Correlation Example: Cricket Chirps

- Crickets make chirping sounds. (http://songsofinsects.com/crickets/striped-ground-cricket) Sometimes faster, sometimes slower. Question: Is the frequency of cricket chirps affected by the temperature?

- An old study (G.W. Pierce, "The Songs of Insects", 1948) measured the rate of chirps (pulses per second) 15 times, at different temperatures (in Celsius). The results were as follows:

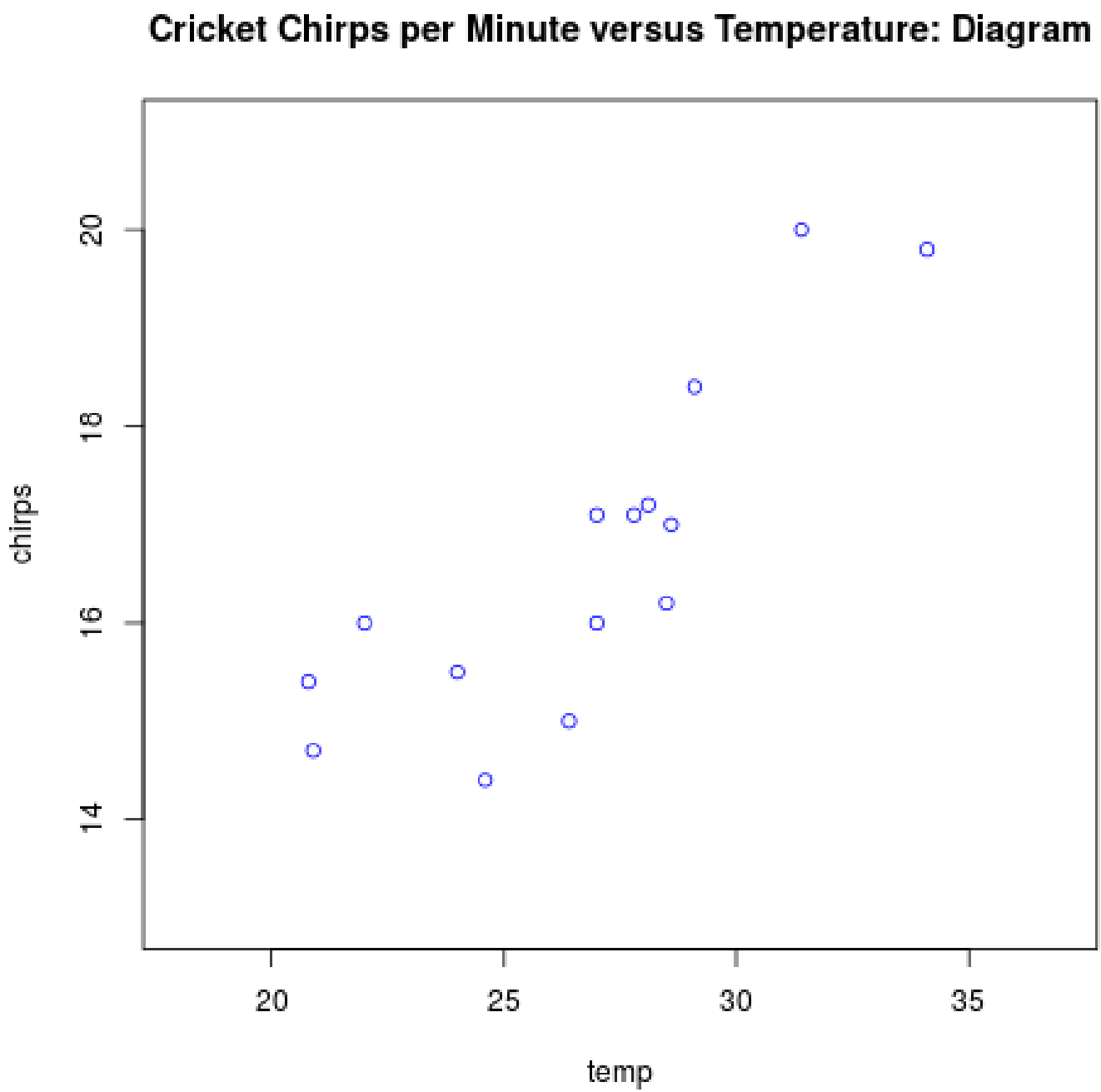
Temperature (C)	31.4	22.0	34.1	29.1	27.0	24.0	20.9
Chirps / Second	20.0	16.0	19.8	18.4	17.1	15.5	14.7

Temp	27.8	20.8	28.5	26.4	28.1	27.0	28.6	24.6
C/S	17.1	15.4	16.2	15.0	17.2	16.0	17.0	14.4

- Does this indicate that temperature affects chirps?
  - How can we test this??

Cricket Chirps (cont'd)

- Are these Yes/No proportions? No, they're general quantities.
- Can we compare two general samples? No, they're two different aspects of the same sample.
- Can any of our previous techniques be applied? Not really . . .
- So what to do?
- One strategy: plot all the values on a graph, of chirps/second versus temperature, to see if there is a pattern.
- Let's try it . . .


Cricket Chirps per Minute versus Temperature: Diagram
chirps
20
18
16
14
20
25
30
35
temp

Cricket Chirps (cont'd)

- So, is there a pattern?? Seems to be. How to test?

- Let X be the temperature (random), and let Y be the cricket chirps/second. We want to see if they are "related".

- First problem: X and Y are in different "units", on different "scales", with different means, different variances, etc. How to adjust them to be comparable? Solution: use Z-scores!

- Write μ_X for the true mean of X, and σ_X for the true sd of X. And μ_Y and σ_Y for Y.

- Then let $Z = (X - \mu_X)/\sigma_X$ be the Z-score for X. And, let $W = (Y - \mu_Y)/\sigma_Y$ be the Z-score for Y. Then Z and W are on the same "scale": they measure how many sd above (or below) the mean, for X and for Y, respectively.

- So now the question is, are Z and W related?

Cricket Chirps (cont'd)

- Question: Are Z and W related? That is, does increasing Z tend to increase (or decrease) W, or does it make no difference?
- Idea: Look at some expected values.
  - $E(Z) = 0$ (since it's a Z-score!). And $E(W) = 0$.
  - If Z and W had no relation (independent), then $E(ZW) = E(Z)\,E(W) = 0 \times 0 = 0$.
  - But if Z tends to get larger when W gets larger, and smaller when W gets smaller, then we might find that $E(ZW) > 0$.
  - Or, if Z tends to get smaller when W gets larger, and larger when W gets smaller, then we might find that $E(ZW) < 0$.
- So, we define the correlation between X and Y as:

$$\rho = \rho_{X,Y} = E(ZW) = E\left[\left(\frac{X-\mu_X}{\sigma_X}\right)\left(\frac{Y-\mu_Y}{\sigma_Y}\right)\right].$$

Estimating the Correlation

- Recall: the correlation **Cor**(X, Y) between X and Y is:

$$\rho = \rho_{X,Y} = E(ZW) = E\left[\left(\frac{X - \mu_X}{\sigma_X}\right)\left(\frac{Y - \mu_Y}{\sigma_Y}\right)\right].$$

  - Can we compute this value?

- Well, given a sample of values $x_1, x_2, \ldots, x_n$ for X, and corresponding sample $y_1, y_2, \ldots, y_n$ for Y, we could try to estimate the correlation as

$$\frac{1}{n}\sum_{i=1}^{n}\left(\frac{x_i - \mu_X}{\sigma_X}\right)\left(\frac{y_i - \mu_Y}{\sigma_Y}\right).$$

- The problem is: we don't know the true means μ_X and μ_Y, nor the true sd σ_X and σ_Y (or the true variances σ_X^2 and σ_Y^2).

- Solution: estimate them too!

