General Quantities, Besides Yes/No

- So far, we have mostly done statistics on Yes/No quantities. (Do you support the government? Is the coin heads? Does the die show 5? Did the roulette spin come up 22? etc.)
- Then we could study proportions or fractions or probabilities, and compute P-values and confidence intervals for them, and (now) 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.) Next!

Example: Baby Weights

- Ten babies born in a hospital (in North Carolina) had the following weights, in pounds: $x_1 = 9.88$, $x_2 = 9.12$, $x_3 = 8.00$, $x_4 = 9.38$, $x_5 = 7.44$, $x_6 = 8.25$, $x_7 = 8.25$, $x_8 = 6.88$, $x_9 = 7.94$, $x_{10} = 6.00$. (Here $n = 10$.)
- What is 95% confidence interval 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 $\overline{x} = (x_1 + x_2 + \ldots + x_{10})/10 = \frac{1}{n}\sum_{i=1}^{n} x_i \doteq 8.11$.
  - But is this close to the true μ? How close?
  - Variability? Confidence interval? Hypothesis test? etc.

Baby Weights (cont'd)

- Well, we could estimate $\sigma^2 = E[(X_i - \mu)^2]$ by the average of the squared differences from $\overline{x}$, i.e. by $s^2 = [(x_1 - 8.11)^2 + (x_2 - 8.11)^2 + \ldots + (x_{10} - 8.11)^2]/10 = \frac{1}{n}\sum_{i=1}^{n}(x_i - \overline{x})^2 \doteq 1.226$.
  - (Controversial! Some people, and R, prefer to divide by $n - 1$, which has some advantages (e.g. "unbiased"). But I think it's fine to divide by n; see my article: www.probability.ca/varmse)
  - Then, could estimate sd by: $s = \sqrt{s^2} \doteq \sqrt{1.226} \doteq 1.11$.
- But how close is $\overline{x}$ to μ? For this, we need to consider the probabilities for what $\overline{x}$ could have been (ignoring its observed value, 8.11).
- Well, if each x_i was random, with mean μ, and variance σ^2, then $x_1 + x_2 + \ldots + x_n$ would have mean $n \times \mu = n\mu$, and variance $n \times \sigma^2 = n\sigma^2$.

Probabilities for Baby Weights

- Now use our mean and variance tricks!
- Then $\overline{x} = (x_1 + x_2 + \ldots + x_n)/n$ would have mean $n\mu/n = \mu$ (same as mean of each x_i), and variance $n\sigma^2/n^2 = \sigma^2/n$ (which is only $1/n$ of the variance of each x_i).
- Then $\overline{x} - \mu$ would have mean 0, and variance σ^2/n, hence sd $\sqrt{\sigma^2/n} = \sigma/\sqrt{n}$.
- So, $(\overline{x} - \mu)/(\sigma/\sqrt{n})$ 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!
- So, $P[-1.96 < (\overline{x} - \mu)/(\sigma/\sqrt{n}) < +1.96] \doteq 0.95$. So, $P[-1.96\,\sigma/\sqrt{n} < \overline{x} - \mu < +1.96\,\sigma/\sqrt{n}] \doteq 0.95$. So, $P[\overline{x} - 1.96\,\sigma/\sqrt{n} < \mu < \overline{x} + 1.96\,\sigma/\sqrt{n}] \doteq 0.95$.
  - 95% confidence interval for μ! Good? Any problems?

Confidence Interval for Baby Weights

- Have confidence interval $[\overline{x} - 1.96\,\sigma/\sqrt{n},\ \overline{x} + 1.96\,\sigma/\sqrt{n}]$.
- Problem: σ is unknown! Could replace it by its estimate, s. This is like a "bold" option (though quite accurate if n is large). Is there also a "conservative" option? No! σ could be very large!
  - Instead, can compensate by using the "t distribution" instead of the normal distribution. ("t test") This corresponds to increasing the factor "1.96" a little bit, depending on the value of n:

n	5	10	20	50	100	200	500
factor	2.78	2.26	2.09	2.01	1.98	1.97	1.96

- In this course, don't worry, just replace σ by s, and use "1.96" for simplicity. So, confidence interval for μ is: $[\overline{x} - 1.96\,s/\sqrt{n},\ \overline{x} + 1.96\,s/\sqrt{n}]$.

Confidence Interval for Baby Weights (cont'd)

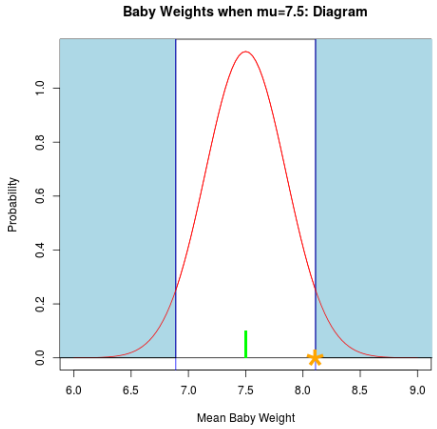
- Confidence interval: $[\overline{x} - 1.96\,s/\sqrt{n},\ \overline{x} + 1.96\,s/\sqrt{n}]$.
- Baby example: $n = 10$, $\overline{x} = 8.11$, $s = 1.11$, so 95% confidence interval for μ is: $[8.11 - 1.96 \times 1.11/\sqrt{10},\ 8.11 + 1.96 \times 1.11/\sqrt{10}] \doteq [7.42,\ 8.80]$.
  - Margin of error is: $1.96\,s/\sqrt{n} = 1.96 \times 1.11/\sqrt{10} \doteq 0.69$.
- Conclusion: We are 95% confident that the true mean baby weight, μ, is between 7.42 pounds and 8.80 pounds.
- Or, if use "2.26" factor instead, then 95% confidence interval becomes: $[8.11 - 2.26 \times 1.11/\sqrt{10},\ 8.11 + 2.26 \times 1.11/\sqrt{10}] \doteq [7.32,\ 8.90]$. (A bit wider, i.e. a bit more uncertainty.)
- But are we sure? Could the true mean baby weight be just 7.5 pounds? P-value? Hypothesis test?

