BIOS 660/BIOS 672 (3 Credits): Probability and Statistical Inference I

Jianwen Cai

https://sakai.unc.edu/portal/site/bios660-bios672-3-credits

Notes 16

Recall: Chebychev Inequality

Let X be a random variable and let $g(x)$ be a non-negative function. Then for any $r > 0$,

$$P[g(X) \geq r] \leq \frac{\mathsf{E}g(X)}{r}$$

Proof:

$$\begin{aligned}
\mathsf{E}g(X) &= \int_{-\infty}^{\infty} g(x)\, f_X(x)\, dx \\
&\geq \int_{\{x:g(x)\geq r\}} g(x)\, f_X(x)\, dx \\
&\geq r \int_{\{x:g(x)\geq r\}} f_X(x)\, dx \\
&= r\, P\{g(X) \geq r\}
\end{aligned}$$

Corollaries

1. Suppose X is a non-negative and g is a positive, non-decreasing function, with $\mathsf{E}[g(X)] < \infty$. Then

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

2. Suppose g is a non-negative symmetric function, increasing on $\mathbb{R}^+$, with $\mathsf{E}[g(X)] < \infty$. Then

$$P\{|X| \geq a\} \leq \frac{\mathsf{E}[g(X)]}{g(a)}$$

Proof: $P\{X \geq a\} = P\{g(X) \geq g(a)\}$, so the results follow from the inequality on the previous slides.

Special cases

Provided that the appropriate expectations exist, for $a > 0$:

$$X \geq 0: \qquad P\{X \geq a\} \leq \frac{\mathsf{E}(e^{tX})}{e^{ta}} \tag{1}$$

$$X \in \mathbb{R}: \qquad P\{|X| \geq a\} \leq \frac{\mathsf{E}(|X|)}{a} \tag{2}$$

$$X \in \mathbb{R},\ p > 0: \qquad P\{|X| \geq a\} \leq \frac{\mathsf{E}(|X|^p)}{a^p} \tag{3}$$

$$\sigma^2 = \mathsf{Var}(X): \qquad P\{|X - \mathsf{E}X| \geq a\sigma\} \leq \frac{1}{a^2} \tag{4}$$

Note:

- (1) is called Chernoff bound, useful when the mgf is easier to compute than the cdf.
- (3) is sometimes called Markov's inequality.
- (4) is sometimes called Chebychev's inequality.

Functional inequalities

First a couple of useful items:

- *L^p spaces*:
  The space called L^p consists of all random variables whose p^{th} absolute power is integrable, i.e., $E(|X|^p) < \infty$.
- *Triangle inequality*:
  For two real or complex numbers a and b,

  $$|a + b| \leq |a| + |b|$$

  Proof: HW

Convex functions

Definition: Let I be an interval on $\mathbb{R}$. A function $g : I \to R$ is *convex* on I if for any $\lambda \in [0, 1]$, and any points x and y in I

$$g\left[\lambda x + (1-\lambda)y\right] \leq \lambda g(x) + (1-\lambda)g(y)$$

Properties:

- A differentiable function g is convex iff it lays above all its tangents.
- A twice differentiable function g is convex iff its second derivative is non-negative.

Definition: Let I be an interval on $\mathbb{R}$. A function $g : I \to R$ is *concave* on I if $-g$ is convex on I.

Examples:

- $g(x) = x^2$ is a convex function for all x.
- $g(x) = \log(x)$ is concave for $x > 0$.

Jensen's Inequality

Let $X \in L^1$ and $g(x)$ be a convex function where $\mathsf{E}[g(X)]$ exists. Then,

$$\mathsf{E}[g(X)] \geq g[\mathsf{E}X]$$

with equality if and only if for every line $a + bx$ tangent to $g(x)$ at $x = \mathsf{E}X$, $P[g(X) = a + bX] = 1$.

Examples:

$$\begin{aligned} g(x) = x^2 \quad &\to \quad \mathsf{E}(X^2) \geq \mathsf{E}^2(X) \\ g(x) = 1/x,\ x > 0 \quad &\to \quad \mathsf{E}(1/X) \geq 1/\mathsf{E}(X),\ X > 0 \end{aligned}$$

Note: The direction of the inequality is reversed if g is concave.

Jensen's Inequality (proof)

Let $l(x) = a + bx$ be the tangent line to $g(x)$ at $g(\mathsf{E}X)$. Then

$$\begin{aligned} \mathsf{E}g(X) &\geq \mathsf{E}(a + bX) \\ &= a + b\mathsf{E}X \\ &= l(\mathsf{E}X) \\ &= g(\mathsf{E}X) \end{aligned}$$

Example

Let $a_1, \ldots, a_n > 0$. Then

$$\left(\frac{1}{n}\sum_{i=1}^{n}\frac{1}{a_i}\right)^{-1} \leq \left(\prod_{i=1}^{n} a_i\right)^{1/n} \leq \frac{1}{n}\sum_{i=1}^{n} a_i$$

Proof: Let X be a rv such that $P(X = a_i) = 1/n$. Since $\log(x)$ is concave,

$$\begin{aligned} \log\left(\prod_{i=1}^{n} a_i\right)^{1/n} &= \frac{1}{n}\sum_{i=1}^{n}\log(a_i) \\ &= \mathsf{E}(\log(X)) \\ &\leq \log(\mathsf{E}(X)) \\ &= \log\left(\frac{1}{n}\sum_{i=1}^{n} a_i\right) \end{aligned}$$

The proof of the second inequality is similar.

Young's Inequality

Let $a, b > 0$ and $p, q > 1$ with $1/p + 1/q = 1$. Then

$$\frac{a^p}{p} + \frac{b^q}{q} \geq ab$$

With equality only if $a^p = b^q$.

Proof: Consider

$$g(a) = \frac{1}{p}a^p + \frac{1}{q}b^q - ab$$

To minimize $g(a)$, differentiate and set equal to 0:

$$\frac{d}{da}g(a) = 0 \to a^{p-1} - b = 0 \to a = b^{1/(p-1)}.$$

Since $g(b^{1/(p-1)}) = 0$, the result follows.

Hölder's inequality

Suppose $X \in L^p$, $Y \in L^q$ where $p, q > 1$ and $1/p + 1/q = 1$. Then

$$\mathsf{E}|XY| \leq [\mathsf{E}|X|^p]^{1/p}\ [\mathsf{E}|Y|^q]^{1/q}$$

with equality if $X^p = cY^q$ for some $c \in \mathbb{R}$.

Proof: Let

$$a = \frac{|X|}{(\mathsf{E}|X|^p)^{1/p}} \quad \text{and} \quad b = \frac{|Y|}{(\mathsf{E}|Y|^q)^{1/q}}$$

By Young's Inequality,

$$\frac{|X|^p}{p\mathsf{E}|X|^p} + \frac{|Y|^q}{q\mathsf{E}|Y|^q} \geq \frac{|XY|}{(\mathsf{E}|X|^p)^{1/p}(\mathsf{E}|Y|^q)^{1/q}}$$

The result follows by taking the expected value of both sides and noting that the expected value of the left-hand side is 1.

Corollaries

- **Cauchy-Schwartz inequality:** Special case where $p = q = 2$.

$$\mathsf{E}|XY| \leq [\mathsf{E}|X|^2]^{1/2}[\mathsf{E}|Y|^2]^{1/2} = \sqrt{\mathsf{E}[X^2]\mathsf{E}[Y^2]}$$

  with equality if $X = cY$ for some $c \in \mathbb{R}$.

- **Lyapunov's inequality:** For $1 \leq r \leq s$ and $X \in L^s$,

$$[\mathsf{E}|X|^r]^{1/r} \leq [\mathsf{E}|X|^s]^{1/s}$$

  Proof:
  Apply Hölder's inequality to $|X|^r$ with $Y = 1$ and $p = s/r$.

Application of Cauchy-Schwartz:

Let ρ represent the correlation between two rvs X and Y, ie,

$$\rho = \frac{\mathsf{Cov}(X, Y)}{\sqrt{\mathsf{Var}(X)\,\mathsf{Var}(Y)}}.$$

Then, $|\rho| \leq 1$, with equality iff $Y - \mu_Y = c(X - \mu_X)$ for some $c \in \mathbb{R}$.

Proof: By the Cauchy-Schwartz Inequality,

$$\mathsf{E}|(X - \mu_x)(Y - \mu_Y)| \leq \{\mathsf{E}(X - \mu_X)^2\}^{\frac{1}{2}}\{\mathsf{E}(Y - \mu_Y)^2\}^{\frac{1}{2}}.$$

Squaring both sides, we obtain

$$(\mathsf{Cov}(X, Y))^2 \leq \sigma_X^2 \sigma_Y^2.$$

Thus, $|\rho| \leq 1$.

Minkowski's inequality

Suppose $X, Y \in L^p$, $p \geq 1$. Then $(X+Y) \in L^p$ and

$$[\mathsf{E}|X+Y|^p]^{1/p} \leq [\mathsf{E}|X|^p]^{1/p} + [\mathsf{E}|Y|^p]^{1/p}$$

Proof:
For $p=1$, the proof follows almost immediately from the triangle inequality (HW).
The case $p > 1$ is more complicated. Consider

$$\begin{aligned}
\mathsf{E}|X+Y|^p &= \mathsf{E}\left(|X+Y|\,|X+Y|^{p-1}\right) \\
&\leq \mathsf{E}\left(|X|\,|X+Y|^{p-1}\right) + \mathsf{E}\left(|Y|\,|X+Y|^{p-1}\right) \\
&\leq [\mathsf{E}|X|^p]^{1/p}\left[\mathsf{E}|X+Y|^{(p-1)q}\right]^{1/q} \\
&+ [\mathsf{E}|Y|^p]^{1/p}\left[\mathsf{E}|X+Y|^{(p-1)q}\right]^{1/q}
\end{aligned}$$

where the last row follows by Hölder's inequality with $1/p + 1/q = 1$.

Order Statistics (*C-B*, Section 5.4; *Gut*, Chapter IV)

Distribution of the Maximum

The *cdf* of $Z = \max(Y_1, \ldots, Y_n)$ is

$$\begin{aligned}
F_Z(z) &= Pr\{Z \leq z\} \\
&= Pr\{Y_1 \leq z, Y_2 \leq z, \ldots, Y_n \leq z\} \\
&= \prod_{j=1}^{n} Pr\{Y_j \leq z\} \qquad \text{(indep)} \\
&= F_Y(z)^n \qquad \text{(ident. distrib.)}
\end{aligned}$$

and thus the density (or pmf) is:

$$f_Z(z) = nF_Y(z)^{n-1} f_Y(z)$$

Distribution of the Minimum

Similarly, consider $W = \min(Y_1, Y_2, \ldots, Y_n)$.

$$\begin{aligned} 1 - F_W(w) &= Pr\{W > w\} \\ &= Pr\{Y_1 > w, Y_2 > w, \ldots, Y_n > w\} \\ &= \prod_{j=1}^{n} Pr\{Y_j > w\} \qquad \text{(indep)} \\ &= (1 - F_Y(w))^n \qquad \text{(ident. distrib)} \end{aligned}$$

Thus

$$F_W(w) = 1 - (1 - F_Y(w))^n$$

and the corresponding density (or pmf) is:

$$f_W(w) = n(1 - F_Y(w))^{n-1} f_Y(w)$$

Example

Suppose $Y_i \sim$ *exp*(λ):

$$f_Y(y) = \lambda e^{-\lambda y} \qquad \text{for } y > 0, \quad 1 - F(y) = e^{-\lambda y}$$

Maximum:

$$f_Z(z) = n(1 - e^{-\lambda z})^{n-1} \lambda e^{-\lambda z} = n\lambda e^{-\lambda z}(1 - e^{-\lambda z})^{n-1}$$

Minimum:

$$f_W(w) = n(e^{-\lambda w})^{n-1} \lambda e^{-\lambda w} = \quad (n\lambda) e^{-n\lambda w}$$

$\Rightarrow$ exponential with parameter $n\lambda$

The next obvious statistic is the range defined as the difference of the maximum and the minimum, but to get its distribution we need the joint distribution of the maximum and the minimum.

Order Statistics

Let $Y_1, Y_2, \ldots, Y_n$ be *iid* with *pdf* $f_Y(x)$.
Order the observations; i.e.

$$Y_{(1)} \le Y_{(2)} \le \cdots \le Y_{(n)}$$

The $Y_{(i)}$ are called *order statistics*. For example, the *minimum* is $Y_{(1)}$ and the *maximum* is $Y_{(n)}$.

We are interested in finding the distribution of an arbitrary $Y_{(i)}$, as well as the joint distributions of sets of $Y_{(i)}$s and $Y_{(j)}$s.

e.g. Range $= Y_{(n)} - Y_{(1)}$
or interquartile range, or joint of median and interquartile range, etc....

rth order statistic

We need to find the density of $Y_{(r)}$ at a value y:

$$\overbrace{\qquad}^{r-1}\ \overbrace{\qquad\qquad}^{1}\ \overbrace{\qquad\qquad}^{n-r}$$
$$y \qquad\qquad y+dy$$

Consider 3 intervals $(-\infty, y), [y, y+dy), [y+dy, \infty)$. The number of observations in each of the intervals follows the tri-nomial distribution

$$f(s_1, s_2, s_3) = \frac{n!}{s_1!s_2!s_3!}\ p_1^{s_1}\ p_2^{s_2}\ p_3^{s_3}$$

The event that $y \le Y_{(r)} < y + dy$ is the event that we have
$(r-1)$ observations are less than y
$(n-r)$ observations are greater than y
1 observation is in interval; $y, y+dy$

In the trinomial distribution, this corresponds to

$$s_1 = r-1, \quad s_2 = 1, \quad s_3 = n-r$$
$$p_1 = F_Y(y), \quad p_2 = f_Y(y)dy, \quad p_3 = 1 - F_Y(y+dy)$$

cont.

Taking the limit as $dy \to 0$, we get:

$$f_{Y_{(r)}}(y) = \frac{n!}{(r-1)!(n-r)!} F_Y(y)^{r-1}[1 - F_Y(y)]^{n-r} f_Y(y)$$
$$= \frac{F_Y(y)^{r-1}[1 - F_Y(y)]^{n-r} f_Y(y)}{B(r, n-r+1)}$$

Gut has a more formal derivation based on deriving the joint density of the order statistics, then integrating out all but the r^{th} order statistic. (see also Casella and Berger, p.228).

Example

$F_Y(y) = y$, that is, $Y \sim$ *Uniform* $(0, 1)$

$$f_{Y_{(r)}}(y) = \frac{y^{r-1}(1-y)^{n-r}}{B(r, n-r+1)}$$

hence, $Y_{(r)}$ follows a *Beta Distribution* with parameters r and $n - r + 1$.

Note: $E[Y_{(r)}] = \frac{r}{n+1}$

Distribution of the median

To simplify, suppose the sample size is odd, $n = 2m + 1$, so that the median corresponds to the $(m + 1)^{\text{th}}$ order statistic.
Setting $r = m + 1$ and $n = 2m + 1$ into the formula derived earlier

$$f_{\text{med}}(y) = f_{Y_{(m+1)}}(y) = \frac{F_Y(y)^m (1 - F_Y(y))^m f_Y(y)}{B(m + 1, m + 1)}$$

If the density $f_Y(y)$ is symmetric around zero, so that E$Y = 0$, then

$$F_Y(-y) = 1 - F_Y(y)$$

and so the density of the median is also symmetric around zero, so that

$$\text{E}[\text{med}(Y_1, \ldots, Y_n)] = 0$$

Joint distribution of $Y_{(r)}, Y_{(s)}$, $r < s$

	Interval	Prob.	# obs $= s_i$
1.	$(-\infty, u]$	$p_1 = F_Y(u)$	$r - 1$
2.	$(u, u + du]$	$p_2 = f_Y(u)du$	1
3.	$(u + du, v]$	$p_3 = F_Y(v) - F_Y(u + du)$	$s - r - 1$
4.	$(v, v + dv]$	$p_4 = f_Y(v)dv$	1
5.	$(v + dv, \infty)$	$p_5 = 1 - F_Y(v + dv)$	$n - s$

This is a multinomial with 5 cells:

$$f(s_1, \ldots, s_5) = \frac{n!}{\prod s_i!} \prod_{i=1}^{5} p_i^{s_i}$$

cont.

Taking limits as du and dv approach 0, we get

$$f_{Y_{(r)},Y_{(s)}}(u,v) = \frac{n!}{(r-1)!(s-r-1)!(n-s)!} F_Y(u)^{r-1}$$
$$\times \left[F_Y(v) - F_Y(u)\right]^{s-r-1} (1 - F_Y(v))^{n-s} f_Y(u) f_Y(v)$$

Example: Suppose $F_Y(x) = x$ (Uniform)

$$\begin{aligned} f_{Y_{(r)},Y_{(s)}}(u,v) &= \frac{n!}{(r-1)!(s-r-1)!(n-s)!} \\ &\quad \times u^{r-1}(v-u)^{s-r-1}(1-v)^{n-s} \end{aligned}$$

for $u < v$.

Joint distribution of all order statistics

Multinomial with $2n+1$ cells, where we have one observation in each interval $[u_i, u_i + du_i)$, $i = 1, \ldots, n$, and zero on the others.

$$f_{Y_{(1)},Y_{(2)},\ldots,Y_{(n)}}(u_1, \ldots, u_n) = n! \prod_{i=1}^{n} f_Y(u_i)$$

for $u_1 < \cdots < u_n$.

