Mathematik 3: Übungsblatt - Statistik 1

1. Aufgabe:

Was genau versteht man unter

a) nominalskalierten

b) ordinalskalierten

c) metrisch

skalierten Merkmalen und nennen Sie je mindestens ein Beispiel.

Lösung:

a) Die Nominalskala ist das informationsärmste Skalenniveau. Wir können diese an unterschiedlichen Merkmalen unterscheiden, jedoch **keine Rangfolge** bilden. Z.B. **Geschlecht**, Postleitzahl.

b) Im Gegensatz zu nominalskalierten Daten können ordinalskalierte Daten zwar in eine **natürliche Reihenfolge** gebracht werden – da allerdings die **Abstände zwischen den einzelnen Werten nicht quantifizierbar** sind, kann mit ihnen nicht "normal gerechnet" werden, obwohl es sich auf den ersten Blick um "normale Zahlen" handelt. Das klassische Beispiel hierfür sind **Schulnoten**. Schulnoten weisen sowohl eine natürliche Reihenfolge (eine 1 ist besser als eine 2, eine 2 ist besser als eine 3 usw.) als auch unterschiedliche Abstände zwischen den einzelnen Werten auf (der Notenbereich der 1 umfasst den Bereich von 92% bis 100% der maximal erreichbaren Punkte, der Notenbereich der 5 dagegen den Bereich von 0% bis 49%). Aus diesem Grund sind Rechenoperationen wie etwa das Addieren oder das Subtrahieren und sogar die arithmetische Mittelwertbildung von Noten nicht sinnvoll.

c) Metrisch skalierte Daten verfügen über eine natürliche Reihenfolge sowie auch über **quantifizierbare Abstände** – mit ihnen kann also ganz "normal" gerechnet werden. In vielen Lehrbüchern wird innerhalb der metrischen Skala – die häufig auch als Kardinalskala bezeichnet wird – zusätzlich noch in die Intervallskala (ohne natürlichen Nullpunkt – z.B. **Temperatur** in Celsius) und in die Verhältnisskala (mit natürlichem Nullpunkt – z.B. Temperatur in Kelvin) unterschieden.

2. Aufgabe:

Zwei konkurrierende Produkthersteller A und B verkaufen über das Internet das gleiche Produkt. Einige Zeit nachdem das Produkt geliefert wurde, fragen sie die Kunden nach ihrer Zufriedenheit und sammeln das Ergebnis für jeden Monat. Dabei ergaben sich in zwei aufeinanderfolgenden Monaten folgende Ergebnisse in der Form: zufriedene Kunden (Gesamtzahl der Befragten):

	Monat 1	Monat 2
A	35 (40)	60 (1025)
B	110 (245)	15 (415)

a) Berechnen Sie für beide Hersteller jeweils die relative Häufigkeit der zufriedenen Kunden in den einzelnen Monaten.

b) Hersteller A wirbt nun damit, dass er "Monat für Monat die zufriedensten Kunden" habe. Was kann B auf diese Aussage erwidern?
Ermitteln Sie dazu den Zufriedenheitsrückgang und vergleichen Sie die Ergebnisse.

Lösung:
Relative Häufigkeit: $h_i = \frac{n_i}{n}$

a) Hersteller A:
Monat 1: $h_{A_1} = \frac{35}{40} = 0,875 \Rightarrow$ **87.5 %** der Kunden sind zufrieden
Monat 2: $h_{A_2} = \frac{60}{1025} = 0,0585 \Rightarrow$ **5.85 %** der Kunden sind zufrieden

Hersteller B:
Monat 1: $h_{B_1} = \frac{110}{245} = 0,45 \Rightarrow$ **45 %** der Kunden sind zufrieden
Monat 2: $h_{B_1} = \frac{15}{415} = 0,036 \Rightarrow$ **3.6 %** der Kunden sind zufrieden

b) Rückgang der Kundenzufriedenheit:
Bei A: $1 - \frac{5,85}{87,5} = 1 - 0.067 = 0,933$ also **93.3%**
Bei B: $1 - \frac{3,6}{45} = 1 - 0.08 = 0,92$ also **92%**

B kann sagen, dass bei ihm die Kundenzufriedenheit nicht in dem Maß zurückgegangen ist wie bei A (um **1,3%** weniger).

3. Aufgabe:

Im Rahmen der Messung der Beliebtheit von verschiedenen Artikeln zeichnet der Betreiber einer Webseite auf, wie lange sich Besucher beim Lesen eines bestimmten Artikels aufhielten. Die letzten 10 Besucher liefern ihm die folgenden Daten (in s), wobei Lesezeiten unter 10 s mit dem Wert 0 abgespeichert werden.

