



3 Statistische Graphik

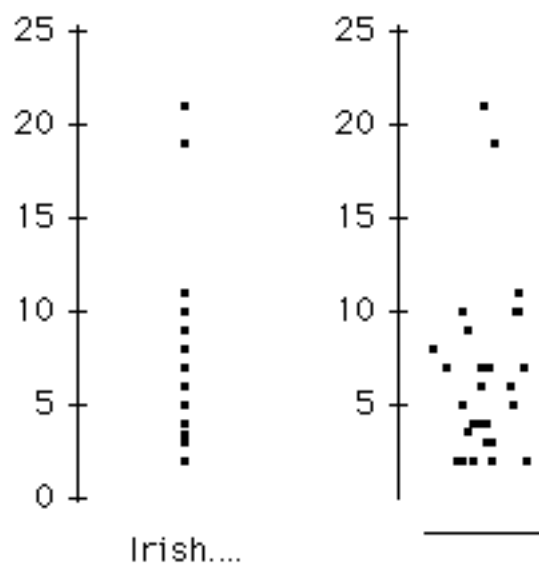
3.1 Graphische Darstellungen von stetigen Variablen

3.1.1 Punktplots („Dotplots“)

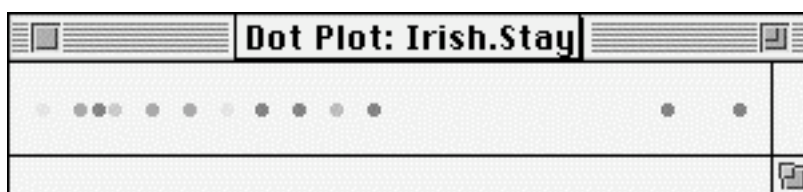
Mit Dotplots kann man kleinere Datensätze gut zusammenfassen und einzelne Fälle identifizieren. Lücken werden hervorgehoben (auch bei größeren Datensätzen).

Alle Fälle sollen einzeln aufgetragen werden. Mehrfache Fälle werden in Software oft nicht dargestellt (das Bild links).

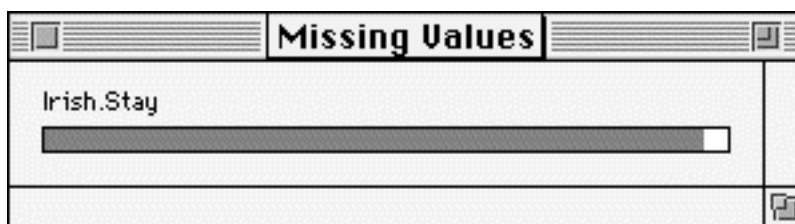
Eine Möglichkeit ist, die mehrfachen Punkte nebeneinander zu zeichnen. Eine andere Möglichkeit ist „jittering“, d.h. jeden Punkt zufällig nach rechts oder links zu verschieben (das Bild rechts).



Eine andere Möglichkeit die mehrfachen Punkte zu berücksichtigen, ist es, diese heller darzustellen:
(in MANET wird eine weitere Graphik automatisch erstellt, wenn es Fälle mit fehlenden Werten gibt):



MANET

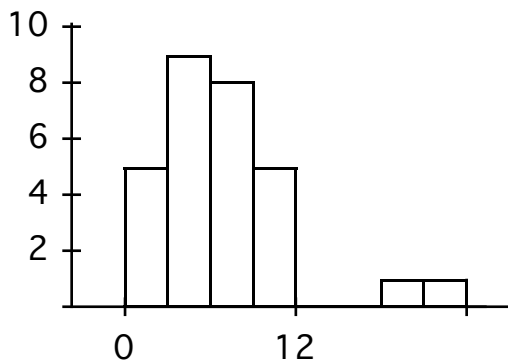


Wenn 1 Punkt gezeigt werden soll, wird er schwarz gezeichnet (wie normal).

