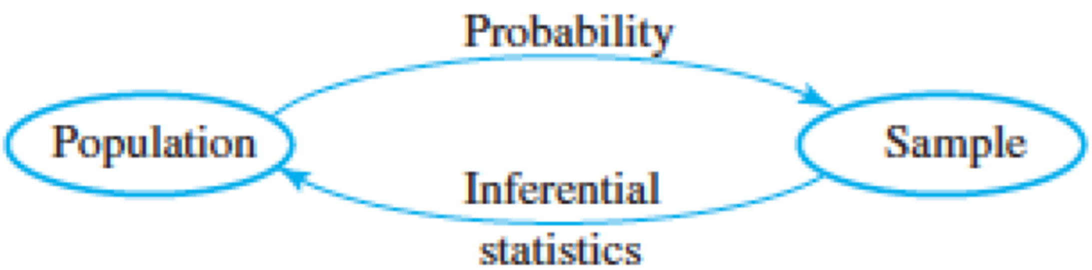
Topic 1. Review: Probability and Statistics

Population and Sample

- **Population** is the whole set of individuals about which we attempt to draw conclusion.
- **Sample** is a part of population which is observed.
- Relation between population and sample

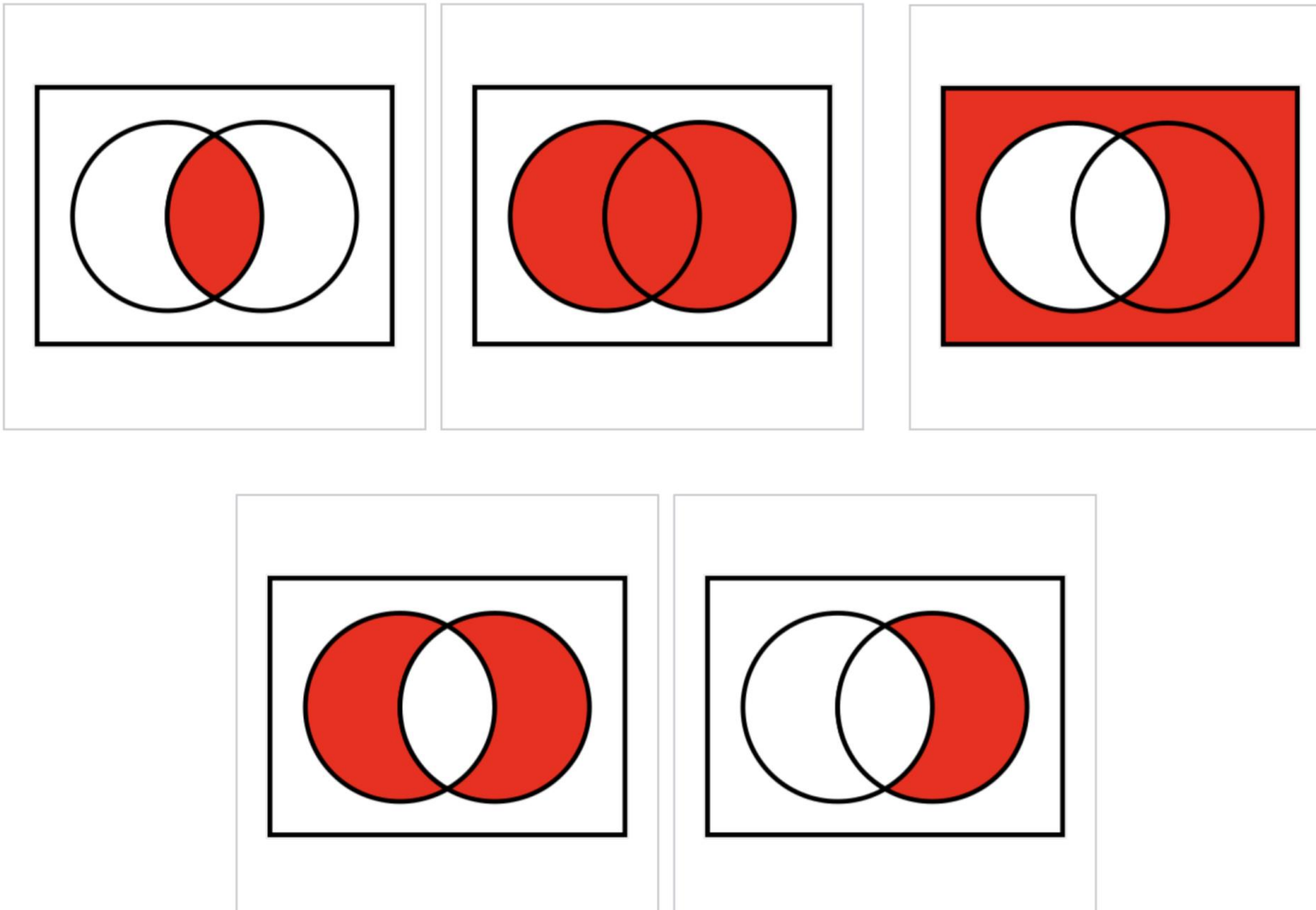


- Some toy examples
  - Flip a fair coin 10K times, what is the probability of observing 5.2K H’s?
  - If 5.2K H’s are observed, is the coin fair?

Basics of Probability

- An **experiment** is any action that generates observations.
- **Sample space** of an experiment, denoted as S, is the set of all possible outcomes of the experiment.
- An **event** is a subset of outcomes in S.
- Set operations: Given two events A and B
  - $A \cup B$: union of A and B
  - $A \cap B$: intersection of A and B
  - A': complement of A

Venn Diagram



Probability

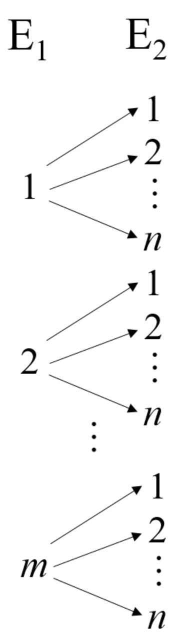
- Probability of an event A is

$$P(A) = \frac{\#\ of\ outcomes\ in\ A}{\#\ of\ outcomes\ in\ S}$$

