Central Limit Thm, Normal Approximations

Engineering Statistics
Section 5.4

Josh Engwer

TTU

23 March 2016

PART I:

CENTRAL LIMIT THEOREM

Expected Value & Variance of a Sum of iid rv's

Proposition

Let $X_1, \ldots, X_n$ be a random sample from a population and $c_1, \ldots, c_n \neq 0$. Then:

- $\mathbb{E}[c_1X_1 + \cdots + c_nX_n] = c_1\mathbb{E}[X_1] + \cdots + c_n\mathbb{E}[X_n]$
- $\mathbb{V}[c_1X_1 + \cdots + c_nX_n] = c_1^2\mathbb{V}[X_1] + \cdots + c_n^2\mathbb{V}[X_n]$

PROOF: CASE I $(n = 2)$: $X_1, X_2 \overset{iid}{\sim}$ (discrete population)

$$\mathbb{E}[c_1X_1 + c_2X_2] := \sum_{(j,k)\in \text{Supp}(X_1,X_2)} \sum (c_1j + c_2k) \cdot p_{(X_1,X_2)}(j,k)$$

$$\overset{iid}{=} \sum_{j\in \text{Supp}(X_1)} \sum_{k\in \text{Supp}(X_2)} (c_1j + c_2k) \cdot p_{X_1}(j) \cdot p_{X_2}(k)$$

$$= c_1 \sum_{j\in \text{Supp}(X_1)} j \cdot p_{X_1}(j) + c_2 \sum_{k\in \text{Supp}(X_2)} k \cdot p_{X_2}(k)$$

$$:= c_1\mathbb{E}[X_1] + c_2\mathbb{E}[X_2]$$

Expected Value & Variance of a Sum of iid rv's

Proposition

Let $X_1, \ldots, X_n$ be a random sample from a population and $c_1, \ldots, c_n \neq 0$. Then:

- $\mathbb{E}[c_1X_1 + \cdots + c_nX_n] = c_1\mathbb{E}[X_1] + \cdots + c_n\mathbb{E}[X_n]$
- $\mathbb{V}[c_1X_1 + \cdots + c_nX_n] = c_1^2\mathbb{V}[X_1] + \cdots + c_n^2\mathbb{V}[X_n]$ *(requires **independence**)*

PROOF: CASE II $(n = 2)$: $X_1, X_2 \overset{iid}{\sim}$ (continuous population)

$$\mathbb{E}[c_1X_1 + c_2X_2] := \iint_{\text{Supp}(X_1,X_2)} (c_1x_1 + c_2x_2) \cdot f_{(X_1,X_2)}(x_1, x_2)\, dx_1dx_2$$

$$\overset{iid}{=} \int_{\text{Supp}(X_1)} \int_{\text{Supp}(X_2)} (c_1x_1 + c_2x_2) \cdot f_{X_1}(x_1) \cdot f_{X_2}(x_2)\, dx_1dx_2$$

$$= c_1 \int_{\text{Supp}(X_1)} x_1 \cdot f_{X_1}(x_1)\, dx_1 + c_2 \int_{\text{Supp}(X_2)} x_2 \cdot f_{X_2}(x_2)\, dx_2$$

$$:= c_1\mathbb{E}[X_1] + c_2\mathbb{E}[X_2]$$

Properties of Sample Mean

Proposition

Let $X_1, \ldots, X_n$ be a random sample from a distribution with mean μ and variance σ^2. Then:

- $\mu_{\overline{X}} = \mathbb{E}[\overline{X}] = \mu$
- $\sigma^2_{\overline{X}} = \mathbb{V}[\overline{X}] = \sigma^2/n$
- $\sigma_{\overline{X}} = \sigma/\sqrt{n}$

PROOF:

$$\begin{aligned}\mathbb{E}[\overline{X}] &= \mathbb{E}\left[\tfrac{1}{n}(X_1 + \cdots + X_n)\right] = \tfrac{1}{n}\left[\mathbb{E}[X_1] + \cdots + \mathbb{E}[X_n]\right] \\ &= \tfrac{1}{n}[\mu + \cdots + \mu] = \tfrac{1}{n}[n\mu] = \mu\end{aligned}$$

$$\begin{aligned}\mathbb{V}[\overline{X}] &= \mathbb{V}\left[\tfrac{1}{n}(X_1 + \cdots + X_n)\right] = \tfrac{1}{n^2}\left[\mathbb{V}[X_1] + \cdots + \mathbb{V}[X_n]\right] \\ &= \tfrac{1}{n^2}[\sigma^2 + \cdots + \sigma^2] = \tfrac{1}{n^2}[n\sigma^2] = \sigma^2/n\end{aligned}$$

$\sigma_{\overline{X}} = \sqrt{\mathbb{V}[X]} = \sqrt{\sigma^2/n} = \sigma/\sqrt{n}$ □

Properties of Sample Total

Proposition

Let $X_1, \ldots, X_n$ be a random sample from a distribution with mean μ and variance σ^2. Then:

- $\mu_{X_1+\cdots+X_n} = \mathbb{E}[X_1 + \cdots + X_n] = n\mu$
- $\sigma^2_{X_1+\cdots+X_n} = \mathbb{V}[X_1 + \cdots + X_n] = n\sigma^2$
- $\sigma_{X_1+\cdots+X_n} = \sigma\sqrt{n}$

PROOF:

$$\mathbb{E}[X_1 + \cdots + X_n] = \mathbb{E}\left[n\overline{X}\right] = n\mathbb{E}\left[\overline{X}\right] = n \cdot \mu = n\mu$$

$$\mathbb{V}[X_1 + \cdots + X_n] = \mathbb{V}\left[n\overline{X}\right] = n^2\mathbb{V}\left[\overline{X}\right] = n^2[\sigma^2/n] = n\sigma^2$$

$$\sigma_{X_1+\cdots+X_n} = \sqrt{\mathbb{V}[X_1 + \cdots + X_n]} = \sqrt{n\sigma^2} = \sigma\sqrt{n} \qquad \square$$

