

FAQ: Programm solara.MP

Beurteilung attributiver Prüfprozesse

Teil 4: Berechnungsmodelle

87298374 0987298374982739
8470 2 **Q-DBM** 7 1545 82138 12
7198723987 987239 98729872
PROCELLA 234 154 13 544 565
9872 2719827 7 27198723987
45 8912 687723 **VIDARA** 27198
21245 666 1214432 329 **Q-QIS**
928 234 345 344 4718723987
0187309 445 455 4877298374
M-QIS DASHBOARD 772728498
81 4981 **DESTRA** 918 2589 23
59 **QS-STAT** 49814981 45598
M-QIS ENGINE 49983 259 1547
7487 29837409872 98374982
73984702 **SOLARA.MP** 987349
9283 120 38 485 0 2 38 49081



CONTENTS

Vorwort	9
1. Beispieldatensatz für dieses Fallbeispiel	10
2. Kappa-Koeffizienten	11
2.1 Kappa nach Cohen	11
2.1.1 Aufbauschema Datentabelle	11
2.1.2 Kreuztabellen-Berechnungsschema	13
2.1.3 Berechnungsschema für den Kappa-Koeffizienten (Cohen)	14
2.1.4 Berechnungsschema für die Varianz des Kappa-Koeffizienten	15
2.1.5 Signifikanztest	16
2.1.5.1 Prüfgröße z	16
2.1.5.2 P-Wert	17
2.2 Kappa nach Cohen - Berechnungsart „AIAG MSA Standard“	18
2.2.1 Vergleichbarkeit aller Prüfer ohne Referenzvergleich („AIAG MSA Standard“)	18
2.2.1.1 Virtuelle Einheiten-Gesamtzahl ($A \times B$ „AIAG MSA Standard“)	18
2.2.1.2 Aufbauschema Datentabelle ($A \times B$ „AIAG MSA Standard“)	19
2.2.1.3 Kreuztabellen ($A \times B$ „AIAG MSA Standard“)	20
2.2.1.4 Kappa-Koeffizient ($A \times B$ „AIAG MSA Standard“)	22
2.2.1.5 Signifikanztest ($A \times B$ „AIAG MSA Standard“)	23
2.2.1.6 Standardabweichung ($A \times B$ „AIAG MSA Standard“)	24
2.2.1.7 Prüfgröße z ($A \times B$ „AIAG MSA Standard“)	25
2.2.1.8 P-Wert ($A \times B$ „AIAG MSA Standard“)	25
2.2.2 Vergleichbarkeit: Prüfer A vs. Referenz („AIAG MSA Standard“)	26
2.2.2.1 Virtuelle Einheiten-Gesamtzahl ($A \times \text{Ref.}$ „AIAG MSA Standard“)	26
2.2.2.2 Aufbauschema Datentabelle ($A \times \text{Ref.}$ „AIAG MSA Standard“)	27
2.2.2.3 Kreuztabellen ($A \times \text{Ref.}$ „AIAG MSA Standard“)	28
2.2.2.4 Kappa-Koeffizient ($A \times \text{Referenz}$ „AIAG MSA Standard“)	29
2.2.2.5 Signifikanztest ($A \times \text{Referenz}$ „AIAG MSA Standard“)	30
2.2.2.6 Standardabweichung ($A \times \text{Ref.}$ „AIAG MSA Standard“)	30
2.2.2.7 Prüfgröße z ($A \times \text{Ref.}$ „AIAG MSA Standard“)	30
2.2.2.8 P-Wert	31
2.2.3 Vergleichbarkeit: Prüfer B vs. Referenz („AIAG MSA Standard“)	32

2.2.3.1	Virtuelle Einheiten-Gesamtanzahl (B × Ref. „AIAG MSA Standard“)	32
2.2.3.2	Aufbauschema Datentabelle (B × Ref. „AIAG MSA Standard“)	32
2.2.3.3	Kreuztabellen (B × Ref. „AIAG MSA Standard“)	33
2.2.3.4	Kappa-Koeffizient (B × Referenz „AIAG MSA Standard“)	35
2.2.3.5	Signifikanztest (B × Referenz „AIAG MSA Standard“)	35
2.2.3.6	Standardabweichung (B × Ref. „AIAG MSA Standard“)	36
2.2.3.7	Prüfgröße z (B × Ref. „AIAG MSA Standard“)	36
2.2.3.8	P-Wert (B × Ref. „AIAG MSA Standard“)	36
2.3	Kappa nach Cohen - Berechnungsart „AIAG MSA Extended“	37
2.3.1	Virtuelle Anzahl Einheiten für die Wiederholbarkeit innerhalb eines Prüfers	37
2.3.2	Wiederholbarkeit Prüfer A („AIAG MSA Extended“)	39
2.3.2.1	Virtuelle Einheiten-Gesamtanzahl (A „AIAG MSA Extended“)	39
2.3.2.2	Aufbauschem Datentabelle (A „AIAG MSA Extended“)	39
2.3.2.3	Kreuztabelle (A „AIAG MSA Extended“)	40
2.3.2.4	Kappa-Koeffizient (A „AIAG MSA Extended“)	41
2.3.2.5	Signifikanztest (A „AIAG MSA Extended“)	42
2.3.2.6	Standardabweichung (A „AIAG MSA Extended“)	42
2.3.2.7	Prüfgröße z (A „AIAG MSA Extended“)	42
2.3.2.8	P-Wert (A „AIAG MSA Extended“)	42
2.3.3	Wiederholbarkeit Prüfer B („AIAG MSA Extended“)	42
2.3.3.1	Virtuelle Einheiten-Gesamtzahl (B „AIAG MSA Extended“)	42
2.3.3.2	Aufbauschema Datentabelle (B „AIAG MSA Extended“)	43
2.3.3.3	Kreuztabellen (B „AIAG MSA Extended“)	44
2.3.3.4	Kappa-Koeffizient (B „AIAG MSA Extended“)	45
2.3.3.5	Signifikanztest Kappa-Koeffizient (B „AIAG MSA Extended“)	46
2.3.3.6	Standardabweichung (B „AIAG MSA Extended“)	46
2.3.3.7	Prüfgröße z (B „AIAG MSA Extended“)	46
2.3.3.8	P-Wert (B „AIAG MSA Extended“)	46
2.3.4	Vergleichbarkeit: alle Prüfer ohne Referenz („AIAG MSA Extended“)	46
2.3.4.1	Virtuelle Einheiten-Gesamtzahl (A × B „AIAG MSA Extended“)	46
2.3.4.2	Aufbau Datentabelle (A × B „AIAG MSA Extended“)	48
2.3.4.3	Kreuztabellen (A × B „AIAG MSA Extended“)	49
2.3.4.4	Kappa-Koeffizient (A × B „AIAG MSA Extended“)	50

2.3.4.5	Signifikanztest Kappa-Koeffizient ($A \times B$ „AIAG MSA Extended“)	51
2.3.4.6	Standardabweichung ($A \times B$ „AIAG MSA Extended“)	51
2.3.4.7	Prüfgröße z ($A \times B$ „AIAG MSA Extended“)	51
2.3.4.8	P-Wert ($A \times B$ „AIAG MSA Extended“)	51
2.3.5	Vergleichbarkeit Prüfer A vs. Referenz („AIAG MSA Extended“)	52
2.3.5.1	Virtuelle Einheiten-Gesamtzahl ($A \times \text{Ref.}$ „AIAG MSA Extended“)	52
2.3.5.2	Aufbau Datentabelle ($A \times \text{Ref.}$ „AIAG MSA Extended“)	53
2.3.5.3	Kreuztabellen ($A \times \text{Ref.}$ „AIAG MSA Extended“)	54
2.3.5.4	Kappa-Koeffizient ($A \times \text{Ref.}$ „AIAG MSA Extended“)	56
2.3.5.5	Signifikanztest ($A \times \text{Ref.}$ „AIAG MSA Extended“)	56
2.3.5.6	Standardabweichung ($A \times \text{Ref.}$ „AIAG MSA Extended“)	57
2.3.5.7	Prüfgröße z ($A \times \text{Ref.}$ „AIAG MSA Extended“)	57
2.3.5.8	P-Wert ($A \times \text{Ref.}$ „AIAG MSA Extended“)	57
2.3.6	Vergleichbarkeit Prüfer B vs. Referenz („AIAG MSA Extended“)	58
2.3.6.1	Virtuelle Anzahl der Einheiten ($B \times \text{Ref.}$ „AIAG MSA Extended“)	58
2.3.6.2	Aufbau Datentabelle ($B \times \text{Ref.}$ „AIAG MSA Extended“)	59
2.3.6.3	Kreuztabellen ($B \times \text{Ref.}$ „AIAG MSA Extended“)	60
2.3.6.4	Kappa-Koeffizient ($B \times \text{Ref.}$ „AIAG MSA Extended“)	62
2.3.6.5	Signifikanztest ($B \times \text{Ref.}$ „AIAG MSA Extended“)	62
2.3.6.6	Standardabweichung ($B \times \text{Ref.}$ „AIAG MSA Extended“)	62
2.3.6.7	Prüfgröße z ($B \times \text{Ref.}$ „AIAG MSA Extended“)	62
2.3.6.8	P-Wert ($B \times \text{Ref.}$ „AIAG MSA Extended“)	62
2.3.7	Vergleichbarkeit alle Prüfer vs. Referenz („AIAG MSA Extended“)	64
2.3.7.1	Virtuelle Einheiten-Gesamtzahl ($A \times B \times \text{Ref.}$ „AIAG MSA Extended“)	64
2.3.7.2	Aufbauschema Datentabelle ($A \times B \times \text{Ref.}$ „AIAG MSA Extended“)	65
2.3.7.3	Kreuztabellen ($A \times B \times \text{Ref.}$ „AIAG MSA Extended“)	67
2.3.7.4	Kappa-Koeffizient ($A \times B \times \text{Ref.}$ „AIAG MSA Extended“)	68
2.3.7.5	Signifikanztest Kappa-Koeffizient ($A \times B \times \text{Ref.}$ „AIAG MSA Extended“)	69
2.3.7.6	Standardabweichung ($A \times B \times \text{Ref.}$ „AIAG MSA Extended“)	69
2.3.7.7	Prüfgröße z ($A \times B \times \text{Ref.}$ „AIAG MSA Extended“)	69
2.3.7.8	P-Wert ($A \times B \times \text{Ref.}$ „AIAG MSA Extended“)	69
2.4	Kappa nach Cohen - Berechnungsart „Standard Calculation“	70
2.4.1	Wiederholbarkeit Prüfer A („Standard Calculation“)	70

2.4.1.1	Virtuelle Einheiten-Anzahl (A „Standard Calculation“)	70
2.4.1.2	Aufbau Datentabelle (A „Standard Calculation“)	71
2.4.1.3	Kreuztabellen (A „Standard Calculation“)	72
2.4.1.4	Kappa-Koeffizient (A „Standard Calculation“)	74
2.4.1.5	Signifikanztest (A „Standard Calculation“)	74
2.4.1.6	Standardabweichung (A „Standard Calculation“)	74
2.4.1.7	Prüfgröße z (A „Standard Calculation“)	74
2.4.1.8	P-Wert (A „Standard Calculation“)	74
2.4.2	Wiederholbarkeit Prüfer B („Standard Calculation“)	75
2.4.2.1	Aufbau Datentabelle (B „Standard Calculation“)	76
2.4.2.2	Kreuztabellen (B „Standard Calculation“)	77
2.4.2.3	Kappa-Koeffizient (B „Standard Calculation“)	79
2.4.2.4	Signifikanztest (B „Standard Calculation“)	79
2.4.2.5	Standardabweichung (B „Standard Calculation“)	79
2.4.2.6	Prüfgröße z (B „Standard Calculation“)	79
2.4.2.7	P-Wert (B „Standard Calculation“)	79
2.4.3	Vergleichbarkeit Prüfer A vs. Referenz („Standard Calculation“)	80
2.4.3.1	Virtuelle Einheiten-Gesamtzahl (A × Ref. „Standard Calculation“)	80
2.4.3.2	Aufbau Datentabelle (A × Ref. „Standard Calculation“)	81
2.4.3.3	Kreuztabellen (A × Ref. „Standard Calculation“)	81
2.4.3.4	Kappa-Koeffizient (A × Ref. „Standard Calculation“)	84
2.4.3.5	Signifikanztest Kappa-Koeffizient (A × Ref. „Standard Calculation“)	84
2.4.3.6	Standardabweichung (A × Ref. „Standard Calculation“)	84
2.4.3.7	Prüfgröße z (A × Ref. „Standard Calculation“)	84
2.4.3.8	P-Wert (A × Ref. „Standard Calculation“)	84
2.4.4	Vergleichbarkeit Prüfer B vs. Referenz („Standard Calculation“)	85
2.4.4.1	Virtuelle Einheiten-Gesamtzahl (B × Ref. „Standard Calculation“)	85
2.4.4.2	Aufbau Datentabelle (B × Ref. „Standard Calculation“)	86
2.4.4.3	Kreuztabellen (B × Ref. „Standard Calculation“)	87
2.4.4.4	Kappa-Koeffizient (B × Ref. „Standard Calculation“)	89
2.4.4.5	Signifikanztest (B × Ref. „Standard Calculation“)	89
2.4.4.6	Standardabweichung (B × Ref. „Standard Calculation“)	89
2.4.4.7	Prüfgröße z (B × Ref. „Standard Calculation“)	89

2.4.4.8	P-Wert ($B \times \text{Ref.} \mid \text{„Standard Calculation“}$)	89
2.4.5	Vergleichbarkeit aller Prüfer vs. Referenz („Standard Calculation“)	91
2.4.5.1	Virtuelle Einheiten-Gesamtzahl ($A \times B \times \text{Ref.} \mid \text{„Standard Calculation“}$)	91
2.4.5.2	Aufbau Datentabelle ($A \times B \times \text{Ref.} \mid \text{„Standard Calculation“}$).....	92
2.4.5.3	Kreuztabellen ($A \times B \times \text{Ref.} \mid \text{„Standard Calculation“}$)	94
2.4.5.4	Kappa-Koeffizient ($A \times B \times \text{Ref.} \mid \text{„Standard Calculation“}$)	95
2.4.5.5	Signifikanztest ($A \times B \times \text{Ref.} \mid \text{„Standard Calculation“}$)	96
2.4.5.6	Standardabweichung ($A \times B \times \text{Ref.} \mid \text{„Standard Calculation“}$).....	96
2.4.5.7	Prüfgröße z ($A \times B \times \text{Ref.} \mid \text{„Standard Calculation“}$)	96
2.4.5.8	P-Wert ($A \times B \times \text{Ref.} \mid \text{„Standard Calculation“}$)	96
2.5	Kappa nach Fleiss	96
2.5.1	Schritt-schema der Kappa-Bestimmung:	96
2.5.1.1	Aufbauschema Datentabelle	97
2.5.1.2	Anteil der beobachteten Übereinstimmungen.....	98
2.5.1.3	Erwarteter Anteil Übereinstimmungen	98
2.5.1.4	Vom Zufall bereinigte Anteile beobachteter Übereinstimmungen	98
2.5.1.5	Kappa-Koeffizient nach Fleiss	99
2.5.1.6	Standardabweichung	99
2.5.1.7	Prüfgröße z	100
2.5.1.8	P-Wert.....	100
2.5.2	Wiederholbarkeit Prüfer A	101
2.5.2.1	Aufbau Datentabelle (A Fleiss).....	101
2.5.2.2	Anteil beobachteter Übereinstimmungen (A Fleiss).....	101
2.5.2.3	Erwarteter Anteil Übereinstimmungen (A Fleiss)	102
2.5.2.4	Vom Zufall bereinigte Anteile (A Fleiss).....	103
2.5.2.5	Kappa-Koeffizient (A Fleiss)	103
2.5.2.6	Signifikanztest (A Fleiss)	103
2.5.2.7	Standardabweichung (A Fleiss).....	103
2.5.2.8	Prüfgröße z (A Fleiss)	104
2.5.2.9	P-Wert (A Fleiss)	104
2.5.3	Wiederholbarkeit Prüfer B	104
2.5.3.1	Aufbau Datentabelle (B Fleiss).....	105
2.5.3.2	Anteil beobachteter Übereinstimmungen (B Fleiss).....	106

2.5.3.3	Erwarteter Anteil zufälliger Urteilsübereinstimmungen (B Fleiss)	107
2.5.3.4	Kappa-Koeffizient (B Fleiss)	107
2.5.3.5	Signifikanztest (B Fleiss)	107
2.5.3.6	Standardabweichung (B Fleiss).....	107
2.5.3.7	Prüfgröße z (B Fleiss)	108
2.5.3.8	P-Wert (B Fleiss)	108
2.5.4	Vergleichbarkeit aller Prüfer (ohne Referenz)	108
2.5.4.1	Erwarteter Anteil zufälliger Übereinstimmungen (A x B Fleiss)	108
2.5.4.2	Anteil beobachteter Übereinstimmungen (A x B Fleiss)	109
2.5.4.3	Kappa-Koeffizient (A x B Fleiss).....	111
2.5.4.4	Signifikanztest (A x B Fleiss).....	111
2.5.4.5	Standardabweichung (A x B Fleiss)	111
2.5.4.6	Prüfgröße z (A x B Fleiss)	111
2.5.4.7	P-Wert (A x B Fleiss).....	111
2.5.5	Vergleichbarkeit Prüfer A vs. Referenz	113
2.5.5.1	Aufbau Datentabelle (A x Ref. Fleiss)	113
2.5.5.2	Anteil beobachteter Übereinstimmungen (A x Ref. Fleiss)	114
2.5.5.3	Erwarteter Anteil Übereinstimmungen (A x Ref. Fleiss)	117
2.5.5.4	Kappa-Koeffizient (A vs. Ref. Fleiss).....	118
2.5.5.5	Signifikanztest (A x Ref. Fleiss).....	119
2.5.5.6	Standardabweichung (A x Ref. Fleiss)	120
2.5.5.7	Prüfgröße z (A x Ref. Fleiss).....	121
2.5.5.8	P-Wert (A x Ref. Fleiss)	121
2.5.6	Vergleichbarkeit Prüfer B vs. Referenz	121
2.5.6.1	Aufbau Datentabelle (B vs. Ref. Fleiss).....	121
2.5.6.2	Anteil beobachteter Übereinstimmungen (B vs. Ref. Fleiss).....	123
2.5.6.3	Durch den Zufall erwarteter Anteil Übereinstimmungen (B vs. Ref. Fleiss)	125
2.5.6.4	Kappa-Koeffizient (B vs. Ref. Fleiss).....	125
2.5.6.5	Signifikanztest Kappa-Koeffizient (B x Ref. Fleiss).....	127
2.5.6.6	Standardabweichung (B x Ref. Fleiss)	127
2.5.6.7	Prüfgröße z (B x Ref. Fleiss).....	128
2.5.6.8	P-Wert (B x Ref. Fleiss)	128
2.5.7	Vergleichbarkeit alle Prüfer vs. Referenz	129

2.5.7.1	Aufbau Datentabelle (Vergleichbarkeit aller Prüfer vs. Referenz Fleiss)	129
2.5.7.2	Signifikanztest Kappa-Koeffizient ($A \times B \times \text{Ref.}$ Fleiss)	129
2.5.7.3	Standardabweichung ($A \times B \times \text{Ref.}$ Fleiss)	129
2.5.7.4	Prüfgröße z ($A \times B \times \text{Ref.}$ Fleiss)	129
2.5.7.5	P-Wert ($A \times B \times \text{Ref.}$ Fleiss)	129
3.	Prüfsystem-Effektivität	131
3.1	Zählung der Anzahl Objekte mit übereinstimmenden Urteilen	131
3.2	Prüfsystem-Effektivität ohne einen Referenz-Vergleich	132
3.2.1	Prüfsystem-Effektivität innerhalb der einzelnen Prüfer (ohne Referenz)	132
3.2.2	Prüfsystem-Effektivität über alle Prüfer hinweg (ohne Referenz)	133
3.3	Prüfsystem-Effektivität mit Referenz-Vergleich	135
3.3.1	Prüfsystem-Effektivität – Einzelne Prüfer vs. Referenz	135
3.3.1.1	Vergleich der Urteile – Prüfer A gegen Referenz	135
3.3.1.2	Vergleich der Urteile – Prüfer B gegen Referenz	135
3.3.2	Prüfsystem-Effektivität – Alle Prüfer vs. Referenz	136
4.	Anhang	137
4.1	Verwendete Symbole und deren Bedeutung	137
5.	Literatur	138

VORWORT

In diesem Dokument beschreiben wir die Methoden, welche im Q-DAS Produkt solara.MP für die Messsystemanalyse mit nominalen oder ordinalen Merkmalen zur Verfügung stehen.

Dies sind im Speziellen die folgenden Verfahren der Übereinstimmungsanalyse:

- Kappa-Koeffizient nach Cohen (drei verschiedene Berechnungsarten)
- Kappa-Koeffizient nach Fleiss
- Prüfsystem-Effektivität

Das Dokument enthält bewusst viele Wiederholungen. Dadurch sollen Sie als Leser den Inhalt für eine bestimmte Berechnungsart erfassen können, ohne dass Sie ständig durch Verweise auf vorhergehende Abschnitte im Lesefluss gestört werden. Die einzigen Ausnahmen von dieser Regel betreffen das erste Kapitel, in dem der allen Daten- und Berechnungsbeispielen zugrunde liegende Versuch beschrieben ist sowie die Einleitungen zu den jeweiligen Berechnungsarten, die Sie zumindest überfliegen sollten.

1. BEISPIELDATENSATZ FÜR DIESES FALLBEISPIEL

Für das Vorführen aller Berechnungsschritte beziehen wir uns in diesem Dokument stets auf den hier vorzustellenden Mini-Beispieldatensatz, der folgende Eigenschaften aufweist:

Tabelle 1: Eigenschaften des Versuchsplans zum Beispieldatensatz

Eigenschaft	Wert
Anzahl der Prüfer	$N_a = 2$
Anzahl der Prüfobjekte	$N_o = 5$
Anzahl der Durchgänge	$N_t = 2$
Anzahl der Urteilkategorien	$N_c = 2$

Fünf Prüfobjekte wurden von zwei Prüfern – A und B genannt – jeweils zweimal mit einer Lehre geprüft. Dieselben Prüfobjekte wurden zuvor im Feinmessraum vermessen und anhand der dort ermittelten Messergebnisse wurden die „Referenzurteile“ abgeleitet. Zusammenfassend sind alle vergebenen Urteile zu den fünf Objekten in der Tabelle 2 dargestellt:

Tabelle 2: Beispieldatensatz, der als Grundlage für das Vorführen der Berechnungsschritte im gesamten Dokument dient

Objekt-Nr. i	Referenzurteil	Prüfer A		Prüfer B	
		Durchgang 1	Durchgang 2	Durchgang 1	Durchgang 2
1	Okay	Okay	Okay	Not okay	Not okay
2	Not okay	Okay	Not okay	Okay	Not okay
3	Okay	Not okay	Okay	Okay	Okay
4	Not okay	Okay	Okay	Okay	Okay
5	Okay	Okay	Okay	Okay	Okay



Wichtig: Dieses Datenbeispiel mit nur fünf Prüfobjekten ist bewusst bezüglich der Datenmenge und der dargestellten Urteile extrem gehalten, damit jeder Rechenschritt mit einem Taschenrechner oder einer Tabellenkalkulation vom Leser leicht nachvollzogen werden kann.

Für die praktische Anwendung der in diesem Dokument beschriebenen Verfahren empfehlen wir dringend, **mindestens 50 Einheiten** zu prüfen. Es gilt die Regel: Je mehr Einheiten wir prüfen, desto besser ist die statistische Aussagekraft.

2. KAPPA-KOEFFIZIENTEN

Die beiden Kappa-Koeffizienten nach *Jacob Cohen* und *Joseph L. Fleiss* sind zwei der am weitesten verbreiteten **Kenngößen, mit denen das Ausmaß der Urteils-Übereinstimmung zwischen mindestens zwei urteilenden Instanzen gemessen wird**. Mit dem Begriff *Instanzen* bezeichnen wir hier nicht allein einzelne Personen, sondern verstehen darunter z.B. auch Gruppenentscheidungen unabhängiger Entscheidungsgremien sowie Urteile programmierter Algorithmen in einem automatisiert eingesetzten Prüfsystem.

2.1 Kappa nach Cohen

Dieser Koeffizient ist der historisch ältere der beiden in solara.MP angebotenen Kappa-Koeffizienten¹, mit denen der Grad der Übereinstimmung der Urteile mehrerer Prüfer (oder allgemein: urteilender Instanzen) über die gleichen N_o Objekte ausgedrückt wird. Der Koeffizient nach *Jakob Cohen* ist in der Anwendung beschränkt: Wir können nur die Urteile von **genau zwei Instanzen** miteinander vergleichen.

Tabelle 3: Verfügbare Vergleichsergebnisse innerhalb der Berechnungsmethoden für den Kappa-Koeffizienten nach Cohen

Betrachtete Übereinstimmung	Berechnungsmethode für den Kappa-Koeffizient nach <i>Jacob Cohen</i>		
	AIAG MSA Standard	AIAG MSA Extended	Standard Calculation
Wiederholbarkeit der <i>einzelnen</i> Prüfer ohne Referenz		×	× ⁽¹⁾
Vergleichbarkeit <i>aller</i> Prüfer ohne Referenz	×	×	
Vergleichbarkeit der <i>einzelnen</i> Prüfer mit der Referenz	×	×	×
Vergleichbarkeit <i>aller</i> Prüfer mit der Referenz		×	×

⁽¹⁾ Nur möglich und verfügbar, wenn exakt 2 Prüfdurchgänge ausgeführt wurden.

2.1.1 Aufbauschema Datentabelle

Der Kappa-Koeffizient nach Jacob Cohen wurde für den Vergleich von exakt zwei Urteilsspalten entwickelt. Aus diesem Grund gilt für alle Kappa-Auswertungen (nach Jacob Cohen!) ein **Aufbauschema** für die Urteilswerte, dass aus **einer Tabelle mit zwei Urteils-Spalten und N_o Zeilen** besteht. **Jede Spalte enthält die Urteile von einer urteilenden Instanz. Jede Zeile enthält die Urteile zu einem bestimmten Prüfobjekt**. Die Urteile in einer einzelnen Zeile sind immer eindeutig einem bestimmten Prüfobjekt zuordenbar. Ebenso sind die Urteile in einer einzelnen Spalte immer einer prüfenden Instanz – z.B. einem Prüfer – eindeutig zuordenbar.

¹ Es gibt noch weitere Kappa-Koeffizienten und auch weitere Maße der Übereinstimmungsanalyse. Eine gute Übersicht enthält die Monografie des Autors *Kilem L. Gwet*: „Handbook of Inter-Rater Reliability“, vierte Auflage, ISBN 978-0-970-80628-4

Tabelle 4: Allgemeiner Aufbau einer Datentabelle für die Auswertung mit dem Kappa-Koeffizienten nach Jacob Cohen

Objekt	Urteilende Instanz 1	Urteilende Instanz 2
1	Okay	Okay
2	Okay	Not okay
⋮	⋮	⋮
N_o	Not okay	Not okay

Wie wir gerade erwähnt haben, beruht das Berechnungsverfahren für den Kappa-Koeffizient nach Cohen auf einer Datentabelle mit genau zwei Urteilsspalten. Da das Verfahren für den Vergleich der Urteile von genau zwei Prüfern entwickelt wurde, ist in dem ursprünglichen Ansatz von Cohen ein mehrfach wiederholtes Beurteilen derselben Einheiten durch mehrere urteilende Instanzen genau so wenig vorgesehen wie der Vergleich von mehr als zwei urteilenden Instanzen [Coh]. In der industriellen Anwendung sind daher verschiedene heuristische Modifikationen für das Bestimmen des Kappa-Koeffizienten nach Cohen entstanden, bei denen die beiden im ursprünglichem Ansatz unberücksichtigt gebliebenen Bestandteile – mehr als zwei Prüfer und mehr als ein Durchgang pro Prüfer – pragmatisch integriert worden sind.

Wie wir in den Abschnitten zu den einzelnen Berechnungsverfahren noch sehen werden, haben diese heuristischen Modifikationen Auswirkungen auf die Datentabelle für die Auswertung: Darin entspricht die Gesamtzahl der dargestellten Einheiten (Zeilen) in der Regel nicht der tatsächlich im Versuch verwendeten Anzahl an Teilen N_o . Vielmehr werden die Ergebnisse aus mehreren Prüfurfgängen fallweise wie Prüfergebnisse an virtuellen neuen Einheiten behandelt. Dadurch ergibt sich für die Auswertung eine virtuelle Gesamtanzahl Einheiten N_{vo} , die oft deutlich größer ist als die tatsächlich vorhandene Anzahl Einheiten N_o . Das konkrete Vorgehen ist in den betreffenden Abschnitten beschrieben und daher werden wir das Thema hier zunächst nicht weiter betrachten.

Hinweis: Die Darstellung der Urteile innerhalb der Wertemaske des Programms solara.MP orientiert sich stets an dem ursprünglich verwendeten Versuchsaufbau. Daher entspricht die Anordnung der Urteile in der Wertemaske in der Regel nicht dem beschriebenen Aufbau der Datentabelle für die Auswertung.

2.1.2 Kreuztabellen-Berechnungsschema

Hier betrachten wir den - sicherlich auch am häufigsten angewandten - Fall einer Bewertung mit **zwei Urteilskategorien** „Okay“ und „Not okay“.

Zählen wir die Häufigkeit des Auftretens der vier möglichen **Urteils kombinationen in den Zeilen** einer Datentabelle (Aubauschema siehe Tabelle 4), so erhalten wir als Ergebnis eine **2 × 2-Kreuztabelle der beobachteten absoluten Häufigkeit der Urteils kombinationen**:

Tabelle 5: 2 × 2-Kreuztabelle für eine Bewertung mit zwei Urteilskategorien „Okay“ und „Not okay“

Beobachtete absolute Häufigkeiten		Zweite urteilende Instanz		Zeilensumme
		Okay	Not okay	
Erste urteilende Instanz	Okay	a	b	a + b
	Not okay	c	d	c + d
Spaltensumme		a + c	b + d	a + b + c + d

Teilen wir jeden Wert in der Tabelle 5 durch die Summe aller Urteils kombinationen $a + b + c + d$, so erhalten wir die **Tabelle der beobachteten Anteile der Urteils kombinationen**:

Tabelle 6: Berechnungsschema für eine 2 × 2-Kreuztabelle der beobachteten relativen Häufigkeiten aller Urteils kombinationen

Beobachtete relative Häufigkeiten		Zweite urteilende Instanz		Zeilensumme
		Okay	Not okay	
Erste urteilende Instanz	Okay	$\frac{a}{a + b + c + d}$	$\frac{b}{a + b + c + d}$	$\frac{a + b}{a + b + c + d}$
	Not okay	$\frac{c}{a + b + c + d}$	$\frac{d}{a + b + c + d}$	$\frac{c + d}{a + b + c + d}$
Spaltensumme		$\frac{a + c}{a + b + c + d}$	$\frac{b + d}{a + b + c + d}$	$\frac{a + b + c + d}{a + b + c + d}$

Bilden wir die Summe mit den Werten in den grün hinterlegten Zellen, so erhalten wir den **beobachteten Anteil gleicher Urteile** p_o :

$$p_o = \frac{a + d}{a + b + c + d}$$

Multiplizieren wir die Werte in den Randsummen - das sind die Spalten- und die Zeilensummen - der vorstehenden Tabelle 6 miteinander, so erhalten wir die **Tabelle der durch den Zufall erwarteten Anteile der Urteils kombinationen**.

Tabelle 7: Berechnungsschema für eine 2×2 -Kreuztabelle der durch den Zufall erwarteten relativen Häufigkeiten aller Urteils

Erwartete relative Häufigkeiten		Zweite urteilende Instanz		Zeilensumme
		Okay	Not okay	
Erste urteilende Instanz	Okay	$\frac{(a+b) \cdot (a+c)}{(a+b+c+d)^2}$	$\frac{(a+b) \cdot (b+d)}{(a+b+c+d)^2}$	$\frac{a+b}{a+b+c+d}$
	Not okay	$\frac{(c+d) \cdot (a+c)}{(a+b+c+d)^2}$	$\frac{(c+d) \cdot (b+d)}{(a+b+c+d)^2}$	$\frac{c+d}{a+b+c+d}$
Spaltensumme		$\frac{a+c}{a+b+c+d}$	$\frac{b+d}{a+b+c+d}$	$\frac{a+b+c+d}{a+b+c+d}$

Bilden wir die Summe mit den Werten in den beiden grün hinterlegten Tabellenzellen, so erhalten wir den **durch den Zufall erwarteten Anteil gleicher Urteile** p_e .

$$p_e = \frac{(a+b) \cdot (a+c) + (c+d) \cdot (b+d)}{(a+b+c+d)^2}$$

2.1.3 Berechnungsschema für den Kappa-Koeffizienten (Cohen)

Jacob Cohen ging bei seinen Überlegungen davon aus, dass in dem beobachteten Anteil übereinstimmender Urteile ein gewisser Anteil Übereinstimmungen enthalten ist, der allein durch den Zufall zustande gekommen ist. Aus diesem Grund hat er das folgende Berechnungsschema für das Bereinigen des zufälligen Anteils gleicher Urteile eingeführt:

Im ersten Bereinigungsverfahren wird aus dem beobachteten Anteil Übereinstimmungen der erwartete Anteil zufälliger Übereinstimmungen herausgerechnet:

$$p_o - p_e$$

Im zweiten Schritt bereinigen wir den größtmöglichen Anteil beobachteter Übereinstimmungen von dem erwarteten Anteil zufälliger Übereinstimmungen:

$$1 - p_e$$

Das Verhältnis aus diesen beiden bereinigten Anteilswerten ist der Kappa-Koeffizient:

$$\kappa_{\text{Cohen}} = \frac{p_o - p_e}{1 - p_e}$$

2.1.4 Berechnungsschema für die Varianz des Kappa-Koeffizienten

Würden wir den gleichen Versuchsaufbau der nominalen Messsystemanalyse sehr oft wiederholt ausführen, so würden die aus den Versuchsergebnissen berechneten Kappa-Koeffizienten von Versuch zu Versuch zufällig andere Werte annehmen. Eine näherungsweise Abschätzung der dabei auftretenden Zufallsstreuung des Kappa-Koeffizienten gibt die in [VMB] angegebene Varianzformel:

$$s_k^2 = \frac{1}{N_{vo}(1 - p_e)^2} \cdot \left\{ \sum_{i=1}^{N_c=2} p_{i.} \cdot p_{.i} \cdot [1 - (p_{i.} + p_{.i})]^2 + \sum_{i=1}^{N_c=2} \sum_{\substack{j=1 \\ j \neq i}}^{N_c=2} p_{i.} \cdot p_{.j} \cdot (p_{i.} + p_{.j})^2 - p_e^2 \right\}$$

Hierin werden die Spalten- und Zeilensummen aus der Tabelle 7 des vorhergehenden Abschnittes verwendet:

Spaltensummen (1 = Okay, 2 = Not okay)

$$p_{1.} = \frac{a + c}{N_{vo}}$$

$$p_{2.} = \frac{b + d}{N_{vo}}$$

Zeilensummen (1=Okay, 2 = Not okay):

$$p_{.1} = \frac{a + b}{N_{vo}}$$

$$p_{.2} = \frac{c + d}{N_{vo}}$$

Die Größe p_e ist die durch den Zufall erwartete Anteilssumme übereinstimmender Urteile:

$$p_e = \frac{(a + c) \cdot (a + b) + (c + d) \cdot (b + d)}{N_{vo}^2}$$

mit

N_{vo} = Virtuelle Gesamtzahl der Einheiten ($N_{vo} = a + b + c + d$), die sich aus den tatsächlich verwendeten Einheiten und virtuell hinzugefügten Einheiten zusammen setzt (wird gleich erklärt).

Erwähnenswert ist, dass je nach der verwendeten Berechnungsart die Anzahl der rechnerisch berücksichtigten Einheiten – die virtuelle Gesamtzahl der Einheiten N_{vo} – nicht identisch ist mit der tatsächlich im Versuch verwendeten Anzahl der Einheiten N_o . In solchen Fällen deuten wir z.B. die in einem zweiten (dritten, ...) Prüfdurchgang ermittelten Urteile so, als hätten wir diese durch das Bewerten „virtuell neuer Objekte“ erhalten.

2.1.5 Signifikanztest

Für den Signifikanztest stellen wir zunächst die Null- und Alternativhypothese auf:

Nullhypothese H_0 : Der Kappa-Koeffizient der Grundgesamtheit ist gleich Null.

Praktische Deutung dieser Nullhypothese: Übereinstimmende Urteile der Prüfer entstehen nur durch den Zufall.

Alternativhypothese H_1 : Der Kappa-Koeffizient der Grundgesamtheit ist ungleich Null.

Praktische Deutung der Alternativhypothese: Übereinstimmende Urteile der Prüfer entstehen nicht (allein) durch den Zufall, sondern sind systematischer Natur. Geprüft wird die Nullhypothese.

2.1.5.1 Prüfgröße z

Dazu bilden wir zunächst die Prüfgröße $z_{\kappa_{Cohen}}$ bestimmen, indem wir den Kappa-Koeffizienten κ_{Cohen} durch die Standardabweichung s_κ teilen:

$$z_{\kappa_{Cohen}} = \frac{\kappa_{Cohen}}{s_\kappa}$$

2.1.5.2 P-Wert

Mit der Prüfgröße $z_{\kappa_{Cohen}}$ berechnen wir den P-Wert:

$$P_{\kappa_{Cohen}} = 1 - G(z_{\kappa_{Cohen}})$$

$G(z_{\kappa_{Cohen}})$ = Verteilungsfunktion der Standardnormalverteilung

Die Entscheidung des Tests leiten wir durch den Vergleich des P-Wertes mit dem gewählten Signifikanzniveau α ab: Ist der P-Wert kleiner als das gewählte Signifikanzniveau α , so wird die Nullhypothese verworfen.

Wurde für das Signifikanzniveau α kein spezifischer Wert gewählt, können wir zur Orientierung ggf. das klassische Beurteilungsschema verwenden:

Tabelle 8: Klassische Methode der Testergebnis-Ermittlung durch den Vergleich des P-Wertes mit den drei vorgegebenen Signifikanzschwellen

P-Wert	Aussage des Tests	Anzahl Sterne
$P > 0,05$	Kein Verwerfen der Nullhypothese H_0 . Der Test ist nicht signifikant (Nullhypothese wird beibehalten).	Kein Stern
$0,05 \geq P > 0,01$	Verwerfen der Nullhypothese H_0 auf dem Signifikanzniveau kleiner oder gleich $\alpha = 5\%$. Das Testergebnis ist indifferent .	*
$0,01 \geq P > 0,001$	Verwerfen der Nullhypothese H_0 auf dem Signifikanzniveau kleiner oder gleich $\alpha = 1\%$. Das Testergebnis ist signifikant .	**
$0,001 \geq P$	Verwerfen der Nullhypothese H_0 auf dem Signifikanzniveau kleiner oder gleich $\alpha = 0,1\%$. Das Testergebnis ist hoch signifikant .	***

2.2 Kappa nach Cohen - Berechnungsart „AIAG MSA Standard“

Wie wir einleitend im Abschnitt 2.1 erwähnt haben, ist das Berechnungsverfahren für den Kappa-Index nach *Cohen* auf den Vergleich der Urteile von *exakt zwei Beurteilern* ausgelegt. Im Originalansatz ist weder das wiederholte Beurteilen derselben Einheiten vorgesehen, noch der Vergleich von mehr als zwei urteilenden Instanzen [Coh]. In der industriellen Anwendung sind aus diesem Grund verschiedene *heuristische Modifikationen* für das Bestimmen des Kappa-Koeffizienten entstanden, mit denen die zwei „Probleme“ – mehr als zwei Prüfer und mehr als ein Durchgang – pragmatisch berücksichtigt sind.

Bei der Berechnungsart „AIAG MSA Standard“ behandelt das Programm solara.MP die einzelnen Prüfdurchläufe so, als wären die bei einem zweiten (dritten,...) Prüfdurchgang bestimmten Urteile weitere Prüfergebnisse zu „virtuell neuen Einheiten“.

Wir erhalten mit der Berechnungsmethode „AIAG MSA Standard“ folgende Auswertungsergebnisse:

Tabelle 9: Verfügbare Auswertungsergebnisse bei der Berechnungsmethode „AIAG MSA Standard“

Ohne Referenzvergleich	Keine Wiederholbarkeit für jede der urteilenden Instanzen allein
	Vergleichbarkeit für alle zwei-elementigen Kombinationen der urteilenden Instanzen, wie z.B. AxB, AxC, BxC...
Mit Referenzvergleich	Vergleichbarkeit aller urteilenden Instanzen mit den Referenz-Urteilen, wie z.B. AxRef, BxRef, CxRef, ...
	Keine Vergleichbarkeit aller Instanzen gemeinsam mit der Referenz

2.2.1 Vergleichbarkeit aller Prüfer ohne Referenzvergleich („AIAG MSA Standard“)

Mit der Vergleichbarkeit wollen wir herausfinden, wie groß der Grad der Übereinstimmung zwischen den Urteilen der Prüfer A und B ist.

2.2.1.1 Virtuelle Einheiten-Gesamtanzahl ($A \times B$ | „AIAG MSA Standard“)

Wie schon im Abschnitt 2.1.1 beschrieben, kann der Kappa-Koeffizient nach *Jacob Cohen* nur mit Urteilen von genau zwei bewertenden Instanzen bestimmt werden. Auch bei dem Berechnungsverfahren „AIAG MSA Standard“ wenden wir Heuristiken für das Umwandeln der Versuchsdaten (siehe Tabelle 2 auf der Seite 10) in das Schema einer zweispaltigen Urteilstabelle an.

Allgemein gilt:

Wenn jeder Prüfer (=urteilende Instanz) jedes der N_o Einheiten insgesamt N_t mal beurteilt, so ergibt sich die *virtuelle Gesamtzahl der Einheiten* N_{vo} wie folgt:

$$N_{vo} = N_t \cdot N_o$$

N_t = Anzahl der Durchläufe je urteilender Instanz (= wie oft jeder Prüfer alle Einheiten geprüft hat)

N_o = Die Anzahl der im Versuch tatsächlich verwendeten Einheiten



Wichtig: Durch das Umdeuten des zweiten Durchgangs als „zusätzliche virtuelle Teile sechs bis zehn“ verwenden wir in den folgenden Berechnungen die „neue virtuelle Gesamtzahl der Prüfeinheiten“ $N_{vo} = 10$ statt der ursprünglich tatsächlich verwendeten fünf Prüfobjekte ($N_o = 5$)!

2.2.1.2 Aufbauschema Datentabelle ($A \times B$ | „AIAG MSA Standard“)

Aus der Tabelle 2 auf der Seite 10 entnehmen wir die Urteile der beiden Prüfer A und B und wandeln diese Daten wie folgt in das zweiseitige Tabellenlayout um.

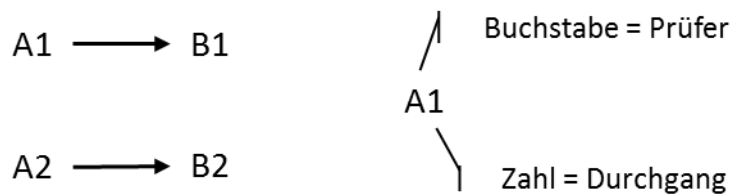


Abbildung 1: Datenaufbau-Schema

Durch diese Anordnung erhalten wir die virtuelle Gesamtzahl der Einheiten $N_{vo} = 2 \cdot 5 = 10$:

Tabelle 10: Beispiel-Daten – Vergleich der Urteile des Prüfers A mit den Urteilen des Prüfers B

Virtuelle Einheiten-Nr.	Nr. der Einheit im Versuch	Prüfer A	Prüfer B
		Alle Urteile	Alle Urteile
1	Einheit 1 Durchgang 1	Okay	Not okay
2	Einheit 2 Durchgang 1	Okay	Okay
3	Einheit 3 Durchgang 1	Not okay	Okay
4	Einheit 4 Durchgang 1	Okay	Okay
5	Einheit 5 Durchgang 1	Okay	Okay
6	Einheit 1 Durchgang 2	Okay	Not okay
7	Einheit 2 Durchgang 2	Not okay	Not okay
8	Einheit 3 Durchgang 2	Okay	Okay
9	Einheit 4 Durchgang 2	Okay	Okay
10	Einheit 5 Durchgang 2	Okay	Okay

Wie wir anhand der Tabelle 10 sehen, wurde der zweite Prüfdurchgang für die Einheiten eins bis fünf *umgedeutet* zu den Prüfergebnissen der „virtuell neuen Einheiten“ mit den Nummern sechs bis zehn.

2.2.1.3 Kreuztabellen (A × B | „AIAG MSA Standard“)

Auf der Grundlage dieser Interpretation ermitteln wir nun die Häufigkeiten der einzelnen Urteilskombinationen durch Zählen. D.h., wir zählen die Anzahl des Vorkommens der unterschiedlichen Urteilskombinationen in den einzelnen Zeilen der Tabelle 10 und übertragen das Ergebnis der Zählung in die **Kreuztabelle der absoluten Häufigkeit beobachteter Urteilskombinationen**:

Tabelle 11: Kreuztabelle der absoluten Häufigkeit der beobachteten Urteilskombinationen für die Vergleichbarkeit der Urteile der Prüfer A und B nach der Berechnungsmethode „AIAG

Vergleichbarkeit Prüfer A vs. B abs. Anzahl		Prüfer B		Summe
	Kategorie	Okay	Not okay	
Prüfer A	Okay	1	1	2
	Not okay	2	6	8
Summe		3	7	10

Teilen wir jeden Zahlenwert in der Tabelle 11 durch die virtuelle Gesamtzahl der Einheiten $N_{vo} = 10$, so erhalten wir die folgende **Kreuztabelle der Anteile beobachteter Urteilskombinationen**:

Tabelle 12: Kreuztabelle der Anteile beobachteter Urteilskombinationen
für die Vergleichbarkeit der Urteile der Prüfer A und B nach der Berechnungsmethode „AIAG MSA Standard“

Vergleichbarkeit Prüfer A vs. B beob. Anteile		Prüfer B		Summe
	Kategorie	Okay	Not okay	
Prüfer A	Okay	0,10	0,10	0,20
	Not okay	0,20	0,60	0,80
Summe		0,30	0,70	1,00

Mit den grün hinterlegten Zellenwerten der Tabelle 12 bilden wir die **Summe des beobachteten Anteils gleicher Urteile** p_o :

$$p_o = 0,1 + 0,6 = 0,7$$

Multiplizieren wird die Spaltensummen mit den Zeilensummen der Tabelle 12, erhalten wir die in der Tabelle 13 dargestellte **Kreuztabelle der Anteile erwarteter Urteils kombinationen**, die durch den Zufall entstanden sind.

Tabelle 13: Kreuztabelle der Anteile der durch den Zufall erwarteten Urteils kombinationen für die Vergleichbarkeit der Urteile der Prüfer A und B nach der Berechnungsmethode „AIAG MSA Standard“

Vergleichbarkeit Prüfer A vs. B erw. Anteile		Prüfer B		Summe
	Kategorie	Okay	Not okay	
Prüfer A	Okay	0,06	0,14	0,20
	Not okay	0,24	0,56	0,80
Summe		0,30	0,70	1,00

Mit den Werten in den grün hinterlegten Zellen der Tabelle 13 bilden wir die Summe und erhalten so den **erwarteten Anteil gleicher Urteile, der allein durch den Zufall entstanden ist**:

$$p_e = 0.06 + 0.56 = 0.62$$

2.2.1.4 Kappa-Koeffizient (A x B | „AIAG MSA Standard“)

Wir berechnen den **beobachteten Anteil gleicher Urteile, der vom zufälligen Anteil gleicher Urteile bereinigt ist**:

$$p_o - p_e = 0,7 - 0,62 = 0,08$$

Anschließend bestimmen wir den **(theoretisch) größtmöglichen Wert für den beobachteten Anteil gleicher Urteile, der vom zufällig erwarteten Anteil gleicher Urteile bereinigt ist**:

$$1 - p_e = 0,38$$

Teilen wir die beiden bereinigten Anteilswerte durcheinander, so erhalten wir als Ergebnis den **Kappa-Koeffizienten**:

$$\kappa_{A \times B \text{ (Cohen | AIAG MSA Standard)}} = \frac{p_o - p_e}{1 - p_e} = \frac{0,08}{0,38} \approx 0.211$$

Im nächsten Abschnitt betrachten wir den Signifikanztest für den gerade berechneten Kappa-Koeffizienten.

2.2.1.5 Signifikanztest (A x B | „AIAG MSA Standard“)

Für die Signifikanzprüfung wird ein statistischer Test durchgeführt. Dazu führen wir die folgenden Berechnungsschritte durch:

- Aufstellen der Null- und Alternativhypothese
- Wahl des Signifikanzniveaus
- Standardabweichung s für den Kappa-Koeffizienten bestimmen
- Prüfgröße z für den Kappa-Koeffizienten bestimmen
- P-Wert zum Signifikanztest ermitteln

Wir beginnen mit dem Aufstellen der Hypothesen:

Nullhypothese H_0 : Der Kappa-Koeffizient der Grundgesamtheit ist gleich Null

Die Aussage der Nullhypothese deuten wir praktisch: Alle übereinstimmenden Urteile sind durch den Zufall zustande gekommen.

Alternativhypothese H_1 : Der Kappa-Koeffizient der Grundgesamtheit ist ungleich Null.

Die Aussage der Alternativhypothese deuten wir praktisch: Die übereinstimmenden Urteile sind nicht (alle) zufallsbedingt.

2.2.1.6 Standardabweichung (A x B | „AIAG MSA Standard“)

Das allgemeine Schema für das Bestimmen der Standardabweichung für den Kappa-Koeffizienten nach Cohen ist im Abschnitt 2.1.4 beschrieben. Aufbauend auf dem dort angegebenen Rechenschema bestimmen wir nun den Istwert der Standardabweichung des Kappa-Koeffizienten nach Cohen:

Tabelle 14: Beispiel-Datensatz – 2x2-Kreuztabelle für die beobachtete absolute Häufigkeit der Urteilkombinationen der Prüfer A und B

Prüfer A	Prüfer B		Zeilensumme
	Okay	Not okay	
Okay	a = 1	b = 1	a + b = 2
Not okay	c = 2	d = 6	c + d = 8
Spaltensumme	a + c = 3	b + d = 7	a+b+c+d = 10

$$\begin{aligned}
 p_{1.} &= \frac{a+c}{N_{vo}} & p_{2.} &= \frac{b+d}{N_{vo}} & p_{.1} &= \frac{a+b}{N_{vo}} & p_{.2} &= \frac{c+d}{N_{vo}} \\
 p_{1.} &= \frac{3}{10} = 0,3 & p_{2.} &= \frac{7}{10} = 0,7 & p_{.1} &= \frac{2}{10} = 0,2 & p_{.2} &= \frac{8}{10} = 0,8
 \end{aligned}$$

$$p_e = \frac{(a+c) \cdot (a+b) + (c+d) \cdot (b+d)}{N_{vo}^2} = \frac{3 \cdot 2 + 8 \cdot 7}{10^2} = 0,62$$

$$s_k^2 = \frac{1}{N_{vo}(1-p_e)^2} \cdot \left\{ \sum_{i=1}^{N_c=2} p_{i.} \cdot p_{.i} \cdot [1 - (p_{i.} + p_{.i})]^2 + \sum_{i=1}^{N_c=2} \sum_{\substack{j=1 \\ j \neq i}}^{N_c=2} p_{i.} \cdot p_{.j} \cdot (p_{i.} + p_{.j})^2 - p_e^2 \right\}$$

$$s_{\kappa_{A \times B}(\text{Cohen} | \text{AIAG MSA Standard})}^2 = \frac{1}{10(1-0,62)^2} \cdot \left\{ \begin{aligned} &0,3 \cdot 0,2 \cdot [1 - (0,3 + 0,2)]^2 \\ &+ 0,7 \cdot 0,8 \cdot [1 - (0,7 + 0,8)]^2 \\ &+ 0,3 \cdot 0,8 \cdot (0,2 + 0,7)^2 \\ &+ 0,7 \cdot 0,2 \cdot (0,8 + 0,3)^2 \\ &- 0,62^2 \end{aligned} \right\} = 0,093074792$$

$$s_{\kappa_{A \times B}(\text{Cohen} | \text{AIAG MSA Standard})} = \sqrt{0,093074792} \approx 0,305081616$$

2.2.1.7 Prüfgröße z (A × B | „AIAG MSA Standard“)

Teilen wir den Wert des Kappa-Koeffizienten durch die Standardabweichung, so erhalten wir die Prüfgröße z_k :

$$z_{\kappa_{A \times B}(\text{Cohen} | \text{AIAG MSA Standard})} = \frac{\kappa_{A \times B}(\text{Cohen} | \text{AIAG MSA Standard})}{s_{\kappa_{A \times B}(\text{Cohen} | \text{AIAG MSA Standard})}} = \frac{0,210\,526\,316}{0,305\,081\,616} \approx 0,69$$

2.2.1.8 P-Wert (A × B | „AIAG MSA Standard“)

Auf der Grundlage der Prüfgröße z ermitteln wir mit der Verteilungsfunktion der Standardnormalverteilung den P-Wert:

$$P_{\kappa_{A \times B}(\text{Cohen} | \text{AIAG MSA Standard})} = 1 - G(z) = 1 - G(0,69) \approx 1 - 0,755 = 0,245$$

mit

$G(z)$ = Verteilungsfunktion der Standardnormalverteilung

Interpretation des Testergebnisses

Aufgrund des viel zu kleinen Stichprobenumfangs ist die Power des Tests (= Nullhypothese verwerfen, wenn diese nicht zutrifft) viel zu gering und daher das Testen an sich hier unsinnig. Dennoch vollziehen wir den formalen Akt des „Test-Rituals“ mit dem P-Wert.

Da der berechnete P-Wert größer ist als die Signifikanzschwelle $\alpha = 0,05$ (entspricht 5 %), wird die Nullhypothese nicht verworfen (=beibehalten). Damit gilt der aus der Stichprobe berechnete Kappa-Koeffizient als rein zufällig vom Wert $\kappa_{\text{Grundgesamtheit}} = 0$ abweichend.

2.2.2 Vergleichbarkeit: Prüfer A vs. Referenz („AIAG MSA Standard“)

Wir wollen feststellen, wie gut die Urteile des Prüfers A mit den Referenzbewertungen übereinstimmen. Es gilt, dass der Kappa-Koeffizient nach *Jacob Cohen* nur mit *exakt zwei Urteilsspalten* bestimmt werden kann. Daher müssen wir die Daten aus dem Versuch (Tabelle 2 auf der Seite 10) in das erforderliche zweispaltige Tabellenschema überführen.

2.2.2.1 Virtuelle Einheiten-Gesamtzahl ($A \times \text{Ref.}$ | „AIAG MSA Standard“)

Wir erzeugen das zweispaltige Tabellenlayout aus den Versuchsdaten wie folgt:

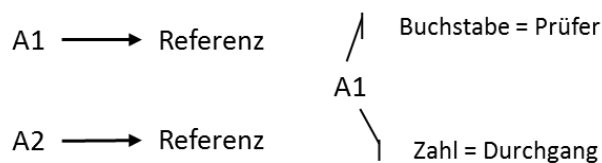


Abbildung 2: Datenaufbau-Scheme für den Vergleich Prüfer A vs. Referenz (Cohen | AIAG MSA Standard)

Anhand dieser Anordnung der Daten bestimmen wir die *virtuelle Gesamtanzahl der Einheiten* N_{vo} :

$$N_{vo} = N_t \cdot N_o = 2 \cdot 5 = 10$$

mit

N_t = Anzahl der Durchgänge

N_o = Anzahl der im Versuch verwendeten Einheiten

2.2.2.2 Aufbauschema Datentabelle (A × Ref. | „AIAG MSA Standard“)

Das Ergebnis unserer neuen Datenanordnung ist in der Tabelle 15 zu sehen:

Tabelle 15: Datentabelle zur Auswertung der Vergleichbarkeit des Prüfers A mit der Referenz nach der Berechnungsmethode „AIAG MSA Standard“

Virtuelle Einheiten-Nr.	Nr. der Einheit im Versuch	Prüfer A	Referenz
		Alle Urteile	Alle Urteile
1	Einheit 1 Durchgang 1	Okay	Okay
2	Einheit 2 Durchgang 1	Okay	Not okay
3	Einheit 3 Durchgang 1	Not okay	Okay
4	Einheit 4 Durchgang 1	Okay	Not okay
5	Einheit 5 Durchgang 1	Okay	Okay
6	Einheit 1 Durchgang 2	Okay	Okay
7	Einheit 2 Durchgang 2	Not okay	Not okay
8	Einheit 3 Durchgang 2	Okay	Okay
9	Einheit 4 Durchgang 2	Okay	Not okay
10	Einheit 5 Durchgang 2	Okay	Okay

2.2.2.3 Kreuztabellen (A x Ref. | „AIAG MSA Standard“)

Durch das Auszählen der Urteilskombinationen in den Zeilen der Tabelle 15 erhalten wir die **Kreuztabelle der absoluten Häufigkeit beobachteter Urteilskombinationen**:

Tabelle 16: Kreuztabelle der absoluten Häufigkeit beobachteter Urteilskombinationen für den Vergleich des Prüfers A mit der Referenz nach der Berechnungsmethode „AIAG MSA Standard“

Vergleichbarkeit Prüfer A vs. Ref abs. Anzahl		Referenz		Summe
	Kategorie	Okay	Not okay	
Prüfer A	Okay	5	3	8
	Not okay	1	1	2
Summe		6	4	10

Teilen wir jeden Zellenwert in der Tabelle 16 durch die virtuelle Gesamtanzahl der Einheiten $N_{vo} = 10$, erhalten wir die **Kreuztabelle der Anteile beobachteter Urteilskombinationen**:

Tabelle 17: Kreuztabelle der Anteile beobachteter Urteilskombinationen für den Vergleich des Prüfers A mit der Referenz nach der Berechnungsmethode „AIAG

Vergleichbarkeit Prüfer A vs. Ref beob. Anteil		Referenz		Summe
	Kategorie	Okay	Not okay	
Prüfer A	Okay	0,50	0,30	0,80
	Not okay	0,10	0,10	0,20
Summe		0,60	0,40	1,00

Bilden wir mit den beiden grün hinterlegten Zellenwerten in der Tabelle 17 die Summe, so erhalten wir den **Anteil beobachteter Übereinstimmungen p_o** :

$$p_o = 0,50 + 0,10 = 0,60$$

Tabelle 18: Kreuztabelle der durch den Zufall erwarteten Anteile der Urteils kombinationen für den Vergleich Prüfer A vs. Referenz nach der Berechnungsmethode „AIAG MSA Standard“

Vergleichbarkeit Prüfer A vs. Ref erw. Anteil		Referenz		Summe
	Kategorie	Okay	Not okay	
Prüfer A	Okay	0,48	0,32	0,80
	Not okay	0,12	0,08	0,20
Summe		0,60	0,40	1,00

Bilden wir mit den Werten in den grün hinterlegten Zellen der Tabelle 18 die Summe, so erhalten wir den **Anteil der durch Zufall erwarteten gleichen Urteils kombinationen p_e** :

$$p_e = 0,48 + 0,08 = 0,56$$

2.2.2.4 Kappa-Koeffizient (A x Referenz | „AIAG MSA Standard“)

Wir berechnen den **beobachteten Anteil gleicher Urteile, der vom zufälligen Anteil gleicher Urteile bereinigt ist**:

$$p_o - p_e = 0,60 - 0,56 = 0,04$$

Anschließend bestimmen wir den **(theoretisch) größtmöglichen Wert für den Anteil beobachteter gleicher Urteile, der vom zufällig erwarteten Anteil gleicher Urteile bereinigt ist**:

$$1 - p_e = 1 - 0,56 = 0,44$$

Teilen wir die beiden bereinigten Anteilswerte durcheinander, so erhalten wir als Ergebnis den **Kappa-Koeffizienten**:

$$K_{A \times Ref \text{ (Cohen | AIAG MSA Standard)}} = \frac{p_o - p_e}{1 - p_e} = \frac{0,04}{0,44} = 0,0909$$

2.2.2.5 Signifikanztest (A × Referenz | „AIAG MSA Standard“)

Der statistische Test durchläuft die folgenden Schritte:

- Aufstellen der Null- und Alternativhypothese
- Wahl des Signifikanzniveaus
- Standardabweichung s für den Kappa-Koeffizienten bestimmen
- Prüfgröße z für den Kappa-Koeffizienten bestimmen
- P-Wert zum Signifikanztest ermitteln

Wir beginnen mit dem Aufstellen der Hypothesen:

Nullhypothese H_0 : Der Kappa-Koeffizient der Grundgesamtheit ist gleich Null

Die Aussage der Nullhypothese deuten wir praktisch: Alle übereinstimmenden Urteile sind durch den Zufall zustande gekommen.

Alternativhypothese H_1 : Der Kappa-Koeffizient der Grundgesamtheit ist ungleich Null.

Die Aussage der Alternativhypothese deuten wir praktisch: Die übereinstimmenden Urteile sind nicht (alle) zufallsbedingt.

Hier wählen wir für das Signifikanzniveau $\alpha = 5\%$.

2.2.2.6 Standardabweichung (A × Ref. | „AIAG MSA Standard“)

Die Standardabweichung berechnen wir gemäß der Bestimmungsgleichung in dem Abschnitt 2.1.4 auf der Seite 15, wobei wir für die Berechnung die Zeilen- und Spaltensummen aus der Tabelle 17 verwenden:

$$s_{\kappa_{A \times \text{Ref}} (\text{Cohen} | \text{AIAG MSA Standard})} = 0,281672$$

2.2.2.7 Prüfgröße z (A × Ref. | „AIAG MSA Standard“)

Unsere Prüfgröße z erhalten wir, indem wir den Kappa-Wert durch die zuvor berechnete Standardabweichung teilen:

$$z_{\kappa_{A \times \text{Ref}} (\text{Cohen} | \text{AIAG MSA Standard})} = \frac{\kappa_{A \times \text{Ref}} (\text{Cohen} | \text{AIAG MSA Standard})}{s_{\kappa_{A \times \text{Ref}} (\text{Cohen} | \text{AIAG MSA Standard})}} = \frac{0,09090909}{0,281672} = 0,3227$$

2.2.2.8 P-Wert

Mit der Verteilungsfunktion der Standardnormalverteilung und mit der zuvor berechneten Prüfgröße bestimmen wir den P-Wert zum Test:

$$P_{\kappa_{A \times Ref}(Cohen | AIAG MSA Standard)} = 1 - G\left(z_{\kappa_{A \times Ref}(Cohen | AIAG MSA Standard)}\right) = 1 - G(0,3227) \approx 0,37$$

Interpretation des Testergebnisses

Aufgrund des viel zu kleinen Stichprobenumfangs ist die Power des Tests (= Nullhypothese verwerfen, wenn diese nicht zutrifft) viel zu gering und daher das Testen an sich hier unsinnig. Dennoch vollziehen wir den formalen Akt des „Test-Rituals“ mit dem P-Wert.

Der P-Wert ist größer als das gewählte Signifikanzniveau α . Aus diesem Grund wird die Nullhypothese nicht verworfen. Die beobachteten Übereinstimmungen sind vermutlich rein zufälliger Natur.

2.2.3 Vergleichbarkeit: Prüfer B vs. Referenz („AIAG MSA Standard“)

Wir fragen uns bei diesem Vergleich, wie gut die Urteile des Prüfers B mit den Referenzurteilen übereinstimmen.

2.2.3.1 Virtuelle Einheiten-Gesamtanzahl ($B \times \text{Ref.}$ | „AIAG MSA Standard“)

Die virtuelle Gesamtanzahl der Einheiten N_{vo} ergibt sich aus der Anzahl der Bewertungs-Durchgänge N_t und der Anzahl verwendeter Einheiten N_o im Versuch:

$$N_{vo} = N_t \cdot N_o = 2 \cdot 5 = 10$$

2.2.3.2 Aufbauschema Datentabelle ($B \times \text{Ref.}$ | „AIAG MSA Standard“)

Aus der Tabelle 2 auf der Seite 10 entnehmen wir die Urteile des Prüfers B und die Referenzurteile. Die Werte transformieren wir anschließend in die Form des zweispaltigen Tabellenlayout.

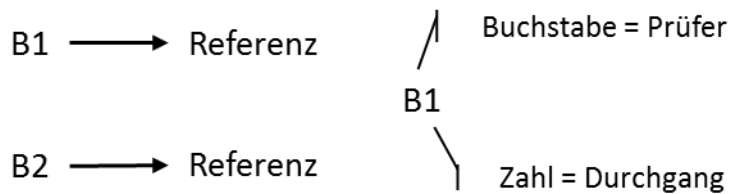


Abbildung 3: Datenaufbau-Schema für den Vergleich Prüfer B vs. Referenz (Cohen | AIAG MSA Standard)

Das Ergebnis der Transformation ist die folgende Tabelle mit der virtuellen Gesamtzahl = 10 Einheiten:

Tabelle 19: Datentabelle zur Auswertung der Vergleichbarkeit des Prüfers B mit der Referenz nach der Berechnungsmethode „AIAG MSA Standard“

Virtuelle Einheiten-Nr.	Nr. der Einheit im Versuch	Prüfer B	Referenz
		Alle Urteile	
1	Teil 1 Durchgang 1	Not okay	Okay
2	Teil 2 Durchgang 1	Okay	Not okay
3	Teil 3 Durchgang 1	Okay	Okay
4	Teil 4 Durchgang 1	Okay	Not okay
5	Teil 5 Durchgang 1	Okay	Okay
6	Teil 1 Durchgang 2	Not okay	Okay
7	Teil 2 Durchgang 2	Not okay	Not okay
8	Teil 3 Durchgang 2	Okay	Okay
9	Teil 4 Durchgang 2	Okay	Not okay
10	Teil 5 Durchgang 2	Okay	Okay

2.2.3.3 Kreuztabellen (B x Ref. | „AIAG MSA Standard“)

Wir bestimmen die Häufigkeit der Urteilskombinationen in den Zeilen der Tabelle 19 und erhalten so die folgende **Kreuztabelle der absoluten Häufigkeit der beobachteten Urteilskombinationen**:

Tabelle 20: Kreuztabelle der absoluten Häufigkeit beobachteter Urteilskombinationen für den Vergleich des Prüfers B mit der Referenz nach der Berechnungsart „AIAG MSA Standard“

Vergleichbarkeit Prüfer B vs. Ref abs. Anzahl		Referenz		Summe
	Kategorie	Okay	Not okay	
Prüfer A	Okay	4	3	7
	Not okay	2	1	3
Summe		6	4	10

Teilen wir die Zahlenwerte in der Tabelle 20 durch die virtuelle Gesamtzahl der Einheiten $N_{vo} = 10$, erhalten wir die **Kreuztabelle der Anteile beobachteter Urteilskombinationen**:Tabelle 21

Tabelle 21: Kreuztabelle der Anteile beobachteter Urteilskombinationen für den Vergleich des Prüfers B mit der Referenz nach der Berechnungsart „AIAG MSA Standard“

Vergleichbarkeit Prüfer B vs. Ref beob. Anteil		Referenz		Summe
	Kategorie	Okay	Not okay	
Prüfer A	Okay	0,40	0,30	0,70
	Not okay	0,20	0,10	0,30
Summe		0,60	0,40	1,00

Bilden wir die Summe mit den Werten in den grün hinterlegten Zellen der Tabelle 21, so erhalten wir den **Anteil beobachteter gleicher Urteilskombinationen** p_o :

$$p_o = 0,40 + 0,10 = 0,50$$

Multiplizieren wir die Zeilen- mit den Spaltensummen, so erhalten wir die **Kreuztabelle der durch den Zufall erwarteten Anteile der Urteils kombinationen**.

Tabelle 22: Kreuztabelle der Anteile der durch Zufall erwarteten Urteils kombinationen für den Vergleich des Prüfers B mit der Referenz nach der Berechnungsart „AIAG MSA Standard“

Vergleichbarkeit Prüfer B vs. Ref erw. Anteil		Referenz		Summe
	Kategorie	Okay	Not okay	
Prüfer A	Okay	0,42	0,28	0,70
	Not okay	0,18	0,12	0,30
Summe		0,60	0,40	1,00

Bilden wir mit den Werten in den grün hinterlegten Zellen in der Tabelle 22 die Summe, so erhalten wir den **erwarteten Anteil der durch Zufall gleichen Urteils kombinationen** p_e :

$$p_e = 0,42 + 0,12 = 0,54$$

2.2.3.4 Kappa-Koeffizient (B x Referenz | „AIAG MSA Standard“)

Wir berechnen den Anteil beobachteter, gleicher Urteile, der vom zufälligen Anteil gleicher Urteile bereinigt ist:

$$p_o - p_e = 0,50 - 0,54 = -0,04$$

Anschließend bestimmen wir den (theoretisch) größtmöglichen Wert für den Anteil beobachteter gleicher Urteile, der vom zufällig erwarteten Anteil gleicher Urteile bereinigt ist:

$$1 - p_e = 1 - 0,54 = 0,46$$

Teilen wir die beiden bereinigten Anteilswerte durcheinander, so erhalten wir als Ergebnis den Kappa-Koeffizienten:

$$\kappa_{B \times Ref} \text{ (Cohen | AIAG MSA Standard)} = \frac{p_o - p_e}{1 - p_e} = \frac{-0,04}{0,46} = -0,086\,956\,52$$

2.2.3.5 Signifikanztest (B x Referenz | „AIAG MSA Standard“)

Der statistische Test durchläuft die folgenden Schritte:

- Aufstellen der Null- und Alternativhypothese
- Wahl des Signifikanzniveaus
- Standardabweichung s für den Kappa-Koeffizienten bestimmen
- Prüfgröße z für den Kappa-Koeffizienten bestimmen
- P-Wert zum Signifikanztest ermitteln

Wir beginnen mit dem Aufstellen der Hypothesen:

Nullhypothese H_0 : Der Kappa-Koeffizient der Grundgesamtheit ist gleich Null

Die Aussage der Nullhypothese deuten wir praktisch: Alle übereinstimmenden Urteile sind durch den Zufall zustande gekommen.

Alternativhypothese H_1 : Der Kappa-Koeffizient der Grundgesamtheit ist ungleich Null.

Die Aussage der Alternativhypothese deuten wir praktisch: Die übereinstimmenden Urteile sind nicht (alle) zufallsbedingt.

Hier wählen wir für das Signifikanzniveau $\alpha = 5\%$.

2.2.3.6 Standardabweichung (B x Ref. | „AIAG MSA Standard“)

Die Standardabweichung berechnen wir gemäß der Bestimmungsgleichung in dem Abschnitt 2.1.4 auf der Seite 15, wobei wir für die Berechnung die Zeilen- und Spaltensummen aus der Tabelle 21 verwenden:

$$S_{\kappa_{B \times \text{Ref}} (\text{Cohen} | \text{AIAG MSA Standard})} \approx 0,308\,665$$

2.2.3.7 Prüfgröße z (B x Ref. | „AIAG MSA Standard“)

Unsere Prüfgröße z erhalten wir, indem wir den Kappa-Wert durch die zuvor berechnete Standardabweichung teilen:

$$Z_{\kappa_{B \times \text{Ref}} (\text{Cohen} | \text{AIAG MSA Standard})} = \frac{\kappa_{B \times \text{Ref}} (\text{Cohen} | \text{AIAG MSA Standard})}{S_{\kappa_{B \times \text{Ref}} (\text{Cohen} | \text{AIAG MSA Standard})}} = \frac{-0,086\,956\,52}{0,308\,665} \approx -0,282$$

2.2.3.8 P-Wert (B x Ref. | „AIAG MSA Standard“)

Mit der Verteilungsfunktion der Standardnormalverteilung und mit der zuvor berechneten Prüfgröße bestimmen wir den P-Wert zum Test:

$$P_{\kappa_{B \times \text{Ref}} (\text{Cohen} | \text{AIAG MSA Standard})} = 1 - G(-0,281\,716\,7) \approx 0,61$$

Interpretation des Testergebnisses

Aufgrund des viel zu kleinen Stichprobenumfangs ist die Power des Tests (= Nullhypothese verwerfen, wenn diese nicht zutrifft) viel zu gering und daher das Testen an sich hier unsinnig. Dennoch vollziehen wir den formalen Akt des „Test-Rituals“ mit dem P-Wert.

Der P-Wert ist größer als das gewählte Signifikanzniveau α . Aus diesem Grund wird die Nullhypothese nicht verworfen. Die beobachteten Übereinstimmungen sind vermutlich rein zufälliger Natur.

2.3 Kappa nach Cohen - Berechnungsart „AIAG MSA Extended“

Das Verfahren zur Berechnung des Kappa-Koeffizienten nach *Jacob Cohen* ist für den Vergleich der Urteile bei *exakt zwei urteilenden Instanzen* ausgelegt. Das originale Konzept berücksichtigt weder den Vergleich von mehr als zwei Prüfern, noch sind mehrere Beurteilungsdurchläufe vorgesehen. Aus diesem Grund sind in der industriellen Praxis heuristische Modifikationen eingeführt worden, um mit mehreren Durchläufen und urteilenden Instanzen pragmatisch umgehen zu können. Anders ausgedrückt: Es wurden verschiedene Ansätze entwickelt, wie die Versuchsergebnisse in ein Tabellenschema mit *genau zwei zu vergleichenden Urteilsspalten* überführt werden.

Auch die Berechnungsart „AIAG MSA Extended“ ist ein derartig heuristisch angepasstes Konzept, bei dem pragmatisch mehrere Beurteilungs-Durchgänge durch das Einführen „virtueller Einheiten“ gelöst wurde. Haben im konkreten Anwendungsfall beispielsweise die Prüfer mehrere Prüfdurchgänge ausgeführt, so werden die im zweiten (dritten, ...) Durchgang gefällten Urteile so behandelt, als wären diese an virtuell neuen Einheiten entstanden.

In unserem Daten-Beispiel (siehe Kapitel 1 auf der Seite 10) hat jeder der Prüfer zwei Prüfdurchläufe ausgeführt. Somit ist das Berechnen der Wiederholbarkeit innerhalb eines Prüfers (ausnahmsweise) ohne das Umdeuten des zweiten Prüfdurchgangs als neue „virtuelle Einheiten“ durchführbar, da in diesem Fall genau eine Tabelle mit exakt zwei Urteilsspalten vorhanden ist. Die Sachlage ändert sich, wenn jeder der beiden Prüfer insgesamt drei oder mehr Prüfdurchgänge ausgeführt hat.

2.3.1 Virtuelle Anzahl Einheiten für die Wiederholbarkeit innerhalb eines Prüfers

Damit wir auch in den Fällen mit mehr als zwei Durchläufen das Schema einer Zweispalten-Datentabelle für die Auswertung erhalten, müssen wir zunächst herausfinden, wie viele zwei-elementige Paarkombinationen wir mit der ausgeführten Anzahl Prüfdurchgänge (N_t) bilden können:

$$N_{vo} = \binom{N_t}{2} \cdot N_o = \frac{N_t!}{(N_t - 2)! \cdot 2!} \cdot N_o = \frac{N_t \cdot (N_t - 1)}{2} \cdot N_o = \frac{N_t^2 - N_t}{2} \cdot N_o$$

N_t = Anzahl der Durchläufe, also wie oft jeder einzelne Prüfer jede der Einheiten geprüft hat

N_o = Anzahl der im Versuch verwendeten Einheiten

Um ein besseres Gefühl für die Anzahl der entstehenden Paarkombinationen zu bekommen, sind in der Tabelle 23 für einige vorgegebene Anzahlen an Prüfdurchgängen N_t die entstehenden Anzahlen an Paarkombinationen gelistet. Die Einträge in dieser Tabelle sind wie folgt zu deuten:

Wir wollen eine Tabelle mit genau zwei Urteilsspalten erhalten. Die Zahlenwerte links vom Symbol „×“ stehen für die Urteilswerte der Prüfdurchgänge, die wir in die linke Spalte unserer (Zwei-Urteilsspalten-) Datentabelle zu schreiben haben und die Zahlenwerte rechts vom Symbol „×“ stehen für die Urteile der Prüfdurchgänge, die wir in die rechte Spalte unserer (Zwei-Urteilsspalten-) Datentabelle einzutragen haben.

Tabelle 23: Ansicht der zu kombinierenden Urteile in Abhängigkeit von der Anzahl der Prüfdurchläufe N_t für die Berechnungsart „AIAG MSA Extended - innerhalb der Prüfer“

N_t	Anzahl der Kombination	Nummern der Prüfdurchgänge, die wir zu Paaren zusammen fassen (Urteilsübereinstimmung innerhalb eines Prüfers)
2	1	1×2
3	3	$1 \times 2 \mid 1 \times 3 \mid 2 \times 3$
4	6	$1 \times 2 \mid 1 \times 3 \mid 1 \times 4 \mid 2 \times 3 \mid 2 \times 4 \mid 3 \times 4$
5	10	$1 \times 2 \mid 1 \times 3 \mid 1 \times 4 \mid 1 \times 5 \mid 2 \times 3 \mid 2 \times 4 \mid 2 \times 5 \mid 3 \times 4 \mid 3 \times 5 \mid 4 \times 5$

Hinweis: Das Zeichen × ist hier nicht als Multiplikation zu lesen, sondern als „vergleichen mit“.

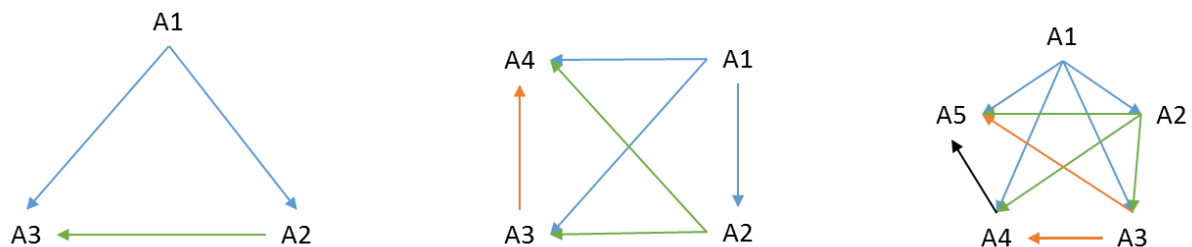


Abbildung 4: Datenaufbau-Schema für die Wiederholbarkeit bei drei, vier oder fünf Prüfdurchgängen

Auf diese Weise erhalten wir die folgende Gesamtzahl virtueller Prüfeinheiten:

$$N_{vo} = \binom{N_t}{2} \cdot N_o$$

N_t = Anzahl der Durchläufe, also wie oft der Prüfer alle Einheiten geprüft hat

N_o = Anzahl der im Versuch verwendeten Einheiten

2.3.2 Wiederholbarkeit Prüfer A („AIAG MSA Extended“)

Für die Wiederholbarkeit betrachten wir allein die Prüfurteile des Prüfers A. Wir wollen wissen, wie gut die Urteile des Prüfers A aus dem ersten Durchgang mit seinen Urteilen aus dem zweiten Durchgang übereinstimmen.

2.3.2.1 Virtuelle Einheiten-Gesamtanzahl (A | „AIAG MSA Extended“)

Wir bestimmen die virtuelle Gesamtanzahl der Einheiten gemäß der Beschreibung im vorhergehenden Abschnitt wie folgt:

$$N_{vo} = \binom{N_t}{2} \cdot N_o = \binom{2}{2} \cdot 5 = 5$$

2.3.2.2 Aufbauschem Datentabelle (A | „AIAG MSA Extended“)

Wir entnehmen aus der Tabelle 2 auf der Seite 10 die Urteile des Prüfers A und transformieren die Daten in das Schema des zweispaltigen Tabellenlayouts.

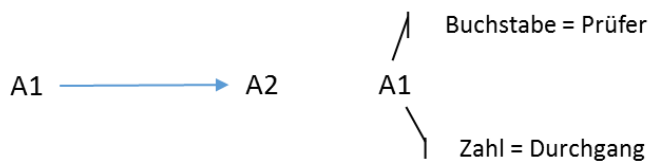


Abbildung 5: Datenaufbau-Schema für die Wiederholbarkeit Prüfer A

In diesem Fall ergibt sich das zweispaltige Schema recht natürlich, indem wir in die erste Spalte die Urteile des ersten Prüfdurchgangs schreiben und in die zweite Spalte die Urteile des zweiten Prüfdurchgangs.

Tabelle 24: Beispiel-Datensatz – Urteile des Prüfers A

Objekt-Nr.	Prüfer A	
	Durchgang 1	Durchgang 2
1	Okay	Okay
2	Okay	Not okay
3	Not okay	Okay
4	Okay	Okay
5	Okay	Okay

2.3.2.3 Kreuztabelle (A | „AIAG MSA Extended“)

Zählen wir die Häufigkeit der in den Zeilen der Tabelle 24 befindlichen Urteils-Kombinationen, so erhalten wir die folgende Kreuztabelle:

Tabelle 25: Beispiel-Datensatz – für den **Prüfer A beobachtete absolute Häufigkeit der Urteils kombinationen** durch Vergleich der Urteile aus dem ersten Durchgang mit den Urteilen aus dem zweiten Durchgang

Wiederholbarkeit Prüfer A abs. Anzahl		Prüfer A 2. Durchgang		Summe
	Kategorie	Okay	Not okay	
Prüfer A 1. Durchgang	Okay	3	1	4
	Not okay	1	0	1
Summe		4	1	5

Dividieren wir jeden Zellenwert in der Tabelle 25 durch die virtuelle Gesamtzahl der Einheiten $N_{VO} = 5$, erhalten wir die folgende Tabelle der beobachteten Anteile der Urteils kombinationen:

Wiederholbarkeit Prüfer A beob. Anteile		Prüfer A 2. Durchgang		Summe
	Kategorie	Okay	Not okay	
Prüfer A 1. Durchgang	Okay	0,6	0,2	0,8
	Not okay	0,2	0,0	0,2
Summe		0,8	0,2	1,0

Aus der Tabelle 26 entnehmen wir die Summe der beobachteten Anteile übereinstimmender Urteile. Dazu bilden wir die Summe mit den Werten aus den beiden grün hinterlegten Zellen. Diese Summe ist also der *beobachtete Anteil gleicher Urteile*:

$$p_o = 0,6 + 0,0 = 0,6$$

Multiplizieren wir die Spaltensummen mit den Zeilensummen der Tabelle 26, erhalten wir die Tabelle der *durch den Zufall erwarteten Anteile* der Urteils kombinationen:

Tabelle 27: Beispiel-Datensatz – erwartete relative Häufigkeit der Urteils kombinationen für die Berechnungsart „AIAG MSA Extended – innerhalb des Prüfers A“

Wiederholbarkeit Prüfer A erw. Anteile		Prüfer A 2. Durchgang		Summe
	Kategorie	Okay	Not okay	
Prüfer A 1. Durchgang	Okay	0,64	0,16	0,8
	Not okay	0,16	0,04	0,2
Summe		0,8	0,2	1,0

Bilden wir die Summe mit den beiden grün hinterlegten Werten in dieser Tabelle, so erhalten wir den **erwarteten Anteil gleicher Urteile, der durch den Zufall entstanden ist:**

$$p_e = 0,64 + 0,04 = 0,68$$

2.3.2.4 Kappa-Koeffizient (A | „AIAG MSA Extended“)

Wir bereinigen zunächst den beobachteten Anteil übereinstimmender Entscheidungen von dem Anteil der durch Zufall übereinstimmenden Urteile:

$$p_o - p_e = 0,60 - 0,68 = -0,08$$

Den größtmögliche Wert für den beobachteten Anteil übereinstimmender Entscheidungen, der nicht durch den Zufall zustande kommt, ermitteln wir wie folgt:

$$1 - p_e = 1 - 0,68 = 0,32$$

Teilen wir die beiden zuletzt berechneten Anteilswerte durcheinander, erhalten wir den **Kappa-Koeffizienten nach Cohen für die Wiederholbarkeit des Prüfers A:**

$$\kappa_{A(\text{Cohen} | \text{AIAG MSA Extended})} = \frac{p_o - p_e}{1 - p_e} = \frac{-0,08}{0,32} = -0,25$$

2.3.2.5 Signifikanztest (A | „AIAG MSA Extended“)

Die zu prüfende Hypothese lautet:

Nullhypothese H_0 : Der Kappa-Koeffizient der Grundgesamtheit ist gleich Null

Praktische Deutung der Nullhypothese: Die übereinstimmenden Urteile sind vermutlich allein durch den Zufall entstanden.

Alternativhypothese H_1 : Der Kappa-Koeffizient der Grundgesamtheit ist ungleich Null.

Praktische Deutung der Alternativhypothese: Die übereinstimmenden Urteile sind nicht (allein) durch den Zufall zustande gekommen. Der Urteile folgen einer erkennbaren Systematik.

Für die Prüfung der Nullhypothese benötigen wir die Prüfgröße z , die wir aus der Division des Kappa-Koeffizienten durch die Standardabweichung des Kappa-Koeffizienten erhalten.

2.3.2.6 Standardabweichung (A | „AIAG MSA Extended“)

Für den Signifikanztest berechnen wir zunächst die **Standardabweichung des Kappa-Koeffizienten**: gemäß der Bestimmungsgleichung im Abschnitt 2.1.4.

$$s_{\kappa_{A(\text{Cohen} | \text{AIAG MSA Extended})}} = 0,447\,214$$

2.3.2.7 Prüfgröße z (A | „AIAG MSA Extended“)

Teilen wir den Kappa-Koeffizienten durch die Standardabweichung, erhalten wir die **Prüfgröße für den Signifikanztest**:

$$z_{\kappa_{A(\text{Cohen} | \text{AIAG MSA Extended})}} = \frac{\kappa_{A(\text{Cohen} | \text{AIAG MSA Extended})}}{s_{\kappa_{A(\text{Cohen} | \text{AIAG MSA Extended})}}} = \frac{-0,25}{0,447\,214} \approx -0,559$$

2.3.2.8 P-Wert (A | „AIAG MSA Extended“)

Für diese Prüfgröße bestimmen wir mit der Verteilungsfunktion der Standardnormalverteilung den **P-Wert**:

$$P_{\kappa_{A(\text{Cohen} | \text{AIAG MSA Extended})}} = 1 - G(-0,559) \approx 0,72$$

Interpretation des Testergebnisses anhand des P-Wertes

Aufgrund des viel zu kleinen Stichprobenumfangs ist die Power des Tests (= Nullhypothese verwerfen, wenn diese nicht zutrifft) viel zu gering und daher das Testen an sich hier unsinnig. Dennoch vollziehen wir den formalen Akt des „Test-Rituals“ mit dem P-Wert.

Wir vergleichen den P-Wert mit dem Signifikanzniveau $\alpha = 0,05$. Der berechnete P-Wert ist größer als das gewählte Signifikanzniveau. Aus diesem Grund wird die **Nullhypothese H_0 nicht verworfen**: „Der Kappa-Koeffizient der Grundgesamtheit ist gleich Null“. Das heißt, die beobachteten Übereinstimmung sind rein zufälliger Natur.

2.3.3 Wiederholbarkeit Prüfer B („AIAG MSA Extended“)

Mit der Wiederholbarkeit wollen wir herausfinden, wie gut die Urteile des Prüfers B aus dem ersten Durchgang mit den Urteilen aus dem zweiten Durchgang übereinstimmen.

2.3.3.1 Virtuelle Einheiten-Gesamtzahl (B | „AIAG MSA Extended“)

Gemäß der im Abschnitt 2.3.1 dargestellten Formel bestimmen wir die virtuelle Gesamtzahl der Einheiten wie folgt:

$$N_{vo} = \binom{N_t}{2} \cdot N_o = \binom{2}{2} \cdot 5 = 5$$

2.3.3.2 Aufbauschema Datentabelle (B | „AIAG MSA Extended“)

Wir entnehmen aus der Tabelle 2 auf der Seite 10 die Urteile für den Prüfer B.

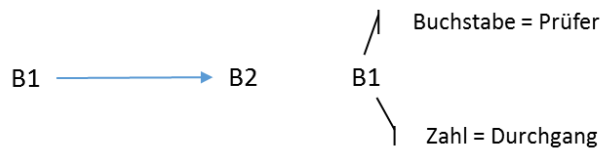


Abbildung 6: Datenaufbau-Schema für die Wiederholbarkeit Prüfer B

In diesem Fall erhalten wir das Schema der zweispaltigen Analysetabelle ohne Aufwand, da das Übertragen der Urteile aus den beiden Prüfdurchgängen ganz natürlich das zweispaltige Layout ergibt:

Tabelle 28: Beispiel-Datensatz – Urteile des Prüfers B

Einheiten Nr.	Prüfer B	
	Durchgang 1	Durchgang 2
1	Not okay	Not okay
2	Okay	Not okay
3	Okay	Okay
4	Okay	Okay
5	Okay	Okay

2.3.3.3 Kreuztabellen (B | „AIAG MSA Extended“)

Wir zählen die Häufigkeit der Urteilskombinationen in den Zeilen der Tabelle 28 und erhalten so die folgende **Kreuztabelle der absoluten Häufigkeit der beobachteten Urteilskombinationen**:

Tabelle 29: Kreuztabelle der absoluten Häufigkeit der beobachteten Urteilskombinationen für die Wiederholbarkeit des Prüfers B nach der Berechnungsmethode „AIAG MSA Extended“

Wiederholbarkeit Prüfer B abs. Anzahl		Prüfer B 2. Durchgang		Summe
	Kategorie	Okay	Not okay	
Prüfer B 1. Durchgang	Okay	3	1	4
	Not okay	0	1	1
Summe		3	2	5

Teilen wir jeden Zahlenwert in der Tabelle 29 durch die virtuelle Gesamtanzahl $N_{vo} = 5$, so erhalten wir die **Tabelle der beobachteten Anteile der Urteilskombinationen**:

Tabelle 30: Beispiel-Datensatz – für den Prüfer B **beobachtete relative Häufigkeit** der Urteilskombinationen aus dem Vergleich des ersten mit dem zweiten Prüfdurchgang

Wiederholbarkeit Prüfer B beob. Anteil		Prüfer B 2. Durchgang		Summe
	Kategorie	Okay	Not okay	
Prüfer B 1. Durchgang	Okay	0,6	0,2	0,8
	Not okay	0,0	0,2	0,2
Summe		0,6	0,4	1,0

Bilden wir die Summe mit den Werten in den grün hinterlegten Feldern, so erhalten wir den beobachteten Anteil gleicher Urteile p_o .

$$p_o = 0,6 + 0,2 = 0,8$$

Multiplizieren wir die Zeilensummen mit den Spaltensummen der Tabelle 30, so ergibt sich die **Tabelle der erwarteten Anteile der Urteilskombinationen, die durch den Zufall entstanden sind**:

Tabelle 31: Beispiel-Datensatz – für den Prüfer B **erwarteter Anteil** der Urteilskombinationen aus dem Vergleich des ersten mit dem zweiten Prüfdurchgang

Wiederholbarkeit Prüfer B erw. Anteil		Prüfer B 2. Durchgang		Summe
	Kategorie	Okay	Not okay	
Prüfer B 1. Durchgang	Okay	0,48	0,32	0,8
	Not okay	0,12	0,08	0,2
Summe		0,6	0,4	1,0

Bilden wir die Summe mit den Werten in den grün hinterlegten Feldern, so erhalten wir **den erwarteten Anteil gleicher Urteile p_e , der durch den Zufall entstanden ist.**

$$p_e = 0,48 + 0,08 = 0,56$$

2.3.3.4 Kappa-Koeffizient (B | „AIAG MSA Extended“)

Mit p_o und p_e bestimmen wir den **Anteil beobachteter übereinstimmender Entscheidungen, der vom zufällig entstandenen Anteil gleicher Urteile bereinigt ist:**

$$p_o - p_e = 0,8 - 0,56 = 0,24$$

Den **theoretisch größtmöglichen Wert für den Anteil übereinstimmender Urteile, der nicht durch den Zufall entstanden ist**, berechnen wir wie folgt:

$$1 - p_e = 1 - 0,56 = 0,44$$

Mit den beiden zuletzt berechneten Anteilswerten bestimmen wir schließlich den **Kappa-Koeffizienten für den Prüfer B:**

$$\kappa_{B(\text{Cohen} \mid \text{AIAG MSA Extended})} = \frac{p_o - p_e}{1 - p_e} = \frac{0,24}{0,44} \approx 0,545\,454$$

2.3.3.5 Signifikanztest Kappa-Koeffizient (B | „AIAG MSA Extended“)

Zu prüfende Hypothese:

Nullhypothese H_0 : Der Kappa-Koeffizient der Grundgesamtheit ist gleich Null

Praktische Deutung der Nullhypothese: Die übereinstimmenden Urteile sind vermutlich allein durch den Zufall entstanden.

Alternativhypothese H_1 : Der Kappa-Koeffizient der Grundgesamtheit ist ungleich Null.

Praktische Deutung der Alternativhypothese: Die übereinstimmenden Urteile sind nicht (allein) durch den Zufall zustande gekommen. Der Urteile folgen einer erkennbaren Systematik.

Für die Prüfung der Nullhypothese benötigen wir die Prüfgröße z , die wir aus der Division des Kappa-Koeffizienten durch die Standardabweichung des Kappa-Koeffizienten erhalten.

2.3.3.6 Standardabweichung (B | „AIAG MSA Extended“)

Würden wir den Versuch mehrfach wiederholt ausführen, so würde der Wert des Kappa-Indexes von Versuch zu Versuch streuen. Eine näherungsweise Abschätzung der Zufallsstreuung erhalten wir durch das Bestimmen der Standardabweichung des Kappa-Koeffizienten.

Den Wert der **Standardabweichung des Kappa-Koeffizienten** berechnen wir mit der Berechnungsgleichung aus dem Abschnitt 2.1.4.

$$s_{\kappa(\text{Cohen} | \text{AIAG MSA Extended})} = 0,398\ 344$$

2.3.3.7 Prüfgröße z (B | „AIAG MSA Extended“)

Teilen wir den Kappa-Index durch die Standardabweichung, so erhalten wir die **Prüfgröße z_{κ}** .

$$z_{\kappa B(\text{Cohen} | \text{AIAG MSA Extended})} = \frac{0,545\ 454}{0,398\ 344} \approx 1,369$$

2.3.3.8 P-Wert (B | „AIAG MSA Extended“)

Den Wert der Prüfgröße $z_{\kappa(\text{Cohen} | \text{AIAG MSA Extended})}$ setzen wir in die Verteilungsfunktion der Standardnormalverteilung und erhalten gemäß der folgenden Beziehung den **P-Wert**:

$$P_{B(\text{Cohen} | \text{AIAG MSA Extended})} = 1 - G(z_{\kappa(\text{Cohen} | \text{AIAG MSA Extended})}) = 1 - G(1,369) \approx 0,086$$

Interpretation des P-Wertes

Aufgrund des viel zu kleinen Stichprobenumfangs ist die Power des Tests (= Nullhypothese verwerfen, wenn diese nicht zutrifft) viel zu gering und daher das Testen an sich hier unsinnig. Dennoch vollziehen wir den formalen Akt des „Test-Rituals“ mit dem P-Wert.

Wir vergleichen den P-Wert mit dem Signifikanzniveau $\alpha = 0,05$. Der berechnete P-Wert ist größer als das gewählte Signifikanzniveau. Aus diesem Grund **verwerfen wir die Nullhypothese H_0 nicht**: „Der Kappa-Koeffizient der Grundgesamtheit ist gleich Null“.

2.3.4 Vergleichbarkeit: alle Prüfer ohne Referenz („AIAG MSA Extended“)

Wir wollen wissen, wie gut die Urteile der einzelnen Bewertungsinstanzen miteinander übereinstimmen. Aber, mit dem Kappa-Koeffizienten nach *Jacob Cohen* **können wir nicht direkt die Urteile von mehr als zwei Bewertungsinstanzen miteinander vergleichen**. Dieser Kappa-Index wurde von seinem Schöpfer für den Vergleich der Urteile bei *exakt zwei Beurteilungsinstanzen* entwickelt. Aus diesem Grund führen wir den Vergleich für alle zwei-elementigen Paarungen der Beurteilungsinstanzen durch.

2.3.4.1 Virtuelle Einheiten-Gesamtzahl ($A \times B$ | „AIAG MSA Extended“)

Die Einschränkung, dass wir nur zwei-elementige Paare (= zwei Spalten mit Urteilen) für das Bestimmen des Kappa-Koeffizienten nach Cohen verwenden können, führt uns zu der Frage, wie wir mit den mehrfachen Bewertungs-Durchgängen umzugehen haben. Der *heuristische* Ansatz bei der Berechnungsmethode „**AIAG MSA Extended**“ ist pragmatisch: Wir erzeugen das zweiseitige Schema, mit allen Kombinationen, die wir mit den einzelnen Prüfdurchgängen beider Prüfer bilden können. Die dabei entstehen Anzahl Zeilen, die über die

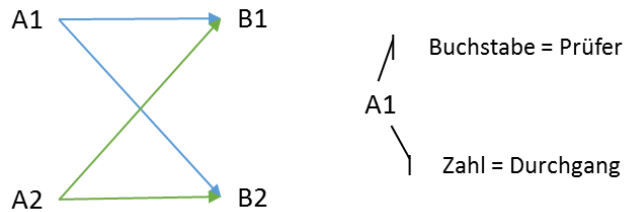
Anzahl der im Versuch verwendeten Einheiten hinaus geht, deuten wir auch hier als virtuell neue Einheiten. Somit entspricht die Anzahl der Zeilen in unserer Auswertungs-Datentabelle der virtuellen Gesamtanzahl an Einheiten, die wir wie folgt ermitteln.

Anzahl der Durchläufe der ersten Bewertungsinstanz mal Anzahl der Durchläufe der zweiten Bewertungsinstanz mal Anzahl der im Versuch verwendeten Einheiten:

$$N_{vo} = N_t \cdot N_t \cdot N_o = 2 \cdot 2 \cdot 5 = 20$$

2.3.4.2 Aufbau Datentabelle (A × B | „AIAG MSA Extended“)

Wir entnehmen der Tabelle 2 auf der Seite 10 die Urteile der beiden Prüfer A und B und bilden alle möglichen paarweisen Vergleiche:



Durch diese paarweisen Urteilsvergleiche erhalten wir insgesamt 20 virtuellen Einheiten:

Tabelle 32: Beispi-Datensatz - Wertetabelle Prüfer A vs. Prüfer B für die Berechnungsart AIAG MSA Extended

Virtuelle Einheiten-Nr.	Einheiten-Nr. im Versuch	Prüfer A	Prüfer B
1	Teil 1 A - 1. Durchgang vs. B - 1. Durchgang	Okay	Not okay
2	Teil 2 A - 1. Durchgang vs. B - 1. Durchgang	Okay	Okay
3	Teil 3 A - 1. Durchgang vs. B - 1. Durchgang	Not okay	Okay
4	Teil 4 A - 1. Durchgang vs. B - 1. Durchgang	Okay	Okay
5	Teil 5 A - 1. Durchgang vs. B - 1. Durchgang	Okay	Okay
6	Teil 1 A - 1. Durchgang vs. B - 2. Durchgang	Okay	Not okay
7	Teil 2 A - 1. Durchgang vs. B - 2. Durchgang	Okay	Not okay
8	Teil 3 A - 1. Durchgang vs. B - 2. Durchgang	Not okay	Okay
9	Teil 4 A - 1. Durchgang vs. B - 2. Durchgang	Okay	Okay
10	Teil 5 A - 1. Durchgang vs. B - 2. Durchgang	Okay	Okay
11	Teil 1 A - 2. Durchgang vs. B - 1. Durchgang	Okay	Not okay
12	Teil 2 A - 2. Durchgang vs. B - 1. Durchgang	Not okay	Okay
13	Teil 3 A - 2. Durchgang vs. B - 1. Durchgang	Okay	Okay
14	Teil 4 A - 2. Durchgang vs. B - 1. Durchgang	Okay	Okay
15	Teil 5 A - 2. Durchgang vs. B - 1. Durchgang	Okay	Okay
16	Teil 1 A - 2. Durchgang vs. B - 2. Durchgang	Okay	Not okay
17	Teil 2 A - 2. Durchgang vs. B - 2. Durchgang	Not okay	Not okay
18	Teil 3 A - 2. Durchgang vs. B - 2. Durchgang	Okay	Okay
19	Teil 4 A - 2. Durchgang vs. B - 2. Durchgang	Okay	Okay

20	Teil 5 A - 2. Durchgang vs. B - 2. Durchgang	Okay	Okay
----	--	------	------

2.3.4.3 Kreuztabellen (A × B | „AIAG MSA Extended“)

Wir zählen die Häufigkeit der Urteilskombinationen in den Zeilen der Tabelle 32 und erhalten so die folgende **Kreuztabelle der absoluten Häufigkeit der beobachteten Urteilskombinationen**.

Tabelle 33: Kreuztabelle der absoluten Häufigkeit beobachteter Urteilskombinationen für die Prüfer A und B nach der Berechnungsmethode „AIAG MSA Extended“

Vergleichbarkeit Prüfer A & B abs. Anzahl		Prüfer B		Summe
	Kategorie	Okay	Not okay	
Prüfer A	Okay	11	5	16
	Not okay	3	1	4
Summe		14	6	20

Wir teilen jeden Zahlenwert in der Tabelle 33 durch die virtuelle Einheiten-Gesamtzahl $N_{vo} = 20$ und erhalten als Ergebnis die **Kreuztabelle der Anteile beobachteter Urteilskombinationen**:

Tabelle 34: Kreuztabelle der Anteile beobachteter Urteilskombinationen für den Vergleich der Prüfer A und B nach der Berechnungsmethode „AIAG MSA Extended“

Vergleichbarkeit Prüfer A & B beob. Anteile		Prüfer B		Summe
	Kategorie	Okay	Not okay	
Prüfer A	Okay	0,55	0,25	0,80
	Not okay	0,15	0,05	0,20
Summe		0,70	0,30	1,00

Mit den Werten in den grün hinterlegten Zellen berechnen wir die **Summe der beobachteten Anteile übereinstimmender Urteile**:

$$p_o = 0,55 + 0,05 = 0,60$$

Multiplizieren wir die Spalten- und Zeilensummen der beobachteten Anteile der Urteils kombinationen, so erhalten wir als Ergebnis die **Kreuztabelle der Anteile erwarteter Urteils kombinationen, die durch den Zufall entstehen**.

Tabelle 35: Kreuztabelle der Anteile erwarteter Urteils kombinationen, die durch den Zufall entstehen, bestimmt für den Vergleich der Prüfer A und B nach der Berechnungsmethode „AIAG MSA

Vergleichbarkeit Prüfer A & B erw. Anteile		Prüfer B		Summe
	Kategorie	Okay	Not okay	
Prüfer A	Okay	0,56	0,24	0,80
	Not okay	0,14	0,06	0,20
Summe		0,70	0,30	1,00

Bilden wir die Summe mit den Werte in den grün hinterlegten Tabellenzellen, so entsteht der **durch den Zufall erwartete Anteil übereinstimmender Urteile**.

$$p_e = 0,56 + 0,06 = 0,62$$

2.3.4.4 Kappa-Koeffizient (A x B | „AIAG MSA Extended“)

Zunächst bereinigen wir den Anteil beobachteter, übereinstimmender Urteile vom Anteil der durch den Zufall erwarteten übereinstimmenden Urteile:

$$p_o - p_e = 0,60 - 0,62 = -0,02$$

Des weiteren Bestimmen wir den größtmöglichen Wert für den Anteil beobachteter gleicher Urteile, der vom zufälligen Anteil übereinstimmender Urteile bereinigt ist:

$$1 - p_e = 1 - 0,62 = 0,38$$

Den Kappa-Koeffizienten erhalten wir durch das Dividieren der beiden zuvor berechneten Anteile:

$$\kappa_{A \times B (\text{Cohen} | \text{AIAG MSA Extended})} = \frac{p_o - p_e}{1 - p_e} = \frac{-0,02}{0,38} \approx -0,052 \ 631$$

2.3.4.5 Signifikanztest Kappa-Koeffizient (A x B | „AIAG MSA Extended“)

Zu prüfende Hypothese:

Nullhypothese H0: Der Kappa-Koeffizient der Grundgesamtheit ist gleich Null

Praktische Deutung der Nullhypothese: Die übereinstimmenden Urteile sind vermutlich allein durch den Zufall entstanden.

Alternativhypothese H1: Der Kappa-Koeffizient der Grundgesamtheit ist ungleich Null.

Praktische Deutung der Alternativhypothese: Die übereinstimmenden Urteile sind nicht (allein) durch den Zufall zustande gekommen. Der Urteile folgen einer erkennbaren Systematik.

Für die Prüfung der Nullhypothese benötigen wir die Prüfgröße z, die wir aus der Division des Kappa-Koeffizienten durch die Standardabweichung des Kappa-Koeffizienten erhalten.

2.3.4.6 Standardabweichung (A x B | „AIAG MSA Extended“)

Würden wir den Versuch mehrfach wiederholt ausführen, so würde der Wert des Kappa-Indexes von Versuch zu Versuch streuen. Eine näherungsweise Abschätzung der Zufallsstreuung erhalten wir durch das Bestimmen der Standardabweichung des Kappa-Koeffizienten.

Den Wert der **Standardabweichung des Kappa-Koeffizienten** berechnen wir mit der im Abschnitt 2.1.4. angegebenen Beziehungsgleichung:

$$s_{\kappa_{A \times B}} = 0,215\,725\,3$$

2.3.4.7 Prüfgröße z (A x B | „AIAG MSA Extended“)

Teilen wir den Kappa-Index durch die Standardabweichung, so erhalten wir die **Prüfgröße z_{κ}** .

$$z_{\kappa_{A \times B}(\text{Cohen} | \text{AIAG MSA Extended})} = \frac{-0,052\,631}{0,215\,725\,3} \approx 1,369$$

2.3.4.8 P-Wert (A x B | „AIAG MSA Extended“)

Den Wert der Prüfgröße $z_{\kappa_{A \times B}(\text{Cohen} | \text{AIAG MSA Extended})}$ setzen wir in die Verteilungsfunktion der Standardnormalverteilung ein und erhalten gemäß der folgenden Beziehung den **P-Wert**:

$$P_{\kappa_{A \times B}(\text{Cohen} | \text{AIAG MSA Extended})} = 1 - G\left(z_{\kappa_{A \times B}(\text{Cohen} | \text{AIAG MSA Extended})}\right) = 1 - G(1,369) \approx 0,086$$

Interpretation des Testergebnisses anhand des P-Wertes

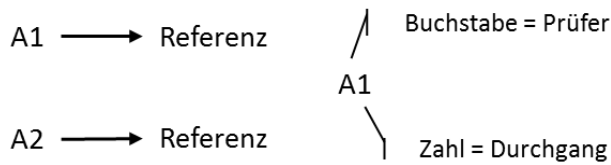
Da der P-Wert größer ist als das gewählte Signifikanzniveau $\alpha = 5\%$, wird die Nullhypothese verworfen.

2.3.5 Vergleichbarkeit Prüfer A vs. Referenz („AIAG MSA Extended“)

Wir wollen bei diesem Vergleich erfahren, wie gut die Urteile des Prüfer A mit den Referenz-Urteilen übereinstimmen.

2.3.5.1 Virtuelle Einheiten-Gesamtzahl (A x Ref. | „AIAG MSA Extended“)

Wir kombinieren die Urteile aus beiden Prüfdurchgängen mit den Referenzwerten:



Aus diesem Paarungsschema ergibt sich allgemein, dass die *virtuelle Gesamtzahl der Einheiten* N_{vo} sich aus dem Produkt der *Anzahl Durchläufe* N_t und der *Anzahl Versuchseinheiten* N_o ergibt.

Konkret für unser Daten-Beispiel erhalten wir die folgende *virtuelle Gesamtzahl Einheiten*:

$$N_{vo} = N_t \cdot N_o = 2 \cdot 5 = 10$$

2.3.5.2 Aufbau Datentabelle (A × Ref. | „AIAG MSA Extended“)

Aus der Tabelle 2 auf der Seite 10 entnehmen wir die Urteile des Prüfers A und der Referenz und transformieren diese gemäß dem Aufbauschema einer Tabelle mit zwei Spalten für die Urteile der beiden Urteilsinstanzen.

Tabelle 36: Datentabelle für die Analyse der Vergleichbarkeit der Urteile des Prüfers A mit den Referenz-Urteilen nach der Berechnungsmethode „AIAG MSA Extended“

Virtuelle Einheiten-Nr.	Nr. der Einheiten im Versuch	Urteile Prüfer A	Referenz-Urteil
1	Einheit 1 Durchgang 1	Okay	Okay
2	Einheit 2 Durchgang 1	Okay	Not okay
3	Einheit 3 Durchgang 1	Not okay	Okay
4	Einheit 4 Durchgang 1	Okay	Not okay
5	Einheit 5 Durchgang 1	Okay	Okay
6	Einheit 1 Durchgang 2	Okay	Okay
7	Einheit 2 Durchgang 2	Not okay	Not okay
8	Einheit 3 Durchgang 2	Okay	Okay
9	Einheit 4 Durchgang 2	Okay	Not okay
10	Einheit 5 Durchgang 2	Okay	Okay

2.3.5.3 Kreuztabellen (A × Ref. | „AIAG MSA Extended“)

Zählen wir zeilenweise die Häufigkeit der einzelnen Urteils kombinationen in der Tabelle 36, so erhalten wir die **Kreuztabelle der absoluten Häufigkeit beobachteter Urteils kombinationen**:

Tabelle 37: Kreuztabelle der absoluten Häufigkeit beobachteter Urteils kombinationen für die Vergleichbarkeit der Urteile des Prüfers A mit den Referenz-Urteilen nach der Berechnungsmethode „AIAG MSA Extended“

Vergleichbarkeit Prüfer A vs. Ref. abs. Anzahl		Referenz-Urteile		Summe
	Kategorie	Okay	Not okay	
Prüfer A	Okay	5	1	6
	Not okay	3	1	4
Summe		8	2	10

Teilen wir jeden der Zahlenwerte in der Tabelle 37 durch die *virtuelle Gesamtzahl der Einheiten* N_{v0} , erhalten wir die **Kreuztabelle der Anteile beobachteter Urteilskombinationen**.

Tabelle 38: Kreuztabelle der Anteile beobachteter Urteilskombinationen für die Vergleichbarkeit der Urteile des Prüfers A mit den Referenz-Urteilen nach der Berechnungsmethode „AIAG MSA Extended“

Vergleichbarkeit Prüfer A vs. Ref. beob. Anteile		Referenz-Urteile		Summe
	Kategorie	Okay	Not okay	
Prüfer A	Okay	0,50	0,10	0,60
	Not okay	0,30	0,10	0,40
Summe		0,80	0,20	1,00

Bilden wir die Summe mit den Werten in den grün hinterlegten Tabellenzellen, so erhalten wir den Anteil beobachteter Urteilsübereinstimmungen:

$$p_o = 0,50 + 0,10 = 0,60$$

Multiplizieren wir die Zeilensummen und Spaltensummen miteinander, so erhalten wir die Kreuztabelle der Anteile erwarteter Urteilskombinationen, die durch den Zufall entstanden sind:

Tabelle 39: Kreuztabelle der Anteile erwarteter Urteilskombinationen, die durch den Zufall entstanden sind, für die Vergleichbarkeit der Urteile des Prüfers A mit den Referenz-Urteilen nach der Berechnungsmethode „AIAG MSA Extended“

Vergleichbarkeit Prüfer A vs. Ref. erw. Anteile		Referenz-Urteile		Summe
	Kategorie	Okay	Not okay	
Prüfer A	Okay	0,48	0,12	0,60
	Not okay	0,32	0,08	0,40
Summe		0,80	0,20	1,00

Bilden wir die Summe mit den beiden grün hinterlegten Zellenwerten aus der Tabelle 39, so erhalten wir den **Anteil erwarteter Urteilsübereinstimmungen, der durch den Zufall entstanden ist**.

$$p_e = 0,48 + 0,08 = 0,56$$

2.3.5.4 Kappa-Koeffizient (A x Ref. | „AIAG MSA Extended“)

Zunächst bestimmen wir den Anteil der beobachteten, übereinstimmenden Urteile, der vom Anteil zufällig übereinstimmender Urteile bereinigt ist:

$$p_o - p_e = 0,60 - 0,56 = 0,04$$

Anschließend bestimmen wir den größtmöglichen Wert für den Anteil beobachteter, übereinstimmender Urteile, der vom Anteil zufällig übereinstimmender Urteile bereinigt ist:

$$1 - p_e = 1 - 0,56 = 0,44$$

Teilen wir die beiden zuletzt berechneten Anteilswerte durcheinander, so erhalten wir den Kappa-Koeffizienten:

$$\kappa_{A \times Ref \text{ (Cohen | AIAG MSA Extended)}} = \frac{0,04}{0,44} = 0,09\overline{09}$$

2.3.5.5 Signifikanztest (A x Ref. | „AIAG MSA Extended“)

Zu prüfende Hypothese:

Nullhypothese H0: Der Kappa-Koeffizient der Grundgesamtheit ist gleich Null

Praktische Deutung der Nullhypothese: Die übereinstimmenden Urteile sind vermutlich allein durch den Zufall entstanden.

Alternativhypothese H1: Der Kappa-Koeffizient der Grundgesamtheit ist ungleich Null.

Praktische Deutung der Alternativhypothese: Die übereinstimmenden Urteile sind nicht (allein) durch den Zufall zustande gekommen. Der Urteile folgen einer erkennbaren Systematik.

Für die Prüfung der Nullhypothese benötigen wir die Prüfgröße z, die wir aus der Division des Kappa-Koeffizienten durch die Standardabweichung des Kappa-Koeffizienten erhalten.

2.3.5.6 Standardabweichung (A × Ref. | „AIAG MSA Extended“)

Würden wir den Versuch mehrfach wiederholt ausführen, so würde der Wert des Kappa-Indexes von Versuch zu Versuch streuen. Eine näherungsweise Abschätzung der Zufallsstreuung erhalten wir durch das Bestimmen der Standardabweichung des Kappa-Koeffizienten.

Den Wert der **Standardabweichung des Kappa-Koeffizienten** berechnen wir mit der Berechnungsgleichung aus dem Abschnitt 2.1.4.

$$S_{\kappa_{A \times \text{Ref}}(\text{Cohen} | \text{AIAG MSA Extended})} = 0,281\,672$$

2.3.5.7 Prüfgröße z (A × Ref. | „AIAG MSA Extended“)

Teilen wir den Kappa-Koeffizienten durch die Standardabweichung des Kappa-Koeffizienten, erhalten wir die Prüfgröße z:

$$Z_{\kappa_{A \times \text{Ref}}(\text{Cohen} | \text{AIAG MSA Extended})} = \frac{\kappa_{A \times \text{Ref}}(\text{Cohen} | \text{AIAG MSA Extended})}{S_{\kappa_{A \times \text{Ref}}(\text{Cohen} | \text{AIAG MSA Extended})}} = \frac{0,090\,909}{0,281\,672} = 0,323$$

2.3.5.8 P-Wert (A × Ref. | „AIAG MSA Extended“)

Den P-Wert bestimmen wir gemäß der folgenden Beziehung:

$$P_{\kappa_{A \times \text{Ref}}} = 1 - G\left(Z_{\kappa_{A \times \text{Ref}}(\text{Cohen} | \text{AIAG MSA Extended})}\right) = 1 - G(0,323) \approx 0,373$$

Interpretation des Testergebnisses anhand des P-Wertes

Aufgrund des viel zu kleinen Stichprobenumfangs ist die Power des Tests (= Nullhypothese verwerfen, wenn diese nicht zutrifft) viel zu gering und daher das Testen an sich hier unsinnig. Dennoch vollziehen wir den formalen Akt des „Test-Rituals“ mit dem P-Wert.

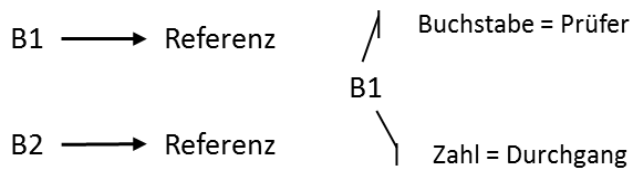
Der P-Wert wird mit dem gewählten Signifikanzniveau (hier: $\alpha = 5\%$) verglichen. Hier ist der P-Wert größer als das Signifikanzniveau und die Nullhypothese wird daher **nicht** verworfen. Die übereinstimmenden Urteile sind zufälliger Natur.

2.3.6 Vergleichbarkeit Prüfer B vs. Referenz („AIAG MSA Extended“)

Wir wollen wissen, wie gut die Urteile des Prüfers B mit den Referenz-Urteilen übereinstimmen.

2.3.6.1 Virtuelle Anzahl der Einheiten (B x Ref. | „AIAG MSA Extended“)

Wir kombinieren jeden der beiden Prüfdurchgänge des Prüfers mit den Referenzwerten.



Auf diese Weise erhalten wir die virtuelle Gesamtzahl Einheiten:

$$N_{vo} = N_t \cdot N_o = 2 \cdot 5 = 10$$

2.3.6.2 Aufbau Datentabelle (B × Ref. | „AIAG MSA Extended“)

Wir entnehmen aus der Tabelle 2 auf der Seite 10 die Urteile des Prüfers B und die Referenz-Urteile und transformieren die Werte in das Schema der folgenden Tabelle, in der zwei Spalten für die Urteile beider Urteilsinstanzen enthalten sind.

Tabelle 40: Daten-Beispiel – Vergleich der Urteile des Prüfers B mit den Referenz-Urteilen für die Berechnungsmethode „AIAG MSA Extended“

Objekt-Nr.	Nr. der Einheiten im Versuch	Urteile Prüfer B	Referenz- urteil
1	Einheit 1 Durchgang 1	Not okay	Okay
2	Einheit 2 Durchgang 1	Okay	Not okay
3	Einheit 3 Durchgang 1	Okay	Okay
4	Einheit 4 Durchgang 1	Okay	Not okay
5	Einheit 5 Durchgang 1	Okay	Okay
6	Einheit 1 Durchgang 2	Not okay	Okay
7	Einheit 2 Durchgang 2	Not okay	Not okay
8	Einheit 3 Durchgang 2	Okay	Okay
9	Einheit 4 Durchgang 2	Okay	Not okay
10	Einheit 5 Durchgang 2	Okay	Okay

2.3.6.3 Kreuztabellen (B × Ref. | „AIAG MSA Extended“)

Wir zählen die Häufigkeit der Urteilskombinationen in den Zeilen der Tabelle 40 und erhalten so die

Kreuztabelle der absoluten Häufigkeit der Urteilskombinationen:

Tabelle 41: Kreuztabelle der absoluten Häufigkeit der Urteilskombinationen für den Vergleich des Prüfers B mit der Referenz nach der Berechnungsmethode „AIAG MSA Extended“

Vergleichbarkeit Prüfer B vs. Ref. abs. Anzahl		Referenz-Urteile		Summe
	Kategorie	Okay	Not okay	
Prüfer B	Okay	4	2	6
	Not okay	3	1	4
Summe		7	3	10

Teilen wir jeden Zahlenwert in der Tabelle 41 durch die virtuelle Gesamtzahl der Einheiten $N_{vo} = 10$, so erhalten wir die folgende **Kreuztabelle der Anteile beobachteter Urteils kombinationen**:

Tabelle 42: Kreuztabelle der Anteile beobachteter Urteils kombinationen für den Vergleich des Prüfers B mit der Referenz nach der Berechnungsmethode „AIAG MSA Extended“

Vergleichbarkeit Prüfer B vs. Ref. beob. Anteil		Referenz-Urteile		Summe
	Kategorie	Okay	Not okay	
Prüfer B	Okay	0,40	0,20	0,60
	Not okay	0,30	0,10	0,40
Summe		0,70	0,30	1,00

Mit den Werten in den grün hinterlegten Zellen der Tabelle 42 bilden wir den Anteil beobachteter Urteilsübereinstimmungen:

$$p_o = 0,40 + 0,10 = 0,50$$

Multiplizieren wir die Zeilensummen mit den Spaltensummen (beide aus der Tabelle 42), so erhalten wir die folgende **Kreuztabelle mit den Anteilen der durch den Zufall erwarteten Urteils kombinationen**:

Tabelle 43: Kreuztabelle der Anteile erwarteter Urteilskombinationen, die aufgrund des Zufalls entstehen, für den Vergleich des Prüfers B mit der Referenz nach der Berechnungsmethode „AIAG MSA Extended“

Vergleichbarkeit Prüfer B vs. Ref. erw. Anteil		Referenz-Urteile		Summe
	Kategorie	Okay	Not okay	
Prüfer B	Okay	0,42	0,18	0,60
	Not okay	0,28	0,12	0,40
Summe		0,70	0,30	1,00

Bilden wir die Summe mit den Werten in den grün hinterlegten Tabellenzellen, so erhalten wir den **Anteil übereinstimmender Urteils kombinationen, der aufgrund des Zufalls zu erwarten ist**.

$$p_e = 0,42 + 0,12 = 0,54$$

2.3.6.4 Kappa-Koeffizient (B x Ref. | „AIAG MSA Extended“)

Zunächst bilden wir den **Anteil beobachteter Übereinstimmungen, der von dem Anteil zufällig erwarteter Übereinstimmungen bereinigt ist**:

$$p_o - p_e = 0,50 - 0,54 = -0,04$$

Anschließend bestimmen wir den **größtmöglichen Wert für den Anteil beobachteter Übereinstimmungen, der von dem Anteil zufällig erwarteter Übereinstimmungen bereinigt ist**:

$$1 - p_e = 1 - 0,54 = 0,46$$

Teilen wir die beiden zuletzt berechneten Anteilswerte durcheinander, so erhalten wir den Kappa-Koeffizienten:

$$\kappa_{B \times Ref(Cohen | AIAG MSA Extended)} = \frac{p_o - p_e}{1 - p_e} = \frac{-0,04}{0,46} \approx -0,087$$

2.3.6.5 Signifikanztest (B x Ref. | „AIAG MSA Extended“)

Zu prüfende Hypothese:

Nullhypothese H0: Der Kappa-Koeffizient der Grundgesamtheit ist gleich Null

Praktische Deutung der Nullhypothesen: Die übereinstimmenden Urteile sind vermutlich allein durch den Zufall entstanden.

Alternativhypothese H1: Der Kappa-Koeffizient der Grundgesamtheit ist ungleich Null.

Praktische Deutung der Alternativhypothese: Die übereinstimmenden Urteile sind nicht (allein) durch den Zufall zustande gekommen. Der Urteile folgen einer erkennbaren Systematik.

Für die Prüfung der Nullhypothese benötigen wir die Prüfgröße z, die wir aus der Division des Kappa-Koeffizienten durch die Standardabweichung des Kappa-Koeffizienten erhalten.

2.3.6.6 Standardabweichung (B x Ref. | „AIAG MSA Extended“)

Würden wir den Versuchs mehrfach wiederholt ausführen, so würde der Wert des Kappa-Indexes von Versuchs zu Versuch streuen. Eine näherungsweise Abschätzung der Zufallsstreuung erhalten wir durch das Bestimmen der Standardabweichung des Kappa-Koeffizienten.

Den Wert der **Standardabweichung des Kappa-Koeffizienten** berechnen wir mit der Berechnungsgleichung aus dem Abschnitt 2.1.4.

$$s_{\kappa_{B \times Ref(Cohen | AIAG MSA Extended)}} = 0,308\ 665$$

2.3.6.7 Prüfgröße z (B x Ref. | „AIAG MSA Extended“)

Teilen wir den Kappa-Koeffizienten durch die Standardabweichung des Kappa-Koeffizienten, erhalten wir die Prüfgröße z:

$$z_{\kappa_{B \times Ref(Cohen | AIAG MSA Extended)}} = \frac{\kappa_{B \times Ref(Cohen | AIAG MSA Extended)}}{s_{\kappa_{B \times Ref(Cohen | AIAG MSA Extended)}}} = \frac{-0,087}{0,308\ 665} \approx -0,282$$

2.3.6.8 P-Wert (B x Ref. | „AIAG MSA Extended“)

Den P-Wert bestimmen wir gemäß der folgenden Beziehung:

$$P_{\kappa_{B \times Ref}} = 1 - G\left(z_{\kappa_{B \times Ref(Cohen | AIAG MSA Extended)}}\right) = 1 - G(-0,282) \approx 0,611$$

Interpretation des Testergebnisses anhand des P-Wertes

Aufgrund des viel zu kleinen Stichprobenumfangs ist die Power des Tests (= Nullhypothese verwerfen, wenn diese nicht zutrifft) viel zu gering und daher das Testen an sich hier unsinnig. Dennoch vollziehen wir den formalen Akt des „Test-Rituals“ mit dem P-Wert.

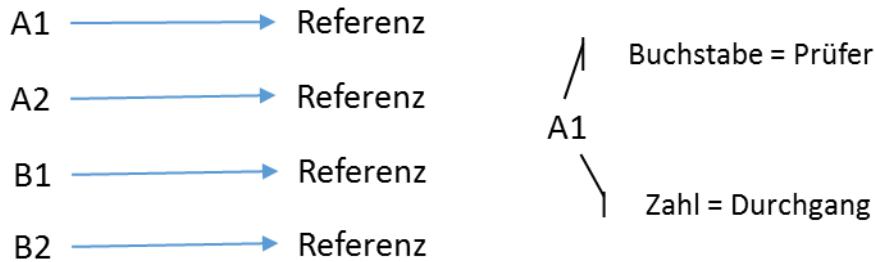
Da der P-Wert größer ist als das hier gewählte Signifikanzniveau $\alpha = 5\%$ wird die Nullhypothese nicht abgelehnt (also beibehalten). Wir deuten die beobachteten Urteilsübereinstimmungen allein aufgrund des Zufalls entstanden.

2.3.7 Vergleichbarkeit aller Prüfer vs. Referenz („AIAG MSA Extended“)

Wie einleitend im Abschnitt 2.1.1 beschrieben, wurde der Kappa-Koeffizient nach Jacob Cohen für den Vergleich der Urteile von exakt zwei Urteilsinstanzen entwickelt. Wir müssen die Daten so anordnen, dass zwei Urteilsspalten vorhanden sind.

2.3.7.1 Virtuelle Einheiten-Gesamtzahl ($A \times B \times \text{Ref}$ | „AIAG MSA Extended“)

Für jeden Prüfer kombinieren wir die Ergebnisse aus beiden Durchläufen mit den Referenzwerten (Daten aus der Tabelle 2 auf der Seite 10):



Für das Zahlenbeispiel mit fünf Einheiten und zwei Durchgängen je Prüfer, die von den beiden Prüfern A und B jeweils in zwei Bewertungsgängen beurteilt werden. Das ergibt die virtuelle Gesamtzahl Einheiten:

$$N_{vo} = \sum_{i=1}^{N_a} N_o \cdot N_t = 5 \cdot 2 + 5 \cdot 2 = 20$$

2.3.7.2 Aufbauschema Datentabelle (A × B × Ref. | „AIAG MSA Extended“)

Das Ergebnis der zuvor beschriebenen Transformation der Versuchsdaten aus der Tabelle 2 zur Datentabelle für die Analyse finden wir in der nachfolgend dargestellten Tabelle.

Tabelle 44: Analyse-Datentabelle für den Vergleich der Prüferurteile mit den Referenzbewertungen nach der Berechnungsmethode „AIAG MSA Extended“

Objekt-Nr.	Nr. der Einheiten im Versuch	Referenz- Urteile	Urteile der Prüfer A & B
1	A Einheit 1 Durchgang 1	Okay	Okay
2	A Einheit 2 Durchgang 1	Not okay	Okay
3	A Einheit 3 Durchgang 1	Okay	Not okay
4	A Einheit 4 Durchgang 1	Not okay	Okay
5	A Einheit 5 Durchgang 1	Okay	Okay
6	A Einheit 1 Durchgang 2	Okay	Okay
7	A Einheit 2 Durchgang 2	Not okay	Not okay
8	A Einheit 3 Durchgang 2	Okay	Okay
9	A Einheit 4 Durchgang 2	Not okay	Okay
10	A Einheit 5 Durchgang 2	Okay	Okay
11	B Einheit 1 Durchgang 1	Okay	Not okay
12	B Einheit 2 Durchgang 1	Not okay	Okay
13	B Einheit 3 Durchgang 1	Okay	Okay
14	B Einheit 4 Durchgang 1	Not okay	Okay
15	B Einheit 5 Durchgang 1	Okay	Okay
16	B Einheit 1 Durchgang 2	Okay	Not okay
17	B Einheit 2 Durchgang 2	Not okay	Not okay

18	B Einheit 3 Durchgang 2	Okay	Okay
19	B Einheit 4 Durchgang 2	Not okay	Okay
20	B Einheit 5 Durchgang 2	Okay	Okay

2.3.7.3 Kreuztabellen (A × B × Ref. | „AIAG MSA Extended“)

Wir zählen die Häufigkeit der Urteilskombinationen in den Zeilen der Tabelle 44 und erhalten daraus die **Kreuztabelle der absoluten Häufigkeit beobachteter Urteilskombinationen**.

Tabelle 45: Kreuztabelle der absoluten Häufigkeit beobachteter Urteilskombinationen für den Vergleich aller Prüfer mit der Referenz nach der Berechnungsmethode „AIAG MSA Extended“

Vergleichbarkeit Prüfer vs. Ref. abs. Anzahl		Referenz-Urteile		Summe
	Kategorie	Okay	Not okay	
Alle Prüfer	Okay	9	6	15
	Not okay	3	2	5
Summe		12	8	20

Teilen wir jeden Zahlenwert der Tabelle 45 durch die virtuelle Gesamtzahl der Einheiten $N_{vo} = 20$, erhalten wir die **Kreuztabelle der Anteile beobachteter Urteilskombinationen**.

Tabelle 46: Kreuztabelle der Anteile beobachteter Urteilskombinationen für den Vergleich aller Prüfer mit der Referenz nach der Berechnungsmethode „AIAG MSA Extended“

Vergleichbarkeit Prüfer vs. Ref. beob. Anteil		Referenz-Urteile		Summe
	Kategorie	Okay	Not okay	
Alle Prüfer	Okay	0,45	0,30	0,75
	Not okay	0,15	0,10	0,25
Summe		0,60	0,40	1,00

Bilden wir mit den Werten in den grün hinterlegten Tabellenzellen die Summe, so erhalten wir den beobachteten Anteil übereinstimmender Urteile:

$$p_o = 0,45 + 0,10 = 0,55$$

Multiplizieren wir die Zeilen- und Spaltensummen der Tabelle 46 miteinander, so erhalten wir die **Kreuztabelle der Anteile für die Urteils kombinationen, die wir aufgrund des Zufalls erwarten**.

Tabelle 47: Kreuztabelle der Anteile erwarteter Urteil kombinationen

für den Vergleich aller Prüfer mit der Referenz nach der Berechnungsmethode „AIAG MSA Extended“

Vergleichbarkeit Prüfer vs. Ref. erw. Anteil		Referenz-Urteile		Summe
	Kategorie	Okay	Not okay	
Alle Prüfer	Okay	0,45	0,30	0,75
	Not okay	0,15	0,10	0,25
Summe		0,60	0,40	1,00

Bilden wir mit den Zahlenwerten in den grün hinterlegten Zellen der Tabelle 47 die Summe, so erhalten wir den **Anteil übereinstimmender Urteile, den wir aufgrund des Zufalls erwarten**:

$$p_e = 0,45 + 0,10 = 0,55$$

2.3.7.4 Kappa-Koeffizient (A × B × Ref. | „AIAG MSA Extended“)

Zunächst bilden wir den **Anteil beobachteter Übereinstimmungen, der vom Anteil zufällig erwarteter Übereinstimmungen bereinigt ist**:

$$p_o - p_e = 0,55 - 0,55 = 0,00$$

Anschließend bestimmen wir den **größtmöglichen Wert für den Anteil beobachteter Übereinstimmungen, der vom Anteil zufällig erwarteter Übereinstimmungen bereinigt ist**:

$$1 - p_e = 1 - 0,55 = 0,45$$

Teilen wir die beiden zuletzt berechneten Anteilswerte durcheinander, so erhalten wir den Kappa-Koeffizienten:

$$K_{A \times B \times Ref(Cohen | AIAG MSA Extended)} = \frac{p_o - p_e}{1 - p_e} = \frac{0,00}{0,45} = 0,00$$

2.3.7.5 Signifikanztest Kappa-Koeffizient (A x B x Ref. | „AIAG MSA Extended“)

Zu prüfende Hypothese:

Nullhypothese H0: Der Kappa-Koeffizient der Grundgesamtheit ist gleich Null

Praktische Deutung der Nullhypothese: Die übereinstimmenden Urteile sind vermutlich allein durch den Zufall entstanden.

Alternativhypothese H1: Der Kappa-Koeffizient der Grundgesamtheit ist ungleich Null.

Praktische Deutung der Alternativhypothese: Die übereinstimmenden Urteile sind nicht (allein) durch den Zufall zustande gekommen. Der Urteile folgen einer erkennbaren Systematik.

Für die Prüfung der Nullhypothese benötigen wir die Prüfgröße z, die wir aus der Division des Kappa-Koeffizienten durch die Standardabweichung des Kappa-Koeffizienten erhalten.

2.3.7.6 Standardabweichung (A x B x Ref. | „AIAG MSA Extended“)

Würden wir den Versuchs mehrfach wiederholt ausführen, so würde der Wert des Kappa-Indexes von Versuchs zu Versuch streuen. Eine näherungsweise Abschätzung der Zufallsstreuung erhalten wir durch das Bestimmen der Standardabweichung des Kappa-Koeffizienten.

Den Wert der **Standardabweichung des Kappa-Koeffizienten** berechnen wir mit der Berechnungsgleichung aus dem Abschnitt 2.1.4.

$$S_{K_{A \times B \times \text{Ref}}(\text{Cohen} | \text{AIAG MSA Extended})} = 0,2108185$$

2.3.7.7 Prüfgröße z (A x B x Ref. | „AIAG MSA Extended“)

Teilen wir den Kappa-Koeffizienten durch die Standardabweichung des Kappa-Koeffizienten, erhalten wir die Prüfgröße z:

$$z_{K_{A \times B \times \text{Ref}}(\text{Cohen} | \text{AIAG MSA Extended})} = \frac{K_{A \times B \times \text{Ref}}(\text{Cohen} | \text{AIAG MSA Extended})}{S_{K_{A \times B \times \text{Ref}}(\text{Cohen} | \text{AIAG MSA Extended})}} = \frac{0,00}{0,2108185} \approx 0,00$$

2.3.7.8 P-Wert (A x B x Ref. | „AIAG MSA Extended“)

Den P-Wert bestimmen wir gemäß der folgenden Beziehung:

$$P_{K_{A \times B \times \text{Ref}}} = 1 - G\left(z_{K_{A \times B \times \text{Ref}}(\text{Cohen} | \text{AIAG MSA Extended})}\right) = 1 - G(0) = 0,50$$

Interpretation des Testergebnisses

Aufgrund des viel zu kleinen Stichprobenumfangs ist die Power des Tests (= Nullhypothese verwerfen, wenn diese nicht zutrifft) viel zu gering und daher das Testen an sich hier unsinnig. Dennoch vollziehen wir den formalen Akt des „Test-Rituals“ mit dem P-Wert.

Der P-Wert ist größer als das gewählte Signifikanzniveau $\alpha = 5\%$, wird die Nullhypothese nicht verworfen. Beobachtete Übereinstimmungen in der Stichprobe sind vermutlich nur durch den Zufall entstanden.

2.4 Kappa nach Cohen - Berechnungsart „Standard Calculation“

Wie einleitend im Abschnitt 2.1.1 beschrieben, ist der Kappa-Koeffizient für den Vergleich der Urteile von exakt zwei Urteilsinstanzen konstruiert worden. Wollen wir mehr als zwei Instanzen miteinander vergleichen oder haben wir mehrere Prüfdurchläufe ausgeführt, so müssen wir die Versuchsdaten in das Schema einer Datentabelle mit zwei Urteilsspalten transformieren. Auch die hier beschriebenen Transformationen sind heuristischer Natur, die im Original-Aufsatz von Jacob Cohen [Coh] nicht enthalten sind.

Mit der Berechnungsart „Standard Calculation“ können wir die folgenden Ergebnis-Inhalte berechnen:

Berechnungen **ohne Referenzvergleich**

- Wiederholbarkeit innerhalb der einzelnen Prüfer (Allein bei zwei Prüfdurchläufen, **bei mehr als zwei Durchläufen erfolgt keine Berechnung der Wiederholbarkeit**)
- **Keine Vergleichbarkeit aller Prüfer untereinander**

Berechnungen **mit Referenzvergleich**

- Vergleich der einzelnen Prüfer zur Referenz
- Vergleich aller Prüfer gemeinsam zur Referenz

2.4.1 Wiederholbarkeit Prüfer A („Standard Calculation“)

Wir fragen uns, wie gut die Urteile der einzelnen Prüfer im ersten und zweiten Prüfdurchgang übereingestimmt haben. Speziell mit den Daten unseres Fallbeispiels ergibt sich für die Wiederholbarkeit innerhalb der einzelnen Prüfer aufgrund der zwei Prüfdurchgänge „ganz natürlich“ eine Tabelle mit zwei Urteilsspalten.

Sollte in einem anderen Anwendungsfall jeder Prüfer mehr als zwei Prüfdurchgänge ausgeführt haben, so ist die Berechnung der Wiederholbarkeit nicht möglich, da mehr als zwei Spalten mit Urteilswerten vorliegen (Bei dieser Berechnungsart („Standard Calculation“) ist für die Wiederholbarkeit keine Transformation der Versuchsdaten in das zweisepaltige Aufbauschema vorgesehen).

2.4.1.1 Virtuelle Einheiten-Anzahl (A | „Standard Calculation“)

Bei dieser Berechnungsart werden keine Transformationen der Versuchsdaten in das zweisepaltige Schema der Analyse-Tabelle ausgeführt. Die Anzahl der Einheiten entspricht der im Versuch verwendeten Anzahl Einheiten.

2.4.1.2 Aufbau Datentabelle (A | „Standard Calculation“)

Wie zuvor erwähnt ist Berechnung der Wiederholbarkeit für die Berechnungsart „Standard Calculation“ nur möglich, wenn jeder Prüfer *exakt zwei Prüfdurchgänge* ausgeführt hat. D.h., wir entnehmen aus der Tabelle der Versuchdaten (siehe Tabelle 2 auf der Seite 10) allein die Urteile des Prüfers A:

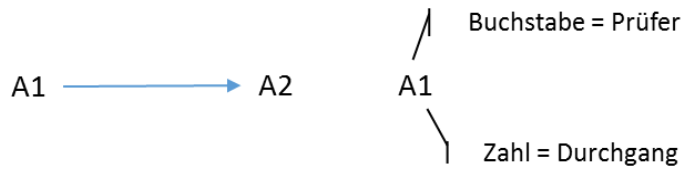


Abbildung 11. Datenaufbau-Schema für die Wiederholbarkeit Prüfer A (Cohen | Standard Calculation)

Mit diesem Vergleich erhalten wir die folgende Tabelle:

Tabelle 48: Datentabelle für die Wiederholbarkeit des Prüfers A nach der Berechnungsart „Standard Calculation“

Virtuelle Einheiten-Nr.	Prüfer A	Prüfer A
	1. Durchgang	2. Durchgang
1	Okay	Okay
2	Okay	Not okay
3	Not okay	Okay
4	Okay	Okay
5	Okay	Okay

2.4.1.3 Kreuztabellen (A | „Standard Calculation“)

Zählen wir die Häufigkeit der Urteils kombinationen in den Zeilen der Tabelle 48, so erhalten wir die folgende

Kreuztabelle der absoluten Häufigkeit der Urteils kombinationen des Prüfers A:

Tabelle 49: Kreuztabelle der absoluten Häufigkeit der Urteils kombination für die Wiederholbarkeit des Prüfers A nach der Berechnungsmethode „Standard Calculation“

Wiederholbarkeit Prüfer A abs. Anzahl		Prüfer A 2. Durchgang		Summe
	Kategorie	Okay	Not okay	
Prüfer A 1. Durchgang	Okay	3	1	4
	Not okay	1	0	1
Summe		4	1	5

Teilen wir jeden Zahlenwert in der Tabelle 49 durch die Anzahl der Einheiten $N_o = 5$, so erhalten wir die **Kreuztabelle der Anteile beobachteter Urteils kombinationen für die Wiederholbarkeit** des Prüfers A:
Tabelle 50: Kreuztabelle der Anteile beobachteter Urteils kombinationen für die Wiederholbarkeit des Prüfers A
nach der Berechnungsart „Standard Calculation“

Wiederholbarkeit Prüfer A beob. Anteil		Prüfer A 2. Durchgang		Summe
	Kategorie	Okay	Not okay	
Prüfer A 1. Durchgang	Okay	0,60	0,20	0,80
	Not okay	0,20	0,00	0,20
Summe		0,80	0,20	1,00

Bilden wir die Summe mit den Werten in den grün hinterlegten Zellen der Tabelle 50, so erhalten wir den beobachteten Anteil übereinstimmender Urteile:
 $p_o = 0,60 + 0,00 = 0,60$

Multiplizieren wir die Zeilen- mit den Spaltensummen der Tabelle 50 miteinander, so erhalten wir die **Kreuztabelle des Anteils erwarteter Urteils kombinationen, die aufgrund des Zufalls entstanden sind:**

Wiederholbarkeit Prüfer A erw. Anteil		Prüfer A 2. Durchgang		Summe
	Kategorie	Okay	Not okay	
Prüfer A 1. Durchgang	Okay	0,64	0,16	0,80
	Not okay	0,16	0,04	0,20
Summe		0,80	0,20	1,00

Bilden wir die Summe mit den Werten in den grün hinterlegten Zellen, so erhalten wir den erwarteten Anteil an zufällig übereinstimmenden Urteilen.
 $p_e = 0,64 + 0,04 = 0,68$

2.4.1.4 Kappa-Koeffizient (A | „Standard Calculation“)

Zunächst bilden wir den **Anteil beobachteter Übereinstimmungen, der von dem Anteil zufällig erwarteter Übereinstimmungen bereinigt ist**:

$$p_o - p_e = 0,60 - 0,68 = -0,08$$

Anschließend bestimmen wir den **größtmöglichen Wert für den Anteil beobachteter Übereinstimmungen, der von dem Anteil zufällig erwarteter Übereinstimmungen bereinigt ist**:

$$1 - p_e = 1 - 0,68 = 0,32$$

Teilen wir die bereinigten Anteilswerte durcheinander, so erhalten wir den Kappa-Koeffizienten:

$$\kappa_{A(\text{Cohen}|\text{Standard Calculation})} = \frac{p_o - p_e}{1 - p_e} = \frac{-0,08}{0,32} = -0,25$$

2.4.1.5 Signifikanztest (A | „Standard Calculation“)

Zu prüfende Hypothese:

Nullhypothese H0: Der Kappa-Koeffizient der Grundgesamtheit ist gleich Null

Praktische Deutung der Nullhypothesen: Die übereinstimmenden Urteile sind vermutlich allein durch den Zufall entstanden.

Alternativhypothese H1: Der Kappa-Koeffizient der Grundgesamtheit ist ungleich Null.

Praktische Deutung der Alternativhypothese: Die übereinstimmenden Urteile sind nicht (allein) durch den Zufall zustande gekommen. Der Urteile folgen einer erkennbaren Systematik.

Für die Prüfung der Nullhypothese benötigen wir die Prüfgröße z, die wir aus der Division des Kappa-Koeffizienten durch die Standardabweichung des Kappa-Koeffizienten erhalten.

2.4.1.6 Standardabweichung (A | „Standard Calculation“)

Würden wir den Versuch mehrfach wiederholt ausführen, so würde der Wert des Kappa-Indexes von Versuch zu Versuch streuen. Eine näherungsweise Abschätzung der Zufallsstreuung erhalten wir durch das Bestimmen der Standardabweichung des Kappa-Koeffizienten.

Den Wert der **Standardabweichung des Kappa-Koeffizienten** berechnen wir mit der Berechnungsgleichung aus dem Abschnitt 2.1.4.

$$s_{\kappa_{A(\text{Cohen}|\text{Standard Calculation})}} = 0,447\ 214$$

2.4.1.7 Prüfgröße z (A | „Standard Calculation“)

Teilen wir den Kappa-Koeffizienten durch die Standardabweichung des Kappa-Koeffizienten, erhalten wir die Prüfgröße z:

$$z_{\kappa_{A(\text{Cohen}|\text{Standard Calculation})}} = \frac{\kappa_{A(\text{Cohen}|\text{Standard Calculation})}}{s_{\kappa_{A(\text{Cohen}|\text{Standard Calculation})}}} = \frac{-0,25}{0,447\ 214} \approx -0,559$$

2.4.1.8 P-Wert (A | „Standard Calculation“)

Den P-Wert bestimmen wir gemäß der folgenden Beziehung:

$$P_{\kappa_A} = 1 - G\left(z_{\kappa_{A(\text{Cohen}|\text{Standard Calculation})}}\right) = 1 - G(-0,559) = 0,712$$

Interpretation des Testergebnisses anhand des P-Wertes

Aufgrund des viel zu kleinen Stichprobenumfangs ist die Power des Tests (= Nullhypothese verwerfen, wenn diese nicht zutrifft) viel zu gering und daher das Testen an sich hier unsinnig. Dennoch vollziehen wir den formalen Akt des „Test-Rituals“ mit dem P-Wert.

Da der P-Wert größer ist als das gewählte Signifikanzniveau $\alpha = 5\%$, wird die Nullhypothese nicht verworfen (= beibehalten). Die in der Stichprobe beobachteten Übereinstimmungen sind vermutlich allein aufgrund des Zufalls entstanden.

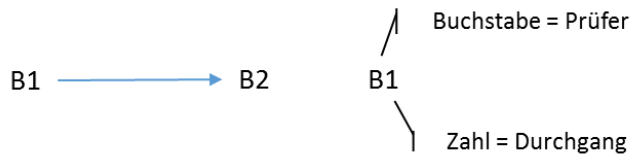
2.4.2 Wiederholbarkeit Prüfer B („Standard Calculation“)

Wir können die Wiederholbarkeit für den Prüfer B bestimmen, da im Datenbeispiel genau zwei Prüfdurchgänge ausgeführt wurden.

Werden in einem anderen Anwendungsfall mehr als zwei Prüfdurchgänge je Prüfer ausgeführt, so ist eine Berechnung der Wiederholbarkeit nicht mehr möglich.

2.4.2.1 Aufbau Datentabelle (B | „Standard Calculation“)

Zur Erinnerung: Das Berechnen der Wiederholbarkeit im Rahmen der Berechnungsmethode „Standard Calculation“ ist nur möglich, wenn *exakt zwei Prüfdurchgänge* vorliegen. Im Fallbeispiel haben wir exakt zwei Prüfdurchgänge je Prüfer und führen mit diesen beiden Prüfdurchgängen den Vergleich aus. Die Anzahl der Einheiten wird bei diesem Vergleich nicht virtuell erhöht ($N_o = 5$).



Mit diesem Vergleichs-Schema erzeugen wir die Zwei-Urteilsspalten-Datentabelle für die Analyse:

Tabelle 51: Datentabelle für die Wiederholbarkeit des Prüfers B nach der Berechnungsmethode „Standard Calculation“

Objekt-Nr.	Prüfer B	
	1. Instanz = Durchgang 1	2. Instanz = Durchgang 2
1	Not okay	Not okay
2	Okay	Not okay
3	Okay	Okay
4	Okay	Okay
5	Okay	Okay

2.4.2.2 Kreuztabellen (B | „Standard Calculation“)

Zählen wir die Häufigkeit der Urteils kombination in den Zeilen der Tabelle 51, so erhalten wir die **Kreuztabelle der absoluten Häufigkeit der Urteils kombinationen**:

Tabelle 52: Kreuztabelle der absoluten Häufigkeit der Urteils kombinationen für die Wiederholbarkeit des Prüfers B nach der Berechnungsmethode „Standard Calculation“

Wiederholbarkeit Prüfer B abs. Anzahl		Prüfer B 2. Durchgang		Summe
	Kategorie	Okay	Not okay	
Prüfer B 1. Durchgang	Okay	3	1	4
	Not okay	0	1	1
Summe		3	2	5

Teilen wir jeden Zahlenwert in der Tabelle 52 durch die Anzahl der Einheiten $N_o = 5$, so erhalten wir die **Kreuztabelle der Anteile beobachteter Urteilskombinationen**:

Tabelle 53: Kreuztabelle der Anteile beobachteter Urteilskombinationen für die Wiederholbarkeit des Prüfers B nach der Berechnungsmethode „Standard Calculation“

Wiederholbarkeit Prüfer B beob. Anteil		Prüfer B 2. Durchgang		Summe
	Kategorie	Okay	Not okay	
Prüfer B 1. Durchgang	Okay	0,60	0,20	0,80
	Not okay	0,00	0,20	0,20
Summe		0,60	0,40	1,00

Bilden wir die Summe mit den beiden grün hinterlegten Zellenwerten der Tabelle 53, erhalten wir den Anteil beobachteter gleicher Urteile.

$$p_o = 0,60 + 0,20 = 0,80$$

Multiplizieren wir die Zeilen- und Spaltensummen der Tabelle 53 miteinander, erhalten wir die **Kreuztabelle für den Anteil erwarteter Urteilskombinationen, die durch den Zufall entstanden sind**.

Tabelle 54: Kreuztabelle der Anteile zufällig erwarteter Urteilskombinationen für die Wiederholbarkeit des Prüfers B

Wiederholbarkeit Prüfer B erw. Anteil		Prüfer B 2. Durchgang		Summe
	Kategorie	Okay	Not okay	
Prüfer B 1. Durchgang	Okay	0,48	0,32	0,80
	Not okay	0,12	0,08	0,20
Summe		0,60	0,40	1,00

Bilden wir die Summe mit den Zahlenwerten in den beiden grün hinterlegten Tabellenzellen, so erhalten wir den Anteil erwarteter, gleicher Urteile.

$$p_e = 0,48 + 0,08 = 0,56$$

2.4.2.3 Kappa-Koeffizient (B | „Standard Calculation“)

Zunächst bilden wir den **Anteil beobachteter Übereinstimmungen, der vom Anteil zufällig erwarteter Übereinstimmungen bereinigt ist**:

$$p_o - p_e = 0,80 - 0,56 = 0,24$$

Anschließend bestimmen wir den **größtmöglichen Wert für den Anteil beobachteter Übereinstimmungen, der vom Anteil zufällig erwarteter Übereinstimmungen bereinigt ist**:

$$1 - p_e = 1 - 0,56 = 0,44$$

Teilen wir die beiden zuletzt berechneten Anteilswerte durcheinander, so erhalten wir den Kappa-Koeffizienten:

$$\kappa_{B(\text{Cohen}|\text{Standard Calculation})} = \frac{p_o - p_e}{1 - p_e} = \frac{0,24}{0,44} = 0,5454$$

2.4.2.4 Signifikanztest (B | „Standard Calculation“)

Zu prüfende Hypothese:

Nullhypothese H0: Der Kappa-Koeffizient der Grundgesamtheit ist gleich Null

Praktische Deutung der Nullhypothesen: Die übereinstimmenden Urteile sind vermutlich allein durch den Zufall entstanden.

Alternativhypothese H1: Der Kappa-Koeffizient der Grundgesamtheit ist ungleich Null.

Praktische Deutung der Alternativhypothese: Die übereinstimmenden Urteile sind nicht (allein) durch den Zufall zustande gekommen. Der Urteile folgen einer erkennbaren Systematik.

Für die Prüfung der Nullhypothese benötigen wir die Prüfgröße z, die wir aus der Division des Kappa-Koeffizienten durch die Standardabweichung des Kappa-Koeffizienten erhalten.

2.4.2.5 Standardabweichung (B | „Standard Calculation“)

Würden wir den Versuch mehrfach wiederholt ausführen, so würde der Wert des Kappa-Indexes von Versuch zu Versuch streuen. Eine näherungsweise Abschätzung der Zufallsstreuung erhalten wir durch das Bestimmen der Standardabweichung des Kappa-Koeffizienten.

Den Wert der **Standardabweichung des Kappa-Koeffizienten** berechnen wir mit der Berechnungsgleichung aus dem Abschnitt 2.1.4.

$$s_{\kappa_{B(\text{Cohen}|\text{Standard Calculation})}} = 0,398\ 344$$

2.4.2.6 Prüfgröße z (B | „Standard Calculation“)

Teilen wir den Kappa-Koeffizienten durch die Standardabweichung des Kappa-Koeffizienten, erhalten wir die Prüfgröße z:

$$z_{\kappa_{B(\text{Cohen}|\text{Standard Calculation})}} = \frac{\kappa_{B(\text{Cohen}|\text{Standard Calculation})}}{s_{\kappa_{B(\text{Cohen}|\text{Standard Calculation})}}} = \frac{0,545\ 454}{0,398\ 344} \approx 1,369$$

2.4.2.7 P-Wert (B | „Standard Calculation“)

Den P-Wert bestimmen wir gemäß der folgenden Beziehung:

$$P_{\kappa_B} = 1 - G\left(z_{\kappa_{B(\text{Cohen}|\text{Standard Calculation})}}\right) = 1 - G(1,369) = 0,085\ 5$$

Interpretation des Testergebnisses anhand des P-Wertes

Aufgrund des viel zu kleinen Stichprobenumfangs ist die Power des Tests (= Nullhypothese verwerfen, wenn diese nicht zutrifft) viel zu gering und daher das Testen an sich hier unsinnig. Dennoch vollziehen wir den formalen Akt des „Test-Rituals“ mit dem P-Wert.

Da der P-Wert größer ist als das gewählte Signifikanzniveau $\alpha = 5\%$, wird die Nullhypothese nicht verworfen (= beibehalten). Die in der Stichprobe beobachteten Übereinstimmungen sind vermutlich allein aufgrund des Zufalls entstanden.

2.4.3 Vergleichbarkeit Prüfer A vs. Referenz („Standard Calculation“)

Wir wollen wissen, wie gut die Urteile vom Prüfer A mit den Referenz-Urteilen übereingestimmt haben. Dazu entnehmen wir aus der Versuchsdaten-Tabelle die Urteile vom Prüfer A und auch die Referenz-Urteile.

2.4.3.1 Virtuelle Einheiten-Gesamtzahl ($A \times \text{Ref.}$ | „Standard Calculation“)

Wie im Abschnitt 2.1.1 erwähnt, ist der Kappa-Koeffizient nach Jacob Cohen nur mit genau zwei Urteilsspalten möglich. Daher transformieren wir die mehrspaltigen Versuchsergebnisse in eine Tabelle mit zwei Urteilsspalten:

Um die zwei Spalten zu erhalten, kombinieren wir die Urteile aus jedem Prüfdurchgang mit den Referenzurteilen.

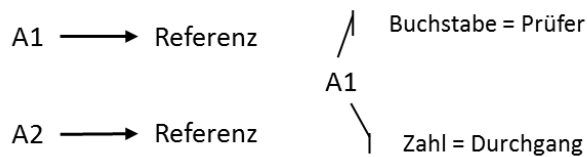


Abbildung 13: Datenaufbau-Schema für den Vergleich Prüfer A vs. Referenz (Cohen | Standard Calculation)

Dadurch erhalten wir die folgende virtuelle Anzahl Einheiten:

$$N_{vo} = N_t \times N_o = 2 \times 5 = 10$$

2.4.3.2 Aufbau Datentabelle (A × Ref. | „Standard Calculation“)

Tabelle 55 Datentabelle für die Vergleichbarkeit Prüfer A vs. Referenz nach der Berechnungsmethode „Standard Calculation“

Virtuelle Einheiten-Nr.	Ursprüngliche Einheiten-Nr.	Prüfer A	Referenz
		Alle Urteile	
1	1. Einheit 1. Durchgang	Okay	Okay
2	2. Einheit 1. Durchgang	Okay	Not okay
3	3. Einheit 1. Durchgang	Not okay	Okay
4	4. Einheit 1. Durchgang	Okay	Not okay
5	5. Einheit 1. Durchgang	Okay	Okay
6	1. Einheit 2. Durchgang	Okay	Okay
7	2. Einheit 2. Durchgang	Not okay	Not okay
8	3. Einheit 2. Durchgang	Okay	Okay
9	4. Einheit 2. Durchgang	Okay	Not okay
10	5. Einheit 2. Durchgang	Okay	Okay

2.4.3.3 Kreuztabellen (A × Ref. | „Standard Calculation“)

Zählen wir die Häufigkeit der Urteils kombinationen in den Zeilen der Tabelle 55, so erhalten wir die **Kreuztabelle der absoluten Häufigkeit der beobachteten Urteils kombinationen**:

Tabelle 56: Kreuztabelle der beobachteten absoluten Häufigkeiten für den Vergleich Prüfer A vs. Referenz nach der Berechnungsmethode „Standard Calculation“

Wiederholbarkeit Prüfer A vs. Ref. abs. Anzahl		Referenz		Summe
	Kategorie	Okay	Not okay	
Prüfer A alle Urteile	Okay	5	3	8
	Not okay	1	1	2
Summe		6	4	10

Teilen wir jeden Zahlenwert in der Tabelle 56 durch die virtuelle Einheiten-Gesamtzahl $N_{vo} = 10$, so erhalten wir die **Kreuztabelle der beobachteten Anteile der Urteils kombinationen**:

Tabelle 57: Kreuztabelle der beobachteten Anteile gleicher Urteils kombinationen für den Vergleich Prüfer A vs. Referenz nach der Berechnungsmethode „Standard Calculation“

Wiederholbarkeit Prüfer A vs. Ref. beob. Anteil		Referenz		Summe
	Kategorie	Okay	Not okay	
Prüfer A alle Urteile	Okay	0,50	0,30	0,80
	Not okay	0,10	0,10	0,20
Summe		0,60	0,40	1,00

Bilden wir die Summe mit den grün hinterlegten Zellenwerten, so erhalten wir den **Anteil beobachteter Übereinstimmungen**:

$$p_o = 0,50 + 0,10 = 0,60$$

Multiplizieren wir die Zeilen- mit den Spaltensummen der Tabelle 57, so erhalten wir die **Kreuztabelle der durch den Zufall erwarteten Anteile Urteils kombinationen**:

Tabelle 58: Kreuztabelle der durch den Zufall erwarteten Anteile an Urteils kombinationen für den Vergleich Prüfer A vs. Referenz nach der Berechnungsmethode „Standard Calculation“

Wiederholbarkeit Prüfer A vs. Ref. erw. Anteil		Referenz		Summe
	Kategorie	Okay	Not okay	
Prüfer A alle Urteile	Okay	0,48	0,32	0,80
	Not okay	0,12	0,08	0,20
Summe		0,60	0,40	1,00

Bilden wir die Summe mit den Werten in den grün hinterlegten Zellen der Tabelle 58, so erhalten wir den **erwarteten Anteil zufälliger Urteils übereinstimmungen**:

$$p_e = 0,48 + 0,08 = 0,56$$

2.4.3.4 Kappa-Koeffizient (A x Ref. | „Standard Calculation“)

Zunächst bilden wir den **Anteil beobachteter Übereinstimmungen, der vom Anteil zufällig erwarteter Übereinstimmungen bereinigt ist**:

$$p_o - p_e = 0,60 - 0,56 = 0,04$$

Anschließend bestimmen wir den **größtmöglichen Wert für den Anteil beobachteter Übereinstimmungen, der vom Anteil zufällig erwarteter Übereinstimmungen bereinigt ist**:

$$1 - p_e = 1 - 0,56 = 0,44$$

Teilen wir die beiden zuletzt berechneten Anteilswerte durcheinander, so erhalten wir den Kappa-Koeffizienten:

$$\kappa_{A \times Ref(Cohen | Standard Calculation)} = \frac{p_o - p_e}{1 - p_e} = \frac{0,04}{0,44} = 0,0909$$

2.4.3.5 Signifikanztest Kappa-Koeffizient (A x Ref. | „Standard Calculation“)

Zu prüfende Hypothese:

Nullhypothese H0: Der Kappa-Koeffizient der Grundgesamtheit ist gleich Null

Praktische Deutung der Nullhypothesen: Die übereinstimmenden Urteile sind vermutlich allein durch den Zufall entstanden.

Alternativhypothese H1: Der Kappa-Koeffizient der Grundgesamtheit ist ungleich Null.

Praktische Deutung der Alternativhypothese: Die übereinstimmenden Urteile sind nicht (allein) durch den Zufall zustande gekommen. Der Urteile folgen einer erkennbaren Systematik.

Für die Prüfung der Nullhypothese benötigen wir die Prüfgröße z, die wir aus der Division des Kappa-Koeffizienten durch die Standardabweichung des Kappa-Koeffizienten erhalten.

2.4.3.6 Standardabweichung (A x Ref. | „Standard Calculation“)

Würden wir den Versuch mehrfach wiederholt ausführen, so würde der Wert des Kappa-Indexes von Versuch zu Versuch streuen. Eine näherungsweise Abschätzung der Zufallsstreuung erhalten wir durch das Bestimmen der Standardabweichung des Kappa-Koeffizienten.

Den Wert der **Standardabweichung des Kappa-Koeffizienten** berechnen wir mit der Berechnungsgleichung aus dem Abschnitt 2.1.4.

$$s_{\kappa_{A \times Ref(Cohen | Standard Calculation)}} = 0,281672$$

2.4.3.7 Prüfgröße z (A x Ref. | „Standard Calculation“)

Teilen wir den Kappa-Koeffizienten durch die Standardabweichung des Kappa-Koeffizienten, erhalten wir die Prüfgröße z:

$$z_{\kappa_{A \times Ref(Cohen | Standard Calculation)}} = \frac{\kappa_{A \times Ref(Cohen | Standard Calculation)}}{s_{\kappa_{A \times Ref(Cohen | Standard Calculation)}}} = \frac{0,0909}{0,281672} \approx 0,323$$

2.4.3.8 P-Wert (A x Ref. | „Standard Calculation“)

Den P-Wert bestimmen wir gemäß der folgenden Beziehung:

$$P_{\kappa_{A \times Ref(Cohen | Standard Calculation)}} = 1 - G\left(z_{\kappa_{A \times Ref(Cohen | Standard Calculation)}}\right) = 1 - G(0,323) = 0,37$$

Interpretation des Testergebnisses anhand des P-Wertes

Aufgrund des viel zu kleinen Stichprobenumfangs ist die Power des Tests (= Nullhypothese verwerfen, wenn diese nicht zutrifft) viel zu gering und daher das Testen an sich hier unsinnig. Dennoch vollziehen wir den formalen Akt des „Test-Rituals“ mit dem P-Wert.

Da der P-Wert größer ist als das gewählte Signifikanzniveau $\alpha = 5\%$, wird die Nullhypothese nicht verworfen (= beibehalten). Die in der Stichprobe beobachteten Übereinstimmungen sind vermutlich allein aufgrund des Zufalls entstanden.

2.4.4 Vergleichbarkeit Prüfer B vs. Referenz („Standard Calculation“)

Mit diesem Vergleich wollen wir das Ausmaß der Urteilsübereinstimmung zwischen dem Prüfer B und der Referenz beurteilen.

2.4.4.1 Virtuelle Einheiten-Gesamtanzahl ($B \times \text{Ref.}$ | „Standard Calculation“)

Wie im Abschnitt 2.1.1 erwähnt, ist der Kappa-Koeffizient nach Jacob Cohen nur mit genau zwei Urteilsspalten möglich. Daher kombinieren wir die Ergebnisse aus jedem der Prüfdurchgänge mit den Referenzwerten gemäß dem folgenden Schema:

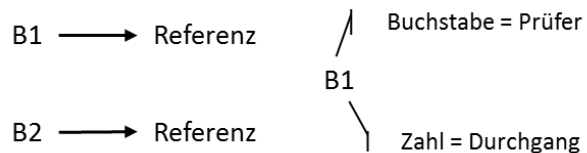


Abbildung 14: Datenaufbau-Schema für die Vergleichbarkeit Prüfer $B \times \text{Referenz}$

Für die Daten unsers Fallbeispiels entsteht dadurch die virtuelle Anzahl Einheiten:

$$N_{vo} = N_t \times N_o = 2 \times 5 = 10$$

2.4.4.2 Aufbau Datentabelle (B x Ref. | „Standard Calculation“)

Wir entnehmen aus der Tabelle 2 auf der Seite 10 die Urteile des Prüfers B und tragen diese entsprechend dem zuvor gezeigten Schema ein.

Tabelle 59: Datentabelle für die Vergleichbarkeit Prüfer B vs. Referenz nach der Berechnungsmethode „Standard Calculation“

Virtuelle Einheiten-Nr.	Ursprüngliche Einheiten-Nr.	Prüfer B	Referenz
		Alle Urteile	
1	1. Einheit 1. Durchgang	Not okay	Okay
2	2. Einheit 1. Durchgang	Okay	Not okay
3	3. Einheit 1. Durchgang	Okay	Okay
4	4. Einheit 1. Durchgang	Okay	Not okay
5	5. Einheit 1. Durchgang	Okay	Okay
6	1. Einheit 2. Durchgang	Not okay	Okay
7	2. Einheit 2. Durchgang	Not okay	Not okay
8	3. Einheit 2. Durchgang	Okay	Okay
9	4. Einheit 2. Durchgang	Okay	Not okay
10	5. Einheit 2. Durchgang	Okay	Okay

2.4.4.3 Kreuztabellen (B × Ref. | „Standard Calculation“)

Zählen wir die Häufigkeit der Urteils kombinationen in den Zeilen der Tabelle 59, so erhalten wir die folgende Kreuztabelle:

Tabelle 60: Kreuztabelle der absoluten Häufigkeit der Urteils kombinationen für die Vergleichbarkeit Prüfer B vs. Referenz nach der Berechnungsmethode „Standard Calculation“

Wiederholbarkeit Prüfer B vs. Ref. abs. Anzahl		Referenz		Summe
	Kategorie	Okay	Not okay	
Prüfer B alle Urteile	Okay	4	3	7
	Not okay	2	1	3
Summe		6	4	10

Teilen wir jeden Zahlenwert in der Tabelle 60 durch die virtuelle Gesamtzahl der Einheiten $N_{v0} = 10$, so erhalten wir die **Kreuztabelle der beobachteten Anteile der Urteils kombinationen**.

Tabelle 61: Kreuztabelle der beobachteten Anteile der Urteils kombinationen für die Vergleichbarkeit Prüfer B vs. Referenz nach der Berechnungsmethode „Standard Calculation“

Wiederholbarkeit Prüfer B vs. Ref. beob. Anteil		Referenz		Summe
	Kategorie	Okay	Not okay	
Prüfer B alle Urteile	Okay	0,40	0,30	0,70
	Not okay	0,20	0,10	0,30
Summe		0,60	0,40	1,00

Bilden wir die Summe mit den Werten in den grün hinterlegten Zellen der Tabelle 61 und erhalten so den beobachteten Anteil übereinstimmender Urteile:

$$p_o = 0,40 + 0,10 = 0,50$$

Multiplizieren wir die Zeilen- und Spaltensumme der Tabelle 61 miteinander, erhalten wir die **Kreuztabelle der durch den Zufall erwarteten Anteile an Urteils kombinationen**:

Tabelle 62: Kreuztabelle der durch den Zufall erwarteten Urteils kombinationen für die Vergleichbarkeit Prüfer B vs. Referenz nach der Berechnungsmethode „Standard Calculation“

Wiederholbarkeit Prüfer B vs. Ref. erw. Anteil		Referenz		Summe
	Kategorie	Okay	Not okay	
Prüfer B alle Urteile	Okay	0,42	0,28	0,70
	Not okay	0,18	0,12	0,30
Summe		0,60	0,40	1,00

Bilden wir die Summe mit den Werten in den grün hinterlegten Zellen der Tabelle 61, erhalten wir den zufallsbedingt erwarteten Anteil übereinstimmender Urteile:

$$p_e = 0,42 + 0,12 = 0,54$$

2.4.4.4 Kappa-Koeffizient (B x Ref. | „Standard Calculation“)

Zunächst bilden wir den **Anteil beobachteter Übereinstimmungen, der vom Anteil zufällig erwarteter Übereinstimmungen bereinigt ist**:

$$p_o - p_e = 0,50 - 0,54 = -0,04$$

Anschließend bestimmen wir den **größtmöglichen Wert für den Anteil beobachteter Übereinstimmungen, der vom Anteil zufällig erwarteter Übereinstimmungen bereinigt ist**:

$$1 - p_e = 1 - 0,54 = 0,46$$

Teilen wir die beiden zuletzt berechneten Anteilswerte durcheinander, so erhalten wir den Kappa-Koeffizienten:

$$\kappa_{B \times Ref(Cohen | Standard Calculation)} = \frac{p_o - p_e}{1 - p_e} = \frac{-0,04}{0,46} = -0,086\,565\,2$$

2.4.4.5 Signifikanztest (B x Ref. | „Standard Calculation“)

Zu prüfende Hypothese:

Nullhypothese H0: Der Kappa-Koeffizient der Grundgesamtheit ist gleich Null

Praktische Deutung der Nullhypothesen: Die übereinstimmenden Urteile sind vermutlich allein durch den Zufall entstanden.

Alternativhypothese H1: Der Kappa-Koeffizient der Grundgesamtheit ist ungleich Null.

Praktische Deutung der Alternativhypothese: Die übereinstimmenden Urteile sind nicht (allein) durch den Zufall zustande gekommen. Der Urteile folgen einer erkennbaren Systematik.

Für die Prüfung der Nullhypothese benötigen wir die Prüfgröße z, die wir aus der Division des Kappa-Koeffizienten durch die Standardabweichung des Kappa-Koeffizienten erhalten.

2.4.4.6 Standardabweichung (B x Ref. | „Standard Calculation“)

Würden wir den Versuchs mehrfach wiederholt ausführen, so würde der Wert des Kappa-Indexes von Versuchs zu Versuch streuen. Eine näherungsweise Abschätzung der Zufallsstreuung erhalten wir durch das Bestimmen der Standardabweichung des Kappa-Koeffizienten.

Den Wert der **Standardabweichung des Kappa-Koeffizienten** berechnen wir mit der Berechnungsgleichung aus dem Abschnitt 2.1.4.

$$s_{\kappa_{B \times Ref(Cohen | Standard Calculation)}} = 0,308\,665$$

2.4.4.7 Prüfgröße z (B x Ref. | „Standard Calculation“)

Teilen wir den Kappa-Koeffizienten durch die Standardabweichung des Kappa-Koeffizienten, erhalten wir die Prüfgröße z:

$$z_{\kappa_{B \times Ref(Cohen | Standard Calculation)}} = \frac{\kappa_{B \times Ref(Cohen | Standard Calculation)}}{s_{\kappa_{B \times Ref(Cohen | Standard Calculation)}}} = \frac{-0,086\,565\,2}{0,308\,665} \approx -0,281\,718$$

2.4.4.8 P-Wert (B x Ref. | „Standard Calculation“)

Den P-Wert bestimmen wir gemäß der folgenden Beziehung:

$$P_{\kappa_{B \times Ref(Cohen | Standard Calculation)}} = 1 - G\left(z_{\kappa_{B \times Ref(Cohen | Standard Calculation)}}\right) = 1 - G(-0,281) \approx 0,61$$

Interpretation des Testergebnisses anhand des P-Wertes

Aufgrund des viel zu kleinen Stichprobenumfangs ist die Power des Tests (= Nullhypothese verwerfen, wenn diese nicht zutrifft) viel zu gering und daher das Testen an sich hier unsinnig. Dennoch vollziehen wir den formalen Akt des „Test-Rituals“ mit dem P-Wert.

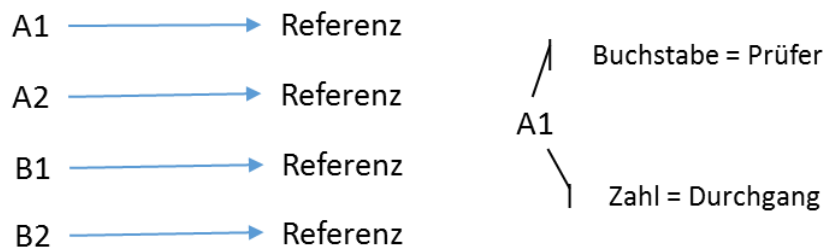
Da der P-Wert größer ist als das gewählte Signifikanzniveau $\alpha = 5 \%$, wird die Nullhypothese nicht verworfen (= beibehalten). Die in der Stichprobe beobachteten Übereinstimmungen sind vermutlich allein aufgrund des Zufalls entstanden.

2.4.5 Vergleichbarkeit aller Prüfer vs. Referenz („Standard Calculation“)

Mit diesem Vergleich wollen wir das Ausmaß der Urteilübereinstimmung zwischen allen Prüfern mit der Referenz feststellen.

2.4.5.1 Virtuelle Einheiten-Gesamtzahl ($A \times B \times \text{Ref.}$ | „Standard Calculation“)

Wie im Abschnitt 2.1.1 erwähnt, ist der Kappa-Koeffizient nach Jacob Cohen nur mit genau zwei Urteilsspalten möglich. Daher kombinieren wir die Ergebnisse aus den einzelnen Prüfdurchgängen mit den Referenzwerten gemäß dem folgenden Schema:



Daraus ergibt sich die folgende virtuelle Anzahl Einheiten:

$$N_{vo} = N_a \times N_t \times N_o = 2 \times 2 \times 5 = 20$$

2.4.5.2 Aufbau Datentabelle (A × B × Ref. | „Standard Calculation“)

Wir entnehmen aus der Tabelle 2 auf der Seite 10 die Urteile der Prüfer A und B und auch die Referenzurteile. Alle Prüferurteile schreiben wir untereinander in eine Spalte und die Referenzurteile in die zweite Spalte (4 ×).

