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

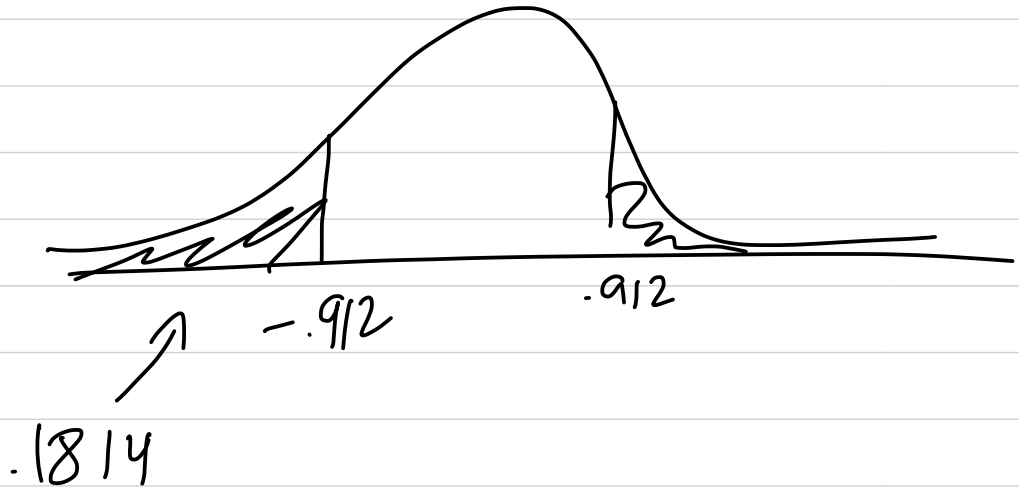
The distribution of F is not Normal, because it is discrete (b/c it only takes on integer values).

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.

Def: A level C confidence interval for a parameter has two parts.

① An interval calculated from some data, of the form

estimate ± 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

$5 \pm .2$ with 95% 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

$\sigma \approx 7.5$

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

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

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

⋮

$26.8 \pm .6$

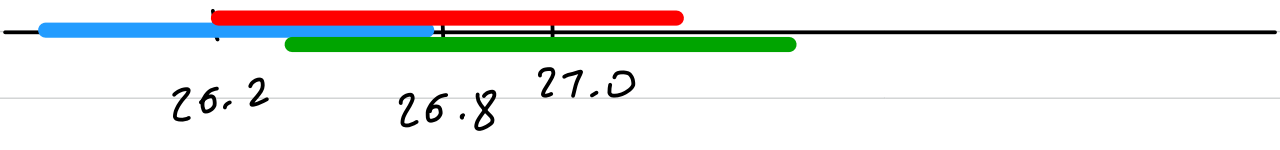
$27.0 \pm .6$

$26.2 \pm .6$

⋮

95% will contain μ

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% ± 5%, so between 21% and 31%

What does this mean?

