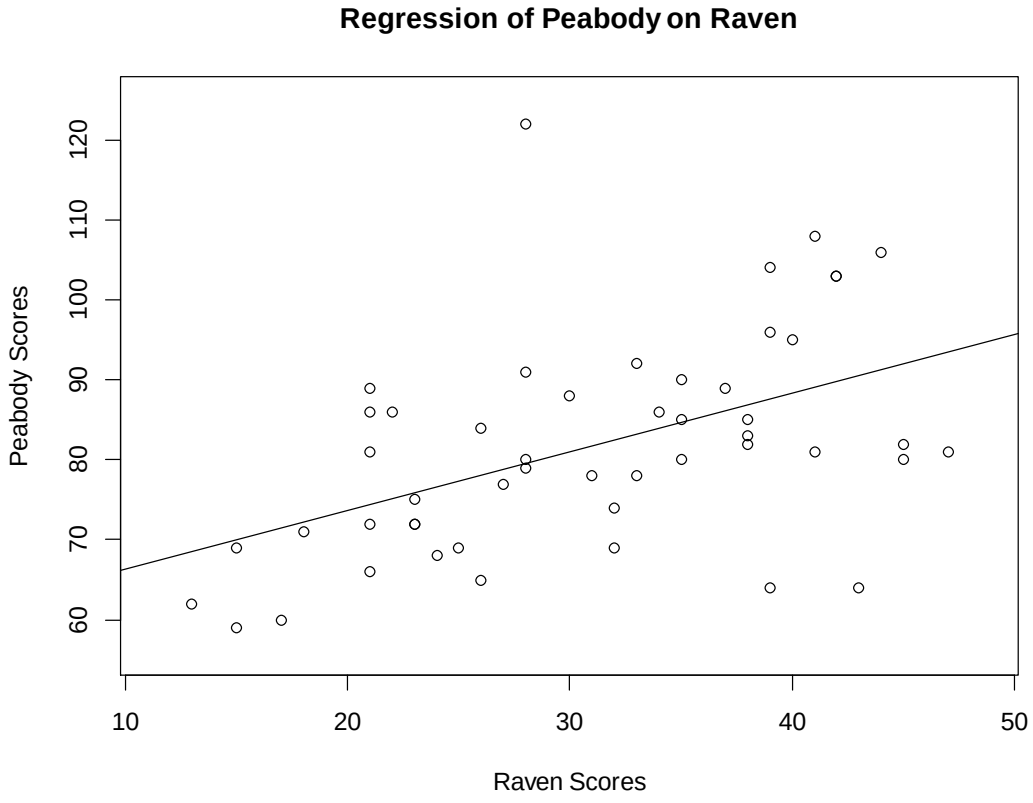
Sample of Homework Three

Jack L. Vevea

Part One: Linear regression

1. Here is a good scatterplot showing Peabody conditioned on Raven for my sample:



I produced that plot using the following R commands:

```
> attach(JackStatlab)
> RAinc <- .05* (max(CTRA)-min(CTRA))
> PBinc <- .05*(max(CTPEA)-min(CTPEA))
> par(pin=c(6,4))
> plot(CTRA, CTPEA, main="Regression of Peabody on Raven",
+      xlab="Raven Scores", ylab="Peabody Scores",
+      xlim=c(min(CTRA)-RAinc,max(CTRA)+RAinc),
+      ylim=c(min(CTPEA)-PBinc,max(CTPEA)+PBinc))
> abline(lm(CTPEA~CTRA)$coef)
```

