# Review problems

Modules 7-12

Stat 251

updated 2021

(1) Ithaca, NY is located in upstate New York and averages around 37" of rain each year, with standard deviation of 3.25". Rainfall in Ithaca, NY tends to follow an approximately normal distribution. Say that a student attends Conrnell University in Ithaca and is there for 4 years.
   (a) Define the Central Limit Theorem.
   (b) Describe the sampling distribution of the sample mean of rainfall in Ithaca, NY. Include the mean of the sampling distribution of the sample mean and the standard deviation of the sampling distribution of the sample mean (standard error).
   (c) What is the probability that during the 4 years, we see an average less than 32"?
   (d) What are the wettest 7% of years?

(2) Ithaca, NY is located in upstate New York and averages around 37" of rain each year, with standard deviation of 3.25". Rainfall in Ithaca, NY tends to follow an approximately normal distribution. Say that a student attends Conrnell University in Ithaca and is there for 7 years (undergraduate and graduate degrees).
   (a) Describe the sampling distribution of the total rainfall in Ithaca, NY. Include the total of the sampling distribution and the standard deviation of the sampling distribution of the total (standard error)
   (b) What is the probability that during the 7 years, we see a total precipitation of less than 220"?
   (c) What is the $93^{rd}$ percentile for total precipitation?

(3) A survey of purchasing agents from 250 randomly selected industrial companies found that 25% of the buyers reported higher levels of new orders in January than in earlier months.
   (a) Describe the sampling distribution of the proportion of buyers in the US with higher levels of new orders in January. Include the mean of the sampling distribution and the standard deviation of the sampling distribution (standard error).
   (b) What is the probability that the sample proportion is more than 26%?
   (c) What is the probability that the sample proportion is less than 20%?

(4) A company that manufactures coffee for use in commercial machines monitors the caffeine content in its coffee. The company randomly selected 50 samples of coffee every hour from its production line and determines the caffeine content. From historical data, the caffeine content is known to to be approximately normal. The caffeine content should usually be 107 mg.
   (a) Is there sufficient evidence that the mean caffeine content is more than the usual amount?
   (b) What kind of error could have been made? Define the error and explain it in the context of the scenario.
   (c) Estimate $\mu$, the true average caffeine content of the coffee, with 95% confidence. Interpret.
   (d) Suppose that the company would prefer a bound for the next batch to be 1.15. What sample size would be needed to get a bound of 1.15, while maintaining 95% confidence?

```
t.test(coffee,mu=107,alternative='g')
```

```
    One Sample t-test
```

```
data:  coffee
t = 4.375, df = 49, p-value = 3.17e-05
alternative hypothesis: true mean is greater than 107
95 percent confidence interval:
 109.8615      Inf
sample estimates:
mean of x
 111.6393
```

```
t.test(coffee)
```

```
    One Sample t-test

data:  coffee
t = 105.28, df = 49, p-value < 2.2e-16
alternative hypothesis: true mean is not equal to 0
95 percent confidence interval:
 109.5083 113.7704
sample estimates:
mean of x
 111.6393
```

(5) It is thought that more than 70% of all faults in transmission lines are caused by lightning. In a random sample of 200 faults from a large data base, 151 are due to lightning.
   (a) Is there sufficient evidence that the proportion of faults in transmission due to lightning strikes is different from 70%? Conduct a hypothesis test.
   (b) What kind of error could have been made? Define the error and explain it in the context of the scenario.
   (c) Estimate $p$, the true proportion of faults in transmission due to lightning strikes, with 95% confidence. Interpret.
   (d) Suppose that next sample should have a bound of 3%. What sample size would be needed to get a bound of 3%, while maintaining 95% confidence?

```
t.test(lightning,mu=.7)
```

```
    One Sample t-test

data:  lightning
t = 1.804, df = 199, p-value = 0.07275
alternative hypothesis: true mean is not equal to 0.7
95 percent confidence interval:
 0.6948788 0.8151212
sample estimates:
mean of x
    0.755
```

(6) In 1882 Michelson measured the speed of light (usually denoted as $c$ in Einstein's equation $E = mc^2$). He reported the results of 23 random trials with a mean of 29756.22 $km/sec$ and standard deviation of 107.12 $km/sec$. [Note that the actual speed of light is actually 299,792,458 metres per second but remember, these results are from experiments done in *1882*; use the units of measurement that Michelson used ($km/sec$).]
   (a) Suppose previous experiments of Michelson found that the speed of light was 29750 $km/sec$. Is there sufficient evidence from his experiments that the speed of light is significantly different from

the previous result of 29750? Let $\alpha = 0.02$.

