Probability and Mathematical Statistics in Data Science

Lecture 18: Section 7.1: Sum of Independent Random Variables

7.1: Sums of Independent Random Variables

- Recall that expectation is additive, which we used many times.

$$(E(X+Y)=E(X)+E(Y))$$

- What about $Var(X+Y)$? Well, it depends.

- Consider tossing a fair coin 10 times. Let H be the number of heads and T be the number of tails in 10 tosses. Then $H+T=10$.

- Note that $Var(H), Var(T) \neq 0$, but $Var(H+T) = Var(10) = 0$

7.1: Sums of Independent Random Variables

- But now let H_1 be the number of heads in the first 5 tosses, and H_2 the number of heads in the last 5 tosses.

- Will we have that $Var(H_1 + H_2) = 0$?

7.1: Sums of Independent Random Variables

- It turns out that if X and Y are ***independent***, then we have that

$$Var(X+Y) = Var(X) + Var(Y)$$

and

$$Var(X-Y) = Var(X) + Var(Y)$$

Sums of iid random variables

- Let $X_1, X_2, \ldots, X_n$ be independent and identically distributed random variables with mean μ and variance σ^2. Define S_n to be their sum:

$$S_n = X_1 + X_2 + \cdots + X_n.$$

- We already know that $E(S_n) = \sum E(X_k) = n\mu$.

- Now we can further say that:

$$Var(S_n) = Var(X_1 + X_2 + \cdots + X_n)$$

$$= Var(X_1) + \cdots + Var(X_n) = n\sigma^2$$

$$SD(S_n) = \sqrt{n}\,\sigma$$

Proposition

Let $X_1, X_2, \ldots, X_n$ have mean values $\mu_1, \ldots, \mu_n$, respectively, and variances $\sigma_1^2, \ldots, \sigma_n^2$, respectively.

1. Whether or not the X_i's are independent,

$$E(a_1X_1 + a_2X_2 + \cdots + a_nX_n) = a_1E(X_1) + a_2E(X_2) + \cdots + a_nE(X_n)$$
$$= a_1\mu_1 + \cdots + a_n\mu_n \quad (5.8)$$

2. If $X_1, \ldots, X_n$ are independent,

$$V(a_1X_1 + a_2X_2 + \cdots + a_nX_n) = a_1^2V(X_1) + a_2^2V(X_2) + \cdots + a_n^2V(X_n)$$
$$= a_1^2\sigma_1^2 + \cdots + a_n^2\sigma_n^2 \quad (5.9)$$

and

$$\sigma_{a_1X_1+\cdots+a_nX_n} = \sqrt{a_1^2\sigma_1^2 + \cdots + a_n^2\sigma_n^2} \quad (5.10)$$

Example

A shipping company handles containers in three different sizes: (1) 27 ft^3 ($3 \times 3 \times 3$), (2) 125 ft^3, and (3) 512 ft^3. Let X_i ($i = 1, 2, 3$) denote the number of type i containers shipped during a given week. With $\mu_i = E(X_i)$ and $\sigma_i^2 = V(X_i)$, suppose that the mean values and standard deviations are as follows:

$$\mu_1 = 200 \qquad \mu_2 = 250 \qquad \mu_3 = 100$$
$$\sigma_1 = 10 \qquad \sigma_2 = 12 \qquad \sigma_3 = 8$$

a. Assuming that X_1, X_2, X_3 are independent, calculate the expected value and variance of the total volume shipped. [*Hint*: Volume $= 27X_1 + 125X_2 + 512X_3$.]

