Häufigkeitstabellen
Die Prozedur FREQ

Ein-, zwei- und höherdimensionale Häufigkeiten

Eindimensionale Zufallsvariablen

$$X: \quad \begin{pmatrix} x_0 & x_1 & \cdots & x_n & \cdots \\ p_0 & p_1 & \cdots & p_n & \cdots \end{pmatrix}$$

Die p_i sind zu schätzen:

$$\hat{p}_i = \frac{n_i}{N}$$

N: Stichprobenumfang n_i: relative Häufigkeiten

```
PROC FREQ <Optionen>;
  TABLES variablenliste </Optionen>;
RUN;
```

Descr_Freq_Banknote.sas Descr_Freq.sas

Zweidimensionale diskrete Zufallsgrößen

Einführendes Beispiel

3maliges Werfen einer Münze

X: Anzahl von Blatt nach 3 Würfen

Y: Anzahl von Blatt nach 2 Würfen

Element von Ω	X	Y
BBB	3	2
BBZ	2	2
BZB	2	1
BZZ	1	1
ZBB	2	1
ZBZ	1	1
ZZB	1	0
ZZZ	0	0

Zweidimensionale diskrete Zufallsgrößen

Einführendes Beispiel (Fortsetzung)

Besetzungswahrscheinlichkeiten

$X \mid Y$	0	1	2	
0	$\frac{1}{8}$	0	0	$\frac{1}{8}$
1	$\frac{1}{8}$	$\frac{1}{4}$	0	$\frac{3}{8}$
2	0	$\frac{1}{4}$	$\frac{1}{8}$	$\frac{3}{8}$
3	0	0	$\frac{1}{8}$	$\frac{1}{8}$
	$\frac{1}{4}$	$\frac{1}{2}$	$\frac{1}{4}$	1

$$X : \begin{pmatrix} 0 & 1 & 2 & 3 \\ \frac{1}{8} & \frac{3}{8} & \frac{3}{8} & \frac{1}{8} \end{pmatrix}$$

$$Y : \begin{pmatrix} 0 & 1 & 2 \\ \frac{1}{4} & \frac{1}{2} & \frac{1}{4} \end{pmatrix}$$

Tabelle der zweidimensionalen Wahrscheinlichkeiten

$X \mid Y$	y_1	y_2	$\cdots$	y_j	$\cdots$	y_N	
x_1	p_{11}	p_{12}	$\cdots$	p_{1j}	$\cdots$	p_{1N}	$p_{1.}$
x_2	p_{21}	p_{22}	$\cdots$	p_{2j}	$\cdots$	p_{2N}	$p_{2.}$
$\cdots$							
x_i	p_{i1}	p_{i2}	$\cdots$	p_{ij}	$\cdots$	p_{iN}	$p_{i.}$
$\cdots$							
x_M	p_{M1}	p_{M2}	$\cdots$	p_{Mj}	$\cdots$	p_{MN}	$p_{M.}$
	$p_{.1}$	$p_{.2}$	$\cdots$	$p_{.j}$	$\cdots$	$p_{.N}$	1

Zweidimensionale diskrete Zufallsgrößen

Zweidimensionale Zufallsvariable

Seien X, Y Zufallsgrößen. Das Paar (X, Y) heißt zweidimensionale Zufallsvariable.

Zweidimensionale diskrete Zufallsgrößen

Zweidimensionale Zufallsvariable

Seien X, Y Zufallsgrößen. Das Paar (X, Y) heißt zweidimensionale Zufallsvariable.

Seien X und Y diskret und (x_i, y_j) die möglichen Ergebnisse von (X, Y), $\quad i = 1, \ldots, M, j = 1, \ldots, N$.

gemeinsame Wahrscheinlichkeitsfunktion von (X, Y)

$$p_{ij} = P(X = x_i, Y = y_j),$$

Zweidimensionale diskrete Zufallsgrößen

Zweidimensionale Zufallsvariable

Seien X, Y Zufallsgrößen. Das Paar (X, Y) heißt zweidimensionale Zufallsvariable.

