Confidence Intervals

using mtcars dataset

Lecture by Christine Mallot - DSTI S21

Benoit Mialet

20/09/2021

I) Sample and population variance

```r
library(prettyR) # for method valid.n()
```

The two following estimators for σ^2 are used to build confidence intervals (CI).

I.1) Sample variance

- If we work on a sample of the whole population, **expectation μ is unknown** and estimated by $\overline{X_n}$. In this case, **sample variance $\hat{\sigma_n}^2$** is an unbiased and consistent estimator for σ^2:

$$\hat{\sigma_n}^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \overline{X_n})^2$$

```r
X <- mtcars$wt          # random variable
var(X, na.rm=TRUE)      # sample variance
```

```
## [1] 0.957379
```

```r
sd(X, na.rm=TRUE)       # sample standard deviation
```

```
## [1] 0.9784574
```

I.2) Population variance

- If **expectation μ is known**, **population variance σ^2** is an unbiased estimator for σ^2:

$$\tilde{\sigma_n}^2 = \frac{1}{n}\sum_{i=1}^{n}(X_i - \mu)^2$$

```r
X <- mtcars$wt                    # random variable
mu <- 3.2                         # expectation (arbitrary value)
pop_var <- 1/(valid.n(X, na.rm=TRUE))*sum((X - mu)^2)     # population variance
pop_sd <- sqrt(pop_var)                                   # population standard deviation
pop_var
```

```
## [1] 0.9277584
```

```r
pop_sd
```

```
## [1] 0.9632022
```

I) Confidence interval of the expectation μ

```r
library(prettyR) # for method valid.n()
```

- some terms:
  - X: random variable
  - $\overline{X_n}$: empirical mean for n individuals
  - σ^2: variance
  - $\hat{\sigma}^2$: sample variance (estimator for σ^2)
  - $\tilde{\sigma}^2$: population variance (estimator for σ^2)
  - $1-\alpha$: confidence level
  - $Z_{(1-\alpha/2)}$: quantile of order $1-\alpha/2$ for a Normal random variable
  - $t_{(1-\alpha/2;n-)}$: quantile of order $1-\alpha/2$ for a Student random variable with $n-1$ degree of freedom

I.1) If X is not assumed to be gaussian, and variance σ^2 unknown (most frequent case):

Asymptotic symmetric CI, thanks to the Central Limit Theorem, is:

$$\left[\overline{X_n} - \frac{\hat{\sigma_n}}{\sqrt{n}} Z_{(1-\alpha/2)}; \overline{X_n} + \frac{\hat{\sigma_n}}{\sqrt{n}} Z_{(1-\alpha/2)}\right]$$

With $\overline{X_n}$ the empirical mean and $\hat{\sigma_n}$ the population standard deviation

```r
X <- mtcars$wt          # random variable
alpha <- 0.05           # confidence level = 1-alpha = 0.95
IC_min <- mean(X, na.rm=TRUE)-sd(X, na.rm=TRUE)/sqrt(valid.n(X, na.rm=TRUE))*qnorm(1-alpha/2)
IC_max <- mean(X, na.rm=TRUE)+sd(X, na.rm=TRUE)/sqrt(valid.n(X, na.rm=TRUE))*qnorm(1-alpha/2)
#Asymptotic CI:
mean(X)
```

```
## [1] 3.21725
```

```
IC_min
```

```
## [1] 2.878238
```

```
IC_max
```

```
## [1] 3.556262
```

I.2) If X is assumed to be gaussian, and variance σ^2 known:

Asymptotic symmetric CI is:

$$\left[\overline{X_n} - \frac{\sigma}{\sqrt{n}} Z_{1-\alpha/2}; \overline{X_n} + \frac{\sigma}{\sqrt{n}} Z_{1-\alpha/2}\right]$$

