STAT 22000 Lecture Slides
Binomial Distributions

Yibi Huang
Department of Statistics
University of Chicago

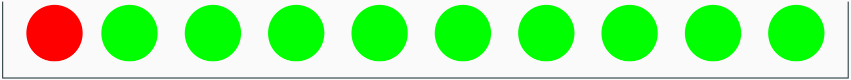
Coverage

- Binomial distribution 3.3.

Please skip section 3.2 and 3.5

Binomial distribution

Five draws are made at random with replacement from a box containing one red ball and 9 green balls.



What is the probability that the first two draws are Red and the next 3 are Green?

$$\begin{aligned}P(\mathsf{R\,R\,G\,G\,G}) &= P(\mathsf{R})\cdot P(\mathsf{R})\cdot P(\mathsf{G})\cdot P(\mathsf{G})\cdot P(\mathsf{G})\\ &= 0.1\times 0.1\times 0.9\times 0.9\times 0.9.\end{aligned}$$

The outcomes of the 5 draws are independent of each other as the draws are made with replacement. So we can use the multiplication rule for independent events here.

What is the probability of getting **exactly two Reds in 5 draws**?

Is it also equal to

$$0.1 \times 0.1 \times 0.9 \times 0.9 \times 0.9?$$

There are 10 possible orderings of the 2 Reds and the 3 Greens.

Possible Orders	Probability
RRGGG	$0.1 \times 0.1 \times 0.9 \times 0.9 \times 0.9 = (0.1)^2(0.9)^3$
RGRGG	$0.1 \times 0.9 \times 0.1 \times 0.9 \times 0.9 = (0.1)^2(0.9)^3$
RGGRG	$0.1 \times 0.9 \times 0.9 \times 0.1 \times 0.9 = (0.1)^2(0.9)^3$
RGGGR	$0.1 \times 0.9 \times 0.9 \times 0.9 \times 0.1 = (0.1)^2(0.9)^3$
GRRGG	$0.9 \times 0.1 \times 0.1 \times 0.9 \times 0.9 = (0.1)^2(0.9)^3$
GRGRG	$0.9 \times 0.1 \times 0.9 \times 0.1 \times 0.9 = (0.1)^2(0.9)^3$
GRGGR	$0.9 \times 0.1 \times 0.9 \times 0.9 \times 0.1 = (0.1)^2(0.9)^3$
GGRRG	$0.9 \times 0.9 \times 0.1 \times 0.1 \times 0.9 = (0.1)^2(0.9)^3$
GGRGR	$0.9 \times 0.9 \times 0.1 \times 0.9 \times 0.1 = (0.1)^2(0.9)^3$
GGGRR	$0.9 \times 0.9 \times 0.9 \times 0.1 \times 0.1 = (0.1)^2(0.9)^3$

P(exactly 2 Reds in 5 draws) is the sum of the probabilities of the 10 cases above because the 10 cases are *disjoint*. So

$$\mathrm{P}(\text{exactly 2 Reds in 5 draws}) = 10 \times (0.1)^2(0.9)^3.$$

What is P(getting **exactly k Reds in n draws**)?

We have to consider all possible orderings of the k Reds and the $n - k$ Greens.

Possible Orders	Probability
$\underbrace{\text{RRR}\ldots\text{R}}_{k}\,\underbrace{\text{G}\ldots\text{G}}_{n-k}$	$\underbrace{0.1\times\ldots\times0.1}_{k}\times\underbrace{0.9\times\ldots\times0.9}_{n-k} = (0.1)^k(0.9)^{n-k}$
$\text{RGR}\ldots\text{R}\,\text{G}\ldots\text{G}$	$0.1\cdot0.9\cdot0.1\ldots0.1\times0.9\ldots0.9 = (0.1)^k(0.9)^{n-k}$
$\vdots$	$\vdots$

Note

- the events for different orderings are *disjoint*, and
- each occurs with identical probability $(0.1)^k(0.9)^{n-k}$.

