# Lektion 5

# Statistische Entscheidungstheorie

LERNZIELE

Nach Abschluss dieser Lektion werden Sie Folgendes gelernt haben...

- die grundlegenden Elemente der statistischen Entscheidungstheorie einschließlich Verlustfunktion, Entscheidungsfunktion und Risikofunktion.

- die Definitionen von Minimax-Risiko, Bayes-Risiko, Minimax-Entscheidungsfunktionen und Bayes-Entscheidungsfunktionen.

- die Definition der Zulässigkeit einer Entscheidungsfunktion.

- das Stein-Paradoxon zusammen mit einer Veranschaulichung anhand des James-Stein-Schätzers für mehrfache Mittelwerte.

# 5. Statistische Entscheidungstheorie

### Einführung

Wir möchten einen Spam-Filter entwickeln, der eingehende E-Mails automatisch als "Spam" oder "kein Spam" klassifiziert. Wenn dieser Filter einen Fehler macht, entsteht den Benutzer:innen ein Verlust (verschwendete Zeit). Wenn der Filter eine Spam-E-Mail durchlässt, müssen die Benutzer:innen Zeit aufwenden, um die E-Mail zu löschen (eine Sekunde). Wenn der Filter eine Nicht-Spam-E-Mail blockiert, müssen die Benutzer:innen den Spam-Ordner durchsuchen, um die "gute" E-Mail zu finden, und verlieren viel Zeit (100 Sekunden). Auf die statistische Entscheidungstheorie angewandt, wird der Spam-Filter als Entscheidungsfunktion bezeichnet, die versucht, den wahren Zustand (Spam oder nicht Spam) einer zuvor nicht gesehenen (zufälligen) Beobachtung (einer E-Mail) zu erraten. In dieser Lektion werden wir lernen, wie man die Qualität solcher Entscheidungsfunktionen misst.

In Abschnitt 1 werden die grundlegenden Elemente der statistischen Entscheidungstheorie einschließlich der Entscheidungs-, Verlust- und Risikofunktionen vorgestellt. Wir werden sehen, wie diese Elemente verwendet werden können, um Fragen darüber zu stellen, wie gut unsere Entscheidungsfunktion (unser Schätzer) funktioniert. In Abschnitt 5.2 betrachten wir einige Möglichkeiten zur Definition einer optimalen Entscheidungsfunktion für einen Zielparameter von Interesse. Insbesondere betrachten wir das Minimax-Risiko zusammen mit der zugehörigen Minimax-Entscheidungsfunktion (Schätzer). Wir definieren und diskutieren auch eine Bayesianische Betrachtung, bei der wir das Bayes-Risiko und die entsprechende Bayes-Entscheidungsfunktion (Bayes-Schätzer) definieren, die dieses Bayes-Risiko minimiert. Schließlich wird in Abschnitt 5.3 ein wichtiges Konzept, die Zulässigkeit, besprochen, die eine wünschenswerte Eigenschaft einer Entscheidungsfunktion ist. Entgegen der Intuition werden wir sehen, dass, wenn wir drei oder mehr unabhängige Zielparameter gleichzeitig schätzen wollen, der "gewöhnlich beste" Schätzer für jeden Parameter nicht die beste Option ist. Dieses eher paradoxe Ergebnis wird als Stein-Paradoxon bezeichnet. Abschnitt 5.3 endet mit dem James-Stein-Schätzer (Entscheidungsfunktion) für die Schätzung von drei oder mehr Mittelwerten einer unabhängigen Gauß-Verteilung. Das Ergebnis mag Sie überraschen.

## 5.1 Die Risikofunktion

Betrachten wir das Problem der Feststellung, ob es sich bei einer E-Mail um Spam handelt. Wir wollen dies tun, indem wir eine Entscheidung δ auf der Grundlage einer Beobachtung x treffen. In diesem Fall enthält x Informationen über die E-Mail, wie den Absender, die Betreffzeile oder den Text. Der Zustand der E-Mail ist die wahre Klassifizierung, die wir in der Variablen y kodieren können. Wir wählen y = 0, um "Spam" anzuzeigen und y = 1, um "gut" anzuzeigen. Unsere Aufgabe ist es, eine Entscheidungsfunktion δ zu finden, die einer gegebenen E-Mail, die durch x repräsentiert wird, eine Null ("Spam") oder 1 ("gut") zuweist. Wenn wir glauben, dass die E-Mail "Spam" ist, dann ist δ x = 0. Eine korrekte Entscheidung entspricht dem (wahren) Zustand. Dies kann in zwei Fällen geschehen:

1. Die E-Mail ist Spam, und die Entscheidungsfunktion sagt y = δ x = 0 voraus

2. Die E-Mail ist gut und die Entscheidungsfunktion sagt y = δ x = 1 voraus

Auch bei einer falschen Entscheidung gibt es zwei Fälle:

1. Die E-Mail ist Spam, aber die Entscheidungsfunktion stuft sie als gut ein: 0 = y ≠ δ x = 1

**Verlustfunktion**

Diese Funktion misst die Qualität einer Entscheidung

gegen den wahren Zustand. Es ist eine nicht-negative Funktion.

Da sie auf einer zufälligen Beobachtung beruht, ist die Verlustfunktion eine Zufallsvariable.

