Lösung 16.3

Analog zu Aufgabe 16.1 werden die Daten durch folgenden Befehl eingelesen:

```
> kredit<-read.table("c:\\compaufg\\kredit.txt",header=T)
```

Der Datensatz ist nun als Data Frame Objekt `kredit` in S-Plus verfügbar.

(a)

Säulendiagramme der diskreten Variablen *boni*, *moral*, *zweck*, *geschl*, *famst* und *konto* erhält man durch folgende Aufrufe (vgl. Aufgabe 16.1 (a)):

```
> par(mfrow=c(3,2))
> hist(factor(boni),prob=T,xlab="Bonitaet",ylim=c(0,1))
> hist(factor(moral),prob=T,xlab="Zahlungsmoral",ylim=c(0,1))
> hist(factor(zweck),prob=T,xlab="Verwendungszweck",ylim=c(0,1))
> hist(factor(geschl),prob=T,xlab="Geschlecht",ylim=c(0,1))
> hist(factor(famst),prob=T,xlab="Familienstand",ylim=c(0,1))
> hist(factor(konto),prob=T,xlab="Laufendes Konto",ylim=c(0,1))
```

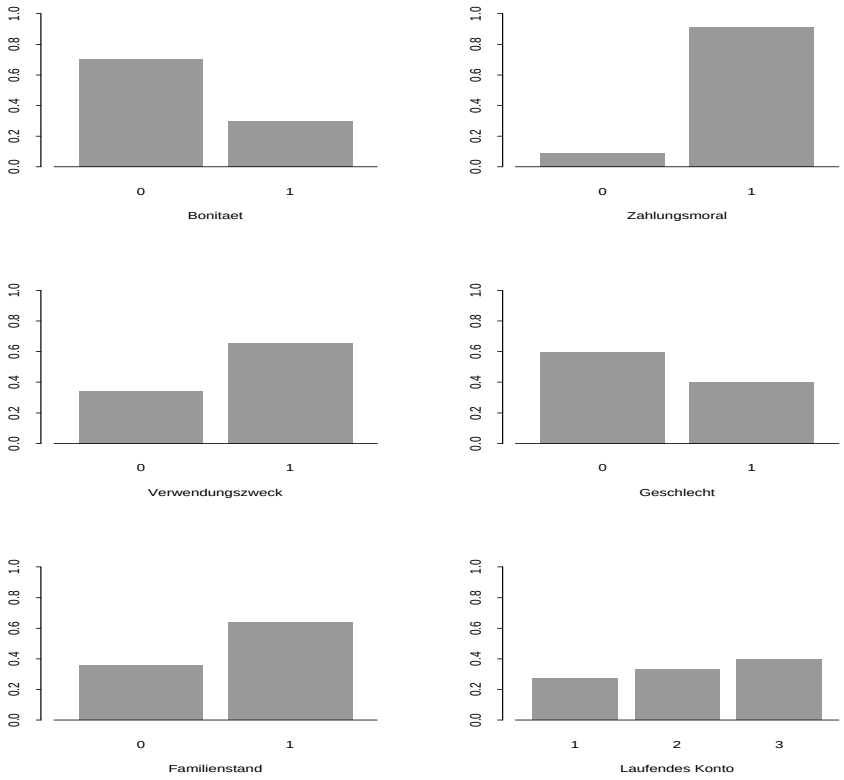

Abbildung 1: Graphische Veranschaulichung für die Variablen *boni*, *moral*, *zweck* *geschl*, *famst*, *konto*

Bei der Interpretation der Säulendiagramme ist Vorsicht geboten: Es handelt sich um eine geschichtete Stichprobe, in der 30 % der Kredite nicht

zurückgezahlt wurden. Dies spiegelt sich exakt im Säulendiagramm für die Bonität wider.
Für die stetigen Variablen *laufz* und *hoehe* bieten sich das Histogramm, der Box-Plot und der Kerndichteschätzer an. Die Funktionsaufrufe lauten (vgl. Aufgabe 16.1 (a)):

```
> par(mfrow=c(3,2))
> hist(hoehe,ylab="Anteil",xlab="Hoehe",prob=T,ylim=c(0,1))
> hist(laufz,ylab="Anteil",xlab="Laufzeit",ylim=c(0,1))
> boxplot(hoehe,ylab="Hoehe")
> boxplot(laufz,ylab="Laufzeit")
> plot(density(hoehe),type="l",xlab="Hoehe",ylab="Anteil")
> plot(density(laufz,width=10),type="l",xlab="Laufzeit",ylab="Anteil")
```

