Statistics Notebook

by Tommy Yang

Updated: September 11, 2016

Contents

Chapter 1: Exploring Data

1. Exploring Data

1.1 Analyzing Categorical Data

Definition 1.1 – Individuals and variables.

- **Individuals** are the objects described by a set of data. Individuals may be people, animals, or things.
- A **variable** is any characteristic of an individual. A variable can take different values for different individuals.

Definition 1.2 – Categorical variable and quantitative variable.

- A **categorical variable** places an individual into one of several groups or categories.
- A **quantitative variable** takes numerical values for which it makes sense to find an average.

Definition 1.3 – Distribution.

- The **distribution** of a variable tells us what values the variable takes and how often it takes these values.

Definition 1.4 – Distribution of a categorical variable.

- The distribution of a categorical variable lists the categories and gives the count (**frequency**) or percent (**relative frequency**) of individuals that fall within each category.
- **Pie charts** and **bar graphs** display the distribution of a categorical variable. *(All bars in a bar graph should have the same width; a change in area could be **misleading**)*
- A **two-way table** of counts organizes data about two categorical variables measured for the same set of individuals.

- The **marginal distribution** of one of the categorical variables in a two-way table of counts is the distribution of values of that variable among all individuals described by the table.
- A **conditional distribution** of a variable describes the values of that variable among individuals who have a specific value of another variable.
- An **association** is a relationship between two variables if knowing the value of one variable helps predict the value of the other.

1.2 Displaying Quantitative Data with Graphs

Definition 1.1 – Describing Shape.

$$\textbf{Describing Shape}\begin{cases}\text{Type}\begin{cases}\text{Unimodal}\\\text{Bimodal}\\\text{Mutimodal}\end{cases}\\\text{Shape}\begin{cases}\text{Symmetric}\\\text{Skewed to the right}\\\text{Skewed to the left}\end{cases}\\\text{Center}\implies\text{Midpoint}(\textit{median})\\\text{Spread}\implies\text{The range from }\textit{Maximum}\text{ to }\textit{Minimum}\\\text{Outliers}\implies\begin{cases}>Q_3+(1.5\times IQR)\\<Q_1-(1.5\times IQR)\end{cases}\end{cases}\tag{1.1a}$$

Definition 1.2 – Histogram.

- **Histogram** is an estimate of the probability distribution of a continuous variable (quantitative variable).

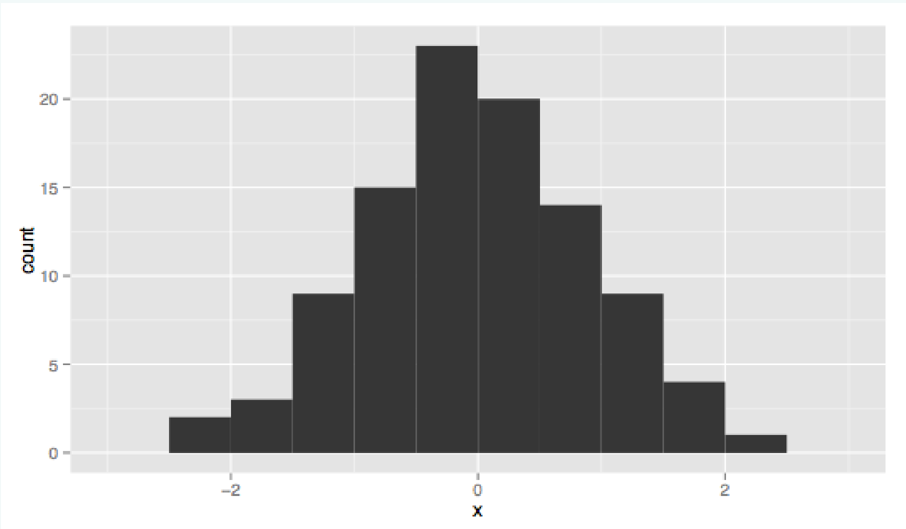

Figure 1.1: A symmetric, unimodal histogram

- **Remember**: histogram are for quantitative data; bar graphs are for categorical data. Also, be sure to use relative frequency histograms when comparing data sets of different sizes.

Definition 1.3 – Dotplot.

- A **Dotplot** is a representation of a distribution consists of group of data points plotted on a simple scale. Dotplots are used for continuous, quantitative, univariate data.

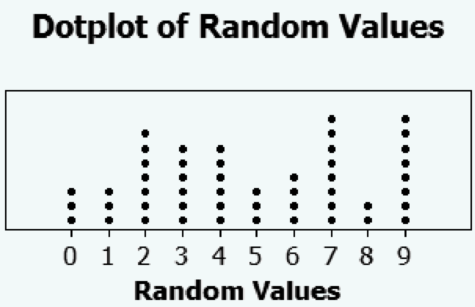

Figure 1.2: A dotplot of 50 random values from 0 to 9.

