8 Laws of large numbers

8.1 Introduction

We first start with the idea of "standardizing a random variable." Let X be a random variable with mean μ and variance σ^2. Then $Z = (X - \mu)/\sigma$ will be a random variable with mean 0 and variance 1. We refer to this procedure of subtracting off the mean and then dividing by the standard deviation as standardizing the random variable. In particular, if X is a normal random variable, then when we standardize it we get the standard normal distribution.

The following comes up very often, especially in statistics. We have an experiment and a random variable X associated with it. We repeat the experiment n times and let $X_1, X_2, \cdots, X_n$ be the values of the random variable we get. We can think of the n-fold repetition of the original experiment as a sort of super-experiment and $X_1, X_2, \cdots, X_n$ as random variables for it. We assume that the experiment does not change as we repeat it, so that the X_j are indentically distributed. (Recall that this means that X_i and X_j have the same pmf or pdf.) And we assume that the different perfomances of the experiment do not influence each other. So the random variables $X_1, \cdots, X_n$ are independent.

In statstics one typically does not know the pmf or the pdf of the X_j. The statistician's job is to take the random sample $X_1, \cdots, X_n$ and make conclusions about the distribution of X. For example, one would like to know the mean, $\mathbf{E}[X]$, of X. The simplest way to estimate it is to look at the "sample mean" which is defined to be

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

Note that $\overline{X}_n$ is itself a random variable. Intuitively we expect that as $n \to \infty$, $\overline{X}_n$ will converge to $\mathbf{E}[X]$. What exactly do we mean by "convergence" of a sequence of random variables? And what can we say about the rate of convergence and the error? These questions are the focus of this chapter.

We already know quite a bit about $\overline{X}_n$. Its mean is $\mu = \mathbf{E}[X]$. And its variance is σ^2/n where σ^2 is the common variance of the X_j. The first theorems in this chapter will say that as $n \to \infty$, $\overline{X}_n$ converges to the constant μ in some sense. Results like this are called a "law of large numbers."

We will see two of them corresponding to two different notions of convergence. The other big theorem of this chapter is the central limit theorem. Suppose we shift and rescale $\overline{X}_n$ by

$$\frac{\overline{X}_n - \mu}{\sigma/\sqrt{n}}$$

Note that the above random variable has mean zero and variance one. The CLT says that it converges to a standard normal under some very mild assumptions on the distribution of X.

8.2 Weak law of large numbers

If we roll a fair six-sided die, the mean of the number we get is 3.5. If we roll the die a large number of times and average the numbers we get (i.e., compute $\overline{X}_n$), then we do not expect to get exactly 3.5, but rather something close. So we could ask if $|\overline{X}_n - 3.5| < 0.01$. This is an event (for the super-experiment), so we can consider its probability:

$$|\overline{X}_n - 3.5| < 0.01$$

In particular we might expect this probability to go to zero as $n \to \infty$. This motivates the following definition.

Definition 1. *Let Y_n be a sequence of random variables, and Y a random variable, all defined on the same probability space. We say Y_n converges to Y in probability if for every $\epsilon > 0$,*

$$\lim_{n\to\infty} \mathbf{P}(|Y_n - Y| > \epsilon) = 0$$

Theorem 1. *(***Weak law of large numbers***) Let X_j be an i.i.d. sequence with finite mean and variance. Let $\mu = \mathbf{E}[X_j]$. Then*

$$\overline{X}_n = \frac{1}{n}\sum_{j=1}^{n} X_j \to \mu \quad \textit{in probability}$$

There are better versions of the theorem in the sense that they have weaker hypotheses (you don't need to assume the variance is finite). There is also a stronger theorem that has a stronger form of convergence (strong law of large numbers).

We will eventually prove the theorem, but first we introduce another notion of convergence.

Definition 2. *Let Y_n be a sequence of random variables with finite variance and Y a random variable with finite variance. We say that Y_n converges to Y in mean square if*

$$\lim_{n\to\infty} \mathbf{E}[(Y_n - Y)^2] = 0$$

In analysis this is often called convergence in L^2.

Proposition 1. *Let X_n be a sequence of i.i.d. random variables with finite variance. Let $\mu = \mathbf{E}[X_n]$. Then X_n coverges to μ in mean square.*

Proof. We have to show

$$\lim_{n\to\infty} \mathbf{E}[(\overline{X}_n - \mu)^2] = 0$$

But since the mean of $\overline{X}_n$ is μ, $\mathbf{E}[(\overline{X}_n - \mu)^2]$ is the variance of $\overline{X}_n$. We know that this variance is σ^2/n which obviously goes to zero as $n \to \infty$. □

Next we show that convergence in mean square implies convergence in probability. The tool to show this is the following inequality:

Proposition 2. *(***Chebyshev's inequality***)*

$$\mathbf{P}(|X| \geq a) \leq \frac{\mathbf{E}[X^2]}{a^2}$$

Proof. To make things concrete we assume we have a continuous RV. Then letting $f(x)$ be the pdf of X,

$$\mathbf{E}[X^2] = \int_{-\infty}^{\infty} x^2 \, f(x) \, dx$$

Since the integrand is non-negative,

$$\begin{aligned} \mathbf{E}[X^2] &\geq \int_{|x|\geq a} x^2 \, f(x) \, dx \geq \int_{|x|\geq a} a^2 \, f(x) \, dx \\ &= a^2 \int_{|x|\geq a} f(x) \, dx = a^2 \mathbf{P}(|X| \geq a) \end{aligned}$$

Thus we have the inequality in the proposition. □

Proposition 3. *Let Y_n is a sequence of random variables with finite variance and Y is a random variable with finite variance. Suppose Y_n converges to Y in mean square. Then it converges in probability.*

Proof. Let $\epsilon > 0$. We must show

$$\lim_{n\to\infty} \mathbf{P}(|Y_n - Y| > \epsilon) = 0$$

By Chebyshev's inequality,

$$\mathbf{P}(|Y_n - Y| > \epsilon) \leq \frac{\mathbf{E}[(Y_n - Y)^2]}{\epsilon^2}$$

By hypothesis $\mathbf{E}[(Y_n - Y)^2] \to 0$ as $n \to \infty$. So for a fixed ϵ, $\mathbf{E}[(Y_n - Y)^2]/\epsilon^2 \to 0$ as $n \to \infty$. □

8.3 Central limit theorem

Let X_n be an i.i.d. sequence with finite variance. Let μ the their common mean and σ^2 their common variance. As before we let $\overline{X}_n = \frac{1}{n}\sum_{j=1}^n X_j$. Define

$$Z_n = \frac{\overline{X}_n - \mu}{\sigma/\sqrt{n}} = \frac{\sum_{j=1}^n X_j - n\mu}{\sqrt{n}\sigma}$$

Note that $\mathbf{E}[Z_n] = 0, var(Z_n) = 1$. The best way to remember the definition is that it is $\overline{X}_n$ shifted and scaled so that it has mean 0 and variance 1. The central limit theorem says that the distribution of Z_n converges to a standard normal. There are several senses in which it might converge, so we have to make this statement more precise. We might ask if the density function of Z_n converges to that of a standard normal, ie., $\frac{1}{\sqrt{2\pi}}\exp(-z^2/2)$. We do not assume that X_n is a continuous RV. If it is a discrete RV, then so is Z_n. So it does not even have a density function. Instead we look at the probability that Z_n is in some interval $[a, b]$.

Theorem 2. *(***Central limit theorem***) Let X_n be an i.i.d. sequence of random variables with finite mean μ and variance σ^2. Define Z_n as above. Then for all $a < b$*

$$\lim_{n\to\infty} \mathbf{P}(a \leq Z_n \leq b) = \int_a^b \frac{1}{\sqrt{2\pi}} e^{-z^2/2}\, dz$$

If we take $a = -\infty$, then the theorem says that the cdf of Z_n converges pointwise to the cdf of the standard normal. This is an example of what is called "convergence in distribution" in probability. However, we caution the reader that the general definition of convergence in distribution involves some technicalities.

This is the most important theorem in the course and plays a major role in statistics. In particular much of the theory of confidence intervals and hypothesis testing is the central limit theorem in disguise.

