Algorithms and Inference

Statistical Learning

From Efron and Hastie (2016), chapter 1

Algorithms and Inference

Statistics is the science of *learning from experience*, particularly experience that arrives a little bit at a time

- the successes and failures of a new experimental drug
- the uncertain measurements of an asteroid's path toward Earth
- etc.

There are two main aspects of statistical analysis:

- the *algorithmic* aspect
- the *inferential* aspect

The distinction begins with the most basic, and most popular, statistical method, averaging.

Suppose we have observed $y_1, \ldots, y_n$, realizations of the random variables $Y_1, \ldots, Y_n$ i.i.d. Y, applying to some phenomenon of interest, where we choose that the *parameter of interest* is $\mathbb{E}(Y)$.

Averaging is the *algorithm*

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

How accurate is that number?

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

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

It is a surprising, and crucial, aspect of statistical theory that the same data that supplies an estimate can also assess its accuracy.

Of course, $\widehat{\text{se}}$ is itself an algorithm, which could be (and is) subject to further inferential analysis concerning its accuracy.

The point is that the algorithm comes *first* and the inference follows at a *second* level of statistical consideration.

In practice this means that algorithmic invention is a more free-wheeling and adventurous enterprise, with inference playing catch-up as it strives to assess the accuracy, good or bad, of some hot new algorithmic methodology.

Basic Gaussian model

In the basic Gaussian model with $Y_1, \dots, Y_n$ i.i.d. $Y \sim \mathcal{N}(\mu, \sigma^2)$, where

- the unknown parameter μ is the *parameter of interest*, and we wish to assess the relation of the data to that value
- σ^2 is the *nuisance parameter*, which can be either known or unknown.

One well-known *point estimator* for μ is

$$\hat{\mu} = \frac{1}{n}\sum_{i=1}^{n} Y_i \sim N(\mu, \sigma^2/n)$$

with standard error

$$\text{se}(\hat{\mu}) = \sqrt{\text{Var}(\hat{\mu})} = \sigma\sqrt{1/n}$$

If σ^2 is unknown, we use the estimator of the standard error

$$\widehat{\text{se}}(\hat{\mu}) = \hat{\sigma}\sqrt{1/n}$$

where the estimator for σ^2 is

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

which is independent from $\hat{\mu}$.

A *pivot* for inference about μ is

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

with $\Pr(-t_{n-1}^{1-\alpha/2} \le T \le t_{n-1}^{1-\alpha/2}) = 1-\alpha$ where $t_{n-1}^{1-\alpha/2}$ is the $1-\alpha/2$ quantile of the Student $\mathcal{T}$ distribution with $n-1$ degrees of freedom.

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

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

The *coverage* probability of the confidence interval in (1) is then

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

The following simulation corroborates the coverage property of the confidence interval:

```
sim = function(alpha=0.05, n=5, mu=0, sigma=1){
  y = rnorm(n, mean=mu, sd=sigma)
  hatmu = mean(y)
  hatse = sqrt( var(y) / n )
  k = qt(alpha/2, df = n-1, lower.tail = F)
  ci = hatmu + c(-1,1) * k * hatse
  cover = (mu >= ci[1] & mu <= ci[2])
```

```r
  return(cover)
}
set.seed(123)
mean( replicate(100, sim(alpha=0.05) ))
```

```
## [1] 0.96
```

Suppose now that the target of inference is a future realization y of Y.

The point predictor for Y is $\hat{Y} = \frac{1}{n}\sum_{i=1}^{n} Y_i$ (which is equal to $\hat{\mu}$), and the $1-\alpha$ *prediction interval* is given by

$$[\underline{Y}, \overline{Y}] = \hat{Y} \pm t_{n-1}^{1-\alpha/2}\hat{\sigma}\cdot\sqrt{1+1/n} \tag{2}$$

The prediction interval in (2) can be obtained from the pivot

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

and it has the property that

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

The following simulation corroborates the coverage property of the predition interval:

```r
sim = function(alpha=0.05, n=5, mu=0, sigma=1){
  y = rnorm(n, mean=mu, sd=sigma)
  haty = mean(y)
  k = qt(alpha/2, df = n-1, lower.tail = F)
  pi = haty + c(-1,1) * k * sqrt( var(y) * (1 + (1/n)) )
  y = rnorm(1, mean=mu, sd=sigma)
  cover = (y >= pi[1] & y <= pi[2])
  return(cover)
}
set.seed(123)
mean( replicate(100, sim(alpha=0.05) ))
```

```
## [1] 0.93
```

References

- Efron and Hastie (2016) Computer-Age Statistical Inference: Algorithms, Evidence, and Data Science, Cambridge University Press