General Quantities: Summary So Far

- Have data values $x_1, x_2, \ldots, x_n$.
- Can estimate true mean μ by $\overline{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$.
- Then can estimate true variance σ^2 by $s^2 = \frac{1}{n}\sum_{i=1}^{n}(x_i - \overline{x})^2$.
- Then can estimate true sd σ by $s = \sqrt{s^2}$. ("bold")
- Then $\overline{x}$ is approximately normal, with mean μ, and variance σ^2/n, so sd $\sigma/\sqrt{n} \approx s/\sqrt{n}$.
- So, $(\overline{x} - \mu)/(s/\sqrt{n})$ is approximately standard normal.
- So, $P[-1.96 < (\overline{x} - \mu)/(s/\sqrt{n}) < +1.96] \approx 0.95$.
- So, $P[\overline{x} - 1.96\,s/\sqrt{n} < \mu < \overline{x} + 1.96\,s/\sqrt{n}] \approx 0.95$.
- So, 95% C.I. for μ is $[\overline{x} - 1.96\,s/\sqrt{n},\ \overline{x} + 1.96\,s/\sqrt{n}]$.
- Baby weights: $n = 10$, $\overline{x} \doteq 8.11$, $s \doteq 1.11$, C.I. $\doteq [7.42,\ 8.80]$.

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 σ^2, then $\overline{x} = (x_1 + x_2 + \ldots + x_n)/n$ would have mean μ and variance σ^2/n.
- But the observed value of $\overline{x}$ was 8.11.
- So, the P-value is the probability, assuming that $\mu = 7.5$, that the value of $\overline{x}$ would have been 8.11 or more, or 6.89 or less (two-sided) (since $8.11 = 7.5 + 0.61$, and $6.89 = 7.5 - 0.61$).
- Now, if $\mu = 7.5$, then $\overline{x}$ has mean 7.5, and variance σ^2/n.
  - Once again, to proceed, replace σ (unknown) by s.
  - So, assume the variance is $s^2/n \doteq 1.226/10$.
  - So, assume the sd is $\sqrt{s^2/n} = s/\sqrt{n} \doteq 1.11/\sqrt{10} \doteq 0.35$.



- So, under the null, $\overline{x}$ has mean 7.5, and sd about 0.35.
- Now, once again, we "should" use the t-distribution instead of a normal (i.e., there's slightly more uncertainty), but for simplicity we'll just use a normal.
- So, the P-value is the probability that the random quantity $\overline{x}$, which is approximately normal(!), and has mean 7.5, and sd approximately 0.35, will be 8.11 or more, or 6.89 or less (two-sided).
- In R: pnorm(8.11, 7.5, 0.35, lower.tail=FALSE) + pnorm(6.89, 7.5, 0.35, lower.tail=TRUE). Answer is: 0.08135857. 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 $Z = (8.11 - 7.5)/0.35 \doteq 1.74$, so P-value $= P(Z > 1.74) + P(Z < -1.74) = (1 - P(Z < 1.74)) + (1 - P(Z < 1.74)) \doteq 2 \times (1 - 0.9591) \doteq 0.0818$.

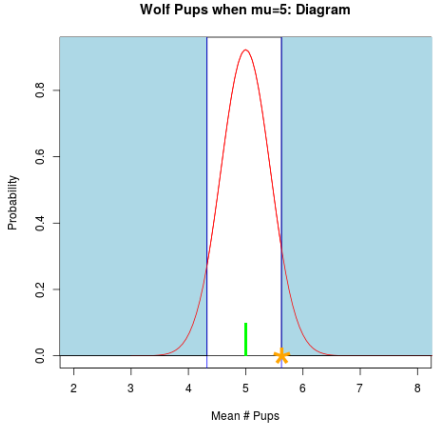
- Let's try another test! For the baby example, suppose instead we want to test the null hypothesis that $\mu = 7.2$, versus the alternative hypothesis that $\mu > 7.2$ (one-sided).
- Now, under the null, $\overline{x}$ has mean 7.2, and sd approximately $s/\sqrt{n} = 1.11/\sqrt{10} \doteq 0.35$.
- So, the P-value is the probability that the random quantity $\overline{x}$, which is approximately normal(!), and has mean 7.2, and sd ≈ 0.35, will be 8.11 or more. [Not or 6.29 or less, since just one-sided.]
- In R: pnorm(8.11, 7.2, 0.35, lower.tail=FALSE). Answer is: 0.004661188. Much less than 0.05! So, can reject the null!
  - (Or, using table: P-value $= P(Z > (8.11 - 7.2)/0.35) = P(Z > 2.6) = 1 - P(Z < 2.6) \doteq 1 - 0.9953) \doteq 0.0047$.)
- Conclusion: Based on the ten baby weights studied, the true mean baby birth weight, μ, is more than 7.2 pounds.

