544 Stats Final, Fall 2006

Lada Adamic
SI 544
Dec 12, 2006

This is your final exam. Turn in all R code, label your plots, and put in a sentence or two highlighting the results of each step of your analysis. Mention where appropriate things like p-values, what your null-hypothesis is, and whether you reject it or accept it. For each statistical method you choose, briefly justify your selection and comment on whether the assumptions needed in order to be able to use the method are met. This exam is open book/lecture notes/solution sets/web. You are to work on this without the assistance of others. Good luck!

1 Tennis balls

A. (15 pts) The box says that the balls should weigh 57.5 grams with a standard deviation of 0.5 grams. You have a high precision digital scale handy and have a few minutes to kill before your next class, so you decide to weigh 10 balls to see if they meet specifications. These weights are saved in the file tennisballweights.txt. What is your null hypothesis? Can you reject it?

We expect that this is a job for the t-test, since the weight data looks pretty normally distributed:

```
> ballweights = read.table("tennisballweights.txt",head=T)
> qqnorm(ballweights$weight)
> ballweights$weight
     V1
1  58.31 2  57.69 3  57.41 4  57.50 5  57.66 6  57.86 7  58.16 8
57.78 9  58.44 10 57.03

> t.test(ballweights,mu=57.5)

        One Sample t-test

data:  ballweights t = 2.0865, df = 9, p-value = 0.06656 alternative
hypothesis: true mean is not equal to 57.5 95 percent confidence
interval:
 57.47609 58.09191
sample estimates: mean of x
   57.784
```

The qqnorm plot tells us that we are OK to use the t-test, but the t-test turns up being not significant at that the $\alpha = 0.05$ level (p=0.07). So we can't say for certain if these balls differ in weight.

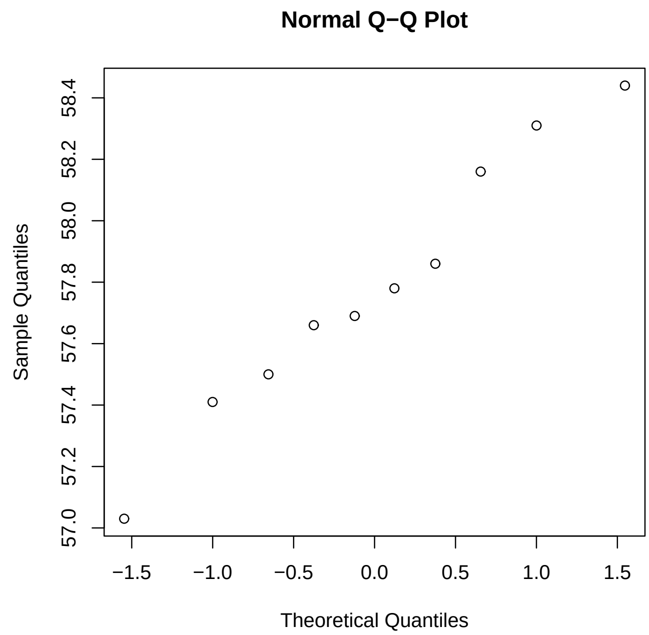

Figure 1: QQ norm plot of ball weights to test that the sample weights are normally distributed.

B. (10 pts) If the balls are actually 0.5 grams heavier on average, how many balls would you need to weigh in order to detect the difference at the $\alpha = 0.05$ *level with power* $\beta = 0.80$. *Comment on the result with respect to your analysis in 1A, assuming that the sample in actuality was drawn from a box of balls that were 0.5 grams heavier on average.*

```
> power.t.test(delta=0.5,sd=0.5,sig.level=0.05,power=0.8,type="one.sample")

     One-sample t test power calculation

              n = 9.937864
          delta = 0.5
             sd = 0.5
      sig.level = 0.05
          power = 0.8
    alternative = two.sided
```

The power test tells us that already 10 balls should have been enough to detect a difference 80% of the time. However, our sample happened to have a lower mean (57.78 grams) than the box as a whole, so we fell in the 20% of cases where we could not reject the null hypothesis that the true mean was 57.5 grams, even though that hypothesis was false.

