Confidence intervals

Statistical Computing and Empirical Methods
Unit EMATM0061, Data Science MSc

Rihuan Ke
rihuan.ke@bristol.ac.uk

Teaching Block 1, 2024

University of BRISTOL

What we will cover today

We will introduce confidence intervals as an important concept for quantifying uncertainty.

We will introduce how to compute confidence intervals for Gaussian data, and discuss the Student's t confidence intervals for approximately Gaussian data.

We will show the importance of the Gaussian assumption and how it can be checked.

We will introduce Wilson's method for confidence intervals on proportions.

We will introduce a powerful non-parametric alternative known as Bootstrap

What can we infer from our sample statistics?

Suppose we want to know the average flipper length μ for the population of Adelie penguins.

We collect a sample of size $n = 151$, and compute sample statistics

$$\text{Sample mean: } \overline{X} = 190, \qquad \text{Sample variance: } s^2 = 42.8.$$

Assuming that the sample is i.i.d. Gaussian random variables, we know that $\overline{X}$ is a consistent, minimum variance unbiased, maximum likelihood estimate of μ

However, it is rarely the case that $\overline{X} = \mu$.

Question: Can we say with confidence that μ is within some specific range of possible values?

Today's focus

Suppose that $X_1, \cdots, X_n$ are i.i.d. random variables. We are interested in the population parameter $\theta \in \mathbb{R}$.

Given the sample, we want to find a range of possible values for θ such that we can be certain that the true value of θ will fall in the given range with a probability of our choosing.

This range of estimates is an alternative to the estimates in the form of a single value (point estimates), e.g., the sample mean, sample variance, MLE, $\cdots$

1. Confidence intervals

Suppose that $X_1, \cdots, X_n$ are i.i.d. random variables. We are interested in the population parameter $\theta \in \mathbb{R}$.

Confidence intervals

If we have sample statistics $L_n \equiv L_n(X_1, \cdots, X_n)$ and $U_n \equiv U_n(X_1, \cdots, X_n)$ satisfying

$$\mathbb{P}[L_n(X_1, \cdots, X_n) < \theta < U_n(X_1, \cdots, X_n)] \geq \gamma,$$

then we refer to (L_n, U_n) as $\gamma \times 100\%$-level confidence interval.

Note that L_n and U_n are random variables themselves and θ is a number, so we can discuss the probability of the event $\{L_n(X_1, \cdots, X_n) < \theta < U_n(X_1, \cdots, X_n)\}$.

$\gamma \in [0, 1]$, and γ is referred to as the confidence level of the confidence interval.

2. Confidence intervals with Gaussian data

If we have sample statistics $L_n \equiv L_n(X_1, \cdots, X_n)$ and $U_n \equiv U_n(X_1, \cdots, X_n)$ satisfying $\mathbb{P}[L_n(X_1, \cdots, X_n) < \theta < U_n(X_1, \cdots, X_n)] \geq \gamma$, then we refer to (L_n, U_n) as $\gamma \times 100\%$-level confidence interval.

Again, we model our sample $X_1, \cdots, X_n$ with a parametric model, then based on the parametric model we find the L_n and U_n.

Let's consider one of the simplest cases where the sample is modelled as i.i.d. Gaussian random variables,

$$X_1, \cdots, X_n \sim \mathcal{N}(\mu, \sigma^2).$$

Assume that we want to have a confidence interval for μ.

We can generate a confidence interval for μ based on the Student's t-distribution (see the next slide).

Confidence intervals with Gaussian data

To develop a confidence interval for μ, we first construct a random variable

$$T := \frac{\overline{X} - \mu}{S/\sqrt{n}}$$

where $\overline{X}$ is the sample mean and S is the sample standard deviation.

The distribution of T is described by the following result:

Lemma

Suppose that $X_1, \cdots, X_n \sim \mathcal{N}(\mu, \sigma^2)$ are i.i.d. random variables. Let $\overline{X} := \frac{1}{n}\sum_{i=1}^{n} X_i$ and $S^2 := \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \overline{X})^2$. Let

$$T := \frac{\overline{X} - \mu}{S/\sqrt{n}}$$

Then the random variable is t-distributed with $n-1$ degree of freedom.