The chance that the actual proportion of the population addicted

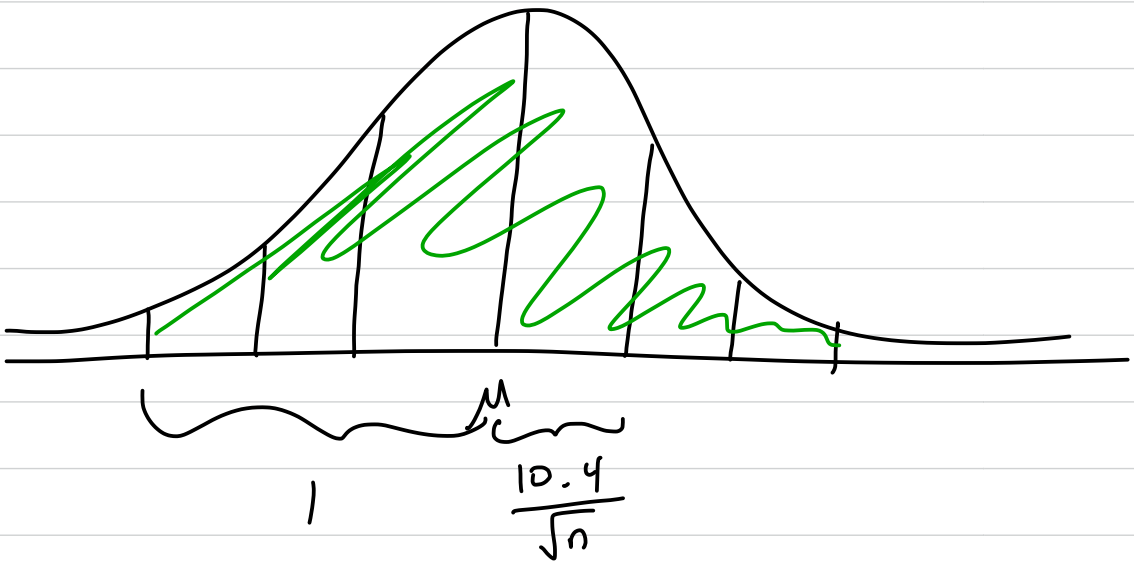
to coffee is between 21% and 31% is 95%

$N(\mu, 10.4)$

Sample of size n

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

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



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

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

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

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

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

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

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

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

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

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

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

$P(F) = .5$

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

