Chapter 5
Variance

Higher moments

Defn: The **$\boldsymbol{k}$th moment** of r.v. X is

$$\boldsymbol{E}[X^k] = \sum_x x^k \cdot \boldsymbol{P}\{X = x\}$$

Can we say

$$\boldsymbol{E}[X^2] = \boldsymbol{E}[X] \cdot \boldsymbol{E}[X]?$$

This doesn’t work because X is not independent of X.

Example:

$X \sim Geometric(p)$.

Derive $\boldsymbol{E}[X^2]$.

Higher moments

Defn: The ***k*th moment** of r.v. X is

$$\boldsymbol{E}[X^k] = \sum_x x^k \cdot \boldsymbol{P}\{X = x\}$$

Example:

$X \sim Geometric(p)$.

Derive $\boldsymbol{E}[X^2]$.

$$\boldsymbol{E}[X^2] = \sum_{i=1}^{\infty} i^2 p_X(i)$$

$$= \sum_{i=1}^{\infty} i^2 (1-p)^{i-1} \cdot p$$

Not obvious how to compute this sum

2^{nd} Moment of Geometric

$X \sim Geometric(p)$.

Let's try conditioning

Derive $\boldsymbol{E}[X^2]$. Condition on value of 1st flip, Y.

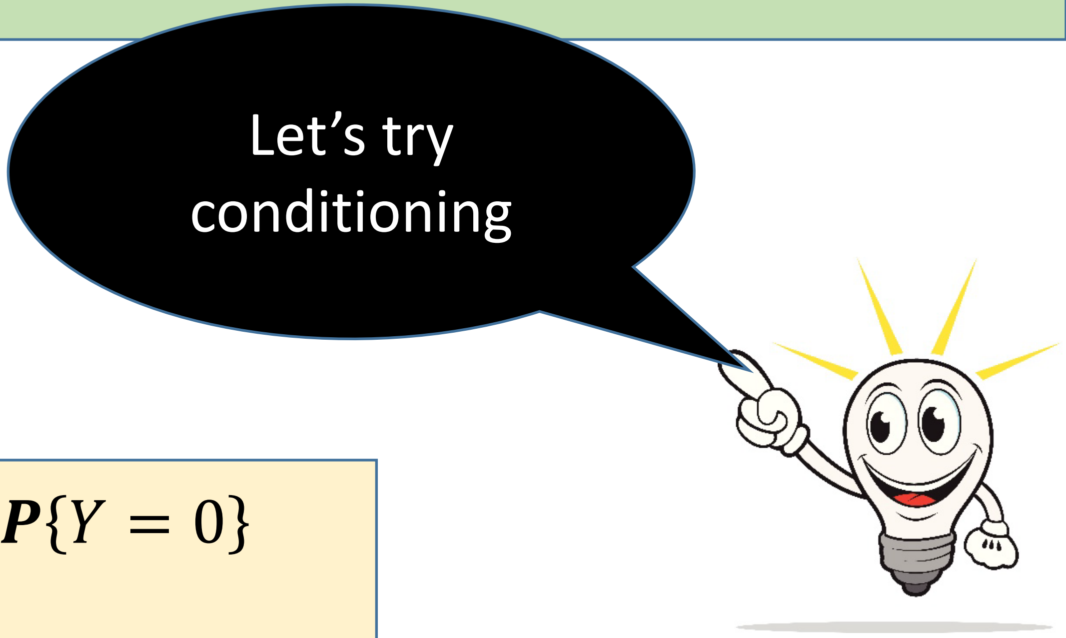
What is this?

$$\boldsymbol{E}[X^2] = \boldsymbol{E}[X^2 \mid Y = 1] \cdot \boldsymbol{P}\{Y = 1\} + \boldsymbol{E}[X^2 \mid Y = 0] \cdot \boldsymbol{P}\{Y = 0\}$$

$$= 1 \cdot p + \boxed{\boldsymbol{E}[X^2 \mid Y = 0]} \cdot (1 - p)$$

2^{nd} Moment of Geometric

$X \sim Geometric(p)$.



Derive $\boldsymbol{E}[X^2]$. Condition on value of 1st flip, Y.

$$\boldsymbol{E}[X^2] = \boldsymbol{E}[X^2 \mid Y = 1] \cdot \boldsymbol{P}\{Y = 1\} + \boldsymbol{E}[X^2 \mid Y = 0] \cdot \boldsymbol{P}\{Y = 0\}$$

$$= 1 \cdot p + \boldsymbol{E}[X^2 \mid Y = 0] \cdot (1 - p)$$

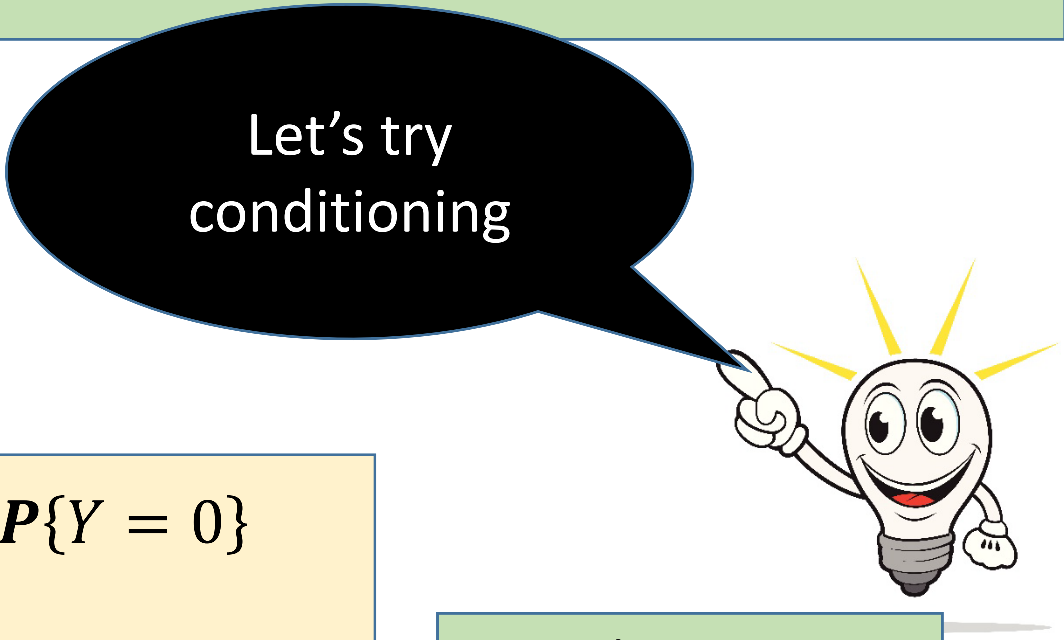
$$[X \mid Y = 0] = X + 1$$

$$[X^2 \mid Y = 0] = (X + 1)^2$$

$$= 1 \cdot p + \boldsymbol{E}[(1 + X)^2] \cdot (1 - p)$$

2^{nd} Moment of Geometric

$X \sim Geometric(p)$. 



Derive $\boldsymbol{E}[X^2]$. Condition on value of 1st flip, Y.

$$\boldsymbol{E}[X^2] = \boldsymbol{E}[X^2 \,|Y = 1] \cdot \boldsymbol{P}\{Y = 1\} + \boldsymbol{E}[X^2 \,|Y = 0] \cdot \boldsymbol{P}\{Y = 0\}$$

