STA258H5
University of Toronto Mississauga

Al Nosedal and Omid Jazi

Winter 2023

NORMAL APPROXIMATION TO THE BINOMIAL DISTRIBUTION.

Definition

A statistic is a function of the observable random variables in a sample and known constants.

Sampling Distribution

Because all statistics are functions of the random variables observed in a sample, all statistics are random variables. Consequently, all statistics have probability distributions, which we will call their **sampling distributions**.

Review From STA256

Bernoulli Distribution (Discrete)

1 trial → outcome is a success ($X=1$) — p
or a failure ($X=0$) — $1-p$

$X=x$	0	1
$P(X=x)$	$1-p$	p

PMF: $f(x) = p^x (1-p)^{1-x}$

Binomial

n trials → outcome: X successes (a success occurs with prob p)
(indep) $n-x$ failures (a failure occurs with prob $1-p$)

PMF:

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

- $\binom{n}{x}$: # combinations of x successes and $n-x$ failures
- p^x: prob of x successes
- $(1-p)^{n-x}$: prob of $n-x$ failures

Mean: $\mu = E(x) = np$

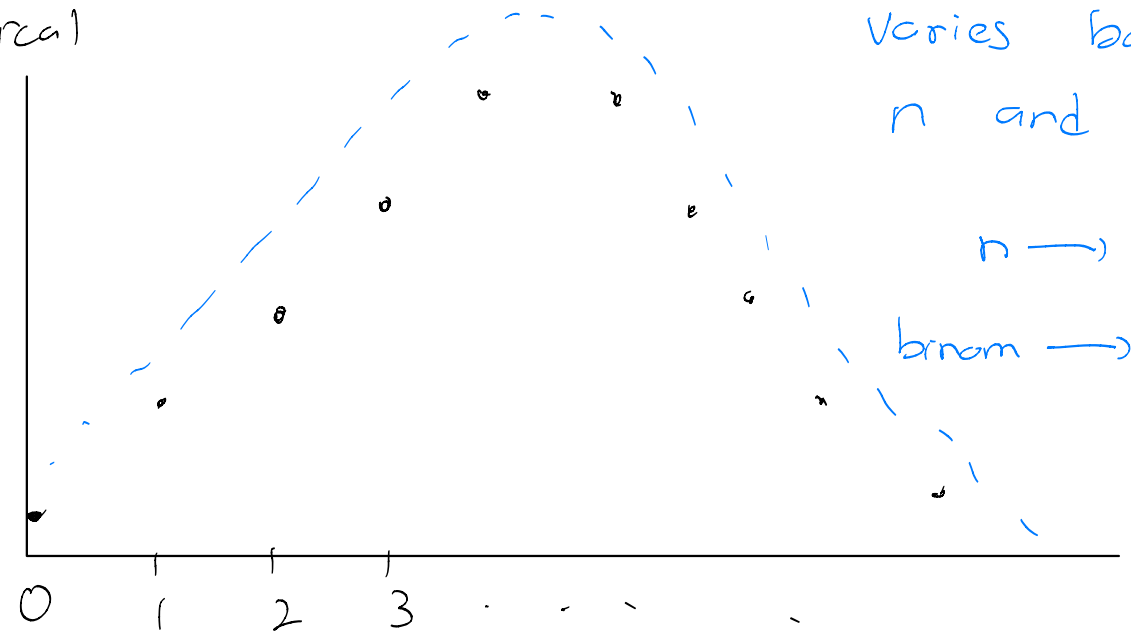
Variance: $\sigma^2 = Var(x) = np(1-p)$

Binomial Distribution (Discrete)

PMF:

$$P(X=x) = \binom{n}{x} p^x (1-p)^{n-x} \qquad x = 0, 1, \ldots, n$$

Graphical



Varies based on n and p

$n \rightarrow$ larger

binom $\rightarrow$ "curve"

CDF

$$F(x) = P(X \leq x) = \sum_{j=1}^{x} \binom{n}{j} p^j (1-p)^{n-j}$$

calculations can be time consuming for large n

under certain conditions we can approximate binomial probabilities with the normal distrib

Bernoulli Distribution (Binomial($n = 1$, p))

Random experiment: Rolling a die once.
Random variable:

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

$p = \text{P(rolling a six)}$
$\mu = E(X_i) = p$
$\sigma^2 = V(X_i) = p(1-p)$

Example

Consider determining the sampling distribution of the sample total $S_n = X_1 + X_2 + ... + X_n$. Suppose that a random sample of size n is taken from a Bernoulli(p). Then,

$$\begin{aligned}
M_{S_n}(t) &= E[e^{tS_n}] \\
&= E[e^{t(X_1+X_2+...+X_n)}] \\
&= E[e^{tX_1}e^{tX_2}...e^{tX_n}] \quad \text{(independence)} \\
&= E[e^{tX_1}]E[e^{tX_2}]...E[e^{tX_n}] \\
&= M_{X_1}(t)M_{X_2}(t)...M_{X_n}(t) \\
&= [pe^t + (1-p)][pe^t + (1-p)]...[pe^t + (1-p)] \\
&= [pe^t + (1-p)]^n
\end{aligned}$$

Example (cont.)

On comparing $M_{S_n}(t)$ with the moment-generating function of a Binomial random variable, we see that S_n must have a Binomial distribution with parameters n and p.

Example

So, we can 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 Y. The distribution of Y 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

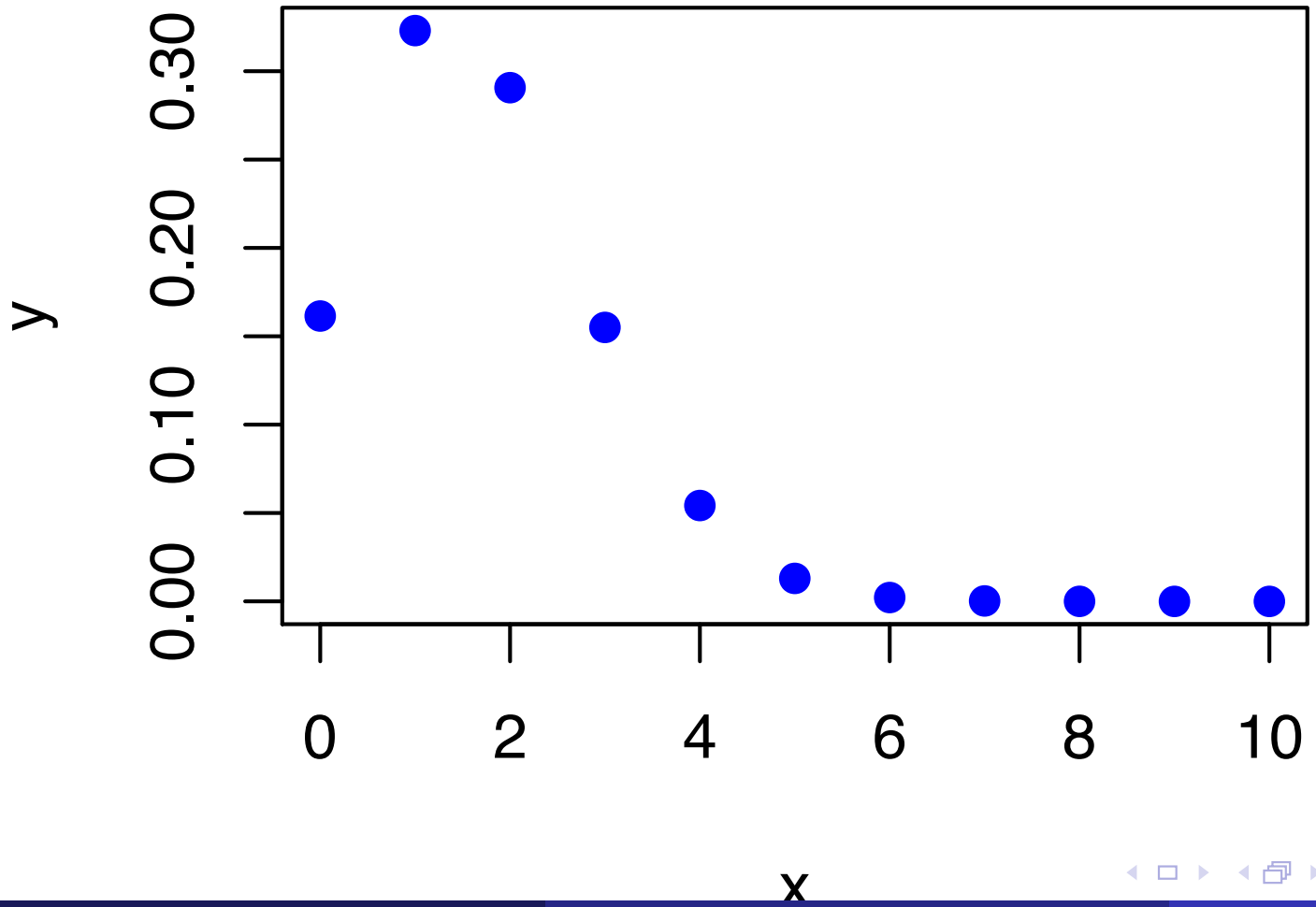
$$p(y) = \frac{n!}{y!(n-y)!} p^y (1-p)^{n-y}, \ \ y = 0, 1, 2, ..., n \text{ and } 0 \leq p \leq 1.$$

