Chi-Quadrat-Verteilung

Die Verteilung einer Summe $X_1^2 + X_2^2 + \ldots + X_n^2$, wobei $X_1, \ldots, X_n$ unabhängige standardnormalverteilte Zufallsvariablen sind, heißt χ^2-Verteilung mit n Freiheitsgraden.

Eine $N(0,1)$-verteilte Zufallsvariable hat die Dichte

$$\varphi(x) = \frac{1}{\sqrt{2\pi}}\, e^{-\frac{1}{2}x^2}$$

Sei nun $\chi_1^2 = X^2$. Für die Verteilungsfunktion gilt:

$$F_{\chi_1^2}(x) = P(X^2 < x) = P(-\sqrt{x} < X < \sqrt{x}) = 2\,\Phi(\sqrt{x}) - 1$$

und für die Dichte:

$$f_{\chi_1^2}(x) = F'_{\chi_1^2}(x) = 2\,\varphi(\sqrt{x})\,\frac{1}{2\sqrt{x}} = \frac{1}{\sqrt{2\pi}} \cdot \frac{1}{\sqrt{x}} \cdot e^{-\frac{x}{2}} = \frac{1}{\sqrt{2\pi}} \cdot x^{-\frac{1}{2}} \cdot e^{-\frac{x}{2}}, \quad x > 0$$

Tasten wir uns weiter vor: $\chi_2^2 = X_1^2 + X_2^2$

$$\begin{aligned}
f_{\chi_2^2}(x) &= \int_0^x f_{\chi_1^2}(y) \cdot f_{\chi_1^2}(x-y)\; dy \qquad \text{(Faltung)} \\
&= \int_0^x \frac{1}{\sqrt{2\pi}}\, y^{-\frac{1}{2}}\, e^{-\frac{y}{2}} \cdot \frac{1}{\sqrt{2\pi}}\, (x-y)^{-\frac{1}{2}}\, e^{-\frac{x-y}{2}}\; dy \\
&= \frac{1}{2\pi}\, e^{-\frac{x}{2}} \underbrace{\int_0^x \frac{1}{\sqrt{y(x-y)}}\; dy}_{} \qquad \text{(Substitution } y = x(1-u^2)\text{)} \\
&\qquad\qquad = \int_1^0 \frac{-2xu\; du}{\sqrt{(1-u^2)x^2u^2}} = \int_0^1 \frac{2\; du}{\sqrt{1-u^2}} = 2\arcsin u\Big|_0^1 = \pi \\
&= \frac{1}{2}\, e^{-\frac{x}{2}}
\end{aligned}$$

Das Weitere ist wesentlich einfacher: $\chi_3^2 = X_1^2 + X_2^2 + X_3^2$

$$\begin{aligned}
f_{\chi_3^2}(x) &= \int_0^x f_{\chi_1^2}(y) \cdot f_{\chi_2^2}(x-y)\; dy \\
&= \int_0^x \frac{1}{\sqrt{2\pi}}\, y^{-\frac{1}{2}}\, e^{-\frac{y}{2}} \cdot \frac{1}{2}\, e^{-\frac{x-y}{2}}\; dy \\
&= \frac{1}{\sqrt{2\pi} \cdot 2}\; e^{-\frac{x}{2}} \int_0^x y^{-\frac{1}{2}}\; dy \\
&= \frac{1}{\sqrt{2\pi}}\; x^{\frac{1}{2}}\, e^{-\frac{x}{2}}
\end{aligned}$$

χ^2-Verteilung Fortsetzung

$\chi_4^2 = X_1^2 + X_2^2 + X_3^2 + X_4^2$.

$$f_{\chi_4^2}(x) = \int_0^x f_{\chi_2^2}(y) \cdot f_{\chi_2^2}(x-y)\ dy$$

$$= \int_0^x \frac{1}{2}\, e^{-\frac{y}{2}} \cdot \frac{1}{2}\, e^{-\frac{x-y}{2}}\ dy$$

