Zur Konzeption einfachster Untersuchungen: Effektgrösse, Teststärke und optimale Stichprobengrösse

Inhaltsverzeichnis

Zur Konzeption einfachster Untersuchungen: Effektgrösse, Teststärke und optimale Stichprobengrösse 2
Lernhinweise 2
Einführung 3
Einführung - Teil 1 .....
Einführung - Teil 2 .....
Einführung - Teil 3 .....
Einführung - Teil 4 .....
Theorie (1 - 6) 12
1. Spezifische Alternativhypothese und Effektgrösse 12
2. Alpha- und Beta-Fehler-Risiko 15
3. Teststärke [power] 17
4. Entscheid zwischen Arbeitshypothese und spezifischer Alternativhypothese 17
Teil 1 .....
Teil 2 .....
Teil 3 .....
5. Optimale Stichprobengrösse 21
6. Beurteilung eines effektiv nachgewiesenen Effektes 24
Fallbeispiele (1 - 4) 25
1. Unabhängige Stichproben - Vergleich der Mittelwerte 26
Teil 1 - Unabhängige Stichproben - Vergleich der Mittelwerte .....
Teil 2 - Unabhängige Stichproben - Vergleich der Mittelwerte .....
Teil 3 - Unabhängige Stichproben - Vergleich der Mittelwerte .....
2. Abhängige Stichproben - Vergleich der Mittelwerte 27
Teil 1 - Abhängige Stichproben - Vergleich der Mittelwerte .....
Teil 2 - Abhängige Stichproben - Vergleich der Mittelwerte .....
Teil 3 - Abhängige Stichproben - Vergleich der Mittelwerte .....
Teil 4 - Abhängige Stichproben - Vergleich der Mittelwerte .....
Teil 5 - Abhängige Stichproben - Vergleich der Mittelwerte .....
3. Intervall- oder proportional skalierte Daten - Stochastischer Zusammenhang 30
Teil 1 - Intervall- oder proportional skalierte Daten - Stochastischer Zusammenhang .....
Teil 2 - Intervall- oder proportional skalierte Daten - Stochastischer Zusammenhang .....
Teil 3 - Intervall- oder proportional skalierte Daten - Stochastischer Zusammenhang .....
Teil 4 - Intervall- oder proportional skalierte Daten - Stochastischer Zusammenhang .....
4. Nominal skalierte Daten - Stochastischer Zusammenhang 33
Teil 1 - Nominal skalierte Daten - Stochastischer Zusammenhang .....
Teil 2 - Nominal skalierte Daten - Stochastischer Zusammenhang .....
Teil 3 - Nominal skalierte Daten - Stochastischer Zusammenhang .....
Teil 4 - Nominal skalierte Daten - Stochastischer Zusammenhang .....

Zur Konzeption einfachster Untersuchungen: Effektgrösse, Teststärke und optimale Stichprobengrösse

Lernhinweise

Da schon bei einfachsten Untersuchungen die Resultate statistischer Datenanalysen massgeblich durch das Untersuchungsdesign (z.B. die Wahl der Grösse der untersuchten Stichprobe) mitbestimmt sind, soll in diesem Lernschritt ein Thema besprochen werden, das im Zusammenhang mit jedem Projekt aus der empirischen Forschung relevant ist und zumindest im Detail überdacht werden muss.

Benötigte Vorkenntnisse

- Grundprinzip entscheidungsstatistischer Verfahren
- Parametrische Prüfverfahren

Sie können Ihr Vorwissen in einem kleinen Quiz (siehe unten) prüfen?

Geschätzte Bearbeitungszeit
Nach Konsultation der in Ihrem Curriculum vorgesehenen Vorbereitungslektüre können Sie diesen Lernschritt in ca. 2 Stunden bearbeiten.

Hinweise zur Bearbeitung
Den Studierenden im Grundkurs zur Statistik wird empfohlen, den Lernschritt in der vorgegebenen Struktur zu berarbeiten.

Quiz zu den benötigten Vorkenntnissen
Welche der folgenden Aussagen sind richtig?
1. Parametrische Prüfverfahren setzen intervall- oder proportionalskalierte Daten voraus.
❍ richtig
❍ falsch

2. Unterscheiden sich zwei Stichprobenmittelwerte auf dem 5%-Niveau signifikant, so bedeutet dies, dass bei der Interpretation des Unterschiedes in den Mittelwerten der Zufall mit einer Irrtumswahrscheinlichkeit < 5% ausgeschlossen werden kann.
❍ richtig
❍ falsch

3. Unterscheiden sich zwei Stichprobenmittelwerte auf dem 5%-Niveau nicht signifikant, so ist bewiesen, dass der Unterschied in den Mittelwerten nur dem Zufall zuzuschreiben ist.
❍ richtig
❍ falsch

4. Wird die Arbeitshypothese H_0 auf dem 5%-Niveau abgelehnt, so ist die Wahrscheinlichkeit, dass H_0 irrtümlicherweise abgelehnt wird, kleiner als 5%.

- ❍ richtig
- ❍ falsch

5. Ist die Überschreitungswahrscheinlichkeit der Prüfgrösse, unter Annahme der Gültigkeit von H_0, grösser als 5%, so nehmen wir H_0 an.

- ❍ richtig
- ❍ falsch

6. Das

α

-Fehler-Risiko entspricht dem Risiko, dass H_0 irrtülicherweise angenommen wird.

- ❍ richtig
- ❍ falsch

7. Das

α

-Fehler-Risiko entspricht der Wahrscheinlichkeit, dass H_0 irrtümlicherweise abgelehnt wird.

- ❍ richtig
- ❍ falsch

8. Bei Ablehnung von H_0 entspricht die Irrtumswahrscheinlichkeit p dem

α

-Fehler-Risiko.

- ❍ richtig
- ❍ falsch

9. Bei Annahme von H_0 kann immer eine Irrtumswahrscheinlichkeit angegeben werden.

- ❍ richtig
- ❍ falsch

Einführung

An einem einfachen Beispiel wollen wir uns als erstes die grundlegenden Probleme in Erinnerung rufen, die sich stellen, wenn wir im Rahmen eines entscheidungsstatistischen Verfahrens der Arbeitshypothese H_0 eine unspezifische Alternativhypothese H_1 gegenüberstellen.

Alle in vorhergehenden Lernschritten besprochenen entscheidungsstatistischen Verfahren stützten sich auf eine Arbeitshypothese H_0 und eine unspezifische Alternativhypothese H_1. An einem einfachen Beispiel wollen wir uns als erstes die grundlegenden Probleme in Erinnerung rufen, die sich stellen, wenn wir im Rahmen eines entscheidungsstatistischen Verfahrens der Arbeitshypothese H_0 eine unspezifische Alternativhypothese H_1 gegenüberstellen.

Vergleich des Mittelwertes eines Merkmals, das in einer Stichprobe erhoben wurde, mit dem bekannten Merkmalsmittelwert einer Population

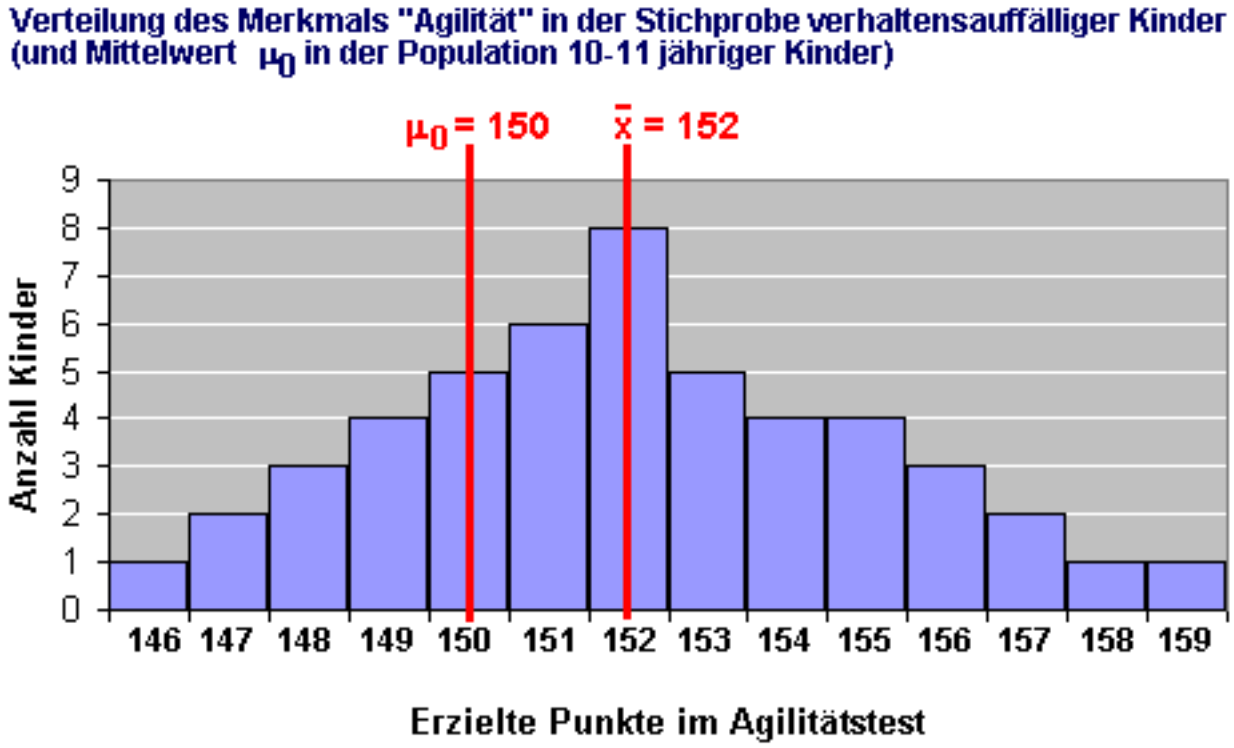
In einer Untersuchung verhaltensauffälliger Kinder stellte sich u.a. die Frage, inwieweit sich diese Kinder auch bezüglich ihrer Agilität von unauffälligen Kindern unterscheiden. So wurde eine Stichprobe von 49 verhaltensauffälligen Kindern zwischen 10 und 11 Jahren mit einem bewährten Agilitätstest untersucht.
Die Auswertung der Testergebnisse zeigt, dass die Testresultate mit den Kennwerten

$\bar{x}$

=152,

σ

=9 eingipflig verteilt sind.



