Caltech Department of Mathematics

Ma 3/103 KC Border
Introduction to Probability and Statistics Winter 2017

Lecture 7: The Law of Averages

Relevant textbook passages:

Pitman [9]: Section 3.3

Larsen–Marx [8]: Section 4.3

7.1 Law of Averages

The "Law of Averages" is an informal term used to describe a number of mathematical theorems that relate averages of sample values to expectations of random variables.

Given random variables $X_1, \dots, X_n$ on a probability space $(S, \mathcal{E}, P)$, each point $s \in S$ yields a list $X(s)_1, \dots, X(s)_n$. If we average these numbers we get $\bar{X}(s) = \sum_{i=1}^n X_i(s)/n$, the **sample average** associated with the point s. The sample average, since it depends on s, is also a random variable.

In later lectures, I'll talk about how to determine the distribution of a sample average, but we already have a case that we can deal with. If $X_1, \dots, X_n$ are independent Bernoulli random variables, their sum has a Binomial distribution, so the distribution of the sample average is easily given. First note that the sample average can only take on the values k/n, for $k = 0, \dots, n$, and that

$$P\left(\bar{X} = k/n\right) = \binom{n}{k} p^k (1-p)^{n-k}.$$

Figure 7.1 shows the probability mass function of $\bar{X}$ for the case $p = 1/2$ with various values of n. Observe the following things about the graphs.

- The sample average $\bar{X}$ is always between 0 and 1, and it is simply the fraction of successes in sequence of trials.

- If the frequency interpretation of probability is to make sense, then as the sample size grows, it should converge to the probability of success, which in this case is 1/2.

- What can we conclude about the probability that $\bar{X}$ is near 1/2? As the sample size becomes larger, the heights (which measure probability) of the dots shrink, but there are more and more of them close to 1/2. Which effect wins?

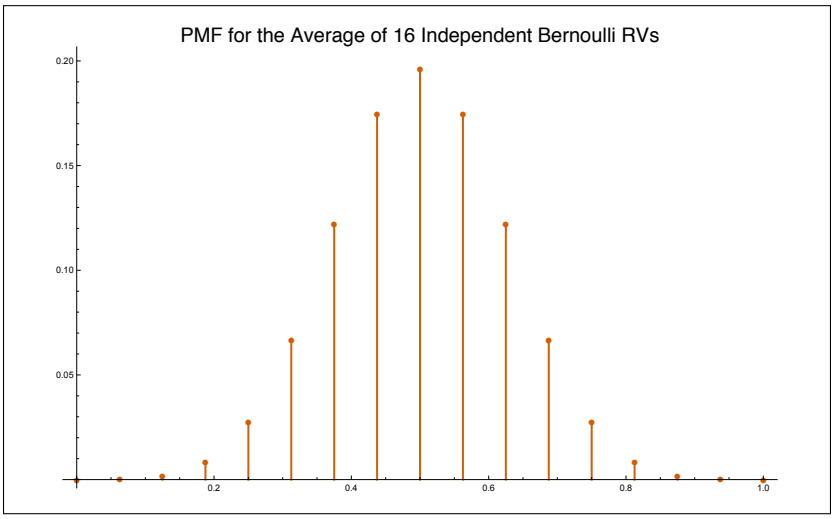
What happens for other kinds of random variables? Fortunately we do not need to know the details of the distribution to prove a Law of Averages. But we start with some preliminaries.

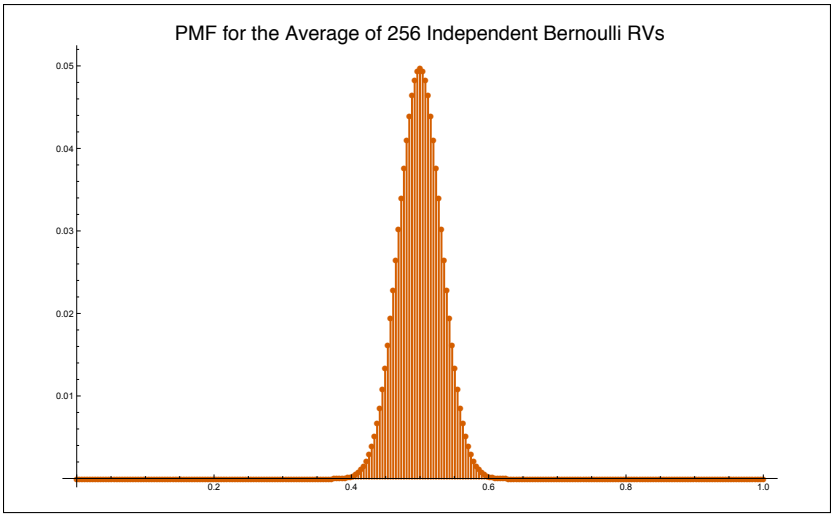
7.2 Preliminary Inequalities

7.2.1 Markov

Markov's Inequality bounds the probability of large values of a nonnegative random variable in terms of its expectation. It is a very crude bound, but it is just what we need.

Pitman [9]: p. 174


PMF for the Average of 16 Independent Bernoulli RVs
0.20
0.15
0.10
0.05
0.2
0.4
0.6
0.8
1.0


PMF for the Average of 256 Independent Bernoulli RVs
0.05
0.04
0.03
0.02
0.01
0.2
0.4
0.6
0.8
1.0

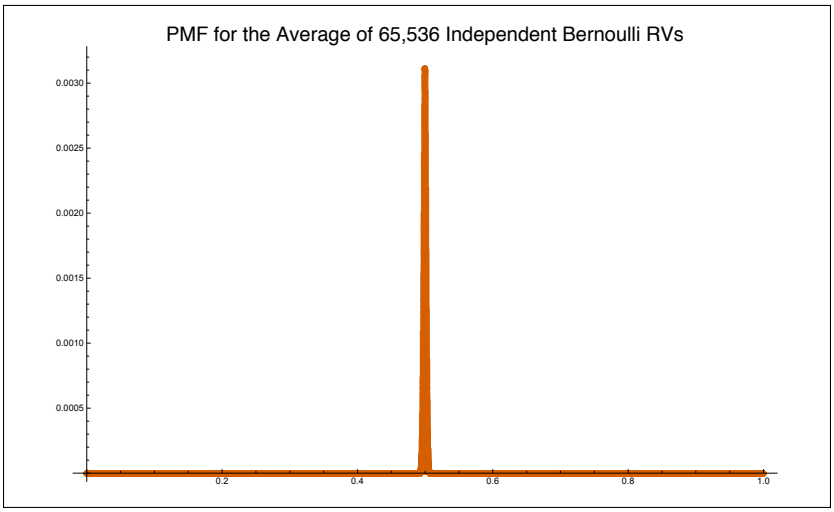

PMF for the Average of 65,536 Independent Bernoulli RVs
0.0030
0.0025
0.0020
0.0015
0.0010
0.0005
0.2
0.4
0.6
0.8
1.0

Figure 7.1.

7.2.1 Proposition (Markov's Inequality) *Let X be a nonnegative random variable with finite mean μ. For every $a > 0$,*

$$P(X \geqslant a) \leqslant \frac{\mu}{a}.$$

Proof: Use the fact that **expectation is a positive linear operator**. Recall that this implies that if $X \geqslant Y$, then $\boldsymbol{E}\, X \geqslant \boldsymbol{E}\, Y$.

Let $\mathbf{1}_{[a,\infty)}$ be the indicator of the interval $[a, \infty)$. Note that

$$X\mathbf{1}_{[a,\infty)}(X) = \begin{cases} X & \text{if } X \geqslant a \\ 0 & \text{if } X < a. \end{cases}$$

Since $X \geqslant 0$ this implies

$$X \geqslant X\mathbf{1}_{[a,\infty)}(X) \geqslant a\mathbf{1}_{[a,\infty)}(X).$$

Okay, so

$$\begin{aligned} \boldsymbol{E}\, X &\geqslant \boldsymbol{E}\big(X\mathbf{1}_{[a,\infty)}(X)\big) \\ &\geqslant \boldsymbol{E}\big(a\mathbf{1}_{[a,\infty)}(X)\big) \\ &= a\,\boldsymbol{E}\big(\mathbf{1}_{[a,\infty)}(X)\big) \\ &= aP\left(\mathbf{1}_{[a,\infty)}(X) = 1\right) \\ &= aP\left(X \geqslant a\right). \end{aligned}$$

Divide by $a > 0$ to get $P(A) \leqslant \boldsymbol{E}\, X/a$. ∎

7.2.2 Chebychev

Chebychev's Inequality bounds the deviation from the mean in terms of the number of standard deviations. **Pitman [9]:** p. 191

7.2.2 Proposition (Chebychev's Inequality, version 1) *Let X be a random variable with finite mean μ and variance σ^2 (and standard deviation σ). For every $a > 0$,*

$$P(|X - \mu| \geqslant a) \leqslant \frac{\sigma^2}{a^2}.$$

Proof:

$$\begin{aligned} P(|X-\mu| \geqslant a) &= P\left((X - \boldsymbol{E}\, X)^2 \geqslant a^2\right) && \text{square both sides} \\ &\leqslant \frac{\boldsymbol{E}(X - \boldsymbol{E}\, X)^2}{a^2} && \text{Markov's Inequality} \\ &= \frac{\sigma^2}{a^2} && \text{definition of variance,} \end{aligned}$$

where the inequality follows from Markov's Inequality applied to the random variable $(X - \boldsymbol{E}\, X)^2$ and the constant a^2. ∎

Recall that given a random variable X with finite mean μ and variance σ^2, the **standardization** of X is the random variable X^* defined by **Pitman [9]:** p. 190

$$X^* = \frac{X - \mu}{\sigma}.$$

Recall that $\boldsymbol{E}\, X^* = 0$ and $\boldsymbol{Var}\, X^* = 1$.

Then we may rewrite Chebychev's Inequality as:

Pitman [9]: p. 191

7.2.3 Proposition (Chebychev's Inequality, version 2) *Let X be a random variable with finite mean μ and variance σ^2 (and standard deviation σ), and standardization X^*. For every $k > 0$,*

$$P(|X^*| \geqslant k) \leqslant \frac{1}{k^2}.$$

7.3 Sums and averages of iid random variables

Pitman [9]: pp. 193–195

For each n define a new random variable S_n by

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

These are the **partial sums** of the sequence. Assume $\mu = \boldsymbol{E}\, X_1 = \boldsymbol{E}\, X_i$ for any i since the X_i are randomly distributed. Since expectation is a linear operator, we have

$$\boldsymbol{E}\, S_n = n\mu, \qquad \text{so } \boldsymbol{E}\left(\frac{S_n}{n}\right) = \mu.$$

Let $\sigma^2 = \boldsymbol{Var}\, X_1$ ($= \boldsymbol{Var}\, X_i$ for any i since the X_i are identically distributed). Since the random variables X_i are independent, we have

$$\boldsymbol{Var}\, S_n = n\sigma^2 \qquad \text{and} \qquad \mathrm{SD}\, S_n = \sqrt{n}\sigma$$

Recall that for any random variable Y, we have $\boldsymbol{Var}(aY) = a^2\, \boldsymbol{Var}\, Y$. Thus

$$\boldsymbol{Var}\left(\frac{S_n}{n}\right) = \frac{1}{n^2}\, \boldsymbol{Var}\, S_n = \frac{n\sigma^2}{n^2} = \frac{\sigma^2}{n}.$$

Thus the standard deviation (the square root of the variance) of S_n/n is given by

$$\text{standard deviation } \frac{S_n}{n} = \frac{\sigma}{\sqrt{n}}.$$

(Pitman [9, p. 194] calls this the **Square Root Law**.)

> Note that the variance of the **sum** of n independent and identically distributedrandom variables grows linearly with n, so the standard deviation grows like $\sqrt{n}$; but the variance of the **average** is proportional to $1/n$, so the standard deviation is proportional to $1/\sqrt{n}$.

Aside: This is confusing to many people, and can lead to bad decision making. The Nobel prize-winning economist Paul Samuelson [10] reports circa 1963 that he offered "some lunch colleagues to bet each \$200 to \$100 that the side of coin *they* specified would not appear at the first toss."

A "distinguished colleague" declined the bet, but said, "I'll take you on if you promise to let me make 100 such bets." Samuelson explains that, "He and many others give something like the following explanation. 'One toss is not enough to make it reasonably sure that the law of averages will turn out in my favor. But in a hundred tosses of a coin, the law of large numbers will make it a darn good bet. I am so to speak, virtually sure to come out ahead in such a sequence ..."

Samuelson points out that this is *not* what the Law of Large Numbers guarantees. He says that his colleague "should have asked for to subdivide the risk and asked for a sequence of 100 bets each of which was \$1 against \$2."

Let's compare means, standard deviations and the probability of losing money,

Bet	Expected Value	Std. Dev.	Probability of being a net Loser
\$200 v. \$100 on one coin flip	\$50	\$111.80	0.5
100 bets of \$200 v. \$100	\$5000	\$1,118.03	0.0003
100 bets of \$2 v. \$1	\$50	\$11.10	0.0003

Betting \$2 to \$1 has expectation \$0.50 and standard deviation \$1.11, so 100 such bets has mean \$50 and standard deviation \$11.10, which is a lot less risky than the first case.

7.4 The Weak Law of Large Numbers

It follows that as n gets large the variance of the partial sums S_n grows unboundedly, while the variance of the average S_n/n is shrinking to zero. That means that the averages are getting more and more concentrated around their mean μ. This is what is commonly referred to as the **Law of Averages**. There are a few versions. Here is one.

7.4.1 The Weak Law of Large Numbers, version 1 *Let $X_1, X_2, \dots$ be a sequence of independent and identically distributed random variables, with common expectation μ and variance $\sigma^2 < \infty$. Define the partial sums*

$$S_n = \sum_{i=1}^{n} X_n, \qquad n = 1, 2, 3, \dots .$$

Let $k > 0$ be given. Then

$$P\left(\left|\frac{S_n}{n} - \mu\right| \geqslant \varepsilon\right) \leqslant \frac{1}{n}\frac{\sigma^2}{\varepsilon^2}.$$

Proof: Note that if S is the sample space for a single experiment, this statement never requires consideration of a sample space more complicated than S^n. The result is simply Chebychev's Inequality (version 1) applied to the random variable S_n/n, which has mean μ and variance σ^2/n:

$$P\left(\left|\frac{S_n}{n} - \mu\right| \geqslant \varepsilon\right) \leqslant \frac{\sigma^2/n}{\varepsilon^2}.$$

∎

An important fact about this theorem is it tells us how fast in n the probability of the deviation of the average from the expected value goes to zero: it's bounded by a constant times $1/n$.

7.5 ⋆ Convergence in probability

Not in the textbooks.

7.5.1 Definition *A sequence $Y_1, Y_2, \dots$ of random variables (not necessarily independent) on a common probability space $(S, \mathcal{E}, P)$* ***converges in probability*** *to a random variable X if*

$$\big(\forall \varepsilon > 0\big) \Big[\lim_{n\to\infty} P\left(|Y_n - Y| \geqslant \varepsilon\right) = 0; \Big]$$

or equivalently

$$\big(\forall \varepsilon > 0\big) \Big[\lim_{n\to\infty} P\left(|Y_n - Y| \leqslant \varepsilon\right) = 1; \Big]$$

or equivalently

$$\big(\forall \varepsilon > 0\big) \big(\forall \delta > 0\big) \big(\exists N\big) \big(\forall n \geqslant N\big) \big[P\left(|Y_n - Y| > \varepsilon\right) < \delta\big],$$

in which case we may write

$$Y_n \xrightarrow{P} Y,$$

or

$$\operatorname*{plim}_{n\to\infty} Y_n = Y.$$

(The symbol plim *is pronounced "p-lim.")*

This allows us to rewrite the WLLN:

7.5.2 The Weak Law of Large Numbers, version 2 *Let $X_1, X_2, \dots$ be a sequence of independent an identically distributed random variables, with common expectation μ and variance $\sigma^2 < \infty$. Define the partial sums*

$$S_n = \sum_{i=1}^{n} X_n, \qquad n = 1, 2, 3, \dots .$$

Then

$$\operatorname*{plim}_{n\to\infty} \frac{S_n}{n} = \mu.$$

Note that this formulation throws away the information on the rate of convergence.

7.6 ⋆ The Strong Law of Large Numbers

There is another version of the Law of Averages that in some ways strengthens the Weak Law, and is called the **Strong Law of Large Numbers** or sometimes **Kolmogorov's Strong Law of Large Numbers**. The proof of the Strong Law is a *tour de force* of "hard analysis," chock full εs and δs and clever approximations and intermediate lemmas. That is one reason why I won't prove it in this course. One of the ways that the Strong Law strengthens the Weak Law is that it drops the restriction that the variables have finite variance.

Not in the textbooks.

7.6.1 Kolmogorov's Strong Law of Large Numbers *Let $X_1, X_2, \dots$ be a sequence of independent and identically distributed random variables, with common finite expectation μ. Define the partial sums*

$$S_n = \sum_{i=1}^{n} X_n, \qquad n = 1, 2, 3, \dots .$$

Then

$$\frac{S_n}{n} - \mu \to 0 \text{ with probability one.}$$

This statement is taken from Feller [5, Theorem 1, p. 238] and the proof may be found in pages 238–240. (His statement is for the case $\mu = 0$, but it trivially implies this apparent generalization.)

The next result is a converse to the strong law that shows that finiteness of the expectation cannot be dispensed with. It may be found in Feller [5, Theorem 4, p. 241].

7.6.2 Theorem *Let $X_1, X_2, \dots$ be a sequence of independent and identically distributed random variables. If $\boldsymbol{E}\lvert X_i\rvert = \infty$, then for any numerical sequence c_n,*

$$\limsup_{n\to\infty} \left| \frac{S_n}{n} - c_n \right| = \infty \text{ with probability one.}$$

One implication of this is that if the expectation is not finite, the averages will become arbitrarily large infinitely often with probability one.

7.7 ⋆ Sample spaces for independent and identically distributed random variables

Let $X_1, X_2, \dots$ be a sequence of independent and identically distributed random variables.

WHOA! What kind of sample space are we talking about here? Well let's start out with a probability space $(S, \mathcal{E}, P)$ and a random variable X on S. To get n repeated trials of the experiment, we take as our sample space S^n with the probability P^n.[1] When S is finite (or discrete) P^n is defined by

Not in the textbooks.

$$P^n\big((s_1, s_2, \dots, s_n)\big) = P(s_1)P(s_2)\cdots P(s_n).$$

One way to get a finite sequence $X_1, \dots, X_n$ of independent and identically distributed random variables is to take as your sample space S^n with probability P^n and for each point $\boldsymbol{s} = (s_1, \dots, s_n)$ in S^n define

$$X_i(\boldsymbol{s}) = X(s_i).$$

This is fine as long as we always get to work with some finite n. This is adequate for the Weak Law of Large Numbers, but it won't do for the Strong Law. For that we want a probability measure on the set of infinite sequences, $S^\infty = S \times S \times \cdots$. "What's the big deal?" you may ask. For a sequence $\boldsymbol{s} = (s_1, s_2, \dots)$ why don't we just set the probability of $\boldsymbol{s}$ to $P(s_1)P(s_2)P(s_3)\cdots$? Well, unless X is degenerate this will always be zero (assuming that we mean the limit of the finite products), which is not very helpful. Moreover, even if S has only two elements (as in the Bernoulli trials case), S^∞ is uncountably infinite, which we know means measure-theoretic trouble. Nevertheless, there is a meaningful way (provided S is not too weird) to put a probability measure P^∞ on S^∞, so that defining $X_i(\boldsymbol{s}) = X(s_i)$ makes $X_1, X_2, \dots$ a sequence of independent and identically distributed random variables on S^∞, which all have the same distribution as X. This result is known as the **Kolmogorov Extension Theorem**, after Andrey Kolmogorov (Андре́й Никола́евич Колмого́ров) Proving it requires a course in measure theory and topology, but you can find an excellent proof in Roko Aliprantis and KC Border [1, Theorem 15.6, p. 522].

With all these P^ns and P^∞ running around, when dealing with sequences of independent and identically distributed random variables, I may simply write Prob, knowing that they all agree.

[1] The set of events is called the product σ-algebra $\mathcal{E}^n$. and I won't go into details about that now.

7.8 ⋆ Comparing the Weak and Strong Laws

In order to understand the difference between the Weak and Strong Laws, let's confine our attention to the case where the expectation is zero. In order to understand the Weak Law, we only needed to compute probabilities on the finite product space S^n, which given independence we get my multiplying probabilities defined by P. In order to understand the Strong Law we have to realize that the sample space is $\Omega = S^\infty$, the space of all infinite sequences sample outcomes. A typical element in Ω is an infinite sequence $\boldsymbol{\omega} = (s_1, s_2, \dots)$ if outcomes in S. Let us assume as above that the random variables $X_1, X_2, \dots$ are defined on the common sample space $\Omega = S^\infty$ by $X_i\big((s_1, s_2, \dots)\big) = X(s_i)$. This space has a probability measure P^∞ that is consistent with P, but we typically cannot compute probabilities as "infinite products."

Not in the textbooks.

The Strong Law says that the set of infinite sequences $\boldsymbol{\omega} = (s_1, s_2, \dots)$ for which $S_n(\boldsymbol{\omega})/n$ becomes small (recall we are in the mean = 0 case) at some point N and stays small for all $n \geqslant N$ has probability one under the measure P^∞. The catch is that the Strong Law does not give a hint as to how large N is. The Weak Law says that with probability $\big(1 - (1/n)\big))(\sigma^2/\varepsilon^2)$, the average S_n/n is within ε of the mean zero. By taking n large enough we can make this probability as close to one as we want, but we can't be sure. In fact, there can be (infinitely many) sequences $\boldsymbol{\omega}$ for which S_n/n becomes arbitrarily large for infinitely many n. (It turns out that the set of these sequences have probability zero under P^∞.)

So one difference is that the Weak Law tells us how big n has to be to get the degree of (still positive) uncertainty we wish, but the Strong Law assures us that S_n/n will surely become small and stay small, but we just don't know when.

Kai Lai Chung [4, p. 233] points out that the relevance of the two versions is matter of dispute among probabilists. For instance, William Feller [5, p. 237] argues that,

> "In practice one is rarely interested in the probability $P\left(n^{-1}|S_n| > \varepsilon\right)$ for any particular large value of n. A more interesting question is whether $n^{-1}|S_n|$ will ultimately become and remain small, that is, whether $n^{-1}|S_n| < \varepsilon$ simultaneously for all $n \geqslant N$. Accordingly we ask for the probability of the event that $n^{-1}|S_n| \to 0$."

On the other hand, B. L. van der Waerden [15, p. 100] claims that the Strong Law of Large Numbers "hardly plays any role in mathematical statistics."

One thing I do know is that Claude Shannon's [12, 13] theory of information and coding relies only on the Weak Law. If you want to be well-educated in the 21st century, I strongly recommend you take Professor Michelle Effros's course **EE/Ma 126: Information Theory**.

7.9 ⋆ Convergence in probability vs. almost-sure convergence

7.9.1 Definition *A property is said to hold* ***almost surely****, abbreviated* **a.s.***, if the set of outcomes for which it holds has probability one.*

We can rephrase the SLLN as $S_n/n \to \boldsymbol{E}\, X$a.s..

More generally,

for any sequence $Y_1, Y_2, \dots$ on the common probability space $(S, \mathcal{E}, P)$, we say that Y_n **converges almost surely** to Y, written $Y_n \to Y$a.s. if

$$P\{s \in S : Y_n(s) \to Y(s)\} = 1.$$

I am including this discussion of the difference between convergence in probability and almost-sure convergence in theses notes, but **I do not plan to go over this material in class or examine you on it**. It is here for the benefit of those who want to understand the material more fully.

Let

$$\begin{aligned} A_n(\varepsilon) &= (|Y_n - Y| \leqslant \varepsilon) \\ &= \{s \in S : |Y_n(s) - Y(s)| \leqslant \varepsilon\}. \end{aligned}$$

We can rewrite convergence in probability as

$$\begin{aligned} &\operatorname*{plim}_{n\to\infty} Y_n = Y \\ &\iff \big(\forall k\big)\ \big(\forall m\big)\ \big(\exists M\big)\ \big(\forall n \geqslant M\big)\ \big[\,P\big(A_n(1/k)\big) > 1 - (1/m)\,\big] \end{aligned} \tag{1}$$

Now observe that the event

$$\begin{aligned} (Y_n \to Y) &= \{s \in S : Y_n(s) \to Y(s)\} \\ &= \{s \in S : \big(\forall \varepsilon > 0\big)\ \big(\exists N\big)\ \big(\forall n \geqslant N\big)\ \big[\,\big|Y_n(s) - Y(s)\big| < \varepsilon\big\}\,\big] \\ &= \{\in S : \big(\forall k\big)\ \big(\exists N\big)\ \big(\forall n \geqslant N\big)\ \big[\,\big|Y_n(s) - Y(s)\big| < 1/k\big\}\,\big] \end{aligned}$$

Now we use the cool trick of rewriting this as

$$= \bigcap_{k=1}^{\infty} \bigcup_{N=1}^{\infty} \bigcap_{n=N}^{\infty} A_n(1/k).$$

In other words,

$$Y_n \to Y \text{a.s.} \iff P\Big(\bigcap_{k=1}^{\infty} \bigcup_{N=1}^{\infty} \bigcap_{n=N}^{\infty} A_n(1/k)\Big) = 1. \tag{2}$$

The following argument uses the simple facts that for any countably additive probability measure P, if $P(\bigcap_j A_j) = 1$, then $P(A_j) = 1$ for all j; and if $P(\bigcup_{j=1}^{\infty} A_n) = 1$, then for each $\delta > 0$, there exists N such that $P(\bigcup_{j=1}^{N} A_j) > 1 - \delta$.

So observe that

$$\begin{aligned} &P\Big(\bigcap_{k=1}^{\infty} \bigcup_{N=1}^{\infty} \bigcap_{n=N}^{\infty} A_n(1/k)\Big) = 1 \\ &\implies \big(\forall k\big)\ \left[P\Big(\bigcup_{N=1}^{\infty} \bigcap_{n=N}^{\infty} A_n(1/k)\Big) = 1\right] \\ &\implies \big(\forall k\big)\ \big(\forall m\big)\ \big(\exists M\big)\ \left[P\Big(\bigcup_{N=1}^{M} \bigcap_{n=N}^{\infty} A_n(1/k)\Big) > 1 - (1/m)\right] \end{aligned}$$

But $\bigcup_{N=1}^{M} \bigcap_{n=N}^{\infty} A_n(1/k) = \bigcap_{n=M}^{\infty} A_n$ (because letting $E_N = \bigcap_{n=N}^{\infty} A_n(1/k)$, we have $E_1 \subset E_2 \subset \cdots \subset E_M$), so

$$\begin{aligned} &\implies \big(\forall k\big)\ \big(\forall m\big)\ \big(\exists M\big)\ \left[P\Big(\bigcap_{n=M}^{\infty} A_n(1/k)\Big) > 1 - (1/m)\right] \\ &\implies \big(\forall k\big)\ \big(\forall m\big)\ \big(\exists M\big)\ \big(\forall n \geqslant M\big)\ \big[\,P\big(A_n(1/k)\big) > 1 - (1/m)\,\big] \end{aligned}$$

so by (1)

$$\implies \operatorname{plim} Y_n = Y.$$

So we have proven the following proposition.

7.9.2 Proposition

$$Y_n \xrightarrow{\text{a.s.}} Y \implies Y_n \xrightarrow{P} Y.$$

7.9.1 Converse Not True

Consider the sequence defined by

$$\begin{aligned} &Y_1 = \mathbf{1}_{[0,\frac{1}{2})},\ Y_2 = \mathbf{1}_{[\frac{1}{2},1)} \\ &Y_3 = \mathbf{1}_{[0,\frac{1}{4})},\ Y_4 = \mathbf{1}_{[\frac{1}{4},\frac{1}{2})},\ Y_5 = \mathbf{1}_{[\frac{1}{2},\frac{3}{4})},\ Y_6 = \mathbf{1}_{[\frac{3}{4},1)} \\ &Y_7 = \mathbf{1}_{[0,\frac{1}{8})},\ \text{etc.} \end{aligned}$$

Then $Y_n \xrightarrow{P} 0$, but $Y_n \xrightarrow{\text{a.s.}\!\!\!\!\!\!\diagup} 0$.

However we do have the following result.

7.9.3 Proposition *If $Y_n \xrightarrow{P} Y$, then for some subsequence, $Y_{n_k} \xrightarrow{\text{a.s.}} Y$.*

7.10 ⋆ Convergence of Empirical CDFs

A corollary of the Law of Large Numbers is that the empirical distribution function of a sample of independent and identically distributed random variables converges to the ideal (theoretical) cumulative distribution function. In fact, this may be the best way to judge how well your data conforms to your theoretical model.

Not in the textbooks.

7.10.1 Definition *Given random variables $X_1, X_2, \dots, X_n, \dots$, for each n, the **empirical cumulative distribution function** F_n evaluated at x is defined to be the fraction of the first n random variables that have a value $\leqslant x$. That is,*

$$F_n(x) = \frac{|\{i : i \leqslant n \ \&\ X_i \leqslant x\}|}{n},$$

where, as you may recall the absolute value signs around a set denote the number of its elements.

*This makes each $F_n(x)$ a random variable, or each F_n a **random function**, so more pedantically, letting S denote the sample space on which the random variables are defined, what I should have written is*

$$F_n(x)(s) = \frac{|\{i : i \leqslant n \ \&\ X_i(s) \leqslant x\}|}{n}.$$

Recalling that the indicator function $\mathbf{1}_{(-\infty,x]}(y)$ is equal to one if $y \leqslant x$ and zero otherwise, we may rewrite this as

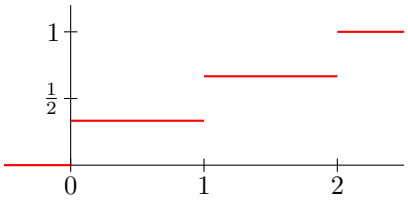
$$F_n(x) = \frac{\sum_{i=1}^n \mathbf{1}_{(-\infty,x]}(X_i)}{n}$$

7.10.2 Example Let X_i, $i = 1, 2$, be independent uniform random variables on the finite set $\{0, 1, 2\}$. (One way to get such a random variable is to roll a die, divide the result by 3, and take the remainder. Then 1 or 4 gives $X_i = 1$, 2 or 5 gives $X_i = 2$, and 3 or 6 gives $X_i = 0$.)

