# Practice Exam #3 (Sta-209, S24)

Ryan Miller

The following information will appear verbatim on the first page of Exam 3.

**Directions**

- Answer each question using *no more than specified number of sentences* and not attempt to avoid these guidelines by using run-on sentences. Answers that are unnecessarily verbose may result in point loss.
- Do not include superfluous information in your answers, you may be penalized if you make an inaccurate statement even if you go on to provide a correct answer.

**Formula Sheet**

**Definitions**:

- **Risk**: relative frequency of an event/outcome
- **Relative Risk**: ratio of the risks across two groups
- **Odds**: ratio of how often an event/outcome is observed relative to how often it is not observed
- **Odds ratio**: ratio of odds across two groups

**Formulas**:

$$\bar{x} = \tfrac{1}{n}\sum_{i=1}^{n} x_i$$

$$s = \sqrt{\tfrac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2}$$

$$r = \tfrac{1}{n-1}\sum_{i=1}^{n}\left(\tfrac{x_i-\bar{x}}{s_x}\right)\left(\tfrac{y_i-\bar{y}}{s_y}\right)$$

| Statistic | Standard Error | Conditions |
|---|---|---|
| $\hat{p}$ | $\sqrt{\frac{p(1-p)}{n}}$ | $np \geq 10$ and $n(1-p) \geq 10$ |
| $\bar{x}$ | $\frac{\sigma}{\sqrt{n}}$ | normal population or $n \geq 30$ |
| $\hat{p}_1 - \hat{p}_2$ | $\sqrt{\frac{p_1(1-p_1)}{n_1} + \frac{p_2(1-p_2)}{n_2}}$ | $n_i p_i \geq 10$ and $n_i(1-p_i) \geq 10$ for $i \in \{1, 2\}$ |
| $\bar{x}_1 - \bar{x}_2$ | $\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}$ | normal populations or $n_1 \geq 30$ and $n_2 \geq 30$ |

Chi-squared test statistic: $X^2 = \sum_{j=1}^{k} \frac{(\text{observed}_j - \text{expected}_j)^2}{\text{expected}_j}$

## Question #1 (conceptual questions)

**Part A**: Suppose we are interested in building a linear regression model that predicts daily ozone concentration based upon three quantitative explanatory variables: temperature, wind speed, and solar radiation. Identify which of the following statements must be **true** (there may be more than 1 true statement):

A) The model: $\widehat{Ozone} = b_0 + b_1 Temp + b_2 Wind$ will have a smaller sum of squared residuals than the model $\widehat{Ozone} = b_0 + b_1 Solar$
B) The model: $\widehat{Ozone} = b_0 + b_1 Temp + b_2 Wind$ will have a smaller sum of squared residuals than the model $\widehat{Ozone} = b_0 + b_1 Temp$
C) The model: $\widehat{Ozone} = b_0 + b_1 Temp + b_2 Temp^2$ will have a smaller sum of squared residuals than the model $\widehat{Ozone} = b_0 + b_1 Wind$

State which statements are true and briefly explain the reasoning or thought process you used to determine whether a statement was true or false.

**Part B**: For each of the following scenarios state the name of the appropriate hypothesis test. You do not need to explain your answers.

- i: Using a sample Grinnell students from the science division to see if the racial/ethnic distribution of science students at Grinnell differs from the distribution of the entire student body that is published by the college.
- ii: Conducting a randomized experiment to determine if fertilizer A produces a higher average crop yield than fertilizer B.
- iii: Using a poll that asks $n = 200$ voters if they will vote for candidate A or candidate B to see if there is evidence that candidate A will receive a majority of votes in the election.
- iv: Conducting an experiment where 4 different brands of feed supplements are given to piglets with the intent of determining whether there is an association between type of feed supplement during youth and the adult weight of a pig.

**Part C**: Recall that one-way ANOVA can be described as a comparison between two models using the observed sample data. With this in mind, answer the following questions:

- i: Suppose we are interested in how each model involved in one-way ANOVA will predict the value of the outcome variable for a new observation. Briefly describe what the prediction will be based upon for each model.
- ii: The models in one-way ANOVA involve the Normal distribution. Briefly describe the role of the Normal distribution in these models.
- iii: Suppose we perform one-way ANOVA and reject the null hypothesis. We check the model's assumptions and they are verified as reasonable. Is this the end of our analysis or is there more that we should do? If this is end, briefly describe what we'd conclude from the test (in generic terms). If more should be done, briefly describe what you'd do next.

