Statistics Final Exam - answers

(I created this more for my own uses in grading the exam more than as what you needed to do in answering the questions. In particular, R is not the only calculator to do this stuff, and explaining things with plots or sketches could help to convince me that you understand the concepts. - Jim)

1

A study of 20 randomly selected students found that they spent an average of $16.92/week on gas. The standard deviation of the sample was $3.10. Find the 99% confidence interval of the true mean.

In [2]:

```
average = 16.92     # estimate of parent population mean
sigma = 3.10        # estimate of parent population standard deviation
N = 20              # sample size

# This study gives an estimate of the population true mean. The best estimate of of the me
an is 16.92 ,
# and its standard error (i.e. standard deviation of the population of such studies, each
a mean of 20) is
error = sigma / sqrt(N)      # estiate of standard deviation of (estimate of parent popula
tion mean)
error
```

Out[2]:

```
0.693181073024935
```

In [5]:

```
# The 99% confidence range is that part of the normal distribution which
# contains 99% of the probability. Of that 1%, half (0.5%) is above the confidence
# interval, and half (0.5%) is below. We can use the R qnorm() function
# to find the value of z (normal curve x axis in units of standard deviations)
# where the cutoff is for 99%.
qnorm(1 - 0.005)      # What z value has 0.995% of the values below it?
```

Out[5]:

```
2.5758293035489
```

In [7]:

```
# The confidence interval is therefore (average ± 2.58 * error) which is
2.58 * error   # plus or minus
```

Out[7]:

```
1.78840716840433
```

In [9]:

```
c(average - 2.58 * error, average + 2.58 * error)   # range
```

Out[9]:

```
15.1315928315957   18.7084071684043
```

The confidence interval is ($ 16.92 ± $ 1.79) or ($15.13 to $18.71).

2

A variable is normally distributed with mean 52.1 and standard deviation 4.6. What is the probability that a randomly chosen variable lies between 58 and 61? What value marks the 90th percentile?

In [10]:

```
average = 52.1
sigma = 4.6

# First calculate the corresponding z scores.
z_high = (61 - average)/sigma
z_low = (58 - average)/sigma

# Then use R's pnorm function (or an online normal distribution calculator)
# to find the probability between those two values
pnorm(z_high) - pnorm(z_low)
```

Out[10]:

```
0.07330611202004
```

The probability being between 58 and 61 is 7.3% .

In [12]:

```
# To find the 90% percentile, I first find the z value with the qnorm()
# function, and then turn that back into the scale of the original variable
z_90 = qnorm(0.90)  # What z value has 90% of the probability below it?
z_90
```

Out[12]:

```
1.2815515655446
```

In [13]:

```
average + 1.281 * sigma   # The variable value at the 90% percentile
```

Out[13]:

```
57.9926
```

The 90% percentile is at 58.

3

Here are the three year rates of return on various mutual funds:

```
5.37 4.31 4.13 8.58 5.99 7.90 9.11 6.11
3.06 14.48 12.50 8.33 10.10 8.21 6.83 10.94
2.34 0.97 8.33 8.89 6.07 6.50 5.99 9.38
0.05 13.88 3.71 10.07 9.88 4.93 6.38 10.34
2.27 11.91 11.69 12.06 9.84 7.75 2.86 6.68
```

