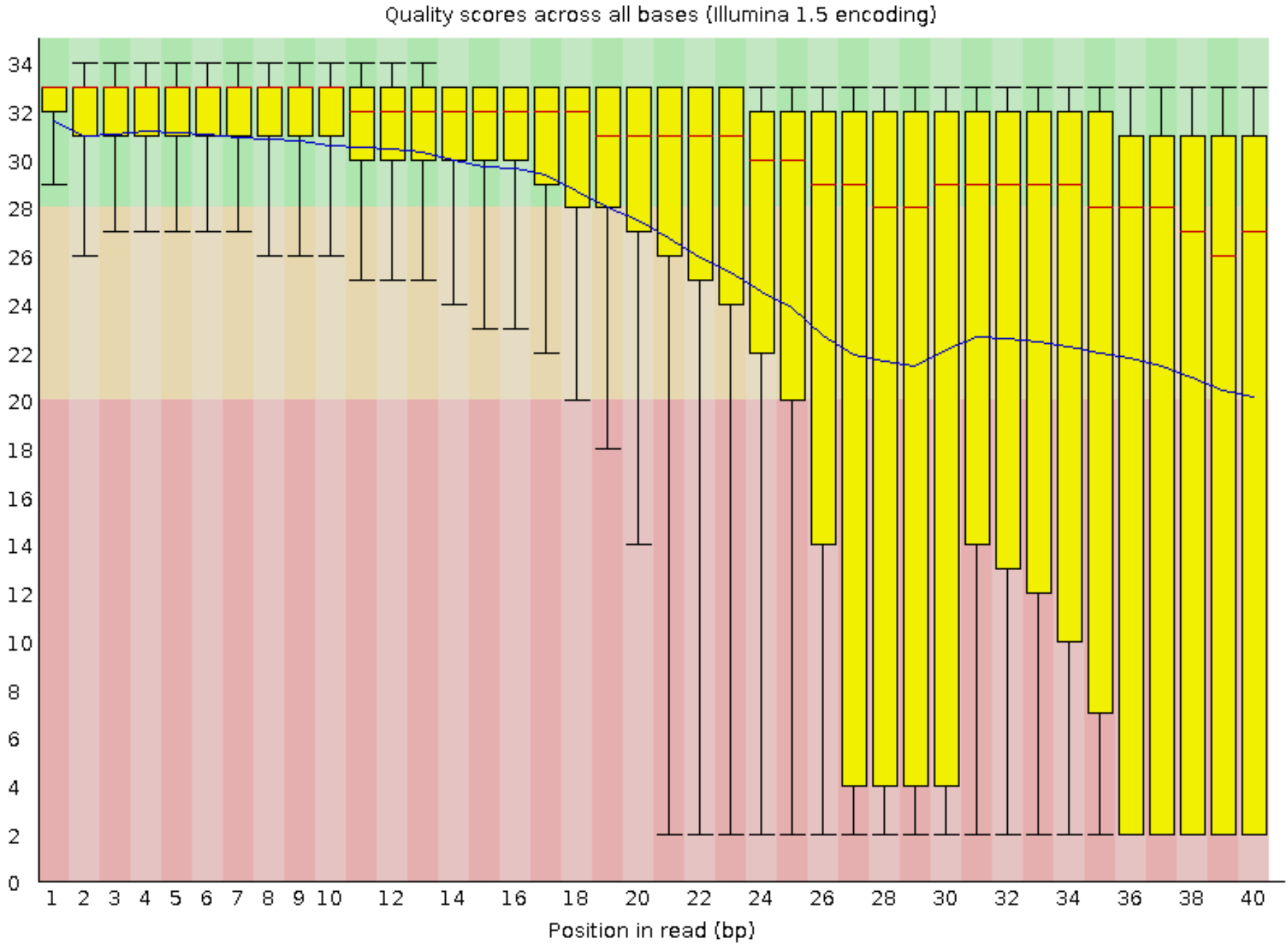
Boxplot（盒形图/箱线图）

Yi Xianfu
xfyin@sibs.ac.cn

Institute of Health Sciences
Shanghai Institute for Biological Science
Chinese Academy of Science

December 19, 2011


Quality scores across all bases (Illumina 1.5 encoding)
34
32
30
28
26
24
22
20
18
16
14
12
10
8
6
4
2
0
1 2 3 4 5 6 7 8 9 10 12 14 16 18 20 22 24 26 28 30 32 34 36 38 40
Position in read (bp)

Outline

Boxplot | 简介

Definition

箱线图（Boxplot）也称箱须图（Box-whisker Plot），是利用数据中的五个统计量：最小值、第一四分位数、中位数、第三四分位数与最大值来描述数据的一种方法。它也可以粗略地看出数据是否具有有对称性，分布的离散程度等信息；特别适用于对几个样本的比较。

箱线图美中不足之处在于它不能提供关于数据分布偏态和尾重程度的精确度量；对于批量较大的数据集，箱线图反映的形状信息更加模糊；用中位数代表总体平均水平有一定的局限性等等。所以，应用箱线图最好结合其它描述统计工具如均值、标准差、偏度、分布函数等来描述数据集的分布形状。

Boxplot | 简介

Definition

箱线图（Boxplot）也称箱须图（Box-whisker Plot），是利用数据中的五个统计量：最小值、第一四分位数、中位数、第三四分位数与最大值来描述数据的一种方法。它也可以粗略地看出数据是否具有有对称性，分布的离散程度等信息；特别适用于对几个样本的比较。