0	15	0	315	223	90	45	0	0	247

a) Berechnen Sie Median und Mittelwert dieser Messreihe.

b) Zeichnen Sie die empirische Verteilungsfunktion dieser Messreihe.

c) Der nächste Leser des Artikels verließ während er den Artikel geöffnet hatte aus nicht näher bekannten Gründen seinen PC für eine längere Zeitspanne. Dadurch wird für ihn eine Lesezeit von 4'269 s notiert. Wie verändert dies die statistischen Kennzahlen aus Aufgabenteil a)?

Lösung:

Geordnete Verteilung:

n	1	2	3	4	5	6	7	8	9	10
x_n	0	0	0	0	15	45	90	223	247	315

$\Rightarrow \quad n = 10$

```
> Lesezeit=c(0,15,0,315,223,90,45,0,0,247)
> Lesezeit=sort(Lesezeit)
```

a) Arithmetischer Mittelwert:
$\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_n = \frac{1}{10}\Big(4 \cdot 0 + 15 + 45 + 90 + 223 + 247 + 315\Big) =$ **93.5**

```
> mean(Lesezeit)
[1] 93.5
```

Median da n gerade:
$\Rightarrow \quad \bar{x}_M = \tilde{x} = \frac{1}{2}\Big(x_{\frac{n}{2}} + x_{\frac{n}{2}+1}\Big) = \frac{1}{2}\Big(x_5 + x_6\Big) = \frac{1}{2}\Big(15 + 45\Big) =$ **30**

```
> median(Lesezeit)
[1] 30.0
```

```
> summary(Lesezeit)
Min.   1st Qu.   Median Mean   3rd Qu.   Max.
0.0    0.0       30.0   93.5   189.8     315
```

b) Die Empirische Verteilungsfunktion erhält man, indem man den Datenvektor sortiert, die Summe bildet (`sum(Lesezeit)` → `[1] 935`) und dann die jeweiligen Datenpunkte (geteilt durch 935) nacheinander aufträgt.

Lösung in R:

```
> edf(Lesezeit)
```

Die Funktion `edf` muss man sich selbst erstellen.
Siehe Anhang des Übungsblattes.

c) Erweiterung der geordneten Tabelle:

n	1	2	3	4	5	6	7	8	9	10	11
x_n	0	0	0	0	15	45	90	223	247	315	4269

$\Rightarrow \; n = 11$

```
> Lesezeit=c(0,0,0,0,15,45,90,223,247,315,4269)
> Lesezeit=sort(Lesezeit)
```

Arithmetischer Mittelwert:

$$\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_n = \frac{1}{11}\Big(4 \cdot 0 + 15 + 45 + 90 + 223 + 247 + 315 + 4269\Big) \approx \mathbf{473.1}$$

```
> mean(Lesezeit)
[1] 473.0909
```

Median da n ungerade:

$\Rightarrow \; \bar{x}_M = \tilde{x} = x_{\frac{n+1}{2}} = x_{\frac{12}{2}} = x_6 = \mathbf{45}$

```
> median(Lesezeit)
[1] 45
> summary(Lesezeit)
Min.   1st Qu.   Median Mean   3rd Qu.   Max.
0.0    0.0       45.0   473.1  235.0     4269.0
```

4. Aufgabe:

Bei einer Mathematikarbeit ergibt sich die nachfolgende tabellierte Notenverteilung

Note	1	2	3	4	5	6
Schüler/innen	3	7	11	8	2	1

Konstruieren Sie einen Box-Plot für die Verteilung und benennen Sie alle für die Konstruktion benötigten Größen.

Geordnete Verteilung:

n	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24	25	26	27	28	29	30	31	32
x_n	1	1	1	2	2	2	2	2	2	2	3	3	3	3	3	3	3	3	3	3	3	4	4	4	4	4	4	4	4	5	5	6

$\Rightarrow \; n = 32$

Erforderliche Größen:

Unteres Quantil: $n \cdot p = 32 \cdot 0.25 = 8 \quad \Rightarrow \; x_8 = \mathbf{2}$

Oberes Quantil: $n \cdot q = 32 \cdot 0.75 = 24 \; \Rightarrow \; x_{24} = \mathbf{4}$

Median (Zentralwert) da n gerade:

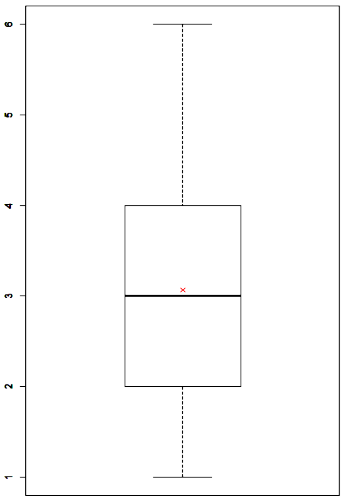
$$\Rightarrow \; \bar{x}_M = \tilde{x} = \frac{1}{2}\left(x_{\frac{n}{2}} + x_{\frac{n}{2}+1}\right) = \frac{1}{2}\left(x_{16} + x_{17}\right) = \frac{1}{2}(3+3) = \mathbf{3}$$

Arithmetischer Mittelwert: $\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_n = \frac{1}{32}\left(3 \cdot 1 + 7 \cdot 2 + 11 \cdot 3 + 8 \cdot 4 + 5 \cdot 2 + 6 \cdot 1\right) \approx \mathbf{3.07}$

Lösung in R:

```
> Schuelernoten=c(1,1,1,2,2,2,2,2,2,2,
3,3,3,3,3,3,3,3,3,3,4,4,4,4,4,4,4,4,5,5,6)
> boxplot(Schuelernoten)
> points(mean(Schuelernoten),pch=4,col="red")

> summary(Schuelernoten)
Min.   1st Qu.   Median Mean   3rd Qu.   Max.
1.000 2.000    3.000  3.065 4.000    6.000
```



5. Aufgabe:

Die Bestimmung der Körpergröße von 11 Personen erbringt folgende Ergebnisse:

1.62 m	1.71 m	1.61 m	1.82 m	1.75 m	1.77 m	1.82 m	1.55 m	1.74 m	1.63 m	1.83 m

Bestimmen Sie folgende Größen:

a) arithmetisches Mittel

b) Median

c) Varianz

d) Standardabweichung

Geordnete Verteilung:

n	1	2	3	4	5	6	7	8	9	10	11
x_n	1.55	1.61	1.62	1.63	1.71	1.74	1.75	1.77	1.82	1.82	1.83

$\Rightarrow\ n = 11$

```
> Groesse=c(1.55,1.61,1.62,1.63,1.71,1.74,1.75,1.77,1.82,1.82,1.83)
```

a) Arithmetischer Mittelwert:

$\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_n = \frac{1}{11}\left(1.62+1.71+1.61+1.82+1.75+1.77+1.82+1.55+1.74+1.63+1.83\right) \approx$ **1.714**

```
> mean(Groesse)
[1] 1.713636
```

b) Median da n ungerade:

$\Rightarrow\ \bar{x}_M = \tilde{x} = x_{(\lfloor np+1 \rfloor)} = x_{(\lfloor \frac{11}{2}+1 \rfloor)} = x_{(\underbrace{\lfloor (5.5+1 \rfloor}_{\text{6.5 abrunden}})} = x_6 = \mathbf{1.74}$

```
> summary(Groesse)
Min.  1st Qu.  Median Mean  3rd Qu.  Max.
1.550 1.625    1.740  1.714 1.795    1.830
```

(auch der direkte Befehl `median(Groesse)` möglich)

c) Varianz: $\text{VAR} = s^2 = \frac{1}{n-1}\left(\sum_{i=1}^{n} x_i^2 - n\cdot\bar{x}^2\right) = \frac{1}{10}\left((1.55)^2 - 11\cdot(1.714)^2 + ...\right) \approx \mathbf{0.00948}$

```
> var(Groesse)
[1] 0.009465455
```

d) Standardabweichung: $s = \sqrt{\text{VAR}} = \sqrt{0.009465455} = \mathbf{0.09729057}$

```
> sqrt(var(Groesse))
[1] 0.09729057
```

(auch der direkte Befehl `sd(Groesse)` möglich)

6. Aufgabe:

Kurz nach Semesterstart liefert die Suche nach einem WG-Zimmer in Friedrichshafen folgende Preise für die Gesamtmiete eines Zimmers (in €):

435	400	400	325	410	420	525	385	440	370
352	365	530	432	358	300	385	370	355	300

a) Was ist der Median-Preis für die Miete eines WG-Zimmers in dieser Messreihe?

b) Was ist das arithmetische Mittel der Messreihe?

c) Betrachten Sie die Klassen

$[0, 200)$	$[200, 300)$	$[300, 350)$	$[350, 400)$	$[400, 500)$	$[500, 800]$

Erstellen Sie ein Histogramm der Messreihe bezüglich der Klasseneinteilung. Ist die gegebene Klasseneinteilung sinnvoll?

