How to Use R Studio in AP Statistics

HATS

Cathy Poliak, Ph.D.
cathy@math.uh.edu
Office Fleming 11c

Department of Mathematics
University of Houston

December 3, 2016

UNIVERSITY of HOUSTON
DEPARTMENT OF MATHEMATICS

Outline

UNIVERSITY of HOUSTON
DEPARTMENT OF MATHEMATICS

R and R-studio

- Open source software (free) for statistical analysis.
- R download: `https://cran.cnr.berkeley.edu/`
- R-studio download: `https://www.rstudio.com/products/rstudio/download/`
- Help in R-studio: Right hand bottom panel.

UNIVERSITY of HOUSTON
DEPARTMENT OF MATHEMATICS

How To Input Data

- Preloaded data
- Packages: Mosaic data
- Text file: You can input a data set into Excel then save as a Text (Tab delimited *.txt) file. Then open the dataset in R using the Tools » Import Dataset » From Textfile
- Directly into R: x=c(1,5,6,10)
- Examples to download
  - Grades: `https://www.math.uh.edu/~cathy/Math3339/data/grades.txt`
  - ERA: `https://www.math.uh.edu/~cathy/Math3339/data/Era.txt`
  - Stress: `https://www.math.uh.edu/~cathy/Math3339/data/Stress.txt`

UNIVERSITY of HOUSTON
DEPARTMENT OF MATHEMATICS

Example for Basic Statistics

```
scores=c(8,12,17,22,40,43,49,51,67,68,68,72,75,80,81,
83,84,84,84,85,87,90,91,92,93,98,101,101,101,104)
mean(scores)
[1] 71.03333
median(scores)
[1] 82
> sd(scores)
[1] 28.12225
> var(scores)
[1] 790.8609
> fivenum(scores)
[1]   8  51  82  91 104
```

UNIVERSITY of HOUSTON
DEPARTMENT OF MATHEMATICS

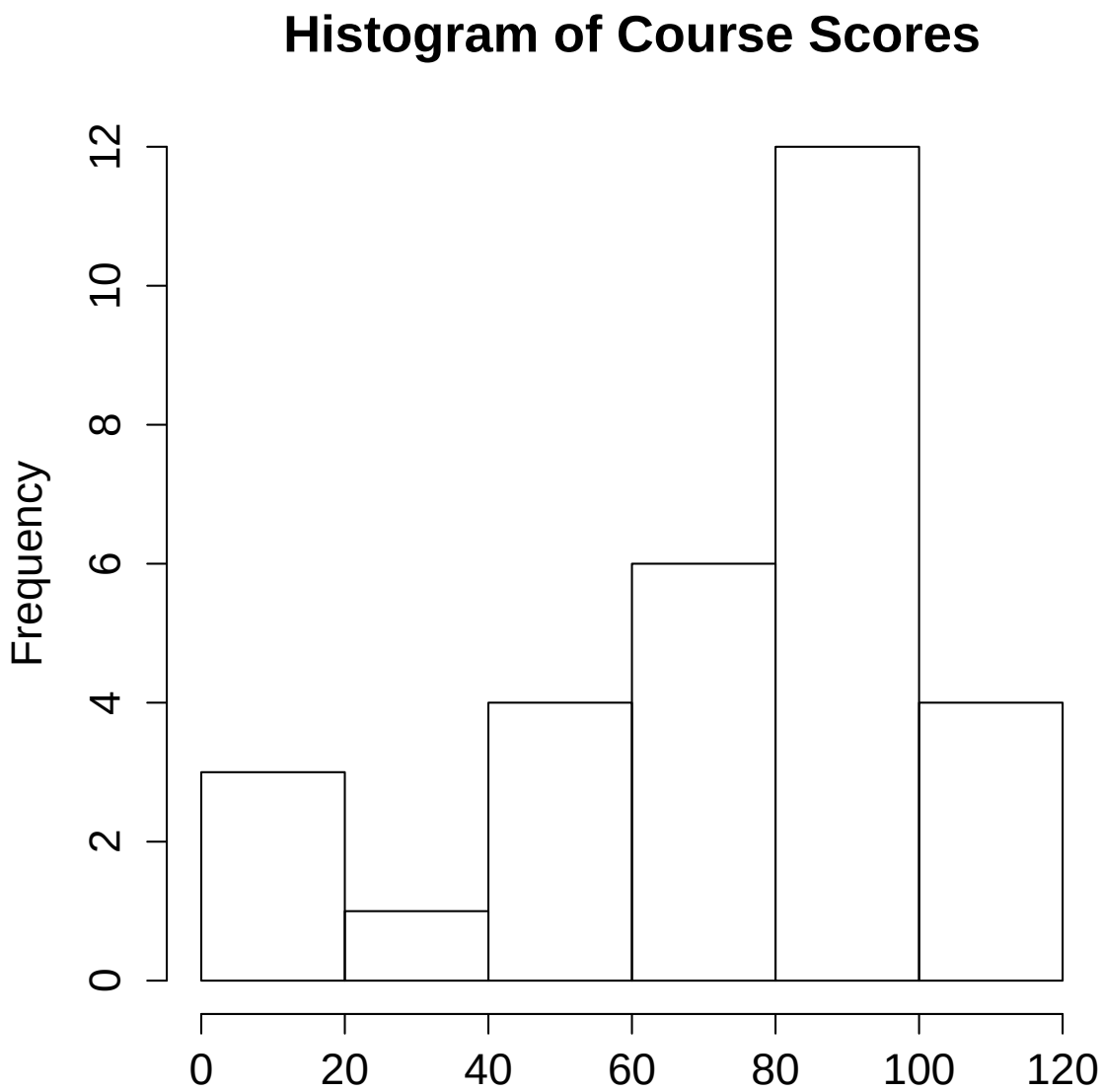
Stem-and-Leaf Plot

```
> stem(scores)

  The decimal point is 1 digit(s) to the right of the |

   0 | 8
   1 | 27
   2 | 2
   3 |
   4 | 039
   5 | 1
   6 | 788
   7 | 25
   8 | 01344457
   9 | 01238
  10 | 1114
```