2. Die E-Mail ist gut, aber die Entscheidungsfunktion stuft sie als Spam ein: 1 = y ≠ δ x = 0

Sobald wir eine Entscheidungsfunktion haben, wollen wir bewerten, wie gut sie funktioniert. Zu diesem Zweck wollen wir jedem Paar y, δ, das den wahren Zustand und die Entscheidung enthält, einen "Verlust" zuweisen. Wenn die Werte des Paares übereinstimmen, haben wir keinen Verlust erlitten. Wenn die Werte des Paares nicht übereinstimmen, haben wir eine falsche Entscheidung getroffen und erleiden einen (positiven) Verlust. Eine **Verlustfunktion**, L, nimmt ein solches Paar und gibt eine (nicht-negative) reelle Zahl zurück. Wie gewünscht, ist L y, δ = 0, wenn y = δ ist, und wenn y ≠ δ ist, ist L y, δ > 0. Im Beispiel der E-Mail-Klassifizierung ist eine gängige Verlustfunktion der Null-Eins-Verlust (0-1), der wie folgt definiert ist:

Xxx

In unserem Beispiel können sowohl der Zustand (y) als auch die Entscheidung (δ) jeweils einen Wert von 0 (Spam) oder 1 (gut) annehmen. Die Null-Eins-Verlustfunktion kann für jedes mögliche Paar berechnet werden:

L 0, 0 = L 1, 1 = 0undL 0, 1 = L 1, 0 = 1

Manchmal ist es sinnvoll, diese Werte in einer Verlustmatrix zusammenzufassen:

Tabelle 27: Die Null-Eins-Verlustmatrix für die binäre Klassifizierung

Die Verlustmatrix ist das analoge Objekt einer Konfusionsmatrix, die für Klassifizierungsprobleme verwendet wird. Jede Zelle dieser Matrix enthält den Wert des Verlustes, der durch den Wert der gewählten Entscheidungsfunktion (Zeilen) gegenüber dem wahren Zustand (Spalten) entsteht. Wenn diese übereinstimmen, ist der Verlust gleich Null. Wenn sie nicht übereinstimmen, ist der Verlust positiv.

Bei dieser Verlustfunktion gehen wir davon aus, dass die Klassifizierung einer E-Mail als Spam, wenn sie gut ist, genauso schlecht ist wie die Einstufung einer E-Mail als gut, wenn sie tatsächlich Spam ist. Dies kann wünschenswert sein oder auch nicht; vielleicht ist die Klassifizierung einer guten E-Mail als Spam mit mehr Kosten verbunden als der umgekehrte Fall. Vielleicht haben Sie eine E-Mail erwartet, die im Spam-Ordner landet, und Sie verbringen mehr Zeit mit der Suche danach. Der umgekehrte Fall ist zwar ärgerlich, aber man kann die E-Mail auch einfach löschen, und es kostet nicht allzu viel Zeit. Um einen solchen unausgewogenen (gewichteten) Verlust zu modellieren, betrachten Sie die folgende Verlustfunktion, die der Klassifizierung einer guten E-Mail als Spam zehnmal mehr Gewicht zuweist als umgekehrt:

xxx

Die entsprechende Verlustmatrix ist gegeben durch

Tabelle 28: Eine gewichtete Verlustmatrix für die binäre Klassifizierung

Erinnern Sie sich daran, dass in unserem Anordnung x eine E-Mail (Beobachtung) ist. Da wir unsere Entscheidungsfunktion für jede beliebige eingehende E-Mail erstellen, ist x eine Realisierung eines Zufallsprozesses (Variable) X. Diese Zufälligkeit macht die Entscheidungsfunktion δ ebenfalls zufällig, auch wenn sie ein gewisses Muster haben muss (sonst würden wir nur zufällig raten). Diese Zufälligkeit wirkt sich auch auf die Verlustfunktion aus. Wenn wir die Verlustfunktion explizit schreiben, wird anhand von L y, δ X deutlich, dass es sich um eine Zufallsvariable handelt. Daher können wir die Verlustfunktion L nicht für alle Eingaben X bewerten. Daher können wir die "Güte" einer Entscheidungsfunktion quantifizieren, indem wir den Erwartungswert der Verlustfunktion analysieren: E L y, δ X . Dieser Erwartungswert ist genau die Definition der **Risikofunktion**. Mit anderen Worten: Bei einer Verlustfunktion L, einer Entscheidungsfunktion δ und einer Zufallsvariablen X, deren Werte wir beobachten werden, ist die Risikofunktion gegeben durch (Jiao et al., 2013)

**Risikofunktion**

Dies ist der erwartete Verlust für eine gegebene Verlustfunktion und eine gegebene Entscheidungsfunktion.