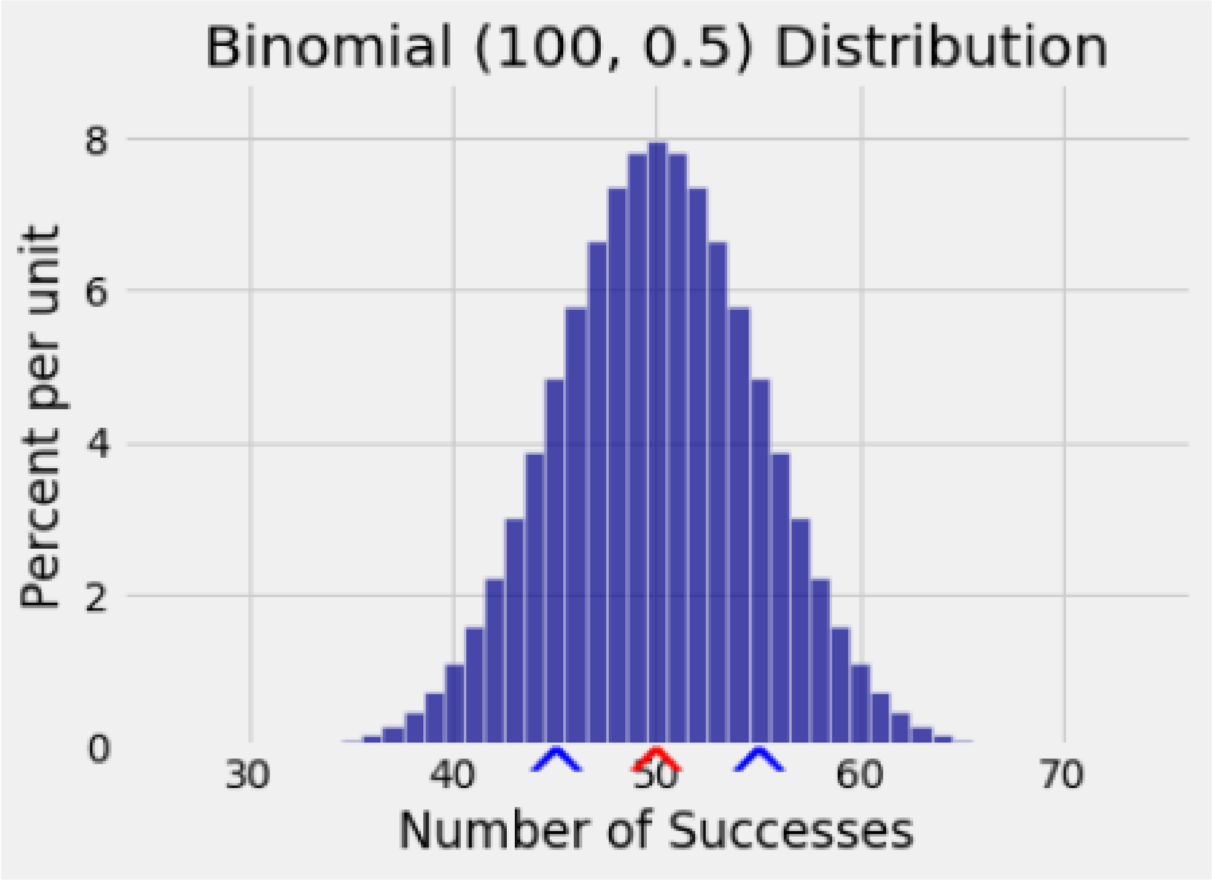
b. Would your calculations necessarily be correct if the X_i's were not independent? Explain.