UNIVERSITY of HOUSTON
DEPARTMENT OF MATHEMATICS

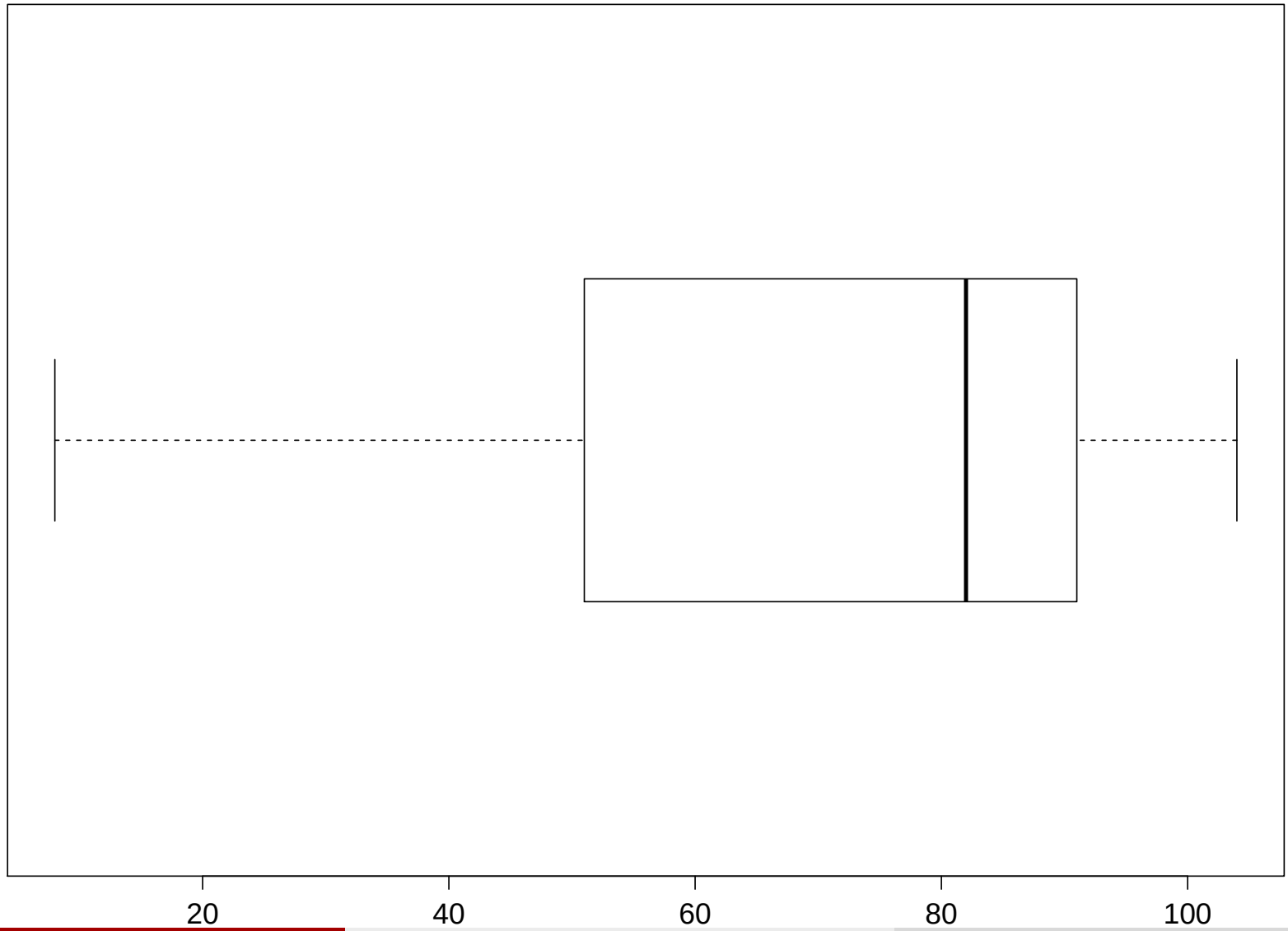
Histogram

hist(scores)

