MATH 180A: INTRO TO PROBABILITY
(FOR DATA SCIENCE)

www.math.ucsd.edu/~tkemp/180A

Today: § 8.1 – 8.4

Next: § 9.1 – 9.2

Homework 7 Due **Wednesday, Nov 27**

Exam 1: graded & released. (Mean 63%, St. Dev. 20%)

↳ Topics to focus on reviewing:

- Chebyshev's Inequality
- Transforming densities
- MGF (esp. when it determines the distribution)

We can now come back to questions about sums of random variables – in the context of their <u>joint</u> distribution.

Let X, Y be two (let's say discrete) random variables.

$$\mathbb{E}(\underbrace{X+Y}_{g(X,Y)}) = \sum_{k,\ell} g(k,\ell)\, p_{X,Y}(k,\ell) \leftarrow \quad p_{X,Y}(k,\ell) = \mathbb{P}(X=k, Y=\ell)$$

$$= \sum_{k,\ell} (k+\ell)\, p_{X,Y}(k,\ell) = \sum_{k,\ell} k\, p_{X,Y}(k,\ell) + \sum_{k,\ell} \ell\, p_{X,Y}(k,\ell)$$

$$= \sum_{k} k \underbrace{\sum_{\ell} p_{X,Y}(k,\ell)}_{p_X(k)} + \sum_{\ell} \ell \underbrace{\sum_{k} p_{X,Y}(k,\ell)}_{p_Y(\ell)}$$

$$= \sum_{k} k\, p_X(k) + \sum_{\ell} \ell\, p_Y(\ell) = \mathbb{E}(X) + \mathbb{E}(Y).$$

<u>Theorem:</u> For any random variables $X_1, X_2, \dots, X_n$,

$$\mathbb{E}(X_1 + \dots + X_n) = \mathbb{E}(X_1) + \mathbb{E}(X_2) + \dots + \mathbb{E}(X_n).$$

Eg. $S \sim \text{Bin}(n,p)$. $\quad P(S=k) = \binom{n}{k} p^k (1-p)^{n-k} \quad 0 \le k \le n$.

This means $S = X_1 + X_2 + \cdots + X_n$ where $X_1, \ldots, X_n \sim \text{Ber}(p)$

$$\therefore \mathbb{E}(S) = \mathbb{E}(X_1) + \mathbb{E}(X_2) + \cdots + \mathbb{E}(X_n) \qquad \mathbb{E}(X_j) = \mathbb{P}(X_j = 1) = p.$$

$$= p + p + \cdots + p \qquad (n \text{ of them})$$

$$= np.$$

A binomial is a sum of Bernoullis (indicator r.v.'s). Lots of problems can be solved when we can express desired events in terms of sums of indicators.

Eg. Suppose we put 200 balls randomly into 100 boxes. What is the expected number of empty boxes?

$X_i := \mathbb{1}\{\text{Box } i \text{ is empty}\}$ $\quad X := \#\text{of empty boxes} = \sum_{i=1}^{100} X_i$

$1 \le i \le 100$

$$\therefore \mathbb{E}(X) = \sum_{i=1}^{100} \mathbb{E}(X_i) = 100(0.99)^{200}$$

$$\mathbb{E}(X_i) = \mathbb{P}(X_i = 1) = \mathbb{P}(\text{Box } i \text{ is empty}) = (0.99)^{200} \qquad \doteq 13.4$$

Eg. Your favorite cereal (chocolate frosted sugar bombs) comes with a Pokémon figurine. There are n ~~20~~ to collect. What is the expected number of boxes you need to buy to collect them all?

X = # of boxes you need to collect them all

$Y \sim \mathrm{Geom}(p)$

$\mathbb{E}(Y) = 1/p$

X_1 = " " 1st one $= 1$

X_2 = # of boxes after the X_1^{th} needed to collect the 2nd $\sim \mathrm{Geom}\left(\frac{n-1}{n}\right)$

⋮

X_j = " " X_{j-1}^{th} " j^{th} $\sim \mathrm{Geom}\left(\frac{n-j+1}{n}\right)$

$X = X_1 + X_2 + \cdots + X_n$

$$\mathbb{E}(X) = \mathbb{E}(X_1) + \mathbb{E}(X_2) + \cdots + \mathbb{E}(X_n) = 1 + \frac{n}{n-1} + \frac{n}{n-2} + \cdots + \frac{n}{1}$$

$$= n\underbrace{\left(1 + \frac{1}{2} + \frac{1}{3} + \cdots + \frac{1}{n}\right)}_{H_n} \approx n \ln n + \gamma n + O\left(\tfrac{1}{n}\right)$$

$\gamma \sim 0.577$ E.M. const.

