MATH 338

FINAL EXAM

SOFTWARE PORTION

MON/WED/THURS, DEC 11/13/14, 2017

Your name: _____

Your scores (to be filled in by Dr. Wynne):

Problem 1: ____/5

Problem 2: ____/7

Problem 3 (optional): ____/4

Problem 4 (optional): ____/5

Total: ____/21

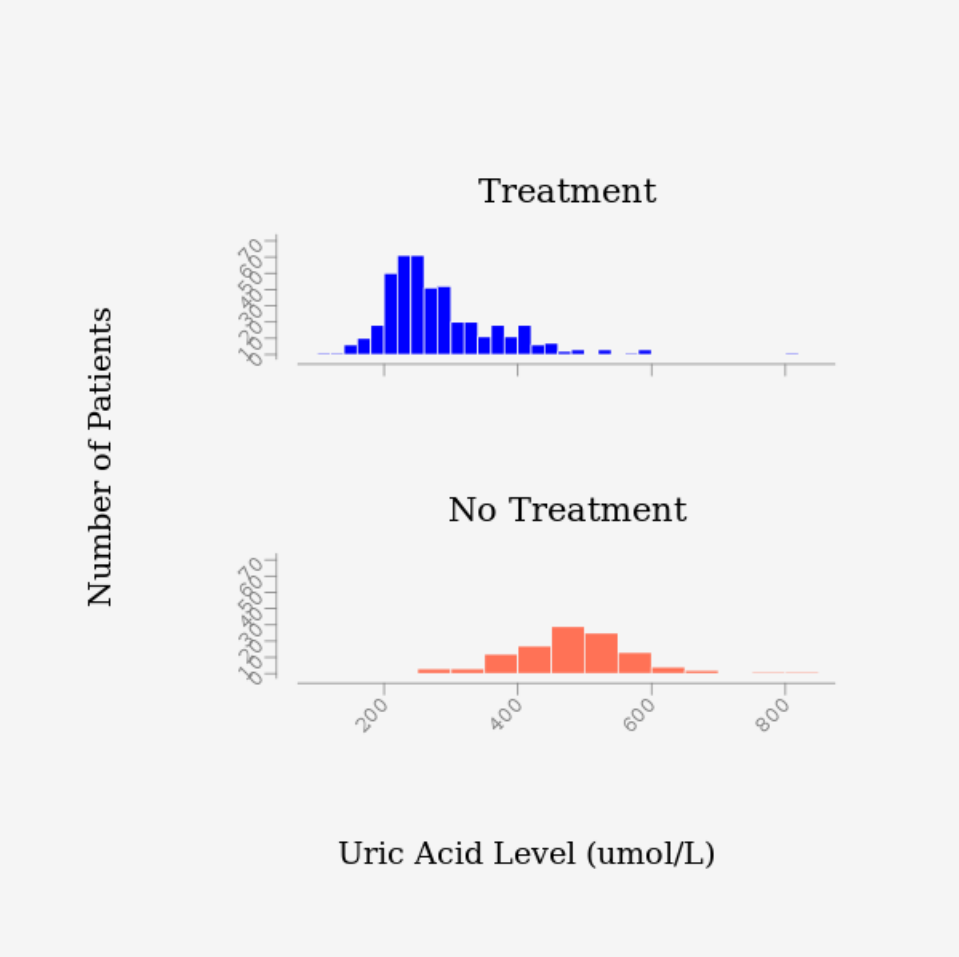
You may refer to your textbook, any notes/code you wrote, anything on Titanium, and software help menus. You may ask Dr. Wynne to clarify what a question is asking for, or to help you troubleshoot RGuroo errors and/or debug your R code. You may not ask other people for help or use online resources other than those on Titanium or the software itself.

