3 Joint Distributions

Discrete Joint Probability Function:

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

for $(x,y) \in R_{X,Y}$
Properties:
(1) $f_{X,Y}(x,y) \geq 0$ for any $(x,y) \in R_{X,Y}$
(2) $f_{X,Y}(x,y) = 0$ for any $(x,y) \notin R_{X,Y}$
(3)

$$\sum_{i=1}^{\infty}\sum_{j=1}^{\infty} f_{X,Y}(x_i, y_j)$$

$$= \sum_{i=1}^{\infty}\sum_{j=1}^{\infty} P(X = x_i, Y = y_j) = 1$$

Equivalently, $\sum\sum_{(x,y)\in R_{X,Y}} f(x,y) = 1$
(4) Let A be any subset of $R_{X,Y}$, then

$$P((X,Y) \in A) = \sum\sum_{(x,y)\in A} f_{(X,Y)}(x,y)$$

Continuous Joint Probability Function:

$$P((X,Y) \in D) = \iint_{(x,y)\in D} f_{X,Y}(x,y)dydx$$

for any $D \subset \mathbb{R}^2$.
More specifically,

$$P(a \leq X \leq b, c \leq Y \leq d)$$

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

Properties:
(1) $f_{X,Y}(x,y) \geq 0$ for any $(x,y) \in R_{X,Y}$
(2) $f_{X,Y}(x,y) = 0$ for any $(x,y) \notin R_{X,Y}$
(3)

$$\int_{-\infty}^{\infty}\int_{-\infty}^{\infty} f_{X,Y}(x,y)dxdy = 1$$

Equivalently,

$$\iint_{(x,y)\in R_{X,Y}} f_{X,Y}(x,y)dxdy = 1$$

Marginal Probability Distribution:
Discrete:

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

Continuous:

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

$f_X(x)$ is a **probability function**.
Conditional Distribution:

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

is defined for $f_X(x) > 0$.

$$f_{Y|X}(y|x)$$

is a probability function for y but not for x, so there is **no** requirement that

$$\int_{-\infty}^{\infty} f_{Y|X}(y|x)dx = 1$$

or

$$\sum_x f_{Y|X}(y|x) = 1$$

If $f_X(x) > 0$,

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

If $f_Y(y) > 0$,

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

For a c.r.v,

$$P(Y \leq y | X \leq x) = \int_{-\infty}^{y} f_{Y|X}(y|x)dy$$

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

(similarly for d.r.v)
Independent Random Variables:
Random variables $X_1, \dots, X_n$ are independent iff for any $x_1, \dots, x_n$,

$$f_{X_1,\dots,X_n}(x_1,\dots,x_n) = f_{x_1}(x_1)\dots f_{X_n}(x_n)$$

This applies regardless of whether X, Y are continuous or discrete.

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

$$R_{X,Y} = \{(x,y) | x \in R_X; y \in R_Y\} = R_X \times R_Y$$

If $R_{X,Y}$ is not a product space, then X, Y are not independent.
Properties:
(1)

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

and

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

(2) For arbitrary functions g_1, g_2, $g_1(X)$ and $g_2(Y)$ are independent.
(3) If $f_X(x) > 0$, then

$$f_{Y|X}(y|x) = f_Y(y)$$

If $f_Y(y) > 0$, then

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

Expectation:

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

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

Covariance:

$$\begin{aligned}
cov(X,Y) &= E[(X - E(X))(Y - E(Y))] \\
&= E[(X - \mu_X)(Y - \mu_Y)] \\
&= \sum_x \sum_y (x - \mu_X)(y - \mu_Y) f_{X,Y}(x,y) \\
&= \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} (x - \mu_X)(y - \mu_Y) f_{X,Y}(x,y)dxdy
\end{aligned}$$

Properties:
(1)

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

$$\begin{aligned}
&E[(X - \mu_X)(Y - \mu_Y)] \\
&= E[XY - Y\mu_X - X\mu_Y + \mu_X\mu_Y] \\
&= E(XY) - \mu_X E(Y) - \mu_Y E(X) + \mu_X\mu_Y \\
&= E(XY) - \mu_X\mu_Y - \mu_Y\mu_X + \mu_X\mu_Y \\
&= E(XY) - \mu_X\mu_Y
\end{aligned}$$