- Counting techniques
  - **Product rule:** the number of outcomes for a composite event is the product of the numbers of outcomes for each simple event
  - **Permutation:** an ordered sequence of k objects from n distinct objects is a permutation; the number = P_n^k
  - **Combination:** any unordered k objects from n distinct objects is a combination; the number = C_n^k
  - $P_n^k = C_n^k \times k!$

Product Rule

- Two experiments: If Experiment 1 has m outcomes, and out of each outcome of Experiment 1, Experiment 2 has n outcomes, then together there are $m \times n$ outcomes.



$$\begin{array}{cccc} (1, 1) & (1, 2) & \ldots & (1, n) \\ (2, 1) & (2, 2) & \ldots & (2, n) \\ & & \vdots & \\ (m, 1) & (m, 2) & \ldots & (m, n) \end{array}$$

Example: Roll a die twice

$m = 6, n = 6$

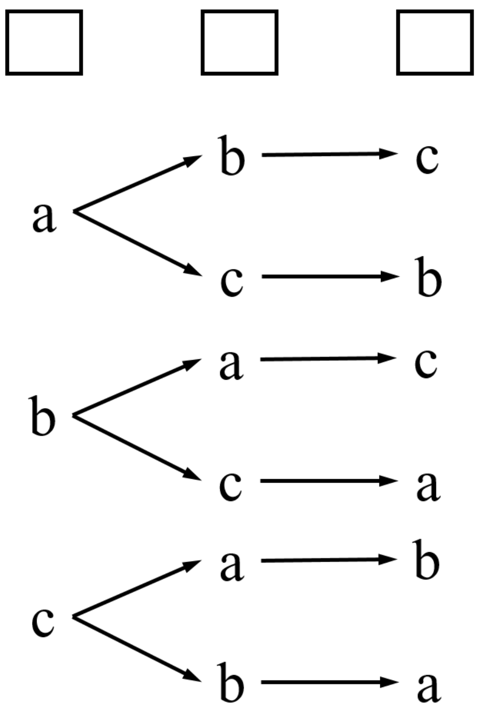
$$\begin{array}{cccc} (1, 1) & (1, 2) & \ldots & (1, 6) \\ (2, 1) & (2, 2) & \ldots & (2, 6) \\ & \vdots & & \\ (6, 1) & (6, 2) & \ldots & (6, 6) \end{array}$$

