# Unknown Proportions and Means

Decisions with Data | Inference for proportions

STAT5002

*The University of Sydney*

Apr 2025

THE UNIVERSITY OF SYDNEY

# 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: $\chi^2$-test

# Outline

## Last week

- Central Limit Theorem
- Confidence Interval (for unknown proportion)

## Today

- A revision
- More on Confidence Intervals (for unknown mean)
- Hypothesis test for unknown proportion

# Central Limit Theorem

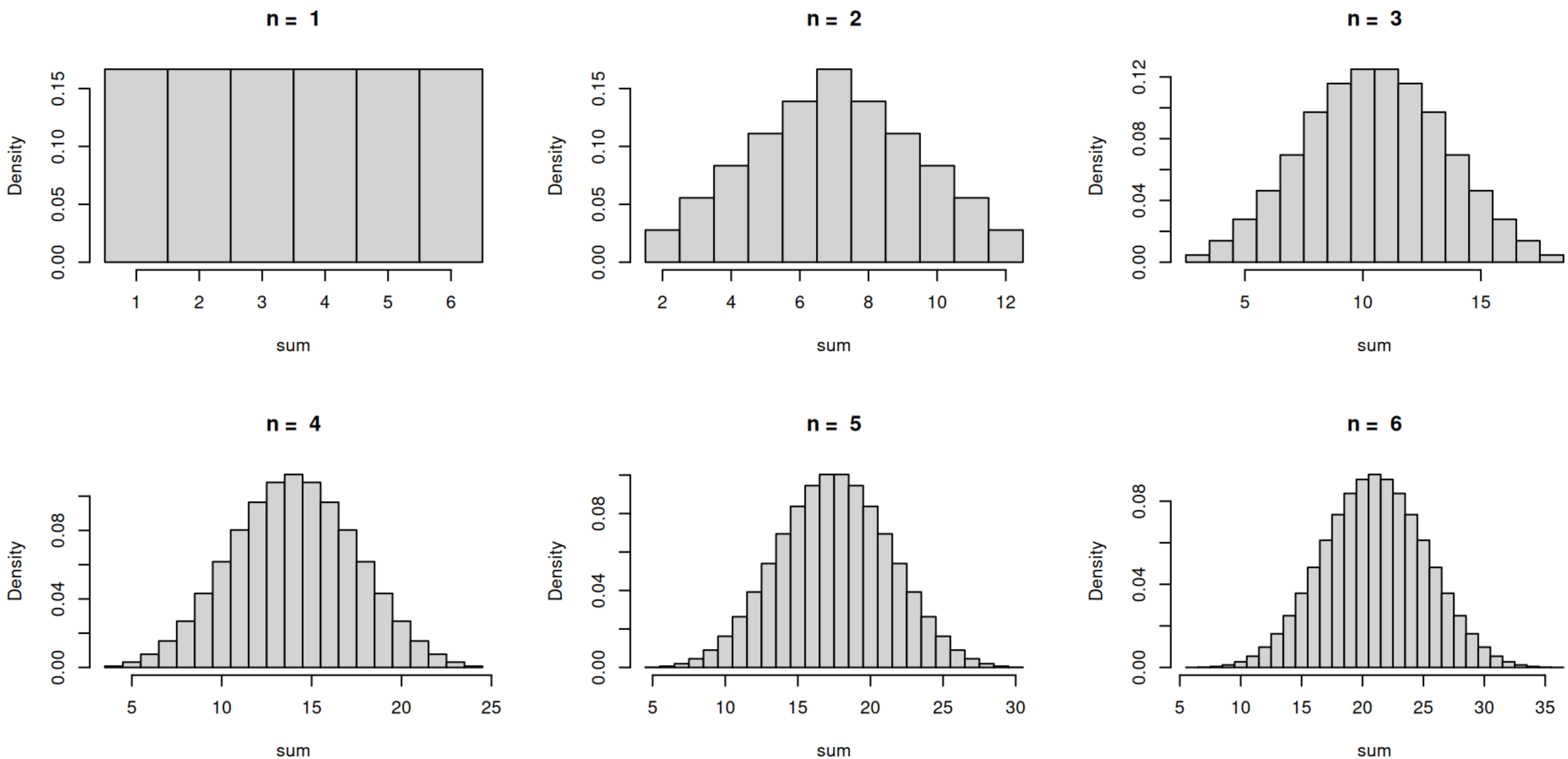
# Example: rolling a 6-sided die

- Suppose we are interested in rolling a 6-sided die $n$ times. How does the sum of the rolls behave?
- This is like taking a random sample of size $n$ from the box

