Confidence intervals

Parametric and non-parametric methods for quantifying uncertainty.

Henry W J Reeve

henry.reeve@bristol.ac.uk

Statistical Computing & Empirical Methods (EMATM0061)

MSc in Data Science, Teaching block 1, 2021.

What will we cover today?

- We will introduce the concept of a confidence interval for quantifying uncertainty.
- We will introduce Student’s t confidence intervals for approximately Gaussian data.
- We 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 the Bootstrap.

What can we infer from our sample statistics?

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

We have access to a sample of measurements of flipper lengths with

Sample size: $n = 151$ Sample mean: $\overline{X} = 190$ Sample variance: $s^2 = 42.8$

What can we infer about the population mean μ ?

What can we infer from our sample statistics?

Sample size: $n = 151$ Sample mean: $\overline{X} = 190$ Sample variance: $s^2 = 42.8$

What can we infer about the population mean μ ?

We know that $\overline{X} = 190$ is a consistent, minimum variance unbiased, maximum likelihood estimate…

Can we quantify the uncertainty of our estimate for μ ?

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

Can we reject the hypothesis that $\mu = 200$?

Confidence intervals

Can we give a range of values which we are confident that our population parameter θ lies within?

Suppose we want to estimate a population parameter $\theta \in \mathbb{R}$ from a sample $X_1, \cdots, X_n$,

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

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

We refer to (L_n, U_n) as the $\gamma \times 100\%$ -level **confidence interval**.

γ is referred to the confidence level of the confidence interval (L_n, U_n).

A Gaussian model for our data?

To answer these questions we would like to model our sample $X_1, \cdots, X_n$ with a parametric model.

As a continuous feature we first propose to model the sample as i.i.d. draws from a univariate Gaussian.

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

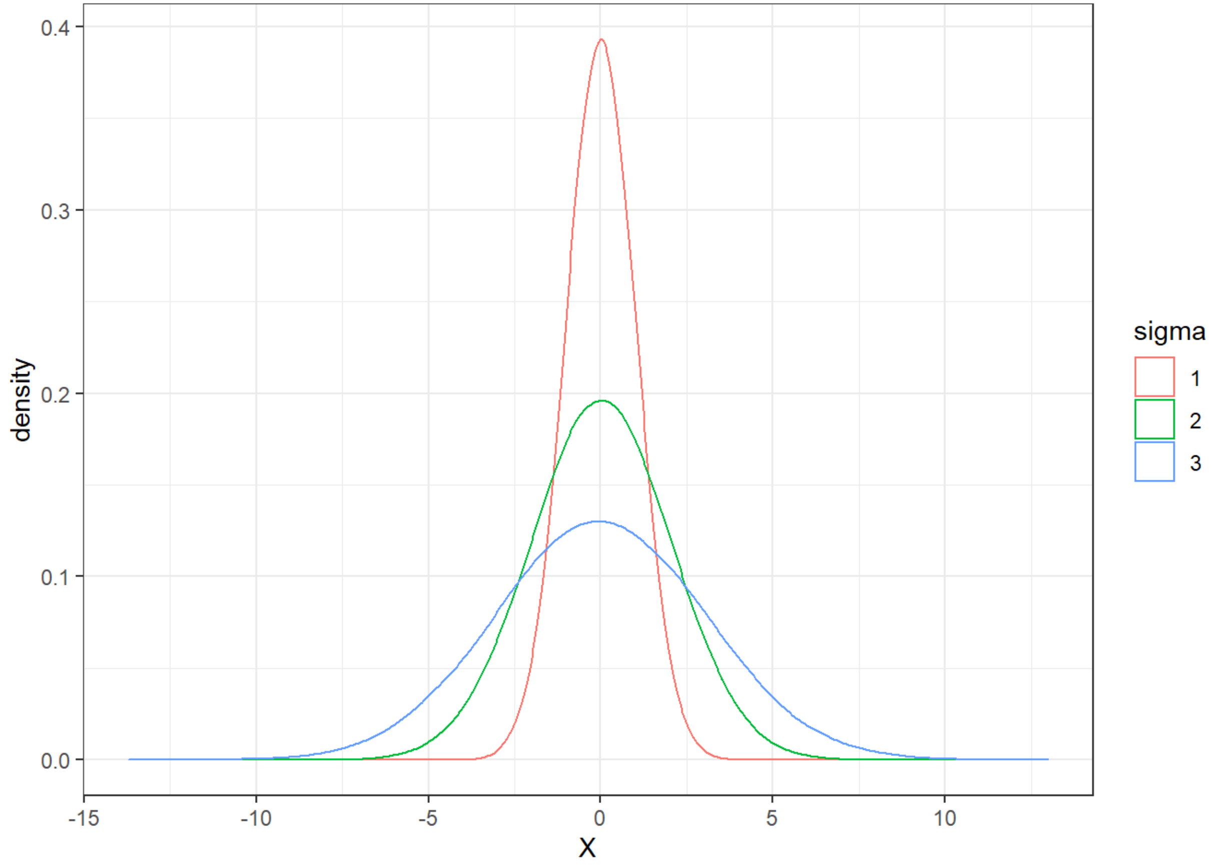
Is the i.i.d. (independent and identically distributed) assumption reasonable?

