AP Statistics – Lesson Notes - Chapter 21: Type I, II Errors; Power of a Test

Does our analysis always come to the correct conclusion?

Example: The DV counseling department records show that, in the past, 78% of DV seniors attend college in-state. We ask an SRS of 50 seniors this year and 42 of them say they are staying in-state. *Do we have reason to believe that more seniors are staying in-state this year?*

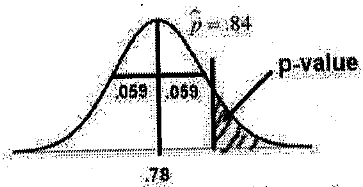
H_0: *Percentage of seniors in–state is 78%* $(p = .78)$

H_A: *Percentage of seniors in–state is greater than 78%* $(p > .78)$

$\hat{p} = \frac{42}{50} = .84 \quad \mu_{\hat{p}} = p = .78$

$\sigma_{\hat{p}} = \sqrt{\frac{pq}{n}} = \sqrt{\frac{(.78)(.22)}{50}} = .059$

$z = \frac{.84 - .78}{.059} = 1.02$



p-value $= P(z > 1.02 \mid H_0 \ true)$

$= normalcdf(.84, 999, .78, .059)$

$= .15$

With p-value=.15, we don't reject H_0. Percentage in-state is still 78%.

What happens if sample size increases?

Example: The DV counseling department records show that, in the past, 78% of DV seniors attend college in-state. We ask an SRS of 200 seniors this year and 168 of them say they are staying in-state. *Do we have reason to believe that more seniors are staying in-state this year?*

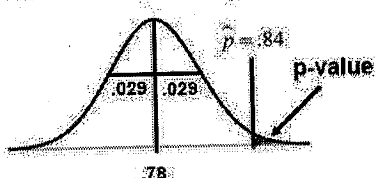
H_0: *Percentage of seniors in–state is 78%* $(p = .78)$

H_A: *Percentage of seniors in–state is greater than 78%* $(p > .78)$

$\hat{p} = \frac{168}{200} = .84 \quad \mu_{\hat{p}} = p = .78$

$\sigma_{\hat{p}} = \sqrt{\frac{pq}{n}} = \sqrt{\frac{(.78)(.22)}{200}} = .029$

$z = \frac{.84 - .78}{.029} = 2.07$



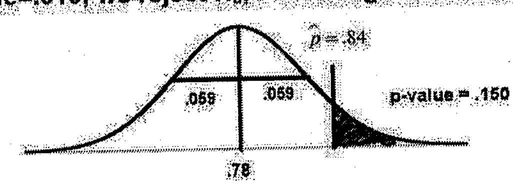
p-value $= P(z > 2.07 \mid H_0 \ true)$

$= normalcdf(.84, 999, .78, .029)$

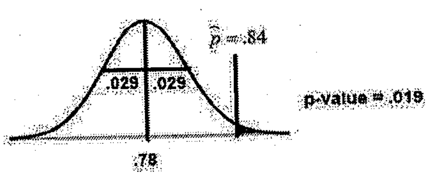
$= .019$

With p-value=.019, we reject H_0. Percentage in-state is greater than 78%.

n = 50



n = 200



There was a difference all along, but this same difference was unusual for n=200, but not unusual for n=50.

The n=50 came to the wrong conclusion.

Types of Error

With the larger sample size, we determined that the percentage of seniors staying in-state did increase. But there was nothing 'wrong' with our first analysis which determined that there was not sufficient statistical evidence to reject the null hypothesis, and concluded that the percentage had not changed.

Although we didn't do anything wrong, the first analysis *failed to reject the null hypothesis, even though the null hypothesis was false.* This called an **Error.**

In fact, there are two types of errors:

__Type I Error__: The Null Hypothesis is actually true (there is 'nothing to detect') but our data happens to be far from Ho so we incorrectly reject Ho.

__Type II Error__: The Null Hypothesis is actually false (there is 'something to detect') but our data happens to be close to Ho so we incorrectly fail to reject Ho.

