STA 4504/5503 - Practice set 1 (with solutions)

True or False

1. In 2×2 tables, statistical independence is equivalent to a population odds ratio value of $\theta = 1.0$.
   TRUE

2. A British study reported in the New York Times: (Dec. 3, 1998) stated that of smokers who get lung cancer, "women were 1.7 times more vulnerable than men to get small-cell lung cancer." The number 1.7 is a sample odds ratio.
   FALSE

3. Using data from the Harvard Physician's Health Study, we find a 95% confidence interval for the relative risk relating having a heart attack to drug (placebo, aspirin) to be (1.4, 2.3). If we had formed the table with aspirin in the first row (instead of placebo), then the 95% confidence interval would have been (1/2.3, 1/1.4) = (.4, .7).
   TRUE

4. Pearson's chi-squared test of independence treats both the rows and the columns of the contingency table as nominal scale; thus, if either or both variables are ordinal, the test ignores that information.
   TRUE

5. For testing independence with random samples, Pearson's X^2 statistic and the likelihood-ratio G^2 statistic both have exact chi-squared distributions for any sample size, as long as the sample was randomly selected.
   FALSE

6. Fisher's exact test is a test of the null hypothesis of independence for 2×2 contingency tables that fixes the row and column totals and uses a hypergeometric distribution for the count in the first cell. For a one-sided alternative of a positive association (i.e., odds ratio > 1), the p-value is the sum of the probabilities of all those tables that have count in the first cell at least as large as observed, for the given marginal totals.
   TRUE

7. The difference of proportions, relative risk, and odds ratio are valid measures for summarizing 2×2 tables for either prospective or retrospective (e.g., case-control) studies.
   FALSE

8. An ordinary regression model that treats the response Y as having a normal distribution is a case of a generalized linear model, with normal (a.k.a. Gaussian) random component, identity link function and assuming the same predictors, i.e. same systematic component.
   TRUE

Open Ended Problems

1. Each of 100 multiple-choice questions on an exam has five possible answers but only one correct response. For each question, a student randomly selects one response as the answer.

   (a) Specify the probability distribution of the student's number of correct answers on the exam, identifying the parameter(s) for that distribution.
   Answer: $X \sim \text{Bin}(100, 0.20)$ where $n = 100$ is the number of trials and $\pi = 0.20$ is the probability of randomly choosing a correct response

   (b) Would it be surprising if the student made at least 50 correct responses? Explain your reasoning.
   Answer: Yes,

$$
\begin{aligned}
P(X \geq 50) &= 1 - P(X \leq 49) \\
&= \texttt{1-pbinom(49,100,0.20)} \\
&= \texttt{2.139244e-11}
\end{aligned}
$$

2. Consider the following data from a women's health study (MI is myocardial infarction, i.e., heart attack).

		MI	
		Yes	No
Oral Contraceptives	Used	23	34
	Never Used	35	132

   (a) Construct a 95% confidence interval for the population odds ratio.
   Answer:

```
> exp(log(23*132/(35*34))+c(-1,1)*1.96*sqrt(1/23+1/34+1/35+1/132))
[1] 1.335599 4.873415
```

   (b) Does it seem plausible that the variables are independent? Explain.
   Answer: The CI for θ does is strictly greater than 1 so the odds of MI for the used group are at least 33.56% larger and at most 387.34% from the the "never used" group.

3. For adults who sailed on the Titanic on its fateful voyage, the odds ratio between gender (female, male) and survival (yes, no) was 11.4.

   (a) What is wrong with the interpretation, "The probability of survival for females was 11.4 times that for males."
   Answer: Being an odds ratio it should be the odds of survival for females was 11.4 times that for males.

   (b) When would the quoted interpretation be approximately correct? Why?
   Answer: When the conditional probabilities of success are really small then,

$$\frac{\frac{\pi_1}{1-\pi_1}}{\frac{\pi_2}{1-\pi_2}} \approx \frac{\pi_1}{\pi_2}$$

   because the denominators $1 - \pi_i$ are close to 1.

