# Two-sample hypothesis testing, I

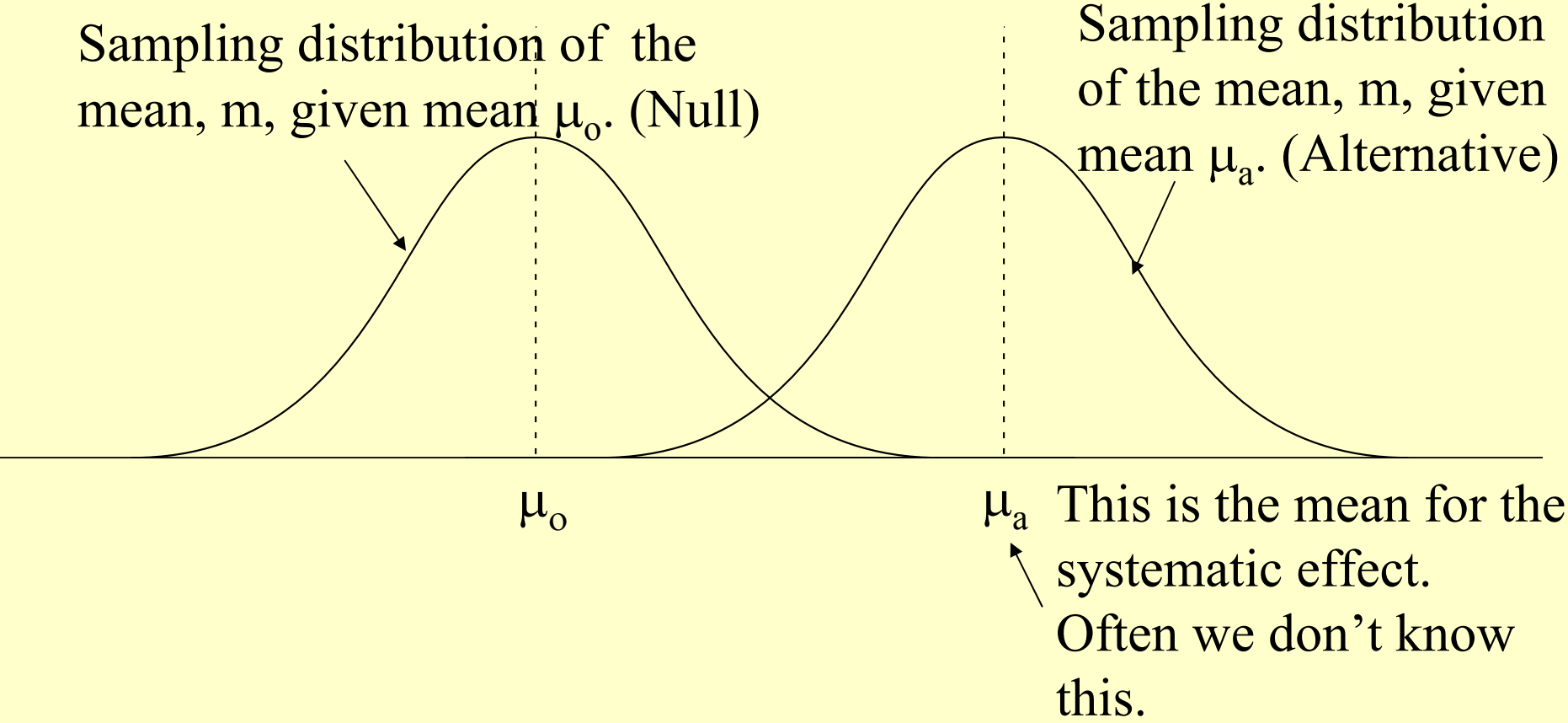
9.07

3/09/2004

- But first, from last time…

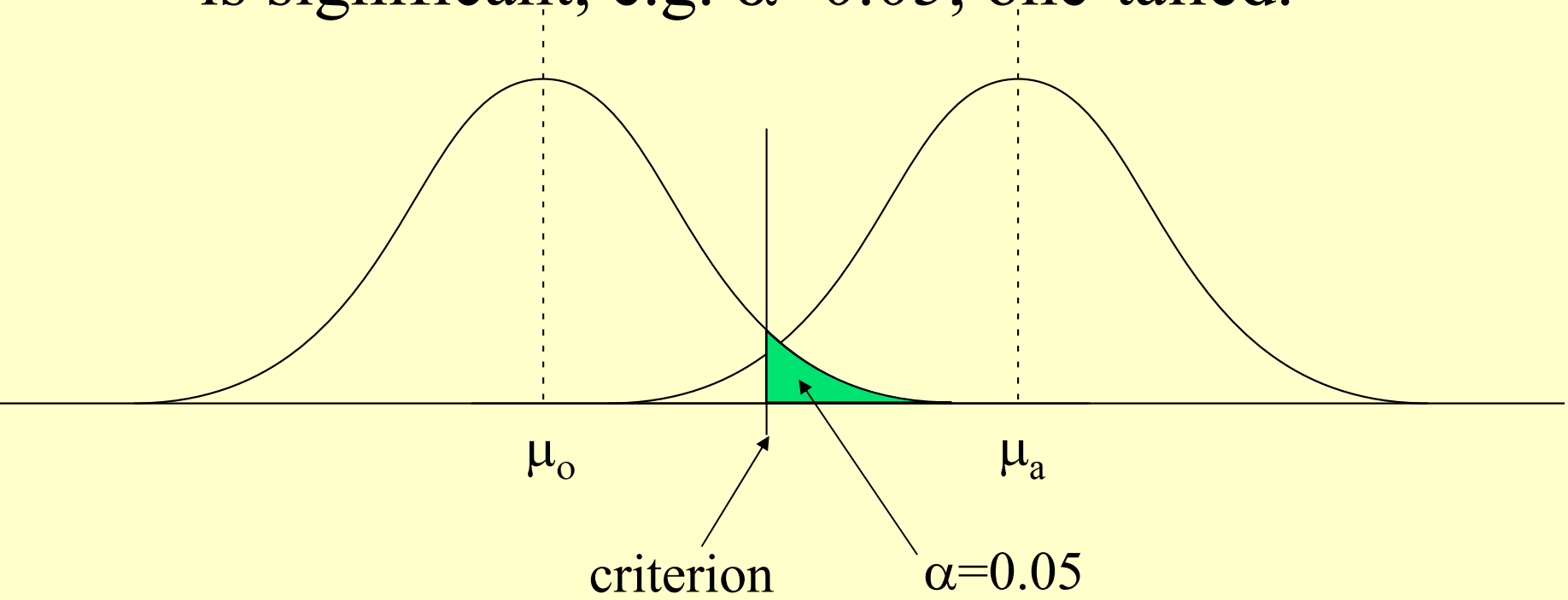
# More on the tradeoff between Type I and Type II errors

- The null and the alternative:



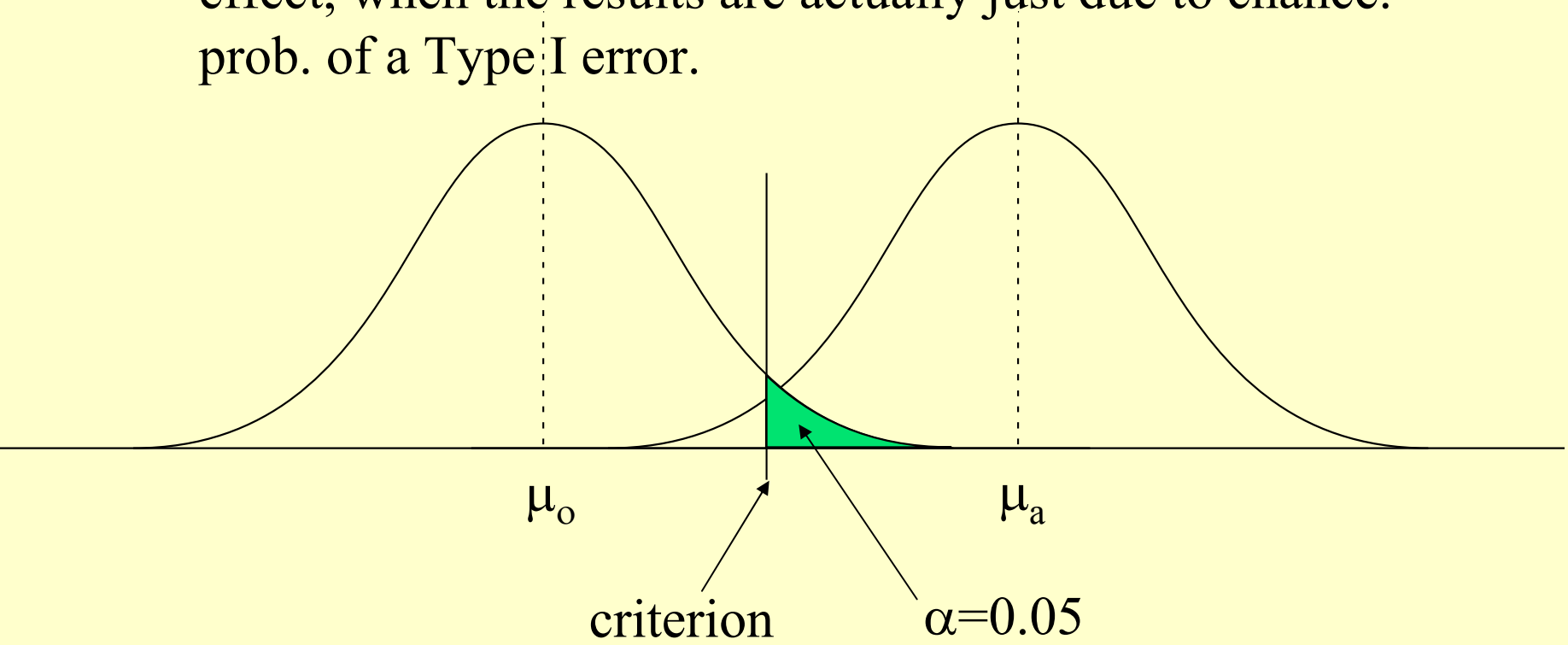
# More on the tradeoff between Type I and Type II errors

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



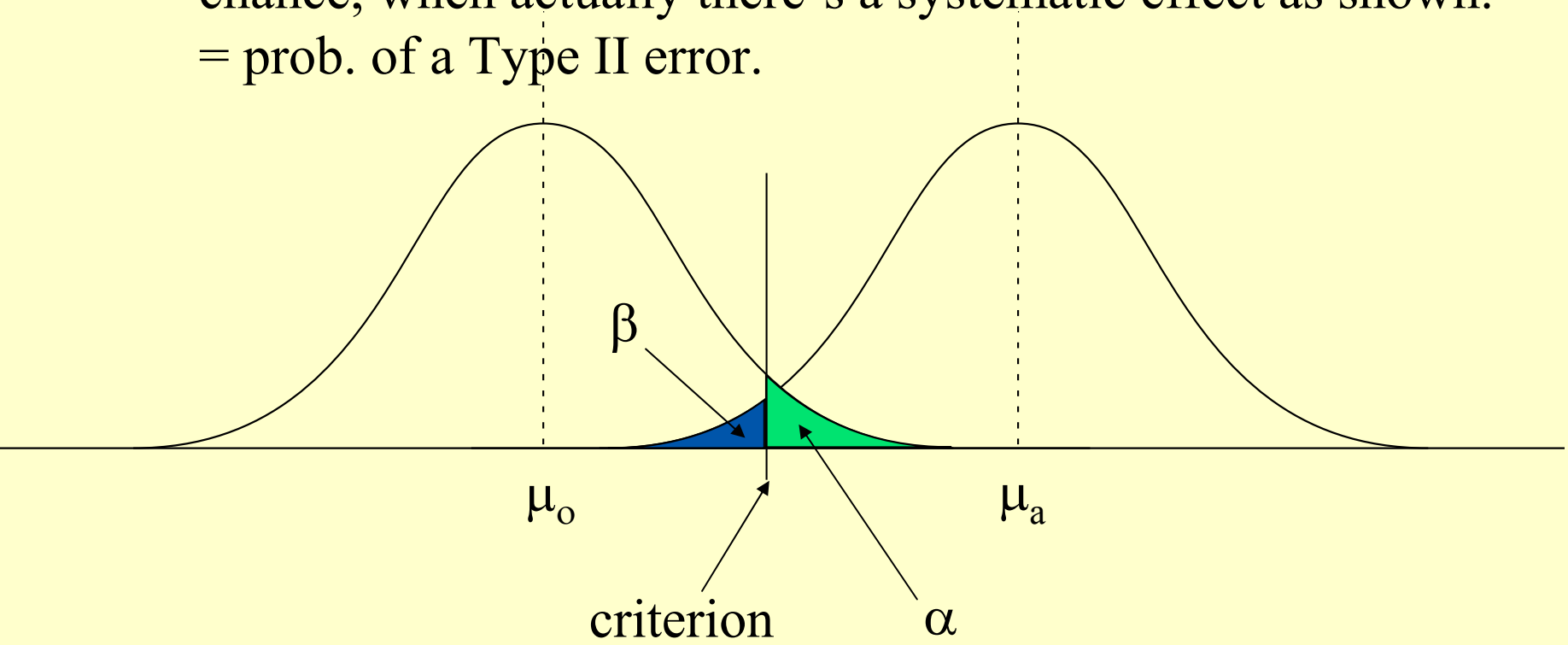
# More on the tradeoff between Type I and Type II errors

- Note that $\alpha$ is the probability of saying there's a systematic effect, when the results are actually just due to chance. = prob. of a Type I error.



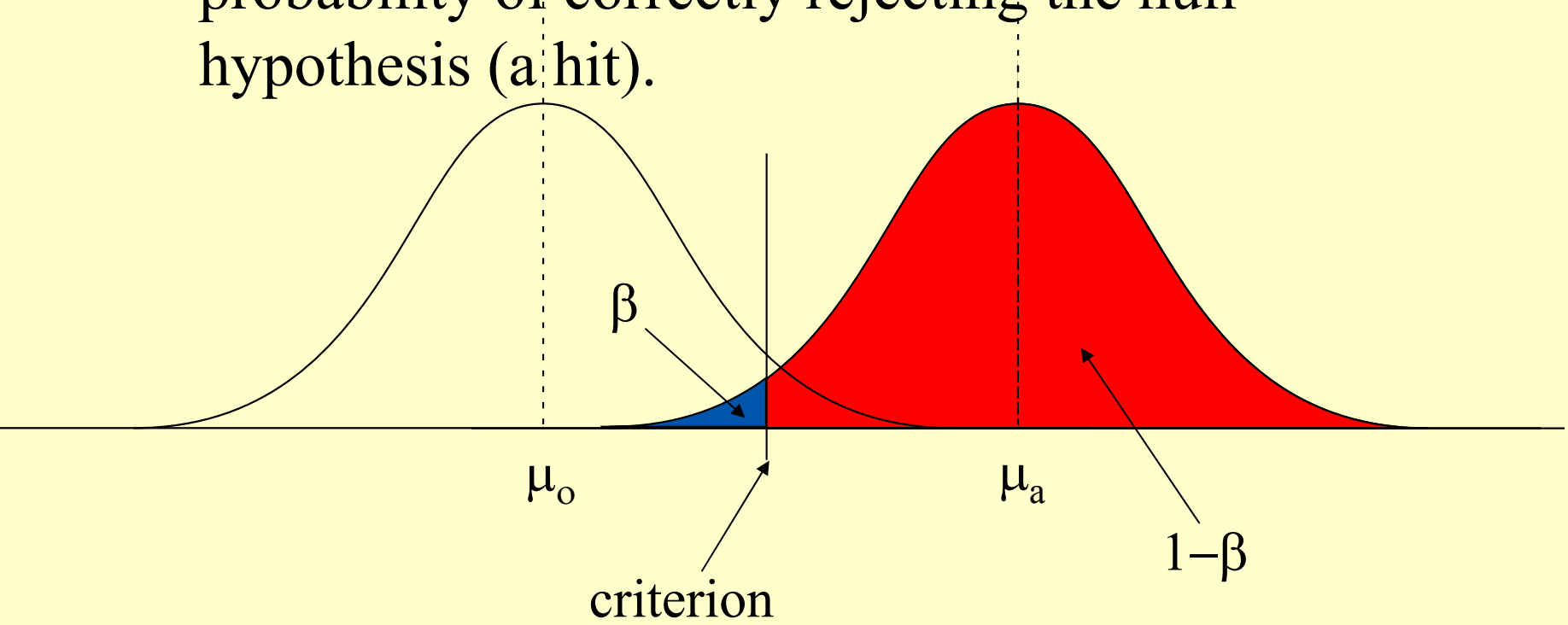
# More on the tradeoff between Type I and Type II errors

