Lectures 16: Statistics review and bias estimates

error propagation for nonlinear functions of fit parameters

with material from https://www.scss.tcd.ie/Rozenn.Dahyot/

statistics reviewed

The Empirical density function

Statistical inference concerns learning from experience: we observe a random sample $\mathbf{x} = (x_1, x_2, \cdots, x_n)$ and wish to infer properties of the complete population $\mathcal{X}$ that yielded the sample. A complete knowledge is obtained from the population density function $F(.)$ from which $\mathbf{x}$ has been generated $F \rightsquigarrow \mathbf{x} = (x_1, x_2, \cdots, x_n)$

Definition

The empirical density function $\hat{F}(.)$ is defined as:

$$\hat{F}(x) = \frac{1}{n} \sum_{i=1}^{n} \delta(x - x_i)$$

where $\delta(\cdot)$ is the Dirac delta function. So the probability of $x = x_j$ is :

$$\int \hat{F}(x_j)\, dx = \int \frac{1}{n} \sum_{i=1}^{n} \delta(x_j - x_i)\, dx = \begin{cases} \frac{1}{n}, & x_j \in \{x_1, \cdots, x_n\} \\ 0, & \text{otherwise} \end{cases}$$

statistics reviewed

Parameters

Definition

A parameter, θ, is a function of the probability density function (p.d.f.) F, e.g.:

$$\theta = t(F)$$

if θ is the mean

$$\theta = \mathbb{E}_F(x) = \int_{-\infty}^{+\infty} x \; F(x)dx = \mu_F$$

if θ is the variance

$$\theta = \mathbb{E}_F[(x - \mu_F)^2] = \int_{-\infty}^{+\infty} (x - \mu_F)^2 \; F(x)dx = \sigma_F^2$$

statistics reviewed

Statistics or estimates

Definition

A statistic (also called estimates, estimators) $\hat{\theta}$ is a function of $\hat{F}$ or the sample $\mathbf{x}$, e.g.:

$$\hat{\theta} = t(\hat{F})$$

or also written $\hat{\theta} = s(\mathbf{x})$.

if $\hat{\theta}$ is the mean:

$$\begin{aligned}
\hat{\theta} &= t(\hat{F}) &&= \int_{-\infty}^{+\infty} x \, \hat{F}(x) dx \\
& &&= \int_{-\infty}^{+\infty} x \, \frac{1}{n} \sum_{i=1}^{n} \delta(x - x_i) \, dx \\
& &&= \frac{1}{n} \sum_{i=1}^{n} x_i \\
& &&= s(\mathbf{x}) = \overline{x}
\end{aligned}$$

statistics reviewed

Statistics or estimates

if $\hat{\theta}$ is the variance:

$$\begin{aligned}\hat{\theta} &= \int_{-\infty}^{+\infty} (x - \overline{x})^2 \, \hat{F}(x) dx \\ &= \frac{1}{n} \sum_{i=1}^{n} (x_i - \overline{x})^2 \\ &= \hat{\sigma}^2\end{aligned}$$

The Plug-in principle

Definition

The Plug-in estimate of a parameter $\theta = t(F)$ is defined to be:

$$\hat{\theta} = t(\hat{F}).$$

The function $\theta = t(F)$ of the probability density function F is estimated by the same function $t(.)$ of the empirical density $\hat{F}$.

- $\bar{x}$ is the plug-in estimate of μ_F.
- $\hat{\sigma}$ is the plug-in estimate of σ_F

