Statistics 101 Summer I 2011
Final Examination

Adam Kapelner, Instructor

June 30, 2011, 8:30-10:30AM

First Name Adam Last Name Kapelner

University of Pennsylvania's Code of Academic Integrity

Since the University is an academic community, its fundamental purpose is the pursuit of knowledge. Essential to the success of this educational mission is a commitment to the principles of academic integrity. Every member of the University community is responsible for upholding the highest standards of honesty at all times. Students, as members of the community, are also responsible for adhering to the principles and spirit of the following Code of Academic Integrity.

Activities that have the effect or intention of interfering with education, pursuit of knowledge, or fair evaluation of a students performance are prohibited. Examples of such activities include but are not limited to the following definitions:

Cheating Using or attempting to use unauthorized assistance, material, or study aids in examinations or other academic work or preventing, or attempting to prevent, another from using authorized assistance, material, or study aids. Example: altering a graded exam and resubmitting it for a better grade, etc.

I acknowledge and agree to uphold the University of Pennsylvania's Code of Academic Integrity.

signature 6/30/11 date

Instructions

This exam is two hours and closed-book. A cheat sheet of 2 pages 2-sided *is* allowed. You may use a graphing calculator of your choice. Please read the questions carefully. I advise you to *skip difficult problems until you have finished the exam*, then loop back and plug in all the holes. I also advise you to use pencil.

The exam is 100 points total. Partial credit will be granted for incomplete answers on most of the questions. Good luck!

Problem 1 Our hospital system in Philadelphia consists of St Joseph's Hospital, Penn Presbyterian Medical Center, Northeastern Hospital, and Mercy Fitzgerald Hospital. We are interested in studying Cirrhosis of the liver and alcoholism. We take a simple random sample of 235 patients out of the pooled patient population of the four hospitals and surveyed them.

In the survey, we asked the patients' age, gender, if they are or were alcoholics, whether or not they had Cirrhosis presently or in the past, whether or not their cholestoral was "high" or "low", and recorded the day of the week and hour of the day the survey was completed.

57 said they either presently or were alcoholics for some time in their life. 9 of these alcoholics also have some form of Cirrhosis. There were also 10 patients who had Cirrhosis but had no history of ever drinking to excess.

(a) [2 pt] Draw the first two rows of the raw data table from the compiled surveys followed by a "..." to indicate more rows. Make up values that are reasonable for each of the variables. Indicate column headers clearly. Add an "ID" column whose values are the record number beginning with the number 1 and make "ID" the first column.

(1) Cols correct

(1) row values

ID	Age	Gender	Alcoholic?	Cirrhosis?	Cholesterol	Wkd	Hr
1	47	M	Y	Y	H	M	2:00 pm
2	53	F	N	N	L	F	3:00 pm
⋮							

(b) [2 pt] What are the data types of each of the columns in the data table? List them in order of the columns in the drawing in the previous part from left to right separated by a comma.

(-1) for two mistakes

(-2) for more than 5 mistakes

nominal, interval (or ordinal), nominal, nominal, nominal, ordinal, nominal (ordinal), interval

(c) [3 pt] Draw a contingency table of frequencies for cirrhosis vs. alcoholism with cirrhosis as the columns. Calculate all the marginal probabilities as percentages and round to one decimal place. Write the marginal probabilities in the margin next to the marginal frequencies. Write the sample size n in the bottom ~~left~~ RIGHT corner margin.

(1) for row & col headers

(1) correct freq's

(1) correct marginal probs

Cirrhosis?

Alcoholism?

	C	C^C		
A	9	48	57	(.243)
A^C	10	168	178	(.757)
	19	216	235	
	(.081)	(.919)		

(d) [5 pt] According to Wikipedia:

> About 12% of American adults have had an alcohol dependence problem at some time in their life.

Write hypotheses and run a test at $\alpha = 5\%$ to see if the alcoholism rate is greater in our hospital system. Use the p_{val} approach. Ignore any problems you may see with running this test, we will explore the problems afterwards.

