Single sample hypothesis testing, II

9.07

3/02/2004

Outline

- Very brief review
- One-tailed vs. two-tailed tests
- Small sample testing
- Significance & multiple tests II: Data snooping
- What do our results mean?
- Decision theory and power

Brief review

- Null and alternative hypothesis
  - Null: only chance effects
  - Alternative: systematic + chance effects
- Assume the null is true
- Given this assumption, how likely is it that we'd see values at least as extreme as the ones we got?
- If it's highly unlikely, reject the null hypothesis, and say the results are statistically significant.
  - The results are due to a combination of chance and a systematic effect.

Key Concepts

- H_0 and H_a are contradictory (mutually exclusive)
- Support for H_a can only be obtained indirectly -- by rejecting H_0
- Rationale:
  - We can never prove anything true, but we can prove something false
  - We know the value of the parameter given H_0 but not given H_a

Why bother with H_a at all?

- The alternative hypothesis describes the condition that is contrary to the null hypothesis, and this can be directional or non-directional
  - Directional: The effect only occurs in a specific direction -- increases or decreases
  - Non-directional: The effect may be greater or less than a population parameter

Outline

- Very brief review
- One-tailed vs. two-tailed tests
- Small sample testing
- Significance & multiple tests II: Data snooping
- What do our results mean?
- Decision theory and power

A Tale of Two Tails

- Directional hypotheses are called one-tailed
  - We are only interested in deviations at one tail of the distribution
- Non-directional hypotheses are called two-tailed
  - We are interested in any significant deviations from H_0

The p-value for a test of $H_o: \mu = \mu_o$ against:

$H_a: \mu > \mu_o$ is prob

$H_a: \mu < \mu_o$ is prob

$H_a: \mu \neq \mu_o$ is prob

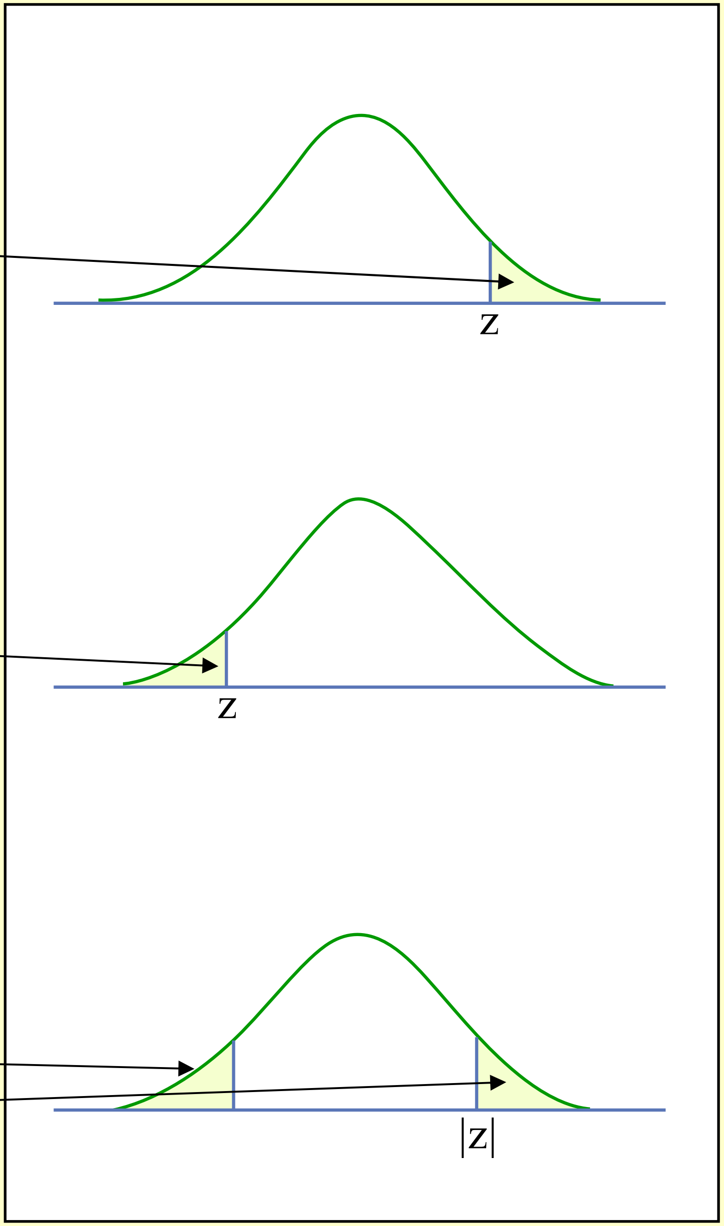
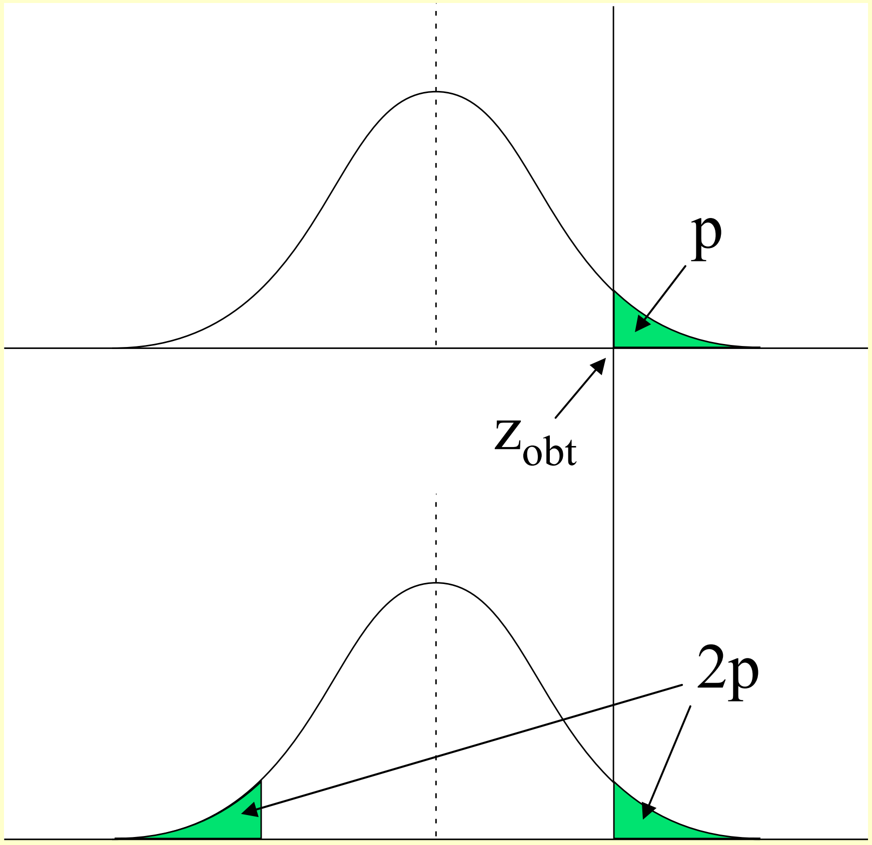


Figure by MIT OCW.

How do you decide to use a one- or two-tailed approach?

- A one-tailed approach is more liberal -- it is more likely to declare a result significant.
  - $t_{crit} = 1.69$ 5%, one-tailed
  - $t_{crit} = 2.03$ 5%, two-tailed
- There's no one right answer as to which test to use. People will debate this point.



