Statistics 571 Final Examination

Hanlon/Larget, Fall 2010

Name: _____

Please circle the lecture section *in which you are registered*: Hanlon Larget

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	40	
2	30	
3	40	
4	40	
5	50	
Total	200	

1. (40 points) In a randomized, double-blind, experiment the drug tamoxifen was given to 6681 women and a placebo was given to 6707 other women. After four years there were 89 cases of breast cancer in the tamoxifen group, compared with 175 in the placebo group. The following table summarizes the results.

	Placebo	Tamoxifen	Total
Cancer	175	89	264
No Cancer	6532	6592	13124
Total	6707	6681	13388

(a) Compute the relative risk of cancer in the placebo group versus the tamoxifen group. Interpret the result in the context of the problem.

(b) Compute a 95% confidence interval for the odds ratio of cancer in the placebo group versus the tamoxifen group. Interpret the result in the context of the problem.

(c) Conduct a χ^2-test to test for independence between treatment (placebo or tamoxifen) and the development of cancer. State the value of the test statistic and indicate its approximate sampling distribution including degrees of freedom. Interpret the results in the context of the problem.

2. (30 points, 3 points per part) For each statement, circle TRUE or FALSE. If FALSE, explain why or make a small change to correct it. Very brief explanations are sufficient.

(a) Circle either TRUE or FALSE (and explain/correct if FALSE):

Statistical inference about a population of interest based on a sample can only be justified when the sample is selected using a formal random selection procedure from a complete list of individuals in the population of interest.

(b) Circle either TRUE or FALSE (and explain/correct if FALSE):

In a random sample of size 100 from a population, the sample proportion of individuals with a specific phenotype is 0.14. A null genetic model predicts that the population proportion of individuals with this phenotype is $H_0: p = 0.25$. The p-value for the test with alternative hypothesis $H_A: p < 0.25$ is correctly calculated as $\binom{100}{14}(0.25)^{14}(0.75)^{86}$.

(c) Circle either TRUE or FALSE (and explain/correct if FALSE):

When choosing between a paired t-test and a two independent sample t-test, one should choose the test that results in the smaller p-value.

(d) Circle either TRUE or FALSE (and explain/correct if FALSE):

When a p-value equals 0.03, there is 3 percent chance that the null hypothesis is correct.

(e) Circle either TRUE or FALSE (and explain/correct if FALSE):

In a random sample from a single population, a 95% confidence interval for μ, the population mean of a quantitative variable, is $48 < \mu < 63$. This implies that about 95% of all individuals in the population have values of the quantitative variable between 48 and 63.

(f) Circle either TRUE or FALSE (and explain/correct if FALSE):

The purpose of including blocking in the design of an experiment is to help control for the effects of some potentially confounding variables.

(g) Circle either TRUE or FALSE (and explain/correct if FALSE):

A statistician is selecting an ANOVA model for a quantitative response variable Y and two categorical explanatory variables A and B, each of which has two levels. If the lines of an interaction plot cross, it follows that the interaction term in a two-way ANOVA model with an interaction will be statistically significant (have a p-value less than 0.05).

(h) Circle either TRUE or FALSE (and explain/correct if FALSE):

In a simple linear regression model, the equation of the fitted line is $\hat{Y} = 2.5 - 1.3X$. This implies that the correlation coefficient between X and Y is negative.

(i) Circle either TRUE or FALSE (and explain/correct if FALSE):

A student uses paper and pencil and the textbook equations to calculate the correlation coefficient between two quantitative variables X and Y and finds $r = 1.06$. This implies a very strong linear relationship between X and Y.

(j) Circle either TRUE or FALSE (and explain/correct if FALSE):

After fitting a simple linear regression model, a plot of residuals versus values of the explanatory variable exhibit an up/down/up pattern from left to right. This is evidence of nonconstant variance.

3. (40 points, 8 points for each part) For each part: (1) state the statistical method studied this semester most suitable for analyzing the data from one of these options: (*inference for proportions; contingency table analysis; paired t-test; independent sample t-test; ANOVA; or regression analysis*); (2) note two important assumptions for the resulting data analysis to be valid; and (3) name the most appropriate probability distribution upon which inference (such as p-values or the widths of confidence intervals) would depend (*for example, binomial with* $n = 12$ *and* $p = 0.5$*, normal, t with 20 degrees of freedom, F with 2 and 30 degrees of freedom, or chi-square with 4 degrees of freedom*). **Do not actually conduct the analysis.**

   (a) Researchers conducted a randomized, double-blind, clinical trial to compare treatments in which some patients with schizophrenia were given the drug clozapine and others were given haloperidol. After one year 61 of 163 patients in the clozapine group showed clinically important improvement in the symptoms, compared with 51 of out 159 in the haloperidol group.

   (b) A biologist was interested in the relationship between the velocity at which beluga whale swims and the tail-beat frequency of the whale. A sample of 19 whales was studied and measurements were made on swimming velocity, measured in units of body-lengths of the whale per second and tail-beat frequency, measured in units of hertz.

(c) A researcher captured male damselflies and randomly assigned them to one of three groups. For those in the first group the sizes of red spots on the wing were artificially enlarged with red ink. For those in the second group the wing spots were enlarged with clear ink. The third group served as a control. The damselflies were then released into a contaminated area. The numbers surviving in each of the three groups 22 days later were determined. There were 312 damselflies in each of the three groups. After 22 days, there were 41 survivors in the "red ink" group, 49 survivors in the "clear ink" group, and 57 survivors in the control group.

