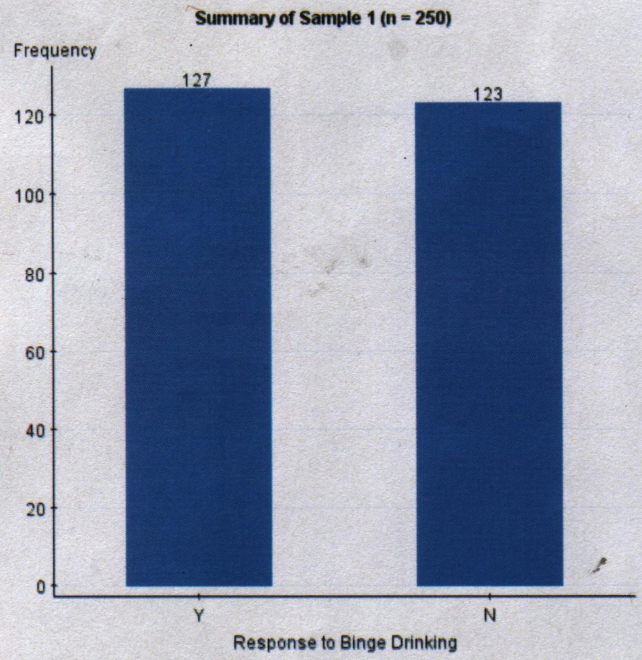
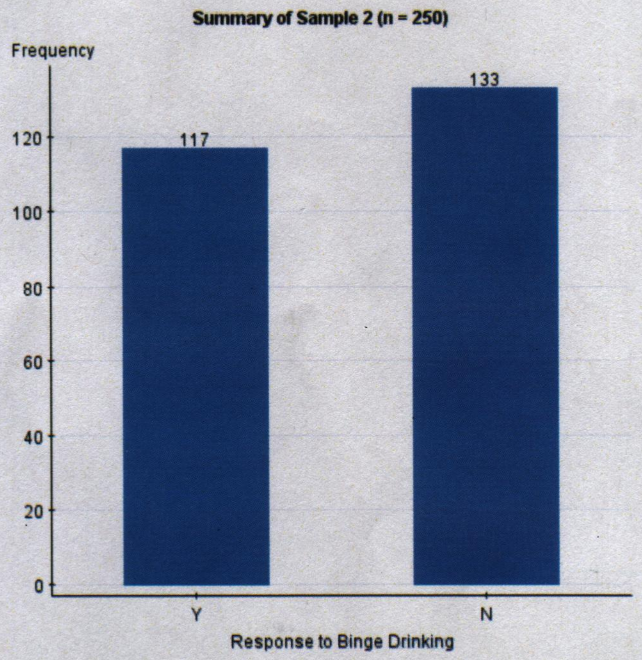
Packet 3: Sampling Distribution of the Sample Proportion Textbook pages: 399 – 406

After completing this material, you should be able to:

- explain what the symbols $\hat{p}$, p, $\mu_{\hat{p}}$ and $\sigma_{\hat{p}}$ represent.
- describe the sampling distribution of the sample proportion by discussing its shape, mean, and standard deviation.
- find probabilities associated with various sample proportions based on the sampling distribution.
- ~~make inferences from the probability and explain the reasoning.~~

A university is concerned with the percentage of its students who binge drink. Two different campus offices take samples in order to investigate the severity of the problem. Students in each sample were asked whether or not they had engaged in binge drinking (5 drinks at a sitting for men, 4 for women) in the past month. Results from the two surveys are summarized in the relative frequency bar graphs below:





For each of the two samples, determine the sample proportion who responded that they had engaged in binge drinking over the past month.

Sample 1:

$\hat{p} = \frac{\text{part}}{\text{whole}} = \frac{127}{250} = .508$

Sample 2:

$\hat{p} = \frac{117}{250} = .468$

Notation Alert

(You *must* remember this notation!!)

p - population proportion (parameter)

$\hat{p}$ - sample proportion (statistic)

p-hat

Why are the two sample proportions different?