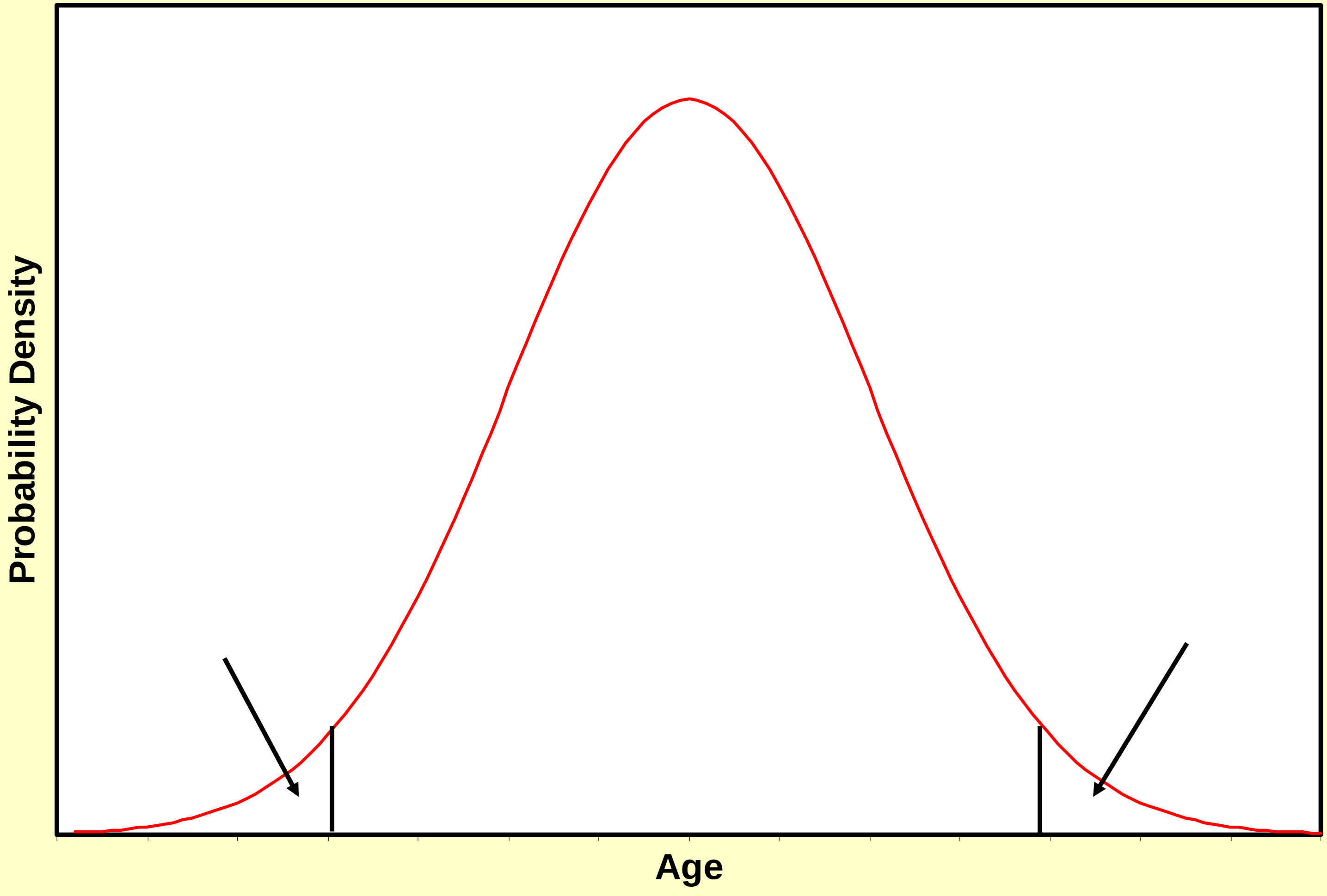
One Tail or Two? The moderate approach:

- If there's a strong, prior, theoretical expectation that the effect will be in a particular direction (A>B), then you may use a one-tailed approach. Otherwise, use a two-tailed test.
- Because only an A>B result is interesting, concentrate your attention on whether there is evidence for a difference in that direction.
  - E.G. does this new educational reform improve students' test scores?
  - Does this drug reduce depression?

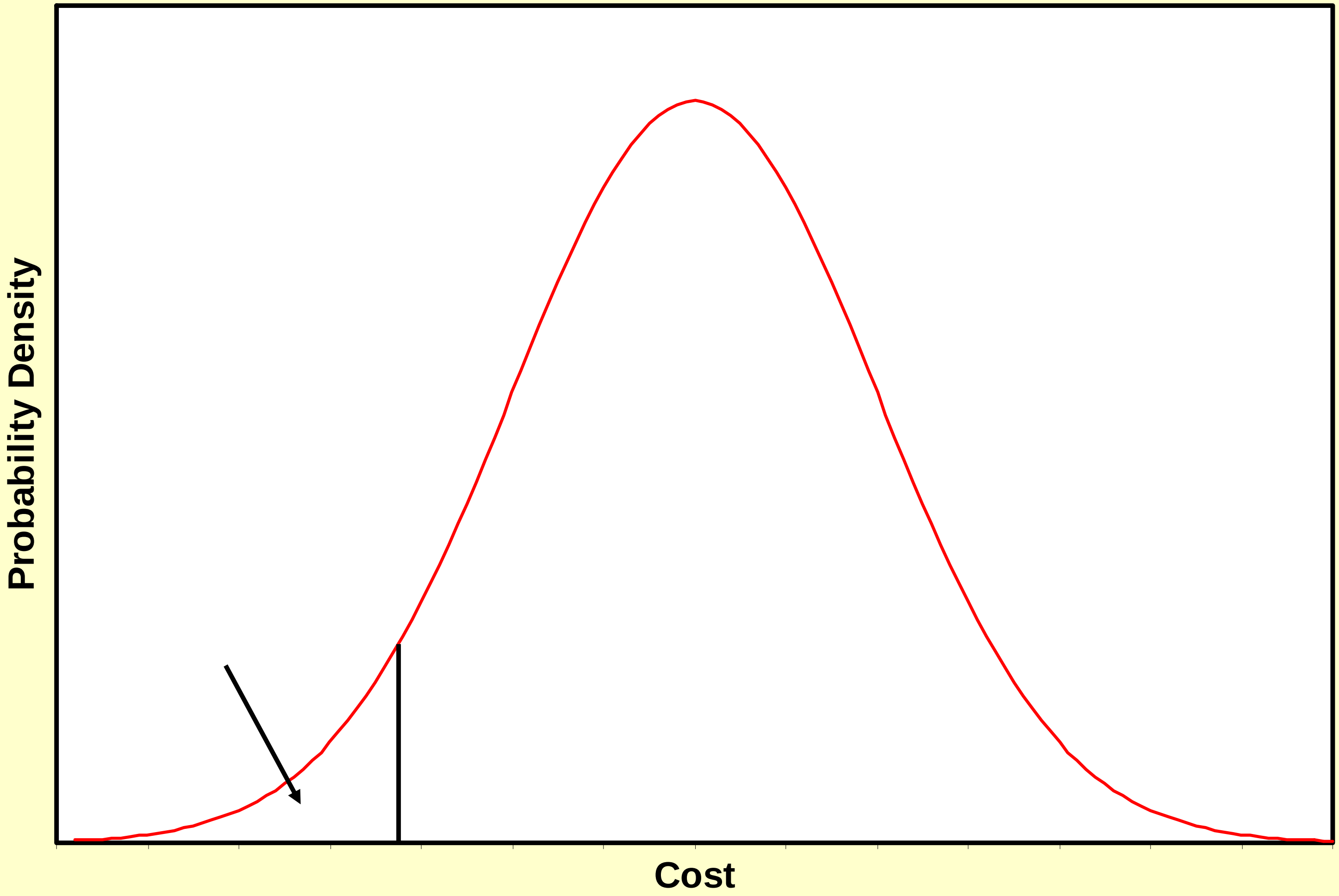
Examples of the moderate approach

- Is the age of this class different than the average age at MIT?
- Do you pay less for an education at a state university than you do at an Ivy League college?
- Is this class more boring than the norm for an MIT class?

