Statistical Inference

Lecture 09a

ANU - RSFAS

Last Updated: Mon May 1 11:11:53 2017

Non-Parametric Methods

- So far we have assumed a particular parametric model and examined the estimation of parameters in those models.
- Now we will no longer make that assumption.
- Parameters are just surrogates for some numerical characteristic of the population.
- We can think about building estimation procedures through some "function" of the underlying population distribution.
- Let's denote this quantity:

$$\theta(F)$$

where $F = F(x)$ is a is the CDF of the population distribution of numerical characteristics.

Non-Parametric Methods

- The empirical distribution function $\hat{F}$ is the CDF of a new discrete random variable, say X^*.
- It can be shown that $\hat{F}$ is a sufficient statistic for F (based on a random sample), so
- $\hat{F}$ and X^* mimics the relationship between F and the X.
- This leads to studying $(\hat{F}, X^*)$ to learn about (F, X).

$$\hat{F}(x) = \frac{1}{n} \sum_{i=1}^{n} \mathbb{I}_{(x_i \leq x)}$$

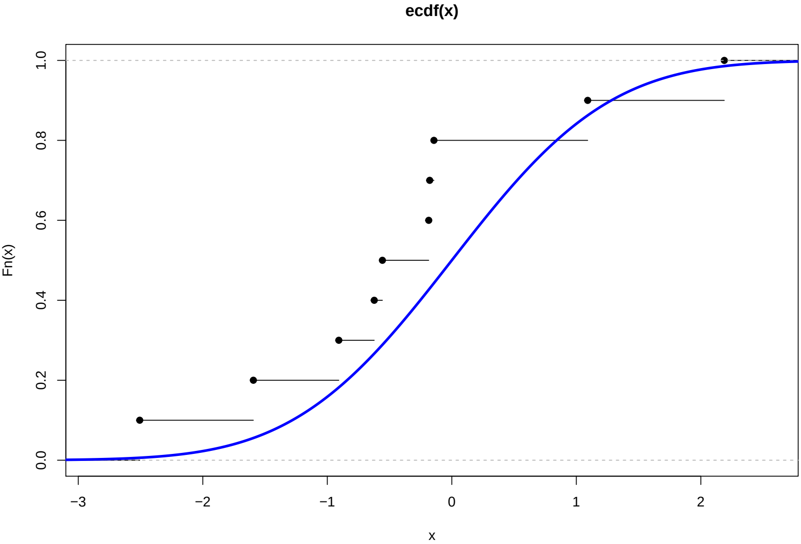
Non-Parametric Methods

```r
set.seed(1001)
x <- rnorm(10)
sort(x)
```

```
##  [1] -2.5065362 -1.5937133 -0.9074604 -0.6229437 -0.5573113
##  [7] -0.1775473 -0.1435595  1.0915017  2.1886481
```

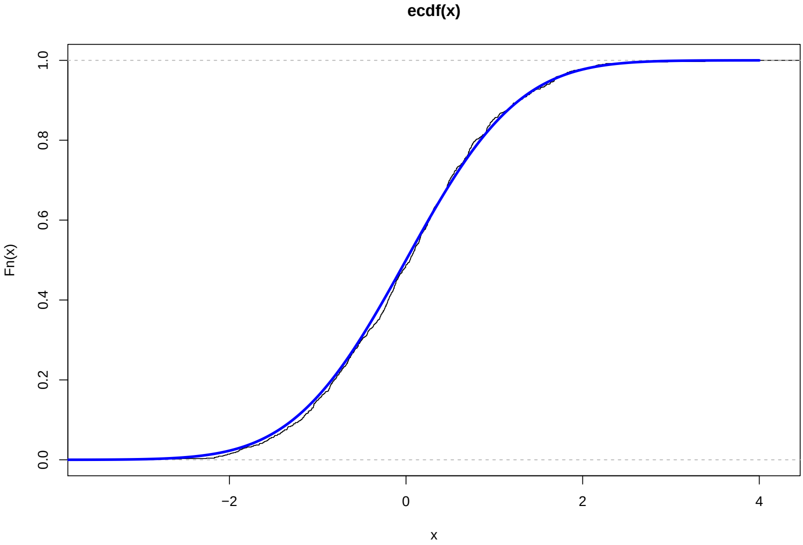
```r
F <- ecdf(x)
plot(F)
z <- seq(-4,4, by=0.001)
lines(z, pnorm(z), lwd=3)
```

