

Practice Final Exam

From a May 2000 L.A. Times:

"Pregnant women traveling to developing countries should take special precautions to ward off mosquitoes because they are twice as likely to be bitten by the sometimes-disease-carrying insects, according to British researchers.

"Dr. Steve Lindsay and his colleagues at the University of Durham in northern England studied 36 pregnant women and 36 non-pregnant women in rural Gambia. Each night, three women from each group slept alone under a bed net. Researchers counted the number of mosquitoes attracted to each woman. The team reported in Saturday's Lance (<http://www.thelancet.com>) that the pregnant women attracted twice as many mosquitoes."

a) Is this a controlled study or an observational study, and why?

Observational: the "treatment" is pregnancy. The observed response variable is the number of mosquitoes attracted to a woman. Because researchers did not assign some woman to be pregnant and others not, this cannot be classified as a controlled stud.

b) Suppose we let X be a random variable representing the number of mosquitoes attracted to a pregnant woman. Is X binomial? Why or why not?

No. Let's go through the steps.

1) Outcomes of trials are yes/no.

A "trial" here would be whether or not a mosquito is attracted to a woman. So you could say that the mosquito either lands on the net or does not.

2) There are a fixed number of trials. No. There could be any number of mosquitoes on the net.

So it fails to be binomial.

2) The table below shows the number of transfer applicants to UCLA in Fall '99, and the number of these that ended up enrolling in Fall '99. The data come from the Student Affairs Information & Research Office. Transfer applicants are students coming in from other colleges, including other UC's, Cal States, and community colleges.

	Did not Enroll	Enrolled	Total Applicants
Native American	59	16	75
African American	252	66	318
Latino	289	106	395
Chicano	665	214	879
Asian	1916	595	2511
White	2780	870	3650
Other/unknown	788	231	1019
Total	6749	2098	8847

a) What is the probability that a randomly selected transfer student will enroll at UCLA? Let A be the event "transfer student enrolls". There are 8847 transfer students, and 2098 enrolled. Therefore $P(A) = 2098/8847 = .23714$

b) Let A be the event that a randomly selected transfer applicant is Chicano, and let B be the event that he or she enrolls. Find $P(A)$, $P(B)$.

$$P(A) = 879/8847 = .0993$$

$$P(B) = 214/879 = .243$$

(Note: B is a somewhat ambiguous event. When I gave this final, it required some elaboration. What it is asking is, what's the probability a randomly selected Chicano applicant will enroll.

c) What is the probability a randomly selected transfer applicant is Chicano and will enroll?

$$P(\text{Chicano and Enrolls}) = 214/8847 = .024$$

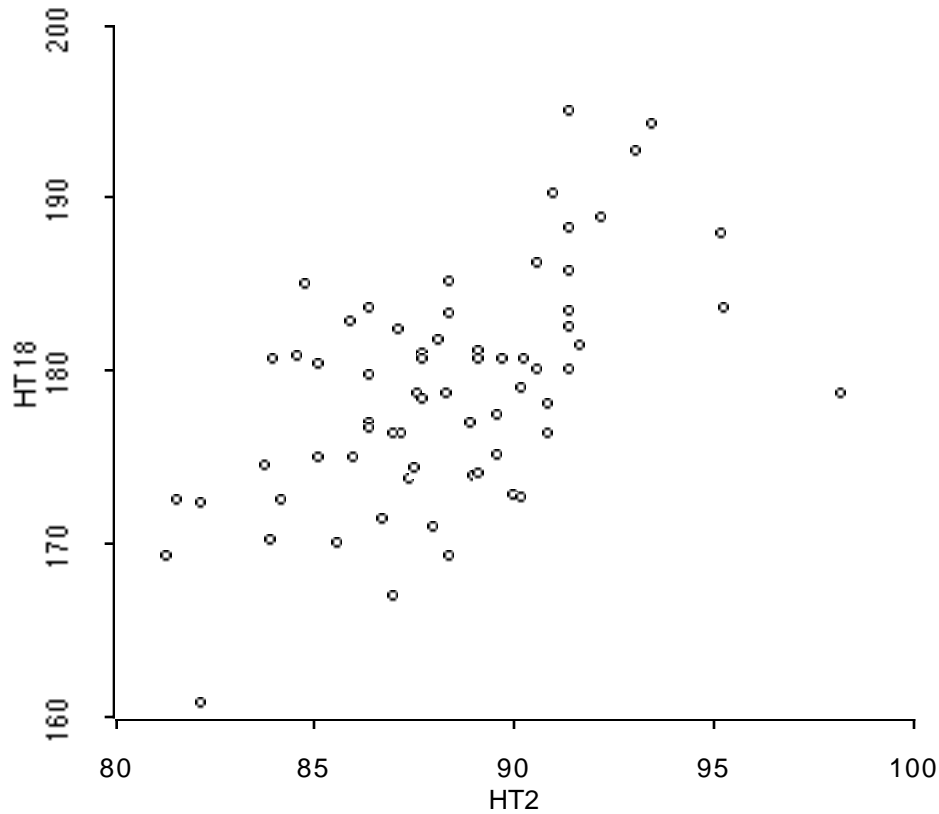
d) According to these data, are the events that the selected transfer applicant is Chicano mutually exclusive from the event that he or she enrolls? Or these events independent? Explain each of your answers.

