Nonparametric Inference and Bootstrap – Q-Q plots; Kolmogorov test

Heike Hofmann

October 11, 2013
Outline of Presentation

1. Q-Q plots
2. Warm-up example
3. Kolmogorov-Smirnov Test
4. Lilliefors test for normality
The quantile $Q$ corresponding to probability $p$ is defined as

$$Q(p) = F^{-1}(p) = \inf \{ x : F(x) \geq p \}$$

- for continuous distribution function $F$ straightforward
- problematic for empirical estimate $S_n$ of $F$
Quantile definitions

- $\text{Freq}(X_k \leq \hat{Q}_1(p)) = \lceil pn \rceil$ and $\text{Freq}(X_k \geq \hat{Q}_1(1 - p)) = \lfloor pn + 1 \rfloor$ (shown left)
- $\text{Freq}(X_k \leq \hat{Q}_2(p)) = \text{Freq}(X_k \geq \hat{Q}_2(1 - p)) = \lceil pn \rceil$ (shown right)
- more complicated ...
- R has 9 different definitions of quantiles behind its \texttt{quantile} function
Q-Q plots

- For a sample \( x_1, \ldots, x_n \) compute sample quantiles, plot in a scatterplot against a) theoretical quantiles of a hypothesized distribution, or b) quantiles of a second sample.
- We can say that the sample is consistent with the theoretical distribution or the two samples come from the same distribution, if the points line up along the line of identity in the Q-Q plot.
QQ-plots from scratch

- Comparing one sample against a theoretical distribution

```r
x <- sort(rnorm(20))
seqx <- seq_along(x)/21
qplot(x, qnorm(seqx)) + geom_smooth(method="lm") + geom_abline(colour="red")
```

![QQ-plot example](image)
QQ-plot diagnostics

Relationship to identity line tells about location shifts (parallel to identity) or differences in scale (different tail behavior; identity crosses sample)
Q-Q plots

- For two samples of the same length, plot sorted values in a scatterplot.

Low and high average temperatures in Ames:

```r
sort(temps$Low)
sort(temps$High)
```
The following observations come from a stress test that measures on pieces of string of length 6 cm.
0.6, 1.2, 1.8, 1.8, 2.4, 2.4, 2.4, 3, 3, 3, 3, 3, 3, 3.6, 3.6, 4.2, 4.2, 4.2, 4.2, 4.2

Do these observations come from a uniform distribution on (0,6)?
... and now in R:

```r
x <- c(0.6, 1.2, 1.8, 1.8, 2.4, 2.4, 2.4, 3, 3, 3, 3, 3, 3, 3.6, 3.6, 4.2, 4.2, 4.2, 4.2, 4.2)
qplot(x, seq(0, 1, length=length(x))) + geom_abline(intercept=0, slope=1/6)
```

How much deviation is too much deviation?
Kolmogorov Smirnov test

Let $x_1, x_2, ..., x_n$ be an i.i.d sample

Let $F$ be a continuous distribution

$H_0$: $x_1, x_2, ..., x_n \sim F$ versus $H_1$: the sample is not from $F$

idea of Kolmogorov-Smirnov (KS) test: compare $S_n$ to $F$, i.e. test statistic

$$D := \sup \{|F(x) - S_n(x)| \text{ for all } x\}$$
Kolmogorov-Smirnov

The difference between $F$ and $S_n$ is maximized in one of the steps
Need to only look at $|F(x_i) - S_n(x_i)|$ or $|F(x_i) - S_n(x_i - 1)|$

<p>| | | | | | | | | | | |</p>
<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>d1</td>
<td>-0.05</td>
<td>-0.10</td>
<td>-0.10</td>
<td>-0.10</td>
<td>-0.05</td>
<td>-0.05</td>
<td>-0.05</td>
<td>0.15</td>
<td>0.15</td>
<td>0.15</td>
</tr>
<tr>
<td>d2</td>
<td>-0.15</td>
<td>-0.20</td>
<td>-0.10</td>
<td>-0.20</td>
<td>-0.05</td>
<td>-0.05</td>
<td>-0.15</td>
<td>0.15</td>
<td>0.15</td>
<td>0.15</td>
</tr>
</tbody>
</table>

<p>| | | | | | | | | | | |</p>
<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>d1</td>
<td>0.15</td>
<td>0.15</td>
<td>0.15</td>
<td>0.15</td>
<td>0.15</td>
<td>0.30</td>
<td>0.30</td>
<td>0.30</td>
<td>0.30</td>
<td>0.30</td>
</tr>
<tr>
<td>d2</td>
<td>0.15</td>
<td>0.15</td>
<td>0.05</td>
<td>0.15</td>
<td>0.05</td>
<td>0.30</td>
<td>0.30</td>
<td>0.30</td>
<td>0.30</td>
<td>0.00</td>
</tr>
</tbody>
</table>
$D$ has Kolmogorov-Smirnov distribution, see below for $n=20$ (black), 10 (red), and 5 (blue)
ks.test(x, punif, min = 0, max = 6)

## Warning: ties should not be present for the Kolmogorov-Smirnov test

## One-sample Kolmogorov-Smirnov test

## data: x
## D = 0.3, p-value = 0.05465
## alternative hypothesis: two-sided
KS-test for normality

Kolmogorov only tests for exact distribution, i.e. does not adjust for different parameters in the same family

```r
z <- rnorm(20, mean = 1)
ks.test(z, pnorm)

## One-sample Kolmogorov-Smirnov test
## data: z
## D = 0.5028, p-value = 3.334e-05
## alternative hypothesis: two-sided
```

```r
ks.test(z, pnorm, mean = 1)

## One-sample Kolmogorov-Smirnov test
```
KS-test for normality

kn.test(z, pnorm, mean = 1)

##
## One-sample Kolmogorov-Smirnov test
##
## data:  z
## D = 0.1887, p-value = 0.4228
## alternative hypothesis: two-sided
Lilliefors test for normality

Instead of $X$ test $(X - \hat{\mu}_X)/\hat{\sigma}_X$ for standard normality

Lilliefors test:

```r
library(nortest)
lillie.test(x)
```

```output
## Lilliefors (Kolmogorov-Smirnov) normality test
##
data:  x
## D = 0.1729, p-value = 0.1213
```