Statistics 571 Midterm 1

Hanlon/Larget, Fall 2011

Name: ____

Instructions:

1. You may use a calculator, but you may not use a laptop computer or phone.
2. The examination is open book, open notes, but not open neighbor. You may use any course handouts including lecture notes and homework solutions.
3. Do all of your work in the space provided. Use the backs of pages if necessary, indicating clearly that you have done so (so the grader can easily find your complete answer).

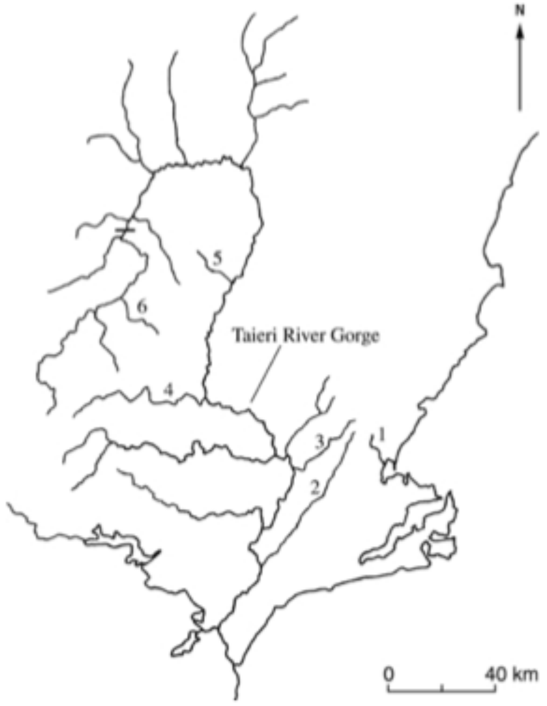
For Graders' Use:

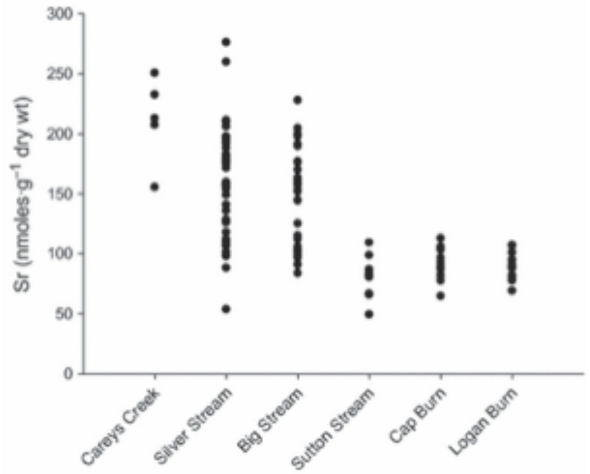
Question	Possible Score	Score
1	15	
2	25	
3	20	
4	25	
5	15	
Total	100	

1. **(15 points) Read the questions on the following page before reading this background material to understand better what information you need to respond to the questions.**

Researchers interested in brown trout from the Taieri River catchment in New Zealand collected brown trout fish eggs from redds (hollows in river beds scooped by fish to spawn) at multiple locations along six tributaries. The tributaries were selected because they spanned different regions of the river system and had relatively large brown trout populations. Some female brown trout live in residence in the freshwater river system for their entire lives. Others are *anadromous*, which means that they live primarily in the ocean, but come into river systems to breed. A primary question of the research was to determine if the strontium concentration in eggs could be used to determine the fraction of spawn in a given tributary that are from anadromous fish. The map on the left shows the Taieri catchment and collection sites: 1. Careys Creek; 2. Silverstream; 3. Big Stream; 4. Sutton Stream; 5. Cap Burn; and 6. Logan Burn. The first three sites, are accessible to the sea by trout without any barriers and are assumed to be populated by a mix of freshwater-resident and anadromous fish. Site 6, Logan Burn, is separated by the Paerau weir (solid bar on the map), which is thought to be impassable for trout, from lower portions of the river system; thus, Logan Burn is thought to be populated exclusively by freshwater-resident fish. Sites 4 and 5, Sutton Stream and Cap Burn, are located below the Paerau weir, but above the Taieri River Gorge, a natural barrier about 20 km long that consists of many small rapids and waterfalls.

A total of 107 redds were sampled from the six tributaries over a three month period in 2003. Researchers collected eggs from redds from locations spread along the full length of each tributary and took care using GPS not to sample the same redd twice at different times. Redds thought to have eggs from more than one fish (identifiable by differences in size and or color) were excluded. From each redd, researchers collected from 30–40 eggs and subsequently measured the strontium (Sr) concentration (nmole per gram dried egg mass) per redd. The graph on the right displays this data.





(a) The researchers treat the sampled redds from each tributary as a random sample from the population of brown trout redds in the tributary during the study period. Is this a reasonable statistical assumption? **Explain with one sentence.**

(b) Researchers predict that fish eggs collected from the stream that is not one of the sample sites, but empties into the Taieri River below the Taieri River Gorge between the mouths of the Silverstream and Big Stream (sites 2–3) will have a mean strontium level greater than 125 nmole per gram. Is this inference better justified by random sampling of tributaries or background scientific knowledge? **Explain with one sentence.**

(c) A statistical model based on the data from the right graph uses tributary to predict strontium concentration. For each of these two variables, circle each appropriate classification.

strontium concentration	EXPLANATORY	CATEGORICAL	EXPERIMENTAL
	RESPONSE	QUANTITATIVE	OBSERVATIONAL
tributary	EXPLANATORY	CATEGORICAL	EXPERIMENTAL
	RESPONSE	QUANTITATIVE	OBSERVATIONAL

2. **(25 points)** The box shown below contains eight mice, each with a different phenotype. For each mouse, we record the number of spots on its nose and its color, either dark (D) or light (L). For example, a mouse labeled (1L) has one spot and is light colored. We draw a mouse at random from the box.

