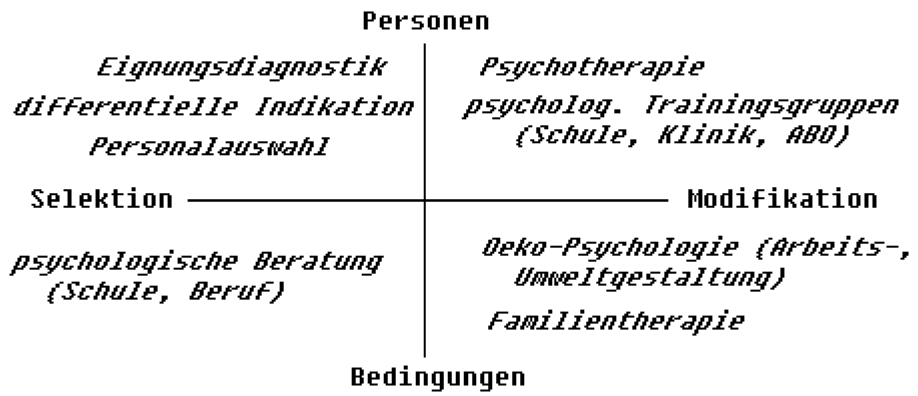
8. Entscheidungstheorie: Selektion oder Klassifikation

Tests sollen uns helfen Entscheidungen zu treffen. Wenn wir Tests anwenden, müssen wir vor der Gesellschaft und vor den beteiligten Menschen rechtfertigen können, dass die Testanwendung wirklich etwas gebracht hat. Die Messung und Vorhersage kann man durch formale, psychometrische Kriterien wie Reliabilität und Validität beschreiben. Diagnostische Entscheidungen müssen allerdings anhand ihres Nutzens für Individuen und Institutionen in der Gesellschaft bewertet werden. Man kann die Diagnosen in den Anwendungsgebieten der Psychologie im Schema in Abbildung 18 grob klassifizieren.

Abbildung 18: Diagnostische Entscheidungen in der Psychologie



Legende: Diagnostische Entscheidungen können in den Raum zwischen Selektion von Personen oder Bedingungen versus Modifikation von Personen oder Bedingungen eingeordnet werden.

Die Einteilung in Abbildung 18 soll helfen, diagnostische Entscheidungen nicht nur im Zusammenhang mit Selektion von Personen zu sehen. Die Personalentscheidungs-problematik stellt aber den leichtesten Fall für entscheidungstheoretische Ueberlegungen dar und soll deshalb weiter dargestellt werden.

8.1. Kategorisierung von Personalentscheidungsproblemen

Abbildung 19: Vier Fälle bei der Personalentscheidung

	Zurückweisung möglich	Zurückweisung nicht möglich
Eine Stelle ist offen	Selektion	Annahme
Zwei oder mehr Stellen sind offen	multiple Selektion	Klassifikation

Legende: Bei einer zu besetzenden Stelle, wenn Bewerber zurückgewiesen werden können spricht man von "Selektion". Bei der "multiplen Selektion" sind mehrere offene Stellen zu besetzten. Ist einen Ablehnung von Bewerbern nicht möglich spricht man bei einer Stelle von "Annahme" und bei mehreren Stellen von "Klassifikation".

Wie bei der klassischen Validität wird in praktisch allen vier Fällen die multiple Regression als formales Modell verwendet, um Aussagen über die Güte der Entscheidungen zu treffen. Die Anwendungen des Regressionsmodells variieren allerdings je nach Typ der Entscheidung.

