

## Some Basic Results in Probability & Statistics

- Linear Algebra
- Probability
- Random Variables
- Common Statistical Distributions
- Statistical Estimation
- Statistical Inference about Normal Distributions

2

## Linear Algebra

- Summation and Product Operators

$$\sum_{i=1}^n x_i = x_1 + x_2 + \cdots + x_n; \quad \prod_{i=1}^n Y_i = Y_1 \cdot Y_2 \cdots Y_n$$

$$\sum_{i=1}^n \sum_{j=1}^p x_{ij} = \sum_{i=1}^n \{x_{i1} + \cdots + x_{ip}\} = x_{11} + \cdots + x_{1p} + \cdots + x_{n1} + \cdots + x_{np}$$

- Matrix: a rectangular display and organization of data. You can treat matrix as data with two subscripts, e.g.  $x_{ij}$ , the first subscript is row index and the second is the column index. We note the matrix as  $X_{n \times p} = (x_{ij})$ , and call it a n by p matrix.

3

## Matrix Operations

- Transpose: reverse the row and column index. So  $t(X)_{ij} = x_{ji}$ .
- Summation: element-wise summation
- Product: for  $X_{n \times p} = (x_{ij})$ ;  $B_{p \times m} = (\beta_{jk})$ , their product  $Y = XB = (y_{ik})$  is a n by m matrix with  $y_{ik} = \sum_{j=1}^p x_{ij}\beta_{jk}$ .
- Identity matrix  $I$ : square ( $n = p$ ), diagonal equal to 1 and 0 elsewhere.
- Inverse: the product of a matrix  $X$  and its inverse  $X^{-1}$  is identity matrix.
- Trace: for square matrix  $X_{n \times n}$ ,  $tr(X) = \sum_{i=1}^n x_{ii}$ .

4

## Some Notes about Matrix

- When doing matrix product  $XB$ , always make sure the number of columns of  $X$  and rows of  $B$  are equal.
- Matrix product has orders,  $XB$  and  $BX$  are different. For inverse matrix we have  $XX^{-1} = X^{-1}X = I$ . So only square matrix has inverse.
- Only square matrix has trace, and  $tr(XB) = tr(BX)$ .
- If  $X^{-1} = t(X)$ , we call  $X$  an orthogonal matrix.

5

## Probability

- Sample space, events (sets) A,B
- Basic rules

$$\Pr(\Omega) = 1; \quad \Pr(\Phi) = 0$$

$$\Pr(A \cup B) = \Pr(A) + \Pr(B) - \Pr(A \cap B)$$

$$\Pr(A \cap B) = \Pr(A) \Pr(B|A) = \Pr(B) \Pr(A|B)$$

- Complementary events:  $\Pr(\bar{A}) = 1 - \Pr(A)$

6

## Random Variables

- A mapping (function)  $Y$  from sample space to  $R^1$ . For continuous random variables, the distribution and density functions are defined as  $F(y) = \Pr(Y \leq y)$ ;  $f(y) = \lim_{\epsilon \rightarrow 0} \{F(y+\epsilon) - F(y)\} / \epsilon$ .
- Joint, Marginal, and Conditional Probability Distributions

$$\Pr(y_i) = \sum_j \Pr(y_i, z_j); \quad \Pr(y_i|z_j) = \Pr(y_i, z_j) / \Pr(z_j)$$

- Expectation:  $E(Y) = \sum_i y_i \Pr(y_i) = \int y f(y) dy$
- Variance:  $Var(Y) = E[Y - E(Y)]^2 = E(Y^2) - E(Y)^2$

7

## Random Variables: Contd.

- Covariance:  $\text{Cov}(Y, Z) = E[Y - E(Y)][Z - E(Z)] = E(YZ) - E(Y)E(Z)$
- Correlation:  $\rho(Y, Z) = \frac{\text{Cov}(Y, Z)}{\sqrt{\text{Var}(Y)\text{Var}(Z)}}$
- Independent Random Variables

$$\begin{aligned} Y \text{ and } Z \text{ are independent} &\Leftrightarrow \Pr(y_i, z_j) = \Pr(y_i) \Pr(z_j) \\ &\Rightarrow \text{Cov}(Y, Z) = 0 \end{aligned}$$

- Central Limit Theorem: If  $Y_1, \dots, Y_n$  are iid (independent and identically distributed) random variables with mean  $\mu$  and variance  $\sigma^2$ , then the sample mean  $\bar{Y} = \sum_{i=1}^n Y_i/n$  is approximately  $N(\mu, \sigma^2/n)$  when the sample size  $n$  is reasonably large.

