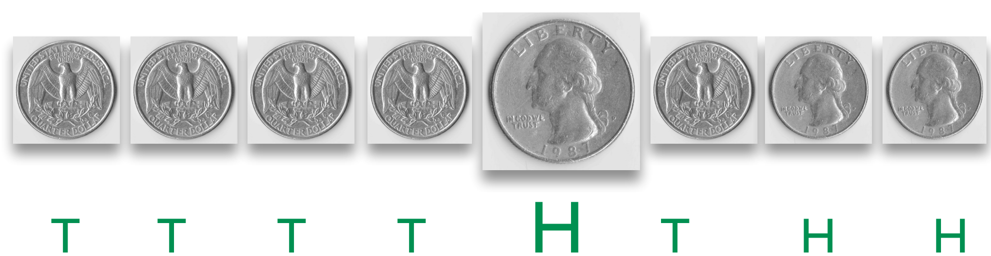
6. random variables



let X = index of

A *random variable* is a *numeric function* of the outcome of an experiment, not the outcome itself.

(Technically, neither random nor a variable, but...)

Ex.

Let H be the *number* of Heads when 20 coins are tossed

Let T be the *total* of 2 dice rolls

Let X be the *number* of coin tosses needed to see 1st head

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

Outcome	*X = #H*	*P(X)*
TT	0	P(X=0) = 1/4
TH	1	P(X=1) = 1/2
HT	1	
HH	2	P(X=2) = 1/4

20 balls numbered 1, 2, ..., 20

Draw 3 without replacement

Let X = the maximum of the numbers on those 3 balls

What is P(X ≥ 17)

$$
\begin{aligned}
P(X=20) &= \binom{19}{2} \Big/ \binom{20}{3} = \frac{3}{20} = 0.150 \\
P(X=19) &= \binom{18}{2} \Big/ \binom{20}{3} = \frac{18 \cdot 17/2!}{20 \cdot 19 \cdot 18/3!} \approx 0.134 \\
&\vdots
\end{aligned}
$$

$$\sum_{i=17}^{20} P(X=i) \approx 0.508$$

Alternatively:

$$P(X \geq 17) = 1 - P(X < 17) = 1 - \binom{16}{3} \Big/ \binom{20}{3} \approx 0.508$$

Flip a (biased) coin repeatedly until 1[st] 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) = $(1\text{-}p)^2p$

...

$$\sum_{i\geq 0} x^i = \frac{1}{1-x},$$

when $|x| < 1$

memorize me!

Check that it is a valid probability distribution:

1) $\forall i \geq 1, P(\{X = i\}) \geq 0$

2) $P\left(\bigcup_{i\geq 1}\{X = i\}\right) = \sum_{i\geq 1}(1-p)^{i-1}p = p\sum_{i\geq 0}(1-p)^i = p\frac{1}{1-(1-p)} = 1$

A *discrete* random variable is one taking on a *countable* number of possible values.

Ex:

X = sum of 3 dice, $3 \le X \le 18$, $X \in N$

Y = number of 1st head in seq of coin flips, $1 \le Y$, $Y \in N$

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

Definition: If *X* is a discrete random variable taking on values from a countable set $T \subseteq \mathcal{R}$, then

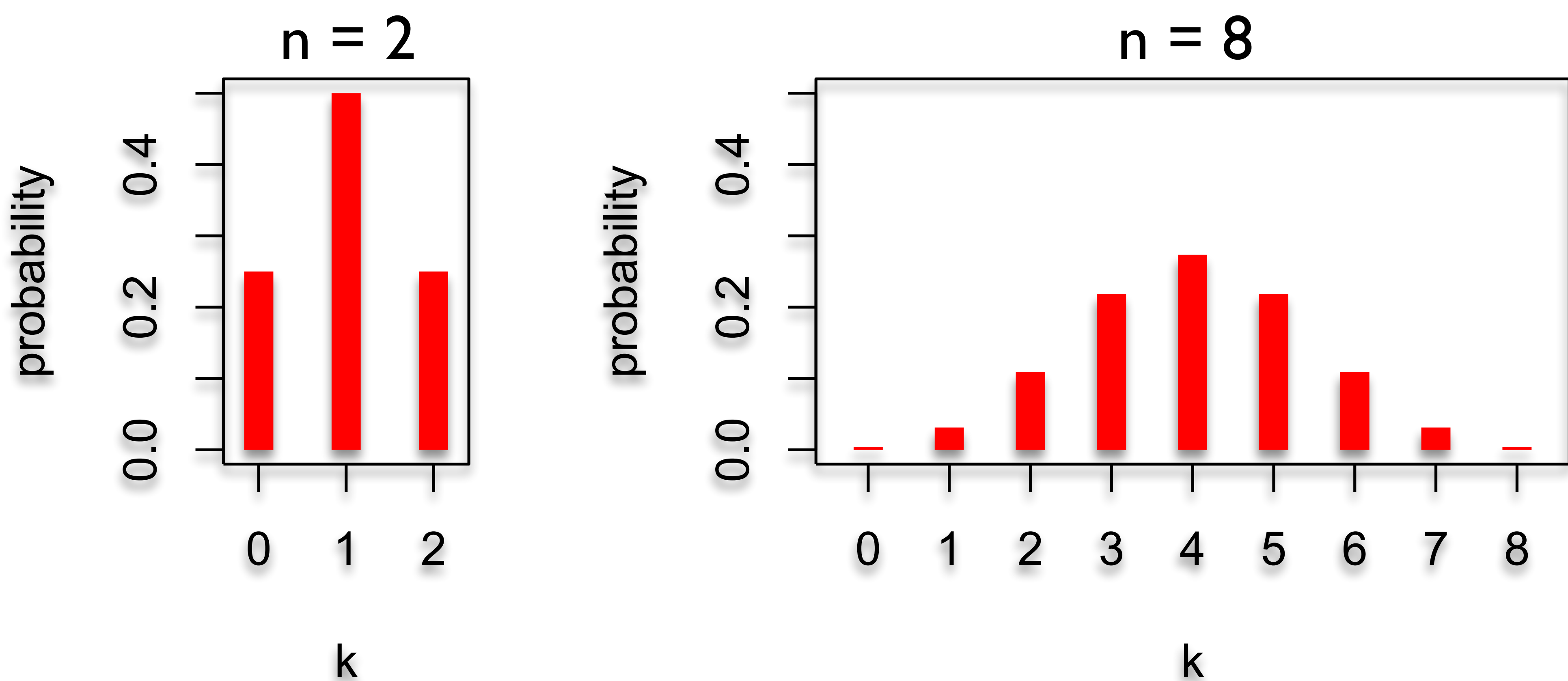
$$p(a) = \begin{cases} P(X = a) & \text{for } a \in T \\ 0 & \text{otherwise} \end{cases}$$

is called the *probability mass function*. Note: $\sum_{a \in T} p(a) = 1$

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

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

Probability mass function (p = ½):



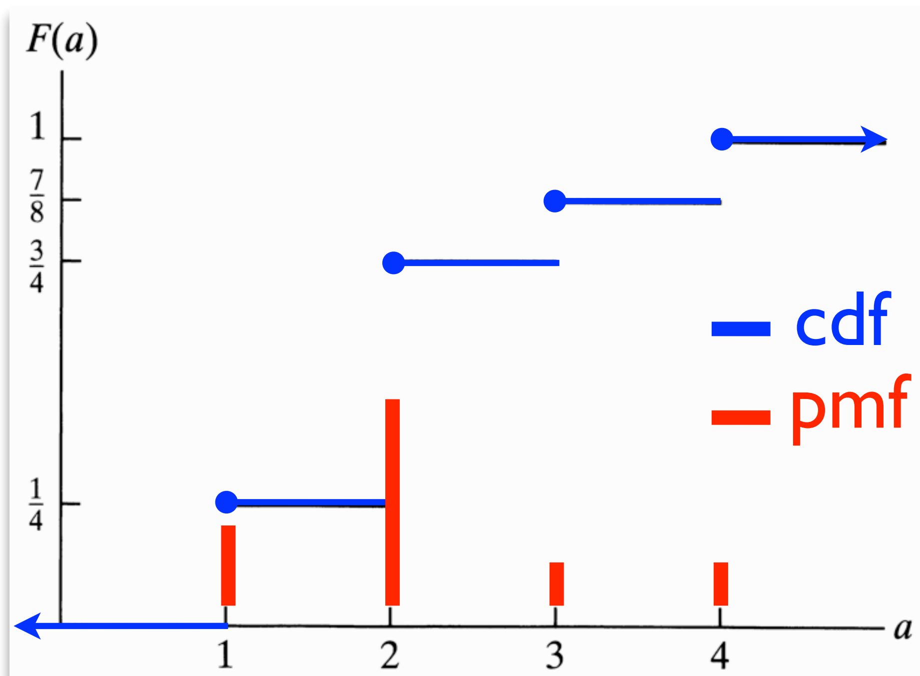
The *cumulative distribution function* for a random variable X is the function $F: \mathcal{R} \rightarrow [0,1]$ defined by

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

Ex: if X has probability mass function given by:

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

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



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

Why use random variables?

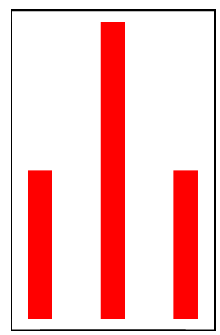
A. Often we just care about numbers

If I win $1 per head when 20 coins are tossed, what is my average winnings? What is the most likely number? What is the probability that I win < $5? ...

B. It cleanly abstracts away unnecessary detail about the experiment/sample space; PMF is all we need.

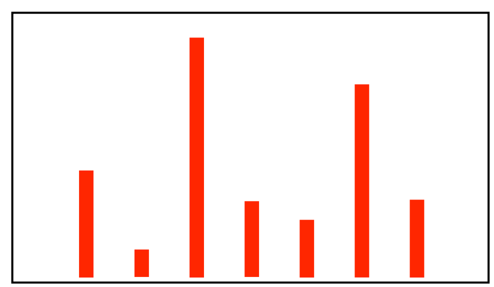
Outcome	*x=#H*	*P(X)*
TT	0	P(X=0) = 1/4
TH	1	P(X=1) = 1/2
HT	1	
HH	2	P(X=2) = 1/4

→



Flip 7 coins, roll 2 dice, and throw a dart; if dart landed in sector = dice roll mod #heads, then *X* = ...

→



expectation

For a discrete r.v. X with p.m.f. p(•), the *expectation of X*, aka *expected value* or *mean*, is

$$\boxed{E[X] = \Sigma_x\, xp(x)}$$

average of random values, *weighted* by their respective probabilities

For the equally-likely outcomes case, this is just the average of the possible random values of X

For *un*equally-likely outcomes, it is again the average of the possible random values of X, weighted by their respective probabilities

Ex 1: Let X = value seen rolling a fair die $p(1), p(2), ..., p(6) = 1/6$

$$E[X] = \sum_{i=1}^{6} ip(i) = \frac{1}{6}(1+2+\cdots+6) = \frac{21}{6} = 3.5$$

Ex 2: Coin flip; X = +1 if H (win \$1), -1 if T (lose \$1)

$$E[X] = (+1) \cdot p(+1) + (-1) \cdot p(-1) = 1 \cdot (1/2) + (-1) \cdot (1/2) = 0$$

For a discrete r.v. X with p.m.f. $p(\bullet)$, the *expectation of X*, aka *expected value* or *mean*, is

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

average of random values, weighted by their respective probabilities

Another view: A 2-person gambling game. If X is how much you win playing the game once, how much would you expect to win, on average, per game, when repeatedly playing?

Ex 1: Let X = value seen rolling a fair die $p(1), p(2), ..., p(6) = 1/6$

If you win X dollars for that roll, how much do you expect to win?

$$E[X] = \sum_{i=1}^{6} ip(i) = \frac{1}{6}(1 + 2 + \cdots + 6) = \frac{21}{6} = 3.5$$

Ex 2: Coin flip; X = +1 if H (win \$1), -1 if T (lose \$1)

$$E[X] = (+1) \bullet p(+1) + (-1) \bullet p(-1) = 1 \bullet (1/2) + (-1) \bullet (1/2) = 0$$

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

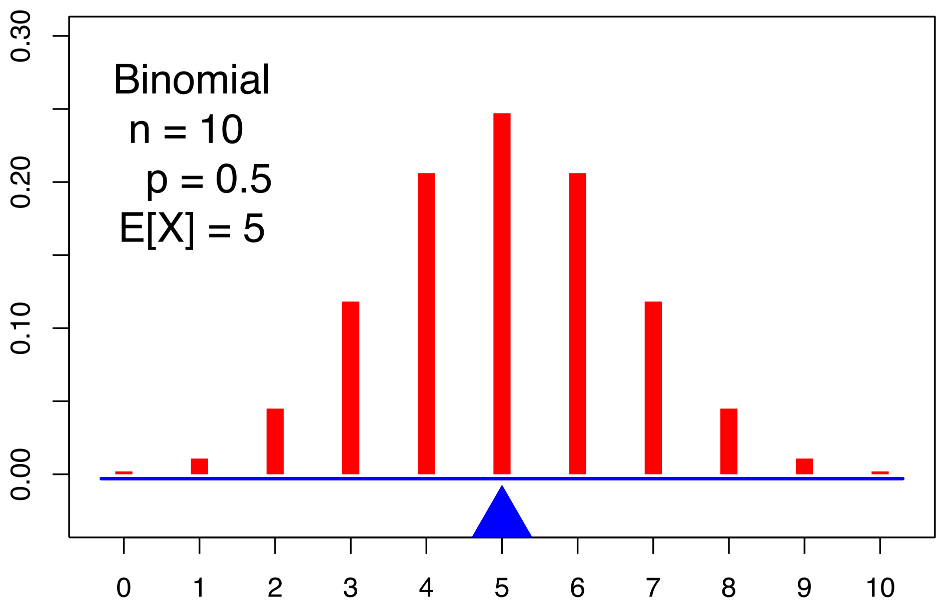
For a discrete r.v. X with p.m.f. $p(\bullet)$, the *expectation of X*, aka *expected value* or *mean*, is