By the Addition Rule, P(exactly k Reds in n draws) equals

$$(\text{\# of ways to order } k \text{ Reds and } n-k \text{ Greens}) \times (0.1)^k(0.9)^{n-k}$$

Factorial

The notation $n!$, read n *factorial*, is defined as

$$n! = 1 \times 2 \times 3 \times \ldots \times (n-1) \times n$$

e.g.,

$$\begin{aligned} 1\,! &= 1, & 3\,! &= 1 \times 2 \times 3 = 6, \\ 2\,! &= 1 \times 2 = 2, & 4\,! &= 1 \times 2 \times 3 \times 4 = 24. \end{aligned}$$

By convention,

$$0\,! = 1.$$

Binomial Coefficients

The number of ways to order k Reds and $n-k$ Greens equals

$$\binom{n}{k} = \frac{n!}{k!(n-k)!}$$

- $\binom{n}{k}$ is read as "*n choose k*', also denoted as ${}_nC_k$, or C_k^n.

e.g.,

$$\binom{5}{2} = \frac{5\,!}{2\,! \times (5-2)\,!} = \frac{5 \times 4 \times 3 \times 2 \times 1}{(2 \times 1)(3 \times 2 \times 1)} = \frac{5 \times 4}{2 \times 1} = 10,$$

$$\binom{n}{n} = \frac{n\,!}{n\,! \times 0\,!} = \frac{n\,!}{n\,! \times 1} = 1$$

You can also use R for these calculations:

```
> choose(5,2)
[1] 10
```

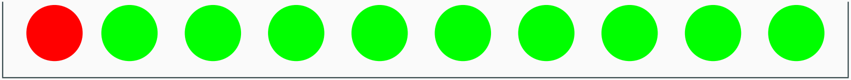
- $\binom{n}{0} = \dfrac{n\,!}{0\,! \times n\,!} = 1$
  $\Rightarrow$ there is only 1 way to order 0 Reds and n Greens

- $\binom{n}{n} = \dfrac{n\,!}{n\,! \times 0\,!} = 1$
  $\Rightarrow$ there is only 1 way to order n Reds and 0 Green

- $\binom{n}{1} = \dfrac{n\,!}{1\,! \times (n-1)\,!} = n$
  $\Rightarrow$ there are n ways to order 1 Red and $n - 1$ Greens

- $\binom{n}{n-1} = \dfrac{n\,!}{(n-1)\,! \times 1\,!} = n$
  $\Rightarrow$ there are n ways to order $n - 1$ Reds and 1 Green

When n draws are made at random with replacement from a box contains one red ball and 9 green ones,



the probability to get exactly k Reds (and $n - k$ Greens) equals

$$\binom{n}{k}(0.1)^k(0.9)^{n-k}.$$

Such calculation can be generalized to other similar problems and the general formula is called the *Binomial Formula*.

Bernoulli Trials

A random trial having only 2 possible outcomes (Success, Failure) is called a *Bernoulli trial*, e.g.,

- whether a coin lands heads or tails when tossing a coin
- whether one gets a six or not a six when rolling a die
- whether a drug works on a patient or not
- whether a electronic device is defected
- whether a subject answers Yes or No to a survey question

Binomial Formula

Suppose n **independent** Bernoulli trials are to be performed, each of which results in

- a *success* with probability p and
- a *failure* with probability $1 - p$.

The probability of getting k successes and $n - k$ failures in n Bernoulli trials is given by

(# of ways to order the k successes and $n - k$ failures) $\times p^k(1-p)^{n-k}$

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

Binomial Distribution

Suppose n **independent** Bernoulli trials are to be performed, each of which results in

- a *success* with probability p and
- a *failure* with probability $1 - p$.

If we define

$$X = \text{the number of successes that occur in the } n \text{ trials},$$

then X is said to have a *binomial distribution* with parameters (n, p), denoted as

$$X \sim Bin(n, p).$$

with the probability distribution

$$\mathrm{P}(X = k) = \binom{n}{k} p^k (1 - p)^{n-k}, \quad k = 0, 1, \ldots, n$$

Conditions Required to be Binomial

Condition that needs to be met for the binomial formula to be applicable:

1. each trial outcome must be classified as a *success* or a *failure*
2. the probability of success, p, must be the same for each trial
3. the number of trials, n, must be fixed
4. the trials must be independent

Binomial or Not?

Rolling a die 10 times, what is the probability of getting exactly 3 aces?

- a trial: whether one gets an ace when rolling a die once
- prob. of success $p = 1/6$
- number of trials $n = 10$
- the trials (rolls) are independent

So, it's okay to use the Binomial formula.

$$\begin{aligned}
\mathrm{P}(\text{3 aces in 10 rolls}) &= \frac{10!}{3!\,7!}\left(\frac{1}{6}\right)^3\left(1-\frac{1}{6}\right)^7 \\
&= \frac{10\times 9\times 8\times (7!)}{(3\times 2\times 1)(7!)}\left(\frac{1}{6}\right)^3\left(\frac{5}{6}\right)^7 \\
&= 120\left(\frac{1}{6}\right)^3\left(\frac{5}{6}\right)^7 \approx 0.155
\end{aligned}$$

Binomial or Not?

