MATH2411: Applied Statistics

Instructor: Dr. Jing YAO, HKUST

Notes by Joyce Shen

1 Descriptive Statistics

The two most important functions of descriptive statistics are **communicate information** and **support reasoning about data**.

1.1 Data

Data are a collection of measurements and identifiers, with a variety of different forms (including numerical, character, etc). Each **datum** is also an observation. A **variable** often refers to a characteristic of interest; variables can be categorical/qualitative, which yield descriptive responses and includes both **nominal** (consisting of distinct, unrankable categories), and **ordinal** (rankable) variables, or quantitative, which are numerical responses that are either **discrete** or **continuous**. Understanding the type of variable allows for better selection of statistical analysis (averaging nominal data is invalid).

1.2 Data Presentation

1.2.1 Numerical

Quantitive data measures **center** (location) and **spread** (variability). Suppose we have n (sorted) data labeled with $x_1, x_2, ...x_n$. The center can be measured with the **sample mean:**

$$\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i = \frac{x_1 + x_2 + ... + x_n}{n} \tag{1}$$

or the **sample median:**

$$\tilde{x} = \begin{cases} x_{\frac{n+1}{2}} & \text{if } n \text{ is odd} \\ \frac{1}{2}(x_{\frac{n}{2}} + x_{\frac{n}{2}+1}) & \text{if } n \text{ is even} \end{cases} \tag{2}$$

The **trimmed mean** is another option, computed by trimming away a certain percent of both the largest and the smallest set of values in sample. The mean is the most sensitive to outliers, followed by the trimmed mean, with the median as the least sensitive. However, the median uses the least information (only one or two data points) compared to the other two.

The spread can be measured by the **sample variance:**

$$s_{n-1}^2 = \frac{1}{n-1} \sum_{i=1}^{n} (x_i - \bar{x})^2 \tag{3}$$

the **sample standard deviation** (square root of the variance), the **sample range** ($x_max - x_min$), or the **interquartile range** (Q3 - Q1).

1.2.2 Tabular and Graphical

The nature of the data and goal of visualization determines what to use. For quantitative data, the following tools are typically used. **Line charts** are good for time-series data, **frequency tables** arrange into numerical and non-overlapping categories, and **histograms** are a graphical version of the frequency tables. **Boxplots** or box-and-whisker plots use the five summary statistics (minimum, lower quartile, median, upper quartile, maximum). **Quartiles** split the ordered data into four equal segments. To find the quartiles, let n be the total number of data, p be 0.25, 0.5, and 0.75 for Q1, Q2, and Q3 respectively. If $np + 0.5$ is an integer, say m, then m is the quartile. If not, then average the two closest numbers to m to find the quartile. Boxplots are commonly used to detect **outliers** (data points either $\leq Q1 - 1.5 * IQR$ or $\geq Q3 + 1.5 * IQR$). **Scatterplots** are used to depict data that comes in pairs and visualizes the relationship between the two variables.

For categorical data, a frequency table can also be used, as well as **pie charts** and **bar charts**.

1.3 Sampling Methods

Sampling correctly is very important; when bias is present, data collected may be inaccurate. **Selection bias** (sample of convenience) makes it more likely to sample certain people over others. **Non-response bias** occurs when people are less likely to respond at certain times, and **voluntary response bias** tends to collect only the most extreme individuals.

Sampling itself can be broken down into **probability sampling**, which involves **random selection**, and allows strong statistical inferences about the population to be made. This includes **simple random sampling**, **systematic sampling**, **stratified sampling** (SRS from groups), and **cluster sampling** (SRS of groups). **Non-probability sampling**, on the other hand, includes **convenience sampling**, **voluntary response sampling**, **purposive sampling**, and **snowball sampling**.

2 Probability

2.1 Sample Space; Events

Probability is a study of randomness. Basic terms include **random experiment** (process which generates a set of data i.e. observations, where the outcomes of the observations cannot be known BEFOREHAND), **sample space** (set of all possible outcomes of a given random experiment, often denoted by S), **sample point** (member of the sample space), and **event** (subset of sample space, usually the outcomes we are interested in, often denoted by A).

There are a variety of event operations, including

- **union**, denoted by $\cup$, where $A \cup B = \{s | s \in A \text{ or } s \in B\}$
- **intersection**, denoted by $\cap$, where $A \cap B = \{s | s \in A \text{ and } s \in B\}$. If two sets have no common element, then their intersection is **disjoint**, and their intersection is the **empty set**, denoted as Φ in this class (and $\{\}$ or $\emptyset$ in other places). If two or more events are **pairwise disjoint**, then all of the events are **mutually exclusive**.
- **complement**, written A^c, $\bar{A}$, or A', is $\{s | s \in S, s \notin A\}$
- **difference**, where $A - B = \{s | s \in A \text{ and } s \notin B\} = A \cap B^c$

There are also properties of operations on events, including the

- **commutative laws**: $A \cup B = B \cup A$ and $A \cap B = B \cap A$, ie order doesn't matter
- **associative laws**: $A \cup (B \cup C) = (A \cup B) \cup C$ and $A \cap (B \cap C) = (A \cap B) \cap C$, ie order of evaluation doesn't matter with the same operation
- **distributive laws**: $A \cup (B \cap C) = (A \cup B) \cap (A \cup C)$ and $A \cap (B \cup C) = (A \cap B) \cup (A \cap C)$
- **DeMorgan's Laws**: $(A \cup B)^c = A^c \cap B^c$ and $(A \cap B)^c = A^c \cup B^c$

2.2 Probability

Now to define probability properly:

Definition 2.1 (Probability)**.** A probability on the sample space S is an assignment of a value, say $P(E)$, to an event E in S via a function $P(.)$ s.t.:

1. $P(E) \geq 0$, for any event E in S (i.e., that it is nonnegative)
2. $P(S) = 1$, (i.e., unit measure)
3. For any sequence of mutually exclusive events $E_1, E_2, ...$ $P(\bigcup\limits_{i=1}^{\infty} E_i) = \sum_{i=1}^{\infty} P(E_i)$ (i.e., countable additivity)

Note that a set is **countable** if it is finite (can be enumerated) or can be described. It is **countably infinite** if it has the same cardinality as the set of natural numbers.

2.2.1 Properties of Probability

Some properties we will cover include:

1. The **probability of the empty set** is 0: $P(\emptyset) = 0$.
2. **Finite additivity**: For any sequence of n mutually exclusive events $E_1, E_2, ..., E_n$, $P(\bigcup\limits_{i=1}^{\infty} E_i) = \sum_{i=1}^{\infty} P(E_i)$. (This is essentially the third point of the definition, except with n events instead of ∞.)
3. **Monotonicity**: $P(A) \leq P(B)$, if $A \subseteq B$
4. The **numeric bound**: $0 \leq P(A) \leq 1$
5. The **complement rule**: $P(A^c) = 1 - P(A)$
6. The **addition law**: $P(A \cup B) = P(A) + P(B) - P(A \cap B)$ (This is the in-and-out rule covered in Discrete)

2.3 Counting Rules

Theorem 2.1. *Consider a sample space $S = s_1, s_2, ...s_N$ with N equally likely outcomes, where N is a finite positive integer. Let E be any event in S. Then*

$$P(E) = \frac{\textit{number of outcomes in E}}{N}$$

To compute probabilities for more complex situations, we must develop systematic ways of counting outcomes.

2.3.1 Addition Principle

If there are m_1 outcomes the 1 way, m_2 outcomes the 2 way, ... m_n outcomes for n total ways, then the total number of outcomes $N = m_1 + m_2 + ... + m_n$.

2.3.2 Multiplication Principle

If there are m_1 outcomes the 1 step, m_2 outcomes the 2 step, ... m_n outcomes for n total steps, then the total number of outcomes $N = m_1 m_2 ... m_n$.

2.3.3 Permutation and Combination

Definition 2.2 (Permutation)**.** A permutation is an ordered arrangement of all or part of a set of objects

Theorem 2.2. *The number of permutations of n objects is $n!$.*

Theorem 2.3. *The number of permutations of n distinct objects taken r at a time is*

$$\frac{n!}{(n-r)!}$$

Theorem 2.4 (Circular Permutation)**.** *The number of permutations of n objects arranged in a circle is $(n-1)!$.*

Definition 2.3 (Combination)**.** The number of ways of selecting r objects from n without regard to order.

Theorem 2.5. *The number of combinations of n distinct objects taken r at a time is*

$$\binom{n}{r} = \frac{n!}{(n-r!)r!}$$

Theorem 2.6 (Partition)**.** *The number of ways that n distinct objects can be grouped into k classes with n_i objects in the i-th class, $i = 1, ...k$ and $\sum_{i=1}^{k} n_i = n$ is*

$$\binom{n}{n_1, n_2, ..., n_k} = \frac{n!}{n_1!n_2!...n_k!}$$

2.4 Conditional Probability

Sometimes the knowledge/information that an event B has occurred influences the probability that A will occur. We can "update" the probability of an event A happening after obtaining some "additional" information from B. **Conditional probability** is the probability of one event given that another event (conditional event) has already occurred (denoted $P(A|B)$).

