Statistisches Testen I

Was ist ein statistischer Test?

Ein statistischer Test ist ein Entscheidungsverfahren, das über eine a priori formulierte „Vermutung“ (Hypothese) nach der Sammlung von empirischer Information (Daten) eine konkrete Entscheidung trifft.

„Entscheidung“ meint hier, dass die Vermutung entweder abgelehnt wird oder nicht.

Die getroffene Entscheidung kann richtig oder falsch sein.

Ein statistischer Test erlaubt es, die Häufigkeit einer Fehlentscheidung im Sinne der ungerechtfertigten Ablehnung der Vermutung zu „kontrollieren“.

Beispiele

1. Randomisierte kontrollierte Therapiestudie zur Behandlung des schweren Atemnotsyndroms von Frühgeborenen mit zwei Dosierungsschemata eines natürlichen Surfactantpräparats

   Ergebnisse:

Therapiearm	Umfang	Mortalität
einfach	176	37 (= 21%)
mehrfach	167	21 (= 13%)

   Frage: Mit welchem Dosierungsschema soll in der Zukunft ein schweres Atemnotsyndrom bei Frühgeborenen behandelt werden?

2. Laut Herstellerangabe enthält der oral-wirksame ACE-Hemmer Xanef zur Blutdrucksenkung 10mg Enalapril pro Tablette. Zur Qualitätskontrolle des Herstellers überprüft ein pharmakologisches Institut den Wirkstoffgehalt von 100 zufällig ausgewählten Xanef-Tabletten.

   Ergebnis:

   Der Mittelwert der 100 Wirkstoffgehaltsangaben lag bei 9.81mg, die Standardabweichung der Messungen bei 0.76mg.

   Frage: Hat der Hersteller ein Produktionsproblem bei der Fabrikation von Xanef-Tabletten?

Statistisches Testen

Allgemeines Vorgehen:

- Anhand des Vorwissens wird eine „Vermutung“ formuliert.
- Daten werden im Rahmen einer Studie, eines Experiments o.ä. erhoben.
- Aufgrund der Daten wird nach Anwendung einer statistischen Regel eine Entscheidung über die Vermutung getroffen.
- Diese Entscheidung kann falsch sein.
- Die Wahrscheinlichkeit für eine Fehlentscheidung wird durch Vorgabe eines „Signifikanzniveaus“ nach oben begrenzt.

Wie formuliert man statistische Vermutungen?

Eine „statistische Vermutung“ wird durch das Aufstellen von Nullhypothese (H_0) und zugehöriger Alternative (H_1) formalisiert. Die Nullhypothese enthält dabei stets die zu widerlegende Aussage, die Alternative beinhaltet die zu beweisende Aussage.

H_0: ⟨was zu widerlegen ist⟩

H_1: ⟨was zu beweisen ist⟩

Fortsetzung der Beispiele

1. **Nullhypothese (H_0):** „Die Mortalitäts-wahrscheinlichkeiten der beiden Therapiearme unterscheiden sich nicht.“

   Alternative (H_1): „Die Mortalitäts-wahrscheinlichkeiten der beiden Therapiearme unterscheiden sich.“ oder „Die Mortalitäts-wahrscheinlichkeit der mehrfach-behandelten Gruppe ist größer/kleiner als die der einfach behandelten Gruppe.“

2. **Nullhypothese (H_0):** „Der Wirkstoffgehalt einer Xanef-Tablette ist 10 mg.“

   Alternative (H_1): „Der Wirkstoffgehalt einer Xanef-Tablette ist von 10 mg verschieden.“ oder „Der Wirkstoffgehalt einer Xanef-Tablette ist größer/kleiner als 10 mg.“

Ein- und zweiseitiges statistisches Testen

Es gibt verschiedene Formulierungen der Alternative H_1 beim statistischen Testen. Ist eine Abweichung von der Nullhypothese (H_0) nur in eine Richtung („größer“/„kleiner“) von Interesse, so spricht man von einem einseitigen statistischen Test. Ist eine Abweichung von der Nullhypothese (H_0) in beide Richtungen („unterschiedlich“ vom in H_0 spezifizierten Zustand) von Interesse, so spricht man von einem zweiseitigen statistischen Test.

