

Statistics 151 Final Lecture V01, Version 1 April 21, 2009

Instructor: Paul Cartledge

1. *Read all the instructions carefully.*
2. *This is a closed book exam.*
3. *You may use the formula sheets and the tables provided and a non-programmable calculator only.*
4. *You have 3 hours to complete the exam.*
5. *The exam is out of a total of 75 marks and has 13 pages.*
6. *Show all your work for the long answer section to receive full credit.*
7. *Use the backs of pages for scrap work.*
8. *Make sure your name and signature are on the front and that your ID number is on the top of page two.*
9. *Make sure you mark all your multiple choice responses on the response page (page 2), as THESE will be taken as your answers.*
10. *When asked to “carry out a full analysis in detail”, set up the hypotheses, briefly discuss assumptions, calculate the test statistic, state the distribution of the test statistic (i.e. t_9 or z), approximate the p -value, and state your conclusion in plain English. If no significance level is stated, use the judgment approach.*

Name: **SOLUTION**

Signature: _____

Submit the entire exam booklet, the tables, AND your formula sheet.

Multiple Choice	15 questions	15	
Long Answer Q1	2 parts	10	
Long Answer Q2		10	
Long Answer Q3	7 parts	20	
Long Answer Q4	2 parts	10	
Long Answer Q5		10	
BONUS		4 or 2	
Total		75	

ID: _____

Part 1: Multiple Choice

In each multiple choice question, choose the answer you think is closest to being correct. There are no deductions for incorrect guesses. Mark your choices clearly in the answer section below by writing the letter corresponding to your chosen answer. Make sure your answers are correctly located and clearly marked. These will be your marked answers. Each correct answer is worth 1 mark.

Response Section

- | | |
|----------|-----------|
| 1. _____ | 9. _____ |
| 2. _____ | 10. _____ |
| 3. _____ | 11. _____ |
| 4. _____ | 12. _____ |
| 5. _____ | 13. _____ |
| 6. _____ | 14. _____ |
| 7. _____ | 15. _____ |
| 8. _____ | |

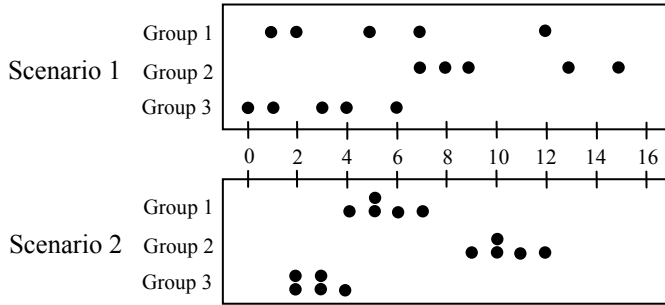
1. Is the proportion of people who want to see *Harry Potter and the Half-Blood Prince* significantly greater in Edmonton as opposed to Calgary? If the test statistic is $z = 0.74$, which of the following pieces of information correspond to the appropriate hypothesis test?

- A) $H_a: p_{Edm} > p_{Calg}$, $p\text{-value} = 0.2296$, the proportions may still be the same.
- B) $H_a: p_{Edm} \neq p_{Calg}$, $p\text{-value} = 0.7704$, the proportions may still be the same.
- C) $H_a: p_{Edm} > p_{Calg}$, $p\text{-value} = 0.2296$, Edmonton has a significantly higher proportion.
- D) $H_a: p_{Edm} \neq p_{Calg}$, $p\text{-value} = 0.4592$, the proportions may still be the same.

2. At Hogwart’s School of Witchcraft and Wizardry, each student belongs to only one of four houses: Gryffindor (G), Slytherin (L), Ravenclaw (R), and Hufflepuff (H). Which of the following probability statements CORRECTLY relates the events described?