Computing the mean knowing F

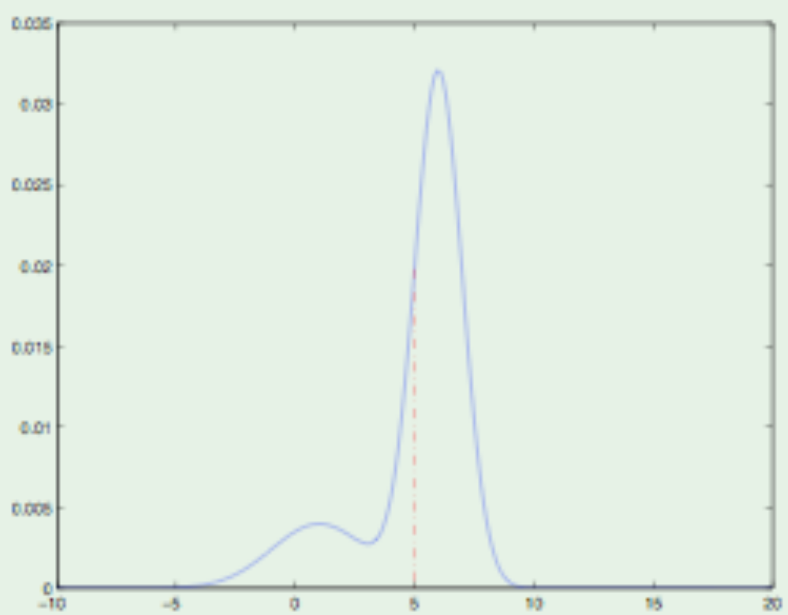
Example A

Lets assume we know the p.d.f. F:

$$F(x) = 0.2\,\mathcal{N}(\mu{=}1,\sigma{=}2) + 0.8\,\mathcal{N}(\mu{=}6,\sigma{=}1)$$

Then the mean is computed:

$$\begin{aligned} \mu_F = \mathbb{E}_F(x) &= \int_{-\infty}^{+\infty} x\, F(x)\; dx \\ &= 0.2 \cdot 1 + 0.8 \cdot 6 \\ &= 5 \end{aligned}$$



statistics reviewed

Estimating the mean knowing the observations **x**

Example A

Observations $\mathbf{x} = (x_1, \cdots, x_{100})$:

7.0411	4.8397	5.3156	6.7719	7.0616
5.2546	7.3937	4.3376	4.4010	5.1724
7.4199	5.3677	6.7028	6.2003	7.5707
4.1230	3.8914	5.2323	5.5942	7.1479
3.6790	0.3509	1.4197	1.7585	2.4476
−3.8635	2.5731	−0.7367	0.5627	1.6379
−0.1864	2.7004	2.1487	2.3513	1.4833
−1.0138	4.9794	0.1518	2.8683	1.6269
6.9523	5.3073	4.7191	5.4374	4.6108
6.5975	6.3495	7.2762	5.9453	4.6993
6.1559	5.8950	5.7591	5.2173	4.9980
4.5010	4.7860	5.4382	4.8893	7.2940
5.5741	5.5139	5.8869	7.2756	5.8449
6.6439	4.5224	5.5028	4.5672	5.8718
6.0919	7.1912	6.4181	7.2248	8.4153
7.3199	5.1305	6.8719	5.2686	5.8055
5.3602	6.4120	6.0721	5.2740	7.2329
7.0912	7.0766	5.9750	6.6091	7.2135
4.9585	5.9042	5.9273	6.5762	5.3702
4.7654	6.4668	6.1983	4.3450	5.3261

From the samples, the mean can be computed:

$$\bar{x} = \frac{\sum_{i=1}^{100} x_i}{100}$$

$$= 4.9970$$

statistics reviewed

Accuracy of estimates $\hat{\theta}$

We can compute an estimate $\hat{\theta}$ of a parameter θ from an observation sample $\mathbf{x} = (x_1, x_2, \cdots, x_n)$. But how accurate is $\hat{\theta}$ compared to the real value θ ?

Our attention is focused on questions concerning the probability distribution of $\hat{\theta}$. For instance we would like to know about:

- its standard error
- its confidence interval
- its bias
- etc.