$$= 1 \cdot p + \boldsymbol{E}[X^2 \,|Y = 0] \cdot (1 - p)$$

$$= 1 \cdot p + \boldsymbol{E}[(1 + X)^2] \cdot (1 - p)$$

$$= 1 \cdot p + \boldsymbol{E}[1 + 2X + X^2] \cdot (1 - p)$$

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

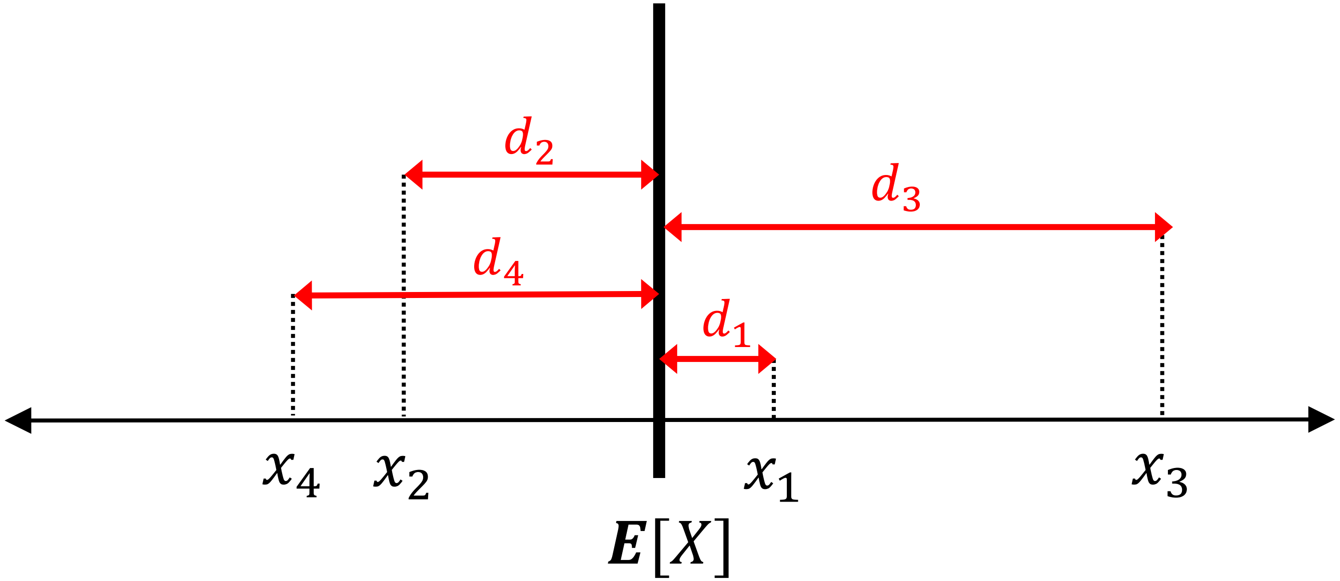
Result:

$$\boldsymbol{E}[X^2] = \frac{2 - p}{p^2}$$

Variance

Defn: The **variance** of r.v. X is the expected squared difference of X from its mean:

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



What is Var(-X)?

What is Var(2 + X)?

$$\boldsymbol{Var}(X) = \frac{d_1^2 + d_2^2 + d_3^2 + d_4^2}{4}$$

Choosing between Microsoft and a Startup

Work at Microsoft ➔ Earnings $= 10^5$

Work at Startup $= \begin{cases} 10^7 & \text{w.p. } 1\% \\ 0 & \text{w.p. } 99\% \end{cases}$

Determine the mean and variance in each case.

$\boldsymbol{E}[\text{Money at Microsoft}] = 10^5$

$\boldsymbol{Var}(\text{Money at Microsoft}) = \boxed{0}$

$\boldsymbol{E}[\text{Money at Startup }] = 10^5$

$\boldsymbol{Var}(\text{Money at Startup})$

$= \boldsymbol{E}[(\text{Money} - 10^5)^2]$

$= (10^7 - 10^5)^2 \cdot 0.01 + (0 - 10^5)^2 \cdot 0.99$

$\approx 10^{14} \cdot 0.01 + 10^{10} \cdot 0.99 \quad \approx \boxed{10^{12}}$

Variance of Bernoulli(p)

X = value of the coin flip $= \begin{cases} 1 & \text{w.p. } p \\ 0 & \text{o.w.} \end{cases}$

Probability p of heads

Recall: $\boldsymbol{E}[X] = p$

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

$$= \boldsymbol{E}[X^2 - 2Xp + p^2]$$

$$= \boldsymbol{E}[X^2] - 2p\boldsymbol{E}[X] + p^2$$

$$= p \cdot 1^2 - 2p \cdot p + p^2$$

$$= p - p^2 = p(1-p)$$

Remember! Variance of Bernoulli(p) is $p(1-p)$.

Conditioning on Variance is NOT allowed

X = value of the coin flip $= \begin{cases} 1 & \text{w.p. } p \\ 0 & \text{o.w.} \end{cases}$

Probability p of heads

Recall: $\boldsymbol{E}[X] = p$

$$\begin{aligned} \boldsymbol{Var}(X) &= \boldsymbol{Var}(X \mid X = 1) \cdot p + \boldsymbol{Var}(X \mid X = 0) \cdot (1 - p) \\ &= 0 \cdot p + 0 \cdot (1 - p) \\ &= 0 \end{aligned}$$

Alternative definitions of variance?

Potential new defn:

$$\boldsymbol{E}[X - \boldsymbol{E}[X]]$$

What's wrong with this?

Potential new defn:

$$\boldsymbol{E}[|X - \boldsymbol{E}[X]|]$$

Legitimate, but lacking linearity property, coming soon.

Potential new defn:

$$\sqrt{\boldsymbol{E}[(X - \boldsymbol{E}[X])^2]}$$

This has a name! std(X)

Standard deviation of X

Defn: The **standard deviation** of a r.v. X is:

$$\sigma_X = \boldsymbol{std}(X) = \sqrt{\boldsymbol{E}[(X - \boldsymbol{E}[X])^2]}$$

We often write:

$$\boldsymbol{Var}(X) = \sigma_X^2$$

The need for a different variation metric

Suppose we measure a quantity, first in cm (r.v. X) and then in mm (r.v. Y):

$$X = \begin{cases} 1 & \text{w.p. } \frac{1}{3} \\ 2 & \text{w.p. } \frac{1}{3} \\ 3 & \text{w.p. } \frac{1}{3} \end{cases} \qquad Y = \begin{cases} 10 & \text{w.p. } \frac{1}{3} \\ 20 & \text{w.p. } \frac{1}{3} \\ 30 & \text{w.p. } \frac{1}{3} \end{cases}$$

Feels like they should have same variance, since they're the same quantity, but they don't:

$$\boxed{\boldsymbol{Var}(X) = \frac{2}{3}} \qquad \boxed{\boldsymbol{Var}(Y) = \frac{200}{3}}$$

Need a new metric!

Squared coefficient of variation

Defn 5.6: The **squared coefficient of variation** of a r.v. X is:

$$C_X^2 = \frac{\mathbf{Var}(X)}{\mathbf{E}[X]^2}$$

The coeff of variation is popular because it's scale invariant!

$$X = \begin{cases} 1 & \text{w.p. } \frac{1}{3} \\ 2 & \text{w.p. } \frac{1}{3} \\ 3 & \text{w.p. } \frac{1}{3} \end{cases}$$

$$Y = \begin{cases} 10 & \text{w.p. } \frac{1}{3} \\ 20 & \text{w.p. } \frac{1}{3} \\ 30 & \text{w.p. } \frac{1}{3} \end{cases}$$

$\mathbf{E}[X] = 2 \qquad \mathbf{Var}(X) = \frac{2}{3}$

$$\boxed{C_X^2 = \frac{1}{6}}$$

$\mathbf{E}[Y] = 20 \qquad \mathbf{Var}(Y) = \frac{200}{3}$

$$\boxed{C_Y^2 = \frac{1}{6}}$$

Equivalent definition of variance

Theorem 5.7:

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

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

$$= \boldsymbol{E}[X^2 - 2X\boldsymbol{E}[X] + \boldsymbol{E}[X]^2]$$

$$= \boldsymbol{E}[X^2] - 2\,\boldsymbol{E}[X]\boldsymbol{E}[X] + \boldsymbol{E}[X]^2$$

$$= \boldsymbol{E}[X^2] - \boldsymbol{E}[X]^2$$

Linearity of Variance

Theorem 5.8: Let X and Y be random variables where $X \perp Y$. Then

$$\boldsymbol{Var}(X+Y) = \boldsymbol{Var}(X) + \boldsymbol{Var}(Y)$$

$$\boldsymbol{Var}(X) = \boldsymbol{E}[(X+Y)^2] - \boldsymbol{E}[X+Y]^2$$

$$= \boldsymbol{E}[X^2] + \boldsymbol{E}[Y^2] + 2\boldsymbol{E}[XY] - \boldsymbol{E}[X]^2 - \boldsymbol{E}[Y]^2 - 2\,\boldsymbol{E}[X]\boldsymbol{E}[Y]$$

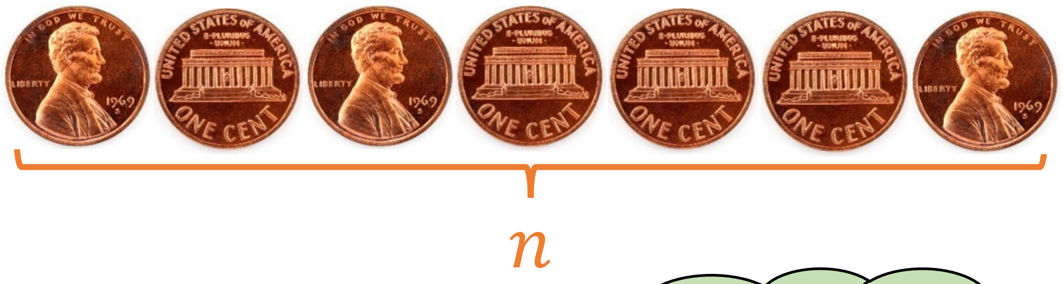
$$= \boldsymbol{Var}(X) + \boldsymbol{Var}(Y) + \underbrace{2\boldsymbol{E}[XY] - 2\boldsymbol{E}[X]\boldsymbol{E}[Y]}_{0}$$

$$= \boldsymbol{Var}(X) + \boldsymbol{Var}(Y)$$

Where did we use $X \perp Y$?

