ST2334 Summary Notes

AY23/24 Sem 1, github.com/gerteck

1. Basic Probability Concepts

- **Sample Space:** S All possible outcomes of stat. expt.
- **Null Event**: Event that contains no element, empty set, $\varnothing$
- **Axioms of Probability**:
  For any event X, $0 \leq P(X) \leq 1$. $P(S) = 1$.
  If $A \cap B = \emptyset$ (Mut Excl), $P(A \cup B) = P(A) + P(B)$.
- Finite sample space with equally likely outcomes: $P(A) = (\frac{\#samplepointsA}{\#totalsamplepointsS})$. (e.g. birthday problem)

Event Operation & Relationships

- **Event Operations:** Union, Intersection, Complement.
- **Event Relationships:** Contained: $A \subset B$
  Equivalence: $A \subset B$ with $A \supset B \rightarrow A = B$
  Mutually Exclusive: $A \cap B = \emptyset$.
- **De Morgan's Law:** $(A \cup B)' = A' \cap B'$ and $(A \cap B)' = A' \cup B'$

Counting Methods

- Multiplication Principle: (Sequential Events)
- Addition Principle: (Pairwise Disjoin sets)
- **Permutation**: ${}_nP_r = \frac{n!}{(n-r)!}$
- **Combination**: $\binom{n}{r} = \frac{n!}{(n-r)!r!}$

Conditional Probability

- Understand conditional as reduced sample space.

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

Independence

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

Law of Total Probability

- **Partition:** If $A_1, \cdots, A_n$ mutually exclusive, $\bigcup_{i=1}^{n} A_i = S$, then $A_1, \cdots, A_n$ are partitions.
- If $A_1, \cdots, A_n$ are partitions of S, then for any event B:

$$P(B) = \sum_{i=1}^{n} P(B \cap A_i) = \sum_{i=1}^{n} P(B|A_i)P(A_i)$$

Bayes' Theorem

Let $A_1, \cdots, A_n$ be partitions of S. For any event B:

$$P(A_k|B) = \frac{P(B|A_k)P(A_k)}{\sum_{i=1}^{n} P(B|A_k)P(A_i)}$$

For when $n = 2$, $\{A, A'\}$ becomes a partition of S.

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

2. Random Variables

A function X, which assigns a real number to every s ∈ S is called a random variable.
- **Range space**: $Rx = \{x | x = X(s), s \in S\}$
- Likewise, the set X ∈ A, for A being a subset of R, is also a subset of $S : s \in S : X(s) \in A$.

Probability Distribution

Two main types of RV used in practice: discrete and continuous.

- Probability assigned to each possible X
- Given RV X with range of R_x:
  Discrete: Numbers in R_x are finite or countable
  Continuous: R_x is interval

(Discrete) Probability Mass Function $f(x)$:

$$f(x) \begin{cases} P(X = x), & \text{for } x \in R_X \\ 0, & \text{for } x \notin R_X \end{cases}$$

1. $f(x_i) = P(X = x_i) \geq 0$ for $x_i \in R_x$
2. $f(x_i) = 0$ for $x_i \notin R_x$
3. $\sum_{i=1}^{\infty} f(x_i) = 1$ (PSum = 1)
4. $\forall B \subseteq \mathbb{R}, P(X \in B) = \sum_{x_i \in B \cap R_x} f(x_i)$

(Continuous) Probability Density Function $f(x)$:

- Given R_x is interval. Quantifies probability that X is in some range.
- $p.f.$ must satisfy:
  1. $f(x) \geq 0$, $f(x) = 0$ for $x \notin R_x$
  2. No need $f(x) \leq 1$ (Concerned with area)
  3. $\int_{R_x} f(x)dx = 1$ (Integration over R_X = 1)
  4. $\forall a, b$ s.t. $a \leq b, P(a \leq X \leq b) = \int_a^b f(x)dx$
- **Note:** $P(X = x_0) = \int_{x_0}^{x_0} f(x)dx = 0$
- Hence, to check if a function is a pdf,
  1. $f(x) \geq 0$ for $x \in R_x$, $f(x) = 0$ for $x \notin R_x$
  2. $\int_{R_x} f(x)dx = 1$.

Cumulative Distribution Function

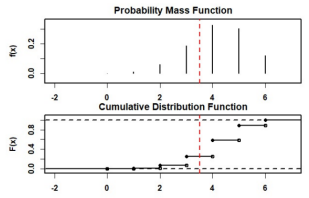
Describes distribution of a RV X: cumulative distribution function (cdf), applicable for discrete or continuous RV.

$$F(x) = P(X \leq x)$$

$F(x)$ is non-decreasing and $0 \leq F(x) \leq 1$
- Probability fn & cumulative distribution fn have one-to-one correspondence. For any probability fn given, the cdf is uniquely determined, vice versa.

CDF Discrete RV: Step Function $F(x)$

$$F(x) = \sum_{t \in R_x; t \leq x} f(t)$$

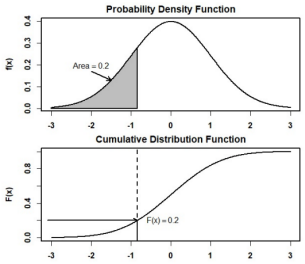


- $P(a \leq X \leq b) = P(X \leq b) - P(X < a)$
- $P(a \leq X \leq b) = F(b) - F(a-)$
- $P(a \leq X \leq b) = F(b) - \lim_{x \to a^-} F(x)$
- $0 \leq f(x) \leq 1$
- c.d.f has to be **right continuous** (• —)

CDF Continuous RV: $F(x)$

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

$$impt: f(x) = \frac{d(F(x))}{dx}$$



- $P(a \leq X \leq b) = P(a < X < b) = F(b) - F(a)$
- $0 \leq f(x)$.
  e.g. $f(x) = 3x^2$ is a valid $p.f.$ since $\int_{R_x} f(x)dx = 1$

Expectation μ & Variance σ

Expectation of Random Variable: μ

- **Mean of discrete RV**:

$$\mu = E(X) = \sum_{x \in R_x} x_i f(x_i)$$

- **E.g.**: X discrete RV with p.m.f. $f(x)$ and range R_X
  $\mu = E(g(x)) = \sum_{x \in R_x} g(x)f(x)$

- **Mean of continuous RV**:

$$\mu = E(X) = \int_{x \in R_x} xf(x)dx$$

- **E.g.**: X continuous RV with p.d.f. $f(x)$ and range R_X
  $\mu = E(g(x)) = \int_{x \in R_x} g(x)f(x)dx$

- **Properties of Expectation:**
- $E(aX + b) = aE(X) + b$
- Linearity of expectation: $E(X + Y) = E(X) + E(Y)$

Variance of Random Variable: σ

$$\sigma_X^2 = V(X) = E[(X - \mu_X)^2]$$

- **Variance of discrete RV**:

$$V(X) = \sum_{x \in R_x} (x - \mu_X)^2 f(x)$$

- **Variance of continuous RV**:

$$V(X) = \int_{x \in R_x} (x - \mu_X)^2 f(x)dx$$

- $V(X) \geq 0$ and $V(X) = 0$ when X is a constant
- $V(aX + b) = a^2V(X)$
- **alt. form:** $V(X) = E(X^2) - (E(X))^2$
- **Standard Deviation:** $\sigma_X = \sqrt{V(X)}$

3. Joint Distributions

- Consider more than 1 RV simultaneously,
- Given sample space S. Let X and Y be functions mapping $s \in S \to \mathbb{R}$: (X, Y) is 2D random vector.

Range spc: $R_{X,Y} = \{(x, y) | x = X(s), y = Y(s), s \in S\}$

- **Discrete 2D RV:**
  # of possible values of $(X(s), Y(s))$ finite / countable
- **Continuous 2D RV**:
  # of possible values of $(X(s), Y(s))$ assume any value in some region of the Euclidean space $\mathbb{R}^2$
- If both X and Y are discrete/continuous, then (X, Y) is discrete/continuous respectively.

Joint Probability Function

- **Joint Probability (mass) function, 2D discrete RV**:

$$f_{X,Y}(x, y) = P(X = x, Y = y)$$

  – $f_{X,Y}(x, y) \geq 0$ for any $(x, y) \in R_{X,Y}$
  – $f_{X,Y}(x, y) = 0$ for any $(x, y) \notin R_{X,Y}$
  – $\sum_{i=1}^{\infty} \sum_{j=1}^{\infty} P(X = x_i, Y = y_i) = 1$
  – Let $A \subseteq R_{X,Y}$.
  $P((X, Y) \in A) = \sum \sum_{(x,y) \in A} f_{X,Y}(x, y)$

- **Joint Probability (density) function, 2D cont. RV**:

$$P(a \leq X \leq b, c \leq Y \leq d) = \int_a^b \int_c^d f_{X,Y}(x, y)dydx$$

  – $f_{X,Y}(x, y) \geq 0$ for any $(x, y) \in R_{X,Y}$
  – $f_{X,Y}(x, y) = 0$ for any $(x, y) \notin R_{X,Y}$
  – $\int_{-\infty}^{\infty} \int_{-\infty}^{\infty} f_{X,Y}(x, y)dxdy = 1$
  or equivalently:
  – $\int \int_{(x,y) \in R_{X,Y}} f_{X,Y}(x, y)dxdy = 1$

Marginal Probability Function

Marginal distribution of X is individual distribution of X, ignoring the value of Y. "Projection" of 2D function $f_{X,Y}(x, y)$ to 1D function.

Let (X, Y) be 2D RV with joint probability function $f_{X,Y}(x, y)$:

$$\text{If } Y \text{ is } \textbf{discrete}, f_X(x) = \sum_y f_{X,Y}(x, y)$$

$$\text{If } Y \text{ is } \textbf{continuous}, f_X(x) = \int_{-\infty}^{\infty} f_{X,Y}(x, y)dy$$

- $f_Y(y)$ defined similarly
- $f_X(x)$ is a $p.f.$, satisfies all properties of prob. fn.