To help remember...

Null Hypothesis is:

Decision:	True	False
Reject	Type I Error α	OK (power) $1-\beta$
Not Reject	OK	Type II Error β

Which type of error is more serious? It depends upon the scenario...

__For the DV Seniors in-state example:__

H_0: *Percentage of seniors in-state is* 78% $(p = .78)$

H_A: *Percentage of seniors in-state is not* 78% $(p \neq .78)$

Null Hypothesis is:

	True	False
Reject	Type I Error α	OK (power) $1-\beta$
Not Reject	OK	Type II Error β

Type I Error: The true (population) percentage of Seniors in-state was still 78%, but our sample just happened to be far from this so we conclude that the percentage staying in state is different.

Type II Error: The true (population) percentage of Seniors in-state **is actually** different from 78%, but we did not detect this, and mistakenly concluded that the percentage did not change.

Neither of these errors would be particularly terrible.

__For a jury trial:__

H_0: *The defendant is not guilty.*

H_A: *The defendant is guilty.*

Null Hypothesis is:

	True	False
Reject	Type I Error α	OK (power) $1-\beta$
Not Reject	OK	Type II Error β

Type I Error: The defendant is actually innocent, but the evidence convinces us they are guilty and they are wrongly sent to prison.

Type II Error: The defendant is actually guilty, but the evidence convinces us they are innocent so they are wrongly set free.

Most people would likely say that the Type I Error is more serious in this scenario.

Type I Error

Use your phone's web browser access: www.mrfelling.com/sa4

...and enter the following:

Population Proportion: 0.78

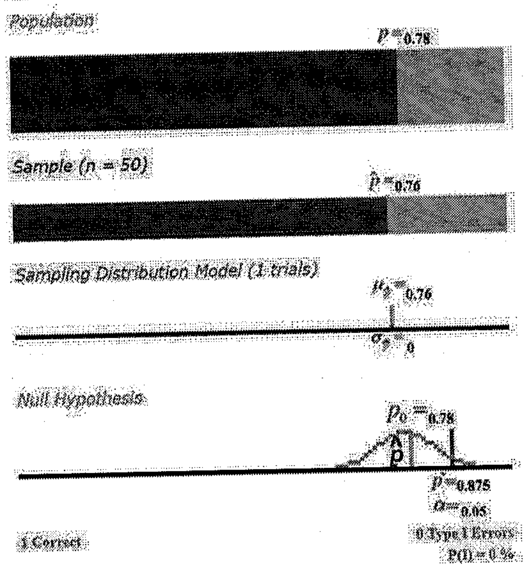
Sample size: 50

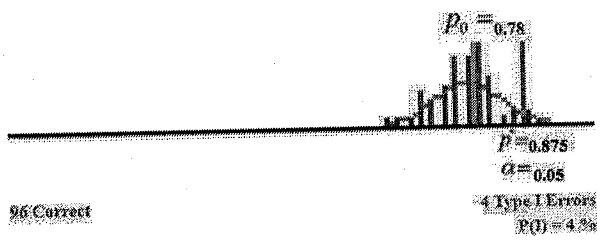
Significance Level (alpha): 0.05

Null Hypothesis is ◉ **TRUE**
Null Hypothesis is ○ **FALSE**
$p_0 =$ ____

This year's % of seniors staying in state is likely not exactly 78%, but probably fairly close to 78% →

This sample is one of the possible $\hat{p}$ values in the Sampling Distribution of Sample Proportions



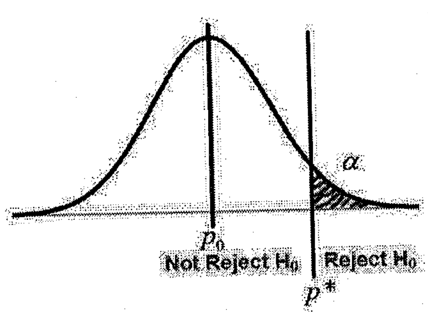


If we took many samples, each would have a proportion which would vary due to natural sampling variation. Sometimes, just due to chance, we would get a proportion which is far enough away from the Ho value that p-value < 0.05 and we would reject Ho even though it is actually true.

