CHAPTER PROBLEM

Why was the *Literary Digest* poll so wrong?

Founded in 1890, the *Literary Digest* magazine was famous for its success in conducting polls to predict winners in presidential elections. The magazine correctly predicted the winners in the presidential elections of 1916, 1920, 1924, 1928 and 1932. In the 1936 presidential contest between Alf Landon and Franklin D. Roosevelt, the magazine sent out ten million ballots and received 1,293,669 ballots for Landon and 972,897 ballots for Roosevelt, so it appeared that Landon would capture 57% of the vote. The size of this poll is extremely large when compared to the sizes of other typical polls, so it appeared that the poll would correctly predict the winner once again. James A. Farley, Chairman of the Democratic National Committee at the time, praised the poll by saying this: "Any sane person cannot escape the implication of such a gigantic sampling of popular opinion as is embraced in *The Literary Digest* straw vote. I consider this conclusive evidence as to the desire of the people of this country for a change in the National Government. *The Literary Digest* poll is an achievement of no little magnitude. It is a poll fairly and correctly conducted." Well, Landon received 16,679,583 votes to the 27,751,597 votes cast for Roosevelt. Instead of getting 57% of the vote as suggested by *The Literary Digest* poll, Landon received only 37% of the vote. *The Literary Digest* magazine suffered a humiliating defeat and soon went out of business. In that same 1936 presidential election, George Gallup used a much smaller poll of 50,000 subjects, and he correctly predicted that Roosevelt would win. How could it happen that the larger *Literary Digest* poll could be wrong by such a large margin? What went wrong? As you learn about the basics of statistics in this chapter, we will return to the *Literary Digest* poll and explain why it was so wrong in predicting the winner of the 1936 presidential contest.

MATH 103 CHAPTER 1 HOMEWORK			
1.1	NA	**1.4**	1,3, 4, 5, 6, 8, 9, 10, 12, 13, 15-19, 21, 24, 25, 28, 30
1.2	1-18, 23, 26, 28	**1.5**	1-4,6, 9, 11, 12, 13, 15, 16, 18, 19, 21-26, 27, 29, 31
1.3	1-32, 34		

1.1 REVIEW AND PREVIEW

The *Literary Digest* poll and George Gallup's poll both used _____ data. Polls collect _____ from a _____ part of a larger group so that we can learn something about the _____ group.

DEFINITION

Data are _____ of _____ (such as measurements, genders, survey responses).

Statistics is the _____ of planning _____ and _____, obtaining _____, and then _____, _____, _____, _____, and drawing _____ based on the _____.

A **population** is the complete collection of all _____ (scores, people, measurements, and so on) to be studied.

A **census** is the collection of _____ from _____ member of the population.

A **sample** is a _____ of members selected from a _____.

The *Literary Digest* poll was a _____ of 2.3 million respondents. What would the population consist of?

Remember—garbage in, garbage out! Sample data must be collected through a process of _____ selection. If sample data are not collected in an appropriate way, the data may be completely _____!

1.2 STATISTICAL THINKING

Key Concept…
When conducting a statistical analysis of data we have collected or analyzing a statistical analysis done by someone else, we should not rely on blind acceptance of mathematical calculations. We should consider these factors:

- Context of the data
- Source of the data
- Sampling method
- Conclusions
- Practical implications

650	24249	0
1050	20666	0
967	19413	0
500	21992	0
1700	21399	0
2000	22022	0
1100	25859	0
1300	20390	0
1400	23738	0
2250	23294	0
800	19063	0
3500	30131	0
1200	18698	0
1250	25348	0

2250	25642	1
3000	23074	1
1750	28349	1
1525	24644	1
1500	23245	1
1500	24378	1
1250	23246	1
1200	23695	1
1600	23258	1
425	19325	1
1450	20397	1
900	17256	1
675	19545	1
1450	20780	1

Description: These data for the 1991 season of the National Football League were reported by the Associated Press.

