STAT 22000 Lecture Slides
Binomial Distributions

Yibi Huang
Department of Statistics
University of Chicago

Coverage

- Binomial distribution (3.4)

Please skip section 3.3 and 3.5.

Binomial distribution

Bernoulli Trials

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

- Tossing a coin, the outcomes are head or tail.
- Whether a drug works on a patient or not.
- Whether a electronic device is defected
- Whether a randomly selected eligible voter will vote in the 2016 general election or not

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

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

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.,

$$1\,! = 1, \qquad 3\,! = 1 \times 2 \times 3 = 6,$$
$$2\,! = 1 \times 2 = 2, \qquad 4\,! = 1 \times 2 \times 3 \times 4 = 24.$$

By convention,

$$0\,! = 1.$$

Binomial Coefficients

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

- which is the number of ways to choose k items, regardless of order, from a total of n distinct items
- $\binom{n}{k}$ is read as "*n choose k*", also denoted as ${}_nC_k$, or C_k^n.

e.g.,

$$\binom{6}{2} = \frac{6!}{4! \times 2!} = \frac{6 \times 5 \times 4 \times 3 \times 2 \times 1}{(4 \times 3 \times 2 \times 1)(2 \times 1)} = \frac{6 \times 5}{2 \times 1} = 15,$$

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

You can also use R for these calculations:

```
> choose(6,2)
[1] 15
```

- $\binom{n}{0} = 1$
  because there is only 1 way of getting 0 success in n trials

- $\binom{n}{n} = 1$
  because there is only 1 way of getting n successes in n trials

- $\binom{n}{1} = n$
  because there are n ways of getting 1 success in n trials

- $\binom{n}{n-1} = n$
  because there are n ways of getting $n - 1$ successes (i.e., 1 failures) in n trials

Binomial Formula

If Y has the binomial distribution $Bin(n, p)$, the probability to have k successes in n trials, $P(Y = k)$, is given as

$$P(Y = k) = \binom{n}{k} p^k (1-p)^{n-k} \quad \text{for } k = 0, 1, 2, \ldots, n.$$

Why is the binomial probability formula true?

See the next slide for an example.

Why is the Binomial Probability Formula True?

Let $Y = \#$ of success in 4 indep. trials, each with prob. p of success. So $Y \sim Bin(n = 4, p)$.

- There are 6 possible ways to get 2 successes ($Y = 2$):

  SSFF SFSF SFFS FSSF FSFS FFSS

- As trials are *independent*, by the multiplication rule,

$$\begin{aligned} P(\text{SSFF}) &= P(\text{S})P(\text{S})P(\text{F})P(\text{F}) \\ &= p \cdot p \cdot (1-p) \cdot (1-p) = p^2(1-p)^2 \\ P(\text{SFSF}) &= P(\text{S})P(\text{F})P(\text{S})P(\text{F}) \\ &= p \cdot (1-p) \cdot p \cdot (1-p) = p^2(1-p)^2 \\ &\vdots \end{aligned}$$

All 6 ways occur with prob. $p^2(1-p^2)$, because all have 2 successes and 2 failures

So $P(Y = 2) = (\#\text{ of ways}) \times (\text{prob. of each way}) = 6 \cdot p^2(1-p)^2$

Why is the Binomial Formula True?

In general, for $Y \sim Bin(n, p)$

$$
\begin{aligned}
P(Y = k) = &(\text{Number of ways to have exactly } k \text{ success}) \\
&\times P(\text{success in all the first } k \text{ trials} \\
&\qquad \text{and none of the last } n - k \text{ trials}) \\
= &(\text{Number of ways to choose } k \text{ out of } n) \times p^k (1-p)^{n-k} \\
= &\binom{n}{k} p^k (1-p)^{n-k}
\end{aligned}
$$

Conditions Required to be Binomial

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

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

Binomial or Not?

A SRS of 50 from all UC undergrads are asked whether they are eligible to vote in the 2016 presidential 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 "*will you vote in the 2016 presidential election?*"

Suppose the truth is that 60% of the 75 members would say "yes."
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.

Binomial or Not?

To study the prevalence rate of lung cancer among smokers, smokers are sampled until the number of lung cancer cases in the sample reaches 10. Let

$N =$ total number of smokers sampled, and

$p =$ proportion of lung cancer patients among smokers.