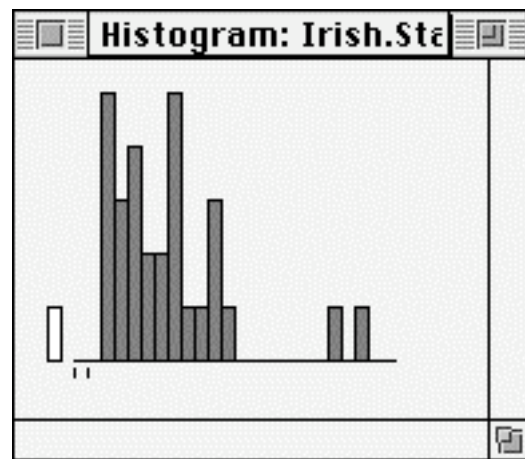
Wenn k Punkte an derselben Bildschirmstelle gezeigt werden sollen, wird diese Stelle mit einer Helligkeit gezeichnet. L ist ein Parameter des Plots. $k = 1$ wird schwarz und $k \geq L$ wird weiß.

3.1.2 Histogramme

Die horizontale Achse wird in Klassen eingeteilt. Für jede Klasse wird die relative Häufigkeit als Rechteck über der Klasse aufgetragen.



Data Desk



MANET

In MANET wird eine weitere Säule für fehlende Werte gezeichnet (in weiß auf der linken Seite).

Üblicherweise sind alle Klassen gleich breit. Dann stellt die Höhe die relative Häufigkeit dar. Im allgemeineren (seltenen) Fall sind die Klassenintervalle unterschiedlich: Dann stellt die Fläche die relative Häufigkeit dar. Da es schwierig ist, Flächen zu vergleichen, werden gleich breite Klassenintervalle bevorzugt.



Histogramme sind graphische Darstellungen von Häufigkeitstabellen für stetige Variablen. Sie geben eine Übersicht einer Verteilungsform.

Hauptparameter: Anfangspunkt („Ankerpunkt“)

 Klassenbreite

 Ziele: Genug Klassen, aber nicht zuviele

 Sinnvolle Klassengrenzen

Formatierungsparameter:

 Skalierung

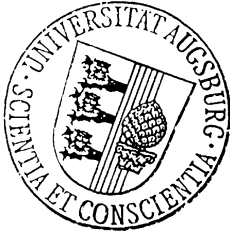
 Größe

 Aspekt

 Beschriftung

 Legende

 Farbe/Schattierung



Heuristische Regeln für die Bestimmung der Anzahl der Klassen bzw. der Klassenbreite.

n = Größe des Datensatzes

s = Standardabweichung des Datensatzes

IQ = InterQuartilAbstand

(a) $1 + \log_2 n$ # Klassen (Sturges)

(b) $3.5 * s * n^{(-1/3)}$ Klassenbreite (Scott)

(c) Klassenbreite

Keine Heuristik wird alle Datensätze gut einteilen.

Beispiele:

(1) $n = 100$ Min = 0.4 Max = 9.5

(2) $n = 100$ Min = 0.4 Max = 10.1

(3) $n = 101$ $x_{(1)} = 0.4$ (Min) $x_{(100)} =$ $x_{(101)} =$ (Max)



Statt zu hoffen, daß ein Histogramm alle Informationen darstellen könnte, ist es besser, sich mehrere Histogramme der selben Daten anzuschauen. Data Desk, R und MONDRIAN bieten verschiedene Möglichkeiten dazu.

(1) Data Desk

Vergrößert/verkleinert das Fenster beim Drücken der Option-Taste. Man weiß nie, wie es ausfallen wird.

(2) R

Parameter müssen im "hist" Befehl gesetzt werden.