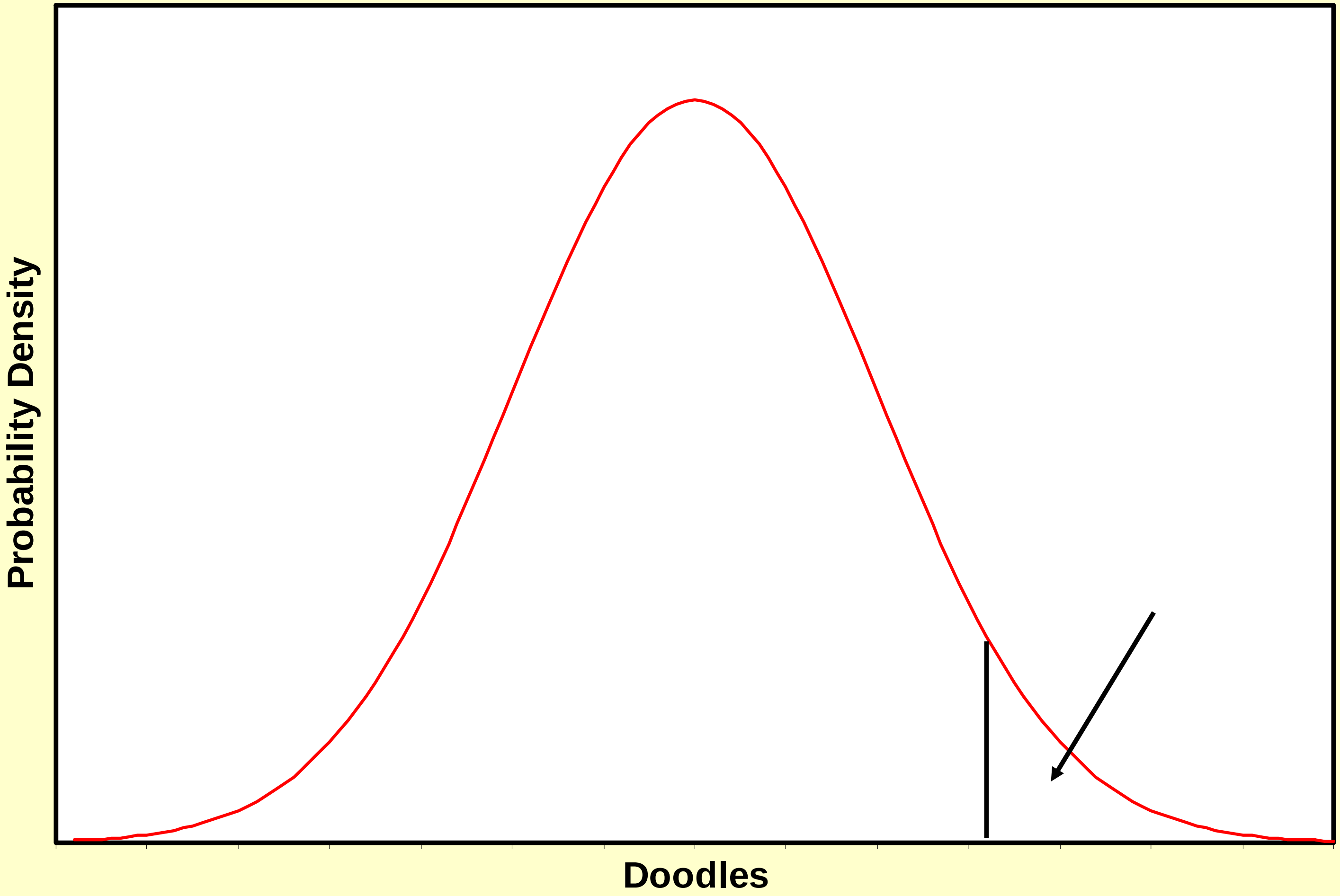
Age Distribution



Cost of an Ivy Education



Number of Doodles



One tail or two? The moderately conservative approach:

- The problem with the moderate approach is that you probably would actually find it interesting if the result went the other way, in many cases.
  - If the new educational reform leads to *worse* test scores, we'd want to know!
  - If the new drug actually *increases* symptoms of depression, we'd want to know!

One tail or two? The moderately conservative approach:

- Only use a one-tailed test if you not only have a strong hypothesis about the directionality of the results (A>B) but if it could also be argued that a result in the “wrong tail” (A<B) is meaningless, and might as well be due to chance.
- Put another way, only use a one-tailed test if you would not have been tempted, if the result went the “wrong” way, to switch to a two-tailed test (or switch the direction of your one-tailed test).
- It’s tough to meet this criterion.

The moderately conservative approach: a possible example

- It’s known how well students typically do on a intro statistics class.
- You test a new self-paced study guide, in addition to the instruction the students usually get, and have reason to believe this will improve how well they do in class.
- You might well consider any evidence that the students do *worse* as simply due to chance. After all, the students are getting the exact same instruction as they usually do – the study guide is extra.
- The moderately conservative approach would allow a one-tailed test in this case.

One tail or two: The conservative approach

- Always use two-tailed tests.

- More on one- vs. two-tailed tests later in the lecture.

Outline

- Very brief review
- One-tailed vs. two-tailed tests
- Small sample testing
- Significance & multiple tests II: Data snooping
- What do our results mean?
- Decision theory and power

Significance testing for small samples

- z-test is for known standard error, or large sample size (N>30)
- As you might imagine, for small sample sizes, we can again use the t-distribution instead, resulting in a t-test.

Example t-test

- A researcher needs to calibrate a spectrophotometer used to measure carbon monoxide (CO) concentration in the air.
- This is done by measuring the CO concentration in a special manufactured gas sample ("span gas"), known to have a precisely controlled concentration of 70 ppm.
- If the machine reads close to 70 ppm, it's ready for use. If not, it needs to be adjusted.

Spectrophotometer calibration

- One day the technician makes five readings on the span gas: 78, 83, 68, 72, 88.
- Can these readings have occurred by chance, if the machine is set properly, or do they show bias, i.e. that the machine needs to be adjusted?
- H_0: $\mu = 70$ ppm
- H_a: $\mu \neq 70$ ppm

Calculate the test statistic

- As before (with the z-test) we calculate the test statistic,
  t_{obt} = (observed – expected)/SE
- Under H_0, expected = μ = 70 ppm
- Observed = m = 77.8 ppm
- We don't know the SE of the mean, given H_0, but we can estimate it by SD/sqrt(N). But for this small sample size (N=5), we then need to use a t-test instead of a z-test.
- SD $\approx$ 8.07 ppm
  - Note this is the SD estimate where we divide by N-1, not N

Calculate the test statistic

- $m = 77.8$ ppm, $SE = 8.07/\text{sqrt}(5) \approx 3.61$ppm
- $t_{obt} = (77.8 - 70)/3.61 \approx 2.2$

Find the p-value

- $t_{obt} = 2.2$, d.f. = 4
- From the table in the back of your book, it looks like we're dealing with the 5% column.

Degrees of freedom	10%	5%	1%
1	3.08	6.31	31.82
2	1.89	2.92	6.96
3	1.64	2.35	4.54
4	1.53	2.13	3.75
5	1.48	2.02	3.36

Find the p-value

- However, this 5% is the area under one tail of the t-distribution.
- Recall the alternative hypothesis:
  - H_a: $\mu \neq 70$ ppm
  - We are interested in whether the spectrophotometer is off in either direction from 70 ppm.
  - This means we should be doing a 2-tailed t-test.
  - Note your book does a 1-tailed test, which doesn't really match H_a.
- $p = 2(0.05) = 0.10$
- This isn't much evidence against the null hypothesis, so we might decide not to calibrate.

Report the results

- “The spectrophotometer readings ($M=77.8$, $SD=8.07$) were not significantly different from those expected from a calibrated machine ($t(4)=2.2$, $p=0.10$, two-tailed).”

Outline

- Very brief review
- One-tailed vs. two-tailed tests
- Small sample testing
- Significance & multiple tests II: Data snooping
- What do our results mean?
- Decision theory and power

Significance and multiple tests (from the last lecture)

- Point of testing is to distinguish between real differences and chance variation.
- Does statistical significance mean that the result cannot be explained by chance variation?
  - No. Once in a while, an event that is unlikely to occur due to chance can actually occur.
  - We talked about this with confidence intervals – roughly 1 in 20 times, the true mean fell outside of the 95% confidence interval.