Can we reasonably assume our data is generated by a Gaussian distribution?

Let’s assume so, for now….. we will return to this question later….

Then we can generate a confidence interval for μ based on the Student’s t-distribution.

Gaussian random variables

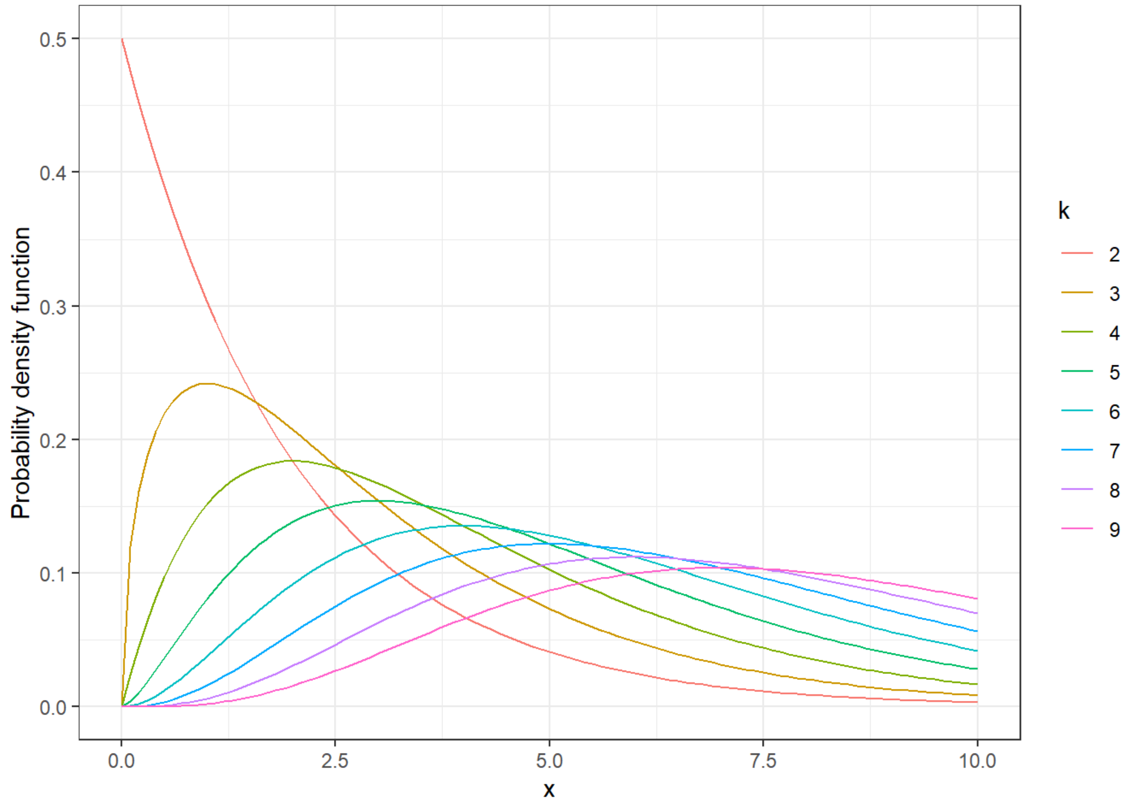


Given $X \sim \mathcal{N}\left(\mu, \sigma^2\right)$ we have $\mathbb{E}[X] = \mu$ and $\mathrm{Var}(X) = \sigma^2$

Chi squared distribution

A random variable Q is said to be chi-squared with k degrees of freedom $Q \sim \chi^2(k)$ if

