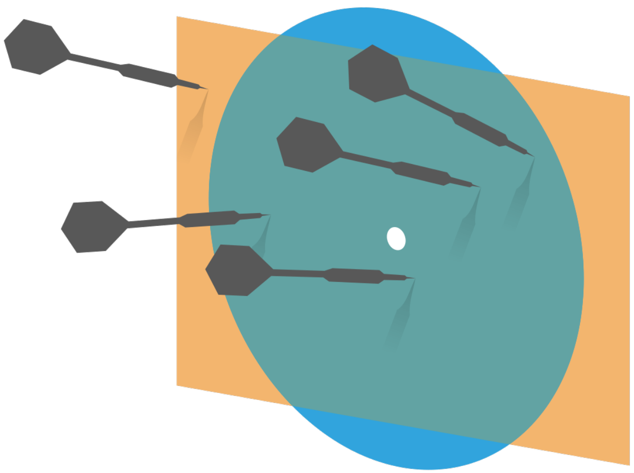
Probability and Statistics for Computer Science



Credit: wikipedia

"The weak law of large number gives us a very valuable way of thinking about expectations." ---Prof. Forsythe

* HW4 is released! Due on 2/27.

It has Python coding prob. so

Start early please!!

Discussion on Wed. will show demo.

* Added material on the course

website.

Last time

- Random Variable
  - *Expected value*
  - *Variance & covariance*

Content

- Random Variable
  - *Review with questions* Expectations
  - *The weak law of large numbers*

Form pair with your neighbor and get the sum of the die of your individual roll and write on the paper to submit through the student collectors.

The number of each roll is the Missing number of the down face.

Example:

$$\boxed{\begin{array}{ll} A: 3 & B: 2 \\ S: 3+2=5 & \end{array}}$$

If you don't have a team, do it twice, submit the sum.

Example:

$$\boxed{\begin{array}{ll} A_1: 2 & A_2: 1 \\ S: 2+1=3 & \end{array}}$$

Expected value

- The **expected value** (or **expectation**) of a random variable X is

$$E[X] = \sum_x x \underline{P(x)}$$

The expected value is a weighted sum of the values X can take

Linearity of Expectation

- For random variables X and Y and constants k,c
  - Scaling property
    $$E[kX] = kE[X]$$
  - Additivity
    $$E[X+Y] = E[X] + E[Y]$$
  - And $E[kX + c] = kE[X] + c$

Expected value of a function of X

- If f is a function of a random variable X, then $Y = f(X)$ is a random variable too
- The expected value of $Y = f(X)$ is

$$E[Y] = E[f(X)] = \sum_x f(x)P(x)$$

Q:

What is $E[E[X]]$?

A. $E[X]$
B. 0
C. Can’t be sure

Q:

What is $E[E[X]]$?

A. $E[X]$
B. 0
C. Can’t be sure

Probability distribution

- Given the random variable ***X***, what is

E[2|***X***| +1]?

$E[f(x)]$

$f(x) = 2|x| + 1$

A. 0
B. 1
C. 2
D. $\boxed{3}$
E. 5

if $x = -1$

$f(x) = 3$

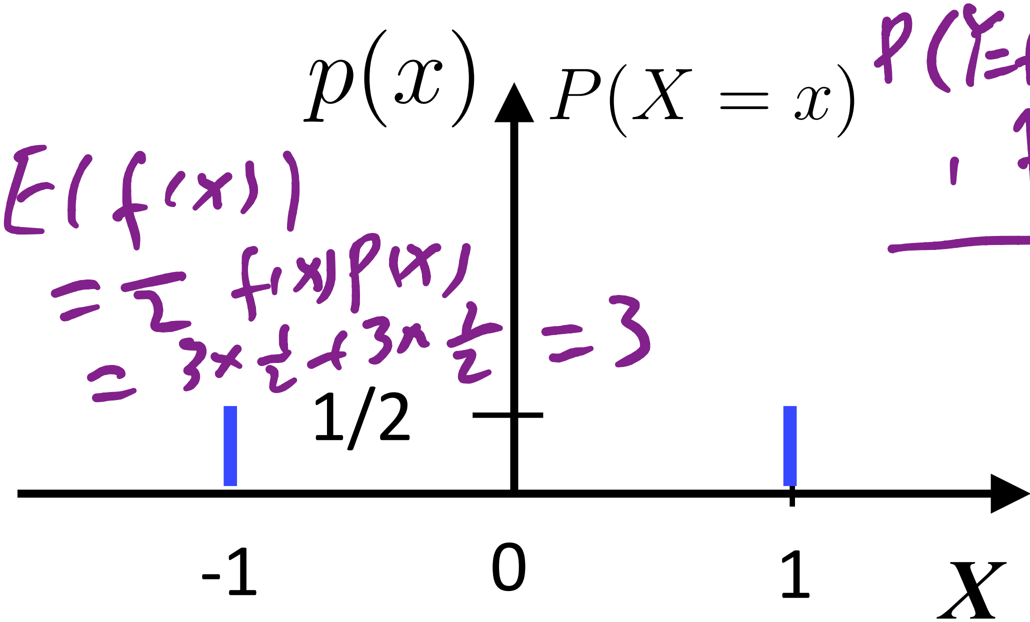
if $x = 1$

$f(x) = 3$

$E(f(x))$

$= \sum f(x) p(x)$

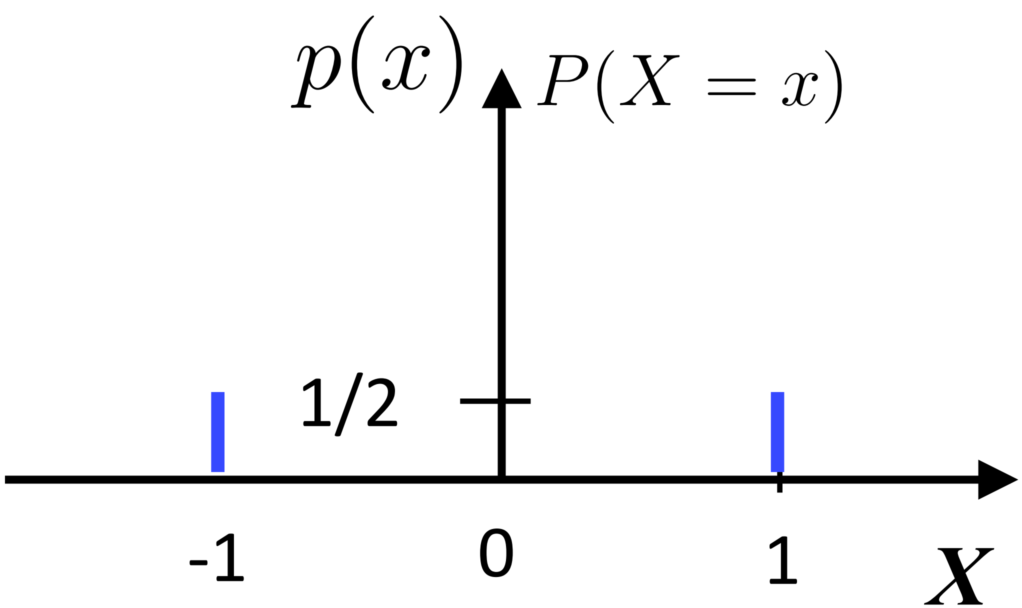
$= 3 \times \frac{1}{2} + 3 \times \frac{1}{2} = 3$



$P(Y = f(x))$

Q.

