6.262: Discrete Stochastic Processes 2/7/11

Lecture 2: More review; the Bernoulli process

Outline:

- Expectations
- Indicator random variables
- Multiple random variables
- IID random variables
- Laws of large numbers in pictures
- The Bernoulli process
- Central limit theorem for Bernoulli

Expectations

The distribution function of a rv X often contains more detail than necessary. The expectation $\overline{X} = \mathsf{E}[X]$ is sometimes all that is needed.

$$\mathsf{E}[X] = \sum_i a_i \mathsf{p}_X(a_i) \qquad \text{for discrete } X$$

$$\mathsf{E}[X] = \int x \mathsf{f}_X(x) dx \qquad \text{for continuous } X$$

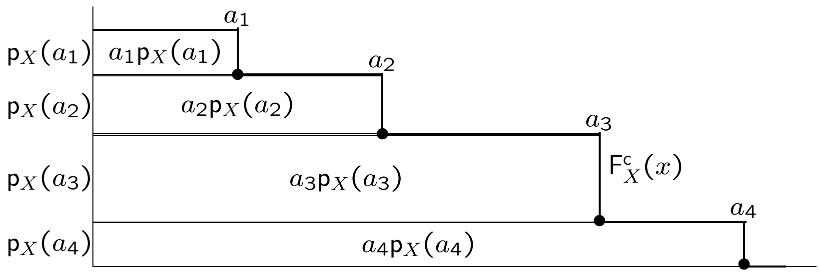
$$\mathsf{E}[X] = \int \mathsf{F}^{\mathsf{c}}_X(x)\, dx \qquad \text{for arbitrary nonneg } X$$

$$\mathsf{E}[X] = \int_{-\infty}^{0} \mathsf{F}_X(x)\, dx + \int_0^{\infty} \mathsf{F}^{\mathsf{c}}_X(x)\, dx \qquad \text{for arbitrary } X.$$

Almost as important is the standard deviation,

$$\sigma_X = \sqrt{\mathsf{E}\left[(X - \overline{X})^2\right]}$$

Why is $\mathsf{E}[X] = \int \mathsf{F}^c_X(x)\,dx$ for arbitrary nonneg X? Look at discrete case. Then $\int \mathsf{F}^c_X(x)\,dx = \sum_i a_i \mathsf{p}_X(a_i)$.



If X has a density, the same argument applies to every Riemann sum for $\int_x x\mathsf{f}_X(s)\,dx$ and thus to the limit.

It is simpler and more fundamental to take $\int \mathsf{F}^c_X(x)\,dx$ as the general definition of $\mathsf{E}[X]$. This is also useful in solving problems

Indicator random variables

For every event A in a probabiity model, an indicator rv $\mathbb{I}_A$ is defined where $\mathbb{I}_A(\omega) = 1$ for $\omega \in A$ and $\mathbb{I}_A(\omega) = 0$ otherwise. Note that $\mathbb{I}_A$ is a binary rv.

$$\mathsf{p}_{\mathbb{I}_A}(0) = 1 - \Pr\{A\}; \qquad \mathsf{p}_{\mathbb{I}_A}(1) = \Pr\{A\}.$$

$$\mathsf{E}[\mathbb{I}_A] = \Pr\{A\} \qquad \sigma_{\mathbb{I}_A} = \sqrt{\Pr\{A\}\,(1 - \Pr\{A\})}$$

Theorems about rv's can thus be applied to events.

Multiple random variables

Is a random variable (rv) X specified by its distribution function $\mathsf{F}_X(x)$?

No, the relationship between rv's is important.

$$\mathsf{F}_{XY}(x,y) = \Pr\Big\{\{\omega : X(\omega) \le x\} \bigcap \{\omega : Y(\omega) \le y\}\Big\}$$

The rv's $X_1, \dots, X_n$ are independent if

$$\mathsf{F}_{\vec{X}}(x_1, \dots x_n) = \prod_{m=1}^{n} \mathsf{F}_{X_m}(x_m) \quad \text{for all } x_1, \dots, x_n$$

This product form carries over for PMF's and PDF's.

For discrete rv's, independence is more intuitive when stated in terms of conditional probabilities.

$$\mathsf{p}_{X|Y}(x|y) = \frac{\mathsf{p}_{XY}(x,y)}{\mathsf{p}_Y(y)}$$

Then X and Y are independent if $\mathsf{p}_{X|Y}(x|y) = \mathsf{p}_X(x)$ for all sample points x and y. This essentially works for densities, but then $\Pr\{Y = y\} = 0$ (see notes). This is not very useful for distribution functions.

