

1 Introduction

Example: California's 2003 Recall Election, with 7,974,834 voters.

In California in October 2003, a special election was held to determine whether Governor Gray Davis should be recalled from office. In the same election, voters were asked to vote for the candidate who would replace Davis as governor if he were recalled. The four candidates included the Hollywood actor Arnold Schwarzenegger. If a majority voted yes for the recall, the candidate who received the most votes would be the new governor.

The exit poll on which TV networks relied for their projections found from sampling 3160 voters (0.040% of voters):

Recall Governor Davis (54%),

Do not recall Governor Davis (46%)

Is it reasonable to conclude that Governor Davis would be recalled, based on that exit poll?

□

Why do we conduct surveys?

What is a **population**?

What is a **sample**?

Sample surveys (Math 325) differs from **experimental design** (Math 321).

What are some useful examples of surveys?

Why is the United States Census important?

Why does the U.S. Bureau of Labor Statistics conduct surveys?

Example: Quotes from Shere Hite's book *Women and Love: A Cultural Revolution in Progress* (1987):

- (a) 84% of women are "not satisfied emotionally with their relationships" (p. 804).
- (b) 70% of all women "married five or more years are having sex outside of their marriages" (p. 856).
- (c) 95% of women "report forms of emotional and psychological harassment from men with whom they are in love relationships" (p. 810).
- (d) 84% of women "report forms of condescension from the men in their love relationships" (p. 809).

Example: News statement in 1993: "Twenty-two percent of Americans doubt that the Holocaust ever occurred."

Question from the Roper organization: "Does it seem possible or does it seem impossible to you that the Nazi extermination of the Jews never happened?"

Responses were: 22% said "it seemed possible," 12% said they did not know, 65% said it was "impossible it never happened."

Later, question from the Gallup organization: "The term *Holocaust* usually refers to the killing of millions of Jews in Nazi death camps during World War II. In your opinion, did the Holocaust: definitely happen, probably happen, probably did not happen, or definitely not happen?"

Responses were: 83% said Holocaust definitely happened, 13% said Holocaust probably happened, 1% said Holocaust definitely did not happen.

“When statistics are not based on strictly accurate calculations, they mislead instead of guide. The mind easily lets itself be taken in by the false appearance of exactitude which statistics retain in their mistakes, and confidently adopts errors clothed in the form of mathematical truth.” – Alexis de Tocqueville, *Democracy in America*

Using R or Splus {see www.r-project.org}

To download *R*: Click “CRAN”, “UNC-Chapel Hill” (or some other site), “Windows”, “base”, “Download R 2.9.1 for Windows”. Click “Save” and then “Save” again. The saving should take less than five minutes using high-speed internet. Click “Open” and “OK”, and continue to click “Next” until the process is complete.

Alternatively, use *Rweb* by going to <http://bayes.math.montana.edu/Rweb/Rweb.general.html>

```
> 11:20 # Generate numbers from 11 to 20
> c(11:20) # Also, generate numbers from 11 to 20. “c” means combine.
> x = c(11:20) # Save numbers from 11 to 20 as the variable x.

> x # List contents of x.
> x[4]
> x[4:7]
> y = 3 * x^2 + 5
> y
> plot( x, y )
> c(5:-3, 10:15, 20, 6)
> ls( ) # Lists all variables.
```

Example: Consider data in exercise 4.40 on p. 111.

```
> x1 = scan( "d:/SAS/DATASETS.SAS/EXER4_40.DAT" ) # Save the data as x1.
```

```
> x1 = scan( "http://www.math.jmu.edu/~garrenst/math325.dir/datasets/EXER4_40.DAT" )  
# Scan the data from the website.
```

```
> x1 # Note that data are listed as a vector.
```

```
> # 'x1' may be converted to a matrix using the 'matrix' command, but we will skip  
this detail.
```

```
> x1 = read.table( "d:/SAS/DATASETS.SAS/EXER4_40.DAT" ) # Save the data as  
the matrix x1.
```

```
> x1 = read.table( "http://www.math.jmu.edu/~garrenst/math325.dir/datasets/EXER4_40.DAT" )  
# Save the data as the matrix x1.
```

```
> help( "scan" )
```

```
> ?scan
```

```
> source("http://www.math.jmu.edu/~garrenst/math325.dir/Rmacros")
```

