Statistik I — bis jetzt

K1 Einführung

K2 Beschreibende Statistik

2.1 Fragen

2.2 Beschreibung eines Datensatzes

2.3 Datentypen, Daten Schwierigkeiten

2.4 Schwierigkeiten mit Daten

2.5 Beschreibung kategorialer Variablen

2.6 Beschreibung von stetigen Variablen

2.7 Statistiken/Kenngrößen

K3 Statistische Graphik

3.1 Graphische Darstellungen stetiger Variablen
(Punktplots (“Dotplots”), Histogramme, Boxplots)

. . . und heute

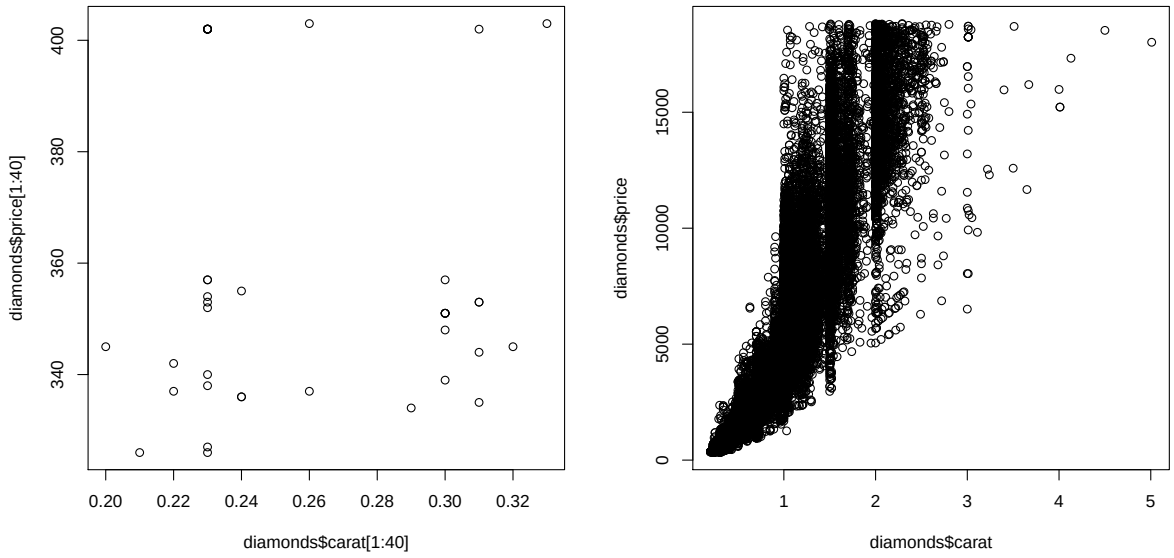
3.2 Bivariat stetig — Streudiagramme

3.3 Kategoriale Variablen
Säulendiagramme, Spineplots, Kreisdiagramme

3.4 Zeitreihen

3.2 Streudiagramme