Conditional Distribution

Let (X, Y) be 2D RV with joint probability function $f_{X,Y}(x, y)$. Then $\forall x$ s.t. $f_X(x) > 0$: (X marg prob fn.)
Conditional probability function of Y given X = x:

$$f_{Y|X}(y|x) = \frac{f_{X,Y}(x, y)}{f_X(x)}$$

- Intuition: Distribution of Y given $X = x$
- Only defined for x s.t. $f_X(x) > 0$
- $f_{Y|X}(y|x)$ is a $p.f.$ if we fix x, satisfies prop. of prob.fn.
- But, $f_{Y|X}(y|x)$ is not a $p.f.$ for x: No need for sum / integral over x = 1. Hence,
  If $f_X(x) > 0$: $f_{X,Y}(x, y) = f_X(x)f_{Y|X}(y|x)$
  If $f_Y(y) > 0$: $f_{X,Y}(x, y) = f_Y(y)f_{X|Y}(x|y)$
- **Probability Y $\leq y$, Average Y given $X = x$**
- $P(Y \leq y | X = x) = \int_{-\infty}^{y} f_{Y|X}(y|x)dy$
- $E(Y|X = x) = \int_{-\infty}^{\infty} y f_{Y|X}(y|x)dy$

Independent Random Variables

$$X \perp Y : \forall x, y, f_{X,Y}(x,y) = f_X(x) f_Y(y)$$

- Necessary condition: $R_{X,Y}$ must be a product space. i.e. $R_{X,Y} = \{(x,y) | x \in R_X; y \in R_y\} = R_X \times R_Y$ Else, dependent.

Properties of Independent RV

For X, Y independent RV:

- If $A, B \subseteq \mathbb{R}$, then events $X \in A$ and $Y \in B$ are independent:

$$P(X \in A; Y \in B) = P(X \in A) P(Y \in B)$$

$$P(X \leq x; Y \leq y) = P(X \leq x) P(Y \leq y)$$

- Then, $g_1(X)$ and $g_2(Y)$ are **independent**, for arbitrary g.
- **Conditional distribution** given Independence:

$$f_X(x) > 0 \rightarrow f_{Y|X}(y|x) = f_Y(y)$$

$$f_Y(y) > 0 \rightarrow f_{X|Y}(x|y) = f_X(x)$$

To check independence

1. $R_{X,Y}$ is a product space. i.e. R_X does not depend on Y, vice versa. (e.g. $0 < y < x$ is NOT a product space)
2. Additionally, $f_{X,Y}(x,y)$ **= some** $C * g_1(x) g_2(y)$ **where** g_1 **depends on** x **only and** g_2 **depends on** y **only.**

Marginal Distribution under Independence

- Since, $f_{X,Y}(x,y) = f_X(x) f_Y(y)$ for independent RV, we derive marginal distribution by standardising $g_1(x)$ and $g_2(y)$.
- For discrete: $f_X(x) = \frac{g_1(x)}{\sum_{t \in R_X} g_1(t)}$
- For continuous: $f_X(x) = \frac{g_1(x)}{\int_{t \in R_X} g_1(t) dt}$

Expectation of a Random Vector

Given **2 variable function** $g(x,y)$**:**

- If (X,Y) is discrete:

$$E(g(X,Y)) = \sum_x \sum_y g(x,y) f_{X,Y}(x,y)$$

- If (X,Y) is continuous:

$$E(g(X,Y)) = \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} g(x,y) f_{X,Y}(x,y) dy dx$$

- If $X \perp Y$:

$$E(XY) = E(X)E(Y)$$

- (If $X \perp Y$, follows that $cov(X,Y) = 0$). However, converse not always true.

Covariance

- For random variables X, Y:

$$cov(X,Y) = E((X - E(X))(Y - E(Y)))$$

- If (X,Y) both **discrete**:

$$cov(X,Y) = \sum_x \sum_y (x - \mu_X)(y - \mu_Y) f_{X,Y}(x,y)$$

- If (X,Y) both **continuous**:

$$cov(X,Y) = \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} (x - \mu_X)(y - \mu_Y) f_{X,Y}(x,y) dx dy$$

- **Alt:** $cov(X,Y) = E(XY) - E(X)E(Y)$
- **Hence, for** $X \perp Y \rightarrow cov(X,Y) = 0$. (However, converse not always true).
- **Properties of covariance:**
- $cov(aX + b, cY + d) = (ac) cov(X,Y)$
- $V(aX + bY) = a^2 V(X) + b^2 V(Y) + 2ab * cov(X,Y)$
- $X \perp Y \rightarrow V(X \pm Y) = V(X) + V(Y)$

4.1 Special Probability Distributions

- **Discrete Distributions**: Study whole classes of discrete RVs that arise frequently in applications.

Discrete Uniform Distribution

- If X has values $x_1, x_2, \cdots, x_k$ with **equal probability**

$$f(x)\begin{cases}\frac{1}{k}, & \text{for } x = x_1, x_2, ..., x_k \\ 0, & \text{otherwise}\end{cases}$$

- **Expectation**: $\mu_X = E(X) = \sum_{i=1}^{k} x_i f_X(x_i) = \frac{1}{k}\sum x_i$
- **Variance**:
  $\sigma_X^2 = V(X) = E(X^2) - (E(X))^2 = \frac{1}{k}\sum x_i^2 - \mu_X^2$

Bernoulli, $Ber(p)$

- **Bernoulli Trial**: Random experiment has 2 possible outcomes (success and failure).
- **Bernoulli Random Variable**: X represents number of success in a single Bernoulli Trial. X has only two possible values: 1, or 0.
- **Probability mass function:** Let $0 \leq p \leq 1$ be the probability of success in Bernoulli trial

$$f_X(x) = P(X = x) = \begin{cases} p & x = 1 \\ 1-p & x = 0 \\ 0 & otherwise \end{cases}$$

- $f_X(x) = p^x(1-p)^{1-x}$ for $x = 0$ or 1
- Bernoulli distr. is case of binomial distr. where $n = 1$.
- **Notation:** $X \sim Ber(p)$ and $q = 1 - p$

$$f_x(1) = p, f_x(0) = q$$

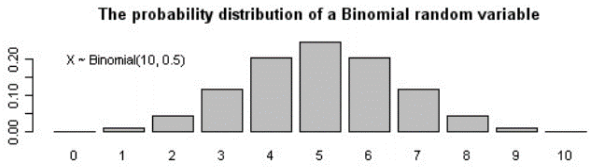
- **Expectation:** $\mu_X = E(X) = p$
- **Variance:** $\sigma_X^2 = V(X) = p(1-p)$
- **Bernoulli Process**: Sequence of repeatedly performed independent and identical Ber. trials.
- Generates sequence of **independent and identically distributed (i.i.d.)** Ber. RVs: $X_1, X_2, \cdots$

Binomial Distribution, $B(n, p)$

- **Binomial RV:** counts **number of successes** in n trials in a Ber. process.
- Given n independent trials with each trial having same probability p of success:

$$P(X = x) = \binom{n}{x} p^x (1-p)^{n-x}$$

- **Notation**: $X \sim B(n, p)$
- $E(X) = np$, $V(X) = np(1-p)$



Negative Binomial Distribution, $NB(k, p)$ (k^{th} success)

- Let X = no. of independent identical distributed Bernoulli(p) trials until k^{th} **success** occurs.
- **Probability mass function of X**:

$$P(X = x) = \binom{x-1}{k-1} p^k (1-p)^{x-k}$$

- **Notation**: $X \sim NB(k, p)$
- $E(X) = \frac{k}{p}$ and $V(X) = \frac{(1-p)k}{p^2}$

Geometric Distribution, $G(p)$ (till 1^{st} success)

- Let X = no. of i.i.d. Bernoulli(p) trials until 1st success occurs.

$$P(X = x) = p(1-p)^{x-1}$$

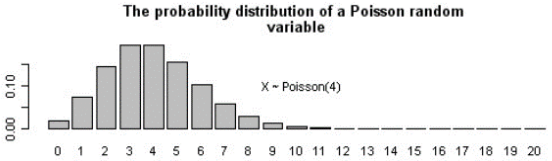
- **Notation**: $X \sim G(p)$
- $E(X) = \frac{1}{p}$ and $V(X) = \frac{1-p}{p^2}$

Poisson Distribution

- **Poisson RV**: Denotes number of events occuring in **fixed period of time or fixed region**, k = no. of occurences.

$$P(X = k) = \frac{e^{-\lambda}\lambda^k}{k!}$$

- **Notation**: $X \sim Poisson(\lambda)$ where $\lambda > 0$ is expected number of occurrences during given period/region
- $E(X) = \lambda$ and $V(X) = \lambda$



The number of infections X in a hospital each week has been shown to follow a Poisson distribution with a mean of 3 infections per week. What is the probability that

(a) there is *no* infection for a week?
(b) there are *less than* 4 infections for a week?

We are given than $X \sim$ Poisson(3). Then required probabilities are

(a) $P(X = 0) = e^{-3}$.

(b) $P(X \leq 3) = e^{-3}\left(1 + 3 + \frac{3^2}{2!} + \frac{3^3}{3!}\right)$.

