Stat 400, section 5.3-5.4 Sampling Distributions & the Central Limit Theorem

notes by Tim Pilachowski

If you haven't done it yet, go to the Stat 400 page and download the handout 5.4 supplement Central Limit Theorem. The homework (both practice and hand-in) homework for section 5.4 will be from that supplement.

From the previous Lecture, 5.3a Populations and Samples:

Random variables $X_1, X_2, \ldots, X_n$ form a (simple) random sample of size n if they meet two (important) requirements:

1. The X_i's are independent random variables.
2. Every X_i has the same probability distribution.

Data is collected from the sample, i.e., the random variables $X_1, X_2, \ldots, X_n$ each receive values $x_1, x_2, \ldots, x_n$. These values are used to calculate sample statistics. The sample statistics we'll be most interested in are:

1. The sample total $T_0 = X_1 + X_2 + \ldots + X_n$.
2. The sample mean $\overline{X} = \frac{1}{n} * T_0$. (The calculated sample mean is symbolized by $\bar{x}$.)
3. The sample variance $S^2 = \frac{1}{n-1} * \sum_{i=1}^{n} \left(X_i - \overline{X}\right)^2$. (The calculated sample variance is symbolized by s^2.)

Note that T_0, $\overline{X}$ and S^2 are themselves random variables. Sections 5.3 and 5.4 focus on what the probability distributions of these random variables look like, and what they can tell us about the overarching population's distribution.

Theory

Probability models exist in a theoretical world where everything is known. If you constructed every possible sample of a specified size n from a given population (Example 1 in the supplement), or were able to toss a coin an infinite number of times (Example 2 in the supplement), you would create what statisticians call a **sampling distribution**.

In the Examples below, we'll use a hypothetical population Ψ consisting of the numbers 10, 20, 30, 40 and 50. The parameter and statistic we'll consider first is the mean.

Example A-1: Calculate the mean and standard deviation of a population Ψ which consists of elements from the set {10, 20, 30, 40, 50} with probabilities given in the table below.

X	10	20	30	40	50
$P(X = x)$	0.4	0.2	0.2	0.0	0.2

Example A-2: Construct a histogram for population Ψ.

We would get the same histogram if we were to consider all possible samples of size n = 1 that could be taken from the population Ψ and calculated each sample's expected value (mean).

Example A-3: Construct all possible samples of size $n = 2$ that can be made from the elements of Ψ, designate the probability of each being picked, and calculate each sample's expected value (mean).

sample	P	$\bar{x}$	sample	P	$\bar{x}$	sample	P	$\bar{x}$	sample	P	$\bar{x}$
10, 10			20, 10			30, 10			50, 10		
10, 20			20, 20			30, 20			50, 20		
10, 30			20, 30			30, 30			50, 30		
10, 50			20, 50			30, 50			50, 50		

Example A-4: Draw the histogram for the sampling distribution from Example A-3.

$\bar{X}$	10	15	20	25	30	35	40	45	50
$P(\bar{X} = \bar{x})$									

Example A-5: Calculate mean and standard deviation of the sampling distribution of Ψ for sample size $n = 2$.

Mean of sampling distribution = $E(\bar{X})$ =

$$10(0.16)+15(0.16)+20(0.20)+25(0.08)+30(0.20)+35(0.08)+40(0.08)+45(0.0)+50(0.04)=$$

Variance of sampling distribution = $V(\bar{X})$ =

$$(10-24)^2(0.16)+(15-24)^2(0.16)+(20-24)^2(0.20)+(25-24)^2(0.08)+(30-24)^2(0.20)+(35-24)^2(0.08)$$
$$+(40-24)^2(0.08)+(45-24)^2(0.0)+(50-24)^2(0.04)=$$

Standard deviation of sampling distribution $=\sigma(\bar{X})=\sigma_{\bar{X}}=$ standard error =

Theory:

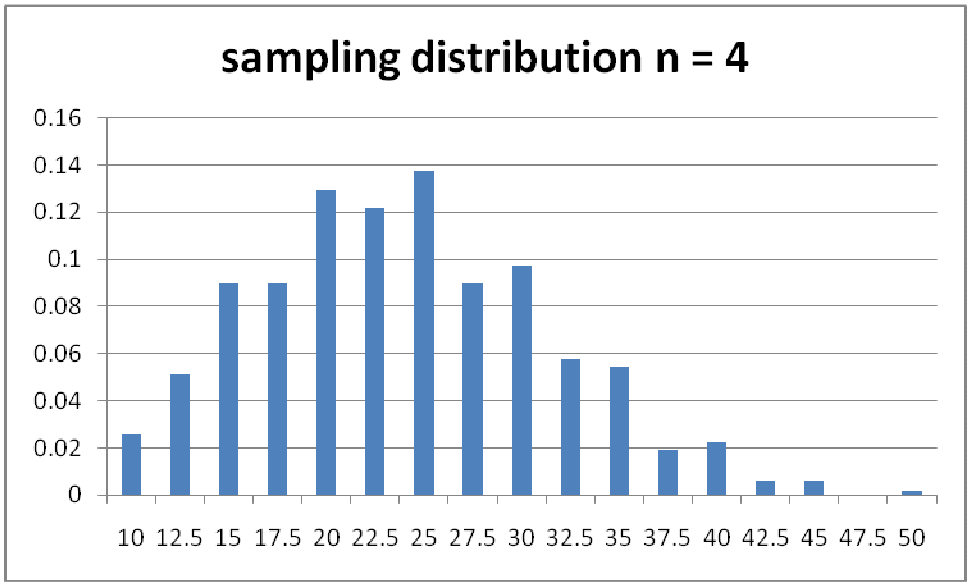
Example A-6: Draw the histogram and calculate the mean and standard deviation of the sampling distribution of Ψ for sample size $n = 4$.

Mean of sampling distribution = $E(\overline{X})$ =

Variance of sampling distribution = $\mathrm{Var}(\overline{X})$ =

Standard deviation of sampling distribution $= \sigma(\overline{X}) = \sigma_{\overline{X}}$ = standard error =

The histogram for this sampling distribution (sample size $n = 4$) looks like this.



Example A-7: Explore the histogram for sampling distributions of Ψ for various sample sizes. (That is, conduct a series of simulation experiments using various values of n.)

We'll use one of two sources:
http://www.chem.uoa.gr/applets/AppletCentralLimit/Appl_CentralLimit2.html
http://www.intuitor.com/statistics/CentralLim.html.

Notes following Examples A:

In Example A-1 we found $E(X)$ and $\sigma(X)$. In Examples A-5 and A-6, we found that $E(\overline{X}) = E(X)$ and $\sigma(\overline{X}) = \frac{\sigma(X)}{\sqrt{n}}$. Will the same be true for any sample size n?

4.3 Example D revisited. Distribution of ACT scores is approximately normal. In 2010 mean score for the ACT = 22.6, with a standard deviation of 4.3. What is the probability that ***a single student chosen at random*** has an ACT score between 22 and 24? *answer*: 0.1850
(source: *Usefulness of High School Average and ACT Scores in Making College Admission Decisions*, retrieved from www.act.org/research/researchers/reports/pdf/ACT_RR2010-2.pdf.)

Theory: sampling distribution for a normally distributed population

4.3 Example D – a new question. Distribution of ACT scores is approximately normal. In 2010 mean score for the ACT = 22.6, with a standard deviation of 4.3. a) If a random sample of 50 students who took the ACT is selected, what is the shape of the resulting sampling distribution? b) What are $E(\overline{X})$ and $\sigma_{\overline{X}}$? c) What is the probability that ***the sample mean*** is between 22 and 24? *answers*: normal; 22.6, ≈ 0.6081; 0.8282

This is fine for a population known to be normally distributed, but can we make any statements about sampling distributions from a population which may not (or definitely do not) have a normal probability distribution?