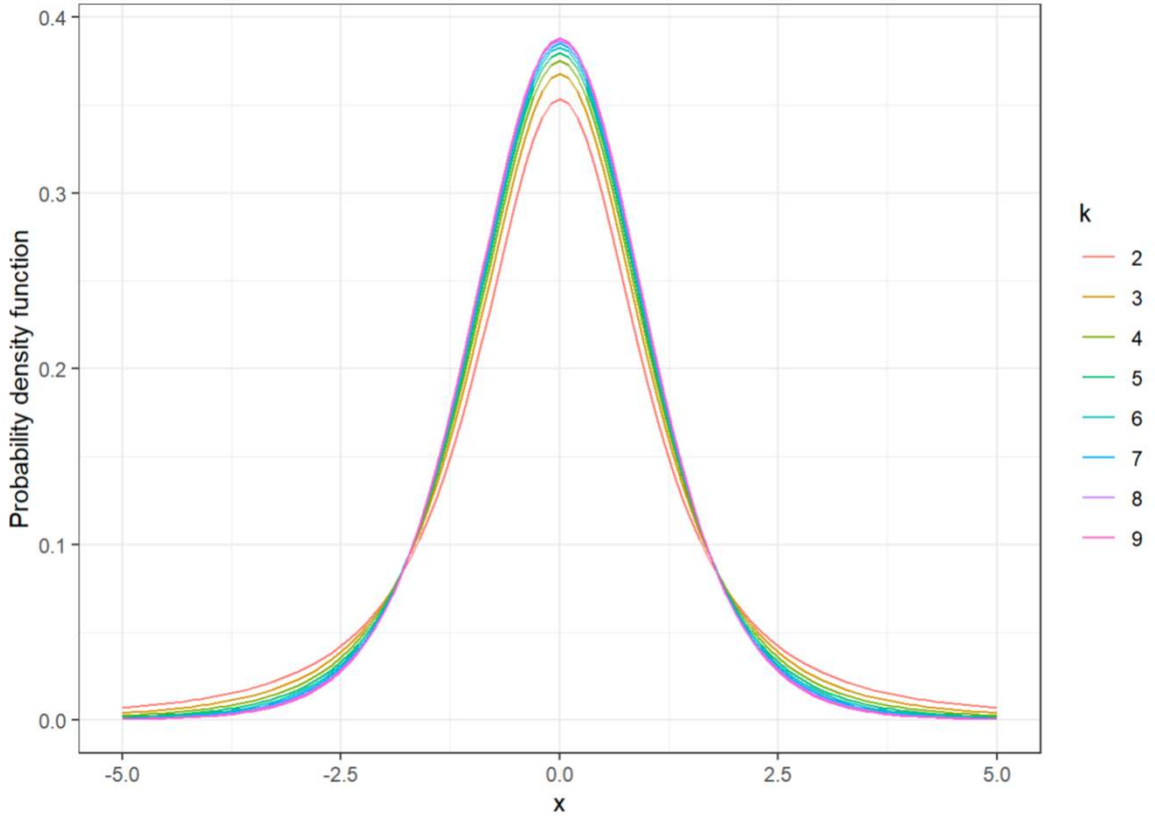
The distribution of T does not depend on μ and σ.

Student's t distribution

A random variable Q is said to be chi-squared with k degrees of freedom if $Q = \sum_{i=1}^{k} Z_i^2$ with independent $Z_1, Z_2, \cdots, Z_k \sim \mathcal{N}(0,1)$.

A random variable T is said to be t distributed with k degrees of freedom if $T = \frac{Z}{\sqrt{Q/k}}$ for two independent random variables $Z \sim \mathcal{N}(0,1)$ and $Q \sim \chi^2(k)$.

The distribution of T is called a Student's t distribution.



A t-distributed random variable has a single parameter k.

The t distribution is symmetric!

Student's t distribution

A random variable T is said to be t distributed with k degrees of freedom if $T = \frac{Z}{\sqrt{Q/k}}$ for two independent random variables $Z \sim \mathcal{N}(0,1)$ and $Q \sim \chi^2(k)$.

Let the cumulative distribution function of T be denoted by $F_k(t) := \mathbb{P}(T < t)$.

Let the quantile function be denoted by $(F_k)^{-1} : [0,1] \to \mathbb{R}$, then for any $\alpha \in [0,1]$, we have

$$\mathbb{P}\big(T < (F_k)^{-1}(\alpha)\big) = \alpha$$

The t distribution is well understood, and the function F_k and $(F_k)^{-1}$ can be computed easily by using R programming, Python, $\cdots$

Confidence intervals with Gaussian data

Lemma

Suppose that $X_1, \cdots, X_n \sim \mathcal{N}(\mu, \sigma)$ are i.i.d. random variables. Let $\overline{X} := \frac{1}{n}\sum_{i=1}^{n} X_i$ and $S^2 := \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \overline{X})^2$. Let

$$T := \frac{\overline{X} - \mu}{S/\sqrt{n}}$$

Then the random variable is t-distributed with $n-1$ degree of freedom.

Given $\alpha \in [0,1]$, compute a quantity $t_{\alpha/2,n-1}$, such that $\mathbb{P}(-t_{\alpha/2,n-1} < T < t_{\alpha/2,n-1}) = 1-\alpha$. (We will show what is $t_{\alpha/2,n-1}$ later).