Variance of Binomial(n, p)

Experiment: Flip a coin, with probability p of Heads, n times



Random Variable X = number of heads

Key Observation:

What is $\boldsymbol{E}[X_i]$?
What is $\boldsymbol{Var}(X_i)$?

$$X = X_1 + X_2 + \cdots + X_n \text{ , where } X_i \sim \text{Bernoulli}(p)$$

Applying Linearity of Variance:

$$\boldsymbol{Var}(X) = \boldsymbol{Var}(X_1) + \boldsymbol{Var}(X_2) + \cdots + \boldsymbol{Var}(X_n)$$

$$= p(1-p) + p(1-p) + \cdots + p(1-p) = np(1-p)$$

Remember! Variance of Binomial(n, p) is $np(1-p)$.

Sums versus copies

Let X_1 and X_2 be independent and identically distributed (i.i.d.) random variables, where $X_1 \sim X_2 \sim X$.

$Y = X_1 + X_2$ versus $Z = 2X$

How do $\boldsymbol{E}[Y]$ and $\boldsymbol{E}[Z]$ compare?

$$\boldsymbol{E}[\mathrm{Y}] = \mathbf{E}[\mathrm{Z}] = 2\mathbf{E}[\mathrm{X}]$$

How do $\boldsymbol{Var}(Y)$ and $\boldsymbol{Var}(Z)$ compare?

$$\boldsymbol{Var}(Y) = \boldsymbol{Var}(X_1) + \boldsymbol{Var}(X_2) = 2\,\boldsymbol{Var}(X)$$

$$\boldsymbol{Var}(Z) = 4\,\boldsymbol{Var}(X)$$

Why does Z yield more variance?

Covariance

<u>Defn 5.11</u>: The **covariance** of two random variables X and Y is:

$$\boldsymbol{Cov}(X, Y) = \boldsymbol{E}[(X - \boldsymbol{E}[X]) \cdot (Y - \boldsymbol{E}[Y])]$$

<u>Intuition</u>: If the large values of X tend to happen with the large values of Y, and the small values of X tend to happen with the small values of Y, then $(X - \boldsymbol{E}[X]) \cdot (Y - \boldsymbol{E}[Y])$ is positive on average, so $\boldsymbol{Cov}(X, Y) > 0$, and we say that X and Y are **positively correlated**.

Likewise if $\boldsymbol{Cov}(X, Y) < 0$, we say that X and Y are **negatively correlated**.

Theorem 5.12: $\boldsymbol{Cov}(X, Y) = \boldsymbol{E}[XY] - \boldsymbol{E}[X]\boldsymbol{E}[Y]$

Central moments

<u>Defn 5.13</u>: The ***k*th central moment** of r.v. X is

$$\mathbf{E}[(X-\mathbf{E}[X])^k] = \sum_x (x-\mathbf{E}[X])^k \cdot \mathbf{P}\{X=x\}$$

Q: What is the 2^{nd} central moment?

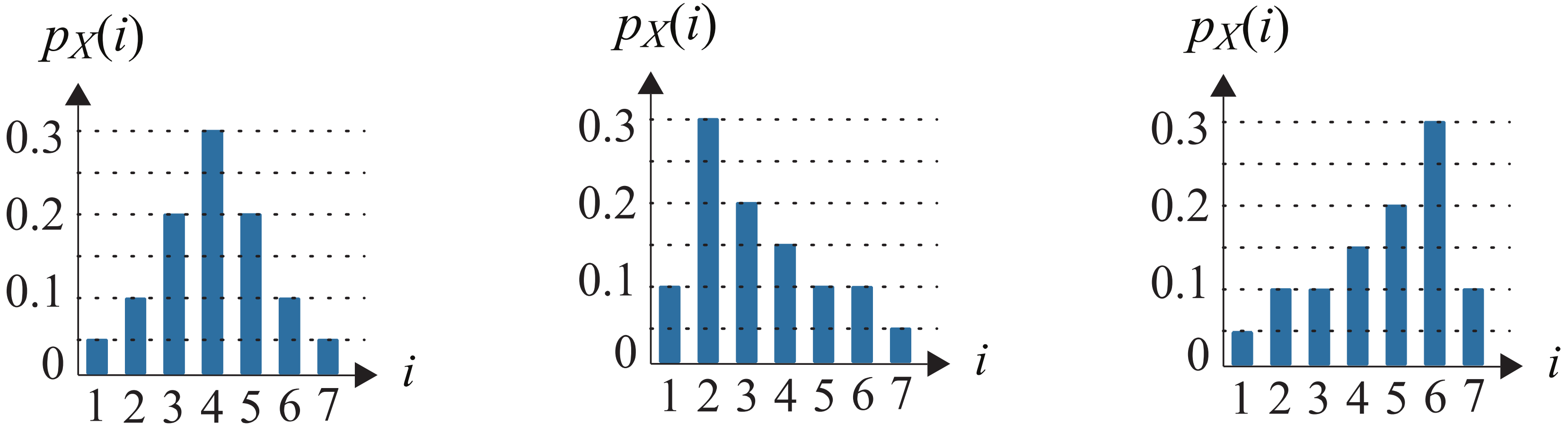
The 2^{nd} central moment is the variance, representing how much the distribution varies from its mean.

Q: What's the difference between the 2^{nd} and 4^{th} central moments?

The 4^{th} central moment is similar to variance, but outliers (those far from the mean) count a lot more!

Third central moment and skew

The **3rd central moment** of r.v. X is $\boldsymbol{E}[(X - \boldsymbol{E}[X])^3]$.
Roughly, the 3rd moment captures the **skew** of the distribution.



(a) (b) (c)

- Zero skew
- Zero 3rd central moment

- Positive skew
- Positive 3rd central moment

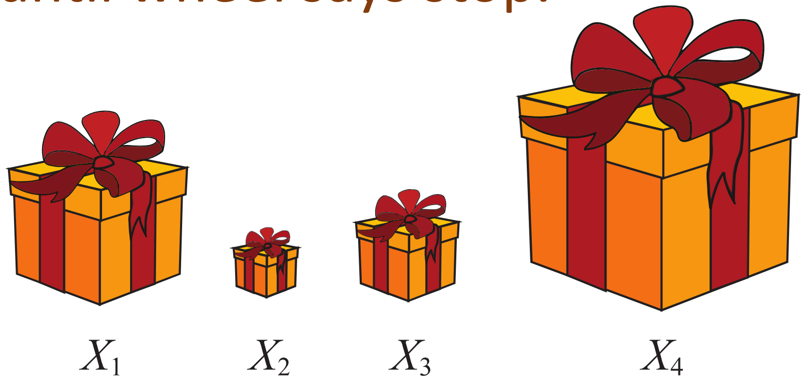
- Negative skew
- Negative 3rd central moment

Sum of random number of random variables

In many applications, we need to add a number of i.i.d. r.v.s, where the total number of r.v.s added is itself a r.v.

$$S = \sum_{i=1}^{N} X_i \qquad N \perp \{X_1, X_2, X_3, \dots\}$$

Get new prize every day, until wheel says stop.



$$\text{Total earnings} = \sum_{i=1}^{N} X_i$$

where $N \sim Geometric\left(\frac{1}{6}\right)$

Sum of random number of random variables

Let $X_1, X_2, X_3, \ldots$ be i.i.d. r.v.s, where $X_i \sim X$. Let $N \perp \{X_1, X_2, X_3, \ldots\}$

$$\text{Let } S = \sum_{i=1}^{N} X_i$$

Q: What is $\boldsymbol{E}[S]$?

Q: Can we apply Linearity of Expectation?

No, because N is not a constant!

Q: Is there a way to make N into a constant?

