Chapter 1

Statistics and the Scientific Method

1.1

a. The population of interest is all salmon released from fish farms located in Norway.
b. The samples are the two batches of salmon released (1,996 and 2,499 in northern and southern Norway, respectively).
c. The migration pattern and survival of salmon released from fish farms.
d. Since the sample is only a small proportion of the whole population, it is necessary to evaluate what the mean weight may be for any other random selection of farmed salmon.

1.2

a. All private water wells.
b. The 100 private water wells in or near the Barnett Shale in Texas.
c. The level of contaminants in the water wells.
d. We want to relate the level of contaminants of the 100 points in the sample to the level in the whole suspect area. Thus we need to know how accurate a portrayal of the population is provided by the 100 points in the sample.

1.3

a. All families that have had option of SNAP (food stamps).
b. 60,782 examined over the time period of 1968 to 2009.
c. Adult health and economic outcomes (specifically, the incidence of metabolic health outcomes and economic self-sufficiency).
d. In order to evaluate how closely the sample families represent the American population over this time period.

1.4

a. All head impacts resulting from playing football over a given period of time.
b. The 1,281,444 head impacts recorded.
c. The number (or percent) of concussions suffered through these impacts.
d. The advances in tackling techniques imply that there is variability in how a tackle is performed. We need to see if our sample was representative of the hits that may be sustained.

1.5

a. The population of interest is the population of those who would vote in the 2004 senatorial campaign.
b. The population from which the sample was selected is registered voters in this state.
c. The sample will adequately represent the population, unless there is a difference between registered voters in the state and those who would vote in the 2004 senatorial campaign.
d. The results from a second random sample of 5,000 registered voters will not be exactly the same as the results from the initial sample. Results vary from sample to sample. With either sample we hope that the results will be close to that of the views of the population of interest.

1.6

a. The professor's population of interest is college freshmen at his university.
b. The sampled population is all freshmen enrolled in HIST 101.
c. Yes, there is a major difference in the two populations. Those enrolled in HIST 101 may not accurately reflect the population of all freshmen at his university. For example, they might be more interested in history.
d. Had the professor lectured on the American Revolution, those students in HIST 101 would be more likely to know which country controlled the original 13 states prior to the American Revolution than other freshmen at the university.

Chapter 2

Using Surveys and Experimental Studies to Gather Data

2.1

a. The explanatory variable is level of alcohol drinking. One possible confounding variable is smoking. Perhaps those who drink more often also tend to smoke more, which would impact incidence of lung cancer. To eliminate the effect of smoking, we could block the experiment into groups (e.g., nonsmokers, light smokers, heavy smokers).
b. The explanatory variable is obesity. Two confounding variables are hypertension and diabetes. Both hypertension and diabetes contribute to coronary problems. To eliminate the effect of these two confounding variables, we could block the experiment into four groups (e.g., hypertension and diabetes, hypertension but no diabetes, diabetes but no hypertension, neither hypertension nor diabetes).

2.2

a. The explanatory variable is the new blood clot medication. The confounding variable is the year in which patients were admitted to the hospital. Because those admitted to the hospital the previous year were not given the new blood clot medication, we cannot be sure that the medication is working or if something else is going on. We can eliminate the effects of this confounding by randomly assigning stroke patients to the new blood clot medication or a placebo.
b. The explanatory variable is the software program. The confounding variable is whether students choose to stay after school for an hour to use the software on the school's computers. Those students who choose to stay after school to use the software on the school's computers may differ in some way from those students who do not choose to do so, and that difference may relate to their mathematical abilities. To eliminate the effect of the confounding variable, we could randomly assign some students to use the software on the school's computers during class time and the rest to stay in class and learn in a more traditional way.

2.3 Possible confounding factors include student-teacher ratios, expenditures per pupil, previous mathematics preparation, and access to technology in the inner city schools. Adding advanced mathematics courses to inner city schools will not solve the discrepancy between minority students and white students, since there are other factors at work.

2.4 There may be a difference in student-teacher ratios, expenditures per pupil, and previous preparation between the schools that have a foreign language requirement and schools that do not have a foreign language requirement.