Definition 1.4 – Stemplot.

- A **Stemplot** is a complicated device for presenting quantitative data in a graphical format, similar to a histogram, to assist in visualizing the shape of a distribution.

Monday section		Wednesday section
9	4	
8 6 1	5	
8 3 0 0	6	2 3 9
9 8 6 6 2 1	7	2 6
9 9 8 8 8 8 7 6 6 4 1	8	1 1 2 3 5 6 8 8 9
5 5 1 0	9	0 2 2 2 5 5 5 5

Figure 1.3: A back-to-back stemplot

1.3 Describing Quantitative Data with Numbers

Definition 1.5 – Mean and median.

- The **Mean** is the arithmetic average of a set of number.

Definition 1.3.1 – Mean.

$$\bar{x} = \frac{\text{sum of observations}}{n} = \frac{x_1 + x_2 + \cdot + x_n}{n} = \frac{\sum x_i}{n} \tag{1.2a}$$

- The **Median** is the midpoint of a distribution, the number such that half the observations are smaller and half are larger.

- **Remember**: the median is a **resistant** measure if center because it is relatively unaffected by extreme observations. The mean is **nonresistant**. Among the measures of spread, the *IQR* is **resistant**, but the standard deviation and range are **nonresistant**.

Definition 1.6 – Measuring Spread.

- **Interquartile range** (IQR)

Definition 1.3.2 – Interquartile range (IQR).

$$IQR = Q_3 - Q_1 \tag{1.3a}$$

- **The five-number summary**

$$\text{The five-number summary} \begin{cases} \text{Minimum} \\ Q_1 \implies \text{The } \textbf{first quartile} \\ \text{Median} \\ Q_3 \implies \text{The } \textbf{third quartile} \\ \text{Maximun} \end{cases} \tag{1.4a}$$

- **Outliers**

Definition 1.3.3 – Outlier.

$$\text{Outlier} \begin{cases} > Q_3 + (1.5 \times IQR) \\ < Q_1 - (1.5 \times IQR) \end{cases} \tag{1.5a}$$

- **Standard Deviation** s_x measures the typical distance of the values in a distribution from the mean.

Definition 1.3.4 – Standard Deviation.

$$S_x = \sqrt{\frac{1}{n-1}\sum(x_i - \bar{x})^2} \tag{1.6a}$$

- **Variance** s_x^2 is the average squared deviation of a set of number.

Definition 1.3.5 – Variance.

$$S_x^2 = \frac{(x_1 - \bar{x})^2 + (x_2 - \bar{x})^2 + \cdots + (x_n - \bar{x})^2}{n-1} = \frac{1}{n-1}\sum(x_i - \bar{x})^2 \tag{1.7a}$$

- **Remember**: The mean and standard deviation are good descriptions for roughly **symmetric** distributions without outliers. The median and *IQR* are a better description for **skewed** distributions.

Chapter 2: Modeling Distributions of Data

2. Modeling Distributions of Data

2.1 Describing Location in a Distribution

Definition 2.1 – Measuring Positions.

- **Percentile**: the ***p*th percentile** of a distribution is the value with p **percent** of the observations less than it.
- **Standardized Score** ($\mathbb{Z}$-Score): if x is an observation from a distribution that has known mean and standard deviation, the **standardized score** for x is

  Definition 2.1.1 – Normal Distribution.

  $$\mathbb{Z} = \frac{x_i - \mu}{\sigma} \tag{2.1a}$$

- A **cumulative relative frequency graph** allows us to examine location within a distribution, beginning by grouping the observations into equal-width classes.
- For a common **transform data** like changing units of measurement:

  $$\textbf{Add} \text{ a constant } a \begin{cases} \textbf{Median, mean, quartiles, and percentiles} \Longrightarrow \underline{\text{increase}} \text{ by } a. \\ \textbf{Spread} \Longrightarrow \text{ do not change.} \end{cases} \tag{2.2a}$$

  $$\textbf{Multiply} \text{ a constant } b \begin{cases} \textbf{Median, mean, quartiles, percentiles} \Longrightarrow \underline{\text{multiply}} \text{ by } b. \\ \textbf{Spread} \Longrightarrow \text{ also } \underline{\text{multiply}} \text{ by } b. \end{cases} \tag{2.3a}$$

- **Neither** of these transformations changes the shape of the distribution.

2.2 Density Curves and Normal Distributions

Definition 2.1 – Density Curves.

