STA258H5

Al Nosedal
and Alison Weir

Winter 2017

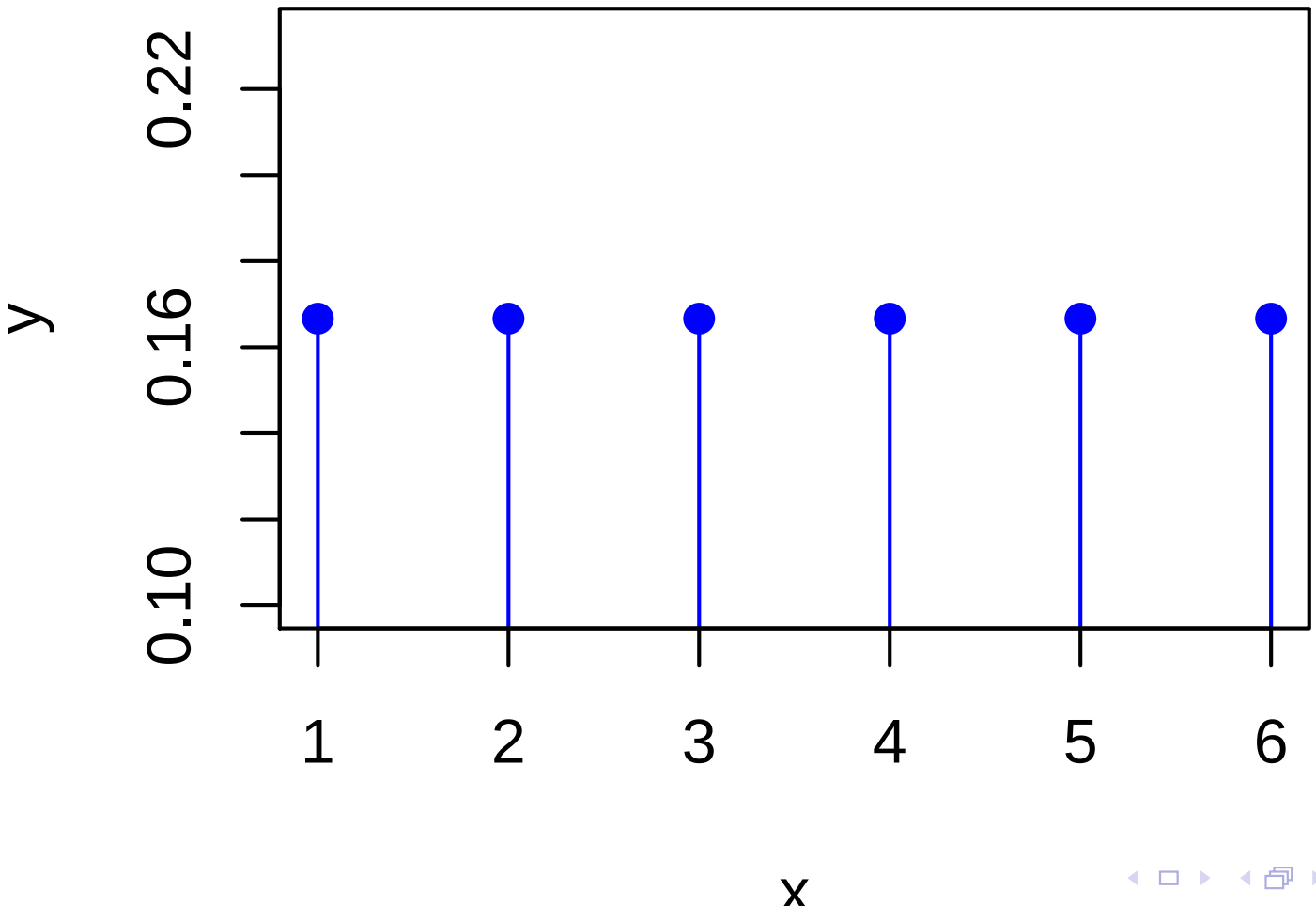
NORMAL APPROXIMATION TO THE BINOMIAL DISTRIBUTION.

Discrete Uniform Distribution

A random variable X has a **discrete uniform distribution** if each of the n values in its range, say, $x_1, x_2, \ldots, x_n$ has equal probability. Then,

$$f(x_i) = \frac{1}{n}$$

Probability mass function (pmf)



Random Variable

A **random variable** is a variable whose value is a numerical outcome of a random phenomenon.
The **probability distribution** of a random variable X tells us what values X can take and how to assign probabilities to those values.

The Binomial setting

- There are a fixed number n of observations.
- The n observations are all **independent**. That is, knowing the result of one observation tells you nothing about the other obsevations.
- Each observation falls into one of just two categories, which for convenience we call "success" and "failure".
- The probability of a success, call it p, is the same for each observation.

Example

Think of rolling a die n times as an example of the binomial setting. Each roll gives either a six or a number different from six. Knowing the outcome of one roll doesn't tell us anything about other rolls, so the n rolls are independent. If we call six a success, then p is the probability of a six and remains the same as long as we roll the same die. The number of sixes we count is a random variable X. The distribution of X is called a **binomial distribution**.

Binomial Distribution

A random variable Y is said to have a **binomial distribution** based on n trials with success probability p if and only if

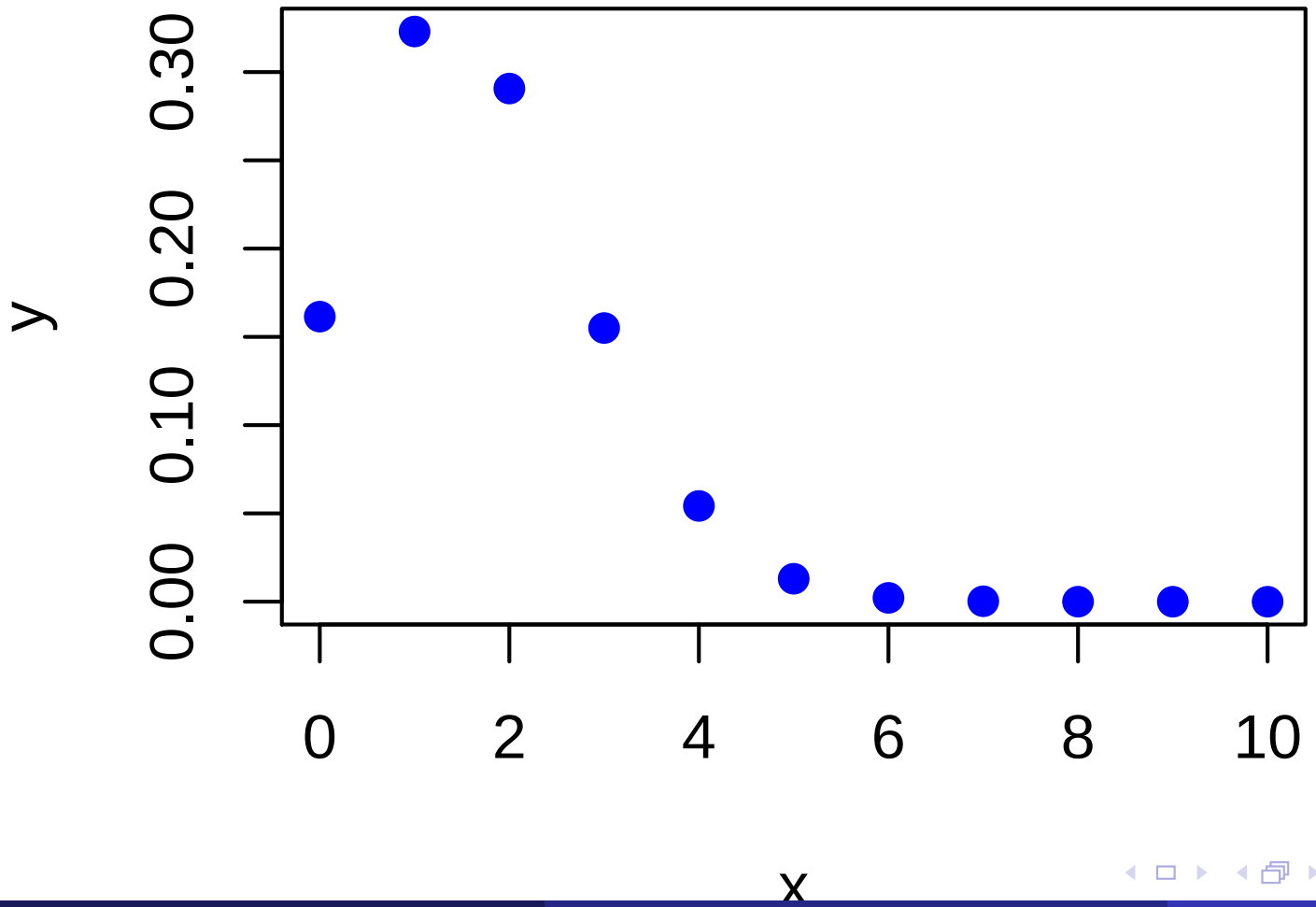
$$p(y) = \frac{n!}{y!(n-y)!} p^y (1-p)^{n-y}, \ y = 0, 1, 2, ..., n \text{ and } 0 \leq p \leq 1.$$

$E(Y) = np$ and $V(Y) = np(1-p)$.

