Bemerkung: Wesentlich schwieriger ist das Testen von Hypothesen über den Verteilungstyp, ohne den Wert der Parameter zu spezifizieren. Zum Beispiel:

$$H_0 : P_{X_1} \in \left\{\mathcal{N}_{\mu,\sigma} \;\middle|\; \mu \in \mathbb{R}, \sigma^2 > 0\right\}$$
$$H_1 : P_{X_1} \notin \left\{\mathcal{N}_{\mu,\sigma} \;\middle|\; \mu \in \mathbb{R}, \sigma^2 > 0\right\}$$

In diesem Falle sei die Testgröße

$$T(X_1, \ldots, X_n) = \sqrt{n} \sup_{t \in \mathbb{R}} \left| \hat{F}(t) - F_{\hat{\mu},\hat{\sigma}^2}(t) \right|$$

Wobei hier $F_{\hat{\mu},\hat{\sigma}^2}$ die Verteilungsfunktion zu $\mathcal{N}_{\hat{\mu},\hat{\sigma}}$ ist. Weiterhin lassen sich Verteilungsfunktionen auch anhand von Simulationen bestimmen.

- Kolmojorow-Smiruov-Test wird als extrem konservativ bezeichnet und verwirft zu selten H_0
- grafische Methode für Anpassungstest auf Normalverteilung

0.1 Zweistichproben-Test

Gegeben seien zwei konkrete Stichproben vom Umfang $n \in \mathbb{N}$ beziehungsweise $m \in \mathbb{N}$ aus zwei Grundgesamtheiten

$$x_1, \ldots, x_n$$
$$y_1, \ldots, y_m$$

Dies tritt zum Beispiel beim Vergleich zwei verschiedener Produktionsverfahren auf. Für die zugehörige mathematische Stichprobe gilt dann

$$X_1, \ldots, X_n \text{ } i.i.d \text{ gemäß } P_{\theta_1}$$
$$Y_1, \ldots, Y_m \text{ } i.i.d \text{ gemäß } P_{\theta_2}$$

Das Ziel ist nun der Vergleich von Merkmalen der Verteilungstypen wie zum Beispiel

- Erwartungswert
- Varianz
- verschiedene Quantile
- Verteilungsfunktion

Für uns soll es hier reichen die Erwartungswerte zweier normalverteilter Grundgesamtheiten mit unbekannter Varianz zu betrachten. Dies wird mithilfe eines Zweistichproben-t-Test ermöglicht. Es sei also folgender statistischer Raum gegeben

$$\left(\mathbb{R}^n \times \mathbb{R}^m, \mathcal{R}_{n+m}, \left\{\mathcal{N}_{\mu_1,\sigma^2} \otimes \mathcal{N}_{\mu_2,\sigma^2} \;\middle|\; \mu_1, \mu_2 \in \mathbb{R}, \sigma^2 > 0\right\}\right)$$

Es ergeben sich dann als Beispiel folgende einseitige Alternativhypothesen

$$H_0 : \mu_1 \leq \mu_2$$
$$H_1 : \mu_1 > \mu_2$$

Es soll dabei folgende Testgröße verwendet werden

$$T(x_1, \ldots, x_n, y_1, \ldots, y_m) = \sqrt{\frac{nm}{n+m}} \cdot \frac{\bar{x} - \bar{y}}{\sqrt{\widehat{\hat{\sigma}^2}}}$$

mit

$$\begin{aligned}\widehat{\hat{\sigma}^2} &= \frac{1}{m+n-2}\left[\sum_{i=1}^{n}(x_i - \bar{x})^2 + \sum_{i=1}^{m}(y_i - \bar{y})^2\right] \\ &= \frac{1}{m+n-2}\left[(n-1)\hat{\sigma}_x^2 + (m-1)\hat{\sigma}_y^2\right]\end{aligned}$$

Es folgt für den Fall $\mu_1 = \mu_2$ durch Rechnung

$$T(X_1, \ldots, X_n, Y_1, \ldots, Y_m) \sim t_{n+m-2}$$

Der kritische Bereich kann also analog zu den vorherigen Hypothesentests gewählt werden.

$$K = [t_{n+m-2,1-\alpha}, \infty)$$

Sollte es doch passieren, dass $\mathrm{var}(X_1) \neq \mathrm{var}(Y_1)$, so empfiehlt sich die Anwendung des Welch-Tests (hier nicht beschrieben). Ein Test zum Vergleich der Varianzen für normalverteilte Grundgesamtheiten ist der sogenannte F-Test (Erwartungswerte müssen nicht gleich sein).

