Statistical Data Analysis, Lecture 6

dr. Dennis Dobler

Vrije Universiteit Amsterdam

International Women's Day + 2, 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 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

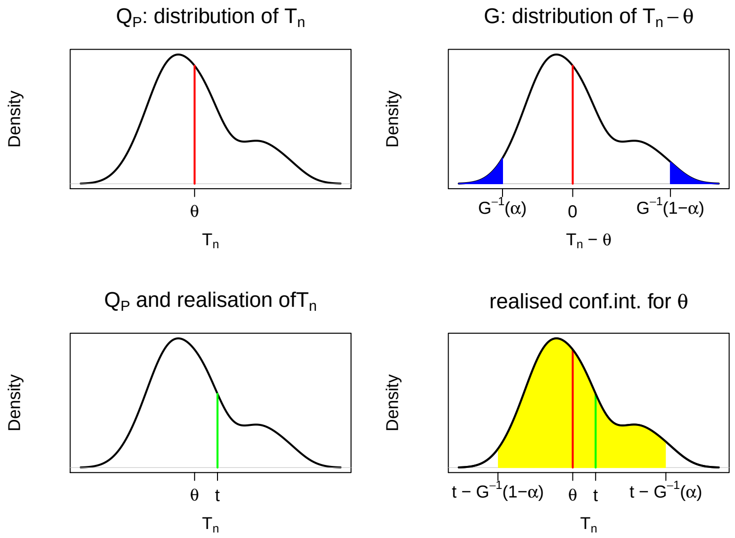
T estimates $\theta \;\Rightarrow\; T - \theta \sim G$ concentrated around 0.

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

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

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

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

The bootstrap confidence interval (1)

In confidence interval $[T - G^{-1}(1-\alpha), T - G^{-1}(\alpha)]$ unknown:

- G, i.e. the distribution of $T - \theta$,
- Q_P, i.e. the distribution of T,
- θ, the parameter of interest.

Hence, estimate G

by empirical distribution of $Z_i^* = T_i^* - T$, $i = 1, \ldots, B$.

$T_1^*, \ldots, T_B^*$ bootstrap realizations (empirical or parametric).

$G^{-1}(\alpha)$ by $Z^*_{([\alpha B])}$.
$G^{-1}(1-\alpha)$ by $Z^*_{([(1-\alpha)B])}$.

R: quantile

The bootstrap confidence interval (2)

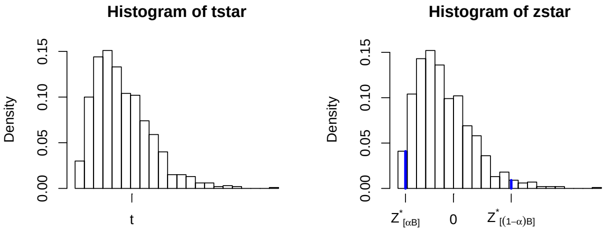
Instead of $[T - G^{-1}(1-\alpha), T - G^{-1}(\alpha)]$ (unknown)

Use $[T - Z^*_{([(1-\alpha)B])}, T - Z^*_{([\alpha B])}]$ (known!)

$$= [2T - T^*_{([(1-\alpha)B])}, 2T - T^*_{([\alpha B])}]$$

because $Z^*_i = T^*_i - T$.

The bootstrap confidence interval (3)




```
> zstar=tstar-tn
> c(tn-quantile(zstar,0.975),tn-quantile(zstar,0.025))
     97.5%       2.5%
-0.1129308 11.3179222
> 2*tn-quantile(tstar,c(0.975,0.025))
     97.5%       2.5%
-0.1129308 11.3179222
```

FYI: Reliability of a confidence interval

Problem Actual coverage probability only $\approx 1 - \alpha$...

Question Which approach is most trustworthy?

Approach Simulate actual coverage probability of confidence interval:

Pick θ & estimator T_n of θ.

Do e.g. $K = 10000$ times:

1. generate $x_1, \ldots, x_n \sim P_\theta$,
2. derive $T_n(x_1, \ldots, x_n)$, generate $T_1^*, \ldots, T_B^*$,
3. construct confidence interval C,
4. is $\theta \in C$?