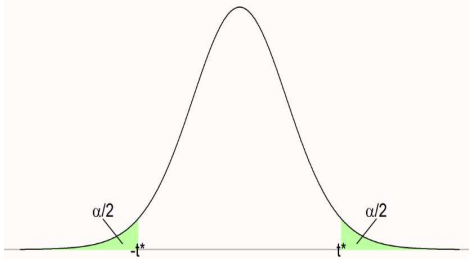
Then by $T := \frac{\overline{X}-\mu}{S/\sqrt{n}}$ we get $\mathbb{P}\left(\overline{X} - \frac{t_{\alpha/2,n-1}}{\sqrt{n}} \cdot S < \mu < \overline{X} + \frac{t_{\alpha/2,n-1}}{\sqrt{n}} \cdot S\right) = 1-\alpha$.

So, $\left(\overline{X} - \frac{t_{\alpha/2,n-1}}{\sqrt{n}} \cdot S, \overline{X} + \frac{t_{\alpha/2,n-1}}{\sqrt{n}} \cdot S\right)$ is a $(1-\alpha)\%$-level confidence interval.

Confidence intervals with Gaussian data

So, $\left(\overline{X} - \frac{t_{\alpha/2,n-1}}{\sqrt{n}} \cdot S, \overline{X} - \frac{t_{\alpha/2,n-1}}{\sqrt{n}} \cdot S\right)$ is a $(1-\alpha)\%$-level confidence interval.

Next, we will show how to compute $t_{\alpha/2,n-1}$:



By symmetric

$$\mathbb{P}(-t_{\alpha/2,n-1} < T < t_{\alpha/2,n-1}) = 1-\alpha \iff \mathbb{P}(T < t_{\alpha/2,n-1}) = 1 - \frac{\alpha}{2}$$

$$\iff t_{\alpha/2,n-1} := (F_{n-1})^{-1}(1-\frac{\alpha}{2}).$$

Example: compute $t_{\alpha/2,n-1}$

```
# compute t_{alpha/2, n-1}
alpha <- 0.05
n <- 80
t = qt(1-alpha/2, df=n-1) # quantile function for t distribution
t
```

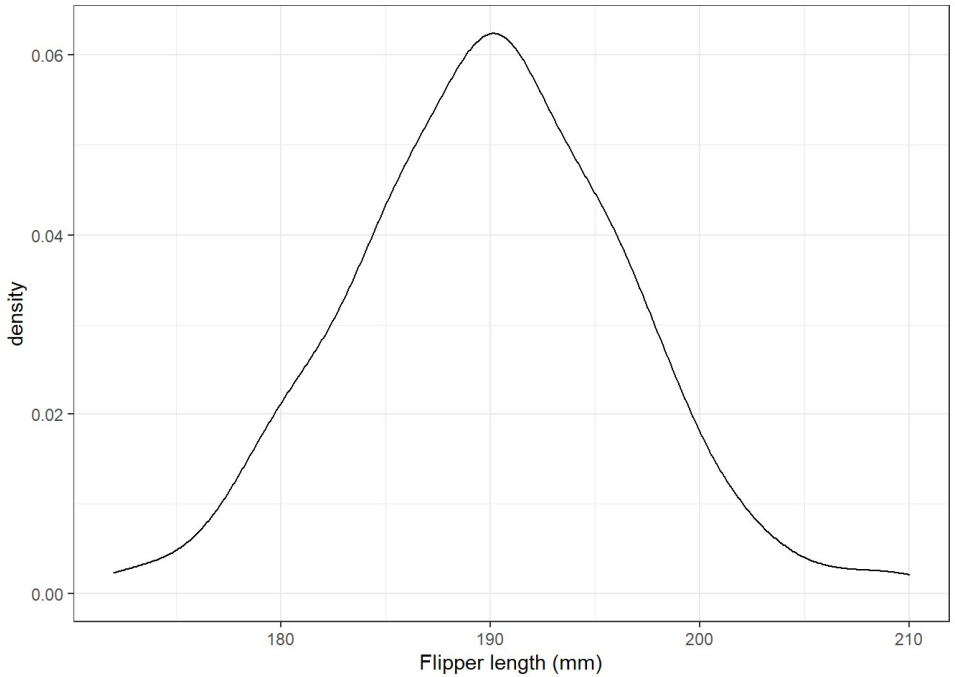
```
## [1] 1.99045
```

3. How to check if a Gaussian model is reasonable

Can we reasonably assume our data is generated by a Gaussian model $X_1, \cdots, X_n \sim \mathcal{N}(\mu, \sigma^2)$?