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

$\sigma = 13$

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

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

Central Limit Theorem: sample of size n, the distribution of $\bar{x}$ is

approximately $N(\mu, \frac{\sigma}{\sqrt{n}})$.

A and B are disjoint

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

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

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

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

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

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

(O, not O) (not O, O),

(.072)(.928) (.928)(.072)

(not O, not O) | (O, O)

.982 · .982 | .072 · .072

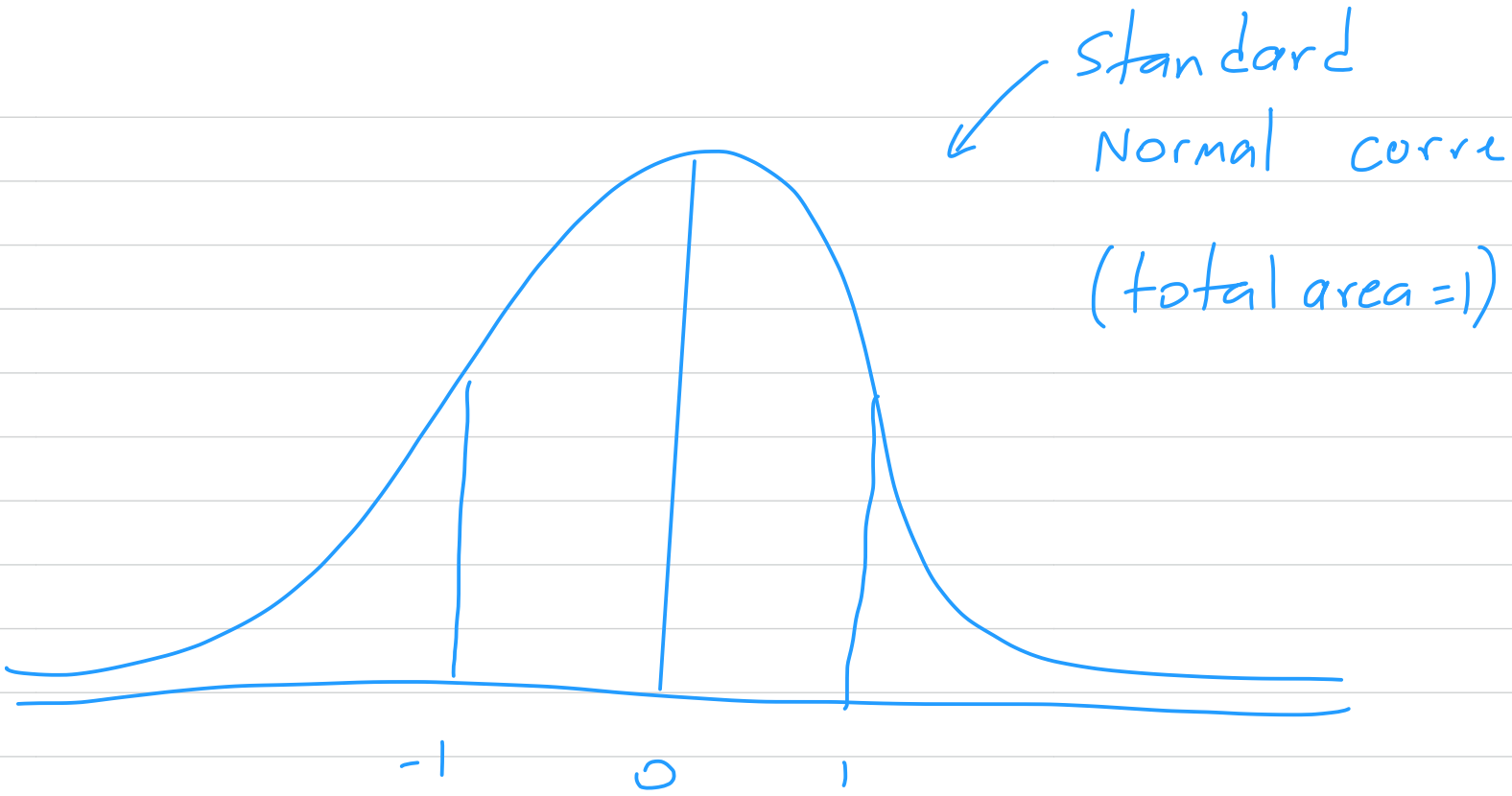
- AND problem: try to find independence
- OR: try disjointness, or (if multiple of the events can happen at the same time), try inverting the event and taking 1 − the new prob. If it's still