2.5 The relative merits of the different types of sampling units depends on the availability of a sampling frame for individuals, the desired precision of the estimates from the sample to the population, and the budgetary and time constraints of the project.

2.6 She could conduct a stratified random sample in which the states serve as the stratum. A simple random sample could then be selected within each state. This would provide information concerning the differences between the states along with the individual opinions of the employees.

2.7

a. All residents in the county.
b. All registered voters.
c. Survey nonresponse – those who responded were probably the people with much stronger opinions than those who did not respond, which then makes the responses not representative of the responses of the entire population.

2.8

a. In the first scenario, people would be more willing to lie about using a biodegradable detergent because there is no follow up to verify and individuals usually prefer to appear environmentally conscious. The second survey has a check in place to verify the answers given are truthful.
b. The first survey would likely yield a higher percentage of those who say they use a biodegradable detergent. The second may anger the individuals who tell the truth as if their honesty is being tested.

2.9

a. Alumni (men only?) who graduated from Yale in 1924.
b. No. Alumni whose addresses were on file 25 years later would not necessarily be representative of their class.
c. Alumni who responded to the mail survey would not necessarily be representative of those who were sent the questionnaires. Income figures may not be reported accurately (intentionally), or may be rounded off to the nearest $5,000, say, in a self-administered questionnaire.
d. Rounding income responses would make the figure $25,111 unlikely. The fact that higher income respondents would be more likely to respond (bragging), and the fact that incomes are likely to be exaggerated, would tend to make the estimate too high.

2.10

a. Simple random sampling.
b. Stratified sampling.
c. Cluster sampling.

2.11

a. Simple random sampling.
b. Stratified sampling.
c. Cluster sampling.

2.12

a. Stratified sampling. Stratify by job category and then take a random sample within each job category. Different job categories will use software applications differently, so this sampling strategy will allow us to investigate that.
b. Systematic random sampling. Sample every tenth patient (starting from a randomly selected patient from the first ten patients). Provided that there is no relationship between the type of patient and the order that the patients come into the emergency room, this will give us a representative sample.

2.13

a. Stratified sampling. We should stratify by type of degree and then sample 5% of the alumni within each degree type. This method will allow us to examine the employment status for each degree type and compare among them.

b. Simple random sampling. Once we find 100 containers we will stop. Still it will be difficult to get a completely random sample. However, since we don't know the locations of the containers, it would be difficult to use either a stratified or cluster sample.

2.14

a. Water temperature and Type of hardener
b. Water temperature: 175 °F and 200 °F; Type of hardener: H_1, H_2, H_3
c. Manufacturing plants
d. Plastic pipe
e. Location on Plastic pipe
f. 2 pipes per treatment
g. Covariates: None
h. 6 treatments: (175 °F, H_1), (175 °F, H_2), (175 °F, H_3), (200 °F, H_1), (200 °F, H_2), (200 °F, H_3)

2.15

This is an example where there are two levels of Experimental units, and the analysis is discussed in Chapter 18.

To study the effect of month:

a. Factors: Month
b. Factor levels: 8 levels of month (Oct - May)
c. Block = each section
d. Experimental unit (Whole plot EU) = each tree
e. Measurement unit = each orange
f. Replications = 8 replications of each month
g. Covariates = none
h. Treatments = 8 treatments (Oct – May)

To study the effect of location:

a. Factors: Location
b. Factor levels: 3 levels of location (top, middle, bottom)
c. Block = each section
d. Experimental unit (Split plot EU) = each location tree
e. Measurement unit = each orange
f. Replications = 8 replications of each location
g. Covariates = none
h. Treatments = 3 treatments (top, middle, bottom)

2.16

a. Factors: Type of drug
b. Factor levels: D_1, D_2, Placebo
c. Blocks: Hospitals
d. Experimental units: Wards
e. Measurement units: Patients
f. Replications: 2 wards per drug in each of the 10 hospitals
g. Covariates: None
h. Treatments: D_1, D_2, Placebo

2.17

a. Factors: Type of treatment
b. Factor levels: D_1, D_2, Placebo
c. Blocks: Hospitals, Wards
d. Experimental units: Patients
e. Measurement units: Patients
f. Replications: 2 patients per drug in each of the ward/hospital combinations
g. Covariates: None
h. Treatments: D_1, D_2, Placebo