(3) MONDRIAN

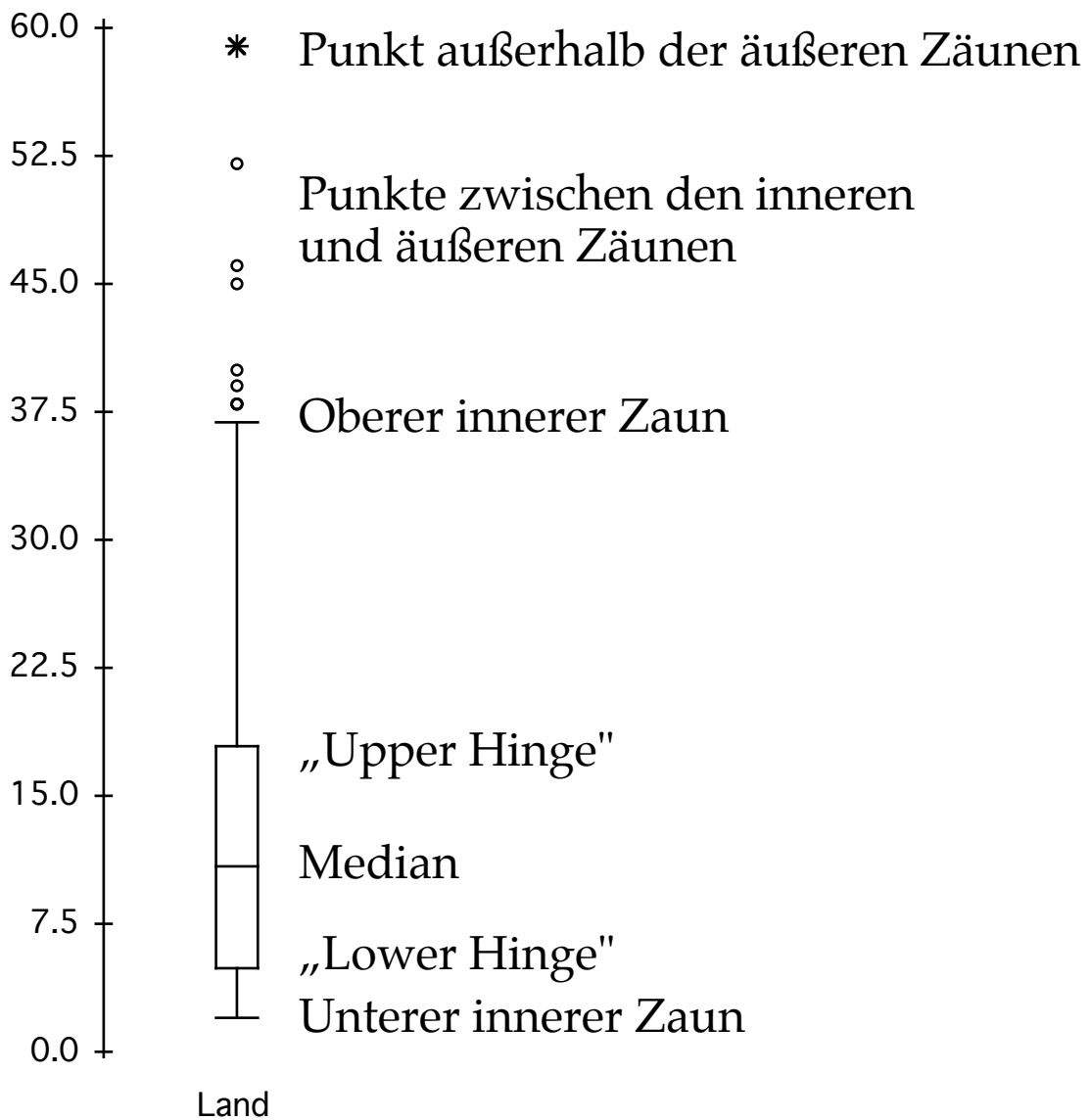
Die Klassenanzahl wird mit den Pfeiltasten auf- und abgestuft. Die vertikale Skalierung muß nachher immer neu gesetzt werden. Parameter können auch über ein Dialogfenster gesetzt werden.



3.1.3 Boxplots

Eine klassische Definition (leider gibt es viele andere).

z.B. Landwirtschaftliche Arbeiter in deutschen Wahlkreisen





Gegeben sind die Daten $\{x_1, x_2, \dots, x_n\}$
oder in aufsteigender Reihenfolge $\{x_{(i)}\}$

Ein Boxplot basiert auf den Rängen der Daten:

der Median (der mittlere Wert des Datensatzes)

$$M = x_{(k+1)} \quad \text{für } n = 2k + 1$$

$$M = (x_{(k)} + x_{(k+1)})/2 \quad \text{für } n = 2k$$

die obere und untere Hinges (fast wie Quartile)

F_o = der Median von $\{x_{(k+1)}, \dots, x_{(n)}\}$

F_u = der Median von $\{x_{(1)}, \dots, x_{(k)}\}$ für $n = 2k$

der Median von $\{x_{(1)}, \dots, x_{(k+1)}\}$ für $n = 2k + 1$

Die inneren Zäune sind die extremsten Werte innerhalb

Die äußeren Zäune sind

Punkte zwischen den inneren und äußeren Zäunen werden einzeln mit 'o' dargestellt („Ausreißer“)