$(n=20)$: $\mathbb{E}(X) = 71.95$

Sums & Variances

~~8.2/8.4~~

$$
\begin{aligned}
\mathrm{Var}(X+Y) &= \mathbb{E}((X+Y)^2) - (\mathbb{E}(X+Y))^2 \\
&= \mathbb{E}((X+Y)^2) - (\mathbb{E}(X)+\mathbb{E}(Y))^2 \\
&= \mathbb{E}(X^2+2XY+Y^2) - (\mathbb{E}(X)^2+2\mathbb{E}(X)\mathbb{E}(Y)+\mathbb{E}(Y)^2) \\
&= \mathbb{E}(X^2)+2\mathbb{E}(XY)+\mathbb{E}(Y^2) - \mathbb{E}(X)^2 - 2\mathbb{E}(X)\mathbb{E}(Y) - \mathbb{E}(Y)^2 \\
&= \underbrace{\mathbb{E}(X^2)-\mathbb{E}(X)^2}_{\mathrm{Var}(X)} + \underbrace{\mathbb{E}(Y^2)-\mathbb{E}(Y)^2}_{\mathrm{Var}(Y)} + 2\underbrace{(\mathbb{E}(XY)-\mathbb{E}(X)\mathbb{E}(Y))}
\end{aligned}
$$

Def: $\mathrm{Cov}(X,Y) = \mathbb{E}[(X-\mathbb{E}(X))(Y-\mathbb{E}(Y))] = \mathbb{E}(XY)-\mathbb{E}(X)\mathbb{E}(Y)$

(calculation)

Note: $\mathrm{Cov}(X,X) = \mathrm{Var}(X)$

Theorem: $\mathrm{Var}(X_1+X_2+\cdots+X_n) = \mathrm{Var}(X_1)+\mathrm{Var}(X_2)+\cdots+\mathrm{Var}(X_n) + \sum_{i\neq j}\mathrm{Cov}(X_i,X_j)$

$= \sum_{i,j}\mathrm{Cov}(X_i,X_j)$

Covariance & Independence

If $X_1, X_2, \ldots, X_n$ are independent, then for $i \neq j$

$$\mathrm{Cov}(X_i, X_j) = \underbrace{\mathbb{E}(X_i X_j)}_{\mathbb{E}(X_i)\mathbb{E}(X_j)} - \mathbb{E}(X_i)\mathbb{E}(X_j) = 0.$$

Corollary: If $X_1, X_2, \ldots, X_n$ are independent

$$\mathrm{Var}(X_1 + X_2 + \cdots + X_n) = \mathrm{Var}(X_1) + \mathrm{Var}(X_2) + \cdots + \mathrm{Var}(X_n)$$

Eg. $S_n \sim \mathrm{Bin}(n, p)$ $\quad S_n = X_1 + X_2 + \cdots + X_n$ $\quad X_j$ iid

$X_j \sim \mathrm{Ber}(p)$

$\mathrm{Var}(S_n) = np(1-p)$ $\quad \therefore \mathrm{Var}(S_n)$

$$= \mathrm{Var}(X_1) + \mathrm{Var}(X_2) + \cdots + \mathrm{Var}(X_n)$$

$$= p(1-p) + p(1-p) + \cdots + p(1-p)$$

$$= np(1-p).$$

$$\therefore \mathrm{Var}(X_j) = \mathbb{E}(X_j^2) - \mathbb{E}(X_j)^2$$

$$= \mathbb{E}(X_j) - \mathbb{E}(X_j)^2$$

$$= p - p^2$$

$$= p(1-p)$$

Independent vs. Uncorrelated

We've seen that independent rv's are uncorrelated.
The converse does not hold.

E.g. $X \sim \text{Unif}\{-1,0,1\}$ (i.e. $P(X=\pm1)=P(X=0)=\frac{1}{3}$)
$Y = X^2$.

$$\mathbb{E}(X) = \tfrac{1}{3}(-1) + \tfrac{1}{3}(0) + \tfrac{1}{3}(1) = 0.$$

$$\begin{aligned}
\text{Cov}(X,Y) &= \mathbb{E}(XY) - \mathbb{E}(X)\mathbb{E}(Y) \\
&= \underbrace{\mathbb{E}(X^3)}_{=\mathbb{E}(\cancel{X}) = 0} - \underbrace{\mathbb{E}(\cancel{X})}_{0}\,\mathbb{E}(X^2) \\
&= 0.
\end{aligned}$$

$X^3 = X$.

Eg. Coupon Collector (Revisited)

Let T_n be the number of cereal boxes it takes to collect n distict toys.

$$T_n = 1 + W_1 + W_2 + \cdots + W_{n-1}$$

$W_k \sim \text{Geom}\left(\frac{n-k}{n}\right)$ are all independent.

$$\text{Var}(T_n) \approx \frac{\pi^2}{6} n^2$$

Reversion to the Mean

Let $X_1, X_2, \ldots, X_n$ be i.i.d. random variables (i.e. sampling, but not just Bernoulli trials.)

Say $\mathbb{E}(X_j) = \mu$, $Var(X_j) = \sigma^2$.

The sample mean $\bar{X}_n = \frac{1}{n}(X_1 + X_2 + \cdots + X_n)$.

$$\mathbb{E}(\bar{X}_n) = \frac{1}{n}\left(\mathbb{E}(X_1) + \cdots + \mathbb{E}(X_n)\right) = \mu$$

indep.

$$Var(\bar{X}_n) = Var\left(\frac{1}{n}\sum_{j=1}^{n} X_j\right) = \frac{1}{n^2} Var\left(\sum_{j=1}^{n} X_j\right)$$

$$= \frac{1}{n^2}\sum_{j=1}^{n} Var(X_j) = \frac{1}{n^2} \cdot n\sigma^2$$

$$= \frac{\sigma^2}{n}.$$