Standard error of $\hat{\theta}$

Definition

The standard error is the standard deviation of a statistic $\hat{\theta}$. As such, it measures the precision of an estimate of the statistic of a population distribution.

$$se(\hat{\theta}) = \sqrt{var_F[\hat{\theta}]}$$

Standard error of $\bar{x}$

We have:

$$\mathbb{E}_F\left[(\bar{x} - \mu_F)^2\right] = \frac{\sum_{i=1}^n \mathbb{E}_F\left[(x_i - \mu_F)^2\right]}{n^2} = \frac{\sigma_F^2}{n}$$

Then

$$se_F(\bar{x}) = [\mathrm{var}_F(\bar{x})]^{1/2} = \frac{\sigma_F}{\sqrt{n}}$$

Plug in estimate of the standard error

Suppose now that F is unknown and that only the random sample $\mathbf{x} = (x_1, \cdots, x_n)$ is known. As μ_F and σ_F are unknown, we can use the previous formula to compute a plug-in estimate of the standard error.

Definition

The estimated standard error of the estimator $\hat{\theta}$ is defined as:

$$\widehat{\mathrm{se}}(\hat{\theta}) = \mathrm{se}_{\hat{F}}(\hat{\theta}) = [\mathrm{var}_{\hat{F}}(\hat{\theta})]^{1/2}$$

Estimated standard error of $\overline{x}$

$$\widehat{\mathrm{se}}(\overline{x}) = \frac{\hat{\sigma}}{\sqrt{n}}$$

Example on the mouse data

Data (Treatment group)	94; 197; 16; 38; 99; 141; 23
Data (Control group)	52; 104; 146; 10; 51; 30; 40; 27; 46

Table: The mouse data [Efron]. 16 mice divided assigned to a treatment group (7) or a control group (9). Survival in days following a test surgery. Did the treatment prolong survival ?

Example on the mouse data

Mean and Standard error for both groups

	$\bar{x}$	$\hat{se}$
Treatment	86.86	25.24
Control	56.22	14.14

Conclusion at first glance

It seems that mice having the treatment survive $d = 86.86 - 56.22 = 30.63$ days more than the mice from the control group.

Example on the mouse data

Stantard error of the difference $\bar{x}_{Treat} - \bar{x}_{Cont}$

$\bar{x}_{Treat}$ and $\bar{x}_{Cont}$ are independent, so the standard error of their difference is $\hat{\text{se}}(d) = \sqrt{\hat{\text{se}}^2_{Treat} + \hat{\text{se}}^2_{Cont}} = 28.93$. We see that:

$$\frac{d}{\hat{\text{se}}(d)} = \frac{30.63}{28.93} = 1.05$$

This shows that this is an insignificant result as it could easily have arised by chance (i.e. if the test was reproduced, it is *likely possible* to measure datasets giving $d = 0$!).
Therefore, we can not conclude with certainty that the treatment improves the survival of the mice.

Confidence interval for $\hat{\theta}$

Definition

Assuming that the estimator $\hat{\theta}$ is normally distributed with unknown expectation θ and variance se^2, then :

$$\mathrm{Prob}\{\hat{\theta} - z^{(1-\alpha)}\mathrm{se} \leq \theta \leq \hat{\theta} - z^{(\alpha)}\mathrm{se}\} = 1 - 2\alpha$$

Therefore $1 - 2\alpha$ % confidence interval for θ is $[\hat{\theta} - z^{(1-\alpha)}\mathrm{se}; \hat{\theta} - z^{(\alpha)}\mathrm{se}]$
Confidence limits are the lower and upper boundaries values of a confidence interval. The confidence level is the probability value $100 \times (1 - 2\alpha)$ % associated with a confidence interval.