(1) correct Hypotheses

$H_0: p = .12$

$H_a: p > .12$

(2) correct z-stat

(-1) t-test

(-1) wrong SE

(1) pval computation

$$z = \frac{\hat{p} - p}{\sqrt{\frac{p(1-p)}{n}}} = \frac{.243 - .12}{\sqrt{\frac{.12 \cdot .88}{235}}} = \frac{.123}{.021} = 5.8$$

$$p_{val} = P(\hat{p} > .243 \mid p = .12) = P(Z > 5.8) \approx 0$$

$\Rightarrow$ Reject H_0

(1) conclusion

(e) [2 pt] Write what the p_{val} means for this example in English.

(1) general definition

(1) def in this problem context

It's the probability our sample has more than 24.3% alcoholics given the true proportion of alcoholics is 12%.

(f) [2 pt] Why is the p_{val} from the previous test so low? Is there cause for alarm? Is Philadelphia in trouble? Discuss.

(1) Saying No

(1) explaining pt. bias

It is low because this is a hospital and we may expect more alcoholics as a proportion. The pval is so low because it is a different population. It is not an SRS from the normal population. Therefore, there is no cause for alarm.

(g) [2 pt] Any association we see in this data between alcoholism and Cirrhosis may occur due to sampling variation. We wish to test the association using a statistical test. Formulate your hypotheses below by writing English sentences.

(1) Correct Ho
(1) Correct Ha

H_0: Alcoholism and Cirrhosis are independent

H_a: Alcoholism and Cirrhosis are not independent

(h) [6 pt] Carry out the test you created in the previous part at $\alpha = 5\%$ using the p_{val} approach (estimate the p_{val} as best you can with the table or use your graphing calculator to calculate it exactly). Show as much work as you can. Write a sentence in English as a concluding remark.

(3) Exp cts table
(1) χ^2 computation
(1) pval comp.
(1) conclusion

Expected Cts

	C	C^C
A	4.626	52.479
A_C	14.409	163.486

Alcoholism & Cirrhosis appear to be dependent

$$Q = \sum\sum \frac{(\text{obs}-\text{exp})^2}{\text{exp}} = \frac{(9-4.626)^2}{4.626} + \frac{(48-52.479)^2}{52.479} + \frac{(10-14.409)^2}{14.409} + \frac{(168-163.486)^2}{163.486} = 5.992$$

$p_{\text{val}} = P(\chi^2_1 > 5.992)$ is between 1% and 2.5% $< 5\% \Rightarrow$ Reject H_0
acc'd to the table

(1) Saying obs study
(1) defining obs study

(i) [2 pt] The hospital administrator wants to publish this experimental data because he thinks he found interesting results. Is this an experiment? If it is, define what an "experiment" is. If it's not, what is it? Then, define whatever that thing is.

It is not an experiment. It is an observational study. An observational study is when you just observe data in the real world and you have no control of which subject gets assigned to which treatment.

(1) saying NO

(2) talking about confounding/lurking var's or how you would need to run an experiment

(j) [3 pt] The administrator saw the unequivocal results of these analyses and he feels that it is obvious that Cirrhosis is caused directly by alcoholism. Based solely on this study, would you agree with him? Discuss.

Since this is an obs. study, we cannot infer causality. There may be a lurking variable confounding the observed association.

(k) [3 pt] Other hospital researchers are interested in those patients whose livers are heathly (*i.e.* they don't have and never had Cirrhosis). Does being an alcoholic matter in that case? Phrased equivalently, is there any difference between the likelihood of being Cirrhosis free if you're not an alcoholic versus the likelihood of being Cirrhosis-free if you are an alcoholic? Frame this as a hypothesis test. Use the notation p_1 and p_2. Tell me the *name* of this test but *do not execute the test.*

(1) Correct Ho

(1) Correct Ha

(1) A name that's mostly correct

$H_0: p_1 - p_2 = 0$

$H_a: p_1 - p_2 \neq 0$

Name of test:
2-prop z-test

(l) [4 pt] Calculate the sample proportions $\hat{p}_1$ and $\hat{p}_2$ which are necessary to investigate this situation. It would help to consult the contingency table you built previously.

(2) $\longrightarrow \hat{p}_1 = P(C^c \mid A^c) \approx \frac{168}{178} = .9438$

(2) $\longrightarrow \hat{p}_2 = P(C^c \mid A) \approx \frac{48}{57} = .8421$

(1) for getting correct prob statements

(m) [2 pt] Assume that we only care about making inference for our hospital system. Are there any *other* issue(s) with running the hypothesis test to test this difference?

