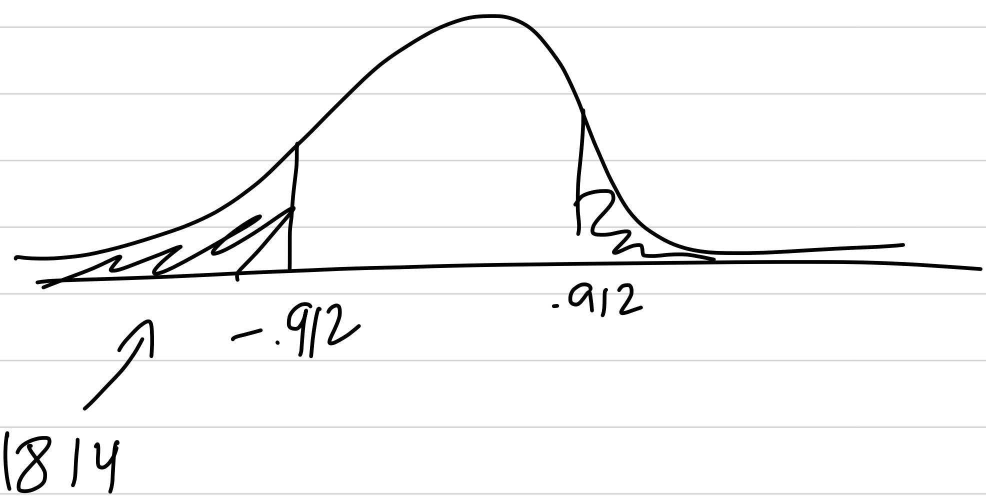
Ex: Let F be the random variable given by the number of fleas on a randomly selected household dog.

The distribution of F is not Normal, because it is discrete (b/c it only takes on integer values).

From studies, the population mean is approximately 2.7 with standard deviation 1.8. What is the approximate probability that a sample of 30 dogs will have a mean of more than 3?

By the Central Limit theorem, the distribution of $\bar{x}$ is approximately $N\left(2.7, \frac{1.8}{\sqrt{30}}\right) = N(2.7, .329)$

$$z = \frac{3 - 2.7}{.329} = .912$$



~18.14% chance of this sample mean being > 3

Chapter 16: Confidence Intervals

Statistical Inference for a Mean: we have an SRS, and the population is large compared to the sample size. We're measuring a variable whose distribution is $N(\mu, \sigma)$. We don't know μ, but we do know σ.

<u>Def</u>: A level C confidence interval for a parameter has two parts.

① An interval calculated from some data, of the form

estimate $\pm$ margin of error

② A confidence level C, which gives the probability that the interval will capture the true parameter value (i.e. the predicted success rate). The most common confidence level is 95%

What does this mean? For example, if you have a confidence interval of

$$\underbrace{5 \pm .2 \text{ with } 95\% \text{ confidence}}$$

we got to these numbers with a method that gives correct results 95% of the time.

$2. \frac{7.5}{\sqrt{654}}$

Population

μ unknown

$\sigma \approx 7.5$

$\xrightarrow{SRS \ n=654}$ $\bar{x} = 26.8 \pm .6$

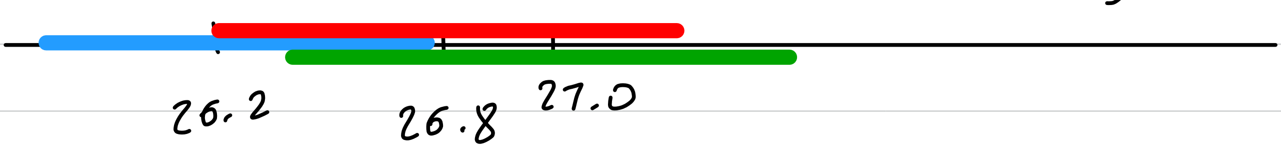
$\xrightarrow{SRS \ n=654}$ $\bar{x} = 27.0 \pm .6$

$\xrightarrow{SRS \ n=654}$ $\bar{x} = 26.2 \pm .6$

$\vdots$

$$\left.\begin{array}{l} 26.8 \pm .6 \\ 27.0 \pm .6 \\ 26.2 \pm .6 \\ \vdots \end{array}\right\} \text{95\% will contain } \mu$$

95% of bands contain μ



Ex: A Gallup poll done in 2015 found that 26% of the 675 coffee drinkers in the sample were addicted to coffee. Here is how Gallup announced their results: "with 95% confidence, the maximum margin of error is ± 5 percentage points".

what is the confidence interval?