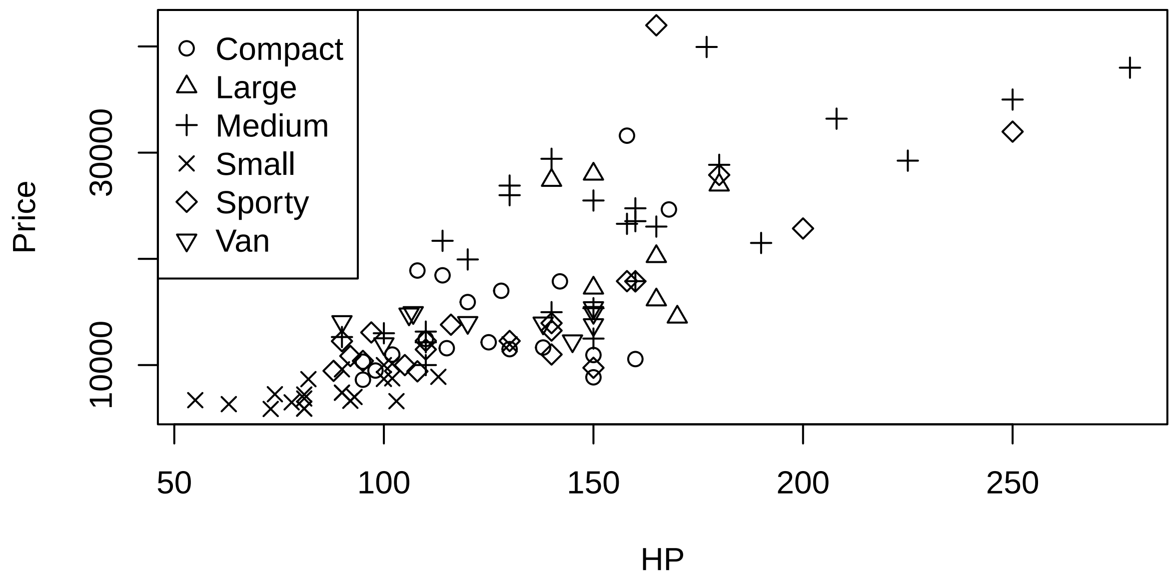
## Question #2

This question will analyze data on 111 different types of cars published in *Consumer Reports*. The overall goal of the analysis is to identify factors associated with price. A few key variables include:

- **Price** - List price (US dollars) with standard equipment
- **Country** - Where the car was manufactured
- **HP** - Net horsepower
- **Type** - A categorical variable describing the general type of vehicle (small, medium, large, compact, sporty, van)

- **Length** - Length of the vehicle (inches)

**Part A**: The plot below depicts the relationship between Price, HP, and Type. Based upon this plot, is HP associated with Price? Is Type associated with Price? Provide a brief explanation of your answers.




**Part B**: Ignoring all other variables, what *statistical approach* would be the most appropriate *hypothesis test* for discovering a possible association between Type and Price? Provide the name of the test and a brief explanation (no more than 1-sentence).

**Part C**: The table below summarizes price by vehicle type. Is any information presented in this table problematic for the validity of the statistical test you identified in Part A? If so, briefly explain what aspect(s) of these data are problematic.

| Type | N | Mean | Median | StdDev |
|---|---|---|---|---|
| Compact | 19 | 14395.368 | 11650.0 | 5938.762 |
| Large | 7 | 21499.714 | 20225.0 | 5825.878 |
| Medium | 26 | 22750.154 | 23170.0 | 8416.809 |
| Small | 22 | 7736.591 | 7239.5 | 1627.928 |
| Sporty | 21 | 15889.810 | 12279.0 | 8539.240 |
| Van | 10 | 14014.300 | 14037.5 | 1126.104 |

