

S1211Q Introduction to Statistics

Lecture 13

Wei Wang

July 23, 2012

CLT

- Theorem:

The Central Limit Theorem (CLT)

Let X_1, X_2, \dots, X_n , be an i.i.d. sequence from a distribution with mean μ and variance σ^2 . Then if n is sufficiently large, the sample mean \bar{X} has approximately a normal distribution with $\mu_{\bar{X}} = \mu$ and $\sigma_{\bar{X}}^2 = \sigma^2/n$; And the sample total has approximately a normal distribution with $\mu_T = n\mu$, $\sigma_T^2 = n\sigma^2$. The larger the value of n , the better the approximation.

- Rule of Thumb: if $n > 30$, the CLT can be used.

Distribution of a Linear Combination

- ▶ Sample mean is a particular case of linear combinations.
- ▶ The expectation and variance of a general linear combination

$$a_1X_1 + a_2X_2 + \dots + a_nX_n$$

is given by the following result.

A key result ***

Let X_1, X_2, \dots, X_n , have mean values $\mu_1, \mu_2, \dots, \mu_n$, respectively, and variances $\sigma_1^2, \sigma_2^2, \dots, \sigma_n^2$, respectively.

- Whether or not the X_i 's are independent,

$$\begin{aligned} E(a_1X_1 + a_2X_2 + \dots + a_nX_n) &= a_1E(X_1) + a_2E(X_2) + \dots + a_nE(X_n) \\ &= a_1\mu_1 + a_2\mu_2 + \dots + a_n\mu_n \end{aligned}$$

- For any X_1, X_2, \dots, X_n ,

$$\text{Var}(a_1X_1 + a_2X_2 + \dots + a_nX_n) = \sum_{i=1}^n \sum_{j=1}^n a_i a_j \text{Cov}(X_i, X_j)$$

If they are independent, then

$$\begin{aligned} &\text{Var}(a_1X_1 + a_2X_2 + \dots + a_nX_n) \\ &= a_1^2 \text{Var}(X_1) + a_2^2 \text{Var}(X_2) + \dots + a_n^2 \text{Var}(X_n) \\ &= a_1^2 \sigma_1^2 + a_2^2 \sigma_2^2 + \dots + a_n^2 \sigma_n^2 \end{aligned}$$

Special Cases

- $E(X+Y) = E(X) + E(Y)$;
- $E(X-Y) = E(X) - E(Y)$;
- $\text{Var}(X+Y) = \text{Var}(X) + \text{Var}(Y) + 2\text{Cov}(X, Y)$
- $\text{Var}(X-Y) = \text{Var}(X) + \text{Var}(Y) - 2\text{Cov}(X, Y)$
- If X and Y are independent, then $\text{Cov}(X, Y) = 0$, and
 $\text{Var}(X+Y) = \text{Var}(X) + \text{Var}(Y)$
 $\text{Var}(X - Y) = \text{Var}(X) + \text{Var}(Y)$

Example

Ex. Show that if $X \sim \text{Bin}(n, p)$, then $E(X) = np$, and $\text{Var}(X) = np(1 - p)$.

