Statistical Inference

Lukáš Lafférs

Matej Bel University, Dept. of Mathematics

- Maximum likelihood
- Bootstrap

Maximum likelihood

Statistical inference deals with the problem of quantifying uncertainty.

By uncertainty we mean the **statistical** uncertainty, not the ~~model~~ uncertainty.

Given the fact that our sample size is limited. How sure/unsure are we regarding our parameter estimate?

Example 1 - Tossing a coin

We observe the following

$$\underbrace{000001000010010000000010000100101000\cdots0001000010000}_{500 \ tosses}$$

97 heads, 403 tails.

These are independent coin flips of a single coin with a fixed probability of showing the head.

$$Pr(C = 97) = \binom{500}{97} p^{97} (1-p)^{403}$$

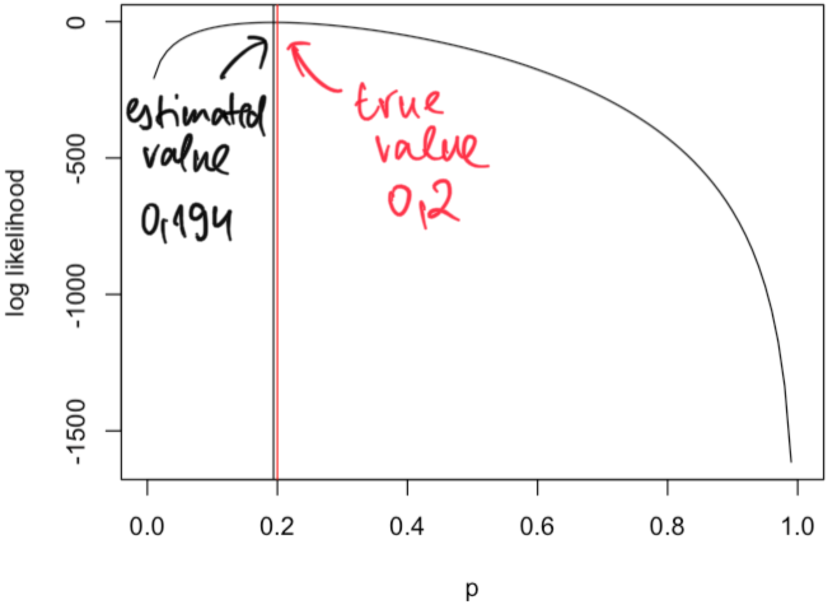
Is it fair?

If $p = 0.5$ we would see 97 heads with probability $9.31491 \cdot 10^{-46}$
(strictly mathematically speaking: not a whole lot)

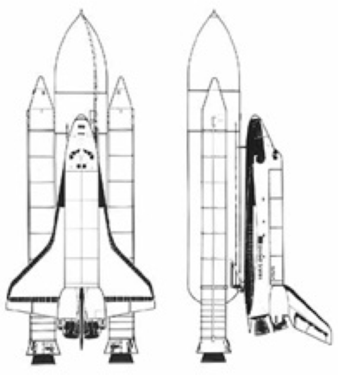
Example 1 - Tossing a coin

What value of p is the most likely?

Find the one that makes $Pr(X = 97)$ most likely.


0
-500
-1000
-1500
log likelihood
0.0
0.2
0.4
0.6
0.8
1.0
p
estimated value
0,194
true value
0,2

Example 2 - Challenger Disaster



Courtesy of NASA.

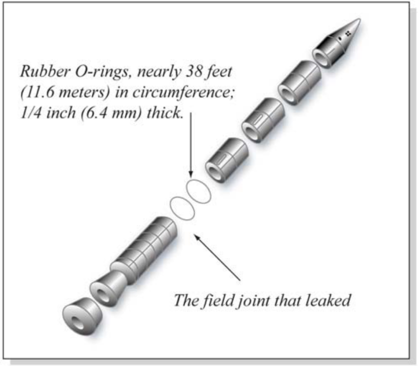
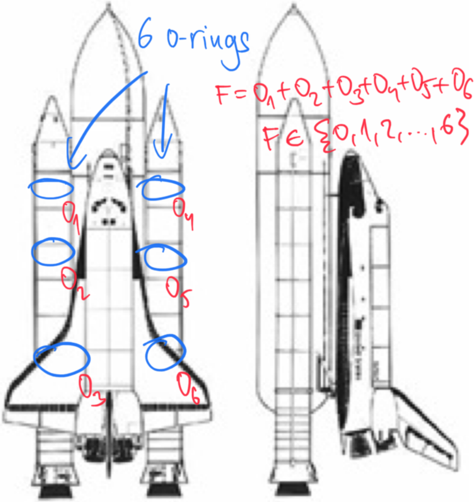
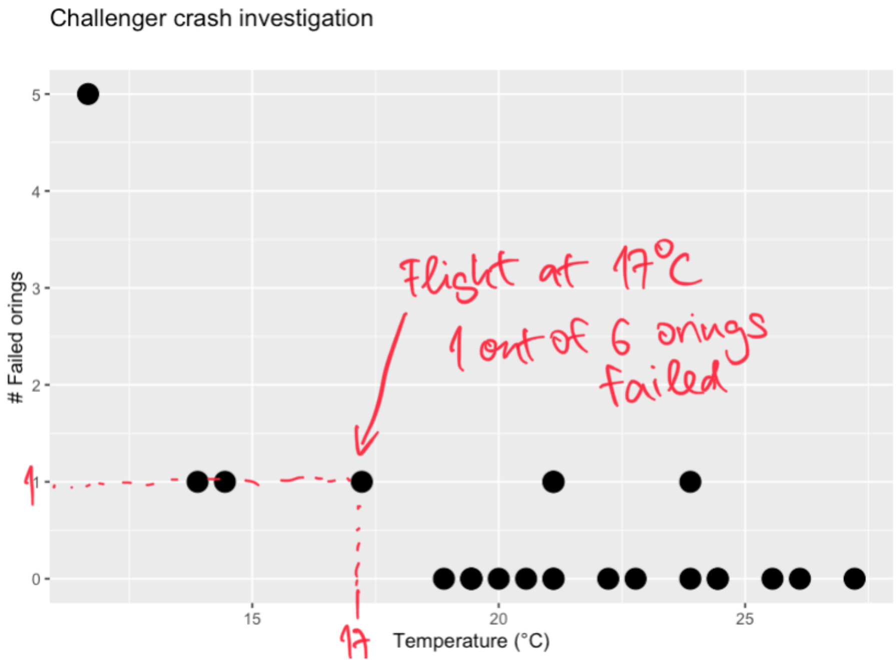


Figure by MIT OCW.



