Grundlagen der Diagnostik

Lerneinheit 8

Urteile & Fehler

Wahr-oder-Falsch Aussagen zur letzten Sitzung

1. Aufmerksamkeitsfehler durch nachlassende Konzentration werden als Beobachterdrift bezeichnet und gelten auch für Interviews.

2. Strukturierte Interviews sind weniger anfällig für Beurteilerfehler durch äußeres Erscheinungsbild/ Impression Management Strategien

3. Das professionelle Erscheinungsbild der Kandidatinnen korreliert durchschnittlich am geringsten mit Raterurteilen.

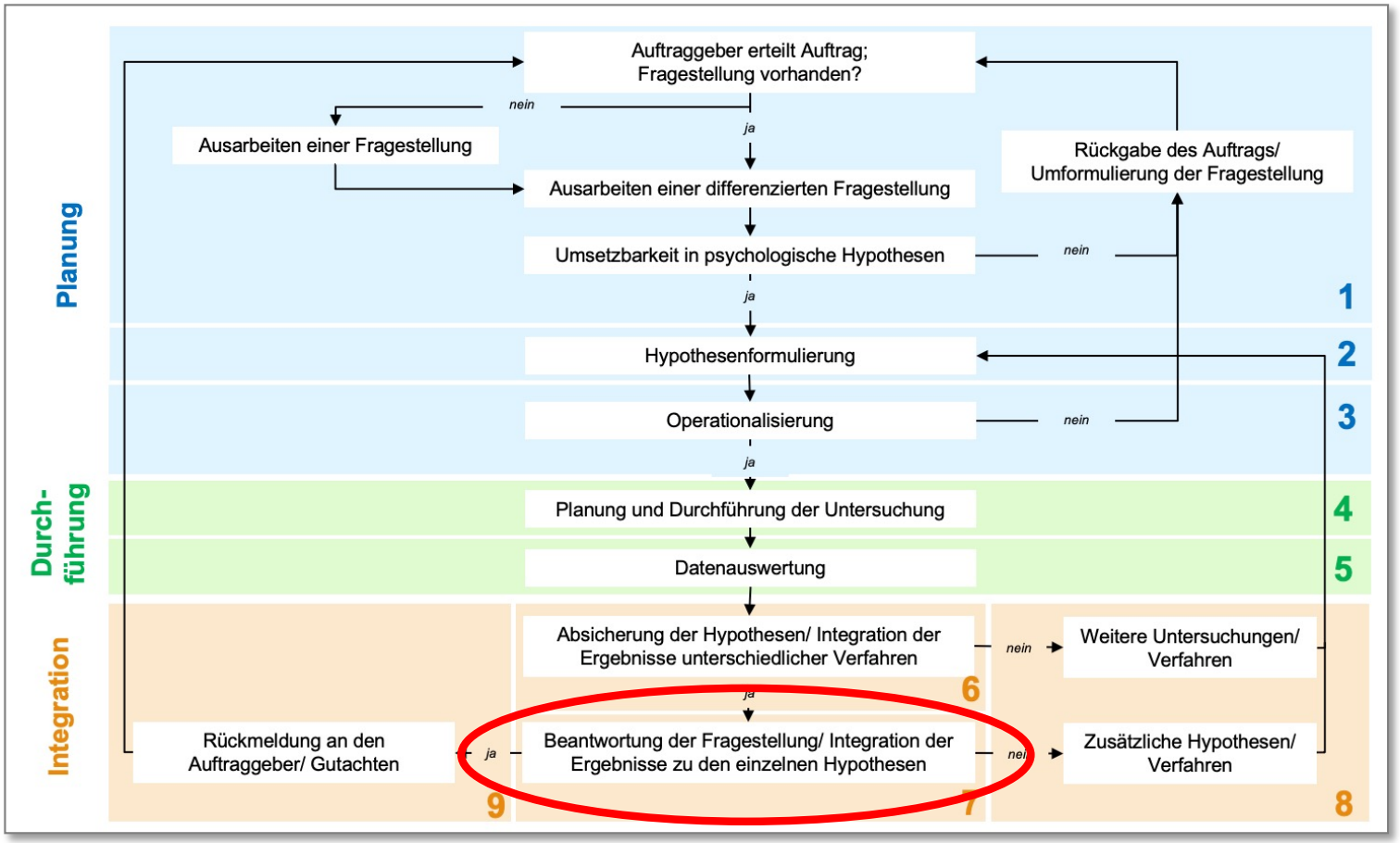
4. Durch Suggestivfragen kann die Konsistenz der Antworten der Kandidatinnen und Kandidaten überprüft werden.

5. Interviewerin notiert die Antwort der Person mit und bewertet anschließend nach vorher festgelegten Regeln

Definition

„Als diagnostisches Urteil wird die Beantwortung einer Fragestellung unter Verwendung von bereits vorliegenden diagnostischen Informationen bezeichnet.“ Schmidt-Atzert und Amelang, 2012, S. 390

→ Qualität eines Urteils kann überprüft werden, wenn ein Goldstandard vorliegt („Kriteriumswerte“, z.B. eine bestätigte psychiatrische Diagnose, der Ausbildungs- oder Berufserfolg, …)



(Abbildung adaptiert nach Ziegler & Bühner, 2012, S.16)

Die heutige Vorlesung

1. Arten der Urteilsbildung

2. Fehler & Güte von Urteilen

3. Optimierung einer Auswahlentscheidung

4. Beispiele

Lernziele

1. Arten der Urteilsbildung

Arten der Urteilsbildung

Klinische Urteilsbildung:

- Individuelle (intuitive) Urteile von Menschen/Diagnostikerinnen
- Freie Kombination der vorhandenen Informationen
- Die Bezeichnung beruht auf überwiegend aus dem klinischen Bereich stammender Forschung

Mechanische Urteilsbildung:

- Anwendung einer feststehenden, zuvor (empirisch) ermittelten Verrechnungsvorschrift („Formel“)
- Kombination der vorhandenen Informationen nach dieser Regel (manchmal auch *„Rationale Urteilsbildung“* genannt)
- Die Verrechnungsvorschrift kann auf einem statistischen Vorhersagemodell basieren (z.B. lineare oder logistische Regression). Hier bietet sich der Begriff *„Statistische Urteilsbildung“* an, es wird jedoch nicht immer klar unterschieden.

(vgl. Ziegler & Bühner, 2012, S.123ff)

Was funktioniert besser? – Klinische vs. Mechanische Urteilsbildung:

Historisches Beispiel (Goldberg, 1965):

- Vorliegende Information: Ausgewählte Skalenwerte von Patientinnen im Minnesota Multiphasic Personality Inventory (MMPI)
- Kriterium: Patientinnen „neurotisch“ oder „psychotisch“ (Psychiaterinnenurteile)

Klinische Urteilsbildung

- Klinikerin betrachtet Profil der Skalenwerte und fällt ein Urteil

Mechanische Urteilsbildung

- Verrechnung einzelner Skalen gemäß Goldberg-Index (Goldberg, 1965):

→ Lügenskala + Paranoia + Schizophrenie − Hysterie − Psychasthenie Index > 45: Klassifikation „psychotisch“

Trefferquote bei entweder psychotischen oder neurotischen Patientinnen:

68% **74%**

Was funktioniert besser? – Klinische vs. Mechanische Urteilsbildung:

Autoren	**Grove et al. (2000)**	**Ægisdóttir et al. (2006)**	**Kuncel et al. (2013)**
Kontext	Psychologischer/medizinischer Bereich	Klinisch-psychologischer Bereich	Beruflicher Bereich
Anzahl Studien	k = 136	k = 67	k = 18
Ergebnisse	▪ In 46% war das mechanische Urteil überlegen ▪ in 48% keine Unterschiede ▪ in 6% war das klinische Urteil überlegen ▪ Mechanische Urteile im Durchschnitt 10% genauer ▪ Klinische Urteile sind anfällig für Fehler (z.B. Verwendung einer unangemessen Gewichtung der Einzelbefunde) und die Anwendung von Heuristiken	▪ Bei *konservativer* Betrachtung (z.B. Ausschluss von Ausreißern): 41 Untersuchungen, in denen mechanische Urteile 13% genauer waren als klinische ▪ Bei Betrachtung *aller* Studien: Unterschied zugunsten der mechanischen Urteile noch größer	Korrelation der mechanischen vs. klinischen Urteile zur Vorhersage von ▪ Leistung im Beruf (k=9): .44 vs .28 ▪ Beruflicher Erfolg (k=5): .42 vs .36 ▪ Ausbildungserfolg (k=2): .31 vs .16

→ **Mechanische Urteile** über verschiedene Kontexte hinweg im Durchschnitt überlegen!

Grove, W. M., Zald, D. H., Lebow, B. S., Snitz, B. E., & Nelson, C. (2000). Clinical versus mechanical prediction: A meta-analysis. *Psychological Assessment, 12*(1), 19.

Was funktioniert besser? – Klinische vs. Mechanische Urteilsbildung:

Studie von Vrieze & Grove (2009):

- Befragung von 491 klinisch tätigen Psychologinnen in den USA
  - Rücklaufquote ca. 37% → 180 verwertete Antworten
  - Durchschnittlicher Anteil diagnostischer Tätigkeit an der Arbeitszeit: 20%
- Urteilsbildung:
  - 98% klinisch
  - 31% mechanisch (4% basierend auf statistischen Modellen)

Vrieze, S. I., & Grove, W. M. (2009). Survey on the use of clinical and mechanical prediction methods in clinical psychology. *Professional Psychology: Research and Practice, 40*(5), 525–531.

Was funktioniert besser? – Klinische vs. Mechanische Urteilsbildung:

Studie von Vrieze & Grove (2009):

Berichtete Gründe, warum Methoden der mechanischen Urteilsbildung (MU) *nicht* genutzt werden:

- 40%: MU nicht verfügbar
- 36%: Nicht ausreichend mit den Methoden vertraut, um sie bequem anzuwenden
- 32%: MU kann niemals alle Faktoren berücksichtigen, die eine Vorhersage beeinflussen
- 32%: Glaube nicht, dass MU so genau ist wie andere Methoden
- 27%: Zu teuer
- 23%: Ineffizient
- 20%: MU kann die Intuition des Klinikers nicht ersetzen
- 19%: Die eigene klinische Erfahrung wird ignoriert
- 19%: Zu schwierig in der Anwendung

Vrieze, S. I., & Grove, W. M. (2009). Survey on the use of clinical and mechanical prediction methods in clinical psychology. *Professional Psychology: Research and Practice, 40*(5), 525–531.

Statistische Urteilsbildung als Spezialfall Mechanischer Urteilsbildung

Beispiel für Mechanische Urteilsbildung (im klassischen Sinne)

- Der Goldberg-Index ist ein klassisches Beispiel für eine empirisch ermittelte Verrechnungsvorschrift, die für eine mechanische Urteilsbildung genutzt werden kann, z.B.
- Entscheidungsregel: Goldberg-Index > 45 → Klassifikation „psychotisch“

Beispiel für Statistische Urteilsbildung

- Eine empirische Verrechnungsvorschrift kann aber auch mithilfe eines statistischen Vorhersagemodells konstruiert werden, z.B.
- Logistisches Regressionsmodell:
  - AV: Diagnose im standardisierten klinischen Interview: „neurotisch“ vs. „psychotisch“
  - UVs: Skalen (oder Items) im MMPI Fragebogen
- Entscheidungsregel: P(AV = „psychotisch“ | UVs) > 0.5
  → Klassifikation „psychotisch“

Herausforderungen Mechanischer Urteilsbildung:

- Mechanische (Statistische) Urteile können nur für Personen getroffen werden, bei denen alle Informationen vorliegen
  - Wenn beispielsweise ein Interview-Score in ein Gesamturteil von einem statistischen Urteilsmodell einfließt, benötigt man diese Information zwingend (außer man nutzt Methoden, die fehlende Werte ersetzen können)
  - Ein klinisches Urteil könnte auch mit weniger diagnostischen Informationen zurechtkommen
- Für die Konstruktion Mechanischer (Statistischer) Urteilsmodelle sind große Fallzahlen bei einheitlicher Fragestellung notwendig
- Einschlägige Forschungsergebnisse (und daraus ableitbare Verrechnungsvorschriften) oftmals nicht vorhanden

Handlungsempfehlungen

Vorteile und Chancen beider Urteilsmodelle nutzen:

- Diagnostikerinnen sollten mechanische (statistische) Urteilsmodelle kennen und für die Urteilsbildung heranziehen (objektiver, reliabler und scheinbar auch valider!)
- In begründeten Fällen kann das mechanische Urteil korrigiert oder durch ein klinisches Urteil ersetzt werden — etwa wenn Zweifel daran bestehen, ob das vorhandene Modell für den betrachteten Einzelfall anwendbar ist (z.B. im Modell unberücksichtigte Vorerkrankung, vermuteter Bias des Modells, etc.)
- Für diese Korrektur kann dann jede zusätzliche Information genutzt werden, die im Modell (noch) nicht enthalten ist

Handlungsempfehlungen

Kontinuierliche Optimierung statistischer Vorhersagemodelle notwendig:

- Empirisch validierte Modelle verwenden (im Gegensatz zu rein rational begründeten)
- Statistische Modelle kreuzvalidieren (d.h. auf neue Daten anwenden, die nicht zur Berechnung der Entscheidungsregel verwendet wurden), kontinuierlich überprüfen und weiterentwickeln (gilt auch für mechanische Urteilsbildung mit empirisch validiertem Index)
- Inhaltliche Nachvollziehbarkeit der verwendeten Modelle anstreben (soweit möglich)
- Moderne Modellierungsansätze nutzen (Stichwort „Maschinelles Lernen“, siehe Vorlesung im Master)