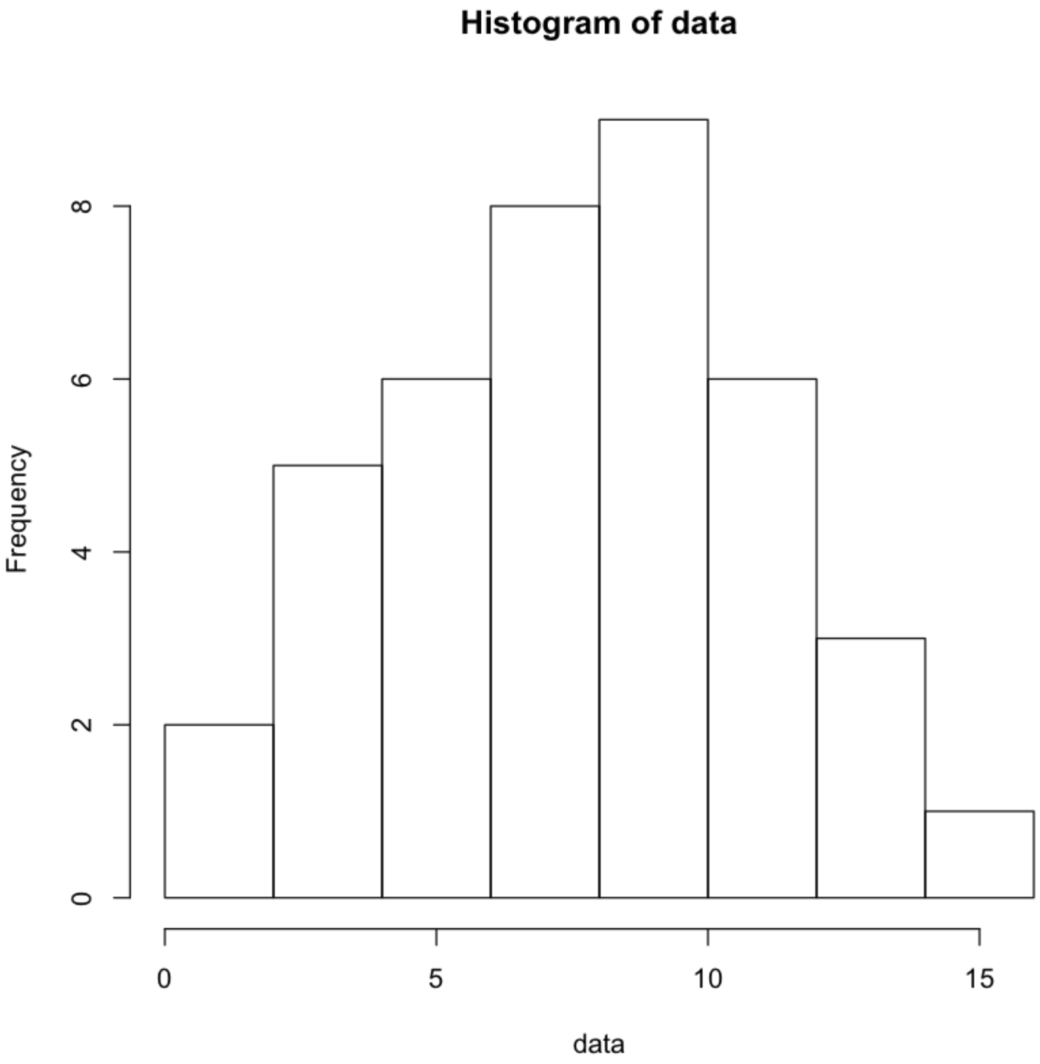
Display these data in an appropriate graph. What are the mean, median, and standard deviation of these data? How are those values calculated?

In [26]:

```
# In spite of the table presentation of the numbers, I am taking
# these to be one list of values, since the "find mean, standard
# deviation, median" instructions are appropriate for a list, not a table.

data = c(5.37, 4.31, 4.13, 8.58, 5.99, 7.90, 9.11, 6.11,
         3.06, 14.48, 12.50, 8.33, 10.10, 8.21, 6.83, 10.94,
         2.34, 0.97, 8.33, 8.89, 6.07, 6.50, 5.99, 9.38,
         0.05, 13.88, 3.71, 10.07, 9.88, 4.93, 6.38, 10.34,
         2.27, 11.91, 11.69, 12.06, 9.84, 7.75, 2.86, 6.68)

# Next I'll show several possible plots, using some of R's built-in functions,
# where are the simplest for this simple data set.
# Just googling "R histogram" or "R box and whiskers" finds these, or
# you could use the ggplot graphing library we did at the beginning of the term.
```