(2)

$$X \perp Y \Rightarrow \text{cov}(X,Y) = 0$$

$$\text{cov}(X,Y) = 0 \not\Rightarrow X \perp Y$$

If $X \perp Y$, then

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

so

$$\begin{aligned}&E(XY)\\&=\sum_i\sum_j x_i y_j f_{X,Y}(x_i,y_j)\\&=\sum_i\sum_j x_i y_j f_X(x_i) f_Y(y_j)\\&=\sum_i x_i f_X(x_i)\sum_j y_j f_Y(y_j)\\&=E(X)E(Y)\end{aligned}$$

(3)

$$\text{cov}(aX+b, cY+d) = ac\cdot\text{cov}(X,Y)$$

$$\text{cov}(X,Y)=\text{cov}(Y,X)$$

$$\text{cov}(X+b,Y)=\text{cov}(X,Y)$$

$$\text{cov}(aX,Y)=a\text{cov}(X,Y)$$

(4)

$$\begin{aligned}&V(aX+bY)\\&=a^2V(X)+b^2V(Y)+2ab\cdot\text{cov}(X,Y)\end{aligned}$$

$$V(aX)=a^2V(X)$$

$$V(X+Y)=V(X)+V(Y)+2\text{cov}(X,Y)$$

4 Distributions

Discrete Uniform Distribution:

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

$$\mu_X=E(X)=\sum_{i=1}^k x_i f_X(x_i)=\frac{1}{k}\sum_{i=1}^k x_i$$

$$\begin{aligned}&\sigma_X^2=V(X)=E(X^2)-[E(X)]^2\\&=\frac{1}{k}(\sum_{i=1}^k x_i^2)-\mu_X^2\end{aligned}$$

Bernoulli Trial: experiment with only 2 possible outcomes, success/fail

Bernoulli Random Variable (BRV): X = no. of success of Bernoulli Trial, p = prob. of success, $0\le p\le 1$.

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

$$f_X(x)=p^x(1-p)^1-x$$

for $x=0,1$

$$X\sim\text{Bernoulli}(p)$$

$$q=1-p,\qquad f_X(1)=p,\qquad f_X(0)=q$$

$$\mu_X=E(X)=p$$

$$\sigma_X^2=V(X)=p(1-p)=pq$$

Bernoulli Process: sequence of independent & identical Bernoulli trials generates a sequence of iid BRV $X_1,X_2,\dots$

Binomial Distribution: n iid Bernoulli trials. Counts the number of successes in n trials. p remains constant and trials are independent.

$$X\sim\text{Bin}(n,p)$$

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

for $x=0,1,2,\dots,n$

$$E(X)=np$$

$$V(X)=np(1-p)=npq$$

Bernoulli distribution is when $n=1$

Negative binomial distribution: no. of iid Bernoulli(p) trials needed until the k-th success occurs

$$X\sim NB(k,p)$$

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

for

$$x=k,k+1,k+2,\dots$$

$$E(X)=\frac{k}{p}$$

$$V(X)=\frac{(1-p)k}{p^2}$$

Geometric distribution: special case of negative binomial distribution where $k=1$

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

$$E(X)=\frac{1}{p}$$

$$V(X)=\frac{1-p}{p^2}$$

Poisson distribution: no. of events occurring in a fixed period of time

$$X\sim\text{Poisson}(\lambda)$$

where $\lambda>0$ is the expected no. of occurrences in the period

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

for $k=0,1,\dots$

$$E(X)=\lambda,\qquad V(X)=\lambda$$

$$\begin{aligned}&P(X\le k)=P(X=0)+\cdots+P(X=k)\\&=\sum_{i=0}^k P(X=i)\end{aligned}$$

$$\begin{aligned}&P(X>k)=1-P(X\le k)\\&=1-\sum_{i=0}^k P(X=i)\end{aligned}$$

Poisson process: no. of occurrences in *any* interval T follows Poisson(αT)

