ETC5242 Class 1

Confidence Interval

授课老师：Joe

ETC5242平时班 – Class 1

- **Week 5**
  - **Central limited theorem**

  - **Bootstrapping for paired variables**

  - **Bootstrapping for independent variables**

- **Week 6**
  - **Maximum likelihood estimate (MLE)**

  - **Bootstrapping for model parameters**

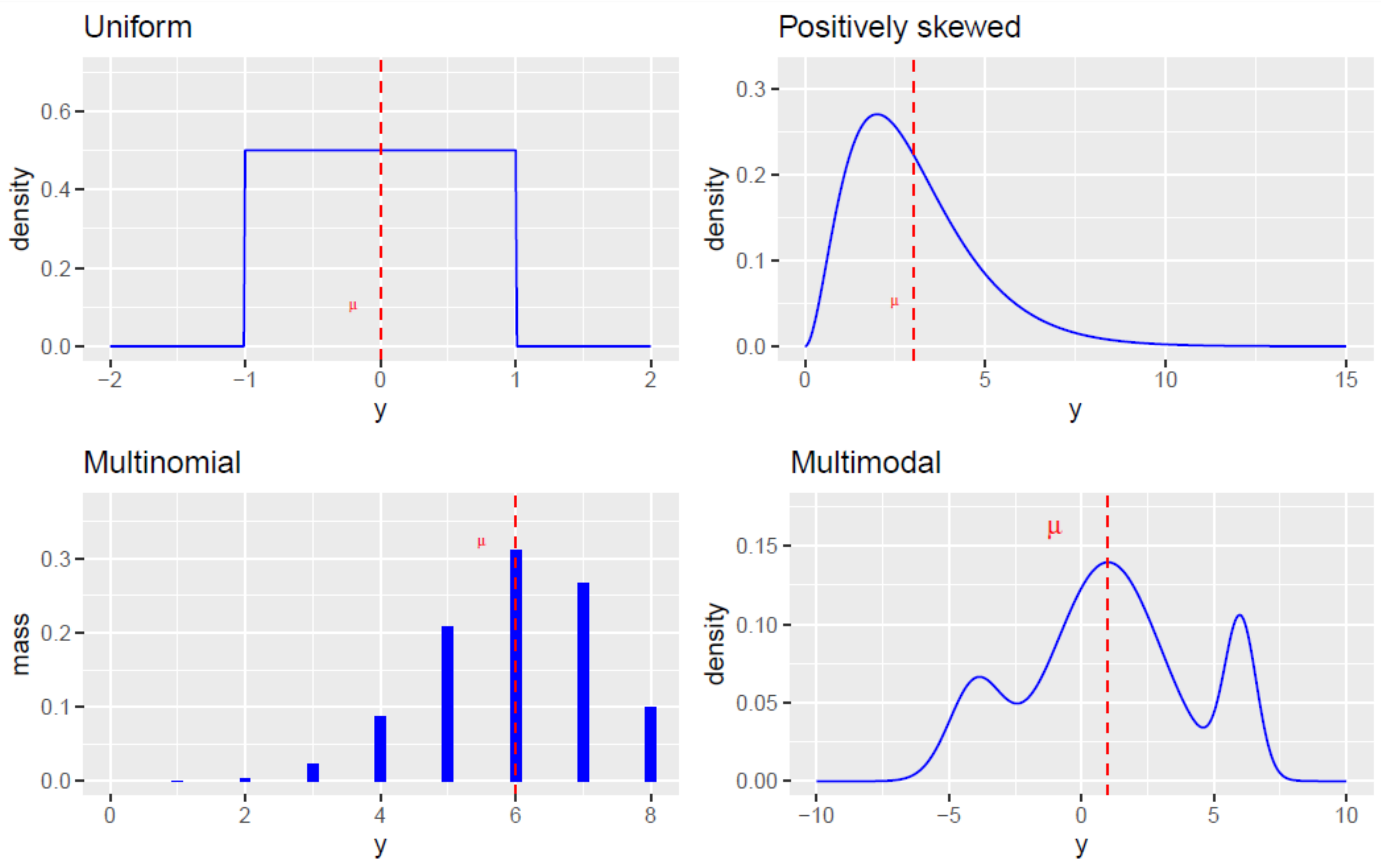
Central limited theorem

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

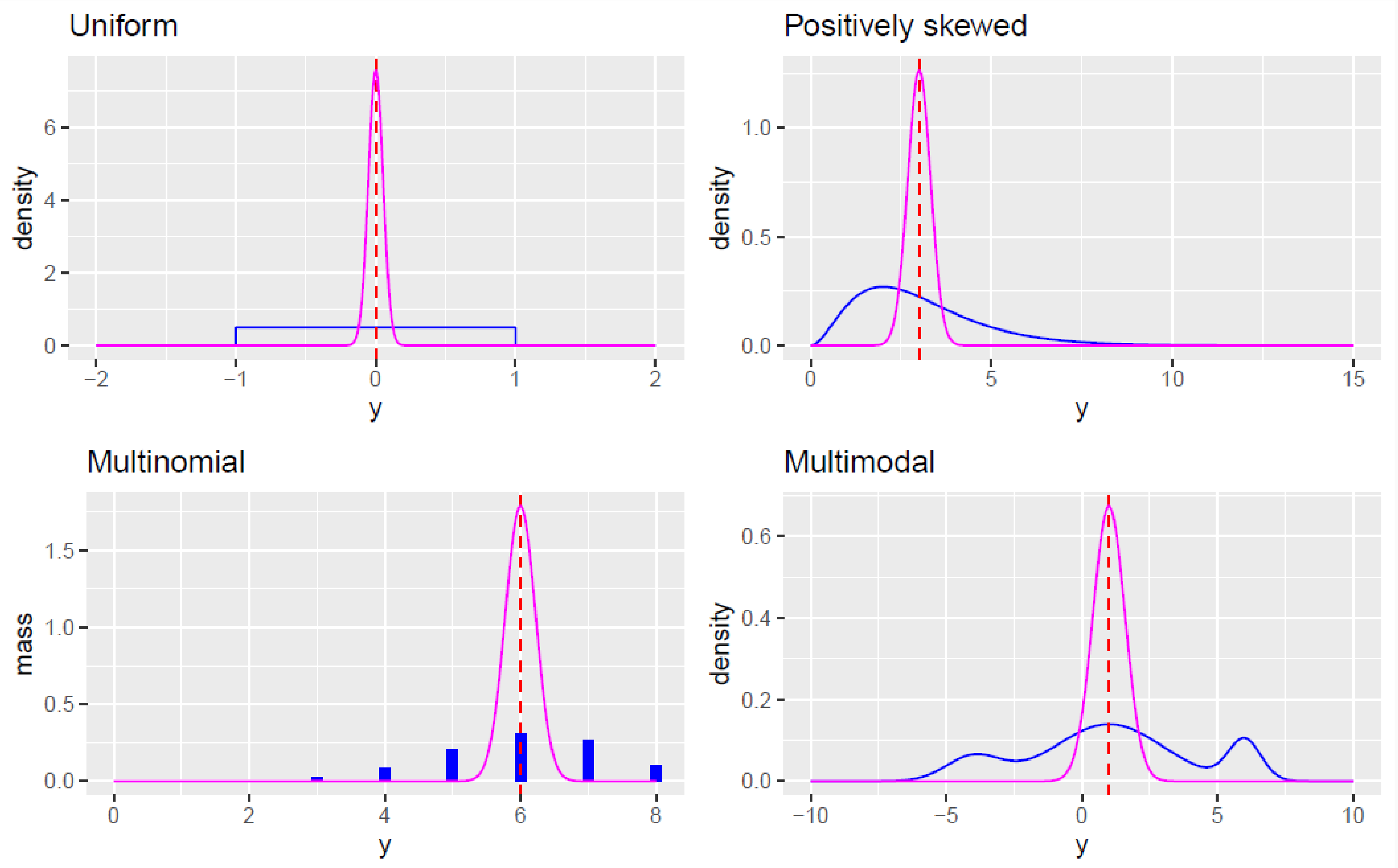
- even if the data from the original population is **not** normally distributed

ETC5242平时班 – Class 1



ETC5242平时班 – Class 1

CLT approximation with n = 30



ETC5242平时班 – Class 1

Sample standard deviation (measures the variation of the sample):

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

Standard error (measures the variation of the standard deviation)

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

Different variables can take on different range of values, so we need to standardize

$$T = \frac{\bar{X} - \mu}{SE} \overset{approx}{\sim} t_{n-1}$$

ETC5242平时班 – Class 1

Hypothesis testing with CLT

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 $\mu = \mu_0$

Confidence interval with CLT

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:

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

ETC5242平时班 – Class 1

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

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

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 $z_{0.975}$

Bootstrap

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

Bootstrap CI for single population mean base on x_bar (Week 5 lab)

1. Generate a Bootstrap sample of B potential $\bar{X}$ values
   - Denote these as $\{\bar{x}^{[1]}, \bar{x}^{[2]}, \dots, \bar{x}^{[B]}\}$
   - $B =$ 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%

Bootstrap CI for single population mean base on x_bar (Week 5 lab)

- 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

ETC5242平时班 – Class 1

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

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

```r
mean(a)
```

```
[1] 5.5
```

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

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

- 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

Confidence interval for difference between two means – paired samples (correlated data)

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

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

First calculate the sample of paired differences:

$DD_n = \{Diff_i = X_{1,i} - X_{2,i},\ \text{for } i = 1, 2, \dots, n\}$

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 $\{\bar{Diff}^{[b]},\ \text{for } b = 1, 2, \dots, B\}$ to obtain a confidence interval for $\mu_{Diff} = \mu_1 - \mu_2$

ETC5242平时班 – Class 1

Confidence interval for difference between two means – independent variables

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 $\{(\bar{x}^{[b]}_{1,obs} - \bar{x}^{[b]}_{2,obs}),\ \text{for } b = 1, 2, \ldots, B\}$

