

Political Poll Headline

More than half of Canadians approve of Trudeau: poll

CHRIS HANNAY

Ottawa — The Globe and Mail

Published Friday, Jan. 15, 2016 8:34AM EST

> The Justin Trudeau honeymoon continues: an **Abacus Data poll** this morning shows 53 per cent of respondents give a thumbs-up to the Liberal government, and even more rate the Prime Minister positively.

“53%”? Really “More Than Half”? Just Luck??

- “Today, 53% say they approve of the performance of the federal government”. Details at: <http://abacusdata.ca/how-do-we-feel-about-the-trudeau-government/>
 - “Our survey was conducted online with 1,500 Canadians aged 18 and over”.
 - (Also: “The margin of error for a comparable probability-based random sample of the same size is $\pm 2.6\%$, 19 times out of 20.” Huh? Later!)
- So does this demonstrate that “more than half” approve?
- Or, could it be that half or less approve, but just by luck they polled more people who approved than who didn’t?
- To consider this properly, need . . . a P-value! How?

Statistical Analysis of Polls

- How to model a poll in terms of probability? What null hypothesis? Etc. Well ...
- Suppose that some (unknown) fraction p of Canadians approve of the government (or, at least, would say that they do).
- The poll selects Canadians uniformly at random (hopefully! online??). It then asks each one of them if they approve (Yes/No).
- Can consider this as repeating a random “experiment” each time: select a Canadian at random, and ask them if they approve (Yes/No). Then repeat many times (e.g. 1,500).
- Kind of like repeatedly flipping a coin, or playing RPS, or guessing cards in the ESP experiment, or repeatedly rolling a fair six-sided die. But this time we don't really know a guess for p . (Before: $p = 1/2$, or $p = 1/3$, or $p = 1/6$, or ...)

Statistical Analysis of the Trudeau Poll

- For the Trudeau poll, one possible null hypothesis: does $p = 50\%$? Versus alternative hypothesis: is $p > 50\%$?
- So, here we define the P-value as the probability, assuming $p = 50\%$, that if we sample 1,500 people and ask each one of them if they approve, that 53% or more will answer Yes.
- Now, 53% of 1,500 is: 795.
- So, the question is, if you repeatedly conduct an experiment, which has probability $p = 50\%$ of getting Yes each time, a total of 1,500 times, then what is the probability that you will get Yes 795 or 796 or 797 or 798 or 799 or 800 or 801 or ... or 1,500 times.
- (Phew!) How to compute? Binomial distribution?
- Or, use normal approximation! What mean? What variance or sd?

P-value for the Trudeau Poll

- Recall: out of 1,500 Canadians, 795 approved. Null: $p = 50\%$.
- P-value? Could use: `pbinom(794, 1500, 0.50, lower.tail=FALSE)`

Ans = 0.01076618

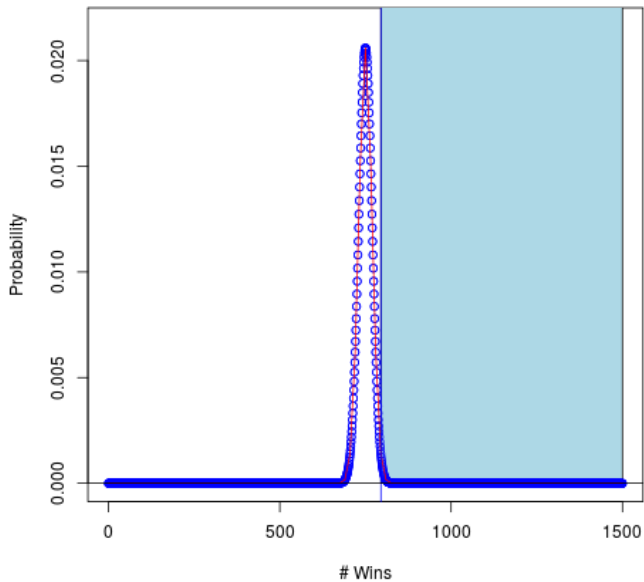
- Or: if you do an experiment $n = 1,500$ times, with probability $1/2$ of success each time (like flipping a fair coin, or RPS, or ...), then mean = $n/2 = 1,500/2 = 750$, and variance = $n/4 = 1,500/4 = 375$. So, $sd = \sqrt{375} \doteq 19.36$.

- So, the P-value is approximately the probability that a normal distribution with mean=750 and variance=375 (i.e., with width = $sd = \sqrt{375} \doteq 19.36$), will equal 795 or more.

- Use the R command for this: `pnorm(794.5, 750, sqrt(375), lower.tail=FALSE)` Ans = 0.01078174. Very close!

- Much less than 0.05. Shows that $p > 50\%$? (But is $p > 51\%$?)

Poll Example: Diagram



New Babies: Boy or Girl?

- When a new baby is born, is it equally likely to be a boy or girl? Or is one more likely than the other?
- Suppose the probability of a new baby being a girl is “ p ”.
- Null hypothesis: $p = 1/2$. Alternative hypothesis: $p \neq 1/2$. (Not sure if $p > 1/2$ or $p < 1/2$, so include both possibilities: this is called a “two-sided” test.)



- To proceed, we need some ... data!
<http://www.statcan.gc.ca/tables-tableaux/sum-som/l01/cst01/health103a-eng.htm>

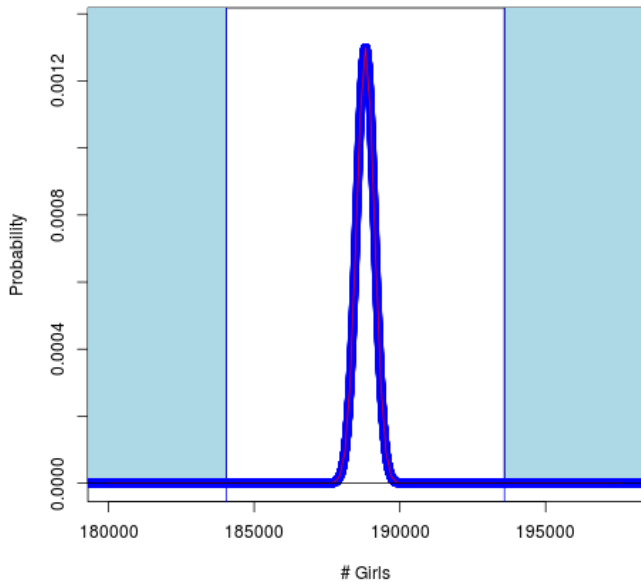
- In 2011, of 377,636 total births in Canada, 184,049 were females (girls). (48.74%) Just luck? Or does this “demonstrate” that girls are less likely?