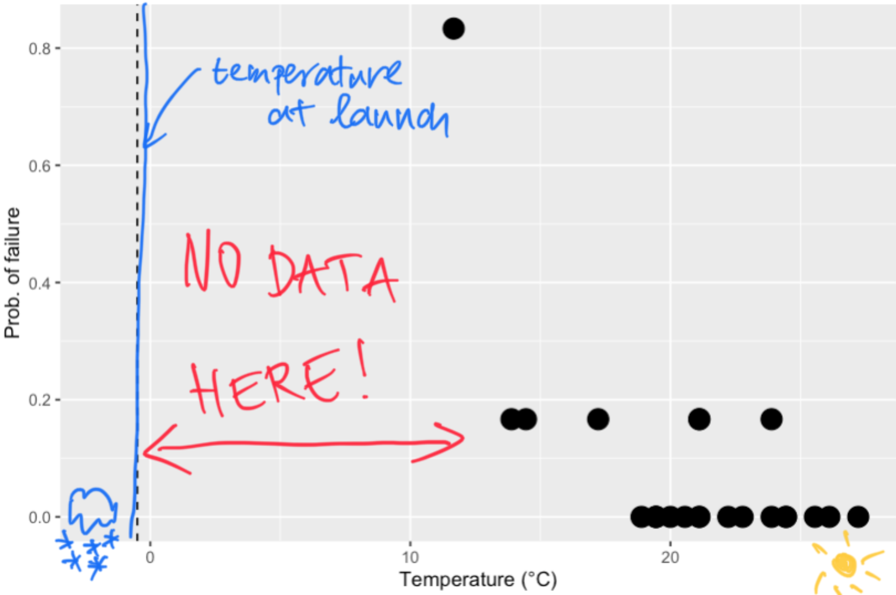
Courtesy of NASA.

- $Y_i \sim Bern(p_i)$
- $Y_i \perp Y_j$
- $F_i = \sum_{i=1}^{6} Y_i \sim Bin(6, p_i)$
- $g(p_i) = \beta_0 + \beta_1 temp_i$



Challenger crash investigation
Failed orings
5
4
3
2
1
0
15
20
25
Temperature (°C)
Flight at 17°C
1 out of 6 orings failed
1
17

Challenger crash investigation



$$p_i = \beta_0 + \beta_1 \, temp_i + \varepsilon_i$$

??? but $\hat{p}_i$ may lie outside $[0, 1]$.

What about

$$g(p_i) = \beta_0 + \beta_1 \, temp_i$$

? E.g.

$$\log\left(\frac{p_i}{1 - p_i}\right) = \beta_0 + \beta_1 \, temp_i$$

OK, but where is the random component ε_i?

- n_i
  number of **independent** Bernoulli trials
- p_i
  probability of **each** Bernoulli trial
- y_i
  number of occurrences of events $\in \{0, 1, ..., n_i\}$

$y_i \sim Bin(n_i, p_i)$

Now, the probabilistic description is complete! Everything is now known, except for unknown parameters β_0, β_1

Challenger data

- $n_i = 6$
  number of o-rings (whose failures are **independent**)
- $p_i = g^{-1}(\beta_0 + \beta_1 temp_i)$
  probability of a failure of **each** o-ring
- y_i
  number of failed o-rings $\in \{0, 1, ..., 6\}$

Event 1: 6 o-rings, 1 failure, 18.3 temperature

$$Pr(Y_1 = 1) = \binom{6}{1} p_1^1 (1 - p_1)^5 \qquad p_1 = g^{-1}(\beta_0 + \beta_1 . 18.3)$$

Event 2: 6 o-rings, 2 failures, 11.3 temperature

$$Pr(Y_2 = 2) = \binom{6}{2} p_2^2 (1 - p_2)^4 \qquad p_2 = g^{-1}(\beta_0 + \beta_1 . 11.3)$$

...

Event n: 6 o-rings, 0 failures, 20.6 temperature

$$Pr(Y_n = 0) = \binom{6}{0} p_n^0 (1 - p_n)^6 \qquad p_n = g^{-1}(\beta_0 + \beta_1 . 20.6)$$

Probability of observing (y, X) given parameter values (β_0, β_1)

$$\begin{aligned}
L(\beta_0, \beta_1 | y, X) &= Pr(Y_1 = y_1) \cdot Pr(Y_2 = y_2) \cdot Pr(Y_3 = y_3) \cdot \cdots \cdot Pr(Y_n = y_n) \\
&= \prod_{i=1}^{n} Pr(Y_i = y_i) \\
&= \prod_{i=1}^{n} \binom{n_i}{y_i} p_i^{y_i} (1 - p_i)^{n_i - y_i} \\
\log L(\beta_0, \beta_1 | y, X) &= \sum_{i=1}^{n} \log \binom{n_i}{y_i} + y_i \log(p_i) + (n_i - y_i) \log(1 - p_i)
\end{aligned}$$

What is the likelihood of observing the data?