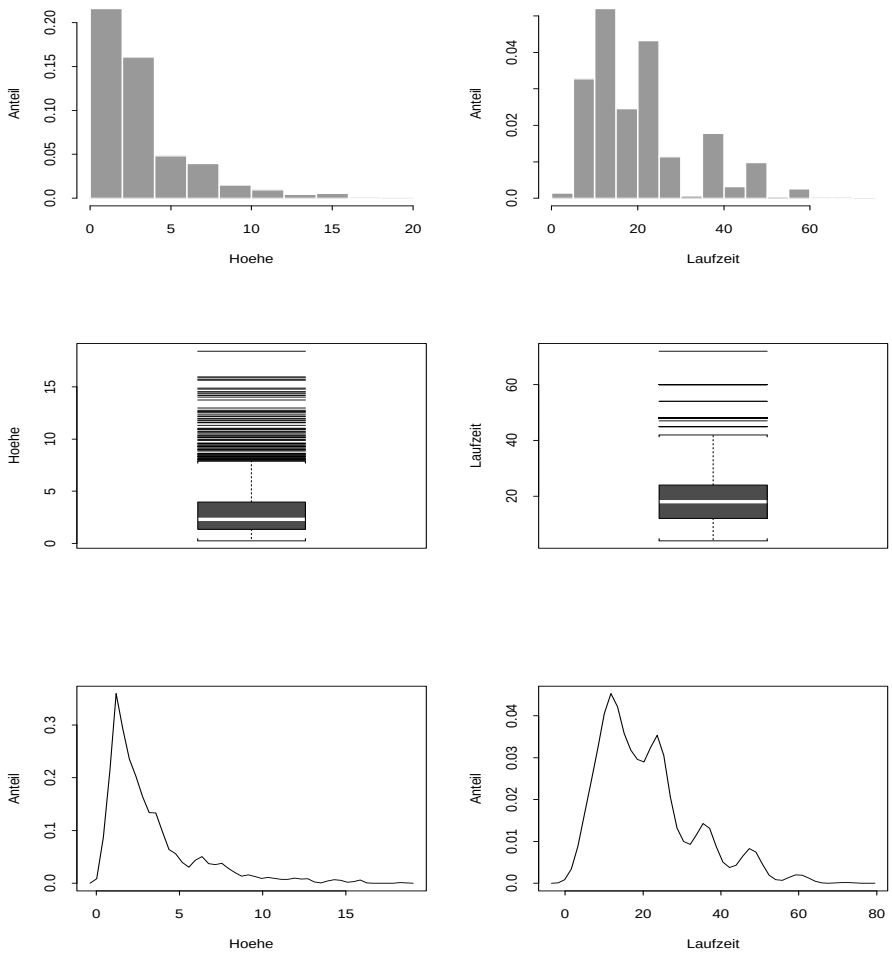

Abbildung 2: Graphische Veranschaulichung für die Variablen *hoehe*, *laufz*

Die Kredithöhe besitzt somit eine extrem linkssteile Verteilung, was bei der Laufzeit nicht so stark ausgeprägt ist. Dies zeigt sich auch an den Ergeb-

nissen aus Teilaufgabe (b). So sind bei der Kredithöhe, im Gegensatz zur Laufzeit, Median und Mittelwert deutlich verschieden.

(b)

Für die stetigen Merkmale *hoehe* und $laufz$ erhält man durch die `summary()`-Funktion die 5-Punkte-Zusammenfassung (vgl. Aufgabe 16.1 (b)). Die Funktionsaufrufe lauten:

```
> summary(hoehe)
> summary(laufz)
```

Quantile berechnet man mit der `quantile()`-Funktion. Zur Berechnung der Varianz und Standardabweichung verwendet man die Funktionen `var()` und `sqrt()`. Die entsprechenden Aufrufe lauten (vgl. Aufgabe 16.1 (b)):

```
>  quantile(hoehe,c(0.05,0.1,0.9,0.95))
>  quantile(laufz,c(0.05,0.1,0.9,0.95))
>  var(hoehe)
>  var(laufz)
>  sqrt(var(hoehe))
>  sqrt(var(laufz))
```

Die folgende Tabelle fasst diese Ergebnisse zusammen:

	hoehe	$laufz$
Minimum	0.25	4.0
5 %-Quantil	0.71	6.0
10 %-Quantil	0.93	9.0
25 %-Quantil	1.37	12.0
Median	2.32	18.0
Mittelwert	3.27	20.9
75 %-Quantil	3.97	24.0
90 %-Quantil	7.18	36.0
95 %-Quantil	9.16	48.0
Maximum	18.42	250.0
Varianz	7.97	145.4
Standardabw.	2.82	72.0

