Once again, this review sheet comes with no guarantee that it even makes sense, let alone will allow you to divine the contents of the final exam. Past performance may be an indicator of future performance though.

This is a review for chapters 13 and 15. The other chapters that we covered in this term are reviewed in two previous review sheets that may be found on the course website.

First, a reminder. The final exam will be Wednesday, May 22 at 12:15. We can take up to 3 hours. This final will be of similar length to the previous in-class exams, although it will cover material from the entire semester. There will be more material from chapters 13 and 15 than others as we haven't been tested on these yet. The exam will be open book, as usual, you can use your calculator, but you are not allowed to bring in outside notes.

Chapter 13:

Chapter 13 if title "Goodness of Fit". In this chapter, we learn how to test if data comes from some specified distribution. Prior to this, we tested in chapters 10 and 12 if the data came from a population with some specific parameter. For example, the t-test assumes the data is normal (nearly so anyway) and tests if the population has a specified mean. The variance test, tests if two data sets (which are both assumed to be normally distributed) have the same variances.

For the tests we learned about in chapter 13 they test to see if the data fits some specified distribution. The simplest one was the Chi-squared test. The other was the Kolmogorov-Smirnov test. We did not cover the Likelihood Ratio test explicitly (although, I could ask such material if I wanted you to derive it as that we did cover), and we did not cover paired tests for homogeneity. (That is, we did cover sections 13-1 to 13-5 but not the last part of 13-4 which coovered rankit plots from page 539 line 3 to the end of the section.)

Chi-Squared tests:

The multinomial model is similar to the binomial model only there are more categories that just 2: success and failure. For example a dice roll has 6 possible outcomes, and if we rolled a die lots of times, the number of 1's, 2's etc. we had would have a multinomial distribution. Operatively, we have a sequence of independent trials, each resulting in one of several categories and the probability of a getting a specific category does not change from trial to trial. The model is specified in terms of the number of trials n, and probabilities of the categories p_1, p_2, ... p_k. The Chi-Squared test allows us to test the hypotheses $H_0 : p_1 = \pi_1, ...p_k = \pi_k$ against the alternative that one or more of these is incorrect. That is we are testing if the data comes from the specified distribution. The test statistic used was

$$\chi^2 = \sum_{i=1}^{k} \frac{(f_i - e_i)^2}{e_i}$$

Where f_i is the frequency of category i for the data and e_i is the expected number we should have found in category i under H_o. This is easy to calculate, it is the probility times

the number of trials or $e_i = n\pi_i$. To do the hypothesis test, we need to know the sampling distribution, which in this instance is chi-squared with $k-1$ degrees of freedom.

A refinement to this is also discussed. First, we can summarize the null hypothesis above as $p = \pi$ in vector form. The refinement, is if we paramterize the π by some values $\pi = \pi(\theta_1, ..., \theta_r)$ and then use the data to predict the values of the θ. If we do this, then the chi-squared statistic should be smaller in value, as we are using the data to fit the paramters. Indeed this is true, and as such the distribution changes. If we fit r parameters, then the chi-squared statistic has $k-1-r$ degrees of freedom.

In section 13-5, another application of this statistic is used to test the hypothesis that the rows of a 2-way contigency table are independent. In general, a 2-way contigency table, is similar to the above based on a multinomial distribution with cell probabilities naturally indexed by π_{ij} where i if the row and j is the column. The null hypotheses of independence is simply stated as $p_{ij} = p_{i\cdot}p_{\cdot j}$ where the latter are the marginal probability. Notice, this parameterizes the values of $p_{i,j}$ by the smaller set of values $p_{i\cdot}$, $p_{\cdot j}$. That is, we have to fit $r-1$ plus $c-1$ values so the new statistic will have $n-1-(r-1)-(c-1)$ degrees of freedom where $n = rc$. This simplifies to $(r-1)(c-1)$. Finally, what to predict the $p_{i\cdot}$ with? Why the m.l.e's which turn out to be the sample proportions. For $p_{i\cdot}$ that is the number of samples in the ith row category divided by the total number of samples. Succinctly, the row sum divided by n.

