COLLEGE OF THE HOLY CROSS
Department of Mathematics and Computer Science

STAT 232 Categorical Data Analysis
Spring 2025 Final Exam

Question	**1**	**2**	3	4	5	6	7	8	9	10	11	12	*EC*	Total
Points	10	10	10	10	10	10	10	10	10	10	10	10	10	100

Your name_____

Duration of the Final Exam is 150 minutes. There are 12 problems. ***The first problem and the second problem are mandatory****. From problems 3 – 12, only 8 problems will be graded. If you solve all Problems 3 – 12, you must cross out the two problems in the boxes above that must not be graded. If you solve all Problems 3 – 12 but do not cross out two problems, only the first ten problems from 3 – 12 will be graded. Show all your work for full credit. Books, notes etc. are prohibited. Calculators, cellphones, earphones, AirPods and cheat sheets are NOT permitted.*

1. Let $Y_1, Y_2, \ldots, Y_n$ denote a random sample from a Bernoulli distribution where $P(Y_i = 1) = p$ and $P(Y_i = 0) = 1 - p$ and assume that the prior distribution for p is Beta(α, β).

   (a) Find the posterior distribution for p.

   (b) Find the Bayes estimators for p and $p(1 - p)$.

2. Table 3.2 (Agresti, pp.76-77) comes from a study of nesting horseshoe crabs. Each female crab in the study had a male crab attached to her in her nest. The study investigated factors that affect whether the female crab had any other males, called satellites, residing nearby her. The response outcome for each female crab is her number of satellites (sat). There are several predictor variables in the data set: color, spine condition, width, and weight. Let $Y = 1$ if a crab has at least one satellite, and let $Y = 0$ otherwise. If we use logistic regression to fit the model for $\pi(x) =$ probability of having a satellite, using weight as the predictor, we get the following R output.

```
> sat=crabsdata$sat
> weight=crabsdata$weight
> yesorno=crabsdata$y
> yesorno
  [1] 1 0 1 0 1 0 0 0 0 0 0 0 1 0 1 1 1 1 0 1 1 1 1 1 1 0 1 1 0 0 1 0 0 1 1 0 0 1
 [39] 0 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 0 0 0 1 1 1 1 1 1 0 0 1 1 1 0 1 0 0
 [77] 1 1 0 1 0 0 0 1 0 1 0 0 0 0 1 0 0 0 0 1 0 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 0 1
[115] 1 0 1 0 1 1 1 1 0 1 1 0 1 1 1 1 1 1 0 1 1 1 0 1 1 1 1 1 1 1 1 0 0 1 1 1 1 1 1 0
[153] 0 0 0 1 0 1 0 1 1 1 1 0 1 0 0 1 1 1 0 0 0
> model44<-glm(yesorno~weight,family=binomial(link=logit))
> summary(model44)

Call:
glm(formula = yesorno ~ weight, family = binomial(link = logit))

Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept)  -3.6947     0.8802  -4.198 2.70e-05 ***
weight        1.8151     0.3767   4.819 1.45e-06 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 225.76  on 172  degrees of freedom
Residual deviance: 195.74  on 171  degrees of freedom
AIC: 199.74

Number of Fisher Scoring iterations: 4
```

(a) Write the logit form of the fitted model, $g(x)$.

(b) Find $\hat{\pi}(x)$ at the weight values 1.20, 2.44, and 5.20 kg, which represent the sample min, mean, and max weights.

(c) At what weight is a female crab equally likely to have/not have a satellite?

(d) Construct a 95% confidence interval to describe the effect of weight on the odds of having a satellite. Interpret your result. **Hint:** $z_{0.05} = 1.645$, $z_{0.025} = 1.96$

(e) Conduct a test of the hypothesis that weight has no effect on having a satellite using the likelihood ratio test and G-statistic. Interpret. **Hint:** $\chi^2_{1,0.05} = 3.84$.

3. (a) The National Fire Incident Reporting Service stated that, among residential fires, 73% are in family homes, 20% are in apartments, and 7% are in other types of dwellings. If four residential fires are independently reported on a single day, what is the probability that two are in family homes, one is in an apartment, and one is in another type of dwelling?