Definition 2.4 (Conditional Probability)**.** Let A and B be two events such that $P(B) > 0$. Then

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

Note that the **multiplicative rule** $P(A \cap B) = P(A|B)P(B)$ is implied.

2.4.1 Law of Total Probability

Theorem 2.7 (Law of Total Probability - General Version)**.** *If $B_1, B_2, ...B_n$ are **mutually exclusive** and **exhaustive** events (i.e., a **partition** of the sample space S) with probabilities between 0 and 1 exclusively, then for any event A we have*

$$P(A) = \sum_{i=1}^{n} P(A|B_i)P(B_i)$$

2.4.2 Bayes' Theorem

Bayes' theorem is a famous theorem in probability relating the conditional probability $P(A|B)$ to its inverse counterpart $P(B|A)$.

Theorem 2.8 (Bayes' Theorem - First form). *If $P(A) > 0$ and $P(B) > 0$, then we have*

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

Theorem 2.9 (Bayes' Theorem - Second form). *If $B_1, B_2, ...B_n$ are* ***mutually exclusive*** *and* ***exhaustive*** *events (i.e., a* ***partition*** *of S) with probabilities between 0 and 1 exclusively, then for any event $B_j (j = 1, ..., n)$, we have*

$$P(B_j|A) = \frac{P(A|B_j)P(B_j)}{\sum_{i=1}^{n} P(A|B_i)P(B_i)}$$

Note that the unconditional probability $P(B_j)$ is the **prior probability** of B_j, as it does not take any information about A into account. The conditional probability $P(B_j|A)$ is the **posterior probability**, as some additional information (the occurence of A) has been taken into account.

2.5 Independence

Note that it is possible that the knowledge / information that an event B has occurred DOES NOT INFLUENCE the probability that A will occur.

Definition 2.5 (Independence). Two events A and B are independent if

$$P(A|B) = P(A), P(B|A) = P(B)$$

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

Note that events A, B, C are **mutually independent** iff they are pairwise independent AND $P(ABC) = P(A)P(B)P(C)$.

3 Random Variables

3.1 Random Variables

Definition 3.1 (Random Variables). A random variable $X : S \mapsto R$ is a numerical valued function defined on a sample space. That is, a number $X(a)$ is assigned to an outcome a in S

By the notion of r.v., now we have a new interpretation of DATA: data are the actual values (realizations) of the corresponding random variable. Keep in mind that a random variable is a "function" rather than a number; the value of X depends on an outcome - or more rigorously,

$$\{X = x\} = \{a \in S | X(a) = x\}$$

There are two types of random variables:

- **Discrete** r.v., which have a finite or countably infinite range; possible values are isolated numbers. Examples include: number of sales in a store, number of defective items randomly selected from all the items produced

- **Continuous** r.v., whose range is an interval over the real line. Examples include: weight, time until failure of a mechanical product

3.2 Discrete Random Variables

3.2.1 Probability Mass and Cumulative Distribution Functions

Definition 3.2 (Probability mass function)**.** The probability mass function (pmf) of a discrete r.v. X, denoted by $p(x)$, is a function that gives us the probability of occurrence for each possible value x of X. It is valid for all possible values x of X.

Properties of a pmf include

- $0 < p(x) \leq 1$, for all $x \in \chi$, and $p(x) = 0$ for all $x \notin \chi$. (Recall χ is the range of x)

- $\sum_{x \in \chi} p(x) = 1$

Note that when writing/specifying a pmf, you must specify the possible values of x. Alternatively, a cumulative distribution function (cdf) can also be used to specify a random variable.

Definition 3.3 (Cumulative distribution function)**.** The cumulative distribution function (cdf) of a random variable X (discrete or continuous), denoted by $F(x)$, is defined as $F(x) = P(X \leq x)$, for all real values x.

Properties of a cdf include

- $F(x)$ is non decreasing

- $0 \leq F(x) \leq 1$, and $\lim_{x \to -\infty} F(x) = 0, \lim_{x \to \infty} F(x) = 1$

For discrete r.v.,

$$F(a) = P(X \leq a) = \sum_{x \leq a} p(x) \text{ for all real values } a$$

Note that the cdf of a discrete r.v. is a STEP FUNCTION with the size $p(a)$ of the jumps at the possible value a.
The cdf is useful when we want the probability that r.v. X falls in an open and closed interval, as for any $a < b$, we have $P(a < X \leq b) = F(b) - F(a)$

3.2.2 Numerical Measures (Expectation and Variance)

Definition 3.4 (Population mean, Expectation)**.** If X is a discrete r.v. with its pmf $p(x)$ and its range χ, then the population mean or **expectation** of X is defined as

$$E(X) = \sum_{x \in \chi} [xp(x)]$$

This population mean is usually denoted μ or μ_x and is a measure of central location. $E(X)$ is essentially a weighted average of all possible outcomes of X. Remember that this is the *population* mean, which is different from the *sample* mean (which is collected from the data).

Suppose X, Y are random variables and a and b are two constants. Then properties of the expectation includes

- $E(b) = b$
- $E(aX) = aE(X)$
- $E(aX + b) = aE(X) + b$
- $E(X + Y) = E(X) + E(Y)$ (Linearity of Expectation)
- $E[g(X)] = \sum_{x \in \chi}[g(x)p(x)]$ where g is a real-valued function

Definition 3.5 (Population variance)**.** If X is a discrete r.v. with its pmf $p(x)$ and its range χ, then the population variance of X is derived as

$$Var(X) = \sum_{x \in \chi} [(x - \mu)^2 p(x)]$$

The population variance is usually denoted σ^2 or σ_x^2 and is a measure of spread or dispersion. Remember that this is the *poplulation* variance, which differs from the *sample* variance. The positive square root of the variance (σ) is called the **population standard deviation** of X.

Suppose X, Y are random variables and a and b are two constants. Then properties of the expectation includes

- $Var(b) = 0$
- $Var(aX) = a^2 Var(X)$
- $Var(aX + b) = a^2 Var(X)$
- If X and Y are independent, then $Var(X + Y) = Var(X) + Var(Y)$
- Let $g(X) = (X - \mu)^2$. Then $Var(X) = E[g(X)] = E[(X - \mu)^2]$
- $Var(X) = E(X^2) - [E(X)]^2$ (i.e., $\sigma^2 = E(X^2) - \mu^2$)

3.3 Continuous Random Variables

3.3.1 Probability Density Function

To specify a continuous random variable, we can use the cumulative distribution function (cdf) or the probability *density* function (pdf). For the **probability density function** $f(x)$,

- $f(x) > 0$ for all $x \in \chi$ and $f(x) = 0$ for all $x \notin \chi$
- $\int_{-\infty}^{\infty} f(x)dx = 1$

Note that the value of the pdf $f(x)$ does NOT give the probability that the corresponding random variable takes on the value x. In general, if X is a continuous r.v., then $P(X = k) = 0$ for any k. To convert between the pdf and cdf, take the derivative and integral.

- **pdf to cdf**: $F(a) = P(X \leq a) = \int_{-\infty}^{a} f(x)dx$
- **cdf to pdf**: $f(a) = \frac{\mathrm{d}}{\mathrm{d}x}F(x)_{x=a}$

Theorem 3.1. *For any $a < b$,*

$$P(a < X \leq b) = F(b) - F(a) = \int_a^b f(x)dx$$

3.3.2 Expectation and Variance

Definition 3.6 (Population mean, Expectation)**.** If X is a continuous r.v. with its pdf $f(x)$, then the population mean or **expectation** of X is defined as

$$E(X) = \int_{-\infty}^{\infty} [xf(x)]dx$$

Definition 3.7 (Population variance)**.** If X is a continuous r.v. with its pdf $f(x)$, then the population variance of X is derived as

$$Var(X) = \int_{-\infty}^{\infty} [(x-\mu)^2 f(x)]dx$$

The properties of the expectation and variance are the same as for discrete random variables.

3.3.3 Chebyshev's Theorem

Theorem 3.2 (Chebyshev's Theorem)**.** *Let X be a random variable with mean μ and variance σ^2. Then for any $t > 0$,*

$$P(|X - \mu| \geq t) \leq \frac{\sigma^2}{t^2}$$

Let $t = k\sigma$, then $P(|X - \sigma| \geq k\sigma) \leq \frac{1}{k^2}$.

3.4 (Some) Probability Distributions

3.4.1 Binomial Distribution - Discrete

To understand a binomial distribution, we need to first look at a **Bernoulli random variable**: a variable that takes on only two values: 0 and 1, with probabilities $1-p$ and p, respectively. Its pmf is thus

$$p(x) = \begin{cases} p^x(1-p)^{1-x} & \text{if } x \text{ is 0 or 1} \\ 0 & \text{if } x \text{ is anything else} \end{cases}$$

A **binomial distribution** can be seen as a series of Bernoulli random variables (if $n = 1$, then X is a Bernoulli r.v.).