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



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

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

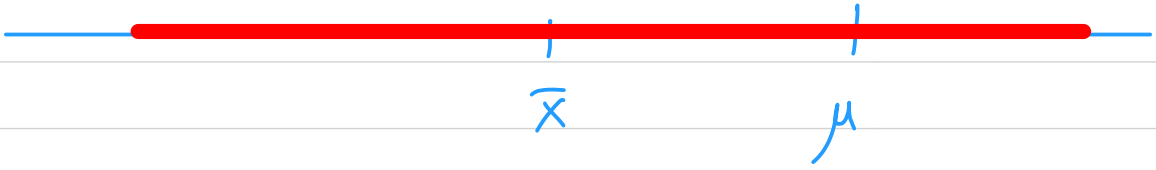
Comment: The most important critical values are :

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

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

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

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

$$\bar{x} \pm z^* \frac{\sigma}{\sqrt{n}}$$ (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.

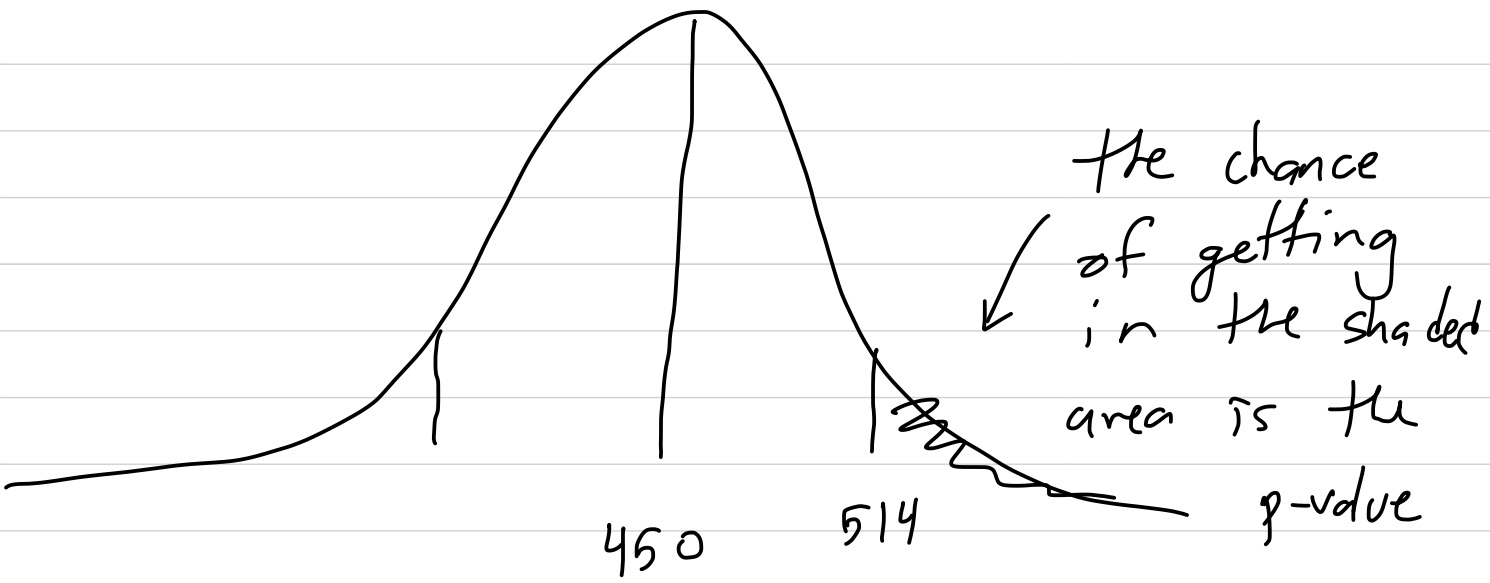
Ex: The null hypothesis is that

$\mu = 450$

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%

Method: 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 <u>reject the null hypothesis</u>. Otherwise, we say we <u>fail to reject the null hypothesis</u>.

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 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 different at a significance level of .05?

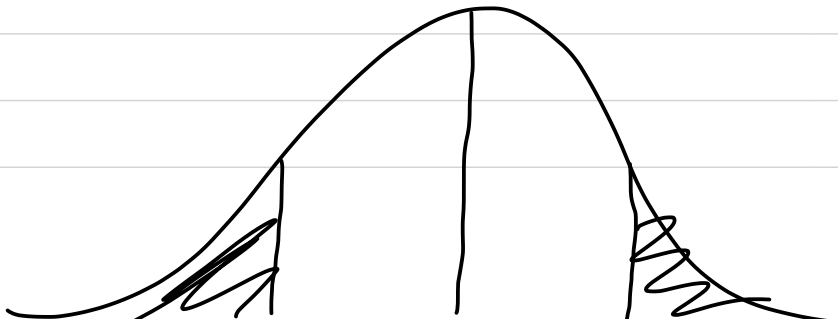
different: want a two-tailed p-value.

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

H_a: $\mu \neq 128$

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

③



-1.09 0 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.

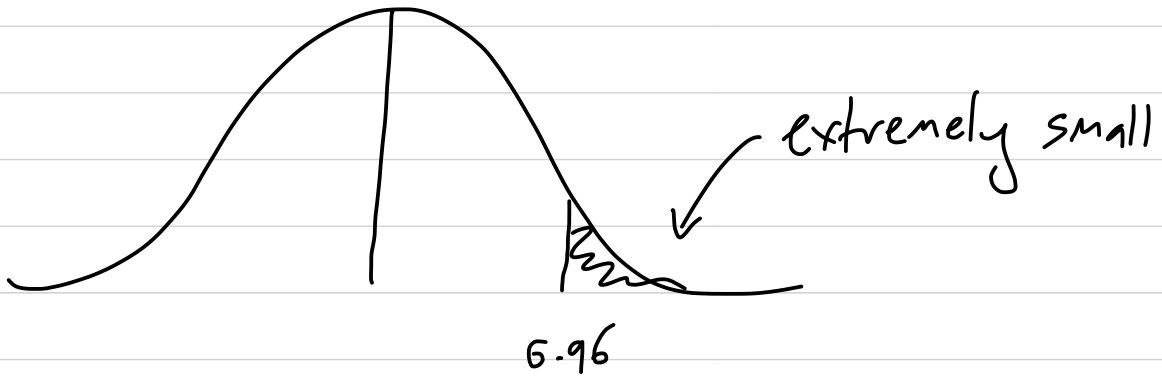
Ex: We wish to determine if NBA players are taller than the male population. The distribution of heights in that population is $N(69.3, 2.8)$. You take an SRS of 25 NBA players and find $\bar{x} = 73.2$. Is this significant at the .05 level?

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 statistically 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.

