1.4 Stichproben aus einer Normalverteilung

Die Normalverteilung ist wohl das am stärksten verbreitete Modell. Stichproben daraus führen zu nützlichen Eigenschaften der Statistiken und ergeben bekannte Stichprobenverteilungen (χ^2_p, t_p, $F_{p,q}$).

1.4.1 Eigenschaften des Stichprobenmittels und der Stichprobenvarianz

Definition 1.4.1: Die Chiquadrat-Verteilung mit p Freiheitsgraden, χ^2_p, entspricht einer Gamma$(p/2, 2)$ Verteilung und hat somit Dichte

$$f(x|p) = \frac{1}{\Gamma(p/2)2^{p/2}} x^{p/2-1} \exp(-x/2) I_{[0,\infty)}(x)\,, \qquad p = 1, 2, \ldots\,.$$

Hierbei wird der Parameter p *Freiheitsgrad* genannt.
Spezielle Wahl $p = 1$:

$$f(x|1) = \frac{1}{\sqrt{2\pi x}} \exp(-x/2)\,, \qquad 0 \leq x\,.$$

Lemma 1.4.1: (Eigenschaften einer χ^2 verteilten Zufallsvariablen)

(a) Falls $Z \sim N(0,1)$, dann ist $Z^2 \sim \chi^2_1$,

(b) Falls $X_1, \ldots, X_n$ unabhängig mit $X_i \sim \chi^2_{p_i}$, dann ist $\sum_i X_i \sim \chi^2_{\sum_i p_i}$.

Satz 1.4.1: Sei $X_1, \ldots, X_n$ eine Zufallsstichprobe aus einer $N(\mu, \sigma^2)$ Verteilung. Dann gilt:

(a) $\overline{X}$ und S^2 sind unabhängig,

(b) $\overline{X} \sim N(\mu, \sigma^2/n)$, (siehe dazu Beispiel 1.2.1)

(c) $(n-1)S^2/\sigma^2 \sim \chi^2_{n-1}$.

Für den Induktionsbeweis von Teil (c) werden wir die Aussagen des folgenden Beispiels verwenden.

Beispiel 1.4.1: Bezeichne $\overline{X}_n$ und S_n^2 die Statistiken basierend auf n Beobachtungen. Kommt eine weitere Beobachtung X_{n+1} dazu, dann resultiert

$$\overline{X}_{n+1} = \frac{1}{n+1}\sum_{i=1}^{n+1} X_i = \frac{1}{n+1}\left(\sum_{i=1}^{n} X_i + X_{n+1}\right) = \frac{1}{n+1}\left(n\overline{X}_n + X_{n+1}\right) .$$

Weiters gilt wegen Satz 1.2.1(b), dass $(n-1)S_n^2 = \sum_{i=1}^n X_i^2 - n\overline{X}_n^2$ hält. Also

$$\begin{aligned}
nS_{n+1}^2 &= \sum_{i=1}^{n+1} X_i^2 - (n+1)\overline{X}_{n+1}^2 \\
&= \sum_{i=1}^{n} X_i^2 + X_{n+1}^2 - (n+1)\left[\frac{1}{n+1}\left(n\overline{X}_n + X_{n+1}\right)\right]^2 \\
&= \sum_{i=1}^{n} X_i^2 - n\overline{X}_n^2 + n\overline{X}_n^2 + X_{n+1}^2 - \frac{1}{n+1}\left(n^2\overline{X}_n^2 + 2nX_{n+1}\overline{X}_n + X_{n+1}^2\right) \\
&= (n-1)S_n^2 + \frac{n}{n+1}\left(\overline{X}_n^2 + X_{n+1}^2 - 2X_{n+1}\overline{X}_n\right) \\
&= (n-1)S_n^2 + \frac{n}{n+1}\left(X_{n+1} - \overline{X}_n\right)^2 .
\end{aligned}$$

Beweis: OBdA nehmen wir $\mu = 0$ und $\sigma^2 = 1$ an.

ad (a): zeige dass $\overline{X}$ und S^2 Funktionen von unabhängigen Zufallsvektoren sind.

1. Schreibe dazu S^2 als Funktion nur von $(X_2 - \overline{X}, X_3 - \overline{X}, \ldots, X_n - \overline{X})$.
2. Zeige dass $\overline{X}$ und $(X_2 - \overline{X}, X_3 - \overline{X}, \ldots, X_n - \overline{X})$ unabhängig sind, d.h. dass deren gemeinsame Dichte entsprechend faktorisiert.

ad (b): bereits im Beispiel 1.2.1 gezeigt.

ad (c): Induktion

1. Betrachte $n = 2$ und zeige $S_2^2 \sim \chi_1^2$.
2. Zeige unter der Annahme, es gelte für $n = k$ gleich $(k-1)S_k^2 \sim \chi_{k-1}^2$, dass damit für $n = k + 1$ gleich $kS_{k+1}^2 \sim \chi_k^2$ folgt.