Coverage probability $\approx$ relative frequency of "$\theta \in C$".

bootstrap tests

Bootstrap test

Situation $X_1, \ldots, X_n \stackrel{i.i.d.}{\sim} P$ (unknown)
Aim: goodness-of-fit hypothesis testing

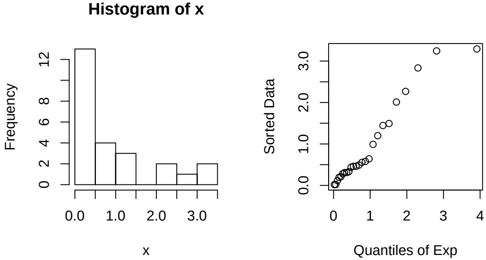
$$H_0 : P \in \mathcal{P}_0 \quad \text{versus} \quad H_1 : P \notin \mathcal{P}_0$$

$\mathcal{P}_0$ collection of distributions.

Test statistic $T \sim Q_P$.
Problem: Q_P unknown for all $P \in \mathcal{P}_0$!

Idea: Bootstrap! Estimate Q_P by $\tilde{Q}_{\tilde{P}}$.

Example (1)



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

$$H_0: \quad X_1, \ldots, X_n \sim \text{Exp}(\lambda) \text{ for some } \lambda > 0$$
$$H_1: \quad X_1, \ldots, X_n \text{ are not exponentially distributed}$$

Possible test statistic: $\quad T = \dfrac{median(X)}{mean(X)} \sim Q_P$

Simulate T under H_0, because $Q_{Exp(\lambda)}$ unknown.

Example (2)

$Q_{Exp(\lambda)}$ independent of λ!

$\Rightarrow$ T is "nonparametric".

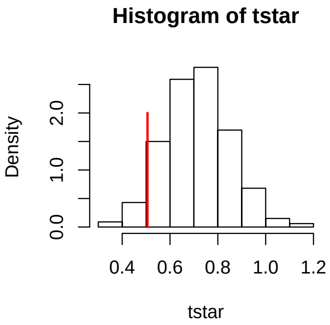
Simulate $Q_{Exp(1)}$ via (parametric) bootstrap: B times

- generate $X_1^*, \ldots, X_n^* \overset{i.i.d.}{\sim} Exp(1)$,
- compute $T^* = median(X_1^*, \ldots, X_n^*)/mean(X_1^*, \ldots, X_n^*)$.

Remark calling this "bootstrap" is actually inappropriate!

Example (3)

```
> median(x)/mean(x)
[1] 0.5058572
> for(i in 1:B) {
+   xstar=rexp(n)
+   tstar[i]=median(xstar)/mean(xstar) }
> p=2*min(sum(tstar<=0.5058572)/B,sum(tstar>=0.5058572)/B)
> p
[1] 0.112
```



two-sided H_0 not rejected

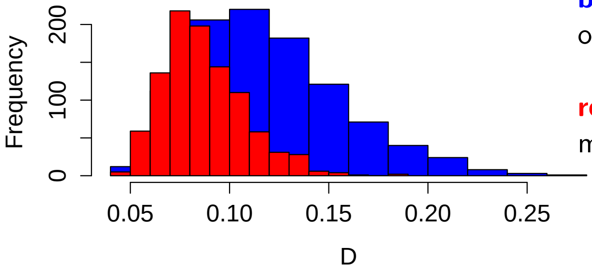
Another example (1)

Remember how not to use Kolmogorov-Smirnov test for composite

$$H_0 : X_1, \dots, X_n \sim N(\mu, \sigma^2) \quad \text{for some} \quad \mu \text{ and } \sigma^2$$

```
> ks.test(x,pnorm,mean(x),sd(x))
```

R-command "tests" simple $H_0 : X_1, \dots, X_n \sim N(\overline{X}, S_X^2)$.



blue: distribution of original KS-statistic D_n

red: distribution of modified KS-statistic $\tilde{D}_n$

Another example (2)

$\tilde{D}_n$: sensible test statistic... but p-value

```
> ks.test(x,pnorm,mean(x),sd(x))$p.val
```

is wrong; calculated from blue distribution of D_n.

Bootstrap to simulate red distribution of $\tilde{D}_n$!

Nonparametric? (see the syllabus and the assignment)

Bootstrap: Warnings

Warning Bootstrap can fail!

- parametric bootstrap & outliers in sample & sensitive parameter estimator $\Rightarrow$ bad bootstrap approximation.
- empirical bootstrap & extreme order statistics: distribution of $X_{(1)} = \min(X_1, \ldots, X_n)$ or $X_{(n)} = \max(X_1, \ldots, X_n)$.
- Heavy-tailed data distribution (Example 4.6 in syllabus: Cauchy distribution.)

to finish

To summarize

Today we discussed

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

Next week Exam preparation