Solution to Statistics 571 Midterm 1

Hanlon/Larget, Fall 2011

1.

(a)

Solution: Yes. While the researchers do not use a formal random sampling procedure, the researchers make efforts to ensure that the redds they sample are representative of the redds in each tributary by sampling redds from the entire length of each stream.

Notes:

i. The problem defines six populations; the female brown trout that spawn in each of six tributaries of the Taieri River catchment in New Zealand during the study period. These six populations are the ones from which the researchers wish to sample. They want to make comparisons among the strontium levels measured in eggs sampled from these streams.

ii. The researchers are not attempting to sample from brown trout populations elsewhere in the Taieri River catchment, elsewhere in the world, or at different times.

iii. The researchers are interested in the strontium concentration in eggs laid by individual female brown trout. Measurements from biological samples that are a mixture of eggs from more than one trout would not be accurate measurements from single fish, and are properly discarded.

iv. It is legitimate to question if there is a difference in the strontium concentration between eggs from sampled redds and eggs from redds that the researchers did not find, but there is no obvious reason from the sampling description that there would be a bias.

(b)

Solution: This inference is best explained by background scientific knowledge because the researchers do not sample any redds from the stream with mouth between the Silverstream and the Big Stream, but they do expect strontium levels in eggs to be similar to those from nearby streams for which trout have similar access to the ocean.

Notes:

i. The tributaries are not randomly sampled.

ii. There is no random sampling justification for arguing that tributaries that were not sampled should be like tributaries that were sampled with regard to strontium concentrations in eggs.

(c)

Solution: Strontium concentration is the response variable of the stated model, is quantitative, and is observational.

Tributary is an explanatory variable in the stated model, is categorical, and is observational.

Notes:

i. Even though the researchers design their study and select from which tributaries to sample, the main point is that for each sampling unit, in this case a biological sample of eggs, the researchers observe from which tributary the eggs came.

ii. The location of the eggs is an intrinsic property of the eggs, not something that the researchers controlled.

iii. Nature decides which eggs are in which tributary, not the researchers; the researchers control what portion of nature they choose to observe.

2.

Solution:

(a) Six of the eight mice have two spots, are dark, or both. $P(N_2 \cup D) = \frac{6}{8}$.

(b) Only one of the six mice has two spots and is dark. $P(N_2 \cap D) = \frac{1}{8}$.

(c) Of the four dark mice, only one has two spots. $P(N_2 \mid D) = \frac{1}{4}$.

(d) N_4 and D are independent because the proportion of mice with four spots is the same among both dark and light mice.
More formally, $\mathsf{P}(N_4 \cap D) = \frac{1}{8}$. Also, $\mathsf{P}(N_4) = \frac{2}{8}$, $\mathsf{P}(D) = \frac{4}{8}$, and $\mathsf{P}(N_4) \times \mathsf{P}(D) = \frac{2}{8} \times \frac{4}{8} = \frac{1}{8}$. As $\mathsf{P}(N_4 \cap D) = \mathsf{P}(N_4) \times \mathsf{P}(D)$, N_4 and D are independent.

(e) Events N_2 and N_4 are mutually exclusive because no mouse has both 2 and 4 spots.
$\mathsf{P}(N_2 \cap N_4) = 0$.

3.

Solution:

(a) The number of seeds that germinate is Binomial$(20, 0.96)$.

$$\binom{20}{19}(0.96)^{19}(0.04)^1 \doteq 0.3683$$

(b) Under the null hypothesis, define X to be the number of seeds that germinate. It follows that $X \sim$ Binomial$(500, 0.96)$. The p-value for the alternative hypothesis $H_a\colon p < 0.96$ equals $\mathsf{P}(X \leq 466)$ which is (3) $\sum_{k=0}^{466} \binom{500}{k}(0.96)^k(0.04)^{500-k}$.

(c) X is not binomial because p is not the same for all trials.

(d) LetX_1 be the number of seeds that germinate among the 20 with $p = 0.96$ and let X_2 be the number that germinate among the 30 with $p = 0.36$. Then $X = X_1 + X_2$. In addition, $X_1 \sim$ Binomial$(20, 0.96)$ so that $\mathsf{E}(X_1) = 20(0.96) = 19.2$ and $X_2 \sim$ Binomial$(30, 0.36)$ so that $\mathsf{E}(X_2) = 30(0.36) = 10.8$. By the linearity of expectation, $\mathsf{E}(X) = \mathsf{E}(X_1) + \mathsf{E}(X_2) = 30$ and it follows that $\mathsf{E}(X/50) = 30/50 = 0.6$.