Poisson Process

- Continuous time process, count number of occurrences within some interval of time. (given **rate** α)
- Properties of **Poisson process with rate parameter** α:
  - Expected no. of occurrences in interval length T: αT
  - No simultaneous occurrences, and no. of occurrences in disjoint intervals independent.
- **No. of occurrences in any interval** T of Poisson process follows $Poisson(\alpha T)$ distribution.
  (**Apply** $X \sim Poisson(\alpha T)$ **directly)**

Poisson Approximation of Binomial Distribution

- Let $X \sim B(n, p)$. Suppose $n \to \infty$ and $p \to 0$ s.t. $\lambda = np$ remains constant.
- Then, approximately, $X \sim Poisson(\lambda)$.

$$\lim_{p \to 0; n \to \infty} P(X = x) = \frac{e^{-np}(np)^x}{x!}$$

- Approximation is good when ($n \geq 20$ and $p \leq 0.05$), or ($n \geq 100$ and $np \leq 10$)
- Use $B(n, p)$: $E(X) = np$, $V(X) = np(1-p) = npq$

4.2 Special Probability Distributions

- **Continuous Distributions**: Many "natural" RVs whose set of possible values **uncountable**. Model with classes of continuous random variables.

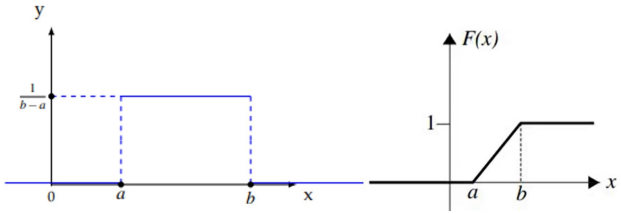
Continuous Uniform Distribution, $U(a,b)$

RV X follows uniform distribution over interval (a,b) if $p.d.f.$ given by:

$$f_X(x) = \begin{cases} \frac{1}{b-a} & a \le x \le b \\ 0 & otherwise \end{cases}$$

- **Notation**: $X \sim U(a,b)$
- $E(X) = \frac{a+b}{2}$ and $V(X) = \frac{(b-a)^2}{12}$ (derive by integration).
- **Cumulative distr. func.** $c.d.f.$ is given by:

$$F_X(x) = \begin{cases} 0 & x < a \\ \frac{x-a}{b-a} & a \le x \le b \\ 1 & x > b \end{cases}$$



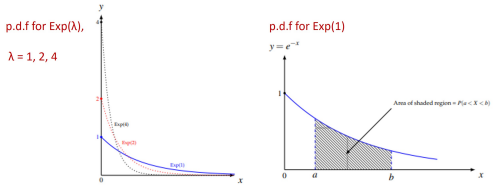
Exponential Distribution, $Exp(\lambda)$

- Continuous counterpart to **geometric distribution**.

X follows exponential distribution, with parameter $\lambda > 0$ if $p.f.$ is given by:

$$f_x(x) = \begin{cases} \lambda e^{-\lambda x} & x \ge 0 \\ 0 & x < 0 \end{cases}$$

- **Notation**: $X \sim Exp(\lambda)$
- $E(X) = \frac{1}{\lambda}$ and $V(X) = \frac{1}{\lambda^2}$



- We can **derive λ from mean / expectation of X**, since $E(X) = \frac{1}{\lambda}$.
- $c.d.f.$ is given by:

$$F_X(x) = P(X \le x) = \begin{cases} 1 - e^{-\lambda x} & x > 0 \\ 0 & x \le 0 \end{cases}$$

- Additionally, $P(X > x) = e^{-\lambda x}$, for $x > 0$.
- **Exponential distribution "Memoryless"**: Suppose X has exponential distribution with $\lambda > 0$. Then for any positive numbers s and t, we have:

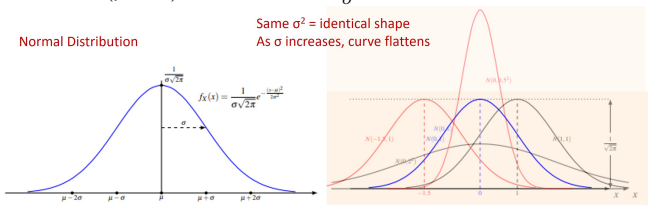
$$P(X > s + t | X > s) = P(X > t)$$

Normal Distribution, $N(\mu, \sigma^2)$

X said to follow normal distribution with mean μ and variance σ^2 if $p.f.$ given by:

$$f_X(x) = \frac{1}{\sqrt{2\pi}\sigma} e^{-(x-\mu)^2/(2\sigma^2)}$$

- **Notation**: $X \sim N(\mu, \sigma^2)$
- $E(X) = \mu$ and $V(X) = \sigma^2$
- $p.f.$ is **bell-shaped curve and symmetric** about $x = \mu$
- Total area under curve is 1
- 2 normal curves are identical in shape if they have same σ^2. They differ in location by $\mu_1 - \mu_2$.
- As σ increases, curve becomes more spread out
- If $X \sigma N(\mu, \sigma^2)$ and let $Z = \frac{X-\mu}{\sigma}$



Standardized Normal Distribution, $Z = N(0,1)$

If $X \sim N(\mu, \sigma^2)$, then $Z \sim N(0,1)$:

$$Z = \frac{X - \mu}{\sigma}$$

- $E(Z) = 0$ and $V(Z) = 1$
- $p.f$ **of Z is given by**:

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

- **Standardizing normal distribution** allows us to use tables to find probabilities:
- For $X \sim N(\mu, \sigma^2)$, compute $P(x_1 < X < x_2)$ by standardization:

$$x_1 < X < x_2 \leftrightarrow \frac{x_1 - \mu}{\sigma} < \frac{X - \mu}{\sigma} < \frac{x_2 - \mu}{\sigma}$$

- Then, $P(z_1 < Z < z_2)$, **use $f_Z(z)$ table to calculate**.
- **Cumulative d.f. of standard Normal**:

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

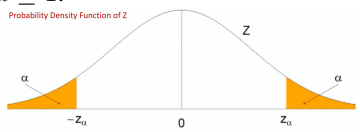
- $P(Z \ge 0) = P(Z \le 0) = \phi(0) = 0.5$
- For any z, $\Phi(z) = P(Z \le z) = P(Z \ge -z) = 1 - \phi(-z)$
- $-Z \sim N(0,1)$
- If $Z \sim N(0,1)$, then $\sigma Z + \mu \sim N(\mu, \sigma^2)$

Quantile

- **Upper Quantile:** x_α that satisfies:

$$P(X \ge x_\alpha) = \alpha$$

- where $0 \le \alpha \le 1$.



e.g. The 0.05th (upper) quantile of $Z \sim N(0,1)$ is 1.645, i.e. $z_{0.05} = 1.645$.

- $P(Z \ge z_\alpha) = P(Z \le -z_\alpha) = \alpha$
- Upper z_α = Lower $z_{1-\alpha}$

Normal Approximation to Binomial Distribution

Let $X \sim B(n,p)$, then as $n \to \infty$:

$$Z = \frac{X - E(X)}{\sqrt{V(X)}} = \frac{X - np}{\sqrt{np(1-p)}} \sim N(0,1)$$

- Approximation is good when $np > 5$ and $n(1-p) > 5$

5. Sampling, Sampling Distributions

Population and Sample

- **Statistical Inference**: Infer about population w. sample.
- **Population:** Totality of all possible obsv / outcomes.
- **Sample:** Subset of population
- Observation can be **numerical or categorical**
- Population can be **Finite or Infinite.**

Random Sampling

- Motivation: Often know what distribution population belongs to, but we not the parameters of distribution. Hence, use sample to estimate the parameters.

Single Random Sample

- **Simple Random Sample (SRS):** Sample of size n. Every subset of n observations (total $\binom{N}{n}$) equal chance of selection.

SRS for Infinite Population

- **For X be RV with certain** $p.f.$ $f_X(x)$:
- Let $X_1, X_2, \cdots, X_n$ be n independent RV with same distribution as X. Then $X_1, \cdots, X_n$ is a **simple random sample** of size n.
- **Joint probability function of $X_1, \cdots, X_n$: (product)**

$$f_{X_1,\cdots,X_n}(x_1, \cdots, x_n) = f_X(x_1) f_X(x_2) \cdots f_X(x_n)$$

Sampling with Replacement (as Infinite)

- **Sampling with replacement** from finite population is considered as sampling from **infinite population**.
- Sample is random if:
  - Every element in population has same probability
  - Successive draws are independent

Sample Distribution of Sample Mean

- **Statistic:** Suppose random sample of n observations is $X_1, \cdots, X_n$. A **statistic** is a function of $X_1, \cdots, X_n$
- **Sample Mean**

$$\bar{X} = \frac{1}{n}\sum_{i=1}^{n} X_i$$

- **Sample Variance**

$$S^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \bar{X})^2$$

- **Statistics are random variables**. If values in random sample observed, calculate **realization** of the statistic. Meaningful to consider distribution of statistics.

Sampling Distribution

Distribution of a statistic

- Mean and variance of $\bar{X}$:

$$E(\bar{X}) = \mu \text{ and } V(\bar{X}) = \frac{\sigma_X^2}{n}$$

μ_X is unknown constant. $\bar{X}$ serves as valid estimator for μ_X. As n increases, accuracy of $\bar{X}$ increases.

- **Standard Error:** Standard deviation of sampling distribution (e.g. $\sigma_{\bar{X}}$), describes how much $\bar{X}$ tends to vary from sample to sample of size n.
- **Law of Large Numbers:** As n increases, $\bar{X}$ converges to μ_X. i.e. For any $\epsilon \in \mathbb{R}$:

$$P(|\bar{X} - \mu| > \epsilon) \to 0 \text{ as } n \to \infty$$

As n increases, probability that sample mean differs from population mean goes to zero.

Central Limit Theorem

$\bar{X}$, **mean of random sample of size** n from population with mean μ and variance σ^2, then as $n \to \infty$:

$$\bar{X} \sim N(\mu, \frac{\sigma^2}{n}) \text{ approximately}$$

- For large n, $\bar{X}$ is approximately normally distributed.
- If random sample is from normal population, $\bar{X}$ is normally distributed no matter value of n
- If very skewed, CLT may not hold even with large n.