2. The estimated linear regression equation has an intercept of 59.049 and a slope of 0.732. I determined that by entering "`lm(CTPEA~CTRA)`" in *R*.
3. (The regression line has already been added to the plot.)
4. The slope of 0.732 indicates that, according to the model, the conditional mean of Peabody increases by 0.732 points for every 1-point increase in Raven.
5. I conducted the hypothesis test about the slope in two ways: first, via the *F* statistic and second via the *t* statistic. (I know that these two are equivalent, but wanted to verify that result for myself.) For the *F* statistic, I first saved the regression output: "`regout <- lm(CTPEA~CTRA)`". Then I entered "`anova(regout)`", which produced the following output:

```
Analysis of Variance Table

Response: CTPEA
          Df Sum Sq Mean Sq F value    Pr(>F)
CTRA       1 2189.0 2189.00  15.541 0.0002616 ***
Residuals 48 6760.8  140.85
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

The *p* value associated with the *F(1, 48)* statistic is .0002616, which is lower than my criterion of .05, so I reject the null hypothesis and conclude that the slope is *not* equal to zero.

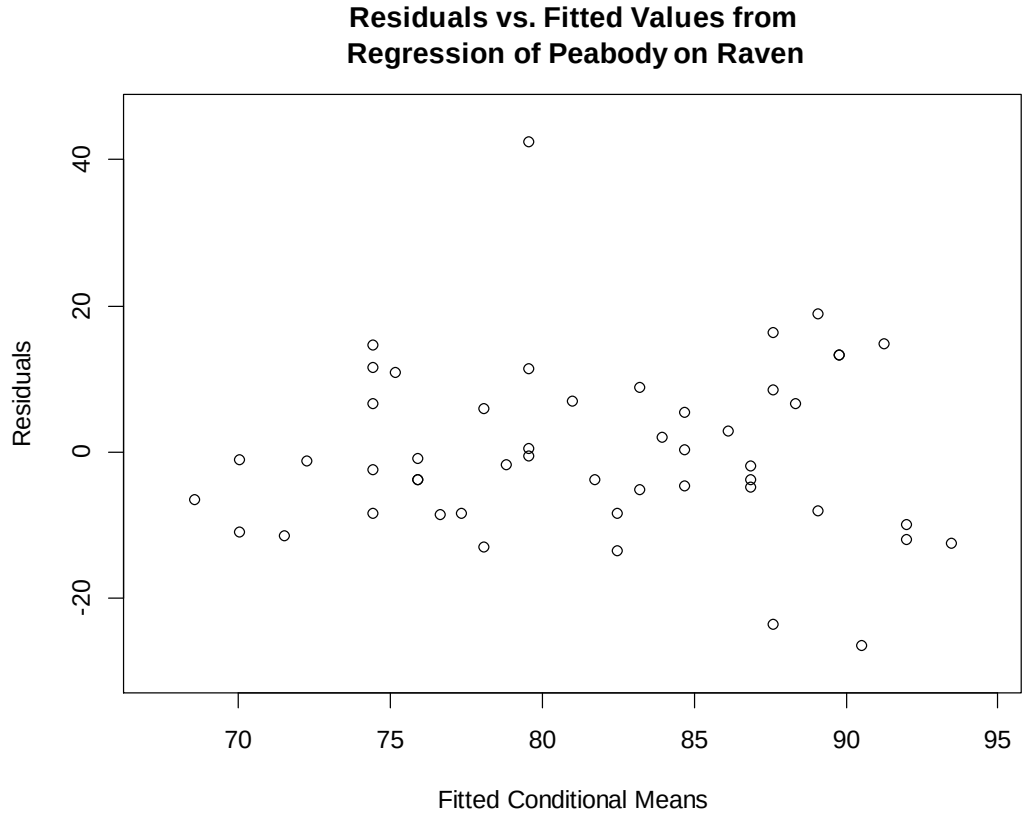
I also used the following code to conduct the test using a *t* statistic: "`summary(regout)`", which produced the following output:

```
Call:
lm(formula = CTPEA ~ CTRA)

Residuals:
    Min      1Q  Median      3Q     Max
-26.520  -8.273  -1.516   6.915  42.459

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  59.0490     5.9663   9.897 3.55e-13 ***
CTRA          0.7319     0.1856   3.942 0.000262 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 11.87 on 48 degrees of freedom
Multiple R-squared:  0.2446,    Adjusted R-squared:  0.2288
F-statistic: 15.54 on 1 and 48 DF,  p-value: 0.0002616
```

The *t(48)* statistic in that output has a *p* value of .000262 (which is the same as the result from the *F* test, rounded to six decimal places). Hence, the conclusion about the slope is the same, as it must be because the tests are exactly equivalent.

6. The assumptions that must be satisfied for those tests to be valid are:
    a. The relationship must be linear.
    b. The errors must be independent.
    c. The vertical variability of errors about the regression line should be the same through the range of fitted values.
    d. The errors must be normally distributed.

7. I assessed the linearity assumption first by examining the scatterplot that I produced earlier. It does seem plausible that the relationship is linear. I also produced the following plot of residuals against fitted values:



That plot was produced using the following *R* code:

```
> resxinc <- .05 * (max(regout$fit)-min(regout$fit))
> resyinc <- .05 * (max(regout$res)-min(regout$res))
> plot(regout$fit, regout$res, main="Residuals vs. Fitted Values
from\n Regression of Peabody on Raven",
+     xlab="Fitted Conditional Means", ylab="Residuals",
+     xlim=c(min(regout$fit)-resxinc,max(regout$fit)+resxinc),
+     ylim=c(min(regout$res)-resyinc,max(regout$res)+resyinc))
```

I see no evidence of a curvilinear relationship in the residuals plot, so by this criterion it also appears that the relationship is linear.

The independence of the errors amounts to the same thing as arguing that the Peabody scores themselves are independent. I cannot assess that by looking at the data, but I am comfortable with the assumption because it seems unlikely that a large study like this one would have collected Peabody data in a way that violated independence.

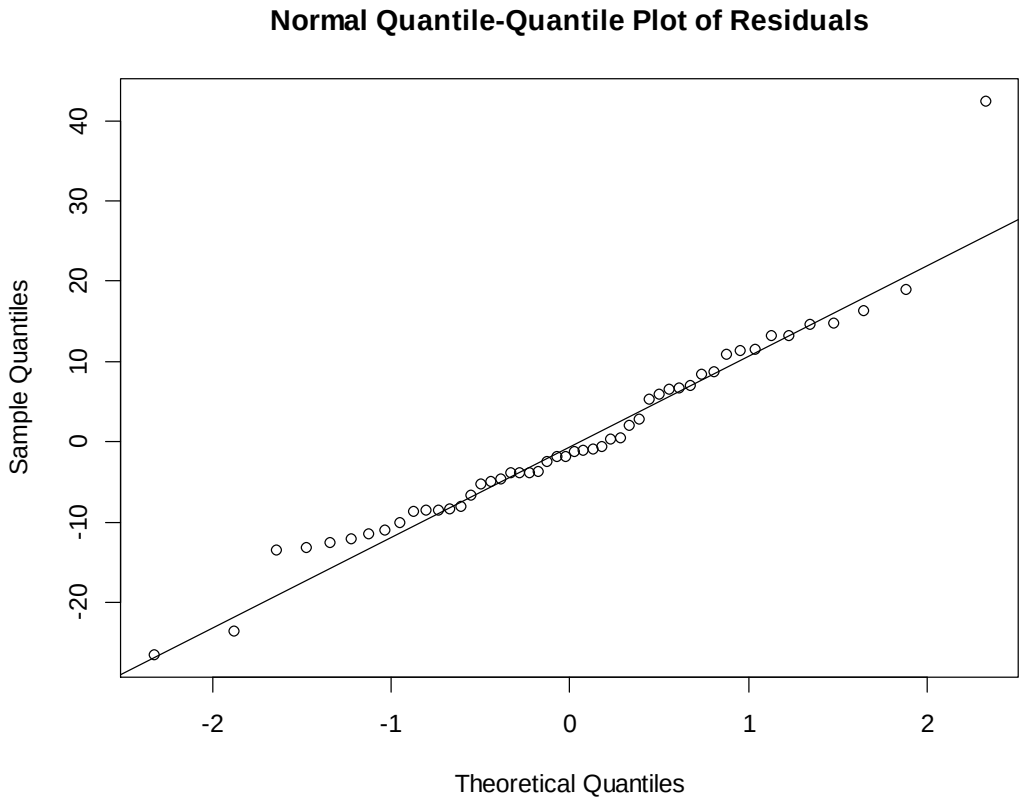
I evaluated the homoscedasticity assumption (i.e., that the vertical variability of errors about the regression line should be the same through the range of fitted values) using the same residuals plot that appears above. Although there may be a *slight* tendency for the variability of the residuals to increase toward the right side of the plot, the increase is not large. Given the overwhelming evidence against the null hypothesis about the slope, I'm not too worried about this assumption; if there *is* heteroscedasticity, it is slight enough that it probably wouldn't represent a reasonable explanation for the large values of the test statistics.

I evaluated the assumption of normally distributed errors by considering the normality of the residuals. Both in a stem-and-leaf plot…