Geordnete Verteilung:

n	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20
x_n	300	300	325	352	355	358	365	370	370	385	385	400	400	410	420	432	435	440	525	530

$\Rightarrow n = 20$

```
> Preise=c(300,325,352,355,358,365,370,370,385,385,400,400,
410,420,432,435,440,525,530)
```

a) Median da n gerade:

$\Rightarrow \bar{x}_M = \tilde{x} = \frac{1}{2}\left(x_{\frac{n}{2}} + x_{\frac{n}{2}+1}\right) = \frac{1}{2}\left(x_{10} + x_{11}\right) = \frac{1}{2}(385 + 385) = \mathbf{385}$

```
> summary(Preise)
Min.   1st Qu.  Median Mean   3rd Qu.  Max.
300.0 361.5    385.0  397.7 426.0    530.0
```

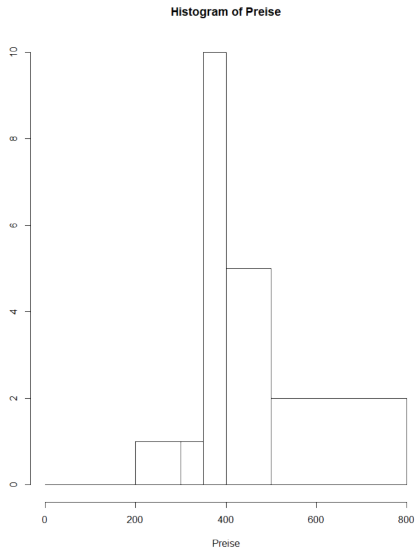
(auch der direkte Befehl `median(Preise)` möglich)

b) Arithmetischer Mittelwert: $\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_n = \frac{1}{20}\left(2 \cdot 300 + 325 + 352 + ...\right) \approx \mathbf{397.74}$

```
> mean(Preise)
[1] 397.7368
```

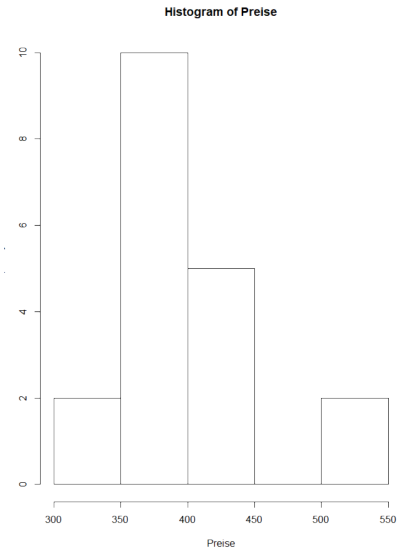
c) Für die Anzahl der Bereiche N gilt die Faustformel aus dem Skript:

$N_{Bereiche} = \sqrt{n} = \sqrt{20} \approx 4,4 \;\Rightarrow\; N_{Bereiche} = \mathbf{5}$




```
> hist(Preise,breaks=c(0,200,300,
350,400,500,800),freq = TRUE)
```

Die Aufteilung sollte bei 300 starten und in 50er Schritten bis 550 gehen. Aufteilungen mit unterschiedlichen Abständen sollten vermieden werden!




```
> hist(Preise)
```

Anhang:

Die Funktion edf (plottet eine empirische Verteilungsfunktion)

```
> Lesezeit=c(0,0,0,0,15,45,90,223,247,315)
> edf(Lesezeit)
```

```
edf <- function(x = NULL, Norm=T, xlab = "", ylab = "", ...)
{
   # Aufgabe:  Die Funktion edf zeichnet die empirische Verteilungsfunktion
   # des Datenvektors x.
   # Inputparameter:
   # x ...  Datenvektor mit den Beobachtungswerten
   # Ergebnis:
   # Graphische Darstellung der empirischen Verteilungsfunktion
   # samt einer angepassten Normalverteilung

    x <- sort(x)
    y <- c(0, cumsum(table(x))/length(x), 1)
    x.min <- floor(min(x) - (max(x) - min(x)) * 0.05)
    x.max <- ceiling(max(x) + (max(x) - min(x)) * 0.05)
    plot(c(x.min, unique(x), x.max), y, type = "s",xlab = xlab,
         ylab = ylab, xlim = c(x.min, x.max), ylim = c(0, 1),
         axes = F, ...)
    if (Norm) lines(c(x.min, unique(x), x.max),
              pnorm(c(x.min, unique(x), x.max),
              mean(x), sqrt(var(x))))
    axis(1, pos = 0)
    axis(2, pos = x.min)
    invisible()
}
```