Definition 3.8 (Binomial Distribution)**.** If X is the discrete r.v. of the *number of successes in n trials*, then we can use a **binomial distribution** to characterize its random behavior. With notation: $X \sim Binomial(n, p)$ where n is the number of trials and p is the probability of success, a binomial distribution has pmf:

$$p(x) = \begin{cases} \binom{n}{x} p^x(1-p)^{n-x} & \text{for } x = \text{1,2,...,n} \\ 0 & \text{if } x \text{ is anything else} \end{cases}$$

A binomial distribution has expectation $E(X) = np$ and variance $Var(X) = np(1-p)$.

The conditions for a binomial distribution include:

- Fixed **finite** number of identical trials (i.e., $n < \infty$)
- Trials are **independent**
- Trials can either be **successful** or **failure**
- Probability of success p is **constant** across trials

It's important to note whether or not a situation follows a binomial distribution before using its pmf. For example, pay attention to selection with(out) replacement and violating the independent trial condition.

3.4.2 Poisson Distribution - Discrete

Definition 3.9 (Poisson Distribution)**.** A **poisson distribution** can be used to determine the probability of counts of the occurence of an event over time (or space). Let X be the *number of occurences of an event over a unit time.* With notation: $X \sim Poisson(\lambda)$ where $\lambda \in (0, \infty)$ is the *rate or average number of occurences* of an event *per unit time*, a poisson distribution has pmf:

$$p(x) = \begin{cases} \frac{\lambda^x e^{-\lambda}}{x!} & \text{for } x = \text{1,2,...} \\ 0 & \text{if } x \text{ is anything else} \end{cases}$$

A poisson distribution has expectation $E(X) = \lambda$ and variance $Var(X) = \lambda$.

The conditions for a poisson distribution include:

- Trials are **independent**
- Average rate (events per unit time) λ is **constant**

Note that when we study the count of occurences of an event over a period of t units of time with the rate λ still representing the number of occurences per *unit* time, use $Y_t \sim Poisson(\lambda t)$.

Theorem 3.3 (Poisson Limit Theorem)**.** *The binomial distribution tends towards the Poisson distribution as $n \to \infty, p \to 0$ and np stays constant. Let p_n be a sequence of real numbers in $[0,1]$ such that the sequence np_n converges to a final limit λ. Then*

$$\lim_{n\to\infty} \binom{n}{k} p_n^k (1-p_n)^{n-k} = \frac{\lambda^k e^{-\lambda}}{k!}$$

I.e., the Poisson distribution with $\lambda = np$ closely approximates the binomial distribution if n is large and p is small.

3.4.3 Normal Distribution - Continuous

The **normal distribution**, also called the **Gaussian distribution**, is the most important continuous probability distribution because of its ubiquity in the real world.

Definition 3.10 (Normal Distribution)**.** A **normal distribution** has notation: $X \sim N(\mu, \sigma^2)$ where $\mu \in (-\infty, \infty)$ is the *mean* and $\sigma^2 \in (0, \infty)$ is the **variance**. A normal distribution has pdf:

$$f(x) = \frac{1}{\sqrt{2\pi b}} e^{-\frac{1}{2b}(x-a)^2} \text{ for all real values x}$$

A normal distribution has expectation $E(X) = \mu$ and variance $Var(X) = \sigma^2$.

Note that all normal distributions have a bell-shaped density curve regardless of the values of μ and σ.

Definition 3.11 (Standard Normal Distribution)**.** The **standard normal distribution** has mean 0 and variance 1, i.e., $N(0,1)$. The random variable following the standard normal distribution is often denoted by Z in probability and statistics. Its probability density function (pdf) is

$$\varphi(z) = \frac{1}{\sqrt{2\pi}} e^{-\frac{z^2}{2}}$$

and its distribution function (cdf) is

$$\Phi(z) = \int_{-\infty}^{z} f(t)dt = \int_{-\infty}^{z} \frac{1}{\sqrt{2\pi}} e^{-\frac{t^2}{2}} dt$$

Standardization n is a process used to transform any normal-distributed r.v., say $X \sim N(\mu, \sigma^2)$ to a standard normal-distributed r.v through

$$Z = \frac{X - \mu}{\sigma} \sim N(0,1)$$

3.5 Moment Generating Function (MGF)

Definition 3.12 (Population Moments)**.** The **r-th (population) moment about the origin** of (the distribution of) X is defined by $E(X^r)$, for $r = 1, 2, ...$ if it exists. This moment can also be seen as $E[(x-0)^2]$. The **r-th (population) central moment** is defined by $E[X - E(X)]^r$, for $r = 1, 2, ...$ if it exists.

The population mean $\mu = E(x)$ is a measure of location of a distribution of a r.v. and the population variance $\sigma^2 = Var(X) = E[(X - \mu)]^2 = E(X^2) - [E(X)]^2$ is a measure of spread of a distribution r.v.

3.5.1 Skewness - 3rd population moment

A variant of the 3rd population moment, called **skewness**: the 3rd population moment of a standardized r.v.:

$$E[(\frac{X - \mu}{\sigma})^3] = \frac{\mu_3}{\sigma^3}$$

Skewness is a *measure of the symmetry of the distribution*:

- Skewness $= 0$ if a distribution is **symmetric** about its **mean** or **median**
- Skewness is **positive** if it has a tail to the **right**
- Skewness is **negative** if it has a tail to the **left**

3.5.2 Kurtosis - 4th population moment

A variant of the 4th population moment, called **kurtosis**: the 4th population moment of a standardized r.v.:

$$E[(\frac{X - \mu}{\sigma})^4]$$

3.5.3 Moment Generating Function (MGF)

Definition 3.13 (Moment generating function)**.** The **moment generating function** of X is defined as:

$$M_x(t) = E(e^{tx}) \begin{cases} \sum_x e^{tx} p_X(x) & \text{if } X \text{ is discrete} \\ \int_{-\infty}^{\infty} e^{tx} f_X(x) dx & \text{if } X \text{ is continuous} \end{cases}$$

...if the expectation exists and is finite for all real t around 0.

Note that

- $M_x(t)$ is a function of t
- Since the exponential function decreases rapidly, the mgf does not always exist (e.g., the Cauchy Distribution)
- When the mgf does exist, we can use it to determine population moments very easily
- (In this course, unless specified...) we assume all mgfs exist and are finite for all real t around 0.

Note that with the MGF, if we evaluate the k-th derivative of $M_X(t)$ at $t = 0$, aka $M_X(0)$, then we would have

$$M_X^k(0) = \frac{\mathrm{d}^k}{\mathrm{d}t^k} M_X(t)_{t=0} = E(X^k) \quad \text{for } k = 1, 2, ...$$

Which demonstrates the relationship between MGF and the k-th moment about the origin.

4 Parameter Estimation

4.1 Basic Concepts

From chapter 3, recall that if the distribution of X is known (CDF, PMF, PDF, etc), then we can answer any *probability* problem of X. However, if the distribution is unknown (either an unknown type or unknown parameters), then we need to use **statistics** to **make an inference** about the distrbution or parameter of interest.

Definition 4.1 (Inferential statistics)**.** Statistical procedures that use data from an observed sample to make a conclusion about a population.

Some key terms in statistics include:

- ***unknown* population**: an unknown distribution of the r.v. X
- **sample**: a collection of data of X
- **parameter**: e.g. μ_x and σ_x^2 of a normal distribution
- **statistic**: a quantity calculated from a sample; e.g. $\bar{x}, s_{n-1}^2, s_n^2$

When the distribution of X is unknown, population parameters such as $E(X)$ (i.e., μ_x) and $Var(X)$ (i.e., σ_x^2) are unknown; if we are interested in their true values, then we use collected data to 'guess' their values *statistically.* For example, we often use $\bar{x}$ to guess μ_x.

Definition 4.2 (Statistic)**.** Any function of the random variables constituting a random sample is called a **statistic** if *it does not depend on the unknown parameter(s).*

Recall that the unknown parameter (e.g., μ_x) is *fixed* – but the statistic (e.g., $\bar{x}$) may be different for each sample. If we have B samples, then we would have B sample means $(\bar{x_1}, \bar{x_2}, ...\bar{x_B})$, which behaves very similarly to the n data $(x_1, x_2, ..., x_n)$ of the random variable X. Then, we can imagine a r.v. whose actual values are the B sample means $\bar{x_1}, \bar{x_2}, ...\bar{x_B}$ – which we can denote $\bar{X}$.

4.1.1 Estimator (n.) vs Estimate (n.)

Essentially, an **estimator** is a random variable (e.g. $\bar{X}$) used to study the theoretical property of the estimation method for the unknown parameter while an **estimate** is a number (e.g. $\bar{x}$) used to provide a numerical value of the unknown parameter according to a *particular* sample (changes with the sample).