Dem Manual zum eingesetzten Test entnehmen wir, dass die Testresultate in der Eichpopulation 10-11jähriger mitteleuropäischer Kinder mit den Verteilungsparametern

μ_0

=150 und

σ_0

=10 näherungsweise normalverteilt sind.
Der um 2 Punkte höhere Mittelwert der Stichprobe verhaltensauffälliger Kinder fällt wohl auf, bevor aber an eine inhaltliche Ausdeutung gedacht werden kann, muss abgeklärt werden, inwieweit bei der Interpretation eines um zwei Punkte höheren Mittelwertes der Zufall ausgeschlossen werden kann.
Formulieren Sie für sich die diesbezügliche Frage in der Terminologie der Statistik und blättern Sie dann weiter.

Aus der Sicht der Statistik lautet unsere Frage: Darf bei einem um 2 Punkte höheren Stichprobenmittelwert ausgeschlossen werden, dass die Stichprobe aus einer Population stammt, die bezüglich des Mittelwertes mit der bekannten Eichpopulation identisch ist?

Was für eine Fragestellung liegt hier vor?

- ❍ einseitig
- ❍ zweiseitig

Für den Vergleich des Stichproben- und des Populationsmittelwertes verfahren wir wie gewohnt:

1. Zusammenstellung der Voraussetzungen, von denen man bei der Wahl des Prüfverfahrens ausgehen kann: Population: N(150,10); Stichprobe:

   $\bar{x}$

   = 152; s = 9; n = 49
2. Prüfgrösse ist

   $\bar{x}$

   .
3. Arbeitshypothese H_0:

   μ

   =

   μ_0

   Unspezifische Alternativhypothese:

   μ

   >

   μ_0
4. Unter der Annahme der Gültigkeit der Arbeitshypothese H_0 kennen wir die Verteilung der Stichprobenmittelwerte: Für eine Stichprobe der Grösse n = 49 sind die Stichprobenmittelwerte mit den Verteilungsparametern

   $\mu_{\bar{x}}$

   und

   $\sigma_{\bar{x}}$

   normalverteilt.

   $$\mu_{\bar{x}} = \mu_0 = 150$$

   $$\sigma_{\bar{x}} = \sigma_0 / \sqrt{n} = 10 / \sqrt{49} = 1{,}43$$
5. Zur Bestimmung der einseitigen Überschreitungswahrscheinlichkeit des Stichprobenmittelwertes

   $\bar{x}$

=152 transformieren wir diesen Mittelwert in die z-Verteilung.

$$z_{\bar{x}} = \frac{\bar{x} - \mu_{\bar{x}}}{\sigma_{\bar{x}}} = \frac{152 - 150}{1,43} = 1,40$$

Der z-Tabelle entnehmen wir die Überschreitungswahrscheinlichkeit eines z-Wertes von 1,40.

Prozentualer Anteil, der unter einem bestimmten z-Wert liegt

z	A%		z	A%		z	A%
- 3,0	0,1		- 1,0	15,9		+	86,4
- 2,9	0,2		- 0,9	18,4		1,1	88,5
- 2,8	0,3		- 0,8	21,2		+	90,3
- 2,7	0,4		- 0,7	24,2		1,2	91,9
- 2,6	0,5		- 0,6	27,4		+	93,3
- 2,5	0,6		- 0,5	30,8		1,3	94,5
- 2,4	0,8		- 0,4	34,5		+	95,5
- 2,3	1,1		- 0,3	38,2		1,4	96,4
- 2,2	1,4		- 0,2	42,1		+	97,1
- 2,1	1,8		- 0,1	46,0		1,5	97,7
- 2,0	2,3		0,0	50,0		+	98,2
- 1,9	2,9		+	54,0		1,6	98,6
- 1,8	3,6		0,1	57,9		+	98,9
- 1,7	4,5		+	61,8		1,7	99,2
- 1,6	5,5		0,2	65,5		+	99,4
- 1,5	6,7		+	69,2		1,8	99,5
- 1,4	8,1		0,3	72,6		+	99,6
- 1,3	9,7		+	75,8		1,9	99,7
- 1,2	11,5		0,4	78,8		+	99,8
- 1,1	13,6		+	81,6		2,0	99,9
			0,5	84,1		+	
			+			2,1	
			0,6			+	
			+			2,2	
			0,7			+	
			+			2,3	
			0,8			+	
			+			2,4	
			0,9			+	
			+			2,5	
			1,0			+	
						2,6	
						+	
						2,7	
						+	
						2,8	
						+	
						2,9	
						+	
						3,0	

Wie gross ist diese einseitige Überschreitungswahrscheinlichkeit?

p(z

$\geq$

1,40) =

- 91.9%
- 83.9%
- 16.1%
- 8.1%

Die einseitige Überschreitungswahrscheinlichkeit beträgt:

$$p\ (z \geq 1{,}40) = 8{,}1\%$$

Bei dieser Überschreitungswahrscheinlichkeit kann unsere Arbeitshypothese nicht abgelehnt werden. Damit kommen wir zum Schluss, dass bei der Interpretation des um 2 Punkte höheren Mittelwertes der Testresultate in der Stichprobe verhaltensauffälliger Kinder der Zufall nicht ausgeschlossen werden kann. Eine Abschätzung des

β

-Fehler-Risikos, d.h. der Wahrscheinlichkeit, dass H_0 irrtümlicherweise angenommen wird, ist bei unspezifisch formulierter Alternativhypothese H_1

$(\mu > \mu_0)$

nicht möglich.
Nachdem wir auf unsere Frage eine Antwort gefunden haben, wollen wir kurz darüber nachdenken, von welcher Grösse diese Antwort abhängig ist, wenn die Populationsparameter

μ_0

und

σ_0

fest vorgegeben sind und sich der Stichprobenmittelwert nicht verändert.

Abhängigkeit der Überschreitungswahrscheinlichkeit der Prüfgrösse von der Stichprobengrösse
Die obenstehenden Formeln zur Berechnung von

$\sigma_{\overline{x}}$

und

$z_{\overline{x}}$

zeigen, dass bei festem

σ_0

und

$\overline{x}$

die für die Überschreitungswahrscheinlichkeit

$$p(z \geq z_{\bar{x}})$$

massgebliche Standardabweichung

$$\sigma_{\bar{x}}$$

der Stichprobenmittelwerte von der Stichprobengrösse n abhängig ist. Prüfen Sie diesen Sachverhalt, indem Sie einige unterschiedliche Stichprobengrössen eingeben und die resultierende Einordnung von

$$\bar{x}$$

in die Prüfverteilung und die sich daraus ergebende Überschreitungswahrscheinlichkeit beobachten.

Dieses Element (Animation, Video etc.) kann in der PDF version nicht dargestellt werden und ist nur in der online Version sichtbar. [link]

Dieses Element (Animation, Video etc.) kann in der PDF version nicht dargestellt werden und ist nur in der online Version sichtbar. [link]

Dieses Element (Animation, Video etc.) kann in der PDF version nicht dargestellt werden und ist nur in der online Version sichtbar. [link]	**Dieses Element (Animation, Video etc.) kann in der PDF version nicht dargestellt werden und ist nur in der online Version sichtbar. [link]**	Wählen Sie die Stichprobengrösse:			
		Dieses Element (Animation, Video etc.) kann in der PDF version nicht dargestellt werden und ist nur in der online Version sichtbar. [link]	$\sigma_{\overline{X}}$	$z_{\overline{X}}$	p
		Dieses Element (Animation, Video etc.) kann in der PDF version nicht dargestellt werden und ist nur in der online Version sichtbar. [link]	**Dieses Element (Animation, Video etc.) kann in der PDF version nicht dargestellt werden und ist nur in der online Version sichtbar. [link]**	**Dieses Element (Animation, Video etc.) kann in der PDF version nicht dargestellt werden und ist nur in der online Version sichtbar. [link]**	**Dieses Element (Animation, Video etc.) kann in der PDF version nicht dargestellt werden und ist nur in der online Version sichtbar. [link]**

Dieses Element (Animation, Video etc.) kann in der PDF version nicht dargestellt werden und ist nur in der online Version sichtbar. [link]

Dieses Element (Animation, Video etc.) kann in der PDF version nicht dargestellt werden und ist nur in der online Version sichtbar. [link]

Dieses Element (Animation, Video etc.) kann in der PDF version nicht dargestellt werden und ist nur in der online Version sichtbar. [link]

Dieses Element (Animation, Video etc.) kann in der PDF version nicht dargestellt werden und ist nur in der online Version sichtbar. [link]

Dieses Element (Animation, Video etc.) kann in der PDF version nicht dargestellt werden und ist nur in der online Version sichtbar. [link]

Ab welcher Stichprobengrösse n wird der Mittelwertsunterschied von 2 Punkten auf dem 5%-Niveau signifikant?