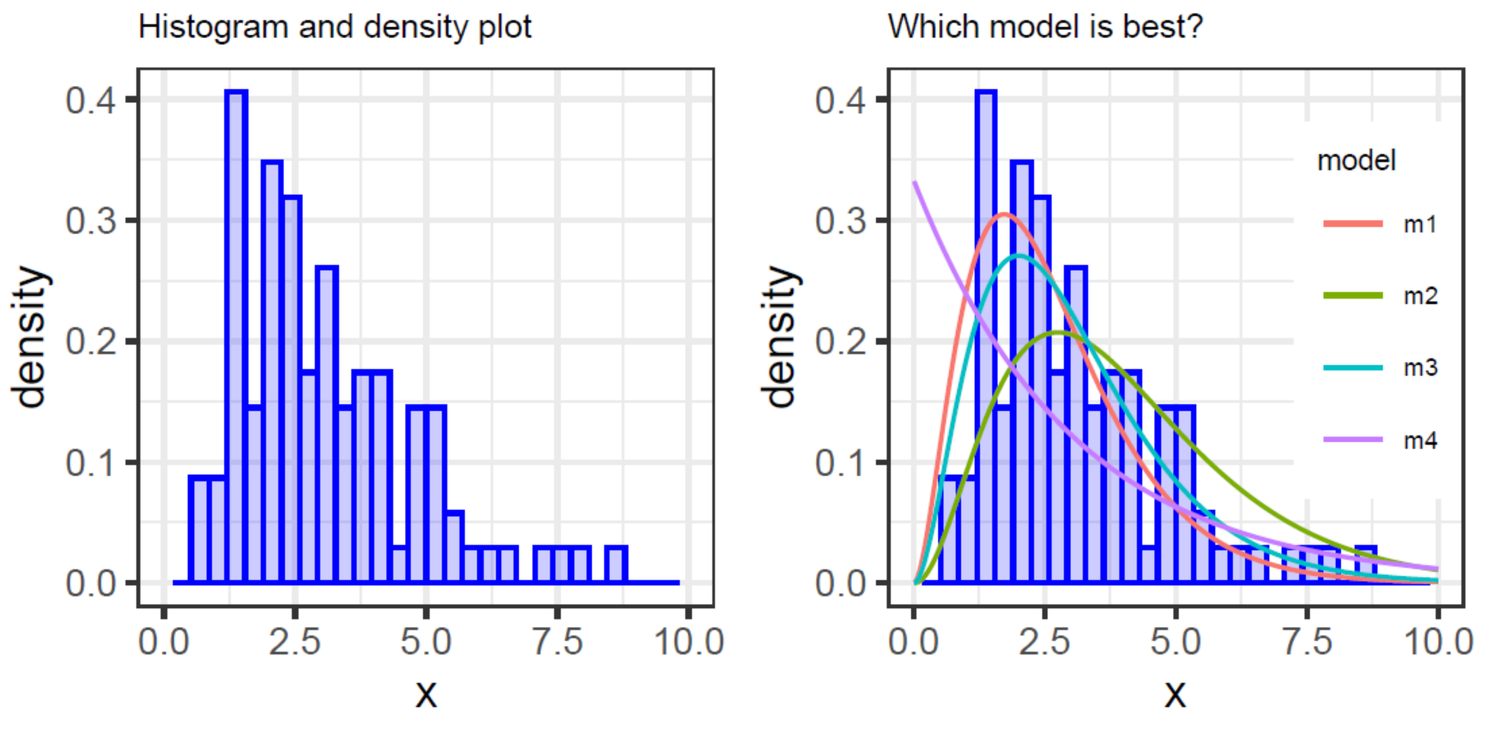
ETC5242平时班 – Class 1

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

However, there is one question related to bootstrap hypothesis testing in the assignment !

ETC5242平时班 – Class 1

- Which distributions might fit this data?
  - A normal distribution? An exponential? A gamma distribution? Something else?



ETC5242平时班 – Class 1

- Assuming the data are a random sample, we need to **choose a model** $F_X(x \mid \theta)$
  - We fit models using the sample and well-established distributional families
- Once we choose a model, we'll need to **estimate** the parameter θ
  - use the **maximum likelihood estimation** (MLE) method
- A fitted model will imply an estimate of the population mean
  - and other features

Likelihood Function

If $x_1, x_2, \ldots, x_n \overset{i.i.d.}{\sim} F_X(x \mid \theta)$, then the likelihood function is

$$\mathcal{L}_n(\theta) = \prod_{i=1}^{n} f_X(x_i \mid \theta)$$

And the **MLE** for θ is

$$\hat{\theta}_{MLE} = \arg\max_{\theta \in \Theta} \mathcal{L}_n(\theta)$$

Gaussian density function (normal distribution)

$$P(x;\mu,\sigma) = \frac{1}{\sigma\sqrt{2\pi}} \exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right)$$

Likelihood(Probability) of observing the three data points, 9, 9.5 and 11 given a particular gaussian density function, But we don’t know the two parameters yet

We want to maximise this joint probability

$$\begin{aligned} P(9, 9.5, 11; \mu, \sigma) = \frac{1}{\sigma\sqrt{2\pi}} \exp\left(-\frac{(9-\mu)^2}{2\sigma^2}\right) \times \frac{1}{\sigma\sqrt{2\pi}} \exp\left(-\frac{(9.5-\mu)^2}{2\sigma^2}\right) \\ \times \frac{1}{\sigma\sqrt{2\pi}} \exp\left(-\frac{(11-\mu)^2}{2\sigma^2}\right) \end{aligned}$$

ETC5242平时班 – Class 1

Optimising the likelihood function

- It is often easier to maximise the **log-likelihood function**

$$\ell_n(\theta) = \ln \mathcal{L}_n(\theta) = \left[\sum_{i=1}^{n} \ln f_X(x_i|\theta)\right]$$

- The **same** $\hat{\theta}_{MLE}$ maximises $\mathcal{L}_n(\theta)$ and $\ell_n(\theta)$
- In simple cases we can solve for $\hat{\theta}_{MLE}$ through differentiation
  - set first derivative of $\ell_n(\theta)$ equal to zero and solve
  - then check the second derivative of $\ell_n(\theta)$ is negative at $\hat{\theta}_{MLE}$
- More generally MLE is found using numerical optimisation on a computer

Very handy in R

```
fit <- fitdistr(x, "gamma")
fit
```

```
    shape       rate   
  3.4697     1.1235  
 (0.4690)   (0.1634)
```

Bootstrapping for confidence interval of model parameters

1. Generate a Bootstrap sample of B potential $\hat{\theta}$ values
   - For each b in $1 : B$
     - resample n draws from the observed data values, with replacement
     - label these values as $\{x_1^{[b]}, x_2^{[b]}, \ldots, x_n^{[b]}\}$
     - compute the MLE $\hat{\theta}^{[b]}$ by maximising $\mathcal{L}_n^{[b]}(\theta)$, constructed from the bootstrap sample
   - Bootstrap sample: $\{\hat{\theta}^{[1]}, \hat{\theta}^{[2]}, \ldots, \hat{\theta}^{[B]}\}$
2. Use the empirical distribution from this Bootstrap sample to approximate the sampling distribution of $\hat{\theta}_{MLE}$
3. Construct an approximate 95% confidence interval by selecting interval from 2. with (empirical) probability (at least) 95%
   - (lower) 2.5% quantile to 97.5% quantile