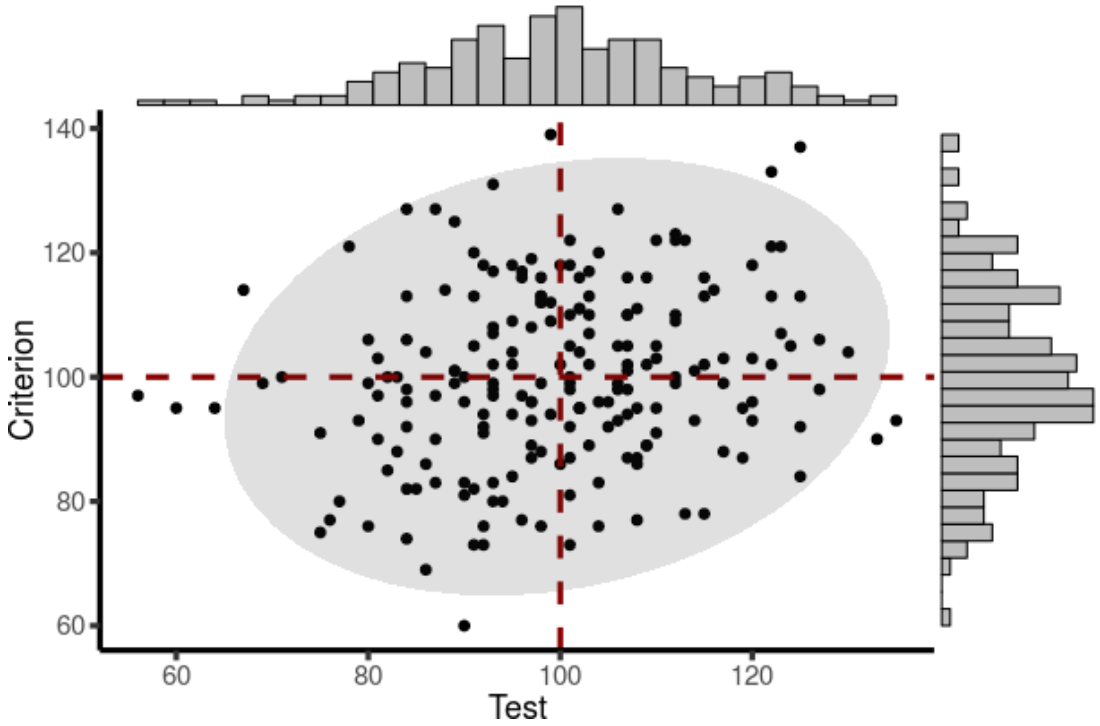
2. Fehler & Güte von Urteilen

Empirische Festlegung von Cut-Off Werten

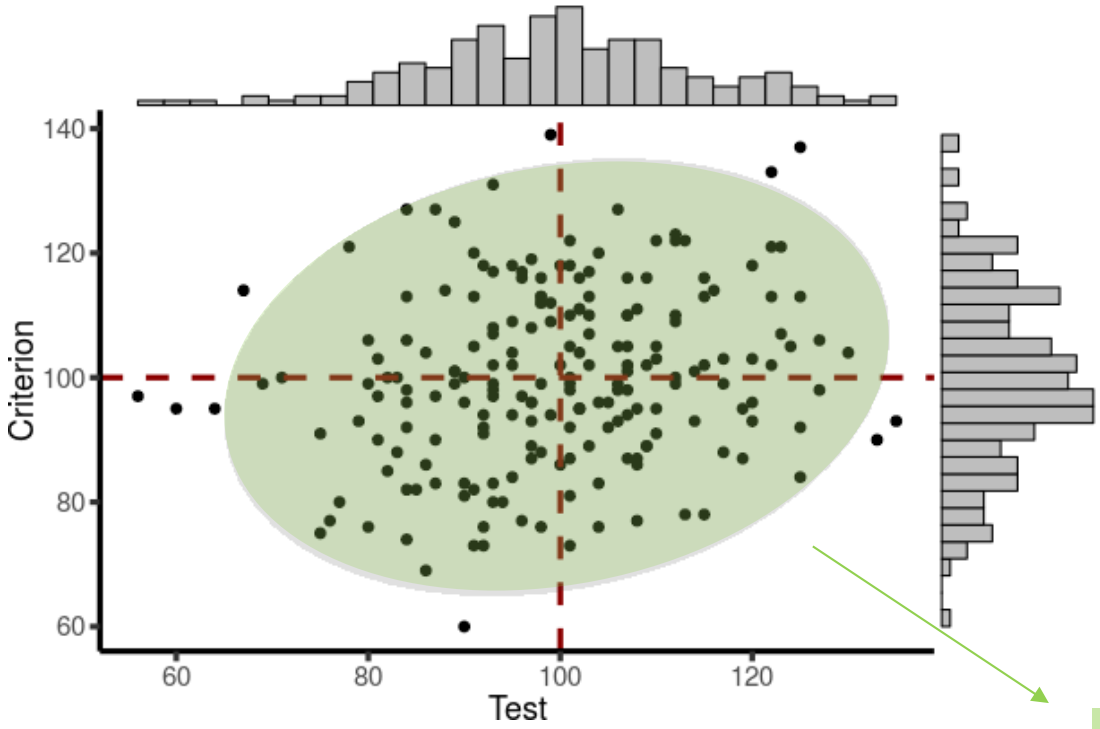
- Zentrales Anliegen diagnostischer Tätigkeit sollte sein, **Fehler bei diagnostischen Entscheidungen zu minimieren**
- Viele Entscheidungen beinhalten eine **Zuordnung auf (zwei) unterschiedliche Klassen**
  - Diagnose in der klinischen Psychologie: Störung ja/nein?
  - Einstellungsentscheidung im Unternehmen: Bewerberin un-/geeignet?

- Wichtige Aufgabe im diagnostischen Prozess:
  Festlegung geeigneter Cut-Off Werte
  - Dabei gilt es verschiedene Urteilsfehler zu berücksichtigen und gegeneinander abzuwägen
  - Diese lernen wir im Folgenden Schritt für Schritt kennen

Visualisierung der Klassenzuordnung



Visualisierung der Klassenzuordnung



Kriterium ≙ Goldstandard:
z.B. tatsächliches Verhalten, wirklicher (wahrer) Zustand

Test ≙ Urteil:
z.B. Testergebnis, diagnostische Entscheidung

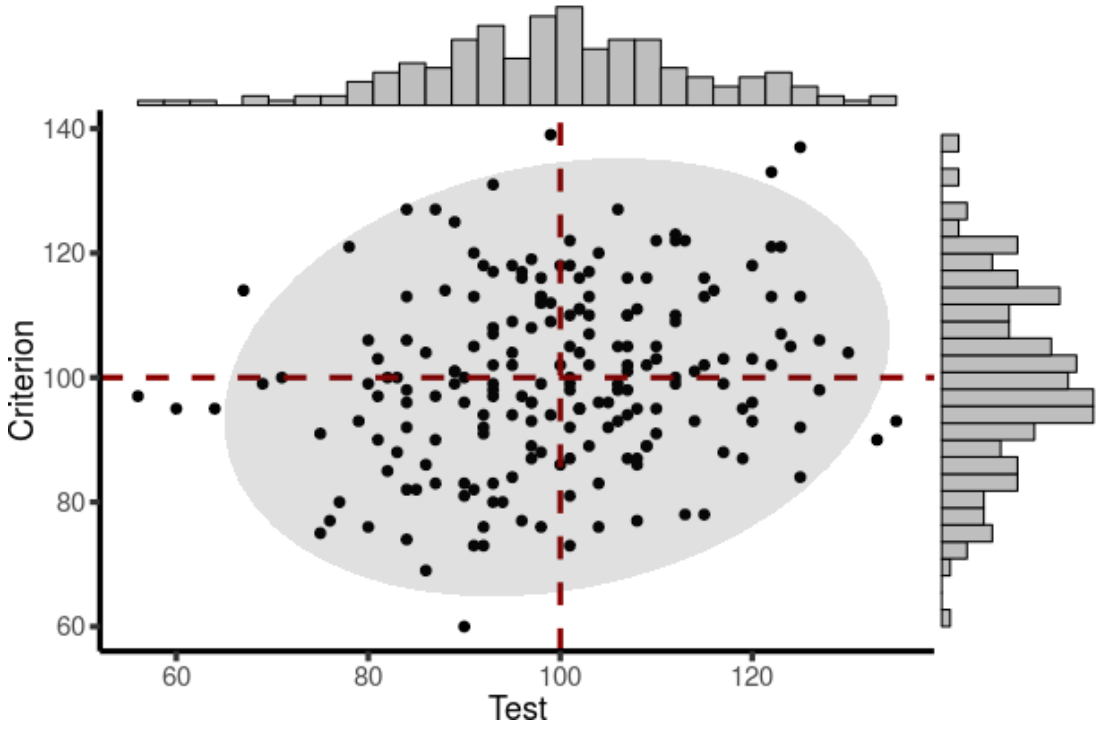
Validität des Tests / der eingesetzten Verfahren:
Je höher die Validität, desto schmaler und näher ist die Punktewolke an der Diagonalen (perfekte Validität → Korrelation r = 1)

Visualisierung der Klassenzuordnung

Einige (vereinfachte) Anwendungsbeispiele…

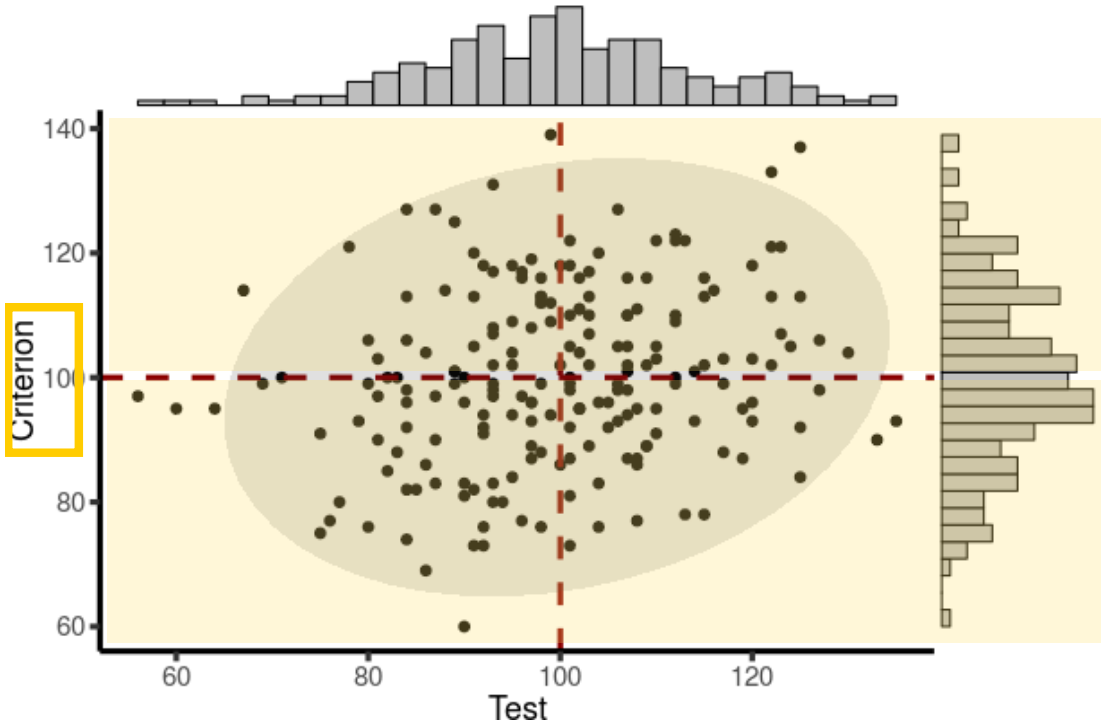
- Personalpsychologie:
  - Testwert: Score Gewissenhaftigkeit Selbstbericht (z.B. NEO-PI-R),
    Kriterium: Score Vorgesetztenbeurteilung nach der Probezeit
  - Testwert: Score im Intelligenztest (z.B. IST 5)
    Kriterium: Monatlicher Umsatz als Verkäuferin nach der Probezeit
- Klinische Psychologie:
  - Testwert: Score Narzissmus Selbstbericht (z.B. SINS)
    Kriterium: Score Narzisstische Persönlichkeitsstörung im Klinischen Interview (z.B. SCID-5-PD)
  - Testwert: Score Depressionsschwere Selbstbericht (z.B. QIDS-SR)
    Kriterium: Score Depressionsschwere Expertenrating (z.B. QIDS-C)

Das Kriterium – ein Goldstandard?



- Die Qualität der Überlegungen zu Urteilsfehlern hängt von der Qualität des Goldstandards ab
- Der Goldstandard (und damit verbunden oft auch der Cut-Off im Kriterium) wird als „Wahrheit“ betrachtet – ob das immer zutreffend ist, muss von Fall zu Fall entschieden werden
- Für alle folgenden Überlegungen nehmen wir an, dass das Kriterium sinnvoll gewählt wurde

Visualisierung der Klassenzuordnung



Realität

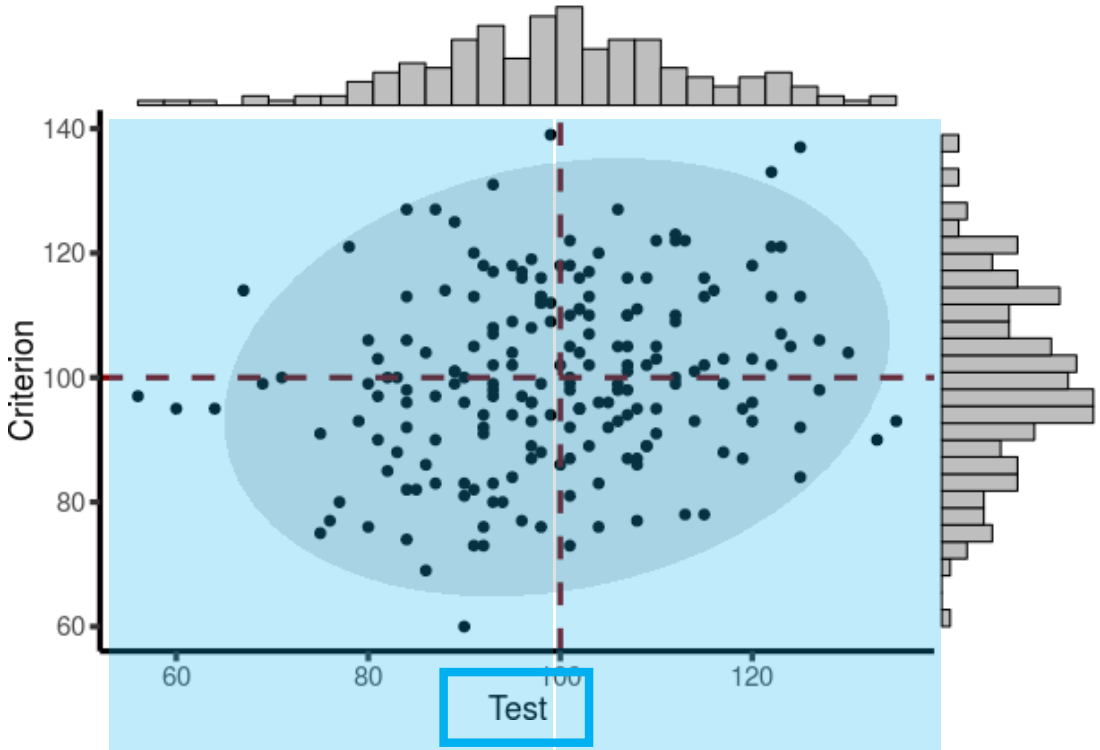
„positiver Zustand“, z.B.
tatsächliches Vorhandensein einer Krankheit
→ d.h. kranke Personen,
Tatsächliche Eignung
→ d.h. geeignete Personen

„negativer Zustand“, z.B.
tatsächliche Abwesenheit einer Krankheit
→ d.h. gesunde Personen,
Tatsächlich keine Eignung
→ d.h. ungeeignete Personen

→ „positiv“ im Sinne von „Vorhanden“, *nicht* im Sinne von „wünschenswert“

→ „negativ“ im Sinne von „Nicht-Vorhanden“, *nicht* im Sinne von „unerwünscht“

Visualisierung der Klassenzuordnung

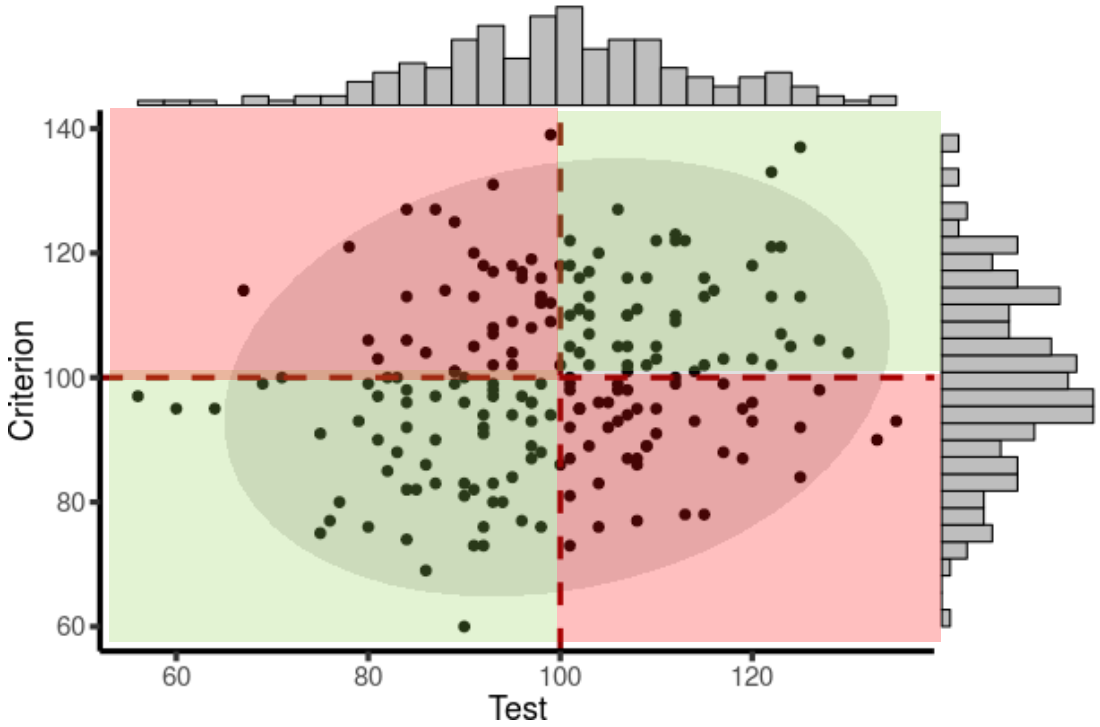


„negatives Urteil“, z.B. keine Krankheits-Diagnose (negative Diagnose), keine Einstellung

„positives Urteil“, z.B. Krankheits-Diagnose (positive Diagnose), Einstellung

Diagnostisches Urteil

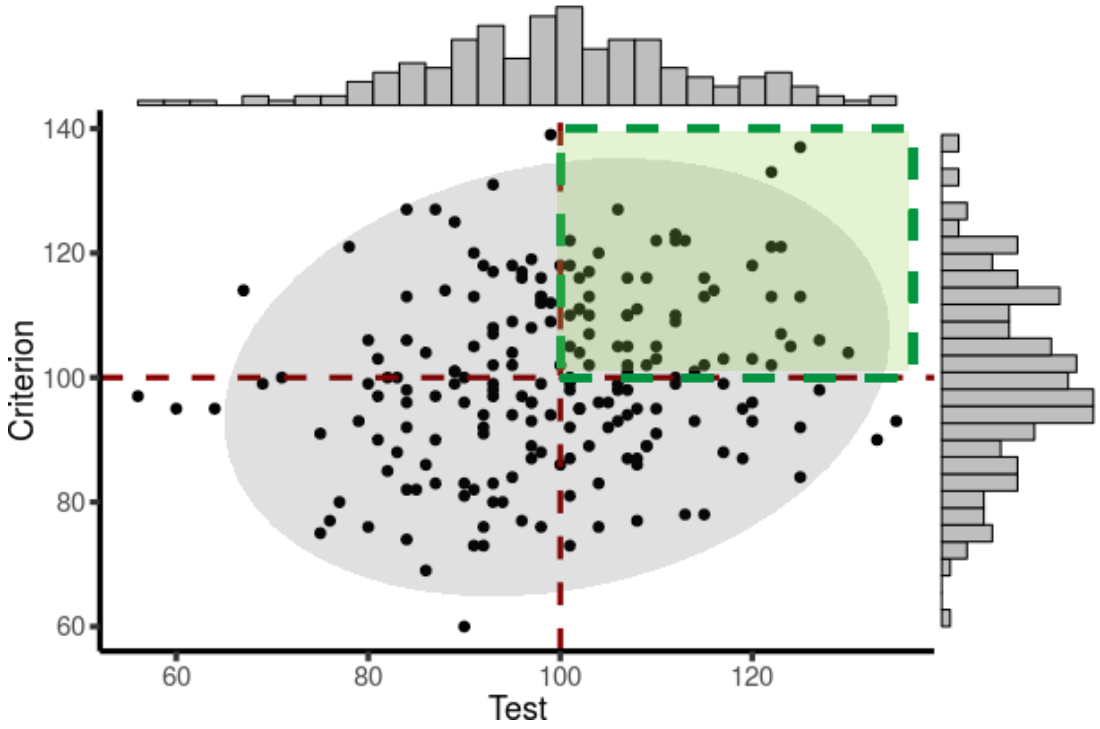
Visualisierung der Klassenzuordnung



- Man kann sich nun vier Kategorien der (Nicht-) Übereinstimmung anschauen
- Jede Einzelentscheidung fällt in eine dieser vier Kategorien

Das Prinzip kommt uns bekannt vor von Beobachter-/Beurteilerübereinstimmung!

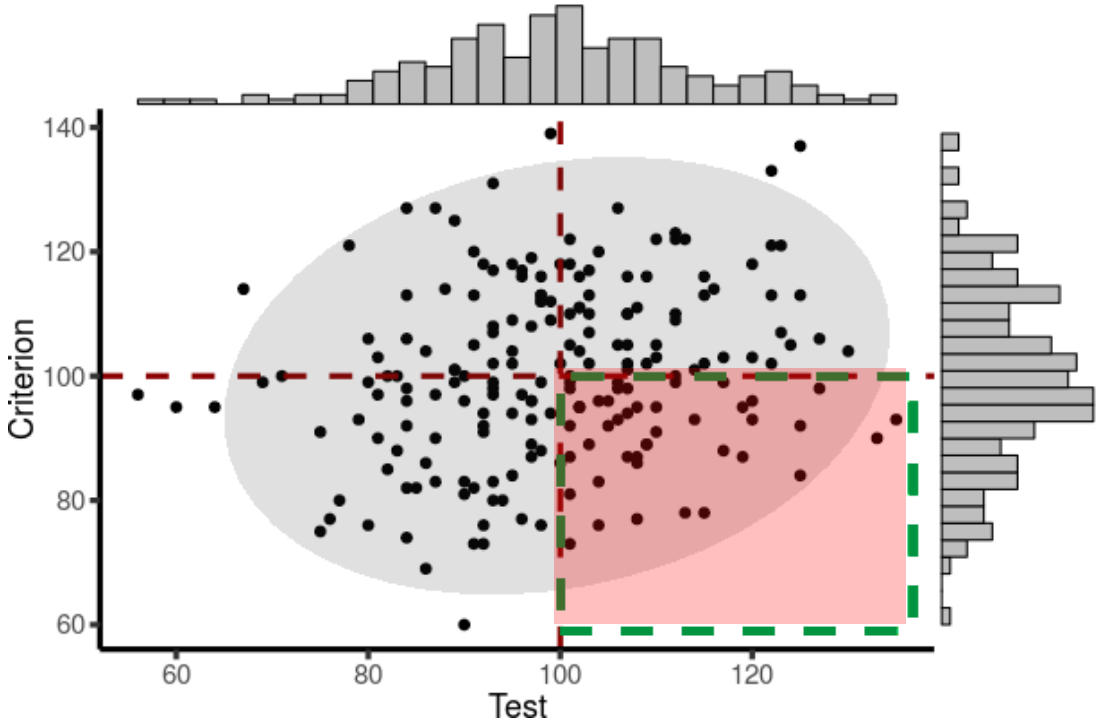
Richtig Positive



- Engl.: **True Positive (TP)**
- Auch: „Hit“
- **Richtige Zuordnung:** im Test und Kriterium jeweils positiv

- Beispiele:
  - korrekterweise als krank identifizierte Kranke
  - korrekterweise als geeignet identifizierte Geeignete

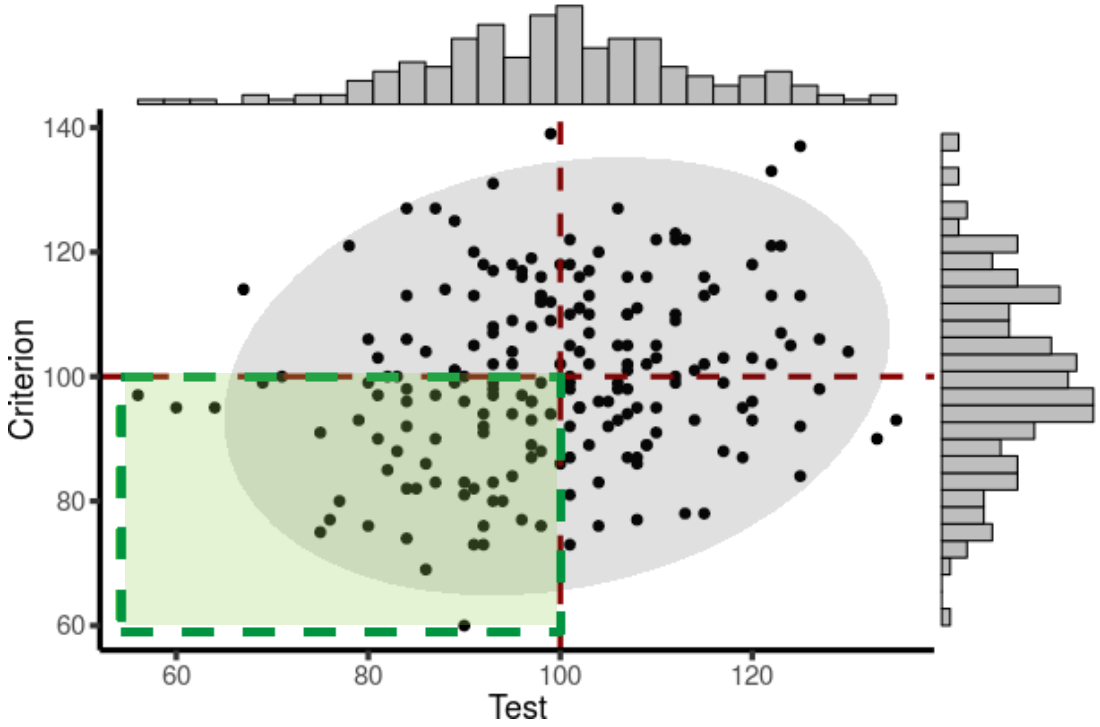
Falsch Positive



- Engl.: **False Positive (FP)**
- Auch: „False Alarm“
- **Falsche Zuordnung:** im Test positiv, im Kriterium negativ