Sample Mean & Sample Total of a Random Sample from a Normal Population

The proceeding properties of sample means & sample variances can be applied to particular population distributions:

Proposition

Let random sample $X_1, \ldots, X_n \overset{iid}{\sim}$ *Normal*(μ, σ^2). *Then:*

- *For any sample size* $n > 1$, $\overline{X} \sim$ *Normal*$(\mu, \sigma^2/n)$
- *For any sample size* $n > 1$, $X_1 + \cdots + X_n \sim$ *Normal*$(n\mu, n\sigma^2)$

PROOF: $X \sim \text{Normal}(\mu, \sigma^2) \implies \mathbb{E}[X] = \mu, \quad \mathbb{V}[X] = \sigma^2$

$$\begin{aligned} \mathbb{E}[\overline{X}] &= \mathbb{E}\left[\tfrac{1}{n}(X_1 + \cdots + X_n)\right] = \tfrac{1}{n}\left[\mathbb{E}[X_1] + \cdots + \mathbb{E}[X_n]\right] \\ &= \tfrac{1}{n}[\mu + \cdots + \mu] = \tfrac{1}{n}[n\mu] = \mu \end{aligned}$$

$$\begin{aligned} \mathbb{V}[\overline{X}] &= \mathbb{V}\left[\tfrac{1}{n}(X_1 + \cdots + X_n)\right] = \tfrac{1}{n^2}\left[\mathbb{V}[X_1] + \cdots + \mathbb{V}[X_n]\right] \\ &= \tfrac{1}{n^2}[\sigma^2 + \cdots + \sigma^2] = \tfrac{1}{n^2}[n\sigma^2] = \sigma^2/n \end{aligned}$$

$\therefore \quad \overline{X} \sim \text{Normal}(\mu, \sigma^2/n)$ $\square$

Sample Mean & Sample Total of a Random Sample from a Normal Population

The proceeding properties of sample means & sample variances can be applied to particular population distributions:

Proposition

Let random sample $X_1, \ldots, X_n \overset{iid}{\sim}$ *Normal*(μ, σ^2). *Then:*

- *For any sample size* $n > 1$, $\overline{X} \sim$ *Normal*$(\mu, \sigma^2/n)$
- *For any sample size* $n > 1$, $X_1 + \cdots + X_n \sim$ *Normal*$(n\mu, n\sigma^2)$

PROOF: $X \sim \text{Normal}(\mu, \sigma^2) \implies \mathbb{E}[X] = \mu, \quad \mathbb{V}[X] = \sigma^2$

$$\mathbb{E}[X_1 + \cdots + X_n] = \mathbb{E}\left[n\overline{X}\right] = n\mathbb{E}\left[\overline{X}\right] = n \cdot \mu = n\mu$$

$$\mathbb{V}[X_1 + \cdots + X_n] = \mathbb{V}\left[n\overline{X}\right] = n^2\mathbb{V}\left[\overline{X}\right] = n^2[\sigma^2/n] = n\sigma^2$$

$$\therefore \quad X_1 + \cdots + X_n \sim \text{Normal}(n\mu, n\sigma^2) \qquad \square$$

Central Limit Theorem (CLT)

The **Central Limit Thm** is considered a fundamental theorem of Statistics:

Theorem

(Central Limit Theorem)

Let $X_1, \ldots, X_n$ be a random sample from a non-normal distribution with mean μ and variance σ^2. Then:

- *The sample mean $\overline{X}$ is approximately normal as follows:*

$$\overline{X} \overset{approx}{\sim} Normal(\mu, \sigma^2/n)$$

- *The sample total $X_1 + \cdots + X_n$ is approximately normal as follows:*

$$X_1 + \cdots + X_n \overset{approx}{\sim} Normal(n\mu, n\sigma^2)$$

Requirement for this normal approximation to be valid: $n > 30$

The larger the sample size n, the better the approximation.

PROOF: Beyond the scope of this course.

PART II:

NORMAL APPROXIMATION TO THE BINOMIAL

The Sum of several iid Bernoulli(p) rv's

Proposition

Let random sample $X_1, \ldots, X_m \overset{iid}{\sim}$ *Bernoulli*(p). *Then:*

- *For any sample size* $n > 1$, $X_1 + \cdots + X_n \sim$ *Binomial*(n, p)

PROOF: $X \sim \text{Bernoulli}(p) \implies \mathbb{E}[X] = p, \quad \mathbb{V}[X] = pq \qquad (q = 1 - p)$

$$\mathbb{E}[X_1 + \cdots + X_n] = \mathbb{E}[X_1] + \cdots + \mathbb{E}[X_n] \overset{iid}{=} n \cdot \mathbb{E}[X_1] = n \cdot p = np$$

$$\mathbb{V}[X_1 + \cdots + X_n] \overset{iid}{=} \mathbb{V}[X_1] + \cdots + \mathbb{V}[X_n] \overset{iid}{=} n \cdot \mathbb{V}[X_1] = n \cdot pq = npq$$

$\therefore \quad X_1 + \cdots + X_n \sim \text{Binomial}(n, p)$ $\square$

Normal Approximation to the Binomial

Corollary

Let $X \sim$ *Binomial*(n, p). *Then:*

$$X \overset{approx}{\sim} \textit{Normal}(\mu = np, \sigma^2 = npq) \qquad (\textit{where } q = 1 - p)$$

Requirement for this normal approximation to be valid: $\min\{np, nq\} \geq 10$

i.e. It's required that both $np \geq 10$ *and* $nq \geq 10$.