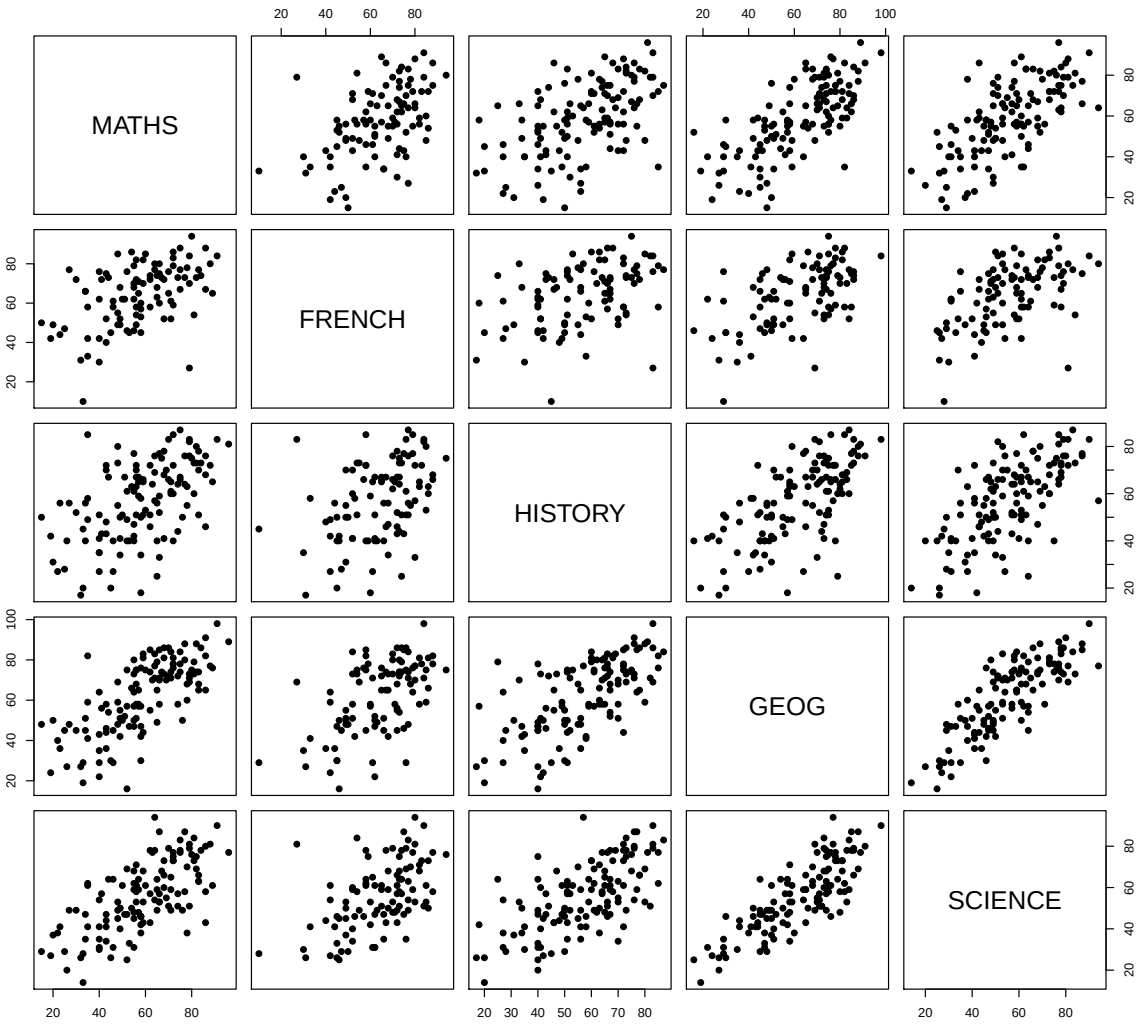
Darstellungen von zwei stetigen Variablen, Eine Variable wird als Y (vertikale Achse) und eine als X (horizontale Achse) genommen. Aus dem Diamantendatensatz wird *Preis* gegen carat geplottet. Links werden die ersten vierzig Fälle gezeigt und rechts alle:



Formatierungsparameter:

- Punkt-Form, -Größe, -Alphablending und -Farbe
- Skalierung, Seitenverhältnis
- Beschriftung, Legende

Für mehrere stetige Variablen kann man eine Matrix von Streudiagrammen zeichnen. (Irische Prüfungsergebnisse: Mathematik, Französisch, Geschichte, Geographie, Naturwissenschaft)

