Exam 1 Review

Chapter 1 (Sections 1.1-1.3)

- *Types of data/variables*
- *Graphical displays*
- *Numerical measures*
- *Features of a distribution*
- *Normal distributions (not covered on this exam)*

Chapter 3 (Sections 3.1.-3.3)

- *Sources of Data*
- *Experiments*
- *Observational studies*
- *Bias and confounding*
- *Sampling designs*
- *Ethics (not covered on this exam)*

Chapter 4

- *Not covered on this exam*

Section 1.1 SUMMARY

- A data set contains information on a number of **cases.** Cases may be customers, companies, subjects in a study, units in an experiment, or other objects.
- For each case, the data give values for one or more **variables**. A variable describes some characteristic of a case, such as a person's height, gender, or salary. Variables can have different **values** for different cases.
- A **label** is a special variable used to identify cases in a data set.
- Some variables are **categorical**, and others are **quantitative**. A categorical variable places each individual into a category, such as male or female. A quantitative variable has numerical values that measure some characteristic of each case, such as height in centimeters or annual salary in dollars.
- The **key characteristics** of a data set answer the questions Who?, What?, and Why?
- Converting a count to a **rate** is an example of **adjusting one variable to create another**.

Section 1.2 SUMMARY

- **Exploratory data analysis** uses graphs and numerical summaries to describe the variables in a data set and the relations among them.
- The **distribution** of a variable tells us what values it takes and how often it takes these values.
- To describe a distribution, begin with a graph. **Bar graphs** and **pie charts** display the distribution of a categorical variable. **Stemplots** and **histograms** display the distributions of a quantitative variable.
- When examining any graph, look for an overall pattern and for clear **deviations** from that pattern.
- **Shape**, **center**, and **spread** describe the overall pattern of a distribution. Some distributions have simple shapes, such as **symmetric** or **skewed**. Not all distributions have a simple overall shape, especially when there are few observations.
- **Outliers** are observations that lie outside the overall pattern of a distribution. Always look for outliers and try to explain them.
- When observations on a variable are taken over time, make a **time plot** that graphs time horizontally and the values of the variable vertically. A time plot can reveal interesting patterns in a set of data.

Section 1.3 SUMMARY

- A numerical summary of a distribution should report its **center** and its **spread** or **variability.**
- The **mean** $\bar{x}$ and the **median** M describe the center of a distribution in different ways. The mean is the arithmetic average of the observations, and the median is their midpoint.
- When you use the median to describe the center of a distribution, describe its spread by giving the **quartiles.** The **first quartile** Q_1 has one-fourth of the observations below it, and the **third quartile** Q_3 has three-fourths of the observations below it.
- The **interquartile range** is the difference between the quartiles. It is the spread of the center half of the data. The **1.5 × *IQR* rule** flags observations more than $1.5 \times IQR$ beyond the quartiles as possible outliers.
- The **five-number summary**—consisting of the median, the quartiles, and the smallest and largest individual observations—provides a quick overall description of a distribution. The median describes the center, and the quartiles and extremes show the spread.
- **Boxplots** based on the five-number summary are useful for comparing several distributions. The box spans the quartiles and shows the spread of the central half of the distribution. The median is marked within the box. Lines extend
- The **variance** s^2 and especially its square root, the **standard deviation** s, are common measures of spread about the mean as center. The standard deviation s is zero when there is no spread and gets larger as the spread increases.
- A **resistant measure** of any aspect of a distribution is relatively unaffected by changes in the numerical value of a small proportion of the total number of observations, no matter how large these changes are. The median and quartiles are resistant, but the mean and the standard deviation are not.
- The mean and standard deviation are good descriptions for symmetric distributions without outliers. They are most useful for the Normal distributions introduced in the next section. The five-number summary is a better exploratory description for skewed distributions.
- **Linear transformations** have the form $x_{\text{new}} = a + bx$. A linear transformation changes the origin if $a \neq 0$ and changes the size of the unit of measurement if $b > 0$. Linear transformations do not change the overall shape of a distribution. A linear transformation multiplies a measure of spread by b and changes a percentile or measure of center m into $a + bm$.
- Numerical measures of particular aspects of a distribution, such as center and spread, do not report the entire shape of most distributions. In some cases, particularly distributions with multiple peaks and gaps, these measures may not be very informative.

Section 3.1 SUMMARY

- **Anecdotal data** come from stories or reports about cases that do not necessarily represent a larger group of cases.
- **Available data** are data that were produced for some other purpose but that may help answer a question of interest.
- A **sample survey** collects data from a **sample** of cases that represent some larger **population** of cases.
- A **census** collects data from all cases in the population of interest.
- In an **observational study**, we observe individuals but we do not attempt to influence their responses.
- In an **experiment**, a treatment or an intervention is imposed, and the responses are recorded.

Section 3.2 SUMMARY

- In an experiment, one or more **treatments** are imposed on the **experimental units** or **subjects**. Each treatment is a combination of **levels** of the explanatory variables, which we call **factors**. **Outcomes** are the measured variables that are used to compare the treatments.
- The **design** of an experiment refers to the choice of treatments and the manner in which the experimental units or subjects are assigned to the treatments.
- The basic principles of statistical design of experiments are **comparison**, **randomization**, and **repetition**.
- The simplest form of control is **comparison**. Experiments should compare two or more treatments in order to prevent **confounding** the effect of a treatment with other influences, such as lurking variables.
- **Randomization** uses chance to assign subjects to the treatments. Randomization creates treatment groups that are similar (except for chance variation) before the treatments are applied. Randomization and comparison together prevent **bias**, or systematic favoritism, in experiments.
- You can carry out randomization by giving numerical labels to the experimental units and using software or a **table of random digits** to choose treatment groups.
- **Repetition** of the treatments on many units reduces the role of chance variation and makes the experiment more sensitive to differences among the treatments.
- Good experiments require attention to detail as well as good statistical design. Many behavioral and medical experiments are **double-blind**. **Lack of realism** in an experiment can prevent us from generalizing its results.
- **Matched pairs** are used to compare two treatments. In some matched pairs designs, each subject receives both treatments in a random order. This is called a cross-over experiment. In others, the subjects are matched in pairs as closely as possible, and one subject in each pair receives each treatment.
- In addition to comparison, a second form of control is to restrict randomization by forming **blocks** of experimental units that are similar in some way that is important to the response. Randomization is then carried out separately within each block.

Section 3.3 SUMMARY

- A sample survey selects a **sample** from the **population** of all individuals about which we desire information. We base conclusions about the population on data about the sample.
- The **sample design** refers to the method used to select the sample from the population. **Probability sample designs** use impersonal chance to select a sample.
- The basic probability sample is a **simple random sample (SRS)**. An SRS gives every possible sample of a given size the same chance to be chosen.
- Choose an SRS using software. This can also be done using a **table of random digits** to select the sample.
- To choose a **stratified random sample**, divide the population into **strata**, groups of individuals that are similar in some way that is important to the response. Then choose a separate SRS from each stratum and combine them to form the full sample.
- **Multistage samples** select successively smaller groups within the population in stages, resulting in a sample consisting of clusters of individuals. Each stage may employ an SRS, a stratified sample, or another type of sample.
- Failure to use probability sampling often results in **bias**, or systematic errors in the way the sample represents the population.
- **Voluntary response samples**, in which the respondents choose themselves, are particularly prone to large bias.
- In human populations, even probability samples can suffer from bias due to **undercoverage** or **nonresponse**, from **response bias** due to the behavior of the interviewer or the respondent, or from misleading results due to **poorly worded questions**.