



Introduction to 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)



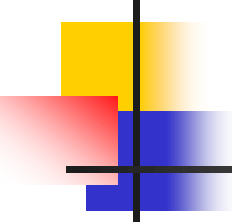
Random variables can be discrete or continuous

- **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.

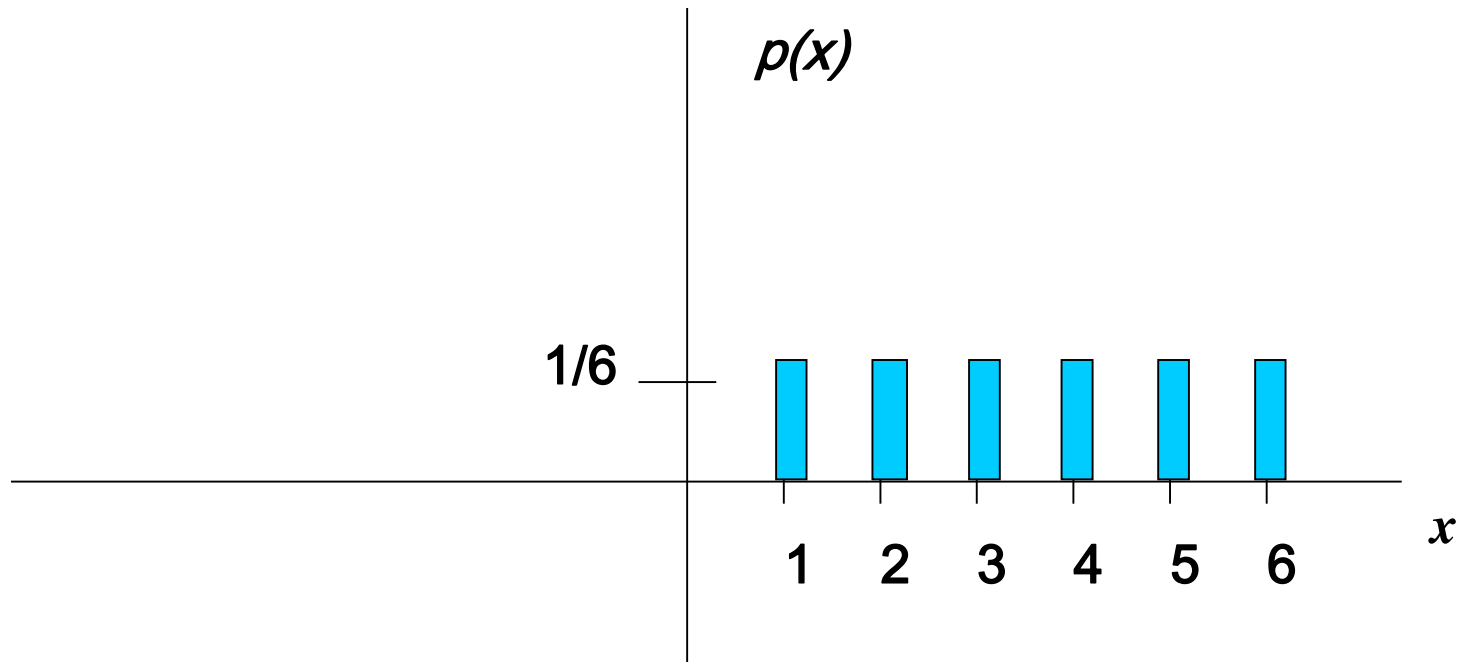


Probability functions

- 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



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

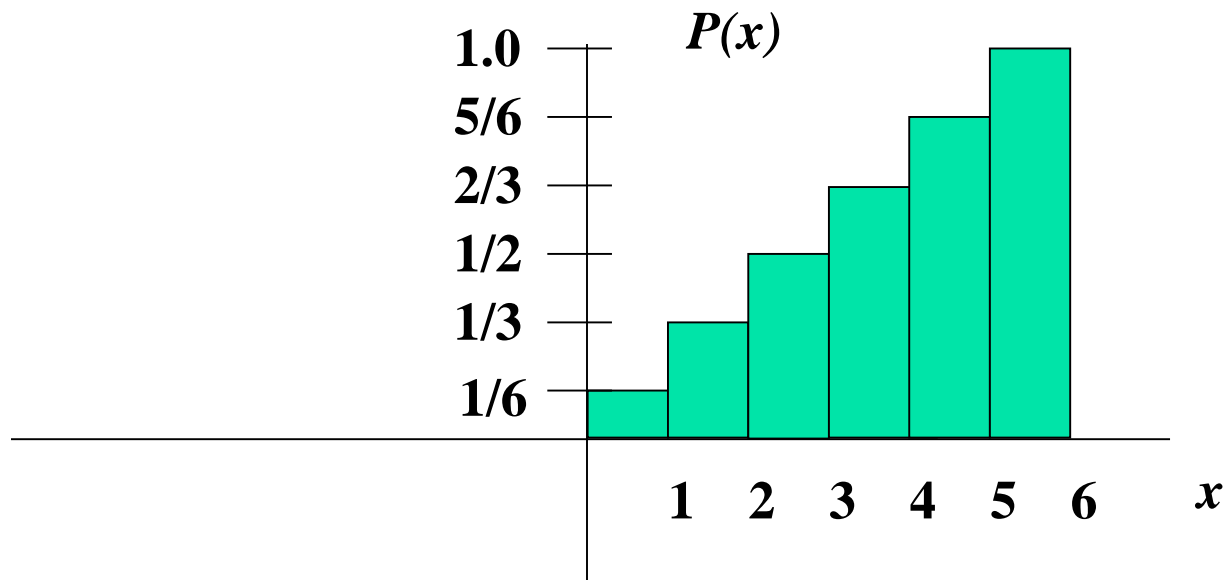


Probability mass function (pmf)

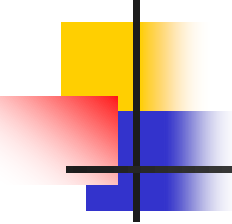
x	$p(x)$
1	$p(x=1)=1/6$
2	$p(x=2)=1/6$
3	$p(x=3)=1/6$
4	$p(x=4)=1/6$
5	$p(x=5)=1/6$
6	$\underline{p(x=6)=1/6}$

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	$P(x \leq 6) = 6/6$



Practice Problem:

- The number of patients seen in the ER in any given hour 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 in a given hour:

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



Review Question 1

If you toss a die, what's the probability that you roll a 3 or less?