Estimating the Correlation (cont'd)

- We can estimate the true means μ_X and μ_Y, by: $\mu_X \approx \overline{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$, $\mu_Y \approx \overline{y} = \frac{1}{n}\sum_{i=1}^{n} y_i$; and the true variances σ_X^2 and σ_Y^2, by: $\sigma_X^2 \approx s_x^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \overline{x})^2$, $\sigma_Y^2 \approx s_y^2 = \frac{1}{n-1}\sum_{i=1}^{n}(y_i - \overline{y})^2$.

- Then, the sample correlation between X and Y is

$$r = r_{xy} = \frac{1}{n-1}\sum_{i=1}^{n}\left(\frac{x_i - \overline{x}}{s_x}\right)\left(\frac{y_i - \overline{y}}{s_y}\right).$$

  – We know all these quantities from our sample. Good!

  – If we use the Z-scores $z_i = (x_i - \overline{x})/s_x$, and $w_i = (y_i - \overline{y})/s_y$, then we can write this more simply as: $r = \frac{1}{n-1}\sum_{i=1}^{n} z_i\, w_i$.

  – (Again, controversy over dividing by n versus $n-1$ ... but R divides by $n-1$, so we'll usually do that.)

Back to Cricket Data

Temperature (C)	31.4	22.0	34.1	29.1	27.0	24.0	20.9
Chirps / Second	20.0	16.0	19.8	18.4	17.1	15.5	14.7

Temp	27.8	20.8	28.5	26.4	28.1	27.0	28.6	24.6
C/S	17.1	15.4	16.2	15.0	17.2	16.0	17.0	14.4

- Write X for temperature, and Y for chirps/second. Then $\overline{x} = \frac{1}{n}\sum_{i=1}^{n} x_i = \frac{1}{15}[31.4 + 22.0 + \ldots + 24.6] \doteq 26.7$. And, $\overline{y} = \frac{1}{n}\sum_{i=1}^{n} y_i = \frac{1}{15}[20.0 + 16.0 + \ldots + 14.4] \doteq 16.7$.

- And, $s_x^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \overline{x})^2 = \frac{1}{14}[(31.4 - 26.7)^2 + (22.0 - 26.7)^2 + \ldots + (24.6 - 26.7)^2] \doteq 13.9$. So, $s_x = \sqrt{s_x^2} \doteq \sqrt{13.9} \doteq 3.7$. Also, $s_y^2 = \frac{1}{n-1}\sum_{i=1}^{n}(y_i - \overline{y})^2 = \frac{1}{14}[(20.0 - 16.7)^2 + (16.0 - 16.7)^2 + \ldots + (14.4 - 16.7)^2] \doteq 2.9$. So, $s_y = \sqrt{s_y^2} \doteq \sqrt{2.9} \doteq 1.7$.

- Then how to compute the sample correlation r? Take "the average of the products of the Z-scores". That is, ...

Cricket Data: Correlation

- For the cricket data,

$$r = r_{xy} = \frac{1}{n-1}\sum_{i=1}^{n} z_i w_i = \frac{1}{n-1}\sum_{i=1}^{n}\left(\frac{x_i-\overline{x}}{s_x}\right)\left(\frac{y_i-\overline{y}}{s_y}\right) =$$

$$\frac{1}{14}\left[\left(\frac{31.4-26.7}{3.7}\right)\left(\frac{20.0-16.7}{1.7}\right) + \left(\frac{22.0-26.7}{3.7}\right)\left(\frac{16.0-16.7}{1.7}\right) + \cdots\right.$$

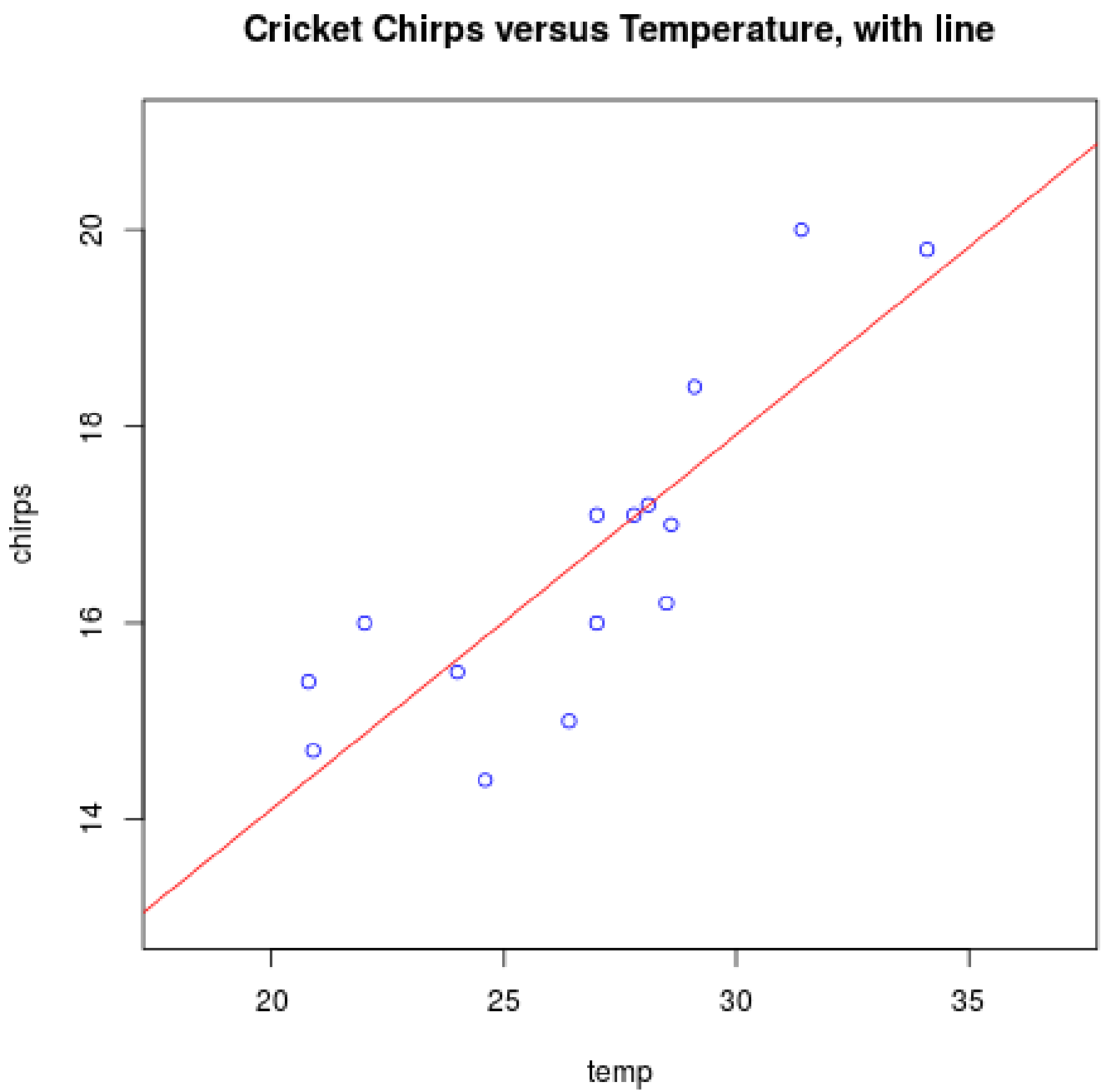
$$\left. + \left(\frac{24.6-26.7}{3.7}\right)\left(\frac{14.4-16.7}{1.7}\right)\right] \doteq 0.84.$$ Phew!

- So, the sample correlation is 0.84. This means that on average, every time the temperature increases by one standard deviation, the cricket chirp rate increases by 0.84 of its standard deviation.

  – That is, every time the temperature increases by s_x, the cricket chirp rate increases by $0.84\, s_y$.

  – Or, every time the temperature increases by one degree, the cricket chirp rate increases by $r_{xy}\, s_y/s_x = 0.84\, s_y/s_x$.

  – Can illustrate with "line of best fit" (more later) . . .


Cricket Chirps versus Temperature, with line
chirps
temp
14
16
18
20
20
25
30
35

Correlation: Discussion

- Conclusion so far: the sample correlation r_{xy} between the temperature in degrees celsius, and the rate of cricket chirps per second, is equal to 0.84.

