STA215
Testing Hypotheses about Proportions

Al Nosedal.
University of Toronto.
Summer 2017

June 17, 2019

Who wants to be a millionaire?

Let's say that one of you is invited to this popular show. As you probably know, you have to answer a series of multiple choice questions and there are four possible answers to each question. Perhaps you also have seen that if you don't know the answer to a question you could either "jump the question" or you could "ask the audience".
Suppose that you run into a question for which you don't know the answer with certainty and you decide to "ask the audience". Let's say that you initially believe that the right answer is **A**. Then you ask the audience and only 2% of the audience shares your opinion. What would you do? Change your initial answer or keep it?

Hypotheses Tests for a Proportion

To test the hypothesis $H_0 : p = p_0$, compute the z_* statistic,
$z_* = \frac{\hat{p} - p_0}{\sqrt{\frac{p_0(1-p_0)}{n}}}$
In terms of a variable Z having the standard Normal distribution, the approximate P-value for a test of H_0 against
$H_a : p > p_0 :$ is : $P(Z > z_*)$
$H_a : p < p_0 :$ is : $P(Z < z_*)$
$H_a : p \neq p_0 :$ is : $2P(Z > |z_*|)$

Example

Consider the following hypothesis test:
$H_0 : p = 0.75$
$H_a : p < 0.75$
A sample of 300 items was selected. Compute the p-value and state your conclusion for each of the following sample results. Use $\alpha = 0.05$.

a. $\hat{p} = 0.68$
b. $\hat{p} = 0.72$
c. $\hat{p} = 0.70$
d. $\hat{p} = 0.77$

Solution a.

$$z_* = \frac{\hat{p} - p_0}{\sqrt{p_0(1-p_0)/n}} = \frac{0.68 - 0.75}{\sqrt{0.75(1-0.75)/300}} = -2.80$$

Using Normal table, P-value $=$

$P(Z < z_*) = P(Z < -2.80) = 0.0026$

P-value$< \alpha = 0.05$, reject H_0.

Solution b.

$$z_* = \frac{\hat{p} - p_0}{\sqrt{p_0(1-p_0)/n}} = \frac{0.72 - 0.75}{\sqrt{0.75(1-0.75)/300}} = -1.20$$

Using Normal table, P-value $=$

$P(Z < z_*) = P(Z < -1.20) = 0.1151$

P-value$> \alpha = 0.05$, do not reject H_0.

Solution c.

$$z_* = \frac{\hat{p} - p_0}{\sqrt{p_0(1-p_0)/n}} = \frac{0.70 - 0.75}{\sqrt{0.75(1-0.75)/300}} = -2.00$$

Using Normal table, P-value $=$

$P(Z < z_*) = P(Z < -2.00) = 0.0228$

P-value$< \alpha = 0.05$, reject H_0.

Example

Consider the following hypothesis test:
$H_0 : p = 0.20$
$H_a : p \neq 0.20$
A sample of 400 provided a sample proportion $\hat{p} = 0.175$.
a. Compute the value of the test statistic.
b. What is the p-value?
c. At the $\alpha = 0.05$, what is your conclusion?
d. What is the rejection rule using the critical value? What is your conclusion?

Solution

a. $z_* = \frac{\hat{p}-p_0}{\sqrt{\frac{p_0(1-p_0)}{n}}} = \frac{0.175-0.20}{\sqrt{\frac{(0.20)(0.80)}{400}}} = -1.25$

b. Using Normal table, P-value $= 2P(Z > |z_*|) = 2P(Z > |-1.25|) = 2P(Z > 1.25) = 2(0.1056) = 0.2112$

c. P-value $> \alpha = 0.05$, we CAN'T reject H_0.

Problem

A study found that, in 2005, 12.5% of U.S. workers belonged to unions. Suppose a sample of 400 U.S. workers is collected in 2006 to determine whether union efforts to organize have increased union membership.

a. Formulate the hypotheses that can be used to determine whether union membership increased in 2006.

b. If the sample results show that 52 of the workers belonged to unions, what is the p-value for your hypothesis test?

c. At $\alpha = 0.05$, what is your conclusion?

Solution

a. $H_0 : p = 0.125$ vs $H_a : p > 0.125$

b. $\hat{p} = \frac{52}{400} = 0.13$

$z_* = \frac{\hat{p} - p_0}{\sqrt{\frac{p_0(1-p_0)}{n}}} = \frac{0.13 - 0.125}{\sqrt{\frac{(0.125)(0.875)}{400}}} = 0.30$

Using Normal table, P-value $=$

$P(Z > z_*) = P(Z > 0.30) = 1 - 0.6179 = 0.3821$

c. P-value $= > 0.05$, do not reject H_0. We cannot conclude that there has been an increase in union membership.

R Code

```
prop.test(52,400,p=0.125,alternative="greater",
correct=FALSE);

##
##  1-sample proportions test without continuity correction
##
## data:  52 out of 400, null probability 0.125
## X-squared = 0.091429, df = 1, p-value = 0.3812
## alternative hypothesis: true p is greater than 0.125
## 95 percent confidence interval:
##  0.1048085 1.0000000
## sample estimates:
##    p
## 0.13
```

Problem

A study by Consumer Reports showed that 64% of supermarket shoppers believe supermarket brands to be as good as national name brands. To investigate whether this result applies to its own product, the manufacturer of a national name-brand ketchup asked a sample of shoppers whether they believed that supermarket ketchup was as good as the national brand ketchup.

Problem (cont.)

a. Formulate the hypotheses that could be used to determine whether the percentage of supermarket shoppers who believe that the supermarket ketchup was as good as the national brand ketchup differed from 64%.
b. If a sample of 100 shoppers showed 52 stating that the supermarket brand was as good as the national brand, what is the p-value?
c. At $\alpha = 0.05$, what is your conclusion?

Solution

a. $H_0 : p = 0.64$ vs $H_a : p \neq 0.64$

b. $\hat{p} = \frac{52}{100} = 0.52$

$z_* = \frac{\hat{p} - p_0}{\sqrt{\frac{p_0(1-p_0)}{n}}} = \frac{0.52 - 0.64}{\sqrt{\frac{(0.64)(0.36)}{100}}} = -2.50$

Using Normal table, P-value $= 2P(Z > |z_*|) = 2P(Z > |-2.50|) = 2P(Z > 2.50) = 2(0.0062) = 0.0124$

c. P-value $= < 0.05$, reject H_0. Proportion differs from the reported 0.64.

R Code

```
prop.test(52,100,p=0.64,alternative="two.sided",
correct=FALSE);

##
##  1-sample proportions test without continuity correction
##
## data:  52 out of 100, null probability 0.64
## X-squared = 6.25, df = 1, p-value = 0.01242
## alternative hypothesis: true p is not equal to 0.64
## 95 percent confidence interval:
##  0.4231658 0.6153545
## sample estimates:
##    p
## 0.52
```

Problem

The National Center for Health Statistics released a report that stated 70% of adults do not exercise regularly. A researcher decided to conduct a study to see whether the claim made by the National Center for Health Statistics differed on a state-by-state basis.

a. State the null and alternative hypotheses assuming the intent of the researcher is to identify states that differ from 70% reported by the National Center for Health Statistics.

b. At $\alpha = 0.05$, what is the research conclusion for the following states:

Wisconsin: 252 of 350 adults did not exercise regularly.

California: 189 of 300 adults did not exercise regularly.

Solution (Wisconsin)

a. $H_0 : p = 0.70$ vs $H_a : p \neq 0.70$

b. Wisconsin $\hat{p} = \frac{252}{350} = 0.72$

$z_* = \frac{\hat{p} - p_0}{\sqrt{\frac{p_0(1-p_0)}{n}}} = \frac{0.72 - 0.70}{\sqrt{\frac{(0.70)(0.30)}{350}}} = 0.82$

Using Normal table, P-value $= 2P(Z > |z_*|) = 2P(Z > |0.82|) = 2P(Z > 0.82) = 2(0.2061) = 0.4122$

c. P-value > 0.05, we don't have enough evidence to reject H_0. There is not enough evidence against the claim made by the National Center for Health Statistics.

One last example: 100-Cup Challenge

A YouTuber goes to her nearest Tim Hortons and buys 100 empty cups. After rolling up the rims, she ends up with 12 winning cups out of the 100 she bought, all of them were food prizes.
If the probability of winning a food prize is supposed to be $\frac{1}{6}$, does she have evidence to claim that the probability of winning a food prize is less than $\frac{1}{6}$?

Probability Models

The **sample space S** of a random phenomenon is the set of all possible outcomes.
An **event** is an outcome or a set of outcomes of a random phenomenon. That is, an event is a subset of the sample space.
A **probability model** is a mathematical description of a random phenomenon consisting of two parts: a sample space S and a way of assigning probabilities to events.