- a. $1/6$
- b. $1/3$
- c. $1/2$
- d. $5/6$
- e. 1.0



Review Question 1

If you toss a die, what's the probability that you roll a 3 or less?

- a. $1/6$
- b. $1/3$
- c. **$1/2$**
- d. $5/6$
- e. 1.0



Review Question 2

Two dice are rolled and the sum of the face values is six? What is the probability that at least one of the dice came up a 3?

- a. $1/5$
- b. $2/3$
- c. $1/2$
- d. $5/6$
- e. 1.0



Review Question 2

Two dice are rolled and the sum of the face values is six. What is the probability that at least one of the dice came up a 3?

- a. $1/5$
- b. $2/3$
- c. $1/2$
- d. $5/6$
- e. 1.0

How can you get a 6 on two dice? 1-5, 5-1, 2-4, 4-2, 3-3

One of these five has a 3.

$\therefore 1/5$

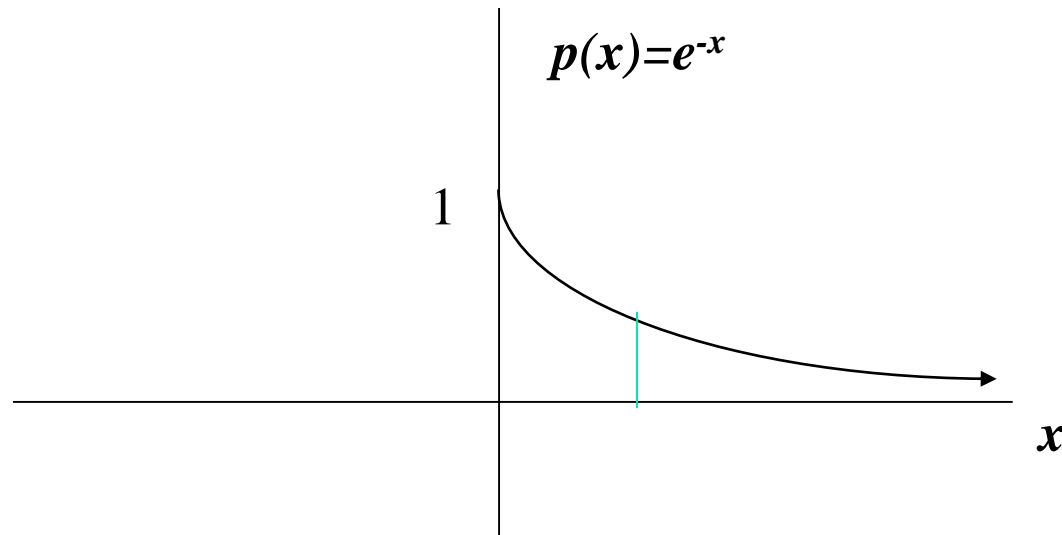


Continuous case

- The probability function that accompanies a continuous random variable is a continuous mathematical function that integrates to 1.
- 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)



