## Matrices and Vectors

- Vector scalar multiplication and addition operations:

$$2\mathbf{u}_{n\times 1} + 3\mathbf{v}_{n\times 1} = 2\begin{pmatrix} u_1 \\ u_2 \\ \dots \\ u_n \end{pmatrix} + 3\begin{pmatrix} v_1 \\ v_2 \\ \dots \\ v_n \end{pmatrix} = \begin{pmatrix} 2u_1 + 3v_1 \\ 2u_2 + 3v_2 \\ \dots \\ 2u_n + 3v_n \end{pmatrix}$$

- The inner (dot or cross) product of two vectors is defined to be

$$\mathbf{u}^t\mathbf{v} = \sum_i u_i v_i = \|\mathbf{u}\| \cdot \|\mathbf{v}\| \cos(\theta),$$

  where $\|\mathbf{u}\|$ denotes the norm of a vector

$$\|\mathbf{u}\| = \sqrt{\mathbf{u}^t\mathbf{u}} = \sqrt{\sum_i u_i^2},$$

  and $\theta$ is the angle between the two vectors.

- A unit vector is a vector whose norm is 1.
- When two vectors are orthogonal, $\cos(\theta) = 0$, therefore $\mathbf{u}^t\mathbf{v} = 0$, denoted by $\mathbf{u} \perp \mathbf{v}$.
- The Euclidean distance between two vectors $\mathbf{u}$ and $\mathbf{v}$ is $\|\mathbf{u} - \mathbf{v}\|$.

- A set of vectors $\mathbf{v}_1, \dots, \mathbf{v}_m$ is said to be linear independent, if

$$c_1\mathbf{v}_1 + \cdots + c_m\mathbf{v}_m = 0 \text{ iff } c_1 = \cdots = c_m = 0.$$

  Otherwise they are linear dependent.

  In other words, if a set of vectors are linear independent, then **no one** can be expressed as a linear combination of the others; if they are linear dependent, then there **exists at least one** vector, say $\mathbf{v}_2$, which can be written as a linear combination of $\mathbf{v}_1, \mathbf{v}_3, \dots, \mathbf{v}_m$.

- If $\mathbf{A}$ is a matrix, its transpose is denoted by $\mathbf{A}^T$. The trace of a square matrix, denoted by $\mathrm{tr}(\mathbf{A})$, is the sum of its diagonal elements.
- If $\mathbf{A}$ is a square matrix, its inverse is another matrix $\mathbf{C}$ such that $\mathbf{AC} = \mathbf{I}$. We usually denote the inverse matrix by $\mathbf{A}^{-1}$. Not all square matrices have an inverse. Only *full rank* or *non-singular* square matrices has an inverse, and the corresponding inverse is *unique*.
- Suppose we have an $n \times p$ matrix $\mathbf{X}$ with $n \geq p$.
  - $\mathbf{X}_{n\times p}$ is not full rank $\Longleftrightarrow$ its columns are linear dependent.
  - $\mathbf{X}_{n\times p}$ is full rank $\Longleftrightarrow$ its columns are linear independent.

  If $\mathbf{X}$ is full rank, then $(\mathbf{X}^t\mathbf{X})$ is a $p \times p$ full-rank square matrix, so $(\mathbf{X}^t\mathbf{X})^{-1}$ exists.

## Means and Variances of Random Variables

- Covariance of $X$ and $Y$

$$\text{Cov}(X,Y) = \mathbb{E}\big[(X-\mu_X)(Y-\mu_Y)\big] = \mathbb{E}(XY) - \mu_X\mu_Y$$

  - Connection with Variance: $\text{Cov}(X,X) = \text{Var}(X)$;

  - Symmetric: $\text{Cov}(X,Y) = \text{Cov}(Y,X)$;

  - Linearity with *scale* change, but invariant with *location* change

$$\text{Cov}(aX+b,Y) = a\text{Cov}(X,Y), \quad \text{Cov}(X+Y,W) = \text{Cov}(X,W) + \text{Cov}(Y,W).$$

$$\begin{aligned} \text{Cov}(aX+bY, cX+dY) &= ac\text{Var}(X) + bd\text{Var}(Y) \\ &\quad +(ad+bc)\text{Cov}(X,Y) \\ \text{Var}(aX+bY) &= a^2\text{Var}(X) + 2ab\text{Cov}(X,Y) + b^2\text{Var}(Y). \end{aligned}$$