Boy or Girl? (cont'd)

- Recall: of 377,636 total births, 184,049 (48.74%) were females.
- If $p = 50\%$, would expect $50\% \times 377,636 = 188,818$ females.
So is 184,049 too few? Or just random luck?
- Here the P-value is the probability, assuming the null hypothesis that $p = 50\%$, that out of 377,636 total births, there would be 184,049 or fewer females. Or, to make it two-sided, we should also include the probability of correspondingly too many females.
 - Here $|\text{observed} - \text{expected}| = |184,049 - 188,818| = 4,769$.
 - So, correspondingly too many females would be the expected number plus 4,769, i.e. $188,818 + 4,769 = 193,587$ or more.
 - So, perhaps we should really ask, what is the probability that out of 377,636 total births, the number of females would be either 184,049 or fewer, or 193,587 or more. (This is approximately twice the probability of just 184,049 or fewer females.)

- Recall: of 377,636 total births, 184,049 (48.74%) were females.
- P-value? Well, if $p = 50\%$ (null hyp.), the number of females has mean $= n/2 = 377,636/2 = 188,818$, and variance $= n/4 = 377,636/4 = 94,409$, so $sd = \sqrt{94,409} \doteq 307.26$.
- So, the P-value is approximately the probability that a normal distribution, with mean 188,818, and $sd \sqrt{94,409}$, is either 184,049 or fewer, or 193,587 or more. `pnorm(184049, 188818, sqrt(94409), lower.tail=TRUE) + pnorm(193587, 188818, sqrt(94409), lower.tail=FALSE)` Ans: 2.5×10^{-54} . Astronomically tiny!!
- So, we REJECT the null hypothesis that $p = 50\%$. Instead, we conclude that $p \neq 50\%$, i.e. boys and girls are not equally likely. (But actually more females than males in Canada: 17,147,459 female, 16,869,134 male. Why? Women live longer on average! [Statscan: 83 versus 79 years, if born today](#). So, more baby boys, but more old women!)

Babies Example: Diagram



P-Value Comparison: Sample Size

- Consider again a RPS competition.
- Suppose you win 11 out of 20 (55%). Impressive? No! P-value: `pnorm(10.5, 20/2, sqrt(20/4), lower.tail=FALSE)`. Answer is: 0.4115316. Not significant at all.
- Suppose you win 22 out of 40 (55%). Impressive? No! P-value: `pnorm(21.5, 40/2, sqrt(40/4), lower.tail=FALSE)`. Answer is: 0.3176281. Not significant at all. (But a bit smaller ...)
- Suppose you win 55 out of 100 (55%). Impressive? Here P-value is: `pnorm(54.5, 100/2, sqrt(100/4), lower.tail=FALSE)`. Answer is: 0.1840601. Not significant. (But smaller ...)
- Suppose you win 550 out of 1000 (55%). Impressive? Here P-value is: `pnorm(549.5, 1000/2, sqrt(1000/4), lower.tail=FALSE)`. Answer is: 0.000872085. Very significant!

Sample Size Comparison (cont'd)

- So, even with the same percentage correct (55%), if the sample size is larger, then the result has a smaller P-value, i.e. is more significant, i.e. is more surprising. Why?
- Intuition: winning 550/1000, is kind of like winning 11/20, repeatedly, 50 times in a row. (Since $11 \times 50 = 550$, and $20 \times 50 = 1000$.)
- More precisely: If you repeat an experiment (e.g. taking a sample) many times, then the average result will get closer and closer to the true value (e.g. $p = 1/2$). “Law of Large Numbers”
- So, with more samples, even minor percentage differences become more significant.
- (The issue of “how big a sample size do we need” is very important in statistical consulting! You’re all experts now!)

Expectation and Variance: More General Formula

- More generally, suppose we repeat an experiment “ n ” times, with probability “ p ” of success each time, and let “ T ” be the total number of successes. What are $E(T)$, $Var(T)$, and $sd(T)$?
 - (We already solved this when $n = 100$ and $p = 0.5$ [RPS], or $n = 12$ and $p = 1/6$ [dice], or $n = 200$ and $p = 1/38$ [roulette], ...)
- Well, let $X_i = 1$ if the i^{th} experiment is a success (e.g. a Head, or a Correct Guess, or a Five, or a “22”, or ...), otherwise $X_i = 0$.
- Then $E(X_i) = 1 \times p + 0 \times (1 - p) = p$.
- And, $Var(X_i) = E[(X_i - p)^2] = [1 - p]^2 \times p + [0 - p]^2 \times (1 - p) = (1 - 2p + p^2)p + p^2(1 - p) = (p - 2p^2 + p^3) + (p^2 - p^3) = p - p^2 = p(1 - p)$.
- Then, $T = X_1 + X_2 + \dots + X_n$.
 - So, mean = $E(T) = E(X_1) + E(X_2) + \dots + E(X_n) = n \times p = np$.

More General Formula (cont'd)

- Know: $T = X_1 + X_2 + \dots + X_n$, and $E(T) = np$.
 - Then, $Var(T) = Var(X_1) + Var(X_2) + \dots + Var(X_n) = n \times p(1 - p) = np(1 - p)$.
 - So, $sd(T) = \sqrt{Var(T)} = \sqrt{np(1 - p)}$.
- RPS Example: Play $n = 100$ times, had probability $p = 0.5$ of winning each time, got mean = 50, and variance = 25. Check: $np = 100 \times 0.5 = 50$ (same), and $np(1 - p) = 100 \times 0.5 \times (1 - 0.5) = 25$ (same). Good!
- Dice Example: Rolled a die $n = 12$ times, with probability $p = 1/6$ of getting a “5” each time. For the number of 5’s, got mean = 2, and variance = $5/3$. Check: $np = 12 \times 1/6 = 2$ (same), and $np(1 - p) = 12 \times 1/6 \times (1 - 1/6) = 5/3$ (same).

More General Formula (cont'd)

- Roulette Example (tried it?): Spun $n = 200$ times, with probability $p = 1/38$ of getting a “22” each time. For number of 22's, get mean $= 200/38$, and variance $= 1850/361$. Check: $np = 200 \times 1/38 = 200/38$, and $np(1 - p) = 200 \times 1/38 \times (1 - 1/38) = 1850/361$.
- So, for any repeated Success/Failure experiment (e.g. Yes/No, or Heads/Tails, or Correct/Incorrect, or Five/Not-Five, or 22/not-22, or ...), repeated n times, with probability of success p each time, the number of successes has mean np , and variance $np(1 - p)$, hence sd $\sqrt{np(1 - p)}$.
- Can use this for normal approximations, P-values, etc.
- We can apply this to polls! For example ...

