Script **generated by TTT**

Title: Esparza: DWT (28.06.2012)

Date: Thu Jun 28 15:04:47 CEST 2012

Duration: 40:21 min

Pages: 16

Einführung in die (induktive) Statistik

$X \sim \mathrm{Bern}(p)$

$X \sim \mathrm{Po}(\lambda)$

$X \sim \mathcal{N}(\mu, \sigma^2)$

Typische Fragestellung der Statistik:

Auf Grund einer Problemmodellierung sind wir interessiert an:

Zufallsexperiment beschrieben durch ZV X.

Problem: Verteilung von X ist nicht vollständig bekannt.

Z.B.: $X \sim \mathcal{N}(\mu, 1)$ für unbekanntes μ.

Gegeben: Wiederholungen $X_1, X_2, \ldots, X_n$ des Experiments

D.h.: $X, X_1, \ldots, X_n$ sind identisch verteilt.

Häufige Annahme: $X_1, \ldots, X_n$ unabhängig.

ergeben Beobachtungen/Stichprobe $x_1, \ldots, x_n$.

▷ Versuche mit Hilfe der Stichprobe Rückschlüsse auf die Verteilung von X zu ziehen.

Beispiel: Anzahl Taxis

Situation:

Beobachten von vorbeifahrenden Taxis.

Jedes Taxi trägt eine eindeutige Nummer.

Nummer des i-ten beobachteten Taxis: $x_i \in \mathbb{N}$.

Fragestellung:

Wie viele Taxis gibt es?

Beispiel: Anzahl Taxis

Modellierung:

Wir treffen folgende (sinnvolle?) Annahmen:

- Beobachtung eines Taxis entspricht Zufallsexperiment X:

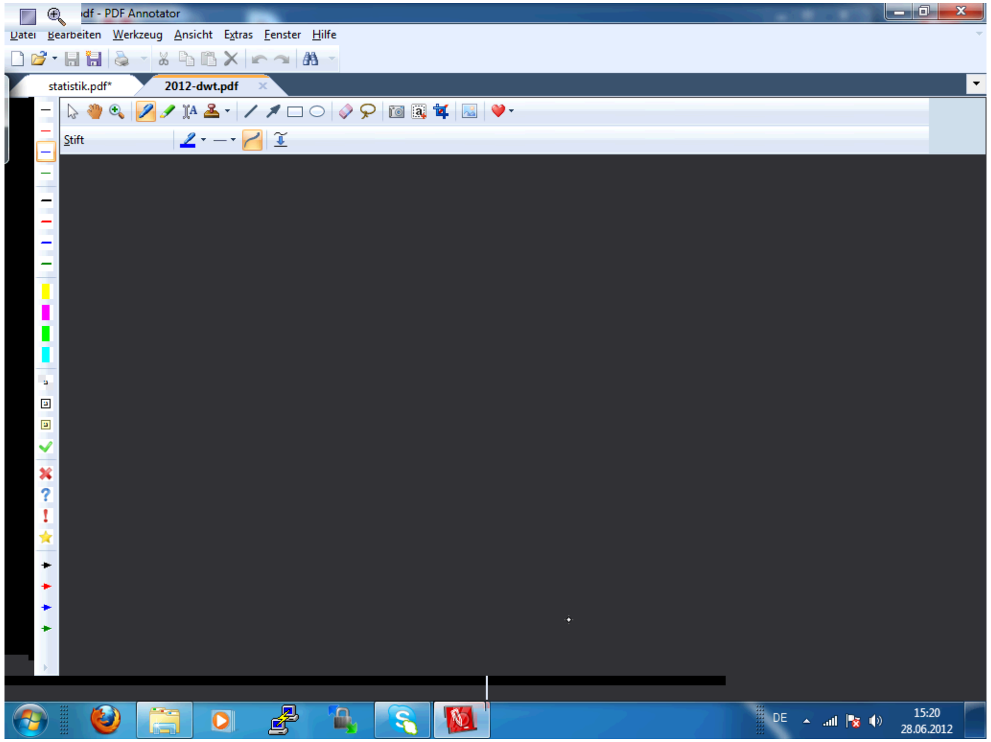
  Ziehung eines zufälligen Wertes aus $[M]$,

  wobei M die gesuchte Anzahl der Taxis (der unbekannte Parameter) ist.
- Beobachtung von n Taxis entspricht n unabhängigen Wiederholungen von X:

  Unabhängige ZVen $X_1, \ldots, X_n$ mit X_i ist gleichverteilt über $[M]$.
- Beobachtete Nummer x_i ist das Ergebnis des Experiments X_i.

Formalisierung des Problems:

Unter diesen Annahmen und gegeben die Stichprobe $x_1, \ldots, x_n$ bestimme einen Schätzwert für M.



3. Konfidenzintervalle

Bei der Verwendung von Schätzvariablen geht man davon aus, dass der erhaltene Schätzwert „nahe“ beim gesuchten Parameter θ liegt. Die Schätzungen werden „besser“, je größer die betrachtete Stichprobe ist. Diese Angaben sind aus quantitativer Sicht natürlich unbefriedigend, da nicht erkennbar ist, wie gut man sich auf den Schätzwert verlassen kann.

Die Lösung dieses Problems besteht darin, statt einer Schätzvariablen U zwei Schätzer U_1 und U_2 zu betrachten. U_1 und U_2 werden so gewählt, dass

$$\Pr[U_1 \leq \theta \leq U_2] \geq 1 - \alpha.$$

Die Wahrscheinlichkeit $1 - \alpha$ heißt Konfidenzniveau und kann dem „Sicherheitsbedürfnis“ angepasst werden.

Wenn wir für eine konkrete Stichprobe die Schätzer U_1 und U_2 berechnen und davon ausgehen, dass $\theta \in [U_1, U_2]$ ist, so ziehen wir höchstens mit Wahrscheinlichkeit α einen falschen Schluss. $[U_1, U_2]$ heißt Konfidenzintervall.

In vielen Fällen verwendet man nur eine Schätzvariable U und konstruiert mittels $U_1 := U - \delta$ und $U_2 := U + \delta$ ein symmetrisches Konfidenzintervall $[U - \delta, U + \delta]$.

Sei X eine $\mathcal{N}(\mu, \sigma^2)$-verteilte Zufallsvariable, und seien $X_1, \ldots, X_n$ n zugehörige Stichprobenvariablen. Gemäß der Additivität der Normalverteilung (siehe Satz 78) ist das Stichprobenmittel $\overline{X}$ ebenfalls normalverteilt mit $\overline{X} \sim \mathcal{N}(\mu, \frac{\sigma^2}{n})$. Wir suchen für $\overline{X}$ ein symmetrisches Konfidenzintervall.

Nach Lemma 72 ist

$$Z := \sqrt{n} \cdot \frac{\overline{X} - \mu}{\sigma}$$

standardnormalverteilt.

Für Z betrachten wir das Konfidenzintervall $[-c, c]$ für ein geeignetes $c > 0$ und setzen

$$\Pr[-c \leq Z \leq c] \stackrel{!}{=} 1 - \alpha.$$

Auflösen nach μ ergibt

$$\Pr\left[\overline{X} - \frac{c\sigma}{\sqrt{n}} \leq \mu \leq \overline{X} + \frac{c\sigma}{\sqrt{n}}\right] \stackrel{!}{=} 1 - \alpha\,.$$

Das gesuchte Konfidenzintervall lautet also

$$K = [\overline{X} - \frac{c\sigma}{\sqrt{n}}, \overline{X} + \frac{c\sigma}{\sqrt{n}}]\,.$$