Non-Parametric Methods



- data (black), truth (blue)

Non-Parametric Methods $n = 1000$



- data (black), truth (blue)

Non-Parametric Methods

- Let's start with a basic question:

$$\begin{aligned}
E\{\hat{F}(x)\} &= E\left\{\frac{1}{n}\sum_{i=1}^{n}\mathbb{I}_{(x_i\leq x)}\right\} \\
&= \frac{1}{n}\sum_{i=1}^{n}E\left\{\mathbb{I}_{(x_i\leq x)}\right\} \\
&= \frac{1}{n}\sum_{i=1}^{n}Pr(X_i\leq x) = \frac{1}{n}\sum_{i=1}^{n}F(x) = F(x)
\end{aligned}$$

Non-Parametric Methods

- Consider again the empirical distribution function:

$$\hat{F} = \frac{1}{n}\sum_{i=1}^{n} \mathbb{I}_{(x_i \leq x)} = \frac{n_x}{n}$$

where n_x is defined as the number of observed data values which are less than or equal to the value x.

Non-Parametric Methods

- Another way to see this — n_x is binomially distributed with n trials and a "success" probability of $p = Pr(X_i \leq x) = F(x)$.

$$E\left(\hat{F}\right) = E\left(\frac{n_x}{n}\right) = \frac{E\left(n_x\right)}{n} = \frac{np}{n} = \frac{nF(x)}{n} = F(x)$$

- So we also have:

$$\begin{aligned} Var\{\hat{F}\} &= Var\left(\frac{n_x}{n}\right) = \frac{1}{n^2} Var(n_x) \\ &= \frac{1}{n^2} np(1-p) = \frac{1}{n} F(x)(1-F(x)) \end{aligned}$$

Non-Parametric Methods

- Typically we are interested in a function of F. Let's consider the mean of X which has distribution F.
- A standard function of interest is:

$$\theta(F) = \int_{-\infty}^{\infty} x f(x) dx$$

- So we are interested in $\theta(F) = E_F(X)$:

$$\hat{\theta} = \theta(\hat{F}) = E_{\hat{F}}(X) = \sum_{x \in \mathcal{X}} x p_{\hat{F}}(x) = \sum_{i=1}^{n} x_i \frac{1}{n} = \bar{x}$$

- Note: $p_{\hat{F}}(x) = \frac{1}{n} \quad \forall x \in \mathcal{X}$.

Non-Parametric Methods - Bootstrap

- We are interested in the behavior of $\hat{\theta}$ under F.
- Let's examine $\hat{\theta}^* = \theta(\hat{F}^*)$ associated with $X_1^*, \ldots, X_n^*$ having distribution $\hat{F}$.
- To examine $\hat{\theta}$ under F, imagine that our observed data forms its own "population" from which we randomly sample according to the "true" distribution $\hat{F}$.
- Construct an estimate of the "true" population parameter $\theta(\hat{F})$ using the "re-sampled" data, arriving at $\theta(\hat{F}^*)$ as the estimator for $\theta(\hat{F})$.

Non-Parametric Methods - Bootstrap

- Since we know the “true” distribution $\hat{F}$, we can determine exactly the bias and variance of $\theta(\hat{F}^*)$.
- So we use the bias and variance of $\theta(\hat{F}^*)$ under $\hat{F}$ as estimators for the bias and variance of $\theta(\hat{F})$ under F.
- This approach is generally referred to as the bootstrap, since we are using the data itself to estimate its behavior under F, effectively "pulling ourselves up by our own bootstraps".

Non-Parametric Methods - Bootstrap

- Define the bootstrap estimators of bias and variance as:

$$\hat{B}_B = E_{\hat{F}}\{\theta(\hat{F}^*)\} - \theta(\hat{F})$$

$$\hat{Var}_B\{\theta(\hat{F}^*)\} = E_{\hat{F}}\{\theta(\hat{F}^*)^2\} - \Big[E_{\hat{F}}\{\theta(\hat{F}^*)\}\Big]^2$$

- These calculations are occasionally possible exactly in the case of simple functionals $\theta(\cdot)$.
- However, generally they are estimated through computational means!
- Note: we have the observed values $x_1, \ldots, x_n$ in our possession, we can easily create realisations of the random sample $X_1^*, \ldots, X_n^*$ simply randomly drawing n values with replacement from the collection $\mathcal{X} = \{x_1, \ldots, x_n\}$.