1.4.2 Hergeleitete Verteilungen: Student's t und Snedecor's F

Mit Satz 1.4.1 ist bekannt, dass für eine Zufallsstichprobe aus $N(\mu, \sigma^2)$ gilt:

$$\overline{X} \sim N(\mu, \sigma^2/n) \qquad \text{und} \qquad (n-1)S^2/\sigma^2 \sim \chi^2_{n-1}\,.$$

In der Praxis ist aber σ^2 unbekannt. Um eine Idee über die Variabilität von $\overline{X}$ zu bekommen (als Schätzer für μ), muss diese Varianz geschätzt werden. Dieser Punkt wurde erstmals von W. S. Gosset (publizierte unter Pseudonym *Student*) anfangs 1900 aufgegriffen (Biometrika, 1908). Er untersuchte die Verteilung von

$$\frac{\overline{X}-\mu}{S/\sqrt{n}} \qquad \text{anstatt von} \qquad \frac{\overline{X}-\mu}{\sigma/\sqrt{n}} \sim N(0,1)\,.$$

Diese Größe bildet die Basis einer statistischen Analyse von μ falls σ^2 unbekannt.

Es gilt

$$\frac{\overline{X}-\mu}{S/\sqrt{n}} = \frac{\frac{\overline{X}-\mu}{\sigma/\sqrt{n}}}{\sqrt{S^2/\sigma^2}} = \frac{N(0,1)}{\sqrt{\chi^2_{n-1}/(n-1)}}.$$

Der Zähler ist $N(0,1)$-verteilt und der Nenner $\sqrt{\chi^2_{n-1}/(n-1)}$ ist unabhängig vom Zähler. Student interessierte sich also u.a. für die Verteilung von

$$\frac{U}{\sqrt{V/p}}$$

mit $U \sim N(0,1)$, $V \sim \chi^2_p$, und U, V unabhängig.

Dies ergibt Student's t-Verteilung.

Definition 1.4.2: Sei $X_1, \ldots, X_n$ eine Zufallsstichprobe aus einer $N(\mu, \sigma^2)$ Verteilung. Die Zufallsvariable

$$T = \frac{\overline{X} - \mu}{S/\sqrt{n}}$$

hat Student's t Verteilung mit $(n-1)$ Freiheitsgraden. Äquivalent hat eine Zufallsvariable T Student's t Verteilung mit p Freiheitsgraden, $T \sim t_p$, falls ihre Dichte geschrieben werden kann als

$$f_T(t|p) = \frac{\Gamma((p+1)/2)}{\Gamma(p/2)} \, \frac{1}{(p\pi)^{1/2}} \, \frac{1}{(1+t^2/p)^{(p+1)/2}}, \quad t \in \mathbb{R}, \; p = 1, 2, \ldots \; .$$

Falls $p = 1$ ist dies die Dichte der Cauchy-Verteilung, die für $n = 2$ resultiert.

Satz 1.4.2: (Eigenschaften der t_p Verteilung)

(a) Für $X \sim t_p$ gilt

$$\mathsf{E}(X) = 0 \quad \text{falls } p > 1, \qquad \mathsf{var}(X) = \frac{p}{p-2} \quad \text{falls } p > 2,$$

(b) Die Momentenerzeugende Funktion existiert im allgemeinen nicht,

(c) Für p Freiheitsgrade existieren nur die ersten $p-1$ Momente, d.h. t_1 hat keinen Erwartungswert, t_2 keine Varianz, ...,

(d) Für $X_p \sim t_p$ gilt

$$\lim_{p\to\infty} f_T(t|p) = \frac{1}{\sqrt{2\pi}} \exp(-t^2/2)\,,$$

an jeder Stelle $x \in \mathbb{R}$, d.h. X_p konvergiert in Verteilung gegen die $N(0,1)$-Verteilung.

Eine weitere wichtige herleitbare Verteilung ist Snedecor's F. Die F-Verteilung, genannt nach Sir Ronald **F**isher, ergibt sich als Verteilung des Quotienten von Stichprobenvarianzen.

Sei dazu $X_1, \ldots, X_n$ eine Zufallsstichprobe aus einer $N(\mu_X, \sigma_X^2)$ Verteilung, und sei $Y_1, \ldots, Y_m$ eine zweite Zufallsstichprobe aus einer $N(\mu_Y, \sigma_Y^2)$ Verteilung unabhängig von $X_1, \ldots, X_n$.

Will man die Populationsvariabilitäten vergleichen, so könnte σ_X^2/σ_Y^2 interessieren. Information darüber steckt in S_X^2/S_Y^2. Die F Verteilung erlaubt diesen Vergleich und gibt uns die Verteilung von

$$\frac{S_X^2/S_Y^2}{\sigma_X^2/\sigma_Y^2} = \frac{S_X^2/\sigma_X^2}{S_Y^2/\sigma_Y^2} = \frac{\chi_{n-1}^2/(n-1)}{\chi_{m-1}^2/(m-1)}\,.$$

Die Quotienten S_X^2/σ_X^2 und S_Y^2/σ_Y^2 sind unabhängige, skalierte χ^2 Variablen.