Significance and multiple tests

- Put another way, a researcher who runs 100 tests can expect to get 5 results which are "statistically significant" ($p<0.05$), and one which is "highly significant" ($p<0.01$), even if the null hypothesis is correct in every case.
- You cannot tell, for sure, whether a difference is real or just coincidence.
  - This is why science requires replicable results. If n independent tests all show a statistically significant result, the probability of this happening due to chance is very small.

A special case of multiple tests: data snooping

- Data snooping = deciding which tests to do once you've seen the data.
- Examples:
  - Disease clusters
  - One-tailed vs. two-tailed tests

Data snooping: Disease clusters

- Liver cancer is rare. The chance of having 2 or more cases in a given town in a year (a "cluster") with 10,000 inhabitants is about 0.5%
- A cluster of liver cancer cases causes a researcher to search for causes, like water contamination.
- But, with a bunch of small towns of this size, looked at over a 10-year time period, it's likely you'll see a few clusters like this. 100 towns x 10 years = 1000 cases. 0.005*1000 = 5.

Data snooping: One-tailed vs. two-tailed significance testing

- This is where you look at your data to see whether your sample average is bigger or smaller than expected, before you choose your statistical test.
- H_0: μ=50
- m = 65, so, uh, H_a: $\mu > 50$. So, I'll do a one-tailed t-test looking at the upper tail…
- This is not allowed, and many statisticians recommend always using two-tailed tests, to guard against this temptation.

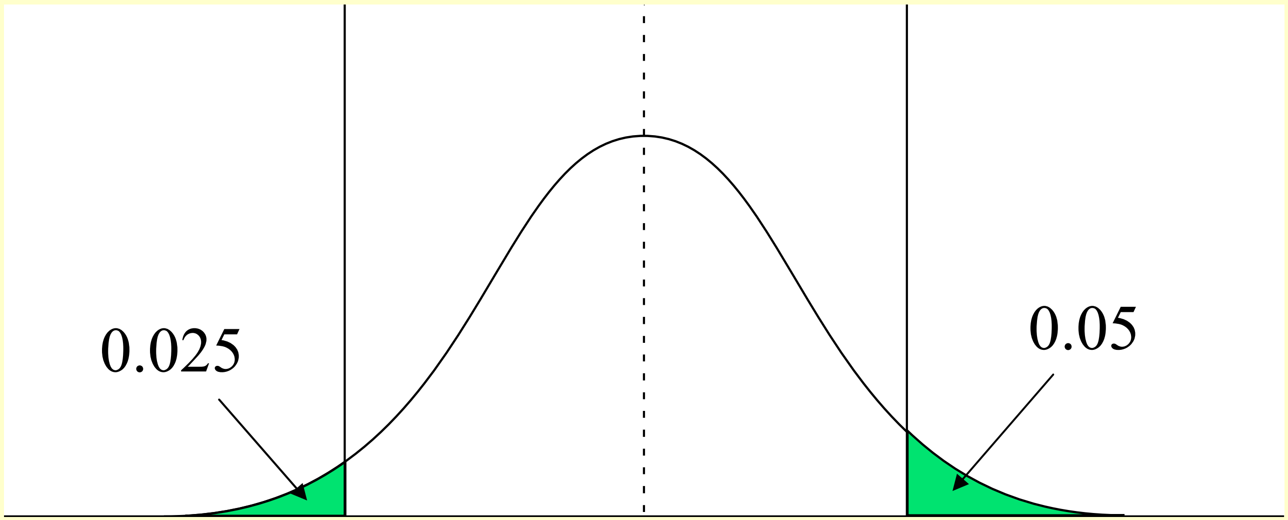
Consequences of data snooping: 1-tailed vs. 2-tailed tests

- Suppose H_0: $\mu = 20$.
- You set $\alpha=0.05$ as your criterion, and initially plan a 1-tailed test (H_a: $\mu > 20$).
- Running the experiment, you find that m=15. Oops, you switch to a 2-tailed test to see if this is significant.
- What is p?

Data snooping & the switch to a 2-tailed test

- Reject the null hypothesis if z_{obt} falls in the 5% region of the upper tail (1-tailed test)
- Or, switching to a 2-tailed test with $\alpha=0.05$, if it falls in the 2.5% region of the lower tail.
- Thus, if z_{obt} passes the test, you should report $p<0.075$, not $p<0.05$.
  - Probably the researcher incorrectly reports $p<0.05$.
- This is like a "one-and-a-half" tailed test.

Switching to a 2-tailed test



Data snooping and the switch to a 1-tailed test

- Similiarly, you might start off assuming you'll do a 2-tailed test, with α=0.05.
  - 2.5% in each of the two tails
- But when you get the data, z_{obt} isn't big enough to fall in the 2.5% region of the upper tail, but is big enough to fall in the 5% region of the upper tail.
- You decide to switch to a 1-tailed test.
- Again, this amounts to a one-and-a-half tailed test.
  - Reject the null hypothesis if z_{obt} falls in the 2.5% region of the lower tail (2-tailed test),
  - Or, switching to a 1-tailed test, if z_{obt} falls in the 5% region of the upper tail.

Correcting for one- vs. two-tailed tests

- If you think a researcher has run the wrong kind of test, it's easy to recalculate the p-value yourself.
- p(one-tailed) = ½ p(two-tailed)
- 1.5 p(one-tailed) = p(1.5-tailed)
- Etc.

A special case of multiple tests: data snooping

- If you’re going to use your data to pick your statistical test, you should really test your conclusions on an independent set of data.
- Then it’s like you used *pilot* data (or other previous experiments) to form your hypothesis, and tested the hypothesis independently on other data. This is allowed.

Outline

- Very brief review
- One-tailed vs. two-tailed tests
- Small sample testing
- Significance & multiple tests II: Data snooping
- What do our results mean?
- Decision theory and power

What do our results mean?

- Significance
- Importance
- Size of the effect
- Does the difference prove the point?

Was the result significant?

- There is no true sharp dividing line between probable and improbable results.
  - There's little difference between $p=0.051$ and $p=0.049$, except that some journals will not publish results at $p=0.051$, and some readers will accept results at $p=0.049$ but not at $p=0.051$.

Was the result important?

- “Significant” does not mean you care about it.
- Some of what “important” means has to do with what you’re studying.

Importance and what you are studying

- Suppose you give children a vocabulary test consisting of 40 words that the child must define. 2 points are given for a correct answer, 1 point for a partially correct answer.
- City kids, ages 6-9, are known to average 26 points on this test.
- Study 2500 rural kids, ages 6-9.
- Rural kids get an average of 25 points. This difference from the expected 26 points is highly significant.
  - We would probably really do a two-sample test here, not a one-sample test. But we don't cover that until next week…

Importance and what you are studying

- But is the result important?
- The z-test only tells us that this one point difference is unlikely to have occurred by chance.
- Suppose you studied the entire population, and found this difference between rural and big city kids. What would this difference mean?
  - A one-point difference in average scores only amounts to partial credit on one word out of a test of 40 words.
  - If anything, the investigators have provided evidence that there is almost no difference between rural and big city kids on this test.

Was the result important?

- The p-value of a test depends upon the sample size.
- z_{obt} = (observed – expected)/SE   (same idea with t_{obt})
- SE has a sqrt(N) in the denominator – as N increases, SE decreases, and z_{obt} (t_{obt}) increases.
  - As N increases, the same difference between observed & expected becomes more significant.
- An important result can be non-significant just because you didn't take a big enough sample.
- A very small, unimportant result can be significant just because the sample size is so big.

Picking N

- As with confidence intervals, we can estimate what sample size we should use, for a given anticipated effect size.
- For the vocabulary test example, suppose an effect is only important if the rural kids' scores are at least 10 points different from the city kids' score of 26.
- How many rural kids should we give the vocabulary test to, if we want to be able to detect a significant difference of this size, with α=0.01?

