STT 315 Spring 2006
Week 6 slides.

These slides cover portions of chapters 5 and 6.

Double bar $\overline{\overline{X}}$ will be used for the sample mean because the single bar does not print clearly enough.

Sample mean is unbiased for population mean μ.
For every sample size n,

$$E\,\overline{\overline{X}} = \mu.$$

That is, the average value taken by the sample mean $\overline{\overline{X}}$ (in all of its possibilities) is identical with the population mean μ. This is true whether sampling with replacement (i.e. independently) or sampling without-replacement instead.

For example, accounts may be audited to determine the error in the balance due. The following are the **same**

E $\overline{\overline{X}}$, the average value taken by the sample mean $\overline{\overline{X}}$.
E X, the average value of a sample of one X (i.e. n=1).
μ, the average of the population distribution.

That E X = μ follows from the fact that a random sample of one (i.e. X) has the same distribution as the population whether sampling with or without replacement.

That E $\overline{\overline{X}}$ = E X follows from the properties of expectation.

$$E\,\overline{\overline{X}} = E\frac{X_1 + \ldots + X_n}{n} = \frac{E\,X_1 + \ldots + E\,X_n}{n} = \frac{n\,E\,X}{n} = E\,X.$$

Since E $\overline{\overline{X}} = \mu$ **regardless of the population** having finite mean μ we

say that $\overline{\overline{X}}$ is an **unbiased** estimator for μ. It carries the meaning that the distribution of $\overline{\overline{X}}$ balances at μ **in all cases**. This is true whether we sample with-replacement or without-replacement.

Another way to write the above is

$\mu_{\overline{\overline{X}}} = \mu$ (the mean of the list of all possible $\overline{\overline{X}}$ is μ).

Variance of $\overline{\overline{X}}$ differs when sampling with/without-replacement.

With replacement. For every sample size n with-replacement

$$\textbf{Variance } \overline{\overline{X}} = \frac{\sigma^2}{n}.$$

We can **prove it** using rules of expectation, and Variance for a sum of independent random variables Letting $X_1, \ldots, X_n$ denote the individual sample scores,

$$\begin{aligned}
\text{Var } \overline{\overline{X}} &= \text{Var}\left(\frac{X_1 + \ldots + X_n}{n}\right) = \left(\frac{1}{n}\right)^2 \text{Var}(X_1 + \ldots + X_n) \\
&= \left(\frac{1}{n}\right)^2 (\text{Var } X_1 + \ldots + \text{Var } X_n), \text{ by independence} \\
&= \left(\frac{1}{n}\right)^2 \text{n Var } X_1, \text{ since all have the same distribution} \\
&= \frac{\sigma^2}{n}
\end{aligned}$$

The s.d. of $\overline{\overline{X}} = \sqrt{\frac{\sigma^2}{n}} = \frac{\sigma}{\sqrt{n}}$. We write $\sigma_{\overline{\overline{X}}} = \frac{\sigma}{\sqrt{n}}$ where $\overline{\overline{X}}$ denotes the sample mean of a sample of n with-replacement .

Without-replacement. For every sample size n selected **without**-replacement from a population of size N, we state without proof that

$$\textbf{Var } \overline{\overline{X}} = \frac{N-n}{N-1}\,\frac{\sigma^2}{n}.$$

The s.d. of $\overline{\overline{X}}$ is thus $\sigma_{\overline{\overline{X}}} = \sqrt{\frac{N-n}{N-1}\,\frac{\sigma^2}{n}} = \sqrt{\frac{N-n}{N-1}}\,\frac{\sigma}{\sqrt{n}}$.

The term $\sqrt{\frac{N-n}{N-1}}$ is called the finite population correction **FPC**.

The Central Limit Theorem (CLT). The CLT asserts that the distribution of the sample average will be **approximated** by a normal distribution having the mean μ and standard deviation $\frac{\sigma}{\sqrt{n}}$ (or $\sqrt{\frac{N-n}{N-1}}\ \frac{\sigma}{\sqrt{n}}$ in the without-replacement case).