Other Sampling Distributions

$\chi^2(n)$ (Chi) Distribution

- Let $Z_1, \cdots, Z_n$ be n independent and identically distributed standard normal RVs.
- A χ^2 RV with n **degrees of freedom** is defined as a RV with same distribution as $Z_1^2 + \cdots + Z_n^2$
- **Notation:** $\chi^2(n)$ with n degrees of freedom
- If $Y \sim \chi^2(n)$, then $E(Y) = n$ and $V(Y) = 2n$
- **For large** n, $\chi^2(n)$ is approximately $N(n, 2n)$
- If Y_1 and Y_2 are independent χ^2 RVs with m **and** n **degrees of freedom respectively,** then $Y_1 + Y_2$ is $\chi^2(m+n)$
- χ^2 distribution is a family of curves. All density functions have long right tail.

Sampling Distribution of S^2

- $E(S^2) = \sigma^2$

Sampling Distribution of $\frac{(n-1)S^2}{\sigma^2}$

If S^2 is variance of random sample of size n from normal population of variance σ^2, then:

$$\frac{(n-1)S^2}{\sigma^2} = \frac{\sum_{i=1}^{n}(X_i - \bar{X})^2}{\sigma^2}$$

has $\chi^2(n-1)$ distribution

Suppose 6 random samples are drawn from a normal population $N(\mu, 4)$. Define the sample variance

$$S^2 = \frac{1}{5}\sum_{i=1}^{n}(X_i - \bar{X})^2.$$

Find c such that $P(S^2 > c) = 0.05$.

Solution:

We know that $\frac{5S^2}{4} \sim \chi^2(5)$. Hence,

$$\begin{aligned} & P(S^2 > c) = 0.05 \\ \Leftrightarrow\ & P(5S^2/4 > 5c/4) = 0.05 \\ \Leftrightarrow\ & 5c/4 = \chi^2(5; 0.05) = 11.07 \\ \Leftrightarrow\ & c = 8.86. \end{aligned}$$

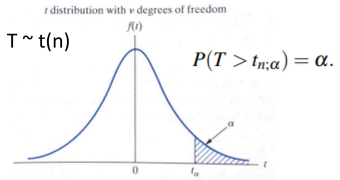
t-Distribution $t(n)$

Suppose $Z \sim N(0,1)$, $U \sim \chi^2(n)$. If Z, U independent:

$$T = \frac{Z}{\sqrt{U/n}} \sim t(n)$$

where $t(n)$ is t-distribution with n degrees of freedom

- **t-Distribution approaches $N(0,1)$ as $n \to \infty$.** When $n \geq 30$, t-dist approx normal, replace by $N(0,1)$.
- **Expectation, Variance:** If $T \sim t(n)$, then $E(T) = 0$ and $V(T) = \frac{n}{n-2}$ for $n > 2$
- Symmetric about vertical axis and resembles standard normal distribution
- **Critical value for t-distribution $t_{n;\alpha}$**: number with right hand tail probability of α.



- If $X_1, \cdots, X_n$ are independent and identically distributed normal RVs with mean μ and variance σ^2, then:

$$t.value = \frac{\bar{X} - \mu}{S/\sqrt{n}} \sim t(n-1)$$

i.e. follows **t distribution** with **n-1 degrees of freedom**.

L-EXAMPLE 5.12

A manufacturer of light bulbs claims that his light bulbs will burn on the average $\mu = 500$ hours. To maintain this average, he tests 25 bulbs each month.

If the computed t value, $\frac{\bar{x}-\mu}{s/\sqrt{n}}$, falls between $-t_{24;0.05}$ and $t_{24;0.05}$, he is satisfied with his claim.

What conclusion should be drawn from a sample that has a mean $\bar{x}$=518 hours and a standard deviation $s = 40$ hours? Assume that the distribution of burning times in hours is approximately normal.

Solution:

From the t-table or software, $t_{24;0.05} = 1.711$.

Therefore, the manufacturer is satisfied with his claim if a sample of 25 bulbs yields a t-value between -1.711 and 1.711.

If $\mu = 500$, then

$$t = \frac{518 - 500}{40/5} = 2.25 > 1.711.$$

Note that if $\mu > 500$, then the value of t computed from the sample would be more reasonable. Hence the manufacturer is likely to conclude that his bulbs are a better product than he thought.

F-Distribution $F(m,n)$

Suppose $U \sim \chi^2(m)$ and $V \sim \chi^2(n)$ independent:

$$F = \frac{U/m}{V/n} \sim F(m,n)$$

i.e. **F-distribution with (m,n) degrees of freedom**

- If $X \sim F(m,n)$, then **mean**:

$$E(X) = \frac{n}{n-2} \text{ for } n > 2$$

and **variance**:

$$V(X) = \frac{2n^2(m+n-2)}{m(n-2)^2(n-4)} \text{ for } n > 4$$

- Values of the F-distribution can be found in the statistical tables or software. The values of interests are $F(m,n;\alpha)$ such that

$$P(F > F(m,n;\alpha)) = \alpha,$$

where $F \sim F(m,n)$.

- It can be shown that

$$F(m,n;1-\alpha) = 1/F(n,m;\alpha).$$

- If $F \sim F(m,n)$, then $1/F \sim F(n,m)$

L-EXAMPLE 5.15

Let S_1^2 and S_2^2 be the sample variances of independent random samples of sizes $n_1 = 25$ and $n_2 = 31$, taken from normal populations with variances $\sigma_1^2 = 10$ and $\sigma_2^2 = 15$ respectively. Find $P(S_1^2/S_2^2 > 1.26)$.

Solution:

Note that

$$\frac{S_1^2/\sigma_1^2}{S_2^2/\sigma_2^2} \sim F(n_1 - 1, n_2 - 1),$$

which gives

$$\frac{S_1^2/10}{S_2^2/15} \sim F(24,30).$$

Thus

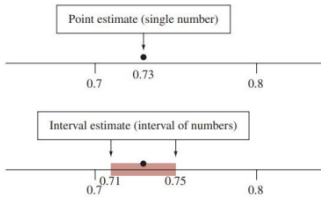
$$P\left(\frac{S_1^2}{S_2^2} > 1.26\right) = P\left(\frac{S_1^2/10}{S_2^2/15} > 1.26 \times \frac{15}{10}\right)$$
$$= P(F > 1.89) = 0.05.$$

Note that here $F \sim F(24,30)$.

06. Estimation

Two types of estimation (of population parameters):

- **Point estimation**: single number calculated to estimate, called point estimator)
- **Interval Estimation**: two numbers calculated to form an interval which the parameter is expected to lie.



Notation

- **Estimator**: An estimator is a rule (usually expressed as a formula) that tells us how to calculate an estimate based on info in sample.
- **Estimate**: Result of Estimator.
- **Concern**: How good is estimator? Criteria for good estimator?
- **Notation**: θ represents parameter of interest. θ can be p, μ, σ, etc.

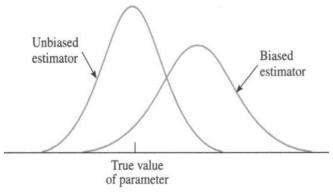
Point Estimation

Unbiased Estimator

Let $\hat{\theta}$ be an estimator of θ. Then $\hat{\theta}$ is unbiased if:

$$E(\hat{\theta}) = \theta$$

- This means, unbiased estimator has mean value equals to the true value of the parameter.



Example

- Let $X_1, ..., X_n$ be random sample from same population with mean μ and variance σ^2. Then, S^2 (sample variance, see formula in sampling), is an **unbiased estimator** of σ^2 as $E(S^2) = \sigma^2$.
- Sample mean $\bar{X}$ also U.E. for mean μ.

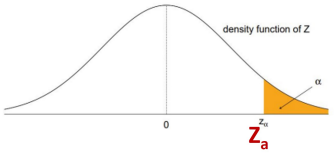
Error of Estimate

As typically $\bar{X} \neq \mu$ (estimator $\neq$ true value). We make use of $\bar{X} - \mu$ to measure difference between estimator and true value of parameter.
Recall if population normal or sufficiently large, $\frac{\bar{X}-\mu}{\sigma/\sqrt{n}}$ follows (approx) standard normal distribution.

Let $\bar{X}$ follow Std. Normal Distribution:

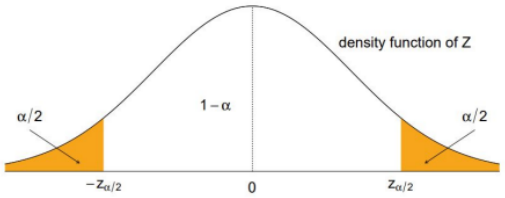
- Let z_α be αth upper quantile of standard normal distribution Z. i.e. $P(Z > z_\alpha) = \alpha$.

Define z_α to be the number with an upper-tail probability of α for the standard normal distribution Z. That is, $P(Z > z_\alpha) = \alpha$.



Then, we have

$$P(-z_{\alpha/2} \leq \frac{\bar{X}-\mu}{\sigma/\sqrt{n}} \leq z_{\alpha/2})$$
$$= P(|\bar{X} - \mu| \leq z_{\alpha/2} \times \frac{\sigma}{\sqrt{n}})$$
$$= 1 - \alpha$$



Hence:
Error $|\bar{X} - \mu|$ is less than $z_{\alpha/2}\frac{\sigma}{\sqrt{n}}$ with probability 1 - α.

