STA218
Inference about a Population

Al Nosedal.
University of Toronto.
Fall 2018

November 20, 2018

Conditions for Inference about mean

- We can regard our data as a *simple random sample* (SRS) from the population. This condition is very important.
- Observations from the population have a *Normal distribution* with mean μ and standard deviation σ. In practice, it is enough that the distribution be symmetric and single-peaked unless the sample is very small. Both μ and σ are unknown parameters.

Standard Error

When the standard deviation of a statistic is estimated from data, the result is called the *standard error* of the statistic. The standard error of the sample mean $\bar{x}$ is $\frac{s}{\sqrt{n}}$.

Travel time to work

A study of commuting times reports the travel times to work of a random sample of 1000 employed adults. The mean is $\bar{x} = 49.2$ minutes and the standard deviation is $s = 63.9$ minutes. What is the standard error of the mean?

Solution

The standard error of the mean is

$$\frac{s}{\sqrt{n}} = \frac{63.9}{\sqrt{10000}} = 2.0207 \text{ minutes}$$

The one-sample t statistic and the t distributions

Draw an SRS of size n from a large population that has the Normal distribution with mean μ and standard deviation σ. The *one-sample t statistic*

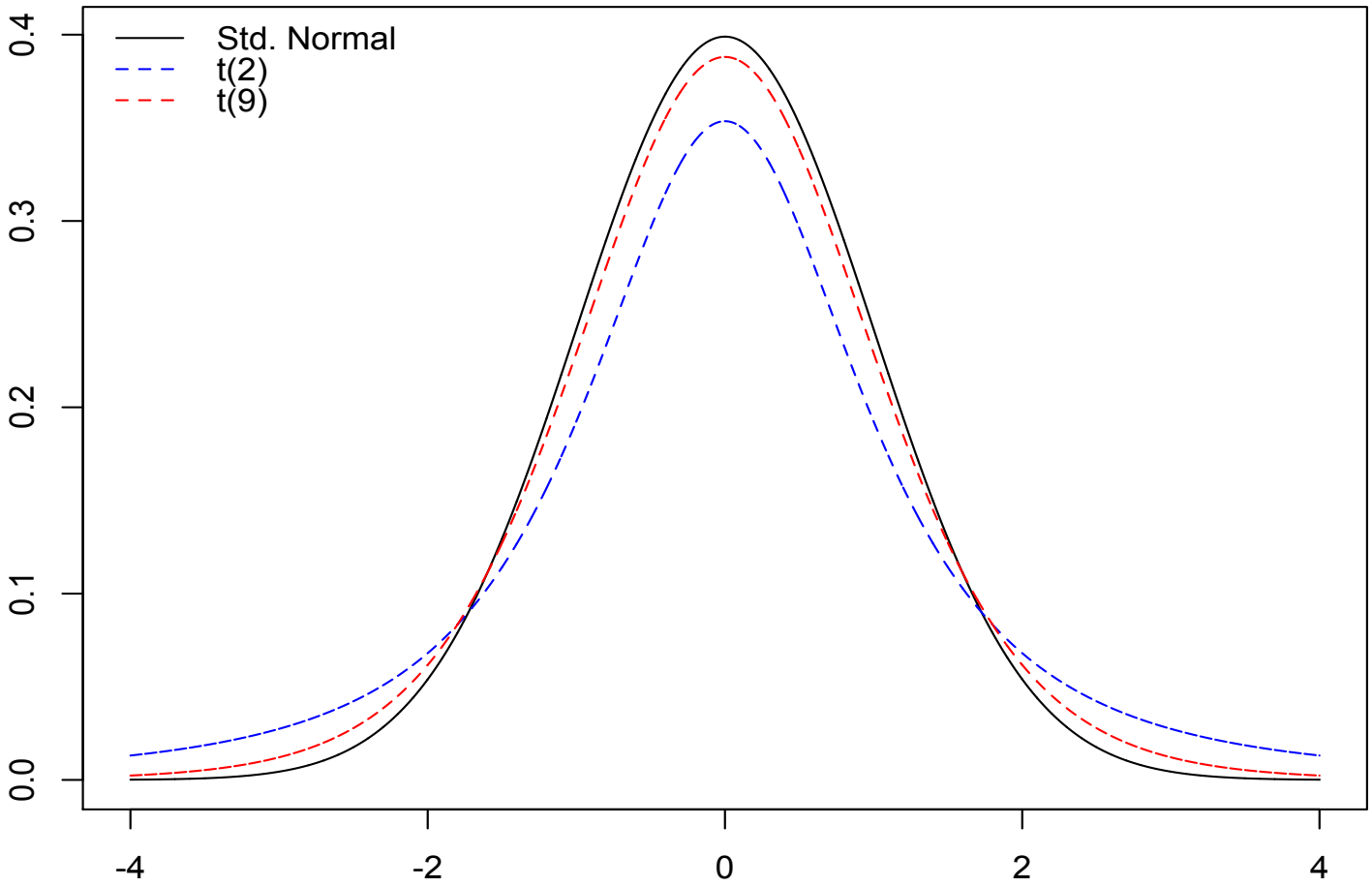
$$t^* = \frac{\bar{x} - \mu}{s/\sqrt{n}}$$

has the *t distribution* with $n - 1$ degrees of freedom.

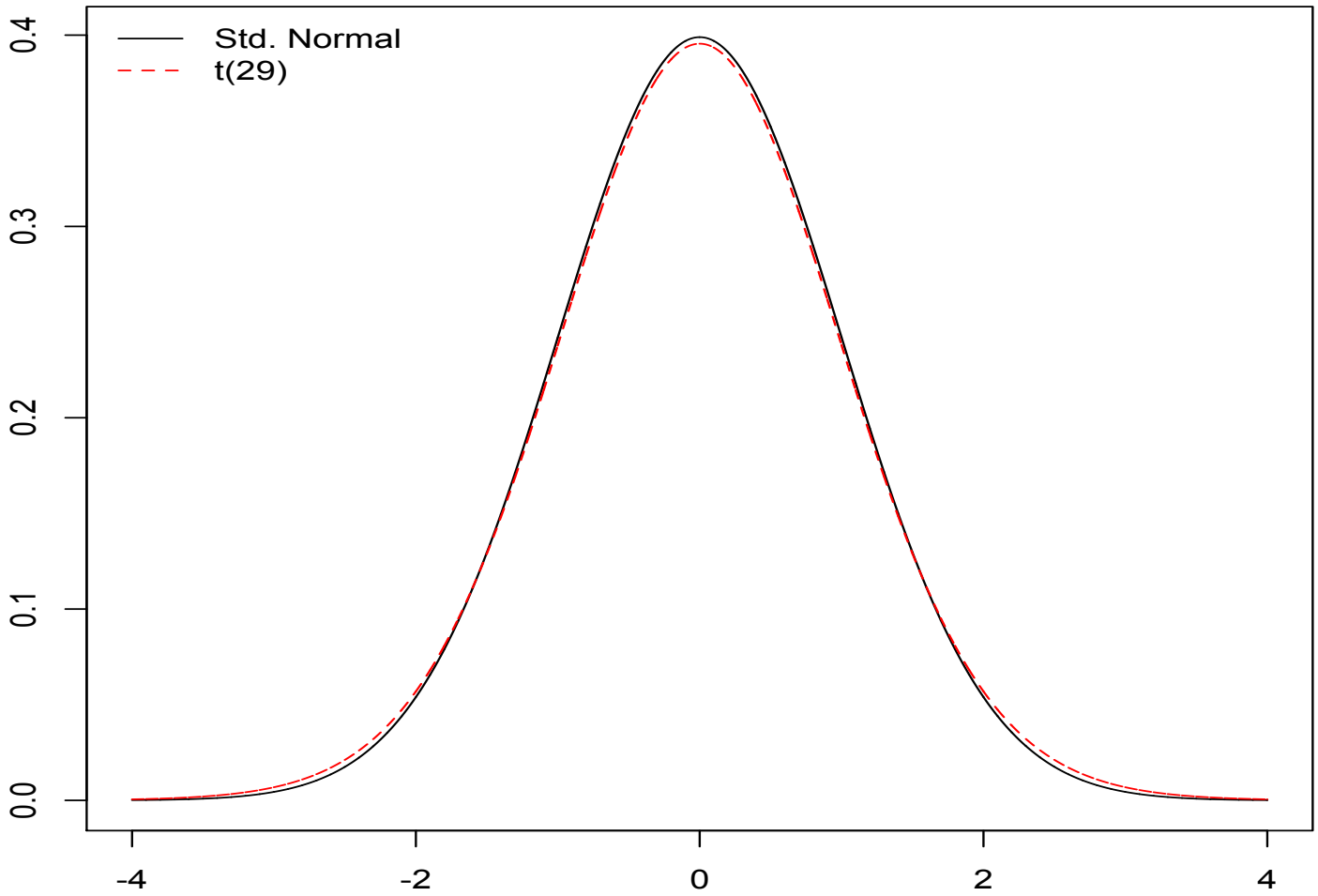
The t distributions

- The density curves of the t distributions are similar in shape to the Standard Normal curve. They are symmetric about 0, single-peaked, and bell-shaped.
- The spread of the t distributions is a bit greater than of the Standard Normal distribution. The t distributions have more probability in the tails and less in the center than does the Standard Normal. This is true because substituting the estimate s for the fixed parameter σ introduces more variation into the statistic.
- As the degrees of freedom increase, the t density curve approaches the $N(0, 1)$ curve ever more closely. This happens because s estimates σ more accurately as the sample size increases. So using s in place of σ causes little extra variation when the sample is large.

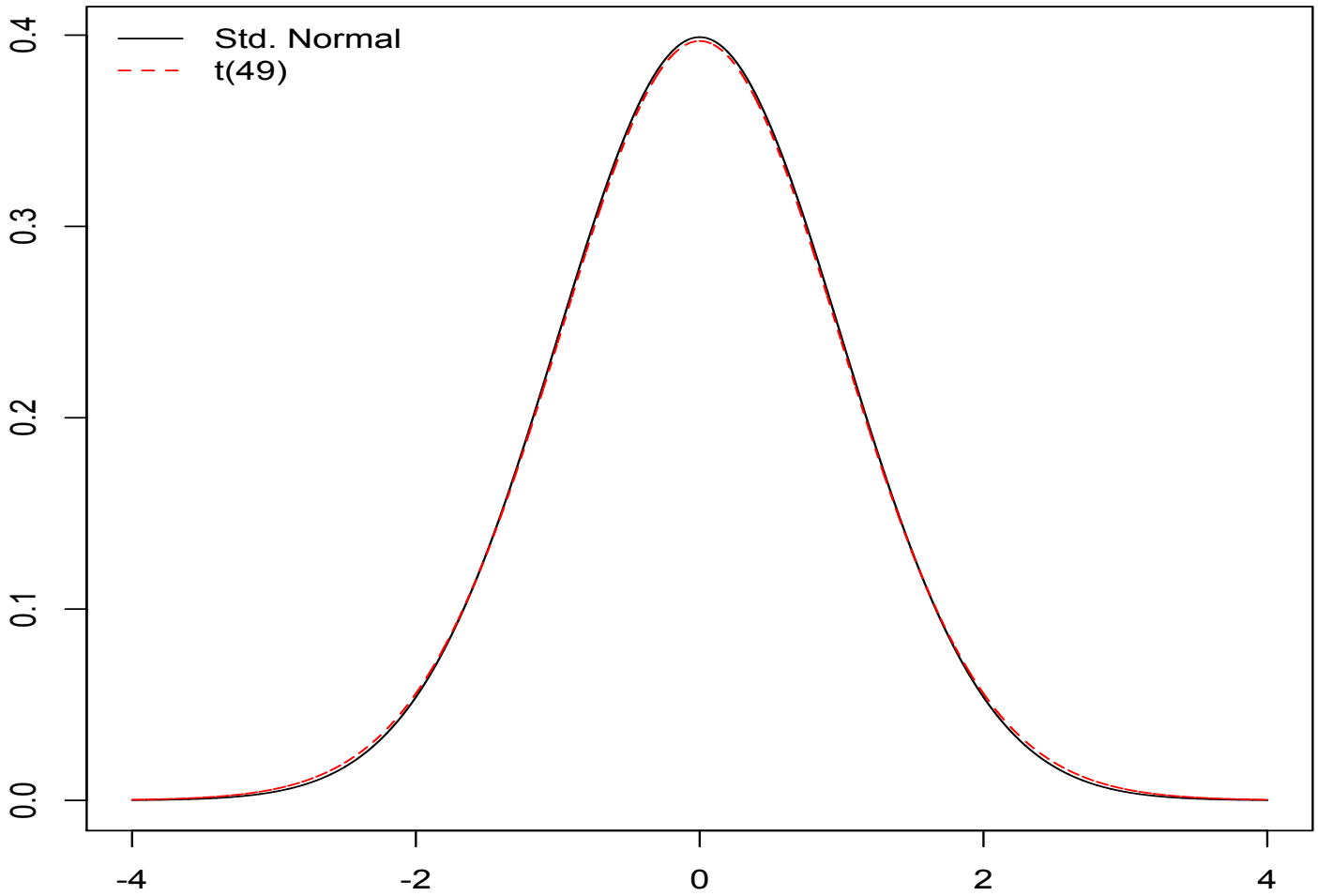
Density curves



Density curves



Density curves



Critical values

Use Table C or software to find
a) the critical value for a one-sided test with level $\alpha = 0.05$ based on the $t(4)$ distribution.
b) the critical value for 98% confidence interval based on the $t(26)$ distribution.

Solution

a) $t^* = 2.132$
b) $t^* = 2.479$

More critical values

You have an SRS of size 30 and calculate the one-sample t-statistic. What is the critical value t^* such that

a) t has probability 0.025 to the right of t^*?

b) t has probability 0.75 to the left of of t^*?

Solution

Here, $\text{df} = 30 - 1 = 29$.

a) $t^* = 2.045$

b) $t^* = 0.683$

The one-sample t confidence interval

Draw an SRS of size n from a large population having unknown mean μ. A level C *confidence interval for* μ is

$$\bar{x} \pm t^* \frac{s}{\sqrt{n}}$$

where t^* is the critical value for the $t(n-1)$ density curve with area C between $-t^*$ and t^*. This interval is exact when the population distribution is Normal and is approximately correct for large n in other cases.

Critical values

What critical value t^* from Table C would you use for a confidence interval for the mean of the population in each of the following situations?

a) A 95% confidence interval based on $n = 12$ observations.

b) A 99% confidence interval from an SRS of 18 observations.

c) A 90% confidence interval from a sample of size 6.

Solution

a) $df = 12 - 1 = 11$, so $t^* = 2.201$.
b) $df = 18 - 1 = 17$, so $t^* = 2.898$.
c) $df = 6 - 1 = 5$, so $t^* = 2.015$.

Example

The following sample data are from a normal population: 10, 8, 12, 15, 13, 11, 6, 5.

a. What is the point estimate of the population mean?

b. What is the point estimate of the population standard deviation?

c. With 95 % confidence, what is the margin of error for the estimation of the population mean?

d. What is the 95 % confidence interval for the population mean?

Solution

a. $\bar{x} = 10$.

b. $s = 3.4641$.

c. margin of error $= t_* \frac{s}{\sqrt{n}} = 2.365(\frac{3.4641}{\sqrt{8}}) = 2.8965$.

d. $(\bar{x} - t_*(\frac{s}{\sqrt{n}}), \bar{x} + t_*(\frac{s}{\sqrt{n}}))$

$(7.1039, 12.896)$.

R Code

```r
# Step 1. Entering data;

dataset=c(10, 8, 12, 15, 13, 11, 6, 5);

# Step 2. Finding CI;

t.test(dataset);
```

R Code

```
##
##  One Sample t-test
##
## data:  dataset
## t = 8.165, df = 7, p-value = 7.999e-05
## alternative hypothesis: true mean is not equal to 0
## 95 percent confidence interval:
##   7.103939 12.896061
## sample estimates:
## mean of x
##        10
```

Example

A simple random sample with $n = 54$ provided a sample mean of 22.5 and a sample standard deviation of 4.4.

a. Develop a 90% confidence interval for the population mean.

b. Develop a 95% confidence interval for the population mean.

c. Develop a 99% confidence interval for the population mean.

Solution

a. $(\bar{x} - t_*(\frac{s}{\sqrt{n}}), \bar{x} + t_*(\frac{s}{\sqrt{n}}))$

$(22.5 - 1.676(\frac{4.4}{\sqrt{54}}), 22.5 + 1.676(\frac{4.4}{\sqrt{54}}))$

$(21.496, 23.503)$.

b. $(22.5 - 2.009(\frac{4.4}{\sqrt{54}}), 22.5 + 2.009(\frac{4.4}{\sqrt{54}}))$

$(21.297, 23.703)$.

c. $(22.5 - 2.678(\frac{4.4}{\sqrt{54}}), 22.5 + 2.678(\frac{4.4}{\sqrt{54}}))$

$(20.896, 24.103)$.

Example

Sales personnel for Skillings Distributors submit weekly reports listing the customer contacts made during the week. A sample of 65 weekly reports showed a sample mean of 19.5 customer contacts per week. The sample standard deviation was 5.2. Provide a 90% and 95% confidence intervals for the population mean number of weekly customer contacts for the sales personnel.

Solution

90 % Confidence.

$(\bar{x} - t_*(\frac{s}{\sqrt{n}}), \bar{x} + t_*(\frac{s}{\sqrt{n}}))$

$(19.5 - 1.671(\frac{5.2}{\sqrt{65}}), 19.5 + 1.671(\frac{5.2}{\sqrt{65}}))$

$(18.42, 20.58)$

95 % Confidence.

$(\bar{x} - t_*(\frac{s}{\sqrt{n}}), \bar{x} + t_*(\frac{s}{\sqrt{n}}))$

$(19.5 - 2(\frac{5.2}{\sqrt{65}}), 19.5 + 2(\frac{5.2}{\sqrt{65}}))$

$(18.21, 20.79)$

Exercise

Most owners of digital cameras store their pictures on the camera. Some will eventually download these to a computer or print them using their own printers or a commercial printer. A film-processing company wanted to know how many pictures were stored on computers. A random sample of 10 digital camera owners produced the data given here. Estimate with 95% confidence the mean number of pictures stored on digital cameras.
25 6 22 26 31 18 13 20 14 2.

Solution

μ = mean number of pictures stored on digital cameras. We will estimate μ with a 95% confidence interval.
$\bar{x} = 17.7$, $s = 9.080504$, and $df = 10 - 1 = 9$. From our table, $t^* = 2.262$ (or $t(9) = 2.262$).
margin of error $= 2.262\left(\frac{9.080504}{\sqrt{10}}\right) = 6.49535$.
$LCL = \bar{x} -$ margin of error $= 17.7 - 6.49535 = 11.20465$
$UCL = \bar{x} -$ margin of error $= 17.7 + 6.49535 = 24.19535$

Solution

95 percent confidence interval:

11.20 24.19

R Code

```r
# Step 1. Entering data;

dataset=c(25,6,22,26,31,18,13,20,14,2);

# Step 2. Finding CI;

t.test(dataset, conf.level = 0.95);
```

R Code

```
##
##  One Sample t-test
##
## data:  dataset
## t = 6.164, df = 9, p-value = 0.0001659
## alternative hypothesis: true mean is not equal to 0
## 95 percent confidence interval:
##  11.2042 24.1958
## sample estimates:
## mean of x
##      17.7
```

Ancient air

The composition of the earth's atmosphere may have changed over time. To try to discover the nature of the atmosphere long ago, we can examine the gas in bubbles inside ancient amber. Amber is tree resin that has hardened and been trapped in rocks. The gas in bubbles within amber should be a sample of the atmosphere at the time the amber was formed. Measurements on specimens of amber from the late Cretaceous era (75 to 95 million years ago) give these percents of nitrogen:

63.4 65 64.4 63.3 54.8 64.5 60.8 49.1 51.0

Assume (this is not yet agreed on by experts) that these observations are an SRS from the late Cretaceous atmosphere. Use a 90% confidence interval to estimate the mean percent of nitrogen in ancient air (Our present-day atmosphere is about 78.1% nitrogen).

Solution

μ = mean percent of nitrogen in ancient air. We will estimate μ with a 90% confidence interval.
With $\bar{x} = 59.5888$, $s = 6.2552$, and $t^* = 1.860$ ($df = 9 - 1 = 8$), the 90% confidence interval for μ is

$$59.5888 \pm 1.860 \left(\frac{6.2552}{\sqrt{9}} \right)$$

$$59.5888 \pm 3.8782$$

55.7106 to 63.4670

R Code

```
# Step 1. Entering data;

dataset=c(63.4, 65,64.4,63.3,54.8,64.5,60.8,49.1,51.0);

# Step 2. Finding CI;

t.test(dataset, conf.level = 0.90);
```

R Code

```
##
##  One Sample t-test
##
## data:  dataset
## t = 28.578, df = 8, p-value = 2.43e-09
## alternative hypothesis: true mean is not equal to 0
## 90 percent confidence interval:
##  55.71155 63.46622
## sample estimates:
## mean of x
##  59.58889
```

The one-sample t test

Draw an SRS of size n from a large population having unknown mean μ. To *test the hypothesis* $H_0 : \mu = \mu_0$, compute the *one-sample t statistic*

$$t^* = \frac{\bar{x} - \mu_0}{s/\sqrt{n}}$$

In terms of a variable T having the $t(n-1)$ distribution, the P-value for a test of H_0 against
$H_a : \mu > \mu_0$ is $P(T \geq t^*)$.
$H_a : \mu < \mu_0$ is $P(T \leq t^*)$.
$H_a : \mu \neq \mu_0$ is $2P(T \geq |t^*|)$.
These P-values are exact if the population distribution is Normal and are approximately correct for large n in other cases.

Is it significant?

The one-sample t statistic for testing
$H_0 : \mu = 0$
$H_a : \mu > 0$
from a sample of $n = 20$ observations has the value $t^* = 1.84$.
a) What are the degrees of freedom for this statistic?
b) Give the two critical values t from Table C that bracket t^*. What are the one-sided P-values for these two entries?
c) Is the value $t^* = 1.84$ significant at the 5% level? Is it significant at the 1% level?
d) (Optional) If you have access to suitable technology, give the exact one-sided P-value for $t^* = 1.84$?

Solution

a) df $=$ 20 -1 $=$ 19.
b) $t^* = 1.84$ is bracketed by $t = 1.729$ (with right-tail probability 0.05) and $t = 2.093$ (with right-tail probability 0.025). Hence, because this is a one-sided significance test,
$0.025 <$ P-value < 0.05.
c) This test is significant at the 5% level because the P-value < 0.05. It is not significant at the 1% level because the P-value > 0.01.
d) P-value $= 0.0407$.
(Type: `1-pt(1.84,df=19)` in R console).

d) Using R

```
1-pt(1.84,df=19)
```

```
## [1] 0.04072234
```

Is it significant?

The one-sample t statistic from a sample of $n = 15$ observations for the two-sided test of

$H_0 : \mu = 64$

$H_a : \mu \neq 64$

has the value $t^* = 2.12$.

a) What are the degrees of freedom for t^*?

b) Locate the two-critical values t from Table C that bracket t^*. What are the two-sided P-values for these two entries?

c) is the value $t^* = 2.12$ statistically significant at the 10% level? At the 5% level?

d) (Optional) If you have access to suitable technology, give the exact two-sided P-value for $t^* = 2.12$.

Solution

a) df $=$ 15 -1 $=$ 14.
b) $t^* = 2.12$ is bracketed by $t = 1.761$ (with two-tail probability 0.10) and $t = 2.145$ (with two-tail probability 0.05). Hence, because this is a two-sided significance test, $0.05 <$ P-value < 0.10.
c) This test is significant at the 10% level because the P-value < 0.10. It is not significant at the 5% level because the P-value > 0.05.
d) P-value $= 2(0.0261) = 0.0523$.
(Type: `2*(1-pt(2.12,df=14))` in R console).

d) Using R

```
2*(1-pt(2.12,df=14))
```

```
## [1] 0.05235683
```

Example

$H_0 : \mu = 12$
$H_a : \mu > 12$
A sample of 25 provided a sample mean $\bar{x} = 14$ and a sample standard deviation $s = 4.32$.

a. Compute the value of the test statistic.
b. Use the t distribution table to compute a range for the p-value.
c. At $\alpha = 0.05$, what is your conclusion?