Features of a statistic: *before* observing, $T(X_1, X_2, ..., X_n)$ is a random variable, aka an **estimator (methodology)** – *after* observing, $T(x_1, x_2, ..., x_n)$ is a number, aka an **estimate (number)**.

4.1.2 Sample Moment vs Population Moment

The $k-th$ **sample moment** about the origin is

$$\bar{X^k} = \frac{1}{n}\sum_{i=1}^{n} X_i^k \ (k = 1, 2, ...)$$

The $k-th$ **population moment** about the origin is

$$E(X^k) = \begin{cases} \sum_x [x^k p_X(x)] & \text{if } X \text{ is discrete} \\ \int_{-\infty}^{\infty} [x^k f_X(x)] dx & \text{if } X \text{ is continuous} \end{cases}$$

Note that with **sample** moments, $X_1, X_2, ..., X_n$ are **independent and identically distributed (i.i.d.)** with the same distribution as the population X, so $X_1^k, X_2^k, ..., X_n^k$ are also i.i.d. with the same distribution as X^k.

4.1.3 Numerical Description of the Sample Mean + Sample Variance

Theorem 4.1. *Assume that the mean and variance of the population X exist: $E(X) = \mu$, $Var(X) = \sigma^2$. $X_1, X_2, ..., X_n$ are a random sample from population X. Then*

$$E(\bar{X}) = \mu, Var(\bar{X}) = \frac{\sigma^2}{n}, E(S_{n-1}^2) = \sigma^2$$

where $\bar{X} = \frac{1}{n}\sum_{i=1}^{n} X_i$ is the sample mean and $S_{n-1}^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \bar{X})^2$ is the sample variance.

4.2 Point Estimation

4.2.1 Basics of Point Estimation

Motivation: If the parameter of interest is a population mean (or variance), then we can use sample mean (or variance) to estimate it.
Let the pdf (or pmf) of a r.v. X be $f(x;\theta)$ with an unknown parameter vector $\theta = (\theta_1, \theta_2, ..., \theta_m)^T \in \Theta \subseteq \mathbb{R}^m$ where Θ denotes the corresponding **parameter space** (range of values θ can take on), and $m \geq 1$ represents the *number of unknown parameters* to be estimated. Therefore, we have a family of densities: $\{f(x;\theta) : \theta \in \Theta\}$, from which we need to select one member as the pdf of X. This is equivalent to estimating the parameter vector θ.

We take a random sample $X_1, X_2, ..., X_n$ from a population with a pdf/pmf $f(x;\theta)$ where n is the sample size. If a *statistic* $Y = T(X_1, X_2, ..., X_n)$ is used to estimate the parameter θ, then the statistic is a **point estimator** of θ, where Y is a **random variable**. If the *observations* of $X_1, X_2, ..., X_n$ are $x_1, x_2, ..., x_n$, then $y = T(x_1, x_2, ..., x_n)$ is called a **point estimate** of θ, where y is a **real number**.
A few remarks about point estimation:

- parameter of interest (**estimand**) can be a *function* of the unknown distribution parameter(s) θ (i.e., $h(\theta)$)

4.2.2 Method of Moments Estimation (MME)

Definition 4.3 (Method of Moments Estimation)**.** Suppose that there are m unknown parameters $\theta_1, \theta_2, ..., \theta_m$. If we can rewrite them in the form of m or more moments, i.e.,

$$\begin{cases} \theta_1 = g_1(E(X), E(X^2), ..., E(X^m), ...) \\ \theta_2 = g_2(E(X), E(X^2), ..., E(X^m), ...) \\ ... \\ \theta_m = g_m(E(X), E(X^2), ..., E(X^m), ...) \end{cases}$$

then the **method of moments estimator (MME)**, denoted $(\hat{\theta_1}, \hat{\theta_2}, ..., \hat{\theta_m})$ of $(\theta_1, \theta_2, ..., \theta_m)$ is

$$\begin{cases} \hat{\theta_1} = g_1(\bar{X}, \bar{X^2}, ..., \bar{X^m}, ...) \\ \hat{\theta_2} = g_2(\bar{X}, \bar{X^2}, ..., \bar{X^m}, ...) \\ ... \\ \hat{\theta_m} = g_m(\bar{X}, \bar{X^2}, ..., \bar{X^m}, ...) \end{cases}$$

where

$$\bar{X^k} = \frac{1}{n}\sum_{i=1}^{n}(X_i)^k, \text{ for } k = 1, 2, ..., m, ...$$

Creds to Karl Pearson in the late 1800s, the method of moments estimation is a classical **point-valued estimation approach** in statistics. In general, if there are m unknown parameters to be estimated, the *first* m or more population moments ($E(X^k)$ for $k = 1, 2, ..., m$) are required.

For an equivalent method to get the MME, we can first **equate** the **sample moments** to the corresponding **population moments**, then solve the system of equations to obtain moment estimators of parameters.

Since $E(\bar{X^k}) = E(X^k)$, when n is large, $\bar{X^k} \approx E(X^k)$. By equating the sample moments to the corresponding population moments, you can then solve the system of equations to get the moment estimator of parameters.

The procedure of constructing a method of moments estimator:

1. Calculate low order moments and find expressions for the moments in terms of the parameters
2. Invert the expression found above and find a new expressions for the parameters in terms of the moments
3. Insert the sample moments into the expressions to get the parameters in terms of the sample moments

Note that the method of moments estimation ...

- is quick and easy – it uses only population moments (and relies heavily on the existence of them)
- may not be unique
- is often biased
- works well only when the sample size is sufficiently large
- if $\hat{\theta}_i$ is the MME for θ_i for $i = 1, ..., m$ then $h(\hat{\theta_1}, \hat{\theta_2}, ..., \widehat{\theta_m})$ is the MME for $h(\theta_1, \theta_2, ..., \theta_m)$ where h is a known function (invariance property)

4.2.3 Properties of Point Estimation

Point estimators are **not unique** – we need a way to qualify if an estimate is a ’good’ estimate i.e., how do we measure the ”closeness” of an estimator to the unknown parameter?

- **Unbiased vs Biased**: an estimator is **unbiased** if the expectation of the estimator $\hat{\theta} = \hat{\theta}(X_1, X_2, ..., X_n)$ exists, and $\forall \theta \in \Theta : E(\hat{\theta}) = \theta$. Otherwise, it is **biased**; the **bias** of the estimator $\hat{\theta}$ is denoted $b_n(\hat{\theta}) = E(\hat{\theta}) - \theta$

  Biased Estimator: if $b_n(\hat{\theta}) \neq 0$

  Asymptotic Unbiased Estimator: if $b_n(\hat{\theta}) \neq 0$ but $\lim_{n \to -\infty} b_n(\hat{\theta}) = 0$

- **Efficiency**: efficiency allows us to compare two unbiased estimators for a common unknown parameter – intuitively, the estimator with a smaller variance is better;

  Definition 4.4 (Efficiency)**.** Let $X_1, X_2, ..., X_n$ be a random sample from the population $X \sim f(x;\theta), \theta \in \Theta$ and $\hat{\theta_1} = \hat{\theta_1}(X_1, X_2, ..., X_n)$, and $\hat{\theta_2} = \hat{\theta_2}(X_1, X_2, ..., X_n)$ are two **unbiased** estimators of θ, i.e.,

  $$E(\hat{\theta_1}) = E(\hat{\theta_2}) = \theta(\forall\theta \in \Theta)$$

  If $\forall\theta \in \Theta$ and we have $Var(\hat{\theta_1}) \leq Var(\hat{\theta_2})$, then $\hat{\theta_1}$ is **more efficient** than $\hat{\theta_2}$ – thus, variance can be used to compare different *unbiased* estimators.

  Note that **accuracy** (how close the estimator is to the unknown parameter) and **precision** (how close the estimates are to each other) are different.

4.3 Interval Estimation

Point estimators, regardless of methodology, are unable to provide the *precision* nor *reliatbility* of estimators – therefore, sometimes a *range* of the unknown parameter is more useful (i.e., interval estimation).

Definition 4.5 (Interval Estimation)**.** Assume that the population $X \sim f(x;\theta)$ $(\theta \in \Theta)$, for any $\alpha \in (0,1)$, if there exists two statistics

$$T_1 = T_1(X_1, X_2, ..., X_n), \quad T_2 = T_2(X_1, X_2, ..., X_n), \quad (T_1 \leq T_2)$$

such that $\forall\theta \in \Theta$,

$$P(T_1 \leq \theta \leq T_2) \geq 1 - \alpha$$

Then the **random interval** $[T_1, T_2]$ is the interval desired. A value $[t_1, t_2]$ of the random interval $[T_1, T_2]$ is also called a $100(1-\alpha)\%$ **confidence interval** for θ.
$1 - \alpha$ is the **confidence level** associated with the confidence interval, with T_1, T_2 as the **lower/upper bounds** respectively.