  - Cauchy-Schwarz Inequality

$$|\text{Cov}(X,Y)| \le \sqrt{\text{Var}(X)\text{Var}(Y)}$$

- Correlation Coefficient

$$\rho_{XY} = \frac{\sigma_{XY}}{\sigma_X\sigma_Y} = \frac{\text{Cov}(X,Y)}{\sqrt{\text{Var}(X)\text{Var}(Y)}}.$$

  - $-1 \le \rho_{XY} \le 1$, due to the CS inequality.

  - $\rho_{XY} = \pm 1$ **if and only if** $X$ and $Y$ are linear functions of each other.

  - $\rho_{XY} = 0$ **if and only if** $\text{Cov}(X,Y) = 0$, then we say $X$ and $Y$ are uncorrelated.

  - The magnitude of $\rho_{XY}$ reflects *only the linear dependence* between $X$ and $Y$. So it is possible that $Y = g(X)$ where $g$ is a one-to-one map (i.e., $X$ totally determines $Y$), but $\rho_{XY}$ is small.

  - If $X$ and $Y$ are **independent**, then $\rho_{XY} = 0$, but the reverse doesn't hold.

- Mean and Variance of Sum of Random Variables. Let $X_1, \ldots, X_n$ be $n$ random variables and $a_0, a_1, \ldots, a_n$ be $(n+1)$ constants. Define $U = a_0 + a_1 X_1 + \cdots + a_n X_n$.

$$\mathbb{E}U = \mathbb{E}\big(a_0 + a_1 X_1 + \cdots + a_n X_n\big) = a_0 + a_1 \mathbb{E}X_1 + \cdots + a_n \mathbb{E}X_n$$

$$\begin{aligned}\mathrm{Var}(U) &= \mathrm{Var}\big(a_0 + a_1 X_1 + \cdots + a_n X_n\big) \\ &= \sum_{i=1}^n a_i^2 \mathrm{Var}(X_i) + \sum_{i \neq j} a_i a_j \mathrm{Cov}(X_i, X_j) \\ &= \sum_{i=1}^n a_i^2 \mathrm{Var}(X_i) + 2\sum_{i<j} a_i a_j \mathrm{Cov}(X_i, X_j)\end{aligned}$$

If $X_i$'s are **uncorrelated**, i.e., $\mathrm{Cov}(X_i, X_j) = 0$ for all $i \neq j$, then variance of the sum is equal to the sum of variances.

- Conditional Expectations.

$$\begin{aligned}\mathbb{E}(X|Y=y) &= \sum_x x \cdot P(X=x|Y=y) = \sum_x x \cdot p_{X|Y}(x|y). \\ \mathbb{E}(X|Y=y) &= \int x \cdot f_{X|Y}(x|y)dx\end{aligned}$$

What does the symbol $\mathbb{E}(X|Y)$ mean? You can view it as a function of $Y$, i.e., $\mathbb{E}(X|Y) = g(Y)$ with its value at $Y = y$ given by

$$g(y) = \mathbb{E}(X \mid Y = y).$$

Therefore $\mathbb{E}(X|Y)$ is a random variable. We can talk about its distribution and compute its mean and variance.

Useful properties of conditional expectations and variances

$$\begin{aligned}\mathbb{E}\big[\mathbb{E}(X \mid Y)\big] &= \mathbb{E}X && \text{(Iterative rule)} \\ \mathrm{Var}(X) &= \mathbb{E}(\mathrm{Var}(X|Y)) + \mathrm{Var}(\mathbb{E}(X|Y)) && \text{(Variance decomposition)}\end{aligned}$$

- The mean of a random vector $\mathbf{Z}_{m\times 1}$ is an $m$-by-1 vector with the $i$-th element equal to $\mathbb{E}(Z_i)$.

$$\boldsymbol{\mu}_{m\times 1} = \mathbb{E}[\mathbf{Z}] = \begin{pmatrix} \mathbb{E}Z_1 \\ \cdots \\ \mathbb{E}Z_m \end{pmatrix}.$$