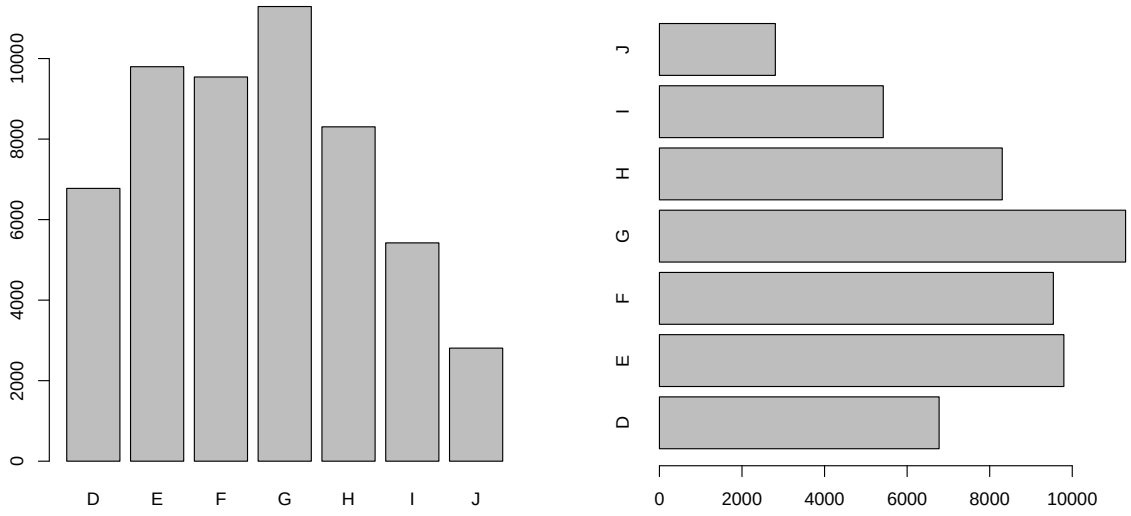


Das Diagramm an der Stelle (i, j) ist das Streudiagramm von X_i gegen $X_j (i \neq j)$. Auf der Diagonale findet man manchmal die Namen der Variablen, manchmal Histogramme oder andere univariate Darstellungen.

3.3 Darstellungen kategorialer Daten

3.3.1 (a) Säulendiagramme

Es gibt eine Säule für jede Kategorie.
Alle Säulen sind gleich breit/hoch.
Die Säulenhöhe/breite stellt die Anzahl der Fälle in der Kategorie dar.
z.B. Farbe aus dem Diamantendatensatz:



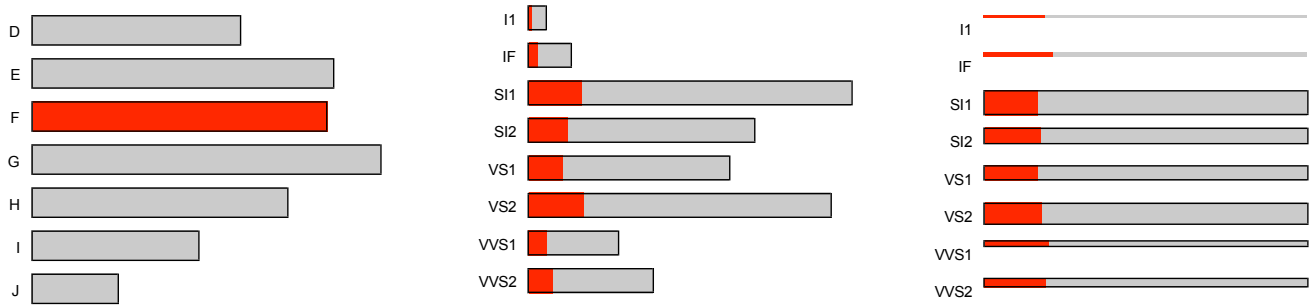
Formatierungsparameter:

- Vertikal/Horizontal
- Skalierung und Abstand zwischen den Säulen
- Farbe/Schattierung
- Beschriftung, Legende

3.3.1 (b) Spineplots

Wie Säulendiagramme, nur anders herum:
Bei vertikalen Spineplots sind die Säulen gleich hoch und die Breite stellt die Anzahl der Fälle in der Kategorie dar. Dadurch können relative Häufigkeiten verglichen werden.