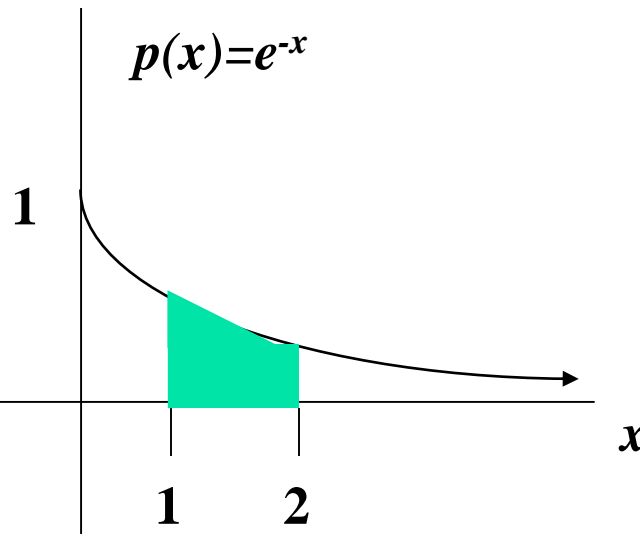
The probability that x is any exact particular value (such as 1.9976) is 0; we can only assign probabilities to possible ranges of x .



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

Clinical example: Survival times after lung transplant may roughly follow an exponential function.

Then, the probability that a patient will die in the second year after surgery (between years 1 and 2) is 23%.

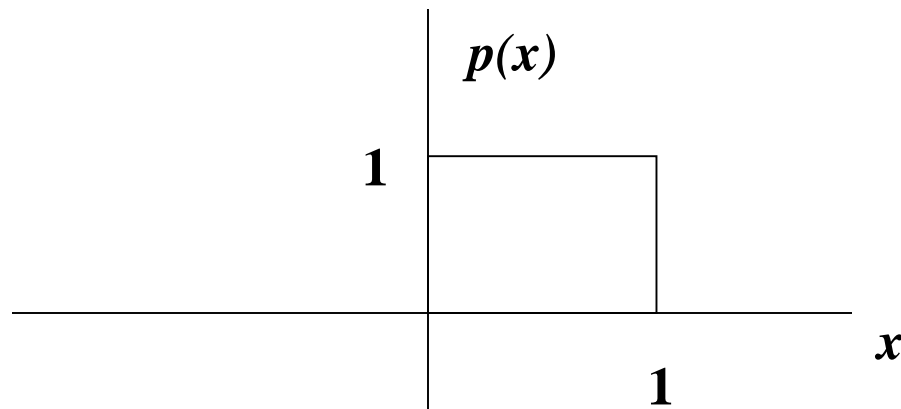


$$P(1 \leq x \leq 2) = \int_1^2 e^{-x} = -e^{-x} \Big|_1^2 = -e^{-2} - (-e^{-1}) = -.135 + .368 = .23$$

Example 2: Uniform distribution

The uniform distribution: all values are equally likely.

$$f(x) = 1, \text{ for } 1 \geq x \geq 0$$

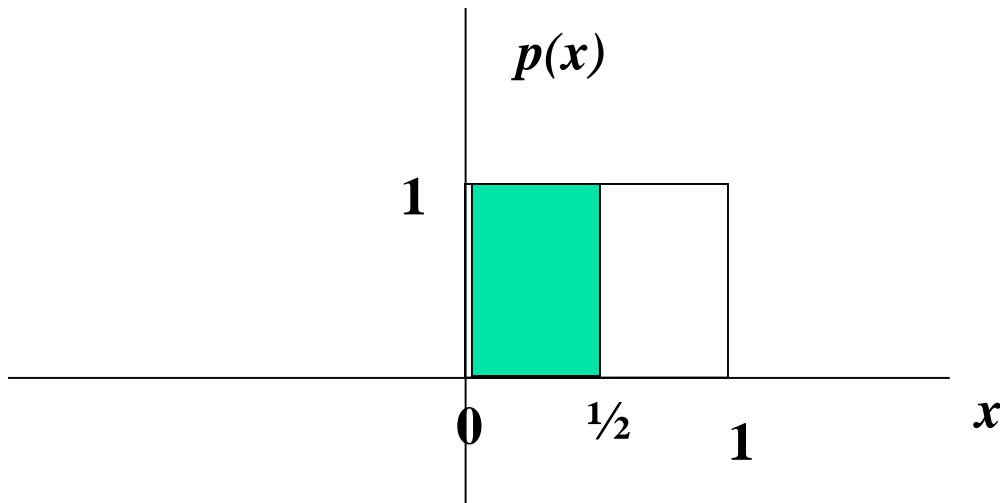


We can see it's a probability distribution because it integrates to 1 (the area under the curve is 1):

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

Example: Uniform distribution

What's the probability that x is between 0 and $1/2$?



Clinical Research Example:
When randomizing patients in an RCT, we often use a random number generator on the computer. These programs work by randomly generating a number between 0 and 1 (with equal probability of every number in between). Then a subject who gets $X < .5$ is control and a subject who gets $X > .5$ is treatment.

$$P(1/2 \geq x \geq 0) = 1/2$$



Expected Value and Variance

- All probability distributions are characterized by an expected value (mean) and a variance (standard deviation squared).