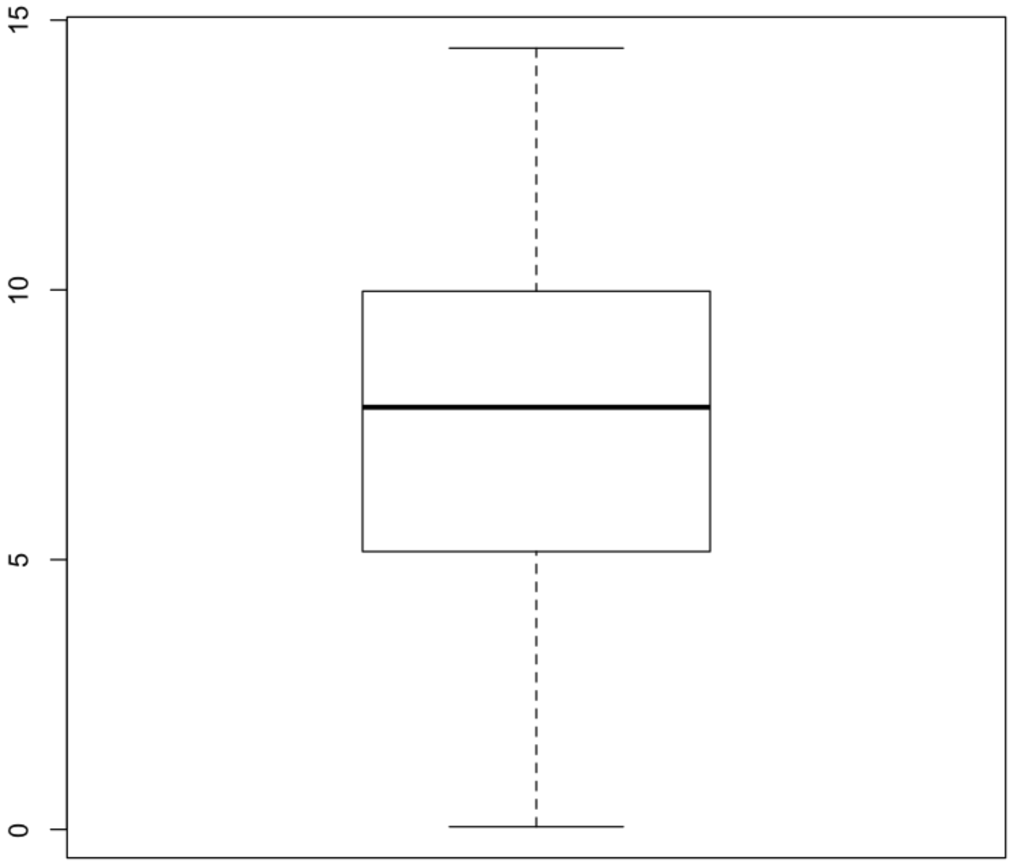
In [27]:

```
# First, to visualize this as a distribution, the most appropriate plot is a historgram
# which shows us how frequent each number is within some bin size.

hist(data)
```