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

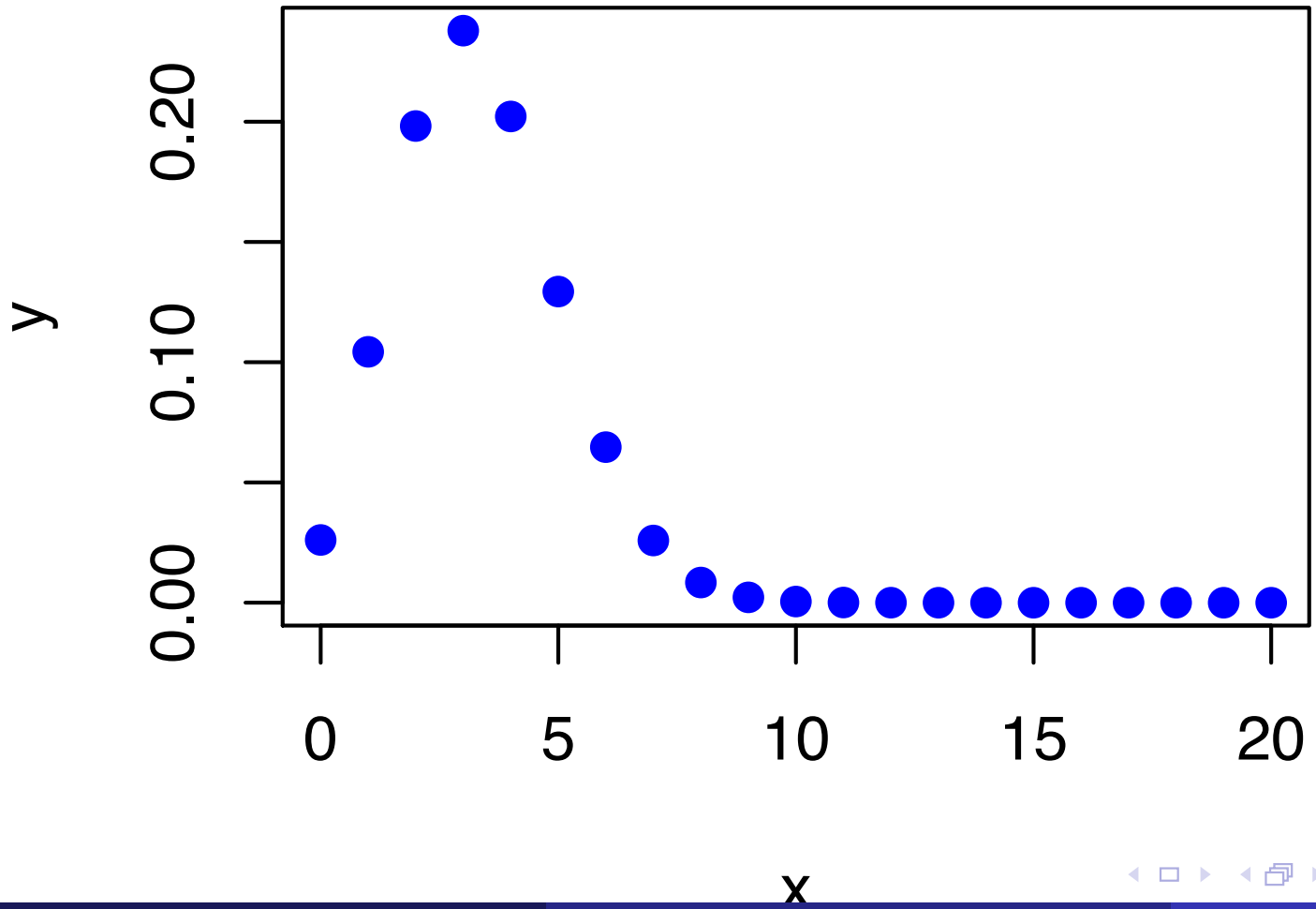
Probability mass function when n=10 and p=1/6

```
## 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);
```