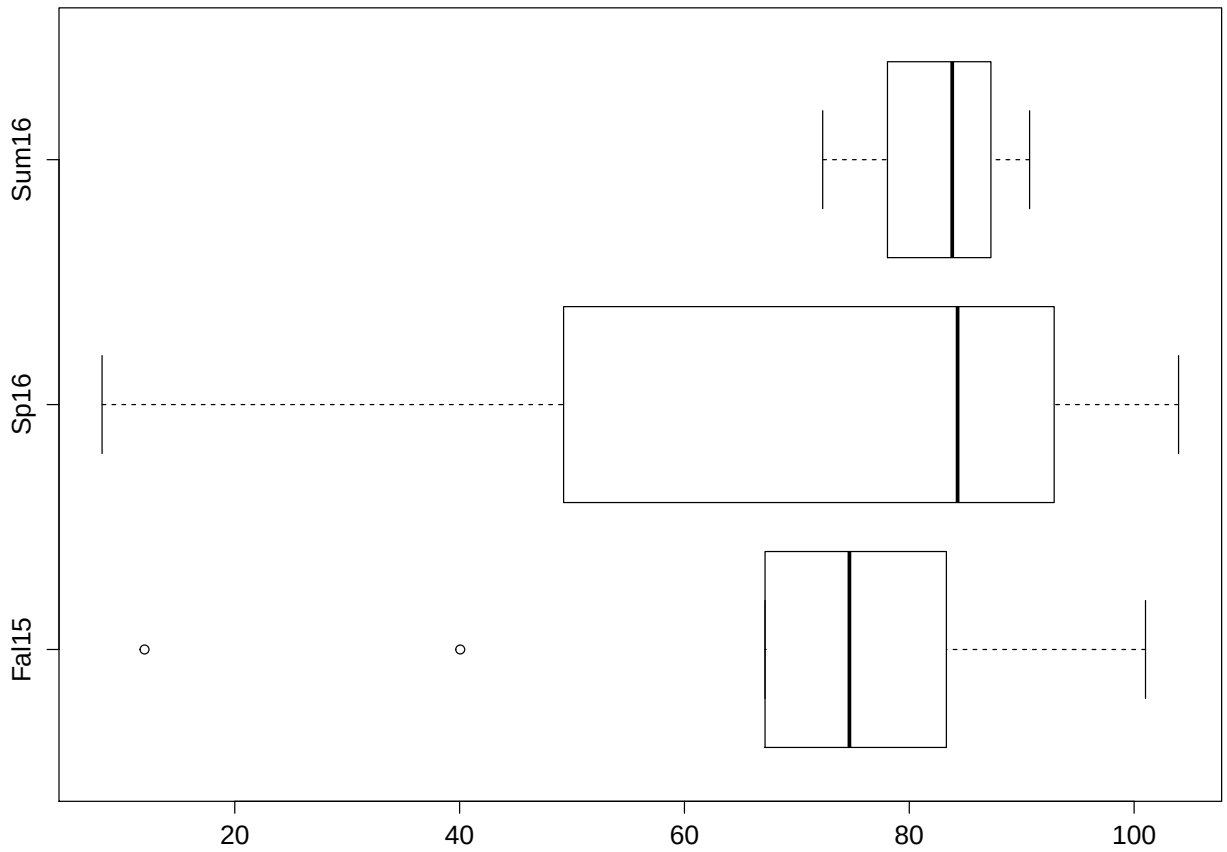


UNIVERSITY of HOUSTON
DEPARTMENT OF MATHEMATICS

Boxplot of Course Scores