Application: that Trudeau Government Poll again

- Recall: out of 1,500 Canadians surveyed in a poll, 795 of them (53%) approved of the government.
- We verified earlier that this “demonstrates” that more than half of Canadians approve. (P-value = 0.01006838.)
 - (Well, at least assuming that it was a truly random sample, everyone replied honestly, etc. Let’s assume that for now.)
- But does it demonstrate that over 51% approve? Let’s test that!
- Null hypothesis: $p = 0.51$, i.e. 51% approve.
 - Alternative hypothesis: $p > 0.51$, i.e. over 51% approve.
- Under the null hypothesis, the number (out of 1,500) who approve has mean = $np = 1,500 \times 0.51 = 765$, and variance = $np(1 - p) = 1,500 \times 0.51 \times 0.49 = 374.85$, so sd = $\sqrt{374.85}$.

Trudeau Government Poll (cont'd)

- Then the P-value should be the probability, assuming the null hypothesis that $p = 0.51$, that out of 1,500 people, 795 or more would approve.
- Use normal approximation, with mean = 765, and $sd = \sqrt{374.85}$:
`pnorm(794.5, 765, sqrt(374.85), lower.tail=FALSE)`
- Answer is: 0.06379456
- This is more than 0.05, so cannot reject the null, i.e. cannot claim that more than 51% approve.
- So what precisely can we claim? More than 50.5%??
- Should we just keep doing different hypothesis tests?
- Better: use confidence intervals. (The poll said: “The margin of error for a comparable probability-based random sample of the same size is +/- 2.6%, 19 times out of 20.” Huh? Later!)

Properties of Mean and Variance

• $E(X) = \sum_x x P(X = x)$, and $Var(X) = \sum_x (x - m)^2 P(X = x)$ (where $m = E(X)$), and $sd(X) = \sqrt{Var(X)}$.

• If you add a constant number (like 10) to a random quantity (like X), then the mean increases by the same amount, and the variance is unchanged, and the sd is unchanged.

• But what if we multiply or divide instead?

• If you multiply a random quantity (like X) by a constant positive number (like 2), then the mean is multiplied by the same amount, and the variance is multiplied by the square of the amount, and the sd is multiplied by the squareroot of the square, i.e. the absolute value, of the amount.

• e.g. if $E(X) = 5$ and $Var(X) = 4$ and $sd(X) = 2$, and $W = 3X$, then $E(W) = 3 \times 5 = 15$, and $Var(W) = 3^2 \times 4 = 36$, and $sd(W) = |3| \times 2 = 6$. Easy!

Standardised Variables (Z-Scores)

- Suppose X is any random quantity, with any mean m and variance v and sd $s = \sqrt{v}$.
- Then the quantity $X - m$ has mean 0, and variance v , and sd s .
- Then the quantity $(X - m)/s$ has mean $0/s = 0$, and variance $v/s^2 = v/v = 1$, and sd $\sqrt{1} = 1$.
- Conclusion: For any random quantity, if you subtract off its mean, and then divide by its sd, you get a quantity with mean=0, and sd=1.
 - So if it was approximately normal before, then now it is approximately standard normal, i.e. normal with mean 0 and var 1.
 - The standardised value is sometimes called a “Z-score”.
- Intuition: the Z-score measures “how many standard deviations is it above (or below) the mean”.

Standardised Variables (cont'd)

- Example: If we roll a fair die, then the result has mean 3.5, and variance $17.5/6 \doteq 2.917$, hence sd $\sqrt{17.5/6} \doteq 1.708$.
- Suppose on our next roll, we get a 6. What is the Z-score?
 - Answer: $(6 - 3.5)/1.708 \doteq 1.464$.
 - Interpretation: “It was about 1.464 standard deviations above the mean.”
- Suppose instead we get a 2. What is the Z-score?
 - Answer: $(2 - 3.5)/1.708 \doteq -0.878$.
 - Interpretation: “It was about 0.878 standard deviations below the mean.”
- This reduces all probabilities, to “standard” ones.
 - And, it reduces normal probabilities to standard normals!

Using Z-Scores to Compare: Elephants and Mice

- Suppose you meet a “large” African elephant which weighs 5000 kg, and a “large” white mouse which weighs 38 g.
- Relatively speaking, which is larger?
 - How to compare them??
- For a relative comparison, should use Z-scores!
 - Apparently(?) African elephants’ weights have mean 4550 kg, and standard deviation 150 kg.
 - And, white mice’s weights have mean 30 g, and standard deviation about 2 g.
- So, the elephant has $Z\text{-score} = (5000 - 4550)/150 = 3.0$.
- And, the mouse has $Z\text{-score} = (38 - 30)/2 = 4.0$.
- So, on a relative scale, the mouse is larger! Z-scores can provide a comparison between two very different quantities.

Converting to Standard Normal: Example

- Suppose a random quantity T is approximately normal, with mean 50 and sd 10. Then in terms of a standard normal, what is the probability that T is more than 70, i.e. what is $P[T > 70]$?

- Well, $P[T > 70] = P[(T - 50)/10 > (70 - 50)/10] = P[(T - 50)/10 > 2] \doteq P[Z > 2]$ (where Z is a standard normal random quantity, i.e. normal with mean 0 and var 1).

- So what? Well, for one thing, we can look up this probability in a standard normal table (e.g. on midterm/exam!), see www.probability.ca/sta130/normaltable.png.

- Need to use that: $P[Z > 2] = 1 - P[Z < 2] = 1 - P[Z \leq 2]$.

- From the table: $P[Z \leq 2] \doteq 0.9772$. So, $P[T > 70] \doteq P[Z > 2] \doteq 1 - 0.9772 = 0.0228$. Check in R: `pnorm(70, 50, 10, lower.tail=FALSE)`. Ans: 0.02275013. Good!

Standard Normal: RPS Example

• RPS games (again!). Suppose you win 59 out of 100 games. Does this show that you have skill?

– We know that the number correct (assuming $p = 1/2$) has mean = 50, variance = 25, and sd = 5.

– So, Z-score is: $(59-50)/5 = 9/5 = 1.8$. (Meaning?)

– P-value = $P[\text{would have won 59 or more games}] \doteq P[\text{normal with mean 50 and sd 5 would be 58.5 or more}] = P[\text{standard normal would be } (58.5-50)/5 \text{ or more}] = P[Z \geq (58.5 - 50)/5] = P[Z \geq 1.7] = 1 - P[Z \leq 1.7]$.

– From table: This is about $1 - 0.9554 = 0.0446$.

– Less than 0.05. So, statistically significant!

– (But a two-sided test would not be significant ... check ...)

