ST 501: Fundamentals of Statistical Inference

Discrete Random Variables (part II)

Minh Tang

Fall 2024

Department of Statistics, North Carolina State University.

Bernoulli random variables

Named after Jacob Bernoulli (not to be confused with other Bernoullis such as Daniel, Johann (I, II, III), Nicholas (I, II), ...).

Definition
A random variable X is said to be a Bernoulli random variable with parameter $p \in [0, 1]$ if X has pmf

$$P(X = 0) = 1 - p; \quad P(X = 1) = p.$$

Note The mean and variance of a Bernoulli rv is

$$\mathbb{E}[X] = 0 \times (1 - p) + p \times 1 = p$$
$$\mathbb{E}[X^2] = 0 \times (1 - p) + 1 \times p = p$$
$$\mathrm{Var}[X] = \mathbb{E}[X^2] - (\mathbb{E}[X])^2 = p - p^2 = p(1 - p)$$

Binomial Random Variables

Definition
A binomial experiment on n trials with success probability p is an experiment with the following properties

- The number of trials n is fixed a priori.
- The trials are all identical. and each trial results in one of two possible outcomes e.g., 0 or 1, F or S.
- The probability of success on each trial is p and the probability of failure is $1 - p$.
- The trials are mutually independent.

Definition

A random variable X is said to be a binomial random variable with parameters n and p if X is the random variable for the number of successes in a binomial experiment on n trials with success probability p.

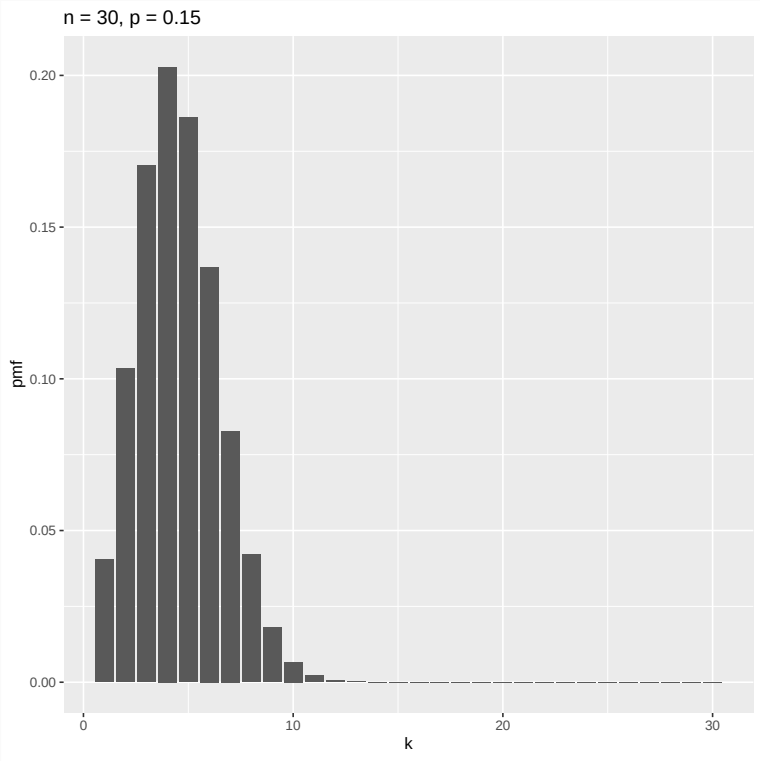
We usually write $X \sim \mathrm{Bin}(n, p)$ to denote a binomial r.v. with parameters n and p.

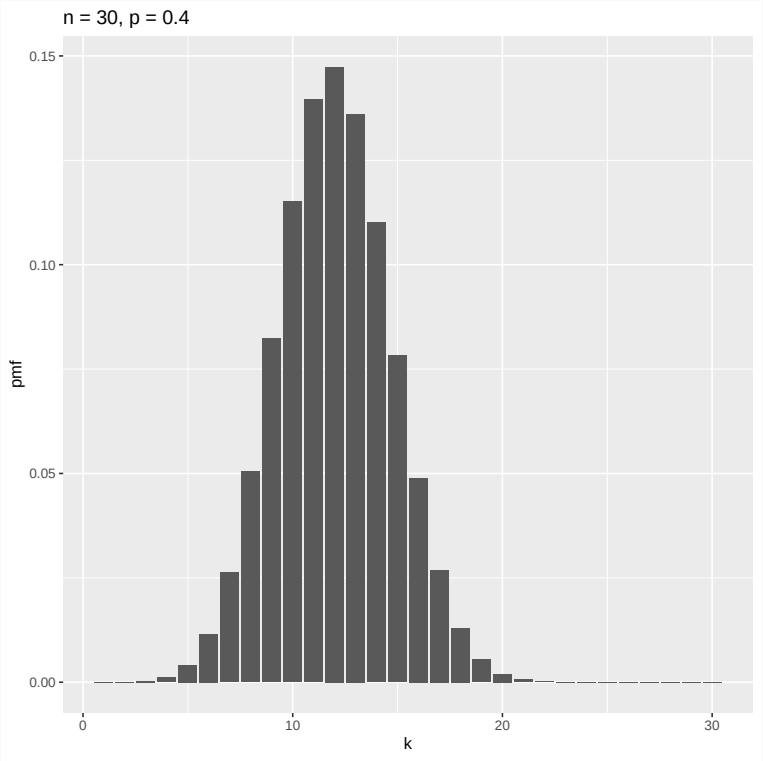
Definition
$X \sim \text{Bin}(n, p)$ if and only if X has pmf p where

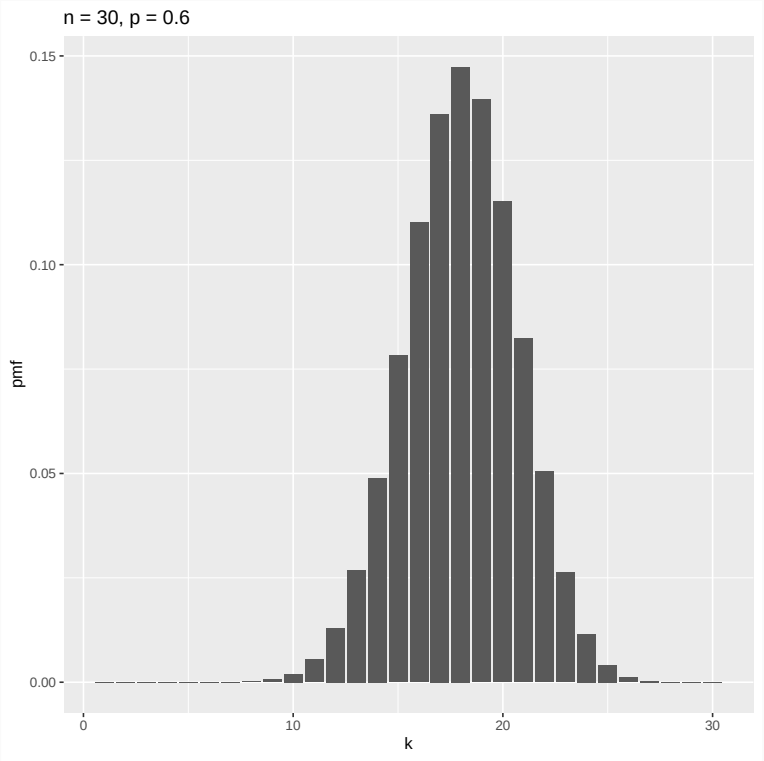
$$p(k) = P(X = k) = \binom{n}{k} p^k (1-p)^{n-k}, \qquad k = 0, 1, 2, \ldots, n.$$

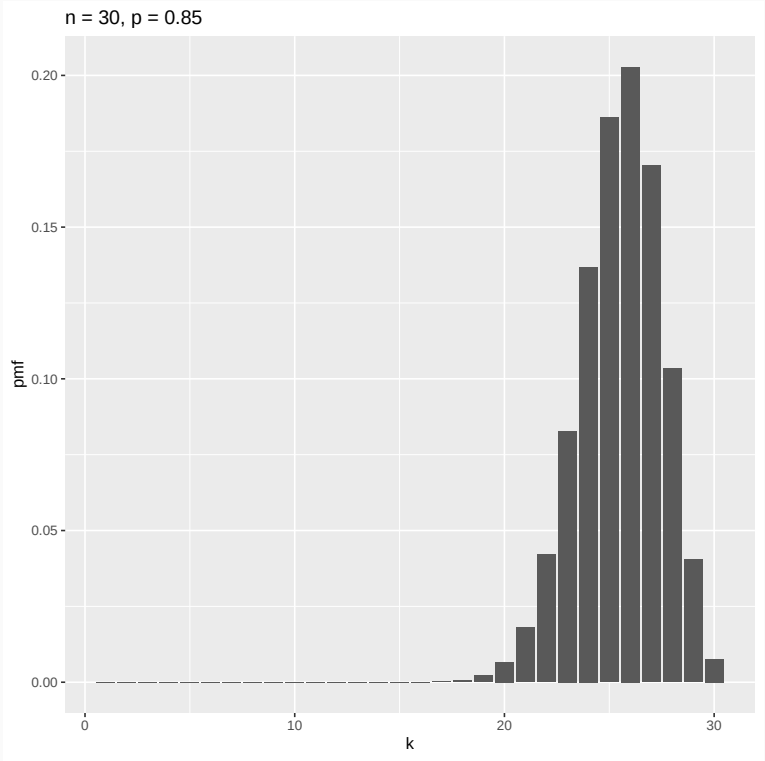
Note p is a valid pmf follows directly from the binomial theorem. More specifically

$$\sum_{k=0}^{n} \binom{n}{k} p^k (1-p)^{n-k} = (p + (1-p))^n = 1^n = 1.$$


n = 30, p = 0.15
pmf
0.20
0.15
0.10
0.05
0.00
0
10
20
30
k


n = 30, p = 0.4
pmf
0.15
0.10
0.05
0.00
0
10
20
30
k


n = 30, p = 0.6
pmf
0.15
0.10
0.05
0.00
0
10
20
30
k


n = 30, p = 0.85
pmf
0.20
0.15
0.10
0.05
0.00
0
10
20
30
k

113. A test for the presence of a certain disease has probability .20 of giving a false-positive reading (indicating that an individual has the disease when this is not the case) and probability .10 of giving a false-negative result. Suppose that ten individuals are tested, five of whom have the disease and five of whom do not. Let X = the number of positive readings that result.

a. Does X have a binomial distribution? Explain your reasoning.