- Beispiele:
  - fälschlicherweise als krank bezeichnete Gesunde
  - fälschlicherweise als geeignet bezeichnete Ungeeignete

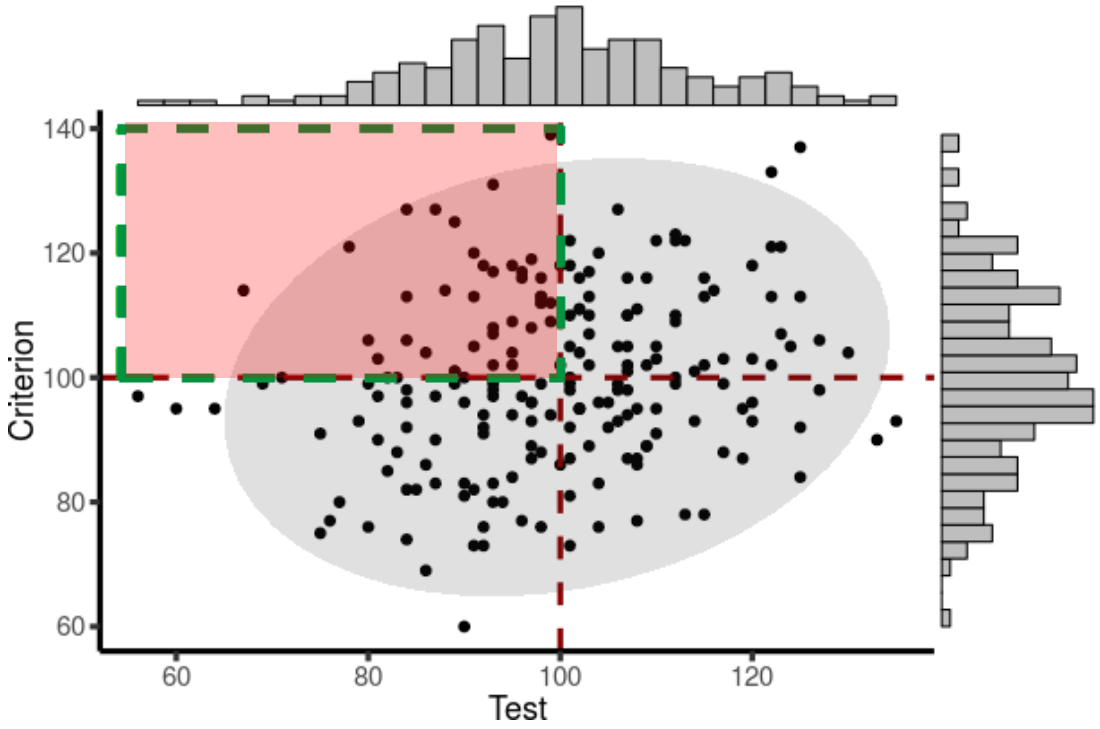
Richtig Negative



- Engl.: **True Negative (TN)**
- Auch: „Correct Rejection“
- **Richtige Zuordnung:** im Test und Kriterium jeweils negativ

- Beispiele:
  - korrekterweise als gesund identifizierte Gesunde
  - korrekterweise als ungeeignet identifizierte Ungeeignete

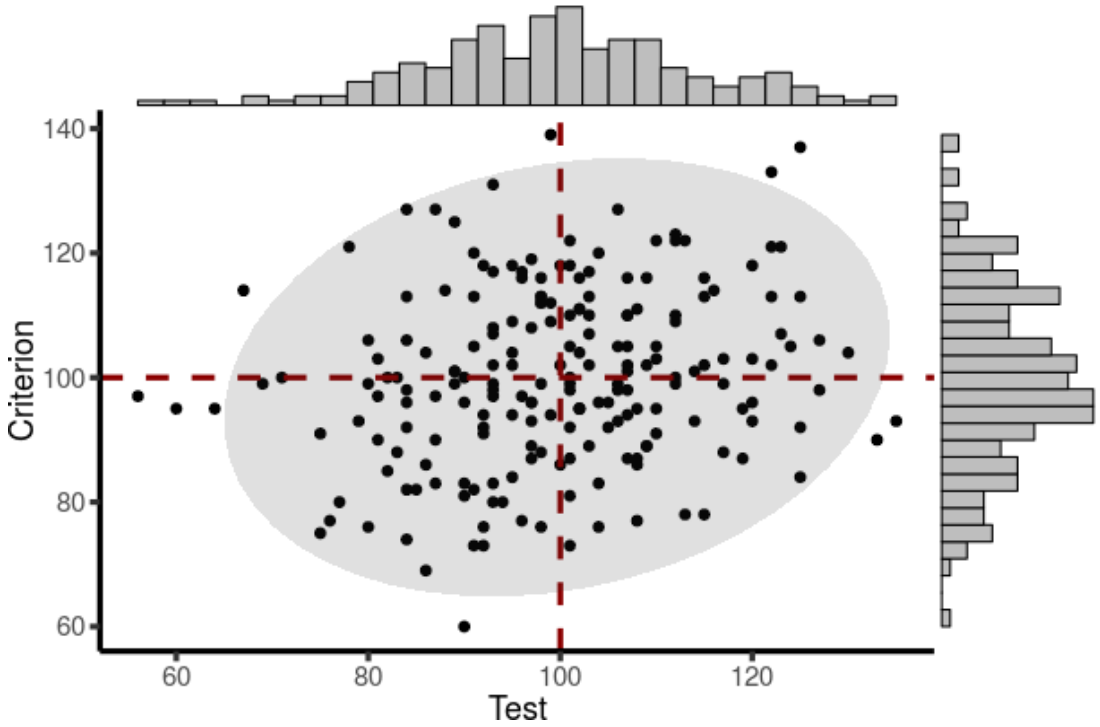
Falsch Negative



- Engl.: **False Negative (FN)**
- Auch: „Miss“
- **Falsche Zuordnung:** im Test negativ, im Kriterium positiv

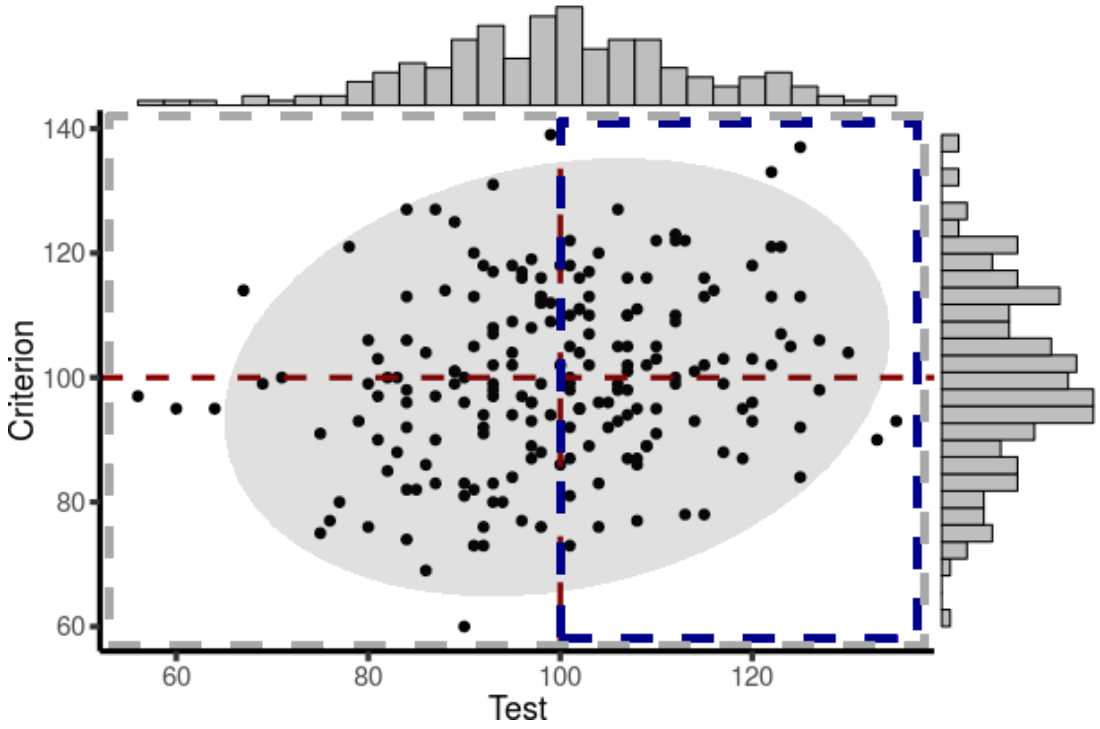
- Beispiele:
  - fälschlicherweise als gesund bezeichnete Kranke
  - fälschlicherweise als ungeeignet bezeichnete Geeignete

Wahrscheinlichkeiten und Kennwerte



- Man kann sich nun basierend auf der Häufigkeit der Fälle in bestimmten Kategorien die Wahrscheinlichkeit für verschiedene Situationen anschauen
- Je größer die Stichprobe, desto genauer die Schätzungen der Kennwerte

Selektionsrate

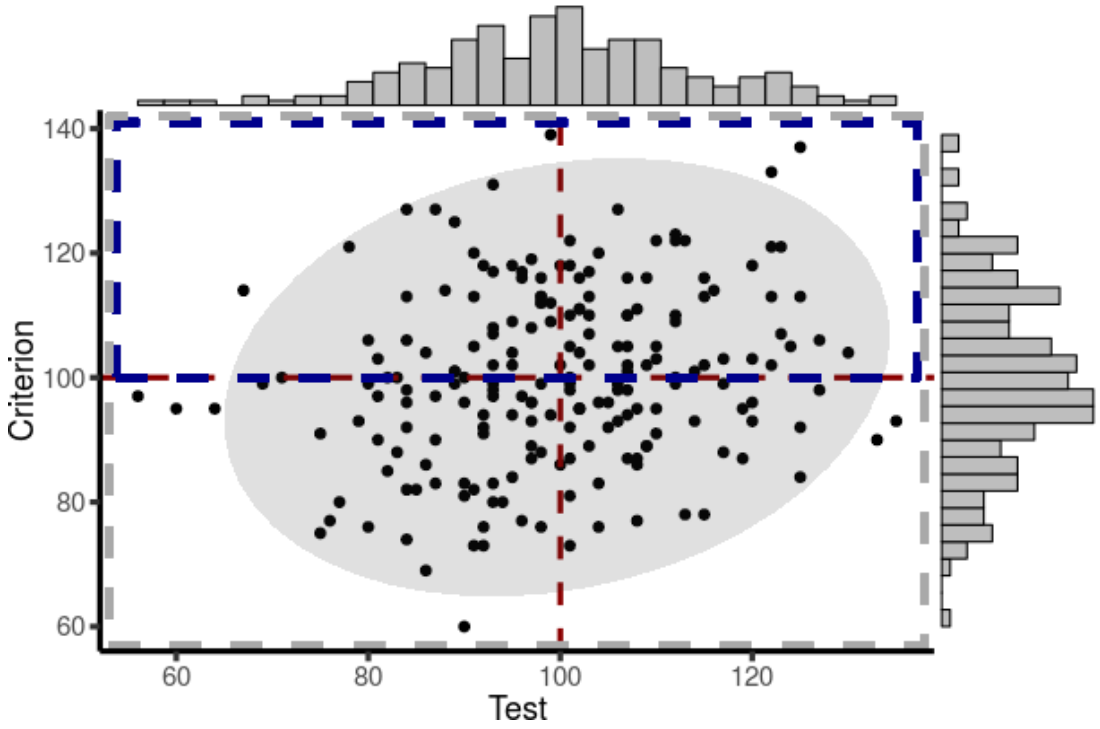


- Engl.: **Selection ratio**
- Abhängig vom Cut-Off-Wert

- Beispiele:
  - Anteil der als krank bezeichneten Personen an allen Personen
  - Anteil der eingestellten Personen an allen Personen

$$\textbf{P(positives Urteil)} = \frac{(TP+FP)}{N}$$

Prävalenz/Basisrate



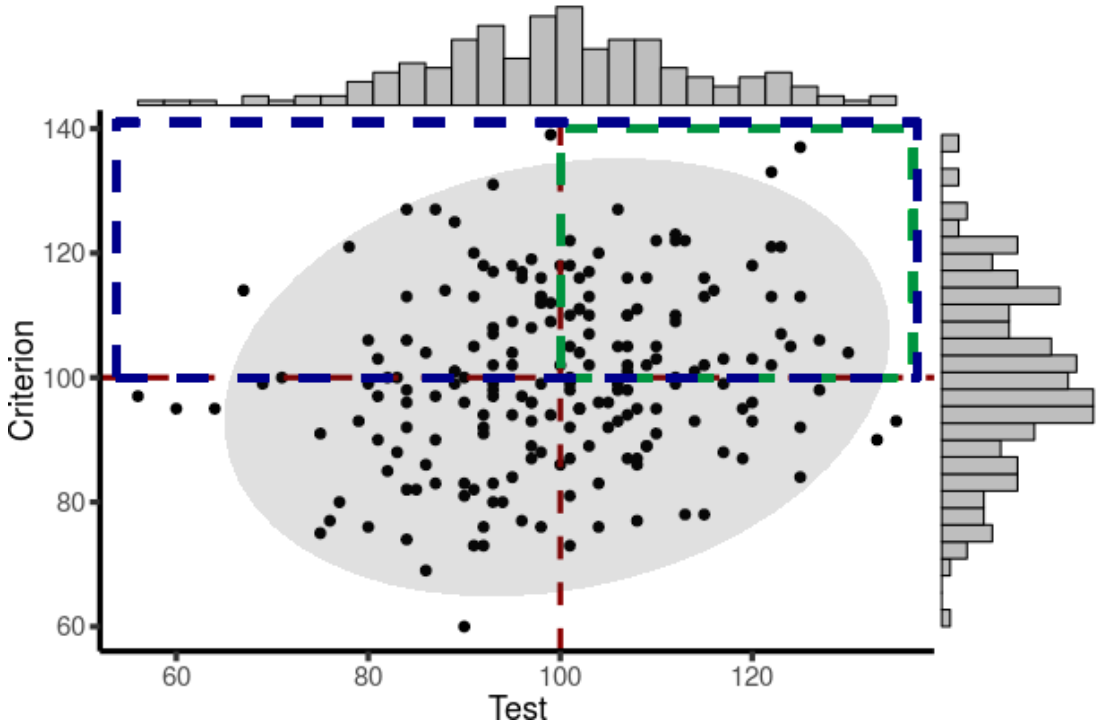
- Engl.: **Prevalence**
- Anteil der Personen in der Population mit einem positiven Zustand
- Auch: “Natürlicher Eignungskoeffizient” („success without use of test“) → wie gut (Anzahl korrekter Diagnosen) wäre eine Klassifikation ohne Test