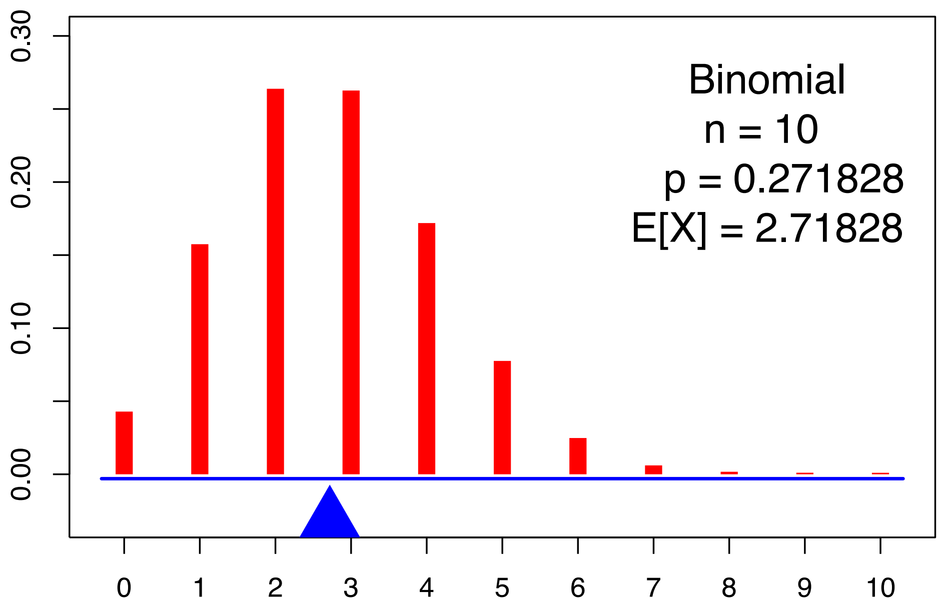
$$E[X] = \Sigma_x \, x p(x)$$

average of random values, weighted by their respective probabilities

A third view: $E[X]$ is the "balance point" or "center of mass" of the probability mass function

Ex: Let X = number of heads seen when flipping 10 coins





Let X be the number of flips up to & including 1st head observed in repeated flips of a biased coin. If I pay you $1 per flip, how much money would you expect to make?

$$\begin{aligned} P(H) &= p; \quad P(T) = 1-p = q \\ p(i) &= pq^{i-1} \quad \leftarrow \textit{PMF} \\ E[X] &= \textstyle\sum_{i\geq 1} ip(i) = \sum_{i\geq 1} ipq^{i-1} = p\sum_{i\geq 1} iq^{i-1} \qquad (*) \end{aligned}$$

A calculus trick:

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

$dy^0/dy = 0$

So $(*)$ becomes:

$$E[X] = p\sum_{i\geq 1} iq^{i-1} = \frac{p}{(1-q)^2} = \frac{p}{p^2} = \frac{1}{p}$$

E.g.:

p=1/2; on average head every 2nd flip

p=1/10; on average, head every 10th flip.

How much would you pay to play?

(To geo)

Let X be the number of heads observed in n repeated flips of a biased coin. If I pay you $1 per head, how much money would you expect to make?

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

E.g.:

p=1/2; on average, n/2 heads

p=1/10; on average, n/10 heads

How much would you pay to play?

(compare to slide 24, slide 56)

Way 1

Calculating $E[g(X)]$:

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

X = sum of 2 dice rolls

i	$p(i) = P[X=i]$	$i \cdot p(i)$
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

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

$Y = g(X) = X \bmod 5$

j	$q(j) = P[Y = j]$	$j \cdot q(j)$
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

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

Way 2

expectation of a *function* of a random variable

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

X = sum of 2 dice rolls

i	$p(i) = P[X=i]$	$g(i) \bullet p(i)$
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

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

$Y = g(X) = X \bmod 5$

j	$q(j) = P[Y = j]$	$j \bullet q(j)$
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

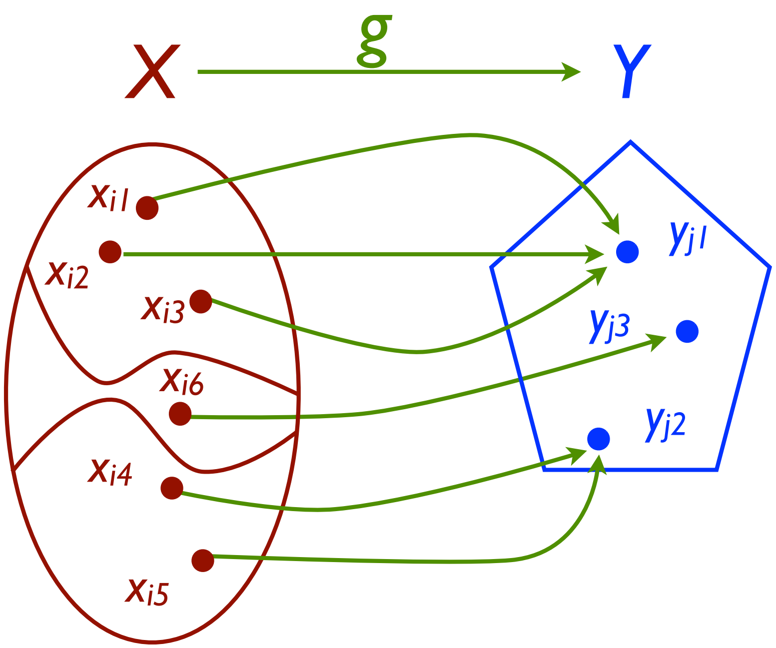
$E[Y] = \Sigma_j\, jq(j) =$ 72/36 $= 2$

BT pg.84-85

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, then flip 2 coins:

HH	*A* wins \$2
HT	Each takes back \$1
TH	
TT	*B* wins \$2

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

$P(X = +1) = 1/4$
$P(X = 0) = 1/2$
$P(X = -1) = 1/4$

What is $E[X]$?

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

What is $E[X^2]$?

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

Big Deal Note:
$E[X^2] \neq E[X]^2$

Linearity of expectation, I

For any constants *a*, *b*: $\boxed{E[aX + b] = aE[X] + b}$

Proof:

$$
\begin{aligned}
E[aX+b] &= \sum_x (ax+b)\cdot p(x) \\
&= a\sum_x xp(x) + b\sum_x p(x) \\
&= aE[X] + b
\end{aligned}
$$

A & *B* each bet \$1, then flip 2 coins:

HH	*A* wins \$2
HT	Each takes back \$1
TH	
TT	*B* wins \$2

Let X = *A*'s net gain: +1, 0, -1, resp.:

$P(X = +1) = 1/4$
$P(X = 0) = 1/2$
$P(X = -1) = 1/4$

From slide 18

What is $E[X]$?

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

What is $E[X^2]$?

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

What is $E[2X+1]$?

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

Example:
Caezzo's Palace Casino offers the following game: They flip a biased coin (P(Heads) = 0.10) until the first Head comes up. "You're on a hot streak now! The more Tails the more you win!" Let X be the number of flips up to & including 1st head. They will pay you \$2 per flip, i.e., $2X$ dollars. They charge you \$25 to play.

Q: Is it a fair game? On average, how much would you expect to win/lose per game, if you play it repeatedly?

A: Not fair. Your net winnings per game is $2X - 25$, and
$E[2X - 25] = 2E[X] - 25 = 2(1/0.10) - 25 = -5$,
i.e., you lose \$5 per game on average

Linearity, II

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

$$\boxed{E[X+Y] = E[X] + E[Y]}$$

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: 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 of $E[X] = \sum_x x \cdot P(X = x)$

Then:

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

A & *B* each bet \$1, then flip 2 coins:

HH	*A* wins \$2
HT	Each takes back \$1
TH	
TT	*B* wins \$2

Let X = *A*'s net gain: +1, 0, -1, resp.:

$$P(X = +1) = 1/4$$
$$P(X = 0) = 1/2$$
$$P(X = -1) = 1/4$$

From slide 18

What is $E[X]$?

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

What is $E[X^2]$?

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

What is $E[X^2+2X+1]$?

$$E[X^2 + 2X + 1] = E[X^2] + 2E[X] + 1 = 1/2 + 2 \cdot 0 + 1 = 1.5$$

Example

X = # of heads in *one* coin flip, where $P(X=1) = p$.

What is $E(X)$?

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

Let X_i, $1 \leq i \leq n$, be # of H in flip of coin with $P(X_i=1) = p_i$

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

$E[\Sigma_i X_i] = \Sigma_i E[X_i] = \Sigma_i p_i$

Special case: $p_1 = p_2 = ... = p$:

$E[\text{\# of heads in } n \text{ flips}] = pn$

☜ Compare to slide 14

Note:

Linearity is special!

It is *not* true in general that

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

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

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

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

•

•

•

variance

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

$\ldots$ Time passes $\ldots$

Alice (yawning) says "let's raise the stakes"

$$Y = \begin{cases} +1000 & \text{if Heads} \\ -1000 & \text{if Tails} \end{cases}$$

E[Y] = 0, as before.

Are you (Bob) equally happy to play the new game?

E[X] measures the "average" or "central tendency" of X. What about its *variability*?

If E[X] = μ, then E[|X-μ|] seems like a natural quantity to look at: how much do we expect (on average) X to deviate from its average.

Unfortunately, it's a bit inconvenient mathematically; following is nicer/easier/much more common.

Definitions

The *variance* of a random variable X with mean $E[X] = \mu$ is

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

often denoted σ^2.

The *standard deviation* of X is

$$\sigma = \sqrt{\mathrm{Var}[X]}$$

The *variance* of a random variable X with mean $E[X] = \mu$ is $Var[X] = E[(X-\mu)^2]$, often denoted σ^2.

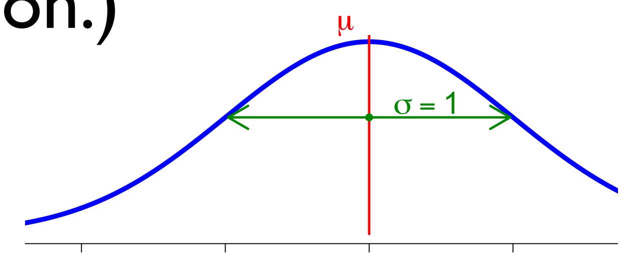
I: Square always ≥ 0, and exaggerated as X moves away from μ, so $Var[X]$ emphasizes *deviation* from the mean.

II: Numbers vary a lot depending on exact distribution of X, but it is common that X is

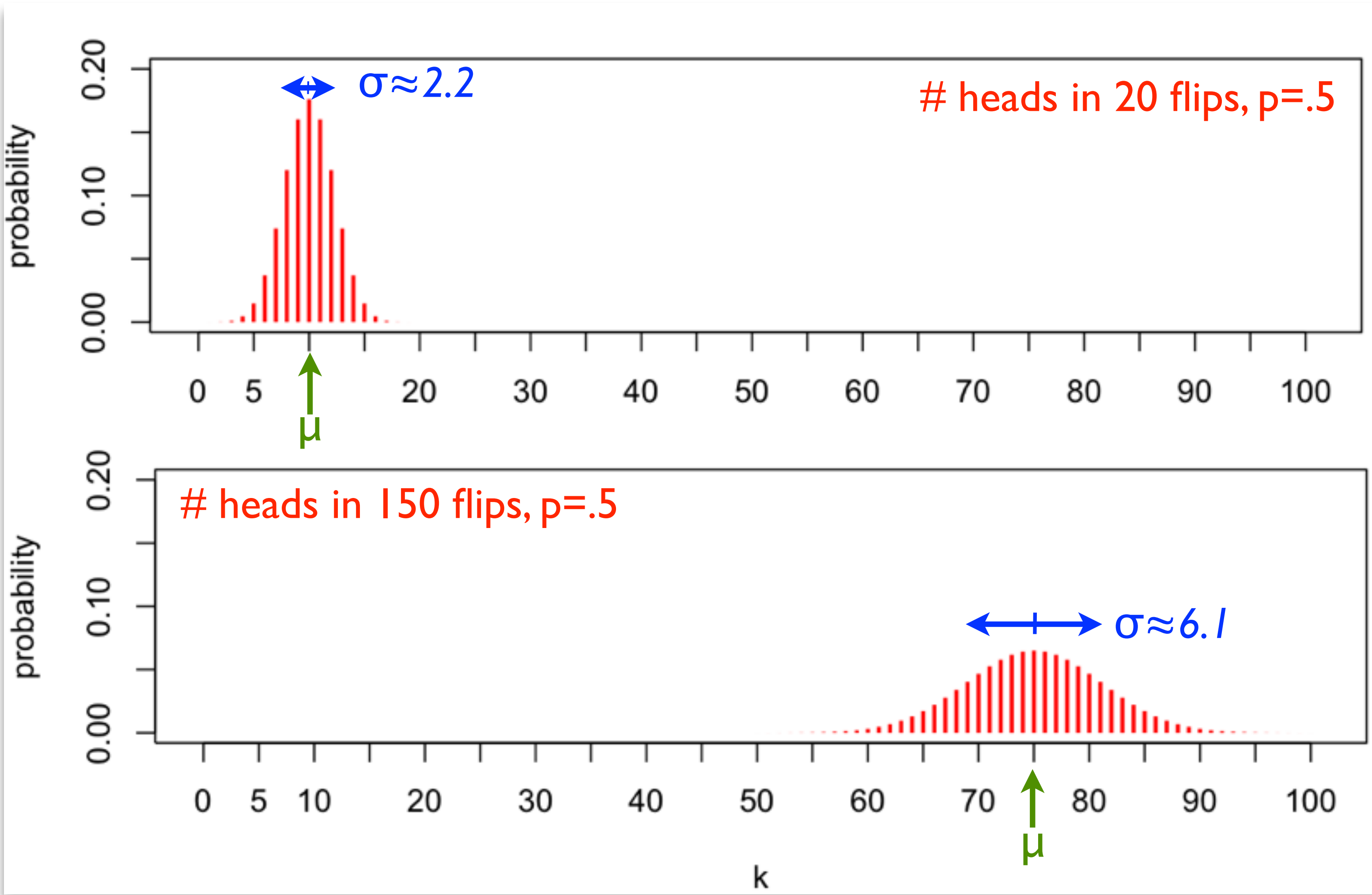
within $\mu \pm \sigma$ ~66% of the time, and

within $\mu \pm 2\sigma$ ~95% of the time.