Permutation

- # ways to arrange n distinct objects:

$$P_n^k = \frac{n!}{(n-r)!} = n \times (n-1) \times (n-2) \ldots \times (n-r+1)$$

Example: Number of ways to order letters {a, b, c}



$$P_3^3 = 3! = 3 \times 2 \times 1 = 6$$

Combinations

- # ways to select r objects from a set of n objects:

$$C_n^k = \binom{n}{r} = \frac{n!}{r!\,(n-r)!}$$

Example: Number of ways to select 3 from {a, b, c, d, e}

$$\{\text{a, b, c}\},\ \{\text{a, b, d}\},\ \{\text{a, b, e}\},\ \{\text{a, c, d}\},\ldots$$

$$C_5^3 = \binom{5}{3} = \frac{5!}{3! \times 2!} = \frac{5 \times 4 \times 3 \times 2!}{3! \times 2!} = \frac{5 \times 4 \times 3}{3 \times 2 \times 1} = 10$$

Example of Probability Calculation

- A pair of fair dice is rolled. What is the probability that the second die lands on a higher value than the first?

Let S: All possible outcomes of rolling two dice
E: Second die has higher value than the first die

11	12	13	14	15	16
21	22	23	24	25	26
31	32	33	34	35	36
41	42	43	44	45	46
51	52	53	54	55	56
61	62	63	64	65	66

$$P(E) = \frac{\#\ of\ Outcomes\ in\ E}{total\ \#\ of\ outcomes\ in\ S}$$

$$= \frac{15}{36}$$

Axioms of Probability

- $P(A') = 1 - P(A)$
- If A and B are mutually exclusive, then

$$P(A \cap B) = P(\emptyset) = 0$$

- $P(A \cup B) = P(A) + P(B) - P(A \cap B)$
- Generalizing to three events

$$\begin{aligned} P(A \cup B \cup C) = P(A) + P(B) + P(C) \\ -P(A \cap B) - P(A \cap C) - P(B \cap C) \\ +P(A \cap B \cap C) \end{aligned}$$

Example of Probability Calculation

- There are 30 psychiatrists and 24 psychologists attending a certain conference. Three of these 54 people are randomly chosen to take part in a panel discussion. What is the probability that at least one psychologist is chosen?

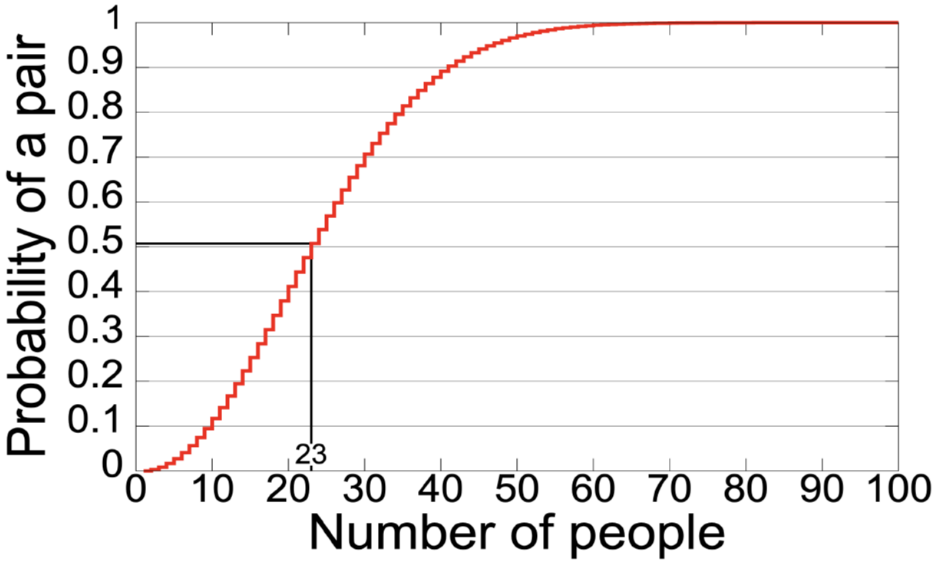
Let A: At least one psychologist is chosen

$$P(A) = 1 - P(A') = 1 - \frac{\binom{30}{3}}{\binom{54}{3}} = 1 - \frac{30 \times 29 \times 28}{54 \times 53 \times 52} = 0.84$$