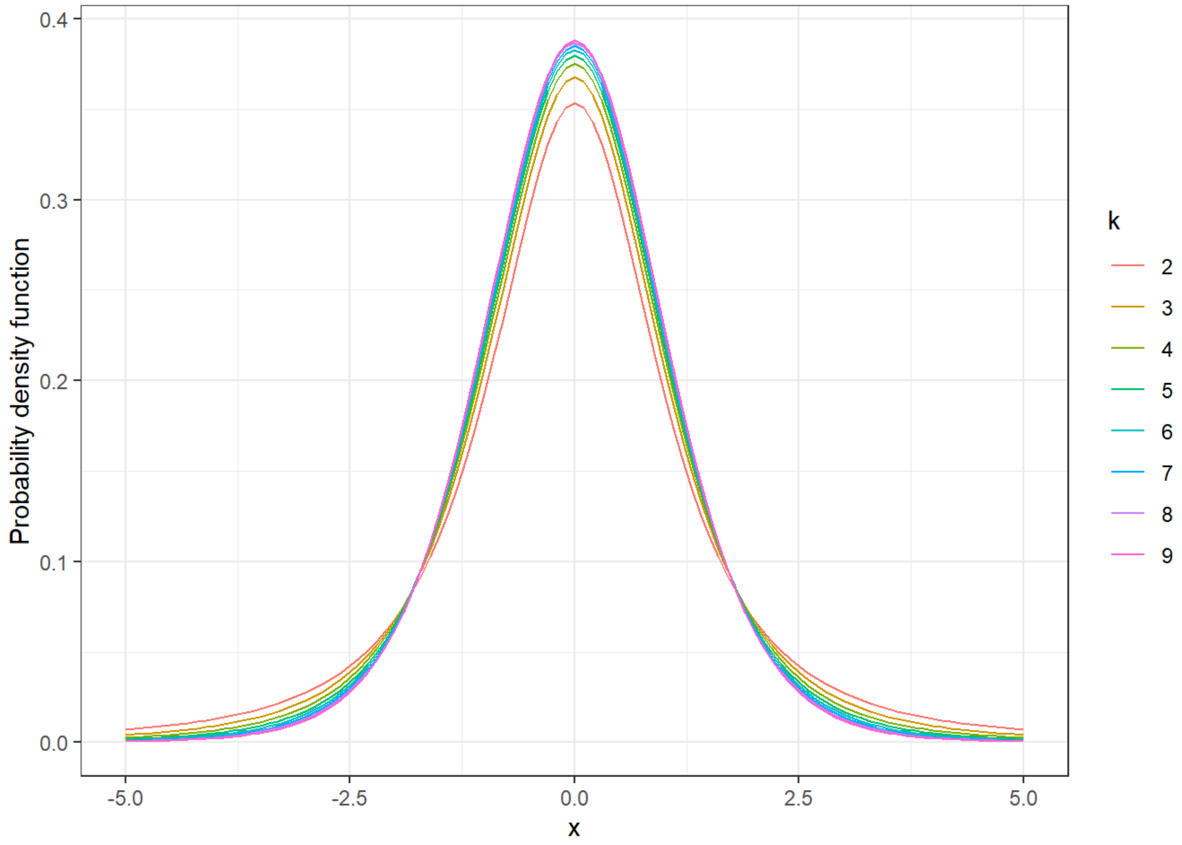
$$Q = \sum_{i=1}^{k} Z_i^2$$ with $Z_1, \cdots, Z_k \sim \mathcal{N}(0,1)$ independent and identically distributed.



Student’s t distribution

A random variable T is said to be t distributed with k degrees of freedom if $T = \dfrac{Z}{\sqrt{Q/k}}$

for two independent random variables $Z \sim \mathcal{N}(0,1)$ and $Q \sim \chi^2(k)$.



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 $Q \sim \chi^2(k)$.

The t distribution has probability density function:

$$f_k(x) = \frac{\Gamma\left(\frac{k+1}{2}\right)}{\sqrt{k\pi}\,\Gamma\left(\frac{k}{2}\right)} \left(1 + \frac{x^2}{k}\right)^{-\frac{k+1}{2}}$$

Cumulative distribution function:

$$F_k(t) := \mathbb{P}\left(T < t\right) = \int_{-\infty}^{t} f_k(x)dx.$$

Quantile function:

$\left(F_k\right)^{-1}$ which satisfies $\mathbb{P}\left(T < \left(F_k\right)^{-1}(\alpha)\right) = \alpha$

All of these functions have efficient implementations in R, python (scipy) and many other languages.

Student's t distribution

Lemma 1. *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* $S^2 := \frac{1}{n-1}\left(X_i - \overline{X}\right)^2$. *Then the random variable*

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

is t-distributed with $n-1$ *degrees of freedom.*

This lemma follows from a result known as Cochran's theorem.

Note that by dividing through by the sample variance the distribution only depends on n.

Computing Student's t confidence intervals

Suppose $X_1, \cdots, X_n \sim \mathcal{N}(\mu, \sigma^2)$ are independent and identically distributed.

Then $T := \dfrac{\overline{X} - \mu}{S/\sqrt{n}}$ is t-distributed with $n-1$ degrees of freedom.