Yes! We can condition on the value of N

Sum of random number of random variables

Let $X_1, X_2, X_3, \ldots$ be i.i.d. r.v.s, where $X_i \sim X$. Let $N \perp \{X_1, X_2, X_3, \ldots\}$

$$\text{Let } S = \sum_{i=1}^{N} X_i$$

$$\begin{aligned}\boldsymbol{E}[S] &= \sum_{n=1}^{\infty} \boldsymbol{E}[S \mid N = n] \cdot \boldsymbol{P}\{N = n\} \\ &= \sum_{n=1}^{\infty} \boldsymbol{E}\left[\sum_{i=1}^{n} X_i \mid N = n\right] \cdot \boldsymbol{P}\{N = n\} \\ &= \sum_{n=1}^{\infty} n\boldsymbol{E}[X] \cdot \boldsymbol{P}\{N = n\} = \boldsymbol{E}[X] \cdot \boldsymbol{E}[N]\end{aligned}$$

Sum of random number of random variables

Let $X_1, X_2, X_3, \ldots$ be i.i.d. r.v.s, where $X_i \sim X$.   Let $N \perp \{X_1, X_2, X_3, \ldots\}$

$$\text{Let } S = \sum_{i=1}^{N} X_i$$

Q: Can we get $\boldsymbol{Var}(S)$ similarly?

$$\boldsymbol{Var}(S) = \sum_{n=1}^{\infty} \boldsymbol{Var}[S \mid N = n] \cdot \boldsymbol{P}\{N = n\}$$

This is WRONG!
There's no Total Law of Variance

Sum of random number of random variables

Let $X_1, X_2, X_3, \ldots$ be i.i.d. r.v.s, where $X_i \sim X$. Let $N \perp \{X_1, X_2, X_3, \ldots\}$

$$\text{Let } S = \sum_{i=1}^{N} X_i$$

Q: Instead derive $\boldsymbol{E}[S^2]$

$$\boldsymbol{E}[S^2] = \sum_{n=1}^{\infty} \boldsymbol{E}[S^2 \mid N = n] \cdot \boldsymbol{P}\{N = n\}$$

$$= \sum_{n=1}^{\infty} \boldsymbol{E}[(X_1 + X_2 + \cdots + X_n)^2] \cdot \boldsymbol{P}\{N = n\}$$

$$= \sum_{n=1}^{\infty} (n\boldsymbol{E}[X_1^2] + (n^2 - n)\boldsymbol{E}[X_1 X_2]) \cdot \boldsymbol{P}\{N = n\}$$

$$= \boldsymbol{E}[X^2] \cdot \boldsymbol{E}[N] + \boldsymbol{E}[X]^2 \cdot (\boldsymbol{E}[N^2] - \boldsymbol{E}[N]) = \boldsymbol{E}[N]\boldsymbol{Var}(X) + \boldsymbol{E}[N^2]\boldsymbol{E}[X^2]$$

Sum of random number of random variables

Summary Theorem 5.14:

We'll do this much more easily when we get to transforms!

Let $X_1, X_2, X_3, \ldots$ be i.i.d. r.v.s, where $X_i \sim X$.

Let $S = \sum_{i=1}^{N} X_i$, where $N \perp \{X_1, X_2, X_3, \ldots\}$

Then

$$\boldsymbol{E}[S] = \boldsymbol{E}[N] \cdot \boldsymbol{E}[X]$$

$$\boldsymbol{E}[S^2] = \boldsymbol{E}[N] \cdot \boldsymbol{Var}(X) + \boldsymbol{E}[N^2] \cdot \boldsymbol{E}[X]^2$$

$$\boldsymbol{Var}(S) = \boldsymbol{E}[N] \cdot \boldsymbol{Var}(X) + \boldsymbol{Var}(N) \cdot \boldsymbol{E}[X]^2$$

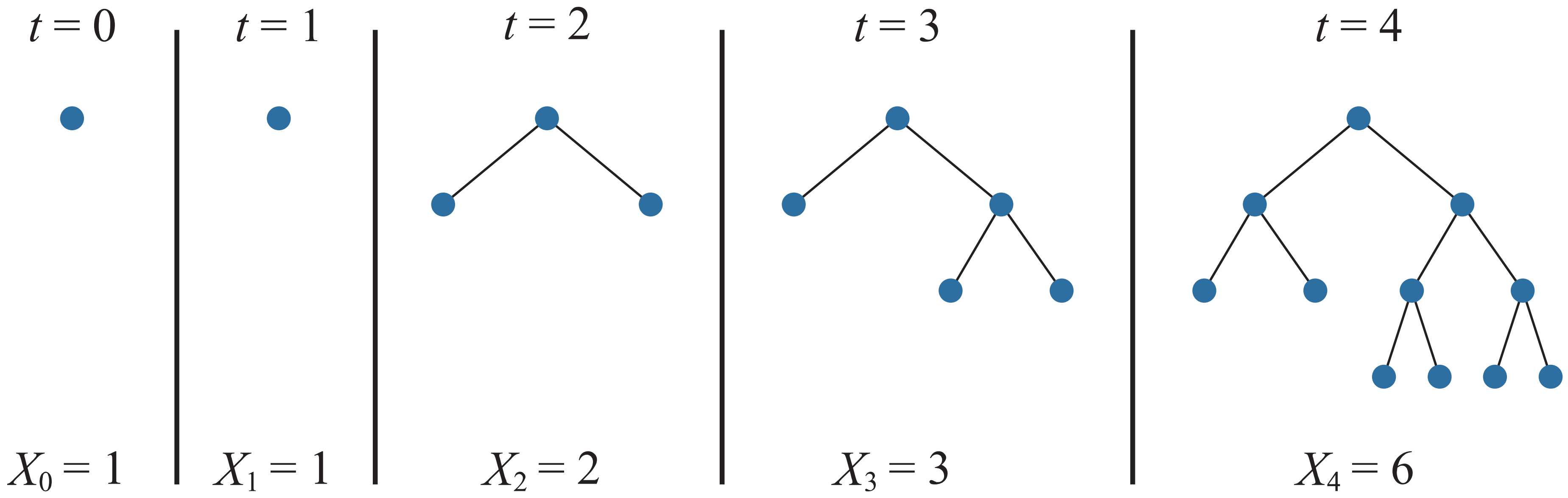
Example: Epidemic growth modeling

At each time step, every leaf independently either:

- forks off 2 children, w.p. $\frac{1}{2}$
- stays inert w.p. $\frac{1}{2}$

X_t is number of leaves in tree after t steps.

Q: What is $\boldsymbol{E}[X_t]$
What is $\boldsymbol{Var}(X_t)$?



Example: Epidemic growth modeling

At each time step, every leaf independently either:

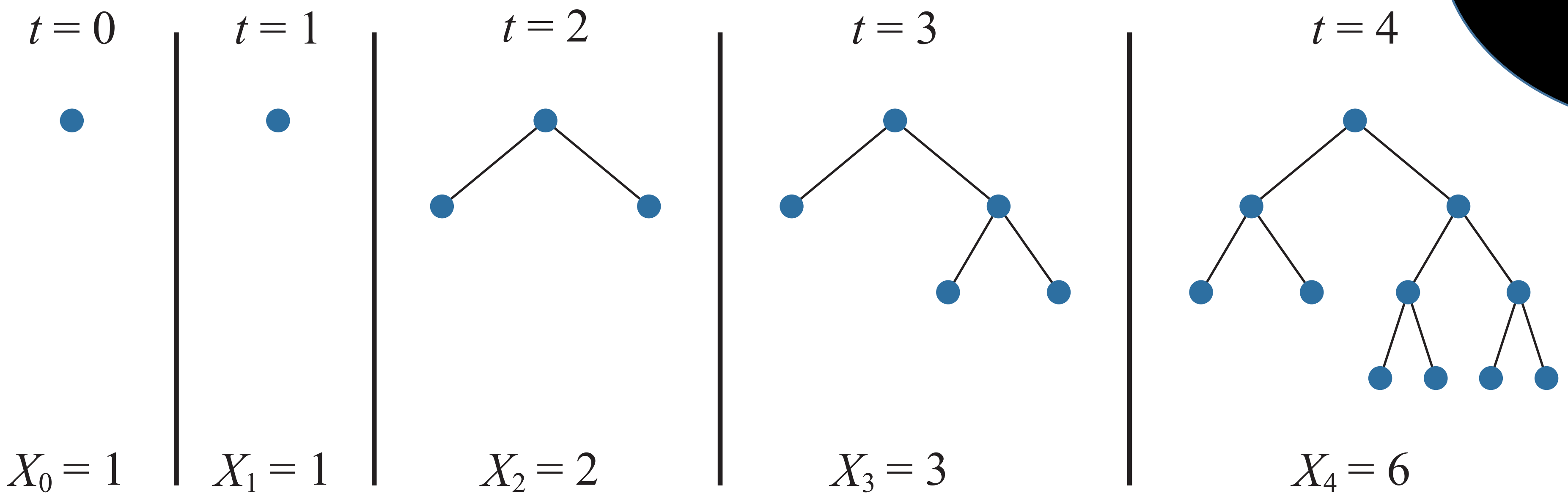
- forks off 2 children, w.p. $\frac{1}{2}$
- stays inert w.p. $\frac{1}{2}$

X_t is number of leaves in tree after t steps.