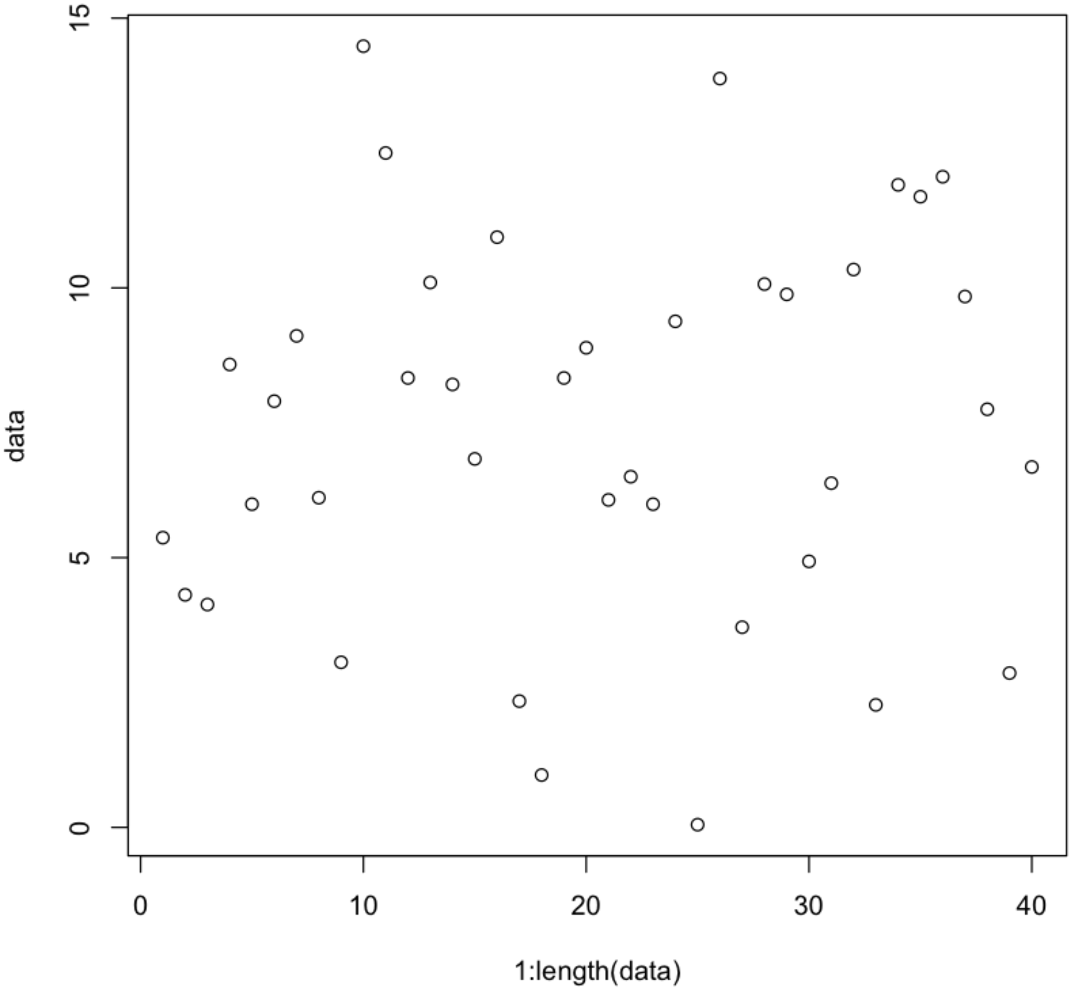
In [32]:

```
# Second, another reasonable way to see the data
# is with a box and whisker's plot that shows the range.

boxplot(data)
```



In [29]:

```
# Here's one more way to look at the data ... but not a good one.
#
# For this data, I'd call this an *inappropriate* plot,
# since the numbers along the x axis don't mean anything.
#
# This just shows that there is no trend in how the numbers
# were arranged sequentially.
#
# This type of plot is helpful when there are two (x,y) variables
# to compare and look for correlations, or different groups of
# variables to compare. That is not the case here.

plot(1:length(data), data)
```



In [33]:

```
# The mean is one sort of "central" value.
mean(data)
```

Out[33]: 7.468

In [34]:

```
# The mean is calculated by adding up the numbers and dividing by how many there are.
sum(data) / length(data)
```

Out[34]:

```
7.468
```

In [35]:

```
# The median is the middle value - half above, half below.
# It is another "central" measure, which is changed less by an outlier way away from the o
thers.
# Typically though the mean and median are similar.
median(data)
```

Out[35]:

```
7.825
```

In [41]:

```
# To see that this is the middle, put them in order and count up half way.
sorted_data = sort(data)
sorted_data
```

Out[41]:

```
0.05  0.97  2.27  2.34  2.86  3.06  3.71  4.13  4.31  4.93  5.37  5.99  5.99  6.07  6.11
6.38  6.5  6.68  6.83  7.75  7.9  8.21  8.33  8.33  8.58  8.89  9.11  9.38  9.84  9.88
10.07  10.1  10.34  10.94  11.69  11.91  12.06  12.5  13.88  14.48
```

In [42]:

```
# The media is between these two values.
N = length(data)
c( sorted_data[N/2] , sorted_data[N/2 + 1] )
```

Out[42]:

```
7.75  7.9
```

In [43]:

```
# The standard deviation is a measure of the width of the distribution.
sd(data)
```