(b) What kind of error could have been made? Define the error and explain it in the context of the scenario.

(c) Estimate $\mu$, the true speed of light with 98% confidence. Interpret.

```
t.test(make.it.so,mu=29750,conf.level=.98)
```

```
    One Sample t-test

data:  make.it.so
t = 1.3576, df = 22, p-value = 0.1884
alternative hypothesis: true mean is not equal to 29750
98 percent confidence interval:
 29722.12 29843.66
sample estimates:
mean of x
 29782.89
```

(7) *pvalues*: Find the *pvalue* of the following scenarios and state whether or not the null hypothesis should be rejected for each one

(a) $H_0 : \mu = 5$ vs. $H_a : \mu < 5$, $z_{calc} = -1.55$

(b) $H_0 : \mu = 5$ vs. $H_a : \mu > 5$, $z_{calc} = 1.55$, $\alpha = 0.10$

(c) $H_0 : \mu = 5$ vs. $H_a : \mu \neq 5$, $z_{calc} = -1.55$

(8) *Difference of 2 independent means test and CI* $\mu_1 - \mu_2$: Researchers speculate that drivers who do not wear a seatbelt are more likely to speed than drivers who do wear one. A random sample of 40 drivers was taken. In the experiment, the people were clocked to see how fast they were driving (mph) and then were stopped to see whether or not they were wearing a seatbelt.

(a) Is there sufficient evidence that the average speed for non-seatbelt wearers differs from those drivers that do wear a seatbelt? Let $\alpha = 0.10$

(b) Estimate the true difference in means of the speeds of drivers who do not wear seatbelts as compared to those who wear seatbelts with 90% confidence and interpret

(c) State the kind of error that could have been made. **DESCRIBE IT IN CONTEXT OF THE DATA**

```
t.test(mph~belt,conf.level=.9)
```

```
    Welch Two Sample t-test

data:  mph by belt
t = 2.5321, df = 37.524, p-value = 0.01565
alternative hypothesis: true difference in means is not equal to 0
90 percent confidence interval:
  2.053407 10.244202
sample estimates:
 mean in group No mean in group Yes
          73.0519           66.9031
```

(9) *Paired t-test and CI (dependent samples)* $\mu_d$: Many freeways have service (or logo) signs that give information on attractions, camping, lodging, food, and gas services prior to off-ramps. An article reported that in one investigation, six sites along Virginia interstate highways where service signs are posted were selected randomly. For each site, crash data was obtained for a three-year period before distance information was added to the service signs for a one-year period afterward.

(a) Is there sufficient evidence that there is a decrease in accidents after the signs added the distance

information? Let $\alpha = 0.01$

(b) Estimate the true mean difference in accidents before and after the signage change with 99% confidence and interpret

(c) State the kind of error that could have been made. **DESCRIBE IT IN CONTEXT OF THE DATA**

```
t.test(before,after,paired=T,conf.level=.99,alternative='l')
```

```
    Paired t-test

data:  before and after
t = -0.15141, df = 5, p-value = 0.4428
alternative hypothesis: true difference in means is less than 0
99 percent confidence interval:
     -Inf 68.12903
sample estimates:
mean of the differences
              -3.209914
```

```
t.test(before,after,paired=T,conf.level=.99)
```

```
    Paired t-test

data:  before and after
t = -0.15141, df = 5, p-value = 0.8856
alternative hypothesis: true difference in means is not equal to 0
99 percent confidence interval:
 -88.69426  82.27443
sample estimates:
mean of the differences
              -3.209914
```

(10) *Difference of two proportions test and CI*: A hospital administrator suspects that the delinquency rate in the payment of hospital bills has increased over the past year. Hospital records show that the bills of 48 of 1284 persons admitted in the month of April have been delinquent for more than 90 days. This number compares to 34 of 1002 persons admitted during the same month one year ago.

(a) Is there sufficient evidence that there is an increase in delinquency rate in the payment of hospital bills over the last year?