First, we can do a density plot and check if the data looks Gaussian....

```
ggplot(data=filter(penguins, species=="Adelie"),
       aes(x=flipper_length_mm)) + geom_density() + theme_bw() +
  xlab("Flipper length (mm)")
```



Here, a Gaussian model seems reasonable:

- The data looks unimodal (a single peak)
- The data looks symmetric about its mean

3. How to check if a Gaussian model is reasonable

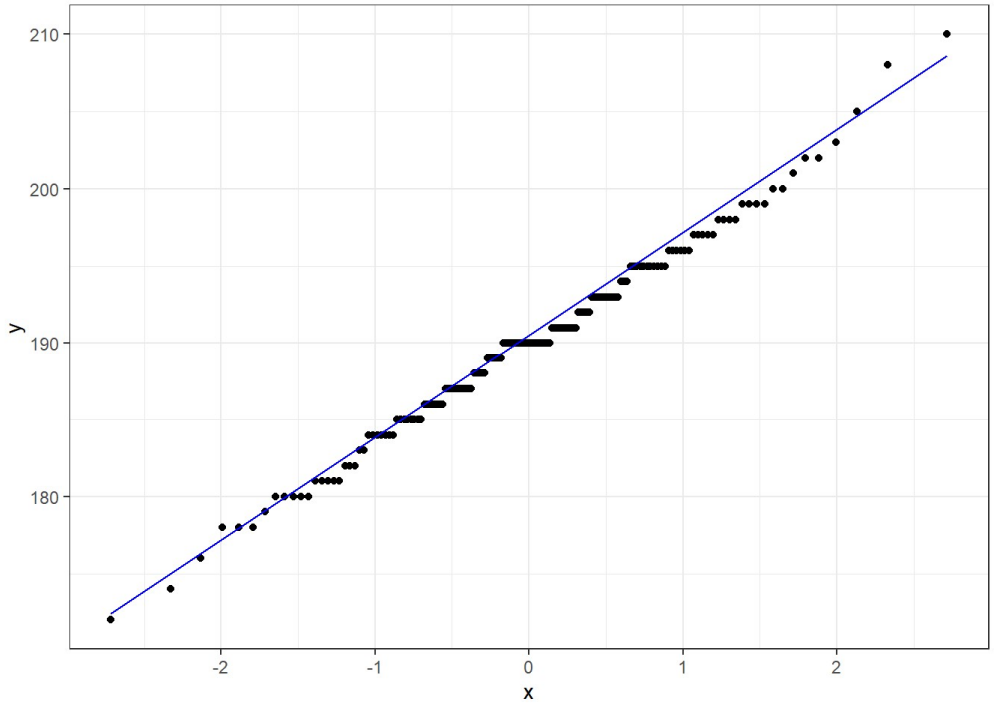
Can we reasonably assume our data is generated by a Gaussian model $X_1, \cdots, X_n \sim \mathcal{N}(\mu, \sigma^2)$?

Our second check that a Gaussian model is reasonable is a quantile-quantile plot.

The QQ-plot is a plot that compares the quantiles in the sample (y-axis) with theoretical quantiles from a Gaussian (x-axis), i.e., it is the plot of the points

$$\left(F^{-1}(q), F_X^{-1}(q)\right) \text{ for } q \in [0,1]$$

```
ggplot(data=filter(penguins, species=="Adelie"),
       aes(sample=flipper_length_mm)) +
  theme_bw() + stat_qq() + stat_qq_line(color="blue")
```



If our QQ plot points lie close to a straight line?

If so then our assumption of Gaussian data is reasonable..

4. Confidence intervals with non-Gaussian data

In practice, our data is rarely exactly Gaussian.

However, when the sample size is large, we have the central limit theorem, which approximates sample distribution with Gaussian distributions:

Theorem (Lindeberg—Lévy) The central limit theorem

Let $X : \Omega \to \mathbb{R}$ be a random variable with expectation $\mu = \mathbb{E}(X)$ and variance $\sigma^2 = \mathrm{Var}(X)$. Let $X_1, \cdots, X_n : \Omega \to \mathbb{R}$ be a sequence of independent copies of X. Let $Z \sim \mathcal{N}(0, 1)$ be a standard Gaussian random variable. Then for all $x \in \mathbb{R}$,

$$\lim_{n\to\infty} \mathbb{P}\left\{ \sqrt{\frac{n}{\sigma^2}} \cdot \left(\frac{1}{n} \sum_{i=1}^{n} X_i - \mu \right) \leq x \right\} = \mathbb{P}(Z \leq x).$$

Therefore, for large n, the sample mean $\frac{1}{n}\sum_{i=1}^{n} X_i$ behaves like the Gaussian distribution $\mathcal{N}(\mu, \frac{\sigma^2}{n})$.

So one can derive (approximate) confidence intervals based on this limiting behaviour, for non-Gaussian data.

5. Binomial proportion confidence interval

Suppose our data sample is a binary sequence $(X_i)_{i=1}^n$ in $\{0,1\}^n$.