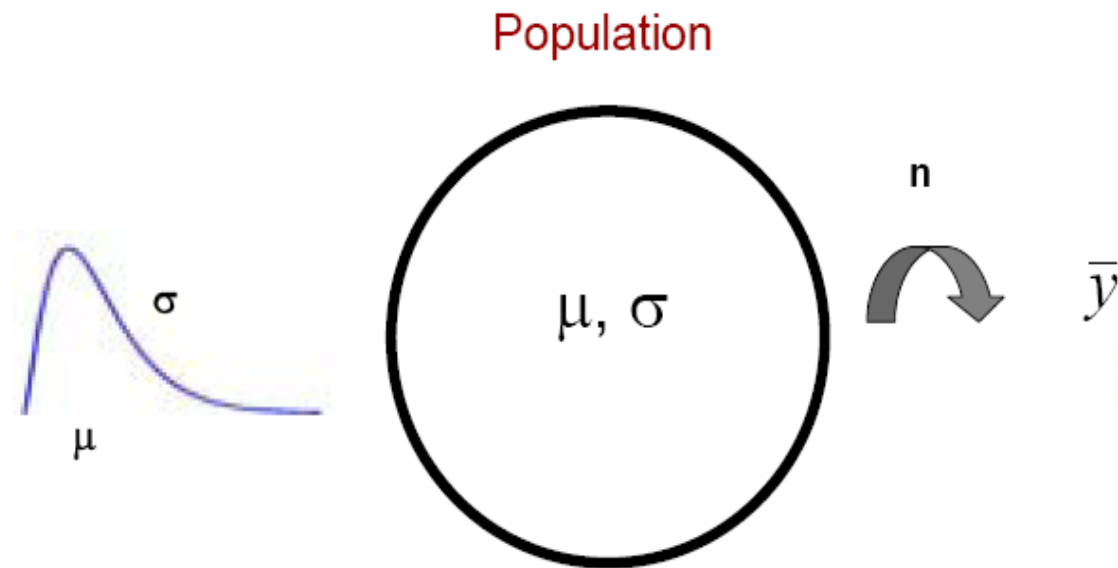
Ex. Show that if X is a negative binomial rv with pmf $nb(x; r, p)$, then $E(X) = r(1-p)/p$,
 $\text{Var}(X) = r(1 - p)/p^2$.

Statistical Inference

- From the previous two examples, we know that quite often, we need to infer the truth (**population**) from some partial information (**sample**).
- Question: why do we need a model?
- **Statistical inference** comprises the use of statistics and random sampling to make inferences concerning some unknown aspect of a population.
- A **point estimate** of a parameter θ is a single number that can be regarded as a sensible value for θ . A point estimate is obtained by selecting a suitable statistic and computing its value from the given sample data. The selected statistic is called the **point estimator** of θ .

Sampling scheme for a Mean

- Usually our problem set up will be as illustrated in the graph.



- The actual sample observations y_1, y_2, \dots, y_n (**realizations**) are assumed to be the result of a random sample Y_1, Y_2, \dots, Y_n (**random variables**) from a certain distribution.

Estimating probability

Ex. A biased coin has probability p of having heads and p is unknown. Suppose we flipped the coin for 100 times and had 73 heads. What is your best guess for p ?

Naturally, people would use estimator $\hat{p} = \frac{\text{number of heads}}{\text{number of flips}} = \frac{73}{100} = 0.73$

In other words, we are using the **sample proportion** to estimate the **population probability**.

Is this a good estimator? Are there any other estimators?

Guidelines on Estimation

1. Decide **what** (population mean/variance, proportion, etc.) to estimate and **how** (sample mean/variance, proportion, etc.) to estimate.
2. Obtain the sample, and calculate the estimator based on “**how**” from 1.
3. Report the estimator as well as its associated **error**.

More estimators

- In most problems, there will be **more than one** reasonable estimators.

Ex. Use `rnorm` to generate 20 normal random variables with mean 3 and standard deviation 1. Let's denote them as x_1, x_2, \dots, x_{20} . Now we pretend we do NOT know anything about the true underlying distribution. Come up with some estimators for the mean parameter.

(mean, trimmed-mean, median, partial mean)

Measure of a good Estimator

- Our estimator $\hat{\theta}$ is in fact a function of the sample x_i 's, therefore, it is also a random variable. For some samples, $\hat{\theta}$ may yield a value larger than θ , whereas for other samples $\hat{\theta}$ may underestimate θ .
- The quantity $\hat{\theta} - \theta$ characterize the error of estimation. A good estimator should result in small estimation errors.
- A commonly used measure of accuracy is the **mean square error**.

$$\text{MSE} = E(\hat{\theta} - \theta)^2$$

- However, since MSE will generally depend on the value of θ , finding an estimator with smallest MSE is typically **NOT** possible.