- A **density curve** is that

$$\begin{cases}\text{is always on or above the horizontal axis, and}\\ \text{has area exactly 1 underneath it.}\end{cases} \tag{2.4a}$$

Definition 2.2 – Normal Distribution.

- A **Normal Distribution** is described by a Normal density curve.
  The **mean** of a Normal distribution μ is at the center of the symmetric **Normal curve**.
  The **standard deviation** σ is the distance from the center to the change-of-curvature points on either side.

- We **abbreviate** the Normal distribution with mean μ and standard deviation σ as $N(\mu, \sigma)$.

  Definition 2.2.1 – Normal Distribution.

$$f(x|\mu,\sigma) = \frac{1}{\sigma\sqrt{2\pi}}e^{-\frac{(x-\mu)^2}{2\sigma^2}} \tag{2.5a}$$

- **The 68-95-99.7 rule**:
  For a Normal distribution with mean μ and standard deviation σ:
  68% of the observations fall within σ of the mean μ.
  95% of the observations fall within 2σ of the mean μ.
  99.7% of the observations fall within 3σ of the mean μ.

  Definition 2.2.2 – The 68-95-99.7 rule.

$$\text{Three Sigma}\begin{cases}\int_{-1}^{1}\frac{1}{\sqrt{2\pi}}e^{-\frac{1}{2}x^2}\mathrm{d}x = \mathbf{68.2\%} & \text{(within 1 SD)}\\[2ex] \int_{-2}^{2}\frac{1}{\sqrt{2\pi}}e^{-\frac{1}{2}x^2}\mathrm{d}x = \mathbf{95.4\%} & \text{(within 2 SD)}\\[2ex] \int_{-3}^{3}\frac{1}{\sqrt{2\pi}}e^{-\frac{1}{2}x^2}\mathrm{d}x = \mathbf{99.7\%} & \text{(within 3 SD)}\end{cases} \tag{2.6a}$$

- The **standard Noraml distribution** is the Normal distribution with mean 0 and standard deviation 1.

Definition 2.2.3 – Standard Normal Distribution.

$$\mathbb{Z} = \frac{x_i - \mu}{\sigma} \Rightarrow \text{Standardization} = \frac{\text{Obs} - \text{Mean}}{\text{SD}} \tag{2.7a}$$

$$f(x|0,1) = \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}x^2} \tag{2.7b}$$

$$\int_{-\infty}^{\infty} \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}x^2} \mathrm{d}x = 1 \tag{2.7c}$$

Chapter 3: Describing Relationships

3. Describing Relationships

3.1 Scatterplots and Correlation

Definition 3.1 – Scatterplots.

- A **scatterplot** displays the relationship between two quantitative variables measured on the same individuals.
- Explanatory and response variable:

$$\begin{cases} x \Longrightarrow \textbf{Explanatory variable} \\ y \Longrightarrow \textbf{Response variable} \end{cases} \tag{3.1a}$$

Definition 3.2 – Describing Scatterplots.

- **Describing Scatterplots**

$$\text{Scatterplots} \begin{cases} \textbf{Direction} \text{ (gengeral point trend)} \begin{cases} \text{up to right: positive association} \\ \text{down: negative association} \end{cases} \\ \\ \textbf{Form} \text{ (shape of direction)} \begin{cases} \text{Linear} \\ \text{Curved} \end{cases} \\ \\ \textbf{Strength} \text{ (how closely graph fits form)} \begin{cases} \text{Strong} \\ \text{Weak} \end{cases} \\ \\ \textbf{Outlier} \text{ (a departure falls outside the overall relationship)} \end{cases} \tag{3.2a}$$

Definition 3.3 – Measuring Linear Association.

- The **correlation** r measures the direction and strength of the linear relationship between two quantitative variables.

