Unknown Proportions

Decisions with Data | Inference for proportions

STAT5002

The University of Sydney

Mar 2025

Decisions with Data

Topics 8 and 9: Confidence intervals and the z-test

Topic 10: The t-test

Topic 11: The two-sample test

Topic 12: χ^2-test

Outline

Today

- Confidence Interval

Next week

- Hypothesis Test
- Review

Important special case: 0-1 box

An important example is where the box only contains $\boxed{0}$ and $\boxed{1}$.

Let $0 \le p \le 1$ denote the proportion of $\boxed{1}$s in the box, and N be the size of the box. Then, the box contains

$$\boxed{\underbrace{\boxed{0}\cdots\boxed{0}}_{(1-p)N \text{ of these}} \quad \underbrace{\boxed{1}\cdots\boxed{1}}_{pN \text{ of these}}}$$

- The mean of the box $\mu = \frac{pN}{N} = p$;
- The mean square of the box is also p, and so the SD of the box is

$$\sigma = \sqrt{\text{mean.sq.} - (\text{mean})^2} = \sqrt{p - p^2} = \sqrt{p(1-p)}\,,$$

  only depending on p.

Taking n draws from the box, then

- $E(S)$, $E(\bar{X})$, $SE(S)$ and $SE(\bar{X})$ only depends on p and n.
- $\bar{X}$ is also the sample proportion of $\boxed{1}$s

Prediction intervals

Prediction intervals

A $\gamma\%$ (two-sided) prediction interval for the sample sum S is an interval $[a, b]$ in which there is a $\gamma\%$ chance that S lands in $[a, b]$:

$$P(a \leq S \leq b) = \frac{\gamma}{100}$$

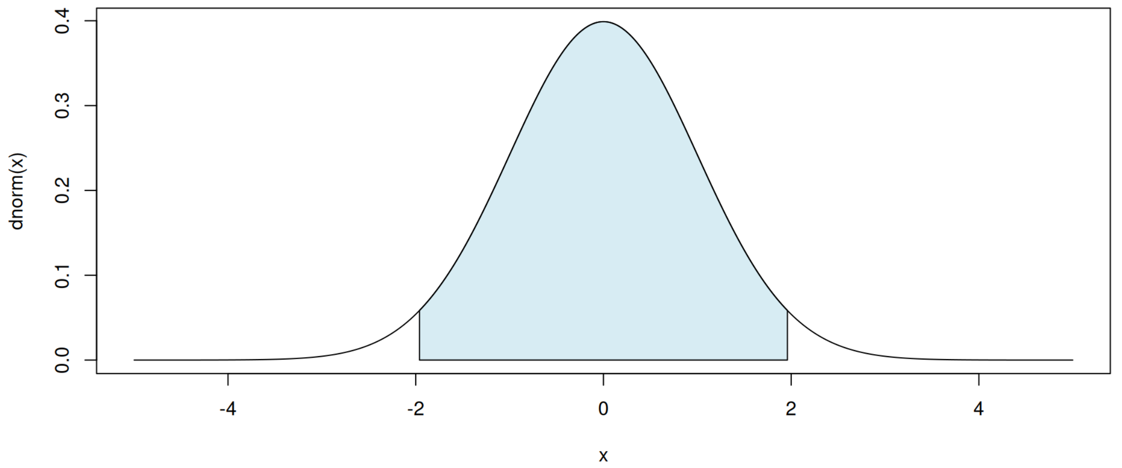
A $\gamma\%$ (two-sided) prediction interval for the sample mean $\bar{X}$ is an interval $[c, d]$ in which there is a $\gamma\%$ chance that $\bar{X}$ lands in $[c, d]$:

$$P(c \leq \bar{X} \leq d) = \frac{\gamma}{100}$$

How can we find $[a, b]$ or $[c, d]$?

Standard normal curve

Suppose X follows a general normal curve with mean $E(X)$ and SD $SE(X)$, then its standard unit $Z = \frac{X - E(X)}{SE(X)}$ follows the standard normal curve.




```r
round(qnorm(2.5/100), 2)
```

```
[1] -1.96
```

Under the standard normal curve

- Approximately 2.5% is to the left of -1.96 and 2.5% is to the right of 1.96.
- In other words 95% is between -1.96 and 1.96 (blue area).

Derivation for the sample mean $\bar{X}$

By CLT, $\bar{X}$ is approximately normal with mean $E(\bar{X})$ and SD $SE(\bar{X})$.

- Equivalently, $\frac{\bar{X}-E(\bar{X})}{SE(\bar{X})}$ is approximately standard normal $N(0,1)$

$$P(c \leq \bar{X} \leq d) = P\left(\underbrace{\frac{c - E(\bar{X})}{SE(\bar{X})}}_{=-1.96} \leq \frac{\bar{X} - E(\bar{X})}{SE(\bar{X})} \leq \underbrace{\frac{d - E(\bar{X})}{SE(\bar{X})}}_{=1.96}\right) = 95\%$$

So with $c = E(\bar{X}) - 1.96 \times SE(\bar{X})$ and $d = E(\bar{X}) + 1.96 \times SE(\bar{X})$, 95% of the time the sample mean $\bar{X}$ lands in $[c, d]$.

0-1 box

- Approximately, the 95% prediction interval for the sample mean $\bar{X}$ is

$$[E(\bar{X}) - 1.96 \times SE(\bar{X}), E(\bar{X}) + 1.96 \times SE(\bar{X})].$$

- For the 0-1 box with proportion p getting a $\boxed{1}$ and a sample size n. We have
  - $E(\bar{X}) = \mu = p$
  - $SE(\bar{X}) = \frac{\sigma}{\sqrt{n}} = \sqrt{\frac{p(1-p)}{n}}$
- The 95% prediction interval for the sample mean $\bar{X}$ is

$$\left[p - 1.96 \times \sqrt{\frac{p(1-p)}{n}}, p + 1.96 \times \sqrt{\frac{p(1-p)}{n}}\right].$$

- **Note** for other proportions $\gamma\%$, the value 1.96 needs to be adjusted.

