



# Probability Distributions

---



# Random Variable

---

- A random variable  $x$  takes on a defined set of values with different probabilities.
  - For example, if you roll a die, the outcome is random (not fixed) and there are 6 possible outcomes, each of which occur with probability one-sixth.
  - For example, if you poll people about their voting preferences, the percentage of the sample that responds “Yes on Proposition 100” is also a random variable (the percentage will be slightly differently every time you poll).
- Roughly, probability is how frequently we expect different outcomes to occur if we repeat the experiment over and over (“frequentist” view)



# Random variables can be discrete or continuous

---

- **Discrete** random variables have a countable number of outcomes
  - Examples: Dead/alive, treatment/placebo, dice, counts, etc.
- **Continuous** random variables have an infinite continuum of possible values.
  - Examples: blood pressure, weight, the speed of a car, the real numbers from 1 to 6.



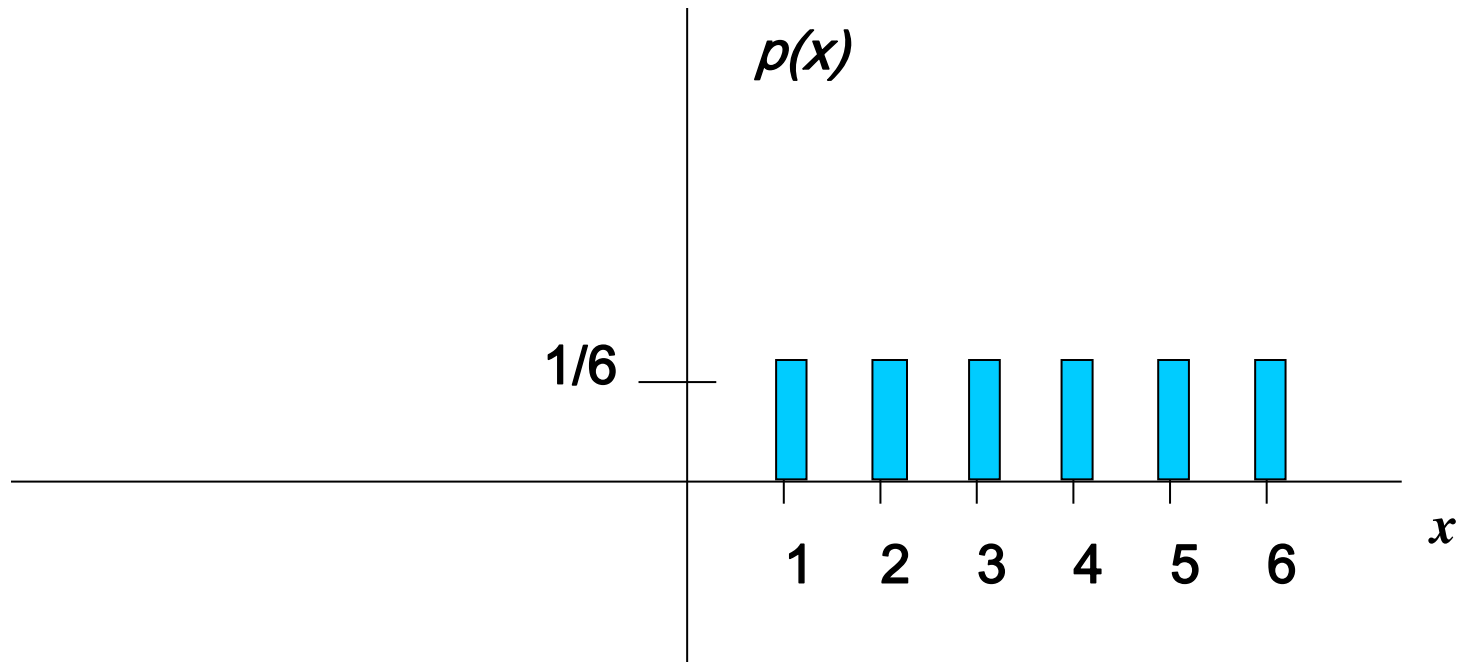
# Probability functions

---

- A probability function maps the possible values of  $x$  against their respective probabilities of occurrence,  $p(x)$
- $p(x)$  is a number from 0 to 1.0.
- The area under a probability function is always 1.



# Discrete example: roll of a die



$$\sum_{\text{all } x} P(x) = 1$$

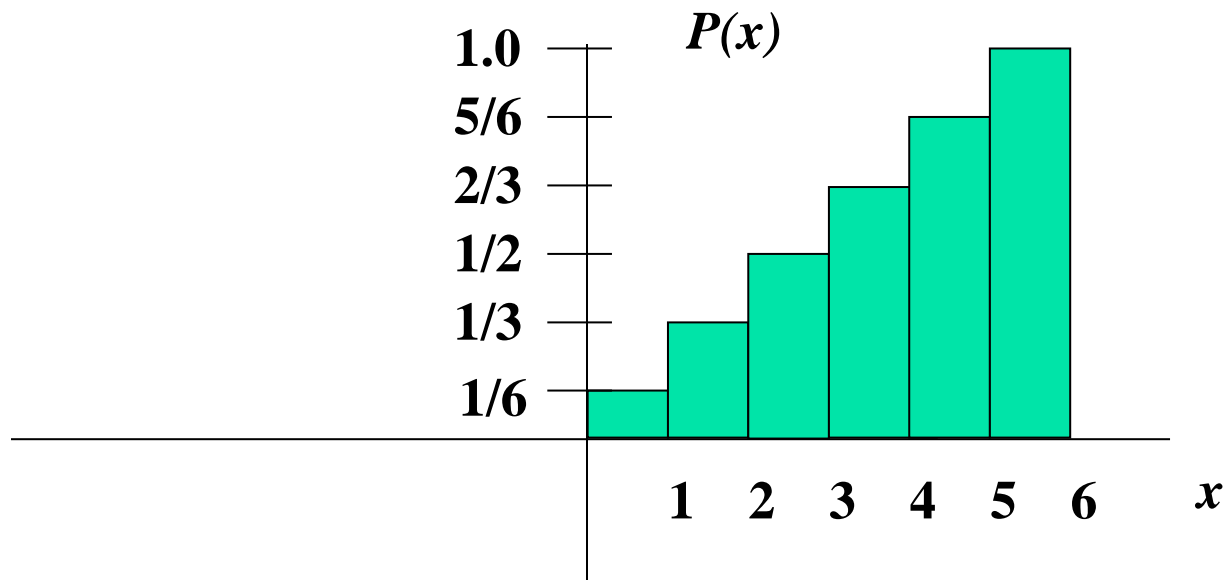


# Probability mass function (pmf)

$x$	$p(x)$
1	$p(x=1)=1/6$
2	$p(x=2)=1/6$
3	$p(x=3)=1/6$
4	$p(x=4)=1/6$
5	$p(x=5)=1/6$
6	$\underline{p(x=6)=1/6}$

1.0

# Cumulative distribution function (CDF)



# Cumulative distribution function

$x$	$P(x \leq A)$
1	$P(x \leq 1) = 1/6$
2	$P(x \leq 2) = 2/6$
3	$P(x \leq 3) = 3/6$
4	$P(x \leq 4) = 4/6$
5	$P(x \leq 5) = 5/6$
6	$P(x \leq 6) = 6/6$





# Examples

---

1. What's the probability that you roll a 3 or less?

$$P(x \leq 3) = 1/2$$

2. What's the probability that you roll a 5 or higher?

$$P(x \geq 5) = 1 - P(x \leq 4) = 1 - 2/3 = 1/3$$



# Practice Problem

---

Which of the following are probability functions?

- a.  $f(x) = .25$  for  $x = 9, 10, 11, 12$
- b.  $f(x) = (3-x)/2$  for  $x = 1, 2, 3, 4$
- c.  $f(x) = (x^2 + x + 1)/25$  for  $x = 0, 1, 2, 3$



# Answer (a)

---

a.  $f(x) = .25$  for  $x = 9, 10, 11, 12$

<b><math>x</math></b>	<b><math>f(x)</math></b>
<b>9</b>	<b>.25</b>
<b>10</b>	<b>.25</b>
<b>11</b>	<b>.25</b>
<b>12</b>	<b><u>.25</u></b>

**1.0**

**Yes, probability  
function!**



## Answer (b)

b.  $f(x) = (3-x)/2$  for  $x=1,2,3,4$

$x$	$f(x)$
<b>1</b>	<b><math>(3-1)/2=1.0</math></b>
<b>2</b>	<b><math>(3-2)/2=.5</math></b>
<b>3</b>	<b><math>(3-3)/2=0</math></b>
<b>4</b>	<b><math>(3-4)/2=-.5</math></b>

Though this sums to 1, you can't have a negative probability; therefore, it's not a probability function.



## Answer (c)

c.  $f(x) = (x^2 + x + 1)/25$  for  $x=0,1,2,3$

<b>x</b>	<b>f(x)</b>
<b>0</b>	<b>1/25</b>
<b>1</b>	<b>3/25</b>
<b>2</b>	<b>7/25</b>
<b>3</b>	<b><u>13/25</u></b>

**24/25**

Doesn't sum to 1. Thus,  
it's not a probability  
function.



# Practice Problem:

---

- The number of ships to arrive at a harbor on any given day is a random variable represented by  $x$ . The probability distribution for  $x$  is:

$x$	10	11	12	13	14
$P(x)$	.4	.2	.2	.1	.1

Find the probability that on a given day:

- exactly 14 ships arrive  $p(x=14) = .1$
- At least 12 ships arrive  $p(x \geq 12) = (.2 + .1 + .1) = .4$
- At most 11 ships arrive  $p(x \leq 11) = (.4 + .2) = .6$



# Practice Problem:

---

You are lecturing to a group of 1000 students. You ask them to each randomly pick an integer between 1 and 10. Assuming, their picks are truly random:

- What's your best guess for how many students picked the number 9?

Since  $p(x=9) = 1/10$ , we'd expect about  $1/10^{\text{th}}$  of the 1000 students to pick 9. 100 students.

- What percentage of the students would you expect picked a number less than or equal to 6?

Since  $p(x \leq 6) = 1/10 + 1/10 + 1/10 + 1/10 + 1/10 + 1/10 = .6$   
60%



# Important discrete distributions in epidemiology...

---

- Binomial

- Yes/no outcomes (dead/alive, treated/untreated, smoker/non-smoker, sick/well, etc.)

- Poisson

- Counts (e.g., how many cases of disease in a given area)





# Continuous case

---

- The probability function that accompanies a continuous random variable is a continuous mathematical function that integrates to 1.
- The probabilities associated with continuous functions are just areas under the curve (integrals!).
- Probabilities are given for a range of values, rather than a particular value (e.g., the probability of getting a math SAT score between 700 and 800 is 2%).



# Continuous case

---

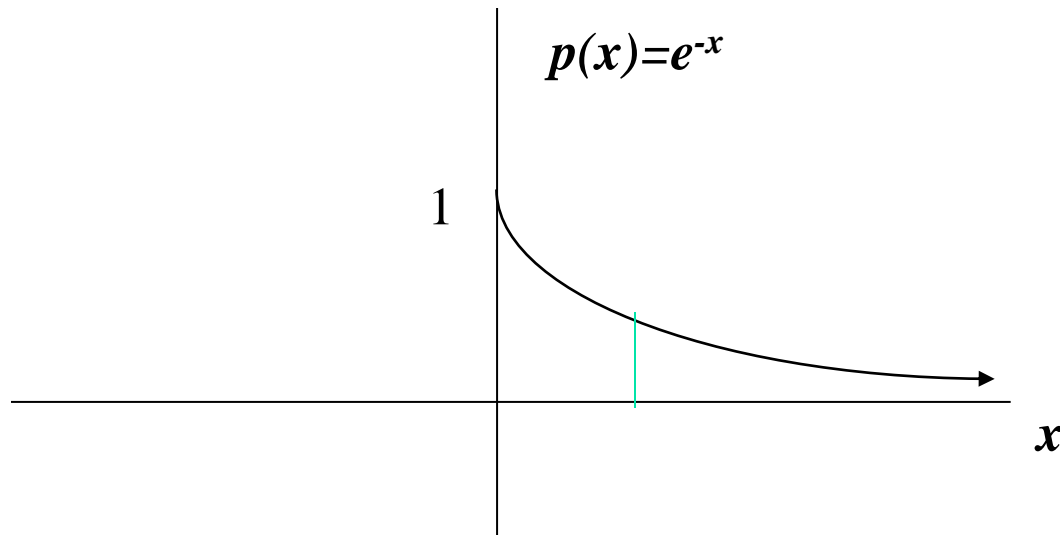
- For example, recall the negative exponential function (in probability, this is called an “exponential distribution”):

$$f(x) = e^{-x}$$

- This function integrates to 1:

$$\int_0^{+\infty} e^{-x} = -e^{-x} \Big|_0^{+\infty} = 0 + 1 = 1$$

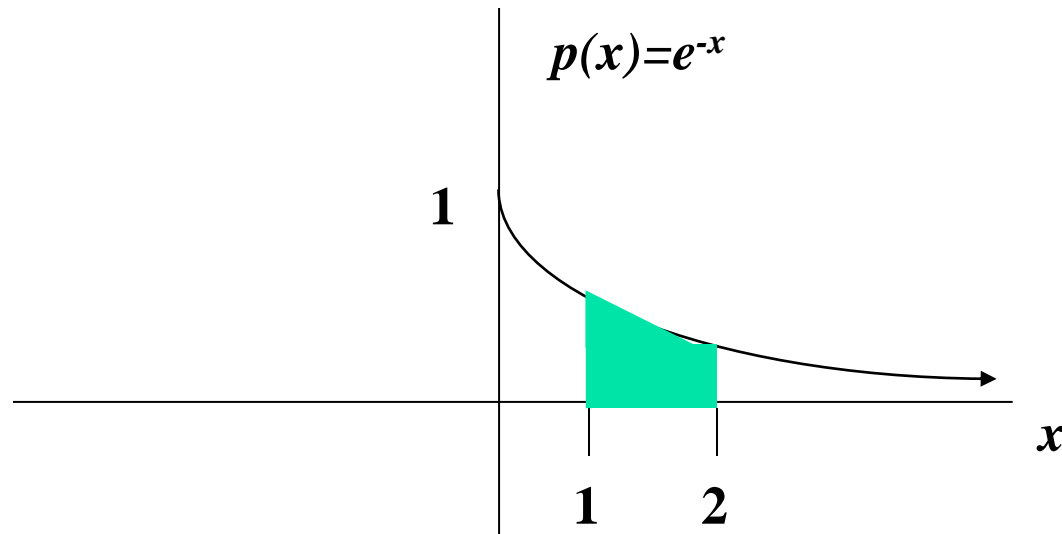
# Continuous case: “probability density function” (pdf)



The probability that  $x$  is any exact particular value (such as 1.9976) is 0; we can only assign probabilities to possible ranges of  $x$ .



For example, the probability of  $x$  falling within 1 to 2:



$$P(1 \leq x \leq 2) = \int_1^2 e^{-x} = -e^{-x} \Big|_1^2 = -e^{-2} - (-e^{-1}) = -.135 + .368 = .23$$



# Cumulative distribution function

---

As in the discrete case, we can specify the “cumulative distribution function” (CDF):

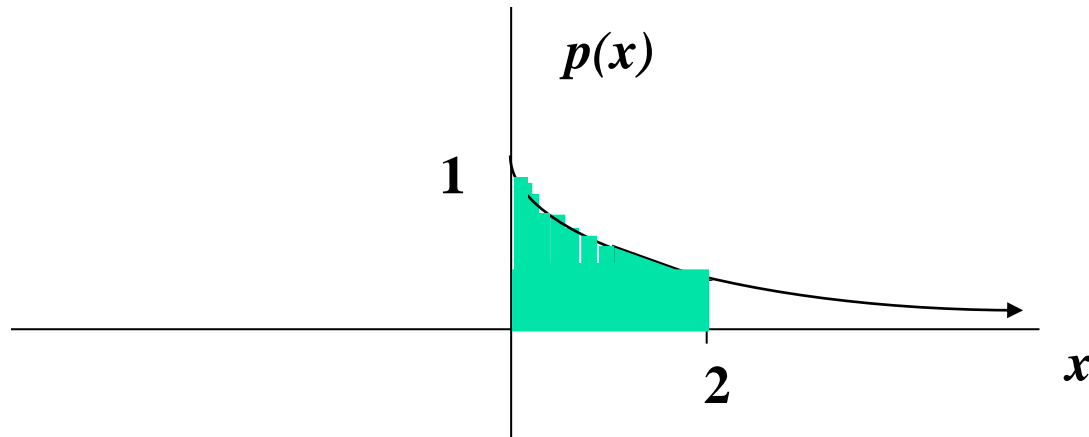
The CDF here =  $P(x \leq A) =$

$$\int_0^A e^{-x} = -e^{-x} \Big|_0^A = -e^{-A} - (-e^0) = -e^{-A} + 1 = 1 - e^{-A}$$



# Example

---

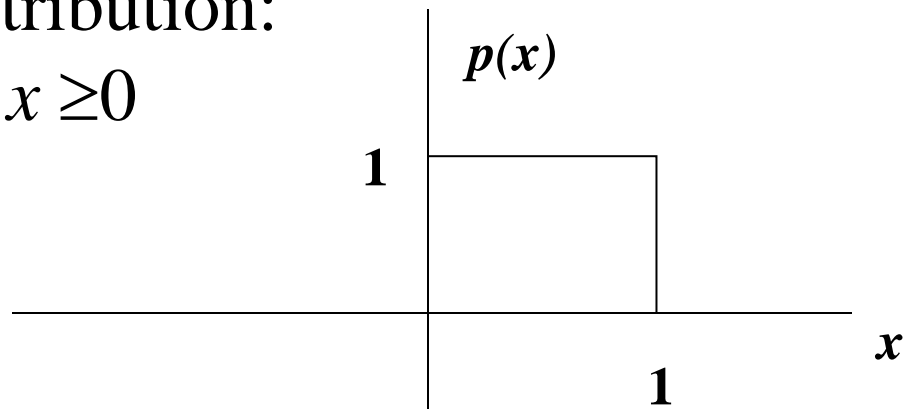


$$P(x \leq 2) = 1 - e^{-2} = 1 - .135 = .865$$

# Example 2: Uniform distribution

The uniform distribution: all values are equally likely

The uniform distribution:  
 $f(x) = 1$ , for  $1 \geq x \geq 0$

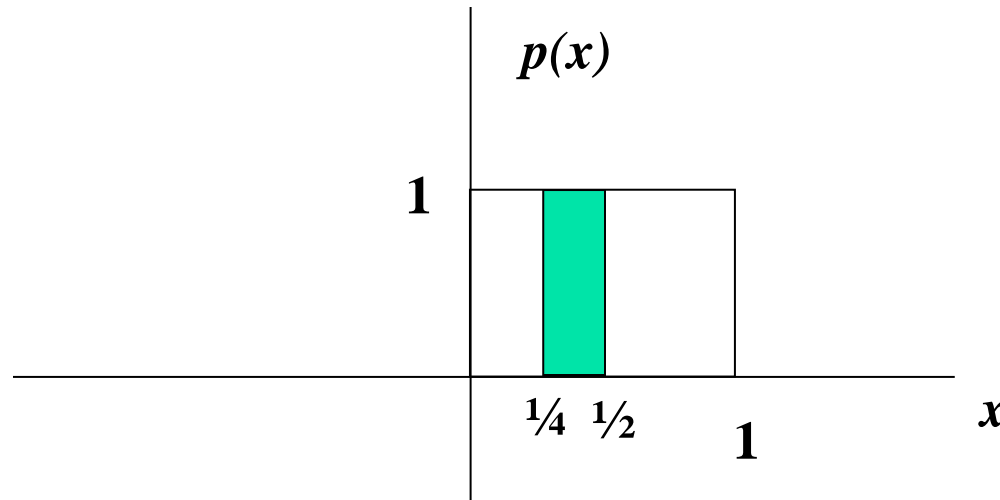


We can see it's a probability distribution because it integrates to 1 (the area under the curve is 1):

$$\int_0^1 1 = x \Big|_0^1 = 1 - 0 = 1$$

# Example: Uniform distribution

What's the probability that  $x$  is between  $1/4$  and  $1/2$ ?



$$P(1/2 \geq x \geq 1/4) = 1/4$$





# Practice Problem

---

4. Suppose that survival drops off rapidly in the year following diagnosis of a certain type of advanced cancer. Suppose that the length of survival (or time-to-death) is a random variable that approximately follows an exponential distribution with parameter 2 (makes it a steeper drop off):

$$\text{probability function : } p(x = T) = 2e^{-2T}$$

$$[\text{note : } \int_0^{+\infty} 2e^{-2x} = -e^{-2x} \Big|_0^{+\infty} = 0 + 1 = 1]$$

**What's the probability that a person who is diagnosed with this illness survives a year?**



# Answer

---

The probability of dying within 1 year can be calculated using the cumulative distribution function:

Cumulative distribution function is:

$$P(x \leq T) = -e^{-2x} \Big|_0^T = 1 - e^{-2(T)}$$

The chance of surviving past 1 year is:  $P(x \geq 1) = 1 - P(x \leq 1)$

$$1 - (1 - e^{-2(1)}) = .135$$



## Expected value, or mean

---

- Expected value is just the weighted average or mean ( $\mu$ ) of random variable  $x$ .
- Imagine placing the masses  $p(x)$  at the points  $X$  on a beam; the balance point of the beam is the expected value of  $x$ .

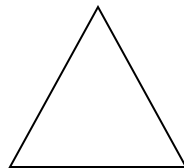


# Example: expected value

---

- Recall the following probability distribution of ship arrivals:

$x$	10	11	12	13	14
$P(x)$	.4	.2	.2	.1	.1



$$\sum_{i=1}^5 x_i p(x) = 10(.4) + 11(.2) + 12(.2) + 13(.1) + 14(.1) = 11.3$$



# Expected value, formally

---

**Discrete case:**


$$E(X) = \sum_{\text{all } x} x_i p(x_i)$$

**Continuous case:**

$$E(X) = \int_{\text{all } x} x_i p(x_i) dx$$

# Empirical Mean is a special case of Expected Value...

Sample mean, for a sample of  $n$  subjects: =

$$\bar{X} = \frac{\sum_{i=1}^n x_i}{n} = \sum_{i=1}^n x_i \left( \frac{1}{n} \right)$$


**The probability (frequency) of each person in the sample is  $1/n$ .**



# Expected value, formally

---

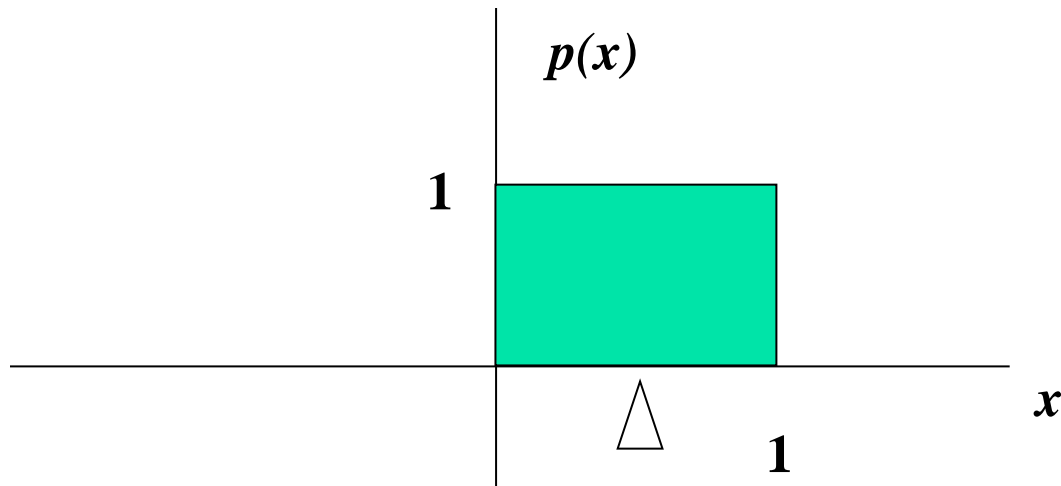
**Discrete case:**

$$E(X) = \sum_{\text{all } x} x_i p(x_i)$$

**Continuous case:**

$$E(X) = \int_{\text{all } x} x_i p(x_i) dx$$

# Extension to continuous case: uniform distribution



$$E(X) = \int_0^1 x(1)dx = \frac{x^2}{2} \Big|_0^1 = \frac{1}{2} - 0 = \frac{1}{2}$$





# Symbol Interlude

---

- $E(X) = \mu$ 
  - these symbols are used interchangeably



# Expected Value

---

- Expected value is an extremely useful concept for good decision-making!



# Example: the lottery

---

- The Lottery (also known as a tax on people who are bad at math...)
- A certain lottery works by picking 6 numbers from 1 to 49. It costs \$1.00 to play the lottery, and if you win, you win \$2 million after taxes.
- *If you play the lottery once, what are your expected winnings or losses?*



# Lottery

**Calculate the probability of winning in 1 try:**

$$\frac{1}{\binom{49}{6}} = \frac{1}{\frac{49!}{43!6!}} = \frac{1}{13,983,816} = 7.2 \times 10^{-8}$$

**“49 choose 6”**

Out of 49 numbers,  
this is the number  
of distinct  
combinations of 6.

**The probability function (note, sums to 1.0):**

$x\$$	$p(x)$
-1	.999999928
+ 2 million	$7.2 \times 10^{-8}$



# *Expected Value*

## The probability function

$x\$$	$p(x)$
-1	.999999928
+ 2 million	$7.2 \times 10^{-8}$

## Expected Value

$$\begin{aligned} E(X) &= P(\text{win}) * \$2,000,000 + P(\text{lose}) * -\$1.00 \\ &= 2.0 \times 10^6 * 7.2 \times 10^{-8} + .999999928 (-1) = .144 - .999999928 = -\$ .86 \end{aligned}$$

Negative expected value is never good!

You shouldn't play if you expect to lose money!



## Gambling (or how casinos can afford to give so many free drinks...)

---

A roulette wheel has the numbers 1 through 36, as well as 0 and 00. If you bet \$1 that an odd number comes up, you win or lose \$1 according to whether or not that event occurs. If random variable  $X$  denotes your net gain,  $X=1$  with probability  $18/38$  and  $X=-1$  with probability  $20/38$ .

$$E(X) = 1(18/38) - 1(20/38) = -\$0.053$$

On average, the casino wins (and the player loses) 5 cents per game.

The casino rakes in even more if the stakes are higher:

$$E(X) = 10(18/38) - 10(20/38) = -\$0.53$$

If the cost is \$10 per game, the casino wins an average of 53 cents per game. If 10,000 games are played in a night, that's a cool \$5300.

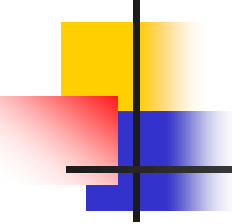


## \*\*A few notes about Expected Value as a mathematical operator:

---

If  $c$  = a constant number (i.e., not a variable) and  $X$  and  $Y$  are any random variables...

- $E(c) = c$
- $E(cX) = cE(X)$
- $E(c + X) = c + E(X)$
- $E(X + Y) = E(X) + E(Y)$


$$E(c) = c$$

---

$$E(c) = c$$

Example: If you cash in soda cans in CA, you always get 5 cents per can.

Therefore, there's no randomness. You always expect to (and do) get 5 cents.




$$E(cX) = cE(X)$$

---

$$E(cX) = cE(X)$$

Example: If the casino charges \$10 per game instead of \$1, then the casino expects to make 10 times as much on average from the game (See roulette example above!)


$$E(c + X) = c + E(X)$$

---

$$E(c + X) = c + E(X)$$

Example, if the casino throws in a free drink worth exactly \$5.00 every time you play a game, you always expect to (and do) gain an extra \$5.00 regardless of the outcome of the game.


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

---

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

Example: If you play the lottery twice, you expect to lose:  $-\$.86$   
+  $-\$.86$ .



# Variance/standard deviation

---

“The average (expected) squared distance (or deviation) from the mean”

$$\sigma^2 = \text{Var}(x) = E[(x - \mu)^2] = \sum_{\text{all } x} (x_i - \mu)^2 p(x_i)$$

*\*\*We square because squaring has better properties than absolute value. Take square root to get back linear average distance from the mean (= "standard deviation").*



# Variance, formally

---

**Discrete case:**

$$Var(X) = \sigma^2 = \sum_{\text{all } x} (x_i - \mu)^2 p(x_i)$$

**Continuous case:**

$$Var(X) = \sigma^2 = \int_{-\infty}^{\infty} (x_i - \mu)^2 p(x_i) dx$$



# Symbol Interlude

---

- $\text{Var}(X) = \sigma^2$ 
  - these symbols are used interchangeably



# Handy calculation formula!

---

**Handy calculation formula (if you ever need to calculate by hand!):**

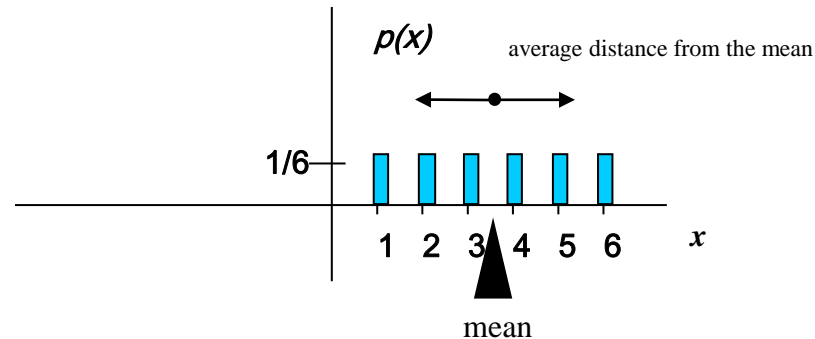
$$Var(X) = \sum_{\text{all } x} (x_i - \mu)^2 p(x_i) = \sum_{\text{all } x} x_i^2 p(x_i) - (\mu)^2$$

**Intervening algebra!**

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

For example, what's the variance and standard deviation of the roll of a die?

$x$	$p(x)$
1	$p(x=1)=1/6$
2	$p(x=2)=1/6$
3	$p(x=3)=1/6$
4	$p(x=4)=1/6$
5	$p(x=5)=1/6$
6	$p(x=6)=1/6$
1.0	

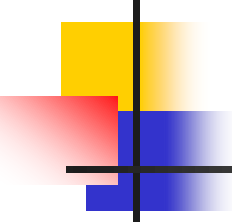


$$E(x) = \sum_{\text{all } x} x_i p(x_i) = (1)\left(\frac{1}{6}\right) + 2\left(\frac{1}{6}\right) + 3\left(\frac{1}{6}\right) + 4\left(\frac{1}{6}\right) + 5\left(\frac{1}{6}\right) + 6\left(\frac{1}{6}\right) = \frac{21}{6} = 3.5$$

$$E(x^2) = \sum_{\text{all } x} x_i^2 p(x_i) = (1)\left(\frac{1}{6}\right) + 4\left(\frac{1}{6}\right) + 9\left(\frac{1}{6}\right) + 16\left(\frac{1}{6}\right) + 25\left(\frac{1}{6}\right) + 36\left(\frac{1}{6}\right) = 15.17$$

$$\sigma_x^2 = \text{Var}(x) = E(x^2) - [E(x)]^2 = 15.17 - 3.5^2 = 2.92$$
$$\sigma_x = \sqrt{2.92} = 1.71$$

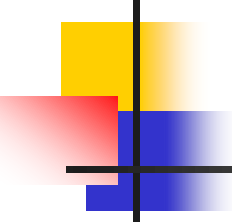



$$\text{Var}(c) = 0$$

---

$$\text{Var}(c) = 0$$

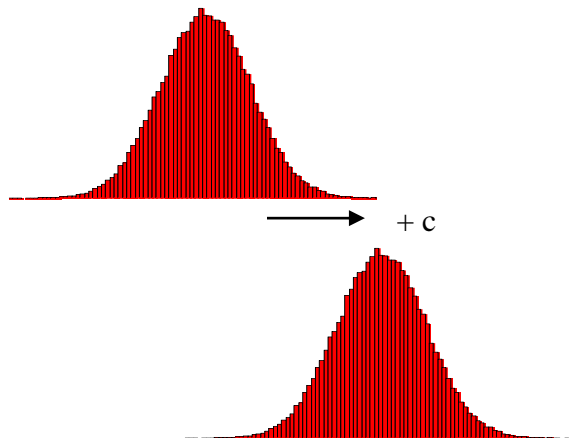
Constants don't vary!


$$\text{Var}(c+X) = \text{Var}(X)$$

---

$$\text{Var}(c+X) = \text{Var}(X)$$

Adding a constant to every instance of a random variable doesn't change the variability. It just shifts the whole distribution by  $c$ . If everybody grew 5 inches suddenly, the variability in the population would still be the same.