For the repeated experiments there are 9 possible outcomes:

$$(0,0), (0,1), (0,2), (1,0), (1,1), (1,2), (2,0), (2,1), (2,2).$$

Each of these outcomes has an empirical cumulative distribution function F_2 associated with it. There however only 6 distinct functions, since different points in the sample space may give rise to the same empirical cdf. Table 7.1 shows the empirical cdf associated to each point in the sample space. You should check that for each x if you look at the value of the empirical cdf at x and weight them by the probabilities associated with each empirical cdf, the resulting function is the cdf of the distribution of each X_i:



□

Now assume that $X_1, X_2, \dots, X_n$ are independent and identically distributed, with common cumulative distribution function F. Since

$$\boldsymbol{E}\, \mathbf{1}_{(-\infty,x]}(X_i) = \text{Prob}\,(X_i \leqslant x) = F(x),$$

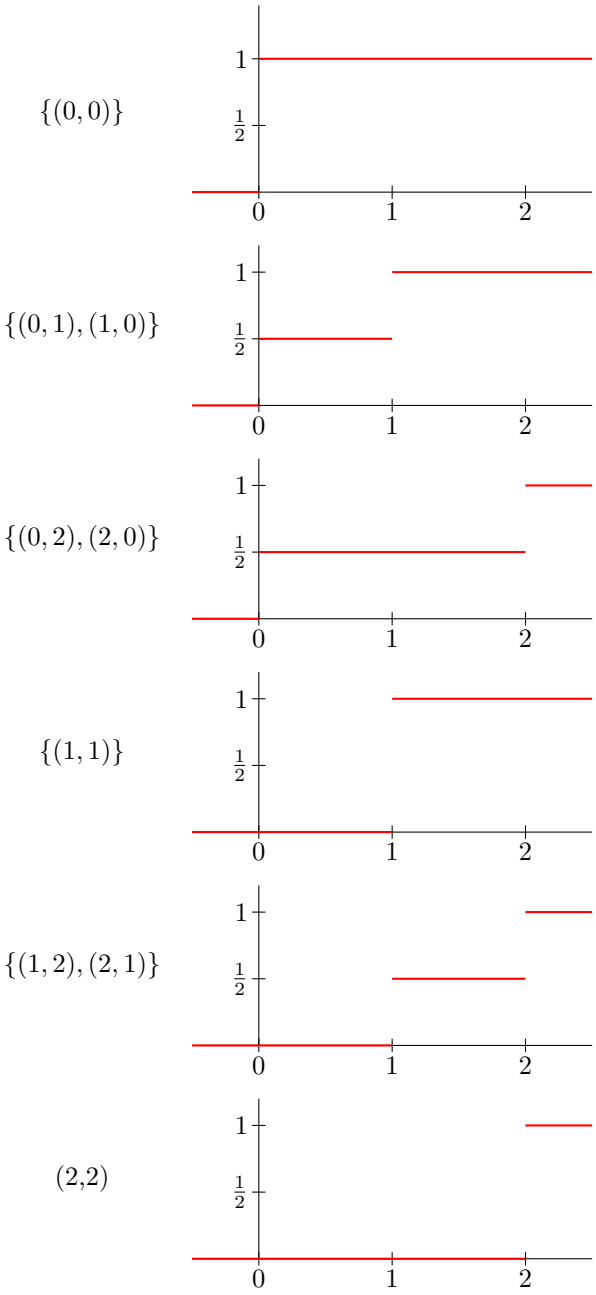


Table 7.1. Empirical cdfs at different points in the sample space for Example 7.10.2.

and since expectation is a linear operator, for each x we have

$$\boldsymbol{E}\, F_n(x) = \boldsymbol{E}\, \frac{\sum_{i=1}^n \mathbf{1}_{(-\infty,x]}(X_i)}{n} = F(x).$$

Now what is the variance of an indicator function? An indicator $\mathbf{1}_A$ is just a Bernoulli random variable with probability of success $P(A)$. Thus its variance is $P(A) - P(A)^2$, which is certainly finite. It follows from the Weak Law of Large Numbers that for each n we have the following

$$\operatorname{Prob}\left(|F_n(x) - F(x)| \geqslant \varepsilon\right) \leqslant \frac{F(x) - F(x)^2}{n\varepsilon^2},$$

and from the Strong Law that

$$\operatorname{Prob}\left(F_n(x) \to F(x)\right) = 1.$$

This result is a weaker version of the following theorem on uniform convergence, which we shall not prove. You can find it for example, in Kai Lai Chung's book [3], Theorem 5.5.1, p. 133.

7.10.3 Glivenko–Cantelli Theorem *Assume $X_1, X_2, \ldots, X_n, \ldots$ are independent and identically distributed, with common cumulative distribution function F. Then*

$$\operatorname{Prob}\left(\sup_{x\in\boldsymbol{R}} |F_n(x) - F(x)| \xrightarrow[n\to\infty]{} 0\right) = 1.$$

I shall omit the proof. It uses the Borel–Cantelli Lemma, [2, p. 44]. The practical implication of the Glivenko–Cantelli Theorem is this:

- Since with probability one, the empirical cdf converges to the true cdf,
- we can estimate the cdf **nonparametrically**.
- Moreover, since the empirical cdf is also the cdf of a uniform random variable on the sample values,
- with probability one, as the sample size n becomes large, **resampling** by drawing sample points independently at random is almost the same as getting new independent sample points.
- This resampling scheme is called the **bootstrap**, and is the basis for many nonparametric statistical procedures.
- The Glivenko–Cantelli Theorem also guarantees that we can find rules for creating **histograms** of our data that converge to the underlying density from which it is drawn.

7.11 ⋆ Histograms and densities

A **histogram** of a set of numbers is a bar graph that helps to visualize their distribution. The numbers are placed in **bins** or **classes** (that is, nonoverlapping intervals), and the height of the bar indicates the number of data in that bin. The histogram was named and popularized by the statistician Karl Pearson.