**Part D**: The table below displays the *coefficient estimates* of a linear regression model that uses both Type and HP to predict a vehicle's Price. Use the information in this table to answer the following questions (I - III)

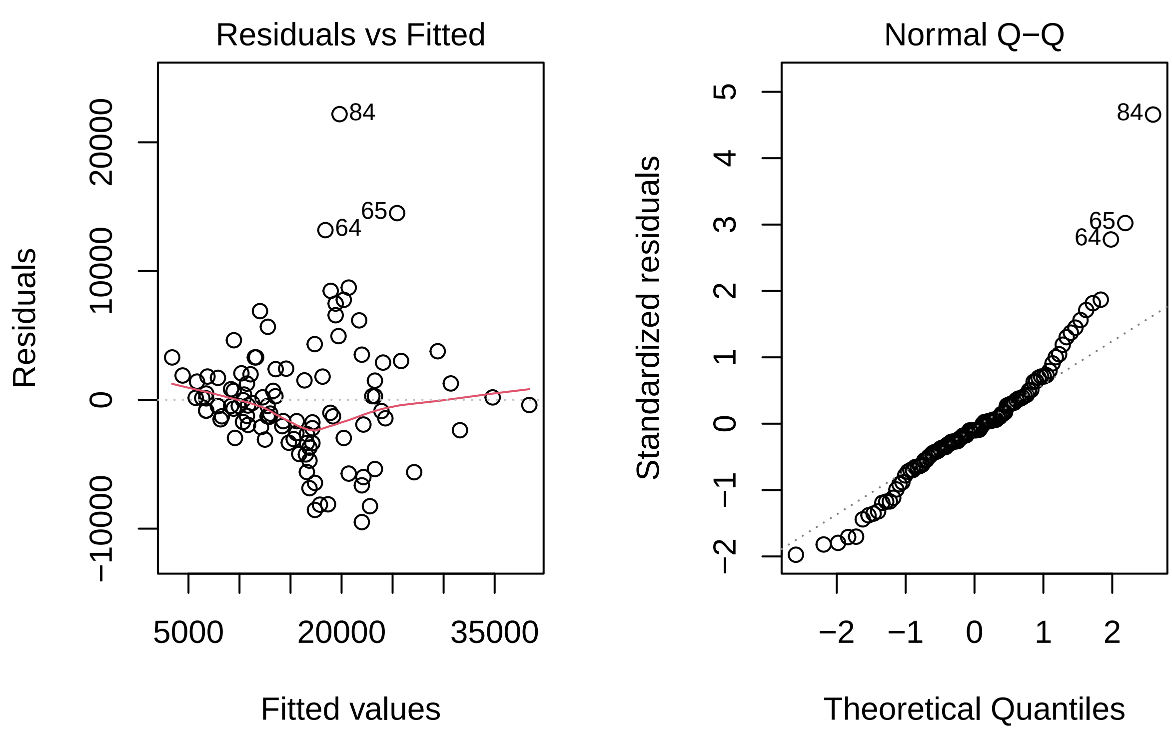
| | Coefficient | Std. Error | t statistic | p-value |
|---|---|---|---|---|
| (Intercept) | -1842.01 | 2197.32 | -0.84 | 0.40 |
| HP | 128.23 | 14.91 | 8.60 | 0.00 |
| TypeLarge | 2825.67 | 2224.20 | 1.27 | 0.21 |
| TypeMedium | 4593.94 | 1543.07 | 2.98 | 0.00 |
| TypeSmall | -1810.14 | 1635.77 | -1.11 | 0.27 |
| TypeSporty | 488.56 | 1556.85 | 0.31 | 0.75 |
| TypeVan | -248.79 | 1915.62 | -0.13 | 0.90 |

I) The intercept of this model is -1842.01, what does this value mean? Should we care that this value isn't statistically significant?

