Statistics

Overview

- *Statistics* is the science of collecting, organizing, analyzing, and interpreting data in order to make decisions.
- *Data* consists of information coming from observations, counts, measurements, or responses.
- A *population* is the collection of all outcomes, responses, measurements, or counts that are of interest.
- A *sample* is a subset of the population.
- A *parameter* is a numerical description of a population characteristic.
- A *statistic* is a numerical description of a sample characteristic.

Branches of Statistics

- *Descriptive statistics* is the branch of statistics that involves the organization, summarization, and display of data.
- *Inferential statistics* is the branch of statistics that involves using a sample to draw conclusions about a population. A basic tool in the study of inferential statistics is probability.

Data Classification

- Types of data:
  - *Qualitative data* consist of attributes, labels, or nonnumerical entries.
  - *Quantitative data* consist of numerical measurements or counts.
- Levels of measurement:
  - Nominal: categorized using names, labels, or qualities.
  - Ordinal: can be arranged in order or ranked.
  - Interval: can be ordered and meaningful differences between entries can be calculated.
  - Ratio: similar to interval, but there is a zero entry that is an inherent zero (implies none).

Measures of Central Tendency

- The *mean* of a data set is the sum of the data entries divided by the number of entries.
  - Population mean:
$$\mu = \frac{\sum x}{N}$$
  - Sample mean:
$$\bar{x} = \frac{\sum x}{n}$$
- The *median* of a data set is the value that lies in the middle of the data when the data is in sorted order.
- The *mode* of a data set is the data entry that occurs with the greatest frequency.

Measures of Central Tendency

- An *outlier* is a data entry that is far removed from the other entries in the data set.
- A *weighted mean* is the mean of a data set whose entries have varying weights. A weighted mean is given by:

$$\bar{x} = \frac{\sum x \cdot w}{\sum w}$$

where w is the weight of each entry x.

Measures of Variation

- The *range* of a data set is the difference between the maximum and minimum data entries in the set.
- The *deviation* of an entry x in a population data set is the difference between the entry and the mean μ of the data set.

$$\text{Deviation of x} = x - \mu$$

- The *population variance* of a population data set of N entries is

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

where the symbol σ is a lowercase Greek letter Sigma.

Measures of Variation

- The *population standard deviation* of a population data set of N entries is the square root of the population variance

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

Finding Population Variance and Standard Deviation

1. Find the mean of the population data set.	$\mu = \frac{\sum x}{N}$
2. Find the devation of each entry.	$x - \mu$
3. Square each deviation.	$(x - \mu)^2$
4. Add to get the *sum of squares*	$SS_x = \sum (x - \mu)^2$
5. Divide by N to get the *population variance*.	$\sigma^2 = \frac{\sum (x-\mu)^2}{N}$
6. Find the square root of the variance to get the *population standard deviation*.	$\sigma = \sqrt{\frac{\sum (x-\mu)^2}{N}}$

Measures of Variation

- The *sample variance* and *sample standard deviation* of a sample data set of n entries are

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

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

Measures of Variation Symbols

	Population	Sample
Variance	σ^2	s^2
Standard deviation	σ	s
Mean	μ	$\bar{x}$
Number of entries	N	n
Deviation	$x - \mu$	$x - \bar{x}$
Sum of squares	$\sum(x - \mu)^2$	$\sum(x - \bar{x})^2$

Populations and Parameters

- A *population* is any collection of objects about which information is desired.
- A *parameter* is any summary number, such as an average or percentage, that describes the entire population.
- Problem: we usually cannot know the real value of a population parameter so we need to estimate the parameter.

Samples and Statistics

- A *sample* is a representative group drawn from the population.
- A *statistic* is any summary number that describes the sample.
- We can use the known value of a sample statistic to learn about the unknown value of the population parameter.

Learning About Population Parameters

- There are two ways to learn about a population parameter:
  1. Confidence intervals are used to estimate parameters.
  2. Hypothesis tests are used to draw conclusions about the value of a parameter.

