S&DS 241 Lecture 12

Mean, variance, deviation inequalities

B-H: 4.6, 10.1.3

Recall from Lecture 5: Expectation

The expectation (aka expected value or mean) of a discrete random variable X is

$$\boxed{E(X) = \sum_{x \in \mathcal{X}} x p_X(x)}$$

where $p_X(x) = P(X = x)$ is the PMF.

Properties of expectation

- X is a nonnegative random variable $\implies E(X) \geq 0$

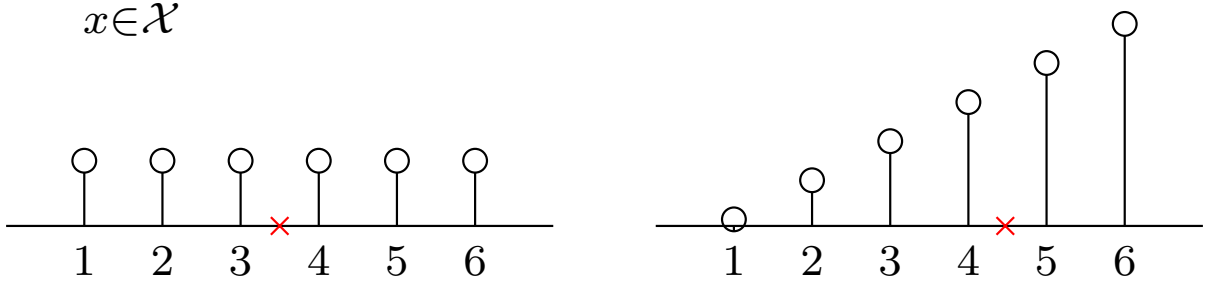
Properties of expectation

- X is a nonnegative random variable $\implies E(X) \geq 0$
- Linearity of expectation: For any constants a, b and any random variables X and Y (not necessarily independent!)

$$E(aX + bY) = aE(X) + bEY$$

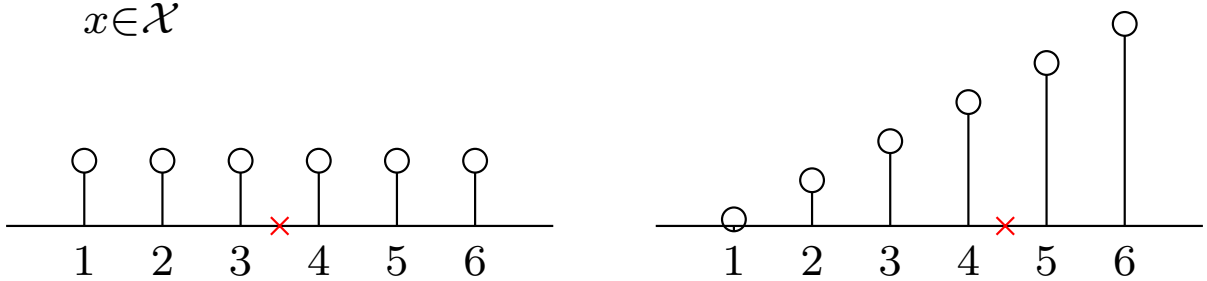
Interpretation of expectation: Center of gravity

$$E(X) = \sum_{x \in \mathcal{X}} x p_X(x) = \text{average of values weighted by PMF}$$



Interpretation of expectation: Center of gravity

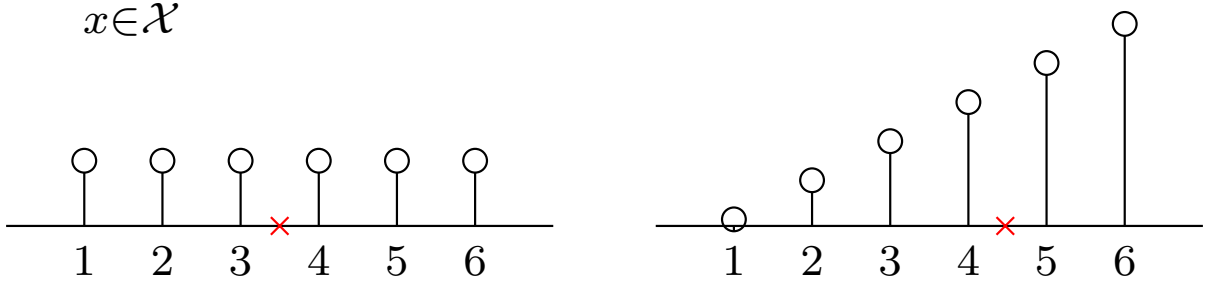
$$E(X) = \sum_{x \in \mathcal{X}} x p_X(x) = \text{average of values weighted by PMF}$$



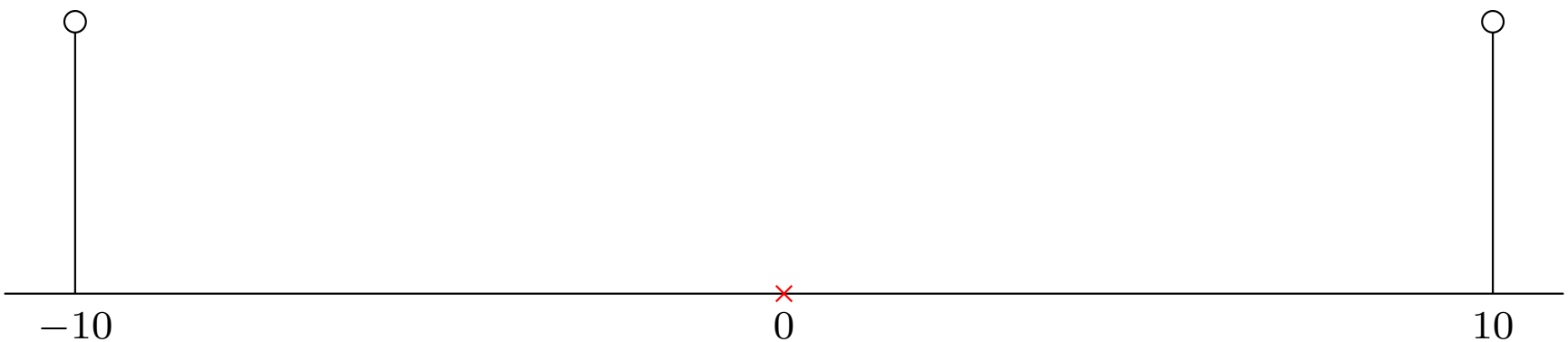
- Expected value $\neq$ typical value!
- How close a random variable is to its expectation depends on many things, e.g., variance

Interpretation of expectation: Center of gravity

$$E(X) = \sum_{x \in \mathcal{X}} x p_X(x) = \text{average of values weighted by PMF}$$

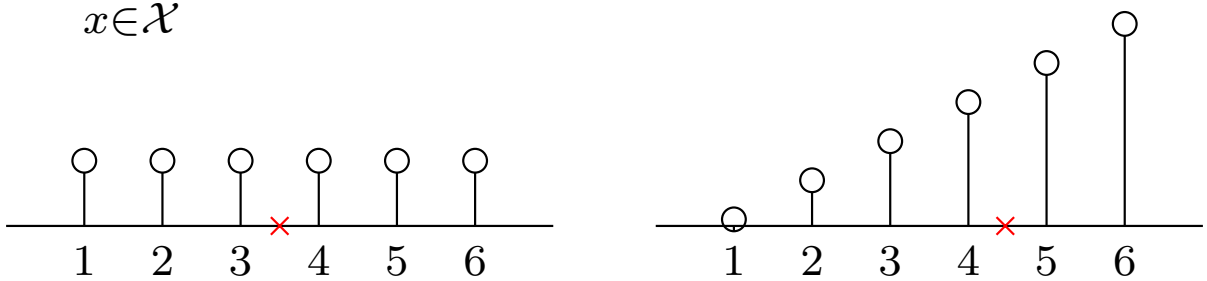


- Expected value $\neq$ typical value!
- How close a random variable is to its expectation depends on many things, e.g., variance

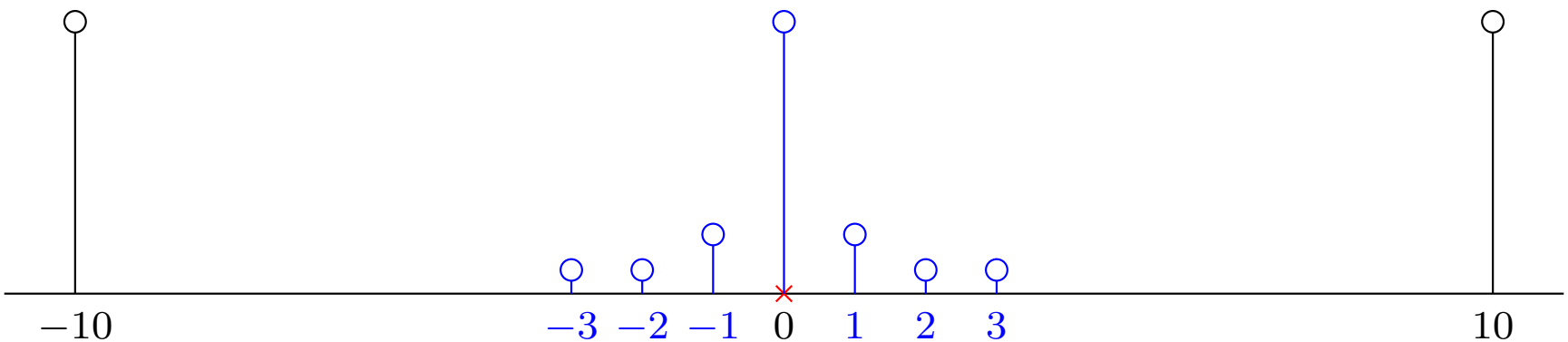


Interpretation of expectation: Center of gravity

$$E(X) = \sum_{x \in \mathcal{X}} x p_X(x) = \text{average of values weighted by PMF}$$



- Expected value $\neq$ typical value!
- How close a random variable is to its expectation depends on many things, e.g., variance



Variance: definition

Variance of a random variable X:

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

- Other notations: $\sigma_X^2, V(X)$
- Variance $=$ **mean-squared** deviation from the expectation
- Significance: measures the uncertainty of a random variable/the spread of a distribution

Standard deviation

- Variance is quadratic in nature: for example

	Units
X	ft
$E(X)$	ft
$\mathrm{Var}(X)$	sqft

Standard deviation

- Variance is quadratic in nature: for example

	Units
X	ft
$E(X)$	ft
$\mathrm{Var}(X)$	sqft

- Standard deviation of X:

$$\mathrm{SD}(X) = \sigma_X = \sqrt{\mathrm{Var}(X)}$$

which is the root-mean-squared deviation from expectation

Standard deviation

- Variance is quadratic in nature: for example

	Units
X	ft
$E(X)$	ft
$\mathrm{Var}(X)$	sqft
$\mathrm{SD}(X)$	ft

- Standard deviation of X:

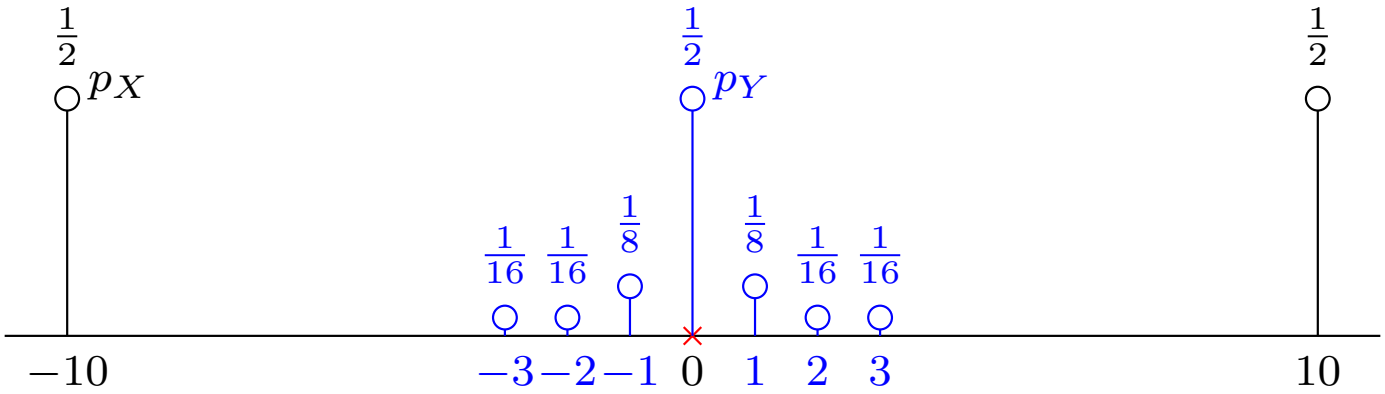
$$\mathrm{SD}(X) = \sigma_X = \sqrt{\mathrm{Var}(X)}$$

which is the root-mean-squared deviation from expectation

Intuition

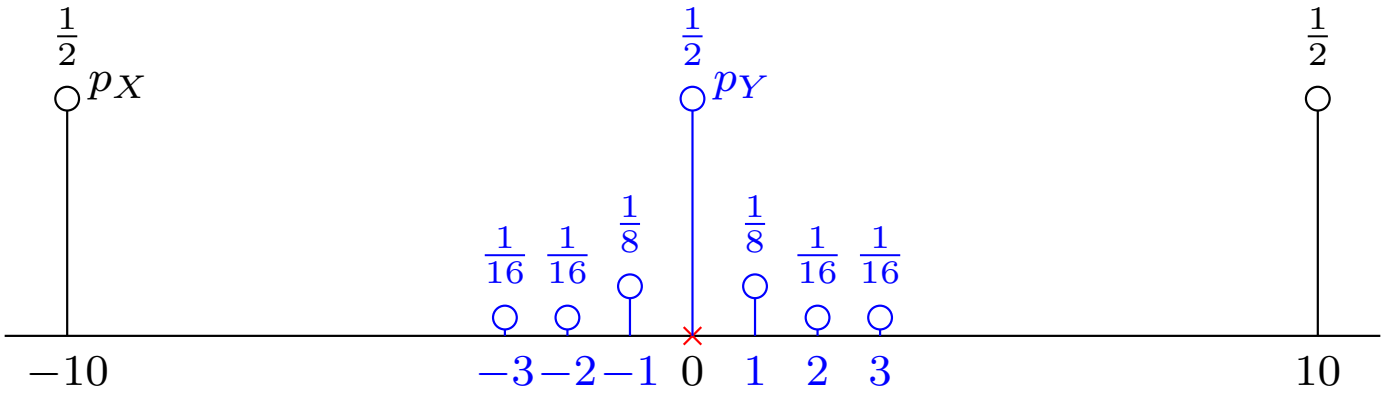
- Smaller $\mathrm{Var}(X)$ or $\mathrm{SD}(X)$ $\implies$ PMF of X is more concentrated around the mean
- Larger $\mathrm{Var}(X)$ or $\mathrm{SD}(X)$ $\implies$ PMF of X is more spread out

Example



- Mean: $E(X) = EY = 0$

Example

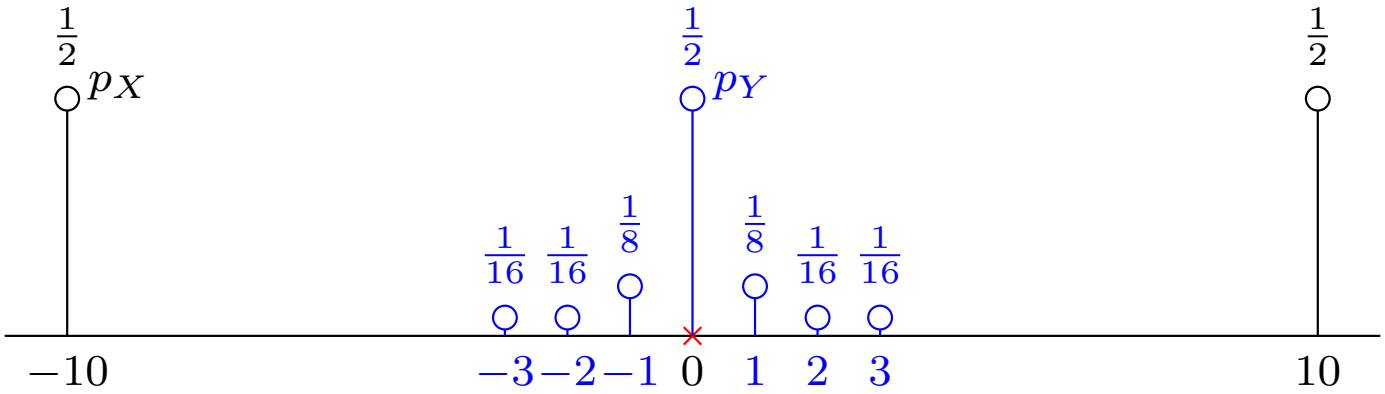


- Mean: $E(X) = EY = 0$
- Variance:

$$\mathrm{Var}(X) = E(X^2) \overset{\text{LOTUS}}{=\!=} \frac{1}{2} \times (-10)^2 + \frac{1}{2} \times 10^2 = 100$$

$$\mathrm{Var}(Y) = E(Y^2) \overset{\text{LOTUS}}{=\!=} \frac{1^2 \times 2}{8} + \frac{2^2 \times 2 + 3^2 \times 2}{16} = \frac{7}{4}$$

Example



- Mean: $E(X) = EY = 0$
- Variance:

$$\mathrm{Var}(X) = E(X^2) \overset{\text{LOTUS}}{=\!=} \frac{1}{2} \times (-10)^2 + \frac{1}{2} \times 10^2 = 100$$

$$\mathrm{Var}(Y) = E(Y^2) \overset{\text{LOTUS}}{=\!=} \frac{1^2 \times 2}{8} + \frac{2^2 \times 2 + 3^2 \times 2}{16} = \frac{7}{4}$$

- Standard deviation:

$$\mathrm{SD}(X) = 10, \quad \mathrm{SD}(Y) \approx 1.32$$

Indeed, Y is much more concentrated than X

Alternative formula of variance

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

Proof.

Call $\mu = E(X)$. By linearity of expectation,

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

□

Alternative formula of variance

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

Proof.

Call $\mu = E(X)$. By linearity of expectation,

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

□

- First moment: $E(X)$
- Second moment: $E(X^2)$
- kth moment: $E(X^k)$

Alternative formula of variance

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

Proof.

Call $\mu = E(X)$. By linearity of expectation,

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

□

- First moment: $E(X)$
- Second moment: $E(X^2)$
- kth moment: $E(X^k)$
- Variance = second moment $-$(first moment)^2

Example: dice

Let X be the outcome of a fair die. Find $\mathrm{Var}(X)$.

$$E(X) = \frac{1}{6}(1+2+3+4+5+6) = \frac{7}{2}$$
$$E(X^2) = \frac{1}{6}(1^2+2^2+3^2+4^2+5^2+6^2) = \frac{91}{6}$$

Therefore

$$\mathrm{Var}(X) = \frac{91}{6} - \left(\frac{7}{2}\right)^2 = \frac{35}{12}$$
$$\mathrm{SD}(X) = \sqrt{\frac{35}{12}} \approx 1.71$$

Question

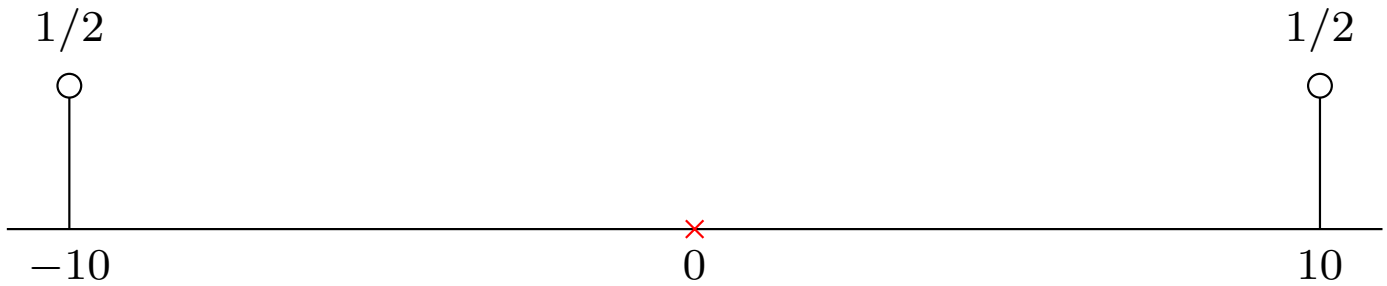
Suppose X takes values in $[-10, 10]$.

- What is the largest possible variance?

Question

Suppose X takes values in $[-10, 10]$.

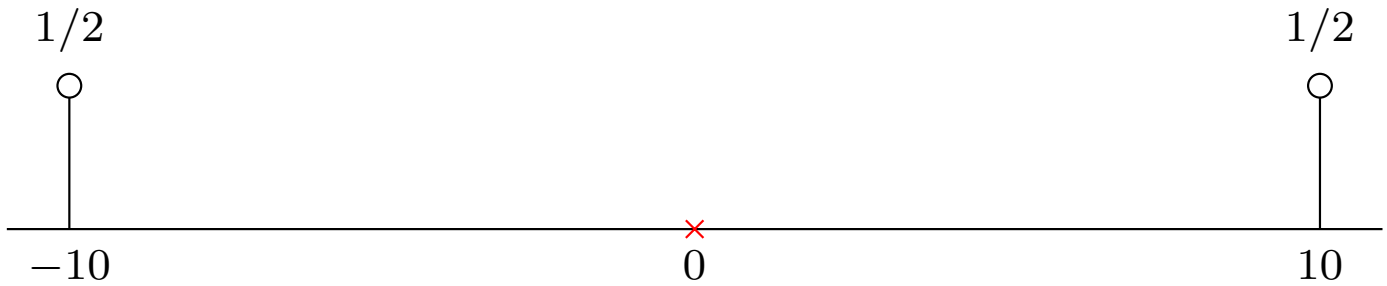
- What is the largest possible variance? 100



Question

Suppose X takes values in $[-10, 10]$.

- What is the largest possible variance? 100

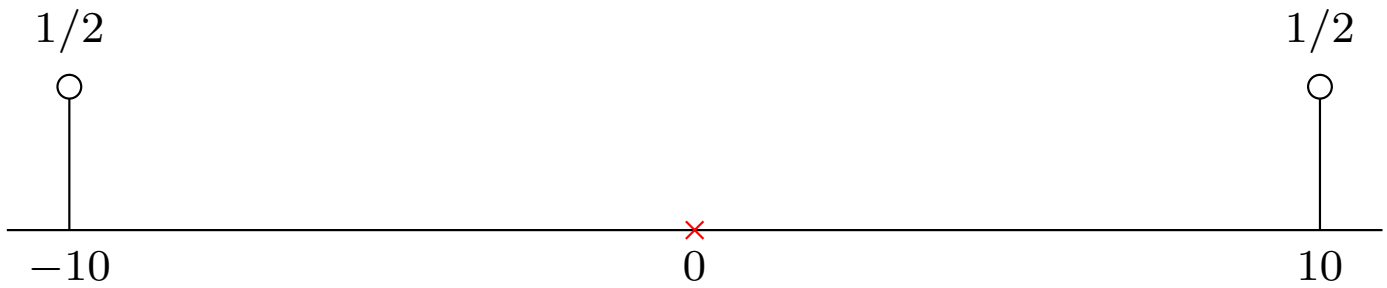


- What is the smallest possible variance?

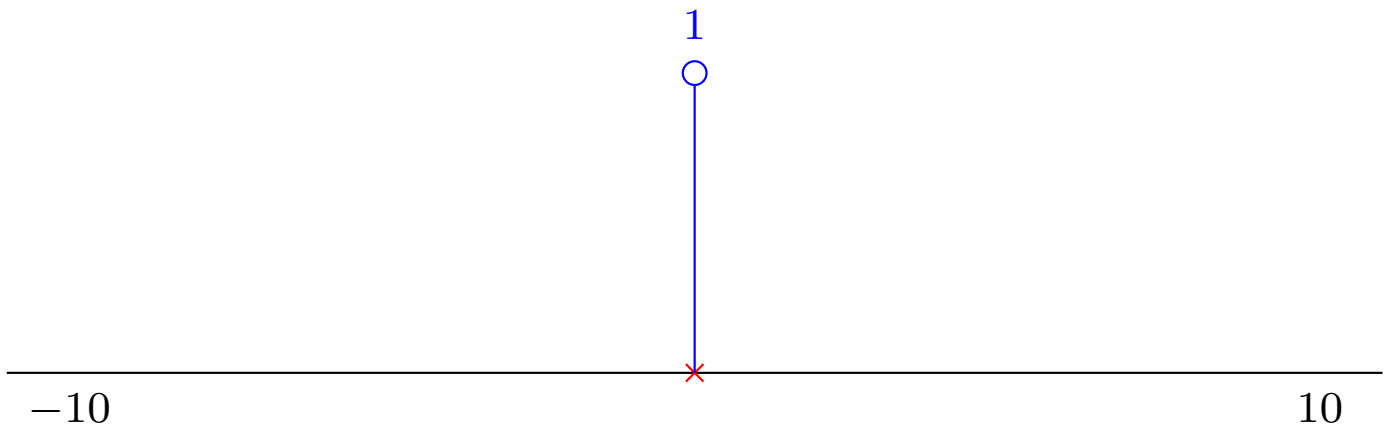
Question

Suppose X takes values in $[-10, 10]$.

- What is the largest possible variance? 100



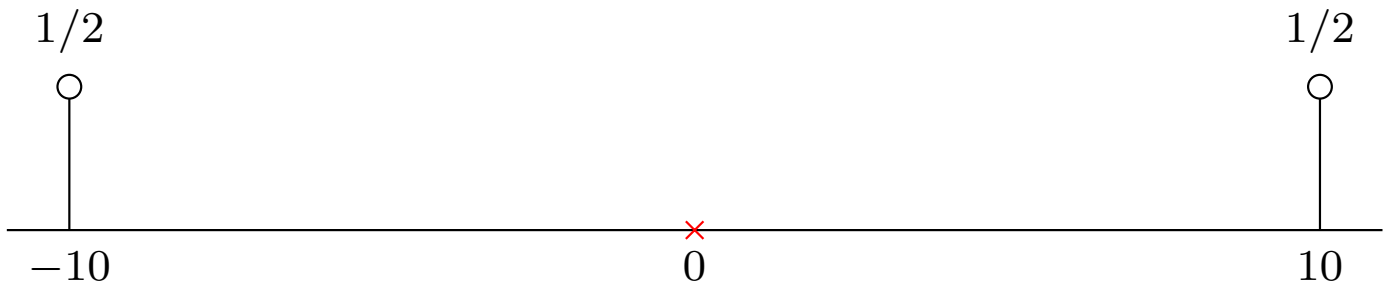
- What is the smallest possible variance? 0



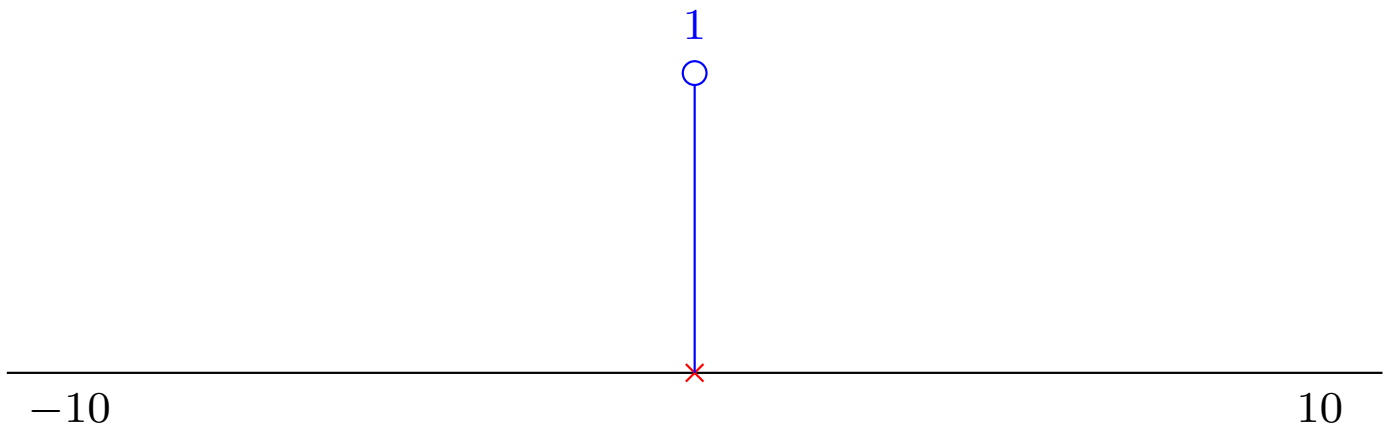
Question

Suppose X takes values in $[-10, 10]$.

- What is the largest possible variance? 100



- What is the smallest possible variance? 0



Fact: $\mathrm{Var}(X) = 0 \Leftrightarrow X =$ constant (no randomness)

Properties of mean, variance and std dev

- Shift does not change variance

$$\mathrm{Var}(X + b) = \mathrm{Var}(X)$$

- Scaling:

$$\mathrm{Var}(aX) = a^2\mathrm{Var}(X)$$

- More generally:

$$\boxed{\begin{aligned} E(aX + b) &= aE(X) + b \\ \mathrm{Var}(aX + b) &= a^2\mathrm{Var}(X) \\ \mathrm{SD}(aX + b) &= |a|\mathrm{SD}(X) \end{aligned}}$$

Sum of independent variables

- For any X and Y

$$E(X+Y) = E(X) + E(Y)$$

thanks to linearity of expectation

Sum of independent variables

- For any X and Y

$$E(X+Y) = E(X) + E(Y)$$

  thanks to linearity of expectation

- For independent X and Y

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

Sum of independent variables

- For any X and Y

$$E(X+Y) = E(X) + E(Y)$$

  thanks to linearity of expectation

- For independent X and Y

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

  This can fail without independence,

Sum of independent variables

- For any X and Y

$$E(X+Y) = E(X) + E(Y)$$

  thanks to linearity of expectation

- For independent X and Y

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

  This can fail without independence, e.g., $Y = -X$.

Sum of independent variables

For independent X and Y

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

Proof.

Since shifting does not change variance, we can assume, without loss of generality, that $E(X) = E(Y) = 0$.

Sum of independent variables

For independent X and Y

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

Proof.

Since shifting does not change variance, we can assume, without loss of generality, that $E(X) = E(Y) = 0$. Then

$$\begin{aligned}
&\mathrm{Var}(X+Y) \\
&= E((X+Y)^2) && E(X+Y) = 0 \\
&= E(X^2 + Y^2 + 2XY) \\
&= E(X^2) + E(Y^2) + 2E(XY) && \text{linearity of expectation} \\
&= E(X^2) + E(Y^2) && \text{independence so } E(XY) = E(X)E(Y) \\
&= \mathrm{Var}(X) + \mathrm{Var}(Y) && \square
\end{aligned}$$

Variance of sum of independent random variables

Let $X_1, \ldots, X_n$ be independent random variables. Then

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

Variance of sum of iid random variables

Let $X_1, \ldots, X_n$ be iid random variables. Then

$$\mathrm{Var}(X_1 + X_2 + \cdots + X_n) = n \cdot \mathrm{Var}(X_1)$$
$$\mathrm{SD}(X_1 + X_2 + \cdots + X_n) = \sqrt{n} \cdot \mathrm{SD}(X_1)$$

in contrast

$$\mathrm{SD}(X_1 + X_1 + \cdots + X_1) = n \cdot \mathrm{SD}(X_1)$$

Note: $\sqrt{n} \ll n$ for large n. Why does this happen?

Variance of sum of iid random variables

Let $X_1, \ldots, X_n$ be iid random variables. Then

$$\mathrm{Var}(X_1 + X_2 + \cdots + X_n) = n \cdot \mathrm{Var}(X_1)$$
$$\mathrm{SD}(X_1 + X_2 + \cdots + X_n) = \sqrt{n} \cdot \mathrm{SD}(X_1)$$

in contrast

$$\mathrm{SD}(X_1 + X_1 + \cdots + X_1) = n \cdot \mathrm{SD}(X_1)$$

Note: $\sqrt{n} \ll n$ for large n. Why does this happen? Example: random walk ($X_i = \pm 1$)

- $X_1 + X_1 + \cdots + X_1$: all steps are aligned (either all $+$ or all $-$). This leads to SD on the order of n.
- $X_1 + X_2 + \cdots + X_n$: some are $+$ and some $-$. Cancellation leads to SD on the order of $\sqrt{n}$.

Statistical application: sample average

Let $X_1, \ldots, X_n$ be iid random variables with mean μ and variance σ^2.

- Sample average:
$$\overline{X} = \frac{1}{n}(X_1 + \ldots + X_n)$$
- Then
$$E(\overline{X}) = \mu, \quad \mathrm{Var}(\overline{X}) = \frac{\sigma^2}{n} \ll \sigma^2 \quad \text{for large } n$$
- More data (larger sample size) $\implies$ less uncertainty

Summary

- $\mathrm{Var}(X) = \mathrm{SD}(X)^2 = E((X - E(X))^2) \geq 0$
- $\mathrm{Var}(X) = 0 \Leftrightarrow X =$ constant
- $\mathrm{Var}(X) = E(X^2) - (E(X))^2$
- $\mathrm{Var}(aX + b) = a^2 \mathrm{Var}(X)$
- $\mathrm{Var}(X + Y) \overset{\text{independence}}{=\!=\!=} \mathrm{Var}(X) + \mathrm{Var}(Y)$