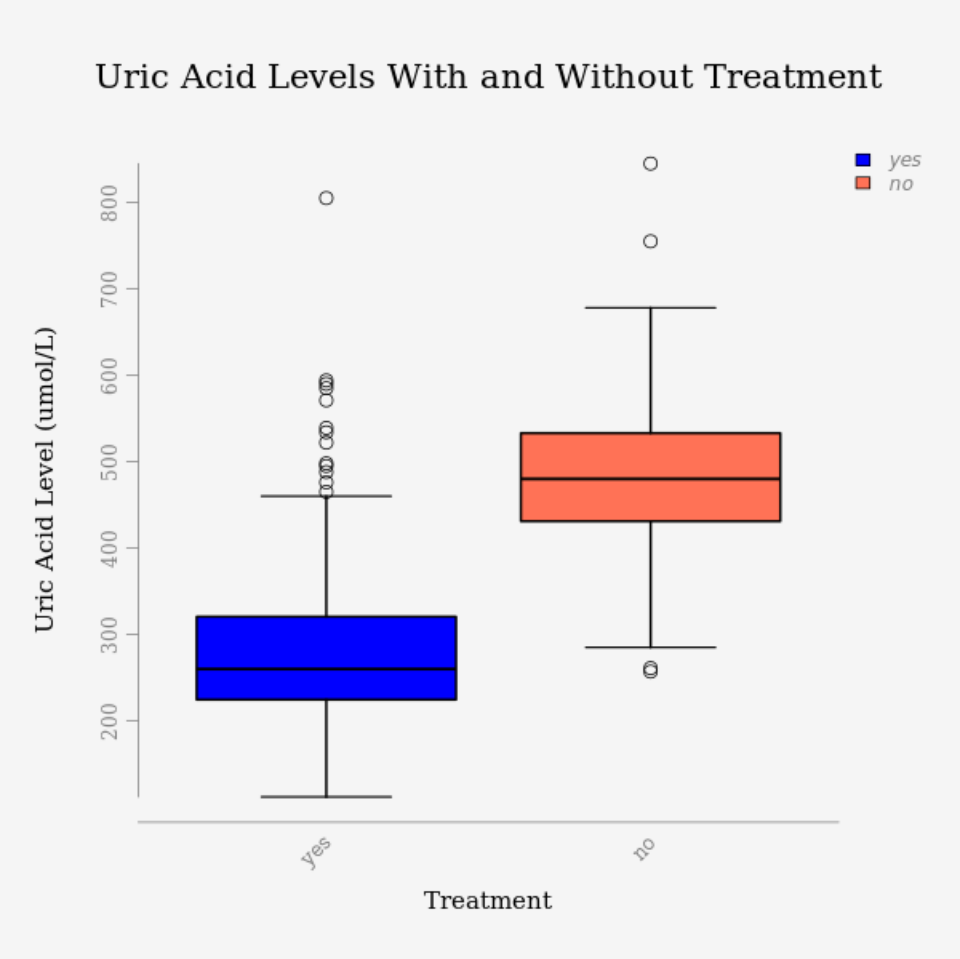
For full credit, include all R code (if using RStudio), graphs, and output. Save your answers as a .docx or .pdf file and upload the file to Titanium.

Problem 1. Abhishek and colleagues (2017) investigated a group of 550 patients with acute gout attacks. In the gout.csv data set, the variable urate_lowering_treatment indicates whether the patient was on a drug to lower uric acid levels (in μmol/L). The patient's measured uric acid level is recorded as the variable serum_uric_acid. Note that some patients may have missing values for one or both of these variables.

A. [2 pts] Graphically display the distribution of serum uric acid levels in patients with and without treatment.

RGUROO POSSIBILITIES:

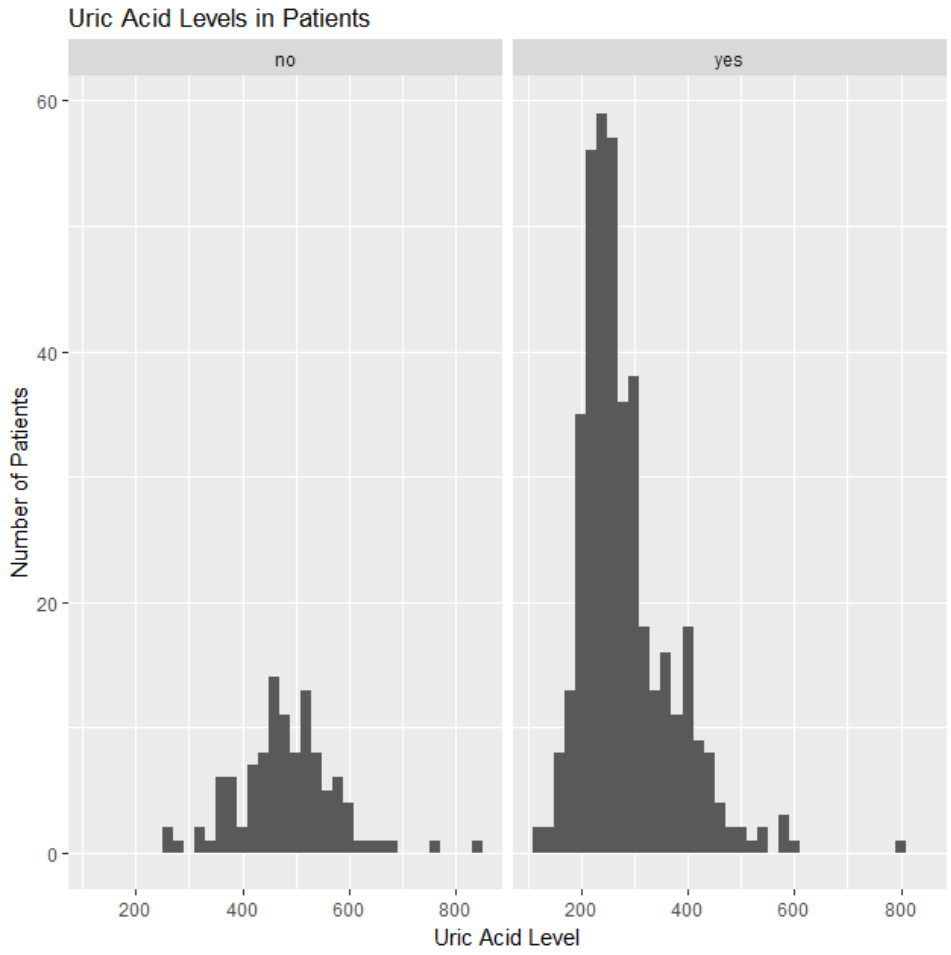




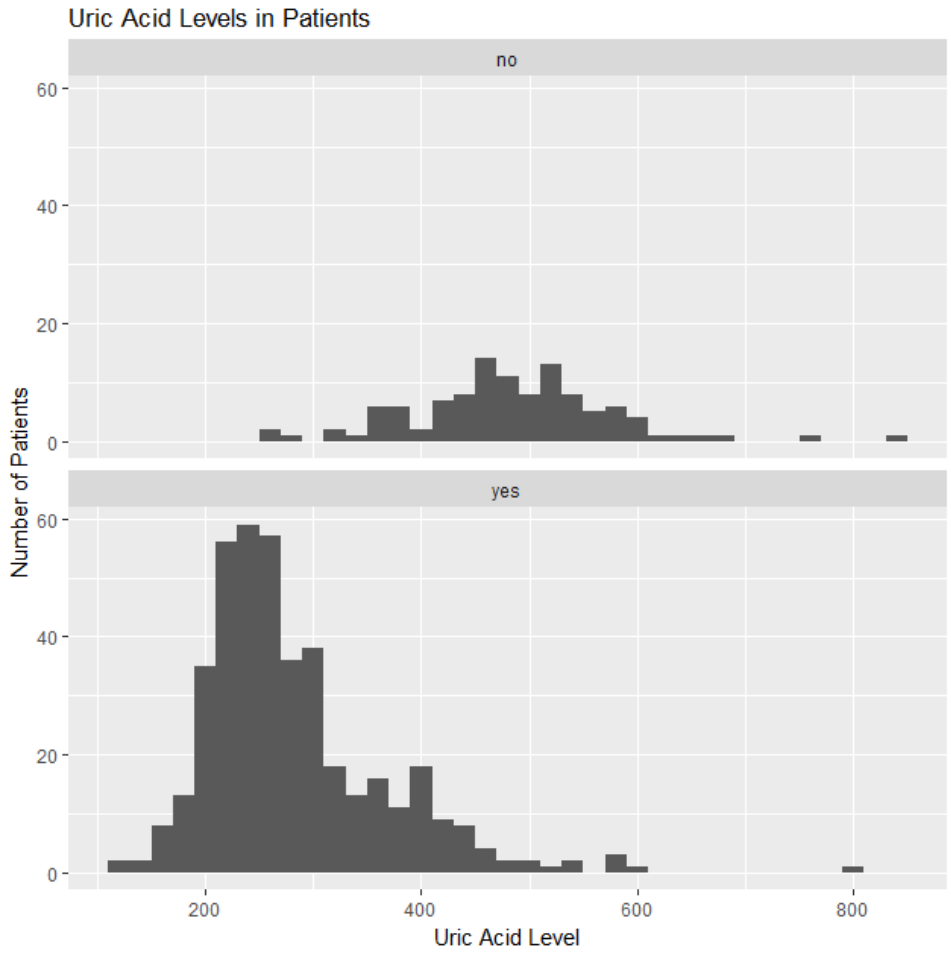
Obviously I expect your graphs to be a little less customized than these (for instance, with the "no" level on top/left).