- This means that the true correlation $\rho_{X,Y}$ between the temperature in degrees celsius, and the rate of cricket chirps per second, is probably: approximately 0.84.

- This means that the correlation between the temperature in degrees celsius, and the rate of cricket chirps per minute (not second), is also approximately 0.84. (Since correlation involves standardised variables, it is unaffected by e.g. multiplying everything by 60.)

- And, the correlation between the temperature in degrees fahrenheit, and the rate of cricket chirps per second, is also approximately 0.84. (Correlation is unaffected by adding any constants, or multiplying by any positive constants.)

Correlation Calculations in R

- Computing the sample correlation r_{xy} requires calculating lots of things: $\overline{x}, \overline{y}, s_x, s_y, z_i, w_i, r_{xy} = \frac{1}{n-1}\sum_{i=1}^{n} z_i w_i$.
  - Lots of work!
- R can do this automatically . . . with e.g. cor(temp, chirps). (Just like R can do mean, var, sd, etc.)
- So, in statistics applications, usually we don't need to do all this calculation by hand.
  - (But you might need to, for example, on an exam!)
- If we try cor(temp, chirps) in R, the answer is: 0.8360942.
  - Very close to 0.84.

Cricket Data: Correlation (cont'd)

- Is 0.84 a lot?
  - Well, the correlation is largest if Y is completely determined by X, e.g. when $Y = X$. In that case, $\rho_{X,Y} = E\left[\left(\frac{X-\mu_X}{\sigma_X}\right)\left(\frac{Y-\mu_Y}{\sigma_Y}\right)\right] = E\left[\left(\frac{X-\mu_X}{\sigma_X}\right)\left(\frac{X-\mu_X}{\sigma_X}\right)\right] = E\left[\left(\frac{X-\mu_X}{\sigma_X}\right)^2\right] = (1/\sigma_X^2)E\left[(X-\mu_X)^2\right] = (1/Var(X))\ Var(X) = 1.$
  - Summary: the largest possible correlation is: 1, which occurs if e.g. $Y = X$. (So, if correlation is near 1, then Y mostly increases with X.) Similarly, the smallest (i.e., most negative) possible correlation is: -1, which occurs if e.g. $Y = -X$. (So, if correlation is near -1, then Y mostly decreases when X increases.)
- So, yes, 0.84 seems like a lot. But does it actually demonstrate a correlation? Or, is it just ... luck?
  - How to test? What probabilities? Coming next! But first ...

Rough Guidelines for Interpreting Correlation

- How to interpret correlation? Hard to say; depends on context! Here's one suggestion, taken from: http://www.statstutor.ac.uk/resources/uploaded/pearsons.pdf

Range of r_{xy}	Relationship between X and Y
0.80 to 1.00	very strong positive correlation
0.60 to 0.79	strong positive correlation
0.40 to 0.59	moderate positive correlation
0.20 to 0.39	weak positive correlation
0.00 to 0.19	very weak positive correlation
−0.19 to −0.00	very weak negative correlation
−0.39 to −0.20	weak negative correlation
−0.59 to −0.40	moderate negative correlation
−0.79 to −0.60	strong negative correlation
−1.00 to −0.80	very strong negative correlation

- Rough guidelines only . . . debatable . . .

More Correlation Guidelines

- Or, here's another slightly different interpretation, taken from: https://explorable.com/statistical-correlation

Range of r_{xy}	Relationship between X and Y
0.50 to 1.00	strong positive correlation
0.30 to 0.50	moderate positive correlation
0.10 to 0.30	weak positive correlation
−0.10 to 0.10	none or very weak correlation
−0.30 to −0.10	weak negative correlation
−0.50 to −0.30	moderate negative correlation
−1.00 to −0.50	strong negative correlation

- Which interpretation is more correct? Hard to say! Some "judgement" is required.

Probabilities for Correlation

- Recall: For cricket chirps versus temperature, the sample correlation is $r_{xy} = 0.84$. (Strong positive correlation.) And, $r_{xy} = \frac{1}{n}\sum_{i=1}^{n} z_i w_i$, where $z_i = (x_i - \overline{x})/s_x$ and $w_i = (y_i - \overline{y})/s_y$ are the corresponding Z-scores.

- To draw statistical inferences about correlation, we need to know the probabilities for r_{xy}.

- Well, r_{xy} is an average of different products $z_i w_i$.

  – And, each such product has mean $E(z_i w_i) \approx E(ZW) = E\left[\left(\frac{X-\mu_X}{\sigma_X}\right)\left(\frac{Y-\mu_Y}{\sigma_Y}\right)\right]$, which equals $\rho_{X,Y}$, i.e. equals the true correlation between X and Y.

  – So, $E(r_{xy}) \approx \rho_{X,Y}$. That is, the sample correlation r_{xy} has mean approximately equal to the true correlation $\rho_{X,Y}$. (Just like how $\overline{x}$ has mean μ_X, and s_x has mean approximately σ_X.)

Probabilities for Correlation (cont'd)

- So r_{xy} has mean approximately $\rho_{X,Y}$. But what about the variance and sd of r_{xy}?

- First of all, what is $Var(z_i w_i)$? It should equal $Var(ZW)$. But what is that? Hard! Know $E(Z) = 0$ and $Var(Z) = 1$, but ...

- Assume for now that X and Y are actually independent, i.e. they do not affect each other at all. Then Z and W are also independent. Then the true correlation of X and Y is $\rho_{X,Y} = E(ZW) = E(Z)\,E(W) = (0)(0) = 0$.

  - In particular, $E(ZW) = \rho_{X,Y} = 0$, i.e. $\mu_{ZW} = 0$.

  - Then $Var(ZW) = E[(ZW - \mu_{ZW})^2] = E[(ZW - 0)^2] = E[(ZW)^2] = E[Z^2\,W^2] = E(Z^2)\,E(W^2) = (1)\,(1) = 1$.

- So, in the independent case, $E(z_i w_i) \approx 0$, and $Var(z_i w_i) \approx 1$.

Probabilities for Correlation (cont'd)

- Recall: if X and Y are independent, then each $z_i w_i$ has variance ≈ 1.
- Then $Var\left(\sum_{i=1}^{n} z_i w_i\right) \approx 1 + 1 + \ldots + 1 = n \times 1 = n$.
- So what about $Var(r_{xy})$? Well, $Var(r_{xy}) \approx Var\left(\frac{1}{n}\sum_{i=1}^{n} z_i w_i\right) = \frac{1}{n^2} Var\left(\sum_{i=1}^{n} z_i w_i\right) \approx \frac{1}{n^2}(n) = 1/n$.
- Summary: in the independent case, $Var(r_{xy}) \approx 1/n$. Hence, $sd(r_{xy}) \approx 1/\sqrt{n}$.
- FACT: Even if X and Y are not independent, still approximately $Var(r_{xy}) \approx 1/n$. (This is rather subtle, and there is no general formula. One approach is to consider the "Fisher transformation" $\text{arctanh}(r_{xy}) := \frac{1}{2}\ln\left(\frac{1+r_{xy}}{1-r_{xy}}\right)$, see R's cor.test, which has variance approximately $1/n$ in the general case. But still only approximate! So, let's not worry about this, and just use that $Var(r_{xy}) \approx 1/n$.)

P-Values for Correlation

- For the crickets example, suppose want to test the null hypothesis that $\rho_{X,Y} = 0$, versus the alternative hypothesis that $\rho_{X,Y} \neq 0$. (two-sided)

- We know that if $\rho_{X,Y} = 0$, then r_{xy} would have mean 0 and sd approximately $1/\sqrt{n} = 1/\sqrt{15} \doteq 0.258$. And approximately normal.

- But the observed value of r_{xy} was 0.84.

- So, the P-value is the probability that a normal random quantity, with mean 0, and sd $1/\sqrt{n} = 1/\sqrt{15}$, is 0.84 or more, or -0.84 or less (two-sided). In R: pnorm(0.84, 0, 1/sqrt(15), lower.tail=FALSE) + pnorm(−0.84, 0, 1/sqrt(15), lower.tail=TRUE). Answer is: 0.001140706.