Xxx

Da der Erwartungswert die Zufälligkeit von X absorbiert, ist die Risikofunktion eine deterministische Funktion des Zustands y. Wenn X eine diskrete Verteilung hat, dann

xxx

Wenn X eine kontinuierliche Verteilung hat, dann

Xxx

wobei f die Wahrscheinlichkeitsdichtefunktion (PDF) von X ist. Um dieses (möglicherweise mehrdimensionale) Integral müssen Sie sich nicht allzu sehr kümmern. Wir werden uns in dieser Lektion nicht mit der Bewertung solcher Größen beschäftigen. Stattdessen werden wir uns darauf konzentrieren, zu verstehen, wie die Risikofunktion definiert und quantifiziert wird und im nächsten Abschnitt, wie sie optimiert werden kann. Zusammenfassend lässt sich sagen, dass wir angesichts einer Verlustfunktion und der Verteilung der Beobachtungen, deren Zustand wir vorhersagen wollen, die Qualität unserer Entscheidungsfunktion durch die Risikofunktion quantifizieren.

Wir werden nun ein vertrauteres Beispiel betrachten und untersuchen, wie die Schlüsselkonzepte in diesem Abschnitt angewendet werden. Nehmen wir an, dass X einer Gauß-Verteilung N μ, 1 mit unbekanntem Mittelwert μ und Einheitsstandardabweichung σ = 1 folgt. Wir nehmen an, dass unsere Beobachtung die Form von n unabhängigen Beobachtungen aus dieser Verteilung hat: x = x1, ..., xn aus X = X1, ...,Xn iid aus N μ, 1 . (Wir erinnern uns, dass iid für "unabhängig und identisch verteilt" steht.) In diesem Fall ist der wahre Zustand der wahre Mittelwert μ. Eine offensichtliche Möglichkeit, diesen Mittelwert über eine Zufallsstichprobe zu schätzen, besteht darin, den Stichprobenmittelwert zu verwenden. Für die Beobachtung schreiben wir x = 1nΣi = 1nxi, und für die Stichprobe schreiben wir X = 1nΣi = 1nXi. Mit anderen Worten: Eine Option für die Entscheidungsfunktion ist der Stichprobenmittelwert. δ x = x für eine bestimmte Beobachtung, und δ X = X für die Zufallsstichprobe. Eine Verlustfunktion, die sich an den kleinsten Quadraten orientiert, ist einfach die quadratische Differenz zwischen dem wahren Mittelwert und der Entscheidung: L μ, δ = μ - δ 2. Wenn wir den Verlust für eine bestimmte Beobachtung x berechnen, ist die Größe L μ, δ = μ - δ x 2 eine Zahl. Wenn wir jedoch den Verlust einer Zufallsstichprobe betrachten, ist die Größe L μ, δ = μ - δ X 2 eine Zufallsvariable.

Die erste Version ist recht einfach. Angenommen, wir beobachten x = 1, 3, 3, 2, 1 (hier n = 5). Dann ist δ x = x = 1 + 3 + 3 + 2 + 1 5 = 2. Wenn der wahre Zustand μ = 1 . 5 lautet, dann lautet der Verlust L 1 . 5, 2 = 1 . 5 - 2 2 = 0 . 25. Für die Zufallsvariable L wollen wir die Zufälligkeit anhand einer Simulation untersuchen. Wir legen den Stichprobenumfang auf n = 30 fest und nehmen an, dass der wahre Zustand μ = 1 . 5. Wir erzeugen also 1000 zufällige Beobachtungen (mit jeweils 30 Zahlen) aus N 1 . 5, 1 . Als Nächstes berechnen wir für jede Beobachtung die Entscheidung δ, die einfach der Stichprobenmittelwert ist. Zu diesem Zeitpunkt haben wir 1000 Stichprobenmittelwerte. Als Nächstes berechnen wir für jeden Stichprobenmittelwert den Verlust, so wie wir es für die (kleine) Stichprobe am Anfang dieses Abschnitts getan haben. Zu diesem Zeitpunkt haben wir 1000 Verlustwerte. Ein Histogramm dieser Verlustwerte ist in der folgenden Abbildung dargestellt. Zusätzlich zeigen wir den mittleren Verlust der Stichprobe, eine Schätzung des Wertes der Risikofunktion R 1 . 5, δ , mit einer roten vertikalen Linie.

Abbildung 45: Quadratischer Verlust für ein simuliertes Beispiel (Histogramm)

Der Wert 0 . 033 ist eine Schätzung für den Risikowert bei dem Zustand μ = 1 . 5. Wir können dieses Experiment für viele Zustände wiederholen und das Risiko als Funktion des Zustands auftragen. Die folgende Abbildung zeigt die Risikofunktion für verschiedene Werte von μ. Dies ist die primäre Art und Weise, in der die Risikofunktion untersucht werden muss, d. h. als Funktion des Zustands.

Abbildung 46: Schätzungen der Risikofunktion für den quadratischen Verlust und die Stichprobenmittelwert-Entscheidungsfunktion mit Hilfe von Simulationen

Wenn die Entscheidung δ nicht mit dem wahren Mittelwert μ übereinstimmt, liegt das daran, dass entweder unsere Schätzung den Mittelwert unterschätzt (δ < μ) oder ihn überschätzt (δ > μ). In bestimmten Situationen kann eine dieser beiden Möglichkeiten schlechter sein als die andere. Ähnlich dem gewichteten Verlust, den wir für das E-Mail-Klassifizierungsproblem diskutiert haben, können wir eine Verlustfunktion definieren, die diese beiden Szenarien unterschiedlich modelliert. Nehmen wir an, dass eine Unterschätzung des Mittelwerts 100-mal schlimmer ist als eine Überschätzung des Mittelwerts um den gleichen Betrag. Die Verlustfunktion, die dieses Ungleichgewicht kodiert, kann wie folgt definiert werden:

xxx

Unter Verwendung dieser Verlustfunktion und der Entscheidungsfunktion (Stichprobenmittelwert) kann die Risikofunktion durch Simulation geschätzt werden. Wir führen diese Simulation wie zuvor durch und fassen das Ergebnis in der folgenden Abbildung zusammen.