Example A multiple-choice test with 20 questions has four possible answers for each question. A completely unprepared student picks the answer for each question at random and independently. Suppose the instructor has decided that it will take at least 12 correct answers to pass this test. What is the probability that the student pass the test ?

Example A multiple-choice test with 20 questions has four possible answers for each question. A completely unprepared student picks the answer for each question at random and independently. Suppose the instructor has decided that it will take at least 12 correct answers to pass this test. What is the probability that the student pass the test ?

A. Let X be the random variable denoting the number of correct answers. Then $X \sim \mathrm{Bin}(20, 0.25)$. We want $P(X \geq 12)$.

$$P(X = 12) = \binom{20}{12} \times 0.25^{12} \times 0.75^{8} = \frac{125970 \times 3^{8}}{4^{20}} \approx 0.000752,$$

$$P(X = 13) = \binom{20}{13} \times 0.25^{13} \times 0.75^{7} \approx 0.000154,$$

$$P(X = 14) = \binom{20}{14} \times 0.25^{14} \times 0.75^{6} \approx 0.000026,$$

$$P(X \geq 12) = \sum_{k=12}^{20} P(X = k) \approx 0.00094.$$

Example A company order supplies from M distributors and wishes to place n orders. Assume that the company places the orders in a manner that allows every distributor an equal chance of getting any one order. Find the probability that distributor III gets at least $\ell \leq n$ orders.

Example A company order supplies from M distributors and wishes to place n orders. Assume that the company places the orders in a manner that allows every distributor an equal chance of getting any one order. Find the probability that distributor III gets at least $\ell \leq n$ orders.

A. Let X be the random variable for the number of orders placed to distributor III. Then $X \sim \mathrm{Bin}(n, 1/M)$ and hence

$$P(X \geq \ell) = \sum_{k=\ell}^{n} \binom{n}{k} \left(\frac{1}{M}\right)^k \left(1-\frac{1}{M}\right)^{n-k} = \frac{1}{M^n} \sum_{k=\ell}^{n} \binom{n}{k} (M-1)^{n-k}.$$

Example A large lot of electrical fuses is supposed to contain only 5% defective fuses. Suppose we sample $n = 25$ fuses at random from this lot. Find the probability that there are at least 3 defective fuses in this sample.

Example A large lot of electrical fuses is supposed to contain only 5% defective fuses. Suppose we sample $n = 25$ fuses at random from this lot. Find the probability that there are at least 3 defective fuses in this sample.

A. Let X be the r.v. for the number of defective fuses in the sample. Then $X \sim \mathrm{Bin}(25, 0.05)$. We are interested in $P(X \geq 3)$. We have

$$P(X = 0) = \binom{25}{0} \times 0.05^0 \times 0.95^{25} \approx 0.277,$$
$$P(X = 1) = \binom{25}{1} \times 0.05^1 \times 0.95^{24} \approx 0.365,$$
$$P(X = 2) = \binom{25}{2} \times 0.05^2 \times 0.95^{23} \approx 0.23,$$
$$P(X \geq 3) \approx 0.13.$$

Note In general, for $X \sim \text{Bin}(n, p)$, there is no closed form expression for $P(X \geq k)$.

We usually compute $P(X \geq k)$ using statistical software and/or consulting tables. Accurate approximations of $P(X \geq k)$ are also available.

107. Forty percent of seeds from maize (modern-day corn) ears carry single spikelets, and the other 60% carry paired spikelets. A seed with single spikelets will produce an ear with single spikelets 29% of the time, whereas a seed with paired spikelets will produce an ear with single spikelets 26% of the time. Consider randomly selecting ten seeds.

a. What is the probability that exactly five of these seeds carry a single spikelet and produce an ear with a single spikelet?

b. What is the probability that exactly five of the ears produced by these seeds have single spikelets? What is the probability that at most five ears have single spikelets?

For part (a), the probability that a seed carry a single spikelet and produce an ear with a single spikelet is $0.4 \times 0.29 = 0.116$.

As there are ten seeds, we are interested in $P(X = 5)$ where $X \sim \text{Bin}(10, 0.116)$. The answer is

$$P(X = 5) = \binom{10}{5}(0.116)^5(0.884)^5 \approx 0.00286.$$

For part (b), the probability that a randomly selected seed produce an ear with a single spikelet is

$$0.4 \times 0.29 + 0.6 \times 0.26 = 0.272.$$

We are thus interested in $P(Y = 5)$ and $P(Y \leq 5)$ where $Y \sim \text{Bin}(10, 0.272)$.

Expectations and variances of binomial r.v.

Proposition
Let $X \sim \mathrm{Bin}(n, p)$. Then

$$\mathbb{E}[X] = np,$$
$$\mathbb{E}[X^2] = n^2p^2 + np(1-p),$$
$$\mathrm{Var}[X] = np(1-p).$$

Proof (brute force) We start with the observation that

$$k\binom{n}{k} = n\binom{n-1}{k-1}$$

We then have

$$\begin{aligned}
\mathbb{E}[X] &= \sum_{k=1}^{n} k\binom{n}{k}p^k(1-p)^{n-k} \\
&= \sum_{k=1}^{n} n\binom{n-1}{k-1}p^k(1-p)^{n-k} \\
&= \sum_{k=1}^{n} np\binom{n-1}{k-1}p^{k-1}(1-p)^{n-1-(k-1)} \\
&= \sum_{\ell=0}^{n-1} np\binom{n-1}{\ell}p^{\ell}(1-p)^{n-1-\ell} \qquad (\text{let } \ell = k-1) \\
&= np\sum_{\ell=0}^{n-1}\binom{n-1}{\ell}p^{\ell}(1-p)^{n-1-\ell} = np.
\end{aligned}$$

We next evaluate $\mathbb{E}[X^2] = \mathbb{E}[X(X-1)] + \mathbb{E}[X]$. Note that

$$\begin{aligned}
\mathbb{E}[X(X-1)] &= \sum_{k=2}^{n} k(k-1)\tbinom{n}{k} p^k (1-p)^{n-k} \\
&= \sum_{k=0}^{n} n(n-1)\tbinom{n-2}{k-2} p^k (1-p)^{n-k} \\
&= \sum_{k=2}^{n} n(n-1)p^2 \tbinom{n-2}{k-2} p^{k-2} (1-p)^{n-2-(k-2)} \\
&= n(n-1)p^2 \sum_{\ell=0}^{n-2} \tbinom{n-2}{\ell} p^{\ell} (1-p)^{n-2-\ell} = n(n-1)p^2.
\end{aligned}$$

We therefore have

$$\mathbb{E}[X^2] = \mathbb{E}[X(X-1)] + \mathbb{E}[X] = n(n-1)p^2 + np = n^2p^2 + np(1-p)$$
$$\mathrm{Var}[X] = \mathbb{E}[X^2] - (\mathbb{E}[X])^2 = n^2p^2 + np(1-p) - (np)^2 = np(1-p).$$

Sum of independent binomial r.v.

Q. Let $X \sim \mathrm{Bin}(n, p)$ and $Y \sim \mathrm{Bin}(m, p)$. Suppose X and Y are independent, i.e., $P(X = k, Y = \ell) = P(X = k)P(Y = \ell)$ for all k, ℓ. Let $Z = X + Y$. What is the pmf for Z ?

Sum of independent binomial r.v.

Q. Let $X \sim \mathrm{Bin}(n,p)$ and $Y \sim \mathrm{Bin}(m,p)$. Suppose X and Y are independent, i.e., $P(X=k, Y=\ell) = P(X=k)P(Y=\ell)$ for all k, ℓ. Let $Z = X+Y$. What is the pmf for Z ?

Choose an integer z with $0 \le z \le n+m$. Then

$$\begin{aligned} P(Z=z) &= \sum_{k=0}^{n} P(X=k, Y=z-k) \\ &= \sum_{k=0}^{n} P(X=k)P(Y=z-k) \\ &= \sum_{k=0}^{n} \binom{n}{k} p^k (1-p)^{n-k} \times \binom{m}{z-k} p^{z-k} (1-p)^{m-z+k} \\ &= p^z (1-p)^{n+m-z} \sum_{k=0}^{n} \binom{n}{k}\binom{m}{z-k} = \binom{n+m}{z} p^z (1-p)^{n+m-z}. \end{aligned}$$

In summary Z is a $\mathrm{Bin}(n+m,p)$ random variable.