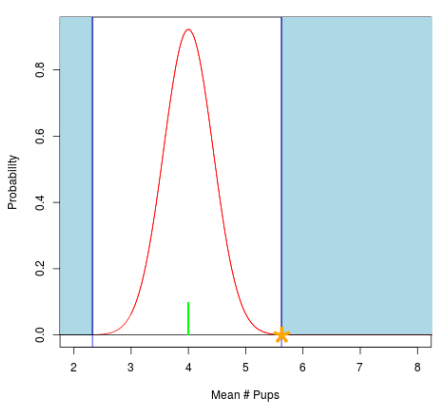
General Quantites Example: Wolf Pups

- A study of endangered wolves in the southwestern United States sampled 16 wolf dens, and found the following numbers of pups (baby wolves) in them: 5, 8, 7, 5, 3, 4, 3, 9, 5, 8, 5, 6, 5, 6, 4, 7.
- Here the sample mean is $\overline{x} = \frac{1}{n}\sum_{i=1}^{n} x_i = \frac{1}{16}(5 + 8 + 7 + \ldots + 4 + 7) = 5.625$.
- And, the sample variance is $s^2 = \frac{1}{n}\sum_{i=1}^{n}(x_i - \overline{x})^2 = \frac{1}{16}\sum_{i=1}^{16}(x_i - 5.625)^2 = \frac{1}{16}([5 - 5.625]^2 + [8 - 5.625]^2 + [7 - 5.625]^2 + \ldots + [4 - 5.625]^2 + [7 - 5.625]^2) \doteq 2.984$.
  - (If divide by $n - 1$ instead of n, get 3.183.)
- So, the sample sd is $s = \sqrt{s^2} = \sqrt{2.984} \doteq 1.727$.
- Then a 95% confidence interval for the true mean number μ of pups per den is: $[\overline{x} - 1.96\, s/\sqrt{n},\ \overline{x} + 1.96\, s/\sqrt{n}] = [5.625 - 1.96 \times 1.727/\sqrt{16},\ 5.625 + 1.96 \times 1.727/\sqrt{16}] \doteq [4.78, 6.47]$.

Wolf Pups (cont'd)

- Conclusion: We are 95% confident that the true mean number of pups per wolf den is between 4.78 and 6.47.
- Could the true mean, μ, be equal to 5?
- The P-value for this is the probability that a normal (approx.) random variable with mean 5, and sd $s/\sqrt{n} \doteq 1.727/\sqrt{16} \doteq 0.432$, is 5.625 or more, or 4.375 or less (since $5.625 - 5 = 5 - 4.375$).
- In R: pnorm(5.625, 5, 0.432, lower.tail=FALSE) + pnorm(4.375, 5, 0.432, lower.tail=TRUE). Answer is: 0.1479644. More than 0.05! So, cannot reject the null!
- Conclusion: Based on the available data from the 16 wolf dens, the true mean number of pups, μ, could indeed be equal to 5.
- Could it be 4?
- First, some diagrams ...





Wolf Pups (cont'd)

- Could μ be 4?
- The P-value for that is the probability that a normal (approx.) random variable with mean 4, and sd $s/\sqrt{n} \doteq 1.727/\sqrt{16} \doteq 0.432$, is 5.625 or more, or 2.375 or less (since $5.625 - 4 = 4 - 2.375$).
- In R: pnorm(5.625, 4, 0.432, lower.tail=FALSE) + pnorm(2.375, 4, 0.432, lower.tail=TRUE). Answer is: 0.0001688474. Much less than 0.05! So, can reject the null! So, μ is not equal to 4.
- Conclusion: Based on the available data from the 16 wolf dens, the true mean number of pups, μ, is not equal to 4.
- SUMMARY: We can compute confidence intervals and P-values for general quantities, similar to for Yes/No proportions. The main differences are: we estimate the mean by $\overline{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$ instead of $\hat{p}$, and estimate the individual variance by $s^2 = \frac{1}{n}\sum_{i=1}^{n}(x_i - \overline{x})^2$ instead of $\hat{p}(1 - \hat{p})$.

Connection between General Quantities and Proportions

- Recall: For proportions, we estimate the mean by $\hat{p}$, and estimate the individual variance (bold option) by $\hat{p}(1-\hat{p})$.
- But for general quantities, we estimate the mean by $\overline{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$, and estimate the individual variance by $s^2 = \frac{1}{n}\sum_{i=1}^{n}(x_i - \overline{x})^2$.
- What is the connection between these two cases?
- Suppose we write $x_i = 1$ for each Yes, and $x_i = 0$ for each No.
  - Then, $\overline{x} = \frac{1}{n}\sum_{i=1}^{n} x_i = \frac{1}{n}$(number of Yes in sample) = proportion of Yes in sample $= \hat{p}$. Same as before!
  - And, $s^2 = \frac{1}{n}\sum_{i=1}^{n}(x_i - \overline{x})^2 = \frac{1}{n}\sum_{i=1}^{n}(x_i - \hat{p})^2$ $= \hat{p}(1-\hat{p})^2 + (1-\hat{p})(0-\hat{p})^2 = \hat{p}(1-\hat{p})$. Also same as before!
- So, these two cases aren't so different, after all.