Solution

a. $t_* = \frac{\bar{x}-\mu_0}{s/\sqrt{n}} = \frac{14-12}{4.32/\sqrt{25}} = 2.31$

b. Degrees of freedom $= n - 1 = 24$.

P-value $= P(T > t_*) = P(T > 2.31)$

Using t-table, P-value is between 0.01 and 0.02.

Exact P-value $= 0.0149$ (using R).

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

Example

$H_0 : \mu = 18$
$H_a : \mu \neq 18$
A sample of 48 provided a sample mean $\bar{x} = 17$ and a sample standard deviation $s = 4.5$.

a. Compute the value of the test statistic.
b. Use the t distribution table to compute a range for the p-value.
c. At $\alpha = 0.05$, what is your conclusion?

Solution

a. $t_* = \frac{\bar{x}-\mu_0}{s/\sqrt{n}} = \frac{17-18}{4.5/\sqrt{48}} = -1.54$

b. Degrees of freedom $= n - 1 = 47$.

P-value $= 2P(T > |t_*|) = 2P(T > |-1.54|) = 2P(T > 1.54)$

Using t-table, P-value is between 0.10 and 0.20.

Exact P-value $= 0.1303$ (using R).

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

Example

The Employment and Training Administration reported the U.S. mean unemployment insurance benefit of $ 238 per week. A researcher in the state of Virginia anticipated that sample data would show evidence that the mean weekly unemployment insurance benefit in Virginia was below the national level.

a. Develop appropriate hypotheses such that rejection of H_0 will support the researcher's contention.

b. For a sample of 100 individuals, the sample mean weekly unemployment insurance benefit was $231 with a sample standard deviation of $80. What is the p-value?

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

Solution

a. $H_0 : \mu = 238$ vs $H_a : \mu < 238$.

b. $t_* = \frac{\bar{x}-\mu_0}{s/\sqrt{n}} = \frac{231-238}{80/\sqrt{100}} = -0.88$

Degrees of freedom $= n - 1 = 99$.

Using t table, P-value is between 0.15 and 0.20

c. P-value > 0.05, we CAN'T reject H_0. Cannot conclude mean weekly benefit in Virginia is less than the national mean.

Example

The National Association of Professional Baseball Leagues, Inc., reported that attendance for 176 minor league baseball teams reached an all-time high during the 2001 season. On a per-game basis, the mean attendance for minor league baseball was 3530 people per game. Midway through the 2002 season, the president of the association asked for an attendance report that would hopefully show that the mean attendance for 2002 was exceeding the 2001 level.

Example (cont.)

a. Formulate hypotheses that could be used to determine whether the mean attendance per game in 2002 was greater than the previous year's level.
b. Assume that a sample of 92 minor league baseball games played during the first half of the 2002 season showed a mean attendance of 3740 people per game with a sample standard deviation of 810. What is the p-value?
c. At $\alpha = 0.01$, what is your conclusion?

Solution

a. $H_0 : \mu = 3530$ vs $H_a : \mu > 3530$

b. $t_* = \frac{\bar{x} - \mu_0}{s/\sqrt{n}} = \frac{3740 - 3530}{810/\sqrt{92}} = 2.49$

Degrees of freedom $= n - 1 = 91$. Using t-table, P-value is between 0.005 and 0.01.

Exact P-value $= 0.007$ (using R).

c. Since P-value $< \alpha = 0.01$, we reject H_0. We conclude that the mean attendance per game has increased.

Exercise

A courier service advertises that its average delivery time is less than 6 hours for local deliveries. A random sample of times for 12 deliveries to an address across town was recorded. These data are shown here. Is this sufficient evidence to support the courier's advertisement, at the 5% level of significance?
3.03 6.33 6.50 5.22 3.56 6.76
7.98 4.82 7.96 4.54 5.09 6.46.

Solution

$\mu =$ average delivery time for local **all** deliveries.

Step 1. $H_0 : \mu = 6$ vs $H_a : \mu < 6$.

Step 2. $t_* = \frac{\bar{x} - \mu_0}{s/\sqrt{n}} = \frac{5.69 - 6}{1.58/\sqrt{12}} = -0.6796$

Step 3. Degrees of freedom $= n - 1 = 11$.

Using t-table, P-value > 0.25.

Exact P-value $= 0.2538$ (using R).

Step 4. Since P-value $> \alpha = 0.05$, we CAN'T reject H_0. There is not enough evidence to support the courier's advertisement.

R Code

```r
# Step 1. Entering data;

dataset=c(3.03, 6.33,6.50,5.22,3.56,6.76,
7.98,4.82,7.96,4.54,5.09,6.46);

# Step 2. Finding CI;

t.test(dataset, alternative="less", mu=6);
```

R Code

```
##
##  One Sample t-test
##
## data:  dataset
## t = -0.68499, df = 11, p-value = 0.2538
## alternative hypothesis: true mean is less than 6
## 95 percent confidence interval:
##      -Inf 6.506806
## sample estimates:
## mean of x
##    5.6875
```

"Pepsi" problem

A market research consultant hired by the Pepsi-Cola Co. is interested in determining the proportion of UTM students who favor Pepsi-Cola over Coke Classic. A random sample of 100 students shows that 40 students favor Pepsi over Coke. Use this information to construct a 95% confidence interval for the proportion of all students in this market who prefer Pepsi.

Bernoulli Distribution

$$x_i = \begin{cases} 1 & \text{i-th person prefers Pepsi} \\ 0 & \text{i-th person prefers Coke} \end{cases}$$

$\mu = E(x_i) = p$
$\sigma^2 = V(x_i) = p(1-p)$
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 $N\left(p, \sqrt{\frac{p(1-p)}{n}}\right)$

Interval Estimate of p

Draw a simple random sample of size n from a population with unknown proportion p of successes. An (approximate) confidence interval for p is:

$$\hat{p} \pm z_* \left(\sqrt{\frac{\hat{p}(1-\hat{p})}{n}} \right)$$

where z_* is a number coming from the Standard Normal that depends on the confidence level required.
Use this interval only when:
1) n is "large" and
2) $n\hat{p} \geq 5$ and $n(1-\hat{p}) \geq 5$.

Example

A simple random sample of 400 individuals provides 100 Yes responses.

a. What is the point estimate of the proportion of the population that would provide Yes responses?

b. What is the point estimate of the standard error of the proportion, $\sigma_{\hat{p}}$?

c. Compute the 95% confidence interval for the population proportion.

Solution

a. $\hat{p} = \frac{100}{400} = 0.25$

b. Standard error of $\hat{p} = \sqrt{\frac{(\hat{p})(1-\hat{p})}{n}} = \sqrt{\frac{(0.25)(0.75)}{400}} = 0.0216$

c. $\hat{p} \pm z_* \left(\sqrt{\frac{(\hat{p})(1-\hat{p})}{n}} \right)$

$0.25 \pm 1.96(0.0216)$

$(0.2076, 0.2923)$

Example

A simple random sample of 800 elements generates a sample proportion $\hat{p} = 0.70$.

a. Provide a 90% confidence interval for the population proportion.

b. Provide a 95% confidence interval for the population proportion.

Solution

a. $\hat{p} \pm z_* \left(\sqrt{\frac{(\hat{p})(1-\hat{p})}{n}} \right)$

$0.70 \pm 1.65 \left(\sqrt{\frac{(0.70)(1-0.70)}{800}} \right)$

$0.70 \pm 1.65(0.0162)$

$(0.6732, 0.7267)$

b. $0.70 \pm 1.96(0.0162)$

$(0.6682, 0.7317)$

Problem

A survey of 611 office workers investigated telephone answering practices, including how often each office worker was able to answer incoming telephone calls and how often incoming telephone calls went directly to voice mail. A total of 281 office workers indicated that they never need voice mail and are able to take every telephone call.

a. What is the point estimate of the proportion of the population of office workers who are able to take every telephone call?

b. At 90% confidence, what is the margin of error?

c. What is the 90% confidence interval for the proportion of the population of office workers who are able to take every telephone call?

Solution

a. $\hat{p} = \frac{281}{611} = 0.46$

b. Margin of error $=$

$z_* \sqrt{\frac{(\hat{p})(1-\hat{p})}{n}} = 1.65\sqrt{\frac{(0.46)(0.54)}{611}} = 1.65(0.0201) = 0.0332$

c. $\hat{p} \pm z_* \left(\sqrt{\frac{(\hat{p})(1-\hat{p})}{n}} \right)$

$0.46 \pm .0332$

$(0.4268, 0.4932)$

R Code

```
prop.test(281,611,conf.level=0.90,correct=FALSE);

# correct = FALSE;
# this is telling R NOT to use
# Yates' continuity correction;
```

R Code

```
##
##  1-sample proportions test without continuity correction
##
## data:  281 out of 611, null probability 0.5
## X-squared = 3.9296, df = 1, p-value = 0.04744
## alternative hypothesis: true p is not equal to 0.5
## 90 percent confidence interval:
##  0.4269866 0.4931705
## sample estimates:
##         p
## 0.4599018
```

Nielsen Ratings

Statistical techniques play a vital role in helping advertisers determine how many viewers watch the shows that they sponsor. There are several companies that sample television viewers to determine what shows they watch, the best known of which is the A. C. Nielsen firm. The Nielsen Ratings are based on a sample of randomly selected families. A device attached to the family television keeps track of the channels the television receives. The ratings then produce the proportions of each show from which sponsors can determine the number of viewers and the potential value of any commercials.

Nielsen Ratings

The results for the 18-to 49-year-old group on Thursday, March 7, 2013, for the time slot 8:00 p.m. to 8:30 p.m. have been recorded using the following codes:

Network	Show	Code
ABC	Shark Tank	1
CBS	Big Bang Theory	2
CW	The Vampire Diaries	3
Fox	American Idol	4
NBC	Community	5
Television turned off		6

CBS would like to use the data to estimate how many Americans aged 18 to 49 were tuned to its program *Big Bang Theory*.

R Code

```r
#Step 1. Entering data;

# importing data;

# url of ratings;
url="https://mcs.utm.utoronto.ca/~nosedal/data/rating.txt"

ratings_data= read.table(url,header=TRUE);

names(ratings_data);

# first 6 observations from file
ratings_data[1:6, ]
```

R Code

```
## [1] "ViewerNumber" "TV.Program"
##   ViewerNumber TV.Program
## 1            1          6
## 2            2          6
## 3            3          6
## 4            4          6
## 5            5          6
## 6            6          6
```

R Code

```
all.programs=ratings_data$TV.Program;

# I want you to see the first 6 observations;

all.programs[1:6];
```

R Code

```
## [1] 6 6 6 6 6 6
```

R Code

```
# Recall that Big Bang Theory's code is 2;

big.bang=all.programs[all.programs==2];

# First 6 observations from big.bang

big.bang[1:6]
```

R Code

```
## [1] 2 2 2 2 2 2
```

R Code

```
## CI for p;

sample.size=length(all.programs);

sample.size;

successes=length(big.bang);

successes;

prop.test(successes,sample.size,conf.level=0.95,
correct=FALSE);
```

R Code

```
## [1] 5000
## [1] 275
##
##  1-sample proportions test without continuity correction
##
## data:  successes out of sample.size, null probability 0.
## X-squared = 3960.5, df = 1, p-value < 2.2e-16
## alternative hypothesis: true p is not equal to 0.5
## 95 percent confidence interval:
##  0.04901565 0.06166761
## sample estimates:
##     p
## 0.055
```

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
$H_a : p > p_0 : \text{is} : P(Z > z_*)$
$H_a : p < p_0 : \text{is} : P(Z < z_*)$
$H_a : p \neq p_0 : \text{is} : 2P(Z > |z_*|)$
Use this test when the sample size n is so large that both np_0 and $n(1 - p_0)$ are 10 or more.

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);

# R uses a different test statistic;
# but you will get the same P-value;
```

R Code

```
##
##  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);

# R uses a different test statistic;
# but you will get the same P-value;
```

R Code

```
##
##  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 YouTube user 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

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

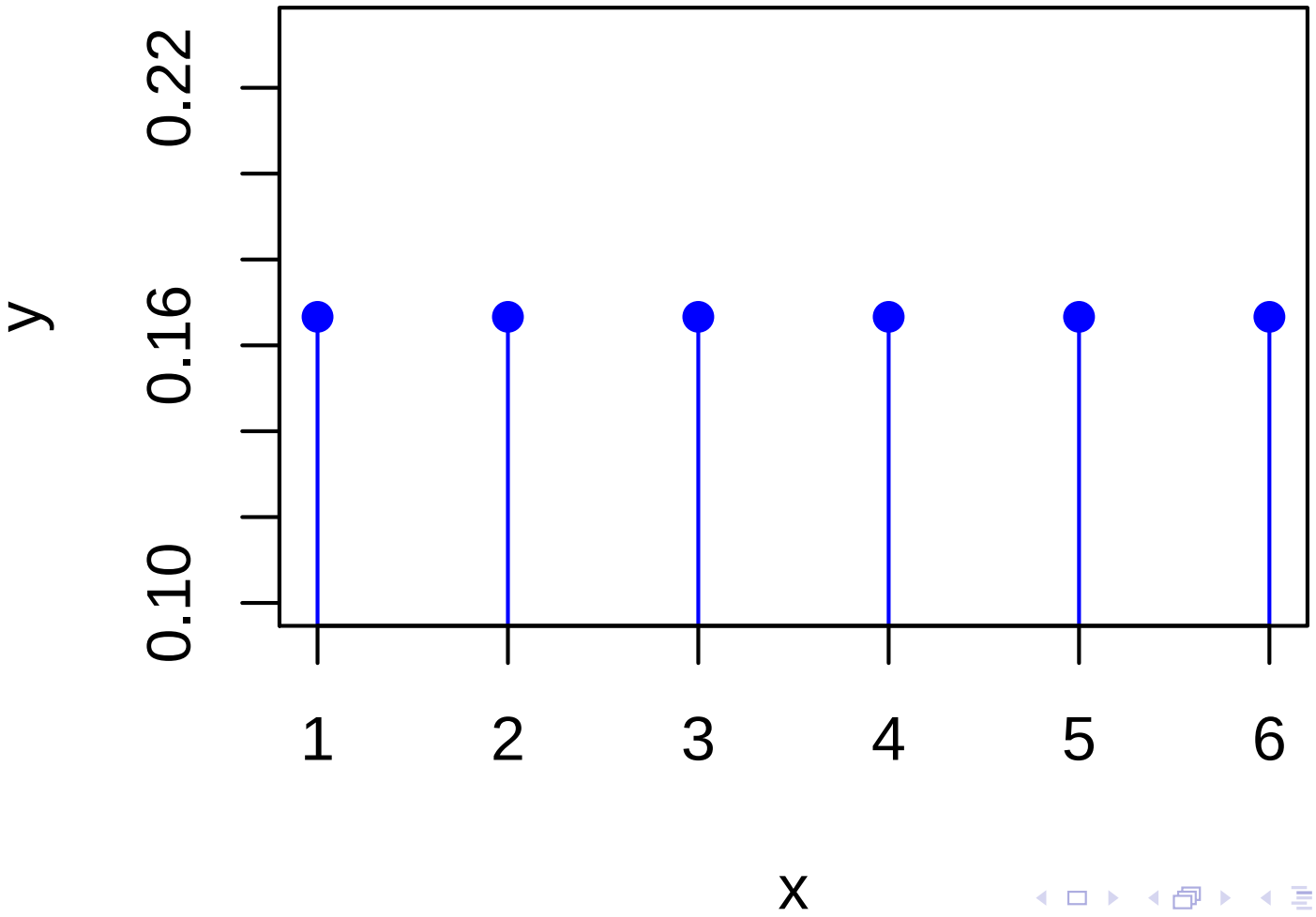
"Roll a 6" is an event, that contains one of these 6 outcomes.

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,

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

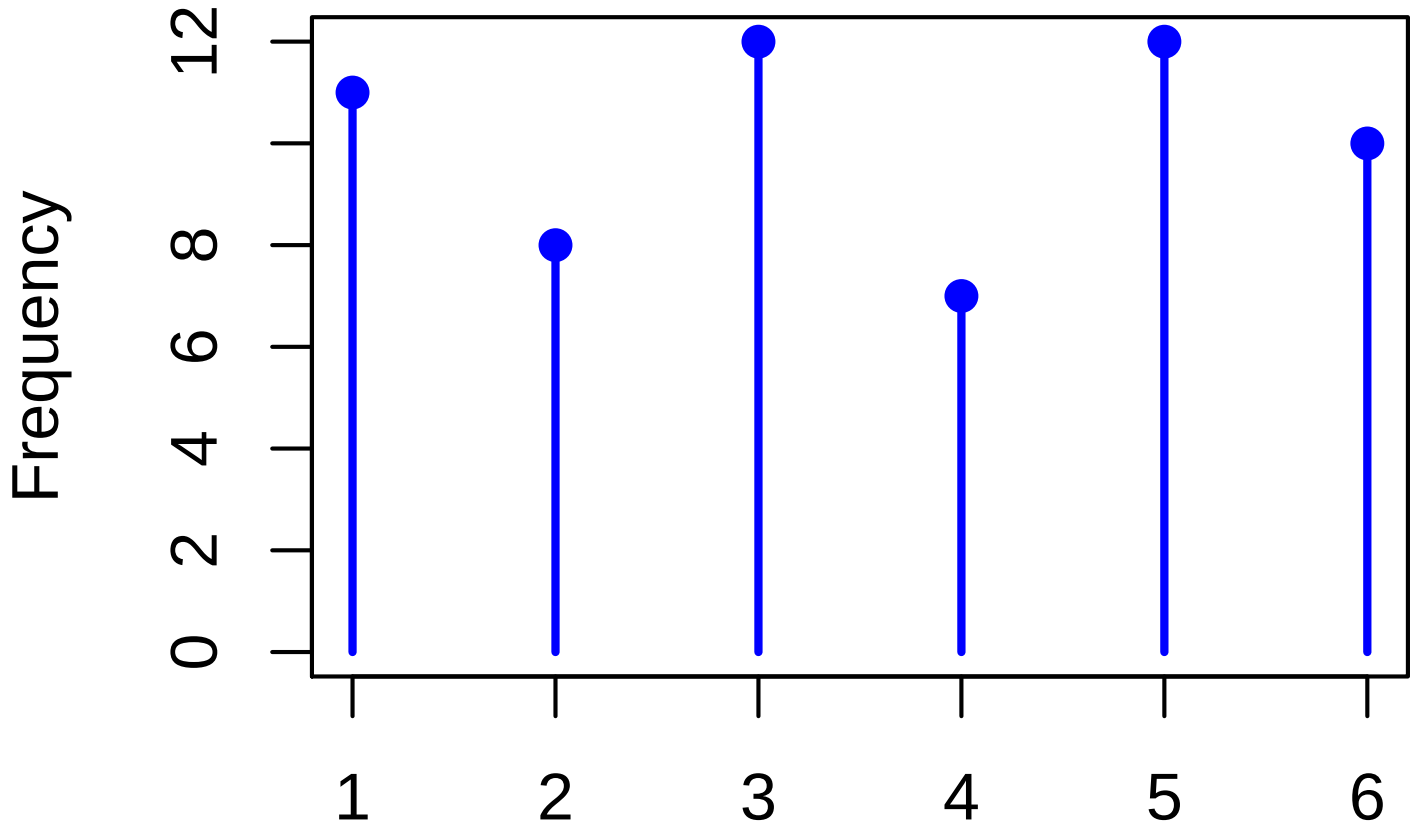
Probability mass function (pmf)



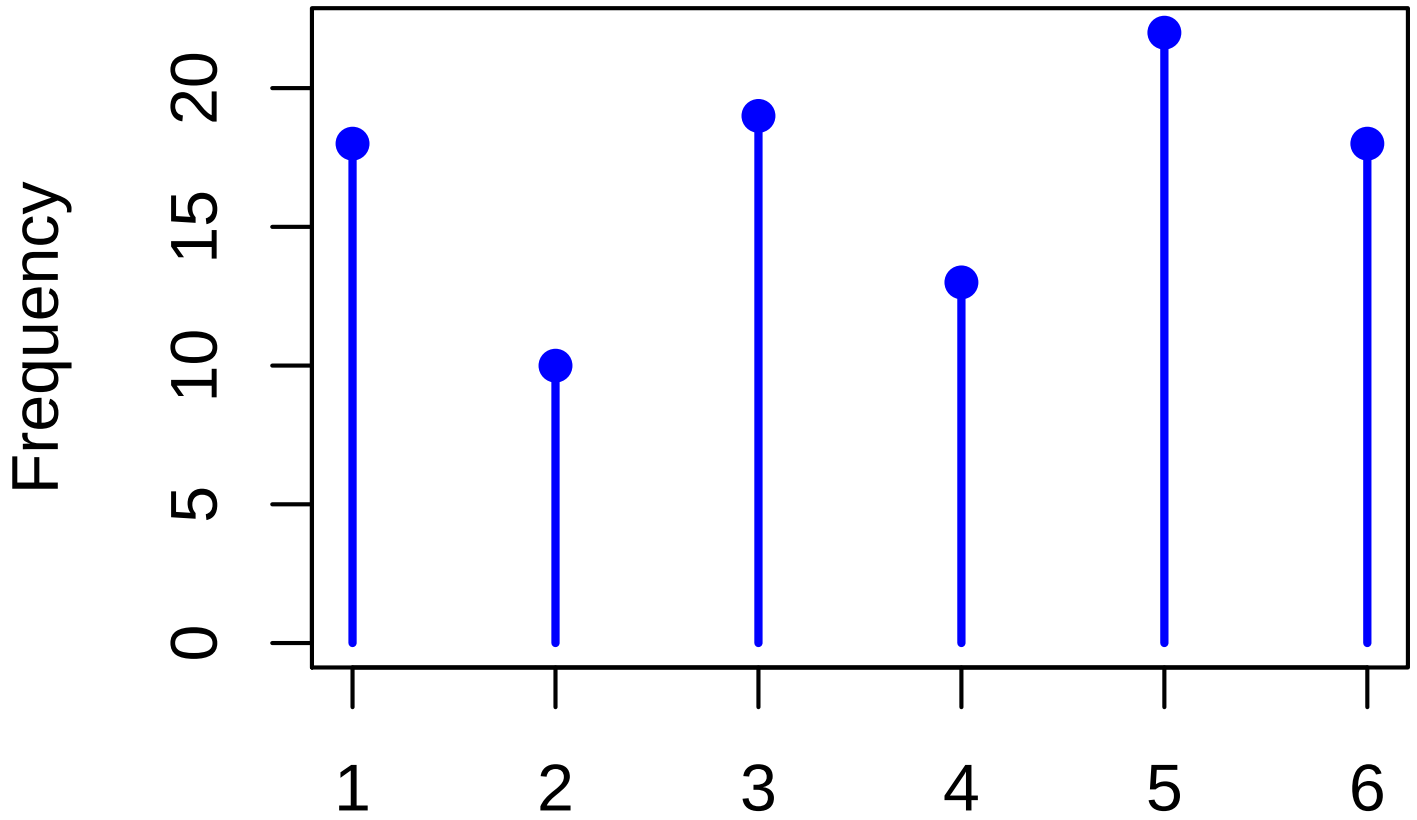
Six simulations

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

sample(die,1,replace=TRUE);

## [1] 4

sample(die,6,replace=TRUE);

