6in,4in

The final exam is comprehensive. Here are some practice problems that may look familiar, except for the new one on multiple regression. A multiple regression model of the birthweight by the variables

The new material on the exam will be only on multiple regression. You should understand what the F statistic in the output of `summary` tests ($H_0 : \beta_1 = \cdots = \beta_p = 0$, H_A : not true). And you should understand that `anova()` is used to test two *nested* models for whether a new term is necessary.

Problem 1 (0 points):

The data set babies contains data on birthweights of babies and information about the mother and father.

A simple model of birthweight `wt` using both the moms weight `wt1` and dad's weight `dwt` is given by

```
> res.1 = lm(wt ~ wt1 + dwt, data = babies)
> summary(res.1)


Call:
lm(formula = wt ~ wt1 + dwt, data = babies)

Residuals:
     Min       1Q   Median       3Q      Max
-64.3716 -11.1797   0.4119  11.3420  56.4245

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 1.184e+02  9.740e-01 121.610   <2e-16 ***
wt1         5.712e-03  3.511e-03   1.627    0.104
dwt         4.911e-04  1.277e-03   0.385    0.701
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 18.23 on 1233 degrees of freedom
Multiple R-Squared: 0.002307,        Adjusted R-squared: 0.0006883
F-statistic: 1.425 on 2 and 1233 DF,  p-value: 0.2408
```

1. Are any of the t-tests (the ones below coefficients) flagged as significant?
2. Is the p-value of the F statistic big or small?

We can compare this model to the more complicated model also involving gestation time (`gestation`), and mom's and dad's height (`ht`,`dht`).

```
> res.2 = lm(wt ~ wt1 + dwt + gestation + ht + dht, data = babies)
> summary(res.2)


Call:
lm(formula = wt ~ wt1 + dwt + gestation + ht + dht, data = babies)

Residuals:
     Min       1Q   Median       3Q      Max
-63.6719 -11.1692   0.4676  11.2909  56.3773

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 75.888818  11.085027   6.846 1.20e-11 ***
wt1         -0.005480   0.004368  -1.254   0.2099
dwt         -0.003281   0.004642  -0.707   0.4798
gestation    0.013362   0.006862   1.947   0.0517 .
```

1. Is the bigger model "better?"
2. Which variables might be considered unimportant in the bigger model based on the flagged variables?

Problem 2 (newone points):
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.

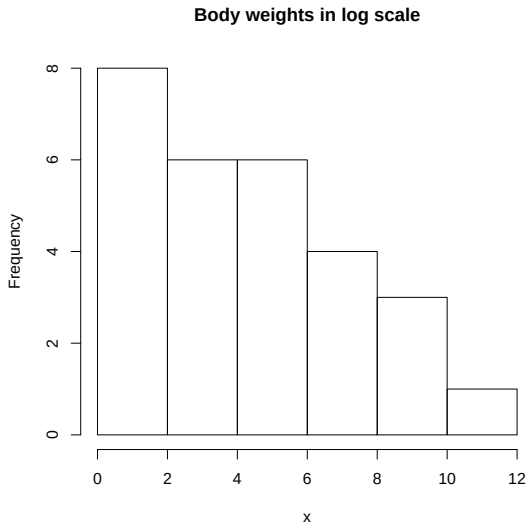
Problem 3 (5 points):
The following stem-and-leaf diagram records the salaries of the 2004 Yankees in dollars.

```
The decimal point is 7 digit(s) to the right of the |

0 | 000001111122233334
0 | 678999
1 | 1222
1 | 669
2 | 2
```

1. Describe the shape of the distribution.
2. Based on the stem and leaf diagram, find the values of the minimum, median, and maximum *Use the correct units.*
3. There is some truncation of the data to fit this choice of stem. Find the value corresponding to Enrique Wilon's salary of $700,00.

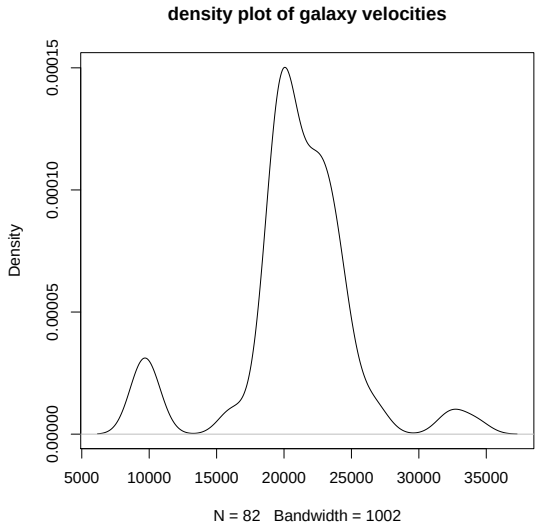
Problem 4 (5 points):
A data set on body weights for various animals yields the following histogram



There are 28 animals in the data set.

1. What percent of the animals measure 4 or fewer? (None are exactly 4)
2. Estimate as accurately as you can the median body weight measurement.

Problem 5 (5 points):
The density plot shows measurements of galaxy velocities.



1. Estimate the mean velocity of the data set.

2. 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 6 (10 points):
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. Compute the 20% trimmed mean
3. Compare these values to the median.
4. Compute the range.
5. Compute the IQR.

Problem 7 (10 points):
This data set on body weights and brain weights for various animals has a Pearson correlation of -0.06873.

```
> animals
                  body brain
Horse              521 655.0
Brachiosaurus     8700 154.5
Mouse           0.0230   0.4
Potar monkey        10 115.0
Cat                3.3  25.6
```

1. Draw a scatterplot
2. Compute the Spearman correlation coefficient
3. Explain why the Pearson and Spearman correlation coefficients have the similar/different values that they do.

Problem 8 (10 points):
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 coefficients of the least squares regression line with voltage as the response variable and current as the predictor variable.
2. Find the residual of the data point $(100, 15)$.

Problem 9 (5 points):
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 10 (5 points):
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 11 (5 points):
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 12 (5 points):
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 13 (5 points):
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 14 (5 points):
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 15 (5 points):
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 16 (5 points):
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.*

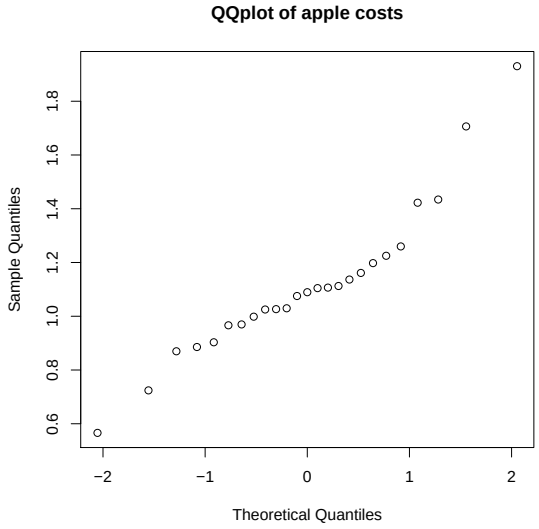
Problem 17 (5 points):
According to a New York Times article 8.7% of American adults have type-II diabetes. Is this the same for children? Suppose a random sample of 1000 American children showed that 50 had type-II diabetes.

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 18 (5 points):
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

$$\bar{x} = \$1.11, \quad s = \$0.28, \quad n = 25$$

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 19 (5 points):
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 $\widehat{p} = 1/2$.