GISC7310: ADVANCED GIS DATA ANALYTICS

SAMPLE FINAL EXAM FORMAT

Name: ....

I pledge to honor the rules of academic honesty. In particular, I will not communicate information about this exam to students who have not yet taken the exam.

Signature: ....

Notes and Instructions:

1. This exam counts for 34 % of your overall course grade.
2. In total, this exam consists of *xx* points spread over *xx* tasks and *xx* True/False questions.
3. The exam is designed to be complete in 2 hours (but 2 hours 45 minutes are given).
4. Limit the length of your answers to the provided space.
5. This is an open book exam. You can use your text book, your lecture notes (including PDFs on the computer), labs, and a pocket calculator
   But the use of the internet, online dictionaries or any communication resources are ***absolutely prohibited***.

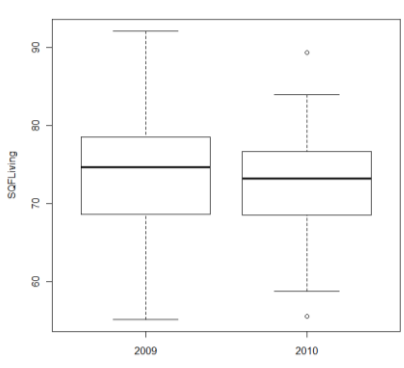
Part 1: Applied Analysis

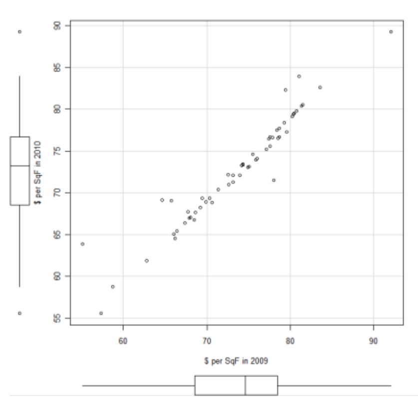
Task 1: Difference in Mean Test [6 points]

Below you find output for several tests for the difference of means. The ***underlying problem*** is that you need to decide whether the appraised home values of 60 homes in "The Crown" neighborhood have ***changed*** from 2009 to 2010. The appraised values are measured by

$$[\$ValuePerSqFoot]_{Year} = [Value\ of\ Improvement\ in\ \$]_{Year}\ /\ [Living\ Area\ in\ ft^2]_{Year}$$

Some descriptive plots may guide you in your decision making process:





Paired *t*-test:

```
> t.test(TheCrown$SQF2009, TheCrown$SQF2010, alternative='two.sided', conf.level=.95, paired=TRUE)

        Paired t-test

data:  TheCrown$SQF2009 and TheCrown$SQF2010
t = 3.4188, df = 59, p-value = 0.001146
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:  0.3589748 1.3722073
sample estimates: mean of the differences 0.865591
```

F-test for equality of variances:

```
> var.test(SQFLiving ~ Year, alternative='two.sided', conf.level=.95)

        F test to compare two variances

data:  SQFLiving by Year
F = 1.1725, num df = 59, denom df = 59, p-value = 0.5431
alternative hypothesis: true ratio of variances is not equal to 1
95 percent confidence interval:  0.7003443 1.9628674
sample estimates: ratio of variances 1.172469
```

Independent *t*-test assuming equal variances

```
> t.test(SQFLiving~Year, alternative='two.sided', conf.level=.95, var.equal=TRUE)

        Two Sample t-test

data:  SQFLiving by Year
t = 0.7263, df = 118, p-value = 0.4691
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval: -1.494594  3.225776
sample estimates: mean in group 2009 mean in group 2010
                            73.52549           72.65990
```

Independent *t*-test assuming unequal variances

```
> t.test(SQFLiving~Year, alternative='two.sided', conf.level=.95, var.equal=FALSE)

        Welch Two Sample t-test

data:  SQFLiving by Year
t = 0.7263, df = 117.261, p-value = 0.4691
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval: -1.494748  3.225930
sample estimates:
mean in group 2009 mean in group 2010
          73.52549           72.65990
```

Task 1: Formulate in ***precise*** statistical terms the ***null*** and the ***alternative*** hypotheses. Make sure that you specify either directed or undirected hypotheses base on the above research question. [2 points]

Task 2: ***Justify*** whether you use a matched pairs or an independent sample *t*-test? [2 point]

Task 3: Based on the *p*-value of your selected test procedure, ***draw your conclusions*** about the differences in the appraisal values for 2009 and 2010 at an error probability of $\alpha = 0.01$. Use statistical terminology in your answer. [2 points]

Part 3 True and False Questions with Rationale (20 points)

*Mark either the **true** or **false** box to indicate whether the following statements are true or false (1 point). Briefly justify your choice (1 point).*

1. [TRUE] [FALSE]: The relationship between the Wald, likelihood ratio and the score tests is $W \leq LR \leq S$.

   ....

2. [TRUE] [FALSE]: Let $\boldsymbol{y}_1$ and $\boldsymbol{y}_2$ be two vectors, which each holds counts. The rate of $\boldsymbol{y}_1$ in dependence of $\boldsymbol{x}$ can be estimated in logistic regression with `glm(cbind(y1,y2)~x, family=binomial(logit))`.

   ....