Example: Suppose $F_Y(x) = x$ (Uniform)

$$f_{X_{(1)},X_{(2)},\ldots,X_{(n)}}(u_1, \ldots, u_n) = n! \qquad u_1 < \cdots < u_n$$

Distribution of the range

Setting $r = 1$ and $s = n$ in the joint dist. of the r$^{\text{th}}$ and s$^{\text{th}}$ order statistics gives the joint dist. of the *min* and *max*:

$$f_{Y_{(1)},Y_{(n)}}(u,v) = \frac{n!}{(n-2)!}[F_Y(v) - F_Y(u)]^{n-2} f_Y(u) f_Y(v)$$

Now, do a transformation to $R = Y_{(n)} - Y_{(1)}$ and $W = Y_{(1)}$. Note that the Jacobian is 1. What is the range?

Hence,

$$f_{W,R}(w,r) = n(n-1)[F_Y(w+r) - F_Y(w)]^{n-2} f_Y(w) f_Y(w+r)$$

The density of R can be obtained by integrating out W:

$$f_R(r) = \int_{-\infty}^{\infty} f_{W,r}(w,r)\, dw$$

Example:

Suppose $Y \sim U[0,1]$, i.e. $F_Y(x) = x$

$$\begin{aligned} f_R(r) &= \int_0^{1-r} n(n-1) r^{n-2}\, dw \\ &= n(n-1) r^{n-2}(1-r) \end{aligned}$$

Note that R has a Beta distribution.

$$\begin{aligned} \mathsf{E}(R) &= n(n-1)\int_0^1 r \cdot r^{n-2}(1-r)dr \\ &= n(n-1)\left[\frac{1}{n} - \frac{1}{n+1}\right] \\ &= \frac{n-1}{n+1} \end{aligned}$$

What happens when $n = 2$ and $n \to \infty$?

Agenda

In the last part of the course, we discuss

- Convergence of random variables. Several different kinds
  - Convergence in probability
  - Almost sure convergence
  - Convergence in distribution
  - Convergence in L^p
  - Complete convergence
- Weak law of large numbers
- Strong law of large numbers
- Central limit theorems

The moment inequalities will be useful in proving these results.

Material is in *C-B*, Section 5.5, and *Gut*, Chapter VI.

Modes of Convergence

There are five modes of convergence. If $X_n \to X$ by any of these modes, then X is unique (see Section 2, Chapter VI of *Gut*).

1. *Convergence in Probability* $\quad X_n \xrightarrow{P} X$

   For any $\epsilon > 0$, $\quad \lim_{n\to\infty} P\{|X_n - X| < \epsilon\} = 1$

   Or equivalently,
   for any $\epsilon > 0$, $\quad \lim_{n\to\infty} P\{|X_n - X| > \epsilon\} = 0$

2. *Convergence "almost surely" (a.s.)*, denoted $X_n \xrightarrow{a.s.} X$.
   Also called *Convergence with Prob. 1*

   For any $\epsilon > 0$, $\quad P\{\lim_{n\to\infty} |X_n - X| < \epsilon\} = 1$
   Or
   $$P\Big\{\lim_{n\to\infty} X_n = X\Big\} = 1$$
   Or
   $$P\Big\{\omega : \lim_{n\to\infty} X_n(\omega) = X(\omega)\Big\} = 1$$

Notes

The distinction between *Convergence almost surely* and *Convergence in probability* is subtle.

We'll see how the Markov inequality and Chebychev's inequality can often be used to establish convergence in probability. Establishing convergence a.s. is often more difficult.

Almost sure convergence is stronger than convergence in probability. Or equivalently,

$$X_n \xrightarrow{a.s.} X \Rightarrow X_n \xrightarrow{p} X.$$

Convergence in Distribution

3. *Convergence in Distribution* $X_n \xrightarrow{d} X$

   Also called *convergence in law* or *weak convergence*.

   $\lim_{n\to\infty} F_{X_n}(x) = F_X(x)$ for all <u>continuity</u> points of $F_X(x)$

Notes:

- Recall that cdfs can have at most a countable number of discontinuities.
- Theorem (no proof):

$$X_n \xrightarrow{d} X \quad \Leftrightarrow \quad \forall \text{ bounded continuous functions } g, \; \mathsf{E}g(X_n) \to \mathsf{E}g(X) \text{ as } n \to \infty.$$

Convergence in Distrib. cont.

- Convergence in distribution is the weakest convergence and does not imply the other modes. E.g. $X_n \sim N(0,1)$ and $Y = -X$.
- An exception is the following important special case:

Suppose $X_n \xrightarrow{d} X$ where X has the degenerate distribution (i.e. $P\{X = a\} = 1$). Then, $X_n \xrightarrow{d} X \;\Rightarrow\; X_n \xrightarrow{P} X$.

