Topic 7: Convergence of Random Variables

Rohini Somanathan

Course 003, 2016

The Inference Problem

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

- We now look at how to generate information about the probability space by analyzing a sample of outcomes. This process is referred to as statistical inference.

- Inference procedures are parametric when we make assumptions about the probability space 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). If we make no such assumptions our procedures are nonparametric.

- Our focus in this course is on parametric inference. This involves both the estimation of population parameters and testing hypotheses about them.

Defining a Statistic

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

Notice that:

- a statistic is itself a random variable

- we've considered several functions of random variables, whose distributions are well defined such as :

$Y = \frac{X - \mu}{\sigma}$ *where* $X \sim N(\mu, \sigma^2)$. We showed that $Y \sim N(0, 1)$.

$Y = \sum_{i=1}^{n} X_i$, where each X_i has a bernoulli distribution with parameter p was shown to have a binomial distribution with parameters n and p.

$Y = \sum_{i=1}^{n} X_i^2$ where each X_i has a standard normal distribution was shown to have a χ_n^2 distribution etc...

- Only some of these functions of random variables are statistics. This distinction is important because statistics 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.

Rohini Somanathan

Markov's Inequality

We begin with some useful inequalities which provide us with distribution-free bounds on the probability of certain events and are useful in proving the law of large numbers, one of our two main large sample theorems.

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 $t > 0$,

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

Proof. (for discrete distributions)

$E(X) = \sum_x x f(x) = \sum_{x < t} x f(x) + \sum_{x \geq t} x f(x)$ All terms in these summations are non-negative by assumption, so we have

$$E(X) \geq \sum_{x \geq t} x f(x) \geq \sum_{x \geq t} t f(x) = t P(X \geq t)$$

$\square$

This inequality obviously holds for $t \leq E(X)$ (why?). Its main interest is in bounding the probability 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. This is true irrespective of the distribution of X- this is what makes the result powerful.

Chebyshev's Inequality

This is a special case of Markov's inequality and relates the variance of a distribution to the probability associated with deviations from the mean.

Chebyshev's Inequality: Let X be a random variable with finite variance σ^2 and mean μ. Then, for every $t > 0$,

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

or equivalently,

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

Proof. Use Markov's inequality with $Y = (X - \mu)^2$ and use t^2 in place of the constant t. Then Y takes only non-negative values and $E(Y) = Var(X) = \sigma^2$. $\square$

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?

Probability bounds ..an example

Chebyshev's Inequality can, in principle be used for computing bounds for the probabilities of certain events. In practice this is not often used because the bounds it provides are much larger than the actual probabilities of tail events as seen in the following 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 $\sigma^2 = \frac{(b-a)^2}{12} = 1$. 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 $\frac{1}{t^2} = \frac{4}{9}$ which is much higher. If $t = 2$, the exact probability is 0, while our bound is $\frac{1}{4}$.

The sample mean

Recall the mean and variance of the sample mean:

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

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

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

So we know the following about the sample mean

- Its expectation is equal to that of the population.

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

- The larger the sample, the lower the variance of $\overline{X}_n$.

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 and 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.

- This calculation does not use any information on the distribution of $\bar{X}$ and therefore often gives us a much larger number than we would get if this information was available. That's the trade-off between the more efficient parametric procedures and more robust non-parametric ones.

Sample size and precision: Bernoulli example

- We want to pick a random sample size of size n from a Bernoulli distribution and use the sample mean to estimate the parameter p. How big should n be for the sample mean of $(X_1, \ldots X_n)$ to lie within .1 of the population mean with probability equal to k?

- If we don't use the distribution of X_i and apply Chebyshev's Inequality, we get:
  $P(|\overline{X}_n - \mu| \le .1) \ge 1 - \frac{\sigma^2_{\overline{X}_n}}{\frac{1}{100}}$ or $P(|\overline{X}_n - \mu| \le .1) \ge 1 - \frac{p(1-p)}{n} * 100$
  Setting the RHS equal to k, we get:

$$n = \frac{100p(1-p)}{1-k}$$

- With $\sigma^2_{\overline{X}_n} = \frac{1}{4}$, ($p = \frac{1}{2}$) and $k = .7$, we choose $n = 84$ so that our sample mean to lies in the interval $[.4, .6]$, with probability equal to .7. What if $p \ne \frac{1}{2}$?

- Now let's use the fact that X_i is Bernoulli. We want

$$P(p - .1 \le \frac{\sum X_i}{n} \le p + .1) \ge k$$
$$P(np - .1n \le \sum X_i \le np + .1n) \ge k$$

- When $n = 15$ and $p = .5$, $np = 7.5$ and for the Binomial distribution $F(9) - F(6) = .7$ so we can use this sample size for the desired level of precision in our estimate of $\overline{X}_n$.

Convergence of Real Sequences

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

- One desirable property 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

- A sequence of random variables is a sequence for which the set A is a collection of random variables and the function defining the sequence puts these random variables in a specific order.

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

- We now 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.

- 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 $F_n(y) \rightarrow F(y)$ 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 $\text{plim } Y_n = Y$. 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

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

$\square$

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 $\mathbf{E}X_i = \mu$ 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:

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

In this case, $\sum\limits_{i=1}^{n} X_i$ 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 $\sum\limits_{i=1}^{n} X_i$ 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 $\sum\limits_{i=1}^{n} X_i \sim \chi^2_n$ 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:

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