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

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

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

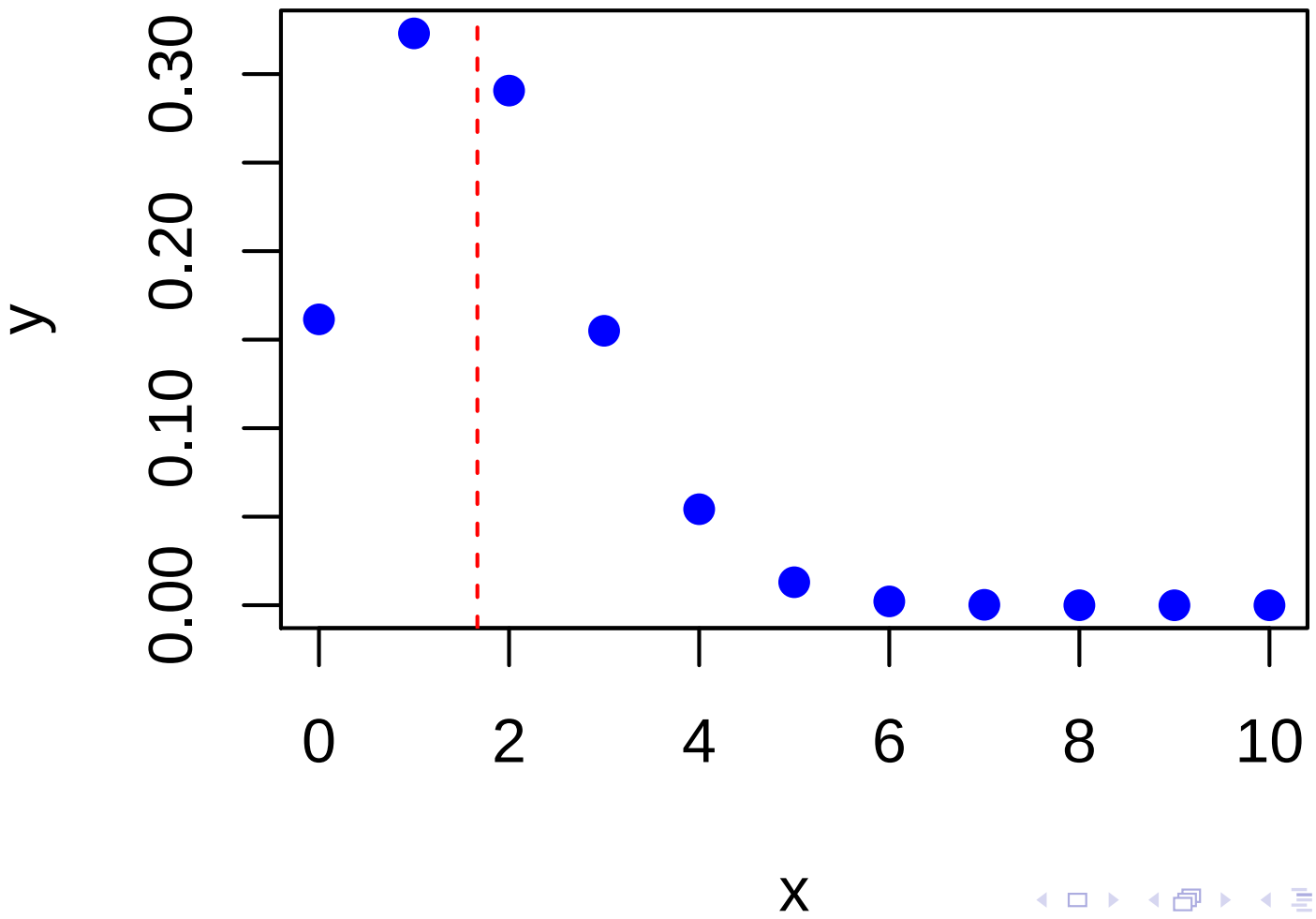
A few simulations

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

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

## [1] 4

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

## [1] 3

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

## [1] 0
```

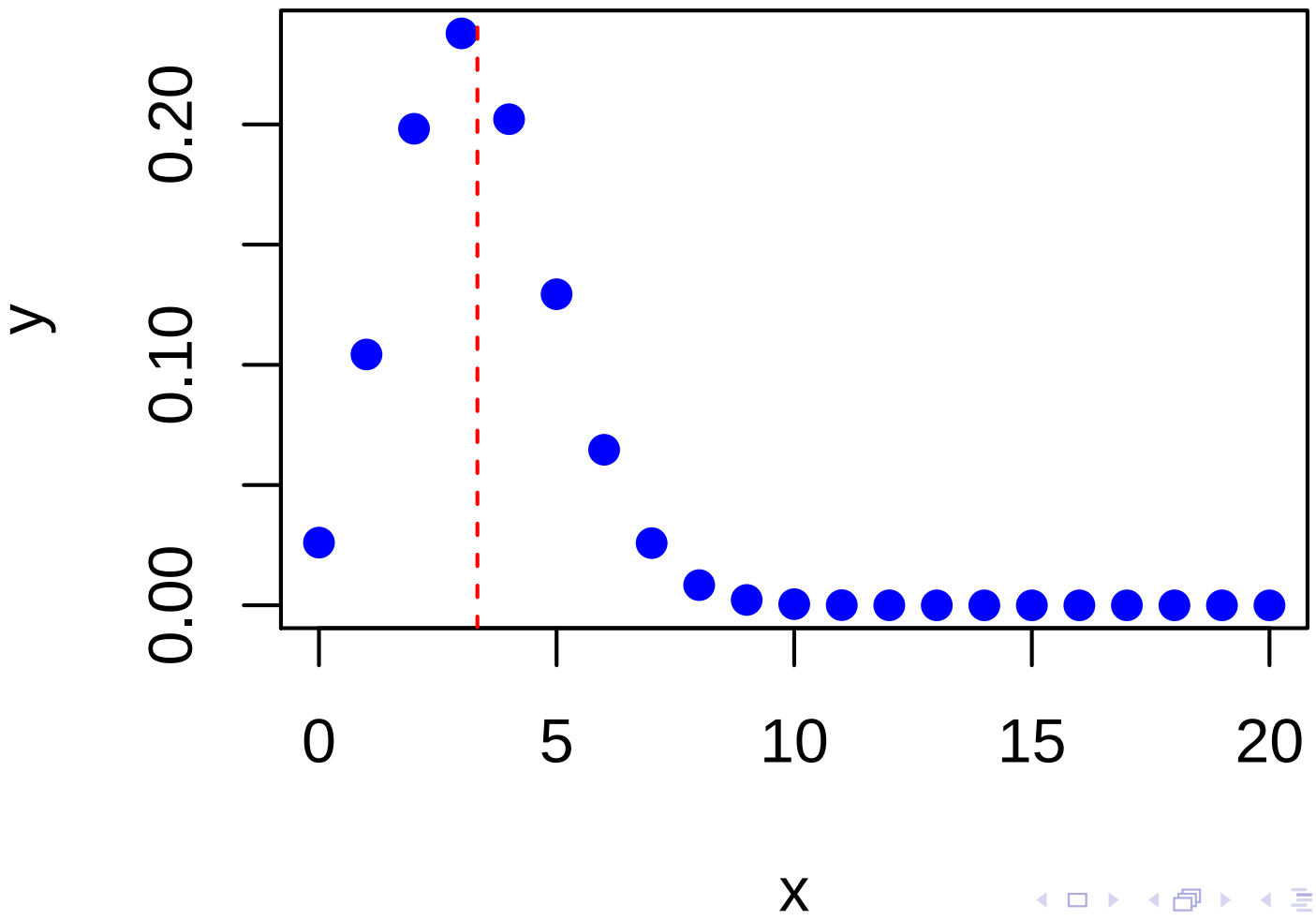
