Review for Final Exam

The Three-Step Process in Statistical Analysis.

Step 1: Collecting the Data **Chapter 3**

- *Sampling design*
- *Experimental design*
- *Observational study*

Step 2: Summarizing/Organizing the Data **Chapter 1**

- *Descriptive statistics*
- *Graphical displays*
- *Numerical measures*

Step 3: Drawing Conclusions from the Data **Chapters 4-8**

- *Inferential Statistics*
- *Estimation of unknown parameters*
- *Hypothesis testing*

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** $\overline{x}$ 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 **$1.5 \times IQR$ 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.

Some Basic Terminology

Subject/Case: The object (person or thing) upon which we are collecting data (information)

Population: The collection of all subjects/cases of interest to our study

Sample: The collection of subjects/cases actually used in our study

Variable: A characteristic that varies from subject to subject (case to case)

Label: A special variable used in some data sets to distinguish the different cases

Data: The collection of observed values (observations) for one or more variables recorded for all subjects in the sample.

Distribution: The pattern of variability displayed by the data of a variable. The distribution displays the possible values and the frequency of each value.

Two Different Types of Variables

Quantitative Variable: A numerical characteristic that represents a quantity. For this variable is meaningful to:

- Average its values
- Arrange values in order

ex: Age, Length of Employment

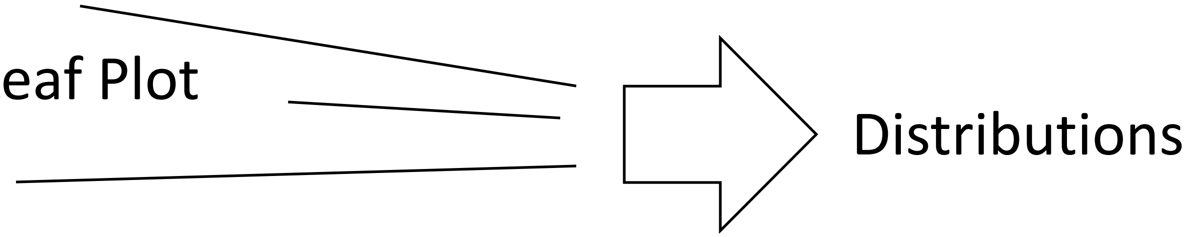
Categorical Variable: A non-numerical or numerical characteristic represented by two or more categories (not representing a quantity)

ex: Gender, Ethnicity, Social Security Number

Characterizing Data Using Graphical Methods

Quantitative Variables

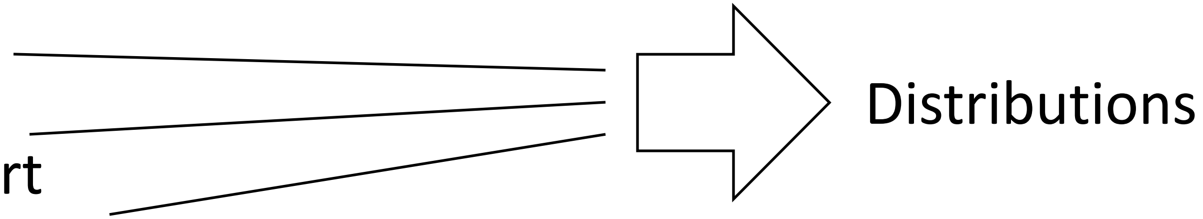
- Histogram
- Stem-and-Leaf Plot
- Time Plot



Distributions

Categorical Variables

- Pie Chart
- Bar Chart
- Pareto Chart



Distributions

Main Features of a Distribution

Symmetrical: Both sides of the distribution are approximately identical mirror images. There is a *line* of symmetry.

Skewed: One tail is stretched out longer than the other. The direction of skewness is on the side of the longer tail. (Right skewed vs. left skewed)

Bimodal/Unimodal: The distribution is bimodal (unimodal) if the distribution has two (one) major peak(s). The two largest classes are separated by one or more classes. Often implies two populations are sampled.