- ❍ n=25
- ❍ n=49
- ❍ n=64
- ❍ n=100
- ❍ n=144

Was können wir unserem Beispiel bis hierher entnehmen?

Es sind zwei an sich bekannte Tatsachen, auf die wir etwas näher eingehen wollen:

1. Wenn wir bei einer unspezifischen Alternativhypothese H_1

   $(\mu > \mu_0)$

   die Arbeitshypothese H_0

   $(\mu = \mu_0)$

   annehmen, so ist das

   β

   -Fehler-Risiko nicht einschätzbar. In anderen Worten: Wir können keine Angabe machen, mit welcher Wahrscheinlichkeit H_0 irrtümlicherweise beibehalten wird.
2. Die Frage, ob bei der Interpretation eines um 2 Punkte höheren Stichprobenmittelwertes der Zufall ausgeschlossen werden darf, ist auch von der Stichprobengrösse n abhängig.
   In anderen Worten: Ob unsere Arbeitshypothese H_0 verworfen werden kann, ist nicht nur vom Unterschied zwischen dem Stichproben- und dem Populationsmittelwert abhängig, sondern auch von der Stichprobengrösse n!
   Dies bedeutet: Bei genügend grossen Stichproben wird jede noch so kleine Differenz zwischen dem Stichproben- und dem Populationsmittelwert signifikant.

Forschungsresultate, die anhand entscheidungsstatistischer Verfahren gewonnen wurden, sind aus diesem Grunde bezüglich der Grösse der nachgewiesenen Unterschiede und der in der spezifischen Untersuchungen verwendeten Stichprobengrössen zu hinterfragen.
Im Theorieteil wollen wir uns das hierfür notwendige Wissen erarbeiten.

Theorie (1 - 6)

Übersicht

- 1. Spezifische Alternativhypothese und Effektgrösse
- 2. Das Alpha- und Beta-Fehler-Risiko
- 3. Die Teststärke [power]
- 4. Entscheid zwischen Arbeitshypothese und spezifischer Alternativhypothese
- 5. Optimale Stichprobengrösse
- 6. Beurteilung eines effektiv nachgewiesenen Effektes

1. Spezifische Alternativhypothese und Effektgrösse

Im Rahmen unseres Einführungsbeispiels soll gezeigt werden, was wir unter einer spezifischen Alternativhypothese verstehen und was wir gewinnen, wenn wir der Arbeitshypothese H_0 eine spezifische Alternativhypothese gegenüberstellen.
Wir erinnern uns: In der Eichpopulation 10-11jähriger mitteleuropäischer Kinder sind die Leistungen in einem eingeführten Agilitätstest mit den Parametern

μ_0

= 150,

σ_0

= 10 näherungsweise normalverteilt.
evor wir nun die mittlere Leistung einer Stichprobe verhaltensauffälliger Kinder von 10-11 Jahren (

$\bar{x}$

= 152, s = 9; n = 49) mit dem Mittelwert der Eichpopulation vergleichen, muss aus inhaltlicher Sicht entschieden werden, wie gross der Unterschied zwischen dem Mittelwert

μ_0

der Eichpopulation und dem Mittelwert

μ_1

, der hinter der Stichprobe stehenden Population verhaltensauffälliger Kinder sein muss, damit von einem inhaltlich relevanten Unterschied zwischen Population und Stichprobe gesprochen werden kann.
Von einem Experten der Sonderpädagogik und Kenner des betreffenden Testverfahrens erfahren wir, dass der höhere Mittelwert der Stichprobe nur vernünftig interpretiert werden kann, wenn diese aus einer Population stammt, deren Mittelwert

μ_1

um mindestens 3 Punkte über demjenigen der Eichpopulation liegt.
Wir benutzen diese inhaltliche Information für die Formulierung einer **spezifischen Alternativhypothese H_1** die postuliert, dass die Stichprobe aus einer Population mit dem Mittelwert

μ_1

153 und der Standardabweichung

σ_1

=

σ_0

= 10 stammt. Damit stellen wir unserer Arbeitshypothese eine ganz spezifische Alternative gegenüber:
Arbeitshypothese H_0:

$\mu = \mu_0 = 150$

;
Spezifische Alternativhypothese H_1:

$\mu \geq \mu_1 = 153$

Mit der spezifischen Alternativhypothese haben wir den minimalen Effekt festgelegt, der für eine sinnvolle inhaltliche Interpretation erforderlich ist.
Für die formale Beschreibung dieses minimalen Effektes benutzen wir die **Effektgrösse**

ε_{soll}

. **[effect size]**.
Die Effektgrösse beschreibt die aus inhaltlichen Gründen geforderte minimale Differenz (

μ_1

-

μ_0

) die, im Hinblick auf die Vergleichbarkeit verschiedener Untersuchungen, mit der Standardabweichung

σ_0

des untersuchten Merkmals normiert wird. Für unseren einfachen Fall ist die Effektgrösse

ε_{soll}

wie folgt definiert:

$$\varepsilon_{soll} = \frac{(\mu_1 - \mu_0)}{\sigma_0} = \frac{(153 - 150)}{10} = 0{,}30$$

Eine Effektgrösse dieser Ausprägung wird als gering bezeichnet. Dies deshalb, weil bei einem Merkmal mit einer Standardabweichung von 10 Punkten für eine sinnvolle Interpretation nur gerade eine Differenz von 3 Punkten notwendig ist: Schon ein - im Vergleich zur Standardabweichung des Merkmals - geringer Effekt, so sagen wir, ist inhaltlich relevant.

Mit der Wahl einer spezifischen Alternativhypothese H_1 haben wir festgelegt, wie gross der Unterschied zwischen

μ_1

und

μ_0

sein muss, damit wir von einem inhaltlich bedeutungsvollen Effekt sprechen können.

In der Praxis erfordert die sachgerechte Wahl von H_1 eine intensive inhaltliche Auseinandersetzung mit der Definition und Charakteristik des interessierenden Merkmals. Der Einsatz lohnt sich aber. Ein festgelegter minimaler Effekt erleichtert die Konzeption einer Untersuchung und die Interpretation der Ergebnisse ganz entscheidend.

Um Effekte über verschiedene Untersuchungen vergleichbar zu machen, normieren wir die Differenz (

μ_1

-

μ_0

) mit der Standardabweichung des Merkmals und kommen so zur **Effektgrösse**

ε_{soll}

.

Diese Definition der Effektgrösse

ε_{soll}

haben wir uns im Rahmen eines einfachen Beispiels zum Vergleich eines Stichprobenmittelwertes mit einem Populationsmittelwert überlegt. Zudem sind wir von der idealisierten Bedingung ausgegangen, dass sich die Verteilungen des Merkmals unter H_0 und H_1 nur im Mittelwert und nicht in der Varianz unterscheiden.

Effektgrössen können im Zusammenhang mit allen entscheidungsstatistischen Verfahren definiert werden. Da in der Praxis aber die idealisierte Bedingung identischer Varianzen unter H_0 und H_1 nicht erfüllt ist, müssen die Bestimmungsformeln korrigiert werden. Wir ersparen uns die Herleitung dieser Formeln und begnügen uns — neben dem grundsätzlichen Verständnis — damit, die entsprechenden Formeln in der Tabelle "Effektgrösse und optimaler Stichprobenumfang" nachzusehen.

Werfen Sie einen Blick auf diese Tabelle.

Sie stellen fest, dass die Tabelle auch die im Forschungsalltag etablierte **Klassifikation für die Ausprägung der Effektgrösse** zeigt.

Die Angaben zu den optimalen Stichprobenumfängen können wir an dieser Stelle noch nicht verstehen, dies wird sich bald ändern.

Bevor wir aber darauf eingehen können, müssen wir noch kurz auf einen weiteren Vorteil spezifischer Alternativhypothesen zu sprechen kommen.

2. Alpha- und Beta-Fehler-Risiko

Dieses Element (Animation, Video etc.) kann in der PDF version nicht dargestellt werden und ist nur in der online Version sichtbar. [link]

Mit der Definition einer spezifischen Alternativhypothese H_1 wird die Einschätzung des

β

-Fehler-Risikos bezüglich H_0 möglich!

Wir wollen uns dies wiederum für den Vergleich eines Stichprobenmittelwertes mit einem Populationsmittelwert überlegen:

Ein Merkmal sei in einer Population mit den Parametern

μ_0

und

σ_0

normalverteilt. In einer Stichprobe der Grösse n ist das Merkmal mit den Kennwerten

$\bar{x}$

und s eingipflig verteilt. Es soll nun entschieden werden, ob die Stichprobe aus einer Population mit dem Mittelwert

$\mu = \mu_0$

oder aus einer Population mit einem Mittelwert

$\mu \geq \mu_1$

stammt. Formal muss also zwischen den folgenden Hypothesen entschieden werden:

Arbeitshypothese H_0:

$\mu = \mu_0$

Spezifische Alternativhypothese H_1:

$\mu \geq \mu_1$

Die Abbildung zeigt die Einordnung des Stichprobenmittelwertes x_quer.gif in die unter H_0 und H_1 zu erwartenden Verteilungen der Stichprobenmittelwerte. Vereinfachend ist dabei wiederum angenommen, dass sich die Verteilungen des Merkmals unter H_0 und H_1 nur im Mittelwert unterscheiden.