- Much less than 0.05! Conclusion: The data indicates that the true correlation between temperature and cricket chirp rate is not zero. That is, they are "correlated".

"Correlation Does Not Imply Causation"

- (Mentioned on HW#2.) What does this mean?
- Just because two quantities are truly correlated (i.e., have non-zero true correlation), this does not necessarily mean that the second quantity is caused by the first quantity.
- Other possibilities include: the first quantity causes the second quantity ("reverse causation"); or the two quantities are both caused by some other quantity ("common cause"); or ...
- For cricket example: Does increased temperature cause the crickets to chirp more? Maybe. Other possibilities?
  - Perhaps cricket chirps cause temperature increase? (No!)
  - Perhaps both cricket chirps and temperature increase are caused by some other quantity? (Well, maybe, but what quantity? Perhaps ... sunlight! Except, crickets mostly chirp at night.)
  - So, probably(?) temperature increase causes chirps.

Causation Example: Drowning

- Suppose that in a certain city, the number of people who drown each day is positively correlated with the number of ice cream cones sold each day.

  – Possibility #1: Ice cream cones cause drowning! Surely not!

  – Possibility #2: Drowning causes people to buy ice cream! Surely not!

  – Possibility #3: Drowning and ice cream are both caused by something else. But by what?

  – Perhaps by warm, sunny weather, which makes more people go swimming, and makes more people buy ice cream!

  – Seems likely! Then have correlation, but not causation! How to test this? Could get additional data, about each day's weather, and the number of people who go swimming each day.

Causation Example: Yellow Fingers

- Suppose there is a positive correlation between people who get lung cancer, and people who have yellow stains on their finger.
  - Possibility #1: Yellow fingers cause lung cancer! Surely not!
  - Possibility #2: Lung cancer makes fingers yellow! Surely not!
  - Possibility #3: Lung cancer and yellow finger stains are both caused by something else. But by what?
  - Perhaps by smoking cigarettes, which definitely causes lung cancer, and which might also cause yellow stains on fingers (at least with old-style cigarette filters).
  - Seems likely! How to test? Perhaps change the cigarette filters to a different colour! (Tricky to arrange, over many years . . .)
- Many other similar examples. Have to think about (and explain) the meaning of a correlation.

Example: Ice Cream Sales

- A student monitored the weekly sales (in U.S. dollars), and average temperature (in degrees celsius), at a Southern California ice cream shop, for 12 consecutive weeks during the Summer of 2013.
  - TEMPERATURES (°C): 14.2, 16.4, 11.9, 15.2, 18.5, 22.1, 19.4, 25.1, 23.4, 18.1, 22.6, 17.2.
  - SALES (U.S. $): 215, 325, 185, 332, 406, 522, 412, 614, 544, 421, 445, 408.
- Is there a statistically significant correlation between the two?
- Let's check!
  - Compute the sample correlation! (Guesses?)

Example: Ice Cream Sales (cont'd)

- Let X be temperature, and Y be sales. Then

$\overline{x} = \frac{1}{12}[14.2 + 16.4 + \ldots + 17.2] \doteq 18.7$, and
$\overline{y} = \frac{1}{12}[215 + 325 + \ldots + 408] \doteq 402$. Then
$s_x^2 = \frac{1}{11}[(14.2-18.7)^2+(16.4-18.7)^2+\ldots+(17.2-18.7)^2] \doteq 16.09$,
so $s_x \doteq \sqrt{16.09} \doteq 4.01$. And,
$s_y^2 = \frac{1}{11}[(215-402)^2+(325-402)^2+\ldots+(408-402)^2] \doteq 15887$,
so $s_y \doteq \sqrt{15887} \doteq 126$.

- Hence, $r = r_{xy} = \frac{1}{n-1}\sum_{i=1}^{n} z_i w_i = \frac{1}{11}\sum_{i=1}^{12}\left(\frac{x_i-\overline{x}}{s_x}\right)\left(\frac{y_i-\overline{y}}{s_y}\right) =$
$\frac{1}{11}\left[\left(\frac{14.2-18.7}{4.01}\right)\left(\frac{215-402}{126}\right) + \left(\frac{16.4-18.7}{4.01}\right)\left(\frac{325-402}{126}\right) + \ldots\right.$
$\left. + \left(\frac{17.2-18.7}{4.01}\right)\left(\frac{408-402}{126}\right)\right] \doteq 0.957.$

- Extremely high positive correlation!

- So what can we conclude from this?

Example: Ice Cream Sales (cont'd)

- First Conclusion: Ice cream sales are positively correlated with temperature.

- But does this imply causation? That is, do higher temperatures cause higher ice cream sales?

- First consider other possible explanations:

  – Reverse causation? Perhaps ice cream sales cause higher temperatures? No, ice cream can't affect the temperature.

  – Common cause? I can't think of one . . .

  – Does causation make sense? Yes! Heat makes people hot and thirsty, so they might want more ice cream!

- So, in this case, I would say: Yes, this does imply causation, i.e. higher temperatures do cause people to buy more ice cream.

Example: Smoking and Wealth, by U.S. State

- I found data giving the percentage of adults who smoke, in each of the 50 U.S. states, in 2014; see www.probability.ca/sta130/statesmoke
- And I found their average income per capita in 2012; see www.probability.ca/sta130/stateincome
- Is there a correlation? Positive or negative? Strong or weak? Check in R (www.probability.ca/sta130/stateR). cor(sm,inc): -0.427. Moderate negative correlation! Statistically significant (check)! Why? Does smoking cause people to earn less (causation)? Do lower wages make people smoke more (reverse causation)? Are they both caused by some other factor (common cause)? If so, what other factor? Education?
  - I also found high school completion percentage in each U.S. state: www.probability.ca/sta130/statehigh cor(high,inc): 0.438. cor(high,sm): -0.335. Interpretation??

Confidence Intervals

- Recall that Trudeau Government poll: out of 1,500 Canadians surveyed, 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 it does not demonstrate that over 51% approve. (P-value = 0.06063039.)
- 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?)