Probability Mass Function when n=10 and p=1/6

```
## 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);
```

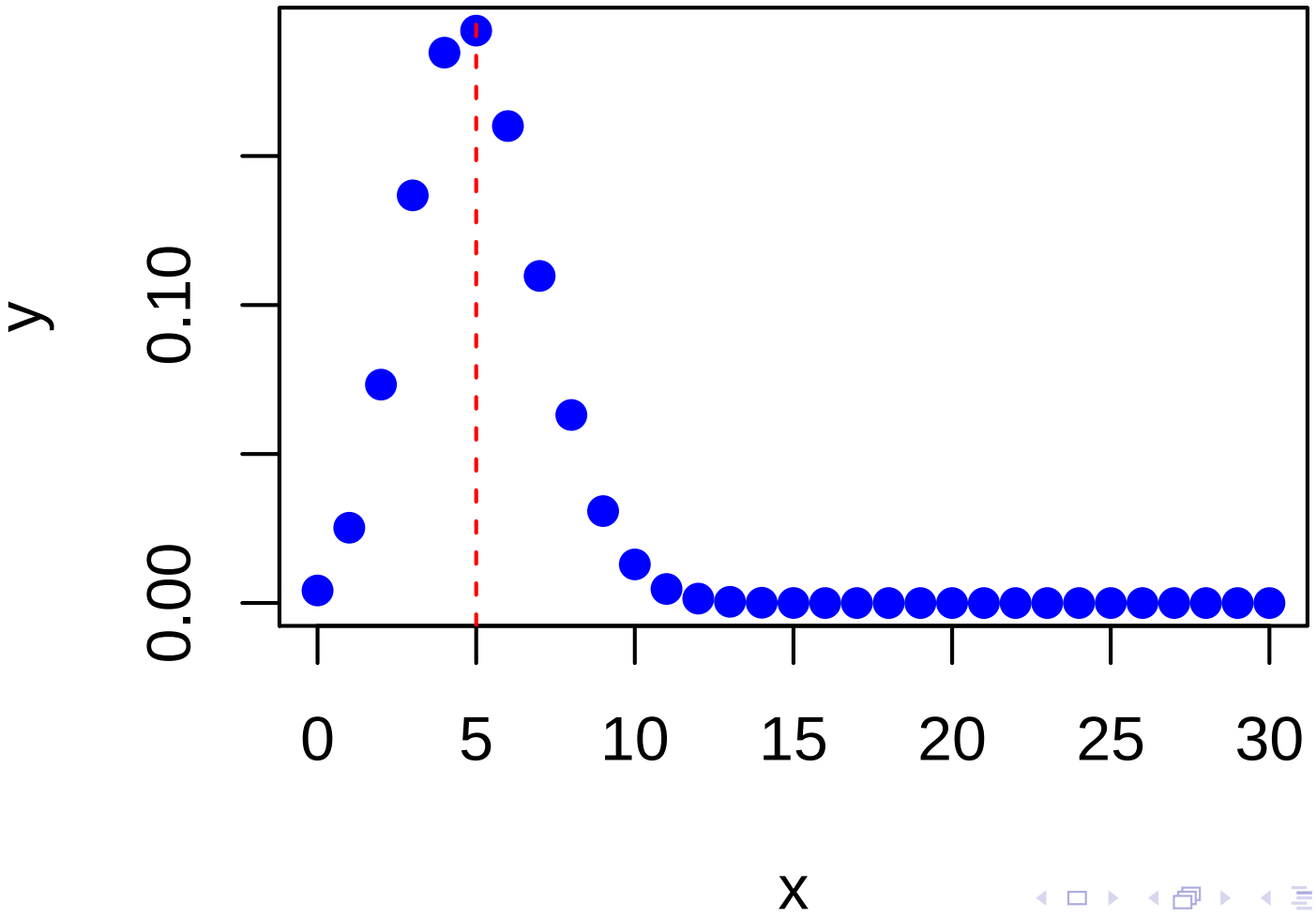
Probability Mass Function when n=10 and p=1/6



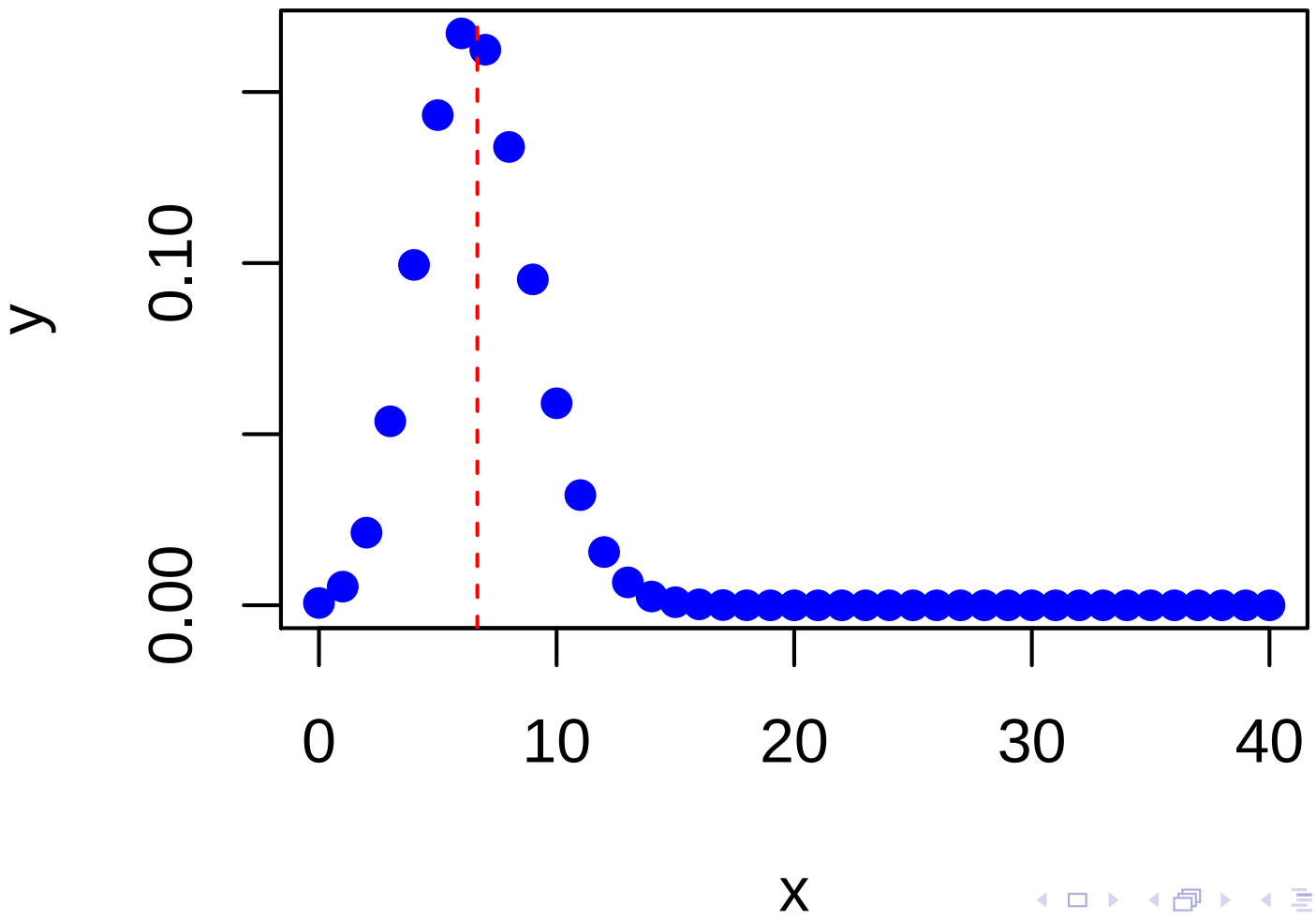
Pmf when n=20 and p=1/6



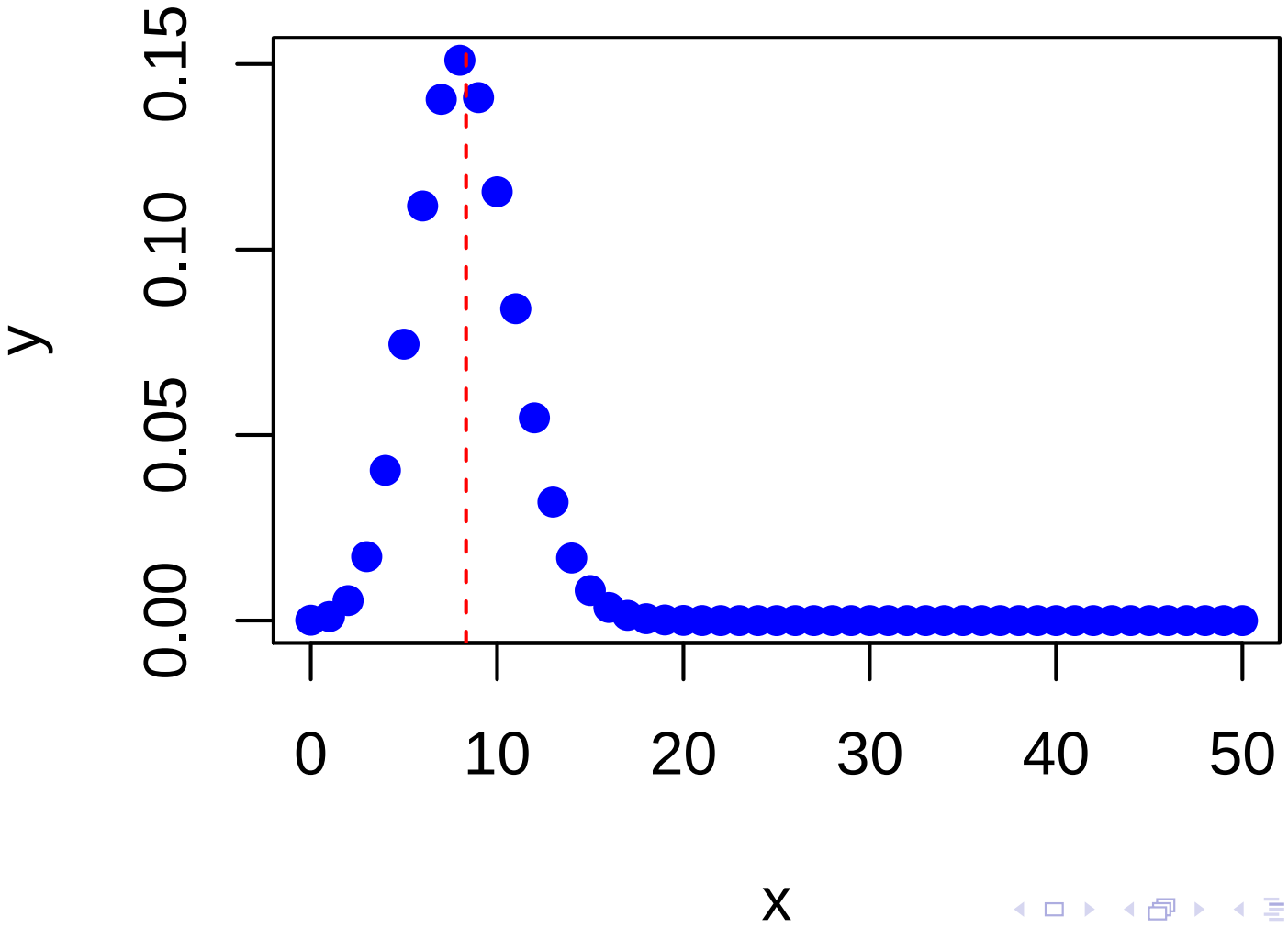
Pmf when n=30 and p=1/6



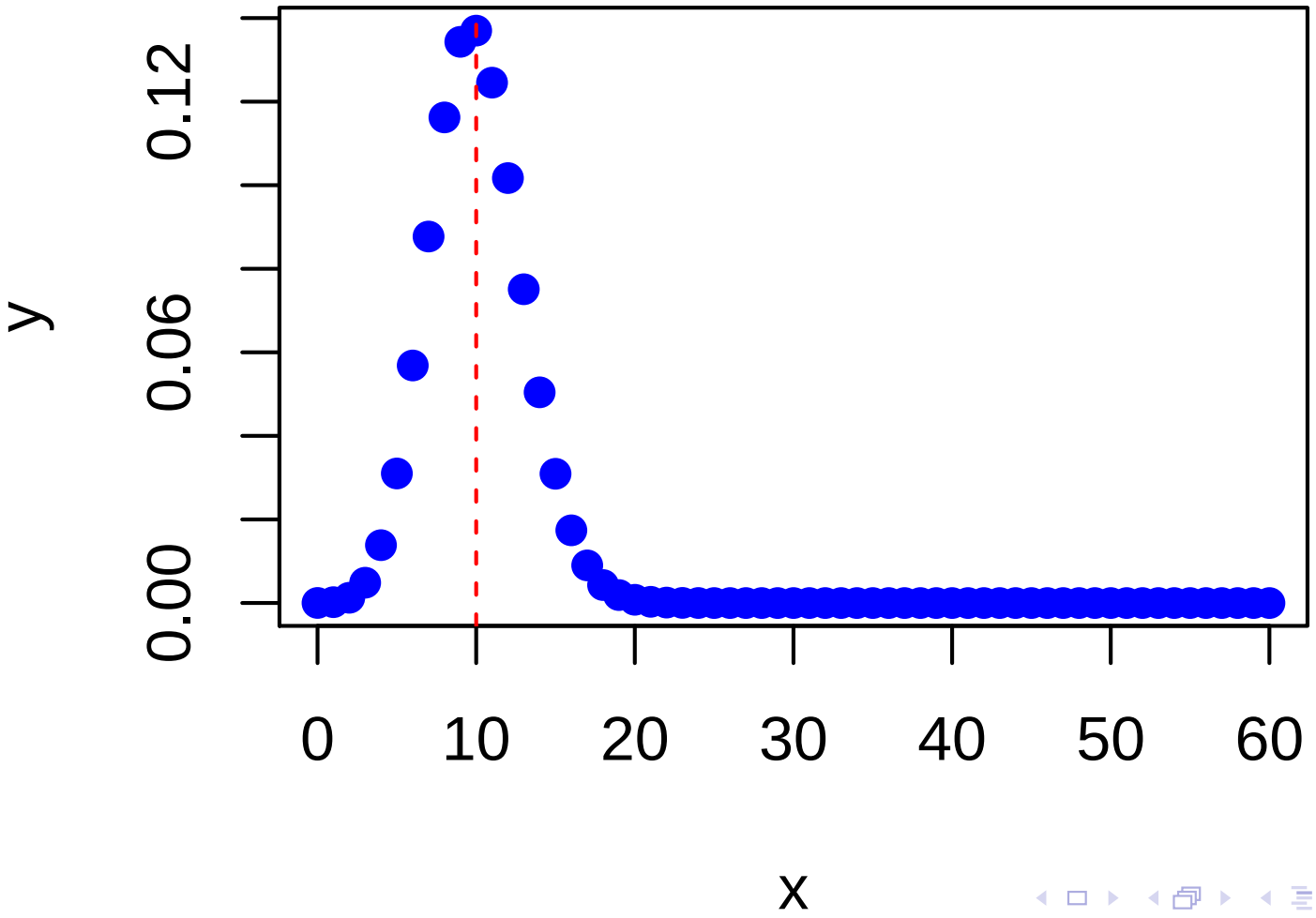
Pmf when n=40 and p=1/6



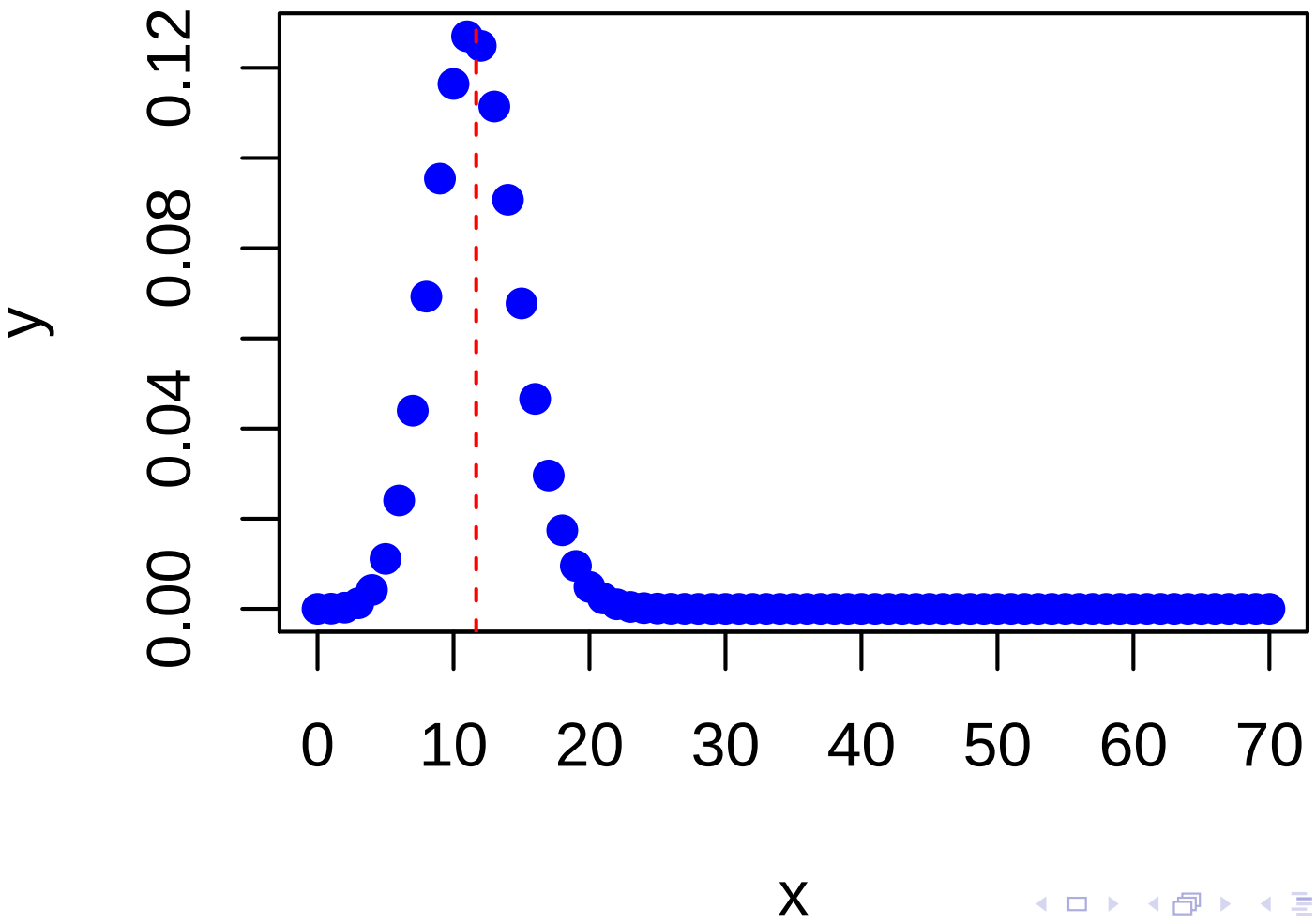
Pmf when n=50 and p=1/6



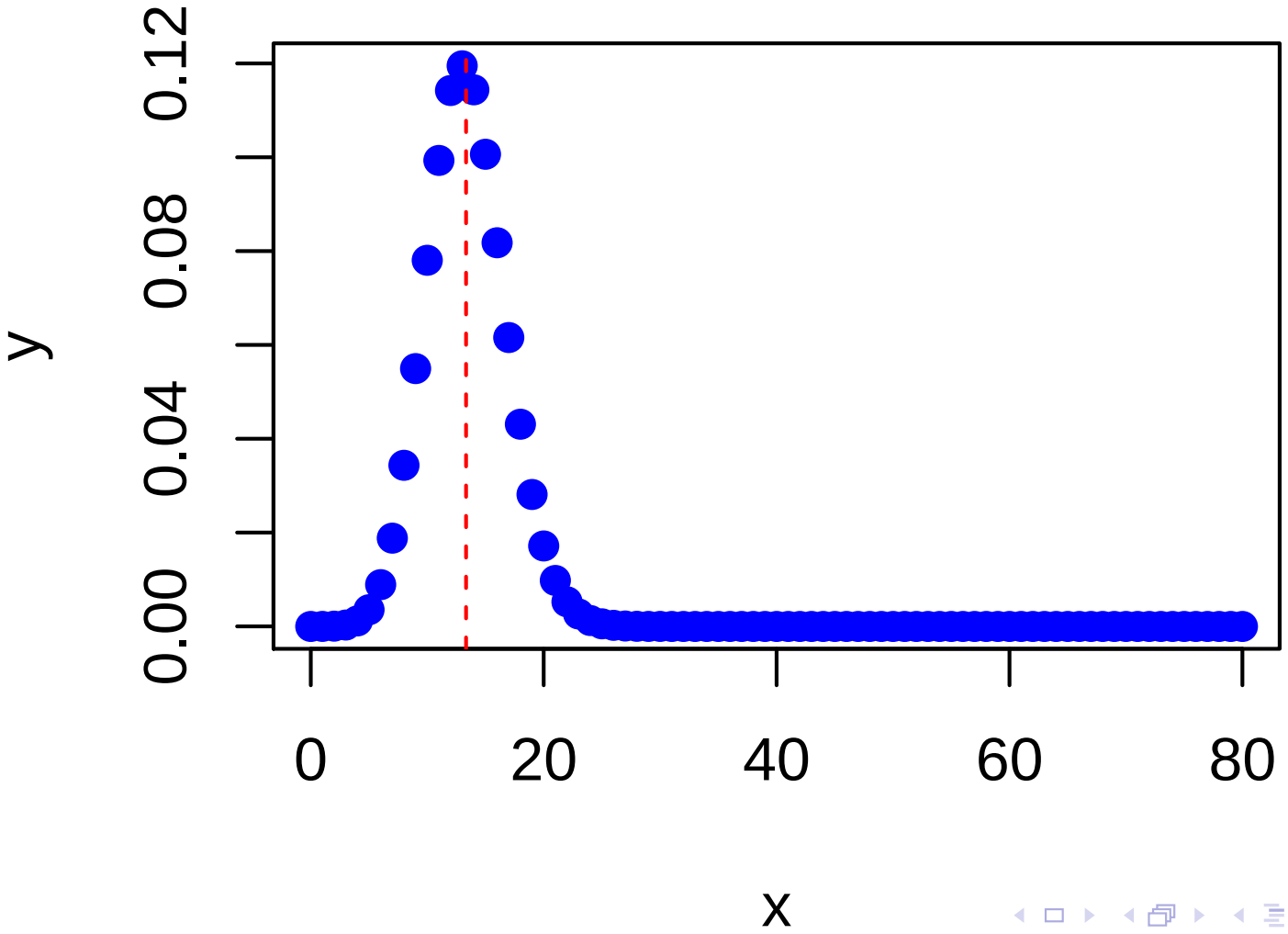
Pmf when n=60 and p=1/6