Comparing Two General Quantities

- For Yes/No proportions (like polls), we also know how to compare two different samples to each other, and get a confidence interval for the difference of the means, or a P-value for testing if the two means are equal. Can we do that with general quantities?
- EXAMPLE: 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: 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 $x_1, x_2, \ldots, x_{22}$ for the birthweights of the $n_1 = 22$ babies whose mothers smoked. And, write $y_1, y_2, \ldots, y_{35}$ for the birthweights of the $n_2 = 35$ babies whose mothers didn't smoke.
- Then can compute the means, $\overline{x} = \frac{1}{22}\sum_{i=1}^{22} x_i$, and $\overline{y} = \frac{1}{35}\sum_{i=1}^{35} y_i$. Obtain: $\overline{x} = 2863$, and $\overline{y} = 3588$.
- So, $\overline{y}$ is larger, and in fact $\overline{y} - \overline{x} = 725$ 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 a confidence interval for $\mu_2 - \mu_1$? And, what is the P-value to test the null hypothesis that $\mu_1 = \mu_2$, against 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 $\overline{y} - \overline{x}$.
  - This quantity has mean $\mu_2 - \mu_1$.
  - But what about the variance?
  - Write σ_1^2 for the true variance of the birthweights of babies whose mothers smoked. And σ_2^2 for those whose mothers didn't.
  - And, write $s_1^2 = \frac{1}{n_1}\sum_{i=1}^{n_1}(x_i - \overline{x})^2 \approx \sigma_1^2$ (sample variance).
  - And $s_2^2 = \frac{1}{n_2}\sum_{i=1}^{n_2}(y_i - \overline{y})^2 \approx \sigma_2^2$.
  - Then $\overline{x}$ has variance $\sigma_1^2/n_1 \approx s_1^2/n_1$.
  - And, $\overline{y}$ has variance $\sigma_2^2/n_2 \approx s_2^2/n_2$.
  - So, $\overline{y} - \overline{x}$ has variance $\sigma_1^2/n_1 + \sigma_2^2/n_2 \approx s_1^2/n_1 + s_2^2/n_2$.
  - So, $\overline{y} - \overline{x}$ has sd $\approx \sqrt{s_1^2/n_1 + s_2^2/n_2}$.

Comparing Birthweights (cont'd)

- Confidence interval? Use mean & variance tricks again!
- Indeed, here $((\overline{y} - \overline{x}) - (\mu_2 - \mu_1))\Big/\sqrt{s_1^2/n_1 + s_2^2/n_2}$ has mean 0, and sd 1, so it is approximately standard normal.
- So, $P[-1.96 < ((\overline{y} - \overline{x}) - (\mu_2 - \mu_1))\Big/\sqrt{s_1^2/n_1 + s_2^2/n_2} < +1.96] \doteq 0.95$.
- Re-arranging (similar to before), $P[\overline{y} - \overline{x} - 1.96\sqrt{s_1^2/n_1 + s_2^2/n_2} < \mu_2 - \mu_1 < \overline{y} - \overline{x} + 1.96\sqrt{s_1^2/n_1 + s_2^2/n_2}] \doteq 0.95$.
  - This gives a 95% confidence interval for $\mu_2 - \mu_1$!
  - Namely, $[\overline{y} - \overline{x} - 1.96\sqrt{s_1^2/n_1 + s_2^2/n_2},\ \overline{y} - \overline{x} + 1.96\sqrt{s_1^2/n_1 + s_2^2/n_2}]$.
- Next, apply this to the birthweight data ...

Comparing Birthweights (cont'd)

- Birthweight data: $n_1 = 22$, $n_2 = 35$, $\overline{x} = 2863$, $\overline{y} = 3588$.
- For this data, we compute that $s_1^2 = \frac{1}{n_1}\sum_{i=1}^{n_1}(x_i - \overline{x})^2 = \frac{1}{22}[(3276-2863)^2 + (1974-2863)^2 + \ldots + (2912-2863)^2] = 873,531.9$. Then $s_1 = \sqrt{873,531.9} \doteq 934.6$.
- Also, $s_2^2 = \frac{1}{n_2}\sum_{i=1}^{n_2}(y_i - \overline{y})^2 = \frac{1}{35}[(3612-3588)^2 + (3640-3588)^2 + \ldots + (3556-3588)^2] = 346,713.6$ (smaller!). Then $s_2 = \sqrt{346,713.6} \doteq 588.8$.
- Hence, $\overline{y} - \overline{x}$ has mean $\mu_2 - \mu_1$, and sd $\approx \sqrt{s_1^2/n_1 + s_2^2/n_2} \doteq \sqrt{873,531.9/22 + 346,713.6/35} \doteq 222.7$.
- So, $((\overline{y} - \overline{x}) - (\mu_2 - \mu_1))\Big/\sqrt{s_1^2/n_1 + s_2^2/n_2}$ has mean 0, and sd 1. Standard normal! (approx.)
- This is what we need!

Comparing Birthweights (cont'd)

- Birthweight data: $n_1 = 22$, $n_2 = 35$, $\overline{x} = 2863$, $\overline{y} = 3588$, $s_1^2 = 873,531.9$, $s_2^2 = 346,713.6$.
- So, $P[3588 - 2863 - 1.96\sqrt{873,531.9/22 + 346,713.6/35} < \mu_2 - \mu_1 < 3588 - 2863 + 1.96\sqrt{873,531.9/22 + 346,713.6/35}] \doteq 0.95$.
- i.e., $P[288.4 < \mu_2 - \mu_1 < 1161.6] \doteq 0.95$.
- i.e., 95% confidence interval is [288.4, 1161.6].
- Conclusion: We are 95% confident that the true mean birthweight of babies whose mothers do not smoke, is between 288.4 and 1,161.6 grams higher than the true mean birthweight of babies whose mothers do smoke.
- Good!
- What about hypothesis tests and P-values?