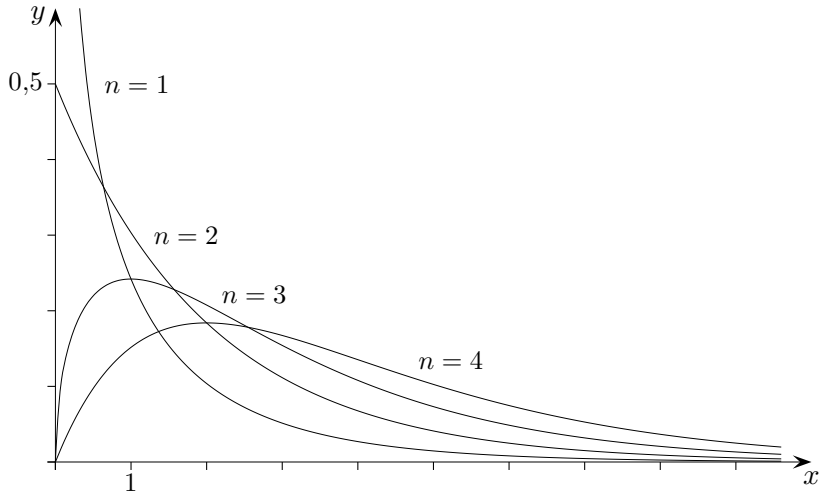
$$= \frac{1}{4}\, x\, e^{-\frac{x}{2}}$$

Wenn wir diese Berechnung weiter fortsetzen, wird das Regelmäßige erkennbar.

Die Dichte der χ_n^2-Verteilung kann schließlich geschlossen durch

$$f_{\chi_n^2}(x) = \frac{1}{2^{\frac{n}{2}}\,(\frac{n}{2}-1)!}\, x^{\frac{n}{2}-1}\, e^{-\frac{x}{2}}, \qquad x > 0$$

angegeben werden.
Dabei ist $n! = n \cdot (n-1)!$ und $(-\frac{1}{2})! = \sqrt{\pi}$.
Eine weitere übersichtliche Darstellung benutzt die Γ-Funktion.



χ^2-Anpassungstest

Ist ein Stichprobenergebnis mit einer gegebenen Verteilung verträglich?

X_1 sei eine binomialverteilte Zufallsvariable. Dann ist $Z = \dfrac{X_1 - \mu}{\sigma}$ näherungsweise $N(0,1)$-verteilt und somit

$$Z^2 = \frac{(X_1 - \mu)^2}{\sigma^2}$$

χ_1^2-verteilt.

Z^2 kann mit $X_2 = n - X_1$ und $p_2 = 1 - p_1$ umgeformt werden:

$$\begin{aligned}
Z^2 &= \frac{(X_1 - np_1)^2}{np_1(1-p_1)} \\
&= \frac{(X_1 - np_1)^2(p_2 + p_1)}{np_1p_2} \\
&= \frac{(X_1 - np_1)^2}{np_1} + \frac{(X_1 - np_1)^2}{np_2} \\
&= \frac{(X_1 - np_1)^2}{np_1} + \frac{(X_1 - n(1-p_2))^2}{np_2} \\
&= \frac{(X_1 - np_1)^2}{np_1} + \frac{(X_1 - (X_1 + X_2) + np_2)^2}{np_2} \\
&= \frac{(X_1 - np_1)^2}{np_1} + \frac{(X_2 - np_2)^2}{np_2}
\end{aligned}$$

Allgemein:
Tritt von k Ereignissen A_i jeweils eines mit der Wahrscheinlichkeit p_i ein und zählt X_i die Häufigkeit von A_i, dann ist die Testgröße

$$T = \frac{(X_1 - np_1)^2}{np_1} + \frac{(X_2 - np_2)^2}{np_2} + \ldots + \frac{(X_k - np_k)^2}{np_k}$$

χ_{k-1}^2-verteilt.

Da $V(X_i) = E(X_i - np_i)^2 = np_i(1-p_i)$ ist, folgt für den Erwartungswert von T:

$$E(T) = \frac{np_1(1-p_1)}{np_1} + \frac{np_2(1-p_2)}{np_2} + \ldots + \frac{np_k(1-p_k)}{np_k} = k - 1$$

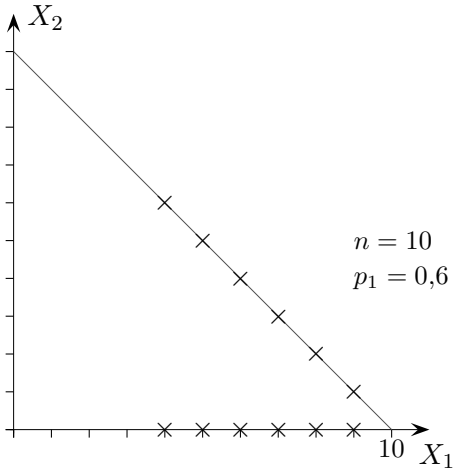
$k-1$ (Anzahl der Freiheitsgrade) der X_i können frei gewählt werden, X_k ergibt sich dann aus $X_1 + X_2 + \ldots + X_k = n$.