NitPick: If $X_1, \dots, X_n$ are independent, then all subsets of $X_1, \dots, X_n$ are independent. (This isn't always true for independent events).

IID random variables

The random variables $X_1, \dots, X_n$ are independent and identically distributed (IID) if for all $x_1, \dots, x_n$

$$\mathsf{F}_{\vec{X}}(x_1, \dots, x_n) = \prod_{k=1}^{n} \mathsf{F}_X(x_k)$$

This product form works for PMF's and PDF's also.

Consider a probability model in which $\mathbb{R}$ is the sample space and X is a rv.

We can always create an extended model in which $\mathbb{R}^n$ is the sample space and $X_1, X_2, \dots, X_n$ are IID rv's. We can further visualize $n \to \infty$ where $X_1, X_2, \dots$ is a stochastic process of IID variables.

We study the sample average, $S_n/n = (X_1 + \cdots + X_n)/n$. The laws of large numbers say that S_n/n 'essentially becomes deterministic' as $n \to \infty$.

If the extended model corresponds to repeated experiments in the real world, then S_n/n corresponds to the arithmetic average in the real world.

If X is the indicator rv for event A, then the sample average is the relative frequency of A.

Models can have two types of difficulties. In one, a sequence of real-world experiments are not sufficiently similar and isolated to correspond to the IID extendied model. In the other, the IID extension is OK but the basic model is not.

We learn about these problems here through study of the models.

Science, symmetry, analogies, earlier models, etc. are all used to model real-world situations.

Trivial example: Roll a white die and a red die. There are 36 sample outcomes, $(i, j), 1 \leq i, j \leq 6$, taken as equiprobable by symmetry.

Roll 2 indistinguishable white dice. The white and red outcomes (i, j) and (j, i) for $i \neq j$ are now indistinguishable. There are now 21 'finest grain' outcomes, but no sane person would use these as sample points.

The appropriate sample space is the 'white/red' sample space with an 'off-line' recognition of what is distinguishable.

Neither the axioms nor experimentation motivate this model, i.e., modeling requires judgement and common sense.

Comparing models for similar situations and analyzing limited and defective models helps in clarifying fuzziness in a situation of interest.

Ultimately, as in all of science, some experimentation is needed.

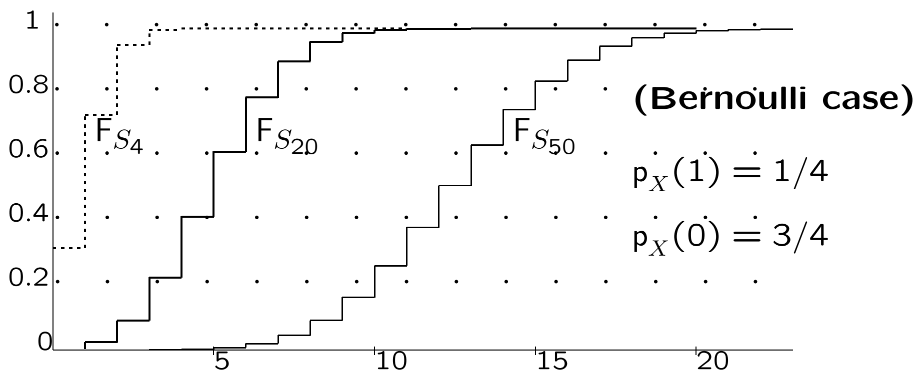
The outcome of an experiment is a sample point, not a probability.

Experimentation with probability requires multiple trials. The outcome is modeled as a sample point in an extended version of the original model.

Experimental tests of an original model come from the laws of large numbers in the context of an extended model.

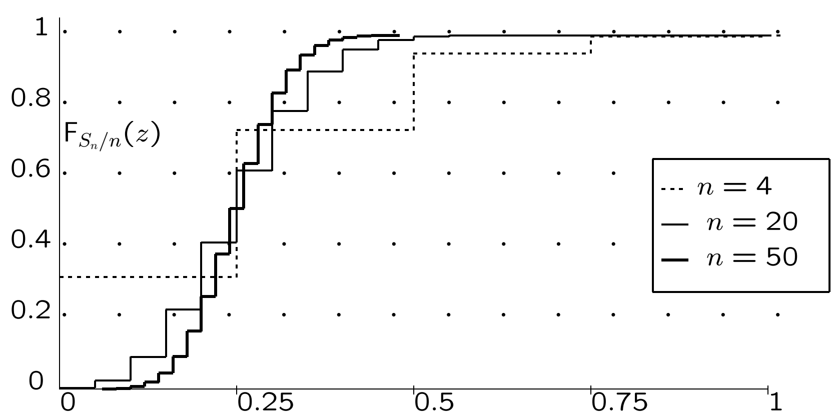
Laws of large numbers in pictures

Let $X_1, X_2, \dots, X_n$ be IID rv's with mean $\overline{X}$, variance σ^2. Let $S_n = X_1 + \cdots + X_n$. Then $\sigma^2_{S_n} = n\sigma^2$.