<u>8.1.1. Selektion</u>

Man geht von folgender multiplen Regressionsgleichung aus:

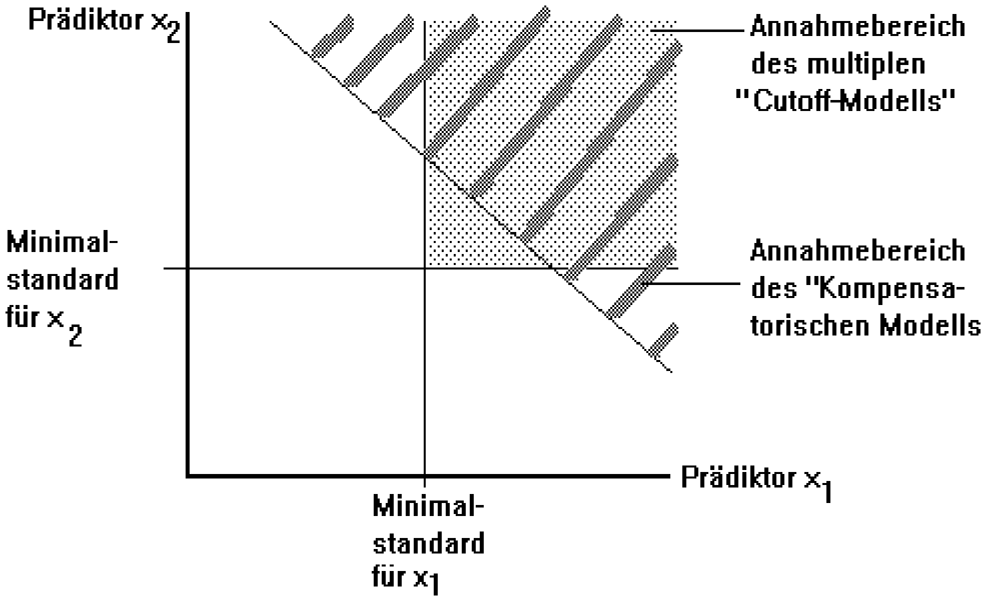
$$y' = b_1 x_1 + b_2 x_2 + ... + b_n x_n$$

y'	... Vorhergesagtes Kriterium (z.B. Belastungsfähigkeit)
$x_{1 \ldots n}$	... Prädiktoren zur Vorhersage des Kriteriums (z.B. Leistungen im Subtest 1 bis n)
$b_{1 \ldots n}$	... Gewichte für die Prädiktoren

Wenn die Prädiktoren niedrig interkorrelieren und gleichzeitig eine hohe Korrelation zum Kriterium aufweisen ergibt sich eine hohe multiple Korrelation $\mathbf{R_{y.12 \ldots n}}$ und damit eine gute Vorhersage z.B. der Eignung einer Person für einen Posten.

Zur Festlegung der Gewichte bei der einfachen Selektion muss man zwischen dem "Cutoff-Modell" und dem "Kompensatorischen Modell" unterscheiden.

Abbildung 20: Cutoff-Modell und Kompensatorisches Modell im Vergleich



Legende: Prädiktor x_1 und x_2 sind die Variablen, die das Kriterium, z. B. Berufseignung vorhersagen sollen. Das Cutoff-Modell legt minimale Anforderungen für die Prädiktoren fest. Beim Kompensatorischen Modell können Schwächen im einen Prädiktor durch Stärken im anderen ausgeglichen werden.

Beim Cutoff-Modell werden Minimalstandards gesucht, das heisst für die Selektion wird eine Mindestausprägung der Gewichte gefordert (Bsp.: für die Ausbildung zum Piloten darf die Sehschärfe des Bewerbers nicht kleiner sein als 0,5 Dioptrien). Wenn man mehr als einen solcher Grenzwerte festlegt spricht man von multiplen Cutoff-Modellen (Der Pilot darf zusätzlich nicht grösser sein als 1,80 Meter sein, da er ansonsten nicht ins Cockpit passt).