Lassen Sie sich die verschiedenen Überschreitungswahrscheinlichkeiten und Fehlerrisikos durch Anklicken der einzelnen Links in die Graphik einzeichnen (beobachten Sie dabei die entsprechenden Einfärbungen) und ergänzen Sie die nachfolgende Folgerung zum b-Fehler-Risiko bezüglich H_0.

Dieses Element (Animation, Video etc.) kann in der PDF version nicht dargestellt werden und ist nur in der online Version sichtbar. [link]	Bei Annahme der Gültigkeit von H_0:
	Dieses Element (Animation, Video etc.) kann in der PDF version nicht dargestellt werden und ist nur in der online Version sichtbar. [link]
	Bei Annahme der Gültigkeit von H_1:
	Dieses Element (Animation, Video etc.) kann in der PDF version nicht dargestellt werden und ist nur in der online Version sichtbar. [link]
Einordnung des Stichprobenmittelwertes $\bar{x}$ in die Verteilungen der Stichprobenmittelwerte unter Annahme der Gültigkeit von H_0 resp. H_1	

Liegt zu einer Arbeitshypothese H_0 eine spezifische Alternativhypothese H_1 vor, so entspricht das auf H1 bezogene

α

-Fehler-Risiko dem

β

-Fehlerrisiko bezüglich H_0.

Folgerung

Entspricht das Risiko, dass H_0 irrtümlicherweise abgelehnt wird, der Wahrscheinlichkeit, dass H_1 irrtümlicherweise abgelehnt oder angenommen wird?

- ❍ abgelehnt
- ❍ angenommen

Liegt also zu einer Arbeitshypothese H_0 eine spezifische Alternativhypothese H_1 vor, so entspricht das auf H_1 bezogene Alpha-Fehler-Risiko dem Beta-Fehler-Risiko bezüglich _____. Damit ist das Beta-Fehler-Risiko bezüglich H_0 bekannt!

- ❍ H_0
- ❍ H_1

So kennen wir für einen konkreten Stichprobenmittelwert - im Hinblick auf die Arbeitshypothese H_0 - sowohl das

α

- als auch das

β

-Fehler-Risiko.

3. Teststärke [power]

Bevor wir unser Beispiel mit diesen neuen Erkenntnissen noch einmal aufgreifen, wollen wir noch kurz auf den Begriff der **Teststärke [power]** eines Prüfverfahrens eingehen. Dabei geht es um folgendes:
Wir wissen, dass das

β

-Fehler-Risiko der Wahrscheinlichkeit entspricht, mit der die Alternativhypothese H1 irrtümlicherweise abgelehnt und damit H0 irrtümlicherweise angenommen wird. Dies bedeutet, dass (1-

β

) der Wahrscheinlichkeit entspricht, mit der eine richtige H1 auch angenommen und H0 korrekterweise abgelehnt wird. Diese Wahrscheinlichkeit (1-

β

) nennt man die Teststärke eines Signifikanztests.
In anderen Worten: Die **Teststärke (1-**

β

) eines Prüfverfahrens gibt die Wahrscheinlichkeit an, mit der eine korrekte Alternativhypothese H1 auch angenommen wird.

4. Entscheid zwischen Arbeitshypothese und spezifischer Alternativhypothese

Kennen wir für einen konkreten Stichprobenmittelwert - im Hinblick auf die Antworthypothese H0 - sowohl das

α

- als auch das

β

-Fehler-Risiko, so stellt sich natürlich die Frage, unter welchen Bedingungen H0 zu Gunsten von H1 abgelehnt werden soll.
Bezüglich des

α

-Fehler-Risikos kennen wir die Antwort: Wir können H0 verwerfen, wenn das

α

-Fehler-Risiko kleiner wird als 5%. Wie steht es nun aber mit dem

β

-Fehler-Risiko? Hierfür gibt es keine etablierten Grenzwerte, letztendlich ist es immer eine inhaltliche Frage, welches

β

-Fehler-Risiko als t ragbar erachtet wird. Eine Faustregel für Entscheidungen ohne spezielle Tragweite kann aber angegeben werden.

Faustregel

- Wir verwerfen H0 zugunsten von H1 bei einem Alpha-Fehler-Risiko
  $\leq$
  5% und einem Beta-Fehler-Risiko
  $\geq$
  20%.
- Wir nehmen H0 an und verwerfen H1 bei einem Alpha-Fehler-Risiko
  $\geq$
  20% und einem Beta-Fehler-Risiko
  $\leq$
  5%.

Zur Veranschaulichung greifen wir unser Beispiel noch einmal auf. Wir prüfen unsere Daten bezüglich einer spezifischen Alternativhypothese.
Wir erinnern uns: In der Eichpopulation 10-11jähriger mitteleuropäischer Kinder sind die Leistungen in einem eingeführten Agilitätstest mit den Parametern

μ_0

= 150,

σ_0

= 10 normalverteilt. Wir vergleichen die mittlere Leistung einer Stichprobe verhaltensauffälliger Kinder von 10-11 Jahren (

$\overline{x}$

= 152, s = 9; n = 49) mit dem Mittelwert der Eichpopulation. Testpsychologische Erfahrung lehrt, dass der höhere Mittelwert in der Stichprobe nur sinnvoll interpretiert werden kann, wenn diese aus einer Population stammt, deren Mittelwert um mindestens 3 Punkte über demjenigen der Eichpopulation liegt.
Wir verfahren zum **Vergleich des Stichproben- und des Populationsmittelwertes** wie gewohnt:

1. Zusammenstellung der Voraussetzungen, von denen man bei der Wahl des Prüfverfahrens ausgehen kann: Population: N(150,10); Stichprobe:
   $\overline{x}$

= 152; s = 9; n = 49;

2. Prüfgrösse ist

$\overline{x}$

.

3. Arbeitshypothese H0:

$$\mu = \mu_0 = 150$$

Spezifische Alternativhypothese H1:

$$\mu \geq \mu_1 = 153$$

Damit definieren wir die Effektgrösse:

$$\varepsilon_{soll} = \frac{(\mu_1 - \mu_0)}{\sigma_0} = \frac{(153 - 150)}{10} = 0{,}30$$

4. Unter der Annahme der Gültigkeit der Hypothesen H0 und H1 kennen wir die Verteilungen der Stichprobenmittelwerte: Für eine Stichprobe der Grösse n = 49 sind die Stichprobenmittelwerte mit den Verteilungsparametern

$\mu_{\overline{x}}$

und

$\sigma_{\overline{x}}$

normalverteilt.

Unter Annahme der Gültigkeit von H0:

$$\mu_{\overline{x}} = \mu_0 = 150$$

$$\sigma_{\overline{x}} = \sigma_0 / \sqrt{n} = 10 / \sqrt{49} = 1{,}43$$

Unter Annahme der Gültigkeit von H1:

$$\mu_{\overline{x}} = \mu_1 = 153$$

$$\sigma_{\overline{x}} = \sigma_0 / \sqrt{n} = 10 / \sqrt{49} = 1{,}43$$

5. Zur Bestimmung der einseitigen Über- resp. Unterschreitungswahrscheinlichkeit des Stichprobenmittelwertes

$\overline{x}$

= 152 transformieren wir diesen Mittelwert für H0 und H1 in die z-Verteilung.

Für H0:

$$z_{\bar{x}} = \frac{\bar{x} - \mu_{\bar{x}}}{\sigma_{\bar{x}}} = \frac{152 - 150}{1{,}43} = 1{,}40$$

Für H1:

$$z_{\bar{x}} = \frac{\bar{x} - \mu_{\bar{x}}}{\sigma_{\bar{x}}} = \frac{152 - 153}{1{,}43} = -0{,}70$$

Bestimmen Sie anhand der einschlägigen z-Tabelle (Verzeichnis der Tabellen) die einseitigen Überschreitungswahrscheinlichkeiten. Geben Sie die Werte gemäss der Tabelle ein (Format 7.0% oder 0.07 oder 0.071).

Alpha-Fehler-Risiko: p(z

$\geq$

1,4) = _____

Beta-Fehler-Risiko: p(z

$\leq$

-0,7) = _____

Entscheiden Sie anhand der ermittelten

α

- und

β

-Fehler-Risiken (8,1% resp. 24,2%) und unserer Faustregel für oder gegen H0:

Faustregel

- Wir verwerfen H0 zugunsten von H1 bei einem

  α

  -Fehler-Risiko

  $\leq$

  5% und einem

  β

  -Fehler-Risiko

  $\geq$

  20%
- Wir nehmen H0 an und verwerfen H1 bei einem

  α

  -Fehler-Risiko

  $\geq$

  20% und einem

  β

-Fehler-Risiko

$\leq$

5%.

Die Arbeitshypothese H0 wird:

- ❍ abgelehnt
- ❍ angenommen
- ❍ weder abgelehnt noch angenommen

Wir können H0 weder annehmen noch ablehnen, d.h. wir können bei den gegebenen Daten nicht zwischen H0 und H1 entscheiden. Eine äusserst unangenehme Situation! Was wir dagegen unternehmen können, zeigt der nächste Absatz "Optimale Stichprobengrösse für den Entscheid zwischen Arbeits- und Alternativhypothese".