Das Standardvorgehen des statistischen Testens in der Medizin ist der zweiseitige Test. Lässt sich jedoch **a priori** stringent begründen, dass eine Abweichung von H_0 nur in eine Richtung möglich ist, so ist in diesen seltenen Ausnahmefällen ein einseitiger Test indiziert.

Fehlentscheidungen beim statistischen Testen

Es gibt zwei unterschiedliche Typen von Fehlentscheidungen:

- **Fehler 1. Art:** Nullhypothese wird abgelehnt, obwohl sie zutrifft
- **Fehler 2. Art:** Nullhypothese wird nicht abgelehnt, obwohl sie nicht zutrifft

Die Wahrscheinlichkeit für einen Fehler 1. Art wird i.d.R. mit α abgekürzt, die Wahrscheinlichkeit für einen Fehler 2. Art mit β. Beim statistischen Testen wird α („Signifikanzniveau“) vorgegeben und durch die Konstruktion des Entscheidungsverfahrens „kontrolliert“. Dies gilt nicht für β.

Illustration der Fehlertypen beim statistischen Testen

		Testentscheidung	
		H_0 nicht ablehnen	H_0 ablehnen
Unbekannte Realität	H_0 richtig	Richtige Entscheidung	**Fehler 1. Art** α
	H_0 falsch	**Fehler 2. Art** β	Richtige Entscheidung

Illustration der Fehlertypen im Beispielkontext

		Testentscheidung	
		H_0 nicht ablehnen	H_0 ablehnen
Unbekannte Realität	kein Unterschied zwischen beiden Therapien	kein Unterschied (richtig)	**signifikanter Unterschied (falsch)**
	eine Therapie ist besser als die andere	**kein Unterschied (falsch)**	signifikanter Unterschied (richtig)

Wie trifft man auf der Basis der Daten eine statistische Entscheidung?

Allgemein: Man berechnet aus den Daten eine „Prüfgröße“ (Teststatistik), die einem Auskunft über die Vermutung gibt. Die Form der Teststatistik wird so gestaltet, dass man ihre Wahrscheinlichkeitsverteilung unter der Nullhypothese angeben kann. Dann weiß man, ob ein aus den Daten berechneter Wert für Teststatistik unter Gültigkeit der Vermutung „wahrscheinlich“ oder „unwahrscheinlich“ ist. Ist er so unwahrscheinlich, dass er kleiner als das zuvor festgesetzte Signifikanzniveau α ist, so entscheidet man sich, die Nullhypothese abzulehnen. Ist die Wahrscheinlichkeit allerdings größer als α, so lehnt man die Nullhypothese nicht ab.

Fortsetzung des Beispiels 2

Hat der Medikamentenproduzent ein Produktionsproblem bei seinen Xanef-Tabletten?
(Erinnerung: N = 100, $\bar{X} = 9.81$, $\sigma = 0.76$)

H_0: Wirkstoffgehalt ist 10 mg

H_1: Wirkstoffgehalt unterscheidet sich von 10 mg

Prüfgröße T = $(10 - 9.81) / 0.76 * \sqrt{100}$

= 2.5

T ist unter H_0 t-verteilt mit 99 Freiheitsgraden.

Signifikanzniveau α sei 0.05

$\Rightarrow$ unter H_0 gilt : $P(|T| \geq c_\alpha) = 0.05$, wobei c_α „kritischer Wert“ heißt. In unserem Fall (siehe Vertafelungen) ist $c_\alpha = 1.99$.

Wahl des Signifikanzniveaus

Standard: $\alpha = 0.05$

Jede Abweichung muss gut begründet sein.