Beim Kompensatorischen Modell kann der Bewerber Schwächen in einem oder mehreren Prädiktoren durch Stärken in anderen ausgleichen. (Bsp.: Zur Auswahl eines Bewerbers für eine Lehrstelle im metallverarbeitenden Gewerbe werden seine Leistungen in einem Schulleistungstest, einem Intelligenztest und in einem Handgeschicklichkeitstest festgestellt. Der von uns untersuchte Bewerber hat grosse Schwächen im Schulleistungstest, die er aber durch gute Leistungen im Intelligenztest ausgleichen kann und somit als geeignet eingestuft wird.)

8.1.2. Multiple Selektion

Bei der multiplen Selektion wird eine Auswahl von Bewerben für verschiedene Stellen getroffen. Dabei sollten sich die verschiedenen Stellen durch spezielle Anforderungen unterscheiden, das heisst die Prädiktoren erhalten verschieden starke Gewichte je nach Stelle und man kann für jede Stelle eine Regressionsgleichung aufstellen:

$$y'_a = {}_a b_1 x_1 + {}_a b_2 x_2 + \ldots + {}_a b_n x_n$$

$$y'_b = {}_b b_1 x_1 + {}_b b_2 x_2 + \ldots + {}_b b_n x_n$$

(Bsp.: Es sollen Schülern Empfehlungen gegeben werden, welche weiterführenden Schulen für sie geeignet wären. *a* wäre das Anforderungsprofil für naturwissenschaftliche Gymnasien, bei dem hohes Gewicht auf Mathematik, Physik und Chemie gelegt wird. *b* ist entsprechend der geisteswissenschaftliche Schultyp, mit hohen Gewichten auf sprachlich, philosophischen Fächern. Zudem kann es aber auch Gewichte geben, die für beide Schulzweige hoch sein sollten, z.B. das Gewicht für das Fach Geographie.)

8.1.3. Annahme

Es werden alle Personen angenommen und nicht differentiell Stellen zugewiesen.

8.1.4. Klassifikation

Für die Klassifikation sollte die Unterschiedlichkeit zwischen Stellen möglichst gross sein und damit auch der Unterschied zwischen den Prädiktoren:

$$(y_a - y_b)' = ({}_a b_1 - {}_b b_1)\, x_1 + ({}_a b_2 - {}_b b_2)\, x_2 + \ldots + ({}_a b_n - {}_b b_n)\, x_n$$

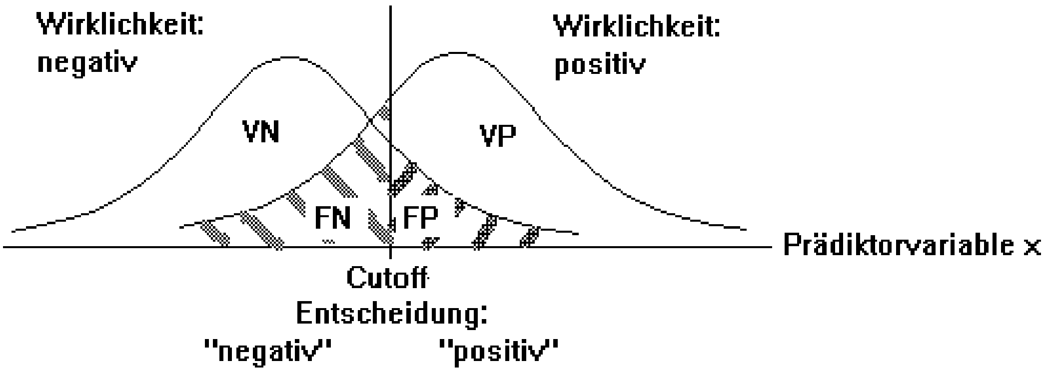
Bei dem oben genannten Beispiel ist das Fach Geographie kein guter Indikator oder Prädiktor für den Erfolg in dem Schulzweig, denn er erlaubt keine Klassifikation der Schüler. Die anderen genannten Fächer sind sehr gut für eine Klassifikation, denn sie erlauben eine differentielle Zuweisung, je nach Leistung im entsprechenden Fach.