Confidence Interval for the Trudeau Poll

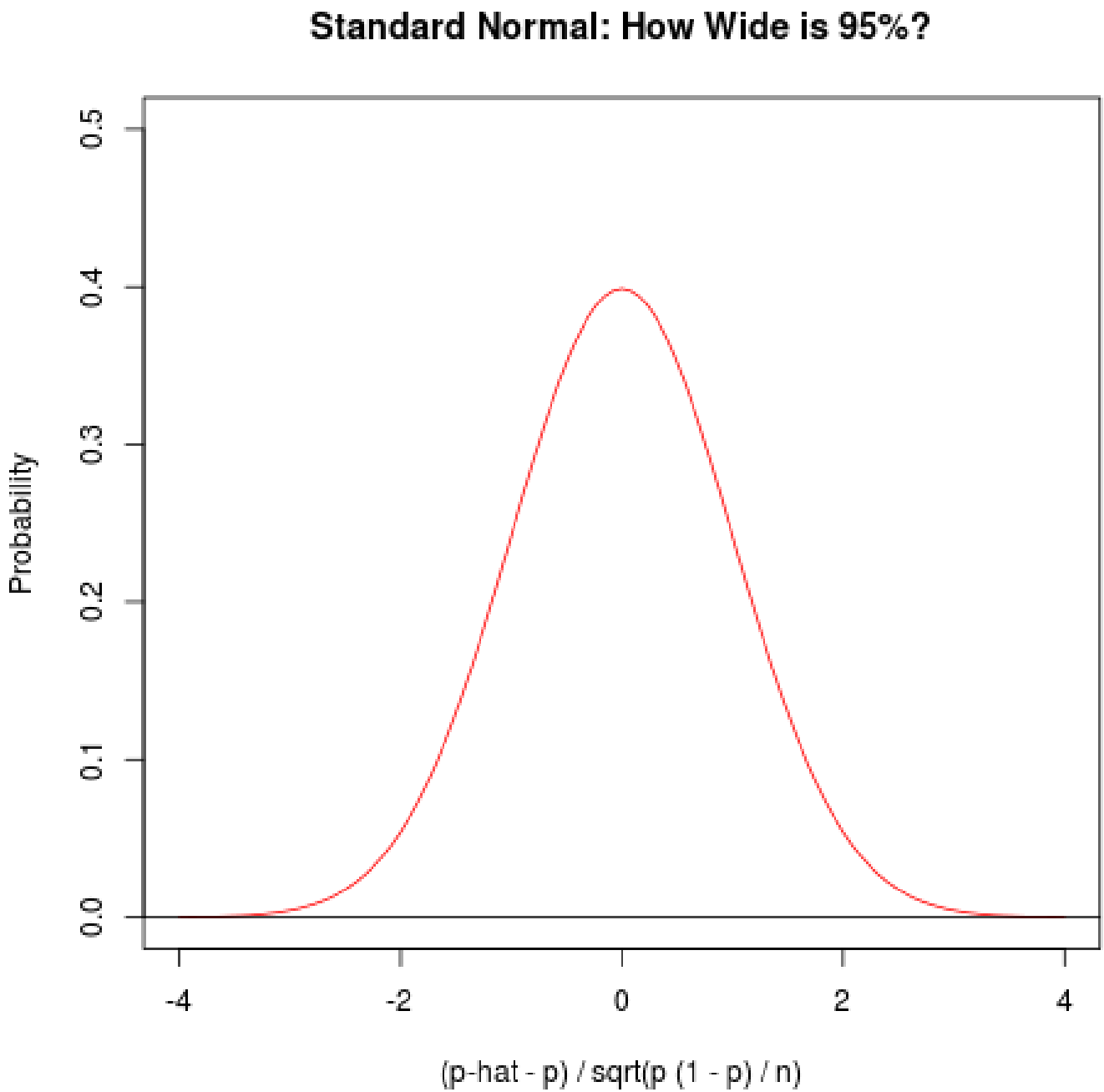
- Recall: of $n = 1,500$ people surveyed, 795 (53%) approve.

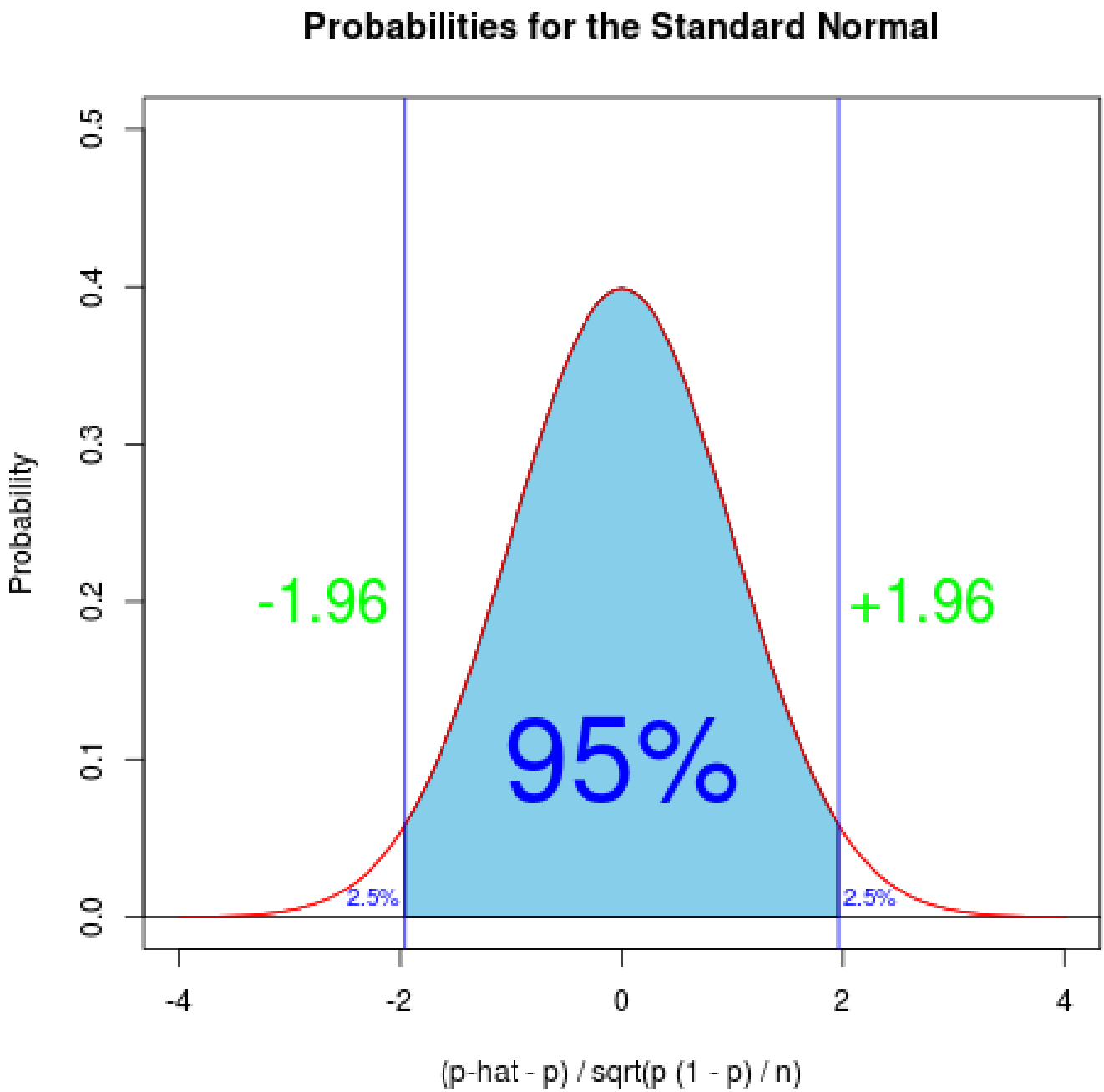
- Suppose the true fraction of all Canadians who approve is equal to some value p (unknown). What are the "plausible" values for p? 40%? (no!) 50%? (no!) 51%? (yes!) 50.5%? 57%??

- Let T = number of respondents (out of 1,500) who approve. Then T is approximately normal, with mean $= np = 1500\,p$, and variance $= np(1-p) = 1500\,p(1-p)$, so sd $= \sqrt{np(1-p)}$.

- Then if $\hat{p} = T/n = T/1,500$, then $\hat{p}$ is approximately normal, with mean $= np \;/\; n = p$, and sd $= \sqrt{np(1-p)} \;/\; n \;= \sqrt{p(1-p)/n}$. Then $\hat{p} - p$ has mean $= 0$, and sd $= \sqrt{p(1-p)/n}$. So $(\hat{p} - p)/\sqrt{p(1-p)/n}$ has mean $= 0$, and sd $= 1$. ("standard normal")

- For the standard normal, we can find the "plausible" values by measuring. What is a 95% range? Let's see ...


Standard Normal: How Wide is 95%?
Probability
0.0
0.1
0.2
0.3
0.4
0.5
-4
-2
0
2
4
(p-hat - p) / sqrt(p (1 - p) / n)


Probabilities for the Standard Normal
Probability
0.0
0.1
0.2
0.3
0.4
0.5
-1.96
+1.96
95%
2.5%
2.5%
-4
-2
0
2
4
(p-hat - p) / sqrt(p (1 - p) / n)

Confidence Interval for Trudeau Poll (cont'd)

- That is: the standard normal has probability about 95% of being between -1.96 and $+1.96$.
  - Check: pnorm(+1.96,0,1,lower.tail=TRUE) − pnorm(-1.96,0,1,lower.tail=TRUE) equals 0.9500042. good!
  - (What if we wanted probability 99%? Replace "1.96" by about: 2.58. Then get: 0.99012.)
- So, if Z has the standard normal distribution (i.e., probabilities), then $P[-1.96 < Z < +1.96] \doteq 0.95$.
So, apply this to: $Z = (\hat{p} - p)/\sqrt{p(1-p)/n}$
- Conclusion: $P[-1.96 < (\hat{p} - p)/\sqrt{p(1-p)/n} < +1.96] \doteq 0.95$.
So, $P[-1.96\sqrt{p(1-p)/n}) < \hat{p} - p < +1.96\sqrt{p(1-p)/n}] \doteq 0.95$.
So, $P[\hat{p} - 1.96\sqrt{p(1-p)/n}) < p < \hat{p} + 1.96\sqrt{p(1-p)/n}] \doteq 0.95$.
  - That is, with probability 95%, p is within $\pm 1.96\sqrt{p(1-p)/n}$ of the observed fraction $\hat{p}$. This is what we want!