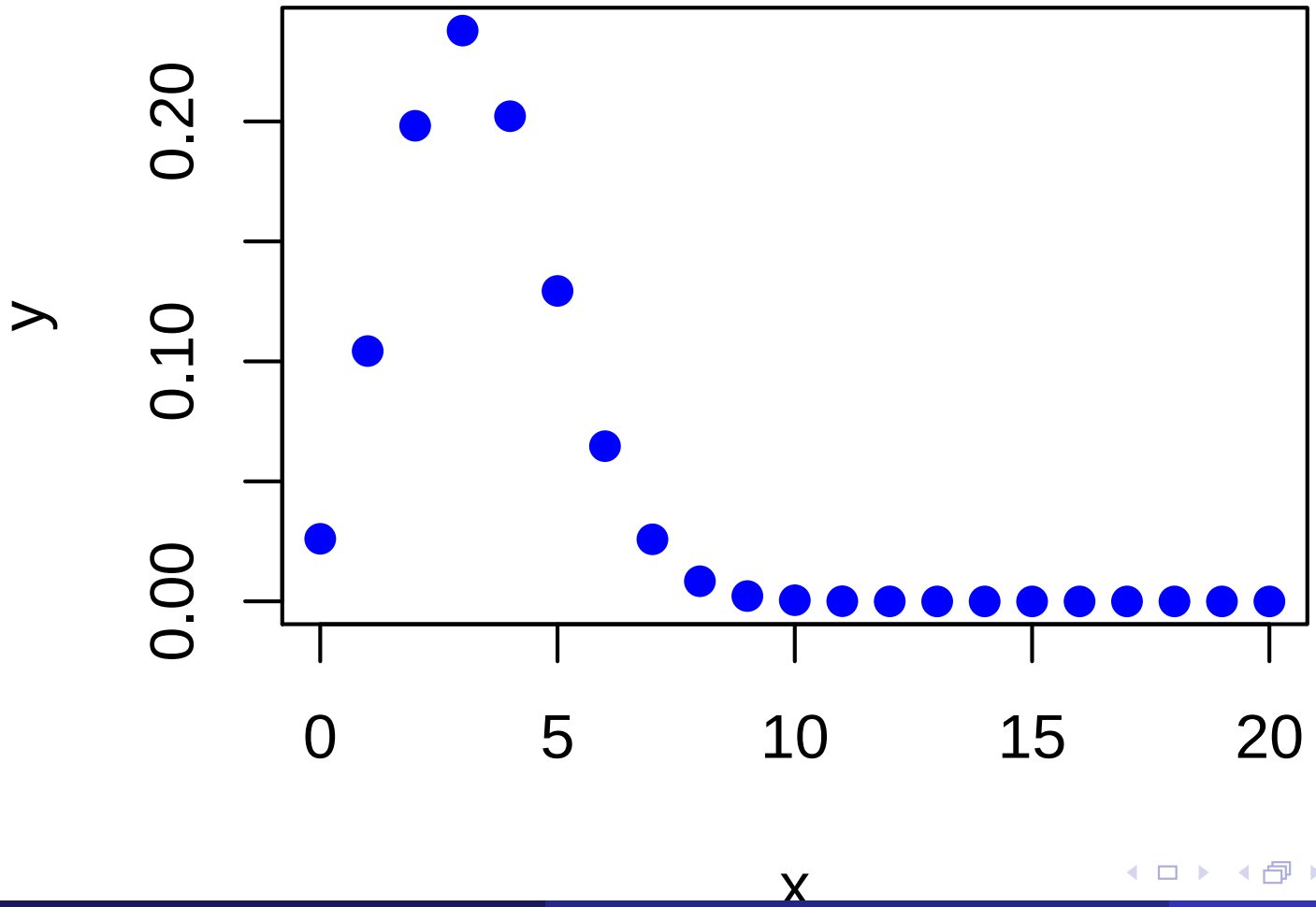
Probability mass function when n=10 and p=1/6

```
## Pmf of Binomial with n=10 and p=1/6.

x<-seq(0,10,by=1);

y<-dbinom(x,10,1/6);

plot(x,y,type="p",col="blue",pch=19);
```

