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:

$$H_0 : \beta = 0$$

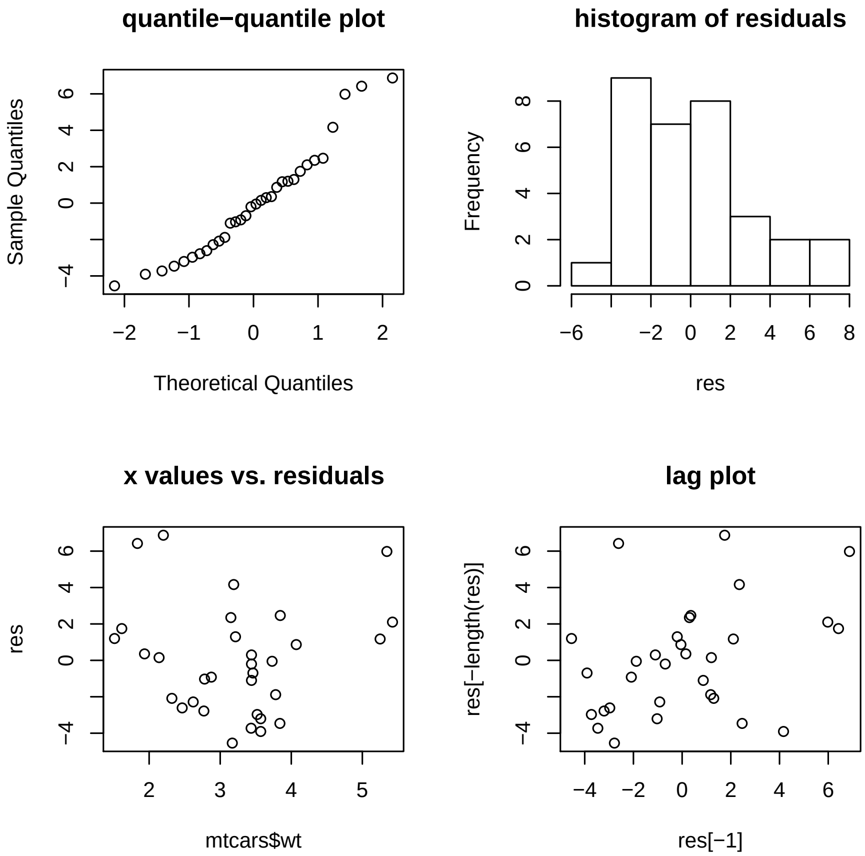
(b) The units of weight are in 1,000 pounds. Perform a two-sided regression test that each extra 1,000 pounds leads to a 5 mile reduction in the miles per gallon variable.

(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
smoke	never	20	5	7
	a few times	15	20	10
	regularly	5	10	8

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

Answers to the questions

1 This is a test of proportion where we assume the returns are independent. The analysis can be carried out using the Z statistic. Using the computer, a chi-squared statistic is used:

```
> prop.test(c(11, 25), c(275, 450))
```

```
^^I2-sample test for equality of proportions with continuity correction
data:  c(11, 25) out of c(275, 450)
X-squared = 0.5766, df = 1, p-value = 0.4476
alternative hypothesis: two.sided
95 percent confidence interval:
 -0.04985863  0.01874752
sample estimates:
    prop 1     prop 2
0.04000000 0.05555556
```

The p-value is 0.46.

2 Assuming normality, then we can use the F statistic to test the equivalence of variances. The observed value is

```
> f.obs = 7^2/12^2
> f.obs
```

```
[1] 0.3402778
```

Which gives a p-value of

```
> 2 * pf(f.obs, df1 = 25 - 1, df2 = 15 - 1)
```

```
[1] 0.01967451
```

The p-value is not strong support for the null.

3 Using the computer we have

```
> x = c(1, 2, 2, 3, 3, 4, 5, 6, 7, 7)
> y = c(2, 3, 3, 2, 3, 3, 2, 3, 4, 5)
> cor(x, y)
```

```
[1] 0.6546537
```

```
> lm(y ~ x)
```

```
Call:
lm(formula = y ~ x)
Coefficients:
(Intercept)            x
     1.8571       0.2857
```

4

1. The output

```
[...]
wt          -5.3445     0.5591  -9.559 1.29e-10 ***
[...]
```

performs the test. It has a tiny *p*-value.

2. We need to do this test by hand. The SE and degrees of freedom are read from the output.

```
> T.obs = (-5.3445 - (-5))/0.5591
> 2 * pt(-abs(T.obs), df = 30)
```

```
[1] 0.5424306
```

The *p*-value is not small.