Number of cases: 28

Variable Names:

1. TEAM: Name of team
2. QB: Salary ($thousands) of regular quarterback
3. TOTAL: Total team salaries ($thousands)
4. NFC: National Football Conference (1) or American Football Conference (0)

The Data:

TEAM	QB	TOTAL	NFC
BILLS	650	24249	0
BENGALS	1050	20666	0
BROWNS	967	19413	0
BRONCOS	500	21992	0
OILERS	1700	21399	0
COLTS	2000	22022	0
CHIEFS	1100	25859	0
RAIDERS	1300	20390	0
DOLPHINS	1400	23738	0
PATRIOTS	2250	23294	0
JETS	800	19063	0
STEELERS	3500	30131	0
CHARGERS	1200	18698	0
SEAHAWKS	1250	25348	0
FALCONS	2250	25642	1
BEARS	3000	23074	1
COWBOYS	1750	28349	1
LIONS	1525	24644	1
PACKERS	1500	23245	1
RAMS	1500	24378	1
VIKINGS	1250	23246	1
SAINTS	1200	23695	1
GIANTS	1600	23258	1
EAGLES	425	19325	1
CARDINALS	1450	20397	1
49ERS	900	17256	1
BUCCANEERS	675	19545	1
REDSKINS	1450	20780	1

Example 1: Refer to the data in the table below. The *x*-values are weights (in pounds) of cars; the *y*-values are the corresponding highway fuel consumption amounts (in mi/gal).

Car Weights and Highway Fuel Consumption Amounts

WEIGHT	4035	3315	4115	3650	3565
FUEL CONSUMPTION	26	31	29	29	30

a. Context of the data.
   i. Are the *x*-values matched with the corresponding *y*-values? That is, is each *x*-value somehow associated with the corresponding *y*-value in some meaningful way?

   ii. If the *x* and *y* values are matched, does it make sense to use the difference between each *x*-value and the *y*-value that is in the same column? Why or why not?

b. Conclusion. Given the context of the car measurement data, what issue can be addressed by conducting a statistical analysis of the values?

c. Source of the data. Comment on the source of the data if you are told the car manufacturers supplied the values. Is there an incentive for car manufacturers to report values that are not accurate?

d. Conclusion. If we use statistical methods to conclude that there is a correlation between the weights of cars and the amounts of fuel consumption, can we conclude that adding weight to a car causes it to consume more fuel?

Example 2: Form a conclusion about statistical significance. Do not make any formal calculations. Either use results provided or make subjective judgements about the results.

One of Gregor Mendel's famous hybridization experiments with peas yielded 580 offspring with 152 of those peas (or 26%) having yellow pods. According to Mendel's theory, 25% of the offspring should have yellow pods. Do the results of the experiment differ from Mendel's claimed rate of 25% by an amount that is statistically significant?

1.3 TYPES OF DATA

Key Concept…
A goal of statistics is to make _____, or generalizations, about a population. In addition to the terms population and sample, we need to know the meanings of the terms _____ and _____. These new terms are used to distinguish between cases in which we have data for an entire _____, and cases in which we have data for a _____ only. We also need to know the difference between _____ data and _____ data, which distinguish between different types of _____.

DEFINITION

A **parameter** is a _____ measurement describing some characteristic of a _____.

A **statistic** is a _____ measurement describing some characteristic of a _____.

Example 3: Determine whether the given value is a statistic or a parameter.

a. 45% of the students in a calculus class failed the first exam.

b. 25 calculus students were randomly selected from all the sections of calculus I. 38% of these student failed the first exam.

DEFINITION

Quantitative (aka numerical) data consist of _____

representing _____ or _____.

Categorical (aka qualitative or attribute) data consist of _____

or _____ that are not numbers representing counts or measurements.

Give 2 examples of

a. Quantitative data

b. Categorical data

DEFINITION

Discrete data result when the number of possible values is either a

_____ number or a _____ number.

Continuous (aka numerical) data result from _____ many

possible values that correspond to some _____ scale

that covers a _____ of values without gaps, interruptions or jumps.

Give 2 examples of

a. Discrete data

b. Continuous data

DEFINITION

The **nominal level of measurement** is characterized by data that consists of

_____, _____, or _____

only. The data cannot be arranged in an _____ scheme (such as low to high).

Give 2 examples of the nominal level of measurement.

DEFINITION

Data are at the **ordinal level of measurement** if they can be

_____ in some _____, but differences (obtained by subtraction) between data values either cannot be determined or are meaningless.