Example

Rolling a fair die (random phenomenon). There are 6 possible outcomes when we roll a die.
The sample space for rolling a die and counting the pips is

$$S = \{1,\ 2,\ 3,\ 4,\ 5,\ 6\}$$

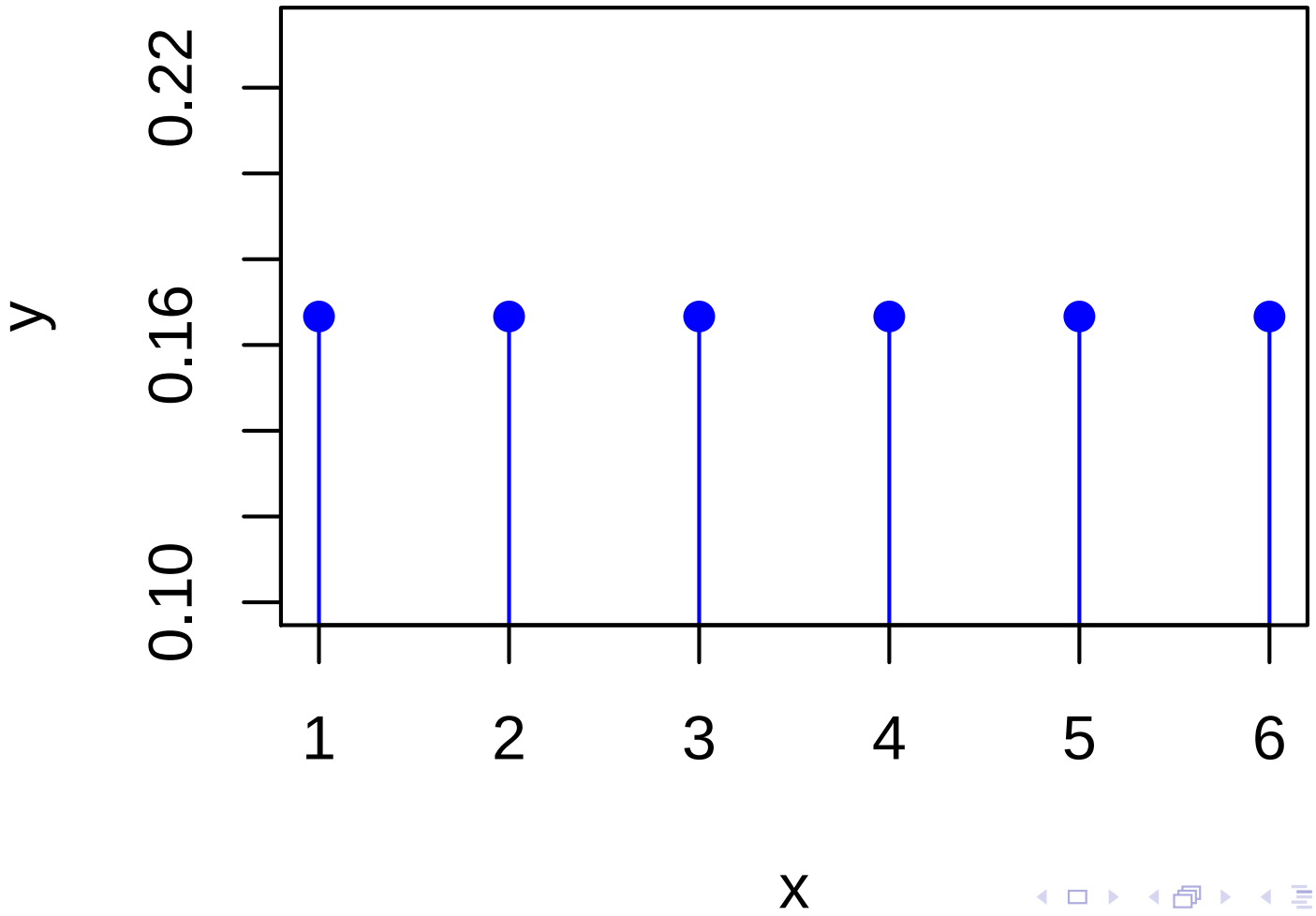
"Roll a 6" is an event, that contains one of these 6 outcomes.

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,

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

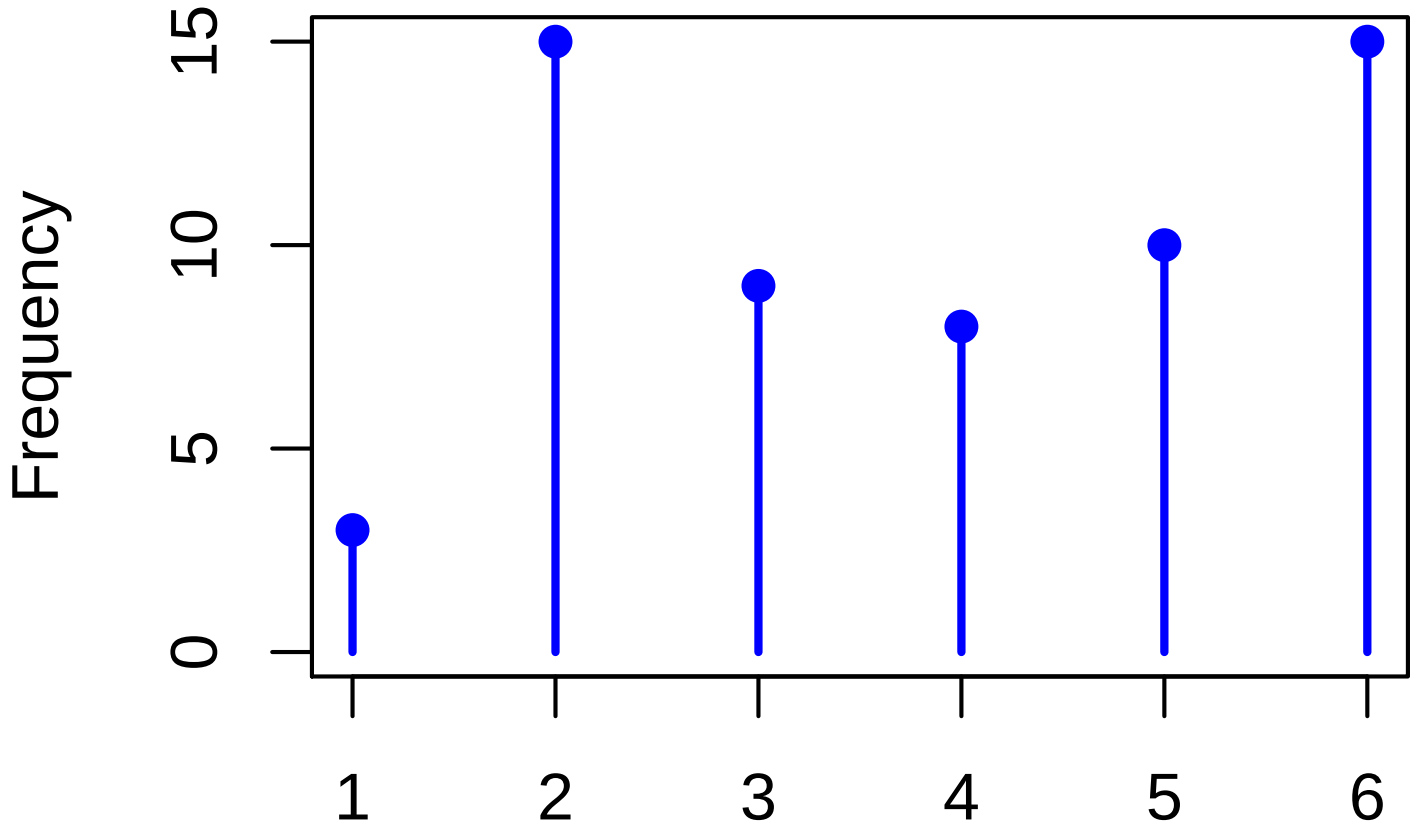
Probability mass function (pmf)