2.18

a. Factors: Type of school
b. Factor levels: Public; Private – non-parochial; Parochial
c. Blocks: Geographic region
d. Experimental units: Classrooms
e. Measurement units: Students in classrooms
f. Replications: 2 classrooms per each type of school in each of the city/region combinations
g. Covariate: Measure of socio-economic status
h. Treatments: Public; Private – non-parochial; Parochial

2.19

a. Factors: Temperature, Type of seafood
b. Factor levels: Temperature (0 °C, 5 °C, 10 °C); Type of seafood (oysters, mussels)
c. Blocks: None
d. Experimental units: Package of seafood
e. Measurement units: Sample from package
f. Replications: 3 packages per temperature
g. Treatments: (0 °C, oysters), (5 °C, oysters), (10 °C, oysters), (0 °C, mussels), (5 °C, mussels), (10 °C, mussels)

2.20

- Randomized complete block design with blocking variable (10 orange groves) and 48 treatments in a $3 \times 4 \times 4$ factorial structure.
- Experimental Units: Plots
- Measurement Units: Trees

2.21

- Randomized complete block design with blocking variable (10 warehouses) and 5 treatments (5 vendors)

2.22

- Randomized complete block design, where blocked by day
- 2-factor structure (where the factors are type of glaze, and thickness)

2.23

a. Design B. The experimental units are not homogeneous since one group of consumers gives uniformly low scores and another group gives uniformly high scores, no matter what recipe is used.

Using design A, it is possible to have a group of consumers that gives mostly low scores randomly assigned to a particular recipe. This would bias this particular recipe. Using design B, the experimental error would be reduced since each consumer would evaluate each recipe. That is, each consumer is a block and each of the treatments (recipes) is observed in each block. This results in having each recipe subjected to consumers who give low scores and to consumers who give high scores.

b. This would not be a problem for either design. In design A, each of the remaining 4 recipes would still be observed by 20 consumers. In design B, each consumer would still evaluate each of the 4 remaining recipes.

2.24

a. "Employee" should refer to anyone who is eligible for *sick days*.
b. Use payroll records. Stratify by employee categories (full-time, part-time, etc.), employment location (plant, city, etc.), or other relevant subgroup categories. Consider systematic selection within categories.
c. Sex (women more likely to be care givers), age (younger workers less likely to have elderly relatives), whether or not they care for elderly relatives now or anticipate doing so in the near future, how many hours of care they (would) provide (to define "substantial"), etc. The company might want to explore alternative work arrangements, such as flex-time, offering employees 4 ten-hour days, cutting back to 3/4-time to allow more time to care for relatives, etc., or other options that might be mutually beneficial and provide alternatives to taking sick days.

2.25

a. Each state agency and some federal agencies have records of licensed physicians, professional corporations, facility licenses, etc. Professional organizations such as the American Medical Association, American Hospital Administrators Association, etc., may have such lists, but they may not be as complete as licensing records.
b. What nursing specialties are available at this time at the physician's offices or medical facilities? What medical specialties/facilities do they anticipate adding or expanding? What staffing requirements are unfilled at this time or may become available when expansion occurs? What is the growth/expansion time frame?
c. Licensing boards may have this information. Many professional organizations have special categories for members who are unemployed, retired, working in fields not directly related to nursing, students who are continuing their education, etc.
d. Population growth estimates may be available from the Census Bureau, university economic growth research, bank research studies (prevailing and anticipated load patterns), etc. Health risk factors and location information would be available from state health departments, the EPA, epidemiological studies, etc.
e. Licensing information should be stratified by facility type, size, physician's specialty, etc., prior to sampling.

2.26

If phosphorous first: [P,N]

[10,40], [10,50], [10,60], then [20,60], [30,60] or
[20,40], [20,50], [20,60], then [10,60], [30,60] or
[30,40], [30,50], [30,60], then [10,60], [10,60]

If nitrogen first: [N,P]

[40,10], [40,20], [40,30], then [50,30], [60,30] or
[50,10], [50,20], [50,30], then [40,30], [60,30] or
[60,10], [60,20], [60,30], then [40,30], [50,30]

So, for example

Phosphorus	30	30	30	10	20
Nitrogen	40	50	60	60	60
Yield	150	170	190	165	185

Recommendation: Phosphorus at 30 pounds, and Nitrogen at 60 pounds.

