The final exam is Wednesday May 21 in the usual class period. The final exam is comprehensive. New material since the second exam includes: correlation and linear regression. There are new formulas used to compute the values r, b_0, b_1 and new test statistics to do tests/confidence intervals on b_0 and b_1.

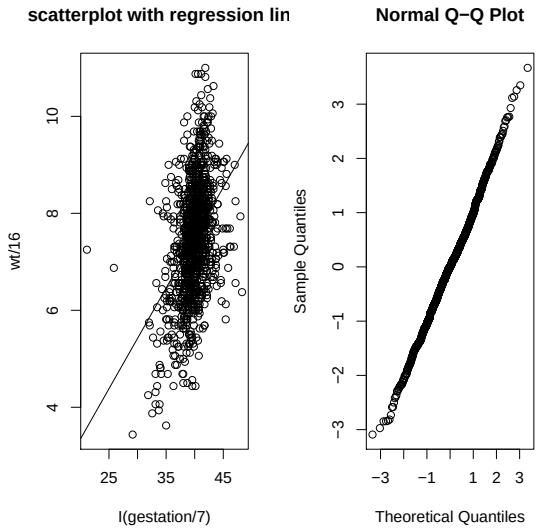
Here are some practice problems. **There are many topics not covered in these problems that may occur on the exam!!!**

Problem 1

The birthweight of a baby is in the variable `wt`, its mother's weight in `wt1`. Test the hypothesis that the mother's weight has an influence on the mean baby weight using a two sided alternative. The summary is

```
Call:
lm(formula = wt ~ wt1, data = babies, subset = wt1 < 800)
Residuals:
     Min       1Q   Median       3Q      Max
-66.0297 -10.9898   0.3189  11.0746  56.0159
Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 102.15029    3.25693  31.364  < 2e-16 ***
wt1           0.13485    0.02499   5.396  8.2e-08 ***
---
Signif. codes:  0 âĂŸ***âĂŹ 0.001 âĂŸ**âĂŹ 0.01 âĂŸ*âĂŹ 0.05 âĂŸ.âĂŹ 0.1 âĂŸ âĂŹ 1
Residual standard error: 18.15 on 1198 degrees of freedom
Multiple R-squared: 0.02373,        Adjusted R-squared: 0.02291
F-statistic: 29.12 on 1 and 1198 DF,  p-value: 8.207e-08
```

Problem 2



A standard rule of thumb is that soon-to-be-born babies grow a half pound per week in the womb. We will test this hypothesis using the weight data in the variable `wt` and the gestation time `gestation`. That is, perform a two-sided test of $H_0 : \beta_1 = 1/2$ using the data summarized with

```
> summary(res)

Call:
lm(formula = wt/16 ~ I(gestation/7), data = babies, subset = gestation <
    350)
Residuals:
      Min        1Q    Median        3Q       Max
-3.090512 -0.690321  0.001664  0.625327  3.668844
Coefficients:
               Estimate Std. Error t value Pr(>|t|)
(Intercept)    -0.81083    0.52841  -1.534    0.125
I(gestation/7)  0.20773    0.01323  15.706   <2e-16 ***
---
Signif. codes:  0 âĂŸ***âĂŹ 0.001 âĂŸ**âĂŹ 0.01 âĂŸ*âĂŹ 0.05 âĂŸ.âĂŹ 0.1 âĂŸ âĂŹ 1
Residual standard error: 1.041 on 1219 degrees of freedom
Multiple R-squared: 0.1683,        Adjusted R-squared: 0.1676
F-statistic: 246.7 on 1 and 1219 DF,  p-value: < 2.2e-16
```

1. From the scatterplot, estimate the correlation coefficient. (Is it positive or negative? Close to 1, -1 or 0?)
2. Find a 95% confidence interval for the slope.
3. The F Statistic is 246.7, which is very significant. Explain what this says and what test if rejected.
4. The ANOVA table for the above is

