STAT 659 — Exam 1 **Name:** _____
Spring 2015

INSTRUCTIONS FOR THE STUDENT:

1. You have exactly 70 minutes to complete the exam.
2. There are 11 pages including this cover sheet and 4 pages of SAS output.
3. Each lettered part of a question is worth 8 points unless otherwise marked.
4. Please answer all questions.
5. Show all your work on the test booklet.
6. Do not discuss or provide any information to any one concerning any of the questions on this exam or your solutions until I post the solutions.
7. You may use a calculator that does not have the capability of phoning, texting, or accessing the internet and one $8\frac{1}{2} \times 11$ formula sheet (you may use both sides). Do not use the textbook or class notes.
8. Carry out tests at level 0.05 unless otherwise stated.
9. Be sure to clearly state the hypotheses, the test statistic and its value, and conclusion for all tests.

I attest that I spent no more than 70 minutes to complete the exam. I used only the materials described above. I did not receive assistance from anyone during the taking of this exam.

Student's Signature_____

INSTRUCTIONS FOR PROCTOR:

(1) Record the time at which the student starts the exam: _____

(2) Record the time at which the student ends the exam: _____

(3) Immediately after the student completes the exam, please scan the exam to a .pdf file and have the student upload it to webassign.

(4) Collect all portions of this exam at its conclusion. Do not allow them to take any portion with them.

(5) Please keep these materials until March 10, at which time you may either dispose of them or return them to the student.

I attest that the student has followed all the INSTRUCTIONS FOR THE STUDENT listed above and that the exam was scanned into a pdf and uploaded to webassign in my presence:

Proctor's Signature_____

Some Chi-Squared Percentiles

	Right-Tail Probability			
df	0.100	0.050	0.025	0.010
1	2.71	3.84	5.02	6.63
2	4.61	5.99	7.38	9.21
3	6.25	7.81	9.35	11.34
4	7.78	9.49	11.14	13.28
5	9.24	11.07	12.83	15.09
6	10.64	12.59	14.45	16.81
7	12.02	14.07	16.01	18.48
8	13.36	15.51	17.53	20.09
9	14.68	16.92	19.02	21.67
10	15.99	18.31	20.48	23.21

Some Normal Percentiles

Right-Tail Probability			
0.100	0.050	0.025	0.010
1.282	1.645	1.960	2.326

1. A study was carried out to investigate the relationship between regular smoking and major depressive disorder. A sample of 3213 individuals were classified according to regular smoking habit (`smoke`) and major depressive disorder (`depress`). Since this relationship may depend on gender, the tables were stratified according to gender (`gender`). Use the accompanying SAS output to help you answer this problem.

   (a) Is Simpson's paradox present for these data? Explain why or why not.

   (b) Determine the values of the test statistics, the $P-$values, and the conclusions for (i) the test of equal odds ratios for between `smoke` and `depress` for males and females and (ii) the test of partial association of `smoke` and `depress`, controlling for gender.

	(i) the test of equal odds ratios	(ii) the test of partial association
Value of test statistic		
$P-$value		
Conclusion		

(c) Report a 95% confidence interval for a common odds ratio between `smoke` and `depress`. Comment on whether it is appropriate to use this interval to summarize these tables.

2. A study was carried out in 5 countries of a possible relationship between age of a woman when her first child is born and the onset breast cancer. Certain researchers believe that the risk of breast cancer increases with age at the birth of the first child. Breast-cancer cases were selected among women in selected hospitals in 5 countries. An independently chosen sample of women without breast cancer was selected among women who did not have breast cancer and who were in the hospital at the same time. Each woman was asked about her age at first birth. The following data were obtained for those women with at least one birth:

	Age at First Birth					Sample
	<20	20–24	25–29	30–34	≥ 35	Size
Cancer Patients	320	1206	1011	463	220	3220
Controls	1422	4432	3904	1555	626	11939

(a) Construct a 95% confidence interval for the proportion of female breast-cancer patients that have their first child at age 30 or later.

	Age at First Birth					Sample
	<20	20–24	25–29	30–34	≥35	Size
Cancer Patients	320	1206	1011	463	220	3220
Controls	1422	4432	3904	1555	626	11939

(b) Construct a 95% confidence interval for the odds ratio for having breast cancer for women in the oldest age group (≥35) relative to those in the youngest age group (<20). Interpret this interval.

(c) Carry out a test of independence between age at first birth and type of patient (cancer or control). (We note that this test can be used to determine whether the proportions of patients in the different age groups differed between the cancer patients and the control patients.) The frequency procedure in SAS produced the following output:

```
Statistics for Table of group by age

Statistic                     DF       Value      Prob

Chi-Square                     4     24.9562    <.0001
Likelihood Ratio Chi-Square    4     24.6488    <.0001
Mantel-Haenszel Chi-Square     1     15.8685    <.0001
Phi Coefficient                      0.0406
Contingency Coefficient              0.0405
Cramer's V                           0.0406

Sample Size = 15159
```

(d) The standardized residuals were obtained for the test of independence in part (c). Interpret these residuals.

	Age at First Birth				
	<20	20–24	25–29	30–34	≥35
Cancer Patients	-3.11	0.35	-1.40	2.01	3.49
Controls	3.11	-0.35	1.40	-2.01	-3.49

3. A cross between white and yellow summer squash gave progeny of the following colors:

Color	White	Yellow	Green
Number of Progeny	143	46	19

Carry out a test of whether these data are consistent with the $12:3:1$ ratio predicted by a certain genetic model (i.e., the probability of white progeny is $12/(12+3+1) = 0.75$, etc.).

4. Researchers studied the incidence of lower respiratory infections in 284 children over a year. Explanatory variables included passive smoking (`passive` =1 if yes), socioeconomic status (`ses`, 3 categories), crowding (`crowding` =1 if yes), race (`race`), exposure time (`risk`), and age (`agegroup`, three categories). The response variable was the number times that the child had a lower respiratory infection during the year.

   (a) A model (Model A) with the predictors `passive, crowding, ses, race, agegroup, risk` was fit to the data. The researchers felt that the effects of `risk` and `ses` might not be useful in the model, so they fit a second model (Model B) omitting these two variables from the model. Carry out a likelihood ratio test to determine whether it is appropriate to simultaneously omit the two effects from Model A.

   (b) Use the Poisson loglinear model (Model A) with all the predictors to answer this part of the problem. What is the estimated effect of simultaneously being in a crowded setting (`crowding`= 1) and being exposed to passive smoke (`passive`= 1) on the mean number of lower respiratory infections relative to being in a noncrowded setting (`crowding`= 0) and not being exposed to passive smoke (`passive`= 0), keeping all other variables constant?

(c) Several Poisson regression models with log link were fit to the data. At each step, the least significant predictor was eliminated from the model, and a model with the remaining predictors was fit to the data. Based on information in the table below, select a reasonable Poisson regression model. Explain your reasoning.

Model	Predictors	Deviance	DF	AIC_C
1	`passive,crowding,ses,race,agegroup,risk`	378.7	275	664.3
2	`passive,crowding,race,agegroup,risk`	380.0	277	661.4
3	`passive,crowding,agegroup,risk`	380.9	278	660.2
4	`crowding,agegroup,risk`	386.3	279	663.5
5	`crowding,agegroup`	399.0	280	674.2
6	`crowding`	404.4	282	675.4

(d) A negative binomial model with log link (Model C) was also fit using the same predictors as Model A. Based on the output of this model and also that of Model A, is there any evidence of lack of fit or inadequacy of the Poisson regression Model A? Give a complete explanation.