$26\% \pm 5\%$, so between 21% and 31%

What does this mean?

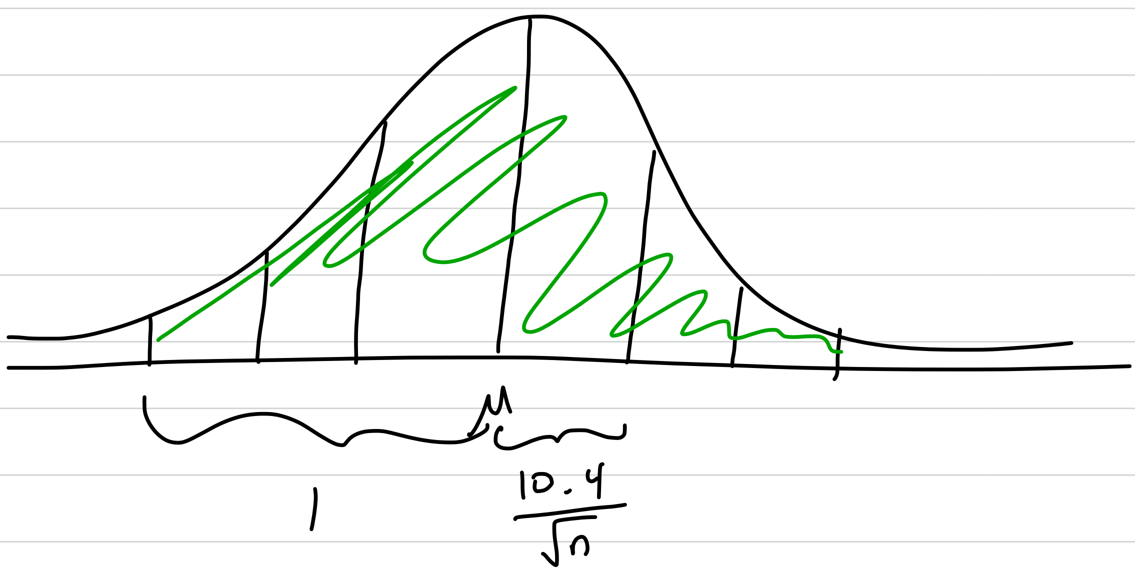
The chance that the actual proportion of the population addicted to coffee is between 21% and 31% is 95%

$N(\mu, 10.4)$

Sample of size n

distribution of $\bar{x}$ is $N\left(\mu, \frac{10.4}{\sqrt{n}}\right)$

$\frac{10.4}{\sqrt{n}}$



$$\frac{10.4}{\sqrt{n}} \cdot 3 = 1$$

distribution of $\bar{x}$ is $N(80, \, ^4/\sqrt{25})$

$$= N(80, .8)$$

$$z = \frac{86-80}{.8} = 7.5$$

$$P(\text{positive} \mid \text{disease}) = \frac{P(\text{positive and disease})}{P(\text{disease})}$$

$$P(\text{disease}) = \frac{\#\text{ patients w/ disease}}{\#\text{ patients}}$$

$$= \frac{574}{1286}$$

$P(\text{positive and disease}) = \dfrac{564}{1286}$

$$\frac{564/1286}{574/1286} = \frac{564}{574} = 98.2\%$$

Priors

$P(A) = .26$

$P(B) = .49$

$P(M) = .2$

$P(D) = .05$

Posterior

$P(A \mid F)$

A = event of getting an associate degree

F = event of the recipient being female

$.61 = P(F \mid A)$

$P(F) = .5$

$$P(A \mid F) = \frac{P(F \mid A)\, P(A)}{P(F)}$$

$$= \frac{(.61)(.26)}{.5} = .317$$

$\sigma = 13$

$\bar{X}$: $n = 7$

$$\bar{x} : N\left(\mu, \frac{13}{\sqrt{7}}\right)$$

Central Limit Theorem: sample of size n, the distribution of $\bar{x}$ is approximately $N\left(\mu, \frac{\sigma}{\sqrt{n}}\right)$.