Hypothesis Test for Comparing Birthweights

- Null hypothesis: $\mu_1 = \mu_2$, i.e. the two true means are equal.
- Under the null hypothesis, $\overline{y} - \overline{x}$ has mean $\mu_2 - \mu_1 = 0$, and sd $\approx \sqrt{s_1^2/n_1 + s_2^2/n_2} \doteq 222.7$ like before.
- So, the two-sided P-value is the probability that a normal, with mean 0, and sd 222.7, would be as large or larger than the observed value 725, or as small or smaller than -725.
- In R: pnorm(725, 0, 222.7, lower.tail=FALSE) + pnorm(−725, 0, 222.7, lower.tail=TRUE). Answer: 0.00113. 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 $\bar{x} = \frac{1}{25}(292 + 360 + \ldots + 1 + 3) = 288.4$ seconds (nearly 5 minutes). And, $\bar{y} = \frac{1}{24}(653 + 73 + \ldots + 17 + 119) \doteq 539.4$ seconds (about 9 minutes).
- Hence, $\bar{y} - \bar{x}$ has observed value $539.4 - 288.4 = 251.0$.
- Also, $s_1^2 = \frac{1}{25}[(292 - 288.4)^2 + (360 - 288.4)^2 + \ldots + (1 - 288.4)^2 + (3 - 288.4)^2] = 166,146.8$, so $s_1 = \sqrt{166,146.8} \doteq 407.6$.
- And, $s_2^2 = \frac{1}{24}[(653 - 539.4)^2 + (73 - 539.4)^2 + \ldots + (17 - 539.4)^2 + (119 - 539.4)^2] = 618,271.8$, so $s_2 = \sqrt{618,271.8} \doteq 786.3$.
- Then $\bar{y} - \bar{x}$ has sd $\approx \sqrt{s_1^2/n_1 + s_2^2/n_2}$ $\doteq \sqrt{166,146.8/25 + 618,271.8/24} \doteq 180.0$.

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 180.0 seconds, is larger than the observed difference, i.e. than 251.0 seconds.
- In R: pnorm(251, 0, 180.0, lower.tail=FALSE). Answer: 0.0816. 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, 180.0, lower.tail=FALSE) + pnorm(−251, 0, 180.0, lower.tail=TRUE). Answer: 0.1632. 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.

Phone Call Example: Confidence Interval

- Recall that here $\bar{y} - \bar{x}$ has mean $\mu_2 - \mu_1$, and sd ≈ 180.0 (as above), and observed value 251.0.
- So, a 95% confidence interval for $\mu_2 - \mu_1$ is: $[\bar{y} - \bar{x} - 1.96\sqrt{s_1^2/n_1 + s_2^2/n_2},\ \bar{y} - \bar{x} + 1.96\sqrt{s_1^2/n_1 + s_2^2/n_2}]$, i.e. $[251.0 - 1.96 \times 180.0,\ 251.0 + 1.96 \times 180.0]$, i.e. $[-101.8,\ 603.8]$.
- Conclusion: Based on the available data, on average females could talk on the phone up to 101.8 seconds less than males, or up to 603.8 seconds more than males; we can't say which.
- So, that's P-values and confidence intervals for comparing two different sets of general quantities (e.g. birth weights when mother smokes or doesn't smoke; cell phone call lengths for males and females). Get it? (More practice on Homework #3.)
- Next: What about "correlations" between quantities?

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 (# chirps / minute) 15 times, at different temperatures (in Celsius). The results were as follows:

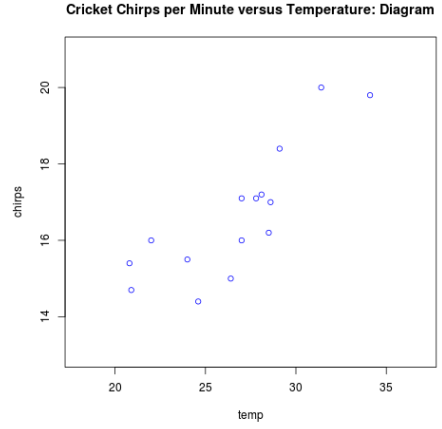
Temperature (C)	31.4	22.0	34.1	29.1	27.0	24.0	20.9
Chirps / Minute	20.0	16.0	19.8	18.4	17.1	15.5	14.7

Temp	27.8	20.8	28.5	26.4	28.1	27.0	28.6	24.6
C/M	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/minute versus temperature, to see if there is a pattern.
- Let's try it ...