Expected value of a random variable

- Expected value is just the average or mean (μ) of random variable x .
- It's sometimes called a "weighted average" because more frequent values of X are weighted more highly in the average.
- It's also how we expect X to behave on-average over the long run ("frequentist" view again).



Expected value, formally

Discrete case:

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

Continuous case:

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



Symbol Interlude

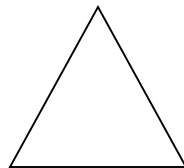
- $E(X) = \mu$
 - these symbols are used interchangeably



Example: expected value

- Recall the following probability distribution of ER arrivals:


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



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

Sample 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

- 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	7.2×10^{-8}



Expected Value

The probability function

$x\$$	$p(x)$
-1	.999999928
+ 2 million	7.2×10^{-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!



Expected Value

If you play the lottery every week for 10 years, what are your expected winnings or losses?

$$520 \times (-.86) = -\$447.20$$



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.



Expected value isn't everything though...

- Take the hit new show “Deal or No Deal”
- Everyone know the rules?
- Let's say you are down to two cases left. \$1 and \$400,000. The banker offers you \$200,000.
- So, Deal or No Deal?



Deal or No Deal...

- This could really be represented as a probability distribution and a non-random variable:

$x\$$	$p(x)$
+1	.50
+\$400,000	.50

$x\$$	$p(x)$
+\$200,000	1.0



Expected value doesn't help...

$x\$$	$p(x)$
+1	.50
+\$400,000	.50

$$\mu = E(X) = \sum_{\text{all } x} x_i p(x_i) = +1(.50) + 400,000(.50) = 200,000$$

$x\$$	$p(x)$
+\$200,000	1.0

$$\mu = E(X) = 200,000$$



How to decide?

Variance!

- **If you take the deal, the variance/standard deviation is 0.**
- **If you don't take the deal, what is average deviation from the mean?**
- **What's your gut guess?**



Variance/standard deviation

$$\sigma^2 = \text{Var}(x) = E(x - \mu)^2$$

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

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



Variance, continuous

Discrete case:

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

Continuous case?:

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



Symbol Interlude


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



Similarity to empirical variance

The variance of a sample: $s^2 =$

$$\frac{\sum_{i=1}^N (x_i - \bar{x})^2}{n-1} = \sum_{i=1}^N (x_i - \bar{x})^2 \left(\frac{1}{n-1} \right)$$



Division by $n-1$ reflects the fact that we have lost a "degree of freedom" (piece of information) because we had to estimate the sample mean before we could estimate the sample variance.



Variance

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

$$\begin{aligned}\sigma^2 &= \sum_{\text{all } x} (x_i - \mu)^2 p(x_i) = \\ &= (1 - 200,000)^2 (.5) + (400,000 - 200,000)^2 (.5) = 200,000^2 \\ \sigma &= \sqrt{200,000^2} = 200,000\end{aligned}$$

Now you examine your personal risk tolerance...



Practice Problem

On the roulette wheel, $X=1$ with probability $18/38$ and $X=-1$ with probability $20/38$.

We already calculated the mean to be $= -\$0.053$. What's the variance of X ?



Answer

$$\begin{aligned}\sigma^2 &= \sum_{\text{all } x} (x_i - \mu)^2 p(x_i) \\&= (+1 - -.053)^2 (18/38) + (-1 - -.053)^2 (20/38) \\&= (1.053)^2 (18/38) + (-1 + .053)^2 (20/38) \\&= (1.053)^2 (18/38) + (-.947)^2 (20/38) \\&= .997 \\ \sigma &= \sqrt{.997} = .99\end{aligned}$$

Standard deviation is \$.99. Interpretation: On average, you're either 1 dollar above or 1 dollar below the mean, which is just under zero. Makes sense!



Review Question 3

The expected value and variance of a coin toss ($H=1$, $T=0$) are?

- a. .50, .50
- b. .50, .25
- c. .25, .50
- d. .25, .25



Review Question 3

The expected value and variance of a coin toss are?

- a. .50, .50
- b. .50, .25**
- c. .25, .50
- d. .25, .25



Important discrete probability distribution: The binomial



Binomial Probability Distribution

- A fixed number of observations (trials), n
 - e.g., 15 tosses of a coin; 20 patients; 1000 people surveyed
- A binary outcome
 - e.g., head or tail in each toss of a coin; disease or no disease
 - Generally called “success” and “failure”
 - Probability of success is p , probability of failure is $1 - p$
- Constant probability for each observation
 - e.g., Probability of getting a tail is the same each time we toss the coin



Binomial distribution

Take the example of 5 coin tosses.
What's the probability that you flip
exactly 3 heads in 5 coin tosses?