Confidence interval

The width of the confidence interval gives us some idea about how uncertain we are about the unknown parameter. A very wide interval may indicate that more data should be collected before anything very definite can be said about the parameter.

percentile $\alpha \times 100$ %	confidence level $(1-2\alpha) \times 100$ %	$z^{(1-\alpha)}$
10	80	1.28155
5	90	1.64485
2.5	95	1.95996
0.5	99	2.57583
0.25	99.5	2.80703
0.05	99.9	3.29053

Table: For a normal p.d.f $z^{(\alpha)} = -z^{(1-\alpha)}$

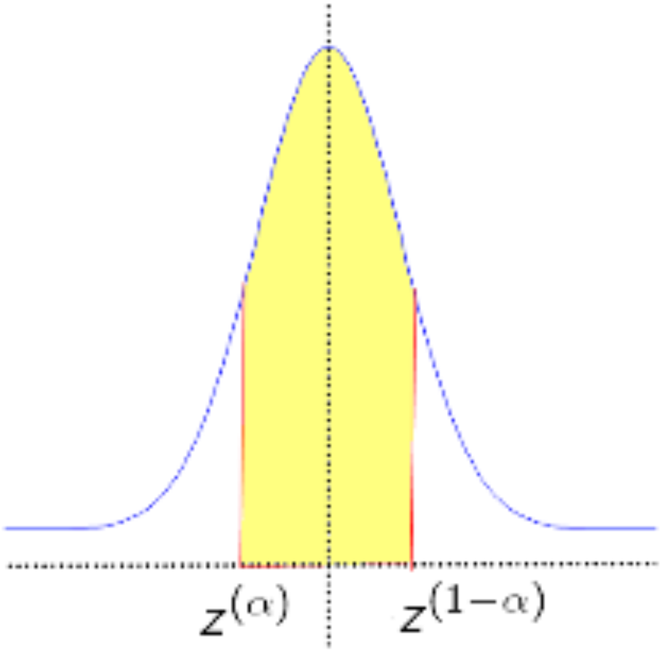


Figure: Density function $\mathcal{N}(0, 1)$.

Example Confidence interval

Confidence interval of the mean

Using the central limit theorem, the estimate $\overline{x}$ is following a normal density function $\mathcal{N}\left(\mu_F, \frac{\sigma_F^2}{n}\right)$. The 90% confidence interval is :

$$\overline{x} \pm 1.645 \frac{\sigma_F}{\sqrt{n}} \text{ estimated by } \pm 1.645 \frac{\hat{\sigma}}{\sqrt{n}}$$

confidence interval of the difference for the mouse data

The difference d in days of survival between the treatment group and the control group has a estimated 90% confidence interval defined as:

$$d = 30.63 \pm 1.645 \times 28.93 = 30.63 \pm 47.5898$$

Bias of $\hat{\theta}$

Definition

The Bias is the difference between the expectation of an estimator $\hat{\theta}$ and the quantity θ being estimated:

$$\mathrm{Bias}_F(\hat{\theta}, \theta) = \mathbb{E}_F(\hat{\theta}) - \theta$$

Bias of the mean $\overline{x}$

we have:

$$\mathbb{E}_F(\overline{x}) = \mathbb{E}_F\left(\frac{\sum_{i=1}^n x_i}{n}\right) = \frac{\sum_{i=1}^n \mathbb{E}_F(x_i)}{n} = \mu_F$$

then:

$$\mathrm{Bias}_F(\overline{x}, \mu_F) = \mathbb{E}_F(\overline{x}) - \mu_F = 0$$

Bias of $\hat{\theta}$

Bias of $\hat{\sigma}^2$

$$\begin{aligned}\hat{\sigma}^2 &= \tfrac{1}{n}\textstyle\sum_{i=1}^n (x_i - \bar{x})^2 = \tfrac{1}{n}\sum_{i=1}^n ((x_i - \mu_F) + (\mu_F - \bar{x}))^2 \\ &= \left(\tfrac{1}{n}\textstyle\sum_{i=1}^n (x_i - \mu_F)^2\right) - (\bar{x} - \mu_F)^2\end{aligned}$$

The first term has an expected value of σ_F^2 and the second term has expected value σ_F^2/n. So the bias of $\hat{\sigma}^2$ is:

$$\mathrm{Bias}_F(\hat{\sigma}^2, \sigma_F^2) = \sigma_F^2 - \frac{\sigma_F^2}{n} - \sigma_F^2 = -\frac{\sigma_F^2}{n}$$

Bias of $\hat{\theta}$

Instead of using $\hat{\sigma}^2$ as an estimate of the variance, you should try to choose an unbiased estimate.