- Given the random variable ***X***, what is E[2|***X***| +1]?



A. 0
B. 1
C. 2
D. 3
E. 5

Q.

- Given the random variable ***X***, what is

E[2|***X***| +1]?

E[ax+b]

= aE[x]+b

$$E[|X|] = 1 \times \frac{1}{2} + 1 \times \frac{1}{2} = 1$$

E(|x|)

$= 1\times\frac{1}{2}$

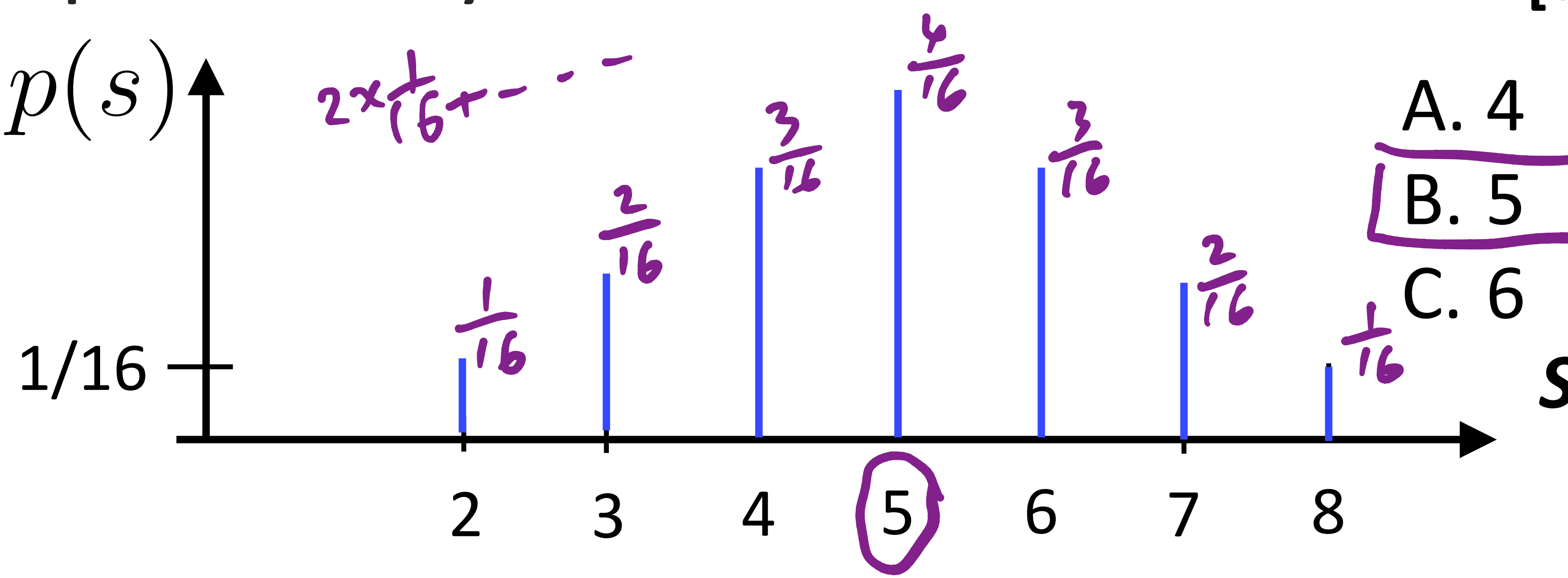
$+1\times\frac{1}{2} = 1$

$$E[2|X| + 1] = \underbrace{2E[|X|] + 1}_{} = \boxed{3}$$

Q.

- Give the random variable ***S*** *in the 4-sided die, whose range is* {2,3,4,5,6,7,8}, probability distribution of ***S*.**

What is **E[S]** ?



A. 4

$\boxed{\text{B. 5}}$

C. 6

A neater expression for variance

- Variance of Random Variable X is defined as:

$f(x) = (X - E[X])^2$

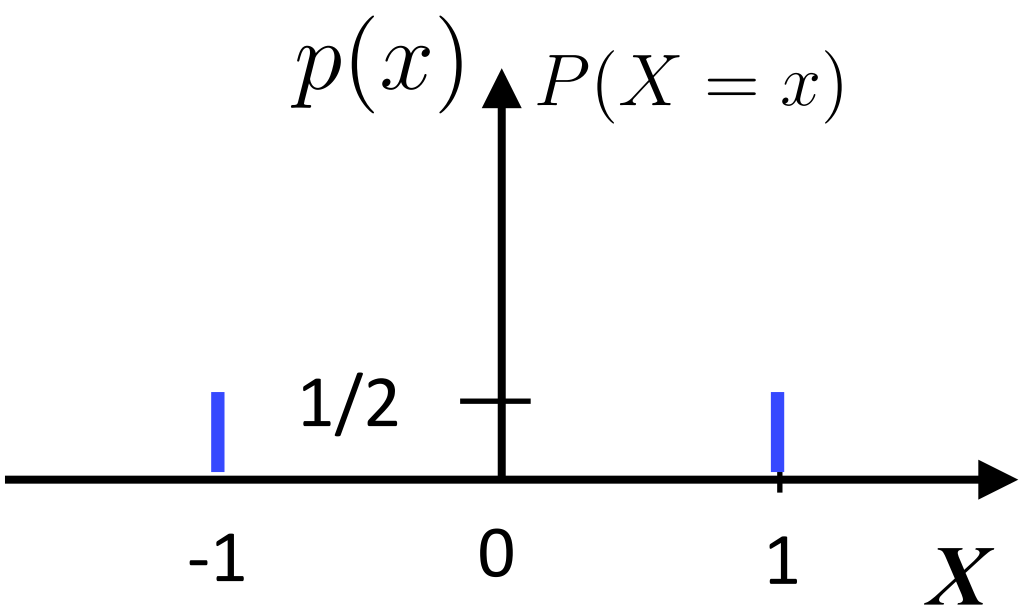
$$var[X] = E[\underline{(X - E[X])^2}]$$

- It's the same as:

$$\boxed{\underline{var[X] = E[X^2] - E[X]^2}}$$

Q.

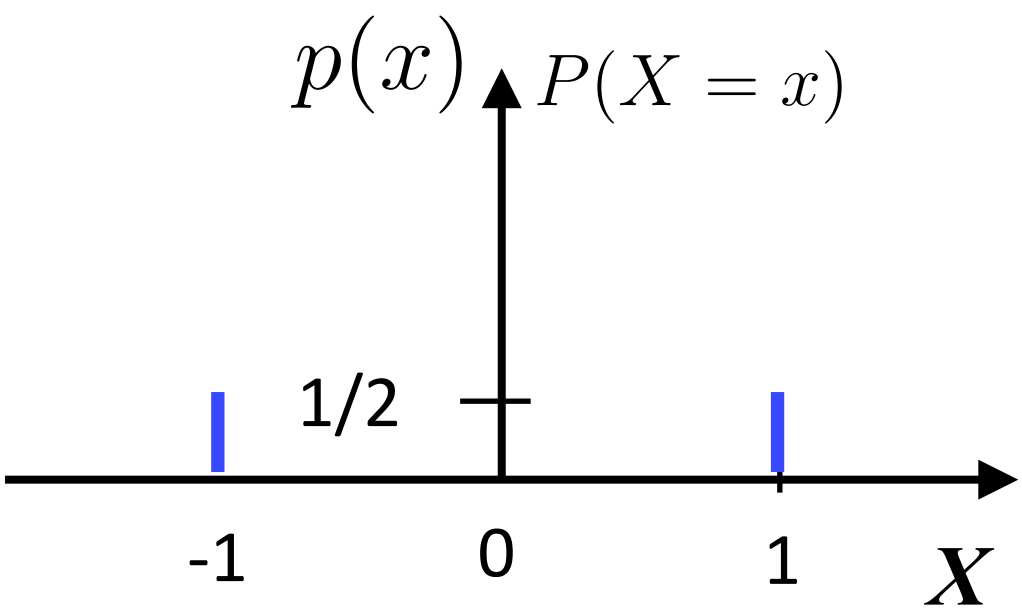
- Given the random variable ***X***, what is var[2|***X***| +1]?



A. 0
B. 1
C. 2
D. 3
E. -1

Q.

- Given the random variable ***X***, what is var[2|***X***| +1]?



A. 0

B. 1

C. 2

D. 3

E. -1

Q.

- Given the random variable ***X***, what is

var[2|***X***| +1]?

$$E[|X|] = 1 \times \frac{1}{2} + 1 \times \frac{1}{2} = 1$$

$$E[X^2] = 1 \times \frac{1}{2} + 1 \times \frac{1}{2} = 1$$

$$E[2|X| + 1] = 2E[|X|] + 1 = 3$$

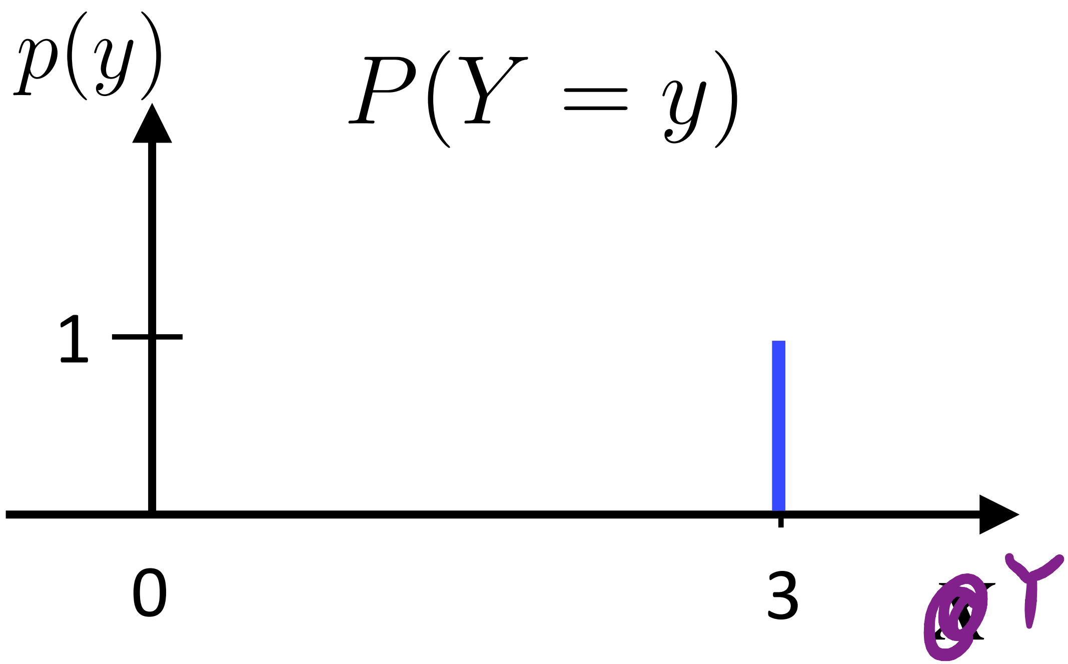
$$var[2|X| + 1] = E[(2|X| + 1)^2] - (E[2|X| + 1])^2$$

$$= E[4X^2 + 4|X| + 1] - 3^2$$

$$= 4 \times 1 + 4 \times 1 + 1 - 9 = 0$$

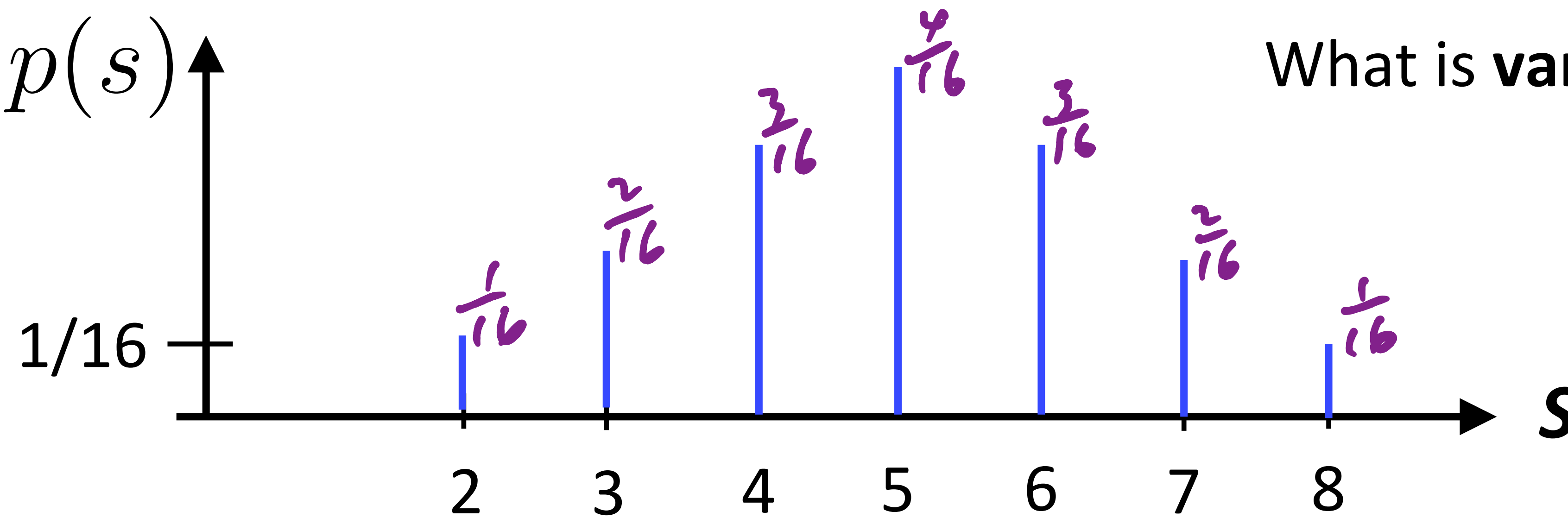
Q.

- Given the random variable ***X***, what is var[2|***X***| +1]? Let ***Y*** = 2|***X***|+1



Q.

- Give the random variable ***S*** *in the 4-sided die, whose range is* {2,3,4,5,6,7,8}, probability distribution of ***S.***



What is **var[S]** ?

$var(S)$

$$= \frac{1}{16}\cdot(2-5)^2 + \frac{2}{16}(3-5)^2 + \frac{3}{16}(4-5)^2 + \frac{4}{16}(5-5)^2 +$$

$$\frac{3}{16}\cdot(6-5)^2 + \frac{2}{16}(7-5)^2 + \frac{1}{16}(8-5)^2$$

$$= \frac{1}{8}\cdot 3^2 + \frac{2}{8}\cdot 2^2 + \frac{3}{8}\cdot 1^2$$

$$= \frac{9}{8} + 1 + \frac{3}{8}$$

$$\underline{= \frac{5}{2}}$$

Q:

- Which of the following is NOT generally true about two independent random variables X and Y?

  A. E[X+Y] = E[X] +E[Y]

  B. var[X+Y] = var[X]+var[Y]

  C. E[XY] = E[X]E[Y]

  D. corr(X,Y) = 0

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

$\text{var} = \text{std}^2$

$\text{var}[X+Y]$
$= \text{var}[X] + \text{var}[Y] + 2\,\text{cov}(X,Y)$, with $\text{cov}(X,Y) \to 0$

$\text{cov}[X,Y] = 0$

$\because \text{cov}(X,Y) = E[XY] - E[X]E[Y]$

Q:

- Which of the following is NOT generally true about two independent random variables X and Y?

  A. E[X+Y] = E[X] +E[Y]

  B. var[X+Y] = var[X]+var[Y]

  C. E[XY] = E[X]E[Y]

  D. corr(X,Y) = 0

  E. std[X+Y] = std[X]+std[Y]

Content

- Random Variable
  - *Review with questions*
  - ***The weak law of large numbers***

Towards the weak law of large numbers

- The weak law says that if we repeat an experiment many times, the average of the observations will "converge" to the expected value
- For example, if you repeat the profit example, the average earning will "converge" to E[X]=20p-10
- The weak law justifies using simulations (instead of calculation) to estimate the expected values of random variables

* Indicator function
* Markov Inequality
* Chebyshev Inequality
* The weak law of large numbers

Markov inequality

- The inequality that was the foundation of many probabilistic theories
- Discovered by Andrei Markov who also invented Markov Chain model (ch 14)

Andrei Markov
1856 - 1922

Indicator functions

- An indicator function for an event A is a function of x such that

$$\mathbb{I}_{[A]}(x) = \begin{cases} \underline{1} & event\ occurs\ for\ the\ value\ x \\ \underline{0} & otherwise \end{cases}$$

x could be a range of possible values

- The expected value of the indicator function is the probability of event A

$$E[\mathbb{I}_{[A]}(x)] = 1 \times \underline{P(A)} + 0 \cdot (1 - P(A)) = P(A)$$

Indicator functions

- An indicator function for an event $\boldsymbol{A}$ is a function of x such that

$$\mathbb{I}_{[\boldsymbol{A}]}(x) = \begin{cases} 1 & event\ occurs\ for\ the\ value\ x \\ 0 & otherwise \end{cases}$$

- The expected value of the indicator function is the probability of event $\boldsymbol{A}$

$$\boxed{E[\,\mathbb{I}_{[A]}\,(x)] = 1 \times P(A) + 0 \times (1 - P(A)) = P(A)}$$

Markov's inequality

- For any random variable X and constant a >0

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

- So, a random variable is unlikely to have the absolute value much larger than the mean of its absolute value

- For example, if a = 10 E[|X|]

$$P(|X| \geq 10E[|X|]) \leq 0.1$$

Proof of Markov's inequality

$$\mathbb{I}_{[|X|\geq a]}(X) = \begin{cases} 1 & if\ |X| \geq a \\ 0 & otherwise \end{cases}$$

$a > 0$

$$\leq \frac{|X|}{a}$$

$$1 \leq \frac{|X|}{a} \quad \because |X| \geq a$$

$$0 \leq \frac{|X|}{a}$$

$$\rightarrow E\left[\mathbb{I}_{(|X|\geq a)}(X)\right] \leq E\left[\frac{|X|}{a}\right]$$

$$LHS = P(|X| \geq a) \leq E\left[\frac{|X|}{a}\right]$$

$$= \frac{1}{a} E[|X|] = \frac{E[|X|]}{a}$$

Proof of Markov's inequality

$$\mathbb{I}_{[|X| \geq a]}(X) = \begin{cases} 1 & if\ |X| \geq a \\ 0 & otherwise \end{cases}$$

$$\leq \frac{|X|}{a}$$

$a > 0$

Proof of Markov's inequality

$$\mathbb{I}_{[|X|\geq a]}(X) = \begin{cases} 1 & if\ |X| \geq a \\ 0 & otherwise \end{cases}$$

$a > 0$

$$\leq \frac{|X|}{a}$$

$$E[\mathbb{I}_{[|X|\geq a]}(X)] \leq \frac{E[|X|]}{a}$$

Proof of Markov's inequality

$$\mathbb{I}_{[|X|\geq a]}(X) = \begin{cases} 1 & if\ |X| \geq a \\ 0 & otherwise \end{cases}$$

$$\leq \frac{|X|}{a}$$

$$E[\mathbb{I}_{[|X|\geq a]}(X)] \leq \frac{E[|X|]}{a}$$

$\downarrow$

LHS =

Proof of Markov's inequality

$$\mathbb{I}_{[|X| \geq a]}(X) = \begin{cases} 1 & if\ |X| \geq a \\ 0 & otherwise \end{cases}$$

$$\leq \frac{|X|}{a}$$

$$E[\mathbb{I}_{[|X| \geq a]}(X)] \leq \frac{E[|X|]}{a}$$

$\downarrow$

LHS = $P(|X| \geq a)$

Proof of Markov's inequality

$$\mathbb{I}_{[|X|\geq a]}(X) = \begin{cases} 1 & if \ |X| \geq a \\ 0 & otherwise \end{cases}$$

$$\leq \frac{|X|}{a}$$

$$E[\mathbb{I}_{[|X|\geq a]}(X)] \leq \frac{E[|X|]}{a}$$

$\downarrow$

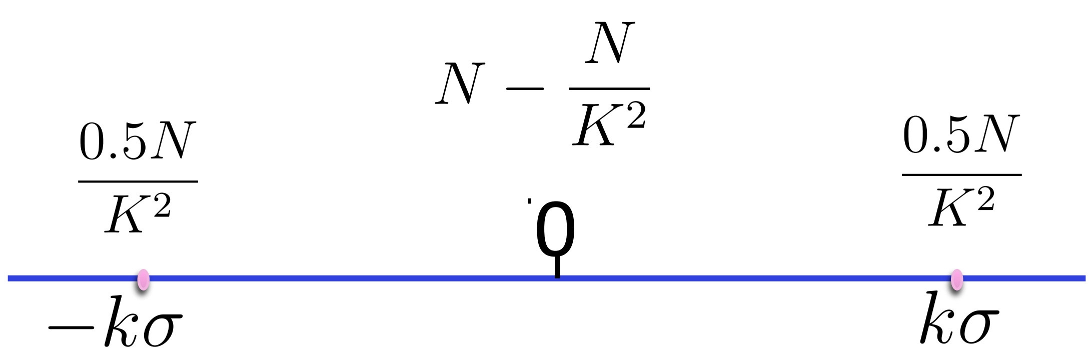
LHS = $P(|X| \geq a) \leq \dfrac{E[|X|]}{a}$

Recall

Standard deviation: Chebyshev's inequality (1st look)

- At most $\frac{N}{k^2}$ items are k standard deviations (σ) away from the mean

- Rough justification: Assume mean =0



$$\frac{\frac{N}{K^2}}{N} = \frac{1}{K^2}$$

$$std = \sqrt{\frac{1}{N}[(N - \frac{N}{k})0^2 + \frac{N}{k^2}(k\sigma)^2]} = \sigma$$

Chebyshev's inequality

- For any random variable X and constant *a* >0

$$P(|X - E[X]| \geq a) \leq \frac{var[X]}{a^2}$$

$a \uparrow$

RHS $\downarrow$

- If we let a = kσ where σ = std[X]

$$P(|X - E[X]| \geq k\sigma) \leq \frac{1}{k^2}$$

$var[x] = \sigma^2$

- In words, the probability that X is greater than k standard deviation away from the mean is small

Proof of Chebyshev's inequality

- Given Markov inequality, a>0

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

$$P(|U| \geq \omega) \leq \frac{E[|U|]}{\omega}$$

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

Proof of Chebyshev's inequality

- Given Markov inequality, a>0

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

- We can write $\boxed{P(|U| \geq w) \leq \frac{E[|U|]}{w}}$

$\omega > 0$

Proof of Chebyshev's inequality

- Given Markov inequality, a>0

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

- We can write $P(|U| \geq w) \leq \frac{E[|U|]}{w}$

$\omega > 0$, Let $\underline{U = (X - E[X])^2}$

Proof of Chebyshev's inequality

- Apply Markov inequality to $U = (X - E[X])^2$

$$P(|U| \geq w) \leq \underbrace{\frac{E[|U|]}{w} = \frac{E[U]}{w}}$$

$U \geq 0$

$|U| = U$

Proof of Chebyshev's inequality

- Apply Markov inequality to $U = (X - E[X])^2$

$$P(|U| \geq w) \leq \frac{E[|U|]}{w} = \frac{E[U]}{w} = \frac{var[X]}{w}$$

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

$= Var[X]$

Proof of Chebyshev's inequality

- Apply Markov inequality to $U = (X - E[X])^2$

$$P(|U| \geq w) \leq \frac{E[|U|]}{w} = \frac{E[U]}{w} = \frac{var[X]}{w}$$

- Substitute $U = (X - E[X])^2$ and $w = a^2$

Proof of Chebyshev's inequality

- Apply Markov inequality to $U = (X - E[X])^2$

$$P(|U| \geq w) \leq \frac{E[|U|]}{w} = \frac{E[U]}{w} = \frac{var[X]}{w}$$

- Substitute $U = (X - E[X])^2$ and $w = a^2$

$$P((X - E[X])^2 \geq a^2) \leq \frac{var[X]}{a^2}$$

$a > 0$

$\text{LHS} = P(|X - E[x]| \geq a) \leq \frac{var[X]}{a^2}$

Proof of Chebyshev's inequality

- Apply Markov inequality to $U = (X - E[X])^2$

$$P(|U| \geq w) \leq \frac{E[|U|]}{w} = \frac{E[U]}{w} = \frac{var[X]}{w}$$

- Substitute $U = (X - E[X])^2$ and $w = a^2$

$$P((X - E[X])^2 \geq a^2) \leq \frac{var[X]}{a^2}$$

Assume $a > 0$

$var[x] < \infty$

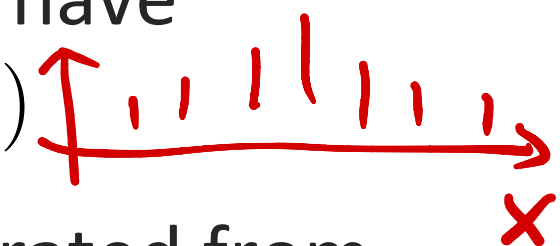
$$\Rightarrow \boxed{P(|X - E[X]| \geq a) \leq \frac{var[X]}{a^2}}$$ ✓

Now we are closer to the law of large numbers

Sample mean and IID samples

- We define the sample mean $\overline{\mathbf{X}}$ of ***N*** random variables $X_1, \ldots, X_N$ to be their average.

- If $X_1, \ldots, X_N$ are ***independent*** and have ***identical*** probability function $P(x)$



  then the numbers randomly generated from them are called **IID** samples

- The **sample mean** is a random variable

$X_i \to S_i$

$E[X_i] = 5$

$E\left[\sum_{i=1}^{50} X_i\right] = 50 \times 5 \nearrow E[X_i]$

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

$E\left[\frac{\sum^{50} X_i}{50}\right] = \frac{1}{50} E\left[\sum_i^{50} X_i\right]$

$E[\bar{X}] = E[X_i] = E[X] = 5$

Sample mean and IID samples

- Assume we have a set of **IID samples** from ***N*** random variables $X_1, \ldots, X_N$ that have probability function $P(x)$

- We use $\overline{\mathbf{X}}$ to denote the **sample mean** of these **IID samples**

$$\overline{\mathbf{X}} = \frac{\sum_{i=1}^{N} X_i}{N}$$

Expected value of sample mean of IID random variables

- By linearity of expected value

$$E[\overline{\mathbf{X}}] = E\left[\frac{\sum_{i=1}^{N} X_i}{N}\right] = \frac{1}{N}\sum_{i=1}^{N} E[X_i]$$

S_i

$X_i \rightarrow$ iid samples.