Set $(\hat{\beta}_0, \hat{\beta}_1)$ in order to maximize $\log L(\beta_0, \beta_1 | y, X)$

Example 3 - waiting time

We observe inter-arrival times of a insurance claims (in days).

$$\underbrace{2.07 \;\; 5.06 \;\; 6.51 \;\; 1.75 \;\; 13.95 \;\; 2.55 \quad \ldots \quad 18.03 \;\; 1.92 \;\; 1.03}_{100 \;\; observations}$$

These may be exponentially distributed.

what value would fit the data best?

Notation

- X random variable
- $X_1, \ldots X_n$ iid from parametric distribution $f(x|\theta)$
- $\theta \in \Theta$ unknown parameter to be estimated. The true value is denoted as θ_0.

Example:

- $X \sim Exp(\lambda)$
- $f(x|\lambda) = \exp(-x/\lambda)/\lambda$
- $\lambda \in [0, \infty)$ unknown parameter to be estimated. The true value is denoted as λ_0.

Likelihood function: $L_n(\theta) \equiv f(X_1|\theta) \cdot \ldots \cdot f(X_n|\theta) = \prod_i f(X_i|\theta)$

- unlike density f it is a function of a parameter θ with data kept fixed
- i.i.d. is crucial

Example:

$$L_n(\lambda) = \prod_i \left(\frac{1}{\lambda} \exp\left(-\frac{X_i}{\lambda} \right) \right) = \frac{1}{\lambda^n} \exp\left(-\frac{n\bar{X}_n}{\lambda} \right)$$

Maximum likelihood estimator: $\hat{\theta} \equiv \arg\max_{\theta} L_n(\theta)$

- what parameter value can rationalise the given data best?
- the estimator is a random variable, because the data is random
- has some favourable statistical properties
- can be computed analytically or numerically

Example:

We need to solve F.O.C.:

$$0 = \frac{\partial}{\partial \lambda} L_n(\lambda) = -n \frac{1}{\lambda^{n+1}} \exp\left(-\frac{n\bar{X}_n}{\lambda}\right) + \frac{1}{\lambda^n} \exp\left(-\frac{n\bar{X}_n}{\lambda}\right) \frac{n\bar{X}_n}{\lambda^2}$$

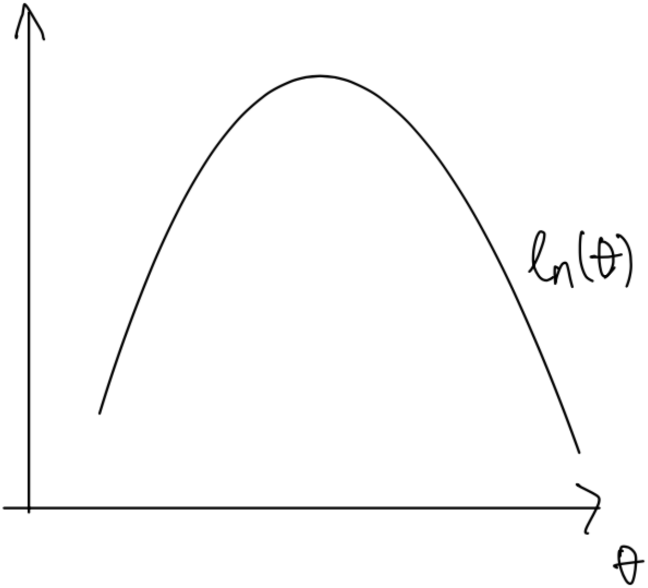
$$\hat{\lambda} = \bar{X}_n$$

Log-likelihood function: $\ell_n(\theta) \equiv \log L_n(\theta) = \sum_i \log f(X_i|\theta)$

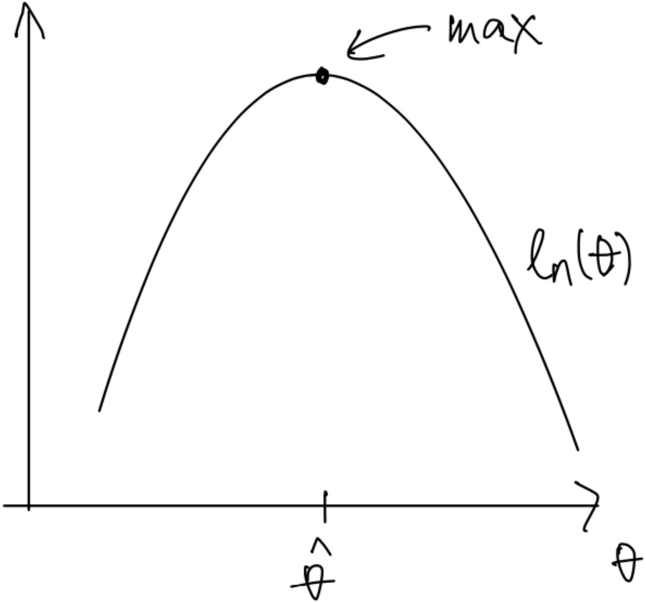
- Numerically more stable.
- $\arg\max_\theta \ell_n(\theta) = \arg\max_\theta L_n(\theta)$

Example:

$$\ell_n(\lambda) = \sum_i \log f(X_i|\theta) = \sum_i \left(-\log\lambda - \frac{X_i}{\lambda} \right) = -n\log\lambda - \frac{n\bar{X}_n}{\lambda}$$



$\ell_n(\theta)$
θ


max
ln(θ)
θ̂
θ

Expected log density $\ell(\theta) \equiv E[\log f(X|\theta)]$

- under correct specification we have likelihood analog principle: $\theta_0 = \arg\max_\theta l(\theta)$

Example:

$$\ell(\theta) = E[\log f(X|\theta)] = E[-\log\lambda - X/\lambda] = -\log\lambda - \frac{E[X]}{\lambda} = -\log\lambda - \frac{\lambda_0}{\lambda}$$

