Last few slides from last time…

Example 3: What is the probability that p' will fall in a certain range, given p?

- Flip a coin 50 times. If the coin is fair (p=0.5), what is the probability of getting an estimate, p', greater than or equal to 0.7 (=35 heads).
- E(P') = 0.5
- Std. error(P') = sqrt((.5)(.5)/50) = .0707
- z = (0.7-0.5)/0.0707 ≈ 2.83
- P(z > 2.83) ≈ (1-0.9953)/2 = 0.0024
  = P(p' > 0.7)
  - Pretty unlikely to get such high estimates of p

More examples on finding the mean and standard deviation of a r.v.

- $x \sim N(\mu_x, \sigma_x)$, $y \sim N(\mu_y, \sigma_y)$
- $Z = x + 4y + 2$
  - $E(Z) = E(x) + E(4y) + E(2) = \mu_x + 4\mu_y + 2$
  - $Var(Z) = var(x) + var(4y) + var(2)$
    $= \sigma_x^2 + 16\,\sigma_y^2$
- $Z = (2x_1 + 2x_2 - y)/5$
  - $E(Z) = (E(2x) + E(2x) - E(y))/5 = 4/5\ \mu_x - 1/5\ \mu_y$
  - $Var(Z) = var(2x/5) + var(2x/5) + var(y/5)$
    $= 8/25\ \sigma_x^2 + \sigma_y^2/25$

Confidence intervals

9.07

2/26/2004

From last time…

- Sampling theory
  - Start with a large *population*.
  - Take a sample of N units from the population, and compute a statistic on that sample.
    - E.G. statistic = sample mean = estimate of the population mean.
  - We imagine doing this many times, and make deductions about how the *sample* statistic (estimator) is distributed.

The purpose of confidence intervals

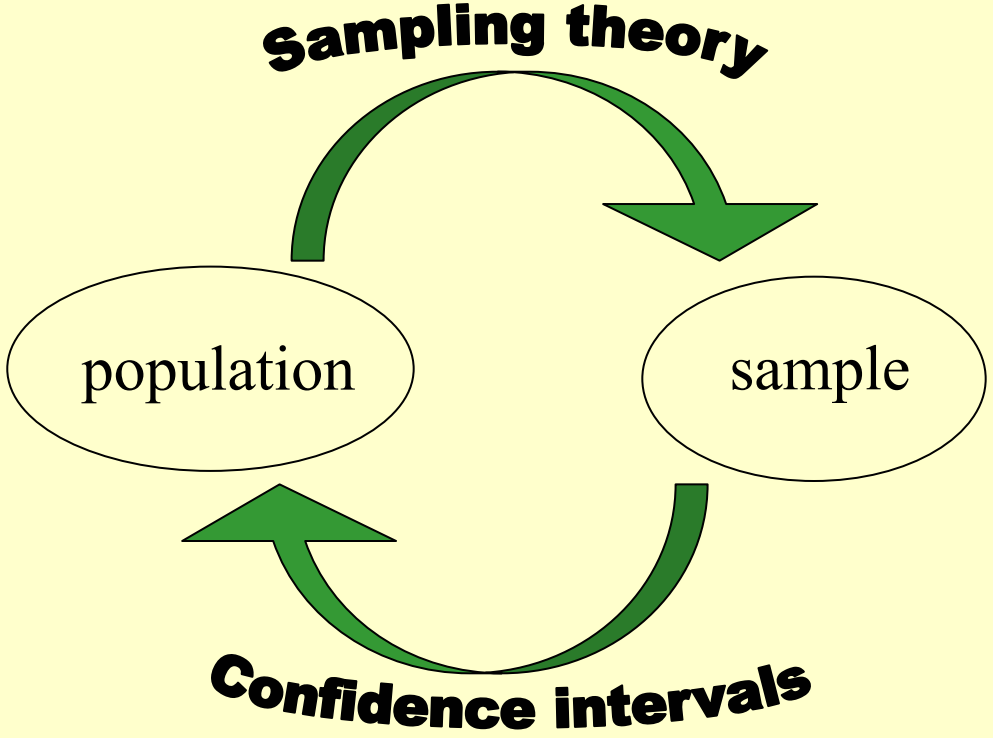
- To estimate an unknown population parameter with an indication of how accurate the estimate is and of how confident we are that our estimate correctly captures the true value of the parameter.

Relationship of confidence intervals to sampling theory

- Confidence intervals
  - Start with one *sample* of N units from a population.
  - We make inferences from this one sample about the parameters of the underlying *population*.

Deduction: Reasoning from a hypothesis to a conclusion.

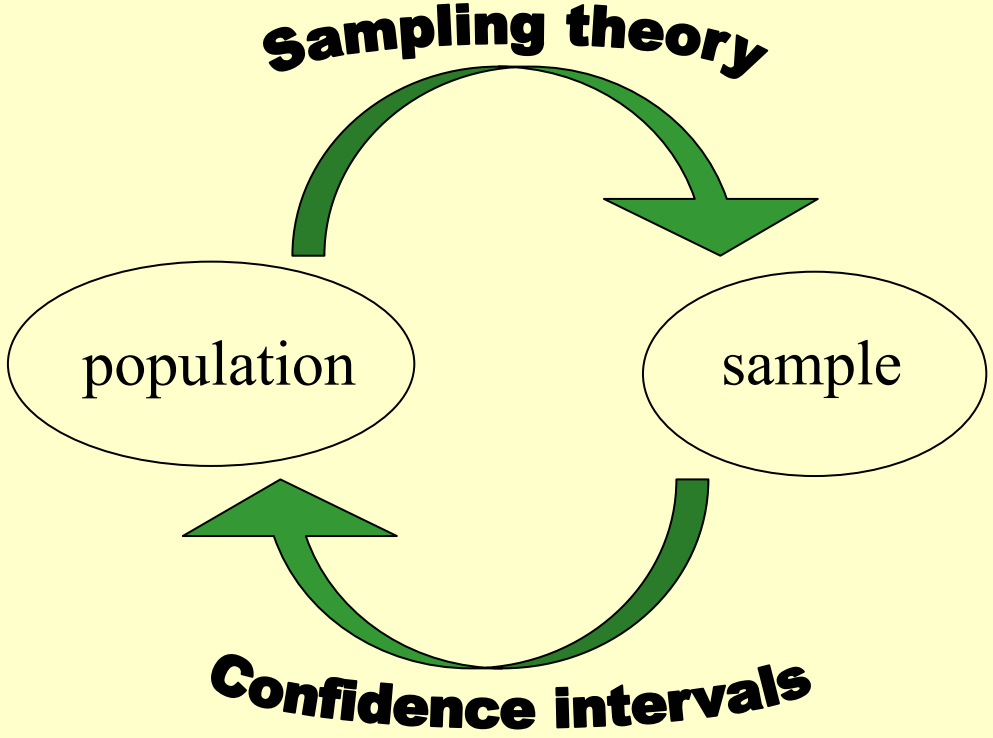
Deduce properties of the *sample statistic* from knowledge of the population.



Induction: Reasoning from a set of observations to a reasonable hypothesis.

Infer properties of the *population* parameter from knowledge of the sample.

Probability that the statistic takes on a particular range of values.