Maximum Error of Estimate

- Given probability $1 - \alpha$: (vary α as desired)

$$E_{max} = z_{\alpha/2}\frac{\sigma}{\sqrt{n}}$$

Determination of Sample Size (so Error $\leq E_0$)

Minimum sample size n we can have, given probability $1 - \alpha$, so that maximum error is E:

$$n \geq (\frac{z_{\alpha/2}\sigma}{E})^2$$

Different Cases for Max Error & Min Sample Size

	Population	σ	n	Statistic	E	n for desired E_0 and α
I	Normal	known	any	$Z = \frac{\bar{X}-\mu}{\sigma/\sqrt{n}}$	$z_{\alpha/2} \cdot \frac{\sigma}{\sqrt{n}}$	$\left(\frac{z_{\alpha/2} \cdot \sigma}{E_0}\right)^2$
II	any	known	large	$Z = \frac{\bar{X}-\mu}{\sigma/\sqrt{n}}$	$z_{\alpha/2} \cdot \frac{\sigma}{\sqrt{n}}$	$\left(\frac{z_{\alpha/2} \cdot \sigma}{E_0}\right)^2$
III	Normal	unknown	small	$T = \frac{\bar{X}-\mu}{S/\sqrt{n}}$	$t_{n-1;\alpha/2} \cdot \frac{s}{\sqrt{n}}$	$\left(\frac{t_{n-1;\alpha/2} \cdot s}{E_0}\right)^2$
IV	any	unknown	large	$Z = \frac{\bar{X}-\mu}{S/\sqrt{n}}$	$z_{\alpha/2} \cdot \frac{s}{\sqrt{n}}$	$\left(\frac{z_{\alpha/2} \cdot s}{E_0}\right)^2$

Interval Estimation

- **Interval Estimator:** rule for calculating from a sample an interval (a, b) in which parameter lies.
- **Confidence Level:** Degree of confidence. Confidence level $(1 - \alpha)$, or the probability that interval contains parameter. i.e. $P = (1 - \alpha)$

$$P(a < \mu < b) = 1 - \alpha$$

- **Confidence Interval:** Interval calculated by interval estimator. i.e. (a, b) is called the $(1 - \alpha)$ confidence interval.

Case 1: σ known, data normal

Previously:

$$P(-z_{\alpha/2} \leq \frac{\bar{X} - \mu}{\sigma/\sqrt{n}} \leq z_{a/2}) = 1 - \alpha$$

By rearranging, the $(1 - \alpha)$ confidence interval (a, b) is:

$$(\bar{X} - z_{\alpha/2}\frac{\sigma}{\sqrt{n}}, \bar{X} + z_{\alpha/2}\frac{\sigma}{\sqrt{n}})$$

Other Cases of Confidence Interval for Pop. Mean

Case	Population	σ	n	Confidence Interval
I	Normal	known	any	$\bar{x} \pm z_{\alpha/2} \cdot \sigma/\sqrt{n}$
II	any	known	large	$\bar{x} \pm z_{\alpha/2} \cdot \sigma/\sqrt{n}$
III	Normal	unknown	small	$\bar{x} \pm t_{n-1;\alpha/2} \cdot s/\sqrt{n}$
IV	any	unknown	large	$\bar{x} \pm z_{\alpha/2} \cdot s/\sqrt{n}$

- n is considered large when $n \geq 30$

Interpreting Confidence Intervals

- We calculate that $X \pm E$ has probability $(1-\alpha)$ of containing μ.
- The probability is a **statement about the procedure** by which we compute the interval — the interval estimator.
- Each time we take a sample, and go through this construction, we get a **different confidence interval**. Sometimes we get a confidence interval that contains μ, and sometimes we get one not containing μ.
- Once an interval is computed, mu **is either in it or not. There is no more randomness.**

- Since μ is typically not known, no way to determine if true parameter in interval. **Confidence is in the method used**. If we repeat procedure of taking sample and computing confidence interval, about $(1-\alpha)$ of confidence intervals will contain the true parameter.

Comparing 2 Populations

We may want to compare the means of two populations, i.e. make statistical inference on $\mu_1 - \mu_2$.

Experimental Design

To compare, we need to take a number of observations from each population. Exp. design is manner in which samples collected from populations.

- **Independent Samples:** Completely randomized
- **Matched Pairs Samples:** Randomization btwn. matched pairs

Independent Samples (Known, Unequal Variance)

For: Random sample of size n_1 from population 1 with μ_1 and σ^2 and random sample of size n_2 from population 2 with μ_2 and σ^2. We define $\delta = \mu_1 - \mu_2$.
Conditions:
1. 2 Samples are independent
2. Population variances are **known and not same**: $\sigma_1^2 \neq \sigma_2^2$
3. Both populations are normal OR $n_1 \geq 30$ and $n_2 \geq 30$

Let $X_1, \cdots, X_{n_1}$ and $Y_1, \cdots, Y_{n_2}$ be random samples,
then:

$$E(\bar{X}) = \mu_1 \, , V(\bar{X}) = \frac{\sigma_1^2}{n_1} \, , E(\bar{Y}) = \mu_2 \, , V(\bar{Y}) = \frac{\sigma_2^2}{n_2}$$

$$E(\bar{X} - \bar{Y}) = \mu_1 - \mu_2 \, , V(\bar{X} - \bar{Y}) = \frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}$$

Thus, by normalizing RV $(\bar{X} - \bar{Y})$ and using assumption 3:

$$Z = \frac{(\bar{X} - \bar{Y}) - (\mu_1 - \mu_2)}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}} \sim N(0,1)$$

Thus, the $100(1-\alpha)\%$ **confidence interval** for $\mu_1 - \mu_2$ is:

$$(\bar{X} - \bar{Y}) \pm z_{\alpha/2} \sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}$$

Independent Samples (Unknown, Unequal Variance)

For: Random sample of size n_1 from population 1 with μ_1 and σ^2 and random sample of size n_2 from population 2 with μ_2 and σ^2, **where:**
1. 2 samples are independent, $n_1 \geq 30$ and $n_2 \geq 30$
2. Population variances are unknown and unequal $\sigma_1^2 \neq \sigma_2^2$.

Since σ_1 and σ_2 unknown, we use standard error:

$$S_1^2 = \frac{1}{n_1 - 1} \sum_{i=1}^{n_1} (X_i - \bar{X})^2 \, , S_2^2 = \frac{1}{n_2 - 1} \sum_{i=1}^{n_2} (Y_i - \bar{Y})^2$$

Thus, by normalizing RV $\bar{X} - \bar{Y}$ and using assumption 1:

$$Z = \frac{(\bar{X} - \bar{Y}) - (\mu_1 - \mu_2)}{\sqrt{\frac{S_1^2}{n_1} + \frac{S_2^2}{n_2}}} \sim N(0,1)$$

Thus, the $100(1-\alpha)\%$ **confidence interval** for $\mu_1 - \mu_2$ is:

$$(\bar{X} - \bar{Y}) \pm z_{\alpha/2} \sqrt{\frac{S_1^2}{n_1} + \frac{S_2^2}{n_2}}$$

Indpt. Samples (Small n, Equal Unknown Variance)

For: Random sample of size n_1 from population 1 with μ_1 and σ^2 and random sample of size n_2 from population 2 with μ_2 and σ^2.

where:

1. 2 samples are independent, $n_1 < 30$ and $n_2 < 30$.
2. Population variances are unknown but equal: $(\sigma_1^2 = \sigma_2^2)$
3. Both populations are **normally distributed**

Thus, by normalizing RV $\bar{X} - \bar{Y}$ and using cond. 1 and 3, and using pooled estimator to estimate σ^2 better:

$$T = \frac{(\bar{X} - \bar{Y}) - (\mu_1 - \mu_2)}{S_p \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}} \sim t_{n_1+n_2-2}$$

where S_p is the pooled sample variance and S_1^2 & S_2^2 are sample variances of samples:

$$S_p^2 = \frac{(n_1 - 1)S_1^2 + (n_2 - 1)S_2^2}{n_1 + n_2 - 2}$$

Thus, the $100(1-\alpha)\%$ confidence interval for $\mu_1 - \mu_2$ is:

$$(\bar{X} - \bar{Y}) \pm t_{n_1+n_2-2;\alpha/2} S_p \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}$$

Indpt. Samples (Large n, Equal Unknown Variance)

Since n is large, we can replace $t_{n_1+n_2-2;\alpha/2}$ with $z_{\alpha/2}$ in the previous formula.

For: Random sample of size n_1 from population 1 with μ_1 and σ^2 and random sample of size n_2 from population 2 with μ_2 and σ^2, **where**:

1. 2 samples are independent, $n_1 \geq 30$ and $n_2 \geq 30$
2. Population variances unknown but equal: $\sigma_1^2 = \sigma_2^2$

By applying CLT on large n, replace $t_{n_1+n_2-2;\alpha/2}$ with $z_{\alpha/2}$. Thus, the $100(1-\alpha)\%$ **confidence interval** for $(\mu_1 - \mu_2)$ is:

$$(\bar{X} - \bar{Y}) \pm z_{\alpha/2} S_p \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}$$

Paired Data

In cases where it makes sense to take matched data instead of independent samples (e.g. couple income, each couple independent of other couples).

For: $(X_1, Y_1), \cdots, (X_n, Y_n)$ are matched pairs, where $X_1, \cdots, X_n$ is random sample from population 1 and $Y_1, \cdots, Y_n$ is random sample from population 2.

where:

1. X_i and Y_i are dependent (within pair),
2. (X_i, Y_i) and (X_j, Y_j) are independent for any $i \neq j$.
3. For matched pairs, we define $D_i = X_i - Y_i$, and $\mu_D = \mu_1 - \mu_2$.
4. We can now treat $D_1, \cdots, D_n$ as random sample from a single population with μ_D and σ_D^2.