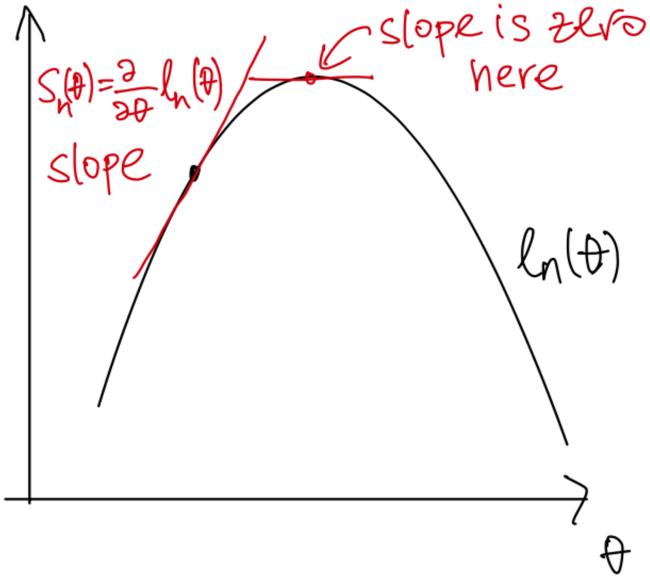
FOC gives $0 = \frac{1}{\lambda} + \frac{\lambda_0}{\lambda^2}$ which has an unique solution $\lambda = \lambda_0$.

Score function: $S_n(\theta) \equiv \frac{\partial}{\partial\theta}\ell_n(\theta) = \sum_i \frac{\partial}{\partial\theta}\log f(X_i|\theta)$

- How sensitive is the likelihood to θ
- for interior solution we have $S_n(\hat{\theta}) = 0$

Example:

$$S_n(\lambda) = \frac{\partial}{\partial\lambda}\left(-n\log\lambda - \frac{n\bar{X}_n}{\lambda}\right) = -\frac{n}{\lambda} + \frac{n\bar{X}_n}{\lambda^2}$$



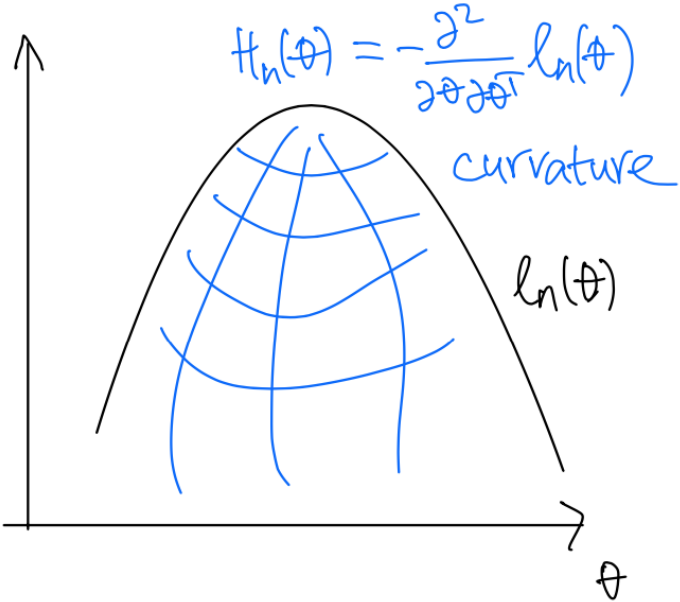
$S_n(\theta) = \frac{\partial}{\partial\theta} l_n(\theta)$
slope
slope is zero here
$l_n(\theta)$
θ

Likelihood Hessian: $H_n(\theta) \equiv -\frac{\partial^2}{\partial\theta\partial\theta^T}\ell_n(\theta) = -\sum_i \frac{\partial^2}{\partial\theta\partial\theta^T}\log f(X_i|\theta)$

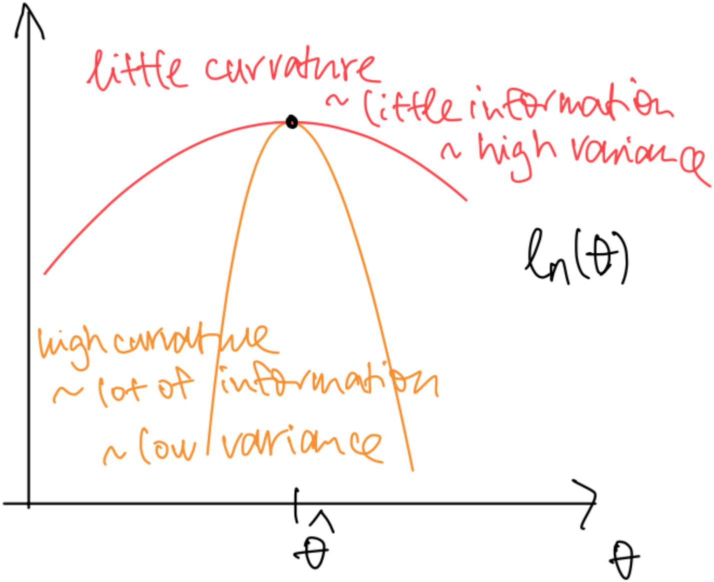
- tells us how curved is the log-likelihood

Example:

$$H_n(\lambda) = -\frac{\partial^2}{\partial\lambda^2}\ell_n(\lambda) = -\frac{\partial}{\partial\lambda}S_n(\lambda) = -\frac{n}{\lambda^2} + \frac{2n\bar{X}_n}{\lambda^3}$$



$H_n(\theta) = -\frac{\partial^2}{\partial\theta\partial\theta^T} l_n(\theta)$
curvature
$l_n(\theta)$
θ



little curvature
~ little information
~ high variance
$\ell_n(\theta)$
high curvature
~ lot of information
~ low variance
$\hat{\theta}$
θ

