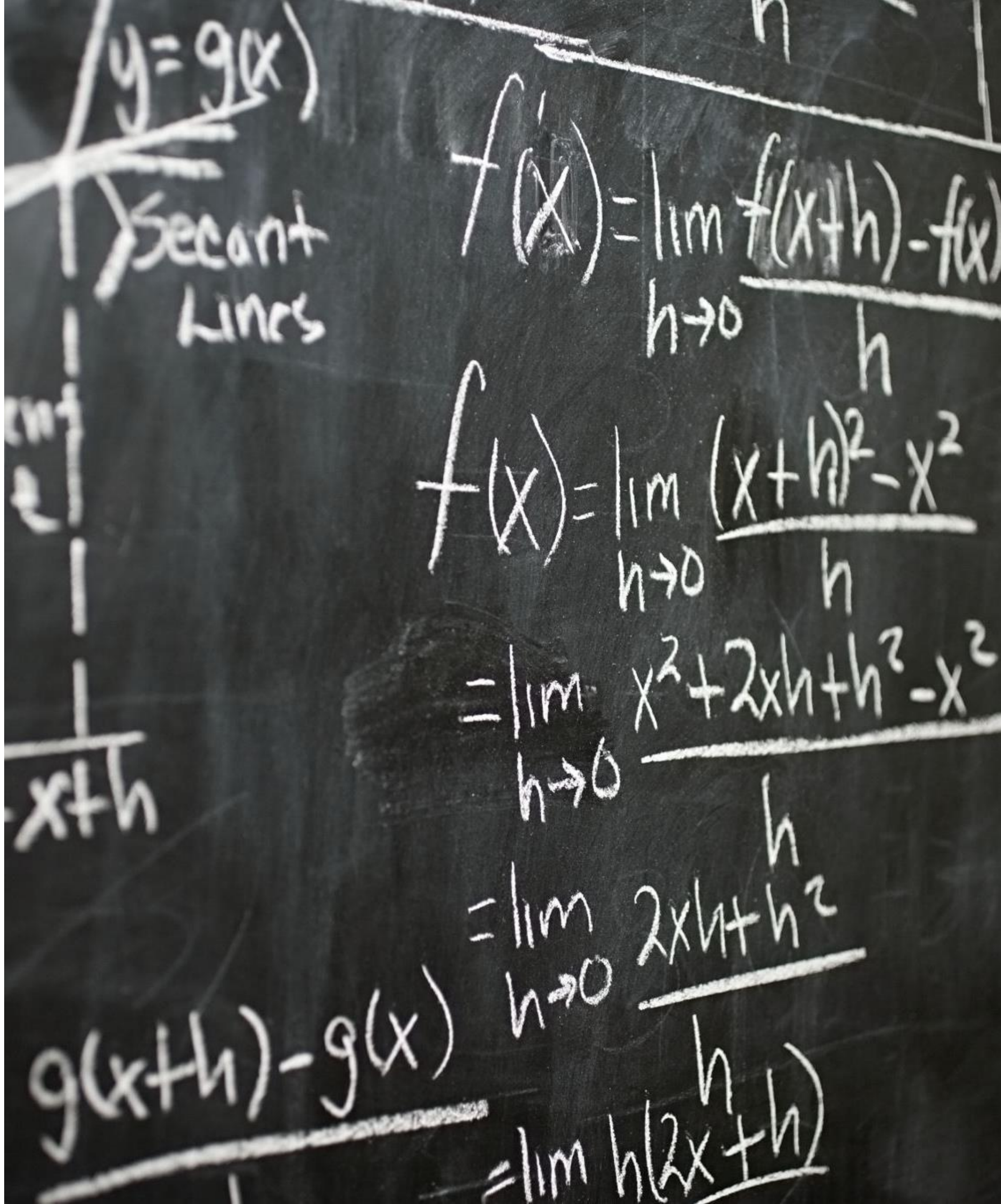
QUANTITATIVE METHODS **STATISTICAL CONCEPTS**

Tomi Heimonen (he/him)

Computing and New Media Technologies
University of Wisconsin-Stevens Point



CONTENT

- Statistical concepts
  - Data types
  - Variable types
- Descriptive statistics
- Fundamentals of statistical inference

BIG PICTURE

- Quantitative research requires **measurement** of something: depending on what is being measured, different **data types** are used.
- Data type influences (in part) the kinds of **statistical methods** can be used to analyze the metrics.
- Measurements are generally a form of **dependent variable** – measured in connection to **independent variables** (e.g., different treatments or conditions).
- Measurements form a **sample** – with specific characteristics, such as minimum and maximum values, mean, and standard deviation, and variance.
- **Descriptive statistics** describe sample characteristics – which can be used to estimate population characteristics (**inferential statistics**).