Compute $t_{\alpha/2,n-1} > 0$ so that $\mathbb{P}\left(-t_{\alpha/2,n-1} < T < t_{\alpha/2,n-1}\right) = 1 - \alpha.$

Hence,

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

$\left(\overline{X} - \frac{t_{\alpha/2,n-1}}{\sqrt{n}} \cdot S, \overline{X} + \frac{t_{\alpha/2,n-1}}{\sqrt{n}} \cdot S\right)$ is a $(1-\alpha) \times 100\%$-level confidence interval.

Computing Student's t confidence intervals in R

To compute $\left(\overline{X} - \frac{t_{\alpha/2,n-1}}{\sqrt{n}} \cdot S, \overline{X} + \frac{t_{\alpha/2,n-1}}{\sqrt{n}} \cdot S\right)$ note that by symmetry

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

$$\iff t_{\alpha,n-1} = \left(F_{n-1}\right)^{-1}\left(1 - \frac{\alpha}{2}\right)$$

<u>Example</u>

We want to compute 95% confidence interval for the population mean with sample vector v.

```
alpha<-0.05
n<-length(v)
t<-qt(1-alpha/2,df=n-1) #This is the quantile function for the t distriution.
l<-mean(v)-t/sqrt(n)*sd(v)
u<-mean(v)+t/sqrt(n)*sd(v)
c(l,u)
```

```
## [1] 188.9021 191.0052
```

What can we infer from our sample statistics?

Sample size: $n = 151$ Sample mean: $\overline{X} = 190$ Sample variance: $s^2 = 42.8$

We have deduced a 95% confidence interval of (188.9, 191.0) for the population mean μ .

Can we reject the hypothesis that $\mu = 200$?

If $\mu \geq 200$ then the probability of observing $\overline{X} + \frac{t_{\alpha/2,n-1}}{\sqrt{n}} \cdot S < 200$ would be less than 5%.

On these grounds we can reasonably reject the hypothesis that $\mu = 200$.

We can formalize this idea with statistical hypothesis testing.

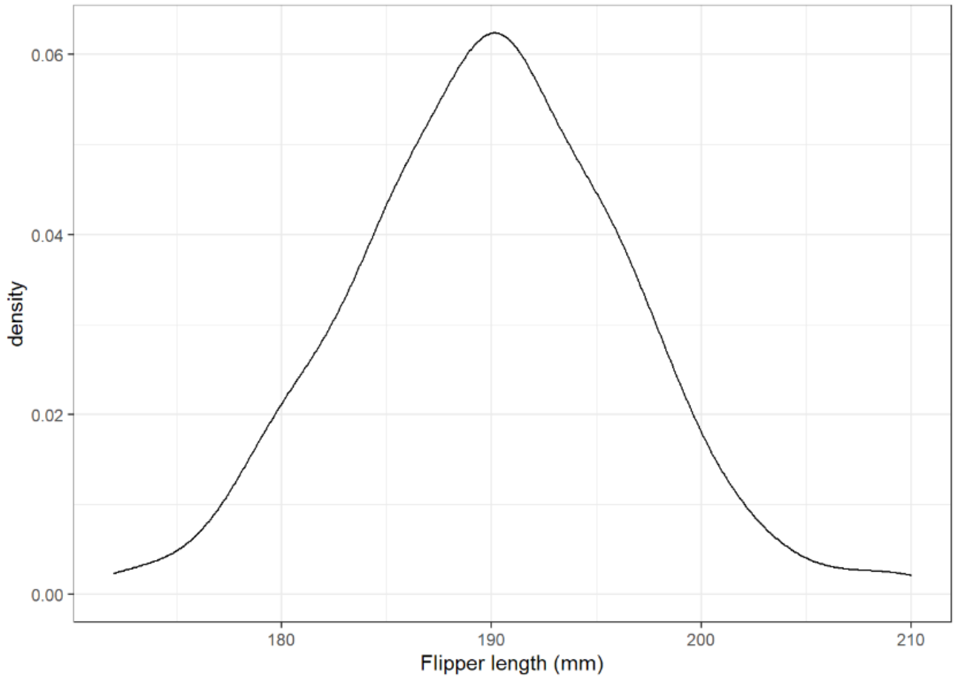
…….. But our conclusions are only as valid as our assumptions!

A Gaussian model for our data?