Variance of common distributions

Bernoulli

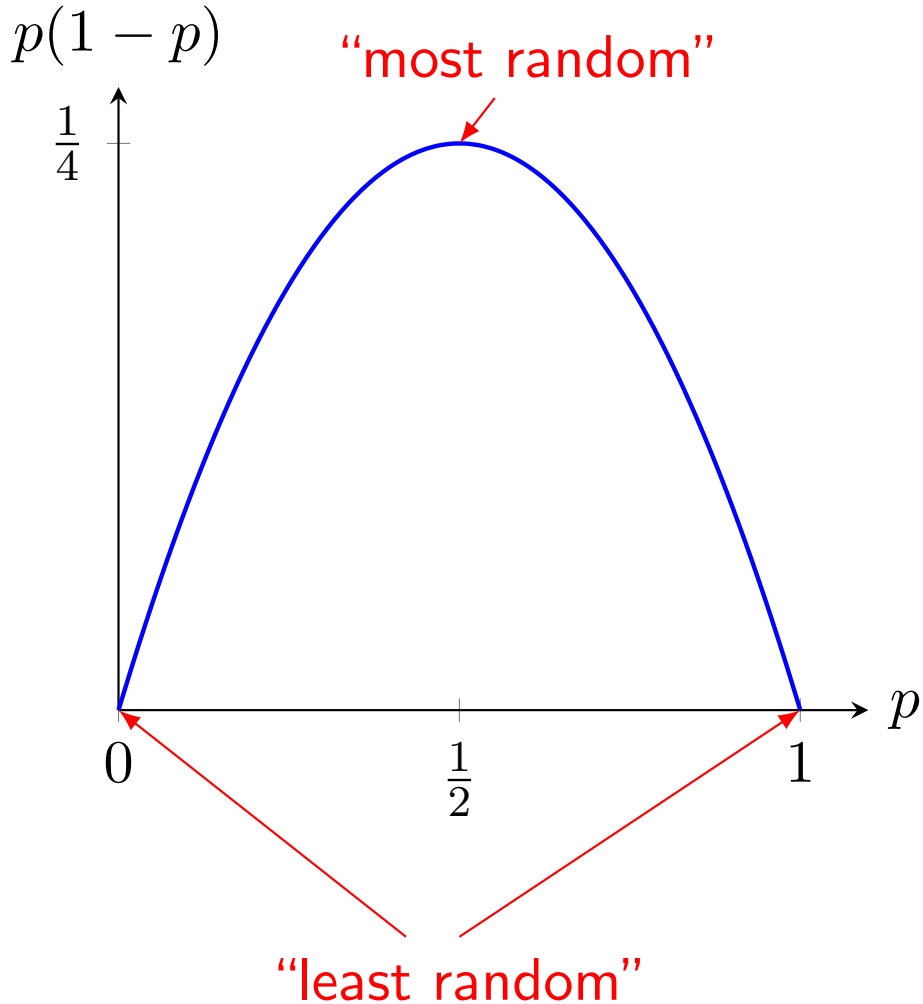
$X \sim \text{Bern}(p)$. Then

$$\text{Var}(X) = E(X^2) - (E(X))^2 = p - p^2 = p(1-p)$$

Bernoulli

$X \sim \text{Bern}(p)$. Then

$$\text{Var}(X) = E(X^2) - (E(X))^2 = p - p^2 = p(1-p)$$



Binomial

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

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

Proof.

$$X = \underbrace{X_1 + \cdots + X_n}_{\text{i.i.d. Bern}(p)} \implies \text{Var}(X) = n\text{Var}(X_1)$$

□

Binomial

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

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

Proof.

$$X = \underbrace{X_1 + \cdots + X_n}_{\text{i.i.d. Bern}(p)} \implies \text{Var}(X) = n\text{Var}(X_1)$$

□

Exercise: Alternatively, find $E(X^2)$ using LOTUS rule and binomial PMF, then apply $E(X^2) - (E(X))^2$.

Poisson

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

$$\text{Var}(X) = \lambda$$

Poisson

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

$$\text{Var}(X) = \lambda$$

Interpretation: Poisson as limiting Binomial

$$\text{Bin}\left(n, \frac{\lambda}{n}\right) \xrightarrow{n\to\infty} \quad \text{Pois}(\lambda)$$

so we expect:

$$\text{Var} = n\frac{\lambda}{n}\left(1 - \frac{\lambda}{n}\right) \xrightarrow{n\to\infty} \quad \lambda$$

Verify using Poisson PMF (B-H Example 4.7.2)

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

Verify using Poisson PMF (B-H Example 4.7.2)

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

Verify using Poisson PMF (B-H Example 4.7.2)

$$\begin{aligned} E(X^2) &= \sum_{k=1}^{\infty} k^2 \frac{e^{-\lambda}\lambda^k}{k!} \\ &= \sum_{k=1}^{\infty} k \frac{e^{-\lambda}\lambda^k}{(k-1)!} \end{aligned}$$

Verify using Poisson PMF (B-H Example 4.7.2)

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

$$= \sum_{k=1}^{\infty} k \frac{e^{-\lambda}\lambda^k}{(k-1)!} = \sum_{k=1}^{\infty} (k-1) \frac{e^{-\lambda}\lambda^k}{(k-1)!} + \sum_{k=1}^{\infty} \frac{e^{-\lambda}\lambda^k}{(k-1)!}$$

Verify using Poisson PMF (B-H Example 4.7.2)

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

$$= \sum_{k=1}^{\infty} k \frac{e^{-\lambda}\lambda^k}{(k-1)!} = \sum_{k=2}^{\infty} (k-1) \frac{e^{-\lambda}\lambda^k}{(k-1)!} + \sum_{k=1}^{\infty} \frac{e^{-\lambda}\lambda^k}{(k-1)!}$$

Verify using Poisson PMF (B-H Example 4.7.2)

$$
\begin{aligned}
E(X^2) &= \sum_{k=1}^{\infty} k^2 \frac{e^{-\lambda}\lambda^k}{k!} \\
&= \sum_{k=1}^{\infty} k \frac{e^{-\lambda}\lambda^k}{(k-1)!} = \sum_{k=2}^{\infty} (k-1) \frac{e^{-\lambda}\lambda^k}{(k-1)!} + \sum_{k=1}^{\infty} \frac{e^{-\lambda}\lambda^k}{(k-1)!} \\
&= \sum_{k=2}^{\infty} \frac{e^{-\lambda}\lambda^k}{(k-2)!} + \sum_{k=1}^{\infty} \frac{e^{-\lambda}\lambda^k}{(k-1)!}
\end{aligned}
$$

Verify using Poisson PMF (B-H Example 4.7.2)

$$
\begin{aligned}
E(X^2) &= \sum_{k=1}^{\infty} k^2 \frac{e^{-\lambda}\lambda^k}{k!} \\
&= \sum_{k=1}^{\infty} k \frac{e^{-\lambda}\lambda^k}{(k-1)!} = \sum_{k=2}^{\infty} (k-1) \frac{e^{-\lambda}\lambda^k}{(k-1)!} + \sum_{k=1}^{\infty} \frac{e^{-\lambda}\lambda^k}{(k-1)!} \\
&= \sum_{k=2}^{\infty} \frac{e^{-\lambda}\lambda^k}{(k-2)!} + \sum_{k=1}^{\infty} \frac{e^{-\lambda}\lambda^k}{(k-1)!} \\
&= \sum_{j=0}^{\infty} \frac{e^{-\lambda}\lambda^{j+2}}{j!} + \sum_{j=0}^{\infty} \frac{e^{-\lambda}\lambda^{j+1}}{j!}
\end{aligned}
$$

Verify using Poisson PMF (B-H Example 4.7.2)

$$
\begin{aligned}
E(X^2) &= \sum_{k=1}^{\infty} k^2 \frac{e^{-\lambda}\lambda^k}{k!} \\
&= \sum_{k=1}^{\infty} k \frac{e^{-\lambda}\lambda^k}{(k-1)!} = \sum_{k=2}^{\infty} (k-1) \frac{e^{-\lambda}\lambda^k}{(k-1)!} + \sum_{k=1}^{\infty} \frac{e^{-\lambda}\lambda^k}{(k-1)!} \\
&= \sum_{k=2}^{\infty} \frac{e^{-\lambda}\lambda^k}{(k-2)!} + \sum_{k=1}^{\infty} \frac{e^{-\lambda}\lambda^k}{(k-1)!} \\
&= \sum_{j=0}^{\infty} \frac{e^{-\lambda}\lambda^{j+2}}{j!} + \sum_{j=0}^{\infty} \frac{e^{-\lambda}\lambda^{j+1}}{j!} \\
&= \lambda^2 \underbrace{\sum_{j=0}^{\infty} \frac{e^{-\lambda}\lambda^j}{j!}}_{=1} + \lambda \underbrace{\sum_{j=0}^{\infty} \frac{e^{-\lambda}\lambda^j}{j!}}_{=1}
\end{aligned}
$$

Verify using Poisson PMF (B-H Example 4.7.2)

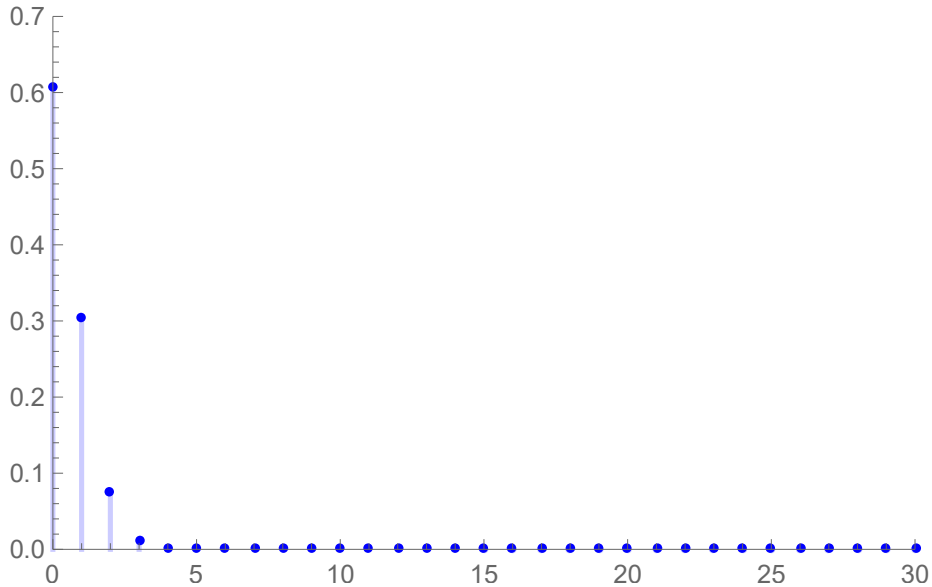
$$
\begin{aligned}
E(X^2) &= \sum_{k=1}^{\infty} k^2 \frac{e^{-\lambda}\lambda^k}{k!} \\
&= \sum_{k=1}^{\infty} k \frac{e^{-\lambda}\lambda^k}{(k-1)!} = \sum_{k=2}^{\infty} (k-1) \frac{e^{-\lambda}\lambda^k}{(k-1)!} + \sum_{k=1}^{\infty} \frac{e^{-\lambda}\lambda^k}{(k-1)!} \\
&= \sum_{k=2}^{\infty} \frac{e^{-\lambda}\lambda^k}{(k-2)!} + \sum_{k=1}^{\infty} \frac{e^{-\lambda}\lambda^k}{(k-1)!} \\
&= \sum_{j=0}^{\infty} \frac{e^{-\lambda}\lambda^{j+2}}{j!} + \sum_{j=0}^{\infty} \frac{e^{-\lambda}\lambda^{j+1}}{j!} \\
&= \lambda^2 \underbrace{\sum_{j=0}^{\infty} \frac{e^{-\lambda}\lambda^j}{j!}}_{=1} + \lambda \underbrace{\sum_{j=0}^{\infty} \frac{e^{-\lambda}\lambda^j}{j!}}_{=1} = \lambda^2 + \lambda
\end{aligned}
$$

Verify using Poisson PMF (B-H Example 4.7.2)

$$
\begin{aligned}
E(X^2) &= \sum_{k=1}^{\infty} k^2 \frac{e^{-\lambda}\lambda^k}{k!} \\
&= \sum_{k=1}^{\infty} k \frac{e^{-\lambda}\lambda^k}{(k-1)!} = \sum_{k=2}^{\infty} (k-1) \frac{e^{-\lambda}\lambda^k}{(k-1)!} + \sum_{k=1}^{\infty} \frac{e^{-\lambda}\lambda^k}{(k-1)!} \\
&= \sum_{k=2}^{\infty} \frac{e^{-\lambda}\lambda^k}{(k-2)!} + \sum_{k=1}^{\infty} \frac{e^{-\lambda}\lambda^k}{(k-1)!} \\
&= \sum_{j=0}^{\infty} \frac{e^{-\lambda}\lambda^{j+2}}{j!} + \sum_{j=0}^{\infty} \frac{e^{-\lambda}\lambda^{j+1}}{j!} \\
&= \lambda^2 \underbrace{\sum_{j=0}^{\infty} \frac{e^{-\lambda}\lambda^j}{j!}}_{=1} + \lambda \underbrace{\sum_{j=0}^{\infty} \frac{e^{-\lambda}\lambda^j}{j!}}_{=1} = \lambda^2 + \lambda
\end{aligned}
$$