- Beispiele:
  - Anteil der kranken Personen in der Population
  - Anteil der geeigneten Personen in der Population

$$\textbf{P(positiver Zustand)} = \frac{(TP+FN)}{N}$$

(vgl. Schmidt-Atzert & Amelang, 2012, S.419)

Güte: Sensitivität

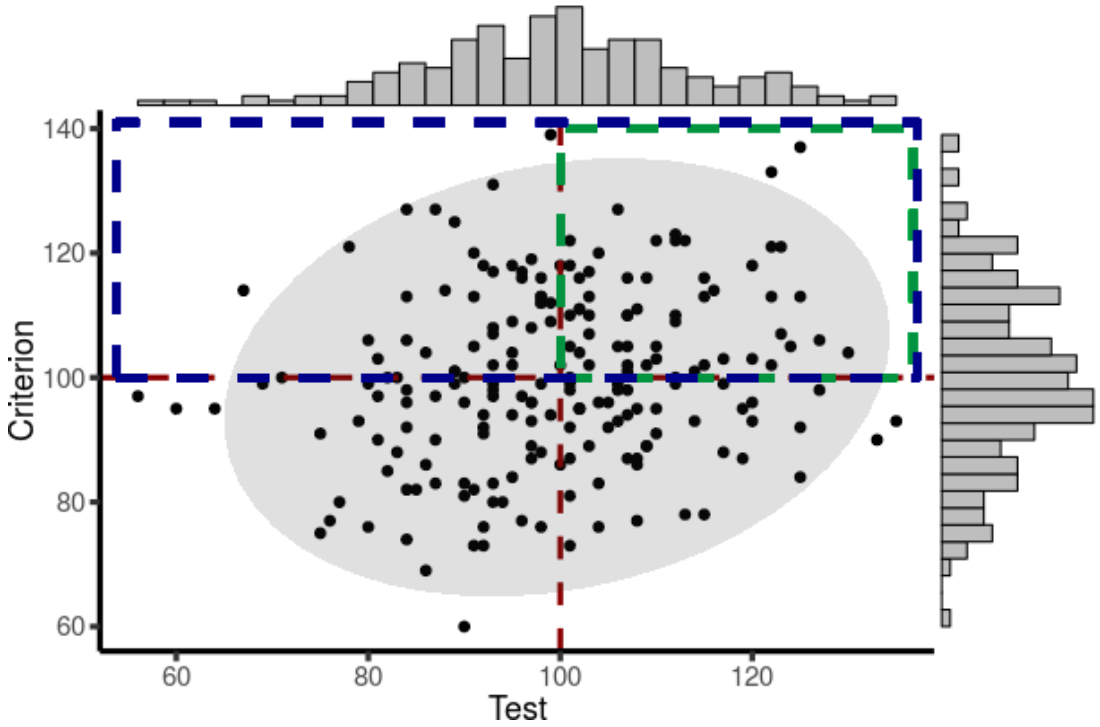


- Engl.: **Sensitivity**/Hit Rate
- Auch: „Trefferquote“
- Eigenschaft des Verfahrens: *Wahrscheinlichkeit, mit der ein vorliegender positiver Zustand als solcher erkannt wird*
- z.B. Anteil der richtig identifizierten Kranken in der Gruppe der Kranken

$$\textbf{P(positives Urteil | positiver Zustand)} = \frac{TP}{(TP+FN)}$$

Sprich: „Wahrscheinlichkeit für ein positives Urteil gegeben eines positiven Zustands“

Güte: Sensitivität

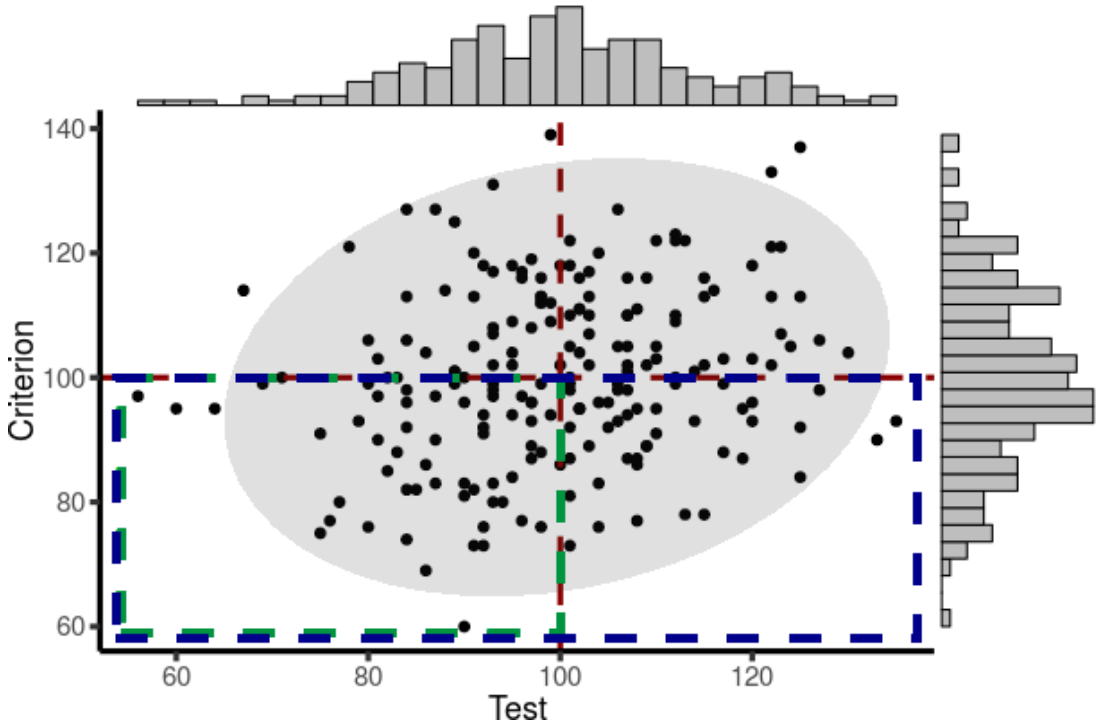


- Hohe Sensitivität
  → Wenige False-Negatives!
- Vorsicht: In Ziegler & Bühner (2012), S.133ff werden die Begriffe „Trefferquote“ bzw. „Hit Rate“ fälschlicherweise für den positiven Prädiktionswert (siehe später) statt der Sensitivität verwendet.

$$\text{P(positives Urteil | positiver Zustand)} = \frac{TP}{(TP+FN)}$$

(vgl. Schmidt-Atzert & Amelang, 2012, S.417)

Güte: Spezifität

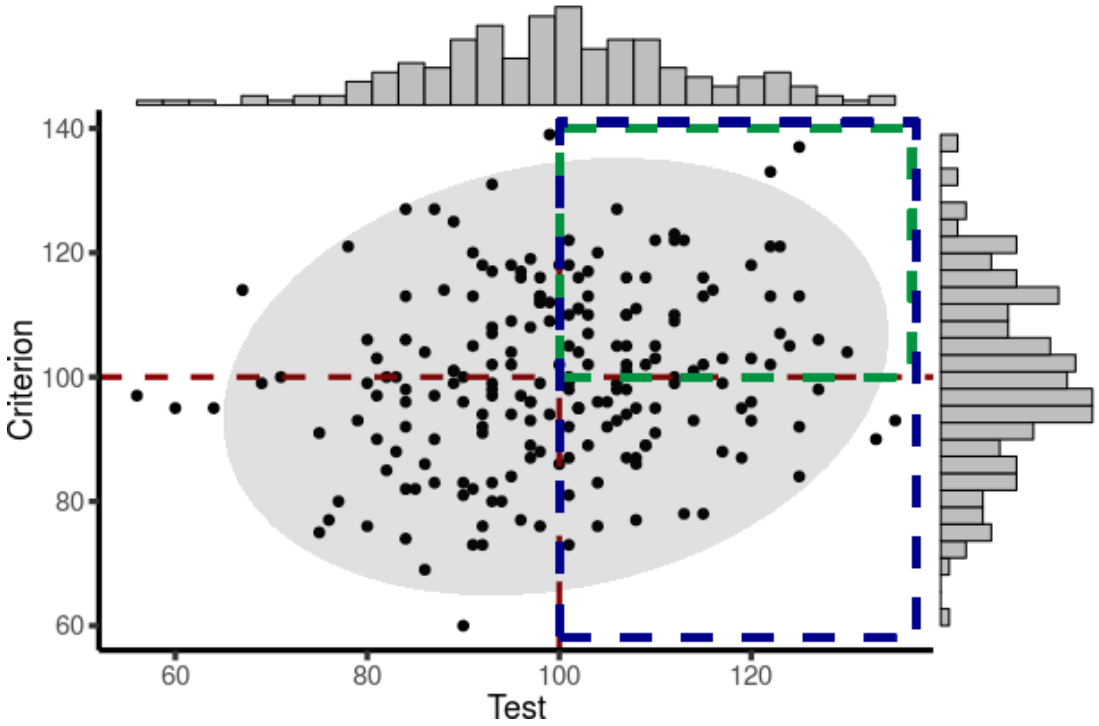


- Engl.: **Specificity**
- Eigenschaft des Verfahrens: *Wahrscheinlichkeit, mit der ein vorliegender negativer Zustand als solcher erkannt wird*
- z.B. Anteil der richtig identifizierten Gesunden in der Gruppe der Gesunden
- Hohe Spezifität
  → Wenige False-Positives!

$$\textbf{P(negatives Urteil | negativer Zustand)} = \frac{TN}{(FP+TN)}$$

(vgl. Schmidt-Atzert & Amelang, 2012, S.418)

Güte: Positiver Prädiktionswert (PPV)

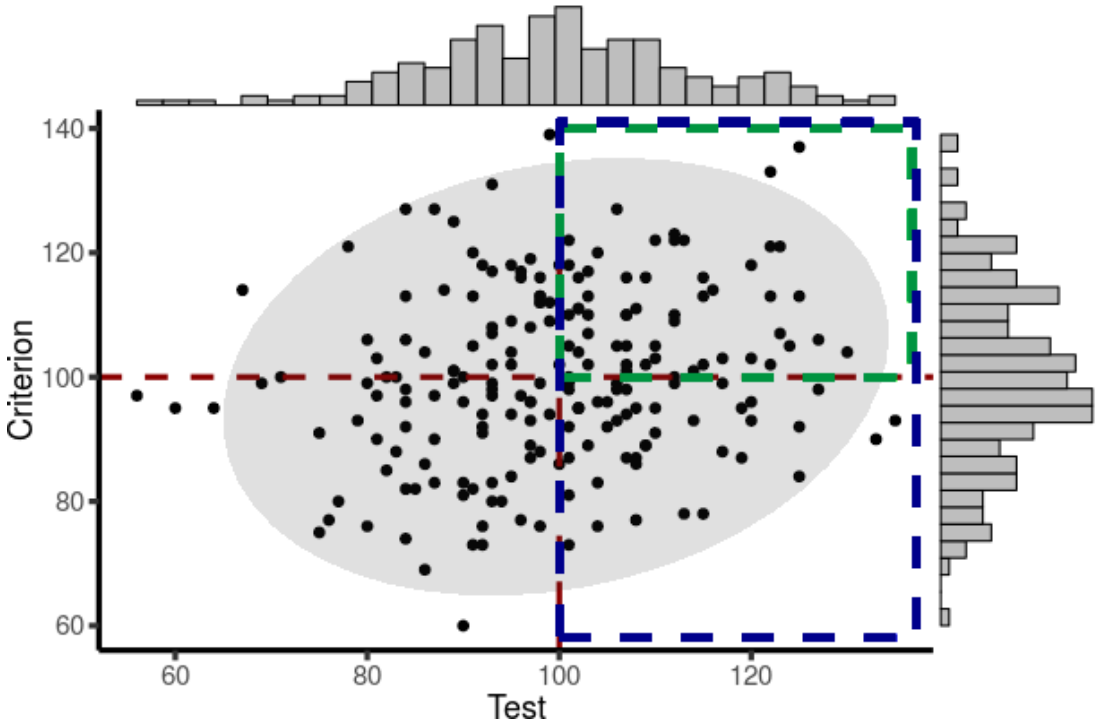


- Engl.: **Positive predictive value**, **Precision**
- Hängt u.a. von Prävalenz ab
- Gibt Aufschluss über Relevanz des Verfahrens: *Wahrscheinlichkeit, mit der ein positives Urteil zutreffend ist*
- z.B. Anteil der richtig identifizierten Kranken an allen als krank bezeichneten Personen