```
> stem(regout$res)

  The decimal point is 1 digit(s) to the right of the |

  -2 | 74
  -1 | 3322110
  -0 | 98888755544442221111
   0 | 00235677789
   1 | 112335569
   2 |
   3 |
   4 | 2
```

...and in a normal Quantile-Quantile plot…

```
> qqnorm(regout$res, main="Normal Quantile-Quantile Plot of Residuals")
> qqline(regout$res)
```



...it appears reasonable to assume that the residuals (and, by implication, the errors) do not dramatically depart from a normal distribution. Hence, I am comfortable with the assumption that the errors are normally distributed.

Part Two: Using simulation to learn about probability distributions

I am not going to provide an example of this task for the exponential distribution mentioned in the homework because it would give away the findings that I want you to discover for yourself. Instead, I am doing the exercise using a chi-square distribution and varying the degrees of freedom (df). **It is very important that you realize my example here uses a different distribution from the one you are supposed to use! Wherever you see "`rchisq`" in my example, you should be using "`rexp`".**

I found it useful to construct a table of the different df values I tried, along with the resulting means and variances from each (very large) sample. My basic code (here, for df=1) looked like this:

```
x <- rchisq(500000, 1)
mean(x)
var(x)
pskew(x)
hist(x)
```

Note that I did not bother to "improve" these histograms beyond continuing to use the wide aspect ratio I used in Part One. You also need not do improvements here.

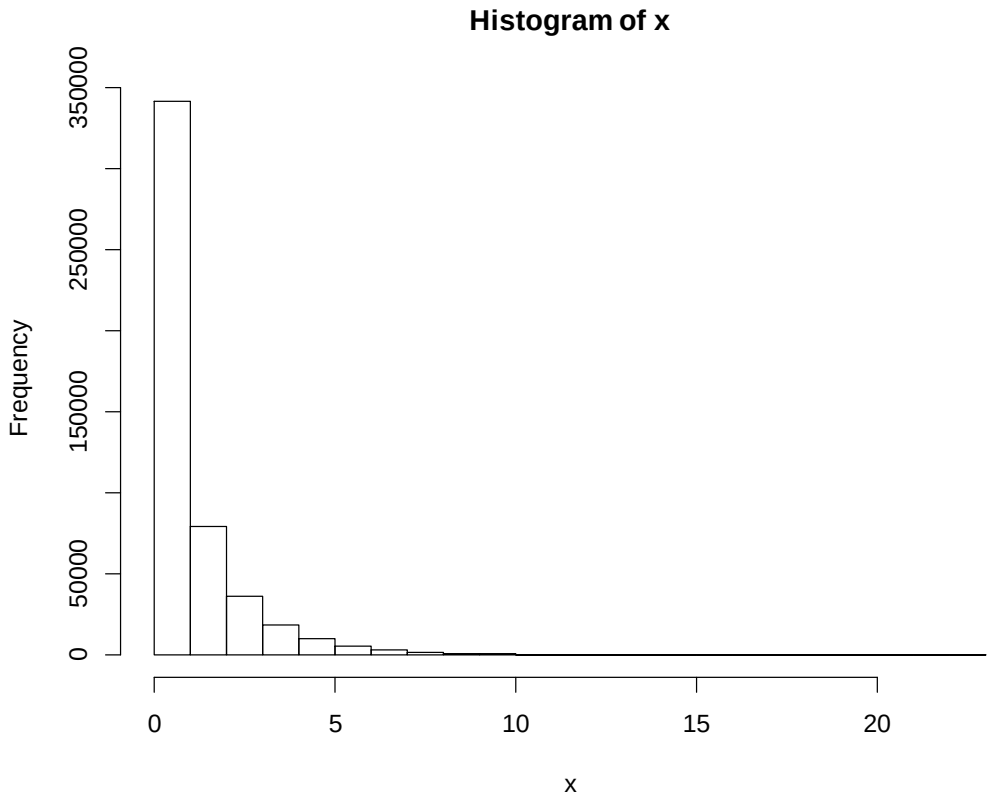
I tried the values 1, 2, 5, 10, and 50 for degrees of freedom, resulting in the following table:

df	*Mean*	*Variance*	*Pearson's Skew*
1	0.998	1.999	1.157
2	2.000	4.026	0.918
5	4.997	9.971	0.617
10	10.004	20.033	0.441
50	50.000	100.160	0.198

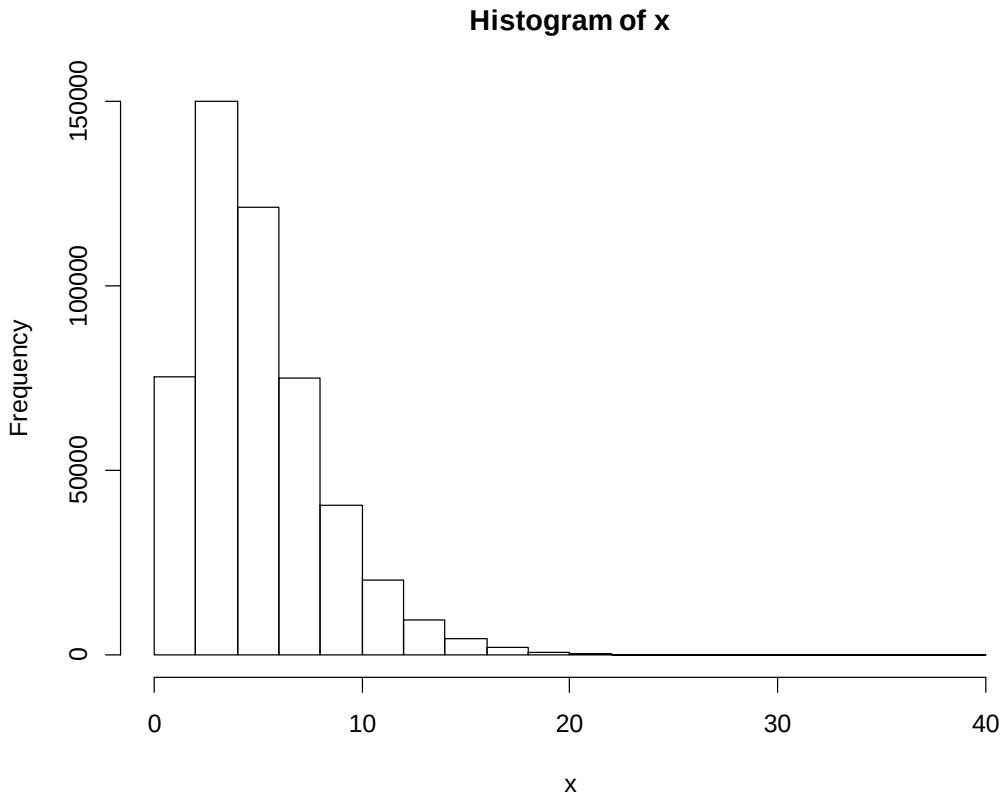
I am including three of the histograms here: 1 df, 5 df, and 50 df. That is sufficient to show the pattern I detected for the skew.

I note that the mean is always almost exactly the same as the degrees of freedom. From that, I conclude that the mean of the chi-square distribution is equal to its degrees of freedom. Similarly, the variance of the chi-square distribution appears to be twice the degrees of freedom. As for the skew, both the statistical evidence of Pearson's skew index and the graphical evidence from the histograms shows that for small degrees of freedom, there is positive skew, but as the df increases, the distribution becomes more symmetric.

Histogram for df=1:



Histogram for df=5:



Histogram for df=50: