FALL 2013
Stat 8003: Statistical Methods I
Lecture 5

Jichun Xie

1 Features of Distributions

1.1 Variance

Variance is a measure of spread of the distribution in the population.

Definition.

$$\texttt{Var}(X) = \texttt{E}\{X - \texttt{E}(X)\}^2 = \texttt{E}(X^2) - (\texttt{E}X)^2.$$

- Discrete: $\texttt{Var}(X) = \sum_x \{X - \texttt{E}(X)\}^2 \texttt{P}(X = x)$.
- Continuous: $\texttt{Var}(X) = \int_x \{X - \texttt{E}(X)\}^2 f_X(x)\,\mathrm{d}x$.

Properties.

1. For some constants a, b, $\texttt{Var}(aX + b) = a^2\texttt{Var}(X)$.
2. If Y_i, $i = 1, \ldots, n$, are independent of each other, $\texttt{Var}(\sum_{i=1}^n Y_i) = \sum_{i=1}^n \texttt{Var}(Y_i)$.
3. For a pair of random variables X and Y,

$$\texttt{Var}(X + Y) = \texttt{Var}(X) + \texttt{Var}(Y) + 2\texttt{Cov}(X, Y),$$

where

$$\texttt{Cov}(X, Y) = \texttt{E}\{(X - \texttt{E}X)(Y - \texttt{E}Y)\} = \texttt{E}(XY) - \texttt{E}X\texttt{E}Y.$$

Why are we interested in variance? It provides a description of the precision of a statistic. For two unbiased estimates ($\texttt{E}X = \theta$, the parameter of interest) of the same parameter, a statistic with lower variance is more interesting than the one with a higher variance. In addition, sometime variance itself is of interest since it characterize the population.

Example.

In lots of situations, we can use the formula $\mathtt{Var}(X) = \mathtt{E}(X^2) - (\mathtt{E}X)^2$ to calculate the variance. In Last Lecture, we discussed how to calculate $\mathtt{E}X$ for some commonly seen distributions. Now we can use the similar techniques to caculcate the variance.

1. Binomial distribution: $X \sim \text{Binom}(n, p)$. Then

$$\begin{aligned}\mathtt{E}(X^2) &= \mathtt{E}X(X-1) + \mathtt{E}X = n(n-1)p^2 + np \\ \mathtt{Var}(X) &= \mathtt{E}X^2 - (\mathtt{E}X)^2 = n(n-1)p^2 + np - (np)^2 = np(1-p)\end{aligned}$$

2. Poisson distribution: $X \sim \text{Poisson}(\lambda)$. Then

$$\begin{aligned}\mathtt{E}(X^2) &= \mathtt{E}X(X-1) + \mathtt{E}X = \lambda^2 + \lambda \\ \mathtt{Var}(X) &= \mathtt{E}X^2 - (\mathtt{E}X)^2 = \lambda\end{aligned}$$

3. Exponential distribution: $X \sim \text{Exp}(\lambda)$, with $f(x) = \lambda \exp(-\lambda x)$.

$$\begin{aligned}\mathtt{E}(X^2) &= \int_0^\infty x^2 \lambda \exp(-\lambda x)\,\mathrm{d}x = \frac{2}{\lambda^2} \\ \mathtt{Var}(X) &= \mathtt{E}(X^2) - (\mathtt{E}X)^2 = \frac{1}{\lambda^2}\end{aligned}$$

4. Normal distribution: $X \sim \text{N}(\mu, \sigma^2)$. We know that $\mathtt{E}X = \mu$. Then

$$\mathtt{Var}(X) = \mathtt{E}(X-\mu)^2 = \int_{-\infty}^{\infty} (x-\mu)^2 \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right)\,\mathrm{d}x = \sigma^2$$

5. At one genetic location, suppose there are two possible nucleotide types, denoted as "A" and "a". For each person, the possible genetic combination at this location is "AA", "Aa" or "aa". "A" is called a wide allele and "a" is called a minor allele. Let

$$X = \begin{cases} 0 & \text{if "AA"} \\ 1 & \text{if "Aa"} \\ 2 & \text{if "aa"} \end{cases}$$

Now suppose there are two genetic locations. At location 1, the allele types are "AA", "Aa" and "aa"; at location 2, the allele types are "BB", "Bb" and "bb". Now suppose in the population, at one allele, $\mathtt{P}(A) = 0.6$ and $\mathtt{P}(B) = 0.95$. X_1 is the count of "a" alleles at location 1 and X_2 is the count of "b" alleles at location 2. What is $\mathtt{E}(X_1)$, $\mathtt{E}(X_2)$, $\mathtt{Var}(X_1)$ and $\mathtt{Var}(X_1)$? How to interprete the results?

1.2 Other Moments

For a random variable X, define its r-th moment as $\mathtt{E}(X^r)$, and define its r-th center moment as $\mathtt{E}(X-\mu)^r$, for $r = 1, 2, \ldots$

Skewness. The third central moment $\mathtt{E}(X-\mu)^3$ measures the lack of symmetry.

$$\gamma_1 = \mathtt{E}(X-\mu)^3/\sigma^3,$$

where $\sigma^2 = \mathtt{Var}(X)$. $\gamma_1 = 0$ stands for a symmetric distribution.

Kurtosis. The fourth central moment $\mathtt{E}(X-\mu)^4$ measures the pointness of the distribution.

$$\gamma_2 = \mathtt{E}(X-\mu)^4/\sigma^4 - 3.$$

For Normal distribution, $\gamma_2 = 0$.

2 Estimation

2.1 Introduction

Statistica Inference:

population distribution	$\Leftarrow$	(inference)	samples
	$\Rightarrow$	(probability sample)	

What are we trying to estimate? It depends on the question of interest.

- pdf/cdf: defining the probability distribution – lots of information
- A subset of information given by pdf/cdf
- Location or scale parameters (*e.g.*, mean, median, mode)

Two basic types of inference

1. Estimation
   - Point estimation (direct estimation)
   - Confidence interval (precision of the estimation)
2. Hypothesis testing
   - Decision making

2.2 Parameter Estimation

Under some cases, we have some aspects of the random variables that is of interest, *e.g.*, $\mathtt{E}(X)$, $\mathtt{Var}(X)$, *etc.*

Basic procedures.

1. data collection
2. formulate statistical model
3. calculate statistic, also a random variable
4. use pdf or statistic for inference

How do we formulate a statistic model?

Parametric Model.

- make lots of assumptions
- choose of probability density function indexed by parameters as the basis of the model
- most precise when we get the distribution right
- if wrong, bad properties

Semi-parameteric model.

- *e.g.*, moment based methods: assume $\mathtt{E}(Y) = \mu$, $\mathtt{Var}(Y) = \sigma^2$
- relaxed some assumptions

Non-parametric model.

- no assumptions about specific mathematical form of distributions
- some assumptions, *e.g.*, symmetric distributions
- less efficient if we know the distribution but don't use the information

How to estimate? For estimation problem, one key point is to find a proper "statistic". A statistic is a function of the data which we my use for inference. It is a r.v. with its own pdf and cdf, *e.g.*, sample $Y_1, \ldots, Y_n$. Possible (not inclusive) include:

- $Y_{(1)}$, the minimum of the sample
- $\bar{Y} = \frac{1}{n}\sum_{i=1}^{n} Y_i$, sample mean
- M_Y, the sample median ($Y_{(n/2+1/2)}$ for n odd, and $\frac{1}{2}[Y_{n/2} + Y_{(n/2_1/2)}]$ for n even, where $Y_{(i)}$ are the order statistics)

2.3 Method of Moments

Suppose $Y \sim f_Y(y)$, and $Y_1, \ldots, Y_n$ are n *i.i.d.* observations of Y. Recall that the r-th (population) moment of Y is $\mathtt{E}Y^r$. The r-th sample moment of Y is

$$\widehat{\mathtt{E}Y^r} = \frac{1}{n}\sum_{i=1}^{n} Y_i^r.$$

Sample moments are often "good" estimators of population moments (Weak Law of Large Numbers).

Suppose θ is the parameter of interest. Assume a parametric family of distributions is

$$\mathcal{F} = \{f_Y(y;\theta),\ \theta \in \Theta\}$$

Our estimator, $\hat{\theta}$, can be a function of the sample moments.

Example.

Suppose $Y_1, \ldots, Y_n \sim \mathrm{N}(\mu, 1)$, and $\mathtt{E}(Y_i) = \mu$. What is the moment estimator of μ?

Answer: $\bar{Y}$

Method. More generally, suppose Y is a continous r.v., with distribution function

$$f_Y(y;\theta_1, \ldots, \theta_K)$$

The first K moments of Y (assuming they exist) are:

$$\mathtt{E}(Y^k) = \int y^k f_Y(y;\theta_1, \ldots \theta_K)\,\mathrm{d}y, \quad k = 1, \ldots, K$$

Equate the population and sample moments – we have K equations in K unknowns, and we can solve for them.

Example.

Suppose $Y_1, \ldots, Y_n \sim \mathrm{N}(\mu, \sigma^2)$. What are the MOM estimators of μ and σ^2?

$$\mathtt{E}Y_i = \mu \tag{1}$$
$$\mathtt{E}(Y_i^2) = \mu^2 + \sigma^2 \tag{2}$$

Set the expectation equal to the sample estimates

$$\hat{\mu} = \bar{Y}$$
$$\hat{\mu}^2 + \hat{\sigma}^2 = \frac{1}{n}\sum_{i=1}^{n} Y_i^2$$

Solving the equations, we have

$$\hat{\mu}_n = \bar{Y}$$
$$\hat{\sigma}^2 = \frac{1}{n}\sum_{i=1}^{n}(Y_i - \bar{Y})^2$$

Pros of MOM estimators.

Under some regularity conditions,

- The moment estimator $\hat{\theta}$ exist
- The estimator is consistent: $\hat{\theta} \xrightarrow{\mathtt{P.}} \theta$.
- The estimator vector is a function of the sample means, and thus is asymptotically multivariate normal.

Cons of MOM estimators.

- MOM is not unique.
- Sometimes MOM doesn't make sense.

Why MOM is not unique? Think about Poisson.

$$\mathtt{E}X = \lambda$$
$$\mathtt{E}X^2 = \lambda + \lambda^2$$

2.4 Maximum Likelihood

Definition of Likellihood. Likelihood is a joint density of the data viewed as *a function of the parameter* that characterize the family of distributions.

Example. Consider the following situation for a density of Y indexed by θ:

θ	Y			$\sum_y \mathtt{P}(Y = y)$
	-1	0	1	
1	0.2	0.3	0.5	1
2	0.7	0.2	0.1	1
3	0.2	0.6	0.2	1

Now consider taking a single sample of Y's. The observed value of Y is $Y = 0$. What is the most likely value of θ?

Definition of MLE.

More generally, suppose $Y_1, \ldots, Y_n \sim f_Y(y; \theta)$, *i.i.d.*, where $\theta \in \Theta$. If the observed values are $y_1, ..., y_n$ in the sample, then the likelihood

$$L_n(\theta) = \prod_{i=1}^{n} f_Y(y_i; \theta).$$

The maximum likelihood estimator is

$$\hat{\theta} = \max_{\theta \in \Theta} L_n(\theta).$$

Note that people often work with

$$l_n(\theta) = \log L_n(\theta) = \sum_{i=1}^{n} \log f(y_i; \theta)$$

and maximize $l_n(\theta)$ instead.

How do we find MLE's?

- graphing (simple, intuitive)
- calculus (exponential families)

- numerical techniques (more complicated problems,*e.g.*, Newton Raphson and EM algorithm)

Example.

1. Suppose the number of category 4 hurricans in each 5-season interval for the past 5 years are: $1, 4, 6, 9, 10$, following a distribution of Poi(λ). What is the MLE of λ?

$$L_n(\lambda) = \prod_{i=1}^{n} \lambda^{Y_i} \frac{\exp(-\lambda)}{Y_i!}$$
$$l_n(\lambda) = \sum_{i=1}^{n} Y_i \log(\lambda) - n\lambda - \sum_{i=1}^{n} \log(Y_i!)$$

Set $l_n'(\lambda) = 0$, we have $\hat{\lambda} = \bar{Y}$. Check second derivative $l_n''(\lambda) < 0$, thus we have a MLE.

Now the observed data are $(Y_1, \ldots, Y_5) = (1, 4, 6, 9.10)$. The sample mean $\bar{Y} = 6.2$. Then the rate of C4 hurricanes is 6.2 per 5-season interval, or equivalently, 1.2 category 4 curricane per season interval.

Properties of MLE's. Suppose $\hat{\theta}$ is the MLE of θ. Under some regularity conditions,

1. The MLE is consistent, $\hat{\theta} \xrightarrow{\mathtt{P}} \theta$.
2. The MLE is transform equivalent. If $\hat{\theta}$ is the MLE of θ, and g is a one-to-one function. Then $g(\hat{\theta})$ is the MLE of $g(\theta)$.
3. The MLE is asymptotically optimal or efficient. Among all well-behaved estimators, the MLE has the smallest variance in large sample.
4. The MLE is asymptotically normal

$$\sqrt{I_n(\theta)}(\hat{\theta}_{\text{mle}} - \theta) \xrightarrow{D.} N(0, 1),$$

where $I_n(\theta)$ is the variance of $\hat{\theta}$

$$I_n(\theta) = -\mathtt{E}\left[\left\{ \frac{\mathrm{d}^2}{\mathrm{d}\theta^2} l_n(\theta) \right\}\right]$$

From the last time, we know that $\mathtt{Var}(\hat{\theta}_{\text{mle}}) = I_n^{-1}(\theta)$. This item is useful for computing confidence interval and constructing hypothesis testing.

Example. We are interested in estimating the rate of becoming pregnant among smokers. The data are shown in the table below.

Cycle	1	2	3	4	5	6	7	8	9	10	11	12	> 12
Pregnancies	29	16	17	4	3	9	4	5	1	1	1	3	7

Suppose the probability p of being pregnant remains constant for all smokers. How to estimate p?

Note that in the table there are two types of women. Most women became pregnant in the study and the number of menstrual cycles that it took for this to occur is listed in the table below. For example, 29 women were pregnant during their first cycle. On the other hand seven women were not pregnant after 12 cycles.

Let Y_i, $i = 1, \ldots, n$ be the number of cycles up to and including the one in which the ith woman became pregnant. We can assume Y_i follows Geo(p).

$$\mathsf{P}(Y_i = y_i) = p(1-p)^{y_i - 1}, \quad y = 1, \ldots, \infty$$

And thus,

$$\mathsf{P}(Y_i > 12) = \sum_{y_i=13}^{\infty} p(1-p)^{y_i-1} = \frac{p(1-p)^{12}}{1-(1-p)} = (1-p)^{12}.$$

The likelihood function is

$$\begin{aligned} L_n(p) &= \prod_{k=1}^{12} \{p(1-p)^{k-1}\}^{n_k} \{(1-p)^{12}\}^{n_{13}} \\ &= \prod_{k=1}^{12} p^{n_k} (1-p)^{(k-1)n_k} (1-p)^{12 n_{13}} \end{aligned}$$