Normal: A symmetrical distribution is mounded about the mean and becomes sparse at the extremes

Outliers are isolated observations that deviate significantly from the “main” population.

Characterizing (Quantitative) Data Using Numerical Methods

Measures of Central Tendency

- Sample Mean
- Sample Median
- Mode

Measures of Variation

- Range
- Sample Variance or Standard Deviation
- Inter Quartile Range (IQR)

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**.

SECTION 4.1 SUMMARY

- A **random phenomenon** has outcomes that we cannot predict but that nonetheless have a regular distribution in very many repetitions.
- The **probability** of an event is the proportion of times the event occurs in many repeated trials of a random phenomenon.
- Trials are **independent** if the outcome of one trial does not influence the outcome of any other trial.

SECTION 4.2 SUMMARY

- A **probability model** for a random phenomenon consists of a sample space *S* and an assignment of probabilities *P*.
- The **sample space *S*** is the set of all possible outcomes of the random phenomenon. Sets of outcomes are called **events**. *P* assigns a number *P*(*A*) to an event *A* as its probability.
- The **complement** A^c of an event *A* consists of exactly the outcomes that are not in *A*. Events *A* and *B* are **disjoint** if they have no outcomes in common. Events *A* and *B* are **independent** if knowing that one event occurs does not change the probability we would assign to the other event.
- Any assignment of probability must obey the rules that state the basic properties of probability:

SECTION 4.3 SUMMARY

- A **random variable** is a variable taking numerical values determined by the outcome of a random phenomenon. The **probability distribution** of a random variable *X* tells us what the possible values of *X* are and how probabilities are assigned to those values.
- A random variable *X* and its distribution can be **discrete** or **continuous**.
- A **discrete random variable** has possible values that can be given in an ordered list. The probability distribution assigns each of these values a probability between 0 and 1 such that the sum of all the probabilities is exactly 1. The probability of any event is the sum of the probabilities of all the values that make up the event.
- A **continuous random variable** takes all values in some interval of numbers. A **density curve** describes the probability distribution of a continuous random variable. The probability of any event is the area under the curve and above the values that make up the event.
- **Uniform distributions** are continuous probability distributions that are very similar to equally likely discrete distributions.
- **Normal distributions** are one type of continuous probability distribution.
- You can picture a probability distribution by drawing a **probability histogram** in the discrete case or by graphing the density curve in the continuous case.

SECTION 4.4 SUMMARY

- The probability distribution of a random variable X, like a distribution of data, has a **mean** μ_X and a **standard deviation** σ_X.

- The **law of large numbers** says that the average of the values of X observed in many trials must approach μ.

- The **mean** μ is the balance point of the probability histogram or density curve. If X is **discrete** with possible values x_i having probabilities p_i, the mean is the average of the values of X, each weighted by its probability:

$$\mu_X = x_1 p_1 + x_2 p_2 + \cdots$$

- The **variance** σ_X^2 is the average squared deviation of the values of the variable from their mean. For a discrete random variable,

$$\sigma_X^2 = (x_1 - \mu_X)^2 p_1 + (x_2 - \mu_X)^2 p_2 + \ldots$$

- The **standard deviation** σ_X is the square root of the variance. The standard deviation measures the variability of the distribution about the mean. It is easiest to interpret for Normal distributions.

- The **mean and variance of a continuous random variable** can be computed from the density curve, but to do so requires more advanced mathematics.

- The means and variances of random variables obey the following rules. If a and b are fixed numbers, then

Section 4.5 SUMMARY

- The **complement** A^c of an event A contains all outcomes that are not in A. The **union** $\{A \text{ or } B\}$ of events A and B contains all outcomes in A, in B, and in both A and B. The **intersection** $\{A \text{ and } B\}$ contains all outcomes that are in both A and B, but not outcomes in A alone or B alone.

- The **conditional probability** $P(B|A)$ of an event B, given an event A, is defined by

  $$P(B|A) = \frac{P\left(A \text{ and } B\right)}{P\left(A\right)}$$

  when $P(A) > 0$. In practice, conditional probabilities are most often found from directly available information.

- The essential general rules of elementary probability are

  Legitimate values:$0 \leq P(A) \leq 1$ for any event A

  Total probability 1: $P(S) = 1$

  Complement rule: $P(A^c) = 1 - P(A)$

  Addition rule: $P(A \text{ or } B) = P(A) + P(B) - P(A \text{ and } B)$

  Multiplication rule: $P(A \text{ and } B) = P(A)P(B|A)$

- If A and B are **disjoint**, then $P(A \text{ and } B) = 0$. The general addition rule for unions then becomes the special addition rule, $P(A \text{ or } B) = P(A) + P(B)$.

- A and B are **independent** when $P(B|A) = P(B)$. The multiplication rule for intersections then becomes $P(A \text{ and } B) = P(A)P(B)$.

- In problems with several stages, draw a **tree diagram** to organize use of the multiplication and addition rules.

SECTION 5.1 SUMMARY

- A number that describes a population is a **parameter**. A number that describes a sample (is computed from the sample data) is a **statistic**. The purpose of sampling or experimentation is usually **inference**: use sample statistics to make statements about unknown population parameters.

- A statistic from a probability sample or a randomized experiment has a **sampling distribution** that describes how the statistic varies in repeated data productions. The sampling distribution answers the question "What would happen if we repeated the sample or experiment many times?" Formal statistical inference is based on the sampling distributions of statistics.

- A statistic as an estimator of a parameter may suffer from **bias** or from high **variability**. Bias means that the center of the sampling distribution is not equal to the true value of the parameter. The variability of the statistic is described by the spread of its sampling distribution. Variability is usually reported by giving a **margin of error** for conclusions based on sample results.

- Properly chosen statistics from randomized data production designs have no bias resulting from the way the sample is selected or the way the experimental units are assigned to treatments. We can reduce the variability of the statistic by increasing the size of the sample or the size of the experimental groups.

SECTION 5.2 SUMMARY

- The **population distribution** of a variable is the distribution of its values for all members of the population.

- The **sample mean** $\bar{x}$ of an SRS of size n drawn from a large population with mean μ and standard deviation σ has a sampling distribution with mean and standard deviation

$$\begin{aligned} \mu_{\bar{x}} &= \mu \\ \sigma_{\bar{x}} &= \frac{\sigma}{\sqrt{n}} \end{aligned}$$