Abbildung 47: Schätzungen der Risikofunktion für den gewichteten quadratischen Verlust und die Stichprobenmittelwert-Entscheidungsfunktion mit Hilfe von Simulationen

Wie Sie sehen können, hat die Risikofunktion für den gewichteten quadratischen Verlust viel höhere Werte als die des (gewöhnlichen) quadratischen Verlusts. Dies legt nahe, dass wir eine Entscheidungsfunktion wählen sollten, die den Mittelwert überschätzt. Man könnte also δ x = x + c für einen positiven Wert c wählen. Die folgende Abbildung zeigt ähnliche Simulationen für verschiedene Werte von c.

Die horizontale Achse zeigt den wahren Mittelwert der Gauß-Verteilung und die vertikale Achse hat die Risikowerte (erwartete Verluste). Jeder Punkt im Diagramm stellt eine beobachtete Stichprobe und das damit verbundene Risiko dar. Die verschiedenen Farben stehen für verschiedene Entscheidungsfunktionen, die durch c parametrisiert sind.

Abbildung 48: Schätzungen der Risikofunktion für den gewichteten quadratischen Verlust und die verzerrte Mittelwert-Entscheidungsfunktion für verschiedene Verzerrungen anhand von Simulationen

Wie Sie sehen können, schneiden einige verzerrte Schätzer c > 0 besser ab als der unverzerrte Schätzer c = 0 . Unter den von uns getesteten Schätzern scheint c = 0 . 2 der beste zu sein. Auch eine zu starke Verzerrung c = 0 . 6 schneidet schlechter ab als der unverzerrte Schätzer.

Für das Problem der Mittelwertschätzung für eine Gauß-Verteilung mit unbekannter Standardabweichung haben wir gesehen, wie die Wahl der Verlustfunktion die Verwendung einer Entscheidungsfunktion gegenüber einer anderen diktiert. Bevor wir diesen Abschnitt abschließen, wollen wir noch eine letzte Verlustfunktion betrachten. Manchmal hängt der Verlust, der entsteht, wenn eine Entscheidung getroffen wird, die vom wahren Zustand abweicht, auch von der Größe des wahren Zustands ab. Hier ist ein Beispiel für eine Verlustfunktion, die einen solchen Fall modelliert:

xxx

Bei einer solchen Verlustfunktion gehen wir davon aus, dass der entstehende Verlust bei einer Differenz von z. B. μ - δ = 1 größer ist, wenn μ klein ist (im absoluten Wert), als wenn μ groß ist (im absoluten Wert). Hier einige Zahlen zur Veranschaulichung. Angenommen, der wahre Mittelwert ist μ = 1 und unsere Entscheidung ist δ = 1 . 5, d. h. sie liegt um 0 . 5 daneben. Dann beträgt der Verlust L 1, 1,5 = 1 - 1,5 2 12 + 1 = 0,125. Nehmen wir nun an, der wahre Zustand ist μ = 10, und wir liegen um 0,5 d. h. daneben, d. h. unsere Entscheidung ist δ = 10 . 5. Dann ist der Verlust L 10, 10,5 = 10 - 10,5 2

102 + 1

≈ 0 . 003. Wie zuvor wollen wir die Risikofunktion anhand einer Simulation untersuchen. Die folgende Abbildung zeigt die Ergebnisse.

Auf der horizontalen Achse befinden sich die Werte des wahren Zustands, d. h. der wahre Mittelwert der Gauß-Verteilung. Auf der vertikalen Achse befinden sich die Risikowerte, d. h. der erwartete Verlust. Jeder Punkt im Diagramm steht für eine beobachtete Stichprobe. Wie Sie sehen können, ist mit kleineren (absoluten) Werten des Mittelwerts ein höheres Risiko verbunden.

Abbildung 49: Schätzungen der Risikofunktion für den quadratischen Verlust, gewichtet nach Größe des Mittelwerts und des Stichprobenmittelwerts als Entscheidungsfunktion anhand von Simulationen

Die früheren Risikofunktionen schienen als Funktionen des Mittelwerts mehr oder weniger konstant zu sein. Diese Risikofunktion hingegen ist gekrümmt. Schließlich können wir die Ideen der unterschiedlichen Bestrafung von Unter- und Überschätzungen zusammen mit der Größe des Mittelwerts in unserer Verlustfunktion kombinieren:

xxx

Wie zuvor bevorzugen wir, motiviert durch die Verlustfunktion, eine Überschätzung gegenüber einer Unterschätzung und könnten eine Entscheidungsfunktion in Betracht ziehen, die eine (möglicherweise) verzerrte Schätzung des Mittelwerts ist: δ x = x + c, c ≥ 0. Wiederum werden wir Simulationen verwenden, um die Risikofunktionen für verschiedene Verzerrungen, c, zu untersuchen.