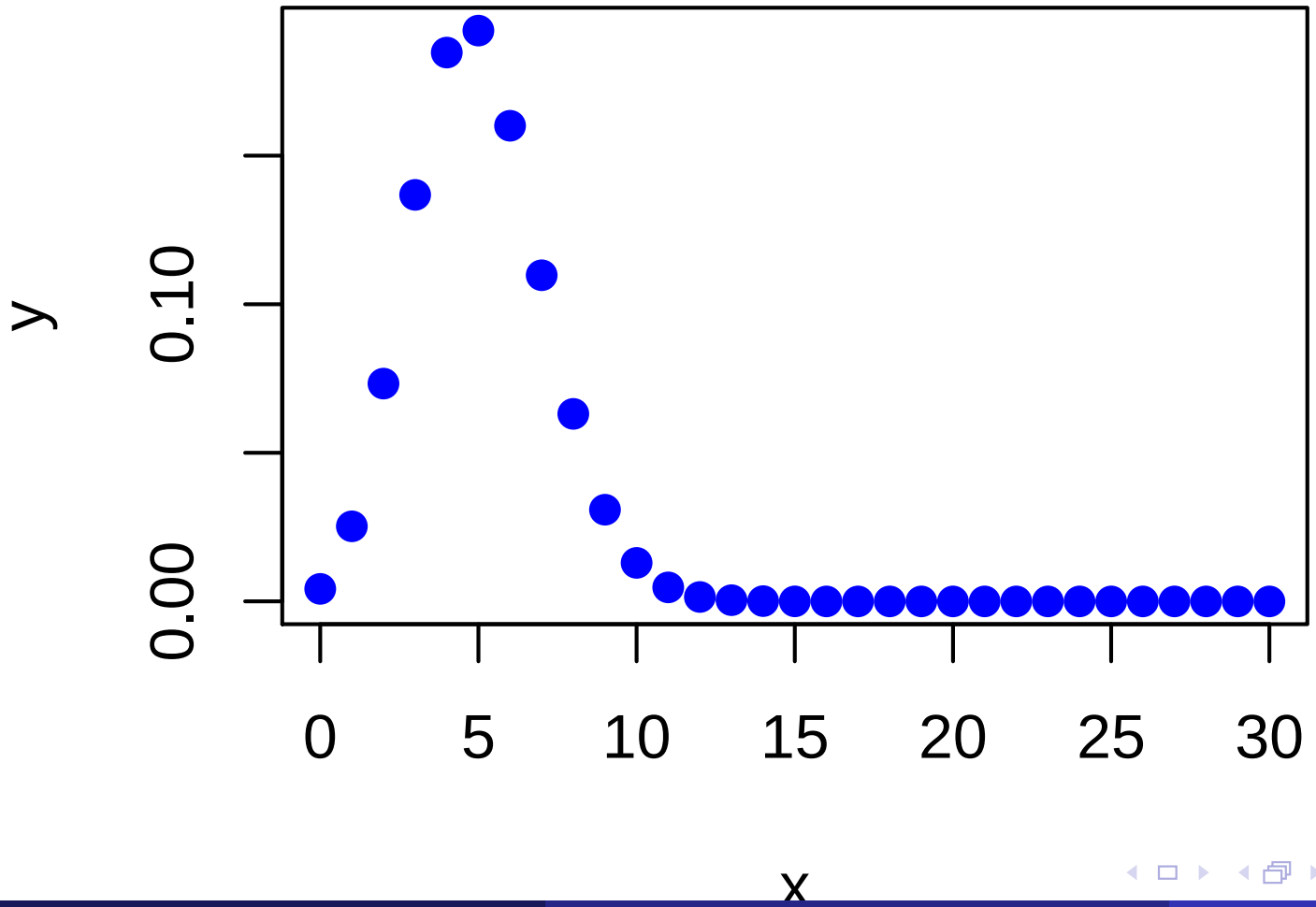
PMF when n=10 and p=1/6



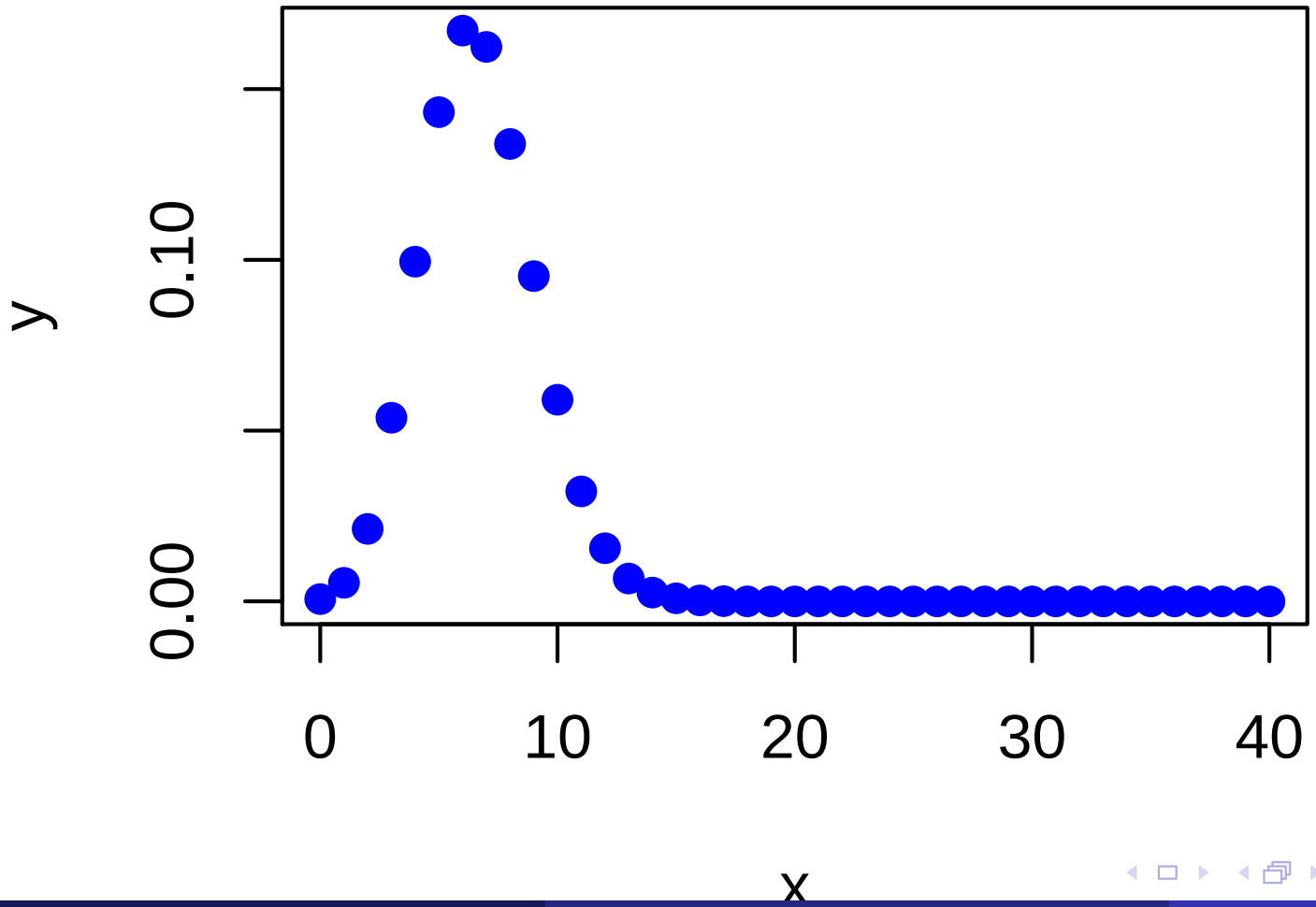
PMF when n=20 and p=1/6



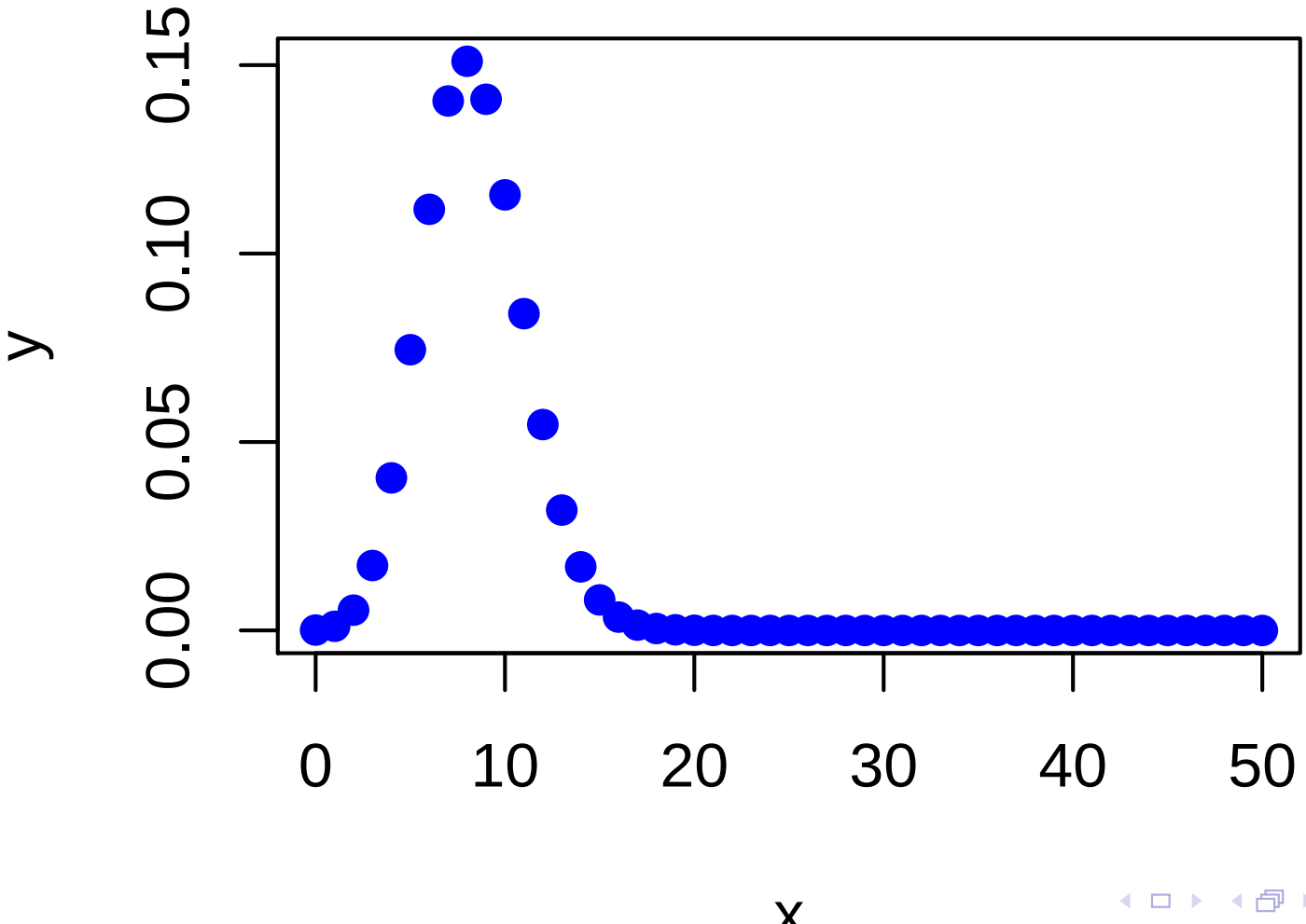
PMF when n=30 and p=1/6



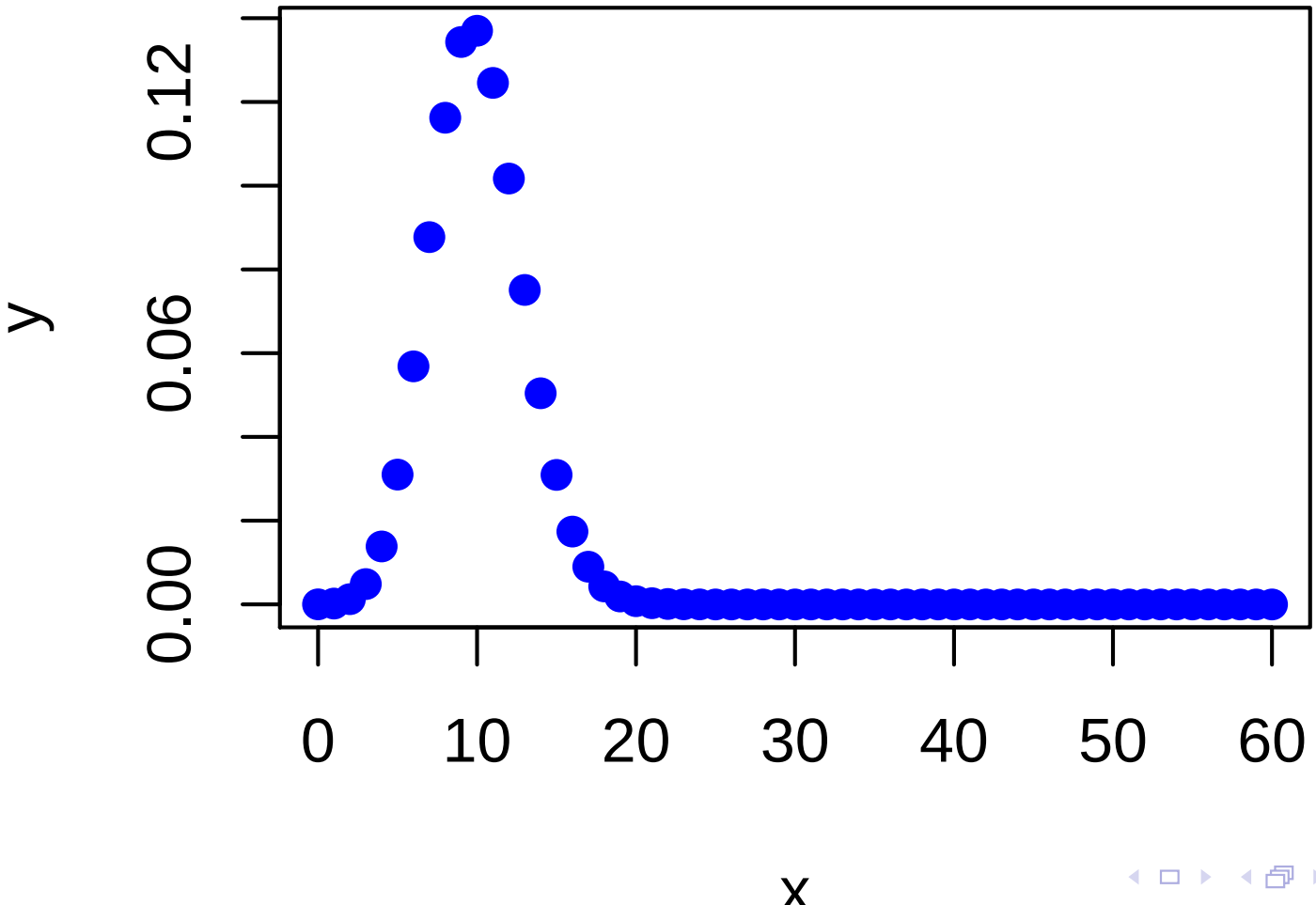
