STAT 25100 Lecture 7
Discrete Random Variables (Chapter 4)

Yibi Huang
Department of Statistics
University of Chicago

Random Variables

Random Variables

- So far we have considered probabilities for **events** (subsets) in a space space.
- But, sample spaces can be complex, e.g.,
  - Coin tosses: sequences like `TTHHTTTHTHTTTTH...`
  - Poker: all possible orders of 52 cards
- Usually, we are interested in a specific **numerical** property of the outcome
  - Number of tosses to first heads
  - Number of aces in a poker hand
- Such a numerical property of a random phenomenon is a called *random variable*.

Random Variable

Formally speaking, a *random variable* is a real-valued function on the sample space S, mapping each outcome in S to a real number:

$$\begin{array}{rcl} S & \xrightarrow{X} & \mathbb{R} \\ \text{outcome} & \longmapsto & x = X(\text{outcome}) \end{array}$$

Ex 1. $X =$ number of heads in 3 coin tosses.
Sample space $S = \{\mathtt{HHH}, \mathtt{HHT}, \mathtt{HTH}, \mathtt{HTT}, \mathtt{THH}, \mathtt{THT}, \mathtt{TTH}, \mathtt{TTT}\}$. Then

$$\begin{array}{llll} X(\mathtt{HHH}) = 3, & X(\mathtt{HHT}) = 2, & X(\mathtt{HTH}) = 2, & X(\mathtt{HTT}) = 1, \\ X(\mathtt{THH}) = 2, & X(\mathtt{THT}) = 1, & X(\mathtt{TTH}) = 1, & X(\mathtt{TTT}) = 0 \end{array}$$

Ex 2. $Y =$ nnumber of tosses to first heads.
$S = \{\mathtt{H}, \mathtt{TH}, \mathtt{TTH}, \mathtt{TTTH}, \mathtt{TTTTH}, \ldots\}$ Then

$$Y(\mathtt{H}) = 1,\ Y(\mathtt{TH}) = 2,\ Y(\mathtt{TTH}) = 3,\ Y(\mathtt{TTTH}) = 4, \ldots$$

Discrete and Continuous Random Variable

There are two types of random variables:

- *Discrete random variables* can only take a finite or countable infinite number of possible values
  - Example: Number of heads obtained, number of batteries replaced last year
- *Continuous random variables* take real values (can be decimal)
  - Example: lifetime of a battery, someone's blood pressure

Distribution of a Discrete Random Variable

Coin Example

Let X = number of heads in 4 tosses of a fair coin.

$$\mathrm{P}(X=0) = \mathrm{P}(\{\mathtt{TTTT}\}) = 1/16$$
$$\mathrm{P}(X=1) = \mathrm{P}(\{\mathtt{HTTT}, \mathtt{THTT}, \mathtt{TTHT}, \mathtt{TTTH}\}) = 4/16$$
$$\mathrm{P}(X=2) = \mathrm{P}(\{\mathtt{HHTT}, \mathtt{HTHT}, \mathtt{HTTH}, \mathtt{THHT}, \mathtt{THTH}, \mathtt{TTHH}\}) = 6/16$$
$$\mathrm{P}(X=3) = \mathrm{P}(\{\mathtt{HHHT}, \mathtt{HHTH}, \mathtt{HTHH}, \mathtt{THHH}\}) = 4/16$$
$$\mathrm{P}(X=4) = \mathrm{P}(\{\mathtt{HHHH}\}) = 1/16$$

Each possible value of X and its probability:

Possible Value x of X	0	1	2	3	4
Probability $\mathrm{P}(X=x)$	$\frac{1}{16}$	$\frac{4}{16}$	$\frac{6}{16}$	$\frac{4}{16}$	$\frac{1}{16}$

Note: these probabilities add up to 1:

$$\frac{1}{16} + \frac{4}{16} + \frac{6}{16} + \frac{4}{16} + \frac{1}{16} = 1$$

Probability Mass Function (PMF)

Probability Mass Function (PMF)

The *probability mass function* (PMF) of a random variable X is a function $p(x)$ that assigns each possible value x_i a probability $\mathrm{P}(X = x_i)$.