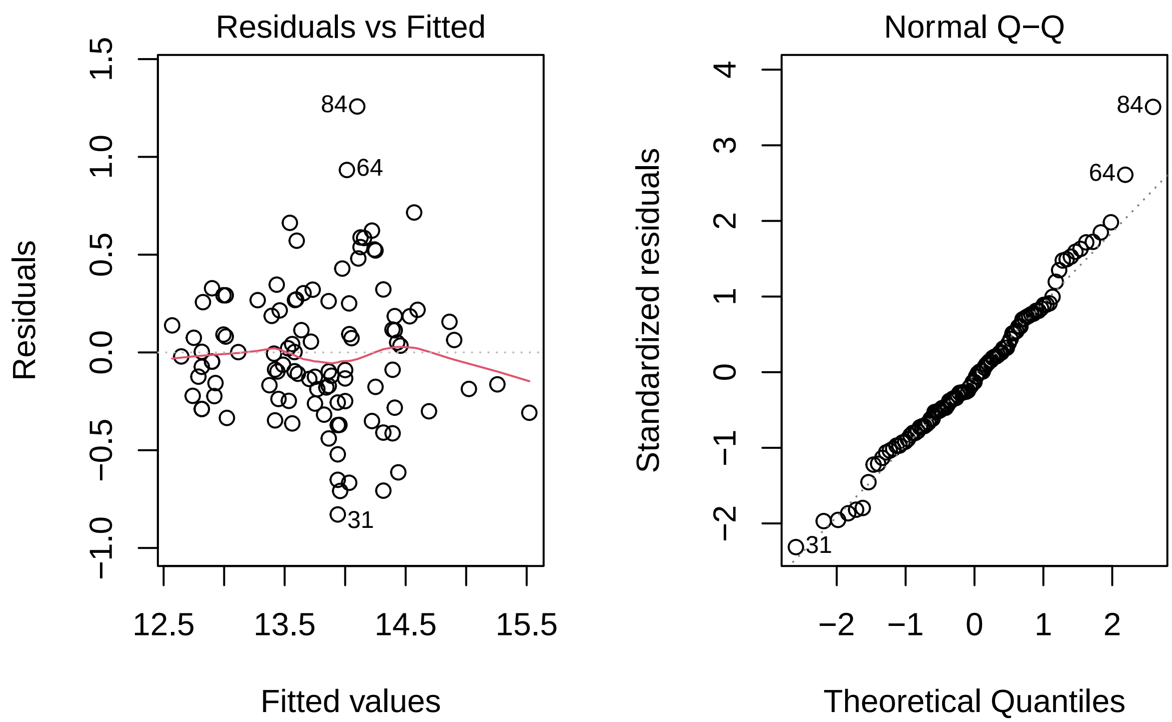
II) Provide a one sentence interpretation of the coefficient for "TypeMedium", be specific.

III) True or False, in this model the effect of HP on price differs depending on the type of vehicle. You do not need to explain your answer.

**Part E**: Below are two `R` plots related to the model described in Part D, `Price ~ HP + Type`. Based upon what you see in these plots, do you believe $p$-values calculated for these data will be valid/reliable? Briefly explain.



**Part F**: The plots show results after transforming the response variable `Price` using a log-transformation, making the model: `log2(Price) ~ HP + Type`. When compared with the model from Parts D-E, are you more comfortable trusting the *p*-values produced by statistical tests that use this model? Briefly explain why or why not.