5. Optimale Stichprobengrösse

In jedem entscheidungsstatistischen Verfahren sind - wie in unserem Beispiel - auf Grund inhaltlicher Überlegungen sowohl die Effektgrösse

ε_{soll}

als auch das

α

- und

β

-Fehler-Risiko festgelegt. Das einzige, was wir verändern können, ist der Umfang unserer Untersuchung, d.h. die Grösse n der untersuchten Stichprobe.
Zu diesem Thema haben wir in der Einführung zum Lernschritt eine Erkenntnis gewonnen: Wir stellten fest, dass die Signifikanz der Differenz zwischen einem Stichproben- und einem Populationsmittelwert von der Stichprobengrösse n abhängig ist. Bei genügend grosser Stichprobe werden auch kleinste Differenzen, die inhaltlich absolut irrelevant sein können, signifikant.
Als erstes nahmen wir diese Erkenntnis zum Anlass, um eine spezifische Alternativhypothese und damit die Effektgrösse

ε_{soll}

einzuführen. Jetzt nutzen wir dieses Wissen, um dem Gedankengang unseres Statistikers wenigstens im Grundsatz folgen zu können. Er erklärt uns folgendes:
Bei festgelegter Effektgrösse

ε_{soll}

und vorgewähltem

α

- und

β

-Fehler-Risiko existiert ein sogenannter optimaler Stichprobenumfang nopt, mit dem zwischen H0 und H1 eindeutig entschieden werden kann. Für den idealisierten Fall, den wir bisher zur Vereinfachung immer angenommen haben, und der davon ausgeht, dass sich die Verteilungen des Merkmals unter H0 und H1 nur im Mittelwert unterscheiden, für diesen idealisierten Fall lässt sich nopt wie folgt berechnen:

$$n_{opt} = \frac{(z_{(1-\alpha)} - z_{\beta})^2}{\varepsilon_{soll}^{\ 2}}$$

Wir schenken unserem Statistiker Vertrauen und verweisen Interessierte für die Herleitung dieser Beziehung auf die Fachliteratur (z.B. Bortz, Kap. 4.8).

Diese einfache Formel gilt wie gesagt für den idealisierten Fall. Für die Prüfung realer Daten sind leider für jedes Prüfverfahren aufwendige Korrekturen notwendig, so dass man sich in der Forschungspraxis wiederum mit tabellarischen Zusammenstellungen behilft. Diese nennen zu jedem Prüfverfahren für verschiedene Effektgrössen, fest gewähltes

α

-Fehler-Risiko und fest gewählte Teststärke (1-

β

) die optimalen Stichprobengrössen.

Diese Angaben dienen in der Praxis als Richtgrössen.

Sie finden eine solche Tabelle unter "Effektgrösse und optimaler Stichprobenumfang", diese ist der Fachliteratur entnommen (z.B. (Bortz 1995) , Kap. 9; (Bortz 1999) , Kap. 5).

Wenn Sie jetzt einen Blick auf diese Tabelle werfen, so werden Sie auch die Angaben zu 'n_{opt}' verstehen.

Bevor wir mit diesem Wissen unser Beispiel ein letztes Mal aufgreifen, muss noch einmal festgehalten werden, was unter dem etwas vagen Begriff **optimaler Stichprobenumfang** zu verstehen ist:

Der optimale Stichprobenumfang nopt gibt an, wie gross eine Stichprobe gewählt werden muss, damit bei vorgewählter Effektgrösse

ε_{soll}

und festgelegten

α

- und

β

-Fehler-Risiken eindeutig zwischen der Arbeitshypothese H0 und der spezifischen Alternativhypothese H1 entschieden werden kann.

Für die Planung einer Untersuchung ist dies eine wichtige Information, möchte man doch einerseits den Untersuchungsaufwand, d.h. den Umfang der zu untersuchenden Stichprobe in Grenzen halten, andererseits aber auch sicherstellen, dass auf Grund der erhobenen Daten eine eindeutige statistische Entscheidung möglich ist.

Zur Veranschaulichung greifen wir unser Einführungs-Beispiel ein letztes Mal auf und bestimmen die Grösse der zu untersuchenden Stichprobe, mit der eindeutig zwischen der Arbeits- und Alternativhypothese entschieden werden kann.

Wir möchten eine Stichprobe verhaltensauffälliger Kinder bezüglich ihrer mittleren Leistung in einem Agilitätstest mit dem in der Eichpopulation ermittelten Populationsmittelwert vergleichen. Dem Testmanual entnehmen wir für die betreffende Altersgruppe: Die Leistungen im Agilitätstest sind mit den Parametern

μ_0

= 150,

σ_0

= 10 normalverteilt. Von einem Sonderpädagogen und Kenner des Agilitätstests wissen wir zudem, dass eine sinnvolle inhaltliche Interpretation eines höheren Leistungsmittelwertes in der Stichprobe nur möglich ist, wenn diese aus einer Population stammt, deren Mittelwert um mindestens 3 Punkte über demjenigen der Eichpopulation liegt. Unsere Untersuchung muss also so angelegt sein, dass zwischen den folgenden Hypothesen eindeutig entschieden werden kann:
Arbeitshypothese H0:

$$\mu = \mu_0 = 150$$

Spezifische Alternativhypothese H1:

$$\mu \geq \mu_1 = 153$$

Damit ist die Effektgrösse e_{soll} definiert:

$$\varepsilon_{soll} = \frac{(\mu_1 - \mu_0)}{\sigma_0} = \frac{(153 - 150)}{10} = 0{,}30$$

Wir möchten für unsere Entscheidung ein

α

-Fehler-Risiko von 5% zulassen, d.h. wir begnügen uns bei der Ablehnung von H0 mit einer Irrtumswahrscheinlichkeit von 5%. Das

β

-Fehler-Risiko legen wir den Gepflogenheiten entsprechend bei 20% fest, womit sich eine Teststärke (1-

β

) von 80% ergibt. Dies bedeutet, dass wir einen tatsächlich höheren Leistungsmittelwert in der Stichprobe mit einer Wahrscheinlichkeit von 80% aufdecken werden.
Nun stellt sich die Frage, wieviele verhaltensauffällige Kinder untersucht werden müssen, damit wir eindeutig zwischen der Arbeitshypothese H0 und der spezifischen Alternativhypothese H1 entscheiden können.
Da wir in diesem Beispiel von idealisierten Bedingungen ausgehen und annehmen, dass sich die Eichpopulation und die hinter der Stichprobe stehende Population verhaltensauffälliger Kinder nur im Leistungsmittelwert, aber nicht in der Varianz unterscheiden, benutzen wir für die Bestimmung der optimalen Stichprobengrösse nopt die obenstehende Formel und kommen zu folgendem Schluss:

$$n_{opt} = \frac{(z_{(1-\alpha)} - z_\beta)^2}{\varepsilon_{soll}^{\ 2}} = \frac{(1{,}65 - (-0{,}84))^2}{0{,}3^2} = 68{,}89$$

Wird also eine Stichprobe von 69 Kindern untersucht, so kann auf dem 5% Signifikanzniveau eindeutig zwischen H0 und H1 entschieden werden.

6. Beurteilung eines effektiv nachgewiesenen Effektes

Nun soll noch kurz auf die Post-hoc-Beurteilung eines effektiv nachgewiesenen Effektes eingegangen werden, die bei Vorliegen der Untersuchungsresultate möglich ist.
Zeitigt z.B. der Vergleich der Mittelwerte aus zwei unabhängigen Stichproben ein signifikantes Resultat, so ist nachgewiesen, dass bei der Interpretation der Differenz in den Stichprobenmittelwerten der Zufall mit einer Irrtumswahrscheinlichkeit $p < 5\%$, 1% oder 0,1% ausgeschlossen werden darf. Mit diesem Befund ist nur sichergestellt, dass das Resultat nicht zufällig zustande gekommen ist. Inwieweit dem nachgewiesenen Effekt aber auch eine inhaltliche Bedeutung zukommt, kann auf Grund des erzielten Signifikanzniveaus nicht entschieden werden.
Obwohl die Frage nach der inhaltlichen Relevanz eines Resultates letztendlich immer nur vor dem Hintergrund der untersuchten Fragestellung und deren Operationalisierung (Wahl des untersuchten Merkmals, Verteilungscharakteristik etc.) beantwortet werden kann, bietet uns die Definition der Effektgrösse eine einfache Möglichkeit zur formalen Beschreibung des in der Untersuchung nachgewiesenen Effektes. Wir benutzen dabei dieselben Definitionsformeln, ersetzen aber die bei der Planung der Untersuchung geforderten minimalen, inhaltlich relevanten Differenzen durch die in der Untersuchung tatsächlich nachgewiesenen Differenzen. Zur Kennzeichnung bezeichnen wir die Post-hoc bestimmte Effektgrösse mit

$$\varepsilon_{eff}$$

.
Wir betrachten die Ermittlung von

$$\varepsilon_{eff}$$

am Beispiel eines Vergleichs der Mittelwerte aus zwei unabhängigen Stichproben. Für alle anderen Prüfverfahren gilt dann dasselbe in analoger Form.
Der Tabelle "Effektgrösse und optimaler Stichprobenumfang" entnehmen wir die Definition der Effektgrösse

$$\varepsilon_{soll}$$

für die Planung der Untersuchung und diejenige für die Post-hoc-Beurteilung des effektiv nachgewiesenen Effektes

$$\varepsilon_{eff}$$

.

2.a Planung einer Untersuchung

$$\varepsilon_{soll} = \frac{(\mu_1 - \mu_2)}{\hat{\sigma}}$$

$$(\mu_1 - \mu_2)$$

Minimale, inhaltlich relevante Differenz. Wir rechnen aus tabellarischen Gründen immer mit einem positiven Wert.