(We'll see the reasons for this soon.)



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



Blue arrows denote the interval $\mu \pm \sigma$

(and note σ *bigger in absolute terms in second ex., but smaller as a proportion of* μ *or max.)*

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

Var[X] = 1

… Time passes …

Alice (yawning) says "let's raise the stakes"

$$Y = \begin{cases} +1000 & \text{if Heads} \\ -1000 & \text{if Tails} \end{cases}$$

E[Y] = 0, as before.

Var[Y] = 1,000,000

Are you (Bob) equally happy to play the new game?

Two games:

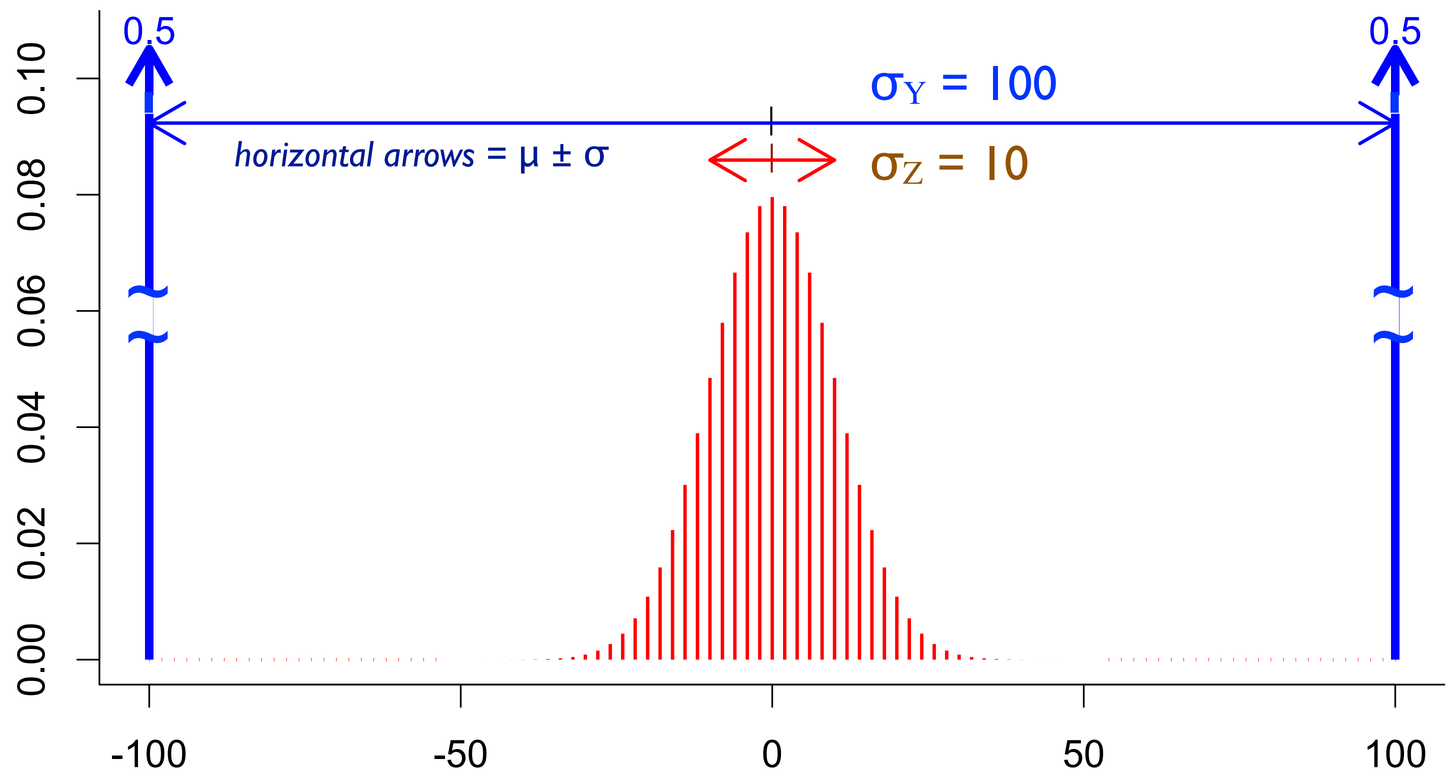
a) flip 1 coin, win Y = \$100 if heads, \$-100 if tails

b) flip 100 coins, win Z = (#(heads) - #(tails)) dollars

Same expectation in both: $E[Y] = E[Z] = 0$

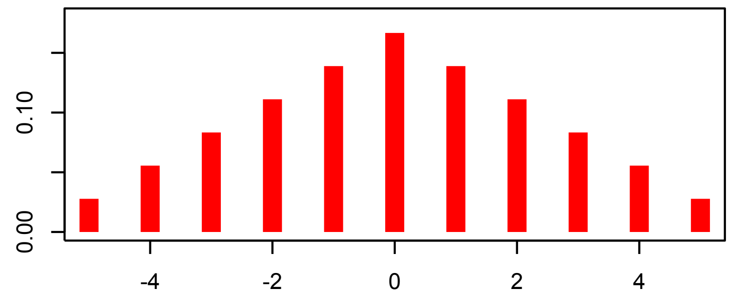
Same extremes in both: max gain = \$100; max loss = \$100

But *variability* is very different:



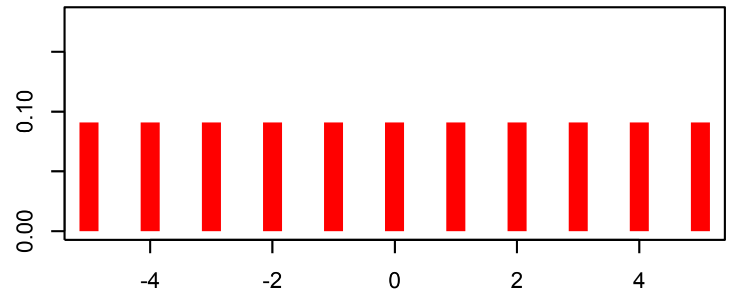
more variance examples

X_1 = sum of 2 fair dice, minus 7



$\sigma^2 = 5.83$

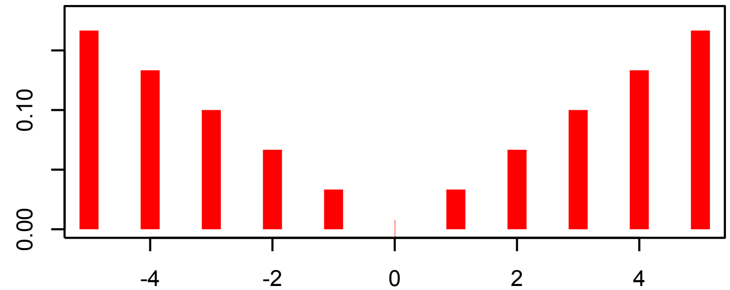
X_2 = fair 11-sided die labeled -5, ..., 5



$\sigma^2 = 10$

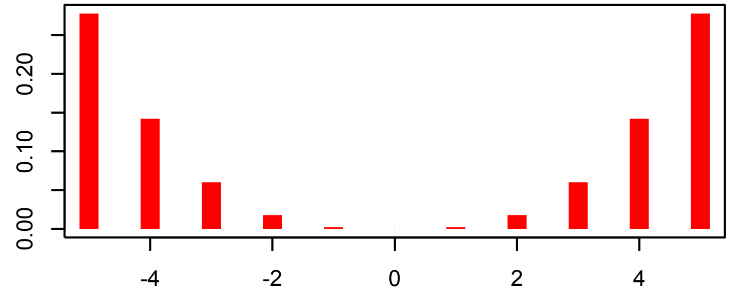
-1, 0, +1

$X_3 = Y$-6•signum(Y), where Y is the *difference* of 2 fair dice, *given* no doubles



$\sigma^2 = 15$

$X_4 = X_3$ when 3 pairs of dice all give same X_3



$\sigma^2 = 19.7$

NB: Wow, kinda complex; see slide 8

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

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

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

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

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

NOT linear;
insensitive to location (b),
quadratic in scale (a)

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

Ex:

$$X = \begin{cases} +1 & \text{if Heads} \\ -1 & \text{if Tails} \end{cases}$$

E[X] = 0
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

Var[Y] = Var[10^3 X] = 10^6Var[X] = 10^6

In general:

$$\boxed{Var[X+Y] \underset{\wedge\wedge\wedge\wedge\wedge\wedge\wedge}{\neq} Var[X] + Var[Y]}$$

NOT linear

Ex 1:

Let X = ±1 based on 1 coin flip

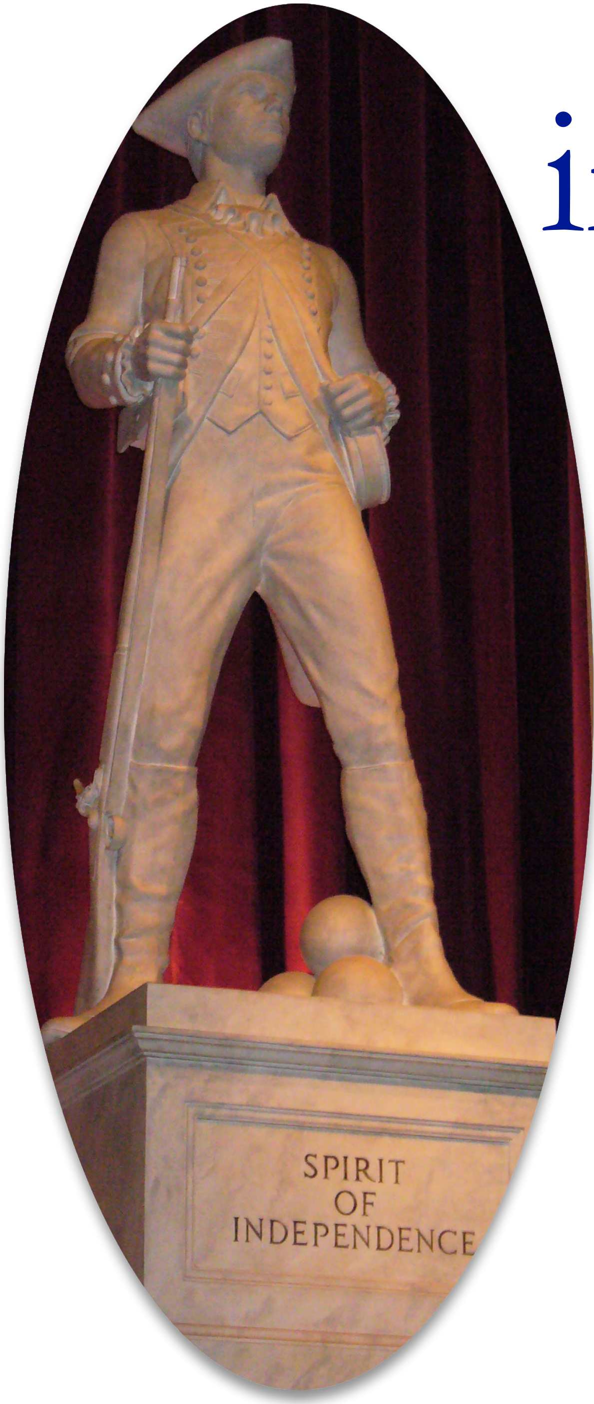
As shown above, E[X] = 0, Var[X] = 1

Let Y = -X; then Var[Y] = $(-1)^2$Var[X] = 1

But X+Y = 0, always, so Var[X+Y] = 0

Ex 2:

As another example, is Var[X+X] = 2Var[X]?



independence

and

joint

distributions

Defn: Random variable X and event E are independent if the event E is independent of the event {X=x} (for any fixed x), i.e.

$$\forall x \; P(\{X = x\} \;\&\; E) = P(\{X=x\}) \cdot P(E)$$

Defn: Two random variables X and Y are independent if the events {X=x} and {Y=y} are independent (for any fixed x, y), i.e.

$$\forall x, y \; P(\{X = x\} \;\&\; \{Y=y\}) = P(\{X=x\}) \cdot P(\{Y=y\})$$

Intuition as before: knowing X doesn't help you guess Y or E and vice versa.

Random variable X and event E are independent if

$$\forall x \; P(\{X = x\} \; \& \; E) = P(\{X=x\}) \bullet P(E)$$

Ex 1: Roll a fair die to obtain a random number $1 \leq X \leq 6$, then flip a fair coin X times. Let E be the event that the number of heads is even.

$P(\{X=x\}) = 1/6$ for any $1 \leq x \leq 6$,

$P(E) = 1/2$

$P(\{X=x\} \; \& \; E) = 1/12$, so they are independent

Ex 2: as above, and let F be the event that the total number of heads = 6.

