Übungsrunde 2, Gruppe 2

LVA 107.369, Übungsrunde 1, Gruppe 2, 24.10.

Markus Nemetz, markus.nemetz@tuwien.ac.at, TU Wien, 10/2006

1 1.2.3.4.

1.1 Angabe

Bestimmen Sie für den Datensatz `bulb.dat` die Quartile, d.h. das 25%-, das 50%- (=Median) und das 75%- Quantil (Fraktil); stützen Sie sich dabei auf die klassierten Daten. (`bulb.r`)

1.2 Wichtige Begriffe

Wichtige Lageparameter sind die **Quartile**:

- **Median** (50%, 0.5 Quantil): Mindestens die Hälfte aller Werte sind kleiner oder gleich dem eines Medians ($x_{0.5}$) und Mindestens die Hälfte aller Werte sind grösser oder gleich dem eines Medians
- **25%-Quantil** (0.25): Wie beim Median, nur liegt die 'Scheide' bei 25%.
- **75%-Quantil** (0.75): Wie beim Median, nur liegt die 'Scheide' bei 75%.

1.3 Lösung der Aufgabenstellung

Listing 1: R: Daten klassifizieren, Quartile berechnen

```
x <- scan("bulb.dat")

n <- length(x)

breaks <- seq(500,1500,by=100)

m <- length(breaks)

#äHufigkeitstabelle

x.hist <- hist(x,breaks=breaks,plot=FALSE) ( haeuf.tab <-
cbind(u.i=breaks[1:(m-1)],o.i=breaks[2:m],
       z.i=x.hist$mids,b.i=diff(breaks),
       H.i=x.hist$counts,h.i=x.hist$counts/length(x),
       F.i=cumsum(x.hist$counts/length(x))) )

pfrac <- function(p) {
  i <- which(haeuf.tab[,"F.i"] >= p)[1]
  u <- haeuf.tab[,"u.i"][i]
  b <- haeuf.tab[,"b.i"][i]
  h <- haeuf.tab[,"h.i"][i]
  F <- haeuf.tab[,"F.i"][i]
  u + (p-(F-h))/h *b }

```

```
x25 <- pfrac(0.25)

x50 <- pfrac(0.50)

x75 <- pfrac(0.75)

x25
[1] 908.6207
x50
[1] 994.8276
x75
[1] 1116.279
```

2 1.2.4.3.

2.1 Angabe

Bestimmen und vergleichen Sie für das Merkmal 'Gewicht' (Datensatz: meddat.dat) die Varianz, die Streuung und den Variationskoeffizienten für Männer und Frauen. (`meddat.r`)

2.2 Wichtige Begriffe

Streuungsparameter (Daten $x_1, \ldots, x_n$)

- **Spannweite**: $x_{(n)} - x_{(1)} = \max_{i=1(1)n} x_i - \min_{i=1(1)n} x_i$
- **Quartilabstand**: $x_{0.75} - x_{0.25}$
- **Mittlere absolute Abweichung** (Median mit Welle unten (x_M); MAD = 'mean absolute deviation')

$$MAD = \frac{1}{n} \sum_{i=1}^{n} |x_i - x_M|$$

- **Mittlere quadratische Abweichung (empirische Varianz)**

$$s^2 := \frac{1}{n} \sum_{i=1}^{n} (x_i - \bar{x})^2$$

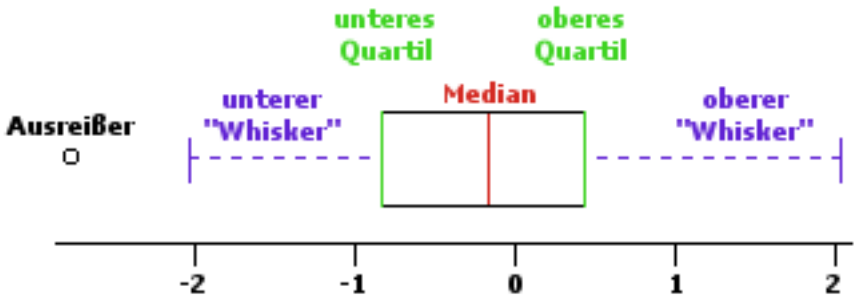
- **Empirische Streuung**

$$s := \sqrt{s^2}$$

- **Empirischer Variationskoeffizient** - dimensionsloses Streumaß

$$VK := \frac{s}{\bar{x}}$$

Der **Boxplot** (auch Box-Whisker-Plot) ist ein Diagramm, das zur graphischen Darstellung einer Reihe numerischer Daten verwendet wird. Er fasst verschiedene Maße der zentralen Tendenz, Streuung und Schiefe in einem Diagramm zusammen. Alle Werte der Fünf-Punkte-Zusammenfassung, also der Median, die zwei Quartile und die beiden Extremwerte, sind dargestellt.