Mutually exclusive: No. It is possible that a transfer applicant be both Chicano and enroll. Otherwise, the answer to part (c) would be 0. It is not 0, so these are not mutually exclusive.

Independent. If so, then $P(\text{Chicano and Enroll}) = P(\text{chicano}) * P(\text{enroll})$. $P(\text{Chicano}) = .0993$. $P(\text{enroll}) = .23714$. So $P(\text{chicano}) * P(\text{enroll}) = .0022$. But from part c, we found $P(\text{chicano and enroll}) = .024$. So these are not equal, and therefore the events are NOT independent.

3) Does a child's height at age two predict his height at age 18? The data below are based on 66 boys and come from "Physical Growth of California Boys and Girls from Birth to

"Eighteen years," by R. D. Tuddenham and M. M. Snyder, 1954, Univ. of Calif. Publications in Child Development, 1, 183-364. All children were born in 1928-29 in Berkeley, CA. All heights are in centimeters. HT18 is height (in cm) at age 18, and HT2 is height at age 2.



Variable	Average	Std. Dev
HT18	178.98	6.5175
HT2	88.371	3.3203

Correlation, r , between height at 18 and height at age two: $r = 0.5708$.

(Continued next page)

a) (12 points) What is the regression line?

$$\text{slope} = \text{SD}(y) * r / \text{SD}(x) = 6.5175 * .5708 / 3.3203 = 1.120$$

$$\text{intercept} = \bar{y} - \text{slope} * \bar{x} = 178.98 - 1.120 * 88.371 = 80.00$$

$$\hat{y} = 80 + 1.120 x$$

or

$$\text{height at 18} = 80 + 1.120 * \text{height at 2 (measured in cm.)}$$

b) Consider the children who are 85 cm tall at age two. What is the estimated average height of these children at age 18?

$$\text{estimated height} = 80 + 1.120 * 85 = 175.2 \text{ cm}$$

c) Interpret the slope of the regression line in the context of this study.

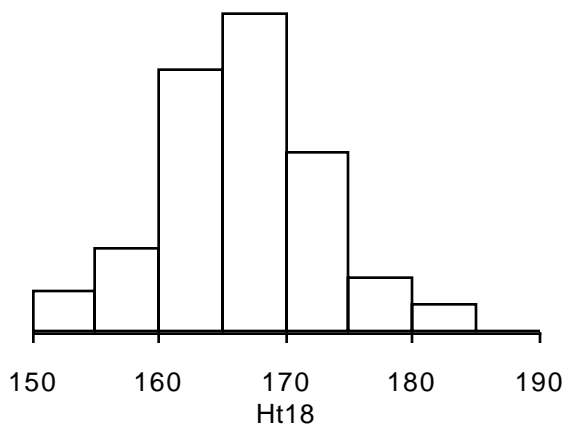
Boys who are one cm taller than others at age two are, on average, 1.12 cm taller than that "other" group at age 18.

4) Equal time for girls: Here are summary statistics for the heights at age 18 (cm) for 70 girls as reported by the same study as in question (3).

Variable	Average	Std. Dev	Minimum	Median	Maximum
Ht18	166.54	6.0749	153.6	166.75	183.2

First Quartile: 163 cm

Third Quartile: 170.2



a) Consider the height of the 18-year-old girls. Which, if either, is a better summary of the center of this distribution: the mean or the median. Why?