A and B are disjoint

$P(\text{clubs or diamonds}) = \frac{12}{52} + \frac{12}{52}$

$P(\text{green}) = .1$

$P(\text{shirt}) = .4$

$P(\text{green or shirt}) = .45$

$P(\geq 1 \text{ type O}) = 1 - P(\text{no type O})$

$1 - (.928)^{10}$

$\underbrace{(0, \text{not } 0)}_{(.072)(.928)}$ $\quad$ $(\text{not } 0, 0)$,

$(.928)(.072)$

$(\text{not } 0, \text{not } 0)$

$.982 \cdot .982$

$(0, 0)$

$.072 \cdot .072$

- AND problem: try to find independence
- OR: try disjointness, or (if multiple of the events can happen at the same time), try inverting the event and taking 1 − the new prob. If it's still

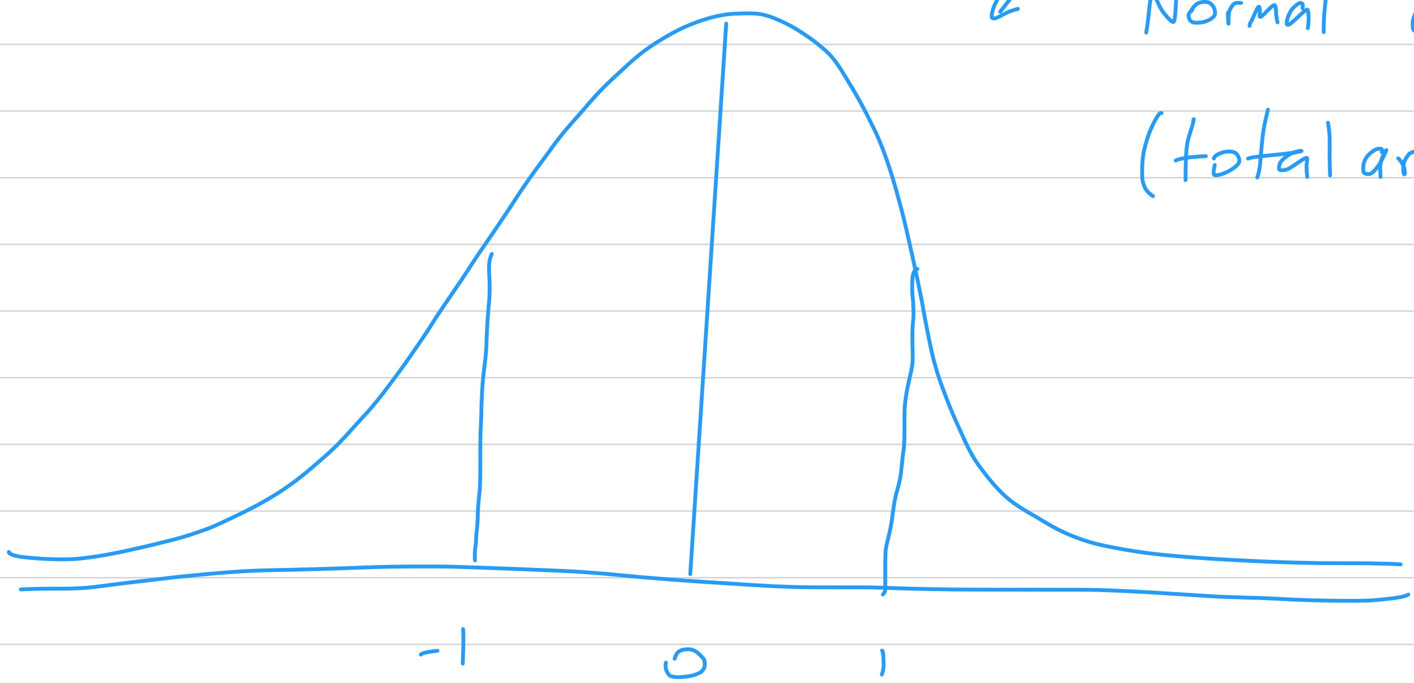
not working, try a Venn diagram

- Conditional probability: directly or Bayes'

$P(\text{heart}) = 13/52$

$P(\text{heart} \mid \text{red}) = 13/26$ ← restriction of S to red cards

Standard Normal curve (total area = 1)