$$\boxed{\boxed{1}\ \boxed{2}\ \boxed{3}\ \boxed{4}\ \boxed{5}\ \boxed{6}}$$

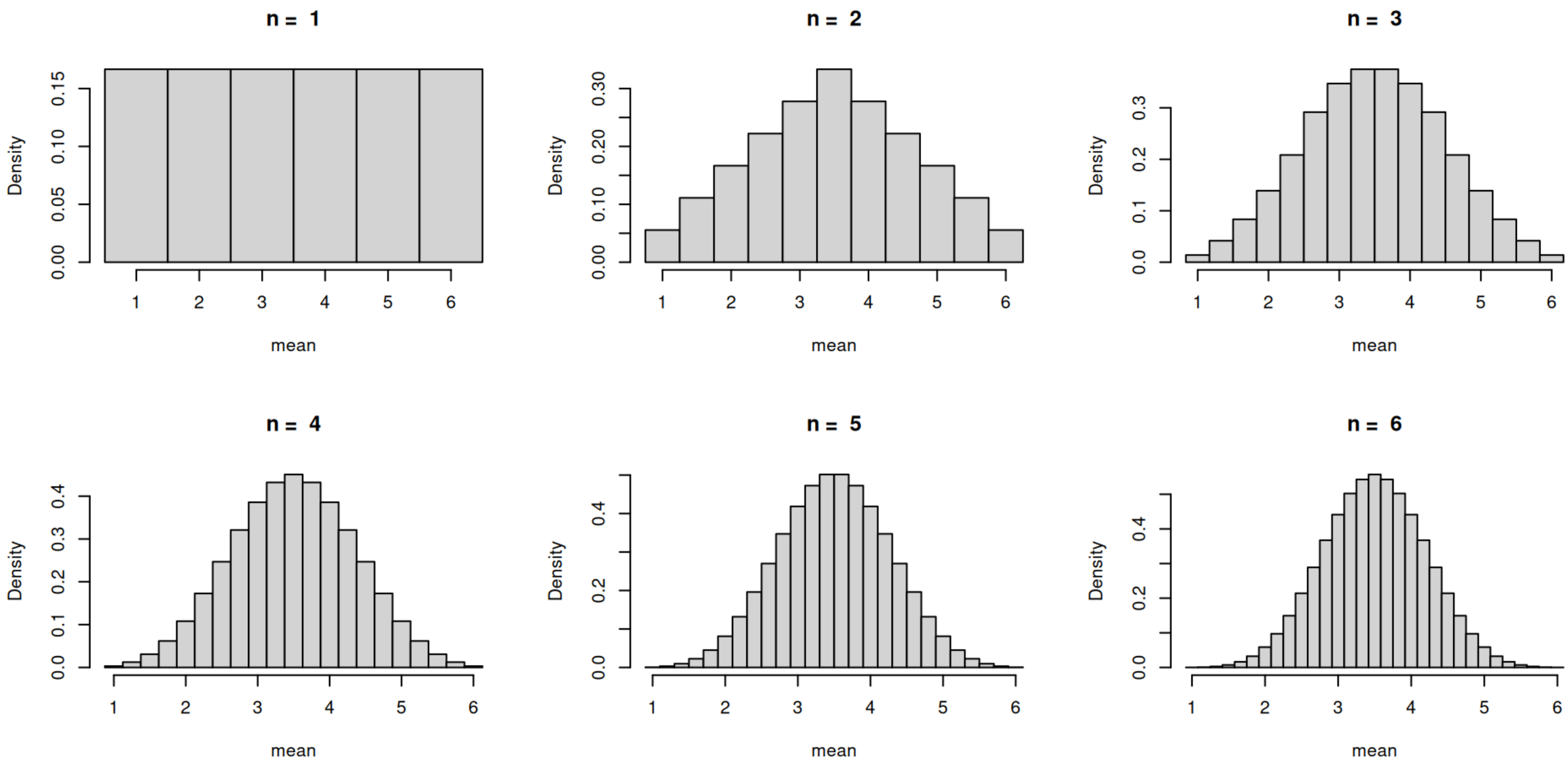
- This box has
  - mean $\mu = 3.5 = \frac{7}{2}$
  - mean square $\frac{1+4+9+16+25+36}{6} = \frac{91}{6}$
  - SD $\sigma = \sqrt{\frac{91}{6} - \left(\frac{7}{2}\right)^2} = \sqrt{\frac{182-(3\times 49)}{12}} = \sqrt{\frac{35}{12}}.$
- We plot the histograms for all possible sums and averages for $n = 1, 2, 3, \ldots$

# Histograms of all possible sums-of-$n$-rolls




For $n = 6$ this is normal-shaped too!

# Histograms of all possible average-of-$n$-rolls




Same shape, but different scaling.

