STA261 Probability and Statistics II
Lecture Notes

Yuchen Wang

January 28, 2019

Contents

1 Normal Distribution Theory

Theorem: Sum of independent normal random variables Suppose $X_i \sim N(\mu_i, \sigma_i^2)$ for $i = 1, 2, ..., n$ and that they are independent random variables. Let $Y = (\Sigma_i a_i X_i) + b$ for some constants $\{a_i\}$ and b. Then

$$Y \sim N((\Sigma_i a_i \mu_i) + b, \Sigma_i a_i^2 \sigma_i^2)$$

Corollary: The distribution of the sample mean of normal random variables Suppose $X_i \sim N(\mu, \sigma^2)$ for $i = 1, 2, ..., n$ and that they are independent random variables, If $\bar{X} = (X_1 + ... + X_n)/n$, then $\bar{X} \sim N(\mu, \sigma^2/n)$

Theorem: The covariance of sums of normal random variables Suppose $X_i \sim N(\mu_i, \sigma_i^2)$ for $i = 1, 2, ..., n$ and also that the $\{X_i\}$are independent. Let $U = \Sigma_{i=1}^n a_i X_i$ and $V = \Sigma_{i=1}^n b_i X_i$ for some constants $\{a_1\}$ and $\{b_i\}$. Then $Cov(U, V) = \Sigma_i a_i b_i \sigma^2$. Furthermore, $Cov(U, V) = 0$ if and only if U and V are independent.

2 Expectation and Covariance

2.1 Expectation -Discrete case

Definition of expectation Let X be a discrete random variable, taking on distince values $x_1, x_2, ...$, with $p_i = P(X = x_i)$. Then the *expected value* (or *mean* or *mean value*) of X, written E(X) (or μ_x), is defined by

$$E(X) = \Sigma_i x_i p_i$$

Theorem: expectation involving nested functions

1. Let X be a discrete random variable, and let $g : \mathbb{R} \to \mathbb{R}$ be some function such that the expectation of the random variable $g(X)$ exists. Then
$$E(g(X)) = \Sigma_x g(x) P(X = x)$$

2. Let X and Y be discrete random variables, and let $h : \mathbb{R}^2 \to \mathbb{R}$ be some function such that the expectation of the random variable $h(X, Y)$ exists. Then
$$E(h(X, Y)) = \Sigma_{x,y} h(x, y) P(X = x, Y = y)$$

Theorem: Linearity of expected values Let X and Y be discrete random variables, let a and b be real numbers, and put $Z = aX + bY$. Then

$$E(Z) = aE(X) + bE(Y)$$

Theorem: Expectation of product of independent r.v Let X and Y be discrete random variables that are independent. Then

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

Monotonicity Let X and Y be discrete random variables, and suppose that $X \leq Y$ (Remember that this means $X(s) \leq Y(s)$ for all $s \in S$) Then $E(X) \leq E(Y)$.

2.2 Expectation - Continuous case

Definition of expectation Let X be an absolutely continuous random variable, with density function f_X. Then the *expected value* of X is given by

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

Theorem: expectation involving nested functions

1. Let X be a an absolutely continuous random variable with density function f_X, and let $g : \mathbb{R} \to \mathbb{R}$ be some function such that the expectation of the random variable $g(X)$ exists. Then

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

2. Let X and Y be discrete random variables, and let $h : \mathbb{R}^2 \to \mathbb{R}$ be some function such that the expectation of the random variable $h(X, Y)$ exists. Then

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

Theorem: Linearity of expected values Let X and Y be jointly absolutely continuous random variables, let a and b be real numbers. Then

$$E(aX + bY) = aE(X) + bE(Y)$$

Monotonicity Let X and Y be jointly continuous random variables, and suppose that $X \leq Y$ (Remember that this means $X(s) \leq Y(s)$ for all $s \in S$) Then $E(X) \leq E(Y)$.