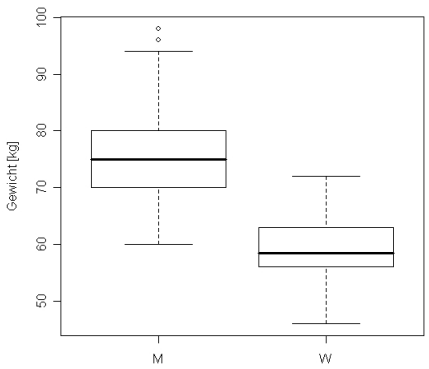


Als 'Box' wird das durch die Quartile bestimmte Rechteck bezeichnet. Sie umfasst 50% der Daten. Durch die Länge der Box ist der Interquartilsabstand (interquartile range, IQR) abzulesen. Dies ist ein Maß der Streuung, welches durch die Differenz des oberen und unteren Quartils bestimmt ist. Als weiteres Quantil ist der Median in der Box eingezeichnet, welcher durch seine Lage innerhalb der Box einen Eindruck von der Schiefe der den Daten zugrunde liegenden Verteilung vermittelt.
Als 'Whisker' werden die horizontalen Linien bezeichnet. In der Literatur finden sich drei verschiedene Definitionen über die Länge der Whisker:
Variante 1: Die Länge der Whisker beträgt maximal das 1,5-fache des Interquartilsabstands (1,5IQR) und wird immer durch einen Wert aus den Daten bestimmt. Werte, die über dieser Grenze liegen, werden separat in das Diagramm eingetragen und als Ausreißer bezeichnet. Gibt es keine Werte außerhalb der Whisker, so wird die Länge des Whiskers durch den maximalen bzw. minimalen Wert festgelegt.
Häufig werden Ausreißer, die zwischen 1,5*IQR und 3*IQR liegen als 'milde' Ausreißer bezeichnet und Werte, die über 3*IQR liegen als 'extreme' Ausreißer. Diese werden dann auch unterschiedlich im Diagramm gekennzeichnet.
Variante 2: Die Länge der Whisker entspricht der Differenz zwischen dem Minimum und dem unteren Quartil bzw. zwischen dem oberen Quartil und dem Maximum. Ausreißer werden in dieser Variante nicht dargestellt; Minimum und Maximum sind sofort erkennbar.
Variante 3: Berechnung des unteren Whisker als 2,5% Quantil. Berechnung des oberen als 97,5% Quantil. Innerhalb der Whiskergrenzen liegen somit 95% aller beobachteten Werte.
Die Behandlung von Ausreißern erfolgt wie in Variante 1.

2.3 Lösung der Aufgabenstellung

Listing 2: R: Varianz, Streuung, Variationskoeffizient

```
meddat <- read.table("meddat.dat",header=TRUE,skip=8)[,-1]

attach(meddat)

        The following object(s) are masked from meddat ( position 3 ) :

         BG GA GE GR GW RA RF

n.m <- length(GW[GE=="M"])

( m.var <- var(GW[GE=="M"])*(n.m-1)/n.m ) [1] 77.73307

( m.sd <- sqrt(m.var) )
[1] 8.816636

( m.vk <- m.sd/mean(GW[GE=="M"]) )
[1] 0.1161651

n.w <- length(GW[GE=="W"])
    ( w.var <- var(GW[GE=="W"])*(n.w-1)/n.w) #
[1] 34.31556

( w.sd <- sqrt(w.var) )
[1] 5.857948

( w.vk <- w.sd/mean(GW[GE=="W"]) )
[1] 0.1001930

#Gewicht - Boxplots
boxplot(GW ~ GE,ylab="Gewicht [kg]")
```



3 1.2.4.5.