NOTE: Remember that normal distributions are symmetric.
If $\min\{np, nq\} < 10$, then the binomial distribution is too skewed.

PROOF:
Let $X_1, \ldots, X_n \sim$ Bernoulli$(p) \implies \mu_{X_1} = \mathbb{E}[X_1] = p$ and $\sigma^2_{X_1} = \mathbb{V}[X_1] = pq$.

Then $X := X_1 + \cdots + X_n \sim$ Binomial(n, p).

Moreover, the CLT asserts that $X_1 + \cdots + X_n \overset{approx}{\sim}$ Normal$(\mu = n\mu_{X_1}, \sigma^2 = n\sigma^2_{X_1})$

$\therefore \quad X \sim$ Binomial$(n, p) \implies X \overset{approx}{\sim}$ Normal$(\mu = np, \sigma^2 = npq)$ $\square$

Normal Approximation to Binomial Probability

Corollary

Let $X \sim$ *Binomial*(n, p). *Then:*

$$\mathbb{P}(X \leq x) = \textit{Bi}(x; n, p) \approx \Phi\left(\frac{x + 0.5 - np}{\sqrt{npq}}\right) \qquad (\textit{where } \; q = 1 - p)$$

Requirement for this normal approximation to be valid: $\min\{np, nq\} \geq 10$

NOTE: The ***continuity correction term*** *"*$+ 0.5$*" improves the approximation.*

PROOF: Assume that the requirement $\min\{np, nq\} \geq 10$ is satisfied. Then:

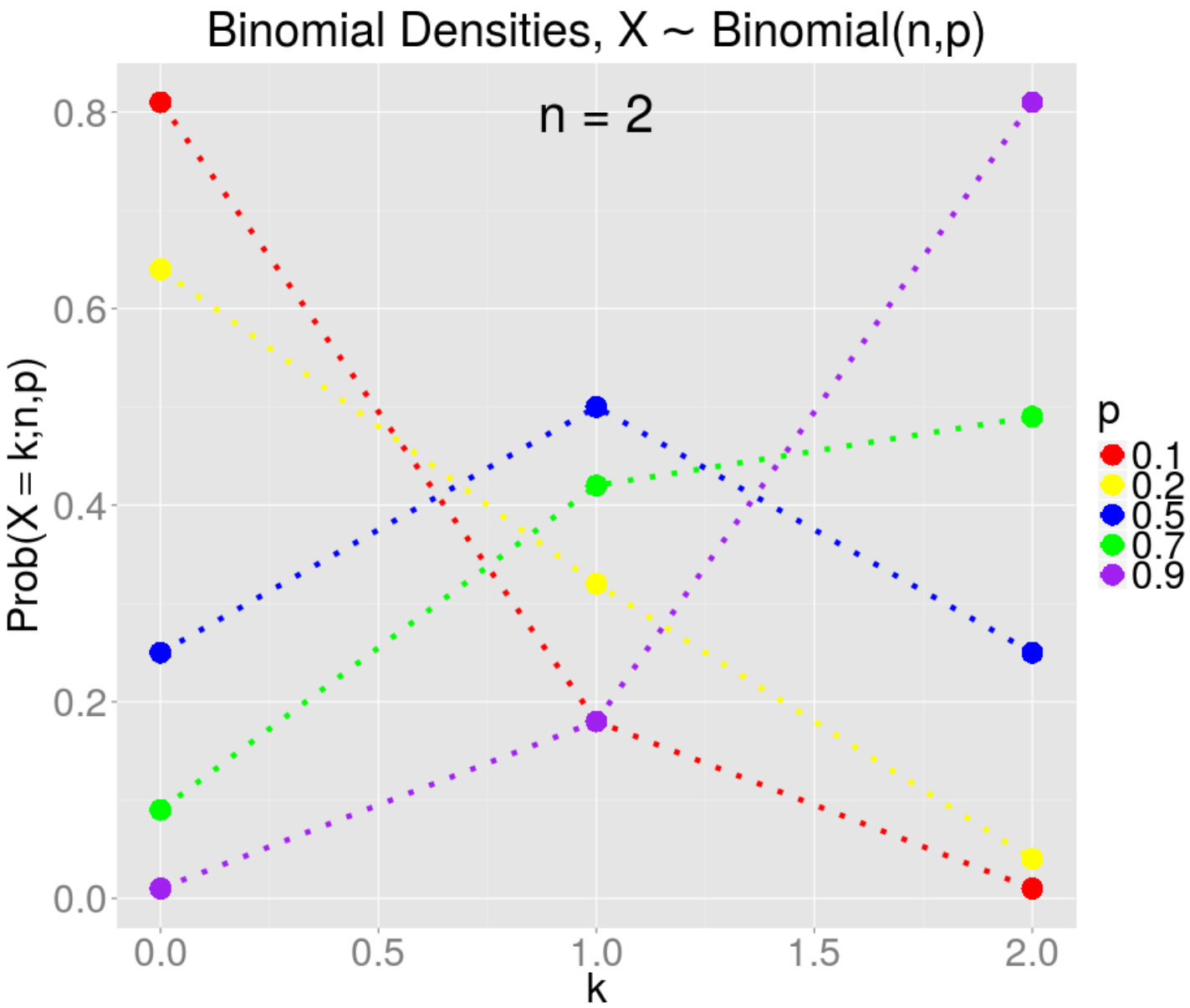
$$X \sim \text{Binomial}(n, p) \implies X \overset{approx}{\sim} \text{Normal}(\mu = np, \sigma^2 = npq)$$

$$\implies \text{Bi}(x; n, p) = \mathbb{P}(X \leq x) \approx \mathbb{P}\left(Z \leq \frac{x - \mu}{\sigma}\right) = \Phi\left(\frac{x - \mu}{\sigma}\right) = \Phi\left(\frac{x - np}{\sqrt{npq}}\right)$$