Optimaler Stichprobenumfang bei einseitigem Test, (1-b) = 0,80 und	**a = 0,05** **a = 0,01**
e = 0,20: schwacher Effekt	nopt = 310, nopt = 503
e = 0,50: mittlerer Effekt	nopt = 50, nopt = 82
e = 0,80: starker Effekt	nopt = 20, nopt = 33

2.b Beurteilung des effektiv nachgewiesenen Effektes

$$\varepsilon_{eff} = \frac{(\bar{x}_1 - \bar{x}_2)}{\hat{\sigma}}$$

$(\bar{x}_1 - \bar{x}_2)$

Differenz der Stichprobenmittelwerte. Wir rechnen aus tabellarischen Gründen immer mit einem positiven Wert.

$\hat{\sigma}$

Schätzwert für die Standardabweichung des Merkmals in der Population

Beurteilung der Effektgrösse wie bei der Planung einer Untersuchung.
Vergleichen Sie die beiden Definitionsformeln, so ist leicht einsehbar, dass für die Bestimmung des effektiv nachgewiesenen Effektes

ε_{eff}

anstelle der in der Planung geforderten minimalen Differenz

$(\mu_1 - \mu_2)$

die tatsächlich nachgewiesene Differenz

$(\bar{x}_1 - \bar{x}_2)$

eingesetzt wird. Für die anschliessende Klassifizierung des effektiv nachgewiesenen Effektes

ε_{eff}

(schwach, mittel, stark) gelten dieselben Richtwerte wie für die Planungsphase.

Fallbeispiele (1 - 4)

Übersicht

- 1. Unabhängige Stichproben - Vergleich der Mittelwerte
- 2. Abhängige Stichproben - Vergleich der Mittelwerte
- 3. Intervall- oder proportional skalierte Daten - Stochastischer Zusammenhang
- 4. Nominal skalierte Daten - Stochastischer Zusammenhang

1. Unabhängige Stichproben - Vergleich der Mittelwerte

Interessieren sich Studierende mehr für die Bildungspolitik als Studentinnen?
Eine Studentin plant eine Studienarbeit, in der untersucht werden soll, inwieweit sich Studenten mehr für die aktuelle Bildungspolitik interessieren als Studentinnen. Zu diesem Zweck möchte sie in zwei etwa gleich grossen Stichproben von Komilitoninnen und Komilitonen erheben, wieviele einschlägige Medienbeiträge (Artikel in der Tagespresse und in studentischen Publikationsorganen) im Durchschnitt pro Monat gelesen werden. Da in ihrem Bekanntenkreis dieses Mass bei beiden Geschlechtern breit streut, vermutet sie, dass die Varianz des Merkmals in den hinter den Stichproben stehenden Populationen relativ hoch ist.
Da unsere Studentin einen allfälligen Unterschied nur dann als relevant erachtet, wenn die zwei Stichproben aus unterschiedlichen Populationen stammen, deren Mittelwerte sich um mindestens die Hälfte der (noch nicht bekannten) Standardabweichung

σ

unterscheiden, wählt sie für ihre Untersuchung die Effektgrösse

ε_{soll}

= 0,5 und entscheidet sich damit für folgende Hypothesen:
Arbeitshypothese

$H_0 : \mu_1 = \mu_2$

Spezifische Alternativhypothese

$H_1 : \mu_1 \geq \mu_2 (= \mu_1 + 0,5 \cdot \sigma)$

Für das

α

-Fehler-Risiko (Wahrscheinlichkeit, dass H0 irrtümlicherweise zu Gunsten von H1 abgelehnt wird) möchte sie nur 1% zulassen, als

β

-Fehler-Risiko (Wahrscheinlichkeit, dass H0 irrtümlicherweise angenommen wird) scheinen ihr die üblichen 20% vertretbar.
Nun stellt sich die Frage, wie gross die beiden Stichproben gewählt werden sollen, damit zwischen H0 und H1 eindeutig entschieden werden kann.
Ein einschlägiger Richtwert findet sich in der Tabelle "Effektgrösse und optimaler Stichprobenumfang".
Welchen Umfang sollten beide Stichproben in etwa haben?
n = _____

Mit welcher Wahrscheinlichkeit darf unsere Studentin damit rechnen, dass sie einen tatsächlichen Unterschied zwischen Studentinnen- und Studentenstichprobe auch aufdeckt?
p = _____

Wir treffen unsere Studentin nach einigen Monaten wieder und erkundigen uns nach dem Resultat ihrer Untersuchung. Sie berichtet uns folgendes:

Untersucht wurde je eine Stichprobe von 82 Studentinnen und Studenten. Die Mittelwerte der pro Monat gelesenen Medienbeiträge zur Bildungspolitik unterscheiden sich in den beiden Stichproben signifikant auf dem 1%-Niveau. In der Stichprobe der Studenten beträgt der Mittelwert 1,21, in der Stichprobe der Studentinnen 0,82. Die Standardabweichungen der Daten in den beiden Stichproben sind nahezu gleich gross, die geschätzte Standardabweichung in der Population

$\hat{\sigma}$

beträgt 0,41.

Bestimmen Sie aus diesen Angaben die effektiv nachgewiesene Effektstärke

ε_{eff}

(Tabelle "Effektgrösse und optimaler Stichprobenumfang").

(Runden Sie jedes Zwischenresultat auf 2 Stellen nach dem Komma und rechnen sie mit diesem Resultat weiter.)

ε_{eff}

= _____

Als was ist der effektiv nachgewiesene Effekt zu beurteilen?

- ❍ schwach
- ❍ mittel
- ❍ stark

Unsere Studentin konnte also einen starken Effekt nachweisen, dennoch ist sie ein wenig konsterniert. Hätte sie mit einem derartigen Resultat gerechnet, wäre sie mit wesentlich kleineren Stichproben ausgekommen und hätte sich einige Arbeit erspart.

Wie gross hätten die Stichproben gewählt werden müssen, wenn man bei gleichen Rahmenbedingungen einen starken Effekt erwartet hätte (Tabelle "Effektgrösse und optimaler Stichprobenumfang")?

Bei einem erwarteten starken Effekt, einem a-Fehler-Risiko von 0,01 und einer Teststärke (1-b) = 0,8 wären Stichproben von je _____ Personen ausreichend gewesen.

2. Abhängige Stichproben - Vergleich der Mittelwerte

Wie gross muss die Schülerstichprobe sein, um über die Wirkung eines Treatments entscheiden zu können?

Im Zusammenhang mit der Überarbeitung der Lehrpläne und Lehrmethoden für Schülerinnen und Schüler des 7. Schuljahres wurde eine neue Lehreinheit zum Thema 'Ausländerfeindlichkeit' entwickelt, die im Rahmen spezieller Arbeitswochen bearbeitet werden kann. Das Ziel der Erfinder dieser Lehreinheit war es, die soziale Wahrnehmung der ausländischen Wohnbevölkerung in positiver Weise zu beeinflussen, d.h. die soziale Toleranz zu erhöhen resp. die soziale Rigidität zu vermindern. Da die Organisation von speziellen Arbeitswochen sowohl in organisatorischer als auch in finanzieller Hinsicht recht aufwendig ist, stellt sich die Frage, inwieweit die neue Lehreinheit die erhoffte Veränderung auch bewirkt.

Zur Konzeption einfachster Untersuchungen: Effektgrösse, Teststärke und optimale Stichprobengrösse

Zur Klärung dieser Frage wird eine empirische Untersuchung geplant. Eine Stichprobe von Schülern soll in einer speziellen Arbeitswochen nach dem neuen Lehrprogramm unterrichtet werden, wobei vor und nach der Arbeitswoche das Konstrukt 'Soziale Rigidität' mit einem eingeführten, standardisierten Fragebogen zu erheben ist. Die erhoffte Einstellungsänderung müsste sich in Form einer Reduktion des Mittelwertes der 'Sozialen Rigidität' nachweisen lassen.

Die geneigte Leserschaft erkennt das Untersuchungsdesign: Geplant ist eine Untersuchung mit Messwiederholung.

Nun stellt sich natürlich die Frage, wieviele Schüler in die Untersuchung einbezogen werden müssen, damit über die Wirkung des Treatments (hier Arbeitswoche) eindeutig entschieden werden kann.

Welche der folgenden Grössen müssen bekannt sein, um die optimale Zahl der zu untersuchenden Schulklassen abzuschätzen? Benutzen Sie für die Beantwortung dieser Frage die Tabelle "Effektgrösse und optimaler Stichprobenumfang" und markieren Sie die zutreffenden Grössen in der ersten Spalte.

- ❑ α -Fehler-Risiko
- ❑ β -Fehler-Risiko
- ❑ Freiheitsgrad
- ❑ Korrelation zwischen den Erhebungsdaten
- ❑ Effektgrösse
- ❑ Gesamtmittelwert der Daten

Die optimale Stichprobengrösse lässt sich abschätzen, wenn die gewünschte Effektgrösse, die geschätzte Korrelation zwischen den Daten der ersten und der zweiten Erhebung, das α -Fehler-Risiko und die Teststärke (1-β) angegeben werden können.