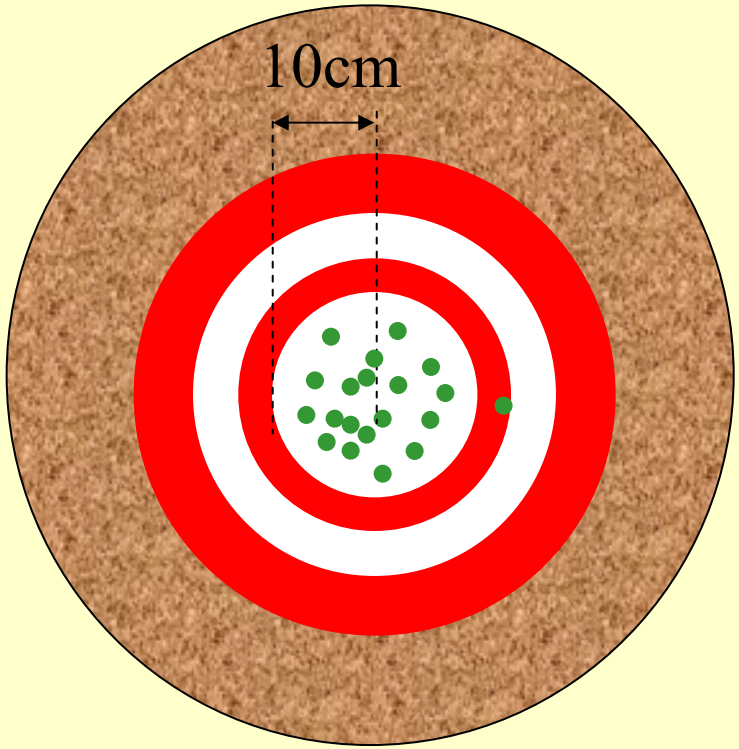


Why confidence vs. probability? The population parameter has a true value, which either is or is not in the given range.

Confidence that the population parameter falls within a particular range of values.

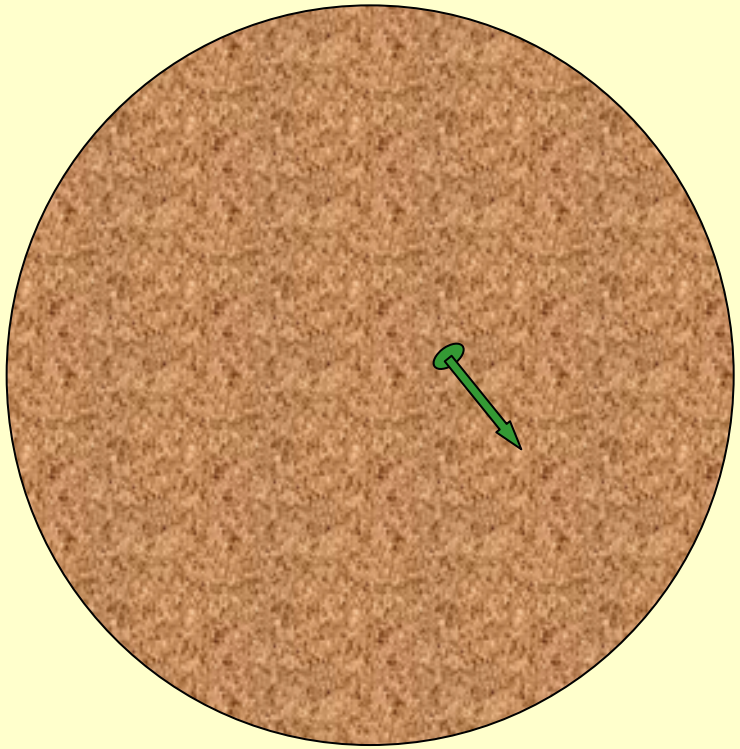
Archery & confidence intervals

- Suppose you have an archer who can hit the 10 cm radius bull's eye 95% of the time.
- One arrow in 20 misses.



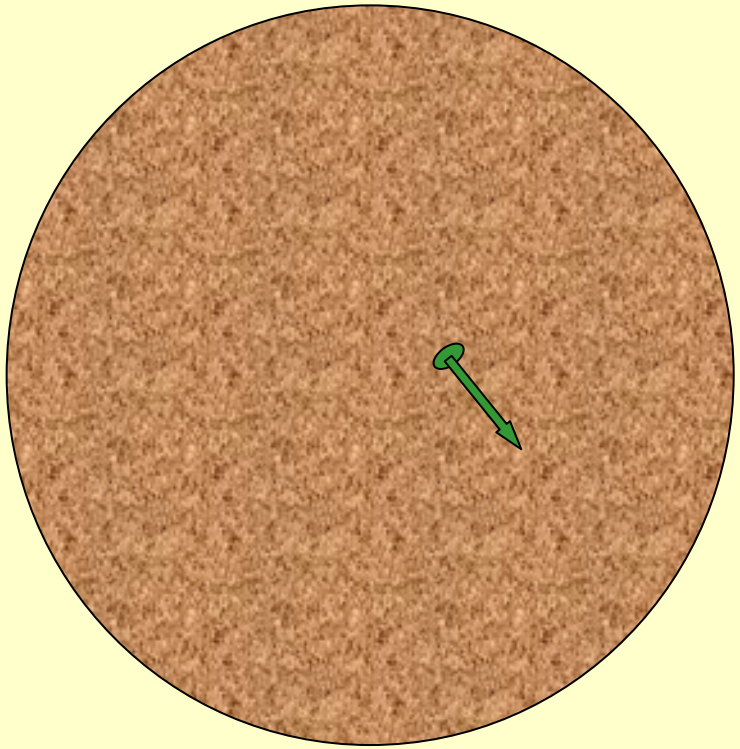
Archery & confidence intervals

- You look at the target from the back. The archer shoots a single arrow.
- Knowing the archer's ability, what can you say about where you expect the center of the bull's eye to be?



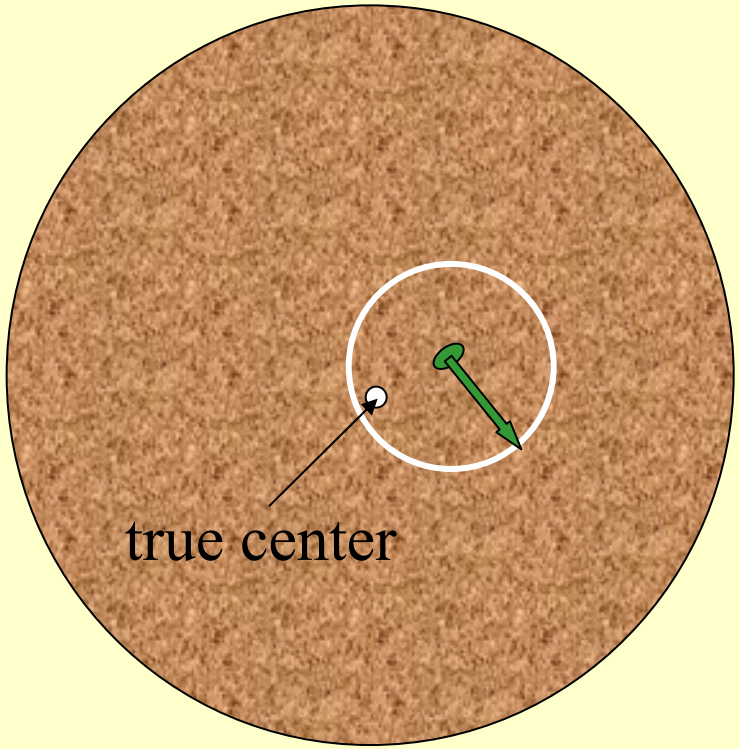
Archery & confidence intervals

- Well, 95% of the time, that arrow is going to be within 10cm of the center of the bull's eye.
- Thus, 95% of the time, the center of the bull's eye is going to be within 10cm of the arrow.



Archery & confidence intervals

- Draw a 10cm circle centered around the arrow.
- If you did this many times, your circles would include ("cover") the center of the target 95% of the time.
- You are 95% *confident* that the center of the target lies in this circle.