Proof:

$$\begin{aligned}
P\{|X_n - a| < \epsilon\} &= F_{X_n}(a+\epsilon) - F_{X_n}(a-\epsilon) \\
\lim_{n\to\infty} P\{|X_n - a| < \epsilon\} &= F(a+\epsilon) - F(a-\epsilon) \\
&= 1 - 0 = 1
\end{aligned}$$

Other modes of convergence

4. *Convergence in r^{th} mean* $(r \geq 1)$ $X_n \xrightarrow{r} X$
   If $E|X_n^r| < \infty$ for all n and
   $$\lim_{n\to\infty} (E|X_n - X|^r) = 0.$$
   Also called *convergence in L^r* and sometimes referred to as *convergence in the L^r* norm. (Some books use L^p)

5. *Complete convergence* (see Chapter VI, section 4 in *Gut*), defined as,
   $$\sum_{n=1}^{\infty} P(|X_n - X| > \epsilon) < \infty \qquad \forall\, \epsilon > 0,$$
   is slightly stronger than a.s. convergence, but much easier to verify.

   Hence, it sometimes provides a relatively easy way to establish a.s. convergence. Some books use this as the definition of a.s. convergence.

Example 1

1. Let X_n be *gamma*(n, n) Then $X_n \xrightarrow{p} 1$.
   Proof:
   Since $E(X_n) = 1$ and Var$(X_n) = 1/n$, we can apply Chebychev's inequality to obtain

$$P(|X_n - 1| > \epsilon) \leq \frac{1}{n\epsilon^2} \to 0 \quad \text{as } n \to \infty.$$

   Therefore $X_n \xrightarrow{p} 1$.

Example 2

2. Suppose $X_n \sim$ *binom*$(n, \lambda/n)$. Then $X_n \xrightarrow{d} X$, where X has a *Poisson*(λ) distribution.
   Proof:

$$F_{X_n}(x) = \sum_{y=0}^{x} \binom{n}{y} \left(\frac{\lambda}{n}\right)^y \left(1 - \frac{\lambda}{n}\right)^{n-y} \to \sum_{y=0}^{x} e^{-\lambda} \frac{\lambda^y}{y!},$$

   as $n \to \infty$. (We saw this before). The RHS is the distribution function of the Poisson with parameter λ.

Example 3

3. Let $X_2, X_3 ...$ be a sequence of binary random variables defined by

$$P(X_n = 1) = 1 - \frac{1}{n} \quad \text{and} \quad P(X_n = n) = \frac{1}{n}$$

If we choose an ϵ smaller than 1, then

$$P(|X_n - 1| > \epsilon) = P(X_n = n) = 1/n \to 0,$$

hence $X_n \xrightarrow{P} 1$.
It turns out that X_n does not converge to 1 almost surely (see page 156 in *Gut*).

Example 4

4. Let $X_2, X_3 ...$ be a binary random variables defined by

$$P(X_n = 1) = 1 - \frac{1}{n^2} \quad \text{and} \quad P(X_n = n) = \frac{1}{n^2}$$

Now, for ϵ small enough,

$$\sum_{n=1}^{\infty} P(|X_n - 1| > \epsilon) = \sum_{n=1}^{\infty} \frac{1}{n^2}$$

which converges (the series $\sum_{n=1}^{\infty} \frac{1}{n^k}$ converges for $k > 1$). I.e., $X_n \to X$ in complete convergence, with $X = 1$. Hence $X_n \xrightarrow{a.s.} 1$.

Relationships among convergence modes

$$\begin{array}{ccccc} X_n \xrightarrow{Compl} X \;\Rightarrow\; X_n \xrightarrow{a.s.} X & \searrow & & & \\ & & X_n \xrightarrow{P} X & \Rightarrow & X_n \xrightarrow{d} X \\ X_n \xrightarrow{r} X & \nearrow & & & \end{array}$$

A silly mneumonic is *All Probabilists Drink*.

Also: If $r \geq s \geq 1$

$$X_n \xrightarrow{r} X \Longrightarrow X_n \xrightarrow{s} X.$$

(Try to prove this one).

In probability and distribution

Theorem: If $X_n \xrightarrow{P} X \Rightarrow X_n \xrightarrow{d} X$

Proof:
For any $\epsilon > 0$:

$$\begin{aligned} F_{X_n}(x) = P\{X_n \leq x\} &= P\{X_n \leq x \cap |X - X_n| \leq \epsilon\} \\ &\quad + P\{X_n \leq x \cap |X_n - X| > \epsilon\} \\ &\leq P\{X \leq x + \epsilon\} + P\{|X_n - X| > \epsilon\} \end{aligned}$$

because

$$\begin{aligned} \{|X - X_n| \leq \epsilon\} &= \{-\epsilon \leq X - X_n \leq \epsilon\} \\ &\subset \{X - X_n \leq \epsilon\} = \{X \leq X_n + \epsilon\} \end{aligned}$$

and $\{X_n \leq x\} \cap \{X \leq X_n + \epsilon\} \subset \{X \leq x + \epsilon\}$.

cont.