Birthday Paradox

- **Question:** What is the probability that in a group of n people at least two share a birthday?
- Let A be the event that at least two share a birthday, then

$$P(A) = 1 - P(A') = 1 - \frac{365 \times 364 \times \cdots \times (365 - n + 1)}{365^n}$$



Conditional Probability

- The conditional probability of A given B occurred is

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

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

- A and B are independent if

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

- Bayes theorem

$$P(A_i|B) = \frac{P(A_i \cap B)}{P(B)} = \frac{P(A_i)P(B|A_i)}{\sum_{k=1}^{K} P(A_k)P(B|A_k)}$$

Bayes Theorem

Drug Testing

- **Question:** Suppose that a drug test produces 99% true positive results for drug users and 99% true negative results for non-drug users. Suppose that 0.5% of people are drug users. What is the probability that a random individual with a positive test is a drug user?

$$P(\text{User}|+) = \frac{P(+|\text{User})P(\text{User})}{P(+)}$$

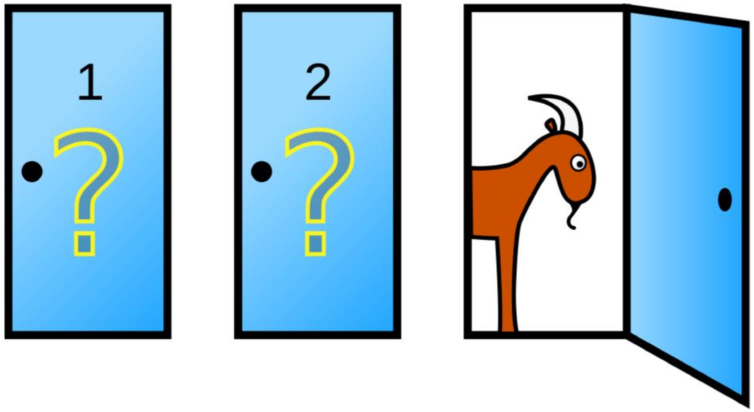
$$= \frac{P(+|\text{User})P(\text{User})}{P(+|\text{User})P(\text{User}) + P(+|\text{Non. user})P(\text{Non. user})}$$

$$= \frac{0.99 \times 0.005}{0.99 \times 0.005 + 0.01 \times 0.995}$$

$$= 33.2\%$$

The Monty Hall Problem

- **Question:** Say you're on a game show where there are three doors. Behind two of the doors, there are goats. Behind one of the doors, there is a brand new car.
- The host says that once you pick a door, he'll open one of the doors you didn't pick to reveal a goat. Then, you can either stay with your door or switch to the last unopened door.
- Do you switch or stay?



Random Variable

- A **random variable** is any characteristic(s) whose value(s) may change from one individual to another.

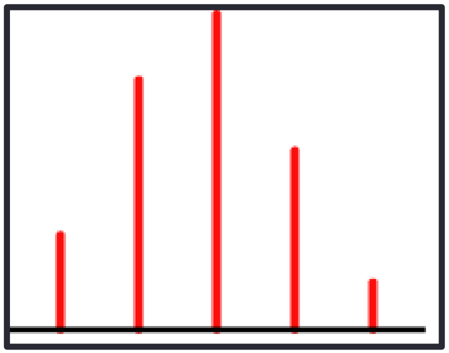
- Descriptive statistics
  - Numerical
    - Location: mean, median, trimmed mean
    - Variability: variance, standard deviation
  - Graphical
    - Histogram: frequency, relative frequency, density
    - Pie chart: proportion
    - Boxplot: median, 1st quantile, 3rd quantile, outlier