Different samples!
We expect variation in different samples.
We expect different $\hat{p}$ values - but how much variation?

What exactly is a **sampling distribution** and why is it important?

The sampling distribution describes how a statistic ($\hat{p}$) varies from sample to sample.

We're going to use the sampling distribution of $\hat{p}$ to inform us about p, so we want to know how much variation to expect.

Example: According to a 2005 newspaper report about financial aid for college students, $\boxed{75\%}$ of all full-time degree-seeking students receive some form of financial aid. Given the recent financial crisis, an economist conjectures that the percentage receiving some form of financial aid has increased. In order to test his conjecture, he plans to sample 265 full-time degree-seeking students to determine the proportion that receive financial aid.

Two numbers are given in the example. Assign the appropriate notation (based on the previous page) to these values.

$75\% \to .75 = p$ proportion who receive financial aid in the population.

$n = 265$ is the sample size for $\hat{p}$. p is a number between 0 & 1

What is the conjecture that the economist is trying to find support for?

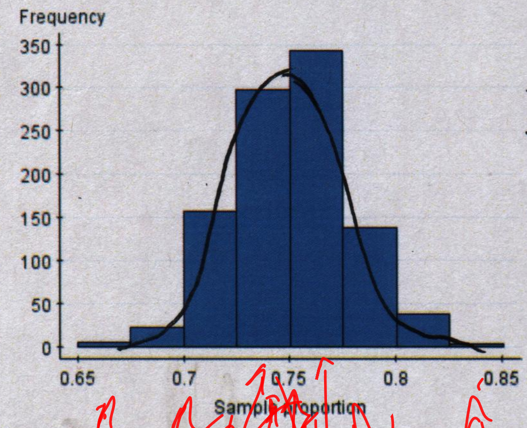
The proportion in the population (p) has actually gone up from .75. ← fulltime, degree seeking college students.

Suppose he takes a sample of 265 college students. Which would give more support to his conjecture – finding that 204 students received financial aid or that 215 students received financial aid? Explain your choice.

$204 \to \frac{204}{265} = \hat{p} = .7698 > .75$

$215 \to \frac{215}{265} = \hat{p} = .8113 \gg .75$ → So this one lends more support, because it's farther away.

Instead of taking a single sample of 265 students, suppose that 1000 different samples of 265 students were taken. A sample proportion from each sample was computed and summarized in the graph below.



What seems familiar about the shape of this graph?

It looks like a bell-shaped distribution – in fact it should look normal.

When asked to **describe** the sampling distribution of the sample proportion, the following three characteristics must be addressed:

- shape: $\hat{p}$ is normally distributed if $np > 10$ and $n(1-p) > 10$.
- mean: $\mu_{\hat{p}} = p$
- std. dev: $\sigma_{\hat{p}} = \sqrt{\frac{p(1-p)}{n}}$ n is the sample size.

Why are we using the notation $\mu_{\hat{p}}$ and $\sigma_{\hat{p}}$ instead of just using μ and σ??

$\mu_{\hat{p}}$ pertains to the variation in $\hat{p}$; as does $\sigma_{\hat{p}}$ (the expected spread) center of the

Back to the example: Use these characteristics to describe the sampling distribution of the sample proportion of full-time degree-seeking students who receive some form of financial aid.