And the log-likelihood is

$$l_n(p) = \sum_{k=1}^{12} n_k \log p + \sum_{k=1}^{12} n_k (k-1) \log(1-p) + 12 n_{13} \log(1-p)$$

Let

$$\begin{aligned} N &= \sum_{k=1}^{12} n_k \\ N_u &= \sum_{k=1}^{12} n_k (k-1) \end{aligned}$$

N is the total number of woman being pregnant up to 12 months. And N_u is the number of total unsuccessful trials of those women who were pregnant before 12 months. Then

$$l_n(p) = N \log(p) + N_u \log(1-p) + 12 n_{13} \log(1-p).$$

Set

$$l_n'(p) = \frac{N}{p} - \frac{N_u}{1-p} - \frac{12n_{13}}{1-p} = 0.$$

Then

$$\hat{p} = \frac{N}{N + N_u + 12n_{13}}$$

Check the second derivative

$$l_n''(p) = -\frac{N}{p^2} - \frac{N_u}{(1-p)^2} - \frac{12n_{13}}{(1-p)^2} < 0.$$

Therefore, $\hat{p}$ is the MLE of p.

2.5 Bias, Variance and MSE

Among all the possible choices, which estimate should we use? What is a good estimate?

- What kind of estimators get the answer close to the truth on average? (Bias)
- What kind of estimators get the answer close to the truth most of the time? (Variance)
- When sample size is large, can we get an estimator approaches to the truth? (Convergence)

2.5.1 Bias

Let T be out estimator. On average (in repeated experiments), we want T to be equal to the parameter of interest θ.

Definition of Bias. Let $T = g(Y_1, \ldots, Y_n)$. Suppose the parameter of interest is θ. Define

$$\texttt{Bias}(T) = \texttt{E}(T) - \theta.$$

If T is unbiased, $\texttt{Bias}(T) = 0$.

Example.

1. Let $Y_1, \ldots, Y_n \sim \text{Bernoulli}(\theta)$, *i.i.d.* Suppose $W = \frac{1}{n}\sum_{i=1}^n Y_i$. Then

$$\texttt{E}(W) = \frac{1}{n}\sum_{i=1}^n \texttt{E}(Y_i) = \theta.$$

W is unbiased.

2. Let $Y_1, \ldots, Y_n \sim N(\mu, \sigma^2)$, *i.i.d.* Assure yourself that $\hat{\mu} = \frac{1}{n}\sum_{i=1}^{n} Y_i$ is unbiased estimator of μ. Suppose μ is known. Assure yourself that $\tilde{\sigma}^2 = \frac{1}{n}\sum_{i=1}^{n}(Y_i - \mu)^2$ is unbiased estimator σ^2.

3. Following Example 2, if μ is unknown, then we cannot use it in the statistic for estimating σ^2. Then

$$\hat{\sigma}^2 = \frac{1}{n}\sum_{i=1}^{n}(Y_i - \bar{Y})^2$$

is biased. Assure yourself that it is true. (Key: in the bracket, add $+\mu - \mu$)

From Example 3, we know that

$$\mathtt{E}\hat{\sigma}^2 = \frac{n-1}{n}\sigma^2.$$

Let $s^2 = \frac{n}{n-1}\hat{\sigma}^2 = \frac{1}{n}\sum_{i=1}^{n}(Y_i - \bar{Y})^2$. Then s^2 is unbiased. And when the sample size n is large, the bias of $\hat{\sigma}^2$ approaches to zero. We call $\hat{\sigma}^2$ asymptotically unbiased.

2.5.2 Efficiency

Let T_1 and T_2 be unbiased estimators of θ. T_1 is more efficient than T_2 if

$$\mathtt{Var}(T_1) < \mathtt{Var}(T_2).$$

The relative efficiency is $R(T_1, T_2) = \texttt{Var}(T_2)/\texttt{Var}(T_1)$. When $R(T_1, T_2) < 1$, T_2 is more efficient than T_1.

