Data

Definition

Categorical Variables represent data that can be divided into different groups.

Common examples of categorical data include ethnicity, income level, education, age group, and gender. Categories are described by words or letters. Categorical data is much harder to analyze mathematically than numerical data.

Definition

Quantitative Variables represents numerical data from population measurement. Quantitative data can be either **discrete** or **continuous**.

Quantitative data includes height, weight, income, age, and cost. A discrete data set can only take on specific values (e.g. integer values). If the data is not restricted to specific values, then the data set is continuous.

Sampling

Gathering the data on an entire population may be virtually impossible due to limitations in time and cost. Instead, a **sample** of the population will be studied. The sample must be representative of the population being sampled in order for any statistical inference to be made. Random sampling is used to to create samples that minimize any bias from the sampling process.

Definition

A **simple random sample (SRS)** is a sampling method in which each member of a population has an equal chance of being selected.

Remark. *In practice, it may be difficult to achieve a random sample. Random number generators can be powerful tools to add randomization to the sampling process. Two other common types of sampling methods are stratified sampling and cluster sampling. Introductory statistics assumes a simple random sampling process.*

Histograms

Definition

A **Histogram** is a bar chart representing how many data points are present in each specified interval.

To construct a histogram from a data set:

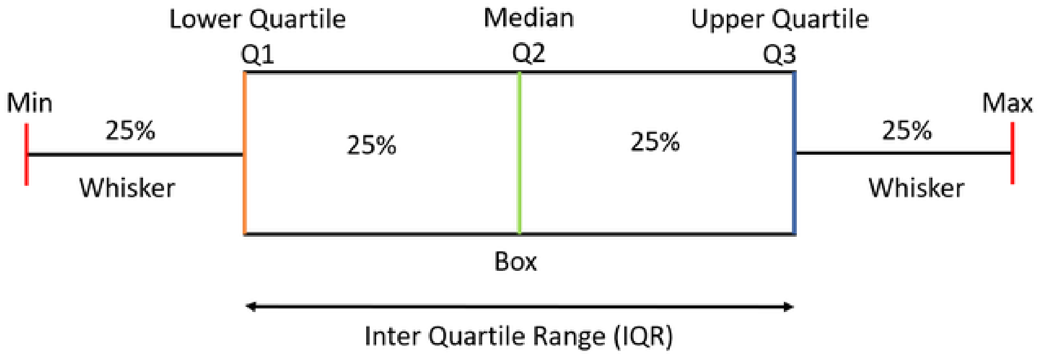
1. Determine number of bins.
2. Find the bin width and determine the bin intervals.
3. Count the number of data points per bin.
4. Plot data point counts vs bins on a bar graph.

Percentiles

Definition

Percentiles measure the location of data relative to the entire data set. The kth percentile is a score below which k percent of the data falls below.

The 25th percentiles is called the **first quartile** Q_1. The 50th percentile is called the **median** or **second quartile** Q_2. The 75th percentile is called the **third quartile** Q_3.



Definition

A **box plot** is a diagram representing five values: the minimum, maximum, and the three quartiles.

To calculate the box plot values on a TI-84 calculator:

1. Press STAT. Choose 1-EDIT. Fill a list with values.
2. Press STAT and arrow to CALC
3. Choose 1-VarStats. Choose list with values entered. ENTER

To plot the boxplot on a TI-84 calculator:

1. STAT EDIT. Fill a list with values.
2. Press 2ND STATPLOT. Turn on the plot. Select box plot icon.
3. Press ZOOM and choose 9-Stat
4. Press TRACE and use arrow keys to view Min, Q_1, Q_2, Q_3 and Max.

Remark. *To calculate percentiles by hand, the data needs to be sorted. The data point that marks the kth percentile is at the kN index of the sorted array. Non-integer indices are rounded up. For example, in a data set $N = 470$, the index of the data point representing the 65th percentile is 306.*

Statistical Mean

Definition

The **mean (arithmetic mean)** is one of the most common measures of center of a data set in statistics. The mean of a data set representing an entire population is called a **population mean** and is denoted μ.

The mean can be calculated by the formula

$$\mu = \frac{1}{N}\sum x$$

where the sum is taken over the entire data set and N represents the population size. The mean of a data set representing a sample from a population is called a **sample mean** and is denoted $\bar{x}$. The formuala for sample mean is identical to the formula for population mean

$$\bar{x} = \frac{1}{n}\sum x$$

except that the sum is taken over the sample data set and n represents the sample size. For datasets described by frequency tables or histograms,

$$\mu = \frac{\sum fm}{\sum f}$$

where f is the frequency of the interval and m is the midpoint of the interval.

Standard Deviation

Definition

The **standard deviation** provides a measure of the overall variation in a data set. The standard deviation can be used to determine whether a data value is close to or far away from the mean.

The **population standard deviation** σ is computed by the formula

$$\sigma = \sqrt{\frac{\sum(x-\mu)^2}{N}}$$

where μ is the population mean and N is the population size. The **sample standard deviation** s has a similar formula

$$s = \sqrt{\frac{\sum(x-\bar{x})^2}{n-1}}$$

where $\bar{x}$ is the sample mean and n is the sample size. Observe that the denominator is $n-1$ not n.

To calculate the descriptive statistics of a data set:

1. Press STAT. Choose 1-EDIT. Fill a list with values.
2. Press STAT and arrow to CALC
3. Choose 1-VarStats. Choose list with values entered. ENTER

The window will list the average, sum of the elements, sum of the squares of the elements, sample standard deviation, population standard deviation, number of elements, minimum, first quartile, median, third quartile, and max value.

Probability

Definition

In an experiment there are different possible outcomes. **Probability** measures how likely an outcome will occur. Probabilities take on values between 0 and 1.

