Statistical Data Analysis, Lecture 5

dr. Dennis Dobler

Vrije Universiteit Amsterdam

3 March 2021

Topics in this course

1. Summarizing data
2. Exploring distributions
3. Density estimation
4. Bootstrap methods
5. Nonparametric tests
6. Analysis of categorical data
7. Multiple linear regression

Chapter 5: The bootstrap

Contents of Chapter 5:

1. Simulation
2. Bootstrap estimators for a distribution
    - parametric bootstrap
    - empirical bootstrap
3. Bootstrap confidence intervals
4. Bootstrap tests

bootstrap techniques

Verb [edit]

pull oneself up by one's bootstraps

1. (*idiomatic*) To begin an enterprise or recover from a setback without any outside help; to succeed only by one's own efforts or abilities. [quotations ▼]

   We can't get a loan, so we'll just have to ***pull ourselves up by our bootstraps****.*

bootstrap techniques

Verb [edit]

pull oneself up by one's bootstraps

1. (*idiomatic*) To begin an enterprise or recover from a setback without any outside help; to succeed only by one's own efforts or abilities. [quotations ▼]

   We can't get a loan, so we'll just have to ***pull ourselves up by our bootstraps****.*

Example (1)

Example $X_1, \dots, X_n \overset{i.i.d.}{\sim} P$ (unknown)

$T_n = \overline{X}$ estimator of $\mu_P = E(X_1)$.

$T_n \sim Q_P$

What is Q_P? What is $var(T_n)$?

Example (1)

Example $X_1, \ldots, X_n \overset{i.i.d.}{\sim} P$ (unknown)

$T_n = \overline{X}$ estimator of $\mu_P = E(X_1)$.

$T_n \sim Q_P$

What is Q_P? What is $var(T_n)$?
P unknown $\Rightarrow$ Q_P unknown!
(Asymptotically: normal...)

Example (1)

Example $X_1, \ldots, X_n \overset{i.i.d.}{\sim} P$ (unknown)

$T_n = \overline{X}$ estimator of $\mu_P = E(X_1)$.

$T_n \sim Q_P$

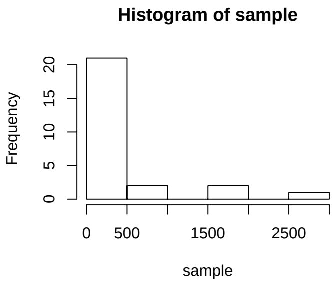
What is Q_P? What is $var(T_n)$?
P unknown $\Rightarrow$ Q_P unknown!
(Asymptotically: normal...)

More involved example:
$D_n =$ test statistic of KS test $\overset{H_0}{\sim} Q_P$ (unknown!)

Use bootstrap to estimate Q_P!

Example (2)

Example $X_1, \dots, X_n$ are data from cloud seeding.

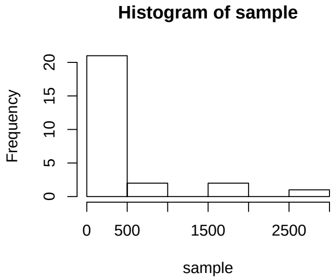



```
> mean(sample)
[1] 441.9846
```

Estimate of μ_P is $\overline{X} = 442$.

Example (2)

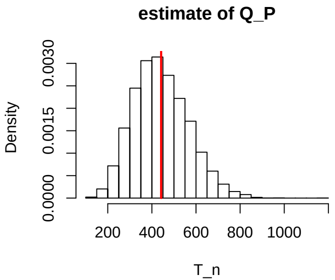
Example $X_1, \ldots, X_n$ are data from cloud seeding.




```
> mean(sample)
[1] 441.9846
```

Estimate of μ_P is $\overline{X} = 442$.

Confidence interval for μ_p??



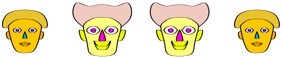
Use bootstrap to estimate Q_P.