α = rate parameter = expected no. of occurrences in a fixed period

Poisson Approximation to Binomial: Let $X\sim\text{Bin}(n,p)$. As $n\to\infty$, $p\to 0$ s.t. $np=\lambda$ is constant, then approx $X\sim\text{Poisson}(np)$

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

The approximation is good when $n\ge 20$ and $p\le 0.05$, or $n\ge 100$ and $np\le 10$

Continuous Uniform Distribution: take any value over the interval (a,b)

$$X\sim U(a,b)$$

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

$$\begin{aligned}&E(X)=\int_a^b x\cdot\frac{1}{b-a}dx=\frac{1}{b-a}[\frac{b^2-a^2}{2}]\\&=\frac{a+b}{2}\end{aligned}$$

$$E(X^2)=\int_a^b x^2\cdot\frac{1}{b-a}dx=\frac{a^2+ab+b^2}{3}$$

$$V(X)=(\frac{a^2+ab+b^2}{3})-(\frac{a+b}{2})^2=\frac{(b-a)^2}{12}$$

(c.d.f.)

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

$$\begin{aligned} F_X(x) &= \int_{-\infty}^{\infty} f_X(t)dt \\ &= \int_{-\infty}^{a} 0dt + \int_{a}^{x} \frac{1}{b-a}dt = \frac{x-a}{b-a} \end{aligned}$$

Exponential Distribution:

waiting time to first sucess

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

$$X \sim \text{Exp}(\lambda)$$

for $\lambda > 0$

$$E(X) = \frac{1}{\lambda}, \qquad V(X) = \frac{1}{\lambda^2}$$

$$\begin{aligned} F_X(x) &= P(X \le x) \\ &= \int_0^x \lambda e^{-\lambda t} dt = 1 - e^{-\lambda x} \end{aligned}$$

for $x \ge 0$

$$F_X(x) = \begin{cases} 1 - e^{-\lambda x} & x \ge 0 \\ 0 & x < 0 \end{cases}$$

Alternative form:

$$f_X(x) = \begin{cases} \frac{1}{\mu} e^{-x/\mu} & x \ge 0 \\ 0 & x < 0 \end{cases}$$

$$\mu = \frac{1}{\lambda}, \qquad E(X) = \mu, \qquad V(X) = \mu^2$$

$$F_X(x) = 1 - e^{-x/\mu}$$

for $x \ge 0$

Memoryless:

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

for any $s, t \in \mathbb{Z}^+$

$$\frac{P(\{X > s+t\} \cap \{X > s\})}{P(X > s)} = \frac{P(X > s+t)}{P(X > s)}$$
$$= \frac{e^{-\lambda(s+t)}}{e^{-\lambda s}} = e^{-\lambda t} = P(X > t)$$

Normal Distribution:

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

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

$$E(X) = \mu, \qquad V(X) = \sigma^2$$

$$\int_{-\infty}^{\infty} f_X(x)dx = 1$$

$f_X(x)$ is a valid p.d.f

Two normal curves are identical in shape if they have the same σ^2, centered at diff positions if μ are different.

As σ increases, the curve flattens.

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

$$E(Z) = 0, \qquad V(Z) = 1$$

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

$$\Phi(z) = \int_{-\infty}^{z} \phi(t)dt$$

$$\begin{aligned} &P(x_1 < X < x_2) \\ &= P(\frac{x_1 - \mu}{\sigma} < \frac{x_1 - \mu}{\sigma} < \frac{x_2 - \mu}{\sigma}) \\ &= \Phi(\frac{x_2 - \mu}{\sigma}) - \Phi(\frac{x_1 - \mu}{\sigma}) \end{aligned}$$

$$P(Z \ge 0) = P(Z \le 0) = \Phi(0) = 0.5$$

For any z,

$$\Phi(z) = P(Z \le z) = P(Z \ge -z) = 1 - \Phi(-z)$$

If

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

then

$$-Z \sim N(0,1)$$

and

$$\sigma Z + \mu \sim N(\mu, \sigma^2)$$

$\alpha^{\textbf{th}}$ **(Upper) Quantile**:

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

for $0 \le \alpha \le 1$

$$z_{0.05} = 1.654, \qquad z_{0.01} = 2.326$$

$$P(Z \ge z_\alpha) = P(Z \le -z_\alpha) = \alpha$$

since $\phi(z)$ is symmetric about 0

Normal Approximation to Binomial:

Conditions: $np > 5$, $n(1-p) > 5$

Let

$$X \sim Bin(n, p)$$

$$E(X) = np, \quad V(X) = np(1-p)$$

As $n \to \infty$,

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

is approx $\sim N(0,1)$

Continuity Correction: add/subtract before standardizing

$$P(X = k) \approx P(k - \frac{1}{2} < X < k + \frac{1}{2})$$

$$P(a \le X \le b) \approx P(a - \frac{1}{2} < X < b + \frac{1}{2})$$

$$P(a < X \le b) \approx P(a + \frac{1}{2} < X < b + \frac{1}{2})$$

$$P(a \le X < b) \approx P(a - \frac{1}{2} < X < b - \frac{1}{2})$$

$$P(a < X < b) \approx P(a + \frac{1}{2} < X < b - \frac{1}{2})$$

$$\begin{aligned} P(X \le c) &= P(0 \le X \le c) \\ &\approx P(-\frac{1}{2} < X < c + \frac{1}{2}) \end{aligned}$$

$$\begin{aligned} P(X > c) &= P(c < X \le n) \\ &\approx P(c + \frac{1}{2} < X < n + \frac{1}{2}) \end{aligned}$$

5 Sample and Sampling Distribution

Population:

all possible outcomes/observations

Sample:

a subset of a population

Finite Population:

finite number of elements in the population

Infinite Population:

infinitely large number of elements. E.g. rolling a pair of dice infinitely many times

Simple Random Sample of size n:

For finite population. Every subset of n observations of the population has the same probability of being selected. $\binom{N}{n}$ ways of choosing a sample of size n

SRS (Infinite Population):

Let X be a random variable with probability function $f_X(x)$. $X_1 \ldots, X_n$ is a random sample of size n from a population with distribution $f_X(x)$. Joint probability function:

$$f_{X_1,\dots,X_n}(x_1,\dots,x_n) = f_{X_1}(x_1)\dots f_{X_n}(x_n)$$

Drawing from a finite population with replacement is the same as sampling from an infinite population.
The sample is random if the probability of being selected remains the same for each draw, and each successive draw is independent.
Statistic:
Given a random sample of n observations $(X_1,\dots,X_n)$, a function of $(X_1,\dots,X_n)$ is a statistic.
E.g. $\bar{X} = \frac{1}{n}\sum_{i=1}^{n} x_i$
The realization of a statistic $\bar{X}$ is the actual observed values $(x_1,\dots,x_n)$, $\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$
A statistic is a random variable.
Sample Variance:

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

(statistic)

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

(realization)
Sampling Distribution:
The probability distribution of a statistic is called a sampling distribution.
Mean and Variance of $\bar{X}$:
For a random sample of size n taken from an infinite population with a mean μ_X and variance σ_X^2, the sampling distribution of the sample mean $\bar{X}$ has a mean μ_X and variance $\frac{\sigma_X^2}{n}$.

$$\mu_{\bar{X}} = E(\bar{X}) = \mu_X$$

$$\sigma_{\bar{X}}^2 = V(\bar{X}) = \frac{\sigma_X^2}{n}$$

As n gets larger, the variance decreases, $\bar{X}$ becomes a better estimator of μ_X.
Standard Error:
the spread of a sampling distribution i.e. it's standard deviation, $\sigma_{\bar{X}}$. It describes how much $\bar{x}$ varies from sample to sample.
Law of Large Numbers:
$\bar{X}$ tends to be closer to μ_X as n increases.
If $X_1,\dots,X_n$ are independent random variables with mean μ and variance σ^2, then for any $\epsilon \in \mathbb{R}$,

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

as $n \to \infty$.
Central Limit Theorem:
If $\bar{X}$ is the mean of a random sample of size n taken from a population with mean μ and finite variance σ^2, then as $n \to \infty$,

$$\frac{\bar{X} - \mu}{\sigma/\sqrt{n}} \to Z \sim N(0,1), \quad \bar{X} \to N(\mu, \frac{\sigma^2}{n})$$