Diese Abbildung ist wie die vorherige Abbildung aufgebaut, nur dass sie jetzt die Schätzungen der Risikowerte (erwartete Verluste) (vertikale Achse) für verschiedene wahre Zustände (horizontale Achse) für verschiedene Entscheidungsfunktionen enthält. Die Punkte, die den verschiedenen Entscheidungsfunktionen entsprechen, sind unterschiedlich eingefärbt. Wie zuvor steht jeder Punkt für eine beobachtete Stichprobe.

Abbildung 50: Schätzungen der Risikofunktion für den nach Richtung und Größe gewichteten quadratischen Verlust und verzerrte Mittelwert-Entscheidungsfunktion für verschiedene Verzerrungen anhand von Simulationen

Die Kernelemente der statistischen Entscheidungstheorie sind folgende:

- die Größe, die wir identifizieren wollen, d. h. den Zustand θ unter allen möglichen Zuständen Θ

- X eine Zufallsvariable, die der zu beobachtenden Information entspricht, und x der beobachtete Wert dieser Zufallsvariablen

- eine Menge von Entscheidungsfunktionen, δ ∈ Δ, wobei jede Entscheidungsfunktion δ eine Entscheidung δ X oder δ x zuweist, die Werte im Zustandsraum Θ annimmt

- eine Verlustfunktion L, die die Qualität einer Entscheidungsfunktion (einer Zufallsvariablen) bewertet

- eine Risikofunktion, die die Verlustfunktion auf eine einzige Zahl reduziert, indem ihr Erwartungswert berechnet wird

Betrachten wir jedes dieser Elemente aus dem letzten Beispiel dieses Abschnitts, d. h. die Schätzung des unbekannten Mittelwerts einer Gauß-Verteilung. Der Zustandsraum beträgt Θ = - ∞ , ∞ , alles reelle Zahlen. Der wahre Zustand ist eine Zahl θ = μ ∈ - ∞ , ∞ . Eine Zufallsbeobachtung X = X1,X2, ...,Xn ist eine Zufallsstichprobe der Größe n, und die beobachteten Werte werden mit x = x1, x2, ..., xn bezeichnet. Als Nächstes betrachten wir den Schätzer des Stichprobenmittelwerts als mögliche Entscheidungsfunktion: δ X = X. Ein Beispiel für eine Verlustfunktion, die wir betrachtet haben, war der quadratische Verlust: L μ, δ = μ - δ X 2 = μ - X 2. Es ist zu beachten, dass es sich hierbei um eine Zufallsvariable handelt. Die Risikofunktion R schließlich ist der erwartete Verlust.

## 5.2 Maximum Likelihood, Minimax und Bayes

Maximum-Likelihood-, Minimax- und Bayes-δ-Funktion für eine gegebene Verlustfunktion L. Wie wir jedoch sehen werden, ist dieser Prozess nicht einfach und es gibt einige verschiedene Ansätze. Wir beginnen mit einigen grundlegenden Beispielen und geben eine Zusammenfassung für den allgemeinen Fall.

Nehmen wir an, wir wollen den Mittelwert einer Gauß-Verteilung N μ, 1 schätzen. Mit anderen Worten: wir haben eine Gauß-Verteilung mit unbekanntem Mittelwert μ und bekannter Standardabweichung σ = 1. Zur Vereinfachung betrachten wir nur eine Stichprobe mit einer einzigen Zahl X N μ, 1, deren beobachteter Wert mit x bezeichnet wird. Außerdem verwenden wir die quadratische Standardverlustfunktion L μ, δ = μ - δ 2. Betrachten wir die Zwei-Entscheidungsfunktion δ1 X = X δ1 x = x und δ2 X = 2 δ2 x = 2 . Die Risikofunktionen lauten R μ, δ1 = E μ - X 2 = Var X = 1 und R μ, δ2 = E μ - 2 2 = μ - 2 2 . Um diese beiden Entscheidungsfunktionen zu vergleichen, vergleichen wir ihre entsprechenden Risikofunktionen. Wir möchten die Entscheidungsfunktion verwenden, die ein geringeres Risiko aufweist. Die Antwort hängt jedoch vom wahren Wert von μ ab. Wenn 1 < μ < 3, dann ist δ2 die bessere Entscheidungsfunktion, weil sie ein geringeres Risiko hat. Ist jedoch μ < 1 oder μ > 3, dann ist δ1 die bessere Entscheidungsfunktion, weil ihr Risiko geringer ist. Wenn μ = 1 oder μ = 3 ist, haben die beiden Entscheidungsfunktionen das gleiche Risiko, so dass man sich für eine von ihnen entscheiden kann. Die folgende Abbildung zeigt einen Graphen der beiden Risikofunktionen.

Auf der horizontalen Achse sind die Werte des wahren Zustands aufgetragen, d. h. der wahre Mittelwert der Gauß-Verteilung. Auf der vertikalen Achse ist das Risiko, also der erwartete Verlust, aufgetragen. Wenn zum Beispiel der wahre Mittelwert Null ist, beträgt der erwartete Verlust (das Risiko) für die erste Entscheidungsfunktion δ1 1 und der erwartete Verlust der zweiten Entscheidungsfunktion δ2 etwa 5.