This remarkable result provides many benefits. For starters, the FPC tells us very precisely the nature of what added precision may be gained by sampling **without**-replacement instead of sampling with replacement. We can see that in most cases the improvement is slight since $\sqrt{\frac{N-n}{N-1}} \approx 1$ when sampling relatively few times n from a large population N.

Proper statement of CLT. Proper statement of the CLT in the with-replacement case is

$$P(\frac{\overline{\overline{X}} - \mu}{\frac{\sigma}{\sqrt{n}}} < z) \to P(Z < z) \text{ for every z, as } n \to \infty.$$

The above statement suits us just fine. It is precisely what we require when justifying our use of the standard normal table.

Sampling **without** replacement is necessarily more complicated. After all, when n reaches the population size N any sample **without** replacement can go no further. Nonetheless

$$P(\frac{\overline{\overline{X}} - \mu}{\sqrt{\frac{N-n}{N-1}}\ \frac{\sigma}{\sqrt{n}}} < z) \approx P(Z < z) \text{ for every z}$$

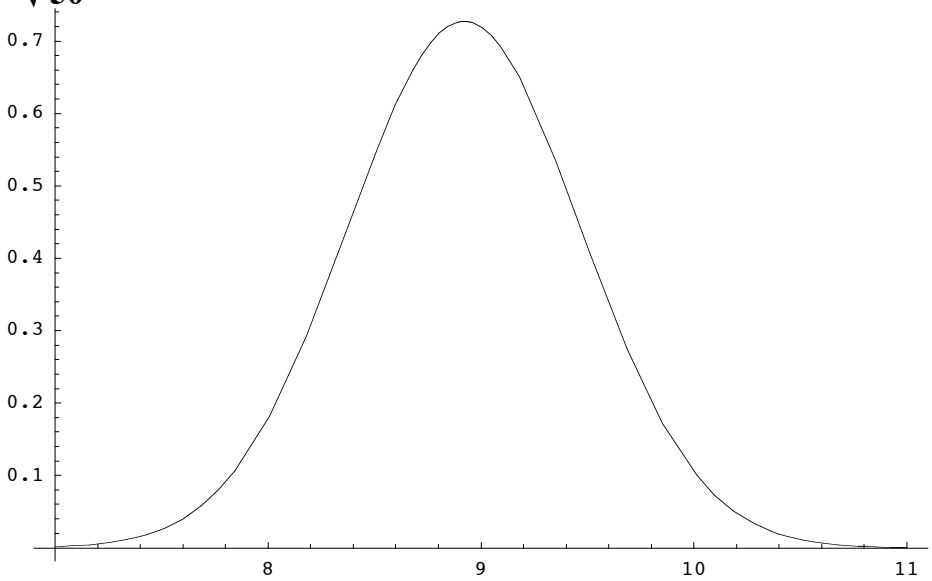
provided n and N-n are both large and other conditions are met.

As a practical matter, regard the central limit theorem as applica-

ble when you are asked to apply it in this course. Your textbook will invite you to use normal approximations in exercises without addressing the question of whether the assumptions of the CLT might be satisfied. It is best that you regard this as "school practice" not to be repeated in serious work.

Example 1. 4207 trucks owned by a delivery service average 8.92 miles per gallon (in a strict test of mpg) with standard deviation 3.88. Not knowing this, the company undertakes a mpg test on a with-replacement sample of 50 trucks.

a. Sketch the approximate distribution of the sample mean mpg $\overline{\overline{X}}$ of the 50 trucks. **ans. The normal density with mean 8.92 and s.d.** $\frac{3.88}{\sqrt{50}}$ **= 0.548715.**



b. To determine the probability that the company underestimates fleet mpg (average) by more than one mpg we need first calculate the standard score of 7.92. What is z = standard score of 7.92 (in the scale of the distribution of $\overline{\overline{X}}$)? **ans.** $\mathbf{z = \frac{7.92-8.92}{0.548715} = -1.82.}$

c. Using (b) determine the probability $P(\overline{\overline{X}} < 7.92)$ that the sample mean $\overline{\overline{X}}$ underestimates $\mu = 8.92$ by more than one mpg. **ans.**

