

Stat 1: Solutions to Practice Final Problems

1. Does taking vitamins prevent colon cancer? A study assigned 864 subjects at random to four groups. One group took beta-carotene, another took vitamins C and E, a third took all three, and the fourth group took only a dummy pill (placebo). After four years, there was no significant difference among the groups in the formation of cancer-related polyps in the colon.

a) (4 pts) *Earlier studies of 1000's of people had shown that people who choose to eat lots of vegetables containing these vitamins (including beta-carotene) were less likely to have colon cancer. Explain why the new study is more trustworthy, despite the relatively small sample size.*

The new study is an experiment, while the earlier study was observational study, in which people chose for themselves what diet to eat. As with all observational studies, we have to worry about what other differences might exist for the two groups. In this case, those who eat more vegetables might be vegetarian, or might do many other things to reduce the risk of cancer (e.g., exercising more, not smoking, etc.).

b) (1 pt) Suppose that the new study reports a P -value of 0.45. The best interpretation of this p -value is _____

- (i) There is a .45 probability that the null hypothesis is true.
- (ii) There is a .45 probability that the null hypothesis is false.
- (iii) There is a .45 probability of finding a difference in polyps as large or larger than they did, assuming there is no difference between the placebo and the other treatments.
- (iv) There is a .45 probability of finding a difference in polyps as small or smaller than they did, assuming there is no difference between the placebo and the other treatments.
- (v) There is a .45 probability of finding the difference in polyps they did, assuming there is a difference between the placebo and the other treatments.

c) (1 pt) The best summary of the results (including the P -value) is _____

- (i) All of the treatments are effective in preventing cancer.
- (ii) The treatments differ in preventing cancer.
- (iii) The observed differences in polyp formation were consistent with there being no differences between the four treatments.
- (iv) The observed differences in polyp formation were too large to be due to chance alone.

2. (2 pts) *If three people each roll a six-sided die, what is the probability that at least one person will roll a 6?*

$$P(\text{at least one } 6) = 1 - P(\text{no sixes}) = 1 - (5/6)^3 = 1 - 0.58 = 0.42.$$

3. The following table lists the ranks and gender of the faculty members at Purdue University.

	Female	Male	Total
Assistant Professor	151	254	405
Associate Professor	154	397	551
Full Professor	99	642	741
Total	404	1293	1697

Imagine choosing a faculty member at random. Let A be the event the person selected is female and let B be the event the person selected is a full professor.

a) (3 pts) Find the marginal and intersection probabilities $P(A)$, $P(B)$ and $P(A \cap B)$.

$$P(A) = 404/1697 = 0.238; \quad P(B) = 741/1697 = 0.437; \quad P(A \cap B) = 99/1697 = 0.0583.$$

b) (2 pts) Find the conditional probabilities $P(A|B)$ and $P(B|A)$.

$$P(A|B) = \frac{P(A \cap B)}{P(B)} = \frac{99/1697}{741/1697} = 99/741 = 0.134.$$

$$P(B|A) = \frac{P(A \cap B)}{P(A)} = \frac{99/1697}{404/1697} = 99/404 = 0.245.$$

c) (2 pts) Are the events A and B independent? Explain how you know.

If A and B were independent, all of the following would be true: $P(B|A) = P(B)$, $P(A|B) = P(A)$, and $P(A \cap B) = P(A)P(B)$. Either these equalities all hold and the events are independent, or else none of them hold and the events are not independent. Here we see that none are true (you would only need to check one of the conditions) so the events are dependent. The proportion of females among full professors is lower than the proportion of females overall. So learning that someone is a full professor lowers the probability they are female.

4. Professor Everson leaves for a holiday and asks his neighbor to water his plants. If the plants are not watered, there is a 0.8 probability they will die. If they are watered, there is still a 0.1 probability they will die. Professor Everson estimates there is a 0.95 probability that his neighbor will remember to water the plants. Assume these probabilities are correct.

a) (2 pts) What is the probability the plants will die? $135/1000 = 0.135$.

b) (2 pts) Professor Everson returns and finds his plants have died. What is the conditional probability that his neighbor watered the plants? $95/135 = 0.704$.

Both parts can be answered most easily by constructing a hypothetical table. Suppose there are 1000 replications of this situation. The assumptions imply that on 950 of these, the plants will be watered (0.95 is the unconditional probability the neighbor remembers to water the plants). So the overall probability that the plants die is 135/1000. The conditional probability that they were watered given that they died is 95/135. Note that it is still quite probable that they were watered even when they die. This is due to the high probability (0.95) we assumed for the probability they would be watered.

	live	die	
watered	855	95	950
not watered	10	40	50
	865	135	1000

Of the 950 times the plants are watered, they will die 95 of those times (the conditional probability of dying given they are watered is 0.1). And of the 50 times they are not watered, they die 40 times (conditional probability of dying given no water is 0.8).