Abbildung 51: Risikofunktionen für zwei Entscheidungsfunktionen: Schätzung des Mittelwerts auf der Grundlage einer einzelnen Beobachtung aus einer Gauß-Verteilung

**Maximales Risiko**

Dies ist der Maximalwert der

Risikofunktion über alle mögliche Werte deswahren Zustands.

Wie Sie sehen können, liegt keine der beiden Risikofunktionen vollständig unter der anderen. Mit anderen Worten: Keine der beiden Risikofunktionen ist gleichmäßig (für alle Werte von μ) niedriger. Zur Klärung benötigen wir eine einzige Zahl für die Risikofunktion, die einer bestimmten Entscheidungsfunktion entspricht. Eine Möglichkeit ist das **maximale Risiko** (Kasy, 2014): R μ, δ = maxμR μ, δ . In unserem Fall ist R μ, δ1 = 1 und R μ, δ2 = ∞. Auf der Grundlage dieser Zusammenfassung der Risikofunktion (in einer einzelnen Zahl) werden wir daher δ1 als die bessere Entscheidungsfunktion wählen.

Eine andere Möglichkeit, eine einzige Zahl für die Risikofunktion einer Entscheidungsfunktion zu erhalten, ist die Berechnung des **Bayes-Risikos**. Wie Sie sich vielleicht erinnern, besteht eine Bayes-Formulierung zur Schätzung eines unbekannten Parameters (in diesem Fall μ) darin, eine A-priori-Verteilung für diesen Parameter zu wählen. Nehmen wir an, wir wählen die Prior μ. Dann ist das Bayes-Risiko definiert durch (Kasy, 2014)

RB δ = Eprior μ R μ, δ = ∫-∞

∞ R μ, δ prior μ dμ

Wenn die Verteilung der Prior diskret ist, würden wir das Integral durch eine Summe ersetzen. Mit anderen Worten: Das Bayes-Risiko behandelt den unbekannten Parameter als Zufallsvariable, gibt ihm eine A-priori-Verteilung und berechnet dann das erwartete Risiko unter Verwendung dieser A-priori-Verteilung. Für unser obiges Beispiel setzen wir einen Standard-Gauß-Prior ein: μ prior = N 0, 1 . Nun können wir das Bayes-Risiko für jeden der beiden Schätzer berechnen:

xxx

Wie Sie sehen können, ist das Bayes-Risiko für δ1 geringer, und wir würden diese Funktion als Entscheidungsfunktion wählen.

Betrachten wir ein anderes Beispiel. Wir wollen die (unbekannte) Erfolgswahrscheinlichkeit aus einer Bernoulli-Verteilung schätzen: Bernoulli p . Wir erinnern uns, dass eine Variable X der Bernoulli-Verteilung folgt, wenn sie genau zwei Werte annehmen kann: 0 mit der Wahrscheinlichkeit 1-p und 1 mit der Wahrscheinlichkeit p. Wir verwenden eine Beobachtung mit n unabhängigen und identisch verteilten Variablen: X = X1, ...,Xn i.i.dBernoulli p . Die realisierten Werte betragen x = x1, ..., xn . Wir werden die beiden nachstehenden Entscheidungsfunktionen anwenden, die einem Beispiel aus Wasserman (2004) entnommen sind.

Berechnen wir nun die Risikofunktionen für jede dieser Entscheidungsfunktionen, damit wir sie vergleichen können:

xxx

Nehmen wir an, wir verwenden eine Beobachtung mit n=400 Zahlen. Die folgende Abbildung zeigt den Graphen zweier Risikofunktionen.

Die horizontale Achse enthält die wahren Werte des wahren Zustands (die Erfolgswahrscheinlichkeit p einer Bernoulli-Verteilung). Die vertikale Achse enthält das Risiko (erwarteter Verlust), das jedem dieser wahren Zustände entspricht. Wenn beispielsweise der wahre Zustand p klein oder groß ist (weg von 0,5), ist das mit der ersten Entscheidungsfunktion δ1 verbundene Risiko geringer. Bei einem Wert von p nahe 0,5 hat die zweite Entscheidungsfunktion δ2 jedoch ein geringeres Risiko.

Abbildung 52: Risikofunktionen für zwei Entscheidungsfunktionen: Schätzung des Wahrscheinlichkeitsparameters der Bernoulli-Verteilung auf der Grundlage einer Stichprobe von zehn Zahlen

Es ist zu beachten, dass keiner der beiden Werte gleichmäßig niedriger ist als der andere. Daher können wir unter Berücksichtigung aller Werte nicht sagen, welche besser ist. Nach dem Vorbild des vorangegangenen Beispiels können wir für jede Risikofunktion eine Zusammenfassung mit einer Zahl berechnen und diejenige mit dem niedrigeren Wert wählen. Berechnen wir zunächst das maximale Risiko für jede der beiden Entscheidungsfunktionen:

xxx

Die bessere Entscheidungsfunktion (diejenige mit dem geringeren maximalen Risiko) ist daher δ2.

