MONASH University

MONASH BUSINESS SCHOOL

Statistical Thinking (ETC2420/ETC5242)

Associate Professor Catherine Forbes

Week 5: Resampling techniques for assessing variability in means

Learning Goals for Weeks 4 and 5

- Review the Central Limit Theorem
- Apply one and two sample t-tests and confidence intervals ✓
- Build Bootstrap confidence interval for numerical data
- Distinguish between independent and paired samples ✓

Assigned reading for Week 5:

- Chapter 4 (skip Section 4.4) in ISRS
- Section 19.1-19.6 in R for Data Science (writing functions and if-else statements)

The Central Limit Theorem (CLT)

- CLT describes the **sampling distribution** of $\bar{X}$, as the sample size **increases**
- The (hypothetical) sampling distribution of the sample mean will become normally distributed
  - even if the data from the original population is **not** normally distributed
- F is the population distribution
  - $E[X_i] = \mu$
  - $Var(X_i) = \sigma^2 < \infty$
- The sample mean: $\bar{X} = \frac{1}{n}\sum_{i=1}^{n} X_i$ is a **point estimator** of μ

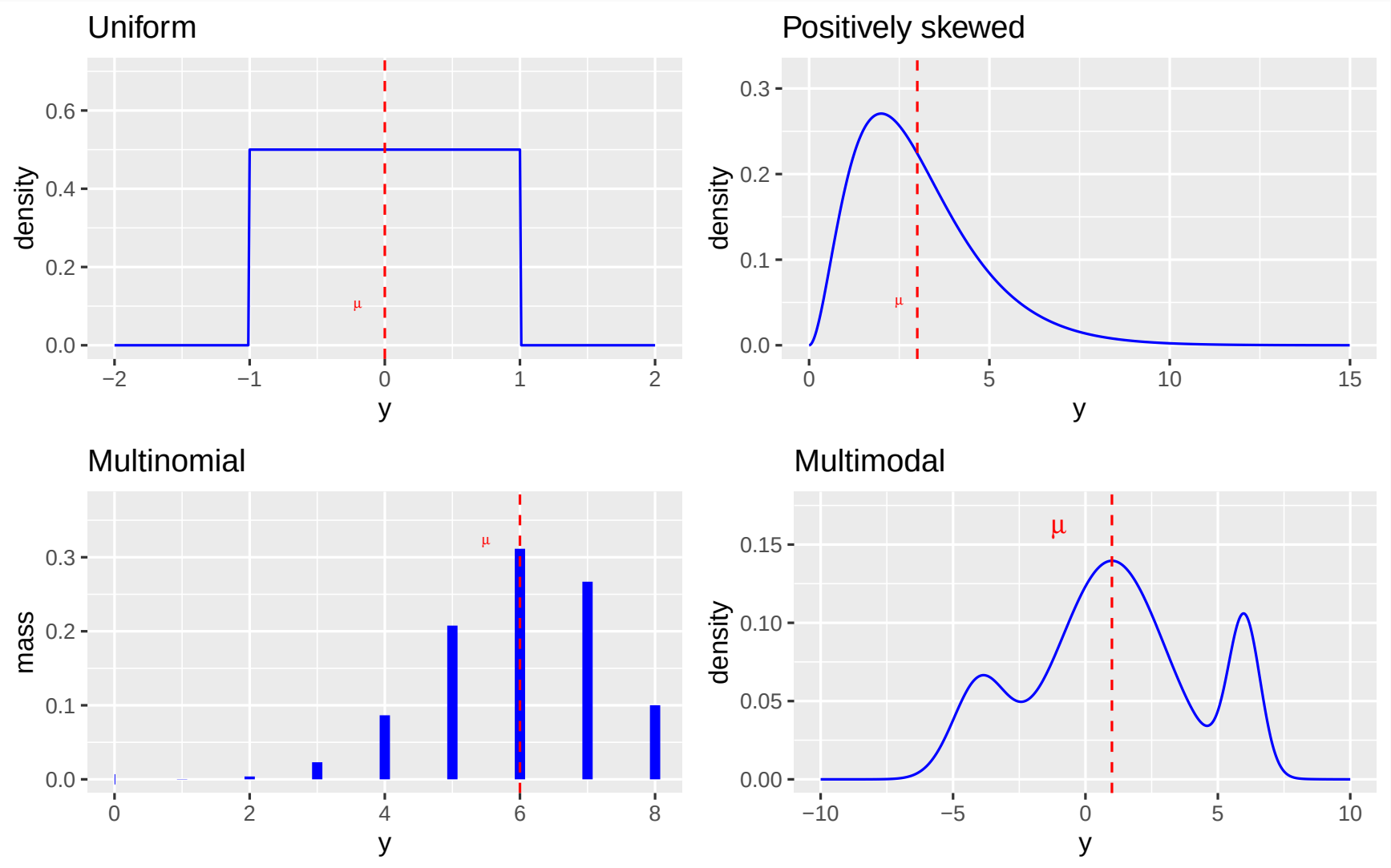
CLT

If $X_1, X_2, \ldots, X_n, \ldots \overset{i.i.d}{\sim} F$, then

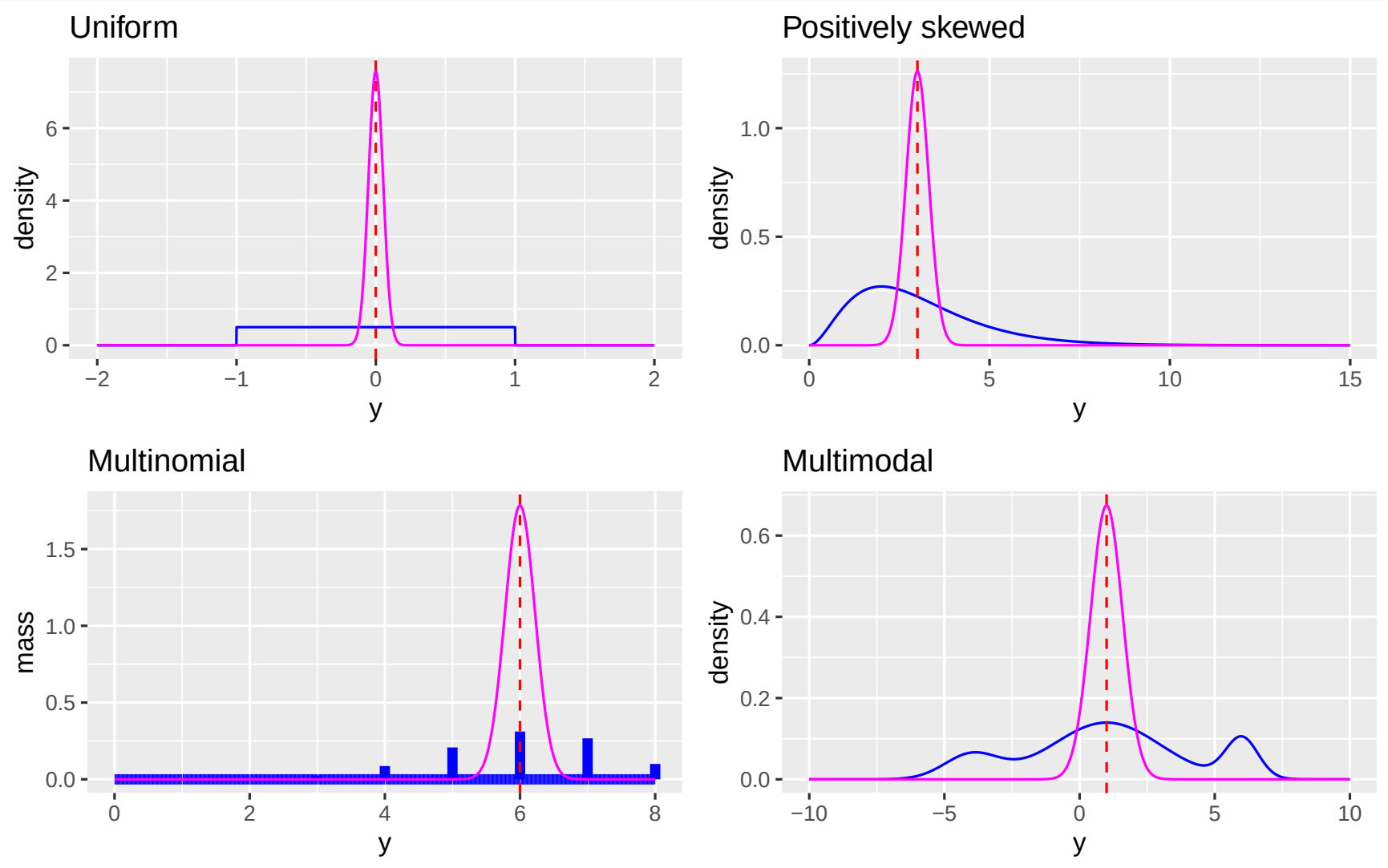
$$\sqrt{n}\left(\bar{X} - \mu\right) \xrightarrow{Dist} N\left(0, \sigma^2\right), \text{ as } n \to \infty$$