Punkte außerhalb der äußeren Zäune werden einzeln mit '*' dargestellt („krasse Ausreißer“)



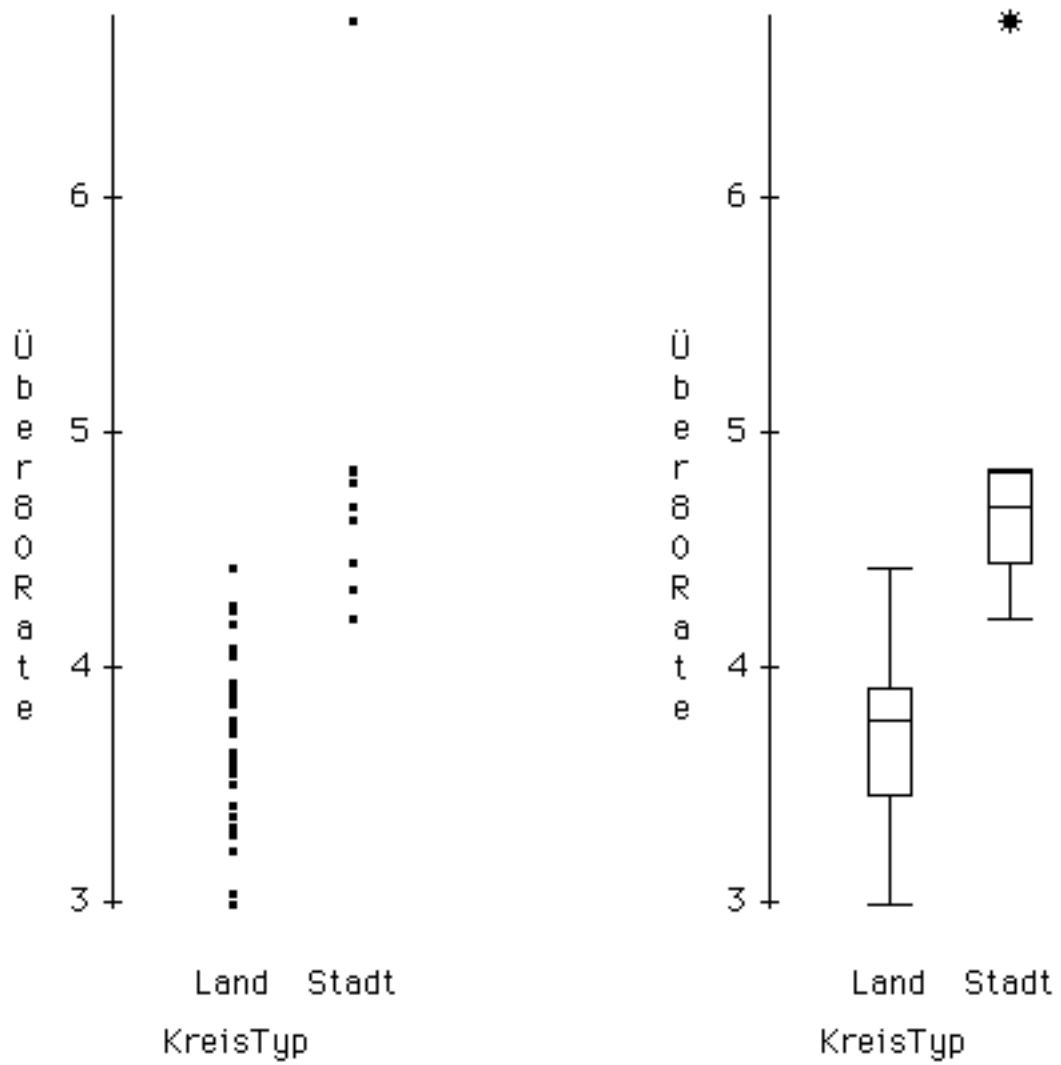
Bevölkerungsanteile über 80, Baden-Württemberg, 1993

