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 joint distribution.

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

\[
E(X + Y) = \sum_{k,l} g(k,l) \cdot P_{X,Y}(k,l) \iff P_{X,Y}(k,l) = P(X = k, Y = l)
\]

\[
E(X + Y) = \sum_{k,l} (k + l) \cdot P_{X,Y}(k,l) = \sum_{k} k \cdot P_{X,Y}(k,l) + \sum_{l} l \cdot P_{X,Y}(k,l)
\]

\[
= \sum_{k} k \cdot \left( \sum_{l} P_{X,Y}(k,l) \right) = \sum_{k} k \cdot P_{X}(k)
\]

\[
= \sum_{l} l \cdot \left( \sum_{k} P_{Y}(l) \right) = \sum_{l} l \cdot P_{Y}(l) = E(X) + E(Y).
\]

**Theorem:** For any random variables $X_1, X_2, \ldots, X_n$,

\[
E(X_1 + \ldots + X_n) = E(X_1) + E(X_2) + \ldots + E(X_n).
\]
Eg. $S \sim \text{Bin}(n,p)$. 

$$P(S = k) = \binom{n}{k} p^k (1-p)^{n-k} \quad 0 \leq k \leq n.$$ 

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

$$\therefore E(S) = E(X_1) + E(X_2) + \cdots + E(X_n) \quad E(X_j) = P(X_j = 1) = p.$$ 

$$= p + p + \cdots + p \quad (n \text{ of } p \text{em})$$ 

$$= np.$$ 

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?

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

$$= \sum_{i=1}^{100} E(X_i) \quad \because E(X_i) = P(X_i = 1) = P(\text{Box } i \text{ is empty}) = (0.9)^2 \to \approx 100 (0.9)^{200}$$ 

$$E(X_i) = P(X_i = 1) = P(\text{Box } i \text{ is empty}) = (0.9)^2 \to \approx 100 (0.9)^{200}$$ 

$$\approx 13.4.$$
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?

$x = \text{# of boxes you need to collect them all}$

$x_1 = "1st one = 1"

$x_2 = \text{# of boxes after the } x_1^{th} \text{ needed to collect the } 2^{nd}$

$\vdots$

$x_j = \text{# of boxes after the } x_{j-1}^{th} \text{ needed to collect the } j^{th}$

$\vdots$

$x_n = \text{# of boxes after the } x_{n-1}^{th} \text{ needed to collect the } n^{th}$

$X = x_1 + x_2 + \cdots + x_n$

$\mathbb{E}(X) = \mathbb{E}(x_1) + \mathbb{E}(x_2) + \cdots + \mathbb{E}(x_n) = 1 + \frac{n}{n-1} + \frac{n}{n-2} + \cdots + \frac{n}{1}$

$\approx n \left( 1 + \frac{1}{2} + \frac{1}{3} + \cdots + \frac{1}{n} \right) \approx n \ln n + \gamma n + O(1)$

$(n = 20) : \mathbb{E}(X) \approx 71.95$
Sums & Variances

\[ \text{Var}(X+Y) = \mathbb{E}((X+Y)^2) - (\mathbb{E}(X+Y))^2 \]
\[ = \mathbb{E}(X^2 + 2XY + Y^2) - (\mathbb{E}(X)^2 + 2\mathbb{E}(X)\mathbb{E}(Y) + \mathbb{E}(Y)^2) \]
\[ = \mathbb{E}(X^2) + 2\mathbb{E}(XY) + \mathbb{E}(Y^2) - \mathbb{E}(X)^2 - 2\mathbb{E}(X)\mathbb{E}(Y) - \mathbb{E}(Y)^2 \]
\[ = \frac{\mathbb{E}(X^2) - \mathbb{E}(X)^2}{\text{Var}(X)} + \frac{\mathbb{E}(Y^2) - \mathbb{E}(Y)^2}{\text{Var}(Y)} + 2(\mathbb{E}(XY) - \mathbb{E}(X)\mathbb{E}(Y)) \]

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

**Note:** \( \text{Cov}(X,Y) = \text{Var}(X) \)

**Theorem:** \( \text{Var}(X_1 + X_2 + \ldots + X_n) = \text{Var}(X_1) + \text{Var}(X_2) + \ldots + \text{Var}(X_n) \)

\[ \Delta = \sum_{i,j} \text{Cov}(X_i,X_j) + \sum_{i \neq j} \text{Cov}(X_i;X_j) \]
Covariance & Independence

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

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

\[\mathbb{E}(X_i)\mathbb{E}(X_j)\]

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

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

E.g., $S_n \sim \text{Bin}(n, p)$  \[S_n = X_1 + X_2 + \cdots + X_n \]

$$\text{Var}(S_n) = np(1-p) \quad \therefore \text{Var}(S_n) = \text{Var}(X_1) + \text{Var}(X_2) + \cdots + \text{Var}(X_n)$$

$$= np(1-p) + np(1-p) + \cdots + np(1-p) \quad \because \text{Var}(X_i) = \mathbb{E}(X_i^2) - \mathbb{E}(X_i)^2$$

$$= np(1-p) + \cdots + np(1-p) \quad \text{Var}(X_j) = \mathbb{E}(X_j^2) - \mathbb{E}(X_j)^2$$

$$= np(1-p) \quad \because \text{Var}(X_j) = \mathbb{E}(X_j^2) - \mathbb{E}(X_j)^2$$
Independent vs. Uncorrelated

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

Eg. \( X \sim \text{Unif}\{-1,0,1\} \) (i.e. \( P(X=\pm 1) = P(X=0) = \frac{1}{3} \) )

\[ E(X) = \frac{1}{3}(-1) + \frac{1}{3}(0) + \frac{1}{3}(1) = 0. \]

\[ X^3 = X. \]

\[ \text{Cov}(X,Y) = E(XY) - E(X)E(Y) \]

\[ = E(X^3) - E(X)E(X^2) \]

\[ = 0. \]
Eg. Coupon Collector (Revisited)

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

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

$W_k \sim \text{Geom}\left( \frac{n-k}{n} \right)$ 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 $\mathbb{E}(X_j) = \mu$, $\text{Var}(X_j) = \sigma^2$.

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

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

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

\[ \sigma^2 / n \]