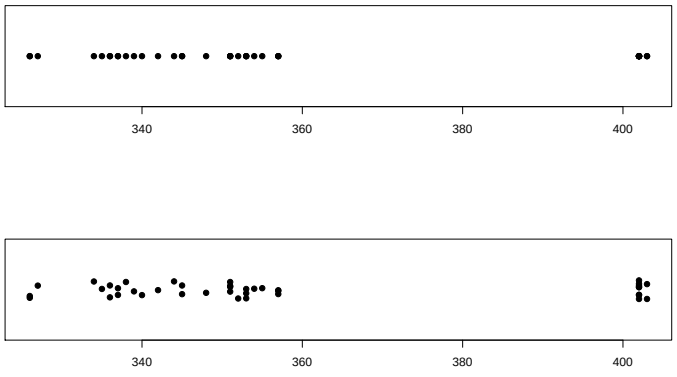
K3 Statistische Graphik

3.1 Graphische Darstellungen von stetigen Variablen

3.1.1 Punktplots (*stripchart* in R)

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 (oben). Mehrfache Punkte können nebeneinander gezeichnet werden. Eine andere Möglichkeit ist “jittering”, d.h. jeden Punkt zufällig nach rechts oder links bzw. nach oben oder unten zu verschieben (unten). Hier werden die Preise der ersten 40 Fälle des Diamanten Datensatzes gezeigt:

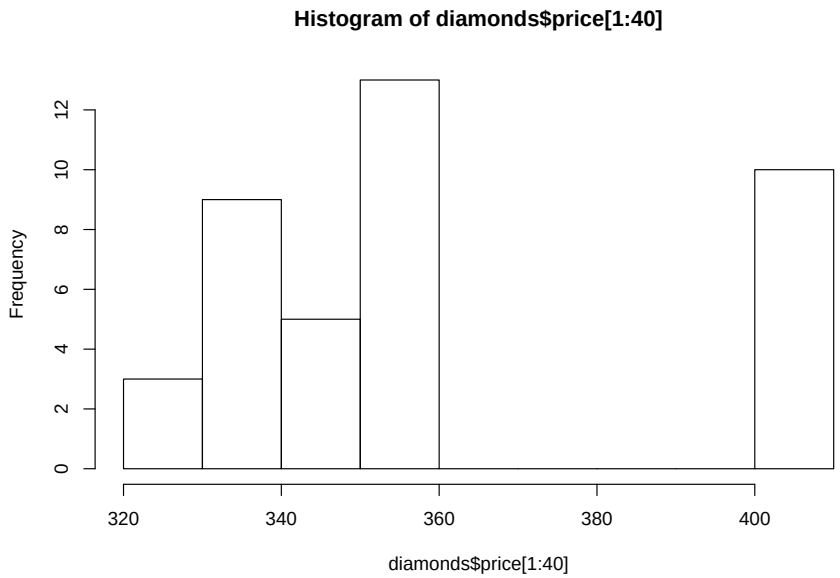


Um mehrfache Punkte sonst zu berücksichtigen, kann die Darstellung mit alphablending abgeändert werden. Jeder Punkt wird mit dem Dunkelheitswert α gezeichnet. Wenn k Punkte an derselben Bildschirmstelle gezeigt werden sollen, wird diese Stelle mit dem Wert

$$\max(1, k * \alpha)$$

gezeichnet. Das geht nicht in R im allgemeinen, aber im Paket *ggplot2* sowie in Mondrian (obwohl Mondrian keine Punktplots direkt anbietet).

3.1.2 Histogramme (*hist* in R) Die horizontale Achse wird in Klassen eingeteilt. Für jede Klasse wird die relative Häufigkeit als Rechteck über der Klasse aufgetragen.



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 stellen Häufigkeitstabellen für stetige Variablen graphisch dar. Sie geben eine Übersicht einer empirischen Verteilungsform.

Hauptparameter:
Anfangspunkt (“Ankerpunkt”), Klassenbreite

Ziele:
Genug Klassen, nicht zuviele, sinnvolle Klassengrenzen

Formatierungsparameter:
Skalierung, Größe, Seitenverhältnis, Beschriftung, Legende, Farbe/Schattierung

Heuristische Regeln für Klassenzahl bzw. Klassenbreite.

$n =$ Größe des Datensatzes
$s =$ Standardabweichung des Datensatzes
$IQ =$ InterQuartilAbstand

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

2. $(3.5sn^{-1/3})$ Klassenbreite (Scott)

3. $(2(IQ)n^{-1/3})$ Klassenbreite

Keine Heuristik wird alle Datensätze gut einteilen:

Beispiele:

	n	Min	Max	
(1)	100	0.4	9.5	
(2)	100	0.4	10.1	
(3)	101	0.4	$x_{(100)} = 9.5$	$x_{(101)} = 87.3$

Statt zu hoffen, daß ein Histogramm alle Informationen darstellen könnte, ist es besser, sich mehrere Histogramme der selben Daten anzuschauen:

R: Parameter müssen gesetzt werden, z.B.:

hist `breaks` bietet mehrere verschiedene Optionen.

truehist benutzt `nbins`, `h`, oder `breaks`.

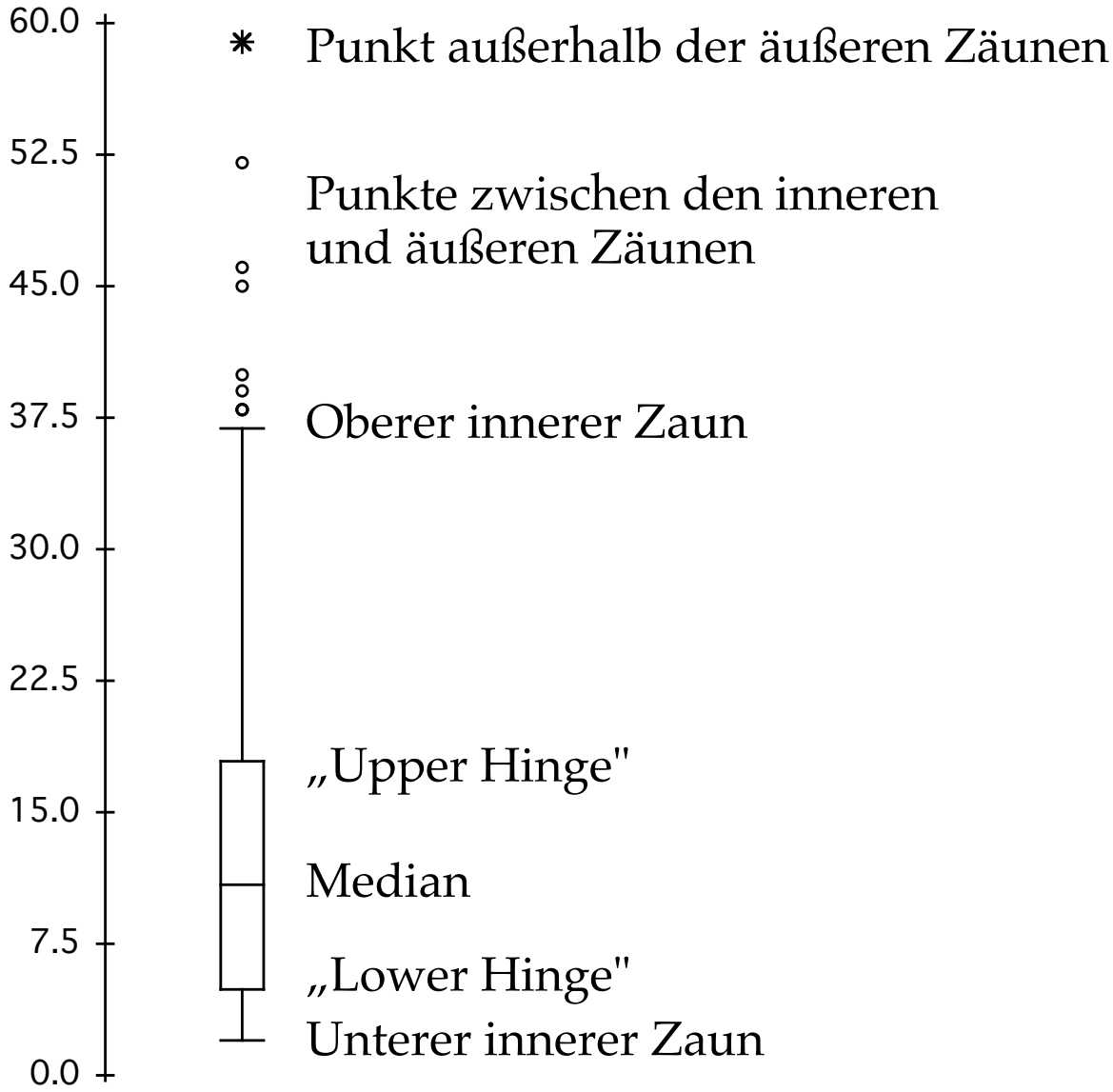
ggplot2 verwendet `binwidth`.

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

3.1.3 Boxplots (*boxplot* in R)

Eine klassiche Definition (leider gibt es viele andere).

z.B. Landwirtschaftliche Arbeiter in deutschen Wahlkreisen



Ein Boxplot basiert auf den Rängen der Daten.

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

Der Median (der mittlere Wert des Datensatzes)

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

$$M = \frac{1}{2}\left(x_{(k)} + x_{(k+1)}\right) \qquad \text{für } n = 2k$$

Oberen und untere Hinge (fast wie Quartile)

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

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

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

Die inneren Zäune sind die extremsten Werte innerhalb

$$F_o - 1.5(F_o - F_u) \text{ und } F_u + 1.5(F_o - F_u)$$

Die äußeren Zäune sind

$$F_o - 3(F_o - F_u) \text{ und } F_u + 3(F_o - F_u)$$

Punkte zwischen der inneren und äußeren Zäunen werden einzeln mit ‘o’ dargestellt (**Ausreißer**).
Punkte außerhalb der äußeren Zäunen werden einzeln mit ‘*’ dargestellt (“krasse” **Ausreißer**).