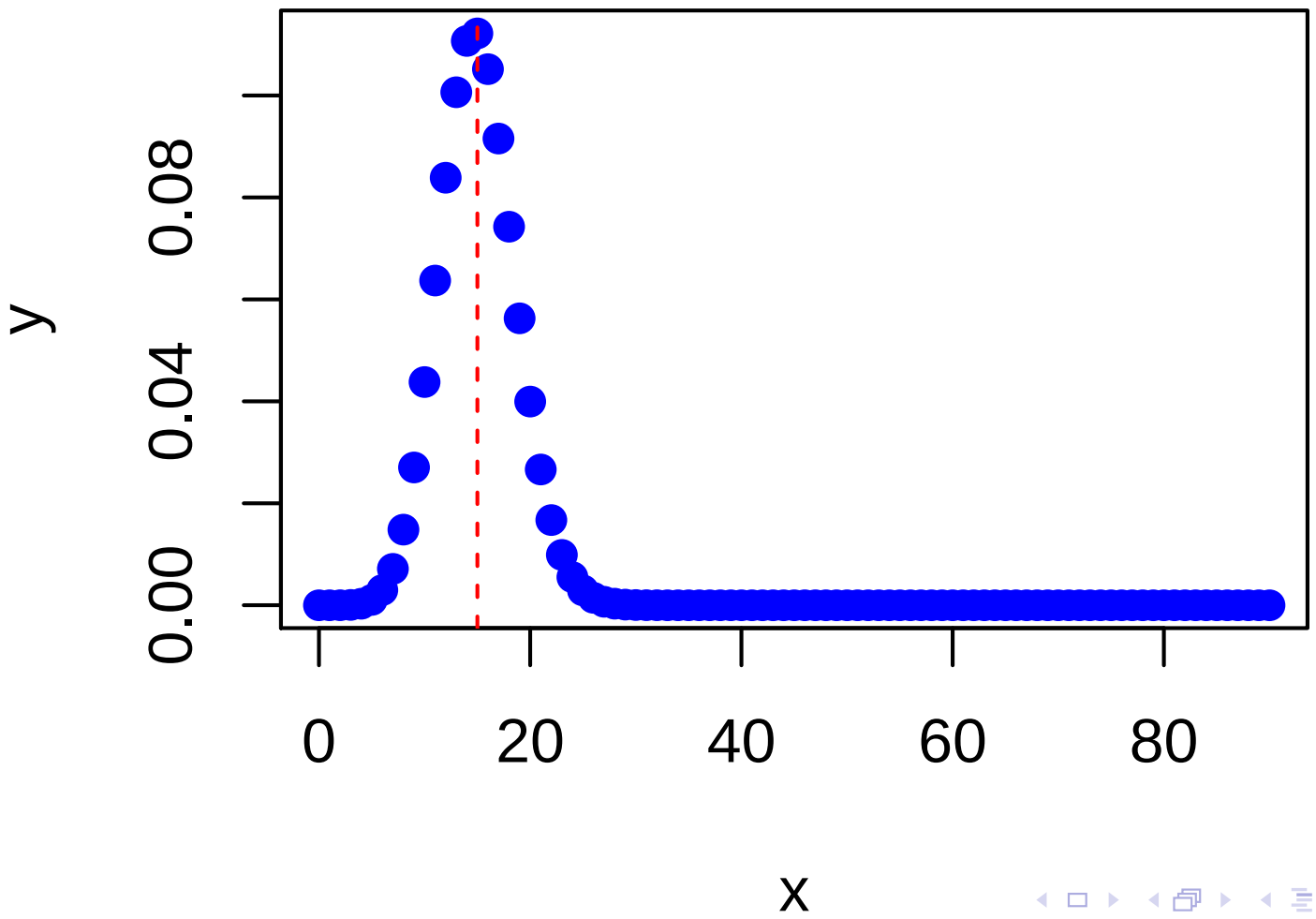
Pmf when n=70 and p=1/6



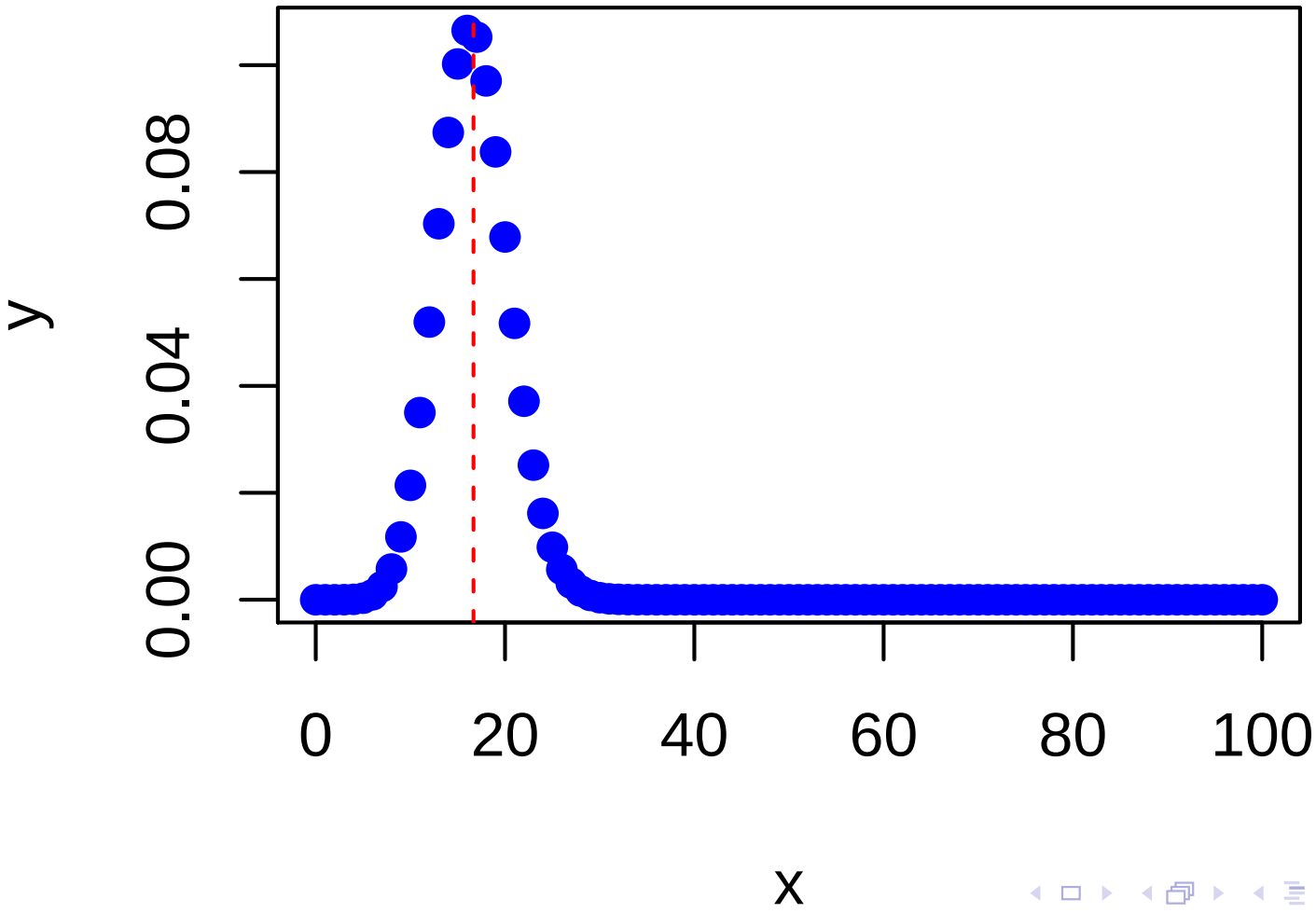
Pmf when n=80 and p=1/6



Pmf n=90 and p=1/6



Pmf when n=100 and p=1/6

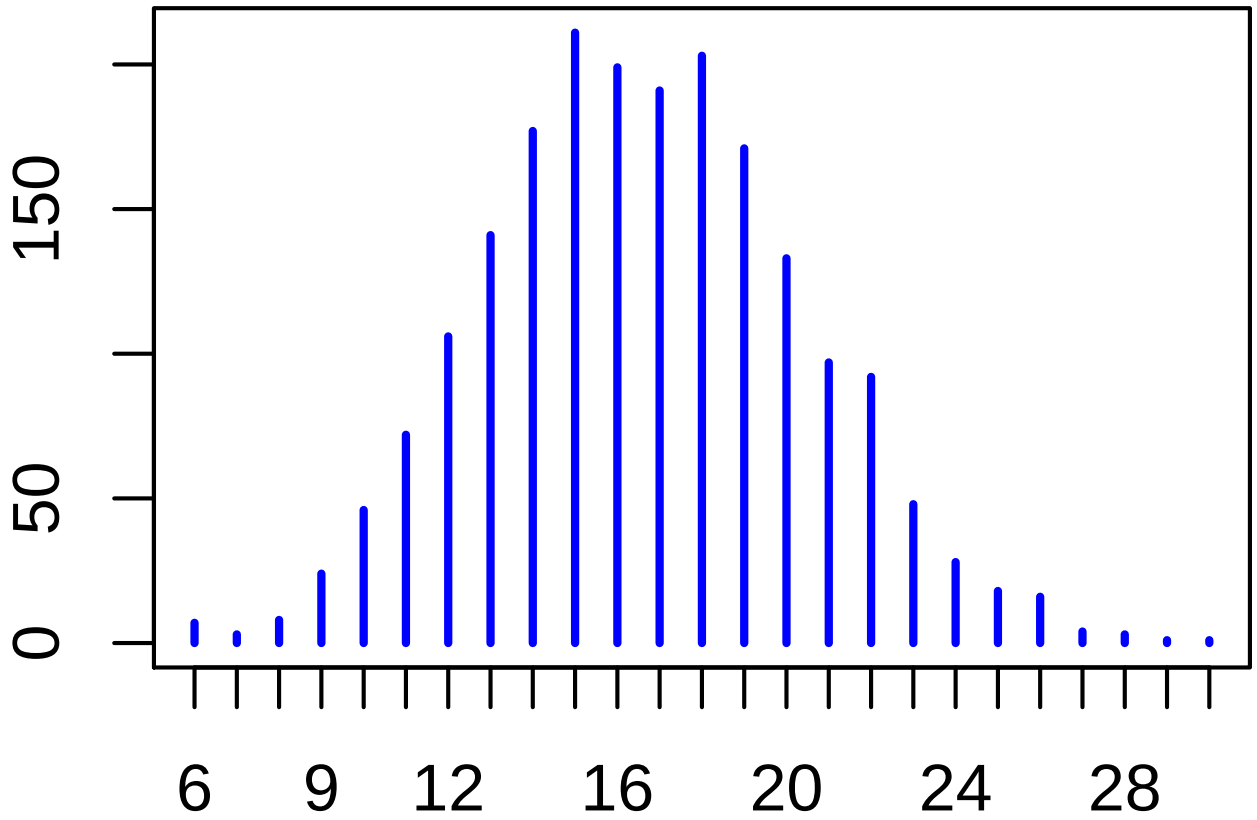


A few values from our pmf (n=100 and p=1/6)

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

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

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



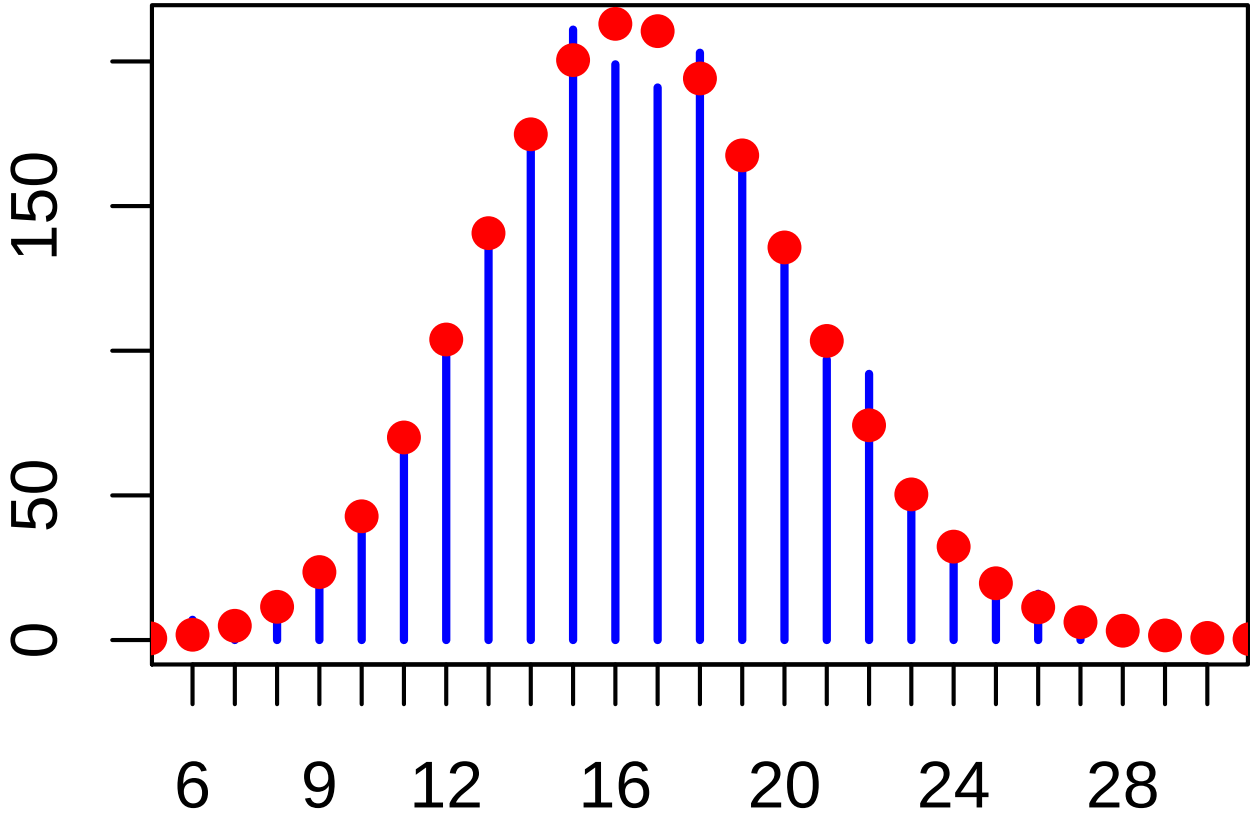
A few values from our simulation