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

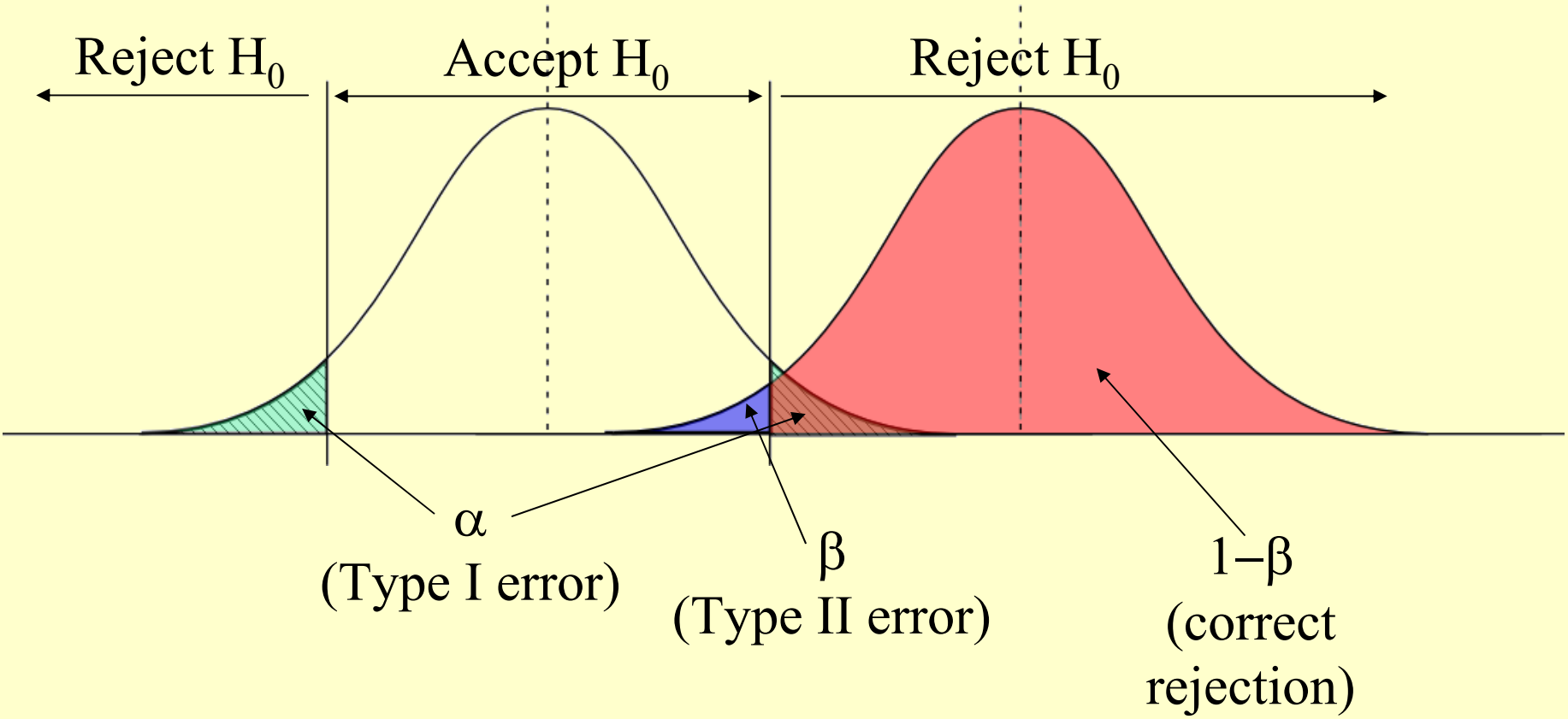


# More on the tradeoff between Type I and Type II errors

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



# For a two-tailed test



# Type I and Type II errors

- Hypothesis testing as usually done is minimizing $\alpha$, 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.

$$1-\beta = P(\text{rejecting } H_0 \mid H_a \text{ true})$$

# Statistical power

- The probability that a significance test at fixed level $\alpha$ will reject the null hypothesis when the alternative hypothesis is true.
  $= 1 - \beta$
- 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.

# Example: why we care about power

- Suppose that factories that discharge chemicals into the water are required to prove that the discharge is not affecting downstream wildlife.
- Null hypothesis: no effect on wildlife
- The factories can continue to pollute as they are, so long as the null hypothesis is not rejected at the 0.05 level.

Cartoon guide to statistics

# Example: why we care about power

- A polluter, suspecting he was at risk of violating EPA standards, could devise a very weak and ineffective test of the effect on wildlife.
- Cartoon guide extreme example: “interview the ducks and see if any of them feel they are negatively impacted.”

