6. random variables

let X = index of

May be interested, not in outcome per se, but in some function of outcome usually numeric. This is called a random variable.

Ex:

Let H be the number of heads when 2 coins tossed

Let T be the total of 2 dice rolls

Let N be the number of coin tosses needed to observe 1st head

Note: even if underlying experiment has "equally likely outcomes", the associated random variable(s) may not

outcome	H	P(H)
TT	0	$P(H=0) = 1/4$
TH	1	$P(H=1) = 1/2$
HT	1	
HH	2	$P(H=2) = 1/4$

numbered balls

Ross 4.1 ex 1b

20 balls numbered 1, 2, .. 20

draw 3 without replacement

let X = max number drawn

what is $P(X \geq 17)$?

$$P(X=20) = \binom{19}{2} \Big/ \binom{20}{3} = \frac{3}{20} = .150$$

$$P(X=19) = \binom{18}{2} \Big/ \binom{20}{3} = \frac{18 \cdot 17/2!}{20 \cdot 19 \cdot 18/3!} = \frac{51}{380} \approx 0.134$$

$$\vdots$$

$$\Sigma = .508$$

Alt:

$$P(X \geq 17) = 1 - P(X < 17) = 1 - \binom{16}{3} \Big/ \binom{20}{3} \approx .508$$

flip a (biased) coin repeatedly until 1st head observed

how many flips ? Let X be that number

$$P(X=1) = P(H) = p$$

$$P(X=2) = P(TH) = (1-p)\,p$$

$$P(X=3) = P(TTH) = (1-p)^2\,p$$

$$\vdots$$

$$P\left(\bigcup_{i=1}^{\infty}\{X=i\}\right) = \sum_{i=0}^{\infty}(1-p)^i p = p\sum_{i=0}^{\infty} q^i = p\cdot\frac{1}{1-q} = p/p = 1$$

$q = 1-p$

If a random variable X takes on only a countable number of possible values, X is said to be discrete

Ex:

X_1 = sum of 3 dice : $3 \leq X \leq 18$, $X \in \mathbb{N}$

X_2 = number of 1st head in seq of coin flips $\quad 1 \leq X_2$, $X_2 \in \mathbb{N}$

X_3 = largest prime factor of $(1+X_3)$ $\quad X_3 \in \{2, 3, 5, 7, 11, \ldots\}$

If X is a discrete random variable taking on values from a countable set $T \subseteq \mathbb{R}$ then

$$p(a) = \begin{cases} P(X=a) & \text{for } a \in T \\ 0 & \text{otherwise} \end{cases}$$

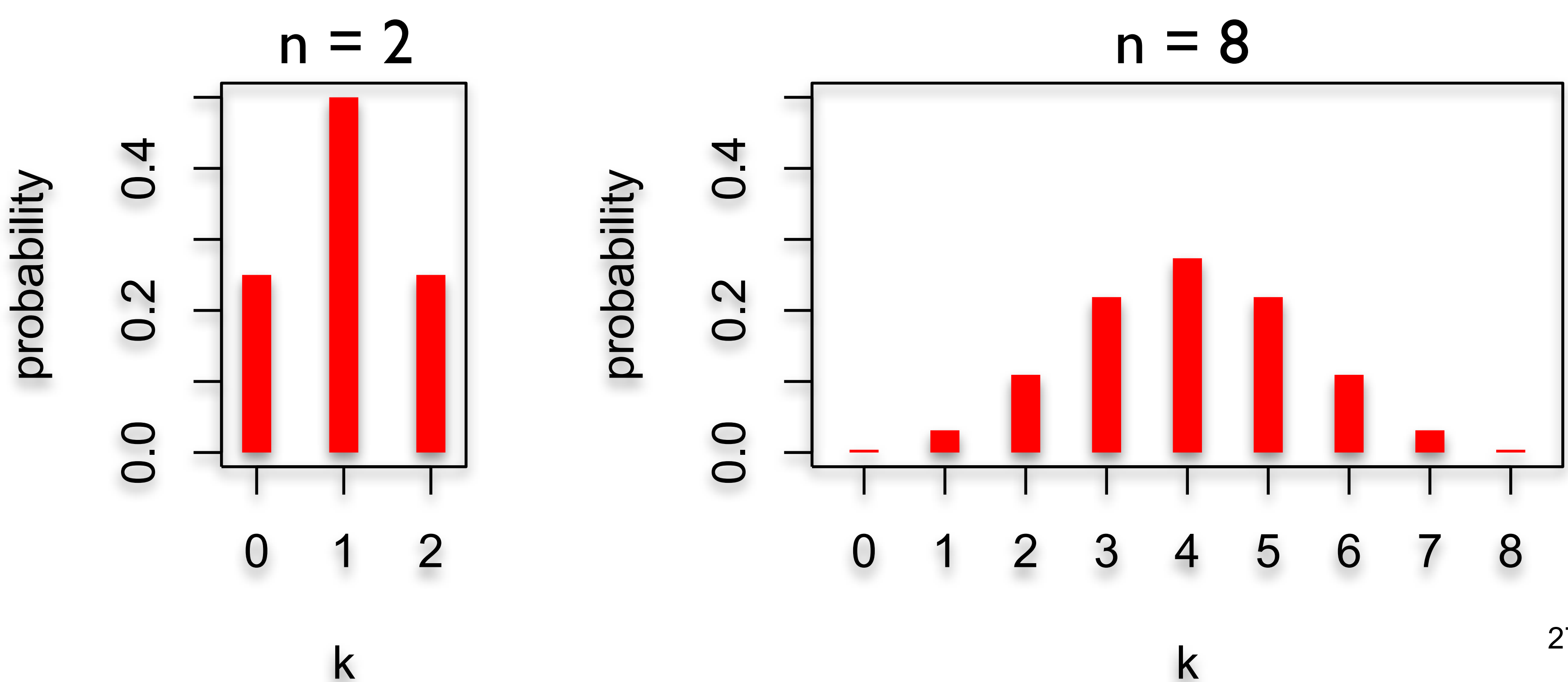
is called the probability mass function for X

Note $\quad \displaystyle\sum_{a \in T} p(a) = 1$

Let X be the number of heads observed in n coin flips

$$P(X=k) = \binom{n}{k} p^k (1-p)^{n-k} \quad \text{where } p = P(H)$$

Probability mass function:



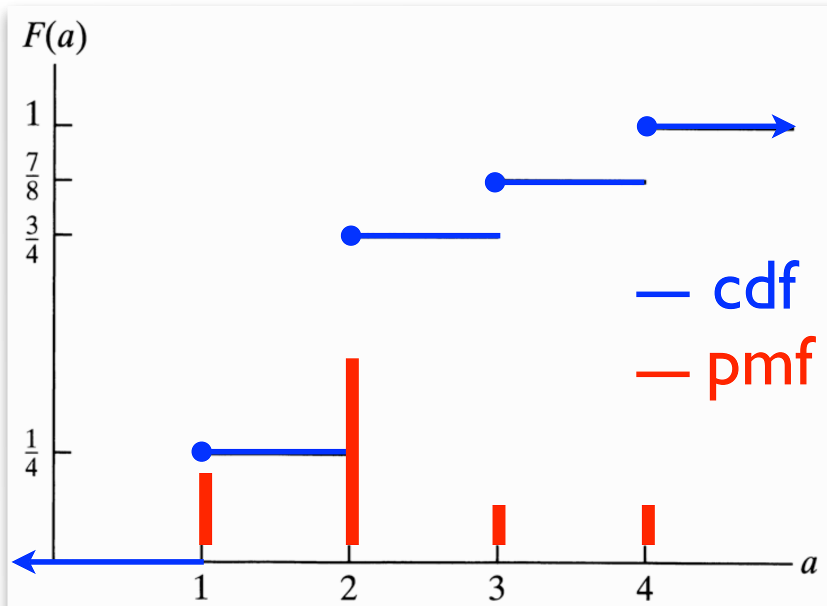
The *cumulative distribution function* for a random variable X is the function *F*: $\mathbb{R} \to [0,1]$ defined by

$$\boxed{F(a) = P[X \leq a]}$$

Ex: if X has probability mass function given by:

$$p(1) = \frac{1}{4} \qquad p(2) = \frac{1}{2} \qquad p(3) = \frac{1}{8} \qquad p(4) = \frac{1}{8}$$

$$F(a) = \begin{cases} 0 & a < 1 \\ \frac{1}{4} & 1 \leq a < 2 \\ \frac{3}{4} & 2 \leq a < 3 \\ \frac{7}{8} & 3 \leq a < 4 \\ 1 & 4 \leq a \end{cases}$$