Example: A computer has a random number generator that generates random numbers uniformly distributed in $[0, 1]$. We run it 100 times and let S_n be the sum of the 100 numbers. Estimate $\mathbf{P}(S_n \geq 52)$. Let X_j be the numbers. So $\mathbf{E}[X_j] = 1/2$ and $var(X_j) = 1/12$. So

$$Z_n = \frac{S_n - n\mu}{\sqrt{n}\sigma} = \frac{S_n - 100 \times 1/2}{10\sqrt{1/12}}$$

So $S_n \geq 52$ is the same as

$$Z_n \geq \frac{52 - 100 \times 1/2}{10\sqrt{1/12}} = \frac{2\sqrt{12}}{10} = 0.6928$$

So $\mathbf{P}(S_n \geq 52) = \mathbf{P}(Z_n \geq 0.6928)$. Now we approximate $\mathbf{P}(Z_n \geq 0.6928)$ by assuming Z_n has a standard normal distribution. Then we can "look up" $\mathbf{P}(Z_n \geq 0.6928)$ in R. We use $\mathbf{P}(Z_n \geq 0.6928) = 1 - \mathbf{P}(Z_n \leq 0.6928)$. In R we get $\mathbf{P}(Z_n \leq 0.6928)$ from $pnorm(0.6928)$. (This is the same as $pnorm(0.6928, 0, 1)$. If you don't specify a mean and variance when you call $pnorm()$ it assumes a standard normal.) So we find

$$\mathbf{P}(S_n \geq 52) = \mathbf{P}(Z_n \geq 0.6928) = 1 - \mathbf{P}(Z_n \leq 0.6928) \approx 1 - 0.7548 = 0.2442$$

Example: We flip a fair coin n times. How large must n be to have $\mathbf{P}(|(fraction\, of\, H) - 1/2| \geq 0.01) \leq 0.05$.

According to R, $qnorm(0.975) = 1.959964$. This means that for a standard normal Z, $\mathbf{P}(Z \leq 1.959964) = 0.975$. By symmetry, $\mathbf{P}(|Z| \leq 1.959964) = 0.95$. So $\mathbf{P}(|Z| \geq 1.959964) = 0.05$. Let X_j be 1 if there is H on jth flip, 0 if it is T. Let $\overline{X}_n$ be the usual. Then $\overline{X}_n$ is the fraction of heads. For a single flip the variance is $p(1-p) = 1/4$. So $\sigma = 1/2$. So

$$Z_n = \frac{\overline{X}_n - 1/2}{\sigma\sqrt{n}} = \frac{\overline{X}_n - 1/2}{\sqrt{n}/2}$$

So $|\overline{X}_n - 1/2| \geq 0.01$ is the same as $|Z_n| \geq \frac{0.01n}{\sigma\sqrt{n}}$, i.e., $|Z_n| \geq 0.02\sqrt{n}$. So

$$\mathbf{P}(|(fraction\, of\, H) - 1/2| \geq 0.01) \leq 0.05 = \mathbf{P}(|Z_n| \geq 0.02\sqrt{n})$$

So it this is to be less than 0.05 we need $0.02\sqrt{n} \geq 2$. So $n \geq 10,000$.

End of ? lecture

Confidence intervals: The following is an important problem in statistics. We have a random variable X (usually called the population). We know its variance σ^2, but we don't know its mean μ. We have a "random sample," i.e., random variables $X_1, X_2, \cdots, X_n$ which are independent random variables which all have the same distribution as X. We want to use our one sample $X_1, \cdots, X_n$ to estimate μ. The natural estimate for μ is the sample mean

$$\overline{X}_n = \frac{1}{n}\sum_{j=1}^{n} X_j$$

How close is $\overline{X}_n$ to the true value of μ ? This is a vague question. We make it precise as follows. For what $\epsilon > 0$ will $\mathbf{P}(|\overline{X}_n - \mu| \leq \epsilon) = 0.95$? We say the $[\overline{X}_n - \epsilon, \overline{X}_n + \epsilon]$ is a 95% confidence interval for μ. (The choice of 95% is somewhat arbitrary. We can use 98% for example.