Note that:

- α is usualy small (e.g., 0.01, 0.05, 0.10), s.t. $1 - \alpha$ is as close to 1 as possible (e.g, 0.99, 0.05, 0.90)
- For a **continuous** population, the interval is chosen s.t. $P(T_1 \leq \theta \leq T_2) = 1 - \alpha$
- For a **discrete** population, the interval is chosen s.t. $P(T_1 \leq \theta \leq T_2)$ is no less than $1 - \alpha$ and as close to $1 - \alpha$ as possible
- If $T_2 - T_1$ is *small*, then the precision is HIGH, but reliability is LOW
- If $T_2 - T_1$ is *large*, then the precision is LOW, but reliability is HIGH

4.3.1 Sampling Distributions

Definition 4.6 (χ^2-Distribution)**.** Let $X_1, X_2, ..., X_n$ be a random sample (i.e., iid) from the population $X \sim N(0,1)$, and

$$Y = X_1^2 + X_2^2 + ... + X_n^2$$

then Y follows the χ^2-distribution with n **degrees of freedom** (number of variables that can change freely), denoted $Y \sim \chi^2(n)$ with pdf:

$$f(y) = \begin{cases} \frac{1}{2^{n/2}\Gamma(n/2)} y^{\frac{n}{2}-1} e^{\frac{-y}{2}} & \text{if } y > 0 \\ 0 & \text{if } y \leq 0 \end{cases}$$

Note that if $Y \sim \chi^2(n)$, then $E(Y) = n, Var(Y) = 2n$. The χ^2-distribution is also **additive**, meaning if $V \sim \chi^2(n_1), W \sim \chi^2(n_2)$ and V and W are independent, then $V + W \sim \chi^2(n_1 + n_2)$.

Definition 4.7 (t-Distribution)**.** If $Z \sim N(0,1), Y \sim \chi^2(n)$ and Z and Y are independent, let

$$W = \frac{Z}{\sqrt{Y/n}}$$

then W follows the t-distribution with n degrees of freedom, denoted as $W \sim t(n)$ with pdf

$$f(w) = \frac{\Gamma[(n+1)/2]}{\sqrt{n\pi}\Gamma(n/2)}(1 + \frac{w^2}{n})^{-(n+1)/2} \quad \text{for } -\infty < w < \infty$$

Properties of this pdf includes

- $f(w) = f(-w)$ (i.e., symmetric about the y-axis)
- $f'(w) > 0$ for $w < 0$ and $f'(w) < 0$ for $w > 0$
- $\lim\limits_{w \to -\infty} f(w) = 0, \lim\limits_{w \to \infty} f(w) = 0$
- As the degrees of freedom increase, the t-distribution approaches the normal density curve

Theorem 4.2. *Let $X_1, X_2, ..., X_n$ be a random sample from the population $X \sim N(\mu_x, \sigma_x^2)$, then*

$$\bar{X} \sim N(\mu_x, \frac{\sigma_x^2}{n})$$

Theorem 4.3. *Let $X_1, X_2, ..., X_n$ be a random sample from the population $X \sim N(\mu_x, \sigma_x^2)$, then*

$$\frac{(n-1)S_{n-1}^2}{\sigma_x^2} \sim \chi^2(n-1)$$

and $\bar{X} = \frac{1}{n}\sum_{i=1}^n X_i$ and $S_{n-1}^2 = \frac{1}{n-1}\sum_{i=1}^n (X_i - \bar{X})^2$ are independent.

Theorem 4.4. *Let $X_1, X_2, ..., X_n$ be a random sample from the population $X \sim N(\mu_x, \sigma_x^2)$, then*

$$\frac{\bar{X} - \mu_x}{S_{n-1}/\sqrt{n}} \sim t(n-1)$$

4.3.2 Interval Estimation for the Population Mean (μ_x)

The key to constructing a **random interval** is to find a **pivotal quantity (pivot)** and its distribution (e.g., $\frac{\bar{X}-\mu_x}{\sigma_x/\sqrt{n}} \sim N(0,1)$).

When estimating the population mean, the population variance (σ_x^2) can either be *known* or *unknown.*
For a **known variance** (σ_x^2):

- **Random Interval:**

$$[\bar{X} - z_{\alpha/2} * \frac{\sigma_x}{\sqrt{n}}, \bar{X} + z_{\alpha/2} * \frac{\sigma_x}{\sqrt{n}}]$$

- **Confidence Interval**: the *100(1-α)%* confidence interval for unknown μ_x is given by

$$[\bar{x} - z_{\alpha/2} * \frac{\sigma_x}{\sqrt{n}}, \bar{x} + z_{\alpha/2} * \frac{\sigma_x}{\sqrt{n}}]$$

  where $z_{\alpha/2}$ is defined as $P(Z > z_{\alpha/2}) = \frac{\alpha}{2}$, or equivalently, $P(Z \leq z_{\alpha/2}) = 1 - \frac{\alpha}{2}$.

For an **unknown variance** (σ_x^2):

- **Random Interval:**

$$[\bar{X} - t_{(n-1),\alpha/2}\frac{S_n - 1}{\sqrt{n}}, \bar{X} + t_{(n-1),\alpha/2}\frac{S_n - 1}{\sqrt{n}}]$$

- **Confidence Interval**: the *100(1-α)%* confidence interval for unknown μ_x is given by

$$[\bar{x} - t_{(n-1),\alpha/2}\frac{S_n - 1}{\sqrt{n}}, \bar{x} + t_{(n-1),\alpha/2}\frac{S_n - 1}{\sqrt{n}}]$$

  where $t_{(n-1),\alpha/2}$ describes $n-1$ degrees of freedom and probability $\alpha/2$.

4.3.3 Interval Estimation for the Population Variance (σ_x^2)

For an **unknown mean** (μ_x):

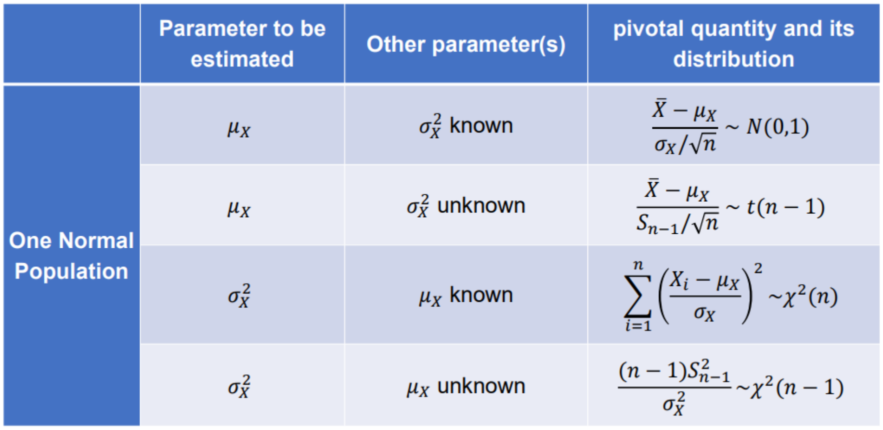
- **Random Interval:**

$$[\frac{(n-1)S_{n-1}^2}{\chi_{n-1,\alpha/2}^2}, \frac{(n-1)S_{n-1}^2}{\chi_{n-1,1-\alpha/2}^2}]$$

- **Confidence Interval**:

$$[\frac{(n-1)s_{n-1}^2}{\chi_{n-1,\alpha/2}^2}, \frac{(n-1)s_{n-1}^2}{\chi_{n-1,1-\alpha/2}^2}]$$

4.3.4 Summary



	Parameter to be estimated	Other parameter(s)	pivotal quantity and its distribution
One Normal Population	μ_X	σ_X^2 known	$\dfrac{\bar{X}-\mu_X}{\sigma_X/\sqrt{n}} \sim N(0{,}1)$
	μ_X	σ_X^2 unknown	$\dfrac{\bar{X}-\mu_X}{S_{n-1}/\sqrt{n}} \sim t(n-1)$
	σ_X^2	μ_X known	$\sum_{i=1}^{n}\left(\dfrac{X_i-\mu_X}{\sigma_X}\right)^2 \sim\chi^2(n)$
	σ_X^2	μ_X unknown	$\dfrac{(n-1)S_{n-1}^2}{\sigma_X^2} \sim\chi^2(n-1)$

Theorem 4.5 (Central Limit Theorem (CLT))**.** *If $\bar{X}$ is the mean of a random sample of size n taken from any population with mean μ_x and finite variance σ_x^2, then the limiting form of the distribution of*

$$Z = \frac{\sum_{i=1}^{n} X_i - n\mu}{\sqrt{n}\sigma_x} = \frac{\bar{X}-\mu_x}{\sigma_x/\sqrt{n}}$$

as $n \to \infty$, is the standard normal distribution $N(0,1)$.