Example: $p = 0.4$

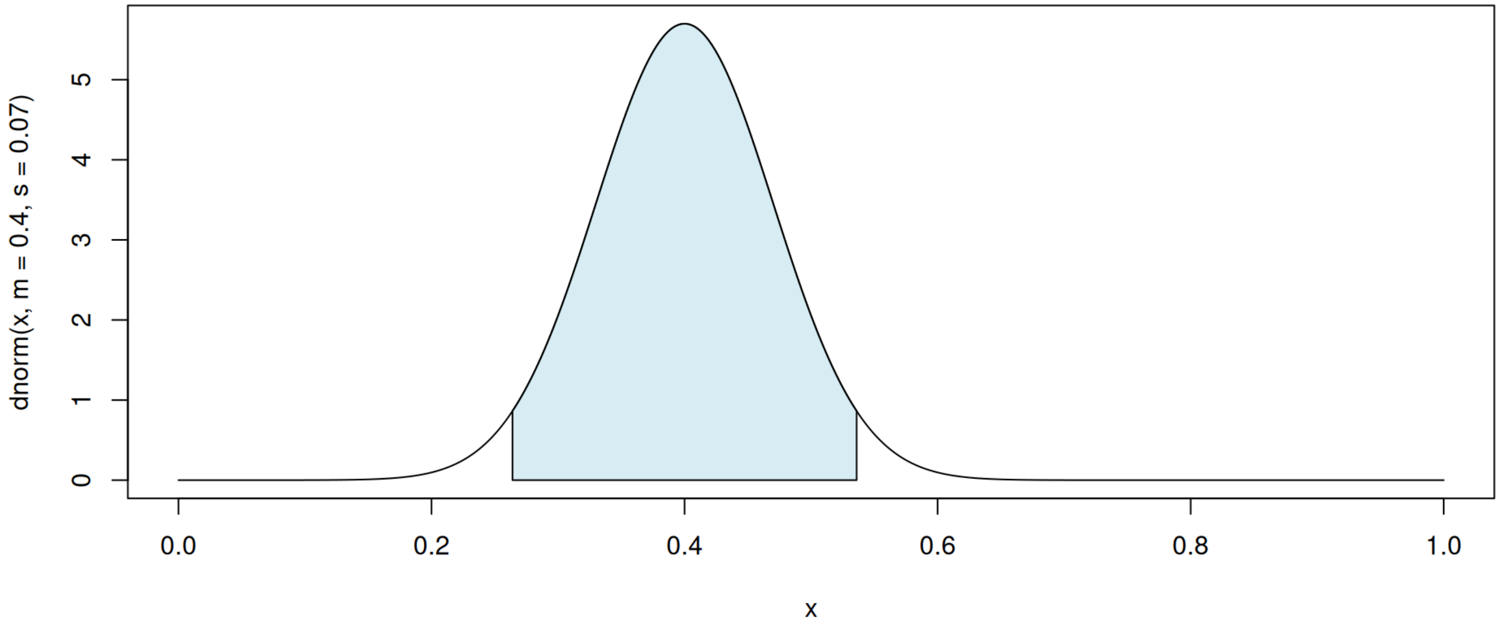
- Suppose we draw $n = 49$ times randomly from a box with $p = 0.4$.
- What is the 95% prediction interval for $\bar{X}$?

Solution:

- The expected value is $E(\bar{X}) = \mu = p = 0.4$;
- The standard error is $SE(\bar{X}) = \frac{\sigma}{\sqrt{n}} = \sqrt{\frac{p(1-p)}{n}} = \sqrt{\frac{1}{49} \times \frac{2}{5} \times \frac{3}{5}} = \frac{\sqrt{6}}{35} \approx 0.07$.
- Substituting into our prediction interval gives us: $[0.4 - 1.96 \times 0.07, 0.4 + 1.96 \times 0.07]$.

Visualisation

- The box of all possible sample proportions (sample mean) looks like a normal curve centred at 0.4, but scaled down by a factor of 0.07:



- Our 95% prediction interval is thus $\mathbf{0.4 \pm (1.96 \times 0.07)}$, i.e. roughly $\mathbf{(0.26, 0.54)}$:

```r
0.4 + c(-1, 1) * 1.96 * 0.07
```

```
[1] 0.2628 0.5372
```

What if $p = 0.2$ instead of 0.4?

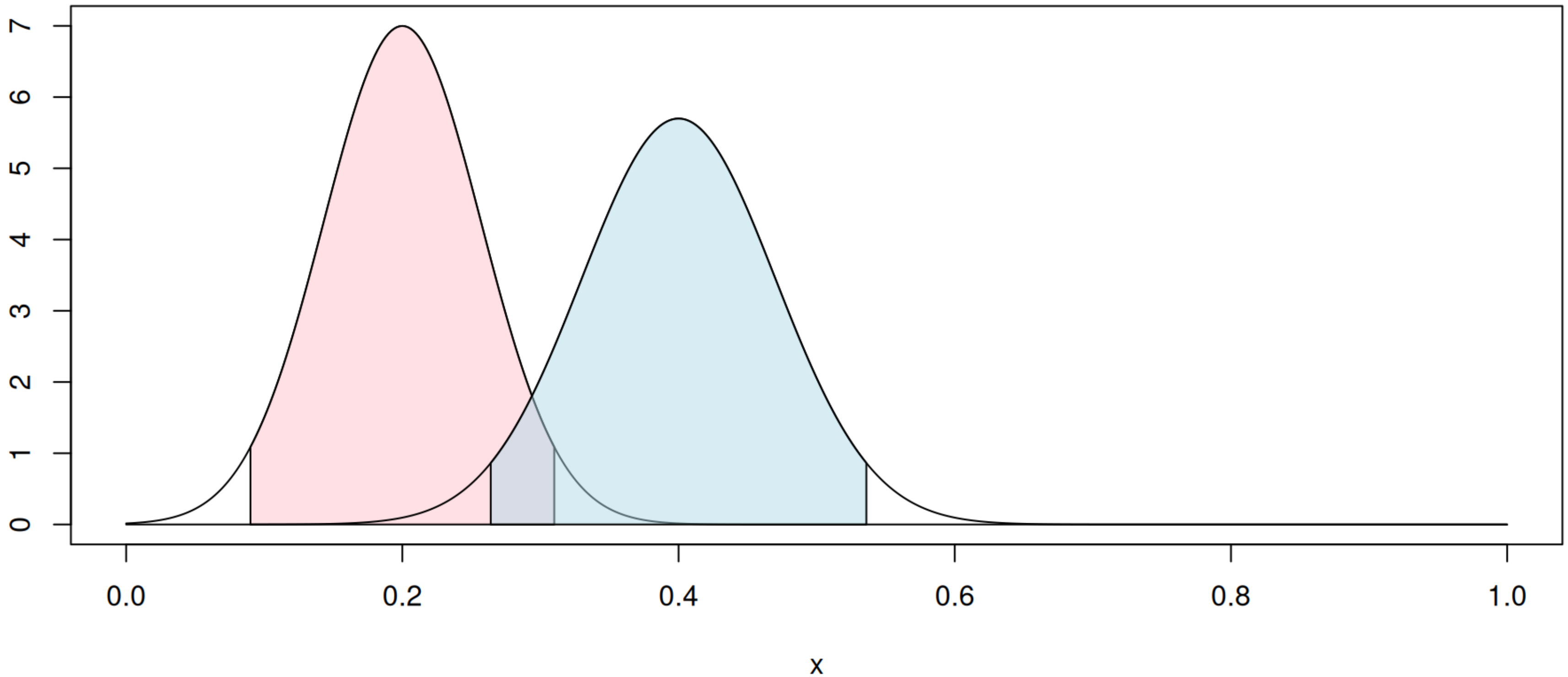
It is interesting to see how this changes if the proportion in the box is 0.2 instead of 0.4. We then get

- $E(\bar{X}) = \mu = p = 0.2$
- $SE(\bar{X}) = \frac{\sigma}{\sqrt{n}} = \sqrt{\frac{p(1-p)}{n}} = \sqrt{\frac{1}{49} \times \frac{1}{5} \times \frac{4}{5}} = \frac{2}{35} = 0.057$

So the box of all possible $\bar{X}$ values has

- Mean $\mathbf{0.2}$, SD $\mathbf{0.057}$, and approximately a normal shape.