This is a **Type I Error** - Ho is actually true, but we happen to have an experiment with an outcome that is unusual just due to chance, so the analysis will come to the wrong conclusion.

Probability of Type I Error

The probability of a Type I error is the chance that our particular sample's proportion falls in the upper 5% if we set $\alpha = .05$:



We reject H_0 if $\hat{p} > p^*$
where p* is some critical value.
This means probability of a Type I Error is alpha:

$$P(\text{Type I error}) = \alpha$$

Type II Error

A Type II Error occurs when the Null Hypothesis is actually false (there is something to detect) but we have an experimental result which is close to Ho so we fail to reject Ho.

Reset, and enter the following:

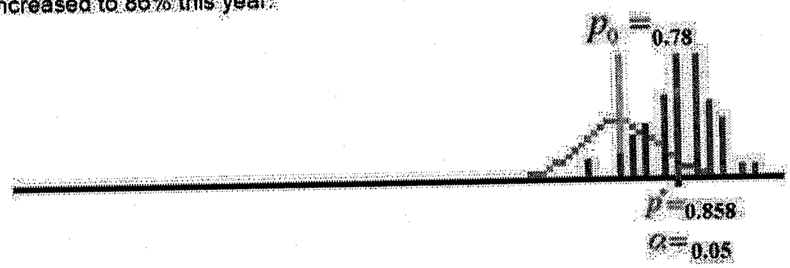
Population Proportion: 0.86

Sample size: 50

Significance Level (alpha): 0.05

Null Hypothesis is ○ **TRUE**
Null Hypothesis is ◉ **FALSE**
$p_0 =$ 0.78

Now we are saying Ho is still 0.78 (we are comparing to the historical 78% of seniors staying in state, so this is the Ho value), but the actual % has actually increased to 86% this year.

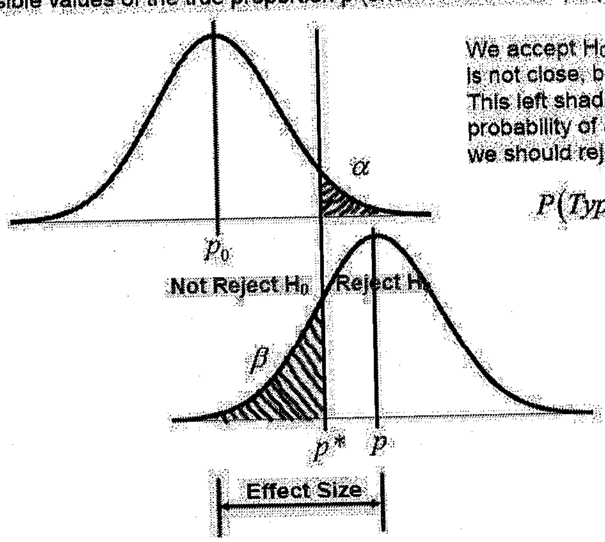


Since Ho is actually false (the % staying in state has increased) notice that the sampling distribution of the samples is centered at the actual population value of 86%, but the normal distribution we are using to find the p-value is still centered at the Ho value of 0.78.

Also, any sample proportion which is far away from Ho is now green because this is the correct conclusion. But proportions which are close to Ho are now incorrect...they happen to be close to Ho so we conclude there is no change, when there actually is, so these are **Type II Errors.**

Probability of Type II Errors

For Type II Errors, H_0 is not true, so the true proportion, p, is actually far from p_0. But we don't know what the true proportion is, only that it is not p_0, so there is a distribution of possible values of the true proportion p (shown is the lower proportion).