$$\textbf{P(positiver Zustand | positives Urteil)} = \frac{TP}{(TP+FP)}$$

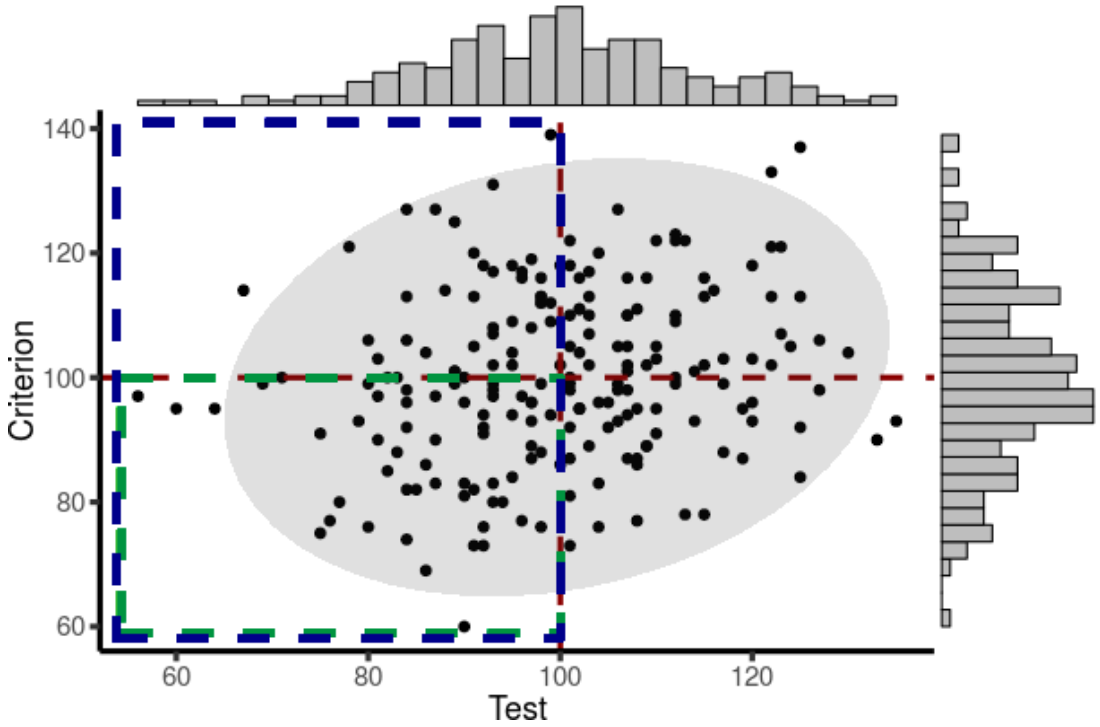
(vgl. Schmidt-Atzert & Amelang, 2012, S.418)

Güte: Positiver Prädiktionswert (PPV)



Vorsicht: In mancher Literatur wird der positive Prädiktionswert auch „Erfolgsrate“ genannt. Aufgrund der Verwechslungsgefahr mit „Trefferquote“ bzw. noch stärker mit der Übersetzung von „Hit Rate“ verwenden wir diesen Begriff hier bewusst nicht (auch da manche Autorinnen diese Begriffe synonym und nicht für zwei verschiedene Maße verwenden). **Bitte verwenden Sie daher (auch im Hinblick auf die Klausur) bitte nicht den Begriff Erfolgsrate, da je nach Kontext unklar sein kann, was damit gemeint ist.**

Güte: Negativer Prädiktionswert (NPV)

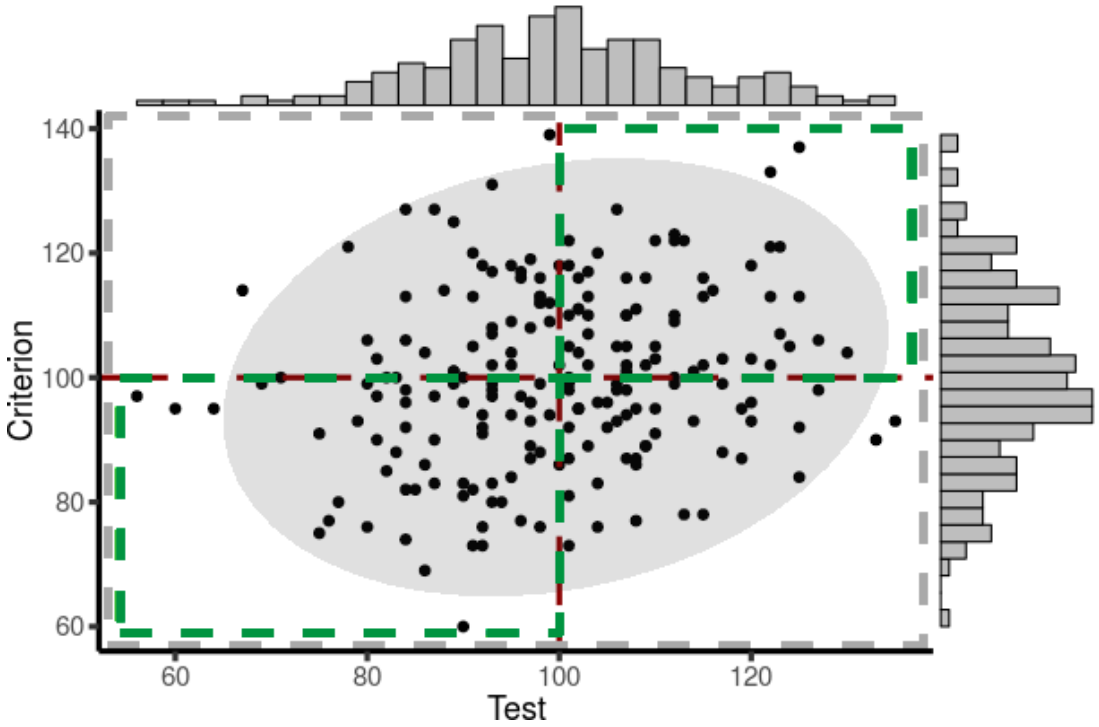


- Engl.: **Negative predictive value**
- Hängt u.a. von Prävalenz ab
- Gibt Aufschluss über Relevanz des Verfahrens: *Wahrscheinlichkeit, mit der ein negatives Urteil zutreffend ist*
- z.B. Anteil der richtig identifizierten Gesunden an allen als gesund bezeichneten Personen

$$\textbf{P(negativer Zustand | negatives Urteil)} = \frac{TN}{(FN+TN)}$$

(vgl. Schmidt-Atzert & Amelang, 2012, S.418)

Güte: Anzahl korrekter Diagnosen (AKD)



- Engl.: **Accuracy**
- Auch: „Genauigkeit“, „Prozentuale Übereinstimmung“ (siehe Beobachterübereinstimmung)
- *Wahrscheinlichkeit, mit der ein gefälltes Urteil zutreffend ist*
- z.B. Anteil der richtigerweise als krank bzw. gesund (geeignet bzw. ungeeignet) identifizierten Personen an allen Personen

- **Problem: Gibt keinen Aufschluss über Unterschiede zwischen negativen / positiven Urteilen**

$$\textbf{P(richtiges Urteil)} = \frac{(TP+TN)}{N}$$

Wann beurteile ich was?

- Ob **Sensitivität vs. Spezifität** bzw. **PPV vs. NPV** wichtig(er) sind, hängt davon ab, **welcher Fehler** im Vordergrund steht und minimiert werden soll
  - False-Negatives → Sensitivität bzw. NPV
  - False-Positives → Spezifität bzw. PPV
- Ob ich mich für die **Eigenschaften** des Verfahrens (Sensitivität, Spezifität) oder die **Relevanz** des Verfahrens (PPV, NPV) interessiere, hängt davon ab, ob ich…
  - …eine generelle Aussage über die Qualität des Verfahrens machen will (Sensitivität, Spezifität)
  - …auf Basis eines konkreten Ergebnisses einen Rückschluss auf ein Merkmal machen und dabei die Prävalenz mitberücksichtigen will (PPV, NPV)

Exkurs: Kennen wir das nicht aus Statistik I und II?

False Positive ≈ Fehler 1. Art = Alpha-Fehler?

False Negative ≈ Fehler 2. Art = Beta-Fehler?

- Im Prinzip, ja, und taucht auch in der Literatur so auf.
- Aber: Die Begrifflichkeiten Alpha-Fehler und Beta-Fehler werden sinnvollerweise nur im Kontext von expliziten Hypothesen verwendet (fälschliche Ablehnung oder fälschliche Beibehaltung der Nullhypothese)

→ Da wir hier keine expliziten Hypothesen vorliegen haben, verwenden wir diese Begriffe bei der Bewertung von Urteilen eher nicht.

3. Optimierung einer Auswahlentscheidung

Optimierung einer Auswahlentscheidung

Art und Umfang der Fehler einer Auswahlentscheidung (gilt für fast alle Arten von diagnostischen Entscheidungen) hängen von mehreren Faktoren ab, die teilweise nur eingeschränkt (direkt) beeinflussbar sind:

(1) Validität

(2) Basisrate

(3) Cut-Off Werte

Taylor, H. C., & Russell, J. T. (1939). The relationship of validity coefficients to the practical effectiveness of tests in selection: discussion and tables. *Journal of Applied Psychology, 23*(5), 565–578.

Um diese Faktoren zu veranschaulichen, verwenden wir im folgenden eine Shiny App, die mit simulierten Daten arbeitet:

http://shinyapps.org/apps/diagnostic_decisions/

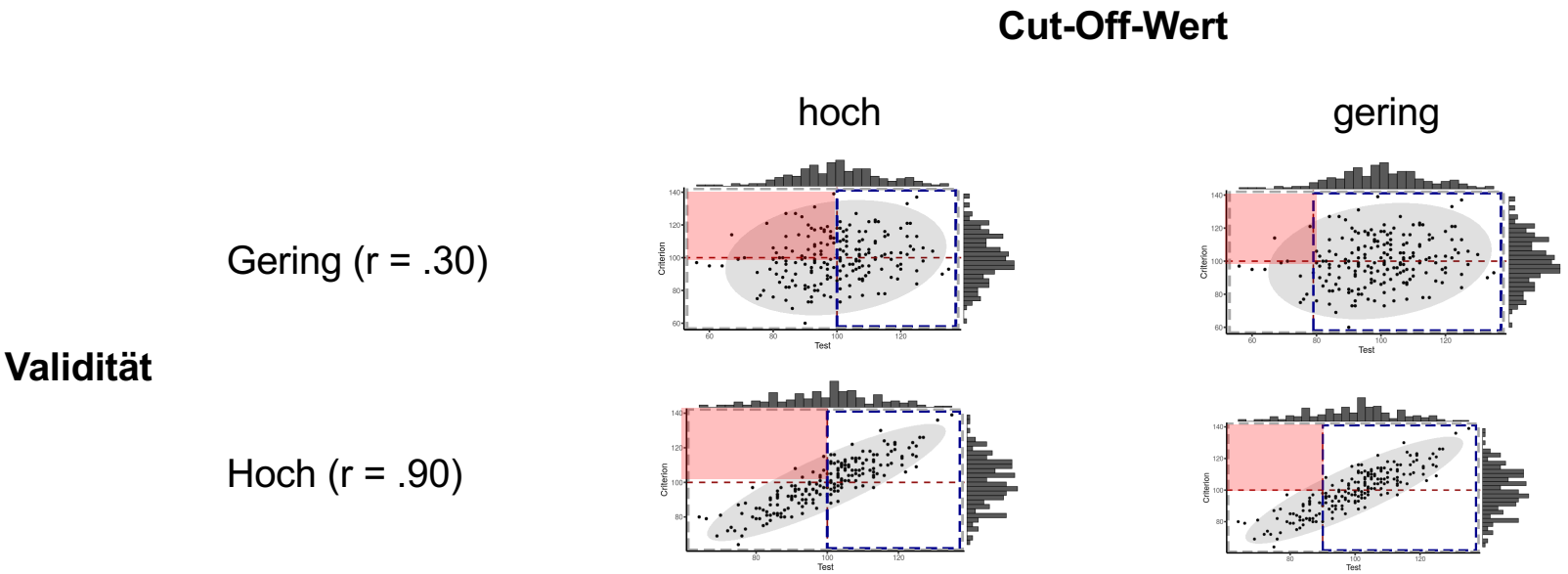
(1) Validität

- Eine höhere Validität der eingesetzten Verfahren verringert alle Arten von Fehlentscheidungen
- Eine höhere Validität verbessert daher prinzipiell erst einmal alle Gütemaße von Urteilen
- **Exkurs:** Damit bei einer Validität (= Korrelation Testwert mit Kriterium) von 1 keine Fehlentscheidungen mehr möglich sind, muss auch der Cut-Off-Wert des Verfahrens zum Cut-Off-Wert des Kriteriums passen.

→ Ggf. durch veränderte Datenerhebung (anderes Instrument) zu verbessern

(2) Basisrate

- Bei einer mehrstufigen Entscheidungsstrategie haben die ersten Entscheidungen maßgeblichen Einfluss auf die Basisrate für nachfolgende Entscheidungen
- Vor allem bei geringer Validität der ersten Entscheidungen ist (oft) eine hohe Selektionsrate (d.h. ein geringer Cut-Off-Wert) wünschenswert, um den Anteil der FN gering zu halten; zur Veranschaulichung:



(vgl. Schmidt-Atzert & Amelang, 2012, S.418f)

(2) Basisrate

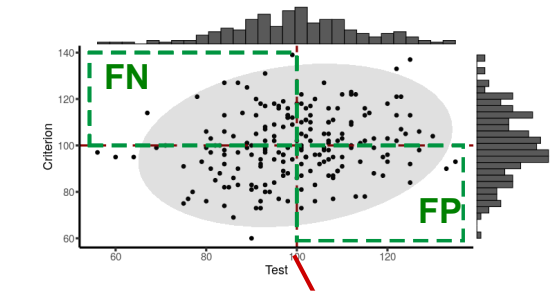
- Eine höhere Basisrate geht mit einem höheren PPV einher, aber einem niedrigeren NPV
- Eine niedrige Basisrate geht mit einem höheren NPV einher, aber einem niedrigeren PPV
- Sensitivität und Spezifität sind (entgegen manchen pauschalen Aussagen in der Literatur) **nicht** im Allgemeinen von der Basisrate unabhängig. Die Details sind aber kompliziert und werden hier nicht im Detail behandelt.
- Bei extrem geringen und extrem hohen Basisraten kann der Einsatz von Testverfahren mit geringer / moderater Validität (insgesamt) mehr Fehler (1 - AKD) bedeuten als eine Zufallsauswahl!