Interval now a bit narrower



- The 95% prediction interval is now roughly $(0.09, 0.31)$, i.e. 0.22 wide (0.28 when $p = 0.4$).

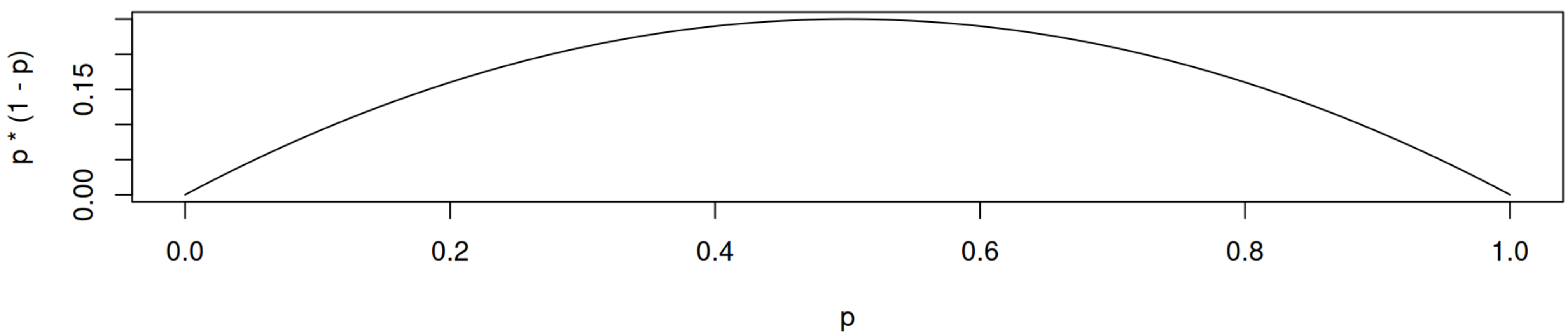
```
1  0.2 + c(-1, 1) * 1.96 * 0.057
```

```
[1] 0.08828 0.31172
```

Size of prediction intervals

- The variability in the sample proportion gets **smaller** as the p in the box gets **further from 0.5**.
- This is precisely reflected in $SE(\bar{X}) = \sqrt{\frac{p(1-p)}{n}}$.
- The function $p \mapsto p(1-p) = p - p^2$ is a quadratic function of p:

```r
p = 0:1000/1000
plot(p, p * (1 - p), type = "l")
```



Simulations

A function for simulating sample mean (proprotion) with p and n as input parameters

```r
sample_proportion = function(p, n) {
    samp = sample(c(0, 1), prob = c(1 - p, p), repl = T, size = n)
    prop = mean(samp)
    return(prop)
}
```

`sample(c(0,1), prob=c(1-p, p), repl=T, size=n)`

- `c(0,1)`: the box
- `prob=c(1-p, p)`: draw the ticket `0` with probability `1-p`, draw the ticket `1` with probability `p`

Repeat the experiment 1000 times for $p = 0.4$ and $p = 0.2$, check the percentage of times the simulated sample means falling outside of the prediction intervals

- The case $p = 0.4$, 2.5-th and 97.5-th percentiles are close to the prediction interval $(0.26, 0.54)$.

```r
p = 0.4
n = 49
props = replicate(1000, sample_proportion(p, n))
too.big = props > (0.4 + 1.96 * 0.07)  # right end
too.small = props < (0.4 - 1.96 * 0.07)  # left end
round(c(sum(too.big)/1000, sum(too.small)/1000), 3)  # proprotion of lower tail and upper tail
```

```
[1] 0.020 0.013
```

```r
round((1000 - sum(too.big) - sum(too.small))/1000, 3)  # proprotion in the interval
```

```
[1] 0.967
```

- The case $p = 0.2$, 2.5-th and 97.5-th percentiles are also close to the prediction interval $(0.09, 0.31)$.

```r
p = 0.2
n = 49
props = replicate(1000, sample_proportion(p, n))
too.big = props > (0.2 + 1.96 * 0.057)  # right end
too.small = props < (0.2 - 1.96 * 0.057)  # left end
round(c(sum(too.big)/1000, sum(too.small)/1000), 3)  # proprotion of lower tail and upper tail
```

```
[1] 0.026 0.026
```

```r
round((1000 - sum(too.big) - sum(too.small))/1000, 3)  # proprotion in the interval
```

```
[1] 0.948
```

Confidence Intervals

0-1 Box: interval of values consistent with each p

- The previous section showed how the sample mean/proportion $\bar{X}$ behaves for a **known** box proportion p.
- For each value p, it is associated with a 95% prediction interval

$$\left[p - 1.96 \times \sqrt{\frac{p(1-p)}{n}}, p + 1.96 \times \sqrt{\frac{p(1-p)}{n}}\right].$$

  which is an **interval of sample means consistent with that p**

  - The interval is centred at p
  - Its width depends on n and p
  - Interval gets wider when the proportion p gets closer to 0.5.
- **Consistency:** with 95% chance, sample means fall into the prediction interval of that p. Those samples means in the interval are consistent to that p.
  - we can use other precentages to define prediction intervals.

What if the box proportion p is unknown?

- Prediction intervals are useful for predicting $\bar{X}$ when $\boldsymbol{p}$ is **known**.
  - ➡ They are however not directly useful if $\boldsymbol{p}$ is **unknown**.
- Other procedures can be derived from prediction intervals.
- They are based on the idea that if observed value $\bar{\boldsymbol{x}}$ lies in the prediction interval for some $\boldsymbol{p}$, it is **consistent** with that value of $\boldsymbol{p}$ (in the 95% prediction interval sense).

Turning things around

What if the “population” proportion $\boldsymbol{p}$ is **unknown**?