(b) Suppose that $E(\hat{\theta}_1) = E(\hat{\theta}_2) = \theta$, $\text{Var}(\hat{\theta}_1) = \sigma_1^2$, and $\text{Var}(\hat{\theta}_2) = \sigma_2^2$. Consider the estimator

$$\hat{\theta}_3 = a\hat{\theta}_1 + (1-a)\hat{\theta}_2.$$

i. Show that $\hat{\theta}_3$ is an unbiased estimator for θ.

ii. If $\hat{\theta}_1$ and $\hat{\theta}_2$ are independent, how should the constant a be chosen in order to minimize the variance of $\hat{\theta}_3$?

4. Information about eight four-cylinder automobiles judged to be among the most fuel efficient in 2006 is given in the following table. Engine sizes are in total cylinder volume, measured in liters (L).

Car	Cylinder Volume (x)	Horsepower (y)
Honda Civic	1.8	51
Toyota Prius	1.5	51
VW Golf	2.0	115
VW Beetle	2.5	150
Toyota Corolla	1.8	126
VW Jetta	2.5	150
Mini Cooper	1.6	118
Toyota Yaris	1.5	106

(a) Find the least-squares line for the data.

(b) Use the least-squares line to estimate the mean horsepower rating for a fuel-efficient automobile with cylinder volume 1.9 L.

5. (a) Suppose that Y is a binomial random variable based on n trials and success probability p. Use the conjugate beta prior with parameters α and β to derive the posterior distribution of $p|y$.

   Hint: You may use your answer to Problem 1.

   (b) Suppose that our prior distribution is Beta(4,16), and we observe 3 heads among 10 coin flips. What is our posterior distribution?

   (c) If we observe heads on the next coin flip, what does our new posterior look like?

   (d) Let $\alpha = 1$ and $\beta = 3$. Find the Bayes estimator for p, and derive its mean and variance.

6. Does a nicotine patch help a person to quit smoking? 150 people have joined a clinical trial that is testing the effectiveness of a nicotine patch on a person's ability to quit smoking. The 150 subjects were divided into two groups, one given the nicotine patch and the other a placebo. After 12 weeks, the number of people who had quit smoking was recorded:

	Still Smoking	Quit Smoking	Total
Patch	58	22	80
No Patch	57	13	70
Total	115	35	150

(a) Construct a 95% confidence interval for the difference in the proportion of people who quit smoking on the nicotine patch compared with a placebo. Does this appear to be a significant difference? Why or why not? **Hint:** $z_{0.05} = 1.645$, $z_{0.025} = 1.96$

(b) At the 5% level of significance, test the claim that the nicotine patch is effective at helping people to quit smoking.

(c) Determine the odds ratio of still smoking while on the nicotine patch. Interpret.

(d) What is the relative risk of smoking if you are on nicotine patch? Interpret.

7. Fine needle aspiration (FNA) is a technique in which a small sample of the tumor is taken using a needle and visually inspected through a microscope. The data below represent 37 FNA slide samples. Slides with smooth ellipsoid-shaped nuclei were classified as "round" and slides with poorly shaped cell nuclei were classified as "concave." A biopsy was also conducted on each of these samples to determine if each was malignant or benign.

	Malignant	Benign	Total
Concave	17	4	21
Round	7	9	16
Total	24	13	37

(a) What are the largest and smallest possible values for n_{11} in this problem?

(b) Let's say that we use a Fisher's Test to find the p-value associated with a test to claim that concave cells are more likely to be malignant. Write down the null hypothesis and alternative hypothesis.

(c) Compute the p-value. You do not have to find the exact value. You may leave your answer in terms of binomial coefficients.

(d) What is the p-value for the test if we instead used the mid-p-value approach

8. A baker is trying to choose between two types of cookies to determine which should be featured in an upcoming marketing ad. The plan is to sample public opinion on the two types of cookies by setting up sampling areas for each flavor in a busy area and asking people at each location if they enjoy the cookie. Suppose that the baker also collected information related to the gender, age, etc. of the person sampling the cookie. Data are given below.

Age	Cookie	Liked	Not Liked
<30	Raspberry Rally	760	140
	Toast-Yay!	600	100
30-49	Raspberry Rally	40	60
	Toast-Yay!	150	150

(a) Calculate the conditional odds ratios for the two flavors, conditioned on age.

(b) Calculate the marginal odds ratio for the two flavors.

(c) Do we have evidence of Simpson's paradox? Why or why not?

9. Suppose that we have two drug treatments, A and B (variable X), and we define a response variable, Y, in terms of success and failure of the treatment. The treatments are taking place at two clinics which we'll label by $Z = 1, 2$.

