

The Univariate Normal Distribution

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

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

- $X \sim N(\mu, \sigma^2)$, then
- $X = \mu + \sigma ZN(\mu, \sigma^2)$. pdf, mgf, mean and variance.
- $aX + b \sim N(a\mu + b, a^2\sigma^2)$. Specially, $\frac{1}{\sigma}(X - \mu) \sim N(0, 1)$.
- Linear combinations of independent normal rv's are normal.
- Z_i iid $\sim N(0, 1)$. 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 $W \sim \text{Ga}(\frac{n}{2}, \frac{1}{2}) = \chi_n^2$.

$$\mathbb{E}(W) = n, \quad \text{Var}(W) = 2n.$$

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

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

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

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

- μ_1 and σ_1^2 are the mean and variance, respectively, of Y_1 ;
- μ_2 and σ_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 ρ 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 $\rho^2 < 1$,

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

- Marginals: $Y_i \sim \mathbf{N}(\mu_i, \sigma_i^2)$ where $i = 1, 2$.
- Conditionals:

$$Y_1 \mid Y_2 = y_2 \quad \sim \quad \mathbf{N}\left(\mu_1 + \rho \frac{\sigma_1}{\sigma_2}(y_2 - \mu_2), (1 - \rho^2)\sigma_1^2\right)$$

$$Y_2 \mid Y_1 = y_1 \quad \sim \quad \mathbf{N}\left(\mu_2 + \rho \frac{\sigma_2}{\sigma_1}(y_1 - \mu_1), (1 - \rho^2)\sigma_2^2\right)$$

- Linear Combinations:

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

- Uncorrelated = Independent: $\rho = 0$ 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,

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

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

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

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

and

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

- 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

$$\mathbf{X} \sim N_p(\boldsymbol{\mu}, \Sigma),$$

with

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

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 Σ .

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 Σ , 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*

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

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

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

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

- Properties of the multivariate normal

- Affine transformations of multivariate normals are still normal:

$$\mathbf{X} \sim \mathbf{N}_n(\boldsymbol{\mu}, \Sigma) \implies A_{m \times n} \mathbf{X} + b_{m \times 1} \sim \mathbf{N}_m(A\boldsymbol{\mu} + b, A\Sigma A^t).$$

- Marginals of a normal are still normal.
- Conditionals of a normal are still normal.

$$\mathbf{X}_1 | \mathbf{X}_2 \sim \mathbf{N}_m \left(\boldsymbol{\mu}_1 + \Sigma_{12} \Sigma_{22}^{-1} (\mathbf{X}_2 - \boldsymbol{\mu}_2), \Sigma_{11} - \Sigma_{12} \Sigma_{22}^{-1} \Sigma_{21} \right)$$

- For multivariate normals, uncorrelated = independent.

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

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

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 $\mathbf{Z} \sim \mathbf{N}_n(\mathbf{0}, \mathbf{I}_n)$, then $\|\mathbf{Z}\|^2 \sim \chi_n^2$.

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

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

If $\mathbf{X} \sim \mathbf{N}_n(\mathbf{0}, \mathbf{H})$ where \mathbf{H} is a projection matrix (i.e., \mathbf{H} is symmetric and idempotent), then $\mathbf{X}^t \mathbf{H} \mathbf{X} \sim \chi_m^2$ 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 $W \sim \chi_n^2$ are independent, then

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

$W_1 \sim \chi_n^2$, $W_2 \sim \chi_m^2$ 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.

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

$$\bar{X} \sim \mathbf{N}\left(\mu, \frac{\sigma^2}{n}\right), \quad \sum_{i=1}^n (X_i - \bar{X})^2 \sim \chi_{n-1}^2,$$

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 θ is called the parameter, e.g., p in $\text{Bern}(p)$ and (θ, σ^2) in $\text{N}(\theta, \sigma^2)$.

Previously, we focus on problems where the value of θ 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 θ 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 θ
- 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 θ 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 θ . 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 θ_0 denotes a fixed value for the parameter θ .
- Prediction

The Maximum Likelihood Estimator

- MLE: the estimator or estimators¹ that maximize the Likelihood function

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

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

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

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.

¹MLE may not be unique.

- Usually the likelihood function (i.e., the joint pdf/pmf) can be factorized into two parts

$$f(x_1, \dots, x_n; \theta) = K_1[s(x_1, \dots, x_n); \theta] \cdot K_2(x_1, \dots, x_n),$$

where $K_2(\cdot)$ does not depend on θ , and $s(x_1, \dots, x_n)$ is a summary statistic (e.g., sample mean/variance) of the data. Note that the conditional distribution of (X_1, \dots, X_n) given S (if it exists) doesn't depend on θ . In other words, the statistic $S(X_1, \dots, X_n)$ has exhausted all the relevant information about the unknown parameter θ contained in this random sample, and this is why we call $S(X_1, \dots, X_n)$ a sufficient statistic for θ .

- *Invariance property of MLE.* Let X_1, \dots, 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 θ . 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 θ , 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 θ is multi-dimensional, then MSE is defined as $\mathbb{E}\|\hat{\theta} - \theta\|^2 = \mathbb{E}\|\hat{\theta} - \mathbb{E}(\hat{\theta})\|^2 + \text{tr}[\text{Cov}(\hat{\theta})]$, where the 2nd term is the trace of the covariance matrix of $\hat{\theta}$.

In Stat410/510, we have learned various methods for deriving estimators (MLE, method of moment, Bayes, etc), and known how to compare estimators by MSE. A natural question: can we find *the best* estimator(s), i.e., the estimator(s) with the smallest MSE? The answer is NO.

Suppose we want to estimate θ , given $X_1, \dots, X_n \sim \mathbf{N}(\theta, 1)$. Consider the following two estimators:

$$\hat{\theta}_1 = \bar{X}, \quad \hat{\theta}_2 = 0.$$

What does the $\text{MSE}(\theta)$, as a function of θ , look like for each of the two estimators? The class of all possible estimators is too large, so it is impossible to find one whose MSE is better than all other estimators for any value of θ . However, it is possible to find the best estimator among a subclass of estimators, e.g., among all *unbiased* estimators. The best unbiased estimators, i.e., the one with the smallest variance, is often referred to as the UMVUE (*uniform minimal variance unbiased estimator*). Later we will learn how to find the UMVUE for linear regression models.

Hypothesis Testing

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

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

or equivalently if all distributions have parameters,

$$H_0 : \theta \in \Theta_0, \quad \text{vs.} \quad H_A : \theta \in \Theta_A.$$

$H_0 : \theta \geq \theta_0$	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 δ is defined to be the probability of rejection when the true parameter is θ , namely,

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

The significant level of a test (usually denoted by α) 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 α .

- 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 α test: If $p\text{-value} < \alpha$, 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 θ . Consider

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

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

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

Similarly we can *invert* tests to obtain CIs. Let $\delta(\mathbf{X})$ to denote the test statistic and RR_{θ_0} denotes the rejection region for level α test with null $H_0 : \theta = \theta_0$. Given data \mathbf{X} , this set

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

is a $100(1 - \alpha)\%$ CI² for θ , 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 θ 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 θ , 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 α 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_ν . The $(1 - \alpha)$ CI for θ 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 θ_0 in the $(1 - \alpha)$ CI, we cannot reject the hypothesis $H_0 : \theta = \theta_0$.

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