Suppose

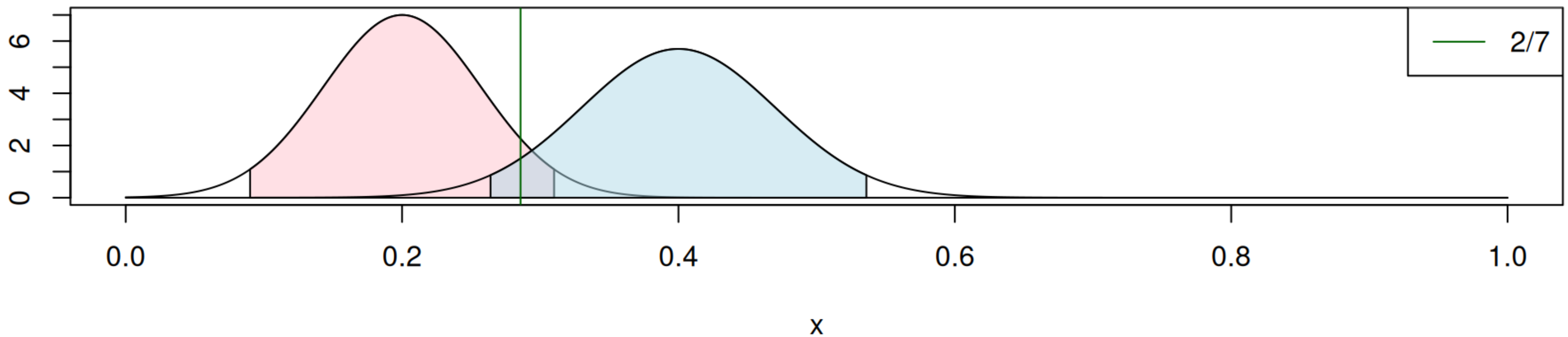
- We have a sample of size $\boldsymbol{n = 49}$ from a box with unknown $\boldsymbol{p}$,
- The observed sample sum is $\boldsymbol{s = 14}$, so that
- The observed sample proportion is $\bar{x} = \frac{s}{n} = \frac{14}{49} = \frac{2}{7} \approx 0.2857$.

We might ask the following question:

Which values of $\boldsymbol{p}$ is this observation consistent with (in the 95% prediction interval sense)?

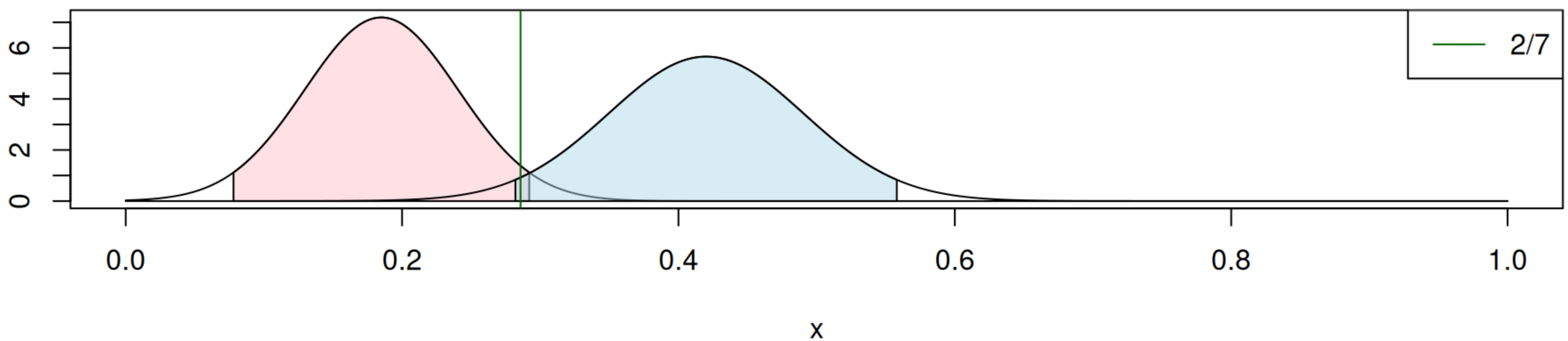
How about both $p = 0.2$ and $p = 0.4$?

- We replicate our graph from before, showing intervals of values consistent with both $p = 0.2$ and $p = 0.4$, when $n = 49$.
- The vertical green line below shows our observed value $\bar{x} = \frac{2}{7}$.



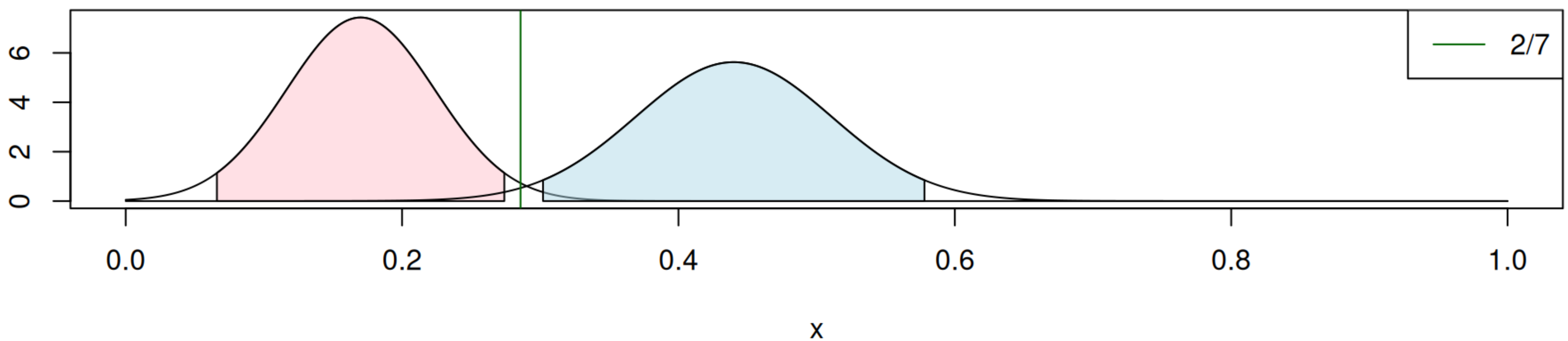
- Note that $\bar{x} = \frac{2}{7}$ is consistent with both $p = 0.2$ and $p = 0.4$.
- What other values of p is the observed value $\frac{2}{7}$ consistent with (in this sense)?

How about both $p = 0.185$ and $p = 0.42$?