Hence,

$$F_{X_n}(x) \le F_X(x+\epsilon) + P\{|X - X_n| > \epsilon\} \tag{5}$$

and as $n \to \infty$, this implies

$$\limsup_{n\to\infty} F_{X_n}(x) \le F_X(x+\epsilon)$$

Now, interchange the roles of X_n and X in (5) and repeat to get

$$F_X(x) \le F_{X_n}(x+\epsilon) + P\{|X_n - X| > \epsilon\}$$

but apply inequality to $x = x - \epsilon$ instead of x, yielding

$$F_X(x-\epsilon) \le F_{X_n}(x) + P\{|X_n - X| > \epsilon\}$$

or

$$F_X(x-\epsilon) - P\{|X_n - X| > \epsilon\} \le F_{X_n}(x)$$

cont.

As $n \to \infty$, this implies

$$F_X(x-\epsilon) \le \liminf_{n\to\infty} F_{X_n}(x)$$

Putting these together, we have

$$F_X(x-\epsilon) \le \liminf_{n\to\infty} F_{X_n}(x) \le \limsup_{n\to\infty} F_{X_n}(x) \le F_X(x+\epsilon)$$

for all $\epsilon > 0$. Therefore, for all x where $F_X(x)$ is continuous,

$$\lim_{n\to\infty} F_{X_n}(x) = F_X(x).$$

Note: If $F_X(x)$ is not continuous at x, then all we can claim is

$$F_X(x-) \le \liminf_{n\to\infty} F_{X_n}(x) \le \limsup_{n\to\infty} F_{X_n}(x) \le F_X(x)$$

rth moment and in Probability

Theorem: $X_n \xrightarrow{r} X \Rightarrow X_n \xrightarrow{P} X$

Proof: By Markov's inequality,

$$\begin{aligned} P(|X_n - X| > \epsilon) &= P(|X_n - X|^r > \epsilon^r) \\ &\leq \frac{E(|X_n - X|^r)}{\epsilon^r} \to 0 \end{aligned}$$

Example: Let $Y_1, \ldots, Y_n$ be iid with common mean μ and variance σ^2. Let $\bar{Y}_n = \sum_{i=1}^n Y_i/n$. Then

$$E(\bar{Y}_n - \mu)^2 = \text{var}(\bar{Y}) = \frac{\sigma^2}{n} \to 0$$

Therefore $\bar{Y}_n \xrightarrow{r=2} \mu$ and so $\bar{Y}_n \xrightarrow{P} \mu$.

This result is the *weak law of large numbers*.

Convergence properties

Convergence in probability

Theorem: If $X_n \xrightarrow{P} X$ and $Y_n \xrightarrow{P} Y$, then

1. $X_n + Y_n \xrightarrow{P} X + Y$
2. $X_n Y_n \xrightarrow{P} XY$
3. If $g(x)$ is a continuous function, then $g(X_n) \xrightarrow{P} g(X)$

Proof:

Slutsky's Theorem

Also known as *Cramer's Theorem* - **VERY VERY USEFUL**

Theorem: If $X_n \xrightarrow{d} X$ and $Y_n \xrightarrow{P} a$, where a is a constant, then

1. $X_n + Y_n \xrightarrow{d} X + a$
2. $Y_n X_n \xrightarrow{d} aX$

Proof: Homework

Example

Let $X_1, \ldots, X_n$ be iid with mean μ, variance σ^2, and finite moments up to fourth order. The Central Limit Theorem (CLT) says that

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

But the empirical variance is a consistent estimator of σ^2, i.e. $S_n^2 \xrightarrow{P} \sigma^2$. Therefore

$$\sqrt{n}\frac{\bar{X}_n - \mu}{S_n} \xrightarrow{d} N(0,1)$$

by Slutsky's Thereom.

This is useful for constructing confidence intervals for μ.

Example

Show that a t-distribution with n degrees of freedom converges in distribution to the standard normal as $n \to \infty$.

Proof: Let $Y_n \sim$ ChiSquare(n). Then $Y_n/n \to 1$ by the Weak Law of Large Number (WLLN).