Confidence Interval for Trudeau Poll: Conclusion?

- With probability 95%, p is within $\pm 1.96\sqrt{p(1-p)/n}$ of $\hat{p}$.
  - Here $1.96\sqrt{p(1-p)/n}$ is the "margin of error" (MOE).
  - And, the interval from $\hat{p} - 1.96\sqrt{p(1-p)/n)}$ to $\hat{p} + 1.96\sqrt{p(1-p)/n}$ is the "(95%) confidence interval".
- Good! Just one problem: "p" is unknown! So we can't calculate this margin of error! Bad! What to do?
- Bold Option: replace p (unknown) by $\hat{p}$. Get $1.96\sqrt{\hat{p}(1-\hat{p})/n} = 196\%\sqrt{\hat{p}(1-\hat{p})/n}$.
- Conservative Option: replace p by $1/2$. Get $1.96\sqrt{\frac{1}{2}(1-\frac{1}{2})/n} = 1.96\sqrt{1/n}/2 = 0.98/\sqrt{n} = 98\%/\sqrt{n}$.
- Do polls use this??

Confidence Interval for Trudeau Poll: Conclusion

- Trudeau poll: $\hat{p} = 0.53$, and $n = 1500$.
  - So, $98\%/\sqrt{n} = 98\%/\sqrt{1500} \doteq 2.530349\%$.
- Or, could use the bold option:
$196\%\sqrt{\hat{p}(1-\hat{p})/n} = 196\%\sqrt{0.53(1-0.53)/1,500} \doteq 2.52579\%$.
Very similar! (since $\hat{p} \approx 0.5$ in this case ...)
- What do they claim? "The margin of error for a comparable probability-based random sample of the same size is +/- 2.6%, 19 times out of 20." Check: $(98\%)/\sqrt{1,500} \doteq 2.530349\% \doteq 2.6\%$.
  - Yep! (Though "rounded up", to be safe ...)
  - Conclude that true support p is between $53\% - 2.6\% = 50.4\%$, and $53\% + 2.6\% = 55.6\%$. Confidence interval is: [50.4%, 55.6%].
- Other polls?

POLITICS August 19, 2015 2:46 pm

Majority of Canadians support decriminalizing marijuana: poll

Ipsos found 65 per cent of Canadians (29 per cent strongly/36 per cent somewhat) support the decriminalization of marijuana so that possession of small amounts would no longer carry a penalty or fine.

was conducted between August 13 and August 17, with a sample of 1,000 Canadians, from Ipsos' online panel. The poll reported above is accurate to within 3.5 percentage points 19 times out of 20.

- Check: $98\%/\sqrt{n} = 98\%/\sqrt{1000} \doteq 3.099\% \doteq 3.1\%$.

(Which they rounded up to 3.5%.) Or, use bold option: $196\%\sqrt{\hat{p}(1-\hat{p})/n} = 196\%\sqrt{0.65(1-0.65)/1,000} \doteq 2.9563\%$. (A bit smaller. They didn't use that.) Let's use 3.1%.

- Then, 95% confidence interval is: [61.9%, 68.1%].

But Do Poll Companies Really Use This?

- To check this, I went to the main web site for Forum Research Inc., a leading Canadian pollster.

- For information about margins of error, they refer you to this web page:

  http://www.forumresearch.com/tools-margin-of-error.asp

- That page gives various margins of error, based on the Sample Size (n) and the Observed Proportion ($\hat{p}$).

- Does it follow our formula? Let's check. Oh yeah!

- (In practice, they usually use the "Observed Proportion = 50%" row – corresponding to the ... conservative option.)

- Can we use margins of error and confidence intervals for other experiments besides polls? Yes! e.g. RPS ...

- Suppose you play RPS 100 times, and win 69 times. Then $\hat{p} = 0.69$, and margin of error $= 98\%/\sqrt{n} = 98\%/\sqrt{100} = 9.8\% = 0.098$ (conservative option), or $1.96\sqrt{\hat{p}(1-\hat{p})/n} = 1.96\sqrt{0.69 \times 0.31/100} \doteq 0.090$ (bold). So, 95% confidence interval (conservative option) for p is: [0.592, 0.788]. Or, bold option: [0.600, 0.780].

- Or, if you play 1000 times, and win 550 times: $\hat{p} = 0.55$, and 95% M.O.E. (cons.) $= 0.98/\sqrt{1000} \doteq 0.031$, and 95% conf. int. (cons.) $= [0.519, 0.581] \doteq [0.52, 0.58] = [52\%, 58\%]$. (Not 50%!)

- Or, if you play 20 times, and win 14 times: $\hat{p} = 0.70$, and 95% M.O.E. (cons.) $= 0.98/\sqrt{20} \doteq 0.22$, and 95% conf. int. (cons.) $= [0.48, 0.92]$. Or, bold option: M.O.E. $= 1.96\sqrt{0.7 \times 0.3/20} \doteq 0.20$, conf. int. $= [0.50, 0.90]$.

- Boy or Girl? Of $n = 377,636$ births, 184,049 were female, so $\hat{p} = 184,049/377,636 \doteq 0.4874$. M.O.E. (cons.) $= 0.98/\sqrt{n} \doteq 0.0016$. So, conf. int. $= [0.4858, 0.4890]$.

Confidence Intervals for General Quantities

- For proportions, have confidence intervals $\hat{p} \pm 1.96\sqrt{p(1-p)/n}$ (bold option) or $\hat{p} \pm 0.98/\sqrt{n}$ (conservative option).
- What about general quantities (like Baby Weights, etc.)?
- Can we use a conservative option? No! No upper bound!
- Can we use a bold option? Yes!
- Need to replace "$\hat{p}$" with a more general estimate of the true mean. We know how to do this: use $\overline{x}$ instead!
- And, we need to replace "$p(1-p)/n$" with a more general estimate of the variance of $\overline{x}$. We know how to do this, with s^2/n instead! So, we estimate variance of $\overline{x}$ by s^2/n.
- So, for a general quantity, the 95% confidence interval is: $[\overline{x} - 1.96\sqrt{s^2/n},\ \overline{x} + 1.96\sqrt{s^2/n}] = [\overline{x} - 1.96s/\sqrt{n},\ \overline{x} + 1.96s/\sqrt{n}]$. Easy!

- Baby Weights (North Carolina) example: $n = 10$.

Found that $\overline{x} = 8.11$ pounds, and $s^2 = 1.362$. So, confidence interval is: $[\overline{x} - 1.96\sqrt{s^2/n},\ \overline{x} + 1.96\sqrt{s^2/n}] = [8.11 - 1.96\sqrt{1.326/10},\ 8.11 + 1.96\sqrt{1.326/10}] = [7.40, 8.82]$ pounds. (So, here MOE $= 1.96\sqrt{1.326/10} \doteq 0.71$.)

  - Conclusion: we are 95% confident that the true mean baby weight in North Carolina is somewhere between 7.40 and 8.82 pounds.

- Student biceps (female): $n = 39$. Found that $\overline{x} = 24.93$ cm, and $s^2 = 4.39$. So, confidence interval is: $[\overline{x} - 1.96\sqrt{s^2/n},\ \overline{x} + 1.96\sqrt{s^2/n}] = [24.93 - 1.96\sqrt{4.39/39},\ 24.93 + 1.96\sqrt{4.39/39}] \doteq [24.3, 25.6]$ cm.