Is it true that

$$P(N = n) = \binom{n}{10} p^{10}(1-p)^{n-10}?$$

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

Example: Voter Turnout

Suppose that the turnout rate for the 2016 presidential election in the Chicago area is 55%. Among a random sample of 10 eligible voters, what is the probability that exactly 6 will vote? Exactly 8 will vote?

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

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

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

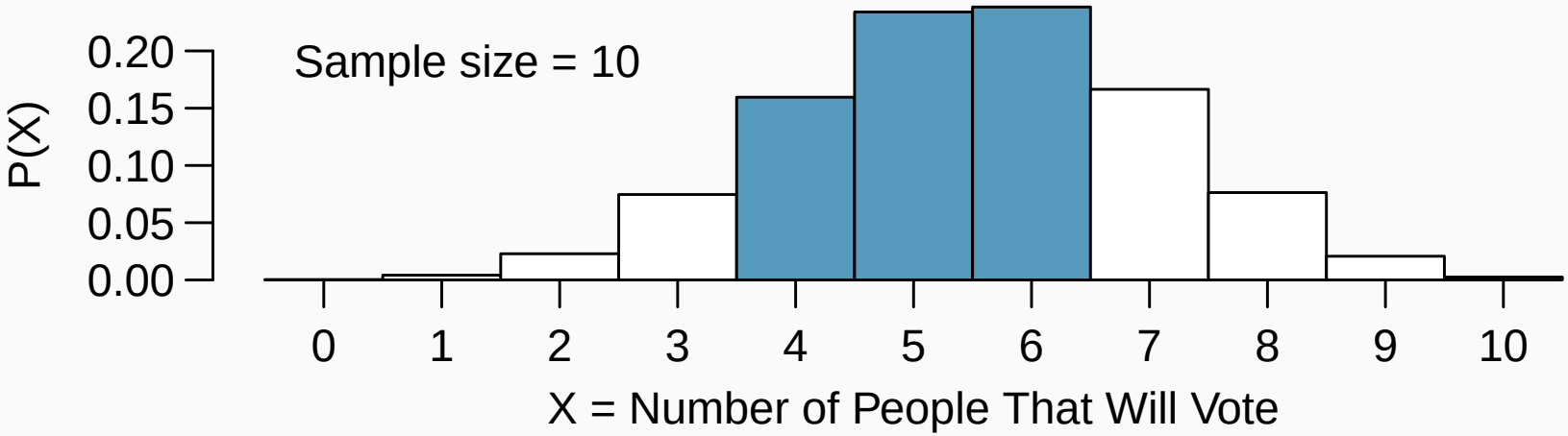
$$P(X = 8) = \binom{10}{8} \times 0.55^8 \times 0.45^2 = 45 \times 0.55^8 \times 0.45^2 \approx 0.0763.$$

```
> 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 will vote?

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



Expected Value of Binomial Distribution

Suppose that the turnout rate for the 2016 presidential election in the Chicago area is 55%. Among a random sample of 100 eligible voters in the Chicago area, how many do you expect will vote?

- Easy enough, $100 \times 0.55 = 55$.
- Or more formally, $\mu = np = 100 \times 0.55 = 55$.