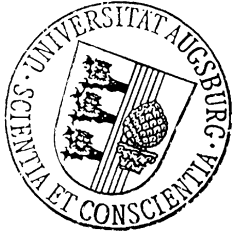
Landkreise		Stadtkreise	
Name	% über 80	Name	% über 80
Böblingen	2.98	Stuttgart Landeshauptstadt	4.79
Esslingen	3.57	Stadt Heilbronn	4.34
Göppingen	4.08	Stadt Baden-Baden	6.75
Ludwigsburg	3.31	Stadt Karlsruhe	4.83
Rems-Murr-Kreis	3.55	Stadt Heidelberg	4.63
Heilbronn	3.32	Stadt Mannheim	4.21
Hohenlohekreis	3.64	Stadt Pforzheim	4.85
Schwäbisch-Hall	3.85	Stadt Freiburg im Breisgau	4.69
Main-Tauber-Kreis	4.26	Stadt Ulm	4.45
Heidenheim	3.90		
Ostalbkreis	3.76		
Karlsruhe	3.22		
Rastatt	3.62		
Neckar-Odenwald-Kreis	3.86		
Rhein-Neckar-Kreis	3.50		
Calw	3.78		
Enzkreis	3.32		
Freudenstadt	4.05		
Breisgau-Hochschwarzwald	3.72		
Emmendingen	3.60		
Ortenaukreis	3.92		
Rottweil	4.24		
Schwarzwald-Baar-Kreis	4.27		
Tuttlingen	3.88		
Konstanz	4.43		
Lörrach	3.90		
Waldshut	3.87		
Reutlingen	3.94		
Tübingen	3.04		
Zollernalbkreis	3.88		
Alb-Donau-Kreis	3.29		
Biberach	3.41		
Bodenseekreis	4.19		
Ravensburg	3.88		
Sigmaringen	3.37		

(Source: Statistisches Landesamt Baden-Württemberg)



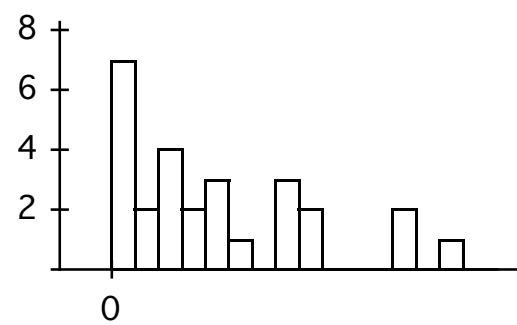
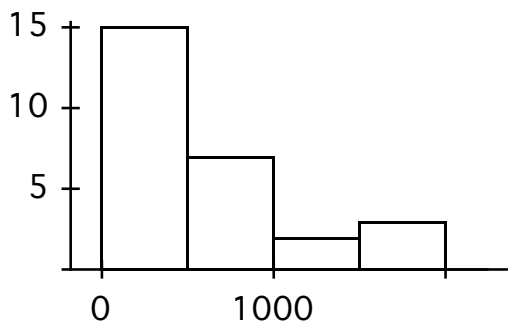
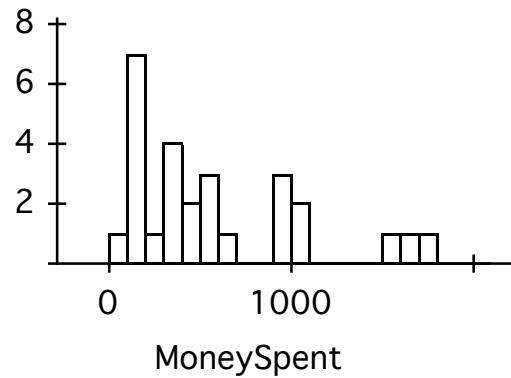
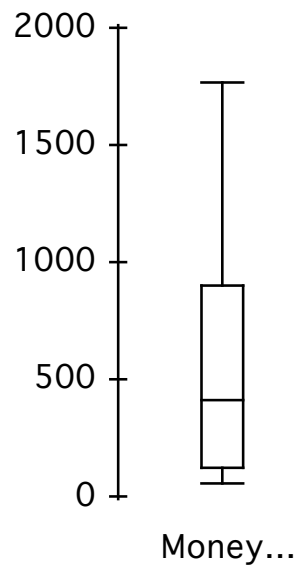
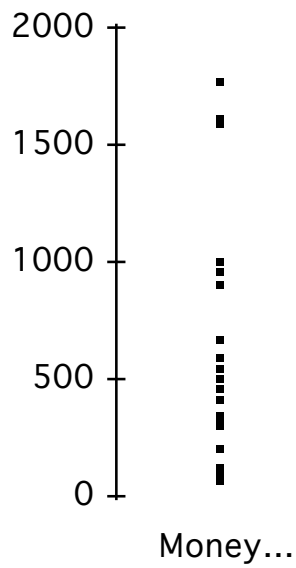
Bevölkerungsanteile über 80, Baden-Württemberg, 1993