(d) Researchers took eight soil samples at each of six locations in Mediterranean pastures. They divided the samples into four pairs and put the soil in pots. One pot from each pair was watered continuously, while the other pot was watered for 13 days, then not watered for 18 days, and then watered again for 30 days. The researchers recorded the number of germinations in each pot during the experiment.

(e) Heat shock proteins (HSPs) are a type of protein produced by some organisms as protection against damage from exposure to high temperature. In the fruit fly *Drosophila melanogaster* the genes that encode HSPs are found on chromosomes that uncoil and appear to puff out. This chromosome puffing can be seen under a microscope. A biologist counted the number of puffs per chromosomal arm from the salivary glands of 40 Drosophila larvae that had been heat shocked at 37°C for 30 minutes, 40 larvae that had been heat shocked for 60 minutes, and 40 control larvae.

4. (40 points) A study was conducted to compare the coefficients of digestibility of dry matter for four diets fed to goats. Six randomly selected goats were assigned to each treatment. Data are indicated on the right. Sample means and standard deviations are given below. Also, calculations yield SS(error) = 250.5 and SS(total) = 689.6.

Diet	A	B	C	D
mean	53.5	49.7	60.2	59.2
SD	3.21	3.14	4.79	2.64
n	6	6	6	6

Source	df	SS	MS	F-statistic	P-value
groups					0.0001211
error		250.5			
Total		689.6			

(a) Complete the ANOVA table.

(b) State the conclusion of the test in the context of the problem.

(c) Compute a 99% confidence interval for the mean difference between diet A and diet C.

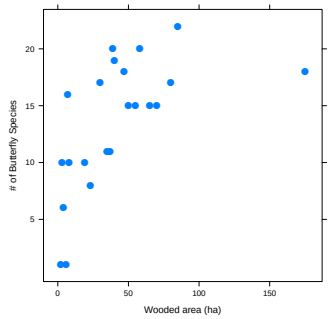
5. (50 points) It is thought by some biologists that increasing fragmentation of woodland regions may result in lower biodiversity of plant and animal species that live mostly in woodland environments. To examine this issue, biologists in England selected 22 woodland sites in eastern England. The sites were purposely chosen to have different characteristics. Some sites were large, some were very small, some had coniferous trees, others deciduous, some were managed, others were unmanaged, and so on. Over a period of time from early May to late October in one calendar year, scientists walked along transects at each site every two weeks and recorded the species of all butterflies that were observed. (More time was spent in larger wooded sites than smaller ones.) A total of 26 species were observed, but the number of species observed at each site ranged from a low of one to a high of 22. The area of each site (in hectares) and the number of different species observed at least once are shown here. The sites are ordered from largest to smallest.

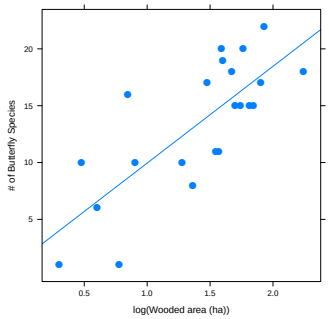
Site	Area (ha)	#Species	Site	Area (ha)	#Species
Markshall Wood	175	18	Weeley Hall Wood	37	11
Lineage Wood	85	22	Dodnash Wood	35	11
Shardlowes Wood	80	17	Arger Fen	30	17
Hintlesham Great Wood	70	15	Bulls Cross Wood	23	8
Stour Wood	65	15	Groton Wood	19	10
Assington Thicks	58	20	Hazel Wood	8	10
Pods Wood	55	15	Wrights Wood	7	16
Layer Wood	50	15	Long Wood	6	1
Bentley Hall Wood	47	18	Walding Wood	4	6
Wolves Wood	40	19	Corner Place	3	10
Coperass Wood	39	20	Stattles Wood	2	1

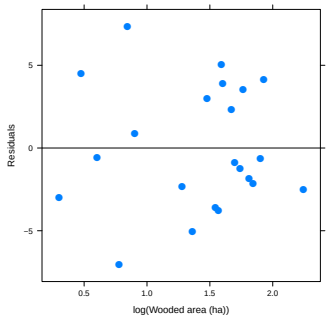
A partial summary of this fitted regression model from R is shown below.

```
Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)     1.438      2.383   0.603    0.553
log10(area)     8.513      1.592   5.349  3.1e-05 ***
---
Residual standard error: 3.836 on 20 degrees of freedom
Multiple R-squared: 0.5885,Adjusted R-squared: 0.568
```

The left scatter plot shows the untransformed variables. The middle scatter plot shows the number of species versus the base 10 logarithm of the area and a fitted regression line. The third plot displays residuals from this regression model versus the base 10 logarithm of area.







(a) Based on examination of the graphs, briefly explain why the choice was made to log transform the explanatory variable area prior to fitting a linear regression model.

(b) According to the regression model, how large of an increase in the number of butterfly species is predicted per increase in the area of a site by a multiplicative factor of 10? Repeat for an increase by a multiplicative factor of 2.

(c) Use the fitted regression model to predict the number of butterfly species in a site with a wooded area of 460 ha selected from the same population of wooded sites in eastern England.

(d) On the basis of the fitted regression model, a 95% prediction interval for the number of butterfly species at a single site with 460 ha wooded area is from 14.9 to 33.3 species. The biologists conclude their abstract with this comment.

> It was predicted that a woodland area of about 460 ha would be required to support all the species recorded in the study.

Briefly discuss *two or more distinct statistical reasons* to question the validity of this confidence interval and the conclusion made by the biologists.

(e) Inference from simple linear regression depends on several assumptions. Use available information to comment on how well the data fits the assumptions of *normality*, *linearity*, *constant variance*, and *independence*.