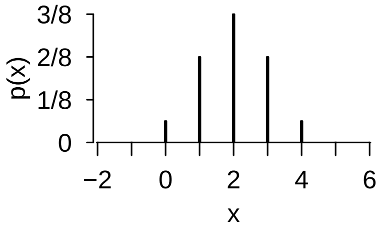
- A PMF $p(x)$ must satisfy $0 \le p(x) \le 1$ and $\sum_x p(x) = 1$.

Example (coin tossing on the previous slide)

Possible Value x of X	0	1	2	3	4
PMF $p(x)$	1/16	4/16	6/16	4/16	1/16

The PMF of X is

$$p(x) = \begin{cases} 1/16 & \text{if } x = 0 \text{ or } 4 \\ 4/16 & \text{if } x = 1 \text{ or } 3 \\ 6/16 & \text{if } x = 2 \\ 0 & \text{if } x \neq 0, 1, 2, 3, 4 \end{cases}$$



Common Discrete Distributions

Bernouli Distribution

A *Bernoulli random variable* X takes only two values, 0 and 1, with probabilities $1 - p$ and p, respectively. Its PMF is thus

$$\begin{aligned} p(1) &= p \\ p(0) &= 1 - p \\ p(x) &= 0, \quad \text{if } x \neq 0 \text{ or } 1 \end{aligned}$$

This is called *Bernoulli distribution* with parameter p, denoted as

$$X \sim \text{Bernoulli}(p).$$

Bernoulli Trials

A random trial with two outcomes: Success (S) or Failure (F) is a *Bernoulli trial*.

- whether a coin lands heads or tails
- whether one gets a six or not a six when rolling a die
- whether a drug works on a patient or not
- whether a electronic device is defected
- whether a subject answers Yes or No to a survey question

If the success probability for a Bernoulli trial is $\mathrm{P}(S) = p$, and let $X = 1$ if the outcome is a Success and 0 if a Failure, then

$$X \sim \text{Bernoulli}(p).$$

Binomial Distributions

Suppose n **independent** Bernoulli trials are performed, each results in

- a *success* with probability p and
- a *failure* with probability $1 - p$.

Define

$$X = \text{number of successes in } n \text{ trials.}$$

X is said to have a *binomial distribution* with parameters (n, p), denoted as

$$X \sim \text{Bin}(n, p).$$

with the probability mass function (PMF)

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

How the Binomial PMF above is obtained? (Next slide)

Consider the case with $n = 5$ trails. Possible outcomes for the event $X = 2$ are the $\binom{5}{2} = 10$ possible orderings of the 2 successes and 3 failures:

Possible Orders
$SSFFF$
$SFSFF$
$SFFSF$
$SFFFS$
$FSSFF$
$FSFSF$
$FSFFS$
$FFSSF$
$FFSFS$
$FFFSS$

As the trials are independent, the probabilities for the outcomes are respectively,

$$\begin{aligned}
\mathrm{P}(SSFFF) &= \mathrm{P}(S)\mathrm{P}(S)\mathrm{P}(F)\mathrm{P}(F)\mathrm{P}(F) \\
&= pp(1-p)(1-p)(1-p) = p^2(1-p)^3, \\
\mathrm{P}(SFSFF) &= \mathrm{P}(S)\mathrm{P}(F)\mathrm{P}(S)\mathrm{P}(F)\mathrm{P}(F) \\
&= p(1-p)p(1-p)(1-p) = p^2(1-p)^3, \\
&\text{etc}
\end{aligned}$$

Observe the 10 outcomes have equal probability $p^2(1-p)^3$ since they all have 2 Successes and 3 Failures.

As the outcomes above are disjoint, $\mathrm{P}(X = 2) = \mathrm{P}(2 \text{ successes in 5 trials})$ is the sum of their probabilities

$$\mathrm{P}(X = 2) = \binom{5}{2} p^2(1-p)^3.$$

In general, for $X \sim \text{Bin}(n, p)$, outcomes in the event $\{X = k\} = \{k \text{ successes in } n \text{ trials}\}$ are the $\binom{n}{k}$ possible orderings of the k successes and $n - k$ failures, each with probability $p^k(1-p)^{n-k}$ to occur. The Binomial PMF is thus

$$p(k) = \mathrm{P}(X = k) = \binom{n}{k} p^k (1-p)^{n-k}, \quad k = 0, 1, \dots, n.$$

Does the Binomial PMF Add Up to 1?