Efficient score: $S \equiv \frac{\partial}{\partial \theta} \log f(X|\theta_0)$

- derivative of a log-likelihood of a single observation
- mean zero random vector
- $E[S] = E\left[\frac{\partial}{\partial \theta} \log f(X|\theta_0)\right] = \frac{\partial}{\partial \theta} E\left[\log f(X|\theta_0)\right] = \frac{\partial}{\partial \theta} \ell(\theta_0) = 0$

Example:

$$S = \frac{\partial}{\partial \lambda} \log f(X|\lambda_0) = -\frac{1}{\lambda_0} + \frac{X}{\lambda_0^2}.$$

$$E[S] = -\frac{1}{\lambda_0} + \frac{E[X]}{\lambda_0^2} = -\frac{1}{\lambda_0} + \frac{\lambda_0}{\lambda_0^2} = 0$$

Fisher information: $J_\theta \equiv E[SS^T]$

- variance of the efficient score S

Example:

$$J_\lambda = \underbrace{E[S^2] = Var[S]}_{E[S]=0} = Var\left[-\frac{1}{\lambda_0} + \frac{X}{\lambda_0^2}\right] = \frac{1}{\lambda_0^4} Var[X] = \frac{1}{\lambda_0^2}$$

Expected Hessian: $H_\theta \equiv -\frac{\partial^2}{\partial\theta\partial\theta^T}\ell(\theta_0)$

- under regularity conditions $H_\theta = -E\left[\frac{\partial^2}{\partial\theta\partial\theta^T}\log f(X|\theta_0)\right]$

Example:

$$H_\theta = -\frac{\partial^2}{\partial\lambda^2}\ell(\lambda)|_{\lambda=\lambda_0} = -\frac{\partial^2}{\partial\lambda^2}\left(-\log\lambda - \frac{\lambda_0}{\lambda}\right)|_{\lambda=\lambda_0} = \frac{1}{\lambda_0^2}$$

Under correct specification of $f(x|\theta)$ (there exists some $\theta_0 \in \Theta$ so that $f(x|\theta_0) = f(x)$), we have Information Matrix Equality:

$$J_\theta = H_\theta$$

Example:

$$J_\lambda = \frac{1}{\lambda_0^2} = H_\lambda$$

MLE has some interesting properties

- invariant to transformations
- asymptotically efficient in the class of unbiased estimators (even for transformations)
- consistent
- asymptotically normal

MLE is invariant to transformations

- $\hat{\theta}$ is the MLE of $\theta \implies \hat{\beta} = h(\hat{\theta})$ is the MLE of $\beta = h(\theta)$

MLE asymptotically achieves Cramer-Rao Lower Bound

- Under (i) correct specification, (ii) support of X not being dependent on θ and (iii) θ_0 lying in the interior of Θ
- For any unbiased $\tilde{\theta}$ we have that

$$Var[\tilde{\theta}] \geq (nJ_\theta)^{-1}$$

- For transformation $\beta = h(\theta)$ (under some more regularity conditions) we get that for any unbiased estimator $\tilde{\beta}$ of β:

$$Var[\tilde{\beta}] \geq \frac{1}{n} H^T J_\theta^{-1} H$$

where $H = \frac{\partial}{\partial \theta} h(\theta_0)^T$.

Average log-likelihood: $\bar{\ell}_n(\theta) \equiv \frac{1}{n}\ell_n(\theta) = \frac{1}{n}\sum_i \log f(X_i|\theta)$

MLE is consistent, $\hat{\theta} \to_P \theta$ under these conditions:

- X_i are i.i.d.
- $E\left|\log f(X|\theta)\right| \leq G(X)$, with $E[G(X)] < \infty$
- $\log f(X|\theta)$ is continuous in θ with probability one
- Θ is compact
- $\forall \theta \neq \theta_0 : l(\theta) < l(\theta_0)$ (so that the parameter θ is identified)

MLE is asymptotically normally distributed

Why? Taylor expansion around θ_0:

$$0 = \frac{\partial}{\partial\theta}\bar{\ell}_n(\hat{\theta}) \approx \frac{\partial}{\partial\theta}\bar{\ell}_n(\theta_0) + \frac{\partial^2}{\partial\theta\partial\theta^T}\bar{\ell}_n(\theta_0)(\hat{\theta} - \theta_0)$$

$$\sqrt{n}(\hat{\theta} - \theta_0) \approx \underbrace{\underbrace{\left(\frac{\partial^2}{-\partial\theta\partial\theta^T}\bar{\ell}_n(\theta_0)\right)^{-1}}_{\to_P\ H_\theta^{-1}} \underbrace{\left(\sqrt{n}\frac{\partial}{\partial\theta}\bar{\ell}_n(\theta_0)\right)}_{\to_D\ N(0, J_\theta)}}_{\to_D\ N(0, H_\theta^{-1} J_\theta H_\theta^{-1}) = N(0, J_\theta^{-1})}$$

OLS is MLE under normal errors

$$y = X\beta + \varepsilon$$

if we assume that $\varepsilon \sim N(0, \sigma^2 I)$
then we get that

$$\hat{\beta}_{MLE} = (X^T X)^{-1} X^T y$$

and

$$\hat{\sigma}^2 = \frac{1}{n} \hat{\varepsilon}^T \hat{\varepsilon}$$

ML (back to linear regression with a single x)

$$y_i = \beta_0 + \beta_1 x_i + \varepsilon_i, \qquad \varepsilon_i \sim N(0, \sigma^2)$$