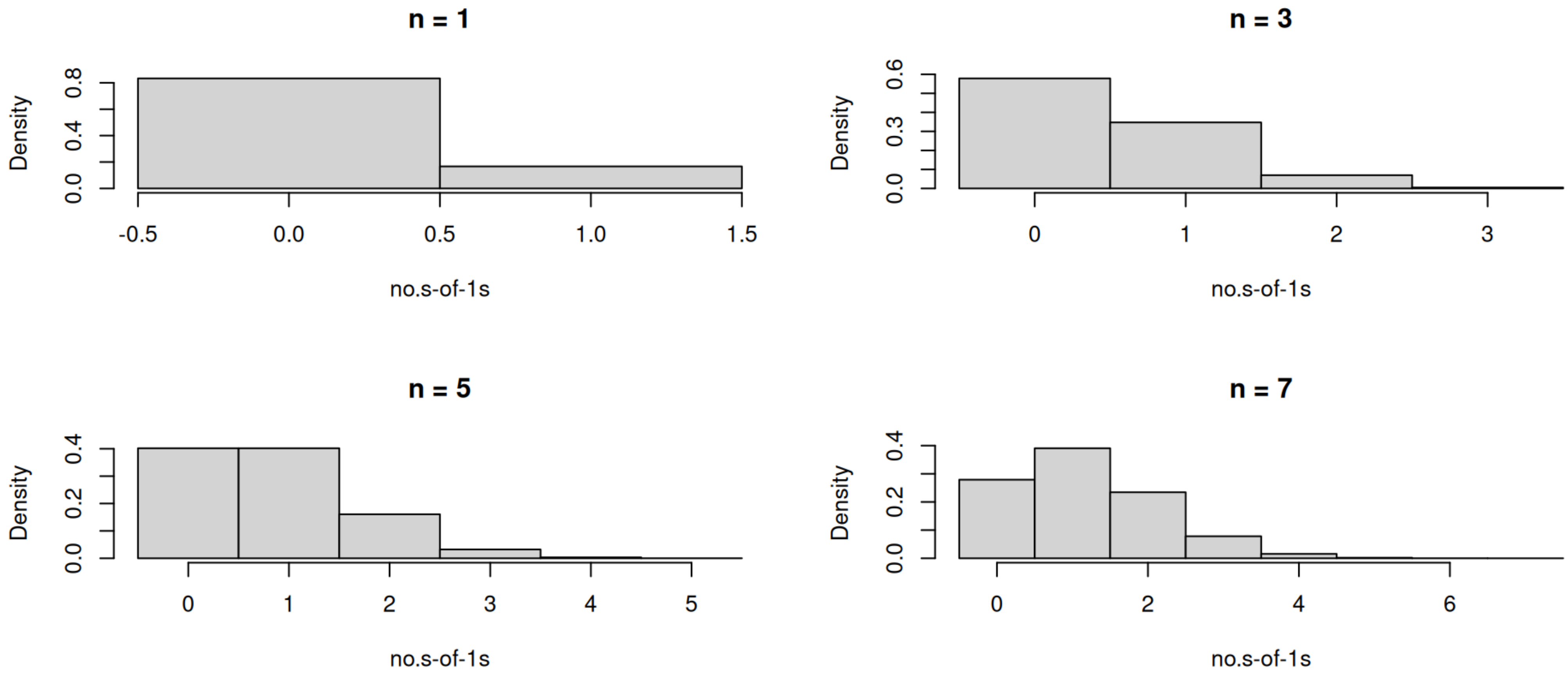
# Asymmetric example

- Consider taking random draws from the following box

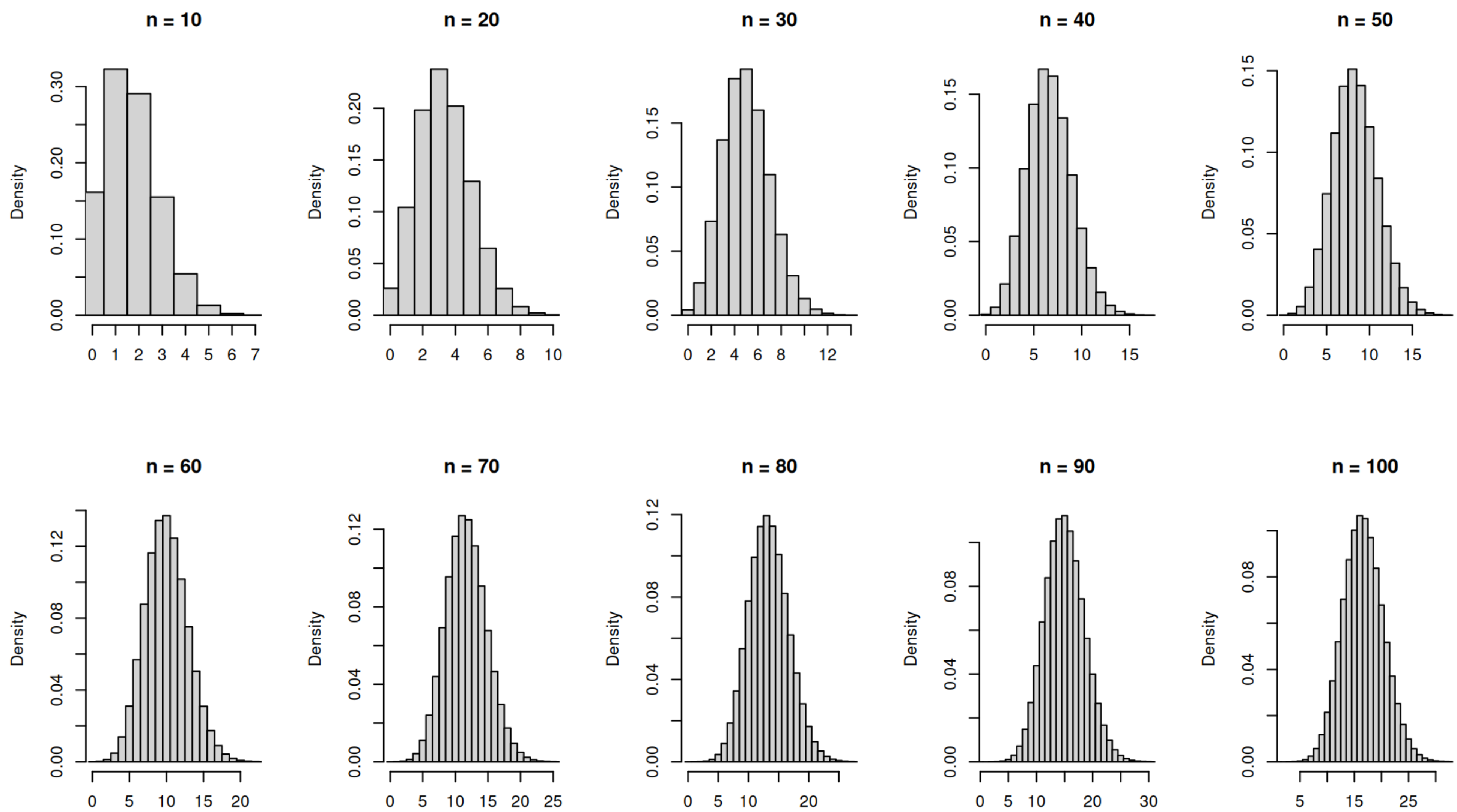
$$\boxed{\boxed{0}\,\boxed{0}\,\boxed{0}\,\boxed{0}\,\boxed{0}\,\boxed{1}}$$

- The number of 1s we get in $n$ random draws from this box is just the sample sum $S$.
- This new box has
    - mean $\mu = \frac{1}{6}$
    - mean square $\frac{1}{6}$
    - SD $\sigma = \sqrt{\frac{1}{6} - \left(\frac{1}{6}\right)^2} = \sqrt{\frac{6-1}{36}} = \frac{\sqrt{5}}{6}$.
- We plot the histograms for all possible sums and averages for $n = 1, 2, 3, \ldots$

# Histograms of all possible no.s-of-1s




Not looking very normal-shaped…what about if we let $n$ get larger?




Again, become normal-shaped as $n$ gets larger

# The Central Limit Theorem

In both cases, the histogram of the sums eventually became normal-shaped with increasing sample size $n$

- For the sysmetric case, this happened for a rather small $n$
- For the unbalanced (assymetric) case, the histograms of all possible sums ("no.s-of-times-we-get-$\boxed{1}$") are not normal-shaped for smaller $n$, as $n$ increases the shape gets closer to a normal.
- This generally holds for box models.