Out[43]:

```
3.50907994370064
```

In [47]:

```
# It is found by
#  (1) subtracting each value from the mean,
#  (2) squaring those,
#  (3) finding the mean of the squares
#  (4) and then taking the square root to get back to the original units.
# Typically that "true" sample standard deviation is multiplied by sqrt(N/(N-1))
# to give a best estimate of the parent population's standard deviation.
# The sd() function in R gives the 1/sqrt(N-1), slightly larger, parent population estimat
e.

# Those steps can be done in R like this :
#

average = mean(data)     # 0: find mean
diffs = data - average   # 1: take differences from mean
diff2 = diffs ** 2       # 2: square them
var = mean(diff2)        # 3: find their mean (also called "variance")
sigma = sqrt(var)        # 4: take the square root to get the sample standard deviation
sigma * sqrt(N/(N-1))      # scale up a bit for parent population estimate
```

Out[47]:

```
3.50907994370064
```

4

You will roll a single fair six-sided die successive times, adding the score as you go. Stop when you have either scored at least 6 or rolled the die 3 times. What's the probability that you take exactly two rolls? What is the expected value of the number of rolls?

In [51]:

```
# On the first roll, there is a 1/6 chance of rolling a 6
# and if so we stop there. So P(1 roll) = 1/6
1/6
```

Out[51]:

```
0.166666666666667
```

In [48]:

```
# On the second roll, we stop if we get 6 or over for the sum.
# We want the probablity of this happening.
# The ways to get 6 or over are
#
#     first roll      second roll
#     ----------      -----------
#     1               5,6
#     2               4,5,6
#     3               3,4,5,6
#     4               2,3,4,5,6
#     5               1,2,3,4,5,6
#
# The probability of each value of the first roll is 1/6.
# The probability of the values for the 2nd roll are (n/6)
# where n is the number of rolls that put us over 6.
#
# We use P(A & B) = P(A) * P(B) to get the probability of each line,
# where P(A) = probability of 1st roll, and P(B) = probability of 2nd roll.
#
#     first roll      second roll           probability per line
#     ----------      -----------           ---------------------
#     1               5,6                   (1/6) * (2/6)
#     2               4,5,6                 (1/6) * (3/6)
#     3               3,4,5,6               (1/6) * (4/6)
#     4               2,3,4,5,6             (1/6) * (5/6)
#     5               1,2,3,4,5,6           (1/6) * (6/6)
#
# The probability of getting 6 or over is the sum of all these,
# since they are disjoint (cannot both happen) and P(x or y) = P(x) + P(y)
# where the x and y are different lines in this table.
#
# So P(2 rolls) = (1/6) * (2/6 + 3/6 + 4/6 + 5/6 + 6/6)
#               = (2 + 3 + 4 + 5 + 6)/36 = 20/36 = 0.555
#
sum(2:6)/36  # probability of exactly 2 rolls.
```

Out[48]:

```
0.555555555555556
```

In [50]:

```
# Since we stop at 3 rolls if we get that far,
# the probability of P(1 roll) + P(2 rolls) + P(3 rolls) = 1,
# so P(3 rolls) = 1 - P(1 roll) - P(2 rolls) = 1 - 1/6 - 20/36
#               = (36 - 6 - 20)/36 = 10/36
10/36
```

Out[50]:

```
0.277777777777778
```

In [53]:

```
# The expected (average) number of rolls = sum ( rolls * P(rolls) )
#                                         = 1 * 1/6 + 2 * (20/36) + 3 * (10/36)
#                                         = (1*6 + 2*20 + 3*10)/36 = 76/36 = 2.11
(1*6 + 2*20 + 3*10)/36
```

Out[53]: 2.11111111111111

5

It is claimed that a particular diet regime helps weightlifters improve their performance. Below are the data for eight weightlifters' maximum achieved bench press in pounds, both before the regime and after the regime has been in place for a month. Is the regime effective? (Assume that the variable is approximately normally distributed.) Include with your argument a plot that supports your conclusion.