Given: $n = 265$ $p = .75$

(1) Normality? $n \cdot p = 265 \cdot .75 = 198.75 > 10$ ✓
$n \cdot (1-p) = 265 \cdot .25 = 66.25 > 10$ ✓

Assume

(2) $\mu_{\hat{p}} = p = .75$, the mean

(3) $\sigma_{\hat{p}} = \sqrt{\frac{.75(1-.75)}{265}} = .0266$

σ_p

The fact that the sampling distribution of the sample proportion is *normally distributed* is important – we know how to find probabilities from the normal distribution. To do this, though, we will need to modify our formula for the z-score to reflect that we are now dealing with sample proportions:

Formula Alert!!
This formula will be given on the formula sheet.

$$z = \frac{\text{obs} - \text{mean}}{\text{std dev.}} = \frac{\hat{p} - \mu_{\hat{p}}}{\sigma_{\hat{p}}} = \boxed{\frac{\hat{p} - p}{\sqrt{\frac{p(1-p)}{n}}}}$$

Let's go back to the example and see how this formula is used ...

$\hat{p} = .7698$

$p = .75$

Assuming the newspaper's claim is correct, find the **probability** of observing 204 or more students on financial aid and the **probability** of finding 215 or more students on financial aid. Of the two probabilities, which one gives more support to the conjecture that the proportion of students on financial aid has increased? Explain.

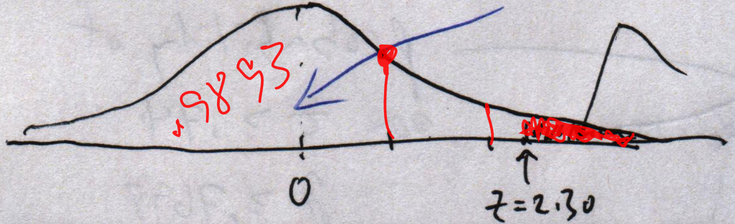
215 of 265 on financial aid: $\hat{p} = \frac{215}{265} = .8113$

$p = .75 \quad n = 265$

$$z = \frac{.8113 - .75}{.0266} = 2.30$$

area from z-table – z = 2.30 = .9893

small enough that we infer that p has gone up.



$1 - .9893 = .0107$, probability of a z value as large or larger than $z = 2.30$.

Using a probability to make a decision (rule of thumb):

If it's unlikely that we get a result as extreme (or more so) than we get (prob < .10) then we'll declare that p is not what was claimed. Otherwise, (prob > .10) we won't declare p incorrect.

Example 2: According to an article on webmd.com, 28.6% of Kentucky residents smoked in 2000. After significant advertising campaigns by the American Cancer Society, a researcher would like to know if the proportion of smokers has decreased. A random sample of 672 Kentucky residents is taken, and each is asked whether or not they smoke.

What is the conjecture that we would like to find evidence to support?

That there are fewer Kentucky residents smoking $(p < .286)$ now than before.

Completely **describe the sampling distribution** for the sample proportion of Kentucky residents that smoke when samples of size 672 are taken.

(1) Shape: is $\hat{p}$ normally distributed? $n = 672$, $p = .286$

$np = 672 \cdot .286 = 192 > 10$

$n(1-p) = 672(1-.286) = 479 > 10$

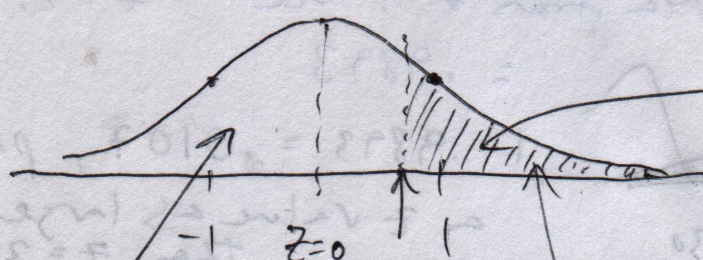
$\hat{p}$ is normal.

