This is a review for the **new** material that will be on the final exam. The final exam is comprehensive.

two sample test of proportion We use a "Z" statistic to test

$$H_0 : p_1 = p_1, \qquad H_A : p_1 \neq p_2$$

This was accidentally asked on the last test, and will likely be asked again

The *F* distribution and comparison of variances In sections 12.7 and 12.8 we learned how to test

$$H_0 : \sigma_1^2 = \sigma_2^2$$

when the populations were normal using the test statistic s_1^2/s_2^2. Its sampling distribution was the *F* distribution.

Regression In chapter 15 we covered 15.1 through 15.6. We did not cover 15.7-9.

The regression model for a data set (x_i, Y_i) (the x_i are not thought of as random, the Y_i are) is that

$$Y_i = \alpha + \beta x_i + \varepsilon_i$$

where ε_i are iid N$(0,\sigma^2)$. Alternatively, we can view the individual Y's as N$(\mu_{y|x},\ \sigma^2)$.

With this model for the data, we can do the following

1. Find $\hat{\alpha}$, $\hat{\beta}$, and $\hat{\sigma^2}$ using maximum likelihood estimators
2. Find the same using the method of least squares. Outside of a $n-2$ they are the same.
3. Find the sampling distribution of these estimators, and find pivotal quantities (think $(\hat{\beta} - \beta)/SE$) that allow us to compute confidence intervals and significance tests.
4. Find the mean square prediction error for predicting Y based on x.

The value of $\hat{\beta}$ is related to the correlation coefficient. The value $\hat{\alpha}$ is not as important, as it can be written in terms of the regression line, $\hat{y} = \hat{\alpha} + \hat{\beta} x$, going through the point $(\widehat{x}, \widehat{y})$.

We define the residuals as "data $-$ fit" or $e_i = y_i - \hat{y}_i$. These are used to assess whether the assumptions of the model seem valid for a data set.

Chi-square statistic We looked at three different cases where the chi-squared statistic:

$$X^2 = \sum \frac{(\text{obs} - \text{exp})^2}{\text{exp}}$$

can be used as a test statistic.

1. For testing the simple hypotheses specifying values of $p_1, p_2, \ldots, p_k$. For example, all $p_k = 1/k$. In this case the sampling distribution of X^2 is chi-squared with $k-1$ degrees of freedom. (Provided the multinomial model applies to the data and n is large enough so that $np_i \geq 5$ for each i.)

2. For testing a more complicated hypothesis

$$H_0 : p_i = g(\theta_1, \theta_2, \ldots, \theta_r), \quad H_A : \text{ not so}$$

In this case we estimate θ_i with the maximum likelihood estimator $\hat{\theta}_i$ and then X^2 is chi-squared with $n-1-r$ degrees of freedom.

3. For testing if two categorical variables are independent we use $\hat{p}_{ij} = \hat{p}_i \hat{q}_j$ (my notation), and the degrees of freedom involve r the number of rows and c the number of columns as $rc - 1 - (r-1) - (c-1)$.

Some sample problems:

1. Are the return rates on these two items similar?

```
         no sold   no return
-----------------------------
eMac       275           11
iMac       450           25
```

Formulate the questions as a significance test, and compute the p-value.

2. A test of different seeds produced measurements of crop height. The measurements are summarized below:

```
            n   xbar   s
----------------------------
seed 1     25    72    7
seed 2     15    67    12
```

Suppose the populations are normally distributed. Perform a significance test to see if the variances are equal.

3. Compute the correlation coefficient and the regression line for the data

```
x   1  2 2 3 3 4 5 6 7 7
---------------------------------
y   2  3 3 2 3 3 2 3 4 5
```

4. A regression has the following summary

```
> summary(lm(mpg ~ wt, data = mtcars))

Call:
lm(formula = mpg ~ wt, data = mtcars)
Residuals:
    Min      1Q  Median      3Q     Max
-4.5432 -2.3647 -0.1252  1.4096  6.8727
Coefficients:
            Estimate Std. Error t value Pr(>|t|)
```

```
(Intercept)  37.2851     1.8776  19.858  < 2e-16 ***
wt           -5.3445     0.5591  -9.559 1.29e-10 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
Residual standard error: 3.046 on 30 degrees of freedom
Multiple R-Squared: 0.7528,        Adjusted R-squared: 0.7446
F-statistic: 91.38 on 1 and 30 DF,  p-value: 1.294e-10
```

This is a model for miles per gallon versus the weight of a vehicle for some cars from 1993.