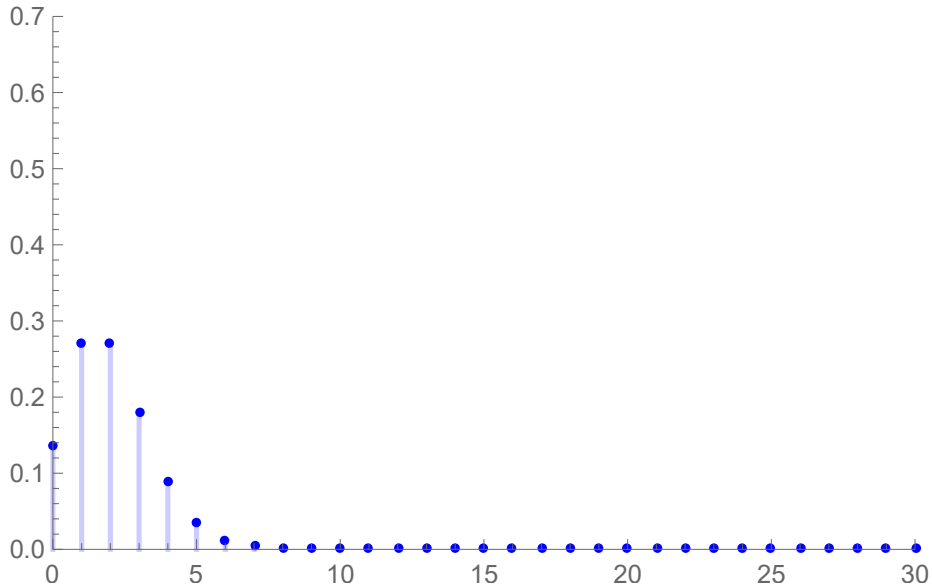
and $$\mathrm{Var}(X) = E(X^2) - (E(X))^2 = \lambda + \lambda^2 - \lambda^2 = \lambda$$

PMF vs expectation



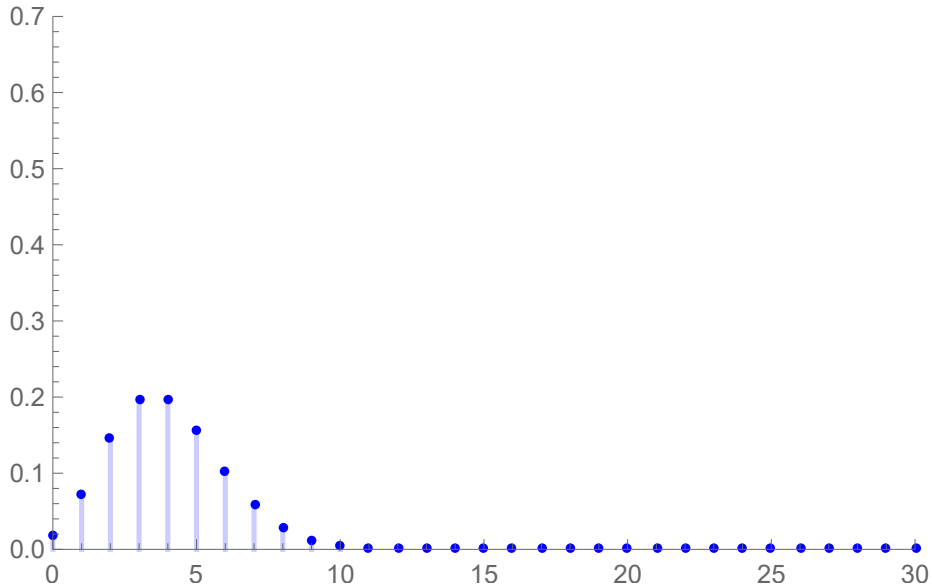
$\text{Pois}(0.5)$

PMF vs expectation



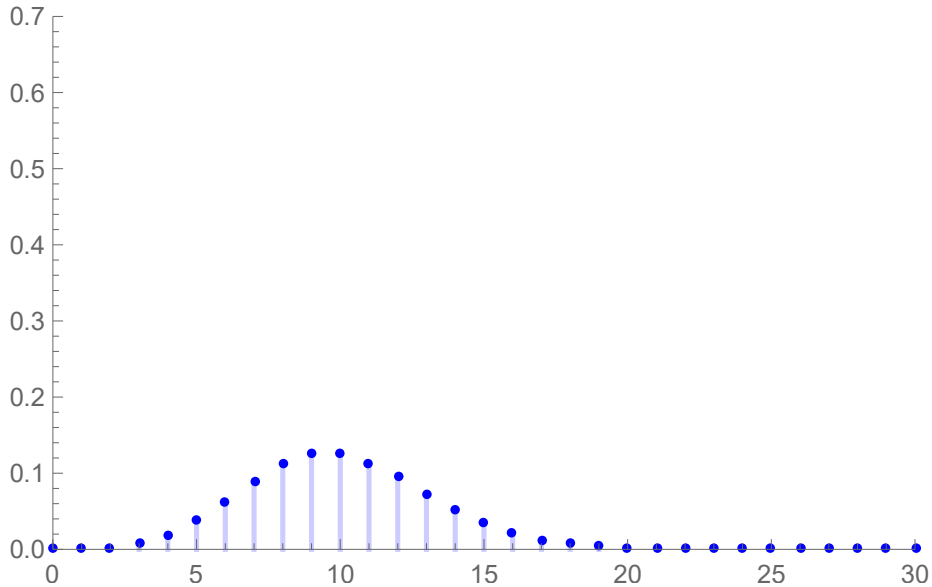
Pois(2)

PMF vs expectation



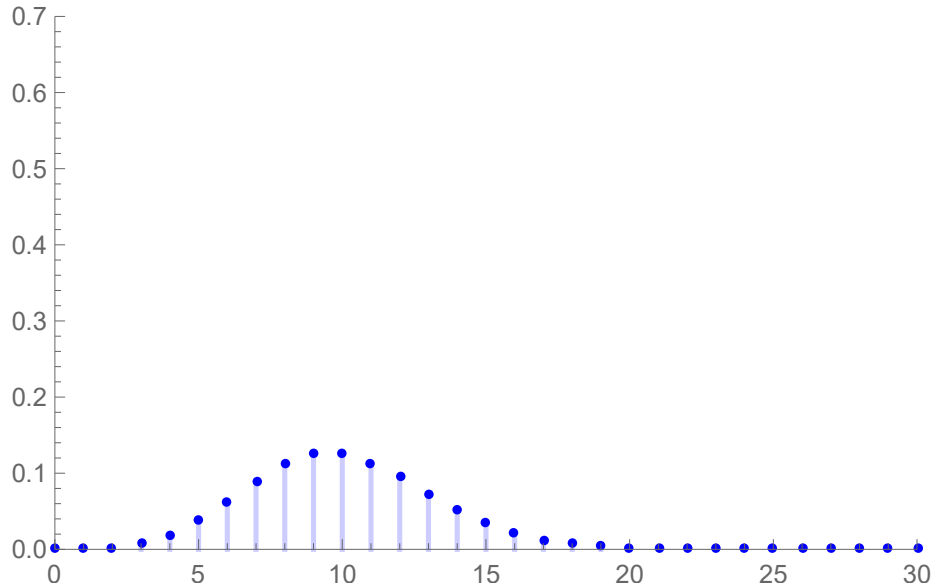
Pois(4)

PMF vs expectation



$\text{Pois}(10)$

PMF vs expectation



Pois(10)

As λ increases, PMF of Pois(λ) shifts to the right and becomes more spread out.

Geometric

$X \sim \text{Geom}(p)$. Then

$$\text{Var}(X) = \frac{1-p}{p^2}$$

Exercise (see B-H Example 4.6.4)

Deviation inequalities

Motivations

In statistics, we want to answer questions like the following:

1. Toss a fair coin 100 times. Getting at least 75 heads is probably unlikely. How unlikely is it?

Motivations

In statistics, we want to answer questions like the following:

1. Toss a fair coin 100 times. Getting at least 75 heads is probably unlikely. How unlikely is it?
2. Toss a coin 100 times. Turns out there are 75 heads, which indicates the coin is likely biased. How biased is it and how confident are we about our estimate?

Motivations

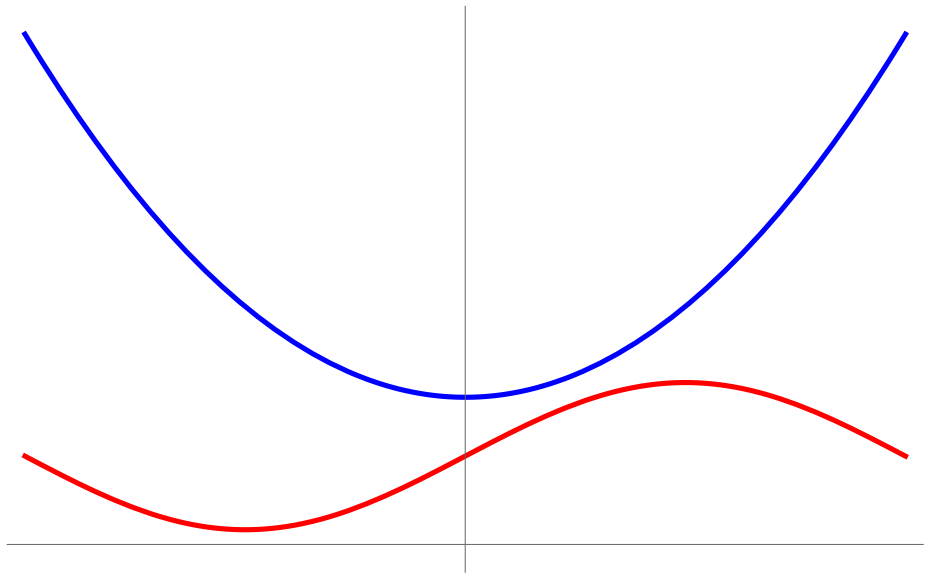
In statistics, we want to answer questions like the following:

1. Toss a fair coin 100 times. Getting at least 75 heads is probably unlikely. How unlikely is it?
2. Toss a coin 100 times. Turns out there are 75 heads, which indicates the coin is likely biased. How biased is it and how confident are we about our estimate?

Useful tools: deviation inequalities (tail bounds)

Preparation

If $f \leq g$ (i.e., $f(x) \leq g(x)$ for any x,)



then for any random variable X

$$E(f(X)) \leq E(g(X))$$

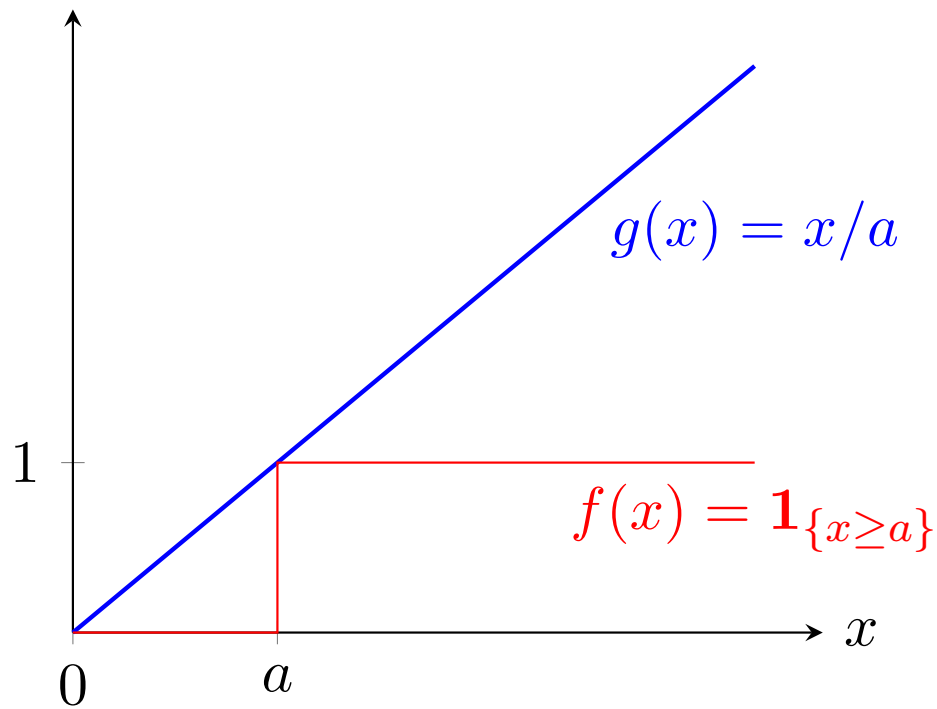
Markov's inequality

Let X be a non-negative random variable and $a > 0$. Then

$$P(X \geq a) \leq \frac{E(X)}{a}$$

Proof.