$$\implies y_i \sim N(\beta_0 + \beta_1 x_i, \sigma^2)$$

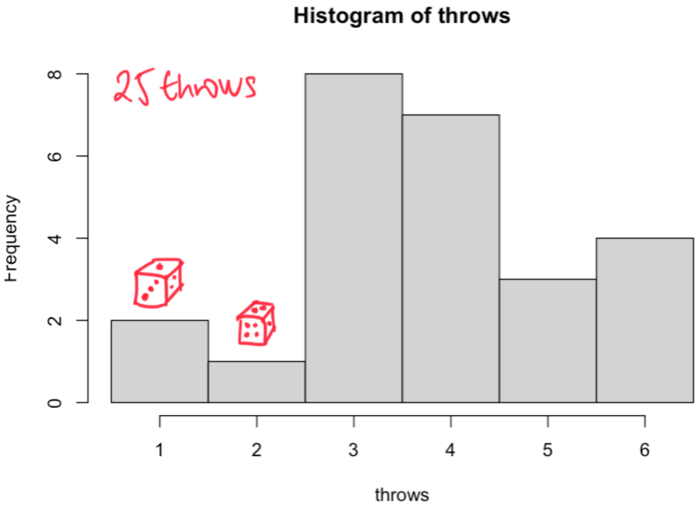
$$L(\beta_0, \beta_1, \sigma^2 | y, X) = \prod_{i=1}^{n} f(y_i | \beta_0, \beta_1, \sigma^2).$$

$$\log L(\beta_0, \beta_1, \sigma^2 | y, X) = \underbrace{-\frac{n}{2} \log 2\pi}_{\text{constant}} \quad \underbrace{-n \log \sigma^2}_{\text{does not depend on } \beta_0, \beta_1} \quad -\frac{1}{2\sigma^2} \sum_{i=1} (y_i - (\beta_0 + \beta_1 x_i))^2$$

- To maximize likelihood = to minimize sum of squares of residuals

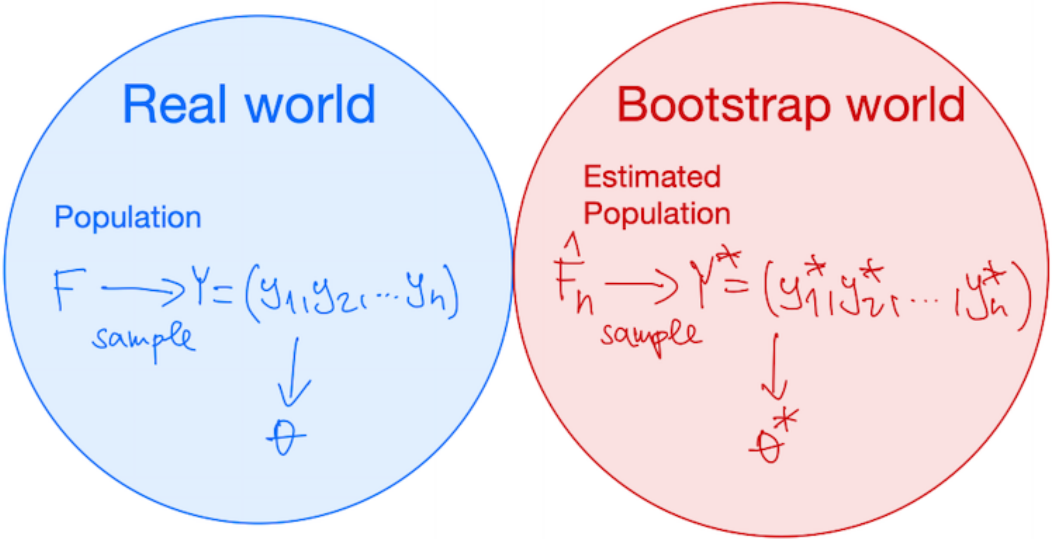
Bootstrap

Example - rolling a dice (again)

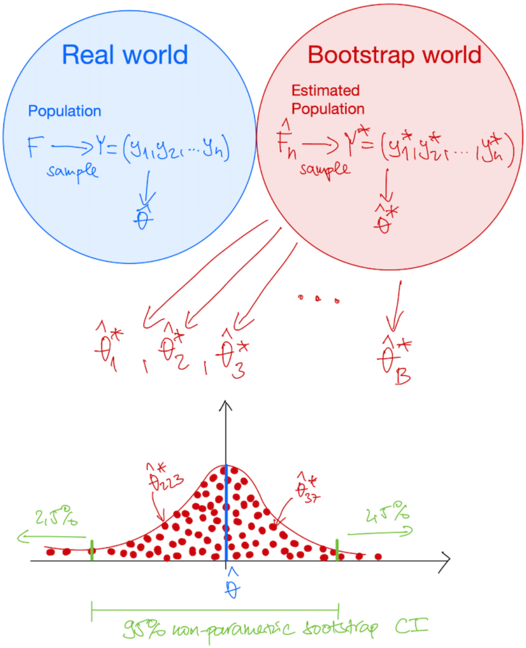


Data is all we have

- $\hat{F}_n \to F$
- we wish to understand sample variation, but we don't have F
- at least we have our data $\hat{F}_n$
- use our $\hat{F}_n$ to simulate new **"bootstrap"** datasets



Real world
Population
$F \xrightarrow{\text{sample}} Y = (y_1, y_2, \cdots y_n)$
θ
Bootstrap world
Estimated Population
$\hat{F}_n \xrightarrow{\text{sample}} Y^* = (y_1^*, y_2^*, \cdots, y_n^*)$
θ^*