(For large n, $\bar{X}$ follows a normal distribution closely. If X is already normally distributed, $\bar{X}$ is also normally distributed, regardless of n)
χ^2 Distribution:
Let Z be a standard normal random variable. A random variable with the same distribution as Z^2 is called a χ^2 random variable, with one degree of freedom.
Let $Z_1,\dots,Z_n$ be n i.i.d standard normal random variables. A random variable with the same distribution as $Z_1^2 + \dots + Z_n^2$ is called a χ^2 random variable with n degrees of freedom, denoted $\chi^2(n)$.
Properties of χ^2 Distribution:
1. If

$$Y \sim \chi^2(n)$$

then $E(Y) = n$ and $V(Y) = 2n$
2. For large n, $\chi^2(n)$ is approximately $N(n, 2n)$
3. If Y_1 and Y_2 are independent χ^2 random variables with m and n degrees of freedom respectively, then $Y_1 + Y_2$ is a χ^2 random variable with $m + n$ degrees of reedom.
4. The χ^2 distribution is a family of curves, each determined by the degrees of freedom n. All the density functions have a long right tail.
$\chi^2(n;\alpha)$: For

$$Y \sim \chi^2(n)$$

$$P(Y > \chi^2(n;\alpha) = \alpha$$

Sampling Distribution of $\frac{(n-1)s^2}{\sigma^2}$:
When $X_i \sim N(\mu, \sigma^2)$ for all i,

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

has a χ^2 distribution with $n-1$ degrees of freedom
t-distribution:

$$Z \sim N(0,1), \quad U \sim \chi^2(n)$$

If Z and U are independent, then

$$T = \frac{Z}{\sqrt{U/n}}$$

follows the t-distribution with n degrees of freedom.
Properties of t-distribution:
1. t-distribution with n degrees of freedom is called *student's t-distribution*, denoted $t(n)$
2. t-distribution approaches $N(0,1)$ as $n \to \infty$. When $n \geq 30$, we can replace it with $N(0,1)$.
3. If $T \sim t(n)$, $E(T) = 0$, $V(T) = n/(n-2)$ for $n > 2$
4. The graph of the t-distribution is symmetric about the vertical axis and resembles the graph of the standard normal distribution.
$t_{n;\alpha}$:

$$T \sim t(n), \quad P(T > t_{n;\alpha}) = \alpha$$

(right tail probability)
If $X_1,\dots,X_n$ are i.i.d normal random variables with mean μ and variance σ^2, then $\frac{\bar{X}-\mu}{S/\sqrt{n}}$ follows a t-distribution with $n-1$ degrees of freedom.
F-distribution:
Suppose $U \sim \chi^2(m)$ and $V \sim \chi^2(n)$ are independent. Then the distribution for the random variables $F = \frac{U/m}{V/n}$ is called a F-distribution with (m,n) degrees of freedom.
Properties of F-distribution:
1. F-distribution with (m,n) degrees of freedom is denoted as $F(m,n)$
2. If $X \sim F(m,n)$, then

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

for $n > 2$

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

3. If $F \sim F(n,m)$, then $1/F \sim F(m,n)$
4.

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

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

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

6 Estimation

Types of Statistical Inference:

1. Estimation of population parameters, 2. Testing hypothesis of parameter values

Types of Estimation:

1. Point Estimate, 2. Interval Estimate

Estimator:

a rule (formula) on how to calculate an **estimate**. E.g. sample mean $\overline{X}$ (estimator), $\overline{x}$ (estimate)

Unbiased Estimator: $E(\hat{\Theta}) = \theta$

Definition of z_α:

$$P(Z > z_\alpha) = \alpha$$

(upper tail probability)

$$\begin{aligned}
&P(-z_{\alpha/2} \le \frac{\bar{X}-\mu}{\sigma/\sqrt{n}} \le z_{\alpha/2}) \\
&= P(\frac{|\bar{X}-\mu|}{\sigma-\sqrt{n}} \le z_{\alpha/2}) \\
&= P(|\bar{X}-\mu| \le z_{\alpha/2} \cdot \frac{\sigma}{\sqrt{n}}) \\
&= 1-\alpha
\end{aligned}$$