8.2. Outcomes of Prediction: Ergebnisse der Vorhersage

Was sind nun die Ergebnisse unserer Testanwendung bzw. Assessment Prozedur? Die Anzahl der korrekten Entscheidungen beinhaltet mehr Informationen, als Regressions- oder Validitätskoeffizienten. Dabei muss man zwischen zwei Populationen unterscheiden: Den in Wirklichkeit geeigneten, positiven und den in Wirklichkeit ungeeigneten, negativen Gruppen von Personen. Die Verhältnisse in der "Wirklichkeit" sind uns allerdings nicht bekannt.

Analog zur Problematik des Vergleichs zweier Populationen (z.B. beim t-Test) kann man sich die Häufigkeitsverteilungen der beiden Gruppen wie in Abbildung 21 vorstellen.

Abbildung 21: Häufigkeitsverteilungen der negativen (N) und positiven (P) Personen, mit den möglichen validen (V) und falschen (F) Entscheidungen.



Legende: Für die Prädiktorvariable x wird ein cutoff festgesetzt. Oberhalb des Wertes fällt man die Entscheidung "positiv". Die "in Wirklichkeit" positiven und negativen Personen verteilen sich bezüglich des Prädiktors nach getrennten Häufigkeitsverteilungen. Die Flächen unter den Verteilungen geben den Anteil der Personen an, für die die Entscheidung valide (V) oder falsch (F) ausfällt, und die in Wirklichkeit positiv (P) oder negativ (N) sind.

Man kann also vier mögliche Ergebnisse der Entscheidung unterscheiden. Die Gegebenheiten in der Wirklichkeit sind uns allerdings nicht bekannt.

Tabelle 10: Vier-Felder-Tafel der möglichen Ergebnisse von Entscheidungen

	Entscheidung: "negativ"	Entscheidung: "positiv"
Wirklichkeit: positiv	falsche Negative FN	valide Positive VP
Wirklichkeit: negativ	valide Negative VN	falsche Positive FP

Die Häufigkeit bzw. Anzahl der Fälle jeder der vier möglichen Entscheidungen kann als Wahrscheinlichkeit (relative Häufigkeit) ausgedrückt werden, indem die absolute Häufigkeit durch die Gesamtzahl der untersuchten Stichprobe teilt.

Tabelle 11: Vier-Felder-Tafel der Wahrscheinlichkeiten für die möglichen Ausgänge von Entscheidungen

	Entscheidung: "negativ"	Entscheidung: "positiv"	
Wirklichkeit: positiv	falsche Negative P(FN)= FN / (P + N)	valide Positive P(VP)= VP / (P + N)	Basisrate: BR= P / (P + N)
Wirklichkeit: negativ	valide Negative P(VN)= VN / (P + N)	falsche Positive P(FP)= FP / (P + N)	Negativrate: 1 - BR= N / (P + N)
	(1 - SR)= N' / (P' + N')	Selektionsrate: SR= P' / (P' + N')	

Dabei erscheinen weitere wichtige Kennwerte als Randsummenwerte:

Die Basisrate entspricht dem Anteil der positiven Fälle in der Population, die Selektionsrate gibt an, wieviel Personen angenommen werden geteilt durch die Anzahl Personen, die untersucht wurden. Je kleiner die Selektionsrate wird, umso effektiver wird die Testanwendung. Die Basisrate ist oftmals nicht bekannt, denn sie macht Aussagen über die Verhältnisse in der Wirklichkeit. Die Basisrate kann aber aus empirischen Untersuchungen (z.B.: durch Vollerhebungen oder repräsentative Stichproben) erschlossen werden.