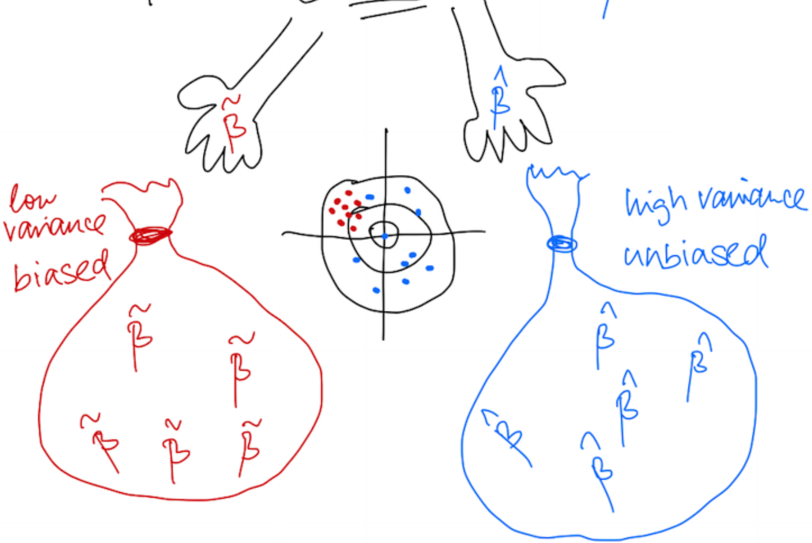


Real world
Population
Bootstrap world
Estimated
Population
sample
sample
95% non-parametric bootstrap CI

Bootstrap in understanding the sample variation

- Suppose we are considering choosing between two different estimators $\tilde{\beta}$ and $\hat{\beta}$
- These may possess different qualities
- The question is: Given that you have to pick only once, which one would you choose??

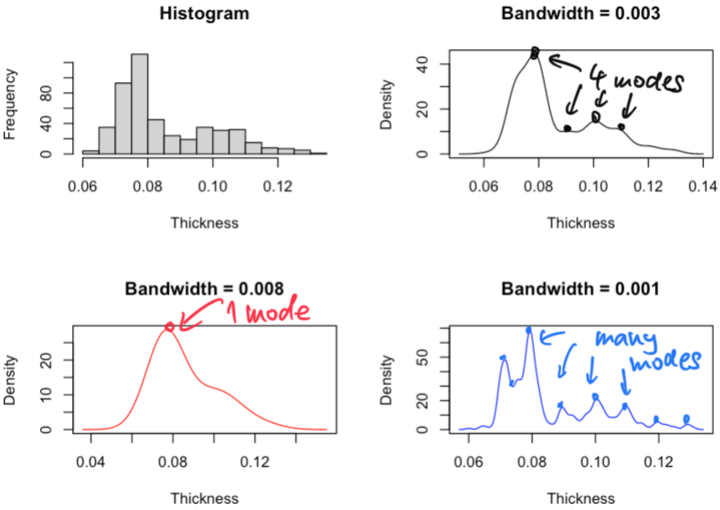
You pick only one $\tilde{\beta}$ or $\hat{\beta}$.



Assume we are in some of the following situations

- **small data sample** $\Longrightarrow$ Asymptotic approximations are unreliable (Ex: $n = 15$ in linear regression)
- our **estimator is complex** and we can't even derive asymptotic approximation (Ex: result of a numerical optimization)
- asymptotic distribution depends on the **unknown parameter** (Ex: $X_1, X_2, \ldots, X_n \sim f(.)$, sample median $\hat{m} \sim N\left(m, \frac{1}{4nf(m)^2}\right)$)
- traditional estimator is based on **dubious assumptions** (Ex: stock returns may have fat tails)

*Example - Stamp thickness



One mode at $\hat{h}_1 = 0.0068$.

H_0 : number of modes $= 1$

Natural candidate is $\hat{f}(t; \hat{h}_1)$.

We will ”improve” $\hat{f}(t; \hat{h}_1)$ so that it has the same variance as our data. The new one is $\hat{g}(\cdot; \hat{h}_1)$ (we applied *variance stabilizing transformation*).

$$ASL_{boot} = P_{\hat{g}(\cdot; \hat{h}_1)} \left(\hat{h}_1^* > \hat{h}_1 \right),$$

(*achieved significance level*) and $\hat{h}_1^*$ is the smallest smoothing parameter so that the distribution is unimodal.

We sample from a smooth distribution $\hat{g}(\cdot;\hat{h}_1)$, not $\hat{F}_n$, hence is it a *smooth bootstrap*.

We need to sample from $\hat{g}(\cdot;\hat{h}_1)$, which has the variance $\hat{\sigma}^2$ and expected value equal to the rv from $\hat{f}(t;\hat{h}_1)$.

How? We achieve this in the following way: we draw a bootstrap sample $y_1^*,\dots,y_n^*$ from $\hat{F}_n$ and set

$$x_i^* = \bar{y}^* + (1+\hat{h}_1^2/\hat{\sigma}^2)^{1/2}(y_i^* - \bar{y}^* + \hat{h}_1\varepsilon_i). \tag{1}$$

Example with Stamps: Algorithm

Step 1 Draw B bootstrap datasets z $\hat{g}(\cdot; \hat{h}_1)$ using (1)

Step 2 For each bootstrap dataset we calculate the smallest smoothing parameter so that the distribution is unimodal.. Denote these B values as $\hat{h}_1(1), ..., \hat{h}_1(B)$.

Krok 3 Aproximate ASL_{boot} using

$$A\hat{S}L_{boot} = \#\{\hat{h}_1^*(b) \geq \hat{h}_1\}/B.$$

For $B = 5000$ we got $A\hat{S}L_{boot} = 0.0002$, which is smaller than 5%, so we reject the null hypothesis that the stamps were printed on one type of paper at the significance level 5%.

Creative choice of the test statistic and null hypothesis improves the properties of the test, e.g. increase the chance of correctly rejecting the null hypothesis, if is untrue (improves power). This is why we chose the parameter of the smoothing parameter value on the edge between uni and bimodal as the null hypothesis.