5. Suppose that Swarthmore women's heights vary according to a Normal distribution with mean $\mu = 64$ inches and standard deviation $\sigma = 2.5$ inches.

- a) (2 pts) If I choose one woman at random, what is the probability I'll choose someone taller than 65 inches?

65 inches is 0.4 standard deviations above the mean ($z = (65 - 64)/2.5 = 0.4$). Looking up $z = 0.4$ in Table B we find that 65.54% of the population is less than this value, so $100 - 65.54 = 34.46\%$ are taller than this (the probability is 0.3446).

- b) (2 pts) If I choose a simple random sample of $n = 4$ women, what is the probability that all four are taller than 65 inches? If you did not answer part a, make up a value to use for answering this part.

The probability that all four women are over 65 inches is the probability for one woman raised to the 4th power: $(0.3446)(0.3446)(0.3446)(0.3446) = (0.3446)^4 = 0.014$.

- c) (2 pts) If I choose a simple random sample of $n = 4$ women, what is the probability that the *average* height will be over 65 inches?

The standard deviation for the average of $n = 4$ heights is $\sigma_{\bar{x}} = 2.5/\sqrt{4} = 1.25$ inches. So 65 is now 0.8 standard deviations above the mean ($z = (65 - 64)/(2.5/\sqrt{4}) = 1/1.25 = 0.8$). The probability of a Normal variable being more than 0.8 standard deviations above the mean is 0.2119 ($100 - 78.81 = 21.19\%$).

- d) (2 pts) Suppose the average height for my SRS of $n = 4$ is $\bar{x} = 65.6$ inches. Use this to find a 95% CI for the mean height of all Swarthmore women (still assuming $\sigma = 2.5$).

There is 0.95 probability that our average will fall within ± 2 sd's ($\sigma_{\bar{x}} = 2.5/\sqrt{4} = 1.25$). So our 95% CI is $\bar{x} \pm 2\sigma/\sqrt{n} = 65.6 \pm 2(1.25) = (63.1, 68.1)$.

- e) (3 pts) State hypotheses for testing the claim that the mean height is $\mu = 64$ inches. Does the sample from part d provide evidence at $\alpha = 0.05$ to reject the claim? Report the P -value and explain your conclusions.

Test $H_o : \mu = 64$ vs $H_a : \mu \neq 64$. We can see from part d that the result will not be significant at the $\alpha = 0.05$ (5%) significance level because the value $\mu = 64$ falls in the 95% CI for μ , the mean height of Swarthmore women. To find the p -value we compute the probability of getting the average of $n = 4$ heights as far from 64 as $\bar{x} = 65.6$, assuming the mean height is $\mu = 64$ inches. The standard score is $z = (65.6 - 64)/(2.5/\sqrt{4}) = 1.6/1.25 = 1.28 \approx 1.3$. The value in table B for $z = 1.3$ is 90.32%. This leaves probability $1 - .9032 = 0.0968$ of getting an average this far above 64 inches. And there is also probability 0.0968 of getting an average 1.3 standard deviations or more below 64 inches. So the overall 2-sided p -value is $2(0.0968) = 0.1936$.

- f) (2 pts) Explain qualitatively how the CI and P -value/significance in parts d and e would change if you saw the same average value ($\bar{x} = 65.6$) from a SRS of $n = 16$ (instead of $n = 4$).

The standard deviation of the average would be smaller (half as large, to be precise) with a larger sample size. This will make the confidence interval narrower (we have a more precise estimate of μ) and the p -value smaller (the same size deviation from 64 would be less likely with a larger sample size).

6. It has been estimated that 66% of all teenagers have a TV set in their room. You take a SRS of $n = 100$ teens in the Philadelphia area and find that 50 have TV sets in their room.

- a) (3 pts) Construct a 95% confidence interval based on this sample, and explain what parameter it is estimating.

The parameter being estimated is p , the proportion of teenagers in the Philadelphia area who have a TV in their room. The sample estimate is $\hat{p} = 50/100 = 0.5$. The quick margin of error for 95% confidence is $1/\sqrt{n} = 1/\sqrt{100} = 0.1$. Because $\hat{p} = 0.5$, this is the same value we get from computing $2\sqrt{\hat{p}(1-\hat{p})/n} = 2\sqrt{.5(.5)/100} = 2(0.05) = 0.1$ (any other value of \hat{p} would give a smaller margin of error than the quick method). The 95% CI is $0.5 \pm 0.1 = (0.4, 0.6)$. So we are 95% confident that between 40% and 60% of teens in the Philly area have TVs in their rooms.

- b) (3 pts) State hypotheses for testing whether $p = 0.66$ in the Philadelphia area. Compute the P -value and explain your conclusion.