Tabelle 63: Datentabelle der Urteils kombinationen für den Vergleich aller Prüfer vs. Referenz nach der Berechnungsmethode „Standard Calculation“

Virtuelle Einheiten-Nr.	Ursprüngliche Einheiten-Nr.	Alle Prüfer	Referenz
1	1. Einheit 1. Durchgang A	Okay	Okay
2	2. Einheit 1. Durchgang A	Okay	Not okay
3	3. Einheit 1. Durchgang A	Not okay	Okay
4	4. Einheit 1. Durchgang A	Okay	Not okay
5	5. Einheit 1. Durchgang A	Okay	Okay
6	1. Einheit 2. Durchgang A	Okay	Okay
7	2. Einheit 2. Durchgang A	Not okay	Not okay
8	3. Einheit 2. Durchgang A	Okay	Okay
9	4. Einheit 2. Durchgang A	Okay	Not okay
10	5. Einheit 2. Durchgang A	Okay	Okay
11	1. Einheit 1. Durchgang B	Not okay	Okay
12	2. Einheit 1. Durchgang B	Okay	Not okay
13	3. Einheit 1. Durchgang B	Okay	Okay
14	4. Einheit 1. Durchgang B	Okay	Not okay
15	5. Einheit 1. Durchgang B	Okay	Okay
16	1. Einheit 2. Durchgang B	Not okay	Okay

17	2. Einheit 2. Durchgang B	Not okay	Not okay
18	3. Einheit 2. Durchgang B	Okay	Okay
19	4. Einheit 2. Durchgang B	Okay	Not okay
20	5. Einheit 2. Durchgang B	Okay	Okay

2.4.5.3 Kreuztabellen (A × B × Ref. | „Standard Calculation“)

Zählen wir die Häufigkeit der Urteils kombinationen in den Zeilen der Tabelle 55, so erhalten wir die **Kreuztabelle der absoluten Häufigkeit der beobachteten Urteils kombinationen**:

Tabelle 64: Kreuztabelle der beobachteten Häufigkeit der Urteils kombinationen für den Vergleich aller Prüfer vs. Referenz nach der Berechnungsmethode „Standard Calculation“

Wiederholbarkeit Prüfer vs. Ref. abs. Anzahl		Referenz		Summe
	Kategorie	Okay	Not okay	
Alle Prüfer	Okay	9	6	15
	Not okay	3	2	5
Summe		12	8	20

Teilen wir jeden Zahlenwert in der Tabelle 65 durch die virtuelle Einheiten-Gesamtzahl, so erhalten wir die folgende Kreuztabelle der beobachteten Anteile der Urteils kombinationen:

Hinweis: Dieselbe Kreuztabelle erhalten wir, wenn wir die Zellenwerte aus den beiden Kreuztabellen der beobachteten Häufigkeiten für die Vergleichbarkeit Prüfer A vs. Referenz (siehe Abschnitt 2.4.3.3) und Prüfer B vs. Referenz (siehe Abschnitt 2.4.4.3) aufsummieren.

Tabelle 65: Kreuztabelle der beobachteten Anteile der Urteils kombinationen für den Vergleich aller Prüfer vs. Referenz nach der Berechnungsmethode „Standard Calculation“

Wiederholbarkeit Prüfer vs. Ref. beob. Anteil		Referenz		Summe
	Kategorie	Okay	Not okay	
Alle Prüfer	Okay	0,45	0,30	0,75
	Not okay	0,15	0,10	0,25
Summe		0,60	0,40	1,00

Bilden wir die Summe mit den Werten in den grün hinterlegten Zellen der Tabelle 65, so erhalten wir den Anteil beobachteter Urteils-Übereinstimmungen:

$$p_o = 0,45 + 0,10 = 0,55$$

Multiplizieren wir die Zeilen- mit den Spaltensummen der Tabelle 65, so erhalten wir die **Kreuztabelle der durch den Zufall erwarteten Urteils kombinationen**:

Tabelle 66: Kreuztabelle der durch den Zufall erwarteten Urteils kombinationen für den Vergleich aller Prüfer vs. Referenz nach der Berechnungsmethode „Standard Calculation“

Wiederholbarkeit Prüfer vs. Ref. erw. Anteil		Referenz		Summe
	Kategorie	Okay	Not okay	
Alle Prüfer	Okay	0,45	0,30	0,75
	Not okay	0,15	0,10	0,25
Summe		0,60	0,40	1,00

Bilden wir die Summe mit den grün hinterlegten Werten in der Tabelle 66, so erhalten wir den erwarteten Anteil zufälliger Urteilsübereinstimmungen:

$$p_e = 0,45 + 0,10 = 0,55$$

2.4.5.4 Kappa-Koeffizient (A x B x Ref. | „Standard Calculation“)

Zunächst bilden wir den **Anteil beobachteter Übereinstimmungen, der vom Anteil zufällig erwarteter Übereinstimmungen bereinigt ist**:

$$p_o - p_e = 0,55 - 0,55 = 0$$

Anschließend bestimmen wir den **größtmöglichen Wert für den Anteil beobachteter Übereinstimmungen, der vom Anteil zufällig erwarteter Übereinstimmungen bereinigt ist**:

$$1 - p_e = 1 - 0,55 = 0,45$$

Teilen wir die beiden zuletzt berechneten Anteilswerte durcheinander, so erhalten wir den Kappa-Koeffizienten:

$$\kappa_{A \times B \times Ref(Cohen | Standard Calculation)} = \frac{p_o - p_e}{1 - p_e} = \frac{0,00}{0,45} = 0,00$$

2.4.5.5 Signifikanztest (A × B × Ref. | „Standard Calculation“)

Zu prüfende Hypothese:

Nullhypothese H0: Der Kappa-Koeffizient der Grundgesamtheit ist gleich Null

Praktische Deutung der Nullhypothese: Die übereinstimmenden Urteile sind vermutlich allein durch den Zufall entstanden.

Alternativhypothese H1: Der Kappa-Koeffizient der Grundgesamtheit ist ungleich Null.

Praktische Deutung der Alternativhypothese: Die übereinstimmenden Urteile sind nicht (allein) durch den Zufall zustande gekommen. Der Urteile folgen einer erkennbaren Systematik.

Für die Prüfung der Nullhypothese benötigen wir die Prüfgröße z, die wir aus der Division des Kappa-Koeffizienten durch die Standardabweichung des Kappa-Koeffizienten erhalten.

2.4.5.6 Standardabweichung (A × B × Ref. | „Standard Calculation“)

Würden wir den Versuch mehrfach wiederholt ausführen, so würde der Wert des Kappa-Indexes von Versuch zu Versuch streuen. Eine näherungsweise Abschätzung der Zufallsstreuung erhalten wir durch das Bestimmen der Standardabweichung des Kappa-Koeffizienten.

Den Wert der **Standardabweichung des Kappa-Koeffizienten** berechnen wir mit der Berechnungsgleichung aus dem Abschnitt 2.1.4.

$$S_{K_{B \times Ref}(Cohen|Standard\ Calculation)} = 0,210\ 819$$

2.4.5.7 Prüfgröße z (A × B × Ref. | „Standard Calculation“)

Teilen wir den Kappa-Koeffizienten durch die Standardabweichung des Kappa-Koeffizienten, erhalten wir die Prüfgröße z:

$$z_{K_{A \times B \times Ref}(Cohen|Standard\ Calculation)} = \frac{K_{A \times B \times Ref}(Cohen|Standard\ Calculation)}{S_{K_{A \times B \times Ref}(Cohen|Standard\ Calculation)}} = \frac{0}{0,210\ 819} \approx 0,00$$

2.4.5.8 P-Wert (A × B × Ref. | „Standard Calculation“)

Den P-Wert bestimmen wir gemäß der folgenden Beziehung:

$$P_{K_{A \times B \times Ref}(Cohen|Standard\ Calculation)} = 1 - G\left(z_{K_{A \times B \times Ref}(Cohen|Standard\ Calculation)}\right) = 1 - G(0,00) \approx 0,50$$

Interpretation des Testergebnisses anhand des P-Wertes

Da der P-Wert größer ist als das gewählte Signifikanzniveau $\alpha = 5\%$, wird die Nullhypothese nicht verworfen (= beibehalten). Die in der Stichprobe beobachteten Übereinstimmungen sind vermutlich allein aufgrund des Zufalls entstanden.

2.5 Kappa nach Fleiss

Der Kappa-Koeffizient nach Joseph L. Fleiss ist von seinem Schöpfer für den Vergleich der Urteilswerte bei zwei oder mehr Urteilsinstanzen gedacht. Es besteht also nicht das Problem, dass nur zwei Urteilsspalten miteinander verglichen werden können.

Alle hier beschriebenen Berechnungsverfahren für den Kappa-Koeffizienten nach *Joseph L. Fleiss* orientieren sich an den Darstellungen in dem Bosch-Heft 10 aus dem Jahr 2010 und an der zugehörigen Microsoft Excel Datei „Übereinstimmungsanalyse.xls“ [H10].

2.5.1 Schrittschema der Kappa-Bestimmung:

Das Berechnungsschema für den Kappa-Koeffizienten nach *Joseph L. Fleiss* umfasst acht Berechnungsschritte, die hier am Beispiel mit zwei Urteilskategorien dargestellt sind.

2.5.1.1 Aufbauschema Datentabelle

Objekt-Nr.	Urteile der urteilenden Instanz 1	...	Urteile der urteilenden Instanz m	Häufigkeit der Urteile in der Kategorie 1	Häufigkeit der Urteile in der Kategorie 2
1	Urteil	...	Urteil	$n_{1.1}$	$n_{1.2}$
\vdots	\vdots	\vdots	Urteil	\vdots	\vdots
N_o	Urteil	...	Urteil	$n_{N_o.1}$	$n_{N_o.2}$

$n_{i,j}$ = Anzahl beobachteter Urteile der Kategorie j für das Objekt Nr. i .

m = Anzahl der zu vergleichenden Urteilsspalten.

2.5.1.2 Anteil der beobachteten Übereinstimmungen

Der Anteil beobachteter Übereinstimmungen wird für jede untersuchte Einheit bestimmt. Mit Blick auf die folgende Tabelle bestimmen wir die Übereinstimmungen also zeilenweise.

Objekt-Nr.	Häufigkeit der Urteile in der Kategorie 1	Häufigkeit der Urteile in der Kategorie 2	Beobachtete Anteile übereinstimmender Urteile für jedes Objekt
1	$n_{1.1}$	$n_{1.2}$	$p_1 = \frac{n_{1.1}(n_{1.1} - 1) + n_{1.2}(n_{1.2} - 1)}{m(m - 1)}$
\vdots	\vdots	\vdots	\vdots
N_o	$n_{N_o.1}$	$n_{N_o.2}$	$p_{N_o} = \frac{n_{N_o.1}(n_{N_o.1} - 1) + n_{N_o.2}(n_{N_o.2} - 1)}{m(m - 1)}$
Summen:	$n_{.1} = \sum_{i=1}^{N_o} n_{i.1}$	$n_{.2} = \sum_{i=1}^{N_o} n_{i.2}$	$\sum_{i=1}^{N_o} p_i$

Der uns eigentlich interessierende **Anteil beobachteter Übereinstimmungen** ergibt sich als Mittelwert der pro Einheit ermittelten Anteile:

$$p_o = \frac{1}{N_o} \sum_{i=1}^{N_o} p_i$$

2.5.1.3 Erwarteter Anteil Übereinstimmungen

Hier ist die Variante für genau zwei Urteilskategorien dargestellt. Existieren mehr als zwei Urteilskategorien, sind entsprechend mehr Anteilswerte zu ermitteln.

Hinweis: m ist in den folgenden Formeln die Anzahl der miteinander zu vergleichenden Instanzen (Anzahl Urteils-Spalten), so dass $m \cdot N_o$ die Anzahl der insgesamt vergebenen Urteile ergibt.

Mittlerer Anteil der vergebenen Urteile in der Kategorie $j=1$:

$$\bar{p}_{.1} = \frac{1}{m \cdot N_o} \sum_{i=1}^{N_o} n_{i.1}$$

Mittlerer Anteil der vergebenen Urteile in der Kategorie $j=2$:

$$\bar{p}_{.2} = \frac{1}{m \cdot N_o} \sum_{i=1}^{N_o} n_{i.2}$$

Durch den Zufall erwarteter Anteil übereinstimmender Urteile:

$$p_e = \bar{p}_{.1}^2 + \bar{p}_{.2}^2$$

2.5.1.4 Vom Zufall bereinigte Anteile beobachteter Übereinstimmungen

Wir bestimmen den Anteil beobachteter Übereinstimmungen, der Anteil zufällig erwarteter Übereinstimmungen bereinigt ist:

$$p_o - p_e$$

Größtmöglicher Wert für den Anteil Übereinstimmungen, der vom Anteil zufälliger Übereinstimmungen bereinigt ist:

$$1 - p_e$$

2.5.1.5 Kappa-Koeffizient nach Fleiss

Der Kappa-Koeffizient nach Fleiss ist analog aufgebaut wie der Kappa-Koeffizient nach Cohen. Also auch hier erhalten wir den Koeffizienten aus dem Verhältnis der beiden vom Zufall bereinigten Anteile:

$$\kappa_{\text{Fleiss}} = \frac{p_o - p_e}{1 - p_e}$$

Besonderheit bei einem Vergleich mit der Referenz

Bei einem Referenz-Vergleich wird jede zwei-elementige Paarung (Jede Urteilsspalte gepaart mit der Referenz-Spalte) gebildet und der Kappa-Wert berechnet. Betrachten wir zum Beispiel den Vergleich der Urteile des Prüfers A mit Referenz: Hat der Prüfer A drei Prüfdurchgänge ausgeführt so wird ausgewertet

- Kappa Berechnung für den Vergleich Durchgang A1 mit Referenz
- Kappa-Berechnung für den Vergleich Durchgang A2 mit Referenz
- Kappa-Berechnung für den Vergleich Durchgang A3 mit Referenz

Der Mittelwert aus diesen drei Kappa-Werten ist Kappa für den Vergleich Prüfer A mit der Referenz.

2.5.1.6 Standardabweichung

Die hier angegebene Berechnungsgrundlage entstammt der Microsoft Excel-Datei „Übereinstimmungsanalyse.xls“ [Quelle: Fa. Bosch].

Darin werden zunächst die beiden Hilfsgrößen h1 und h2 gebildet:

$$h1 = \sum_{j=1}^{N_c} [\bar{p}_j (1 - \bar{p}_j)]$$

$$h2 = \sum_{j=1}^{N_c} [\bar{p}_j (1 - \bar{p}_j) (1 - 2\bar{p}_j)]$$

Mit diesen Hilfsgrößen wird die Standardabweichung bestimmt:

$$se(\kappa) \approx \sqrt{\frac{2(h1^2 - h2)}{m(m-1)N_o h1^2}}$$

m = Anzahl der zu vergleichenden Urteilsinstanzen (Spalten)

Bei einem Vergleich mit der Referenz wird zunächst die Varianz für jede zwei-elementige Paarung (jede einzelne Urteilsspalte mit der Referenz) gebildet.

Betrachten wir zum Beispiel den Vergleich der Urteile des Prüfers A mit Referenz: Hat der Prüfer A drei Prüfdurchgänge ausgeführt so wird ausgewertet

- Varianz Berechnung für den Vergleich Durchgang A1 mit Referenz $var_{\kappa_{A1 \times Ref}}$
- Varianz-Berechnung für den Vergleich Durchgang A2 mit Referenz $var_{\kappa_{A2 \times Ref}}$
- Varianz-Berechnung für den Vergleich Durchgang A3 mit Referenz $var_{\kappa_{A3 \times Ref}}$

Im Beispiel wurden insgesamt $comp = 3$ paarweise Vergleiche ausgeführt. Für das Bestimmen der Varianz Prüfer A vs. Referenz wird folgende Beziehung genutzt.

$$var_{\kappa_{A \times Ref}} = \frac{var_{\kappa_{A1 \times Ref}} + var_{\kappa_{A2 \times Ref}} + var_{\kappa_{A3 \times Ref}}}{comp^2}$$

2.5.1.7 Prüfgröße z

Die Prüfgröße z bestimmen wir, indem wir den Kappa-Koeffizienten durch die Standardabweichung teilen:

$$z_{\kappa_{Fleiss}} = \frac{\kappa_{Fleiss}}{se_{\kappa_{Fleiss}}}$$

2.5.1.8 P-Wert

Mit der Verteilungsfunktion der Standardnormalverteilung berechnen wir den P-Wert:

$$P = 1 - G(z_{\kappa_{Fleiss}})$$

2.5.2 Wiederholbarkeit Prüfer A

Mit der Wiederholbarkeit wollen wir den Grad der Übereinstimmung der Urteile in allen Durchläufen des Prüfers A allein feststellen.

2.5.2.1 Aufbau Datentabelle (A | Fleiss)

Aus der **Tabelle 2** auf der Seite 10 entnehmen wir nur die Urteile des Prüfers A. Anschließend zählen wir zeilenweise, wie oft jede Urteilkategorie vergeben worden ist.

Tabelle 67: Absolute Häufigkeit der beobachteten Urteile je Einheit für die Wiederholbarkeit Prüfer A

Objekt-Nr. i	Prüfer A		Anzahl der Urteile $n_{i,1}$ in der ersten Kategorie „Okay“	Anzahl der Urteile $n_{i,2}$ in der zweiten Kategorie „Not okay“
	Durchgang 1	Durchgang 2		
1	Okay	Okay	$n_{1,1} = 2$	$n_{1,2} = 0$
2	Okay	Not okay	$n_{2,1} = 1$	$n_{2,2} = 1$
3	Not okay	Okay	$n_{3,1} = 1$	$n_{3,2} = 1$
4	Okay	Okay	$n_{4,1} = 2$	$n_{4,2} = 0$
5	Okay	Okay	$n_{5,1} = 2$	$n_{5,2} = 0$
Spaltensumme			$n_{\cdot 1} = 8$	$n_{\cdot 2} = 2$

Die hier berechneten Spaltensummen benötigen wir später für das Bestimmen des durch den Zufall erwarteten Anteils übereinstimmender Urteile.

2.5.2.2 Anteil beobachteter Übereinstimmungen (A | Fleiss)

Wir klären die Frage, wie viele Urteile in einer einzelnen Zeile maximal übereinstimmen können. Wir haben für die Wiederholbarkeit des Prüfers A genau $m = N_t = 2$ Urteilsspalten zu vergleichen.

Betrachten wir den folgenden Bruch, so steht im Nenner die größtmögliche Anzahl an Übereinstimmungen $m \cdot (m - 1)$ und im Zähler befindet sich die tatsächlich beobachtete Anzahl an Übereinstimmungen $n_{ij}(n_{ij} - 1)$, summiert über alle Urteilkategorien.

$$p_i = \left(\sum_{j=1}^{N_c} n_{ij}(n_{ij} - 1) \right) / [m \cdot (m - 1)], i = 1, 2, \dots, N_o$$

Tabelle 68: Bestimmungsschema für die Anteile beobachteter Übereinstimmungen

Objekt-Nr.	Anzahl der „Okay“-Urteile (j = 1)	Anzahl der „Not okay“-Urteile (j = 2)	Anteile beobachteter Übereinstimmungen
i	n_{i1}	n_{i2}	p_i
1	$n_{11} = 2$	$n_{12} = 0$	$p_1 = \frac{2 \cdot (2 - 1) + 0}{2 \cdot (2 - 1)} = 1$
2	$n_{21} = 1$	$n_{22} = 1$	$p_2 = \frac{1 \cdot (1 - 1) + 1 \cdot (1 - 1)}{2 \cdot (2 - 1)} = 0$
3	$n_{31} = 1$	$n_{32} = 1$	$p_3 = \frac{1 \cdot (1 - 1) + 1 \cdot (1 - 1)}{2 \cdot (2 - 1)} = 0$
4	$n_{41} = 2$	$n_{42} = 0$	$p_4 = \frac{2 \cdot (2 - 1) + 0}{2 \cdot (2 - 1)} = 1$
5	$n_{51} = 2$	$n_{52} = 0$	$p_5 = \frac{2 \cdot (2 - 1) + 0}{2 \cdot (2 - 1)} = 1$
Summe beobachteter Anteile übereinstimmender Urteile			$\sum_{i=1}^{N_o=5} p_i = 3$

Den Anteil beobachteter Übereinstimmungen erhalten wir als Mittelwert aus den je Einheit beobachteten Urteilsübereinstimmungen:

$$p_o = \sum_{i=1}^{N_o=5} p_i / N_o = \frac{3}{5} = 0,60$$

Einen gewissen Anteil an Urteilsübereinstimmungen erwarten wir durch den Zufall. Dieser Anteil ist in dem beobachteten Anteil an Urteilsübereinstimmungen enthalten und wird bereinigt.

2.5.2.3 Erwarteter Anteil Übereinstimmungen (A | Fleiss)

Aus der Tabelle 67 übernehmen wir die absolute Häufigkeit der Urteile pro Urteilkategorie.

Absolute Häufigkeit der vergebenen „Okay“-Urteile:

$$n_{\cdot 1} = 8$$

Absolute Häufigkeit der vergebenen „*Not okay*“-Urteile:

$$n_{.2} = 2$$

Im nächsten Schritt berechnen wir dafür die relativen Häufigkeiten (Anteile). Dazu teilen wir die Anzahl der Urteile je Kategorie durch die Anzahl der insgesamt vergebenen Urteile:

Die Anzahl der insgesamt in beiden Durchläufen vergebenen Urteile ist $n_{.1} + n_{.2} = 10$.

Mit diesen Informationen bestimmen wir den durchschnittlichen Anteil „*Okay*“-Urteile :

$$\bar{p}_{.1} = \frac{n_{.1}}{n_{.1} + n_{.2}} = \frac{8}{8 + 2} = 0,8$$

Analog bestimmen wir den durchschnittlichen Anteil „*Not okay*“-Urteile:

$$\bar{p}_{.2} = \frac{n_{.2}}{n_{.1} + n_{.2}} = \frac{2}{8 + 2} = 0,2$$

Mit diesen beiden Anteilswerten bilden wir schließlich den durch Zufall erwarteten Anteil übereinstimmender Urteile:

$$p_e = \bar{p}_{.1}^2 + \bar{p}_{.2}^2 = 0,8^2 + 0,2^2 = 0,68$$

2.5.2.4 Vom Zufall bereinigte Anteile (A | Fleiss)

Zunächst bestimmen wir den **beobachteten Anteil übereinstimmender Urteile, der vom zufällig erwarteten Anteil übereinstimmender Urteile bereinigt ist**:

$$p_o - p_e = 0,60 - 0,68 = -0,08$$

Anschließend bestimmen wir den **maximal möglichen Anteil übereinstimmender Urteile, der vom zufällig erwarteten Anteil übereinstimmender Urteile bereinigt ist**:

$$1 - p_e = 1 - 0,68 = 0,32$$

2.5.2.5 Kappa-Koeffizient (A | Fleiss)

Teilen wir die beiden bereinigten Anteilswerte durcheinander, so erhalten wir den **Kappa-Koeffizienten nach Fleiss**:

$$\kappa_{A(\text{Fleiss})} = \frac{-0,08}{0,32} = -0,25$$

2.5.2.6 Signifikanztest (A | Fleiss)

Zu prüfende Hypothese:

Nullhypothese H_0 : Der Kappa-Koeffizient der Grundgesamtheit ist gleich Null

Praktische Deutung der Nullhypothese: Die übereinstimmenden Urteile sind vermutlich allein durch den Zufall entstanden.

Alternativhypothese H_1 : Der Kappa-Koeffizient der Grundgesamtheit ist ungleich Null.

Praktische Deutung der Alternativhypothese: Die übereinstimmenden Urteile sind nicht (allein) durch den Zufall zustande gekommen. Der Urteile folgen einer erkennbaren Systematik.

Für die Prüfung der Nullhypothese benötigen wir die Prüfgröße z , die wir aus der Division des Kappa-Koeffizienten durch die Standardabweichung des Kappa-Koeffizienten erhalten.

2.5.2.7 Standardabweichung (A | Fleiss)

Wir bestimmen die Hilfsgröße h_1 :

$$h_1 = \bar{p}_{.1}(1 - \bar{p}_{.1}) + \bar{p}_{.2}(1 - \bar{p}_{.2}) = 0,80 \cdot 0,20 + 0,20 \cdot 0,80 = 0,32$$

Wir bestimmen die Hilfsgröße h_2 :

$$h_2 = \bar{p}_{.1}(1 - \bar{p}_{.1})(1 - \bar{p}_{.1}) + \bar{p}_{.2}(1 - \bar{p}_{.2})(1 - 2\bar{p}_{.2}) = 0,80 \cdot 0,20 \cdot (-0,6) + 0,20 \cdot 0,80 \cdot 0,6 = 0$$

Wir berechnen die Standardabweichung:

$$se(\kappa) \approx \sqrt{\frac{2(h1^2 - h2)}{m(m-1)N_o h1^2}}$$

Der Prüfer A hat zwei Prüfdurchgänge ($m=2$ Urteilsspalten) an fünf Einheiten ($N_o = 5$) ausgeführt. Damit berechnen wir:

$$se_A(\kappa) = \sqrt{\frac{2 \cdot (0,32^2 - 0)}{2 \cdot (2-1) \cdot 5 \cdot 0,32^2}} \approx 0,4472136$$

2.5.2.8 Prüfgröße z (A | Fleiss)

Wir teilen den Kappa-Koeffizienten durch die Standardabweichung und erhalten so die Prüfgröße z:

$$z_{\kappa_A} = \frac{\kappa_{A(Fleiss)}}{se_A(\kappa)} = \frac{-0,25}{0,447214} \approx -0,55902$$

2.5.2.9 P-Wert (A | Fleiss)

Mit der Verteilungsfunktion der Standardnormalverteilung berechnen wir den P-Wert:

$$P = 1 - G(z_{\kappa_A}) = 1 - G(-0,55902) \approx 0,712$$

Interpretation des Testergebnisses:

Aufgrund des viel zu kleinen Stichprobenumfangs ist die Power des Tests (= Nullhypothese verwerfen, wenn diese nicht zutrifft) viel zu gering und daher das Testen an sich hier unsinnig. Dennoch vollziehen wir den formalen Akt des „Test-Rituals“ mit dem P-Wert.

Da der P-Wert größer ist als das gewählte Signifikanzniveau $\alpha = 0,05$ (entspricht $\alpha = 5\%$), wird die Nullhypothese nicht verworfen (=beibehalten). Der beobachtete Kappa-Wert spricht für zufällige Urteilsübereinstimmungen.

2.5.3 Wiederholbarkeit Prüfer B

Mit der Wiederholbarkeit wollen wir den Grad der Übereinstimmung der Urteile in allen Durchläufen des Prüfers B allein feststellen.

2.5.3.1 Aufbau Datentabelle (B | Fleiss)

Aus der Tabelle 2 auf der Seite 10 entnehmen wir nur die Urteile des Prüfers B. Anschließend zählen wir zeilenweise, wie oft jede Urteilkategorie vergeben worden ist.

Tabelle 69: Absolute Häufigkeit beobachteter Urteile je Einheit für die Wiederholbarkeit Prüfer B

Objekt-Nr.	Prüfer B		Anzahl der „Okay“-Entscheidungen	Anzahl der „Not okay“-Entscheidungen
	Durchgang 1	Durchgang 2		
1	Not okay	Not okay	0	2
2	Okay	Not okay	1	1
3	Okay	Okay	2	0
4	Okay	Okay	2	0
5	Okay	Okay	2	0
Spaltensumme			$n_{.1} = 7$	$n_{.2} = 3$

Die Spaltensummen benötigen wir später für das Bestimmen der durch den Zufall erwarteten Anteile Übereinstimmungen.

2.5.3.2 Anteil beobachteter Übereinstimmungen (B | Fleiss)

Mit den zeilenweise (=für jede Einheit) ermittelten Urteilshäufigkeiten ermitteln wir den Anteil der beobachteten Übereinstimmungen.

$$p_i = \left(\sum_{j=1}^{N_c} n_{ij}(n_{ij} - 1) \right) / [m \cdot (m - 1)], i = 1, 2, \dots, N_o$$

Die theoretisch größtmögliche Anzahl an Urteilsübereinstimmungen (für eine einzelne Einheit) befindet sich im Nenner $m \cdot (m - 1)$ der vorstehenden Formel.

Im Zähler der p_i -Formel befindet sich die tatsächlich beobachtete Anzahl übereinstimmender Urteile $n_{ij}(n_{ij} - 1)$, als Summe über alle Urteilskategorien. Teilen wir den Zähler durch den Nenner, so erhalten wir den Anteil beobachteter Übereinstimmungen für jede Einheit:

Tabelle 70: Berechnungsschema für das Bestimmen des beobachteten Anteils übereinstimmender Entscheidungen

Objekt-Nr.	Anzahl der „Okay“-Urteile (j = 1)	Anzahl der „Not okay“-Urteile (j = 2)	Anteile beobachteter Übereinstimmungen
<i>i</i>	<i>n_{i1}</i>	<i>n_{i2}</i>	<i>p_i</i>
1	<i>n₁₁</i> = 0	<i>n₁₂</i> = 2	$p_1 = \frac{0 + 2 \cdot (2 - 1)}{2 \cdot (2 - 1)} = 1$
2	<i>n₂₁</i> = 1	<i>n₂₂</i> = 1	$p_2 = \frac{1 \cdot (1 - 1) + 1 \cdot (1 - 1)}{2 \cdot (2 - 1)} = 0$
3	<i>n₃₁</i> = 2	<i>n₃₂</i> = 0	$p_3 = \frac{2 \cdot (2 - 1) + 0}{2 \cdot (2 - 1)} = 1$
4	<i>n₄₁</i> = 2	<i>n₄₂</i> = 0	$p_4 = \frac{2 \cdot (2 - 1) + 0}{2 \cdot (2 - 1)} = 1$
5	<i>n₅₁</i> = 2	<i>n₅₂</i> = 0	$p_5 = \frac{2 \cdot (2 - 1) + 0}{2 \cdot (2 - 1)} = 1$
Summe beobachteter Anteile übereinstimmender Urteile			$\sum_{i=1}^{N_o=5} p_i = 4$

Den Anteil beobachteter Übereinstimmungen erhalten wir als Mittelwert aus den einzelnen beobachteten Urteilsübereinstimmungen:

$$p_o = \sum_{i=1}^{N_o=5} p_i / N_o = \frac{4}{5} = 0,80$$

Ein gewissen Anteil an Urteilsübereinstimmungen erwarten wir durch den Zufall. Dieser Anteil ist in dem beobachteten Anteil an Urteilsübereinstimmungen enthalten. Daher werden wir als nächstes diesen zufallsbedingten Übereinstimmungsanteil bestimmen und bereinigen.

2.5.3.3 Erwarteter Anteil zufälliger Urteilsübereinstimmungen (B | Fleiss)

Aus der Tabelle 69 übernehmen wir die absolute Häufigkeit der Urteile pro Urteilkategorie.

Absolute Häufigkeit der „*Okay*“-Urteile:

$$n_{.1} = 7$$

Absolute Häufigkeit der „*Not okay*“-Urteile:

$$n_{.2} = 3$$

Teilen wir diese Häufigkeiten durch die Gesamtzahl der ausgeführten Bewertungen

$n_{.1} + n_{.2} = 7 + 3 = 10$, so erhalten wir für die beiden Urteilkategorien die folgenden Anteilswerte.

Durchschnittlicher Anteil der „*Okay*“-Urteile:

$$\bar{p}_{.1} = \frac{n_{.1}}{n_{.1} + n_{.2}} = \frac{7}{10} = 0,70$$

Durchschnittlicher Anteil der „*Not okay*“-Urteile:

$$\bar{p}_{.2} = \frac{n_{.2}}{n_{.1} + n_{.2}} = \frac{3}{10} = 0,30$$

Mit diesen beiden Anteilswerten bilden wir schließlich den durch Zufall erwarteten Anteil übereinstimmender Urteile:

$$p_e = \bar{p}_{.1}^2 + \bar{p}_{.2}^2 = 0,70^2 + 0,30^2 = 0,58$$

2.5.3.4 Kappa-Koeffizient (B | Fleiss)

Wir bestimmen den beobachteten Anteil übereinstimmender Urteile, der von dem erwarteten Anteil zufälliger Übereinstimmungen bereinigt ist:

$$p_o - p_e = 0,80 - 0,58 = 0,22$$

Der größtmögliche Anteil beobachteter Urteilsübereinstimmungen ergibt sich wie folgt:

$$1 - p_e = 1 - 0,58 = 0,42$$

Mit diesen beiden bereinigten Anteilswerten bestimmen wir den Kappa-Koeffizienten:

$$\kappa_{B(\text{Fleiss})} = \frac{p_o - p_e}{1 - p_e} = \frac{0,22}{0,42} \approx 0,5238095$$

2.5.3.5 Signifikanztest (B | Fleiss)

Zu prüfende Hypothese:

Nullhypothese H0: Der Kappa-Koeffizient der Grundgesamtheit ist gleich Null

Praktische Deutung der Nullhypothesen: Die übereinstimmenden Urteile entstehen durch Zufall.

Alternativhypothese H1: Der Kappa-Koeffizient der Grundgesamtheit ist ungleich Null.

Praktische Deutung der Alternativhypothese: Die übereinstimmenden Urteile entstehen nicht (allein) durch den Zufall.

2.5.3.6 Standardabweichung (B | Fleiss)

Wir bestimmen die Hilfsgröße h1:

$$h1 = \bar{p}_{.1}(1 - \bar{p}_{.1}) + \bar{p}_{.2}(1 - \bar{p}_{.2}) = 0,70 \cdot 0,30 + 0,30 \cdot 0,70 = 0,42$$

Wir bestimmen die Hilfsgröße h2:

$$h2 = \bar{p}_{.1}(1 - \bar{p}_{.1})(1 - \bar{p}_{.1}) + \bar{p}_{.2}(1 - \bar{p}_{.2})(1 - 2\bar{p}_{.2}) = 0,70 \cdot 0,30 \cdot (-0,4) + 0,30 \cdot 0,70 \cdot 0,4 = 0$$

Wir berechnen die Standardabweichung:

$$se(\kappa) \approx \sqrt{\frac{2(h1^2 - h2)}{m(m-1)N_o h1^2}}$$

Der Prüfer A hat zwei Prüfdurchgänge ($m = 2$ Urteilsspalten) an fünf Einheiten ($N_o = 5$) ausgeführt. Damit berechnen wir:

$$se_B(\kappa) = \sqrt{\frac{2 \cdot (0,42^2 - 0)}{2 \cdot (2 - 1) \cdot 5 \cdot 0,42^2}} \approx 0,4472136$$

2.5.3.7 Prüfgröße z (B | Fleiss)

Teilen wir den Kappa-Koeffizienten durch die Standardabweichung, so erhalten wir die Prüfgröße z:

$$z_{\kappa_B} = \frac{\kappa_{B(Fleiss)}}{se_B(\kappa)} = \frac{0,5238095}{0,4472136} \approx 1,17128$$

2.5.3.8 P-Wert (B | Fleiss)

Mit der Verteilungsfunktion der Standardnormalverteilung bestimmen wir den P-Wert:

$$P = 1 - G(z_{\kappa_B}) = 1 - G(1,17128) = 0,121$$

Interpretation des Testergebnisses (P-Wert)

Aufgrund des viel zu kleinen Stichprobenumfangs ist die Power des Tests (= Nullhypothese verwerfen, wenn diese nicht zutrifft) viel zu gering und daher das Testen an sich hier unsinnig. Dennoch vollziehen wir den formalen Akt des „Test-Rituals“ mit dem P-Wert.

Der P-Wert ist größer als das gewählte Signifikanzniveau $\alpha = 0,05$ (entspricht $\alpha = 5\%$) und daher wird die Nullhypothese nicht verworfen (=beibehalten). Die beobachteten Übereinstimmungen sind vermutlich rein zufälliger Natur.

2.5.4 Vergleichbarkeit aller Prüfer (ohne Referenz)

Wir wollen durch diesen Vergleich erfahren, wie gut die Urteile aller Prüfer übereinstimmen.

2.5.4.1 Erwarteter Anteil zufälliger Übereinstimmungen (A x B | Fleiss)

Zunächst ermitteln wir durch Zählen die Anzahl der Urteile je Objekt und Kategorie $n_{i,j}$. Dafür betrachten wir alle Urteile der beiden Prüfer gemeinsam:

Tabelle 71: Summe beobachteter Übereinstimmungen

Objekt-Nr.	Prüfer A		Prüfer B		Anzahl <i>Okay</i>	Anzahl <i>Not okay</i>
	Durchgang 1	Durchgang 2	Durchgang 1	Durchgang 2		
1	Okay	Okay	Not okay	Not okay	$n_{1.1} = 2$	$n_{1.2} = 2$
2	Okay	Not okay	Okay	Not okay	$n_{2.1} = 2$	$n_{2.2} = 2$
3	Not okay	Okay	Okay	Okay	$n_{3.1} = 3$	$n_{3.2} = 1$
4	Okay	Okay	Okay	Okay	$n_{4.1} = 4$	$n_{4.2} = 0$
5	Okay	Okay	Okay	Okay	$n_{5.1} = 4$	$n_{5.2} = 0$
Summe					$n_{.1} = 15$	$n_{.2} = 5$

Es wurden in Summe $n_{.1} + n_{.2} = 15 + 5 = 20$ Urteile gefällt. Mit diesen Informationen bestimmen wir den durchschnittlichen Anteil der Urteilkategorie „*Okay*“:

$$\bar{p}_{.1} = \frac{n_{.1}}{n_{.1} + n_{.2}} = \frac{15}{20} = 0,75$$

Und den durchschnittlichen Anteil der Urteilkategorie „*Not okay*“:

$$\bar{p}_{.2} = \frac{n_{.2}}{n_{.1} + n_{.2}} = \frac{5}{20} = 0,25$$

Mit diesen beiden durchschnittlichen Anteilen ermitteln wir den durch Zufall erwarteten Anteil übereinstimmender Urteile:

$$p_e = \bar{p}_{.1}^2 + \bar{p}_{.2}^2 = 0,75^2 + 0,25^2 = 0,625$$

2.5.4.2 Anteil beobachteter Übereinstimmungen (A x B | Fleiss)

Den Anteil beobachteter Übereinstimmungen ermitteln wir für *jede Einheit* (=zeilenweise) gemäß der folgenden Beziehung:

$$p_i = \left(\sum_{j=1}^{N_c} n_{ij}(n_{ij} - 1) \right) / [m \cdot (m - 1)], i = 1, 2, \dots, N_o$$

Im Zähler dieser Anteilsformel befindet sich die tatsächlich beobachtete Anzahl Übereinstimmungen $n_{ij}(n_{ij} - 1)$, als Summe über alle Urteilkategorien.

Die größtmögliche Anzahl Urteilsübereinstimmungen $m \cdot (m - 1)$ befindet sich im Nenner, wobei wir für die Anzahl der zu vergleichenden Urteile $m = N_t \cdot N_o = 2 \cdot 2 = 4$ einsetzen

Tabelle 72: Tabelle der Anzahl beobachteten Übereinstimmungen je Urteilkategorie und der Anteile beobachteter Übereinstimmungen je Einheit

Objekt-Nr. <i>i</i>	Anzahl der „Okay“- Urteile	Anzahl der „Not okay“- Urteile	Anteil der beobachteten Übereinstimmungen
1	$n_{1,1} = 2$	$n_{1,2} = 2$	$p_1 = \frac{2 \cdot (2 - 1) + 2 \cdot (2 - 1)}{4 \cdot (4 - 1)} = 0, \overline{33}$
2	$n_{2,1} = 2$	$n_{2,2} = 2$	$p_2 = \frac{2 \cdot (2 - 1) + 2 \cdot (2 - 1)}{4 \cdot (4 - 1)} = 0, \overline{33}$
3	$n_{3,1} = 3$	$n_{3,2} = 1$	$p_3 = \frac{3 \cdot (3 - 1) + 1 \cdot (1 - 1)}{4 \cdot (4 - 1)} = 0, 50$
4	$n_{4,1} = 4$	$n_{4,2} = 0$	$p_4 = \frac{4 \cdot (4 - 1) + 0}{4 \cdot (4 - 1)} = 1$
5	$n_{5,1} = 4$	$n_{5,2} = 0$	$p_5 = \frac{4 \cdot (4 - 1) + 0}{4 \cdot (4 - 1)} = 1$
Spaltensumme:			$\sum_{i=1}^{N_o=5} p_i = \frac{19}{6} = 3, \overline{166}$

Der beobachtete Anteil Urteilsübereinstimmungen entspricht dem Mittelwert aus den für jede Einheit ermittelten Anteilswerten:

$$p_o = \sum_{i=1}^{N_o=5} p_i / N_o = \frac{19}{6 \cdot 5} = 0, \overline{633}$$

2.5.4.3 Kappa-Koeffizient (A × B | Fleiss)

Wir bereinigen den beobachteten Anteil an Urteilsübereinstimmungen von dem Anteil zufälliger Urteilsübereinstimmungen:

$$p_o - p_e = 0,6\overline{33} - 0,625 = 0,008\overline{33}$$

Analog bereinigen wir den größtmöglichen Anteil beobachteter Übereinstimmungen von dem Anteil zufälliger Urteilsübereinstimmungen:

$$1 - p_e = 1 - 0,625 = 0,375$$

Das Verhältnis der beiden bereinigten Anteile ergibt den Kappa-Koeffizienten:

$$\kappa_{A \times B} = \frac{p_o - p_e}{1 - p_e} = \frac{0,008\overline{33}}{0,375} = 0,02\overline{2}$$

2.5.4.4 Signifikanztest (A × B | Fleiss)

Zu prüfende Hypothese:

Nullhypothese H0: Der Kappa-Koeffizient der Grundgesamtheit ist gleich Null

Praktische Deutung der Nullhypothese: Die übereinstimmenden Urteile entstehen durch Zufall.

Alternativhypothese H1: Der Kappa-Koeffizient der Grundgesamtheit ist ungleich Null.

Praktische Deutung der Alternativhypothese: Die übereinstimmenden Urteile entstehen nicht (allein) durch den Zufall.

2.5.4.5 Standardabweichung (A × B | Fleiss)

Wir bestimmen zunächst die Hilfsgröße $h1$:

$$h1 = \bar{p}_1(1 - \bar{p}_1) + \bar{p}_2(1 - \bar{p}_2) = 0,75 \cdot 0,25 + 0,25 \cdot 0,75 = 0,375$$

Als nächstes berechnen wir die zweite Hilfsgröße $h2$:

$$h2 = \bar{p}_1(1 - \bar{p}_1)(1 - \bar{p}_1) + \bar{p}_2(1 - \bar{p}_2)(1 - 2\bar{p}_2) = 0,75 \cdot 0,25 \cdot (-0,5) + 0,25 \cdot 0,75 \cdot 0,5 = 0$$

Die Standardabweichung des Kappa-Koeffizienten ermitteln wir gemäß der folgenden Beziehung:

$$se(\kappa) \approx \sqrt{\frac{2(h1^2 - h2)}{m(m-1)N_o h1^2}}$$

$$se_{A \times B}(\kappa) \approx \sqrt{\frac{2 \cdot (0,375^2 - 0)}{4 \cdot (4-1) \cdot 5 \cdot 0,375^2}} \approx 0,182\,574\,186$$

unter Verwendung der Anzahl zu vergleichender Urteilsspalten $m = N_t \cdot N_o = 2 \cdot 2 = 4$.

2.5.4.6 Prüfgröße z (A × B | Fleiss)

Teilen wir den Kappa-Koeffizienten durch die Standardabweichung, so erhalten wir die Prüfgröße z:

$$z_{\kappa_{A \times B}} = \frac{\kappa_{A \times B | Fleiss}}{se_{A \times B}} \approx \frac{0,022\,222\,222}{0,182\,574\,186} \approx 0,121\,716\,124$$

2.5.4.7 P-Wert (A × B | Fleiss)

Mit der Verteilungsfunktion der Standardnormalverteilung berechnen wir:

$$P = 1 - G(z_{\kappa_{A \times B}}) \approx 1 - G(0,121\,716\,124) \approx 0,452$$

Interpretation

Aufgrund des viel zu kleinen Stichprobenumfangs ist die Power des Tests (= Nullhypothese verwerfen, wenn diese nicht zutrifft) viel zu gering und daher das Testen an sich hier unsinnig. Dennoch betrachten wir hier das „Formal-Ritual“ des Testentscheids.

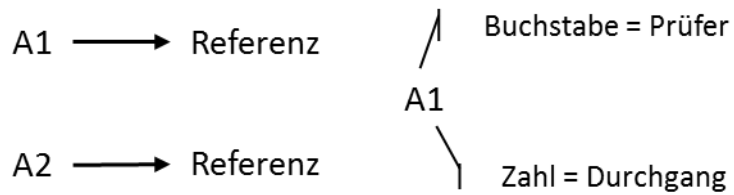
Der P-Wert ist größer als das gewählte Signifikanzniveau $\alpha = 0,05$ (entspricht $\alpha = 5\%$) und daher wird die Nullhypothese nicht verworfen (=beibehalten). Die beobachteten Übereinstimmungen sind vermutlich rein zufälliger Natur.

2.5.5 Vergleichbarkeit Prüfer A vs. Referenz

Mit diesem Vergleich wollen wir den Grad der Übereinstimmung der Urteile des Prüfers A mit den Referenzurteilen feststellen.

2.5.5.1 Aufbau Datentabelle (A × Ref. | Fleiss)

Aus der Tabelle 2 auf der Seite 10 entnehmen wir die Urteile des Prüfers A und die Referenz-Urteile. Bei einem Vergleich des Prüfers mit der Referenz führen wir alle Berechnungsschritte der Kappa-Berechnung für jeden Prüfdurchgang einzeln aus.



Für jeden Prüfdurchgang zählen wir zeilenweise die absoluten Häufigkeiten der vergebenen Urteile.

Prüfer A vs. Referenz, erster Durchgang

Tabelle 73: Absolute Häufigkeit beobachteter Urteile je Einheit für die Vergleichbarkeit Prüfer A vs. Referenz im ersten Prüfdurchgang

Objekt-Nr. i	Referenz-Urteil	Prüfer A	Anzahl der „Okay“-Entscheidungen	Anzahl der „Not okay“-Entscheidungen
		Durchgang 1		
1	Okay	Okay	$n_{1.1} = 2$	$n_{1.2} = 0$
2	Not okay	Okay	$n_{2.1} = 1$	$n_{2.2} = 1$
3	Okay	Not okay	$n_{3.1} = 1$	$n_{3.2} = 1$
4	Not okay	Okay	$n_{4.1} = 1$	$n_{4.2} = 1$
5	Okay	Okay	$n_{5.1} = 2$	$n_{5.2} = 0$
Spaltensumme			$n_{.1} = 7$	$n_{.2} = 3$

Prüfer A vs. Referenz, zweiter Durchgang

Tabelle 74: Absolute Häufigkeiten beobachteter Urteile je Einheit für die Vergleichbarkeit Prüfer A vs. Referenz im zweiten Prüfdurchlauf

Objekt-Nr. i	Referenz-Urteil	Prüfer A	Anzahl der „Okay“-Entscheidungen	Anzahl der „Not okay“-Entscheidungen
1	Okay	Okay	$n_{1.1} = 2$	$n_{1.2} = 0$
2	Not okay	Not okay	$n_{2.1} = 0$	$n_{2.2} = 2$
3	Okay	Okay	$n_{3.1} = 2$	$n_{3.2} = 0$
4	Not okay	Okay	$n_{4.1} = 1$	$n_{4.2} = 1$
5	Okay	Okay	$n_{5.1} = 2$	$n_{5.2} = 0$
Spaltensummen			$n_{.1} = 7$	$n_{.2} = 3$

Die Spaltensummen benötigen wir später für den durch Zufall erwarteten Anteil Übereinstimmungen.

2.5.5.2 Anteil beobachteter Übereinstimmungen (A x Ref. | Fleiss)

Auch diese Berechnungen führen wir für jeden Prüfdurchgang einzeln aus:

Im Zähler der folgenden Anteilsformel befindet sich die tatsächlich beobachtete Anzahl Übereinstimmungen $n_{ij}(n_{ij} - 1)$ für jede Einheit und im Nenner die maximal mögliche Anzahl übereinstimmender Urteile $m \cdot (m - 1)$. In jeder Zeile werden $m = N_t = 2$ Urteile miteinander verglichen.

$$p_i = \left(\sum_{j=1}^{N_c} n_{ij}(n_{ij} - 1) \right) / [m \cdot (m - 1)], i = 1, 2, \dots, N_o$$

Berechnung für den Prüfer A im ersten Durchgang

Mit den beobachteten Übereinstimmungen je Einheit $n_{i,j}$ ermitteln wir den Anteil Übereinstimmungen je Einheit p_i :

Objekt-Nr. i	Anzahl der „Okay“-Urteile	Anzahl der „Not okay“- Urteile	Anteile beobachteter Übereinstimmungen
i	$n_{i,1}$	$n_{i,2}$	p_i
1	$n_{1,1} = 2$	$n_{1,2} = 0$	$p_1 = \frac{2 \cdot (2 - 1) + 0}{2 \cdot (2 - 1)} = 1,00$
2	$n_{2,1} = 1$	$n_{2,2} = 1$	$p_2 = \frac{1 \cdot (1 - 1) + 1 \cdot (1 - 1)}{2 \cdot (2 - 1)} = 0,00$
3	$n_{3,1} = 1$	$n_{3,2} = 1$	$p_3 = \frac{1 \cdot (1 - 1) + 1 \cdot (1 - 1)}{2 \cdot (2 - 1)} = 0,00$
4	$n_{4,1} = 1$	$n_{4,2} = 1$	$p_4 = \frac{1 \cdot (1 - 1) + 1 \cdot (1 - 1)}{2 \cdot (2 - 1)} = 0,00$
5	$n_{5,1} = 2$	$n_{5,2} = 0$	$p_5 = \frac{2 \cdot (2 - 1) + 0}{2 \cdot (2 - 1)} = 1,00$
Summe beobachteter Anteile übereinstimmender Urteile:			$\sum_{i=1}^{N_o=5} p_i = 2$

Den beobachteten Anteil Übereinstimmungen erhalten wir als Mittelwert mit den zeilenweise ermittelten beobachteten Übereinstimmungsanteilen:

$$p_{o(A \times Ref.)1. \text{ Durchlauf}} = \sum_{i=1}^{N_o=5} p_i / N_o = \frac{2}{5} = 0,40$$

Berechnung für den Prüfer B im zweiten Durchgang

Wieder ermitteln wir für jede Einheit den beobachteten Anteil Übereinstimmungen p_i anhand der für jede Einheit und Kategorie gezählten Übereinstimmungen $n_{i,j}$.

Tabelle 76: Bestimmung des Anteils beobachteter Übereinstimmungen für den Prüfer B im zweiten Durchgang

Virtuelle Objekt-Nr.	Anzahl der „Okay“-Urteile (j = 1)	Anzahl der „Not okay“-Urteile (j = 2)	Anteile beobachteter Übereinstimmungen
i	n_{i1}	n_{i2}	p_i
1	$n_{1,1} = 2$	$n_{1,2} = 0$	$p_6 = \frac{2 \cdot (2 - 1) + 0}{2 \cdot (2 - 1)} = 1,00$
2	$n_{2,1} = 0$	$n_{2,2} = 2$	$p_7 = \frac{0 + 2 \cdot (2 - 1)}{2 \cdot (2 - 1)} = 1,00$
3	$n_{3,1} = 2$	$n_{3,2} = 0$	$p_8 = \frac{2 \cdot (2 - 1) + 0}{2 \cdot (2 - 1)} = 1,00$
4	$n_{4,1} = 1$	$n_{4,2} = 1$	$p_9 = \frac{1 \cdot (1 - 1) + 1 \cdot (1 - 1)}{2 \cdot (2 - 1)} = 0,00$
5	$n_{5,1} = 2$	$n_{5,2} = 0$	$p_{10} = \frac{2 \cdot (2 - 1) + 0}{2 \cdot (2 - 1)} = 1,00$
Summe beobachteter Anteile übereinstimmender Urteile			$\sum_{i=1}^{N_o=5} p_i = 4$

Den beobachteten Anteil Übereinstimmungen erhalten wir als Mittelwert der zeilenweise ermittelten Übereinstimmungsanteile:

$$p_{o(A \times Ref. | 2. \text{ Durchlauf})} = \sum_{i=1}^{N_o=5} p_i / N_o = \frac{4}{5} = 0,80$$

Einen gewissen Anteil an Urteilsübereinstimmungen erwarten wir durch den Zufall. Dieser Anteil ist in dem beobachteten Anteil an Urteilsübereinstimmungen enthalten. Daher werden wir als nächstes diesen zufallsbedingten Übereinstimmungsanteil bestimmen und bereinigen.

2.5.5.3 Erwarteter Anteil Übereinstimmungen (A × Ref. | Fleiss)

Die Berechnung der Anteilswerte führen wir für den ersten und zweiten Prüfdurchgang separat aus.

Berechnung für den Prüfer A im ersten Durchgang

Aus der Tabelle 73 entnehmen wir die Spaltensummen. Das sind die absoluten Häufigkeiten der Urteile für jede Urteilkategorie. Zunächst die Absolute Häufigkeit der „*Okay*“-Urteile im ersten Prüfdurchgang:

$$n_{.1} = 7$$

Und die absolute Häufigkeit der „*Not okay*“-Urteile im ersten Prüfdurchgang:

$$n_{.2} = 3$$

Teilen wir diese Häufigkeiten durch die Gesamtzahl der ausgeführten Bewertungen

$n_{.1} + n_{.2} = 7 + 3 = 10$, so erhalten wir für die beiden Urteilkategorien die Anteile.

Durchschnittlicher Anteil der „*Okay*“-Urteile im **ersten Prüfdurchgang**:

$$\bar{p}_{.1} = \frac{n_{.1}}{n_{.1} + n_{.2}} = \frac{7}{7 + 3} = 0,70$$

Durchschnittlicher Anteil der „*Not okay*“- Urteile im **ersten Prüfdurchgang**:

$$\bar{p}_{.2} = \frac{n_{.2}}{n_{.1} + n_{.2}} = \frac{3}{7 + 3} = 0,30$$

Mit diesen beiden Anteilswerten bilden wir schließlich den **durch Zufall erwarteten Anteil übereinstimmender Urteile**:

$$p_{e(A \times Ref., 1. \text{ Durchlauf})} = \bar{p}_{.1}^2 + \bar{p}_{.2}^2 = 0,70^2 + 0,30^2 = 0,58$$

Berechnung für den Prüfer A im zweiten Durchgang

Aus der Tabelle 74 entnehmen wir die Spaltensummen. Das sind die absoluten Häufigkeiten der Urteile für jede Urteilkategorie. Absolute Häufigkeit der „*Okay*“-Urteile im zweiten Prüfdurchgang:

$$n_{.1} = 7$$

Absolute Häufigkeit der „*Not okay*“-Urteile im zweiten Prüfdurchgang:

$$n_{.2} = 3$$

Teilen wir diese Häufigkeiten durch die Gesamtzahl der ausgeführten Bewertungen $n_{.1} + n_{.2} = 7 + 3 = 10$, so erhalten wir für die beiden Urteilstkategorien die folgenden Anteilswerte. Durchschnittlicher Anteil der „*Okay*“-Urteile im **zweiten Prüfdurchgang**:

$$\bar{p}_{.1} = \frac{n_{.1}}{n_{.1} + n_{.2}} = \frac{7}{7 + 3} = 0,70$$

Durchschnittlicher Anteil der „*Not okay*“-Urteile im **zweiten Prüfdurchgang**:

$$\bar{p}_{.2} = \frac{n_{.2}}{n_{.1} + n_{.2}} = \frac{3}{7 + 3} = 0,30$$

Mit diesen beiden Anteilswerten bilden wir schließlich den **durch Zufall erwarteten Anteil übereinstimmender Urteile**:

$$p_{e(A \times Ref, | 2. \text{ Durchlauf})} = \bar{p}_{.1}^2 + \bar{p}_{.2}^2 = 0,70^2 + 0,30^2 = 0,58$$

2.5.5.4 Kappa-Koeffizient (A vs. Ref. | Fleiss)

Wir bestimmen den beobachteten Anteil übereinstimmender Urteile, der vom erwarteten Anteil zufälliger Übereinstimmungen bereinigt ist.

Berechnung für Prüfer A im ersten Durchgang

$$p_{o(A \times Ref, | 1. \text{ Durchlauf})} - p_{e(A \times Ref, | 1. \text{ Durchlauf})} = 0,40 - 0,58 = -0,18$$

Der größtmögliche Anteil beobachteter Urteilsübereinstimmungen ergibt sich wie folgt:

$$1 - p_{e(A \times Ref, | 1. \text{ Durchlauf})} = 1 - 0,58 = 0,42$$

Mit diesen beiden bereinigten Anteilswerten bestimmen wir den Kappa-Koeffizienten:

$$\kappa_{A \times Ref(1. \text{ Durchlauf})} = \frac{p_{o(A \times Ref, | 1. \text{ Durchlauf})} - p_{e(A \times Ref, | 1. \text{ Durchlauf})}}{1 - p_{e(A \times Ref, | 1. \text{ Durchlauf})}} = \frac{-0,18}{0,42} \approx -0,428\,571$$

Berechnung für Prüfer A im zweiten Durchgang

Der vom zufällig erwarteten Anteil Übereinstimmungen bereinigte Anteil beobachteter Urteile:

$$p_{o(A \times \text{Ref.}|2. \text{ Durchlauf})} - p_{e(A \times \text{Ref.}|2. \text{ Durchlauf})} = 0,80 - 0,58 = 0,22$$

Der größtmögliche Anteil beobachteter Urteilsübereinstimmungen, der vom Anteil zufällig erwarteter Übereinstimmungen bereinigt ist, ergibt sich wie folgt:

$$1 - p_{e(A \times \text{Ref.}|2. \text{ Durchlauf})} = 1 - 0,58 = 0,42$$

Mit den bereinigten Anteilen bestimmen wir den Kappa-Koeffizienten:

$$\kappa_{A \times \text{Ref.}(2. \text{ Durchlauf})} = \frac{p_{o(A \times \text{Ref.}|2. \text{ Durchlauf})} - p_{e(A \times \text{Ref.}|2. \text{ Durchlauf})}}{1 - p_{e(A \times \text{Ref.}|2. \text{ Durchlauf})}} = \frac{0,22}{0,42} \approx 0,52381$$

Mit den Kappa-Koeffizienten aus den beiden Prüfdurchgängen bilden wir den Mittelwert. Das Ergebnis ist der gesuchte Kappa-Koeffizient für den Vergleich des Prüfers A mit der Referenz:

$$\kappa_{A \times \text{Ref.}} = \frac{\kappa_{(A \times \text{Ref.}|1. \text{ Durchgang})} + \kappa_{(A \times \text{Ref.}|2. \text{ Durchgang})}}{2} = \frac{-0,42857 + 0,52381}{2} = 0,04762$$

2.5.5.5 Signifikanztest (A × Ref. | Fleiss)

Zu prüfende Hypothese:

Nullhypothese H0: Der Kappa-Koeffizient der Grundgesamtheit ist gleich Null

Praktische Deutung der Nullhypothesen: Die übereinstimmenden Urteile entstehen durch Zufall.

Alternativhypothese H1: Der Kappa-Koeffizient der Grundgesamtheit ist ungleich Null.

Praktische Deutung der Alternativhypothese: Die übereinstimmenden Urteile entstehen nicht (allein) durch den Zufall.

Wir wählen das Signifikanz-Niveau $\alpha = 0,05$ (entspricht $\alpha = 5\%$).

2.5.5.6 Standardabweichung (A x Ref. | Fleiss)

Zunächst bestimmen wir die Varianz des Kappa-Koeffizienten für den ersten und zweiten Prüfdurchgang separat. Anschließend bestimmen wir die mittlere Varianz und ziehen daraus die Wurzel. Das Ergebnis ist die gesuchte Standardabweichung.

Varianz für den ersten Prüfdurchgang:

Wir bestimmen zunächst die Hilfsgröße $h1$ mit den durchschnittlichen Anteilen aus dem ersten Prüfdurchgang:

$$h1 = \bar{p}_1(1 - \bar{p}_1) + \bar{p}_2(1 - \bar{p}_2) = 0,70 \cdot 0,30 + 0,30 \cdot 0,70 = 0,42$$

Wir bestimmen die Hilfsgröße $h2$ mit den durchschnittlichen Anteilen aus dem ersten Prüfdurchgang:

$$h2 = \bar{p}_1(1 - \bar{p}_1)(1 - \bar{p}_1) + \bar{p}_2(1 - \bar{p}_2)(1 - 2\bar{p}_2) = 0,70 \cdot 0,30 \cdot (-0,4) + 0,30 \cdot 0,70 \cdot 0,3 = 0$$

Mit den beiden Hilfsgrößen bestimmen wir die Varianz des ersten Prüfdurchgangs:

$$var(\kappa_{A1 \times Ref}) \approx \frac{2(h1^2 - h2)}{m(m-1)N_o h1^2} = \frac{2 \cdot 0,42^2}{2 \cdot (2-1) \cdot 5 \cdot 0,42^2} = 0,20$$

Varianz für den zweiten Prüfdurchgang

Wir bestimmen zunächst die Hilfsgröße $h1$ mit den durchschnittlichen Anteilen aus dem zweiten Prüfdurchgang:

$$h1 = \bar{p}_1(1 - \bar{p}_1) + \bar{p}_2(1 - \bar{p}_2) = 0,70 \cdot 0,30 + 0,30 \cdot 0,70 = 0,42$$

Wir bestimmen die Hilfsgröße h_2 mit den durchschnittlichen Anteilen aus dem zweiten Prüfdurchgang:
 $h_2 = \bar{p}_{.1}(1 - \bar{p}_{.1})(1 - \bar{p}_{.1}) + \bar{p}_{.2}(1 - \bar{p}_{.2})(1 - 2\bar{p}_{.2}) = 0,70 \cdot 0,30 \cdot (-0,3) + 0,30 \cdot 0,70 \cdot 0,3 = 0$

Mit den beiden Hilfsgrößen bestimmen wir die Varianz des zweiten Prüfdurchgangs:

$$var(\kappa_{A2 \times Ref}) \approx \frac{2(h_1^2 - h_2)}{m(m-1)N_o h_1^2} = \frac{2 \cdot 0,42^2}{2 \cdot (2-1) \cdot 5 \cdot 0,42^2} = 0,20$$

Varianz des Kappa-Koeffizienten für die Vergleichbarkeit A×Referenzen

Wir bilden den Mittelwert mit den Varianzen aus den beiden Prüfdurchgängen:

$$\overline{var}(\kappa_{A \times Ref}) = \frac{var(\kappa_{A1 \times Ref}) + var(\kappa_{A2 \times Ref})}{N_t^2} = \frac{0,2 + 0,2}{2^2} = 0,1$$

Standardabweichung des Kappa-Koeffizienten

Wir ziehen die Wurzel aus der Varianz und erhalten die Standardabweichung:

$$se_{A \times Ref.}(\kappa_{A \times Ref.}) = \sqrt{0,1} = 0,316\ 228$$

2.5.5.7 Prüfgröße z (A × Ref. | Fleiss)

Wir erhalten die Prüfgröße z, indem wir den Kappa-Koeffizienten durch die Standardabweichung teilen:

$$z_{\kappa_{A \times Ref.}} = \frac{z_{\kappa_{A \times Ref.}}}{se_{A \times Ref.}} \approx \frac{0,047\ 62}{0,316\ 228} \approx 0,150\ 588$$

2.5.5.8 P-Wert (A × Ref. | Fleiss)

Mit der Verteilungsfunktion der Standardnormalverteilung bestimmen wir den P-Wert:

$$P = 1 - G(z_{\kappa_{A \times Ref.}}) \approx 1 - G(0,150588) \approx 0,44$$

Interpretation des Testergebnisses

Aufgrund des viel zu kleinen Stichprobenumfangs ist die Power des Tests (= Nullhypothese verwerfen, wenn diese nicht zutrifft) viel zu gering und daher das Testen an sich hier unsinnig. Dennoch betrachten wir hier das „Formal-Ritual“ des Testentscheids.

Der P-Wert ist größer als das gewählte Signifikanzniveau $\alpha = 0,05$ (entspricht $\alpha = 5\%$) und daher wird die Nullhypothese nicht verworfen (=beibehalten). Die beobachteten Übereinstimmungen sind vermutlich rein zufälliger Natur.

2.5.6 Vergleichbarkeit Prüfer B vs. Referenz

Mit der Vergleichbarkeit Prüfer B vs. Referenz wollen wir den Grad der Übereinstimmung der Urteile mit der Referenz feststellen.

2.5.6.1 Aufbau Datentabelle (B vs. Ref. | Fleiss)

Aus der Tabelle 2 auf der Seite 10 entnehmen wir nur die Urteile des Prüfers B und die Referenz-Urteile. Anschließend zählen wir zeilenweise, wie oft jede Urteilkategorie vergeben worden ist.

Diese Betrachtung führen wir für den ersten und zweiten Prüfdurchgang separat durch:

Tabelle 77: Beobachtete Häufigkeit der Übereinstimmungen für die Vergleichbarkeit Prüfer B im ersten Durchgang vs. Referenz

Objekt-Nr. i	Referenz-Urteil	Prüfer B	Anzahl der „Okay“-Urteile	Anzahl der „Not okay“- Urteile
1	Okay	Not okay	$n_{1.1} = 1$	$n_{1.2} = 1$
2	Not okay	Okay	$n_{2.1} = 1$	$n_{2.2} = 1$
3	Okay	Okay	$n_{3.1} = 2$	$n_{3.2} = 0$
4	Not okay	Okay	$n_{4.1} = 1$	$n_{4.2} = 1$
5	Okay	Okay	$n_{5.1} = 2$	$n_{5.2} = 0$
Spaltensumme:			$n_{.1} = 7$	$n_{.2} = 3$

Tabelle 78: Beobachtete Häufigkeit der Übereinstimmungen für die Vergleichbarkeit Prüfer B im zweiten Durchgang vs. Referenz

Objekt-Nr. i	Referenz-Urteil	Prüfer B	Anzahl der „Okay“-Urteile	Anzahl der „Not okay“- Urteile
1	Okay	Not okay	$n_{1.1} = 1$	$n_{1.2} = 1$
2	Not okay	Not okay	$n_{2.1} = 0$	$n_{2.2} = 2$
3	Okay	Okay	$n_{3.1} = 2$	$n_{3.2} = 0$
4	Not okay	Okay	$n_{4.1} = 1$	$n_{4.1} = 1$
5	Okay	Okay	$n_{5.1} = 2$	$n_{5.1} = 0$
Spaltensumme			$n_{.1} = 6$	$n_{.2} = 4$

Die Spaltensummen verwenden wir später für das Bestimmen des zufällig erwarteten Anteils an Übereinstimmungen.

2.5.6.2 Anteil beobachteter Übereinstimmungen (B vs. Ref. | Fleiss)

Wir bestimmen den Anteil beobachteter Übereinstimmungen für den ersten und zweiten Prüfdurchlauf separat:

$$p_i = \left(\sum_{j=1}^{N_c} n_{ij}(n_{ij} - 1) \right) / [m \cdot (m - 1)], i = 1, 2, \dots, N_o$$

Im Zähler der Anteilsformel befindet sich die Anzahl tatsächlich beobachteter Übereinstimmungen $n_{ij}(n_{ij} - 1)$ und im Nenner befindet sich die maximal mögliche Anzahl Übereinstimmungen $m \cdot (m - 1)$.

Berechnung für den Prüfer B im ersten Durchgang

Wir ermitteln die Anteile der tatsächlich beobachteten Übereinstimmungen für jede Einheit.