# The Central Limit Theorem

- Consider a box model with mean $\mu$ and SD $\sigma$. We take independent random draws $X_1, X_2, \ldots, X_n$ from the box (sample with replacement).
- For a sufficiently large sample size $n$
  - ➠ The sample sum $S = X_1 + \cdots + X_n$ follows a normal curve with mean $n\mu$ and SD $\sqrt{n}\sigma$
  - ➠ The sample mean $\bar{X} = \frac{1}{n}S$ will follow a normal curve with mean $\mu$ and SD $\frac{\sigma}{\sqrt{n}}$
- That is, the z-scores (standard units) of sample sum and sample mean follow the standard normal curve for a sufficiently large $n$.
  - ➠ $\Phi(z) =$ `pnorm(z)`

$$P(S \leq s) = \underbrace{P\left(\frac{S - n\mu}{\sigma\sqrt{n}} \leq \frac{s - n\mu}{\sigma\sqrt{n}}\right)}_{\text{standard normal}} \approx \Phi\left(\frac{s - n\mu}{\sigma\sqrt{n}}\right)$$

  - ➠ With $s = nx$

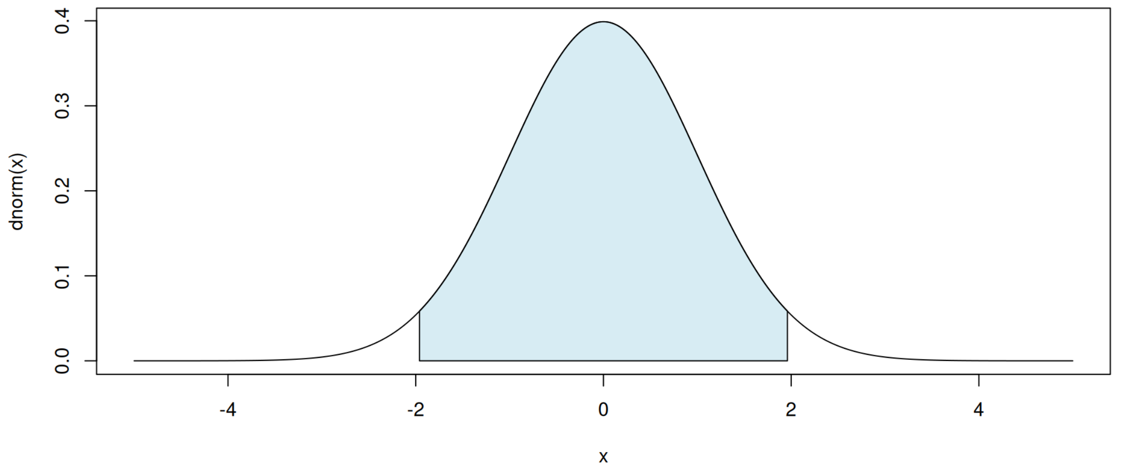
$$P(\bar{X} \leq x) = \underbrace{P\left(\frac{\bar{X} - \mu}{\sigma/\sqrt{n}} \leq \frac{x - \mu}{\sigma/\sqrt{n}}\right)}_{\text{standard normal}} = \Phi\left(\frac{x - \mu}{\sigma/\sqrt{n}}\right) = \Phi\left(\frac{s - n\mu}{\sqrt{n}\sigma}\right)$$

# Q & A

- Is z-score alway normal shaped?
  - ➠ No. As long as we have mean and SD, we can always define the z-socre (standard unit) of data, regardless of shapes of their distributions.
- Why do we use the standard normal in the CLT rather than `pnorm(s, mean = .., sd = ...)`?
  - ➠ Practically, we want to avoid operating with very small or large numbers (for large sample size $n$).
  - ➠ Before we can easily access a computer, we need lookup tables, which only contains results for the standard normal.
  - ➠ Mathematically, the CLT is a statement about convergence as $n \to \infty$, so we need a fixed object (the standard normal distribution) that the sample mean or sample sum converges to.

# More on the standard normal curve

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

# Prediction interval for sample mean

# General formula

- A $\gamma\%$ (two-sided) prediction interval for the sample mean $\bar{X}$ is an interval $[c, d]$ such that

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

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

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

# 0-1 box (define the proportion $p$)

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

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

- The mean of the box $\mu = \frac{pN}{N} = p$ and the SD of the box is $\sqrt{p(1-p)}$ ,; only depend on $p$.
- Taking $n$ draws from the box, then
  - ➠ $E(\bar{X}) = p$ and $SE(\bar{X}) = \sqrt{\frac{p(1-p)}{n}}$; only depend on $p$ and $n$.
  - ➠ $\bar{X}$ is also the sample proportion of $\boxed{1}$s
- The 95% prediction interval for the sample mean $\bar{X}$ is

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

- **Consistency:** with 95% chance, sample means fall into the prediction interval of that $p$. Those samples means in the interval are consistent to that $p$.

# Confidence interval for unknown proportion

# Estimators for unknown proportion

Suppose the true propotion $p_*$ of a 0-1 box is unknown. Given an observed sample mean (proportion) $\bar{x}$, calculated from $n$ independent draws from the box.

Q: how can we estimate the value of $p_*$?

- **Point estimate**: the chance error $\bar{X} - E(\bar{X})$ gets smaller with increasing $n$.
  - ➠ If $n$ is sufficiently large, an observation $\bar{x}$ can be a good estimation to $E(\bar{X}) = p_*$.
- But $\bar{x}$ is just a point, how confident are we about this estimation?
  - ➠ $\bar{x}$ changing from sample to sample, so this point estimate of $p_*$ contains uncertainty.
  - ➠ We gain more information if we can summarize its uncertainty.

# Interval estimate

- **Confidence interval** provides an **interval estimate** of the unknown propotion $p_*$
  - We use here Wilson's confidence interval for unknown proportion – based on the CLT and normal approximation
  - A (Wilson's) **95% confidence interval** consists of all values $p$ consistent with $\bar{x}$ such that

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

which is equivalent to

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

# The R `binom` package

**Q**: Suppose we observe $\bar{x} = \frac{18}{50}$ with $n = 50$ random draws from the 0-1 box. What is the confidence interval for estimating $p_*$?

