Lecture outline for 3 - 15/17 - 10.

This material pertains to chapters 23/24 assigned for MW in the syllabus.

Chapters 23/24 reinforce confidence intervals and tests for a population proportion p, just covered on exam 2, extending those ideas to confidence intervals and tests for the population mean μ.

	for p	for μ	assumptions
z-based CI	$\hat{p} \pm z \sqrt{\frac{\hat{p}\,\hat{q}}{n}}$	$\overline{x} \pm z \frac{s}{\sqrt{n}}$	"large n"
t-based	none	$\overline{x} \pm t_{n-1} \frac{s}{\sqrt{n}}$	for all n > 1 if the population is normal

where s denotes the *sample standard deviation* $s = \sqrt{\frac{\sum_{i=1}^{i=n} (x_i - \overline{x})^2}{n-1}}$.

The respective claims are:

P(p in $\hat{p} \pm z \sqrt{\frac{\hat{p}\,\hat{q}}{n}}$) → P(Z in (-z, z)) as n → ∞ in Bernoulli trials with 0 < p < 1.

P(μ in $\overline{x} \pm z \frac{s}{\sqrt{n}}$) → P(Z in (-z, z)) as n → ∞ in independent samples if $\sigma < \infty$.

P(μ in $\overline{x} \pm t_{n-1} \frac{s}{\sqrt{n}}$) ≡ P(T_{n-1} in (-z, z)) for every n > 1 in independent normal samples.

In all cases we assume with-replacement samples from a fixed population.

See pp. 64/65 for the *sample standard deviation* s which is typically used to estimate the population standard deviation σ.

For data taking only the values 1 (success) or 0 (failure) the population mean μ is the same as the probability of success p. For such data

$$\frac{s}{\sqrt{n}} = \sqrt{\frac{\hat{p}\,\hat{q}}{n-1}}$$ (this is NOT used in the p-case!).

So the z-based CI for p is **almost** a perfect special case of the z-based CI for μ. In fact for large n the difference will be negligible.

Why is n-1 in the denominator above? When estimating a population standard deviation σ by a sample standard deviation s notice that s uses the squared deviations $(x_i - \overline{x})^2$ from sample average (summed over the sample) whereas the calculation of σ uses squared deviations $(x_i - \mu)^2$ from population mean μ (summed over the population). As it happens, using the sample mean always results in a downward bias since the sample mean is "closer to the sample data" than the actual population mean is. Substituting n-1 in the denominator defining s was once thought to be a good compensation for this perceived bias. Evolving statistical tables accommodated the practice and it is now "fixed in stone."

The corresponding test statistics are:

$$\frac{\hat{p} - p_0}{\sqrt{\frac{\hat{p}\,\hat{q}}{n}}} \sim$$ Z distributed if n is large and p_0 is the actual population fraction.

$$\frac{\overline{x} - \mu_0}{\frac{s}{\sqrt{n}}} \sim$$ Z distributed if n is large and the actual population mean is μ_0 and $\sigma < \infty$.

$$\frac{\overline{x} - \mu_0}{\frac{s}{\sqrt{n}}}$$ is **exactly** T_{n-1} distributed if n > 1, the actual population mean is μ_0, and $\sigma < \infty$.

Recitation assignment due 3-16-10.

1. Our class may be considered to be a random sample with-replacement from the msu student body as regards score x = last digit of student number. It is not that the class is a random sample of students for surely that is not the case, rather the last digits are close to being assigned as independent samples from the digits 0 through 9 with equal probability for each. It is unusual in statistical work that we know the population mean but according to our view of last digits we are

confident that μ = (1/10)(0+1+2+ ... + 8+9) = 4.5. The population standard deviation σ is also known to us and is **the square root o**f (1/10)((0-4.5)^2+ ... + (9-4.5)^2) ~ 2.87... .

a. Verify the values given for μ and (to five decimals or more) σ.

b. The standard error for the sample mean $\overline{x}$ is known to be $\frac{\sigma}{\sqrt{n}}$. Since we know σ we also know this standard error. Calculate it for a (with replacement) sample of n = 25.

c. Here is a sample of n = 25 actua**ll**y taken by equal probability with replacement samples from the population {0,1,2,3,4,5,6,7,8,9}. **Make note of the fact that it would make no difference if the population had one million of each of the ten digits since we sample with replacement**.