Standard Normal: Dice Example

- Suppose (again) that you roll a fair die 60 times, and get “2” on 17 of the rolls. Too many 2’s?
- Know that number of 2’s has mean = $np = 60(1/6) = 10$, and variance = $np(1 - p) = 60(1/6)(5/6) = 25/3$, so sd = $\sqrt{25/3}$.
 - So, Z-score is $(17 - 10)/\sqrt{25/3} = 7/\sqrt{25/3} \doteq 2.425$.
 - And, $P[\text{number of “2”} \geq 17] \doteq P[\text{normal with mean 10 and sd } \sqrt{25/3} \text{ is } \geq 16.5] = P[\text{standard normal } \geq (16.5 - 10)/\sqrt{25/3}] \doteq P[\text{standard normal } \geq 2.25] = 1 - P[\text{standard normal } \leq 2.25] \doteq 1 - 0.9878 \text{ (from Table)} = 0.0122$. Less than 0.05! So, yes: too many 2’s!
- Tricks for standard normal table: if Z is standard normal, then $P[Z \leq -0.6] = 1 - P[Z \geq -0.6] = 1 - P[Z \leq +0.6] = 1 - 0.7257 = 0.2743$. (And so on.)

More Standard Normal Practice

• Suppose Z has standard normal probabilities. Using only the standard normal table and a calculator, what is $P[Z < -0.43]$?

– Well, by symmetry, $P[Z < -0.43] = P[Z > 0.43] = 1 - P[Z \leq 0.43] = 1 - 0.6664 = 0.3336$. Check in R?

– `pnorm(-.43, 0, 1, lower.tail=TRUE)` = 0.3335978. Good!

• Suppose X has normal probabilities with mean 80 and sd 6. With just table and calculator, what is $P[X > 90]$?

– Here Z has standard normal probabilities if $Z = (X - 80)/6$.

– Then, $P[X > 90] = P[(X - 80)/6 > (90 - 80)/6] = P[Z > (90 - 80)/6] \doteq P[Z > 1.67] = 1 - P[Z \leq 1.67] \doteq 1 - 0.9525 = 0.0475$.

– Check in R? `pnorm(90, 80, 6, lower.tail=FALSE)` = 0.04779035. Pretty good.

Comparing Polls: Marijuana Support in U.S.

CBS NEWS / April 20, 2015, 7:00 AM

Poll: Support for legal marijuana use reaches all-time high

Should Marijuana Use be Legal?

| | Now | 2/2014 | 4/2013 | 11/2012 | 10/2011 | 7/2009 | 7/1979 |
|-----|-----|--------|--------|---------|---------|--------|--------|
| Yes | 53% | 51% | 45% | 47% | 40% | 41% | 27% |
| No | 43% | 46% | 45% | 47% | 51% | 52% | 69% |

SOURCE: CBS News

- True increases? Just luck? How do we test?

Comparing Polls: Marijuana Support in U.S. (cont'd)

- 2012 poll (<http://www.cbsnews.com/news/poll-nearly-half-support-legalization-of-marijuana/>): out of 1,100 adults surveyed, 47% supported legalisation.
- 2014 poll (<http://www.cbsnews.com/news/majority-of-americans-now-support-legal-pot-poll-says/>): out of 1,018 adults surveyed, 51% supported legalisation.
- 2015 poll (<http://www.cbsnews.com/news/poll-support-for-legal-marijuana-use-reaches-all-time-high/>): out of 1,012 adults surveyed, 53% supported legalisation.
- Which differences are “real”, and which are “just luck”?
- Start with 2012 versus 2014 ...

Comparing Polls (cont'd)

- Write p_1 for the true fraction of support in 2012, and n_1 for the number of people surveyed in 2012, and T_1 for the total number of supporters among the surveyed people in 2012, and $\hat{p}_1 = T_1/n_1$ for the observed fraction of support in 2012. Similarly p_2 and n_2 and T_2 and \hat{p}_2 for 2014.
- Null hypothesis? $p_1 = p_2$. Alternative? $p_1 \neq p_2$.
- Now, T_1 is approximately normal, with mean $n_1 p_1$, and variance $n_1 p_1 (1 - p_1)$, so sd $\sqrt{n_1 p_1 (1 - p_1)}$.
- So, $\hat{p}_1 = T_1/n_1$ is approximately normal, with mean $n_1 p_1/n_1 = p_1$, and variance $n_1 p_1 (1 - p_1)/(n_1)^2 = p_1 (1 - p_1)/n_1$, and sd $\sqrt{p_1 (1 - p_1)/n_1}$.
- Similarly, \hat{p}_2 is approximately normal, with mean p_2 , and variance $p_2 (1 - p_2)/n_2$, and sd $\sqrt{p_2 (1 - p_2)/n_2}$.

Comparing Polls (cont'd)

- What about the negative of \hat{p}_1 , i.e. $-\hat{p}_1$? It has mean $-p_1$, and variance $p_1(1 - p_1)/n_1$ [not $-p_1(1 - p_1)/n_1$, because it involves squares, and $(-1)^2 = 1$]. So, $\text{sd} = \sqrt{p_1(1 - p_1)/n_1}$.
- What about the difference, $\hat{p}_2 - \hat{p}_1$?
 - Its mean is: $p_2 - p_1$. Unknown! But under the null hypothesis, it equals: 0.
 - And, its variance is:
$$\text{Var}(\hat{p}_2) + \text{Var}(-\hat{p}_1) = p_1(1 - p_1)/n_1 + p_2(1 - p_2)/n_2.$$
 - So, its sd is: $\sqrt{p_1(1 - p_1)/n_1 + p_2(1 - p_2)/n_2}$.
 - Note: not $\sqrt{p_1(1 - p_1)/n_1} + \sqrt{p_2(1 - p_2)/n_2}$. You add the variances, but not the standard deviations.
- So how does this help?

Comparing Polls (cont'd)

- We see that under the null hypothesis that $p_1 = p_2$, the quantity $\hat{p}_2 - \hat{p}_1$ should be approximately normal, with mean 0, and sd $\sqrt{p_1(1-p_1)/n_1 + p_2(1-p_2)/n_2}$.
- But even under the null hypothesis, p_1 and p_2 are unknown!
- Bold option: estimate sd by $\sqrt{\hat{p}_1(1-\hat{p}_1)/n_1 + \hat{p}_2(1-\hat{p}_2)/n_2}$.
- Conservative option: estimate sd by its maximum, i.e. $\sqrt{1/2(1-1/2)/n_1 + 1/2(1-1/2)/n_2} = \sqrt{1/n_1 + 1/n_2} / 2$.
- Then can compute a P-value under the null hypothesis.
 - The P-value should be the probability, assuming that $p_1 = p_2$, that $\hat{p}_2 - \hat{p}_1$ would be as “extreme” or more “extreme” than the observed value.
 - That is, the probability that $|\hat{p}_2 - \hat{p}_1|$ would be as large or larger than the observed value.

