MSPP PPOL 501–02 & -06: Fall 2014
Course Notes # 8: Sampling Distribution of the Sample Mean; The Central Limit Theorem
Professor Carolyn Hill

I. REMINDER OF THE DESCRIPTIVE STATISTIC NOTATION WE'VE SEEN SO FAR:

Descriptive Statistic	Population	Sample
Mean	$\mu = \frac{\sum X_i}{N}$	$\overline{X} = \frac{\sum Xi}{n}$
Variance	$\sigma^2 = \frac{\sum (X_i - \mu)^2}{N}$	$s^2 = \frac{\sum (X_i - \overline{X})^2}{n-1}$
Standard Deviation	$\sigma = \sqrt{\frac{\sum (X_i - \mu)^2}{N}}$	$s = \sqrt{\frac{\sum (X_i - \overline{X})^2}{n-1}}$

II. SAMPLING DISTRIBUTION OF the SAMPLE MEAN

- First of all, what is a sampling distribution of the sample mean, and how does this differ from the types of population distributions we have been talking about?

"For a variable *X* and a given sample size, the distribution of the variable $\overline{X}$ is called the sampling distribution of the sample mean" (Weiss, p. 298)

Note: the following example is drawn from Colm O'Muircheartaigh's notes.

Example of a sampling distribution of the sample mean. (note: We'll use a very small population and a very small sample so you can get a feel for what's going on.)

N = 6 elements: A B C D E F

n= 2 elements: What are the possible samples of *n*=2 that could be selected from this population?

AB AC AD AE AF
BC BD BE BF
CD CE CF
DE DF
EF

Thus, 15 samples of size *n*=2 are possible. There is a 1/15 chance that any one of these samples would be selected.

- To make the example concrete, suppose that each letter in the above population represents the amount of income in tens of thousands of dollars:

A = 3
B = 6
C = 4
D = 9
E = 7
F = 7

The mean income for this population is $\mu = 6$, or \$60,000.

- If we draw samples of *n*=2 from this population, what possible sample means $\overline{X}$ might we calculate?

Individuals in the Sample	Values in the Sample	Sample Mean $\overline{X}$
A, B	3, 6	4.5
A, C	3, 4	3.5
A, D	3, 9	6.0
A, E	3, 7	5.0
A, F	3, 7	5.0
B, C	6, 4	5.0
B, D	6, 9	7.5
B, E	6, 7	6.5
B, F	6, 7	6.5
C, D	4, 9	6.5
C, E	4, 7	5.5
C, F	4, 7	5.5
D, E	9, 7	8.0
D, F	9, 7	8.0
E, F	7, 7	7.0

- The possible sample means we would draw range from $\overline{X}$ =3.5 to $\overline{X}$ =8. Each of these sample means would provide an estimate of the population mean μ, but clearly some would give inaccurate estimates (and only one sample – A, D – would give the exact value of the population mean!).

- We can construct a frequency table of the sample means to start moving toward the sampling distribution of the sample mean:

$\overline{X}$	Frequency	Relative Frequency	$\overline{X} * f$
3.5	1	1/15	3.5
4.5	1	1/15	4.5
5.0	3	3/15	15.0
5.5	2	2/15	11.0
6.0	1	1/15	6.0
6.5	3	3/15	19.5
7.0	1	1/15	7.0
7.5	1	1/15	7.5
8.0	2	2/15	16.0
TOTAL	15	15/15	90.0

- The sample mean $\overline{X}$ is a random variable because it varies from sample to sample depending on the result of the sample that is actually drawn (at random!).

- The population mean μ.is the parameter we're trying to estimate with the sample mean $\overline{X}$.

- The sample mean random variable $\overline{X}$ is the estimator of the population mean μ.

- The observed value of $\overline{X}$ for a particular sample is the estimate of the population mean μ obtained from that particular sample.

- "Sampling error": error resulting from using a sample to estimate a population parameter.

→ the Mean of the Sample Mean

- We can describe the sampling distribution of the sample mean and judge its usefulness as an estimator of the population parameter by *calculating the mean and dispersion of the sampling distribution of the random variable* $\overline{X}$.

- For samples of size *n*, the mean of the random variable $\bar{X}$ equals the mean of the random variable X_i : the expected value of the random variable $\bar{X}$ is equal to the population mean for the random variable X_i: $E(\bar{X}) = \mu$

- NOTE: An estimator is an <u>unbiased estimator</u> of the population parameter of interest if the expected value of the estimator is equal to the parameter being estimated.

- The mean of this sampling distribution of sample means = 90/15 = 6.

- Thus, in the case above, we see that $\bar{X}$ is an unbiased estimator of μ since 6=6.

- More generally, whatever population we're sampling from, and however small or large the sample size, <u>the sample mean is always an unbiased estimator of the population mean in simple random sampling.</u> Why is this the case?

* Recall that we can think of the sample mean, $\bar{X}$, as a random variable. Further, recall that the "expected value" of a random variable is like taking the mean of the random variable.