Die Aufrufe für die binären Merkmale *boni*, *moral*, *zweck*, *geschl* und $famst$ lauten (vgl. Aufgabe 16.1 (b)):

```
> summary(factor(boni))/1000
> summary(factor(moral))/1000
> summary(factor(zweck))/1000
> summary(factor(geschl))/1000
> summary(factor(famst))/1000
> summary(factor(konto))/1000
```

Folgende Tabelle fasst die Ergebnisse der binären Merkmale zusammen:

	boni	*moral*	*zweck*	*geschl*	$famst$
0	0.70	0.09	0.34	0.60	0.36
1	0.30	0.91	0.66	0.40	0.64

Für das dreikategoriale Merkmale *konto* ergibt sich folgendes Ergebnis:

	konto
1	0.274
2	0.332
3	0.394

(c)
Kontingenztafeln erhält man durch die `table()`-Funktion. Den entsprechenden Unabhängigkeitstest durch die Funktion `chisq.test`. (Bemerkung: Der optionale Parameter `correct=F` schaltet dabei Yate's Stetigkeitskorrektur für 2×2-Tafeln aus.) :

```
> table(factor(boni),factor(moral))
     0   1
  0 36  664
  1 53  247
> chisq.test(factor(boni),factor(moral),correct=F)
    Pearson's chi-square test without Yates' continuity correction

  data:  factor(boni) and factor(moral)
  X-square = 40.6241, df = 1, p-value = 0
.............................................................
> table(factor(boni),factor(zweck))
     0   1
  0 215 485
  1 128 172
> chisq.test(factor(boni),factor(zweck),correct=F)
    Pearson's chi-square test without Yates' continuity correction

  data:  factor(boni) and factor(zweck)
  X-square = 13.3128, df = 1, p-value = 0.0003
.............................................................
> table(factor(boni),factor(geschl))
     0   1
  0 432 268
  1 166 134
> chisq.test(factor(boni),factor(geschl),correct=F)
    Pearson's chi-square test without Yates' continuity correction

  data:  factor(boni) and factor(geschl)
  X-square = 3.5568, df = 1, p-value = 0.0593
.............................................................
> table(factor(boni),factor(famst))
     0   1
  0 231 469
  1 129 171
> chisq.test(factor(boni),factor(famst),correct=F)
    Pearson's chi-square test without Yates' continuity correction

  data:  factor(boni) and factor(famst)
  X-square = 9.1146, df = 1, p-value = 0.0025
.............................................................
```

```
> table(factor(boni),factor(konto))
     1   2   3
  0 139 213 348
  1 135 119  46
> chisq.test(factor(boni),factor(konto),correct=F)
    Pearson's chi-square test without Yates' continuity correction

  data:  factor(boni) and factor(konto)
  X-square = 116.8513, df = 2, p-value = 0
```

Zusammengefasst:

Test zwischen *boni* und ...	χ^2-Wert	p-Wert
moral	40.62	0.0000
zweck	13.31	0.0003
geschl	3.56	0.0593
famst	9.11	0.0025
konto	116.85	0.0000

Die p-Werte der Unabhängigkeitstests zwischen *boni* und den Variablen *moral*, *zweck*, *famst* und *konto* sind alle kleiner als 1 %, so daß die Hypothese der Unabhängigkeit zwischen den entsprechenden Merkmalen abgelehnt werden muß. Lediglich der Unabhängigkeitstest zwischen *boni* und *geschl* mit einem p-Wert von 5.9 % deutet darauf hin, daß zwischen diesen beiden Merkmalen keine signifikante Abhängigkeit besteht.

(d)

Die bedingten Häufigkeitsverteilungen entnimmt man am besten der Tafel, die man durch die `crosstabs()`-Funktion erhält. Anhand der Variable *geschl* wird im folgenden der daraus resultierende Output erklärt:

```
> crosstabs(~factor(boni)+factor(geschl))
Call:
crosstabs( ~ factor(boni) + factor(geschl))
1000 cases in
table
+----------+
|N         |
|N/RowTotal|
|N/ColTotal|
|N/Total   |
+----------+
factor(boni)|factor(geschl)
            |0      |1      |RowTotl|
  ----------+-------+-------+-------+
  0         |432    |268    |700    |
            |0.62   |0.38   |0.7    |
            |0.72   |0.67   |       |
            |0.43   |0.27   |       |
  ----------+-------+-------+-------+
  1         |166    |134    |300    |
            |0.55   |0.45   |0.3    |
            |0.28   |0.33   |       |
            |0.17   |0.13   |       |
  ----------+-------+-------+-------+
  ColTotl   |598    |402    |1000   |
            |0.6    |0.4    |       |
  ----------+-------+-------+-------+
Test for independence of all factors
    Chi^2 = 3.55683 d.f.= 1 (p=0.0593009)
    Yates' correction not used
```