Boxplot of Course Scores by Session

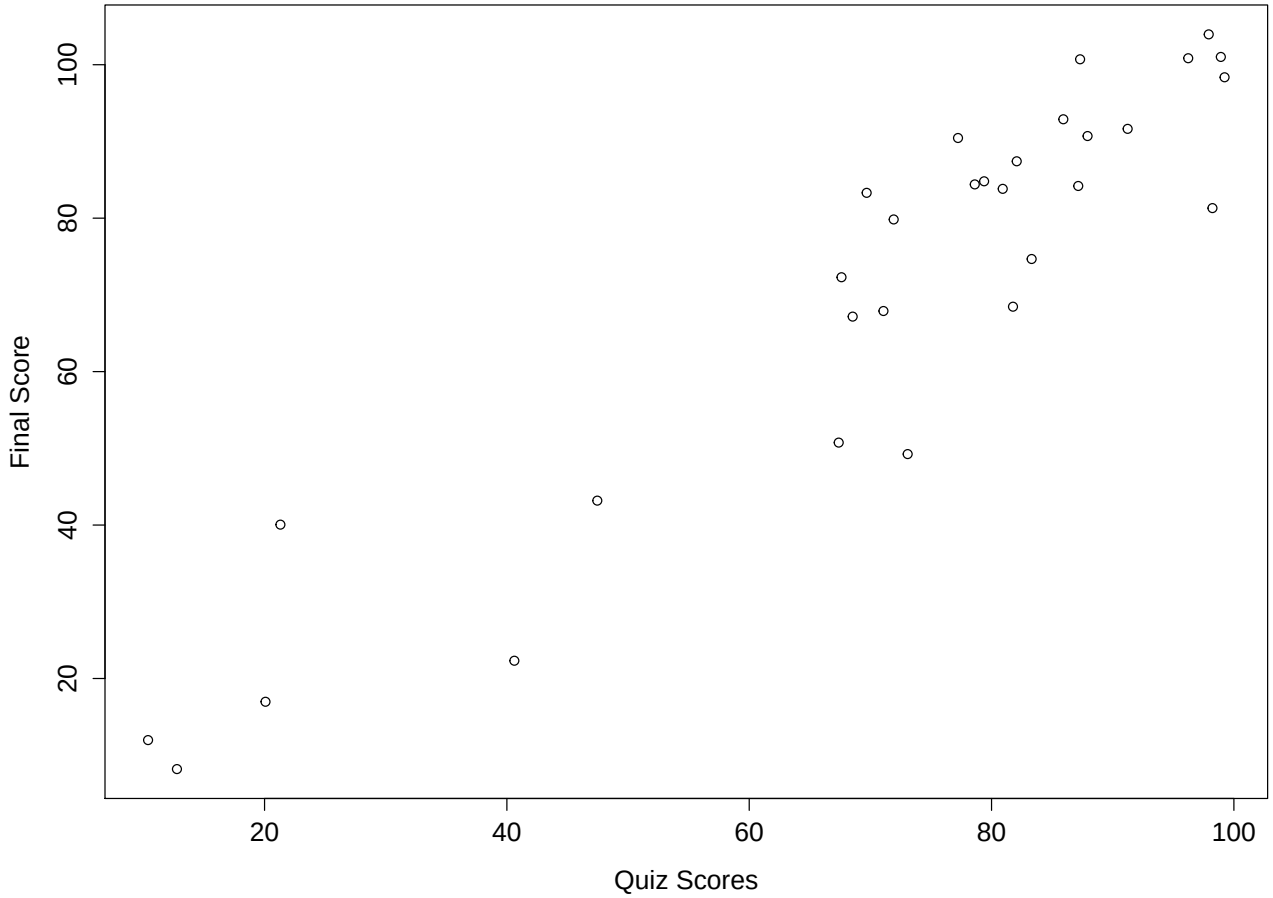



```
boxplot(grades$Score~grades$Session,horizontal=TRUE)
```