Definition 1.4.3: Sei $X_1, \ldots, X_n$ eine Zufallsstichprobe aus einer $N(\mu_X, \sigma_X^2)$ Verteilung und sei $Y_1, \ldots, Y_m$ eine davon unabhängige Zufallsstichprobe aus einer $N(\mu_Y, \sigma_Y^2)$ Verteilung. Die Zufallsvariable

$$F = \frac{S_X^2/\sigma_X^2}{S_Y^2/\sigma_Y^2}$$

hat Snedecor's F Verteilung mit $(n-1)$ und $(m-1)$ Freiheitsgraden, $F \sim F_{n-1,m-1}$. Äquivalent hat eine Zufallsvariable F eine F Verteilung mit p und q Freiheitsgraden, falls ihre Dichte geschrieben werden kann als

$$f_F(x|p,q) = \frac{\Gamma((p+q)/2)}{\Gamma(p/2)\Gamma(q/2)} \left(\frac{p}{q}\right)^{p/2} \frac{x^{p/2-1}}{(1+xp/q)^{(p+q)/2}} I_{[0,\infty)}(x)\,.$$

Allgemein ist die F Verteilung die Verteilung von $(U/p)/(V/q)$, wobei U und V unabhängig sind mit $U \sim \chi_p^2$ und $V \sim \chi_q^2$.

Wie wird nun die F Verteilung verwendet, um Inferenz über das wahre Verhältnis der Populationsvarianzen zu machen?

Die Größe $(S_X^2/\sigma_X^2)/(S_Y^2/\sigma_Y^2)$ hat eine $F_{n-1,m-1}$ Verteilung. Wir berechnen

$$\mathsf{E}(F_{n-1,m-1}) = \mathsf{E}\left(\frac{\chi^2_{n-1}/(n-1)}{\chi^2_{m-1}/(m-1)}\right) = \mathsf{E}\left(\frac{\chi^2_{n-1}}{n-1}\right)\mathsf{E}\left(\frac{m-1}{\chi^2_{m-1}}\right) = 1\mathsf{E}\left(\frac{m-1}{\chi^2_{m-1}}\right).$$

Nun ist für $U \sim \chi^2_p$

$$\mathsf{E}(U^{-1}) = \frac{1}{\Gamma(p/2)2^{p/2}}\int_0^\infty x^{-1}x^{p/2-1}e^{-x/2}dx = \frac{1}{\Gamma(p/2)2^{p/2}}\int_0^\infty x^{(p/2-1)-1}e^{-x/2}dx\,.$$

Der Integrand entspricht dem Kern einer χ^2_{p-2} Dichte und es gilt somit

$$\int_0^\infty x^{(p/2-1)-1}e^{-x/2}dx = \Gamma(p/2-1)2^{p/2-1}\,.$$

Wegen $\Gamma(a) = (a-1)\Gamma(a-1)$ folgt weiters

$$\begin{aligned} \mathsf{E}(U^{-1}) &= \frac{1}{\Gamma(p/2)2^{p/2}}\Gamma(p/2-1)2^{p/2-1} = \frac{1}{2}\frac{\Gamma(p/2-1)}{\Gamma(p/2)} \\ &= \frac{1}{2}\frac{\Gamma(p/2-1)}{(p/2-1)\Gamma(p/2-1)} = \frac{1}{p-2}\,. \end{aligned}$$

Somit ist

$$\mathsf{E}(F_{n-1,m-1}) = \frac{m-1}{m-3} = \mathsf{E}\left(\frac{S_X^2/\sigma_X^2}{S_Y^2/\sigma_Y^2}\right)\,.$$

Dies ist nur dann endlich und positiv, falls $m > 3$.

Für ausreichend großes m gilt daher erwartungsgemäß

$$\frac{S_X^2/\sigma_X^2}{S_Y^2/\sigma_Y^2} \approx \frac{m-1}{m-3} \approx 1\,.$$

Satz 1.4.3: (Eigenschaften der F Verteilung)

(a) Für $X \sim F_{p,q}$ gilt

$$\begin{aligned} \mathsf{E}(X) &= \frac{q}{q-2}\,, \quad (q>2) \\ \mathsf{var}(X) &= \frac{2q^2(p+q-2)}{p(q-2)^2(q-4)}\,, \quad (q>4) \end{aligned}$$

(b) Für $X \sim F_{p,q}$ gilt $1/X \sim F_{q,p}$,

(c) Für $X \sim t_q$ gilt $X^2 \sim F_{1,q}$,

(d) Für $X \sim F_{p,q}$ gilt

$$\frac{\frac{p}{q}X}{1+\frac{p}{q}X} \sim \mathsf{Beta}(p/2, q/2)\,.$$