Unbiased Estimators

- One way to find good estimators, is to restrict our attention just to estimators that have some specified desirable properties and then find the best in this restricted group.
- One popular property is *unbiasedness*.
- A point estimator $\hat{\theta}$ is said to be an *unbiased estimator* of θ if $E(\hat{\theta}) = \theta$ for every possible value of θ . If $\hat{\theta}$ is not unbiased, the difference $E(\hat{\theta}) - \theta$ is called the *bias* of $\hat{\theta}$.

Example

Ex. Recall the unbiased coin example. Is the sample proportion an unbiased estimator of the population probability?

$$\text{estimator } \hat{p} = \frac{\text{number of heads}}{\text{number of flips}} = \frac{73}{100} = 0.73$$

What distribution does “number of heads” follow? What is its expectation?

General Result

- Proposition:

When X is a binomial rv with parameters n and p , the sample proportion $\hat{p} = X/n$ is an unbiased estimator of p .

Example

Ex. Suppose that X , the reaction time to a certain stimulus, has a uniform distribution on the interval from 0 to an unknown upper limit θ . It is desired to estimate θ based on random sample x_1, x_2, \dots, x_n . What is the best guess of θ ? Is your estimator unbiased?

General Result

- Proposition:

Let X_1, X_2, \dots, X_n be an i.i.d. sequence of random samples from a distribution with mean μ and variance σ^2 . Then the estimator

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

is an unbiased estimator of σ^2 .

General Result

- Proposition:

Let X_1, X_2, \dots, X_n be an i.i.d. sequence of random samples from a distribution with mean μ . Then the sample mean \bar{X} is an unbiased estimator of μ . If in addition the distribution is continuous and symmetric, then the sample median M and any trimmed mean are also unbiased estimators of μ .

MVUE

- For unbiased estimators, what are their MSE's?

$$E(\hat{\theta} - \theta)^2 = E(\hat{\theta} - E(\hat{\theta}))^2 = \text{Var}(\hat{\theta})$$

- Among all estimators of θ that are unbiased, choose the one that has minimum variance. The resulting $\hat{\theta}$ is called the **minimum variance unbiased estimator (MVUE)** of θ .
- One needs more knowledge to actually identify if some estimator is really MVUE. But in a special case, we have the following theorem.

Let X_1, X_2, \dots, X_n be an i.i.d. sequence of random samples from a **normal distribution** with mean μ and σ . Then the estimator $\hat{\mu} = \bar{X}$ is the **MVUE** for μ .

The Standard Error

- When reporting a point estimator, one also reports the **standard error** associated with it.
- The **standard error** of an estimator $\hat{\theta}$ is its standard deviation $\sigma_{\hat{\theta}} = \sqrt{\text{Var}(\hat{\theta})}$. If the standard error itself involves unknown parameters whose values can be estimated, substitution of these estimates into $\sigma_{\hat{\theta}}$ yields the **estimated standard error** of the estimator, which we denote as $\hat{\sigma}_{\hat{\theta}}$.
- The associated standard error gives us an idea of how good/accurate the estimators are.

Examples

Ex. What is the standard error of the sample proportion?

Ex. What is the standard error of the sample mean? Given that we know the variance.

Ex. What is the standard error of the sample mean? Given that we don't know the variance.

Methods of Point Estimation

- The definition of unbiasedness does not in general indicate how unbiased estimators can be derived.
- There are two commonly used “constructive” methods for obtaining point estimators: the [method of moments](#) and the [method of maximum likelihood](#).
- Although maximum likelihood estimators are generally preferable to moment estimators because of certain efficiency properties, they often require significantly more computation than do moment estimators.
- It is **NOT** guaranteed that these two methods would yield unbiased estimators.

Methods of Point Estimation

- The definition of unbiasedness does not in general indicate how unbiased estimators can be derived.
- There are two commonly used “constructive” methods for obtaining point estimators: the [method of moments](#) and the [method of maximum likelihood](#).
- Although maximum likelihood estimators are generally preferable to moment estimators because of certain efficiency properties, they often require significantly more computation than do moment estimators.
- It is **NOT** guaranteed that these two methods would yield unbiased estimators.

