STA261 Probability and Statistics II
Lecture Notes

Yuchen Wang

March 27, 2019

Contents

1 Converge in distribution

2 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.

3 Expectation and Covariance

3.1 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 $E(X) \leq E(Y)$.

3.2 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 $E(X) \leq E(Y)$.

3.3 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 $Var(Y) < \infty$

4 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.

5 Different Types of Estimation

5.1 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.

5.2 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 exist
3. The likelihood may not always be differentiable.

6 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**.

7 Population Variance (σ^2)

Definition $\sigma^2 = E[(X-\mu)^2]$ where $\mu = E[X]$.
If we have equally likely N data points in our population, this is equivalent of

$$\sigma^2 = \frac{1}{N}\sum_{i=1}^{N}(X_i - \mu)^2$$

In words: It's the average squared difference of each of the data points (X_i) from the mean (μ)

Estimate σ^2 based on a sample of size n When we are estimating based on the sample of size n, we replace μ by $\bar{X}$, so the numerator is $\sum_{i=1}^{n}(X_i - \bar{X})^2$. We can divide it by $\underline{\text{both } n \text{ or } n-1}$. The latter one is unbiased!
The fraction, $\frac{n-1}{n} \to 1$ as $n \to \infty$. So for large n, both estimator will produce similar estimate. In statistical literature, whenever we say *sample variance* we refer to the *unbiased* one. Hence, from now on,

Definition of sample variance

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

8 Sampling distribution of S^2(under Normal Distribution

Theorem Suppose $X_1, X_2, ..., X_n \sim N(\mu, \sigma^2)$ iid, $\bar{X} = \frac{1}{n}\sum_{i=1}^{n} X_i$ and $S^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \bar{X})^2$.
Then

1. $\bar{X}$ and S^2 are independent, and
2. $\frac{(n-1)S^2}{\sigma^2} \sim \chi^2_{(df=n-1)}$

proof:

Part 1 Let

$$\begin{aligned} U &= \bar{X} = \frac{1}{n}X_1 + ... + \frac{1}{n}X_n \\ V &= X_1 - \bar{X} = X_1 - (\frac{1}{n}X_1 + ... + \frac{1}{n}X_n) \\ &= (1 - \frac{1}{n})X_1 - (\frac{1}{n}X_2 + ... + \frac{1}{n}X_n) \end{aligned}$$

$$\begin{aligned} Cov(\bar{X}, X_1 - \bar{X}) &= Cov(\bar{X}, X_1) - Cov(\bar{X}, \bar{X}) \\ &= Cov(\frac{1}{n}X_1 + ... + \frac{1}{n}X_n, X_1) - \frac{\sigma^2}{n} \\ &= \frac{1}{n}Cov(X_1, X_1) - \frac{\sigma^2}{n} \\ &= \frac{\sigma^2}{n} - \frac{\sigma^2}{n} \\ &= 0 \end{aligned}$$

Hence by E&R theorem, U and V are independent. Similarly, we can show $\bar{X}$ is independent to each $X_i - \bar{X}$ for $i = 1, ..., n$
Therefore, $\bar{X}$ is independent to $\sum_{i=1}^{n}(X_i - \bar{X})^2$
Therefore, $\bar{X}$ is independent to $\frac{\sum_{i=1}^{n}(X_i - \bar{X})^2}{n-1} = S^2$

Part 2

$$\begin{aligned}
\sum_i (X_i-\mu)^2 &= \sum_i (X_i-\bar{X}+\bar{X}-\mu)^2 \\
&= \sum_i (X_i-\bar{X})^2 + \sum_i (\bar{X}-\mu)^2 + 2\sum_i (X_i-\bar{X})(\bar{X}-\mu) \\
&= \sum_i (X_i-\bar{X})^2 + \sum_i (\bar{X}-\mu)^2 + 2(\bar{X}-\mu)\sum_i (X_i-\bar{X}) \\
&= \sum_i (X_i-\bar{X})^2 + \sum_i (\bar{X}-\mu)^2 + 2(\bar{X}-\mu)(\sum_i X_i - n\bar{X}) \\
&= \sum_i (X_i-\bar{X})^2 + \sum_i (\bar{X}-\mu)^2 + 2(\bar{X}-\mu)(n\bar{X} - n\bar{X}) \\
&= \sum_i (X_i-\bar{X})^2 + n(\bar{X}-\mu)^2 \\
\implies \sum_i (X_i-\bar{X})^2 &= \sum_i (X_i-\mu)^2 - n(\bar{X}-\mu)^2
\end{aligned}$$