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

First 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)")
```



A Gaussian model seems reasonable

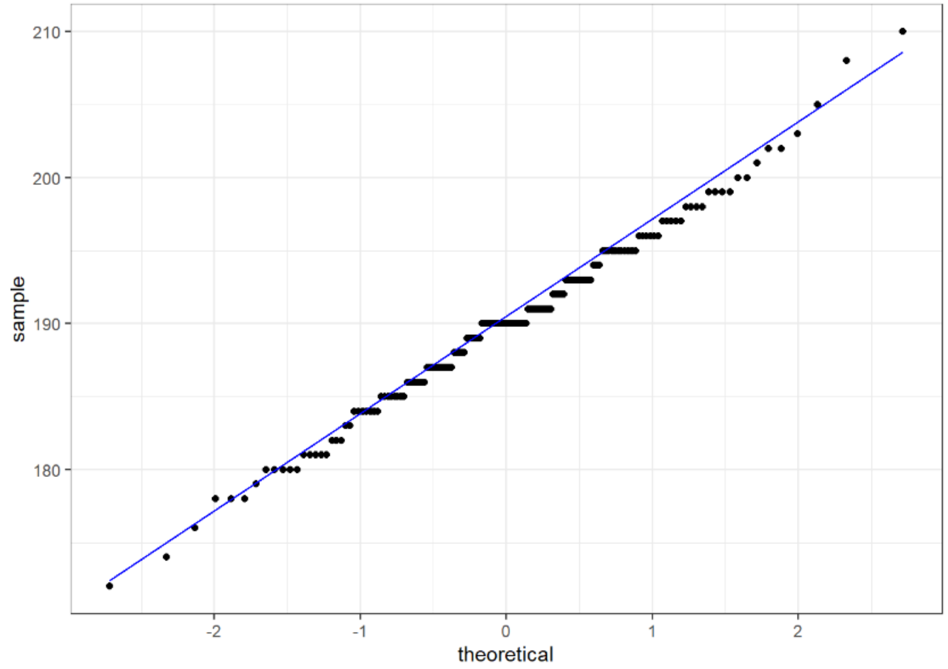
- The data looks unimodal (a single peak)
- The data looks symmetric about its mean.

A Gaussian model for our data?

Our second check that a Gaussian model $X_1, \cdots, X_n \sim \mathcal{N}(\mu, \sigma^2)$ is reasonable is a QQ-plot.

The QQ-plot compares the quantiles in the sample (y-axis) with theoretical quantiles from a Gaussian (x-axis).

```
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..

We will return to what to do otherwise in later lectures…

Non-Gaussian data and the central limit theorem

In practice our data is rarely exactly Gaussian.

Theorem 2 (Lindberg, 1920). *Let $(X_i)_{i=1}^{\infty}$ be a sequence of independent and identically distributed real-valued random variables with mean μ and variance $\sigma^2 < \infty$. Let $Z \sim \mathcal{N}(0,1)$ be a standard Gaussian random variable. Then for all $t > 0$,*

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

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

This justifies using confidence intervals based on Student's t-distribution for large sample size.

Now take a break!

Confidence intervals for proportions

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

<u>Examples</u>

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 $q = \mathbb{P}[X_i = 1]$

Confidence intervals for proportions

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 $q = \mathbb{P}\left[X_i = 1\right]$

From the central limit theorem we have $\frac{1}{n}\sum_{i=1}^{n} X_i$ approximates $\mathcal{N}\left(q, \frac{q(1-q)}{n}\right)$

$$\Rightarrow \sqrt{\frac{n}{q(1-q)}} \cdot \left(\frac{1}{n}\sum_{i=1}^{n} X_i - q\right) \text{ approximates } \mathcal{N}(0,1)$$

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 PropsCIs package to compute confidence intervals via Wilson’s method.

```
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
```

```
library(PropCIs)
```

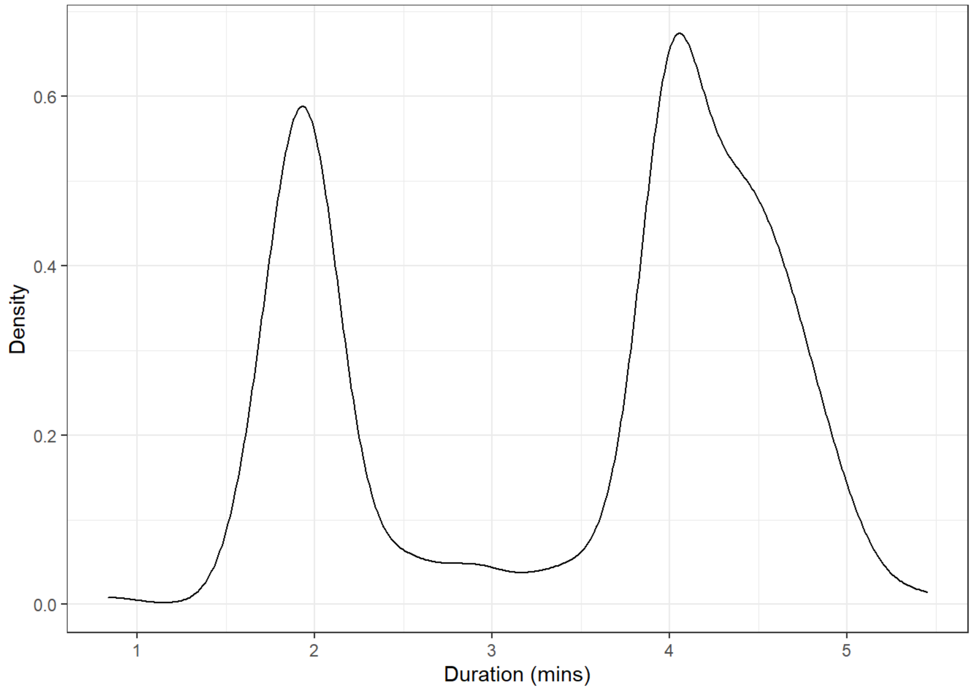
```
alpha<-0.05 # failure probability
num_successes<- sum(driving_test_results) # total passes
sample_size<-length(driving_test_results)

scoreci(x=num_successes, n=sample_size, conf.level=1-alpha) # compute Wilson's confidence intervals
```