$P(\overline{\overline{X}} < 7.92) = P(Z < -1.82) = P(Z > 1.82) = 0.5 - P(0 < Z <$ 1.8 | 2 **).**

z	0.0 | 2
1.8	0.46562

$P(Z > 1.82) = 0.5 - P(0 < Z < 1.82) = 0.5 - 0.46562 = 0.03438.$

Example 2. 4207 trucks owned by a delivery service average 8.92 miles per gallon (in a strict test of mpg) with standard deviation 3.88. Not knowing this, the company undertakes a mpg test on a **without**-replacement sample of 50 trucks.

a. Sketch the approximate distribution of the sample mean mpg $\overline{\overline{X}}$ of the 50 trucks. **ans. The normal density with mean 8.92 and s.d.** $\sqrt{\frac{4702-50}{4702-1}}\,\frac{3.88}{\sqrt{50}}$ **= 0.545848.**

b. To determine the probability that the company underestimates fleet mpg (average) by more than one mpg we need first calculate the standard score of 7.92. What is z = standard score of 7.92 (in the scale of the distribution of $\overline{\overline{X}}$)? **ans.** $\mathbf{z = \frac{7.92-8.92}{0.545848} = -1.83.}$

c. Using (b) determine the probability $P(\overline{\overline{X}} < 7.92)$ that the sample mean $\overline{\overline{X}}$ underestimates $\mu = 8.92$ by more than one mpg. **ans.**

$P(\overline{\overline{X}} < 7.92) = P(Z < -1.83) = P(Z > 1.83) = 0.5 - P(0 < Z <$ 1.8 | 3 **).**

z	0.0 | 3
1.8	0.46562

$P(Z > 1.82) = 0.5 - P(0 < Z < 1.82) = 0.5 - 0.46638 = 0.03362.$

d. Comparing 1(c) with 2(c) which sampling method runs the greater risk of underestimating fleet mpg by more than one mpg? **ans. Sampling with replacement runs the greater risk of underestimating fleet average by more than one mpg. This is because both sampling methods have the same mean of 8.92 = μ but sampling without replacement has the smaller s.d. for $\overline{\overline{X}}$ and therefore tends to underestimate and overestimate) less frequently by any given amount.**

Note: were it not for approximate normality of the distribution of $\overline{\overline{X}}$ it would not follow that the plan with smaller s.d. underestimates less frequently. It is a dividend of "being in control" that simple comparisons such as these can be made.

Example 3. (Dichotomy data). A company has 14,387 retail outlets of which (unknown to them) 4,669 would be flagged for closure if they were audited. The company plans to audit a without-replacement sample of 200 outlets.

a. What fraction p of company outlets would be flagged if all 14387 were audited? **ans. $p = \frac{4669}{14387} = 0.3245$.**