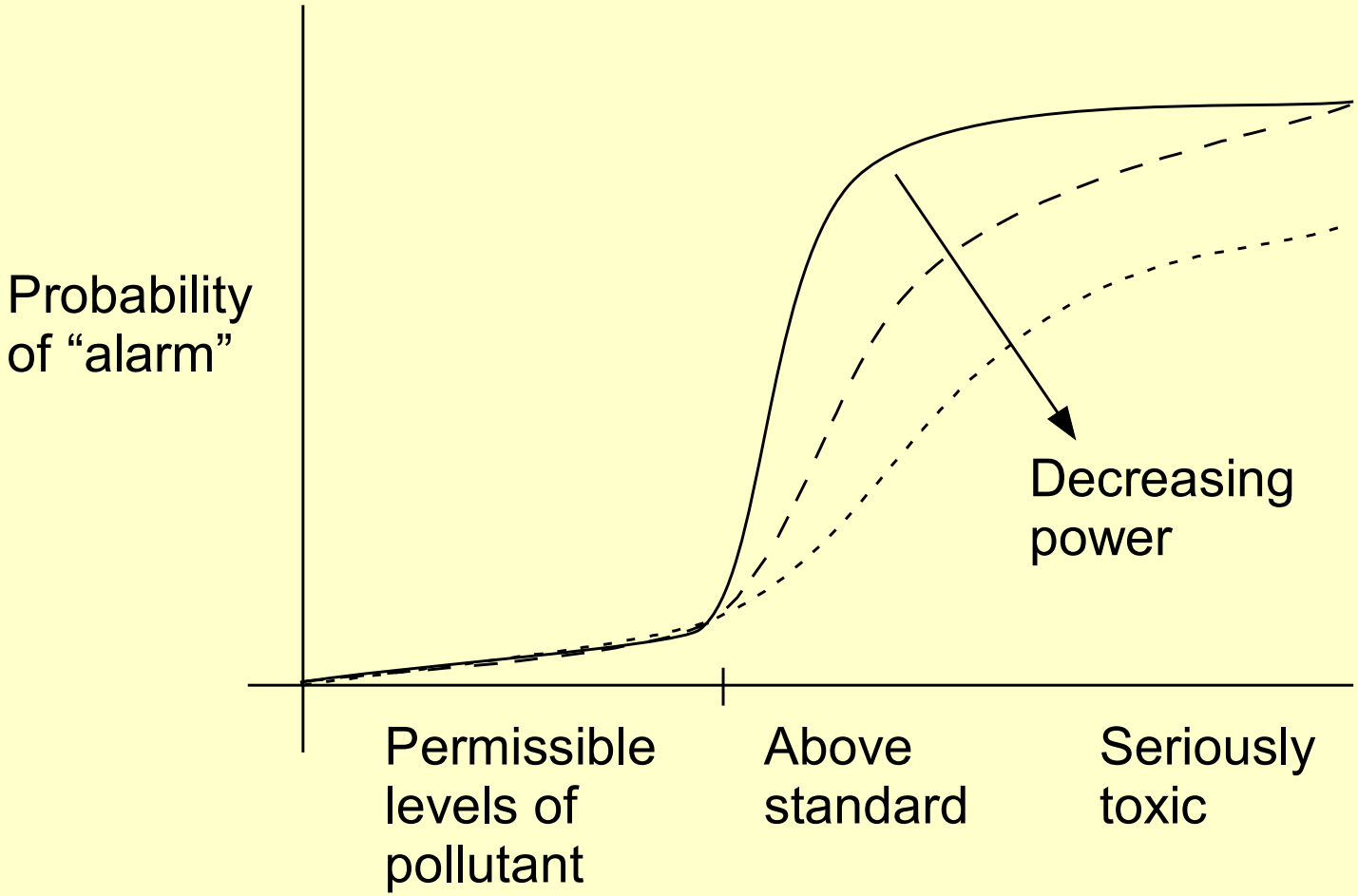
Cartoon guide to statistics

# Example: why we care about power

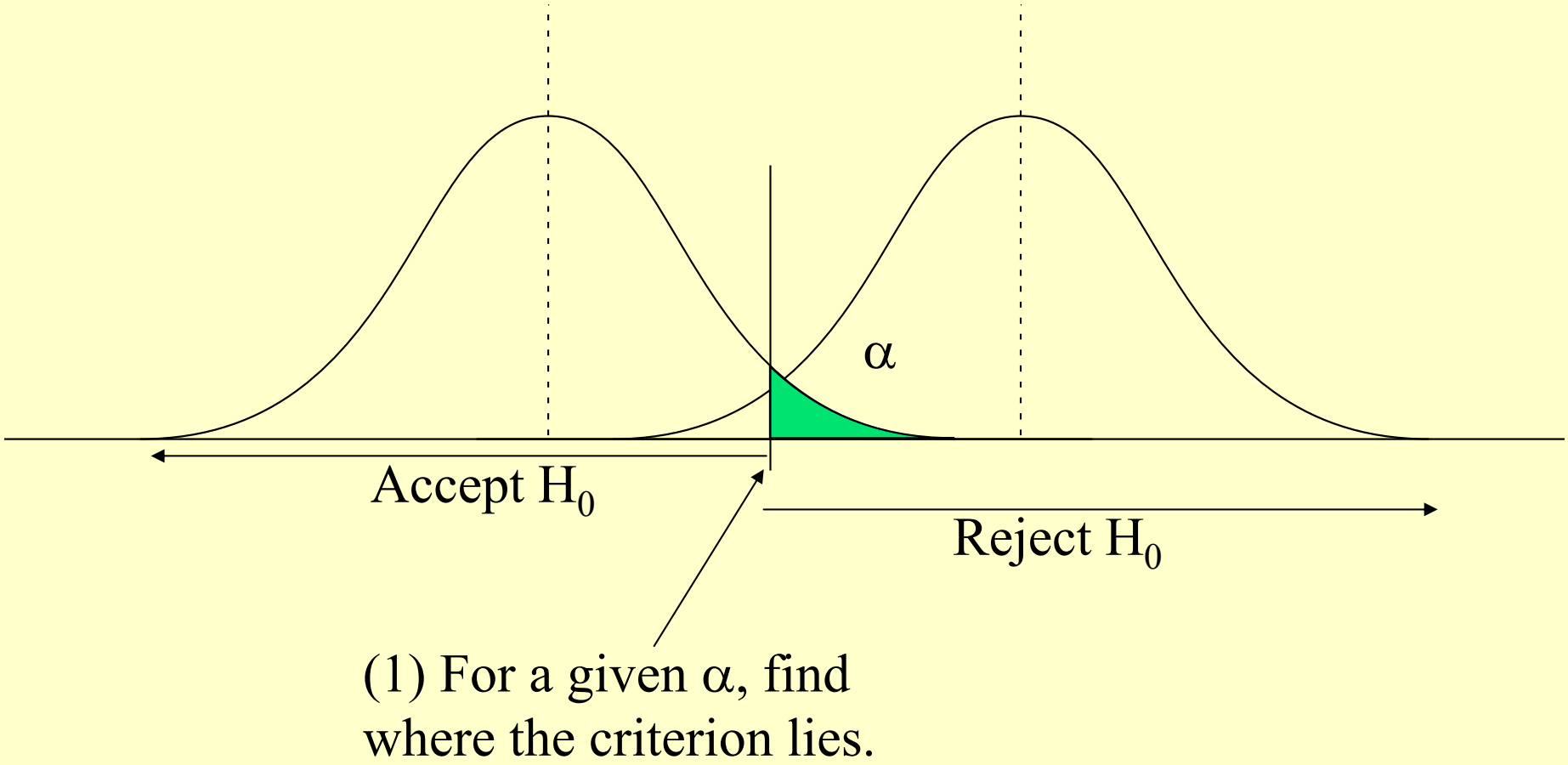
- Just like taking the battery out of the smoke alarm, this test has little chance of setting off an alarm.
- Because of this issue, environmental regulators have moved in the direction of not only requiring tests showing that the pollution is not having significant effects, but also requiring evidence that those tests have a high probability of detecting serious effects of pollution. I.E. they require that the tests have high *power*.