$$\begin{aligned}
\frac{\sum_{i=1}^n (X_i-\mu)^2}{\sigma^2} &= \frac{\sum_i (X_i-\bar{X})^2}{\sigma^2} + \frac{n(\bar{X}-\mu)^2}{\sigma^2} \\
\implies \sum_i (\frac{X_i-\mu}{\sigma})^2 &= \frac{\sum_i (X_i-\bar{X})^2}{\sigma^2} + (\frac{\bar{X}-\mu}{\sigma/\sqrt{n}})^2 \\
&= \frac{(n-1)S^2}{\sigma^2} + (\frac{\bar{X}-\mu}{\sigma/\sqrt{n}})^2 \\
\implies \chi^2_{(n)} &= \frac{(n-1)S^2}{\sigma^2} + \chi^2_{(1)} \\
\implies MGF(\chi^2_{(n)}) &= MGF(\frac{(n-1)S^2}{\sigma^2} + \chi^2_{(1)}) \\
&= MGF(\frac{(n-1)S^2}{\sigma^2}) * MGF(\chi^2_{(1)}) \\
\implies MGF(\frac{(n-1)S^2}{\sigma^2}) &= \frac{MGF(\chi^2_{(n)})}{MGF(\chi^2_{(1)})} \\
&= \frac{(1-2t)^{-\frac{n}{2}}}{(1-2t)^{-\frac{1}{2}}} \\
&= (1-2t)^{-\frac{n-1}{2}}
\end{aligned}$$

which is the MGF of $\chi^2_{(n-1)}$ ■

E&R theorem 4.6.2 $X_1, X_2, ..., X_n \sim N(\mu, \sigma^2) i.i.d.$,, U and V are two different linear combinations of the X_i's, then
$Cov(U, V) = 0 \iff$ U and V are independent.

Note In general, zero covariance doesn't imply independent
Example: $X \sim N(0, 1), Y = X^2$, clearly X and Y are dependent, but

$$\begin{aligned} Cov(X, Y) &= E[XY] - E[X]E[Y] \\ &= E[X^3] - 0 \cdot E[Y] \\ &= E[X^3] \\ &= \int x^3 f(x) dx \\ &= 0 \qquad\qquad \text{(since } x^3 f(x) \text{ is centro-symmetric)} \end{aligned}$$

Unbiasedness of S^2 using the Chi-sq distribution

$$E[\frac{(n-1)S^2}{\sigma^2}] = n - 1$$

$$\implies E[S^2] = \sigma^2$$

This proves S^2 is an unbiased estimator for σ^2 under Normal distribution
There's another way to prove it under any arbitrary distribution with the assumption that X_i's are i.i.d. and μ, σ^2 exists.

9 Some relationships among distributions

1. $\frac{\bar{X}-\mu}{S/\sqrt{n}} \sim t_{(n-1)}$

2. $\frac{\chi^2_{(m)}}{m} \xrightarrow{P} 1$

10 Difference between sample variance and variance of sample mean

variance of sample mean: Expectation of squared difference of sample mean from the true mean

sample variance: average squared difference of each data points in the sample from the sample mean

11 Consistent Estimator

Definition Let T_n be an estimator of parameter θ, T_n is said to be consistent (in probability) if

$$T_n \xrightarrow{P} \theta$$

In words, T_n converges to θ in probability.