DATA TYPES

- **Nominal**: data indicates mutually exclusive categories; ordering of the values is not meaningful
  - Examples: gender, name, favorite ice cream flavor
- **Ordinal**: data values can be ordered; values are not equidistant
  - Example: responses to a survey question: “This website was easy to use” – Agree, Neutral, Disagree
- **Interval**: same as ordinal; differences between values are equidistant
  - Example: temperature (difference between 60 and 70 degrees is the same as between 70 and 80 degrees)
- **Ratio**: same as interval, but there is zero point – ratios of measurements are meaningful
  - Example: weight, height (4 pounds is four times as much as 1 pound)

INDEPENDENT AND DEPENDENT VARIABLES

- Independent variables (IV) are systematically varied (or "manipulated") and dependent variables (DV) are the response measures that are collected.
    - Levels of IV, such as different types of ads shown to consumers, are often called "conditions" or "factors"; one or several IV can be manipulated in the same study.
- Many statistical analyses are designed for examining the effect of independent variables on dependent variables.
    - The objective of the statistical analysis is to examine whether the conditions have a statistically significant effect on the observed outcomes.
    - Example: Effect of vaccine type on treatment success
    - In other words, outcome did not occur purely by random chance.

OTHER TYPES OF VARIABLES

- There are also variables we want to **control** (= keep constant) so that they do not affect the outcome.
  - Examples: temperature, lighting conditions, time of day, …
- The study setting, participants, technology etc. can introduce other unintended variables that, if not controlled for, can **confound** the results.
  - Can in different ways affect the relationship between IV and DV.
  - Example: participants' socioeconomic status when studying educational interventions
  - Often tricky to tease out the impact of these variables!
- Some variables we want to **randomize** to reduce bias and improve generalizability, such as assignment of participants to different conditions.

MEASURING VARIABLES

- Variables can have different ranges of possible values within the context of a specific study.
- **Discrete variables**: values can come from a finite set of possible values.
  - Typically nominal, but also ordinal or interval/ratio data can take discrete form.
  - Examples: ethnic background (nominal), course grade (A-F; ordinal), blood pressure level (when the values are grouped into “high”, “normal”, “low”)
- **Continuous variables**: infinite number of possible values, often within some reasonable range.
  - In practice, continuous variables are always interval/ratio type.
  - Example: the time it takes for an athlete to run a 100-meter dash (bounded by 0)
- The types of the statistical analyses that are appropriate depend on the types of independent and dependent variables.*

DESCRIPTIVE STATISTICS

DESCRIPTIVE STATISTICS

- Descriptive statistics characterize the **main features** of the sample.
- Measures of **central tendency** – where does the “center point” of the data lie?
  - **Mode** – most common value
  - **Median** – the middle value
  - **Arithmetic mean** (“average”) – sum of all values divided by number of values
- Measures of **variability** – how much does the data “spread” around the center point?
  - **Range** – distance between smallest and largest value
  - **Variance** – how far the values are spread around the mean
  - **Standard deviation** – conceptually the same as variance, but expressed in units of the original values
- It is often helpful to also represent the data graphically to identify possible outliers or missing data.

TERMINOLOGY: SAMPLE VS. POPULATION

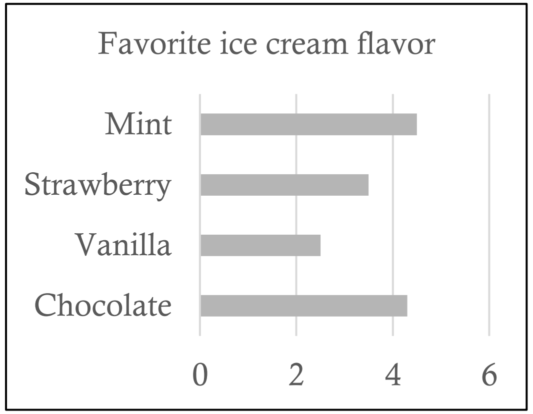
- **Population** = all possible measurements for the given variable
  - Example: current speed of all vehicles on the interstate
- **Sample** = a representative set of measurements from the population
  - Example: current speed of 100 randomly selected vehicles
- A measure of a sample variable is called a **statistic**.
- A measure of a population variable is called a **parameter.**