Example: A Poll

- A pollster draws a random sample of 1000 voters and asks what they think of a candidate
- 550 voters favor the candidate
- What is the true proportion, p, of the population who favor the candidate?
- From our single sample,
  - n=1000, p'=.55

A Poll

- Recall the sampling distribution of P’
  - The distribution of P’ estimates, given the size of the sample, n, and the true population proportion, p.
  - Approximately normal (especially for n=1000).
  - E(P’) = p

$\sigma_{P'} = \text{sqrt}(p(1-p)/ n)$

A Poll

- From last time: what is the range of proportions such that we expect 95% of our p' estimates to fall within this range?
- From the z-tables, and some algebra:

$0.95 = P(-1.96 \leq z \leq 1.96)$

$0.95 = P(-1.96 \leq (p'-p)/ \sigma_{P'} \leq 1.96)$

$0.95 = P(p-1.96* \sigma_{P'} \leq p' \leq p+1.96* \sigma_{P'})$

95% of the p' "arrows" land between $p-1.96* \sigma_{P'}$ and $p+1.96* \sigma_{P'}$

A Poll

- More algebra:

  $0.95 = P(p-1.96* \sigma_{P'} \leq p' \leq p+1.96* \sigma_{P'})$

  $0.95 = P(p'-1.96* \sigma_{P'} \leq p \leq p'+1.96* \sigma_{P'})$
- This says that 95% of the time, the true value of the population parameter, p, lies within $p'-1.96* \sigma_{P'}$ and $p'+1.96* \sigma_{P'}$
- There's just one catch – we don't know p, so we don't know $\sigma_{P'}$ =sqrt(p(1-p)/n).

A Poll

- We don't know $\sigma_{P'}$.
- So, we fudge a bit, and approximate p by p'. Then $\sigma_{P'} \approx \text{sqrt}(p'(1-p')/n)$
- Back to the poll:
  - $p' = 0.55$, $\sigma_{P'} \approx \text{sqrt}(.55*.45/1000) = .0157$
  - We are 95% confident that p, the true % of the voters that favor the candidate, is within the range $p' \pm 1.96\ \sigma_{P'} = 0.550 \pm (1.96)(0.0157) = .550 \pm 0.031$
- Polls will say, "55% favored the candidate, with a *margin of error* of 3%." (Polls typically use a 95% confidence level, just like we did here.)

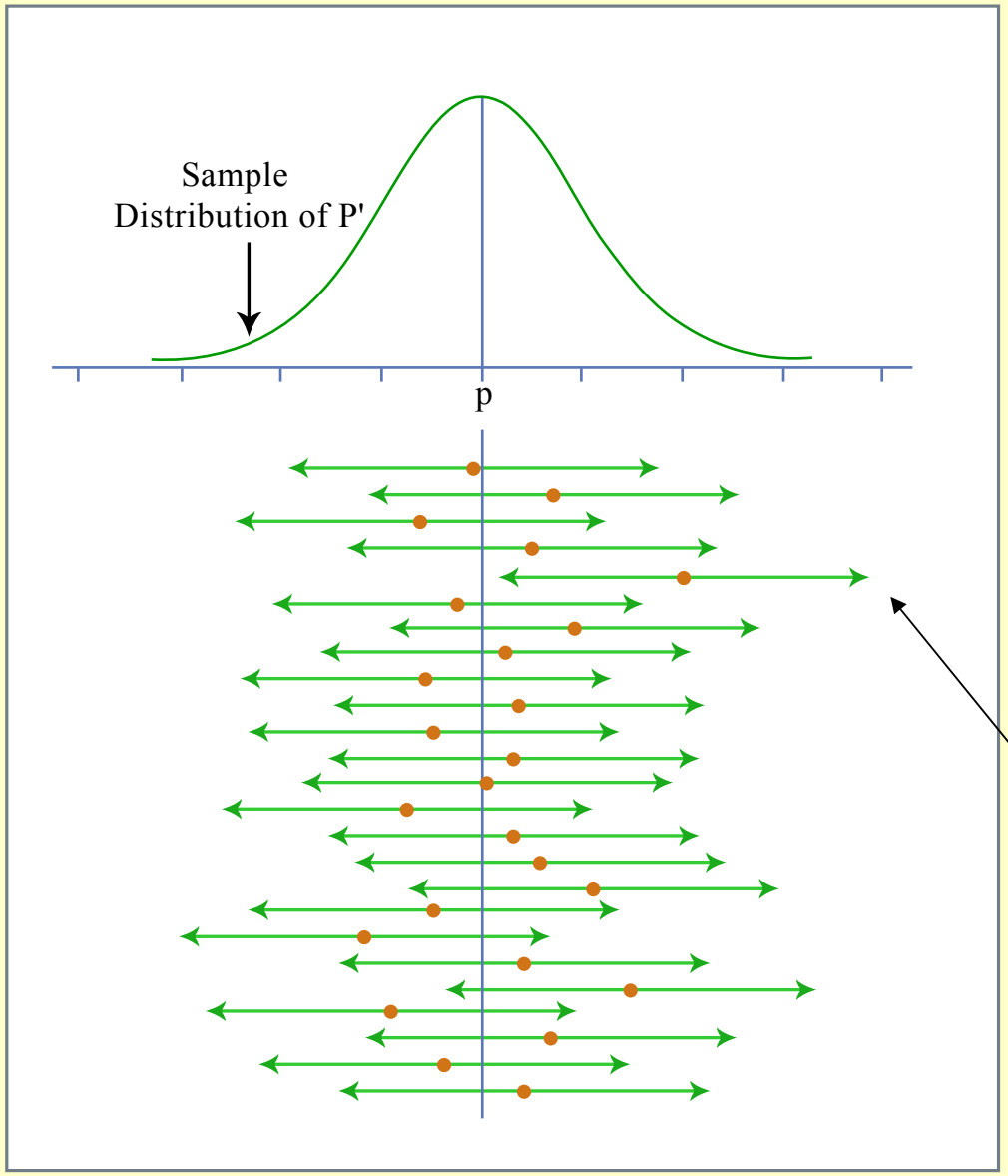


Figure by MIT OCW.

Computer simulation of 20 samples of size n=1000.

Red dots = estimates of p'. Arrows indicate 95% confidence intervals.

All intervals include the actual value of p, except this one. About 5% (1 in 20) of our confidence intervals will not cover the true value of p.

Note a subtle error on that last slide

- The picture of the 20 confidence intervals on the last slide was adapted from a book.
- It actually doesn't quite match the procedure we just followed.
- We *estimated* $\sigma_{P'} \approx \text{sqrt}(p'(1-p')/n)$
- Confidence interval = $\pm$ 1.96 $\sigma_{P'}$
- If we estimate $\sigma_{P'}$ we expect the width of the confidence interval to vary with our estimate of p'.

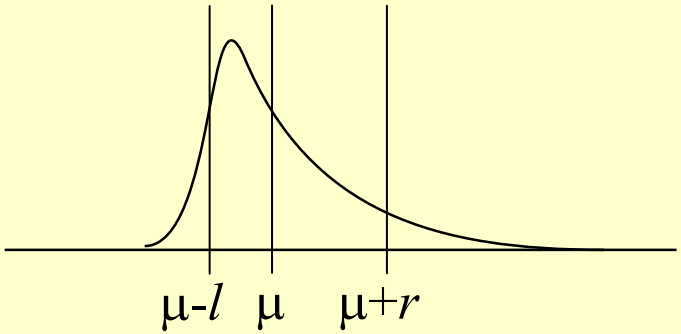
A comment on the symmetry of sampling theory & confidence intervals

- 95% of the p' "arrows" land between $p-1.96\ \sigma_{p'}$ and $p+1.96\ \sigma_{p'}$
- 95% of the time, the true value of the population parameter, p, lies within $p'-1.96\ \sigma_{p'}$ and $p'+1.96\ \sigma_{p'}$

- We get this symmetry because the normal distribution is symmetric about the mean (the parameter we are estimating).