Non-Parametric Methods - Bootstrap

- We create B "bootstrap" (re-sampled) data sets:

$$\{X^*_{1,1}, \ldots, X^*_{n,1}\}, \ldots, \{X^*_{1,b}, \ldots, X^*_{n,b}\}, \ldots, \{X^*_{1,B}, \ldots, X^*_{n,B}\}$$

- With each sample we estimate $\hat{\theta}^b = \theta(\hat{F}^*_b)$, then we can compute:

$$\hat{B}_B = E_{\hat{F}}\{\theta(\hat{F}^*)\} - \theta(\hat{F}) \approx \frac{1}{B}\sum_{b=1}^{B} \hat{\theta}^*_b - \hat{\theta}$$

$$\hat{Var}_B\{\theta(\hat{F}^*)\} \approx \frac{1}{B-1}\sum_{b=1}^{B}\left(\hat{\theta}^*_b - \frac{1}{B}\sum_{b=1}^{B}\hat{\theta}^*_b\right)^2$$

Non-Parametric Methods - Bootstrap

- Note: Casella and Berger suggest that all non unique re-samples (with replacement) should be determined.
- Eg. Suppose we have the following data:

$$2, 3, 9, 12$$

- For the first spot we have 4 possibilities. The second spot we have 4 possibilities, and so on:

$$B = n^n = 4^4 = 256$$

- Typically this is not done! And B is just set to be a large number.

Non-Parametric Methods - Bootstrap

Eg. Law School - Correlation between LSAT and GPA

Suppose that we have observed the following data pairs, which represent the average LSAT (Legal Scholastic Aptitude Test, a common entrance exam for prospective law school students in the United States) and undergraduate GPA (grade point average) scores for the 1973 entering class at a random sample of 15 U.S. Law Schools. We are interested in the an estimate of the correlation between LSAT (Y) and GPA (Z), along with an estimate of it's bias and variance:

$$\theta(F) = \rho_F = \frac{Cov_F(Y, Z)}{\sqrt{Var_F(Y) Var_F(Z)}}$$

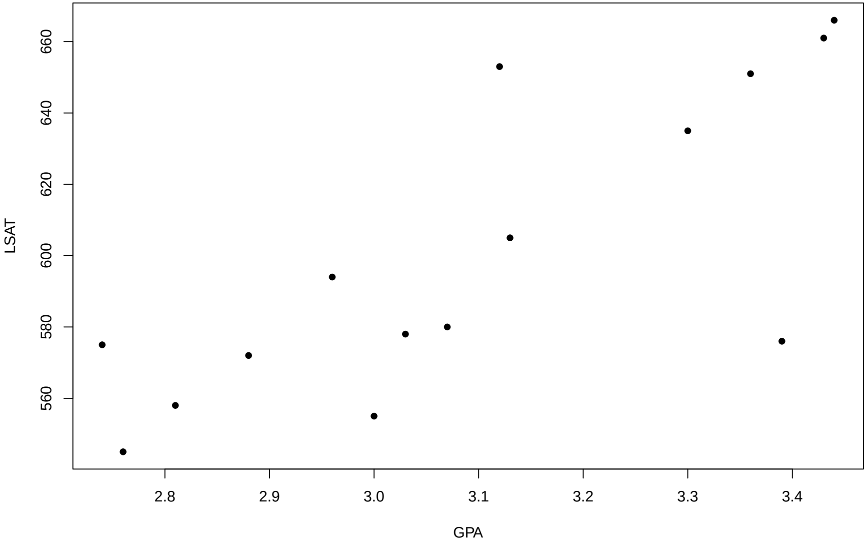
$$\hat{\rho} = \frac{\sum_{i=1}^{n}(y_i - \bar{y})(z_i - \bar{z})}{\sqrt{\sum_{i=1}^{n}(y_i - \bar{y})^2 \sum_{i=1}^{n}(z_i - \bar{z})^2}}$$