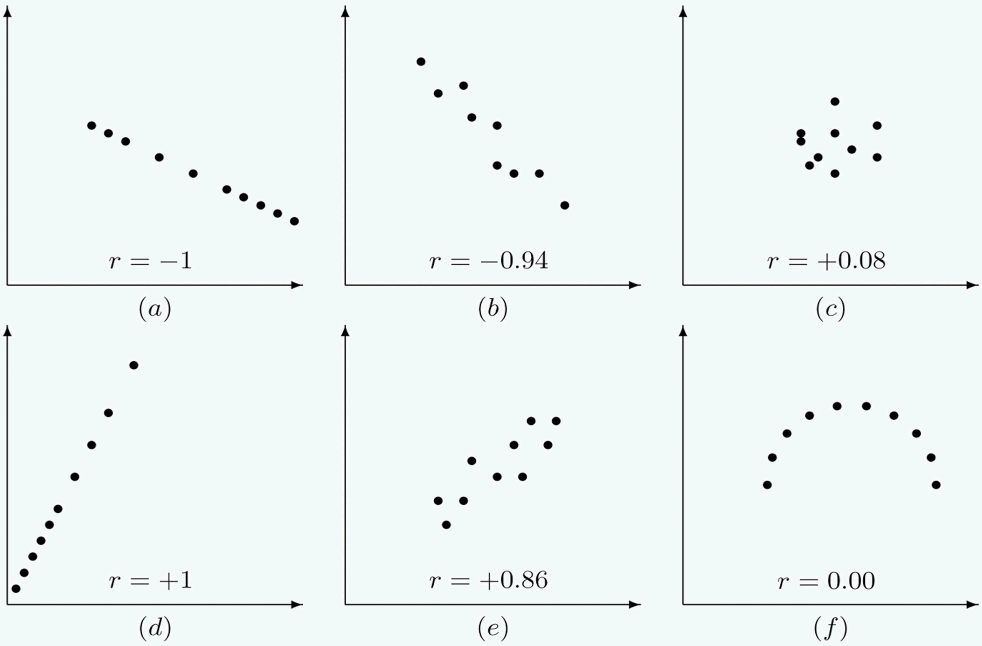

Figure 3.1: Plots with different correlations

Definition 3.1.1 – Correlation.

$$r = \frac{Z_{x_1}Z_{y_1} + Z_{x_2}Z_{y_2} + \cdots + Z_{x_n}Z_{y_n}}{n-1} \tag{3.3a}$$

$$= \frac{1}{n-1}\sum_{i=1}^{n}\left(\frac{x_i - \bar{x}}{S_x}\right)\left(\frac{y_i - \bar{y}}{S_y}\right) \tag{3.3b}$$

$$= \frac{1}{n-1}\sum Z_x Z_y \qquad \left(Z = \frac{x_i - \bar{x}}{S_x}\right) \tag{3.3c}$$

- **Features of Correlation:**

  – $r = 1 \implies$ **linear**, perfect positive correlation.

  – $r > 0 \implies$ **positive** association.

  – $r < 0 \implies$ **negative** association.

  – $r = -1 \implies$ **linear**, perfect negative correlation.

  – $r = 0 \implies$ **doesn't** guarantee there's **no** relationships between two variables, just **No** linear relationship.

- **only** measures the strength of a **linear relationship**.
- can **Not** conclude that change in one variable cause in the other.
- is the **same** when you **inverse** x and y.
- is the **same** when you change the unit.
  e.g: height in inches or meters, weight in pounds or kilograms.
- is **Not** resistant to outliers.
- is **Not** a complete summary of two-variable data.
- both variables must be **quantitative**.

3.2 Least-Squares Regression

Definition 3.1 – Regressions.

- A **Regression line** is a line that describes how a response variable y changes as an explanatory variable x changes.

  Definition 3.2.1 – Regression line, predicted value, slope, y intercept.
  A **regression line** relating y to x has an equation of the form:

  $$\hat{y} = a + bx \tag{3.4a}$$

  - $\hat{y}$ is the **predicted value**.
  - b is the **slop**.
  - a is the y **intercept** (y value when $x = 0$).

- **Extrapolation** is the use of regression line for prediction far outside the interval of values of the explanatory variable x used to obtain the line.

Definition 3.2 – Residuals and the Least-Squares Regression Line.

- A **residual** is the difference between an observed value of the response variable and the value predicted by the regression line.

  Definition 3.2.2 – Residual.

  $$\begin{aligned} \text{residual} &= \text{observed } y - \text{predicted } y \\ &= y - \hat{y} \end{aligned} \tag{3.5a}$$

- The **least-squares regression line** of y on x is the line that makes the sum of the squared residuals as small as possible.

Definition 3.2.3 – Least-squares regression line.
The least-squares regression line is the line $\hat{y} = a + bx$ with **slope**

$$b = r\frac{s_y}{s_x} \tag{3.6a}$$

and y **intercept**

$$a = \bar{y} - b\bar{x} \tag{3.6b}$$

The least-squares regression line always passes through the point $(\bar{x}, \bar{y})$

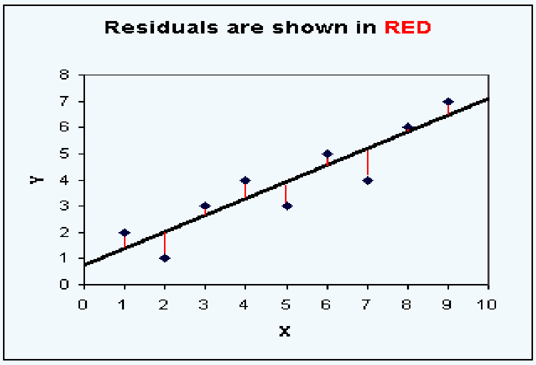

Figure 3.2: A Least-squares regression line

- A **residual plot** is a scatterplot of the residuals against the explanatory variable, helping us access whether a linear model is appropriate.