- From the definition on the previous page, we can solve a quadratic equation (but let's skip this)
- The R package `binom` computes the confidence interval using the `binom.confint()` function.

```r
require(binom)  # this makes sure the binom package is loaded
binom.confint(x = 18, n = 50, method = "wilson")  # note here the argument 'x' is the sample sum
```

```
  method  x  n mean     lower     upper
1 wilson 18 50 0.36 0.2413875 0.4985898
```

- The argument `x = ...` of `binom.confint` is the sample sum

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

- The unknown truth for the population $p_*$ is **not random!**
- The confidence interval is random since it depends on the observed value of $\bar{x}$.
    - ➟ Under repeated sampling from a 0-1 box, the 95% 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.

# 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] 0.965
```

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 ...
```

# Case study: 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 $(0.26, 0.59)$.

# Unknown mean with known SD

# A new box model

For the 0-1 box, $E(X)$ and $SE(X)$ depending on the proportion $p$. Now we consider a new box model.

- We start with an error box with mean $0$ and SD $\sigma$

$$\boxed{\boxed{e_1}\ \boxed{e_2}\ \cdots\ \boxed{e_M}}$$

- Then we build a box model

$$\boxed{\boxed{x_1 = e_1 + \mu}\ \boxed{x_2 = e_2 + \mu}\ \cdots\ \boxed{x_M = e_M + \mu}}$$

- The mean of the new box is $\mu$ and the SD of the new box is $\sigma$ (same as the error box).
- Taking $n$ independent draws (sample with replacement) from the new box, $X_1, X_2, \ldots, X_n$.
  - ➟ What is the 95% prediction interval for the sample mean $\bar{X} = \frac{1}{n}\sum_i X_i$?

# 95% prediction interval

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

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

- We have $E(\bar{X}) = \mu$ and $SE(\bar{X}) = \frac{\sigma}{\sqrt{n}}$, assuming CLT, the 95% prediction interval is

$$\left[\mu - 1.96\frac{\sigma}{\sqrt{n}}, \mu + 1.96\frac{\sigma}{\sqrt{n}}\right]$$

- For a fixed SD $\sigma$, the interval has a fixed size and only changes with $\mu$.
- Suppose $\mu$ **is unknown** but $\sigma$ **is known**, what is the 95% confidence interval for estimating $\mu$?

# Confidence interval for unknown mean with known SD

- A **95% confidence interval** consists of all values $\mu$ consistent with $\bar{x}$ such that

$$\mu - 1.96\frac{\sigma}{\sqrt{n}} \leq \bar{x} \leq \mu + 1.96\frac{\sigma}{\sqrt{n}}\ .$$

- Which is equivalent to

$$-1.96\frac{\sigma}{\sqrt{n}} \leq \bar{x} - \mu \leq 1.96\frac{\sigma}{\sqrt{n}}$$

and

$$\bar{x} + 1.96\frac{\sigma}{\sqrt{n}} \geq \mu \geq \bar{x} - 1.96\frac{\sigma}{\sqrt{n}}\ .$$

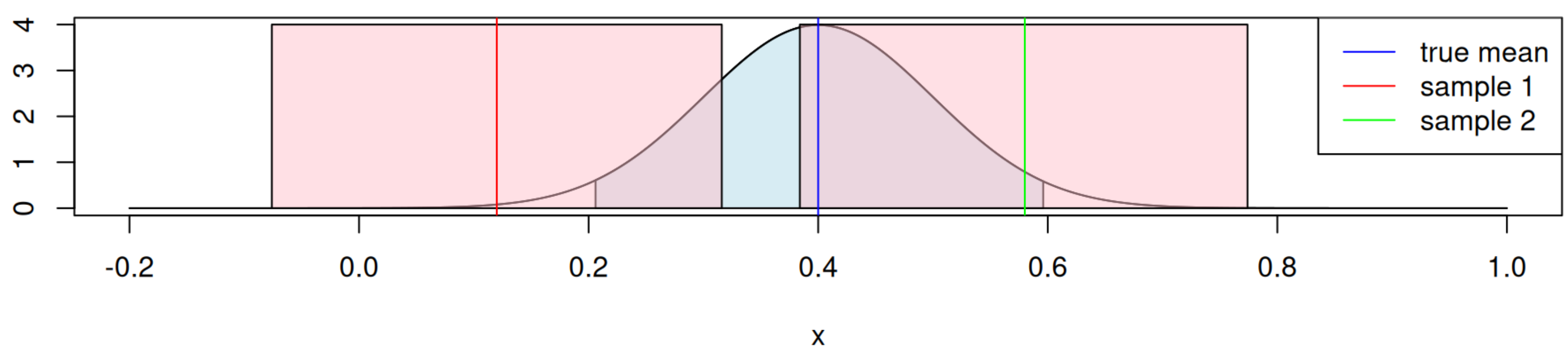
- The 95% confidence interval is $[\bar{x} - 1.96\frac{\sigma}{\sqrt{n}}, \bar{x} + 1.96\frac{\sigma}{\sqrt{n}}]$

# 95% coverage

- It is clear that the 95% confidence interval $[\bar{x} - 1.96\frac{\sigma}{\sqrt{n}}, \bar{x} + 1.96\frac{\sigma}{\sqrt{n}}]$ changes with obervation $\bar{x}$
- Gives a true population mean $\mu_*$
  - Consider any sample mean falls within the 95% prediction interval of $\mu_*$, i.e.,

$$\mu_* - 1.96\frac{\sigma}{\sqrt{n}} \leq \bar{x} \leq \mu_* + 1.96\frac{\sigma}{\sqrt{n}},$$

  we have the distance $|\bar{x} - \mu_*| \leq 1.96\frac{\sigma}{\sqrt{n}}$.
  - The associated confidence interval $[\bar{x} - 1.96\frac{\sigma}{\sqrt{n}}, \bar{x} + 1.96\frac{\sigma}{\sqrt{n}}]$ covers $\mu_*$.