$P(F) = 2^{-6}/6 > 0$, and considering, say, X=4, we have $P(X=4) = 1/6 > 0$ (as above), but $P(\{X=4\} \; \& \; F) = 0$, since you can't see 6 heads in 4 flips. So X & F are *dependent*. (Knowing that X is small renders F impossible; knowing that F happened means X must be 6.)

Two random variables X and Y are independent if the events {X=x} and {Y=y} are independent (for any x, y), i.e.

$\forall x, y$ P({X = x} & {Y=y}) = P({X=x}) • P({Y=y})

Ex: Let X be number of heads in first n of 2n coin flips, Y be number in the last n flips, and let Z be the total. X and Y are independent:

$$P(X = j) = \binom{n}{j} 2^{-n}$$

$$P(Y = k) = \binom{n}{k} 2^{-n}$$

$$P(X = j \wedge Y = k) = \binom{n}{j}\binom{n}{k} 2^{-2n} = P(X = j)P(Y = k)$$

But X and Z are *not* independent, since, e.g., knowing that X = 0 precludes Z > n. E.g., P(X = 0) and P(Z = n+1) are both positive, but P(X = 0 & Z = n+1) = 0.

Often, several random variables are *simultaneously* observed

X = height and Y = weight

X = cholesterol and Y = blood pressure

X_1, X_2, X_3 = work loads on servers A, B, C

Joint probability mass function:

$f_{XY}(x, y) = P(\{X = x\} \,\&\, \{Y = y\})$

Joint cumulative distribution function:

$F_{XY}(x, y) = P(\{X \leq x\} \,\&\, \{Y \leq y\})$

Two joint PMFs

W \ Z	1	2	3
1	2/24	2/24	2/24
2	2/24	2/24	2/24
3	2/24	2/24	2/24
4	2/24	2/24	2/24

X \ Y	1	2	3
1	4/24	1/24	1/24
2	0	3/24	3/24
3	0	4/24	2/24
4	4/24	0	2/24

$P(W = Z) = 3 * 2/24 = 6/24$

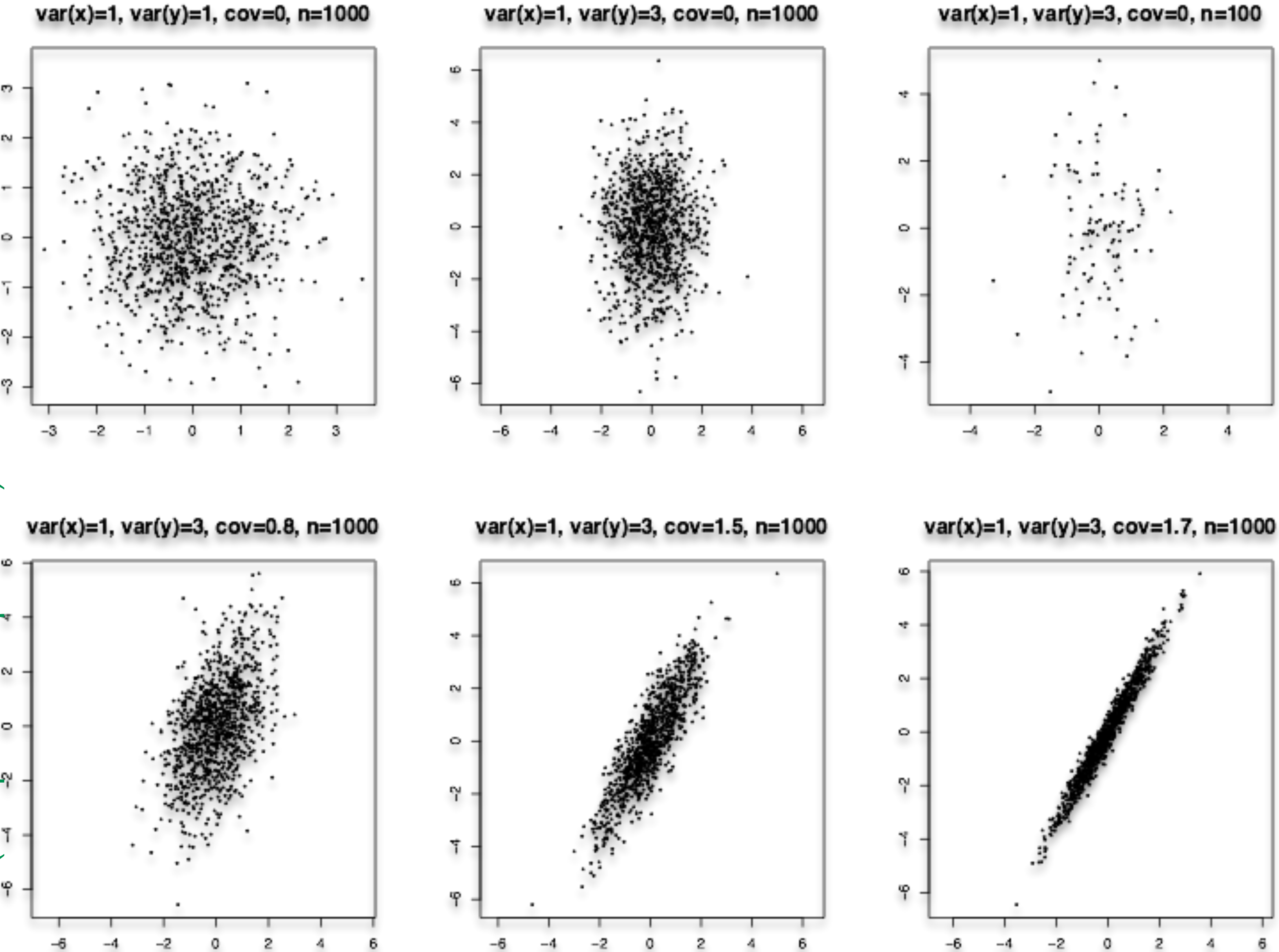
$P(X = Y) = (4 + 3 + 2)/24 = 9/24$

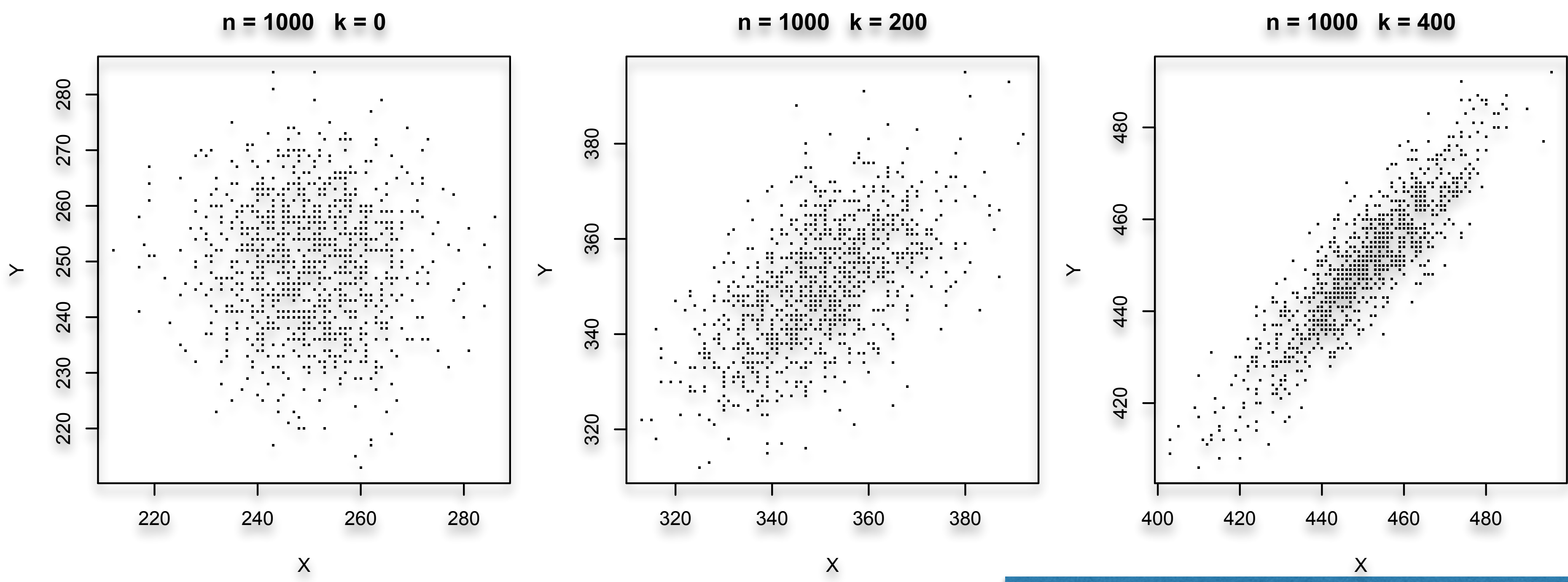
Can look at arbitrary relationships among variables this way

sampling from a joint distribution

Top row; independent variables

bottom row: dependent variables
(a simple linear dependence)

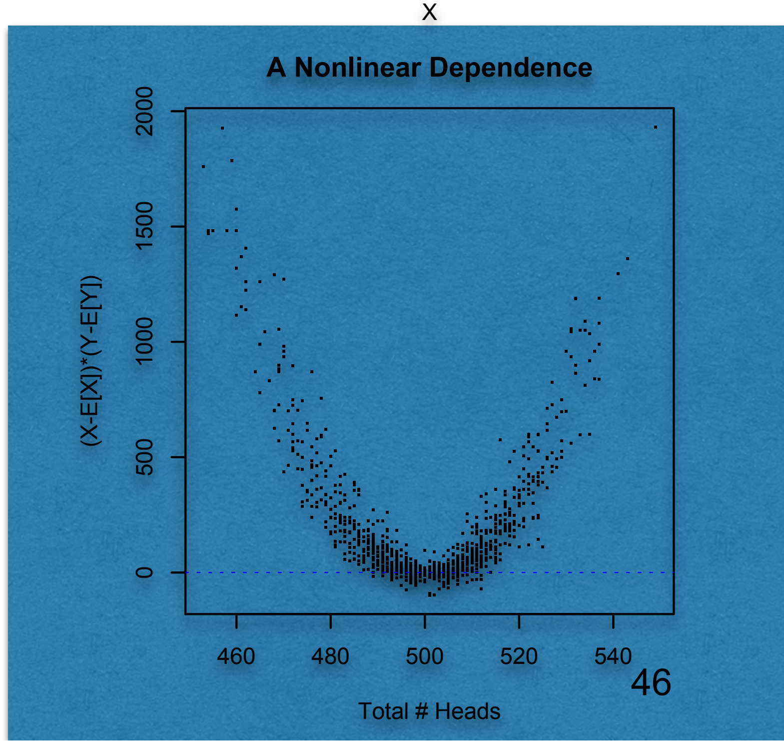




Flip n fair coins

X = #Heads seen in first n/2+k

Y = #Heads seen in last n/2+k



Two joint PMFs

W \ Z	1	2	3	$f_W(w)$
1	2/24	2/24	2/24	6/24
2	2/24	2/24	2/24	6/24
3	2/24	2/24	2/24	6/24
4	2/24	2/24	2/24	6/24
$f_Z(z)$	8/24	8/24	8/24	

X \ Y	1	2	3	$f_X(x)$
1	4/24	1/24	1/24	6/24
2	0	3/24	3/24	6/24
3	0	4/24	2/24	6/24
4	4/24	0	2/24	6/24
$f_Y(y)$	8/24	8/24	8/24	

Marginal PMF of one r.v.: sum over the other (Law of total probability)

$$f_Y(y) = \Sigma_x f_{XY}(x,y)$$

$$f_X(x) = \Sigma_y f_{XY}(x,y)$$

Question: Are W & Z independent? Are X & Y independent?

Repeating the Definition: Two random variables X and Y are independent if the events {X=x} and {Y=y} are independent (for any fixed x, y), i.e.

$$\forall x, y\ P(\{X = x\} \,\&\, \{Y=y\}) = P(\{X=x\}) \cdot P(\{Y=y\})$$

Equivalent Definition: Two random variables X and Y are independent if their *joint* probability mass function is the product of their *marginal* distributions, i.e.

$$\forall x, y\ f_{XY}(x,y) = f_X(x) \cdot f_Y(y)$$

Exercise: Show that this is also true of their *cumulative* distribution functions

A function g(X,Y) defines a new random variable.