Look at 'ls' and 'scan2'.

```
> x.mat[5:10, 2] # To view the audited amount for items 5 to 10.
```

```
> # Now, view only the odd numbered items.
```

```
> odd = (0 : 7) * 2 + 1
```

```
> odd
```

```
> x.mat[odd, ]
```

```
> # Commands use ( ), and variables use [ ].

> colnames(x.mat) = c("item number", "audited amount", "recorded amount", "overstatement")
> x.mat

> # Compute the mean recorded amount.

> x2 = x.mat[, 3]; mean(x2) # Computes the mean recorded amount.
> x3 = apply(x.mat, 2, mean)

> x3 ; x3["recorded amount"] ; x3[3]
> # Suppose a penalty equal to double the overstatement is applied.
> # Append this new variable.
> penalty = 2 * x.mat[, "overstatement"]
> penalty
> x2.mat = cbind(x.mat, penalty)
> x2.mat
> apply( x2.mat, 2, mean ) # Computes the mean for all 5 categories.

> # Using only one or two commands in R, in the 'x.mat' matrix, change the 9th
  recorded amount from 82 to 62, and then fix the overstatement as well.
```

Example: Read in the following data set regarding income (in thousands of dollars): {48, 75, 93, NA, 81, 53}. Compute the mean and standard deviation of this data set, while discarding all values of NA.

> # 'na.rm' also works with the 'apply' command, as in the example using CARS93.DAT and the homework exercise using CLASSSUR.DAT below.

Example: Consider data from CARS93.DAT in Appendix D, pp. 442-445. Compute the average city mpg and the mean number of cylinders among all 57 cars.

> ##### Use 'scan2("CARS93.DAT", T)'.

> ##### Use 'scan2("CARS93.DAT", T, T)'.

> ##### Use 'read.table'.

> ##### Use 'read.table2'.

Continue to work with this data set.

> mpg = y.mat[, 7:8] # for city and highway miles per gallon

> colnames(mpg) = c("city", "highway")

```
> mpgcity = mpg[, 1] # mpg of city

> # Compute the average city mpg among all 57 cars, again!

> # Take a simple random sample of five values of mpgcity withOUT replacement.

> # Suppose for every gallon of gas used for city travel, a driver uses one gallon of gas
  for highway travel.

> # Determine the overall miles per gallon for such a driver, for each car.

> # Again, determine the overall miles per gallon for such a driver, for each car, this
  time using the macro "apply".

> # For each of these numerical variables, compute the average among all of the 57
  cars.

> # Append these means to bottom of the matrix.
```

Example: Enter the data from Table 2.4 on p. 15 into a matrix, and label the rows and columns.

```
> history( ) # View the history.
```

```
> q( ) # Quits R.
```

Homework C1.1: Using the CLASSSUR.DAT data set from appendix D, pp. 440-442, and using *R*, list your source code and your output.

- (a) Scan all of the data into a matrix.
- (b) Using the “apply” command, compute the mean, sum, standard deviation, and variance of each column in the matrix, while discarding all values of NA. (Note that, say, the standard deviation of gender, is not all that meaningful, but don’t worry about that!)
- (c) Attach your answers from part (b) to the bottom of the matrix.
- (d) Print your final matrix.
- (e) Take a simple random sample of 20 GPAs withOUT replacement.
- (f) Take a simple random sample of 100 GPAs WITH replacement.
- (g) Display the height and weight of students #11 through #20 only, in a matrix.
- (h) Add 2 inches to the height of student #12 only, using only one command in *R*.
- (i) List ‘height’ in centimeters, rather than inches, and display the height and weight of students #11 through #20 only, in a matrix.

End of Homework C1.1. □

Homework C1.2: Using the table in exercise 5.29 on p. 165, save the data as a matrix in *R*, provide appropriate row and column names in *R*, and print the

matrix with row and column names in R .

End of Homework C1.2. \square