Maximum Error of Estimate:

$$E = z_{\alpha/2} \cdot \frac{\sigma}{\sqrt{n}}$$

Determination of Sample Size:

Find min sample size such so that max error of estimate is at most E_0 with probability $1-\alpha$

$$z_{\alpha/2} \cdot \frac{\sigma}{\sqrt{n}} \le E_0$$

$$n \ge (\frac{z_{\alpha/2}\sigma}{E_0})^2$$

Case 1: Normal distribution, σ is known, n any size, Statistic:

$$Z = \frac{\bar{X}-\mu}{\sigma/\sqrt{n}}$$

$$E = z_{\alpha/2} \cdot \frac{\sigma}{\sqrt{n}}$$

n for desired E_0 and α:

$$(\frac{z_{\alpha/2}\sigma}{E_0})^2$$

Case 2: Any distribution, σ is known, n is large, Statistic:

$$Z = \frac{\bar{X}-\mu}{\sigma/\sqrt{n}}$$

$$E = z_{\alpha/2} \cdot \frac{\sigma}{\sqrt{n}}$$

n for desired E_0 and α:

$$(\frac{z_{\alpha/2}\sigma}{E_0})^2$$

Case 3: Normal distribution, σ unknown, n is small, Statistic:

$$T = \frac{\bar{X}-\mu}{S/\sqrt{n}}$$

$$E = t_{n-1;\alpha/2} \cdot \frac{s}{\sqrt{n}}$$

n for desired E_0 and α:

$$(\frac{t_{n-1;\alpha/2} \cdot s}{E_0})^2$$

Case 4: Any distribution, σ unknown, n is large, Statistic:

$$Z = \frac{\bar{X}-\mu}{S/\sqrt{n}}$$

$$E = z_{\alpha/2} \cdot \frac{s}{\sqrt{n}}$$

n for desired E_0 and α:

$$(\frac{z_{\alpha/2} \cdot s}{E_0})^2$$

Confidence Interval:

Confidence level $= 1-\alpha =$ how confident you are that the value lies in the interval.

$P(a < \mu < b) = 1-\alpha$, (a,b) is the $(1-\alpha)$ confidence interval.

Sometimes the interval contains μ, sometimes it does not. When the interval is constructed, μ is either in it or not; there is no more randomness

Since μ is usually not known, if we repeat the procedure of collecting a sample and computing the confidence interval, about $1-\alpha$ of them will contain μ

Case 1: σ known, data normal, n any size

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

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

$$X \pm z_{\alpha/2} \cdot \frac{\sigma}{\sqrt{n}}$$

$= (X - z_{\alpha/2} \cdot \frac{\sigma}{\sqrt{n}}, X + z_{\alpha/2} \cdot \frac{\sigma}{\sqrt{n}}) = X \pm E$ is a $(1-\alpha)$ confidence interval

Case 2: σ known, any distribution, n large, confidence interval

$$\bar{x} \pm z_{\alpha/2}\frac{\sigma}{\sqrt{n}}$$

Case 3: σ unknown, data normal, n small, confidence interval

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

Case 4: σ unknown, any distribution, n large, confidence interval

$$\bar{x} \pm z_{\alpha/2}\frac{s}{\sqrt{n}}$$

($n \ge 30$ is considered large)

Experimental Design: the manner of collecting samples from two populations when trying to compare their population parameters

Independent Samples: complete randomization

Assumptions:

Matched Pair Samples: randomization between matched pairs

Case 1: Independent Samples (Known and Unequal Variances):

Assumptions:

1. A random sample of size n_1 from population 1 with mean μ_1 and variance σ_1^2.
2. A random sample of size n_2 from population 2 with mean μ_2 and variance σ_2^2.
3. Two samples are independent.
4. Population variances are known and not equal, $\sigma_1^2 \ne \sigma_2^2$.
5. Either both populations are normal, or both samples are large: $n_1 \ge 30$, $n_2 \ge 30$

$$E(\bar{X}-\bar{Y}) = \mu_1 - \mu = \delta$$

$$V(\bar{X}-\bar{Y}) = \frac{\sigma_1^2}{n_1} - \frac{\sigma_2^2}{n_2}$$

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

$P((\bar{X}-\bar{Y}) - z_{\alpha/2}\sqrt{\frac{\sigma_1^2}{n_1}+\frac{\sigma_2^2}{n_2}} < \mu_1-\mu_2 < (\bar{X}-\bar{Y}) + z_{\alpha/2}\sqrt{\frac{\sigma_1^2}{n_1}+\frac{\sigma_2^2}{n_2}}) = 1-\alpha$