Binomial distribution

Solution:

One way to get exactly 3 heads: HHHTT

What's the probability of this exact arrangement?

$$\begin{aligned} &P(\text{heads}) \times P(\text{heads}) \times P(\text{heads}) \times P(\text{tails}) \times P(\text{tails}) \\ &= (1/2)^3 \times (1/2)^2 \end{aligned}$$

Another way to get exactly 3 heads: THHHT

$$\begin{aligned} \text{Probability of this exact outcome} &= (1/2)^1 \times (1/2)^3 \\ &\times (1/2)^1 = (1/2)^3 \times (1/2)^2 \end{aligned}$$

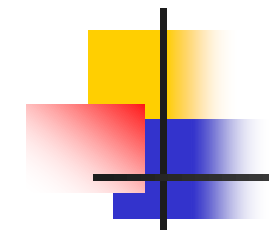


Binomial distribution

In fact, $(1/2)^3 \times (1/2)^2$ is the probability of each unique outcome that has exactly 3 heads and 2 tails.

So, the overall probability of 3 heads and 2 tails is:

$(1/2)^3 \times (1/2)^2 + (1/2)^3 \times (1/2)^2 + (1/2)^3 \times (1/2)^2$
+ for as many unique arrangements as there are—but how many are there??



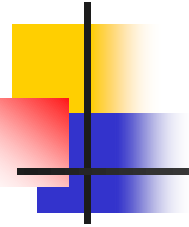
$\binom{5}{3}$ ways to arrange 3 heads in 5 trials

${}_5C_3 = 5!/3!2! = 10$

Outcome	Probability
THHHT	$(1/2)^3 \times (1/2)^2$
HHHTT	$(1/2)^3 \times (1/2)^2$
TTHHH	$(1/2)^3 \times (1/2)^2$
HTTHH	$(1/2)^3 \times (1/2)^2$
HHTTH	$(1/2)^3 \times (1/2)^2$
HTHHT	$(1/2)^3 \times (1/2)^2$
THTHH	$(1/2)^3 \times (1/2)^2$
HTHTH	$(1/2)^3 \times (1/2)^2$
HHTHT	$(1/2)^3 \times (1/2)^2$
THHHT	$(1/2)^3 \times (1/2)^2$
10 arrangements $\times (1/2)^3 \times (1/2)^2$	

The probability of each unique outcome (note: they are all equal)

Factorial review: $n! = n(n-1)(n-2)\dots$

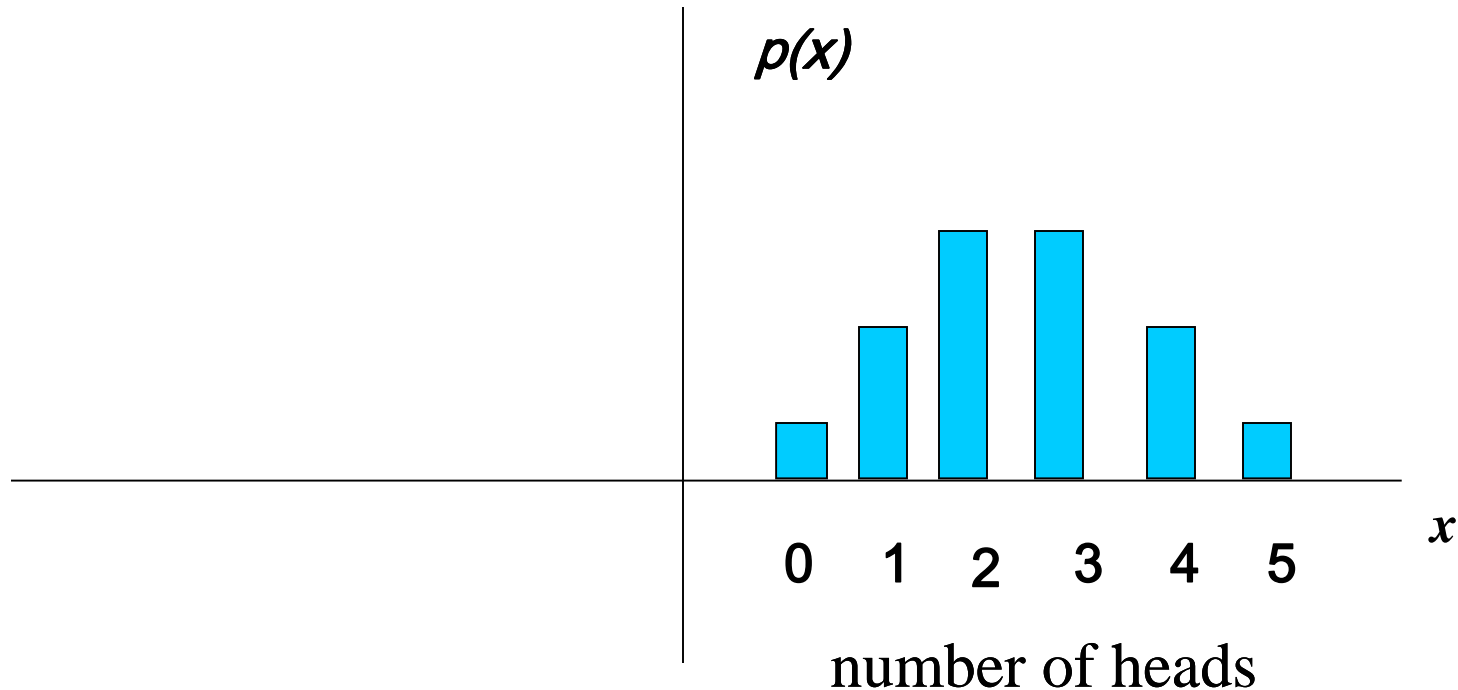


$$\therefore P(3 \text{ heads and } 2 \text{ tails}) = \binom{5}{3} \times P(\text{heads})^3 \times P(\text{tails})^2 =$$
$$10 \times (1/2)^5 = 31.25\%$$