UNIVERSITY of HOUSTON
DEPARTMENT OF MATHEMATICS

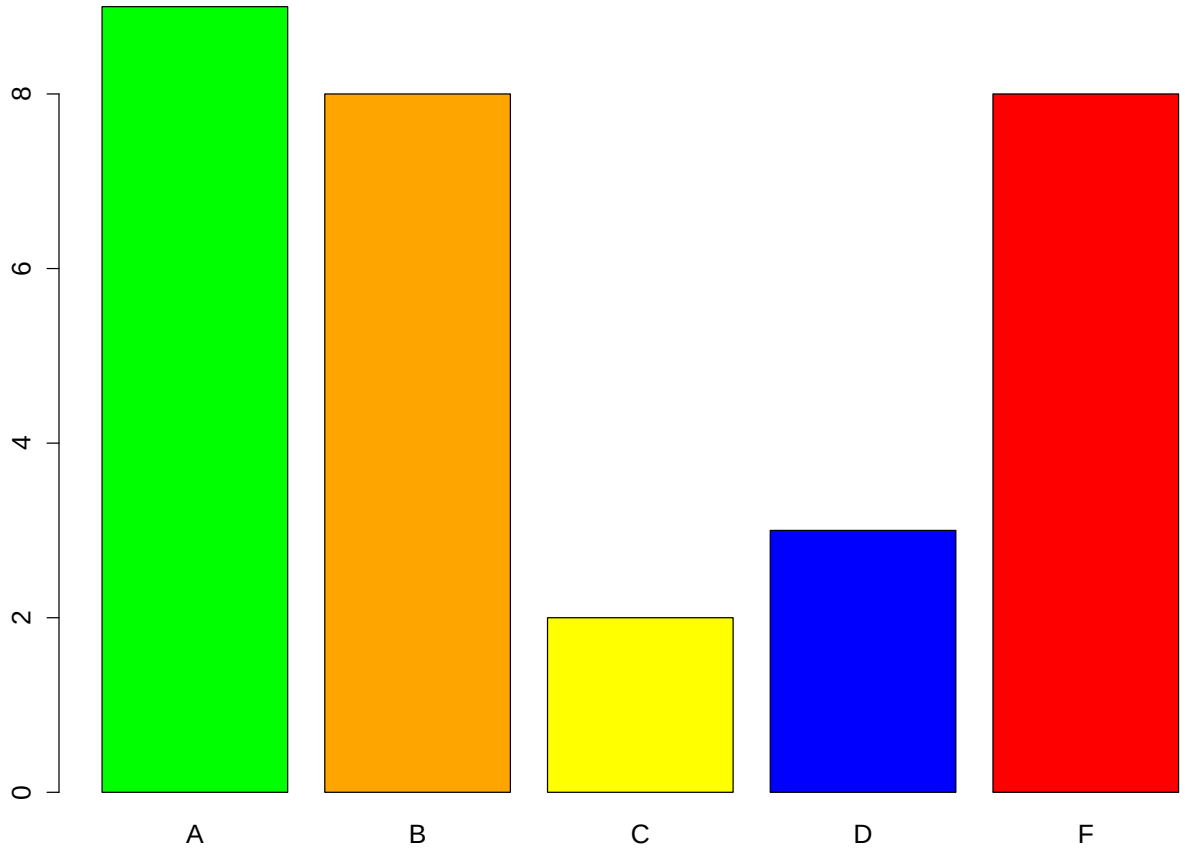
Scatterplot

plot(grades$Quiz,grades$Score,xlab="Quiz Scores",ylab="Final Score")