2.27

	Factor 2		
Factor 1	I	II	III
A	25	45	65
B	10	30	50

2.28

a. Group dogs by sex and age:

Group	Dog
Young female	2, 7, 13, 14
Young male	3, 5, 6, 16
Old female	1, 9, 10, 11
Old male	4, 8, 12, 15

b. Generate a random permutation of the numbers 1 to 16:
15 7 4 11 3 13 8 1 12 16 2 5 6 10 9 14
Go through the list and the first two numbers that appear in each of the four groups receive treatment L_1 and the other two receive treatment L_2.

Group	Dog-Treatment
Young female	2-L_2, 7-L_1, 13, 14-L_2
Young male	3-L_1, 5-L_2, 6-L_1, 16-L_2
Old female	1-L_1, 9-L_2, 10-L_2, 11-L_1
Old male	4-L_1, 8-L_2, 12-L_2, 15-L_1

2.29

a. Bake one cake from each recipe in the oven at the same time. Repeat this procedure r times. The baking period is a block with the four treatments (recipes) appearing once in each block. The four recipes should be randomly assigned to the four positions, one cake per position. Repeat this procedure r times.
b. If position in the oven is important, then position in the oven is a second blocking factor along with the baking period. Thus, we have a Latin square design. To have $r = 4$, we would need to have each recipe appear in each position exactly once within each of four baking periods. For example:

Period 1		Period 2		Period 3		Period 4	
R_1	R_2	R_4	R_1	R_3	R_4	R_2	R_3
R_3	R_4	R_2	R_3	R_1	R_2	R_4	R_1

c. We now have an incompleteness in the blocking variable period since only four of the five recipes can be observed in each period. In order to achieve some level of balance in the design, we need to select enough periods in order that each recipe appears the same number of times in each period and the same total number of times in the complete experiment. For example, suppose we wanted to observe each recipe $r = 4$ times in the experiment. If would be necessary to have 5 periods in order to observe each recipe 4 times in each of the 4 positions with exactly 4 recipes observed in each of the 5 periods.

Period 1		Period 2		Period 3		Period 4		Period 5	
R_1	R_2	R_5	R_1	R_4	R_5	R_3	R_4	R_2	R_3
R_3	R_4	R_2	R_3	R_1	R_2	R_5	R_1	R_4	R_5

2.30

a. The 223 plots of approximately equal sized land from Google Earth (excluding water)
b. If there is some reason to believe the trees in the 'watery' regions differ from those in the other regions, this discrepancy may cause a divide in our sampling frame and the population of all trees in the region.
c. Again, if trees in the watery region tend to have larger trunk diameter, we would underestimate the number of trees with diameter of 12 inches or more.

2.31

a. All cars (and by extension, their tires) in the state.
b. Cars registered in the 4 months in which the sample was taken.
c. 2 potential concerns arise: not all cars in the region are registered and the time of year may lead to ignoring some cars (some people leave the area for the winter). Unregistered cars may have a higher proportion of unsafe tire tread thickness.

2.32

a. All corn fields in the state.
b. All corn fields in the state (if a list is available).
c. Stratified sampling plan in which the number of acres planted in corn determine the strata.
d. No biases appear present.

2.33

a. People are notoriously bad at recall. A telephone interview immediately following the time of interest would likely be best, but nonresponse is often high. Mailed questionnaires would likely be administered too late to be of use and personal interviewing would be intractable to interview in a timely manner.
b. All three are potential avenues. Interviews are more personal but more time consuming. Mailing questionnaires should also work as the editor has a list of his/her clientele, but if he wants to garner information about perspectives of those not reading his/her paper, he/she may need to blanket the city with questionnaires. Telephone interviews may be difficult as finding the numbers of those in the area may be difficult.
c. Again, all three methods would be viable. A mailed questionnaire would be the easiest and cheapest but the response rate would likely be lower.
d. If the county believes they have an accurate list of those with dogs, a mailed questionnaire or telephone interview would work, but using a list of registered dogs may be underrepresenting those who haven't taken good care of their dogs (and thereby underrepresenting the proportion with rabies shots).