  - Conclusion: we are 95% confident that the true mean circumference of female students' biceps is somewhere between 24.3 and 25.6 cm.

Confidence Intervals for Differences of Proportions

- Suppose Poll #1 has sample size n_1, and observed proportion $\hat{p}_1$. Similarly n_2 and $\hat{p}_2$ for Poll #2.
- What is 95% confidence interval for true difference $p_2 - p_1$?
- We can estimate $p_2 - p_1$ by $\hat{p}_2 - \hat{p}_1$, and estimate its sd by $\sqrt{\hat{p}_1(1-\hat{p}_1)/n_1 + \hat{p}_2(1-\hat{p}_2)/n_2}$ (bold option), or $\sqrt{1/n_1 + 1/n_2}\,/2$ (conservative option).
- So, 95% confidence interval is
$[\hat{p}_2 - \hat{p}_1 - 1.96\sqrt{\hat{p}_1(1-\hat{p}_1)/n_1 + \hat{p}_2(1-\hat{p}_2)/n_2},$
$\hat{p}_2 - \hat{p}_1 + 1.96\sqrt{\hat{p}_1(1-\hat{p}_1)/n_1 + \hat{p}_2(1-\hat{p}_2)/n_2}]$ (bold), or
$[\hat{p}_2 - \hat{p}_1 - 0.98\sqrt{1/n_1 + 1/n_2},\ \hat{p}_2 - \hat{p}_1 + 0.98\sqrt{1/n_1 + 1/n_2}]$
(conservative).
- Example: U.S. Marijuana Surveys of 2012 and 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 (i.e., 6%).
- True sd of $\hat{p}_2 - \hat{p}_1$ is $\sqrt{p_1(1-p_1)/n_1 + p_2(1-p_2)/n_2}$.
- Conservative option: estimate sd as $\sqrt{1/n_1 + 1/n_2}\,/\,2 = \sqrt{1/1,100 + 1/1,012}\,/\,2 \doteq 0.022$.
  - So, 95% confidence interval for $p_2 - p_1$ is $[0.06 - 1.96 \times 0.022,\ 0.06 + 1.96 \times 0.022] \doteq [0.017,\ 0.103]$.
  - Conclusion: We are 95% confident that the true U.S. support for legalising marijuana in 2015, minus the support in 2012, is between 0.017 and 0.103, i.e. is between 1.7% and 10.3%.
- Or, bold option: estimate sd as $\sqrt{\hat{p}_1(1-\hat{p}_1)/n_1 + \hat{p}_2(1-\hat{p}_2)/n_2} = \sqrt{0.47(1-0.47)/1,100 + 0.53(1-0.53)/1,012} \doteq 0.0217$. Almost exactly the same (since $\hat{p}_1, \hat{p}_2 \approx 1/2$).

Confidence Intervals for General Differences

- Suppose we have general data sets $x_1, \ldots, x_n$ and $y_1, \ldots, y_m$, with true means μ_1 and μ_2
- What is confidence interval for the difference $\mu_2 - \mu_1$?
- Here we estimate the means by $\overline{x}$ and $\overline{y}$, and estimate variance by s_1^2 and s_2^2, respectively.
- Then can estimate $\mu_2 - \mu_1$ by $\overline{y} - \overline{x}$, and estimate its variance by $s_1^2/n_1 + s_2^2/n_2$.
- So, a 95% confidence interval for $\mu_2 - \mu_1$ is $[\overline{y} - \overline{x} - 1.96\sqrt{s_1^2/n_1 + s_2^2/n_2},\ \overline{y} - \overline{x} + 1.96\sqrt{s_1^2/n_1 + s_2^2/n_2}]$.
- Example: Confidence interval for phone call differences ...

- Recall: data for phone call lengths (in seconds) of male and female students at Hope College (Michigan).

- Males: $n_1 = 25$, $\overline{x} = 288.4$ (seconds), $s_1^2 = 173070$. Females: $n_2 = 24$, $\overline{y} = 539.4$, $s_2^2 = 645153$.

- So, 95% confidence interval is $[\overline{y} - \overline{x} - 1.96\sqrt{s_1^2/n_1 + s_2^2/n_2},\ \overline{y} - \overline{x} + 1.96\sqrt{s_1^2/n_1 + s_2^2/n_2}] = [539.4 - 288.4 - 1.96\sqrt{173070/25 + 645153/24},\ [539.4 - 288.4 + 1.96\sqrt{173070/25 + 645153/24}] \doteq [-109, 611]$.

- Conclusion: We are 95% confident that the difference between the mean time female students talk on the phone, minus the mean time male students talk on the phone, is between -109 and 611 seconds, i.e. between about -1.8 and 10.2 minutes.

- That is, we are 95% confident that on average, female students talk on the phone between 1.8 minutes less than male students, and 10.2 minutes more than male students. (More data?)

Confidence Intervals for Correlations

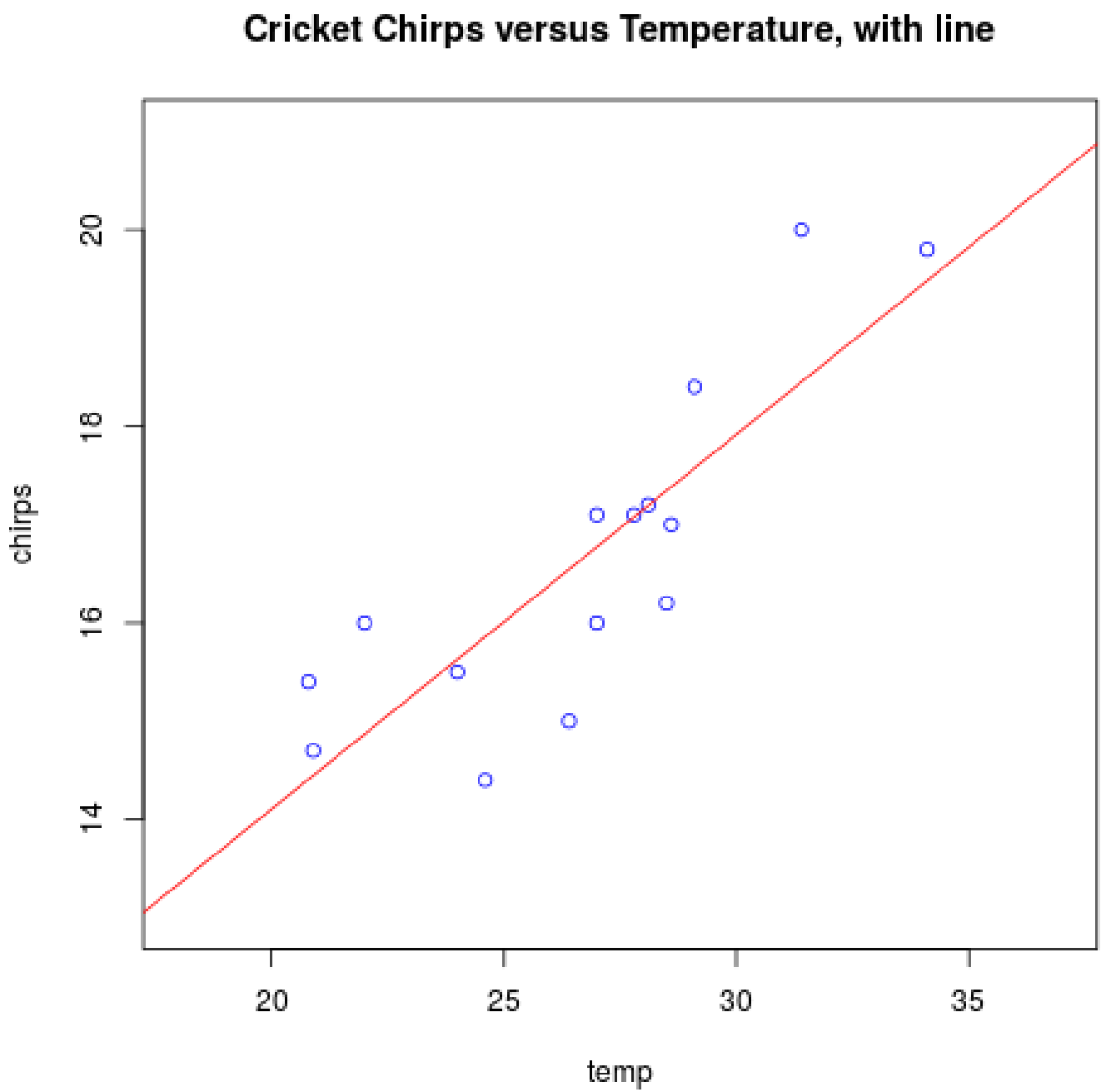
- Recall: for the cricket data, we found the correlation between temperature and chirp rate was $r = 0.84$.
- Confidence interval?
- Here we estimated variance as: $1/n = 1/15$.
- So, confidence interval is $[r - 1.96\sqrt{1/n},\ r + 1.96\sqrt{1/n}] = [0.84 - 1.96/\sqrt{15},\ 0.84 + 1.96/\sqrt{15}] \doteq [0.33,\ 1.35]$.
- But correlation is always ≤ 1, so we could replace this confidence interval by: $[0.33,\ 1]$.
- Conclusion: We are 95% confident that the true correlation between temperature and cricket chirp rate is somewhere between 0.33 and 1, i.e. is more than 0.33.
  - (i.e., moderately positive to strongly positive . . .)