- The covariance of $\mathbf{Z}_{m\times 1}$ is a *symmetric* $m$-by-$m$ matrix with the $(i,j)$-th element equal to $\mathrm{Cov}(Z_i, Z_j)$.

$$\begin{aligned} \Sigma_{m\times m} = \mathrm{Cov}(\mathbf{Z}) &= \mathbb{E}\Big[(\mathbf{Z}-\boldsymbol{\mu})(\mathbf{Z}-\boldsymbol{\mu})^t\Big] \\ &= \begin{pmatrix} \mathrm{Var}(Z_1) & \cdots & \mathrm{Cov}(Z_1, Z_m) \\ \cdots & \cdots & \cdots \\ \mathrm{Cov}(Z_m, Z_1) & \cdots & \mathrm{Var}(Z_m) \end{pmatrix}. \end{aligned}$$

- Affine transformations: $\mathbf{W} = \mathbf{a}_{n\times 1} + \mathbf{B}_{n\times m}\mathbf{Z}$,

$$\mathbb{E}[\mathbf{W}] = \mathbf{a} + \mathbf{B}\boldsymbol{\mu}, \quad \mathrm{Cov}(\mathbf{W}) = \mathbf{B}\boldsymbol{\Sigma}\mathbf{B}^t.$$

Especially, for $W = v_1 Z_1 + \cdots v_m Z_m = \mathbf{v}^t\mathbf{Z}$,

$$\begin{aligned} \mathbb{E}[W] &= \mathbf{v}^t\boldsymbol{\mu} = \sum_{i=1}^m v_i\mu_i, \\ \mathrm{Var}(W) &= \mathbf{v}^t\Sigma\mathbf{v} = \sum_{i=1}^m v_i^2\mathrm{Var}(Z_i) + 2\sum_{i<j} v_i v_j \mathrm{Cov}(Z_i, Z_j). \end{aligned}$$

## The Univariate Normal Distribution

- $Z \sim \mathsf{N}(0,1)$ is called the standard normal rv. $\Phi(z)$ denotes its CDF.

$$\Phi(-z) = 1 - \Phi(z), \quad z > 0.$$

- $X \sim \mathsf{N}(\mu, \sigma^2)$, then
- $X = \mu + \sigma Z \mathsf{N}(\mu, \sigma^2)$. pdf, mgf, mean and variance.
- $aX + b \sim \mathsf{N}(a\mu + b, a^2\sigma^2)$. Specially, $\frac{1}{\sigma}(X - \mu) \sim N(0,1)$.
- Linear combinations of independent normal rv's are normal.
- $Z_i$ iid $\sim \mathsf{N}(0,1)$. What's the distribution of $W = Z_1^2 + \cdots + Z_n^2$?

$$M_W(t) = \mathbb{E}e^{tW} = \prod_{i=1}^{n} \mathbb{E}e^{tZ_i^2} = (1-2t)^{-n/2}.$$

So $W \sim \mathsf{Ga}(\frac{n}{2}, \frac{1}{2}) = \chi_n^2$.

$$\mathbb{E}(W) = n, \qquad \mathrm{Var}(W) = 2n.$$

$$f_W(w) = \frac{(\frac{1}{2})^{n/2}}{\Gamma(\frac{n}{2})} w^{\frac{n}{2}-1} e^{-\frac{w}{2}}, \quad w > 0.$$

- $Z \sim \mathsf{N}(0,1)$ and $W \sim \chi_n^2$ are independent, then $\frac{Z}{\sqrt{W/n}} \sim T_n$ (student $t$-dist with $n$ degrees of freedom).

How to derive the pdf of $T = \frac{Z}{\sqrt{W/n}}$? Consider the following bivariate transformation

$$T = \frac{Z}{\sqrt{W/n}}, \quad V = W.$$

Then

$$f_T(t) = \int f(t,v)dv = \frac{\Gamma(\frac{n+1}{2})}{\sqrt{n\pi}\Gamma(\frac{n}{2})} \left(1 + \frac{t^2}{n}\right)^{-\frac{n+1}{2}}.$$