# Example

Consider the box defined by the file `y.dat` in the R code below. We define an error box using `y`

```
y = scan("y.dat")
y
```

```
  [1]  3  4  5  6  7  8  4  5  6  7  8  9  5  6  7  8  9 10  6  7  8  9 10 11  7
 [26]  8  9 10 11 12  8  9 10 11 12 13  4  5  6  7  8  9  5  6  7  8  9 10  6  7
 [51]  8  9 10 11  7  8  9 10 11 12  8  9 10 11 12 13  9 10 11 12 13 14  5  6  7
 [76]  8  9 10  6  7  8  9 10 11  7  8  9 10 11 12  8  9 10 11 12 13  9 10 11 12
[101] 13 14 10 11 12 13 14 15  6  7  8  9 10 11  7  8  9 10 11 12  8  9 10 11 12
[126] 13  9 10 11 12 13 14 10 11 12 13 14 15 11 12 13 14 15 16  7  8  9 10 11 12
[151]  8  9 10 11 12 13  9 10 11 12 13 14 10 11 12 13 14 15 11 12 13 14 15 16 12
[176] 13 14 15 16 17  8  9 10 11 12 13  9 10 11 12 13 14 10 11 12 13 14 15 11 12
[201] 13 14 15 16 12 13 14 15 16 17 13 14 15 16 17 18
```

```
error_box = y - mean(y)
```

The error box has zero mean and a known SD

```
mean(error_box)  # zero mean
```

```
[1] 0
```

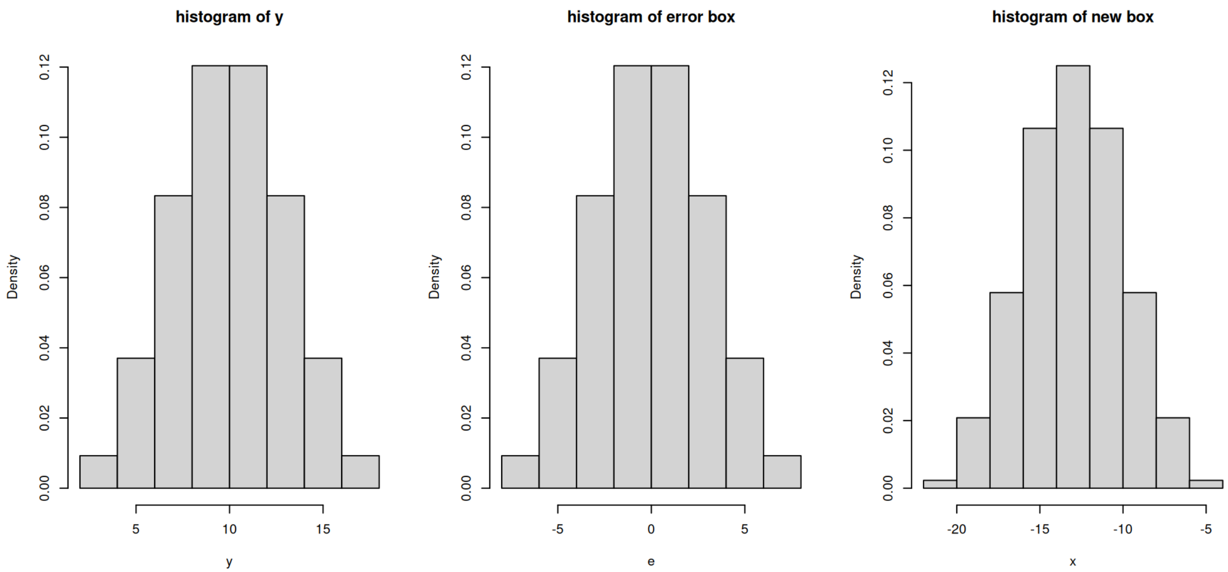
```
sigma = sqrt(mean(error_box^2) - mean(error_box)^2)
sigma
```

```
[1] 2.95804
```

We shift the values of each tickets in `error_box` by a random number to create a `new_box`

```
set.seed(123)
shift = -round(runif(1, 10, 20), 0)  # a randomly generated integer
new_box = shift + error_box
sd(new_box)  # the SD of the new box is the same as the old one
```

```
[1] 2.964911
```