Note If $T_n \xrightarrow{a.s.} \theta$ then T_n is called consistent (almost surely). In this course we will only talk about consistent (in probability)

Proving consistency using LLN LLN tells us, $\bar{X} = \frac{1}{n}\sum X_i \xrightarrow{P} E[X_i]$ for any distribution. Immediately that tells us:

1. If $X_i \overset{iid}{\sim} N(\mu, \sigma^2)$ then $\bar{X}$ is a consistent estimator of μ
2. If $X_i \overset{iid}{\sim} Poisson(\lambda)$ then $\bar{X}$ is a consistent estimator of λ
3. And we can say this for few other known distributions

Goal: prove consistency when the estimator is not simply $\bar{X}$ Still use LLN but with the help of a well known Lemma and the continuous mapping theorem

Slutsky's Lemma We have two different sequence X_n and Y_n
If $X_n \xrightarrow{P} X$ and $Y_n \xrightarrow{P} Y$, then $X_n + Y_n \xrightarrow{P} X + Y$
If $X_n \xrightarrow{P} X$ and $Y_n \xrightarrow{P} Y$, then $X_n Y_n \xrightarrow{P} XY$

Continuous mapping theorem Let $X_n \xrightarrow{P} X$ and $g()$ be a continuous function, then $g(X_n) \xrightarrow{P} g(X)$

Proving S^2 is a consistent estimator of σ^2 ...

MSE consistent An estimator T_n is called MSE consistent if

$$MSE(T_n) \to 0 \text{ as } n \to \infty$$

Example: for $N(\mu, \sigma^2)$ $MSE(\bar{X}) = \sigma^2/n \to 0$ as $n \to \infty$ Therefore $\bar{X}$ is a MSE consistent estimator of μ
In naive words, after you have calculated the MSE of an estimator, just check if it goes to zero for large n

Note MSE consistent $\implies$ consistent (in probability)

12 Efficient Estimator

Definition of Efficiency Let T_1 and T_2 be two different estimators of θ, Efficiency of T_1 relative to T_2 is defined as

$$eff(T_1, T_2) = \frac{var[T_2]}{var[T_1]}$$

Remark

1. $eff(T_1, T_2) > 1 \implies T_1$ has smaller variance $\implies T_1$ is more efficient
2. This comparison is meaningful when T_1 and T_2 are both unbiased or both have the same bias.

Lower bound of the variance of an unbiased estimator This famous inequality provides a lower bound for the variance of all the unbiased estimators. In other words it gives a lower bound of the MSE (since Bias = 0). The estimator whose variance achieves this lower bound is said to be efficient. Before we state the inequality let's define few terms...

Score function, $S(\theta)$ The derivative of the log-likelihood

$$S(\theta) = \frac{\partial l(\theta)}{\partial \theta}$$

For the random variable X, $S(\theta|X = x) = \frac{\partial}{\partial\theta} \ln f_\theta(x)$. For an observed i.i.d sample, it's written as $S(\theta|x_1, x_2, \ldots, x_n)$ with

$$S(\theta|x_1, x_2, \ldots, x_n) = \frac{\partial}{\partial\theta} \sum_i \ln f_\theta(x_i) = \sum_i \frac{\partial}{\partial\theta} \ln f_\theta(x_i) = \sum_i S(\theta|x_i)$$

Fisher Information, $I(\theta)$ The function

$$I(\theta) = var_\theta[S(\theta|X)]$$

It's the amount of information that each observable random variable X contains about θ.
Information of a sample of size $n = var[S(\theta|x_1, x_2, \ldots, x_n)] = nI(\theta)$

A plot showing the randomness of $S(\theta)$ The likelihood function looks different for different data!

One important property of $S(\theta)$ Under some assumptions,

$$E[S(\theta|X = x)] = 0$$

Which implies

$$E[S(\theta|x_1, x_2, \ldots, x_n)] = \sum_i E[S(\theta|x_i)] = 0$$

Cramer-Rao Inequality Let $X_1, X_2, \ldots, X_n$ be i.i.d. with density $f_\theta(x)$, $T(X_1, X_2, \ldots, X_n)$ be an unbiased estimator of θ, Then under some assumptions on $f_\theta(x)$,

$$var[T] \geq \frac{1}{nI(\theta)}$$

$\frac{1}{nI(\theta)}$ is also known as the Cramer-Rao lower bound (CRLB)

Proof of Cramer-Rao Inequality ...

Definition of sufficient statistic A statistic $T(X_1, X_2, \ldots, X_n)$ is said to be sufficient for θ if the conditional distribution of $X_1, X_2, \ldots, X_n$, given $T = t$, does not depend on θ