Rolling a die continuously, is the probability of getting the 3rd aces in the 10th roll equals to

$$\frac{10!}{3!\,7!}\left(\frac{1}{6}\right)^3\left(1-\frac{1}{6}\right)^7?$$

No. The number of trials (sample size) is not determined in advance.

Binomial or Not?

A SRS of 50 from all UC undergrads are asked whether they are eligible to vote in the 2020 election. Let X be the number who reply yes. Is X binomial?

- a trial: a randomly selected student reply yes or not
- prob. of success $p =$ proportion of UC undergrads saying yes
- number of trials $n = 50$
- Strictly speaking, NOT binomial, because SRS draw subjects without replacement — trials are dependent
- Since the sample size 50 is only 1% of the population size (≈ 5000), trials are nearly independent
- So X is approx. binomial, $Bin(n = 50, p)$.

Binomial or Not?

Thirty of the 75 members in a fraternity are selected at random to interview. One question is "*are you eligible to vote in the 2020 election?*"

Suppose the truth is that 60% of the 75 members are eligible.
Let X be the count in your sample who say "yes."
Is X (at least approximately) $\sim Bin(n = 30, p = 0.6)$?

No. The sample size 30 is large relative to the population size 75. The SRS draws are not independent.

Example: Voter Turnout

Suppose that the turnout rate for a certain election in the Chicago area is 55%. Among a random sample of 10 eligible voters, what is the probability that exactly 6 voted? Exactly 8 voted?

Let X = the number of people that will vote in a sample of size 10.

$X \sim Bin(n = 10, p = 0.55)$

$$\mathrm{P}(X = 6) = \binom{10}{6} \times 0.55^6 \times 0.45^4 = 210 \times 0.55^6 \times 0.45^4 \approx 0.238$$

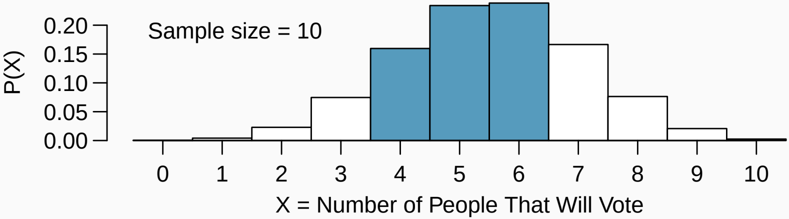
$$\mathrm{P}(X = 8) = \binom{10}{8} \times 0.55^8 \times 0.45^2 = 45 \times 0.55^8 \times 0.45^2 \approx 0.0763.$$

```
> dbinom(6, size = 10, p = 0.55)
[1] 0.2383666
> dbinom(8, size = 10, p = 0.55)
[1] 0.07630255
```

Example: Voter Turnout (Cont'd)

In a sample of size 10, what is the probability that 4 to 6 of them voted?

$$\begin{aligned} \mathrm{P}(4 \le X \le 6) &= \mathrm{P}(X = 4) + \mathrm{P}(X = 5) + \mathrm{P}(X = 6) \\ &= \binom{10}{4} 0.55^4 0.45^6 + \binom{10}{5} 0.55^5 0.45^5 + \binom{10}{6} 0.55^6 0.45^5 \\ &\approx 0.160 + 0.234 + 0.238 = 0.632 \end{aligned}$$



Expected Value & SD of Bin$(n = 1, p)$

A Binomial random variable $X \sim$ Bin(n, p) with $n = 1$ can only take value 0 or 1 with the distribution below

value of X	0	1
probability	$1 - p$	p

The expected value, variance, and SD of Bin$(n = 1, p)$ can be calculated as follows.

$$\begin{aligned}
\mathrm{E}(X) &= \sum\nolimits_{x=0,1} xP(X = x) = 0 \cdot (1 - p) + 1 \cdot p = \boxed{p},\\
\mathrm{V}(X) &= \sum\nolimits_{x=0,1} (x - \underbrace{\mathrm{E}(X)}_{=p})^2 P(X = x)\\
&= (0 - p)^2 \cdot (1 - p) + (1 - p)^2 \cdot p\\
&= p(1 - p)[p + (1 - p)] = p(1 - p)\\
\mathrm{SD}(X) &= \sqrt{\mathrm{V}(X)} = \boxed{\sqrt{p(1 - p)}}
\end{aligned}$$