Geometric random variables

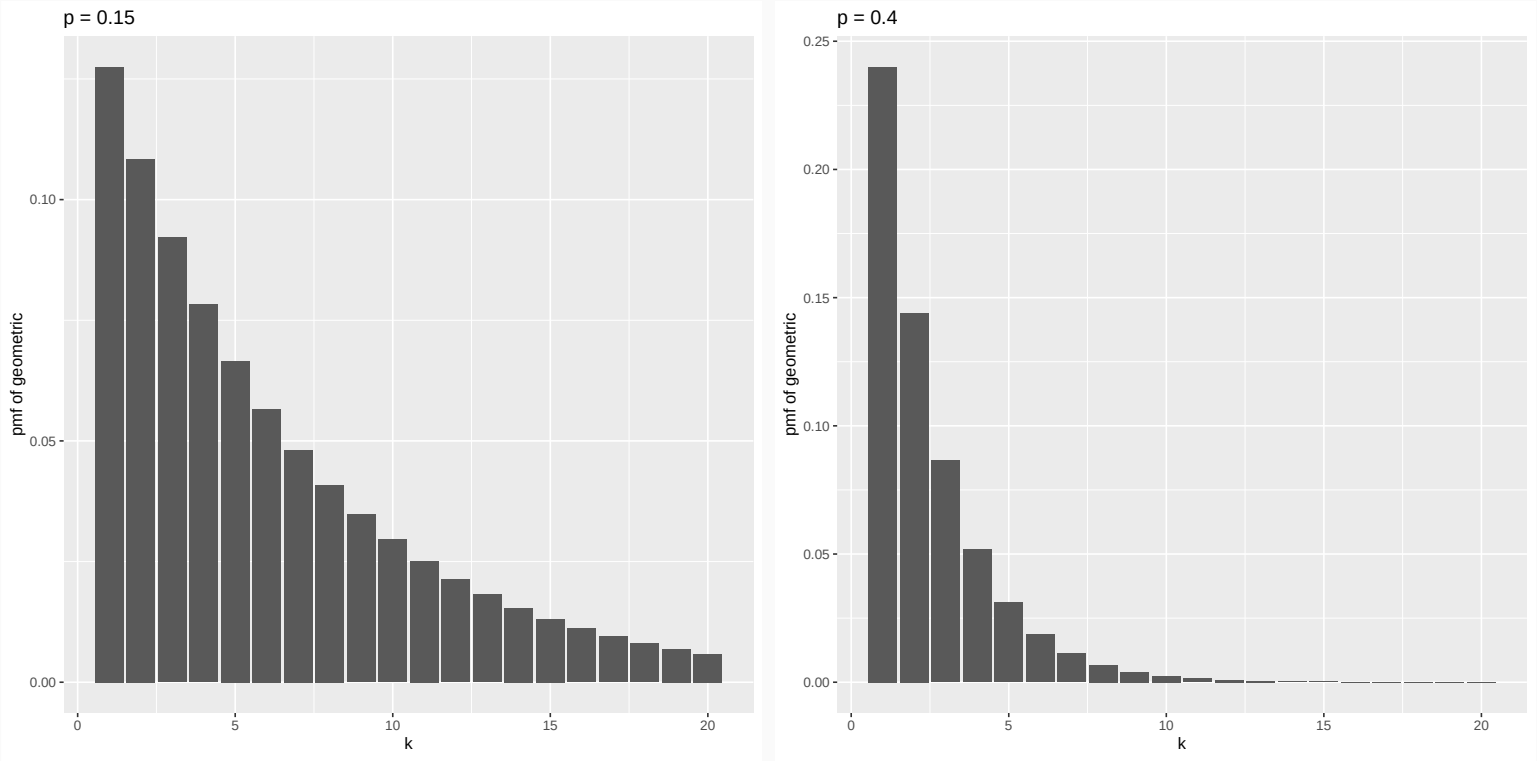
Definition

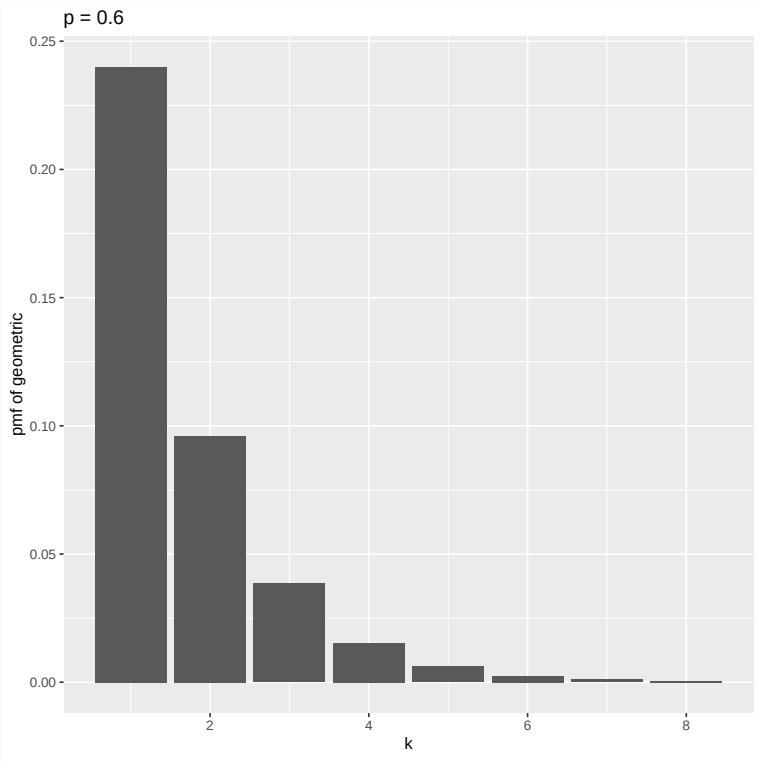
Consider a series of (possibly infinite) independent Bernoulli trials where the probability of success in each trial is identically $p \in [0, 1]$. Let X be the random variable denoting the number of the trial in which the first success occurs. Then X is said to be a geometric random variable with pmf

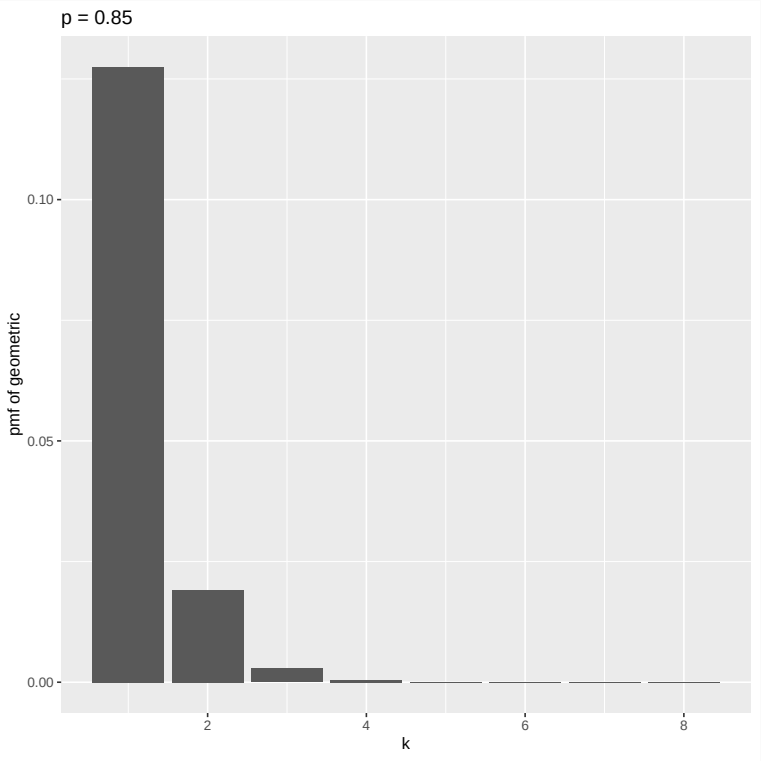
$$p(k) = P(X = k) = (1 - p)^{k-1} p, \qquad k = 1, 2, 3, \ldots$$

Note p is valid pmf follows directly from the formula for a sum of a geometric series, i.e.,

$$\sum_{r=0}^{\infty} q^r = \frac{1}{1 - q} \qquad \text{for } |q| < 1.$$


p = 0.15
pmf of geometric
0.00
0.05
0.10
0
5
10
15
20
k
p = 0.4
pmf of geometric
0.00
0.05
0.10
0.15
0.20
0.25
0
5
10
15
20
k


p = 0.6
0.25
0.20
0.15
0.10
0.05
0.00
pmf of geometric
2
4
6
8
k


p = 0.85
0.10
0.05
0.00
pmf of geometric
2
4
6
8
k

Example Suppose the probability of engine malfunction on a flight during any one-hour period is $p = 0.01$. Find the probability that a single-engine plane will still have a working engine after completing a 14 hours flight.

Example Suppose the probability of engine malfunction on a flight during any one-hour period is $p = 0.01$. Find the probability that a single-engine plane will still have a working engine after completing a 14 hours flight.

A. Let X be the random variable for the number of hours until the first engine failure. Then X is a geometric random variable with $p = 0.01$. We are interested in $P(X \geq 15)$. We have

$$P(X \geq 15) = \sum_{k=15}^{\infty} p(1-p)^{k-1} = \frac{p(1-p)^{14}}{1-(1-p)} = (1-p)^{14} = 0.99^{14} \approx 0.87.$$

Note The above computations yield a useful fact, i.e., for any geometric random variable X with success probability p

$$P(X \geq k) = (1-p)^{k-1},$$

i.e., $X \geq k$ if and only if the first $k-1$ trials are all failures.

Geometric r.v. are memoryless

Proposition
Let X be a geometric r.v. with success probability p. Then for any $k \geq 1, \ell \geq 1$

$$P(X > k + \ell \mid X > k) = \frac{P(X > k + \ell)}{P(X > k)}$$
$$= \frac{(1-p)^{k+\ell}}{(1-p)^k} = (1-p)^\ell = P(X > \ell).$$

That is to say, success are never due anytime soon now, even if you have already waited for quite a long time.

Note Geometric random variables are the only discrete random variable taking values in $\{1, 2, 3, \dots\}$ with the memoryless property.

