SS 2004

Diskrete Strukturen II

Ernst W. Mayr

Fakultät für Informatik

TU München

http://www14.in.tum.de/lehre/2004SS/ds/index.html.de

Gewöhnlich wird K konstruiert, indem man die Zufallsvariablen $X_1, \ldots, X_n$ zu einer neuen Variablen T, der so genannten Testgröße, zusammenfasst. Dann unterteilt man den Wertebereich $\mathbb{R}$ von T in mehrere Bereiche, die entweder zur Ablehnung der Hypothese führen sollen oder nicht. Dabei betrachtet man meist ein einzelnes halboffenes oder abgeschlossenes Intervall und spricht dann von einem einseitigen bzw. von einem zweiseitigen Test.

Die Menge $\widetilde{K} \subseteq \mathbb{R}$ enthalte die Werte von T, die zur Ablehnung der Hypothese führen sollen. Da wir Tests immer über eine Testgröße definieren, werden wir der Einfachheit halber auch $\widetilde{K}$ als Ablehnungsbereich bezeichnen. $\widetilde{K} \subseteq \mathbb{R}$ entspricht direkt dem Ablehnungbereich $K = T^{-1}(\widetilde{K}) \subseteq \mathbb{R}^n$, wie wir ihn oben festgelegt haben.

Die zu überprüfende Hypothese bezeichnen wir mit H_0 und sprechen deshalb auch von der Nullhypothese. Bei manchen Tests formuliert man noch eine zweite Hypothese H_1, die so genannte Alternative. Im Beispiel können wir

$$H_0 : p \geq 1/3 \text{ und } H_1 : p < 1/3$$

setzen.

Manchmal verzichtet man darauf, H_1 anzugeben. Dann besteht die Alternative wie oben einfach darin, dass H_0 nicht gilt. In diesem Fall nennen wir H_1 triviale Alternative.

Ein echter, also nicht-trivialer Alternativtest läge beispielsweise vor, wenn wir ansetzen

$$H_0' : p \geq 1/3 \text{ und } H_1' : p \leq 1/6.$$

Beispiel:

Wir untersuchen eine Festplatte, von der bekannt ist, dass sie zu einer von zwei Baureihen gehört. Die mittleren Zugriffszeiten dieser Baureihen betragen 9ms bzw. 12ms. Wir möchten nun herausfinden, zu welchem Typ die betrachtete Festplatte gehört, indem wir die Zugriffszeit bei n Zugriffen bestimmen. Hier würde man dann ansetzen: $H_0 : \mu \leq 9$ und $H_1 := \mu \geq 12$, wobei μ die mittlere Zugriffszeit bezeichnet.

Fehler bei statistischen Tests

Bei jedem statistischen Test können mit einer gewissen Wahrscheinlichkeit falsche Schlüsse gezogen werden. Dieser Fall tritt beispielsweise ein, wenn H_0 gilt, aber das Ergebnis $\vec{x}$ der Stichprobe im Ablehnungsbereich K liegt.

Dann spricht man von einem Fehler 1. Art.

Analog erhalten wir einen Fehler 2. Art, wenn H_0 nicht gilt und $\vec{x}$ nicht im Ablehnungsbereich liegt.

Fehler 1. Art	:	H_0 gilt, wird aber abgelehnt.
Fehler 2. Art	:	H_0 gilt nicht, wird aber angenommen.

Für die Beurteilung eines Tests ist es wesentlich, mit welcher Wahrscheinlichkeit diese beiden Fehler eintreten können. Ziel ist es natürlich, diese Wahrscheinlichkeiten möglichst klein zu halten. Allerdings sind die Minimierung des Fehlers 1. Art und des Fehlers 2. Art gegenläufige Ziele, so dass ein vernünftiger Ausgleich zwischen beiden Fehlern gefunden werden muss. Wenn man beispielsweise $K = \emptyset$ setzt, so erhält man Wahrscheinlichkeit Null für den Fehler 1. Art, da H_0 immer angenommen wird. Allerdings tritt der Fehler 2. Art dann mit Wahrscheinlichkeit Eins ein, wenn H_0 nicht gilt.

Die Wahrscheinlichkeit für den Fehler 1. Art wird mit α bezeichnet, und man spricht deshalb gelegentlich vom α-Fehler. α heißt auch Signifikanzniveau des Tests.

In der Praxis ist es üblich, sich ein Signifikanzniveau α vorzugeben (übliche Werte hierfür sind $0{,}05$, $0{,}01$ oder $0{,}001$) und dann den Test so auszulegen (also den Ablehnungsbereich K so zu bestimmen), dass die Wahrscheinlichkeit für den Fehler 1. Art den Wert α besitzt.