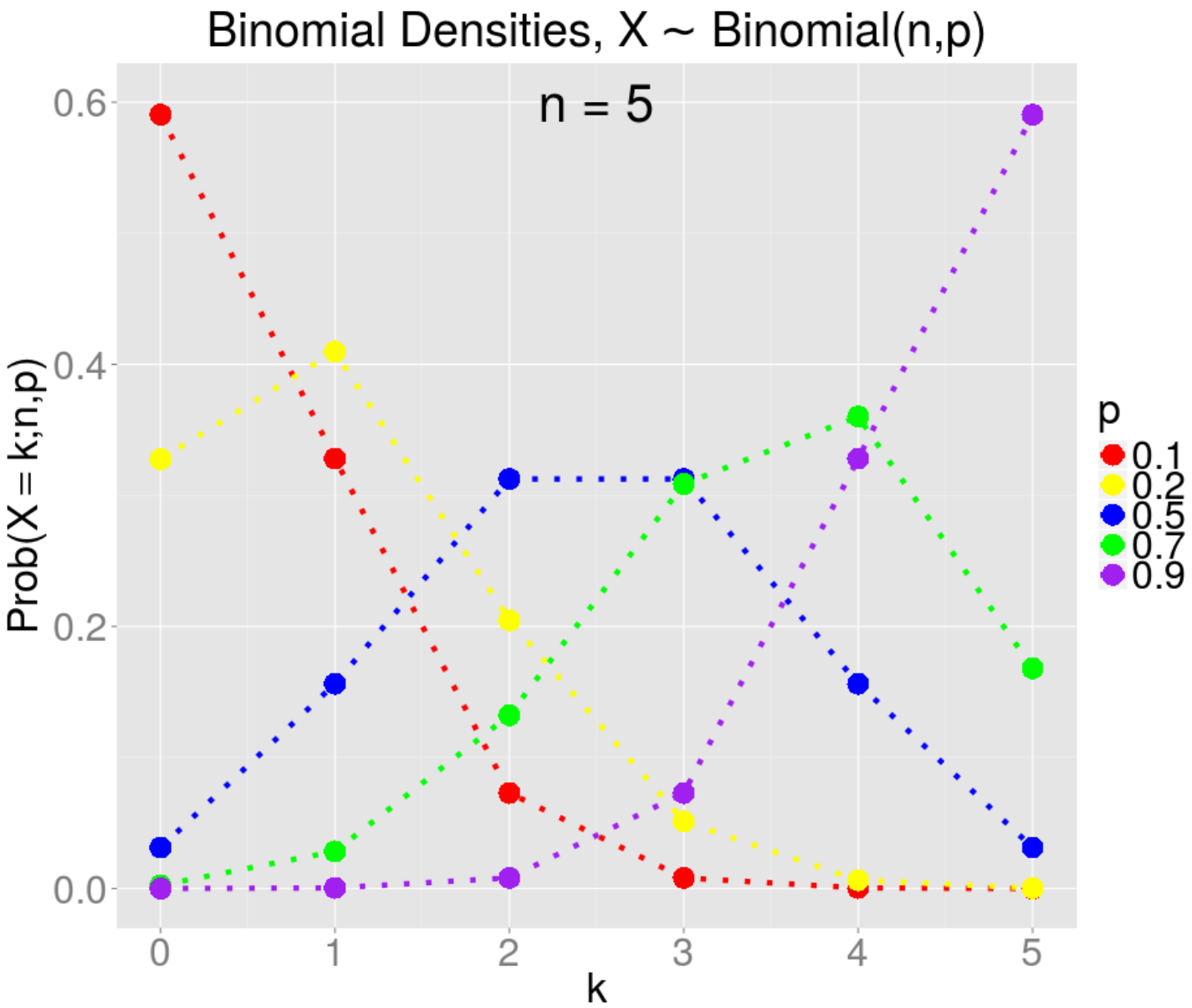
Add continuity correction "$+ 0.5$" in numerator of Φ to better the approximation.