Bar Graphs

```
plot(grades$Grade,col=c("green","orange","yellow","blue","red"))
```



UNIVERSITY of HOUSTON
DEPARTMENT OF MATHEMATICS

Finding Probabilities for Popular Distributions

- For any "named" distribution we can use R to find the probabilities and the quantiles.
- To find $P(X = x) = d \ldots$ (x,list of parameters).
- To find $P(X \leq x) = p \ldots$ (x,list of parameters).
- To find c such that $P(X \leq c) = p$, $c = q \ldots$ (p,list of parameters).

UNIVERSITY of HOUSTON
DEPARTMENT OF MATHEMATICS

Binomial Distribution

Suppose that in a large metropolitan area, 80% of all households have a flat screen television. Suppose you are interested in selecting a group of six households from this area. Let X be the number of households in a group of six households from this area that have a flat screen television.

1. For what proportion of groups will exactly four of the six households have a flat screen television?

```
> dbinom(4,6,.8)
[1] 0.24576
```

2. For what proportion of groups will at most two of the households have a flat screen television?

```
> pbinom(2,6,.8)
[1] 0.01696
```

3. What is the probability that between 2 and 4 inclusive will have a flat screen television?

```
> pbinom(4,6,.8)-pbinom(1,6,.8)
[1] 0.34304
```

UNIVERSITY of HOUSTON
DEPARTMENT OF MATHEMATICS

Normal Distribution

The length of time needed to complete a certain test is normally distributed with mean 77 minutes and standard deviation 11 minutes. Find the probability that it will take between 74 and 80 minutes to complete the test.

```
> pnorm(80,77,11)-pnorm(77,80,11)
[1] 0.2149371
```

UNIVERSITY of HOUSTON
DEPARTMENT OF MATHEMATICS

More Normal Distribution

Part a: Let Z be the standard normal random variable. Calculate the following.

1. P(Z < 2.4) =

```
> pnorm(2.4)
[1] 0.9918025
```

2. P(Z > -1.9) =

```
> 1-pnorm(-1.9)
[1] 0.9712834
```

3. Find c such that P(Z > c) = 0.98

```
> qnorm(0.02)
[1] -2.053749
```

UNIVERSITY of HOUSTON
DEPARTMENT OF MATHEMATICS

More Normal Distribution

Part b: Let X be a normal random variable with a mean of 47 and a standard deviation of 3. Calculate the following.

1. P(X < 50.4) =

```
> pnorm(50.4,47,3)
[1] 0.8714629
```

2. P(43.5 < X < 50.4) =

```
> pnorm(50.4,47,3)-pnorm(43.5,47,3)
[1] 0.7497903
```

3. Find x such that P(X < x) = 0.74

```
> qnorm(0.74,47,3)
[1] 48.93004
```

UNIVERSITY of HOUSTON
DEPARTMENT OF MATHEMATICS

Sampling Distribution of $\bar{X}$

1. Suppose a random sample of 70 measurements is selected from a population with a mean of 35 and a variance of 300. Determine the mean and standard error of $\bar{x}$.

$\mu_{\bar{x}} = 35 \qquad SE(\bar{x}) = \sqrt{\frac{300}{70}} = 2.0702$

2. A random sample of 1024 12-ounce cans of fruit nectar is drawn from among all cans produced in a run. Prior experience has shown that the distribution of the contents has a mean of 12 ounces and a standard deviation of 0.12 ounce. What is the probability that the mean contents of the 1024 sample cans is less than 11.994 ounces?

```
> pnorm(11.994,12,.12/sqrt(1024))
[1] 0.05479929
```

UNIVERSITY of HOUSTON
DEPARTMENT OF MATHEMATICS

Sampling Distribution of $\hat{p}$

1. In a large population, 67% of the households have cable tv. A simple random sample of 256 households is to be contacted and the sample proportion computed. What is the mean and standard deviation (standard error) of the sampling distribution of the sample proportions?