NB: for discrete random variables, be careful about "≤" vs "<"

expectation

discrete r.v. X with p.m.f $p(\,)$

The expectation of X, aka expected value or mean, is

$$E[X] = \sum x\,p(x)$$

average of random values, weighted by their respective probabilities

Ex.

Let X = value seen rolling fair die 4.3ex3a

$$p(1) = p(2) = \cdots = p(6) = 1/6$$

$$E[X] = \sum_{i=1}^{6} i\,p(i) = \tfrac{1}{6}(1+2+\cdots+6) = \tfrac{21}{6} = \tfrac{7}{2} = 3.5$$

Ex.

Suppose you flip a coin; heads – win \$1, tails – lose \$1

X = +1 if heads, −1 if tails

$$E(X) = (+1)\cdot p(+1) + (-1)\cdot p(-1) = +1\left(\tfrac{1}{2}\right) + (-1)\left(\tfrac{1}{2}\right) = 0$$

"a fair game": in repeated play you expect to win as much as you lose. Long term net gain/loss = 0.

flip a (biased) coin repeatedly until 1st head observed

how many flips? Let X be that number

$P(H) = p$, $P(T) = 1-p = q$

$P(i) = q^{i-1}p$

$$E(X) = \sum_{i=1}^{\infty} i\,p(i) = \sum_{i=1}^{\infty} i q^{i-1} p = p \sum_{i \geq 1} i q^{i-1} \qquad (*)$$

A calculus trick:

$$\sum_{i \geq 1} i \cancel{q}\, y^{i-1} = \frac{d}{dy} \sum_{i \geq 1} y^i = \frac{d}{dy} \sum_{i \geq 0} y^i = \frac{d}{dy} \frac{1}{1-y} = \frac{1}{(1-y)^2}$$

$dy^0/dy = 0$

So $(*)$ becomes:

$$p \sum_{i \geq 0} i q^{i-1} = \frac{p}{(1-q)^2} = \frac{p}{p^2} = \frac{1}{p}$$

Eg:

$p = 1/2$ on average, head every 2nd flip

$p = \frac{1}{10}$ on average, head every 10th flip.

expectation of a *function* of a random variable

Calculating E[g(X)]:

Y=g(X) is a new r.v. Calc P[Y=j], then apply defn:

X = sum of 2 dice rolls

i	p(i) = P[X=i]	i•p(i)
2	1/36	2/36
3	2/36	6/36
4	3/36	12/36
5	4/36	20/36
6	5/36	30/36
7	6/36	42/36
8	5/36	40/36
9	4/36	36/36
10	3/36	30/36
11	2/36	22/36
12	1/36	12/36

$E[X] = \Sigma_i \, ip(i) = \boxed{252/36} = 7$

Y = g(X) = X mod 5

j	q(j) = P[Y = j]	j•q(j)
0	4/36+3/36 =7/36	0/36
1	5/36+2/36 =7/36	7/36
2	1/36+6/36+1/36 =8/36	16/36
3	2/36+5/36 =7/36	21/36
4	3/36+4/36 =7/36	28/36

$E[Y] = \Sigma_j \, jq(j) = \boxed{72/36} = 2$

expectation of a *function* of a random variable

Calculating E[g(X)]: Another way – add in a different order, using P[X=...] instead of calculating P[Y=...]

X = sum of 2 dice rolls

i	p(i) = P[X=i]	g(i)•p(i)
2	1/36	2/36
3	2/36	6/36
4	3/36	12/36
5	4/36	0/36
6	5/36	5/36
7	6/36	12/36
8	5/36	15/36
9	4/36	16/36
10	3/36	0/36
11	2/36	2/36
12	1/36	2/36

$E[g(X)] = \Sigma_i\, g(i)p(i) =$ $\boxed{72/36}$ $= 2$

Y = g(X) = X mod 5

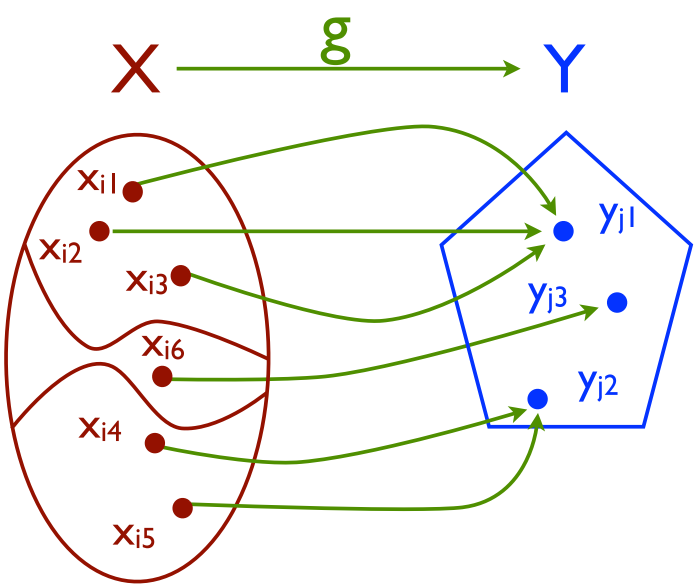
j	q(j) = P[Y = j]	j•q(j)
0	4/36+3/36 =7/36	0/36
1	5/36+2/36 =7/36	7/36
2	1/36+6/36+1/36 =8/36	16/36
3	2/36+5/36 =7/36	21/36
4	3/36+4/36 =7/36	28/36

$E[Y] = \Sigma_j\, jq(j) =$ $\boxed{72/36}$ $= 2$

Above example is not a fluke.

Theorem: if $Y = g(X)$, then $E[Y] = \Sigma_i\, g(x_i)p(x_i)$, where x_i, $i = 1, 2, \ldots$ are all possible values of X.

Proof: Let y_j, $j = 1, 2, \ldots$ be all possible values of Y.



Note that $S_j = \{\, x_i \mid g(x_i)=y_j \,\}$ is a partition of the domain of g.

$$
\begin{aligned}
\sum_i g(x_i)p(x_i) &= \sum_j \sum_{i:g(x_i)=y_j} g(x_i)p(x_i) \\
&= \sum_j \sum_{i:g(x_i)=y_j} y_j p(x_i) \\
&= \sum_j y_j \sum_{i:g(x_i)=y_j} p(x_i) \\
&= \sum_j y_j P\{g(X) = y_j\} \\
&= E[g(X)]
\end{aligned}
$$

A & B each bet \$1; flip 2 coins:

HH : A wins \$2

TH, HT : Each takes back \$1

TT : B wins \$2

Let X be A's net gain: +1, 0, -1

$P(X=1) = 1/4$

$P(X=0) = 1/2$

$P(X=-1) = 1/4$

What is $E[X]$?