Test $H_o : p = 0.66$ vs. $H_a : p \neq 0.66$. We have some evidence for H_a because our sample proportion differed from the null value ($\hat{p} = 0.5 \neq 0.66$). Under H_o , the standard deviation of \hat{p} is $\sigma_{\hat{p}} = \sqrt{0.66(1-0.66)/100} \approx 0.047$. The standard score is then $z = (0.5 - 0.66)/0.047 \approx -3.4$. Table B tells us the probability of getting a value this much smaller than expected is 0.0003. There is also probability 0.0003 that we would get a value 3.4 or more standard deviations above 0.66. So the 2-sided p -value is $2(0.0003) = 0.0006$, meaning there is only 0.06% chance we would get a \hat{p} this far or further from 0.66 if this were the true proportion for Philly teens. This is a small p -value, so we have strong evidence to reject H_o and conclude that the value of p for this population is less than 0.66.

- d) (1 pt) The variation from sample to sample when the poll is repeated is described by the standard deviation of the sampling distribution. We would like this variation to be small, so that repeated polls give almost the same result. To reduce the standard deviation, we could _____

- (i) use an SRS of size less than 100.
→ (ii) use an SRS of size greater than 100.
(iii) use a confidence level less than 95%.
(iv) use a confidence level greater than 95%.
(v) Both (ii) and (iii).
(vi) Both (i) and (iv).

Note that lowering the confidence level will make the CI narrower, but doesn't change the standard deviation of \hat{p} .

7. (3 pts) A friend offers to play a game. You flip two coins and if they both land Heads, your friend will pay you \$2. Otherwise you must pay your friend \$1. Does this game favor you or your friend? Answer by finding the expected value of the amount you would win.

You will either win \$2 (with probability $= (1/2)^2 = 1/4$) or lose \$1 (with probability $1-1/4 = 3/4$). The expected amount that you will win is $(1/4)2 + (3/4)(-1) = -1/4$. So on average you will expect to lose 25 cents per play, so the game favors your friend.

8. In our student survey this semester the correlation between height and shoe length was very close to $r = 0.75$. The average shoe length was about 11 inches, with a standard deviation of about 1 inch. The average height is about 68 inches, with a standard deviation of 4 inches. . . .

a) (1 pt) If I measured in cm instead of inches (1 inch = 2.54 cm) the correlation would be _____

(i) larger. (ii) smaller. →(iii) the same. (iv) we can't tell. The correlation is a unitless summary, so changing units does not affect r .

b) (1 pt) What percent of the variability in heights is explained by the linear regression on shoe length?

The percent of the variability in y explained by a linear fit on x is given by r^2 . In this case we have $r^2 = (0.75)^2 = 0.5625$, so we explain 56.25% more of the variability in heights using the linear fit on shoe length, compared to ignoring shoe lengths and guessing the average height for everyone.

c) (2 pts) Find the least squares prediction equation for predicting a student's height in inches (y) from shoe length in inches (x).

The least squares fit is $\hat{y} = a + bx$, with $b = r(s_y/s_x) = 0.75(4/1) = 3.0$, and $a = \bar{y} - b\bar{x} = 68 - 3(11) = 35$. So the prediction equation is fitted height = $35 + 3(\text{shoe length})$.

d) (2 pts) What height would you predict for a student with a 13 inch shoe?

Plugging in 13 to the equation from c, we get fitted height = $35 + 3(13) = 74$ inches. We could also note that a 13 inch shoe is 2 standard deviations larger than the average shoe length. Our regression to the mean equation tell us that the fitted height is then $0.75(2) = 1.5$ standard deviations above the average height ($r = 0.75$). This corresponds to a height of $68 + 1.5(4) = 74$ inches.

e) (2 pts) What shoe length would you predict for a student who is 74 inches tall?

Given the answer to part d, it is tempting to predict 13 inches. But 74 is the typical height for all people with a 13 inch shoe. Now we want the typical shoe length for all people who are 74 inches tall, which is a different problem (we've switched from predicting height from shoe length to predicting shoe length from height). The new prediction equation is $\hat{x} = a + by$, with $b = r(s_x/s_y) = 0.75(1/4) = 0.1875$, and $a = \bar{x} - b\bar{y} = 11 - 0.1875(68) = -1.75$. The prediction shoe length for someone 74 inches tall is $-1.75 + 0.1875(74) = 12.125$ inches (closer to the mean shoe length than 13).

Again, we could also use the regression to the mean formula. A height of 74 inches is 1.5 standard deviations above the mean height. So our fitted shoe length is $(0.75)(1.5) = 1.125$ standard deviations above the mean shoe length. This gives the same answer: $11 + 1.125(1) = 12.125$ inches.