(a) Perform a two-sided significance test of no effect:

$$H_0 : \beta = 0$$

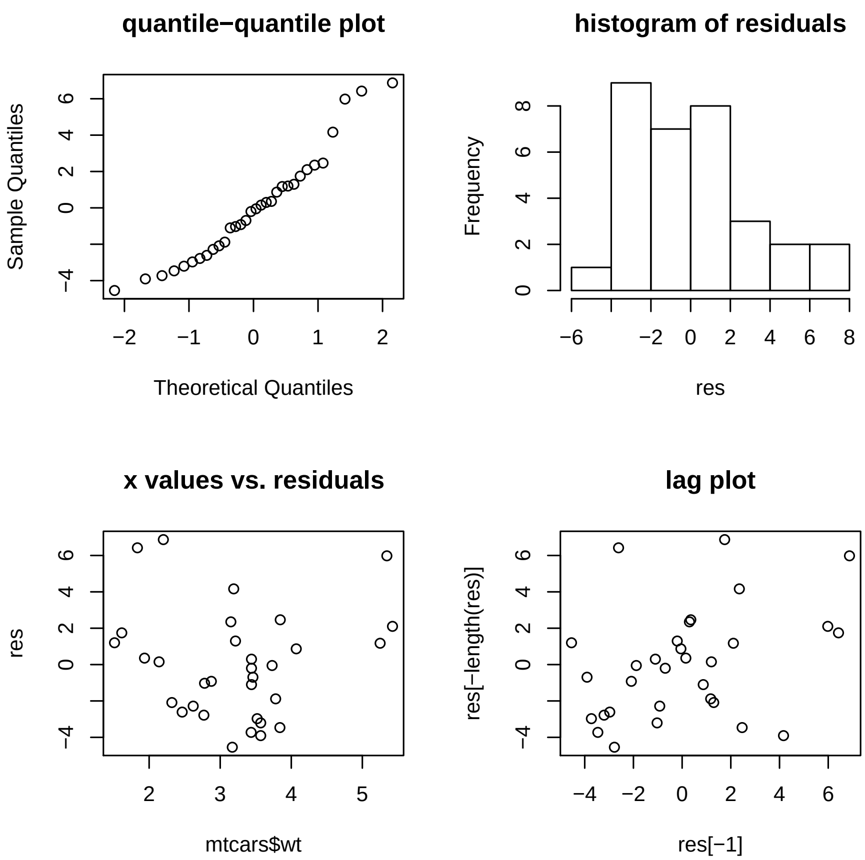
(b) The units of weight are in 1,000 pounds. Perform a two-sided regression test that each extra 1,000 pounds leads to a 5 mile reduction in the miles per gallon variable.

(c) Make a prediction of the mpg for a 2,800 pound MINI Cooper

(d) Make a prediction of the mpg for a 7,000 pound HUMMER H2.

5. The following plots are diagnostic plots of the residuals of the model above. Comment as to the validity of the simple linear regression model.

```
> res = residuals(lm(mpg ~ wt, mtcars))
> par(mfrow = c(2, 2))                # four graphs
> qqnorm(res, main = "quantile-quantile plot")
> hist(res, main = "histogram of residuals")
> plot(mtcars$wt, res, main = "x values vs. residuals")
> plot(res[-1], res[-length(res)], main = "lag plot")
```



6. The game of dreidel involves spinning a top which can land on one of 4 sides. Suppose a game lasts 60 spins and the distribution of values is

```
face    a    b   c    d
-----------------------
        8    7   30   15
```

Perform a significance test to see if the data is consistent with the assumption that the top is fair.

7. Are teen smoking and marijauna usage independent? A collection of 14-16 year olds were asked about their usage of each, The data are collected in the table below:

			marijauna	
		never	a few times	regularly
s				
m	never	20	5	7
o	a few times	15	20	10
k	regularly	5	10	8
e				

Perform a chi-squared test of independence for this data.