Proof of the memoryless property

Let X be a rv taking values in $\{1, 2, \ldots, \}$ such that X is memoryless, i.e.,

$$P(X > k + \ell \mid X > k) = P(X \geq \ell)$$

for all $k, \ell \geq 1$.

Then $P(X > k + \ell) = P(X > k) \times P(X > \ell)$ for all k, ℓ. Let $p_* = P(X > 1)$. Then

$$P(X > 2) = p_*^2,$$
$$P(X > 3) = P(X > 1) \times P(X > 2) = p_*^3,$$
$$P(X > k) = P(X > 1) \times P(X > k - 1) = p_*^k$$

This implies

$$P(X = k) = P(X > k - 1) - P(X > k) = p_*^{k-1}(1 - p_*)$$

for all $k \geq 2$. Now $P(X > 1) = 1 - P(X = 1)$ and thus, with $p = P(X = 1)$ we have

$$P(X = k) = (1 - p)^{k-1}p$$

which is the pmf of geometric rv with success probability p.

Mean & variance of geometric r.v.

Proposition
Let X be a geometric r.v. with success probability p. Then

$$\mathbb{E}[X] = \frac{1}{p}; \quad \mathrm{Var}[X] = \frac{1-p}{p^2}.$$

Deriving $\mathbb{E}[X]$ and $\mathrm{Var}[X]$. We start by noting

$$\begin{aligned}\mathbb{E}[X] &= \sum_{k=1}^{\infty} k(1-p)^{k-1}p = \sum_{k=1}^{\infty}(k-1+1)(1-p)^{k-1}p \\ &= (1-p)\sum_{k=1}^{\infty}(k-1)(1-p)^{k-2}p + \sum_{k=1}^{\infty}(1-p)^{k-1}p \\ &= (1-p)\mathbb{E}[X] + 1\end{aligned}$$

Rearranging the above equation yield

$$p\mathbb{E}[X] = 1 \Longrightarrow \mathbb{E}[X] = p^{-1}.$$

Next for $\mathbb{E}[X^2]$ we have

$$\begin{aligned}\mathbb{E}[X^2] &= \sum_{k=1}^{\infty} k^2(1-p)^{k-1}p = \sum_{k=1}^{\infty}\big((k-1)^2+2k-1\big)(1-p)^{k-1}p \\ &= (1-p)\sum_{k=1}^{\infty}(k-1)^2(1-p)^{k-2}p + 2\mathbb{E}[X] - 1 \\ &= (1-p)\mathbb{E}[X^2] + \frac{2}{p} - 1\end{aligned}$$

Rearranging the above equation yield

$$\mathbb{E}[X^2] = \frac{2-p}{p^2}, \quad \mathrm{Var}[X] = \frac{1-p}{p^2}.$$

Example Alice, Bob and Charlie are at a restaurant. They play a game in which, for each round, they each flip a fair coin. If the three coins all show the same faces (such as HHH or TTT) then the game continue. Otherwise, the person whose coin show a face different from the remaining two people pays the bill.

Q. How many rounds is the game expected to last ?

Example Alice, Bob and Charlie are at a restaurant. They play a game in which, for each round, they each flip a fair coin. If the three coins all show the same faces (such as HHH or TTT) then the game continue. Otherwise, the person whose coin show a face different from the remaining two people pays the bill.

Q. How many rounds is the game expected to last ?

Let X be the number of rounds. Then X is a geometric r.v. with success probability $p = 3/4$ and hence $\mathbb{E}[X] = 4/3$.

Negative binomial r.v.

Definition
Let $G_1, G_2, \ldots, G_r$ be independent geometric random variable with a common success probability p. Then

$$X = G_1 + G_2 + \cdots + G_r$$

is said to be a negative binomial r.v. with parameters r (number of successes) and p (probability of success in each trial). We denote this as $X \sim \mathrm{NB}(r, p)$.

Note If $X \sim \mathrm{NB}(r, p)$ then X denote the number of trials at which the rth success first appears in a possibly infinite series of independent Bernoulli trials.

Proposition
Let $X \sim \mathrm{NB}(r, p)$. Then X has pmf

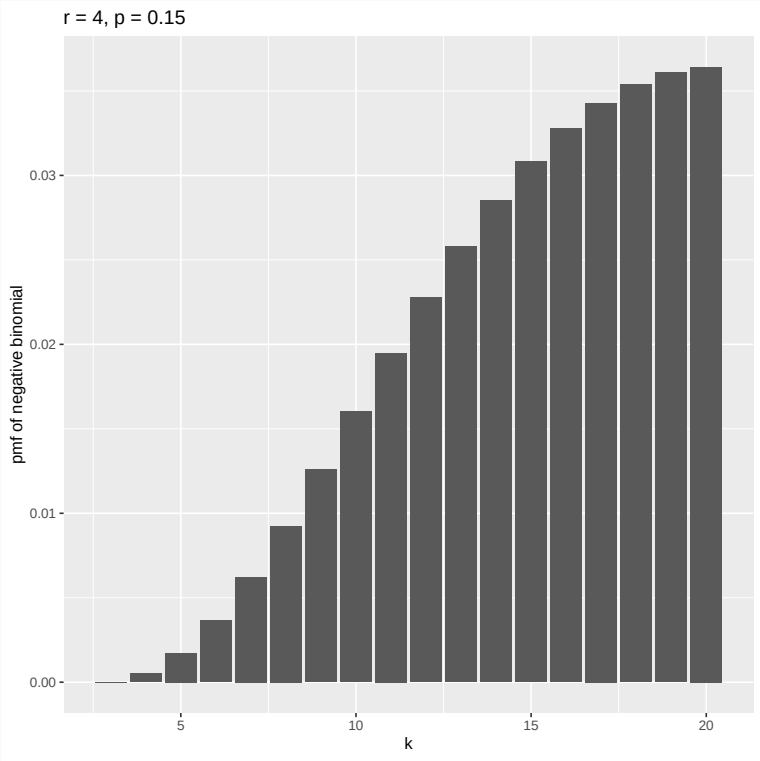
$$p_X(x) = P(X = x) = \binom{x-1}{r-1}(1-p)^{x-r}p^r; \qquad x = r, r+1, r+2, \ldots$$

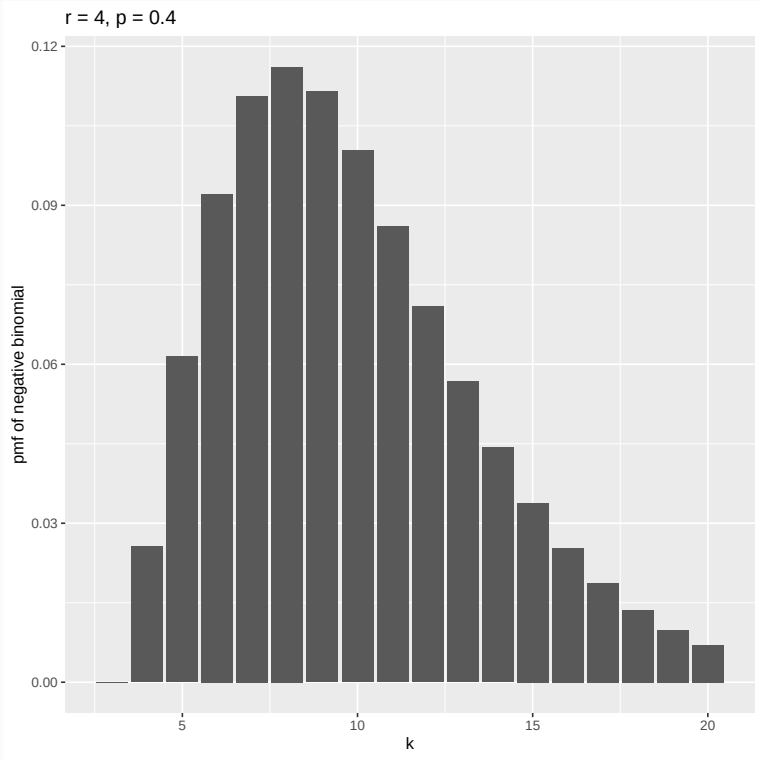
Furthermore, X has mean and variance given by

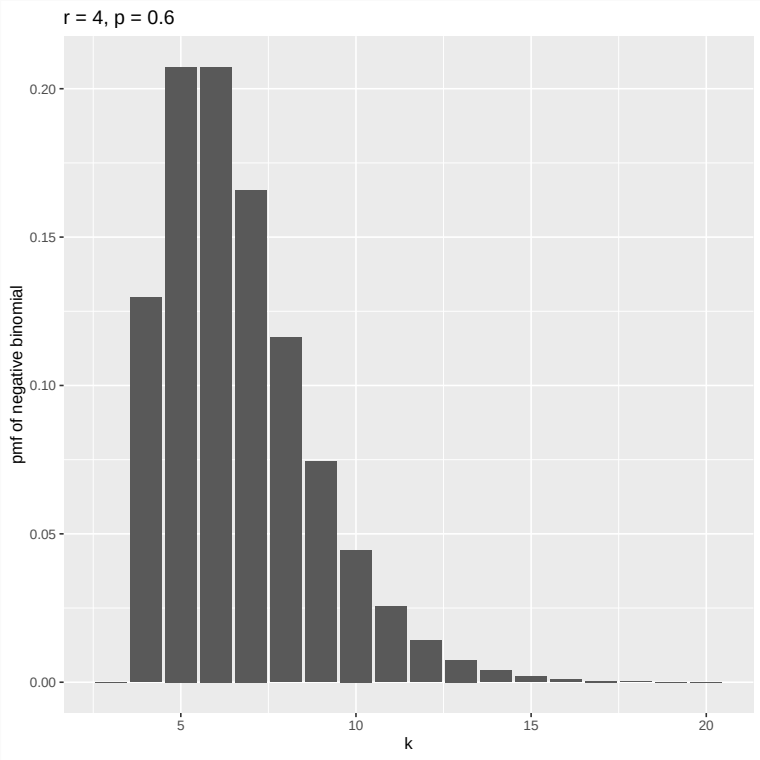
$$\mathbb{E}[X] = \frac{r}{p}; \qquad \mathrm{Var}[X] = \frac{r(1-p)}{p^2}.$$

