*Stat 204* *Exam 3b* *Name:*

---

***All statements should be expressed in complete sentences, and every numerical answer should be justified by showing how it was obtained. Writing R code that will actually calculate the numerical answer is definitely encouraged, followed by displaying the numerical answer your code returns. Recall that the point of this exam is to demonstrate your mastery of complete analytic processes.***

*I pledge that I have neither given nor received unauthorized aid on this exam. Pledged:* .....

Slide your completed exam under my office door, WL120, before the due date and time for your section of Stat 204.

Table 1: **Due Dates and Times**

| Section | Due date and time |
|---|---|
| A (noon) | Tue, May 3, 4 pm |
| B (1 pm) | Tue, May 3, 11 am |

**chi-square HT for goodness of fit**

[3 points] (Peck, resources) *In the United States, professional baseball determines the winner of the year's competition by playing a series of games known as the World Series. The first team to win four out of seven games is declared the winner. A sportscaster believes that by the time the World Series comes around, it is reasonable to suppose that on average the competing teams are equally matched, and that the probability is 0.5 of either team winning any one game. Others disagree with this theory. The table below contains the number of times from 1903 – 2010 that the series lasted 4, 5, 6, or 7 games, as well as the probabilities associated with those outcomes if the sportscaster is correct. Do the data provide sufficient evidence at the 0.05 level that the sportscaster's belief is* ***incorrect****?* [1]

Table 2: **World Series**

| Number of games | 4 | 5 | 6 | 7 |
|---|---|---|---|---|
| p | 0.125 | 0.25 | 0.3125 | 0.3125 |
| Observed x | 20 | 24 | 23 | 35 |

[1] Roxy Peck, "Statistics, Learning from Data,"" resources

For a chi-square HT for goodness of fit:

- Define the univariate categorical variable of interest and its $k$ levels

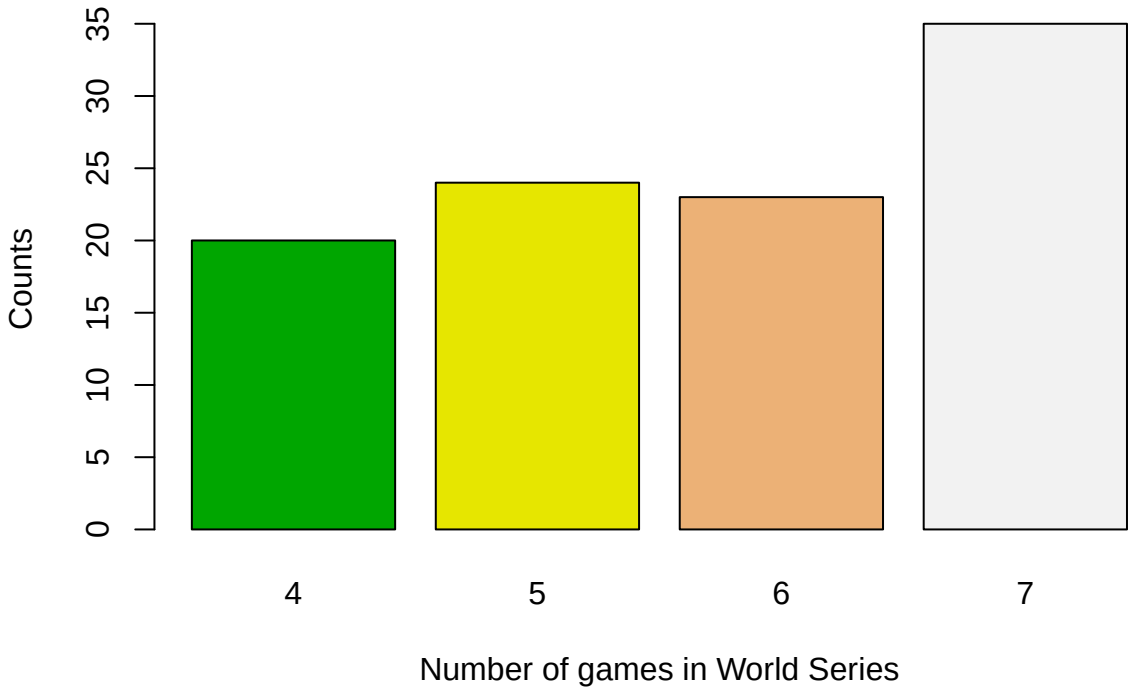
- Define the population proportions for each level, $p_1, \ldots, p_k$