- A) $P(H \cap G \cap L \cap R) = P(H) \times P(G) \times P(L) \times P(R)$
- B) $P(H^C) = P(G) + P(L) + P(R)$
- C) $P(H \cap G \cap L \cap R) = P(H) \times P(G | H) \times P(L | H \cap G) \times P(R | H \cap G \cap L)$
- D) $P(H \cup G \cup L \cup R) = P(H) \times P(G) \times P(L) \times P(R)$

3. Consider the following two scenarios:



Note that for each of the two scenarios, $\bar{y}_1 = 5.4$, $\bar{y}_2 = 10.4$, $\bar{y}_3 = 2.8$.

Which of the following statements is TRUE?

- A) The variance within groups is smaller in the first scenario.
- B) The variance within groups is larger in the first scenario.**
- C) The variance between groups is smaller in the first scenario.
- D) The variance between groups is larger in the first scenario.

Use the following to answer questions 4 – 5:

Do musicians charge too much for ticket prices? Suppose a study was done to investigate the association of ticket price with the number of albums sold. The study consists of 50 random and independent observations. Ticket price (*Price*, measured in US\$) and the number of albums sold (*Album*, measured in millions) are modeled as continuous (numerical) variables. Assume all regression model assumptions hold. The following output was obtained from StatCrunch:

Parameter	Estimate	Std. Err.	DF	T-Stat	P-Value
Intercept	39.394005	6.8168974	48	5.778876	<0.0001
Slope	0.305723	0.032540515	48	9.39515	<0.0001

4. A 98% confidence interval for the average change in mean ticket price when an additional 1 million albums are sold is approximately

- A) 39.700 ± 6.817
- B) 0.306 ± 0.033
- C) 39.700 ± 16.381
- D) 0.306 ± 0.079**

5. If the coefficient of determination is 64.8%, what can you conclude?

- A) $r = 0.805$ and a strong, positive association exists between the two variables.**
- B) $r = -0.805$ and a strong, negative association exists between the two variables.
- C) $r = 0.805$ and a weak, positive association exists between the two variables.
- D) $r = -0.805$ and a weak, negative association exists between the two variables.

Use the following information for questions 6 – 7:

An intern working at Stark Enterprises is trying to empirically determine if one element has a higher melting point than another. Randomly selecting 50 small pieces of the element Fe , the intern measures the temperature at which the piece melts. The intern repeats the process for the element Y . He then obtained the following summary statistics, regarding melting point temperature, from the two elements (units are in Kelvin).

Summary statistic	Fe	Y	Difference
Average	1811.1	1799.5	11.6
Standard Deviation	25.6	25.8	36.3

6. If the intern wants to test if Fe has a significantly higher melting point and found a test statistic of 2.257, what will be the degrees of freedom and the approximate range of his p -value, respectively?

- A) 49 and (0.02, 0.05)
- B) 49 and (0.01, 0.025)
- C) 98 and (0.02, 0.05)
- D) 98 and (0.01, 0.025)

7. If the standard error of the appropriate estimate is 5.140, what is the corresponding 98% confidence interval for the parameter related to this estimate?

- A) (-0.551, 23.751)
- B) (0.049, 24.351)
- C) (-0.751, 23.952)
- D) (-0.151, 24.551)

8. The Mirror of Erised sometimes reveals information to some individuals, but not to others. It was once able to give Harry Potter the Philosopher’s Stone while “you-know-who” assumed he had it, unless there was enough evidence refuting this assumption. Which of the following events would identify a Type II error?

- A) “You-know-who” believes Harry doesn’t have the stone, when he actually does.
- B) “You-know-who” believes Harry has the stone, when he actually doesn’t.
- C) “You-know-who” believes Harry doesn’t have the stone, when he actually doesn’t.
- D) “You-know-who” believes Harry has the stone, when he actually does.