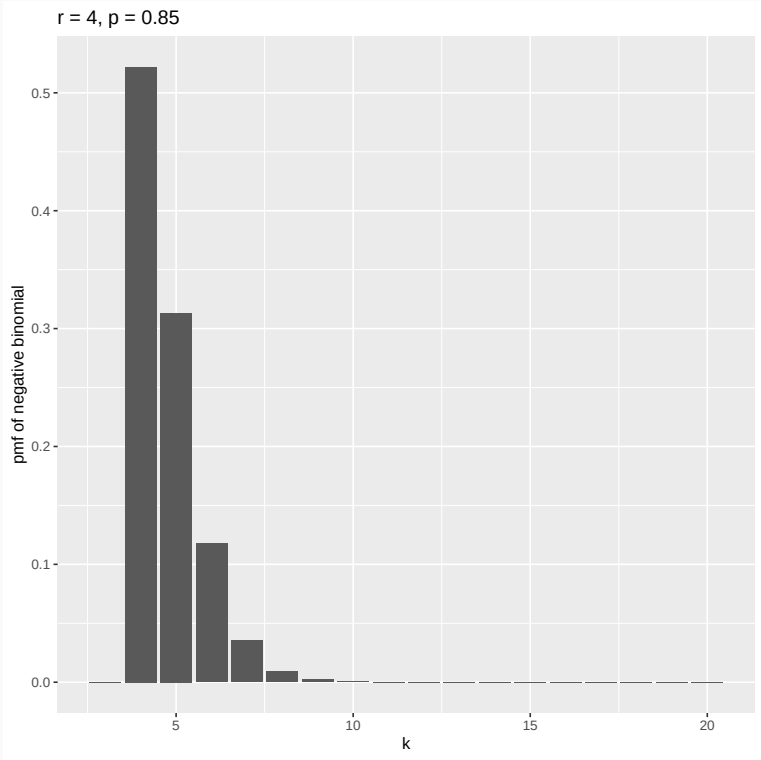
Important Another common definition of a negative binomial r.v. is as $Y = X - r$. Then Y corresponds to the number of failures before observing the rth success. Y then has pmf

$$p_Y(y) = P(Y = y) = \binom{y+r-1}{r-1}(1-p)^y p^r; \qquad y = 0, 1, 2, \ldots$$


r = 4, p = 0.15
pmf of negative binomial
0.00
0.01
0.02
0.03
5
10
15
20
k


r = 4, p = 0.4
pmf of negative binomial
0.00
0.03
0.06
0.09
0.12
5
10
15
20
k


r = 4, p = 0.6
pmf of negative binomial
0.20
0.15
0.10
0.05
0.00
5
10
15
20
k


r = 4, p = 0.85
pmf of negative binomial
0.5
0.4
0.3
0.2
0.1
0.0
5
10
15
20
k

Binomial vs negative binomial

The binomial and the negative binomial distributions share some similarities, but they are, in essence, modeling different things.

- Both are constructed using a sequence of independent Bernoulli rvs with common probability p.
- In a binomial distribution, n, the number of trials or the number of independent Bernoulli rvs is fixed in advance. The rv of interest is then X, the number of successes among these n trials.
- In a negative binomial distribution, n, the number of trials is random and thus not fixed in advance, i.e., for a fixed given $r \geq 1$, we are interested in the rv n that counts the trial number at which the rth success first occurs.

Example In the old days, a male child is considered beneficial to help with physical work and family finances. Suppose a couple will continue to have more children until they have had two male children. Assuming that the probability of a male child is $p = 0.5$, what is the probability that the couple will have at least 6 children ?

Example In the old days, a male child is considered beneficial to help with physical work and family finances. Suppose a couple will continue to have more children until they have had two male children. Assuming that the probability of a male child is $p = 0.5$, what is the probability that the couple will have at least 6 children ?

Let $Y \sim \mathrm{NB}(2, 0.5)$ be the random variable for the number of children. The pmf for Y is

$$P(Y = 0) = P(Y = 1) = 0, \quad P(Y = 2) = \binom{1}{1}0.5^2 = 0.25,$$
$$P(Y = 3) = \binom{2}{1}0.5^3 = 0.25, \quad P(Y = 4) = \binom{3}{1}0.5^4 = 0.1875$$
$$P(Y = 5) = \binom{4}{1}0.5^5 = 0.125, \quad P(Y = 6) = \binom{5}{1}0.5^6 = 0.078$$

In summary, $P(Y \geq 6) \approx 0.19$.

Note There is a close form formula for the tail probability $P(X \geq t)$ when X is a geometric r.v. However there is no simple formula for $X \sim \mathrm{NB}(r, p)$ when $r \geq 2$.

Example How many people do you need to meet before finding 3 other people who have the same birthday as you do ?

Example How many people do you need to meet before finding 3 other people who have the same birthday as you do ?

A. Suppose that the probability that a random person has the same birthday as you do is 1/365, and that the date of birth of the people you meet are somehow independent. We are then interested in the random variable $Y \sim \mathrm{NB}(3, 1/365)$. Thus,

$$P(Y = 3) = \binom{2}{2}(1/365)^3 \approx 2.05 \times 10^{-8},$$
$$P(Y = 10) = \binom{9}{2}(1/365)^3(364/365)^7 \approx 7.26 \times 10^{-7},$$
$$P(Y \geq 100) \approx 0.997,$$
$$P(Y \geq 1000) \approx 0.484,$$
$$P(Y \geq 2000) \approx 0.089.$$

Variants of the geometric series formula

Proposition
Let q be such that $|q| < 1$. Then

$$\sum_{k=0}^{\infty} q^k = \frac{1}{1-q}$$

$$\sum_{k=1}^{\infty} kq^{k-1} = \sum_{k=0}^{\infty} (k+1)q^k = \frac{1}{(1-q)^2}$$

$$\sum_{k=2}^{\infty} k(k-1)q^{k-2} = \sum_{k=0}^{\infty} (k+1)(k+2)q^k = \frac{2}{(1-q)^3}$$

$$\sum_{k=3}^{\infty} k(k-1)(k-2)q^{k-3} = \sum_{k=0}^{\infty} (k+1)(k+2)(k+3)q^k = \frac{6}{(1-q)^4}$$

By taking the ℓ-th derivative on both side of the geometric series formula we have the general result of

Proposition
Let q be such that $|q| < 1$. Then

$$\sum_{k=0}^{\infty} \binom{k+\ell-1}{\ell-1} q^k = \frac{1}{(\ell-1)!} \sum_{k=\ell}^{\infty} k(k-1)\cdots(k-\ell+1) q^{k-\ell}$$
$$= \frac{1}{(1-q)^{\ell}}.$$

Important Example We now verify that the pmf for the negative binomial distribution is valid, i.e., that

$$\sum_{z=r}^{\infty} \binom{z-1}{r-1} p^r (1-p)^{z-r} = 1.$$

Sum of negative binomial random variables

Q. Let $X \sim \mathrm{NB}(r,p)$ and $Y \sim \mathrm{NB}(s,p)$. Suppose X and Y are independent, i.e., $P(X=k,Y=\ell) = P(X=k)P(Y=\ell)$ for all k,ℓ. Let $Z = X+Y$. What is the pmf for Z ?

A. Let z be an arbitrary positive integer. Then

$$\begin{aligned}
P(Z=z) &= \sum_{k=1}^{z-1} P(X=k,Y=z-k) = \sum_{k=1}^{z-1} P(X=k)P(Y=z-k) \\
&= \sum_{k} \binom{k-1}{r-1} p^r (1-p)^{k-r} \binom{z-k-1}{s-1} p^s (1-p)^{z-k-s} \\
&= p^{r+s}(1-p)^{z-r-s} \sum_{k} \binom{k-1}{r-1}\binom{z-k-1}{s-1} \\
&= p^{r+s}(1-p)^{z-r-s} \binom{z-1}{r+s-1}
\end{aligned}$$

which is the pmf of a $\mathrm{NB}(r+s,p)$ random variable.

The above derivations uses the following identity.

Proposition For any positive integers a, b and c we have

$$\sum_{k=1}^{a-1} \binom{k-1}{b-1}\binom{a-k-1}{c-1} = \binom{a-1}{b+c-1}$$

This identity can be proved as follows.

Consider choosing a subset S of $b+c-1$ numbers from the set $\{1, 2, \ldots, a-1\}$. There are $\binom{a-1}{b+c-1}$ possible choices for S.

Now consider the bth largest number in S.

Suppose this number is k for some $k \in \{1, 2, \ldots, a-1\}$.

Then the $b-1$ numbers in S smaller than k are chosen from $\{1, 2, \ldots, k-1\}$ while the $c-1$ numbers in S larger than k are chosen from $\{k+1, k+2, \ldots, a-1\}$. There are thus

$$\binom{k-1}{b-1} \times \binom{a-k-1}{c-1}$$

possible choices for S if the bth largest number is k.

Now k can range from $\{1, 2, \ldots, a-1\}$ and thus we have

$$\sum_{k} \binom{k-1}{b-1} \binom{a-k-1}{c-1}$$

choices for S. The above two counting methods should yield the same answer.

Vandermonde identity

Proposition
For any non-negative integers m, n, r,

$$\sum_{k=0}^{r} \binom{m}{k} \binom{n}{r-k} = \binom{n+m}{r}$$

Trivia Named after the French mathematician Alexandre-Theophile Vandermonde in the 18th century, but supposedly already used in the 14th century in the work of the Chinese mathematician Zhu Shijie.