Statistiken und Graphiken für 'Money Spent'

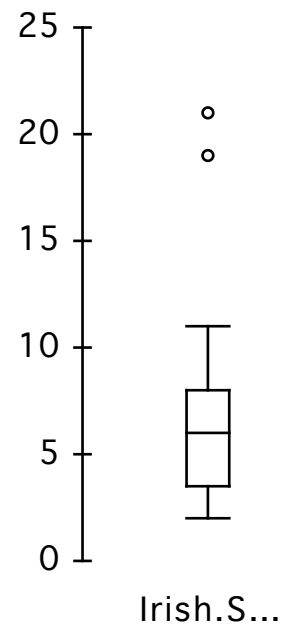
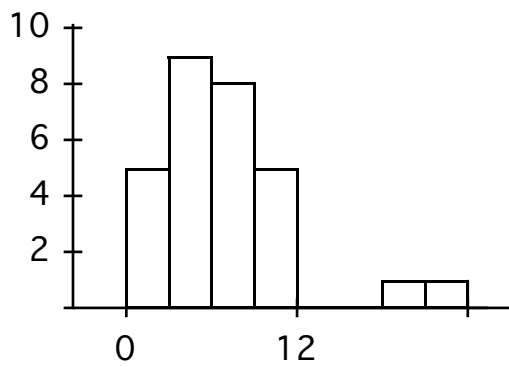
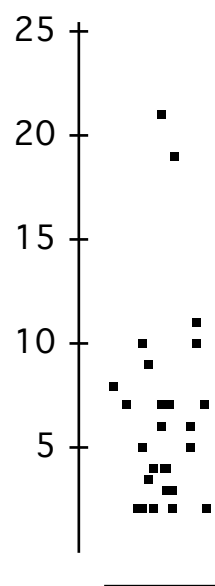
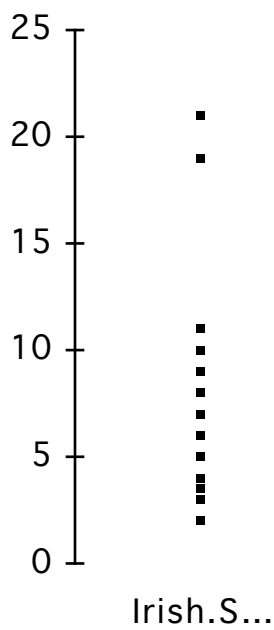
Mean	563.519	Median	415	Missing	3
Min	64	Max	1770		
SD	498.545	IntQRange	775.5		





Statistiken und Graphiken für 'Irish Stay'

Mean	6.569	Median	6	Missing	1
Min	2	Max	21		
SD	4.62	IntQRange	4.875		





3.1.4 Fragen zu univariaten statistischen Graphiken für stetige Variablen

Gibt es Sonderwerte? (Fehler, Ausreißer, Modi...)

Ist die Verteilung symmetrisch?

Gibt es mehrere Modi?

Gibt es Gruppen/Cluster?

Gibt es Lücken?

Gibt es bevorzugte Werte?

Andere Muster oder Struktur?