If n is large we can use the CLT to figure out what ϵ should be. As before we let

$$Z_n = \frac{\overline{X}_n - \mu}{\sigma/\sqrt{n}}$$

So $|\overline{X}_n - \mu| \leq \epsilon$ is equivalent to $|Z_n| \leq \epsilon\sqrt{n}/\sigma$ So we want

$$\mathbf{P}(|Z_n| \leq \epsilon\sqrt{n}/\sigma) = 0.95$$

The CLT says that the distribution for Z_n is approximately that of a standard normal. If Z is a standard normal, then $\mathbf{P}(|Z| \leq 1.96) = 0.95$. So $\epsilon\sqrt{n}/\sigma = 1.96$. So we have found that the 95% confidence interval for μ is $[\mu - \epsilon, \mu + \epsilon]$ where

$$\epsilon = 1.96 * \sigma/\sqrt{n}$$

Example: We assume the lifetime of a certain brand of light bulbs is exponential with a mean of roughly $2,000$ hours. We want a better estimate of the mean so we measure the liftimes of 25 bulbs and compute the sample mean $\bar{X}$ and get $2,132$ hours. Find a 95% confidence interval.

The $1/2$ game : Suppose $X_1, X_2, \cdots$ are i.i.d. and integer valued. For example, we flip a coin repeatedly and let X_n be 1 if we get heads on the nth flip and 0 if we get tails. We want to estimate $\mathbf{P}(a \leq S_n \leq b)$, where $S_n = \sum_{j=1}^n X_j$. We take a and b to be integers. Then

$$\mathbf{P}(a - \delta \leq S_n \leq b + \delta) = \mathbf{P}(a \leq S_n \leq b)$$

for any positive $\delta < 1$. But the CLT will give us different answers depending on which δ we use. What is the best? Answer $\delta = 1/2$. So we approximate $\mathbf{P}(a \leq S_n \leq b)$ by

$$\begin{aligned} \mathbf{P}(a - 1/2 \leq S_n \leq b + 1/2) &= \mathbf{P}(\frac{a - 1/2 - n\mu}{\sigma\sqrt{n}} \leq \frac{S_n - n\mu}{\sigma\sqrt{n}} \leq \frac{b + 1/2 - n\mu}{\sigma\sqrt{n}}) \\ &\approx \mathbf{P}(\frac{a - 1/2 - n\mu}{\sigma\sqrt{n}} \leq Z \leq \frac{b + 1/2 - n\mu}{\sigma\sqrt{n}}) \end{aligned}$$

where Z is a standard normal.

End of Nov 18 lecture

Example: Let X_n be an i.i.d. sequence of standard normals. Let

$$Y_n = \frac{1}{n} \sum_{j=1}^n X_j^2$$

Use the CLT to find n so that $\mathbf{P}(Y_n \geq 1.01) = 5\%$.

Note that the the i.i.d. sequence that we apply the CLT to is X_n^2 not X_n. The mean of X_n^2 is $\mathbf{E}[X_n^2] = 1$. The variance is $\mathbf{E}[X_n^4] - (\mathbf{E}[X_n^2])^2 = 3 - 1 = 2$. The mean of Y_n is $\frac{1}{n} n = 1$. The variance is

$$var(Y_n) = \frac{1}{n^2} n2$$

So to standardize,

$$Z_n = \frac{Y_n - 1}{\sqrt{2/n}}$$

So

$$\begin{aligned}\mathbf{P}(Y_n \geq 1.01) &= \mathbf{P}(\frac{Y_n - 1}{\sqrt{2/n}} \geq \frac{1.01 - 1}{\sqrt{2/n}}) \\ &\approx \mathbf{P}(Z \geq \frac{1.01 - 1}{\sqrt{2/n}})\end{aligned}$$

where Z is a standard normal. R says that $qnorm(0.95) = 1.645$. So $\mathbf{P}(Z \geq 1.645) = 0.05$. So

$$\frac{0.01}{\sqrt{2/n}} = 1.645,$$
$$n = (1.645)^2, 000 = 54, 121$$

8.4 Strong law of large numbers

Suppose X_n is an i.i.d. sequence. As before

$$\overline{X}_n = \frac{1}{n}\sum_{j=1}^{n} X_j$$

The weak law involves probabilities that X_n does certain things. We now ask if $\lim_{n\to\infty} \overline{X}_n$ exists and what does it converge to. Each $\overline{X}_n$ is a random variable, i.e., a function of ω, a point in the sample space. It is possible that this limit converges to μ for some ω but not for other ω.

