Stat 202 - 2015S - W 11 ~~Tuesday~~ Wednesday

~~Recitation~~

Roadmap

Chapter 5 - Sampling Distributions

§5.1 - means
§5.2 - proportions

Chapter 6

Introduction - yesterday
§6.1 Confidence Intervals - skip for now
§6.2 Tests of Significance - better continuation from yesterday
§6.3] Further issues with inference
§6.4]

Chapter 6 involves testing <u>means</u> when you know already what the population standard deviation is.

Chapter 7 involves testing <u>means</u> when you don't know already what the population standard deviation is (more realistic but harder),

Chapter 8 concerns testing proportions
That's all we'll do.

Review and new material / mixed
(Last class was a preview, topic the same)

Sallie Mae compiled data about student debt from applicants of student loans.

Students from Midwest had average debt of $3260. Students from West had average debt of $3817 — thats $557 more! Can we conclude from the data that average debt in these two regions are different. Careful: these numbers are based on a sample draw a different sample and get a different number

Approch: compute probability of obtaining a difference as large or larger than $557 assuming that there are no difference in means.

How would you do that? You would need more information than given in this example. (specifically you need sample sizes)
Lets say the probability is 0.14
then there is a reasonable chance that the data seen could be produced even if means were the same. We have weak evidene that the means are different

Let's say the probability is 0.000000001

Then is probability that the data seen could be produced even if means were the same is negligible.

We have strong evidence that means are not the same,

Stating Hypotheses

Above we asked whether the difference in the observed sample means, is reasonable

the data

assuming that in fact, there is no difference in the population means.

"There is no difference in the population means" is the null hypothesis — a claim that we try to find evidence against.

The statement being tested in a test of significance is called the null hypothesis. The test of significance is designed to assess the strength of the evidence against the null hypothesis. — Usually a statement of "no effect" or "no difference"

Null hypothesis

abbreviated → H_0: there is no difference in the true means

Alternative hypothesis

H_a: there is a difference in true means

Hypotheses always refer to some populations or a model, not to a particular outcome. H_0 and H_a are always stated in terms of population parameters — never sample statistics.

H_a usually expresses the effect we hope to find evidence for.

H_0 is usually the statement that the hoped for effect is not present.

H_a can be one sided of two sided

One sided: $\left\{ \begin{matrix} \mu_1 > \mu_2 \\ \mu_2 > \mu_1 \end{matrix} \right\}$ or written $\mu_1 - \mu_2 > 0$ / $\mu_1 - \mu_2 < 0$

two sided: $\mu_1 \neq \mu_2$ $\quad$ $\mu_1 - \mu_2 \neq 0$

If H_a is $\mu_1 > \mu_2$ we ~~reject~~ ignore data as $\mu_1 < \mu_2$ as insignificant.

The alternative hypothesis should express the hopes or suspicions we bring to the data.

It is cheating to first look at the data and then frame Ha to fit what the data show.

Specifically if you have no reason to know $\mu_1 > \mu_2$ you <u>cannot</u> wait until the data come in then see $\mu_1 > \mu_2$ then frame the alternative hypothesis accordingly

<u>It</u> would be tempting to do that because the p-value would be lower (½ actually). But this is considered an abuse of statistics.

Test statistic

I don't know how to derive the clusterization index from the tree problem, yesterday,

But I do know the test statistic for testing the means

For testing the mean assuming we know already the population standard deviation

Null hypothesis

$H_0: \mu = \mu_0$

↑ population mean

↑ some specified value that we are testing

Here is an example

A large study showed that the mean cholesterol level among college women is 168 mg/dl with standard deviation 27 mg/dl

A more recent study examined sedantary female college students ($n = 71$ females) $\bar{x} = 173.7$

Is there evidence that sedantary females have higher cholesterol?

Null hypothesis

$\mu_0 = 168$

$H_0: \mu = 168$

$H_a: \mu \neq 168$

We have to make an unrealistic assumption because we haven't reached chapter 7,

Assume $\sigma = 27$

We are trying to show H_0 is false we do this by showing that if H_0 is true our data are very unlikely,

If the mean is 168 and standard deviation is 27, what is the probability of seeing $\bar{X} = 173.7$ with $n = 71$.

Test statistic is Z-score

$$Z = \frac{\bar{X} - \text{"mean"}}{\text{"standard deviation"}}$$

"mean" and "standard deviation" refers to the statistics of $\bar{X}$ under Ho

"mean" is 168

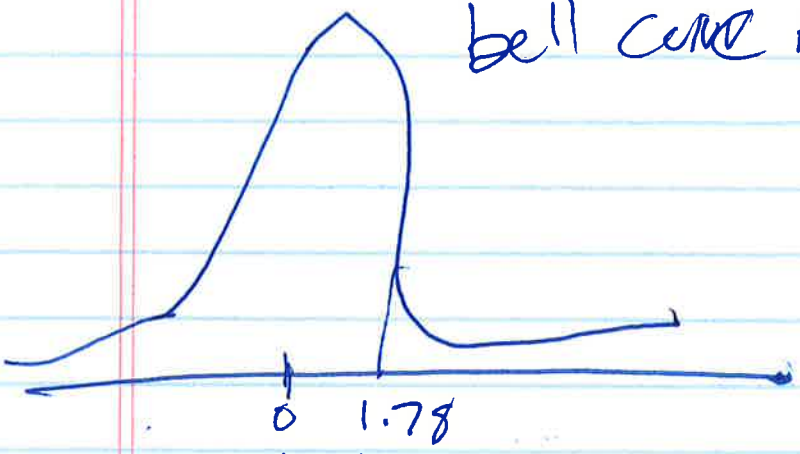
"standard deviation" is $27/\sqrt{71}$

population $\rightarrow \sigma/\sqrt{n}$

$Z \sim N(0,1)$ Standardized sample means are standard normal.

From there you can compute a p value

bell curve for N(0,1)



the test statistic works out to

1.78

The pvalue for a one sided alternative is the area under the bell curve above 1.78. For a two sided alternative it is the area under the bell curve above 1.78 and below −1.78

twice the area twice the pvalue

For a two sided alternative p values double — harder to show evidence against null hypothesis,

In this case we used a two sided alternative and the pvalue was 0.075 — not significant at the traditional level.

Statistical Significance

If the P value is a small or smaller than a number α then we say that the data are **statistically significant at level α**

The traditional level is 0.05
But statisticians are less stringent about the traditional level than they used to

In other words it is now considered more useful to report the p-value than it is to say it's significant (at the 0.05 level) without reporting p-value.