Konstruktion eines einfachen Tests

Wir konstruieren einen Test für den Parameter p einer Bernoulli-verteilten Zufallsvariablen X. Wir setzen

$$H_0 : p \geq p_0, \qquad H_1 : p < p_0.$$

Als Testgröße verwenden wir

$$T := X_1 + \ldots + X_n.$$

Für größere Wahrscheinlichkeiten p erwarten wir auch größere Werte für T. Deshalb ist es sinnvoll, einen Ablehnungsbereich der Art $K := [0, k]$ für T zu wählen, wobei $k \in \mathbb{R}$ geeignet festzulegen ist. Wir konstruieren hier also einen einseitigen Test, während für eine Nullhypothese $H_0 : p = p_0$ sowohl zu kleine als auch zu große Werte von T zur Ablehnung von H_0 führen sollten und somit ein zweiseitiger Test vorzuziehen wäre.

T ist binomialverteilt. Da wir von einem großen Stichprobenumfang n ausgehen, bietet es sich an, die Verteilung von T nach dem Grenzwertsatz von DeMoivre (siehe Korollar 46) durch die Normalverteilung zu approximieren.

Sei

$$\tilde{T} := \frac{T - np}{\sqrt{np(1-p)}}.$$

$\tilde{T}$ ist annähernd standardnormalverteilt.

Wir berechnen für jeden Wert von k das zugehörige Signifikanzniveau α des Tests.

$$\begin{aligned}
\text{Fehlerwahrscheinlichkeit 1. Art} &= \max_{p \in H_0} \Pr_p[T \in K] \\
&= \max_{p \in H_0} \Pr_p[T \leq k] \\
\text{Fehlerwahrscheinlichkeit 2. Art} &= \sup_{p \in H_1} \Pr_p[T \notin K] \\
&= \sup_{p \in H_1} \Pr_p[T > k]
\end{aligned}$$

Für den Fehler 1. Art α erhalten wir

$$\begin{aligned}
\alpha &= \max_{p \geq p_0} \Pr_p[T \leq k] = \Pr_{p=p_0}[T \leq k] \\
&= \Pr_{p=p_0}\left[\tilde{T} \leq \frac{k - np}{\sqrt{np(1-p)}}\right] \\
&= \Pr\left[\tilde{T} \leq \frac{k - np_0}{\sqrt{np_0(1-p_0)}}\right] \approx \Phi\left(\frac{k - np_0}{\sqrt{np_0(1-p_0)}}\right).
\end{aligned}$$

Unter Verwendung der Quantile der Standardnormalverteilung ergibt sich damit:

- Ist k so gewählt, dass $(k - np_0)/\sqrt{np_0(1-p_0)} = z_\alpha$, so ist das Signifikanzniveau gleich α.
- Ist das gewünschte Signifikanzniveau α des Tests vorgegeben, so erhält man den Wert $k = k(n)$ in Abhängigkeit vom Umfang n der Stichprobe durch

$$k = z_\alpha \cdot \sqrt{np_0(1-p_0)} + np_0. \tag{3.4.1}$$

Kleinere Werte für k verkleinern zwar den Fehler 1. Art, vergrößern jedoch den Annahmebereich und damit die Wahrscheinlichkeit für einen Fehler 2. Art.

Verhalten der Testfehler

Wie verhalten sich die möglichen Testfehler des konstruierten Verfahrens? Was geschieht beispielsweise, wenn p nur geringfügig kleiner als p_0 ist?

In diesem Fall betrachten wir beim Fehler 2. Art die Wahrscheinlichkeit

$$\Pr_{p=p_0-\varepsilon}[T \geq k] \approx \Pr_{p=p_0}[T \geq k] \approx 1 - \alpha \, .$$

Wenn sich also die „wahren“ Verhältnisse nur minimal von unserer Nullhypothese unterscheiden, so werden wir diese „im Zweifelsfall“ annehmen.

Bei echten Alternativtests werden für hinreichend große Stichproben und einen geeignet eingestellten Ablehnungsbereich beide Testfehler klein.

Beispiel: Die Abbruchrate p der Transaktionen in einem Online-Datenbanksystem wurde bereits früher einmal ermittelt. Allerdings sind die entsprechenden Daten verloren gegangen und die Entwickler erinnern sich nur noch, dass das Ergebnis entweder $p = 1/3$ oder $p = 1/6$ lautete. Unter dieser Annahme würde man den Test wie folgt ansetzen:

$$H_0 : p \geq 1/3, \qquad H_1' : p \leq 1/6.$$

Für den Fehler 2. Art erhält man nun:

$$\begin{aligned} \text{Fehlerwahrsch. 2. Art} &= \max_{p \leq 1/6} \Pr_p[T > k] \\ &\approx 1 - \Phi\left(\frac{k - (1/6) \cdot n}{\sqrt{(1/6) \cdot (5/6) n}}\right). \end{aligned}$$