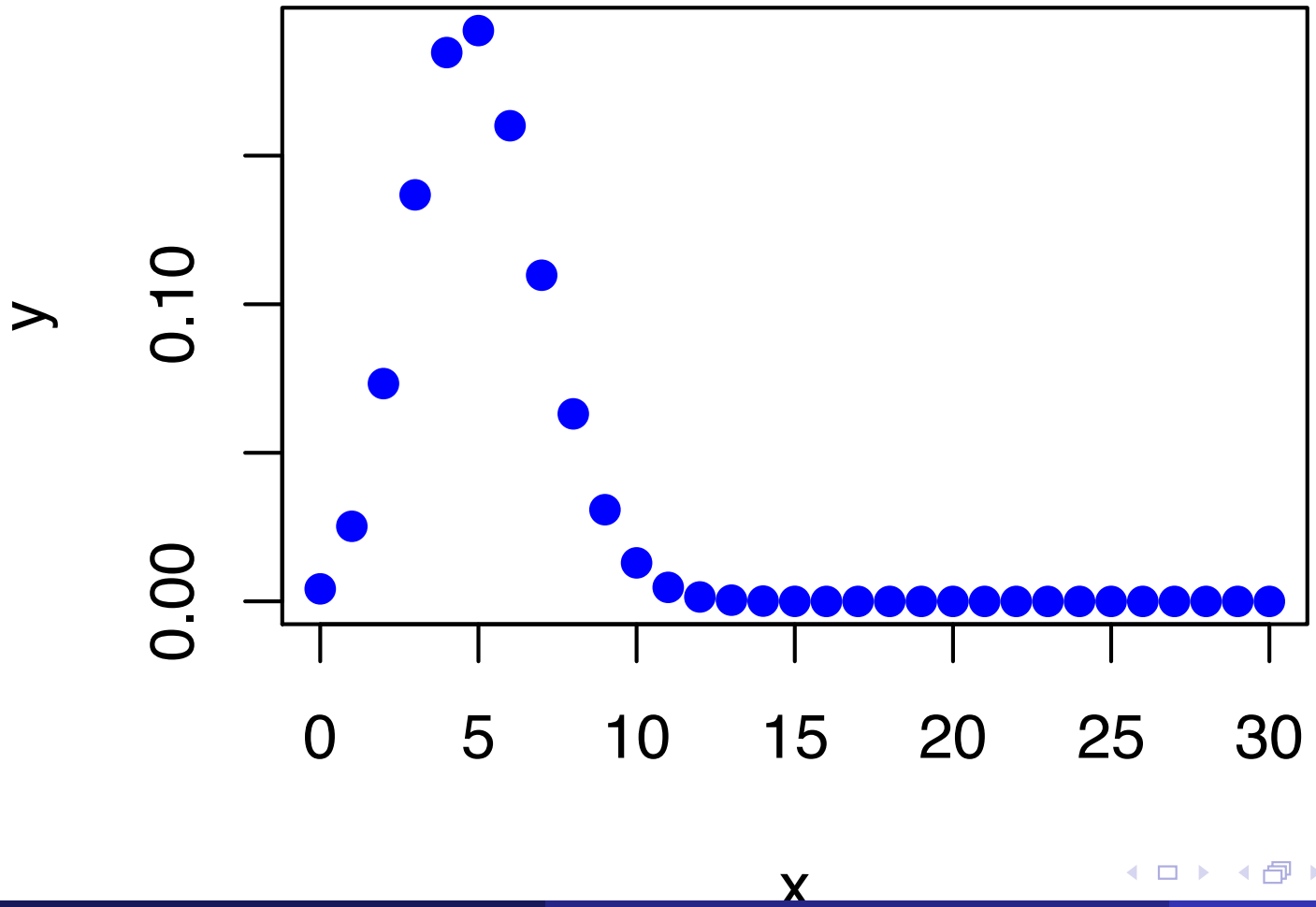
PMF when n=10 and p=1/6



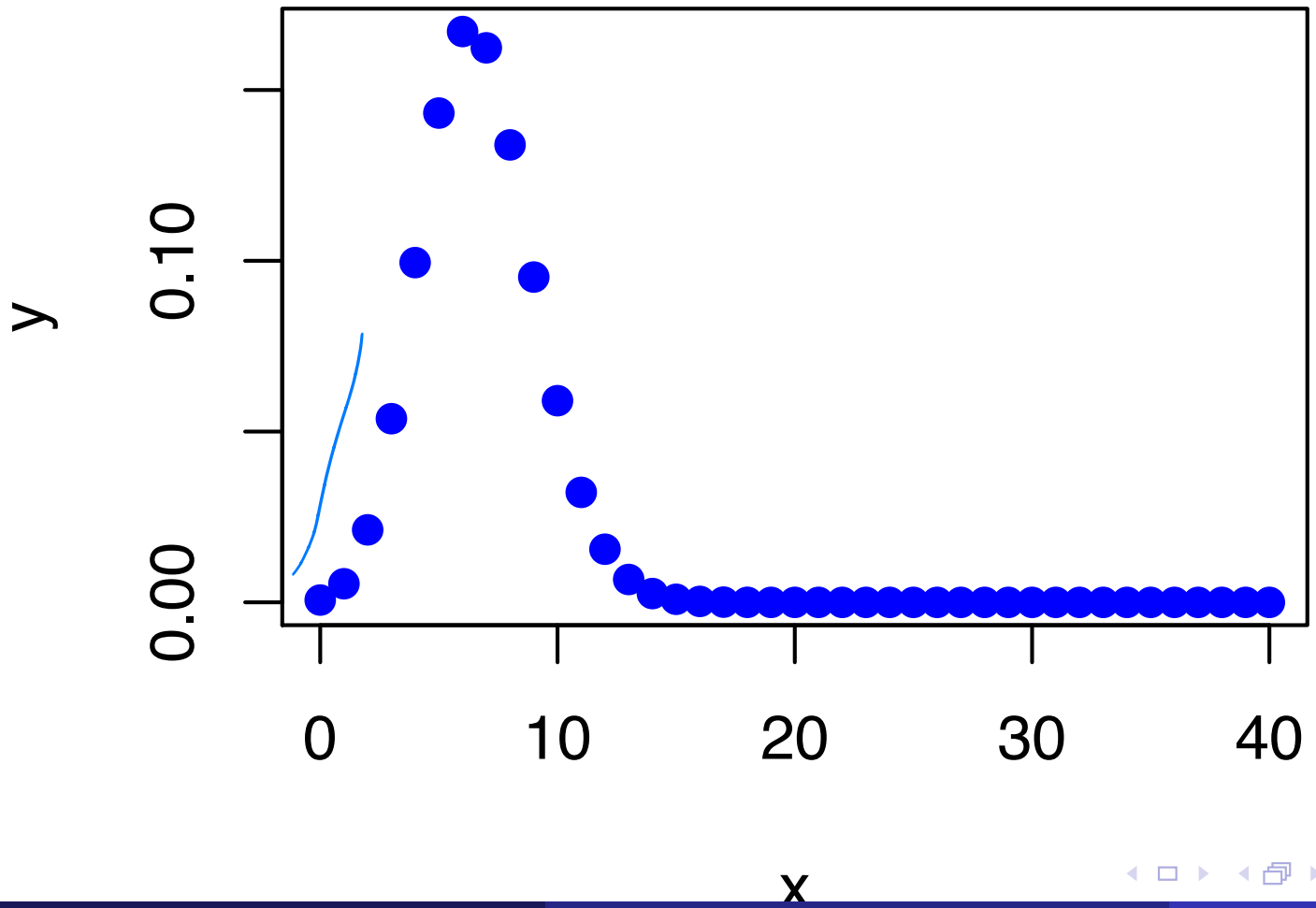
PMF when n=20 and p=1/6



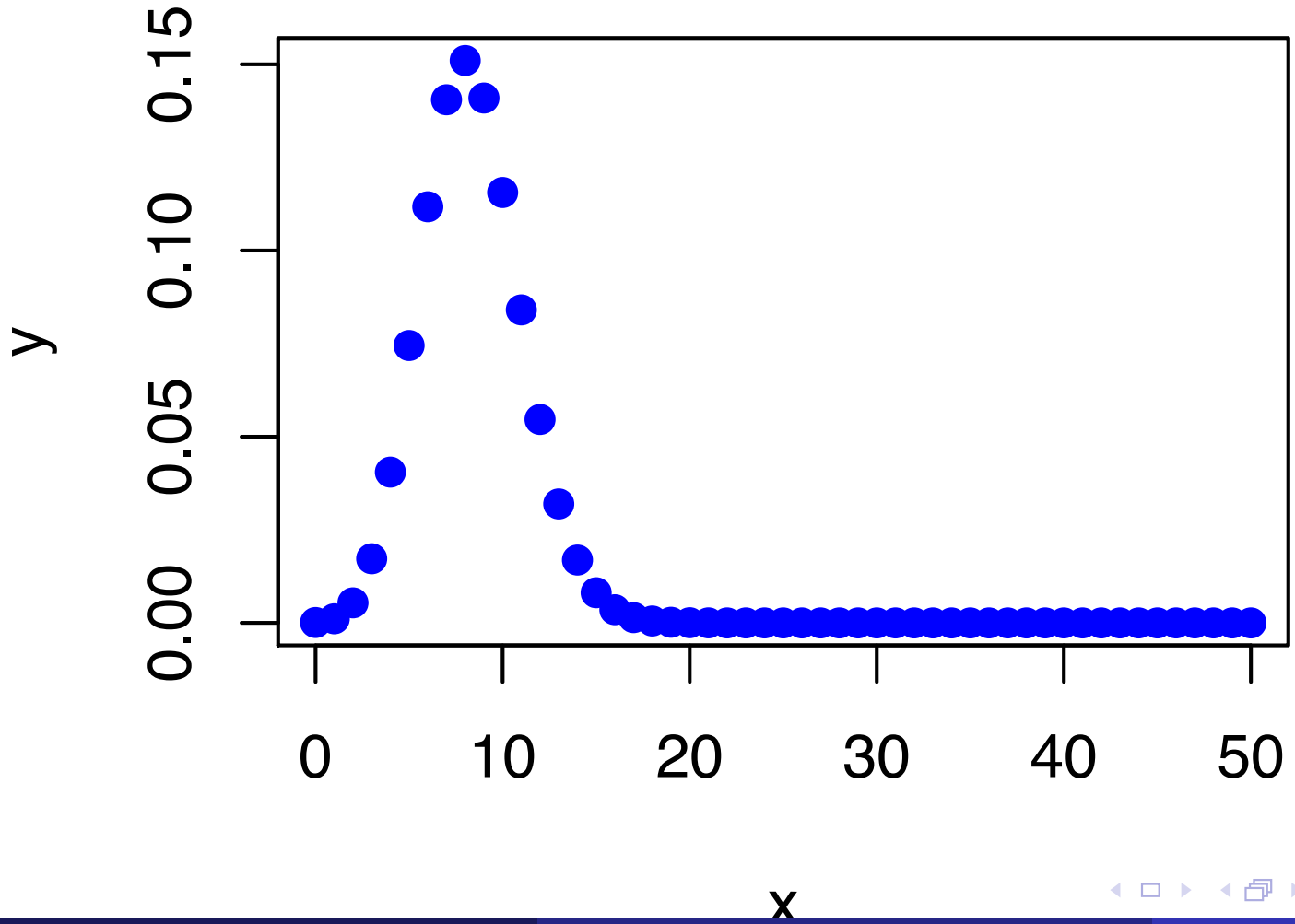
PMF when n=30 and p=1/6



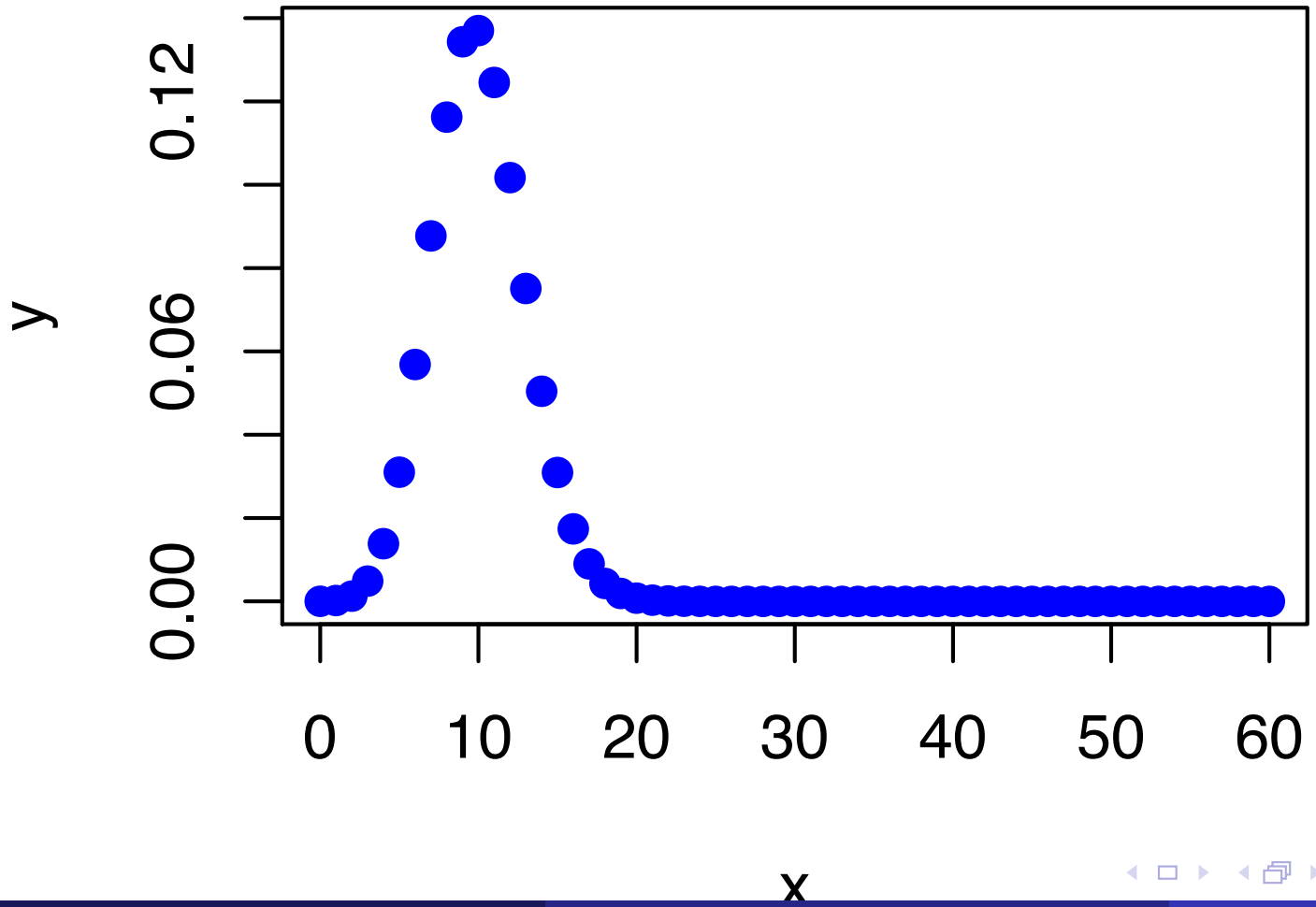
PMF when n=40 and p=1/6



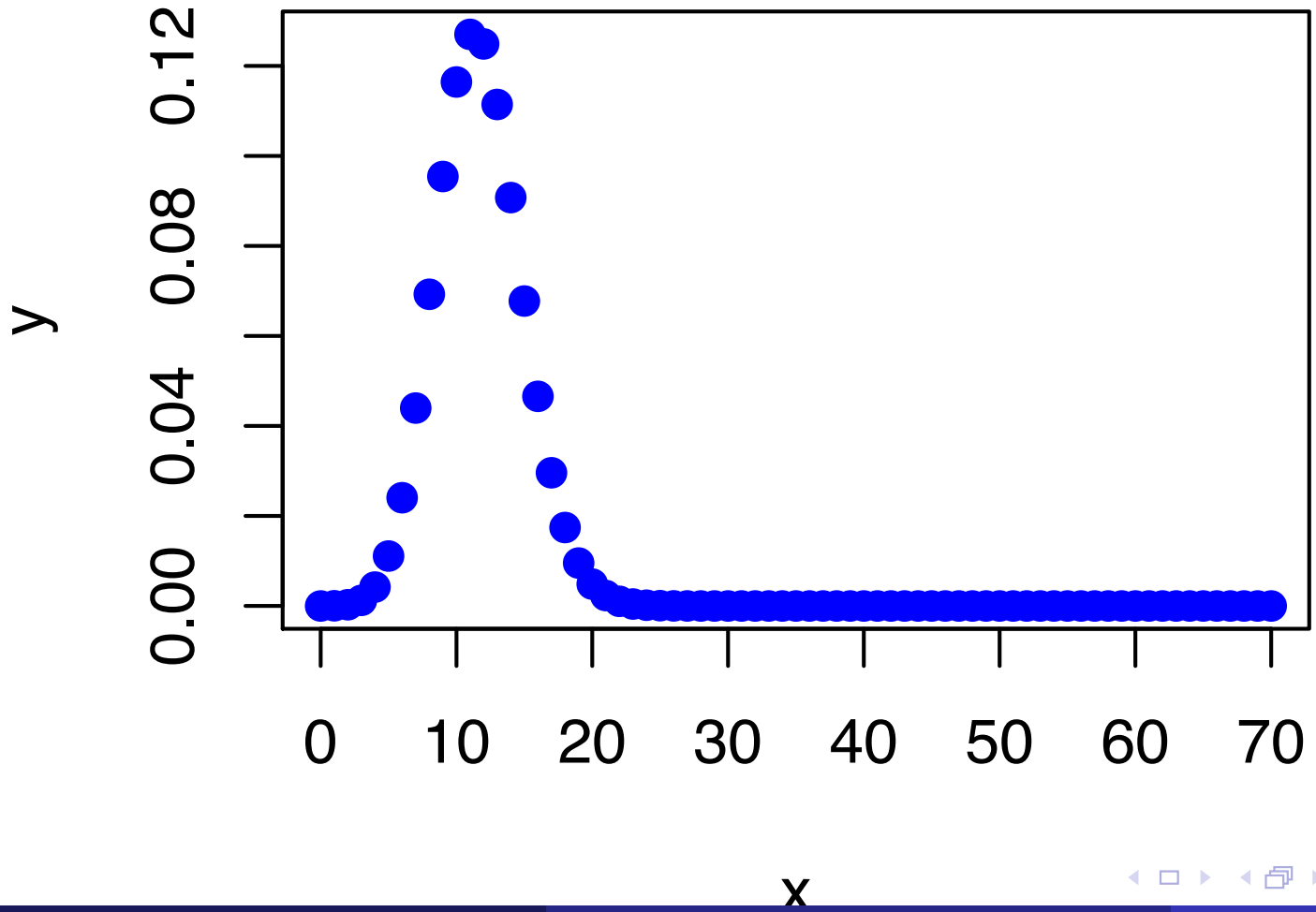
PMF when n=50 and p=1/6



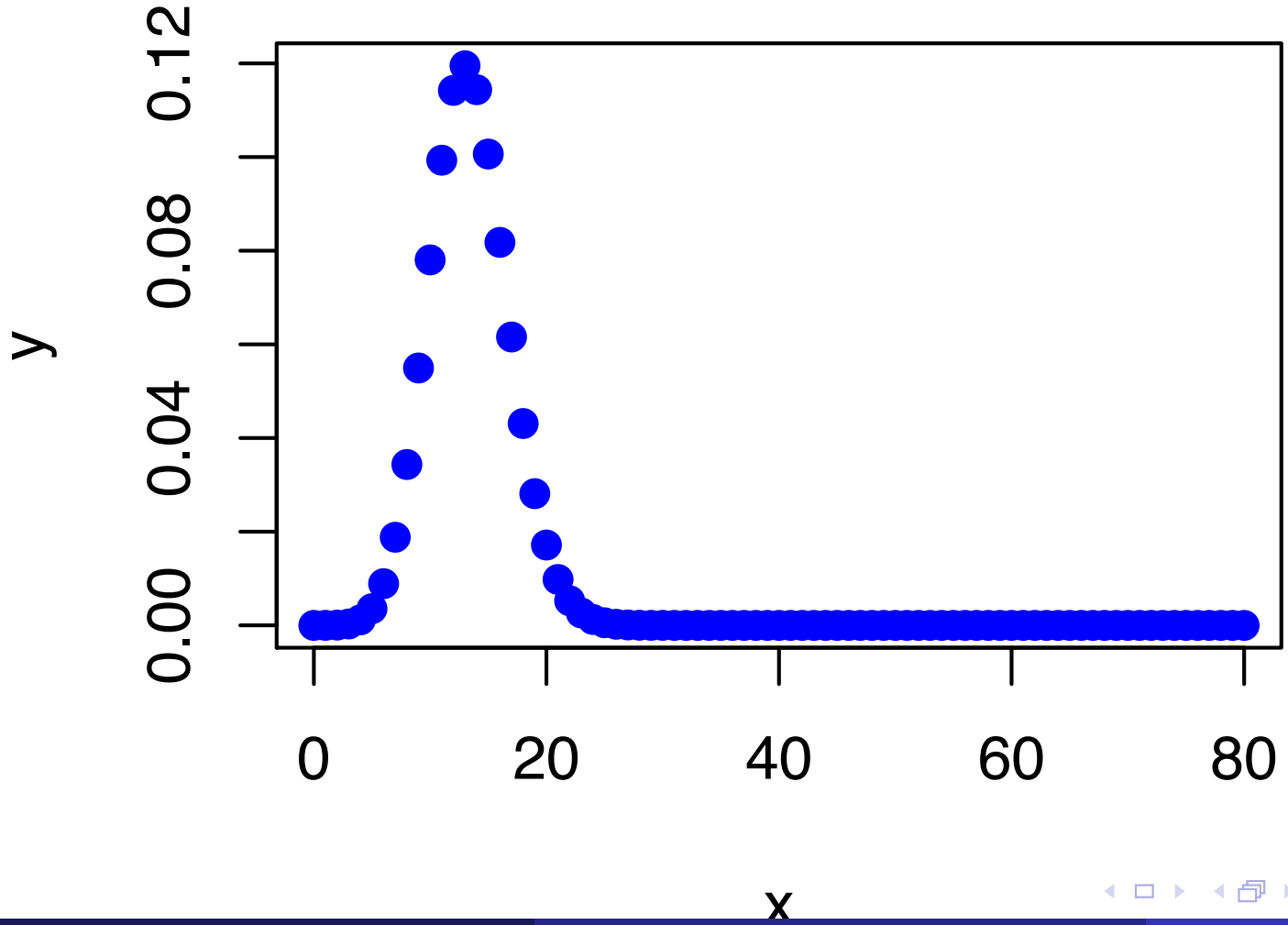
