01/26/06 Discrete random variables (cont.)

Final Exam: May 8 3-6 PM

Office hours: Friday 01/27: 3-4 PM

Regular office hours starting next week:

Wednesdays 2-3 PM

Fridays 4-5 PM

Readings: Karlin + Taylor Sec. 1.1

Kloeden + Platen Secs 1.1-1.4

- include random variable simulation

Discrete random variables:

- state space (range) of r.v. X is discrete

Examples:

1. Binomial
2. Uniform

3. Geometric distribution

$$\text{Prob}(X = j) = (1-q)\, q^j \text{ for}$$

$j \geq 0$ and an integer

$(S = \mathbb{Z}_{\geq 0})$

q is a real number $0 < q < 1$

Interpretation: In a repeated sequence of Bernoulli trials with failure probability $= q$ (success probability $p = 1-q$) then X is the number of failures before the first success.

Genome sequencing example:

G A T T A C A

B. Mishra

primer — C T A dAx

G

A dA T

G

dA

T C

G

dA A

C

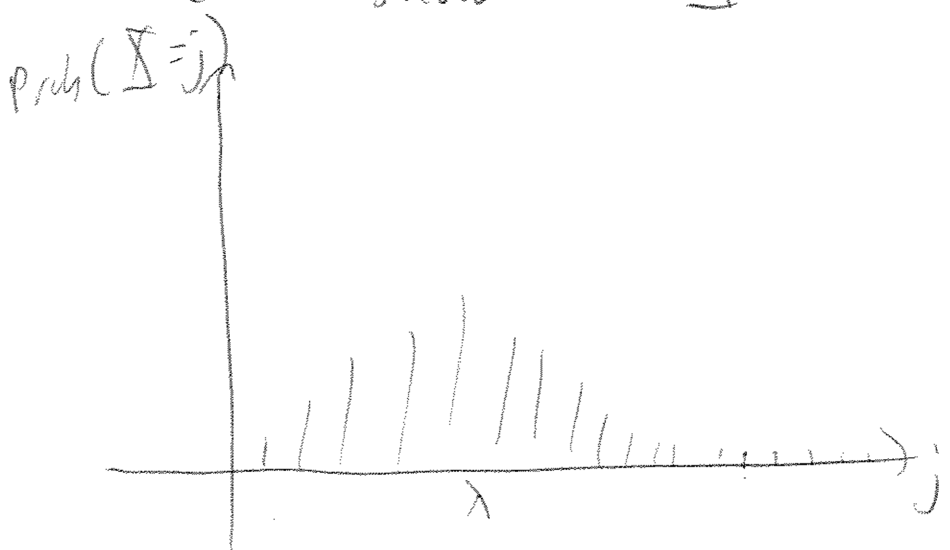
4. Poisson distribution

$S = \mathbb{Z}_{\geq 0}$

$$\text{Prob}(X = j) = \frac{\lambda^j}{j!} e^{-\lambda} \quad \text{for } j \in \mathbb{Z}_{\geq 0}$$

λ is a positive real parameter

Can show $\langle X \rangle = \lambda$.



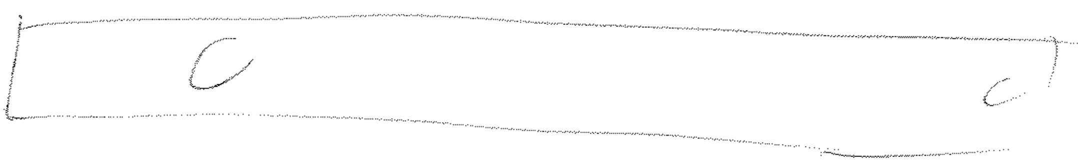
Poisson distribution "most natural" distribution on $\mathbb{Z}_{\geq 0}$

(like Gaussian for $\mathbb{R}$)

Central limit theorem for "rare incidents"

Poisson distribution models the # incidents that occur over a finite sample, given they occur with a prescribed density

Defects in material



Relations between several random variables

I will use <u>random vectors</u> to represent collections of random variables.

X_1 = # actual terrorists in the sample

X_2 = # friends of terrorists

X_3 = # CIA agents tapped

X_4 = # leaks to press

$$\vec{X} = \begin{pmatrix} X_1 \\ X_2 \\ X_3 \\ X_4 \end{pmatrix}$$

State space:

$$S = \mathbb{Z}_{\geq 0}^4$$

Characterize $\vec{X}$ for discrete space.

Specify

$$p_{\vec{x}} = \text{Prob}(\vec{X} = \vec{x}) \quad \text{for } \vec{x} \in S$$

$$= \text{Prob}(X_1 = x_1 \text{ and } X_2 = x_2 \ldots \text{ and } X_n = x_n)$$

Joint probability

This is complete information but sometimes TMI

How characterize in terms of simpler functions?