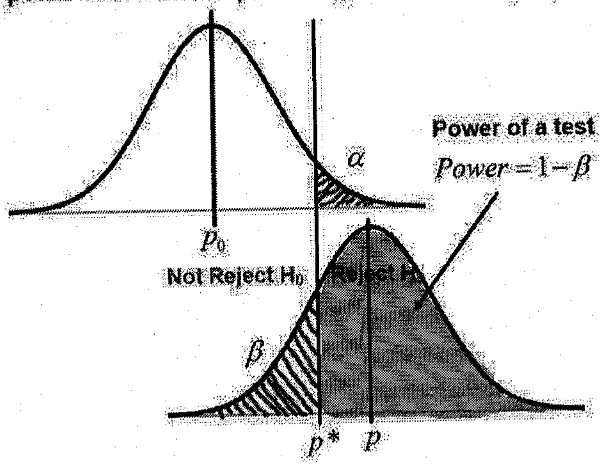


We accept H_0 if $\hat{p} > p^*$ but now this is not close, but far from the true value, p. This left shaded area represents the probability of accepting H_0, even though we should reject. We call this 'beta':

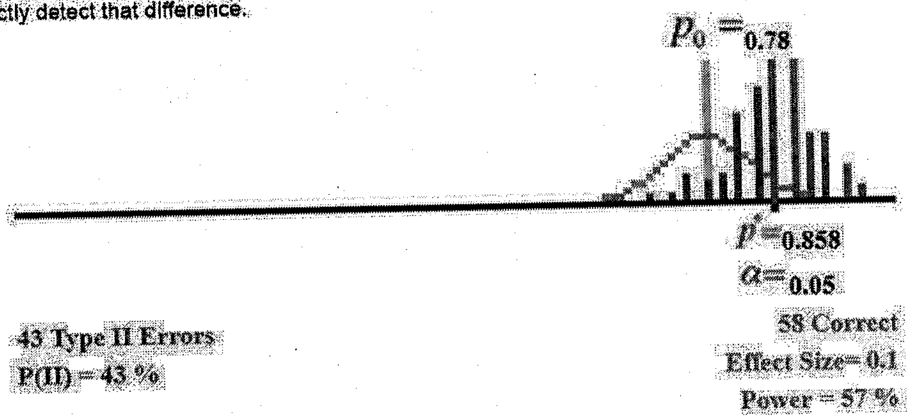
$$P(\text{Type II error}) = \beta$$

Power of a Test

One more important definition...
The power of a test is the probability that it correctly rejects a false null hypothesis.



The power of the test is the probability that, if there is a difference to detect, the analysis will correctly detect that difference.



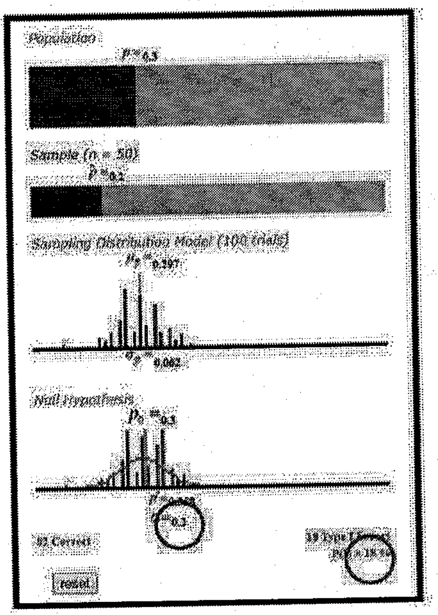
You can use the www.mrfelling.com/sa4 app to play around with scenarios. Try setting the effect size differently (the difference between population proportion and null hypothesis proportion). Try different sample sizes.

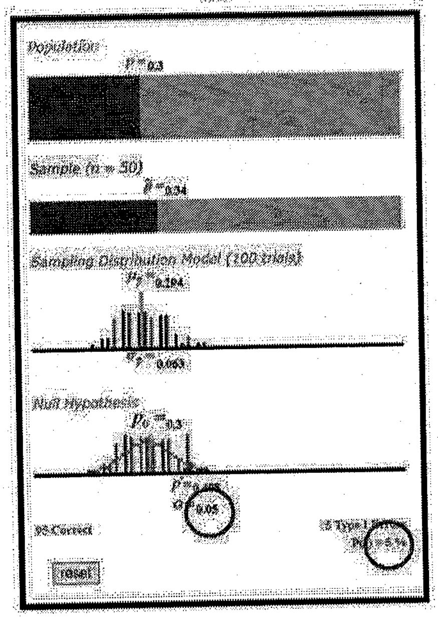
To help remember...