PMF when n=60 and p=1/6



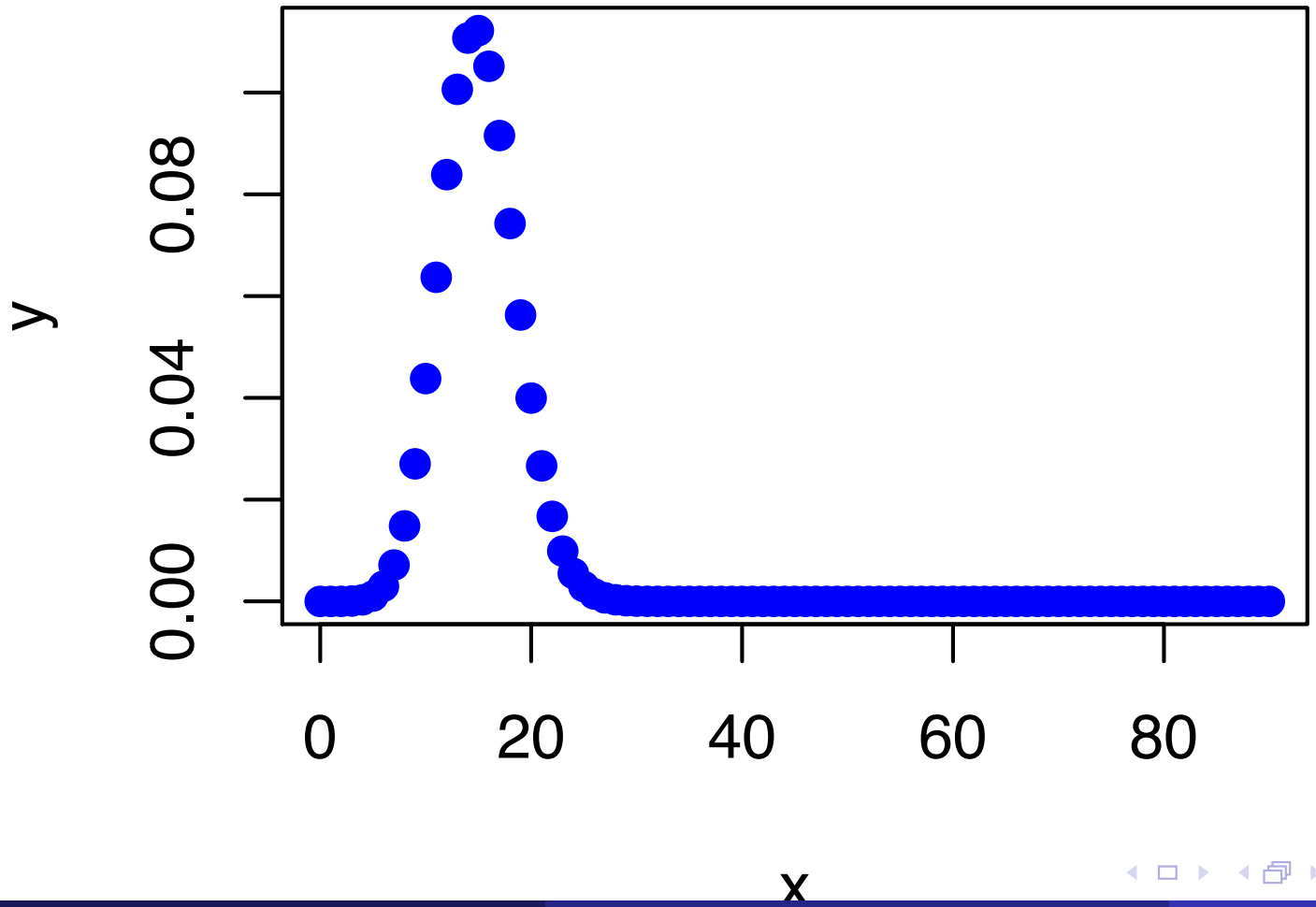
PMF when n=70 and p=1/6



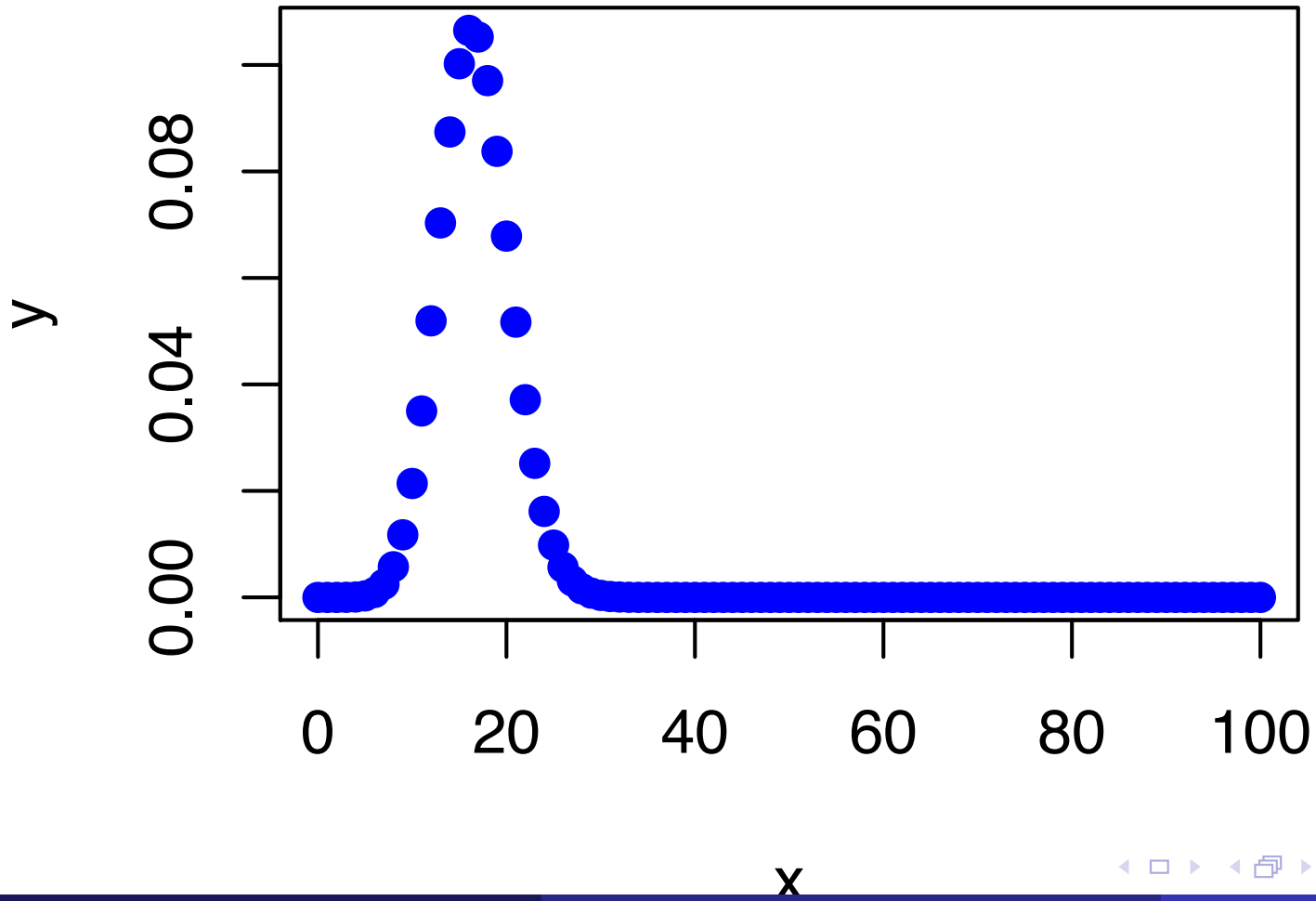
PMF when n=80 and p=1/6



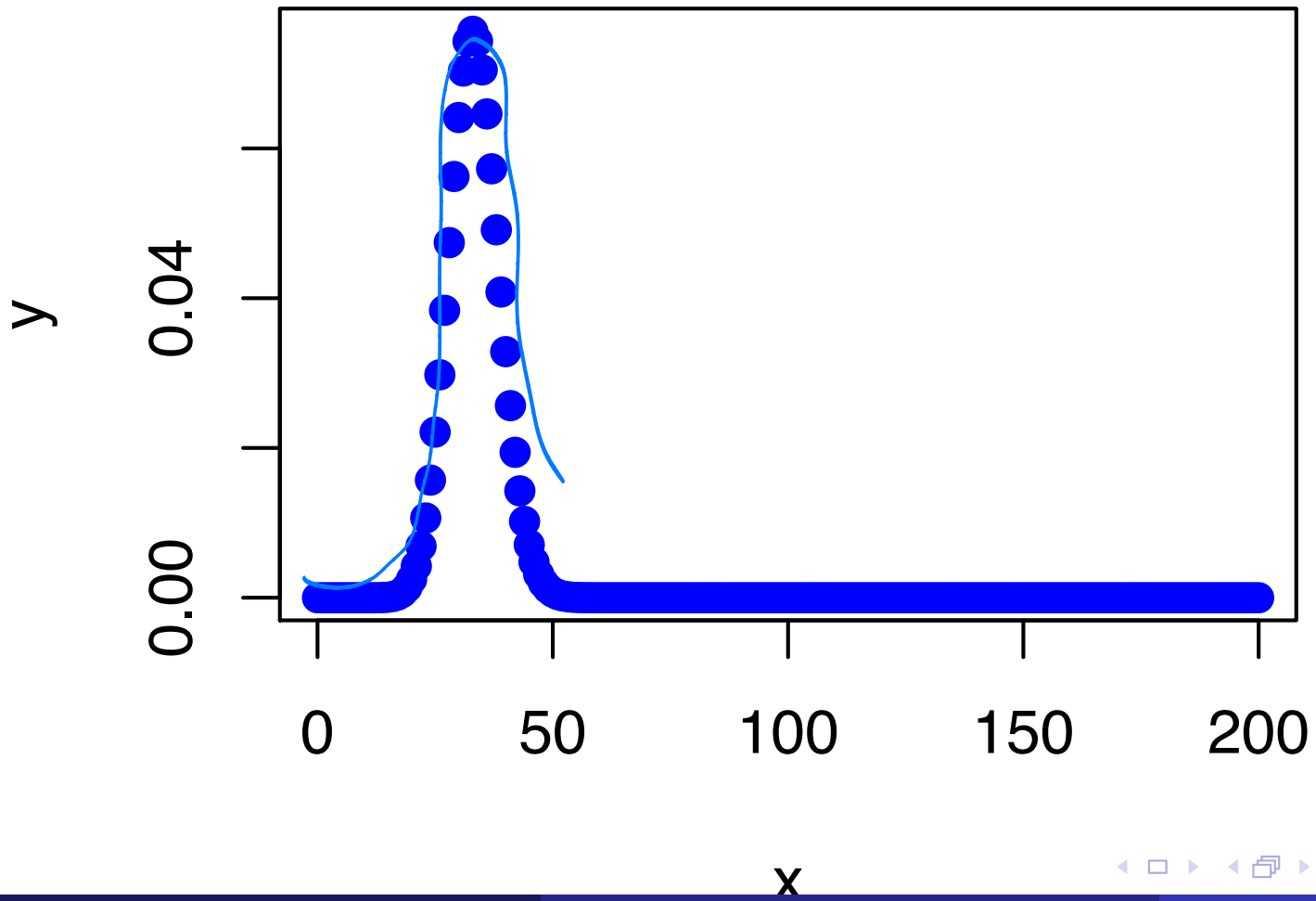
PMF when n=90 and p=1/6



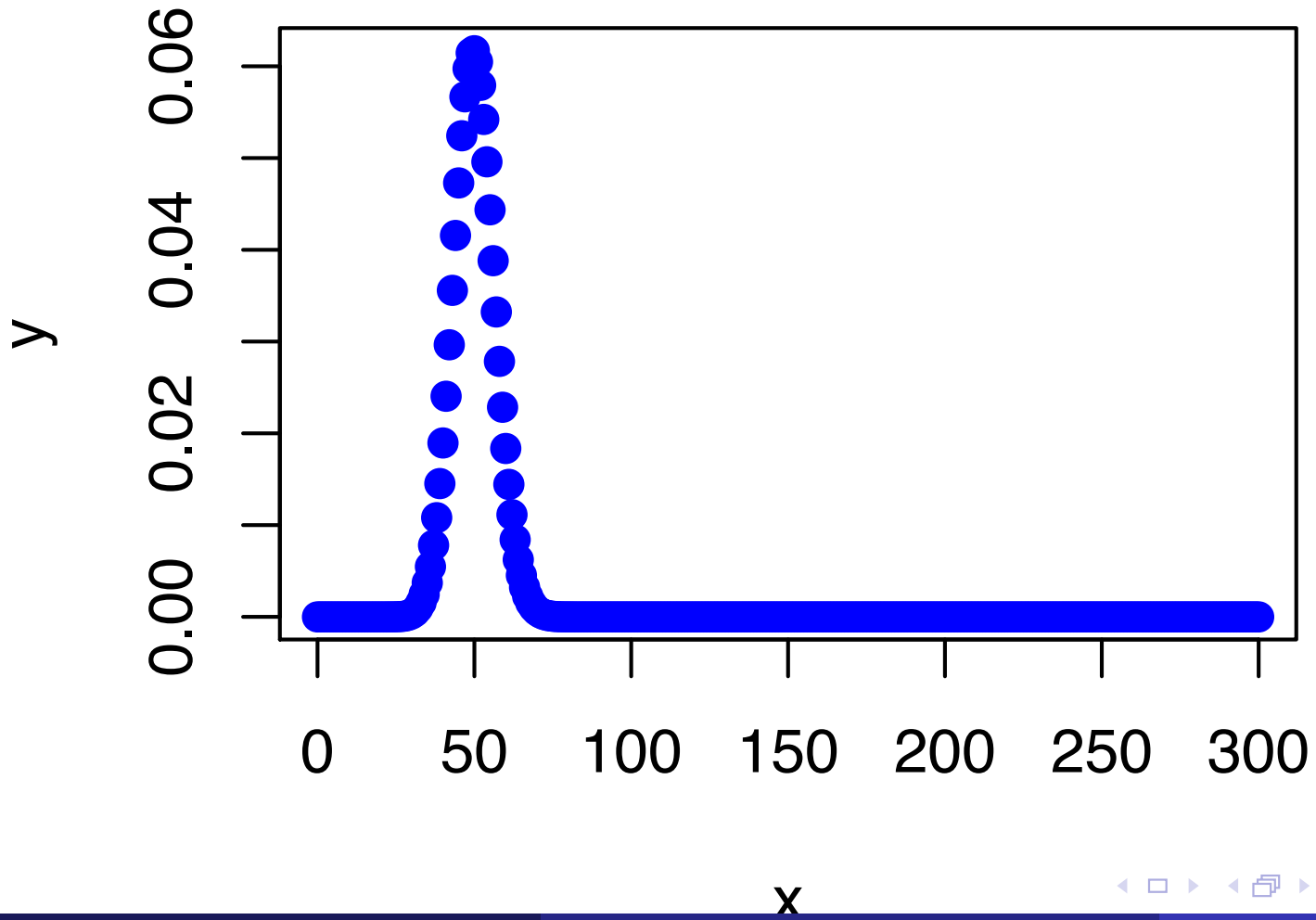
PMF when n=100 and p=1/6



PMF when n=200 and p=1/6



PMF when n=300 and p=1/6