Example.

Suppose $Y_1, \ldots, Y_n \sim \text{N}(\mu, \sigma^2)$, *i.i.d.*. Let $W_1 = \bar{Y}$ and $W_2 = Y_1$. Note that $\texttt{E}(W_1) = \texttt{E}(W_2) = \mu$. They are both unbiased. However $\texttt{Var}(W_1) = \sigma^2/n$ and $\texttt{Var}(W_2) = \sigma^2$. It leads to $R(W_1, W_2) = n \geq 1$, so that W_1 is more efficient than W_2 unless $n = 1$.

2.5.3 Mean Square Error

Suppose that W_1 and W_2 are two statistics of the parameter of interest θ. At least of them is biased. How to compare their performance?

Example. Suppose $Y_1, \ldots, Y_n \sim N(\mu, \sigma^2)$, *i.i.d.*, with both μ and σ^2 unknown. We just discussed that there are two estimators of σ^2: $\hat{\sigma}^2$ and s^2. We know that

$$\texttt{Bias}(\hat{\sigma}) = -\frac{1}{n}\sigma^2, \ \texttt{Bias}(s^2) = 0$$
$$\texttt{Var}(\hat{\sigma}) = \frac{(n-1)^2}{n^2}\texttt{Var}(s^2)$$

Which one should we use?

Definition of Mean Square Error. Suppose $T = g(Y_1, \ldots, Y_n)$ is a statistic for the parameter of interest θ. Then

$$\texttt{MSE}(T) = \texttt{E}(T - \theta)^2.$$

Note that

$$\texttt{MSE}(T) = \texttt{E}\{(T - \texttt{E}T) + (\texttt{E}T - \theta)\}^2 = \texttt{E}(T - \texttt{E}T)^2 + (\texttt{E}T - \theta)^2 = \texttt{Var}(T) + \texttt{Bias}(T)^2.$$

MSE is a combined measure of Bias and MSE.

Example.

In the above example, what are the MSE of $\hat{\sigma}^2$ and s^2? We know the biases for both. We only need to get the variances. To get the variance, we need the following result:

If $Y_1, \ldots, Y_n \sim \text{N}(\mu, \sigma^2)$, *i.i.d.*, then

$$\frac{1}{\sigma^2}\sum_{i=1}^{n}(Y_i - \bar{Y})^2 \sim \chi^2(n-1).$$

For $Z \sim \chi^2(k)$, $\texttt{E}(Z) = k$ and $\texttt{Var}(Z) = 2k$.

Then it is easy to see that

$$\texttt{Var}(s^2) = \frac{2}{n-1}\sigma^4$$
$$\texttt{Var}(\hat{\sigma}^2) = \frac{2(n-1)}{n^2}\sigma^4$$

Then

$$\texttt{MSE}(s^2) = \{\texttt{Bias}(s^2)\}^2 + \texttt{Var}(s^2) = \frac{2}{n-1}\sigma^4$$
$$\texttt{MSE}(\hat{\sigma}^2) = \{\texttt{Bias}(\hat{\sigma}^2)\}^2 + \texttt{Var}(\hat{\sigma}^2) = \frac{2n-1}{n^2}\sigma^4$$

To compare their MSE,

$$\frac{\texttt{MSE}(\hat{\sigma}^2)}{\texttt{MSE}(s^2)} = \frac{(2n-1)(n-1)}{2n^2} = \frac{(n-1/2)(n-1)}{n^2} < 1.$$

Therefore, in terms of MSE, $\hat{\sigma}^2$ is more efficient than s^2. When the sample size is large, they are asymptotically equivalently efficient.

$$\lim_{n\to\infty} \frac{\texttt{MSE}(\hat{\sigma}^2)}{\texttt{MSE}(s^2)} = 1.$$

Summary of Criterion of Comparison:

1. Unbiasedness
2. Relative Efficiency
3. MSE