A legitimate PMF $p(k)$ must add up to 1, $\sum_k p(k) = 1$.
Does the Binomial PMF add up to 1?

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

Does the Binomial PMF Add Up to 1?

A legitimate PMF $p(k)$ must add up to 1, $\sum_k p(k) = 1$.
Does the Binomial PMF add up to 1?

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

Yes, using the Binomial Expansion

$$(a+b)^n = \sum_{k=0}^{n} \binom{n}{k} a^k b^{n-k},$$

plugging in $a = p$ and $b = 1 - p$, we get

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

Sum of i.i.d. Bernoulli Random Variables is Binomial

If $X_1, X_2, \ldots, X_n$ are i.i.d. Bernoulli(p), then

$$X_1 + X_2 + \cdots + X_n \sim \text{Bin}(n, p).$$

where "i.i.d." $=$ independent and identically distributed.

Sum of i.i.d. Bernoulli Random Variables is Binomial

If $X_1, X_2, \ldots, X_n$ are i.i.d. Bernoulli(p), then

$$X_1 + X_2 + \cdots + X_n \sim \text{Bin}(n, p).$$

where "i.i.d." $=$ independent and identically distributed.

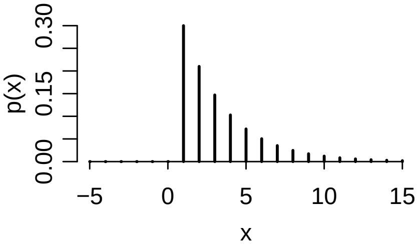
Moreover, if $X \sim \text{Bin}(m,\, p)$ and $Y \sim \text{Bin}(\ell,\, p)$ are independent, then

$$X + Y \sim \text{Bin}(m + \ell,\, p).$$

Geometric Distribution

A sequence of indep. Bernoulli trials are performed, each with success probability p.
Let X = number of trials required to get the first Success.
The PMF of X is

$$
\begin{aligned}
p(k) = \mathrm{P}(X = k) &= \mathrm{P}(\overbrace{F \dots F}^{k-1 \text{ F's}} S) \quad \text{by indep.} \\
&= \mathrm{P}(F) \cdots \mathrm{P}(F)\mathrm{P}(S) \\
&= \underbrace{(1-p) \cdots (1-p)}_{k-1 \text{ copies}} p \\
&= (1-p)^{k-1} p,
\end{aligned}
$$



if x is a positive integer and $p(k) = 0$ if not, denoted as

$$X \sim \text{Geometric}(p).$$

We say X has a *geometric distribution*, since the PMF is a geometric sequence.

Does the Geometric PMF Add Up to 1?

Does $\sum_{k=1}^{\infty} p(k) = \sum_{k=1}^{\infty} (1-p)^{k-1} p = 1$?

Does the Geometric PMF Add Up to 1?

Does $\sum_{k=1}^{\infty} p(k) = \sum_{k=1}^{\infty} (1-p)^{k-1} p = 1$?

Recall the geometric series

$$\begin{aligned} \sum_{k=0}^{\infty} ax^k &= a + ax + ax^2 + \cdots ax^k + \cdots \\ &= \frac{a}{1-x} \quad \text{if } |x| < 1. \end{aligned}$$

Does the Geometric PMF Add Up to 1?

Does $\sum_{k=1}^{\infty} p(k) = \sum_{k=1}^{\infty} (1-p)^{k-1} p = 1$?

Recall the geometric series

$$\begin{aligned}\sum_{k=0}^{\infty} ax^k &= a + ax + ax^2 + \cdots ax^k + \cdots \\ &= \frac{a}{1-x} \quad \text{if } |x| < 1.\end{aligned}$$

The sum of the Geometric PMF

$$\begin{aligned}\sum_{k=1}^{\infty} p(k) &= \sum_{k=1}^{\infty} (1-p)^{k-1} p \\ &= p + (1-p)p + (1-p)^2 p + \cdots + (1-p)^{k-1} p + \cdots\end{aligned}$$

is simply the case $a = p$ and $x = 1 - p$ and hence the sum is

$$\frac{a}{1-x} = \frac{p}{1-(1-p)} = \frac{p}{p} = 1.$$

Negative Binomial Distributions

A sequence of indep. Bernoulli trials are performed, each with success probability p. Let X = number of trials required to get the rth Success. For $\{X = k\}$ to occur,