Alternativen:

- $\alpha = 0.01$, falls ein Fehler 1. Art gravierende Konsequenzen hat, ein Fehler 2. Art zu tolerieren ist
- $\alpha = 0.10$, falls ein Fehler 2. Art gravierende Konsequenzen hat, ein Fehler 1. Art zu tolerieren ist

Konsequenz der Entscheidung zwischen ein- und zweiseitigem statistischen Testen

---

Einseitiges Testen: „Kontrolle“ des Fehlers 1. Art (α) nur in eine Richtung notwendig

Zweiseitiges Testen: „Kontrolle“ des Fehlers 1. Art (α) in beide Richtungen notwendig, so dass für jede Richtung nur $\alpha/2$ zur Verfügung stehen

Somit können bei denselben Daten unterschiedliche Testentscheidungen entstehen. Bei einseitigem Testen kommt man eher zu einer Ablehnung der Nullhypothese als bei zweiseitigem Testen. Dies birgt ein manipulatives Potential in sich, so dass man strenge Anforderungen an die Begründung für ein Abweichen vom Standardvorgehen des zweiseitigen Testens stellt.

Was sind p-Werte?

Ein p-Wert ist die Wahrscheinlichkeit dafür, dass die Prüfgröße unter H_0 Werte annimmt, die größer oder gleich dem aus den Daten berechneten Wert der Prüfgröße sind. Statistiker nennen p-Werte daher „Überschreitungs-wahrscheinlichkeiten“.

Fortsetzung des Beispiels 2

T = 2.5, T ist unter H_0 t-verteilt mit 99 Freiheitsgraden

$\Rightarrow P_{Ho}(|T| \geq 2.5) = 0.007$ ist der zugehörige (zweiseitige) p-Wert des statistischen Tests

Kennt man den p-Wert, so kann man auf die Ermittlung kritischer Werte aus Tabellen verzichten.

Begründung: Vorgegeben sei ein Signifikanzniveau (Größenordnung des tolerierten Fehlers 1. Art). Die beiden Entscheidungsregeln

- H_0 ablehnen, falls p-Wert $\leq \alpha$
- H_0 ablehnen, falls Prüfgröße $\geq c_\alpha$

sind äquivalent. Die Popularität von p-Werten resultiert aus der Verwendung von Statistiksoftware. Die Software gibt stets p-Werte als Resultate von statistischen Tests aus.

Interpretation von Testentscheidungen

Allgemeine Beschreibung:

1. Fall: Prüfgröße $\geq c_\alpha$ oder

p-Wert $\leq \alpha$

Entscheidung: H_0 ablehnen

P(Fehler 1. Art) = α

P(Fehler 2. Art) = 0

Interpretation:

Die in den Daten zu beobachtende Abweichung von der Nullhypothese ist auf dem (100•α)%-Niveau ist signifikant.

2. Fall: Prüfgröße $<$ c_α oder

p-Wert $>$ α

Entscheidung: H_0 nicht ablehnen

P(Fehler 1. Art) = 0

P(Fehler 2. Art) = unbekannt

Interpretation:

Die in den Daten zu beobachtende Abweichung von der Nullhypothese ist auf dem (100•α)%-Niveau ist nicht signifikant.

Vorsicht: H_0 nicht abzulehnen, bedeutet **nicht** die Gültigkeit von H_0 (mit irgendeiner Wahrscheinlichkeitsquantifizierung, z.B. 1 - α) statistisch nachgewiesen zu haben.

Statistische Power

Als „statistische Power“ eines Tests bezeichnet man die Wahrscheinlichkeit, mit der ein statistischer Test eine spezifische „richtige“ Alternative unter den Rahmenbedingungen seines Einsatzes (Fallzahl, Signifikanzniveau) auch als solche entdeckt (d.h. die „falsche“ Nullhypothese ablehnt). Formal gilt:

$$\text{Statistische Power} = 1 - \text{Fehler 2. Art}$$

$$= 1 - \beta$$

Die statistische Power eines Tests hängt von folgenden Größen ab:

- **Fallzahl der Studie:**

  je größer die Fallzahl, desto größer die statistische Power
- **Signifikanzniveau des statistischen Tests:**

  je kleiner das Signifikanzniveau, desto kleiner die statistische Power
- **der konkreten Alternative, die vorliegt:**

  je „weiter weg“ von der Nullhypothese die spezifische Alternative, desto größer die statistische Power

Illustration der statistischen Power

Fortsetzung des Beispiels :