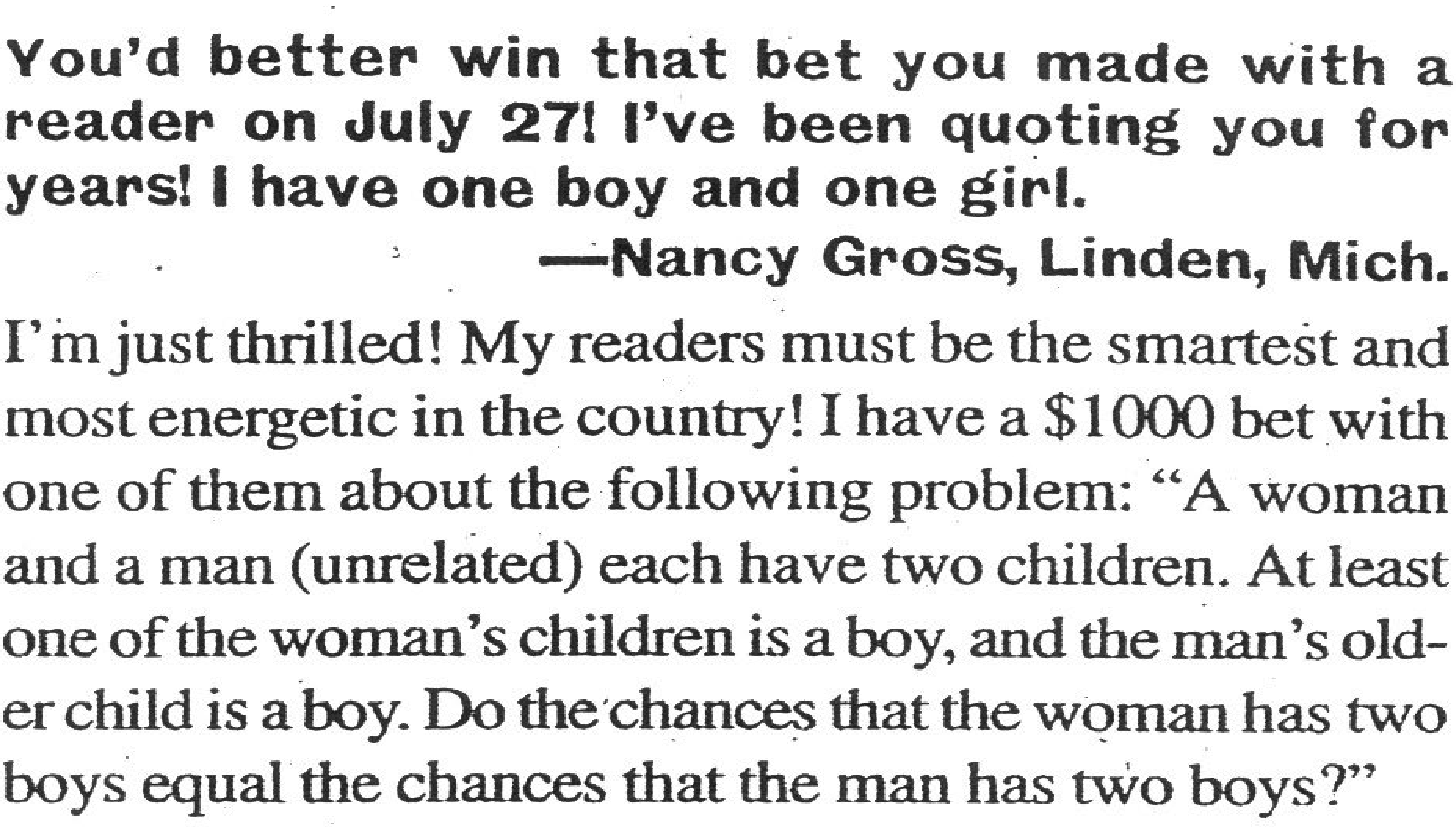
The sample mean $\bar{x}$ is an unbiased estimator of the population mean μ and is less variable than a single observation. The standard deviation decreases in proportion to the square root of the sample size n. This means that to reduce the standard deviation by a factor of C, we need to increase the sample size by a factor of C^2.

- The **central limit theorem** states that, for large n, the sampling distribution of $\bar{x}$ is approximately $N(\mu, \sigma/\sqrt{n})$ for any population with mean μ and finite standard deviation σ. This allows us to approximate probability calculations of $\bar{x}$ using the Normal distribution.

- Linear combinations of independent Normal random variables have Normal distributions. In particular, if the population has a Normal distribution, so does $\bar{x}$.

Additional Examples
Chapters 4-5

$1,000 Challenge



You'd better win that bet you made with a reader on July 27! I've been quoting you for years! I have one boy and one girl.

—Nancy Gross, Linden, Mich.

I'm just thrilled! My readers must be the smartest and most energetic in the country! I have a $1000 bet with one of them about the following problem: "A woman and a man (unrelated) each have two children. At least one of the woman's children is a boy, and the man's older child is a boy. Do the chances that the woman has two boys equal the chances that the man has two boys?"

SOURCE: "Ask Marilyn – The Parade"

CASE: Retirement Annuity

- Upon retiring at age 65, Professor Smith plans to draw $5,000 each month from his retirement savings until his $600,000 funds are depleted.
- His retirement plan, TIAA-CREF is offering him an alternative option that would pay him $5,000 each month for as long as he lives.
- What is the maximum age that TIAA-CREF believes that Mr. Smith will live?

Excel Spreadsheet