Bootstrap - some remarks

- **very general approach** that makes few assumptions
- bootstrapped distribution can be used to construct **standard errors, confidence intervals, bias correction**

*Bootstrap may fail

- Paradox: we wish to use it situations that are complex, but in these, it may be also **difficult to prove** that it "works"
- It may fail if the parameter lies on the **boundary of the parameter space** (Ex: X $N(\mu, 1)$ where $\mu \in [0, \infty]$ - Andrews, 2000)
- If there is **missing support information**: Sample maximum: F_0 has support $[0, \theta_0]$. $\hat{\theta}_n = \max\{X_1, ..., X_n\}$. $\hat{T}_n = n(\hat{\theta}_n - \theta)$, $T_n^* = n(\hat{\theta}_n^* - \hat{\theta}_n)$. $P_n^*(T_n^* = 0) = 1 - (1 - 1/n)^n \to 1 - e^{-1}$ whereas $P(\hat{T}_n = 0) \to 0$.

*What if bootstrap fails?

Subsampling

- we draw smaller bootstrap samples without replacement
- intuition: we sample directly from the true distribution (F_0), not from the estimated one ($\hat{F}_n$)
- more general than bootstrap
- less efficient if the regular bootstrap works
- practical problem - how to choose subsample size?

*Bootstrap: Notation and theory

- $\{X_i, i = 1, \cdots, n\}$ data sample from unknown $F_0 \in \mathscr{I}$
- Sometime we assume some parametric family $F_0(x, \theta_0) = P(X \leq x)$
- Test statistic $\hat{T}_n = T_n(X_1, ..., X_n)$
- $G_n(\tau, F_0) = P(\hat{T}_n \leq \tau)$ denotes the true CDF of test statistic $\hat{T}_n$
- $\hat{T}_n$ je pivotal if $G_n(\tau, F)$ does not depend on F
- $\hat{T}_n$ is asymptotically pivotal if $G_\infty(\tau, F)$ does not depend on F
- how can we estimate $G_n(., F_0)$???
  - e.g. G_∞ - using asymptotic approximation (need large$n$)
  - replacing F_0 with some estimator - **bootstrap**
- let $\hat{F}_n$ denotes estimator of unknown F_0
  - ECDF (empirical cumulative distribution function) - $\hat{F}_n(x) = \frac{1}{n} \sum_{i=1}^{n} I(X_i \leq x)) \rightarrow_{a.s.} F_0(x)$
  - from a parametric family: $F_0(.) = F(., \theta_0)$

Procedure for approximation of $G_n(\tau, F_0)$

Step 1 We generate random sample of size n from $\hat{F}_n$: $\{X_i^* : i = 1, \ldots, n\}$

Step 2 Calculate $\hat{T}_n^* = T_n(X_1^*, \ldots, X_n^*)$

Step 3 Repeat (1) a (2) many times so that we get the empirical distribution of $(\hat{T}_n^* \leq \tau)$

By increasing the number of simulated bootstrap samples B we improve the estimator of $G_n(\tau, \hat{F}_n)$.

So by simulation we only get $\hat{G}_n(\tau, \hat{F}_n)$, if we have enough patience and computing time we can make this estimate arbitrarily good, so that $\hat{G}_n(\tau, \hat{F}_n) \to G_n(\tau, \hat{F}_n)$ for $B \to \infty$.

What does it mean that the bootstrap works?

It means that $G_n(., \hat{F}_n) \rightarrow G_n(., F_0)$

At least we would expect that the approximation would be correct if the sample size grows to infinity.

This property is called consistency

$G_n(t, \hat{F}_n)$ is consistent $\forall \varepsilon > 0, \forall F_0 \in \mathscr{I}$

$$\lim_{n \to \infty} P_n \left[\sup_{\tau} \left| G_n(t, \hat{F}_n) - G_\infty(\tau, F_0) \right| > \varepsilon \right] = 0$$

$$G_n(\tau, \hat{F}_n) \sim G_\infty(\tau, \hat{F}_n) \sim G_\infty(\tau, F_0) \sim G_n(\tau, F_0)$$

Beran and Ducharme (1991) presented sufficient conditions for consistency

- $\hat{F}_n \to F_0$ ($\hat{F}_n$ is a "good" estimator of F_0)
- $G_\infty(\tau, F)$ is continuous in τ for all $F \in \mathscr{I}$ (continuity in τ)
- for every τ and for every sequence H_n, such that $H_n \to F_0$: $G_n(\tau, H_n) \to G_\infty(\tau, F_0)$ ("continuity" in F_0)

$$G_n(\tau, \hat{F}_n) \sim G_\infty(\tau, \hat{F}_n) \sim G_\infty(\tau, F_0) \sim G_n(\tau, F_0)$$

Thank you for your attention!

References

- Very non-technical explanation of MLE in Economics: Lanot, Gauthier. ”Maximum likelihood and economic modeling.” IZA World of Labor 326 (2017).
- MLE is explained in Hansen’s Probability chapter 10 `https://www.ssc.wisc.edu/~bhansen/probability/`
- Appendix A in Faraway (2016) provides reasonable basics: Faraway, Julian J. Extending the linear model with R: generalized linear, mixed effects and nonparametric regression models. CRC press, 2016.
- A book length treatment of the Bootstrap by the inventors (47000 google scholar citations): Efron, Bradley, and Robert J. Tibshirani. An introduction to the bootstrap. CRC press, 1994.
- Bootstrap animations `https://www.stat.auckland.ac.nz/~wild/BootAnim/`
- A very short and succinct explanation of bootstrap and subsampling in a blog post by Larry Wasserman: `https://normaldeviate.wordpress.com/2013/01/19/bootstrapping-and-subsampling-part-i/` and `https://normaldeviate.wordpress.com/2013/01/27/bootstrapping-and-subsampling-part-ii/`
- *A rigourous theory on bootstrap is in chapter 23 in Van der Vaart, Aad W. Asymptotic statistics. Vol. 3. Cambridge university press, 2000.
- *Andrews, Donald WK. ”Inconsistency of the bootstrap when a parameter is on the boundary of the parameter space.” Econometrica (2000): 399-405.
- *Beran, Rudolf, and Gilles R. Ducharme. Asympotic theory for bootstrap methods in statistics. 1991.