$E[X] = \frac{1}{4}\cdot 1 + \frac{1}{2}\cdot 0 + \frac{1}{4}\cdot(-1) = 0$

(in repeated play, average gain = 0)

What is $E[X^2]$?

$E[X^2] = \frac{1}{4}\cdot 1^2 + \frac{1}{2}\cdot 0^2 + \frac{1}{4}\cdot(-1)^2 = \frac{1}{2}$

$\boxed{\text{Note } E(X^2) \neq (E(X))^2}$

Linearity, I

For any constants a, b $\quad E[aX+b] = aE[X]+b$

Proof:

$$E[aX+b] = \sum_x (ax+b)\cdot p(x)$$

$$= a\sum_x x\,p(x) + b\sum_x p(x)$$

$$= a\,E[X] + b$$

Example

In the two-coin game above, what is $E[2X+1]$?

A: $E[2X+1] = 2E[X]+1 = 2\cdot 0+1 = 1$

Ross 4.9

Linearity, II

Let X and Y be two random variables derived from outcomes of a single experiment. Then

$$E[X+Y] = E[X] + E[Y]$$

True even if X,Y dependent

Proof:

Assume the sample space S is countable. (The result is true without this assumption, but I won't prove it.) Let $X(s), Y(s)$ be the values of these r.v.'s for outcome $s \in S$.

Claim: $E[X] = \sum_{s \in S} X(s) \cdot p(s)$

proof is similar to that for "expectation of a function of an r.v." i.e. the events "$X = x$" partition S, so sum above can be rearranged to match the definition $E[X] = \sum_x x \cdot P[X=x]$

Then, let r.v. $Z = X + Y$

$$E[X+Y] = E[Z] = \Sigma_{s\in S} Z[s]\ p(s) = \Sigma_{s\in S}(X[s] + Y[s])\ p(s)$$
$$= \Sigma_{s\in S} X[s]\ p(s) + \Sigma_{s\in S} Y[s]\ p(s) = E[X] + E[Y]$$

Example

X = # heads in one coin flip where $P[X=1]=p$.

What is $E[X]$?

$$E[X] = 1\cdot p + 0(1-p) = p$$

X_i, $1 \le i \le n$ = # heads in flip of a coin with $P[X_i=1]=p_i$

what is the expected number of heads when all are flipped?

$$E\Big[\sum_i X_i\Big] = \sum_i E[X_i] = \sum_i p_i$$

Special case $p_1 = p_2 \cdots = p$

$$E[\text{number of heads}] = np$$

Note

Linearity is special

It is not true in general that

$$E[X \cdot Y] = E[X] \cdot E[Y]$$

$$E[X^2] = E[X]^2$$ ← counterexample above

$$E[X/Y] = E[X]/E[Y]$$

$$E[\operatorname{asinh}(X)] = \operatorname{asinh}(E(X))$$

$$\vdots$$

Alice & Bob are gambling, again. X = Alice's gain per flip

$$X = \begin{cases} +1 & \text{if heads} \\ -1 & \text{if tails} \end{cases}$$

$$E[X] = 0$$

.
.
.
Time
passes
.
.
.

Alice says "Let's raise the stakes"

$$Y = \begin{cases} +1000 & \text{if heads} \\ -1000 & \text{if tails} \end{cases}$$

$E[Y] = 0$ still

Are you [Bob] equally happy to play?

$E[X]$ measures the "average" or "central tendency"

what about its variability

If $E[X] = \mu$

$E[\,|X-\mu|\,]$ is a natural quantity to look at, but mathematically inconvenient

Definition

The variance of a random variable X with mean μ is $E[(X-\mu)^2] = Var[X]$

The standard deviation of X is $\sqrt{Var(X)}$

Alice & Bob are gambling, again. X = Alice's gain per flip

$$X = \begin{cases} +1 & \text{if heads} \\ -1 & \text{if tails} \end{cases}$$

$E[X] = 0$ $\qquad$ $\mathrm{Var}[X] = 1$

⋮

Time passes

⋮

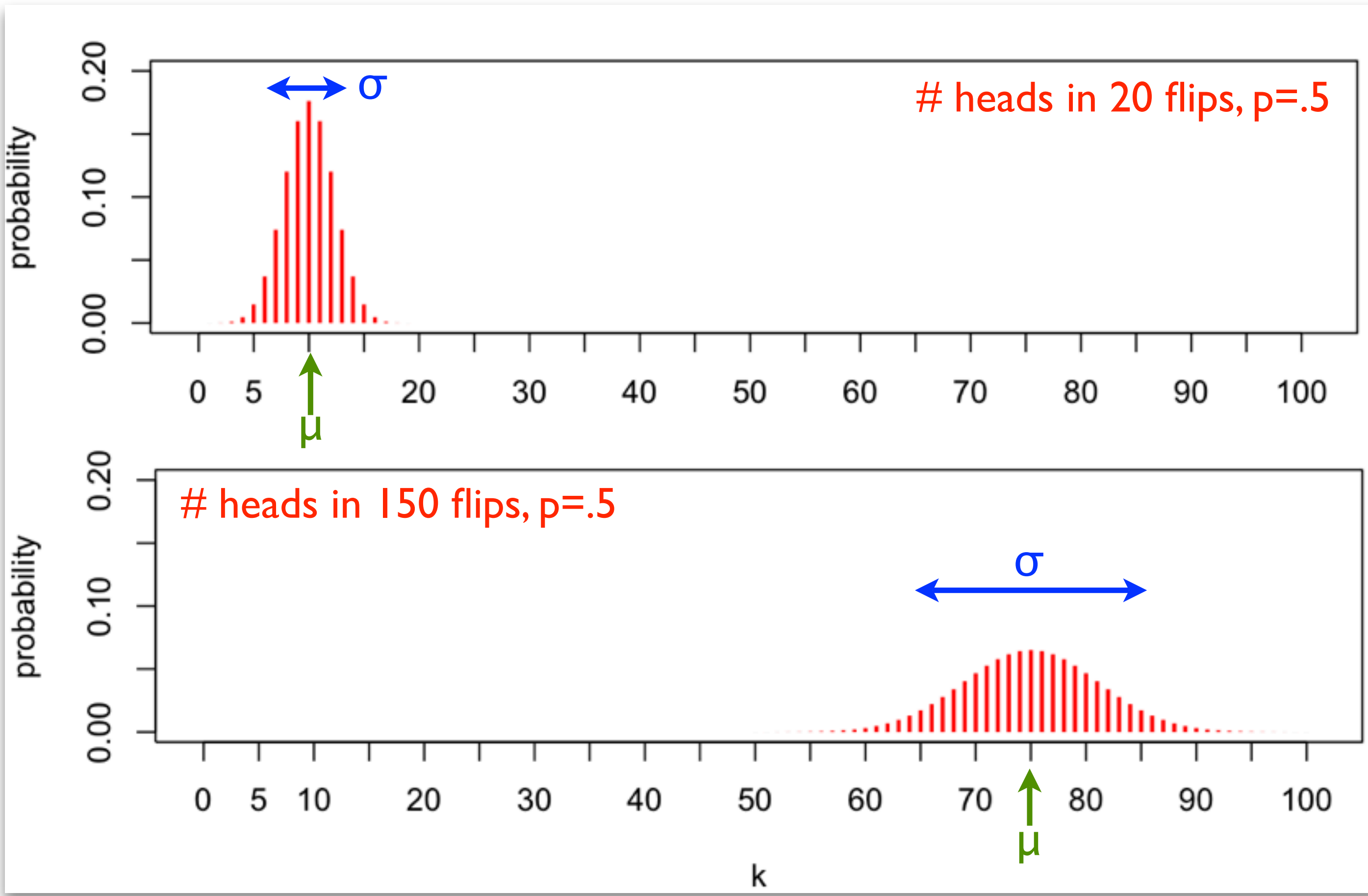
Alice says "Let's raise the stakes"

$$Y = \begin{cases} +1000 & \text{if heads} \\ -1000 & \text{if tails} \end{cases}$$

$E[Y] = 0$ still $\qquad$ $\mathrm{Var}[X] = 1{,}000{,}000$

Are you [Bob] equally happy to play?