Back to the U.S. Marijuana Surveys

- In 2012, had $n_1 = 1,100$, and $\hat{p}_1 = 0.47$. In 2014, had $n_2 = 1,018$, and $\hat{p}_2 = 0.51$.
- So, the observed value of $|\hat{p}_2 - \hat{p}_1|$ was 0.04.
- Under the null hypothesis, that quantity should be approximately normal, with mean 0, and sd $\sqrt{p_1(1-p_1)/n_1 + p_2(1-p_2)/n_2}$.
- Conservative option: estimate sd as $\leq \sqrt{1/n_1 + 1/n_2} / 2 = \sqrt{1/1,100 + 1/1,018} / 2 \doteq 0.0217$.
- So the P-value should be the probability that a normal quantity with mean 0 and sd $\doteq 0.0217$ is as large or larger than 0.04, or as small or smaller than -0.04 . In R: `pnorm(0.04, 0, 0.0217, lower.tail=FALSE) + pnorm(-0.04, 0, 0.0217, lower.tail=TRUE)`. Answer = 0.0658433. [Or use table??] More than 0.05! So, we cannot conclude that more people in the U.S. supported legalisation of marijuana in 2014 than in 2012.

U.S. Marijuana Surveys (cont'd)

- Could perhaps try bold option instead. Would it help much? No, because $\hat{p}_1, \hat{p}_2 \approx 1/2$ anyway.
- Let's instead compare 2012 to 2015! In 2012, had $n_1 = 1,100$, and $\hat{p}_1 = 0.47$. In 2015, had $n_2 = 1,012$, and $\hat{p}_2 = 0.53$. So, the observed value of $|\hat{p}_2 - \hat{p}_1|$ was 0.06. Under the null hypothesis, that quantity should be approximately normal, with mean 0, and sd $\sqrt{p_1(1-p_1)/n_1 + p_2(1-p_2)/n_2}$. Conservative option: estimate sd as $\leq \sqrt{1/n_1 + 1/n_2} / 2 = \sqrt{1/1,100 + 1/1,012} / 2 \doteq 0.02177862$.
- So the P-value should be the probability that a normal quantity with mean 0 and sd $\doteq 0.02177862$ is as large or larger than 0.06, or as small or smaller than -0.06 . In R: `pnorm(0.06, 0, 0.02177862, lower.tail=FALSE) + pnorm(-0.06, 0, 0.02177862, lower.tail=TRUE)`. Answer = 0.005869294. Much less than 0.05 (or 0.01)! Conclusion: more people in the U.S. supported legalisation of marijuana in 2015 than in 2012.

Summary: Comparing Polls

- Suppose Poll #1 has sample size n_1 , and true probability p_1 , and observed fraction \hat{p}_1 . Similarly n_2 and p_2 and \hat{p}_2 for Poll #2.
- \hat{p}_1 is approximately normal, with mean p_1 , and sd $\sqrt{p_1(1-p_1)/n_1}$. Similarly \hat{p}_2 has mean p_2 and sd $\sqrt{p_2(1-p_2)/n_2}$.
- So, the difference, $\hat{p}_2 - \hat{p}_1$, has mean $p_2 - p_1$, and sd $\sqrt{p_1(1-p_1)/n_1 + p_2(1-p_2)/n_2}$. Can estimate this sd by $\sqrt{\hat{p}_1(1-\hat{p}_1)/n_1 + \hat{p}_2(1-\hat{p}_2)/n_2}$ (bold option) or as $\sqrt{1/n_1 + 1/n_2} / 2$ (conservative option).
- To test the null hypothesis $p_1 = p_2$, versus the alternative that $p_1 \neq p_2$, we compute the P-value as the probability that a normal with mean 0 (since if $p_1 = p_2$, then mean = $p_2 - p_1 = 0$), and sd from the above estimate, would be as large or larger than the observed difference, or as small or smaller than the negative of the observed difference.

Student Survey: Romantic Partners by Country

- Of the 44 students born in China, 15 (34.1%) have a romantic partner. ($\hat{p}_1 = 0.341$)
- Of the 15 students born in Canada, 4 (26.7%) have a romantic partner. ($\hat{p}_2 = 0.267$)
- Significant difference?
- Here P-value for the null hypothesis that $p_1 = p_2$, versus the alternative hypothesis that $p_1 \neq p_2$, is the probability that a normal with mean 0, and sd (bold option) $\sqrt{0.341(1 - 0.341)/44 + 0.267(1 - 0.267)/15} \doteq 0.135$, is as large or larger than $0.341 - 0.267 = 0.074$, or as small or smaller than -0.074 .
- So, is this P-value sufficiently small??

Student Survey (cont'd)

- Recall: here the P-value is the probability that a normal with mean 0, and $sd \doteq 0.135$, is as large or larger than 0.074, or as small or smaller than -0.074 .

- In R: `pnorm(0.074, 0, 0.135, lower.tail=FALSE) + pnorm(-0.074, 0, 0.135, lower.tail=TRUE)`. Answer = 0.5835902
Way more than 0.05! Cannot reject the null hypothesis!

- Or, use standard normal table! Here the P-value is the probability that a standard normal is as large or larger than $0.074/0.135 \doteq 0.548$, or as small or smaller than -0.548 . Table: $(1 - 0.7088) \times 2 \doteq 0.5824$. Pretty close (phew!). And, still way more than 0.05.

- Either way, cannot reject the null. Could be just luck!

Student Survey: Continuous Quantities

- Do male students have larger biceps than females?
- `mean(sdata$Q9a[sdata$Q1=="M"], na.rm=TRUE)`: 30.85128
- `mean(sdata$Q9a[sdata$Q1=="F"], na.rm=TRUE)`: 24.93077
- Significant? Let μ_1 be the true mean right bicep circumference of male students, and μ_2 for females.
 - Want to test the null hypothesis that $\mu_1 = \mu_2$, versus the alternative hypothesis that $\mu_1 \neq \mu_2$.
 - Here $\hat{\mu}_1 = 30.85128$ is the estimate of μ_1 from our sample, and $\hat{\mu}_2 = 24.93077$. So, $\hat{\mu}_2 - \hat{\mu}_1 \doteq 5.92$.
 - Then the P-value is: the probability, under the null hypothesis, of observing $\hat{\mu}_2 - \hat{\mu}_1 > 5.92$, or $\hat{\mu}_2 - \hat{\mu}_1 < -5.92$.
 - Under the null hypothesis, $\hat{\mu}_2 - \hat{\mu}_1$ would have mean 0. But what variance??