Example empirical bootstrap

original sample:  $\Longrightarrow \quad \bar{x}$

Example empirical bootstrap

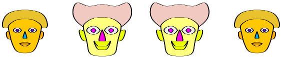
original sample: $\Longrightarrow \bar{x}$



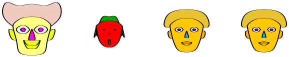
bootstrap sample 1: $\Longrightarrow \bar{x}_1^*$

Example empirical bootstrap

original sample: $\Longrightarrow \quad \bar{x}$



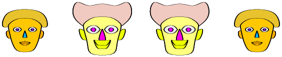
bootstrap sample 1: $\Longrightarrow \quad \bar{x}_1^*$

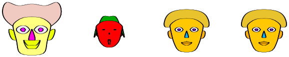


bootstrap sample 2: $\Longrightarrow \quad \bar{x}_2^*$

Example empirical bootstrap

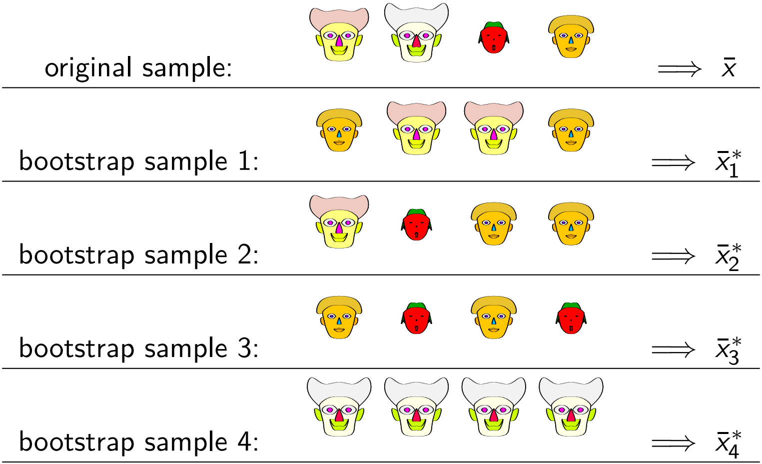
original sample:  $\Longrightarrow \bar{x}$

bootstrap sample 1:  $\Longrightarrow \bar{x}_1^*$

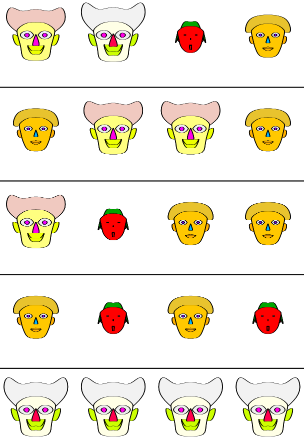
bootstrap sample 2:  $\Longrightarrow \bar{x}_2^*$

bootstrap sample 3:  $\Longrightarrow \bar{x}_3^*$

Example empirical bootstrap



Example empirical bootstrap



original sample: $\Longrightarrow \bar{x}$

bootstrap sample 1: $\Longrightarrow \bar{x}_1^*$

bootstrap sample 2: $\Longrightarrow \bar{x}_2^*$

bootstrap sample 3: $\Longrightarrow \bar{x}_3^*$

bootstrap sample 4: $\Longrightarrow \bar{x}_4^*$

Histogram of $\bar{x}_1^*, \bar{x}_2^*, \ldots, \bar{x}_{1000}^*$ approximately represents Q_P!

Empirical bootstrap set up

3 steps:

1. (Re-)Sampling
2. Recalculate estimator / statistic
3. Distribution

---

0. situation: $X_1, \dots, X_n \overset{i.i.d.}{\sim} P$, and $T_n(X_1, \dots, X_n) \sim Q = Q_P$
1. estimate P by $\tilde{P} = \hat{P}_n$ (empirical distribution)
2. instead of unknown Q_P, aim for: $Q_{\tilde{P}}$.
3. estimate $Q_{\tilde{P}}$ by empirical distribution of $T_1^*, \dots, T_B^* \overset{i.i.d.}{\sim} Q_{\tilde{P}}$.