箱线图美中不足之处在于它不能提供关于数据分布偏态和尾重程度的精确度量；对于批量较大的数据集，箱线图反映的形状信息更加模糊；用中位数代表总体平均水平有一定的局限性等等。所以，应用箱线图最好结合其它描述统计工具如均值、标准差、偏度、分布函数等来描述数据集的分布形状。

Boxplot | 相关概念

- 最小值 min，最大值 max。
- 中位数 median。
- 上四分位数 Q3，下四分位数 Q1。
- 四分位数差 IQR (interquartile range) = Q3 - Q1。
- 内限：Q3+1.5IQR，Q1-1.5IQR。
- 外限：Q3+3IQR，Q1-3IQR。
- 异常值 (outliers)：处于内限以外的数据。
- 温和的异常值 (mild outliers)：在内限与外限之间的异常值。
- 极端的异常值 (extreme outliers)：在外限以外的异常值。

Boxplot | 相关概念

- 最小值 min，最大值 max。
- 中位数 median。
- 上四分位数 Q3，下四分位数 Q1。
- 四分位数差 IQR (interquartile range) = Q3 - Q1。
- 内限：Q3+1.5IQR，Q1-1.5IQR。
- 外限：Q3+3IQR，Q1-3IQR。
- 异常值 (outliers)：处于内限以外的数据。
- 温和的异常值 (mild outliers)：在内限与外限之间的异常值。
- 极端的异常值 (extreme outliers)：在外限以外的异常值。

Boxplot | 相关概念

- 最小值 min，最大值 max。
- 中位数 median。
- 上四分位数 Q3，下四分位数 Q1。
- 四分位数差 IQR (interquartile range) = Q3 - Q1。
- 内限：Q3+1.5IQR，Q1-1.5IQR。
- 外限：Q3+3IQR，Q1-3IQR。
- 异常值 (outliers)：处于内限以外的数据。
- 温和的异常值 (mild outliers)：在内限与外限之间的异常值。
- 极端的异常值 (extreme outliers)：在外限以外的异常值。

Boxplot | 相关概念

- 最小值 min，最大值 max。
- 中位数 median。
- 上四分位数 Q3，下四分位数 Q1。
- 四分位数差 IQR (interquartile range) = Q3 - Q1。
- 内限：Q3+1.5IQR，Q1-1.5IQR。
- 外限：Q3+3IQR，Q1-3IQR。
- 异常值 (outliers)：处于内限以外的数据。
- 温和的异常值 (mild outliers)：在内限与外限之间的异常值。
- 极端的异常值 (extreme outliers)：在外限以外的异常值。

Boxplot | 相关概念

- 最小值 min，最大值 max。
- 中位数 median。
- 上四分位数 Q3，下四分位数 Q1。
- 四分位数差 IQR (interquartile range) = Q3 - Q1。
- 内限：Q3+1.5IQR，Q1-1.5IQR。
- 外限：Q3+3IQR，Q1-3IQR。
- 异常值 (outliers)：处于内限以外的数据。
- 温和的异常值 (mild outliers)：在内限与外限之间的异常值。
- 极端的异常值 (extreme outliers)：在外限以外的异常值。

Boxplot | 相关概念

- 最小值 min，最大值 max。
- 中位数 median。
- 上四分位数 Q3，下四分位数 Q1。
- 四分位数差 IQR (interquartile range) = Q3 - Q1。
- 内限：Q3+1.5IQR，Q1-1.5IQR。
- 外限：Q3+3IQR，Q1-3IQR。
- 异常值 (outliers)：处于内限以外的数据。
- 温和的异常值 (mild outliers)：在内限与外限之间的异常值。
- 极端的异常值 (extreme outliers)：在外限以外的异常值。

Boxplot | 相关概念

- 最小值 min，最大值 max。
- 中位数 median。
- 上四分位数 Q3，下四分位数 Q1。
- 四分位数差 IQR (interquartile range) = Q3 - Q1。
- 内限：Q3+1.5IQR，Q1-1.5IQR。
- 外限：Q3+3IQR，Q1-3IQR。
- 异常值 (outliers)：处于内限以外的数据。
- 温和的异常值 (mild outliers)：在内限与外限之间的异常值。
- 极端的异常值 (extreme outliers)：在外限以外的异常值。

Boxplot | 相关概念

- 最小值 min，最大值 max。
- 中位数 median。
- 上四分位数 Q3，下四分位数 Q1。
- 四分位数差 IQR（interquartile range）= Q3 - Q1。
- 内限：Q3+1.5IQR，Q1-1.5IQR。
- 外限：Q3+3IQR，Q1-3IQR。
- 异常值（outliers）：处于内限以外的数据。
- 温和的异常值（mild outliers）：在内限与外限之间的异常值。
- 极端的异常值（extreme outliers）：在外限以外的异常值。

Boxplot | 相关概念

- 最小值 min，最大值 max。
- 中位数 median。
- 上四分位数 Q3，下四分位数 Q1。
- 四分位数差 IQR（interquartile range）= Q3 - Q1。
- 内限：Q3+1.5IQR，Q1-1.5IQR。
- 外限：Q3+3IQR，Q1-3IQR。
- 异常值（outliers）：处于内限以外的数据。
- 温和的异常值（mild outliers）：在内限与外限之间的异常值。
- 极端的异常值（extreme outliers）：在外限以外的异常值。