1 statistische Methoden für zweidimensionale Stichproben

[1]

Man betrachtet nun sogenannte gepaarte Stichproben

$$(x_1, y_1), \ldots, (x_n, y_n) \in \mathbb{R}^{2n}$$

Es geht also inhaltlich um Paare von Messwerten zu einem bestimmten Zeitpunkt[2]. Verschiedene Fragestellungen sind nun zum Beispiel

Korrelationsanalyse: Sind X_1 und Y_1 unabhängig?

Regressionsanalyse: Wie sieht die funktionale Abhängigkeit von X_1 und Y_1 aus, wenn sie nicht unabhängig sind?

Varianzanalyse: Untersuchung von Parametervektoren.

Diskriminanzanalyse: Trennverfahren zur Zuordnung von Gruppen

Faktorenanalyse: Untersuchung von Kovarianzmatrizen

Clusteranalyse: Klassifikation von Werten

Dennoch werden wir auch hier grundsätzlich Punktschätzungen, Konfidenzschätzungen und Tests betrachten.

[1] Multivariatstatistik

[2] oder auch an einem Objekt

1.1 Regressionsanalyse

Sei $(x_1, y_1), \ldots, (x_n, y_n)$ für $n \in \mathbb{N}$ eine konkrete Stichprobe. Das Modell der zugehörigen mathematischen Stichprobe sei nun

$$(x_1, Y_1), \ldots, (x_n, Y_n)$$

wobei $Y_1, \ldots, Y_n$ Zufallsvariablen sind. Die Wahl des Regressionsansatzes sei intuitiv für alle $i = 1, \ldots, n$ gegeben.

$$y_i = \beta_1 + \beta_2 x_i + u_i$$

Damit folgt auch das stochastische Modell. Es ist ein einfaches lineares Regressionsmodell.

$$Y_i = \beta_1 + \beta_2 x_i + U_i \quad \text{für alle } i = 1, \ldots, n$$

Die $U_1, \ldots, U_n$ sind dabei i.i.d. Zufallsvariablen mit

$$\mathbb{E}U_i = 0, \qquad \operatorname{var} U_i = \sigma^2 > 0$$

Bemerkung:

- Die u_i können nicht beobachtet werden.
- Für die unbekannten Parameter $\beta_1, \beta_2, \sigma^2$ verwendet man Folgendes
  - Punkt- und Konfidenzbereichsschätzungen
  - Testen von Hypothesen (zum Beispiel über (β_1, β_2))
  - Punkt- und Konfidenzschätzungen für Prognosewerte $Y(x)$ an einer Stelle $x \in \mathbb{R}$
  - Test auf Adäquatheit des Modells (zum Beispiel goodness-of-fit-Test oder lack-of-fit-Test)

Es sei nun

$$\underline{\beta} := \begin{pmatrix} \beta_1 \\ \beta_2 \end{pmatrix}$$

Ziel ist es nun eine Punktschätzung für $\underline{\beta}$ durch die Methode der kleinsten Quadrate zu erhalten. Man wählt $\hat{\underline{\beta}} \in \mathbb{R}^2$ als Lösung der Optimierungsaufgabe

$$S(\hat{\underline{\beta}}) = \sum_{i=1}^{n} [y_i - (\beta_1 + \beta_2 x_i)]^2 \stackrel{!}{=} \min_{\underline{\beta} \in \mathbb{R}^2} S(\underline{\beta})$$