Consider what would happen if parameter estimates had an asymmetric distribution

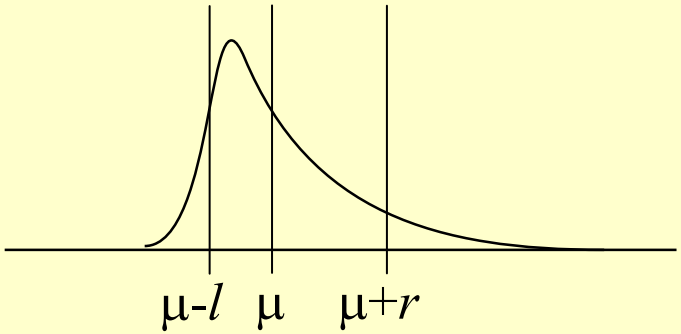
- 95% of estimates of μ fall within an asymmetric region.



- Suppose you observe estimate m.

Consider what would happen if parameter estimates had an asymmetric distribution

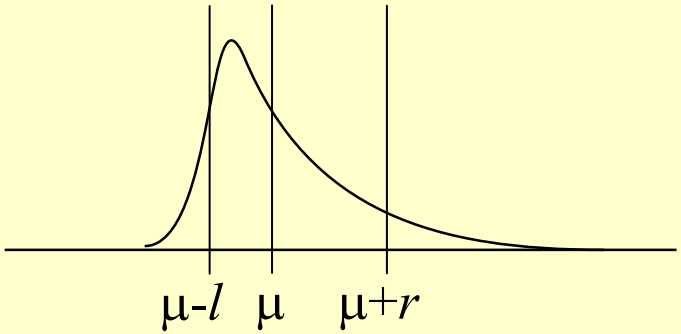
- Suppose you observe estimate m.
- If that estimate is < μ (i.e. to the left), then it must be pretty close to μ.
  - Within *l*.
  - m within *l* of the left of μ, => μ is within *l* of the right of m.



m↔m+*l*

Consider what would happen if parameter estimates had an asymmetric distribution

- Suppose you observe estimate m.
- If that estimate is > μ (i.e. to the right), then m could be farther from μ.
  - Within *r*.
  - m within *r* of the right of μ, => μ is within *r* of the left of m.



m-*r* ⟷ m ↔ m+*l*

A Poll

- Suppose the candidate's campaign manager wants a smaller margin of error and greater confidence
- What can she do?
- There are basically two things she can do:
  - 1. Increase confidence by widening the confidence interval (by going out more SEs)
  - 2. Reduce the margin of error and/or increase confidence by polling more people

Increasing confidence

- By extending the margin of error ± 2SE from p', the pollster was able to arrive at approximately 95% confidence
- Using the standard z table, we can calculate that to get 99% confidence we need to extend the margin of error to ±2.58SE
- In our poll, the 99% confidence interval is $p = .55 \pm .041$
  - With 99% confidence, more than ½ the voters favor the candidate
- In general, as you make the interval wider, you become more confident that the true value falls within your interval.

Reducing the margin of error

- So, we have increased our confidence that our interval straddles the true value of the parameter p to 99%. How do we reduce the margin of error from ±4% to, say, ±1%?
- Sample more people
- $e = z_{crit}\ \mathrm{sqrt}(p'(1-p')/n)$
  - For 99% confidence, $z_{crit} = 2.58$.
    For 95% confidence, $z_{crit} = 1.96$.
- The bigger the value of n, the smaller the e
- What n do we need to get e = 1%?

Aside: Notation

- You will often see the critical value of z, for a certain confidence, C, written $z_{\alpha/2}$, where $\alpha = (1\text{-}C/100)$.

- E.G. for 95% confidence, $z_{0.025} = 1.96$
- For 99% confidence, $z_{0.005} = 2.58$

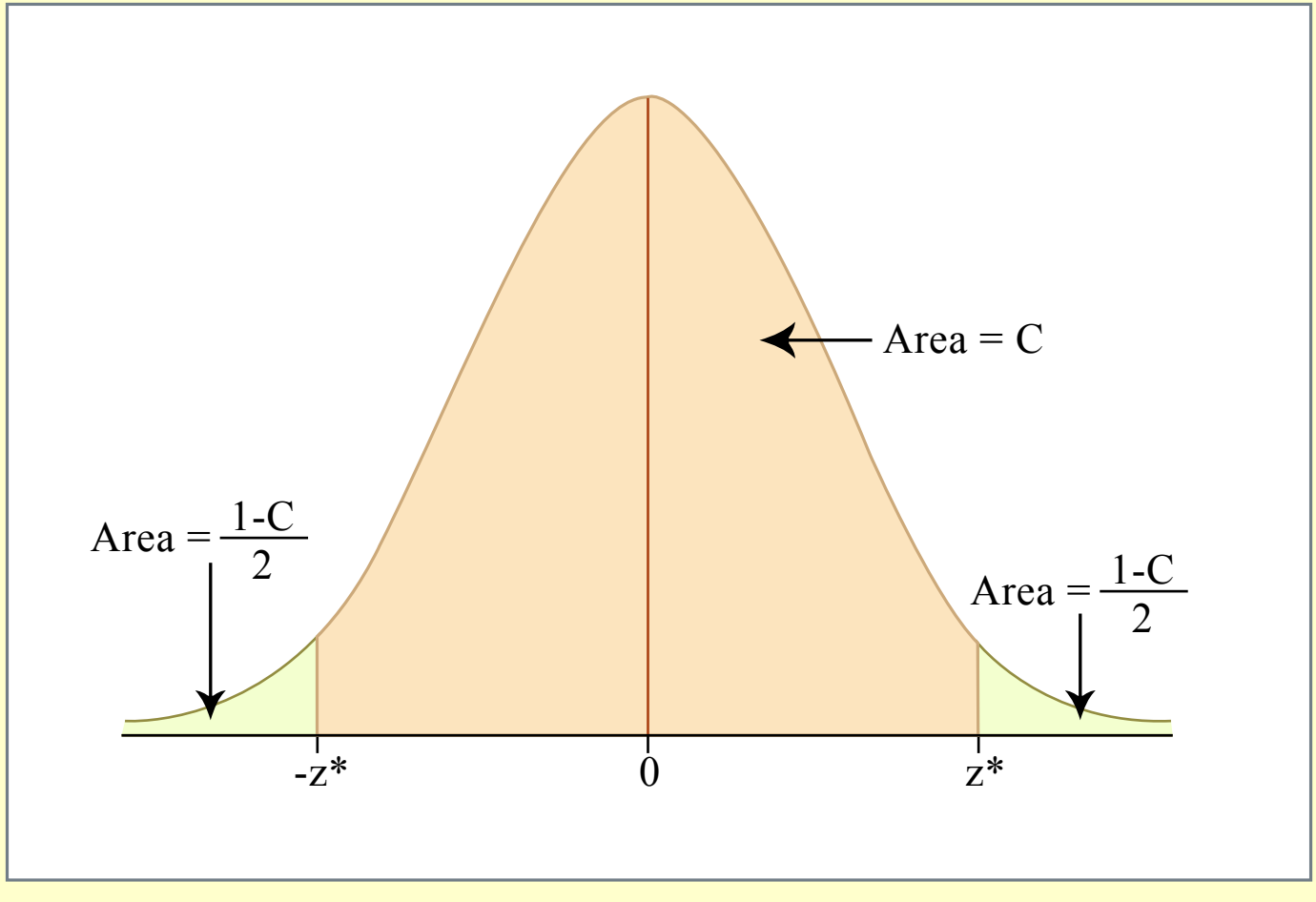


Figure by MIT OCW.