5 Hypothesis Testing

5.1 Concepts of Hypothesis Testing

Unlike the previous chapter where the goal was to use a sample of data to get a point/interval valued estimate of an unknown parameter, hypothesis testing looks at a *hypothesized value of the parameter* to be tested, and then uses sample data to see if the assumption should be rejected or not.

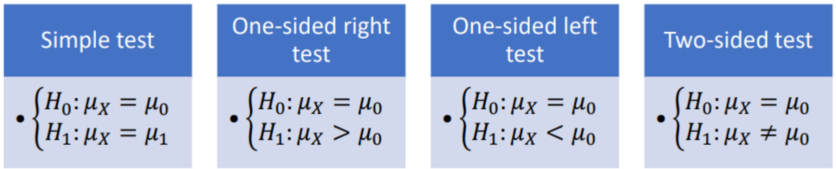
5.1.1 Hypotheses

Definition 5.1 (Statistical Hypothesis)**.** A **statistical hypothesis** is an assertion or conjecture or statement about the population, usually formulated in terms of population parameters.

Definition 5.2 (Null Hypothesis H_0)**.** The **null hypothesis** H_0 is *assumed* to be true and then tested to be rejected or not to be rejected formally. This always contains the $=$ sign (i.e., $=, \leq, \geq$)

Definition 5.3 (Alternative Hypothesis H_1)**.** The **alternative hypothesis** H_1 contains the values of the parameter we would accept if we reject H_0. This never contains the $=$ sign, except in a simple test.

According to the form of the alternative hypothesis, we can have four types of tests:



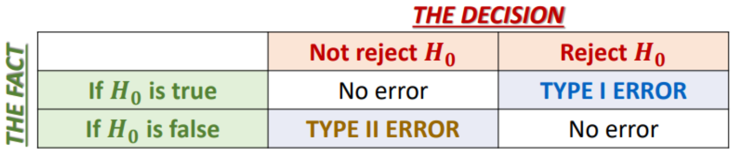
Simple test	One-sided right test	One-sided left test	Two-sided test
• $\begin{cases} H_0: \mu_X = \mu_0 \\ H_1: \mu_X = \mu_1 \end{cases}$	• $\begin{cases} H_0: \mu_X = \mu_0 \\ H_1: \mu_X > \mu_0 \end{cases}$	• $\begin{cases} H_0: \mu_X = \mu_0 \\ H_1: \mu_X < \mu_0 \end{cases}$	• $\begin{cases} H_0: \mu_X = \mu_0 \\ H_1: \mu_X \neq \mu_0 \end{cases}$

Note that in hypothesis testing, we can only ever **reject the null hypothesis** (H_0), never accept it. Instead, we can only ever **fail to reject** H_0.

5.1.2 Errors (Type I and Type II)

With hypothesis testing, there are two possible kinds of errors:

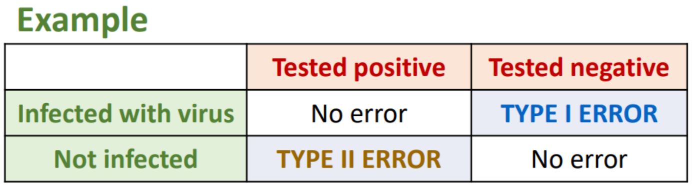
- **Type I**: rejected H_0 when it is TRUE, $P(\text{Type I Error}) = \alpha$
- **Type II**: failed to reject H_0 when it is FALSE, $P(\text{Type II Error}) = \beta$



THE DECISION

THE FACT

	Not reject H_0	Reject H_0
If H_0 is true	No error	TYPE I ERROR
If H_0 is false	TYPE II ERROR	No error



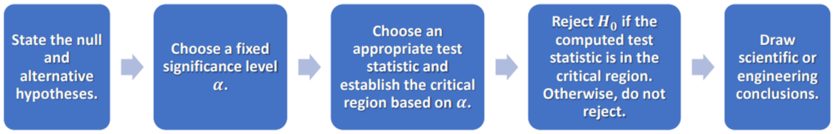
Example

	Tested positive	Tested negative
Infected with virus	No error	TYPE I ERROR
Not infected	TYPE II ERROR	No error

5.1.3 Test Statements

In designing a test statement, we can only control one of the errors: normally guarantee α in a desired low value (often use 0.01, 0.05 or 0.1), and then find a test statement with β as small as possible. Consider how to design a test statement with a restriction of α (called a **significance level**):

- By **critical value**:



State the null and alternative hypotheses.

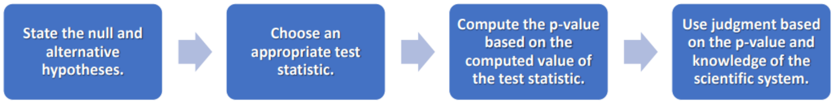
Choose a fixed significance level α.

Choose an appropriate test statistic and establish the critical region based on α.

Reject H_0 if the computed test statistic is in the critical region. Otherwise, do not reject.

Draw scientific or engineering conclusions.

- By **p-value**:


State the null and alternative hypotheses.
Choose an appropriate test statistic.
Compute the p-value based on the computed value of the test statistic.
Use judgment based on the p-value and knowledge of the scientific system.

Definition 5.4 (Power)**.** The **power of a test statement** is defined as $1 - \beta$ i.e., the probability of rejecting H_0 if H_0 is false and is often used to assess the goodness of the test statement.

To increase the power of a test $(1 - \beta)$, then the Type II Error probability β will decrease at the cost of the Type I Error probability α increasing. However, by using *more data*, the power of a test can be decreased w/o a change in α.

5.2 Hypothesis Testing - Normal Case

5.2.1 Formulation of a Test Statement about μ_X

When $X \sim N(\mu_X, \sigma_X^2)$, we will consider a one-sided test (left and right), as well as a two-sided test for σ_x^2 both known and unknown. When σ_X^2 is **known**,

$$\frac{\bar{x} - \mu_X}{\sigma_X/\sqrt{n}} \sim N(0, 1)$$

When σ_X^2 is **unknown**,

$$\frac{\bar{x} - \mu_x}{S_{n-1}/\sqrt{n}} \sim t(n-1)$$

where $\frac{\bar{x} - \mu_x}{S_{n-1}/\sqrt{n}}$ is often called the **t value**.

One-sided right test: Consider $H_0 : \mu_X = \mu_0, H_1 : \mu_x > \mu_0$. Intuitively, we reject H_0 if $\bar{x} > c$

- **Critical Value:** reject H_0 at a significance level α if

$$\bar{x} > \mu_0 + z_\alpha \frac{\sigma_x}{\sqrt{n}} \quad \text{when } \sigma_X^2 \text{ is KNOWN}$$

$$\bar{x} > \mu_0 + t_{n-1,\alpha} \frac{S_{n-1}}{\sqrt{n}} \quad \text{when } \sigma_X^2 \text{ is UNKNOWN}$$

- **P-Value:** reject H_0 at a significance level α if

$$\text{p-value} = P(Z > \frac{\bar{x} - \mu_0}{\sigma_X/\sqrt{n}}) < \alpha$$

where $Z \sim N(0, 1)$ and σ_X^2 is KNOWN

$$\text{p-value} = P(T_{n-1} > \frac{\bar{x} - \mu_0}{s_{n-1}/\sqrt{n}}) < \alpha$$

where $T_{n-1} \sim t(n-1)$ and σ_X^2 is UNKNOWN

One-sided left test: Consider $H_0 : \mu_X = \mu_0, H_1 : \mu_X < \mu_0$. Intuitively, we reject H_0 if $\bar{x} < c$

- **Critical Value:** reject H_0 at a significance level α if

  $$\bar{x} < \mu_0 - z_\alpha \frac{\sigma_x}{\sqrt{n}} \quad \text{when } \sigma_X^2 \text{ is KNOWN}$$

  $$\bar{x} < \mu_0 - t_{n-1,\alpha} \frac{S_{n-1}}{\sqrt{n}} \quad \text{when } \sigma_X^2 \text{ is UNKNOWN}$$

- **P-Value:** reject H_0 at a significance level α if

  $$\text{p-value} = P(Z < \frac{\bar{x} - \mu_0}{\sigma_X/\sqrt{n}}) < \alpha$$

  where $Z \sim N(0, 1)$ and σ_X^2 is KNOWN

  $$\text{p-value} = P(T_{n-1} < \frac{\bar{x} - \mu_0}{s_{n-1}/\sqrt{n}}) < \alpha$$

  where $T_{n-1} \sim t(n-1)$ and σ_X^2 is UNKNOWN

Two-sided test:

- **Critical Value:** reject H_0 at a significance level α if

  $$|\frac{\bar{x} - \mu_0}{\sigma_X/\sqrt{n}}| > z_{\alpha/2} \quad \text{when } \sigma_X^2 \text{ is KNOWN}$$

  $$|\frac{\bar{x_0} - \mu_0}{s_{n-1}/\sqrt{n}}| > t_{n-1,\alpha/2} \quad \text{when } \sigma_X^2 \text{ is UNKNOWN}$$