Outline

Boxplot | 绘制

1. **绘制数轴。**
2. 计算上四分位数（Q3），中位数，下四分位数（Q1）。
3. 计算四分位数差（IQR）。
4. 绘制箱线图的矩形，上限为 Q3，下限为 Q1。在矩形内部中位数的位置画一条横线（中位线）。
5. 在 Q3+1.5IQR 和 Q1-1.5IQR 处画两条与中位线一样的线段，这两条线段为异常值截断点，称为内限；在 Q3+3IQR 和 Q1-3IQR 处画两条线段，称为外限。
   注意：统计软件绘制的箱线图一般都没有标出内限和外限。
6. 在非异常值的数据中，最靠近上边缘和下边缘（即内限）的两个数值处画横线，作为箱线图的触须。
7. 从矩形的两端向外各画一条线段直到不是异常值的最远点（即上一步的触须），表示该批数据正常值的分布区间。
8. 温和的异常值用空心点表示；极端的异常值用实心点（一说用 * ）表示。

Boxplot | 绘制

1. 绘制数轴。
2. 计算上四分位数（Q3），中位数，下四分位数（Q1）。
3. 计算四分位数差（IQR）。
4. 绘制箱线图的矩形，上限为 Q3，下限为 Q1。在矩形内部中位数的位置画一条横线（中位线）。
5. 在 Q3+1.5IQR 和 Q1-1.5IQR 处画两条与中位线一样的线段，这两条线段为异常值截断点，称为内限；在 Q3+3IQR 和 Q1-3IQR 处画两条线段，称为外限。

   注意：统计软件绘制的箱线图一般都没有标出内限和外限。
6. 在非异常值的数据中，最靠近上边缘和下边缘（即内限）的两个数值处画横线，作为箱线图的触须。
7. 从矩形的两端向外各画一条线段直到不是异常值的最远点（即上一步的触须），表示该批数据正常值的分布区间。
8. 温和的异常值用空心点表示；极端的异常值用实心点（一说用 * ）表示。

Boxplot | 绘制

1. 绘制数轴。
2. 计算上四分位数（Q3），中位数，下四分位数（Q1）。
3. 计算四分位数差（IQR）。
4. 绘制箱线图的矩形，上限为 Q3，下限为 Q1。在矩形内部中位数的位置画一条横线（中位线）。
5. 在 Q3+1.5IQR 和 Q1-1.5IQR 处画两条与中位线一样的线段，这两条线段为异常值截断点，称为内限；在 Q3+3IQR 和 Q1-3IQR 处画两条线段，称为外限。
   注意：统计软件绘制的箱线图一般都没有标出内限和外限。
6. 在非异常值的数据中，最靠近上边缘和下边缘（即内限）的两个数值处画横线，作为箱线图的触须。
7. 从矩形的两端向外各画一条线段直到不是异常值的最远点（即上一步的触须），表示该批数据正常值的分布区间。
8. 温和的异常值用空心点表示；极端的异常值用实心点（一说用 * ）表示。

Boxplot | 绘制

1. 绘制数轴。
2. 计算上四分位数（Q3），中位数，下四分位数（Q1）。
3. 计算四分位数差（IQR）。
4. 绘制箱线图的矩形，上限为 Q3，下限为 Q1。在矩形内部中位数的位置画一条横线（中位线）。
5. 在 Q3+1.5IQR 和 Q1-1.5IQR 处画两条与中位线一样的线段，这两条线段为异常值截断点，称为内限；在 Q3+3IQR 和 Q1-3IQR 处画两条线段，称为外限。
   注意：统计软件绘制的箱线图一般都没有标出内限和外限。
6. 在非异常值的数据中，最靠近上边缘和下边缘（即内限）的两个数值处画横线，作为箱线图的触须。
7. 从矩形的两端向外各画一条线段直到不是异常值的最远点（即上一步的触须），表示该批数据正常值的分布区间。
8. 温和的异常值用空心点表示；极端的异常值用实心点（一说用 * ）表示。

Boxplot | 绘制

1. 绘制数轴。
2. 计算上四分位数（Q3），中位数，下四分位数（Q1）。
3. 计算四分位数差（IQR）。
4. 绘制箱线图的矩形，上限为 Q3，下限为 Q1。在矩形内部中位数的位置画一条横线（中位线）。
5. 在 Q3+1.5IQR 和 Q1-1.5IQR 处画两条与中位线一样的线段，这两条线段为异常值截断点，称为内限；在 Q3+3IQR 和 Q1-3IQR 处画两条线段，称为外限。
   注意：统计软件绘制的箱线图一般都没有标出内限和外限。
6. 在非异常值的数据中，最靠近上边缘和下边缘（即内限）的两个数值处画横线，作为箱线图的触须。
7. 从矩形的两端向外各画一条线段直到不是异常值的最远点（即上一步的触须），表示该批数据正常值的分布区间。
8. 温和的异常值用空心点表示；极端的异常值用实心点（一说用 * ）表示。