3, 8, 9, 9, 3, 0, 3, 6, 6, 5, 0, 5, 9, 7, 3, 9, 9, 8, 5, 0, 8, 2, 2, 7, 7

d. From the sample (c) calculate the values of the sample mean $\overline{x}$ and the sample standard deviation s.

e. Compare the sample mean with the actual population mean 4.5. Likewise compare the sample standard deviation s with the actual population standard deviation s. The Job description (so to speak) of the sample mean is to find the population mean. Likewise the sample standard deviation is tasked with finding the actual population standard deviation. Remember, the population could just as well consist of ten million or as many as we like. In view of this it is remarkable if the sample mean comes anywhere near μ and the sample standard deviation comes anywhere near σ! How well did we do?

For the following we have to pretend that we don't know the population, its mean or its standard deviation.

f. The population could be estimated by a histogram (or just a bar graph) showing the frequencies at the distinct values 0 through 9 found in the sample of n = 25 x-scores. Give such a bar graph or histogram.

g. Using the formula $\overline{x} \pm z \frac{s}{\sqrt{n}}$ evaluate it to produce a 95% confidence interval for (unknown) μ based on the above sample of 25.

h. Does your CI (g) cover the actual population mean μ? Ordinarily we would not know the answer since μ is after all not known (the whole point of the sample is to estimate it).

Out of 100,000 samples of n = 25 (independently gathered) around how many samples of 25 would produce a 95% CI failing to cover μ?

Is such a CI exact for 95% or approximately a 95% CI?

Have we said how large n should be for the CI to work as described?

i. To test the hypothesis that μ is actually 4.5 (I've tried to convince you of this, relating my conversations with the registrar's office, etc.) we take $\mu_0 = 4.5$ and evaluate the test statistic

$$\frac{\overline{x} - \mu_0}{\frac{s}{\sqrt{n}}} =$$

which measures **how different $\overline{x}$ is from 4.5** (in terms of the estimated standard deviation $\frac{s}{\sqrt{n}}$ of $\overline{x}$).

j. Just as you did for the P-value of a test of a hypothesis about p (in previous chapters) now do exactly the same with your statistic (i). Use it to determine the P-value of a z-test of

$H_0: \mu = 4.5$ versus $H_A: \mu > 4.5$.

Is this test one-sided or two-sided?

What entry do you make to the z-table?

What is the P-value for this data? Does it seem small enough to lead us to seriously question the null hypothesis I have urged on you?

Sampling from a NORMAL population. If a population is NORMAL the mean μ and standard deviation σ are all that are needed to specify that population completely.

Importantly, our estimates of them, $\overline{x}$ and s, are *statistically independent of one another*. This has deep consequences. Moreover

$$\frac{\overline{x} - \mu}{\frac{s}{\sqrt{n}}}$$

(at the root of CI) actually is free of μ and σ (although they are in the data x, the way the above is constructed nicely cancels them out!).

So when sampling any normal population the distribution of the above quotient does not depend upon the **particular** (normal) population! That distribution does depend upon the sample size n which must be more than 1 for s to be defined. It is tabulated as the "t-distribution" and has many uses in statistics, each with its own "degrees of freedom" calculation linking the particular application to the particular n needed. So instead of n in the t-table we have something called "degrees of freedom."

In our use of the t -CI for μ the "degrees of freedom" is n-1.

2. The following data was obtained as a random sample from a NORMAL population.

76.6464, 104.43, 67.0315, 89.8393, 92.8544, 106.225, 121.926

a. Determine: sample mean

sample standard deviation

b. Determine a 95% CI for the population mean μ. The degrees of freedom are n-1 and if you look at the bottom of the t-table you will find1.96 (the z-score you would use for a z-CI, but not used here for small n). The 1.96 at the table bottom is for "infinite" degrees of freedom (think z) but you need the entry for 95% CI for degrees of freedom n-1. The 95% CI is then

$$\overline{x} \pm t_{n-1} \frac{s}{\sqrt{n}}$$

c. Determine the test statistic for a t-test (not a z-test) of the null hypothesis that μ is 100 versus the alternative that it is less than 100.

d. The t-table is not very complete but you can find a range for the P-value (look for your test statistic in row n-1).