μ = E[X] is about *location*; $\sigma = \sqrt{Var(X)}$ is about *spread*



$$\boxed{\mathrm{Var}(X) = E[X^2] - (E[X])^2}$$

$$\begin{aligned}
\mathrm{Var}(X) &= E[(X - \mu)^2] \\
&= \sum_x (x - \mu)^2 p(x) \\
&= \sum_x (x^2 - 2\mu x + \mu^2) p(x) \\
&= \sum_x x^2 p(x) - 2\mu \sum_x x p(x) + \mu^2 \sum_x p(x) \\
&= E[X^2] - 2\mu^2 + \mu^2 \\
&= E[X^2] - \mu^2
\end{aligned}$$

Example:

What is Var(X) when X is outcome of one fair die?

$$E[X^2] = 1^2\left(\frac{1}{6}\right) + 2^2\left(\frac{1}{6}\right) + 3^2\left(\frac{1}{6}\right) + 4^2\left(\frac{1}{6}\right) + 5^2\left(\frac{1}{6}\right) + 6^2\left(\frac{1}{6}\right)$$
$$= \left(\frac{1}{6}\right)(91)$$

E(X) = 7/2, so

$$\mathrm{Var}(X) = \frac{91}{6} - \left(\frac{7}{2}\right)^2 = \frac{35}{12}$$

$$\boxed{\mathrm{Var}[aX+b] = a^2\,\mathrm{Var}[X]}$$

$$\begin{aligned}\mathrm{Var}(aX + b) &= E[(aX + b - a\mu - b)^2] \\ &= E[a^2(X - \mu)^2] \\ &= a^2E[(X - \mu)^2] \\ &= a^2\mathrm{Var}(X)\end{aligned}$$

Ex:

$$X = \begin{cases} +1 & \text{if heads} \\ -1 & \text{if tails}\end{cases}$$

$E[X] = 0$

$\mathrm{Var}[X] = 1$

$$Y = \begin{cases} +1000 & \text{if heads} \\ -1000 & \text{if tails}\end{cases}$$

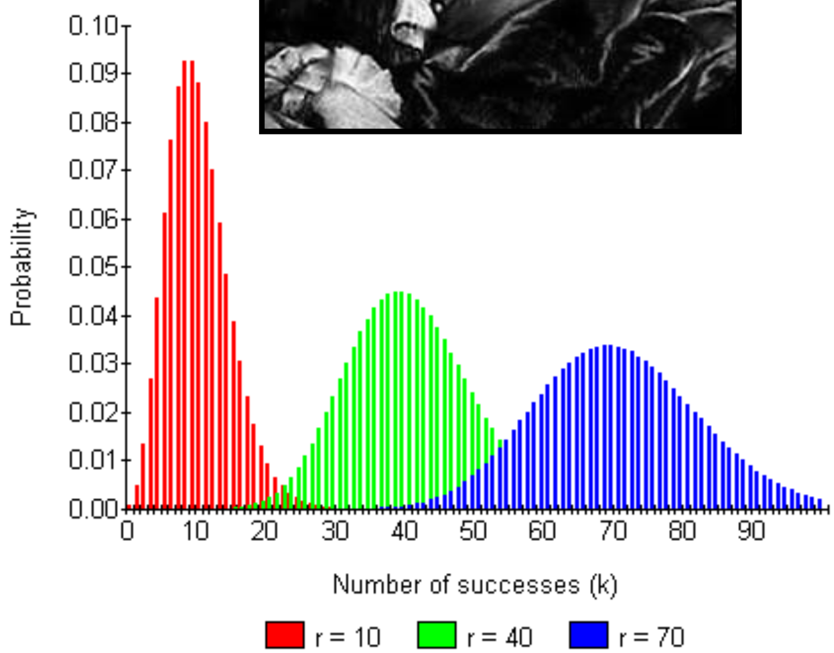
$Y = 1000\ X$

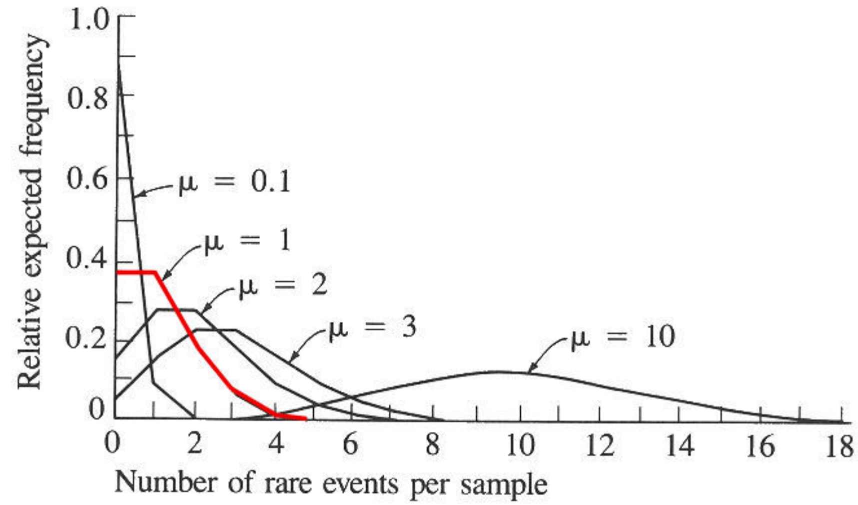
$E[Y] = E[1000\ X] = 1000\ E[x] = 0$

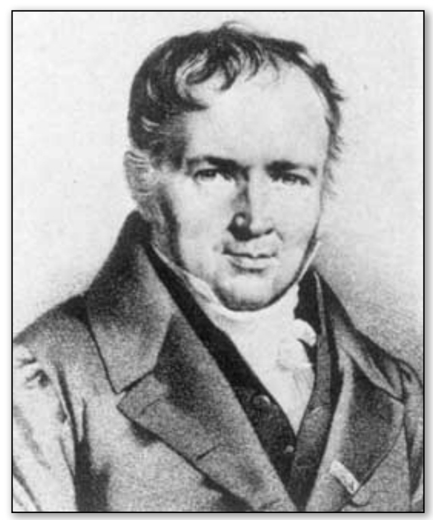
$\mathrm{Var}[Y] = \mathrm{Var}[1000\ X]$

$= 10^6\mathrm{Var}[X] = 10^6$

a zoo of (discrete) random variables







An experiment results in "Success" or "Failure"

X is a random *indicator variable* (1=success, 0=failure)