Chiquadrat-Test

Die verallgemeinerungsfähige Überlegung, die diesem Test zu Grunde liegt, soll erhellt werden.
Wir betrachten einen Einzelversuch mit den beiden Ausgängen A_1 und A_2, die mit den Wahrscheinlichkeiten p_1 und p_2 angenommen werden ($p_1 + p_2 = 1$). Für die n-malige Wiederholung des Einzelversuchs zählen die Zufallsvariablen X_1 und X_2 das Auftreten von A_1 bzw. A_2 ($X_1 + X_2 = n$). Es liegt also eine Bernoullikette vor, allgemein eine multinomiale Verteilung.

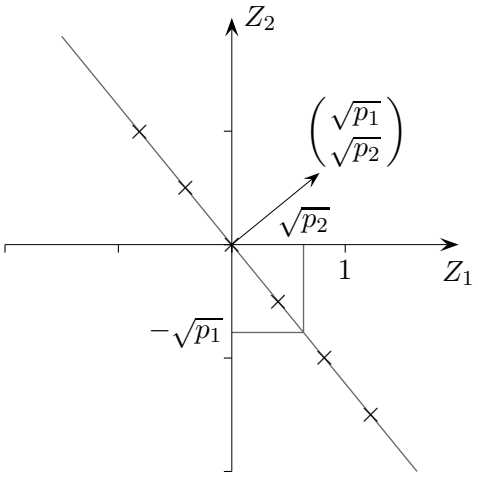
Die Ergebnisse $\begin{pmatrix} X_1 \\ X_2 \end{pmatrix}$

der Zufallsversuche liegen auf der Geraden $x + y = n$.



Wir gehen nun zum Zufallsvektor $\begin{pmatrix} Z_1 \\ Z_2 \end{pmatrix} = \begin{pmatrix} \frac{X_1 - n \cdot p_1}{\sqrt{n \cdot p_1}} \\ \frac{X_2 - n \cdot p_2}{\sqrt{n \cdot p_2}} \end{pmatrix}$

über.



Die Varianz von Z_1 lautet:

$$V(Z_1) = \frac{1}{n \cdot p_1} \cdot n \cdot p_1 \cdot p_2 = p_2$$

entsprechend:

$$V(Z_2) = p_1$$

Hieraus ist zu erkennen, dass $\begin{pmatrix} Z_1 \\ Z_2 \end{pmatrix}$ (genähert) normalverteilt ist, bezogen auf eine Achseneinteilung auf der Geraden, $\mu = 0$, $\sigma = 1$ (Pythagoras und $p_1 + p_2 = 1$). Der Freiheitsgrad 1 wird verständlich.
Anders formuliert: $\sqrt{Z_1^2 + Z_2^2}$ ist $\mathcal{N}(0, 1)$-verteilt. Die Chiquadrat-Testgröße $Z_1^2 + Z_2^2$ ist daher χ_1^2-verteilt.