3.1 Angabe

Zeigen Sie den Verschiebungssatz für die Varianz (Daten: $x_1, \ldots, x_n$):

$$\sum_{i=1}^{n}(x_i - \bar{x}_n)^2 = \sum_{i=1}^{n}(x_i - c)^2 - n \cdot (c - x_n)^2 \qquad c \in \mathbb{R}$$

Speziell für $c = 0$ gilt:

$$s^2 = \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x}_n)^2 = \frac{1}{n}(\sum_{i=1}^{n} x_i^2 - n\bar{x_n}^2) = \frac{1}{n}\sum_{i=1}^{n} x_i^2 - \bar{x}_n^2$$

3.2 Wichtige Begriffe

Der **Verschiebungssatz** (auch Satz von Steiner - nach Ernst Steiner benannt) ist eine Rechenregel für die Ermittlung der Summe quadratischer Abweichungen.

3.3 Lösung der Aufgabenstellung

Es gilt:

$$\sum_{i=1}^{n} x_i = n \cdot \bar{x}_n$$

Linke Seite:

$$\sum_{i=1}^{n}(x_i - \bar{x}_n)^2 =$$

$$\sum_{i=1}^{n}(x_i^2 - x_i\bar{x}_n + \bar{x}_n^2) =$$

$$\sum_{i=1}^{n} x_i^2 - 2\bar{x_n}n\bar{x_n} + n\bar{x_n}^2 = \sum_{i=1}^{n} x_i^2 - n\bar{x}_n^2$$

Rechte Seite:

$$\sum_{i=1}^{n}(x_i - c)^2 - n \cdot (c - x_n)^2 = \qquad c \in \mathbb{R}$$

$$\sum_{i=1}^{n}(x_i^2 - 2cx_i + c^2) - nc^2 + 2nc\bar{x}_n - n\bar{x}_n^2 = \sum_{i=1}^{n} x_i^2 - 2c\sum_{i=1}^{n} x_i + nc^2 - nc^2 + 2c\sum_{i=1}^{n} x_i - n\bar{x}_n^2 =$$

$$\sum_{i=1}^{n} x_i^2 - n\bar{x}_n^2$$

4 1.3.1.5.

4.1 Angabe

In einem Feld der Länge n werden zufällig k ($k \leq n$) Daten abgelegt. Mit welcher Wahrscheinlichkeit kommt es dabei zu Kollisionen (d.h. Mehrfachbelegungen)? Wie groß muß k konkret für $n = 100$ mindestens sein, damit diese Wahrscheinlichkeit größer als 0.5 (0.9) ist?

4.2 Wichtige Begriffe

Klassische Wahrscheinlichkeitsdefinition

- m mögliche Versuchsausgänge - $M = \{a_1, a_2, \ldots, a_m\}$
- Ereignis $A = \{a_{i_1}, \ldots, a_{i_g}\}, i_{i_j} \in \{1, \ldots, m\}$, A enthält g Elemente
- $W(A)$ ist Wahrscheinlichkeit von A
- $W(A) := \frac{g}{m}$

4.3 Lösung der Aufgabenstellung

Die Wahrscheinlichkeit einer Kollision (WK) kann am einfachsten mit der Gegenwahrscheinlichkeit (GW) berechnet werden: also: $WK = 1 - GW$. Die GW (also keine Kollision) ist um einiges leichter zu berechnen:

$$GW = \frac{1 \cdot (n-1)}{n} \cdot \frac{n-2}{n} \cdot \ldots \cdot \frac{n-k+1}{n} = 1 - (\frac{n \cdot (n-1) \cdot \ldots}{n^k})$$

Die Wahrscheinlichkeit einer Kollision bei k Elementen ist also:

$$WK = 1 - \frac{1 \cdot (n-1)}{n} \cdot \frac{n-2}{n} \cdot \ldots \cdot \frac{n-k+1}{n}$$

In Produktschreibweise und der Formeln für Kombinationen (ohne Reihenfolge mit Zurücklegen anhand des Kugelmodells: Entnahme von k Kugeln bei n unterscheidbaren):

$$WK = 1 - \prod_{i=1}^{k} \frac{n-k+1}{n} = 1 - \binom{n-k+1}{k}$$

Wir erhalten dann folgende Werte:

- $WK \geq 0.5 \quad \Rightarrow \quad k = 13$
- $WK \geq 0.9 \quad \Rightarrow \quad k = 22$

Abschätzung durch folgende Formel möglich:

$$K(n) \approx \sqrt{-2 \ln(1-p)}$$

5 1.3.2.2.

