Statistical Methods for the Biological, Environmental, and Health Sciences

STAT 007

Describing, Exploring, and Comparing Data

Chapter 3

Measures of Relative Standing

Section 3-3

- In this section we will:
  - Introduce a standardized score to compare data sets and a rule of thumb to identify significantly low or large values in a data set.
  - Define Outliers and resistant statistics.

- Measures of relative standing describe the location of data values *relative* to other values within the same data set.
- Here we discuss *z*Scores as a measure of relative standing.
- A *z*Score for a value x is found by converting that value to a standardized scale.
- *z*Scores allow us to compare values from different data sets.
- Other measures of position for comparing values within the same data set or between different data sets are quartiles.
- Quartiles are also used to identify potential outlier values.

*z*Scores

- ***z*Scores** (or standard score or standard value): is the number of standard deviations that a given value x is above or below the mean.
- The *z*Score can be computed for a sample: $z = \frac{x-\bar{x}}{s}$ or a population: $z = \frac{x-\mu}{\sigma}$.
- Properties: (i) A *z*score is the number of standard deviations that a given value x is above or below the mean; (ii) *z*Scores are expressed as numbers with no units of measurement; (iii) A data value is *significantly low* if its z score is less than or equal to -2 or the value is *significantly high* if its z score is greater than or equal to $+2$ (iv) If an individual data value is less than the mean, its corresponding z score is a negative number.

Example

Consider the following data of IQ scores:
96; 87; 101; 103; 127; 96; 88; 85; 97; 124
Compute the *z*Scores of each data value and determine whether there are significantly low or high values.
-0.30 -0.92 0.04 0.18 1.83 -0.30 -0.86 -1.06 -0.23 1.63

Outliers

- **Outlier**: is a sample value that lie very far away from the vast majority of the other values in a set of data.
- Outliers can strongly affect the values of some important statistics, such a the mean and standard deviation.
- A statistic is **resistant** if the presence of extreme values or outliers does not cause it to change very much.
- Q_1, Q_2, and Q_3 are the first, second and third **quartiles** of the data set:
  Q_1: divides the first 25% of the *sorted* data from the top 75%.
  Q_2: divides the first 50% of the *sorted* data from the top 50%.
  Q_3: divides the first 75% of the *sorted* data from the top 25%.
- A more specific criteria for identifying outliers:
  values above Q_3 by an amount greater $1.5(Q_3 - Q_1)$
  values below Q_1 by an amount greater $1.5(Q_3 - Q_1)$

Outliers

- Computation of quartiles:
  Q_1: from the sorted data values, is the one in position $L = \frac{25}{100} * n$.
  Q_2: from the sorted data values, is the one in position $L = \frac{50}{100} * n$.
  Q_3: from the sorted data values, is the one in position $L = \frac{75}{100} * n$.
  If L is a whole number, the quartile of interest is $\frac{x_L + x_{L+1}}{2}$.
  If L is not a whole number, round L to the next whole number and the quartile is x_L.

Example

Consider the following data of IQ scores: 96; 87; 101; 103; 127; 96; 88; 85; 97; 124
Note that the ordered data set is: 85; 87; 88; 96; 96; 97; 101; 103; 124; 127;
Are 85 and 127 outliers?

Note that $\frac{25}{100} * 10 = 2.5$, $\frac{50}{100} * 10 = 5$, and $\frac{75}{100} * 10 = 7.5$.

So, $Q_1 = x_3 = 88$, $Q_2 = \frac{x_5 + x_6}{2} = \frac{96+97}{2} = 96.5$, and $Q_3 = x_8 = 103$.

So, outliers are values smaller than $Q_1 - 1.5 * (Q_3 - Q_1) = 88 - 1.5 * 15 = 65.5$ and greater than $Q_3 + 1.5 * (Q_3 - Q_1) = 103 + 1.5 * 15 = 125.5$.

Practice

Look at the exercises at the end of Section 3-3 in page 112

Specially, look at exercises:
1, 4, 5, 6, 7, 9, 10, 11, 12, 13, 14, 15, 16

Probability

Chapter 4

Basic Concepts of Probability

Section 4-1

- In this section we will:
  - Introduce the concept of probability, some basic definitions and notations.
  - Discuss different methods for computing probabilities of events.
  - Discuss how to compute the probability of something not happening (the complement).