(2) $\rightarrow$ Yes: $n(1-\hat{p}_2) = 57(1-.8421) = 9 < 10$

or

Hence, we cannot use the normal approx to the binomial.

(1) possibility of lurking variable

(1) sample too small

Problem 2 Although casinos would like to believe that their games are perfectly designed to give them the edge, there are many documented cases where they make mistakes and actually allow people to play games that are expectation positive (for players).

One of the most famous of these is "five-card-charlies" in blackjack. If the player got five cards without busting, they automatically win. This was discontinued about 50 years ago.

A lesser-known "hole" in the system, which is presently available for player exploitation, is in video-draw poker with deuces wild only in Nevada. According to Wikipedia, playing for 5 credits a hand has a mean return of 100.8% in theory. By "theory" we mean playing the *optimal* hand for every possible combination of cards. The optimal hands are calculated by a computer that can analyze all $\binom{52}{5}$ possible hands and all possible draws. The computer can then print out rules to follow. There are literally thousands of rules.

Someone with a photographic memory who can memorize some of the complicated algorithm may be able to break even, but someone who is really cheating (through unknown means) can really take the casino for a lot of money.

Consider the standard bet for five credits to be \$5. Hands take 12 seconds on average to play.

(a) [1 pt] Why would an expectation-positive game such as 5-credit Deuces Wild Video Draw Poker be bad for the casinos in Nevada?

In the long run, the casino is guaranteed to lose money on such games

(b) [2 pt] Show that the exptected *gain* of one bet (five credits) when playing like the computer under the optimal algorithm is \$0.04. Read the question carefully and you will see this problem is very simple. If you are having trouble, make sure you remember to subtract off the bet amount.

$$\$5 \cdot 100.8\% - \$5 = \boxed{\$0.04}$$

The casino is tracking a suspicious person. He comes in every Monday night alone and uses a different video poker machine, always deuces wild, he plays 5-credit hands, never plays any other games, and never drinks alcohol (unlike your course instructor).

The security cameras capture video of him staring intently at the screen for all of his time there. His photo ID says his name is Bassem. He plays for three hours straight every time he comes in. The casino keeps precise electronic records for each of its video poker machines.

Over the past year (52 weeks), he has won $\bar{x} = \$1.27$ on average per hand with sample standard deviation $s = \$110.85$.

(c) [2 pt] How many video poker hands has he played during the whole year?

(1) Setting up your conversions
(1) execution

$$1\,\text{yr} \cdot 52\,\frac{\text{wk}}{\text{yr}} \cdot 1\,\frac{\text{day}}{\text{wk}} \cdot \frac{3\,\text{hr}}{\text{day}} \cdot 60\,\frac{\text{min}}{\text{hr}} \cdot 60\,\frac{\text{s}}{\text{min}} \cdot \frac{1}{12}\,\frac{\text{plays}}{\text{s}} = \boxed{46{,}800 \text{ plays}}$$

(d) [2 pt] You would like to calculate a 95% confidence interval for his true mean winnings per 5-credit hand ($\bar{x}$ is a realization from $\bar{X}$ which is centered at μ, which is his "true" mean winnings). Are there any theoretical problems with creating this CI?

Partial credit:
(1) pop may not be normal
(1) don't know σ

Outcomes from a video-poker machine are <u>not</u> distributed normally.

(e) [4 pt] Regardless of whether or not there are problems with creating this CI, calculate a 95% confidence interval for his true mean winnings.

(1) structure
(1) substitution
(2) execution

since $n \geq 30$ use $z \approx t$

$$CI_{\mu,95\%} = \left[\bar{x} \pm z\frac{s}{\sqrt{n}}\right] = \left[\$1.27 \pm 1.96 \cdot \frac{\$110.85}{\sqrt{46800}}\right] = \boxed{[\$0.27, \$2.27]}$$

(f) [2 pt] Regardless of whether or not there are problems with creating this CI, calculate a 95% confidence interval for his true *total* winnings for the whole year in Egyptian Pounds where $1 = 5.97E£.

$$CI_{TW\,in\,E£} = (46{,}800)(CI_{\mu,95\%})\left(5.97\,\frac{E£}{\$}\right) = \boxed{[E£\,74{,}232,\ E£\,635{,}938]}$$

(1) multiply by # hands
(1) multiplying by currency conversion

(g) [2 pt] The casino views the highest possble return of 97% on a 5-credit play as the maximum return for a human being (not a computer). This is the return the best players can consistently achieve. This suspicious guy is indubitably in the category of "best player".