The distribution of this sample is fairly symmetric, so the sample median and the average will be nearly equal. So it doesn't matter which you use.

b) Sketch the box plot for height at 18 years. Label important features clearly. (Are there any outliers? If so, you don't have enough information to say precisely where they are, so indicate whether or not there might be, if this is appropriate for this boxplot.)

Can't do this on the web. But your box would have a line through the center at 166.75. The top of the box would be at 170.2, the bottom of the box at 163. The IQR is $170.2 - 163 = 7.2$.

The top whisker will go at $170.2 + 1.5 \cdot \text{IQR}$ or at the maximum value, whichever gives the shortest whisker. $170.2 + 1.5 \cdot \text{IQR} = 181.3$, and this is shorter than the maximum value (at 183.2). So the whisker goes at 181.3, and there is at least one outlier.

The bottom whisker will go at $163 - 1.5 \cdot 7.2 = 152.2$ or the minimum, which is 153.6. This time, the minimum value gives the shorter distance, so the whisker goes at 153.6 and there are no outliers at this end of the box.

c) According to this data, how tall is a girl whose height in standard units is -1.4?

-1.4 means she is 1.4 SDs below average. 1 SD is 6.0749, so 1.4 is 8.59486 inches below average. So she is $166.54 - 8.59486 = 158.04$ cm tall.

d) One inch is 0.394 centimeters (approximately). What is the average height in inches? What is the SD? (Two questions)

166.54 cm is $166.54 \text{ cm} \cdot (.394 \text{ cm/inch}) = 65.6$ inches
The SD is $.394 \cdot 6.0749 = 2.39$ inches.

e) This study also recorded the girls' weight at age 2 and at age 18. Which do you think has the greatest SD: girls' weight at age 2 or girls' weight at age 18? Why?

Age 18 had the greatest SD. As people grow older, there is more time for them to diversify.

5) From an April 2000 L.A. Times Poll:

"The Rampart police corruption scandal is contributing to a malaise in Los Angeles, helping to raise questions about the city's health and image, devastating public impressions

of the Los Angeles Police Department and fueling strong sentiment for the appointment of an independent commission to investigate the crisis...

" Moreover, a majority of city residents rejects the contention--advanced by Police Chief Bernard C. Parks, Mayor Richard Riordan and others--that the misdeeds are the work of a few bad officers. Fifty-one percent of city residents disagree with that notion; instead, they believe the problems are 'symptomatic of a larger problem within the Police Department.' Another 39% believe they are 'isolated incidents and are not representative of the Los Angeles Police Department as a whole.' ...

"The Times Poll interviewed 1,219 city residents from March 29 through April 5. The margin of sampling error is plus or minus 3 percentage points."

a) Suppose that in truth, 50% of the residents of L.A. think that the problems are "isolated incidents and are not representative of the Los Angeles Police Department as a whole." In a random sample of 1219 people (assume they are selected with replacement), how many, then, should we expect to agree with this opinion? What is the SD for this number? (Note: there are two questions.)

Let X represent the number in our sample who agree with this statement.
 $E(X) = np = 1219 \cdot .50 = 609.5$ people.

$SD(X) = \sqrt{n \cdot p \cdot (1-p)} = \sqrt{1219 \cdot .5 \cdot .5} = 17.4571$ people.

b) Suppose that in truth, 50% of the residents of L.A. think that the problems are "isolated incidents and are not representative of the Los Angeles Police Department as a whole." (This is a viewpoint that agrees, more or less, with the LAPD's official position.) In a random sample of 1219 people (assume they are selected with replacement), what percentage of the sample should we expect to agree with this opinion? What is the SD for this percentage? (Note: there are two questions.)

$\hat{p} = X/n$ = proportion of the sample who agree.
 $E(\hat{p}) = p = 609.5/1219 = .50$ or 50%.

$SD(\hat{p}) = SD(X)/n$ or $= \sqrt{p \cdot (1-p)/n} = \sqrt{.5 \cdot .5 / 1219} = 17.4571 = 0.0143$ or 1.43%.

c) Suppose that in truth, 50% of the residents of L.A. think that the problems are "isolated incidents and are not representative of the Los Angeles Police Department as a whole." Using the normal approximation, find the probability that 39% or less of a random sample of 1219 people will agree with this opinion. Do you think it is possible that, although 39% of the sample agrees with the LAPD's viewpoint (that the incidents are isolated), in truth 50% or more of the population agrees? Why?