C. (30 pts) You have three classes to teach that day: beginner, intermediate, and master. You decide to test how well your students do with the "magic" tennis balls, so you design an experiment. In each class of 24, you randomly assign 12 people to practice their serves on one court with the magic tennis balls and the other 12 practice on another court with regular tennis balls. You record *two things about each student's serves: how fast they are on average, and how often the student succeeded in hitting a cone placed on the court. The file tennisdata.txt has the data, with 4 columns, each row corresponding to one student:* ***classLevel***, ***typeOfBall***, ***avServeSpeed***, ***numConesHit****. Do an analysis of variance to figure out how the average serve speed depends on the class level and the type of ball. Do the same for the accuracy of the serve (the number of times the cone was hit). Draw the corresponding interaction plots. Are the magic tennis balls really magic?*

This is a randomized experiment, with the same number of people in each bin (determined by both class level and kind of ball used). The observations are also independent, since the performance of one student is assumed not to influence the performance of the next. We will therefore proceed with doing an anova.

```
> summary(tennisdata)
        classLevel   typeOfBall  avServeSpeed     numConesHit
 beginner    :24   magic  :36   Min.   :  3.16   Min.   : 7.00
 intermediate:24   regular:36   1st Qu.: 35.46   1st Qu.:14.00
 master      :24                Median : 52.38   Median :17.00
                                Mean   : 53.64   Mean   :17.24
                                3rd Qu.: 66.63   3rd Qu.:21.00
                                Max.   :100.66   Max.   :29.00

> anova(lm(avServeSpeed~classLevel*typeOfBall,data=tennisdata))
Analysis of Variance Table

Response: avServeSpeed
                      Df  Sum Sq Mean Sq F value    Pr(>F)
classLevel             2 15158.9  7579.4 24.1500 1.347e-08 ***
typeOfBall             1   315.5   315.5  1.0052    0.3197
classLevel:typeOfBall  2    21.5    10.8  0.0343    0.9663
Residuals             66 20714.0   313.8
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

> anova(lm(numConesHit~classLevel*typeOfBall,data=tennisdata))
Analysis of Variance Table

Response: numConesHit
                      Df Sum Sq Mean Sq F value    Pr(>F)
classLevel             2 483.44  241.72 18.7672 3.525e-07 ***
typeOfBall             1  23.35   23.35  1.8127   0.18279
classLevel:typeOfBall  2 108.11   54.06  4.1968   0.01924 *
Residuals             66 850.08   12.88
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

# interaction plot for serve speed
> interaction.plot(classLevel,typeOfBall,avServeSpeed,ylab="average serve speed",
+ xlab="class level",trace.label="type of ball",cex.axis=1.5,cex.lab=1.5)

# interaction plot for accuracy
> interaction.plot(classLevel,typeOfBall,numConesHit,ylab="average serve speed",
+ xlab="class level",trace.label="type of ball",cex.axis=1.5,cex.lab=1.5)
```

There are several null hypotheses here - that the class level and type of ball individually do not affect speed (and likewise accuracy) of students' serves, and also that there is no interaction term. What we observe is that the type of ball does not seem to affect the speed of the serve ($p > 0.1$), but naturally, the class level does ($p < 10^{-7}$) - the more advanced students have faster serves. For accuracy it's a different story. While class level does influence the accuracy of the serve ($p < 10^{-6}$), the type of ball interacts with the class level ($p < 0.05$). From the interaction plot, we observe that beginning students see the highest improvement in accuracy with the magic tennis balls, while the masters level students actually do worse. The advantages and disadvantages average out over class level, leaving no significant effect overall for the type of ball ($p > 0.1$) but a significant interaction term. The magic tennis balls really are different (if not outright magic) - but it's good magic for beginners as opposed to an impediment for the advanced students.