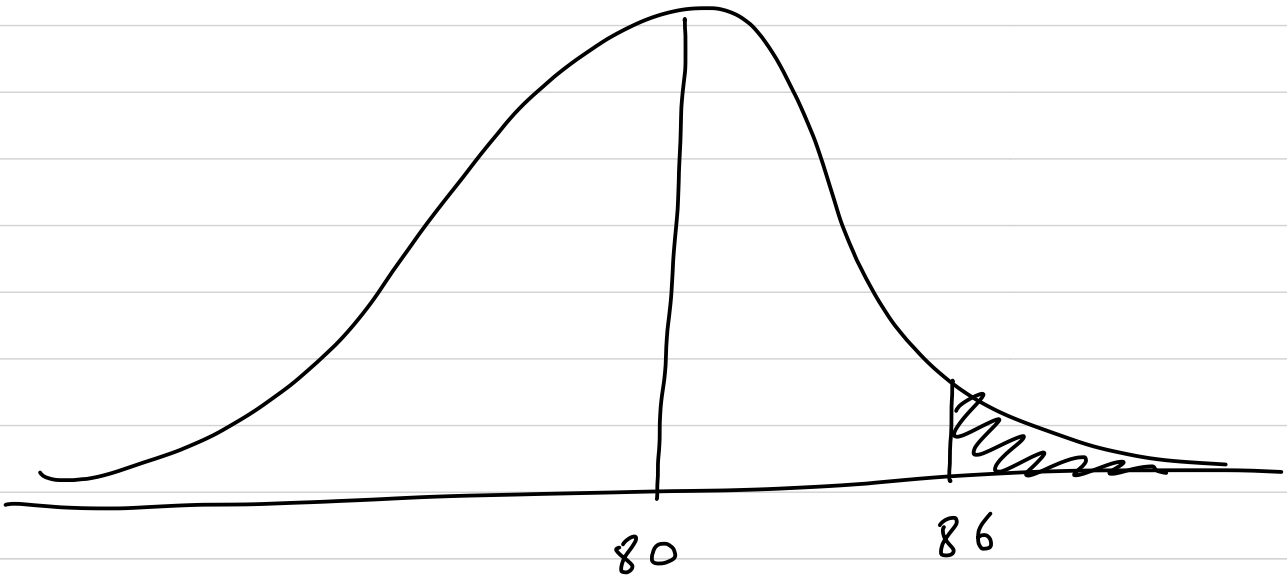
- If rejecting the null hypothesis is a big deal, make α small

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.

$\mu = 80$

$\sigma = 4$



$$z = \frac{86 - 80}{4} = 1.5 \Rightarrow .0668$$

$\bar{X}$ is $N\left(80, \frac{4}{\sqrt{25}}\right) = N(80, .8)$

$$z = \frac{86 - 80}{.8} = 7.5 \Rightarrow \text{effectively } 0$$

. . .

$\mu = 8.8$

$\sigma = 1.0$

$\bar{x}$ is approximately $N\left(8.8, \frac{1}{\sqrt{24}}\right)$

↓

units of beats/5 sec

$$100 \ \frac{\text{beats}}{60 \text{ sec}} = \frac{100}{12} \ \frac{\text{beats}}{5 \text{ sec}} = 8.33$$

$$z = \frac{8.33 - 8.8}{.204} = -2.30 \Rightarrow .0107$$

$N(\mu, .2)$ $n = 6$ Want mean to be 5

5.32 4.88 5.1 4.73 5.15 4.75

$C = 90\%$

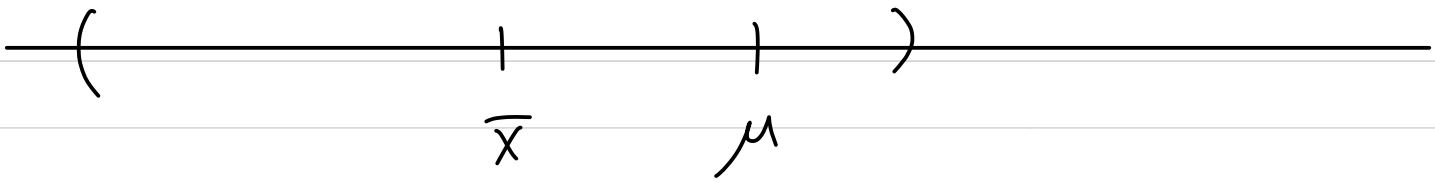
$\bar{x} = 4.988 \rightarrow$ roughly follows $N\left(\mu, \frac{.2}{\sqrt{6}}\right)$

$= N(\mu, .082)$

$z = 1.645$

Confidence interval: $4.988 \pm 1.645(.082)$

$= 4.988 \pm .135$



95%

Comment: A caution about multiple comparisons.

For example, testing 20 different hypotheses with $\alpha = .05$ will, on average, give 1 hypothesis for which we (incorrectly) reject H_0.

Ex: if you try to find a link between jelly beans and acne, with $\alpha = .05$, you probably won't

find statistically significant data to reject the null hypothesis. On the other hand, if you individually test every color for causing acne, you probably will find a link between one or two colors.