- Provides approximate sampling distribution of $\bar{X}$ (for fixed n)
  - $\Rightarrow$ hypothesis test about μ
  - $\Rightarrow$ confidence interval for μ

Some non-normal populations



CLT approximations with n=30



Using the CLT (estimating the SE)

- We don't usually know σ, but we can **estimate** it with the sample standard deviation:

$$s = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(X_i - \bar{X})^2}$$

- And replace $\sigma/\sqrt{n}$ with

$$SE = \frac{s}{\sqrt{n}}$$

- Use approximation:

$$T = \frac{\sqrt{n}\,(\bar{X} - \mu)}{s} = \frac{\bar{X} - \mu}{SE} \overset{approx}{\sim} t_{n-1}$$

- As $n \to \infty$ then $T \xrightarrow{D} N(0, 1)$

Testing hypotheses with the CLT (one sample)

- Use CLT to test $H_0 : \mu = \mu_0$ (= 'null value')
- When H_0 is true: $T_0 = \frac{\bar{X}-\mu_0}{SE} \overset{approx}{\sim} t_{n-1}$ and we test against:

two-sided alternative: $H_1 : \mu \neq \mu_0$

Reject H_0 if $|T_0| \geq t_{n-1,0.975}$

upper one-sided alternative: $H_1 : \mu > \mu_0$

Reject H_0 if $T_0 \geq t_{n-1,0.975}$

lower one-sided alternative: $H_1 : \mu < \mu_0$

Reject H_0 if $T_0 \leq t_{n-1,0.025}$

- Otherwise do not reject H_0 and conclude $\mu = \mu_0$

95% Confidence interval based on T (one sample)

- Start with 95% sampling interval for $\bar{X}$:

$$\Pr\left(t_{n-1,0.025} < \frac{\bar{X} - \mu}{s/\sqrt{n}} < t_{n-1,0.975}\right) = 0.95$$

- Rearrange expression:

$$\Rightarrow \Pr\left(\bar{X} + \frac{s}{\sqrt{n}}\, t_{n-1,0.025} < \mu < \bar{X} + \frac{s}{\sqrt{n}}\, t_{n-1,0.975}\right) = 0.95$$

- 'Plug in' observed: $\bar{X} = \bar{x}_{obs}$ and record observed interval $\Rightarrow$ 95% confidence interval for μ:

$$\left[\bar{x}_{obs} + \frac{s}{\sqrt{n}}\, t_{n-1,0.025},\ \bar{x}_{obs} + \frac{s}{\sqrt{n}}\, t_{n-1,0.975}\right]$$