RSTUDIO POSSIBILITIES:

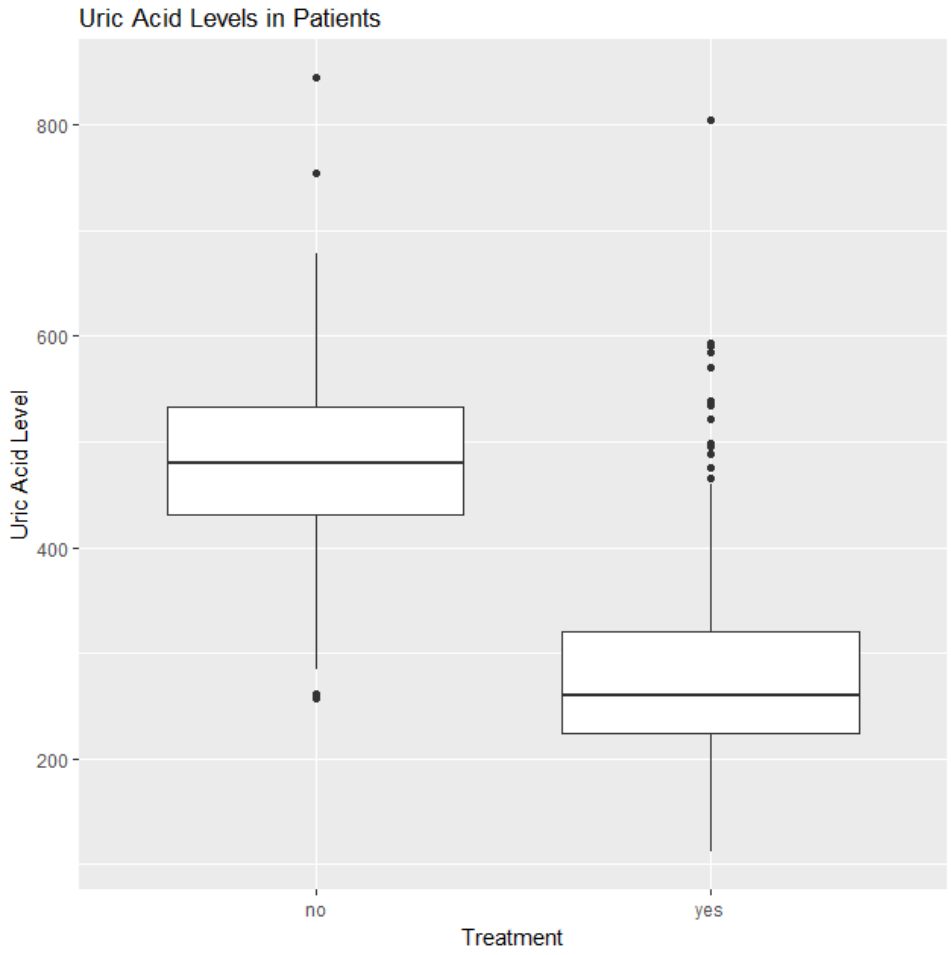
```
> gout_histograms <- ggplot(data = gout, mapping = aes(x =
serum_uric_acid)) +
+   geom_histogram(center = 400, binwidth = 20) +
+   labs(title = "Uric Acid Levels in Patients", x = "Uric Acid
Level", y = "Number of Patients") +
+   facet_wrap(~urate_lowering_treatment)
> print(gout_histograms)
```




```
> gout_histograms <- ggplot(data = gout, mapping = aes(x =
serum_uric_acid)) +
+   geom_histogram(center = 400, binwidth = 20) +
+   labs(title = "Uric Acid Levels in Patients", x = "Uric Acid
Level", y = "Number of Patients") +
+   facet_wrap(~urate_lowering_treatment, ncol = 1)
> print(gout_histograms)
```




```
> gout_boxplots <- ggplot(data = gout, mapping = aes(x =
urate_lowering_treatment, y = serum_uric_acid)) +
+   geom_boxplot() +
+   labs(title = "Uric Acid Levels in Patients", x =
"Treatment", y = "Uric Acid Level")
> print(gout_boxplots)
```