- the kth trial must be a Success,
- the first $k-1$ trials can be $r-1$ successes and $k-r$ failures in any order.

$$\overbrace{\underbrace{\underline{\ ?\ }\ \underline{\ ?\ }\ \cdots\cdots\ \underline{\ ?\ }}_{\substack{r-1 \text{ S's and } k-r \text{ F's} \\ \text{in any order}}}}^{\text{first } k-1 \text{ trials}} \quad \overbrace{\underbrace{\underline{\ S\ }}_{r\text{th S}}}^{k\text{th trial}}$$

The Negative Binomial PMF is thus

$$\begin{aligned} \mathrm{P}(X = k) &= \overbrace{\binom{k-1}{r-1} p^{r-1}(1-p)^{k-r}}^{\text{first } k-1 \text{ trials}} \times \overbrace{p}^{k\text{th trial}} \\ &= \binom{k-1}{r-1} p^{r}(1-p)^{k-r}, \quad k = r, r+1, \ldots. \end{aligned}$$

denoted as $X \sim NB(r, p)$.

Relation Between Negative Binomial & Geometric

If $X_1, X_2, \dots, X_r$ are independent $\sim$ Geometric(p), then

$$X_1 + X_2 + \cdots + X_r \sim \mathsf{NB}(r, p).$$

Relation Between Negative Binomial & Geometric

If $X_1, X_2, \dots, X_r$ are independent $\sim$ Geometric(p), then

$$X_1 + X_2 + \cdots + X_r \sim \text{NB}(r, p).$$

Conversely, let

- X_1 = number of trials needed to get the first Success
- X_2 = number of additional trials to get the 2nd Success after the first one
- $\vdots$
- X_r = number of additional trials to get the rth Success after the previous one

then $X_1, X_2, \dots, X_r$ are independent Geometric(p).

Negative Binomial Expansion (1)

Recall the geometric series: for $|x| < 1$,

$$\frac{1}{1-x} = 1 + x + x^2 + \cdots x^k + \cdots = \sum_{k=0}^{\infty} x^k.$$

Taking derivative on both sides, we get

$$\frac{1}{(1-x)^2} = 1 + 2x + \cdots kx^{k-1} + \cdots = \sum_{k=1}^{\infty} \binom{k}{1} x^{k-1}.$$

Taking the 2nd derivative on both sides:

$$\frac{2}{(1-x)^3} = \sum_{k=2}^{\infty} \binom{k}{1} (k-1) x^{k-2} = \sum_{k=2}^{\infty} 2 \binom{k}{2} x^{k-2}$$

Dividing both sides by 2:

$$\frac{1}{(1-x)^3} = \sum_{k=2}^{\infty} \binom{k}{2} x^{k-2}.$$

Negative Binomial Expansion (2)

By Mathematical Induction, one can show that the mth derivative of $\frac{1}{1-x} = \sum_{k=0}^{\infty} x^k$ divided by $m!$ is

$$\frac{1}{(1-x)^{m+1}} = \sum_{k=m}^{\infty} \binom{k}{m} x^{k-m},$$

called the *Negative Binomial expansion*.

Apply this with $m = r - 1$ and $x = 1 - p$, it shows Negative Binomial PMF sums to 1.

$$\sum_{k=r}^{\infty} \binom{k-1}{r-1} p^r (1-p)^{k-r} = p^r \sum_{k=r}^{\infty} \binom{k-1}{r-1} (1-p)^{k-r}$$
$$= p^r \cdot \frac{1}{(1-(1-p))^{r-1+1}} = \frac{p^r}{p^r} = 1.$$