Moment Estimators

- Definition:

Let X_1, X_2, \dots, X_n be an i.i.d. sample from a pmf or pdf $f(x)$. For $k = 1, 2, 3, \dots$, the moment estimator for the k th population moment, is the k th sample moment, i.e.,

$$\widehat{E(X^k)} = \frac{\sum_{i=1}^n X_i^k}{n}$$

Remarks

- Notice that, the sample k th moment will **converge** to the population k th moment, as sample size $n \rightarrow \infty$, thanks to the **LLN**.
- For any finite sample size n , the sample k th moment in general will **not** be **equal** to the population k th moment. But it is a good candidate for estimation.
- The larger the n , the better the estimation is!
- In practice, if the underlying pmf or pdf has m parameters, namely, we have $f(x; \theta_1, \dots, \theta_m)$, where $\theta_1, \dots, \theta_m$ are parameters whose values are unknown. Then the moment estimators are obtained by **equating the first m sample moments to the corresponding first m population moments and solving for $\theta_1, \dots, \theta_m$** .

Example

Ex. Show that the sample proportion is the moment estimator of the population probability.

Example

Ex. Let X_1, X_2, \dots, X_n be an i.i.d. normal sample, and assume that the underlying normal distribution is $N(\mu, \sigma^2)$ where μ, σ^2 are unknown. How can we construct moment estimators to estimate the two unknown parameters?

As we already know if $X \sim N(\mu, \sigma^2)$, then $E(X) = \mu$, and $E(X^2) = \mu^2 + \sigma^2$.

Therefore, we have two equations:

$$\begin{cases} \hat{\mu} = \sum_{i=1}^n X_i / n \\ \hat{\mu}^2 + \hat{\sigma}^2 = \sum_{i=1}^n X_i^2 / n \end{cases} \longrightarrow \begin{cases} \hat{\mu} = \sum_{i=1}^n X_i / n \\ \hat{\sigma}^2 = \sum_{i=1}^n X_i^2 / n - \bar{X}^2 \end{cases}$$

Is the variance estimator unbiased?

Example

Ex. Let X_1, X_2, \dots, X_n be an i.i.d. sample from exponential distribution with parameter λ which is unknown. How do we estimate λ using moment estimator?

As we already know if $X \sim \text{Exp}(\lambda)$, then $E(X) = 1/\lambda$.

Thus, we have equation $1/\hat{\lambda} = \bar{X} \rightarrow \hat{\lambda} = 1/\bar{X}$.

Is this estimator unbiased?

Maximum Likelihood Est.

- The method of **maximum likelihood** was first introduced by **R.A. Fisher**, a geneticist and statistician, in the 1920s. It is by far the most commonly used method to obtain estimators.
- **Likelihood function** is just another way of looking at the *joint pmf or the pdf*. In particular, let X_1, X_2, \dots, X_n (not necessarily i.i.d.) have joint pmf or pdf

$$f(x_1, x_2, \dots, x_n; \theta_1, \dots, \theta_m)$$

where $\theta_1, \dots, \theta_m$ are parameters whose values are unknown. When x_1, x_2, \dots, x_n are the observed sample values and $f(\cdot)$ is then regarded as a function of $\theta_1, \dots, \theta_m$, it is called the **likelihood function**.

Example

Ex. A biased coin has been flipped for 10 times. Let X_1, X_2, \dots, X_{10} denote the outcomes of the coin flips. Assume the probability of having a head is p (parameter of interest), and the sample we observed is $\{0, 1, 1, 0, 0, 0, 1, 0, 0, 0\}$. Write down the likelihood function for p .

$$f(x_1, x_2, \dots, x_n; p) = f(x_1; p) f(x_2; p) \dots f(x_n; p) = (1-p) p p (1-p) \dots (1-p) = p^3(1-p)^7$$

Idea of **Maximum Likelihood**: can we find a p that can **maximize** the above function?