Discrete Random Variable

- A discrete r.v. X has a finite (countable) number of possible outcomes in S

- **Probability mass function (PMF)**

$$p(x) = P(s \in S{:}\, X(s) = x)$$



- **Cumulative distribution function (CDF)**

$$F(x) = P(X \leq x) = P(s \in S{:}\, X(s) \leq x)$$

- **Expectation**

$$E(X) = \sum_x xp(x)$$

- **Variance**

$$Var(X) = E[X - E(X)]^2 = E(X^2) - [E(X)]^2$$

Popular Discrete Random Variables

- Bernoulli: $X \sim Bern(p)$

$$p(x) = p^x(1-p)^{1-x}; x = 0,1$$

- Binomial: $X \sim Bin(n, p)$

$$p(x) = C_n^x p^x(1-p)^{n-x}; x = 0,1, \dots, n$$

- Geometric, Hypergeometric and Negative Binomial

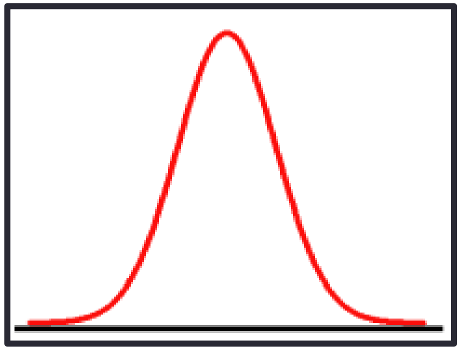
- Poisson: $X \sim Poi(\lambda)$ with $\lambda > 0$

$$p(x) = \frac{\lambda^x}{x!} e^{-\lambda}; x = 0,1, \dots$$

Continuous Random Variable

- A continuous r.v. $X's$ possible outcomes consist of an interval on the real line
- **Probability density function (PDF)**

$$f(x) = \lim_{h \to 0} P(x \leq X \leq x + h)$$



- **Cumulative distribution function (CDF)**

$$F(x) = P(X \leq x) = \int_{-\infty}^{x} f(t)dt$$

- **Expectation**

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

- **Variance**

$$Var(X) = E[X - E(X)]^2 = E(X^2) - [E(X)]^2$$

Popular Continuous Random Variables

- Uniform: $X \sim unif(a, b)$

$$f(x) = \frac{1}{b-a}; x \in [a, b]$$

- Exponential: $X \sim exp(\lambda)$ with $\lambda > 0$

$$f(x) = \lambda e^{-\lambda x}; x > 0$$

- Normal: $X \sim N(\mu, \sigma^2)$

$$f(x) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}; -\infty < x + \infty$$

- Gamma, Beta, Chi-square, Weibull, Lognormal,…

Joint Distribution

- For X and Y, continuous or discrete, its joint pdf is $f(x,y)$:
  - $P\big((X,Y) \in A\big) = \iint_A f(x,y)dydx$
  - Marginal pdf: $f_X(x) = \int f(x,y)dy$
  - X and Y are independent if $f(x,y) = f_X(x)f_Y(y)$
  - Conditional pdf: $f_{Y|X}(y|x) = \frac{f(x,y)}{f_X(x)}$
  - Expectation

$$E(h(X,Y)) = \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} h(x,y)f(x,y)dydx$$

  - Covariance: $cov(X,Y) = E(XY) - E(X)E(Y)$
  - Correlation: $corr(X,Y) = \frac{cov(X,Y)}{\sqrt{var(X)var(Y)}}$

Statistics and Their Distribution

- A **statistic** is a function of data, and thus it is a random variable.
- $X_1, X_2, \dots, X_n$ are called a (simple) random sample if they are i.i.d. (independent and identically distributed).
- For example, $\bar{X} = \frac{X_1 + X_2 + \cdots + X_n}{n}$ is a statistic.
- If $X_1, \dots, X_n$ are i.i.d. $N(\mu, \sigma^2)$, then $\bar{X} \sim N(\mu, \sigma^2/n)$
- **Central limit theorem** (CLT): If $X_1, \dots, X_n$ are i.i.d. with mean μ and variance σ^2,

$$\frac{\sqrt{n}(\bar{X} - \mu)}{\sigma} \xrightarrow{d} N(0,1)$$

Some General Results

- If $X_i \sim N(\mu_i, \sigma_i^2)$ and $X_1, \dots, X_n$ are independent, then

$$Y = \sum_{i=1}^{n} a_i X_i \sim N\left(\sum_{i=1}^{n} a_i \mu_i , \sum_{i=1}^{n} a_i^2 \sigma_i^2\right)$$