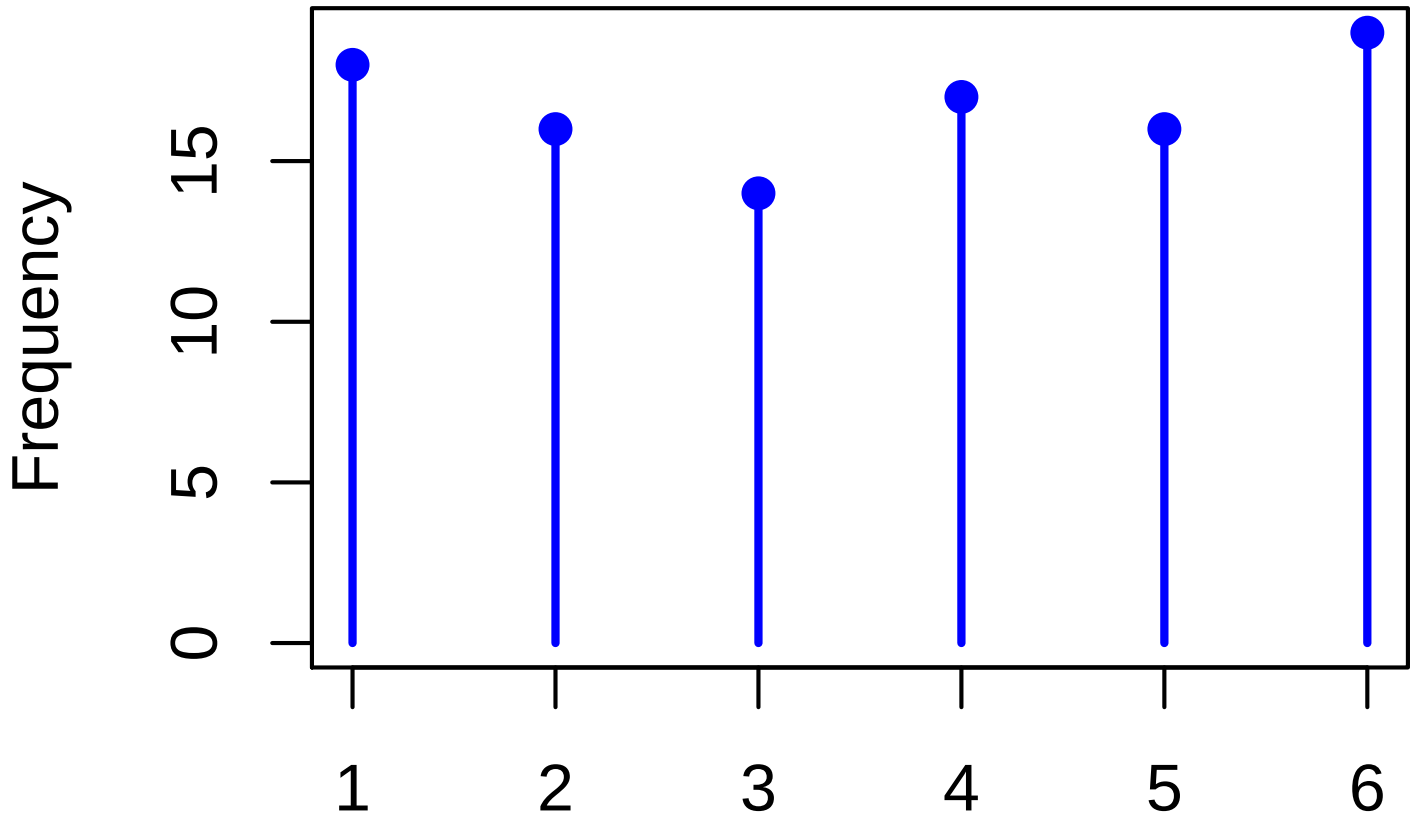
Six simulations

```
die=c(1,2,3,4,5,6);

sample(die,1,replace=TRUE);

## [1] 2

sample(die,6,replace=TRUE);

## [1] 5 3 5 1 1 6
```

60 simulations



100 simulations



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

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

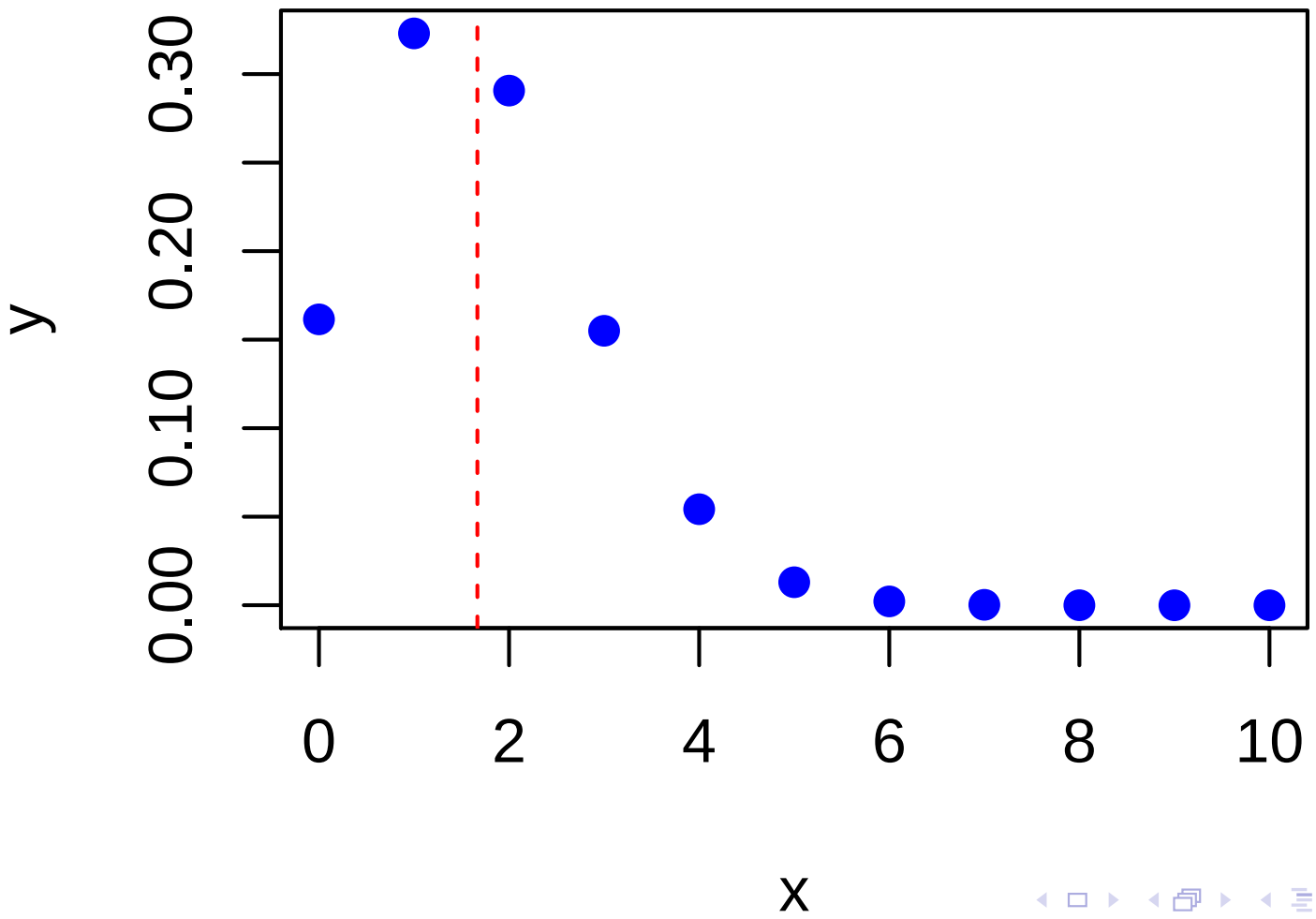
A few simulations

```
## Simulation: Binomial with n=10 and p=1/6.

rbinom(1,size=10,prob=1/6);

## [1] 2

rbinom(1,size=10,prob=1/6);

## [1] 3

rbinom(1,size=10,prob=1/6);

## [1] 2
```

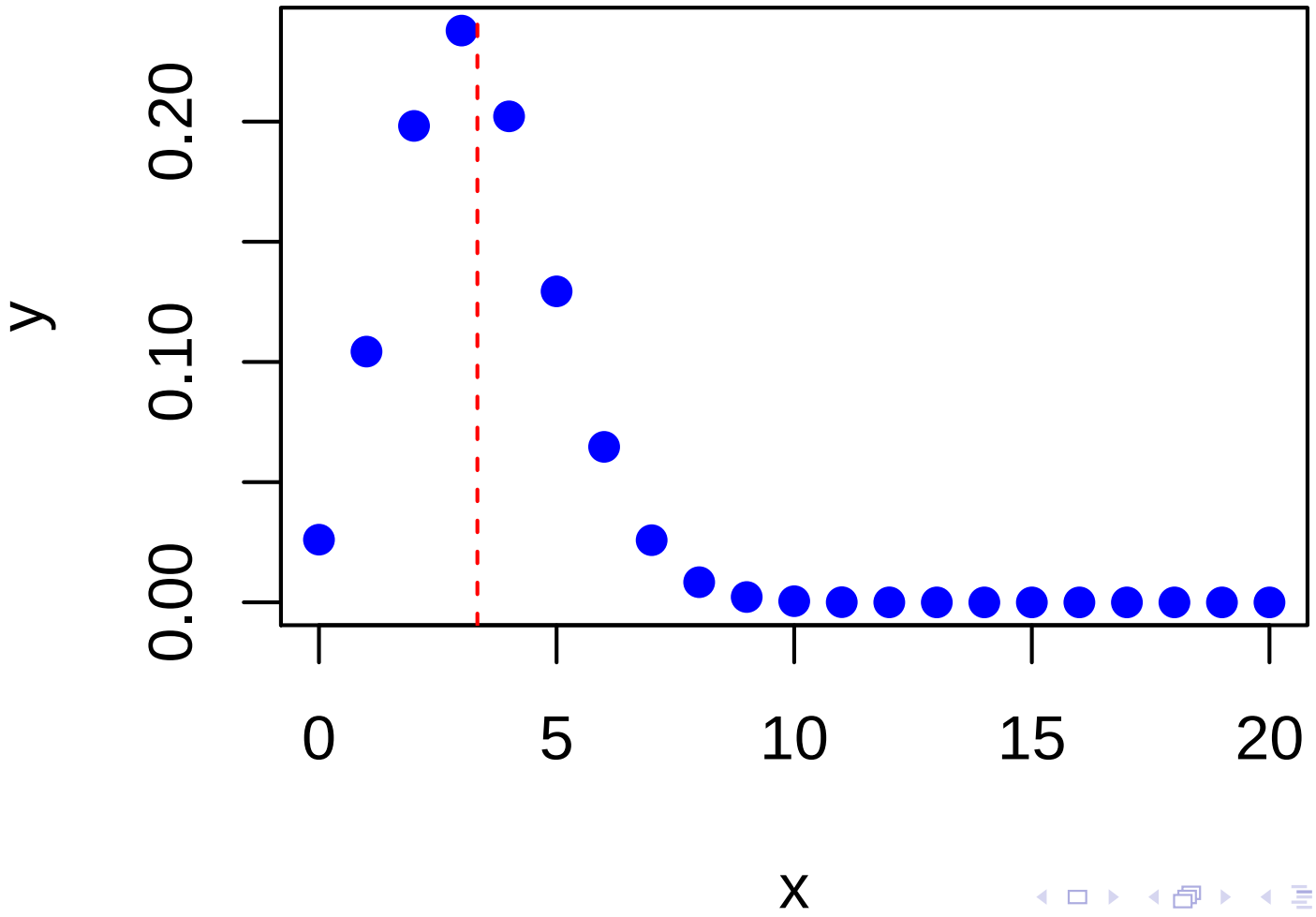
Probability Mass Function when n=10 and p=1/6

```
## Pmf: 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);
```