Cartoon guide to statistics

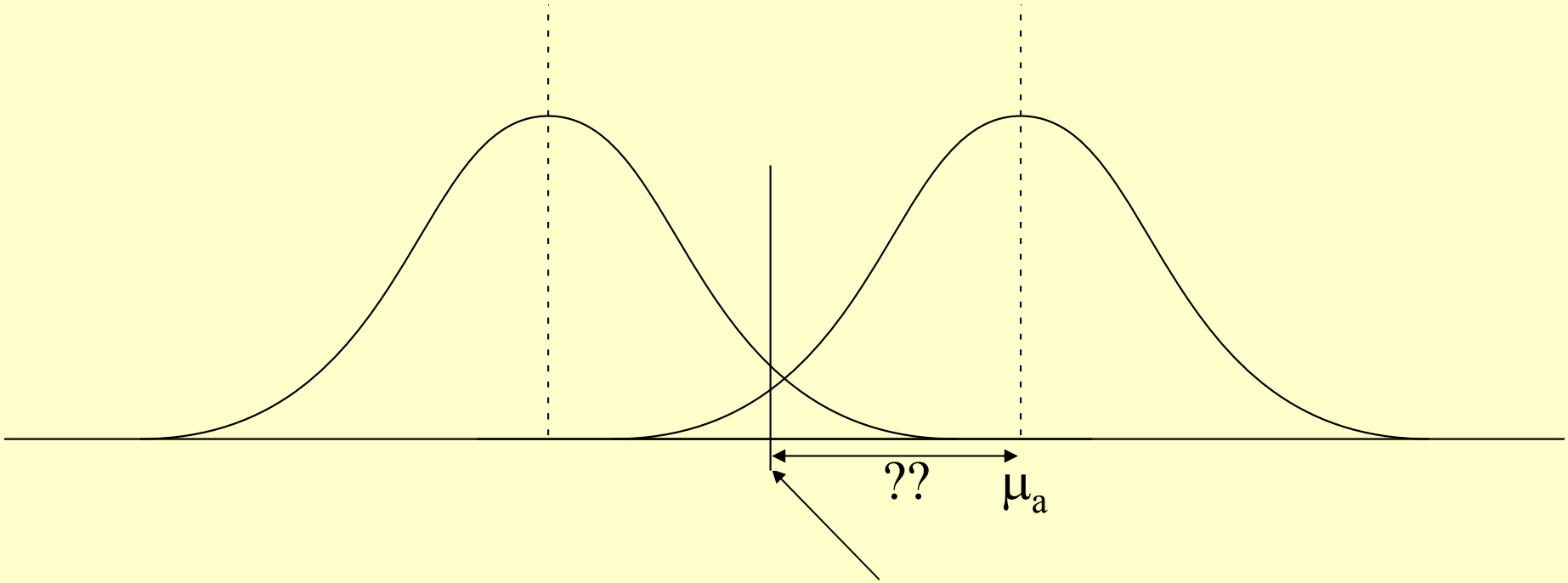
# Power and response curves



# How to compute power (one-sample z-test example)

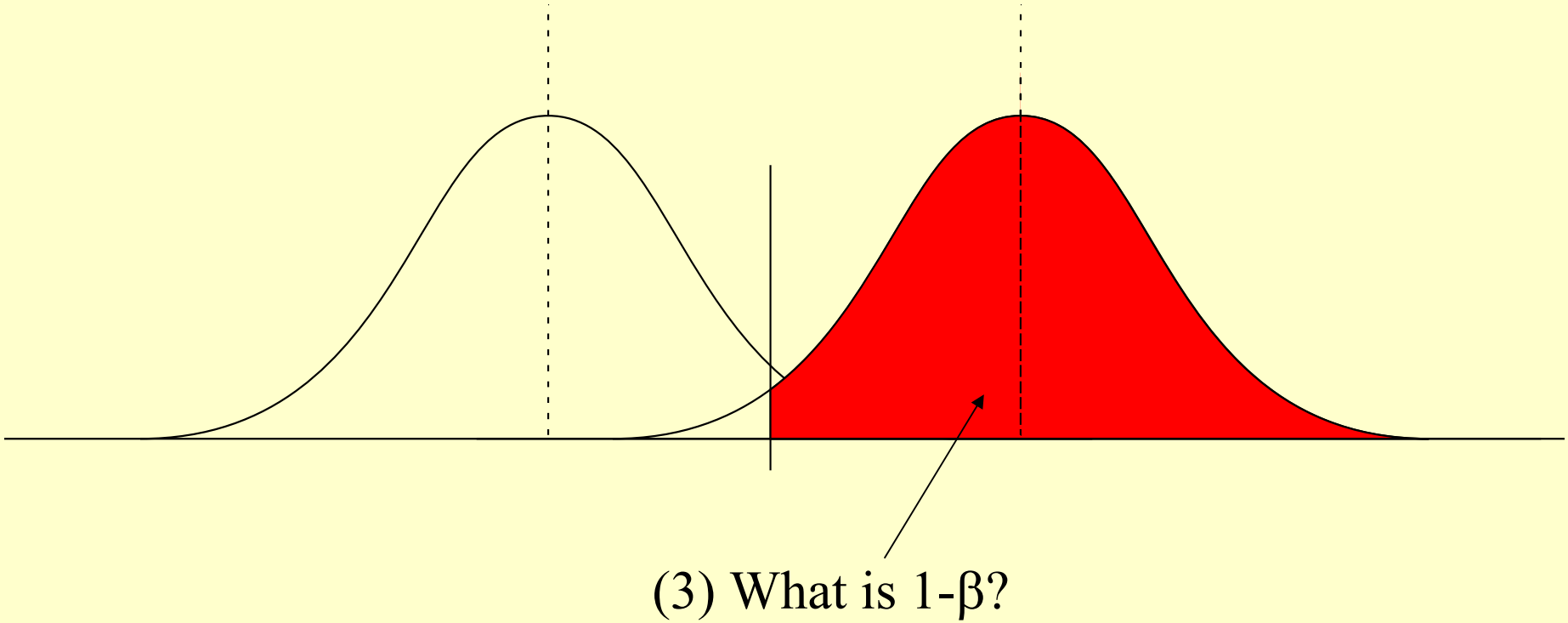


# How to compute power (one-sample z-test example)




(2) How many standard deviations from $\mu_a$ is that criterion? (What's its z-score?)

# How to compute power (one-sample z-test example)



# Computing power: 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?

# How to figure out the power of a z-test

- 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 $\alpha$ to 5%
- Guess the standard deviation is $\sigma=2\%$

# First, find the criterion for rejecting the null hypothesis with $\alpha$=0.05

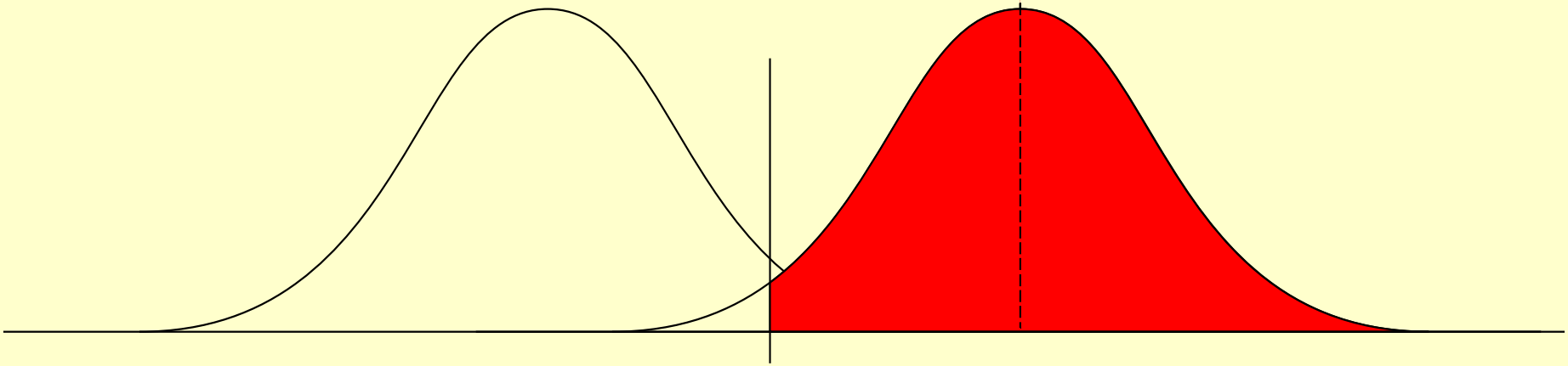
- $H_o$: $\mu$=0%; say n=25 and $\sigma$=2%
- $H_a$: $\mu$>0%

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

  $= (m-0)/(2/5) \geq 1.645$
- 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 +∞
- What is z for m=0.658?