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/minute. 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 $\mathrm{Cor}(X, Y)$ 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: $\mu_X \approx \bar{x} = \frac{1}{n}\sum_{i=1}^n x_i$, $\mu_Y \approx \bar{y} = \frac{1}{n}\sum_{i=1}^n y_i$; and the true variances σ_X^2 and σ_Y^2, by: $\sigma_X^2 \approx s_x^2 = \frac{1}{n}\sum_{i=1}^n (x_i-\bar{x})^2$, $\sigma_Y^2 \approx s_y^2 = \frac{1}{n}\sum_{i=1}^n (y_i-\bar{y})^2$.
- Then, the sample correlation between X and Y is

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

  - We know all these quantities from our sample. Good!
  - If we use the Z-scores $z_i = (x_i-\bar{x})/s_x$, and $w_i = (y_i-\bar{y})/s_y$, then we can write this more simply as: $r = \frac{1}{n}\sum_{i=1}^n z_i\, w_i$.
  - (Some people, and R, divide by $n-1$ instead of n ... let's not worry about that ...)

Back to Cricket Data

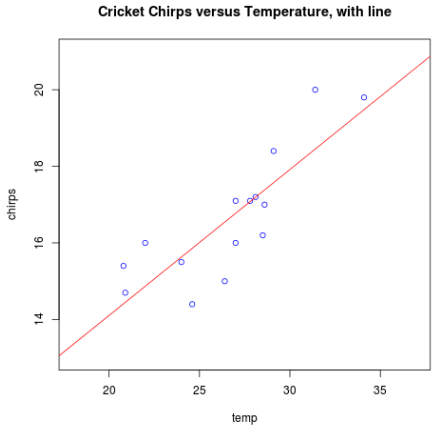
Temperature (C)	31.4	22.0	34.1	29.1	27.0	24.0	20.9
Chirps / Minute	20.0	16.0	19.8	18.4	17.1	15.5	14.7

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

- Write X for temperature, and Y for chirps/minute. Then $\bar{x} = \frac{1}{n}\sum_{i=1}^n x_i = \frac{1}{15}[31.4 + 22.0 + \ldots + 24.6] \doteq 26.7$. And, $\bar{y} = \frac{1}{n}\sum_{i=1}^n y_i = \frac{1}{15}[20.0 + 16.0 + \ldots + 14.4] \doteq 16.7$.
- And, $s_x^2 = \frac{1}{n}\sum_{i=1}^n (x_i-\bar{x})^2 = \frac{1}{15}[(31.4-26.7)^2 + (22.0-26.7)^2 + \ldots + (24.6-26.7)^2] \doteq 13.0$. So, $s_x = \sqrt{s_x^2} \doteq \sqrt{13.0} \doteq 3.6$. Also, $s_y^2 = \frac{1}{n}\sum_{i=1}^n (y_i-\bar{y})^2 = \frac{1}{15}[(20.0-16.7)^2 + (16.0-16.7)^2 + \ldots + (14.4-16.7)^2] \doteq 2.7$. So, $s_y = \sqrt{s_y^2} \doteq \sqrt{2.7} \doteq 1.6$.
- 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, $r = r_{xy} = \frac{1}{n}\sum_{i=1}^n z_i w_i = \frac{1}{n}\sum_{i=1}^n \left(\frac{x_i-\bar{x}}{s_x}\right)\left(\frac{y_i-\bar{y}}{s_y}\right) = \frac{1}{15}\left[\left(\frac{31.4-26.7}{3.6}\right)\left(\frac{20.0-16.7}{1.6}\right) + \left(\frac{22.0-26.7}{3.6}\right)\left(\frac{16.0-16.7}{1.6}\right) + \ldots + \left(\frac{24.6-26.7}{3.6}\right)\left(\frac{14.4-16.7}{1.6}\right)\right] \doteq 0.861$. Phew!
- So, the sample correlation is 0.861. This means that on average, every time the temperature increases by one standard deviation, the cricket chirp rate increases by 0.861 of its standard deviation.
  - That is, every time the temperature increases by s_x, the cricket chirp rate increases by $0.861\, s_y$.
  - Or, every time the temperature increases by one degree, the cricket chirp rate increases by $r_{xy}\, s_y/s_x = 0.861\, s_y/s_x$.
  - Can illustrate with "line of best fit" (more later) ...



Correlation: Discussion

- Conclusion so far: the sample correlation r_{xy} between the temperature in degrees celsius, and the rate of cricket chirps per minute, is equal to 0.861.
- This means that the true correlation $\rho_{X,Y}$ between the temperature in degrees celsius, and the rate of cricket chirps per minute, is probably: approximately 0.861.
- This means that the correlation between the temperature in degrees celsius, and the rate of cricket chirps per second (not minute), is also approximately 0.861. (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 (not minute), is also approximately 0.861. (Correlation is unaffected by adding any constants, or multiplying by any positive constants.)

Correlation Calculations: Aside

- Computing the sample correlation r_{xy} requires calculating lots of things: $\bar{x}, \bar{y}, s_x, s_y, z_i, w_i, r_{xy} = \frac{1}{n}\sum_{i=1}^n z_i w_i$.
  - 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: 0.836.
  - Very close to 0.861, but not quite the same.
  - Why? Because R divides by $n-1$, not by n!