Picking N

- For $\alpha=0.01$, $z_{crit} = 2.58$
- $z_{obt} = (\text{observed} – \text{expected})/SE$
- $SE = SD/\text{sqrt}(N)$
  - Need to approximate SD, either from previous data, or just by taking a guess.
  - Here, we guess SD = 10
- $z_{obt} = 10/(10/\text{sqrt}(N)) = \text{sqrt}(N)$
- A difference of 10 will be highly significant if sqrt(N) > 2.58, which implies we need a sample size of at least 2.58^2, i.e. $N \geq 7$.
  - Note in the example, N=2500!

Does the difference prove the point the study was designed to test?

- No, a test of significance does not check the design of the study. (There are tons of things that could go wrong, here.)
  - Is it a simple random sample, or is there some bias?
    - Did our poll call only phone numbers in the phonebook?
  - Could the result be due to something other than the intended systematic effect?
    - Did drug study subjects figure out whether they had been given the true drug vs. placebo?
  - Is the null hypothesis appropriate?
    - Does it assume that the stimulus levels are randomly selected, when actually they follow a pattern the subject might notice?

Outline

- Very brief review
- One-tailed vs. two-tailed tests
- Small sample testing
- Significance & multiple tests II: Data snooping
- What do our results mean?
- Decision theory and power

Decisions, Decisions...

- Hypothesis testing is an example of the application of *decision theory*
- We want to use the evidence from our sample to decide between two hypotheses
- This involves a trade-off between different types of errors

Decision theory and tradeoffs between types of errors

- Think of a household smoke detector.
- Sometimes it goes off and there's no fire (you burn some toast, or take a shower).
  - A *false alarm*.
  - A *Type I error*.
- Easy to avoid this type of error: take out the batteries!
- However, this increases the chances of a *Type II error*: there's a fire, but no alarm.

Decision theory and tradeoffs between types of errors

- Similarly, one could reduce the chances of a Type II error by making the alarm hypersensitive to smoke.
  - Then the alarm will by highly likely to go off in a fire.
  - But you'll increase your chances of a false alarm = Type I error. (The alarm is more likely to go off because someone sneezed.)
- There is typically a tradeoff of this sort between Type I and Type II errors.

A table

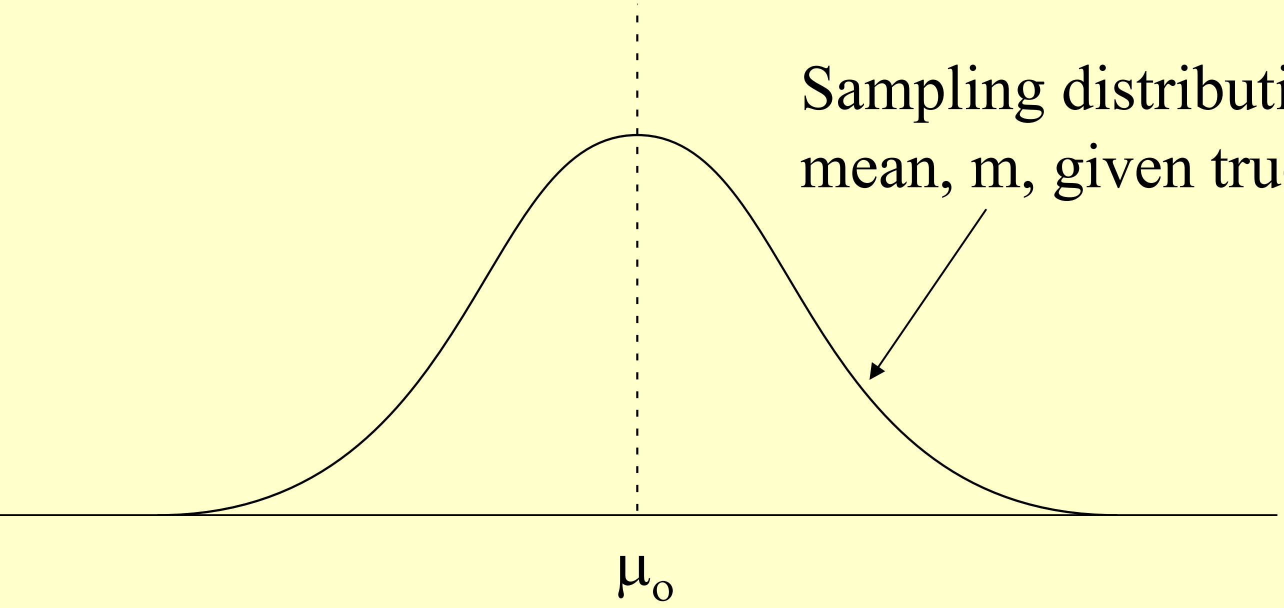
	No fire	Fire
No alarm	No error	Type II
Alarm	Type I	No error

A table

		Truth about the population	
		H_o true (No fire)	H_a true (Fire)
Decision based on sample	Accept H_o (No alarm)	No error (correct null response)	Type II (miss)
	Reject H_o (Alarm)	Type I (false alarm)	No error (hit)

More on the tradeoff between Type I and Type II errors

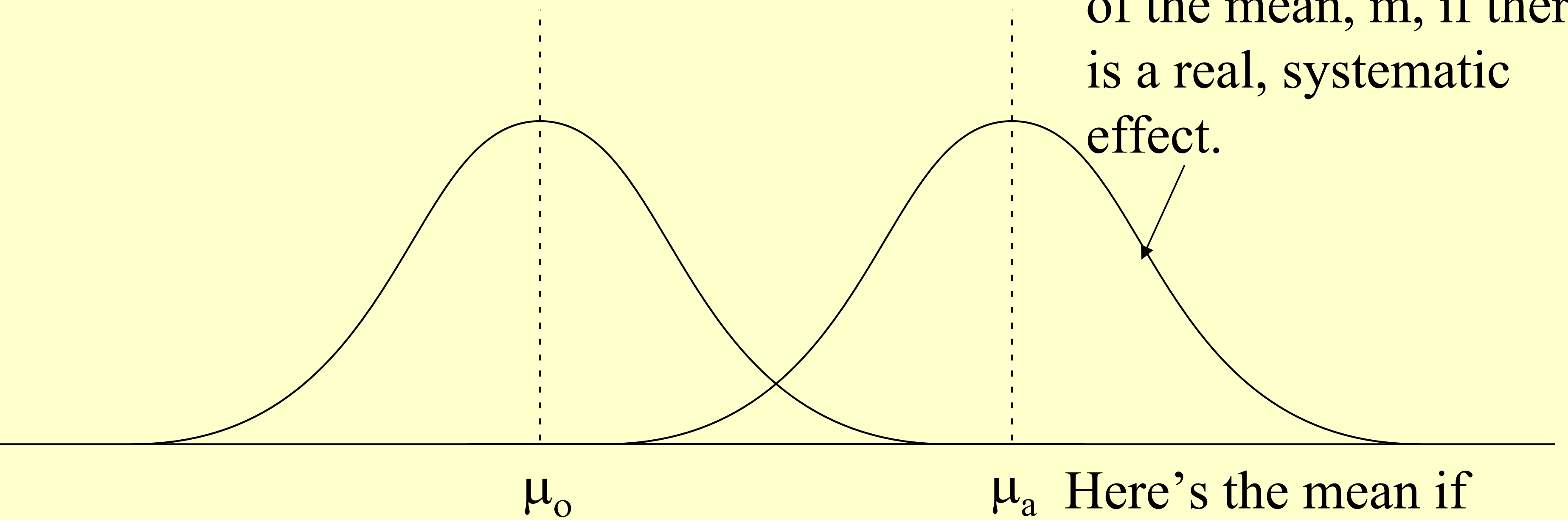
- Consider the null hypothesis, H_0: $\mu=\mu_0$



More on the tradeoff between Type I and Type II errors

- And the alternative:

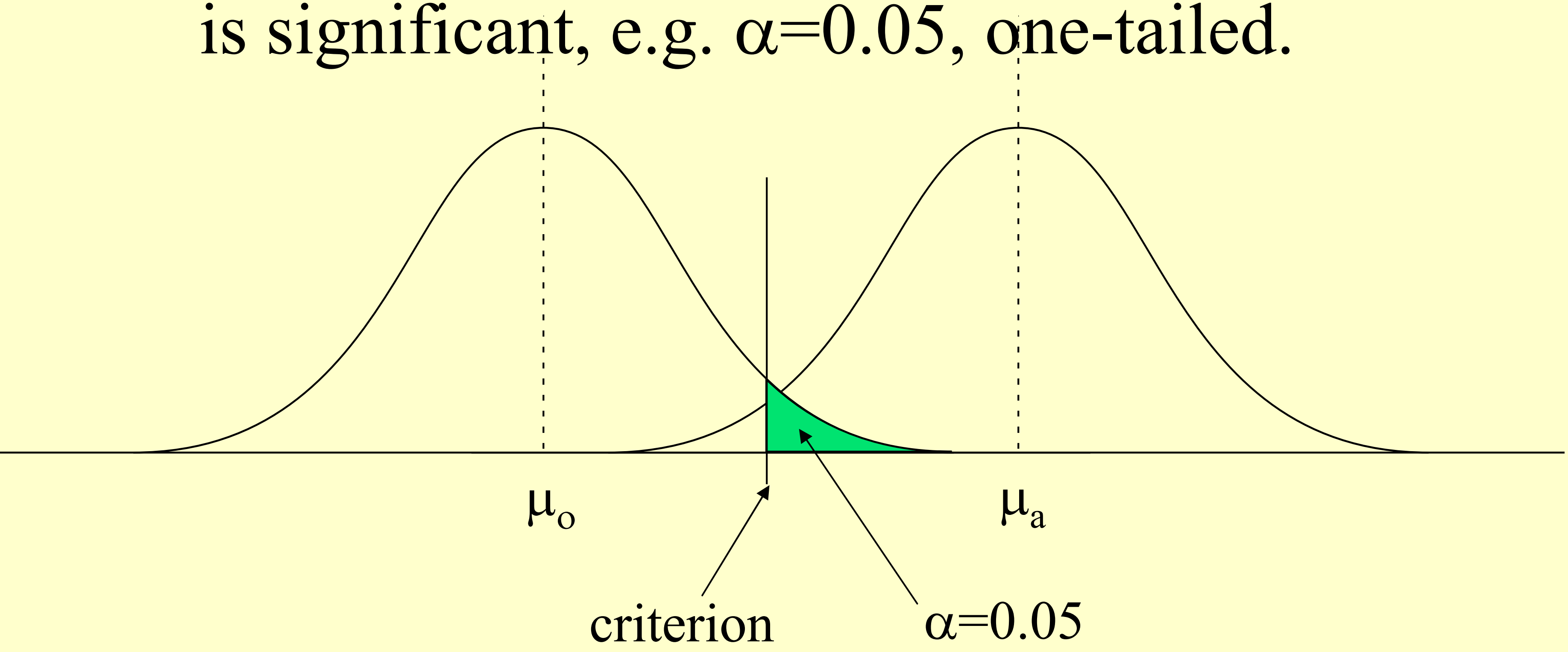
Sampling distribution of the mean, m, if there is a real, systematic effect.



Here's the mean if there's a systematic effect. Often we don't know this.

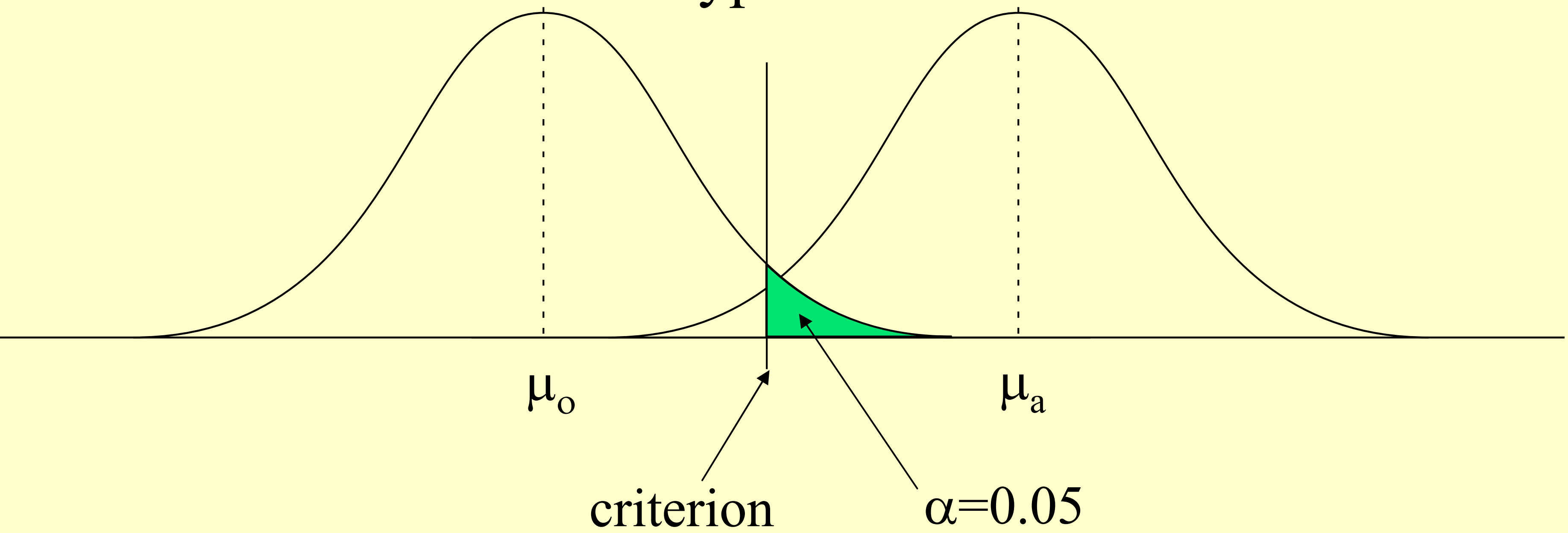
More on the tradeoff between Type I and Type II errors

- We set a criterion for deciding an effect is significant, e.g. α=0.05, one-tailed.



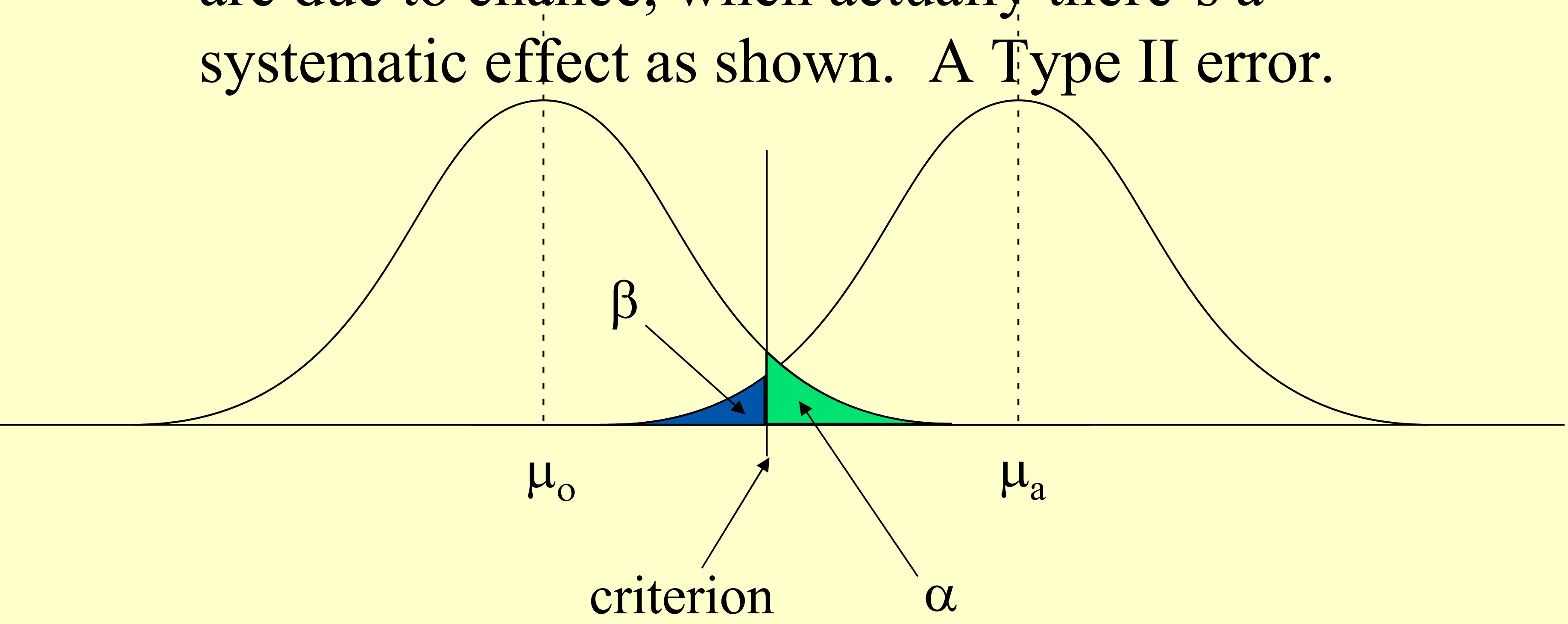
More on the tradeoff between Type I and Type II errors