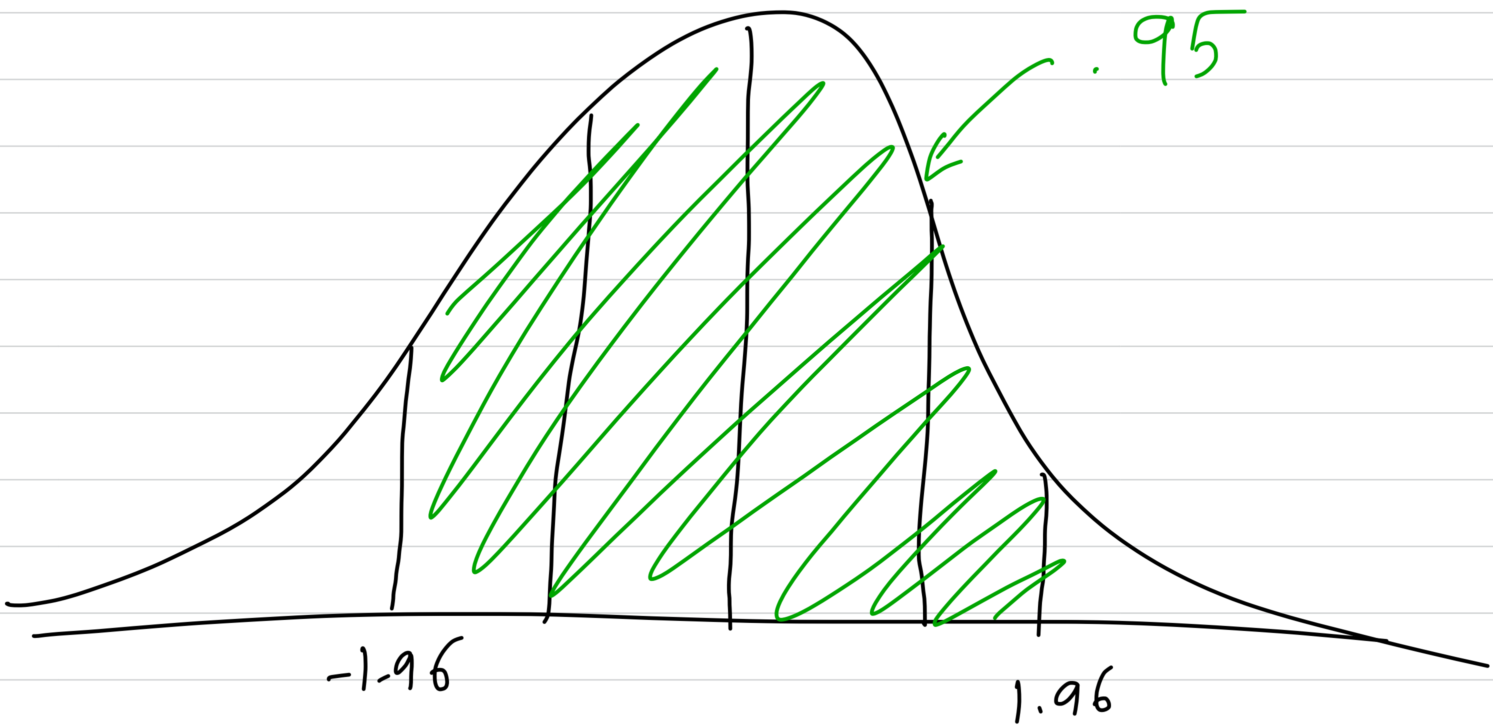
If we want a confidence level of C, we want the area under a portion of the standard Normal curve to be C.

<u>Def</u>: Given a confidence level C, the critical value z^* is the z-score such that the area

between $-z^*$ and z^* is C.

this is typically approximated by $z = 2$

Ex: For $C = .95$, $z^* = 1.96$



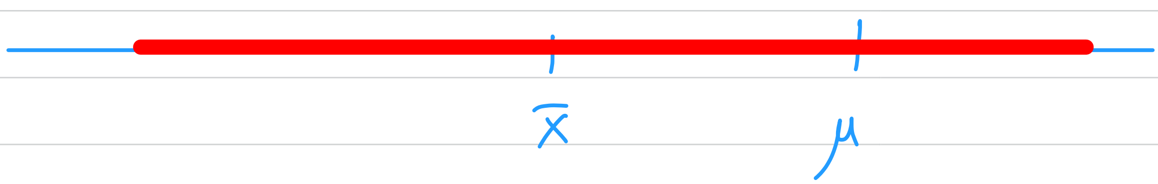
Comment: The most important critical values are :

$C = .9$: $z^* = 1.645$

$C = .95$: $z^* = 1.96$

$C = .99$: $z^* = 2.576$

For example, if you have sample of 500 people from a population with some mean μ and standard deviation 20, then if you want a confidence level of 99%, you need to have confidence interval $2.576 \cdot \left(\frac{20}{\sqrt{500}}\right)$ to either side of the sample mean $\bar{x}$.



Therefore, the confidence interval is

$$\bar{x} \pm z^* \frac{\sigma}{\sqrt{n}} \quad \text{(in a Normal population)}$$

How can we make the margin of error smaller?

- z^* is smaller — but this lowers C
- σ is smaller — but this is outside of our control
- n is larger — warning: n is under a root, so increasing the sample size by a factor of 4 only halves the margin of error.

Chapter 17: Hypothesis Tests

Ex: Suppose we have a distribution of phone prices that is $N(450, 108)$.

We sample 12 customers on their phone prices. We get

480	515	360	580	560	545
550	530	540	580	480	445

$$\bar{X} = 514$$

Assuming that the mean is in fact 450, how likely was this sample?

Def: The null hypothesis, sometimes denoted H_0 (read H-naught) is the proposal that models the status quo: e.g. a claim involving the fact that $\mu = 450$.

An alternative hypothesis, sometimes denoted H_a, is the desired result of an experiment.

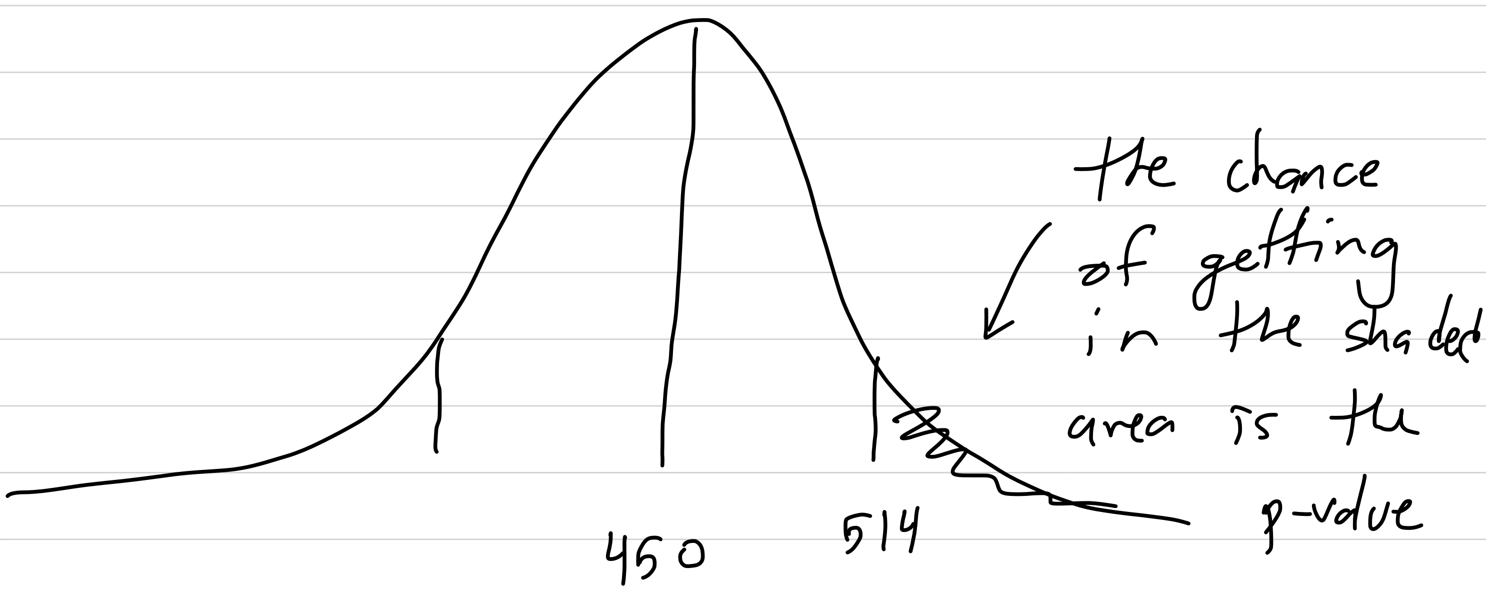
The p-value is the probability that, given that H_0 is true, that we would find a value of our statistic more extreme.