Examples

1. $(X_i)_{i=1}^n$ represents a sequence of test results for a driving test.
2. $(X_i)_{i=1}^n$ represents a sequence of outcomes for a new treatment.

We can model the sequence $(X_i)_{i=1}^n$ as an i.i.d . Bernoulli sequence $X_1, \cdots, X_n \sim \mathcal{B}(q)$.

We would like to estimate a confidence interval for the success probability $q = \mathbb{E}(X_i)$.

The idea is to use the result of the central limit theorem (see next slide).

Binomial proportion confidence interval

Recall that, for large n, the sample mean $\frac{1}{n}\sum_{i=1}^{n} X_i$ behaves like the Gaussian distribution $\mathcal{N}(\mu, \frac{\sigma^2}{n})$.

By the central limit theorem, $\sqrt{\frac{n}{q(1-q)}} \cdot \left(\frac{1}{n}\sum_{i=1}^{n} X_i - q\right)$ approximates $\mathcal{N}(0,1)$.

Or equavilently, we have $\frac{1}{n}\sum_{i=1}^{n} X_i$ approximates $\mathcal{N}\left(q, \frac{q(1-q)}{n}\right)$.

One of the important methods for binomial proportion confidence intervals, called Wilson's method, uses this approximation to create a confidence interval for q based on $\frac{1}{n}\sum_{i=1}^{n} X_i$.

Wilson's method for confidence intervals

We can use the PropCIs package to compute confidence intervals via Wilson's method.

```
library (PropCIs)
```

```
driving_test_results <- c(1,0,1,0,0,0,0,0,0,1,0,0,0,1,0,1,0,1,0,1,0,0,1,0)
mean(driving_test_results)
```

```
## [1] 0.3333333
```

```
alpha <- 0.05
num_successes <- sum(driving_test_results) # total  passes
sample_size <- length(driving_test_results) # sample size

scoreci(x=num_successes, n=sample_size, conf.level=1-alpha)
```

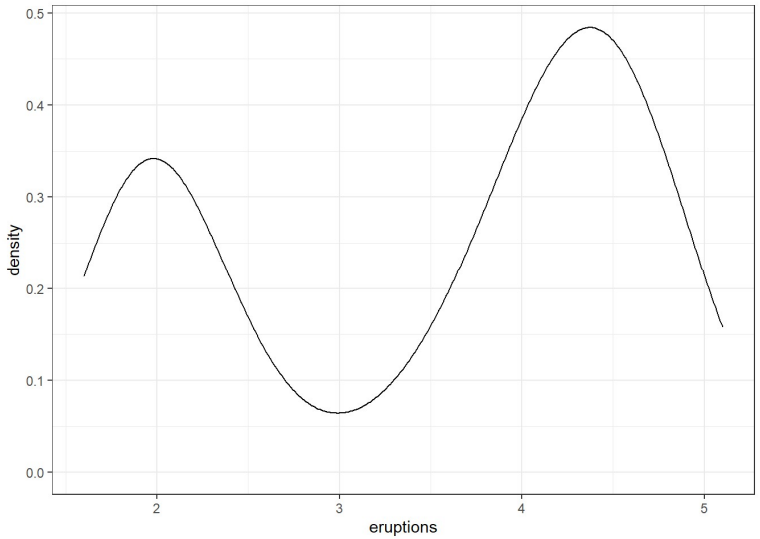
```
##
##
##
## data:
##
## 95 percent confidence interval:
##  0.1797 0.5329
```

6. A flexible method for confidence intervals

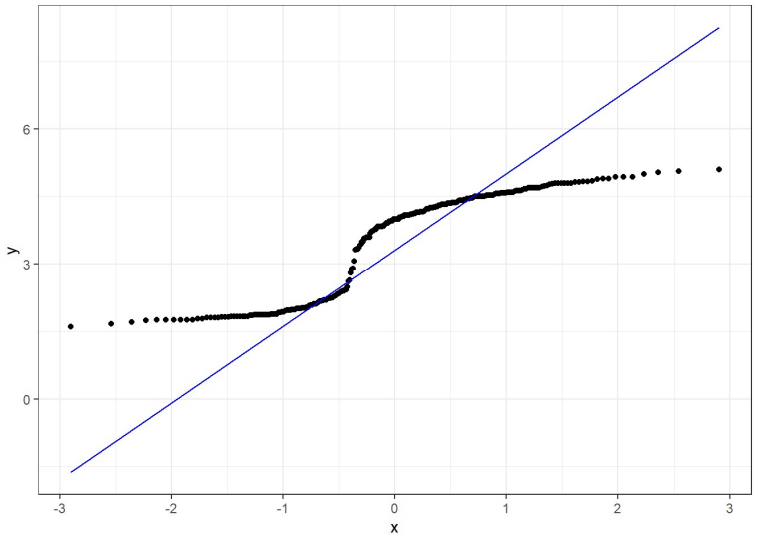
Suppose we are interested in a complex statistic other than the mean...