Its expectation is:

$$E[g(X, Y)] = \Sigma_x \Sigma_y \, g(x, y) \, f_{XY}(x,y)$$

☜ like slide 17

Expectation is linear. E.g., if g is linear:

$$E[g(X, Y)] = E[a\,X + b\,Y + c] = a\,E[X] + b\,E[Y] + c$$

Example:

g(X,Y) = 2X-Y

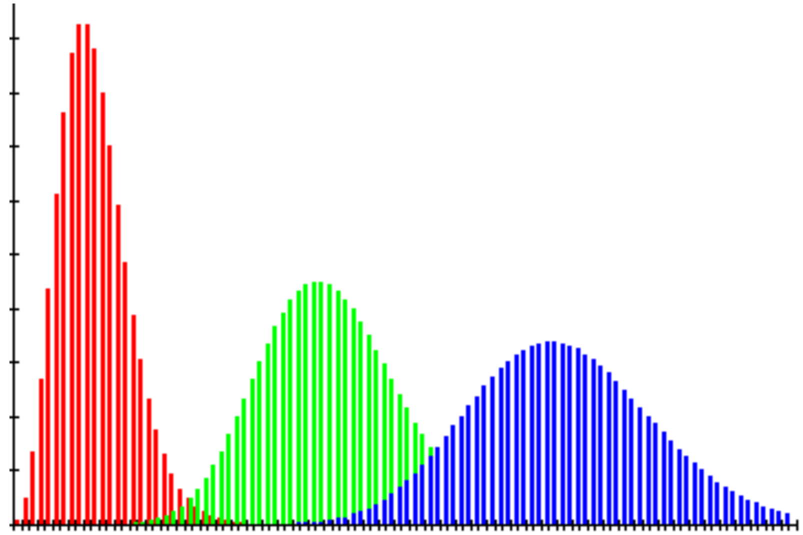
E[g(X,Y)] = 72/24 = 3

E[g(X,Y)] = 2•E[X] - E[Y]

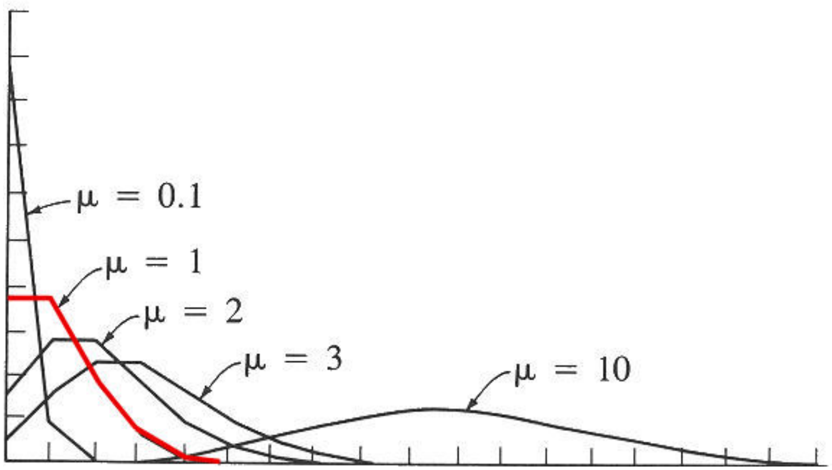
= 2•2.5 - 2 = 3

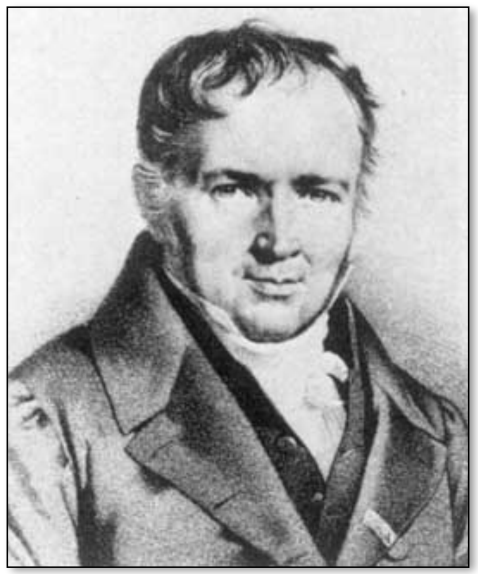
recall both marginals are uniform

X \ Y	1	2	3
1	1 • 4/24	0 • 1/24	-1 • 1/24
2	3 • 0/24	2 • 3/24	1 • 3/24
3	5 • 0/24	4 • 4/24	3 • 2/24
4	7 • 4/24	6 • 0/24	5 • 2/24



a zoo of (discrete) random variables





A discrete random variable X equally likely to take any (integer) value between integers a and b, inclusive, is *uniform*.

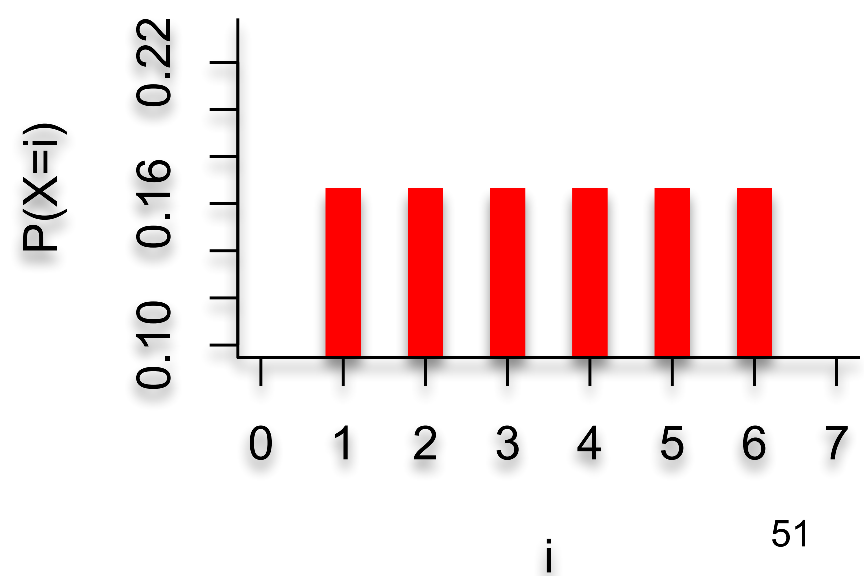
Notation: $X \sim \mathrm{Unif}(a,b)$

Probability: $$P(X = i) = \frac{1}{b - a + 1}$$

Mean, Variance: $$E[X] = \frac{a + b}{2}, \mathrm{Var}[X] = \frac{(b - a)(b - a + 2)}{12}$$

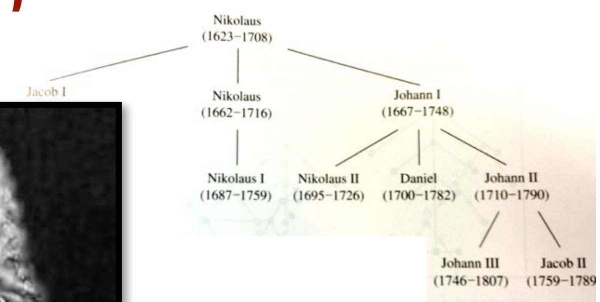
Example: value shown on one roll of a fair die is Unif(1,6):

$P(X=i) = 1/6$
$E[X] = 7/2$
$\mathrm{Var}[X] = 35/12$



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: $X \sim \mathrm{Ber}(p)$
- $E[X] = E[X^2] = p$
- $\mathrm{Var}(X) = E[X^2] - (E[X])^2 = p - p^2 = p(1-p)$



Examples:

- coin flip
- random binary digit
- whether a disk drive crashed

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

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

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

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

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

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

Examples

\# of heads in n coin flips

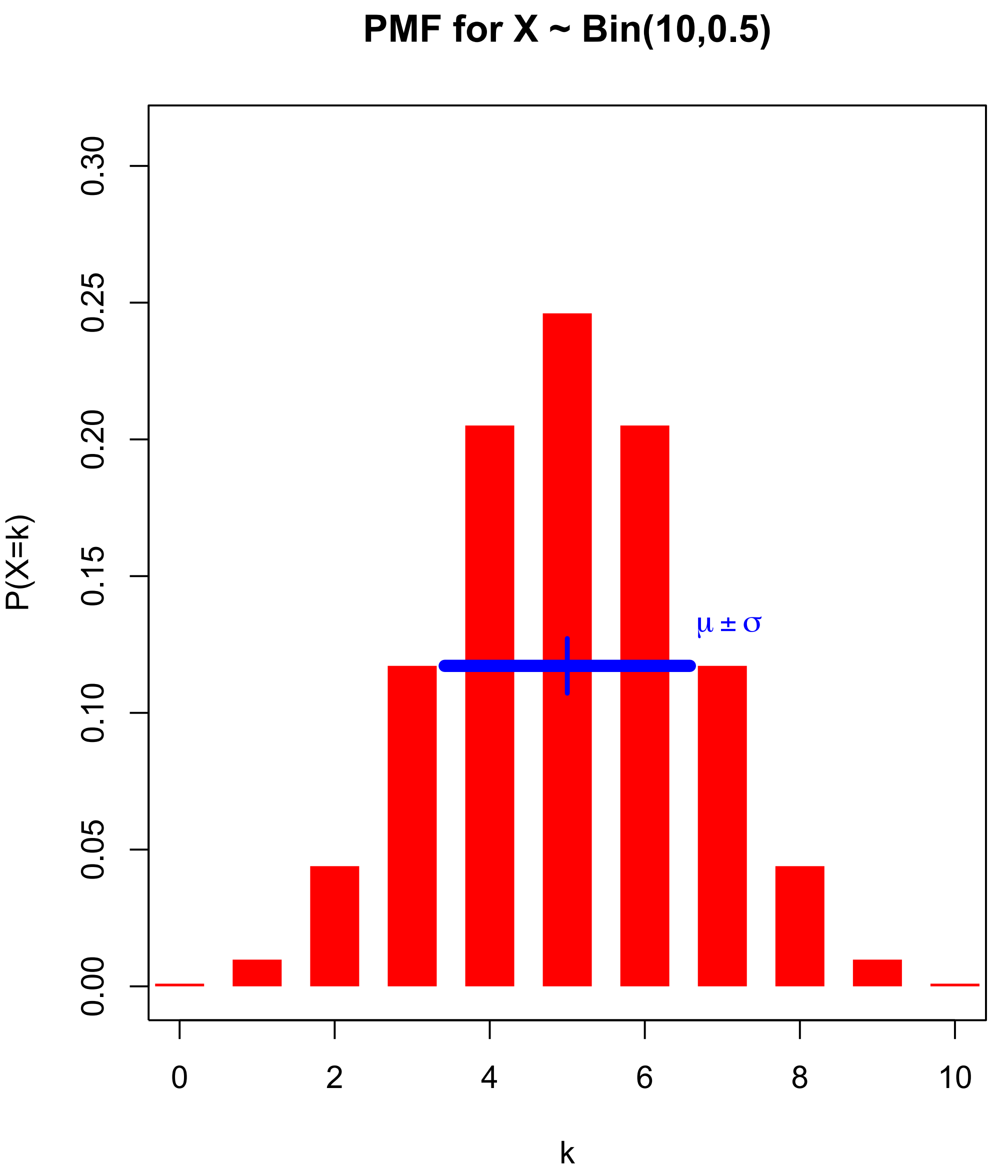
\# of 1's in a randomly generated length n bit string

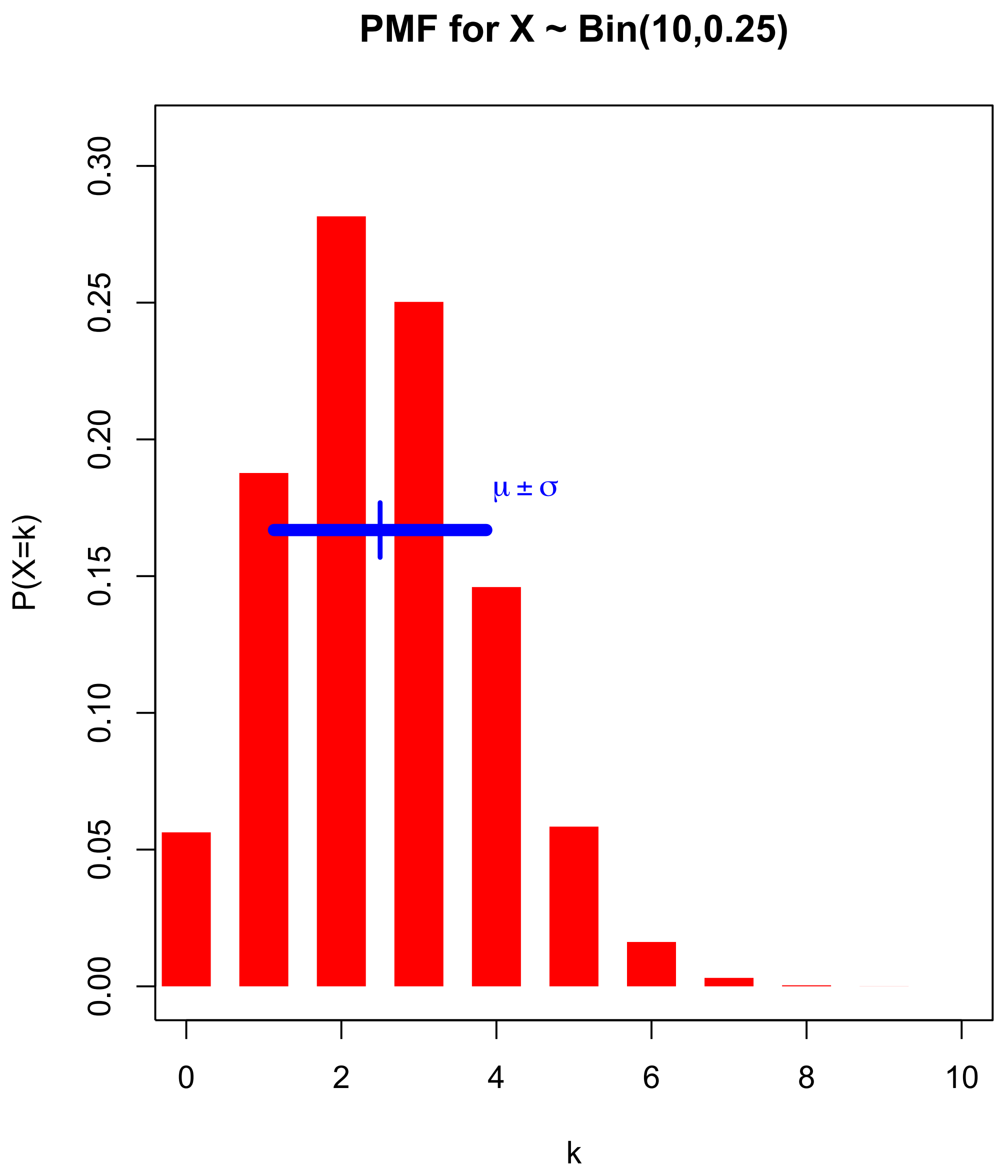
\# of disk drive crashes in a 1000 computer cluster