Proof of Vandermonde identity

Consider selecting r committee members from a set of $n+m$ candidates with n males and m females. This can be done in $\binom{n+m}{r}$ ways.

Equivalently, first let k be the number of females committee members. Then k can take any values in $\{0, 1, 2, \ldots, r\}$.

Given that there are k female committee members, we can select these female committee members from the m female candidates in $\binom{m}{k}$ ways. We then select the remaining $r-k$ male committee members from the remaining n males candidates in $\binom{n}{r-k}$ ways.

These two ways of counting should give the same result.

Hypergeometric distribution

Definition

Let n, M, N be positive integers with $N \geq n$. A random variable X is said to have a hypergeometric distribution with parameters n, M, N if X has pmf

$$P(X = x) = \frac{\binom{M}{x}\binom{N-M}{n-x}}{\binom{N}{n}}, \qquad \text{for } x \geq 0, x \leq M, n - x \leq N - M.$$

$$P(X = x) = \frac{\binom{M}{x}\binom{N-M}{n-x}}{\binom{N}{n}}, \qquad \text{for } x \geq 0, x \leq M, n - x \leq N - M.$$

Interpretation N is the total number of items, of which there are M special items. We sample n items, without replacement, from the N items. The random variable $X \sim \text{Hyper}(n, M, N)$ counts the number of special items in this sample of n items.

Note The above pmf for $X \sim \text{Hyper}(n, M, N)$ is valid is a consequence of the Vandermonde identity.

Note There is no simple closed form formula for $P(X \geq k)$ when $X \sim \text{Hyper}(n, M, N)$.

Example One very popular use of the hypergeometric distribution in statistics is in the analysis of contingency tables for determining whether a sample is representative of the population. Consider the following notional example.

From a pool of six male and five female applicants, seven were selected, and five happened to be men. Is there evidence of gender discrimination ?

	Male	Female	Row Total
Selected	5	2	7
Not Selected	1	3	4
Columns Total	6	5	11

Q. The proportion of males in the applicants pool is $6/11 \approx 54.5\%$, but the proportion of males in the selected group is $5/7 \approx 71.4\%$. Is this difference "large" or "small" ?

We model the number of selected male applicants as $X \sim \mathrm{Hyper}(n, M, N)$ with $M = 6$, $n = 7$, and $N = 11$. We are interested in $P(X \geq 5)$, i.e.,

$$P(X \geq 5) = \frac{\binom{6}{5}\binom{5}{2}}{\binom{11}{7}} + \frac{\binom{6}{6}\binom{5}{1}}{\binom{11}{7}} = \frac{2}{11} + \frac{1}{66} \approx 0.197$$

This example can be easily generalized, i.e., there are M items of a special type among N items; in our sample of n items selected without replacement, there are x items of the special type. Is our selection scheme "biased", e.g., is x/n "very different" from M/N ?

	Special	Not Special	Row Total
Selected	x	$n-x$	n
Not Selected	$M-x$	$N-M-(n-x)$	$N-n$
Columns Total	M	$N-M$	N

Given such a table, and assuming that the sample is really selcted uniformly at random (i.e., no bias), then $X \sim \text{Hyper}(n, M, N)$. Then for an observed value of $X = x$, there are two potential cases

- If $x/n \geq M/N$ then we are interested in $P(X \geq x)$.
- If $x/n \leq M/N$ then we are interested in $P(X \leq x)$.

Binomial vs hypergeometric

The hypergeometric distribution also represents the number of successes in n Bernoulli trials, but these trials are dependent (note that the probability of success of each trial are still identically $p = M/N$). This is due chiefly to the fact that the sampling is done without replacement.

Example A pollster polls $n = 20$ people from population of $N = 10000$ people, where $M = 5500$ people prefers candidate A (and the remaining 4500 people prefers candidate B). Then $X \sim \text{Hyper}(n, M, N)$ is a plausible model, while $X \sim \text{Bin}(n, p)$ with $p = M/N = 0.55$ is another plausible model.

When M and N are large (compared to n), it is generally the case that a binomial r.v. with parameters n and $p = M/N$ and a hypergeometric r.v. with parameters n, M, N are "equivalent".

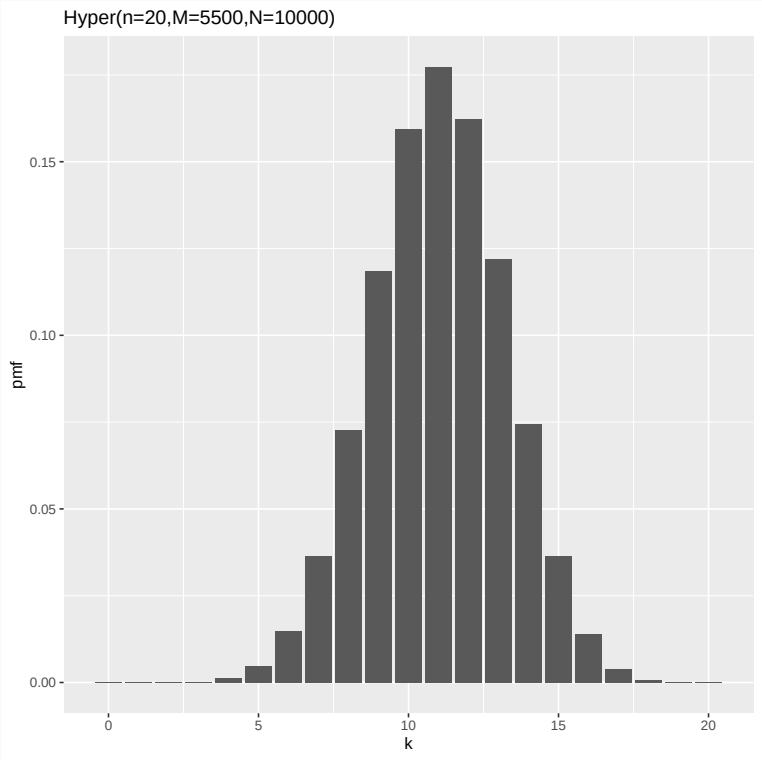
Proposition
Let $X \sim \text{Hyper}(n, M, N)$, where M and N are such that, as $N \to \infty$, $M/N \to p$. Then for any fixed n and x,

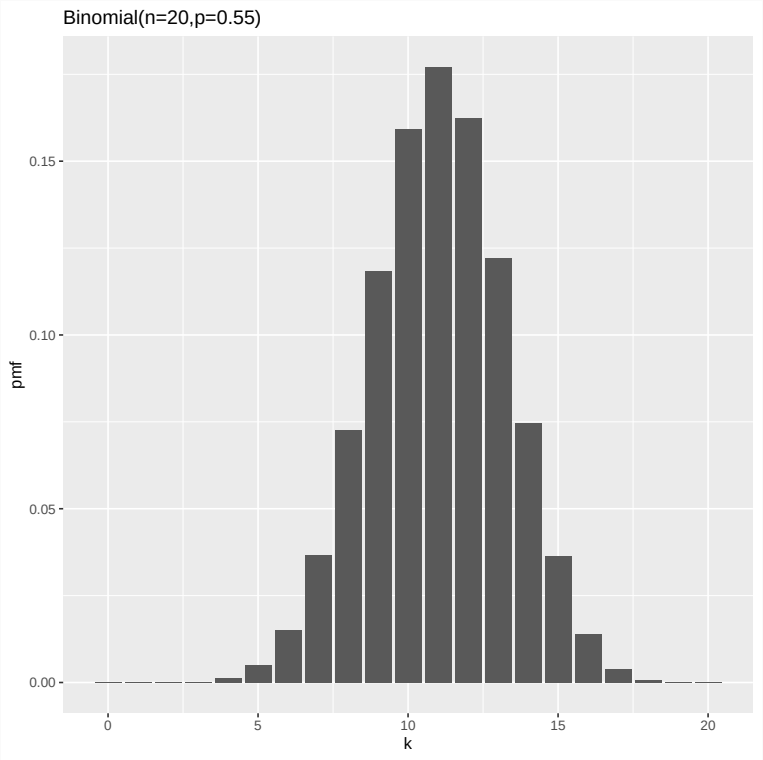
$$P(X = x) = \frac{\binom{M}{x}\binom{N-M}{n-x}}{\binom{N}{n}} \longrightarrow \binom{n}{x} p^x (1-p)^{n-x}$$

as $N \to \infty$.

If M and N are not large compared to n, then the binomial r.v. and the hypergeometric r.v. could be quite different.

Revisiting the pollster example we plot the pmf for $X \sim \text{Hyper}(n = 20, M = 5500, N = 10000)$ versus $X \sim \text{Bin}(n = 20, p = 0.55)$





Mean and variance of hypergeometric r.v.