3. This is simply

```
> 37.2851 - 5.3445 * 2.8
```

```
[1] 22.3205
```

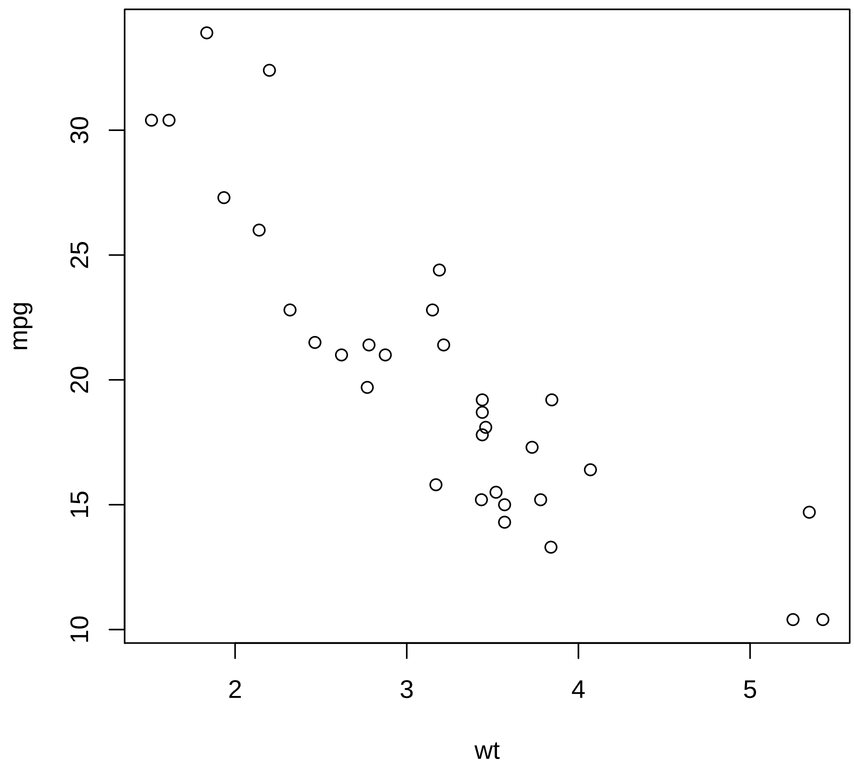
4. Again it is

```
> 37.2851 - 5.3445 * 7
```

```
[1] -0.1264
```

The negative value is to make you think twice about applying a linear model to predict values outside of the range of the data. A plot of the data shows a possible curve:

```
> plot(mpg ~ wt, data = mtcars)
```



5 The qqplot and histogram indicate that the residuals are *basically* a normal sample. However, there is a bit of a trend in the plot of the x values vs. the residuals. THis "U" shape indicates that a curved model for the mean might be more appropriate. The lag plot shows no apparent correlations between successive values.

6 That is, we test if the probability of landing on side i is 1/4. The `chisq.test` function can be used, but we do it by hand here

```
> f = c(8, 7, 30, 15)
> p = c(1/4, 1/4, 1/4, 1/4)
> e = sum(f) * p
> e

[1] 15 15 15 15

> f - e

[1] -7 -8 15  0
```

```
> (f - e)^2/e

[1]  3.266667  4.266667 15.000000  0.000000

> x2 = sum((f - e)^2/e)
> 1 - pchisq(x2, df = 4 - 1)

[1] 5.051616e-05
```

I think someone was cheating.

7 We actually can do this with the computer using

```
> chisq.test(rbind(c(20, 5, 7), c(15, 20, 10), c(5, 10, 8)))
```

but this involves stuff we didn't get a chance to talk about. Rather, we do it by hand.

First the marginals:

```
> s.marginals = c(20 + 5 + 7, 15 + 20 + 10, 5 + 10 + 8)
> m.marginals = c(20 + 15 + 5, 5 + 20 + 10, 7 + 10 + 8)
```

We have the $e_{ij} = R_i C_j / n$ (my notation in class). Notice $n = 100$. We write a for loop to put these numbers in a vector

```
> e = c()
> for (i in 1:3) {
+     for (j in 1:3) {
+         e = c(e, s.marginals[i] * m.marginals[j]/100)
+     }
+ }
> matrix(e, ncol = 3)

     [,1]  [,2] [,3]
[1,] 12.8 18.00 9.20
[2,] 11.2 15.75 8.05
[3,]  8.0 11.25 5.75
```

These match going across rows. The corresponding data is

```
> f = c(20, 5, 7, 15, 20, 10, 5, 10, 8)
> matrix(f, ncol = 3)

     [,1] [,2] [,3]
[1,]   20   15    5
[2,]    5   20   10
[3,]    7   10    8

> x2.obs = sum((f - e)^2/e)
> x2.obs

[1] 12.66304

> 1 - pchisq(x2.obs, df = 3 * 3 - 1 - (3 - 1) - (3 - 1))

[1] 0.01304515
```