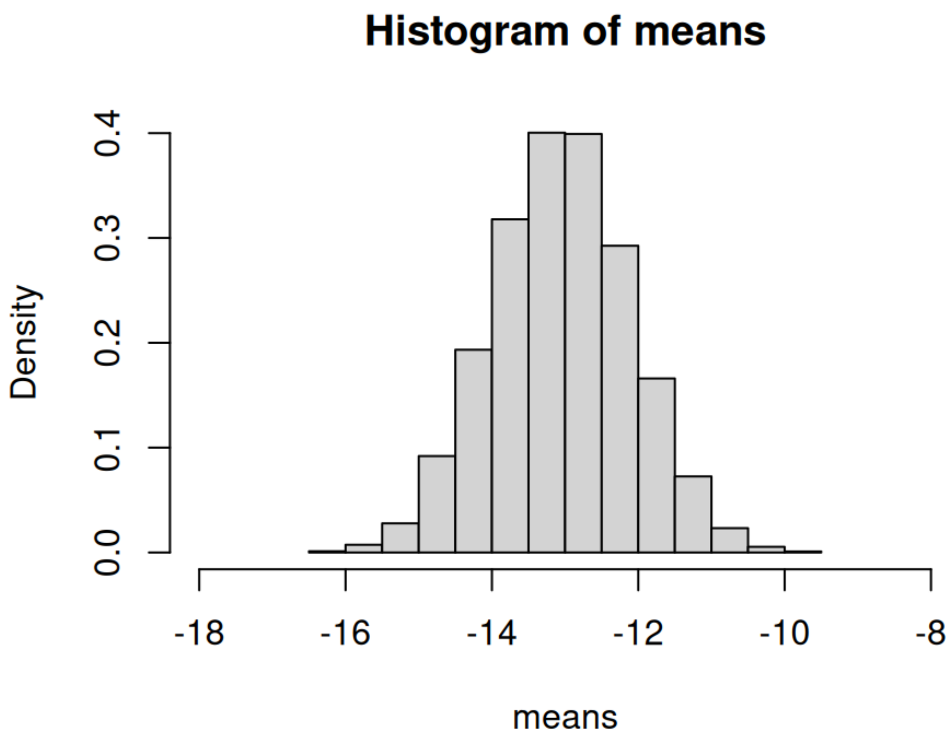


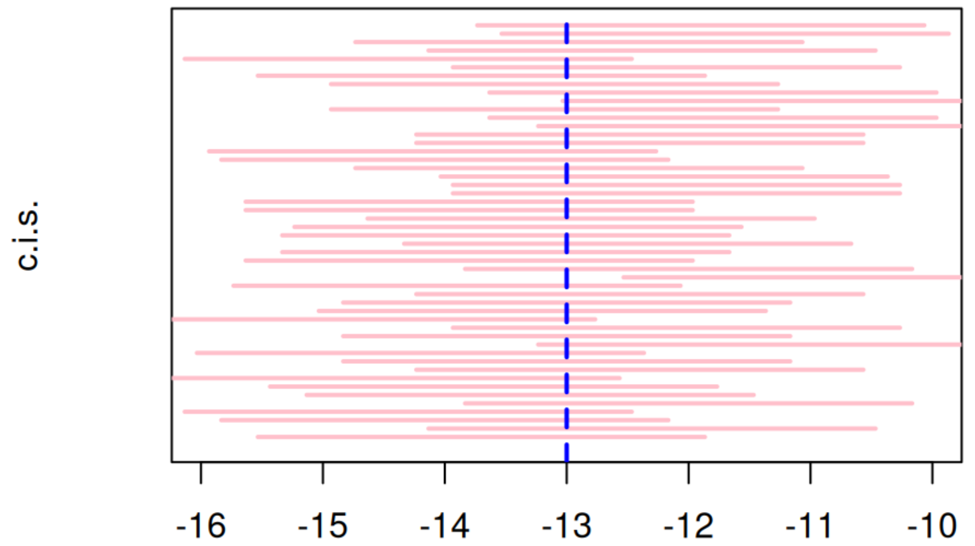

**Draw random samples** from `new_box` and use the 95% confidence interval to estimate its population mean.

```
sample_mean = function(box, n) {
    samp = sample(box, rep = T, size = n)
    return(mean(samp))
}
```

Take $n = 10$ draws:

```r
n = 10
means = replicate(10000, sample_mean(new_box, n))
```

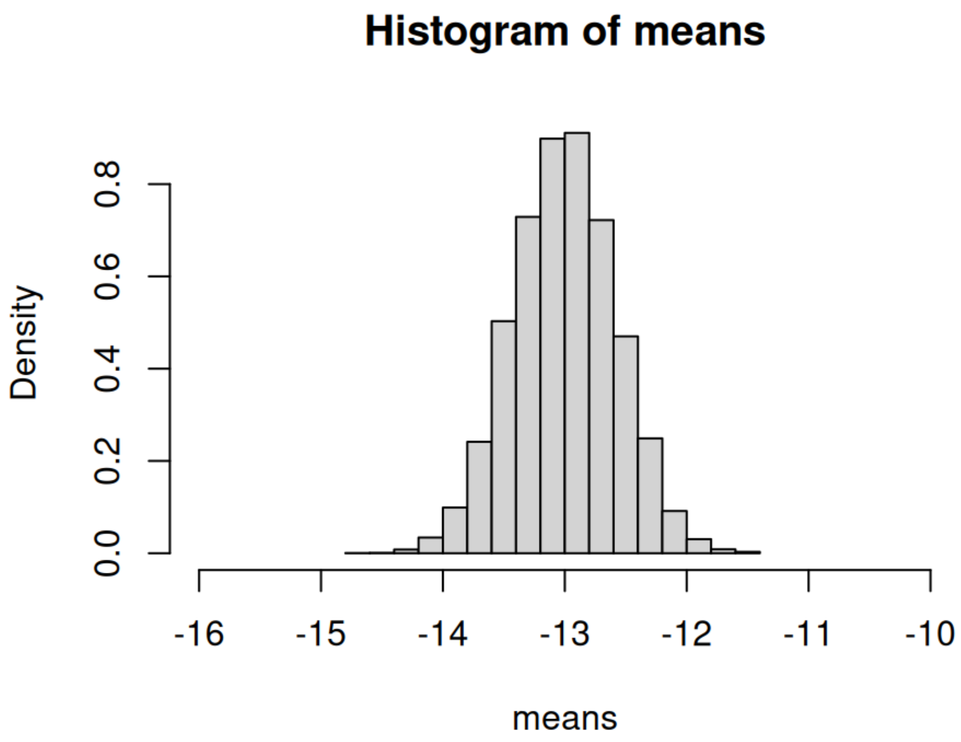


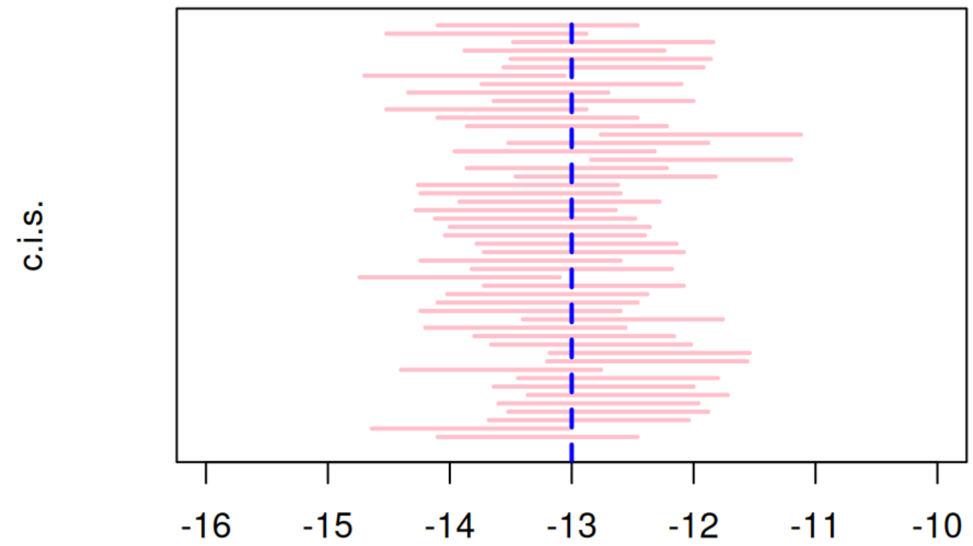




- Left: histogram of sample means.
- Right: confidence intervals (pink) and the true mean (blue line).