Weightlifter	1	2	3	4	5	6	7	8
Before	210	230	182	205	262	253	219	216
After	219	236	179	204	270	250	222	216

In [55]:

```
# Since the numbers are paired into each lifter's weight before and after,
# I just subtract them to get the weight gain for each lifter
# rather than keeping two sets of numbers.

before = c(210, 230, 182, 205, 262, 253, 219, 216)
after = c(219, 236, 179, 204, 270, 250, 222, 216)
gain = after - before
gain
```

Out[55]:

```
9  6  -3  -1  8  -3  3  0
```

In [56]:

```
# Now I set this up as a hypothesis test :
#   H0 = null hypothesis = no weight gain, i.e. mean(gain) = 0
#   HA = motivated alternative hypothesis = mean(gain) > 0    (one tailed test)
#   alpha = significance level = 0.05 (i.e. reject null if pvalue < 0.05)

H0 = 0.0
alpha = 0.05

measured = mean(gain)  # best estimate of parent population mean
sigma = sd(gain)       # best estimate of parent population standard deviation
N = length(gain)       # sample size
error = sigma/sqrt(N)  # standard error, i.e. sd for (estimate of population mean)

z = (measured - H0)/error   # z score for average gain
z
```

Out[56]: 1.38835738622448

In [57]:

```
# So the gain seen is 1.38 standard deviations above the mean.
# This isn't all that unlikely, and so the motivated hypothesis
# isn't looking good.

# Since this is a one tailed test with HA predicting a value above H0,
# the pvalue is the right hand tail of the normal distribution.
pvalue = 1 - pnorm(z)
pvalue
```

Out[57]: 0.0825141228649259

We see that the pvalue (8.2%) is bigger than the alpha significance cutoff (5%), and so we fail to reject the null hypothesis.

In other words, we do *not* have statistical evidence that there is any difference between the "before" and "after" numbers.

A plot illustrating this could be

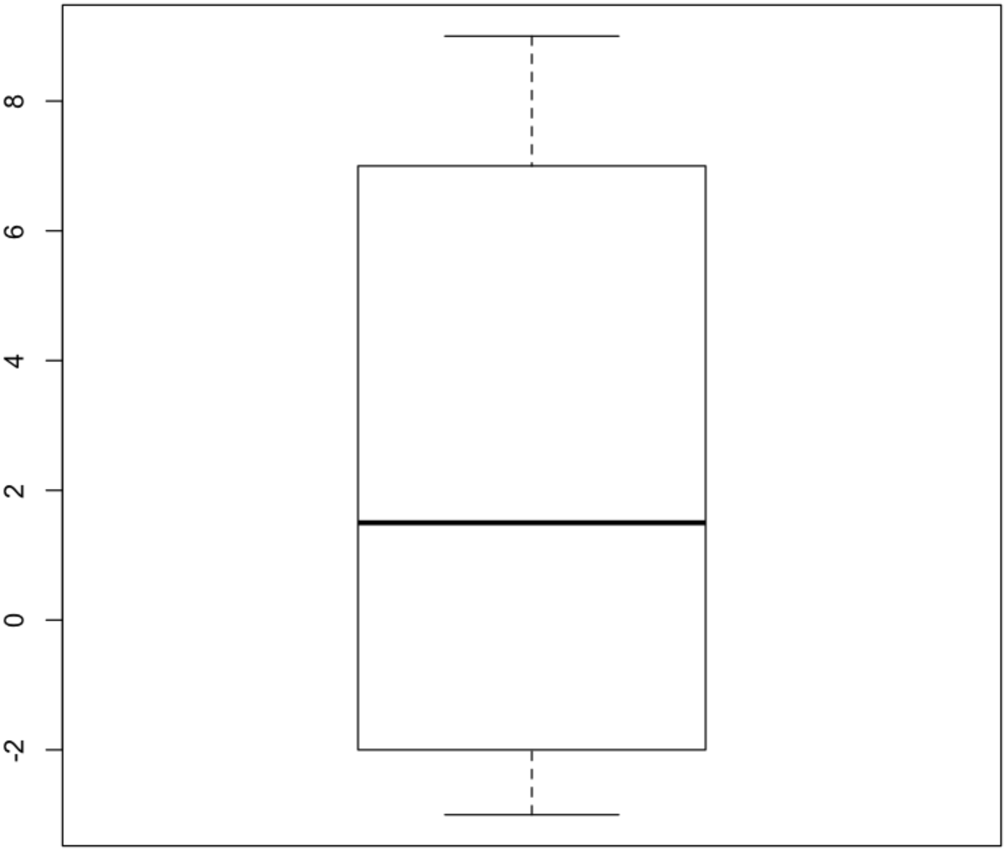
- some sort of bell curve with this value placed in context to see how unlikely it is, or
- an error bar plot with (0 +- 2 * error) error bars and with this observed value within the error bars, or
- or a box and whiskers plot showing that 0 is not all that unlikely