- If $X_1, \dots, X_n$ are independent with mean μ_i and variance σ_i^2, then

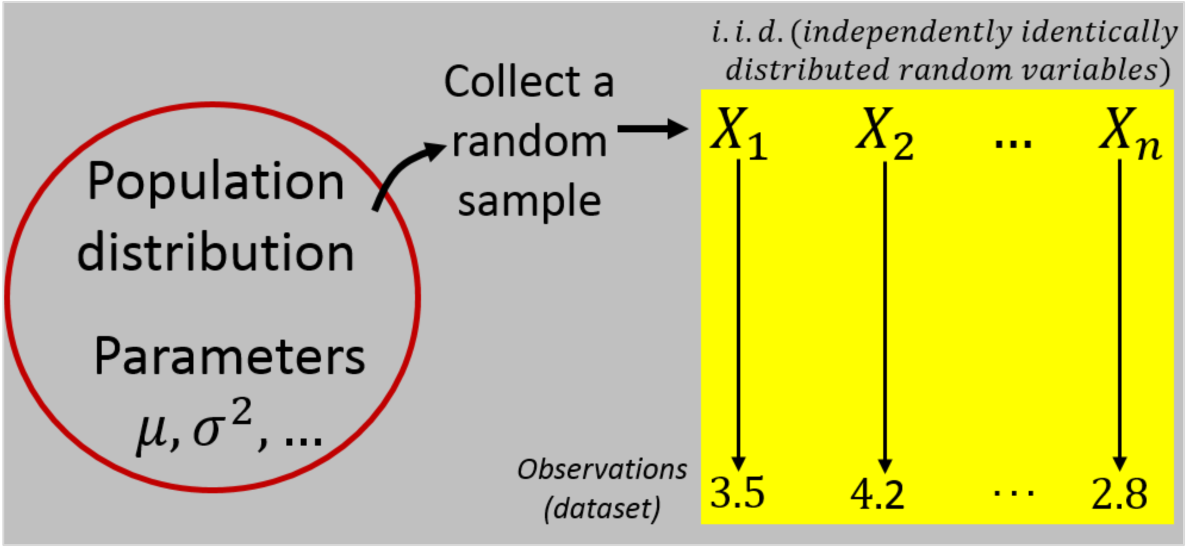
$$E(Y) = \sum_{i=1}^{n} a_i \mu_i \qquad \operatorname{var}(Y) = \sum_{i=1}^{n} a_i^2 \sigma_i^2$$

- If $E(X_i) = \mu_i$, $\operatorname{var}(X_i) = \sigma_i^2$, then

$$E(Y) = \sum_{i=1}^{n} a_i \mu_i \qquad \operatorname{var}(Y) = \sum_{i=1}^{n} a_i^2 \sigma_i^2 + \sum_{i \neq j} a_i a_j \operatorname{cov}(X_i, Y_j)$$

Statistical Inference

- Find truth on the population based on the data obtained from a sample of the population



- **Estimation:** Find estimates of the unknown parameters
  - Point estimation: $\hat{\mu} = 2.5$
  - Confidence interval (CI) estimation: the 95% CI of $\mu = (2.0, 3.0)$
- **Hypothesis testing:** Decisions based on specific hypotheses (e.g., $\mu \leq 2\ vs.\ \mu > 2$)

Point Estimation

- A **point estimate** of a parameter θ is a suitable statistic base on the given sample.
- Consider a random sample from some population:

24.46	25.61	26.25	26.42	26.66	27.15	27.31	27.54	27.74	27.94
27.98	28.04	28.28	28.49	28.50	28.87	29.11	29.13	29.50	30.88

- Estimate the true population mean
  - Sample mean: $\bar{x} = \frac{1}{20}(24.46 + \cdots + 30.88) = 27.79$
  - Sample median: $\tilde{x} = \frac{27.94+27.98}{2} = 27.96$
  - $\frac{min+max}{2} = \frac{24.46+30.88}{2} = 27.67$
  - 10% trimmed mean = $\frac{1}{16}(26.25 + \cdots + 29.13) = 27.84$
- Question: Which estimate is “closer” to the population mean?

Unbiased Estimator

- An estimator $\hat{\theta}$ is said to be unbiased of θ if $E(\hat{\theta}) = \theta$.
- $E(\hat{\theta}) - \theta$ is called the bias of $\hat{\theta}$
- For example, if $X_1, \dots, X_n$ are i.i.d. with mean μ and variance σ^2, then
  - $\bar{X}$ is an unbiased estimator of μ.
  - $\hat{\sigma}^2 = \frac{1}{n-1} \sum_i (X_i - \bar{X})^2$ is an unbiased estimator of σ^2.