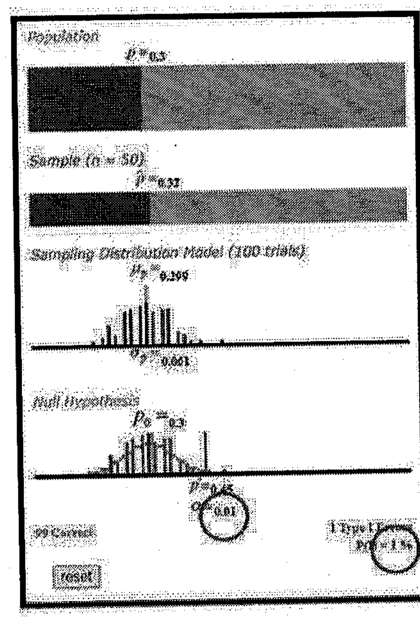
Null Hypothesis is:

Decision:		True	False
	Reject	Type I Error α	OK (power) $1-\beta$
	Not Reject	OK	Type II Error β

To reduce Type I errors... mrfelling.com/sa4

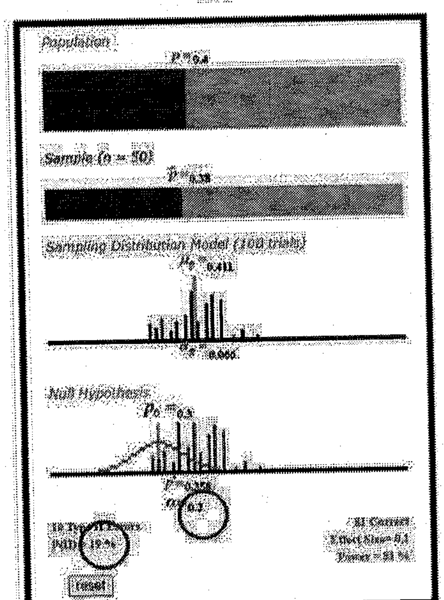


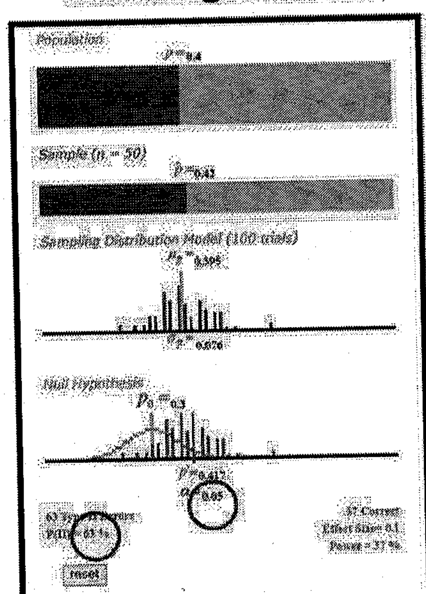


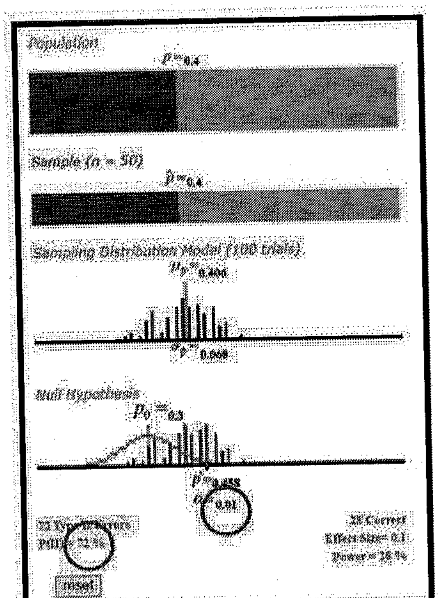


...you can <u>reduce</u> alpha (make it harder to find something significant)