Unter der Tafel sind die Ergebnisse eines Unabhängigkeitstests, analog zu Teilaufgabe (c), ausgegeben. In der Spalte `RowTotl` ist dabei die Randverteilung von *boni*, in der Zeile `ColTotl` die Randverteilung von *geschl* mit den absoluten und relativen Häufigkeiten gegeben. Innerhalb der Zellen sind von oben nach unten folgende Ergebnisse aufgelistet:

- 1. Zeile: absolute Häufigkeit der Kombination (*boni*,*geschl*)
- 2. Zeile: bedingte relative Häufigkeit von *geschl* gegeben *boni*
- 3. Zeile: bedingte relative Häufigkeit von *boni* gegeben *geschl*
- 4. Zeile: relative Häufigkeit der Kombination (*boni*,*geschl*)

Aus der oben berechneten Tafel, die demnach in der 2. Zeile jeder Zelle die für die Fragestellung benötigten Ergebnisse enthält, lassen sich also folgende zwei bedingte Häufigkeitsverteilungen von *geschl* gegeben *boni* entnehmen:

	rel. Häufigkeit bei ... Krediten	
	guten ($boni = 0$)	schlechten ($boni = 1$)
Frauen ($geschl = 0$)	0.62	0.55
Männer ($geschl = 1$)	0.38	0.45
Summe	1.00	1.00

Da sich die bedingten Verteilungen kaum unterscheiden, scheint das Geschlecht die Bonität eines Kunden also nicht stark zu beeinflussen, was auch durch den entsprechenden Unabhängigkeitstest, dessen p-Wert auf Unabhängigkeit schließen läßt, bestätig wird.

Für die bedingte Häufigkeitsverteilung von $famst$ gegeben *boni* ergab sich folgendes Ergebnis:

	rel. Häufigkeit bei ... Krediten	
	guten ($boni = 0$)	schlechten ($boni = 1$)
ledig ($famst = 0$)	0.33	0.43
verheiratet($famst = 1$)	0.67	0.57
Summe	1.00	1.00

Für diese Variablen unterscheiden sich die bedingten Verteilungen, so daß also von einer Abhängigkeit ausgegangen worden kann. Dies wird auch durch das Ergebnis des Unabhängigkeitstests aus Teilaufgabe (c) bestätigt. Da $famst = 0$ bei den schlechten Krediten stärker vertreten ist, scheint es, als ob Ledige ihre Kredite schlechter zurückzahlen.

Für die bedingte Häufigkeitsverteilung von *zweck* gegeben *boni* ergab sich folgendes Ergebnis:

	rel. Häufigkeit bei ... Krediten	
	guten ($boni = 0$)	schlechten ($boni = 1$)
Geschäftlich ($zweck = 0$)	0.31	0.43
Privat ($zweck = 1$)	0.69	0.57
Summe	1.00	1.00

Auch hier unterscheiden sich die Verteilungen so stark, daß man von Abhängigkeit ausgehen kann. Dabei scheint es so, daß Geschäftskredite die Bonität des Kunden mindern, da der Anteil der Geschäftskredite an den guten Krediten kleiner ist als an den schlechten Krediten.

Für die bedingte Häufigkeitsverteilung von *moral* gegeben *boni* ergab sich folgendes Ergebnis:

	rel. Häufigkeit bei ... Krediten	
	guten ($boni = 0$)	schlechten ($boni = 1$)
Schlechte Moral ($moral = 0$)	0.05	0.18
Gute Moral ($moral = 1$)	0.95	0.82
Summe	1.00	1.00

Daraus wird ganz klar ersichtlich, daß eine gute Zahlungsmoral in positivem Zusammenhang zur Bonität steht, da der Anteil der Kreditnehmer mit schlechter Zahlungsmoral bei den schlechten Krediten fast viermal so hoch ist wie bei den guten Kreditnehmern.

Für die bedingte Häufigkeitsverteilung von *konto* gegeben *boni* ergab sich folgendes Ergebnis:

	rel. Häufigkeit bei ... Krediten	
	guten ($boni = 0$)	schlechten ($boni = 1$)
Kein Konto ($konto = 1$)	0.20	0.45
Schlechtes Konto ($konto = 2$)	0.30	0.40
Gutes Konto ($konto = 3$)	0.50	0.15
Summe	1.00	1.00

Auch hier ist der Zusammenhang deutlich zu erkennen; ein gutes laufendes Konto wirkt sich positiv auf die Bonität aus.

(e)

Die enstprechenden Korrelationskoeffizienten erhält man mit der `cor()` - Funktion:

```
> cor(boni,hoehe)
    0.1547401
> cor(boni,laufz)
    0.2149267
```

Beide Korrelationskoeffizienten deuten auf eine schwache Korrelation hin.