Another Example

- Question: Let $X_1, \ldots, X_n (i.i.d) \sim unif(0, \theta)$, and $\hat{\theta} = \max_i \{X_i\}$. Is $\hat{\theta}$ an unbiased estimator of θ?

MVUE

- Among all unbiased estimators, choose the one that has minimum variance.
- The resulting $\hat{\theta}$ is called the **minimum variance unbiased estimator** (MVUE) of θ.
- In the previous example, let $\hat{\theta}_1 = \frac{n+1}{n} \max_i \{X_i\}$ and $\hat{\theta}_2 = 2\bar{X}$. Both are unbiased, but

$$\mathrm{var}\left(\hat{\theta}_1\right) = \frac{\theta^2}{n(n+1)} \leq \mathrm{var}\left(\hat{\theta}_2\right) = \frac{\theta^2}{3n} \text{ when } n \geq 1$$

- If $X_1, \ldots, X_n (i.i.d) \sim N(\mu, \sigma^2)$, then $\hat{\mu} = \bar{X}$ is the MVUE of μ.
- Difficult to find MVUE in general

Method of Moments (MM)

- The kth population moment is $E(X^k)$, and the kth sample moment is $\frac{1}{n}\sum_{i=1}^{n} X_i^k$.
- Let $X_1, \ldots, X_n$ be i.i.d. from $f(x;\theta)$, then the MM estimator is obtained by equating population moments to the corresponding sample moments and solving for θ.
- For example, if $X_1, \ldots, X_n(i.i.d)\sim N(\mu, \sigma^2)$, then

Maximum Likelihood Estimator (MLE)

- Let $X_1, \dots, X_n$ have joint pdf $f(x_1, \dots, x_n; \theta)$, which can also be regarded as a function of θ, called the likelihood function.

- The MLE estimator $\hat{\theta}$ is the value of θ that maximizes the likelihood function, so that

$$f(x_1, \dots, x_n; \theta) \leq f(x_1, \dots, x_n; \hat{\theta}); \text{ for any } \theta$$

- For example, if $X_1, \dots, X_n (i.i.d) \sim N(\mu, \sigma^2)$, then

Another Example

- Question: Let $X_1, \ldots, X_n (i.i.d) \sim unif(0, \theta)$, find the MM and MLE estimators of θ?

Confidence Interval

- A **confidence interval** (CI) is an interval estimate, computed based on certain statistics, that may contain the unknown population parameter with pre-specified probability.

- For example, if $X_1, \ldots, X_n$ is a random sample from $N(\mu, \sigma_0^2)$ with known σ_0^2, the $100(1-\alpha)\%$ CI for μ is

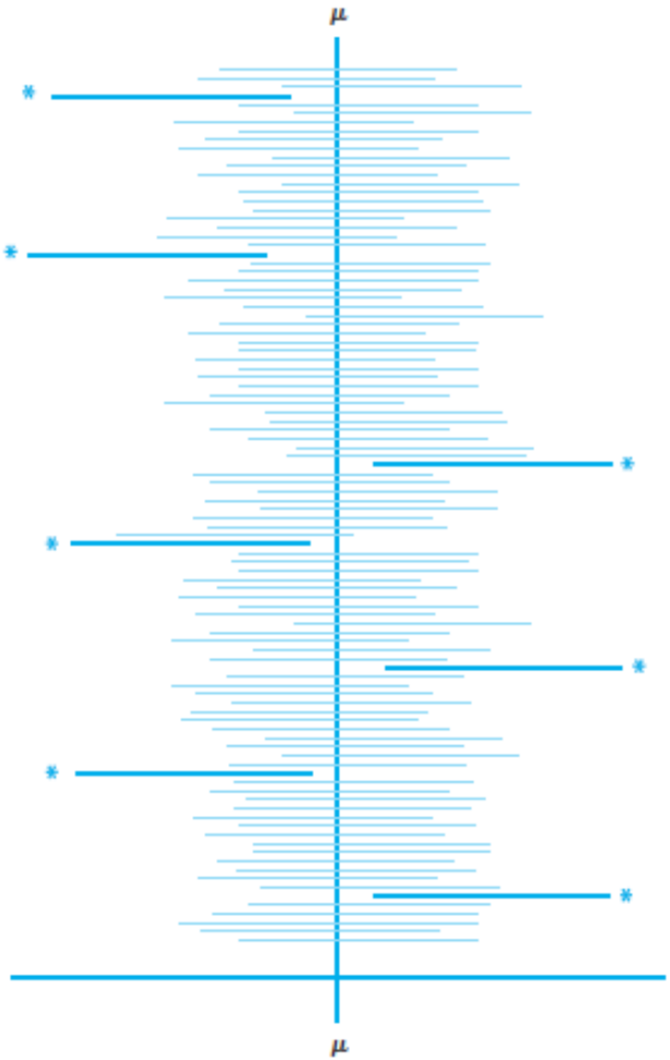
$$\left(\bar{X} - z_{\frac{\alpha}{2}} \frac{\sigma_0}{\sqrt{n}}, \bar{X} + z_{\frac{\alpha}{2}} \frac{\sigma_0}{\sqrt{n}}\right)$$