Let $X \sim \text{Binomial}(n, p)$

$\mu = np$, $\sigma^2 = np(1-p)$ ↓ useful for large n

Binomial probabilities can be approximated by

$$X \overset{\cdot}{\sim} N(\mu = np, \sigma^2 = np(1-p))$$

approx distrib

under the following conditions

1/ conditions for binomial hold (n indep trials, etc.)

2/ $np \geq 10$, AND $n(1-p) \geq 10$ ← Standard

~~np~~ $n(1-p) \geq 10$

3/ Alternate criteria

$$n > 9\left(\frac{\max(p, 1-p)}{\min(p, 1-p)}\right)$$

Example → Slide 39

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

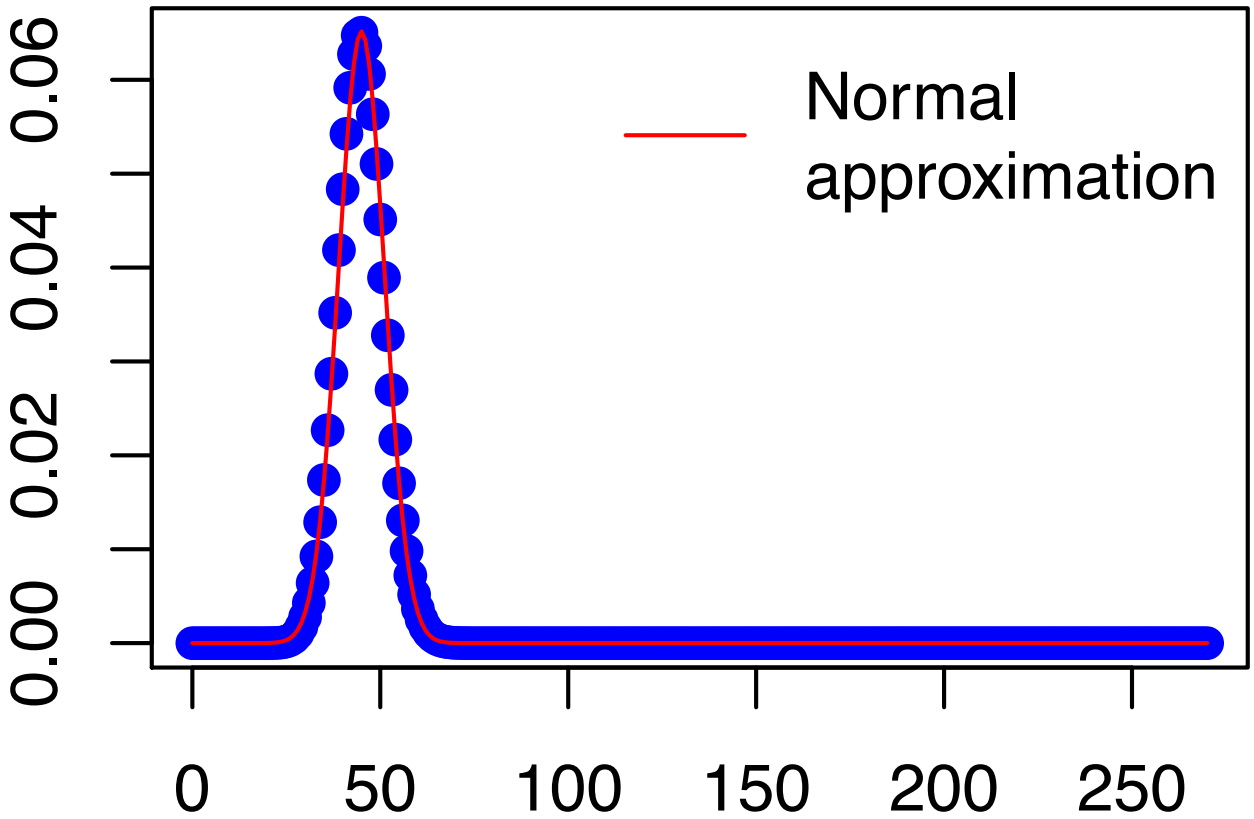
Draw an SRS of size n from a large population that contains proportion p of "successes". Let $\hat{p}$ be the **sample proportion** of successes,