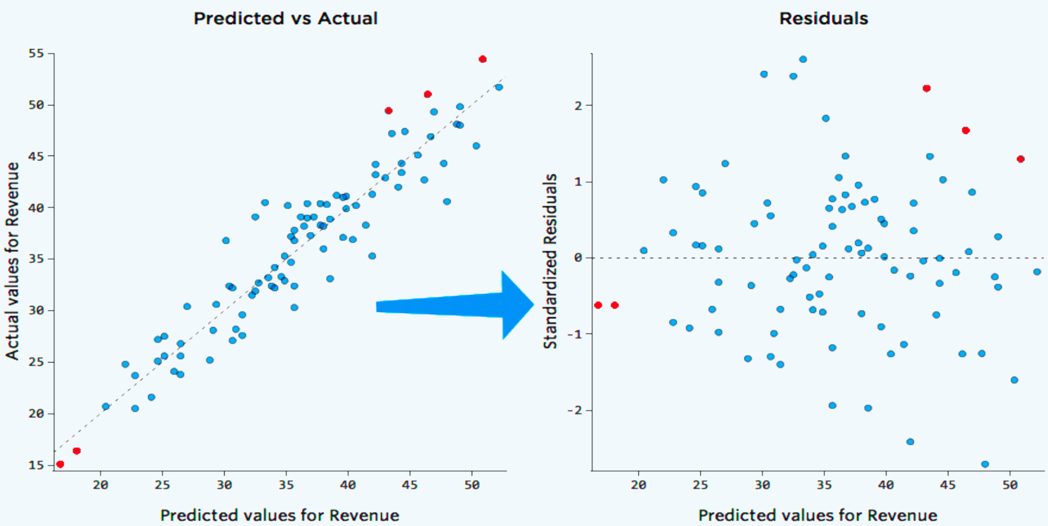

Figure 3.3: A residual plot

Definition 3.3 – The Role of s and r^2 in Regression.

- If we use a least-squares line to predict the values of a response variable y from an explanatory variable x, the **standard deviation of the residuals** s is

Definition 3.2.4 – Standard deviation of the residuals s.

$$s = \sqrt{\frac{\sum \text{residuals}^2}{n-2}} = \sqrt{\frac{\sum (y_i - \hat{y})^2}{n-2}} \tag{3.7a}$$

This value gives the approximate size of a typical **prediction error** (residual)

- The **coefficient of determination** r^2 is the fraction of the variation in the values of y that is accounted for by the least-squares regression line of y on x. (The **percentage** of how well the line fits those data)

Definition 3.2.5 – Coefficient of determination r^2.

$$r^2 = 1 - \frac{\sum \text{residuals}^2}{\sum (y_i - \bar{y})} = 1 - \frac{\sum(\text{Obs}-\text{prediction})^2}{\sum(\text{Obs}-\text{mean Obs})^2} \tag{3.8a}$$

Definition 3.4 – Outliers and influential observations in regression.

- An **outlier** is an observation that lies outside the overall pattern of the other observation. Points that are outliers in the y direction but not the x direction of a scatterplot have large residuals. Other outliers may not have large residuals.

- An observation is **influential** for a statistical calculation if removing it would markedly change the result of the calculation. Outliers in x are often influential for the regression line.

IV

Chapter 4: Designing Studies

4. Designing Studies

4.1 Sampling and Surveys

Definition 4.1 – Population, census,and sample.

- The **population** in a statistical study is the entire group of individuals we want information about.
- A **census** collect data from every individual in the population.
- A **sample** is a subset of individuals in the population from which we actually collect data.
- A **survey** (sample survey) is used to infer statistics of a population.

Definition 4.2 – How to Sample Badly.

- **Convenience sample** is to choose individuals from the population who are easy to reach results.
- A design of a statistical study with **bias** would consistently underestimate or consistently overestimate the value you want to know.
- A **voluntary response sample** consist of people who choose themselves by responding to a general invitation.

Definition 4.3 – How to Sample Well.

- **Random sampling** involves using a chance process to determine which members of a population are included in the sample.
- A **simple random sample** (SRS) of size n is chosen in such a way that every group of n individuals in the population has an equal chance to be selected as the sample.

Definition 4.4 – Other Random Sampling Methods.