Example We flips a fair coin infinitely often. Let X_n be 1 if the nth flip is heads, 0 if it is tails. Then $\overline{X}_n$ is the fraction of heads in the first n flips. We can think of the sample space as sequences of H's and T's. So one possible ω is the sequence with all H's. For this ω, $X_n(\omega) = 1$ for all n. So $\overline{X}_n(\omega) = 1$ for all n. So $\lim_{n\to\infty} \overline{X}_n$ exists, but it equals 1, not the mean of X_n which is 1/2. So it is certainly not true that $\lim_{n\to\infty} \overline{X}_n = \mu$ for all ω. Our counterexample where we get all heads is "atypical." So we might hope that the set of ω for which the limit of the $\overline{X}_n$ is not μ is "small".

Definition 3. *Let Y_n be a sequence of random variables and Y a random variable. We say X_n converges to X with probability one if*

$$\mathbf{P}(\{\omega : \lim_{n\to\infty} Y_n(\omega) = Y(\omega)\}) = 1$$

More succintly,

$$\mathbf{P}(Y_n \to Y) = 1$$

This is a stronger form of convergence than convergence in probability. (This is not at all obvious.)

Theorem 3. *If Y_n converges to Y with probability one, then Y_n converges to Y in probability.*

Proof. Let $\epsilon > 0$. We must show

$$\lim_{n\to\infty} \mathbf{P}(|Y_n - Y| \geq \epsilon) = 0$$

Define

$$E = \{\omega : \lim_{n\to\infty} Y_n(\omega) = Y(\omega)\}$$

Since Y_n converges to Y with probability one, $\mathbf{P}(E) = 1$. Let

$$E_n = \cap_{k=n}^{\infty}\{\omega : |Y_k(\omega) - Y(\omega)| < \epsilon\}$$

As n gets larger, the intersection has fewer sets and so is larger. So E_n is an increasing sequence. If $\omega \in E$, then $\omega \in E_n$ for large enough n. So

$$E \subset \cup_{n=1}^{\infty} E_n$$

Since $\mathbf{P}(E) = 1$, this implies $\mathbf{P}(\cup_{n=1}^{\infty} E_n) = 1$. By continuity of the probability measure,

$$1 = \mathbf{P}(\cup_{n=1}^{\infty} E_n) = \lim_{n\to\infty} \mathbf{P}(E_n)$$

Note that $E_n \subset \{|Y_n - Y| < \epsilon\}$. So $\mathbf{P}(E_n) \leq \mathbf{P}(|Y_n - Y| < \epsilon)$. Since $\mathbf{P}(E_n)$ goes to 1 and probabilities are bounded by 1, $\mathbf{P}(|Y_n - Y| < \epsilon)$ goes to 1. So $\mathbf{P}(|Y_n - Y| \geq \epsilon)$ goes to 0. □

8.5 Proof of central limit theorem

We give a partial proof of the central limit theorem. We will prove that the moment generating function of Z_n converges to that of a standard normal.

Let X_n be an i.i.d. sequence. We assume that $\mu = 0$. (Explain why we can do this.) Since the X_n are identically distributed, they have the same mgf. Call it $m(t)$. So

$$m(t) = M_{X_n}(t) = \mathbf{E}[e^{tX_n}]$$

As before

$$Z_n = \frac{S_n}{\sigma\sqrt{n}}, \quad S_n = \sum_{j=1}^{n} X_j$$

Now

$$M_{Z_n}(t) = \mathbf{E}[\exp(t\frac{S_n}{\sigma\sqrt{n}})] = \mathbf{E}[\exp(\frac{t}{\sigma\sqrt{n}}S_n)] = M_{S_n}(\frac{t}{\sigma\sqrt{n}})$$

Since the X_j are independent, $M_{S_n}(t) = \prod_{j=1}^{n} M_{X_j}(t) = m(t)^n$. So above becomes

$$M_{Z_n}(t) = \left[m(\frac{t}{\sigma\sqrt{n}})\right]^n$$

Now we do a Taylor expansion of $m(t)$ about $t = 0$. We do the expansion to second order:

$$m(t) = m(0) + m'(0)t + \frac{1}{2}m''(0)t^2 + O(t^3)$$

We have

$$\begin{aligned} m(0) &= 1 \\ m'(0) &= \mathbf{E}[X_j] = 0 \\ m''(0) &= \mathbf{E}[X_j^2] = var(X_j) = \sigma^2 \end{aligned}$$