Eine andere Möglichkeit zur Berechnung einer Zusammenfassung in einer einzelnen Zahl besteht darin, das Bayes-Risiko zu berechnen. Dazu müssen wir eine A-priori-Verteilung für p wählen. Wählen wir prior( p) = Beta p; 2, 2 , eine Beta-Verteilung. Als Nächstes behandeln wir p als Zufallsvariable (gemäß dieser A-priori-Verteilung) und berechnen den Erwartungswert der Risikofunktionen laut dieser Prior:

xxx

Somit hat die erste Entscheidungsfunktion ein (geringfügig) besseres Bayes-Risiko.

Wir sind nun bereit, unsere Beobachtungen zusammenzufassen. In den beiden oben betrachteten Beispielen hatten wir im Grunde nur zwei konkurrierende Entscheidungsfunktionen. Nehmen wir nun an, dass Δ alle möglichen Entscheidungsfunktionen enthält, aus denen wir wählen können. Jede betrachtete Entscheidungsfunktion δ ist also ein Element von Δ. Als Nächstes bezeichnen wir den Parameter, den wir schätzen (den wahren Zustand) als θ. Im ersten Beispiel hatten wir θ = μ und im zweiten Beispiel haben wir θ = p. Als Nächstes betrachten wir die Sammlung aller möglichen Werte von θ als Θ. Im Falle eines unbekannten Mittelwertes für eine Gauß-Verteilung, Θ = - ∞ , ∞ und im Falle des unbekannten Wahrscheinlichkeitsparameters p für die Bernoulli-Verteilung, Θ = 0, 1 . Für eine feste Verlustfunktion wissen wir, dass das Risiko auf dem wahren Wert von θ und der Entscheidungsfunktion δ beruht:

**Minimales Risiko**

Für jede Entscheidungsfunktion,

berechnen wir das Risiko für alle

Werte des wahren Zustands.

Dann, unter all diesen

maximalen Risiken, nehmen wir das kleinste über die verschiedenen Entscheidungsfunktionen hinweg.

xxx

Wenn wir nur zwei Risikofunktionen haben, haben wir das Maximum der Risikofunktion ermittelt, indem wir den wahren Parameter variierten. Dann haben wir unter allen maximalen Risikowerten den kleineren ausgewählt. Wenn wir dies für alle δ ∈ Δ tun wollen, dann haben wir das **Minimax-Risiko** definiert durch

Rminimax = min

**Minimax-Risiko**

Für jede Entscheidungsfunktion berechnen wir das schlimmste (maximale) Risiko für alle Werte des wahren Zustands. Dann nehmen wir von all diesen maximalen Risiken das kleinste über die verschiedenen Entscheidungsfunktionen hinweg.

δ∈Δ

max

θ∈Θ

R μ, δ

Sie können sich wahrscheinlich denken, warum dies Minimax genannt wird! Auch in unseren Beispielen haben wir die maximalen Risiken für jede der Entscheidungsfunktionen berechnet: R δi = max

θ ∈ Θ

R θ, δi für

i = 1, 2. Dann wählten wir das kleinere der beiden Risiken: min

δi ∈ Δ

R δi mit Δ = δ1, δ2 .

Eine andere Betrachtungsweise ist das "beste (min) schlimmste (max) Risiko". Schließlich wird die Entscheidungsfunktion, die diesen Minimax-Wert erreicht, als Minimax-Entscheidungsfunktion (Schätzer) bezeichnet:

xxx

Anders ausgedrückt:

xxx

Das Bayes-Risiko wurde definiert als der Erwartungswert des Risikos in Bezug auf eine A-priori-Verteilung des unbekannten Parameters. Für unsere Beispiele haben wir das Bayes-Risiko für jede der beiden Entscheidungsfunktionen berechnet und dann diejenige mit dem niedrigeren Bayes-Risiko gewählt. Wenn wir viele Entscheidungsfunktionen in Δ haben, tun wir im Grunde das Gleiche. Wir wollen

**Bayes-Entscheidungsfunktion**

Diese Entscheidungsfunktion minimiert das Bayes-Risiko.

xxx

Die Entscheidungsfunktion, die dieses Minimum erreicht, wird als Bayes-Entscheidungsfunktion (oder Bayes-Schätzer) bezeichnet. Wenn wir diese **Bayes-Entscheidungsfunktion** mit δBayes bezeichnen, dann

Xxx

Anders ausgedrückt:

Xxx

## 5.3 Zulässigkeit und Stein-Paradoxon

Bei zwei Entscheidungsfunktionen (Schätzern), δ1 und δ2, für einen unbekannten Parameter, θ, sagen wir, dass δ1 gegenüber δ2 zulässig ist, wenn die mit δ1 assoziierte Risikofunktion gleichmäßig kleiner als die mit δ2 assoziierte Risikofunktion (oder gleich) ist. Das bedeutet, dass R θ, δ1 ≤ R θ, δ2 für jeden möglichen Wert von θ ist. Wir betrachten noch einmal den Graphen der Risikofunktionen zweier Entscheidungsfunktionen aus dem letzten Abschnitt (letzte Abbildung in Abschnitt 5.2 oben). Es ist zu beachten, dass keines der Diagramme der beiden Risikofunktionen vollständig unter dem anderen liegt. Die Risikofunktion, die δ1 entspricht, ist nur für Werte von θ = p, die weit von p = 1/2 entfernt sind, niedriger als die von δ2. Betrachten wir nun die verschiedenen Risikofunktionen in der letzten Abbildung von Abschnitt 5.1.