1 pt for a set of histograms or side-by-side boxplots, 1 pt for making the graph correctly with correct graph/axis titles

B. [3 pts] Estimate with 95% confidence the population mean difference in serum uric acid level between gout patients on a urate-lowering treatment and gout patients not on a urate-lowering treatment. Interpret your estimate. Assume that your answer in part (A) suggests that your choice of inference is appropriate.

0.5 pts: We wish to estimate the value of $\mu_1 - \mu_2$ using a two-sample t confidence interval for difference in means.

0.5 pts: one of the following

The interval estimate is (182.05, 220.46) if you defined population 1 to be individuals without treatment and population 2 to be individuals with treatment.

The interval estimate is (-220.46, -182.05) if you defined population 1 to be individuals with treatment and population 2 to be individuals without treatment.

1 pt: We are 95% confident that the mean serum uric acid level in patients on a urate-lowering treatment is between 185.05 and 220.46 μmol/L lower than the mean serum uric acid level in patients not on the treatment.

1 pt for, at minimum, one of the following:

Rguroo output:

Confidence Interval - t Distribution

95% Confidence interval

Variable	DF	Lower CL	Upper CL	Mean	Margin of Error
serum_uric_acid (no) - serum_uric_acid (yes)	160.691	182.054	220.459	201.257	19.2022

RStudio code/output:

```
> t.test(serum_uric_acid ~ urate_lowering_treatment, data =
gout)

        Welch Two Sample t-test

data:  serum_uric_acid by urate_lowering_treatment
t = 20.698, df = 160.69, p-value < 2.2e-16
alternative hypothesis: true difference in means is not equal to
0
95 percent confidence interval:
 182.0544 220.4589
sample estimates:
 mean in group no mean in group yes
         483.9000          282.6434
```

Problem 2. The milk.csv file contains the concentration of trace elements in human milk fed to a sample of infants from three different countries (Argentina, Poland, and United States). All element concentrations are in µg/L.

A. [2 pts] Report the least-squares regression equation predicting arsenic (As) concentration in milk from the copper (Cu) concentration in milk.

1 pt EQUATION: As = 3.044 + 0.005 (Cu)

or $\widehat{As} = 3.044 + 0.005\ (Cu)$

1 pt for one of the outputs below

Rguroo Parameter Estimates table:

Parameter Estimates

Variable	Parameter Estimate	Standard Error	t Value	Pr > |t|
(Intercept)	3.04423	0.382754	7.95349	4.76536e-11
Cu	0.00478968	0.00188539	2.54042	0.0135907

RStudio code/output:

```
> milk_lm <- lm(As ~ Cu, data = milk)
> summary(milk_lm)

Call:
lm(formula = As ~ Cu, data = milk)

Residuals:
    Min      1Q  Median      3Q     Max
-1.4478 -0.7198 -0.1987  0.4446  4.2962

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 3.044233   0.382754   7.953 4.77e-11 ***
```

```
Cu          0.004790   0.001885   2.540   0.0136 *
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 1.103 on 62 degrees of freedom
Multiple R-squared:  0.09428,   Adjusted R-squared:  0.07967
F-statistic: 6.454 on 1 and 62 DF,  p-value: 0.01359
```

B. [1 pt] In this model, is copper concentration a significant predictor of arsenic concentration, at the 5% significance level? Why or why not? Assume that it is valid to do this inference.