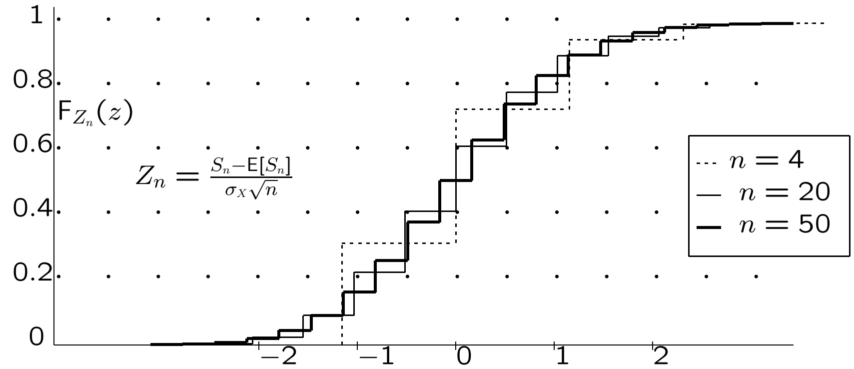
The center of the distribution varies with n and the spread (σ_{S_n}) varies with $\sqrt{n}$.

The sample average is S_n/n, which is a rv of mean $\overline{X}$ and variance σ^2/n.



The center of the distribution is $\overline{X}$ and the spread decreases with $1/\sqrt{n}$.

Note that $S_n - n\overline{X}$ is a zero mean rv with variance $n\sigma^2$. Thus $\frac{S_n - n\overline{X}}{\sqrt{n}\sigma}$ is zero mean, unit variance.



Central limit theorem:

$$\lim_{n\to\infty} \left[\Pr\left\{ \frac{S_n - n\overline{X}}{\sqrt{n}\,\sigma} \le y \right\} \right] = \int_{-\infty}^{y} \frac{1}{\sqrt{2\pi}} \exp\left(\frac{-x^2}{2} \right) dx.$$

The Bernoulli process

$S_n = Y_1 + \cdots Y_n \quad p_Y(1) = p > 0, \quad p_Y(0) = 1 - p = q > 0$

The n-tuple of k 1's followed by $n-k$ 0's has probability $p^k q^{n-k}$.

Each n tuple with k ones has this same probability. For $p < 1/2$, $p^k q^{n-k}$ is largest at $k = 0$ and decreasing in k to $k = n$.

There are $\binom{n}{k}$ n-tuples with k 1's. This is increasing in k for $k < n/2$ and then decreasing. Altogether,

$$p_{S_n}(k) = \binom{n}{k} p^k q^{n-k}$$

$$\mathsf{p}_{S_n}(k) = \binom{n}{k} p^k q^{n-k}$$

To understand how this varies with k, consider

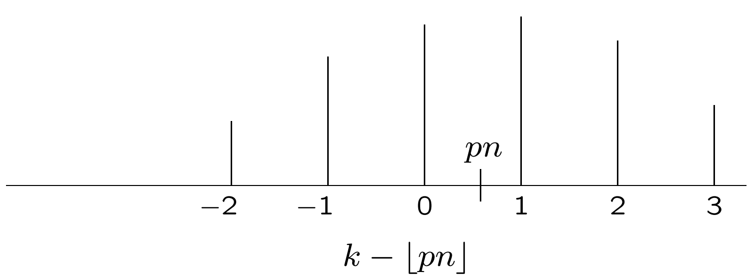
$$\begin{aligned}
\frac{\mathsf{p}_{S_n}(k{+}1)}{\mathsf{p}_{S_n}(k)} &= \frac{n!}{(k{+}1)!(n{-}k{-}1)!} \, \frac{k!(n{-}k)!}{n!} \, \frac{p^{k+1}q^{n-k-1}}{p^k q^{n-k}} \\
&= \frac{n-k}{k{+}1} \, \frac{p}{q}
\end{aligned}$$