- Now there is **no probability remaining**! Only **"confidence"**
  - **Before observed, $\bar{X}$ is random**
  - **After observed, $\bar{x}_{obs}$ is no longer random**
  - **μ is always fixed!**

Student-t quantiles for CLT-based confidence intervals

- The notation $t_{df,\alpha}$ refers to the lower α quantile of the student t distribution with df degrees of freedom:

$$\Pr\left(T \leq t_{df,\alpha}\right) = \alpha$$

- If the degrees of freedom df is "large", then $t_{df,\alpha} \approx z_\alpha$, the lower α quantile of the $N(0,1)$ distribution, i.e.
  - $t_{0.025,n-1} \to z_{0.025} = -1.96$ as $n \to \infty$, and
  - $t_{0.975,n-1} \to z_{0.975} = +1.96$ as $n \to \infty$

- In **R**, use
  - `qt(0.025,(n-1))` for $t_{0.025,n-1}$, and `qt(0.975,(n-1))` for $t_{0.975,n-1}$
- And note that
  - `qnorm(0.025)` is $z_{0.025}$, and `qnorm(0.975)` is $z_{0.975}$

Confidence intervals via a “Bootstrap” approach

- **Bootstrap** techniques provide alternative approaches to constructing a confidence interval
- The basic idea: Replicate “hypothetical” data sets (Bootstrap samples) by re-sampling observed values **with replacement**
- There are several Bootstrap approach variations. Here we consider one referred to the **Bootstrap percentile interval** approach

The Bootstrap CI for single population mean, based on $\bar{X}$

An approximate 95% confidence interval for a single population mean is obtained in three steps:

1. Generate a Bootstrap sample of B potential $\bar{X}$ values
   - Denote these as $\{\bar{x}^{[1]}, \bar{x}^{[2]}, \ldots, \bar{x}^{[B]}\}$
   - $B =$ should be a large number (e.g. $B = 1000$)
2. Use the empirical distribution from this Bootstrap sample to approximate the sampling distribution of $\bar{X}$
   - give each $\bar{x}^{[b]}$ equal weight $= 1/B$, and
   - approximate

$$\hat{\Pr}(\bar{X} \leq c) = \frac{\text{number of } [\bar{x}^{[b]} \leq c]}{B}$$

3. Construct an approximate 95% confidence interval by selecting interval from 2. with (empirical) probability (at least) 95%

How to generate a Bootstrap sample

- How to calculate $\bar{x}^{[b]}$?
- For each b in $1 : B$
  - resample n draws from the D_n set, with replacement
  - label these values as $\{x_1^{[b]}, x_2^{[b]}, \ldots, x_n^{[b]}\}$
  - compute the average $\bar{x}^{[b]} = \frac{1}{n} \sum_{i=1}^{n} x_i^{[b]}$

- In **R** use (with replace = TRUE) either:
  - **sample()**, or
  - **slice_sample()**

Resampling in R with sample()

```
a <- c(1:10)
a
```

```
 [1]  1  2  3  4  5  6  7  8  9 10
```

```
mean(a)
```

```
[1] 5.5
```

```
atil <- sample(a, replace = TRUE)
atil
```

```
 [1]  5  8  7  7  2 10  8 10 10  5
```

```
mean(atil)
```

```
[1] 7.2
```

Resampling in R with slice_sample()

```r
df <- tibble(a = c(1:10), b = letters[1:10])
mean(df$a)
```

```
[1] 5.5
```

