Nonparametric Inference and Bootstrap – Order Statistics

Heike Hofmann

October 2, 2013
Outline of Presentation

1. Warmup question
2. Order statistics: minimum, maximum, median, interquartile range
3. Empirical distribution function
4. Mini-excursion into exploratory data analysis
A library has on its shelves 114 books on statistics. Based on a sample of 12 books we want to test the hypothesis that the median number of pages of statistics books is $\theta = 225$. Out of the sample of 12 books 3 books have fewer than 225 pages. Is that consistent with the null hypothesis, or do we have to revise it? (assume $H_1 : \theta \neq 225$) based on the page numbers of the sample of 12 books find a confidence interval for the median number of pages in a statistics book at a level of not less than 95%:

$$126, 142, 156, 228, 245, 246, 370, 419, 433, 454, 478, 503$$
Order statistics and ranks are fundamental to a lot of non-parametric procedures: the rank test used ranks and their sums, in the Binomial test we did not use ranks, but we made use of the order of the sample.

The maximum might be of interest in worst-case scenarios, such as for example flood levels – how high does the dam have to be to withstand the floods for the next 50 years. What is the probability of a flood exceeding, say 5m, in the future?

Similarly, minima are used for times until breakdown – e.g. how long should the warranty for a product be? What is the time until the first breakdown?

Both of the above questions would be answered with extreme value distributions in a parametric setting.

For a non-parametric setting, both values are just examples of order statistics.
Order Statistics

Let $x_1, x_2, \ldots, x_n$ be a sample of $n$ observations from a continuous distribution $F$.

Continuous distribution implies, that we will assume that all values $x_i$ are different.

Then denote by $x(1), x(2), \ldots, x(n)$ the ordered sample, i.e.

$$x(1) < x(2) < \ldots < x(n)$$

In particular, $x(1)$ denotes the minimum of the sample, $x(n)$ its maximum.

The median is defined as $x(n+1)/2$ if $n$ is odd, and the average of the middle two values, in case $n$ is even: $n = 2m$, median is $0.5 \cdot (x(m) + x(m+1))$. 
Range of the sample: \( x_{(n)} - x_{(1)} \)

Interquartile Range: first quartile \( q_1 \) is defined as the median of the lower half of the sample (less than or equal to the median); the third quartile \( q_3 \) is defined as the median of the upper half of the sample. 

\[
\text{IQR} = q_3 - q_1
\]

is a good measure of the variability in a sample

\[
\begin{align*}
x & \leftarrow \text{rnorm}(100) \\
\text{sd}(x) & \\
[1] & 1.162 \\
\text{IQR}(x) & \\
[1] & 1.632
\end{align*}
\]
Interquartile Range versus standard deviation

For a normal distribution, the ratio of sd to IQR is about 0.7
Tukey definition of outlying value: any value outside $(q_1 - 1.5 \cdot IQR, q_3 + 1.5 \cdot IQR)$ is an outlier.

Interquartile-range is a more robust measure of variability than range, because range is based on maximum and minimum; both of these statistics are influenced by outliers.
Sample distribution function $S_n$

The Sample distribution function or empirical cumulative distribution function (ECDF) is defined as

$$S_n(x) = \frac{1}{n} \cdot (\# \text{ of values } \leq x)$$

$S_n(x) = 0$ for all $x < x_{(1)}$

$S_n(x) = 1$ for all $x \geq x_{(n)}$

$S_n(x) = \frac{i}{n}$ for all $x_{(i)} \leq x < x_{(i+1)}$ for $i = 1, \ldots, n - 1$
Sample distribution function $S_n$

```r
x <- rnorm(8)
sort(x)

[1] -0.6129 -0.5847 -0.3214 0.4599 0.5451 1.2764 1.2978 2.5001

plot(ecdf(x))
```
Different distributions

```r
x <- sort(runif(100))
y <- sort(rexp(100, rate=1))
z <- sort(rnorm(100))
qplot(x, seq_along(x)/100, geom="step") +
  geom_step(aes(y, seq_along(y)/100), colour="darkred") +
  geom_step(aes(z, seq_along(z)/100), colour="darkblue") + ylab("ECDF")
```

![ECDF plot](image)
Sample distribution – 3 theoretical results

- The expected value of $S_n(x)$ is $F(x)$:
  \[ E[S_n(x)] = F(x) \]
- Variance of $S_n(x)$:
  \[ \text{Var}[S_n(x)] = F(x)(1 - F(x))/n \]
- $S_n(x)$ is a consistent estimator of $F(x)$ for any $x$, i.e. $S_n(x)$ converges in probability to $F(x)$ as $n$ goes to infinity.

By looking at the empirical distribution function we can learn quite a bit about the actual distribution of a sample. This is one of the main motivations behind Exploratory Data Analysis (EDA).
Elements of EDA

- Numerical summary statistics
- Graphical representations: boxplot, histogram, scatterplot, ...
- Empirical cumulative distribution functions, comparisons with theoretical distributions (e.g. Q-Q-plot)
Crabs data

crabs <- read.csv(url)

## Warning: unable to resolve 'www.ggobi.org'
## Error: cannot open the connection

head(crabs)
qplot(RW, data=crabs, facets=species~sex, geom="histogram", binwidth=1)

## Error: object 'crabs' not found
```r
qplot(species, FL, data = crabs, fill = sex, geom = "boxplot")
```

```
## Error: object 'crabs' not found
```
```r
qplot(RW, CW, color = species:sex, data = crabs)

## Error: object 'crabs' not found
```
```r
qplot(FL, index/50, color=species:sex, data=crabs, geom="step")
```

```r
## Error: object 'crabs' not found
```