**Part G**: Below are statistical results found using `R`. Based upon what is given, state the null hypothesis of the test that was performed in words and provide a one-sentence conclusion describing the results of the test in regard to the null hypothesis.

```
mod0 <- lm(log2(Price) ~ HP, data = car90)
mod1 <- lm(log2(Price) ~ HP + Type, data = car90)
anova(mod0, mod1)
```

```
## Analysis of Variance Table
##
## Model 1: log2(Price) ~ HP
## Model 2: log2(Price) ~ HP + Type
##   Res.Df    RSS Df Sum of Sq      F    Pr(>F)
## 1    103 19.333
## 2     98 13.347  5    5.9861 8.7906 6.427e-07 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

## Question #3

In the mid-1860s, Joseph Lister, a Professor of Surgery at the Glasgow Royal Infirmary, conducted an experiment to investigate his hypothesis that harmful micro-organisms were the cause of deadly infections that frequently occurred after surgery. Lister randomly assigned 75 surgical patients to receive either his newly developed "sterile" surgery protocol, which entailed wearing clean gloves, gowns, and disinfecting surgical instruments, or a "control" surgery protocol, where no sterilizations steps were taken prior to surgery.

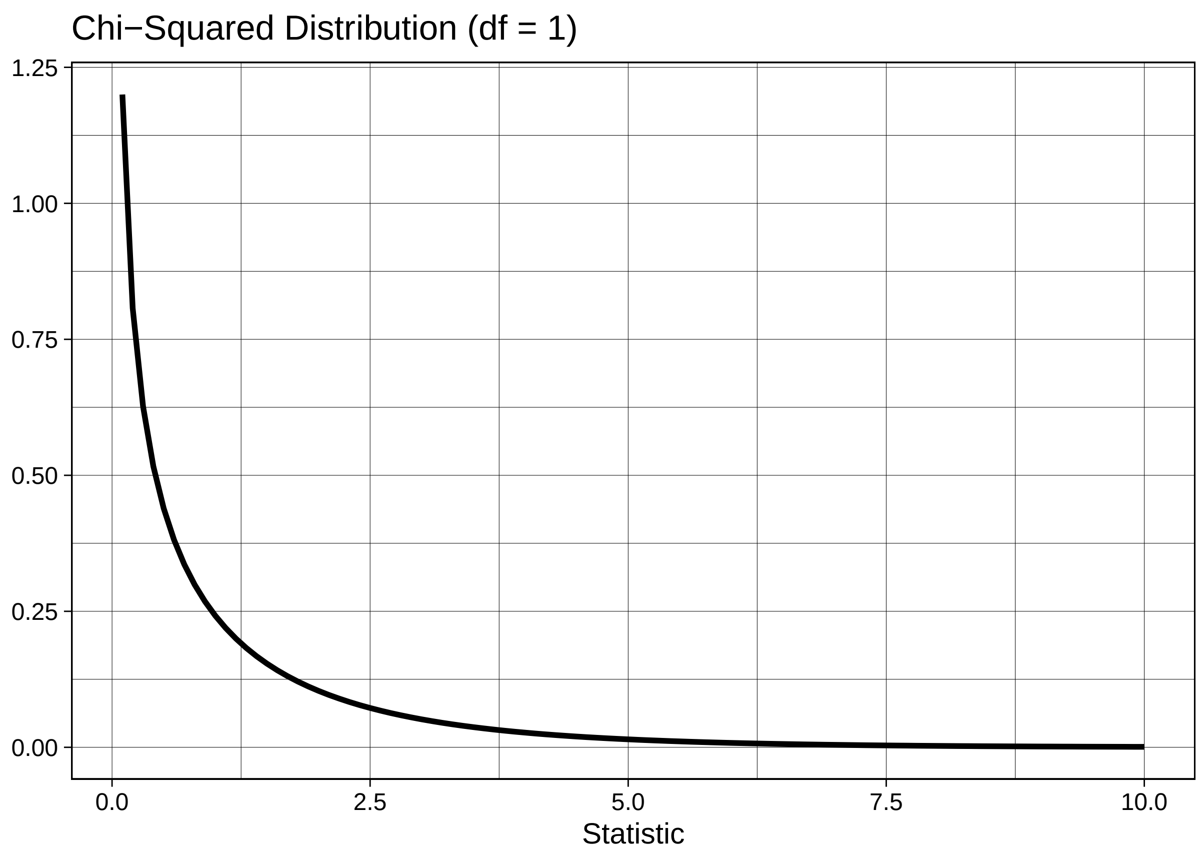
The results of Lister's experiment are summarized below:

| | Died | Survived |
|---|---:|---:|
| Control | 16 | 19 |
| Sterile | 6 | 34 |

**Part A**: If Lister's sterilization protocol made no difference, how many deaths and survivals would you expect in each group (sterile and control)? Provide a table of expected counts.

**Part B**: Use the table of expected counts you found in Part A to calculate a test statistic for a Chi-squared test of association.

**Part C**: Shown below is Chi-squared distribution with 1 degree of freedom. Using your results from Part A/B, shade the region of this curve corresponding to the $p$-value.



**Part D**: Estimate the $p$-value using your response to Part C, and provide a brief conclusion that is consistent with this $p$-value that involves the context of this hypothesis test.