9. Which of the following pairs of hypotheses is VALID?

- A) $H_0: p = 1.5$ $H_a: p > 1.5$
- B) $H_0: \mu = -5$ $H_a: \mu \neq -5$
- C) $H_0: \hat{p} \geq 0.5$ $H_a: \hat{p} < 0.5$
- D) $H_0: \mu > 0$ $H_a: \mu \leq 0$

10. In the upcoming 2009 NHL Stanley Cup Playoffs, the Montréal Canadiens could get through the first round by playing 4, 5, 6, or 7 games. If teams can only win or lose each game, the result of each game is independent of the others, and the probability of Montréal winning a game is 0.574, what is the probability of losing the first two games and winning the next four?

- A) 0.0109 **B) 0.0197** C) 0.9803 D) 0.1815

11. Which of the following confidence intervals is CORRECT?

- A) $\hat{p} \pm t_{\alpha/2, n-1} \times \sqrt{\frac{\hat{p}(1-\hat{p})}{n}}$ B) $\bar{x} \pm t_{\alpha/2, n-1} \times \left(\frac{\sigma}{\sqrt{n}}\right)$
 C) $b \pm t_{\alpha/2, n-1} \times s_b$ **D) $\bar{x}_d \pm t_{\alpha/2, n-1} \times \left(\frac{s_d}{\sqrt{n}}\right)$**

12. It's Tuesday, April 21, 2009 and, for some reason, the proportion of students who are aware they are writing an exam in a large room containing over 1500 desks has suddenly increased dramatically. Suppose the true proportion of people who are aware is 0.175. If a random sample consisted of 150 people, of whom 17 people are aware, then what can be said about the sampling distribution of p ?

- A) It may not be normal with a standard error of 0.0259.
 B) It is approximately normal with a standard error of 0.0259.
C) It is approximately normal with a standard error of 0.0310.
 D) It may not be normal with a standard error of 0.0310.

13. If a 90% confidence interval for the difference between two sample proportions is (0.043, 0.425), what is the 99% confidence interval?

- A) (-0.112, 0.581)
 B) (-0.150, 0.618)
C) (-0.065, 0.533)
 D) (-0.017, 0.485)

Use the following information to answer questions 14 – 15:

A pseudo-scientist has rethought his original theory that an energy field created by all living things is actually a phenomenon denoted by high or low concentrations of “midi-chlorians”. To lay claim to this allegedly superior theory, he randomly and independently samples creatures from five different planets. The table below summarizes average “midi-chlorian” concentration (in counts of 1000s per cell) from each group in his study. Rampaging Wookiees prevented the pseudo-scientist from completing the table so he needs your help to see if his theory works, but note that sample sizes differ because some planets have smaller populations.

Source	df	SS	MS	F
Between		841.635		
Within			0.294	
Total	189			

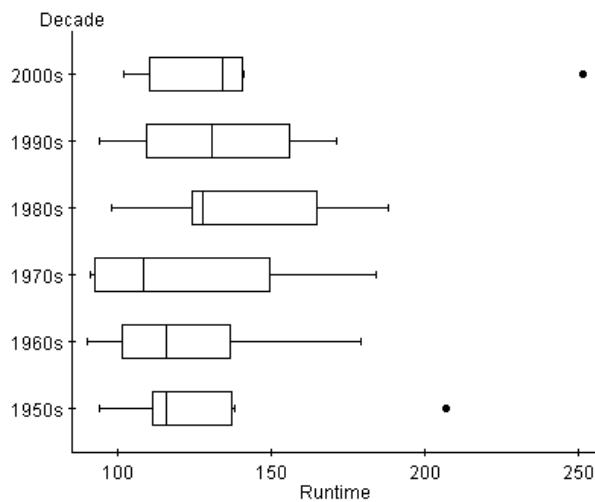
14. From the information above, which of the following is TRUE?

- A) You will not reject H_0 at $\alpha = 0.05$ and conclude that all the means are different.
- B) You will reject H_0 at $\alpha = 0.05$ and conclude that all the means are different.
- C) You will not reject H_0 at $\alpha = 0.05$ and conclude that at least one of the means are different.
- D) You will reject H_0 at $\alpha = 0.05$ and conclude that at least one of the means are different.**

15. Respectively, the test statistic for equality between all teams, the total sum of squares, and degrees of freedom for the denominator are

- A) 715.676, 896.025, 185** B) 715.676, 896.025, 189
- C) 572.541, 895.731, 184 D) 572.541, 895.731, 189

Use the graph below for **Long Answer Question 2**



Part 2: Long Answer

Answer the questions in as much detail as possible to earn full marks. Follow the guidelines on page 1. Clearly mark your answers for visibility and legibility. Mark worth is denoted after each specific question.