# Step 2

- Assuming $\sigma$ 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%

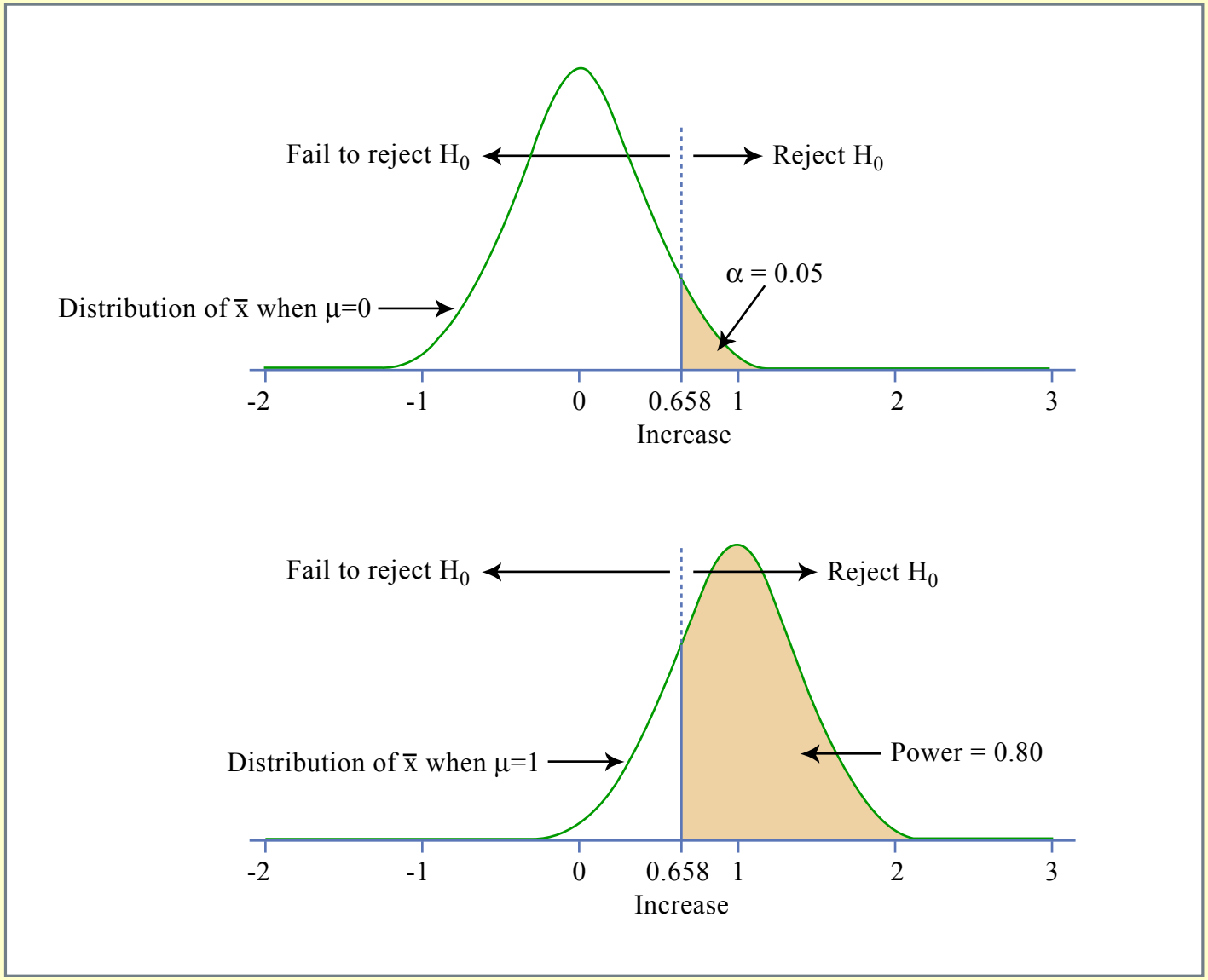
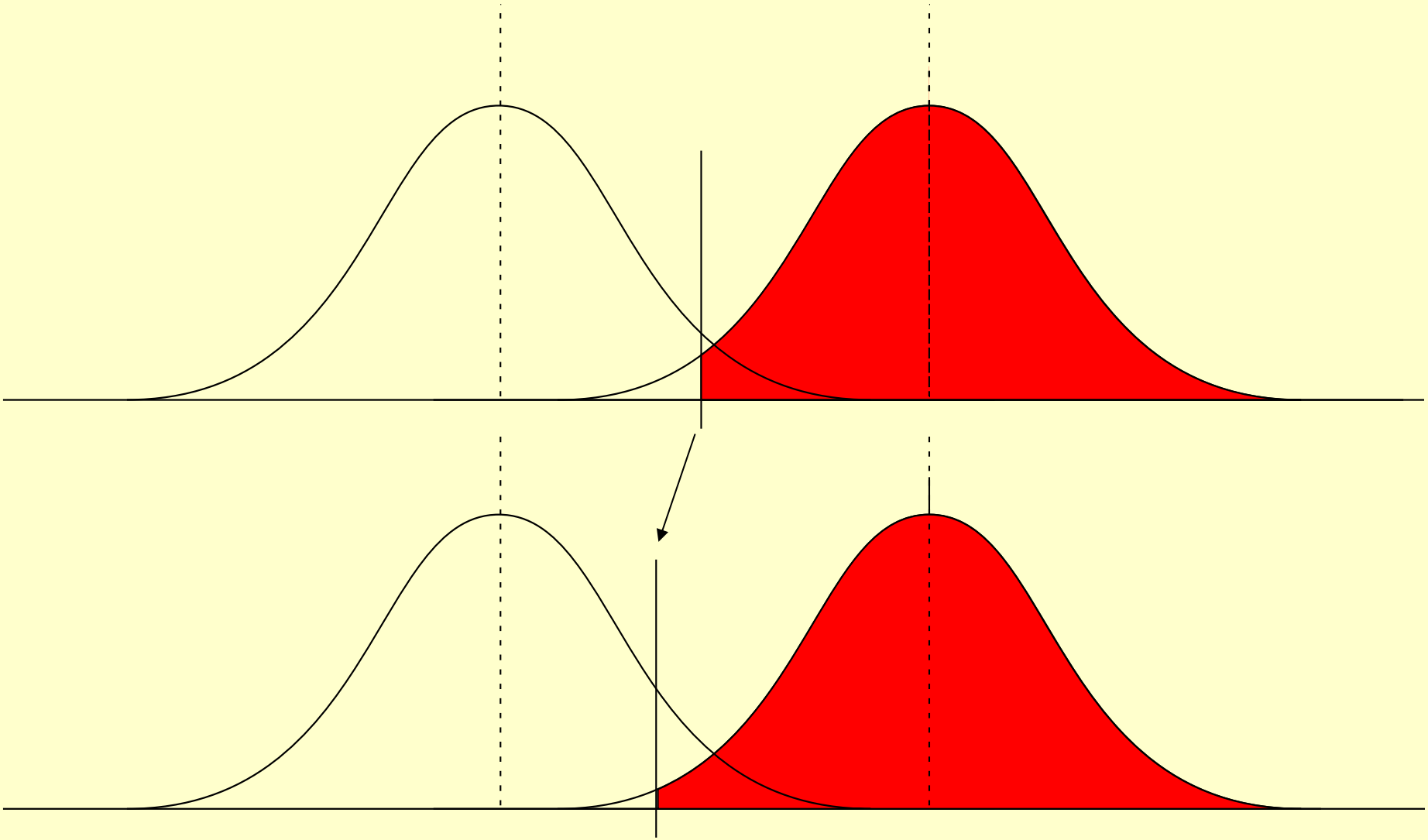



Figure by MIT OCW.

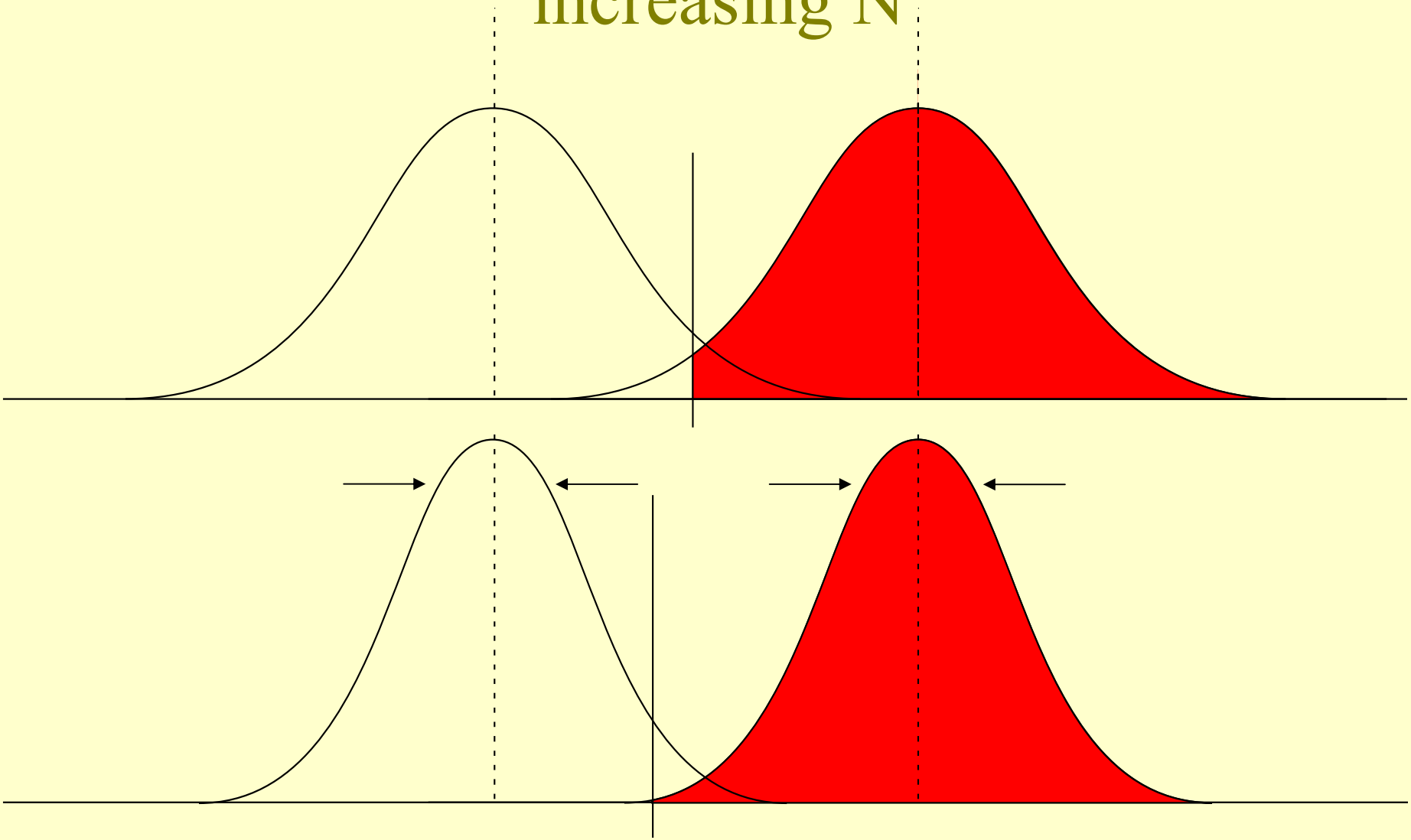
# How to increase power

- Increase $\alpha$
  - 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 $\mu$ in $H_a$ and the in $\mu_o$ in $H_o$.
- Decrease $\sigma$.
- Change to a different kind of statistical test.