- But this doesn't mean in every random sample of 100 people exactly 55 will vote. In some samples this value will be less, and in others more. How much would we expect this value to vary?

Mean and SD of Binomial Distribution

If $X \sim Bin(n, p)$, it can be shown that

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

- Going back to the turnout rate:

$$\sigma = \sqrt{np(1-p)} = \sqrt{100 \times 0.55 \times 0.45} \approx 4.97$$

- We expect 55 out of 100 randomly sampled eligible voters in the Chicago area to vote, with a standard deviation of 4.97.

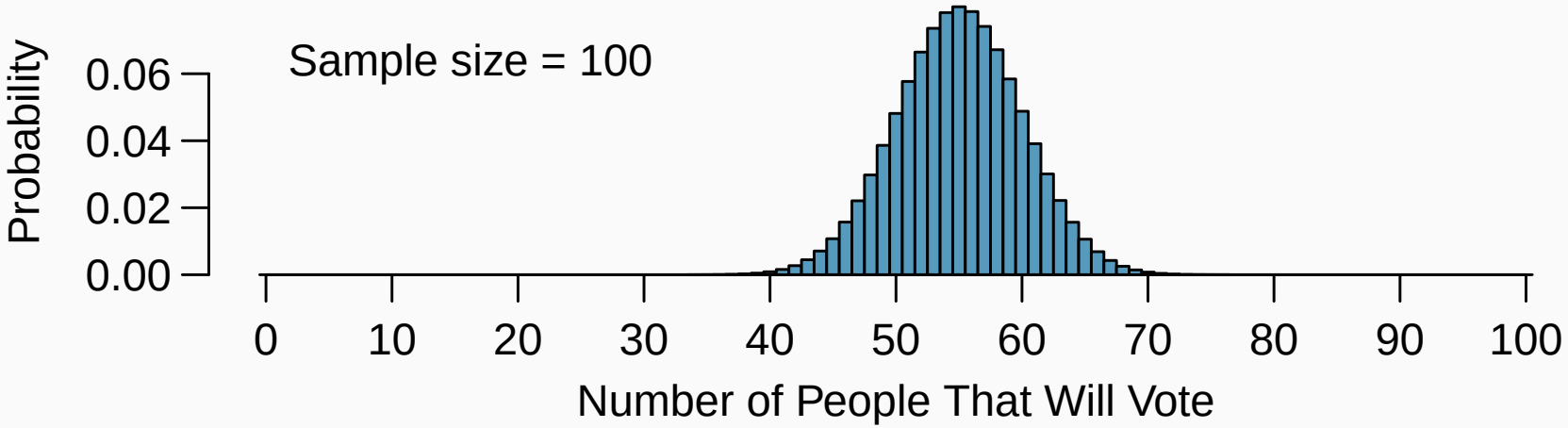
Note: Mean and SD of a binomial might not always be whole numbers, and that is alright, these values represent what we would expect to see on average.

Unusual Observations

Using the notion that *observations that are more than 2 SDs away from the mean are considered unusual* and the mean and the standard deviation we just computed, we can calculate a range for the possible number of subjects that will vote 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 the 2016 presidential election in the Chicago area is 55%. Is it unusual to obtain a sample of size 1000 that only 500 turn out?

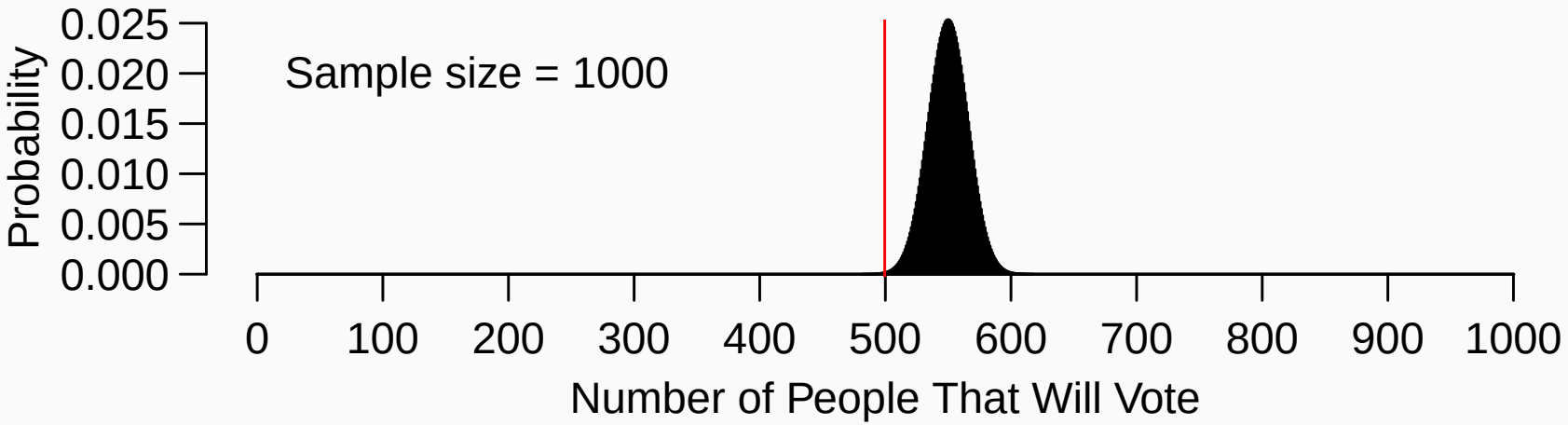
$$\mu = np = 1,000 \times 0.55 = 550$$

$$\sigma = \sqrt{np(1-p)} = \sqrt{1,000 \times 0.55 \times 0.45} \approx 15.73$$

$Z - score$ of observation $Z = \frac{x - expected\ value}{SD} = \frac{500-550}{15.73} = -3.18$

500 is more than 3 SDs below the expected value.

If the true turnout rate is 55%, it'll be unusual 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 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

number of heads obtained – expected number of heads $n \times p$

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.

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

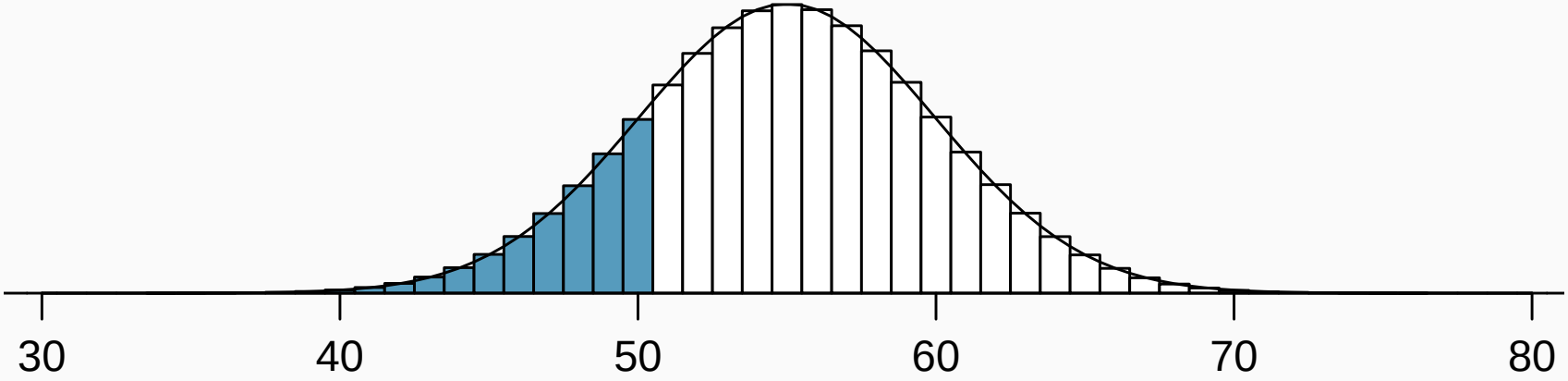
Suppose that the turnout rate for the 2016 presidential election in the Chicago area is 55%. A random sample of 100 eligible voters is taken from the Chicago area. What is the probability that at most 50 of them will vote?

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

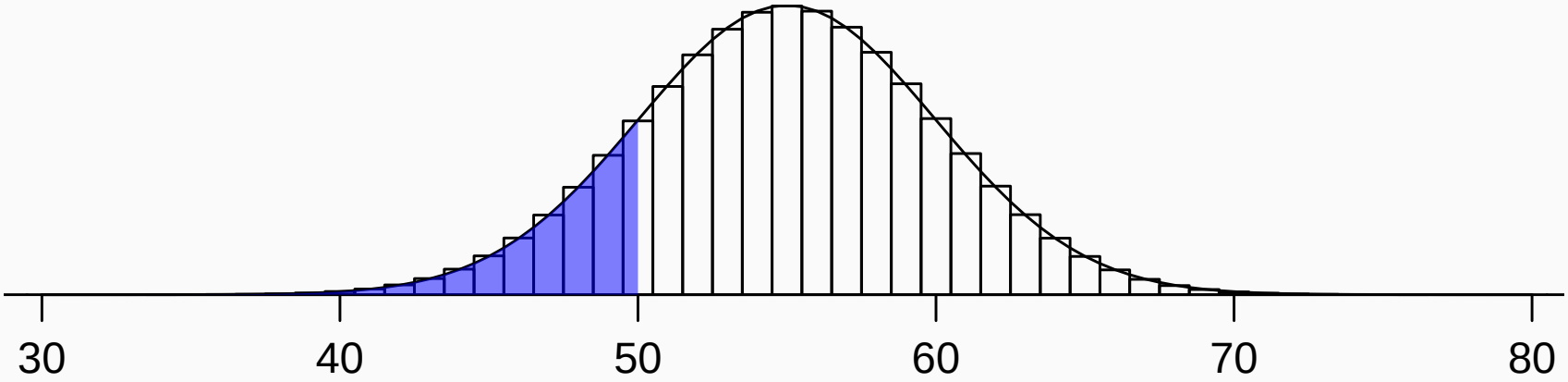
$$\begin{aligned} P(X \leq 50) &= P(X = 0) + P(X = 1) + P(X = 2) + \cdots + P(X = 50) \\ &= \binom{100}{0} \times 0.55^0 \times 0.45^{100} + \binom{100}{1} \times 0.55^1 \times 0.45^{99} \\ &\quad + \binom{100}{2} \times 0.55^2 \times 0.45^{98} + \ldots + \binom{100}{50} \times 0.55^{50} \times 0.45^{50}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