$f \leq g \implies E(f(X)) \leq E(g(X))$. □



Markov's inequality

Let X be a non-negative random variable and $a > 0$. Then

$$P(X \geq a) \leq \frac{E(X)}{a}$$

- Intuition: if the average salary is \$100, then the fraction of people earning \$300+ cannot exceed $1/3$.
- Lesson: A non-negative random variable cannot far exceed its expectation with high probability.

Application: coin

Question

Toss a fair coin 100 times. How unlikely is it to get at least 75 heads?

Application: coin

Question

Toss a fair coin 100 times. How unlikely is it to get at least 75 heads?

Let $X \sim \text{Bin}(100, 1/2)$. Then $E(X) = 50$

$$\text{Markov's inequality} \implies P(X \geq 75) \leq \frac{E(X)}{75} = \frac{2}{3}$$

Application: coin

Question

Toss a fair coin 100 times. How unlikely is it to get at least 75 heads?

Let $X \sim \text{Bin}(100, 1/2)$. Then $E(X) = 50$

$$\text{Markov's inequality} \implies P(X \geq 75) \leq \frac{E(X)}{75} = \frac{2}{3}$$

- Useful estimate, but probably too conservative...
- Can we do better (using more information, e.g., variance)?

Chebyshev's inequality

Let X be an arbitrary random variable and $d > 0$. Then

$$P(|X - E(X)| \geq d) \leq \frac{\mathrm{Var}(X)}{d^2}$$

Chebyshev's inequality

Let X be an arbitrary random variable and $d > 0$. Then

$$P(|X - E(X)| \geq d) \leq \frac{\mathrm{Var}(X)}{d^2}$$

Proof.

Let $\mu = E(X)$. LHS $= P((X - \mu)^2 \geq d^2)$. Apply Markov's inequality to the nonnegative random variable $(X - \mu)^2$. Or, see picture below: $\square$

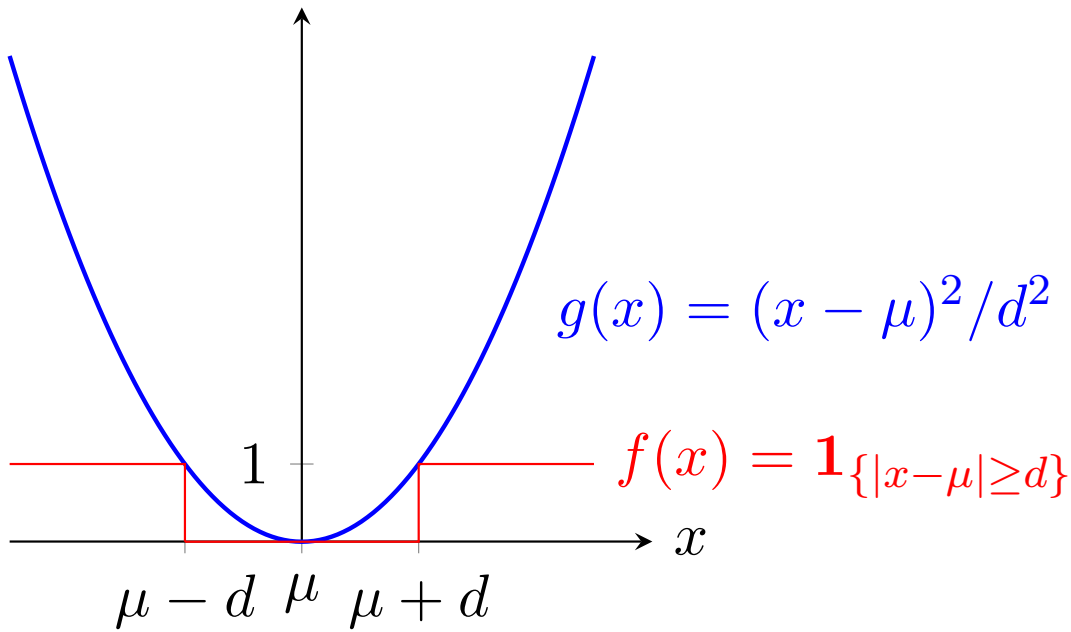
Chebyshev's inequality

Let X be an arbitrary random variable and $d > 0$. Then

$$P(|X - E(X)| \geq d) \leq \frac{\mathrm{Var}(X)}{d^2}$$

Proof.

Let $\mu = E(X)$. LHS $= P((X-\mu)^2 \geq d^2)$. Apply Markov's inequality to the nonnegative random variable $(X-\mu)^2$. Or, see picture below: $\square$



Application: coin

Question

Toss a fair coin 100 times. How unlikely is it to get at least 75 heads?

Application: coin

Question

Toss a fair coin 100 times. How unlikely is it to get at least 75 heads?

Let $X \sim \text{Bin}(100, 1/2)$. Then $E(X) = 50$ and $\text{Var}(X) = 25$

$$\text{Chebyshev} \implies P(X \geq 75) \leq P(|X - 50| \geq 25) \leq \frac{25}{25^2} = 4\%$$

Application: coin

Question

Toss a fair coin 100 times. How unlikely is it to get at least 75 heads?

Let $X \sim \text{Bin}(100, 1/2)$. Then $E(X) = 50$ and $\text{Var}(X) = 25$

$$\text{Chebyshev} \implies P(X \geq 75) \leq P(|X - 50| \geq 25) \leq \frac{25}{25^2} = 4\%$$

- Much better than Markov (we used both mean and variance)
- Actually value: 2.8×10^{-7}

Application: coin

Question

Toss a fair coin 100 times. How unlikely is it to get at least 75 heads?

Let $X \sim \text{Bin}(100, 1/2)$. Then $E(X) = 50$ and $\text{Var}(X) = 25$

$$\text{Chebyshev} \implies P(X \geq 75) \leq P(|X - 50| \geq 25) \leq \frac{25}{25^2} = 4\%$$

- Much better than Markov (we used both mean and variance)
- Actually value: 2.8×10^{-7}
- Central limit theorem (Lec 16) gives accurate estimate: 2.9×10^{-7}

Standard deviation

Equivalent formulation of Chebyshev:

$$P(|X - E(X)| \geq C \cdot \sigma_X) \leq \frac{1}{C^2}$$

Application: Confidence Interval (CI)

Question

Toss a coin 100 times. Turns out there are 75 heads. How biased is the coin and how confident are we about our estimate?

Application: Confidence Interval (CI)

Question

Toss a coin 100 times. Turns out there are 75 heads. How biased is the coin and how confident are we about our estimate?

- Let $X \sim \text{Bin}(n, p)$, where $n = 100$ and the bias p is unknown.
- Observe $X = 75$.

Application: Confidence Interval (CI)

Question

Toss a coin 100 times. Turns out there are 75 heads. How biased is the coin and how confident are we about our estimate?

- Let $X \sim \text{Bin}(n, p)$, where $n = 100$ and the bias p is unknown.
- Observe $X = 75$.
- Empirical frequency of heads is

$$\hat{p} = \frac{X}{n} = 0.75,$$

  a reasonable estimate of p.

Application: Confidence Interval (CI)

Question

Toss a coin 100 times. Turns out there are 75 heads. How biased is the coin and how confident are we about our estimate?

- Let $X \sim \text{Bin}(n, p)$, where $n = 100$ and the bias p is unknown.
- Observe $X = 75$.
- Empirical frequency of heads is

$$\hat{p} = \frac{X}{n} = 0.75,$$

  a reasonable estimate of p.
- More refined estimate: confidence interval. If

$$P(p \in [\hat{p} - \epsilon, \hat{p} + \epsilon]) \geq c, \quad \text{for any } p$$

  we say $[\hat{p} - \epsilon, \hat{p} + \epsilon]$ is a confidence interval with confidence level c, e.g. $c = 95\%$.

Examples

- Conflicting goals: higher confidence and narrower interval.
  - $[0, 1]$ is a CI of confidence level 100%, but not very useful

Examples

- Conflicting goals: higher confidence and narrower interval.
  - $[0, 1]$ is a CI of confidence level 100%, but not very useful
  - Need a tool to assess the chance of being off: let's use Chebyshev inequality

Examples

- Conflicting goals: higher confidence and narrower interval.
  - $[0, 1]$ is a CI of confidence level 100%, but not very useful
  - Need a tool to assess the chance of being off: let's use Chebyshev inequality
- $\mathrm{Var}(\hat{p}) = \mathrm{Var}(X/n) = \frac{1}{n^2}\mathrm{Var}(X) = \frac{p(1-p)}{n}$
- Thus

$$P(p \notin [\hat{p} - \epsilon, \hat{p} + \epsilon]) = P(|\hat{p} - p| > \epsilon) \leq \frac{\mathrm{Var}(\hat{p})}{\epsilon^2} = \frac{p(1-p)}{n\epsilon^2} \leq \frac{1}{4n\epsilon^2}$$

Examples

- Conflicting goals: higher confidence and narrower interval.
  - $[0, 1]$ is a CI of confidence level 100%, but not very useful
  - Need a tool to assess the chance of being off: let's use Chebyshev inequality
- $\mathrm{Var}(\hat{p}) = \mathrm{Var}(X/n) = \frac{1}{n^2}\mathrm{Var}(X) = \frac{p(1-p)}{n}$
- Thus

$$P(p \notin [\hat{p} - \epsilon, \hat{p} + \epsilon]) = P(|\hat{p} - p| > \epsilon) \leq \frac{\mathrm{Var}(\hat{p})}{\epsilon^2} = \frac{p(1-p)}{n\epsilon^2} \leq \frac{1}{4n\epsilon^2}$$

- Since $n = 100$, if we take $\epsilon = 0.1$, then $\frac{1}{4n\epsilon^2} = 0.25$.
- Upon observing $X = 75$, $[0.75 \pm 0.1]$ is a CI of level 75%

What to do if we desire higher confidence?

- Want confidence level 99%. Then

$$P(p \notin [\hat{p} - \epsilon, \hat{p} + \epsilon]) \leq \frac{1}{4n\epsilon^2} \leq 1\% \xrightarrow{n=100} \boxed{\epsilon \geq 0.5}$$

  therefore the CI is $[\hat{p} \pm 0.5]$ — too wide to be very useful.

What to do if we desire higher confidence?

- Want confidence level 99%. Then

$$P(p \notin [\hat{p} - \epsilon, \hat{p} + \epsilon]) \leq \frac{1}{4n\epsilon^2} \leq 1\% \xrightarrow{n=100} \boxed{\epsilon \geq 0.5}$$

  therefore the CI is $[\hat{p} \pm 0.5]$ — too wide to be very useful.

- How to get a CI $[\hat{p} \pm 0.1]$ of level 99%?

What to do if we desire higher confidence?

- Want confidence level 99%. Then

$$P(p \notin [\hat{p} - \epsilon, \hat{p} + \epsilon]) \leq \frac{1}{4n\epsilon^2} \leq 1\% \xrightarrow{n=100} \boxed{\epsilon \geq 0.5}$$

  therefore the CI is $[\hat{p} \pm 0.5]$ — too wide to be very useful.