Boxplot | 绘制

1. 绘制数轴。
2. 计算上四分位数（Q3），中位数，下四分位数（Q1）。
3. 计算四分位数差（IQR）。
4. 绘制箱线图的矩形，上限为 Q3，下限为 Q1。在矩形内部中位数的位置画一条横线（中位线）。
5. 在 Q3+1.5IQR 和 Q1-1.5IQR 处画两条与中位线一样的线段，这两条线段为异常值截断点，称为内限；在 Q3+3IQR 和 Q1-3IQR 处画两条线段，称为外限。
   注意：统计软件绘制的箱线图一般都没有标出内限和外限。
6. 在非异常值的数据中，最靠近上边缘和下边缘（即内限）的两个数值处画横线，作为箱线图的触须。
7. 从矩形的两端向外各画一条线段直到不是异常值的最远点（即上一步的触须），表示该批数据正常值的分布区间。
8. 温和的异常值用空心点表示；极端的异常值用实心点（一说用星号 * ）表示。

Boxplot | 绘制

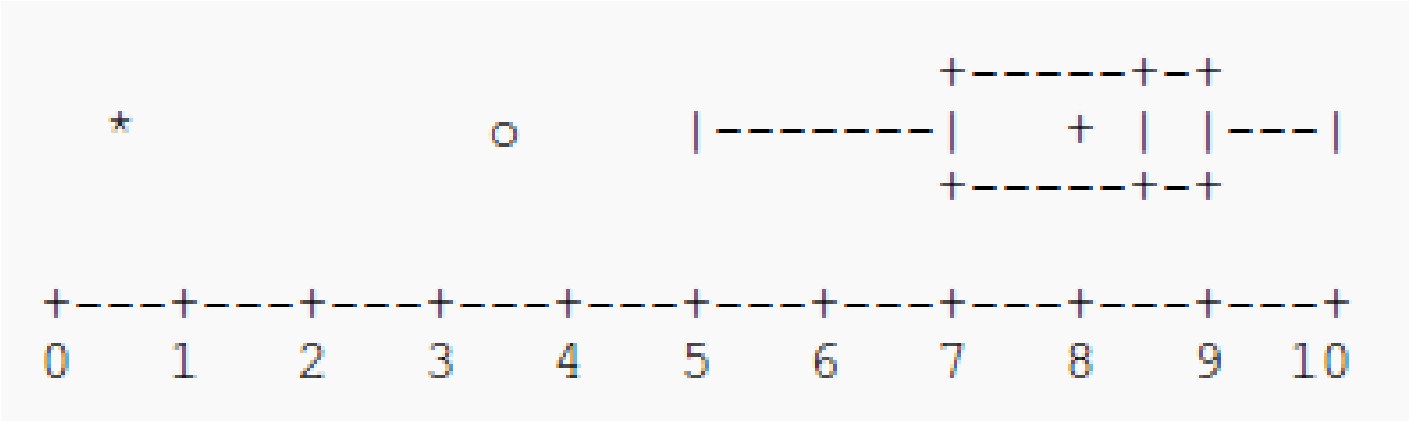
1. 绘制数轴。
2. 计算上四分位数（Q3），中位数，下四分位数（Q1）。
3. 计算四分位数差（IQR）。
4. 绘制箱线图的矩形，上限为 Q3，下限为 Q1。在矩形内部中位数的位置画一条横线（中位线）。
5. 在 Q3+1.5IQR 和 Q1-1.5IQR 处画两条与中位线一样的线段，这两条线段为异常值截断点，称为内限；在 Q3+3IQR 和 Q1-3IQR 处画两条线段，称为外限。
   注意：统计软件绘制的箱线图一般都没有标出内限和外限。
6. 在非异常值的数据中，最靠近上边缘和下边缘（即内限）的两个数值处画横线，作为箱线图的触须。
7. 从矩形的两端向外各画一条线段直到不是异常值的最远点（即上一步的触须），表示该批数据正常值的分布区间。
8. 温和的异常值用空心点表示；极端的异常值用实心点（一说用星号 * ）表示。

Boxplot | 绘制

1. 绘制数轴。
2. 计算上四分位数（Q3），中位数，下四分位数（Q1）。
3. 计算四分位数差（IQR）。
4. 绘制箱线图的矩形，上限为 Q3，下限为 Q1。在矩形内部中位数的位置画一条横线（中位线）。
5. 在 Q3+1.5IQR 和 Q1-1.5IQR 处画两条与中位线一样的线段，这两条线段为异常值截断点，称为内限；在 Q3+3IQR 和 Q1-3IQR 处画两条线段，称为外限。
   注意：统计软件绘制的箱线图一般都没有标出内限和外限。
6. 在非异常值的数据中，最靠近上边缘和下边缘（即内限）的两个数值处画横线，作为箱线图的触须。
7. 从矩形的两端向外各画一条线段直到不是异常值的最远点（即上一步的触须），表示该批数据正常值的分布区间。
8. 温和的异常值用空心点表示；极端的异常值用实心点（一说用星号 * ）表示。