5.1 Angabe

Zwei Wanderer erreichen aus unterschiedlichen Richtungen einen Aussichtspunkt und halten sich dort 10 Minuten (Wanderer 1) bzw. 20 Minuten (Wanderer 2) auf. Ihre Ankunftszeitpunkte liegen - unabhängig voneinander - zufällig zwischen 11 und 12 Uhr.
(a) Mit welcher Wahrscheinlichkeit begegnen sie einander am Aussichtspunkt?
(b) Wie groß ist die Wahrscheinlichkeit, daß sich um 11:30

1. keiner
2. genau einer
3. beide

am Aussichtspunkt befinden?

5.2 Wichtige Begriffe

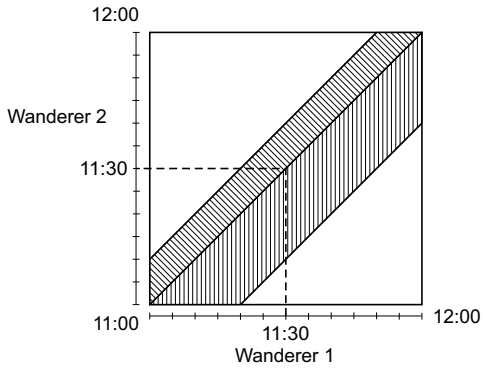
Geometrische Wahrscheinlichkeiten: Merkmalraum $\infty, M \subseteq \mathbb{R}^k$ mit $I(M) < \infty$. Alle Teilräume von M mit gleichem Inhalt sind gleich wahrscheinlich.
Für $A \subseteq M$ definiert man:

$$W(A) := \frac{I(A)}{I(M)}$$

5.3 Lösung der Aufgabenstellung

5.3.1 a(1)

Geometrisch dargestellt wie folgt (günstiger Bereich schraffiert):

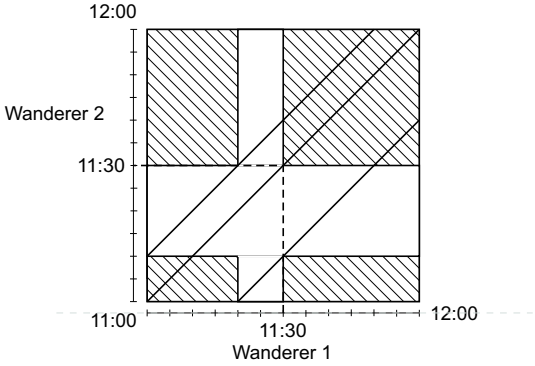


Die Wahrscheinlichkeit wird wie folgt errechnet (in Prozent):

$$W(A) = 100 \cdot \frac{\text{guenstiger Bereich}}{\text{moeglicher Bereich}} = \frac{60 \cdot 60 - (\frac{50 \cdot 50}{2} + \frac{40 \cdot 40}{2})}{60 \cdot 60} = 43.0\dot{5}\%$$

5.3.2 b(1)

Geometrisch dargestellt wie folgt (günstiger Bereich schraffiert):

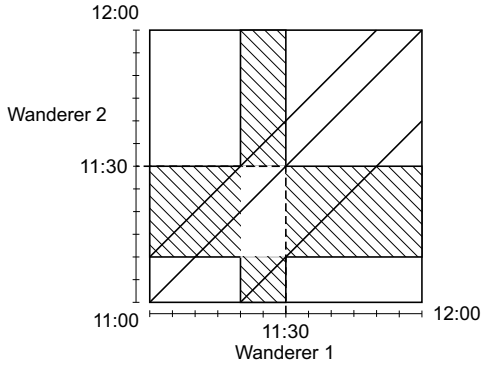


Die Wahrscheinlichkeit wird wie folgt errechnet (in Prozent):

$$W(A) = 100 \cdot \frac{\text{guenstiger Bereich}}{\text{moeglicher Bereich}} = \frac{10 \cdot 20 + 20 \cdot 30 + 30 \cdot 10 + 30 \cdot 30}{60 \cdot 60} = 55.\dot{5}\%$$

