Algorithms and Inference

Aldo Solari

Statistical Learning

Algorithms and inference

- Suppose we have observed $y_1, \ldots, y_n$, realizations of $Y_1, \ldots, Y_n$ i.i.d. Y, and our interest is on $\mathbb{E}(Y) = \mu$
- Averaging is the *algorithm*

$$\bar{y} = \frac{1}{n} \sum_{i=1}^{n} y_i$$

- The standard error provides an *inference* on the algorithm's accuracy

$$\widehat{\mathrm{se}} = \sqrt{\frac{1}{n} \frac{\sum_{i=1}^{n} (y_i - \bar{y})^2}{n-1}}$$

- "It is a surprising, and crucial, aspect of statistical theory that the same data that supplies an estimate can also assess its accuracy" (Efron and Hastie, 2016)

Gaussian model

Model

$Y_1, \ldots, Y_n$ i.i.d. $Y \sim N(\mu, \sigma^2)$

μ is the *parameter of interest*
σ^2 is the *nuisance parameter*

Estimator and its standard error

$\bar{Y} \sim N(\mu, \sigma^2/n)$
$\mathrm{se}(\bar{Y}) = \sqrt{\mathrm{Var}(\bar{Y})} = \sigma\sqrt{1/n}$

Estimator of the standard error

$\hat{\mathrm{se}}(\bar{Y}) = \hat{\sigma}\sqrt{1/n}$

$$\hat{\sigma}^2 = \frac{\sum_{i=1}^{n}(Y_i - \bar{Y})^2}{n-1} \sim \sigma^2\chi^2_{n-1}/(n-1)$$

Confidence interval

Pivotal statistic

$$T = \frac{\bar{Y} - \mu}{\sigma\sqrt{1/n}} \cdot \frac{\sigma}{\hat{\sigma}} \sim \frac{N(0,1)}{\sqrt{\chi^2_{n-1}/(n-1)}} \sim t_{n-1}$$

with $\Pr(-t_{n-1}^{1-\alpha/2} \leq T \leq t_{n-1}^{1-\alpha/2}) = 1 - \alpha$

where $t_{n-1}^{1-\alpha/2}$ is the $1 - \alpha/2$ quantile of the Student t distribution with $n - 1$ degrees of freedom

$1 - \alpha$ **confidence interval for** μ

$$\boxed{[\underline{\mu}, \overline{\mu}] = \bar{Y} \pm t_{n-1}^{1-\alpha/2} \cdot \hat{\mathrm{se}}(\bar{Y})}$$

Coverage

$\Pr([\underline{\mu}, \overline{\mu}] \ni \mu) = 1 - \alpha$

Simulation

```
sim = function(n=25, mu=0, sigma=1, alpha=0.05){
  ys = rnorm(n, mean=mu, sd=sigma)
  bary = mean(ys)
  hatse = sqrt( var(ys) / n )
  k = qt(alpha/2, df = n-1, lower.tail = F)
  ci = bary + c(-1,1) * k * hatse
  cover = (mu >= ci[1] & mu <= ci[2])
  return(cover)
}

set.seed(123)
B = 1000
mean( replicate(B, sim(n=25) ))
```

Outline

1. **Leukemia data**

2. **Kidney data**

Leukemia data

- $n = 72$ leukemia patients: 45 with ALL (acute lymphoblastic leukemia) and 27 with AML (acute myeloid leukemia, a worse prognosis)
- Each patient has genetic activity measured for $p = 7128$ genes
- The histograms in the next slide compare the genetic activities in the two groups for gene 136
- Is the perceived difference genuine, or perhaps, as people like to say, "a statistical fluke"?

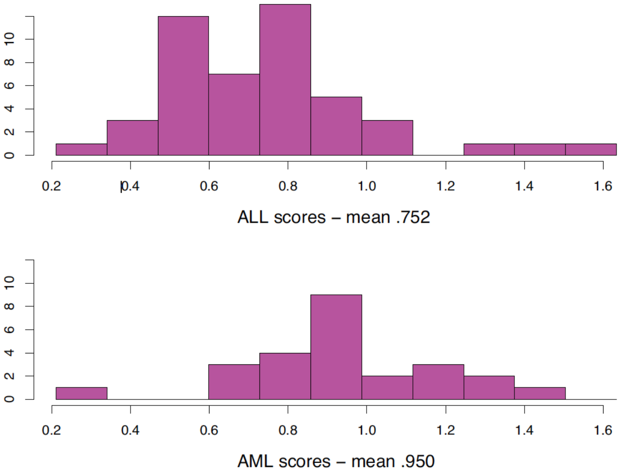


Figure 1.4 Scores for gene 136, leukemia data. Top **ALL** ($n = 47$), bottom **AML** ($n = 25$). A two-sample t-statistic $= 3.01$ with p-value $= .0036$.

Source: Efron and Hastie (2016)

Hypothesis testing

- The classic answer to this question is via a two-sample t-statistic

$$t = \frac{\overline{\mathrm{AML}} - \overline{\mathrm{ALL}}}{\hat{\mathrm{sd}}}$$

- Compare the observed value $t = 3.01$ with the null distribution, i.e. Student's t distribution with 70 degrees of freedom
- The p-value is 0.0036. A small p-value is a statement of statistical surprise: something very unusual has happened if in fact there is no difference in gene 136 expression between ALL and AML patients
- We are less surprised by $t = 3.01$ or $p = 0.0036$ if gene 136 is just one out of thousands candidates that might have produced "interesting" results
- The next slide shows the histogram of the two-sample t-statistics for the 7178 genes. Now $t = 3.01$ looks less unusual; 400 other genes have t exceeding 3.01, about 5.6% of them

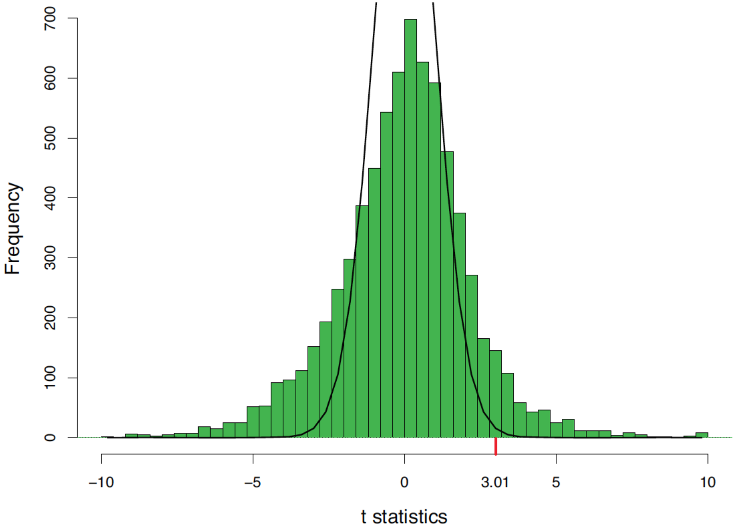


Figure 1.5 Two-sample t-statistics for 7128 genes, leukemia data. The smooth curve is the theoretical null density for the t-statistic.

Outline

1. Leukemia data

2. Kidney data

Kidney data

- Kidney function generally declines with age. The rate of decline is an important question in kidney transplantation: in the past, potential donors past age 60 were prohibited
- $Y = \texttt{tot}$ (kidney function overall score)
- $X = \texttt{age}$ (age in years)
- $(x_1, y_1), \ldots, (x_n, y_n)$ for $n = 157$ healthy volunteers

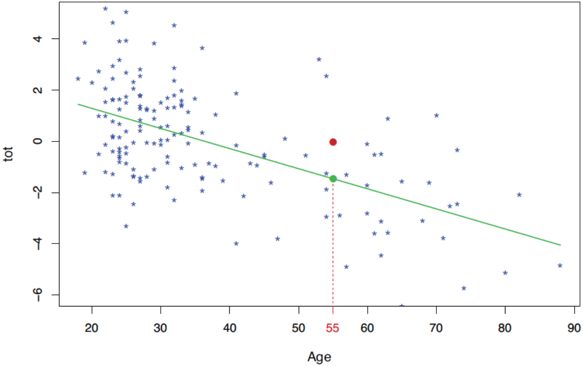


Figure 8.1 Kidney data; a new volunteer donor is aged 55. Which prediction is preferred for his kidney function?

Source: Efron and Hastie (2016)

Linear model

Model

$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}, \quad \boldsymbol{\epsilon} \sim \mathcal{N}(\mathbf{0}, \sigma^2 \mathbf{I}_n)$

$\mathbf{x}^\top = (x_1, \ldots, x_p)$: values of interest

$\mu_x = \mathbf{x}^\top \boldsymbol{\beta}$: parameter of interest

Estimator and its standard error

$$\hat{\mu}_x = \mathbf{x}^\top \hat{\boldsymbol{\beta}} = \mathbf{x}^\top (\mathbf{X}^\top \mathbf{X})^{-1} \mathbf{X}^\top \mathbf{y} \sim N(\mathbf{x}^\top \boldsymbol{\beta}, \sigma^2 \mathbf{x}^\top (\mathbf{X}^\top \mathbf{X})^{-1} \mathbf{x})$$

$$\mathrm{se}(\hat{\mu}_x) = \sigma \sqrt{\mathbf{x}^\top (\mathbf{X}^\top \mathbf{X})^{-1} \mathbf{x}}$$

Estimator of the standard error

$$\hat{\mathrm{se}}(\hat{\mu}_x) = \sigma \sqrt{\mathbf{x}^\top (\mathbf{X}^\top \mathbf{X})^{-1} \mathbf{x}}$$

$$\hat{\sigma}^2 = (\mathbf{y} - \mathbf{X}\hat{\boldsymbol{\beta}})^\top (\mathbf{y} - \mathbf{X}\hat{\boldsymbol{\beta}})/(n-p) \sim \sigma^2 \chi^2_{n-p}/(n-p)$$

Confidence interval

Pivotal statistic

$$T = \frac{\hat{\mu}_x - \mu_x}{\widehat{\mathrm{se}}(\hat{\mu}_x)} \cdot \frac{\sigma}{\hat{\sigma}} \sim \frac{\mathcal{N}(0,1)}{\sqrt{\chi^2_{n-p}/(n-p)}} \sim t_{n-p}$$

with $\Pr(-t_{n-p}^{1-\alpha/2} \leq T \leq t_{n-p}^{1-\alpha/2}) = 1 - \alpha$

$1 - \alpha$ **confidence interval for** μ_x

$$\boxed{[\underline{\mu}_x, \overline{\mu}_x] = \hat{\mu}_x \pm t_{n-p}^{1-\alpha/2} \cdot \widehat{\mathrm{se}}(\hat{\mu}_x)}$$

Coverage

$\Pr([\underline{\mu}_x, \overline{\mu}_x] \ni \mu_x) = 1 - \alpha$

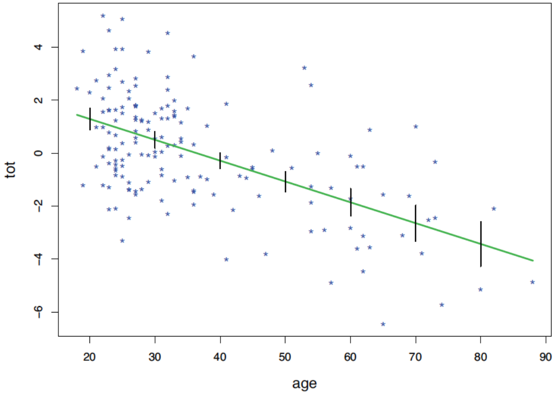


Figure 1.1 Kidney fitness `tot` vs `age` for 157 volunteers. The line is a linear regression fit, showing ±2 standard errors at selected values of `age`.

Source: Efron and Hastie (2016)

Boostrap standard errors

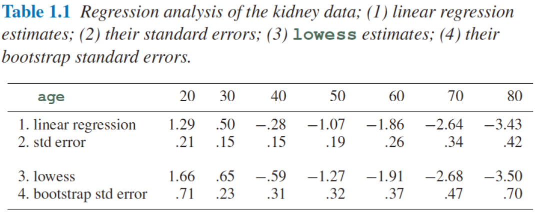
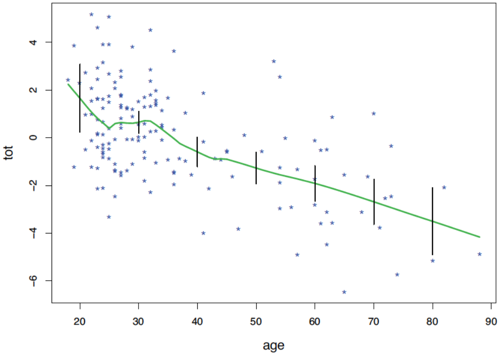


Table 1.1 *Regression analysis of the kidney data; (1) linear regression estimates; (2) their standard errors; (3)* `lowess` *estimates; (4) their bootstrap standard errors.*

age	20	30	40	50	60	70	80
1. linear regression	1.29	.50	−.28	−1.07	−1.86	−2.64	−3.43
2. std error	.21	.15	.15	.19	.26	.34	.42
3. lowess	1.66	.65	−.59	−1.27	−1.91	−2.68	−3.50
4. bootstrap std error	.71	.23	.31	.32	.37	.47	.70



Boostrap replications

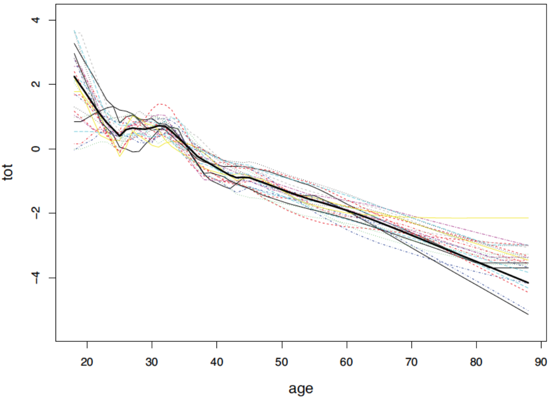


Figure 1.3 25 bootstrap replications of `lowess(x,y,1/3)`.

Source: Efron and Hastie (2016)