 50$.

- so worse than the CLT.