Hypergeometric Distribution

$\boxed{R \text{ Red marbles}, \quad B \text{ Blue marbles}}$

$\swarrow$ $\searrow$

d marbles drawn at random without replacement

$\boxed{X \text{ red}, \quad d-X \text{ blue}}$

$R+B-d$ marbles remained in the original box

$\boxed{R-X \text{ red}, \quad B-(d-X) \text{ blue}}$

Suppose $\boxed{d \text{ draws}}$ are made at random w/o replacement from a box containing R red and B blue marbles. The number of Reds X obtained in d draws has a **hypergeometric distribution:**

$$\begin{aligned}
P(X=x) &= P(x \text{ red},\ d-x \text{ blue}) \\
&= \frac{\binom{\#\text{ of ways to pick } x}{\text{out of R Reds}}\binom{\#\text{ of ways to pick } d-x}{\text{out of B Blues}}}{(\#\text{ of ways to pick } d \text{ out of } R+B \text{ marbles})} = \frac{\binom{R}{x}\binom{B}{d-x}}{\binom{R+B}{d}}
\end{aligned}$$

for $0 \le x \le R$, $0 \le d-x \le B$.

Hypergeometric Distribution

$$\boxed{R \text{ Red marbles}, \quad B \text{ Blue marbles}}$$

$$\swarrow \qquad\qquad\qquad\qquad \searrow$$

d marbles drawn at random without replacement	$R+B-d$ marbles remained in the original box
$\boxed{X \text{ red}, \quad d-X \text{ blue}}$	$\boxed{R-X \text{ red}, \quad B-(d-X) \text{ blue}}$

Suppose $\boxed{d \text{ draws}}$ are made at random w/o replacement from a box containing R red and B blue marbles. The number of Reds X obtained in d draws has a **hypergeometric distribution:**

$$\mathrm{P}(X=x) = \mathrm{P}(x \text{ red}, \ d-x \text{ blue})$$
$$= \frac{\left(\substack{\#\text{ of ways to pick } x \\ \text{out of R Reds}}\right)\left(\substack{\#\text{ of ways to pick } d-x \\ \text{out of B Blues}}\right)}{(\#\text{ of ways to pick } d \text{ out of } R+B \text{ marbles})} = \frac{\binom{R}{x}\binom{B}{d-x}}{\binom{R+B}{d}}$$

for $0 \le x \le R$, $0 \le d-x \le B$.

Q: If the draws are made with replacement, what's the distribution of X?

Hypergeometric Distribution

$$\boxed{R \text{ Red marbles}, \quad B \text{ Blue marbles}}$$

d marbles drawn at random without replacement

$$\boxed{X \text{ red}, \quad d-X \text{ blue}}$$

$R+B-d$ marbles remained in the original box

$$\boxed{R-X \text{ red}, \quad B-(d-X) \text{ blue}}$$

Suppose $\boxed{d \text{ draws}}$ are made at random w/o replacement from a box containing R red and B blue marbles. The number of Reds X obtained in d draws has a **hypergeometric distribution:**

$$\begin{aligned}
\mathrm{P}(X=x) &= \mathrm{P}(x \text{ red},\ d-x \text{ blue}) \\
&= \frac{\left(\begin{smallmatrix}\#\text{ of ways to pick } x\\ \text{out of R Reds}\end{smallmatrix}\right)\left(\begin{smallmatrix}\#\text{ of ways to pick } d-x\\ \text{out of B Blues}\end{smallmatrix}\right)}{(\#\text{ of ways to pick } d \text{ out of } R+B \text{ marbles})} = \frac{\binom{R}{x}\binom{B}{d-x}}{\binom{R+B}{d}}
\end{aligned}$$

for $0 \le x \le R$, $0 \le d-x \le B$.

Q: If the draws are made with replacement, what's the distribution of X?

$X \sim \text{Bin}\left(d,\ p=\frac{R}{R+B}\right)$.

The hypergeometric PMF adds up to 1

$$\sum_x \mathrm{P}(X = x) = \frac{\sum_x \binom{R}{x}\binom{B}{d-x}}{\binom{R+B}{d}} = 1$$

because of **Vandermonde's identity** of the Binomial coefficients.

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

Hypergeometric $\approx$ Binomial if ...

If the number of marbles is large (R, $B \to \infty$ and $\frac{R}{R+B} \to p$), then

$$\mathrm{P}(X = x) = \frac{\binom{R}{x}\binom{B}{d-x}}{\binom{R+B}{d}} \longrightarrow \binom{d}{x} p^x (1-p)^{d-x} \quad \text{for } x = 0, 1, \dots, d,$$

i.e., drawing with or without replacement makes little difference.