Kolmogorov-Smirnov Statistic:

The Kolmogorov-Smirnov statistics are based on the cumulative distribution functions, or c.d.f. These are the integrated densities, or $F(x) = P(X \leq x)$. The *empirical* c.d.f. is based on the data. It is $F_n(x) = \#\{i : X_i \leq x\}/n$. The K.S. tests are based on the fact that when the data come from the distribution F, the *asymptotic* distribution of $D_n = \max \|F_n(x) = F(x)\|$ is

$$P(\sqrt{n}D_n \leq x) \to \sum_j (-1)^j e^{-2j^2x^2}$$

In class, we used just the terms $j = -1, 0, 1$ to get a quick approximation. The book contains tables.

How is this used? To test the hypothesis that H_0 the data is from F, against the alternative H_A that it is not, we can use the test statistic D_n which under H_0 has the above distribution.

This was generalized to test if two samples are from the same distribution $H_0 : F = G$. There is a slightly different asymptotic distribution.

Finally, the book covers a variant to test for normality. If we want to test if the data is normal(0,1) we just use the above. If we want to test the data is normal(μ, σ^2) with *unspecified* values for μ and σ, then the book suggests using $\bar{X}$ for μ and S^2 for σ^2. As usual, when we use the data, to pick the parameters we get less variation. So the distribution of D_n changes. A table is in the book.

Linear Regression:

We covered linear regression in class. We started with the model

$$Y = Z\beta + \epsilon, \quad \text{or} \quad y_i = \alpha + \beta x_i + \epsilon_i$$

where the ϵ_i are assumed to all be normal(0,σ^2). That is, the data (x_i, y_i) are assumed to be in a linear relatiship with slope β and intercept α. However, there is some noise or error that makes this not quite so when we sample.

There are 3 parameters to the model α, β and σ^2. In class, we showed that if we choose α, β to minimize the **squared residuals** (that is minimize $\sum(y_i - a - bx_i)^2$) that we can solve for α and β with the formulas

$$\hat{\beta} = \frac{SS_{xy}}{SS_{xx}} = \frac{\sum(x_i - \bar{x})(y_i - \bar{y})}{\sum(x_i - \bar{x})(x_i - \bar{x})}, \quad \hat{\alpha} = \bar{y} - \hat{\beta}\bar{x}$$

Furthermore, we saw in class, that as y_i is distributed as normal$(\alpha + \beta x, \sigma^2)$ the method of maximum likelihood estimation yields the same values for $\hat{\alpha}$ and $\hat{\beta}$ and a value of

$$\hat{\sigma^2} = \frac{1}{n}\sum(y_i - \hat{y}_i)^2$$

where $\hat{y}_i = \hat{\alpha}|\hat{\beta}x_i$ is the estimate for y_i and the difference $y_i - \hat{y}_i$ is called the **residual**. The above estimate is *biased*, so we use the unbiased one instead. This involves dividing by $n-2$ and not n

$$\hat{\sigma^2} = \frac{1}{n-2}\sum(y_i - \hat{y}_i)^2$$

Now, with these, we proceeded to find the sampling distributions of the three estimators. In summary, they are

- $\hat{\beta}$ is normal$(\beta, \sigma^2/SS_{xx})$. That is normal. unbiased with the specified mean
- $\hat{\alpha}$ is normal$(\alpha, \sigma^2(1/n + \bar{x}^2/SS_{xx}))$. Again, normal, unbiased with specified variance.
- $\hat{\sigma^2}$ has $(n-2)\hat{\sigma^2}/\sigma^2$ is chi-squared with $n-2$ degress of freedom

Testing hypotheses:

To test assumptions about the slope and intercept, we could use the sampling distributions above if we know σ. Unfortunately, we often don't. As such we use S to replace σ. A consequence is the distributions change from normal to t. For example, Under the assumption that $H_0 : \beta = \beta_0$ we have the sampling distribution of

$$t = \frac{\hat{\beta} - \beta_0}{\hat{\sigma^2}/SS_{xx}}$$

is t with $n-2$ degrees of freedom. This can be used to construct confidence intervals for β or to perform tests of significance.