1) Abhängigkeit von der Fallzahl

Statistische Power zur Aufdeckung des Unterschiedes zwischen den beiden Therapiearmen (Mortalität: 21% vs. 13%) bei einem Signifikanzniveau von 5%

Fallzahl	Power
50	0.11
100	0.18
200	0.32
300	0.45
500	0.66
1000	0.92

2) Abhängigkeit vom Signifikanzniveau

Statistische Power zur Aufdeckung des Unterschiedes zwischen den beiden Therapiearmen (Mortalität: 21% vs. 13%) bei einer Fallzahl von 500

Signifikanzniveau	Power
0.001	0.18
0.01	0.43
0.05	0.66
0.1	0.77

3) Abhängigkeit von der konkreten Alternative

Statistische Power zur Aufdeckung des Unterschiedes zwischen den beiden Therapiearmen (Mortalität im einfach-behandelten Therapiearm: 21%) bei einer Fallzahl von 500 und einem Signifikanzniveau von 5%

Mortalität im anderen Arm	Power
17%	0.21
15%	0.42
13%	0.66
11%	0.86
9%	0.97

Statistische Power und Fallzahlplanung

Zusammenhang zwischen statistischer Power (Pow), Fallzahl (n), Signifikanzniveau (α) und spezifischer Alternative (H_1^P) ist evident.

Bislang haben wir Pow als Funktion von n, α und H_1^P aufgefasst. Wir können jedoch diese Beziehung auch nach n „auflösen“ und somit die Fallzahl n als Funktion von Pow, α und H_1^P beschreiben. Dies ist dann die Basis für jede Fallzahlplanung. Das Einsetzen konkreter Werte für die statistische Power Pow, das Signifikanzniveau α und die spezifische Alternative H_1^P in diese Funktion liefert dann die Fallzahl für die vorgegebene Situation.

Konkrete statistische Tests

Generell: Es existieren sehr viele verschiedene statistische Tests, von denen nur wenige Standardtests im Rahmen der Vorlesung konkret vorgestellt werden.

Kriterien für die Unterscheidung zwischen statistischen Tests:

- Skalentyp der auszuwertenden Daten (teilweise zusätzlich: Verteilungsannahmen)
- Strukturelle Designmerkmale der Studie (z.B. Parallel-Gruppen-Design versus Cross-Over-Design in klinischen Prüfungen)

Tests im Parallel-Gruppen-Design

„Parallel-Gruppen-Design“ meint, dass zwei (oder mehrere) Gruppen von unabhängigen Merkmalsträgern hinsichtlich möglicher (Gruppen-)Unterschiede in der Verteilung der beobachteten Merkmalsausprägungen analysiert werden. Statistiker sprechen oft von dieser Situation als „Vergleich zweier (oder mehrerer) unverbundener Stichproben“.

Statistische Tests in dieser Situation:

t-Test für zwei unverbundene Stichproben

U-Test von Mann-Whitney-Wilcoxon

Logrank-Test

Chi-Quadrat-Test

Exakter Test von Fisher

Tests für verbundene Stichproben

„Verbundene Stichproben“ meint, dass in einer Gruppe von Merkmalsträgern die Verteilung von zwei (oder mehreren) an den Merkmalsträgern beobachteten Merkmalsausprägungen analysiert wird (Synonym: „abhängige Stichproben“). Typische Beispiele für dieses Design sind z.B. „Vorher-Nachher-Vergleiche“ oder klinische Prüfungen mit Therapiewechsel im „Cross-over-Design“.

Statistische Tests in dieser Situation:

(„paired“)t-Test für zwei verbundene Stichproben

Wilcoxon-Vorzeichen-Rangsummen-Test

McNemar-Test

Tests im Einstichproben-Design

„Einstichproben-Design“ meint, dass nur eine Gruppe von unabhängigen Merkmalsträgern hinsichtlich der Verteilung der beobachteten Merkmalsausprägungen analysiert wird. Das zuvor vorgestellte Beispiel der Überprüfung des Wirkstoffgehalts der Xanef-Tabletten stammt aus diesem Design.

Statistische Tests in dieser Situation:

Einstichproben-t-Test

Binomialtest