(1L) (2L) (2L) (4L) (4D) (2D) (3D) (3D)

Define the following events:

$$\begin{aligned} N_2 &= \{\text{the mouse has 2 spots}\} \\ N_4 &= \{\text{the mouse has 4 spots}\} \\ D &= \{\text{the mouse is dark}\} \end{aligned}$$

(a) Find $P(N_2 \bigcup D)$.

(b) Find $P(N_2 \bigcap D)$.

(c) Find $P(N_2 \,|\, D)$.

(d) Consider each pair of events (from the three defined above). Which pairs of events are independent, if any? Justify your answer.

(e) Consider each pair of events (from the three defined above). Which pairs of events are mutually exclusive, if any? Justify your answer.

3. **(20 points)** Lehman lovegrass (*Eragrostis lehmanniana Nees*) seeds have an advertised 96% chance of germinating when planted under ideal conditions.

   (a) Assume that the advertised claim is correct. Find the probability that exactly 19 of 20 seeds planted in ideal conditions germinate. Provide an expression equal to this probability and evaluate it as a decimal, accurate to 4 decimal places.

   (b) In a larger study, 466 seeds germinate out of 500 planted in ideal conditions. On the basis of this data, researchers conducted a test of the null hypothesis test that the germination probability is 0.96 versus the alternative that it is lower. Circle the expression below that is equal to the p-value for this hypothesis test.

   (1) $\binom{500}{466}(0.96)^{466}(0.04)^{34}$ (2) $\sum_{k=0}^{466}\binom{500}{k}\left(\frac{466}{500}\right)^k\left(\frac{34}{500}\right)^{500-k}$

   (3) $\sum_{k=0}^{466}\binom{500}{k}(0.96)^k(0.04)^{500-k}$ (4) $\sum_{k=466}^{500}\binom{500}{k}(0.96)^k(0.04)^{500-k}$

   (5) $\mathsf{P}(X \le 466) + \mathsf{P}(X \ge 494)$ where $X \sim \text{Binomial}(500, 0.96)$

   (c) Experiments show that seeds that have passed through the digestive tract of a sheep have a 36% chance of germinating, demonstrating that sheep can disperse the plant. In a new experiment, researchers take 20 regular seeds (assume a 96% germination probability) and 30 seeds extracted from sheep manure (assume a 36% germination probability), plant them, and count the total number that germinate, which we will denote X. Does X have a binomial distribution? If yes, state n and p. If no, briefly explain why not.

   (d) Let $\hat{p}$ be the sample proportion of seeds from the previous experiment $\hat{p} = X/50$. Find the expected value $\mathsf{E}(\hat{p})$.

4. **(25 points)** Marine biologists have noticed that the color of the outermost growth band on a clam tends to be related to the time of year in which the clam dies. A biologist conducted an investigation of whether this is true for the species *Protothaca staminea*. She collected a sample of clam shells from this species and cross-classified them according to (1) month when the claim died and (2) color of the outermost growth band (clear or dark). The data are shown in the following table.

	February	March	April	Total
Clear	8	15	21	44
Dark	17	17	8	42
Total	25	32	29	86

(a) Compute a point estimate and confidence interval for the difference in the proportion of clams having a clear outer growth band between those that died in April and those that died in February. Interpret the confidence interval in the context of the study.

(b) Write down the null and alternative hypothesis for testing the independence of the two categorical variables in this study.

(c) Under the null hypothesis, compute the table of expected counts.

(d) Compute the test statistic X^2 for the χ^2 test of independence.

(e) Compute the test statistic G for the G-test.

(f) The p-value for the G-test is < 0.0001. Interpret the results in the context of this study.

5. **(15 points)** A certain human genetic disease is controlled by a single gene with alleles D and d. When an individual has genotype dd, the person will get the disease when they reach their mid 40s, although there are no symptoms before that time. People with genotypes DD or Dd will not get the disease; however, those with genotype Dd are called *carriers*, because they may pass the disease allele on to their children.

A 30-year-old woman W is the child of two carriers, and by genetic inheritance probabilities, has this probability distribution for her possible genotypes: 0.25 for DD, 0.50 for Dd, and 0.25 for dd. Her husband M is the child of one parent with the disease and one parent with no family history of the disease and is presumed to be a carrier with genotype Dd. This couple has two children, A and B, that are not identical twins and have independent genotypes. The probability distribution of the genotype of a single child from three possible crosses is shown in the following table.

	Child Genotype		
Parent Genotypes	DD	Dd	dd
$Dd \times DD$	0.50	0.50	0
$Dd \times Dd$	0.25	0.50	0.25
$Dd \times dd$	0	0.50	0.50

(a) Given that W is a carrier, what is the probability that both A and B have the disease genotype dd?

(b) What is the probability that both A and B have the disease genotype dd?

(c) Tragically, W dies in her mid 30s in an accident. As a consequence, she does not live into her 40s and it is not known if she would have gotten the disease or not. Given that both A and B exhibit the disease in their 40s, what is the probability that W was a carrier?