Binomial distribution

function:

X = the number of heads tossed in 5 coin
tosses



Binomial distribution, generally

Note the general pattern emerging \rightarrow if you have only two possible outcomes (call them 1/0 or yes/no or success/failure) in n independent trials, then the probability of exactly X “successes”=

The diagram shows the binomial distribution formula $\binom{n}{X} p^X (1-p)^{n-X}$ enclosed in a purple rectangular box. Four arrows point from descriptive text to parts of the formula: one from ' n ' to ' $n = \text{number of trials}$ ', one from ' X ' to ' $X = \# \text{ successes out of } n \text{ trials}$ ', one from ' p ' to ' $p = \text{probability of success}$ ', and one from ' $1-p$ ' to ' $1-p = \text{probability of failure}$ '.

$$\binom{n}{X} p^X (1-p)^{n-X}$$

$n = \text{number of trials}$

$X = \#$
successes
out of n
trials

$p =$
probability of
success

$1-p = \text{probability}$
of failure



Binomial distribution: example

- If I toss a coin 20 times, what's the probability of getting exactly 10 heads?

$$\binom{20}{10} (.5)^{10} (.5)^{10} = .176$$



Binomial distribution: example

- If I toss a coin 20 times, what's the probability of getting 2 or fewer heads?

$$\binom{20}{0} (.5)^0 (.5)^{20} = \frac{20!}{20!0!} (.5)^{20} = 9.5 \times 10^{-7} +$$

$$\binom{20}{1} (.5)^1 (.5)^{19} = \frac{20!}{19!1!} (.5)^{20} = 20 \times 9.5 \times 10^{-7} = 1.9 \times 10^{-5} +$$

$$\binom{20}{2} (.5)^2 (.5)^{18} = \frac{20!}{18!2!} (.5)^{20} = 190 \times 9.5 \times 10^{-7} = 1.8 \times 10^{-4}$$

$$= 1.8 \times 10^{-4}$$



****All probability distributions are characterized by an expected value and a variance:**


If X follows a binomial distribution with parameters n and p : **$X \sim \text{Bin}(n, p)$**

Then:

$$E(X) = np$$

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

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



Note: the variance will always lie between

$0 * N - .25 * N$

$p(1-p)$ reaches maximum at $p=.5$

$P(1-p) = .25$



Practice Problem

- 1. You are performing a cohort study. If the probability of developing disease in the exposed group is .05 for the study duration, then if you (randomly) sample 500 exposed people, how many do you expect to develop the disease? Give a margin of error (+/- 1 standard deviation) for your estimate.
- 2. What's the probability that **at most** 10 exposed people develop the disease?



Answer

1. How many do you expect to develop the disease? Give a margin of error (+/- 1 standard deviation) for your estimate.

$$X \sim \text{binomial}(500, .05)$$

$$E(X) = 500 (.05) = 25$$

$$\text{Var}(X) = 500 (.05) (.95) = 23.75$$

$$\text{StdDev}(X) = \text{square root}(23.75) = 4.87$$

$$\therefore 25 \pm 4.87$$



Answer

2. What's the probability that **at most** 10 exposed subjects develop the disease?

This is asking for a CUMULATIVE PROBABILITY: the probability of 0 getting the disease or 1 or 2 or 3 or 4 or up to 10.

$$P(X \leq 10) = P(X=0) + P(X=1) + P(X=2) + P(X=3) + P(X=4) + \dots + P(X=10) =$$

$$\binom{500}{0} (.05)^0 (.95)^{500} + \binom{500}{1} (.05)^1 (.95)^{499} + \binom{500}{2} (.05)^2 (.95)^{498} + \dots + \binom{500}{10} (.05)^{10} (.95)^{490} < .01$$