For general n, recall a Binomial random variable $X \sim \text{Bin}(n, p)$ is the total number of successes obtained in n independent Bernoulli trials. For each of the n trials, define

$$X_i = \begin{cases} 1 & \text{if success in the } i\text{th trial} \\ 0 & \text{if failure in the } i\text{th trial} \end{cases}$$

$$\begin{aligned} \text{Then } X &= \text{the number of successes obtained in the } n \text{ trials} \\ &= X_1 + X_2 + \ldots + X_n, \end{aligned}$$

The expected value and variance of X are thus

$$\mathrm{E}(X) = \underbrace{\mathrm{E}(X_1)}_{=p} + \underbrace{\mathrm{E}(X_2)}_{=p} + \cdots + \underbrace{\mathrm{E}(X_n)}_{=p} = np$$

$$\mathrm{V}(X) = \underbrace{\mathrm{V}(X_1)}_{=p(1-p)} + \underbrace{\mathrm{V}(X_2)}_{=p(1-p)} + \cdots + \underbrace{\mathrm{V}(X_n)}_{=p(1-p)} = np(1-p)$$

since X_i's are indep. and each with mean p and variance $p(1-p)$ as $X_i \sim \text{Bin}(n = 1, p)$.

Expected Value & SD of Bin(n, p)

For a Binomial random variable $X \sim Bin(n, p)$, the mean and the SD are respectively

$$\mu = \mathrm{E}(X) = np \qquad \sigma = \mathrm{SD}(X) = \sqrt{np(1-p)}$$

Note the SD increases proportionally to $\sqrt{n}$, not n.

Suppose the turnout rate in an election in a city is 55%. Among a random sample of 100 eligible voters taken from the city, how many do you expect will vote? With what SD?

$$\begin{aligned} X &\sim \text{Bin}(n = 100, p = 0.55) \\ \mathrm{E}(X) &= np = 100 \times 0.55 = 55, \\ \mathrm{SD}(X) &= \sqrt{np(1-p)} = \sqrt{100(0.55)(1-0.55)} \approx 4.97 \end{aligned}$$

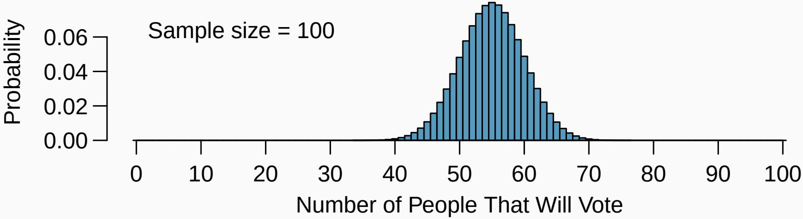
Note: Mean and SD of a binomial might not be whole numbers, and that is alright,

Unusual Observations

As *observations over 2 SDs away from the mean (expected value) are considered unusual*, from the $\mathrm{E}(X)$ and the SD we just computed, we can calculate a range for the possible number of subjects that will turn out in a sample of size 100

$$55 \pm (2 \times 4.97) \approx (45, 65)$$

So, the sample proportion will likely to be between 45% and 65%.



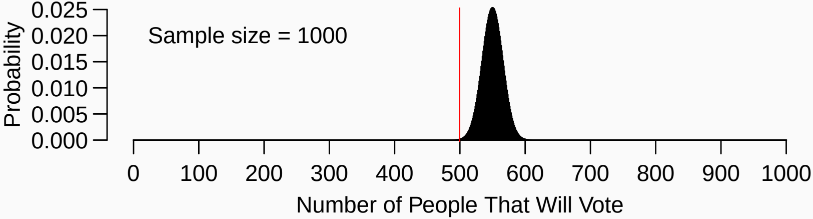
Suppose that the turnout rate for an election in a city is 55%. Is it surprising to obtain a sample of size 1000 that only 500 turn out?

$$\mathrm{E}(X) = np = 1,000 \times 0.55 = 550$$

$$\mathrm{SD}(X) = \sqrt{np(1-p)} = \sqrt{1000 \times 0.55 \times 0.45} \approx 15.73$$

$Z - score$ of observation $Z = (x - \mathrm{E}(X))/\mathrm{SD}(X) = \frac{500-550}{15.73} = -3.18$
500 is more than 3 SDs below the expected value.