Confidence Intervals

- A *confidence interval* is a range of values that we are confident contains the population parameter.
- General form of most confidence intervals:
  sample estimate $\pm$ margin of error
  or
  lower value $<$ sample estimate $<$ upper value

Example: The t-interval for Population Mean

- The t-interval formula:
  sample mean $\pm$ (t-multiplier $\times$ standard error)
  or
  $\bar{x} \pm t_{a/2,n-1}\left(\frac{s}{\sqrt{n}}\right)$
- Notes:
  - The ***t*-multiplier** depends on the sample size through $n-1$ and the confidence level $(1-\alpha) \times 100$ through $\frac{\alpha}{2}$
  - The **standard error** $(\frac{s}{\sqrt{n}})$ is the estimated standard deviation of all possible sample means
  - The term to the right of the $\pm$ is the margin of error
  - The formula assumes that the data is normally distributed

Confidence Levels

- We typically want to be as confident as possible, so high confidence coefficients $(1-\alpha)$ are used.
- Common confidence coefficients:

$(1-\alpha)$	confidence level	$(1-\frac{\alpha}{2})$	$\frac{\alpha}{2}$
0.9	90%	0.95	0.05
0.95	95%	0.975	0.025
0.99	99%	0.995	0.005

Width of t-interval for μ

- The width of the t-interval is:
  $width = 2 \times t_{a/2,n-1}\left(\frac{s}{\sqrt{n}}\right)$
- Consequences of the formula
  - As the sample mean increases, the width stays the same
  - As the sample standard deviation s decreases, the width decreases
  - As the confidence level decreases, the t-multiplier decreases and so the width decreases
  - As the sample size increases, the width decreases

Hypothesis Testing

- The general idea of hypothesis testing involves:
  1. Make an initial assumption
  2. Collect evidence (data)
  3. Based on the evidence, decide whether to reject or not reject the initial assumption

Hypothesis Testing

- When the initial assumption is made we state two competing hypotheses:
  - H_0 – the null hypothesis (initial assumption)
  - H_A – the alternative hypothesis
- We always assume the null hypothesis is true:
  - Logically speaking, it is impossible to prove a hypothesis, but possible to disprove one.
  - Practically speaking, it is challenging to either prove or disprove a hypothesis beyond the slightest doubt.

Hypothesis Testing Example

- Example: The U.S. criminal justice system assumes "the defendant is innocent until proven guilty."
  1. State the initial assumption
     - H_0 – the defendant is not guilty
     - H_A – the defendant is guilty
  2. Collect evidence – the prosecution collects evidence to make the assumption of innocence refutable
  3. Make a decision based on the evidence
     - If the jury finds sufficient evidence, the null hypothesis is rejected; We believe that the defendant is guilty.
     - If there is insufficient evidence, the null hypothesis is not rejected; we believe the defendant is innocent.

Types of Errors

- Type I error: the null hypothesis is rejected when true
- Type II error: the null hypothesis is not rejected when it is false

	H_0	H_A
Do not reject H_0	Ok	Type II error
Reject H_0	Type I error	Ok

Making a Decision

- We wan to determine whether it was likely that we observe the evidence given the initial assumption
  - If it is *likely* then we do not reject the null hypothesis
  - If is is *unlikely* then we reject the null hypothesis
- Ways to determine whether the evidence is likely or unlikely given the initial assumption
  - Critical value approach
  - P-value approach

The Critical Value Approach

1. Specify the null and alternative hypotheses
2. Calculate the value of the test statistic from the sample data
3. Determine the critical value by finding the value of the known distribution of the test statistic such that the probablility of making a Type I error is small
4. Compare the statistic to the critical value:
    - If the test statistic is more extreme in the direction of the alternative hypothesis than the critical value, then reject the null hypothesis
    - Otherwise, do not reject the null hypothesis

The P-value Approach

1. Specify the null and alternative hypotheses
2. Calculate the value of the test statistic from the sample data
3. Using a known distribution of the test statistic calculate the *p-value*: "if the null hypothesis is true, what is the probability that we would observe a more extreme test statistic in the direction of the alternative hypothesis than we did?"
4. Set the probability of making a Type I error to be small (α) and compare that to the p-value
    - If the p-value is less than or equal to α, reject the null hypothesis
    - Otherwise, do not reject the null hypothesis