Bias of $\overline{\sigma}^2$

Let define:

$$\overline{\sigma}^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \overline{x})^2$$

then by computing its bias:

$$\begin{aligned} \mathrm{Bias}_F(\overline{\sigma}^2, \sigma_F^2) &= \mathbb{E}_F(\overline{\sigma}^2) - \sigma_F^2 \\ &= 0 \end{aligned}$$

$\overline{\sigma}$ is an unbiased estimator of the standard deviation.

bootstrap review and bias

bootstrap review and bias

Bootstrap samples and replications

Definition

A bootstrap sample $\mathbf{x}^* = (x_1^*, x_2^*, \cdots, x_n^*)$ is obtained by randomly sampling n times, with replacement, from the original data points $\mathbf{x} = (x_1, x_2, \cdots, x_n)$.

Considering a sample $\mathbf{x} = (x_1, x_2, x_3, x_4, x_5)$, some bootstrap samples can be:

$$\mathbf{x}^{*(1)} = (x_2, x_3, x_5, x_4, x_5)$$
$$\mathbf{x}^{*(2)} = (x_1, x_3, x_1, x_4, x_5)$$

etc.

Definition

With each bootstrap sample $\mathbf{x}^{*(1)}$ to $\mathbf{x}^{*(B)}$, we can compute a bootstrap replication $\hat{\theta}^*(b) = s(\mathbf{x}^{*(b)})$ using the plug-in principle.

bootstrap review and bias

How to compute Bootstrap samples

Repeat B times:

1. A random number device selects integers $i_1, \cdots, i_n$ each of which equals any value between 1 and n with probability $\frac{1}{n}$.
2. Then compute $\mathbf{x}^* = (x_{i_1}, \cdots, x_{i_n})$.

Some matlab code available on the web

See BOOTSTRAP MATLAB TOOLBOX, by Abdelhak M. Zoubir and D. Robert Iskander, http://www.csp.curtin.edu.au/downloads/bootstrap_toolbox.html

bootstrap review and bias

How many values are left out of a bootstrap resample ?

Given a sample $\mathbf{x} = (x_1, x_2, \cdots, x_n)$ and assuming that all x_i are different, the probability that a particular value x_i is left out of a resample $\mathbf{x}^* = (x_1^*, x_2^*, \cdots, x_n^*)$ is:

$$\mathcal{P}(x_j^* \neq x_i, 1 \leq j \leq n) = \left(1 - \frac{1}{n}\right)^n$$

since $\mathcal{P}(x_j^* = x_i) = \frac{1}{n}$. When n is large, the probability $(1 - \frac{1}{n})^n$ converges to $e^{-1} \approx 0.37$.

The Bootstrap algorithm for Estimating standard errors

1. Select B independent bootstrap samples $\mathbf{x}^{*(1)}, \mathbf{x}^{*(2)}, \cdots, \mathbf{x}^{*(B)}$ drawn from $\mathbf{x}$
2. Evaluate the bootstrap replications:

$$\hat{\theta}^*(b) = s(\mathbf{x}^{*(b)}), \quad \forall b \in \{1, \cdots, B\}$$

3. Estimate the standard error $\mathrm{se}_F(\hat{\theta})$ by the standard deviation of the B replications:

$$\hat{\mathrm{se}}_B = \left[\frac{\sum_{b=1}^{B}[\hat{\theta}^*(b) - \hat{\theta}^*(\cdot)]^2}{B-1}\right]^{\frac{1}{2}}$$

where $\hat{\theta}^*(\cdot) = \frac{\sum_{b=1}^{B} \hat{\theta}^*(b)}{B}$

Bootstrap estimate of the standard Error

Example A

From the distribution F: $F(x) = 0.2\,\mathcal{N}(\mu{=}1,\sigma{=}2) + 0.8\,\mathcal{N}(\mu{=}6,\sigma{=}1)$. We draw the sample $\mathbf{x} = (x_1, \cdots, x_{100})$:

$\mathbf{x} =$

7.0411	4.8397	5.3156	6.7719	7.0616
5.2546	7.3937	4.3376	4.4010	5.1724
7.4199	5.3677	6.7028	6.2003	7.5707
4.1230	3.8914	5.2323	5.5942	7.1479
3.6790	0.3509	1.4197	1.7585	2.4476
−3.8635	2.5731	−0.7367	0.5627	1.6379
−0.1864	2.7004	2.1487	2.3513	1.4833
−1.0138	4.9794	0.1518	2.8683	1.6269
6.9523	5.3073	4.7191	5.4374	4.6108
6.5975	6.3495	7.2762	5.9453	4.6993
6.1559	5.8950	5.7591	5.2173	4.9980
4.5010	4.7860	5.4382	4.8893	7.2940
5.5741	5.5139	5.8869	7.2756	5.8449
6.6439	4.5224	5.5028	4.5672	5.8718
6.0919	7.1912	6.4181	7.2248	8.4153
7.3199	5.1305	6.8719	5.2686	5.8055
5.3602	6.4120	6.0721	5.2740	7.2329
7.0912	7.0766	5.9750	6.6091	7.2135
4.9585	5.9042	5.9273	6.5762	5.3702
4.7654	6.4668	6.1983	4.3450	5.3261

We have $\mu_F = 5$ and $\overline{x} = 4.9970$.

bootstrap review and bias

Bootstrap estimate of the standard Error

Example A

1. $B = 1000$ bootstrap samples $\{\mathbf{x}^{*(b)}\}$
2. $B = 1000$ replications $\{\overline{x}^*(b)\}$
3. Bootstrap estimate of the standard error:

$$\widehat{\text{se}}_{B=1000} = \left[\frac{\sum_{b=1}^{1000}[\overline{x}^*(b) - \overline{x}^*(\cdot)]^2}{1000 - 1}\right]^{\frac{1}{2}} = 0.2212$$

where $\overline{x}^*(\cdot) = 5.0007$. This is to compare with $\hat{se}(\overline{x}) = \frac{\hat{\sigma}}{\sqrt{n}} = 0.22$.

Distribution of $\hat{\theta}$

When enough bootstrap resamples have been generated, not only the standard error but any aspect of the distribution of the estimator $\hat{\theta} = t(\hat{F})$ could be estimated. One can draw a histogram of the distribution of $\hat{\theta}$ by using the observed $\hat{\theta}^*(b)$, $b = 1, \cdots, B$.