→ Ggf. durch gutes Personalmarketing und mehrstufige Entscheidungsstrategie beeinflussbar; im klinischen Bereich kaum (kurzfristig) beeinflussbar

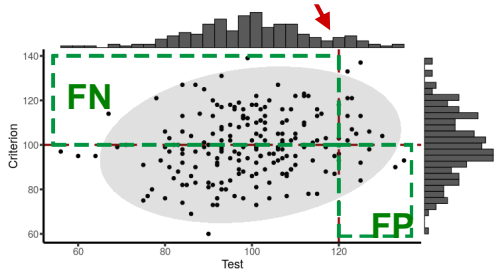
(vgl. Schmidt-Atzert & Amelang, 2012, S.418f)

(3) Testtrennwert („Cut-Off“-Wert)

- Ein höherer Testtrennwert (→ geringere Selektionsrate) geht mit einem höheren PPV und einer höheren Spezifität einher
- Ein niedrigerer Testtrennwert (→ höhere Selektionsrate) geht mit einem höheren NPV und einer höheren Sensitivität einher
- Zur Veranschaulichung:



Verschiebung des Cut-Off-Werts:
100 → 120



→ Mehr FN + weniger FP
→ Höherer Cut-Off führt zu
→ höherem PPV und Spezifität
→ niedrigerem NPV und Sensitivität

→ Veränderung: i.d.R. leicht anpassbar

Wie sollte man den Cut-Off Wert wählen, wenn dieser gegenläufige Effekte auf verschiedene Maße (PPV/Spezifität vs. NPV/Sensitivität) hat?

→ Trade-Off mit Hilfe der **„Receiver Operating Characteristic" = ROC Kurve**

- Hilft bei der Suche nach „optimalen" Cut-Off-Werte, die eine wünschenswerte Gewichtung von Sensitivität und Spezifität vornehmen
- ROC basiert auf der Berechnung von Sensitivität und Spezifität für verschiedene Cut-Off Werte

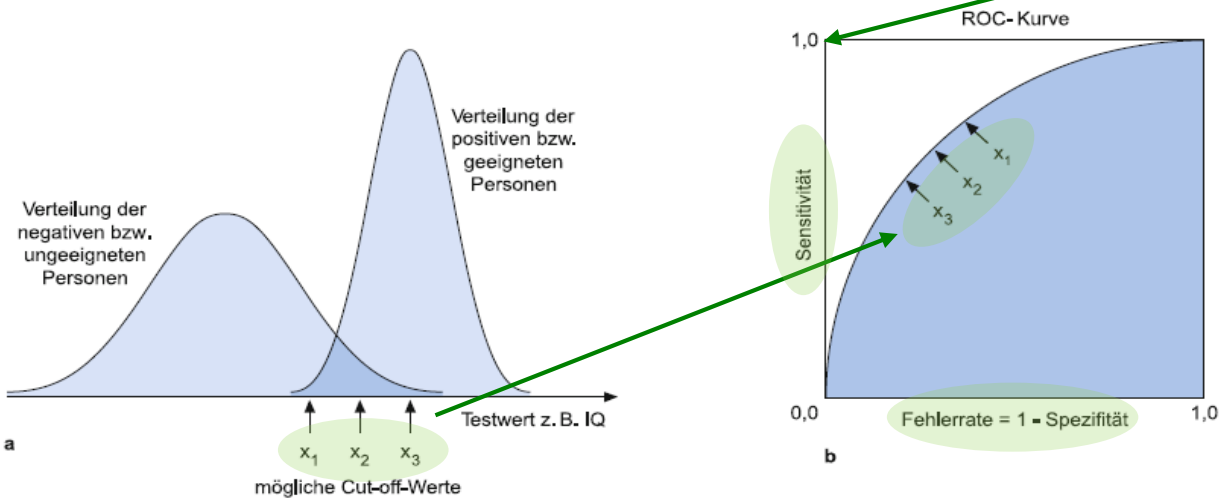
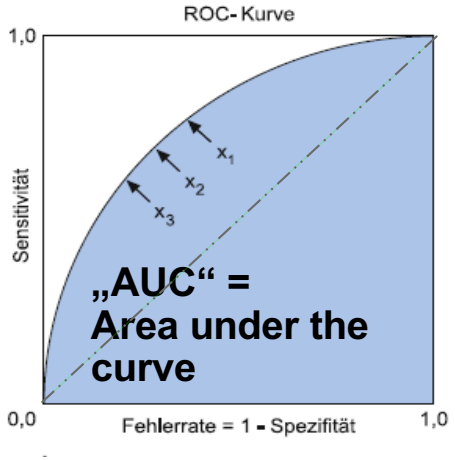


Abb. 6.10 Verteilungen von zwei Gruppen unterschiedlicher Eignung (a) für die eingetragenen Optionen möglicher Trennwerte (b) und ROC-Kurve. (Aus Noack & Petermann, 1992, S. 300. © 1988 Beltz Psychologie in der Verlagsgruppe Beltz, Weinheim/Basel)

Hier würde sich ein perfekt valider Test mit passendem Cutoff-Wert befinden: maximale Sensitivität bei maximaler Spezifität

(Schmidt-Atzert & Amelang, 2012, S.422; Ziegler & Bühner, 2012, S.148)

Ein weiteres Gütekriterium: „Area under the (ROC) curve“ = AUC



- Je weiter weg die Kurve von der Diagonalen ist (AUC = 0.5 → „Zufall“), desto höher die AUC und desto besser ist die Güte des Verfahrens
- = „Diskriminationsfähigkeit“ des Verfahrens: Betrachtet wird dabei die Kombination aus Sensitivität + Spezifität
- AUC ist sinnvoll wenn verschiedene Tests (Testwerte) verglichen werden sollen, aber der konkrete Cut-Off, der für die Testwerte in der Praxis bei verwendet werden noch nicht feststeht oder variieren kann

(Schmidt-Atzert & Amelang, 2012, S.422; Ziegler & Bühner, 2012, S.148)

Ein weiteres Gütekriterium: „Area under the (ROC) curve“ = AUC

- **Theoretischer Wertebereich:** 0 – 1 (wobei Werte < 0.5 für eine mögliche Umpolung von „positiv“ und „negativ“ sprechen würden)
- **Interpretation:** Zieht man zufällig eine Person aus der „positiven“ und eine Person aus der „negativen“ Klasse, entspricht die AUC der Wahrscheinlichkeit, dass die „positive“ Person einen höheren Testwert aufweist als die „negative“ Person.

Comparing Effect Sizes in Follow-Up Studies: ROC Area, Cohen's *d*, and *r*

Marnie E. Rice[1,2] **and Grant T. Harris**[1]

In order to facilitate comparisons across follow-up studies that have used different measures of effect size, we provide a table of effect size equivalencies for the three most common measures: ROC area (AUC), Cohen's d, and r. We outline why AUC is the preferred measure of predictive or diagnostic accuracy in forensic psychology or psychiatry, and we urge researchers and practitioners to use numbers rather than verbal labels to characterize effect sizes.

KEY WORDS: effect size; ROC area; risk assessment; predictive accuracy.

AUC	*d*	r_{pb}
.709	.778	.362
.712	.792	.368
.714	**.800**	**.371**
.716	.806	.374
.719	.820	.379
.722	.834	.385
.726	.849	.391
.729	.863	.396
.732	.877	.402
.736	.891	.407
.739	.905	.412
.742	.919	.418
.745	.933	.423

Table 1. The Relationships Among AUC, and Two More Longstanding Measures of Effect Size, Cohen's *d*, and the Point-Biserial Correlation Coefficient, r_{pb}

(Schmidt-Atzert & Amelang, 2012, S.422; Ziegler & Bühner, 2012, S.148)

- Die hier besprochenen Maßzahlen (Sensitivität, Spezifität, PPV, NPV, Accuracy, AUC) werden auch verwendet, um die Vorhersagegüte von statistischen Vorhersagemodellen mit binärer AV zu quantifizieren (siehe Vorlesung zum Maschinellen Lernen im Master).
- Beispiel: Logistisches Regressionsmodell
  - AV: Diagnose im standardisierten klinischen Interview: „neurotisch“ vs. „psychotisch“
  - UVs: Skalen (oder Items) im MMPI Fragebogen
  - Cut-Off: P(AV = „psychotisch“ | UVs) > 0.5 → Vorhersage „psychotisch“
- In diesem Kontext ergibt sich eine Vierfeldertafel basierend auf den möglichen Kombinationen der Variable *AV: „neurotisch“ vs. „psychotisch“* und der Variable *Vorhersage der AV: „neurotisch“ vs. „psychotisch“*
- Die Vorhersagegüte kann evaluiert werden entweder mit den gleichen Daten, die zur Schätzung des Modells verwendet wurden (in-sample Performance) oder mit neuen Daten (out-of-sample Performance).

Fazit

Wie kann das Ziel erreicht werden, Fehler bei diagnostischen Entscheidungen zu vermeiden?

- Maximierung der Sensitivität? $\frac{TP}{(TP+FN)}$
- Maximierung der Spezifität? $\frac{TN}{(FP+TN)}$
- Maximierung des Positiven Prädiktionswertes? $\frac{TP}{(TP+FP)}$
- Maximierung des Negativen Prädiktionswertes? $\frac{TN}{(TN+FN)}$

→ **Welcher Fehler ist schlimm(er)? Welches Ziel habe ich?**
→ **Abhängig vom Kontext!**

Beispiele zum Fazit

- Im Falle der Personalauswahl ist (oft) der Anteil Geeigneter an den Eingestellten von Interesse (Positiver Prädiktionswert $\frac{TP}{TP+FP}$)
- Bei klinischen Diagnosen oft der Anteil der richtig identifizierten Kranken in der Gruppe der Kranken wichtig, etwa wenn eine Behandlung/Therapie die Überlebenschancen erhöht (Sensitivität $\frac{TP}{TP+FN}$)
- Die Sensitivität kann aber auch bei der Personalauswahl als relevant erachtet werden, etwa wenn möglichst wenige Geeignete an die Konkurrenz verloren gehen sollen
- Im klinischen Kontext wird häufig auch der Anzahl korrekter Diagnosen eine zentrale Bedeutung beigemessen (Accuracy $\frac{TP+TN}{N}$)

(vgl. Schmidt-Atzert & Amelang, 2012, S.420)

Fazit

„[Die Optimierung einer Auswahlentscheidung] stellt somit ein Problem dar, für das es eine eindeutige Lösung nicht gibt, weil sie zugleich ein Werturteil erfordert, das nicht allein wissenschaftlich begründbar ist, sondern **stets auch auf persönlichen, sozialen und ökonomischen Werten sowie auf praktischen Erwägungen beruht**."

Wieczerkowski & Oeveste, 1982, S. 929f

→ Notwendigkeit von individuellen, gesellschaftlichen und finanziellen Kosten- und Nutzenerwägungen
→ Ausblick zur Entscheidungstheorie (VL 11): Ermöglicht quantifizierte Abwägung von Fehlern

Zusammenfassung

- Wenn möglich, ist eine mechanische (statistische) Urteilsbildung der klinischen Urteilsbildung vorzuziehen, um Urteilsfehler zu minimieren
- Bei der Beurteilung der Güte von Urteilen ist ein guter Goldstandard zentral
- Die Korrelation zwischen Testwert und Kriterium spiegelt die Validität des (gesamten) Verfahrens wieder, das zur Urteilsbildung herangezogen wird
- Vier zentrale Kennwerte, um die Güte eines Urteils zu bestimmen:
  - Sensitivität und Negativer Prädiktionswert berücksichtigen dabei den Fehler von Falsch-Negativen (FN) Entscheidungen
  - Spezifität und Positiver Prädiktionswert berücksichtigen dabei den Fehler von Falsch-Positiven (FP) Entscheidungen

- Eine hohe Validität ist zentral, um die Güte des Urteils für alle Kennwerte zu erhöhen
- PPV und NPV berücksichtigen explizit die Basisrate, und erlauben einen Rückschluss auf das Merkmal basierend auf einem Testergebnis
- Die Wahl des Cut-Off-Wertes kann gezielt einen der beiden Fehler (FN, FP) verringern und damit die entsprechenden Kennwerte erhöhen (auf Kosten des anderen Fehlers)
  - Es hängt vom Kontext ab, ob ein Fehler schwerwiegendere Konsequenzen hat als der andere (und welcher Fehler dies ist)
  - Falls Sensitivität und Spezifität ausgewogen sein sollen, hilft die ROC-Kurve bei der Wahl eines Cut-Off-Wertes

4. Beispiele

Fragebogen zu Schluckbeschwerden

Entwicklung eines Fragebogens zur Identifizierung leichter Schluckbeschwerden bei Parkinson-Erkrankung

- Studie mit N = 74 Personen
- Goldstandard / Kriterium:
  - Gesamtwert aus diversen klinischen Schluckbeschwerde-Testergebnissen
  - Kriteriums-Cut-Off-Wert → bestimmt die geschätzte Prävalenz: Personen mit Gesamtwert ≤ 25%-Quantil der Kriteriumswert-Verteilung werden als tatsächlich nicht krank bezeichnet
    - 0.25 * 74 = 19 Personen sind tatsächlich nicht krank
    - 74 - 19 = 55 Personen sind tatsächlich krank
- Korrelation zwischen Testwert und Kriterium (Validität) = .33

Fragebogen zu Schluckbeschwerden

Optimierung des Test-Cutoffs anhand von Sensitivität und Spezifität (ROC-Kurve) ergab folgende Häufigkeiten für die verschiedenen Kategorien:

- TP = 26, TN = 17
- FN = 29, FP = 2

Sanity-Check: Passt das zu unseren vorherigen Werten?