Seien X und Y diskret und (x_i, y_j) die möglichen Ergebnisse von (X, Y), $\quad i = 1, \ldots, M, j = 1, \ldots, N$.

gemeinsame Wahrscheinlichkeitsfunktion von (X, Y)

$$p_{ij} = P(X = x_i, Y = y_j),$$

$$\begin{aligned} p_{ij} &\geq 0 \\ \sum_{i,j} p_{ij} &= 1 \end{aligned}$$

$$p_{i.} := \sum_{j=1}^{N} p_{ij} \qquad p_{.j} := \sum_{i=1}^{M} p_{ij}$$

Zweidimensionale diskrete Zufallsgrößen

Beispiel

Treiben Sie Sport?

X: 0 - nein 1 - ja

Y: 0 - weiblich 1 - männlich

X | Y	0	1	
0	p_{00}	p_{01}	$p_{0.}$
1	p_{10}	p_{11}	$p_{1.}$
	$p_{.0}$	$p_{.1}$	

p_{ij}: unbekannt!

Frage: Ist das Sportverhalten von Männern und Frauen unterschiedlich? Hängt das Sportverhalten vom Geschlecht ab?

Zweidimensionale diskrete Zufallsgrößen

Kontingenztafel

Befragung liefert Häufigkeiten für die einzelnen Felder. Anhand dieser Häufigkeiten werden die Wahrscheinlichkeiten geschätzt!

Die Tabelle der Häufigkeiten heißt Kontingenztafel

X | Y	0	1	# der beobachteten	
0	n_{00}	n_{01}	$n_{0.}$	Nichtsportler
1	n_{10}	n_{11}	$n_{1.}$	Sportler
	$n_{.0}$	$n_{.1}$		
	# der befragten			
	Frauen	Männer		

$$p_{ij} \approx \frac{n_{ij}}{n} = \hat{p}_{ij}$$

Zweidimensionale diskrete Zufallsgrößen

Häufigkeitstabellen in SAS


```
PROC FREQ <Optionen>;
  TABLES variablenliste </Optionen>;
  TABLES vliste1*vliste2 </Optionen>;
  TABLES vliste1*vliste2*varliste3;RUN;
```

Option im Prozedur-Step

ORDER=schlüsselwort, z.B. ORDER=FREQ
wenn die Ausgabe nach Häufigkeiten geordnet.

Optionen der TABLES-Anweisung

MISSING: fehlende Werte werden bei der Berechnung relativer Häufigkeiten mit einbezogen.

OUT=sasfile: Ausgabe der Tabelle in ein SAS-File

Optionen der TABLES-Anweisung

nur für mehrdim. Tabellen

CHISQ:	χ^2-Unabhängigkeitstest
CMH:	u.a. Odds Ratio
MEASURES:	Assozitionsmaße, Korrelationskoeffizient
NO...	keine Ausgabe von:
NOFREQ:	absoluten Häufigkeiten
NOPERCENT:	relativen Häufigkeiten
NOROW:	Zeilenhäufigkeiten
NOCOL:	Spaltenhäufigkeiten

Assoziationsmaße
nur für mehrdim. Tabellen

χ^2-Statistik

$$\sum_{i,j} \frac{(p_{ij} - p_{i.}p_{.j})^2}{p_{i.}p_{.j}}$$

Φ-Koeffizient für 2x2 Tafeln

$$\Phi^2 = \frac{(p_{11}p_{22} - p_{12}p_{21})^2}{p_{1.}p_{2.}p_{.1}p_{.2}}$$

Odds Ratio für 2x2 Tafeln

$$OR = \frac{p_{11}p_{22}}{p_{12}p_{21}}$$

Schätzung: Ersetzen der Wahrscheinlichkeiten durch die jeweiligen relativen Häufigkeiten.

Assoziationsmaße, Beispiel

Mendelsche Kreuzungsversuche

```
DATA Erbsen;
INPUT rund gruen Anzahl;
CARDS;

0 0 101
0 1  32
1 0 315
1 1 108

;
RUN;
```

Assoziationsmaße, Beispiel

Mendelsche Kreuzungsversuche

```
DATA Erbsen;
INPUT rund gruen Anzahl;
CARDS;

0 0 101
0 1  32
1 0 315
1 1 108

;
RUN;
```

```
PROC FREQ;
WEIGHT Anzahl;
TABLES rund*gruen \
  chisq cmh;
RUN;
```

Assoziationsmaße, Beispiel

Mendelsche Kreuzungsversuche

```
DATA Erbsen;
INPUT rund gruen Anzahl;
CARDS;

0 0 101
0 1  32
1 0 315
1 1 108

;
RUN;
```

```
PROC FREQ;
WEIGHT Anzahl;
TABLES rund*gruen \
  chisq cmh;
RUN;
```

$\chi^2 = 0.1163$ Φ-Koeffizient=0.0145.

Zusammenhangsmaße
zwischen Zufallsvariablen X, Y

Erinnerung: Varianz der Zufallsvariablen X

$$\begin{aligned} var(X) &= \mathbf{E}(X - \mathbf{E}X)^2 \\ &= \mathbf{E}[(X - \mathbf{E}X)(X - \mathbf{E}X)] \end{aligned}$$

Zusammenhangsmaße

zwischen Zufallsvariablen X, Y

Erinnerung: Varianz der Zufallsvariablen X

$$\begin{aligned} var(X) &= \mathbf{E}(X - \mathbf{E}X)^2 \\ &= \mathbf{E}[(X - \mathbf{E}X)(X - \mathbf{E}X)] \end{aligned}$$

Kovarianz der Zufallsvariablen X und Y

$$\begin{aligned} Cov(X, Y) &= \mathbf{E}(X - \mathbf{E}X)(Y - \mathbf{E}Y) \\ &= \mathbf{E}(XY) - \mathbf{E}(X)\mathbf{E}(Y) \end{aligned}$$

Zusammenhangsmaße

zwischen Zufallsvariablen X, Y

Erinnerung: Varianz der Zufallsvariablen X

$$\begin{aligned} var(X) &= \mathbf{E}(X - \mathbf{E}X)^2 \\ &= \mathbf{E}[(X - \mathbf{E}X)(X - \mathbf{E}X)] \end{aligned}$$

Kovarianz der Zufallsvariablen X und Y

$$\begin{aligned} Cov(X, Y) &= \mathbf{E}(X - \mathbf{E}X)(Y - \mathbf{E}Y) \\ &= \mathbf{E}(XY) - \mathbf{E}(X)\mathbf{E}(Y) \end{aligned}$$

Korrelation der Zufallsvariablen X und Y

$$Corr(X, Y) = \frac{\mathbf{E}[(X - \mathbf{E}X)(Y - \mathbf{E}Y)]}{\sqrt{var(X) \cdot var(Y)}}$$

Zusammenhangsmaße (2)

Erinnerung: empirische Varianz

$$s_X^2 = \frac{1}{n-1} \sum_{i=1}^{n} (X_i - \overline{X})(X_i - \overline{X})$$

Zusammenhangsmaße (2)

Erinnerung: empirische Varianz

$$s_X^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \overline{X})(X_i - \overline{X})$$

empirische Kovarianz

$$s_{XY} = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \overline{X})(Y_i - \overline{Y})$$

Zusammenhangsmaße (2)

Erinnerung: empirische Varianz

$$s_X^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \overline{X})(X_i - \overline{X})$$

empirische Kovarianz

$$s_{XY} = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \overline{X})(Y_i - \overline{Y})$$

empirische Korrelation, Pearson-Korrelationskoeffizient

$$r_{XY} := \frac{s_{XY}}{s_X s_Y}$$

Pearson-Korrelationskoeffizient
Eigenschaften

- Es gilt stets:

$$-1 \leq r_{XY} \leq 1.$$

Pearson-Korrelationskoeffizient
Eigenschaften

- Es gilt stets:

$$-1 \leq r_{XY} \leq 1.$$

- Der Korrelationskoeffizient ist invariant gegenüber linearen Transformationen

$$x \longrightarrow a + bx$$

Pearson-Korrelationskoeffizient
Eigenschaften

- Es gilt stets:
$$-1 \leq r_{XY} \leq 1.$$

- Der Korrelationskoeffizient ist invariant gegenüber linearen Transformationen
$$x \longrightarrow a + bx$$

- $|r_{XY}| = 1$ gdw. alle Punkte auf einer Geraden liegen, $y = mx + b, m \neq 0$
  $r_{XY} = \quad 1 \rightarrow$ Anstieg > 0
  $r_{XY} = -1 \rightarrow$ Anstieg < 0

Pearson-Korrelationskoeffizient

- Der Korrelationskoeffizient ist also ein Maß für die lineare Abhängigkeit von X und Y.
- $r_{XY} \approx 0 \longrightarrow$ keine lineare Beziehung zwischen X und Y erkennbar, aber es sind durchaus andere Abhängigkeiten möglich!

Realisierung in SAS:

```
PROC CORR PEARSON <DATA = Dateiname>;
  VAR X Y;
RUN;
```

Spearman-Korrelationskoeffizient

Spearman-Rangkorrelationskoeffizient

$$r_S = \frac{\sum_{i=1}^{n}(R_i - \overline{R})(S_i - \overline{S})}{\sqrt{\sum_i (R_i - \overline{R})^2 \sum_i (S_i - \overline{S})^2}}$$

R_i: Rang von X_i in der geordneten Stichprobe $X_{(1)} \leq \ldots \leq X_{(n)}$

S_i: Rang von Y_i in der geordneten Stichprobe $Y_{(1)} \leq \ldots \leq Y_{(n)}$

```
PROC CORR SPEARMAN <DATA=Dateiname>;
  VAR X Y;
RUN;
```

Spearman-Korrelationskoeffizient

$$
\begin{aligned}
r_S &= \frac{\sum_{i=1}^{n}(R_i - \overline{R})(S_i - \overline{S})}{\sqrt{\sum_{i=1}^{n}(R_i - \overline{R})^2 \sum_{i=1}^{n}(S_i - \overline{S})^2}} \\
&= \frac{\sum_{i=1}^{n}(R_i - \frac{n+1}{2})(S_i - \frac{n+1}{2})}{\sqrt{\sum_{i=1}^{n}(R_i - \overline{R})^2 \sum_{i=1}^{n}(S_i - \overline{S})^2}} \\
&= 1 - \frac{6 \cdot \sum_{i=1}^{n}(R_i - S_i)^2}{n \cdot (n^2 - 1)}
\end{aligned}
$$

$$
-1 \leq r_S \leq +1
$$

$|r_S| = 1$ gdw. X_i, Y_i in gleicher oder entgegengesetzter Weise geordnet sind!

Spearman-Korrelationskoeffizient

Beweis der letzten Formel (1)

$$r_S = \frac{\sum_{i=1}^{n}(R_i - \overline{R})(S_i - \overline{S})^2}{\sqrt{\sum_{i=1}^{n}(R_i - \overline{R})^2 \sum_{i=1}^{n}(S_i - \overline{S})}}$$

Nenner:

$$\sum_{i=1}^{n}(R_i - \overline{R})^2 = \sum_{i=1}^{n}(S_i - \overline{S})^2 = \sum_{i=1}^{n}(i - \frac{n+1}{2})^2$$

$$= \sum i^2 - 2 \cdot \frac{n+1}{2} \sum i + n \cdot (\frac{n+1}{2})^2$$

$$= \frac{n \cdot (n+1) \cdot (2n+1)}{6} - \frac{n \cdot (n+1)^2}{2} + \frac{n \cdot (n+1)^2}{4}$$

$$= \frac{n \cdot (n+1)}{12} \cdot [2 \cdot (2n+1) - 3 \cdot (n+1)]$$

$$= \frac{(n-1) \cdot n \cdot (n+1)}{12} = \frac{n \cdot (n^2-1)}{12}$$

Spearman-Korrelationskoeffizient
Beweis der letzten Formel (2)

Zähler:

$$\sum_{i=1}^{n}(R_i - \overline{R})(S_i - \overline{S}) = \sum_{i=1}^{n}(R_i - \frac{n+1}{2})(S_i - \frac{n+1}{2})$$

$$= \sum_{i=1}^{n} R_i S_i - 2 \cdot \frac{n+1}{2} \sum_{i=1}^{n} R_i + n \cdot (\frac{n+1}{2})^2$$

$$= \sum_{i=1}^{n} R_i S_i - \frac{n \cdot (n+1)^2}{4}$$

Damit erhalten wir eine weitere Darstellung für r_S:

$$r_S = 12 \cdot \frac{\sum_{i=1}^{n} R_i S_i - \frac{n \cdot (n+1)^2}{4}}{(n-1) \cdot n \cdot (n+1)}$$

Spearman-Korrelationskoeffizient

Andere Darstellung für den Zähler

Setzen: $d_i := R_i - S_i = (R_i - \frac{n+1}{2}) + (\frac{n+1}{2} - S_i)$

$$\begin{aligned}
\sum d_i^2 &= \sum (R_i - \frac{n+1}{2})^2 + \sum (S_i - \frac{n+1}{2})^2 \\
&\quad -2 \sum (R_i - \frac{n+1}{2})(S_i - \frac{n+1}{2}) \\
&= \frac{(n-1)n(n+1)}{12} + \frac{(n-1)n(n+1)}{12} \\
&\quad -2 \cdot r_S \cdot \frac{(n-1)n(n+1)}{12} \\
&= \frac{(n-1)n(n+1)}{6}(1 - r_S) \\
r_S &= 1 - \frac{6 \sum d_i^2}{(n-1)n(n+1)}
\end{aligned}$$

Spearman-Korrelationskoeffizient

Drei Darstellungen

$$
\begin{aligned}
r_S &= \frac{\sum_{i=1}^{n}(R_i-\overline{R})(S_i-\overline{S})}{\sqrt{\sum_i (R_i-\overline{R})^2 \sum_i (S_i-\overline{S})^2}} \\
&= 12 \cdot \frac{\sum_{i=1}^{n} R_i S_i - \frac{n\cdot(n+1)^2}{4}}{(n-1)n(n+1)} \\
&= 1 - \frac{6\sum (R_i-S_i)^2}{(n-1)n(n+1)}
\end{aligned}
$$

Bem.: Es gilt:

a) $-1 \leq r_S \leq 1$

b) $r_S = 1 \Leftrightarrow R_i = S_i \quad \forall i = 1,\ldots,n$

c) $r_S = -1 \Leftrightarrow R_i = n+1-S_i \quad \forall i = 1,\ldots,n$

Vergleich der Korrelationskoeffizienten

Pearson - Spearman

Vorteile Spearman

- es genügt ordinales Meßniveau
- leicht zu berechnen
- r_S ist invariant gegenüber monotonen Transformationen
- gute Interpretation, wenn $r_S \approx -1, 0, 1$ (wie bei Pearson)
- eignet sich als Teststatistik für einen Test auf Unabhängigkeit
- ist robust gegen Abweichungen von der NV.

Vergleich der Korrelationskoeffizienten

Pearson - Spearman

Nachteile Spearman

- wenn kardinales (stetiges) Meßniveau $\longrightarrow$ Informationsverlust
- schwierige Interpretation,

wenn r_S nicht nahe 0, 1, oder -1
(gilt eingeschränkt auch für Pearson)

Kendalls τ (Konkordanzkoeffizient)

$(X_i, Y_i), i = 1, ..., n$

$$a_{ij} = \begin{cases} 1, \text{ falls } x_i < x_j \wedge y_i < y_j \text{ oder} \\ \qquad\qquad x_i > x_j \wedge y_i > y_j \\ -1, \text{ falls } x_i < x_j \wedge y_i > y_j \text{ oder} \\ \qquad\qquad x_i > x_j \wedge y_i < y_j \\ 0, \text{ sonst} \end{cases}$$

$$= sgn[(X_i - X_j)(Y_i - Y_j)]$$

Falls $a_{ij} = 1$ so heißen die Paare konkordant
Falls $a_{ij} = -1$ " diskordant
Falls $a_{ij} = 0$ " gebunden

Kendalls τ (Konkordanzkoeffizient)

$$\tau = \frac{2 \cdot \sum_{i<j} a_{ij}}{N \cdot (N-1)} = \frac{1}{\binom{N}{2}} \cdot \sum_{i<j} a_{ij}$$

$$= \frac{\sharp \text{ konkordanter Paare - } \sharp \text{ diskordanter Paare}}{\binom{N}{2}}$$

Bem.: einfache Berechnung, wenn neue Paare hinzukommen

Bem.: meist gilt: $|\tau| < |r_S|$. Approximation von τ:

$$\hat{\tau} = \frac{2}{3}\frac{N+1}{N} r_S$$

```
PROC CORR KENDALL;
  VAR X Y; RUN;
```