Or suppose our data deviates strongly from the assumption of a Gaussian or normal distribution

density plot



QQ-plot



Can we still compute confidence intervals?

The Bootstrap method (see next slide)

The Bootstrap method

Suppose we have an i.i.d. sample $X_1, \cdots, X_n \sim P$. We estimate a population parameter θ with a sample statistic $\hat{\theta} = g(X_1, \cdots, X_n)$.

To quantify our uncertainty we wish to understand the distribution of $\hat{\theta} - \theta$.

Motivation:

Consider a special case where $X_1, \cdots, X_n \sim \mathcal{N}(\mu, 1)$. Let $\theta := \mu$ and $\hat{\theta} := \overline{X}$.

To compute the confidence interval for μ we start from the distribution of $\hat{\theta} - \theta = \overline{X} - \mu \sim \mathcal{N}\left(0, \frac{1}{n}\right)$.

Then a confidence interval can be obtained from $\mathbb{P}\left(-\sqrt{\frac{1}{n}} z_{\alpha/2} < \overline{X} - \mu < \sqrt{\frac{1}{n}} z_{\alpha/2}\right) = 1 - \alpha$ for some quantity $z_{\alpha/2}$

The idea of Bootstrap: We do not have direct access to the distribution of $\hat{\theta} - \theta$ or the samples of $\hat{\theta} - \theta$. However, we can try to generate samples to approximate the distribution of $\hat{\theta} - \theta$ (See the next slide).

The Bootstrap method

Suppose we have an i.i.d. sample $X_1, \cdots, X_n \sim P$. We estimate a population parameter θ with a sample statistic $\hat{\theta} = g(X_1, \cdots, X_n)$.

The main steps of the Bootstrap method for confidence intervals are as fellow:

Step 1. We consider an empirical distribution $\hat{P}_n$ which approximates P as follows:

$\hat{P}_n$ is the discrete distribution which assigns probability $\frac{1}{n}$ to each of $X_1, \cdots, X_n$.

So, Sampling from $\hat{P}_n$ is equivalent to randomly sampling from $X_1, \cdots, X_n$ with replacement.

So, in Bootstrap, we "view" $\{X_1, \cdots, X_n\}$ as our population, and we generate subsamples from this population.

The Bootstrap method: main steps

Step 1. We consider an empirical distribution $\hat{P}_n$ which approximates P.

Step 2. We generate multiple samples from $\hat{P}_n$, and compute the associate sample statistics $\tilde{\theta}^1, \cdots, \tilde{\theta}^B$.

$$\begin{array}{ll} \tilde{X}_1^1, \cdots, \tilde{X}_n^1 \sim \hat{P}_n & \tilde{\theta}^1 := g(\tilde{X}_1^1, \cdots \tilde{X}_n^1) \\ \vdots & \vdots \\ \tilde{X}_1^B, \cdots, \tilde{X}_n^B \sim \hat{P}_n & \tilde{\theta}^B := g(\tilde{X}_1^B, \cdots \tilde{X}_n^B) \end{array}$$

Then we view $\{\tilde{\theta}_1, \cdots, \tilde{\theta}_B\}$ as an approximate to the distrubtion of $\hat{\theta}$.

We argue that the behaviour of $\hat{\theta} - \theta$ is **approximately** the same as the behaviour of $\tilde{\theta} - \hat{\theta}$ where $\tilde{\theta}$ is the random variable that represents $\tilde{\theta}_1, \cdots, \tilde{\theta}_B$.

Step 3. We then compute a confidence interval from the distribution of $\tilde{\theta} - \hat{\theta}$, which is an approximation to $\hat{\theta} - \theta$. (See the next slide).

The Bootstrap method: main steps

Step 3. We then compute a confidence interval from the distribution of $\tilde{\theta} - \hat{\theta}$, which is an approximation to $\hat{\theta} - \theta$. (See the next slide).

Suppose we want to compute $(1-\alpha) \times 100\%$-level confidence intervals for the parameter θ.