$E[X_i] = E[X_j] \quad i \neq j$

$= E[X]$

$\sum_i^N E[X]$

$= N \cdot E[X] \Rightarrow E[\bar{X}] = \frac{1}{N} \cdot N \cdot E[X]$

$= E[X]$

Expected value of sample mean of IID random variables

- By linearity of expected value

$$E[\overline{\mathbf{X}}] = E\left[\frac{\sum_{i=1}^{N} X_i}{N}\right] = \frac{1}{N}\sum_{i=1}^{N} E[X_i]$$

- Given each X_i has identical $P(x)$

$$\boxed{E[\overline{\mathbf{X}}] = \frac{1}{N}\sum_{i=1}^{N} E[X] = E[X]}$$

Variance of sample mean of IID random variables

- By the scaling property of variance

$$var[\overline{\mathbf{X}}] = var[\frac{1}{N}\sum_{i=1}^{N} X_i] = \frac{1}{N^2} var[\sum_{i=1}^{N} X_i]$$

$$var[kX] = k^2 var[X]$$

X_i, X_j iid independent random variables

$$var[X+Y] = var[X] + var[Y]$$

$$var[\sum^{N} X_i] = \sum^{N} var[X_i]$$

Variance of sample mean of IID random variables

- By the scaling property of variance

$$var[\overline{\mathbf{X}}] = var[\frac{1}{N}\sum_{i=1}^{N} X_i] = \frac{1}{N^2} var[\sum_{i=1}^{N} X_i]$$

- And by independence of these IID random variables

$$var[\overline{\mathbf{X}}] = \frac{1}{N^2}\sum_{i=1}^{N} var[X_i]$$

$var[X_i] = var[X_j]$

$$= \frac{1}{N^2} N \cdot var[X]$$

$$= \frac{1}{N} var[X]$$

Variance of sample mean of IID random variables

- By the scaling property of variance

$$var[\overline{\mathbf{X}}] = var[\frac{1}{N}\sum_{i=1}^{N} X_i] = \frac{1}{N^2} var[\sum_{i=1}^{N} X_i]$$

- And by independence of these IID random variables

$$var[\overline{\mathbf{X}}] = \frac{1}{N^2}\sum_{i=1}^{N} var[X_i]$$

distribution

- Given each X_i has identical $P(x)$, $var[X_i] = var[X]$

$$\boxed{var[\overline{\mathbf{X}}] = \frac{1}{N^2}\sum_{i=1}^{N} var[X] = \frac{var[X]}{N}}$$

Expected value and variance of sample mean of IID random variables

- The expected value of sample mean is the same as the expected value of the distribution

$$\boxed{E[\overline{\mathbf{X}}] = E[X]}$$

$E[S] = 5$

- The variance of sample mean is the distribution's variance divided by the sample size ***N***

$$\boxed{var[\overline{\mathbf{X}}] = \frac{var[X]}{N}}$$

$var[S] = \frac{5}{2}$

Weak law of large numbers

- Given a random variable X with finite variance, probability distribution function $P(x)$ and the sample mean $\overline{\mathbf{X}}$ of size ***N***.
- For any positive number $\epsilon > 0$

$$\lim_{N \to \infty} P(|\overline{\mathbf{X}} - E[X]| \geq \epsilon) = 0$$

- That is: the value of the mean of **IID** samples is very close with high probability to the expected value of the population when sample size is very large

meaning possible values of X_i

Proof of Weak law of large numbers

- Apply Chebyshev's inequality

$$P(|\overline{\mathbf{X}} - E[\overline{\mathbf{X}}]| \geq \epsilon) \leq \frac{var[\overline{\mathbf{X}}]}{\epsilon^2}$$

$E[\bar{x}]$ $var[\bar{x}]$

$E[\bar{x}] = E[x]$

$var[\bar{x}] = \frac{1}{n} var[x]$

Proof of Weak law of large numbers

- Apply Chebyshev's inequality

$$P(|\overline{\mathbf{X}} - E[\overline{\mathbf{X}}]| \geq \epsilon) \leq \frac{var[\overline{\mathbf{X}}]}{\epsilon^2}$$

- Substitute $E[\overline{\mathbf{X}}] = E[X]$ and $var[\overline{\mathbf{X}}] = \dfrac{var[X]}{N}$

Proof of Weak law of large numbers

- Apply Chebyshev's inequality

$$P(|\overline{\mathbf{X}} - E[\overline{\mathbf{X}}]| \geq \epsilon) \leq \frac{var[\overline{\mathbf{X}}]}{\epsilon^2}$$

$E[S]=5$

- Substitute $E[\overline{\mathbf{X}}] = E[X]$ and $var[\overline{\mathbf{X}}] = \frac{var[X]}{N}$

$$P(|\overline{\mathbf{X}} - E[\mathbf{X}]| \geq \epsilon) \leq \frac{var[\mathbf{X}]}{N\epsilon^2}$$

RHS $\to 0$
as $N \to \infty$

$var[X] < \infty$

Proof of Weak law of large numbers

- Apply Chebyshev's inequality

$$P(|\overline{\mathbf{X}} - E[\overline{\mathbf{X}}]| \geq \epsilon) \leq \frac{var[\overline{\mathbf{X}}]}{\epsilon^2}$$

- Substitute $E[\overline{\mathbf{X}}] = E[X]$ and $var[\overline{\mathbf{X}}] = \frac{var[X]}{N}$

$$P(|\overline{\mathbf{X}} - E[\mathbf{X}]| \geq \epsilon) \leq \frac{var[\mathbf{X}]}{N\epsilon^2} \xrightarrow[N \to \infty]{} 0$$

Proof of Weak law of large numbers

- Apply Chebyshev's inequality

$$P(|\overline{\mathbf{X}} - E[\overline{\mathbf{X}}]| \geq \epsilon) \leq \frac{var[\overline{\mathbf{X}}]}{\epsilon^2}$$

In our expt.
$E[x] = 5$
$var[x] = 5/2$

- Substitute $E[\overline{\mathbf{X}}] = E[X]$ and $var[\overline{\mathbf{X}}] = \frac{var[X]}{N}$

$$P(|\overline{\mathbf{X}} - E[\mathbf{X}]| \geq \epsilon) \leq \frac{var[\mathbf{X}]}{N\epsilon^2} \xrightarrow[N \to \infty]{} 0$$

$E[\bar{x}]$
$var[\bar{x}]$

ϵ small

$$\boxed{\lim_{N \to \infty} P(|\overline{\mathbf{X}} - E[X]| \geq \epsilon) = 0}$$

Weak law of large numbers

- The law of large numbers ***justifies using simulations*** (instead of calculation) to estimate the expected values of random variables

$$\lim_{N\to\infty} P(|\overline{\mathbf{X}} - E[X]| \geq \epsilon) = 0$$

- The law of large numbers also ***justifies using histogram*** of large random samples to approximate the probability distribution function $P(x)$, see proof on

Pg. 353 of the textbook by DeGroot, et al.