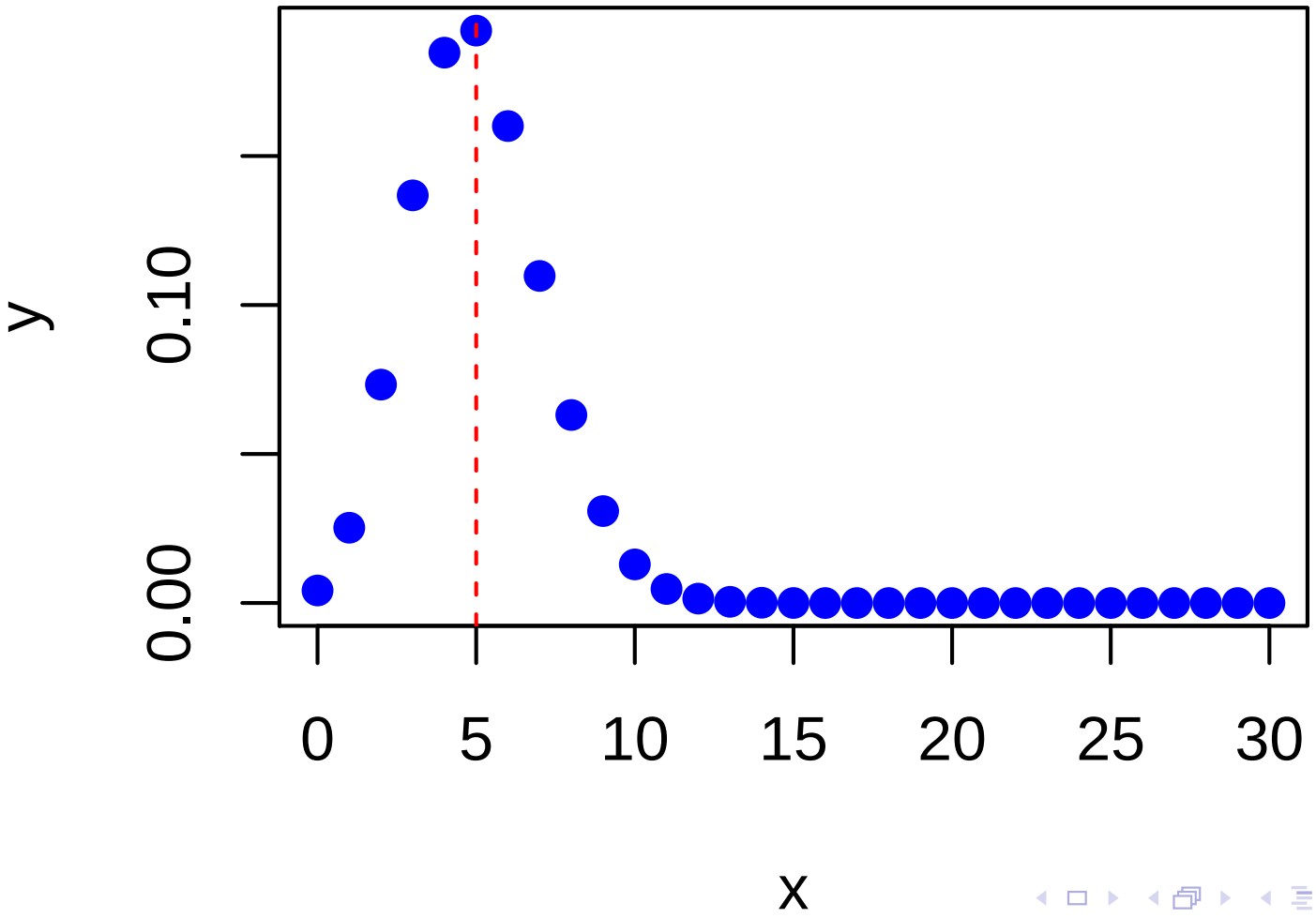
Probability Mass Function when n=10 and p=1/6



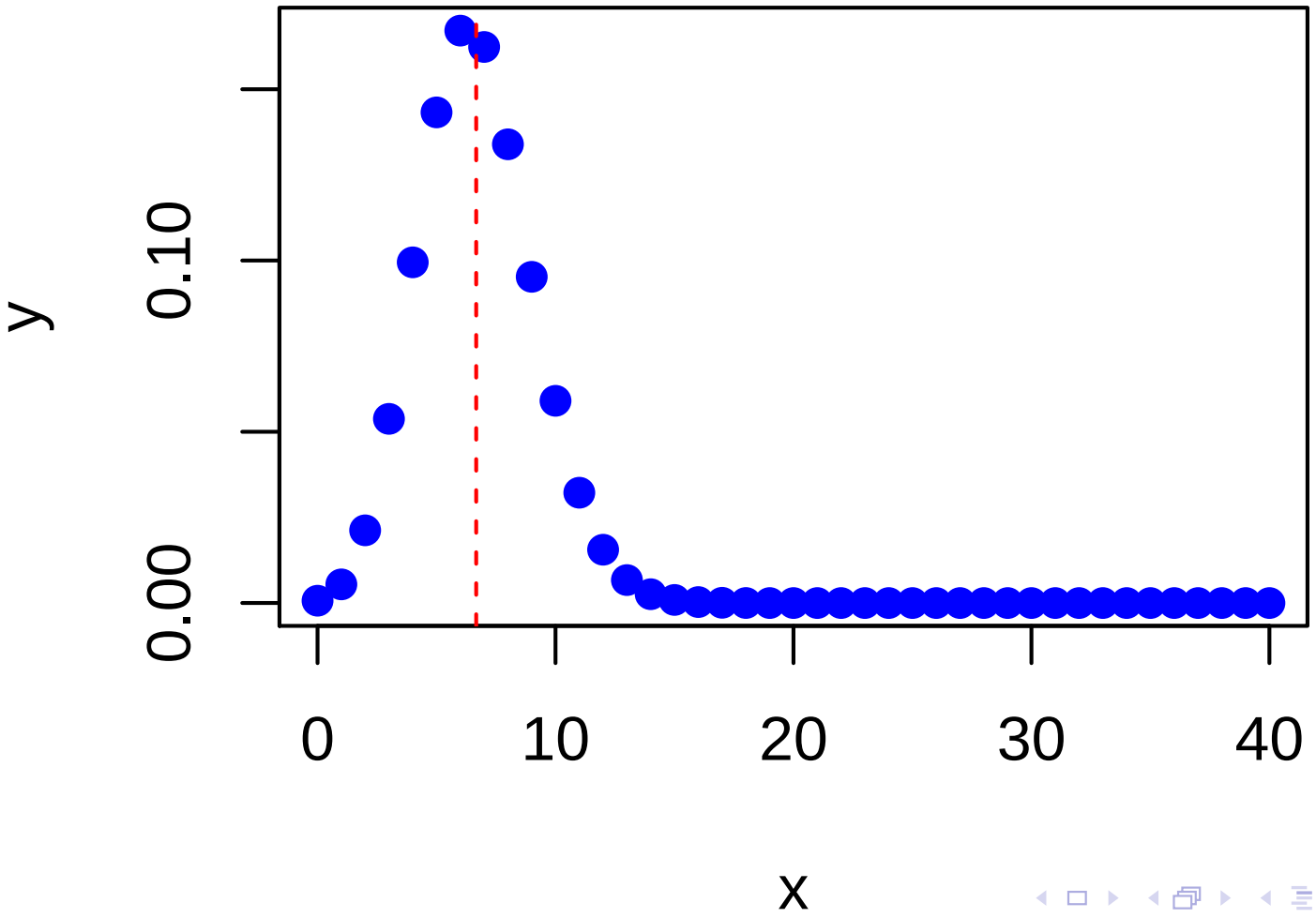
Pmf when n=20 and p=1/6



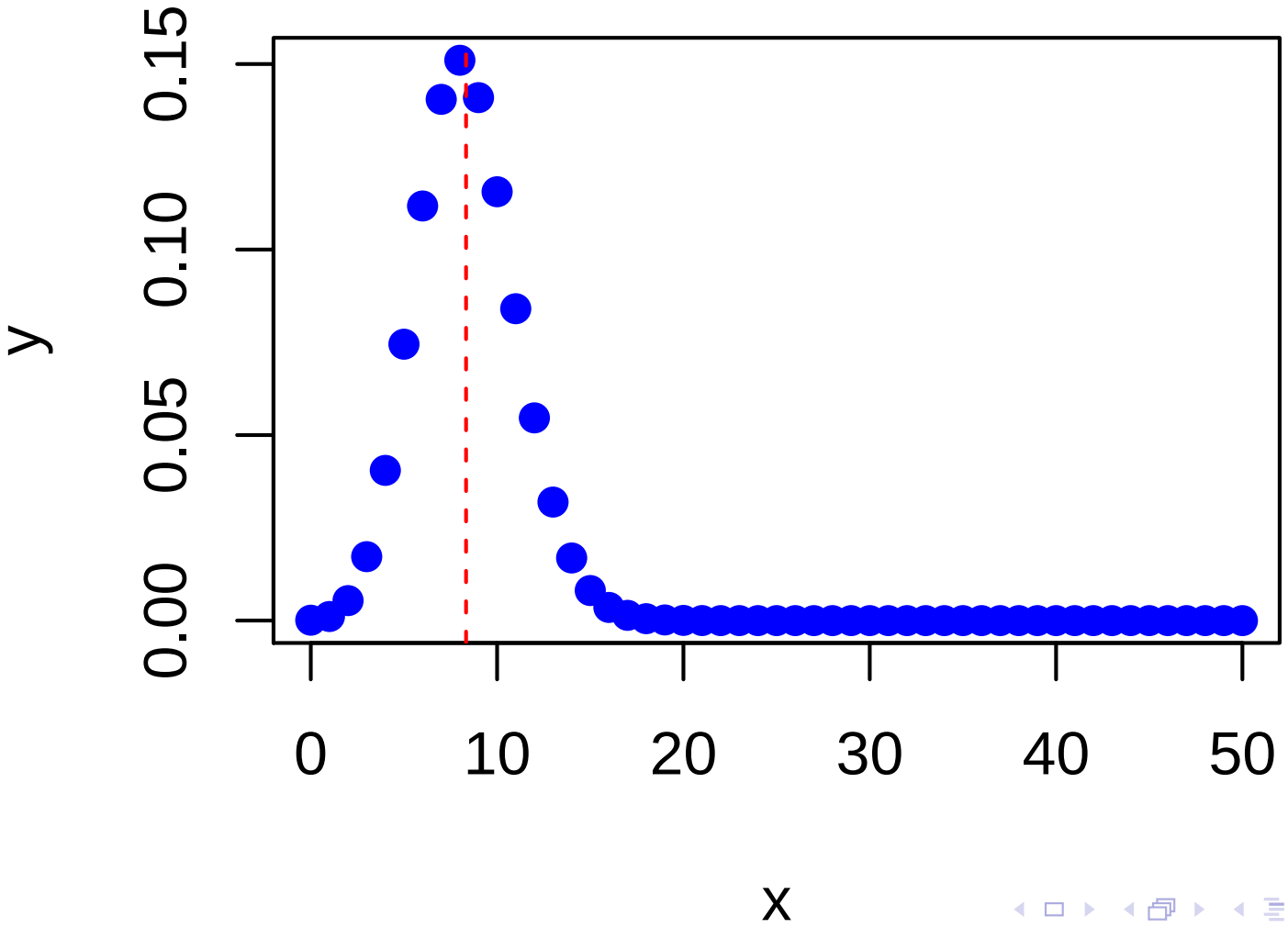
Pmf when n=30 and p=1/6



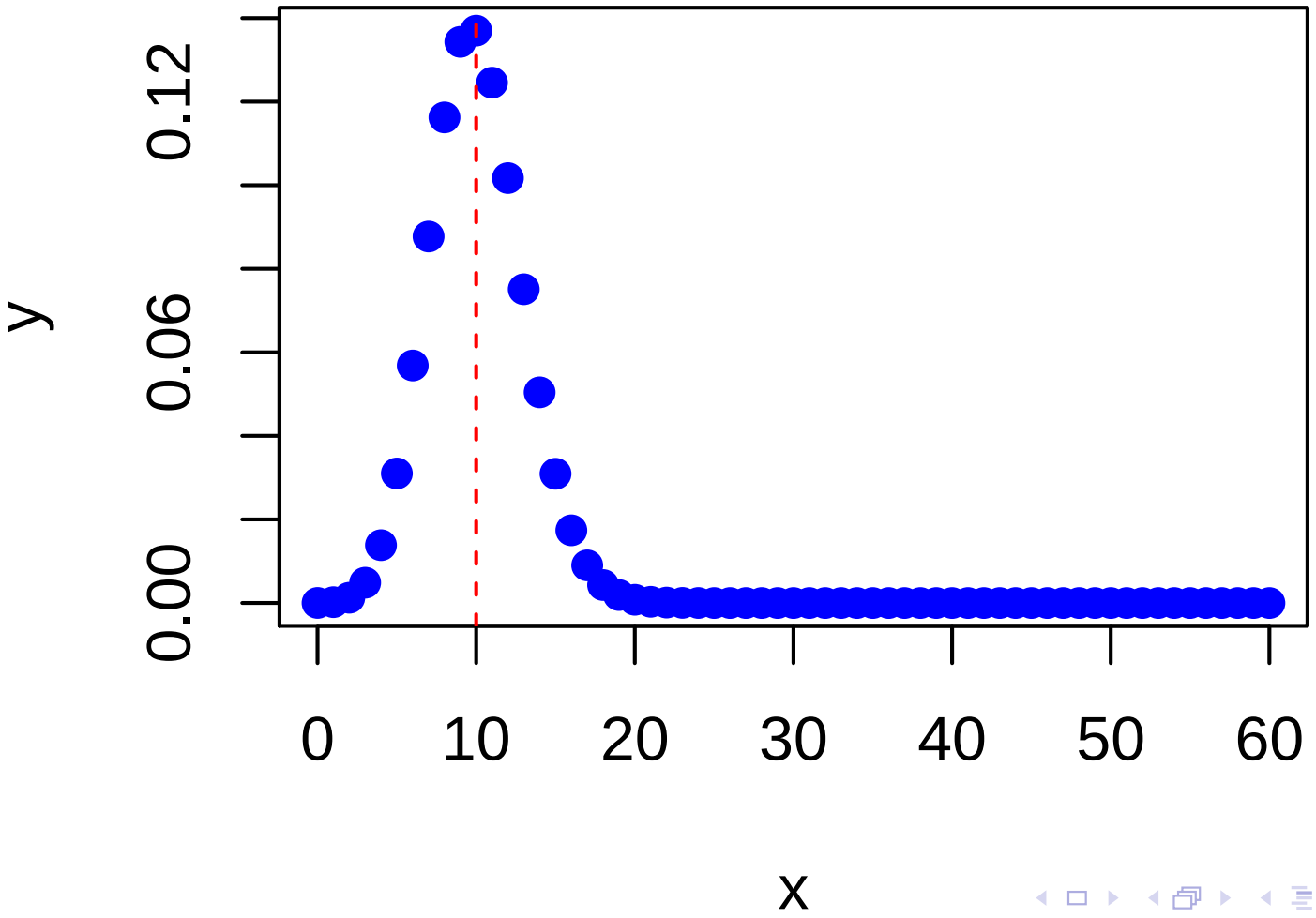
Pmf when n=40 and p=1/6



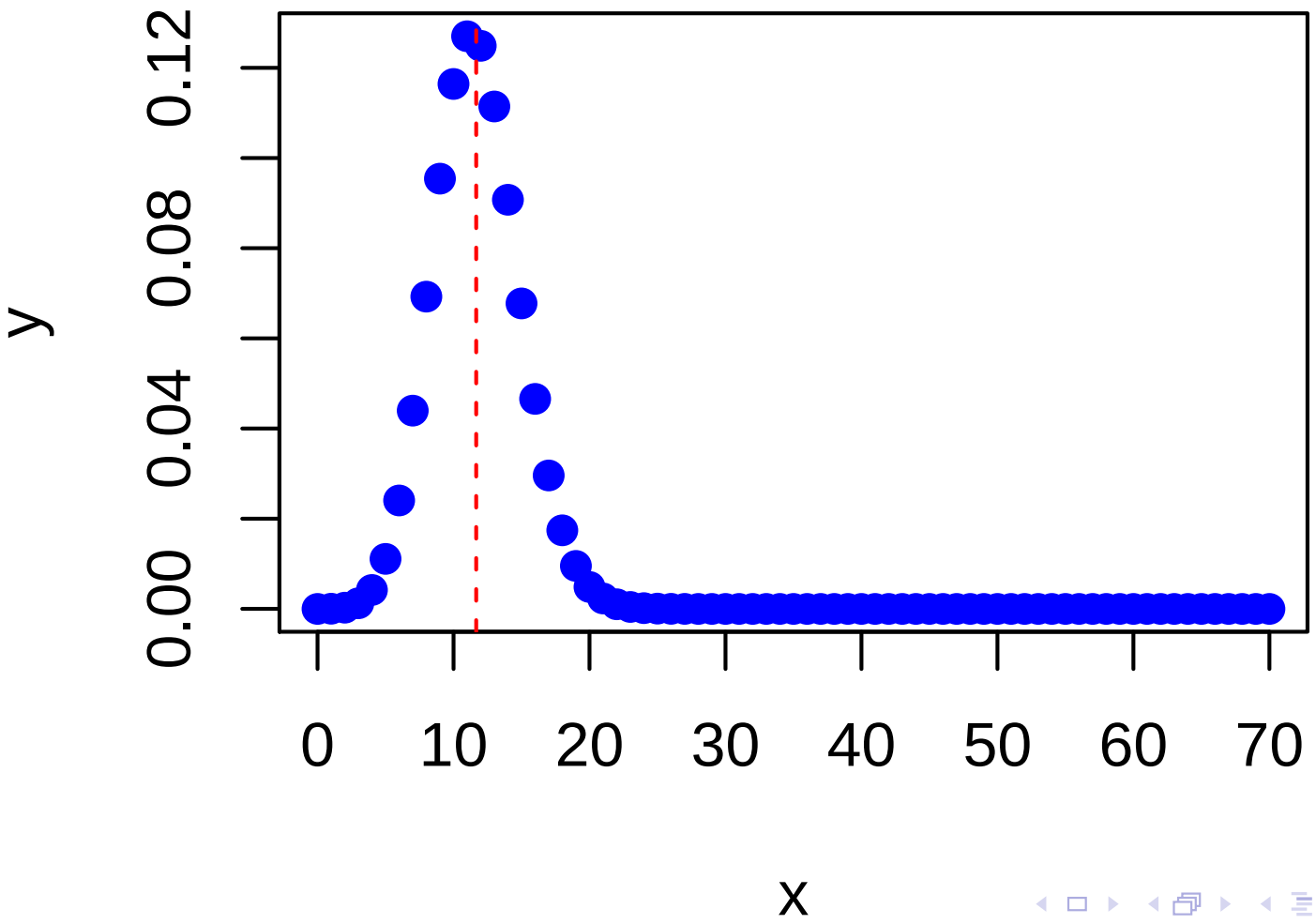
Pmf when n=50 and p=1/6



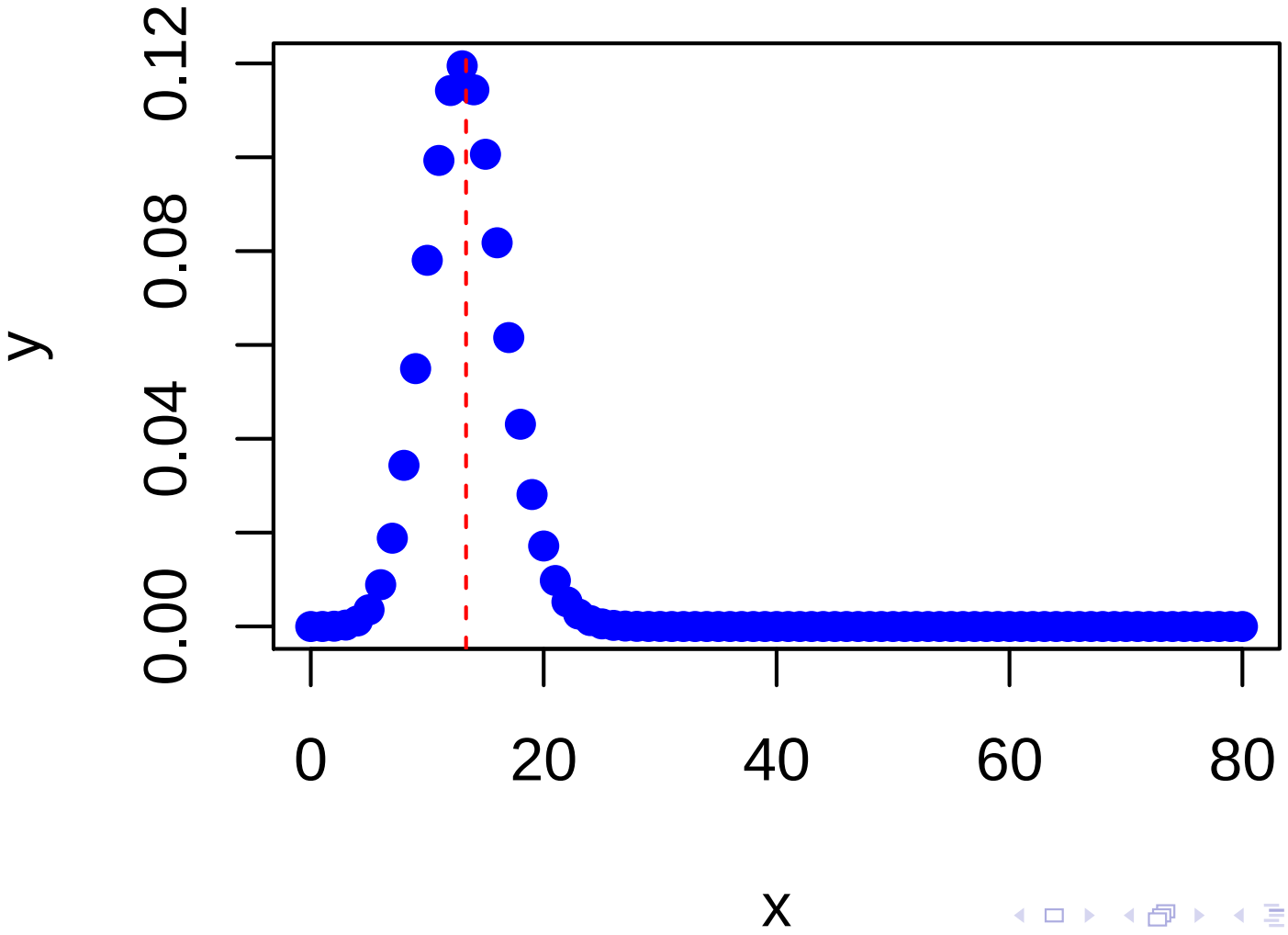
Pmf when n=60 and p=1/6



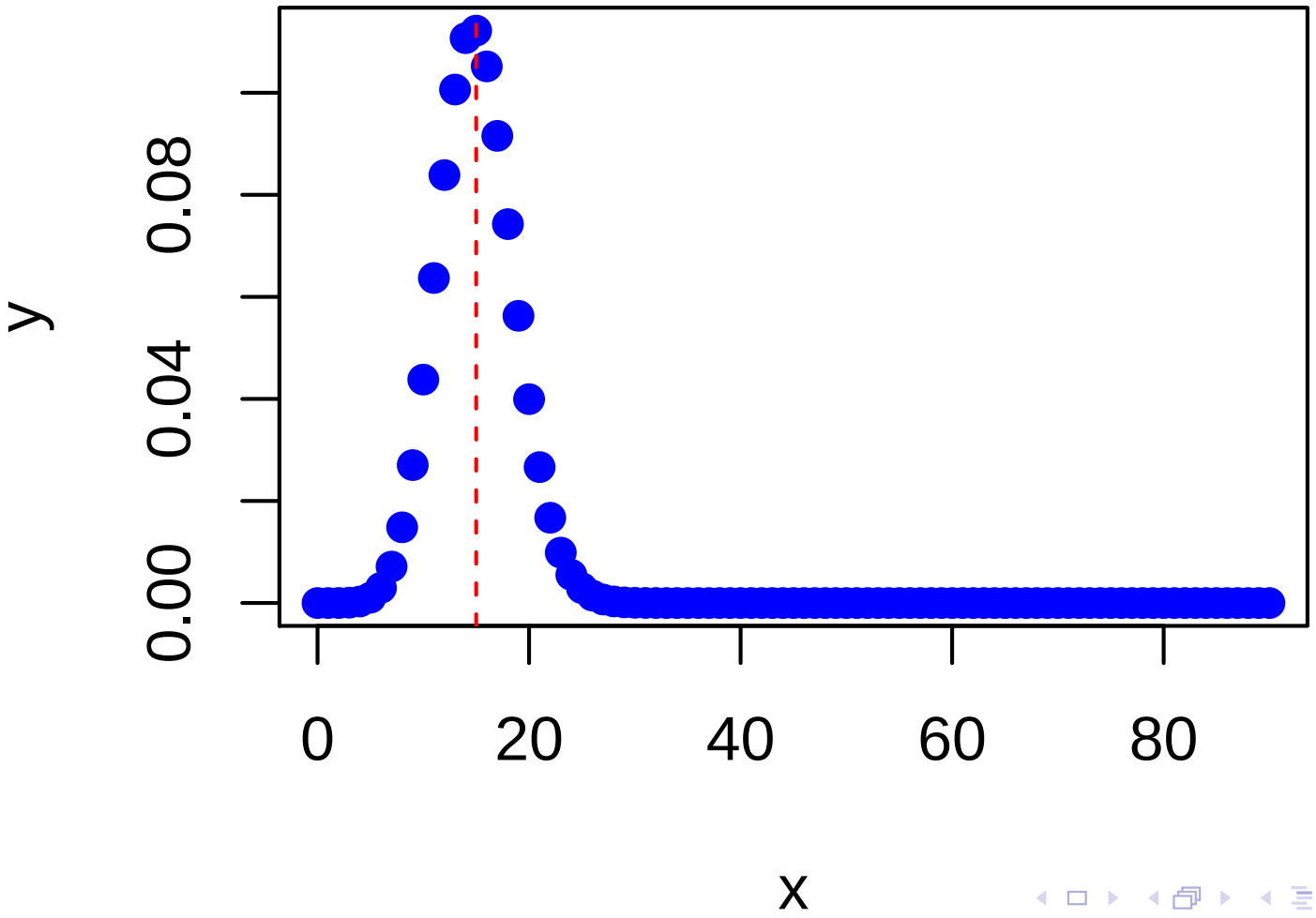
Pmf when n=70 and p=1/6



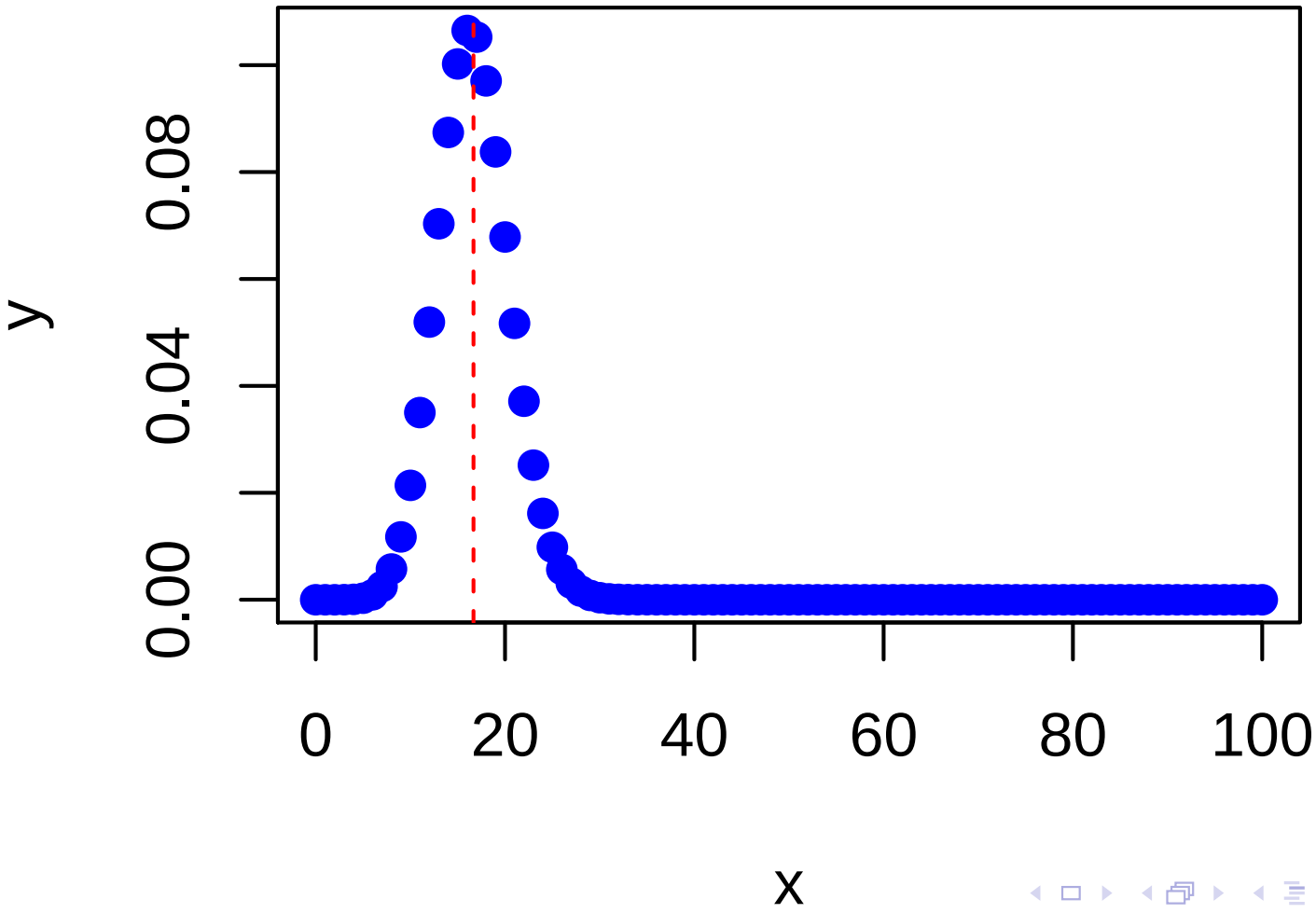
Pmf when n=80 and p=1/6



Pmf n=90 and p=1/6



Pmf when n=100 and p=1/6

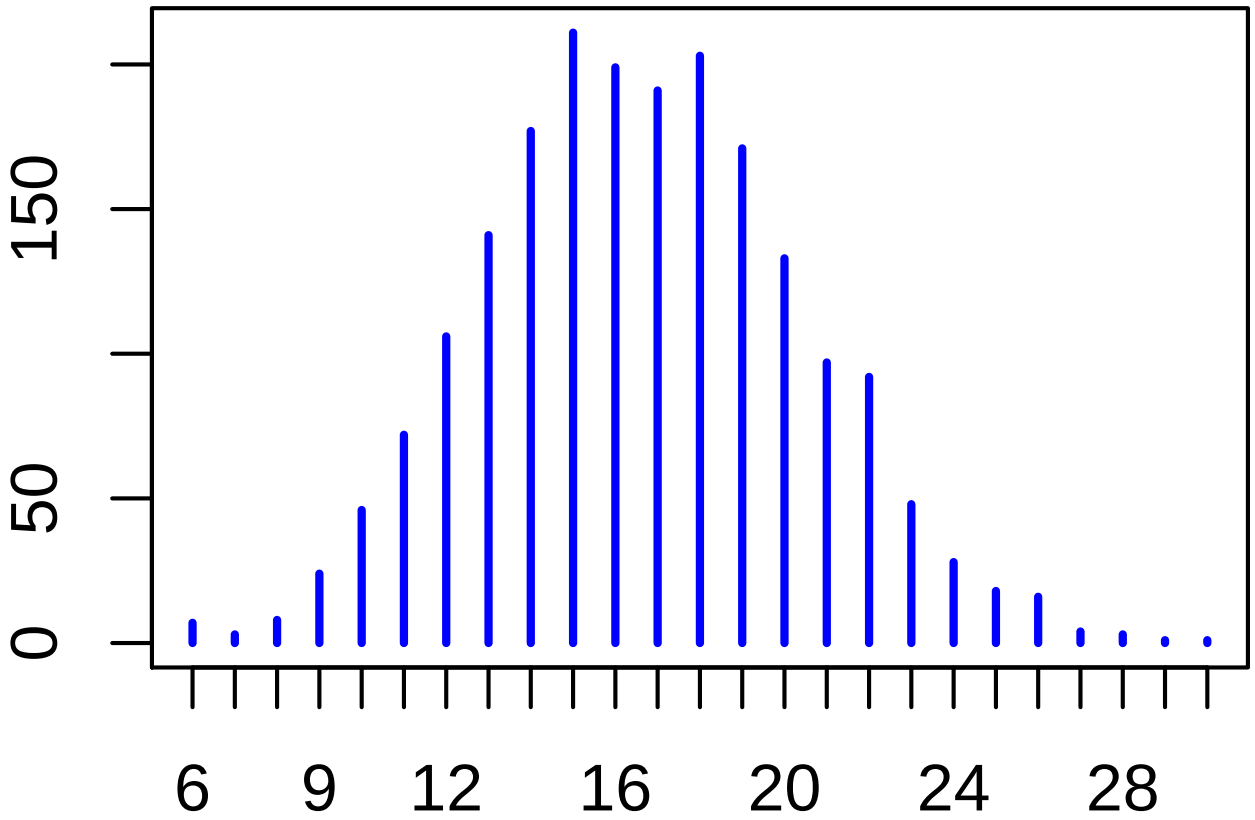


A few values from our pmf (n=100 and p=1/6)

```
dbinom(c(15,16,17,18),size=100,prob=1/6);
```

```
## [1] 0.10023663 0.10650142 0.10524847 0.09706247
```