2.34

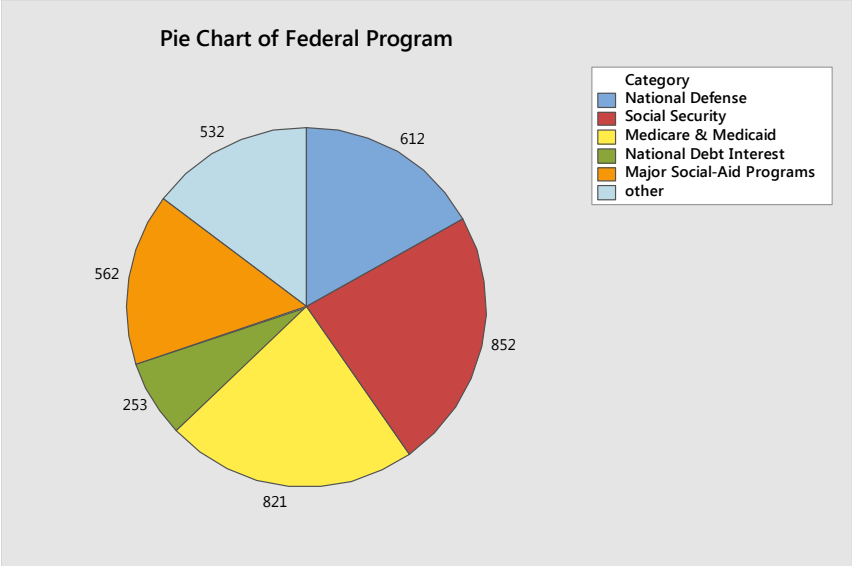
- People who cheat on their taxes are unlikely to admit to it readily. Therefore, the poll likely underestimates the true percentage of people who cheat on their taxes. Garnering truthful responses, even if anonymity is guaranteed, on questions of a personal nature can be a challenge.