Der Zusammenhang der Vier-Felder-Tafel bzw. zwischen der Wirklichkeit und unserer Entscheidung kann durch den Koeffizient Phi (φ) berechnet werden.

$$\varphi_{yy'} = \frac{P(VP) - BR * SR}{\sqrt{BR * (1 - BR) * SR * (1 - SR)}}$$

Dieser Koeffizient Phi (φ) ist nichts anderes, als die Validität unseres Tests, denn wir wollen aufgrund der Testergebnisse (die zu einer Entscheidung führen) die Wirklichkeit vorhersagen.
Durch Umformung erhalten wir die Wahrscheinlichkeit der validen Positiven:

$$P(VP) = BR * SR + \varphi_{yy'} * \sqrt{BR * (1 - BR) * SR * (1 - SR)}$$

Wenn man eine Zufallszuweisung vornimmt, also keinen Test anwendet, ist der Zusammenhang zwischen Entscheidung und Wirklichkeit Null, d.h. der Koeffizient Phi wird Null und man erhält für die Zufallszuweisung:

$$P(VP) = BR * SR \qquad \text{(bei Zufallszuweisung)}$$

8.3. Beispiele für Ergebnisse von Entscheidungen

<u>8.3.1. Entscheidungen ohne Testanwendung (zufällig)</u>

		Basisrate= .60			Basisrate= .70			
		"N"	"P"			"N"	"P"	
Selektions-	P	.06	.54	.60	P	.07	.63	.70
rate = .90	N	.04	.36	.40	N	.03	.27	.30
		.10	.90			.10	.90	
	Prozent valider Entscheidungen: 58%				Prozent valider Entscheidungen: 66%			

		Basisrate= .60			Basisrate= .70			
		"N"	"P"			"N"	"P"	
Selektions-	P	.00	.60	.60	P	.00	.70	.70
rate = 1.00	N	.00	.40	.40	N	.00	.30	.30
		.00	1.00			.00	1.00	
	Prozent valider Entscheidungen: 60%				Prozent valider Entscheidungen: 70%			

<u>8.3.2. Entscheidungen mit Testanwendung (Validität des Tests: Phi = .30)</u>

		Basisrate= .60			Basisrate= .70			
		"N"	"P"			"N"	"P"	
Selektions-	P	.02	.58	.60	P	.03	.67	.70
rate = .90	N	.08	.32	.40	N	.07	.23	.30
		.10	.90			.10	.90	

Prozent valider Entscheidungen: 66%	Prozent valider Entscheidungen: 74%

Es wird deutlich, wie selbst durch einen relativ unreliablen Test die validen Entscheidungen [P(VP) + P(VN)] stark zunehmen.

Was bringt nun aber tatsächlich die Testanwendung?

Um diese Frage zu beantworten brauchen wir einen weiteren Begriff: der Nutzen.
Jedes Ergebnis der Entscheidung hat einen bestimmten Nutzen für die Institution, die die Entscheidung fällt. Der Nutzen der Ergebnisse der Entscheidung müssen festgelegt werden, um bestimmen zu können, ob eine Testanwendung tatsächlich etwas bringt.
Allgemein gilt die Formel für den Nutzen (expectet utility):

$$EU = U_t + N_{VP} * U_{VP} + N_{FN} * U_{FN} + N_{VN} * U_{VN} + N_{FP} * U_{FP}$$

EU ... erwarteter Nutzen (expectet utility)
U_t ... Nutzen der Testanwendung (oft negativ, da Test Geld kostet)
N_{VP} ... Anzahl valider Positiver
U_{VP} ... Nutzen valider Positiver

Angewandt auf das obige Beispiel:

Gegeben sind folgende Informationen:

N = 100	Bewerber werden untersucht
BR = .60	Es gibt 60% positive Fälle in der Population
SR = .90	Es sollen 90% der Bewerber angenommen werden

Die Institution muss nun den Nutzen der Ergebnisse der Vorhersage festlegen:
$U_{VP} = 1.00$
$U_{FN} = -.50$
$U_{VN} = 0$
$U_{FP} = -1.00$