More precise notation

- $\tilde{P}_n$ & $Q_{\tilde{P}_n}$
- $T_{n,1}^*, \dots, T_{n,B}^*$

Bootstrap sampling scheme

3 steps:

1. (Re-)Sampling
2. Recalculate estimator / statistic
3. Distribution

Concretely:

1. B times: generate $X_1^*, \ldots, X_n^* \overset{i.i.d.}{\sim} \tilde{P} = \hat{P}_n$
2. B times: compute $T_i^* = T_n(X_1^*, \ldots, X_n^*), \ i = 1, \ldots, B$
3. empirical distribution of $T_1^*, \ldots, T_B^*$ estimates $Q_{\tilde{P}}$.

Application: sample variance of $T_1^*, \ldots, T_B^*$ estimates $var(T)$.

Bootstrap sampling scheme

Example empirical bootstrap: clouds data

Example

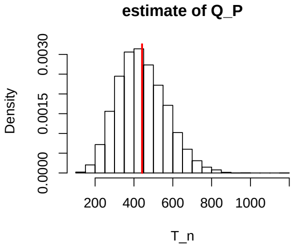
```
> B=1000
> Tstar=numeric(B)

> for(i in 1:B){

+   xstar=sample(clouds[,1], replace=TRUE)

+   Tstar[i]=mean(xstar)
+ }


> hist(Tstar)
> sd(Tstar)
[1] 125.5883
```



1. B times: generate $X_1^*, \ldots, X_n^* \overset{i.i.d.}{\sim} \tilde{P} = \hat{P}_n$
2. B times: compute $T_i^* = T_n(X_1^*, \ldots, X_n^*)$, $i = 1, \ldots, B$
3. empirical distribution of $T_1^*, \ldots, T_B^*$ estimates $Q_{\tilde{P}}$.

Parametric bootstrap set up

3 steps:

1. (Re-)Sampling
2. Recalculate estimator / statistic
3. Distribution

---

0. situation: $X_1, \ldots, X_n \overset{i.i.d.}{\sim} P_\theta$, and $T_n(X_1, \ldots, X_n) \sim Q = Q_{P_\theta}$
1. estimate P_θ by $\tilde{P} = P_{\hat{\theta}}$ (estimated parametric distribution)
2. instead of unknown Q_{P_θ}, aim for: $Q_{\tilde{P}}$.
3. estimate $Q_{\tilde{P}}$ by empirical distribution of $T_1^*, \ldots, T_B^* \overset{i.i.d.}{\sim} Q_{\tilde{P}}$.

Parametric bootstrap set up

3 steps:

1. (Re-)Sampling
2. Recalculate estimator / statistic
3. Distribution

---

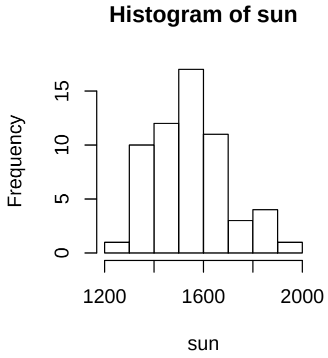
0. situation: $X_1, \dots, X_n \stackrel{i.i.d.}{\sim} P_\theta$, and $T_n(X_1, \dots, X_n) \sim Q = Q_{P_\theta}$
1. estimate P_θ by $\tilde{P} = P_{\hat{\theta}}$ (estimated parametric distribution)
2. instead of unknown Q_{P_θ}, aim for: $Q_{\tilde{P}}$.
3. estimate $Q_{\tilde{P}}$ by empirical distribution of $T_1^*, \dots, T_B^* \stackrel{i.i.d.}{\sim} Q_{\tilde{P}}$.

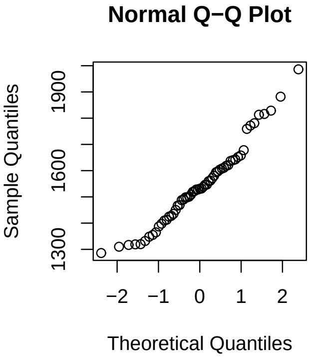
More precise notation

- $\tilde{P}_n$ & $Q_{\tilde{P}_n}$
- $\hat{\theta}_n$, $P_{\hat{\theta}_n}$ & $Q_{P_{\hat{\theta}_n}}$
- $T_{n,1}^*, \dots, T_{n,B}^*$

Example parametric bootstrap (1)

Example Yearly number of sun hours (De Bilt, 1920-1978).
Aim: standard deviation of sample median.