- To get a **stratified random sample**, start by classifying the population into groups of similar individuals, called **strata**. Then choose a separate SRS in each stratum and combine these SRSs to form the sample.
- To get a **cluster sample**, start by classifying the population into groups of individuals taht are located near each other, called **clusters**. Then choose an SRS of the clusters. All individuals in the chosen clusters are included in the sample.

Definition 4.5 – Sample Surveys: What Can Go Wrong?.

- **Undercoverage** occurs whens one members of the population cannot be chosen in a sample.
- **Nonresponse** occurs when an individual chosen for the sample can't be contacted or refuses to participated.

$$\textbf{Wording of questions} \text{ has a big influence on the answer.}$$
$$\underbrace{\text{Incorrect answers} \Longrightarrow \text{Response bias}}$$

4.2 Experiments

Definition 4.1 – Observational Study VS Experiment.

- An **observational study** observes individuals and measures variables of interest but does not attempt to influence the response.
- An **experiment** deliberately imposes some treatment on individuals to measure their responses.
- **Confounding** occurs when two variables are associated in such a way that their effects on a response variable cannot be distinguished from each other.

Definition 4.2 – The Language of Experiments.

- **Treatment** is the condition applied to subjects in an experiment.
- The **experimental units** are the smallest collection of individuals to which treatments are applied. When they are human beings, they often are called **subjects**.

Definition 4.3 – How to Experiment Badly.

- In an experiment, **random assignment** means that experimental units are assigned to treatments using a chance process.

  Principles of Experimental Design:

$$\text{A well experiment}\begin{cases}\textbf{Must} \text{ make a } \boxed{\text{comparison}} \text{ between treatments.}\\ \textbf{Must} \text{ \underline{randomly assign} subjects to treatments.}\\ \textbf{Must} \text{ \underline{control} all other variables that affect the resp-}\\ \text{onse to be the same for all groups.}\\ \textbf{Must} \text{ slect enough subjects} \Longrightarrow \boxed{\text{replication}}.\end{cases} \tag{4.2a}$$

  A good control minimizes confounding and reduces variability in the response.

Definition 4.4 – Completely Randomized Designs.

- In a **completely randomized design**, the experimental units are assigned to the treatments completely by chance.
  All experiments need a control (no treatment or placebo group)

Definition 4.5 – Experiments: What Can Go Wrong?.

- In a **double-blind** experiment, neither the subjects nor those who interact with them and measure the response variable know which treatment a subject received. If one party knows and the other doesn't, then the experiment is **single-blind**.

Definition 4.6 – Inference for Experiments.

- **Statistically significant** is an observed effect so large that it would rarely occur by chance.

Definition 4.7 – Blocking.

- A **block** is a group of experimental units that are known before the experiment to be similar in some way that is expected to affect the response to the treatments.

- In a **randomized block design**, the random assignment of experimental units to treatments is carried out separately within each block.

4.3 Using Studies Wisely

Definition 4.6 – Inference about a population.

- **Inference about a population** requires that the individuals taking part in a study be randomly selected from the population. A well-designed experiment that randomly assigns experimental units to treatments allows **inference about cause and effect**.

- **Lack of realism** in an experiment can prevent us from generalizing its results.

- Any information about the individuals in the study must be kept **confidential**

V

Chapter 5: Probability

5. Probability

5.1 Randomness, Probability, and Simulation

Definition 5.1 – The Idea of Probability.

- **The Law of Large Numbers** says that the proportion of times that a particular outcome occurs in many repetitions will approach a single number, which is the **possibility**.
- **Probability** is a number between 0 and 1 describing the proportion of the time the outcome would occur over the long run.

Definition 5.2 – Simulation.

- A **simulation** is an imitation of chance behavior, most often carried out with random number.

  Four-step process of simulation:

$$\text{Simulation}\begin{cases}\textbf{State}\text{: Ask a question of intertest about some chance process.}\\ \textbf{Plan}\text{: Describe one repetition of process.}\\ \textbf{Do}\text{: Perform many repetitions of the simulation.}\\ \textbf{Conclude}\text{: Use the results of your simulation to answer the}\\ \text{question of interest.}\end{cases} \tag{5.1a}$$

5.2 Probability Rules

Definition 5.1 – Probability Models.

- The **sample space** S of chance process is the set of all possible outcomes.
- A **probability model** is a description of some chance process that consists of two parts: a sample space S and a probability for each outcome.
- An **event** is any collection of outcomes from some chance process.

Definition 5.2 – Basic Rules of Probability.