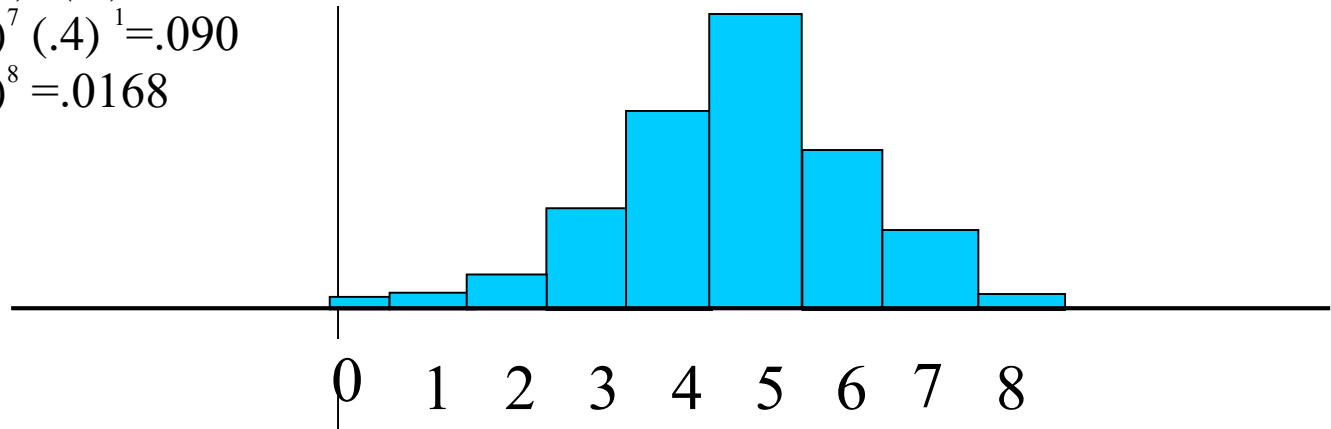
Practice Problem:

You are conducting a case-control study of smoking and lung cancer. If the probability of being a smoker among lung cancer cases is .6, what's the probability that in a group of 8 cases you have:

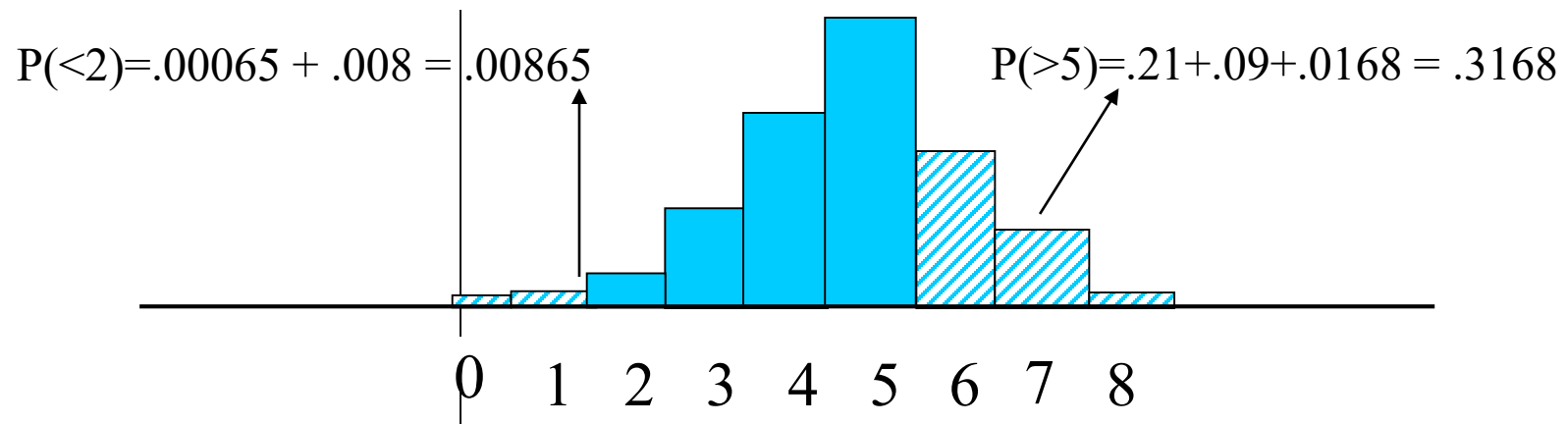
- a. Less than 2 smokers?
- b. More than 5?
- c. What are the expected value and variance of the number of smokers?

Answer

X	P(X)
0	$1(.4)^8 = .00065$
1	$8(.6)^1 (.4)^7 = .008$
2	$28(.6)^2 (.4)^6 = .04$
3	$56(.6)^3 (.4)^5 = .12$
4	$70(.6)^4 (.4)^4 = .23$
5	$56(.6)^5 (.4)^3 = .28$
6	$28(.6)^6 (.4)^2 = .21$
7	$8(.6)^7 (.4)^1 = .090$
8	$1(.6)^8 = .0168$



Answer, continued



$$E(X) = 8 (.6) = 4.8$$

$$\text{Var}(X) = 8 (.6) (.4) = 1.92$$

$$\text{StdDev}(X) = 1.38$$



Review Question 4

In your case-control study of smoking and lung-cancer, 60% of cases are smokers versus only 10% of controls. What is the odds ratio between smoking and lung cancer?