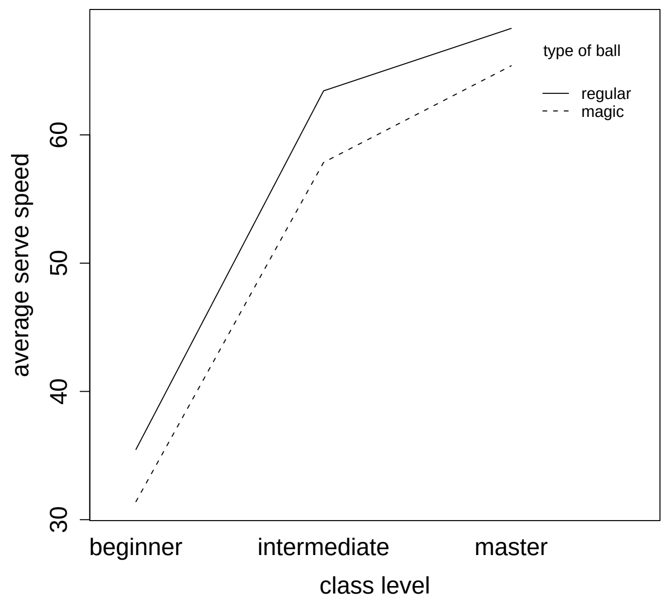

Figure 2: Interaction plot of serve speed depending on class level and type of ball.

D. (20 pts) You'd like to quantify the relationship between the average speed of a student's serve and their accuracy in hitting a cone. Using only the scores of those students who were using the regular balls, fit a linear regression. Plot the data as well as your regression line. Is there a relationship between the two variables?

```
#select just serves with regular tennis balls
> regulardata = tennisdata[typeOfBall=="regular",]
> detach(tennisdata)
> attach(regulardata)

> plot(numConesHit,avServeSpeed,cex.axis=1.5,cex.lab=1.5,
+ xlab="number of times a cone was hit",
+ ylab="average serve speed")
> lm.tennis = lm(avServeSpeed~numConesHit)
> abline(lm.tennis)
> summary(lm.tennis)

Call: lm(formula = avServeSpeed ~ numConesHit)

Residuals:
    Min      1Q  Median      3Q     Max
```

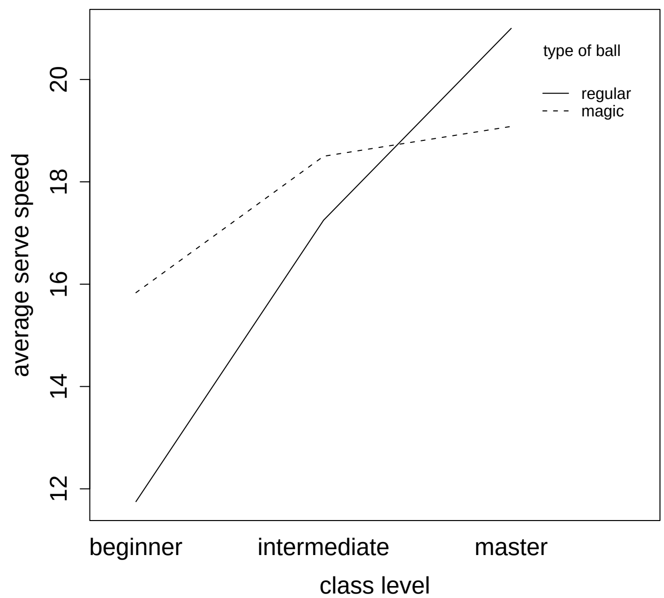

Figure 3: Interaction plot of serve accuracy depending on class level and type of ball.

```
-26.478  -7.943   2.468   8.150  21.260

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  -4.8958     6.7029  -0.730     0.47 numConesHit
3.6380     0.3833   9.491 4.37e-11 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 12.17 on 34 degrees of freedom Multiple
R-Squared: 0.726,      Adjusted R-squared: 0.7179 F-statistic: 90.07
on 1 and 34 DF,  p-value: 4.372e-11
```

The observations should be independent (and we expect them to be since each observation corresponds to a different student), and the residuals appear approximately normally distributed, so we are OK using a linear regression. The linear regression tells us that the two variables are quite correlated (adjusted R-squared is 0.72 and the p-value is $< 10^{-10}$), with faster servers also being more accurate. The regression line shows the best linear fit of serve speed vs. accuracy.