* The definition of a sample mean: $\bar{X} = \dfrac{\sum_{i=1}^{n} X_i}{n} = \dfrac{(X_1 + X_2 + X_3 + X_4 + ... + X_n)}{n}$

* What's the expected value (i.e., the mean) of the sample mean?:

$$E(\bar{X}) = E\left(\frac{\sum_{i=1}^{n} X_i}{n}\right) = E\left(\frac{(X_1 + X_2 + X_3 + X_4 + ... + X_n)}{n}\right)$$

1/n is taken out because it is a constant: the expected value of a constant is a constant.

$$= \frac{1}{n} E(X_1 + X_2 + X_3 + X_4 + ... + X_n)$$

$$= \frac{1}{n} E(X_1) + E(X_2) + E(X_3) + E(X_4) + ... + E(X_n)$$

$$= \frac{1}{n} (\mu + \quad \mu + \quad \mu + \quad \mu + \quad \mu)$$

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

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

The expected value (i.e., mean) of each individual value in the distribution is, by definition, the population mean, μ.

Wah-lah! We've proven that the expected value of Xbar is equal to the population mean. Nowhere here did we say that the sample had to be a certain size, nor did we make any assumptions about the shape of the population distribution.

→ the Standard Deviation of the Sample Mean

For samples of size *n*, the standard deviation of the random variable $\overline{X}$ equals the standard deviation of the random variable X_i, divided by the square root of the sample size. This is known as the "standard error" of the sample mean:

$$\text{SE} = \sigma_{\overline{X}} = \frac{\sigma}{\sqrt{n}}$$

- In addition to the sample mean as a measure of central tendency, and as an unbiased estimator for the population mean of some population of interest, we're also interested in the dispersion of the sampling distribution of sample means.

  -- even if the sample mean as an estimator of the population mean is right "on average," it may not be a desirable measure to make inferences to the population if the dispersion of these sampling means is large

  (remember – we usually only draw *one* sample from the population – what if the sample we draw has a good chance of being quite different from the population mean???)

  -- can use the same idea of the variance of a distribution that we used before. In this case, we find the variance of the sampling distribution of sample means:

$$Var(\overline{X}) = Var\left(\frac{\sum_{i=1}^{n} X_i}{n}\right) = Var\left(\frac{(X_1 + X_2 + X_3 + X_4 + ... + X_n)}{n}\right)$$

$$= \frac{1}{n^2} Var(X_1 + X_2 + X_3 + X_4 + ... + X_n)$$

$1/n^2$ is taken out because it is a constant. If a random variable is multiplied by a constant, the variance increases by the square of that constant.

$$= \frac{1}{n^2}\left[Var(X_1) + Var(X_2) + Var(X_3) + Var(X_4) + ... + Var(X_n)\right] =$$

$$= \frac{1}{n^2}\left[\sigma^2 + \sigma^2 + ... + \sigma^2\right]$$

$$= \frac{1}{n^2}\left[n\sigma^2\right]$$

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

- The standard deviation of the sampling distribution of sample means is just the square root of this value. This term has a special name: the standard error of the sample mean:

$$\text{SE} = \sigma_{\overline{X}} = \frac{\sigma}{\sqrt{n}}$$

- O.K. so what's the relationship between the variance of the sampling distribution and the variance of the population? Before we go on, step back and think:

  The standard error can be *reduced* either (1) by increasing the sample size or (2) by having a smaller standard deviation of the population. Reducing the standard error increases its efficiency, and vice versa.

- *Efficiency* refers to the "lack of dispersion" of the estimator.

III. THE NORMAL DISTRIBUTION AND THE CENTRAL LIMIT THEOREM

- So far, we've talked about *individual observations* (or scores) in some *population of observations*, where the sample or population values on the variable of interest are normally distributed. And, we've just defined the sampling distribution of sample means. This section brings these two ideas together.

- First: few random variables are actually Normally distributed. What are the shapes of some other distributions? (e.g., see Weiss, p. 72)

- The *Central Limit Theorem* provides an important bridge between the (1) nice properties of the normal distribution; and (2) the fact that the distribution of individual elements of many samples or populations are *not* normally distributed; yet we still want to make inferences from the sample to the populations of interest.

****

CLT: Regardless of the shape of the population distribution, the sampling distribution of the sample mean $\bar{X}$ for simple random samples of size *n* will approach the Normal distribution whenever the sample size *n* is sufficiently large.

If the population distribution is itself Normal, the sample mean $\bar{X}$ will have a Normal distribution for any sample size *n.*

Alt:

CLT: In random sampling from any population with $E(X)=\mu$ and $Var(X)=\sigma^2$,

$$\bar{X}_n \sim N(\mu, \frac{\sigma^2}{n}) \qquad \Leftrightarrow \qquad \frac{(\bar{X}-\mu)}{\sigma/\sqrt{n}} \sim N(0,1)$$

****

-- Fundamental and amazing result: no matter what the shape of the population distribution for a the individual values of a variable of interest (skewed or symmetric, rectangular of J-shaped, etc.), the sampling distribution of the sample mean $\bar{X}$ will always approach the Normal distribution if the sample size, *n*, is large enough.

- Important points about the CLT:

(1) Although the CLT applies to population distributions of all shapes, sampling distributions from different population distribution shapes will approach the Normal distribution at different rates. ("the farther the variable under consideration is from being normally distributed, the larger the sample size must for a normal distribution to provide an adequate approximation to the distribution of $\bar{X}$" (Weiss p. 311) (see Figure 7.6, p. 313 in Weiss)

(2) What is a "sufficiently large" sample? A sample size of at least n=30 or so is usually sufficient for a reasonable approximation (approximation improves as sample size increases though).

(3) Simple random sampling is the basis of the CLT. If probability sampling is not used, the CLT does not apply and statistical inference collapses.