Cricket Data: Correlation (cont'd)

- Is 0.861 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, 0.861 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 $r_{xy} = 0.861$. (Strong positive correlation.) And, $r_{xy} = \frac{1}{n}\sum_{i=1}^n z_i w_i$, where $z_i = (x_i - \bar{x})/s_x$ and $w_i = (y_i - \bar{y})/s_y$ 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 $\bar{x}$ 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 $Var\left(\sum_{i=1}^n z_i w_i\right) \approx n \times 1 = n$.
- So what about $Var(r_{xy})$? Well, $Var(r_{xy}) = Var\left(\frac{1}{n}\sum_{i=1}^n z_i w_i\right) = \frac{1}{n^2}\,Var\left(\sum_{i=1}^n z_i w_i\right) \approx \frac{1}{n^2}\,(n) = 1/n$.
- Summary: in the independent case, $Var(r_{xy}) \approx 1/n$.
- 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" $\text{arctanh}(r_{xy}) := \frac{1}{2}\ln\left(\frac{1+r_{xy}}{1-r_{xy}}\right)$, 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 $Var(r_{xy}) \approx 1/n$. See also R's cor.test.)

Confidence Intervals for Correlation

- We're interested in the true correlation, $\rho_{X,Y}$. We can estimate $\rho_{X,Y}$ by the sample correlation, r_{xy}. We've argued that r_{xy} has mean approximately $\rho_{X,Y}$, and variance approximately $1/n$, hence sd approximately $\sqrt{1/n} = 1/\sqrt{n}$.
- How to get confidence intervals? Standardise!
- It follows that $(r_{xy} - \rho_{X,Y})/(1/\sqrt{n})$ has mean approximately 0, and sd approximately 1. And, if n is reasonably large, then the probabilities for r_{xy} are approximately normal, so that $(r_{xy} - \rho_{X,Y})/(1/\sqrt{n})$ is approximately standard normal. Good!
- Then, $P[-1.96 < (r_{xy} - \rho_{X,Y})/(1/\sqrt{n}) < +1.96] \doteq 0.95$.
- Re-arranging (just like before), $P[r_{xy} - 1.96/\sqrt{n} < \rho_{X,Y} < r_{xy} + 1.96/\sqrt{n}] \doteq 0.95$.
  - This gives a 95% confidence interval for $\rho_{X,Y}$!

Confidence Intervals for Crickets

- Summary: a 95% confidence interval for the true correlation $\rho_{X,Y}$ is given by the interval: $[r_{xy} - 1.96/\sqrt{n},\ r_{xy} + 1.96/\sqrt{n}]$.
- In the cricket example, $n = 15$, and the sample correlation was $r_{xy} \doteq 0.861$.
- So, a 95% confidence interval for $\rho_{X,Y}$ is: $[r_{xy} - 1.96/\sqrt{n},\ r_{xy} + 1.96/\sqrt{n}] = [0.861 - 1.96/\sqrt{15},\ 0.861 + 1.96/\sqrt{15}] \doteq [0.355, 1.367]$.
- But correlation is always ≤ 1, so we could replace this confidence interval by: [0.355, 1].
- Conclusion: We are 95% confident that the true correlation between temperature and cricket chirp rate is somewhere between 0.355 and 1, i.e. is more than 0.355. (i.e., moderate to strong ...)
- And what about P-values?

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 0.861.
- So, the P-value is the probability that a normal random quantity, with mean 0, and sd $1/\sqrt{n} = 1/\sqrt{15}$, is 0.861 or more, or -0.861 or less (two-sided). In R: pnorm(0.861, 0, 1/sqrt(15), lower.tail=FALSE) + pnorm(−0.861, 0, 1/sqrt(15), lower.tail=TRUE). Answer is: 0.0008541031.
- 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 $s_x^2 = \frac{1}{12}[(14.2-18.7)^2 + (16.4-18.7)^2 + \ldots + (17.2-18.7)^2] \doteq 14.75$, so $s_x \doteq \sqrt{14.75} \doteq 3.84$. And, $s_y^2 = \frac{1}{12}[(215-402)^2 + (325-402)^2 + \ldots + (408-402)^2] \doteq 14563$, so $s_y \doteq \sqrt{14563} \doteq \$120.7$.
- Hence, $r = r_{xy} = \frac{1}{n}\sum_{i=1}^n z_i w_i = \frac{1}{n}\sum_{i=1}^n \left(\frac{x_i - \overline{x}}{s_x}\right)\left(\frac{y_i - \overline{y}}{s_y}\right) = \frac{1}{12}\left[\left(\frac{14.2-18.7}{3.84}\right)\left(\frac{215-402}{120.7}\right) + \left(\frac{16.4-18.7}{3.84}\right)\left(\frac{325-402}{120.7}\right) + \ldots + \left(\frac{17.2-18.7}{3.84}\right)\left(\frac{408-402}{120.7}\right)\right] \doteq 0.957$.
- 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: Cigarettes versus Latitude

- I looked up the average latitude, and average number of cigarettes smoked per adult per year, for 12 northern countries:

Country	Cigarettes	Latitude
Canada	809	56.1
U.States	1028	37.1
Mexico	371	23.6
U.Kingdom	750	55.4
France	854	46.2
Germany	1045	51.2
Spain	1757	40.5
Greece	2996	39.1
Russia	2786	61.5
China	1711	35.9
Japan	1841	36.2
S.Korea	1958	35.9

Cigarettes versus Latitude (cont'd)