Question 1 (10 marks):

In the Triwizard Tournament, underage wizards are not allowed. Just to be sure, the Muggles Studies teacher decided to take a random sample of 32 entries and recorded their age. The teacher found a sample mean age of 16.236 years and a sample standard deviation of 0.509 years.

A) Is the average age more than 16 years? Carry out a full analysis in detail. (6 marks)

$$H_0: \mu \leq 16 \quad H_a: \mu > 16$$

Assumptions: $n = 32 > 30$, random sample, σ unknown \rightarrow use t -distribution

$$t = \frac{\bar{X} - \mu}{s / \sqrt{n}} = \frac{16.236 - 16}{0.509 / \sqrt{32}} = \frac{0.236}{0.0900} = 2.623 \sim t_{n-1} = t_{31} \quad (\text{use } df = 30 \text{ or } 40 \text{ in table})$$

$$\begin{aligned} df = 30 \\ 2.457 < t = 2.623 < 2.750 \\ 0.01 > p\text{-value} > 0.005 \end{aligned}$$

$$\begin{aligned} df = 40 \\ 2.423 < t = 2.623 < 2.704 \\ 0.01 > p\text{-value} > 0.005 \end{aligned}$$

$p\text{-value} < 0.01$

\rightarrow strong to convincing evidence against H_0 .

\rightarrow reject H_0 .

\rightarrow the average age is significantly more than 16 years.

B) Suppose the mean age of all entries is actually 16.500 years and that the population standard deviation is 0.661 years. For another random sample of 64 entries, what is the probability that the average is greater than 16.410 years? (4 marks)

$$\begin{aligned} P(\bar{X} > 16.410) &= P\left(\frac{\bar{X} - \mu}{\sigma / \sqrt{n}} > \frac{16.410 - 16.500}{0.661 / \sqrt{64}}\right) = P\left(Z > \frac{-0.09}{0.0826}\right) = P(Z > -1.09) \\ &= P(Z < 1.09) = 0.8621 \end{aligned}$$

Question 2 (10 marks):

A film historian, who once took STAT 151, decided to use this vastly enjoyable and applicable wealth of knowledge to compare the lengths of films from multiple decades. N films were randomly sampled, an equal number to each of the past six decades. The table below is ANOVA output for comparing film lengths (in minutes) from each decade. (Insert joke here about a famous movie character creating an incomplete table for you.)

Source	df	SS	MS	F
Between	5	2697.665	539.533	0.440
Within	42	51504.002	1226.286	
Total	47	54201.667		

Is there any evidence of a difference between decades? Carry out a full analysis in detail. (Make sure to involve the graph on page 6.) Use a significance level of 0.05.

$$H_0: \mu_{50s} = \mu_{60s} = \mu_{70s} = \mu_{80s} = \mu_{90s} = \mu_{2000s}$$

$$H_a: \text{at least one } \mu \text{ is different}$$

Assumptions: The samples are random and are likely independent. The boxplot shows that normality is clearly a problem with at least 3 of the decades. The boxplot also shows that the IQR of the 1970s seems to be double the IQR of the 1950s, so the spreads do not appear to be approximately equal. Overall, the assumptions do not appear to be met, but if this question is worth 10 marks, I probably should keep going, even though my results might not be valid. If they aren't valid, I blame Michael Bay.

$$F_0 = \frac{MSB}{MSW} = \frac{539.533}{1226.286} = 0.440$$

$$F_0 \sim F_{5,42} \rightarrow \approx F_{5,40}$$

$$F_{5,40} = 2.45 > 0.440 = F_0$$

$$0.05 < p\text{-value}$$

$\alpha = 0.05 < p\text{-value}$
 \rightarrow do not reject H_0 .