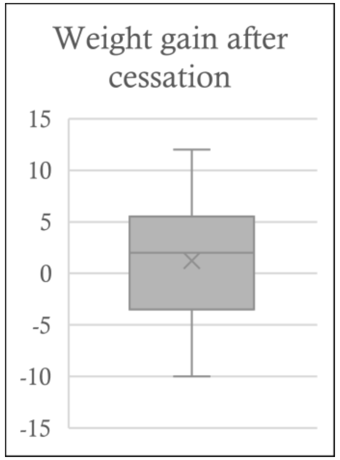
MEASURES OF VARIABILITY FOR POPULATION AND SAMPLES

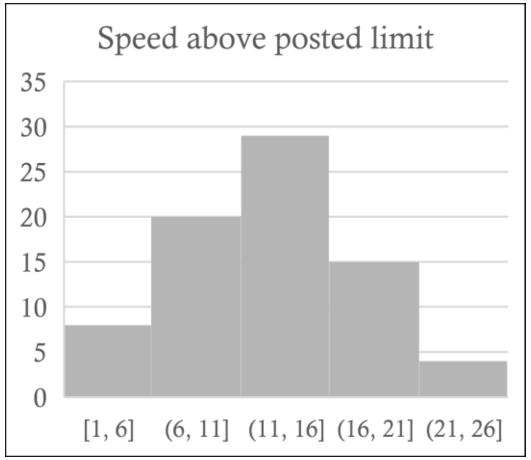
- Range of values in the sample can be used to "quality control" the data.
  - Remove **outliers** – values at extreme ends of the range – and verify **coding** – are all data values correctly inputted?
- Variance measures the spread of the data set:
  - Formula: $S^2 = \frac{\sum_{i=1}^{n}(x_i - \bar{x})^2}{n-1}$
  - x_i= i^{th} value, $\bar{x}$= sample mean, n = number of values
- Standard deviation expresses dispersion in the same units as the mean.
  - Formula: $s = \sqrt{s^2}$
- In a normal distribution (variables that follow a symmetrical distribution around the mean), about 68% of values fall within one standard deviation of the mean and 99.7% within three standard deviations.

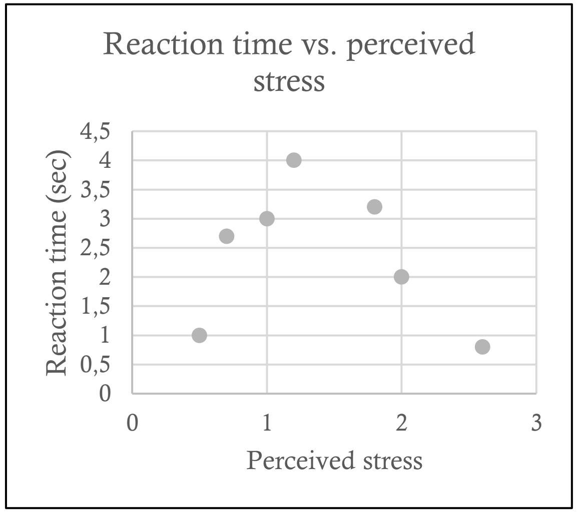
GRAPHING DESCRIPTIVE STATISTICS

- **Bar chart**: series of rectangular bars proportional in size to the values they represent – useful for summarizing categorical data.
- **Histogram**: bar chart that represents the frequency distribution of continuous data
- **Box plot**: graphical summary of the distribution based on min, max, median and 25th and 75th percentiles.
- **Scatter plot**: Visualizes data along two axes – useful when trying to observe any relationships between two variables







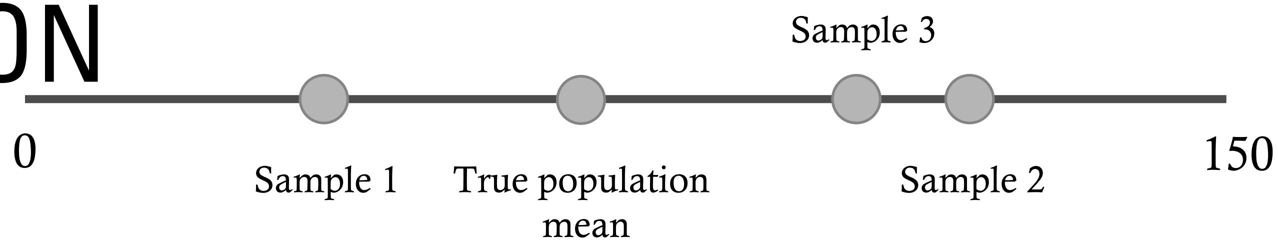


STATISTICAL INFERENCE

STATISTICAL INFERENCE

- Inferential statistics is concerned with **drawing conclusions about a population** based on a sample.
- Why is this relevant? Because researchers rarely have access to the whole population – due to its size, cost or other factors.
- General process for statistical inference:
  - Draw a random sample from the population.
  - Calculate a **sample statistic**.
  - Estimate the **population parameter** (generally one of the measures of central tendency) based on the sample statistic and, if known, population characteristics.
  - Estimates can be **point estimates** (e.g., sample mean) or **interval estimates** (confidence interval).