- Is there a statistically significant correlation between the two? (Guesses? Discussion?)
- Let X be cigarettes, and Y be latitude. Then $\overline{x} = \frac{1}{12}[809 + 1028 + \ldots + 1958] \doteq 1492$, and $\overline{y} = \frac{1}{12}[56.1 + 37.1 + \ldots + 35.9] \doteq 43.2$. Then $s_x^2 = \frac{1}{12}[(809 - 1492)^2 + (1028 - 1492)^2 + \ldots + (1958 - 1492)^2] \doteq 625,120$, so $s_x \doteq \sqrt{625,120} \doteq 790$. And, $s_y^2 = \frac{1}{12}[(56.1-43.2)^2+(37.1-43.2)^2+\ldots+(35.9-43.2)^2] \doteq 110.35$, so $s_y \doteq \sqrt{110.35} \doteq 10.5$.
- Hence, $r = r_{xy} = \frac{1}{n}\sum_{i=1}^n z_i w_i = \frac{1}{n}\sum_{i=1}^n \left(\frac{x_i-\overline{x}}{s_x}\right)\left(\frac{y_i-\overline{y}}{s_y}\right) = \frac{1}{12}\left[\left(\frac{809-1492}{790}\right)\left(\frac{56.1-43.2}{10.5}\right) + \left(\frac{1028-1492}{790}\right)\left(\frac{37.1-43.2}{10.5}\right) + \ldots + \left(\frac{1958-1492}{790}\right)\left(\frac{35.9-43.2}{10.5}\right)\right] \doteq 0.109$.
- Weak, positive correlation. Why? Or is it just luck?

Cigarettes versus Latitude (cont'd)

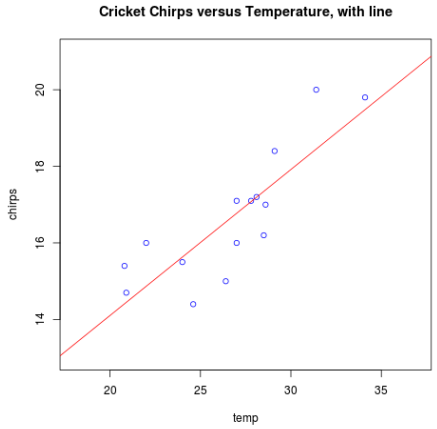
- Correlation $r_{xy} = 0.109$. Is it just luck? Use statistics! e.g. hypothesis tests (P-values), and confidence intervals!
- 95% confidence interval for the true correlation $\rho_{X,Y}$: $[r_{xy} - 1.96/\sqrt{n},\ r_{xy} + 1.96/\sqrt{n}] = [0.109 - 1.96/\sqrt{12},\ 0.109 + 1.96/\sqrt{12}] \doteq [-0.457,\ 0.675]$. Could be positive or negative.
- P-value for null hypothesis that $\rho_{X,Y} = 0$, versus the alternative hypothesis that $\rho_{X,Y} \neq 0$: Probability that a normal with mean 0, and sd $1/\sqrt{12}$, is 0.109 or more, or -0.109 or less. In R: pnorm(0.109, 0, 1/sqrt(12), lower.tail=FALSE) + pnorm(-0.109, 0, 1/sqrt(12), lower.tail=TRUE). Answer is: 0.7057374. Much more than 0.05. Could be just luck!
- Conclusion: The given data do not demonstrate any correlation between countries' cigarette consumption and latitude.

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, from: https://www.tobaccofreekids.org/research/factsheets/pdf/0176.pdf
- And I found their average income per capita in 2012: http://www.infoplease.com/ipa/A0104652.html
- 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 found high school completion percentage in each U.S. state: http://census.gov/prod/2012pubs/p20-566.pdf. cor(high,inc): 0.438. cor(high,sm): -0.335. Interpretation??

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 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!



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: $r^2 = (0.861)^2 \doteq 0.741$ (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 ~ 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 $\mathrm{Cor}(Y, X_1)$, $\mathrm{Cor}(Y, X_2)$, 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??

Correlation and Regression – More Examples

- Countries: www.probability.ca/sta130/countryR. Try various correlations (cor) and linear regressions (lm). Values? coefs? sd? R^2? Interpretation? Causation?
- www.probability.ca/sta130/SAT.txt Data for SAT scores in Verbal and in Math, by state, together with Percentage of high school students taking the SATs, and also the average public school teacher salaries. Try: lm(satm∼satv), lm(pay∼satm), lm(pay∼satv), lm(pay∼satm+satv), lm(perc∼pay). coefs? sd? R^2? Interpretation? Causation?
- Twin birth weights: www.probability.ca/sta130/twindata.txt
- A certain famous current politician. (image) www.probability.ca/sta130/Rtrump Data from March 1, 2016 Georgia primary vote, county by county. Which variables have a significant effect on "fracvotes"?

Possible Interpretations of Correlations

- Suppose two quantities X and Y have a sample correlation which is far from 0.
- Suppose the corresponding P-value is < 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!

Let's Study Students!

- We'll use you as a sample of university students!
- Get an index card, and a (paper) ruler. On your index card, write down the following information about yourself: **(1)** Male or female? **(2)** Born in what country? **(3)** Live on campus, or off? **(4)** Right-handed, or left, or both? **(5)** Currently wearing glasses or contact lenses or neither? **(6)** Your height (in feet and inches? inches only? centimeters?). **(7)** The number of credit cards your currently have with you. **(8)** Circumference of your wrist (at the smallest point) in cm. **(9)** Circumference of your **(a)** right and **(b)** left flexed bicep (at the largest point) in cm. **(10)** # siblings?
- What statistical questions can we ask, given this data? Comparisons of two proportions? Comparisons of two general quantities? Correlations? Think of at least one question of each type. www.probability.ca/sta130/studentdata.txt Then we will investigate them! www.probability.ca/sta130/studentdataR