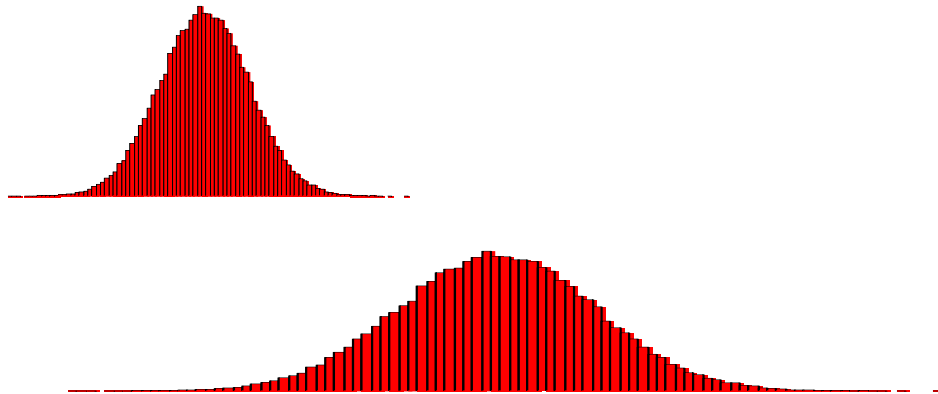



$$\text{Var}(cX) = c^2 \text{Var}(X)$$

---

$$\text{Var}(cX) = c^2 \text{Var}(X)$$

Multiplying each instance of the random variable by  $c$  makes it  $c$ -times as wide of a distribution, which corresponds to  $c^2$  as much variance (deviation squared). For example, if everyone suddenly became twice as tall, there'd be twice the deviation and 4 times the variance in heights in the population.




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

---

$\text{Var}(X + Y) = \text{Var}(X) + \text{Var}(Y)$  ***ONLY IF X and Y are independent!!!!!!!***

With two random variables, you have more opportunity for variation, unless they vary together (are dependent, or have covariance):  $\text{Var}(X + Y) = \text{Var}(X) + \text{Var}(Y) + 2\text{Cov}(X, Y)$



# Practice Problem

---

Find the variance and standard deviation for the number of ships to arrive at the harbor (recall that the mean is 11.3).

$x$	10	11	12	13	14
$P(x)$	.4	.2	.2	.1	.1



# Answer: variance and std dev

---

$x^2$	100	121	144	169	196
$P(x)$	.4	.2	.2	.1	.1

$$E(x^2) = \sum_{i=1}^5 x_i^2 p(x_i) = (100)(.4) + (121)(.2) + 144(.2) + 169(.1) + 196(.1) = 129.5$$

$$Var(x) = E(x^2) - [E(x)]^2 = 129.5 - 11.3^2 = 1.81$$

$$stddev(x) = \sqrt{1.81} = 1.35$$

**Interpretation: On an average day, we expect 11.3 ships to arrive in the harbor, plus or minus 1.35. This gives you a feel for what would be considered a usual day!**



# Practice Problem

---

You toss a coin 100 times. What's the expected number of heads? What's the variance of the number of heads?



# Answer: expected value

---

Intuitively, we'd probably all agree that we expect around 50 heads, right?

Another way to show this→

Think of tossing 1 coin.  $E(X=\text{number of heads}) = (1) P(\text{heads}) + (0)P(\text{tails})$

$$\therefore E(X=\text{number of heads}) = 1(.5) + 0 = .5$$

If we do this 100 times, we're looking for the sum of 100 tosses, where we assign 1 for a heads and 0 for a tails. (these are 100 "independent, identically distributed (i.i.d)" events)

$$\begin{aligned} E(X_1 + X_2 + X_3 + X_4 + X_5 + \dots + X_{100}) &= E(X_1) + E(X_2) + E(X_3) + E(X_4) + E(X_5) + \dots + \\ E(X_{100}) &= \\ 100 E(X_1) &= 50 \end{aligned}$$





# Answer: variance

---

What's the variability, though? More tricky. But, again, we could do this for 1 coin and then use our rules of variance.

Think of tossing 1 coin.

$$E(X^2 = \text{number of heads squared}) = 1^2 P(\text{heads}) + 0^2 P(\text{tails})$$

$$\therefore E(X^2) = 1(.5) + 0 = .5$$

$$\text{Var}(X) = .5 - .5^2 = .5 - .25 = .25$$

Then, using our rule:  $\text{Var}(X+Y) = \text{Var}(X) + \text{Var}(Y)$  (coin tosses are independent!)

$$\begin{aligned} \text{Var}(X_1 + X_2 + X_3 + X_4 + X_5 + \dots + X_{100}) &= \text{Var}(X_1) + \text{Var}(X_2) + \text{Var}(X_3) + \\ &\quad \text{Var}(X_4) + \text{Var}(X_5) + \dots + \text{Var}(X_{100}) = \end{aligned}$$

$$\begin{aligned} 100 \text{Var}(X_1) &= 100 (.25) = 25 \\ \text{SD}(X) &= 5 \end{aligned}$$

Interpretation: When we toss a coin 100 times, we expect to get 50 heads plus or minus 5.