or something similar that shows that this result is not that unlikely.

For example a screenshot of the bell curve from http://onlinestatbook.com/2/calculators/normal_dist.html (http://onlinestatbook.com/2/calculators/normal_dist.html) with the appropriate numbers, as we have seen in some of the homeworks.

In [70]:

```
# Here's one simple plot : a box-and-whisker's plot of the weight gain.
# Since the y value of 0 (the null hypothesis) is inside the box,
# it is not particularly unlikely.
#
# (A spiffier plot would have a horizontal line added at y=0)
# See for example http://stackoverflow.com/questions/6181155/r-draw-a-line-on-the-same-box
plot-graph
# which is one result from googling "R superimpose line on box plot".
# I didn't take the time to play with that.)
#
# The problem with this presentation is that it does not show the 1/sqrt(N)
# narrowing of the range of uncertainty of the mean.
#

boxplot(gain)
```



6

A town has two candidates running for office, John Smith and Jane Doe. Describe how you would design and implement a survey to see if John is more popular than Jane. Invent some reasonable data that you might get, draw a conclusion from that data, and explain what you did and why you reached that conclusion. As usual, a plot of some sort wouldn't hurt. What assumptions are you making, and how might you check them? What can you say about the probability that your conclusion is incorrect?

Now that I'm doing this, I think I probably put in too many of these sort of problems - this one seems like more of the same.

There are a number of reasonable answers to this one - I left it pretty open.

Here's one answer.

```
* Create a written survey asking for people to check (a) John, (b) Jane, (c) no preference.
* Assuming a town like Marlboro, put the survey in the town post office and ask for responses.
```

Getting a large random sample is always tricky. The procedure above will bias towards people with stronger opinions and/or willing to take time to fill out the survey, and assumes "post office visitors" = "registered votors". This approach isn't perfect but is simple.

I will analyze the data by counting a kind of john_popularity

```
* "john"    = +1
* "jane"    = -1
* "neither  =  0
```

The null hypothesis is that the value is 0.0, and the motivated hypothesis is (one sided) > 0, and the significance cutoff is 0.05.

Invent some data :

```
* N = 40 people respond to survey
* john = 20, jane = 10, neither = 10
```

The the analysis is follows the pattern of the lifter's problem.

In [69]:

```
# Here are the votes : 20 for John (1), 10 for neither (0), 10 for jane (-1)
popularity = c(1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,0,0,0,0,0,0,0,0,0,0,-1,-1,-1,-1,-1,
-1,-1,-1,-1,-1)

# Same procedure as several problems previously :

N_pop = length(popularity)
H0 = 0
avg_pop = mean(popularity)
sigma_pop = sd(popularity)
error_pop = sigma_pop/sqrt(N_pop)  # standard error
z_pop = (avg_pop - H0)/error_pop
p_pop = 1 - pnorm(z_pop)
p_pop
```

Out[69]:

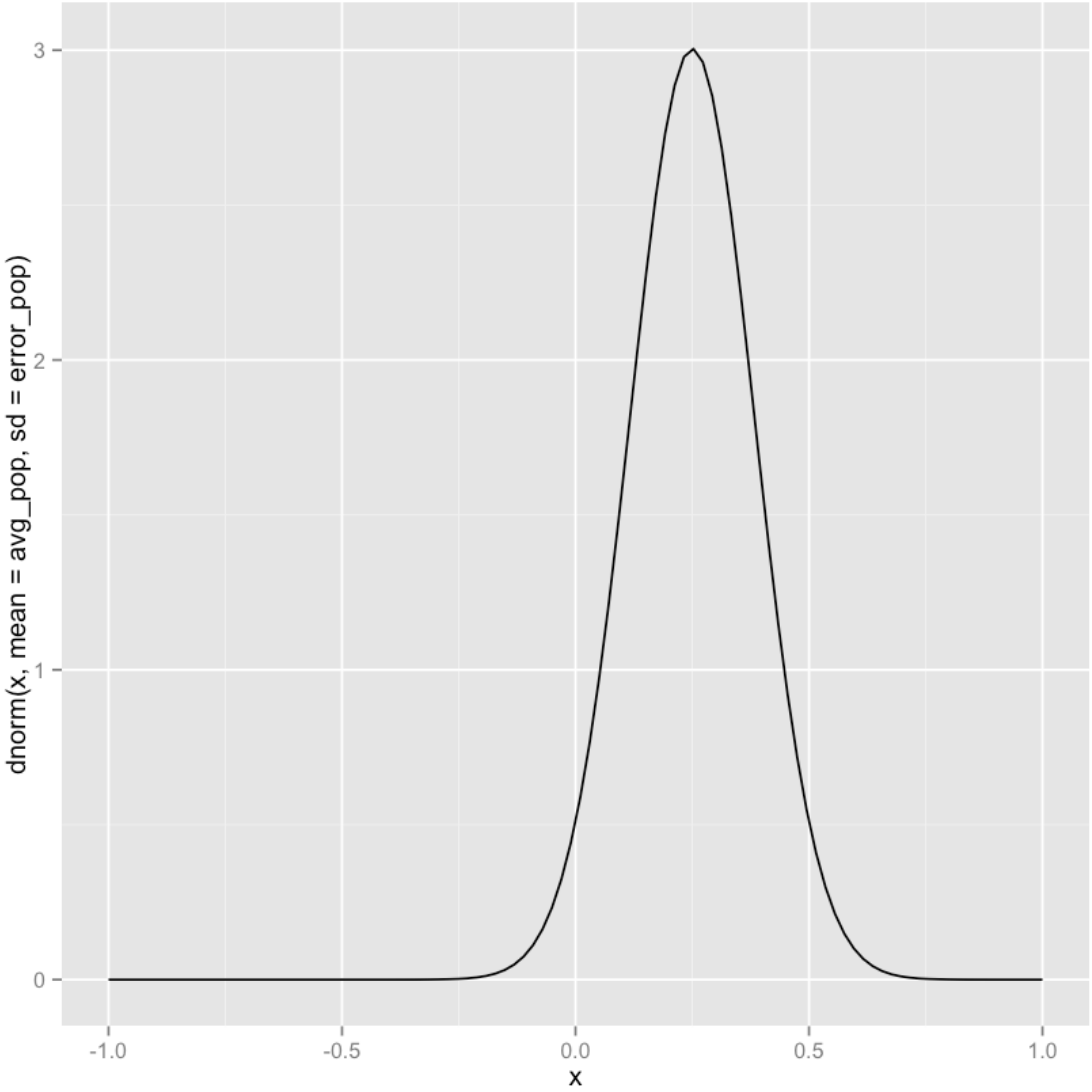
```
0.0298544027370304
```

Since 0.03 < 0.05, we reject the null hypothis and conclude that John is, indeed, more popular.

If we're wrong then we've made a type 1 error : then the null hypothisis would be true (same popularity) and the probability of that is the pvalue, 3%.

In [81]:

```
# Here's one plot using R's dnorm() function to plot the normal distribution.
# I've put the peak at the measure value (avg_pop) and the width
# to be the error of the that measurement. The null hypothesis H0 is at 0,
# so this gives a visualization of how unlikely 0 is. The pvalue is the
# area to the left of 0 under the curve.

x=seq(-1,1,length=100)
qplot(x, dnorm(x, mean=avg_pop, sd=error_pop), geom="line")

# Making something like this with one of the online normal distribution tools
# would have been OK too.
```


dnorm(x, mean = avg_pop, sd = error_pop)
x