Take $n = 100$ draws:

```r
n = 50
means = replicate(10000, sample_mean(new_box, n))
```

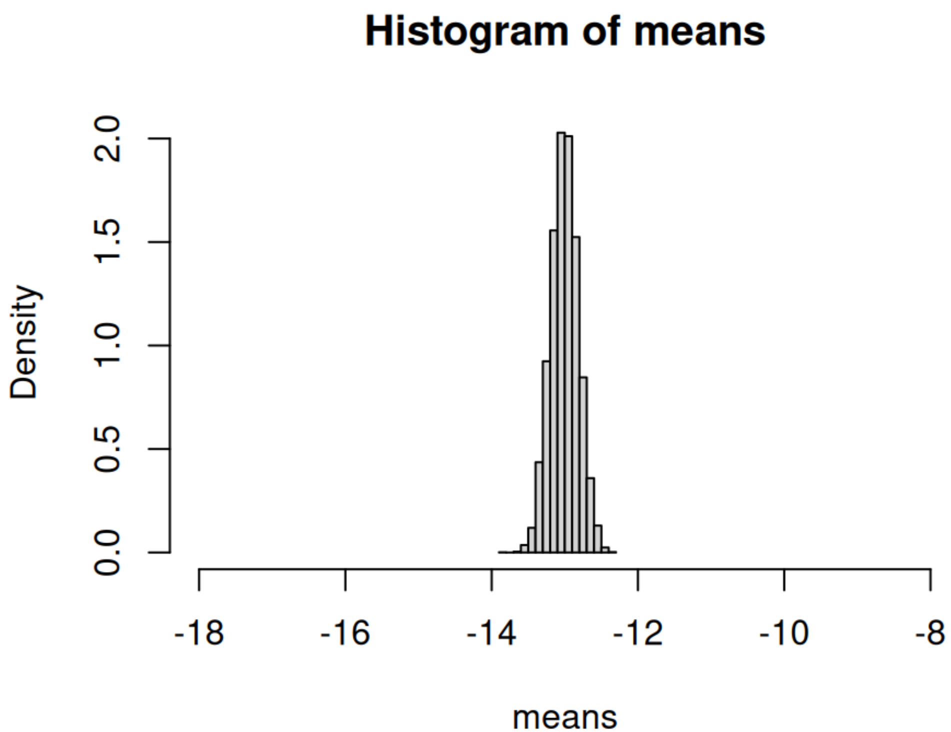


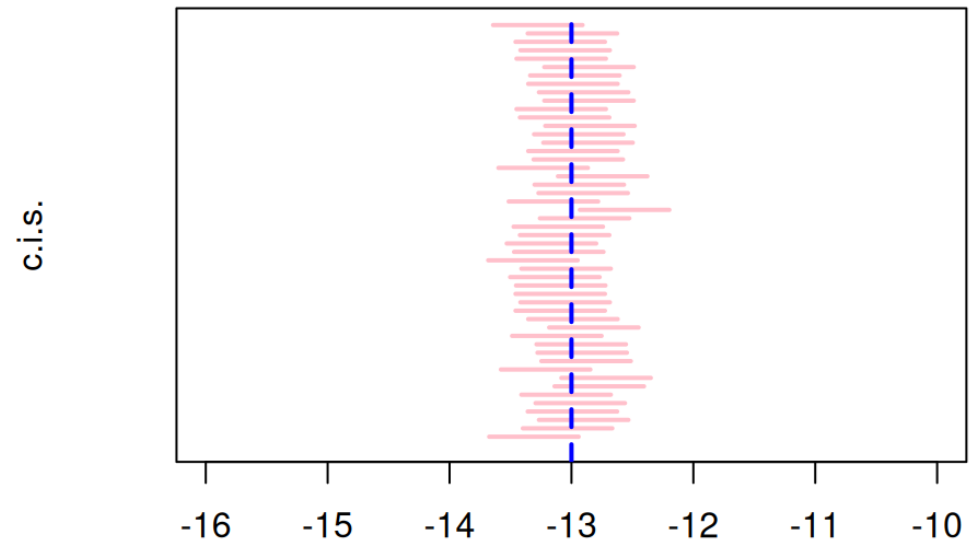




- Left: histogram of sample means.
- Right: confidence intervals (pink) and the true mean (blue line).

Take $n = 1000$ draws:

```
n = 250
means = replicate(10000, sample_mean(new_box, n))
```




- Left: histogram of sample means.
- Right: confidence intervals (pink) and the true mean (blue line).

# Summary

- With inceasing sample size $n$
  - the variations of sample means become smaller ($SE(\bar{X}) = \frac{\sigma}{\sqrt{n}}$)
  - The 95% confidence intervals become narrower $\bar{X} \pm 1.96\frac{\sigma}{\sqrt{n}}$
- The true mean is always fixed!
- Only 95% of confidence intervals covers the true mean
  - We observed some misses.

# Question:

- **Q:** Given a box with a known SD $50$ and a sample mean $\bar{x}$ calculated from a $n = 100$ independent draws from the box. What is the 90% confidence interval?

```r
round(qnorm(0.9), 2)
```

```
[1] 1.28
```

```r
round(qnorm(0.95), 2)
```

```
[1] 1.64
```

# Answer:

- **Q:** Given a box with a known SD $50$ and a sample mean $\bar{x}$ calculated from a $n = 100$ independent draws from the box. What is the 90% confidence interval?

```r
round(qnorm(0.9), 2)
```

```
[1] 1.28
```

```r
round(qnorm(0.95), 2)
```

```
[1] 1.64
```

$$[\bar{x} - 1.64 \times 5, \bar{x} + 1.64 \times 5]$$

- Why 1.64?
  - 5% below $-1.64$ and 5% above $1.64$ under the standard normal curve.
  - the middle 90% percent is bounded between $\pm 1.64$ under the standard normal curve.
- Why times 5? -$SE(\bar{X}) = \frac{\sigma}{\sqrt{n}} = \frac{50}{10} = 5$