Recall the various versions of Example A done earlier. Population Ψ did not have a normal distribution, and was not even symmetric. However, the shape of the sampling distribution took on a shape close to that of a normal distribution as *n* increased.

Enter the Central Limit Theorem:

Given a population with mean μ_X and standard deviation σ_X:

1) As the sample size *n* increases, or as the number of trials *n* approaches infinite, the shape of a sampling distribution becomes increasingly like a normal distribution.

2) The mean of sampling distribution = the mean of the population, $E(\overline{X}) = \mu_X$.

3) The standard deviation of sampling distribution = standard error, $\sigma_{\overline{X}} = \frac{\sigma_X}{\sqrt{n}}$.

The proof of (1) in the Central Limit Theorem requires "moment generating functions" which we do not have yet, and which we may (but probably won't) get to before the end of the semester.

For statistics, a sample size of 30 is usually large enough to use the normal distribution probability table for hypothesis tests and confidence intervals. For Lecture examples, and for homework exercises from the handout, we'll use the normal distribution table to find various probabilities for sample statistics.

The ***Central Limit Theorem*** tells us, in short, that a sampling distribution is often close to a normal distribution.

What does this mean for random sampling? It tells us that 68% of the time, a random sample will give us a result—a statistic—within 1 standard deviation of the "true" parameter. We would expect that 95% of the time, a random sample will give a statistic within 2 standard deviations of the population parameter, and 99.7% of the time, a random sample will give a statistic within 3 standard deviations of the population parameter.

In statistics, the Central Limit Theorem is the justification for constructing confidence intervals and conducting hypothesis tests.

Example B. A population has mean $\mu = 150$ and standard deviation $\sigma = 22$. For a random sample of size 47, calculate a) the expected value of the sample mean and b) the standard error. Find the following probabilities: c) $P\left(145 < \overline{X} < 153\right)$ and d) $P\left(\overline{X} > 154\right)$.

answers: 150, $\frac{22}{\sqrt{47}}$, 0.7644, 0.1056

WARNING*:** Example B is ***not like those we did in section 4.3! That is, it does not find probabilities for single values of *X* for a normally-distributed population, which uses $Z = \frac{X - \mu_X}{\sigma_X}$. (In fact, because we do not know the probability distribution, we cannot specify probabilities for individual subjects.)

Rather, Example B looks at one sample and finds probabilities involving the mean of that sample, $\overline{X}$, considered as part of all the hypothetical samples which could have been constructed (the sampling distribution). Therefore, for example B we used $Z = = \frac{\overline{X} - \mu_X}{\sigma_{\overline{X}}} = \frac{\overline{X} - \mu_X}{\sigma_X / \sqrt{n}}$.

Also, in some homework exercises, you'll need to use the skills from previous sections to determine μ_X and σ_X.

Example C. A random variable X has probability density function $f(x)=\begin{cases}6x(1-x) & 0\le x\le 1\\ 0 & \text{otherwise}\end{cases}$.

a) What are the expected value and standard deviation for a single randomly-chosen value of X?

b) You randomly select a sample of size n = 100. What is the expected value for the sample mean, $E(\overline{X})$? What is the standard error, $\sigma_{\overline{X}}$, for the sampling distribution?

c) What is the probability that a single randomly-chosen subject from this population will exhibit a value of at least 0.55?

d) You randomly select a sample of size n = 100. What is the probability that the sample mean will be at least 0.55?

e) You randomly select a sample of size n = 100. There is a 25% probability that the sample mean will be below what value?

answers: $\frac{1}{2}$, $\frac{1}{2\sqrt{5}}$; $\frac{1}{2}$, $\frac{1}{20\sqrt{5}}$; 0.42525; 0.0125; 0.485

Go to the supplement for more examples worked out. There are ones similar to each of the homework exercises.