All techniques derived for single population can be used:

Consider the statistic:

$$T = \frac{\bar{D} - \mu_D}{S_D/\sqrt{n}}, \text{ where } \bar{D} = \frac{\sum_{i=1}^n D_i}{n} \text{ and}$$

$$S_D^2 = \frac{\sum_{i=1}^n (D_i - \bar{D})^2}{n-1}$$

If $n < 30$ and population is normally distributed:

$$T \sim t_{n-1}$$

Thus, if $n < 30$ and the population is normally distributed, the $100(1-\alpha)\%$ **confidence interval** for μ_D is:

$$\bar{d} \pm t_{n-1;\alpha/2} \frac{S_D}{\sqrt{n}}$$

Else, if $n \geq 30$:

$$T \sim N(0,1)$$

Thus, if $n \geq 30$, the $100(1-\alpha)\%$ **confidence interval** for μ_D is:

$$\bar{d} \pm z_{\alpha/2} \frac{S_D}{\sqrt{n}}$$

07. Hypothesis Testing

Both null and alternative hypothesis are statements about a population. Outcome of hypo. testing is to either **reject or fail to reject** the null hypothesis.

Steps for Hypothesis Testing

Step 1: Null Hypothesis and Alternative Hypothesis

- **Null Hypothesis** H_0: Parameter takes some value
- **Alternative Hypothesis** H_1: Parameter falls in alt. range
- Often, let hypothesis we want to prove be alt. hypothesis, as it states null hypothesis is false, often in a particular way.
- **2-Sided Test:** If H_1 is "Parameter $\neq H_0$ value"
- **Right-Sided Test:** If H_1 is "Parameter is $> H_0$ value"
- **Left-Sided Test:** If H_1 is "Parameter is $< H_0$ value"

Step 2: Level of Significance

	Do not reject H_0	Reject H_0
H_0 is true	Correct Decision	**Type I error**
H_0 is false	**Type II error**	Correct Decision

- **Level of Significance:** α, Probability of making type I error, rejecting H_0 when it is true. i.e.

$$\alpha = P(\text{Type I error}) = P(\text{Reject } H_0 | H_0 \text{ is true})$$

As type I is serious error, set small α, e.g. $\alpha = 0.05, 0.01$

Let

$$\beta = P(\text{Type II error} = P(\text{Do Not Reject } H_0 | H_0 \text{ is false}))$$

- **Power of the Test:** $(1-\beta) = P(\text{Reject } H_0 | H_0 \text{ is false})$

Step 3: Identify Test Statistic, its Distribution, and the Rejection Region / criteria

- **Test Statistic:** quantify how unlikely to observe sample, assuming null hypothesis H_0 is true.
- At significance level α, decision rule can found, divides set of possible values of test statistic into rejection (critical) region and acceptance region.

Step 4: Calculation & Conclusion

Given test statistic, determine if it is **in the rejection region**:

- If yes, sample too improbable, **reject** H_0, fail to reject H_1
- Otherwise, **do not reject** H_0, fall back to og. assumption.

Hypotheses for testing Popln. Mean

Case 1: Known Variance

Given that **population variance σ^2 is known** and **underlying distribution is normal OR $n \geq 30$.**

Steps:

1. Set null and alternative hypotheses. e.g.

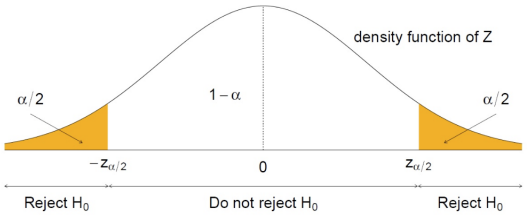
$$H_0 : \mu = \mu_0 \text{ vs } H_1 : \mu \neq \mu_0$$

2. Set level of significance (e.g. $\alpha = 0.05$)
3. With σ^2 known and population normal (or $n \geq 30$), the test statistic is (assume H_0 true):

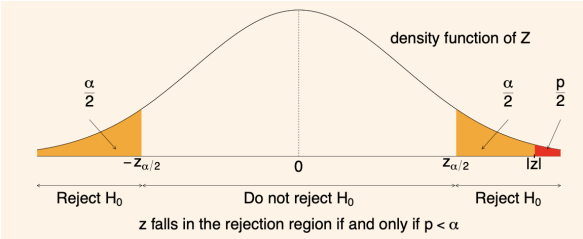
$$Z = \frac{\bar{X} - \mu_0}{\sigma/\sqrt{n}} \sim N(0,1)$$

Rejection region, where we let observed value of Z be z:

- $H_1 : \mu \neq \mu_0$: $z < -z_{\alpha/2}$ or $z > z_{\alpha/2}$
- $H_1 : \mu < \mu_0$: $z < -z_\alpha$
- $H_1 : \mu > \mu_0$: $z > z_\alpha$



- **p-Value:** Conditional probability that test statistic as extreme as observed value, given H_0 true.
- $H_1 : \mu \neq \mu_0$: $p = 2P(Z < -|z|)$
- $H_1 : \mu < \mu_0$: $p = P(Z < -|z|)$
- $H_1 : \mu > \mu_0$: $p = P(Z > |z|)$



4. • **Rejection region:** If z is inside rejection region, reject H_0. Otherwise do not reject.
   • **p-Value**: If p is less than α, reject H_0. Otherwise do not reject.

Case 2: Unknown Variance

Given that:

1. Population variance is unknown
2. Underlying distribution is normal

• Test statistic:

$$T = \frac{\bar{X} - \mu_0}{S/\sqrt{n}} \sim t_{n-1}$$

• Rejection region:
  – $H_1 : \mu \neq \mu_0$: $t < -t_{n-1;\alpha/2}$ or $t > t_{n-1;\alpha/2}$
  – $H_1 : \mu < \mu_0$: $t < -t_{n-1;\alpha}$
  – $H_1 : \mu > \mu_0$: $t > t_{n-1;\alpha}$
• When $n \geq 30$, we can replace t_{n-1} by Z

Two-sided Tests & Confidence Intervals

The **two-sided hypothesis test** procedure is equivalent to finding a $100(1-\alpha)\%$ **confidence interval** for μ.

• When confidence interval contains μ_0, H_0 will not be rejected at level α.
• Similarly, when confidence interval does not contain μ, then t falls within rejcetion region and so H_0 will be rejected.

Comparing Means: Independent Samples

• Given 2 independent samples from 2 populations, interested in testing

$$H_0 : \mu_1 - \mu_2 = \delta_0$$

against a suitable alternative hypothesis.

Rejection Regions and p-Values

H_1	Rejection Region	p-value
$\mu_1 - \mu_2 > \delta_0$	$z > z_\alpha$	$P(Z > \lvert z \rvert)$
$\mu_1 - \mu_2 < \delta_0$	$z < -z_\alpha$	$P(Z < -\lvert z \rvert)$
$\mu_1 - \mu_2 \neq \delta_0$	$z > z_{\alpha/2}$ or $z < -z_{\alpha/2}$	$2P(Z > \lvert z \rvert)$

Case 1: Known Variance

Consider case where:
1. Population variances are known
2. Underlying distributions are normal OR $n_1 \geq 30$ and $n_2 \geq 30$

• Test statistic:

$$Z = \frac{(\bar{X} - \bar{Y}) - \delta_0}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}} \sim N(0,1)$$

Case 2: Unknown Variance

Consider case where:
1. Population variances are unknown
2. $n_1 \geq 30$ and $n_2 \geq 30$

• Test statistic:

$$Z = \frac{(\bar{X} - \bar{Y}) - \delta_0}{\sqrt{\frac{S_1^2}{n_1} + \frac{S_2^2}{n_2}}} \sim N(0,1)$$

Case 3: Unknown but Equal Variance

Consider case where:
1. Population variances are unknown but equal
2. Underlying distributions are normal
3. $n_1 < 30$ and $n_2 < 30$

• Test statistic:

$$Z = \frac{(\bar{X} - \bar{Y}) - \delta_0}{S_p\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}} \sim t_{n_1+n_2-2}$$

Comparing Means: Paired Data

• Obtain difference, then use methods from single samples.
• Define

$$D_i = X_i - Y_i.$$

• For $H_0 : \mu_D = \mu_{D_0}$, test statistic:

$$T = \frac{\bar{D} - \mu_{D_0}}{S_D/\sqrt{n}}$$

  – If $n < 30$ and population is normally distributed,

$$T \sim t_{n-1}$$

  – If $n \geq 30$, $T \sim N(0,1)$

08. Additional Formulae & Misc

Integration by Parts

$$\int udv = uv - \int vdu$$

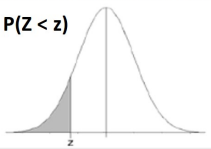
• How to choose u? LIPET

Geometric Series

$$\begin{aligned} s_n &= ar^0 + ar^1 + \cdots + ar^{n-1}, \\ rs_n &= ar^1 + ar^2 + \cdots + ar^n, \\ s_n - rs_n &= ar^0 - ar^n, \\ s_n(1-r) &= a(1-r^n), \\ s_n &= a\left(\frac{1-r^n}{1-r}\right), \text{ for } r \neq 1. \end{aligned}$$

Standard Normal Distribution Probabilities Table

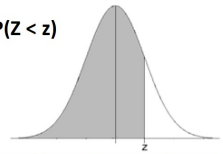
Lower-tail probability: P(Z < z)



z	.00	.01	.02	.03	.04	.05	.06	.07	.08	.09
−3.4	0.0003	0.0003	0.0003	0.0003	0.0003	0.0003	0.0003	0.0003	0.0003	0.0002
−3.3	0.0005	0.0005	0.0005	0.0004	0.0004	0.0004	0.0004	0.0004	0.0004	0.0003
−3.2	0.0007	0.0007	0.0006	0.0006	0.0006	0.0006	0.0006	0.0005	0.0005	0.0005
−3.1	0.0010	0.0009	0.0009	0.0009	0.0008	0.0008	0.0008	0.0008	0.0007	0.0007
−3.0	0.0013	0.0013	0.0013	0.0012	0.0012	0.0011	0.0011	0.0011	0.0010	0.0010
−2.9	0.0019	0.0018	0.0018	0.0017	0.0016	0.0016	0.0015	0.0015	0.0014	0.0014
−2.8	0.0026	0.0025	0.0024	0.0023	0.0023	0.0022	0.0021	0.0021	0.0020	0.0019
−2.7	0.0035	0.0034	0.0033	0.0032	0.0031	0.0030	0.0029	0.0028	0.0027	0.0026
−2.6	0.0047	0.0045	0.0044	0.0043	0.0041	0.0040	0.0039	0.0038	0.0037	0.0036
−2.5	0.0062	0.0060	0.0059	0.0057	0.0055	0.0054	0.0052	0.0051	0.0049	0.0048
−2.4	0.0082	0.0080	0.0078	0.0075	0.0073	0.0071	0.0069	0.0068	0.0066	0.0064
−2.3	0.0107	0.0104	0.0102	0.0099	0.0096	0.0094	0.0091	0.0089	0.0087	0.0084
−2.2	0.0139	0.0136	0.0132	0.0129	0.0125	0.0122	0.0119	0.0116	0.0113	0.0110
−2.1	0.0179	0.0174	0.0170	0.0166	0.0162	0.0158	0.0154	0.0150	0.0146	0.0143
−2.0	0.0228	0.0222	0.0217	0.0212	0.0207	0.0202	0.0197	0.0192	0.0188	0.0183
−1.9	0.0287	0.0281	0.0274	0.0268	0.0262	0.0256	0.0250	0.0244	0.0239	0.0233
−1.8	0.0359	0.0351	0.0344	0.0336	0.0329	0.0322	0.0314	0.0307	0.0301	0.0294
−1.7	0.0446	0.0436	0.0427	0.0418	0.0409	0.0401	0.0392	0.0384	0.0375	0.0367
−1.6	0.0548	0.0537	0.0526	0.0516	0.0505	0.0495	0.0485	0.0475	0.0465	0.0455
−1.5	0.0668	0.0655	0.0643	0.0630	0.0618	0.0606	0.0594	0.0582	0.0571	0.0559
−1.4	0.0808	0.0793	0.0778	0.0764	0.0749	0.0735	0.0721	0.0708	0.0694	0.0681
−1.3	0.0968	0.0951	0.0934	0.0918	0.0901	0.0885	0.0869	0.0853	0.0838	0.0823
−1.2	0.1151	0.1131	0.1112	0.1093	0.1075	0.1056	0.1038	0.1020	0.1003	0.0985
−1.1	0.1357	0.1335	0.1314	0.1292	0.1271	0.1251	0.1230	0.1210	0.1190	0.1170
−1.0	0.1587	0.1562	0.1539	0.1515	0.1492	0.1469	0.1446	0.1423	0.1401	0.1379
−0.9	0.1841	0.1814	0.1788	0.1762	0.1736	0.1711	0.1685	0.1660	0.1635	0.1611
−0.8	0.2119	0.2090	0.2061	0.2033	0.2005	0.1977	0.1949	0.1922	0.1894	0.1867
−0.7	0.2420	0.2389	0.2358	0.2327	0.2296	0.2266	0.2236	0.2206	0.2177	0.2148
−0.6	0.2743	0.2709	0.2676	0.2643	0.2611	0.2578	0.2546	0.2514	0.2483	0.2451
−0.5	0.3085	0.3050	0.3015	0.2981	0.2946	0.2912	0.2877	0.2843	0.2810	0.2776
−0.4	0.3446	0.3409	0.3372	0.3336	0.3300	0.3264	0.3228	0.3192	0.3156	0.3121
−0.3	0.3821	0.3783	0.3745	0.3707	0.3669	0.3632	0.3594	0.3557	0.3520	0.3483
−0.2	0.4207	0.4168	0.4129	0.4090	0.4052	0.4013	0.3974	0.3936	0.3897	0.3859
−0.1	0.4602	0.4562	0.4522	0.4483	0.4443	0.4404	0.4364	0.4325	0.4286	0.4247
−0.0	0.5000	0.4960	0.4920	0.4880	0.4840	0.4801	0.4761	0.4721	0.4681	0.4641

Standard Normal Distribution Probabilities Table

Lower-tail probability: P(Z < z)

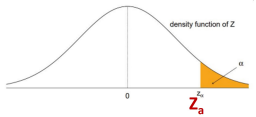


z	.00	.01	.02	.03	.04	.05	.06	.07	.08	.09
0.0	0.5000	0.5040	0.5080	0.5120	0.5160	0.5199	0.5239	0.5279	0.5319	0.5359
0.1	0.5398	0.5438	0.5478	0.5517	0.5557	0.5596	0.5636	0.5675	0.5714	0.5753
0.2	0.5793	0.5832	0.5871	0.5910	0.5948	0.5987	0.6026	0.6064	0.6103	0.6141
0.3	0.6179	0.6217	0.6255	0.6293	0.6331	0.6368	0.6406	0.6443	0.6480	0.6517
0.4	0.6554	0.6591	0.6628	0.6664	0.6700	0.6736	0.6772	0.6808	0.6844	0.6879
0.5	0.6915	0.6950	0.6985	0.7019	0.7054	0.7088	0.7123	0.7157	0.7190	0.7224
0.6	0.7257	0.7291	0.7324	0.7357	0.7389	0.7422	0.7454	0.7486	0.7517	0.7549
0.7	0.7580	0.7611	0.7642	0.7673	0.7704	0.7734	0.7764	0.7794	0.7823	0.7852
0.8	0.7881	0.7910	0.7939	0.7967	0.7995	0.8023	0.8051	0.8078	0.8106	0.8133
0.9	0.8159	0.8186	0.8212	0.8238	0.8264	0.8289	0.8315	0.8340	0.8365	0.8389
1.0	0.8413	0.8438	0.8461	0.8485	0.8508	0.8531	0.8554	0.8577	0.8599	0.8621
1.1	0.8643	0.8665	0.8686	0.8708	0.8729	0.8749	0.8770	0.8790	0.8810	0.8830
1.2	0.8849	0.8869	0.8888	0.8907	0.8925	0.8944	0.8962	0.8980	0.8997	0.9015
1.3	0.9032	0.9049	0.9066	0.9082	0.9099	0.9115	0.9131	0.9147	0.9162	0.9177
1.4	0.9192	0.9207	0.9222	0.9236	0.9251	0.9265	0.9279	0.9292	0.9306	0.9319
1.5	0.9332	0.9345	0.9357	0.9370	0.9382	0.9394	0.9406	0.9418	0.9429	0.9441
1.6	0.9452	0.9463	0.9474	0.9484	0.9495	0.9505	0.9515	0.9525	0.9535	0.9545
1.7	0.9554	0.9564	0.9573	0.9582	0.9591	0.9599	0.9608	0.9616	0.9625	0.9633
1.8	0.9641	0.9649	0.9656	0.9664	0.9671	0.9678	0.9686	0.9693	0.9699	0.9706
1.9	0.9713	0.9719	0.9726	0.9732	0.9738	0.9744	0.9750	0.9756	0.9761	0.9767
2.0	0.9772	0.9778	0.9783	0.9788	0.9793	0.9798	0.9803	0.9808	0.9812	0.9817
2.1	0.9821	0.9826	0.9830	0.9834	0.9838	0.9842	0.9846	0.9850	0.9854	0.9857
2.2	0.9861	0.9864	0.9868	0.9871	0.9875	0.9878	0.9881	0.9884	0.9887	0.9890
2.3	0.9893	0.9896	0.9898	0.9901	0.9904	0.9906	0.9909	0.9911	0.9913	0.9916
2.4	0.9918	0.9920	0.9922	0.9925	0.9927	0.9929	0.9931	0.9932	0.9934	0.9936
2.5	0.9938	0.9940	0.9941	0.9943	0.9945	0.9946	0.9948	0.9949	0.9951	0.9952
2.6	0.9953	0.9955	0.9956	0.9957	0.9959	0.9960	0.9961	0.9962	0.9963	0.9964
2.7	0.9965	0.9966	0.9967	0.9968	0.9969	0.9970	0.9971	0.9972	0.9973	0.9974
2.8	0.9974	0.9975	0.9976	0.9977	0.9977	0.9978	0.9979	0.9979	0.9980	0.9981
2.9	0.9981	0.9982	0.9982	0.9983	0.9984	0.9984	0.9985	0.9985	0.9986	0.9986
3.0	0.9987	0.9987	0.9987	0.9988	0.9988	0.9989	0.9989	0.9989	0.9990	0.9990
3.1	0.9990	0.9991	0.9991	0.9991	0.9992	0.9992	0.9992	0.9992	0.9993	0.9993
3.2	0.9993	0.9993	0.9994	0.9994	0.9994	0.9994	0.9994	0.9995	0.9995	0.9995
3.3	0.9995	0.9995	0.9995	0.9996	0.9996	0.9996	0.9996	0.9996	0.9996	0.9997
3.4	0.9997	0.9997	0.9997	0.9997	0.9997	0.9997	0.9997	0.9997	0.9997	0.9998

Let $\bar{X}$ follow Std. Normal Distribution:

- Let z_α be αth upper quantile of standard normal distribution Z. i.e. $P(Z > z_\alpha) = \alpha$.

Define z_α to be the number with an upper-tail probability of α for the standard normal distribution Z. That is, $P(Z > z_\alpha) = \alpha$.