$P(\hat{p} \leq .39) = P(Z \leq (.39 - .50)/.0143) = P(Z \leq -7.69)$
This is "off the chart", and so the probability is very close to 0.

6) Suppose that the weight of two-year old children follows a normal distribution with mean 13 kilograms (kg) and SD 1.6 kg.

a) What is the probability that a randomly selected two-year old will weigh more than 14 kilograms?

Let X represent the weight of a randomly selected two-year-old.

$$P(X > 14) = P(Z > (14-13)/1.6) = P(Z > .625) = 1 - P(Z \leq .625) = 1 - .7324 = .2676$$

Or, depending on how you round-off .625: $1 - .7357 = .2643$.

b) Take a random sample (with replacement) of 10 two-year olds. What is the expected value of their average weight?

Let \bar{X} represent the average of the 10 observations.

$$E(\bar{X}) = 13\text{kg}.$$

c) Take a random sample (with replacement) of 10 two-year olds. What is the SD of their average weight?

$$SD(\bar{X}) = \sigma/\sqrt{n} = 1.6/\sqrt{10} = 0.5060 \text{ kg}$$

Thus, we expect our average to be 13 kg, give-or-take .5060 kg.

MORE

d) Take a random sample (with replacement) of 10 two-year olds. What is the probability that the average weight will be more than 14 kg?

$$P(\bar{X} \geq 14) = P(Z \geq (14 - 13)/.5060) = P(Z \geq 1.98) = 1 - P(Z < 1.98) = 1 - .9761 = 0.0239$$

Note that the calculations are essentially the same as for part (a). Only the SD of \bar{X} is smaller (.506 compared to 1.6), and so \bar{X} is less likely to be far away from the mean value of 13.

e) Suppose we want the probability that the average weight of a random sample of two-year olds will be greater than 14 kg to be 1%. How many two-year olds should we sample (with replacement.)? In other words, we want $P(\text{avg} > 14) = 0.01$.

$$P(\bar{X} > 14) = 0.01$$

$$\text{means } P(Z > (14 - 13) / (1.6/\sqrt{n})) = 0.01$$

$$\text{means } P(Z < 1/(1.6/\sqrt{n})) = .99$$

From the table, we find that if $z = 2.33$, then $P(Z < 2.33) = .99$.

$$\text{So it must be that } 2.33 = 1/(1.6/\sqrt{n})$$

Doing some algebra:

$$(2.33 * 1.6) = \sqrt{n}$$

$$\sqrt{n} = 3.728$$

So $n = (3.728)^2 = 13.898$ so you need at least 14 observations.

The algebra might be less confusing if you write it out by hand, rather than rely on reading what's on the screen.

7) pH level is a scale that determines the acidity of a liquid. The scale ranges from 0 (very acidic) to 14 (no acidity). A score of 7 is considered "neutral." The mean pH level of patients who are having morning surgery is 3.94, with an SD of 2.51. This is important, because patients with a high levels of stomach acid (low pH) run the risk of getting something called Mendelson's Syndrome, an often fatal condition that results when anesthetized patients inhale the contents of their stomach. Do afternoon surgery patients have lower pH levels? (If so, they could be at greater risk than morning patients). A random sample of 49 afternoon surgery patients had an average pH of 2.93. Assume the population SD was still 2.51. Let μ represent the mean pH level of the population of afternoon patients.

a) Write the null and alternative hypotheses for this study.

$$H_0: \text{mean afternoon patients} = 3.94$$

$$H_a: \text{mean afternoon patients} < 3.94$$

b) What is the observed value of the test statistic?

This is somewhat different from the problems we've done in class. In this case, you are told that you know the population SD is 2.51. So there is no need to estimate it with the sample standard deviation. (And you're not given enough data to estimate it anyways.) So rather than use the T statistic, we can use the Z-statistic:

$$Z = (\bar{X} - \mu) / (\sigma / \sqrt{n})$$

$$Z(\text{observed}) = (2.93 - 3.94) / (2.51 / \sqrt{49}) = -2.82$$

In "english", this means that the observed value of \bar{X} was 2.82 standard deviations lower than what was expected.

c) What is the p-value for the test statistic?

$$p\text{-value} = P(\bar{X} < 2.93) = P(Z < -2.82) = 0.0024$$

d) Using a significance level of 5%, what would your conclusion be?

Reject the null hypothesis: there is evidence that afternoon patients have a significantly lower mean pH level than morning patients. We reject because the p-value is less than 0.05.