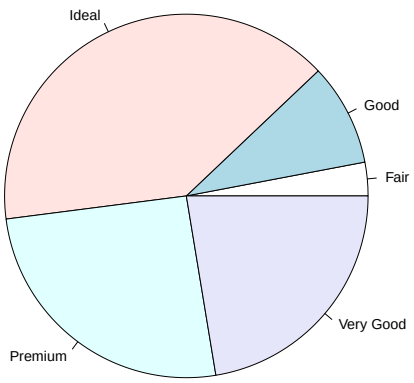
z.B. *clarity* (Klarheit) aus dem Diamantendatensatz:

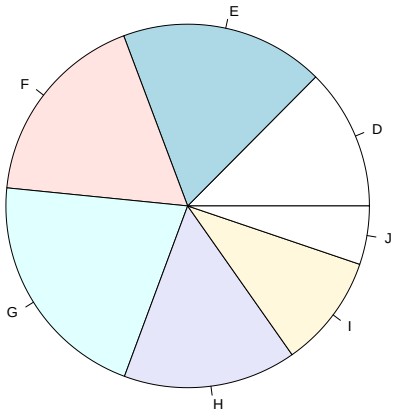


Im ersten Säulendiagramm sind die Diamanten mit *Farbe* F ausgewählt worden. Im zweiten werden die absoluten Häufigkeiten von allen, bzw. von den F-farbigen, dargestellt. Im dritten Plot erkennt man, dass die relativen Häufigkeiten von *Farbe* F für die verschiedenen Kategorien von *Klarheit* ungefähr gleich sind.

3.3.1 (c) Kreisdiagramme (Tortendiagramme)

Der Kreis wird in Flächen zerlegt.
Jedes Stück ist der Anzahl der Fälle proportional.





Vergleiche

- Säulendiagramme
  Am besten für direkte Vergleiche der Kategorien
  Höhe $\sim n_i$

- Spine Plots
  Zum Vergleichen von selektierten Anteilen
  Breite $\sim n_i$

- Kreisdiagramme
  Betont Anteile eines Ganzen
  Winkel $\sim n_i$

3.3.2 Zwei kategorialen Variablen

z.B. Geschlecht und USA (Ja/Nein) bei den Irischen Daten
Eine Kontingenztafel

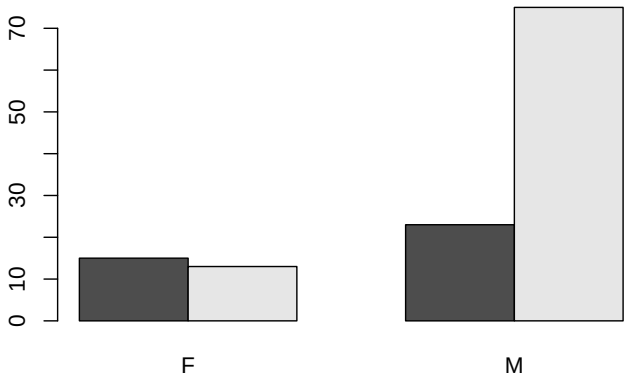
```
          F   M
US       15  23
nonUS    13  75
```

Im allgemeinen gibt es r Zeilen (engl. "row"), c Spalten (engl. "columns") und $r \times c$ Zellen insgesamt.

Darstellungsvariante:
Die Daten können als Zahlen, Zeilenprozente, Spaltenprozente oder, seltener, als Gesamtprozente angegeben werden.
Die Reihenfolge der Zeilen bzw. Spalten kann geändert werden.
Die Zeilenvariable und die Spaltenvariable können vertauscht werden.