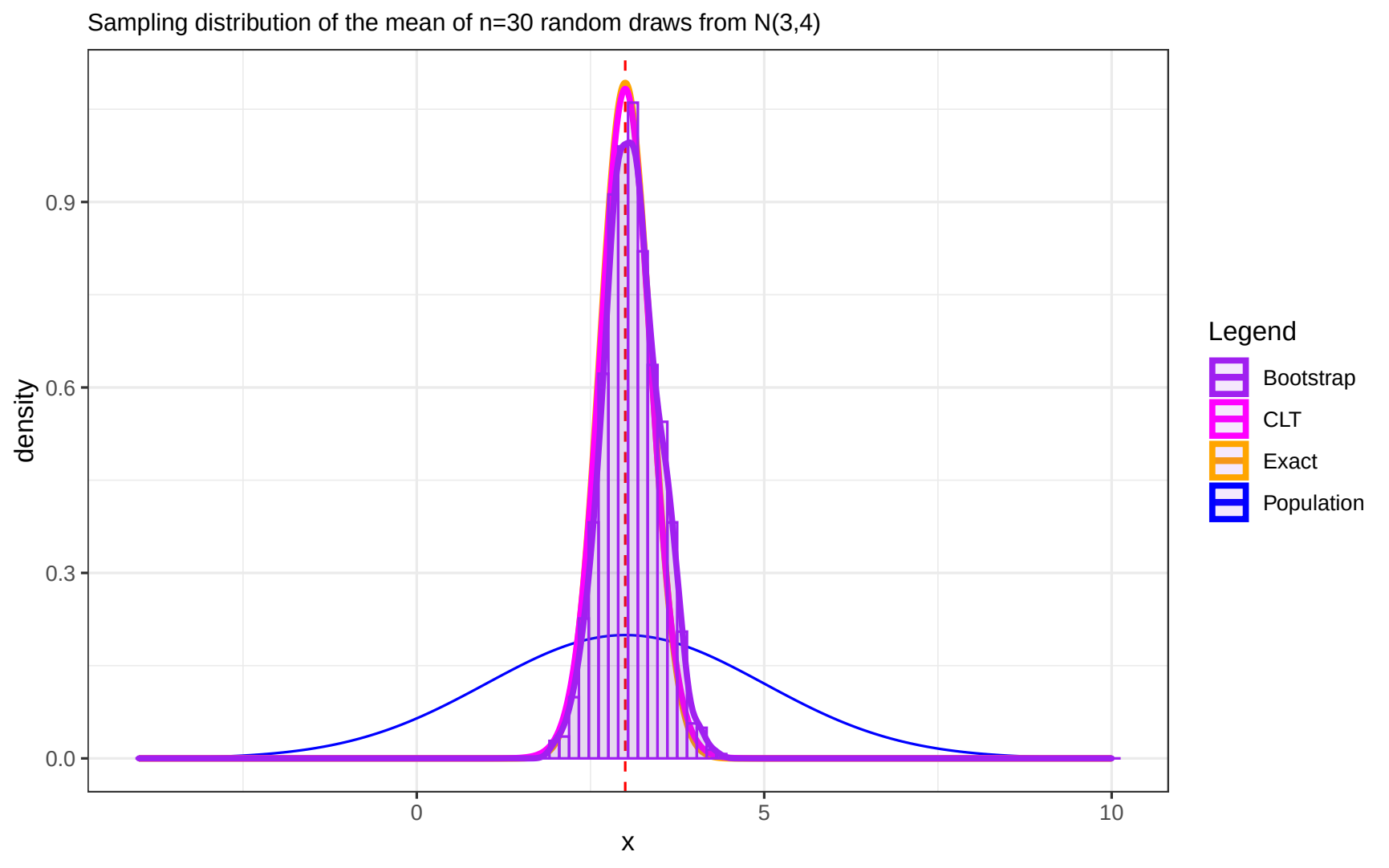
```r
dftil <- slice_sample(df, n = nrow(df), replace = TRUE)
glimpse(dftil)
```

```
Rows: 10
Columns: 2
$ a <int> 5, 8, 7, 7, 2, 10, 8, 10, 10, 5
$ b <chr> "e", "h", "g", "g", "b", "j", "h", "j", "j", "e"
```

```r
mean(dftil$a)
```

```
[1] 7.2
```