Give 2 examples of the ordinal level of measurement.

DEFINITION

The **interval level of measurement** is like the _____ level,

with the additional property that the _____ between any two data values is meaningful. However, data at this level do not have a natural zero starting point (where none of the quantity is present).

Give 2 examples of the interval level of measurement.

DEFINITION

The **ratio level of measurement** is like the _____ level,

with the additional property that there is a natural _____ starting place (where zero indicates that none of the quantity is present). For

values at this level, _____ and _____ are both meaningful.

Give 2 examples of the ratio level of measurement.

LEVELS OF MEASUREMENT

RATIO		
INTERVAL		
ORDINAL		
NOMINAL		

1.4 CRITICAL THINKING

Key Concept…

This section is the first of many in which we focus on the _____ of _____ obtained by studying data. The aim of this section is to improve our skills in _____ information based on _____. This section shows how to use _____ sense to think _____ about data and statistics.

*"**Lies, damned lies, and statistics**" is a phrase describing the persuasive power of numbers, particularly the use of statistics to bolster weak arguments, and the tendency of people to disparage statistics that do not support their positions. It is also sometimes colloquially used to doubt statistics used to prove an opponent's point.*

DEFINITION

A **voluntary response sample (aka self-select sample)** is one in which _____ themselves _____ whether to be included.

Give three examples of voluntary response samples.

CORRELATION AND CAUSALITY

Another way to _____ statistical data is to find a statistical association between two variables and to conclude that one of the variables _____ (or directly affects) the other variable.

_____ DOES NOT IMPLY CAUSALITY!

REPORTED RESULTS

When collecting data from people, it is better to take the measurements yourself instead of asking subjects to report results.

Give two situations in which people might falsely report results.

SMALL SAMPLES

Conclusions should not be based on samples that are far too small.

PERCENTAGES

Some studies will cite _____ or _____ percentages. Keep in mind that 100% of a quantity is ALL of the quantity. If there are references made to percentages which exceed 100%, such references are often not justified.

PERCENTAGE REVIEW

"of" means multiply

Percent means per hundred so $n\% = \frac{n}{100}$

Percentage of: Change the % to $\frac{1}{100}$ then multiply.

Fraction to percentage: Change the fraction to a decimal by dividing the

_____ by the _____,

then multiply by 100 and put in the percent symbol.

Decimal to percentage: Multiply the decimal by 100 and put in the percent symbol.

Percentage to decimal: Remove the percent symbol and divide by 100.

Example 4: Perform the indicated operation.

a. 12% of 1200

b. Write 5/8 as a percentage.

c. Write 8.5% as a decimal

d. Write 15% as a simplified fraction

LOADED QUESTIONS

If survey questions are not worded carefully, the results of a study can be misleading. Survey questions can be _____ or intentionally _____ to elicit a desired response.

ORDER OF QUESTIONS

Sometimes survey questions are unintentionally loaded by such factors as the order of items being considered.

NONRESPONSE

A _____ occurs when someone either refuses to respond or is unavailable. Why do you think that more and more people are refusing to participate in polls?

MISSING DATA

Results can sometimes be dramatically affected by missing data. This can be due to a random occurrence such as a subject dropping out of a study for reasons unrelated to the study. Some data are missing due to special factors such as the tendency of people with low incomes to be less likely to report their income.

SELF-INTEREST STUDY

Some parties with interests to promote will sponsor studies. We should be wary of surveys in which the sponsor can enjoy monetary gains from the results.

PRECISE NUMBERS

Numbers which are estimates should be rounded. 2,234,786 should be rounded to 2 million.

DELIBERATE DISTORTIONS

1.5 COLLECTING SAMPLE DATA

Key Concept…

The method used to collect sample data influences the quality of our statistical analysis. Of particular importance is the _____ _____ _____. **If sample data are not collected in the appropriate way, the data may be so completely useless that no amount of statistical torturing can salvage them.**

DEFINITION

In an **observational study**, we _____ and measure specific characteristics, but we don't attempt to _____ the subjects being studied.

In an **experiment**, we apply some _____ and then proceed to _____ its _____ on the subjects. Subjects in experiments are called experimental units.