So the Taylor expansion is

$$m(t) = 1 + \frac{1}{2}\sigma^2 t^2 + O(t^3)$$

So

$$\begin{aligned} \left[m(\frac{t}{\sigma\sqrt{n}})\right]^n &= \left[1 + \frac{1}{2}\sigma^2\frac{t^2}{\sigma^2 n} + O(t^3\sigma^3 n^{-3/2})\right]^n \\ &= \left[1 + \frac{t^2}{2n} + O(t^3\sigma^3 n^{-3/2})\right]^n \end{aligned}$$

We want to show that this converges to the mgf of the standard normal which is $\exp(\frac{1}{2}t^2)$. So we need to show the ln of the above converges to $\frac{1}{2}t^2$. The ln

of the above is

$$\begin{aligned}\ln(M_{Z_n}(t)) &= n\ln\left[1+\frac{t^2}{2n}+O(t^3\sigma^3 n^{-3/2})\right] \\ &= n\left[\frac{t^2}{2n}+O(t^3\sigma^3 n^{-3/2})+\cdots\right] \\ &= \frac{t^2}{2}+O(t^3\sigma^3 n^{-1/2})+\cdots\end{aligned}$$

which converges to $t^2/2$.

The following theorem (which we do not prove) completes the proof of the central limit under an additional hypothesis about momemnt generating functions.

Theorem 4. *(***Continuity theorem***) Let X_n be a sequence of random variables, X a RV. Let F_{X_n} and F_X be their cdf's. Let M_{X_n} and M_X be their moment generating functions. Suppose there is an $a > 0$ such that $M_{X_n}(t)$ and $M_X(t)$ exist for $|t| < a$. Suppose that for $|t| < a$, $M_{X_n}(t) \to M_X(t)$. Then for all x where $F(x)$ is continuous,*

$$F_{X_n}(x) \to F(x)$$

The central limit theorem only requires that the random variables have a finite second moment, not that their mgf exist. To proof the CLT in this case we use "characteristic functions" instead of mgf's.

Definition 4. *The characteristic function of a random variable X is*

$$\phi_X(t) = \mathbf{E}[e^{itX}] = \mathbf{E}[\cos(tX) + i\sin(tX)]$$

Since $|\cos(tX)|, |\sin(tX)| \leq 1$, the characteristic function is defined for all t for any random variable.

For a continuous random variable with density $f_X(x)$,

$$\phi_X(t) = \int_{-\infty}^{\infty} e^{itx} f_X(x)\,dx$$

This is just the Fourier transform of $f_X(x)$ For a normal random variable the computation of the characteristic function is almost identical to that of

the mfg. We find for a standard normal random variable X that $\phi_Z(t) = \exp(-t^2/2)$.

If X and Y are independent, then

$$\phi_{X+Y}(t) = \phi_X(t) + \phi_Y(t)$$

We also note that for a constant c, the characteristic function of cX is $\phi_{cX}(t) = \phi_X(ct)$.

Now let X have the Cauchy distribution. An exercise in contour integration shows

$$\phi_X(t) = \frac{1}{\pi} \int_{-\infty}^{\infty} \frac{e^{ixt}}{x^2+1}\, dx = e^{-|t|}$$

Now let X_n be an iid sequence of Cauchy random variables. So if we let S_n be their sum, then

$$\phi_{S_n}(t) = \exp -n|t|$$

This is the characteristic function of n times a Cauchy distribution. So

$$\frac{1}{n} \sum_{j=1}^{n} X_j$$

has a Cauchy random variable. Note that the CLT would say

$$\frac{1}{\sqrt{n}} \sum_{j=1}^{n} X_j$$

converges to a normal. So instead of the CLT theorem we see that if we rescale differently then in the limit we get a Cauchy distributions. There are random variables other than Cauchy for which we also get convergence to the Cauchy distribution.

Convergence of Binormial to Poisson ??????????????

8.6 Proof of strong law of large numbers

Borel Cantelli