Proof.

$$\frac{\binom{R}{x}\binom{B}{d-x}}{\binom{R+B}{d}} = \frac{\frac{R!}{x!(R-x)!}\,\frac{B!}{(d-x)!(B-d+x)!}}{\frac{(R+B)!}{d!(R+B-d)!}} = \frac{d!}{x!\,(d-x)!}\,\frac{\frac{R!}{(R-x)!}\,\frac{B!}{(B-d+x)!}}{\frac{(R+B)!}{(R+B-d)!}}$$

$$= \binom{d}{x}\,\underbrace{\frac{R}{R+B}}_{\to p}\,\underbrace{\frac{R-1}{R+B-1}}_{\to p}\cdots\underbrace{\frac{R-x+1}{R+B-x+1}}_{\to p}$$

$$\times\,\underbrace{\frac{B}{R+B-x}}_{\to 1-p}\,\underbrace{\frac{B-1}{R+B-x-1}}_{\to 1-p}\cdots\underbrace{\frac{B-d+x+1}{(R+B-d+1)}}_{\to 1-p}$$

$$\to \binom{d}{x} p^x (1-p)^{d-x}.$$

Poisson Distribution

A random variable X has a Poisson distribution with parameter $\lambda > 0$ if its PMF is

$$\mathrm{P}(X = k) = \frac{\lambda^k}{k!} e^{-\lambda}, \quad k = 0, 1, 2, \ldots$$

denoted as

$$X \sim \text{Poisson}(\lambda).$$

The Poisson PMF sums to 1 because of the Taylor expansion of the exponential function: $e^u = \sum_{k=0}^{\infty} u^k / k!$

$$\sum_{k=0}^{\infty} \mathrm{P}(X = k) = \sum_{k=0}^{\infty} \frac{\lambda^k}{k!} e^{-\lambda} = e^{-\lambda} \underbrace{\sum_{k=0}^{\infty} \frac{\lambda^k}{k!}}_{= e^{\lambda}} = e^{-\lambda} e^{\lambda} = 1.$$

Poisson Approximation to Binomial

If n is huge, p is tiny, but np is moderate,

$$\text{Bin}(n, p) \text{ is approx. Poisson}(\lambda = np).$$

Example:

	$Y \sim \text{Bin}(n = 50,\ p = 0.03)$	$Y \sim \text{Poisson}(\lambda = 50 \times 0.03 = 1.5)$
$\text{P}(Y = 0)$	$(1 - 0.03)^{50} \approx 0.21807$	$e^{-1.5} \approx 0.22313$
$\text{P}(Y = 1)$	$\binom{50}{1}0.03^1(1 - 0.03)^{49} \approx 0.33721$	$1.5e^{-1.5} \approx 0.33470$
$\text{P}(Y = 2)$	$\binom{50}{2}0.03^2(1 - 0.03)^{48} \approx 0.25552$	$\frac{1.5^2}{2!}e^{-1.5} \approx 0.25102$
$\text{P}(Y = 3)$	$\binom{50}{3}0.03^3(1 - 0.03)^{47} \approx 0.12644$	$\frac{1.5^3}{3!}e^{-1.5} \approx 0.12551$
$\text{P}(Y = 4)$	$\binom{50}{4}0.03^4(1 - 0.03)^{46} \approx 0.04595$	$\frac{1.5^4}{4!}e^{-1.5} \approx 0.04707$
$\text{P}(Y = 5)$	$\binom{50}{5}0.03^5(1 - 0.03)^{45} \approx 0.01307$	$\frac{1.5^5}{5!}e^{-1.5} \approx 0.01412$

Proof of Poisson Approximation to Binomial

The Binomial PMF is

$$
\begin{aligned}
\mathrm{P}(X = k) &= \binom{n}{k} p^k (1-p)^{n-k} \\
&= \frac{n!}{k!(n-k)!} \left(\frac{\lambda}{n}\right)^k \left(1 - \frac{\lambda}{n}\right)^{n-k} \qquad \text{(setting } \lambda = np\text{)} \\
&= \frac{\lambda^k}{k!} \underbrace{\frac{n!}{(n-k)!n^k}}_{\to 1} \underbrace{\left(1 - \frac{\lambda}{n}\right)^n}_{\to e^{-\lambda}} \underbrace{\left(1 - \frac{\lambda}{n}\right)^{-k}}_{\to 1} \\
&\to \frac{\lambda^k}{k!} e^{-\lambda} \quad \text{as } n \to \infty \text{ with } np \to \lambda.
\end{aligned}
$$

Example — Fatalities From Horse Kicks (p.45, Textbook)

The number of deaths in a year caused by horse or mule kicks was recorded for each of 10 Prussian cavalry corps over 20 years.
This produced a dataset covering 200 corps-years.

# of Deaths (in a corp in a year)	0	1	2	3	4	Total
Frequency	109	65	22	3	1	200

The count of deaths due to horse kicks in a corp in a given year may have a Poisson distribution because

- $p = \mathrm{P}($a soldier died from horsekicks in a given year$) \approx 0$;
- $n = \#$ of soldiers in a corp, which is large (100's or 1000's);
- whether a soldier was kicked was (at least nearly) independent of whether others were kicked

Example (Fatalities From Horse Kicks — Cont'd)

The fitted Poisson probability to have k deaths from horsekicks for $\lambda = 0.61$ is

$$\mathrm{P}(Y = k) = e^{-\lambda}\frac{\lambda^k}{k!} = e^{-0.61}\frac{(0.61)^k}{k!}, \quad k = 0, 1, 2, \dots .$$

k	Observed Frequency	Relative Frequency	Poisson Probability
0	109	0.545	0.543
1	65	0.325	0.331
2	22	0.110	0.101
3	3	0.015	0.021
4	1	0.005	0.003
Total	200	1	0.999

- $\lambda = 0.61$ is the average of the 200 counts

$$\frac{0 \times 109 + 1 \times 65 + 2 \times 22 + 3 \times 3 + 4 \times 1}{200} = 0.61.$$

When Do Poisson Distributions Come Up?

The number of rare events in a large sample or interval is generally Poisson:

- # of misprints on a page of a book
- # of calls coming into an exchange during a unit of time (if the exchange services a large number of customers who act more or less independently.)
- # of people in a community who survive to age 100
- # of vehicles that pass a marker on a roadway during a unit of time (for light traffic only. In heavy traffic, however, one vehicle's movement may influence another)

Exercise 1 — Identify the Distribution

Roll a fair die 50 times. What is the distribution of the number of $\boxed{\bullet}$ rolled?
Find the probability of no more than 5 $\boxed{\bullet}$.

Exercise 1 — Identify the Distribution

Roll a fair die 50 times. What is the distribution of the number of $\boxed{\bullet}$ rolled?
Find the probability of no more than 5 $\boxed{\bullet}$.

Sol. Let $X =$ number of $\boxed{\bullet}$'s. $X \sim \text{Binomial}(n = 50,\, p = 1/6)$.

$$\mathrm{P}(X \leq 5) = \sum_{k=0}^{5} \mathrm{P}(X = k) = \sum_{k=0}^{5} \binom{50}{k} \left(\frac{1}{6}\right)^{k} \left(\frac{5}{6}\right)^{50-k} \approx 0.139.$$

Exercise 2 — Identify the Distribution

Draw 10 cards from a standard deck without replacement.
Let $X =$ number of Kings obtained.
What's the distribution of X?

Exercise 2 — Identify the Distribution

Draw 10 cards from a standard deck without replacement.
Let $X =$ number of Kings obtained.
What's the distribution of X?

Sol. X has a hypergeometric distribution where

- the Red marbles are the 4 Kings
- the Blue marbles are the remaining 48 cards

The PMF is

$$\mathrm{P}(X = k) = \frac{\binom{4}{k}\binom{48}{10-k}}{\binom{52}{10}}, \quad \text{for } k = 0, 1, 2, 3, 4.$$

Example — Number of Matches in the Matching Problem

N men at a party throw their hats into a basket.
The hats are first mixed up, and each man randomly picks a hat.
Let X = number of men who get their own hat = Number of Matches.

Recall in L04 & L05 we found the PMF of X:

$$p(x) = \mathrm{P}(X = x) = \frac{1}{x!} \sum_{k=0}^{N-x} \frac{(-1)^k}{k!}, \; x = 0, 1, 2, \ldots, N.$$

What's the distribution of X when $N \to \infty$?

$$\lim_{N\to\infty} p(x) = \frac{1}{x!} \sum_{k=0}^{\infty} \frac{(-1)^k}{k!} = \frac{e^{-1}}{x!}, \; x = 0, 1, 2, \ldots .$$

Does the PMF $p(x)$ add up to 1? Let's show using $N = 5$.

$$
\begin{aligned}
p(x) = \sum_{x=0}^{5} \sum_{k=0}^{5-x} \frac{(-1)^k}{x!k!} = \; & 1 + \frac{-1}{0!1!} + \frac{1}{0!2!} + \frac{-1}{0!3!} + \frac{1}{0!4!} + \frac{-1}{0!5!} \\
& \frac{1}{1!0!} + \frac{-1}{1!1!} + \frac{1}{1!2!} + \frac{-1}{1!3!} + \frac{1}{1!4!} + \\
& \frac{1}{2!0!} + \frac{-1}{2!1!} + \frac{1}{2!2!} + \frac{-1}{2!3!} + \\
& \frac{1}{3!0!} + \frac{-1}{3!1!} + \frac{1}{3!2!} + \\
& \frac{1}{4!0!} + \frac{-1}{4!1!} + \\
& \frac{1}{5!0!}
\end{aligned}
$$

Observe the sum along each anti-diagonal is

$$
\frac{1}{n!0!} + \frac{-1}{(n-1)!1!} + \cdots + \frac{(-1)^n}{0!n!} = \sum_{i=0}^{n} \frac{(-1)^i}{(n-i)!i!} = \frac{1}{n!} \underbrace{\sum_{i=0}^{n} \binom{n}{i} (-1)^i 1^{n-i}}_{=(-1+1)^n=0} = 0.
$$

Example 1e (Coupon Collector's Problem, p.132-3, Ross10e)

If each box of breakfast cereals contains a coupon,

- r types of coupons in total
- the coupon in any box is equally likely to be any of the r types,

let $N =$ number of boxes needed to collect all r types of coupons. Find the PMF of N.

Sol: Define $A_j = \{$Type j coupon is NOT collected in the first n boxes.$\}$, $j = 1, \dots, r$.

- $\{N > n\}$ =$\{$at least one type of counpon is not in the first n boxes$\}$ $= \cup_{j=1}^{r} A_j$.
- Are A_j's disjoint?
- $\mathrm{P}(A_j) = \left(\frac{r-1}{r}\right)^n$ (none of the first n boxes include Coupon j)
- $\mathrm{P}(A_{j_1} A_{j_2}) = \left(\frac{r-2}{r}\right)^n$
- $\mathrm{P}(A_{j_1} A_{j_2} \dots A_{j_k}) = \left(\frac{r-k}{r}\right)^n$,

Example 1e (Coupon Collector's Problem, p.132-3, Ross10e) (2)

By the Inclusion-Exclusion Formula,

$$
\begin{aligned}
\mathrm{P}(N > n) &= \mathrm{P}(A_1 \cup A_2 \cup \ldots \cup A_r) \\
&= \sum_{j=1}^{r} \mathrm{P}(A_j) - \sum_{1 \le j_1 < j_2 \le n} \mathrm{P}(A_{j_1} A_{j_2}) + \sum_{1 \le j_1 < j_2 < j_3 \le n} \mathrm{P}(A_{j_1} A_{j_2} A_{j_3}) \\
&\quad + \cdots + (-1)^{r+1} \sum_{1 \le j_1 < j_2 < \ldots < j_r \le n} \mathrm{P}(A_{j_1} A_{j_2} \ldots A_{j_r}) \\
&= r \left(\frac{r-1}{r}\right)^n - \binom{r}{2} \left(\frac{r-2}{r}\right)^n + \binom{r}{3} \left(\frac{r-3}{r}\right)^n - \cdots \\
&= \sum_{i=1}^{r-1} (-1)^{i+1} \binom{r}{i} \left(\frac{r-i}{r}\right)^n
\end{aligned}
$$

The PMF $p(n)$ can be obtained by

$$
p(n) = \mathrm{P}(N = n) = \mathrm{P}(N > n-1) - \mathrm{P}(N > n).
$$