Confidence Interval:

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

Case 2: Independent Samples (Large, Unknown Variances):
Assumptions:
1-3 same as case 1
4. Population variances are unknown and not equal, $\sigma_1^2 \neq \sigma_2^2$
5. Both samples are large, $n_1 \geq 30$, $n_2 \geq 30$

$$S_1^2 = \frac{1}{n_1-1}\sum_{i=1}^{n_1}(X_i-\bar{X})^2$$

$$S_2^2 = \frac{1}{n_2-1}\sum_{i=1}^{n_2}(Y_i-\bar{Y})^2$$

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

Confidence Interval:

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

Case 3: Independent Samples (Small, Equal Variances):
Assumptions:
1-3 same as case 1
4. Population variances are unknown and the same $\sigma_1^2 = \sigma_2^2 = \sigma$
5. Both samples are small $n_1 < 30$, $n_2 < 30$
6. Both populations are normally distributed

$$Z = \frac{(\bar{X}-\bar{Y})-(\mu_1-\mu_2)}{\sigma\sqrt{\frac{1}{n_1}+\frac{1}{n_2}}} \approx N(0,1)$$

Pooled Estimator S_p^2:

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

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

Confidence Interval:

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

Case 4: Large, Equal Variances:
Confidence Interval:

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

We can roughly assume equal variance if $\frac{1}{2} \leq S_1/S_2 \leq 2$ as the statistic is not overly sensitive to small differences between the variances

Paired Data:
Assumptions:
1. $(X_1,Y_1),\ldots,(X_n,Y_n)$ are matched pairs
2. X_i and Y_i are dependent
3. (X_i,Y_i) and (X_j,Y_j) are independent for $i \neq j$
4. For matched pairs, define $D_i = X_i - Y_i$, $\mu_D = \mu_1 - \mu_2$
5. $D_1,\ldots,D_n$ can be treated as random samples from a single population with mean μ_D and variance σ_D^2

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

$$D = \frac{\sum_{i=1}^n D_i}{n}$$

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

If $n < 30$ and the population is normally distributed, $T \sim t_{n-1}$.
If $n \geq 30$ then $T \sim N(0,1)$
$(1-\alpha)$ Confidence Interval:
If $n < 30$

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

If $n \geq 30$

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

7 Hypothesis Testing

Steps:
1. Set null and alternative hypothesis
2. Set level of significance
3. Identify test statistic, it's distribution and rejection criteria
4. Compute the observed test statistic based on the data
5. Conclusion

Type 1 Error: Reject H_0 when H_0 is true
Type 2 Error: Not rejecting H_0 when H_0 is false
Significance Level: Probability of making type 1 error, denoted by α
$\alpha = P(\text{Type 1 error})$
$\beta = P(\text{Type 2 error})$.
Define $1-\beta$ as the *power* of the test

Type I error is considered a serious error, so we want to control the probability of making such an error, by setting the level of significance