Let $\delta_{\alpha/2}$ denote the $\frac{\alpha}{2}$ quantile for $\{\tilde{\theta}_1 - \hat{\theta}, \cdots, \tilde{\theta}_B - \hat{\theta}\}$.
Similar, let $\delta_{1-\alpha/2}$ be the $1 - \frac{\alpha}{2}$ quantile. So when B is large,

$$\mathbb{P}(\tilde{\theta} - \hat{\theta} < \delta_{\alpha/2}) \approx \frac{\alpha}{2}, \qquad \text{and} \qquad \mathbb{P}(\tilde{\theta} - \hat{\theta} < \delta_{1-\alpha/2}) \approx 1 - \frac{\alpha}{2}$$

$$\begin{aligned}
1 - \alpha &\approx \mathbb{P}(\tilde{\theta} - \hat{\theta} < \delta_{1-\alpha/2}) - \mathbb{P}(\tilde{\theta} - \hat{\theta} < \delta_{\alpha/2}) \\
&= \mathbb{P}(\delta_{\alpha/2} < \tilde{\theta} - \hat{\theta} < \delta_{1-\alpha/2}) \\
&\approx \mathbb{P}(\delta_{\alpha/2} < \hat{\theta} - \theta < \delta_{1-\alpha/2}) \qquad \text{by assumption that } \tilde{\theta} - \hat{\theta} \text{ approximates } \hat{\theta} - \theta \\
&= \mathbb{P}(\hat{\theta} - \delta_{\alpha/2} < \theta < \hat{\theta} - \delta_{1-\alpha/2})
\end{aligned}$$

So the empirical Bootstrap level confidence interval: $[\hat{\theta} - \delta_{1-\alpha/2}, \hat{\theta} - \delta_{\alpha/2}]$.

The Bootstrap method: main steps

Step 1. We consider an empirical distribution $\hat{P}_n$ which approximates P.

Step 2. We generate multiple samples from $\hat{P}_n$, and compute the associate sample statistics $\tilde{\theta}^1, \cdots, \tilde{\theta}^B$.

Step 3. We then compute a confidence interval from the distribution of $\tilde{\theta} - \hat{\theta}$, which is an approximation to $\hat{\theta} - \theta$.
The empirical Bootstrap level confidence interval: $[\hat{\theta} - \delta_{1-\alpha/2}, \hat{\theta} - \delta_{\alpha/2}]$.

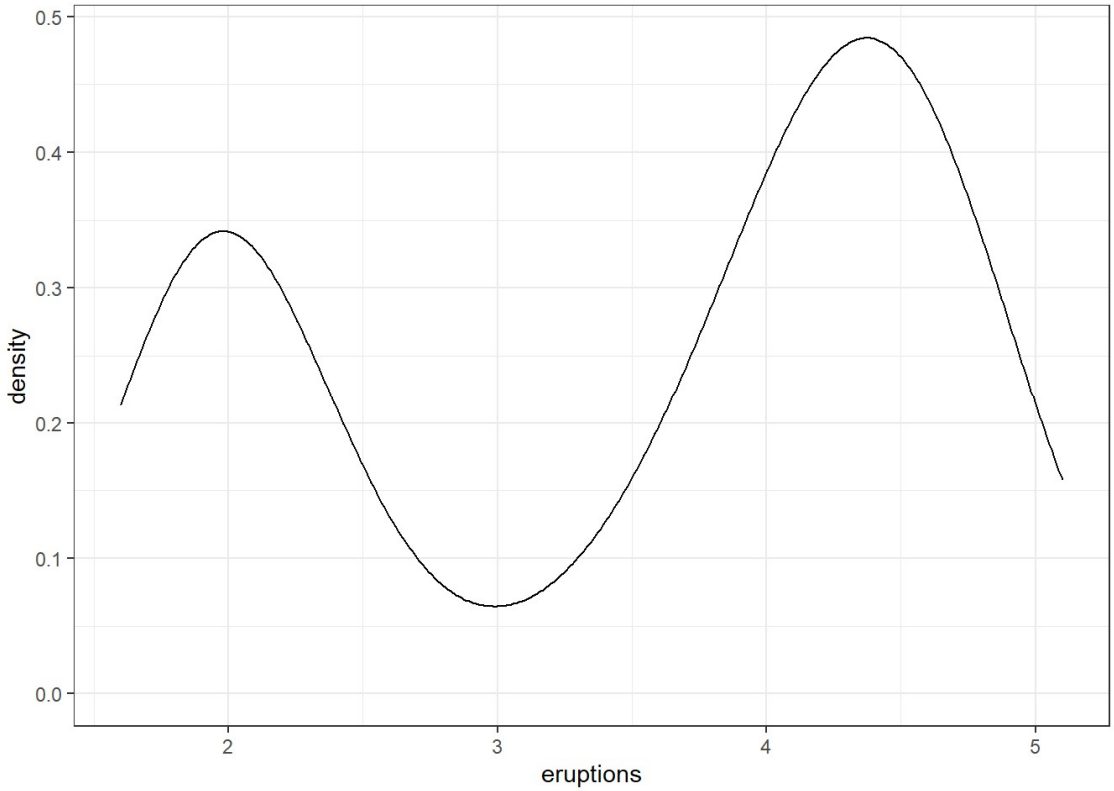
Remarks.

Note in each of the steps we do not assume a specific probabilistic model for P.