The **sample space** of an experiment is the set of all possible outcomes. An **event** is a set of outcomes. There are two ways to build up events from outcomes. The **OR** event, denoted mathematically as $A \cup B$, is an event containing outcome of A or B or both. The **AND** event, denoted mathematically as $A \cap B$, is an event containing outcomes shared by A and B. The **complement** of an event A is denoted $\neg A$ and consists of all the outcomes in the sample space not in event A.

Definition

The **conditional probability** of event A given B, denoted $P(A|B)$, is the probability that event A will occur if the sample space were restricted to event B.

If the outcomes in a finite sample space of size N are equally likely, then then the probability any one outcome x occuring is given by the formula

$$P(x) = \frac{1}{N}$$

The probability of an event A when all outcomes are equally likely is given by the formula

$$P(A) = \frac{|A|}{N}$$

where $|A|$ denotes the number of outcomes in event A. Finding probabilities reduces to a counting problem. Computations of probabilities of more general experiments require applications of the multiplication and addition principles.

Definition

The **multiplication principle** is used for computing probabilites of AND events.

$$P(A \cap B) = P(A)P(B|A)$$

The **addition principle** is used for computing probabilities of OR events.

$$P(A \cup B) = P(A) + P(B) - P(A \cap B)$$

The multiplication principle can also be seen as a definition for conditional probability

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

Independence

Definition

Two events A and B are **mutually exclusive** if the events do not share any outcomes. In mathematical notation, $A \cap B$ is empty.

Remark. *If events A and B are mutually exclusive, the addition principle becomes*

$$P(A \cup B) = P(A) + P(B)$$

A coinflip has a sample space of $\{H, T\}$. The events $\{H\}$ and $\{T\}$ are mutually exclusive. However, if event $\{H\}$ is observed (i.e. a coin toss comes up heads), then the conditional probability of observing event $\{T\}$ is 0. The original probability of event $\{T\}$ was 0.5, but observing event $\{H\}$ changed the probability of event $\{T\}$. Event $\{T\}$ is dependent on event $\{H\}$.

Definition

Two events are **independent** if the observation of one event does not affect the probability of another event and vice versa. Mathematically, the two events must satisfy

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

Alternatively, two events are independent if

$$P(A|B) = P(A|\neg B) \quad \text{and} \quad P(B|A) = P(B|\neg A)$$

If two events are not indepdent, they are said to be **dependent**. Mutually exclusive events are dependent.

Remark. *If events A and B are independent, then the multiplication principle becomes*

$$P(A \cap B) = P(A)P(B)$$

Contingency Tables

Definition

A **contingency table**, also called a **two way frequency table**, represents the observed frequencies of two categorical variables. The data is organized into a table where rows represent the different categories of the first categorical variable and the columns represent the different categories of the second categoical variable.

Row totals and column totals are added to a contingency table to make calculations of conditional probabilities simpler.

	Y_1	Y_2	Y_3	total
X_1	30	40	80	150
X_2	40	70	10	120
total	70	110	90	270

In the contingency table above, there are two categoricl variables X and Y with categories $\{X_1, X_2\}$ and $\{Y_1, Y_2, Y_3\}$ respectively. Conditional probabilities are the ratio of a cell with a row or column total. The conditioned variable determines whether a row or column total is used.

$$P(X_1|Y_2) = \frac{40}{110} \quad \text{and} \quad P(Y_2|X_1) = \frac{40}{150}$$

Discrete Random Variables

Definition

A **random variable** is a variable whose possible values are numerical outcomes of a random process. Random variables can be **continuous** or **discrete**.

Random variables are determined by thier **probability distribution**. For discrete random variables, the probability distribution lists the probability of each outcome of a random variable such that the sum total of probabilities of the outcomes is 1.

The probability distribution of a fair coin would be given by

$$P(H) = 0.5 \quad \text{and} \quad P(T) = 0.5$$

More generally, consider a single experiment with two outcomes, *success* and $failure$, where *success* has a probability of p. Such an experiment is called a **bernoulli trial**. The probability distribution is given by

$$P(success) = p \quad \text{and} \quad P(failure) = 1 - p$$

Random variables with this type or probability distribution are called **bernoulli random variables**.

Expected Value and Variance

Definition

The **expected value** of a random variable X is a weighted average of the possible outcomes. Mathematically,

$$\mathbb{E}[X] = \sum xP(x)$$

Informally, the expected value represents the average outcome of the experiment if the experiment were done many times.

A slot machine has the following probability distribution function for net winnings

payout	-\$2.00	\$1.00	\$10.00	\$100.00
probability	0.90	0.06	0.03	0.01

The expected value (i.e. the expected winnings) of the slot machine is

$$\mathbb{E}[X] = -2.00(0.9) + 1.00(0.06) + 10.00(0.03) + 100.00(0.01) = -0.17$$

Definition

The **variance** of a random variable X is a weighted average of the squared deviations of the possible outcomes from the mean. Mathematically,

$$\sigma^2 = \sum (x - \mu)^2 P(x)$$

Informally, the variance measures how spread out the outcomes are relative to each other. In statistics, its more common to use the standard deviation σ.

Remark. *When the outcomes of a random variable are all equally likely, the expected value reduces to the familiar formula for statistical mean.*

Binomial Random Variables

Definition

When a bernoulli trial with probability p of success is repeated n times, a **binomial experiment** is performed. The outcomes of a binomial experiment are the number of successes in the experiment.

The **binomial distribution**, denoted $B(n, p)$, is the probability distribution of a binomial experiment. The mean and standard deviation of the binomial distribution are

$$\mu = np \quad \text{and} \quad \sigma = \sqrt{np(1-p)}$$

Geometric Random Variables

Definition