Proposition
Let $X \sim \mathrm{Hyper}(n, M, N)$. Then

$$\mathbb{E}[X] = \frac{nM}{N}, \qquad \mathrm{Var}[X] = \frac{nM}{N}\Big(1 - \frac{M}{N}\Big)\frac{N-n}{N-1}.$$

Proof Direct calculations yield

$$\begin{aligned}
\mathbb{E}[X] &= \frac{1}{\binom{N}{n}} \sum_x x \binom{M}{x}\binom{N-M}{n-x} = \frac{1}{\binom{N}{n}} \sum_x M \binom{M-1}{x-1}\binom{N-M}{n-x} \\
&= \frac{M}{\binom{N}{n}} \binom{N-1}{n-1} \qquad \text{(Vandermonde identity)} \\
&= \frac{nM}{N}.
\end{aligned}$$

We next evaluate $\mathrm{Var}[X]$. We first evaluate $\mathbb{E}[X(X-1)]$, i.e.,

$$\begin{aligned}
\mathbb{E}[X(X-1)] &= \frac{1}{\binom{N}{n}} \sum_x x(x-1)\binom{M}{x}\binom{N-M}{n-x} \\
&= \frac{1}{\binom{N}{n}} \sum_x M(M-1)\binom{M-2}{x-2}\binom{N-M}{n-x} \\
&= \frac{M(M-1)}{\binom{N}{n}}\binom{N-2}{n-2} \qquad \text{Vandermonde identity} \\
&= \frac{M(M-1)n(n-1)}{N(N-1)}
\end{aligned}$$

Some further algebraic manipulations yield

$$\begin{aligned}
\mathrm{Var}[X] &= \mathbb{E}[X(X-1)] + \mathbb{E}[X] - (\mathbb{E}[X])^2 \\
&= \frac{Mn(N-M)(N-n)}{N^2(N-1)} = \frac{nM}{N}\left(1-\frac{M}{N}\right)\frac{N-n}{N-1}.
\end{aligned}$$

Example: Capture and recapture

Q. Let N be the number of fish (of a certain type) in a river. Assume that N is unknown and need to be estimated (for example, to prevent over fishing). To estimate N, we first capture M fish (of that type). These fish are then tagged with an identification device and released back to the water. We then capture a second sample of n fish. Let X be the number of tagged fish in the second sample. How big is N ?

Suppose we model $X \sim \mathrm{Hyper}(n, M, N)$ with N unknown. Then

$$\mathbb{E}[X] = \frac{nM}{N}$$

Now assume that $X \approx \mathbb{E}[X]$ (this can be justified formally, see e.g. Chebyshev's inequality in the later part of this lecture slides), we arrive at the estimate

$$\hat{N} = \frac{nM}{X}.$$

Poisson distribution

We came to one of the most famous discrete distribution, named after Siméon Poisson

Definition
Let X be a discrete random variable taking values in $\{0, 1, 2, 3, \dots\}$. Then X is said to have a Poisson distribution with mean/rate parameter λ if X has pmf

$$P(X = k) = \frac{e^{-\lambda}\lambda^k}{k!}, \qquad k = 0, 1, 2, 3, \dots$$

The Poisson distribution is routinely used as a model for

- Modeling the number of times a certain event happens temporally. The rate parameter λ then corresponds to the average number of events happening in a specified unit of time, e.g., the number of phone calls received at a call center over a 24 hour period.
- Modeling the number of times a certain event happens spatially. The rate parameter λ is then the average number of events in a specified unit of area or volume, e.g., the number of burnt trees in 1 square mile of a forest.
- approximating the number of successes in a binomial experiment with n trials and success probability p, where n is comparatively large and p is comparatively small.

Example Cook-In is opening a new store on Hillsborough Street. It is estimated that in any given hour, 20 people will, on average, come in to the store. What is the probability that less than 30 people come in to the store during the time period from 11 am to 1 pm ?

Let X be the random variable for the number of people coming in to Cook-In from 11 am to 1 pm. Suppose we choose to model X as a Poisson r.v. The rate is 20 people per hour, hence we assume that $\lambda = 20 \times 2 = 40$. We therefore have

$$P(X = 0) = \frac{e^{-40} \times 40^0}{0!} = e^{-40} \approx 4.24 \times 10^{-18},$$

$$P(X = 1) = \frac{e^{-40} \times 40^1}{1!} = 40 \times e^{-40} \approx 1.7 \times 10^{-16},$$

$$P(X = 2) = \frac{e^{-40} \times 40^2}{2!} = 800 \times e^{-40} \approx 3.4 \times 10^{-15},$$

$$P(X = 10) = \frac{e^{-40} \times 40^{10}}{10!} \approx 1.22 \times 10^{-8},$$

$$P(X \leq 30) \approx 0.0617$$

Example An entomologist is on a field trip in taking samples from long-leaf pine trees. It is estimated that, on average, there are 5 long-leaf pine trees per 100 square feet in a forest. The entomologist decide to cover a 900 square feet area. If the entomologist takes a sample from each tree, what is the probability that he/she will take more than 60 samples ?

Example An entomologist is on a field trip in taking samples from long-leaf pine trees. It is estimated that, on average, there are 5 long-leaf pine trees per 100 square feet in a forest. The entomologist decide to cover a 900 square feet area. If the entomologist takes a sample from each tree, what is the probability that he/she will take more than 60 samples ?

A. We can model the number of samples as $X \sim \text{Pois}(45)$ where $45 = 5 \times 9$ is the rate parameter for a 900 square feet area. We therefore have

$$P(X \geq 60) = \sum_{k=60}^{\infty} \frac{e^{-45} \times 45^k}{k!} \approx 0.019.$$

Example Suppose a chicken lays, on average, λ eggs per month. Each egg has a probability p of actually developing. A chicken is randomly chosen and observed. What is the probability that the chicken laid at least k eggs that hatch into chicks during that month ?

Let X be the random variable for the number of eggs laid by the chicken. Then $X \sim \text{Pois}(\lambda)$. Let Y be the random variable for the number of eggs that develop. Then Y has pmf

$$\begin{aligned}
P(Y = y) &= \sum_{k=y}^{\infty} P(Y = y, X = k) = \sum_{k=y}^{\infty} P(Y = y \mid X = k) P(X = k) \\
&= \sum_{k=y}^{\infty} \frac{e^{-\lambda} \lambda^k}{k!} \times \binom{k}{y} p^y (1-p)^{k-y} \\
&= \frac{e^{-\lambda} \lambda^y p^y (1-p)^{-y}}{y!} \sum_{k=y}^{\infty} \frac{\lambda^{k-y} (1-p)^{k-y}}{(k-y)!} \\
&= \frac{e^{-\lambda} \lambda^y p^y}{y!} \sum_{\ell=0}^{\infty} \frac{\lambda^{\ell} (1-p)^{\ell}}{\ell!} \\
&= \frac{e^{-\lambda} \lambda^y p^y}{y!} \times e^{\lambda(1-p)} = \frac{e^{-\lambda p} (\lambda p)^y}{y!}.
\end{aligned}$$

Hence, Y is a Poisson random variable with rate parameter λp.

Mean and variance of Poisson r.v.

Proposition
Let $X \sim \text{Pois}(\lambda)$. Then

$$\mathbb{E}[X] = \text{Var}[X] = \lambda.$$

Mean and variance of Poisson r.v.

Proposition
Let $X \sim \mathrm{Pois}(\lambda)$. Then

$$\mathbb{E}[X] = \mathrm{Var}[X] = \lambda.$$

Proof Straightforward computations yield

$$\begin{aligned}\mathbb{E}[X] &= \sum_{k=0}^{\infty} k \frac{e^{-\lambda}\lambda^k}{k!} = \sum_{k=1}^{\infty} \frac{e^{-\lambda}\lambda^k}{(k-1)!} = \lambda \sum_{k=1}^{\infty} \frac{e^{-\lambda}\lambda^{k-1}}{(k-1)!} \\ &= \lambda \sum_{\ell=0}^{\infty} \frac{e^{-\lambda}\lambda^{\ell}}{\ell!} = \lambda.\end{aligned}$$

Similarly, we have

$$\mathbb{E}[X(X-1)] = \sum_{k=0}^{\infty} \frac{k(k-1)e^{-\lambda}\lambda^k}{k!} = \sum_{k=2}^{\infty} \frac{e^{-\lambda}\lambda^k}{(k-2)!} = \lambda^2 \sum_{\ell=0}^{\infty} \frac{e^{-\lambda}\lambda^\ell}{\ell!} = \lambda^2$$

and hence

$$\mathrm{Var}[X] = \mathbb{E}[X^2] - \mathbb{E}[X]^2 = \mathbb{E}[X(X-1)] + \mathbb{E}[X] - \mathbb{E}[X]^2 = \lambda.$$

Poisson approximation to binomial r.v.

Proposition
Let $X_n \sim \text{Bin}(n, p_n)$. Suppose $np_n \to \lambda$ for some $\lambda \in (0, \infty)$ as $n \to \infty$. Let $Y \sim \text{Pois}(\lambda)$. Then for any given k,

$$P(X_n = k) \to P(Y = k)$$

as $n \to \infty$.

Poisson approximation to binomial r.v.

Proposition
Let $X_n \sim \text{Bin}(n, p_n)$. Suppose $np_n \to \lambda$ for some $\lambda \in (0, \infty)$ as $n \to \infty$. Let $Y \sim \text{Pois}(\lambda)$. Then for any given k,

$$P(X_n = k) \to P(Y = k)$$

as $n \to \infty$.