   ```
   > anova(res)

   Analysis of Variance Table
   Response: wt/16
                    Df  Sum Sq Mean Sq F value    Pr(>F)
   I(gestation/7)    1  267.13  267.13  246.68 < 2.2e-16 ***
   Residuals      1219 1320.08    1.08
   ---
   Signif. codes:  0 âĂŸ***âĂŹ 0.001 âĂŸ**âĂŹ 0.01 âĂŸ*âĂŹ 0.05 âĂŸ.âĂŹ 0.1 âĂŸ âĂŹ 1
   ```

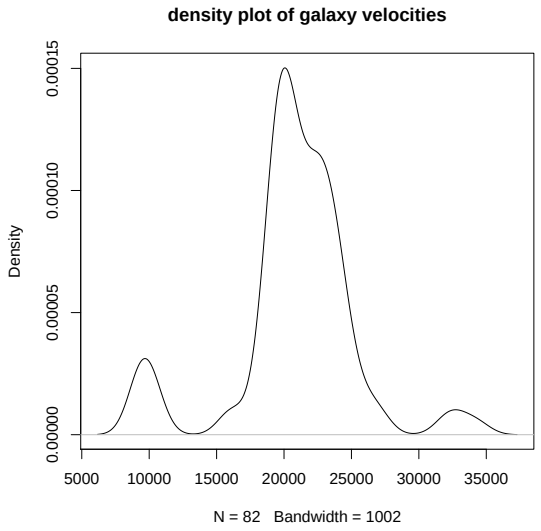
   Here we see the same statistic. Using the notation in class find SSE, SSM and SST. What is n?
5. Using SST, what is the total variation in the response variable (This is $s_y = \sqrt{SST/n-1}$.)
6. What percent of the total variation in the response variable is attributed to the regression model?

7. The error terms in a regression model are supposed to be a random sample from a mean-zero normal population. As such, the residuals should appear to come (more or less) from a normal distribution. above is a quantile-quantile plot of the residuals. Do they seem to satisfy the assumption? Explain.

Problem 3
Which of these models is binomial? For those that are, write as many of n, p and $\mu = np$ that you know.

1. A bag holds 52 balls, 6 are red. A person picks 7 balls without replacement. Let X count the number of red balls chosen.
2. A medical examiner is examining records until they find 100 that satisfy some search criteria. Let X be the number of records satisfying the search criteria
3. Toss a coin 1,000 times. Start with heads each time. Let X be the number of heads tossed.
4. A survey taker has a phone list of 300,000 SI residents. She calls 1,000 randomly chosen numbers (with replacement possible). Let X be the number who would talk to her.
5. A simulation finds 100 95% confidence intervals for μ. Let X be the number that actually contain the population parameter μ.

Problem 4
The density plot shows measurements of galaxy velocities.



1. Estimate the mean and median velocity of the data set.
2. Estimate the IQR of the data.
3. Distinct modes indicate a non-uniformity in the galaxy (voids or superclusters). Is this data set **unimodal**? If yes, say why, if no, describe what the data set shows.

Problem 5
Ten of the Yankees' salaries (in hundred thousands) are

```
3   4   7  10  27  30  35  60 160 214
```

1. Compute the sample mean
2. Compare to the median.
3. Compute the range.
4. Compute the IQR.
5. What is it about this data that makes the standard deviation (and the mean) a poor summary.

Problem 6
A student physics lab has the following data for current versus voltage:

```
current (mA) voltage (V)
========================
 100           15
 200           16
 200           16.5
 300           16.5
```

1. Compute the correlation.
2. Compute the coefficients of the least squares regression line with voltage as the response variable and current as the predictor variable.
3. Predict voltage for a 150 mA current.
4. Find the residual of the data point $(100, 15)$.

Problem 7
The distribution of a random variable X is specified by having possible values 1,2,3,4 and 5 and probabilities specified by

```
k         1    2    3   4    5
-----------------------------
P(X=k)  1/9  2/9  3/9 2/9  ???
```

1. Find the value of $P(X = 5)$.
2. Find the expected value of X. Show your work.

Problem 8
A process is guaranteed to be correct with a probability of 0.95. It is repeated independently 200 times. Let X recored the number of times it is correct.

Based on this, answer the following:

1. What is the expected value of X?
2. What is the standard deviation of X?
3. Write a mathematical expression that evaluates to $P(X = \mu)$. You do not need to give a numeric answer.