Confidence Interval Critical Values, $z_{\alpha/2}$

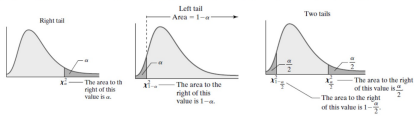
Level of Confidence	Critical Value, $z_{\alpha/2}$
0.90 or 90%	1.645
0.95 or 95%	1.96
0.98 or 98%	2.33
0.99 or 99%	2.575

Hypothesis Testing Critical Values

Level of Significance, α	Left-Tailed	Right-Tailed	Two-Tailed
0.10	− 1.28	1.28	±1.645
0.05	− 1.645	1.645	±1.96
0.01	− 2.33	2.33	±2.575

Chi Squared (χ^2) Distribution Probabilities

	Area to the Right of Critical Value									
d.f.	0.995	0.99	0.975	0.95	0.9	0.1	0.05	0.025	0.01	0.005
1	—	—	0.001	0.004	0.016	2.706	3.841	5.024	6.635	7.879
2	0.010	0.020	0.051	0.103	0.211	4.605	5.991	7.378	9.210	10.597
3	0.072	0.115	0.216	0.352	0.584	6.251	7.815	9.348	11.345	12.838
4	0.207	0.297	0.484	0.711	1.064	7.779	9.488	11.143	13.277	14.860
5	0.412	0.554	0.831	1.145	1.610	9.236	11.070	12.833	15.086	16.750
6	0.676	0.872	1.237	1.635	2.204	10.645	12.592	14.449	16.812	18.548
7	0.989	1.239	1.690	2.167	2.833	12.017	14.067	16.013	18.475	20.278
8	1.344	1.646	2.180	2.733	3.490	13.362	15.507	17.535	20.090	21.955
9	1.735	2.088	2.700	3.325	4.168	14.684	16.919	19.023	21.666	23.589
10	2.156	2.558	3.247	3.940	4.865	15.987	18.307	20.483	23.209	25.188
11	2.603	3.053	3.816	4.575	5.578	17.275	19.675	21.920	24.725	26.757
12	3.074	3.571	4.404	5.226	6.304	18.549	21.026	23.337	26.217	28.300
13	3.565	4.107	5.009	5.892	7.042	19.812	22.362	24.736	27.688	29.819
14	4.075	4.660	5.629	6.571	7.790	21.064	23.685	26.119	29.141	31.319
15	4.601	5.229	6.262	7.261	8.547	22.307	24.996	27.488	30.578	32.801
16	5.142	5.812	6.908	7.962	9.312	23.542	26.296	28.845	32.000	34.267
17	5.697	6.408	7.564	8.672	10.085	24.769	27.587	30.191	33.409	35.718
18	6.265	7.015	8.231	9.390	10.865	25.989	28.869	31.526	34.805	37.156
19	6.844	7.633	8.907	10.117	11.651	27.204	30.144	32.852	36.191	38.582
20	7.434	8.260	9.591	10.851	12.443	28.412	31.410	34.170	37.566	39.997
21	8.034	8.897	10.283	11.591	13.240	29.615	32.671	35.479	38.932	41.401
22	8.643	9.542	10.982	12.338	14.041	30.813	33.924	36.781	40.289	42.796
23	9.260	10.196	11.689	13.091	14.848	32.007	35.172	38.076	41.638	44.181
24	9.886	10.856	12.401	13.848	15.659	33.196	36.415	39.364	42.980	45.559
25	10.520	11.524	13.120	14.611	16.473	34.382	37.652	40.646	44.314	46.928
26	11.160	12.198	13.844	15.379	17.292	35.563	38.885	41.923	45.642	48.290
27	11.808	12.879	14.573	16.151	18.114	36.741	40.113	43.195	46.963	49.645
28	12.461	13.565	15.308	16.928	18.939	37.916	41.337	44.461	48.278	50.993
29	13.121	14.256	16.047	17.708	19.768	39.087	42.557	45.722	49.588	52.336
30	13.787	14.953	16.791	18.493	20.599	40.256	43.773	46.979	50.892	53.672
40	20.707	22.164	24.433	26.509	29.051	51.805	55.758	59.342	63.691	66.766
50	27.991	29.707	32.357	34.764	37.689	63.167	67.505	71.420	76.154	79.490
60	35.534	37.485	40.482	43.188	46.459	74.397	79.082	83.298	88.379	91.952
70	43.275	45.442	48.758	51.739	55.329	85.527	90.531	95.023	100.425	104.215
80	51.172	53.540	57.153	60.391	64.278	96.578	101.879	106.629	112.329	116.321
90	59.196	61.754	65.647	69.126	73.291	107.565	113.145	118.136	124.116	128.299
100	67.328	70.065	74.222	77.929	82.358	118.498	124.342	129.561	135.807	140.169



Properties of P.d.f. & C.d.f

If it is given that a random variable X is either discrete or continuous, its probability function (p.f.) $f(x)$ and cumulative distribution function (c.d.f.) $F(x)$ has the following properties.

- When the random variable X is discrete:
  - The number of values in R_X is finite or countable. Or equivalently: the number of x such that $f(x) > 0$ is finite or countable.
  - The number of possible values of $F(x)$ is finite or countable.
  - $F(x)$ is a step function; and certainly not continuous.
- When the random variable is continuous:
  - R_X is one interval or a collection of multiple intervals; or equivalently, the number of x such that $f(x) > 0$ is NOT countable.
  - The number of possible values of $F(x)$ is NOT countable.
  - $F(x)$ must be a continuous function over the entire real line.

Student t Distribution Probabilities Table

one-tail area		0.25	0.125	0.1	0.075	0.05	0.025	0.01	0.005	0.0005
two-tail area		0.5	0.25	0.2	0.15	0.1	0.05	0.02	0.01	0.001
confidence level		0.5	0.75	0.8	0.85	0.9	0.95	0.98	0.99	0.999
d.f.	1	1.000	2.414	3.078	4.165	6.314	12.706	31.821	63.657	636.619
	2	0.816	1.604	1.886	2.282	2.920	4.303	6.965	9.925	31.599
	3	0.765	1.423	1.638	1.924	2.353	3.182	4.541	5.841	12.924
	4	0.741	1.344	1.533	1.778	2.132	2.776	3.747	4.604	8.610
	5	0.727	1.301	1.476	1.699	2.015	2.571	3.365	4.032	6.869
	6	0.718	1.273	1.440	1.650	1.943	2.447	3.143	3.707	5.959
	7	0.711	1.254	1.415	1.617	1.895	2.365	2.998	3.499	5.408
	8	0.706	1.240	1.397	1.592	1.860	2.306	2.896	3.355	5.041
	9	0.703	1.230	1.383	1.574	1.833	2.262	2.821	3.250	4.781
	10	0.700	1.221	1.372	1.559	1.812	2.228	2.764	3.169	4.587
	11	0.697	1.214	1.363	1.548	1.796	2.201	2.718	3.106	4.437
	12	0.695	1.209	1.356	1.538	1.782	2.179	2.681	3.055	4.318
	13	0.694	1.204	1.350	1.530	1.771	2.160	2.650	3.012	4.221
	14	0.692	1.200	1.345	1.523	1.761	2.145	2.624	2.977	4.140
	15	0.691	1.197	1.341	1.517	1.753	2.131	2.602	2.947	4.073
	16	0.690	1.194	1.337	1.512	1.746	2.120	2.583	2.921	4.015
	17	0.689	1.191	1.333	1.508	1.740	2.110	2.567	2.898	3.965
	18	0.688	1.189	1.330	1.504	1.734	2.101	2.552	2.878	3.922
	19	0.688	1.187	1.328	1.500	1.729	2.093	2.539	2.861	3.883
	20	0.687	1.185	1.325	1.497	1.725	2.086	2.528	2.845	3.850
	21	0.686	1.183	1.323	1.494	1.721	2.080	2.518	2.831	3.819
	22	0.686	1.182	1.321	1.492	1.717	2.074	2.508	2.819	3.792
	23	0.685	1.180	1.319	1.489	1.714	2.069	2.500	2.807	3.768
	24	0.685	1.179	1.318	1.487	1.711	2.064	2.492	2.797	3.745
	25	0.684	1.198	1.316	1.485	1.708	2.060	2.485	2.787	3.725
	26	0.684	1.177	1.315	1.483	1.706	2.056	2.479	2.779	3.707
	27	0.684	1.176	1.314	1.482	1.703	2.052	2.473	2.771	3.690
	28	0.683	1.175	1.313	1.480	1.701	2.048	2.467	2.763	3.674
	29	0.683	1.174	1.311	1.479	1.699	2.045	2.462	2.756	3.659
	30	0.683	1.173	1.310	1.477	1.697	2.042	2.457	2.750	3.646
	35	0.682	1.170	1.306	1.472	1.690	2.030	2.438	2.724	3.591
	40	0.681	1.167	1.303	1.468	1.684	2.021	2.423	2.704	3.551
	45	0.680	1.165	1.301	1.465	1.679	2.014	2.412	2.690	3.520
	50	0.679	1.164	1.299	1.462	1.676	2.009	2.403	2.678	3.496
	60	0.679	1.162	1.296	1.458	1.671	2.000	2.390	2.660	3.460
	70	0.678	1.160	1.294	1.456	1.667	1.994	2.381	2.648	3.435
	80	0.678	1.159	1.292	1.453	1.664	1.990	2.374	2.639	3.416
	100	0.677	1.157	1.290	1.451	1.660	1.984	2.364	2.626	3.390
	500	0.675	1.152	1.283	1.442	1.648	1.965	2.334	2.586	3.310
	1000	0.675	1.151	1.282	1.441	1.646	1.962	2.330	2.581	3.300
	infinity	0.674	1.150	1.282	1.440	1.645	1.960	2.326	2.576	3.291