□

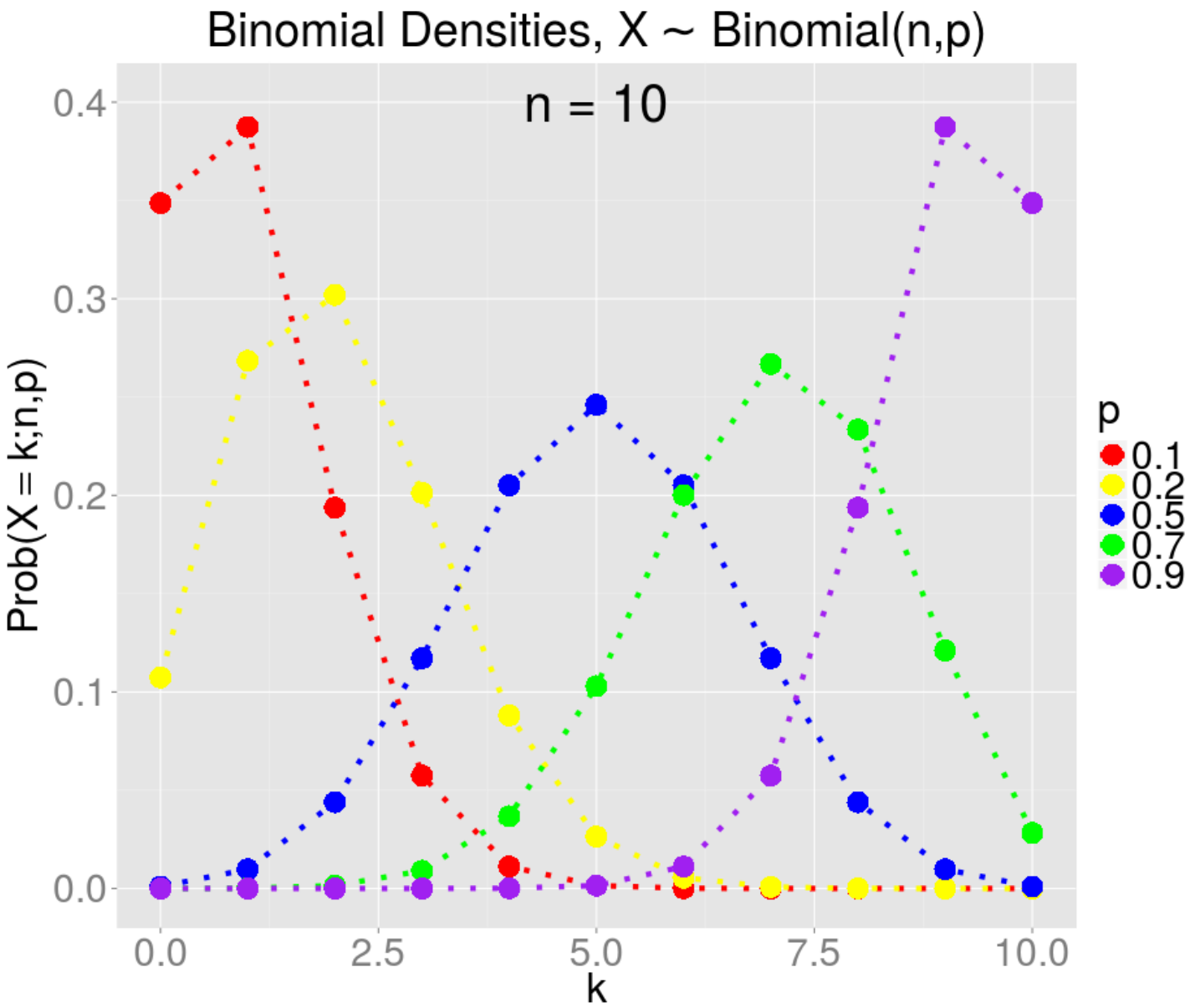
Binomial Density Plots (pmf's) for Sample Size $n = 2$



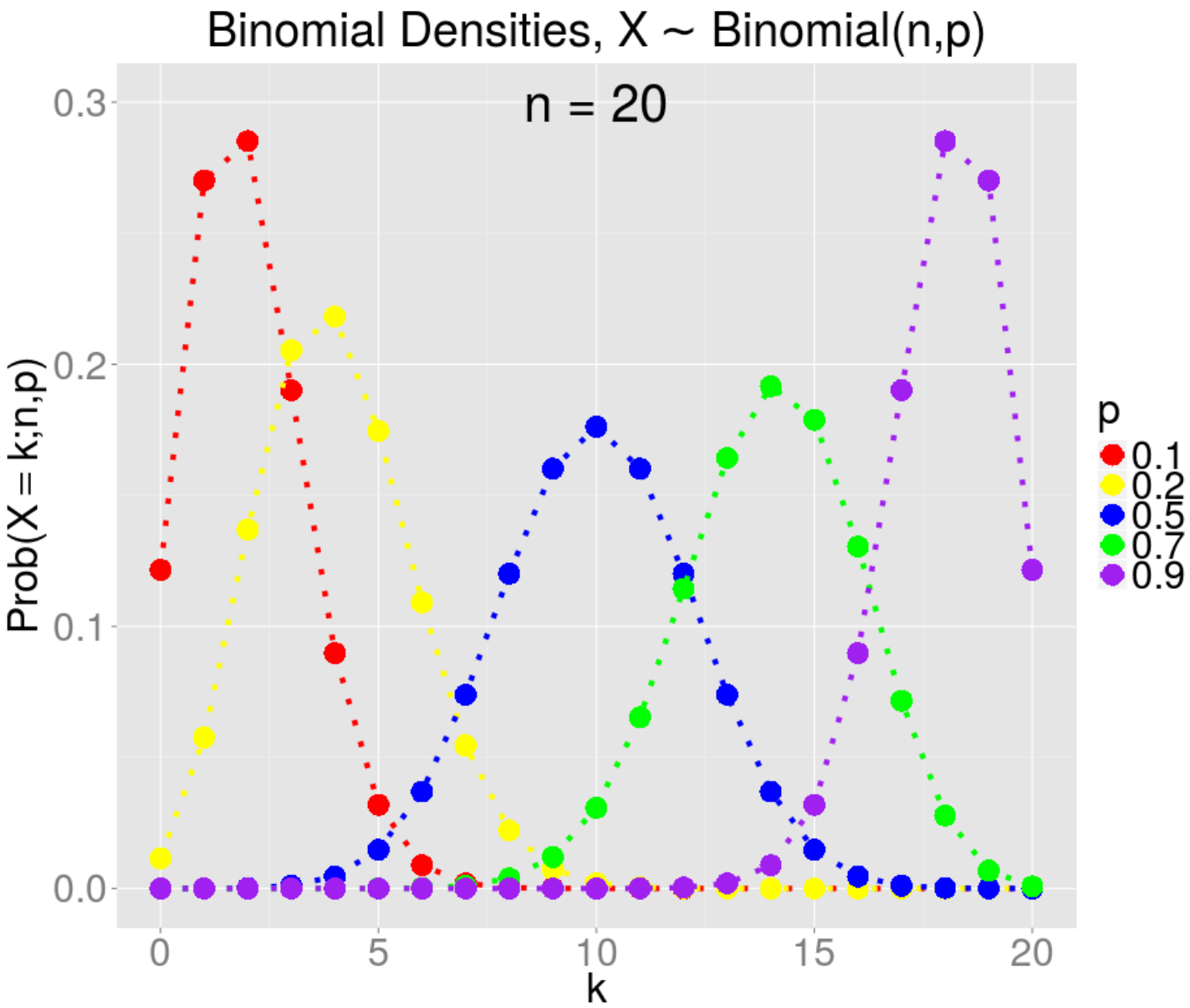
Binomial Density Plots (pmf's) for Sample Size $n = 5$