## The Bivariate Normal Distribution

$(Y_1, Y_2) \sim \mathsf{N}_2(\boldsymbol{\mu}, \Sigma)$, where

$$\boldsymbol{\mu} = \begin{pmatrix} \mu_1 \\ \mu_2 \end{pmatrix}, \quad \Sigma = \begin{pmatrix} \sigma_1^2 & \sigma_{12} \\ \sigma_{12} & \sigma_2^2 \end{pmatrix} = \begin{pmatrix} \sigma_1^2 & \rho\sigma_1\sigma_2 \\ \rho\sigma_1\sigma_2 & \sigma_2^2 \end{pmatrix},$$

- $\mu_1$ and $\sigma_1^2$ are the mean and variance, respectively, of $Y_1$;
- $\mu_2$ and $\sigma_2^2$ are the mean and variance, respectively, of $Y_2$;
- $\sigma_{12} = \rho\sigma_1\sigma_2$ is the covariance of $Y_1$ and $Y_2$ with $\rho$ being the correlation coefficient.

Assume $\rho^2 < 1$, and the joint pdf $f(y_1, y_2)$ is given by

$$\begin{aligned} & \frac{1}{2\pi\sigma_1\sigma_2\sqrt{1-\rho^2}} \times \\ & \exp\left\{-\frac{1}{2(1-\rho^2)}\left[\left(\frac{y_1-\mu_1}{\sigma_1}\right)^2 - 2\rho\left(\frac{y_1-\mu_1}{\sigma_1}\right)\left(\frac{y_2-\mu_2}{\sigma_2}\right) + \left(\frac{y_2-\mu_2}{\sigma_2}\right)^2\right]\right\} \\ = \; & \frac{1}{2\pi|\Sigma|^{1/2}} \exp\left\{-\frac{1}{2}(\mathbf{y}-\boldsymbol{\mu})^t\Sigma^{-1}(\mathbf{y}-\boldsymbol{\mu})\right\}. \end{aligned}$$

Note that $|\Sigma| = \sigma_1^2\sigma_2^2(1-\rho^2)$ and when $\rho^2 < 1$,

$$\Sigma^{-1} = \frac{1}{\sigma_1^2\sigma_2^2(1-\rho^2)}\begin{bmatrix} \sigma_2^2 & -\rho\sigma_1\sigma_2 \\ -\rho\sigma_1\sigma_2 & \sigma_1^2 \end{bmatrix}.$$

- <u>Marginals</u>: $Y_i \sim \mathsf{N}(\mu_i, \sigma_i^2)$ where $i = 1, 2$.
- <u>Conditionals</u>:

$$\begin{aligned} Y_1 \mid Y_2 = y_2 &\sim \mathsf{N}\Big(\mu_1 + \rho\frac{\sigma_1}{\sigma_2}(y_2-\mu_2), (1-\rho^2)\sigma_1^2\Big) \\ Y_2 \mid Y_1 = y_1 &\sim \mathsf{N}\Big(\mu_2 + \rho\frac{\sigma_2}{\sigma_1}(y_1-\mu_1), (1-\rho^2)\sigma_2^2\Big) \end{aligned}$$

- <u>Linear Combinations</u>:

$$aY_1 + bY_2 \sim \mathsf{N}\big(a\mu_1 + b\mu_2, a^2\sigma_1^2 + 2ab\rho\sigma_1\sigma_2 + b^2\sigma_2^2\big)$$

- <u>Uncorrelated = Independent</u>: $\rho = 0$ implies $Y_1$ and $Y_2$ are independent.
- Note: All the statements above assume that the joint distribution of $(Y_1, Y_2)$ is normal. However,

$$Y_1 \sim \mathsf{Norm}, \quad Y_2 \sim \mathsf{Norm} \quad \text{does NOT imply} \quad (Y_1, Y_2) \sim \mathsf{Norm}.$$

So if $Y_1$ and $Y_2$ are marginally normally distributed with correlation 0, we cannot conclude that $Y_1$ and $Y_2$ are independent.