To reduce Type II errors... mrfelling.com/sa4



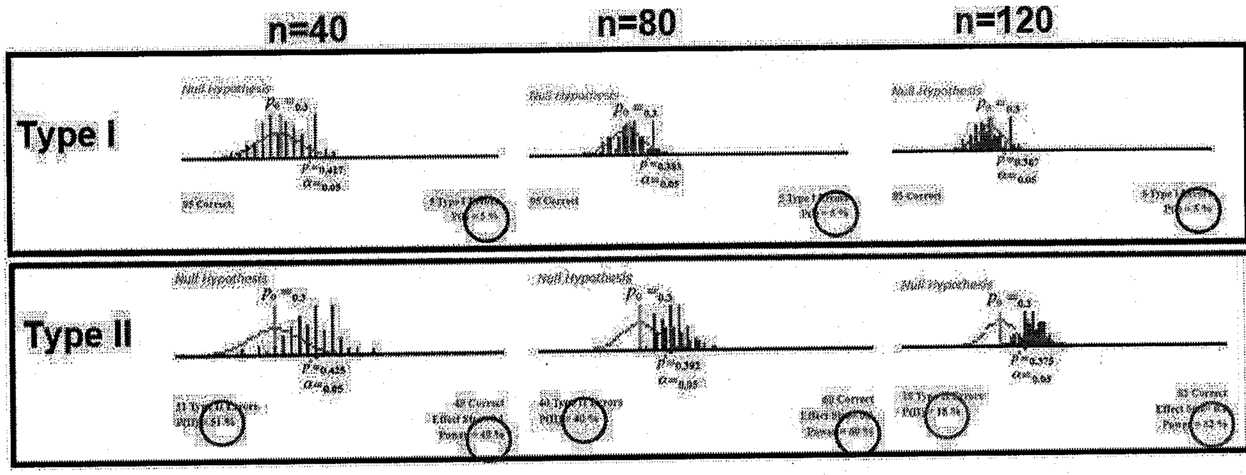




...you can <u>increase</u> alpha (make it easier to find something significant)

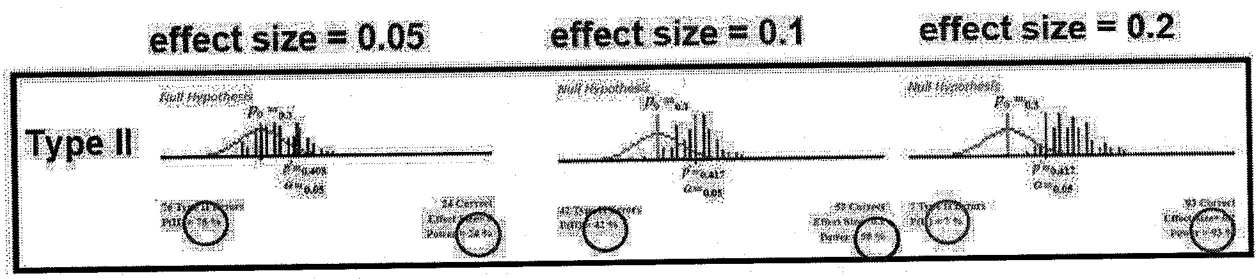
So changing alpha is a trade off between Type I and Type II errors.