0.5 pts: Copper concentration is significant at the 5% level.

0.5 pts: The t-test for significance of the slope corresponding to copper concentration in this model (or the ANOVA F test for significance of the overall model, which has an equivalent null/alternative hypothesis) produces a p-value of 0.0136 < 0.05.

C. [4 pts] Determine whether the population mean arsenic concentration in milk differs between the three countries. Assume that exploratory analysis has already been done and the assumptions for inference have been checked; just do the inference. For this problem you will need to set your own confidence/significance level.

0.5 pts: Use a One-Way ANOVA (or One-Way ANOVA F TEST)

0.5 pts: Because we have a numerical response variable and a categorical explanatory variable with three categories

0.5 pts: H_0: $\mu_{Argentina} = \mu_{Poland} = \mu_{USA}$ and/or H_a: at least one of the three means is different from the others (one or the other hypothesis must be specified for credit)

0.5 pts: Define a significance level (presumably either 0.05 or 0.01)

1 pt: $F(2, 61) = 4.90$, $p = 0.0106$. Therefore, we reject the null hypothesis at the 5% level, and fail to reject at the 1% level. We have (do not have at 1%) significant evidence that there is a difference in the mean arsenic levels in human milk between the three countries. We should (should not) do post-hoc tests to determine which pairs of means are different.

1 pt for one of the outputs below, at minimum

Rguroo output:

ANOVA Table

Source	DF	Sum of Squares	Mean Squares	F Value	Pr > F
Regression	2	11.5377	5.76883	4.90374	0.0105961
Residual	61	71.7612	1.17641		
Total	63	83.2988			

RStudio code/output 1:

```
> milk_anova <- aov(As ~ Country, data = milk)
> summary(milk_anova)
            Df Sum Sq Mean Sq F value Pr(>F)
Country      2  11.54   5.769   4.904 0.0106 *
Residuals   61  71.76   1.176
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

RStudio code/output 2: note that the F test statistic, df, and p-value are at the bottom of the output here

```
> milk_lm2 <- lm(As ~ Country, data = milk)
> summary(milk_lm2)

Call:
lm(formula = As ~ Country, data = milk)

Residuals:
    Min      1Q  Median      3Q     Max
-1.9767 -0.6197 -0.2089  0.2434  4.5665
```

```
Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept)     4.5149     0.2367  19.076  < 2e-16 ***
CountryPoland  -0.6585     0.3274  -2.012  0.04869 *
CountryUSA     -1.0463     0.3389  -3.087  0.00304 **
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 1.085 on 61 degrees of freedom
Multiple R-squared:  0.1385,    Adjusted R-squared:  0.1103
F-statistic: 4.904 on 2 and 61 DF,  p-value: 0.0106
```

RStudio code/output 3:

```
> milk_lm2 <- lm(As ~ Country, data = milk)
> milk_anova2 <- aov(milk_lm2)
> summary(milk_anova2)
            Df Sum Sq Mean Sq F value Pr(>F)
Country      2  11.54   5.769   4.904 0.0106 *
Residuals   61  71.76   1.176
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

Problem 3. [4 pts] A 2017 study investigated the prevalence of drunkenness at Swedish football (soccer) matches. The researchers defined a blood alcohol content of at least 0.1% to be "highly intoxicated." In a random sample of 4420 spectators, 395 had a blood alcohol content (BAC) of at least 0.1%. Construct and interpret a 95% confidence interval for the proportion of all Swedish football (soccer) match spectators who are highly intoxicated.

0.5 pts: Use a large-sample (or one-proportion) z confidence interval.

1 pt: Because we have a single categorical response variable with levels Highly Intoxicated/Not Highly Intoxicated. The B and N assumptions of BINS are met and there is no reason to believe the I and S assumptions are not met (for instance, we assume that their random sample does not contain groups of people who have all been drinking together). We have at least 10 successes and 10 failures in our sample, so we are okay to use z procedures instead of a binomial CI.