\rightarrow There is no evidence of a difference between decades. The average movie length could be the same for all six decades.

Question 3 (20 marks):

While Trey Porrath is usually only concerned about triple-decker “knight buses”, Sean Everuke is more concerned about *speeding* buses. Bus drivers going too fast will result in unhappy and, therefore, fewer passengers. By attaching a device to randomly selected buses, Sean was able to observe the average speed of the bus (x) during the course of the day while a video camera helped him measure the number of people (y) who entered the bus over the same day. Below are some of the statistics Sean was able to calculate. The explanatory variable has units of km/h. Assumptions for regression inference were satisfied.

$$n = 25, \bar{x} = 62.162, \bar{y} = 628.207, \\ s_x = 8.574, s_y = 56.691, r = -0.862$$

A) Determine the change in the response variable when the explanatory variable increases by 1 unit. (1 mark)

$$b = r \left(\frac{s_y}{s_x} \right) = -0.862 \left(\frac{56.691}{8.574} \right) = -5.700$$

B) Test if there is a significant positive linear relationship between the two variables. Carry out a full analysis in detail. (7 marks)

[HINTS: Do part C) first and $S.E.(b) = \frac{s_e}{s_x \sqrt{n-1}}$.]

$H_0: \beta \leq 0 \quad H_a: \beta > 0$

Assumptions: told they are satisfied \rightarrow use t -distribution

$$S.E.(b) = \frac{s_e}{s_x \sqrt{n-1}} = \frac{29.359}{8.574 \sqrt{24}} = \frac{29.359}{42.004} = 0.699$$

$$t = \frac{b}{S.E.(b)} = \frac{-5.700}{0.699} = -8.154 \sim t_{n-2} = t_{23}$$

$-3.485 > t = -8.154$
 $0.999 < p$ -value

There is weak evidence against H_0 . \rightarrow Do not reject H_0 .

There may not be a positive linear relationship between the two variables.

Use the table below to help you answer some parts of Question 3.

	SS	df	MS	F	p-value
Model	57307.72	1	57307.72	66.487	<0.0001
Error	19824.59	23	861.9387		
Total	77132.311	24			

C) What is the estimate of the standard deviation of the errors? (2 marks)

$$\sigma \approx s_e = \sqrt{s_e^2} = \sqrt{861.9387} = 29.359$$

D) Calculate a 90% confidence interval for the intercept. (4 marks)

$$a = \bar{y} - b\bar{x} = 628.207 - (-5.700)(62.162) = 982.500 \quad t_{n-2, \alpha/2} = t_{23, 0.025} = 1.714$$

$$\frac{s_e}{\sqrt{n}} = \frac{29.359}{\sqrt{25}} = 5.872$$

$$\hat{y} \pm t_{\alpha/2} (s_e / \sqrt{n}) = 982.500 \pm (1.714)(5.872) = 982.500 \pm 10.064 = (972.436, 992.564)$$

E) What is the sum of squares for predicting y through x ? (1 mark)

$$SSE = 19824.59$$

F) What is the coefficient of determination? (1 mark)

$$r^2 = (-0.862)^2 = 0.743 \text{ or } 74.3\%$$

G) Calculate a 99.8% confidence interval for the slope. (4 marks)

$$t_{n-2, \alpha/2} = t_{23, 0.001} = 3.485$$

$$b \pm t_{\alpha/2} s_b = -5.700 \pm (3.485)(0.699) = -5.700 \pm 2.436 = (-8.135, -3.264)$$

Question 4 (10 marks):

The following phrase comes from a famous school:

Draco Dormiens Nunquam Titillandus

A random sample of 97 people indicated that 19 people knew where the phrase comes from. A second random sample (independent from the first) indicated that 400 of 500 people knew where the phrase comes from when it was translated.