Partial information:

Marginal probability distribution

$$p_x^{(j)} = \text{Prob}(X_j = x)$$

$$= \sum_{\substack{x_1 \in S \\ x_2 \in S \ldots x_{j-1} \in S,\ x_{j+1} \in S \ldots x_N \in S}} p_{\vec{x}}$$

This figures relations between r.v's.
Look at relation between two subsets of r.v's

Let $\vec{Y}$ and $\vec{Z}$ be random vectors composed of subsets of the random variables in $\vec{X}$

Conditional probability:

$$\text{Prob}(\vec{Y} = \vec{y} \mid \vec{Z} = \vec{z})$$

$$= \frac{\text{Prob}(\vec{Y} = \vec{x} \text{ and } \vec{Z} = \vec{z})}{\text{Prob}(\vec{Z} = \vec{z})}$$

(just like Prob(B | A))

In particular $\vec{Y}$ and $\vec{Z}$ are independent $\Leftrightarrow$

$$\text{Prob}(\vec{Y} = \vec{y} \mid \vec{Z} = \vec{z}) = \text{Prob}(\vec{Y} = \vec{x})$$

$$\Leftrightarrow$$

$$\text{Prob}(\vec{Y} = \vec{x} \text{ and } \vec{Z} = \vec{z}) = \text{Prob}(\vec{Y} = \vec{x}) \times \text{Prob}(\vec{Z} = \vec{z})$$

If X_i and X_j are independent,

then $\mathrm{Cov}(X_i, X_j) = 0$.

This follows from:

Lemma: If X and Y are indep r.v.'s then $\langle f(X)\, g(Y) \rangle$

$$= \langle f(X) \rangle \langle g(Y) \rangle$$

Proof: $\langle f(X) g(Y) \rangle =$

$$\sum_{x \in S_X,\, y \in S_Y} f(x)\, g(y)\, \mathrm{Prob}(X = x, Y = y)$$

$\downarrow$ independence

$$= \sum_{x \in S_X,\, y \in S_Y} f(x)\, g(y)\, \mathrm{Prob}(X = x)\, \mathrm{Prob}(Y = y)$$

$$= \left(\sum_{x \in S_X} f(x)\, \mathrm{Prob}(X = x) \right) \times \left(\sum_{y \in S_Y} g(y)\, \mathrm{Prob}(Y = y) \right)$$

$$= \langle f(X) \rangle \langle g(Y) \rangle$$

Can simplify even more by

$$\mathbb{E}(\vec{Y} = \vec{y} \mid \vec{Z} = \vec{z})$$
$$= \langle \vec{Y} = \vec{y} \mid \vec{Z} = \vec{z} \rangle$$
$$= \sum_{\vec{y} \in S_{\vec{Y}}} \vec{y} \, \mathrm{Prob}(\vec{Y} = \vec{y} \mid \vec{Z} = \vec{z})$$

Conditional expectation

Another way to describe relations between random variables:

Covariance: Choose some X_i, X_j

$$\mathrm{Cov}(X_i, X_j) = \langle (X_i - \mu_i)(X_j - \mu_j) \rangle$$

where $\mu_i = \langle X_i \rangle$, $\mu_j = \langle X_j \rangle$

So $Cov(X_i, X_j) = \langle X_i - \mu_i \rangle \langle X_j - \mu_j \rangle$

if X_i and X_j independent

$= (\mu_i - \mu_i)(\mu_j - \mu_j)$

$= 0$

If $Cov(X_i, X_j) > 0$:

tendency for $X_i - \mu_i > 0 \Leftrightarrow X_j - \mu_j > 0$

"positive correlation"

If $Cov(X_i, X_j) < 0$:

tendency for $X_i - \mu_i > 0 \Leftrightarrow X_j - \mu_j < 0$

"negative correlation"

What about quantifying strength of correlation?

Better to look at correlation coefficient:

$$\rho(X_i, X_j) = \frac{Cov(X_i, X_j)}{\sigma_{X_i} \sigma_{X_j}}$$

standard deviations

Then $\underset{\substack{\uparrow \\ \text{strong} \\ \text{anti-correlation}}}{-1} \leq \rho(X_i, X_j) \leq \underset{\substack{\uparrow \\ \text{strong} \\ \text{positive} \\ \text{correlation}}}{1}$

Numerical simulation of rvs w/ discrete state space.