$\mu_{\hat{p}} = 0.67$ $SE(\hat{p}) =$

```
> sqrt(.67*.33/256)
[1] 0.02938829
```

2. What is the probability that the sampling distribution of sample proportions is less than 73%?

```
> pnorm(.73,.67,sqrt(.67*.33/256))
[1] 0.9794058
```

UNIVERSITY of HOUSTON
DEPARTMENT OF MATHEMATICS

Confidence Intervals

1. A random sample of 64 observations produced a mean value of 73 and standard deviation of 6.5. Determine a 90% confidence interval for the population mean μ.

```
> 73+c(-1,1)*qnorm(1.9/2)*6.5/sqrt(64)
[1] 71.66356 74.33644
```

2. A random sample of 121 observations produced a sample proportion 35%. Determine an approximate 95% confidence interval for the population proportion.

```
> 0.35+c(-1,1)*qnorm(1.95/2)*sqrt(.35*.65/121)
[1] 0.2650143 0.4349857
```

UNIVERSITY of HOUSTON
DEPARTMENT OF MATHEMATICS

How good is a Pitcher for MLB?

- In MLB is the number of wins is attributed to the starting pitcher. Also, the ERA (earned run average) is calculated for the pitcher. Can we use ERA to predict the number of wins that is attributed to a pitcher?
- The following data is from the 2015 baseball season: `https://www.math.uh.edu/~cathy/Math3339/data/Era.txt`
- We will use R to:
  - Construct a scatterplot.
  - Find the LSRL and fit it to the scatterplot.
  - Find r and r^2.
  - Does there appear to be a linear relationship between the two variables? Based on what you found, would you characterized the relationship as positive or negative? Strong or weak?
  - Draw the residual plot.
  - What does the residual plot reveal?
  - `http://insider.espn.com/mlb/insider/story/_/id/13752413/atlanta-braves-pitcher-shelby-miller-terrible-luck-`

UNIVERSITY of HOUSTON
DEPARTMENT OF MATHEMATICS

Stress

A study was conducted to examine the effect of pets in stressful situations. Fifteen subjects were randomly assigned to each of three groups to do a stressful task alone (the control group), with a good friend present, or with their dog present. The subject's mean heart rate (in beats per minutes) during the task is one measure of the effect of stress. The data has is the mean heart rates during stress with a pet (P), with a friend (F) and for the control group (C).

- Make a side by side box plot of the heart rates by the three groups. To do this in R use: plot(Rate Group,data=Stress)
- Does the data suggest that there is a difference among the three groups?
- If there seems to be a difference, complete a Bonferroni pairwise test to determine which or if all the means are different from each other.

UNIVERSITY of HOUSTON
DEPARTMENT OF MATHEMATICS

UNIVERSITY of HOUSTON
DEPARTMENT OF MATHEMATICS

One-Sample T-test

Quart cartons of milk should contain at least 32 ounces. A sample of 22 cartons contained the following amounts in ounces. Does sufficient evidence exist to conclude the mean amount of milk in cartons is less than 32 ounces? The data is: (31.5, 32.2, 31.9, 31.8, 31.7, 32.1, 31.5, 31.6, 32.4, 31.6, 31.8, 32.2, 32.1, 31.8, 31.6, 32.0, 31.6, 31.7, 32.0, 31.9, 31.8, 31.6) $H_0: \mu = 32$ $H_A: \mu < 32$

```
>
milk=c(31.5,32.2,31.9,31.8,31.7,32.1,31.5,31.6,32.4,31.6,31.8,32.2,32.1,31.8,31.6,32,31.6,31.7,3
2,31.9,31.8,31,6)
> t.test(milk,mu=32,alternative="less")

 One Sample t-test

data:  milk
t = -3.0488, df = 21, p-value = 0.00305
alternative hypothesis: true mean is less than 32
95 percent confidence interval:
    -Inf 31.92872
sample estimates:
mean of x
 31.83636
```

UNIVERSITY of HOUSTON
DEPARTMENT OF MATHEMATICS

Two-sample T-test

Is there a difference in the mean miles per gallon of a Honda Civic and a Toyota Prius? The following is data from 5 Honda's and 6 Toyota's:

Honda	32.2	29.8	29.7	29.7	28.1	
Toyota	36.5	33	33	31.7	31	28.8

```
> t.test(c(32.2,29.8,29.7,29.7,28.1),c(36.5,33,33,31.7,31,28.8))

 Welch Two Sample t-test

data:  c(32.2, 29.8, 29.7, 29.7, 28.1) and c(36.5, 33, 33, 31.7, 31, 28.8)
t = -1.9684, df = 8.1315, p-value = 0.08396
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 -5.2759386  0.4092719
sample estimates:
mean of x mean of y
 29.90000  32.33333
```

UNIVERSITY of HOUSTON
DEPARTMENT OF MATHEMATICS

Matched Pair Test

In a experiment on relaxation techniques, subject's brain signals were measured before and after the relaxation exercises with the following results:

Person	1	2	3	4	5
Before	32	38	65	50	30
After	25	35	56	52	24

Is there sufficient evidence to suggest that the relaxation exercise slowed the brain waves? Assume the population is normally distributed.

UNIVERSITY of HOUSTON
DEPARTMENT OF MATHEMATICS

```
> t.test(c(32,38,65,50,30),c(25,35,56,52,24),alternative="greater",paired = T)

 Paired t-test

data:  c(32, 38, 65, 50, 30) and c(25, 35, 56, 52, 24)
t = 2.4045, df = 4, p-value =0.037
alternative hypothesis: true difference in means is greater than 0
95 percent confidence interval:
 0.521537      Inf
sample estimates:
mean of the differences
                    4.6
```

Chi-Square Test

The Blue Diamond Company advertises that their nut mix contains (by weight) 40% cashews, 15% Brazil nuts, 20% almonds and only 25% peanuts. The truth-in-advertising investigators took a random sample (of size 20 lbs) of the nut mix and found the distribution to be as follows: 6 lbs of Cashews, 3 lbs of Brazil nuts, 5 lbs of Almonds and 6 lbs of Peanuts. At the 0.01 level of significance, is the claim made by Blue Diamond true?

1. Calculate the test statistic for this test.

2. Determine the p-value.

3. Give the decision to Reject H_0 or Fail to Reject H_0.

UNIVERSITY of HOUSTON
DEPARTMENT OF MATHEMATICS

```
> chisq.test(c(6,3,5,6),p=c(.4,.15,.2,.25),correct=F)

 Chi-squared test for given
 probabilities

data:  c(6, 3, 5, 6)
X-squared = 0.95, df = 3, p-value = 0.8133

Warning message:
In chisq.test(c(6, 3, 5, 6), p = c(0.4, 0.15, 0.2, 0.25), correct = F) :
  Chi-squared approximation may be incorrect
```

UNIVERSITY of HOUSTON
DEPARTMENT OF MATHEMATICS

Fair Die

A six-sided die is thrown 50 times. The numbers of occurrences of each face are shown below.

Face	1	2	3	4	5	6
Count	12	5	9	11	6	7

Can you conclude that the die is not fair?

```
> chisq.test(c(12,5,9,11,6,7),correct=F)

 Chi-squared test for given
 probabilities

data:  c(12, 5, 9, 11, 6, 7)
X-squared = 4.72, df = 5,
p-value = 0.451
```

UNIVERSITY of HOUSTON
DEPARTMENT OF MATHEMATICS

UNIVERSITY of HOUSTON
DEPARTMENT OF MATHEMATICS

Example

The following table shows three different airlines **row variable** and the number of delayed or on-time flights **column variable** from flightstats.com.

	Delayed	On-time	Total
American	112	843	955
Southwest	114	1416	1530
United	61	896	957
Total	287	3155	3442

- Does on-time performance depend on airline?
- We will use a significance test to answer this question.

UNIVERSITY of HOUSTON
DEPARTMENT OF MATHEMATICS

Chi-square Test Using R

1. Input the data as a matrix.
2. R-code: chisq.test(matrix name,correction=FALSE)

```
> airline<-matrix(c(112,114,61,843,1416,896),nrow=3,ncol=2)
> chisq.test(airline,correct = FALSE)

Pearson's Chi-squared test

data:  airline
X-squared = 20.762, df = 2, p-value =3.102e-05
```

UNIVERSITY of HOUSTON
DEPARTMENT OF MATHEMATICS