General Quantities, Besides Yes/No

- So far, we have mostly done statistics on Yes/No quantities. (Did you win at RPS? Do you support the government? Is the coin heads? Does the die show 5? Is the roulette spin 22? etc.)
- Then we could study proportions or fractions or probabilities, and compute means and variances and P-values for them, and compare them to each other, etc. Good!
- But what about quantities that don't involve just Yes/No? (Medicine: blood pressure, life span, weight gain, etc. Economics: GDP, stock price, company profits, etc. Social policy: number of accidents, amount of congestion, etc. Weather: amount of rain, wind speed, temperature, etc. Environment: global warming, ocean levels, contamination levels, atmospheric concentrations, etc. Sports: number of goals, time of possession, etc. Science: number of particles, speed of chemical reaction, etc. Students: biceps!)

Example: Baby Weights

- Ten babies born in a hospital (in North Carolina) had the following weights, in pounds: $x_1 = 9.88$, $x_2 = 9.12$, $x_3 = 8.00$, $x_4 = 9.38$, $x_5 = 7.44$, $x_6 = 8.25$, $x_7 = 8.25$, $x_8 = 6.88$, $x_9 = 7.94$, $x_{10} = 6.00$.
- What is the variance for the true mean baby weight?
 - It's not a proportion! Can't use previous formulas!
- Well, suppose the weight of babies is random, with some (unknown) mean μ , and some (unknown) sd σ , hence some (unknown) variance σ^2 . What can we say about μ ?
 - Well, we could estimate μ by the average of the data, i.e. by $\bar{x} = (x_1 + x_2 + \dots + x_{10})/10 = \frac{1}{n} \sum_{i=1}^n x_i \doteq 8.11$. (Here $n = 10$.)
 - But is this close to the true μ ? How close?
 - What variance??

Baby Weights (cont'd)

- Well, we could estimate the variance $\sigma^2 = E[(X_i - \mu)^2]$ by the average of the squared differences from \bar{x} , i.e. by $s^2 = [(x_1 - 8.11)^2 + (x_2 - 8.11)^2 + \dots + (x_{10} - 8.11)^2]/10 = \frac{1}{n} \sum_{i=1}^n (x_i - \bar{x})^2 \doteq 1.226$.

- Controversial! Most statisticians, and R, prefer to divide by $n - 1$, which has some advantages (e.g. “unbiased”; later). So, we’ll usually do that, though could instead divide by n . (See my article: www.probability.ca/varmse)

- If we do divide by $n - 1$, then we get $s^2 = [(x_1 - 8.11)^2 + (x_2 - 8.11)^2 + \dots + (x_{10} - 8.11)^2]/9 = \frac{1}{n-1} \sum_{i=1}^n (x_i - \bar{x})^2 \doteq 1.362$. In R: `bw = c(9.88,9.12,8.00,9.38,7.44,8.25,8.25,6.88,7.94,6.00)` and then `var(bw)` gives 1.362204.

- Then, if each x_i has variance s^2 , then $x_1 + x_2 + \dots + x_n$ has variance $s^2 + s^2 + \dots + s^2 = ns^2$, so $\bar{x} = (x_1 + \dots + x_n)/n$ has variance $ns^2/n^2 = s^2/n$.

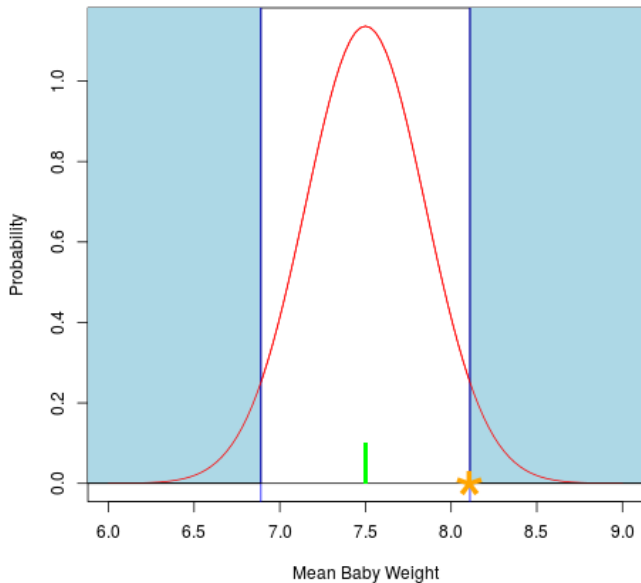
Testing “ μ ” for Baby Weights

- Can we do hypothesis tests for the unknown true mean μ ?
- Here $\bar{x} = (x_1 + x_2 + \dots + x_n)/n$ has mean $n\mu/n = \mu$ (same as mean of each x_i), and variance approximately s^2/n .
- Then $\bar{x} - \mu$ has mean 0, and variance s^2/n , hence sd $\sqrt{s^2/n} = s/\sqrt{n}$.
- So, $(\bar{x} - \mu)/(s/\sqrt{n})$ has mean 0, and sd 1.
- And, it's approximately normal (for reasonably large n), by the Central Limit Theorem.
- So, approximately a standard normal!
- Can use this to compute P-values!

Hypothesis Test for Baby Weights

- For the baby example, suppose want to test the null hypothesis that $\mu = 7.5$, versus the alternative hypothesis that $\mu \neq 7.5$.
- We know that if each x_i has mean μ , and variance s^2 , then $\bar{x} = (x_1 + x_2 + \dots + x_n)/n$ would have mean $\mu = 7.5$ and variance $s^2/n \doteq 1.362/10 = 0.1362$, hence sd $\sqrt{s^2/n} \doteq \sqrt{0.1362} \doteq 0.37$.
- But the observed value of \bar{x} was 8.11.
- So, the P-value is the probability, assuming that $\mu = 7.5$, that the value of \bar{x} would have been 8.11 or more, or 6.89 or less (two-sided) (since $8.11 = 7.5 + 0.61$, and $6.89 = 7.5 - 0.61$).
- And, under the null, \bar{x} has mean 7.5, and sd about 0.37.