- Note that α is the probability of saying there's a systematic effect, when the results are actually just due to chance. A Type I error.



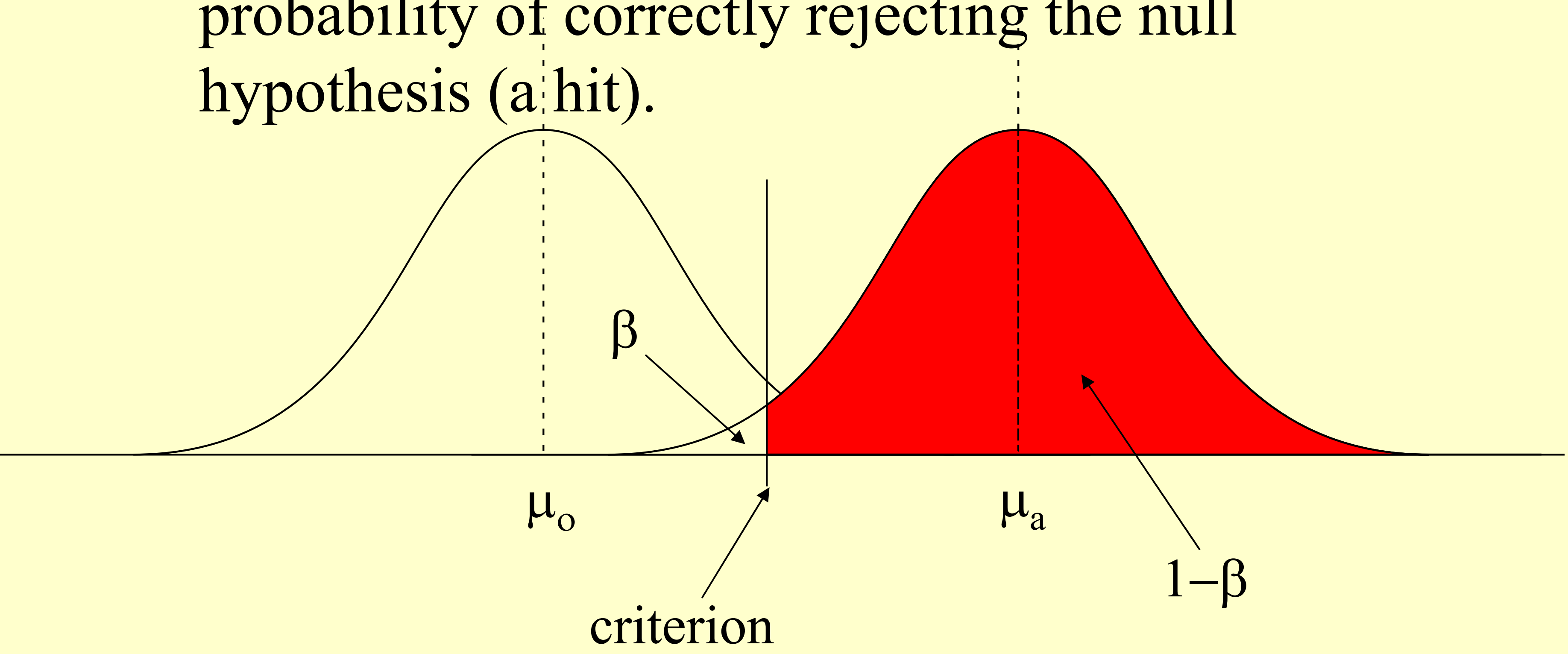
More on the tradeoff between Type I and Type II errors

- Whereas β is the probability of saying the results are due to chance, when actually there's a systematic effect as shown. A Type II error.



More on the tradeoff between Type I and Type II errors

- Another relevant quantity: 1-β. This is the probability of correctly rejecting the null hypothesis (a hit).



Moving the criterion around changes the % of false alarms (α) and "hits" ($1-\beta$)

- A natural tradeoff between Type I and Type II errors.

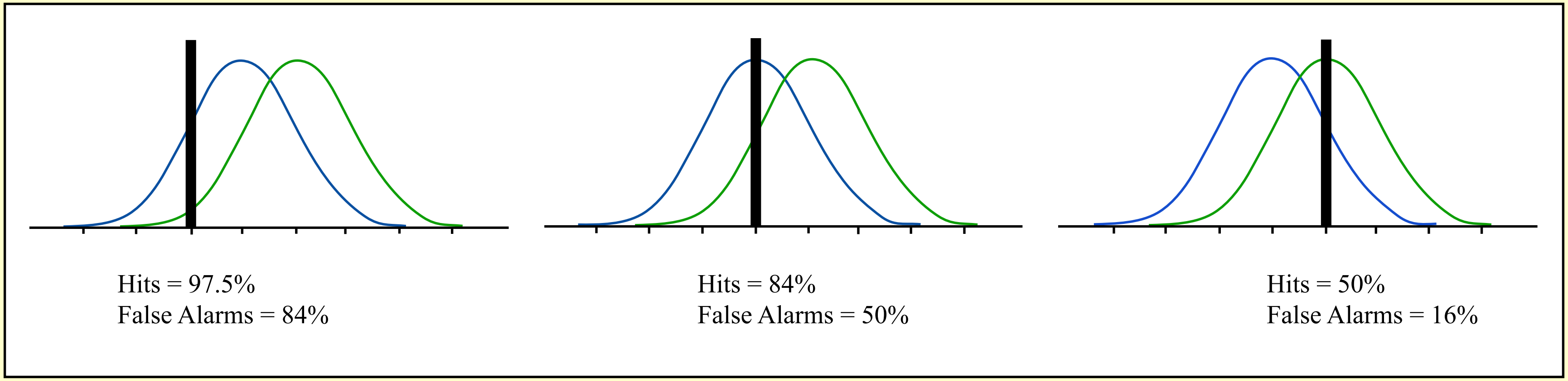


Figure by MIT OCW.

- This is one reason we test $x \geq 14$ instead of $x=14$ (binomial distribution). The latter reduces false alarms, but increases the number of misses.

Type I and Type II errors

- Hypothesis testing as usually done is minimizing α, the probability of a Type I error (false alarm).
- This is, in part, because we don't know enough to maximize $1-\beta$ (hits).
- However, $1-\beta$ is an important quantity. It's known as the *power* of a test.

Statistical power

- The probability that a significance test at fixed level α will reject the null hypothesis when the alternative hypothesis is true.
- In other words, power describes the ability of a statistical test to show that an effect exists (i.e. that H_o is false) when there really is an effect (i.e. when H_a is true).
- A test with weak power might not be able to reject H_o even when H_a is true.