SAMPLING DISTRIBUTION



- Little bit of theory: We generally make inferences based on the **sampling distribution** of the sample statistic of interest.
- Sampling distribution is a **probability distribution** of the statistic based on a large number of samples of size *n* from the population.
  - Sampling distribution gives the likelihood of obtaining a specific value of the statistic when taking a random sample from the population.
- Sampling distributions are used by statistical tests to calculate probabilities of interest – for example, the likelihood that two separate samples were drawn from the same underlying population.

SAMPLING DISTRIBUTION OF THE MEAN

- Distribution of sample means drawn at random from a population; has the following properties:
  - Its mean is the same as the population mean.
  - Its variance is the population variance divided by the sample size.
  - Its standard deviation is the **standard error of the mean** (σ_M): $\sigma \ / \ \sqrt{n}$
  - It is **normally distributed regardless of the shape of the population distribution** if the sample size is sufficiently large (conventionally, when sample size is larger than 30).
- Why does this matter? Many statistical tests are based on a known sampling distributions – this allows us to estimate the probability of obtaining the sample statistic in order to make inferences.

CONFIDENCE INTERVAL

- Confidence interval is an **interval estimate** for an unknown population parameter.
- Confidence level indicates how certain we are that the **true parameter value** falls within the confidence interval.
  - Traditionally, 95% and 99% confidence levels are used.
- Confidence interval for the mean is calculated based on the sampling distribution of the mean using:
  - sample size
  - standard error of the mean = measures precision of sample mean as an estimate of population mean
  - α = alpha, the probability that true population parameter value lies outside the confidence interval – complement of the confidence level

CONFIDENCE INTERVAL WHEN POPULATION PARAMETERS ARE KNOWN

- Confidence interval formula: $\mathrm{M} \mp \mathrm{Z}_{.95} \times \sigma_{\mathrm{M}}$
  - M is the sample mean
  - $\mathrm{Z}_{.95}$ is the number of standard deviations extending from the mean that contain 95% of values in the standard normal distribution (95% = 1.96 standard deviations in a normal distribution)
  - σ_{M} is the standard error of the mean: $\sigma_{\mathrm{M}} = \frac{\sigma}{\sqrt{n}}$
- Example: Confidence interval for a sample of five numbers (2, 3, 5, 6, 9) from a normal distribution with a standard deviation of 2.5
  - Sample size: 5
  - Sample mean: 5
  - Population standard deviation: 2.5
  - Standard error of the mean: $2.5/\sqrt{5} = 1.118$
  - Confidence interval: $5 \mp (1.96 \times 1.118) = 5 \mp 2.19$

Population mean falls somewhere between 2.81 and 7.19.

CONFIDENCE INTERVAL CALCULATION WHEN POPULATION PARAMETERS ARE UNKNOWN

- A more common scenario in experimental research involving people.
- Approach: Use the **t-distribution** to look up the **critical value** based on the **sample degrees of freedom** (sample size – 1) and desired confidence level.
- Formula: $\mathrm{M} \mp \mathrm{t}_{.95} \times \mathrm{s_M}$
- $\mathrm{s_M}$ = estimate of the standard error of the mean = $\frac{\mathrm{s}}{\sqrt{\mathrm{N}}}$
- Example:
  - Sample mean: 35.5
  - Sample variance: 56.8
  - Sample size: 8
  - Critical value of *t* for *df(7)* = 2.365
  - Estimate of the standard error of the mean: $\sqrt{56.8}/\sqrt{8}$ = 2.66
  - Confidence interval: 35.5 ∓ (2.365 × 2.66) = 35.5 ∓ 6.29
  - The confidence interval for the mean ranges from 29.2 to 41.8

INTERPRETING CONFIDENCE INTERVALS

- Does CI represent that there is 95% probability that the true population mean lies within the confidence interval?
    - Not exactly – it means that if we take multiple samples from the population, the true population mean will fall within the CI in 95% of the time.
- In many cases, CI alone is sufficient to make basic inferences with respect to a hypothesis about the population parameter.
- Example:
    - Baseline fitness score before a training program is 45.
    - After completing a training program, we obtain a 95% CI [48, 54] – because 45 does not fall within the CI, the effect is statistically significant → training program is effective.

FROM CONFIDENCE INTERVAL TO MORE COMPLEX SCENARIOS

- Calculating the CI for a single sample is one of the simplest forms of inference – but what if we wanted to compare two different samples, or study multiple IVs or DVs?
- The generalized approach is to carry out **hypothesis testing**, which requires us to
  - State two competing hypotheses – the **null hypothesis** (H_0; status quo) and an **alternative hypothesis** (H_a; an effect, difference or relationship exists in the population)
  - Select the appropriate statistical tests to test the hypotheses and the **significance level** (α) threshold for rejecting the null hypothesis.
  - Design the experiment, collect data and calculate test statistic and probability (**p-value**) of observing the test statistic or more extreme values under the null hypothesis.
  - Based on test results, determine if we can reject the null hypothesis.
  - Interpret and report the results.