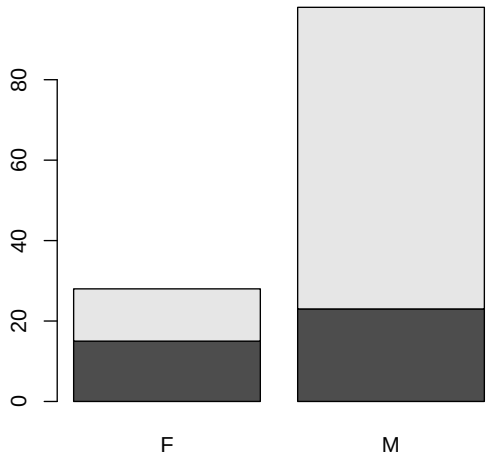
Säulendiagramme für zwei kategorialen Variablen

(a) Nebeneinander



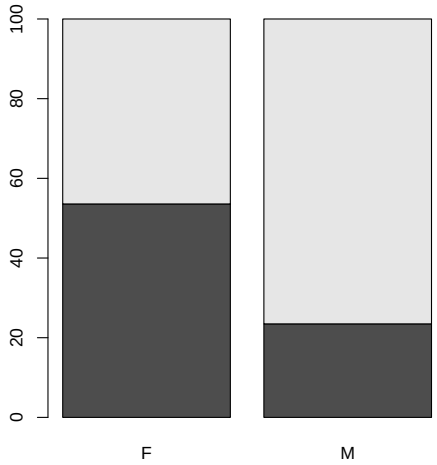
Die vier Kombinationen sind direkt vergleichbar.

(b) Gestapelt (nach absoluten Größen)



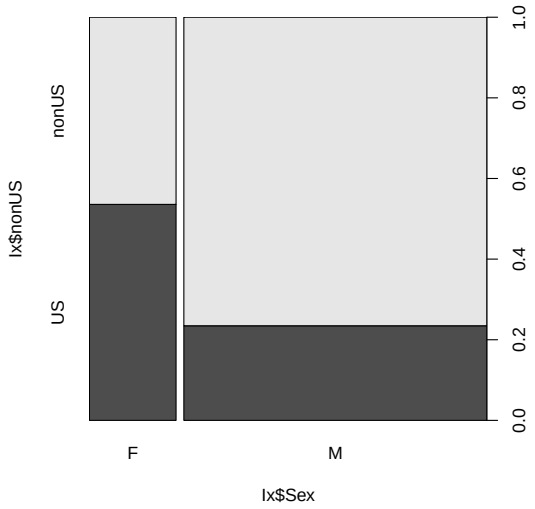
Proportionen werden betont. Bei ungleich großen Gruppen ist es schwierig, die Proportionen innerhalb der Gruppen zu vergleichen.

(c) gestapelt (nach Prozent)



Die Proportionen sind jetzt direkt vergleichbar, nur sind die Gruppengrößen vernachläßigt.