Comment: α is the probability of incorrectly rejecting H_0. β is the probability of failing to reject H_0 when we should have. So $1-\beta$ is the probability of correctly rejecting H_0, and this is called the power of the test.

Type I error ← rejecting H_0

Type II error ← failing to reject H_0

Chapter 20: Inference about the Population Mean

Comment: Right now, we need to know the population standard deviation σ in order to make a confidence interval and do a hypothesis test. In practice, we won't know σ. What we will know is s, the sample standard deviation.

	sample	population
Mean	$\bar{x}$	μ
standard deviation	s	σ

We will use what are called 1-sample t procedures. We can use these when:

- Want to know μ
- The distribution is roughly Normal (rough symmetry, no outliers, mostly clustered around the center)
- Data collected via an SRS
- Population at least 20x as large as the sample.

Recall: If we know σ, the sample standard deviation is $\sigma/\sqrt{n}$.

If we don't know σ, we use something called the standard error.

Def: Let $x_1, x_2, \cdots, x_n$ be a data set with mean $\bar{x}$ and standard deviation s. The standard error of $\bar{x}$ is $s/\sqrt{n}$.

Recall: To find s, first find $\bar{x}$, then take $s^2 = \frac{1}{n-1}\left((x_1-\bar{x})^2 + (x_2-\bar{x})^2 + \cdots + (x_n-\bar{x})^2\right)$.

Notice that this seems kind of suspect: we're a sample's own standard deviation to make claims about the population's mean.

When using the standard error and not the population standard deviation, we need to use a t-score instead of a z-score.

z-score	t-score
$z = \frac{\bar{x} - \mu}{\sigma/\sqrt{n}}$	$t = \frac{\bar{x} - \mu}{s/\sqrt{n}}$

Difference: the distribution of t-scores is different from that of z-scores.

TABLE C *t* distribution critical values

Degrees of freedom	Confidence level *C*											
	50%	60%	70%	80%	90%	95%	96%	98%	99%	99.5%	99.8%	99.9%
1	1.000	1.376	1.963	3.078	6.314	12.71	15.89	31.82	63.66	127.3	318.3	636.6
2	0.816	1.061	1.386	1.886	2.920	4.303	4.849	6.965	9.925	14.09	22.33	31.60
3	0.765	0.978	1.250	1.638	2.353	3.182	3.482	4.541	5.841	7.453	10.21	12.92
4	0.741	0.941	1.190	1.533	2.132	2.776	2.999	3.747	4.604	5.598	7.173	8.610
5	0.727	0.920	1.156	1.476	2.015	2.571	2.757	3.365	4.032	4.773	5.893	6.869
6	0.718	0.906	1.134	1.440	1.943	2.447	2.612	3.143	3.707	4.317	5.208	5.959
7	0.711	0.896	1.119	1.415	1.895	2.365	2.517	2.998	3.499	4.029	4.785	5.408
8	0.706	0.889	1.108	1.397	1.860	2.306	2.449	2.896	3.355	3.833	4.501	5.041
9	0.703	0.883	1.100	1.383	1.833	2.262	2.398	2.821	3.250	3.690	4.297	4.781
10	0.700	0.879	1.093	1.372	1.812	2.228	2.359	2.764	3.169	3.581	4.144	4.587
11	0.697	0.876	1.088	1.363	1.796	2.201	2.328	2.718	3.106	3.497	4.025	4.437
12	0.695	0.873	1.083	1.356	1.782	2.179	2.303	2.681	3.055	3.428	3.930	4.318
13	0.694	0.870	1.079	1.350	1.771	2.160	2.282	2.650	3.012	3.372	3.852	4.221
14	0.692	0.868	1.076	1.345	1.761	2.145	2.264	2.624	2.977	3.326	3.787	4.140
15	0.691	0.866	1.074	1.341	1.753	2.131	2.249	2.602	2.947	3.286	3.733	4.073
16	0.690	0.865	1.071	1.337	1.746	2.120	2.235	2.583	2.921	3.252	3.686	4.015
17	0.689	0.863	1.069	1.333	1.740	2.110	2.224	2.567	2.898	3.222	3.646	3.965
18	0.688	0.862	1.067	1.330	1.734	2.101	2.214	2.552	2.878	3.197	3.611	3.922
19	0.688	0.861	1.066	1.328	1.729	2.093	2.205	2.539	2.861	3.174	3.579	3.883
20	0.687	0.860	1.064	1.325	1.725	2.086	2.197	2.528	2.845	3.153	3.552	3.850
21	0.686	0.859	1.063	1.323	1.721	2.080	2.189	2.518	2.831	3.135	3.527	3.819
22	0.686	0.858	1.061	1.321	1.717	2.074	2.183	2.508	2.819	3.119	3.505	3.792
23	0.685	0.858	1.060	1.319	1.714	2.069	2.177	2.500	2.807	3.104	3.485	3.768
24	0.685	0.857	1.059	1.318	1.711	2.064	2.172	2.492	2.797	3.091	3.467	3.745
25	0.684	0.856	1.058	1.316	1.708	2.060	2.167	2.485	2.787	3.078	3.450	3.725
26	0.684	0.856	1.058	1.315	1.706	2.056	2.162	2.479	2.779	3.067	3.435	3.707
27	0.684	0.855	1.057	1.314	1.703	2.052	2.158	2.473	2.771	3.057	3.421	3.690
28	0.683	0.855	1.056	1.313	1.701	2.048	2.154	2.467	2.763	3.047	3.408	3.674
29	0.683	0.854	1.055	1.311	1.699	2.045	2.150	2.462	2.756	3.038	3.396	3.659
30	0.683	0.854	1.055	1.310	1.697	2.042	2.147	2.457	2.750	3.030	3.385	3.646
40	0.681	0.851	1.050	1.303	1.684	2.021	2.123	2.423	2.704	2.971	3.307	3.551
50	0.679	0.849	1.047	1.299	1.676	2.009	2.109	2.403	2.678	2.937	3.261	3.496
60	0.679	0.848	1.045	1.296	1.671	2.000	2.099	2.390	2.660	2.915	3.232	3.460
80	0.678	0.846	1.043	1.292	1.664	1.990	2.088	2.374	2.639	2.887	3.195	3.416
100	0.677	0.845	1.042	1.290	1.660	1.984	2.081	2.364	2.626	2.871	3.174	3.390
1000	0.675	0.842	1.037	1.282	1.646	1.962	2.056	2.330	2.581	2.813	3.098	3.300
*z**	0.674	0.841	1.036	1.282	1.645	1.960	2.054	2.326	2.576	2.807	3.091	3.291
One-sided *P*	.25	.20	.15	.10	.05	.025	.02	.01	.005	.0025	.001	.0005
Two-sided *P*	.50	.40	.30	.20	.10	.05	.04	.02	.01	.005	.002	.001

Differences between t and z:

The t distribution has more area in the tails (greater variability) (confidence intervals become wider and p-values become larger.

Similarities: both bell-shaped and symmetric.

Method: To produce a level C confidence interval for a population mean μ with unknown standard deviation:

$$\bar{x} \pm t^* \frac{s}{\sqrt{n}} \quad \left(\text{c.f. } \bar{x} \pm z^* \frac{\sigma}{\sqrt{n}}\right)$$

To find t^*, you need to know C and the degrees of freedom (DOF): $DOF = n-1$.

Ex: We find the diameter in cm of 10 snowflakes

9.34	6.08	3.39	1.84	3.19
5.30	7.18	4.11	5.10	7.27

Find a 96% confidence interval for the population mean.

$\overline{x} : 5.28$

$s^2 = \frac{1}{9}\left((9.34-5.28)^2 + \cdots + (7.27-5.28)^2 \right)$

$= 5.09$

$s = 2.26$

$n = 10$, so $DOF = 9$

$C = .96$

$\Rightarrow t^* = 2.398$

$$5.28 \pm 2.398 \cdot \frac{2.26}{\sqrt{10}}$$

$$\boxed{5.28 \pm 1.714}$$

There is a 96% chance that μ is between $5.28 - 1.714$ and $5.28 + 1.714$.

Comment:
- When $n < 15$, we need the data to be completely Normal.
- When $16 < n < 30$, it's okay to have a slight skew
- when $n > 30$, no need for the population to be Normal.