Chapter 3

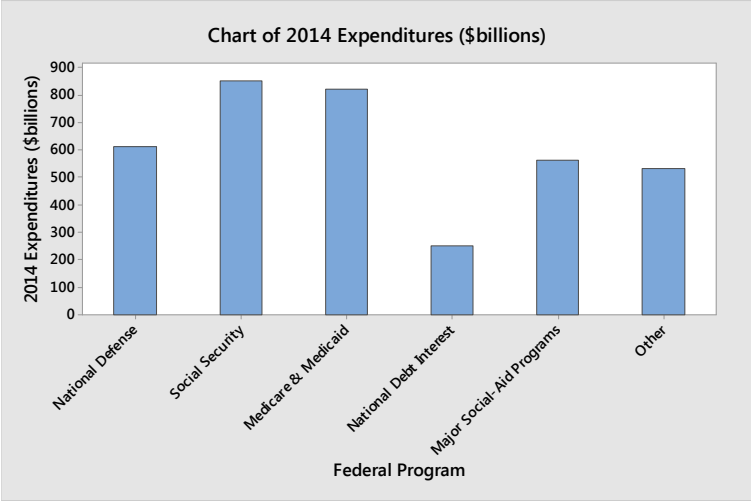
Data Description

3.1

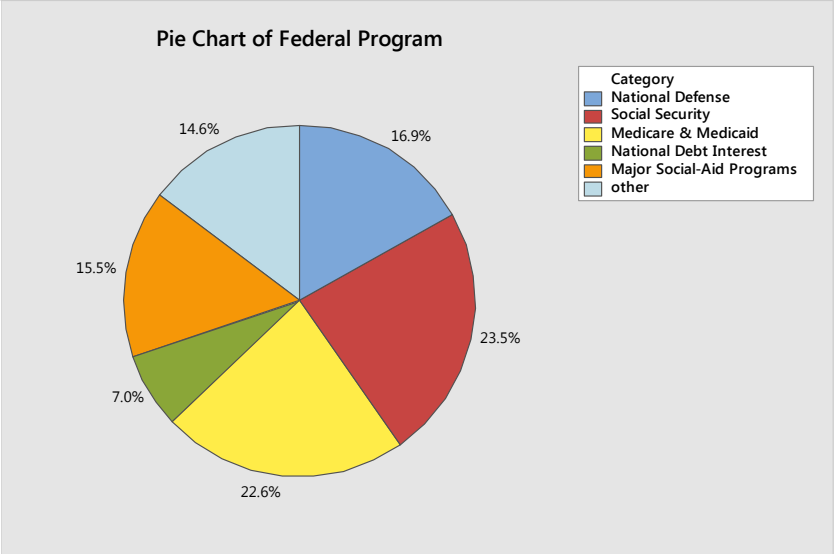
a. The following is a pie chart of the federal expenditures for the 2014 fiscal year (in billions of dollars).

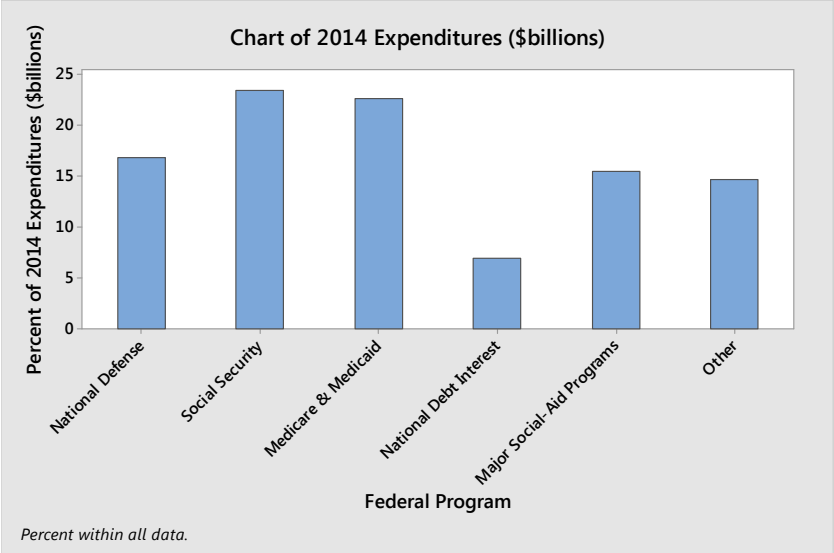


b. The following is a bar chart of the federal expenditures for the 2006 fiscal year (in billions of dollars).



c. The following are a pie chart and bar chart of the federal expenditures for the 2014 fiscal year (in percentages).



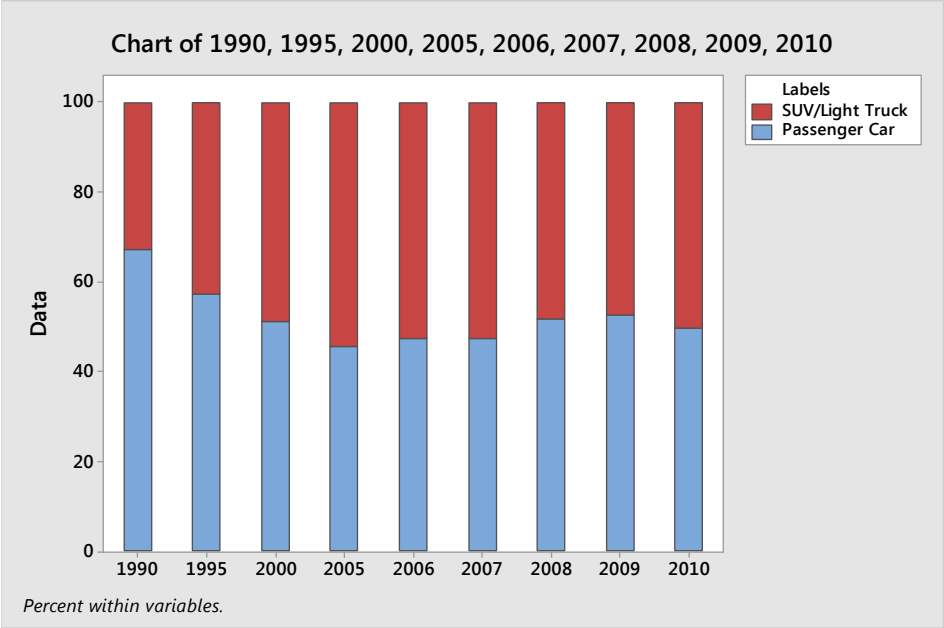


d. The pie chart using percentages is probably most informative to the tax-paying public. Here the tax-paying public can compare the percentages spent by the Federal government for domestic and defense programs as part of a whole.