This is strictly decreasing in k. It also satisfies

$$\frac{\mathsf{p}_{S_n}(k{+}1)}{\mathsf{p}_{S_n}(k)} \begin{cases} < 1 & \textbf{for} \quad k \geq pn \\ \approx 1 & \textbf{for} \quad k < pn < k+1 \\ > 1 & \textbf{for} \quad k+1 \leq pn \end{cases}$$

$$\frac{\mathsf{p}_{S_n}(k{+}1)}{\mathsf{p}_{S_n}(k)} = \frac{n-k}{k{+}1} \, \frac{p}{q} \tag{1}$$

$$\frac{\mathsf{p}_{S_n}(k{+}1)}{\mathsf{p}_{S_n}(k)} \begin{cases} < 1 & \textbf{for} \quad k \geq pn \\ \approx 1 & \textbf{for} \quad k < pn < k+1 \\ > 1 & \textbf{for} \quad k+1 \leq pn \end{cases}$$



In other words, $\mathsf{p}_{S_n}(k)$, for fixed n, is increasing with k for $k < pn$ and decreasing for $k > pn$.

CLT for Bernoulli process

$$\frac{\mathsf{p}_{S_n}(k+1)}{\mathsf{p}_{S_n}(k)} = \frac{n-k}{k+1}\,\frac{p}{q}$$

We now use this equation for large n where k is relatively close to pn. To simplify the algebra, assume pn is integer and look at $k = pn + i$ for relatively small i. Then

$$\begin{aligned}
\frac{\mathsf{p}_{S_n}(pn+i+1)}{\mathsf{p}_{S_n}(pn+i)} &= \frac{n-pn-i}{pn+i+1}\,\frac{p}{q} = \frac{nq-i}{pn+i+1}\,\frac{p}{q} \\
&= \frac{1-\frac{i}{nq}}{1+\frac{i+1}{np}} \\
\ln\left[\frac{\mathsf{p}_{S_n}(pn+i+1)}{\mathsf{p}_{S_n}(pn+i)}\right] &= \ln\left[1-\frac{i}{nq}\right] - \ln\left[1+\frac{i+1}{np}\right]
\end{aligned}$$

Recall that $\ln(1+x) \approx x - x^2/2 + \cdots$ **for** $|x| << 1$.

$$\begin{aligned}
\ln\left[\frac{\mathsf{p}_{S_n}(pn+i+1)}{\mathsf{p}_{S_n}(pn+i)}\right] &= \ln\left[1-\frac{i}{nq}\right] - \ln\left[1+\frac{i+1}{np}\right] \\
&= -\frac{i}{nq} - \frac{i}{np} - \frac{1}{np} + \cdots \\
&= -\frac{i}{npq} - \frac{1}{np} + \cdots
\end{aligned}$$

where we have used $1/p + 1/q = 1/pq$ and the neglected terms are of order i^2/n^2.

This says that these log of unit increment terms are essentially linear in i. We now have to combine these unit incremental terms.

$$\ln\left[\frac{\mathsf{p}_{S_n}(pn+i+1)}{\mathsf{p}_{S_n}(pn+i)}\right] = -\frac{i}{npq} - \frac{1}{np} + \cdots$$

Expressing an increment of j terms as a telescoping sum of j unit increments,

$$\begin{aligned}
\ln\left[\frac{\mathsf{p}_{S_n}(pn+j)}{\mathsf{p}_{S_n}(pn)}\right] &= \sum_{i=0}^{j-1} \ln\left[\frac{\mathsf{p}_{S_n}(pn+i+1)}{\mathsf{p}_{S_n}(pn+i)}\right] \\
&= \sum_{i=0}^{j-1} -\frac{i}{npq} - \frac{1}{np} + \cdots \\
&= -\frac{j(j-1)}{2npq} - \frac{j}{np} + \cdots \approx \frac{-j^2}{2npq}
\end{aligned}$$

where we have used the fact that $1+2+\cdots+j-1 = j((j-1)/2$. We have also ignored terms linear in j since they are of the same order as a unit increment in j.

Finally,

$$\begin{aligned}
\ln\left[\frac{\mathsf{p}_{S_n}(pn+j)}{\mathsf{p}_{S_n}(pn)}\right] &\approx \frac{-j^2}{2npq} \\
\mathsf{p}_{S_n}(pn+j) &\approx \mathsf{p}_{S_n}(pn)\exp\left[\frac{-j^2}{2npq}\right]
\end{aligned}$$

This applies for j both positive and negative, and is a quantized version of a Gaussian distribution, with the unknown scaling constant $\mathsf{p}_{S_n}(pn)$. Choosing this to get a PMF,

$$\mathsf{p}_{S_n}(pn+j) \approx \frac{1}{\sqrt{2\pi npq}}\exp\left[\frac{-j^2}{2npq}\right],$$

which is the discrete PMF form of the central limit theorem. See Section 1.5.3 for a different approach.