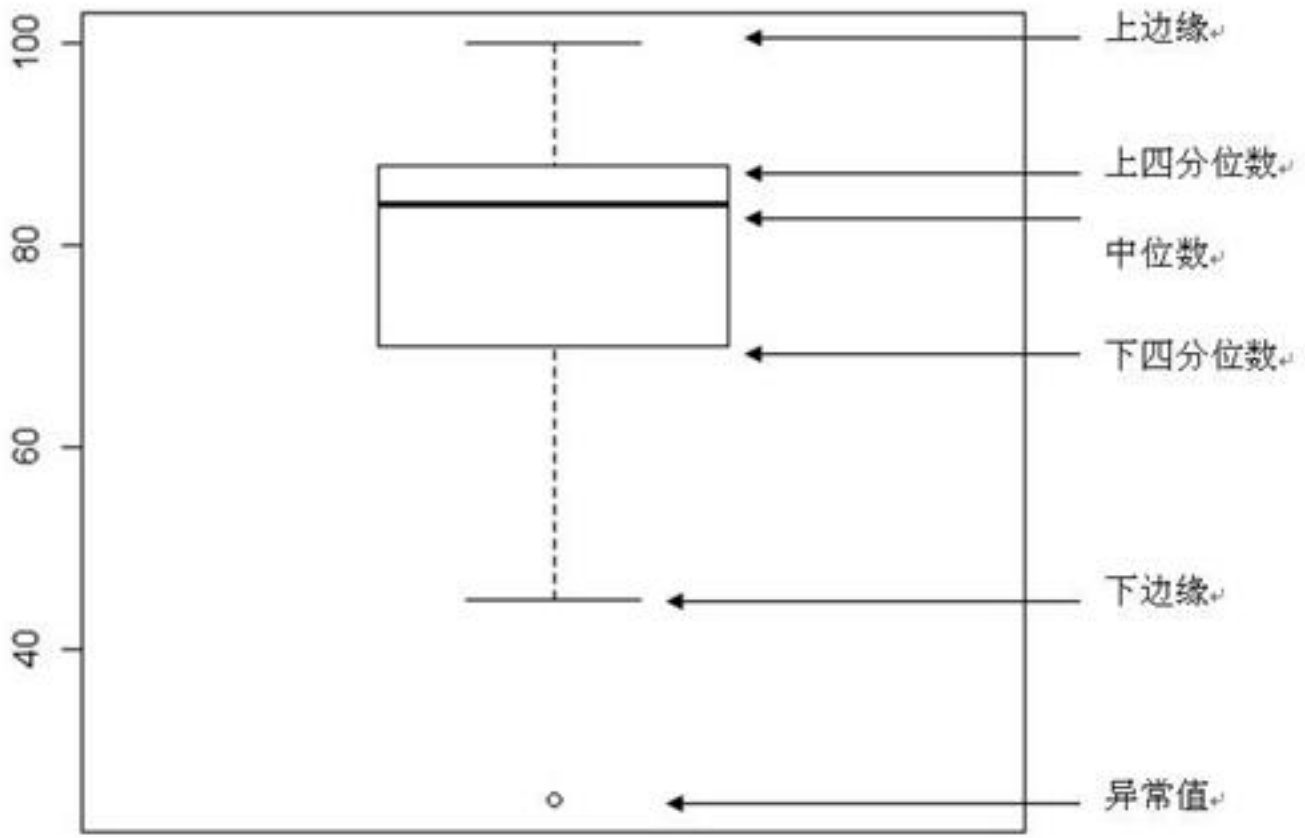
Outline

Boxplot | 图解

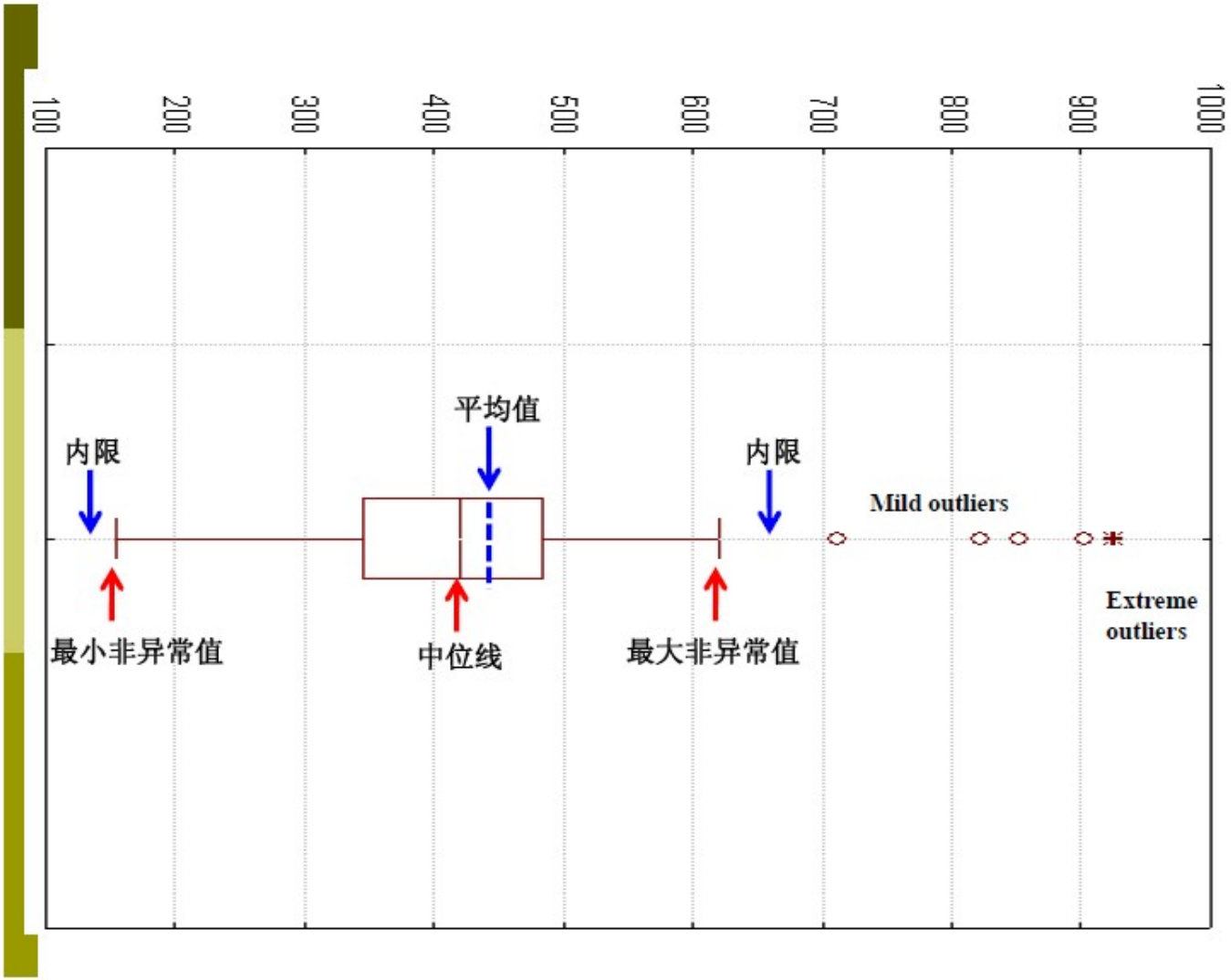


最小值 (min)=0.5；下四分位数 (Q1)=7；中位数 (Med)=8.5；上四分位数 (Q3)=9；最大值 (max)=10；平均值 =8；四分位数差 (interquartile range，四分位间距)=Q3 − Q1=2。

