# 6. Bivariate Häufigkeit

## 6.1. Dichotome Variable

Eine bivariate Häufigkeitsverteilung zweier dichotomer Var, wie z.B. "Geschl" und "Karies" (Geschl_Karies.dta) erhält man mit (siehe 5.)
***tabulate geschl karies*** oder ***tab2 geschl karies***

Zusätzlich zur Vierfeldertafel kann man mit dem Chi²-Test ermitteln, ob ein Zusammenhang zwischen beiden Variablen besteht. Die **Stärke** eines Zusammenhangs kann z.B. mit **Cramer's V** oder **Kendall's taub** beschrieben werden. Bei geringer Fallzahl oder erwarteter Häufigkeiten ≤ 5 sollte mit dem **exakten Test nach Fisher** geprüft werden.
Die erwartete Häufigkeit in den Zellen erhält man mit
***tab geschl karies, expected nofreq*** und im Ergebnis sind alle Werte > 5 (rechte Tabelle)

```
tab geschl karies, expected nof
```

| geschl | karies 0 | 1 | Total |
|---|---|---|---|
| 1 | 13.5 | 27.5 | 41.0 |
| 2 | 12.5 | 25.5 | 38.0 |
| Total | 26.0 | 53.0 | 79.0 |

***tab geschl karies, chi exact taub V*** liefert die gewünschte 2x2-Tafel und die Angaben zur Stärke des Zusammenhangs.

```
tab geschl karies, chi2 exact taub V
```

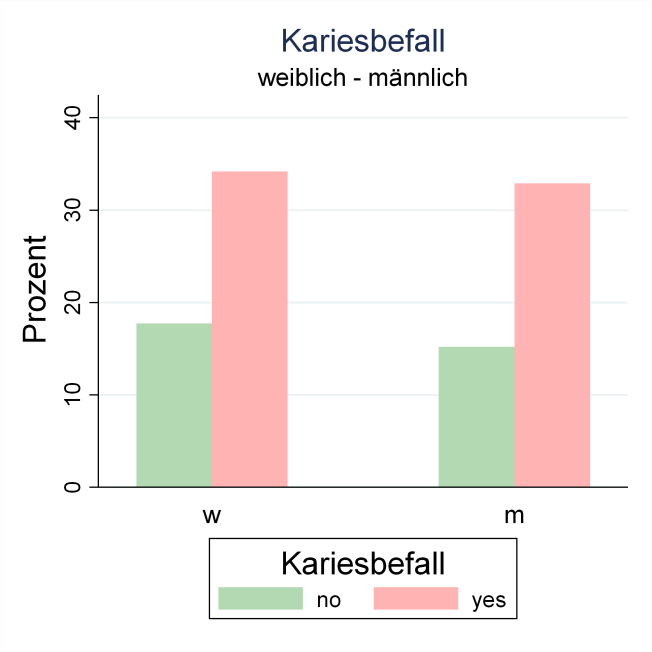
| geschl | karies 0 | 1 | Total |
|---|---|---|---|
| 1 | 14 | 27 | 41 |
| 2 | 12 | 26 | 38 |
| Total | 26 | 53 | 79 |

```
          Pearson chi2(1) =   0.0589   Pr = 0.808
               Cramér's V =   0.0273
          Kendall's tau-b =   0.0273  ASE = 0.112
           Fisher's exact =                 1.000
  1-sided Fisher's exact =                 0.499
```

ASE bezeichnet den approximalen Standardfehler für Kendall's taub . Damit lässt sich auf dem 5%-Niveau prüfen, ob Kendall's taub signifikant von Null verschieden ist. Die Prüfgröße z = taub / ASE ist standardnormalverteilt, der kritische Wert $z_C$ = 1,96 .
Wegen z = 0,0273 / 0,112 = 0,244 ist Kendall's taub nicht signifikant von Null verschieden (0,244 < 1,96). Der zugehörige p-Wert = 0,807 ähnlich dem Wert 0,808 aus dem Chi²-Test.

Ein gruppiertes Säulendiagramm dient zur Darstellung der bivariaten Häufigkeitsverteilung zweier dichotomer Variablen. Hier lässt sich zudem erkennen, dass zwischen Frauen (w) und Männern (m) wenig Veränderung im Verhältnis karies (yes) / nicht karies (no) besteht. Der Kariesbefall in dieser Untersuchung ist damit unabhängig vom Geschlecht.

***graph bar, over(karies) over(geschl)***




## 6.2. Ordinale Variable

Eine bivariate Häufigkeitsverteilung zweier ordinaler Variablen ***zpfl*** und ***dmfkat*** in der Datei kiga_5.dta erhält man mit

***tab zpfl dmfkat , chi taub V***

```
tab zpfl dmfkat , chi taub V
```

| zpfl | Kategorien 1 | 2 | 3 | Total |
|---|---|---|---|---|
| sehr gut | 22 | 0 | 0 | 22 |
| mittel | 95 | 36 | 28 | 159 |
| schlecht | 2 | 3 | 15 | 20 |
| Total | 119 | 39 | 43 | 201 |

```
          Pearson chi2(4) =  52.3350   Pr = 0.000
               Cramér's V =   0.3608
          Kendall's tau-b =   0.4259  ASE = 0.043
```

Ein Zusammenhang zwischen Zahnpflege ***zpfl*** und den dmf-Kategorien ***dmfkat*** ist signifikant mit einer mittleren Stärke von etwa 0,4 .
Die Prüfgröße z = taub / ASE ist standardnormalverteilt, der kritische Wert $z_C$ = 1,96 .Wegen z = 0,4259 / 0,043 = 9,9 ist Kendall's taub signifikant von Null verschieden (9,9 > 1,96). Der zugehörige p-Wert = 0,000 .

Mit ***nptrend*** kann getestet werden, ob die Karies zunimmt mit steigender Verschlechterung der Zahnpflege.

***nptrend dmfkat, by(zpfl)***

```
nptrend dmfkat, by(zpfl)

        zpfl        score          obs    sum of ranks
    sehr gut            1           22            1320
      mittel            2          159           15744
    schlecht            3           20            3237

           z  =  6.39
Prob > |z| = 0.000
```

Auch dieser Test ist signifikant.

Interessiert dagegen der Zusammenhang zwischen ***zpfl*** und ***alter*** , erhält man einen nichtsignifikanten Zusammenhang. Kendall's taub ist nur zufällig von Null verschieden, z = 0,1196 / 0,064 = 1,87 < 1,96  (p = 0,062).

```
tab zpfl alter , chi taub V
```

| | alter | | | |
|---|---|---|---|---|
| zpfl | 3 | 4 | 5 | Total |
| sehr gut | 1 | 8 | 13 | 22 |
| mittel | 18 | 64 | 77 | 159 |
| schlecht | 4 | 9 | 7 | 20 |
| Total | 23 | 81 | 97 | 201 |

```
          Pearson chi2(4) =   3.6553   Pr = 0.455
               Cramér's V =   0.0954
          Kendall's tau-b =  -0.1196  ASE = 0.064
```

Ein weiteres übliches Maß für den Zusammenhang zweier ordinaler Variablen ist der **Rangkorrelationskoeffizient von Spearman** (Spearman's rank correlation coefficient). Man erhält ihn in Stata für die Var ***zpfl***  und ***dmfkat*** mit dem Kommando
***spearman zpfl dmfkat***

```
. spearman zpfl dmfkat

 Number of obs =     201
Spearman's rho =       0.4510

Test of Ho: zpfl and dmfkat are independent
    Prob > |t| =       0.0000
```

Auch dieser Koeffizient zeigt einen signifikanten Zusammenhang mittlerer Stärke.

## 6.3. Metrische Variable

Die Stärke eines linearen Zusammenhangs zwischen zwei metrischen Variablen kann durch den Korrelationskoeffizienten nach Pearson (r) berechnet werden. Für die Variablen ***pefr*** und ***ht*** in lung1984.dta findet man z.B. mit
***correlate pefr ht***

```
. correlate pefr ht
(obs=102)
```

| | pefr | ht |
|---|---|---|
| pefr | 1.0000 | |
| ht | 0.6544 | 1.0000 |

ein r = 0,6544 und damit eine mittlere Stärke.

Möchte man das Signifikanzniveau des Korrelationskoeffizienten ermitteln, dann mit ***pwcorr pefr ht, sig***

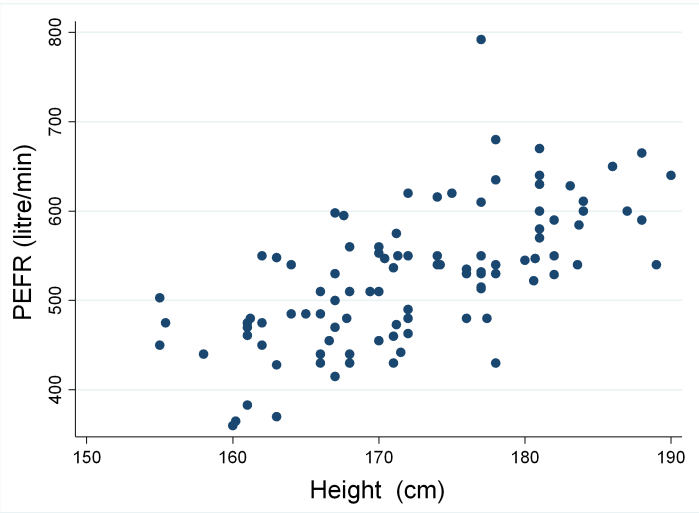
```
pwcorr pefr ht, sig
```

| | pefr | ht |
|---|---|---|
| pefr | 1.0000 | |
| ht | 0.6544<br>0.0000 | 1.0000 |

Mit p = 0,000 ist der Korrelationskoeffizient signifikant von Null verschieden.

Die **Art** des Zusammenhangs (linear oder nicht linear) erkennt man recht gut im Scatter-Plot mit ***graph twoway scatter pefr ht***

```
graph twoway scatter pefr ht
```




oder kurz:
***scatter pefr ht***

In diesem Fall kann ein linearer Zusammenhang der Form $y_i^* = a + b \cdot x_i$ angenommen werden. Die Regressionskoeffizienten **a** und **b** erhält man mit:

***regress pefr ht***

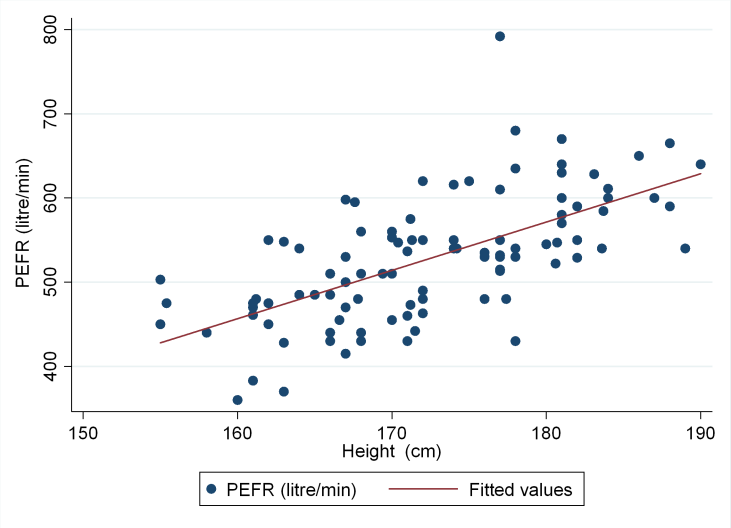
```
regress pefr ht
```

| Source | SS | df | MS | Number of obs | = | 101 |
|---|---|---|---|---|---|---|
| | | | | F(1, 99) | = | 70.50 |
| Model | 227420.91 | 1 | 227420.91 | Prob > F | = | 0.0000 |
| Residual | 319353.553 | 99 | 3225.79347 | R-squared | = | 0.4159 |
| | | | | Adj R-squared | = | 0.4100 |
| Total | 546774.464 | 100 | 5467.74464 | Root MSE | = | 56.796 |

| pefr | Coef. | Std. Err. | t | P>\|t\| | [95% Conf. | Interval] |
|---|---|---|---|---|---|---|
| ht | 5.740113 | .6836338 | 8.40 | 0.000 | 4.383635 | 7.09659 |
| _cons | -461.8401 | 118.0378 | -3.91 | 0.000 | -696.0527 | -227.6274 |

Die Regressionsgleichung lautet nun: $y_i^* = -461{,}84 + 5{,}74 \cdot x_i$
mit den Koeffizienten a = - 461,84 und b = 5,74

***scatter pefr ht || lfit pefr ht*** liefert die Regressionsgerade in der Grafik.




Bei einer Größe von 180 cm erwartet man im Mittel eine Peak expiratory flow rate (PEFR) von etwa 570 l/min. Den genauen Schätzwert erhält man durch Einsetzen in die Regressionsgleichung $y_i^* = -461{,}84 + 5{,}74 \cdot x\ 180 = 571{,}36$.
Mit Bestimmtheitsmaß $R^2 = 0{,}416$ kann von einem brauchbaren linearen Modell ausgegangen werden.
Mit F (1, 99) = 70,5 ist der F-Test signifikant (p = 0,000) und damit der Anstieg der Regressionsgeraden definitiv von Null verschieden.

Zu den Modellvoraussetzungen für eine lineare Einfachregression zählen:

1. Linearer Zusammenhang --> grafische Prüfung mit Scatterplot und Wert von $R^2$ .
2. Normalverteilung der abhängigen Variablen -> grafisch Q-Q-Plot und Shapiro-Wilk Test
3. Normalverteilung der Residuen --> grafisch Q-Q-Plot und Shapiro-Wilk Test
4. Varianzhomogenität der Residuen --> grafisch Residuen-Plot

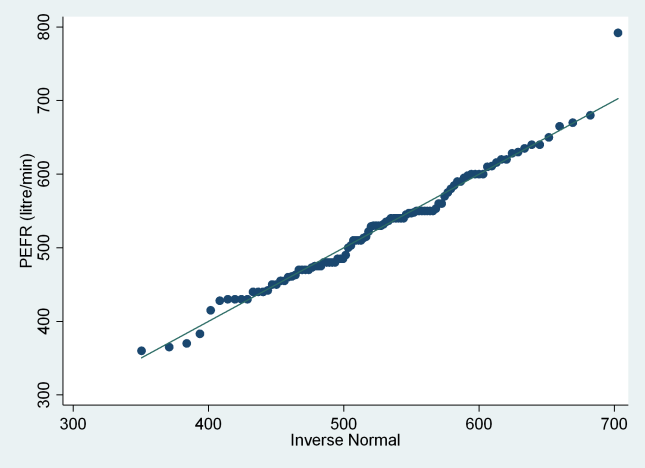
Zu 1.: siehe Grafik Seite 4 und STATA-Ausgabe Regression
Zu 2.: ***swilk pefr*** Shapiro-Wilk Test für ***pefr***
(Die Var ***pefr*** hat einen fehlenden Wert.)

```
swilk pefr
```

Shapiro-Wilk W test for normal data

| Variable | Obs | W | V | z | Prob>z |
|---|---|---|---|---|---|
| pefr | 101 | 0.98247 | 1.460 | 0.840 | 0.20054 |

Q-Q-Plot für ***pefr*** Ergebnis: Die Variable ***pefr*** ist normalverteilt.




Zu 3.
STATA speichert intern verschiedene Ergebnisse der Regression. Direkt im Anschluss an ***regress pefr ht*** erhält man mit ***predict x, residuals*** und anschließend ***swilk x*** folgende Ausgabe für die Residuen

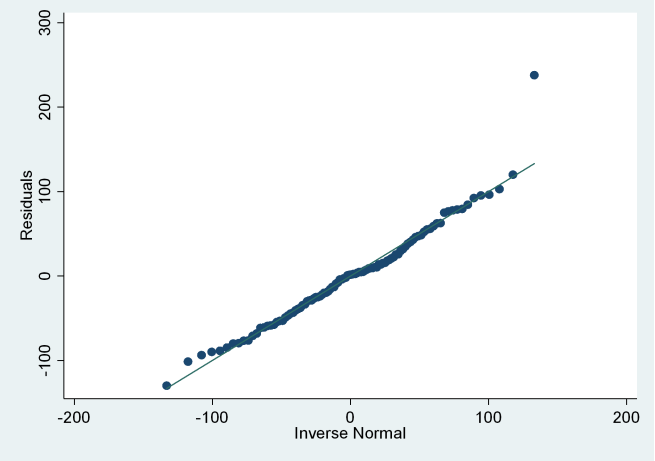
```
swilk x
```

Shapiro-Wilk W test for normal data

| Variable | Obs | W | V | z | Prob>z |
|---|---|---|---|---|---|
| x | 101 | 0.96813 | 2.653 | 2.166 | 0.01517 |

In diesem Fall sind die Residuen nicht normalverteilt, da die abhängige Var ***pefr*** einen extremen Wert aufweist.

Q-Q-Plot für ***x*** Ergebnis: Die Variable ***x*** ist nicht normalverteilt.



Da es sich nur um einen Datenpunkt von 102 Messungen handelt, wird dieser gelöscht und ein Protokollvermerk angefertigt.

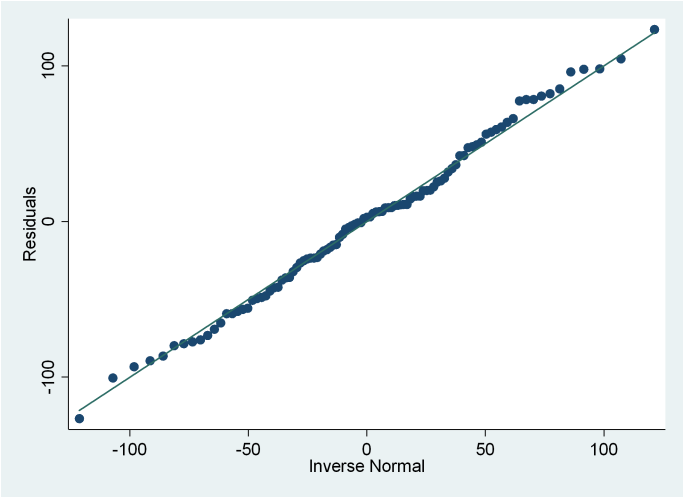
Man erhält für die verbleibenden Daten normalverteilte Residuen

```
swilk x
```

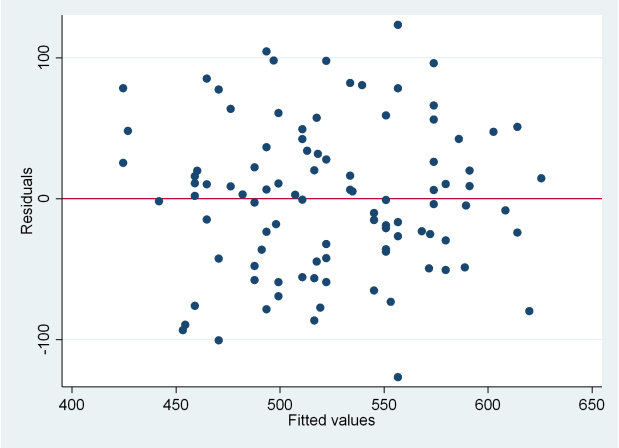
Shapiro-Wilk W test for normal data

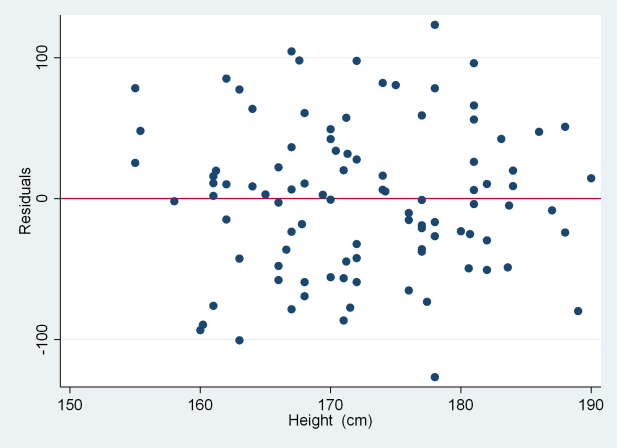
| Variable | Obs | W | V | z | Prob>z |
|---|---|---|---|---|---|
| x | 101 | 0.99178 | 0.684 | -0.843 | 0.80044 |

und einen Q-Q-Plot:




Eine weitere Voraussetzung für die Anwendung des linearen Regressionsmodells ist die Varianzhomogenität (Homoskedastizität), die man grafisch überprüfen kann, indem die Residuen gegen die geschätzten Werte (fitted values) aufgetragen werden.




Eine gleiche Grafik erhält man mit Residuen gegen ***ht*** . Es ist kein Muster zu erkennen, das auf Heteroskedastizität (Abweichung von der Homoskedastizität) hindeuted. Zusätzlich kann der White-Test durchgeführt werden, der bei Signifikanz eine Abweichung von der Homoskedastizität anzeigt.

```
. estat imtest, white

White's test for Ho: homoskedasticity
         against Ha: unrestricted heteroskedasticity

         chi2(2)      =      0.96
         Prob > chi2  =    0.6179
```

Bei einem p-Wert von 0,6179 ist eine Abweichung von der Homoskedastizität nicht erkennbar.

## STATA - Kommandos für bivariate Datenbeschreibung

***tabulate***

***tab v1 v2, expected nofreq***

***tab v1 v2, chi exact taub V***

***graph bar, over(v1) over(v2)***

***nptrend v1, by(v2)***

***spearman v1 v2***

***correlate v1 v2***

***pwcorr v1 v2, sig***

***graph twoway scatter v1 v2***

***regress v1 v2***

***scatter pefr v2 || lfit v1 v2***

***swilk***