(full 1.5 points for binomial exact CI with justification of everything except sample size)

1.5 pts: We are 95% confident that the true proportion of highly intoxicated spectators at Swedish football (soccer) matches is between 0.081 and 0.098.

OR

We are 95% confident that between 8.1% and 9.8% of all Swedish football (soccer) spectators are highly intoxicated.

1 pt for an Rguroo output including one of the following rows in the table, or one of the following pairs of R code and output:

Confidence Interval for One Population Proportion

Success = Yes
Sample Size = 4420
Number of Successes = 395
Proportion of Success = 0.08937
Confidence level = 95%

Method	Lower CL	Upper CL	Midpoint	Width
Binomial (Exact)	0.0811168	0.0981634	0.0896401	0.0170466
Large Sample z	0.0809565	0.0977765	0.0893665	0.0168200
Large Sample z with cc	0.0808434	0.0978896	0.0893665	0.0170463

cc: Continuity correction is used in computing the interval.

RStudio code/output 1:

```
> prop.test(395, 4420)

        1-sample proportions test with continuity correction

data:  395 out of 4420, null probability 0.5
X-squared = 2979.6, df = 1, p-value < 2.2e-16
alternative hypothesis: true p is not equal to 0.5
95 percent confidence interval:
 0.08120095 0.09825481
sample estimates:
         p
0.08936652
```

RStudio code/output 2:

```
> prop.test(395, 4420, correct = F)

        1-sample proportions test without continuity correction

data:  395 out of 4420, null probability 0.5
X-squared = 2981.2, df = 1, p-value < 2.2e-16
alternative hypothesis: true p is not equal to 0.5
95 percent confidence interval:
 0.08130918 0.09813700
sample estimates:
         p
0.08936652
```

RStudio code/output 3:

```
> binom.test(395, 4420)

        Exact binomial test

data:  395 and 4420
number of successes = 395, number of trials = 4420, p-value <
2.2e-16
alternative hypothesis: true probability of success is not equal
to 0.5
95 percent confidence interval:
 0.08111678 0.09816339
sample estimates:
probability of success
            0.08936652
```

Problem 4. [5 pts] This problem expands on the gout problem (Problem 1). The table below may be useful. The row variable is the sex of the gout patient and the column variable shows whether the patient was on a urate-lowering treatment.

	Treatment	No Treatment	Total
Female	32	17	49
Male	384	93	477
Total	416	110	526

Test whether the patient's sex affects whether the patient is on a urate-lowering treatment. Use a significance level of $\alpha = 0.05$.

0.5 pts: use a two-proportion z hypothesis test, chi-squared test for independence, or Fisher exact test

1 pt: We have two populations (male and female) and the BINS assumptions are met within each population. We have at least 5 successes (on treatment) and 5 failures (not on treatment) in each population.

OR

We have two categorical variables, each of which has two levels, and all the cases are independent. Therefore, a chi-squared test of independence is appropriate if we expect at least 5 counts in each cell of the table when the hypothesis of independence is true. Clearly, our expected counts should be at least 5.

OR

One of the above explanations without sample size assumption checking, leading to Fisher's exact test.

1.5 pts for one of the following solutions:

H_0: $p_{male} = p_{female}$ OR the two variables are independent/not associated

H_a: $p_{male} \neq p_{female}$ OR the two variables are dependent/associated

Test statistic: one of the following

$z = -2.49$ or 2.49

$\chi^2 = 6.20$ (without continuity correction) or 5.32 (with continuity correction)

Fisher exact test: no test statistic given (we did not cover odds ratios in this class)

p-value: one of the following

0.0127 (Large sample z and Chi-squared test without continuity correction)

0.0211 (Chi-squared test with continuity correction)

0.0167 (Fisher exact test)

1 pt: At the 5% significance level, we reject the null hypothesis. We conclude that there is an association between sex and being placed on a urate-lowering treatment. (Alternatively, we conclude that there is a difference in the population proportion of males and females who are on a urate-lowering treatment.)