How does sample size affect errors? mrfelling.com/sa4



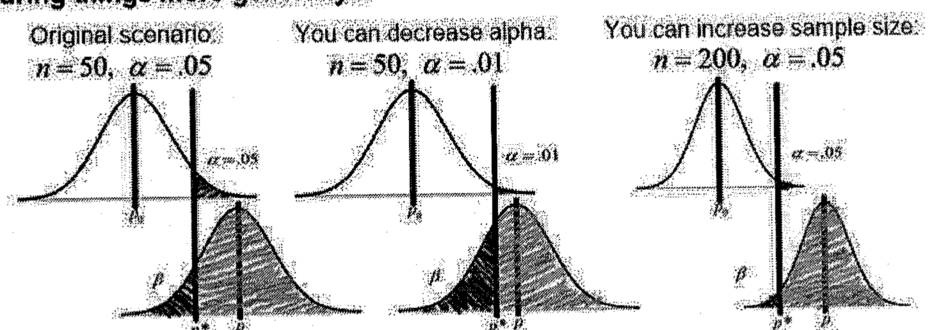
Increasing sample size decreases Type II errors, increases the power of the test, and leaves probability of Type I errors unchanged.

How does effect size affect errors? mrfelling.com/sa4



The larger the effect size (the more the real world is actually different from your null hypothesis) the easier it is to detect that difference. So larger effect sizes always produce lower Type II error, and higher power of the test.

Picturing things more generally...



Decreasing alpha moves the critical p* value further from Ho.

Type I errors decrease, but Type II errors increase, and this also reduces the power of the test.

With larger n, standard deviation decreases, so there is less overlap between the sampling distribution centered at the population and the normal distribution centered at Ho used for finding p-value.

Probability of Type I is still your chosen alpha, but Type II errors are reduced and power of the test is increased.

An example...

A machine produces a mechanical part requiring very tight tolerances, for a tolerance critical application. All parts produced are measured and must be discarded if out of tolerance, reducing profit. If a machine is found to be producing more than 10% of part out of tolerance, it is replaced (at considerable cost). The latest batch of 200 parts from one machine contained 28 which were out of tolerance. Is there sufficient evidence to conclude that this machine's proportion of bad parts is now above 10% (and should be replaced)?

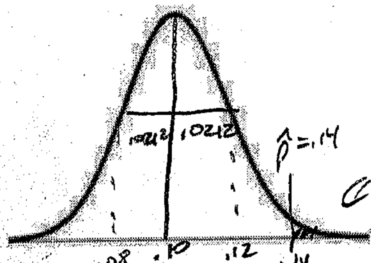
H_0: $p = .10$ (machine waste is acceptable)

H_A: $p > .10$ (machine waste is high, must be replaced)

This sample: $\hat{p} = \frac{28}{200} = .14$

Sampling distribution: $\mu_{\hat{p}} = p_0 = .10$

$\sigma_{\hat{p}} = \sqrt{\frac{pq}{n}} = \sqrt{\frac{(.1)(.9)}{200}} = .0212$



p – value = normalcdf(.14, 999, .10, .0212)

$= .03$

with $\alpha = .05$, p-value $= .03$ is low so we reject Ho.
There is sufficient statistical evidence to conclude the machine waste is high and machine should be replaced

What would constitute a Type I error? What is $P(\text{Type I error})$?

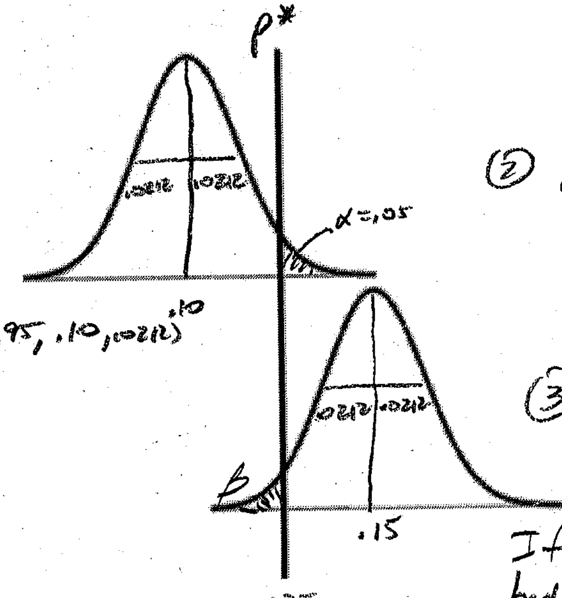
Ho is true but we reject : The machine is actually ok, but our sample leads us to replace it.