Histogram of large random IID samples approximates the probability distribution

- The law of large numbers justifies using histograms to approximate the probability distribution. Given ***N*** IID random variables X_1, …, X_N
  - Let $c_1 < c_2$ be two constants, Define $\boldsymbol{Y_i}$

$$Y_i = \begin{cases} 1 & if\ c_1 \le X_i < c_2 \\ 0 & otherwise \end{cases}$$

  - As we know for indicator function

$$E[Y_i] = P(c_1 \le X_i < c_2)$$

Histogram of large random IID samples approximates the probability distribution

- The law of large numbers justifies using histograms to approximate the probability distribution. Given ***N*** IID random variables X_1, …, X_N
  - Let $c_1 < c_2$ be two constants, Define $\boldsymbol{Y_i}$

$$Y_i = \begin{cases} 1 & if\ c_1 \leq X_i < c_2 \\ 0 & otherwise \end{cases}$$

  - As we know for indicator function

$$E[Y_i] = P(c_1 \leq X_i < c_2) = P(c_1 \leq X < c_2)$$

Histogram of large random IID samples approximates the probability distribution

- The law of large numbers justifies using histograms to approximate the probability distribution. Given ***N*** IID random variables X_1, …, X_N
  - According to the law of large numbers

$$\overline{\mathbf{Y}} = \frac{\sum_{i=1}^{N} Y_i}{N} \xrightarrow{N \to \infty} E[Y_i]$$

  - As we know for indicator function

$$E[Y_i] = P(c_1 \leq X_i < c_2) = P(c_1 \leq X < c_2)$$

Simulation of the sum of two-dice

- http://www.randomservices.org/random/apps/DiceExperiment.html

Assignments

- Finish Chapter 4 of the textbook
- Next time: Continuous random variable, classic known probability distributions

Additional References

- Charles M. Grinstead and J. Laurie Snell "Introduction to Probability"
- Morris H. Degroot and Mark J. Schervish "Probability and Statistics"

See you next time

See You!

Simulation of airline overbooking

- An airline has a flight with **7** seats. They always sell 12 tickets for this flight. If ticket holders show up independently with probability **p**, estimate the following values
  - Expected value of the number of ticket holders who show up
  - Probability that the flight being overbooked
  - Expected value of the number of ticket holders who can't fly due to the flight is overbooked.

Conditional expectation

- Expected value of X conditioned on event A:

$$E[X|A] = \sum_{x \in D(X)} xP(X = x|A)$$

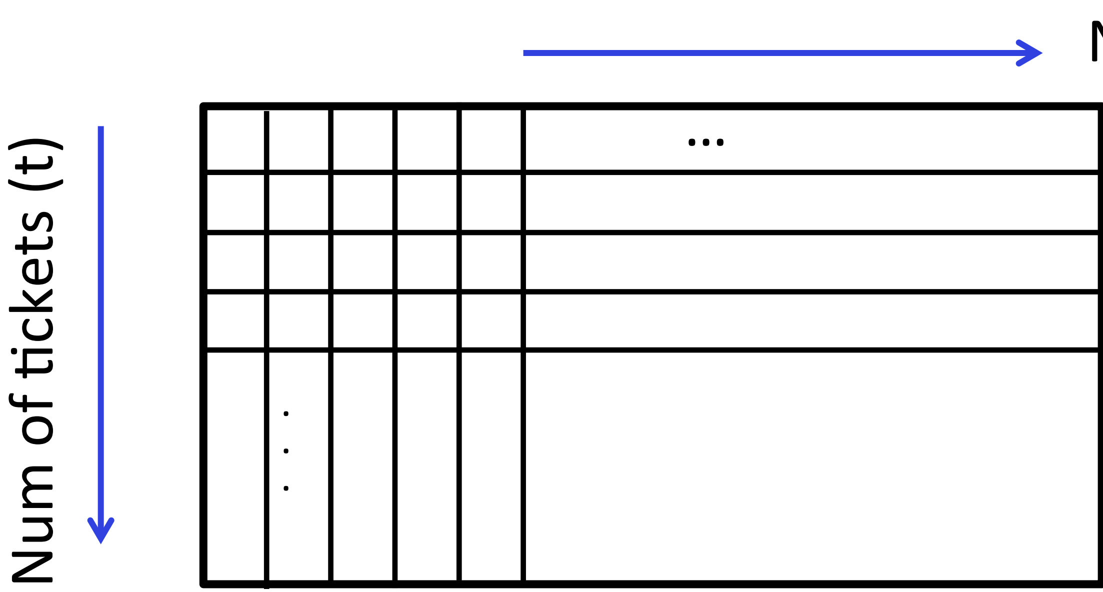
- Expected value of the number of ticketholders not flying

$$E[NF|overbooked] = \sum_{u=s+1}^{t} (u-s) \frac{\binom{t}{u} p^u (1-p)^{t-u}}{\sum_{v=s+1}^{t} \binom{t}{v} p^v (1-p)^{t-v}}$$

Simulate the arrival

- Expected value of the number of ticket holders who show up

nt=100000, t= 12, s=7, p=0.1, 0.2, … 1.0



We generate a matrix of random numbers from uniform distribution in [0,1],
Any number < p is considered an arrival