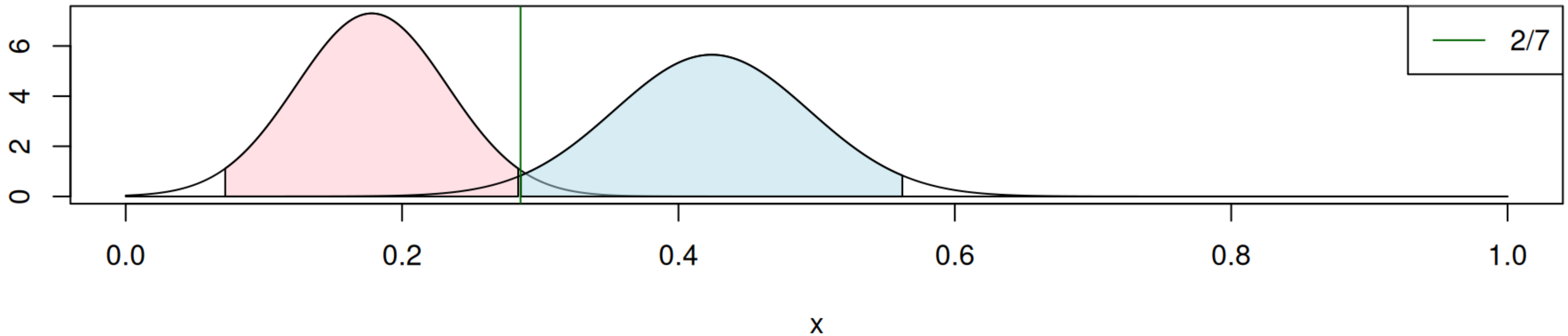
- Note that $\bar{x} = \frac{2}{7}$ is consistent with both $p = 0.185$ and $p = 0.42$.

How about both $p = 0.17$ and $p = 0.44$?



- Note that $\bar{x} = \frac{2}{7}$ is not consistent with both $p = 0.17$ and $p = 0.44$.

Let check what happen if $p = 0.178$ and $p = 0.424$



- For $p = 0.178$, the observed $\bar{x}$ falls on the upper boundary of its 95% prediction interval.
    - ➠ $\bar{x}$ is inconsistent with $p < 0.178$ (outside of its prediction interval).
- For $p = 0.424$ the observed $\bar{x}$ falls on the upper boundary of its 95% prediction interval.
    - ➠ $\bar{x}$ is inconsistent with $p > 0.424$ (outside of its prediction interval).
- 0.178 and 0.424 are "lower" and "upper" values of p so that $\bar{x}$ is consistent with such p's (just on the edge of their prediction intervals).
    - ➠ **For the observed** $\bar{x} = \frac{2}{7}$, $[0.178, 0.424]$ form the **95% confidence interval** for unknown p.

Confidence interval for p

- A **95% confidence interval** for an unknown proportion, based on an observation $\bar{x}$, consists of all values p consistent with $\bar{x}$
  - ➟ $\bar{x}$ lies in the 95% prediction interval for such a p, that is

$$p - 1.96\sqrt{\frac{p(1-p)}{n}} \leq \bar{x} \leq p + 1.96\sqrt{\frac{p(1-p)}{n}}\,.$$

  - ➟ this is called a Wilson's confidence interval for unknown proportion
  - ➟ it is a two-sided interval
- It is thus given by the set

$$\left\{ p : -1.96 \leq \frac{\bar{x} - p}{\sqrt{\frac{p(1-p)}{n}}} \leq 1.96 \right\}.$$

- Endpoints of this interval can be obtained by solving a quadratic equation
  - ➟ We shall not spend time on this, not for assessment
- It's easier to **use the R commands** (which use `s` $= n\bar{x}$ is the sample *sum* as input):

The R `binom` package

- The R package `binom` computes these endpoints using the `binom.confint()` function.
- In our case, we compute the endpoints as follows:

```
require(binom)  # this makes sure the binom package is loaded
binom.confint(x = 14, n = 49, method = "wilson")  # note here the argument 'x' is the sample sum or count
```

```
  method  x  n      mean     lower     upper
1 wilson 14 49 0.2857143 0.1784959 0.4240888
```

- The argument `x = ...` of `binom.confint` is the sample sum or count
- This shows us the "extreme" values of p for which $\bar{x} = \frac{2}{7} \approx 0.2857$ is in the 95% prediction interval are $p = 0.178$ and $p = 0.424$.
- As a "sanity check", we can easily check this

```
0.178 + 1.96 * sqrt(0.178 * 0.822/49)  # upper endpoint of values consistent with 0.178
```

```
[1] 0.2851036
```

```
0.424 - 1.96 * sqrt(0.424 * 0.576/49)  # lower endpoint of values consistent with 0.424
```

```
[1] 0.2856267
```

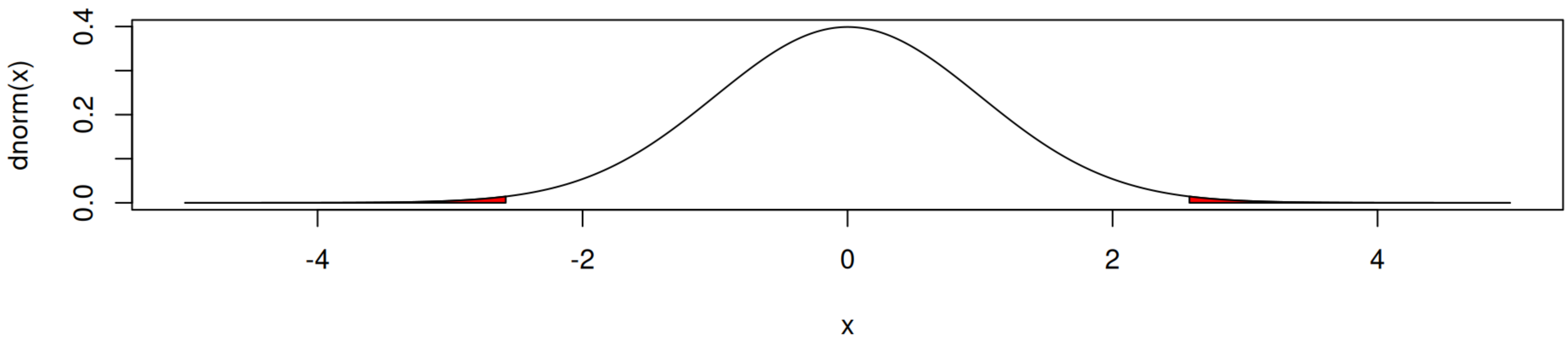
Different confidence levels

- We can change the confidence level by replacing 1.96 with another value.
- E.g., for 99% we should replace `1.96` with

```r
qnorm(0.995)
```

```
[1] 2.575829
```

(which gives 0.5% in the upper tail under the standard normal curve).



Changing confidence level using `binom.confint()`

- Using `binom.confint()` we simply set the `conf.level=` argument to the desired level:

```r
binom.confint(14, 49, conf.level = 0.99, method = "wilson")
```

```
  method  x  n      mean     lower     upper
1 wilson 14 49 0.2857143 0.1531828 0.4693562
```

- As a *sanity check*, we can manually verify that the observed value $\frac{2}{7} \approx 0.2857\ldots$ is "right on the edge" for each of the endpoints 0.153 and 0.469, using the larger multiplier 2.576:

```r
0.469 - 2.576 * sqrt(0.469 * (1 - 0.469)/49)
```

```
[1] 0.285354
```

```r
0.153 + 2.576 * sqrt(0.153 * (1 - 0.153)/49)
```

```
[1] 0.2854754
```

Interpreting the confidence interval

The confidence interval is random (depend on a sample)!

- Suppose there is a true proportion p_* of $\boxed{1}$s, but with a value unknown to us, we can only observe sample means/proportions that are generated by this true proportion p_*.
  - ➟ p_* is not random here – unknown truth for the population.
- For each observed $\bar{x}$, we construct a 95% confidence interval of p's. This gives us an **interval estimate** of the true proportion p_*.
  - ➟ The confidence interval is random since it depends on the observed value of $\bar{x}$.

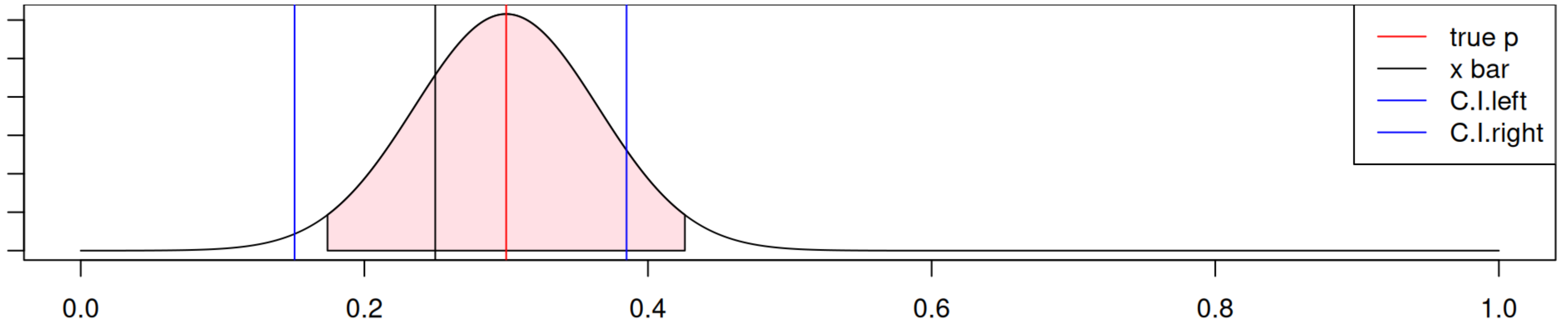
Important notes:

- Under repeated sampling from a 0-1 box, the 95% Wilson confidence interval covers the fixed "true" proportion p_* in (approx.) 95% of samples.
- This is a long-run property of the procedure.
- We don't say with a 95% chance p_* will fall into the confidence interval, as p_* is fixed.
  - ➟ For a *single* data set, we don't know if it has covered the true value or not.
  - ➟ We just know that the procedure you have used is 95% reliable in the long run.

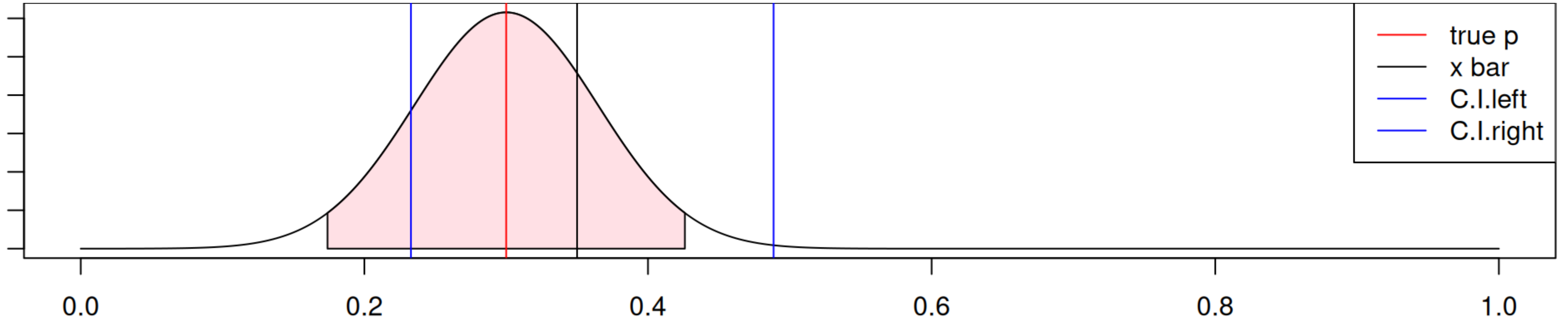
Why the 95% confidence interval covers p_* with 95% chance

Take $p_* = 0.3$ and $n = 50$, its 95% prediction interval is $I_* = [0.173, 0.427]$

- observed value $\bar{x} = 0.25$ in I_* and its confidence interval contains p_*



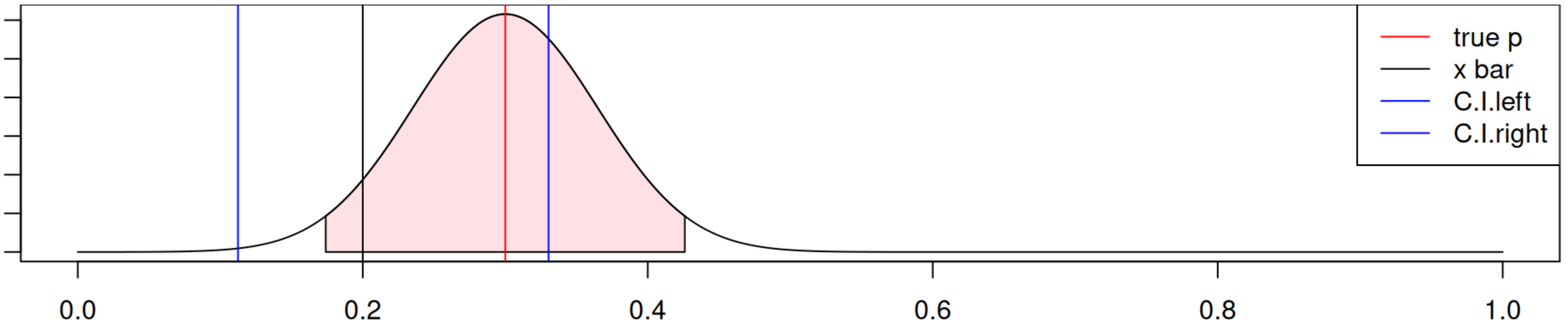
- observed value $\bar{x} = 0.35$ is in I_* and its confidence interval contains p_*