MLE

- The **maximum likelihood estimates** (mle's) $\hat{\theta}_1, \dots, \hat{\theta}_m$ are those values of θ_i 's that maximize the likelihood function, so that

$$f(x_1, \dots, x_n; \hat{\theta}_1, \dots, \hat{\theta}_m) \geq f(x_1, \dots, x_n; \theta_1, \dots, \theta_m) \quad \text{for all } \theta_1, \dots, \theta_m$$

when the X_i 's are substituted in place of the x_i 's.

- **Remark:** the likelihood function tells us how likely the observed sample is as a function of the possible parameter values. Maximizing the likelihood gives the parameter values for which **the observed sample is most likely to have been generated** – that is, the parameter values that “**agree most closely**” with the observed data.
- In practice, in stead of maximizing the likelihood itself, people usually choose to maximize the **log-likelihood function**.

Example

Ex. Let X_1, X_2, \dots, X_n be an i.i.d. sample from exponential distribution with parameter λ which is unknown. Write down the likelihood function for λ . What is the MLE of λ ? Is the MLE unbiased?

Since we have an i.i.d. sample, it is easy to see that the likelihood function is a product of the individual pdf's:

$$f(x_1, \dots, x_n; \lambda) = (\lambda e^{-\lambda x_1}) \dots (\lambda e^{-\lambda x_n}) = \lambda^n e^{-\lambda \sum x_i}$$



$$\log[f(x_1, \dots, x_n; \lambda)] = n \log(\lambda) - \lambda \sum x_i$$



$$\hat{\lambda} = n / \sum X_i$$

Example

Ex. Let X_1, X_2, \dots, X_n be an i.i.d. sample from normal distribution with parameter μ and σ^2 which is unknown. Write down the likelihood function. What are the MLE's of the two parameters? Are they unbiased?


Some complications

- The following is an example that MLE's can't be calculated analytically.

Ex. Let X_1, X_2, \dots, X_n be an i.i.d. sample from Weibull distribution with parameters α and β and pdf

$$f(x; \alpha, \beta) = \begin{cases} \frac{\alpha}{\beta^\alpha} \cdot x^{\alpha-1} \cdot e^{-(x/\beta)^\alpha} & x \geq 0, \\ 0 & \text{otherwise.} \end{cases}$$

by solving equations $\frac{\partial \log(f)}{\partial \alpha} = 0 \quad \frac{\partial \log(f)}{\partial \beta} = 0$


$$\hat{\alpha} = \left[\frac{\sum x_i^{\hat{\alpha}} \cdot \log(x_i)}{\sum x_i^{\hat{\alpha}}} - \frac{\sum \log(x_i)}{n} \right]^{-1} \quad \hat{\beta} = \left(\frac{\sum x_i^{\hat{\alpha}}}{n} \right)^{1/\hat{\alpha}}$$

Some complications

- The following examples show that sometimes calculus cannot be used to obtain MLE's

Ex. Suppose my waiting time for a bus is uniformly distributed on $[0, \theta]$ and the results x_1, \dots, x_n of a random sample from this distribution have been observed. What is the MLE of θ ? Is it unbiased?

Ex. The fish example again. Using MLE approach. (book p.250)

The Invariance Principle

- One of the nice features of MLE's is that, the MLE of a function of parameters, is the function of the MLE's of the parameters.
- More specifically, we have

Let $\hat{\theta}_1, \dots, \hat{\theta}_m$ be the MLE's of the parameters $\theta_1, \dots, \theta_m$. Then the MLE of any function $h(\theta_1, \dots, \theta_m)$ of these parameters is $h(\hat{\theta}_1, \dots, \hat{\theta}_m)$.

Ex. In the normal example, what is the MLE of σ ?

Large Sample Behavior

- The following proposition says, for large samples, it is “**optimal**” to use MLE’s, because it is **asymptotically unbiased** and has the **minimal variance** among all unbiased estimators.
- **Proposition:**

Under very general conditions on the joint distribution of the sample,
When the sample size n is large, the **maximum likelihood estimator** is
Approximately the **MVUE** of the parameter.