$P(I) = \alpha = .05$ (we choose this value)

What would constitute a Type II error? What is $P(\text{Type II error})$? What is the Power of this test?

Ho is false, but we fail to reject : The machine is actually faulty but we don't replace it.

$P(II) = \beta$ to find it, we would need to be told the effect size – how far above 10% the machine is actually out-of-tolerance. (Let's say we were told it was 15%)



① calculate p^* at border for our $\alpha = .05$

p^* = invNorm(.95, .10, .0212)

$p^* = \boxed{.135}$

② calculate β from p^* using p_{actual} for the machine

β = normalcdf(−999, .135, .15, .0212)

$\beta = \boxed{.24}$

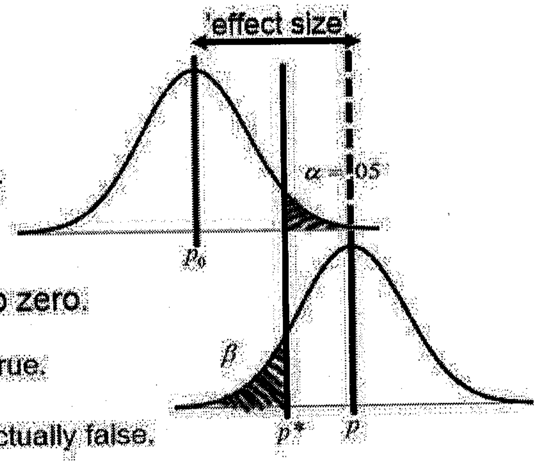
③ Calculate power from β

power $= 1 - \beta = 1 - .24 = \boxed{.76}$

If the machine was actually producing 15% bad parts, this statistical analysis would correctly detect the machine as bad 76% of the time.

Summary

Your analysis may correctly reject a false H_0 or correctly not reject a true H_0. But it is possible that the test will 'fail'. The greater the effect size, the easier it is to correctly 'see' the effect in an analysis.



But it is not possible to reduce the probability of error to zero.

Type I Error: We reject a Null Hypothesis that is actually true.

Type II Error: We do not reject a Null Hypothesis that is actually false.

$P(\text{Type I error}) = \alpha$

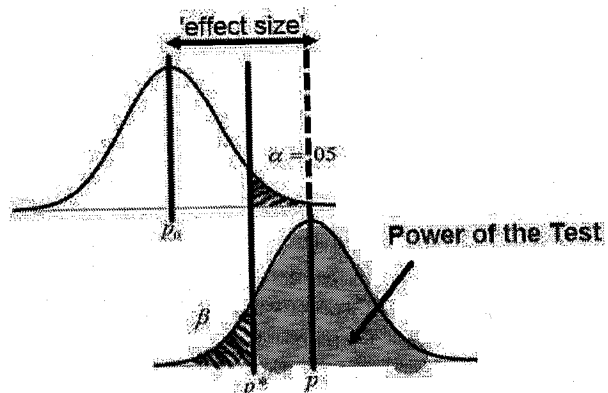
$P(\text{Type II error}) = \beta$

The **power of a test** is the **probability that it correctly rejects a false null hypothesis.**

$Power = 1 - \beta$

Decision:	Null Hypothesis is: True	Null Hypothesis is: False
Reject	Type I Error α	OK (power) $1-\beta$
Not Reject	OK $1-\alpha$	Type II Error β

Things that increase the power of the test (make it more likely that, if there is an effect to detect, the analysis will detect that effect as statistically significant):



1) Increase α $(\alpha \nearrow,\ \beta \searrow,\ power = 1-\beta \nearrow)$

2) Increase n (sample size) $(n \nearrow, \beta \searrow,\ power = 1-\beta \nearrow, \alpha \text{ is unchanged})$

3) Larger effect size (the bigger the effect, the easier it is to detect)

4) Decrease sampling variability by better accounting for sources of variability (block design experiments, stratified random samples, control over other variables)