A 97% return would mean that the mean winnings per hand means $\mu = -\$0.15$ and consider the same $\sigma = \$113$ as we did before.

Any higher return and the casino bans the player (because they assume the player is cheating). Frame a hypothesis test to see if this guy will be booted but *do not execute the test.* Make sure your hypothesis test is framed using μ and not percentage return.

(1) correct Ho $H_0: \mu = -\$0.15$ or $\mu \leq -\$0.15$

(1) correct Ha $H_a: \mu > -\$0.15$

(h) [2 pt] Considering the hypothesis test framed in the previous example, what would a type I error be and what is the cost of making such a mistake? Do not compute the cost in dollars, just explain in English what the cost would be.

(1) what happens We boot him but he's not actually a cheater.

(1) cost explanation Cost: One upset innocent person and a bad reputation

(i) [2 pt] Considering the same hypothesis test, what would a type II error be and what is the cost of making such a mistake? Do not compute the cost in dollars, just explain in English what the cost would be.

(1) what happens We let him continue playing even though he's cheating.

(1) cost explanation Cost: We can lose money in the long run

(j) [7 pt] It is clear that this guy is at least as good as the best players out there. The casino is especially concerned if this guy is playing at the theoretical maximum of 100.8% return just like the computer program. Consider the test at $\alpha = 5\%$. Calculate the probability the casino can detect if he is playing using this theoretically optimal strategy with the aggregated video and machine data they have on him in the past year which was explained before assuming he actually *is* playing with the theoretically optimal strategy.

(1) Hypotheses

(1) picture

(2) $\bar{X}^*$ calculation

(2) z_a calculation

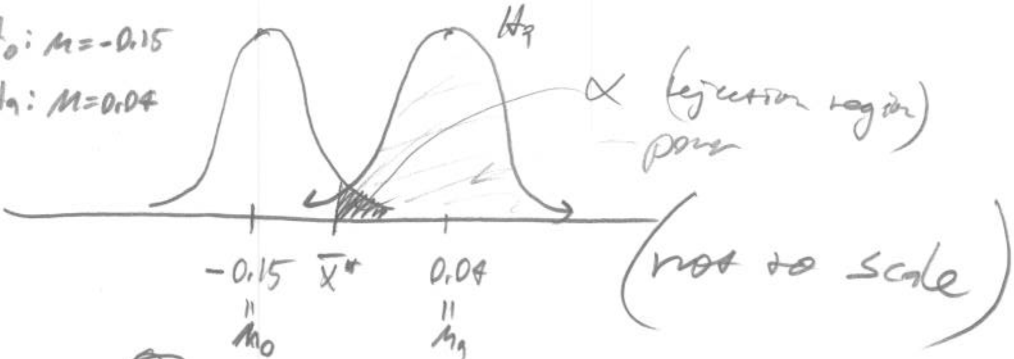
(1) Prob calc.

Please do this problem after you have finished the rest of the exam. Get as far as you can; generous partial credit will be rewarded. I strongly suggest you draw a sketch to scale.

$$\frac{\sigma}{\sqrt{n}} = .5223$$

$H_0: \mu = -0.15$

$H_a: \mu = 0.04$



(4) scaled picture

$$\bar{X}^* = \mu_0 + Z_{0.05} \cdot \frac{\sigma}{\sqrt{n}}$$
$$= -0.15 + 1.645 \cdot .5223$$
$$= 0.709$$

$$\text{Power} = P(H_a \mid H_a)$$
$$= P(Z > 1.281)$$
$$\approx \boxed{10\%}$$

What ~~z~~ value does $\bar{X}^*$ represent on H_a?

$$Z_a = \frac{\bar{X}^* - \mu_a}{\frac{\sigma}{\sqrt{n}}} = \frac{0.709 - 0.04}{\frac{113}{\sqrt{46,800}}} = 1.281$$

(k) [1 pt] Using the result from the previous question, do you think it's difficult to catch cheaters in deuces wild draw poker in Nevada? Explain. If you did not get an answer to the previous question, make up a reasonable answer to it, then answer this question appropriately.