Clinic (Z)	Treatment (X)	Success	Failure
1	A	18	12
	B	12	8
2	A	2	8
	B	8	32
Overall	A	20	20
	B	20	40

(a) Test the null hypothesis of homogeneity of the odds ratio of treatment A compared to treatment B across different clinics. Be sure to state your hypotheses and interpret your result in the context of the problem. Use $\alpha = 0.05$. **Hint:** $P(\chi^2 > 0) = 1$.

(b) Calculate an estimate of the common odds ratio.

(c) Test the null hypothesis that the outcome is independent of treatment, conditional on clinic. Interpret. Use $\alpha = 0.05$.

10. Hypertension is defined as having: Systolic BP $\geq$ 160 mm Hg, Diastolic BP $\geq$ 95 mm Hg

The hypertensive status of 20 patients was evaluated by an automated device and a trained observer. Note that this data is paired, since each person is having their blood pressure measured twice, once by the machine and once by a person.

Hypertensive status of 20 patients as judged by a computer device and a trained observer

	Hypertensive status			Hypertensive status	
Person	Computer device	Trained observer	Person	Computer device	Trained observer
1	–	–	11	+	–
2	–	–	12	+	–
3	+	–	13	–	–
4	+	+	14	+	–
5	–	–	15	–	+
6	+	–	16	+	–
7	–	–	17	+	–
8	+	+	18	–	–
9	+	+	19	–	–
10	–	–	20	–	–

Test the claim that the human and automated device differ in their ability to detect hypertension. Use $\alpha = 0.05$.

11. **THC versus prochlorperazine.** An article in the New England Journal of Medicine described a study on the effectiveness of medications for combatting nausea in patients undergoing chemotherapy treatments for cancer. In the experiment, 157 patients were divided at random into two groups. One group pf 78 patients were given a standard antinausea drug called prochlorperazine, while the other group of 79 patients received THC (the active ingredient in marijuana). Both medications were delivered orally and no patients were told which of the two drugs they were taking. The response measured was whether or not the patient experienced relief from nausea when undergoing chemotherapy. R output of the fitted binary logistic regression model to predict Effectiveness (Yes or No) using type of Drug as a binary predictor is given below.

```
>
Final Exam 9,78) # row totals
> y= c(36,16) # number of 'effective'
> x= c(1,0) # treatment, 1 if THC, 0 if prochlorperazine
> model<- glm(y/n~x, family=binomial, weights= n)
> summary(model)

Call:
glm(formula = y/n ~ x, family = binomial, weights = n)

Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept)  -1.3545     0.2804  -4.831 1.36e-06 ***
x             1.1769     0.3601   3.268  0.00108 **
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 11.345  on 1  degrees of freedom
Residual deviance:  0.000  on 0  degrees of freedom
AIC: 13.211

Number of Fisher Scoring iterations: 3

>
```

(a) Give the logit form of the fitted model.

(b) Use the model to predict the odds and the probability of effectiveness for each of the two drugs.

(c) Find the odds ratio comparing the effectiveness of THC to prochlorperazine based on the binary logistic regression model. Write a sentence that interprets this value in the context of this problem and find a 95% confidence interval for the odds ratio. **Hint:** $z_{0.05} = 1.645$, $z_{0.025} = 1.96$

12. Can telling a joke affect whether or not a waiter in a coffee bar receives a tip from a customer? A study investigated this question at a coffee bar at a famous resort on the west coast of France (dataset: **TipJoke** in the R package Stat2data). Randomly assigned coffee-ordering customers to one of three groups: When receiving the bill one group also received a card telling a joke, another group received a card containing an advertisement for a local restaurant, and a third group received no card at all. Results are summarized below:

	Joke Card	Advertisement Card	No Card	Total
Left a Tip	30	14	16	60
Did note leave a Tip	42	60	49	151
Total	72	74	65	211

In this problem, the explanatory variable is the type of card (if any) given to the customer and the response variable is whether or not the customer left a tip. Let's use "No Card" as the reference value. Here is the R output of the fitted binary logistic regression model

```
>
> data(TipJoke)
> model = glm(Tip ~ Joke + Ad, family = binomial, data = TipJoke)
> summary(model)

Call:
glm(formula = Tip ~ Joke + Ad, family = binomial, data = TipJoke)

Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept)  -1.1192     0.2879  -3.887 0.000101 ***
Joke          0.7828     0.3742   2.092 0.036471 *
Ad           -0.3361     0.4135  -0.813 0.416413
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 251.94  on 210  degrees of freedom
Residual deviance: 242.14  on 208  degrees of freedom
AIC: 248.14

Number of Fisher Scoring iterations: 4

>
```

(a) Compute $\log\big(\hat{OR}(\text{Joke Card, No Card})\big)$ using the first table.