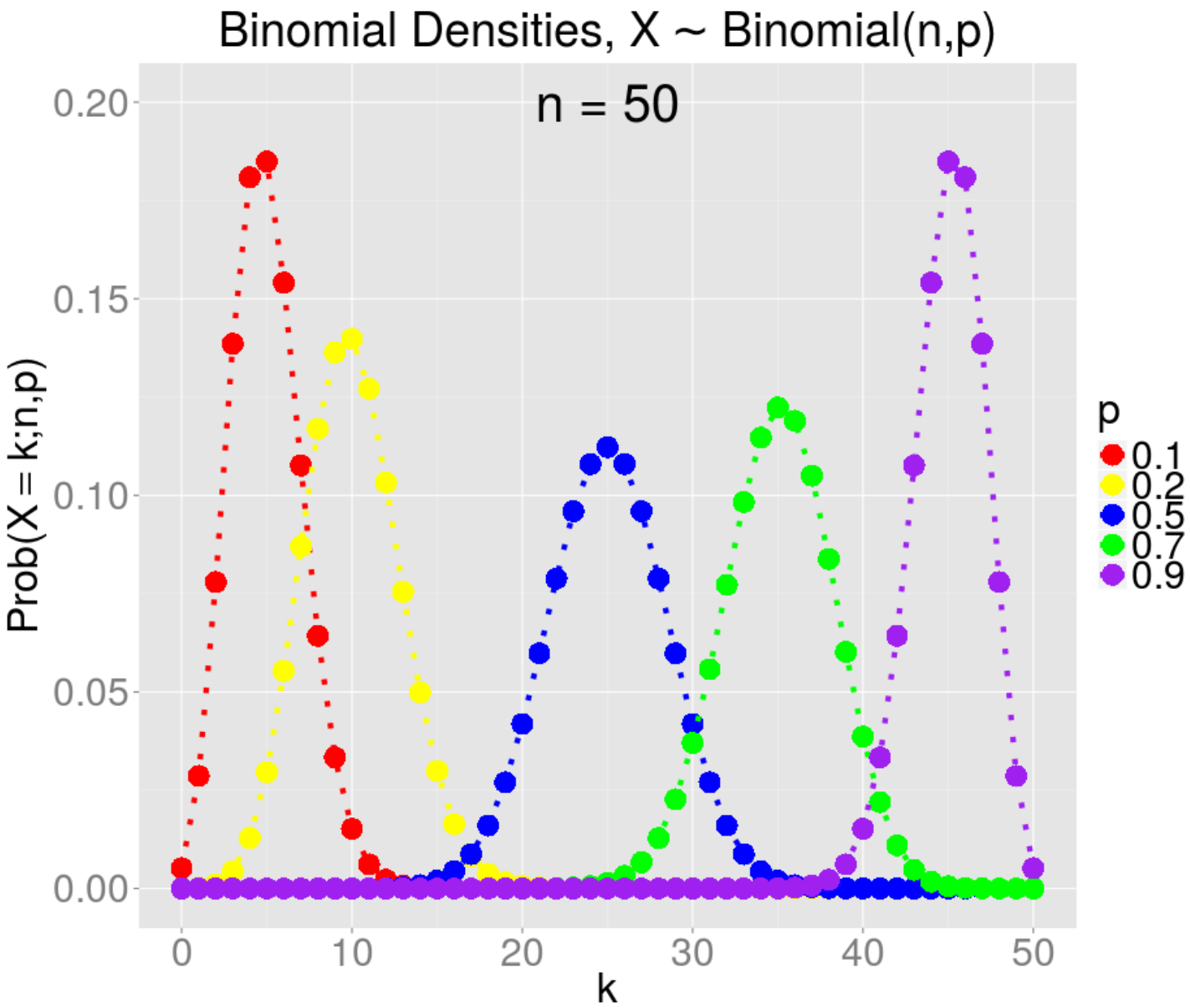
Binomial Density Plots (pmf's) for Sample Size $n = 10$



Binomial Density Plots (pmf's) for Sample Size $n = 20$



Binomial Density Plots (pmf's) for Sample Size $n = 50$



PART III:

NORMAL APPROXIMATION TO THE POISSON

The Sum of several iid Poisson(λ) rv's

The proceeding properties of sample totals can be applied to particular population distributions:

Proposition

Let random sample $X_1, \ldots, X_n \overset{iid}{\sim}$ *Poisson*(λ). *Then:*

- *For any sample size* $n > 1$, $X_1 + \cdots + X_n \sim$ *Poisson*$(n\lambda)$

PROOF: $X \sim \text{Poisson}(\lambda) \implies \mathbb{E}[X] = \lambda, \quad \mathbb{V}[X] = \lambda$

$$\mathbb{E}[X_1 + \cdots + X_n] = \mathbb{E}[X_1] + \cdots + \mathbb{E}[X_n] \overset{iid}{=} n \cdot \mathbb{E}[X_1] = n \cdot \lambda = n\lambda$$

$$\mathbb{V}[X_1 + \cdots + X_n] \overset{iid}{=} \mathbb{V}[X_1] + \cdots + \mathbb{V}[X_n] \overset{iid}{=} n \cdot \mathbb{V}[X_1] = n \cdot \lambda = n\lambda$$

$\therefore \quad X_1 + \cdots + X_n \sim \text{Poisson}(n\lambda)$ $\square$

Normal Approximation to the Poisson

Corollary

Let $X \sim$ *Poisson*(λ). *Then:*

$$X \overset{approx}{\sim} \textit{Normal}(\mu = \lambda, \sigma^2 = \lambda)$$

Requirement for this normal approximation to be valid: $\lambda > 20$

NOTE: Remember that normal distributions are symmetric.
If $\lambda \leq 20$, then the binomial distribution is too skewed.