If the true turnout rate is 55%, it'll be SURPRISING to obtain a sample of size 1000 that only 500 turn out.



Law of Large Numbers Revisit

Flipping a coin n times, the number of heads obtained $H(n)$ has a Bin(n, p) distribution, where p is the prob. for the coin to land heads in a single flip. The size of the difference

$$H(n) - np = \text{observed count of heads} - \text{expected count of heads}$$

is about $\sqrt{np(1-p)}$, the SD of $Bin(n, p)$, which

- will increase as n goes up,
- but is small relative to n.

For example, tossing a fair coin ($p = 0.5$) n times,

- for $n = 100$ tosses, the observed count of heads can be off from the expected count $np = 100 \times 0.5 = 50$ by about $\sqrt{np(1-p)} = \sqrt{100(0.5)(0.5)} = 5$, which is 5% difference in percentage.
- for $n = 10000$ tosses, the observed count of heads can be off from the expected count $np = 10000 \times 0.5 = 5000$ by a greater difference $\sqrt{np(1-p)} = \sqrt{10000(0.5)(0.5)} = 50$, which is, however, only 0.5% difference in percentage.

Normal Approximation to the Binomial Distribution

Shapes of Binomial Distributions

For this activity you will use a web applet. Go to

https:// gallery.shinyapps.io/ dist_calc/

and choose [Binomial] in the drop down menu on the left.

- Set n to 20 and the p to 0.15. Describe the shape of the distribution of $Bin(n = 20, p = 0.15)$.
- Keeping p constant at 0.15, and increase n, what happens to the shape of the distribution?
- Keeping n constant at 30, and change p, what happens to the shape of the distribution?

Normal Approximation to the Binomial

The shape of the binomial distribution can be approximated by a normal distribution

$$Bin(n, p) \approx N\left(\mu = np, \quad \sigma = \sqrt{np(1-p)} \right)$$

as long as n is large enough.

Example: Voter Turnout

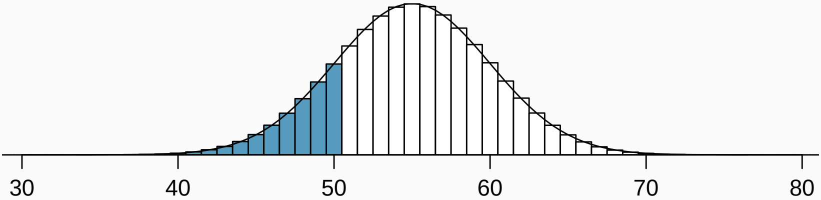
Suppose that the turnout rate for a certain election in some city is 55%. A random sample of 100 eligible voters is taken from the city. What is the probability that at most 50 of them will vote?

$X \sim Bin(n = 100, p = 0.55)$

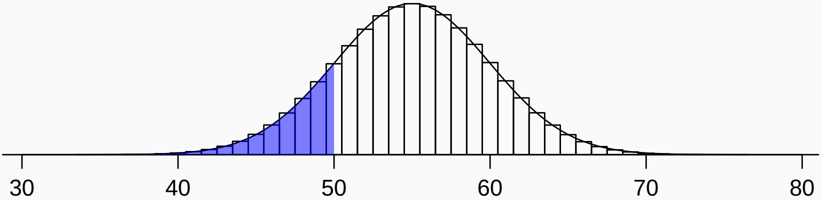
$$\begin{aligned} P(X \leq 50) &= P(X = 0) + P(X = 1) + P(X = 2) + \cdots + P(X = 50) \\ &= \binom{100}{0} \times 0.55^0 \times 0.45^{100} + \binom{100}{1} \times 0.55^1 \times 0.45^{99} \\ &\quad + \binom{100}{2} \times 0.55^2 \times 0.45^{98} + \ldots + \binom{100}{50} \times 0.55^{50} \times 0.45^{50} \end{aligned}$$

That's an awful lot of work...

Exact probability based on binomial formula (area of blue region in the histogram)



is approximated by the area of the blue shaded region under the normal curve.