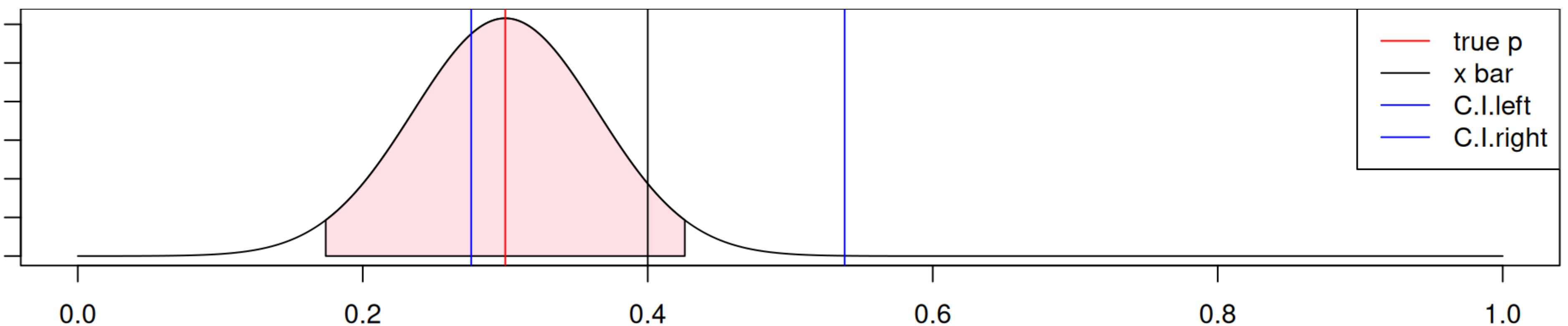
More examples

$p_* = 0.3$ and $n = 50$, its 95% prediction interval is $I_* = [0.173, 0.427]$

- observed value $\bar{x} = 0.2$ is in I_* and its confidence interval contains p_*



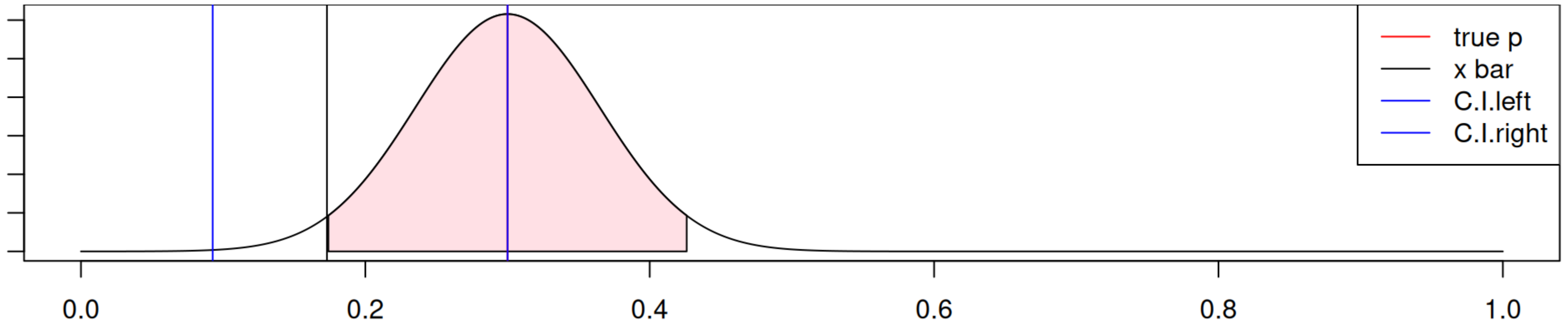
- observed value $\bar{x} = 0.4$ is in I_* and its confidence interval contains p_*



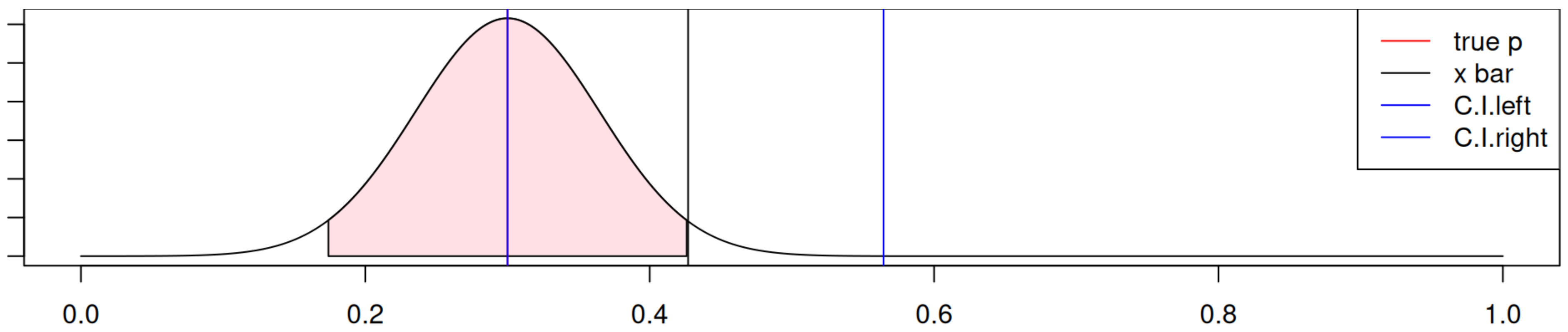
Boundary cases

$p_* = 0.3$ and $n = 50$, its 95% prediction interval is $I_* = [0.173, 0.427]$

- observed $\bar{x} = 0.173$ is on the edge of I_* and the true value p_* is on the edge of the confidence interval



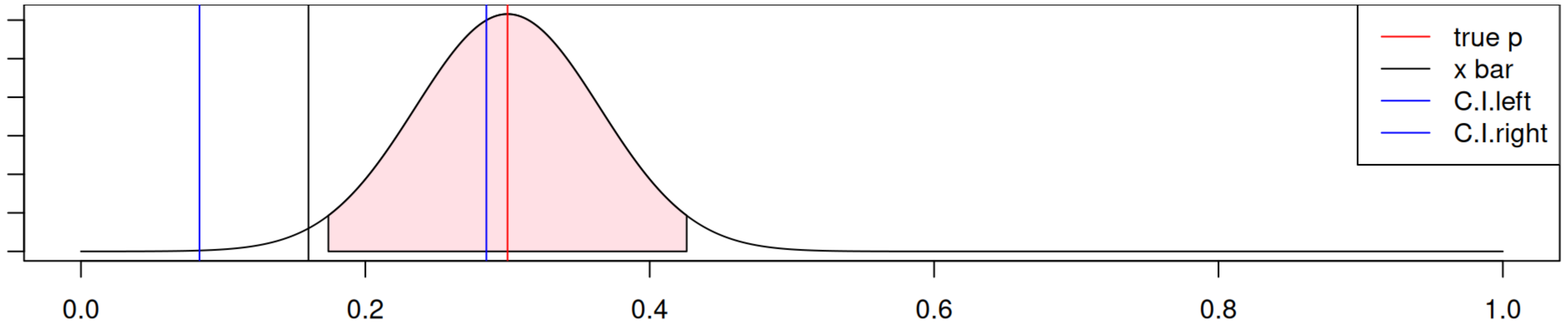
- observed $\bar{x} = 0.427$ is on the edge of I_* and the true value p_* is on the edge of the confidence interval