Extra credit(4pts) Calculate a prediction interval, such that if a student hits the cone 20 out of 40 times, you can predict with 95% confidence that their average serve speed will be within this interval.

```
> predict(lm(avServeSpeed~numConesHit),int="p",data.frame(numConesHit=20))
          fit     lwr      upr
[1,] 67.86383 42.6561 93.07157
```

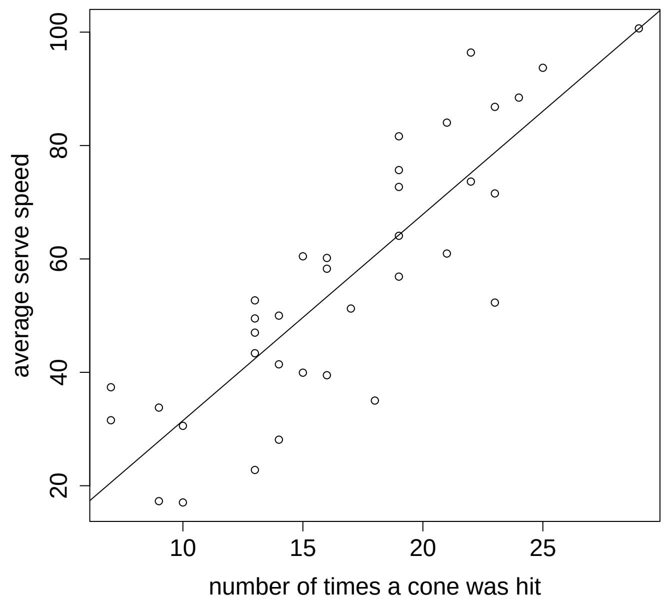

Figure 4: Linear regression of average serve speed vs. accuracy for regular tennis balls.

The 95% prediction interval corresponding to this data point is given as [42.66,93.07], meaning that a person with the given accuracy will with 95% probability have a serve speed somewhere between 43 and 93 mph. If you had used all the data to make the prediction, you would have gotten something like 39 to 91 mph (this is OK since I didn't specify which to use).

2 Money can't buy you happiness, or can it?

*(25 pts) The Pew Spyware survey asked the respondents about their income, reported in the column **INC** of the Spyware.sav data file (remember, it is in SPSS format). You would like to know if people who are better off financially are happier, or at least happier with the state of the country. You remember that the very first question, **Q1** was: "Overall, are you satisfied or dissatisfied with the way things are going in this country today?". Use both a plot and a statistical test to discuss the relationship between the two variables.*

```
> spyware = read.spss("pew/Spyware.sav")
> incandsat = table(spyware$INC,spyware$Q1)
> incandsat

                                  Satisfied Dissatisfied Don't know/Refused
  Less than $10,000                      34           87                 13
  $10,000 to under $20,000               31          114                 15
  $20,000 to under $30,000               66          126                 23
  $30,000 to under $40,000               76          120                 17
  $40,000 to under $50,000               59           82                 13
```

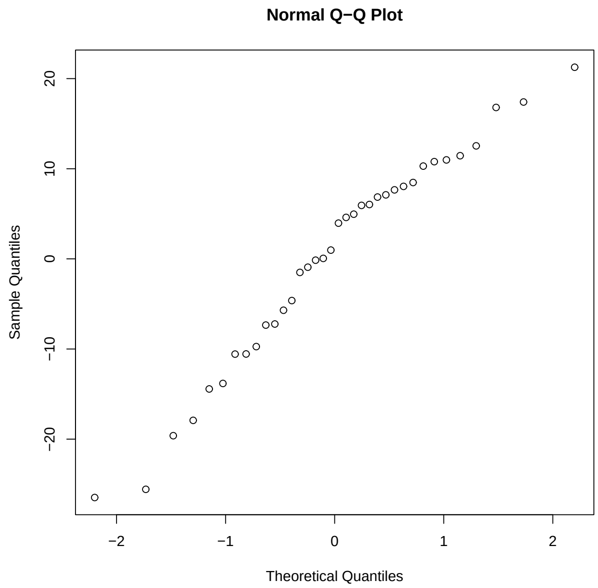

Figure 5: QQ plot of residuals to linear regression

```
  $50,000 to under $75,000               107          130                   18
  $75,000 to under $100,000               88           98                   12
  $100,000 or more                        99           91                   11
  (DO NOT READ) Don't know/Refused       154          240                   77
```