Das α - und das β -Fehler-Risiko können einfach gewählt werden, Die Bestimmung der Effektgrösse und die Schätzung der zu erwartenden Korrelation zwischen den Erhebungsdaten erfordert die folgenden Informationen und inhaltlichen Vorüberlegungen:

1. Das Konstrukt 'Soziale Rigidität' wird mit einem standardisierten Fragebogen erhoben, der die individuelle 'Soziale Rigidität' auf einer Intervallskala einschätzt. Dem Testmanual entnehmen wir, dass die Ausprägungen dieses Konstruktes in der Eichpopulation von Jugendlichen im Alter von 13-14 Jahren mit dem Mittelwert 40 und der Standardabweichung 6 näherungsweise normalverteilt sind.

2. Kenner des eingesetzten Tests gehen davon aus, dass von einer inhaltlich bedeutungsvollen Veränderung gesprochen werden darf, wenn sich die Mittelwerte der hinter den beiden Erhebungen stehenden Populationen um mindestens 3 Punkte unterscheiden, was der Hälfte der erwarteten Standardabweichung entspricht.
3. Wie bei allen Untersuchungen mit Messwiederholungen ist davon auszugehen, dass die Daten der beiden Erhebungen korreliert sind. Da dazu praktisch keine Vorinformationen vorliegen, schätzen wir konservativ:

   $\hat{r}$

   = 0,6

Auf Grund dieser Annahmen können Sie nun die geforderte Effektgrösse bestimmen und anschliessend die optimale Stichprobengrösse abschätzen. Benutzen Sie für die Beantwortung der folgenden Fragen die Tabelle "Effektgrösse und optimaler Stichprobenumfang".

Wie gross ist die geforderte Effektgrösse ?
(Runden Sie jedes Zwischenresultat auf 2 Stellen nach dem Komma und rechnen sie mit diesem Resultat weiter.)

ε_{soll}

= _____

Wie gross soll die Stichprobe sein, wenn für das Alpha-Fehler-Risiko 1% und die Teststärke (1 – Beta) =0,8 gewählt wird?

n = _____

Mit einer Stichprobe von 14 Schülern kann über das Vorhandensein des geforderten starken Effektes eindeutig entschieden werden.

Nun machen wir einen Zeitsprung und betrachten die Ergebnisse der Untersuchung. Diese können wie folgt zusammengefasst werden:

1. Die Mittelwerte der ersten und zweiten Erhebung unterscheiden sich um -2,3 Punkte

   $[(\bar{x}_1 - \bar{x}_2) = -2,3]$

   und

   $p \leq 5\%$

   mit signifikant.
2. Die Standardabweichung des Konstruktes 'Soziale Rigidität' fällt in der Stichprobe etwas grösser aus als im Manual angegeben. Wir schätzen sigma.gif auf dieser Grundlage:

   $\hat{\sigma}$

   = 8,6
3. Auch die Korrelation zwischen den Daten der ersten und zweiten Erhebung ist grösser als angenommen: r = 0,76

Bestimmen Sie aus diesen Angaben die Grösse des effektiv nachgewiesenen Effektes

ε_{eff}

. (Runden Sie jedes Zwischenresultat auf 2 Stellen nach dem Komma und rechnen sie mit diesem Resultat weiter.)

ε_{eff}

= _____

Der effektiv nachgewiesene Effekt ist damit geringer als der bei der Planung geforderte.
Als was ist der effektiv nachgewiesene Effekt zu beurteilen?

- ❍ schwach
- ❍ mittel
- ❍ stark

Die Untersuchung hat mit

ε_{eff}

= 0,55 nur einen mittleren Effekt nachweisen können. Dieses Resultat stellt die Schulbehörde vor ein nicht ganz einfaches Problem. Einerseits konnte im Rahmen eines überschaubaren Schulversuches und mit einer Irrtumswahrscheinlichkeit $p<5\%$ nachgewiesen werden, dass die neue Lerneinheit die Einstellungen der Schülerinnen und Schüler tatsächlich in der erhofften Weise beeinflusst. Andererseits liess sich die gewünschte Effektgrösse nicht nachweisen. Belegbar ist nur ein mittlerer Effekt. Die Schulbehörde hat nun zu entscheiden, ob die Lehreinheit mit dem Ziel einer Erhöhung der Effektivität nochmals überarbeitet werden soll, oder ob man sich im Hinblick auf die Komplexität der Thematik auch mit einer mittleren Effektgrösse zufrieden geben soll.

3. Intervall- oder proportional skalierte Daten - Stochastischer Zusammenhang

Externe Validität eines Testverfahrens

Zur Selektion von Führungskräften aus einer meist grossen Zahl von Bewerberinnen und Bewerbern werden häufig mehrtägige, äusserst kostenaufwendige Auswahlverfahren, sog. Assessments, eingesetzt. Im Rahmen dieser Assessments werden die Kandidatinnen und Kandidaten von Experten bezüglich verschiedener Fähigkeiten wie z.B. "Problemmanagement", "planerisches Denken", "soziale Kompetenz" etc. eingeschätzt und benotet.

Ein Doktorand hat nun ein PC-gestütztes Erhebungsinstrument entwickelt, mit dem das Konstrukt "planerisches Denken" mit geringem Aufwand eingeschätzt werden kann. Nach Abschluss der Entwicklungsarbeit und zahlreichen Vorversuchen mit studentischen Probandinnen und Probanden muss unser Doktorand den Nachweis erbringen, dass man mit seinem neuen Instrument zu einer Einschätzung des Konstruktes "planerisches Denken" kommt, die mit den Expertenurteilen aus den viel teureren Assessments vergleichbar ist.

Zu diesem Zweck plant er eine Untersuchung, in der eine Stichprobe von Bewerberinnen und Bewerbern sowohl im Rahmen des Assessments einer kooperierenden Firma als auch anhand des neuen Erhebungsinstrumentes bezüglich des Konstruktes "planerisches Denken" eingeschätzt werden. Wenn sich

die Hoffnungen unseres Doktoranden erfüllen, müssten die beiden Einschätzungen hoch korrelieren. Bei der Vorbesprechung mit dem Personalchef der kooperierenden Firma stellt sich die Frage nach der Stichprobengrösse: Wieviele Kandidatinnen und Kandidaten müssen in die Untersuchung einbezogen werden, damit ein aussagekräftiges Resultat gewonnen werden kann?
Wir begleiten unseren Doktoranden bei seinen Überlegungen zur Wahl der optimalen Stichprobengrösse.
Doch erst eine Zwischenfrage: Wann wird eine Stichprobengrösse als optimal bezeichnet
Markieren Sie die zutreffende Aussage in der ersten Spalte:

- ❍ Stichprobengrösse, bei der sicher für H0 entschieden werden kann
- ❍ Stichprobengrösse, bei der sicher H1 angenommen und H0 abgelehnt werden kann
- ❍ Stichprobengrösse, bei der zwischen H0 und H1 eindeutig entschieden werden kann
- ❍ Stichprobengrösse, bei der H1 sicher zu Gunsten von H0 abgelehnt werden kann

Wir sprechen von einem optimalen Stichprobenumfang, wenn eindeutig zwischen H0 und H1 entschieden werden kann.
Die Nullhypothese H0 und die Alternativhypothese H1 formuliert unser Doktorand wie folgt:
Nullhypothese H0:

$$\rho = 0$$

;
Alternativhypothese H1:

$$\rho \geq 0,7$$

Nun zur Abschätzung des optimalen Stichprobenumfangs:
Welche der folgenden Grössen müssen bekannt sein, um die optimale Stichprobengrösse abzuschätzen? Benutzen Sie für die Beantwortung dieser Frage die Tabelle "Effektgrösse und optimaler Stichprobenumfang" und markieren Sie die zutreffenden Grössen in der ersten Spalte.

- ❑ Freiheitsgrad
- ❑ Effektgrösse
- ❑ Gesamtmittelwert der Daten
- ❑ Teststärke
- ❑ a-Fehler-Risiko

Diese gerichtete, spezifische Alternativhypothese H1 bringt zum Ausdruck, dass nur dann von einem relevanten stochastischen Zusammenhang gesprochen werden soll, wenn die untersuchte Stichprobe aus einer Population stammt, in der die beiden Leistungseinschätzungen mit

$$\rho \geq 0,7$$

korreliert sind.
Die optimale Zahl der zu untersuchenden Kandidatinnen und Kandidaten lässt sich abschätzen, wenn die gewünschte Effektgrösse, das

α

-Fehler-Risiko und die Teststärke (1-

β

) angegeben werden können.
Das

α

- und das beta.gif-Fehler-Risiko können den Gepflogenheiten entsprechend gewählt werden, und auch die Effektgrösse

ε_{soll}

lässt sich leicht bestimmen. Der Tabelle "Effektgrösse und optimaler Stichprobenumfang" entnehmen wir, dass in diesem Fall

ε_{soll}

der minimalen, inhaltlich relevanten Korrelation

$\hat{r}$

entspricht.
Da unser Doktorand nur eine Korrelation

$\hat{r}$

grösser 0,7 als inhaltlich relevant erachtet, arbeitet er mit einer Effektstärke

ε_{soll}

= 0,7.
Überlegen Sie sich anhand der Tabelle "Effektgrösse und optimaler Stichprobenumfang" , wie gross die Stichprobe gewählt werden soll, wenn für das

α

-Fehler-Risiko = 0,01 und die Teststärke (1 -

β

) = 0,8 gewählt wird?
Sie stellten fest, dass unsere Tabelle nur Angaben zu Effektgrössen bis 0,5 macht. Für diese Effektgrösse wäre eine Stichprobe von 36 Personen optimal. Für die von unserem Probanden geforderte Effektstärke von 0,7 gibt es keine Angaben; wir müssen uns mit einer Abschätzung begnügen. Dazu folgende Überlegung: Der Tabelle entnehmen wir, dass der optimale Stichprobenumfang mit grösser werdender Effektstärke abnimmt. So verstehen wir, dass unser Doktorand dem Personalchef erläutert, dass eine Stichprobe von ca. 30 Probandinnen und Probanden ausreichen sollte.
Auch die im Anschluss an die Untersuchung vorzunehmende Beurteilung des tatsächlich nachgewiesenen Effektes fällt in diesem Fall recht einfach aus. So ergab sich in der Untersuchung unseres Doktoranden eine auf dem 1%-Niveau signifikante Korrelation r = 0,76. Die effektiv nachgewiesene Effektgrösse

ε_{eff}

ist somit 0,76 und liegt über der in der Planungsphase geforderten minimalen Effektgrösse

ε_{soll}

.
Unser Doktorand freut sich, und wir schliessen das Fallbeispiel 3.