E[X] = pn

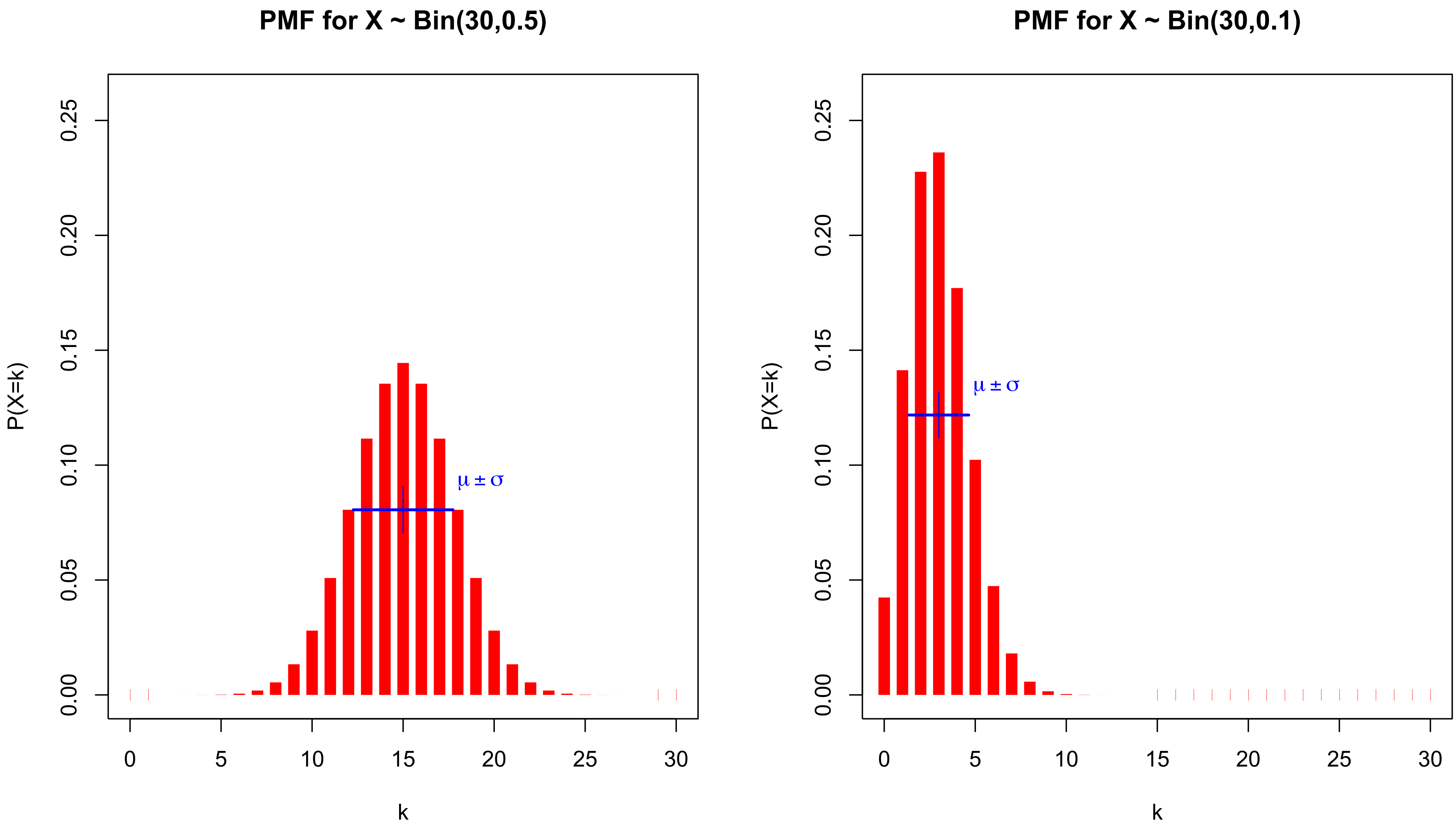
Var(X) = p(1-p)n

←(proof below, twice)


PMF for X ~ Bin(10,0.5)
P(X=k)
0.00
0.05
0.10
0.15
0.20
0.25
0.30
μ ± σ
0
2
4
6
8
10
k


PMF for X ~ Bin(10,0.25)
P(X=k)
0.00
0.05
0.10
0.15
0.20
0.25
0.30
μ ± σ
0
2
4
6
8
10
k

binomial pmfs



mean and variance of the binomial (I)

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

☜ generalizes slide 14

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

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

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

using $j = i - 1$

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

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

where $Y \sim Bin(n-1, p)$

$k = 1$ gives: $\boxed{E[X] = np}$; $k = 2$ gives: $\boxed{E[X^2] = np((n-1)p+1)}$

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

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

$$= np(1-p)$$

Theorem: If X & Y are *independent*, then E[X•Y] = E[X]•E[Y]

any dist, not just binomial

Proof:

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

$$\begin{aligned} E[X \cdot Y] &= \sum_i \sum_j x_i \cdot y_j \cdot P(X = x_i \wedge Y = y_j) \\ &= \sum_i \sum_j x_i \cdot y_j \cdot P(X = x_i) \cdot P(Y = y_j) \\ &= \sum_i x_i \cdot P(X = x_i) \cdot \left(\sum_j y_j \cdot P(Y = y_j) \right) \\ &= E[X] \cdot E[Y] \end{aligned}$$

independence

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

variance of independent r.v.s is additive

(Bienaymé, 1853)

Theorem: If X & Y are *independent*, (any dist, not just binomial) then

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

Proof: Let

$$\begin{aligned} \widehat{X} &= X - E[X] & \widehat{Y} &= Y - E[Y] \\ E[\widehat{X}] &= 0 & E[\widehat{Y}] &= 0 \\ Var[\widehat{X}] &= Var[X] & Var[\widehat{Y}] &= Var[Y] \end{aligned}$$

Var(aX+b) = a²Var(X)

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

previous slide

Q: Why are $\widehat{X}, \widehat{Y}$ independent?
A: See HW

(Bienaymé, 1853)

Theorem: If X & Y are *independent*, (any dist, not just binomial) then

Var[X+Y] = Var[X]+Var[Y]

Alternate Proof:

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

slide 57

If $Y_1, Y_2, \ldots, Y_n \sim \text{Ber}(p)$ and independent,

then $X = \sum_{i=1}^{n} Y_i \sim \text{Bin}(n, p)$.

$$\boxed{E[X] = np}$$

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

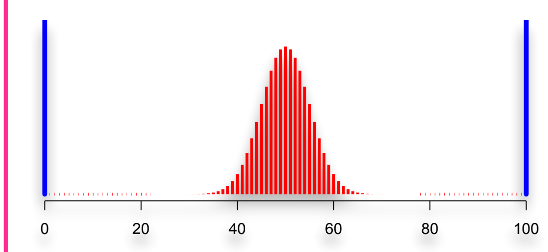
$$\boxed{\text{Var}[X] = np(1-p)}$$

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

If $Y_1, Y_2, \ldots, Y_n \sim \text{Ber}(p)$ and independent,

then $X = \sum_{i=1}^{n} Y_i \sim \text{Bin}(n, p)$.

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



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

Indp random fluctuations tend to cancel when added; dependent ones may reinforce; "nY_7": no such cancelation; much variation

Note :

$$E\left[\sum_{i=1}^{n} Y_i\right] = \sum_{i=1}^{n} E\left[Y_i\right] = nE[Y_7] = E[nY_7]$$

but

Q. Why the big difference? A.

$$\text{Var}\left[\sum_{i=1}^{n} Y_i\right] = \sum_{i=1}^{n} \text{Var}\left[Y_i\right] = n\text{Var}[Y_7] \ll \text{Var}[nY_7] = n^2\text{Var}[Y_7]$$

A RAID-like 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)

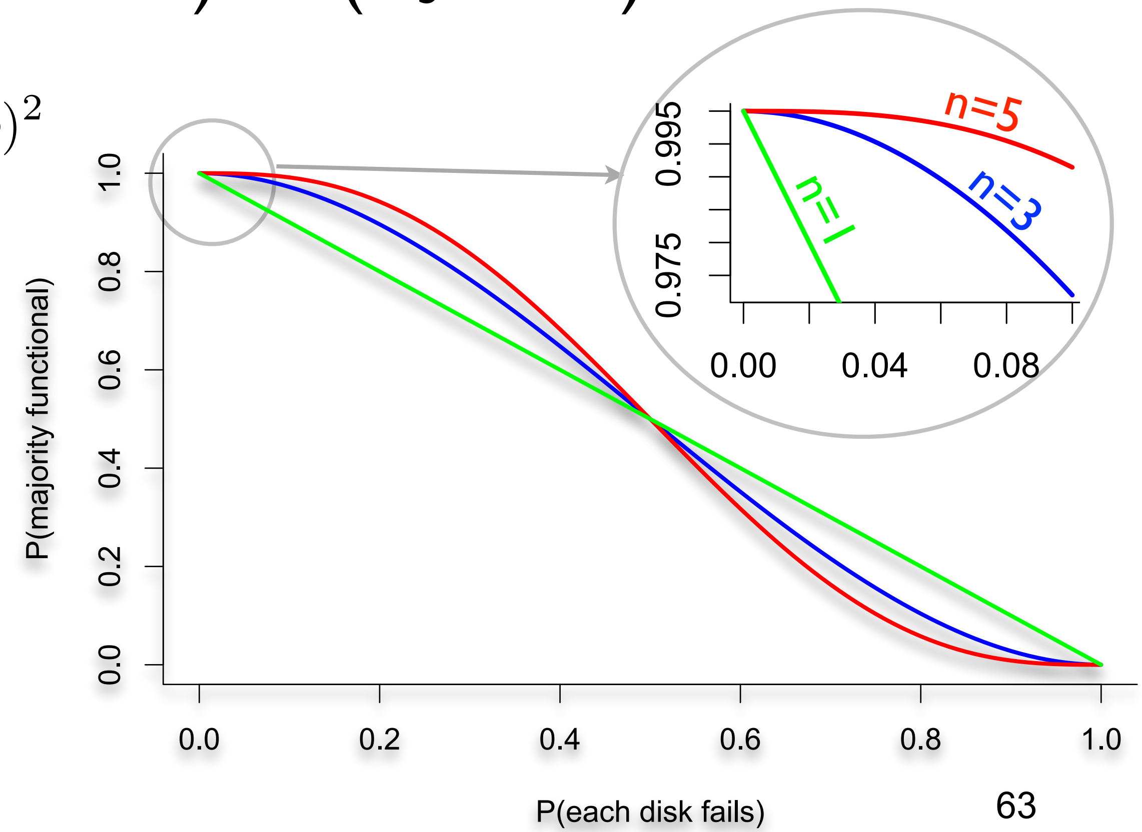
$$\binom{5}{0}p^0(1-p)^5 + \binom{5}{1}p^1(1-p)^4 + \binom{5}{2}p^2(1-p)^3$$

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 iff p < 1/2



Goal: send a 4-bit message over a noisy communication channel.

Say, 1 bit in 10 is flipped in transit, independently.

What is the probability that the message arrives correctly?

Let X = # of errors; X ~ Bin(4, 0.1)

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

$$P(X = 0) = \binom{4}{0}(0.1)^0(0.9)^4 = 0.6561$$

Can we do better? Yes: error correction via redundancy.

E.g., send every bit in triplicate; use majority vote.

Let Y = # of errors in one trio; Y ~ Bin(3, 0.1); P(a trio is OK) =

$$P(Y \leq 1) = \binom{3}{0}(0.1)^0(0.9)^3 + \binom{3}{1}(0.1)^1(0.9)^2 = 0.972$$

If X' = # errors in triplicate msg, X' ~ Bin(4, 0.028), and

$$P(X' = 0) = \binom{4}{0}(0.028)^0(0.972)^4 = 0.8926168$$

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

Z = 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, but not all, multi-bit errors can be detected, but not corrected.)

P(correctable message received) = $P(Z \leq 1)$

Hamming example

"Parity(x,y,z)" is perhaps best defined as (x+y+z+1) mod 2

- I.e., make sure that there are an odd number of one-bits among x,y,z,parity. Why? "Stuck at zero" faults are a common error mode in digital systems, so it's best if the parity check on 000 is 1. I.e., 0001 is OK but 0000 would be recognized as faulty.

Suppose the message you want to send is '1011'

Instead, you send '1011 1 0 1' (via rules on prev slide)

If your partner receives a 1-bit corruption of this, e.g.,

*0*011 <u>1</u> <u>0</u> 1

then both underlined parity bits are incorrect: the quadruples defined above (incl the parity bit) have even parity, but should have odd parity. Studying the rules on the prev slide, this is the *ONLY* single bit corruption displaying this pattern, so you know to "correct" the initial 0 bit to 1, recovering the 1011 message.

Exercise: try all 6 other single bit errors; you should see that each has a distinct pattern of "parity errors," hence is correctable.

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

$$P(Z \leq 1) = \binom{7}{0}(0.1)^0(0.9)^7 + \binom{7}{1}(0.1)^1(0.9)^6 \approx 0.8503$$

Recall, uncorrected success rate is

$$P(X = 0) = \binom{4}{0}(0.1)^0(0.9)^4 = 0.6561$$

And triplicate code success rate is:

$$P(X' = 0) = \binom{4}{0}(0.028)^0(0.972)^4 = 0.8926168$$

Hamming code is nearly as reliable as the triplicate code, with 5/12 ≈ 42% fewer bits. (& better with longer codes; overhead is O(logn) bits for n bit messages.)