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

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 $N\left(p, \sqrt{\frac{p(1-p)}{n}}\right)$

proportion (percentage fraction < 1)

A consequence of normal approx of binomial is the derivation of the sampling distribution of the sample proportion ($\hat{p}$)

$$\hat{p} = \frac{\#\text{ successes } (x)}{\text{sample size } (n)}$$

Recall for $X \sim \text{Binom}(n, p)$

$$X \overset{\cdot}{\sim} N(\mu = np,\ \sigma^2 = np(1-p))$$

$\text{var}(aX) = a^2 \text{var}(X)$

μ: mean $E(X)$; σ^2: $\text{var}(X)$

$$Z = \frac{X - \mu}{\sigma}$$

Let $\hat{p} = \frac{X}{n}$

Mean $\hat{p}$: $E(\hat{p}) = E\left(\frac{X}{n}\right) = \frac{1}{n} E(X) = \frac{1}{n} \cdot np = p$

var $\hat{p}$: $\text{var}(\hat{p}) = \text{var}\left(\frac{X}{n}\right) = \frac{1}{n^2} \text{var}(X) = \frac{1}{n^2} np(1-p) = \frac{p(1-p)}{n}$

By CLT (for suff large n)

$$\hat{p} \sim N\left(\mu_{\hat{p}} = p,\ \sigma^2_{\hat{p}} = \frac{p(1-p)}{n}\right)$$

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

Example

$p = 0.52$

$n = 300$

PERCENTAGE
Hint: consider a prob related to proportions

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?

$\hat{p} > 0.5$

Example (Slide 27)

$p = 0.52, \quad n = 300$ want $P(\hat{p} > 0.50)$

Distribution of $\hat{p}$ is

$$\hat{p} \sim N\left(\mu_{\hat{p}} = p, \ \sigma_{\hat{p}}^2 = \frac{p(1-p)}{n}\right)$$

$$\hat{p} \sim N\left(\mu_{\hat{p}} = 0.52, \ \sigma_{\hat{p}}^2 = \frac{13}{15625}\right)$$

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

$$P(\hat{p} > 0.50)$$

$$= P\left(\frac{\hat{p} - p}{\sqrt{\frac{p(1-p)}{n}}} > \frac{0.50 - p}{\sqrt{\frac{p(1-p)}{n}}}\right)$$

$p = 0.52$

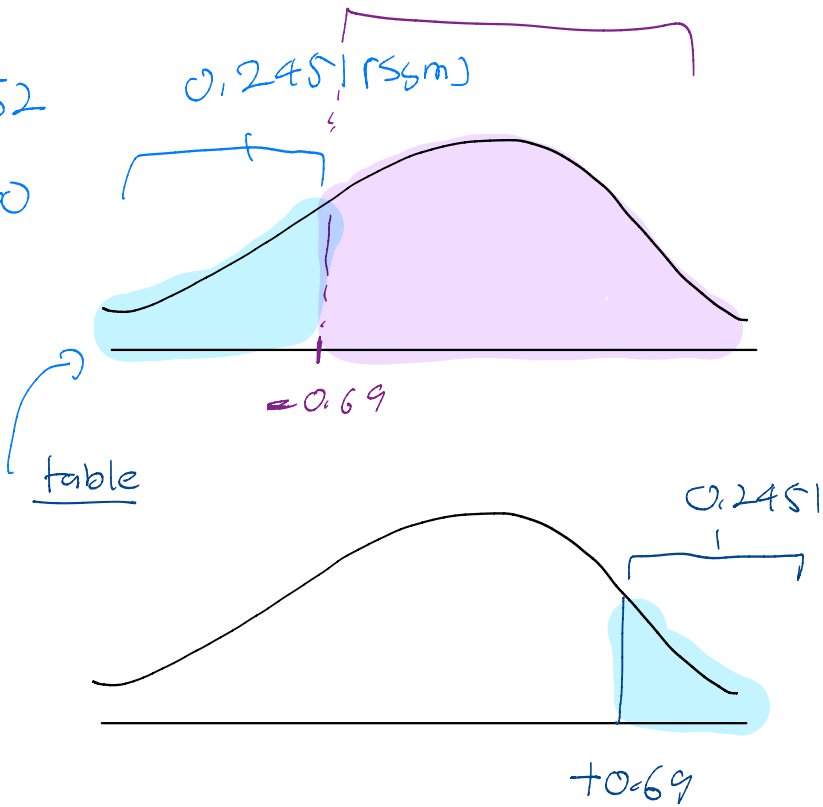
$n = 300$

Z

$$= P(Z > -0.69)$$

$$= 1 - 0.2451$$

$$= 0.7549$$



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.