4. Nominal skalierte Daten - Stochastischer Zusammenhang

Schulabschluss und Mediennutzung

In einer studentischen Untersuchung zur Mediennutzung soll untersucht werden, inwieweit zwischen dem Schulabschluss von 20-30-Jährigen einerseits und den Präferenzen bei der Wahl von Fernsehfilmen ein stochastischer Zusammenhang besteht. Zu diesem Zweck soll eine Stichprobe von Probandinnen und Probanden bezüglich des letzten Schulabschlusses und des Typs der von ihnen am häufigsten ausgewählten Fernsehfilme befragt werden.

Die Studierenden wählten ein sehr einfaches Untersuchungsdesign, das mit folgender 3x4-Feldertafel charakterisiert werden kann:

		Bevorzugter Filmtyp			
		Krimi	Soap Opera	Dokumentar-film	Actionfilm
Letzter Schulabschluss	Volksschule				
	Lehrabschluss				
	Hochschul- oder Fachhochschulabschluss				

Als nächstes beschäftigt unsere Studentengruppe die Frage, wieviele Probandinnen und Probanden befragt werden sollen.

Als erstes muss geklärt werden, wovon die optimale Stichprobengrösse im gegeben Fall abhängig ist. Benutzen Sie die Tabelle "Effektgrösse und optimaler Stichprobenumfang" um dies zu klären.

Benutzen Sie für die Beantwortung dieser Frage die Tabelle "Effektgrösse und optimaler Stichprobenumfang" und markieren Sie die zutreffenden Grössen in der ersten Spalte.

- ❑ Freiheitsgrad
- ❑ Effektgrösse
- ❑ Verteilung der Daten in der Population
- ❑ Teststärke
- ❑ α -Fehler-Risiko
- ❑ minimale Differenzen der absoluten Besetzungshäufigkeiten

Die optimale Stichprobengrösse kann abgeschätzt werden, wenn der Freiheitsgrad der k*l-Feldertafel, die Effektstärke, das α -Fehler-Risiko und die Teststärke (1-β) bekannt sind.

Der Freiheitsgrad df ist durch die Ausgestaltung unserer 3*4-Feldertafel vorgegeben, df = (k-1)*(l-1) = 6. Das α

-Fehler-Risiko und die Teststärke können nach den Gepflogenheiten gewählt werden. In Anbetracht der kulturellen Brisanz der Fragestellung möchten die Studierenden nur ein a-Fehler-Risiko von 0,01 eingehen, für die Teststärke (1-

β

) wählen sie wie üblich 0,8.
Die Bestimmung der Effektstärke

ε_{soll}

stellt uns nun aber vor ein unlösbares Problem: Vor der Befragung ist unklar, wie die Zellen unserer 3*4-Felder-Tafel besetzt sein werden. So kann weder die Nullhypothese H0 noch eine spezifische Alternativhypothese H1 explizit formalisiert werden. Wir können nur die Nullhypothese verbal formulieren, sie lautet: Zwischen den beiden Merkmalen "Schulabschluss" und "Bevorzugter Filmtyp" besteht kein stochastischer Zusammenhang. Für die Bestimmung der Effektgrösse reicht dies nicht aus, wir müssen die Effektgrösse

ε_{soll}

auf der Grundlage inhaltlicher Überlegungen festlegen.
Die Studierenden machen sich in dieser Hinsicht kundig und finden in der Literatur Untersuchungen mit ähnlichen Fragestellungen aber völlig anderen Designs. In diesen älteren Studien zeigte sich meist ein signifikanter aber nicht speziell ausgeprägter Zusammenhang zwischen dem Ausbildungsniveau der Befragten einerseits und den von ihnen bevorzugten Medienbeiträgen andererseits.
Auf der Grundlage dieses Vorwissens scheint es unseren Studierenden vertretbar, für die geplante Untersuchung einen mittleren Effekt zu erwarten. Sie wählen gemäss Tabelle "Effektgrösse und optimaler Stichprobenumfang"

ε_{soll}

= 0,3.
Damit sind alle Angaben zusammengetragen, der Umfang der zu untersuchenden Stichprobe kann abgeschätzt werden. Wieviele Personen müssen gemäss Tabelle "Effektgrösse und optimaler Stichprobenumfang" gesamthaft befragt werden?
nopt = _____

Gesamthaft müssen also 210 Personen befragt werden. Beim gegebenen Design mit drei Ausprägungen für die Dimension "Schulabschluss" wird man im Idealfall je 70 Personen mit Volksschulabschluss, mit abgeschlossener Berufslehre und mit Hoch- oder Fachhochschulabschluss befragen.
Damit ist die Frage der Studentengruppe nach der Zahl der zu befragenden Personen beantwortet, wir können eine Pause machen, bis die Untersuchungsergebnisse vorliegen.
Zeitsprung: Die Untersuchungsergebnisse liegen vor.

Dieses Element (Animation, Video etc.) kann in der PDF version nicht dargestellt werden und ist nur in der online Version sichtbar. [link]

Die Studentengruppe zeigt Ihnen stolz die Untersuchungsergebnisse. Anhand der in die 3*4-Feldertafel eingetragenen absoluten Häufigkeiten der Präferenzen konnte tatsächlich ein stochastischer Zusammenhang auf dem 1%-Niveau nachgewiesen werden.

		Bevorzugter Filmtyp			
		Krimi	Soap Opera	Dokumentar-film	Actionfilm
Letzter Schulabschluss	Volksschule	14	23	11	22
	Lehrabschluss	11	13	17	29
	Hochschul- oder Fachhochschulabschluss	26	13	19	12

Nun stellt sich die Frage nach der Grösse des tatsächlich nachgewiesenen Effektes

ε_{eff}

. Wir konsultieren als erstes die Tabelle "Effektgrösse und optimaler Stichprobenumfang" und finden dabei die folgenden Angaben:

5.b Beurteilung des effektiv nachgewiesenen Effektes

Liegen die Untersuchungsdaten vor, so ist die tatsächliche Besetzung der Zellen bekannt, zudem kann die bei Gültigkeit von H0 zu erwartende Besetzung bestimmt werden: Daraus lässt sich

ε_{eff}

bestimmen.

$$\varepsilon_{eff} = \sqrt{\sum_{i=1}^{m} \frac{(P_{0i} - P_{1i})^2}{P_{0i}}}$$

P_{0i} steht für die gemäss H0 zu erwartenden **relativen** Besetzungshäufigkeiten der m Zellen Diese berechnen sich wie folgt: P_{0i} = (Zeilensumme*Spaltensumme)/n^2

n: Gesamtzahl der Beobachtungen

P_{1i} steht für die relativen Besetzungshäufigkeiten, die sich in der Untersuchung ergaben

m = k*l: Anzahl Zellen der k*l-Felder-Tafel

df = (k – 1)*(l – 1): Freiheitsgrad

Beurteilung der Effektgrösse wie bei der Planung einer Untersuchung

Zur Bestimmung von

ε_{eff}

müssen also die **relativen** Besetzungshäufigkeiten P_{0i} ermittelt werden, die bei Gültigkeit von H0 zu erwarten sind, dann sind die in der Untersuchung festgestellten **relativen** Besetzungshäufigkeiten P_{1i} zu berechnen.

Um Ihnen den Arbeitsaufwand etwas zu erleichtern, wurden diese Daten in die 3*4-Feldertafel eingetragen. In jeder Zelle stehen dabei drei Zahlen nach dem folgenden Schema:

P_{0i}
P_{1i}
$\frac{(P_{0i} - P_{1i})^2}{P_{0i}}$

		Bevorzugter Filmtyp			
		Krimi	Soap Opera	Dokumentar-film	Actionfilm
Letzter Schulabschluss	Volksschule	0,08 0,07 0,0013	0,08 0,11 0,0113	0,07 0,05 0,0057	0,10 0,10 0,0000
	Lehrabschluss	0,08 0,05 0,0113	0,08 0,06 0,0050	0,07 0,08 0,001	0,10 0,14 0,0160
	Hochschul- oder Fachhochschulabschluss	0,08 0,12 0,0200	0,08 0,06 0,0050	0,07 0,11 0,0057	0,10 0,11 0,0160

Benutzen Sie diese Zahlen, um die effektiv nachgewiesene Effektgrösse

ε_{eff}

zu bestimmen. (Runden Sie das Schlussresultat auf 2 Stellen nach dem Komma.)

ε_{eff}

= _____

Wie wird die nachgewiesene Effektgrösse

ε_{eff}

klassifiziert?

Was für ein Effekt konnte nachgewiesen werden?

- ❍ schwacher
- ❍ mittlerer
- ❍ starker