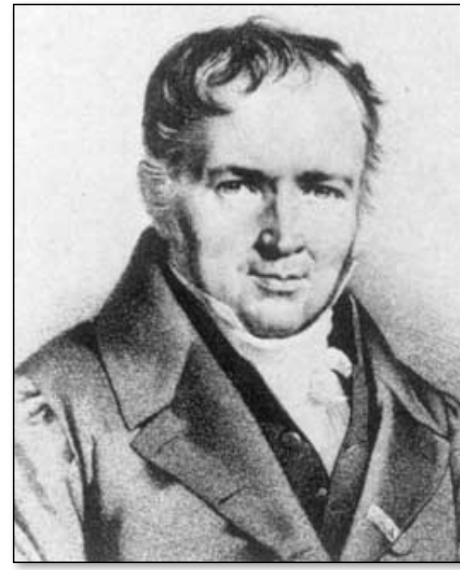
Sending a bit string over the network

- $n = 4$ bits sent, each corrupted with probability 0.1
- X = # of corrupted bits, $X \sim \text{Bin}(4, 0.1)$
- 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

Suppose "events" happen, independently, at an *average* rate of λ per unit time. Let X be the *actual* number of events happening in a given time unit. Then X is a *Poisson* r.v. *with parameter* λ (denoted X ~ Poi(λ)) and has distribution (PMF):



Siméon Poisson, 1781-1840

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

Examples:

\# of alpha particles emitted by a lump of radium in 1 sec.

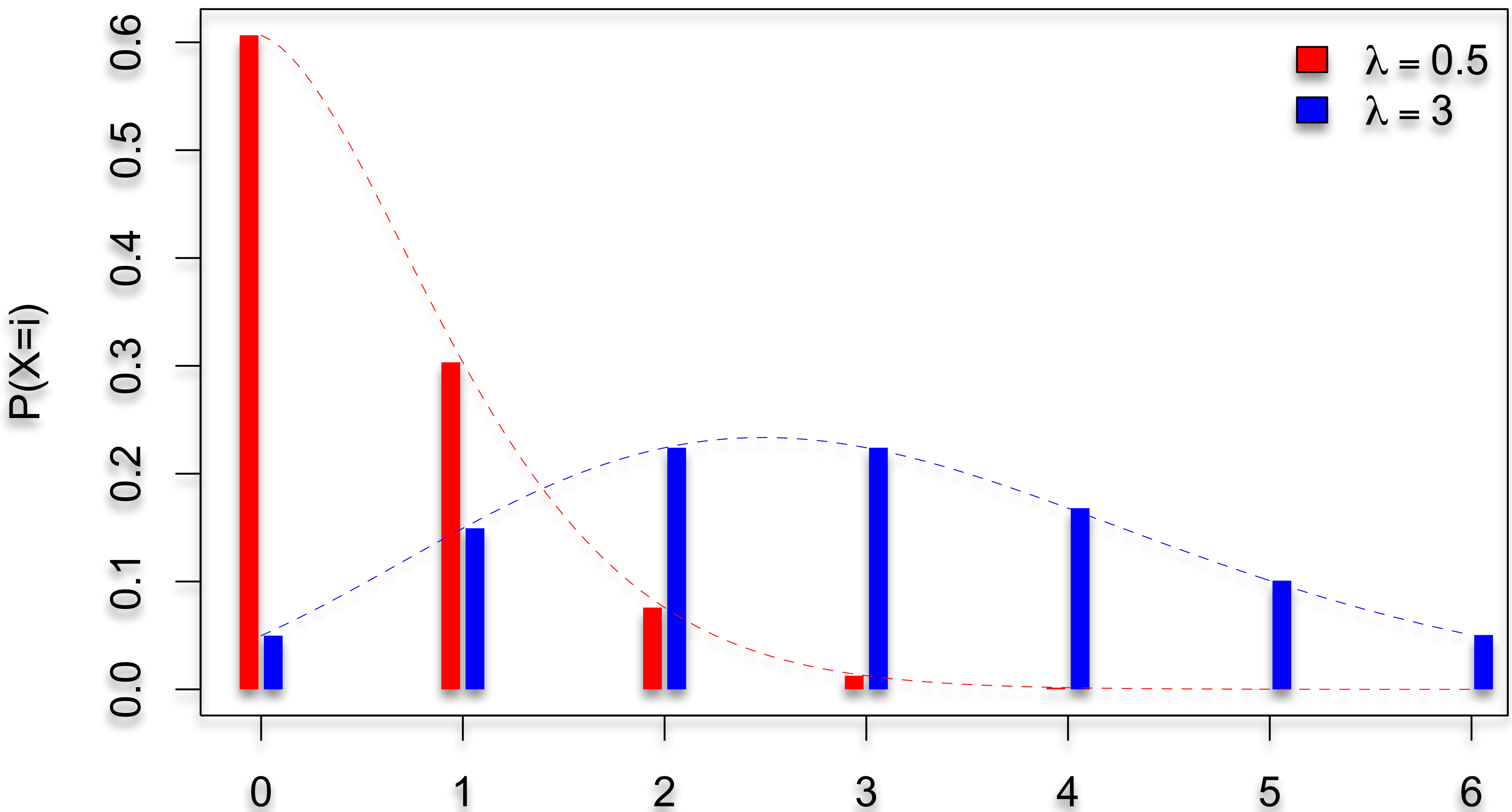
\# of traffic accidents in Seattle in one year

\# of babies born in a day at UW Med center

\# of visitors to my web page today

See B&T Section 6.2 for more on theoretical basis for Poisson.

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



X is a Poisson r.v. with parameter λ if it has PMF:

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

Is it a valid distribution? Recall Taylor series:

$$e^{\lambda} = \frac{\lambda^0}{0!} + \frac{\lambda^1}{1!} + \cdots = \sum_{0 \le i} \frac{\lambda^i}{i!}$$

So

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

$$
\begin{aligned}
E[X] &= \sum_{0 \le i} i \cdot e^{-\lambda} \frac{\lambda^i}{i!} \\
&= \sum_{1 \le i} i \cdot e^{-\lambda} \frac{\lambda^i}{i!} \\
&= \lambda e^{-\lambda} \sum_{1 \le i} \frac{\lambda^{i-1}}{(i-1)!} \\
&= \lambda e^{-\lambda} \sum_{0 \le j} \frac{\lambda^j}{j!} \\
&= \lambda e^{-\lambda} e^{\lambda} \\
&= \lambda
\end{aligned}
$$

i = 0 term is zero

j = i-1

As expected, given definition in terms of "average rate λ"

(Var[X] = λ, too; proof similar, see B&T example 6.20)

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

Different interpretations of "moderate," e.g.

$n > 20$ and $p < 0.05$

$n > 100$ and $p < 0.1$

Formally, Binomial is Poisson in the limit as $n \to \infty$ (equivalently, $p \to 0$) while holding $np = \lambda$

X ~ Binomial(n,p)

$$
\begin{aligned}
P(X=i) &= \binom{n}{i} p^i (1-p)^{n-i} \\
&= \frac{n!}{i!(n-i)!} \left(\frac{\lambda}{n}\right)^i \left(1-\frac{\lambda}{n}\right)^{n-i}, \text{ where } \lambda = pn \\
&= \frac{n(n-1)\cdots(n-i+1)}{n^i} \frac{\lambda^i}{i!} \frac{(1-\lambda/n)^n}{(1-\lambda/n)^i} \\
&= \underbrace{\frac{n(n-1)\cdots(n-i+1)}{(n-\lambda)^i}} \frac{\lambda^i}{i!} \underbrace{(1-\lambda/n)^n} \\
&\approx \quad 1 \quad \cdot \frac{\lambda^i}{i!} \cdot \quad e^{-\lambda}
\end{aligned}
$$

I.e., Binomial ≈ Poisson for large n, small p, moderate i, λ.

Handy: Poisson has only 1 parameter–the expected # of successes

Recall example of sending bit string over a network

Send bit string of length $n = 10^4$

Probability of (independent) bit corruption is $p = 10^{-6}$

$X \sim \text{Poi}(\lambda = 10^4 \cdot 10^{-6} = 0.01)$

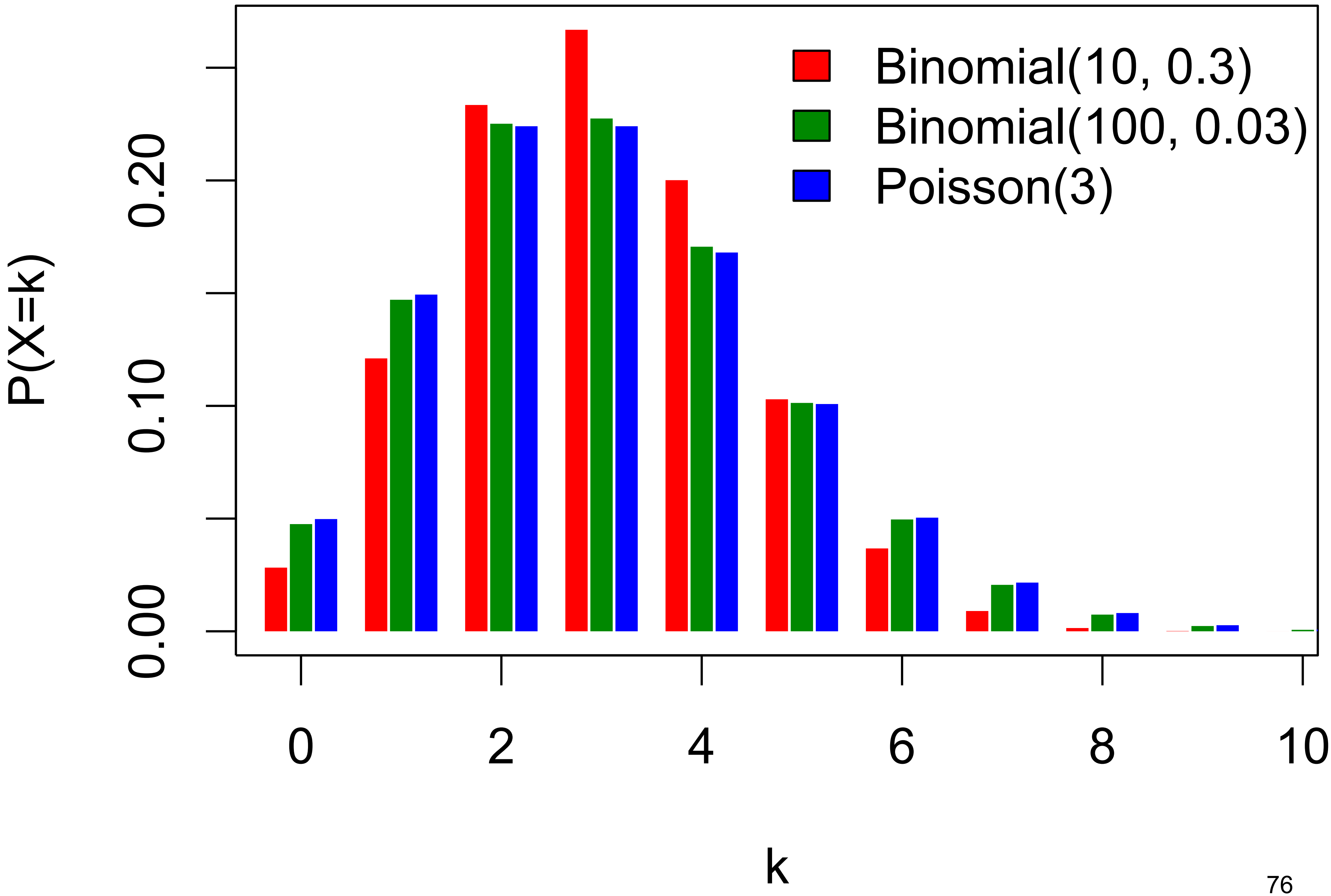
What is probability that message arrives uncorrupted?

$$P(X=0) = e^{-\lambda}\frac{\lambda^0}{0!} = e^{-0.01}\frac{0.01^0}{0!} \approx 0.9900498\underline{34}$$

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

$P(Y=0) \approx 0.9900498\underline{29}$

I.e., Poisson approximation (here) is accurate to ~5 parts per billion


Binomial(10, 0.3)
Binomial(100, 0.03)
Poisson(3)
P(X=k)
0.00
0.10
0.20
0
2
4
6
8
10
k

Recall: if $Y \sim \text{Bin}(n,p)$, then:

$E[Y] = pn$

$\text{Var}[Y] = np(1-p)$

And if $X \sim \text{Poi}(\lambda)$ where $\lambda = np$ ($n \to \infty$, $p \to 0$) then

$E[X] = \lambda = np = E[Y]$

$\text{Var}[X] = \lambda \approx \lambda(1-\lambda/n) = np(1-p) = \text{Var}[Y]$

Expectation and variance of Poisson are the same (λ)

Expectation is the same as corresponding binomial

Variance almost the same as corresponding binomial

Note: when two different distributions share the same mean & variance, it suggests (but doesn't prove) that one may be a good approximation for the other.

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.

$P(X > b) < 1/(24*60*60) \approx 1.2 \times 10^{-5}$

$P(X = b) = e^{-1}/b!$ $\quad \Sigma_{i\geq 8} P(X=i) \approx P(X=8) \approx 10^{-5}$, so $b \approx 8$

Above necessary but not sufficient; also check prob of 10 arrivals in 2 seconds, 12 in 3, etc. See BT p366 for a possible approach to fully solving it.

In a series $X_1, X_2, \ldots$ of Bernoulli trials with success probability p, let Y be the index of the first success, i.e.,

$$X_1 = X_2 = \ldots = X_{Y-1} = 0 \ \& \ X_Y = 1$$

Then Y is a *geometric* random variable with parameter p.