- State $H_0, H_a$ and $\alpha$

```
alpha <- 0.05
```

- Calculate the sample counts for each level. View the data.

```
n.games <- 4:7
p <- c(0.125, 0.25, 0.3125, 0.3125)
obs <- c(20, 24, 23, 35)
barplot(obs,
        names.arg=4:7, col=terrain.colors(4),
        ylab="Counts", xlab="Number of games in World Series")
```




- Calculate the test statistic $X^2$, $df$, and the P-value $p.value$

```
chisq.test(obs, p=p)
```

```
##
##  Chi-squared test for given probabilities
##
## data:  obs
## X-squared = 6.9882, df = 3, p-value = 0.07227
```

- On a sketch of the appropriate $\chi^2$ distribution, locate the test statistic $X^2$ and shade the region whose area is $p.value$

- State the formal conclusion of the HT and explain how you reached that conclusion

- State the conclusion in context. Mention the level of significance.

**HT and CI concerning $\beta$**

[4 points] (Peck, resources) *A veterinary graduate student is studying the relationship between the weight (in pounds) of one year-old golden retrievers (y) and the amount of dog food (in pounds) the dog is fed each day (x). A random sample of 10 one-year-old golden retrievers yielded the following data.* [1]

[1] Peck, Statistics, Learning from Data, resources

*Carry out a model utility test as outlined below.*

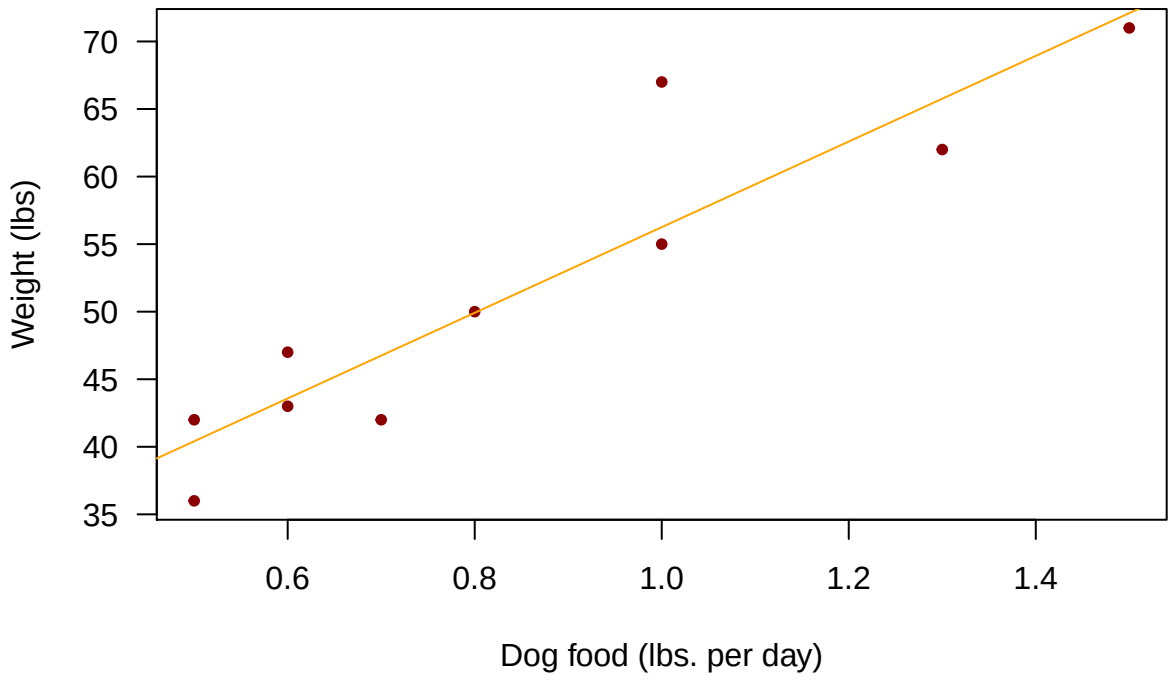
*Calculate and interpret a 95% confidence interval for $\beta$.*

