Topic 7: Inequalities and Limit Theorems

Rohini Somanathan

Course 003, 2018

The Inference Problem

So far, our starting point has been a given probability space $(S, \mathcal{F}, \mathbb{P})$.

In the remainder of this course, we will make inferences about the probability space by analyzing a sample of outcomes. This is known as statistical inference.

Statistical inference involves both the estimation of population parameters and testing hypotheses about them.

We focus on parametric inference - we make assumptions about the probability distributions from which our sample is drawn (for example, each sample observation represents an outcome of a normally distributed random variable with unknown mean and unit variance).

When no such assumptions are made, inference is nonparametric.

Defining a Statistic

Definition: *Any real-valued function* $T = r(X_1, \ldots, X_n)$ *is called a statistic.*

A statistic is a random variable, but not all functions of random variables are statistics.

In which of the following examples is Y a statistic?

$$Y = \frac{X - \mu}{\sigma}$$

$$Y = \sum_{i=1}^{n} X_i.$$

$$Y = \sum_{i=1}^{n} X_i^2.$$

Statistics are useful because they have sample counterparts.

In a problem of estimating an unknown parameter, θ, our estimator will be a statistic whose value can be regarded as an estimate of θ.

It turns out that for large samples, the distributions of some statistics, such as the sample mean, are well-known.

Markov's Inequality

When we don't know the distribution of a random variable, we can estimate probabilities either through simulation or by distribution-free bounds such as those provided by the following inequalities.

Markov's Inequality: *Let* X *be a random variable with density function* $f(x)$ *such that* $P(X \geq 0) = 1$. *Then for any given number* $a > 0$,

$$P(X \geq a) \leq \frac{E(X)}{a}$$

Proof. (for discrete distributions) $E(X) = \sum_x xf(x) = \sum_{x<a} xf(x) + \sum_{x \geq a} xf(x)$

All terms in these summations are non-negative by assumption, so we have

$$E(X) \geq \sum_{x \geq a} xf(x) \geq \sum_{x \geq a} af(x) = aP(X \geq a)$$

□

This inequality obviously holds for $a \leq E(X)$. It is useful in bounding probabilities in the tails. For example, if the mean of X is 1, the probability of X taking values bigger than 100 is less than .01.

Chebyshev's Inequality

Chebyshev's Inequality: Let X be a random variable with finite variance σ^2 and mean μ. Then, for every $a > 0$,

$$P(|X - \mu| \geq a) \leq \frac{\sigma^2}{a^2}$$

or equivalently,

$$P(|X - \mu| < a) \geq 1 - \frac{\sigma^2}{a^2}$$

Proof. Use Markov's inequality with $Y = (X - \mu)^2$ and use a^2 in place of the constant a. Then Y takes only non-negative values and $E(Y) = Var(X) = \sigma^2$. □

In particular, this tells us that for any random variable, the probability that values taken by the variable will be more than 2 standard deviations away from the mean cannot exceed $\frac{1}{4}$ and 3 standard deviations away cannot exceed $\frac{1}{9}$

$$P(|X - \mu| \geq 3\sigma) \leq \frac{1}{9}$$

For most distributions, this upper bound is considerably higher than the actual probability of this event. What are these probabilities for a Normal distribution?

This inequality will be used in the proof of the law of large numbers, one of our two main large sample theorems.

Probability bounds ..an example

Chebyshev's Inequality can, in principle, be used for computing bounds for the probabilities of certain events. We will only use it if we know the distribution of the random variable (the latter allow us to compute actual probabilities of tail events).

Example:

Let the density function of X be given by $f(x) = \frac{1}{(2\sqrt{3})} I_{(-\sqrt{3},\sqrt{3})}(x)$. In this case $\mu = 0$ and $\sigma^2 = \frac{(2\sqrt{3})^2}{12} = 1$. If $t = \frac{3}{2}$, then

$$Pr(|X - \mu| \geq \frac{3}{2}) = Pr(|X| \geq \frac{3}{2}) = 1 - \int_{-\frac{3}{2}}^{\frac{3}{2}} \frac{1}{2\sqrt{3}} dx = 1 - \frac{\sqrt{3}}{2} = .13$$

Chebyshev's inequality gives us $\frac{1}{a^2} = \frac{4}{9}$ which is much higher.

If $a = 2$, the exact probability is 0, while our bound is $\frac{1}{4}$.

The sample mean

A particularly important statistic is the sample mean:

$$\overline{X}_n = \frac{1}{n}(X_1 + \cdots + X_n)$$

Recall its mean and variance:

$$E(\overline{X}_n) = \frac{1}{n}\sum_{i=1}^{n} E(X_i) = \frac{1}{n}.n\mu = \mu$$

$$Var(\overline{X}_n) = \frac{1}{n^2}Var(\sum_{i=1}^{n} X_i) = \frac{1}{n^2}\sum_{i=1}^{n} Var(X_i) = \frac{1}{n^2}n\sigma^2 = \frac{\sigma^2}{n}$$

This tells us that the sample mean has the following properties:

has the same expectation as that of the population.

is more concentrated around its mean value μ than was the original distribution.

its variance is decreasing in sample size, n.

Sample size and precision of the sample mean

We can use Chebyshev's Inequality to ask how big a sample we should take, if we want to ensure a certain level of precision in our estimate of the sample mean.

Suppose the random sample is picked from a distribution which unknown mean and variance equal to 4.

We want to ensure an estimate which is within 1 unit of the real mean with probability .99. So we want $\Pr(|\bar{X} - \mu| \geq 1) \leq .01$.

Applying Chebyshev's Inequality we get $\Pr(|\bar{X} - \mu| \geq 1) \leq \frac{4}{n}$. Since we want $\frac{4}{n} = .01$ we take $n = 400$.

Note: We have not used any information on the distribution of $\bar{X}$ and this is probably much larger than what we actually need. That's the trade-off between the more efficient parametric procedures and more robust non-parametric ones.

Convergence of Real Sequences

We would like our estimators to be well-behaved. What does this mean?

One desirable property of an estimator is that our estimates get closer to the parameter that we are trying to estimate as our sample gets larger. We're going to make precise this notion of getting closer.

Recall that a sequence is just a function from the set of natural numbers $\mathbb{N}$ to any set A (Examples: $y_n = 2^n$, $y_n = \frac{1}{n}$)

A real number sequence $\{y_n\}$ converges to y if for every $\epsilon > 0$, there exists $N(\epsilon)$ for which $n \geq N(\epsilon) \implies |y_n - y| < \epsilon$. In such as case, we say that $\{y_n\} \to y$ Which of the above sequences converge?

If we have a sequence of functions $\{f_n\}$, the sequence is said to converge to a function f if $f_n(x) \to f(x)$ for all x in the domain of f.

In the case of matrices, a sequence of matrices converges if each the sequences formed by $(i,j)^{th}$ elements converge, i.e. $Y_n[i,j] \to Y[i,j]$.

Sequences of Random Variables

In a sequence of random variables, each term is a random variable

Examples: $Y_n = \frac{1}{n}\sum_{i=1}^{n} X_i$, where $X_i \sim N(\mu, \sigma^2)$, or $Y_n = \sum_{i=1}^{n} X_i$, where $X_i \sim \text{Bernoulli}(p)$

We need to modify our notion of convergence, since the sequence $\{Y_n\}$ no longer defines a given sequence of real numbers, but rather, many different real number sequences, depending on the realizations of $X_1, \ldots, X_n$.

Convergence questions can no longer be verified unequivocally since we are not referring to a given real sequence, but they can be assigned a probability of occurrence based on the probability space for random variables involved.

The most relevant random variable sequence for us is one in which the n^{th} element is our estimator for a sample size n. We want to know whether this estimator converges to the parameter we are estimating.

There are several types of random variable convergence discussed in the literature. We'll focus on two of these:

- Convergence in Distribution
- Convergence in Probability

Convergence in Distribution

Definition: *Let* $\{Y_n\}$ *be a sequence of random variables, and let* $\{F_n\}$ *be the associated sequence of cumulative distribution functions. If there exists a cumulative distribution function* F *such that* $F_n(y) \to F(y)$ *then* F *is called the limiting CDF of* $\{Y_n\}$*. Denoting by* Y *the r.v. with the distribution function* F*, we say that* Y_n *converges in distribution to the random variable* Y *and denote this by* $Y_n \xrightarrow{d} Y$*.*

The notation $Y_n \xrightarrow{d} F$ is also used to denote $Y_n \xrightarrow{d} Y \sim F$

Convergence in distribution holds if there is convergence in the sequence of densities ($f_n(y) \longrightarrow f(y)$) or in the sequence of MGFs ($M_{Y_n}(t) \longrightarrow M_Y(t)$). In some cases, it may be easier to use these to show convergence in distribution.

Result: *Let* $X_n \xrightarrow{d} X$*, and let the random variable* $g(X)$ *be defined by a function continous function* $g(.)$ *Then* $g(X_n) \xrightarrow{d} g(X)$

Example: Suppose $Z_n \xrightarrow{d} Z \sim N(0,1)$, then $2Z_n + 5 \xrightarrow{d} 2Z + 5 \sim N(5,4)$ (why?)

Convergence in Probability

This concept formalizes the idea that we can bring the outcomes of the random variable Y_n arbitrarily close to the outcomes of the random variable Y for large enough n.

Definition: *The sequence of random variables* $\{Y_n\}$ *converges in probability to the random variable* Y *iff*

$$\lim_{n\to\infty} P(|y_n - y| < \epsilon) = 1 \quad \forall \quad \epsilon > 0$$

We denote this by $Y_n \xrightarrow{p} Y$ *or* $\operatorname{plim} Y_n = Y$. *This justifies using outcomes of* Y *as an approximation for outcomes of* Y_n *since the two are very close for large* n.

Notice that while convergence in distribution is a statement about the distribution functions of Y_n and Y whereas convergence in probability is a statement about the joint density of *outcomes*, y_n and y.

Distribution functions of very different experiments may be the same: an even number on a fair die and a head on a fair coin have the same distribution function, but the outcomes of these random variables are unrelated.

Therefore $Y_n \xrightarrow{p} Y$ implies $Y_n \xrightarrow{d} Y$. In the special case where $Y_n \xrightarrow{d} c$, we also have $Y_n \xrightarrow{p} c$ and the two are equivalent.

The Weak Law of Large Numbers

Consider now the convergence of the random variable sequence whose n^{th} term is given by:

$$\overline{X}_n = \frac{1}{n}\sum_{i=1}^{n} X_i$$

WLLN: *Let $\{X_n\}$ be a sequence of i.i.d. random variables with finite mean μ and variance σ^2. Then* $\overline{X}_n \xrightarrow{p} \mu$.

Proof. Using Chebyshev's Inequality,

$$P(|\bar{X} - \mu| < \epsilon) \geq 1 - \frac{\sigma^2}{n\epsilon^2}$$

Hence

$$\lim_{n\to\infty} P(|\bar{X} - \mu| < \epsilon) = 1 \text{ or } plim\bar{X} = \mu$$

□

The WLLN will allow us to use the sample mean as an estimate of the population mean, under very general conditions.

Central Limit Theorems

Central limit theorems specify conditions under which sequences of random variables converge in distribution to known families of distributions.

These are very useful in deriving asymptotic distributions of test statistics whose exact distributions are either cumbersome or difficult to derive.

There are a large number of theorems which vary by the assumptions they place on the random variables (dependent or independent, identically or non-indentically distributed).

The Lindberg-Levy CLT: *Let* $\{X_n\}$ *be a sequence of i.i.d. random variables with* $EX_i = \mu$ *and* $var(X_i) = \sigma^2 \in (0, \infty) \ \forall i$. *Then*

$$\frac{\overline{X}_n - \mu}{\frac{\sigma}{\sqrt{n}}} = \frac{\sqrt{n}(\overline{X}_n - \mu)}{\sigma} \xrightarrow{d} N(0,1)$$

In words: Whenever a random sample of size n (large) is taken from any distribution with mean μ and variance σ^2, the sample mean $\overline{X}_n$ will have a distribution which is approximately normal with mean μ and variance $\frac{\sigma^2}{n}$.

Lindberg-Levy CLT..applications

Approximating Binomial Probabilities via the Normal Distribution: Let $\{X_n\}$ be a sequence of i.i.d. Bernoulli random variables. Then, by the LLCLT:

$$\frac{\frac{\sum X_i}{n} - p}{\sqrt{\frac{p(1-p)}{n}}} = \frac{\sum_{i=1}^{n} X_i - np}{\sqrt{np(1-p)}} \xrightarrow{d} N(0,1) \text{ and } \sum_{i=1}^{n} X_i \overset{a}{\sim} N(np, np(1-p))$$

In this case, $\sum_{i=1}^{n} X_i$ is of the form $aZ_n + b$ with $a = \sqrt{np(1-p)}$ and $b = np$ and since Z_n converges to Z in distribution, the asymptotic distribution $\sum_{i=1}^{n} X_i$ is normal with mean and variance given above (based on our results on normally distributed variables).

Approximating χ^2 Probabilities via the Normal Distribution: Let $\{X_n\}$ be a sequence of i.i.d. chi-square random variables with 1 degree of freedom. Using the additivity property of variables with gamma distributions, we have $\sum_{i=1}^{n} X_i \sim \chi^2_n$ Recall that the mean of gamma distribution is $\frac{\alpha}{\beta}$ and its variance is $\frac{\alpha}{\beta^2}$. For a χ^2_n random variable, $\alpha = \frac{\nu}{2}$ and $\beta = \frac{1}{2}$. Then, by the LLCLT:

$$\frac{\frac{\sum X_i}{n} - 1}{\sqrt{\frac{2}{n}}} = \frac{\sum_{i=1}^{n} X_i - n}{\sqrt{2n}} \xrightarrow{d} N(0,1) \text{ and } \sum_{i=1}^{n} X_i \overset{a}{\sim} N(n, 2n)$$