$P(X=1) = p$ and $P(X=0) = 1-p$

X is called a *Bernoulli* random variable: $X \sim Ber(p)$

$E[X] = p$

$Var(X) = E[X^2] - (E[X])^2 = p - p^2 = p(1-p)$

Examples:

coin flip

random binary digit

whether a disk drive crashed

Jacob (aka James, Jacques) Bernoulli, 1654 – 1705

Consider n independent random variables $Y_i \sim \mathrm{Ber}(p)$

$X = \Sigma_i Y_i$ is the number of successes in n trials

X is a *Binomial* random variable: $X \sim \mathrm{Bin}(n,p)$

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

By Binomial theorem, $\sum_{i=0}^{n} P(X=i) = 1$

Examples

\# of heads in n coin flips

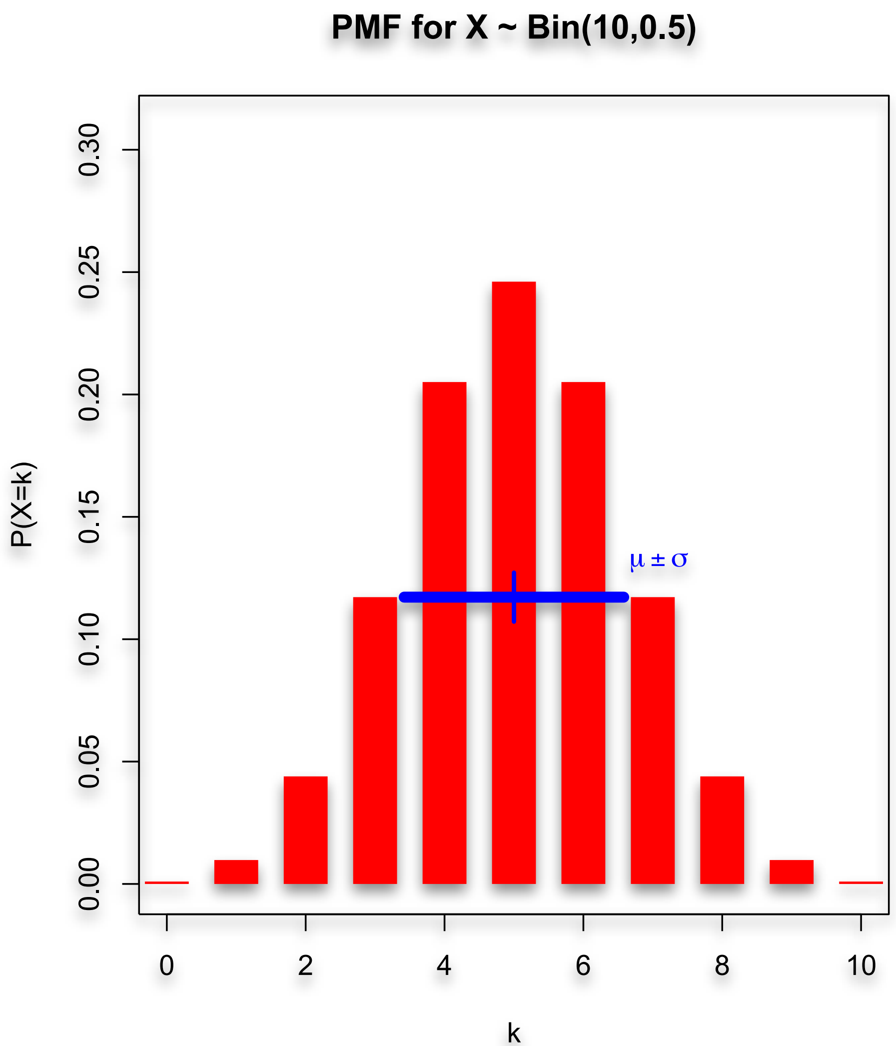
\# of 1's in a randomly generated length n bit string

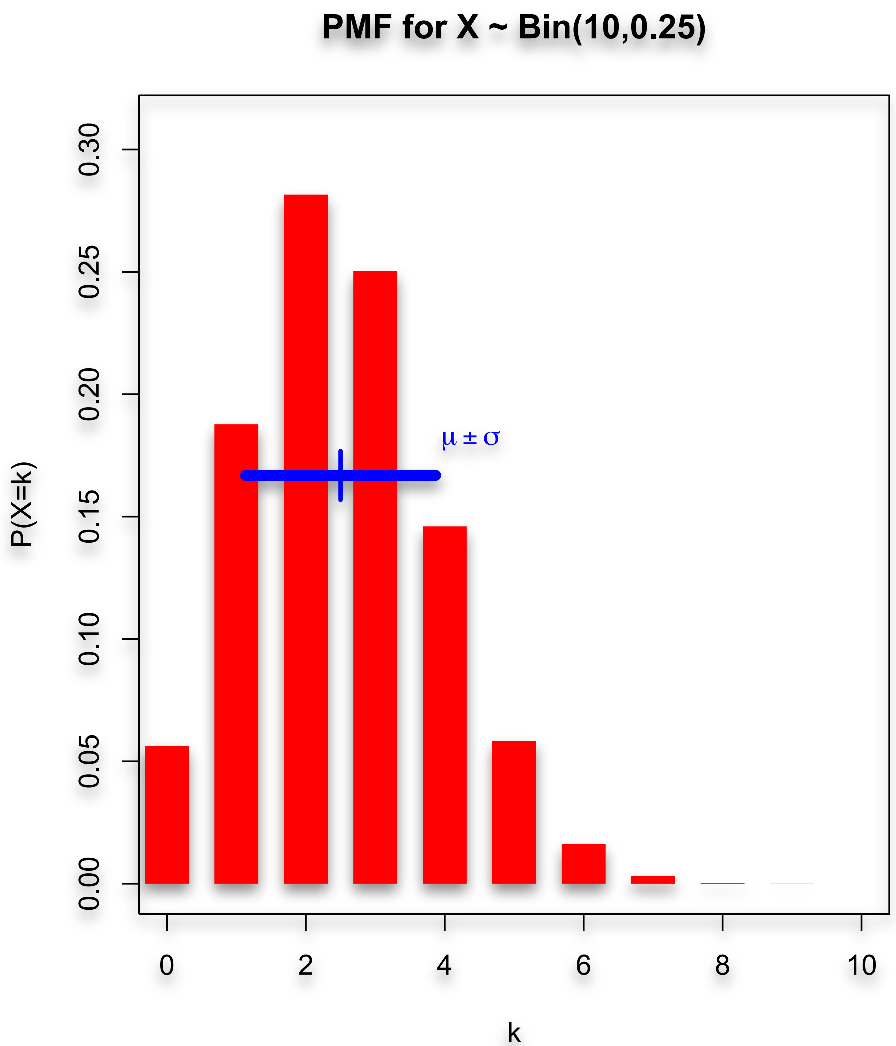
\# of disk drive crashes in a 1000 computer cluster