Computer generally has a pseudorandom number generator

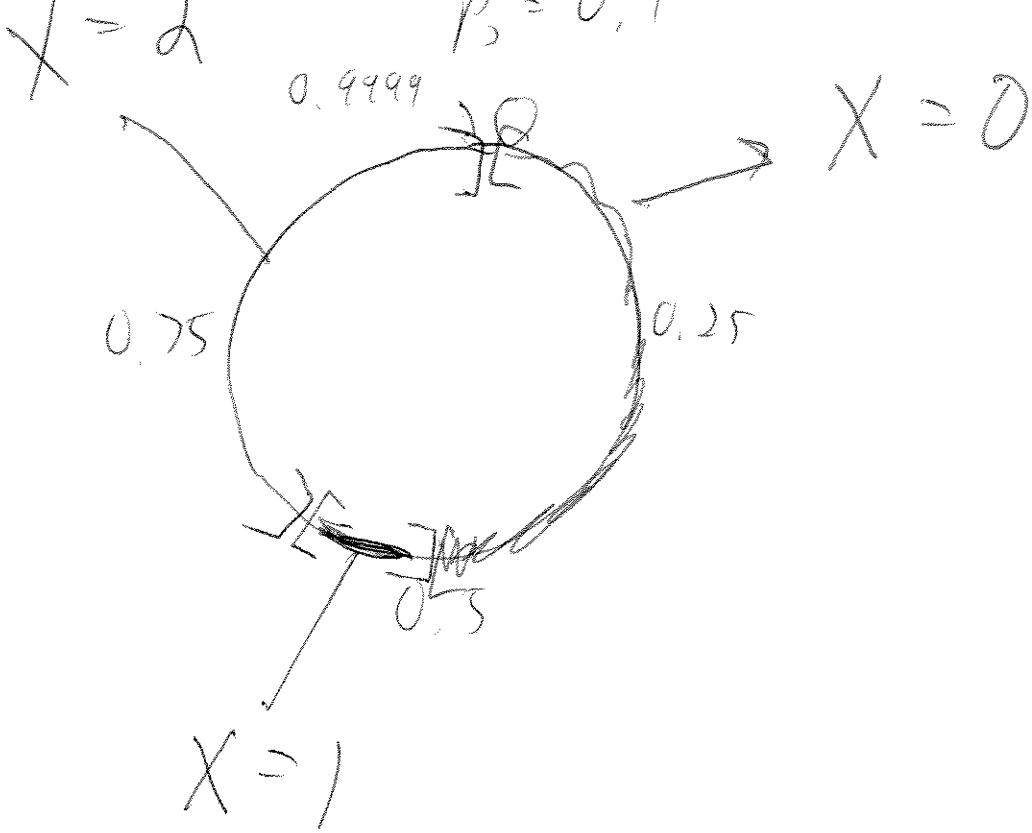
- Sophisticated tests for good pseudorandom number generators
  - L'Ecuyer

- supplies a random number uniformly on $[0,1]$.

How simulate a discrete r.v.?

$$\text{Prob}(X=j) = p_j$$

$$p_0 = 0.5$$
$$p_1 = 0.1$$
$$p_2 = 0.4$$



This works for any discrete rv but can get cumbersome for large state spaces.

- Sometimes can be more efficient by being clever

Binomial distribution:

$$\mathrm{Prob}(X = j) = \binom{N}{j} p^j (1-p)^{N-j}$$

for $j = 0, 1, 2, \ldots, N$.

For large N...

- easier to direct simulate Bernoulli trials

Infinite state space (geometric dist)

- easier to directly simulate

In other words, sometimes a random variable is most easily simulated by simulating a related stochastic process.

- Poisson

Random variables with continuous state space

What is the difference from discrete state space?

- practically: No longer meaningful to base everything on Prob $(X = x)$
- measure theory becomes technical

Continuous state space

- physical locations $(\mathbb{R}^3)$
- ~~stoc~~ financial models
- physical fields (temperature, pressure windspeed, ...)

Measure theory technicalities:

One wants to know

$\text{Prob}(X \in B)$ where B is a subset of state space S.

"Measure" is a means of assigning numbers to sets, obeying certain rules that wind up being what is natural for probabilities.

μ is measure:

$$\mu(\emptyset) = 0$$

$$\mu\left(\bigcup_{j=1}^{\infty} B_j\right) = \sum_{j=1}^{\infty} \mu(B_j)$$

if B_j disjoint

countable additivity

$$\mu(B) \geq 0$$

Annoying fact: Can't define self-consistent measure on $\mathbb{R}^3$ for <u>all</u> subsets.

So have to declare a collection of "measurable subsets" of S

In Euclidean space, a commonly used collection of measurable sets (σ-algebra)

$\mathcal{B}$ is the σ-algebra generated by all open cubes.
- closed under countable unions + intersections
- also all subsets of zero measure sets are measurable

Kloeden + Platen Sec. 2.1 – 2.2
- online

Billingsley, <u>Probability and Measure</u>

Sinai, <u>Probability Theory</u>

Folland, <u>Real Analysis</u>

Dudley, <u>Real Analysis and Probability</u>