1 pt for an Rguroo output including one of the following tables, or one of the following pairs of R code and output:

Test of Hypothesis Treatment
Method: Large Sample z Test (Pooled Standard Error)

Success = Treatment
Population 1 = Female, Population 2 = Male
Sample Size: Female = 49, Male = 477
Number of Successes: Female = 32, Male = 384
Proportion of Success: Female = 0.6531, Male = 0.805
Significance level = 5%
Research Hypothesis H1: Proportion of 'Female - Male' is not equal to 0

Proportion Female	Proportion Male	Difference	Standardized Obs Stat	P-value	95% Lower CL	95% Upper CL
0.653061	0.805031	-0.151970	-2.49095	0.0127401	-0.271545	-0.0323951

Test is significant at 5% level.

Test of Hypothesis Treatment
Method: Chi-Squared Test without continuity correction

Success = Treatment
Population 1 = Female, Population 2 = Male
Sample Size: Female = 49, Male = 477
Number of Successes: Female = 32, Male = 384
Proportion of Success: Female = 0.6531, Male = 0.805
Significance level = 5%
Research Hypothesis H1: Proportion of 'Female - Male' is not equal to 0

Proportion Female	Proportion Male	Difference	Standardized Obs Stat	P-value	95% Lower CL	95% Upper CL
0.653061	0.805031	-0.151970	6.20485	0.0127401	-0.289907	-0.0140331

Test is significant at 5% level.

Test of Hypothesis Treatment
Method: Chi-Squared Test with continuity correction

Success = Treatment
Population 1 = Female, Population 2 = Male
Sample Size: Female = 49, Male = 477
Number of Successes: Female = 32, Male = 384
Proportion of Success: Female = 0.6531, Male = 0.805
Significance level = 5%
Research Hypothesis H1: Proportion of 'Female - Male' is not equal to 0

Proportion Female	Proportion Male	Difference	Standardized Obs Stat	P-value	95% Lower CL	95% Upper CL
0.653061	0.805031	-0.151970	5.32002	0.0210818	-0.301160	-0.00278085

Test is significant at 5% level.

Test of Hypothesis Treatment Method: Fisher Exact Test

Success = Treatment
Population 1 = Female, Population 2 = Male
Sample Size: Female = 49, Male = 477
Number of Successes: Female = 32, Male = 384
Proportion of Success: Female = 0.6531, Male = 0.805
Significance level = 5%
Research Hypothesis H1: Proportion of 'Female - Male' is not equal to 0

Proportion Female	Proportion Male	Difference	P-value
0.653061	0.805031	-0.151970	0.0166837

Test is significant at 5% level.

RStudio code/output 1:

```
> prop.matrix <- matrix(c(32, 17, 384, 93), nrow = 2, ncol = 2,
byrow = T)
> prop.test(prop.matrix)


        2-sample test for equality of proportions with
continuity correction
```

```
data:  prop.matrix
X-squared = 5.32, df = 1, p-value = 0.02108
alternative hypothesis: two.sided
95 percent confidence interval:
 -0.301159595 -0.002780849
sample estimates:
   prop 1    prop 2
0.6530612 0.8050314
```

RStudio code/output 2

```
> prop.matrix <- matrix(c(32, 17, 384, 93), nrow = 2, ncol = 2,
byrow = T)
> prop.test(prop.matrix, correct = F)

        2-sample test for equality of proportions without
continuity correction

data:  prop.matrix
X-squared = 6.2048, df = 1, p-value = 0.01274
alternative hypothesis: two.sided
95 percent confidence interval:
 -0.28990730 -0.01403315
sample estimates:
   prop 1    prop 2
0.6530612 0.8050314
```

RStudio code/output 3:

```
> prop.matrix <- matrix(c(32, 17, 384, 93), nrow = 2, ncol = 2,
byrow = T)
> fisher.test(prop.matrix)

        Fisher's Exact Test for Count Data

data:  prop.matrix
p-value = 0.01668
alternative hypothesis: true odds ratio is not equal to 1
95 percent confidence interval:
 0.2343017 0.9166819
sample estimates:
odds ratio
 0.4566872
```