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 the 2016 presidential election in the Chicago area is 55%. A random sample of 100 eligible voters is taken from the Chicago area. 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 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) $n = 150, p = 0.05$

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

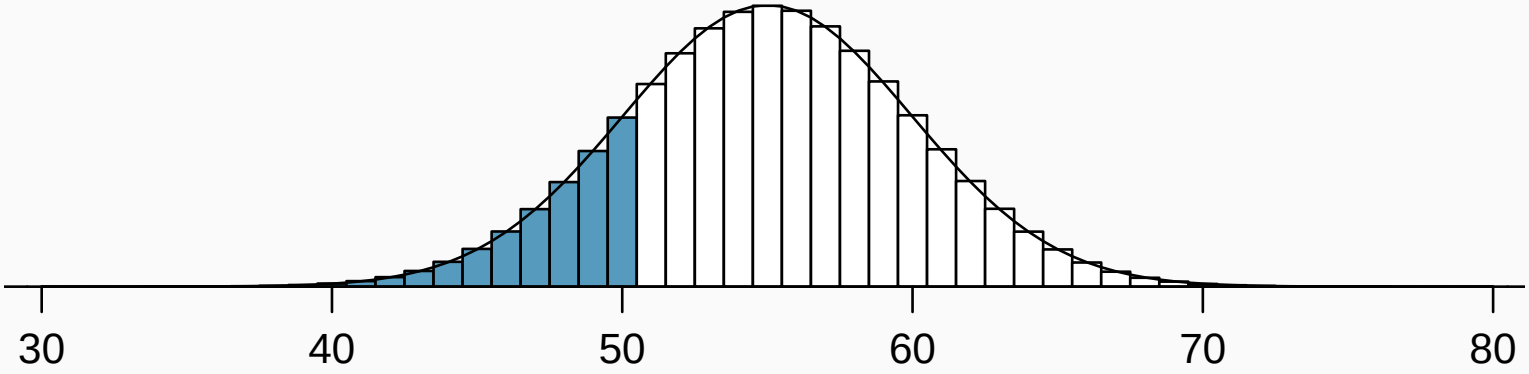
(d) $n = 500, p = 0.015$

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

Continuity Correction of the Normal Approximation to Binomial

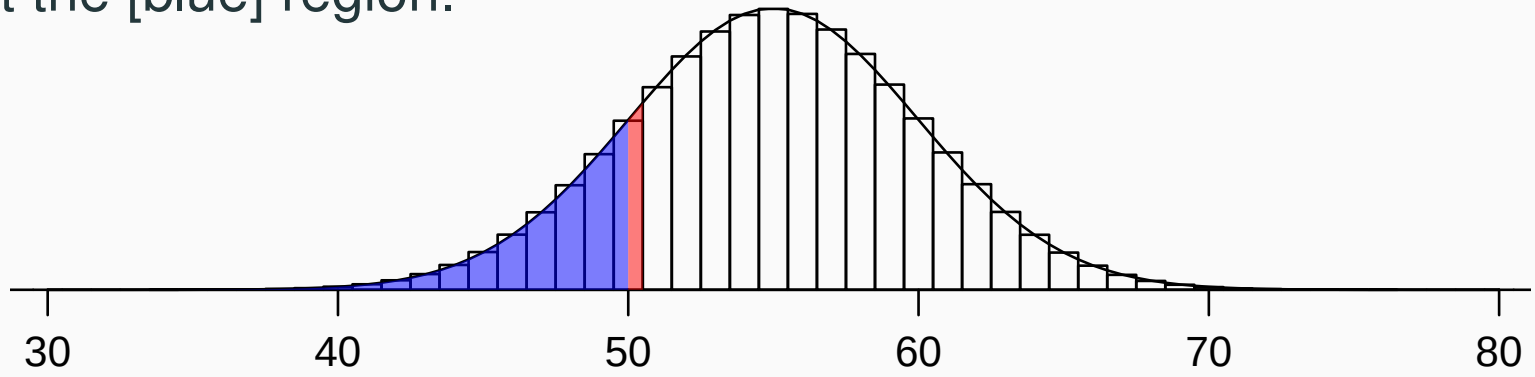
Continuity Correction of the Normal Approx. to Binomial

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 the 2016 presidential election in the Chicago area is 55%. A random sample of 100 eligible voters is taken from the Chicago area. 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 50.5 is $(50.5 - 55)/4.97 \approx -0.90$. By normal approximation, the probability is about

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

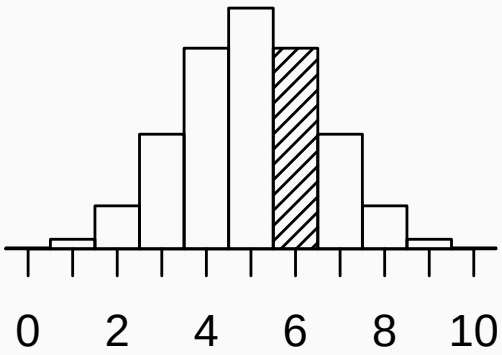
which is closer to exact probability 0.1827

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

How to Adjust the Endpoints?

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

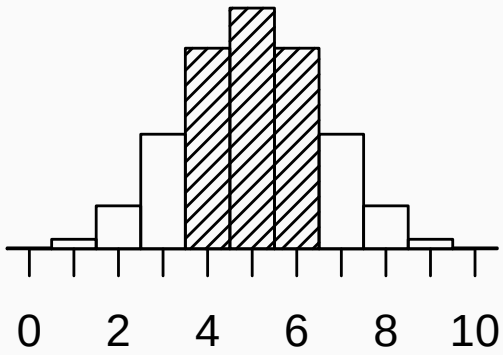
$P(X = 6)$



$P(5.5 < X < 6.5)$

3 to 7 exclusive

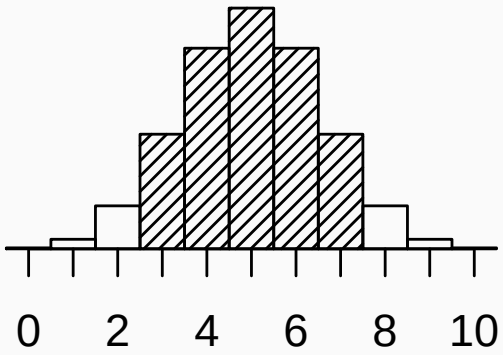
$P(3 < X < 7)$



$P(3.5 < X < 6.5)$

3 to 7 inclusive

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



$P(2.5 < X < 7.5)$

	$P(X = 6)$	$P(3 < X < 7)$	$P(3 \leq X \leq 7)$
binomial formula	0.2051	0.6563	0.8906
normal w/o correction	0	0.7941	0.7941
normal w/ correction	0.2045	0.6572	0.8862

Continuity Correction of the Normal Approx. to Binomial

- Continuity correction can improve the accuracy of normal approximation for binomial $Bin(n, p)$ when n is relatively small, but the improvement is negligible when n is large
- Continuity correction may be used for finding probability over a *small interval*, even when n is large E.g., for $X \sim Bin(1000, 0.5)$, the probability of $520 \leq X \leq 525$ is
  - 0.05535 using binomial formula
  - 0.04603 using normal approximation w/o correction
  - 0.05533 using normal approximation with correction, i.e.,

$$P(520 \leq X \leq 525) = P(519.5 \leq X \leq 525.5).$$