Variance of the Binomial distribution

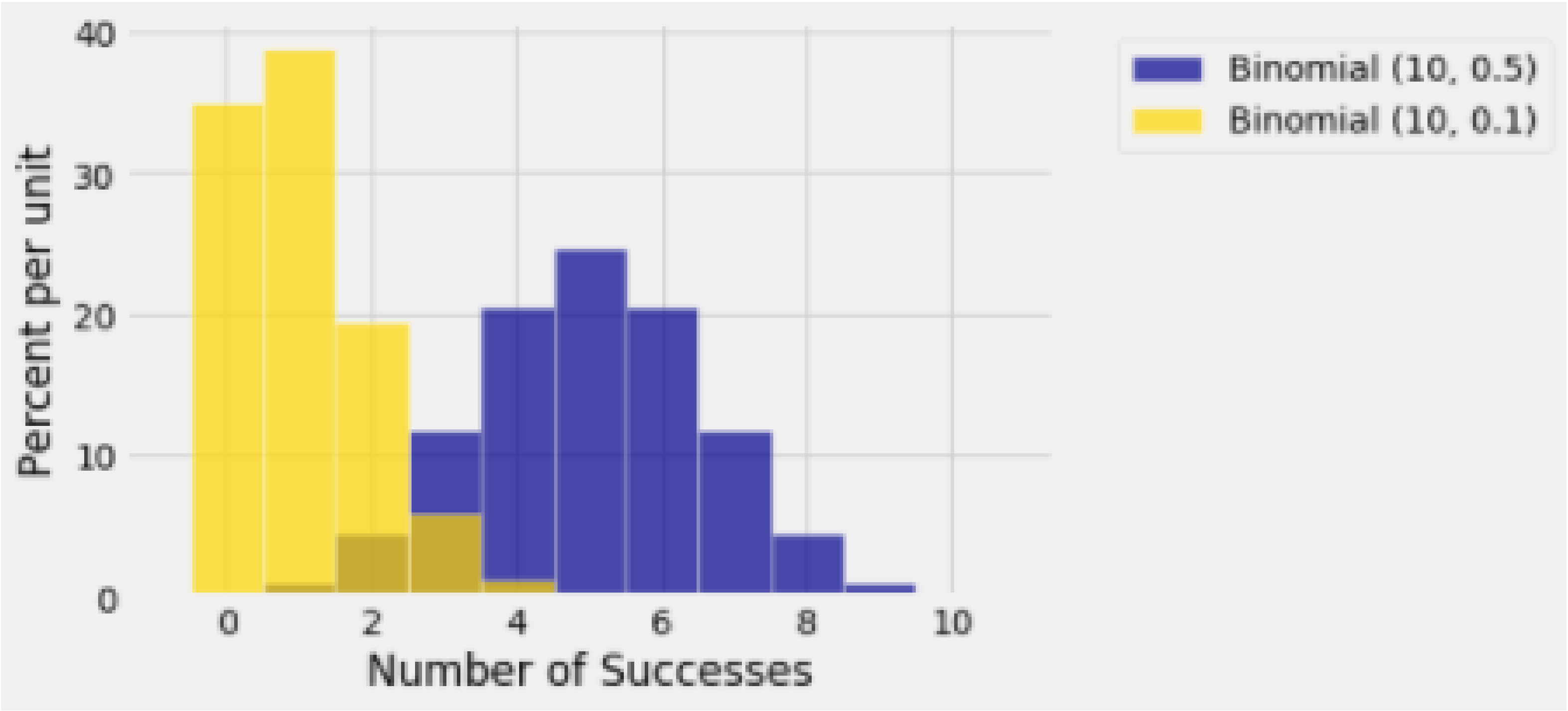
- Recall that a binomial random variable $X \sim Bin(n, p)$ is the sum of n iid Bernoulli(p) random variables $I_1, I_2, \ldots, I_n$ where I_k is the indicator of success on the kth trial.

- What are the mean and variance of I_k? And therefore, what are the mean and variance of X?

Variance of the Binomial distribution



Variance of the Binomial distribution



Variance of Poisson (μ)

- Recall that one way to get the Poisson rv is by approximating the Binomial(n, p) distribution when n is large and p is small. $(\mu = np)$

- SD of the binomial distribution is $\sqrt{np(1-p)}$.

- Note that if p is small, $(1-p) \approx 1$, and we can say that $np(1-p) \approx np$.

- This gives us that the SD of the Poisson(μ) distribution is $\sqrt{\mu}$

Variance of Geometric(p) and Negative Binomial

- **Geometric:** Number of trials until the first success
- The variance of the geometric distribution is $\frac{1-p}{p^2}$
- **Negative Binomial:** Number of trials until we have r successes
- The variance of the negative binomial distribution is $\frac{r(1-p)}{p^2}$

Exercise 7.4.5

The number of typos on the cover page of an exam has a distribution given by

x	0	1
$P(X = x)$	0.8	0.2

The number of misprints in the rest of the exam has the Poisson(3) distribution, independently of the cover page.

Find the expectation and SD of the total number of misprints on the exam.