8

## Common Statistical Distribution

- Normal Distribution  $N(\mu, \sigma^2)$ : density  $\frac{1}{\sqrt{2\pi\sigma^2}} \exp\{-\frac{(y-\mu)^2}{2\sigma^2}\}$ , where  $\mu$  and  $\sigma^2$  are the mean and variance for  $Y$ . We have  $E(Y) = \mu$ ,  $E(Y - \mu)^2 = \sigma^2$ ,  $E(Y - \mu)^4 = 3\sigma^4$ . More generally

$$E(Y - \mu)^{2k-1} = 0; \quad E(Y - \mu)^{2k} = \sigma^{2k} (2k - 1)!!$$

where  $(2k - 1)!! = (2k - 1) \times (2k - 3) \times \dots \times 3 \times 1$ .

- Linear functions of normal random variables are still normal.  $(Y - \mu)/\sigma$  is standard normal with mean 0 and variance 1.  $\phi(\cdot)$  and  $\Phi(\cdot)$  are commonly used to code the standard normal density and distribution functions.

9

## Common Statistical Distribution: Contd.

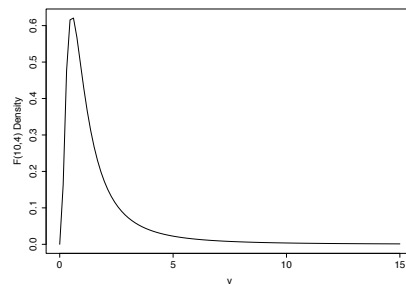
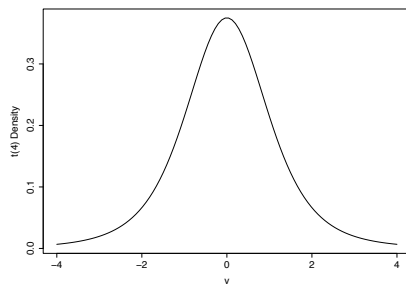
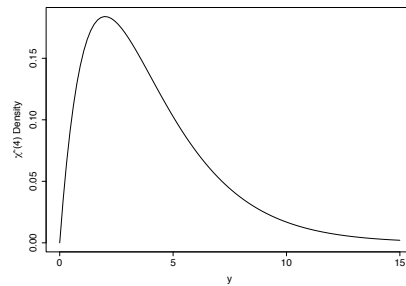
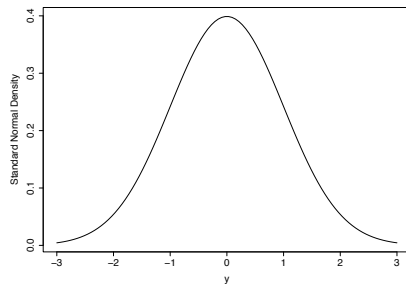
- $\chi^2$  Random Variable:  $\chi^2(n) = \sum_{i=1}^n z_i^2$ , where  $z_i$  are iid standard normal random variables and  $n$  is called the degree of freedom. We have

$$E(\chi^2(n)) = n; \quad \text{Var}(\chi^2(n)) = 2n$$

- $t$  Random Variable:  $t(n) = z/\sqrt{\chi^2(n)/n}$ , where  $z$  is standard normal and independent of  $\chi^2(n)$ .
- $F$  Random Variable:  $F(n, m) = \frac{\chi^2(n)/n}{\chi^2(m)/m}$ , where  $\chi^2(n)$  and  $\chi^2(m)$  are two independent  $chi^2$  random variables.

10

## Common Distribution Densities



11

## Statistical Estimations

- Estimator Properties: an estimator  $\hat{\theta}$  is a function of the sample observations  $(y_1, \dots, y_n)$ , which estimates some parameter  $\theta$  associated with the distribution of  $Y$ .
- Estimation Technique:
  - Maximum Likelihood Estimation
  - Least Squares Estimation
  - A lot of others .....