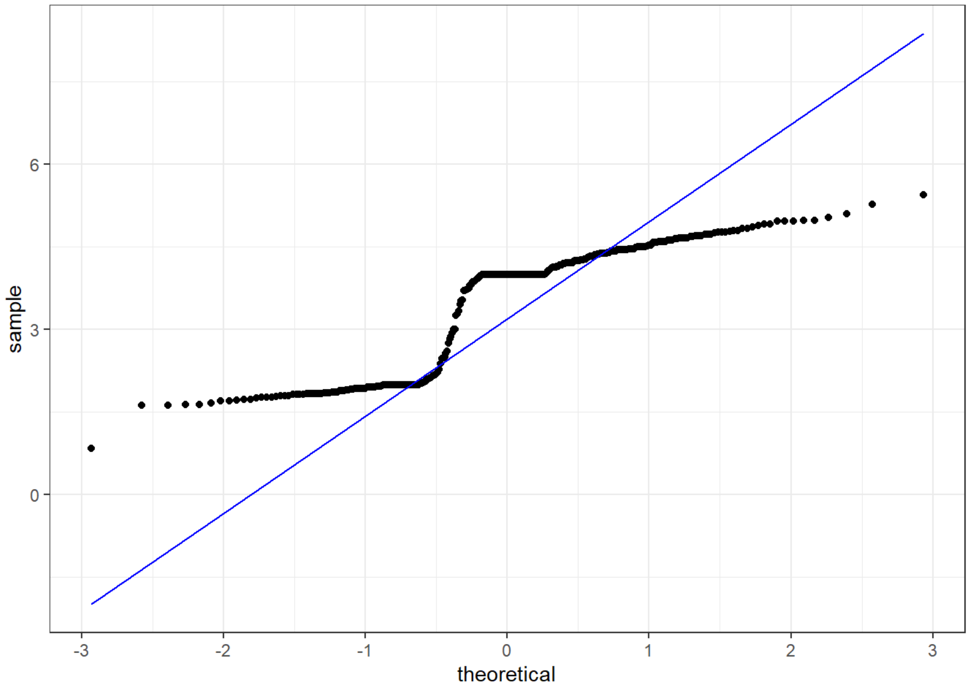
```
## 95 percent confidence interval:
##  0.1797 0.5329
```

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





Can we still compute confidence intervals?

The Bootstrap method

Suppose we have an independent and identically distributed sample $X_1, \cdots, X_n \sim \mathrm{P}$.

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 $\hat{\theta} - \theta$.

In an ideal world we would generate multiple samples and study their distribution about θ :

$$
\begin{array}{ll}
X_1^1, \cdots, X_n^1 \sim \mathrm{P} & \hat{\theta}^1 - \theta = g(X_1^1, \cdots, X_n^1) - \theta \\
\vdots & \vdots \\
X_1^B, \cdots, X_n^B \sim \mathrm{P} & \hat{\theta}^B - \theta = g(X_1^B, \cdots, X_n^B) - \theta.
\end{array}
$$

This is impossible since we don't know θ and we only have access to a single sample $X_1, \cdots, X_n$.

The Bootstrap method

In an ideal world we would generate multiple samples and study their distribution about θ :

$$X_1^1, \cdots, X_n^1 \sim \mathrm{P} \qquad \hat{\theta}^1 - \theta = g(X_1^1, \cdots, X_n^1) - \theta$$

$$\vdots \qquad\qquad\qquad \vdots$$

$$X_1^B, \cdots, X_n^B \sim \mathrm{P} \qquad \hat{\theta}^B - \theta = g(X_1^B, \cdots, X_n^B) - \theta.$$

We generate an empirical distribution $\hat{\mathrm{P}}_n$ which approximates P as follows:

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

Sampling from $\hat{\mathrm{P}}_n$ is equivalent to randomly sampling from $X_1, \cdots, X_n$ <u>with replacement</u>.