Tabelle 79: Bestimmung des beobachteten Anteils Übereinstimmungen für den

Vergleich des Prüfers B im ersten Durchgang mit der Referenz

Virtuelle Objekt-Nr.	Anzahl der „Okay“-Urteile (j = 1)	Anzahl der „Not okay“-Urteile (j = 2)	Anteile beobachteter Übereinstimmungen
i	n_{i1}	n_{i2}	p_i
1	$n_{1,1} = 1$	$n_{1,2} = 1$	$p_1 = \frac{1 \cdot (1 - 1) + 1 \cdot (1 - 1)}{2 \cdot (2 - 1)} = 0,00$
2	$n_{2,1} = 1$	$n_{2,2} = 1$	$p_2 = \frac{1 \cdot (1 - 1) + 1 \cdot (1 - 1)}{2 \cdot (2 - 1)} = 0,00$
3	$n_{3,1} = 2$	$n_{3,2} = 0$	$p_3 = \frac{2 \cdot (2 - 1) + 0}{2 \cdot (2 - 1)} = 1,00$
4	$n_{4,1} = 1$	$n_{4,2} = 1$	$p_4 = \frac{1 \cdot (1 - 1) + 1 \cdot (1 - 1)}{2 \cdot (2 - 1)} = 0,00$
5	$n_{5,1} = 2$	$n_{5,2} = 0$	$p_5 = \frac{2 \cdot (2 - 1) + 0}{2 \cdot (2 - 1)} = 1,00$
Summe beobachteter Anteile übereinstimmender Urteile			$\sum_{i=1}^{N_o=5} p_i = 2$

Den beobachteten Anteil Übereinstimmungen erhalten wir als Mittelwert der zeilenweise ermittelten Übereinstimmungsanteile:

$$p_{o(B \times Ref | 1. \text{ Durchlauf})} = \sum_{i=1}^{N_o=5} p_i / N_o = \frac{2}{5} = 0,40$$

Berechnung für den Prüfers B im zweiten Durchgang

Tabelle 80: Bestimmung des beobachteten Anteils Übereinstimmungen für den Vergleich des Prüfers B im zweiten Durchgang mit der Referenz

Virtuelle Objekt-Nr.	Anzahl der „Okay“-Urteile	Anzahl der „Not okay“-Urteile	Anteile beobachteter Übereinstimmungen
1	$n_{1,1} = 1$	$n_{1,2} = 1$	$p_1 = \frac{1 \cdot (1 - 1) + 1 \cdot (1 - 1)}{2 \cdot (2 - 1)} = 0,00$
2	$n_{2,1} = 0$	$n_{2,2} = 2$	$p_2 = \frac{0 + 2 \cdot (2 - 1)}{2 \cdot (2 - 1)} = 1,00$
3	$n_{3,1} = 2$	$n_{3,2} = 0$	$p_3 = \frac{2 \cdot (2 - 1) + 0}{2 \cdot (2 - 1)} = 1,00$
4	$n_{4,1} = 1$	$n_{4,2} = 1$	$p_4 = \frac{1 \cdot (1 - 1) + 1 \cdot (1 - 1)}{2 \cdot (2 - 1)} = 0,00$
5	$n_{5,1} = 2$	$n_{5,2} = 0$	$p_5 = \frac{2 \cdot (2 - 1) + 0}{2 \cdot (2 - 1)} = 1,00$
Summe beobachteter Anteile übereinstimmender Urteile			$\sum_{i=1}^{N_o=5} p_i = 3$

Den beobachteten Anteil Übereinstimmungen erhalten wir als Mittelwert der zeilenweise ermittelten Übereinstimmungsanteile:

$$p_{o(B \times Ref | 2. \text{ Durchlauf})} = \sum_{i=1}^{N_o=5} p_i / N_o = \frac{3}{5} = 0,60$$

Einen gewissen Anteil an Urteilsübereinstimmungen erwarten wir durch den Zufall. Dieser Anteil ist in dem beobachteten Anteil an Urteilsübereinstimmungen enthalten. Daher werden wir als nächstes diesen zufallsbedingten Übereinstimmungsanteil bestimmen und bereinigen.

2.5.6.3 Durch den Zufall erwarteter Anteil Übereinstimmungen (B vs. Ref. | Fleiss)

Wir bestimmen den durch Zufall erwarteten Anteil für jeden Prüfdurchgang separat:

Berechnung für den Prüfer B im ersten Prüfdurchgang:

Aus der entnehmen wir die Spaltensummen für den ersten Prüfdurchgang. Das sind die absoluten Häufigkeiten der Urteile für jede Urteilkategorie.

Absolute Häufigkeit der „*Okay*“-Urteile im ersten Prüfdurchgang:

$$n_{.1} = 7$$

Absolute Häufigkeit der „*Not okay*“-Urteile im ersten Prüfdurchgang:

$$n_{.2} = 3$$

Teilen wir diese Häufigkeiten durch die Gesamtzahl der ausgeführten Bewertungen

$n_{.1} + n_{.2} = 7 + 3 = 10$, so erhalten wir die durchschnittlichen Anteile je Urteilkategorie.

Durchschnittlicher Anteil der „*Okay*“-Urteile im ersten Prüfdurchgang:

$$\bar{p}_{.1} = \frac{n_{.1}}{n_{.1} + n_{.2}} = \frac{7}{7 + 3} = 0,70$$

Durchschnittlicher Anteil der „*Not okay*“-Urteile im ersten Prüfdurchgang:

$$\bar{p}_{.2} = \frac{n_{.2}}{n_{.1} + n_{.2}} = \frac{3}{7 + 3} = 0,30$$

Mit diesen beiden Anteilswerten bilden wir schließlich den **durch Zufall erwarteten Anteil übereinstimmender Urteile**:

$$p_{e(B \times Ref., 1. \text{ Durchlauf})} = \bar{p}_{.1}^2 + \bar{p}_{.2}^2 = 0,70^2 + 0,30^2 = 0,58$$

Berechnung für den Prüfer B im zweiten Durchgang:

Die gleichen Rechenschritte führen wir auch mit den Urteilen aus dem zweiten Durchgang des Prüfers B aus.

Absolute Häufigkeit der „*Okay*“-Urteile im zweiten Prüfdurchgang:

$$n_{.1} = 6$$

Absolute Häufigkeit der „*Not okay*“-Urteile im zweiten Prüfdurchgang:

$$n_{.2} = 4$$

Teilen wir diese Häufigkeiten durch die Gesamtzahl der ausgeführten Bewertungen

$n_{.1} + n_{.2} = 6 + 4 = 10$, so erhalten wir die durchschnittlichen Anteilswerte je Urteilkategorie.

Durchschnittlicher Anteil „*Okay*“-Urteile im zweiten Prüfdurchgang:

$$\bar{p}_{.1} = \frac{n_{.1}}{n_{.1} + n_{.2}} = \frac{6}{6 + 4} = 0,60$$

Durchschnittlicher Anteil „*Not okay*“-Urteile im zweiten Prüfdurchgang:

$$\bar{p}_{.2} = \frac{n_{.2}}{n_{.1} + n_{.2}} = \frac{4}{6 + 4} = 0,40$$

Mit diesen beiden Anteilswerten bilden wir schließlich den **durch Zufall erwarteten Anteil übereinstimmender Urteile**:

$$p_{e(B \times Ref., 2. \text{ Durchlauf})} = \bar{p}_{.1}^2 + \bar{p}_{.2}^2 = 0,60^2 + 0,40^2 = 0,52$$

2.5.6.4 Kappa-Koeffizient (B vs. Ref. | Fleiss)

Wir bestimmen den beobachteten Anteil übereinstimmender Urteile, der von dem erwarteten Anteil zufälliger Übereinstimmungen bereinigt ist.

Berechnung für den Prüfer B, erster Durchgang:

$$p_{o(B \times Ref., 1. \text{ Durchlauf})} - p_{e(B \times Ref., 1. \text{ Durchlauf})} = 0,40 - 0,58 = -0,18$$

Der größtmögliche Anteil beobachteter Urteilsübereinstimmungen ergibt sich wie folgt:

$$1 - p_{e(B \times Ref. | 1. \text{ Durchlauf})} = 1 - 0,58 = 0,42$$

Mit diesen beiden bereinigten Anteilswerten bestimmen wir den Kappa-Koeffizienten:

$$\kappa_{B \times Ref(1. \text{ Durchlauf})} = \frac{p_{o(B \times Ref. | 1. \text{ Durchlauf})} - p_{e(B \times Ref. | 1. \text{ Durchlauf})}}{1 - p_{e(B \times Ref. | 1. \text{ Durchlauf})}} = \frac{-0,18}{0,42} = -\frac{3}{7}$$

Berechnung für den Prüfer B, zweiter Durchgang:

$$p_{o(B \times Ref. | 2. \text{ Durchlauf})} - p_{e(B \times Ref. | 2. \text{ Durchlauf})} = 0,60 - 0,52 = 0,08$$

Der größtmögliche Anteil beobachteter Urteilsübereinstimmungen ergibt sich wie folgt:

$$1 - p_{e(B \times Ref. | 2. \text{ Durchlauf})} = 1 - 0,52 = 0,48$$

Mit diesen beiden bereinigten Anteilswerten bestimmen wir den Kappa-Koeffizienten:

$$\kappa_{B \times Ref(2. \text{ Durchlauf})} = \frac{p_{o(B \times Ref. | 2. \text{ Durchlauf})} - p_{e(B \times Ref. | 2. \text{ Durchlauf})}}{1 - p_{e(B \times Ref. | 2. \text{ Durchlauf})}} = \frac{0,08}{0,48} = \frac{1}{6}$$

Aus den Kappa-Werten beider Durchläufe bilden wir den Mittelwert und erhalten:

$$\kappa_{B \times Ref} = \frac{\kappa_{B \times Ref(1. \text{ Durchlauf})} + \kappa_{B \times Ref(2. \text{ Durchlauf})}}{2} = \frac{-3 \cdot 6 + 1 \cdot 7}{2 \cdot 6 \cdot 7} = -\frac{11}{84} \approx -0.13095$$

2.5.6.5 Signifikanztest Kappa-Koeffizient (B x Ref. | Fleiss)

Zu prüfende Hypothese:

Nullhypothese H0: Der Kappa-Koeffizient der Grundgesamtheit ist gleich Null

Praktische Deutung der Nullhypothesen: Die übereinstimmenden Urteile entstehen durch Zufall.

Alternativhypothese H1: Der Kappa-Koeffizient der Grundgesamtheit ist ungleich Null.

Praktische Deutung der Alternativhypothese: Die übereinstimmenden Urteile entstehen nicht (allein) durch den Zufall.

Wir wählen das Signifikanz-Niveau $\alpha = 0,05$ (entspricht $\alpha = 5\%$).

2.5.6.6 Standardabweichung (B x Ref. | Fleiss)

Zunächst bestimmen wir für den ersten und zweiten Prüfdurchgang separat die Varianz des Kappa-Koeffizienten. Anschließend bilden wir den Mittelwert der beiden Varianzen. Die Wurzel aus dieser gemittelten Varianz ergibt schließlich die gesuchte Standardabweichung.

Varianz des Kappa-Koeffizienten aus dem ersten Prüfdurchgang:

Wir bestimmen zunächst die Hilfsgröße $h1$ mit den durchschnittlichen Anteilen aus dem ersten Prüfdurchgang:

$$h1 = \bar{p}_{.1}(1 - \bar{p}_{.1}) + \bar{p}_{.2}(1 - \bar{p}_{.2}) = 0,70 \cdot 0,30 + 0,30 \cdot 0,70 = 0,42$$

Wir bestimmen die Hilfsgröße $h2$ mit den durchschnittlichen Anteilen aus dem ersten Prüfdurchgang:

$$h2 = \bar{p}_{.1}(1 - \bar{p}_{.1})(1 - \bar{p}_{.1}) + \bar{p}_{.2}(1 - \bar{p}_{.2})(1 - 2\bar{p}_{.2}) = 0,70 \cdot 0,30 \cdot (-0,4) + 0,30 \cdot 0,70 \cdot 0,4 = 0$$

Mit den beiden Hilfsgrößen bestimmen wir die Varianz des ersten Prüfdurchgangs:

$$var(\kappa_{B1 \times Ref}) \approx \frac{2(h1^2 - h2)}{m(m-1)N_o h1^2} = \frac{2 \cdot 0,42^2}{2 \cdot (2-1) \cdot 5 \cdot 0,42^2} = 0,20$$

Varianz des Kappa-Koeffizienten aus dem zweiten Prüfdurchgang:

Wir bestimmen zunächst die Hilfsgröße $h1$ mit den durchschnittlichen Anteilen aus dem zweiten Prüfdurchgang:

$$h1 = \bar{p}_{.1}(1 - \bar{p}_{.1}) + \bar{p}_{.2}(1 - \bar{p}_{.2}) = 0,60 \cdot 0,40 + 0,40 \cdot 0,60 = 0,48$$

Wir bestimmen die Hilfsgröße $h2$ mit den durchschnittlichen Anteilen aus dem zweiten Prüfdurchgang:

$$h2 = \bar{p}_{.1}(1 - \bar{p}_{.1})(1 - \bar{p}_{.1}) + \bar{p}_{.2}(1 - \bar{p}_{.2})(1 - 2\bar{p}_{.2}) = 0,60 \cdot 0,40 \cdot (-0,2) + 0,40 \cdot 0,60 \cdot 0,2 = 0$$

Mit den beiden Hilfsgrößen bestimmen wir die Varianz des zweiten Prüfdurchgangs:

$$var(\kappa_{B2 \times Ref}) \approx \frac{2(h1^2 - h2)}{m(m-1)N_o h1^2} = \frac{2 \cdot 0,48^2}{2 \cdot (2-1) \cdot 5 \cdot 0,48^2} = 0,20$$

Mittelwert der beiden Varianzen

Wir bilden den Mittelwert mit den Varianzen aus den beiden Prüfdurchgängen:

$$\overline{var}(\kappa_{B \times Ref}) = \frac{var(\kappa_{B1 \times Ref}) + var(\kappa_{B2 \times Ref})}{N_t^2} = \frac{0,2 + 0,2}{2^2} = 0,1$$

Standardabweichung des Kappa-Koeffizienten

Wir ziehen die Wurzel aus der Varianz und erhalten die Standardabweichung:

$$se_{B \times Ref.}(\kappa_{B \times Ref.}) = \sqrt{0,1} = 0,316\,228$$

2.5.6.7 Prüfgröße z (B x Ref. | Fleiss)

Wir erhalten die Prüfgröße z, indem wir den Kappa-Koeffizienten durch die Standardabweichung teilen:

$$z_{\kappa_{B \times Ref.}} = \frac{\kappa_{B \times Ref.}}{se_{B \times Ref.}} \approx \frac{-0,130\,95}{0,316\,228} \approx -0,414$$

2.5.6.8 P-Wert (B x Ref. | Fleiss)

Mit der Verteilungsfunktion der Standardnormalverteilung berechnen wir den P-Wert:

$$P = 1 - G(z_{\kappa_{B \times Ref.}}) \approx 1 - G(-0,414\,108) \approx 0,661$$

Interpretation des Testergebnisses

Aufgrund des viel zu kleinen Stichprobenumfangs ist die Power des Tests (= Nullhypothese verwerfen, wenn diese nicht zutrifft) viel zu gering und daher das Testen an sich hier unsinnig. Dennoch betrachten wir hier das „Formal-Ritual“ des Testentscheids.

Der P-Wert ist größer als das gewählte Signifikanzniveau $\alpha = 0,05$ (entspricht $\alpha = 5\%$) und daher wird die Nullhypothese nicht verworfen (=beibehalten). Die beobachteten Übereinstimmungen sind vermutlich rein zufälliger Natur.

2.5.7 Vergleichbarkeit aller Prüfer vs. Referenz

Mit der Vergleichbarkeit aller Prüfer vs. Referenz wollen wir den Grad der Übereinstimmung der Urteile mit der Referenz feststellen.

2.5.7.1 Aufbau Datentabelle (Vergleichbarkeit aller Prüfer vs. Referenz | Fleiss)

Für das Bestimmen des Kappa-Koeffizienten für die Vergleichbarkeit aller Prüfer vs. Referenz bilden wir den Mittelwert aus allen Einzel-Vergleichen der einzelnen Prüfer vs. Referenz. In unserem Fallbeispiel sind das die beiden Vergleiche A mit Referenz und B mit Referenz:

$$\kappa_{A \times B \times Ref} = \frac{\kappa_{A \times Ref} + \kappa_{B \times Ref}}{2} = \frac{0,04762 + (-0,13095)}{2} = -0,04167$$

2.5.7.2 Signifikanztest Kappa-Koeffizient (A x B x Ref. | Fleiss)

Wir fassen die Varianzen aus dem Vergleich der einzelnen Prüfer mit der Referenz zu einem Mittelwert zusammen:

$$\overline{var}_{A \times B \times Ref} = \frac{var_{A1 \times Ref} + var_{A2 \times Ref} + var_{B1 \times Ref} + var_{B2 \times Ref}}{4^2} = \frac{4 \cdot 0,2}{16} = \frac{1}{20} = 0,05$$

2.5.7.3 Standardabweichung (A x B x Ref. | Fleiss)

Die Standardabweichung erhalten wir durch das Ziehen der Wurzel aus dem Mittelwert der Varianzen:

$$se_{A \times B \times Ref}(\kappa) = \sqrt{\overline{var}_{A \times B \times Ref}} = \sqrt{0,05} \approx 0,223\ 607$$

2.5.7.4 Prüfgröße z (A x B x Ref. | Fleiss)

Wir erhalten die Prüfgröße z, indem wir den Kappa-Koeffizienten durch die Standardabweichung teilen:

$$z_{A \times B \times Ref} = \frac{\kappa_{A \times B \times Ref}}{se_{A \times B \times Ref}(\kappa)} \approx \frac{-0,04167}{0,223\ 607} \approx -0,186\ 354$$

2.5.7.5 P-Wert (A x B x Ref. | Fleiss)

Mit der Verteilungsfunktion der Standardnormalverteilung berechnen wir den P-Wert:

$$P = 1 - G(z_{A \times B \times Ref}) \approx 1 - G(-0,186\ 354) \approx 0,573\ 916$$

Interpretation des Testergebnisses

Aufgrund des viel zu kleinen Stichprobenumfangs ist die Power des Tests (= Nullhypothese verwerfen, wenn diese nicht zutrifft) viel zu gering und daher das Testen an sich hier unsinnig. Dennoch betrachten wir hier das „Formal-Ritual“ des Testentscheids.

Der P-Wert ist größer als das gewählte Signifikanzniveau $\alpha = 0,05$ (entspricht $\alpha = 5\%$) und daher wird die Nullhypothese nicht verworfen (=beibehalten). Die beobachteten Übereinstimmungen sind vermutlich rein zufälliger Natur.

3. PRÜFSYSTEM-EFFEKTIVITÄT

Bei der Prüfsystem-Effektivität wird die Anzahl der Prüfobjekte gezählt, für die eine vollständige Übereinstimmung aller Urteile vorliegt. Hierbei verstehen wir unter der vollständigen Übereinstimmung der Urteile für ein bestimmtes Prüfobjekt, dass alle Prüfer das betreffende Objekt in allen Prüfdurchläufen der gleichen Urteilkategorie zugeordnet haben.

Allgemeine Bestimmungsgleichung für den Anteilswert

Wir zählen die Anzahl der Objekte, bei denen alle Urteile vollständig übereinstimmen.

N_e = Anzahl der Objekte, bei denen ausschließlich übereinstimmende Urteile vorliegen

N_o = Anzahl aller Objekte

$$p = \frac{N_e}{N_o}, \quad \text{bzw.} \quad p = \frac{N_e}{N_o} \cdot 100 \%$$

Allgemeine Bestimmungsgleichungen für den Vertrauensbereich des Anteilswertes

Für das Bestimmen der Grenzwerte des zweiseitigen Vertrauensbereiches (Konfidenzbereiches) für den Anteilswert zum Vertrauensniveau (Konfidenzniveau) $1 - \alpha$ verwenden wir die *inverse Verteilungsfunktion* (Quantil-Funktion) der Beta-Verteilung B^{-1} wie folgt:

Tabelle 81: Bestimmungsgleichungen für die Grenzen des zweiseitigen Vertrauensbereiches des Anteilswertes auf Basis der Quantilsfunktion (inverse Verteilungsfunktion)

Vertrauensbereich	Gleichung	Sonderfall
Unterer Grenzwert	$p_{lo} = B^{-1}\left(\frac{\alpha}{2}, N_e, N_o - N_e + 1\right)$	$p_{lo} = 0$ für $N_e = 0$
Oberer Grenzwert	$p_{up} = B^{-1}\left(1 - \frac{\alpha}{2}, N_e + 1, N_o - N_e\right)$	$p_{up} = 1$ für $N_e = N_o$

3.1 Zählung der Anzahl Objekte mit übereinstimmenden Urteilen

Für das Zählen der Anzahl Prüfobjekte, bei denen übereinstimmende Urteile gefällt wurden, unterscheiden wir zwei Zählarten:

1. Prüfer-Effektivität *ohne* einen Referenz-Vergleich
2. Prüfer-Effektivität *mit* einem Referenz-Vergleich

Die zweite Art des Zählens ist im Programm natürlich nur dann verfügbar, wenn wir für jedes Prüfobjekt das zugehörige Referenz-Urteil eingegeben haben (in der Wertemaske).

3.2 Prüfsystem-Effektivität ohne einen Referenz-Vergleich

Bei der Prüfsystem-Effektivität ohne einen Referenz-Vergleich unterscheiden wir weitere zwei Unterarten des Zählens:

1. Zählung der Objekte mit gleichen Urteilswerten für jeden Prüfer einzeln betrachtet
2. Zählung der Objekte mit gleichen Urteilswerten über alle Prüfer hinweg betrachtet.

Festlegung: Wir gehen davon aus, dass die Urteile für ein Prüfobjekt in einer Tabellenzeile (der Wertemaske) stehen.

Somit gilt:

- Übereinstimmung = alle Werte innerhalb einer Zeile sind gleich
- Nichtübereinstimmung = mindestens ein Wert innerhalb einer Zeile ist verschieden

3.2.1 Prüfsystem-Effektivität innerhalb der einzelnen Prüfer (ohne Referenz)

Hier betrachten wir jeden der Prüfer einzeln für sich. Wir zählen die Anzahl solcher Objekte, bei denen der gerade betrachtete Prüfer in allen Durchläufen das gleiche Urteil gefällt hat. Mit der Bezeichnung „innerhalb der Prüfer“ ist gemeint, dass wir die Entscheidungen (in allen Durchläufen) eines einzelnen Prüfers allein für sich betrachten.

Betrachtung für den Prüfer A

Tabelle 82: Beispiel-Datensatz – Urteile des Prüfers A über fünf Objekte in zwei Prüfdurchläufen

Objekt-Nr.	Prüfer A		Übereinstimmende Urteile in der Zeile?
	Durchgang 1	Durchgang 2	
1	Okay	Okay	Ja
2	Okay	Not okay	Nein
3	Not okay	Okay	Nein
4	Okay	Okay	Ja
5	Okay	Okay	Ja

Wie wir anhand der Tabelle 82 sehen können, hat der Prüfer A zwei der fünf Objekte in den beiden Durchläufen unterschiedlich beurteilt. Anders ausgedrückt: der Prüfer A hat für $N_{eA} = 3$ Objekte der insgesamt $N_o = 5$ Objekte ausschließlich gleiche Entscheidungen gefällt. Anhand dieser Zählungen bestimmen wir für den Prüfer A den Anteil der Objekte mit gleichen Entscheidungen p_A :

$$p_A = \frac{N_{eA}}{N_o} \cdot 100 \% = \frac{3}{5} \cdot 100 \% = 60 \%$$

mit

N_{eA} = Anzahl der Objekte, zu denen Prüfer A ausschließlich gleiche Urteile gefällt hat

N_o = Anzahl aller Objekte

Betrachtung für den Prüfer B

Führen wir die gleiche Vorgehensweise für den Prüfer B durch, so erhalten wir das folgende Ergebnis:

Tabelle 83: Beispiel-Datensatz – Urteile des Prüfers B über fünf Objekte in zwei Prüfdurchläufen

Objekt-Nr.	Prüfer B		Übereinstimmende Urteile in der Zeile?
	Durchgang 1	Durchgang 2	
1	Not okay	Not okay	Ja
2	Okay	Not okay	Nein
3	Okay	Okay	Ja
4	Okay	Okay	Ja
5	Okay	Okay	Ja

Wie wir anhand der Tabelle 83 sehen können, hat der Prüfer B bei $N_{eB} = 4$ Prüfobjekten in den zwei Prüfdurchläufen ein übereinstimmendes Urteil vergeben. Mit diesem Ergebnis berechnen wir den Anteil übereinstimmender Urteile p_B innerhalb des Prüfers B wie folgt:

$$p_B = \frac{N_{eB}}{N_o} \cdot 100 \% = \frac{4}{5} \cdot 100 \% = 80 \%$$

dabei ist

N_{eB} = Anzahl der Objekte, zu denen Prüfer B ausschließlich übereinstimmende Urteile gefällt hat

N_o = Anzahl aller Objekte

Die fünf Prüfobjekte stellen eine Stichproben-Zufallsauswahl aus einer (potenziell) unendlich großen Grundgesamtheit dar. Aus diesem Grund ist der berechnete Anteil übereinstimmender Urteile unsicher. Wie groß der „wahre“ Anteilswert übereinstimmender Entscheidungen in der Grundgesamtheit sein könnte, klären wir mit dem zweiseitigen $1 - \alpha$ -Vertrauensbereich (Konfidenzintervall):

Tabelle 84: Beispiel-Datensatz - Bestimmungsgleichungen für den zweiseitigen $(1 - \alpha)$ -Vertrauensbereich des Anteils übereinstimmender Urteile innerhalb des Prüfers A bzw. innerhalb des Prüfers B

Prüfer	Grenze	Parameter a	Parameter b	Quantil der Beta-Verteilung
A	p_{loA}	$N_{eA} = 3$	$N_o - N_{eA} + 1 = 3$	$B^{-1}\left(\frac{\alpha}{2} = 0,025; a; b\right) \approx 0,147$
	p_{upA}	$N_{eA} + 1 = 4$	$N_o - N_{eA} = 2$	$B^{-1}\left(1 - \frac{\alpha}{2} = 0,975; a; b\right) \approx 0,947$
B	p_{loB}	$N_{eB} = 4$	$N_o - N_{eB} + 1 = 2$	$B^{-1}\left(\frac{\alpha}{2} = 0,025; a; b\right) \approx 0,284$
	p_{upB}	$N_{eB} + 1 = 5$	$N_o - N_{eB} = 1$	$B^{-1}\left(1 - \frac{\alpha}{2} = 0,975; a; b\right) \approx 0,995$

Die Vertrauensbereiche (Konfidenzbereiche) für den Anteil übereinstimmender Urteile innerhalb der Prüfer A und B sind also (nach dem Multiplizieren mit 100 %):

$$p_{loA} = 14,7 \% \leq p_A \leq p_{upA} = 94,7 \%$$

$$p_{loB} = 28,4 \% \leq p_B \leq p_{upB} = 99,5 \%$$

3.2.2 Prüfsystem-Effektivität über alle Prüfer hinweg (ohne Referenz)

Für diese Art des Zählens betrachten wir die Urteile aller Prüfer ohne das Referenzurteil zu berücksichtigen. Wieder erkennen wir eine Übereinstimmung daran, dass alle Werte innerhalb einer Zeile exakt gleich sind.

Tabelle 85: Beispiel-Datensatz – Anzahl übereinstimmender Entscheidungen über alle Prüfer hinweg (ohne Referenzurteile)

Objekt-Nr.	Prüfer A		Prüfer B		Übereinstimmende Urteile in der Zeile?
	Durchgang 1	Durchgang 2	Durchgang 1	Durchgang 2	
1	Okay	Okay	Not okay	Not okay	Nein
2	Okay	Not okay	Okay	Not okay	Nein
3	Not okay	Okay	Okay	Okay	Nein
4	Okay	Okay	Okay	Okay	Ja
5	Okay	Okay	Okay	Okay	Ja

Wie wir anhand der Tabelle 85 sehen, haben beide Prüfer gemeinsam zwei der fünf Prüfobjekte übereinstimmend beurteilt. Daraus ergibt sich der Anteil übereinstimmender Entscheidungen innerhalb aller Prüfer:

$$p_{AB} = \frac{N_{e_{AB}}}{N_o} = \frac{2}{5} = 0,4$$

Tabelle 86: Beispiel-Datensatz – Bestimmung des zweiseitigen $(1 - \alpha)$ -Vertrauensbereiches (Konfidenzbereiches) für den Anteil übereinstimmender Entscheidungen innerhalb aller Prüfer

Prüfer	Grenze	Parameter a	Parameter b	Quantil der Beta-Verteilung
A & B	$p_{lo_{AB}}$	$N_{e_{AB}} = 2$	$N_o - N_{e_{AB}} + 1 = 4$	$B^{-1}\left(\frac{\alpha}{2} = 0,025; a; b\right) \approx 0,053$
	$p_{up_{AB}}$	$N_{e_{AB}} + 1 = 3$	$N_o - N_{e_{AB}} = 3$	$B^{-1}\left(1 - \frac{\alpha}{2} = 0,975; a; b\right) \approx 0,853$

Multiplizieren wir das Ergebnis aus der Tabelle 86 mit 100 %, so erhalten wir das folgende Ergebnis für den zweiseitigen $(1 - \alpha)$ -Vertrauensbereich für den Anteil übereinstimmender Urteile innerhalb aller Prüfer (hier: für die gemeinsame Betrachtung der beiden Prüfer A und B):

$$p_{lo_{AB}} = 5,3 \% \leq p_{AB} \leq p_{up_{AB}} = 85,3 \%$$

3.3 Prüfsystem-Effektivität mit Referenz-Vergleich

Ist in einem Datensatz das Referenzurteil zu jedem der Prüfobjekte eingegeben worden, so können wir zusätzlich die folgenden zwei Effektivitätsbetrachtungen ausführen:

- Vergleich der Urteile jedes einzelnen Prüfers mit den Referenz-Urteilen
- Vergleich der Urteile aller Prüfer gemeinsam mit den Referenz-Urteilen

3.3.1 Prüfsystem-Effektivität – Einzelne Prüfer vs. Referenz

In diesem Abschnitt vergleichen wir die Urteile von jedem Prüfer einzeln mit den Referenzurteilen.

3.3.1.1 Vergleich der Urteile – Prüfer A gegen Referenz

Hier vergleichen wir allein die Urteile des Prüfers A mit den Referenz-Urteilen. Als Übereinstimmungen werden wieder nur solche Zeilen gezählt, bei denen alle Zellen den gleichen Wert enthalten.

Tabelle 87: Beispiel-Datensatz – Prüfer A gegen Referenz

Objekt-Nr.	Referenzurteil	Prüfer A		Übereinstimmende Urteile in der Zeile?
		Durchgang 1	Durchgang 2	
1	Okay	Okay	Okay	Ja
2	Not okay	Okay	Not okay	Nein
3	Okay	Not okay	Okay	Nein
4	Not okay	Okay	Okay	Nein
5	Okay	Okay	Okay	Ja

3.3.1.2 Vergleich der Urteile – Prüfer B gegen Referenz

Wir vergleichen nur die Urteile des Prüfers B mit den Referenzurteilen. Wie zuvor gilt, dass eine Übereinstimmung gegeben ist, wenn alle Werte innerhalb einer Zeile exakt gleich sind.

Tabelle 88: Beispiel-Datensatz – Prüfer B gegen Referenz

Objekt-Nr.	Referenzurteil	Prüfer B		Übereinstimmende Urteile in der Zeile?
		Durchgang 1	Durchgang 2	
1	Okay	Not okay	Not okay	Nein
2	Not okay	Okay	Not okay	Nein
3	Okay	Okay	Okay	Ja
4	Not okay	Okay	Okay	Nein
5	Okay	Okay	Okay	Ja

Mit Blick auf die Tabelle 88 sehen wir, dass zwei von fünf Prüfobjekten übereinstimmend beurteilt wurden. Damit ergibt sich der Anteil übereinstimmender Entscheidungen zu:

$$p_{BRef} = \frac{N_{e_{BRef}}}{N_o} = \frac{2}{5} = 0,4$$

Das Ergebnis für den zweiseitigen $(1 - \alpha)$ -Vertrauensbereich haben wir in den vorhergehenden Beispielen schon mehrfach durchgerechnet, so dass wir uns hier auf die Wiedergabe des Endergebnisses beschränken:

$$p_{lo_{BRef}} = 5,3 \% \leq p_{BRef} \leq p_{up_{BRef}} = 85,3 \%$$

3.3.2 Prüfsystem-Effektivität – Alle Prüfer vs. Referenz

In diesem Abschnitt vergleichen wir die Urteile aller Prüfer gemeinsam mit den Referenzurteilen. Auch gilt hier wieder die Regel: Eine Übereinstimmung ist gegeben, wenn *alle* Werte in einer Zeile gleich sind.

Tabelle 89: Beispiel-Datensatz – Beide Prüfer gegen Referenz

Objekt-Nr.	Referenz-urteil	Prüfer A		Prüfer B		Übereinstimmende Urteile in der Zeile?
		D. 1	D. 2	D. 1	D. 2	
1	Okay	Okay	Okay	Not okay	Not okay	Nein
2	Not okay	Okay	Not okay	Okay	Not okay	Nein
3	Okay	Not okay	Okay	Okay	Okay	Nein
4	Not okay	Okay	Okay	Okay	Okay	Nein
5	Okay	Okay	Okay	Okay	Okay	Ja

Aus der Tabelle 89 entnehmen wir, dass die beiden Prüfer ein einziges Prüfobjekt der insgesamt fünf Prüfobjekte übereinstimmend beurteilt haben. Damit ergibt sich der Anteil übereinstimmender Entscheidungen für die Prüfer gegen Referenz zu:

$$p_{ABRef} = \frac{N_{e_{ABRef}}}{N_o} = \frac{1}{5} = 0,2$$

Tabelle 90: Beispiel-Datensatz – Zweiseitiger $(1 - \alpha)$ -Vertrauensbereich für den Anteil übereinstimmender Entscheidungen für den Vergleich aller Prüfer gegen die Referenz

Prüfer	Grenze	Parameter a	Parameter b	Quantil der Beta-Verteilung
A & B & Ref	$p_{lo_{ABRef}}$	$N_{e_{ABRef}} = 1$	$N_o - N_{e_{ABRef}} + 1 = 5$	$B^{-1}\left(\frac{\alpha}{2} = 0,025; a; b\right) \approx 0,005$
	$p_{up_{ABRef}}$	$N_{e_{ABRef}} + 1 = 2$	$N_o - N_{e_{ABRef}} = 4$	$B^{-1}\left(1 - \frac{\alpha}{2} = 0,975; a; b\right) \approx 0,72$

$$p_{lo_{ABRef}} = 0,5 \% \leq p_{ABRef} \leq p_{up_{ABRef}} = 71,6 \%$$

4. ANHANG

4.1 Verwendete Symbole und deren Bedeutung

N_o = Anzahl aller Objekte (engl. *objects*)

N_e = Anzahl der Objekte, für die ausschließlich gleiche Urteile vergeben wurden

N_a = Anzahl der urteilenden Instanzen (wie Personen, Gremien oder Prüfsysteme, engl. *appraiser*)

N_t = Anzahl der Prüfläufe (engl. *trials*)

N_c = Anzahl der Urteilskategorien, also die Anzahl der Stufenwerte der verwendeten Urteilswerte-Skala (engl. *categories*)

G = Verteilungsfunktion der Standardnormalverteilung

B^{-1} = Quantilfunktion (inverse Verteilungsfunktion) der Beta-Verteilung

α = Gewähltes Signifikanzniveau (und verbleibende Irrtumswahrscheinlichkeit)

5. LITERATUR

[Coh]

Jacob Cohen

A Coefficient of Agreement for nominal scales

Educational and psychological measurement

Volume 20, No. 1, 1960, pages 37-46)

DOI: 10.1177/001316446002000104

[H10]

Qualitätsmanagement in der Bosch-Gruppe – Technische Statistik#

Heft 10 – Fähigkeit von Mess- und Prüfprozessen

Ausgabe 05.2010

Robert Bosch GmbH

C/QMM

Postfach 300220

D-70442 Stuttgart

[VMB]

Jürgen Bortz, Gustav Lietner, Klaus Boehnke

Verteilungsfreie Methoden in der Biostatistik

3. Korrigierte Auflage

Springer Medizin Verlag, 2008

ISBN 978-3-540-74706-2

Kapitel 9 Übereinstimmungsanalyse, Formel (9.4) auf der Seite 452