Normal Approximation to the Binomial

Suppose that the turnout rate for a certain election in some city is 55%. A random sample of 100 eligible voters is taken from the city. What is the probability that at most 50 of them will vote?

$$Bin(n = 100, p = 0.55) \approx N(\mu = np = 55, \sigma = \sqrt{np(1-p)} = 4.97)$$

The Z-score of 50 is $(50 - 55)/4.97 \approx -1$. By normal approximation, the probability is about

$$P(X \leq 50) \approx P(Z < -1) = 0.1587.$$

The exact probability calculated using Binomial formula is

```
> pbinom(50, size = 100, p = 0.55)
[1] 0.182728
```

How Large is Large Enough?

- The size of n required depends on p. The closer p is to 0 or 1, the larger n needs to be
- A rule of thumb: n needs to be so large that the expected number of successes and failures are both at least 10.

$$np \geq 10 \qquad \text{and} \qquad n(1-p) \geq 10$$

Practice

Below are four pairs of Binomial distribution parameters. Which distribution can be approximated by the normal distribution?

(a) $n = 100, p = 0.95$

$n(1 - p) = 100 \times 0.05 = 5 < 10$

(b) $n = 25, p = 0.45$ .... *Answer*

$np = 25 \times 0.45 = 11.25 > 10;$

$n(1 - p) = 25 \times 0.55 = 13.75 > 10$

(c) $n = 150, p = 0.05$

$np = 150 \times 0.05 = 7.5 < 10$

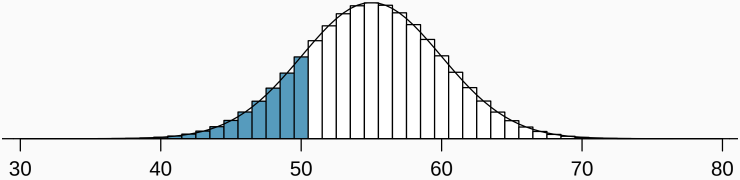
(d) $n = 500, p = 0.015$

$np = 500 \times 0.015 = 7.5 < 10$

Continuity Correction of the Normal Approximation to Binomial

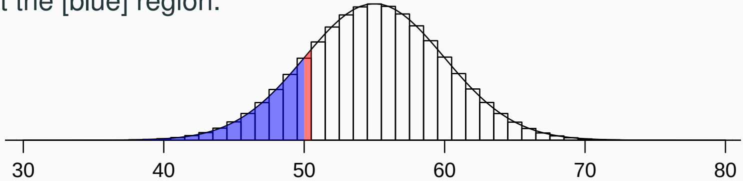
Continuity Correction of the Normal Approx. to Binomial

Exact probability (area of blue region in the histogram):



Observe the right end point of the blue region is at 50.5, not 50.

Better to approximate with the [blue+red] region below, rather than just the [blue] region.



Example (Voter Turnout)

Suppose that the turnout rate for a certain election in some city is 55%. A random sample of 100 eligible voters is taken from the city. What is the probability that at most 50 of them will turn out?

$$Bin(n = 100, p = 0.55) \approx N(\mu = np = 55, \sigma = \sqrt{np(1-p)} = 4.97)$$

The Z-score of 50.5 is $(50.5 - 55)/4.97 \approx -0.90$. By normal approximation, the probability is about

$$P(X \leq 50) = P(X \leq 50.5) \approx P(Z < -0.90) = 0.1841.$$

which is closer to exact probability 0.1827

```
> pbinom(50, size = 100, p = 0.55)
[1] 0.182728
```

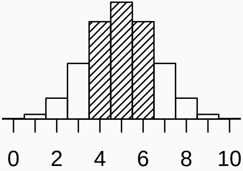
How to Adjust the Endpoints? $+0.5$ or -0.5?

Just make sure the interval contains the same set of integers before and after the adjustment.

For finding probabilities about $X \sim Bin(n = 10, p = 0.5)$ using the normal approx., what values should you convert to z-scores, and why?

3 to 7 exclusive

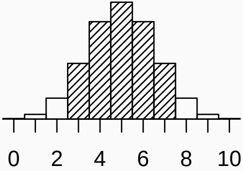
$P(3 < X < 7)$



$P(3.5 < X < 6.5)$

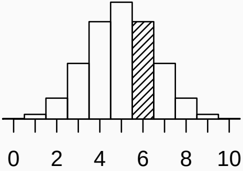
3 to 7 inclusive

$P(3 \le X \le 7)$



$P(2.5 < X < 7.5)$

$P(X = 6)$



$P(5.5 < X < 6.5)$