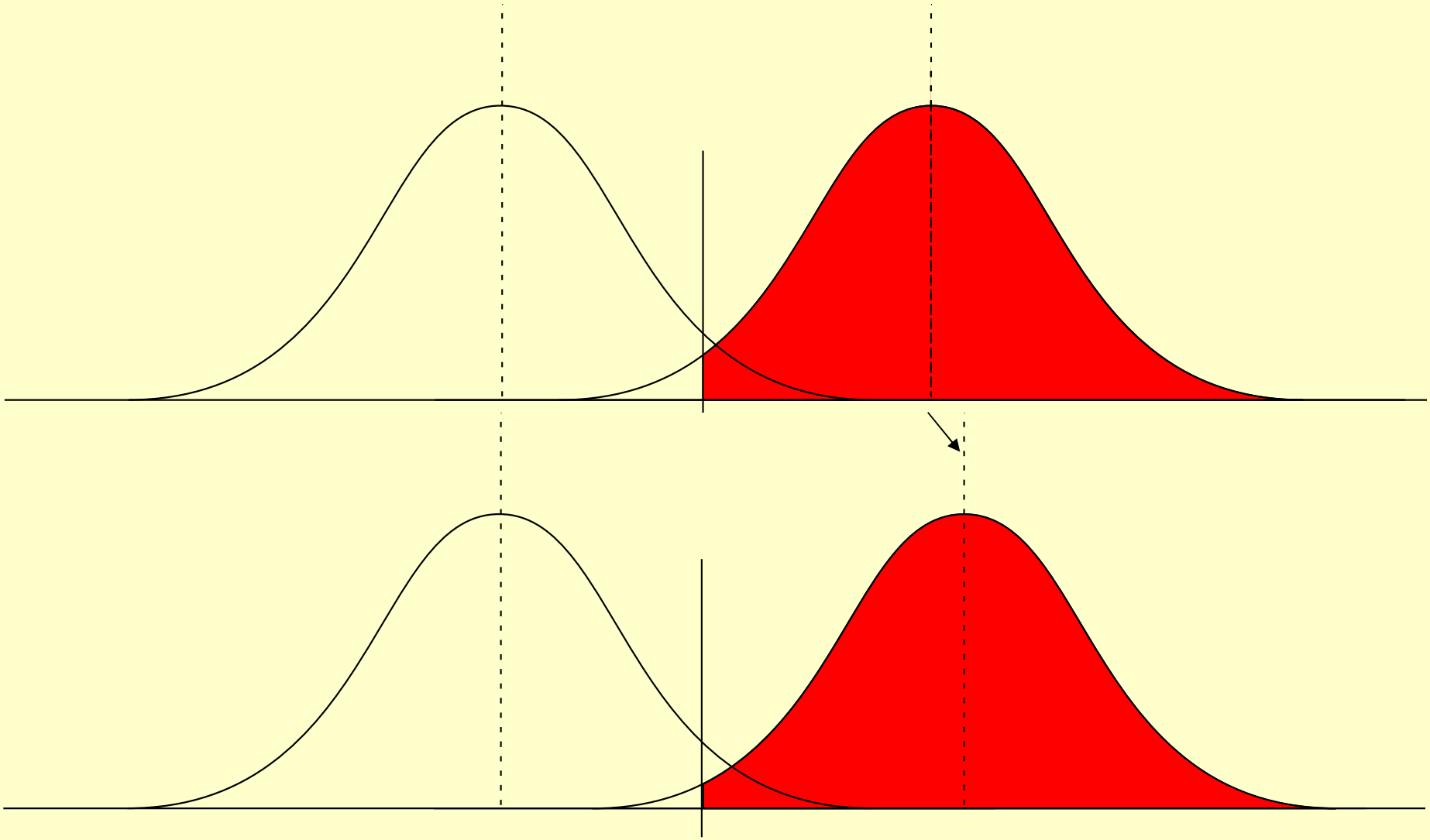
# Increase α



# Reduce SE by either reducing SD, or increasing N



# Increase the difference in means



- OK, on to two-sample hypothesis testing…

# One-sample vs. two-sample hypothesis testing

- One-sample
  - Is our sample different from what is expected (either theoretically, or from what is known about empirically about the population as a whole)
- Two-sample
  - Is sample A different from sample B? E.G. is the mean under condition A significantly different from the mean under condition B?

Most of the time, we end up doing two-sample tests, because we don't often have expectations against which we can compare one sample.

# Example one-sample situations

- Is this a fair coin, given the observed # heads?
  - Compare with theory
- Is performance on this task significantly different from chance?
  - Compare with, e.g., 50%
- Does the gas mileage for this car match the manufacturer's stated mileage of 30 mpg?

# Example two-sample questions

- Does taking a small dose of aspirin every day reduce the risk of heart attack?
  - Compare a group that takes aspirin with one that doesn't.
- Do men and women in the same occupation have different salaries.
  - Compare a sample of men with a sample of women.
- Does fuel A lead to better gas mileage than fuel B?
  - Compare the gas mileage for a fleet of cars when they use fuel A, to when those same cars use fuel B.

# Recall the logic of one-sample tests of means (and proportions)

- $x_i$, i=1:n, are drawn from some distribution with mean $\mu_0$ and standard deviation $\sigma$.
- We measure the sample mean, m, of the $x_i$.
- For large enough sample sizes, n, the sampling distribution of the mean is approximately normal, with mean $\mu_0$ and standard deviation (standard error) $\sigma/\mathrm{sqrt}(n)$.

# Recall the logic of one-sample tests of means (and proportions)

- State the null and alternative hypotheses – these are hypotheses about the sampling distribution of the mean.
  - $H_0: \mu=\mu_0$; $H_a: \mu\neq\mu_0$
- How likely is it that we would have observed a sample mean at least as different from $\mu_0$ as m, if the true mean of the sampling distribution of the mean is $\mu_0$?
- Since the sampling distribution of the mean is approximately normal, it's easy to answer this question using z- or t-tables.

# The logic of hypothesis testing for two independent samples

- The logic of the two-sample t-test (also called the independent-groups t-test) is an extension of the logic of single-sample t-tests
- Hypotheses (for a two-tailed test):

| one-sample | two-sample |
| --- | --- |
| $H_0: \mu=\mu_0$ | $H_0: \mu_1 = \mu_2$ |
| $H_a: \mu\neq\mu_0$ | $H_a: \mu_1 \neq \mu_2$ |

There are versions of $H_0$ where, e.g. $\mu_1=2\mu_2$, but this is rare…

# Two-Sample t-test

- The natural statistic for testing the hypothesis $\mu_1 = \mu_2$ is the difference between the sample means, $m_1 - m_2$
- What is the mean, variance, and shape of the sampling distribution of $m_1 - m_2$, given that $\mu_1 = \mu_2$, and that $m_1$ and $m_2$ are independent means of independent samples?
  - You did this on a homework problem last week. Here's a slightly more general version.

# Mean and variance of the sampling distribution of the difference between two means

- $m_1$ = mean of $n_1$ samples from a distribution with mean $\mu=\mu_1$, standard deviation $\sigma_1$.
- $m_2$ = mean of $n_2$ samples from a distribution with mean $\mu=\mu_2$, standard deviation $\sigma_2$.
- $E(m_1) = 1/n_1\ (n_1 \cdot \mu) = \mu = E(m_2)$
- $var(m_1) = (1/n_1)^2\ (n_1\sigma_1^2) = \sigma_1^2/n_1$
- $var(m_2) = \sigma_2^2/n_2$

# Mean and variance of the sampling distribution of the difference between two means

- $E(m_1) = \mu = E(m_2)$
- $var(m_1) = \sigma_1^2/n_1$
- $var(m_2) = \sigma_2^2/n_2$
- So, what is the mean and variance of $m_1 - m_2$?
- $E(m_1 - m_2) = \mu - \mu = 0$
- $var(m_1 - m_2) = \sigma_1^2/n_1 + \sigma_2^2/n_2$

# Shape of the sampling distribution for the difference between two means

- If the $m_1$ and $m_2$ are normally distributed, so is their difference.
- $m_1$ and $m_2$ are often at least approximately normal, for large enough $n_1$ and $n_2$.
- So, again, use z- or t-tables:
  - How likely we are to observe a value of $m_1 - m_2$ at least as extreme as the one we did observe, if the null hypothesis is true ($H_0$: $\mu_1 - \mu_2 = 0$)?
- This is how we test whether there is a significant difference between two independent means.