It is very difficult to catch cheaters. Even with a player who plays regularly for a whole year, there's only a 10% chance of catching him.

Problem 3 We have two mutually indistinguishable dice, call them die A and die B. They were invented by the MI6 British intelligence agency and designed to *never* roll doubles. Using these dice, spies from Ghana can penetrate El Salvadorian and Colombian drug gangs who play dice games where rolling doubles can get you killed.

Here's how the dice work: if die A rolls a 1, the die B will roll a 2; if A rolls a 2, die B will roll a 3; if A rolls a 3, die B will roll a 4; if A rolls a 4, die B will roll a 5; if A rolls a 5, die B will roll a 6; if A rolls a 6, *die B will roll a 1.* The probability A rolls any of the six faces is still one-sixth as usual.

In homework 2, we calculated a fair die's expectation to be 3.5 and its variance to be 2.92. Use these values in the questions below; do not rederive them.

(a) [2 pt] Are these dice independent? Prove or disprove using a mathematical statement.

(1) saying NO

No:

(1) making a correct prob statement

$$0 = P(X_A = 4, X_B = 4) \neq P(X_A = 4)P(X_B = 4) = \frac{1}{36}$$

(b) [1 pt] Are these dice identically distributed? Answering "Yes" or "No" without justification is okay.

Yes

(c) [2 pt] Calculate the expectation of the sum of the two faces when both are rolled together

works regardless of ind./dep.

(2) saying 2M

$$E[X_A + X_B] = E[X_A] + E[X_B] = 3.5 + 3.5 = \boxed{7}$$

(d) [5 pt] Calculate the covariance between the two die.

(1) $Cov = E[XY] - \mu_X \mu_Y$

$$Cov[X_A, X_B] = E[X_A X_B] - \mu_A \mu_B = 12.67 - (3.5)^2 = \boxed{.417}$$

(1) mult table

(2) probs table

let $M = X_A X_B$

(1) calculation of $E[M]$

$$\Rightarrow E[M] = \frac{1}{6}(2+6+12+20+30+6) = \frac{1}{6}(76) = 12.67$$

	1	2	3	4	5	6
1		2				
2			6			
3				12		
4					20	
5						30
6	6					

$P(X_A = x_A, X_B = x_B)$

	1	2	3	4	5	6
1		1/6				
2			1/6			
3				1/6		
4					1/6	
5						1/6
6	1/6					

(e) [2 pt] Calculate the correlation of these two dice. Write one sentence explaining why it's positive (or negative).

(1) computation

(1) explanation

$$\rho = \frac{Cov[X_A, X_B]}{SD[X_A]\,SD[X_B]} = \frac{0.4167}{\sqrt{2.92 \cdot 2.92}} = \boxed{.142}$$

It's positive since die B usually rolls one more than die A

Problem 4 A oenophile (wine lover) tabulated information on a random sample of 49 wine bottles sold in an online store. Consider the offerings of this online store to be representative of wine consumption in America.

For each bottle, he recorded its price and a rating score assigned by the renowned Robert Parker. Upon visualizing the data, the scatterplot diagram appeared to be football shaped with big anomalies. Here are the summary statistics: the average price is $40 with an sample standard deviation of $18. The average Parker number is 91 points with an sample standard deviation of 2 points. The correlation is .75.

getting:

(1) b_0

(1) b_1

(1) $\hat{y}$

(1) getting the answer

(a) [4 pt] A wine with a parker number of 87 points is expected to sell at what price?

$$b_0 = \bar{y} - b_1\bar{x} = \$40 - 6.75 \cdot 91 = -574.25$$

$$b_1 = r\frac{s_y}{s_x} = (.75)\frac{\$18}{2} = 6.75$$

$$\hat{y} = b_0 + b_1 x = (-574.25) + (6.75)(87) = \boxed{\$13}$$

(b) [2 pt] What was the explanatory variable and what was the response variable you used in the previous question?

(1) response

(1) expl

parker rating and selling price, respectively.

(c) [4 pt] The root mean squared error for the model you built in (a) is 11.9.

You come across a new wine with a Parker number of 95 points that is selling for $76. What fraction of wines with Parker Number of 95 points sell for $76 or less?