A Poll

- Rearranging
  - $n = [z_{0.005}^2\, p^* (1-p^*)]/e^2$
    - p^* is a guess at the true proportion. This is a calculation we do before we do the survey -- we don't have our sample yet!
    - Guess $p^* = 0.5$
  - $n = (2.58)^2(.5)^2/(.01)^2 = 16{,}641$
- 1000 voters gave a 3% error with 95% confidence. To get 1% error with 99% confidence the candidate needs to sample 16,641 voters!

Polls, and increasing n

- 16,641 voters is a lot more surveys than 1000!
- Often, it's not worth it to increase n.
  - Polls have response bias – voters may not tell the truth when polled, they may not vote, or they may refuse to take part in the poll (and maybe the ones that refuse to take part are more likely to vote for the other guy). Poor people may not have phones and not be counted.
  - These effects can be large, and taking a larger sample doesn't help.
  - Gallup & other polling organizations instead try to reduce bias, by taking into account non-response, screening out eligible voters who are not likely to vote, and so on.

- The margin of error of a confidence interval decreases as
  - The confidence level decreases
  - The sample size n increases
  - The population standard deviation decreases

Confidence Intervals for μ

- Exactly the same logic works for estimating the mean value of a population
- Recall that the distribution of sample means is approximately normal with mean μ and standard error $\sigma/\text{sqrt}(n)$
- Just as before, σ is unknown, so we use s/sqrt(n) as an estimate -- the sample standard error

Confidence Intervals for μ

- So, by analogy, the interval

  $m \pm 1.96\ SE(m) = m \pm 1.96\ s/sqrt(n)$
  covers the true mean, μ, with about .95 probability

  $m \pm z_{\alpha/2}\ SE(m) = m \pm z_{\alpha/2}\ s/sqrt(n)$
  covers the true mean, μ, with about $1-\alpha$ probability

Student's t (again)

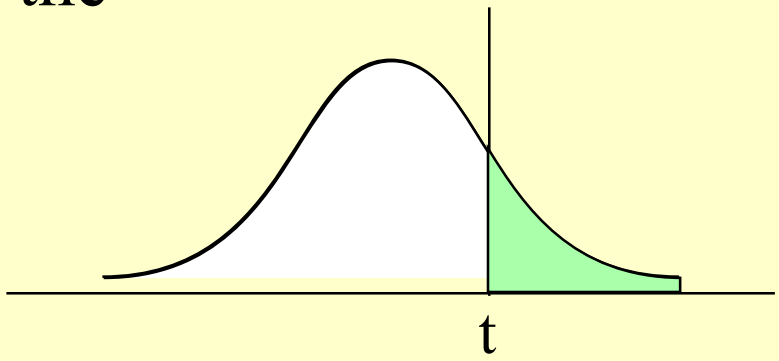
- As we saw in the last lecture, if you don't know σ, you need to approximate it with the sample standard deviation, s.
- For large samples, the approximation is good, and it's as if you know σ.
- For small samples, use a t-distribution instead of a z-distribution.
- Computing confidence intervals is the same as in the z-distribution case, except you look up t_{crit} in the t-distribution table, instead of z_{crit} in the z-table.

Alcohol consumption: t-table example

- A researcher surveys 10 students, asking them how much alcohol they drink per week. The students consume an average of 3.4 units of alcohol/week, with a standard deviation of 4 units of alcohol.
- Assuming this was a simple random sample of MIT students, estimate with 95% confidence the mean alcohol/week consumed by an MIT student.

Alcohol consumption: t-table example

- We want an answer of the form
  mean ± margin of error
- The t-tables at the back of the book give the area under just one tail.
- If we want the area between the tails to be 95%, we look up
  α = area in one tail
  = (1-.95)/2 = 0.025 = 2.5%
- $t_{.025}(n-1) = t_{.025}(9) = 2.26$



Alcohol consumption: t-table example

- $P(-2.26 \leq t \leq 2.26) = 0.95$
  $= P(-2.26 \leq (m-3.4)/(s/\text{sqrt}(N)) \leq 2.26)$
  $= P(-2.26 \leq (m-3.4)/(4/\text{sqrt}(10)) \leq 2.26$
- So, with 95% confidence, the actual mean alcohol consumption is

  3.4 ± 2.26 ($4/\text{sqrt}(10) \approx 3.4 \pm 2.9$ units

Making inferences about the population mean

- You survey 25 students taking a class with 3-0-9 credit units. For such a class, they are supposed to spend about 9 hours in homework and preparation for the class per week. The students spend on average 11 hours studying for this class, with a standard deviation of 4 hours.
- We want to know if the average number of hours spent studying for this class could still be 9 hours, or if this survey is evidence that the students are spending too much time on the class, and the class should be reclassified.

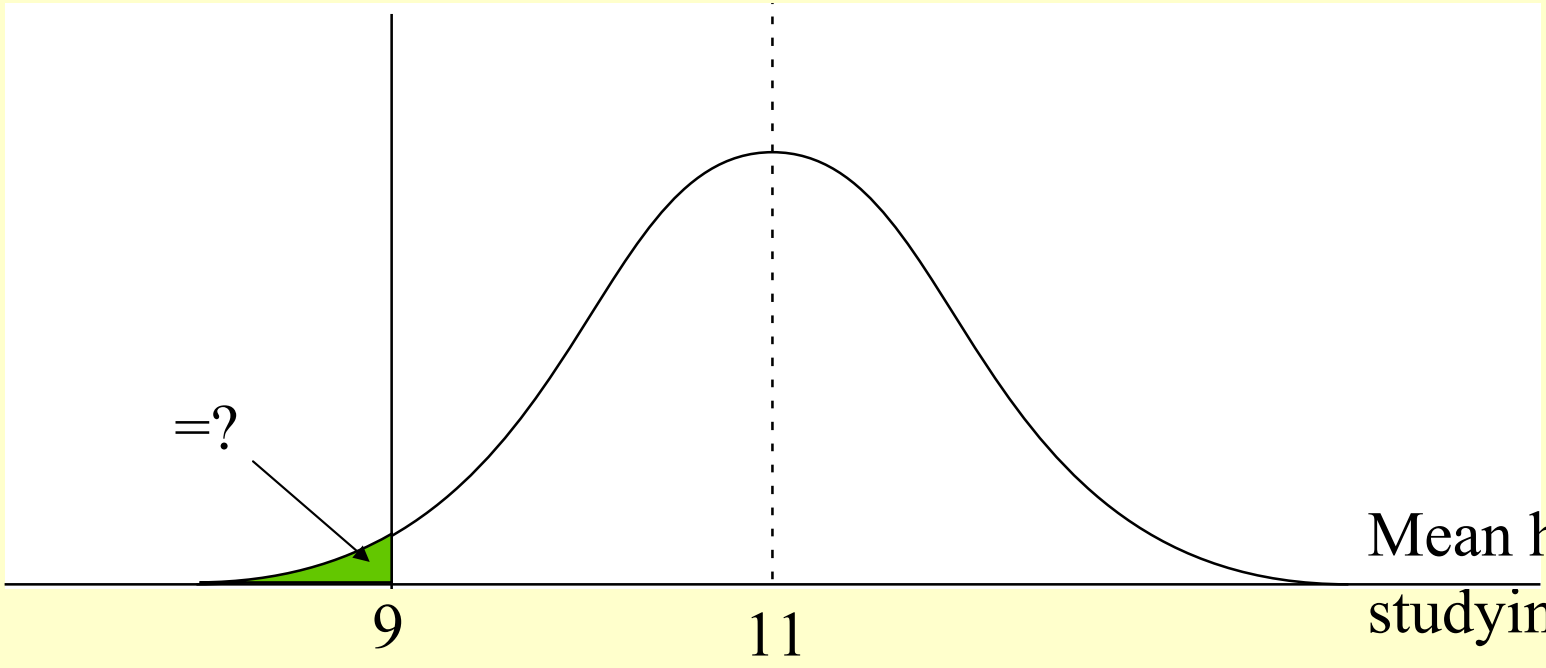
Note that this is sort of the reverse of the problem from last class.

t-table vs. z-table

- We'll do this problem both ways (N=25 is on the border of what's acceptable for the z-table).
- In research, you'll almost always see t-tables used when estimating the mean, rather than trying to decide if N is big enough.