Empirical Bootstrap method gives approximate confidence intervals under general conditions.

The Bootstrap method: example

Example: Suppose that we want to compute a 99% level confidence interval for the median for the duration of eruptions (in mins) for the volcano data set.



The Bootstrap method: example

In R, bootstrap confidence intervals can be compute by using the `boot` package.

```r
library(boot) # load the library
```

```r
set.seed(123) # set random seed
geyser = faithful # the volcano data set
```

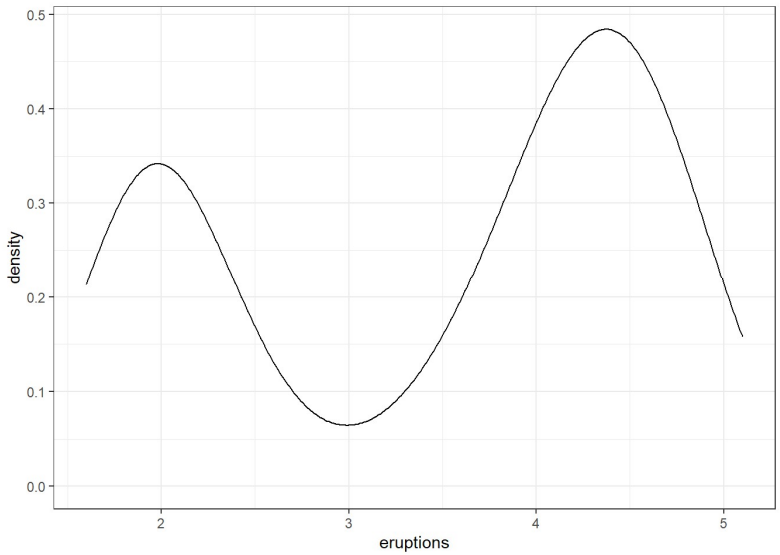
```r
# 1. define a function which computes the median of a column of interest
compute_median <- function(df, indicies,col_name){
  sub_sample <- slice(df, indicies) %>% pull(all_of(col_name)) # extract subsample
  return (median(sub_sample, na.rm=TRUE))# return median
}
```

```r
# 2. use the boot function to generate the bootstrap statistics
results <- boot(data=geyser, statistic=compute_median, col_name="eruptions", R=10000)

# 3. compute the 99% confidence interval for the median
boot.ci(boot.out = results, type = "basic", conf=0.99)
```

The Bootstrap method: example

```r
# 2. use the boot function to generate the bootstrap statistics
results <- boot(data=geyser, statistic=compute_median, col_name="eruptions", R=10000)

# 3. compute the 99% confidence interval for the median
boot.ci(boot.out = results, type = "basic", conf=0.99)
```

```
## BOOTSTRAP CONFIDENCE INTERVAL CALCULATIONS
## Based on 10000 bootstrap replicates
##
## CALL :
## boot.ci(boot.out = results, conf = 0.99, type = "basic")
##
## Intervals :
## Level      Basic
## 99%   ( 3.858,  4.266 )
## Calculations and Intervals on Original Scale
```



Bootstrap vs. parametric confidence intervals

✓ The Bootstrap method has several advantages over parametric methods:

- Non-parametric i.e. does not require strong distributional assumptions e.g. Gaussian data.
- Applies to any statistical estimator e.g. median, trimmed mean etc

✗ The Bootstrap method also some has drawbacks relative to parametric methods:

- Computationally expensive - less of a concern with modern hardware.
- Parametric methods typically outperform the Bootstrap methods when the assumptions hold.

7. General guidelines for confidence intervals

Always check the assumptions of whatever confidence intervals you're using.

➡ If you are interested in the population mean and your data is approximately Gaussian

- A good option is the Student t confidence intervals.
- **Remark:** The larger the sample size the less concerned you need to be about departures from Gaussianity! (because of the central limit theorem)

➡ If you are interested in the population mean and your data is approximately Bernoulli

- A good option is Wilson's score interval

➡ If your data is highly non-Gaussian or you are interested in another statistic either:

a) Use a bespoke confidence interval for a specific setting but always check assumptions, or

b) Use the Bootstrap method!

What have we covered?

We introduced the concept of a confidence interval for quantifying uncertainty.

We discussed visual methods for checking if your data can be modelled as Gaussian.

We introduced Student's t based confidence intervals for approximately Gaussian data.

… but departures from Gaussian behaviour are less of a concern for large sample sizes!

We introduced Wilson's method for confidence intervals on proportions with Bernoulli variables.

We introduced the powerful Bootstrap method for non-parametric confidence intervals.

University of BRISTOL

Thanks for listening!

Dr. Rihuan Ke

rihuan.ke@bristol.ac.uk

Statistical Computing and Empirical Methods
Unit EMATM0061, MSc Data Science