(b) Estimate the true difference of proportions of delinquent bills over the last year with 95% confidence and interpret

(c) State the kind of error that could have been made. **DESCRIBE IT IN CONTEXT OF THE DATA**

```
t.test(current,past,alternative='l')
```

```
    Welch Two Sample t-test

data:  current and past
t = 0.4426, df = 2194.2, p-value = 0.671
alternative hypothesis: true difference in means is less than 0
95 percent confidence interval:
       -Inf 0.01628168
sample estimates:
 mean of x  mean of y
0.03738318 0.03393214
```

```
t.test(current,past)
```

```
    Welch Two Sample t-test

data:  current and past
t = 0.4426, df = 2194.2, p-value = 0.6581
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 -0.01183959  0.01874168
sample estimates:
 mean of x  mean of y
0.03738318 0.03393214
```

(11) *Book Mediums*: (not the psychic kind of medium) A professor of an introductory college class uses an open-source textbook for the class. Of interest is the proportions of students that will either purchase a hard copy, print the book online, or just use the downloaded PDF format to read on a device. From earlier semesters, 60% bought a hard copy of the book, 25% printed it online, and 15% used a downloaded PDF format on their devices. At the end of the semester, the professor asks the students to complete a survey and indicate what format of the book they used. Of the 126 students, 71 bought a hard copy, 30 printed it, and 25 downloaded PDF to use.

(a) Is there evidence that the students used similar mediums for the book?

(b) State the kind of error that could have been made; describe in context.

```
read.it
```

```
       type counts probs
1 Hard copy     71  0.60
2   Printed     30  0.25
3       PDF     25  0.15
```

```
chisq.test(counts,p=probs)$expected
```

```
[1] 75.6 31.5 18.9
```

```
chisq.test(counts,p=probs)
```

```
    Chi-squared test for given probabilities

data:  counts
X-squared = 2.3201, df = 2, p-value = 0.3135
```

(12) *Independence Test* $\chi^2$: In Star Trek fandom, there is a running joke that characters on the show who wear a red shirt are doomed, just another statistic. Shirt colors can be only blue, gold, or red; fatalities can be only dead or alive.

(a) Is there sufficient evidence determine whether there is an association between shirt color and deaths?

(b) State the kind of error that could have been made. **DESCRIBE IT IN CONTEXT OF THE DATA**

```
star.trek; cat('Star Trek survival by shirt colour')
```

```
        Shirt.Colour
Survival Blue Red Gold Total
   Alive  129 215   46   390
```

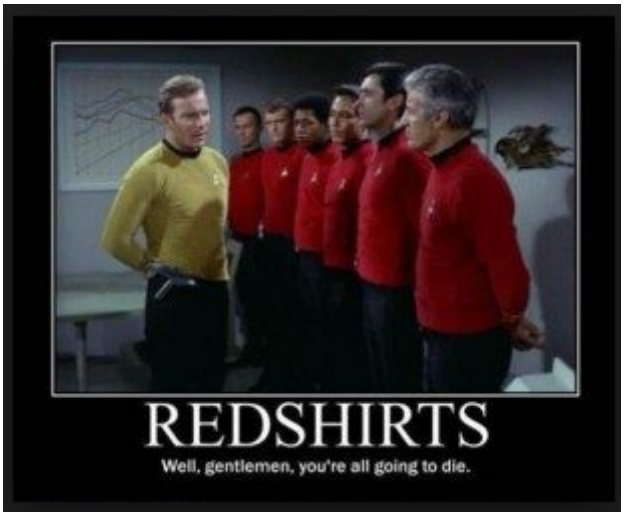



Figure 1: Red Shirt of Doooom

```
   Dead     7  24    9    40
   Total  136 239   55   430
```

```
Star Trek survival by shirt colour
```

```
chisq.test(spock)$expected
```

```
        Shirt.Colour
Survival      Blue       Red      Gold
   Alive 123.34884 216.76744 49.883721
   Dead   12.65116  22.23256  5.116279
```