Baby Weights when $\mu=7.5$: Diagram



- Recall: under the null, \bar{x} has mean 7.5, and sd about 0.37.
- So, the P-value is the probability that the random quantity \bar{x} , which is approximately normal(!), and has mean 7.5, and sd approximately 0.37, will be 8.11 or more, or 6.89 or less (two-sided).
- In R: `pnorm(8.11, 7.5, 0.37, lower.tail=FALSE) + pnorm(6.89, 7.5, 0.37, lower.tail=TRUE)`. Answer is: 0.09921964. More than 0.05! So, cannot reject the null! So, μ could indeed be 7.5!
- What if you didn't have R, only a standard normal probability table? Well, here the Z-score is $Z = (8.11 - 7.5)/0.37 \doteq 1.65$, so $P\text{-value} = P(Z > 1.65) + P(Z < -1.65) = (1 - P(Z < 1.65)) + (1 - P(Z < 1.65)) \doteq 2 \times (1 - 0.9505) = 0.0990$.
- Quite close!

- Let's try another test! For the baby example, suppose instead we want to test the null hypothesis that $\mu = 7.2$, versus the alternative hypothesis that $\mu > 7.2$ (one-sided).
- Under this null, \bar{x} has mean 7.2, and sd still about 0.37.
- So, the P-value is the probability that the random quantity \bar{x} , which is approximately normal with mean 7.2 and $\text{sd} \approx 0.37$, will be 8.11 or more. [Not or 6.29 or less, since just one-sided.]
- In R: `pnorm(8.11, 7.2, 0.37, lower.tail=FALSE)`. Answer is: 0.006957321. Much less than 0.05! So, can reject the null!
 - [Or, using table: $P\text{-value} = P(Z > (8.11 - 7.2)/0.37) = P(Z > 2.46) = 1 - P(Z < 2.46) \doteq 1 - 0.9931 \doteq 0.0069.$]
- Conclusion: Based on the ten baby weights studied, the true mean baby birth weight, μ , is more than 7.2 pounds.

Back to Biceps!

- `mean(sdata$Q9a[sdata$Q1=="M"], na.rm=TRUE)`: 30.85128
- `mean(sdata$Q9a[sdata$Q1=="F"], na.rm=TRUE)`: 24.93077
- If the male bicep measurements are x_1, x_2, \dots, x_n (where $n = 41$), then estimate their mean by $\bar{x} = \frac{1}{n} \sum_{i=1}^n x_i$, and then estimate their variance by $s_1^2 = \frac{1}{n-1} \sum_{i=1}^n (x_i - \bar{x})^2$. R does this all for us: `var(sdata$Q9a[sdata$Q1=="M"], na.rm=TRUE) = 16.18941`. Then $Var(\hat{\mu}_1) = n s_1^2 / (n)^2 = s_1^2 / n = 16.18941 / 41 \doteq 0.395$.
- Similarly, if the female bicep measurements are y_1, y_2, \dots, y_m , (where $m = 39$), then estimate the mean by $\bar{y} = \frac{1}{m} \sum_{i=1}^m y_i$, and then estimate the variance by $s_2^2 = \frac{1}{m-1} \sum_{i=1}^m (y_i - \bar{y})^2$. `var(sdata$Q9a[sdata$Q1=="F"], na.rm=TRUE) = 4.388502`. Then $Var(\hat{\mu}_2) = m s_2^2 / (m)^2 = s_2^2 / m = 4.388502 / 39 \doteq 0.113$.

Biceps: Hypothesis Tests

- Are male student biceps less than 32 cm on average?
- Under the null hypothesis, \bar{x} would have mean 32, and variance $\doteq 0.395$, so $\text{sd } \sqrt{0.395} \doteq 0.628$.
- Then the P-value is the probability that this would be less than the observed 30.85.
 - `pnorm(30.85, 32, 0.628, lower.tail=TRUE)`. Ans = 0.03353459. Less than 0.05. So, can reject the null.
 - Conclusion: On average, male right bicep circumferences are less than 32 cm.
- (Are female biceps larger than 24 cm? You check!)

Biceps: Comparisons

- Do male students have larger biceps than females, on average?
 - Observed difference = $\hat{\mu}_2 - \hat{\mu}_1 = 30.85128 - 24.93077 \doteq 5.92$.
- Variance? Similar to before, $Var(\hat{\mu}_2 - \hat{\mu}_1) = Var(\hat{\mu}_1) + Var(\hat{\mu}_2) = s_1^2/n + s_2^2/m \doteq 16.18941/41 + 4.388502/39 \doteq 0.5074$. So, $sd(\hat{\mu}_2 - \hat{\mu}_1) \doteq \sqrt{0.5074} \doteq 0.7123$.
- So, the P-value is (approximately) the probability that a normal random quantity with mean 0, and sd 0.7123, is larger than the observed value 5.92, or smaller than -5.92 .
 - $pnorm(5.92, 0, 0.7123, lower.tail=FALSE) + pnorm(-5.92, 0, 0.7123, lower.tail=TRUE) \doteq 9.5 \times 10^{-17}$. Significant? Yes! So, the data indicate that on average, male students do have larger biceps than female students.

Another Student Comparison

- What about student heights (Q6, in cm) by gender?
- There (check!) $\hat{\mu}_1 = 177.89$, and $\hat{\mu}_2 = 165.46$.
 - So, difference = 12.43.
- And, variance (check!) $\doteq 38.41/41 + 67.14/39 \doteq 2.66$.
- So, sd = $\sqrt{2.66} \doteq 1.63$.
- Then, P-value (check!) $\doteq 2.5 \times 10^{-14}$.
- So, males are significantly taller than females on average!
- And now a different general-quantity comparison ...

Comparing Baby Weights re Smoking

- QUESTION: Is the birth weight of a baby affected by whether or not the baby's mother smoked during pregnancy?
- Study from a social club in Kentucky:
(<http://www.statcrunch.com/app/index.php?dataid=467619>)
 - Birth weights (in grams) from the 22 babies whose mothers smoked: 3276, 1974, 2996, 2968, 2968, 5264, 3668, 3696, 3556, 2912, 2296, 1008, 896, 2800, 2688, 3976, 2688, 2002, 3108, 2030, 3304, 2912.
 - Birth weights (in grams) from the 35 babies whose mothers didn't smoke: 3612, 3640, 3444, 3388, 3612, 3080, 3612, 3080, 3388, 4368, 3612, 3024, 2436, 4788, 3500, 4256, 3640, 4256, 4312, 4760, 2940, 4060, 4172, 2968, 2688, 4200, 3920, 2576, 2744, 3864, 2912, 3668, 3640, 3864, 3556.
 - Conclusion??