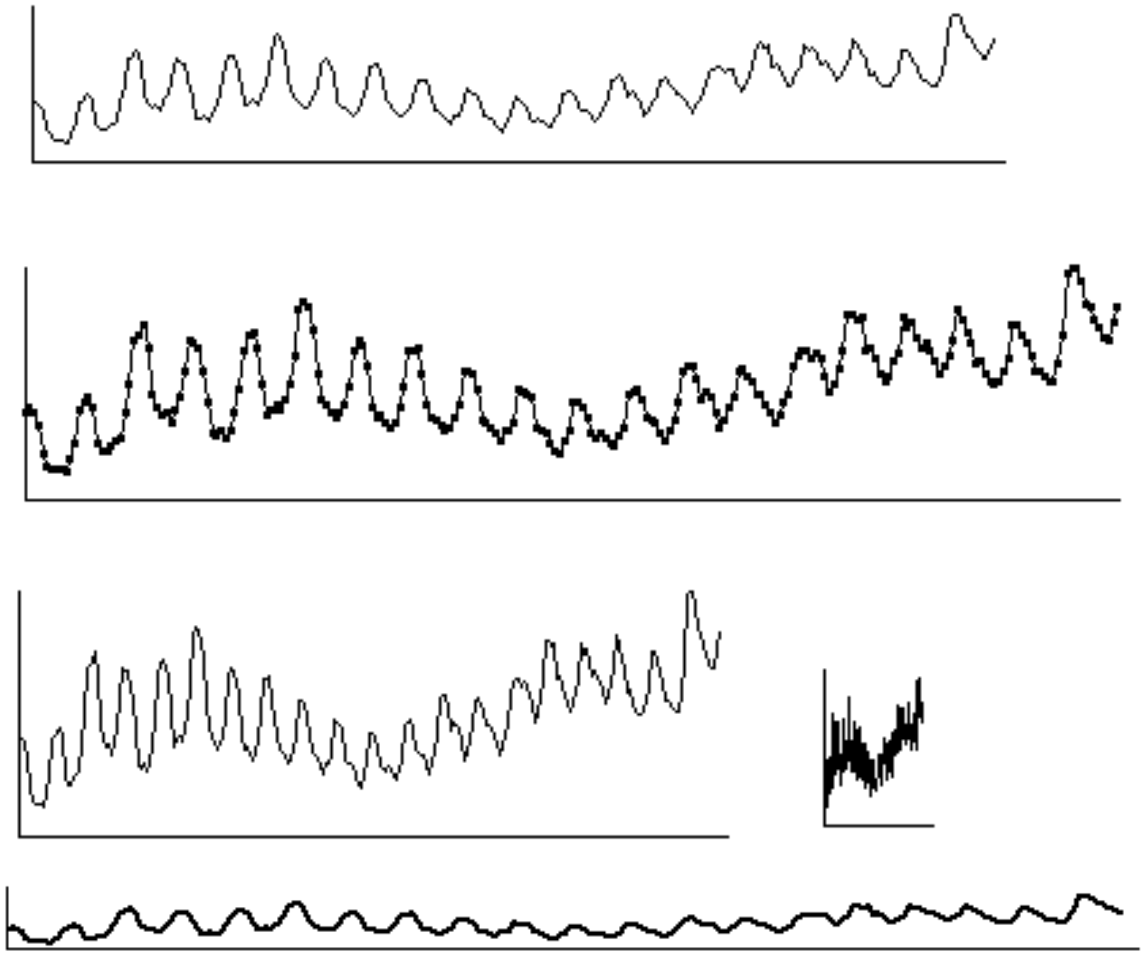
(d) gestapelt (nach Breite): Spineplots



Hier sieht man den Vorteil des Spineplots: Proportionen vergleichen zu können, ohne die Gruppengrößen zu vernachläßigen.