(2) $\hat{Y}$ distr

(1) z-stat

(1) probability

$$\hat{Y} \sim N(b_0 + b_1 x, RMSE^2) = N(\$67, \$11.90^2)$$

$$P(\hat{Y} < \$76) = P\left(Z < \frac{76-67}{11.9}\right) = P(Z < .756) \boxed{\approx \text{between } 77\% \text{ and } 78\%}$$

(d) [3 pt] You come to a party with 10 bottles of wine all of Parker number 91. What is the probability that 3 of the 10 bottles have prices greater than \$40?

$(\bar{X}, \bar{Y})$ is on the regression line hence $\hat{y}(91) = \$40$

(1) prob is $\frac{1}{2}$

$\Rightarrow P(\hat{Y} > 40) = \frac{1}{2}$ since PDF Z is symmetric

(1) binomial

$P(S_{10} = 3) = \binom{10}{3}\left(\frac{1}{2}\right)^3\left(\frac{1}{2}\right)^7 = \boxed{11.7\%}$

(1) execution

(e) [2 pt] The oenophile does the same thing next year and realizes that wines from the previous year with low Parker ratings magically on the average now have higher ratings, and wines with high Parker ratings from last year magically on the average now have lower ratings. This is definitely strange since winemakers in general don't change too much about their wines year-to-year. Explain this strange observation to the best of your ability.

(1) regression to the mean

(1) some reasonable explanation

This is regression to the mean. Bad wines from last year were bad and unlucky, but this year they're on average more lucky. Good wines from last year were good and lucky, but this year they're on average less lucky.

Problem 5 Honda takes the safety of its vehicles very seriously. A major safety feature are functional airbags. The most important thing about airbags is how quickly they inflate. According to Wikipedia:

> The burning propellant generates inert gas which rapidly inflates the airbag in approximately 20 to 30 milliseconds. An airbag must inflate quickly in order to be fully inflated by the time the forward-traveling occupant reaches its outer surface. Typically, the decision to deploy an airbag in a frontal crash is made within 15 to 30 milliseconds after the onset of the crash, and both the driver and passenger airbags are fully inflated within approximately 60-80 milliseconds after the first moment of vehicle contact.

Honda is concerned that airbags manufactured in their factory in their Uttar Pradesh, India plant (near New Delhi) are different from the airbags manufactured in their El Salto, Mexico plant (near Guadalajara).

s_2 X_1 n_1 x_2 s_1 n_2

They can't crash too many cars because it is expensive. They crash-test 5 cars from the Mexico plant and measure an average inflation time of 68.4ms with sample standard deviation 9.4ms and they crash-test 6 cars from the India plant and measure an average inflation time of 76.0ms with sample standard deviation of 8.3ms. Their years of crash-test data indicate that airbag inflation time is normally distributed.

(a) [2 pt] Frame Honda's concern as a hypothesis test.

India : smp 1
Mexico : smp 2

(1) correct Ho

(1) correct Ha

$$H_0 : \mu_1 - \mu_2 = 0$$

$$H_a : \mu_1 - \mu_2 \neq 0$$

(b) [6 pt] Run this test at $\alpha = 5\%$ using the *critical value* approach. Write a one-sentence summary that should be in the report sent to the technical manager in Japan.

$\frac{s_2^2}{s_1^2} = 1.28 < 3$

$\Rightarrow$ pooled t-test is appropriate

$$t = \frac{\bar{X}_1 - \bar{X}_2}{s_p\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}}$$

$$s_p = \sqrt{\frac{(n_1-1)s_1^2 + (n_2-1)s_2^2}{n_1+n_2-2}} = \sqrt{77.54} = 8.806$$

$$= \frac{76.0 - 68.4}{8.806\sqrt{\frac{1}{6} + \frac{1}{5}}}$$

$$df = n_1 + n_2 - 2 = 9$$

$$= \frac{7.6}{5.33} = 1.425 \not> t_{9, 2.5\%} = 2.262$$

$\Longrightarrow$ fail to reject

(1) pooled test
(1) SE calc
(1) correct t-statistic
(1) correct t^*
(1) correct concl.
(1) correct concluding sentence

<u>Report Excerpt</u>: The airbag inflation times between the Mexican and Indian plant do not appear to be different.

[illegible] reject