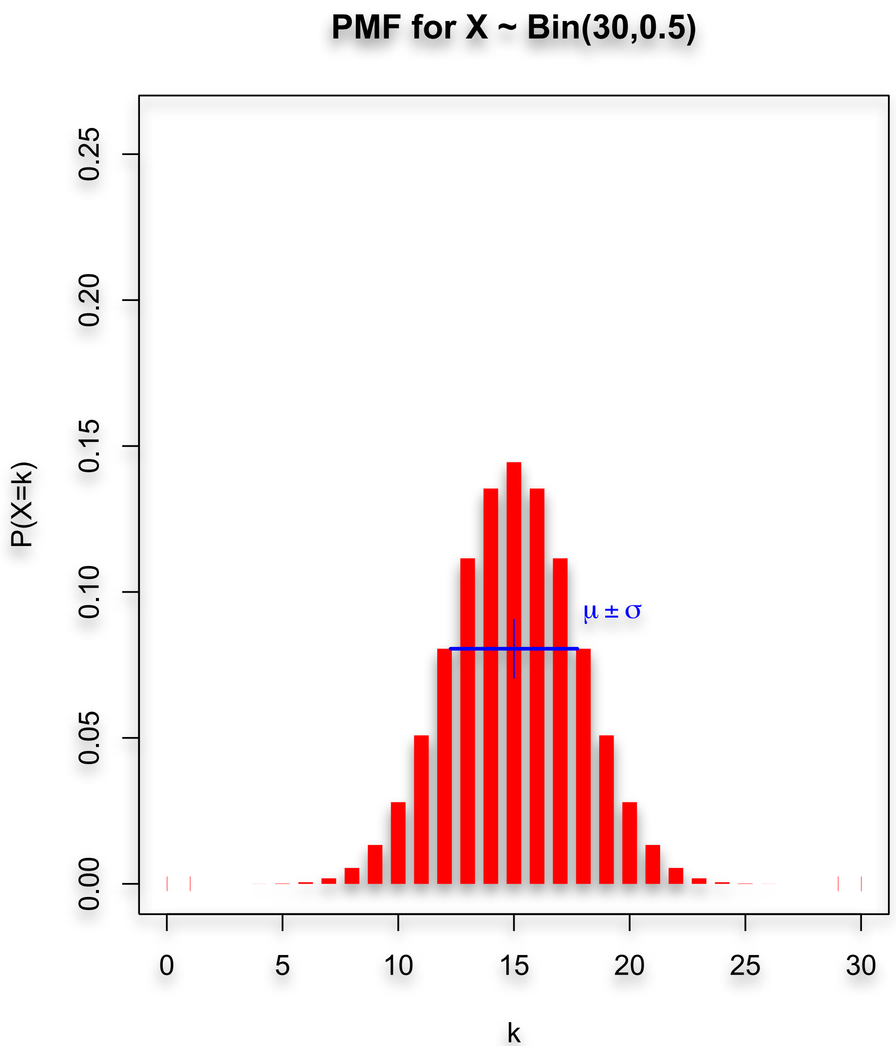
$E[X] = pn$

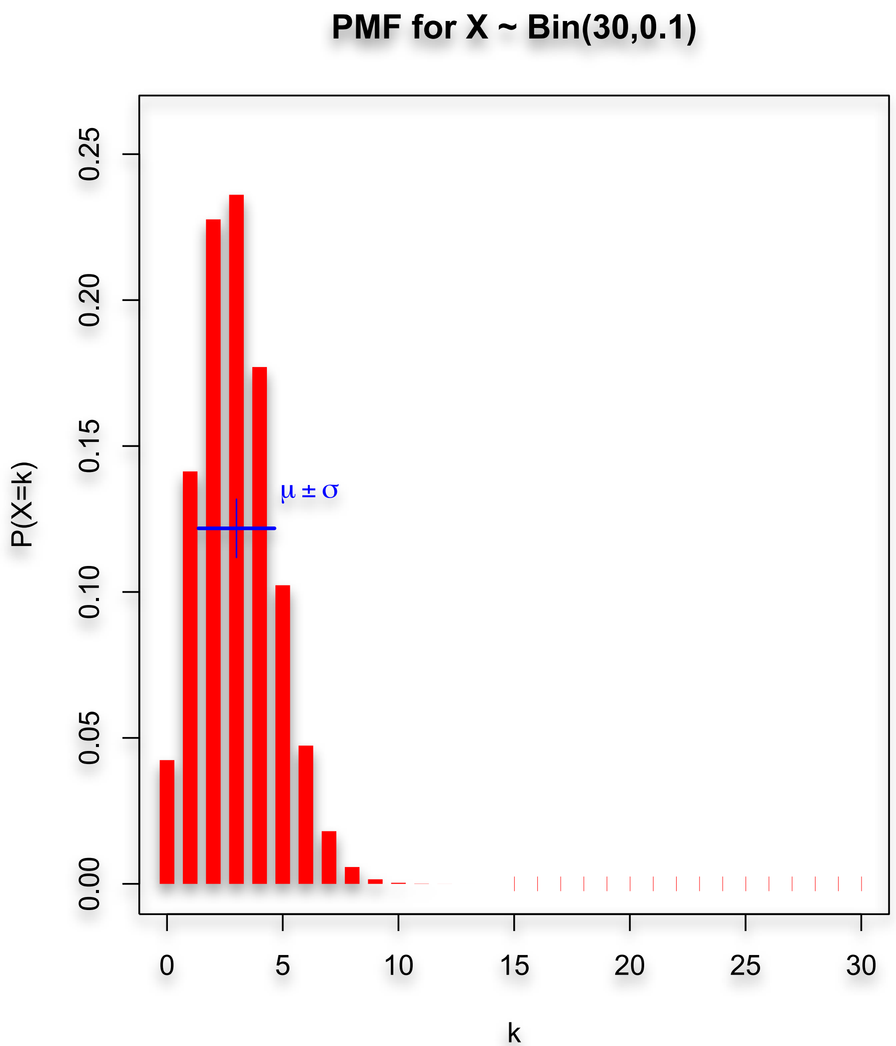
$\mathrm{Var}(X) = p(1-p)n$ ←(proof below, twice)

binomial pmfs






PMF for X ~ Bin(30,0.5)
P(X=k)
0.00
0.05
0.10
0.15
0.20
0.25
μ ± σ
0
5
10
15
20
25
30
k


PMF for X ~ Bin(30,0.1)
P(X=k)
0.00
0.05
0.10
0.15
0.20
0.25
μ ± σ
0
5
10
15
20
25
30
k

$$E[X^k] = \sum_{i=0}^{n} i^k \binom{n}{i} p^i (1-p)^{n-i}$$

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

using

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

$$E[X^k] = np\sum_{i=1}^{n} i^{k-1} \binom{n-1}{i-1} p^{i-1} (1-p)^{n-i}$$

letting
j = i-1

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

$$= npE[(Y+1)^{k-1}]$$

where Y is a binomial random variable with parameters $n-1, p$.

k=1 gives: $E[X] = np$

hence: $\mathrm{Var}(X) = E[X^2] - (E[X])^2$

$$= np[(n-1)p+1] - (np)^2$$

$$= np(1-p)$$

Theorem: If X & Y are INDEPENDENT

Then $E(X \cdot Y) = E(X) \cdot E(Y)$

Proof:

Let x_i, y_i, $i = 1, 2, \ldots$ be the possible values of X, Y

$$E(XY) = \sum_{i,j} x_i \cdot y_j \cdot P(X = x_i \wedge Y = y_j)$$

independence

$$= \sum_{i,j} x_i \cdot y_j \cdot P(X = x_i) \cdot P(Y = y_j)$$

$$= \sum_i x_i \cdot P(X = x_i) \cdot \sum_j y_j \cdot P(Y = y_j)$$

$$= E[X] \cdot E[Y]$$

Note: *NOT* true in general; see earlier example $E[X^2] \neq E[X]^2$

variance of *independent* r.v.s is additive

Theorem if X & Y are INDEPENDENT

then $Var(X+Y) = Var(X) + Var(Y)$ (Bienaymé, 1853)

Proof

Let $\hat{X} = X - E[X]$ $\quad \hat{Y} = Y - E[Y]$

$E[\hat{X}] = 0$ $\quad E[\hat{Y}] = 0$

$Var[\hat{X}] = Var[X]$ $\quad Var[\hat{Y}] = Var[Y]$ ← recall $Var(aX+b) = a^2 Var(X)$

$$
\begin{aligned}
Var(X+Y) &= Var(\hat{X}+\hat{Y}) \\
&= E[(\hat{X}+\hat{Y})^2] \\
&= E[\hat{X}^2 + 2\hat{X}\hat{Y} + \hat{Y}^2] \\
&= E[\hat{X}^2] + 2E[\hat{X}\hat{Y}] + E[\hat{Y}^2] \\
&= Var(\hat{X}) + 0 + Var(\hat{Y}) \\
&= Var(X) + Var(Y)
\end{aligned}
$$

Note:

"$Var(X+Y) = Var(X) + Var(Y)$" is not true in general.