(b) Compute $\log\big(\hat{OR}(\text{Joke Card, No Card})\big)$ using the R output.

(c) Can you explain why the two answers in the previous two problems agree?

EXTRA CREDIT PROBLEM **Pregnancy Test**. A drug company is developing a new pregnancy-test kit for use in an outpatient basis. The company uses the pregnancy test on 100 women who are known to be pregnant; of these 95 test positive. The company uses the pregnancy test on 100 other women who are known to not be pregnant; of these 99 test negative. Define sensitivity and specificity for this problem and then calculate their values. .

True/False?

1. We found that a 95% confidence interval for the odds ratio relating having a heart attack (yes, no) to drug (placebo, aspirin) is (1.44, 2.33). If we had formed the table with aspirin in the first row (instead of placebo), then the 95% confidence interval would have been (1/2.33, 1/1.44)= (0.43, 0.69).

2. Suppose that income (high, low) and gender are conditionally independent, given type of job (secretarial, construction, service, professional, etc.). Then, income and gender are also independent in the 2×2 marginal table

WORKSHEET

CI for difference of proportions

$$(\hat{p}_1 - \hat{p}_2) \pm Z_{\alpha/2} \cdot \sqrt{\frac{\hat{p}_1(1-\hat{p}_1)}{n_1} + \frac{\hat{p}_2(1-\hat{p}_2)}{n_2}}$$

Test Statistic

$$\frac{\hat{p}_1 - \hat{p}_2}{\sqrt{\frac{\overline{p}(1-\overline{p})}{n_1} + \frac{\overline{p}(1-\overline{p})}{n_2}}}$$

$$\overline{p} = \frac{x_1 + x_2}{n_1 + n_2}$$

CI for log odds ratio

$$\ln(\hat{\theta}) \pm Z_{\alpha/2} \cdot (\text{SE})$$

$$\text{SE} = \sqrt{\frac{1}{n_{11}} + \frac{1}{n_{12}} + \frac{1}{n_{21}} + \frac{1}{n_{22}}}$$

CI for odds ratio

$$(e^{\ln(\hat{\theta}) - Z_{\alpha/2} \cdot (\text{SE})}, e^{\ln(\hat{\theta}) + Z_{\alpha/2} \cdot (\text{SE})})$$

Test Statistic

$$Z = \frac{\ln(\hat{\theta})}{\text{SE}}$$

Breslow-Day Test

$$\chi^2 = \sum_i \sum_j \sum_k \frac{(n_{ijk} - \hat{\mu}_{ijk})^2}{\hat{\mu}_{ijk}}, \quad \text{df} = k - 1$$

The Mantel-Haenszel estimate of the common odds ratio, θ

Let each partial table (indexed by k) have the form:

	$Y = 1$	$Y = 0$
$X = 1$	a_k	b_k
$X = 0$	c_k	d_k

$$\hat{\theta}_{MH} = \frac{\sum_{k=1}^{K} \frac{a_k d_k}{n_k}}{\sum_{k=1}^{K} \frac{b_k c_k}{n_k}}, \quad n_k = a_k + b_k + c_k + d_k$$

Cochran-Mantel-Haenszel Test for Conditional Independence

$$\chi^2_{CMH} = \frac{\sum_{k=1}^{K} \left(n^*_{ijk} - \frac{n_{i+k}\, n_{+jk}}{n_k} \right)^2}{\sum_{k=1}^{K} \frac{n_{1+k}\, n_{+1k}\, n_{2+k}\, n_{+2k}}{n_k^2 (n_k - 1)}}, \qquad \text{df} = 1$$

McNemar Test

Normal Version of McNemar Test

$$\chi^2 = \frac{\left(\left| n_A - \frac{n_D}{2} \right| - \frac{1}{2} \right)^2}{\frac{n_D}{4}}, \quad \text{df=1}$$

Exact Version of McNemar Test

1. $2 \sum_{k=0}^{n_A} (n_D\, k) \left(\frac{1}{2} \right)^{n_D}$ when $n_A < n_D/2$.

2. $2 \sum_{k=n_A}^{n_D} (n_D\, k) \left(\frac{1}{2} \right)^{n_D}$ when $n_A > n_D/2$.

3. $p = 1$ if $n_A = n_D/2$