PMF when n=40 and p=1/6



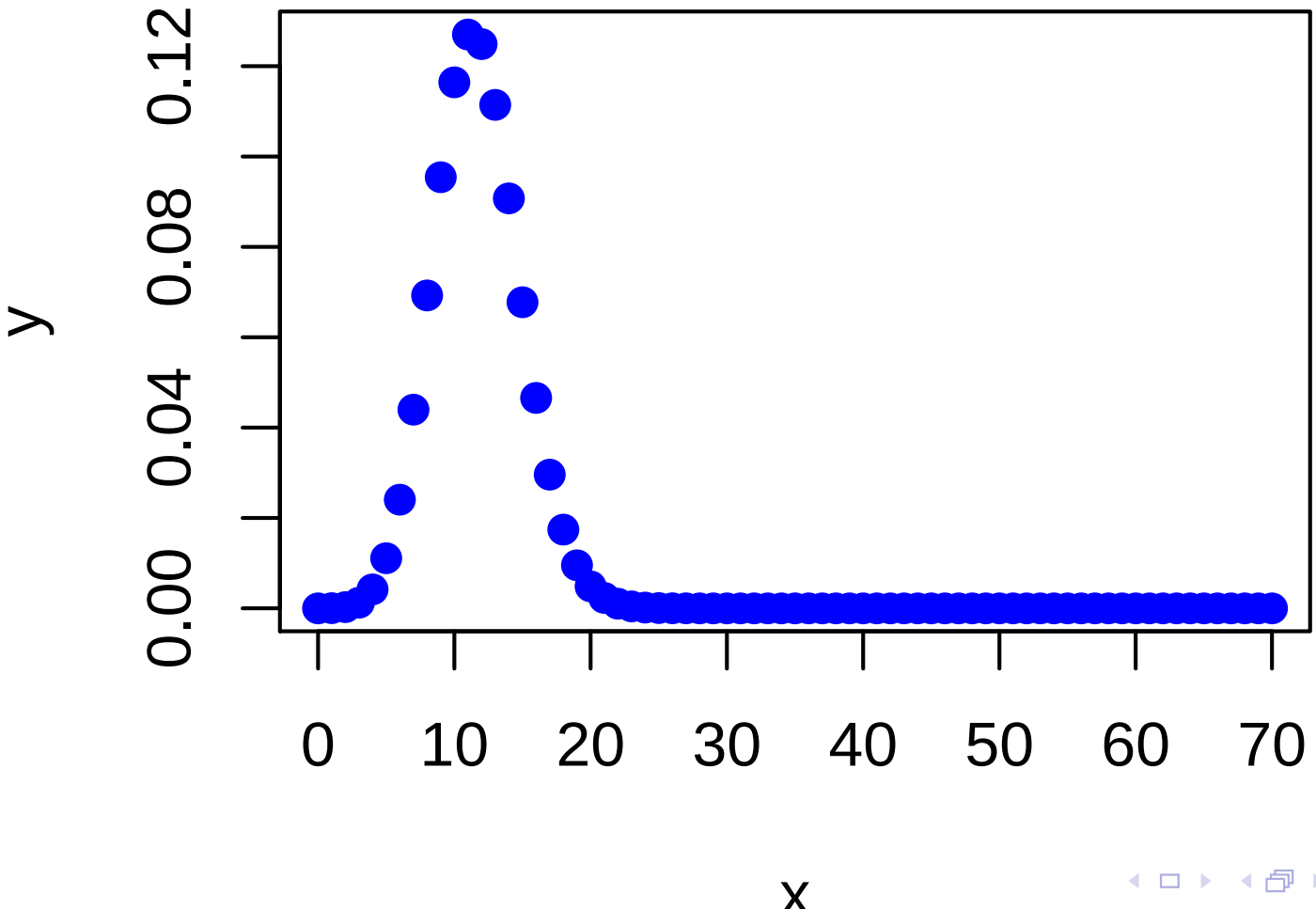
PMF when n=50 and p=1/6



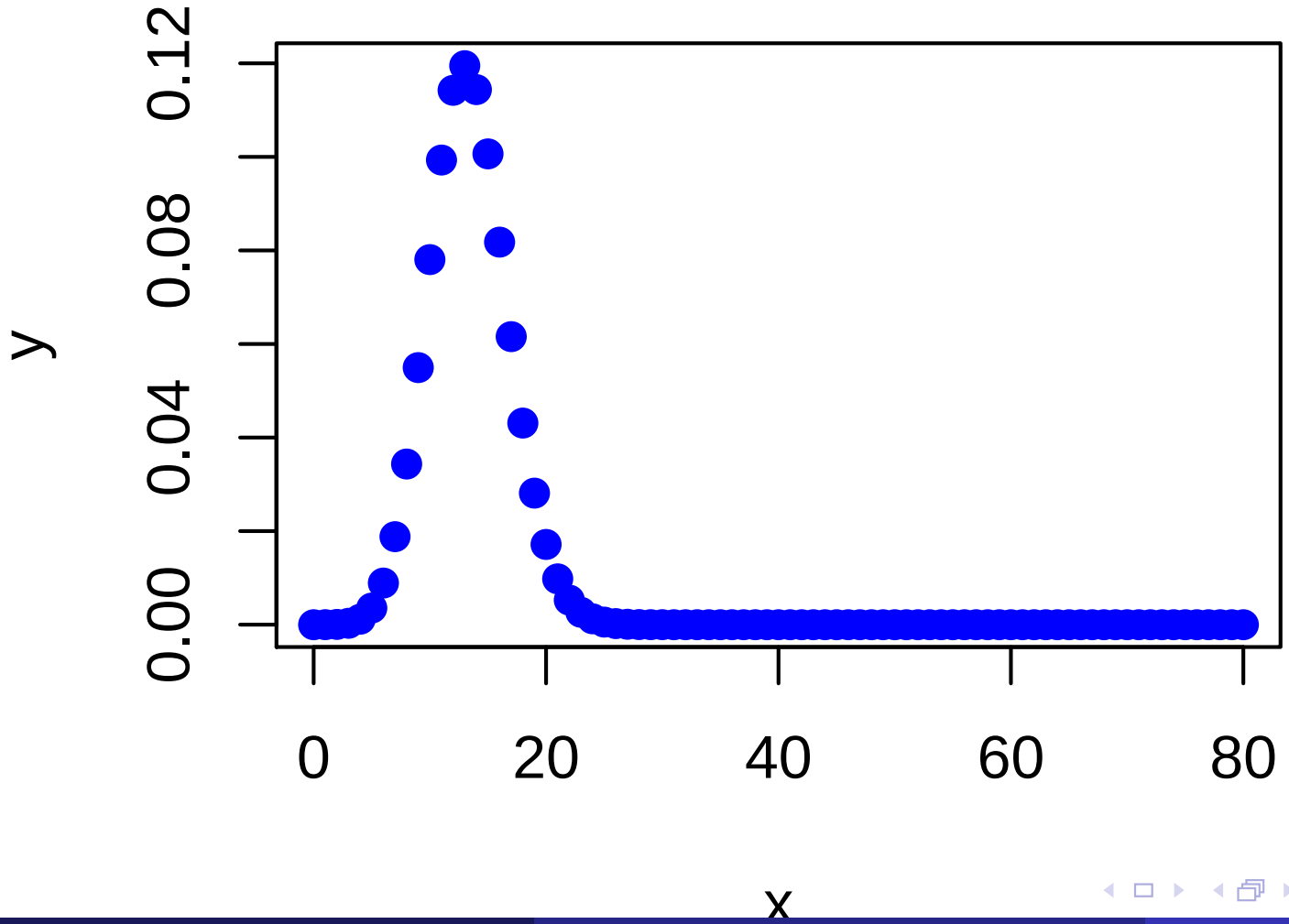
PMF when n=60 and p=1/6



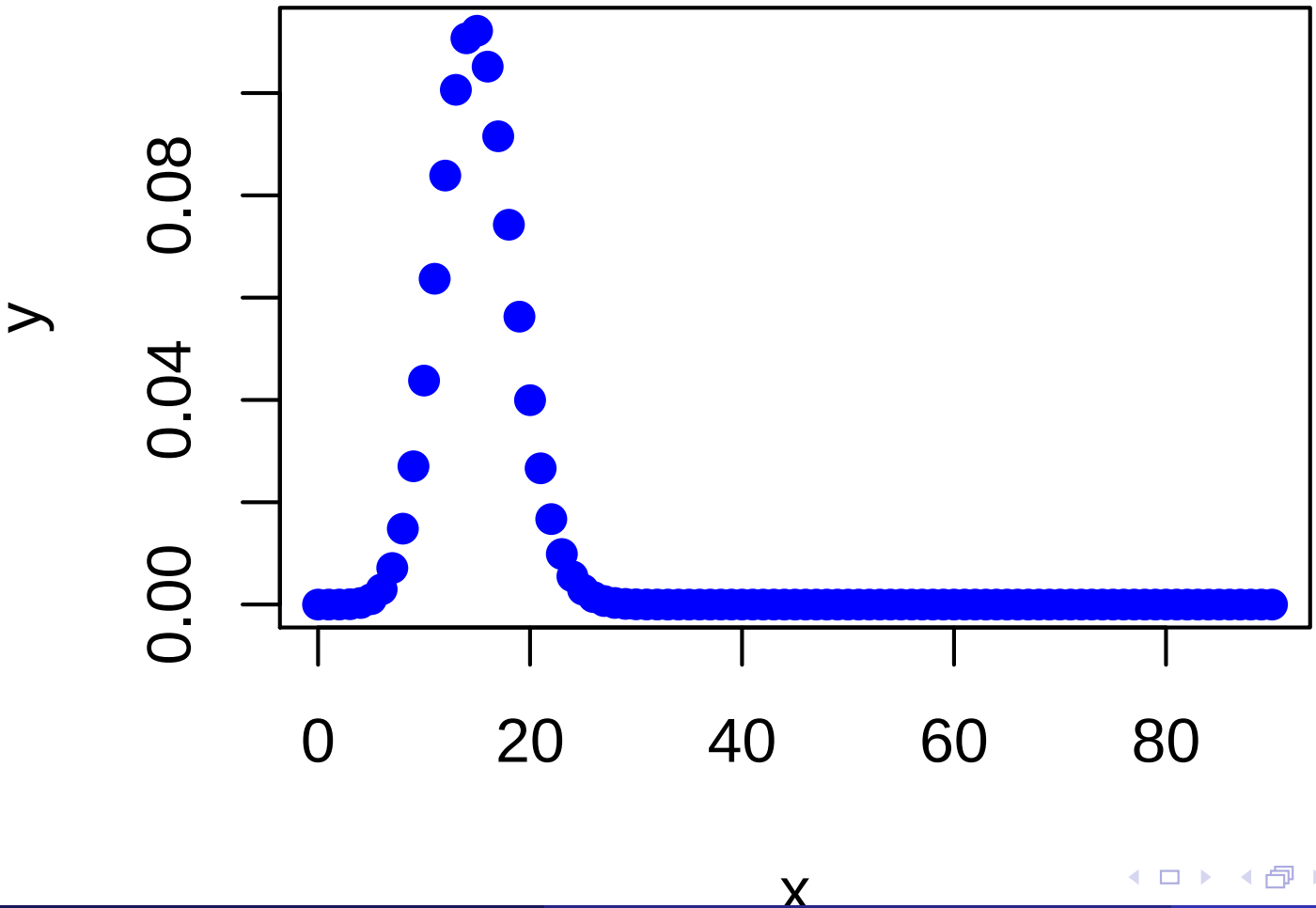
PMF when n=70 and p=1/6



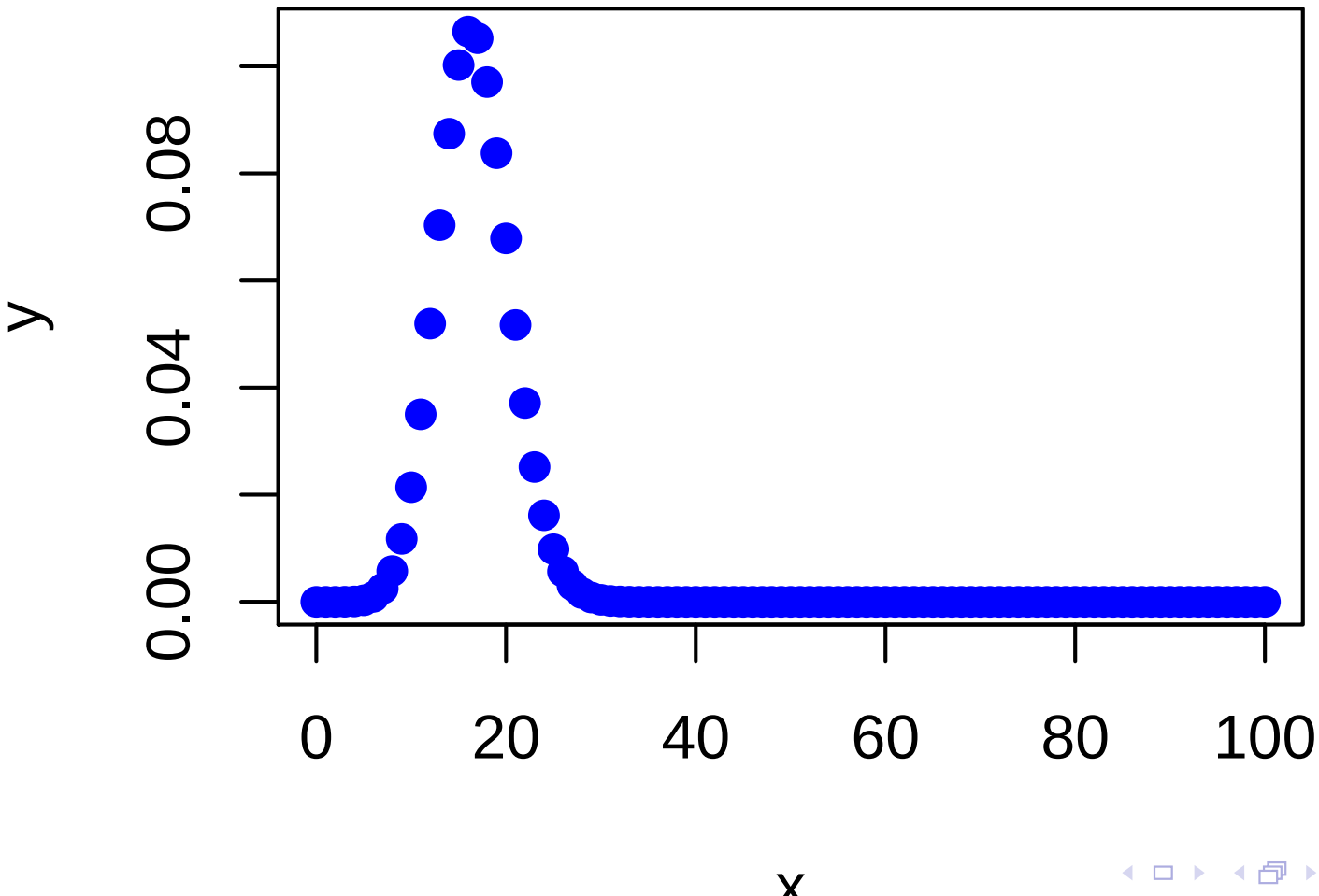