Non-Parametric Methods - Bootstrap

```
## law school data
LSAT <- c(576, 578, 555, 605, 545, 635, 666, 661,
          653, 572, 558, 580, 651, 575, 594)
GPA <- c(3.39, 3.03, 3.00, 3.13, 2.76, 3.30, 3.44,
         3.43, 3.12, 2.88, 2.81, 3.07, 3.36, 2.74, 2.96)

D <- data.frame(LSAT, GPA)
n <- nrow(D)


##
plot(GPA, LSAT, pch=16)
rho.hat <- cor(LSAT, GPA)
rho.hat
```



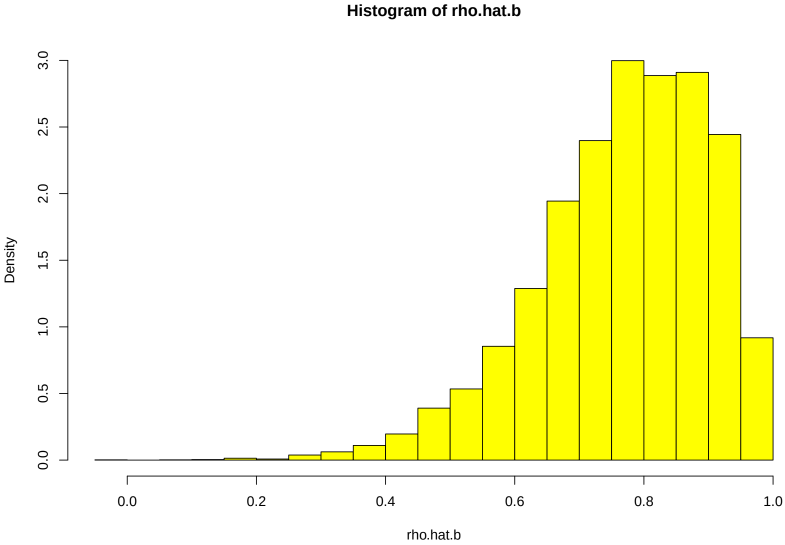
```
## [1] 0.7763745
```

```r
## Bootstrap
#####################################################
## Take a single bootstrap sample
## As it is random it is different than in the notes
set.seed(2000)
S <- sample(1:n, n, replace = TRUE)
D.s <- D[S,]

### Let's do B samples
B <- 10000
D.B <- array(list(), B)
rho.hat.b <- rep(0, B)

for(b in 1:B){
S <- sample(1:n, n, replace = TRUE)
D.B[[b]] <- D[S,]

##
rho.hat.b[b] <- cor(D.B[[b]]$LSAT, D.B[[b]]$GPA)
    }
hist(rho.hat.b, col="yellow", prob=TRUE)
```


Histogram of rho.hat.b
Density
0.0
0.5
1.0
1.5
2.0
2.5
3.0
0.0
0.2
0.4
0.6
0.8
1.0
rho.hat.b

- Estimate, estimate of the bias of the estimate, estimate of the variance of the estimate.

```
rho.hat
```

```
## [1] 0.7763745
```

```
Bias.B.hat <- mean(rho.hat.b) - rho.hat
Bias.B.hat
```

```
## [1] -0.007520911
```

```
Var.B.hat <- var(rho.hat.b)
Var.B.hat
```

```
## [1] 0.01784111
```

Some Thoughts

- The idea behind the bootstrap is powerful and extremely intuitively appealing.
- Why, then, has the bootstrap not replaced parametric approaches?
    - As implemented, the bootstrap method yields a different answer every time (of course, the differences will be very small if B is large).
    - Another drawback is that if θ is complicated to calculate (perhaps because it is implicitly defined as the solution to an equation, just as the MLE was) then computing its value for each of B re-sampled data sets is computationally quite expensive and time consuming.
    - If we truly believe the parametric structure we have set up, then the parametric estimators have nice optimal properties.

Some Thoughts

- The bootstrap is a very flexible and widely applicable approach which deserves more attention than it currently gets among statistical practitioners
- The bootstrap can even be extended to circumstances beyond the *iid* setting on which we have focused here.
- But a word of warning on complicated (non iid settings):
    - We cannot always guarantee that using the bootstrap paradigm (replacing F by $\hat{F}$ and $\hat{F}$ by $\hat{F}^*$) to estimate bias and variance will yield valid estimates.