Examples:

Number of coin flips until first head

Number of blind guesses on LSAT until I get one right

Number of darts thrown until you hit a bullseye

Number of random probes into hash table until empty slot

Number of wild guesses at a password until you hit it

$P(Y=k) = (1-p)^{k-1}p$; Mean $1/p$; Variance $(1-p)/p^2$

☝ see slide 13; see also slide 83, BT p105 for slick alt. proof

Recall: conditional probability

$$P(X \mid A) = P(X \,\&\, A)/P(A)$$

A note about notation: When X is an r.v., take this as either shorthand for "$\forall x$ P(X=x ..." or as defining the conditional PMF p(x|A) from the joint PMF

Conditional probability is a probability, i.e.

1. it's nonnegative
2. it's normalized
3. it's happy with the axioms, etc.

Define: The *conditional expectation* of X

$$E[X \mid A] = \sum_x x \cdot p(X = x \mid A)$$

I.e., the value of r.v. X averaged over outcomes *where I know event A happened*

Recall: the law of total probability

$$p(X) = p(X \mid A) \bullet P(A) + p(X \mid A^c) \bullet P(A^c)$$

I.e., unconditional probability is the weighted average of conditional probabilities, weighted by the probabilities of the conditioning events

Again, "$\forall x$ P(X=x ..." or "unconditional PMF is weighted avg of conditional PMFs"

The Law of Total Expectation

$$E[X] = E[X \mid A] \bullet P(A) + E[X \mid A^c] \bullet P(A^c)$$

I.e., unconditional expectation is the weighted average of conditional expectations, weighted by the probabilities of the conditioning events

Proof of the Law of Total Expectation:

$$
\begin{aligned}
E[X] &= \sum_x x p(x) \\
&= \sum_x x\left(p(x \mid A)P(A) + p(x \mid \overline{A})P(\overline{A})\right) \\
&= \sum_x x p(x \mid A)P(A) + \sum_x x p(x \mid \overline{A})P(\overline{A}) \\
&= \left(\sum_x x p(x \mid A)\right) P(A) + \left(\sum_x x p(x \mid \overline{A})\right) P(\overline{A}) \\
&= E[X \mid A]P(A) + E[X \mid \overline{A}]P(\overline{A})
\end{aligned}
$$

$X \sim \text{geo}(p)$

$$E[X] = E[X \mid X=1] \cdot P(X=1) + E[X \mid X>1] \cdot P(X>1)$$

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

$$\vdots$$

simple algebra

memorylessness: after flipping one tail, *remaining* waiting time until 1st head is exactly the same as starting from scratch

$$E[X] = 1/p$$

E.g., if p=1/2, expect to wait 2 flips for 1st head;
p=1/10, expect to wait 10 flips.

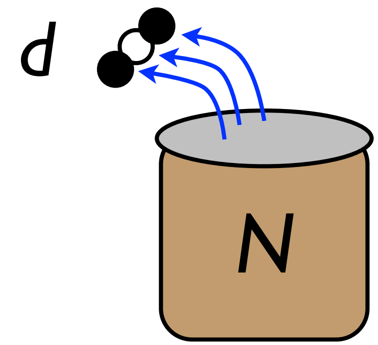
(Similar derivation for variance: $(1-p)/p^2$)

cf. slide 79

balls in urns – the hypergeometric distribution

B&T, exercise 1.61

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_j be 0/1 indicator for j-th ball is white, $X = \Sigma X_j$

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

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

like binomial (almost)

$N \approx 22500$ 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

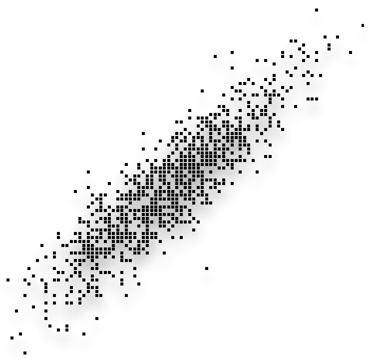
GO Categories Enriched in Genes Associated with Myotube-Increased Peaks

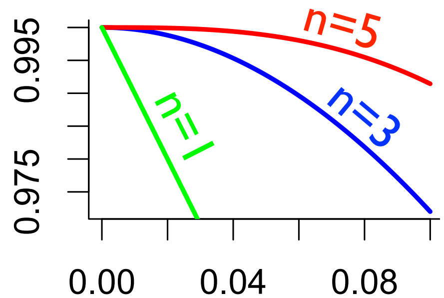
GOID	Term	P Value	OR[a]	Count[b]	Size[c]	Ont[d]
GO:0005856	cytoskeleton	2.05E-11	2.40	94	490	CC
GO:0043292	contractile fiber	6.98E-09	5.85	22	58	CC
GO:0030016	myofibril	1.96E-08	5.74	21	56	CC
GO:0044449	contractile fiber part	2.58E-08	5.97	20	52	CC
GO:0030017	sarcomere	4.95E-08	6.04	19	49	CC
GO:0008092	[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.

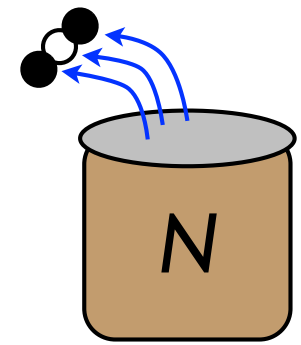
E.g., if you draw 1588 balls from an urn containing 490 white balls and ≈22000 black balls, P(94 white) $\approx 2.05\times 10^{-11}$

So, are genes flagged by this experiment specifically related to muscle development? This doesn't prove that they are, but it does say that there is an exceedingly small probability that so many would cluster in the "muscle development" group purely by chance.





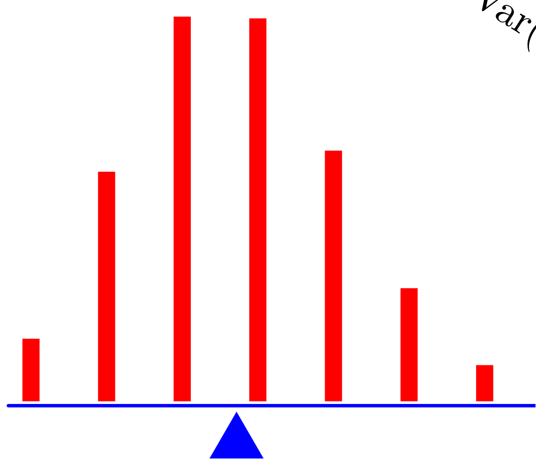
$$E[X+Y] = E[X] + E[Y]$$



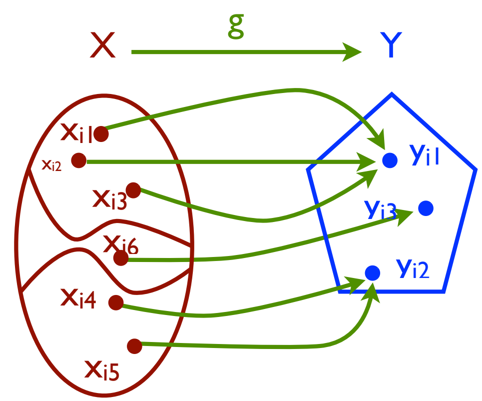
$$P(X=i) = e^{-\lambda}\frac{\lambda^i}{i!}$$

$$\sum_{i=-\infty}^{\infty} =$$

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



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



RV: a numeric function of the outcome of an experiment

Probability Mass Function $p(x)$: prob that RV $= x$; $\Sigma p(x)=1$

Cumulative Distribution Function $F(x)$: *probability that RV* $\leq x$

Generalize to *joint* distributions; *independence* & *marginals*

Expectation:

mean, average, "center of mass," fair price for a game of chance

of a random variable: $E[X] = \Sigma_x \, xp(x)$

of a function: if $Y = g(X)$, then $E[Y] = \Sigma_x \, g(x)p(x)$

(probability)-weighted average

linearity:

$E[aX + b] = aE[X] + b$

$E[X+Y] = E[X] + E[Y]$; even if *dependent*

this interchange of "order of operations" is quite special to linear combinations. E.g., $E[XY] \neq E[X] \bullet E[Y]$*, in general (but see below)*

Conditional Expectation:

$E[X \mid A] = \sum_x x \cdot P(X=x \mid A)$

Law of Total Expectation

$E[X] = E[X \mid A] \cdot P(A) + E[X \mid \neg A] \cdot P(\neg A)$

Variance:

$Var[X] = E[\,(X-E[X])^2\,] = E[X^2] - (E[X])^2]$

Standard deviation: $\sigma = \sqrt{\mathrm{Var}[X]}$

$\mathrm{Var}[aX+b] = a^2\,\mathrm{Var}[X]$ "Variance is insensitive to location, quadratic in scale"

If X & Y are *independent*, then

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

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

(These two equalities hold for *indp* rv's; but not in general.)

Important Examples:

Uniform(a,b): $P(X = i) = \frac{1}{b-a+1}$ $\mu = \frac{a+b}{2}, \sigma^2 = \frac{(b-a)(b-a+2)}{12}$

Bernoulli: $P(X = 1) = p, P(X = 0) = 1\text{-}p$ $\mu = p, \quad \sigma^2 = p(1\text{-}p)$

Binomial: $P(X = i) = \binom{n}{i} p^i (1-p)^{n-i}$ $\mu = np, \sigma^2 = np(1\text{-}p)$

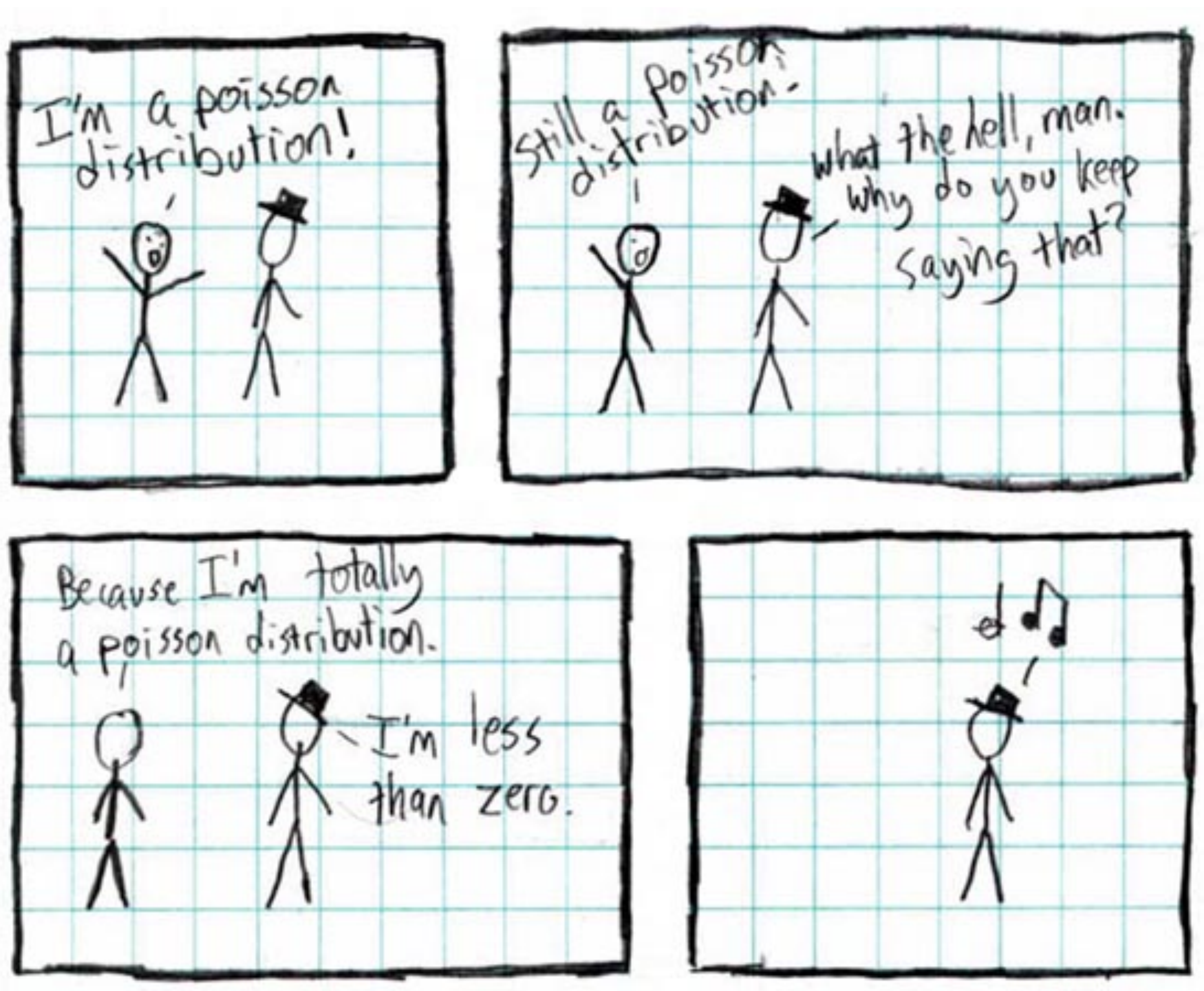
Poisson: $P(X = i) = e^{-\lambda} \frac{\lambda^i}{i!}$ $\mu = \lambda, \quad \sigma^2 = \lambda$

$Bin(n,p) \approx Poi(\lambda)$ where $\lambda = np$ fixed, $n \to \infty$ (and so $p = \lambda/n \to 0$)

Geometric $P(X = k) = (1\text{-}p)^{k\text{-}1} p$ $\mu = 1/p, \sigma^2 = (1\text{-}p)/p^2$

Many others, e.g., hypergeometric, negative binomial, …

Poisson distributions have no value over negative numbers



http://xkcd.com/12/