E.g.: For any random variable X, let $Y = -X$.

Then $Var(X) = Var(Y)$, so $Var(X) + Var(Y) = 2Var(X)$

but $Var(X+Y) = 0$.

mean, variance of binomial r.v.s

If $Y_1, Y_2, \ldots, Y_n \sim Ber(p)$ Then $X = \sum_{i=1}^{n} Y_i \sim Bin(n, p)$

$E(X) = E(\sum_{i=1}^{n} Y_i) = nE[Y_1] = np$

$Var(X) = Var(\sum_{i=1}^{n} Y_i) = nVar[Y_1] = np(1-p)$

Ross 4.6 ex 6f

A disk array consists of n drives, each of which will fail independently with probability p.

Suppose it can operate effectively if at least one-half of its components function, e.g., by "majority vote."

For what values of p is a 5-component system more likely to operate effectively than a 3-component system?

X_5 = # failed in 5-component system ~ Bin(5, p)

X_3 = # failed in 3-component system ~ Bin(3, p)

X_5 = # failed in 5-component system ~ Bin(5, p)
X_3 = # failed in 3-component system ~ Bin(3, p)

P(5 component system effective) = P(X_5 < 5/2)

$$\binom{5}{0}p^0(1-p)^5 + \binom{5}{1}p^1(1-p)^4 + \binom{5}{2}p^2(1-p)^3$$

P(3 component system effective) = P(X_3 < 3/2)

$$\binom{3}{0}p^0(1-p)^3 + \binom{3}{1}p^1(1-p)^2$$

Calculation:

5-component system is better if and only if $p < 1/2$

The Hamming(7,4) code:

Have a 4-bit string to send over the network (or to disk)

Add 3 "parity" bits, and send 7 bits total

If bits are $b_1b_2b_3b_4$ then the three parity bits are

$parity(b_1b_2b_3)$, $parity(b_1b_3b_4)$, $parity(b_2b_3b_4)$

Each bit is independently corrupted (flipped) in transit with probability 0.1

X = number of bits corrupted ~ Bin(7, 0.1)

The Hamming code allow us to *correct* all 1 bit errors.

(E.g., if b_1 flipped, 1st 2 parity bits, but not 3rd, will look wrong; the only single bit error causing this symptom is b_1. Similarly for any other single bit being flipped. Some multi-bit errors can be detected, but not corrected, but not arbitrarily many.)

P(correctable message received) = P(X=0) + P(X=1)

Using error-correcting codes: X ~ Bin(7, 0.1)

$$P(X = 0) = \binom{7}{0}(0.1)^0(0.9)^7 \approx 0.4783$$

$$P(X = 1) = \binom{7}{1}(0.1)^1(0.9)^6 \approx 0.3720$$

$$P(X = 0) + P(X = 1) \approx 0.8503$$

What if we didn't use error-correcting codes?

X ~ Bin(4, 0.1)

P(correct message received) = P(X=0)

$$P(X = 0) = \binom{4}{0}(0.9)^4 \approx 0.6561$$

Using error correction improves reliability by 30% !

Sending a bit string over the network

n = 4 bits sent, each corrupted with probability 0.1

X = # of corrupted bits, X ~ Bin(4, 0.1)

In real networks, large bit strings (length $n \approx 10^4$)

Corruption probability is very small: $p \approx 10^{-6}$

$X \sim \text{Bin}(10^4, 10^{-6})$ is unwieldy to compute

Extreme n and p values arise in many cases

bit errors in file written to disk

of typos in a book

# of elements in particular bucket of large hash table # of server crashes per day in giant data center

facebook login requests sent to a particular server

poisson random variable

Siméon Poisson, 1781-1840

- X is a **Poisson** random variable: X ~ Poi(λ)
    - X takes values 0, 1, 2, …
    - and, for a given parameter λ,
    - has distribution (PMF):

$$P(X = i) = e^{-\lambda} \frac{\lambda^i}{i!}$$

- Note Taylor series: $e^\lambda = \frac{\lambda^0}{0!} + \frac{\lambda^1}{1!} + \cdots = \sum_{i=0}^{\infty} \frac{\lambda^i}{i!}$

so … $\sum_{i=0}^{\infty} P(X = i) = \sum_{i=0}^{\infty} e^{-\lambda} \frac{\lambda^i}{i!} = e^{-\lambda} \sum_{i=0}^{\infty} \frac{\lambda^i}{i!} = e^{-\lambda} e^{\lambda} = 1$

poisson random variable is binomial in the limit

- Poisson approximates binomial when n is large, p is small, and $\lambda = np$ is "moderate"

- Different interpretations of "moderate"
  - $n > 20$ and $p < 0.05$
  - $n > 100$ and $p < 0.1$

- Formally, Poisson is Binomial as

  $n \to \infty$ and $p \to 0$, where $np = \lambda$

binomial → poisson in the limit

$X \sim Bin(p, n)$

$$P(X=i) = \binom{n}{i} p^i (1-p)^{n-i}$$

$$= \frac{n!}{i!\,(n-i)!} \left(\frac{\lambda}{n}\right)^i \left(1-\frac{\lambda}{n}\right)^{n-i} \qquad \lambda = pn$$

$$= \frac{n(n-1)\cdots(n-i+1)}{n^i} \frac{\lambda^i}{i!} \frac{(1-\lambda/n)^n}{(1-\lambda/n)^i}$$

$$= \underbrace{\frac{n(n-1)\cdots(n-i+1)}{(n-\lambda)^i}}_{\approx 1} \frac{\lambda^i}{i!} \underbrace{(1-\lambda/n)^n}_{\approx e^{-\lambda}}$$