Import the data.

```
x <- c(0.5, 1.0, 0.6, 1.0, 1.3, 1.5, 0.5, 0.7, 0.8, 0.6)  #  dog food (lbs. per day)
y <- c(42, 67, 47, 55, 62, 71, 36, 42, 50, 43)            #  weight (lbs)
```

View the data.

```
plot(x, y,
     pch=20, las=1, col="darkred",
     xlab="Dog food (lbs. per day)", ylab="Weight (lbs)")
retriever.lm <- lm(y ~ x)
abline(retriever.lm, col="orange")
```




Linear model.

$$\hat{y} = a + bx$$

- What are the values of $a$ and $b$ for this linear model?

```
retriever.lm.summ <- summary(retriever.lm)
retriever.lm.summ
```

```
##
## Call:
## lm(formula = y ~ x)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -4.747 -3.134 -0.838  1.215 10.746
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)   24.563      4.288   5.729 0.000440 ***
## x             31.690      4.709   6.729 0.000148 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 4.86 on 8 degrees of freedom
## Multiple R-squared:  0.8499, Adjusted R-squared:  0.8311
## F-statistic: 45.28 on 1 and 8 DF,  p-value: 0.0001482
```

Test for model utility.

- State $H_0, H_a$ and *alpha*.

```
alpha <- 0.05
```

- Using R's values for $b$ and $SE_b$, show R code for calculating $t, df$, and $p.value$, and indicate the resulting values.

```
retriever.lm.summ$coefficients
```

```
##             Estimate Std. Error  t value     Pr(>|t|)
## (Intercept) 24.56338   4.287852 5.728598 0.0004396659
## x           31.69014   4.709367 6.729172 0.0001481886
```

```
#  b, beta, se.b
b <- 31.69014
beta <- 0
se.b <- 4.709367
#  t


#  n and df
```

```
#  p.value


#
```

- On a sketch of the appropriate $t$ distribution, mark the test statistic, $t$, and shade the region whose area is $p.value$.

- State the formal conclusion of the hypothesis test for model utility and explain how you reached that conclusion.

- Interpret the conclusion in context. Mention the level of significance.

95% CI for $\beta$

- Show R code for calculating $point.estimate, t.star$ and $ci$, and indicate the resulting values.

```
#  point.estimate


#  t.star


#  ci


#
```

- Sketch the CI.

- Interpret the CI in context. Mention the significance level.

**ANOVA**

[3 points] (Peck, 17.8, p.16) *In an experiment to investigate the performance of five different brands of spark plugs intended for use on a 125-cc motorcycle, six plugs of each brand were tested, and the number of miles (at a constant speed) until failure was observed. A partially completed ANOVA table is given. Fill in the missing entries, and test the relevant hypotheses using a 0.05 level of significance.*

| source | df | SS | MS | F |
|---|---|---|---|---|
| treatments | | | | |
| error | | 235419.04 | | |
| total | | 310500.76 | | |

For a single factor ANOVA F test for equality of more than two means:

- Define the response variable and the $k$ population or treatment subgroups

- State $H_0, H_a$, and $\alpha$

  $$H_0 : \text{the five means are the same}$$

  $$H_a : \text{at least two means differ}$$

  $alpha = 0.05$

- Calculate $N, k, df_1, df_2$

```
#  N


#  k


#  df1


#  df2


#
```

Complete the ANOVA table.

Show appropriate R code for calculating each table entry and the resulting values. See Peck, chapter 17, p.14.

```
#  df.total


#  SS.treatments


#  SS.error


#  MS.treatments


#  MS.error


#  F


#
```

- Calculate the P-value *p.value*

Show R code to calculate *p.value* and the value it returns.

```
#  p.value


#
```

- On a sketch of the appropriate $F$ distribution, locate the test statistic $F$ and shade the region whose area is *p.value*

- State the formal conclusion of the HT and explain how you reached that conclusion.

- State the conclusion in context. Mention the confidence level.