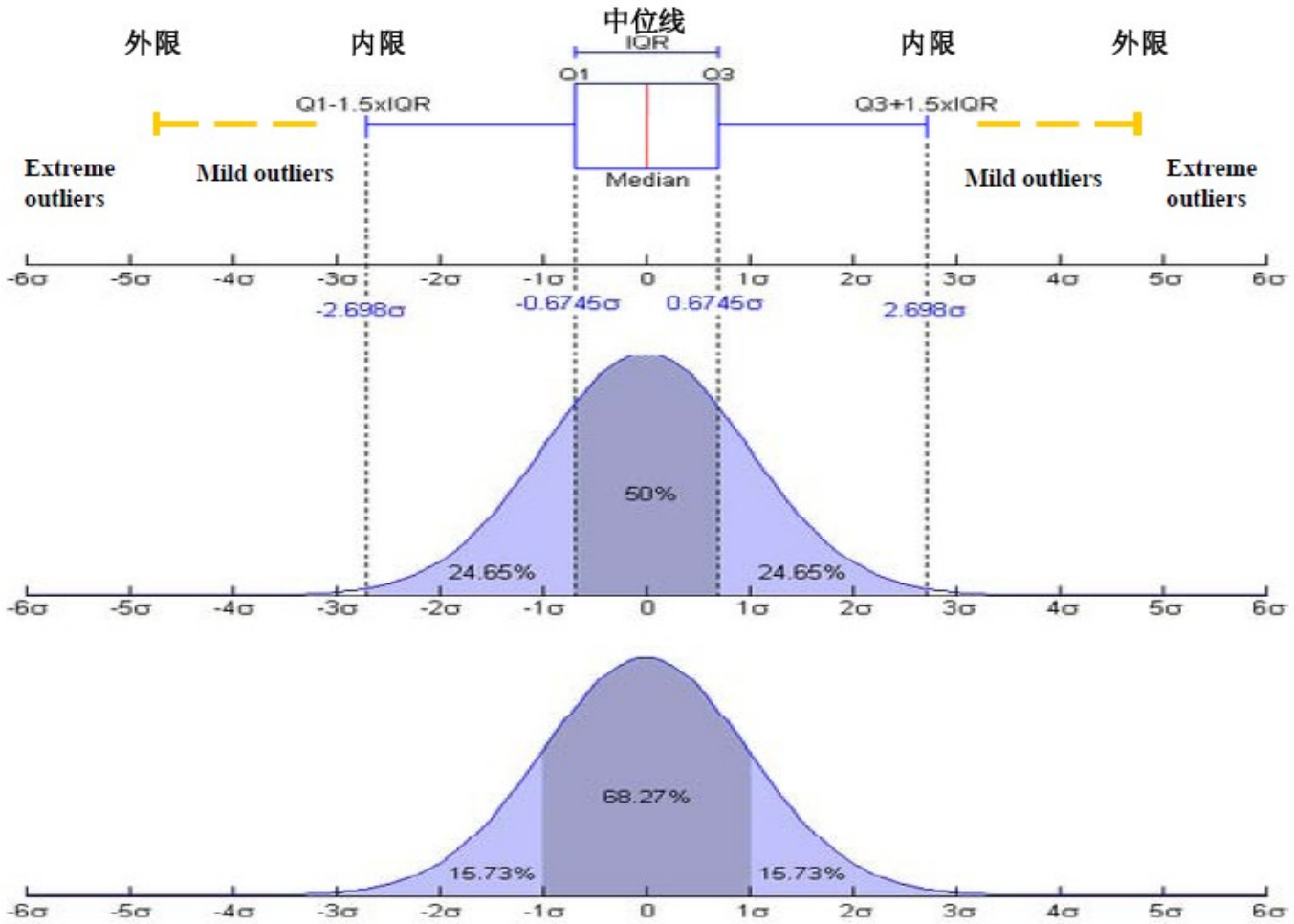
Boxplot | 图解



Boxplot | 图解


100
200
300
400
500
600
700
800
900
1000
内限
平均值
内限
Mild outliers
Extreme outliers
最小非异常值
中位线
最大非异常值