3.2

a. Pie charts would not be appropriate to display these data. We would not be able to see trends over time.

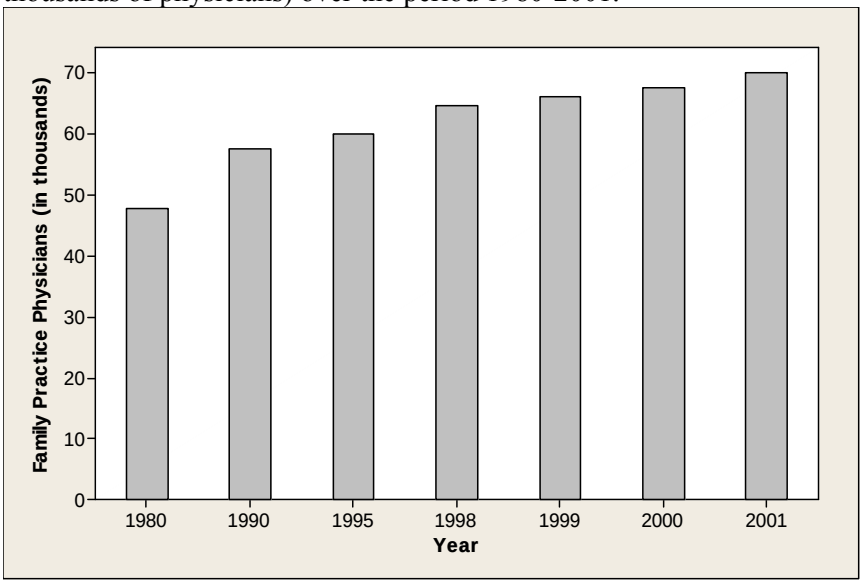
b. The following bar chart shows the changes across the 20 years in the public's choice in vehicle.



c. It appears that the percentage of passenger cars has decreased over the period 1990-2010. If there was a substantial increase in gasoline prices, we would expect the percentage of passenger cars to increase.

3.3

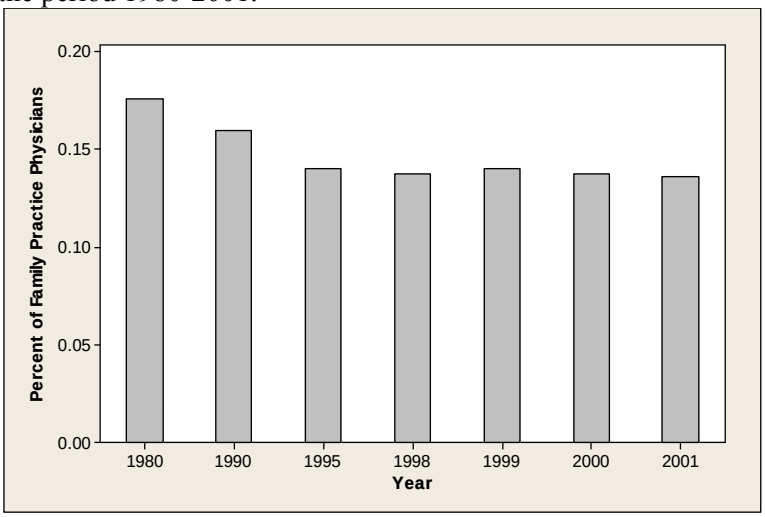
a. The following bar chart shows the increase in the number of family practice physicians (in thousands of physicians) over the period 1980-2001.



b. The percent of office-based physicians who are family practice physicians over the period 1980-2001 can be seen in the following table.