```
## 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

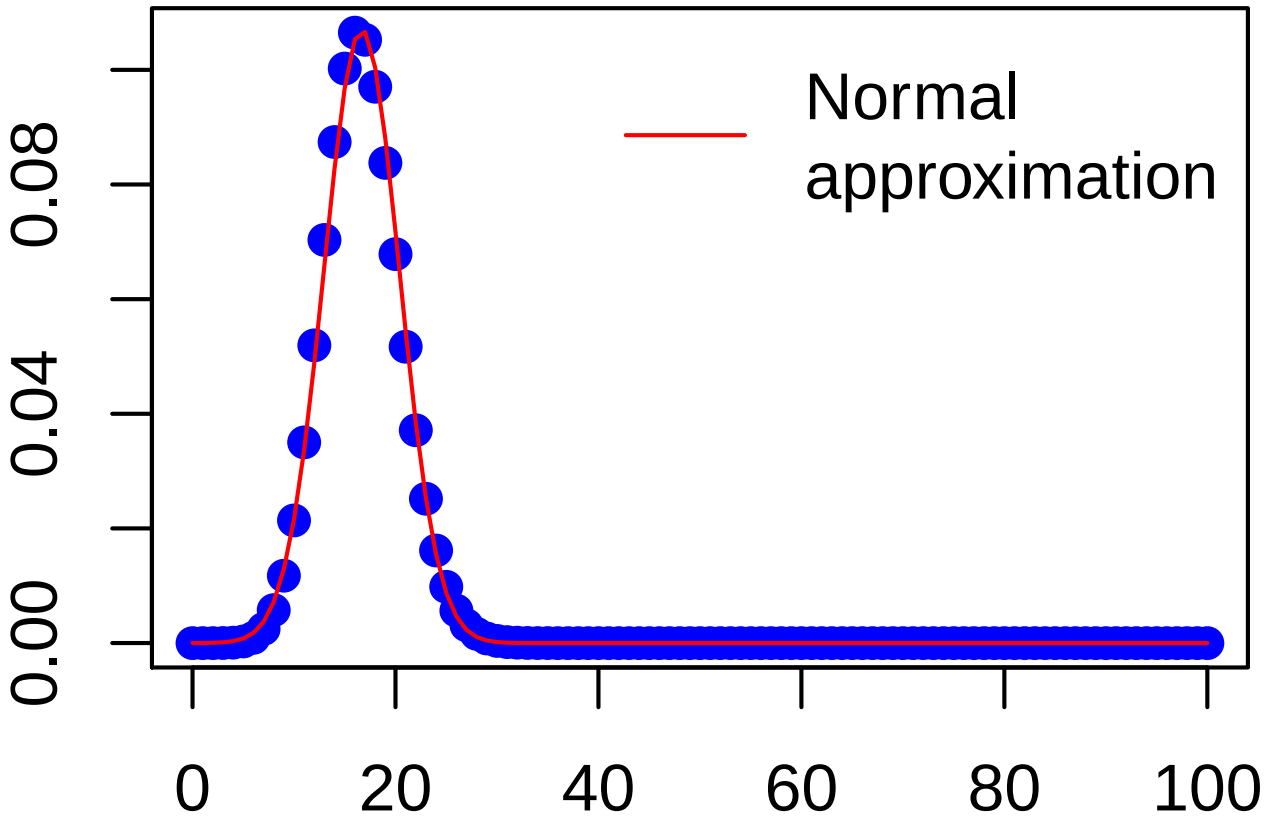
Draw an SRS of size n from a large population that contains proportion p of "successes". Let $\hat{p}$ be the **sample proportion** of successes,

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

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
$H_a : p > p_0 : \text{is} : P(Z > z_*)$
$H_a : p < p_0 : \text{is} : P(Z < z_*)$
$H_a : p \neq p_0 : \text{is} : 2P(Z > |z_*|)$

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)

$z_* = \frac{\hat{p} - p_0}{\sqrt{\frac{p_0(1-p_0)}{n}}} = \frac{0.12 - 0.1667}{\sqrt{\frac{(0.12)(0.88)}{100}}} = -1.25$

(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
```