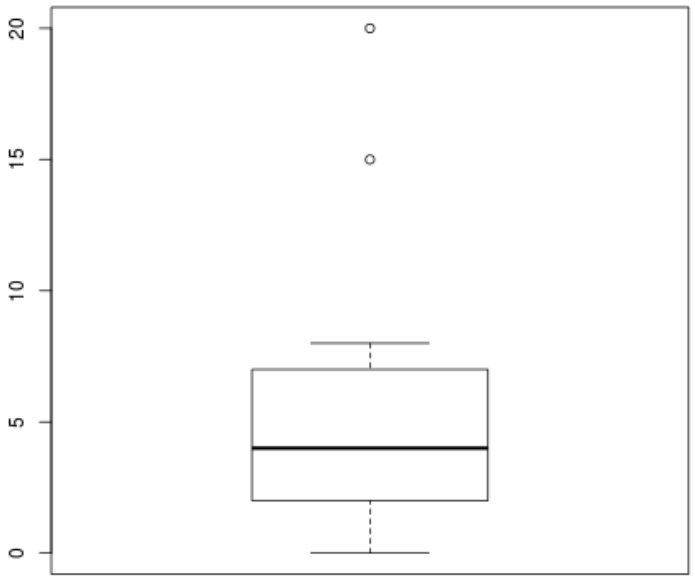
Boxplot | 图解


外限
内限
中位线
IQR
内限
外限
Q1
Q3
Q1-1.5xIQR
Q3+1.5xIQR
Extreme outliers
Mild outliers
Median
Mild outliers
Extreme outliers
-6σ -5σ -4σ -3σ -2σ -1σ 0 1σ 2σ 3σ 4σ 5σ 6σ
-2.698σ
-0.6745σ
0.6745σ
2.698σ
50%
24.65%
24.65%
68.27%
15.73%
15.73%

Outline

Boxplot | 总览数据

```
> x <- c(0,4,15, 1, 6, 3, 20, 5, 8, 1, 3)
> sort(x)
 [1]  0  1  1  3  3  4  5  6  8 15 20
> summary(x)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
      0       2       4       6       7      20
> boxplot(x)
```



Boxplot | 总览数据 | 步骤详述

1. The median and the quartiles are used to construct the box.
2. Multiply the IQR by 1.5. 1.5 x IQR = 1.5 (5) = 7.5. (STEP = 7.5)
3. Add the STEP to the third quartile, obtaining 3rd Quartile + STEP = 7 + 7.5 = 14.5.
   - Any data beyond 14.5 is plotted using an empty circle. (15 and 20)
   - Locate the largest data point below 14.5. (8). This is where the end of the upper whisker is drawn.
4. Subtract the STEP from the first quartile, obtaining 1st Quartile - STEP = 2 - 7.5 = -5.5.

Boxplot | 总览数据 | 步骤详述

1. The median and the quartiles are used to construct the box.
2. Multiply the IQR by 1.5. 1.5 x IQR = 1.5 (5) = 7.5. (STEP = 7.5)
3. Add the STEP to the third quartile, obtaining 3rd Quartile + STEP = 7 + 7.5 = 14.5.
   - Any data beyond 14.5 is plotted using an empty circle. (15 and 20).
   - Locate the largest data point below 14.5. (8). This is where the end of the upper whisker is drawn.
4. Subtract the STEP from the first quartile, obtaining 1st Quartile - STEP = 2 - 7.5 = -5.5.

Boxplot | 总览数据 | 步骤详述

1. The median and the quartiles are used to construct the box.
2. Multiply the IQR by 1.5. 1.5 x IQR = 1.5 (5) = 7.5. (STEP = 7.5)
3. Add the STEP to the third quartile, obtaining 3rd Quartile + STEP = 7 + 7.5 = 14.5.
   - Any data beyond 14.5 is plotted using an empty circle. (15 and 20).
   - Locate the largest data point below 14.5. (8). This is where the end of the upper whisker is drawn.
4. Subtract the STEP from the first quartile, obtaining 1st Quartile - STEP = 2 - 7.5 = -5.5.

Boxplot | 总览数据 | 步骤详述

1. The median and the quartiles are used to construct the box.
2. Multiply the IQR by 1.5. 1.5 x IQR = 1.5 (5) = 7.5. (STEP = 7.5)
3. Add the STEP to the third quartile, obtaining 3rd Quartile + STEP = 7 + 7.5 = 14.5.
   - Any data beyond 14.5 is plotted using an empty circle. (15 and 20).
   - Locate the largest data point below 14.5. (8). This is where the end of the upper whisker is drawn.
4. Subtract the STEP from the first quartile, obtaining 1st Quartile - STEP = 2 - 7.5 = -5.5.
   - Any data below -5.5 is plotted using an empty circle. (NO).
   - Locate the smallest data point that occurs above -5.5. (0). This is where the end of the lower whisker is drawn.