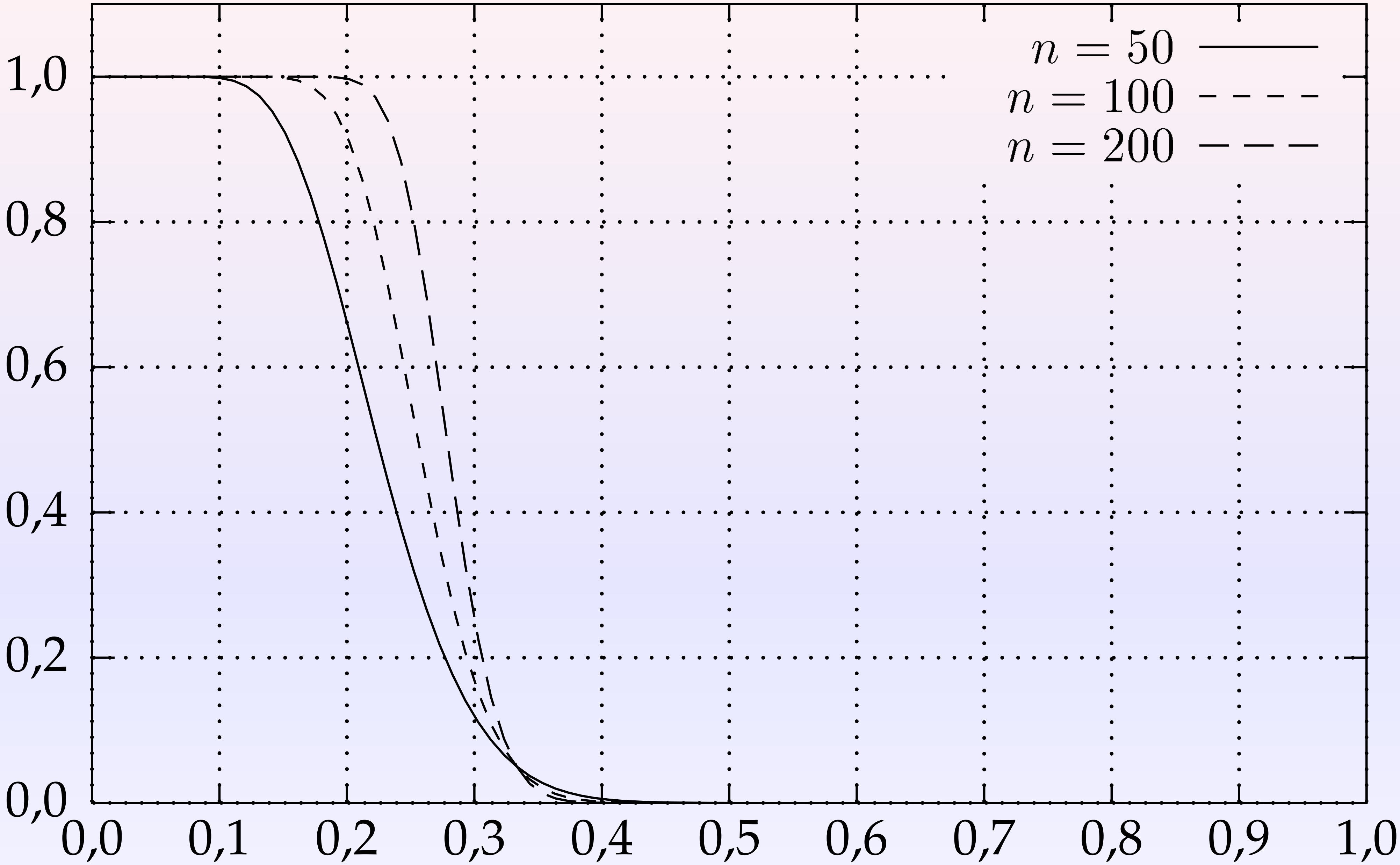
Mit den obigen Werten $k = 25$ und $n = 100$ ergibt sich mit

$$\Phi\left(\frac{150 - 100}{\sqrt{5} \cdot 10}\right) = \Phi(\sqrt{5}) \approx 0{,}9871$$

ein Fehler 2. Art der Größe $0{,}0129$, während sich für die triviale Alternative $H_1 : p \leq 1/3$ ein Wert von etwa $0{,}95$ ergibt.

Die so genannte Gütefunktion g gibt allgemein die Wahrscheinlichkeit an, mit der ein Test die Nullhypothese verwirft. Für unser hier entworfenes Testverfahren gilt

$$g(n,p) = \Pr_p[T \in K] = \Pr_p[T \le k] \approx \Phi\left(\frac{k - np}{\sqrt{np(1-p)}}\right).$$



Gütefunktion $g(n, p)$ für verschiedene Werte von n

Man erkennt deutlich, dass für alle n der Wert von $k = k(n)$ genau so gewählt wurde, dass $g(n, 1/3) = 0{,}05$ gilt. Dies wird durch den in Gleichung 3.4.1 angegebenen Ausdruck erreicht.

Für Werte von p größer als $1/3$ wird $H_0 : p \geq 1/3$ mit hoher Wahrscheinlichkeit angenommen, während für Werte deutlich unter $1/3$ die Hypothese H_0 ziemlich sicher abgelehnt wird.

Ferner ist auffällig, dass g für größere Werte von n schneller von Eins auf Null fällt. Daran erkennt man, dass durch den Test die Fälle „H_0 gilt“ und „H_0 gilt nicht“ umso besser unterschieden werden können, je mehr Stichproben durchgeführt werden. Für Werte von p, bei denen $g(n, p)$ weder nahe bei Eins noch nahe bei Null liegt, kann der Test nicht sicher entscheiden, ob die Nullhypothese abzulehnen ist.