Simulate the arrival

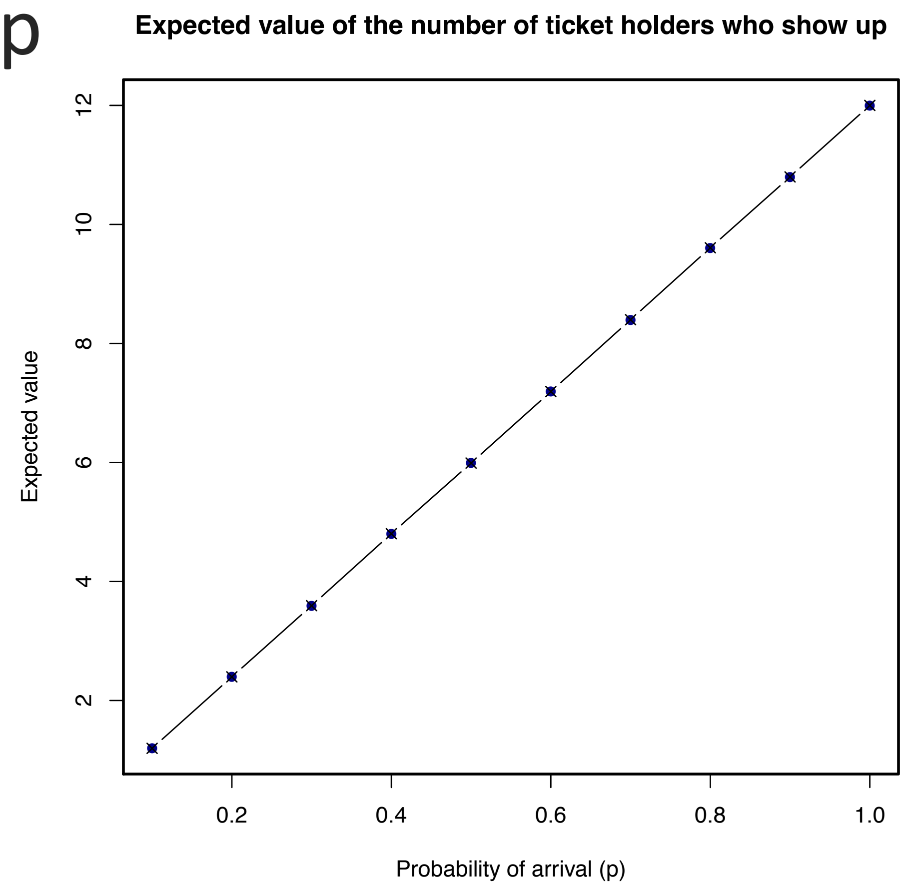
- Expected value of the number of ticket holders who show up

```
numTrials <-  100000
numTickets <- 12
numSeats <- 7
n_in <-10
for (i in 1:n_in ){
    p <- i/10
    m4<- matrix(runif(numTickets*numTrials,min=0,max=1), numTickets,numTrials)
    arrivals <- apply(m4, 2, p=p, cnt)
    numberArrivalExpec <- mean(arrivals)
    df[i,] <-   c(p,numberArrivalExpec)
}
```

Simulate the arrival

- Expected value of the number of ticket holders who show up

nt=100000, t= 12,
s=7, p=0.1, 0.2, … 1.0



Simulate the expected probability of overbooking

- Expected probability of the flight being overbooked

t= 12, s=7, p=0.1, 0.2, … 1.0

- **Expected probability** is equal to the **expected value of indicator function**. Whenever we have Num of arrival > Num of seats, we mark it with an indicator function. Then estimate with the sample mean of indicator functions.

Simulate the expected probability of overbooking

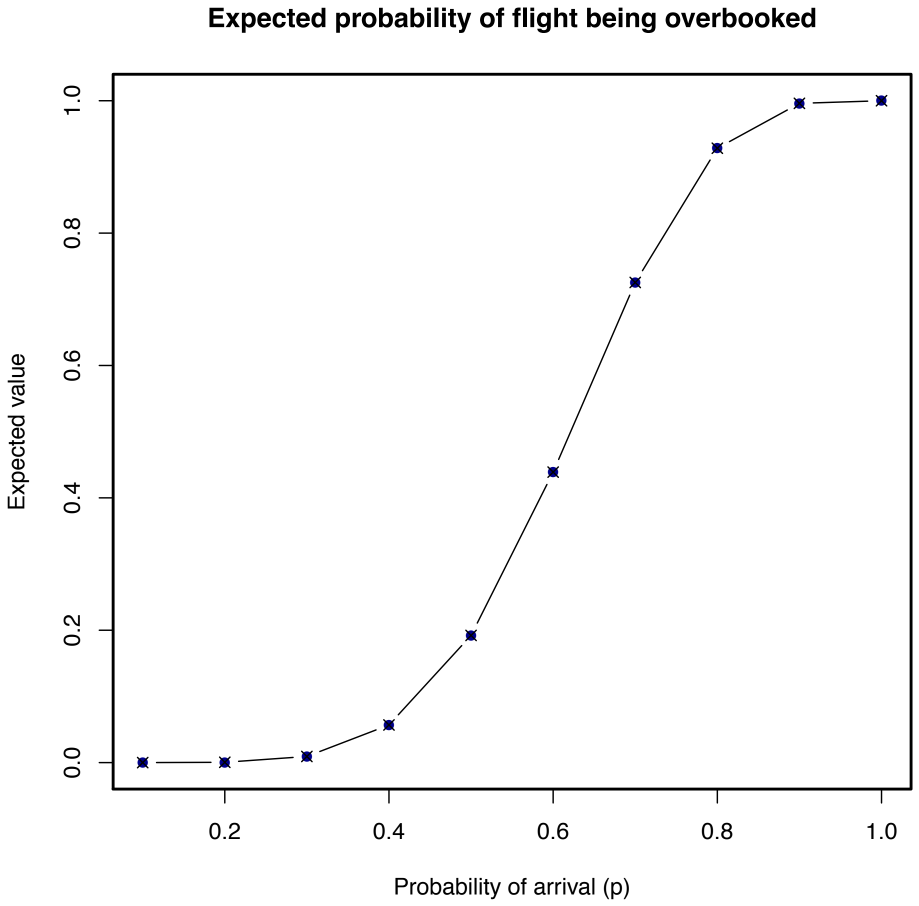
- Expected probability of the flight being overbooked

```
numTrials <-  100000
numTickets <- 12
numSeats <- 7
n_in <-10
for (i in 1:n_in ){
    p <- i/10
     m4<- matrix(runif(numTickets*numTrials,min=0,max=1), numTickets,numTrials)
    arrivals <- apply(m4, 2, p=p, cnt)
    indicatorOverbooked <- ifelse(arrivals > numSeats,1,0)
    df2[i,] <-  c(p,mean(indicatorOverbooked))
}
```

Simulate the expected probability of overbooking

- Expected probability of the flight being overbooked

nt=100000,
t= 12, s=7,
p=0.1, 0.2, … 1.0



Simulate the expected value of the number of grounded ticket holders given overbooked

- Expected value of the number of ticket holders who can’t fly due to the flight being overbooked

```
numTrials <-  200000
numTickets <- 12
numSeats <- 7
n_in <-10
for (i in 1:n_in ){
    p <- i/10
     m4<- matrix(runif(numTickets*numTrials,min=0,max=1), numTickets,numTrials)
    arrivals <- apply(m4, 2, p=p, cnt)
    indicatorOverbooked <- ifelse(arrivals > numSeats,1,0)
    numberGrd <- arrivals[which(indicatorOverbooked>0)]-numSeats
    df3[i,] <-  c(p,mean(numberGrd))
}
```

Simulate the expected value of the number of grounded ticket holders given overbooked

- Expected value of the number of ticket holders who can't fly due to the flight being overbooked

Nt=200000,
t= 12, s=7,
p=0.1, 0.2, … 1.0