Für Z betrachten wir das Konfidenzintervall $[-c, c]$ für ein geeignetes $c > 0$ und setzen

$$\Pr[-c \leq Z \leq c] \stackrel{!}{=} 1 - \alpha.$$

Auflösen nach μ ergibt

$$\Pr\left[\overline{X} - \frac{c\sigma}{\sqrt{n}} \leq \mu \leq \overline{X} + \frac{c\sigma}{\sqrt{n}}\right] \stackrel{!}{=} 1 - \alpha\,.$$

Das gesuchte Konfidenzintervall lautet also

$$K = [\overline{X} - \frac{c\sigma}{\sqrt{n}}, \overline{X} + \frac{c\sigma}{\sqrt{n}}]\,.$$

Den Parameter c wählen wir wie folgt:

$$\Pr[-c \leq Z \leq c] = \Phi(c) - \Phi(-c) \stackrel{!}{=} 1 - \alpha.$$

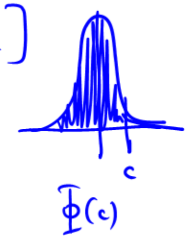
Wegen der Symmetrie von Φ gilt $\Phi(-x) = 1 - \Phi(x)$ und wir erhalten

$$\Phi(c) - \Phi(-c) = 2 \cdot \Phi(c) - 1 \stackrel{!}{=} 1 - \alpha \iff \Phi(c) = 1 - \frac{\alpha}{2},$$

also

$$c = \Phi^{-1}\left(1 - \frac{\alpha}{2}\right).$$

$\Pr[Z \leq c] - \Pr[Z \leq -c]$



Den Parameter c wählen wir wie folgt:

$$\Pr[-c \leq Z \leq c] = \Phi(c) - \Phi(-c) \stackrel{!}{=} 1 - \alpha.$$

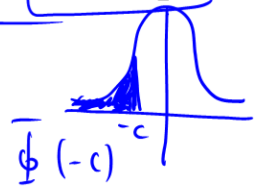
Wegen der Symmetrie von Φ gilt $\Phi(-x) = 1 - \Phi(x)$ und wir erhalten

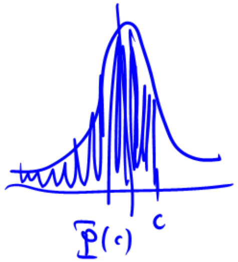
$$\Phi(c) - \Phi(-c) = 2 \cdot \Phi(c) - 1 \stackrel{!}{=} 1 - \alpha \iff \Phi(c) = 1 - \frac{\alpha}{2},$$

also

$$\boxed{c = \Phi^{-1}\left(1 - \frac{\alpha}{2}\right).}$$

$\Phi(c) = 1 - \frac{\alpha}{2}$





Definition 98

X sei eine stetige Zufallsvariable mit Verteilung F_X. Eine Zahl x_γ mit

$$F_X(x_\gamma) = \gamma$$

heißt γ-Quantil von X bzw. der Verteilung F_X.

Definition 99

Für die Standardnormalverteilung bezeichnet z_γ das γ-Quantil.

Damit können wir das gesuchte Konfidenzintervall angeben durch

$$K = \left[\overline{X} - \frac{z_{(1-\frac{\alpha}{2})}\sigma}{\sqrt{n}}, \overline{X} + \frac{z_{(1-\frac{\alpha}{2})}\sigma}{\sqrt{n}}\right].$$

4. Testen von Hypothesen

4.1 Einführung

Bislang haben wir versucht, Parameter von Verteilungen zu schätzen. In der Praxis ist man jedoch oft an der eigentlichen Kenntnis dieser Parameter gar nicht interessiert, sondern man möchte gewisse, damit zusammenhängende Behauptungen überprüfen.

Im Folgenden stellen wir die Bestandteile eines statistischen Tests anhand eines abstrakten Beispiels vor. Wir betrachten dazu eine Zufallsvariable X mit $\Pr[X = 1] = p$ und $\Pr[X = 0] = 1 - p$. Durch einen Test soll überprüft werden, ob $p < 1/3$ oder $p \geq 1/3$ gilt.