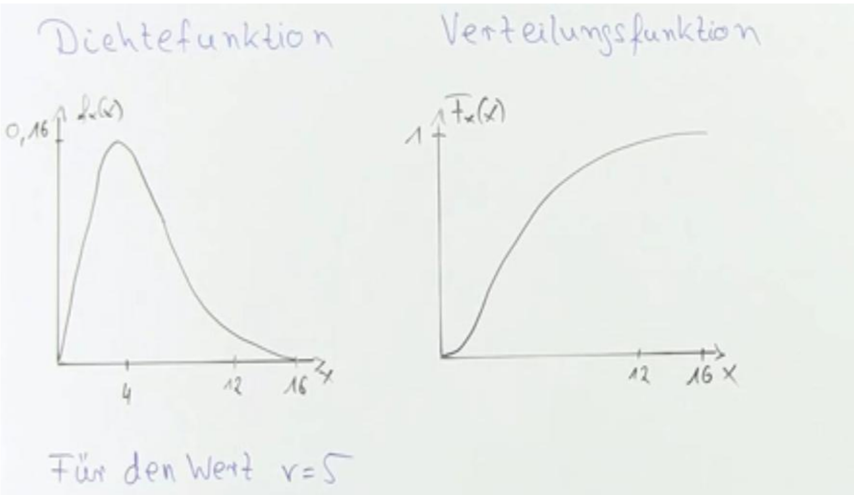
Chi-Quadrat
Anpassungstest &Unabhängigkeitstest

- Induktive Statistik
- Eingeführt von Helmert (1876) &Pearson (1900)

Formale Definition:

Sind $X_i\,(i=1,2,\ldots,n)$ stochastisch unabhängige und $N(0;1)$-verteilte ZV, dann genügt die ZV $Y^2 = X_1^2 + X_2^2 + \ldots + X_n^2$, die sich aus der Summe der n quadrierten ZV X_i^2 ergibt, einer Chi-Quadrat-Verteilung mit $v = n$ Freiheitsgraden.

Dichte & Verteilungsfunktion (für v=5)



Die Dichte f_n der χ_n^2-Verteilung mit n Freiheitsgraden hat die Form:

$$f_n(x) = \begin{cases} \dfrac{x^{\frac{n}{2}-1} e^{-\frac{x}{2}}}{2^{\frac{n}{2}} \Gamma(\frac{n}{2})} & x > 0 \\ 0 & x \le 0 \end{cases}$$

Dabei steht $\Gamma(r)$ für die Gammafunktion. Die Werte von $\Gamma(\frac{n}{2})$ kann man auch berechnen mit

$$\Gamma(\tfrac{1}{2}) = \sqrt{\pi}\,, \quad \Gamma(1) = 1\,,$$
$$\Gamma(r+1) = r \cdot \Gamma(r) \ \text{ mit } r \in \mathbb{R}^+.$$

Erwartungswert; Varianz und die Aproximationsmöglichkeiten:

$$E(Y) = v$$

$$Var(Y) = 2v$$

Approximationsmöglichkeiten:

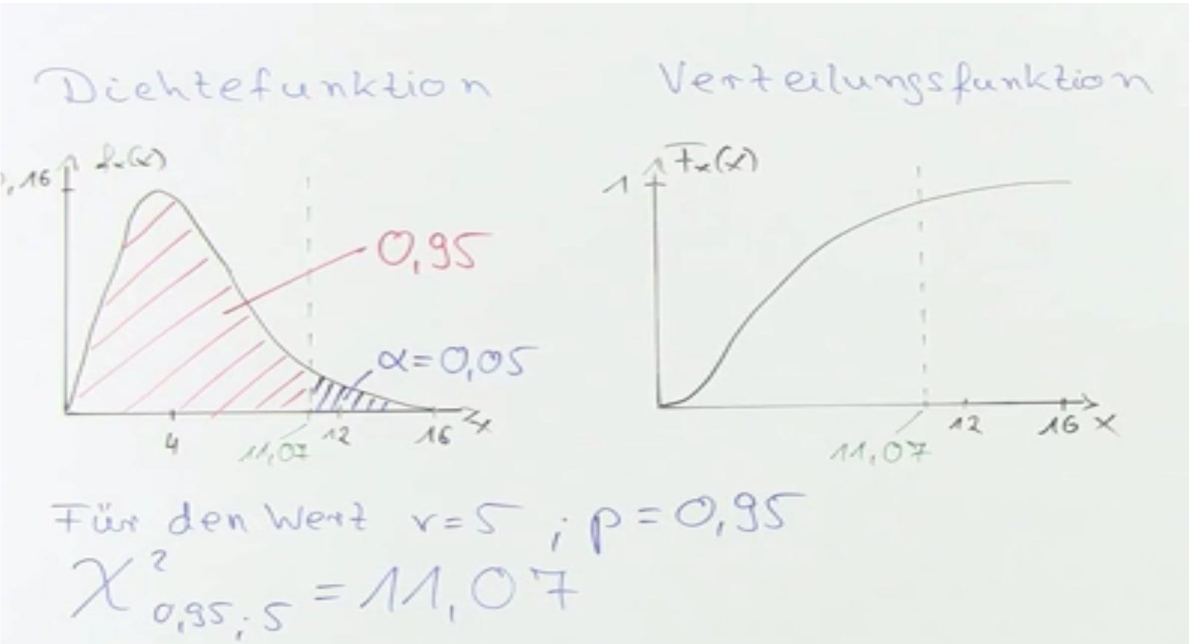
$$Y \sim \chi^2(v)$$

a) $v \geq 30 \Rightarrow Z = \sqrt{2Y} - \sqrt{2v-1}$

näherungsweise standardnormalverteilt

b) $v \geq 100 \Rightarrow N(v; 2v)$

Das Quantil teilt die Dichte-und Verteilungsfunktion in 2 Teilbereiche.
Einmal alle Werte zwischen 0-11,07 die mit einer Möglichkeit von 95% auftreten und alle Werte größer 11,06 die mit einer Möglichkeit von 5% auftreten.
Diese 5% nennt man auch Signifikanzniveau oder Irrtumsmöglichkeit. Den Wert 11,07 kann man entweder von der Verteilungsfunktion ablesen oder von der Chi-Quadrat-Verteilungstabelle mit v=5 und p=0,95.



Chi-Quadrat-Anpassungstest

- Vergleich unserer Stichprobe mit einer hypothetischen
  -> also liegt hier eine bestimmte Verteilung vor?
- Vorgehensweise:

1. Aufgabenstellung

-> Zufallsstichprobe $(X_1, X_2, \ldots, X_n)$

-> Überprüfung, ob Verteilung $F_x(x)$ mit hypothetischer Verteilung $F_0(x)$ übereinstimmt

2. Voraussetzungen

-> zufällige Entnahme der Stichprobenelemente

3. H_0

$H_0: F_x(x) = F_0(x)$

$H_0: f_x(x) = f_0(x)$

4. Testgröße

$$\chi^2_* = \sum_{j=1}^{m} \frac{(ho_j - he_j)^2}{he_j} \qquad he_j = n\, f_x(x_j)$$

$$\chi^2_* \sim \chi^2(m-1)$$

5. Annahmekennzahlen

$c_0 = \chi^2(1-\alpha; m-1)$

6. Testentscheidung

Ablehnung H_0, falls:

$\chi^2_* > c_0$

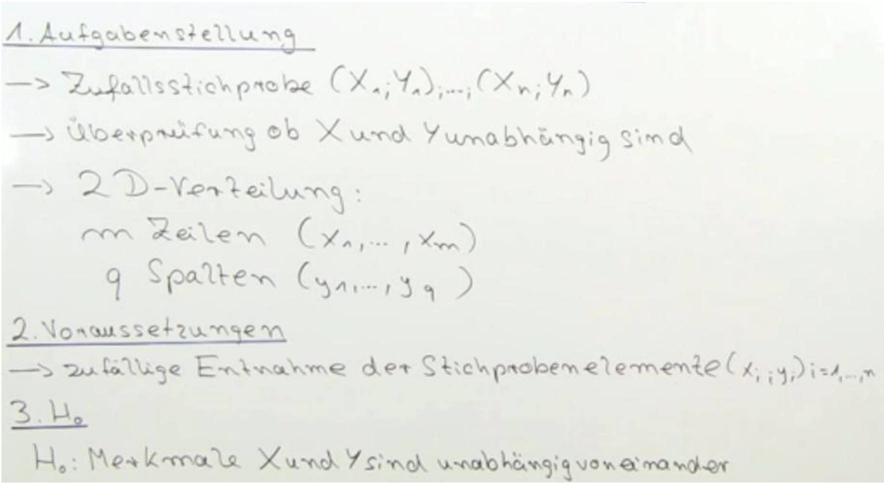
ho_j= beobachtete Werte

he_j = erwartete Werte

m= Anzahl der Klassen (Z.B. beim Würfel gibt es 6 mögliche Ereignisse, also wäre hier m=6)

Chi-Quadrat-Unabhängigkeitstest

- Wollen Herausfinden ob die Merkmale X und Y stochastisch unabhängig sind
- Vorgehensweise:



1. Aufgabenstellung

—> Zufallsstichprobe $(X_1; Y_1);\dots;(X_n; Y_n)$

—> Überprüfung ob X und Y unabhängig sind

—> 2D-Verteilung:

m Zeilen $(x_1, \dots, x_m)$

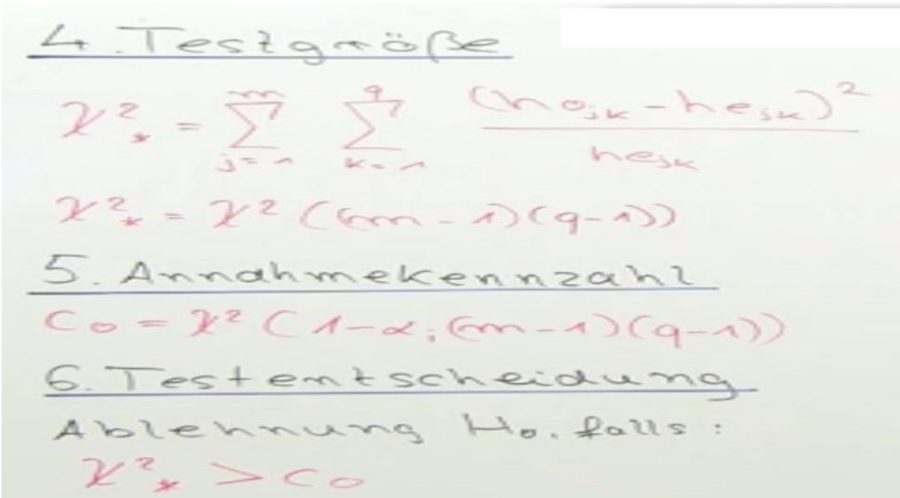
q Spalten $(y_1, \dots, y_q)$

2. Voraussetzungen

—> zufällige Entnahme der Stichprobenelemente $(x_i; y_i)\ i=1,\dots,n$

3. H_0

H_0: Merkmale X und Y sind unabhängig voneinander



4. Testgröße

$$\chi^2_* = \sum_{j=1}^{m} \sum_{k=1}^{q} \frac{(ho_{jk} - he_{jk})^2}{he_{jk}}$$

$$\chi^2_* = \chi^2((m-1)(q-1))$$

5. Annahmekennzahl

$$c_0 = \chi^2(1-\alpha; (m-1)(q-1))$$

6. Testentscheidung

Ablehnung H_0 falls:

$$\chi^2_* > c_0$$

ho_{jk} = beobachtete Werte

he_{jk} = erwarteten Werte (erst, wenn die Randsummen gegeben sind können wir diese Werte berechnen. Weil wir unteranderem davon ausgehen, dass die erwarteten Werte stochastisch unabhängig sind)

formal : $he_{jk} = (h_{+k} * h_{j+}) /n$, wobei h_{+k} und h_{j+} , die oben erwähnten Randsummen sind.

m= Anzahl der Klassen für das Merkmal X

q = Anzahl der Klassen für das Merkmal Y