Man kann den Nutzen einer zufälligen Entscheidung (ohne Test) berechnen:

$EU = U_t + 54 * 1.00 + 6 * (-.50) + 4 * 0 + 36 * (-1.00) = U_t + 15$

Wenn kein Test angewandt wird ist der Nutzen der Testanwendung natürlich 0 ($U_t = 0$), so bleibt ein erwarteter Nutzen von 15.
Entsprechend ist der Nutzen der Entscheidung auf der Grundlage eines Tests mit einer Validität von .30:

$EU = U_t + 58 * 1.00 + 2 * (-.50) + 8 * 0 + 32 * (-1.00) = U_t + 25$

Es ergibt sich ein Nutzen von 25. Man muss allerdings noch die Aufwendung für den Einsatz des Tests abrechnen (Ut = -5), so dass man auf einen erwarteten Nutzen von 20 kommt.

8.4. Der inkrementelle Nutzen

Der Zuwachs an Nutzen durch die Testanwendung ergibt sich aus der Differenz zwischen Nutzen der Anwendung des Tests und Nutzen der zufälligen Entscheidung ohne Test.
Ist die Differenz negativ => keine Anwendung
Ist die Differenz positiv => Anwendung des Tests

In unserem Beispiel ist die Differenz positiv, der Test sollte also als Entscheidungshilfe eingesetzt werden.

8.5. Uebungsaufgaben zur Entscheidungstheorie

1. Bei der Entscheidungstheorie gibt es die folgenden Grössen: Basisrate, Selektionsrate, (1 - BR), (1 - SR), die Wahrscheinlichkeiten von validen Positiven, falschen Positiven, validen Negativen, falschen Negativen, vorhergesagter Erfolg, vorhergesagter Misserfolg, tatsächlicher Erfolg und tatsächlicher Misserfolg. Zeichne das Diagramm (Vierfeldertafel) und trage alle Begriffe ein.

2. Zeichne für zwei Variablen x_1 und x_2, die für die Auslese verwendet werden, ein Kompensatorisches und ein multiples Cutoff-Modell. Welches Selektionsmodell findest Du besser?

3. Die Anzahl "falscher Negativer" wird grösser, wenn der Cutoff-Punkt bezüglich des Prädiktors so gesetzt wird, dass ...
   ❑ ... nur wenige Personen darüber liegen (restriktiv).
   ❑ ... viele darüber liegen (nicht restriktiv).
   ❑ ... das hat nichts miteinander zu tun.

4. Eine gesellschaftliche Institution darf aufgrund ihrer Satzung keine Bewerber ablehnen, sie hat jedoch zwei unterschiedliche Berufsrichtungen (A und B) zu besetzen. Testpsychologen haben für die beiden Richtungen eine Testbatterie zusammengestellt und empirisch validiert, die zu folgenden Regressionsgleichungen führte:

   für A: $Y'_A = .30 * X_1 + .60 * X_2 + .40X_3$
   für B: $Y'_B = .70 * X_1 + .25 * X_2 + .40X_3$

   a) Um welche Entscheidungsform handelt es sich hier?

   b) Welche Testskalen X_i haben differentielle Validität, d.h. bringen etwas zur unterschiedlichen Zuweisung zu den Richtungen (Mehrfachnennung möglich).
   ❑ X1 ❑ X2 ❑ X3

   c) Welche Testskalen X_i haben keine differentielle Validität, d.h. bringen nichts zur unterschiedlichen Zuweisung zu den Richtungen (Mehrfachnennung möglich)..
   ❑ X1 ❑ X2 ❑ X3

5. Die Anwendung einer Testbatterie bei einer Stichprobe ergab folgende Daten:

FN: .20	VP: .60
VN: .10	FP: .10

a) Berechne die Basisrate (BR) und die Selektionsrate (SR).

b) Gib die Prozent valider Entscheidungen an.