Problem 9
Let Z be a normally distributed random variable with $\mu = 0$ and $\sigma = 1$. Find the following probabilities:

1. $P(0 \leq Z \leq 1.23)$,
2. $P(Z \geq 2.13)$.

Problem 10
Let Z be a normally distributed random variable with $\mu = 0$ and $\sigma = 1$. Find a value b so that

$$P(-b \leq Z \leq b) = 0.80.$$

Problem 11
The average time for a penquin egg to hatch is 69 days with a standard deviation of 5 days. Assume the distribution of times is normally distributed.

Based on these assumptions answer the following

1. The probability that a randomly chosen egg hatches in 72 or more days.
2. The probability that a randomly chosen egg hatches between 64 and 74 days.

Problem 12
Let X be normally distributed with mean 0 and standard deviation $1/\sqrt{9}$. Let Y be normally distributed with mean 1 and standard deviation $1/\sqrt{9}$.

Find

1. $P(X \geq 0.7)$
2. $P(Y \leq 0.7)$

Problem 13
Historically, a certain type of strawberry bush has yielded 100 strawberries with a standard deviation of 75. The distribution of yields is skewed right, but the tails are not too long.

If 25 such strawberry plants are planted, find the probability that the average yield per plant is less than 90.

Your answer should show your work, give the probability, and *state any additional assumptions you made about the data.*

Problem 14

To evaluate a new survey taker, a company sends the person out to survey a population where it is known that the population proportion is $p = 0.25$. The survey taker is supposed to take a random sample of size 250 from the population. The survey taker's sample proportion was $\widehat{p} = 0.20$.

Find the probability that a random sample of size 200 from this population would produce a value for $\widehat{p}$ of 0.20 *or less*.

Your answer should show your work, give the probability, and *state any additional assumptions you made about the data.*

Problem 15

According to a New York Times article 8.7% of American adults have type-II diabetes. Is this the same for children? Suppose a random sample of 1000 American children showed that 50 had type-II diabetes.

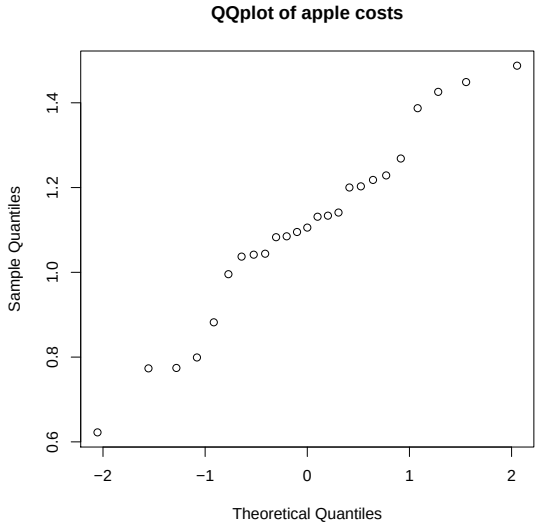
Based on this, find a 95% confidence interval for the proportion of all American children with type-II diabetes. Does your answer include the figure for adults?

Your answer should show your work, state your answers, and describe any assumptions you made about the data.

Problem 16

In a New York Times article on tpye-II diabetes in children, a statement is made that it is cheaper to buy french fries than fruit. Is this so? Suppose a box of french fries at a fast-food outlet costs 0.99. To see if an apple is more expensive, a random sample of stores was taken. From each store, a single apple was purchased. Due to the variations in the weights of the apples and the different price per pounds, there was a range in the prices for apples.

A quantile-quantile plot of the data is shown below:



The sample is summarized by

$$\bar{x} = \$1.1, \quad s = \$0.21, \quad n = 25$$

Based on the sample, find a 90% confidence interval for the population mean price per apple. Does it include \$0.99?

Your answer should show your work, state your answers, and describe any assumptions you made about the data.

Problem 17

A survey designer would like to ensure that his survey has a *margin of error* of no more than 2 percentage points for a 95% confidence interval. How large must n be so that this will be true? Assume the worst case scenario of $\widehat{p} = 1/2$.