3.1.5 Diagramm Vergleiche

Dot Plots Histog.. Box Plots

Kleine

Große

Ausreißer

Identifikation

Symmetrie

Moden

Lücken

Verteilungs
Vergleiche

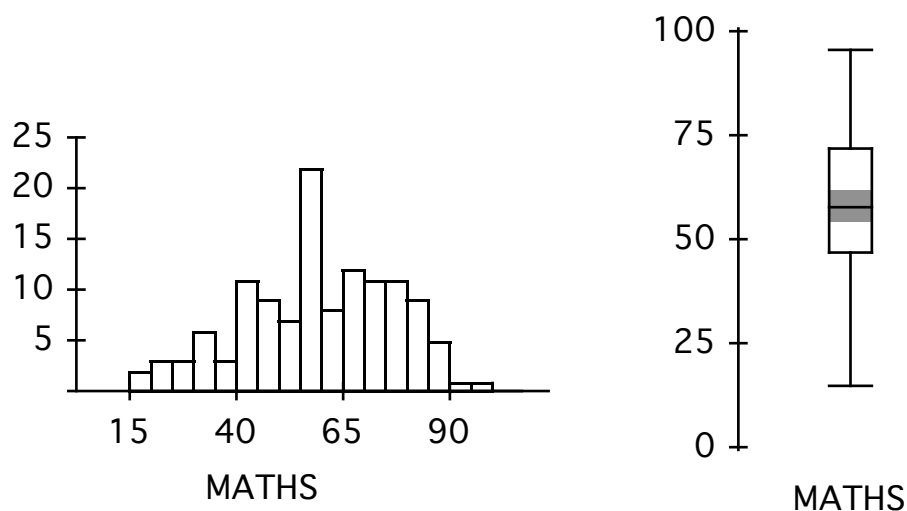


3.2 Analysen mit statistischen Graphiken

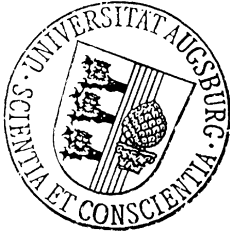
3.2.1 Noten aus einer Irischen Schule

126 SchülerInnen haben Klausuren in bis zu 9 Fächern geschrieben. Jede Note liegt zwischen 0 und 100. Um nicht durchzufallen, muß man mindestens 40 erreichen. Namen sind zufällig geändert worden.

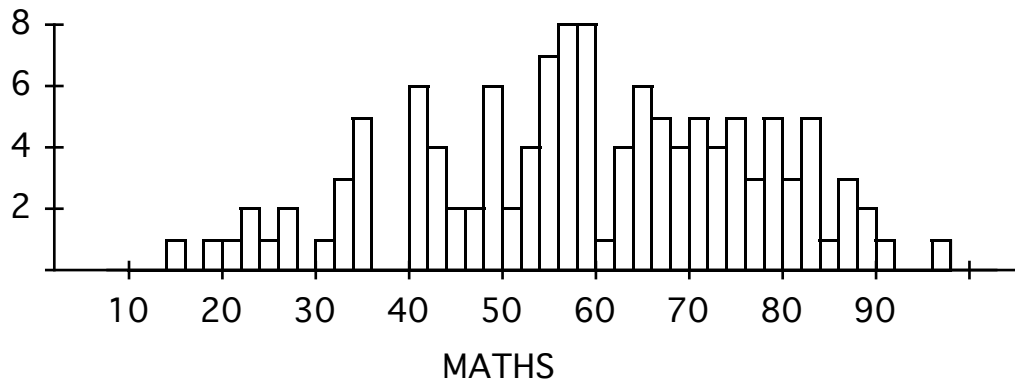
Die Noten in Mathematik sind in einem Histogramm und einem Boxplot dargestellt (beide aus Data Desk):



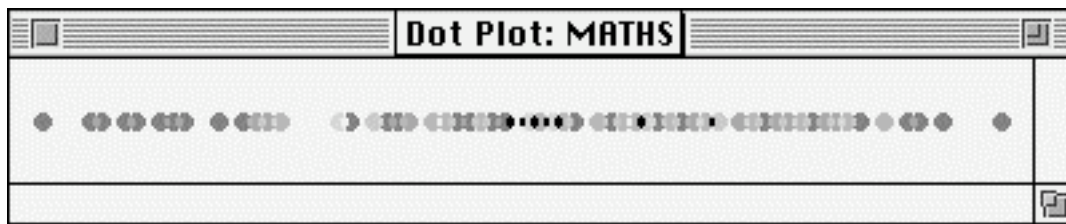
Der Boxplot zeigt eine symmetrische, eigentlich uninteressante Verteilung. Das Histogramm zeigt mehr Struktur und deutet auf etwas merkwürdiges um 40 hin.



Mit einer genaueren Klasseneinteilung wird eine Lücke unter 40 im Histogramm betont:



Wie auch mit einem Dotplot (aus MANET):



Beide Graphiken führen zum selben Schluß, daß zweifelhafte Resultate vermieden werden.

Eine Bestätigung findet man in den Noten der anderen Fächern, wie z.B. in der Geschichte-Prüfung:

