Introduction to Statistics Note

2024 Spring Semester

21 CST H3Art

Chapter 1: Looking at Data——Distributions

1.1 Data

Statistics is the science of learning from data.

Cases are the objects described by a set of data. Cases may be customers, companies, experimental subjects, or other objects.

A **variable** is a special characteristic of a case.

A **label** is used in some data sets to provide additional information about a variable.

Different cases can have different **values** of a variable.

A **categorical** variable places each case into one of several groups, or categories. Sometimes also referred to as a nominal variable (i.e., used for naming). Example: first language, ethnicity.

A **quantitative** variable takes numerical values for which arithmetic operations such as adding and averaging make sense.

The **distribution** of a variable tells us the values that a variable takes and how often it takes each value.

1.2 Displaying Distributions with Graphs

The distribution of a **categorical variable** lists the categories and gives the **count** or **percent** of individuals who fall into each category:

- **Pie charts** show the distribution of a categorical variable as a “pie” whose slices are sized by the counts or percents for the categories.
- **Bar graphs** represent categories as bars whose heights show the category counts or percents.

The distribution of a **quantitative variable** tells us what values the variable takes on and how often it takes those values:

- **Histograms** show the distribution of a quantitative variable by using bars. The height of a bar represents the number of individuals whose values fall within the corresponding class.
- **Stemplots** (or **stem-and-leaf plots**) separate each observation into a stem and a leaf that are then plotted to display the distribution while maintaining the original values of the variable.

A distribution is **symmetric** if the right and left sides (or tails) of the graph are approximately mirror images of each other.

A distribution is **skewed to the right** (**right-skewed**) if the right side of the graph (containing the half of the observations with larger values) is much longer than the left side.

It is skewed to the **left** (**left-skewed**) if the left side of the graph is much longer than the right side.

An important kind of deviation is an **outlier**. Outliers are observations that lie outside the overall pattern of a distribution.

1.3 Describing Distributions with Numbers

The most common measure of center (or central tendency) is the **arithmetic average**, or **mean**.

$$\bar{x} = \frac{\text{sum of observations}}{n} = \frac{x_1 + x_2 + \cdots + x_n}{n} = \frac{1}{n}\sum_{i=1}^{n} x_i = \frac{1}{n}\sum x_i$$

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

The **mean** and **median** of a roughly **symmetric** distribution are close together. If the distribution is exactly **symmetric**, the mean and median are exactly **the same**.

In a **skewed distribution**, the **mean** is usually **farther out in the long tail** than is the median.

A measure of center alone can be **misleading**. And a useful numerical description of a distribution requires **both a measure of center and a measure of spread** (or variability), so we introduce **quartiles**:

- Arrange the observations in increasing order and locate the **median** M.
- The **first quartile** Q_1 is the median of the observations located to the left of the median in the ordered list.
- The **third quartile** Q_3 is the median of the observations located to the right of the median in the ordered list.
- The **interquartile range** (IQR) is defined as: $\text{IQR} = Q_3 – Q_1$.

The **five-number summary** of a distribution consists of the **smallest observation**, the **first quartile**, the **median**, the **third quartile**, and the **largest observation**, written in order from smallest to largest:

$$\min \ Q_i \ M \ Q_3 \ \max$$

The median and quartiles divide the distribution roughly into quarters. This leads to a new way to display quantitative data, the **boxplot**:

- Draw and label a number line that includes the range of the distribution.
- Draw a central box from Q_1 to Q_3.
- Note the **median** M inside the box.
- Extend lines (whiskers) from the box out to the **minimum** and **maximum** values that are **not outliers**.

The $1.5 \times \text{IQR}$ Rule for Outliers: Call an observation an outlier if it falls more than $1.5 \times \text{IQR}$ **above the third quartile** or **below the first quartile**.

The most common measure of spread looks at how far each observation is from the mean. This measure is called the **standard deviation**(s_x).

$$\text{variance} = s_x^2 = \frac{(x_1 - \bar{x})^2 + \cdots + (x_n - \bar{x})^2}{n-1} = \frac{1}{n-1}\sum (x_i - \bar{x})^2$$

$$\text{standard deviation} = s_x = \sqrt{\frac{1}{n-1}\sum (x_i - \bar{x})^2}$$

s_x measures spread about the mean and should be used only when the **mean** is **the measure of center**.

s_x is **not** resistant to outliers.

Numerical summaries do not fully describe the shape of a distribution.

$$\text{ALWAYS PLOT YOUR DATA!}$$

Changing the Unit of Measurement:

- **Multiplying** each observation by a positive number b multiplies both **measures of center (mean, median)** and **spread** (IQR, s_x) by b.
- **Adding** the same number a (positive or negative) to each observation adds a to **measures of center** and to **quartiles**, but it does **not** change measures of spread (IQR, s_x).

1.4 Density Curves and Normal Distributions

A **density curve** is a curve that:

- is always **on** or **above** the horizontal axis
- has an **area** of exactly 1 underneath it

Distinguishing the **Median** and **Mean** of a **Density Curve**:

- The **median** of a density curve is the **equal-areas point**——the point that divides the area under the curve in half.
- The **mean** of a density curve is the **balance point**, that is, the point at which the curve would balance if made of solid material.
- The **mean** of a **skewed curve** is **pulled away from the median** in the direction of the **long tail**.

A **Normal distribution** is described by a **Normal density curve**. Any particular Normal distribution is completely specified by two numbers: its **mean** μ and **standard deviation** σ.

- The **mean** of a Normal distribution is the center of the symmetric Normal curve.
- We abbreviate the Normal distribution with mean μ and standard deviation σ as $N(\mu, \sigma)$.

The 68-95-99.7 Rule:

- Approximately 68% of the observations fall within σ of μ.
- Approximately 95% of the observations fall within 2σ of μ.
- Approximately 99.7% of the observations fall within 3σ of μ.

If a variable x has a distribution with **mean** μ and **standard deviation** σ, then the standardized value of x, or its z-score, is:

$$z = \frac{x - \mu}{\sigma}$$

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

One way to assess if a distribution is indeed approximately Normal is to plot the data on a **Normal quantile plot**.

- If the distribution is indeed Normal, the plot will show a **straight line**, indicating a **good match** between the data and a Normal distribution.
- Systematic **deviations** from a straight line indicate a **non-Normal distribution**.
- **Outliers** appear as points that are far away from the overall pattern of the plot.