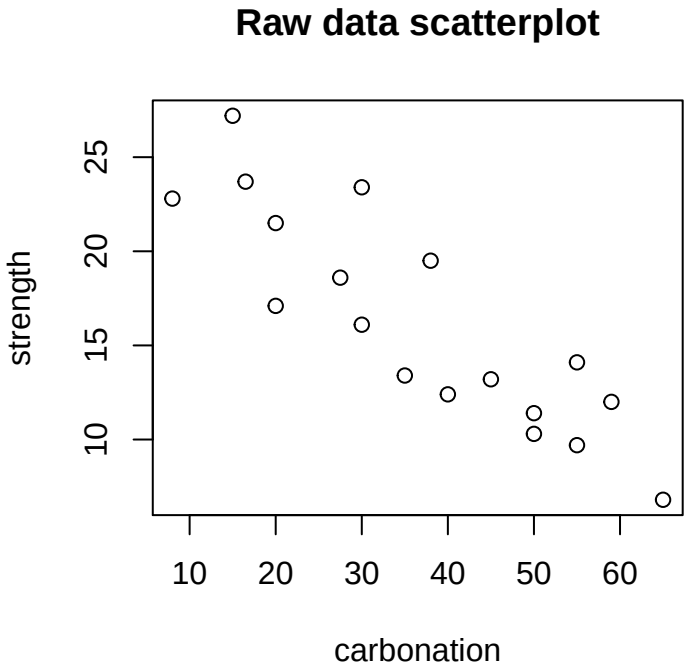
```
chisq.test(spock)
```

```
    Pearson's Chi-squared test

data:  spock
X-squared = 6.1886, df = 2, p-value = 0.04531
```

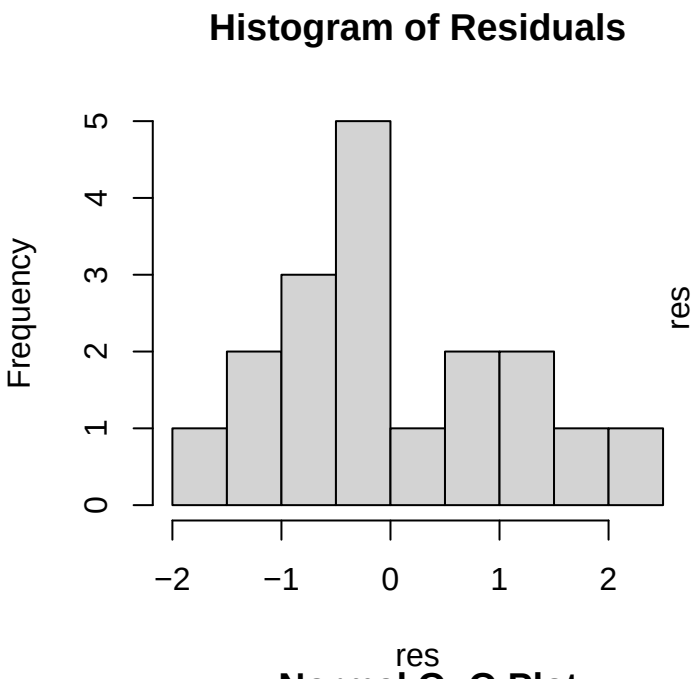
(13) *SLR (simple linear regression)*: Corrosion of steel reinforcing bars is the most important durability problem for reinforcing structures[1]. Carbonation of concrete results form a chemical reaction that lowers the pH value by enough to initiate corrosion of the rebar. Data on the carbonation depth ($mm$) and strength ($MPa$) for a sample of core specimens was taken from a particular building, and all the regression output is provided. We are interested in modeling the strength
   (a) State the (population) model equation and define its components
   (b) Looking at the raw data scatterplot, does it appear as if there is a linear relationship? Positive or negative slope?
   (c) State the regression equation using the provided output. Using the regression equation, estimate the strength when the carbonation depth is 8 $mm$ and estimate it again when the depth is 20 $mm$
   (d) Calculate the residuals for both of your estimates in part c. The observed value for 8 $mm$ is 22.8 $MPa$ ($(8, 22.8)$) and for 20 $mm$ is 17.1 $MPa$ ($(20, 17.1)$)
   (e) Interpret slope and intercept *in context* of the data. If something does not make sense in context, state it and describe why