- a. 2.5
- b. 13.5
- c. 15.0
- d. 6.0
- e. .05



Review Question 4

In your case-control study of smoking and lung-cancer, 60% of cases are smokers versus only 10% of controls. What is the odds ratio between smoking and lung cancer?

- a. 2.5
- b. 13.5**
- c. 15.0
- d. 6.0
- e. .05

$$\frac{\frac{.6}{.4}}{\frac{.1}{.9}} = \frac{3}{2} \times \frac{9}{1} = \frac{27}{2} = 13.5$$



Review Question 5

What's the probability of getting exactly 5 heads in 10 coin tosses?

a.

$$\binom{10}{0} (.50)^5 (.50)^5$$

b.

$$\binom{10}{5} (.50)^5 (.50)^5$$

c.

$$\binom{10}{5} (.50)^{10} (.50)^5$$

d.

$$\binom{10}{10} (.50)^{10} (.50)^0$$



Review Question 5

What's the probability of getting exactly 5 heads in 10 coin tosses?

a.

$$\binom{10}{0} (.50)^5 (.50)^5$$

b.

$$\binom{10}{5} (.50)^5 (.50)^5$$

c.

$$\binom{10}{5} (.50)^{10} (.50)^5$$

d.

$$\binom{10}{10} (.50)^{10} (.50)^0$$



Review Question 6

A coin toss can be thought of as an example of a binomial distribution with $N=1$ and $p=.5$. What are the expected value and variance of a coin toss?

- a. .5, .25
- b. 1.0, 1.0
- c. 1.5, .5
- d. .25, .5
- e. .5, .5



Review Question 6

A coin toss can be thought of as an example of a binomial distribution with $N=1$ and $p=.5$. What are the expected value and variance of a coin toss?

- a. **.5, .25**
- b. 1.0, 1.0
- c. 1.5, .5
- d. .25, .5
- e. .5, .5



Review Question 7

If I toss a coin 10 times, what is the expected value and variance of the number of heads?

- a. 5, 5
- b. 10, 5
- c. 2.5, 5
- d. 5, 2.5
- e. 2.5, 10



Review Question 7

If I toss a coin 10 times, what is the expected value and variance of the number of heads?

- a. 5, 5
- b. 10, 5
- c. 2.5, 5
- d. **5, 2.5**
- e. 2.5, 10



Review Question 8

In a randomized trial with $n=150$, the goal is to randomize half to treatment and half to control. The number of people randomized to treatment is a random variable X . What is the probability distribution of X ?

- a. $X \sim \text{Normal}(\mu=75, \sigma=10)$
- b. $X \sim \text{Exponential}(\mu=75)$
- c. $X \sim \text{Uniform}$
- d. $X \sim \text{Binomial}(N=150, p=.5)$
- e. $X \sim \text{Binomial}(N=75, p=.5)$



Review Question 8

In a randomized trial with $n=150$, every subject has a 50% chance of being randomized to treatment. The number of people randomized to treatment is a random variable X . What is the probability distribution of X ?

- a. $X \sim \text{Normal}(\mu=75, \sigma=10)$
- b. $X \sim \text{Exponential}(\mu=75)$
- c. $X \sim \text{Uniform}$
- d. **$X \sim \text{Binomial}(N=150, p=.5)$**
- e. $X \sim \text{Binomial}(N=75, p=.5)$



Review Question 9

In the same RCT with $n=150$, if 69 end up in the treatment group and 81 in the control group, how far off is that from expected?

- a. Less than 1 standard deviation
- b. 1 standard deviation
- c. Between 1 and 2 standard deviations
- d. More than 2 standard deviations



Review Question 9

In the same RCT with $n=150$, if 69 end up in the treatment group and 81 in the control group, how far off is that from expected?

- a. Less than 1 standard deviation
- b. **1 standard deviation**
- c. Between 1 and 2 standard deviations
- d. More than 2 standard deviations

Expected = 75

81 and 69 are both 6 away from the expected.

Variance = $150(.25) = 37.5$

Std Dev $\cong 6$

Therefore, about 1 SD away from expected.



Proportions...

- The binomial distribution forms the basis of statistics for proportions.
- A proportion is just a binomial count divided by n .
 - For example, if we sample 200 cases and find 60 smokers, $X=60$ but the observed proportion $=.30$.
- Statistics for proportions are similar to binomial counts, but differ by a factor of n .



Stats for proportions

For binomial: $\mu_x = np$

$$\sigma_x^2 = np(1-p)$$

$$\sigma_x = \sqrt{np(1-p)}$$

Differs by
a factor of
n.

For proportion:

$$\mu_{\hat{p}} = p$$

$$\sigma_{\hat{p}}^2 = \frac{np(1-p)}{n^2} = \frac{p(1-p)}{n}$$

Differs
by a
factor
of n.

P-hat stands for "sample
proportion."

$$\sigma_{\hat{p}} = \sqrt{\frac{p(1-p)}{n}}$$



It all comes back to normal...

- Statistics for proportions are based on a normal distribution, because the binomial can be approximated as normal if $np > 5$