A N(3, 4) sample of size n=30



Bootstrap 95% confidence interval

- Take off 2.5% from each tail of the Bootstrap empirical distribution
- Just sort the $\{\bar{x}_{obs}^{[b]}\}$ values and find
  - the lower 2.5% quantile $\Rightarrow L_{\bar{x}_{obs}}$
  - the lower 97.5% quantile $\Rightarrow U_{\bar{x}_{obs}}$
- And then $[L_{\bar{x}_{obs}}, U_{\bar{x}_{obs}}]$ is an approximate 95% confidence interval for μ
- For $N(3,4)$ example: 95% CIs (approximate and exact)

L_{Boot}	U_{Boot}	L_{CLT}	U_{CLT}	L_{Exact}	U_{Exact}
2.34	3.83	2.25	3.90	2.28	3.72

- Exact available since $X_1, X_2, \ldots, X_n \overset{i.i.d.}{\sim} N(3,4)$ then $\bar{X} \sim N(3, \frac{4}{30})$

Testing hypotheses with a Bootstrap approach

- It is possible to use a Bootstrap approach to test $H_0 : \mu = \mu_0$ (= 'null value')
- But the process is more involved and easy to get wrong
- See Hall, P. and Wilson, S. R. (1991) Two Guidelines for Bootstrap Hypothesis Testing, *Biometrics*, Vol. 47, No. 2, pp. 757-762

- $\Rightarrow$ We will not pursue Bootstrap hypothesis tests in this unit.

Bootstrap for paired samples

- Like with the CLT, we can apply the Bootstrap to paired data

$$\{(X_{1,i}, X_{2,i}),\ \text{for } i = 1, 2, \ldots, n\}$$

- First calculate the sample of paired differences:
  $DD_n = \{Diff_i = X_{1,i} - X_{2,i},\ \text{for } i = 1, 2, \ldots, n\}$
- Then apply the **single population Bootstrap** method to the DD_n sample
  - for each b in $1 : B$
    - resample n draws from the DD_n set, with replacement
    - compute the average $\bar{Diff}^{[b]}$
  - Use the empirical sample of $\{\bar{Diff}^{[b]},\ \text{for } b = 1, 2, \ldots, B\}$ to obtain a confidence interval for $\mu_{Diff} = \mu_1 - \mu_2$

Bootstrap for the different in two independent samples

- For unpaired data $D1_{n_1} = \{X_{1,i},\ \text{for}\ i = 1, 2, ..., n_1\}$ and $D2_{n_2} = \{X_{2,j},\ \text{for}\ j = 1, 2, ..., n_2\}$, we can use the Bootstrap to build the relevant confidence interval
- For each b,
  - resample with replacement n_1 observations from $D1_{n_1}$ to produce $\bar{x}^{[b]}_{1,obs}$,
  - resample with replacement n_2 observations from $D2n_2$ to produce $\bar{x}^{[b]}_{2,obs}$, and
  - calculate $(\bar{x}^{[b]}_{1,obs} - \bar{x}^{[b]}_{2,obs})$
- And compute an approximate 95% confidence interval using the lower 2.5% and 97.5% quantiles of $\{(\bar{x}^{[b]}_{1,obs} - \bar{x}^{[b]}_{2,obs}),\ \text{for}\ b = 1, 2, \ldots, B\}$

- Again we will not attempt hypothesis tests using a Bootstrap approach in this setting.

What's the real advantage of the Bootstrap approach?

- Both CLT and Bootstrap approaches **do not** require knowledge of the true underlying population distribution
- Both CLT and Bootstrap approaches are (relatively) **easy** to implement
- However, **the CLT only works for the sampling distribution of** $\bar{X}$ (for single population)
- Whereas **we can apply the Bootstrap approach to any point estimator**!
- e.g. for
  - single population median
  - parameters of assumed models models
  - to assess single population asymmetry: e.g. mean - median
  - differences in the medians of two independent samples
  - and more!
- Use **resampling** of the available **data** to assess uncertainty in the estimator
- (If you are interested in frequentist inference)