By Slutsky's Theorem, if $X \sim$ Normal$(0, 1)$, then

$$\frac{X}{\sqrt{\frac{Y_n}{n}}} \xrightarrow{d} \text{Normal}(0, 1)$$

Convergence in distribution

Theorem: Suppose $X_n \xrightarrow{d} X$ and $Y_n \xrightarrow{d} Y$. Suppose further that X_n and Y_n are independent for all n, and that X and Y are independent. Then

$$X_n + Y_n \xrightarrow{d} X + Y$$

Proof: Omitted (use characteristic functions)

Example

Let $X_n \sim Bin(n_x, p_x)$ with $n_x p_x \to \lambda_x$ as $n_x \to \infty$.
Let $Y_n \sim Bin(n_y, p_y)$ with $n_y p_y \to \lambda_y$ as $n_y \to \infty$, indep. of X_n.
Then

$$X_n \xrightarrow{d} Po(\lambda_x), \qquad Y_n \xrightarrow{d} Po(\lambda_y)$$

and

$$X_n + Y_n \xrightarrow{d} Po(\lambda_x + \lambda_y)$$

The Delta Method

Approximate mean and variance

Suppose we know the distribution of X and want to get the distribution of $Y = g(X)$. The general method is to use the Jacobian transformation. But if the distribution of X is "well concentrated" around its mean $\mu = \mathsf{E}X$, we can approximate the mean and variance of Y as follows.

The Taylor expansion of $g(X)$ around μ is

$$g(X) = g(\mu) + g'(\mu)(X-\mu) + g''(\mu)(X-\mu)^2 + \dots$$

Therefore

$$\begin{aligned} \mathsf{E}[g(X)] &= g(\mu) + \mathsf{E}g''(\mu)(X-\mu)^2 + \dots \\ &\approx g(\mu) \end{aligned}$$

Similarly

$$\begin{aligned} \mathsf{Var}[g(X)] \approx \mathsf{E}[g(X) - g(\mu)]^2 &= \mathsf{E}[g'(\mu)(X-\mu)]^2 \\ &= \mathsf{E}[g'(\mu)]^2 \mathsf{Var}X \end{aligned}$$

Example

Let $X \sim N(\mu, \sigma^2)$ and $Y = \exp(X)$. The exact mean and variance of Y are

$$\mathsf{E}Y = e^{\mu+\sigma^2/2}, \qquad \mathsf{Var}Y = (e^{\sigma^2} - 1)e^{2\mu+\sigma^2}$$

The first order Taylor expansion gives

$$\mathsf{E}(Y) = e^{\mu}, \quad \mathsf{Var}(Y) = e^{2\mu}\sigma^2$$

The Delta method

Let Y_n be a sequence of rvs such that $\sqrt{n}(Y_n - \theta) \xrightarrow{d} N(0, \sigma^2)$. For a given function g and a specific value of θ, suppose $g'(\theta)$ exists and is nonzero. Then

$$\sqrt{n}[g(Y_n) - g(\theta)] \xrightarrow{d} N\big(0, \sigma^2[g'(\theta)]^2\big)$$

Proof: The Taylor expansion of $g(Y_n)$ around $Y_n = \theta$ is

$$g(Y_n) = g(\theta) + g'(\theta)(Y_n - \theta) + R_2$$

where $R_2 \to 0$ as $Y_n \to \theta$. Apply Slutsky's Theorem to

$$\sqrt{n}[g(Y_n) - g(\theta)] = g'(\theta)\sqrt{n}(Y_n - \theta)$$

Example

Let X_i iid with mean μ and variance σ^2. The CLT gives that

$$\sqrt{n}(\bar{X}_n - \mu) \xrightarrow{d} N(0, \sigma^2)$$

Now let $g(x) = e^x$, where $g'(x) = e^x > 0$ for all x. The Delta method gives that

$$\sqrt{n}\big(e^{\bar{X}_n} - e^{\mu}\big) \xrightarrow{d} N(0, \sigma^2 e^{2\mu})$$

Let $Y_i = \exp(X_i)$, then $e^{\bar{X}_n}$ is the geometric average of the Y_i. So we have an approximation for the distribution of the geometric average.

Second-order Delta method

Let Y_n be a sequence of rvs such that $\sqrt{n}(Y_n - \theta) \xrightarrow{d} N(0, \sigma^2)$. For a given function g and a specific value of θ, suppose $g'(\theta) = 0$ and $g''(\theta)$ exists and is nonzero. Then

$$n[g(Y_n) - g(\theta)] \xrightarrow{d} \frac{\sigma^2 g''(\theta)}{2}\chi^2_1$$

Proof: See C&B.