Example A

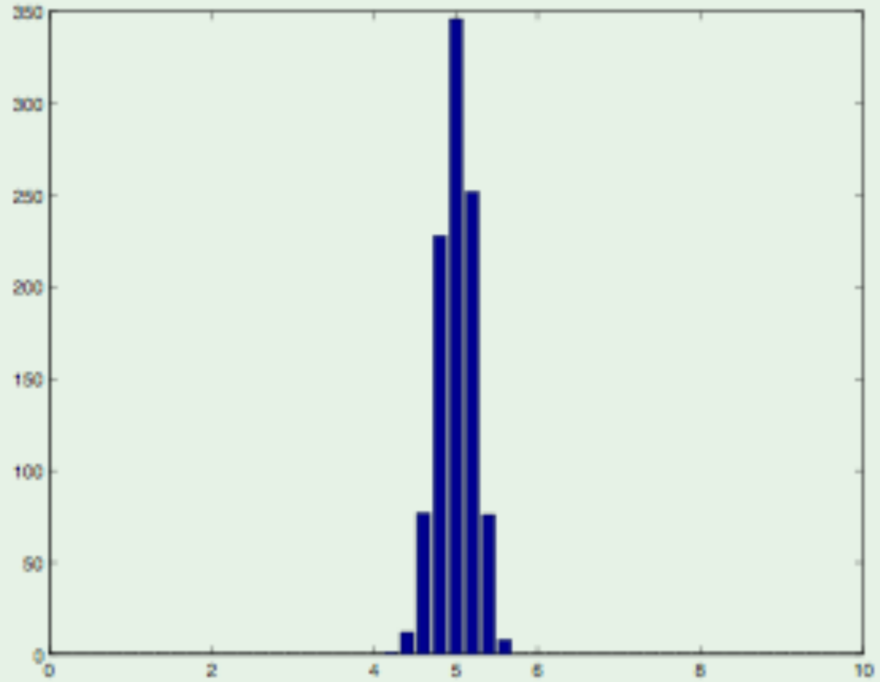


Figure: Histogram of the replications $\{\overline{x}^*(b)\}_{b=1\cdots B}$.

bootstrap review and bias

Bootstrap estimate of the standard error

Definition

The ideal bootstrap estimate $\mathrm{se}_{\hat{F}}(\theta^*)$ is defined as:

$$\lim_{B\to\infty} \hat{\mathrm{se}}_B = \mathrm{se}_{\hat{F}}(\theta^*)$$

$\mathrm{se}_{\hat{F}}(\theta^*)$ is called a non-parametric bootstrap estimate of the standard error.

Bootstrap estimate of the standard Error

How many B in practice ?

you may want to limit the computation time. In practice, you get a good estimation of the standard error for B in between 50 and 200.

Example A

B	10	20	50	100	500	1000	10000
$\widehat{se}_B$	0.1386	0.2188	0.2245	0.2142	0.2248	0.2212	0.2187

Table: Bootstrap standard error w.r.t. the number B of bootstrap samples.

bootstrap review and bias

Bootstrap estimate of bias

Definition

The bootstrap estimate of bias is defined to be the estimate:

$$\widehat{\text{Bias}}_{\hat{F}}(\hat{\theta}) = \mathbb{E}_{\hat{F}}[s(\mathbf{x}^*)] - t(\hat{F})$$

$$= \theta^*(\cdot) - \hat{\theta}$$

Example A

B	10	20	50	100	500	1000	10000
$\mathbb{E}_{\hat{F}}(\overline{x}^*)$	5.0587	4.9551	5.0244	4.9883	4.9945	5.0035	4.9996
$\widehat{\text{Bias}}$	0.0617	-0.0419	0.0274	-0.0087	-0.0025	0.0064	0.0025

Table: $\widehat{\text{Bias}}$ of $\overline{x}^*$ ($\overline{x} = 4.997$ and $\mu_F = 5$).

bootstrap review and bias

Bootstrap estimate of bias

1. B independent bootstrap samples $\mathbf{x}^{*(1)}, \mathbf{x}^{*(2)}, \cdots, \mathbf{x}^{*(B)}$ drawn from $\mathbf{x}$
2. Evaluate the bootstrap replications:

$$\hat{\theta}^*(b) = s(\mathbf{x}^{*(b)}), \quad \forall b \in \{1, \cdots, B\}$$

3. Approximate the bootstrap expectation :

$$\hat{\theta}^*(\cdot) = \frac{1}{B}\sum_{b=1}^{B} \hat{\theta}^*(b) = \frac{1}{B}\sum_{b=1}^{B} s(\mathbf{x}^{*(b)})$$

4. the bootstrap estimate of bias based on B replications is:

$$\widehat{\text{Bias}}_B = \hat{\theta}^*(\cdot) - \hat{\theta}$$

Confidence interval

Definition

Using the bootstrap estimation of the standard error, the $100(1-2\alpha)\%$ confidence interval is:

$$\theta = \hat{\theta} \pm z^{(1-\alpha)} \cdot \widehat{\mathrm{se}}_B$$

Definition

If the bias in not null, the bias corrected confidence interval is defined by:

$$\theta = (\hat{\theta} - \widehat{\mathrm{Bias}}_B) \pm z^{(1-\alpha)} \cdot \widehat{\mathrm{se}}_B$$

bootstrap review and bias

Can the bootstrap answer other questions?