Boxplot | 总览数据 | 步骤详述

1. The median and the quartiles are used to construct the box.
2. Multiply the IQR by 1.5. 1.5 x IQR = 1.5 (5) = 7.5. (STEP = 7.5)
3. Add the STEP to the third quartile, obtaining 3rd Quartile + STEP = 7 + 7.5 = 14.5.
   - Any data beyond 14.5 is plotted using an empty circle. (15 and 20).
   - Locate the largest data point below 14.5. (8). This is where the end of the upper whisker is drawn.
4. Subtract the STEP from the first quartile, obtaining 1st Quartile - STEP = 2 - 7.5 = -5.5.
   - Any data below -5.5 is plotted using an empty circle. (NO).
   - Locate the smallest data point that occurs above -5.5. (0). This is where the end of the lower whisker is drawn.

Boxplot | 总览数据 | 步骤详述

1. The median and the quartiles are used to construct the box.
2. Multiply the IQR by 1.5. 1.5 x IQR = 1.5 (5) = 7.5. (STEP = 7.5)
3. Add the STEP to the third quartile, obtaining 3rd Quartile + STEP = 7 + 7.5 = 14.5.
   - Any data beyond 14.5 is plotted using an empty circle. (15 and 20).
   - Locate the largest data point below 14.5. (8). This is where the end of the upper whisker is drawn.
4. Subtract the STEP from the first quartile, obtaining 1st Quartile - STEP = 2 - 7.5 = -5.5.
   - Any data below -5.5 is plotted using an empty circle. (NO).
   - Locate the smallest data point that occurs above -5.5. (0). This is where the end of the lower whisker is drawn.

Boxplot | 总览数据 | 步骤详述

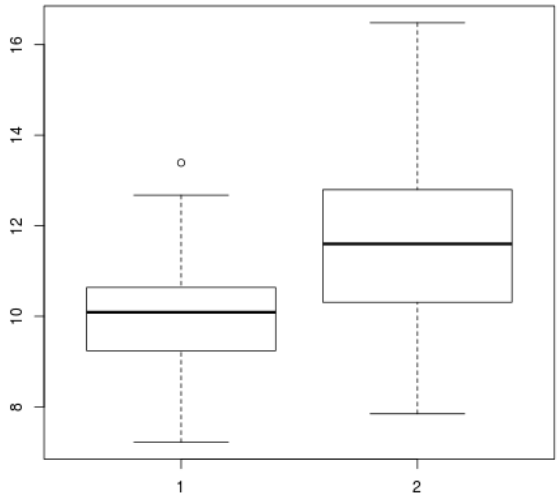
1. The median and the quartiles are used to construct the box.
2. Multiply the IQR by 1.5. 1.5 x IQR = 1.5 (5) = 7.5. (STEP = 7.5)
3. Add the STEP to the third quartile, obtaining 3rd Quartile + STEP = 7 + 7.5 = 14.5.
   - Any data beyond 14.5 is plotted using an empty circle. (15 and 20).
   - Locate the largest data point below 14.5. (8). This is where the end of the upper whisker is drawn.
4. Subtract the STEP from the first quartile, obtaining 1st Quartile - STEP = 2 - 7.5 = -5.5.
   - Any data below -5.5 is plotted using an empty circle. (NO).
   - Locate the smallest data point that occurs above -5.5. (0). This is where the end of the lower whisker is drawn.

Boxplot | 总览数据 | 步骤详述

1. The median and the quartiles are used to construct the box.
2. Multiply the IQR by 1.5. 1.5 x IQR = 1.5 (5) = 7.5. (STEP = 7.5)
3. Add the STEP to the third quartile, obtaining 3rd Quartile + STEP = 7 + 7.5 = 14.5.
   - Any data beyond 14.5 is plotted using an empty circle. (15 and 20).
   - Locate the largest data point below 14.5. (8). This is where the end of the upper whisker is drawn.
4. Subtract the STEP from the first quartile, obtaining 1st Quartile - STEP = 2 - 7.5 = -5.5.
   - Any data below -5.5 is plotted using an empty circle. (NO).
   - Locate the smallest data point that occurs above -5.5. (0). This is where the end of the lower whisker is drawn.

Boxplot | 数据比较

```
> a <- rnorm(100,mean=10,sd=1)
> b <- rnorm(100,mean=12,sd=2)
> summary(a)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
  7.215   9.240  10.090   9.936  10.630  13.390
> summary(b)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
  7.847  10.320  11.600  11.720  12.780  16.480
> boxplot(a,b)
```



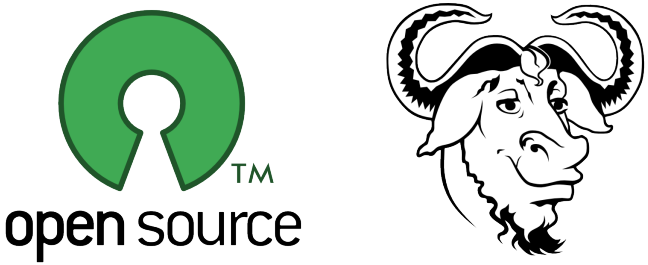
Outline

Boxplot | 参考

- 箱线图
- 什么是箱线图
- 箱线图 (Box plot)
- 箱形图
- Box plot
- Box Plot
- Boxplot
- Box Plots
- Box Plot: Display of Distribution
- Visual Presentation of Data by Means of Box Plots
- Boxplots in R


INSTITUTE OF HEALTH SCIENCES
IHS
RESEARCH · EDUCATION · INNOVATION
健康科学研究所

Powered by



TEX LATEX XETEX Beamer

Thanks for your attention!

Any questions?