For large n, moderate λ, i

$$P(X=i) \approx e^{-\lambda} \frac{\lambda^i}{i!}$$, i.e. Binomial ≈ Poisson

sending data on a network, again

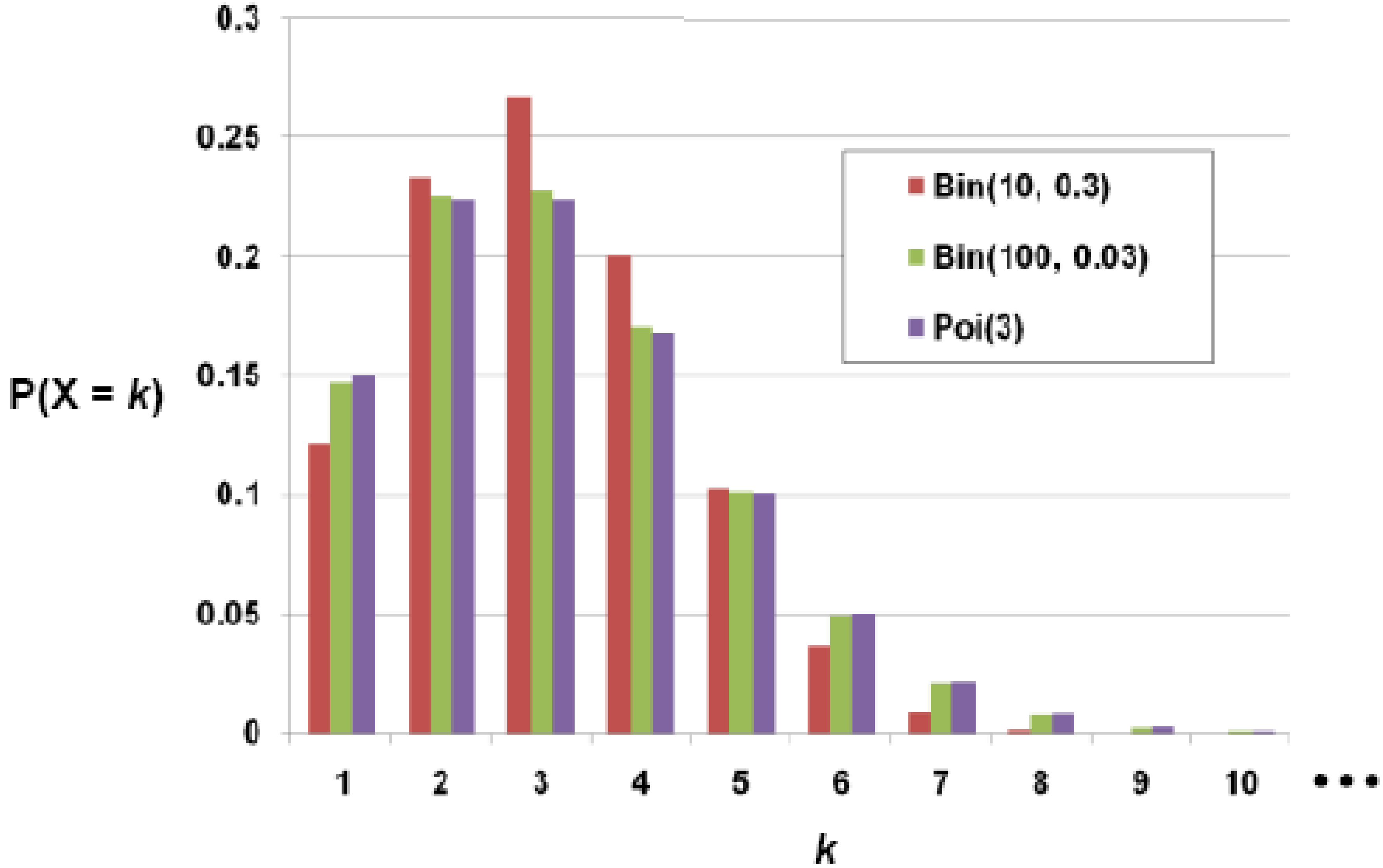
- Recall example of sending bit string over a network
  - Send bit string of length n = 10^4
  - Probability of (independent) bit corruption is p = 10^{-6}
  - $X \sim \text{Poi}(\lambda = 10^4 \cdot 10^{-6} = 0.01)$
  - What is probability that message arrives uncorrupted?

$$P(X=0) = e^{-\lambda}\frac{\lambda^i}{i!} = e^{-0.01}\frac{(0.01)^0}{0!} \approx 0.990049834$$

  - Using $Y \sim \text{Bin}(10^4, 10^{-6})$:

$$P(Y=0) \approx 0.990049829$$

Bin(10, 0.3), Bin(100, 0.03) vs. Poi(3)



expectation and variance of a poisson

- Recall: $Y \sim \text{Bin}(n,p)$
  - $E[Y] = pn$
  - $Var[Y] = np(1-p)$

- $X \sim \text{Poi}(\lambda)$ where $\lambda = np$ $(n \to \infty, p \to 0)$
  - $E[X] = np = \lambda$
  - $Var[X] = np(1-p) = \lambda(1-0) = \lambda$
  - Expectation and variance of a Poisson are the same

Suppose a server can process 2 requests per second

Requests arrive at random at an average rate of 1/sec

Unprocessed requests are held in a *buffer*

Q. How big a buffer do we need to avoid ever dropping a request?

A. Infinite

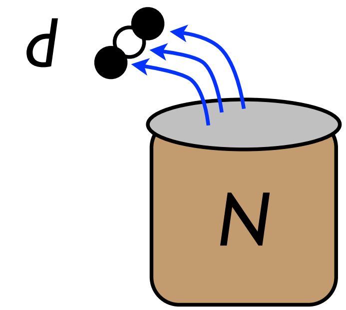
Q. How big a buffer do we need to avoid dropping a request more often than once a day?

A. (approximate) If X is the number of arrivals in a second, then X is poisson(λ=1). We want b s.t.
$P(X > b) < 1/(24*60*60) \approx 10^{-5}$

$P(X = b) = e^{-1}/b!$ $P(X=8) \approx \Sigma_{i>7} P(X=i) \approx 1.02e\text{-}05$

balls in urns – the hypergeometric distribution

Draw *d* balls (without replacement) from an urn containing *N*, of which *w* are white, the rest black.



Let *X* = number of white balls drawn

$$P(X=i) = \frac{\binom{w}{i}\binom{N-w}{d-i}}{\binom{N}{d}}, \quad i = 0, 1, \ldots, d$$

(note: n choose k = 0 if k < 0 or k > n)

E[X] = dp, where p = w/N (the fraction of white balls)

proof: Let X_i be 0/1 indicator for i-th ball is white, $X = \Sigma X_i$

The X_i are dependent, but $E[X] = E[\Sigma X_i] = \Sigma E[X_i] = dp$

Var[x] = dp(1-p)(1-(d-1)/(N-1))

$N \approx 22500$ human genes, many of unknown function

Suppose in some experiment, d =1588 of them were observed (say, they were all switched on in response to some drug)

A big question: What are they doing?

One idea: The Gene Ontology Consortium (www.geneontology.org) has grouped genes with known functions into categories such as "muscle development" or "immune system." Suppose 26 of your d genes fall in the "muscle development" category.

Just chance?

Or call Coach & see if he wants to dope some athletes?

Hypergeometric: GO has 116 genes in the muscle development category. If those are the white balls among 22500 in an urn, what is the probability that you would see 26 of them in 1588 draws?

Cao, et al., Developmental Cell 18, 662–674, April 20, 2010

data mining

Table 2. Gene Ontology Analysis on Differentially Bound Peaks in Myoblasts versus Myotubes

GO Categories Enriched in Genes Associated with Myotube-Increased Peaks

GOID	Term	P Value	OR[a]	Count[b]	Size[c]	Ont[d]
GO:0005856	cytoskeleton	2.05E-11	2.40	94	490	CC
GO:0043292	contractile fiber	6.98E-09	5.85	22	58	CC
GO:0030016	myofibril	1.96E-08	5.74	21	56	CC
GO:0044449	contractile fiber part	2.58E-08	5.97	20	52	CC
GO:0030017	sarcomere	4.95E-08	6.04	19	49	CC
GO:0008092	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	MF
GO:0007519	skeletal muscle development	2.50E-06	4.13	20	65	BP
GO:0015629	actin cytoskeleton	4.73E-06	3.08	27	111	CC
GO:0003779	actin binding	[illegible]	[illegible]	[illegible]	159	MF
GO:0006936	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	BP
GO:0044430	cytoskeletal part	[illegible]	[illegible]	[illegible]	294	CC
GO:0031674	I band	2.27E-05	5.67	12	32	CC
GO:0003012	muscle system process	2.54E-05	4.11	16	52	BP
GO:0030029	actin filament-based process	2.89E-05	2.73	27	119	BP
GO:0007517	muscle development	5.06E-05	2.69	26	116	BP

probability of seeing this many genes from a set of this size by chance according to the hypergeometric distribution.

E.g., if you draw 1588 balls from an urn containing 490 white balls and ≈22000 black balls, P(94 white) $\approx 2.05 \times 10^{-11}$

A differentially bound peak was associated to the closest gene (unique Entrez ID) measured by distance to TSS within CTCF flanking domains. OR: ratio of predicted to observed number of genes within a given GO category. Count: number of genes with differentially bound peaks. Size: total number of genes for a given functional group. Ont: the Geneontology. BP = biological process, MF = molecular function, CC = cellular component.