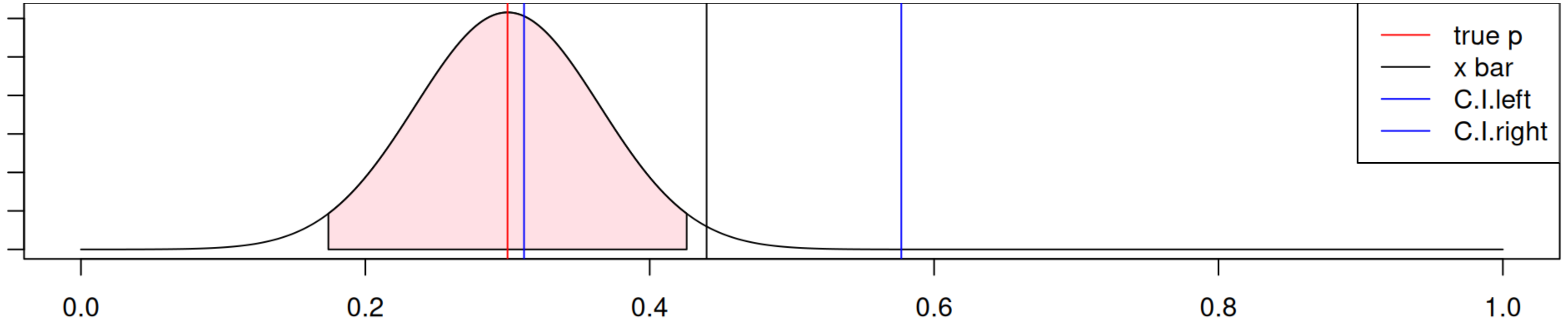
Data outside of I_*

$p_* = 0.3$ and $n = 50$, its 95% prediction interval is $I_* = [0.173, 0.427]$

- observed $\bar{x} = 0.16$ is outside of I_* and the true value p_* is outside of the confidence interval



- observed $\bar{x} = 0.44$ is outside of I_* and the true value p_* is outside of the confidence interval



Quick summary

- For a sample mean $\bar{x}$ in the 95% prediction interval of the (unknown) true proportion p_*
  - ➟ its confidence interval covers the true proportion p_*
  - ➟ this is given by the definition of the confidence interval – covering all p values consistent with $\bar{x}$ (including p_* in this case)
- The chance the sample mean $\bar{X}$ falling into the 95% prediction interval of the (unknown) true proportion p_* is 95%
  - ➟ so 95% of the associated confidence intervals cover the true proportion p_*

Demonstration with random sampling

- Let's see how the Wilson confidence interval works when repeatedly sampling from a box with a known p

```r
is.in.ci = function(truep, n) {
    samp = sample(c(0, 1), prob = c(1 - truep, truep), replace = T, size = n)
    s = sum(samp)
    c.i = binom.confint(s, n, method = "wilson")  # calculate the c.i.
    return(truep >= c.i$lower & truep <= c.i$upper)  # check if true p is in c.i.
}
```

```r
truep = 0.3
n = 50
results = replicate(1000, is.in.ci(truep, n))
sum(results)/1000
```

```
[1] 0.948
```

We see that close to 95% of the time, the interval covers the "true" value of $p = 0.3$.

Case study

Rainfall

- The file `march2024.csv` has daily weather observations from the Canterbury Racecourse weather station for March 2024.

```r
mar.2024 = read.csv("data/march2024.csv", skip = 5)
str(mar.2024)
```

```
'data.frame':	31 obs. of  22 variables:
 $ X                                : logi  NA NA NA NA NA NA ...
 $ Date                             : chr  "2024-03-1" "2024-03-2" "2024-03-3" "2024-03-4" ...
 $ Minimum.temperature..degC.       : num  21.6 23.2 16.6 19.9 14.1 15.2 17.9 20.6 16.1 16.9 ...
 $ Maximum.temperature..degC.       : num  27.9 24.6 32.8 22.5 25.7 29.5 26.9 29.3 29.2 29.3 ...
 $ Rainfall..mm.                    : num  0 0 1 0.2 0 0 0 0 0 0 ...
 $ Evaporation..mm.                 : logi  NA NA NA NA NA NA ...
 $ Sunshine..hours.                 : logi  NA NA NA NA NA NA ...
 $ Direction.of.maximum.wind.gust.  : chr  "SSE" "SSE" "SSE" "SSE" ...
 $ Speed.of.maximum.wind.gust..km.h.: int  37 43 37 44 39 30 46 37 35 46 ...
 $ Time.of.maximum.wind.gust        : chr  "23:01" "08:42" "16:57" "09:23" ...
 $ X9am.Temperature..degC.          : num  23.5 24.6 21.8 20.7 20.3 21.6 24.3 24.8 23.3 23 ...
 $ X9am.relative.humidity....       : int  85 80 80 59 61 73 83 76 76 84 ...
 $ X9am.cloud.amount..oktas.        : logi  NA NA NA NA NA NA ...
 $ X9am.wind.direction              : chr  "S" "SSE" "NW" "SSE" ...
 $ X9am.wind.speed..km.h.           : chr  "6" "20" "7" "19" ...
 $ X9am.MSL.pressure..hPa.          : logi  NA NA NA NA NA NA ...
 $ X3pm.Temperature..degC.          : num  27.4 22.1 30.8 21.6 24.6 27.2 26.3 28.2 28.6 28.2 ...
 $ X3pm.relative.humidity....       : int  68 91 37 60 46 57 71 53 45 45 ...
 $ X3pm.cloud.amount..oktas.        : logi  NA NA NA NA NA NA ...
```

Rainfall

```r
mar.2024$Rain
```

```
 [1]  0.0  0.0  1.0  0.2  0.0  0.0  0.0  0.0  0.0  0.0   NA  0.2   NA  0.0  4.2
[16]  1.0 35.6  1.2  6.2  0.2  0.6  0.0  0.2  0.2  0.2  0.0  0.0  0.0  0.0  0.0
[31]  0.0
```

- What proportion of days in March have rain?
- Suppose we can model the presence or absence of rain as being like a random sample from a 0-1 box with an unknown proportion p of 1s.
- What is a 95% Wilson confidence interval for p?

```r
rain = na.omit(mar.2024$Rain)
s = sum(rain > 0)
s
```

```
[1] 13
```

```r
binom.confint(s, 31, method = "wilson")
```

```
  method  x  n      mean     lower     upper
1 wilson 13 31 0.4193548 0.2641554 0.5923374
```

- The data is thus consistent with the “true” p being anywhere in the range $(0.26, 0.59)$.