An example

- Can a 6-month exercise program increase the mineral content of young women's bones? A change of 1% or more would be considered important.
- What is the power of this test to detect a change of 1% if it exists, given that we study a sample of 25 subjects?
  - Again, you'd probably really run this as a two-sample test…

How to figure out the power of a significance test (p. 471)

- Ho: $\mu=0\%$ (i.e. the exercise program has no effect on bone mineral content)
- Ha: $\mu>0\%$ (i.e. the exercise program has a beneficial effect on bone mineral content).
- Set α to 5%
- Guess the standard deviation is $\sigma=2\%$

First, find the criterion for rejecting the null hypothesis with α=0.05

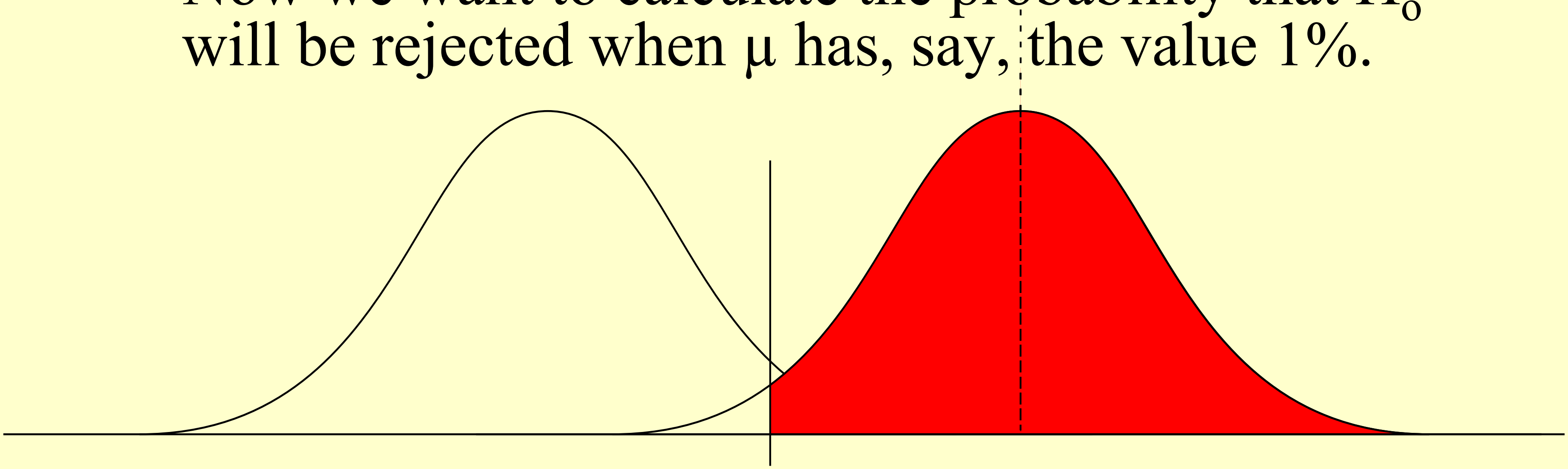
- H_o: μ=0%; say n=25 and σ=2%
- H_a: μ>0%

- The z-test will reject H_o at the α =.05 level when: $z=(m-\mu_o)/(\sigma/\text{sqrt}(n))$

  $= (m-0)/(2/5) \geq 1.645$
- So $m \geq 1.645(2/5) \rightarrow m \geq 0.658\%$ is our criterion for deciding to reject the null.

Step 2

- Now we want to calculate the probability that H_o will be rejected when μ has, say, the value 1%.



- We want to know the area under the normal curve from the criterion (m=0.658) to $+\infty$
- What is z for m=0.658?

Step 2

- Assuming σ for the alternative is the same as for the null, $\mu_a=1$

  $z_{crit} = (0.658-1)/(2/sqrt(25) = -0.855$

- $Pr(z \geq -.855) = .80$
- So, the power of this test is 80%. This test will reject the null hypothesis 80% of the time, if the true value of the parameter $\mu = 1\%$

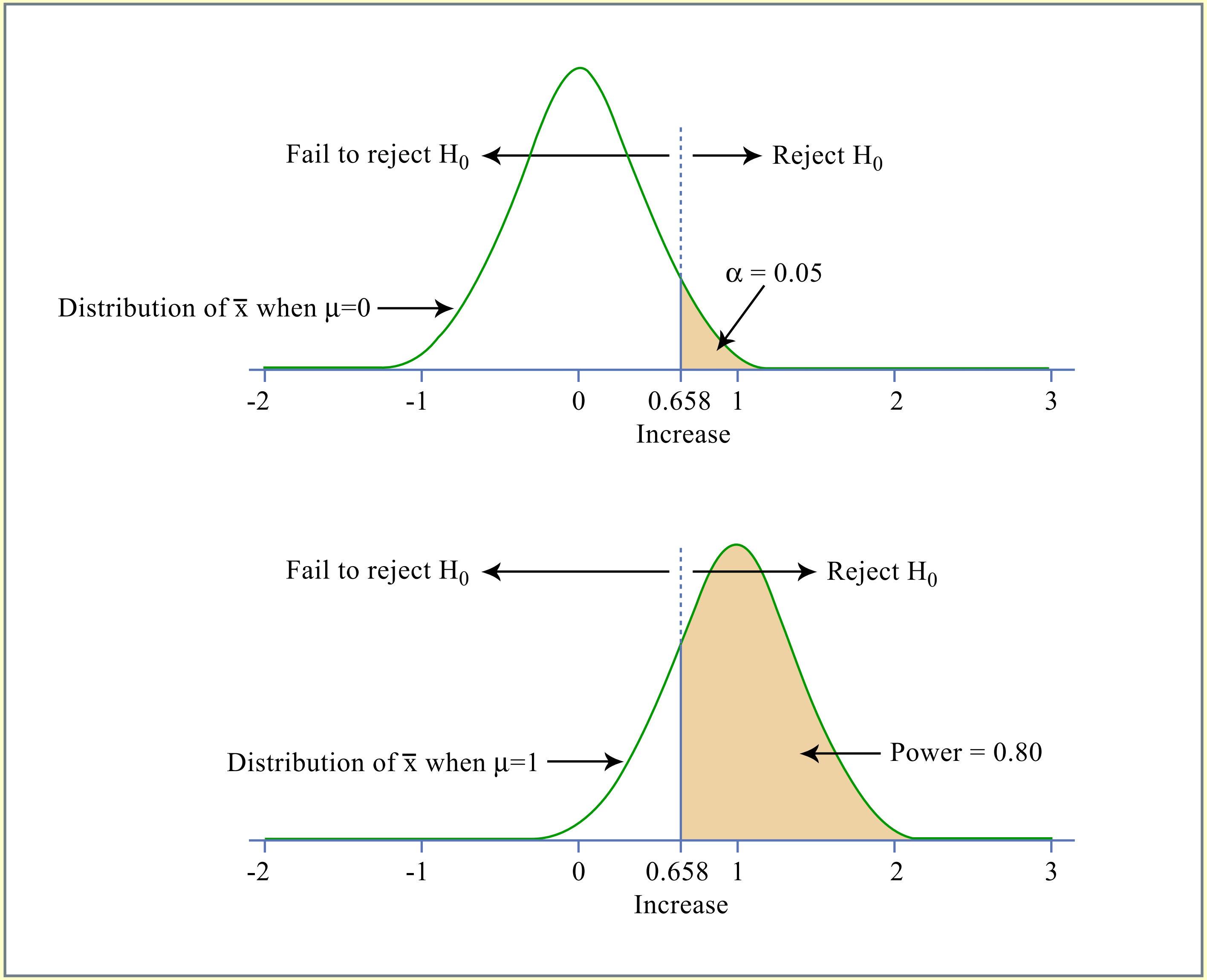


Figure by MIT OCW.

How to increase power

- Increase α
  - Make the smoke alarm more sensitive. Get more false alarms, but more power to detect a true fire.
- Increase n.
- Increase the difference between the μ in H_a and the in μ_o in H_o.
- Decrease σ.