- N(tatsächlich nicht krank) = 19 = TN + FP = 17+2
- N(tatsächlich krank) = 55 = TP + FN = 26 + 29

→ Ja!

Fragebogen zu Schluckbeschwerden

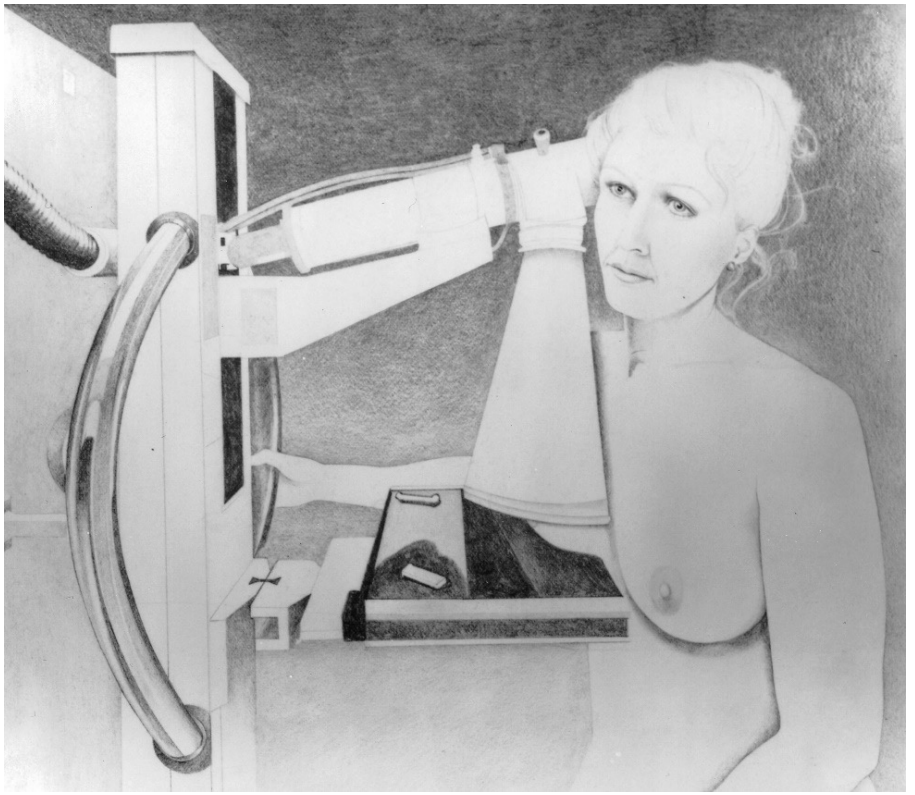
Lohnt sich der Einsatz dieses Fragebogens? Vergleich Prävalenz mit der Anzahl korrekter Diagnosen:

- **Prävalenz:** $\frac{TP+FN}{N} = \frac{26+29}{74} =$ **.74**
  → Bei Klassifikation aller Patienten als krank: 74% der Diagnosen korrekt
- Anzahl korrekter Diagnosen mit diesem Test:
  $\boldsymbol{AKD} = \frac{TP+TN}{N} = \frac{26+17}{74} =$ **.58**
  → Klassifikation aller Patienten anhand des Tests: 58% der Diagnosen korrekt

In diesem Fall müsste die Validität des Verfahrens r ≥ .89 sein, um eine bessere Klassifikation zu erzielen, als bei der pauschalen Bezeichnung aller Patienten als krank (Schönemann, 1997, S. 192)

Mammographie-Screening

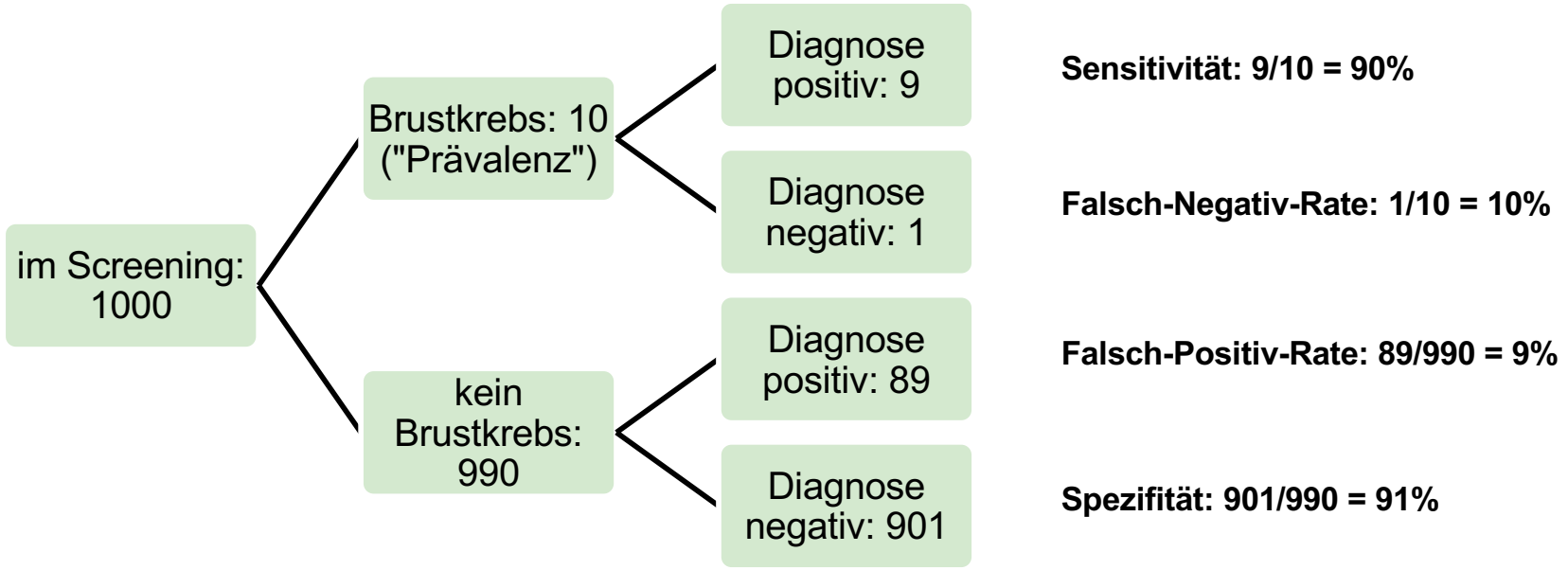
- Die Wahrscheinlichkeit, dass eine Frau Brustkrebs hat beträgt 1% (Prävalenz)
- Wenn eine Frau Brustkrebs hat, erkennt das Screening das mit 90% Wahrscheinlichkeit (Sensitivität)
- Wenn eine Frau keinen Brustkrebs hat, beträgt die Wahrscheinlichkeit, dass das Testergebnis (korrekt) negativ ist, 91% (Spezifität)



Gigerenzer, G.: Das Einmal-eins der Skepsis. Über den richtigen Umgang mit Zahlen und Risiken. Berliner Taschenbuch Verlag, Berlin 2004; http://www.spektrum.de/pdf/gug-09-10-s034-pdf/1006101?file

Mammographie-Screening

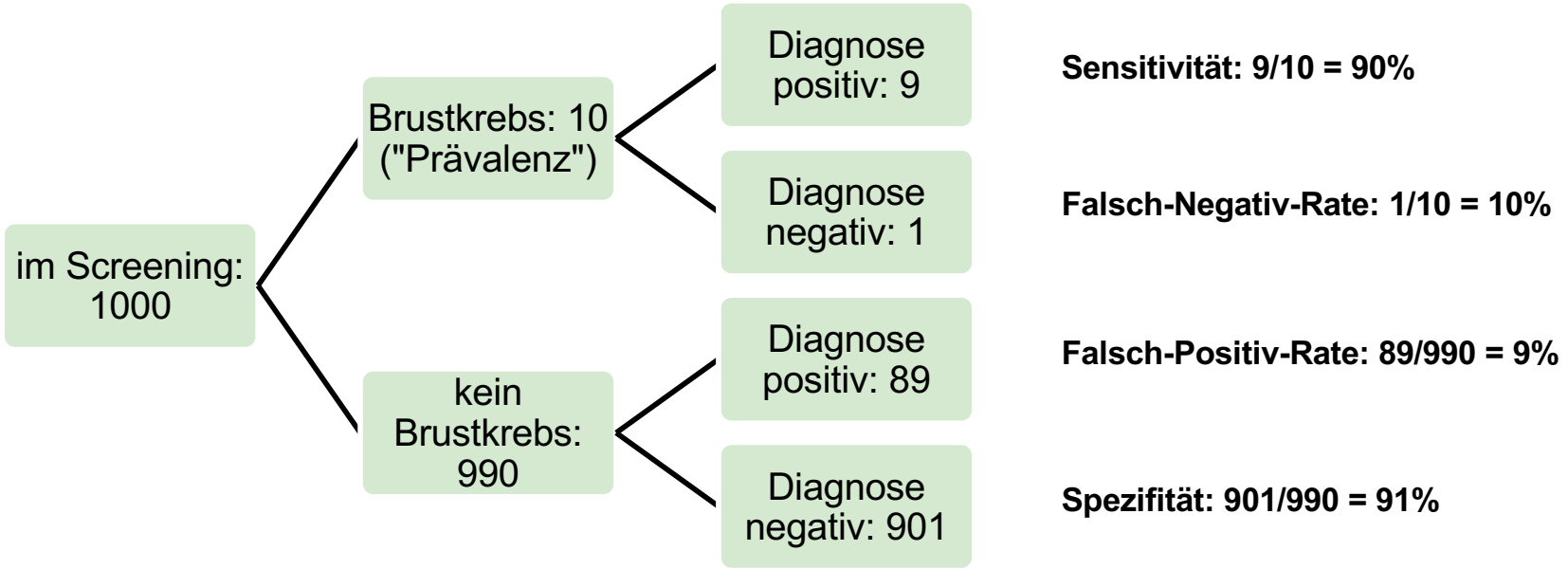
Prävalenz = 1%, Sensitivität = 90%, Spezifität = 91%



Gigerenzer, G.: Das Einmal-eins der Skepsis. Über den richtigen Umgang mit Zahlen und Risiken. Berliner Taschenbuch Verlag, Berlin 2004; http://www.spektrum.de/pdf/gug-09-10-s034-pdf/1006101?file

Mammographie-Screening

Sie gehen zum Screening und erhalten ein positives Ergebnis: Wie hoch ist die Wahrscheinlichkeit, dass Sie tatsächlich Brustkrebs haben?



$$P(\text{Brustkrebs} \mid \text{positive Diagnose}) = PPV = \frac{TP}{(TP+FP)} = \frac{9}{(9+89)} = 9.2\%$$

Gigerenzer, G.: Das Einmal-eins der Skepsis. Über den richtigen Umgang mit Zahlen und Risiken. Berliner Taschenbuch Verlag, Berlin 2004; http://www.spektrum.de/pdf/gug-09-10-s034-pdf/1006101?file

Mammographie-Screening

- Sie gehen zum Screening und erhalten ein positives Ergebnis: Wie hoch ist die Wahrscheinlichkeit, dass Sie tatsächlich Brustkrebs haben?
  - 60% von 160 befragten Gynäkologinnen gaben eine Wahrscheinlichkeit von **> 80%** an (Gigerenzer et al., 2007)
  - Positiver Prädiktionswert liegt aber nur bei 9.2%: Von 10 Frauen mit positivem Screening-Befund hat etwa 1 tatsächlich Brustkrebs
- Hier wird der Einfluss der Basisrate auf den PPV sichtbar!

Merke: Die Inferenz über ein Merkmal (hier: Brustkrebs oder nicht) **hängt nicht nur vom Testergebnis** (Diagnose) **und der Qualität des Tests** (Sensitivität und Spezifität) **ab, sondern auch von der Basisrate** (Prävalenz)

Gigerenzer, G.: Das Einmal-eins der Skepsis. Über den richtigen Umgang mit Zahlen und Risiken. Berliner Taschenbuch Verlag, Berlin 2004; http://www.spektrum.de/pdf/gug-09-10-s034-pdf/1006101?file

lisa lendway
@lisalendway

Any other statisticians have to look up sensitivity and specificity every time? I can never remember which is which. I much prefer true positive rate and true negative rate.

Tweet übersetzen

2:58 nachm. · 11. Apr. 2020 · Twitter for iPhone

39 Retweets **595** „Gefällt mir"-Angaben

icastico
@enemiesnet

Antwort an @JeffZemla @AndrewM_Webb und 2 weitere

As a musician, sensitivity is easy for me to conceptualize. A sensitive microphone will not miss the target sound if it is there - but it may pick up unwanted background noise.

Tweet übersetzen

6:14 vorm. · 12. Apr. 2020 · Twitter Web App

Tommy Carpenito
@CarpenitoThomas

Antwort an @lisalendway

I always think- "I would want to be sensitive giving someone a test result back who actually had the disease." And then I say "I guess specificity is the other."

Tweet übersetzen

3:33 vorm. · 12. Apr. 2020 · Twitter for iPhone

1 „Gefällt mir"-Angabe

Vita
@DataVizVita

Antwort an @lisalendway

When I took a statistics course I found a strange association, but very easy to remember. Toilet paper (TP) is related to sensitivity.

Tweet übersetzen

11:24 nachm. · 11. Apr. 2020 · Twitter Web App

https://twitter.com/lisalendway/status/1248958632131207169