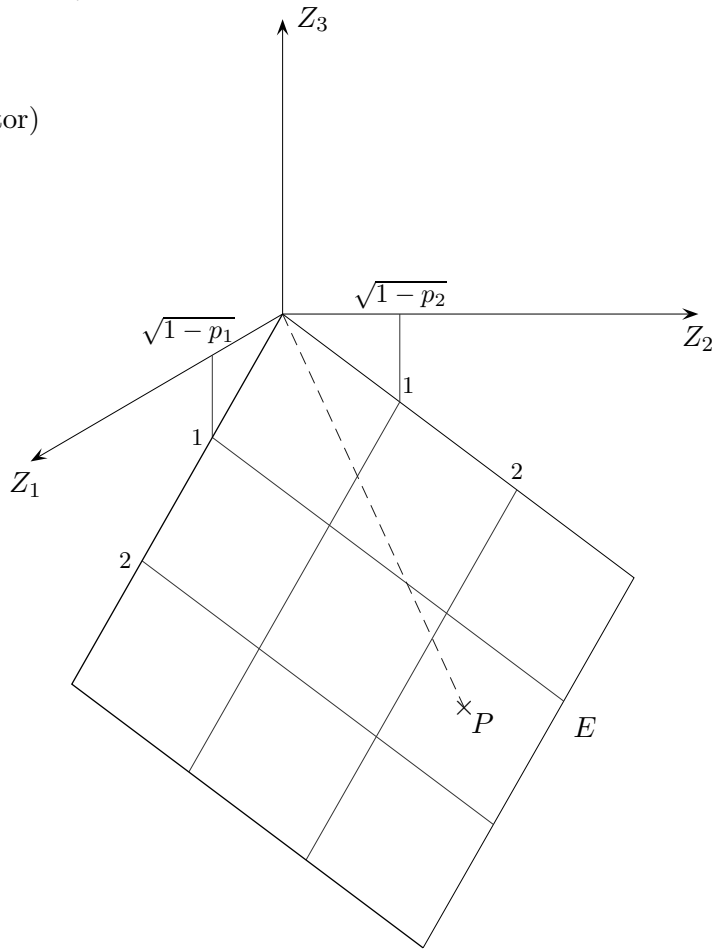
Chiquadrat-Test Fortsetzung

Wir können uns auch den Test mit 2 Freiheitsgraden veranschaulichen. Hierzu betrachten wir einen Einzelversuch mit den 3 Ausgängen A_1, A_2 und A_3, die mit den Wahrscheinlichkeiten p_1, p_2 und p_3 angenommen werden ($p_1+p_2+p_3 = 1$, genau ein A_i tritt ein). Für die n-malige Wiederholung des Einzelversuchs zählen die Zufallsvariablen X_1, X_2 und X_2 das Auftreten von A_1, A_2 und A_3, $X_1 + X_2 + X_3 = n$.

Der Zufallsvektor $\overrightarrow{OP} = \begin{pmatrix} Z_1 \\ Z_2 \\ Z_3 \end{pmatrix} = \begin{pmatrix} \frac{X_1 - n \cdot p_1}{\sqrt{n \cdot p_1}} \\ \frac{X_2 - n \cdot p_2}{\sqrt{n \cdot p_2}} \\ \frac{X_3 - n \cdot p_3}{\sqrt{n \cdot p_3}} \end{pmatrix}$

liegt stets in einer Ebene mit der HNF:
($\overrightarrow{OP}$ steht senkrecht auf dem Normalenvektor)

$$E: \quad \begin{pmatrix} \sqrt{p_1} \\ \sqrt{p_2} \\ \sqrt{p_3} \end{pmatrix} \vec{x} = 0$$



Die Varianz von Z_i beträgt:

$$V(Z_i) = \frac{1}{n \cdot p_i} \cdot n \cdot p_i \cdot (1 - p_i) = 1 - p_i$$

Um die Verteilung des Längenquadrats $|\overrightarrow{OP}|^2 = Z_1^2 + Z_2^2 + Z_3^2$ (= Chiquadrat-Testgröße)
zu ermitteln, kann durch einen Vergleich mit der multinomialen Verteilung von $\overrightarrow{OP}$ gezeigt werden, dass die beiden rechtwinkligen Koordinaten von P, hinsichtlich der Ebene E, $\mathcal{N}(0,1)$-verteilt sind ($n \to \infty$).
Damit ist $|\overrightarrow{OP}|^2$ χ^2_2-verteilt.

Die heuristischen Überlegungen wären noch treffender, wenn sich die Einheiten des Ebenen-Koordinatensystems durch senkrechte Projektion der Standardabweichungen von Z_1 und Z_2 ergäben und die Koordinatenachsen auf E (als Schnittgeraden mit der xz- bzw. yz-Ebene) orthogonal wären. Beides ist jedoch nicht der Fall.
Es kommt nicht auf die Lage der Koordinatenachsen in E an: Das Längenquadrat, d. h. das Skalarprodukt $\overrightarrow{OP}^2$, ist stets gleich. So kann eine zweidimensionale Standardnormalverteilung mit einer orthogonalen Matrix so auf E abgebildet werden, dass der Vektor $\begin{pmatrix} 0 \\ 0 \\ 1 \end{pmatrix}$ in $\begin{pmatrix} \sqrt{p_1} \\ \sqrt{p_2} \\ \sqrt{p_3} \end{pmatrix}$ übergeht.
Die Übereinstimmung (für $n \to \infty$) dieser Verteilung mit der Verteilung von $\begin{pmatrix} Z_1 \\ Z_2 \\ Z_3 \end{pmatrix}$ wird mit Hilfe von charakteristischen Funktionen nachgewiesen.