- Probabilities are used in processes that involve *uncertainty*.
- Probabilities are values between 0 and 1 (both included).
- Probabilities are a key concept in the statistical method of *hypothesis testing* (Chapter 8).

  For example: *such* gender selection method increases chances of having a baby girl.

- Statisticians make decisions using (sample) data by rejecting explanations (such as chance) based on very low probabilities.

  For example: from out of 100 births, it is observed that 75 babies are girls, in couples using *such* gender selection method. The probability of such an *event* due to chance is less than 0.000001. Then, we reject the explanation of chance being the reason for such an event and conclude that is the gender selection method the one that increases the chances of having a baby girl.

Basics of Probability

- In probability, we deal with procedures that produce outcomes.
- Some definitions:
  - An **event** is any collection of outcomes (or results) of a procedure.
  - A **simple event** is an outcome or an event that cannot be further broken down into simpler components.
  - The **sample space** for a procedure consists of all possible *simple events*.

Example

Consider the procedure of *sequentially* sampling at random three students from a class and noting their gender. Let "g" denote a girl and let "b" denote a boy.

a) Describe an event.
b) Describe a simple event.
c) Describe the sample space.

Some notation and facts

- P denotes a probability.
- A, B, and C denote specific events.
- $P(A)$ denotes the "probability of event A occurring".
- $0 \leq P(A) \leq 1$.
- If an event A cannot occur (*impossible event*), then $P(A) = 0$.
- If an event A is a *certain event*, then $P(A) = 1$.

Finding the Probability of an Event

- Let A be an event of interest.
- **Relative Frequency Approximation**: conduct (or observe) a procedure and count the number of times event A occurs.

$$P(A) = \frac{\text{number of times } A \text{ occurred}}{\text{number of times the procedure was repeated}}.$$

  As the number of times the procedure is repeated, the relative frequency probability of an event tends to approach the actual probability (Law of Large Numbers).

- **Classical Approach**: *assume* that a procedure has n different simple events that are *equally likely*. If event A can occur in s different ways, then

$$P(A) = \frac{\text{number of ways } A \text{ occurrs}}{\text{number of different simple events}} = \frac{s}{n}.$$

- **Subjective Approach**: $P(A)$, is *estimated* by using knowledge of the relevant circumstances.

Finding the Probability of an Event

Example

Consider the procedure of *sequentially* sampling at random three students from a class and noting their gender. Let "g" denote a girl and let "b" denote a boy.
Let A be the event that 2 or more girls are sampled.

a) Assume that this procedure was repeated 10 times and the following simple events were observed: *ggg*, *gbb*, *ggb*, *bbb*, *ggb*, *gbg*, *bgb*, *bbb*, *gbg*, *bgg*. Use the frequentist approach to compute $P(A)$. What happens when the procedure is repeated more times?

b) Use the classical approach to compute $P(A)$. Specify the underlying assumptions.

c) Use the subjective approach to compute $P(A)$. Explain your reasoning.

Complementary Events

- Sometimes we need to find the probability that an event A *does not* occur.
- The **complement** of event A, denoted by $\overline{A}$, consists of all outcomes in which event A does not occur.
- The following relationship between A and $\overline{A}$ holds:

$$P(\overline{A}) = 1 - P(A).$$

Example

In recent years, there were about 3 million skydiving jumps and 21 of them resulted in deaths.

Use the frequentist approach to find the probability of *not* dying when making a skydiving jump.

Practice

Look at the exercises at the end of Section 4-1 in page 128

Specially, look at exercises:
1, 2, 3, 4, 5, 6, 7, 8, 13-20, 25, 26, 27, 29, 30, 31, 33, 34, 35, 36

Addition Rule and Multiplication Rule

Section 4-2

- In this section we will:
  - Introduce the addition rule to compute $P(A \text{ or } B)$.
  - Introduce the multiplication rule to compute $P(A \text{ and } B)$.
  - Discuss disjoint and independent events.