Simulation: 2000 YouTubers, n=100, and p=1/6



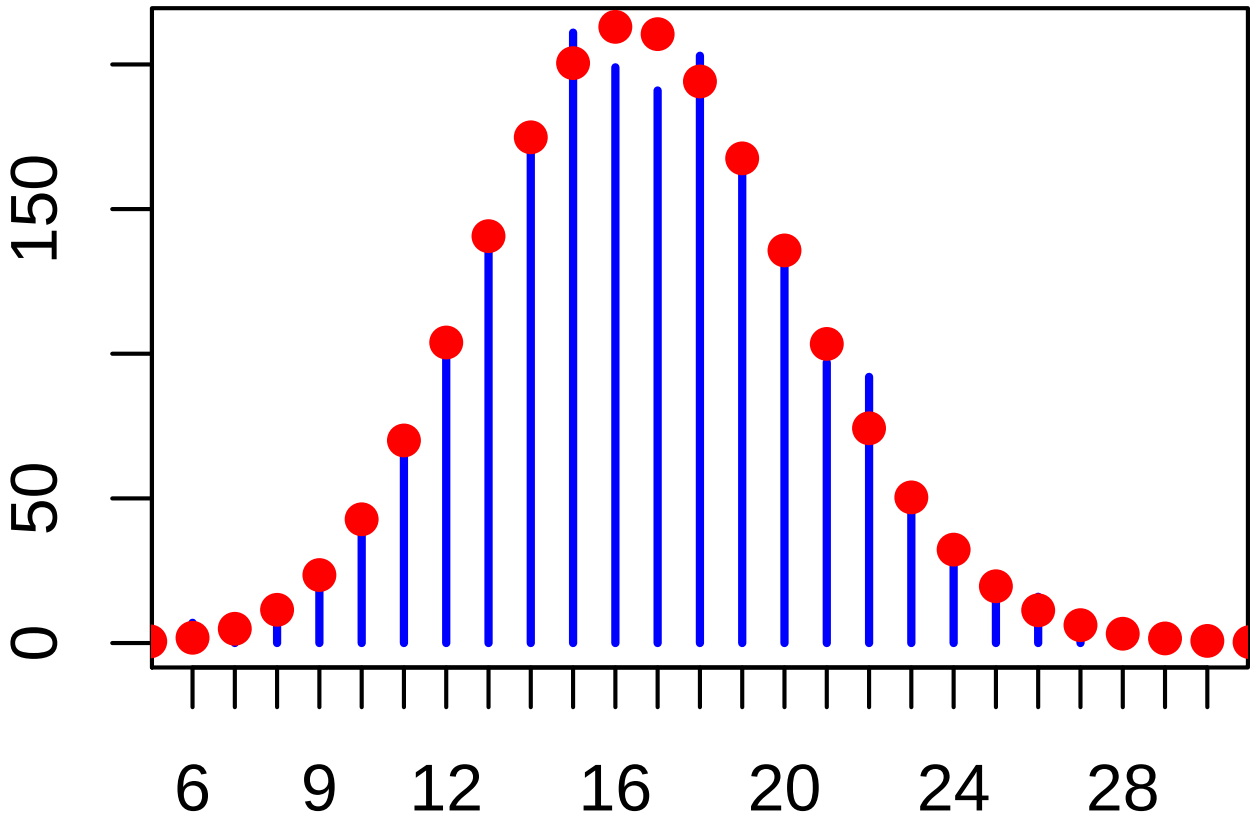
A few values from our simulation

```
## vec.prop
##   6   7   8   9  10  11  12
##   7   3   8  24  46  72 106
## [1] 266
## [1] 0.133
```

P-value

It turns out that our P-value for this simulation is:
0.133

Simulation vs Theoretical pmf



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:

- 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

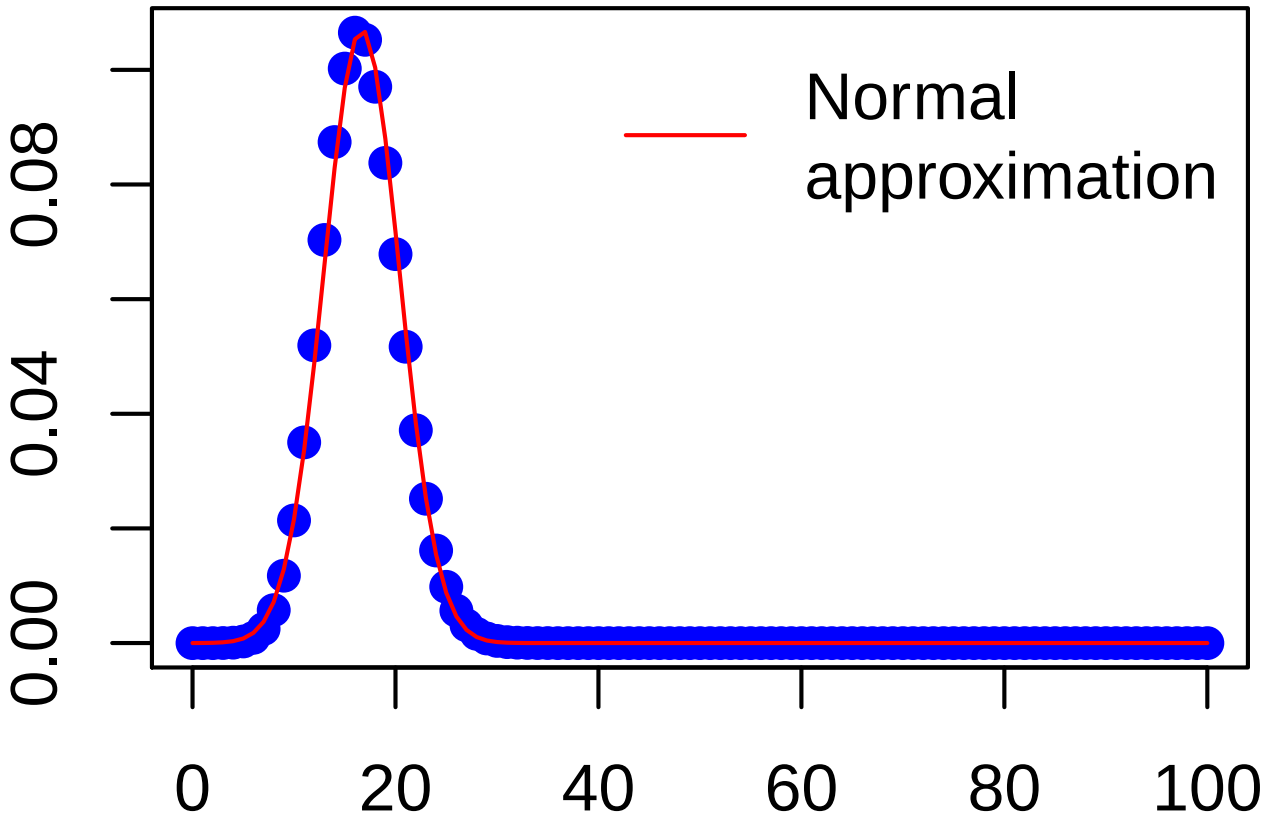
Draw an SRS of size n from a large population that contains proportion p of "successes". Let $\hat{p}$ be the **sample proportion** of successes,

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

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



Hypotheses Tests for a Proportion

To test the hypothesis $H_0 : p = p_0$, compute the z_* statistic,
$z_* = \frac{\hat{p} - p_0}{\sqrt{\frac{p_0(1-p_0)}{n}}}$
In terms of a variable Z having the standard Normal distribution, the approximate P-value for a test of H_0 against
$H_a : p > p_0 :$ is : $P(Z > z_*)$
$H_a : p < p_0 :$ is : $P(Z < z_*)$
$H_a : p \neq p_0 :$ is : $2P(Z > |z_*|)$

Solution

Step 1. $H_0 : p = \frac{1}{6}$ vs $H_a : p < \frac{1}{6}$

$\hat{p} = \frac{12}{100} = 0.12$

Step 2. (Without continuity correction)

$z_* = \frac{\hat{p} - p_0}{\sqrt{\frac{p_0(1-p_0)}{n}}} = \frac{0.12 - 0.1667}{\sqrt{\frac{(0.1667)(0.8333)}{100}}} = -1.2521$

(WITH continuity correction)

$P[X \leq 12] \approx P[X \leq 12.5] = P[\frac{X}{n} \leq 0.125] =$

$P\left[\frac{\bar{X} - \mu}{\sigma/n} \leq \frac{0.125 - 0.1667}{0.0373}\right]$

$= P[Z \leq -1.1179]$

Step 3. Using Normal table, P-value $=$

$P(Z < z_*) = P(Z < -1.1179) \approx 0.1314$

P-value is not small enough to provide evidence against H_0, we can't reject H_0. We conclude that there is not evidence to claim that probability of winning a food prize is less than $\frac{1}{6}$.

R Code

```
prop.test(12,100,p=1/6,alternative="less");

##
##  1-sample proportions test with continuity correction
##
## data:  12 out of 100, null probability 1/6
## X-squared = 1.25, df = 1, p-value = 0.1318
## alternative hypothesis: true p is less than 0.1666667
## 95 percent confidence interval:
##  0.0000000 0.1894571
## sample estimates:
##    p
## 0.12
```