Inferences about the mean

- We want to know: what is the probability that the true average # hours spent studying is 9 or less, given that in our sample of 25 students, the average is 11 hours.



Inferences about the mean: z-table

- 9 hrs is how many std. errors away from the mean of 11 hrs?
- Assume the standard deviation of the population equals the standard deviation of the data.
- Std. error of the mean = std. deviation of the data/sqrt(N) = 4/sqrt(25) = 4/5
- $z = (9-11)/(4/5) = -2.5$
- $p(z<-2.5) = (1-0.9876)/2 = 0.0062$

Inferences about the mean: z-table

- $p(z<-2.5) = (1-0.9876)/2 = 0.0062$
- Put another way, we are >99% confident that the true mean hrs spent studying is >9 hrs.
- Probably this class is too much work for 3-0-9 credit hours.

Inferences about the mean: t-table

- 9 hrs is how many std. errors away from the mean of 11 hrs?
- s/sqrt(N) = 4/sqrt(25) = 4/5
- t = (9-11)/(4/5) = -2.5
- Degrees of freedom = 24
- From our t-table, this roughly corresponds to a one-tailed area of 1%
- $p(t<-2.5) \approx 0.001$

Inferences about the mean: t-table

- $p(t<-2.5) \approx 0.001$
- We are about 99% confident that the mean is greater than 9 hrs.
- This is slightly less confident than we were with the z-test, because the t-test takes into account that we don't really know the standard deviation of the population.
- It still seems highly unlikely that the true mean hrs spent studying is 9 hrs or fewer. Probably this class is too much work for 3-0-9 credit hours.

Another example of confidence intervals for proportions

- In an experiment, a subject is asked to discriminate which of two displays contains a target shape. On 50% of the trials the first display contains the target. On the remaining 50% of the trials, the target is in the second display.
- The subject does this task 50 times, and gets the answer right 65% of the time.
- What is the 99% confidence interval for the subject's performance? Does the interval contain 50% correct (chance performance)?

Percent correct on an experiment

- $p' = 0.65 = E(P')$
- Standard error = $\sigma_{P'} = \text{sqrt}(p'(1-p')/n) = \text{sqrt}(0.65 \cdot 0.35/50) \approx 0.0675$
- From our z-tables, $z_{crit} \approx 2.58$
- $P(0.65 - z_{crit}\ \sigma_{P'} \leq p \leq 0.65 + z_{crit}\ \sigma_{P'}) = 0.99$
- So, the 99% confidence interval is approximately

  $0.48 \leq p \leq 0.82$

Percent correct on an experiment

- With 99% confidence, $0.48 \leq p \leq 0.82$
- This range does include the 50% correct point (p=0.50). So, you *cannot* say, with 99% confidence, that you expect the subject to consistently perform better than chance on this task.
- What about the 95% confidence limits?
- $z_{crit} = 1.96$
- $0.65 - z_{crit}\, \sigma_{P'} \leq p \leq 0.65 + z_{crit}\, \sigma_{P'}$
- $0.52 \leq p \leq 0.78$ You are 95% confident that the subject's performance indicates that their abilities on the task are better than chance performance.

95% confidence limits on performance of the experimental task

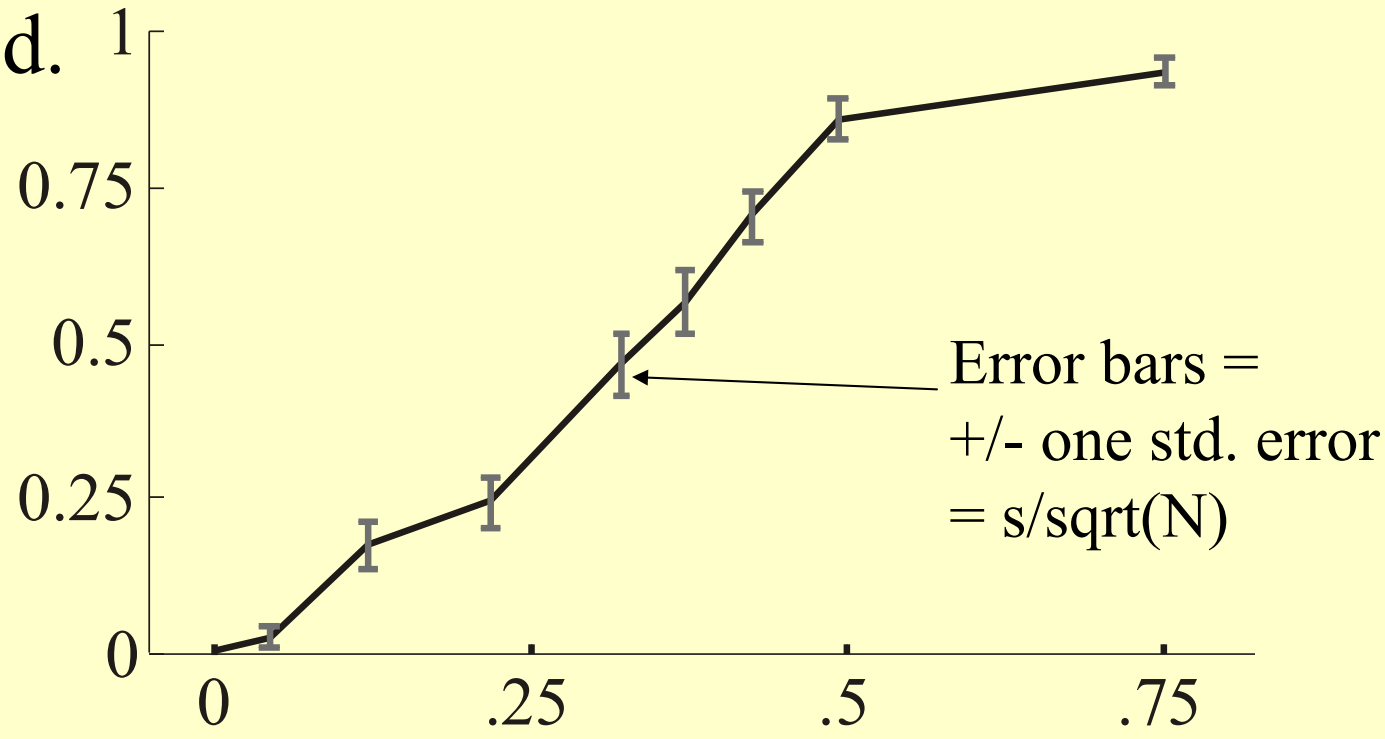
- $z_{crit} = 1.96$
- $0.65 - z_{crit}\,\sigma_{P'} \leq p \leq 0.65 + z_{crit}\,\sigma_{P'}$
- $0.52 \leq p \leq 0.78$
- You are 95% confident that the subject's performance indicates that their abilities on the task are better than chance performance.