The Bootstrap method

In an ideal world we would generate multiple samples and study their distribution about θ :

$$X_1^1, \cdots, X_n^1 \sim \mathrm{P} \qquad \hat{\theta}^1 - \theta = g(X_1^1, \cdots, X_n^1) - \theta$$

$$\vdots \qquad \vdots$$

$$X_1^B, \cdots, X_n^B \sim \mathrm{P} \qquad \hat{\theta}^B - \theta = g(X_1^B, \cdots, X_n^B) - \theta.$$

Instead we use the empirical distribution $\hat{\mathrm{P}}_n$ to generate multiple Bootstrap proxies for $\hat{\theta} - \theta$:

$$\tilde{X}_1^1, \cdots, \tilde{X}_n^1 \sim \hat{\mathrm{P}}_n \qquad \tilde{\theta}^1 - \hat{\theta} = g(\tilde{X}_1^1, \cdots, \tilde{X}_n^1) - \hat{\theta}$$

$$\vdots \qquad \vdots$$

$$\tilde{X}_1^B, \cdots, \tilde{X}_n^B \sim \hat{\mathrm{P}}_n \qquad \tilde{\theta}^B - \hat{\theta} = g(\tilde{X}_1^B, \cdots, \tilde{X}_n^B) - \hat{\theta}.$$

The Bootstrap method

Instead we use the empirical distribution $\hat{\mathrm{P}}_n$ to generate multiple Bootstrap proxies for $\hat{\theta} - \theta$:

$$\tilde{X}_1^1, \cdots, \tilde{X}_n^1 \sim \hat{\mathrm{P}}_n \qquad \tilde{\theta}^1 - \hat{\theta} = g(\tilde{X}_1^1, \cdots, \tilde{X}_n^1) - \hat{\theta}$$

$$\vdots \qquad \vdots$$

$$\tilde{X}_1^B, \cdots, \tilde{X}_n^B \sim \hat{\mathrm{P}}_n \qquad \tilde{\theta}^B - \hat{\theta} = g(\tilde{X}_1^B, \cdots, \tilde{X}_n^B) - \hat{\theta}.$$

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

Let $\hat{\delta}_{\alpha/2}$ and $\hat{\delta}_{1-\alpha/2}$ denote the $\frac{\alpha}{2}$ and $1 - \frac{\alpha}{2}$ quantiles for $\tilde{\theta}_1 - \hat{\theta}_1, \cdots, \tilde{\theta}_B - \hat{\theta}_B$.

When B is sufficiently large (by the law of large numbers) we have,

$$\mathbb{P}\left(\tilde{\theta} - \hat{\theta} < \hat{\delta}_{\alpha/2}\right) \lessapprox \frac{\alpha}{2} \quad \text{and} \quad \mathbb{P}\left(\tilde{\theta} - \hat{\theta} \leq \hat{\delta}_{1-\alpha/2}\right) \gtrapprox 1 - \frac{\alpha}{2} .$$

The Bootstrap method

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

Let $\hat{\delta}_{\alpha/2}$ and $\hat{\delta}_{1-\alpha/2}$ denote the $\frac{\alpha}{2}$ and $1-\frac{\alpha}{2}$ quantiles for $\tilde{\theta}_1-\hat{\theta}_1,\cdots,\tilde{\theta}_B-\hat{\theta}_B$.

When B is large, $\mathbb{P}\left(\tilde{\theta}-\hat{\theta}<\hat{\delta}_{\alpha/2}\right)\lessapprox\frac{\alpha}{2}$ and $\mathbb{P}\left(\tilde{\theta}-\hat{\theta}\le\hat{\delta}_{1-\alpha/2}\right)\gtrapprox 1-\frac{\alpha}{2}$

$$\begin{aligned}
1-\alpha &\lessapprox \mathbb{P}\left(\tilde{\theta}-\hat{\theta}\le\hat{\delta}_{1-\alpha/2}\right)-\mathbb{P}\left(\tilde{\theta}-\hat{\theta}<\hat{\delta}_{\alpha/2}\right)\\
&=\mathbb{P}\left(\hat{\delta}_{\alpha/2}\le\tilde{\theta}-\hat{\theta}\le\hat{\delta}_{1-\alpha/2}\right)\\
&\approx\mathbb{P}\left(\hat{\delta}_{\alpha/2}\le\hat{\theta}-\theta\le\hat{\delta}_{1-\alpha/2}\right)\\
&=\mathbb{P}\left(\hat{\theta}-\hat{\delta}_{1-\alpha/2}\le\theta\le\hat{\theta}-\hat{\delta}_{\alpha/2}\right).
\end{aligned}$$

The Bootstrap method

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

Let $\hat{\delta}_{\alpha/2}$ and $\hat{\delta}_{1-\alpha/2}$ denote the $\frac{\alpha}{2}$ and $1-\frac{\alpha}{2}$ quantiles for $\tilde{\theta}_1-\hat{\theta}_1,\cdots,\tilde{\theta}_B-\hat{\theta}_B$.