- **P-Value:** reject H_0 at a significance level α if

  $$\text{p-value} = P(|Z| > |\frac{\bar{x} - \mu_0}{\sigma_X/\sqrt{n}}|) = 2P(Z > |\frac{\bar{x} - \mu_0}{\sigma_X/\sqrt{n}}|) < \alpha$$

  where $Z \sim N(0, 1)$ and σ_X^2 is KNOWN

  $$\text{p-value} = P(|T_{n-1}| > |\frac{\bar{x} - \mu_0}{s_{n-1}/\sqrt{n}}|) = 2P(T_{n-1} > |\frac{\bar{x} - \mu_0}{s_{n-1}/\sqrt{n}}|) < \alpha$$

  where $T_{n-1} \sim t_{n-1}$ and σ_X^2 is UNKNOWN

5.2.2 Formulation of a Test Statement about σ_X^2

When $X \sim N(\mu_X, \sigma_X^2)$, we will consider a one-sided test (left and right), as well as a two-sided test for μ_X both known and unknown. When μ_X is **unknown**,

$$\frac{(n-1)S_{n-1}^2}{\sigma_X^2} \sim \chi^2(n-1)$$

One-sided right test: Consider $H_0 : \mu_X = \mu_0, H_1 : \mu_x > \mu_0$. Intuitively, we reject H_0 if $\bar{x} > c$

- **Critical Value:** reject H_0 at a significance level α if

$$\frac{(n-1)s_{n-1}^2}{\sigma_0^2} > \chi_{n-1,\alpha}^2 \quad \text{when } \mu_x \text{ is UNKNOWN}$$

- **P-Value:** reject H_0 at a significance level α if

$$\text{p-value} = P(U_{n-1} > \frac{(n-1)s_{n-1}^2}{\sigma_0^2}) < \alpha$$

where $U_{n-1} \sim \chi^2(n-1)$ and μ_x is UNKNOWN

One-sided left test: Consider $H_0 : \mu_X = \mu_0, H_1 : \mu_X < \mu_0$. Intuitively, we reject H_0 if $\bar{x} < c$

- **Critical Value:** reject H_0 at a significance level α if

$$\frac{(n-1)s_{n-1}^2}{\sigma_0^2} < \chi_{n-1,1-\alpha}^2 \quad \text{when } \mu_x \text{ is UNKNOWN}$$

- **P-Value:** reject H_0 at a significance level α if

$$\text{p-value} = P(U_{n-1} < \frac{(n-1)s_{n-1}^2}{\sigma_0^2}) < \alpha$$

where $U_{n-1} \sim \chi^2(n-1)$ and μ_x is UNKNOWN

Two-sided test:

- **Critical Value:** reject H_0 at a significance level α if

$$\frac{(n-1)s_{n-1}^2}{\sigma_0^2} < \chi_{n-1,1-\alpha/2}^2 \text{ OR } \frac{(n-1)s_{n-1}^2}{\sigma_0^2} > \chi_{n-1,\alpha/2}^2$$

when μ_X is UNKNOWN

- **P-Value:** reject H_0 at a significance level α if

$$\text{p-value} = 2min\{P(U_{n-1} < \frac{(n-1)s_{n-1}^2}{\sigma_0^2}), P(U_{n-1} > \frac{(n-1)s_{n-1}^2}{\sigma_0^2})\} < \alpha$$

where $U_{n-1} \sim \chi_{n-1}^2$ and μ_x is UNKNOWN

6 Simple Linear Regression

6.1 Simple Linear Regression and Least Squares

Definition 6.1 (Simple Linear Regression)**. Simple Linear Regression** is a statistical model used to study the relationship between y and x if they are related **linearly**.

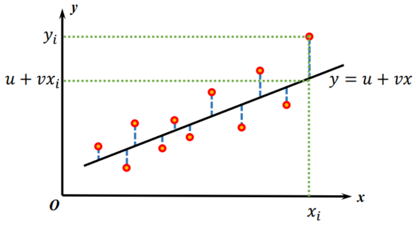
A **scatter plot** is used to visualize *paired* data of x and y, denoted

$$\{(x_i, y_i), i = 1, ..., n\}$$

This course will only consider two variables (a **response variable** y to an **explanatory variable** x) which are linearly related and statistically fitted to

$$Y = \beta_0 + \beta_1 x + \varepsilon$$

Note that ε is *random* and used to measure all uncertainty of the model (e.g., a measurement error) and the **regression coefficients** β_0, β_1 are unknown.



6.1.1 Least Squares Approach

With the function $S(\cdot, \cdot)$ of the total squared difference between the data and a straight line:

$$S(u, v) = \sum_{i=1}^{n} [y_i - (u + vx_i)]^2$$

We try to find a point (a, b) at which $S(\cdot, \cdot)$ is at a minimum:

$$a = \bar{y} - b\bar{x}$$

where a and b are the **least squared estimates**, β_0 and β_1 respectively (in $Y = \beta_0 + \beta_1 x + \varepsilon$). The **estimated regression line**: $\hat{y} = a + bx$

If we substitute x_i (for $i = 1, ..., n$) to the fitted regression line, then we can calculated the fitted value $\hat{y}_i$ of the ith observation y_i. The **residual** e_i of y_i is

$$e_i = y_i - \hat{y}_i$$

The **(residual) sum of squares (SSE)** is used to measure of evaluating the goodness of a simple linear regression model, defined

$$SSE = \sum_{i=1}^{n} [y_i - \hat{y}_i]^2$$

6.2 Correlation

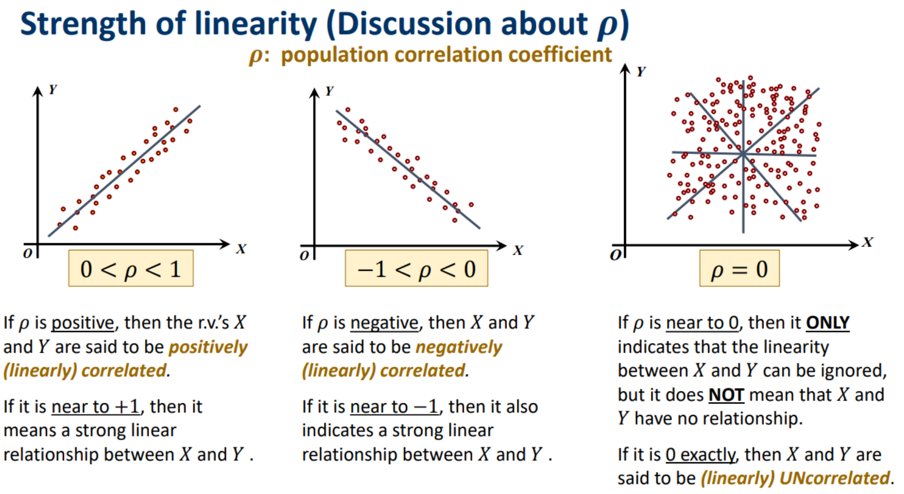
6.2.1 Strength of Linearity

When a relationship is **linear**, Pearson's correlation coefficient r (dimensionless value between -1 and 1) can be used to quantify the strength/degree of **linearity**.

$$r = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2}\sqrt{\sum_{i=1}^{n}(y_i - \bar{y})^2}}$$

r is used to estimate a population quantity called the **population correlation coefficient** ρ of two random variables X and Y:

$$\rho = \frac{E[(X - \mu_X)(Y - \mu_Y)]}{\sigma_X \sigma_Y}$$


Strength of linearity (Discussion about ρ)

ρ: population correlation coefficient

$0 < \rho < 1$

If ρ is positive, then the r.v.'s X and Y are said to be *positively (linearly) correlated*.

If it is near to $+1$, then it means a strong linear relationship between X and Y .

$-1 < \rho < 0$

If ρ is negative, then X and Y are said to be *negatively (linearly) correlated*.

If it is near to -1, then it also indicates a strong linear relationship between X and Y .

$\rho = 0$

If ρ is near to 0, then it **ONLY** indicates that the linearity between X and Y can be ignored, but it does **NOT** mean that X and Y have no relationship.

If it is 0 exactly, then X and Y are said to be *(linearly) UNcorrelated*.

To construct a confidence interval for ρ, a **point estimator** (r – calculated by replacing all $(x_i, y_i), \bar{x}, \bar{y}$ with $(X_i, Y_i), \bar{X}, \bar{Y}$) and an **exact (or approximated) distribution** is needed.