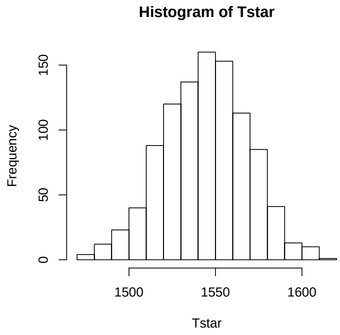




Example parametric bootstrap (2)

Assume normally distributed numbers.

```
> median(sun)
[1] 1531
> mean(sun)
[1] 1543.8
> sd(sun)
[1] 153.945
> var(sun)
[1] 23698.97
> length(sun)
[1] 59
> B=1000
> Tstar=numeric(B)
> for(i in 1:B)
+   xstar=rnorm(59, 1543.8, 153.945)
+   Tstar[i]=median(xstar)
+
> hist(Tstar)
```




```
> sd(Tstar)
[1] 24.3612
```

bootstrap errors

Two types of bootstrap errors

1. Estimate P by $\tilde{P}$ (and Q_P by $Q_{\tilde{P}}$)
2. Estimate $Q_{\tilde{P}}$ by empirical distribution of $T_1^*, \dots, T_B^* \overset{i.i.d.}{\sim} Q_{\tilde{P}}$.

Error 1: (usually) unavoidable.

Wrong parametric distribution P_θ? Big error!

$\tilde{P} = \hat{P}_n$ usually safer choice.

Error 2: depends on B.

Large $B \Leftrightarrow$ small Error 2.

E.g. $B = 1000$.

Example sun hours (1)

Aim: variance of sample *mean* of sun hours, $var(\overline{X}_n)$.

Option 1

Assuming $X_1, \ldots, X_{59} \sim N(\mu, \sigma^2)$:

estimate P by $P_{\hat{\theta}} = N(\hat{\mu}, \hat{\sigma}^2) = N(1544, 23699)$.

Theory: $\overline{X}_n \sim Q_P = N(\mu, \sigma^2/n)$

$\Rightarrow$ no need for bootstrap samples X^* and T^*!

$\Rightarrow$ Use $\hat{\sigma}^2/n = 23699/59 = 402$ to estimate $var(\overline{X})$.

Example sun hours (2)

Aim: variance of sample *mean* of sun hours, $var(\overline{X}_n)$.

Option 2

Assuming $X_1, \ldots, X_{59} \sim N(\mu, \sigma^2)$:

estimate P by $P_{\hat{\theta}} = N(\hat{\mu}, \hat{\sigma}^2) = N(1544, 23699)$.

Use $B = 1000$ bootstrap samples (cf. slide 13). Got estimate 399.

Example sun hours (3)

Aim: variance of sample *mean* of sun hours, $var(\overline{X}_n)$.

Option 3
Don't assume normality & use empirical bootstrap.

```
> var(bootstrap(sun,mean,1000))
[1] 380.055
```

Note: Outcome varies (mainly Error 2):

```
> var(bootstrap(sun,mean,1000))
[1] 425.2614
> var(bootstrap(sun,mean,10000))
[1] 402.4628
> var(bootstrap(sun,mean,10000))
[1] 390.7697
```

Example sun hours (4)

Summarizing the 3 options:

1. Parametric with normal theory (without T^*'s)
2. Parametric with bootstrap sampling (with T^*'s)
3. Empirical bootstrap sampling

Which option (1,2,3) is the best?

Example sun hours (4)

Summarizing the 3 options:

1. Parametric with normal theory (without T^*'s)
2. Parametric with bootstrap sampling (with T^*'s)
3. Empirical bootstrap sampling

Which option (1,2,3) is the best?

Option 1 only possible in special cases.
E.g., not possible for s.d. of sample median (unknown distribution)

SW-test: p=0.08 $\Rightarrow$ safest option: empirical bootstrap.

bootstrap confidence intervals

Idea

Set-up: parameter θ unknown, estimator $T \sim Q_P$ (Q_P unknown).

Accuracy of T:

- bias(T)
- var(T) or sd(T)
- confidence interval C for θ: $\mathrm{P}(C \ni \theta) = 1 - \alpha$
- ...

Confidence interval C based on Q_P.
Use bootstrap approximation $\tilde{Q}_{\tilde{P}}$!

More precisely: "T_n, $\tilde{P}_n$"

The confidence interval, before bootstrapping

T estimates θ $\Rightarrow$ $T - \theta \sim G$ concentrated around 0.

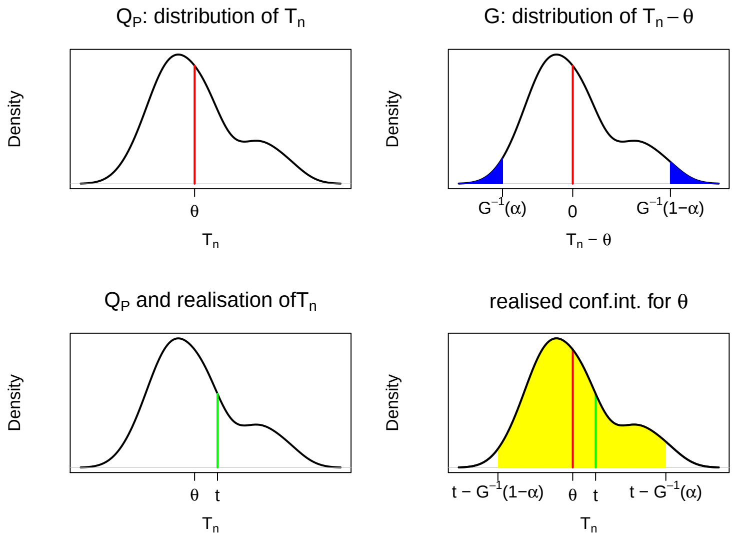
$$\mathrm{P}(G^{-1}(\alpha) \leq T - \theta \leq G^{-1}(1-\alpha)) \geq 1 - 2\alpha$$

$$\Leftrightarrow \quad \mathrm{P}(T - G^{-1}(1-\alpha) \leq \theta \leq T - G^{-1}(\alpha)) \geq 1 - 2\alpha.$$

$\Rightarrow$ $[T - G^{-1}(1-\alpha), T - G^{-1}(\alpha)]$ is $(1-2\alpha)$ confidence interval for θ.

Idea: use bootstrap to estimate quantiles of G! (Next lecture.)

In pictures


Q_P: distribution of T_n
Density
θ
T_n
G: distribution of $T_n - \theta$
Density
$G^{-1}(\alpha)$
0
$G^{-1}(1-\alpha)$
$T_n - \theta$
Q_P and realisation ofT_n
Density
θ
t
T_n
realised conf.int. for θ
Density
$t - G^{-1}(1-\alpha)$
θ
t
$t - G^{-1}(\alpha)$
T_n

Some R-code

Underlying R-code...

```
> plot(d,main=expression(paste("realised conf.int. for ",theta)),
+ xlab=expression(T[n]),lwd=2,yaxt="n",xaxt="n")
> axis(1,0,expression(theta))
> axis(1,1,"t")
> axis(1,-4,expression(paste("t - ",G^{-1},"(1-",alpha,")") ) )
> axis(1,5,expression(paste("t - ",G^{-1},"(",alpha,")") ) )
> lines(xval,yval,type="h",col="yellow")
```

`expression`: for mathematical symbols
`paste`: for combining variables and text
`axis`: for plotting tickmarks along axes
`lines`/`plot` with `type="h",col="..."`: for coloured graphs

to finish

To summarize

Today we discussed

- Simulation
- Bootstrap estimators for a distribution
  - parametric bootstrap
  - empirical bootstrap
- Bootstrap confidence intervals
- Bootstrap tests

Next week Bootstrap confidence intervals
Bootstrap tests