- How to get a CI $[\hat{p} \pm 0.1]$ of level 99%? Increase the sample size!

$$\frac{1}{4n\epsilon^2} \leq 1\% \xrightarrow{\epsilon=0.1} \boxed{n \geq 2500}$$

What to do if we desire higher confidence?

- Want confidence level 99%. Then

$$P(p \notin [\hat{p} - \epsilon, \hat{p} + \epsilon]) \leq \frac{1}{4n\epsilon^2} \leq 1\% \xRightarrow{n=100} \boxed{\epsilon \geq 0.5}$$

  therefore the CI is $[\hat{p} \pm 0.5]$ — too wide to be very useful.

- How to get a CI $[\hat{p} \pm 0.1]$ of level 99%? Increase the sample size!

$$\frac{1}{4n\epsilon^2} \leq 1\% \xRightarrow{\epsilon=0.1} \boxed{n \geq 2500}$$

  (Better tools than Chebyshev $\implies n \geq 169$)

- Lesson: More data leads to more accurate estimate.
  Here $\mathrm{Var}(\hat{p}) \propto \frac{1}{n}$.

Same story: Polling

Same story: Polling

- Poll n people out of a population uniformly at random with replacements. Let X be the number of surveyed people who would vote for Biden.
- Then $X \sim \text{Bin}(n, p)$, where $p =$ fraction of people in the entire population voting for Biden.
- Suppose we want $[\hat{p} \pm 0.01]$ to have confidence 95%

$$\frac{1}{4n\epsilon^2} \leq 5\% \xrightarrow{\epsilon=0.01} \boxed{n \geq 50000}$$

Same story: Polling

- Poll n people out of a population uniformly at random with replacements. Let X be the number of surveyed people who would vote for Biden.
- Then $X \sim \text{Bin}(n, p)$, where $p =$ fraction of people in the entire population voting for Biden.
- Suppose we want $[\hat{p} \pm 0.01]$ to have confidence 95%

$$\frac{1}{4n\epsilon^2} \leq 5\% \xrightarrow{\epsilon=0.01} \boxed{n \geq 50000}$$

Surprise

The number of people needed to poll to reach a desired accuracy and confidence does NOT depend on the population size!!!

Same story: Polling

- Poll n people out of a population uniformly at random with replacements. Let X be the number of surveyed people who would vote for Biden.
- Then $X \sim \text{Bin}(n, p)$, where $p =$ fraction of people in the entire population voting for Biden.
- Suppose we want $[\hat{p} \pm 0.01]$ to have confidence 95%

$$\frac{1}{4n\epsilon^2} \leq 5\% \xrightarrow{\epsilon=0.01} \boxed{n \geq 50000}$$

Surprise

The number of people needed to poll to reach a desired accuracy and confidence does NOT depend on the population size!!!

So what's the catch?

Same story: Polling

- Poll n people out of a population uniformly at random with replacements. Let X be the number of surveyed people who would vote for Biden.
- Then $X \sim \text{Bin}(n, p)$, where $p =$ fraction of people in the entire population voting for Biden.
- Suppose we want $[\hat{p} \pm 0.01]$ to have confidence 95%

$$\frac{1}{4n\epsilon^2} \leq 5\% \xrightarrow{\epsilon=0.01} \boxed{n \geq 50000}$$

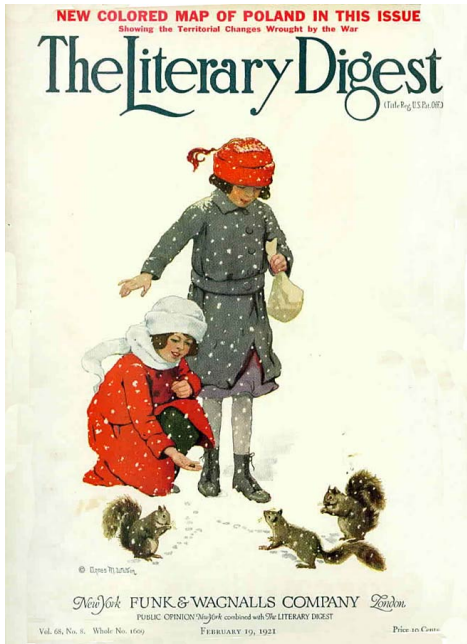
Surprise

The number of people needed to poll to reach a desired accuracy and confidence does NOT depend on the population size!!!

So what's the catch?

True story

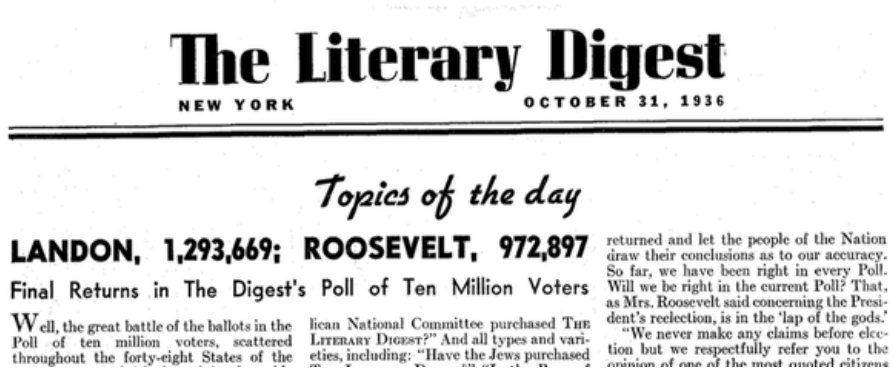
Polling before phone and internet...



By mailing out millions of postcards to readers and simply counting the returns, *The Literary Digest* correctly predicted four presidential elections in a roll (1920,1924,1928,1932)

1936: Landon vs Roosevelt

- Literary Digest predicted Landon would win by 57% based on a sample of size 2.3 million



The Literary Digest

NEW YORK OCTOBER 31, 1936

Topics of the day

LANDON, 1,293,669; ROOSEVELT, 972,897

Final Returns in The Digest's Poll of Ten Million Voters

Well, the great battle of the ballots in the Poll of ten million voters, scattered throughout the forty-eight States of the

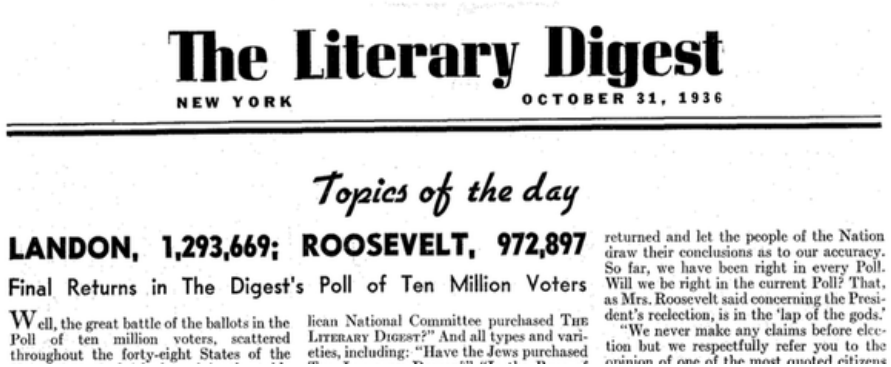
lican National Committee purchased THE LITERARY DIGEST?" And all types and varieties, including: "Have the Jews purchased

returned and let the people of the Nation draw their conclusions as to our accuracy. So far, we have been right in every Poll. Will we be right in the current Poll? That, as Mrs. Roosevelt said concerning the President's reelection, is in the 'lap of the gods.'

"We never make any claims before election but we respectfully refer you to the

1936: Landon vs Roosevelt

- Literary Digest predicted Landon would win by 57% based on a sample of size 2.3 million



The Literary Digest

NEW YORK OCTOBER 31, 1936

Topics of the day

LANDON, 1,293,669; ROOSEVELT, 972,897

Final Returns in The Digest's Poll of Ten Million Voters

Well, the great battle of the ballots in the Poll of ten million voters, scattered throughout the forty-eight States of the

lican National Committee purchased THE LITERARY DIGEST?" And all types and varieties, including: "Have the Jews purchased

returned and let the people of the Nation draw their conclusions as to our accuracy. So far, we have been right in every Poll. Will we be right in the current Poll? That, as Mrs. Roosevelt said concerning the President's reelection, is in the 'lap of the gods.'

"We never make any claims before election but we respectfully refer you to the

- George Gallup's company conducted a survey of much smaller size 50000 and predicted FDR would win by 56%

As it turns out...

- FDR won by a landslide (61%)

As it turns out...

- FDR won by a landslide (61%)
- Literary Digest had their final issue in 1938

As it turns out...

- FDR won by a landslide (61%)
- Literary Digest had their final issue in 1938
- Gallup went on to make a career in polling

As it turns out...

- FDR won by a landslide (61%)
- Literary Digest had their final issue in 1938
- Gallup went on to make a career in polling
- "The Poll That Changed Polling"

As it turns out...

- FDR won by a landslide (61%)
- Literary Digest had their final issue in 1938
- Gallup went on to make a career in polling
- "The Poll That Changed Polling"

As it turns out...

- FDR won by a landslide (61%)
- Literary Digest had their final issue in 1938
- Gallup went on to make a career in polling
- "The Poll That Changed Polling"

So why was Literary Digest so wrong?

As it turns out...

- FDR won by a landslide (61%)
- Literary Digest had their final issue in 1938
- Gallup went on to make a career in polling
- "The Poll That Changed Polling"

So why was Literary Digest so wrong? Selection bias!

Why was Literary Digest so wrong?

- Unbeknownst to the magazine, their subscribers tend to be more affluent Americans who favor Republicans.
- Gallup noticed this bias, and polled a much smaller but demographically representative of the population

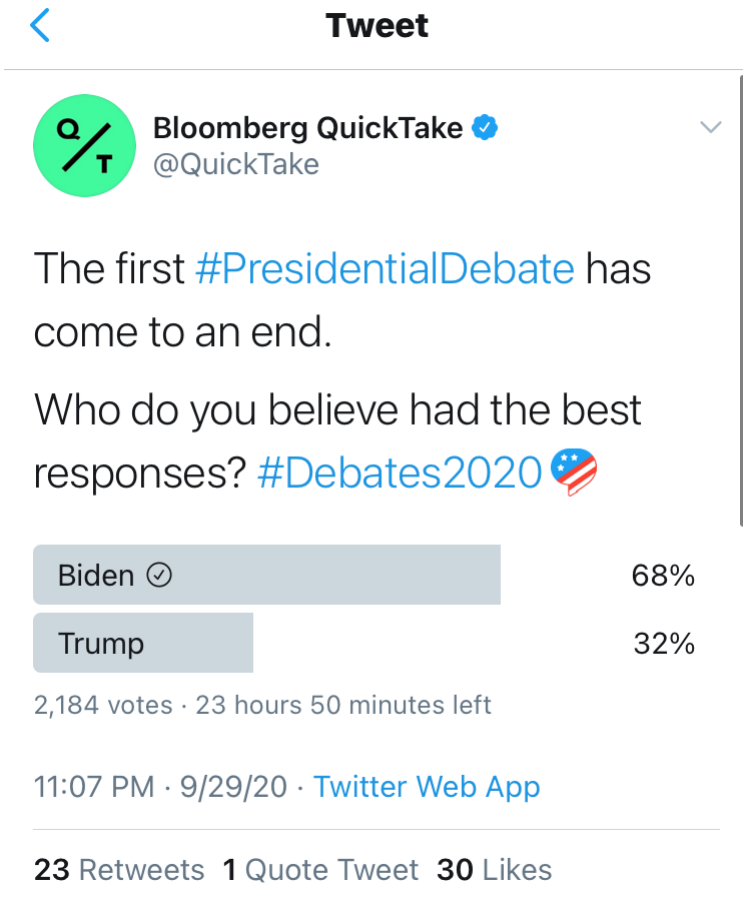
Why was Literary Digest so wrong?