Again, let's review the basic form for confidence intervals (for means and proportions)

- $m \pm 1.96\ SE(m)$
  covers the true mean, μ, with about .95 probability
- $m \pm z_{\alpha/2}\ SE(m) = m \pm z_{\alpha/2}\ s/sqrt(n)$
  covers the true mean, μ, with about $1-\alpha$ probability
  - $z_{\alpha/2}$ -> $t_{\alpha/2}$, for small samples
- $p' \pm 1.96\ SE(P')$
  covers the true proportion, p, with about .95 probability
- $p' \pm z_{\alpha/2}\ SE(P') = p' \pm z_{\alpha/2}\ sqrt(pq/n)$
  covers the true proportion, p, with about $1-\alpha$ probability

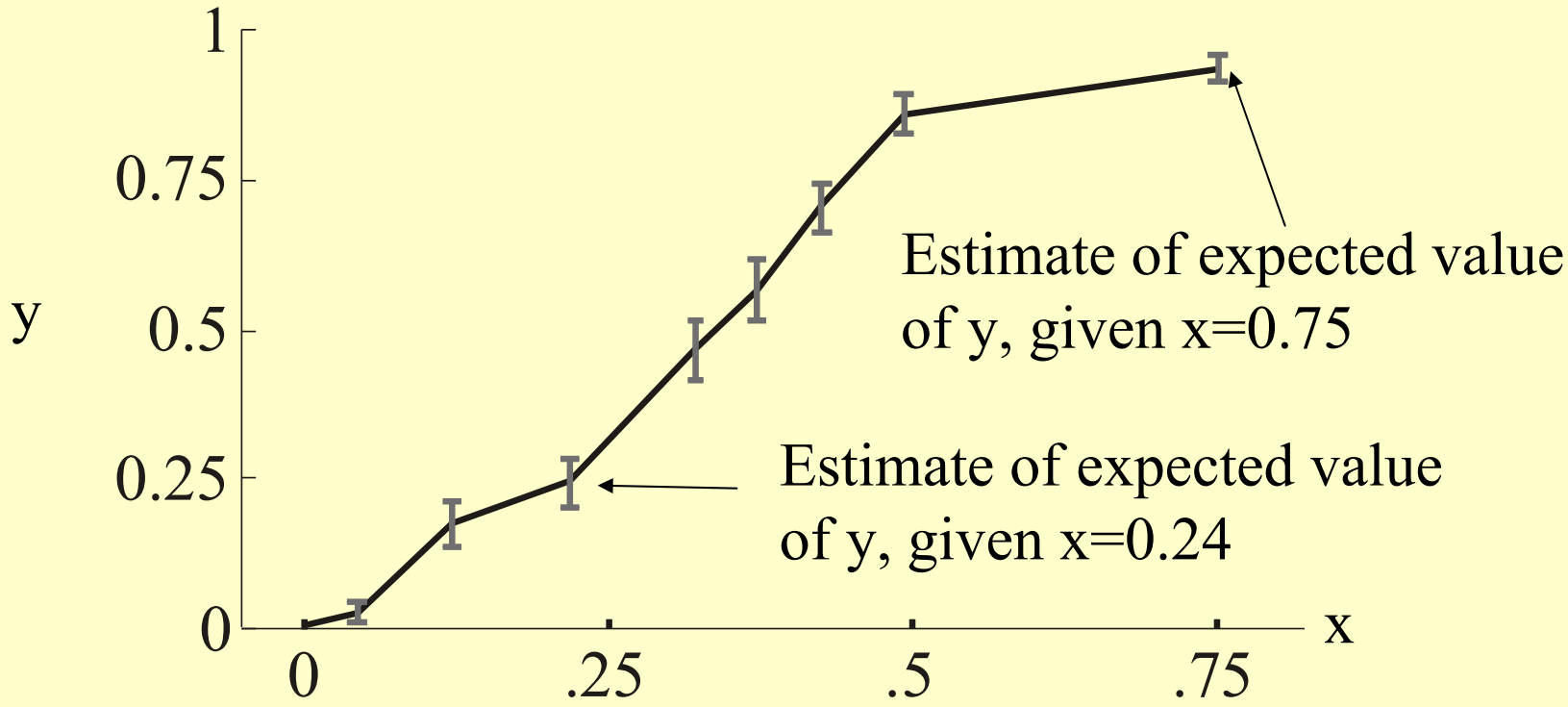
Standard error bars and plots

- You may have heard of standard error before this class, and you've certainly seen it used.



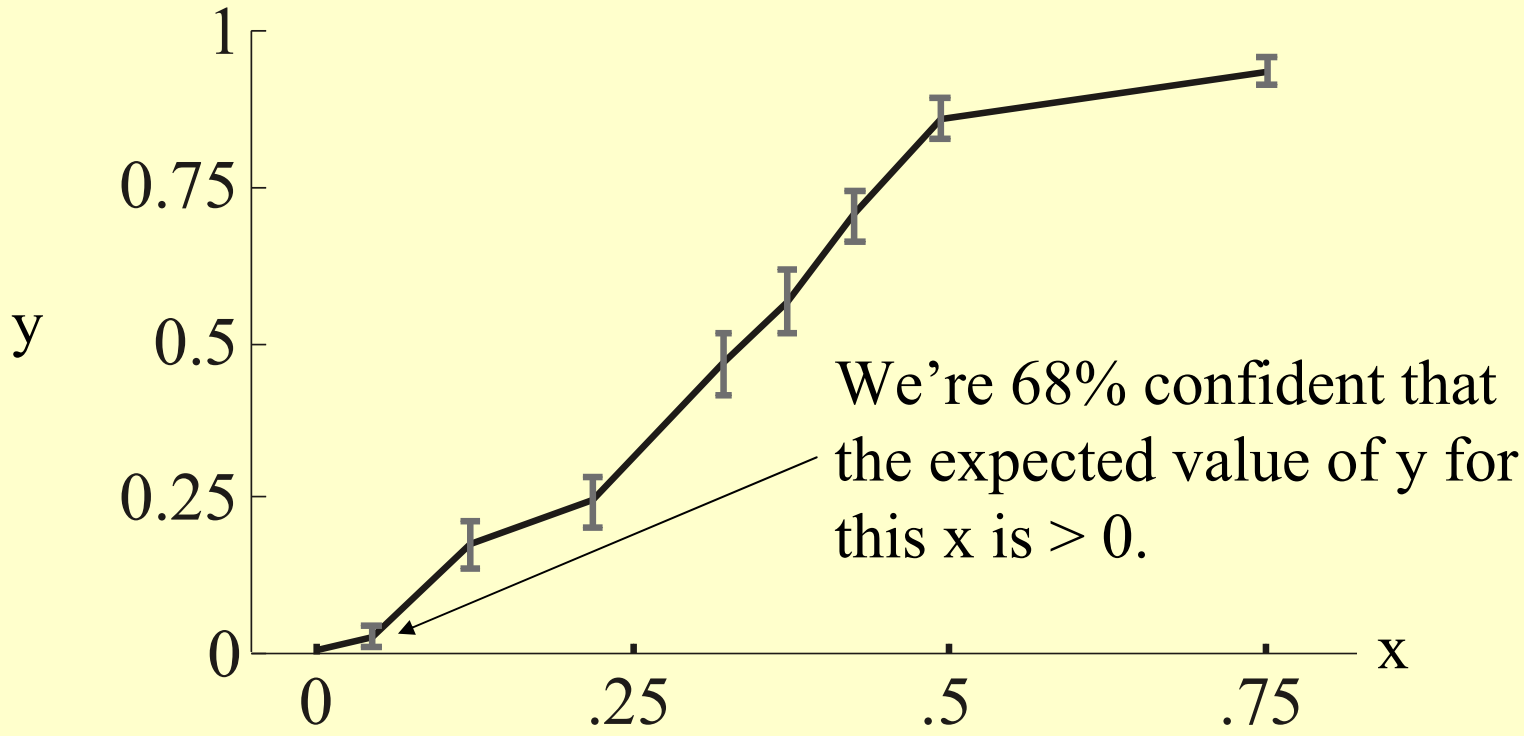
Some notes on error bars

- Think of each data point as being an estimate of some parameter of the population.