Lucky Correlations?

- Suppose two quantities X and Y have a sample correlation which is far from 0, with P-value < 0.05. Then perhaps:
  - X causes Y? (directly or indirectly)
  - Y causes X?
  - X and Y are both caused by a third quantity?
  - It's still just luck! Could it be??
- Example: http://tylervigen.com/spurious-correlations Huh?
  - Would we have P-value < 0.05 in these cases? Yep!
  - But still "spurious". Why? They tested too many correlations before finally finding a significant one! "Multiple testing (comparisons) problem". What to do? Demand smaller P-values? Do follow-up studies? Challenging!
- See also: https://imgs.xkcd.com/comics/significant.png

Another Perspective: Regression

- (Actually "simple linear regression", also called "ordinary least squares (OLS) regression", or the "line of best fit".)
- Suppose the quantities X and Y have true correlation ρ.
- Then $E(ZW) = \rho$, i.e. $E\left[\left(\frac{X-\mu_X}{\sigma_X}\right)\left(\frac{Y-\mu_Y}{\sigma_Y}\right)\right] = \rho$.
- Intuitively, this means that $W = \rho Z + L$, where L is "leftover" randomness, independent of Z and X, with mean 0.
- That is, $\frac{Y-\mu_Y}{\sigma_Y} = \rho\left(\frac{X-\mu_X}{\sigma_X}\right) + L$.
- Solving, $Y = (\rho\,\sigma_Y/\sigma_X)X + (\mu_Y - \mu_X\rho\sigma_Y/\sigma_X) + \sigma_Y L$.
- That is, $Y = \beta_1 X + \beta_0 + e$, where: $\beta_1 = \rho\,\sigma_Y/\sigma_X$ ("regression coefficient"), $\beta_0 = \mu_Y - \mu_X\rho\,\sigma_Y/\sigma_X$ ("intercept"), and $e = \sigma_Y L$ ("error term"; mean=0). Approximate this by $Y = b_1 X + b_0$, where $b_1 = r_{xy}\,s_y/s_x$, and $b_0 = \overline{y} - \overline{x}r_{xy}s_y/s_x$. This is the same line of best fit as before!


Cricket Chirps versus Temperature, with line
chirps
temp
14
16
18
20
20
25
30
35

Coefficient of Determination

- Recall: $Y = \beta_1 X + \beta_0 + e$, where $\beta_1 = \rho\,\sigma_Y/\sigma_X$, and $\beta_0 = \mu_Y - \mu_X \rho\,\sigma_Y/\sigma_X$ is some constant, and e is independent of X with mean 0. (Check: $\mu_Y = E(Y) = \beta_1\mu_X + \beta_0 + 0$? Yep!)
- From this formula, $Var(Y) = (\beta_1)^2 Var(X) + 0 + Var(e)$.
- Question: How much of $Var(Y)$ is "explained" or "caused" by changes in X? Well, $(\beta_1)^2 Var(X)$ of it.
- So, what fraction of $Var(Y)$ is "explained" by changes in X? Well, a fraction $[(\beta_1)^2\, Var(X)]/Var(Y) = [(\rho\,\sigma_Y/\sigma_X)^2\,\sigma_X^2]/\sigma_Y^2 = \rho^2$. Approximate this by $(r_{xy})^2$, i.e. by r^2.
- Definition: The "coefficient of determination", when regressing Y against X, is given by r^2 ("R squared"). It measures how well Y is "explained" by X, i.e. how well the line fits the data. Minimum possible value is 0, maximum is 1. Crickets: $r^2 = (0.84)^2 \doteq 0.71$ (pretty large, i.e. temperature "explains" chirps pretty well).

Regression's "Least Squares" Property

- Recall our regression "line of best fit": $Y = b_1 X + b_0$, where $b_1 = r\, s_y/s_x$, and $b_0 = \overline{y} - \overline{x} s_y/s_x$. Why these b_1 and b_0?
  - Suppose we used some line, $Y = aX + c$. ("linear model")
  - Then for each data value x_i, this model would "predict" a corresponding Y value of $Y = ax_i + c$.
  - But the "real" corresponding data value is y_i.
  - So, we want $ax_i + c$ to be close to y_i.
  - The sum of squares of the errors is : $\sum_{i=1}^{n} (y_i - ax_i - c)^2$.
- FACT: The choices $a = b_1$ and $c = b_0$ (as above) are the choices which minimise this sum of squares of errors.
  - "ordinary least squares estimate" (OLS)
- See also R's function lm, e.g. lm(chirps $\sim$ temp).

Multiple Regression

- Sometimes a quantity Y might depend on multiple other quantities $X_1, X_2, \ldots, X_p$, not just a single X.
  - We can still compute **Cor**(Y, X_1), **Cor**(Y, X_2), etc.
  - But if the different X_i depend on each other, then the interpretation of these correlations gets complicated.
- Use multiple regression: $Y = \beta_1 X_1 + \beta_2 X_2 + \ldots + \beta_p X_p + \beta_0 + e$, where again e has mean 0. (If $p = 1$, then it's the same as before.)
- Can again find estimates b_j of the coefficients β_j from the data, by minimising the sum of squares. Requires multivariable calculus. We'll just trust R's lm function for this! Interpretation?
- U.S. Smoking/Wealth again (www.probability.ca/sta130/stateR). Try lm(sm∼inc), and lm(sm∼high), and lm(sm∼inc+high) (perhaps with summary(...)). What can we conclude??

- More Examples: Student data! Relationships between height (Q6), bicep (Q9b), wrist (Q8b)? www.probability.ca/sta130/studentdata.txt

- Countries: www.probability.ca/sta130/countrydata.txt

Try various correlations (cor) and linear regressions (lm). Values? coefs? sd? R^2? Interpretation? Causation? www.probability.ca/sta130/countryR

- www.probability.ca/sta130/SAT.txt SAT scores (verbal and math) by state, plus percent taking SATs, and teacher salaries. Try: lm(satm∼satv), lm(pay∼satm), lm(pay∼satv), lm(pay∼satm+satv), lm(perc∼pay). coefs? R^2? Causation?

- Twin birth weights: www.probability.ca/sta130/twindata.txt

- A certain politician, and their fraction of Georgia primary votes won, by county: www.probability.ca/sta130/trumpdata.txt

Which variables affect their "fracvotes"? Significant? Causation?

- Country inequality: www.probability.ca/sta130/inequality.txt