# Example: z-test for large samples

- A mathematics test was given to 1000 17-year-old students in 1978, and again to another 1000 students in 1992.
- The mean score in 1978 was 300.4. In 1992 it was 306.7. Is this 6.3 point difference real, or likely just a chance variation?
- $H_0$: $\mu_1 - \mu_2 = 0$
- As before, we compute

$$z_{obt} = (\text{ observed } - \text{ expected }) / \text{ SE}$$

difference difference of the difference

# Example 1: Is there a significant difference between math scores in 1978 vs. 1992?

- Observed difference – expected difference = 6.3 – 0.0 = 6.3
- SE(difference) = $\text{sqrt}(\sigma_1^2/n_1 + \sigma_2^2/n_2)$
- As usual, we don't know $\sigma_1$ or $\sigma_2$, but can estimate them from the data.
- SD(1978) = 34.9; SD(1992) = 30.1
- So, SE(diff) = $\text{sqrt}(34.9^2/1000 + 30.1^2/1000)$
  $\approx 1.5$

# Example 1: Is there a significant difference between math scores in 1978 vs. 1992?

- Observed – expected = 6.3
- SE(diff) ≈ 1.5
- Therefore, $z_{obt} = 6.3/1.5 \approx 4.2$
  - 4.2 SD's from what we expect!
- From the tables at the back of the book, p ≈ .00003. We reject the null hypothesis, and decide that the difference is real.

# Example 2: Test for significant difference in proportions

- Note: as we've said before, the sampling distribution of the proportion is approximately normal, for sufficiently large n
  - $np \geq 10, nq \geq 10$
- So, the distribution of the difference between two proportions should be approximately normal, with mean $(p_1 - p_2)$, and variance $(p_1q_1/n_1 + p_2q_2/n_2)$
- For sufficiently large $n_i$, we can again use the z-test to test for a significant difference.

# Example 2: Is there a significant difference in computer use between men and women?

- A large university takes a survey of 200 male students, and 300 female students, asking if they use a personal computer on a regular basis.
- 107 of the men respond “yes” (53.5%), compared to 132 of the women (44.0%)
- Is this difference real, or a chance variation?
- $H_0$: $p_{men} - p_{women} = 0$

# Example 2: Is there a significant difference in computer use between men and women?

- As before, we need to compute
  $z_{obt}$ = (observed – expected)/SE
- Observed – expected =
  (53.5% - 44.0%) – 0% = 9.5%
- SE(difference) =
  sqrt(0.535·0.465/200 + 0.44·0.56/300)·100
  (SE for males)$^2$ (SE for females)$^2$
  ≈ 4.5%
- So, $z_{obt}$ ≈ 9.5/4.5 ≈ 2.1.
  The difference is significant at the p=0.04 level.

# When can you use the two-sample z-test?

- Two large samples, or $\sigma_1$ and $\sigma_2$ known.
- The two samples are independent
- Difference in mean or proportion
- Sample mean or proportion can be considered to be normally distributed
  - Use z-tests, not t-tests, for tests of proportions. If n is too small for a z-test, it's dicey to assume normality, and you need to look for a different technique.

# When do you not have independent samples (and thus should run a different test)?

- You ask 100 subjects two geography questions: one about France, and the other about Great Britain. You then want to compare scores on the France question to scores on the Great Britain question.
  - These two samples (answer, France, & answer, GB) are not independent – someone getting the France question right may be good at geography, and thus more likely to get the GB question right.

# When do you not have independent samples (and thus should run a different test)?

- You test a number of patients on a traditional treatment, and on a new drug. Does the new drug work better?
  - Some patients might be spontaneously improving, and thus their response to the old and new treatments cannot be considered independent.
- We'll talk next lecture about how to handle situations like this.

# Small sample tests for the difference between two independent means

- For two-sample tests of the difference in mean, things get a little confusing, here, because there are several cases.
- As you might imagine, you can use a t-test instead of a z-test, for small samples.
- Case 1: The sample size is small, and the standard deviations of the populations are *equal*.
- Case 2: The sample size is small, and the standard deviations of the populations are *not equal*.

# Case 1: Sample size is small, standard deviations of the two populations are equal

- This works much like previous examples, except:
  - Use a t-test
  - Need to compute SE a different way
  - Degrees of freedom = $n_1 + n_2 - 2$
- Recall the earlier expression for the standard error of the difference in means:

  $SE(\text{difference}) = \text{sqrt}(\sigma_1^2/n_1 + \sigma_2^2/n_2)$
- If $\sigma_1^2 = \sigma_2^2 = \sigma^2$, this becomes:

  $SE(\text{difference}) = \text{sqrt}(\sigma^2 (1/n_1 + 1/n_2))$

# Case 1: Sample size is small, standard deviations of the two populations are equal

- SE(difference) = sqrt($\sigma^2$ ($1/n_1 + 1/n_2$))
- However, as usual, we don't know $\sigma^2$. But, we do have *two* estimates of it: $s_1^2$ and $s_2^2$.
- We use a *pooled* estimate of $\sigma^2$:

  est. $\sigma^2 = [(n_1 - 1)s_1^2 + (n_2 - 1)s_2^2]/(n_1 + n_2 - 2)$

- This is like an average of estimates $s_1^2$ and $s_2^2$, weighted by their degrees of freedom, $(n_1 - 1)$ and $(n_2 - 1)$

# OK, we're ready for an example

- Two random samples of subjects perform a motor learning task, for which they are given scores.
- Group 1 (5 subjects): rewarded for each correct move.
- Group 2 (7 subjects): punished for each incorrect move.
- Does the kind of motivation matter? Use $\alpha=0.01$.

# Effect of reward on motor learning

- $H_0$: $\mu_1 - \mu_2 = 0$
- $H_a$: $\mu_1 - \mu_2 \neq 0$
- Assume, for now, that the experimenter has reason to believe that the variances of the two populations are equal.
- $m_1 = 18$, $m_2 = 20$
- $s_1^2 = 7.00$, $s_2^2 = 5.83$

# Effect of reward on motor learning

- $n_1 = 5$, $n_2 = 7$
- $m_1 = 18$, $m_2 = 20$
- $s_1^2 = 7.00$, $s_2^2 = 5.83$
- Estimate

  est. $\sigma^2 = [(n_1 - 1)s_1^2 + (n_2 - 1)s_2^2]/(n_1 + n_2 - 2)$

  $= [4 \cdot 7 + 6 \cdot 5.83]/10 \approx 6.3$
- So, SE = sqrt(est. $\sigma^2$ $(1/n_1 + 1/n_2)$)

  = sqrt(6.3 (1/5 + 1/7)) = 1.47

# Effect of reward on motor learning

- SE = 1.47
- Now we calculate $t_{obt}$, and compare with $t_{crit}$:

  $t_{obt} = (diff_{observed} - diff_{expected})/SE$
  $= [(m_1 - m_2) - 0]/SE$
  $= -2/1.47 = -1.36$

  $t_{crit}$, for a two-tailed test, d.f.=10, and $\alpha$=0.01: 3.169
- Comparing $t_{obt}$ to $t_{crit}$, we do not reject the null hypothesis.

# Computing confidence intervals for the difference in mean

- Anytime we do a z- or t-test, we can turn it around and get a confidence interval for the true parameter, in this case the difference in mean, $\mu_1 - \mu_2$
- Usual form for confidence intervals:

  true parameter = observed $\pm$ $t_{crit} \cdot$ SE
- Here, the 99% confidence interval is:

  $\mu_1 - \mu_2 = (m_1 - m_2) \pm 3.169 \cdot 1.47$

  $= -2 \pm 4.66$, or approx from -6.66 to 2.66
- The 99% confidence interval covers $\mu_1 - \mu_2 = 0$, again indicating that we cannot reject this null hypothesis.