	1980	1990	1995	1998	1999	2000	2001
Percent Family Practice	17.6	16.0	14.0	13.8	14.0	13.8	13.6

The following bar chart shows the percent of office-based physicians who are family practice physicians over the period 1980-2001.

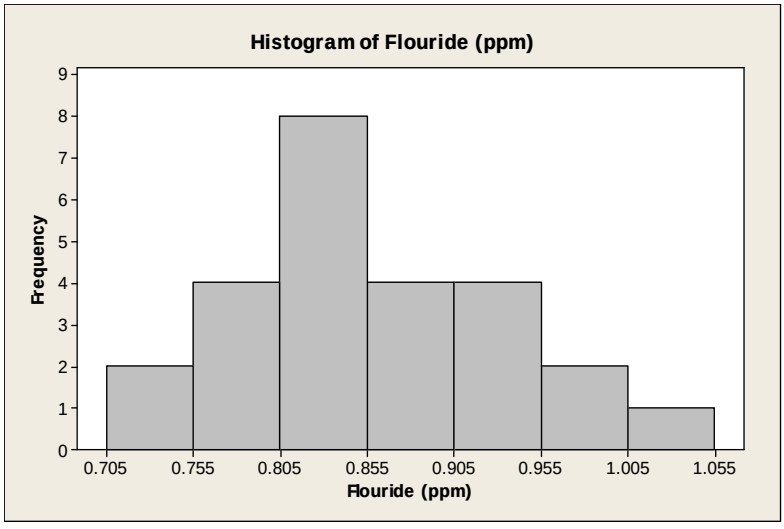


c. While the number of family practice physicians increased over the period 1980-2001, the percent of total office-based physicians who are family practice physicians decreased over the same period.

3.4

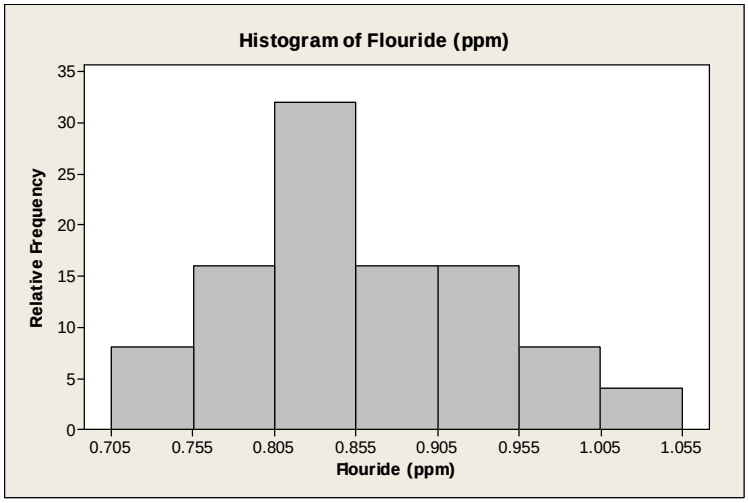
a. Range = 1.05 – 0.72 = 0.33

b. The frequency histogram should be plotted with 7 classes ranging from 0.705 to 1.055. The intervals should have width 0.05.



c. Relative frequencies are given below. Plot relative frequencies versus class intervals.

Class	Class Interval	Frequency (f_i)	Relative Frequency (f_i/25)
1	0.705-0.755	2	0.08
2	0.755-0.805	4	0.16
3	0.805-0.855	8	0.32
4	0.855-0.905	4	0.16
5	0.905-0.955	4	0.16
6	0.955-1.005	2	0.08
7	1.005-1.055	1	0.04
Total		$n = 25$	1.00



d. The probability is 7/25 = 0.28 that the fluoride reading would be greater than 0.90 ppm. Thus, we would predict that 28% of the days would have a reading greater than 0.90 ppm.

3.5 Two separate bar graphs could be plotted, one with Lap Belt Only and the other with Lap and Shoulder Belt. A single bar graph with the Lap Belt Only value plotted next to the Lap and Shoulder Belt for each value of Percentage of Use is probably the most effective plot. This plot would clearly demonstrate that the increase in the number of lives saved by using a shoulder belt increased considerably as the percentage use increased.