<u>Hint</u>:

$$X_t = \sum_{i=1}^{X_{t-1}} Y_i$$

What is Y_i?



See book for solution

Tail bounds

Defn: The **tail** of random variable X is $\boldsymbol{P}\{X > x\}$.

Example: T denotes response time at a web service.
Want to ensure the fraction of people with response time $> 0.5s$ is not too high.

Want an **upper bound** on $\boldsymbol{P}\{T > 0.5\}$. This is called a **tail bound**.

Tail bounds

Another Example: n items are hashed into a table of size n.
Assume each item ends up in a random bucket.
Ideally, we have 1 item per bucket.
What is the fraction of time that your search time $> k$?
(i.e., what's the probability your bucket has $> k$ items?)

Let $N =$ #items in bucket 1

How is N distributed?

$N \sim Binomial\left(n, \frac{1}{n}\right)$

$$\boldsymbol{P}\{N > k\} = \sum_{i=k+1}^{n} \boldsymbol{P}\{N = i\} = \sum_{i=k+1}^{n} \binom{n}{i} \left(\frac{1}{n}\right)^{i} \left(1 - \frac{1}{n}\right)^{n-i}$$

We don't know how to compute such bounds in general.

Point: We'll see that just knowing the mean and variance suffices for a tail **bound**.
In some cases, the mean alone suffices (although this bound is quite weak).

Markov's inequality

Theorem: (Markov's inequality) If r.v. X is non-negative, then $\forall a > 0$,

$$\boldsymbol{P}\{X \geq a\} \leq \frac{\boldsymbol{E}[X]}{a}$$

$$\boldsymbol{E}[X] = \sum_{x=0}^{\infty} x \cdot p_X(x) \geq \sum_{x=a}^{\infty} x \cdot p_X(x)$$

$$\geq \sum_{x=a}^{\infty} a \cdot p_X(x)$$

$$= a \sum_{x=a}^{\infty} p_X(x) = a \cdot \boldsymbol{P}\{X \geq a\}$$

Chebyshev's inequality

Theorem: (Chebyshev's inequality) Let X be any r.v. with finite mean, μ, and finite variance. Then $\forall a > 0$,

$$\boldsymbol{P}\{|X-\mu| \geq a\} \leq \frac{\boldsymbol{Var}(X)}{a^2}$$

$$\boldsymbol{P}\{|X-\mu| \geq a\} = \boldsymbol{P}\{(X-\mu)^2 \geq a^2\}$$

Q: Can you see how to apply Markov's inequality here?

$$\leq \frac{\boldsymbol{E}[(X-\mu)^2]}{a^2}$$

$$= \frac{\boldsymbol{Var}(X)}{a^2}$$

Chebyshev's inequality

Theorem: (Chebyshev's inequality) Let X be any r.v. with finite mean, μ, and finite variance. Then $\forall a > 0$,

$$\boldsymbol{P}\{|X-\mu| \geq a\} \leq \frac{\boldsymbol{Var}(X)}{a^2}$$

Example:

$$N \sim Binomial\left(n, \frac{1}{n}\right)$$

Provide upper bound on: $\boldsymbol{P}\{N \geq 6\}$

$$\boldsymbol{P}\{N \geq 6\} \leq \boldsymbol{P}\{|N-1| \geq 5\}$$

$$\leq \frac{\boldsymbol{Var}(N)}{25}$$

$$\leq \frac{1}{25}$$

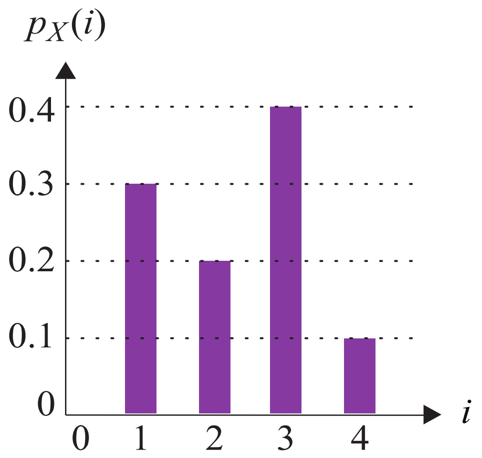
Stochastic dominance

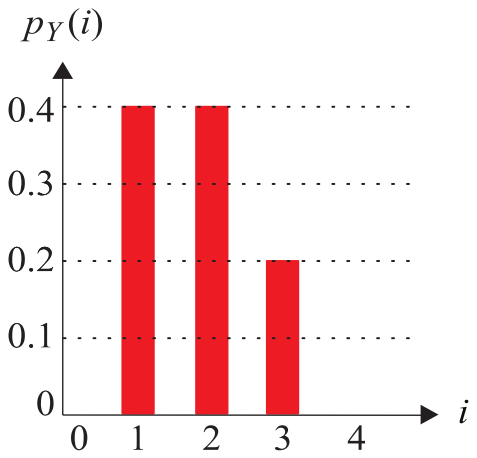
<u>Defn 5.18</u>: Given two random variables X and Y, if

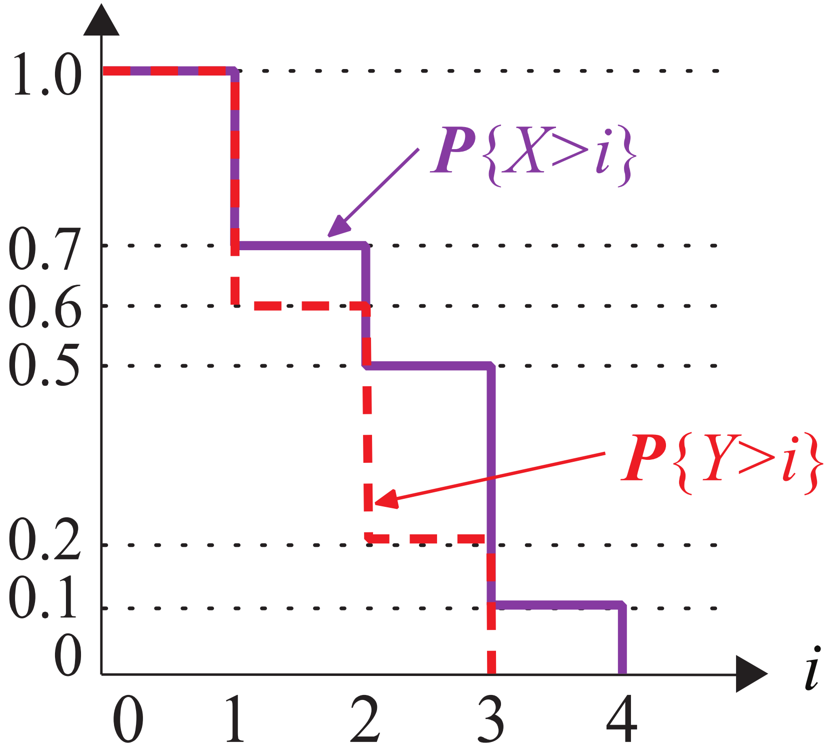
$$\boldsymbol{P}\{X > i\} \geq \boldsymbol{P}\{Y > i\}, \qquad \forall i$$

we say that X **stochastically dominates** Y:

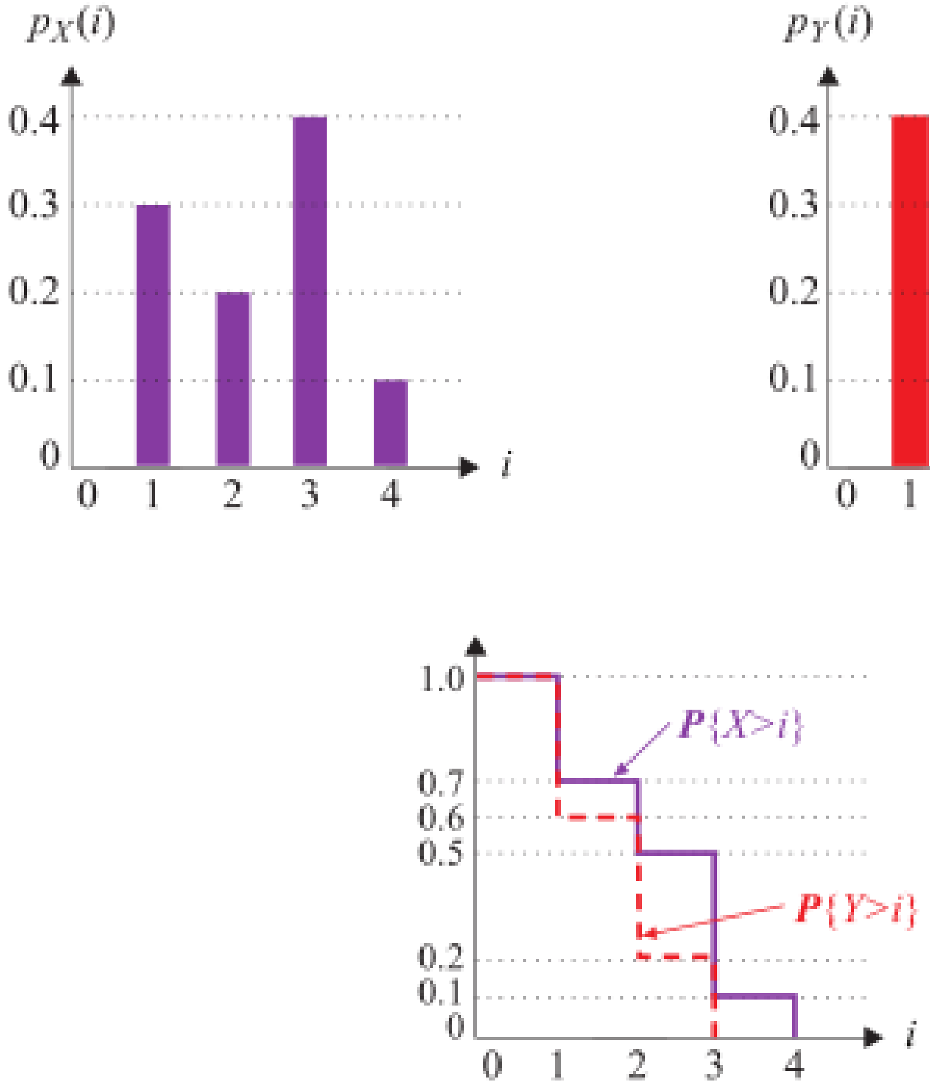
$$X \geq_{st} Y$$





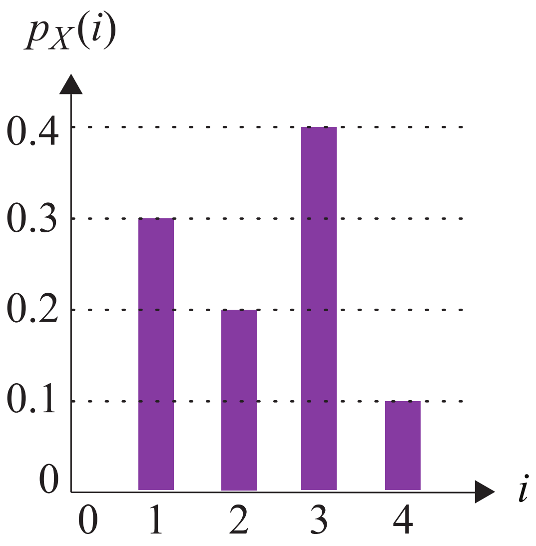


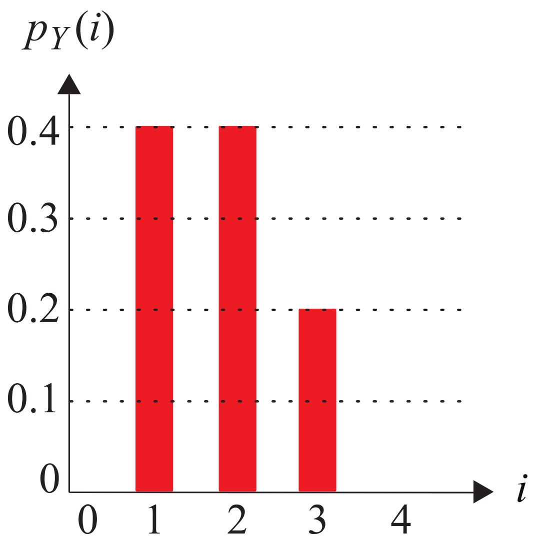
Test page – EPS File converted to png

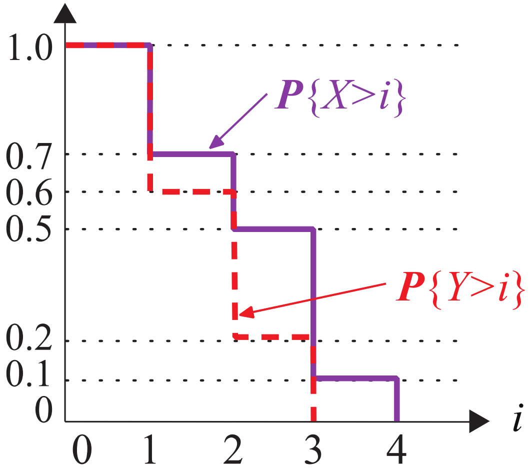


"Introduction to Probability for Computing", Harchol-Balter '24

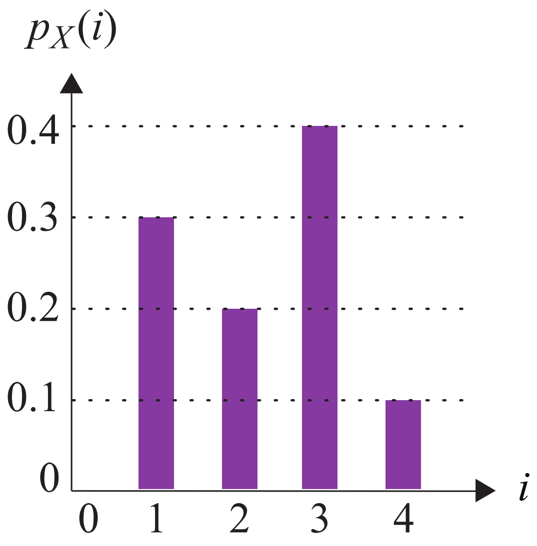
Test page – EPS File converted to EMF

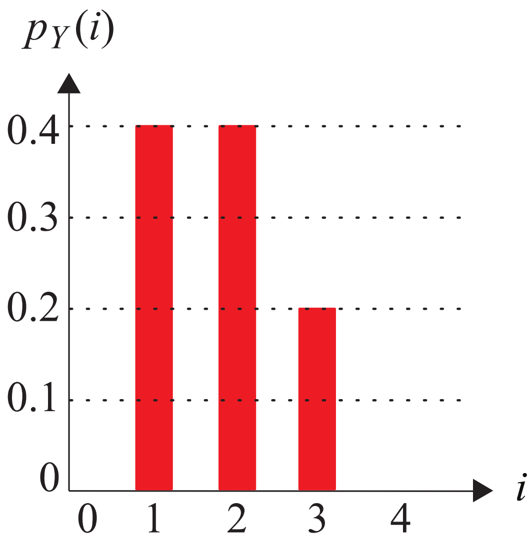


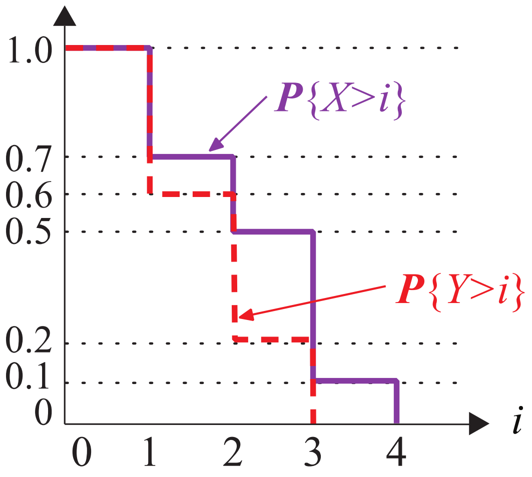




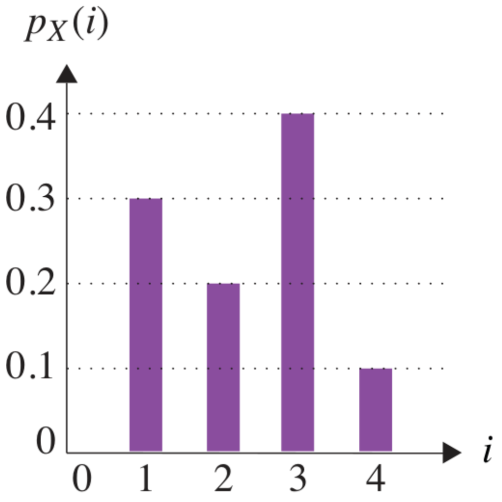
Test page– pdf file converted to EMF

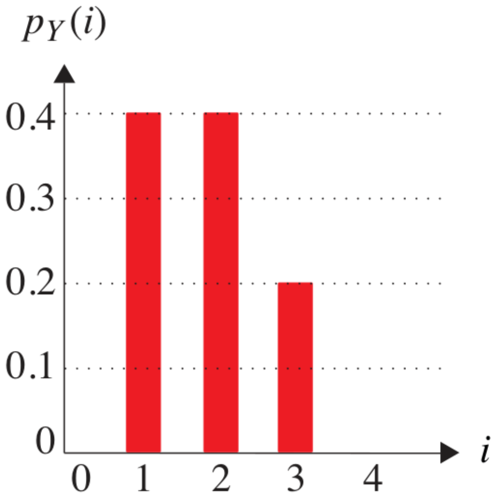


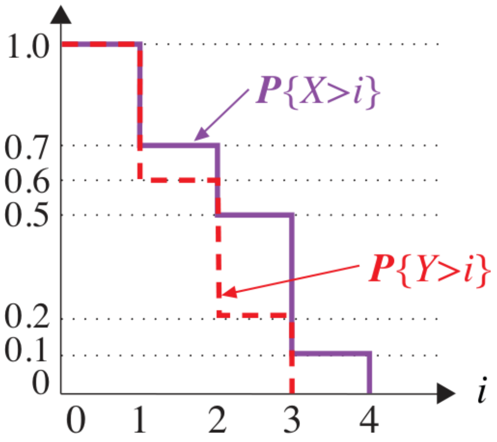




Test page– pdf file converted to png







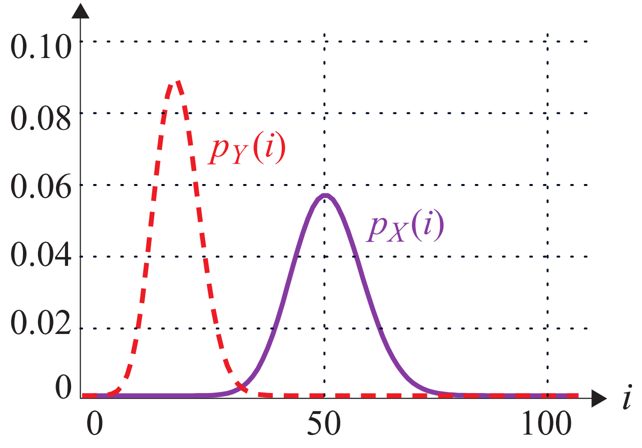
"Introduction to Probability for Computing", Harchol-Balter '24

Stochastic dominance

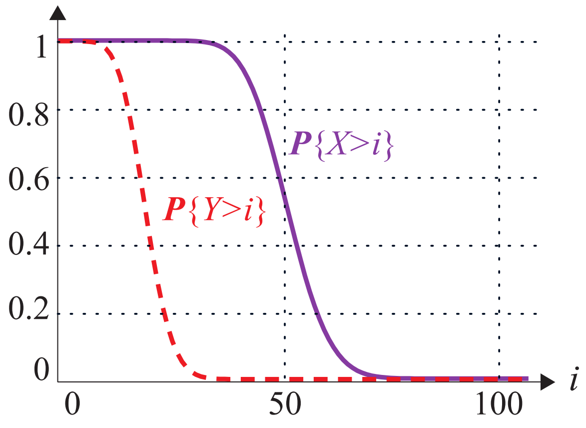
X = Number pairs of shoes owned by women ~ $Poisson(\lambda = 27)$

Y = Number pairs of shoes owned by men ~ $Poisson(\lambda = 12)$

<u>Most, but not all, women have more shoes than men</u>



<u>But women stochastically dominate men w.r.t. shoes</u>



Jensen's inequality: motivation

We already know that

$$\boldsymbol{E}[X^2] \geq \boldsymbol{E}[X]^2$$

Is it also the case that

$$\boldsymbol{E}[X^3] \geq \boldsymbol{E}[X]^3 \text{ ?}$$

$$\boldsymbol{E}[X^4] \geq \boldsymbol{E}[X]^4 \text{ ?}$$

$$\boldsymbol{E}[X^{4.5}] \geq \boldsymbol{E}[X]^{4.5} \text{ ?}$$

This is a consequence of Jensen's inequality! (Exercise 5.32)