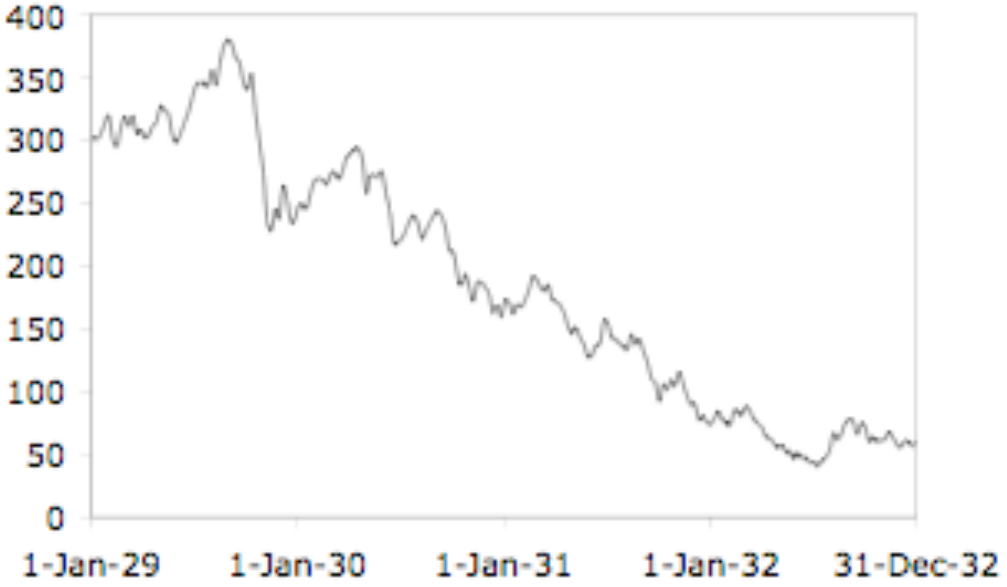
3.4 Zeitreihen

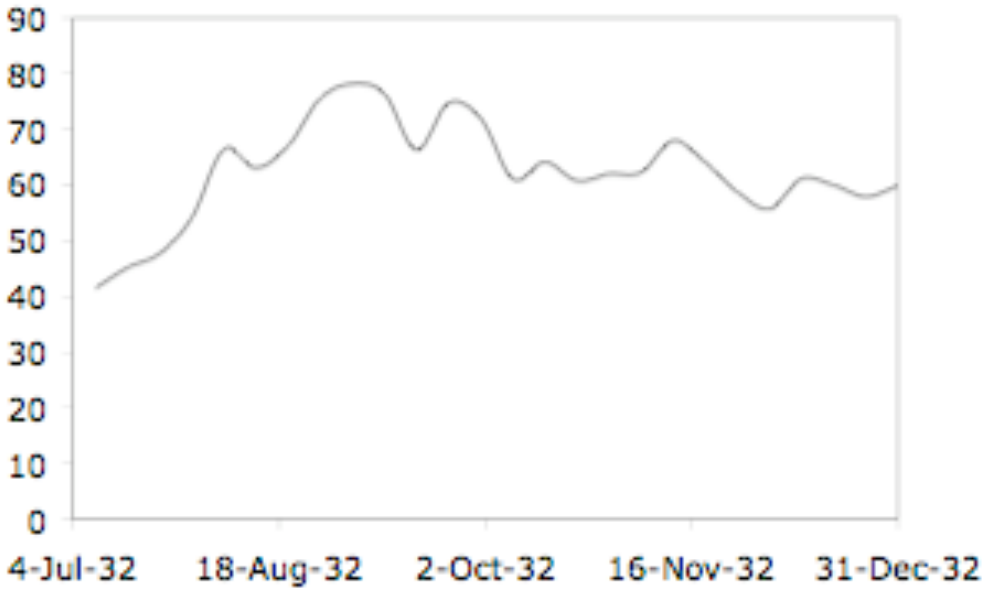
Zeitabhängige Beobachtungen bilden eine Zeitreihe $\{x_t\}$
z.B. Temperatur, Aktienpreise, Verkaufszahlen
Die Werte werden wie in einem Streudiagramm dargestellt, mit Zeit immer auf der horizontalen Achse und Verbindungslinien zwischen den aufeinanderfolgenden Punkten.



Fünf Abbildungen derselben monatlichen Arbeitslosenzahlen aus Kanada von 1952 bis 1971.

Bei einer Zeitreihengraphik sind u.a. wichtig:
Darstellung (Punkte, Linien, Glättungen)
Bildgröße, Seitenverhältnis
Skalierung, Ursprung, Zeitbeschriftungen
Untere und obere Grenzen
z.B. Dow Jones 1929-1932 und Juli bis Dez 1932





Statistik und die Nachrichten — AZ 23.4.2012

s.4 Betreuungsgeld Eine solche Regelung würde nach Expertenschätzungen jährlich mit sechs bis sieben Milliarden Euro zusätzlich zu Buche schlagen.

s.6 Spionage belaster Industrie Deutsche Firmen verlieren durch Industrieespionage jedes Jahr viele Milliarden Euro.

(Webseite 25.4.2012) Solarien sind so gefährlich wie die Sonne am Äquator 224000 Menschen, rechnet Nettekoven vor, erkranken jedes Jahr an Hautkrebs, 26000 davon am besonders gefährlichen malignen Melanom. Und die Zahlen steigen stetig.

Wie werden solche Zahlen geschätzt?