PROOF:
Let $X_1, \ldots, X_n \sim \text{Poisson}(\lambda/n) \implies \mu_{X_1} = \mathbb{E}[X_1] = \lambda/n$ and $\sigma^2_{X_1} = \mathbb{V}[X_1] = \lambda/n$.

Then $X := X_1 + \cdots + X_n \sim \text{Poisson}(\lambda)$.

Moreover, the CLT asserts that $X_1 + \cdots + X_n \overset{approx}{\sim} \text{Normal}(\mu = n\mu_{X_1}, \sigma^2 = n\sigma^2_{X_1})$

$\therefore \quad X \sim \text{Poisson}(\lambda) \implies X \overset{approx}{\sim} \text{Normal}(\mu = \lambda, \sigma^2 = \lambda)$ $\square$

Normal Approximation to Poisson Probability

Corollary

Let $X \sim$ *Poisson*(λ). *Then:*

$$\mathbb{P}(X \leq x) = \textit{Pois}(x; \lambda) \approx \Phi\left(\frac{x + 0.5 - \lambda}{\sqrt{\lambda}}\right)$$

Requirement for this normal approximation to be valid: $\lambda > 20$

NOTE: The ***continuity correction term*** *"*$+ 0.5$*" improves the approximation.*

PROOF: Assume that the requirement $\lambda > 20$ is satisfied. Then:

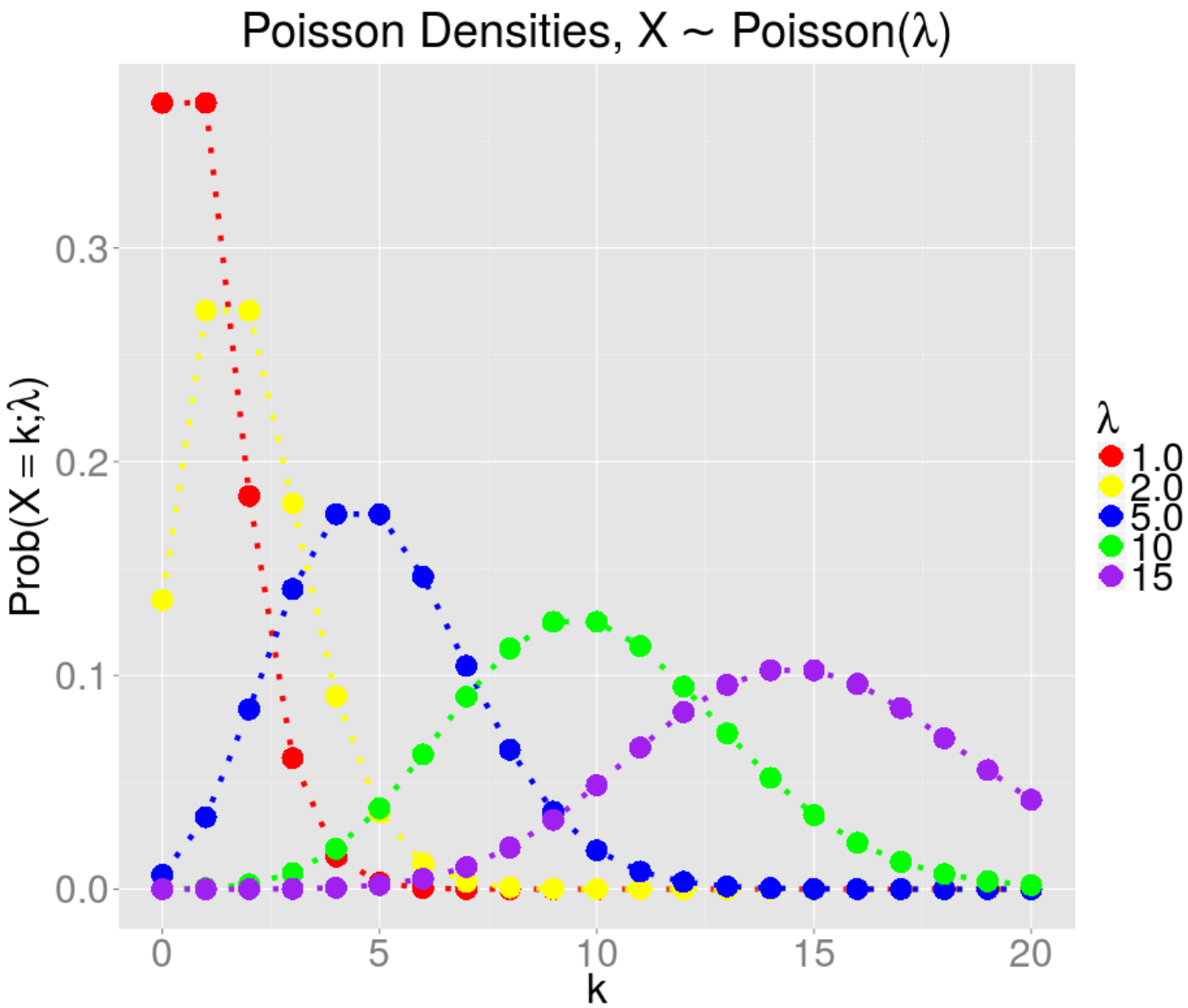
$$X \sim \text{Poisson}(\lambda) \implies X \overset{approx}{\sim} \text{Normal}(\mu = \lambda, \sigma^2 = \lambda)$$

$$\implies \text{Pois}(x; \lambda) = \mathbb{P}(X \leq x) \approx \mathbb{P}\left(Z \leq \frac{x - \mu}{\sigma}\right) = \Phi\left(\frac{x - \mu}{\sigma}\right) = \Phi\left(\frac{x - \lambda}{\sqrt{\lambda}}\right)$$

Add continuity correction "$+ 0.5$" in numerator of Φ to better the approximation.