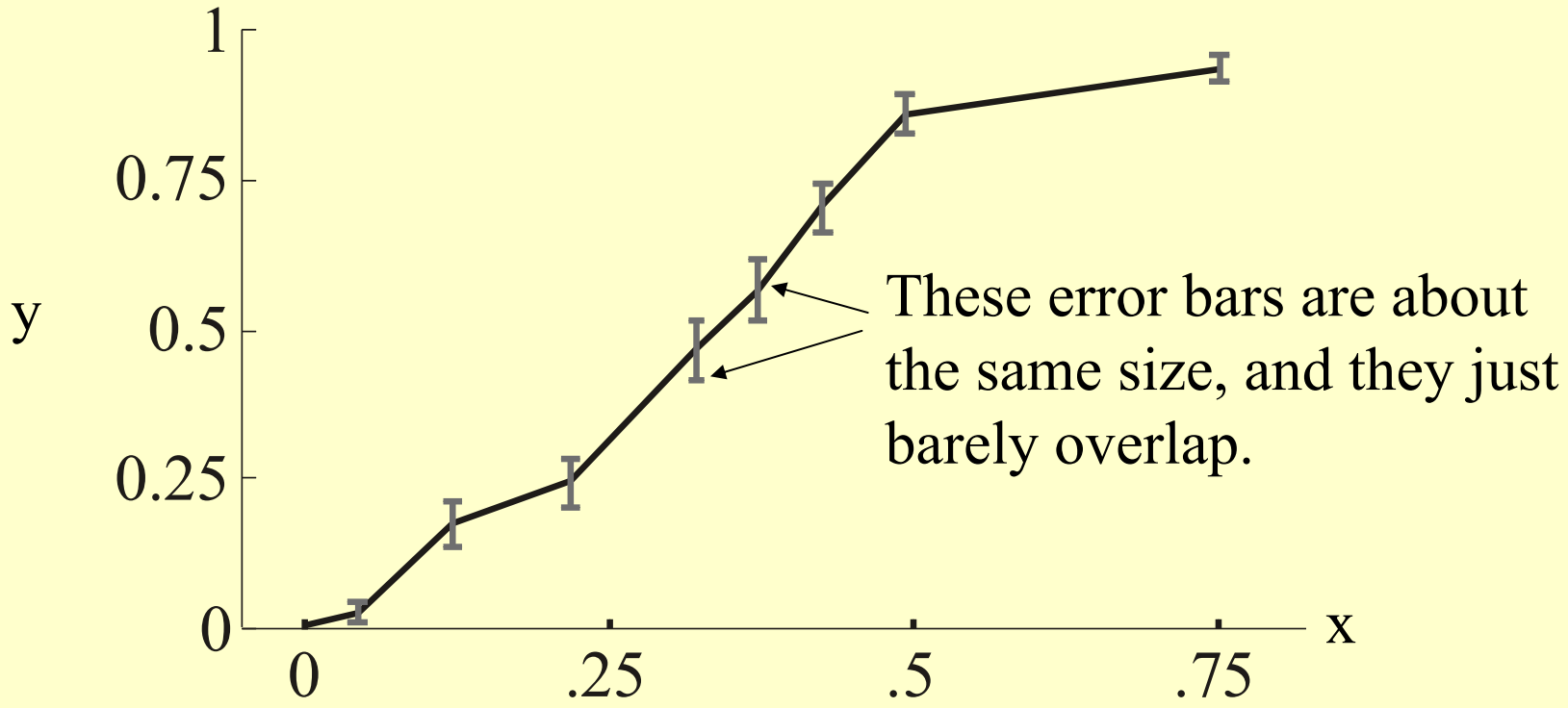
Some notes on error bars

- Then, we can make some inferences, just based on the graph.



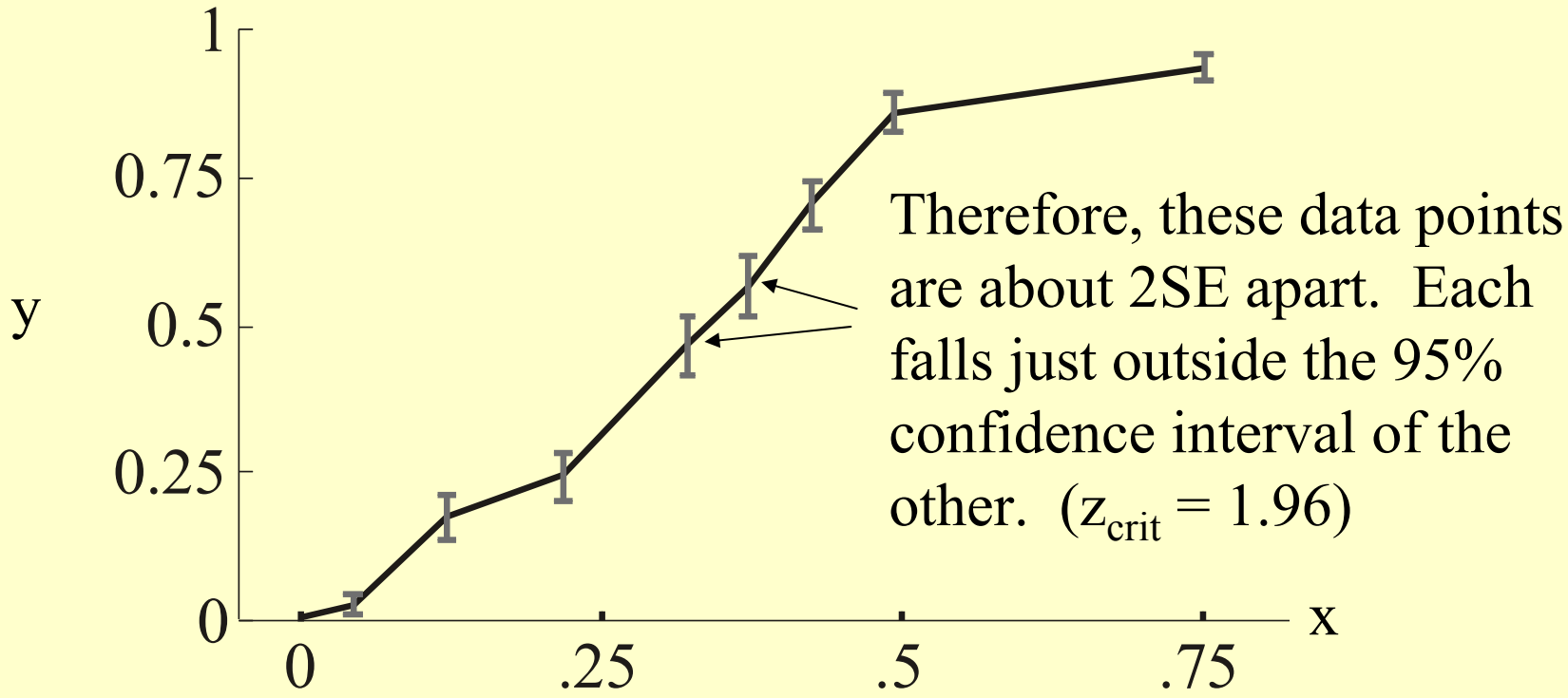
Some notes on error bars

- Then, we can make some inferences, just based on the graph.



Some notes on error bars

- Then, we can make some inferences, just based on the graph.



Some notes on error bars

- Then, we can make some inferences, just based on the graph.