The mouse data

Data (Treatment group)	94; 197; 16; 38; 99; 141; 23
Data (Control group)	52; 104; 146; 10; 51; 30; 40; 27; 46

Table: The mouse data [Efron]. 16 mice divided assigned to a treatment group (7) or a control group (9). Survival in days following a test surgery. Did the treatment prolong survival ?

Can the bootstrap answer other questions?

The mouse data

- Remember in the first lecture, we compute $d = \bar{x}_{Treat} - \bar{x}_{Cont} = 30.63$ with a standard error $\hat{se}(d) = 28.93$. The ratio was $d/\hat{se}(d) = 1.05$ (an insignificant result as measuring $d = 0$ is likely possible).
- Using bootstrap method
  1. B bootstrap samples $\mathbf{x}^{*(b)}_{Treat} = (x^{*(b)}_{Treat\ 1}, \cdots, x^{*(b)}_{Treat\ 7})$ and $\mathbf{x}^{*(b)}_{Cont} = (x^{*(b)}_{Cont\ 1}, \cdots, x^{*(b)}_{Cont\ 9})$, $\forall 1 \leq b \leq B$
  2. B bootstrap replications are computed: $d^*(b) = \bar{x}^{*(b)}_{Treat} - \bar{x}^{*(b)}_{Cont}$
  3. The bootstrap standard error is computed for $B = 1400$: $\hat{se}_{B=1400} = 26.85$.
  4. The ratio is $d/\hat{se}_{1400}(d) = 1.14$.
- This is still not a significant result.

The Law school example

School	1	2	3	4	5	6	7	8
LSAT (X)	576	635	558	578	666	580	555	661
GPA (Y)	3.39	3.30	2.81	3.03	3.44	3.07	3.00	3.43

School	9	10	11	12	13	14	15
LSAT (X)	651	605	653	575	545	572	594
GPA (Y)	3.36	3.13	3.12	2.74	2.76	2.88	2.96

Table: Results of law schools admission practice for the LSAT and GPA tests. It is believed that these scores are highly correlated. Compute the correlation and its standard error.

bootstrap review and bias

Correlation

The correlation is defined :

$$\text{corr}(X, Y) = \frac{\mathbb{E}[(X - \mathbb{E}(X)) \cdot (Y - \mathbb{E}(Y))]}{(\mathbb{E}[(X - \mathbb{E}(X))^2] \cdot \mathbb{E}[(Y - \mathbb{E}(Y))^2])^{1/2}}$$

Its typical estimator is:

$$\widehat{\text{corr}}(\mathbf{x}, \mathbf{y}) = \frac{\sum_{i=1}^{n} x_i \; y_i - n \; \bar{x} \; \bar{y}}{[\sum_{i=1}^{n} x_i^2 - n\bar{x}^2]^{1/2} \cdot [\sum_{i=1}^{n} y_i^2 - n\bar{y}^2]^{1/2}}$$

bootstrap review and bias

The Law school example

- The estimated correlation is $\widehat{\text{corr}}(\mathbf{x}, \mathbf{y}) = .7764$ between LSAT and GPA.
- Precise theoretical formula for the standard error of the estimator is unavailable.

Non-parametric Bootstrap estimate of the standard error

B	25	50	100	200	400	800	1600	3200
$\widehat{\text{se}}_B$	.140	.142	.151	.143	.141	.137	.133	.132

Table: Bootstrap estimate of standard error for $\widehat{\text{corr}}(\mathbf{x}, \mathbf{y}) = .776$.

The standard error stabilizes to $\text{se}_{\hat{F}}(\widehat{\text{corr}}) \approx .132$.