- Unbeknownst to the magazine, their subscribers tend to be more affluent Americans who favor Republicans.
- Gallup noticed this bias, and polled a much smaller but demographically representative of the population
- To drive his point home, he even predicted the result of the Literary Digest poll to within about 1%

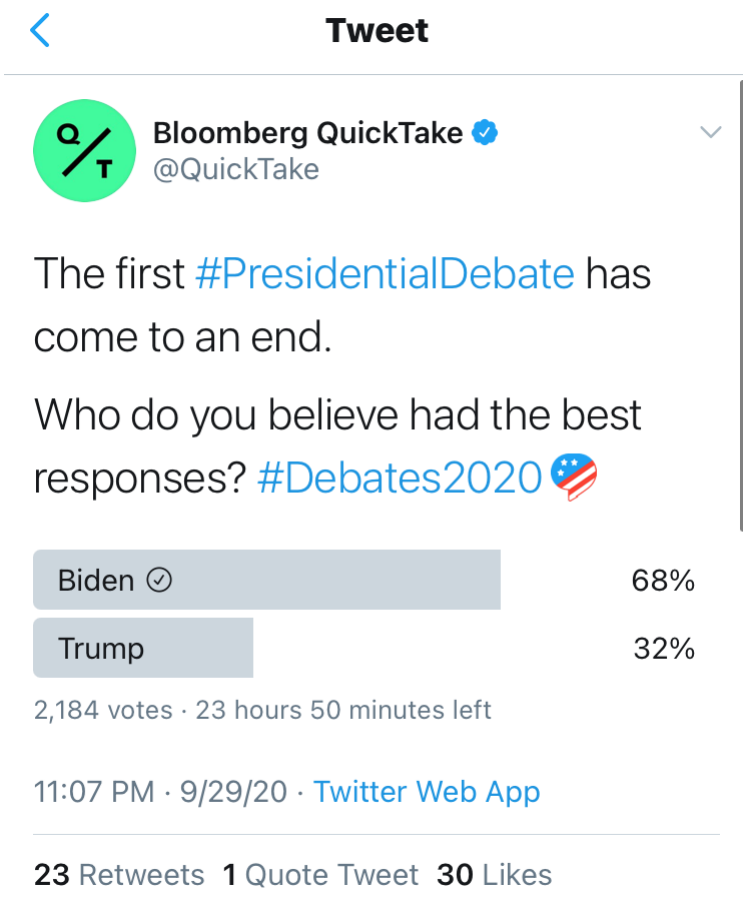
Why was Literary Digest so wrong?

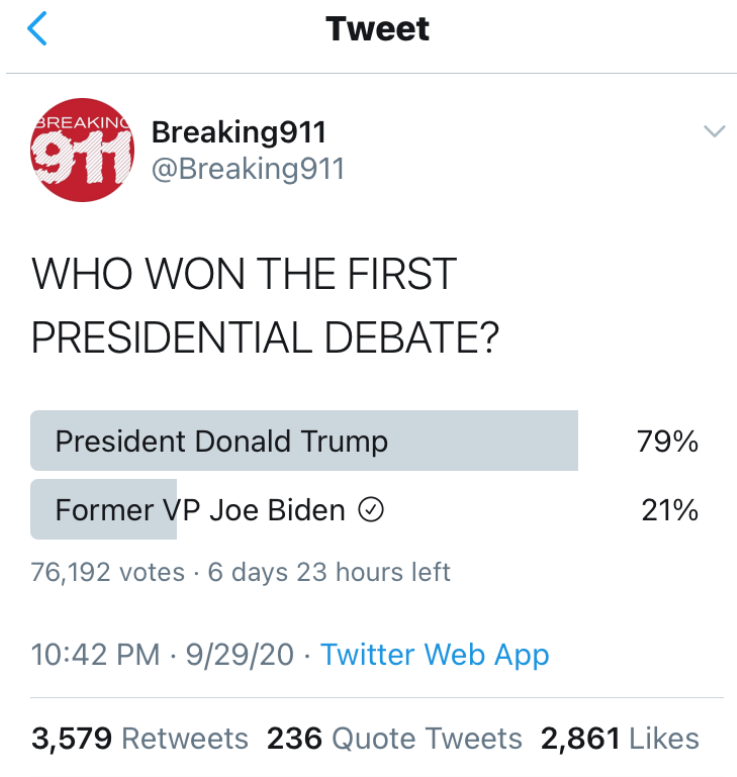
- Unbeknownst to the magazine, their subscribers tend to be more affluent Americans who favor Republicans.
- Gallup noticed this bias, and polled a much smaller but demographically representative of the population
- To drive his point home, he even predicted the result of the Literary Digest poll to within about 1%
- There is lots of science behind opinion polling and survey sampling
  - Selection bias
  - Non-response bias
  - Predict turnout
  - ...

Contemporary story of selection bias



Contemporary story of selection bias

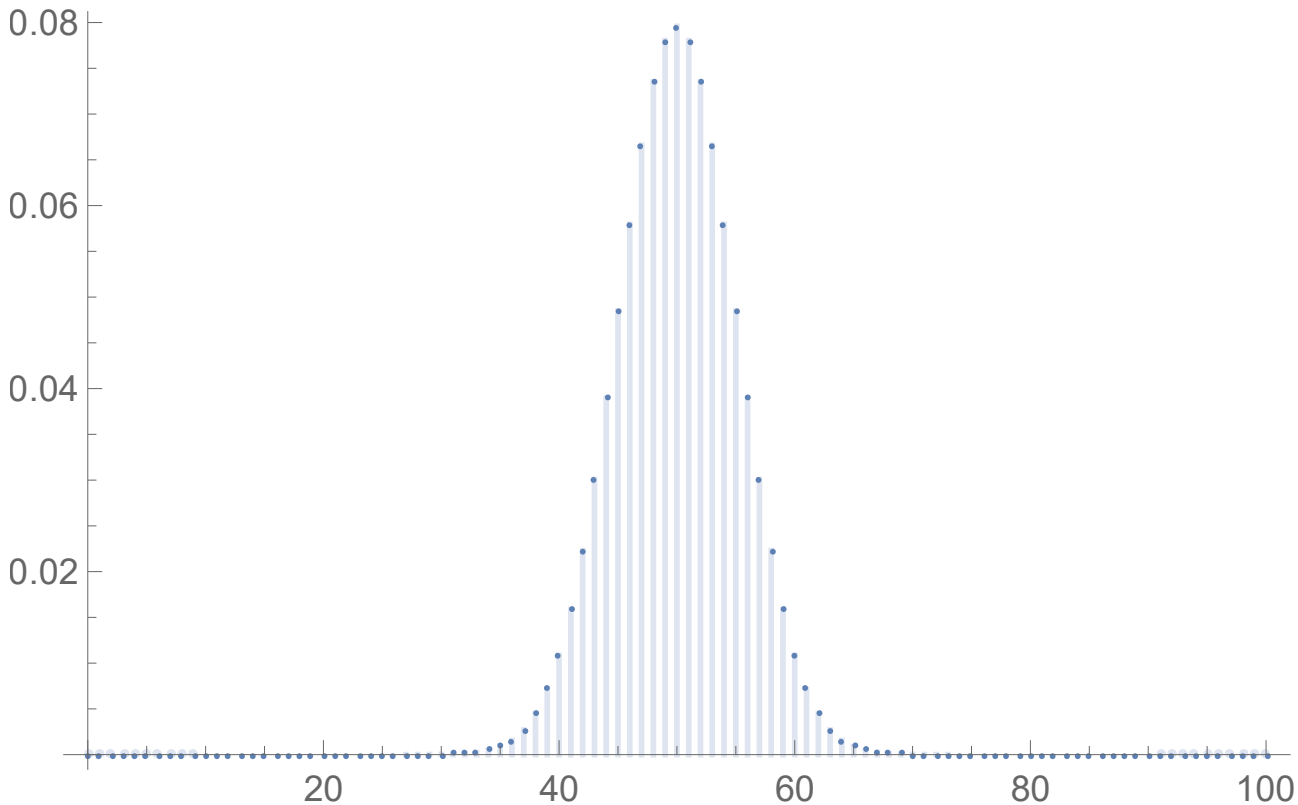




Echo chamber effect in social media

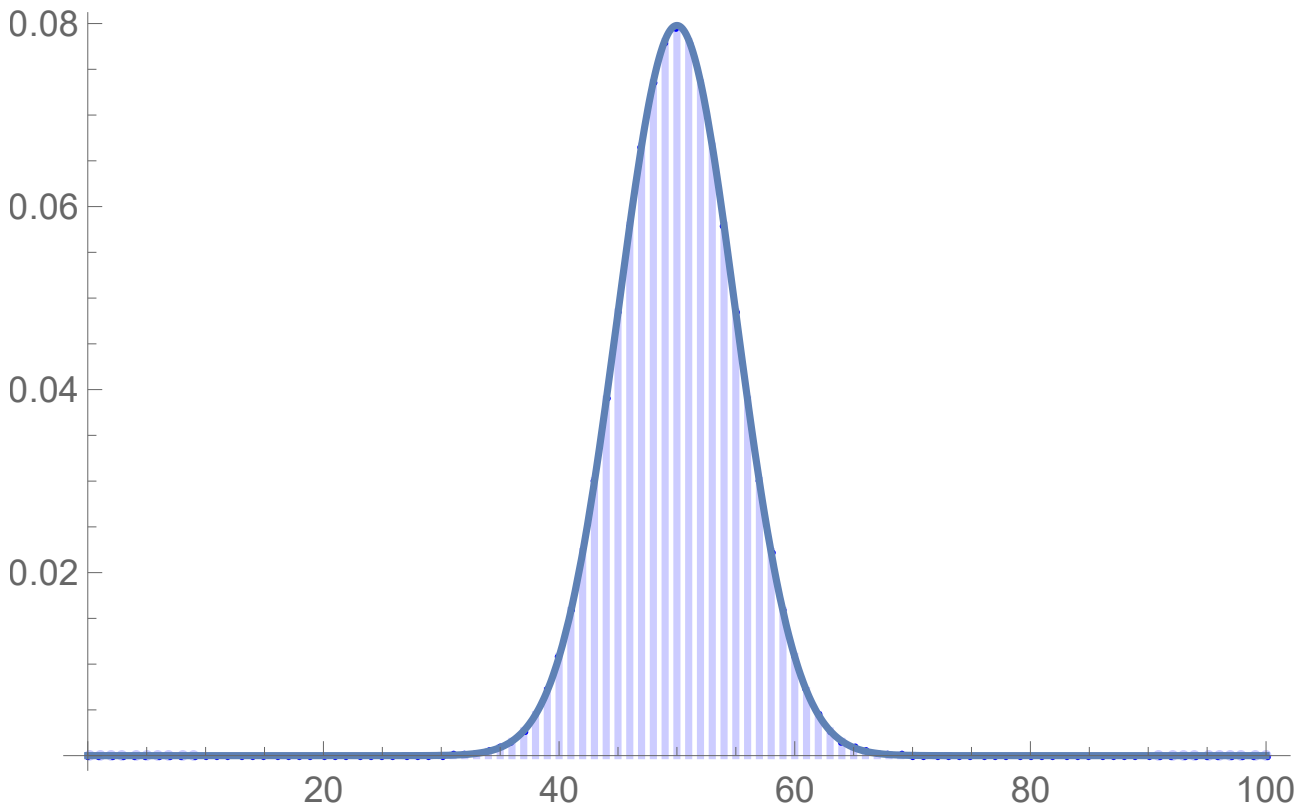
A bit preview

How to obtain better estimate than Chebyshev:



A bit preview

How to obtain better estimate than Chebyshev:



A bit preview

How to obtain better estimate than Chebyshev:

continuous approximation!