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

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

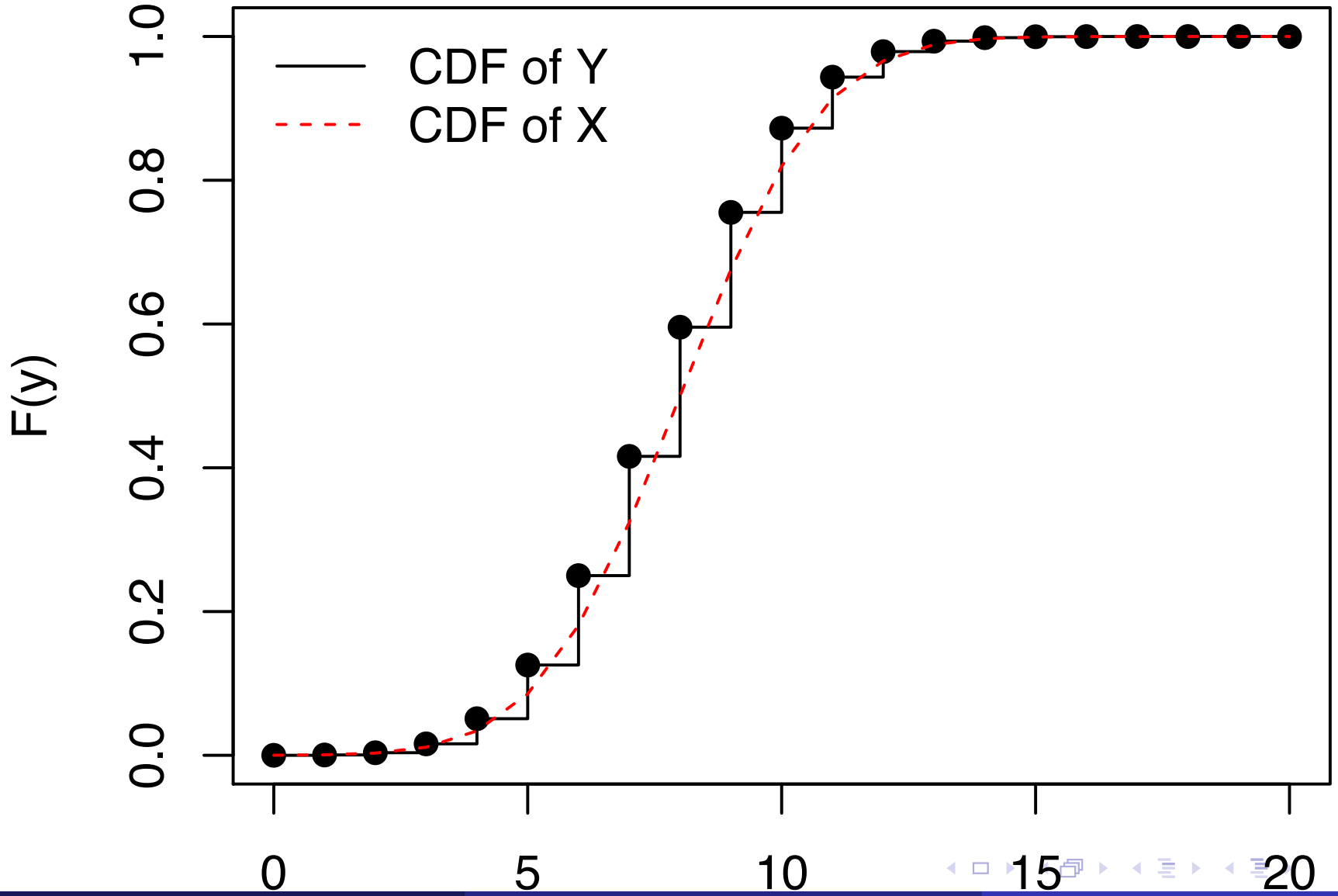
$$= P(Z \leq 0) = 0.5$$

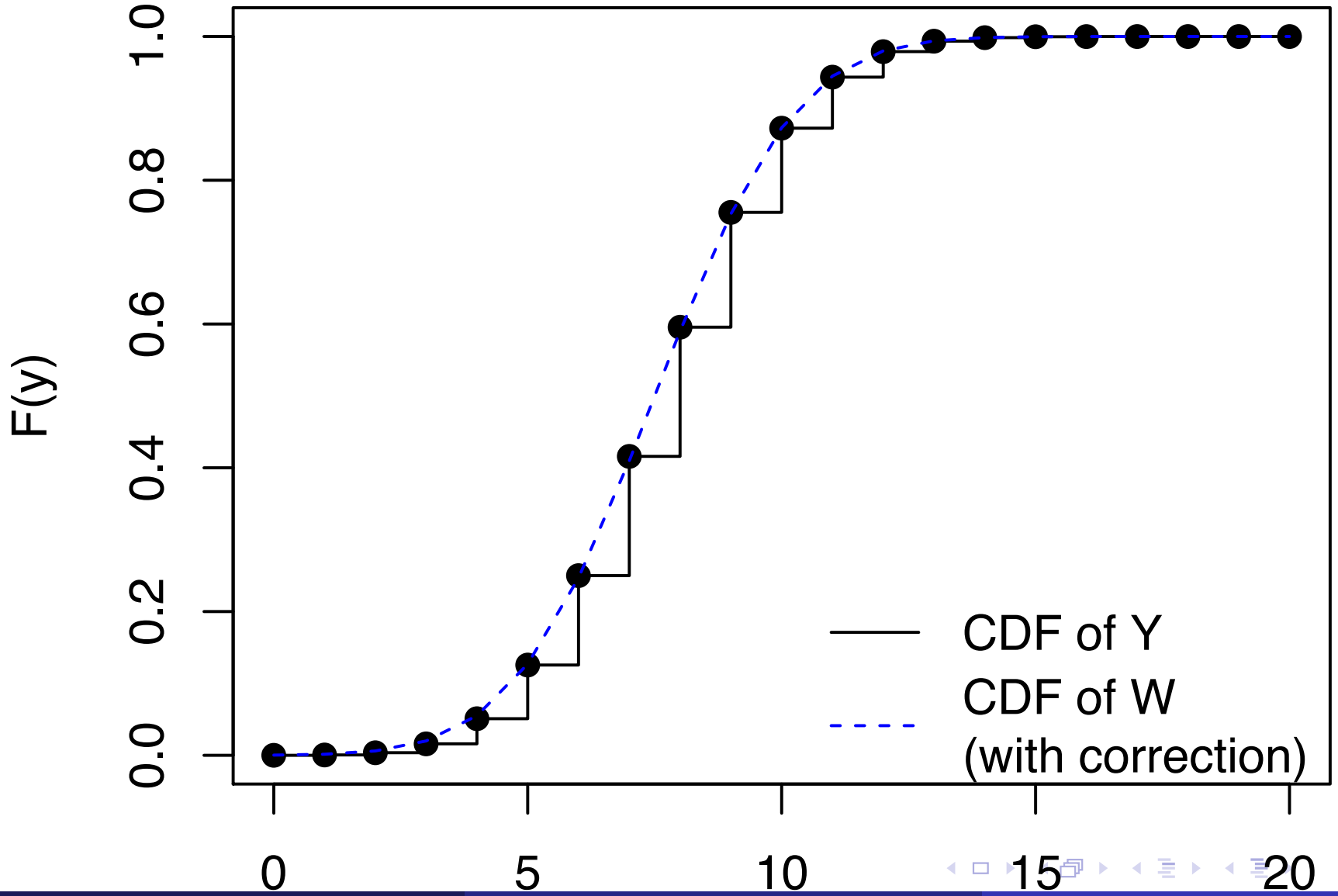
Continuity Correction (cont.)

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

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

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


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

(large) $n=65$

$p=0.51$

hint: use normal approx

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?

$X=40$

Use normal approximation

Example (slide 39)

large

$X \sim \text{Binomial}(n = 65,\ p = 0.51)$

Normal approx

$X \overset{\cdot}{\sim} N(\mu = np,\ \sigma^2 = np(1-p))$

$X \overset{\cdot}{\sim} N(\mu = 33.15,\ \sigma^2 = 16.496)$

$\sigma = 4.06$

$P(X = 40)$

$= P(40 - 0.5 \leq X \leq 40 + 0.5)$

$= P(39.5 \leq X \leq 40.5)$

$= P\left(\frac{39.5 - \mu}{\sigma} \leq \frac{X - \mu}{\sigma} \leq \frac{40.5 - \mu}{\sigma}\right)$

$= Z$

$\mu = 33.15$

$\sigma = 4.06$

$Z = \frac{X - \mu}{\sigma}$

$= P(1.56 \leq Z \leq 1.81)$

$= 0.0594 - 0.0352$

$= 0.0242$ (check)

	Binomial Probability	Continuity Correction	Normal Approximation
1	$P(X = x)$	$P(x - 0.5 \leq X \leq x + 0.5)$	$P\left(\frac{x - 0.5 - \mu}{\sigma} \leq Z \leq \frac{x + 0.5 - \mu}{\sigma}\right)$
2	$P(X \leq x)$	$P(X \leq x + 0.5)$	$P\left(Z \leq \frac{x + 0.5 - \mu}{\sigma}\right)$
3	$P(X < x)$	$P(X \leq x - 0.5)$	$P\left(Z \leq \frac{x - 0.5 - \mu}{\sigma}\right)$
4	$P(X \geq x)$	$P(X \geq x - 0.5)$	$P\left(Z \geq \frac{x - 0.5 - \mu}{\sigma}\right)$
5	$P(X > x)$	$P(X \geq x + 0.5)$	$P\left(Z \geq \frac{x + 0.5 - \mu}{\sigma}\right)$
6	$P(a \leq X \leq b)$	$P(a - 0.5 \leq X \leq b + 0.5)$	$P\left(\frac{a - 0.5 - \mu}{\sigma} \leq Z \leq \frac{b + 0.5 - \mu}{\sigma}\right)$

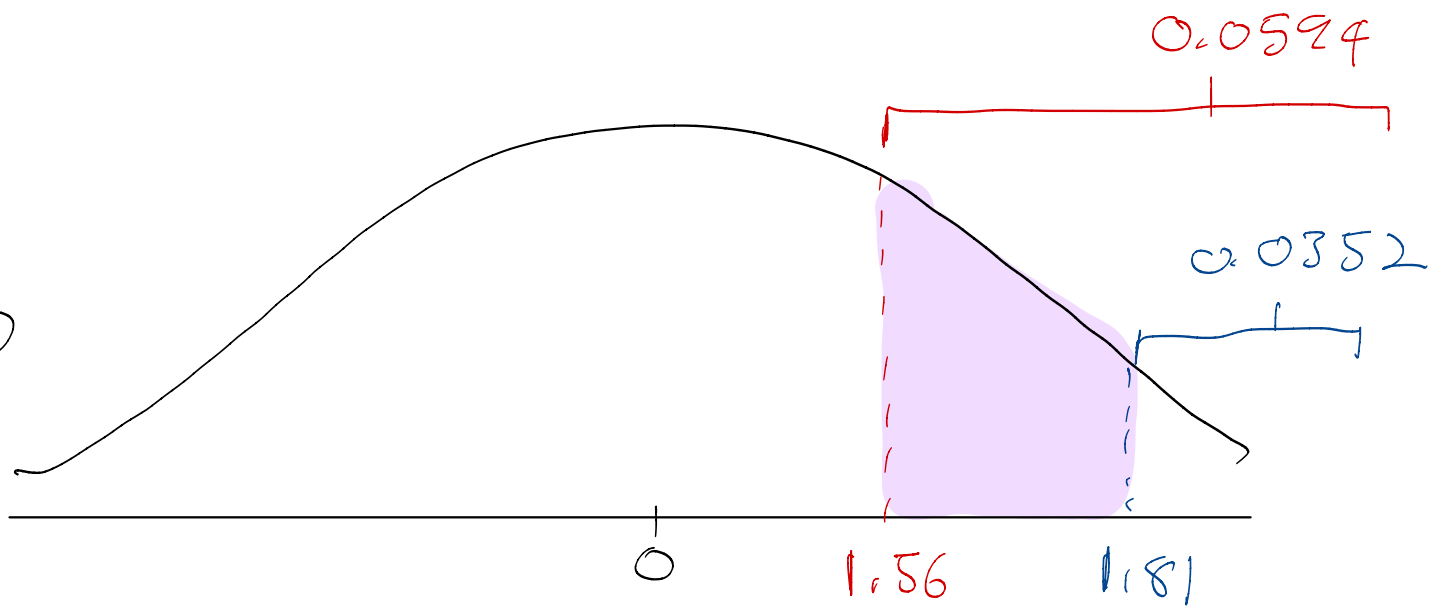
Table 1: Summary of Continuity Corrections for Binomial Distribution

Check

verify

$np \geq 10$

$np(1-p) \geq 10$



directly

$$P(X \leq 40) = P(X=0) + P(X=1) + \dots + P(X=40)$$

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

ASIDE

approx

$$\approx P\left(Z \leq \frac{40.5 - \mu}{\sigma}\right)$$

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 $np \geq 10$ and $n(1-p) \geq 10$.
2. Another criterion is that the Normal approximation is adequate if $n > 9\left(\frac{\text{larger of p and q}}{\text{smaller of p and q}}\right)$
3. The expected value: $E(n\hat{p}) = np$.
4. 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.