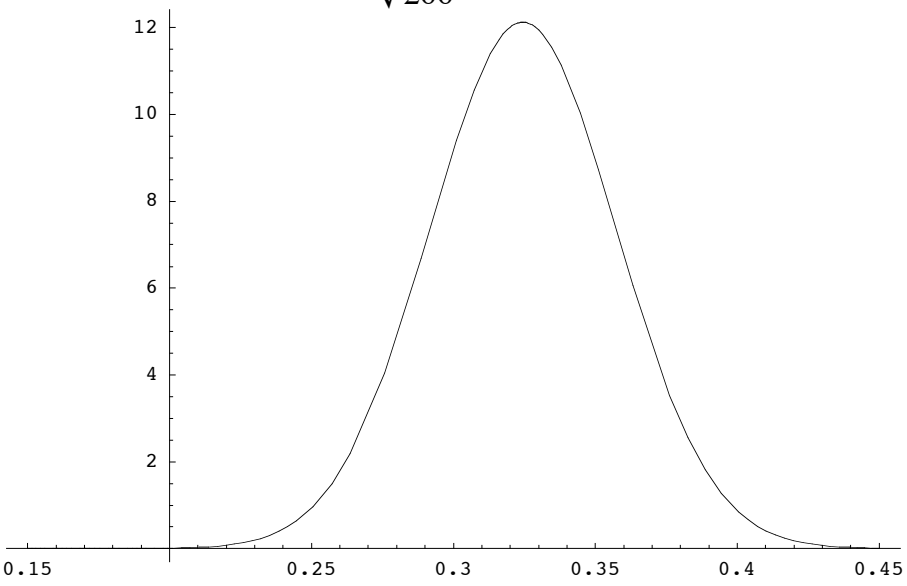
b. For what score x is p the population mean μ ? **ans. The score x defined by x = 1 if "would be flagged" and x = 0 if "would not be flagged." Such a 0-1 score x is given the name "indicator variable" because it indicates one of two cases. Clearly $\mu = \frac{0\,(14387-4669) + 1\,(4669)}{14387} = 0.3245 = p$ as requested in the question.**

What does indicator x do for us? It identifies proportions as averages of indicators and hence under the control of the CLT.

c. What is the population s.d. σ for the score x? **ans. It is just the s.d. of a Bernoulli random variable X with p = 0.3245 which is σ = $\sqrt{0.3245\,(1 - 0.3245)}$ = 0.4682.**

d. Sketch the approximate distribution of the sample proportion $\hat{p}$ (*i.e.* the fraction flagged in the sample) if the company goes ahead with its plan to audit 200 outlets chosen without-replacement. ans. $\hat{p}$ is just the sample mean $\overline{\overline{X}}$ for the indicator x of "flagged." This sample mean is from a population with mean 0.3245 and population s.d. 0.4682. By the CLT the approximate distribution of the sample mean is a normal with mean 0.3245 and s.d.

$$\sqrt{\frac{14387-200}{14387-1}}\ \frac{0.4682}{\sqrt{200}} = 0.0329.$$



e. Approximate the probability that the company estimate $\hat{p}$ will exceed the true p = 0.3245 by a half percentage point or more. ans.

$$P(\hat{p} > 0.3245 + 0.005) \approx P(Z > \frac{0.3295 - 0.3245}{0.0329})$$
$$= P(Z > \boxed{0.1}\boxed{5}) = 0.5 - P(0 < Z < 0.5\boxed{0.1}\boxed{5})$$
$$= 0.5 - 0.0596 = 0.5962 = 0.44.$$

So there is a 44% chance.

z	0.0 5
0.1	0.0596

The sampling distribution of $\overline{\overline{X}}$ when the population itself has a normal distribution. In this case, due to the property that sums of independent normal r.v. are also normal,

$$P(\frac{\overline{\overline{X}} - \mu}{\frac{\sigma}{\sqrt{n}}} < z) = P(Z < z) \text{ for every } z.$$

So there is no requirement that the sample size n be large in order for the CLT to apply. Therefore the calculations of example 1 would all be exact, not approximations, if the distribution of mpg over the population was itself normal.

Sampling distribution of "Studentized" ratio $\frac{\overline{\overline{X}} - \mu}{\frac{s}{\sqrt{n}}}$. This useful expression has replaced the population s.d. σ by its random sample estimate s in the denominator. So the quotient now has randomness in both the numerator and the denominator.

What are the uses of the Studentized ratio? The Studentized ratio is used to form what is known as a **confidence interval for μ** that is valid for every sample size n > 2, provided the population distirbution is normal (in control). A confidence interval is an interval fabricated from sample data that will contain the unknown population mean μ with a specified probability. If this interval is wide we feel the information provided by the sample is not particularly strong. If it is narrow we feel the information is strong.

Case of a normal population. As shown by "Student" for samples

from a normal population

$$P\left(\frac{\overline{\overline{X}} - \mu}{\frac{s}{\sqrt{n}}} < z\right) = P(T < z) \text{ for every t and every } n > 1,$$

where T denotes a random variable having the Student's T distribution with degrees of freedom n-1. The distribution of T does not depend upon μ or σ and is tabulated for each n > 1. Here is what the T table looks like.

Degrees of freedom	$t_{0.025}$
1	12.706
∞	1.960

That is, P(12.706 < T < 12.706) = 0.95 and P(T > 12.706) = 0.025 for T with one degree of freedom (n = 2). So if we have a sample of n = 2 from a normal population having mean σ and s.d. σ, both unknown, we are able to say that

$$P\left(-12.706 < \frac{\overline{\overline{X}} - \mu}{\frac{s}{\sqrt{n}}} < 12.706\right) = 0.95.$$

Example 4. Use the T table to give a 95% confidence interval for μ from a sample of n = 2 from a normal population. **ans.**

0.95 = P(-12.706 < T < 12.706)

$= P\left(\overline{\overline{X}} - 12.706 \frac{s}{\sqrt{2}} < \mu < \overline{\overline{X}} - 12.706 \frac{s}{\sqrt{2}}\right)$

no matter what μ or σ may be. So the interval $\left(\overline{\overline{X}} - 12.706 \frac{s}{\sqrt{2}}, \overline{\overline{X}} - 12.706 \frac{s}{\sqrt{2}}\right)$ is a 95% confidence interval for μ based on a sample of only n = 2. That is, it contains the unknown population mean μ with specified probability 0.95.

Interpretation of the 96% confidence interval is a little tricky. Suppose the data is {3.44, 3.18}. Then $\overline{\overline{X}} = \frac{3.44 + 3.18}{2} = 3.31$ and s =

0.184. So the 95% confidence interval for μ is

$$\bar{\bar{X}} \pm 12.706 \frac{s}{\sqrt{2}}$$
$$= 3.31 \pm 12.706 \frac{0.184}{\sqrt{2}}$$
$$= 3.31 \pm 1.65315.$$

We cannot say that P(μ belongs to 3.31 ± 1.65315) = 0.95 since there is nothing random in the probability statement! What we can say is that we followed the 95% confidence interval method. Our interval should contain μ in 95% of cases where it is correctly used. As for this case, we simply don't know if we've hit μ or missed it. But we followed the 95% prescription. In 95% of cases we would be correct in assuming that μ is in the 95% confidence interval.

"Student's" important insight heralds a major benefit of being "in control" since it means that business activity can be monitored even with samples as small as n = 2 provided the population distribution is normal. There is no need to insist upon large n just to give concise meaning to the sample results if the population distribution is normal.

Can we construct confidence intervals for μ without assuming that the population distribution is normal? Yes, but the confidence interval will only be approximately valid and requires large sample size

CLT for Studentized ratio $\frac{\bar{\bar{X}} - \mu}{\frac{s}{\sqrt{n}}}$. As alluded to above, for large n, when sampling with replacement, the "studentized" ratio is approxmately normally distributed irrespective of the population provided the population variance is finite.

$$P(\frac{\bar{\bar{X}} - \mu}{\frac{s}{\sqrt{n}}} < z) \to P(Z < z) \text{ for every z as n} \to \infty.$$

For samples without-replacement

$$P\left(\frac{\overline{\overline{X}} - \mu}{\sqrt{\frac{N-n}{N-1}}\ \frac{s}{\sqrt{n}}} < z\right) \approx P(Z < z) \text{ for every z.}$$

This requires both n and N-n to be large plus additional assumptions not specified here.

Example 5. A sample of n = 50 is selected from a population that is not necessarily normal and whose mean μ and s.d. σ are not known. Give a 95% confidence interval for μ. **ans.**

$$\mathbf{P\left(\ \mu \text{ belongs to the interval } \left(\overline{\overline{X}} - 1.96\ \frac{s}{\sqrt{50}}, \overline{\overline{X}} + 1.96\ \frac{s}{\sqrt{50}}\right)\right)}$$

$$\mathbf{\approx P(-1.96 < Z < 1.96) = 0.95.}$$