- For any event A, $0 \leqslant P(A) \leqslant 1$
- If S is the sample space in a probability model, $P(S) = 1$.
- $P(A) = \dfrac{\text{number of outcomes corresponding to event A}}{\text{total number of outcomes in sample space}}$
- **Complement rule**: $P(A') = 1 - P(A)$ $\quad P(A) = P(A \cap B) + P(A \cap B')$.
- **Mutually exclusive** (disjoint): two events A and B have no outcomes in common and so can never occur together–that is, if $P(A \cap B) = 0$.
  In other words: $P(A \cup B) = P(A) + P(B)$.

Definition 5.3 – General Addition Rule For Two Events.

- If A and B are any two events resulting from some chance process, the

  Definition 5.2.1 – General Addition Rule For Two Events.

  $$P(A \cup B) = P(A) + P(B) - P(A \cap B) \tag{5.2a}$$

5.3 Conditional Probability and Independence

Definition 5.3 – Conditional Probability.

- **Conditional Probability** is the possibility that one event happens given that another event is already known to have happened. Suppose wen know that event A has happened. Then the probability that event B happens given that event A has happened is denoted by $P(B|A)$.

Definition 5.4 – Calculation Conditional Probabilities.

- To find the conditional probability $P(A|B)$, use formula

$$P(A|B) = \frac{P(A \cap B)}{P(B)} \tag{5.3a}$$

- The conditional probability $P(B|A)$ is given by

$$P(B|A) = \frac{P(B \cap A)}{P(A)} \tag{5.4a}$$

Definition 5.5 – General Multiplication Rule.

- The probability that event A and$B$ both occur can be found using the **General Multiplication Rule**

$$P(A \cap B) = P(A) \cdot P(B|A) \tag{5.5a}$$

where $P(B|A)$ is the conditional probability taht event B occurs given that event A has already occurred.

Definition 5.6 – Independent events.

- When events A and B are **independent**:

$$P(A|B) = P(A) \quad \text{and} \quad P(B|A) = P(B) \tag{5.6a}$$

where $P(B|A)$ is the conditional probability that event B occurs given that event A has already occurred.

Definition 5.3.1 – Multiplication rule for independent events.

$$P(A \cap B) = P(A) \cdot P(B) \tag{5.7a}$$

VI

Chapter 6: Random Variables

6. Random Variables

6.1 Discrete and Continuous Random Variables

Definition 6.1 – Random variable and probability distribution.

- A **random variable** takes numerical values that describe the outcomes of some chance process.
- The **probability distribution** of a random variable gives its possible values and their probability.

Definition 6.2 – Discrete Random Variables.

- A **Discrete random variable** X takes a **fixed** set of possible values with gaps between. The probability distribution of a discrete random variable X lists the values x_i and their probabilities p_i:

Value:	x_1	x_2	x_3	$\cdots$
Probability:	p_1	p_2	p_3	$\cdots$

  The probabilities p_i must satisfy two requirements:

  - Every probability p_i is a number between 0 and 1.
  - The sum of the probabilities is 1: $p_1 + p_2 + p_3 + \cdots + p_n = 1$

Definition 6.3 – Mean (Expected Value) of a Discrete Random Variable.

- Suppose that X is a discrete random variable with probability distribution

Value:	x_1	x_2	x_3	$\cdots$
Probability:	p_1	p_2	p_3	$\cdots$

  To find the **mean (expected value)** of X, multiply each possible value by its probability, the add all the products:

Definition 6.1.1 – Mean (Expected Value).

$$\begin{aligned}\mu_x = E(X) &= x_1p_1 + x_2p_2 + x_3p_3 + \cdots \\ &= \sum x_ip_i\end{aligned} \tag{6.1a}$$

Definition 6.4 – Variance and standard deviation of a discrete random variable.

- Suppose that X is a discrete random variable with probability distribution

Value:	x_1	x_2	x_3	$\cdots$
Probability:	p_1	p_2	p_3	$\cdots$

and that μ_x is the mean of X. The **variance** of X is

Definition 6.1.2 – Variance of a discrete random variable.

$$\begin{aligned}\mathrm{Var}(X) = \sigma_x^2 &= (x_1-\mu_x)^2p_1 + (x_2-\mu_x)^2p_2 + (x_3-\mu_x)^2p_3 + \cdots \\ &= \sum (x_i-\mu_x)^2p_i\end{aligned} \tag{6.2a}$$

The **standard deviation** of X, σ_x, is the square root of the variance.

Definition 6.1.3 – Standard deviation of a discrete random variable.

$$\sigma_x = \sqrt{\sum (x_i-\mu_x)^2p_i} \tag{6.3a}$$

Definition 6.5 – Continuous Random Variables.

- A **Continuous random variable** X takes all values in an interval of numbers. The probability distribution of X is described by a density curve.

6.2 Transforming and Combining Random Variables

Definition 6.1 – Linear Transformations.