(2) mean: $\mu_{\hat{p}} = p = .286$

(3) std. dev.: $\sigma_{\hat{p}} = \sqrt{\frac{p(1-p)}{n}} = .0174$

$\hat{p} = .7698$ $\qquad z = \dfrac{\hat{p} - p}{\sigma_{\hat{p}}} = \dfrac{.7698 - .75}{0.0266}$

$= 0.74$



probability of a $z \geq .74$

$\hat{p} \geq .7698$

observed ≥ 204

$z = .74$

z-table:

area $= .7704$ $\qquad 1 - .7704 = \boxed{.2296}$

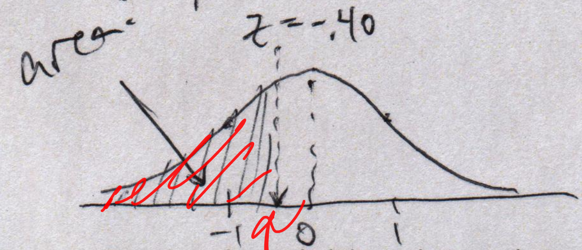
Not too odd.

I wouldn't conclude that financial aid use is up from 75%.

Suppose that a sample of 672 residents is taken, and 27.9% smoke. Assuming the advertising campaigns have had no effect, what is the probability of observing a smoking percentage of 27.9% or less in this sample?

$\hat{p} = .279$

$$z = \frac{\hat{p} - p}{\sigma_{\hat{p}}} = \frac{.279 - .286}{.0174} = -.40$$



area from the z-table = .3446 > .10

Based on the probability, what conclusion would be reached concerning the conjecture which was made?

Since the probability of a z-score that negative (or more so) is > .10, we don't have enough evidence to claim p has gone down.

Example 3: According to census estimates in 2010, approximately $\boxed{72\%}$ of all US citizens were registered to vote. In the recent presidential election, a push was made to encourage those who were not registered to vote to do so. To see if this has had an effect on voter registration, a random sample of $\boxed{300}$ American adults is taken, and the proportion who are registered to vote is recorded.

$n = 300$ $p = .72$

What variable was recorded? Is this a categorical or quantitative variable?

Voting registration status (Y/N)

Categorical.

What is the conjecture that we would like to find evidence for?

The rate of voter registration has gone up from $p = .72$.

Completely **describe the sampling distribution** for the sample proportion of potential voters that are registered to vote if samples of size 300 are taken.

(1) shape: is $\hat{p}$ normally distributed?

$n \cdot p = 300 \cdot .72 = 216 > 10$

$n \cdot (1-p) = 300 \cdot (1 - .72) = 300 \cdot .28 = 84 > 10$

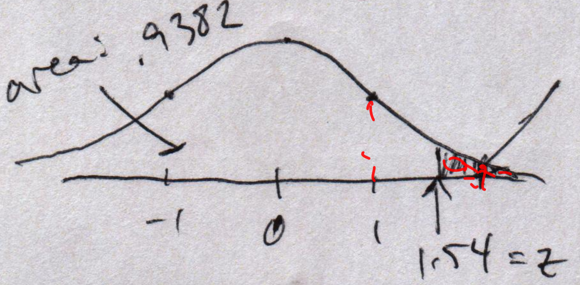
assume $\hat{p}$ normal

(2) mean: $\mu_{\hat{p}} = p = .72$

(3) std. dev.: $\sigma_{\hat{p}} = \sqrt{\frac{p(1-p)}{n}} = .0259$

Suppose that a sample of 300 potential voters is taken, and 228 registered voters are found. Assuming the claim is true, what is the probability of obtaining a sample where more than 228 individuals are registered voters?

$\hat{p} = \frac{228}{300} = .76$ $\qquad z = \frac{\hat{p} - p}{\sigma_{\hat{p}}} = \frac{.76 - .72}{.0259} = 1.54$



$1 - .9382 = \boxed{.0618} < .10$

Based on the probability, what conclusion would be reached about the conjecture which was made?

That's pretty odd, pretty rare: it's suspiciously high (at the .10 level). Since the probability is less than .10 of a value that high or higher, we going to conclude that p has increased: voter registration is up.

Therefore our campaign (or others) have successfully increased voter reg.