Step 3: The test statistic serves to quantify just how unlikely it is to observe the sample, assuming the null hypothesis is true.
Step 4 and 5: Once a sample is taken, the value of the test statistic is obtained. We check if it is within our rejection region. If it is, our sample was too improbable assuming H_0 is true, hence we reject H_0. If it is not, we did not accomplish anything. We failed to reject H_0 and hence fall back to our original assumption of H_0.
In the latter case, we did not "prove" that H_0 is true. Hence, it is prudent to use the term "fail to reject H_0" instead of "accept H_0".
***p*-value**: The probability of obtaining a test statistic at least as extreme ($\leq$ or $\geq$) than the observed sample value, given H_0 is true. I.e. observed level of significance.
If the observed statistic is z, for a two sided test, a worse result is $Z > |z|$ or $Z < -|z|$, i.e. $|Z| > |z|$. So p-value $= P(|Z| > |z|) = 2P(Z > |z|) = 2P(Z < -|z|)$
For $H_1 : \mu < \mu_0$, p-value is $P(Z < -|z|)$
For $H_1 : \mu > \mu_0$, p-value is $P(Z > |z|)$
Note that p-value is smaller than the level of significance iff the test statistic is in the rejection region. If $p < \alpha$, reject H_0. Else if $p \geq \alpha$, do not reject H_0.
It is better to report the p-value than indicate whether H_0 is rejected. For example $p = 0.01$ gives much stronger evidence than $p = 0.049$, while $p = 0.049$ and $p = 0.05$ give practically the same amount of evidence as H_0

Known Variance:
Population variance σ^2 is known
Distribution is normal OR $n \geq 30$
$H_0 : \mu = \mu_0$

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

For $H_1 : \mu \neq \mu_0$, rejection region is $z < -z_{\alpha/2}$ or $z > z_{\alpha/2}$
For $H_1 : \mu < \mu_0$, rejection region is $z < -z_{\alpha/2}$

For $H_1 : \mu > \mu_0$, rejection region is $z > z_{\alpha/2}$

Unknown Variance:

Population variance is unknown

Distribution is normal

$H_0 : \mu = \mu_0$

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

For $H_1 : \mu \neq \mu_0$, rejection region is $t < -t_{n-1;\alpha/2}$ or $t > t_{n-1;\alpha/2}$

For $H_1 : \mu < \mu_0$, rejection region is $t < -t_{n-1;\alpha/2}$

For $H_1 : \mu > \mu_0$, rejection region is $t > t_{n-1;\alpha/2}$

When $n \geq 30$, we can replace t_{n-1} with Z

Confidence Interval: The two-sided hypothesis is equivalent to finding a $(1-\alpha)$ confidence interval for μ.

The $(1-\alpha)$ confidence interval for μ is

$$(\bar{x} - t_{\alpha/2}\frac{s}{\sqrt{n}}, \bar{x} + t_{\alpha/2}\frac{s}{\sqrt{n}})$$

If it contains μ, then

$$\bar{x} - t_{\alpha/2}\frac{s}{\sqrt{n}} \leq \mu_0 \leq \bar{x} + t_{\alpha/2}\frac{s}{\sqrt{n}}$$

Then

$$-t_{\alpha/2} \leq \frac{\bar{x} - \mu_0}{s/\sqrt{n}} \leq t_{\alpha/2}$$

$t = \frac{\bar{x}-\mu_0}{s/\sqrt{n}}$ satisfies $-t_{\alpha/2} \leq t \leq t_{\alpha/2}$

The rejection region is $t < -t_{\alpha/2}$ or $t > t_{\alpha/2}$

This means when the confidence interval contains μ_0, H_0 is not rejected at level α.

When the confidence interval does not contain μ_0, then $t < -t_{\alpha/2}$ or $t > t_{\alpha/2}$, thus t falls within the rejection region and H_0 is rejected

The confidence interval can be used to perform two-sided tests

Independent Samples 1:

Population variances known

Distributions are normal OR $n_1, n_2 \geq 30$

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

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

$H_1 : \mu_1 - \mu_2 > \delta_0$, Rejection region: $z > z_\alpha$, p-value: $P(Z > |z|)$

$H_1 : \mu_1 - \mu_2 < \delta_0$, Rejection region: $z < -z_\alpha$, p-value: $P(Z < -|z|)$

$H_1 : \mu_1 - \mu_2 \neq \delta_0$, Rejection region: $z > z_{\alpha/2}$ or $z < -z_{\alpha/2}$, p-value: $2P(Z > |z|)$

Independent Samples 2:

Population variances unknown

Distributions are normal

n_1, n_2 are small < 30

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

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

Paired Data:

$D_i = X_i - Y_i$

$H_0 : \mu_D = \mu_{D_0}$

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

If $n < 30$, $T \sim t_{n-1}$

If $n \geq 30$, $T \sim N(0,1)$