Formula sheet for final exam

[This will be the formula sheet, unless additional formula are requested via email.]

Some basic summary statistics:

$$\bar{x} = \frac{\sum x_i}{n}, \quad s = \sqrt{\frac{\sum (x_i - \bar{x})^2}{n-1}}, \quad IQR = Q_3 - Q_1, \qquad z = \frac{x_i - \bar{x}}{s} \text{ or } z = \frac{x_i - \mu}{\sigma}$$

The median is the "middle" point, suitably defined.

A distribution of a random variable is a specification of the chance of a given event. For discrete distributions this is done by defining $P(X = k)$ for each appropriate k. For continuous distributions, this is done by defining $P(X \leq z)$ using a density.

For a finite random variable:

$$\mu = \sum kP(X = k), \quad \sigma^2 = \sum (k - \mu)^2 P(X = k).$$

A special case is the binomial with parameters n and p. For this we have

$$P(X = k) = \binom{n}{k} p^k (1-p)^{n-k}, \quad \mu = np, \quad \sigma = \sqrt{np(1-p)}.$$

The central limit theorem states that if $x_1, x_2, \ldots, x_n$ is a random sample from a population with mean μ and standard deviation σ, then the sample mean is approximately normal with mean μ and standard deviation $\sigma/\sqrt{n}$.

Under assumptions, $(1-\alpha)100\%$ CIs for the population proportion p based on $\widehat{p}$ and the population mean μ based on $\bar{x}$ are respectively

$$\widehat{p} \pm z^* \sqrt{\widehat{p}(1-\widehat{p})/n} \quad \text{and} \quad \bar{x} \pm t^* s/\sqrt{n},$$

where t^* and z^* and related to α by

$$P(-z^* \leq Z \leq z^*) = 1 - \alpha, \quad P(-t^* \leq t_{n-1} \leq t^*) = 1 - \alpha.$$

Test statistics The following test statistics may prove useful. For confidence intervals, there may be slight modifications. Under the proper assumptions, statistics labeled Z have a normal sampling distribution and ones labeled T have a t-distribution with corresponding degrees of

freedom.

$$Z = \frac{\widehat{p} - p}{\sqrt{p(1-p)/n}} \qquad Z = \frac{\widehat{p}_1 - \widehat{p}_2}{\sqrt{\widehat{p}(1-\widehat{p})}\sqrt{1/n_1 + 1/n_2}}$$

$$T = \frac{\text{observed} - \text{expected}}{\mathsf{SE}} \qquad T = \frac{\bar{x} - \mu}{s/\sqrt{n}} \quad n-1 \text{ d.f.}$$

$$T = \frac{\bar{x}_1 - \bar{x}_2}{s_p\sqrt{1/n_1 + 1/n_2}}, \quad n_1 + n_2 - 2 \text{ d.f.} \qquad T = \frac{\bar{x}_1 - \bar{x}_2}{\sqrt{s_1^2/n_1 + s_2^2/n_2}} \quad \text{smaller of } n_1 - 1, n_2 - 1$$

(Use the pooled standard deviation when you assume $\sigma_1 = \sigma_2$, otherwise, use the other value to find SE.)

Regression The Pearson correlation and regression coefficients are given by

$$r = \frac{\sum(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum(x_i - \bar{x})^2 \sum(y_i - \bar{y})^2}}, \qquad b_1 = \frac{\sum(x_i - \bar{x})(y_i - \bar{y})}{\sum(x_i - \bar{x})^2}, \quad b_0 = \bar{y} - \hat{\beta}_1 \bar{x}.$$

The statistic used to estimate σ^2 is

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

This test statistic is used for testing the β's:

$$T = \frac{b_i - \beta_i}{\mathsf{SE}} \quad n-2 \text{ d.f..}$$

The formula for SE is not given. If you need it, it will be supplied for you in computer output.

Finding *p*-values You will be asked to find a *p*-value for many of these questions. You have a copy of the normal table and for the tail of the *t* distribution attached to the exam. If you are using the *t* table, you may not be able to give the exact *p*-value. In this case, use the table to give an upper and lower value for the *p*-value. For instance, something such as $0.10 > p > 0.05$.