A) Is the proportion of people who know where the phrase comes from greater than 18%? Carry out a full analysis in detail. (6 marks)

$$H_0: p \leq 0.18 \quad H_a: p > 0.18$$

Assumptions: $np_0 = 97(0.18) = 17.46 > 15$ $n(1 - p_0) = 97(1 - 0.18) = 79.54 > 15$
and random sample \rightarrow use z-distribution

$$z = \frac{\hat{p} - p_0}{\sqrt{\frac{p_0(1 - p_0)}{n}}} = \frac{(19/97) - 0.18}{\sqrt{\frac{0.18(1 - 0.18)}{97}}} = \frac{0.0159}{0.0390} = 0.41 \sim z$$

$$p\text{-value} = P(Z > 0.41) = 1 - P(Z < 0.41) = 1 - 0.6591 = 0.3409$$

$p\text{-value} > 0.1$

\rightarrow weak evidence against H_0 .

\rightarrow do not reject H_0 .

\rightarrow the proportion of people who know where the phrase comes from could be less than or equal to 18%.

B) Calculate an 80% confidence interval for the population proportion of people who do NOT know where the phrase comes from when it is translated. (4 marks)

$$\begin{aligned} \hat{p} \pm z_{\alpha/2} \sqrt{\frac{\hat{p}(1 - \hat{p})}{n}} &\quad \rightarrow \quad \frac{100}{500} \pm (1.282) \sqrt{\frac{(0.2)(1 - 0.2)}{500}} \\ &\quad \rightarrow \quad 0.200 \pm (1.282)(0.0179) \\ &\quad \rightarrow \quad 0.200 \pm 0.0229 \\ &\quad \rightarrow \quad (0.177, 0.223) \end{aligned}$$

Question 5 (10 marks):

A YouTube-obsessed AND musically-inclined doctor from Chicago has invented a new kind of treatment for patients needing more exercise. Accordingly, department stores have noticed an increased demand for multiple treadmills in the Illinois area. To make sure it's not a coincidence, a marketing expert measures the monthly income of the fitness department of a large store in Illinois. In the same month, the expert also measures the monthly income of a fitness department of a large store in Iowa. The expert repeats this measuring process for 39 more months and obtains the following summary statistics (units are in thousands of \$US) for all months.

Summary statistic	Illinois	Iowa	Difference
Average	141.41	99.23	42.18
Standard Deviation	39.8	81.9	88.8

NOTE: Please note that the third column summarize the differences from the original observations. By choosing a test, you will be using certain columns of the above table, not all of them.

Based on statistical evidence, is the state of Illinois selling more treadmills? Carry out a full analysis in detail.

$$H_0: \mu_{ILL} - \mu_{IOWA} \leq 0 \qquad H_a: \mu_{ILL} - \mu_{IOWA} > 0$$

Assumptions: **independent** & random samples, n_1 & $n_2 = 40 > 30$, normality not mentioned, $81.9/39.8 > 2 \rightarrow$ two independent sample t -test with unequal variance

$$se = \sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}} = \sqrt{\frac{39.8^2}{40} + \frac{81.9^2}{40}} = 14.398$$

$$t = \frac{\bar{x}_1 - \bar{x}_2 - (\mu_1 - \mu_2)}{se} = \frac{141.41 - 99.23 - 0}{14.398} = \frac{42.18}{14.398} = 2.930 \sim t_{df} = t_{39} \approx t_{40} \text{ (OR } t_{30})$$

$$\begin{aligned} df &= 30 \\ 2.750 &> t = 2.930 > 3.385 \\ 0.005 &> p\text{-value} > 0.001 \end{aligned}$$

$$\begin{aligned} df &= 40 \\ 2.704 &< t = 2.623 < 3.307 \\ 0.005 &> p\text{-value} > 0.001 \end{aligned}$$

There is strong to convincing evidence against $H_0 \rightarrow$ Reject H_0 .

The state of Illinois is selling more treadmills. (And maybe dance videos.)