4.

Solution:

(a) $\hat{p}_1 = \frac{21}{29} \doteq 0.724$. $\hat{p}_2 = \frac{8}{25} \doteq 0.32$. The point estimate is $\hat{p}_1 - \hat{p}_2 = 0.404$.
The standard error is

$$\sqrt{\frac{(0.724)(0.276)}{29} + \frac{(0.32)(0.68)}{25}} \doteq 0.125$$

A 95% confidence interval is

$$0.404 - 1.96(0.125) < p_1 - p_2 < 0.404 + 1.96(0.125)$$

which simplifies to

$$0.159 < p_1 - p_2 < 0.649$$

We are 95% confident that the proportion of dead clams in April with clear outermost growth bands is between 0.159 and 0.649 larger than the proportion of dead clams in February with clear outermost growth bands.

(b)

H_0: the color of the outermost growth bands and the month of death are independent
H_a: the color of the outermost growth bands and the month of death are not independent

(c)

	February	March	April	Total
Clear	12.8	16.4	14.8	44
Dark	12.2	15.6	14.2	42
Total	25	32	29	86

(d)

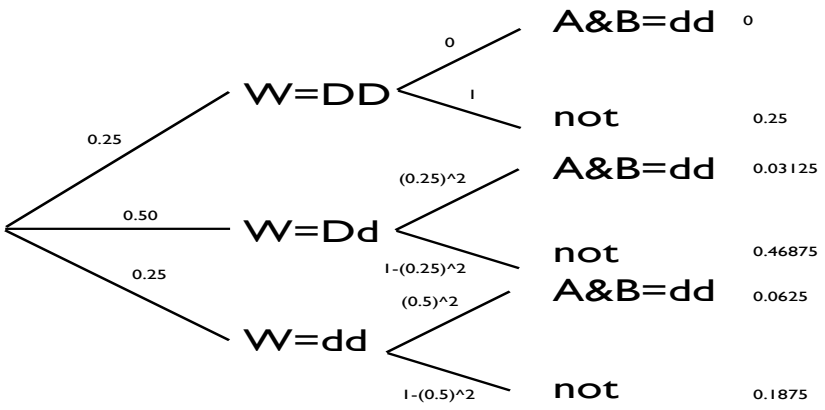
$$X^2 = \frac{(8-12.8)^2}{12.8} + \cdots + \frac{(8-14.2)^2}{14.2} \doteq 9.15$$

(e)

$$G = 2\left(8\ln\left(\frac{8}{12.8}\right) + \cdots + 8\ln\left(\frac{8}{14.2}\right)\right) \doteq -9.43$$

(f) There is very strong evidence ($p < 0.0001$, $G = -9.43$, G-test of independence, 2 df) that the color of the outermost growth band is not independent of the month of death for these clams.

5.



Solution:

(a) $\mathsf{P}(A = dd \cap B = dd \,|\, W = Dd) = \mathsf{P}(A = dd \,|\, W = Dd) \times \mathsf{P}(B = dd \,|\, W = Dd) = (0.25)^2 = \frac{1}{16} = 0.0625.$

(b) Use the law of total probability (add probability of paths in the tree).

$$\begin{aligned}
\mathsf{P}(A = dd \cap B = dd) &= \mathsf{P}(A = dd \cap B = dd \,|\, W = DD)\mathsf{P}(W = DD) \\
&\quad + \mathsf{P}(A = dd \cap B = dd \,|\, W = Dd)\mathsf{P}(W = Dd) \\
&\quad + \mathsf{P}(A = dd \cap B = dd \,|\, W = dd)\mathsf{P}(W = dd) \\
&= (0)(0.25) + (0.25)^2(0.5) + (0.5)^2(0.25) \\
&= 0.09375 = \frac{3}{32}
\end{aligned}$$

(c) Bayes' Theorem.

$$\begin{aligned}
\mathsf{P}(W = Dd \,|\, A = dd \cap B = dd) &= \frac{\mathsf{P}(W = Dd \cap (A = dd \cap B = dd))}{\mathsf{P}(A = dd \cap B = dd)} \\
&= \frac{(0.5)(0.25)^2}{0.09375} \\
&= \frac{1/32}{3/32} \\
&= \frac{1}{3} \doteq 0.3333
\end{aligned}$$

Notes:

i. The genotypes of A and B are conditionally independent given the genotypes of their parents.

ii. If the two children did not have the same mother, then finding the probability that A had genotype dd and squaring would have been correct for finding the probability that A and B both had genotype dd, but they are full siblings.