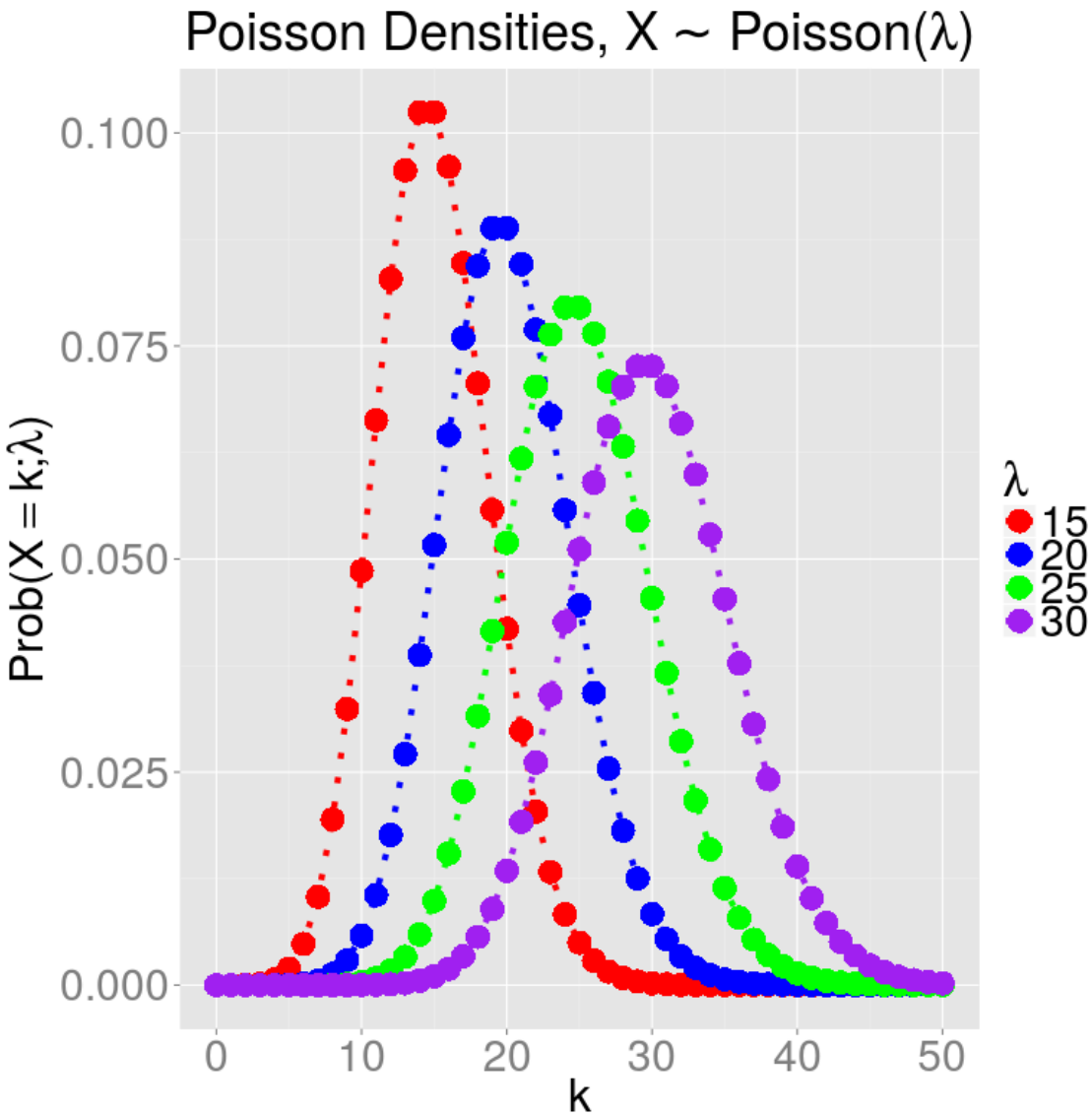
□

Poisson Density Plots (pmf's)



Notice the Poisson density curves for $\lambda = 1, 2, 5, 10, 15$ are skewed.

Poisson Density Plots (pmf's)



Notice the Poisson density curves for $\lambda \geq 20$ are nearly symmetric.

Textbook Logistics for Section 5.4

- Difference(s) in Notation:

CONCEPT	TEXTBOOK NOTATION	SLIDES/OUTLINE NOTATION
Probability of Event	$P(A)$	$\mathbb{P}(A)$
Support of a r.v.	"All possible values of X"	$\text{Supp}(X)$
pmf of a r.v.	$p_X(x)$	$p_X(k)$
Expected Value of r.v.	$E(X)$	$\mathbb{E}[X]$
Variance of r.v.	$V(X)$	$\mathbb{V}[X]$
Sample Total	T_o	$\sum X_k$
pmf of Sample Mean	$p_{\overline{X}}(\overline{x})$	$p_{\overline{X}}(k)$
pmf of Sample Variance	$p_{S^2}(s^2)$	$p_{S^2}(k)$

- Ignore Lognormal Approximation (bottom of pg 236)
  - The Lognormal distribution was part of Section 4.5 that was skipped

Fin

Fin.