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

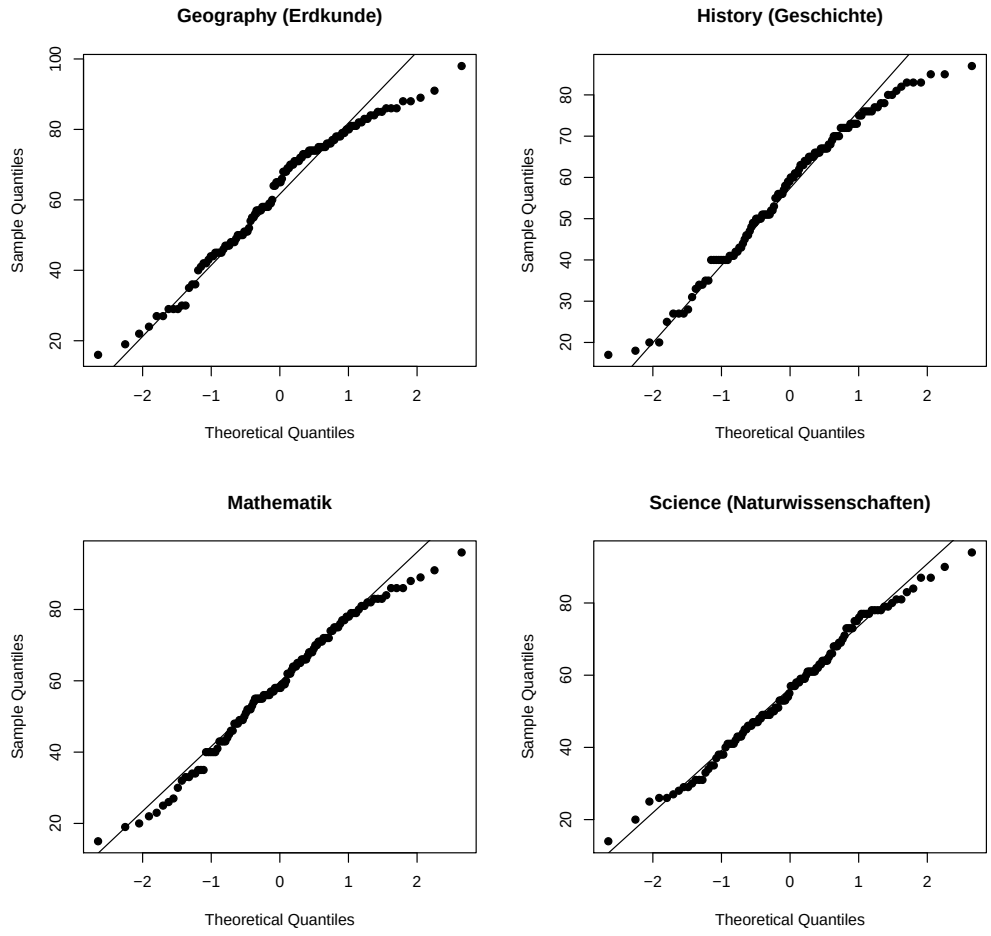
	Punktplots	Histogramme	Boxplots
n klein			
s groß			
Ausreißer			
Identifikation			
Symmetrie			
Moden			
Lücken			
Verteilungs Vergleiche			

3.1.6 QQ-Plot

QQ-Plots: zur Beurteilung der Anpassung der Daten einer Verteilung. Es werden die sortierten Beobachtungen $\{x_{(i)}\}$ gegen die entsprechenden Quantilen der hypothetischen theoretischen Verteilung, G, geplottet:

$$G^{-1}(p_i) \qquad p_i = \frac{i - 0.5}{n} \quad \text{für } i = 1, \ldots, n$$

Die mit *qqline* berechneten Linie geht durch die erste und dritte Quartile.



Die Software Mondrian

- Herunterladen: *www.rosuda.org/Mondrian*

- Daten laden: *File* $>$ *Open*
  Datensätze: tabgetrennte Textdateien (mit Kopfzeilen)
  Fehlende Werte werden mit “NA” angegeben.

- Variablen auswählen:
  Klick, Umschalttaste-Klick (alle dazwischen auswählen),
  Befehlstaste-Klick (Toggle Auswahl)

- Plots: *Plot* $>$ *. . .*

- Interaktiv arbeiten:

  $\rightarrow$ Abfragen mit *ctrl*

  $\rightarrow$ Punkte, Flächen auswählen mit Klick oder Drag (ziehen)

  $\rightarrow$ Plot Optionen: *ctrl-Click* auf Fensterhintergrund

  $\rightarrow$ Hilfe: *Help* $>$ *Reference Card*

3.1.7 Analysen mit statistischen Graphiken

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.

Ziele der Analyse:

- Welche Schüler/Schülerinnen sind die Besten?
- Sind sie gut in allen Fächern?
- Bei welchen Fächern werden die besten Noten erzielt?
- Sind die Resultate in verschiedenen Fächern assoziiert?
- Welche Strukturen haben die Notenverteilungen?