## The Multivariate Normal Distribution

- Let $\mathbf{Z} = (Z_1, \ldots, Z_n)$ where $Z_i$'s are iid $\sim \mathsf{N}(0,1)$ rv's. Then $\mathbf{Z}$ follows a multivariate normal distribution, denoted by $\mathsf{N}_n(\mathbf{0}, \mathbf{I}_n)$, with

$$f(\mathbf{z}) = \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi}} e^{-z_i^2} = \frac{1}{(2\pi)^{n/2}} \exp\Big\{ -\frac{1}{2}\sum_{i=1}^{n} z_i^2 \Big\} = \frac{1}{(2\pi)^{n/2}} \exp\Big\{ -\frac{1}{2}\mathbf{z}^t\mathbf{z} \Big\},$$

$$M(\mathbf{t}) = \mathbb{E}[\exp\{\mathbf{t}^t\mathbf{Z}\}] = \prod_{i=1}^{n} \mathbb{E} e^{t_i Z_i} = \exp\Big\{ \frac{1}{2}\|\mathbf{t}\|^2 \Big\},$$

and

$$\mathbb{E}(\mathbf{Z}) = \mathbf{0}, \qquad \mathrm{Cov}(\mathbf{Z}) = \mathbf{I}_n.$$

- We can define a general multivariate normal distribution via affine transformations.

$$\mathbf{X}_{p\times 1} = \boldsymbol{\mu}_{p\times 1} + B_{p\times n}\mathbf{Z}_{n\times 1},$$

then $\mathbf{X}$ is multivariate normal with mean $\boldsymbol{\mu}$ and covariance matrix $\Sigma = BB^t$, denoted by

$$\mathbf{X} \sim \mathsf{N}_p(\boldsymbol{\mu}, \Sigma),$$

with

$$M_{\mathbf{X}}(\mathbf{t}) = \exp\Big\{ \mathbf{t}^t\boldsymbol{\mu} + \frac{1}{2}\mathbf{t}^t\Sigma\mathbf{t} \Big\}.$$

From the expression of its mgf above, we know that a multivariate normal distribution is completely determined by its mean $\boldsymbol{\mu}$ and covariance matrix $\Sigma$.

Why don't we define $\mathbf{X}$ via its pdf?

- Recall the definition of the covariance matrix for a random vector. Any covariance matrix $\Sigma_{p\times p}$ should be symmetric and *nonnegative definite*, i.e.,

$$\mathbf{a}^t\Sigma\mathbf{a} \geq 0, \text{ for any } \mathbf{a} \in \mathbb{R}^p.$$

The $B$ matrix in $\Sigma = BB^t$ can be viewed as the square root of $\Sigma$, but there are many such square roots. So it is possible to obtain $\mathbf{X}_1$ and $\mathbf{X}_2$ from two different transformations, but they end up having the same distribution (see the example in John's notes).

Any symmetric nonnegative definite matrix has a *spectral decomposition*

$$\Sigma = \Gamma^t\Lambda\Gamma, \qquad \Lambda = \mathrm{diag}(\lambda_i)_{i=1}^p,$$

where $\lambda_1 \geq \lambda_2 \cdots \geq \lambda_p \geq 0$, and $\Gamma_{n\times n}$ is a orthonormal matrix, i.e., $\Gamma\Gamma^t = \mathbf{I}_n$.

- If $\lambda_p > 0$, then $\Sigma$ is positive definite, so $|\Sigma| > 0$ and $\Sigma^{-1} = \Gamma\Lambda^{-1}\Gamma^t$ exists. Then the pdf of $\mathsf{N}_p(\boldsymbol{\mu}, \Sigma)$ is given by

$$f(\mathbf{x}) = \frac{1}{(2\pi)^{p/2}|\Sigma|^{1/2}} \exp\Big\{ -\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu})^t\Sigma^{-1}(\mathbf{x}-\boldsymbol{\mu}) \Big\}.$$

- Properties of the multivariate normal
  - Affine transformations of multivariate normals are still normal:
    $$\mathbf{X} \sim \mathsf{N}_n(\boldsymbol{\mu}, \Sigma) \Longrightarrow A_{m\times n}\mathbf{X} + b_{m\times 1} \sim \mathsf{N}_m(A\boldsymbol{\mu} + b, A\Sigma A^t).$$
  - Marginals of a normal are still normal.
  - Conditionals of a normal are still normal.
    $$\mathbf{X}_1|\mathbf{X}_2 \sim \mathsf{N}_m\Big(\boldsymbol{\mu}_1 + \Sigma_{12}\Sigma_{22}^{-1}(\mathbf{X}_2 - \boldsymbol{\mu}_2), \Sigma_{11} - \Sigma_{12}\Sigma_{22}^{-1}\Sigma_{21}\Big)$$
  - For multivariate normals, uncorrelated = independent.

Note: All the statements above assume that the joint distribution is normal. For example,

$$\mathbf{X}_1 \sim \mathsf{Norm}, \quad \mathbf{X}_2 \sim \mathsf{Norm} \quad \text{does NOT imply} \quad (\mathbf{X}_1, \mathbf{X}_2) \sim \mathsf{Norm}.$$

So if $\mathbf{X}_1$ and $\mathbf{X}_2$ are marginally normally distributed with correlation 0, we cannot conclude that $\mathbf{X}_1$ and $\mathbf{X}_2$ are independent.

## Distributions Related to Normal

- If $\mathbf{Z} \sim \mathsf{N}_n(\mathbf{0}, \mathbf{I}_n)$, then $\|\mathbf{Z}\|^2 \sim \chi^2_n$.

  If $\mathbf{Z} \sim \mathsf{N}_n(\mathbf{0}, \sigma^2\mathbf{I}_n)$, then $\|\mathbf{Z}\|^2/\sigma^2 \sim \chi^2_n$.

  If $\mathbf{X} \sim \mathsf{N}_n(\boldsymbol{\mu}, \Sigma)$ and $\Sigma^{-1}$ exists, then $(\mathbf{X} - \boldsymbol{\mu})^t\Sigma^{-1}(\mathbf{X} - \boldsymbol{\mu}) \sim \chi^2_n$.

  If $\mathbf{X} \sim \mathsf{N}_n(\mathbf{0}, \mathbf{H})$ where $\mathbf{H}$ is a projection matrix (i.e., $\mathbf{H}$ is symmetric and idempotent), then $\mathbf{X}^t\mathbf{H}\mathbf{X} \sim \chi^2_m$ with $m = \text{trace}(\mathbf{H})$. Examples of $\mathbf{H}$:

$$\begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 0 \end{pmatrix}, \quad \begin{pmatrix} \frac{1}{2} & \frac{1}{2} & 0 \\ \frac{1}{2} & \frac{1}{2} & 0 \\ 0 & 0 & 0 \end{pmatrix}.$$

- $Z \sim N(0,1)$ and $W \sim \chi^2_n$ are independent, then

$$\frac{Z}{\sqrt{W}} \sim T_n \text{ (student } t\text{-dist)}.$$

  $W_1 \sim \chi^2_n$, $W_2 \sim \chi^2_m$ and they are independent, then

$$\frac{W_1}{W_2} \sim F_{n,m}.$$

  Chi-square and Student $t$-dist have one df (degree of freedom) and F-dist has two dfs.

- $X_1, \ldots, X_n \sim \mathsf{N}(\mu, \sigma^2)$, then $\bar{X}$ and $(X_1 - \bar{X}, \ldots, X_n - \bar{X})$ are independent,

$$\bar{X} \sim \mathsf{N}(\mu, \frac{\sigma^2}{n}), \quad \sum_{i=1}^n (X_i - \bar{X})^2 \sim \chi^2_{n-1},$$

  therefore

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

  where $S^2 = \sum_i (X_i - \bar{X})^2/(n-1)$ is the sample variance.

## Some Basic Concepts of Statistical Inference

Suppose we have a rv $X$ that has a pdf/pmf denoted by $f(x;\theta)$ or $p(x;\theta)$, where $\theta$ is called the parameter, e.g., $p$ in $\mathsf{Bern}(p)$ and $(\theta,\sigma^2)$ in $\mathsf{N}(\theta,\sigma^2)$.

Previously, we focus on problems where the value of $\theta$ is given, and we calculate various quantities related to the distribution, e.g., the mean, the variance, and various probabilities.

Now we focus on problems where $\theta$ is unknown and we try to estimate various (unknown) quantities related to this distribution, after observing a random sample $(X_1,\dots,X_n)$ from this distribution.

Some jargons:

- Parameter $\theta$
- Random sample: $(X_1,\dots,X_n)$ iid.
- Statistic $T = T(X_1,\dots,X_n)$: a function of the sample, which is also random.
- Estimator $\hat{\theta} = \hat{\theta}(X_1,\dots,X_n)$ of $\theta$ is a function of the sample, i.e., a statistic. Given an observed sample $(X_1 = x_1,\dots,X_n = x_n)$, the value of $\hat{\theta}(x_1,\dots,x_n)$ is called an estimate of $\theta$. So, an estimator is a random variable, while an estimate is a real number.
- Hypothesis testing: decide between the null hypothesis $H_0 : \theta \geq \theta_0$ and the alternative hypothesis $H_a : \theta < \theta_0$ where $\theta_0$ denotes a fixed value for the parameter $\theta$.
- Prediction

## Descriptive Statistics

- Given a set of random samples, $(x_1,\dots,x_n)$, its sample mean/variance are defined to be
$$\bar{x} = \frac{1}{n}\sum_i x_i, \quad s_X^2 = \frac{1}{n-1}\sum_i (x_i - \bar{x})^2.$$
- Given $n$ pairs of random samples, $(x_i, y_i)_{i=1}^n$, the sample covariance is defined to be
$$s_{XY} = \frac{1}{n}\sum_i (x_i - \bar{x})(y_i - \bar{y}).$$
Assuming neither $s_X^2$ nor $s_Y^2$ is zero (i.e., neither $x_i$'s nor $y_i$'s are constant), then the sample correlation is defined to be
$$r = \frac{s_{XY}}{\sqrt{s_X^2 s_Y^2}}.$$

## The Maximum Likelihood Estimator

- MLE: the estimator or estimators[1] that maximize the Likelihood function

$$L(\theta; \mathbf{x}) = f(x_1, \ldots, x_n; \theta) = \prod_{i=1}^{n} f(x_i; \theta).$$

- How to derive MLE? Write out the log likelihood function

$$\ell(\theta) = \log \Big[ \prod_{i=1}^{n} f(x_i; \theta) \Big] = \sum_{i=1}^{n} \log f(x_i; \theta).$$

  Find the maximum of $\ell(\theta)$: $\hat{\theta} = \arg\max_\theta \ell(\theta)$

  - Solve $\ell'(\theta) = 0$ (if no constraints);
  - Otherwise, need to check whether the boundary points are the maximum.

- *Invariance property of MLE.* Let $X_1, \ldots, X_n$ be a random sample with the pdf $f(x; \theta)$. Let $\eta = g(\theta)$ be a parameter of interest. Suppose $\hat{\theta}$ is the mle of $\theta$. Then $g(\hat{\theta})$ is the mle of $g(\theta)$.

## Bias, Variance, and MSE

- An estimator is called unbiased if $\mathbb{E}(\hat{\theta}) = \theta$.
- For an estimator $\hat{\theta}$ of $\theta$, define the Mean Squared Error of $\hat{\theta}$ by

$$\text{MSE}(\hat{\theta}) = \mathbb{E}(\hat{\theta} - \theta)^2 = \mathbb{E}[\hat{\theta} - \mathbb{E}(\hat{\theta})]^2 + \text{Var}(\hat{\theta}) = \text{Bias}^2 + \text{Var}$$

  Specially, if $\hat{\theta}$ is unbiased, then $\text{MSE}(\hat{\theta}) = \text{Var}(\hat{\theta})$.

  If $\theta$ is multi-dimensional, then MSE is defined as $\mathbb{E}\|\hat{\theta}-\theta\|^2 = \mathbb{E}|\hat{\theta}-\mathbb{E}(\hat{\theta})\|^2+\text{tr}\big[\text{Cov}(\hat{\theta})\big]$, where the 2nd term is the trace of the covariance matrix of $\hat{\theta}$.

[1] MLE may not be unique.

## Hypothesis Testing

- Given data $\mathbf{X} = (X_1, \ldots, X_n) \sim P$, we want to test

$$\text{(null) } H_0: \quad P \in \mathcal{P}_0, \quad vs. \quad \text{(alternative) } H_A: \quad P \in \mathcal{P}_A,$$

or equivalently if all distributions have parameters,

$$H_0: \quad \theta \in \Theta_0, \quad vs. \quad H_A: \quad P \in \Theta_A.$$

| | | | |
|---|---|---|---|
| $H_0: \theta \geq \theta_0$ | vs. | $H_a: \theta < \theta_0$ | Left-tailed |
| $H_0: \theta \leq \theta_0$ | vs. | $H_a: \theta > \theta_0$ | Right-tailed |
| $H_0: \theta = \theta_0$ | vs. | $H_a: \theta \neq \theta_0$ | Two-tailed (or two-sided) |

Our decision is usually made based on a test statistic $\delta(\mathbf{X})$: if $\delta(\mathbf{X})$ is in some region (aka, the *Rejection Region*), reject $H_0$, otherwise do not reject.

- Two types of errors

| | $H_0$ True | $H_0$ False |
|---|---|---|
| Do NOT Reject $H_0$ | ⌣̈ | Type II Error |
| Reject $H_0$ | Type I Error | ⌣̈ |

- The power of a test $\delta$ is defined to be the probability of rejection when the true parameter is $\theta$, namely,

$$\beta(\theta) = P_\theta(\delta(\mathbf{X}) \in \text{Rejection Region}).$$

The significant level of a test (usually denoted by $\alpha$) is defined to be $\max_{\theta \in \Theta_0} \beta(\theta)$; for two-tailed test, the significant level is $\beta(\theta_0)$.

We usually fix the level of a test, say 5% test or 1% test, and then decide the Rejection Region, i.e., the Rejection Region depends on $\alpha$.

- The $p$-value (observed level of significance ) is the probability, computed assuming that $H_0$ is true, of obtaining a value of the test statistic as extreme as, or more extreme than, the observed value.

Use $p$-value to perform a level $\alpha$ test: If $p$-value $< \alpha$, reject; otherwise, do not reject $H_0$.

- Connection between CIs and Hypothesis Tests. Suppose $(L(\mathbf{X}), U(\mathbf{X}))$ is a $100(1-\alpha)\%$ CI for $\theta$. Consider

$$H_0: \theta = \theta_0, \quad vs \quad H_A: \theta \neq \theta_0. \tag{1}$$

Define a test: reject $H_0$ if $\theta_0 \in (L(\mathbf{X}), U(\mathbf{X}))$. Then this is a test with significant level $\alpha$, since

$$\mathbb{P}_{\theta_0}(\text{ Reject } H_0) = \mathbb{P}_{\theta_0}\big((L(\mathbf{X}), U(\mathbf{X})) \text{ does not cover } \theta_0\big) = \alpha.$$

Similarly we can *invert* tests to obtain CIs. Let $\delta(\mathbf{X})$ to denote the test statistic and $\mathrm{RR}_{\theta_0}$ denotes the rejection region for level $\alpha$ test with null $H_0 : \theta = \theta_0$. Given data $\mathbf{X}$, this set

$$B = \{\theta_0 : \delta(\mathbf{X}) \neq \mathrm{RR}_{\theta_0}\} \tag{2}$$

is a $100(1-\alpha)\%$ CI[2] for $\theta$, since

$$\mathbb{P}_\theta(\theta \in B) = \mathbb{P}(\text{Not Reject } H_0 \mid H_0 \text{ is true}) = 1 - \alpha.$$

Here is the interpretation of this 95% CI (2): it contains $\theta$ values for which the null would not be rejected given data $\mathbf{X}$.

In most tests we face in Stat425, we are testing $H_0 : \theta = \theta_0$ and $H_a : \theta \neq \theta_0$, and we usually have an unbiased estimator of $\theta$, denoted by $\hat{\theta}$, which has standard error $sd(\hat{\theta})$. The test statistic takes the following form

$$\frac{\hat{\theta} - \theta_0}{sd(\hat{\theta})} \sim T_\nu, \text{ under } H_0.$$

So for a level $\alpha$ test,

$$\text{Reject } H_0, \text{ if } \left| \frac{\hat{\theta} - \theta_0}{sd(\hat{\theta})} \right| > t_\nu^{(\alpha/2)},$$

where $t_\nu^{(\alpha/2)}$ is the $(1-\alpha/2)$ quantile of $T_\nu$. The $(1-\alpha)$ CI for $\theta$ is

$$\hat{\theta} \pm t_\nu^{(\alpha/2)} sd(\hat{\theta}).$$

It is easy to check that, given $\hat{\theta}$ (i.e., given a data set), for any $\theta_0$ in the $(1-\alpha)$ CI, we cannot reject the hypothesis $H_0 : \theta = \theta_0$.

[2] This set, if not an interval, is the credible region.