When B is large, $$1-\alpha \lessapprox \mathbb{P}\left(\hat{\theta}-\hat{\delta}_{1-\alpha/2}\leq\theta\leq\theta-\hat{\delta}_{\alpha/2}\right)$$

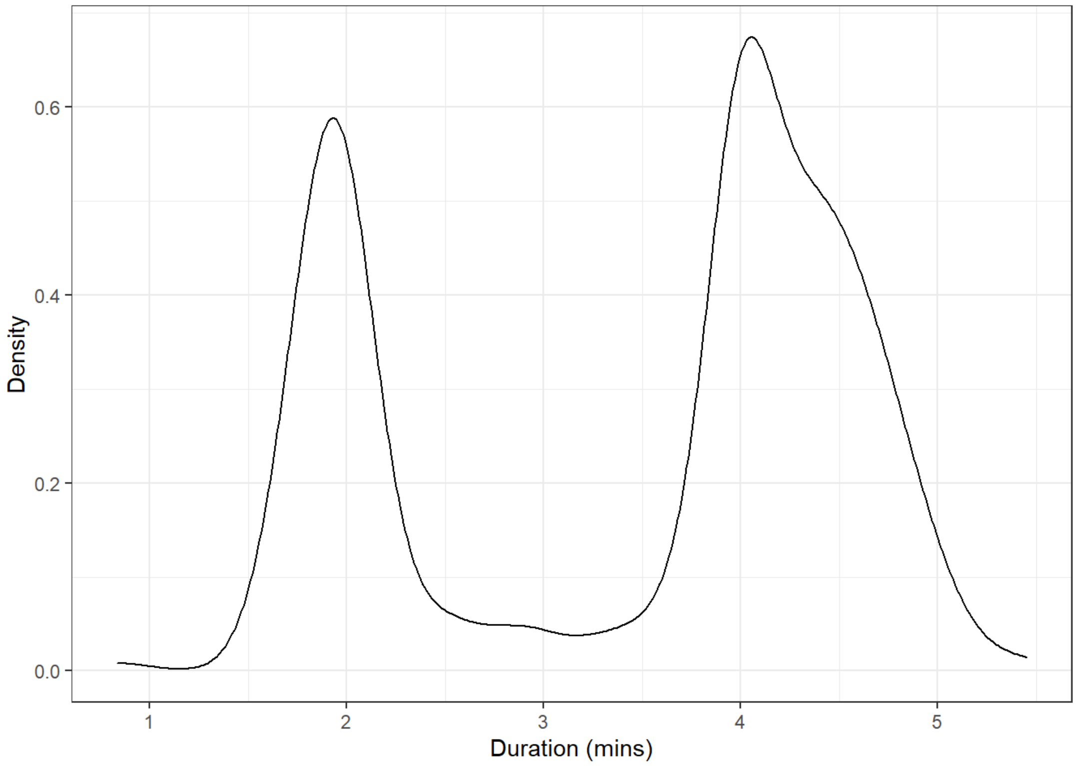
The empirical Bootstrap $(1-\alpha)\times 100\%$ -level confidence interval for the parameter θ is

$$[\hat{\theta}-\hat{\delta}_{1-\alpha/2},\hat{\theta}-\hat{\delta}_{\alpha/2}]$$

Note that empirical Bootstrap method gives *approximate* confidence intervals under general conditions.

The Bootstrap method

We want to compute a 99%-level confidence interval for the median for the volcano data set.



The Bootstrap method

We want to compute a 99%-level confidence interval for the median for the volcano data set.

```
library(boot) # load the library
set.seed(123) # set random seed

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

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

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

The Bootstrap method

```
library(boot) # load the library
set.seed(123) # set random seed

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

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

# 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%   ( 4.000,  4.033 )
## 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:

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

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!

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

  ⇒ A good option is Wilson's score interval.

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

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

  b) Use the Bootstrap methodology!

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!

Henry W J Reeve

henry.reeve@bristol.ac.uk

Statistical Computing & Empirical Methods (EMATM0061)

MSc in Data Science, Teaching block 1, 2021.