PMF when n=80 and p=1/6



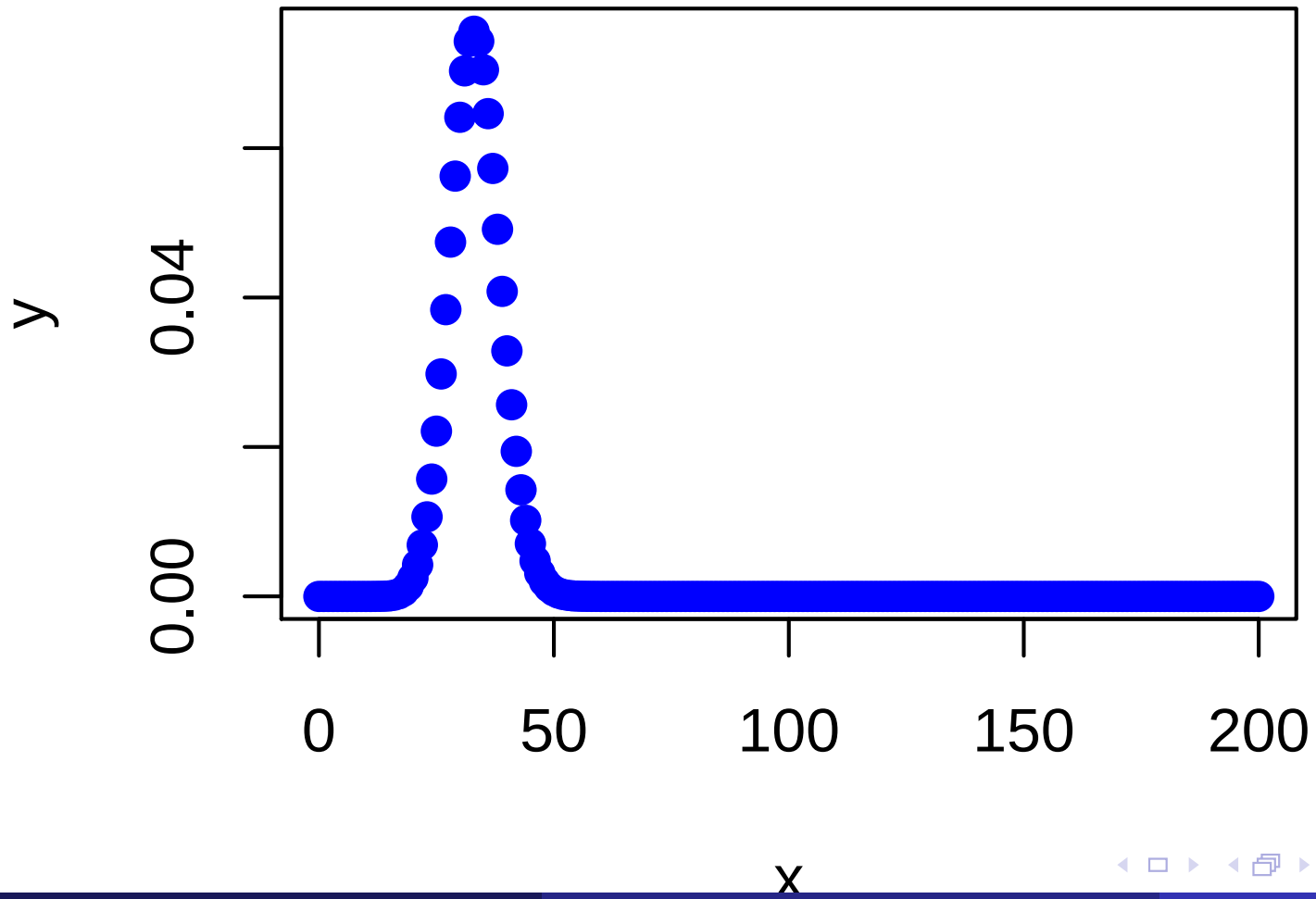
PMF when n=90 and p=1/6



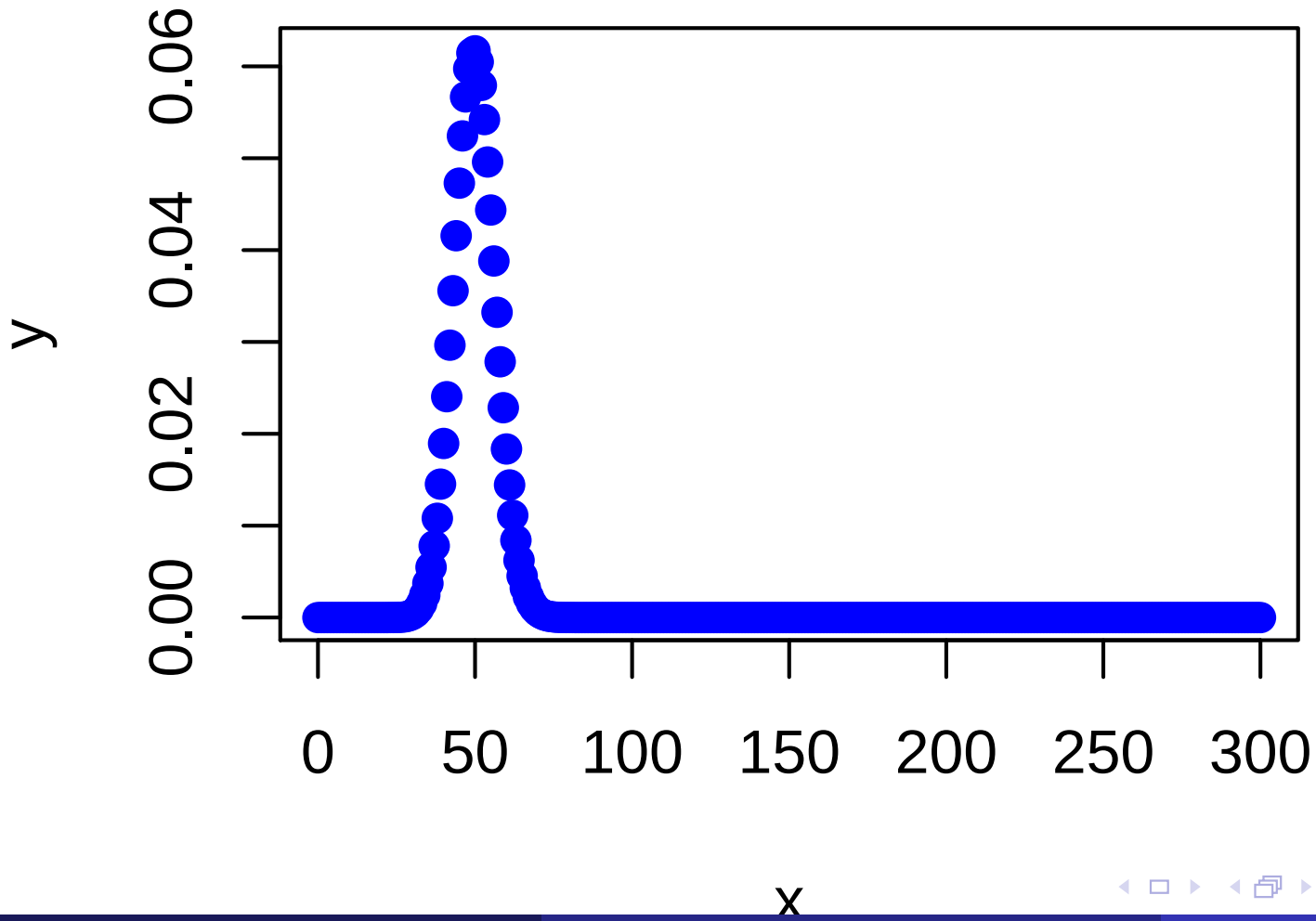
PMF when n=100 and p=1/6



PMF when n=200 and p=1/6



PMF when n=300 and p=1/6



Sampling Distribution of a sample proportion

Draw an Simple Random Sample (SRS) of size n from a large population that contains proportion p of "successes". Let $\hat{p}$ be the **sample proportion** of successes,

$$\hat{p} = \frac{\text{number of successes in the sample}}{n}$$

Then:

- The **mean** of the sampling distribution of $\hat{p}$ is p.
- The **standard deviation** of the sampling distribution is

$$\sqrt{\frac{p(1-p)}{n}}.$$

Sampling Distribution of a sample proportion

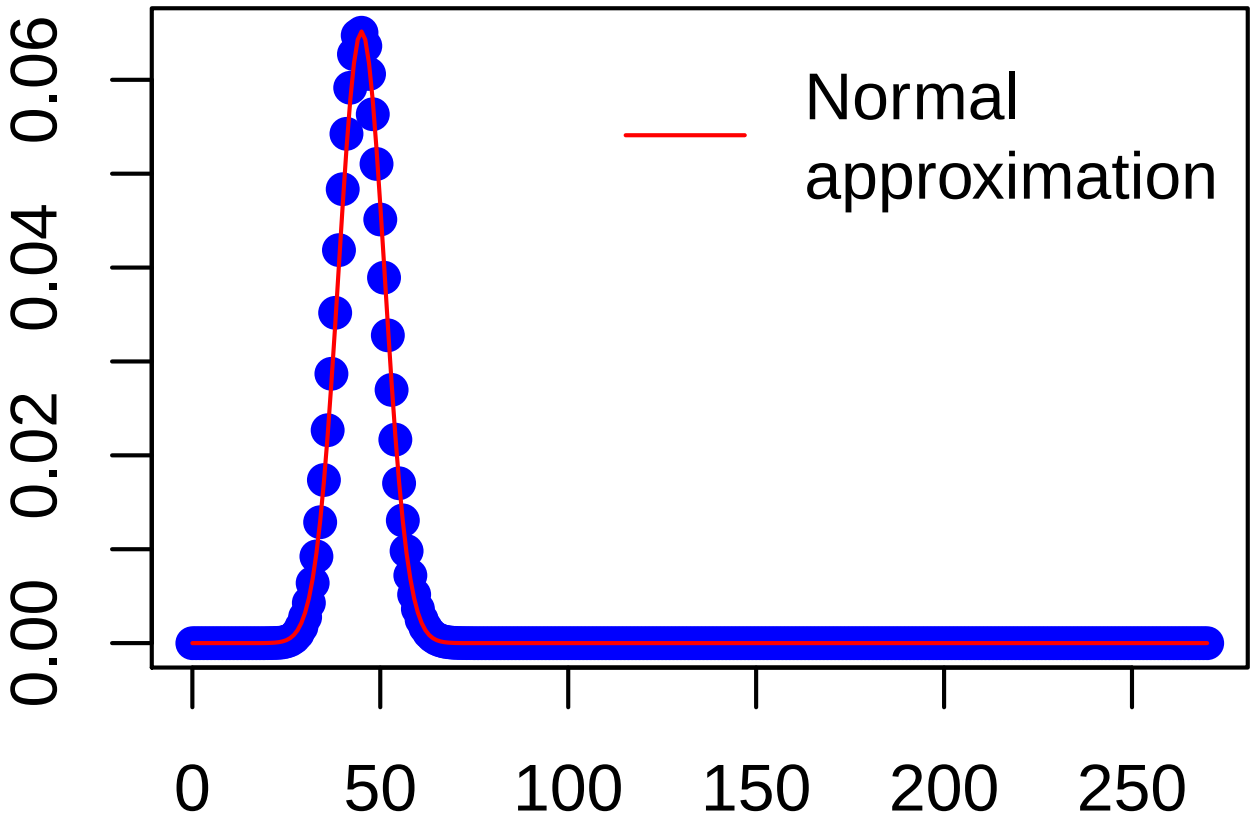
Draw an SRS of size n from a large population that contains proportion p of "successes". Let $\hat{p}$ be the **sample proportion** of successes,

$$\hat{p} = \frac{\text{number of successes in the sample}}{n}$$

Then:

- As the sample size increases, the sampling distribution of $\hat{p}$ becomes **approximately Normal.** That is, for large n, $\hat{p}$ has approximately the $N\left(p, \sqrt{\frac{p(1-p)}{n}}\right)$ distribution.

Binomial with Normal Approximation



Bernoulli Distribution (Binomial with $n = 1$)

$$x_i = \begin{cases} 1 & \text{i-th roll is a six} \\ 0 & \text{otherwise} \end{cases}$$

$\mu = E(x_i) = p$
$\sigma^2 = V(x_i) = p(1-p)$
Let $\hat{p}$ be our estimate of p. Note that $\hat{p} = \frac{\sum_{i=1}^{n} x_i}{n} = \bar{x}$. If n is "large", by the Central Limit Theorem, we know that:
$\bar{x}$ is roughly $N(\mu, \frac{\sigma}{\sqrt{n}})$, that is,
$\hat{p}$ is roughly $N\left(p, \sqrt{\frac{p(1-p)}{n}}\right)$

Example

In the last election, a state representative received 52% of the votes cast. One year after the election, the representative organized a survey that asked a random sample of 300 people whether they would vote for him in the next election. If we assume that his popularity has not changed, what is the probability that more than half of the sample would vote for him?

Solution (Normal approximation)

We want to determine the probability that the sample proportion is greater than 50%. In other words, we want to find $P(\hat{p} > 0.50)$.
We know that the sample proportion $\hat{p}$ is roughly Normally distributed with mean $p = 0.52$ and standard deviation $\sqrt{p(1-p)/n} = \sqrt{(0.52)(0.48)/300} = 0.0288$.
Thus, we calculate

$$\begin{aligned} P(\hat{p} > 0.50) &= P\left(\frac{\hat{p}-p}{\sqrt{p(1-p)/n}} > \frac{0.50-0.52}{0.0288}\right) \\ &= P(Z > -0.69) = 1 - P(Z < -0.69) \quad (Z \text{ is symmetric}) \\ &= P(Z > -0.69) = 1 - P(Z > 0.69) \\ &= 1 - 0.2451 = 0.7549. \end{aligned}$$

If we assume that the level of support remains at 52%, the probability that more than half the sample of 300 people would vote for the representative is 0.7549.

R code (Normal approximation)

Just type in the following:

```
1- pnorm(0.50, mean = 0.52, sd = 0.0288);

## [1] 0.7562982
```

Recall that, `pnorm` will give you the area to the left of 0.50, for a Normal distribution with mean 0.52 and standard deviation 0.0288.

Solution (using Binomial)

We want to determine the probability that the sample proportion is greater than 50%. In other words, we want to find $P(\hat{p} > 0.50)$. We know that $n = 300$ and $p = 0.52$.

Thus, we calculate

$$P(\hat{p} > 0.50) = P\left(\frac{\sum_{i=1}^{n} x_i}{n} > 0.50\right)$$
$$= P(\sum_{i=1}^{300} x_i > 150)$$
$$= 1 - P(\sum_{i=1}^{300} x_i \leq 150)$$

(it can be shown that $Y = \sum_{i=1}^{300} x_i$ has a Binomial distribution with $n = 300$ and $p = 0.52$).

$$= 1 - F_Y(150)$$

R code (using Binomial distribution)

Just type in the following:

```
1- pbinom(150, size = 300, prob=0.52);

## [1] 0.7375949
```

Recall that, `pbinom` will give you the CDF at 150, for a Binomial distribution with $n = 300$ and $p = 0.52$.

Solution (using continuity correction)

We have that $n = 300$ and $p = 0.52$.
Thus, we calculate

$$P(\hat{p} > 0.50) = P\left(\frac{\sum_{i=1}^{n} x_i}{n} > 0.50\right)$$
$$= P(\sum_{i=1}^{300} x_i > 150)$$
$$= 1 - P(\sum_{i=1}^{300} x_i \leq 150)$$

(it can be shown that $Y = \sum_{i=1}^{300} x_i$ has a Binomial distribution with $n = 300$ and $p = 0.50$).

$$\approx 1 - P(\sum_{i=1}^{300} x_i \leq 150.5) \quad \text{(continuity correction)}$$
$$= 1 - P(\frac{\sum_{i=1}^{300} x_i}{n} \leq \frac{150.5}{300})$$
$$= 1 - P(\hat{p} \leq 0.5017)$$
$$= 1 - P(Z \leq -0.6354) \quad \text{(Why?)}$$

R code (Normal approximation with continuity correction)

Just type in the following:

```
1- pnorm(0.5017, mean = 0.52, sd = 0.0288);

## [1] 0.7374216
```

Recall that, `pnorm` will give you the area to the left of 0.5017, for a Normal distribution with mean 0.52 and standard deviation 0.0288.

Continuity Correction

Suppose that Y has a Binomial distribution with $n = 20$ and $p = 0.4$. We will find the exact probabilities that $Y \leq y$ and compare these to the corresponding values found by using two Normal approximations. One of them, when X is Normally distributed with $\mu_X = np$ and $\sigma_X = \sqrt{np(1-p)}$. The other one, W, a shifted version of X.

Continuity Correction (cont.)