Comparing Birthweights With or Without Smoking

- How can we compare them?
- Well, write x_1, x_2, \dots, x_{22} for the birthweights of the $n_1 = 22$ babies whose mothers smoked. And, write y_1, y_2, \dots, y_{35} for the birthweights of the $n_2 = 35$ babies whose mothers didn't smoke.
- Then can compute the means, $\bar{x} = \frac{1}{22} \sum_{i=1}^{22} x_i$, and $\bar{y} = \frac{1}{35} \sum_{i=1}^{35} y_i$. Obtain: $\bar{x} = 2863$, and $\bar{y} = 3588$.
- So, \bar{y} is larger, and in fact $\bar{y} - \bar{x} = 725$ grams.
- Does this prove anything? Or is it just ... luck?
- Write μ_1 for the true mean birthweight of babies whose mothers smoked, and μ_2 for those whose mothers didn't smoke.
 - Then what is the P-value to test the null hypothesis that $\mu_1 = \mu_2$, versus the alternative hypothesis $\mu_1 \neq \mu_2$ (two-sided), or $\mu_2 > \mu_1$ (one-sided)?

Comparing Birthweights (cont'd)

- Well, let's consider $\bar{y} - \bar{x}$.
 - This quantity has mean $\mu_2 - \mu_1$.
 - But what about the variance?
 - Here $s_1^2 = \frac{1}{n_1-1} \sum_{i=1}^{n_1} (x_i - \bar{x})^2 \approx \sigma_1^2 = 915128.7$.
 - And, $s_2^2 = \frac{1}{n_2-1} \sum_{i=1}^{n_2} (y_i - \bar{y})^2 \approx \sigma_2^2 = 356911.1$.
 - Then \bar{x} has variance $\approx s_1^2/n_1 \doteq 915128.7/22 \doteq 41597$.
 - And, \bar{y} has variance $\approx s_2^2/n_2 \doteq 356911.1/35 \doteq 10197$.
 - So, $\bar{y} - \bar{x}$ has var $\approx s_1^2/n_1 + s_2^2/n_2 \doteq 41597 + 10197 = 51794$.
 - So, $\bar{y} - \bar{x}$ has sd $\approx \sqrt{s_1^2/n_1 + s_2^2/n_2} \doteq \sqrt{51794} \doteq 228$.
- P-value?

Hypothesis Test for Comparing Birthweights

- Null hypothesis: $\mu_1 = \mu_2$, i.e. the two true means are equal.
- Under the null hypothesis, $\bar{y} - \bar{x}$ has mean $\mu_2 - \mu_1 = 0$, and sd $\approx \sqrt{s_1^2/n_1 + s_2^2/n_2} \doteq 228$ like before.
- So, the two-sided P-value is the probability that a normal, with mean 0, and sd 228, would be as large or larger than the observed value 725, or as small or smaller than -725 .
- In R: `pnorm(725, 0, 228, lower.tail=FALSE) + pnorm(-725, 0, 228, lower.tail=TRUE)`. Answer: 0.001473642. Much less than 0.05! So, can reject the null! So, μ_1 and μ_2 are not equal.
- Conclusion: The data demonstrates that the true mean birthweight for babies whose mother smokes, is not equal to the true mean birthweight for babies whose mother does not smoke. (Consistent with other, larger studies, e.g. of 34,799 births in Norway, and 347,650 births in Washington State.)

Another Example: Phone Calls

- Some students at Hope College (Michigan) surveyed 25 male and 25 female students. For each student, they checked how many seconds their last cell phone call was.
- Male data: 292, 360, 840, 60, 60, 900, 60, 328, 217, 1565, 16, 58, 22, 98, 73, 537, 51, 49, 1210, 15, 59, 328, 8, 1, 3.
- Female data: 653, 73, 10800, 202, 58, 7, 74, 75, 58, 168, 354, 600, 1560, 2220, 2100, 56, 900, 481, 60, 139, 80, 72, 2820, 17, 119.
- Do females talk on the phone for longer than males do?
- Note: one data value is much larger than all the others, namely 10800. This is exactly three hours. Perhaps(?) this was the default/max reading, and the phone had e.g. accidentally been left on? I decided to omit that value. (“outlier”) So, female data: 653, 73, 202, 58, 7, 74, 75, 58, 168, 354, 600, 1560, 2220, 2100, 56, 900, 481, 60, 139, 80, 72, 2820, 17, 119.

Phone Call Example (cont'd)

- Here $n_1 = 25$ and $n_2 = 24$.
- Then $\bar{x} = \frac{1}{25}(292 + 360 + \dots + 1 + 3) = 288.4$ seconds (nearly 5 minutes). And, $\bar{y} = \frac{1}{24}(653 + 73 + \dots + 17 + 119) \doteq 539.4$ seconds (about 9 minutes).
- Hence, $\bar{y} - \bar{x}$ has observed value $539.4 - 288.4 = 251.0$.
- Also, $s_1^2 = \frac{1}{25-1}[(292 - 288.4)^2 + (360 - 288.4)^2 + \dots + (1 - 288.4)^2 + (3 - 288.4)^2] \doteq 173070$, so $s_1 = \sqrt{173070} \doteq 416$.
- And, $s_2^2 = \frac{1}{24-1}[(653 - 539.4)^2 + (73 - 539.4)^2 + \dots + (17 - 539.4)^2 + (119 - 539.4)^2] \doteq 645153$, so $s_2 = \sqrt{645153} \doteq 803$.
- Then $\bar{y} - \bar{x}$ has sd $\approx \sqrt{s_1^2/n_1 + s_2^2/n_2} \doteq \sqrt{173070/25 + 645153/24} \doteq 183.9$.

Phone Call Example (cont'd)

- So, what is the P-value for the null hypothesis that the true means are equal, i.e. that $\mu_1 = \mu_2$, versus the alternative hypothesis that $\mu_1 < \mu_2$ (one-sided)?
- It is the probability that a normal random value with mean 0 seconds, and sd 183.9 seconds, is larger than the observed difference, i.e. than 251.0 seconds.
- In R: `pnorm(251, 0, 183.9, lower.tail=FALSE)`. Answer: 0.0861. Over 0.05! Cannot reject the null! So, μ_1 and μ_2 could be equal.
- (For two-sided test, would instead use `pnorm(251, 0, 183.9, lower.tail=FALSE) + pnorm(-251, 0, 183.9, lower.tail=TRUE)`. Answer: 0.1723. Much more than 0.05! So, still cannot reject.)
- Conclusion: the available data does not demonstrate that females talk on the phone longer than males do. (More data?)

Statistical Tests: Discussion

- So, now we know how to compare two different categories of general quantities, like biceps for males versus females (in addition to comparing two proportions, like fraction of Canadian versus Chinese students with romantic partners).
- But what about relationships between general quantities which do not divide into categories?
- For example, do students with larger biceps also tend to have larger wrists?
- Not a comparison of proportions. Not a comparison of two different categories of general quantities. Rather, a “relationship” between two general quantities.
- For this we need: correlation!