Ex: The null hypothesis is that $\mu = 450$

The alternative hypothesis, H_a, is

that $\mu > 450$. What is the p-value of this sample?

Here, H_a is <u>one-sided</u>: we only care about getting samples to one side of the observation (here, that's to the right)



This is a sample of 12 people, so the sample standard deviation is $108/\sqrt{12} = 31.18$.

So the z-score of 514 is

$$z = \frac{514 - 450}{31.18} = 2.05$$

from z-score table

proportion of .0202 above this z-score. So the p-value is $p = .0202$, and this means if the mean is truly $\mu = 450$, then this sample only had a 2% chance to occur. We say that this sample is statistically significant at level 2%

<u>Method</u>: How to perform a hypothesis test.

1. Write down the null and alternative hypotheses.

2. Find the test statistic $z = \frac{\bar{x} - \mu}{\sigma / \sqrt{n}}$

3. Find the p-value: can be left, right, or two-tailed depending on H_a.

4. Pick a value of α. If $p < \alpha$, we say that we reject the null hypothesis. Otherwise, we say we fail to reject the null hypothesis.

Remark: this is how the scientific process works: you can't directly prove things, only disprove them, and it's only when you fail to disprove something repeatedly that you're forced to accept it.

Ex: The systolic blood pressure of adult males is approximately $N(128, 15)$. The medical director of a large company wants to determine if the company's executives have a $\boxed{\text{different}}$ mean blood pressure from the general population. The medical records from 72 male executives found the mean blood pressure to be $\bar{x} = 126.07$. Is there sufficient evidence to conclude that the blood pressure of the executives is $\boxed{\text{different}}$ at a significance level of .05?

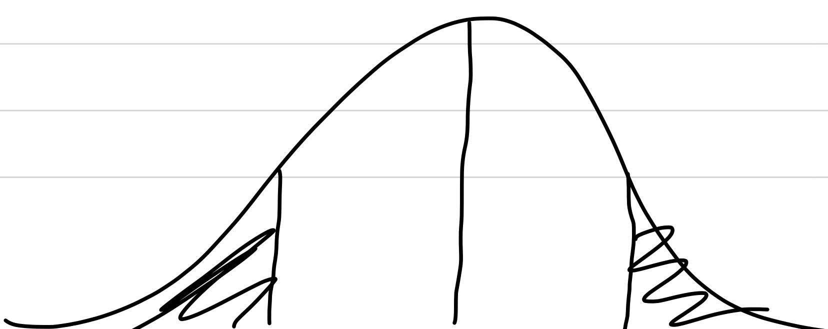
different: want a two-tailed p-value.

① H_0: $\mu = 128$ (without other info, we should assume that the male executives have the same distribution as the general population)

H_a: $\mu \neq 128$

② $z = \dfrac{126.07 - 128}{15/\sqrt{72}} = -1.09.$

③



$-1.09 \quad 0 \quad 1.09$

Area of the shaded regions:

$p = 2(.1379) = .276$

④ Is it true that $p < .05$? No!

We fail to reject the null hypothesis.

Ex: We wish to determine if NBA players are taller than the male population. The distribution of heights in that population is $N(69.3, 2.8)$. You take an SRS of 25 NBA players and find $\bar{x} = 73.2$. Is this significant at the .05 level?

① $H_0: \mu = 69.3$

$H_a: \mu > 69.3$

② $z = \dfrac{73.2 - 69.3}{2.8/\sqrt{25}} = 6.96$

③

proportion: $\dfrac{6.52 \cdot 10^{-23}}{1.77}$

④ It is stastically significant at a level of .05 (and much less), so we reject the null hypothesis.

Chapter 18: Considerations when doing Inference

Question: when can we create a confidence interval with z^*? When can a perform a hypothesis test on the mean μ?

- Large sample (generally > 30)
  - ↳ can get away with smaller if the distribution is Normal and you have no outliers
- SRS

- Need to know σ (!)

Cautions: The value for α is a value judgement – usually .05.

- If rejecting the null hypothesis is a big deal, make α small

  Ex: rejecting Newton's laws of physics

The p-value being significant does not mean that the difference between H_0 and H_a is large – it just says there's a difference.