With $\overline{X_n}$ the empirical mean

```
X <- mtcars$wt          # random variable
alpha <- 0.05           # confidence level = 1-alpha = 0.95
sd <- 0.9               # standard deviation (arbitrary value)
IC_min <- mean(X, na.rm=TRUE)-sd/sqrt(valid.n(X, na.rm=TRUE))*qnorm(1-alpha/2)
IC_max <- mean(X, na.rm=TRUE)+sd/sqrt(valid.n(X, na.rm=TRUE))*qnorm(1-alpha/2)
#Asymptotic CI:
mean(X)
```

```
## [1] 3.21725
```

```
IC_min
```

```
## [1] 2.905422
```

```
IC_max
```

```
## [1] 3.529078
```

I.3) If X is gaussian, and variance σ^2 unknown:

Exact symmetric CI is:

$$\left[\overline{X_n} - \frac{\hat{\sigma_n}}{\sqrt{n}} t_{(1-\alpha/2;n-1)}; \overline{X_n} + \frac{\hat{\sigma_n}}{\sqrt{n}} t_{(1-\alpha/2;n-1)}\right]$$

With $\overline{X_n}$ the empirical mean and $\hat{\sigma_n}$ the sample standard deviation

```
X <- mtcars$wt          # random variable
alpha <- 0.05           # confidence level = 1-alpha = 0.95
IC_min <- mean(X, na.rm=TRUE)-sd(X, na.rm=TRUE)/sqrt(valid.n(X, na.rm=TRUE))*qt((1-alpha/2), (valid.n(X
IC_max <- mean(X, na.rm=TRUE)+sd(X, na.rm=TRUE)/sqrt(valid.n(X, na.rm=TRUE))*qt((1-alpha/2), (valid.n(X

#Exact CI:
mean(X)
```

```
## [1] 3.21725
```

```
IC_min
```

```
## [1] 2.864478
```

```
IC_max
```

```
## [1] 3.570022
```

I.4) If X is gaussian, and variance σ^2 known:

Exact symmetric CI is:

$$\left[\overline{X_n} - \frac{\sigma}{\sqrt{n}} Z_{1-\alpha/2} ; \overline{X_n} + \frac{\sigma}{\sqrt{n}} Z_{1-\alpha/2}\right]$$

With $\overline{X_n}$ the empirical mean

```
X <- mtcars$wt         # random variable
alpha <- 0.05          # confidence level = 1-alpha = 0.95
sd <- 0.9              # standard deviation (arbitrary value)
IC_min <- mean(X, na.rm=TRUE)-sd/sqrt(valid.n(X, na.rm=TRUE))*qnorm((1-alpha/2), (valid.n(X, na.rm=TRUE
IC_max <- mean(X, na.rm=TRUE)+sd/sqrt(valid.n(X, na.rm=TRUE))*qnorm((1-alpha/2), (valid.n(X, na.rm=TRUE
#Exact CI:
mean(X)
```

```
## [1] 3.21725
```

```
IC_min
```

```
## [1] -2.026648
```

```
IC_max
```

```
## [1] 8.461148
```

II) Confidence interval of the variance σ^2

(Gaussian distribution only)

```
library(prettyR) # for method valid.n()
```

II.1) If expectation μ is unknown (most often case):

We use **sample variance** $\hat{\sigma_n}^2$ (see I.1)). Confidence interval is:

$$\left[\frac{\hat{\sigma_n}^2.n}{C_{(1-\alpha_2;n-1)}};\frac{\hat{\sigma_n}^2.n}{C_{(\alpha_1;n-1)}}\right]$$

With $\alpha_1 + \alpha_2 = \alpha$

```
X <- mtcars$wt          # random variable
alpha <- 0.05           # confidence level = 1-alpha = 0.95
n <- (valid.n(X, na.rm=TRUE))
var(X)                  # sample variance
```

```
## [1] 0.957379
```

```
# exact CI
IC_min <- var(X)*n/qchisq((1-alpha/2), n-1)
IC_max <- var(X)*n/qchisq((alpha/2), n-1)
IC_min
```

```
## [1] 0.635184
```

```
IC_max
```

```
## [1] 1.746769
```

II.2) If expectation μ is known:

We use **population variance** $\tilde{\sigma_n}^2$ (see I.2)). Confidence interval is:

$$\left[\frac{\tilde{\sigma_n}^2.n}{C_{(1-\alpha_2;n)}};\frac{\tilde{\sigma_n}^2.n}{C_{(\alpha_1;n)}}\right]$$

With $\alpha_1 + \alpha_2 = \alpha$

```
X <- mtcars$wt          #variable
alpha <- 0.05           # confidence level = 1-alpha = 0.95
mu <- 3.2               # expectation (arbitrary value)
pop_var <- 1/(valid.n(X, na.rm=TRUE))*sum((X - mu)^2)  # population variance

# exact CI
IC_min <- pop_var*n/qchisq((1-alpha/2), n)
IC_max <- pop_var*n/qchisq((alpha/2), n)
pop_var
```

```
## [1] 0.9277584
```

```
IC_min
```

```
## [1] 0.6000001
```

```
IC_max
```

```
## [1] 1.623129
```

III) Confidence interval of a proportion

Several methods exists, most often used are:

II.1) Asymptotic confidence interval

$$\left[\overline{X_n} - \frac{1}{2\sqrt{n}} Z_{1-\alpha/2}; \overline{X_n} + \frac{1}{2\sqrt{n}} Z_{1-\alpha/2}\right]$$

With $\overline{X_n}$ the estimation of the proportion

```
X <- mtcars$wt          # random variable
alpha <- 0.05           # confidence level = 1-alpha = 0.95
X_n_bar <- 0.3          # estimation of the proportion
n <- (valid.n(X, na.rm=TRUE))
IC_min <- X_n_bar - 1/(2*sqrt(n)*qnorm(1-alpha/2))
IC_max <- X_n_bar + 1/(2*sqrt(n)*qnorm(1-alpha/2))
X_n_bar
```

```
## [1] 0.3
```

```
IC_min
```

```
## [1] 0.2549031
```

```
IC_max
```

```
## [1] 0.3450969
```

II.2) Agresti-coull Confidence interval

$$\left[\overline{X_n} - \sqrt{\frac{\overline{X_n}.(1-\overline{X_n})}{n}} Z_{1-\alpha/2}; \overline{X_n} + \sqrt{\frac{\overline{X_n}.(1-\overline{X_n})}{n}} Z_{1-\alpha/2}\right]$$

With $\overline{X_n}$ the estimation of the proportion

```
X <- mtcars$wt          # random variable
alpha <- 0.05           # confidence level = 1-alpha = 0.95
X_n_bar <- 0.3          # estimation of the proportion
n <- (valid.n(X, na.rm=TRUE))
IC_min <- X_n_bar - sqrt(X_n_bar*(1-X_n_bar)/n)*qnorm(1-alpha/2)
IC_max <- X_n_bar + sqrt(X_n_bar*(1-X_n_bar)/n)*qnorm(1-alpha/2)
X_n_bar
```

```
## [1] 0.3
```

```
IC_min
```

```
## [1] 0.1412248
```

```
IC_max
```

```
## [1] 0.4587752
```

IV) Confidence interval of the parameter λ of an exponential distribution

$$\left[\frac{1-\frac{Z_{1-\alpha/2}}{\sqrt{n}}}{\overline{X_n}};\frac{1+\frac{Z_{1-\alpha/2}}{\sqrt{n}}}{\overline{X_n}}\right]$$

With $\overline{X_n}$ the empirical mean

```
X <- seq(from = 0, to = 8, by = 0.02)   # (sequence, for simulation)
X <- dexp(X, rate = 2)#  simulation of a random variable
alpha <- 0.05         # confidence level = 1-alpha = 0.95
lambda_hat <- 1/mean(X)  # estimation of lambda
n <- (valid.n(X, na.rm=TRUE))
IC_min <- (1-qnorm(1-alpha/2)/sqrt(n))/mean(X)
IC_max <- (1+qnorm(1-alpha/2)/sqrt(n))/mean(X)
lambda_hat
```

```
## [1] 7.861718
```

```
IC_min
```

```
## [1] 7.092245
```

```
IC_max
```

```
## [1] 8.631191
```