6.2.2 Fisher Z-Transformation

Taking the randomness of data into account, Fisher showed that the random variable Z_{Fisher} follows an approximated normal distribution with mean $\frac{1}{2}ln(\frac{1+\rho}{1-\rho})$ and variance $\frac{1}{n-3}$.

$$Z_{\text{Fisher}} = \frac{1}{2}ln(\frac{1+r}{1-r})$$

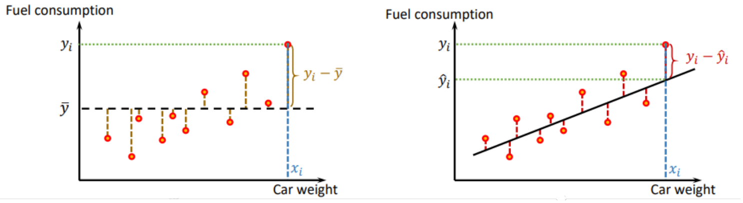
With this, we can construct a confidence interval for $\frac{1}{2}ln(\frac{1+\rho}{1-\rho})$ and then back-transform to get a C.I. for ρ.

6.2.3 Overall Fitting

To assess the goodness of our fitted regression line $\hat{y} = a + bx$, we can use the **coefficient of determination (R-squared)**.

$$\sum_{i=1}^{n}(y_i - \bar{y})^2 = \sum_{i=1}^{n}(\hat{y}_i - \bar{y})^2 + \sum_{i=1}^{n}(y_i - \hat{y}_i)^2$$

The **total** sum of squares (SST) = **regression** sum of squares (RSS) + sum of squared **error** (SSE). The *total* sum of squares refers to the total variability, while the sum of squared *error* referes to variability *unexplained* by the model.



The RSS is the variability *explained* by the model, with the **R-squared** value measuring the proportion of overall fit. The R^2 value is between 0 and 1.

$$R^2 = \frac{RSS}{SST} = 1 - \frac{SSE}{SST}$$

6.3 Inferences w/ Regression Coefficients; Prediction

6.3.1 Statistical Inferences about β_0 and β_1

With model $Y = \beta_0 + \beta_1 x + \varepsilon$, the *least-squared ESTIMATES* of β_0 and β_1 are,

$$a = \bar{y} - b\bar{x}, \quad b = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{n}(x_i - x)^2} = \frac{S_{XY}}{S_{XX}}$$

To study their estimation, we need r.v. to get the *least-squared ESTIMATORS*:

$$\hat{\beta}_0 = \bar{Y} - \hat{\beta}_1\bar{x}, \quad \hat{\beta}_1 = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(Y_i - \bar{Y})}{\sum_{i=1}^{n}(x_i - x)^2}$$

This model assumes (based on the random errors):

1. Errors are **independent**
2. Erros have **constant variance** σ^2
3. Erros have 0 mean
4. Errors follow a **normal distribution** (i.e., $\varepsilon \sim N(0, \sigma^2), i = 1, ..., n$ iid.)

Under these assumptions, then

$$\hat{\beta_0} \sim N(\beta_0, \frac{\sigma^2 \sum_{i=1}^n x_i^2}{nS_{XX}}), \hat{\beta_1} \sim N(\beta_1, \frac{\sigma^2}{S_{XX}})$$

where $S_{XX} = \sum_{i=1}^n (x_i - \bar{x})^2$. Both estimators $(\hat{\beta_0}, \hat{\beta_1})$ are unbiased esimators; if σ^2 is *known*, we can use the estimators directly to construct a confidence interval and form test statements. However, if σ^2 is *unknown*, it will need to be estimated:

Definition 6.2 (Mean Squared Error (MSE))**.** The **mean squared error (MSE)** is the unbiased estimator of σ^2:

$$S^2 = \frac{\sum_{i=1}^n E_i^2}{n-2} = \frac{\sum_{i=1}^n (Y_i - \hat{Y_i})^2}{n-2}$$

For simplicity, its actual value s^2 is also called the MSE:

$$s^2 = \frac{\sum_{i=1}^n e_i^2}{n-2} = \frac{\sum_{i=1}^n (y_i - \hat{y_i})^2}{n-2} = \frac{SSE}{n-2} = \frac{S_{YY} - bS_{XY}}{n-2}$$

After replacing the unknown σ^2 with the MSE, then the $100(1-\alpha)$**% confidence interval** is given by:

$$\beta_0 = (\bar{y} - b\bar{x}) \pm t_{n-2,\alpha/2}\sqrt{\frac{s^2 \sum_{i=1}^n x_i^2}{nS_{XX}}}$$

$$\beta_1 = b \pm t_{n-2,\alpha/2}\sqrt{\frac{s^2}{S_{XX}}}$$

Hypothesis testing, when σ^2 is unknown:
For the slope, β_1:

- One sided **right** test: consider $H_0 : \beta_1 = b_1, H_1 : \beta_1 > b_1$ – intuitively, reject H_0 if $b > c$
  - By critical value (t-value),

    $$\frac{b - b_1}{\frac{s}{\sqrt{S_{XX}}}} > t_{n-2,\alpha}$$

  - By p-value,

    $$P(T_{n-2} > \frac{b - b_1}{\frac{s}{\sqrt{S_{XX}}}}) < \alpha$$

- One sided **left** test: consider $H_0 : \beta_1 = b_1, H_1 : \beta_1 < b_1$ – intuitively, reject H_0 if $b < c$
  - By critical value (t-value),
$$\frac{b - b_1}{\frac{s}{\sqrt{S_{XX}}}} < -t_{n-2,\alpha}$$
  - By p-value,
$$P(T_{n-2} < \frac{b - b_1}{\frac{s}{\sqrt{S_{XX}}}}) < \alpha$$
- **Two-sided** test: consider $H_0 : \beta_1 = b_1, H_1 : \beta_1 \neq b_1$ – instinctively, reject H_0 if $b < c_1$ or $b > c_2$
  - By critical value (*absolute* t-value),
$$|\frac{b - b_1}{\frac{s}{\sqrt{S_{XX}}}}| > t_{n-2,\alpha/2}$$
  - By p-value,
$$2P(T_{n-2} > |\frac{b - b_1}{\frac{s}{\sqrt{S_{XX}}}}|) < \alpha$$

Similarly, for the intercept, β_0:

- One sided **right** test: consider $H_0 : \beta_0 = b_0, H_1 : \beta_0 > b_0$ – intuitively, reject H_0 if $b > c$
  - By critical value (t-value),
$$\frac{a - b_0}{s\sqrt{\frac{\sum_{i=1}^n x_i^2}{nS_{XX}}}} > t_{n-2,\alpha}$$
  - By p-value,
$$P(T_{n-2} > \frac{a - b_0}{s\sqrt{\frac{\sum_{i=1}^n x_i^2}{nS_{XX}}}}) < \alpha$$
- One sided **left** test: consider $H_0 : \beta_0 = b_0, H_1 : \beta_0 < b_0$ – intuitively, reject H_0 if $b < c$
  - By critical value (t-value),
$$\frac{a - b_0}{s\sqrt{\frac{\sum_{i=1}^n x_i^2}{nS_{XX}}}} < -t_{n-2,\alpha}$$
  - By p-value,
$$P(T_{n-2} < \frac{a - b_0}{s\sqrt{\frac{\sum_{i=1}^n x_i^2}{nS_{XX}}}}) < \alpha$$

- **Two-sided** test: consider $H_0 : \beta_0 = b_0, H_1 : \beta_0 \neq b_0$ – instinctively, reject H_0 if $b < c_1$ or $b > c_2$
  - By critical value (*absolute* t-value),

$$|\frac{a - b_0}{s\sqrt{\frac{\sum_{i=1}^n x_i^2}{nS_{XX}}}}| > t_{n-2,\alpha/2}$$

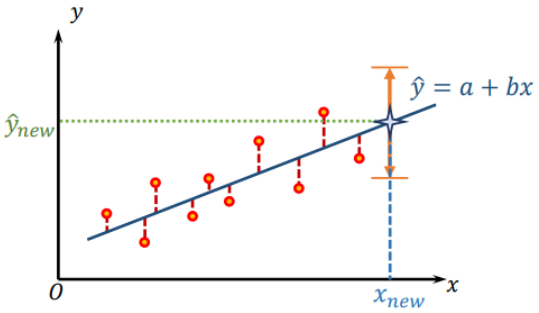
  - By p-value,

$$2P(T_{n-2} > |\frac{a - b_0}{s\sqrt{\frac{\sum_{i=1}^n x_i^2}{nS_{XX}}}}|) < \alpha$$

6.3.2 Prediction

With a fitted regression line, the next step is **prediction**: given a new observation of the *explanatory variable*, x_{new}, the corresponding value of the *response variable*, y_{new} is unknown, and can be predicted with

$$\hat{y}_{\text{new}} = a + bx_{\text{new}}$$



Using a random counterpart, the $100(1-\alpha)\%$ **prediction interval** for y_{new} can be determined with:

$$\hat{y}_{\text{new}} \pm t_{n-1,\alpha/2} s\sqrt{1 + \frac{1}{n} + \frac{(x_{\text{new}} - \bar{x})^2}{S_{XX}}}$$