12

## Estimator Properties

- Unbiasedness:  $E(\hat{\theta}) = \theta$
- Consistency:  $\lim_{n \rightarrow \infty} \Pr(|\hat{\theta} - \theta| \geq \epsilon) = 0; \forall \epsilon > 0$
- Sufficiency:  $\Pr(y_1, \dots, y_n | \hat{\theta})$  doesn't depend on  $\theta$
- Minimum variance estimator :  $\text{Var}(\hat{\theta}) \leq \text{Var}(\tilde{\theta}); \forall \tilde{\theta}$

13

## Maximum Likelihood Estimators (MLE)

Maximum Likelihood is a general method of finding estimators. Suppose  $(y_1, \dots, y_n)$  are  $n$  iid samples from distribution  $f(y; \theta)$  with parameter  $\theta$ . The “probability of observing these samples” is

$$L(\theta) = \prod_{i=1}^n f(y_i; \theta);$$

which is called the likelihood function. Maximize  $L(\theta)$  with respect to  $\theta$  yields the MLE

$$\hat{\theta} = \underset{\theta}{\operatorname{argmax}} L(\theta).$$

Under very general conditions, MLE’s are consistent and sufficient.

14

## MLE for Normal Distributions

Suppose  $(y_1, \dots, y_n)$  are iid samples from normal distribution  $N(\mu, \sigma^2)$ . What’s the MLE for parameters  $\mu$  and  $\sigma^2$ ?

$$L(\mu, \sigma^2) = \prod_{i=1}^n \frac{1}{\sqrt{2\pi\sigma^2}} \exp \left\{ -\frac{(y_i - \mu)^2}{2\sigma^2} \right\}$$

Maximize  $L(\mu, \sigma^2)$  is equivalent to maximize  $\log(L(\mu, \sigma^2))$ , the “Log Likelihood”, and we can easily get the following MLE:

$$\hat{\mu} = \frac{\sum_{i=1}^n y_i}{n}; \quad \hat{\sigma}^2 = \frac{\sum_{i=1}^n (y_i - \bar{y})^2}{n}$$

15

## Least Squares Estimators (LS)

LS is another general method of finding estimators. The sample observations are assumed to be of the form  $y_i = f_i(\theta) + \epsilon_i$ ;  $i = 1, \dots, n$ , where  $f_i(\theta)$  is a known function of the parameter  $\theta$  and the  $\epsilon_i$  are random variables, usually assumed to have expectation  $E(\epsilon_i) = 0$ . LS estimators are obtained by minimizing the sum of squares

$$Q = \sum_{i=1}^n (y_i - f_i(\theta))^2$$

Here  $L_2$  distance is used; more generally  $L_q$  distance can be considered.

## Hypothesis Testing

Hypothesis testing is concerned with the state of population, which is usually characterized by some parameters, e.g. we're interested in testing the mean and variance of a normal distribution. There are several components

- Null hypothesis  $H_0$ : the postulated “default” state (value)
- Alternative hypothesis  $H_a$ : “abnormal” state
- Test statistics: the empirical information from observed data (usually some functions of data)
- Rejection rules: Type-I error  $\alpha = \Pr(\text{reject } H_0 | H_0 \text{ true})$  and Type-II error  $1 - \beta = \Pr(\text{don't reject } H_0 | H_0 \text{ false})$

## P-value

P-value for a hypothesis test is defined as **the probability that the sample outcome is more extreme than the observed one when  $H_0$  is true.**

Large P-values support  $H_0$  while small P-values support  $H_a$ . A test can be carried out by comparing the P-value with the specified type-I error  $\alpha$ . If P-value  $< \alpha$ , then  $H_0$  is rejected.

Note that the calculation of P-value depends on the rejection rules: the selection of rejection regions, which defines what is “more extreme”.

P-value is usually a function of the test statistic. It is just another test statistic and has uniform distribution when  $H_0$  is true.

## One Sample Inference about Normal Distribution

- Test  $H_0 : \sigma = \sigma_0$  vs  $H_a : \sigma \neq \sigma_0$ , under  $H_0$ ,

$$T = \frac{\sum_{i=1}^n (y_i - \bar{y})^2}{\sigma_0^2} \sim \chi^2(n-1).$$

Control Type-I error at level  $\alpha$ , rejection regions are constructed as  $(\chi^2(\alpha/2, n-1), \chi^2(1-\alpha/2, n-1))$ .

- Test  $H_0 : \mu = \mu_0$  vs  $H_a : \mu \neq \mu_0$ , under  $H_0$ ,

$$T = \sqrt{n-1} \frac{\hat{\mu} - \mu_0}{\hat{\sigma}}.$$

Control Type-I error at  $\alpha$ , choose rejection regions as  $(t(\alpha/2, n-1), t(1-\alpha/2, n-1))$ . This test is commonly known as one sample t-test.