- Consider a procedure with some sample space, and consider A and B two events.
- In this section we will discuss how to compute the probability that A occurs, or B occurs, or A and B occur. This is the probability of the "new" event A *or* B: $P(A \text{ or } B)$.
- Be cautious to not *double count* events!
- We will also discuss how to compute the probability that A and B occur. This is, the probability of the "new" event A *and* B: $P(A \text{ and } B)$.
- Be cautious to identify if event A somehow affects the probability of event B.

Addition Rule: **A or B**

- The *Addition Rule* is used to compute the probability of the event *A or B*. This is *P*(*A or B*).
- Intuitively: To find **P(A or B) add** the probability of event *A* and the probability of event *B*, and if there is any overlap that causes double-counting, subtract the probability of outcomes that are included twice.
- Formally: **P(A or B)** $= P(A) + P(B) - P(A \text{ and } B)$.

Example

TABLE 4-1 Results from Drug Tests of Job Applicants

	Positive Test Result (Test shows drug use.)	Negative Test Result (Test shows no drug use.)
Subject Uses Drugs	45 (True Positive)	5 (False Negative)
Subject Does Not Use Drugs	25 (False Positive)	480 (True Negative)

- **Drug Testing of Job Applicants**: The table includes results form 555 adults in the U.S..
- Assume that one subject is randomly chosen, find the probability of selecting a subject who had a positive test result or uses drugs.

Addition Rule for events that are disjoint

- **Definition**: Events A and B are **disjoint** (or mutually exclusive) if they cannot occur at the same time. (Disjoint events do not overlap.)
- If events A and B are disjoint, then $P(A \text{ and } B) = 0$, therefore $P(A \text{ or } B) = P(A) + P(B)$.

Example

Discuss whether the following events are disjoint or not and how does this translate into a mathematical formula.

a) Event A: randomly select a subject for a clinical trial who is a male.
Event B: randomly select a subject for a clinical trial who is a female.

b) Event A: randomly select a subject taking a statistics course.
Event B: randomly select a subject who is a female.

Addition Rule and complementary events

- Principle: we are certain that either and event occurs or it does not occur.
- Use the addition rule to translate the previous principle into a mathematical expression. And show that
  - $P(A) + P(\overline{A}) = 1$.
  - $P(\overline{A}) = 1 - P(A)$.
  - $P(A) = 1 - P(\overline{A})$.

Example

Based on a journal article, the probability of randomly selecting someone who has sleepwalked is 0.292 (based on data from "Prevalence and Comorbidity of Nocturnal Wandering in the U.S. General Population," by Ohayon et al., *Neurology*, Vol. 78, No. 20).

If a person is randomly selected, find the probability of getting someone who has not sleepwalked.

Multiplication Multiplication: **A and B**

- The *Multiplication Rule* is used to compute the probability of the event *A and B*, where events *A* and *B* occur in different trials. Notation: $P(A \text{ and } B)$.
- Intuitively: To find **P(A and B) multiply** the probability of event *A* by the probability of event *B*, but be sure that the probability of event *B* is found by assuming that event *A* has already occurred.
- Formally: **P(A and B)** $= P(A)P(B \mid A)$.
- Events *A* and *B* are **independent** if the occurrence of one does not affect the probability of the occurrence of the other.
- If *A* and *B* are independent events, then $P(A \text{ and } B) = P(A)P(B)$
- If events *A* and *B* are not independent, they are said to be **dependent**.

Multiplication Multiplication: **A and B**

Example

Drug Testing of Job Applicants: Consider only the 50 test results from the subjects who use drugs. The number of subjects that had a positive test result is 45 and the number of subjects that had a negative test result is 5.

a) If 2 of these 50 subjects are randomly selected *with replacement*, find the probability that the first selected person had a positive test result and the second selected person had a negative test result.

b) Repeat part a) by assuming that the two subjects are randomly selected *without replacement*.

Practice

Look at the exercises at the end of Section 421 in page 143

Specially, look at exercises:
1, 2, 4, 5, 6, 7, 9-20, 21-24, 27