If the data are the results of independent and identically distributed draws from a density f, and if the histogram is normalized so that the total area of the bars is one, then the histogram can be used to approximate the density. As the sample size gets larger, we would expect that the normalized histogram would get closer to the density. In order for this to happen, the width of bins must shrink with the sample size.

This raises the question of how to choose the width and number of the bins. There is also the issue of making sure that data do not fall on bin boundaries, but that is relatively straightforward.

- Herbert Sturges [14] argued that if the sample size n is of the form 2^k, and the data are approximately Normal, we can choose $k+1$ bins so that the number in bin j should be close to the binomial coefficient $\binom{n}{j}$. "For example, 16 items would be divided normally into 5 classes, with class frequencies, $1, 4, 6, 4, 1$." For sample sizes n that are not powers of two, he argued that the range of the data should be divided into the number of bins that is the largest "convenient" integer (e.g., a multiple of 5) that is less than or equal to

$$1 + \log_2 n.$$

- David Freedman and Persi Diaconis [6] argue that one should try to minimize the L_2-norm of the difference between the normalized histogram and the density. This depends on unknown parameters, but they argued that the following rule is simple and approximately correct:

> Choose the cell width as twice the interquartile range of the data, divided by the cube root of the sample size.

- David Scott [11] argues that the mean-square error minimizing bin width for large n is given by

$$\left(\frac{6}{n \int_{-\infty}^{\infty} f'(x)^2 \, dx} \right)^{1/3}.$$

- Matt Wand [16] derived more sophisticated methods for binning data based on kernel estimation theory.

- Kevin Knuth [7] has recently suggested an algorithmic method for binning based on a Bayesian procedure. It reportedly works better for densities that are not unimodal.

Each of these methods will create histograms that approximate the density better and better as the sample size increases. Both R and Mathematica have options for using the Sturges, Freedman–Diaconis, and Scott rules in plotting histograms. Mathematica implements Wand's method and it is available as an R package (`dpih`). Mathematica also implements Knuth's rule.

Bibliography

[1] C. D. Aliprantis and K. C. Border. 2006. *Infinite dimensional analysis: A hitchhiker's guide*, 3d. ed. Berlin: Springer–Verlag.

[2] L. Breiman. 1968. *Probability*. Reading, Massachusetts: Addison Wesley.

[3] K. L. Chung. 1974. *A course in probability theory*, 2d. ed. Number 21 in Probability and Mathematical Statistics. Orlando, Florida: Academic Press.

[4] ——— . 1979. *Elementary probability theory with stochastic processes*. Undergraduate Texts in Mathematics. New York, Heidelberg, and Berlin: Springer–Verlag.

[5] W. Feller. 1971. *An introduction to probability theory and its applications*, 2d. ed., volume 2. New York: Wiley.

[6] D. Freedman and P. Diaconis. 1981. On the histogram as a density estimator: L_2 theory. *Zeitschrift für Wahrscheinlichkeitstheorie und Verwandte Gebiete* 57(4):453–476. DOI: 10.1007/BF01025868

[7] K. H. Knuth. 2013. Optimal data-based binning for histograms. arXiv:physics/0605197v2 [physics.data-an] http://arxiv.org/pdf/physics/0605197v2.pdf

[8] R. J. Larsen and M. L. Marx. 2012. *An introduction to mathematical statistics and its applications*, fifth ed. Boston: Prentice Hall.

[9] J. Pitman. 1993. *Probability*. Springer Texts in Statistics. New York, Berlin, and Heidelberg: Springer.

[10] P. A. Samuelson. 1963. Risk and uncertainty: A fallacy of large numbers. *Scientia* 98:108–113. http://https://www.casact.org/pubs/forum/94sforum/94sf049.pdf

[11] D. W. Scott. 1979. On optimal and data-based histograms. *Biometrika* 66(3):605–610. http://www.jstor.org/stable/2335182

[12] C. E. Shannon. 1948. A mathematical theory of communication [Introduction, Parts I and II]. *Bell System Technical Journal* 27(3):379–423. This issue includes the Introduction, Part I: Discrete Noiseless Systems, and Part II: The Discrete Channel with Noise. Part III is in [13].
http://www.alcatel-lucent.com/bstj/vol27-1948/articles/bstj27-3-379.pdf

[13] ——— . 1948. A mathematical theory of communication: Part III: Mathematical preliminaries. *Bell System Technical Journal* 27(4):623–656.
http://www.alcatel-lucent.com/bstj/vol27-1948/articles/bstj27-4-623.pdf

[14] H. A. Sturges. 1926. The choice of a class interval. *Journal of the American Statistical Association* 21(153):65–66. http://www.jstor.org/stable/2965501

[15] B. L. van der Waerden. 1969. *Mathematical statistics*. Number 156 in Grundlehren der mathematischen Wissenschaften in Einzeldarstellungen mit besonderer Berücksichtigung der Anwendungsgebiete. New York, Berlin, and Heidelberg: Springer–Verlag. Translated by Virginia Thompson and Ellen Sherman from *Mathematische Statistik*, published by Springer-Verlag in 1965, as volume 87 in the series Grundlerhen der mathematischen Wissenschaften.

[16] M. P. Wand. 1997. Data-based choice of histogram binwidth. *American Statistician* 51(1):59–64. DOI: 10.1080/00031305.1997.10473591