Theorem: Let X be any positive r.v. Then $\forall a \in Reals$,

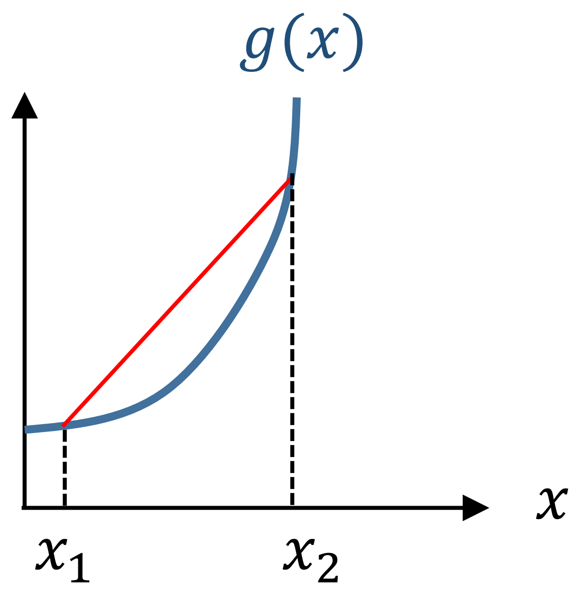
$$\boldsymbol{E}[X^a] \geq \boldsymbol{E}[X]^a$$

Jensen's inequality

<u>Defn 5.21</u>: A function $g(x)$ is **convex** on interval S if, for any $x_1, x_2 \in S$, and any $\alpha \in [0,1]$, we have:

$$g(\alpha x_1 + (1-\alpha)x_2) \leq \alpha g(x_1) + (1-\alpha)g(x_2)$$

The curve always lies below the line segment.



$g(x)$ is convex on S iff $g''(x) \geq 0,\ \forall x \in S$.

Jensen's inequality

Defn 5.22: A function $g(x)$ is **convex** on interval S if, for any $x_1, x_2, \dots, x_n \in S$, and any $\alpha_1, \alpha_2, \dots, \alpha_n \in [0,1]$, where $\sum_i \alpha_i = 1$, we have:

$$g(\alpha_1 x_1 + \alpha_2 x_2 + \cdots + \alpha_n x_n) \leq \alpha_1 g(x_1) + \alpha_2 g(x_2) + \cdots + \alpha_n g(x_n)$$

$$g(p_X(x_1)x_1 + \cdots + p_X(x_n)x_n) \leq p_X(x_1)g(x_1) + \cdots + p_X(x_n)g(x_n)$$

$$X = \begin{cases} x_1 & \text{w.p.} \quad p_X(x_1) \\ x_2 & \text{w.p.} \quad p_X(x_2) \\ \vdots & \\ x_n & \text{w.p.} \quad p_X(x_n) \end{cases}$$

$$\Rightarrow \quad g(\boldsymbol{E}[X]) \leq \boldsymbol{E}[g(X)]$$

Jensen's inequality

Theorem 5.23: (Jensen's inequality) If $g(x)$ is **convex** on interval S and X takes on values on interval S, then:

$$g(\boldsymbol{E}[X]) \leq \boldsymbol{E}[g(X)]$$

$$g(p_X(x_1)x_1 + \cdots + p_X(x_n)x_n) \leq p_X(x_1)g(x_1) + \cdots + p_X(x_n)g(x_n)$$