2.3 Variance, Covariance and Correlation

Definition of variance The *variance* of a random variable X is the quantity

$$\sigma_x^2 = Var(X) = E((X - \mu_X)^2)$$

where σ_X is the *standard deviation* of X.

Theorem Let X be any r.v. with $\mu_X = E(X)$ and variance Var(X). Then the following hold true:

1. $Var(X) \geq 0$
2. If a and b are real numbers, $Var(aX + b) = a^2 Var(X)$
3. $Var(X) = E(X^2) - (\mu_X)^2 = E(X^2) - E(X)^2$
4. $Var(X) \leq E(X^2)$

Definition of covariance

$$Cov(X, Y) = E((X - \mu_X)(Y - \mu_Y))$$

Theorem: Linearity of covariance Let X, Y and z be three r.v.s. Let a and b be real numbers. Then

$$Cov(aX + bY.Z) = aCov(X, Z) + bCov(Y, Z)$$

Theorem Let X and Y be r.v.s. Then

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

Theorem If X and Y are independent, then

$$Cov(X, Y) = 0$$

.

Theorem

1. For any r.v.s X and Y,
$$Var(X+Y) = Var(X) + Var(Y) + 2Cov(X,Y)$$
2. More generally, for any r.v.s $X_1, ..., X_n$,
$$Var(\Sigma_i X_i) = \Sigma_i Var(X_i) + 2\Sigma_{i<j} Cov(X_i, X_j)$$

Corollary

1. If X and Y are independent, then $Var(X+Y) = Var(X) + Var(Y)$
2. If $X_1, ...X_n$ are independent, then $Var(\Sigma_{i=1}^n X_i = \Sigma_{i=1}^n Var(X_i)$

Definition The *correlation* of two r.v.s X and Y is given by
$$Corr(X,Y) = \frac{Cov(X,Y)}{Sd(X)Sd(Y)}$$
provided $Var(X) < \infty$ and $Var(Y) < \infty$

3 Types of Inferences

Estimation:

1. Point estimation: Based on the sample observations, calculating a particular value as an estimate of the parameter θ
2. Interval estimation: Calculating a range of values that is likely to contain the parameter θ

Hypothesis testing Based on the sample, assess whether a hypothetical value θ_0 is a plausible value of the parameter θ or not.

4 Different Types of Estimation

4.1 Method of Moments Estimation

Let $X_1, X_2, ..., X_n$ are independently and identically distributed (i.i.d.) random variables.
Let the k^{th} population moment be
$$\mu_k = E[X^k]$$

k^{th} sample moment based on sample

$$\hat{\mu} = \frac{1}{n}\sum_{i=1}^{n} X_i^k$$

We use $\hat{\mu}_k$ as an estimator of μ_k
In other words, we use the sample moments as estimators of the population moments.

4.2 Maximum Likelihood Estimation

Definition of Likelihood Function Suppose $X_1, X_2, ..., X_n$ has a joint density or mass function $f(x_1, x_2, ..., x_n|\theta)$
We observe sample, $X_1 = x_1, X_2 = x_2, ..., X_n = x_n$
Given the sample, the likelihood function of θ, noted as $L(\theta|x_1, x_2, ..., x_n)$, is defined as

$$L(\theta|x_1, x_2, ..., x_n) = f(x_1, x_2, ..., x_n|\theta)$$

Often written as $L(\theta)$, is a function of θ.
If X follows a discrete distribution, it gives the probability of observing the sample as a function of the parameter θ
If $X_1, X_2, ..., X_n$ are i.i.d. then their joint density is the product of marginal densities, $f_\theta(x)$
Hence, in i.i.d. case we write

$$L(\theta) = \Pi_{i=1}^{n} f_\theta(x_i)$$

Comments

1. $L(\theta)$ is NOT a pdf or pmf of θ
2. Likelihood introduces a belief ordering on parameter space, Ω
3. For $\theta_1, \theta_2 \in \Omega$, we believe in θ_1 as the true value of θ over θ_2 whenever $L(\theta_1) > L(\theta_2)$
4. Which means, the data is more likely to come from f_{θ_1} than f_{θ_2}
5. The value $L(\theta)$ is very small for every value of θ
6. So often, we are interested in the likelihood ratios:

$$\frac{L(\theta_1)}{L(\theta_2)}$$

Maximum Likelihood Estimation

1. Let's say we are interested in a point estimate of θ
2. A sensible choice will be to pick $\hat{\theta}$ that maximizes $L(\theta)$
3. So $\hat{\theta} satisfies L(\hat{\theta} \geq L(\theta)$ for all $\theta \in \Omega$
4. $\hat{\theta}$ is called the maximum likelihood estimate (MLE) of θ

Computation of the MLE

1. Define, log-likelihood function, $l(\theta) = \ln L(\theta)$
2. $\ln(x)$ is a 1-1 increasing function of $x > 0 \implies L(\hat{\theta}) \geq L(\theta)$ for $\theta \in \Omega$ iff $l(\hat{\theta}) \geq l(\theta)$
3. In other words, if $L(\theta)$ is maximized at $\hat{\theta}$ then $l(\theta)$ will also be maximized at $\hat{\theta}$
4. Therefore,
$$l(\theta) = \ln\left(\Pi_{i=1}^{n} f_\theta(x_i)\right) = \sum_{i=1}^{n} \ln f_\theta(x_i)$$
5. The obvious benefit: It's much easier to differentiate a sum than a product
6. Solve the equation, $\frac{\partial l(\theta)}{\partial \theta} = 0$ for θ
7. Say, $\hat{\theta}$ is the solution. But it's still not the MLE
8. Need to check whether or not
$$\frac{\partial^2 l(\theta)}{\partial \theta^2}|_{\theta=\hat{\theta}} < 0$$

Properties of MLE

1. MLE is not unique
2. MLE may not exists
3. The likelihood may not always be differentiable.

5 Sampling Distribution of an Estimator

1. Recall: An Estimator (T) is a random variable (infinite number of sample means)

2. If we repeat the sampling procedure and keep calculating T for each set of sample and finally draw a density histogram based on the T values we get the sampling distribution of T

3. **Standard error:** Standard deviation of an estimator is called the standard error (SE)

Definition of Mean Squared Error Let $\psi(\theta)$ be any real valued function of θ, suppose T is an estimator of $\psi(\theta)$

$$MSE_\theta(T) = E_\theta[(T - \psi(\theta))^2]$$

Corollary

$$MSE_\theta(T) = Var_\theta(T) + (E_\theta(T) - \psi(\theta))^2$$

proof:

$$\begin{aligned}
MST(T) &= E[(T - \psi(\theta))^2] \\
&= E[(T - E(T) + E(T) - \psi(\theta))^2] \\
&= E[(T - E(T))^2 + (E(T) - \psi(\theta))^2 + 2(T - E(T))(E(T) - \psi(\theta))] \\
&= E[(T - E(T))^2] + (E(T) - \psi(\theta))^2 + 2E[T - E(T)](E(T) - \psi(\theta)) \\
&= E[(T - E(T))^2] + (E(T) - \psi(\theta))^2 \\
&\qquad\qquad \text{(Since } E[T - E(T)] = E(T) - E(T) = 0) \\
&= Var(T) + (E(T) - \psi(\theta))^2 \\
&= Var(T) + Bias^2(T)
\end{aligned}$$

∎

Bias The bias of an estimator T of $\psi(\theta)$ is given by

$$E_\theta(T) - \psi(\theta)$$

Unbiased estimator: When the bias of an estimator is zero, it's called unbiased

Remark

1. For unbiased estimators,

$$MSE_\theta(T) = Var_\theta(T)$$

2. If all the other properties are similar, then an unbiased estimator is preferred over a biased estimator.

3. In practice, often an biased estimator with lower variance is preferred over an unbiased estimator with really high variance. **We minimize MSE**.