Confidence intervals for prediction based on the regression line:

One goal of finding a regression line is to make predictions about new measurements. For example, the line may be used to make a prediction for another sample, or it may be used to make a prediction for an average. There is a subtley here. To make a prediction for the average value of y for a given x we use $\hat{y}$, and the variance is given by

$$var(\hat{y}) = \sigma^2(\frac{1}{n} + \frac{(x-\bar{x})^2}{SS_{xx}}$$

However, if we have to make a prediction for a given individual – and not an average – the variance we would expect to be greater. In this case. If we gather a regression line, then take one more data point with x value x_0, then the predicted value for y_0 would still be $\hat{y} = \hat{\alpha} + \hat{\beta}x_0$, but the variance is now

$$var(\hat{y}) = \sigma^2(1 + \frac{1}{n} + \frac{(x-\bar{x})^2}{SS_{xx}}$$

That is, there is a new value of 1 slipped in there. It doesn't look like much, but it makes a big difference in terms of what you see. Here is a concrete example from page 614. The data is

```
chirp   20   16   20   18   17   16   15   17   15    16
temp    89   72   93   84   81   75   70   82   69    83
> sum((chirp - mean(chirp))^2)
[1] 30                          # this is SS_xx
```

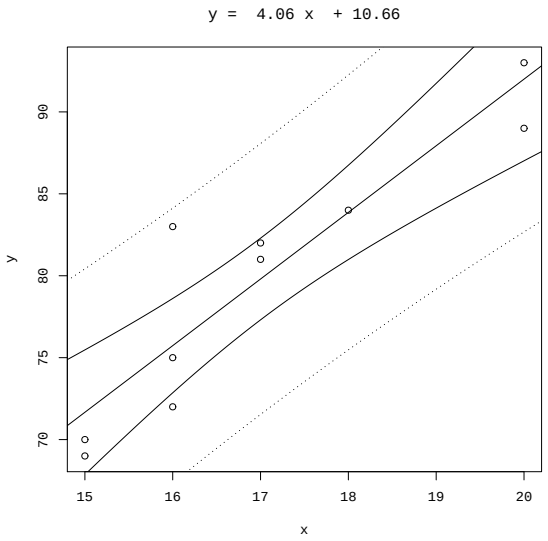
This is data which tries to predict the outside temperature by the number of chirps per sec of crickets.

The output from the computer is

```
> summary(lm(temp ~ chirp))

Call:
lm(formula = temp ~ chirp)

Residuals:
   Min     1Q Median     3Q    Max
-3.733 -2.417 -0.300  1.150  7.267

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  10.6667    10.6639   1.000 0.346473
```

```
chirp         4.0667     0.6241   6.517 0.000185 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 3.418 on 8 degrees of freedom
Multiple R-Squared: 0.8415, Adjusted R-squared: 0.8217
F-statistic: 42.47 on 1 and 8 DF,  p-value: 0.0001849
```

And the graph of the regression line with *both* confidence intervals is



Notice, the difference and similarities between the two. In particular, they both are bending. This is due to the $(x - \bar{x})^2$ term. As we get farther from the mean of x the confidence intervals get larger. As for differences, they say different things. In particular, if you want to know what the average temperature is outside when there are 19 chirps, then the 95% confidence interval is about (83,89). Whereas, if a given night you hear 19 chirps, and you want to predict the temperature for that night, then the range for a 95% confidence interval if from 78 to over 95.

Review Problems:

1. For the chirp example, find a 80% confidence interval for the value of β.
2. For the chirp example, do a test of significance to determine if β is 9 against the alternative that it is more than 9.
3. For the chirp example, compute the Standard error of $\hat{\alpha}$. Compare to that found by the computer.
4. Find the two 95% confidence intervals as above when $x = 16$. Explain why there is a difference.
5. Do a K.S. test to decide if this data comes from exp(2).

```
0.60 0.53 0.73 0.97 0.52 0.54 0.64 0.06 0.95 0.47
```

6. Do a K.S. test to dedide if these two rows of data come from the same distribution

```
x: 76 66 65 76  69
y: 89 90 83 103 102  77
```

7. Are these rows independent?

```
      y
 x 5 23 13 7
   8 25 10 4
```

Does this table of data satisfy the hypothesis that $p = (a, b, c, b, a)$? (That is, $p_1 = p_5$, $p_2 = p_4$.)

```
category 1  2  3 4 5
         -----------
count    5 10 15 8 7
```