For example,

$$P(Y \leq 8) = 0.5955987$$

As previously stated, we can think of Y as having approximately the same distribution as X.

$$P(Y \leq 8) \approx P(X \leq 8)$$

$$= P\left[\frac{X-np}{\sqrt{np(1-p)}} \leq \frac{8-8}{\sqrt{20(0.4)(0.6)}}\right]$$

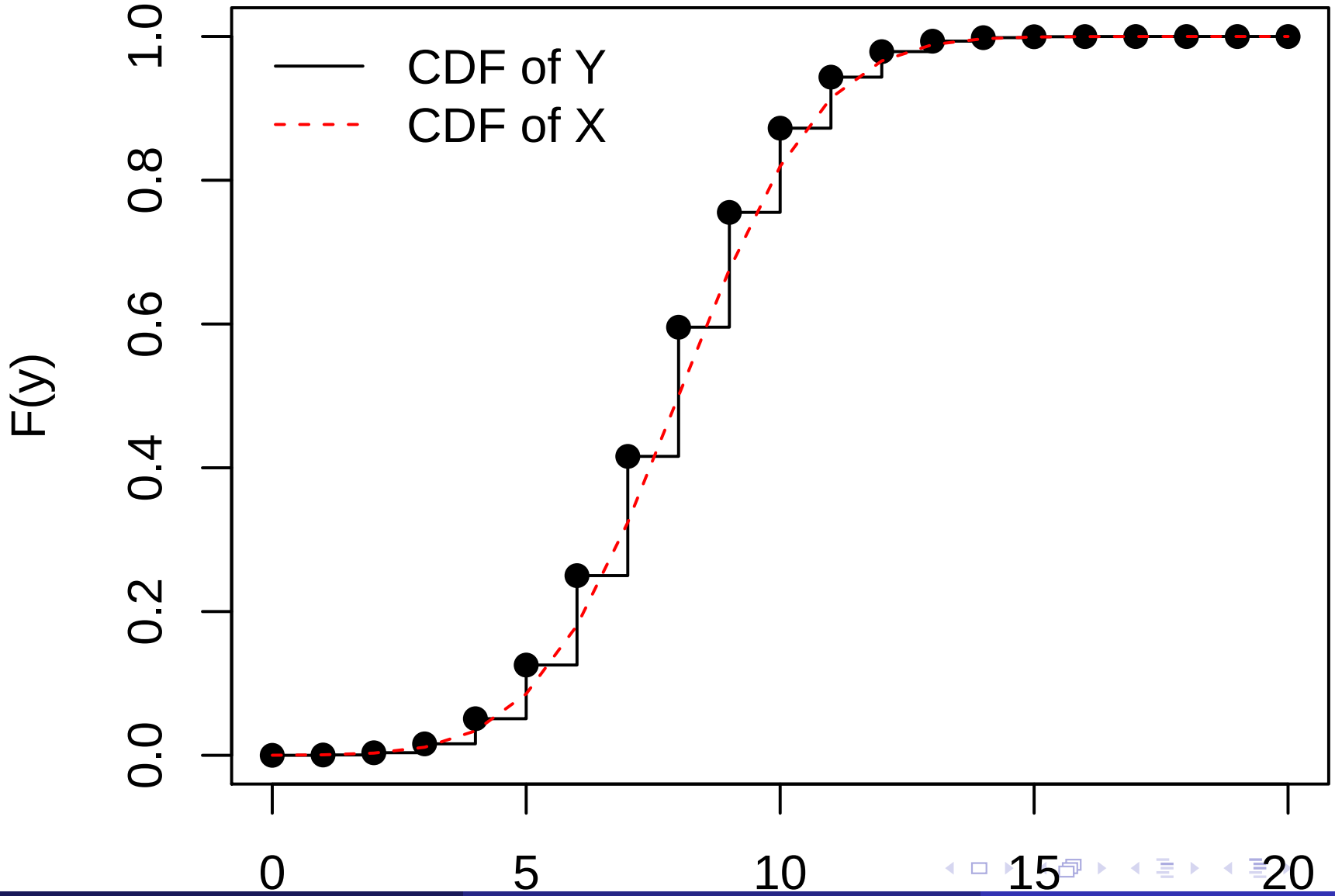
$$= P(Z \leq 0) = 0.5$$

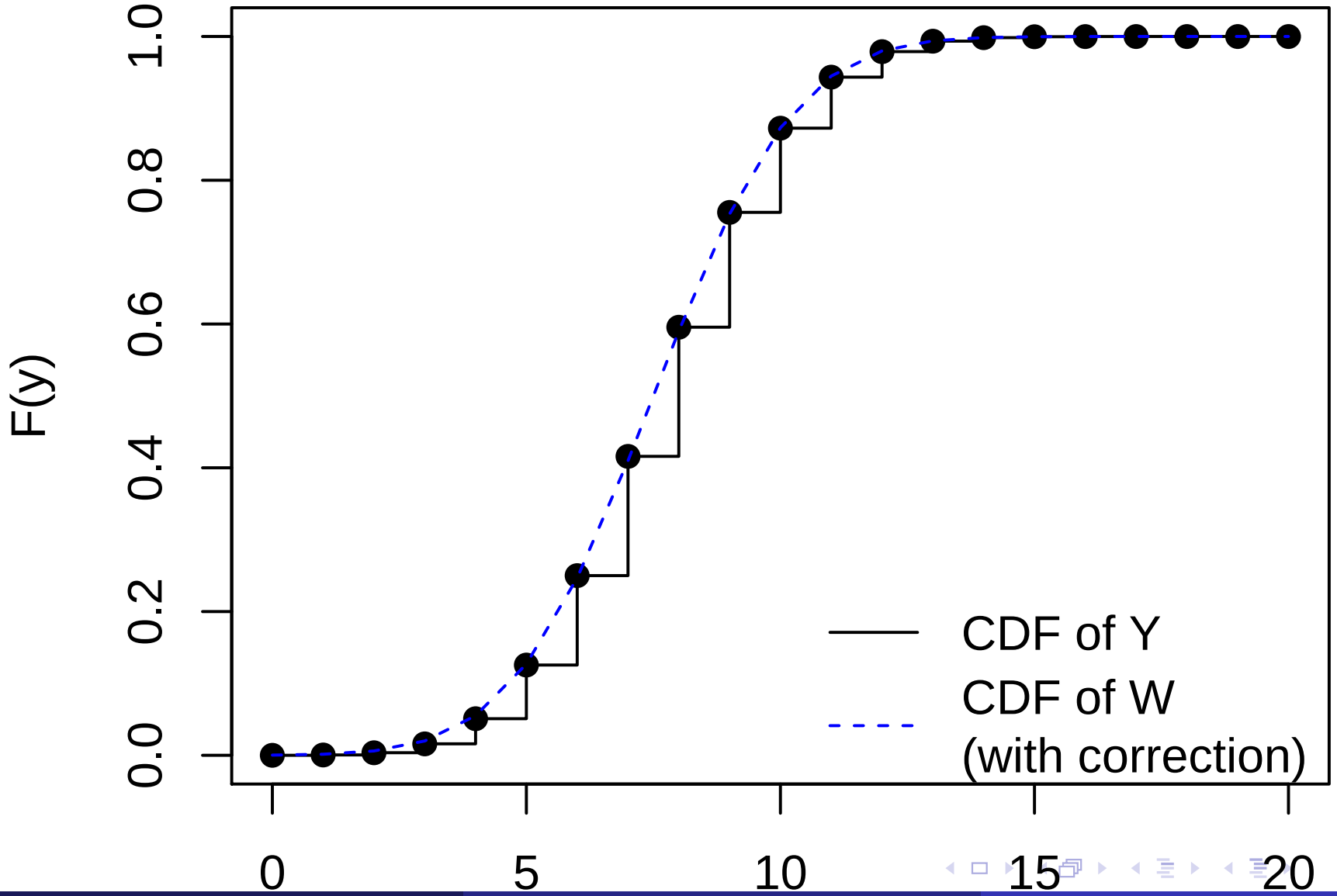
Continuity Correction (cont.)

$$P(Y \leq 8) \approx P(W \leq 8.5)$$

$$= P\left[\frac{W-np}{\sqrt{np(1-p)}} \leq \frac{8.5-8}{\sqrt{20(0.4)(0.6)}}\right]$$

$$= P(Z \leq 0.2282) = 0.5902615$$


CDF of Y
CDF of X
F(y)
0.0
0.2
0.4
0.6
0.8
1.0
0
5
10
15
20


F(y)
0.0
0.2
0.4
0.6
0.8
1.0
0
5
10
15
20
CDF of Y
CDF of W
(with correction)

Example

Fifty-one percent of adults in the U. S. whose New Year's resolution was to exercise more achieved their resolution. You randomly select 65 adults in the U. S. whose resolution was to exercise more and ask each if he or she achieved that resolution. What is the probability that exactly forty of them respond yes?

Example

Fifty-one percent of adults in the U. S. whose New Year's resolution was to exercise more achieved their resolution. You randomly select 65 adults in the U. S. whose resolution was to exercise more and ask each if he or she achieved that resolution. What is the probability that fewer than forty of them respond yes?

Normal Approximation to Binomial

Let $X = \sum_{i=1}^{n} Y_i$ where $Y_1, Y_2, ..., Y_n$ are iid Bernoulli random variables. Note that $X = n\hat{p}$.

1. $n\hat{p}$ is approximately Normally distributed provided that np and $n(1-p)$ are greater than 5.
2. The expected value: $E(n\hat{p}) = np$.
3. The variance: $V(\hat{np}) = np(1-p) = npq$.

Why bother with approximating?

- Calculations may be less tedious.
- Calculations will be made easier and quicker.