We have tabular data with categorical variables - looks like a job for a χ^2 test. All the column counts are high enough to make using a χ^2 test OK, but I would like to omit the non-responses, so as to just test the relationship between income and satisfaction with the country, and not whether people were willing to answer the questions.

```
> answeronly = incandsat[1:8,1:2]
> answeronly

                            Satisfied Dissatisfied
  Less than $10,000                34           87
  $10,000 to under $20,000         31          114
  $20,000 to under $30,000         66          126
  $30,000 to under $40,000         76          120
  $40,000 to under $50,000         59           82
  $50,000 to under $75,000        107          130
  $75,000 to under $100,000        88           98
  $100,000 or more                 99           91
```

Now we're ready to do a χ^2 test:

```
> chisq.test(answeronly)

        Pearson's Chi-squared test

data:  answeronly
X-squared = 49.3675, df = 7, p-value = 1.922e-08
```

The χ^2 is high enough, and the p-value is definitely significant ($p < 10^{-7}$) at the 99% percent level, so we can reject the null hypothesis that there is no effect of income on satisfaction. Better off people are more satisfied. We can also observe this from a barplot, where the number of respondents who are satisfied and have low income is less than those whose income is higher.

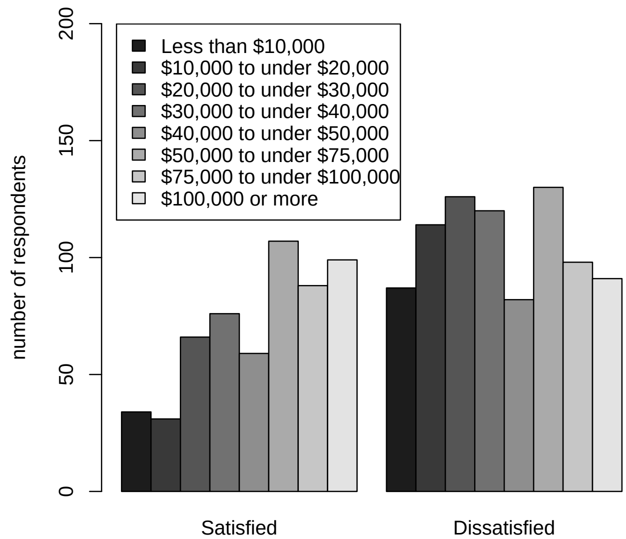

Figure 6: Bar plot of people's satisfaction with the state of the country vs. their income level.

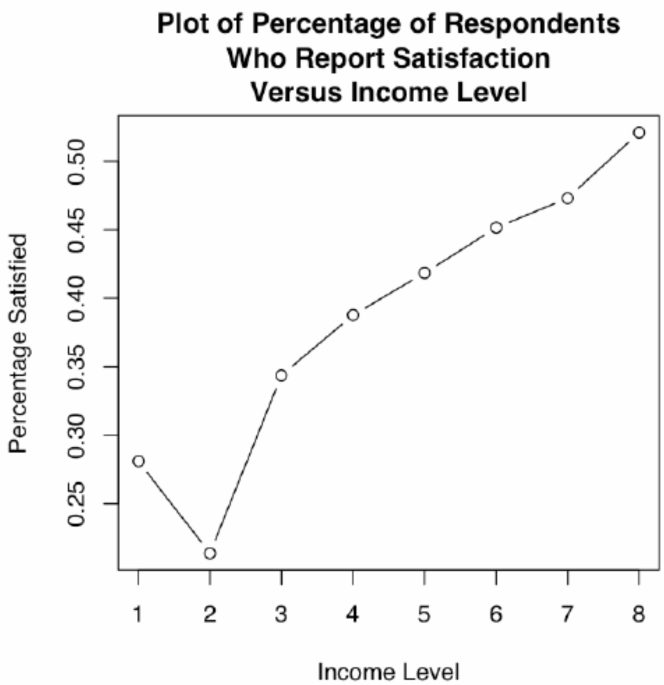

Figure 7: Perry Wong's neat graphical view of the data.