Proof We recall the following approximation result

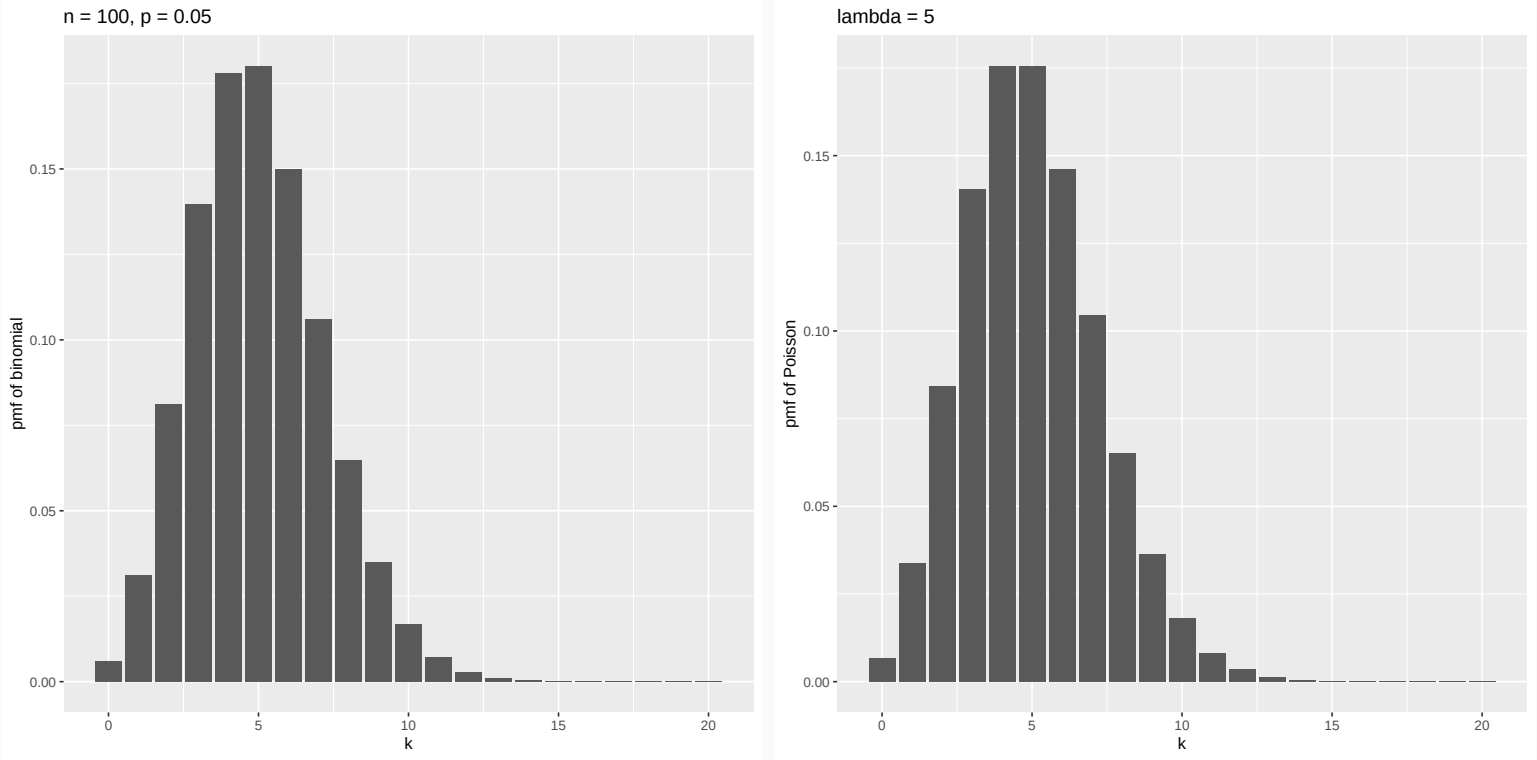
$$\lim_{n\to\infty} \left(1 - \frac{\lambda}{n}\right)^n = e^{-\lambda}$$

We therefore have, since $np_n \to \lambda$,

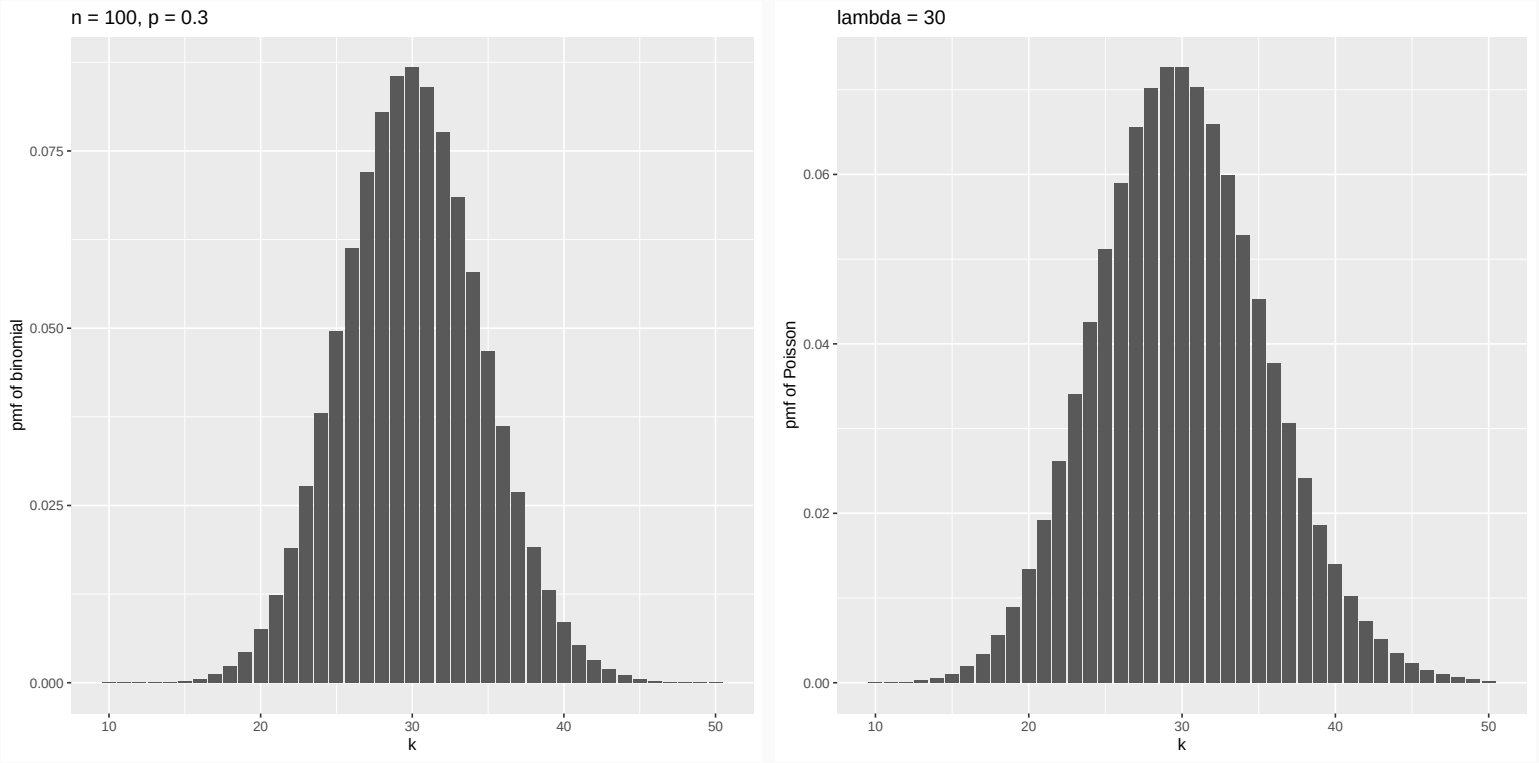
$$\lim_{n\to\infty} (1 - p_n)^n = \lim_{n\to\infty} \left(1 - \frac{np_n}{n}\right)^n = \lim_{n\to\infty} \left(1 - \frac{\lambda}{n}\right)^n = e^{-\lambda}.$$

Therefore, for any fixed k,

$$
\begin{aligned}
P(X_n = k) &= \binom{n}{k} p_n^k (1-p_n)^{n-k} \\
&= \frac{1}{k!} \times (n(n-1)\cdots(n-k+1)) p_n^k (1-p_n)^n \frac{1}{(1-p_n)^k} \\
&= \frac{(np_n)^k}{k!} \left(\tfrac{(n-1)}{n} \tfrac{(n-2)}{n} \cdots \tfrac{(n-k+1)}{n} \right) (1-p_n)^n \left(\frac{1}{1-p_n} \right)^k \\
&\longrightarrow \frac{e^{-\lambda}\lambda^k}{k!}.
\end{aligned}
$$



The pmf of a binomial r.v. with $n = 100$ trials and success probability $p = 0.05$ is well appproximated by that of a Poisson r.v. with rate parameter $\lambda = np = 5$.



The pmf of a binomial r.v. X with $n = 100$ trials and success probability $p = 0.3$ is no longer well appproximated by that of a Poisson r.v. with rate parameter $\lambda = np = 30$ (look at the pmf around $X = 20$ and $X = 40$).

Sums of independent Poisson r.v.

Proposition
Let X and Y be independent Poisson random variables with rate parameter λ and ν, respectively. Then $Z = X + Y$ is also a Poisson r.v. with rate parameter $\lambda + \nu$.

Proof The pmf of Z is given by

$$\begin{aligned}
P(Z=z) &= \sum_{k=0}^{z} P(X=k, Y=z-k) \\
&= \sum_{k=0}^{z} P(X=k)P(Y=z-k) \\
&= \sum_{k=0}^{z} \frac{e^{-\lambda}\lambda^k}{k!}\frac{e^{-\nu}\nu^{z-k}}{(z-k)!} \\
&= e^{-(\lambda+\nu)} \sum_{k=0}^{z} \frac{\lambda^k \nu^{z-k}}{k!(z-k)!} \\
&= \frac{e^{-(\lambda+\nu)}}{z!} \sum_{k=0}^{z} \frac{z!}{k!(z-k)!}\lambda^k \nu^{z-k} = \frac{e^{-(\lambda+\nu)}(\lambda+\nu)^z}{z!}
\end{aligned}$$

which is simply the pmf of a $\mathrm{Pois}(\lambda+\nu)$ random variable.

The previous result implies a very beautiful property of Poisson random variables, namely infinite divisibility.

Proposition
Let $X \sim \mathrm{Pois}(\lambda)$. Then for any integer $n \geq 1$, there exists random variables $X_1, \ldots, X_n$ such that the X_i are identically distributed $X_i \sim \mathrm{Pois}(\lambda/n)$ and $X = X_1 + \cdots + X_n$.

In fact, the Poisson distribution is the unique discrete distribution that is infinitely divisible.