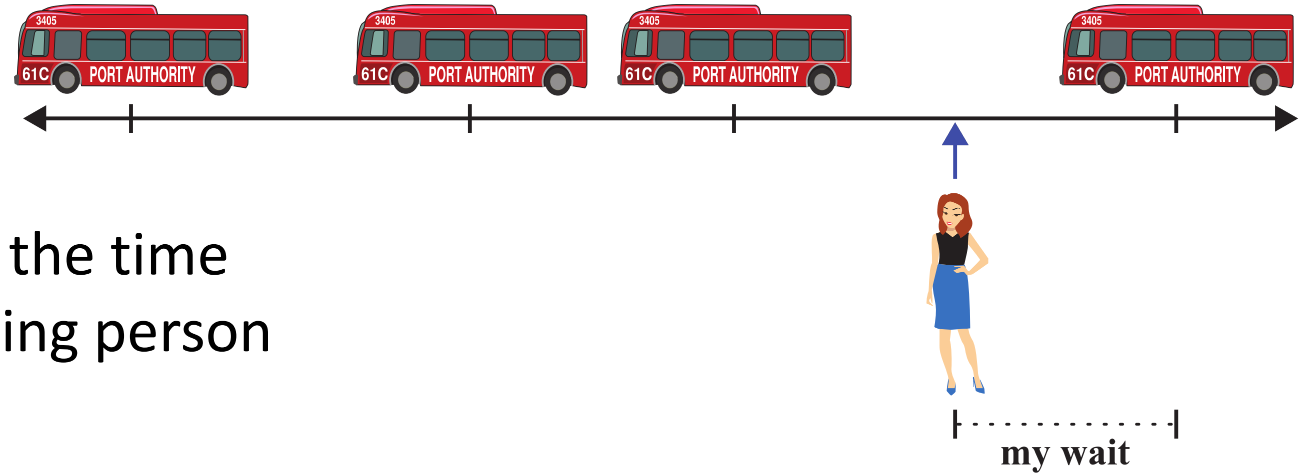
$$X = \begin{cases} x_1 & \text{w.p.} \quad p_X(x_1) \\ x_2 & \text{w.p.} \quad p_X(x_2) \\ \vdots & \\ x_\text{n} & \text{w.p.} \quad p_X(x_n) \end{cases}$$

$$\Rightarrow g(\boldsymbol{E}[X]) \leq \boldsymbol{E}[g(X)]$$

Inspection paradox: Consequence of high variability

<u>Defn</u>: The **inspection paradox** says that, in high-variability settings, the mean seen by a random observer can be very different from the true mean.

Mean time between buses is 10 minutes.



However if there is some variability in the time between buses, then a randomly arriving person will wait more than 5 minutes.

Expected wait can even be >10 minutes!

How can this be?

Inspection paradox: Consequence of high variability

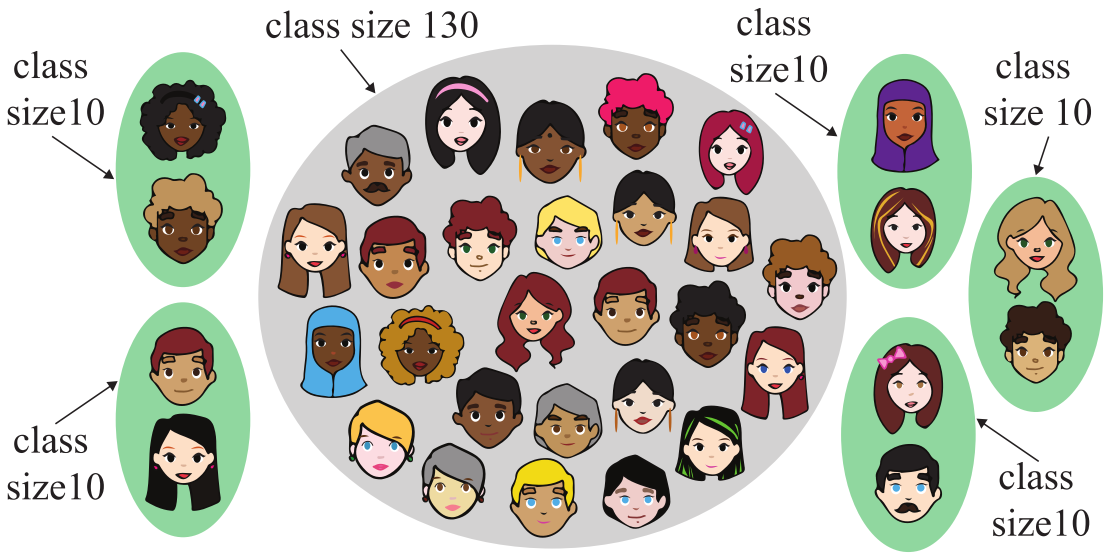
Defn: The **inspection paradox** says that, in high-variability settings, the mean seen by a random observer can be very different from the true mean.

Average class size reported by students is 100.

But the dean claims average class size is 30.

No one is lying.

How can this be?



Inspection paradox: Consequence of high variability

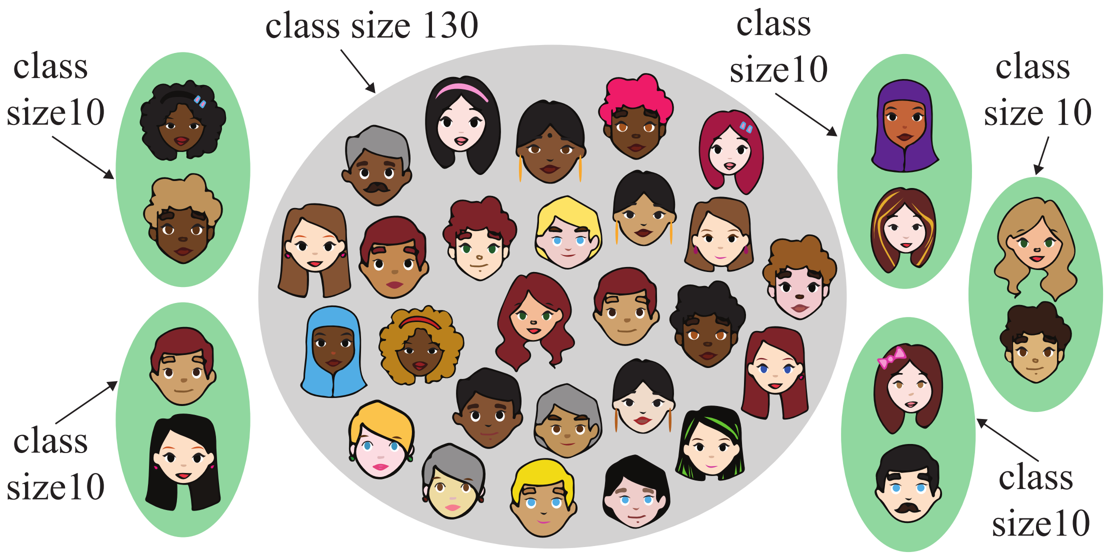
Defn: The **inspection paradox** says that, in high-variability settings, the mean seen by a random observer can be very different from the true mean.

Average class size reported by students is 100.

But the dean claims average class size is 30.

No one is lying.

How can this be?



180 students in 6 classes ➔ 30 students/class.

Avg observed class size $= \frac{50}{180} \cdot 10 + \frac{130}{180} \cdot 130 \approx 97$

Inspection paradox: Consequence of high variability

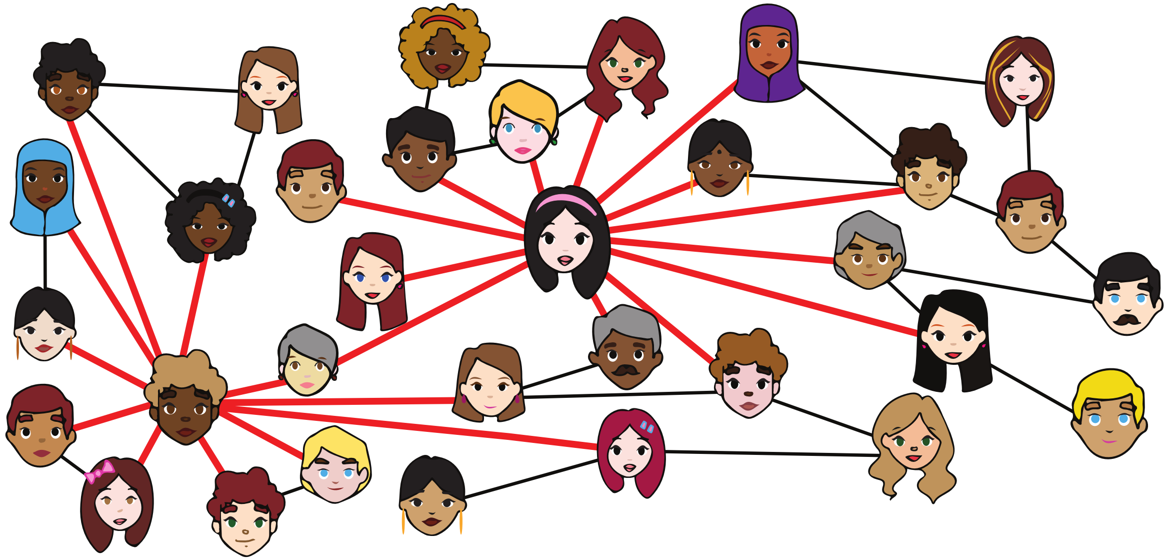
<u>Defn</u>: The **inspection paradox** says that, in high-variability settings, the mean seen by a random observer can be very different from the true mean.

The average Facebook user has 44 friends.

But the average friend of a Facebook user has 104 friends.

In fact, with probability 76%, your friend is more popular than you are.

How can this be?



Most people have few friends.

A few people are very popular with many friends.

Which classification most likely describes you?

Which most likely describes your friend?