- Adding (or subtracting) each value of a random variable by a positive number a:

  Definition 6.2.1 – Effect On A Random Variable of Adding (or subtracting) by A Constant.

  – Mean, Medeian, Quartiles, and percentiles $+ (-)\ a$

  – Does not change Range, IQR, Standard deviation

- Multiplying (or dividing) each value of a random variable by a positive number b:

  Definition 6.2.2 – Effect On A Random Variable of Multiplying (or Dividing) by A Constant.

  – Mean, Medeian, Quartiles, and percentiles $\times (\div)\ b$

  – Range, IQR, and Standard deviation $\times (\div)\ b$

  – Does not change the slop of the distribution

Definition 6.2 – Combining Random Variables.

- Mean of the Sum of Random Variables:

$$E(T) = \mu_T = \mu_x + \mu_y \tag{6.6a}$$

- Range of the Sum of Random Variables:

$$\text{range of } T = \text{range of } X + \text{range of } Y \tag{6.7a}$$

- Variance of the Sum of Random Variables:

$$\sigma_T^2 = \sigma_X^2 + \sigma_Y^2 \tag{6.8a}$$

- Mean of the Difference of Random Variables:

$$\mu_D = E(D) = \mu_x - \mu_y \tag{6.9a}$$

- Variance of the Difference of Random Variables:

$$\sigma_D^2 = \sigma_X^2 + \sigma_Y^2 \tag{6.10a}$$

6.3 Binomial and Geometric Random Variables

Definition 6.6 – Binomial Settings and Binomial Random Variables.

- A **Binomial setting** consists of n independent trials of the same chance process, each resulting in a success or a failure, with probability of success p on each trail.

$$\textbf{BINS}\begin{cases}\text{Trails can be classified as "success" or "failure."}\\ \text{Trails must be independent.}\\ \text{The number of trails } n \text{ must be fixed.}\\ \text{There is the same probabilty } p \text{ of success on each trail.}\end{cases}$$

- The count X of successes is a **Binomial Random Variable**. Its probability distribution is a **Binomial Distribution**.

Definition 6.7 – Binomial Probabilities.

- The **Binomial Coefficient**

  Definition 6.3.1 – Binomial Coefficient.

  $$\binom{n}{k} = \frac{n!}{k!(n-k)!} \tag{6.12a}$$

  counts the number of ways k successes can be arranged among n trails. The **factorial** of n is

  $$n! = n(n-1)(n-2)\cdots(3)(2)(1) \tag{6.13a}$$

  for positive whole numbers n, and $0! = 1$

- **Binomial Probability Formula**

  Definition 6.3.2 – Binomial Probability.

  $$P(X = K) = \binom{n}{k}p^k(1-p)^{n-k} \tag{6.14a}$$

  If a count X of successes has the binomial distribution with number of trials n and probability of success p, the **mean** and **standard deviation** of X are

  Definition 6.3.3 – Mean and Standard deviation of Binomial distribution.

  $$\mu_x = np \tag{6.15a}$$

  $$\sigma_x = \sqrt{np(1-p)} \tag{6.15b}$$

Definition 6.8 – Binomial Distributions in Statistical Sampling.

- **10% Condition**
  When taking an simple random sample of size n from a population of size N, we can use a binomial distribution to model the count of successes in the sample as long as $n \leq \frac{1}{10}N$.

- **The Large Counts Condition**
  Suppose that a count X of successes has the binomial distribution with n trails and success probability p. When n is large, the distribution of X is approximately Normal with

  $$\text{mean: } \mu_x = np \text{ and standard deviaion: } \sigma_X = \sqrt{np(1-p)} \tag{6.16a}$$

  As an approximation, we will use the Normal approximation when$n$ is no longer that

  $$np \geq 10 \qquad \text{and} \qquad n(1-p) \geq 10 \tag{6.17a}$$

  That is, the expected number of **successes** and **failures** are both at **least 10**. We refer to this as the **Large Counts condition**.

Definition 6.9 – Geometric Random Variables.

- A **Geometric setting** consists of repeated trails of the same chance process in which the probability p of successes is the same on each trail, and the goal is to count the number of trails it takes to get one success.

- If Y = the number of trails required to obtain the first success, then Y is a **Geometric probability** that Y takes any value is

  Definition 6.3.4 – Geometric Probability.

  $$P(Y = K) = (1-p)^{k-1}p \tag{6.18a}$$

  The **mean** (expected value) of a geometric random variable Y is

  Definition 6.3.5 – Mean of Geometric Probability.

  $$\mu_Y = E(Y) = \frac{1}{p} \tag{6.19a}$$

  which is the expected number of trails required to get the first success.

Index