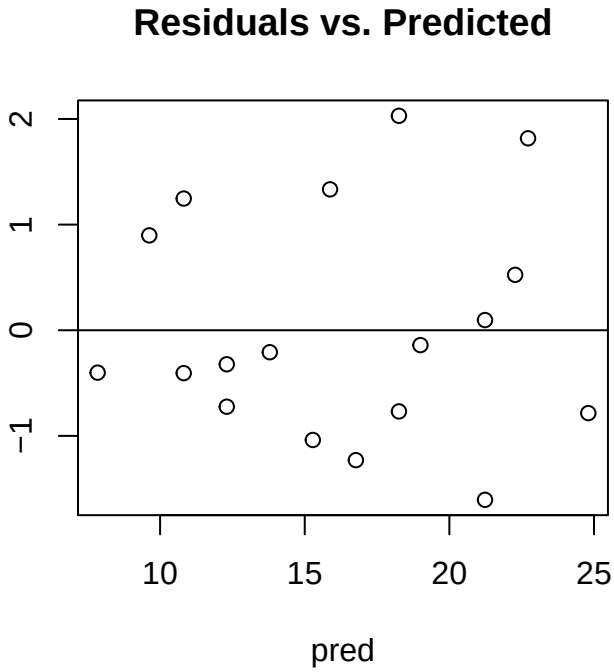
[1] "The Carbonation of Conrete Structures in the Tropical Environment of Singapore" (*Magazine of Concrete Research*, 1996: 293-300)

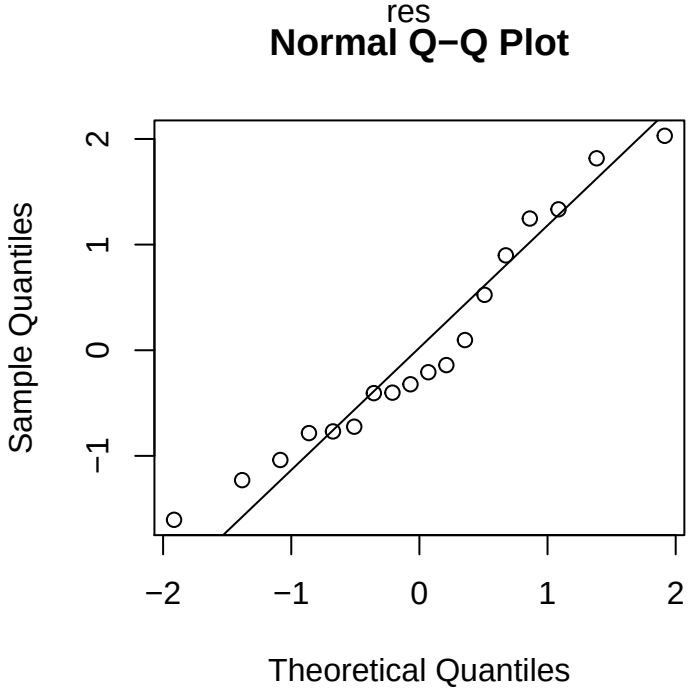
(f) What is the coefficient of determination, $R^2$? List the value and interpret in context
(g) What is the correlation, $r$? List the value and interpret in context
(h) Is the slope significant? Conduct hypothesis test; include hypotheses, test statistic, $df$, $pvalue$, results, and conclusion
(i) List assumptions of regression (words or symbols) and check assumptions (assumptions 1, 2, and 4). Briefly describe how they are/are not met
(j) Using parts b, f, g, h, and i, is this a good model? Reference those to verify your claim (briefly describe)




```
Call:
lm(formula = strength ~ carbonation)

Residuals:
    Min      1Q  Median      3Q     Max
-4.1317 -2.0043 -0.7488  2.1366  5.1439

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 27.18294    1.65135  16.461 1.88e-11 ***
carbonation -0.29756    0.04116  -7.229 2.01e-06 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 2.864 on 16 degrees of freedom
Multiple R-squared:  0.7656,    Adjusted R-squared:  0.7509
F-statistic: 52.25 on 1 and 16 DF,  p-value: 2.013e-06
```


Histogram of Residuals
Frequency
0
1
2
3
4
5
−2
−1
0
1
2
res


Residuals vs. Predicted
res
−1
0
1
2
10
15
20
25
pred


Normal Q−Q Plot
Sample Quantiles
−1
0
1
2
−2
−1
0
1
2
Theoretical Quantiles