We draw an SRS of size n from a subset of a large population with mean μ_0

To test the null hypothesis $\mu = \mu_0$, which says that the mean of the mean of the subset is the mean of the population, we compute $t = \dfrac{\bar{x} - \mu_0}{s/\sqrt{n}}$.

disjoint

$$P(A \text{ or } B) = P(A) + P(B)$$

$$P(A \text{ and } B) = P(A)P(B)$$

independent

B : blue

E : 8 sides

$P(B) = .25$ $\quad$ $P(E) = .1$

$$P(E \mid B) = .5$$

$$P(B \mid E) = \frac{P(E \mid B)\,P(B)}{P(E)} = \frac{(.5)(.25)}{(.1)}$$

Bayes' theorem

$P(A \text{ and } B)$ $P(A \text{ or } B)$

know one, find the other

- $P(A \mid B) = \dfrac{P(A \text{ and } B)}{P(B)}$

$P(A \text{ and } B) = .02$

$P(A) = .2$

$P(B) = .1$

$.02 = P(A)P(B)$

When is $P(A \text{ and } B) = P(A)P(B)$?

When they're independent

B : blue $\quad P(B) = .25$

E : 8-sided $\quad P(E) = .1$

$P(E \mid B) = .5$

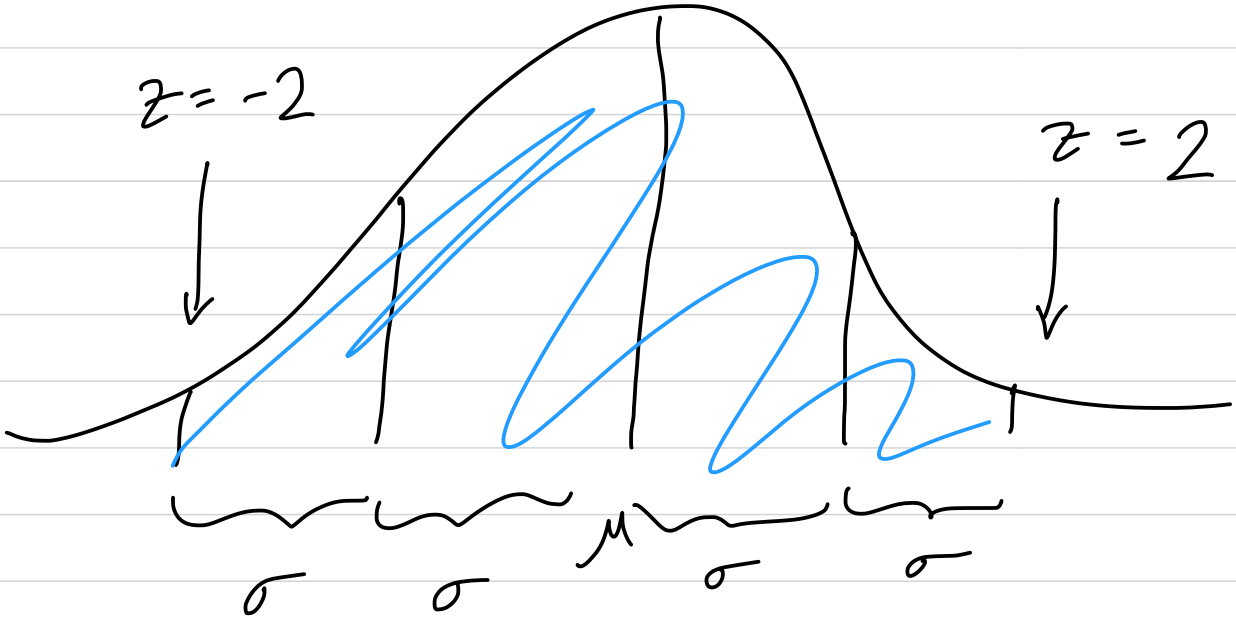
$$P(B|E) = \frac{P(E|B)\,P(B)}{P(E)}$$

$N(\mu, .75)$ $\quad n = 30 \quad$ $\bar{x} = 2.7$

$C = 95\%$

$\bar{X}$ is $N\left(\mu, \frac{.75}{\sqrt{30}}\right)$

What interval of z-scores contains 95% of the data?



$z^* = 2$ for $C = 95\%$

$$\bar{x} \pm z^* \frac{\sigma}{\sqrt{n}} = 2.7 \pm 2 \cdot \frac{.75}{\sqrt{30}}$$

$$= 2.7 \pm .274$$