   (c) The odds of survival for females equaled 2.9. For each gender, find the proportion who survived.
   Answer:

$$\frac{p_F}{1-p_F} = 2.9 \quad \Longrightarrow \quad p_F = 2.9/3.9 \approx 0.7436$$

4. Explain two ways in which the generalized linear model extends the ordinary regression model that is commonly used for quantitative response variables.
   Answer:

   (i) Random component can be any member of the exponential family of distributions

   (ii) Link function can be any continuous function whose first derivative exists

5. In a recent General Social Survey, gender was cross-classified with party identification. The output below shows some results.

```
> gp
          party
gender dem indep rep
female 279    73 225
male   165    47 191

> addmargins(gp)
          party
gender dem indep rep Sum
female 279    73 225 577
male   165    47 191 403
Sum    444   120 416 980

> gp.chisq <- chisq.test(gp)
> gp.chisq
Pearson's Chi-squared test
data: gp
X-squared = 7.0095, df = 2, p-value = 0.03005

> gp.chisq$expected
          party
gender dem     indep    rep
female 261.42 70.653 244.93
male   182.58 49.347 171.07

> round(myadjresids(gp), 2)
          party
gender dem  indep   rep
female 2.29  0.46 -2.62
male  -2.29 -0.46  2.62
```

(a) Explain what the numbers in the "expected" table represent. Show how to obtain 261.42.

Answer: "Expected values" are the cell counts that should have occured under the assumption of independence

$$
\begin{aligned}
\hat{\mu}_{ij} &= n\hat{\pi}_{ij} \\
&= n\hat{\pi}_{i+}\hat{\pi}_{+j} \\
&= n(n_{i+}/n)(n_{+j}/n) \\
&= (n_{i+}n_{+j})/n
\end{aligned}
$$

And for cell (1,1) that is 577(444)/980.

(b) Explain how to interpret the p-value given for the Chi-square statistic.
Answer: That is the p-value for testing the null that gender and party affiliation are independent. Since the p-value, the area to the right of 7.0095 under a χ^2_2 is 0.03005 is small we reject the null assumption of independence.

(c) Explain how to interpret the output of the last command (standardized adjusted residuals). Which counts were significantly higher than one would expect if party identification were independent of gender?
Answer: The only standardized residuals that are greater than 2 in absolute value fall

- for democrats that have higher than expected (under independence) counts for females, and consequently lower than expected males
- for republicans that have lower than expected counts for females and higher for males
- no significant relationship for independents

6. From a simulated data set containing information on 6047 customers such as whether the customer defaulted, is a student, and the average balance carried by the customer (ranging from 0 to 2700).

```
> model1 <- glm(default ~ balance, family = "binomial", data = train)
> summary(model1)
Coefficients:
              Estimate Std. Error z value Pr(>|z|)
(Intercept) -1.101e+01  4.887e-01  -22.52   <2e-16 ***
balance      5.669e-03  2.949e-04   19.22   <2e-16 ***
---
    Null deviance: 1723.03  on 6046  degrees of freedom
Residual deviance:  908.69  on 6045  degrees of freedom
AIC: 912.69
```

(a) Write out the model for predicting the probability of default and estimate the probability for a customer with a balance of 973.
Answer:

$$\hat{\pi}(973) = \frac{e^{-11.01+0.005669(973)}}{1+e^{-11.01+0.005669(973)}} = \dots$$

Note: If we had the model in `R` we could have used the appropriate `predict` function.

(b) Test whether balance is a significance predictor. Note there are two ways.
Answer: $\mathrm{H}_0 : \beta = 0$

- Wald p-value `<2e-16`
- LRT p-value `1-pchisq(1723.03-908.69,1)`≈ 0

(c) Using this output is there evidence of overdispersion? Explain.
Answer: Without being able to fit a beta-binomial we have to utilize X^2/df but since we don't have X^2 we have to make do with G^2 and hence 908.69/6045 is not greater than 1.

(d) Can we perform a goodness of fit test? True of False?
Answer: FALSE, as the we add more customers and their balance, which is a continuous non grouped variable, the number of unique predictor levels will also increase (and not remained anywhere close to fixed).