5.3.3 b(2)

Geometrisch dargestellt wie folgt (günstiger Bereich schraffiert):

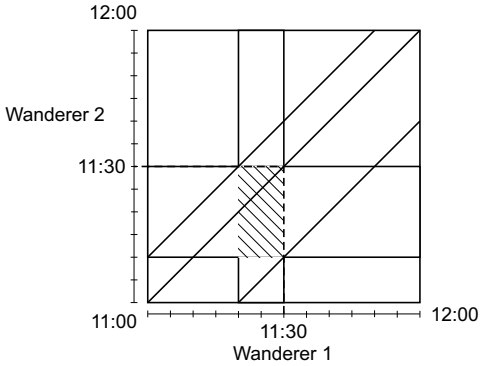


Die Wahrscheinlichkeit wird wie folgt errechnet (in Prozent):

$$W(A) = 100 \cdot \frac{\text{guenstiger Bereich}}{\text{moeglicher Bereich}} = \frac{10 \cdot 10 + 20 \cdot 20 + 30 \cdot 20 + 30 \cdot 10}{60 \cdot 60} = 38.\dot{8}\%$$

5.3.4 b(3)

Geometrisch dargestellt wie folgt (günstiger Bereich schraffiert):



Die Wahrscheinlichkeit wird wie folgt errechnet (in Prozent):

$$W(A) = 100 \cdot \frac{\text{guenstiger Bereich}}{\text{moeglicher Bereich}} = \frac{10 \cdot 20}{60 \cdot 60} = 5.\dot{5}\%$$

6 1.3.2.5.

6.1 Angabe

Bestimmen Sie die Wahrscheinlichkeit, daß die Wurzeln der quadratischen Gleichung $x^2 + 2ax + b = 0$ reell sind, wenn bekannt ist, daß die Koeffizienten mit gleicher Wahrscheinlichkeit aus dem Rechteck $|a| \leq A$, $|b| \leq B$ stammen. Bestimmen Sie unter diesen Bedingungen auch die Wahrscheinlichkeit dafür, daß die Wurzeln beide positiv sind.

6.2 Wichtige Begriffe

Geometrische Wahrscheinlichkeit, siehe 1.3.2.2.!

6.3 Lösung der Aufgabenstellung

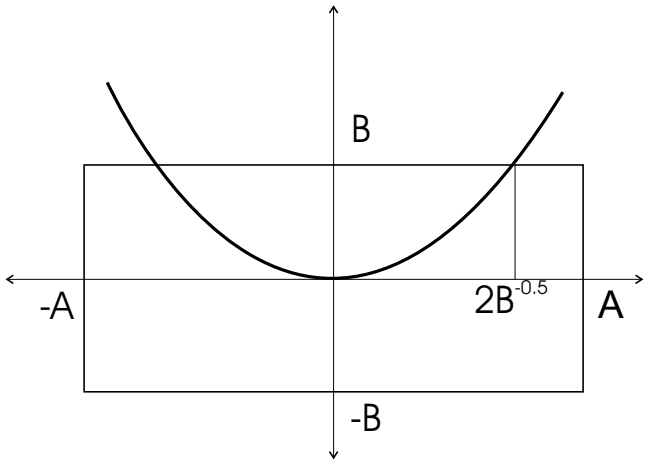
Gehe von $x^2 + 2ax + b = 0$. Lösungen der Gleichung ergeben sich aus:

$$x_{1,2} = -a \pm \sqrt{a^2 - b}$$

Es gilt:

- reell: $a^2 > g$
- gleich (unter Wurzel Null): $a^2 = b$
- komplex: $a^2 - b$

Schematisch graphisch dargestellt:



Wir unterscheiden nun Fall 1 (ein Wert für A) und Fall 2 (ein Wert für B). Die Wahrscheinlichkeit für den reellen Fall beträgt dann:

$$\frac{2AB + \int_0^A \frac{a^2}{4}\, da}{4AB} = \frac{1}{2} + \frac{A^2}{4B}$$

Für den komplexen Fall:

$$\frac{\int_0^A \frac{a^2}{2}\, da}{4AB}$$