Give one example of an

a. Observational study

b. Experiment

DEFINITION

A **simple random sample** of *n* subjects is selected in such a way that every possible sample of the same size *n* has the same chance of being chosen.

DEFINITION

In a **random sample**, members from the _____ are selected

in such a way that each _____ member in the population has

an _____ chance of being selected.

A **probability sample** involves selecting members from a _____

in such a way that each member of the population has a _____
(but not necessarily the same) chance of being selected.

DEFINITION

In **systematic sampling**, we select some _____ point and then select every *k*th (such as every 20th) element in the population.

With **convenience sampling**, we simply use results that are very _____ to get.

With **stratified sampling**, we _____ the population into at least two different subgroups (aka strata) so that subjects within the same

subgroup share the same characteristics, such as _____ or _____

bracket, then we draw a sample from each _____.

In **cluster sampling**, we first _____ the population area into

sections or _____, then _____

select some of those clusters, and then choose _____ the members from those selected clusters.

Example 5: Identify which type of sampling is used: random, systematic, convenience, stratified, or cluster.

a. Every 8th driver is stopped and interviewed at a sobriety checkpoint.

b. In a neighborhood, specific streets are randomly selected and all residents on the selected streets are polled.

c. At Mira Costa College, 500 male students and 500 female students are randomly selected to participate in a study.

d. Ms. Gracey surveyed the students in her class.

e. Telephone numbers are randomly generated. Those people are selected to be interviewed.

DEFINITION

In a **cross-sectional study**, data are _____,

_____, and _____ at one point in time.

In a **retrospective (aka case-control) study**, data are collected from the

_____ by going back through time (through examination of records, interviews, etc).

In a **prospective (aka longitudinal or cohort) study**, data are collected in the

_____ from groups sharing common factors (called cohorts).

Give one example of a

a. Cross-sectional study

b. Retrospective study

c. Prospective study

DESIGN OF EXPERIMENTS

RANDOMIZATION

Subjects are assigned to different groups through a process of random selection.

REPLICATION

Replication is the repetition of an experiment on more than one subject. Use a sample size that is large enough to let us see the true nature of any effects, and obtain the sample using an appropriate method, such as one based on randomness.

BLINDING

Blinding is a technique in which the subject doesn't know whether he or she is receiving the treatment or the placebo. In a double-blind experiment, both the subject and the investigator do not know whether the subject received the treatment or the placebo.

DEFINITION

Confounding occurs in an experiment when you are not able to distinguish among the _____ of different _____.

COMPLETELY RANDOMIZED EXPERIMENTAL DESIGN

Assign subjects to different treatment groups through a process of _____ selection.

RANDOMIZED BLOCK DESIGN

A **block** is a group of subjects that are _____, but blocks differ in ways that might affect the _____ of an experiment. If testing one or more treatments within different blocks, use this experimental design.

1. Form blocks (or groups) of subjects with similar characteristics.
2. Randomly assign treatments to the subjects within each block.

RIGOROUSLY CONTROLLED DESIGN

Carefully assign subjects to different treatment groups, so that those given each treatment are _____ in ways that are important to the _____.

MATCHED PAIRS DESIGN

Compare exactly two treatment groups (such as treatment and placebo) by using subjects matched in pairs that are somehow related or have similar characteristics.

SUMMARY

1. Use _____ to assign subjects to different groups.
2. Use _____ by repeating the experiment on enough subjects so that effects of treatments or other factors can be clearly seen.

3. _____ the effects of _____ by using such techniques as blinding and a completely randomized experimental design.

DEFINITION

A **sampling error** is the difference between a _____ result and the true _____ result; such an error results from chance sample fluctuation.

A **nonsampling error** occurs when the sample data are incorrectly _____, recorded, or _____ (such as by selecting a biased sample, using a defective measurement instrument, or copying the data incorrectly).

Example 6: Identify the type of observational study (cross-sectional, retrospective, or prospective)

a. Physicians at the Mount Sinai Medical Center plan to study emergency personnel who worked at the site of the terrorist attacks in New York City on September 11, 2001. They plan to study these workers from now until several years into the future.

b. University of Toronto researchers studied 669 traffic crashes involving drivers with cell phones. They found that cell phone use quadruples the risk of a collision.