- For another example, the $100(1-\alpha)\%$ CI for μ of a normal distribution with unknown σ is

$$\left(\bar{X} - t_{\frac{\alpha}{2},n-1} \frac{s}{\sqrt{n}}, \bar{X} + t_{\frac{\alpha}{2},n-1} \frac{s}{\sqrt{n}}\right)$$

Interpretation of CI

- CI is a random interval whose endpoints are random.
- More specifically, get 100 random samples independently from the population, and construct CI for each sample.
- Out of the 100 CIs, about $100(1-\alpha)$ of them will cover μ.



A General Procedure for CI

- In general, let $X_1, \ldots, X_n$ be a random sample on which $\hat{\theta}$ is constructed to estimate a parameter θ, satisfying
  - approximately normal
  - unbiased (at least approximately)
  - $\mathrm{var}(\hat{\theta}) = \sigma_{\hat{\theta}}^2$ is available

Then $P\left(-z_{\alpha/2} \le \frac{\hat{\theta}-\theta}{\sigma_{\hat{\theta}}} \le z_{\alpha/2}\right) \approx 1-\alpha$, and thus an approximate $100(1-\alpha)\%$ CI for θ is

$$\left(\hat{\theta} - z_{\frac{\alpha}{2}}\sigma_{\hat{\theta}}, \hat{\theta} + z_{\frac{\alpha}{2}}\sigma_{\hat{\theta}}\right)$$

Hypothesis Test

- A **hypothesis** is a claim about certain characteristics of a probability distribution.
- A **hypothesis test** is a method for using data to decide between two competing claims about a population characteristic.
- The **null hypothesis**, denoted by H_0, is the claim that is initially assumed to be true.
- The **alternative hypothesis**, denoted by H_a, is the competing claim that is contradictory to H_0.

Testing Procedure

- Usually, for H_0: $\theta = \theta_0$, three possible alternatives:
  - H_a: $\theta > \theta_0$
  - H_a: $\theta < \theta_0$
  - H_a: $\theta \neq \theta_0$

- A **test procedure** is a rule based on sample data, for deciding whether to reject H_0
  - A test statistic: a function of the sample data on which the decision is to be based
  - Rejection region: the set of all values of the test statistic for which H_0 will be rejected.

Error Types

- The probability of Type I error is also known as the **significance level**, usually denoted by α.
- The probability of Type II error is denoted by β, and $1 - \beta$ is known as the **power**.
- Typically, increasing the significance level α results in a smaller value of β for any parameter value consistent with H_a.
- Construct a test procedure to minimize β, provided that its significance level is controlled by a pre-specified α.

	H_0 true	H_0 false
Fail to reject	$\surd$	Type II error
Reject	Type I error	$\surd$

P-value

- The **p-value** is the smallest level of significance at which H_0 will be rejected when the test is used on a given dataset.
- If p-value < α, then reject H_0; otherwise, fail to reject H_0.
- The **p-value** is the probability, calculated assuming that H_0 is true, of obtaining a value of the test statistic at least as contradictory to H_0 as the value calculated from the available sample data.
- Usually, α is set as 0.05, but use it with caution!