**Zulässige Entscheidungsfunktion**

Die Risikofunktion dieser Entscheidungsfunktion wird von jeder anderen Entscheidungsfunktion dominiert. Das heißt, ihr Risiko (erwarteter Verlust) ist für jeden möglichen Wert des wahren Zustands geringer als das jeder anderen Entscheidungsfunktion.

Es zeigt sich, dass die Entscheidungsfunktion, die c = 0 . 2 entspricht, unter jeder anderen Risikofunktion liegt. Daher wird δ0 . 2 X = X + 0 . 2 als für jede andere Risikofunktion **zulässig** bezeichnet. Ihr Graph liegt vollständig unter jeder anderen Risikofunktion. Mit anderen Worten: Das mit dieser Entscheidungsfunktion verbundene Risiko ist für jeden Wert des Parameters θ = μ geringer als andere Risikofunktionen. Eine der schönen Eigenschaften von Bayes-Schätzern ist die folgende Tatsache: Wenn δ ein eindeutig bestimmter Bayes-Schätzer ist, dann ist er zulässig, wenn eine A-priori-Verteilung für den unbekannten Parameter θ vorliegt. Mit anderen Worten: Die Risikofunktion eines eindeutigen Bayes-Schätzers ist für jeden Wert des Parameters θ nicht größer als jede andere Risikofunktion (die einer beliebigen anderen Entscheidungsfunktion entspricht)! Wir werden uns nicht mit dem Beweis dieser Tatsache befassen, aber es ist eine wichtige Tatsache, die man sich merken sollte.

Angenommen, wir haben drei unverbundene Größen, von denen bekannt ist, dass sie (unabhängigen) Gauß-Verteilungen mit bekannter (Einheits-)Varianz, aber unbekannten Mittelwerten folgen. Mit anderen Worten: Wir haben Xi N μi, 1 für i = 1, 2, 3, und X1, X2, X3 sind unabhängig. Wir beobachten eine Realisierung von jedem Gauß-Verteilung und wollen diese Realisierung verwenden, um die drei unbekannten Mittelwerte zu schätzen. Die natürlichste Schätzung ist der Wert der Beobachtung selbst. Die natürliche Entscheidungsfunktion ist δ X1,X2,X3 = X1,X2,X3 . Mit anderen Worten: Unser bester Schätzwert für den (unbekannten) Mittelwert μ1 für die erste Gauß-Verteilung ist X1, und Ähnliches gilt für die beiden anderen. Um diese Entscheidungsfunktion zu bewerten, verwenden wir den quadratischen Standardverlust:

xxx

Die zugehörige Risikofunktion, der Erwartungswert dieses Verlustes in Bezug auf die gemeinsame Verteilung von X1,X2,X3 , kann aufgrund der Unabhängigkeit Term für Term berechnet werden:

xxx

Man könnte meinen, dass dies eine ziemlich gute Entscheidungsfunktion ist. Betrachten wir nun die (James-Stein-) Entscheidungsfunktion, die durch δ′ X1,X2,X3 = 1 - 1

S2 X1,X2,X3 gegeben ist, wobei

S2 = X1 2 + X2 2 + X3 2 . Mit anderen Worten: Diese Entscheidungsfunktion schätzt den ersten Mittelwert anhand von 1 - 1 S2 X1 und die anderen Mittelwerte in ähnlicher Weise. Es zeigt sich, dass das mit dieser Entscheidungsfunktion verbundene Risiko Folgendes beträgt:

xxx

Es ist zu beachten, dass 1 S2 > 0 mit der Wahrscheinlichkeit 1 ist. Daher ist E 1 S2 > 0 und somit R μ, δ′ < 3. Da dies für alle Werte von μ = μ1, μ2, μ3 gilt, ist die James-Stein-Entscheidungsfunktion (Schätzung) für die drei Mittelwerte in der Tat zulässig! Das Ergebnis ist aus folgendem Grund recht paradox. Um μ1 zu schätzen, verwendet unsere Entscheidungsfunktion den Schätzer

xxx

der von X2 und X3 abhängt. X1 ist jedoch unabhängig von X2 und X3! In einem konkreten Beispiel ist μ1 der durchschnittliche Teepreis in China, μ2 ist die durchschnittliche Temperatur auf der Marsoberfläche und μ3 ist das durchschnittliche Menge Futter in Kilogramm, die ein Eisbär verzehrt. Das Stein-Paradoxon besagt: Wenn wir diese drei Durchschnittswerte gleichzeitig schätzen wollen, brauchen wir zum Beispiel die Informationen über die Temperaturen auf dem Mars und den Nahrungsverbrauch von Eisbären als Grundlage für unsere Schätzung des durchschnittlichen Teepreises in China! Auf diese Weise wird das Risiko (der erwartete quadratische Verlust) verringert. Die sich daraus ergebende Schätzung wird im Durchschnitt besser abschneiden als eine Entscheidungsfunktion (Schätzung), deren Schätzungen für die drei Mittelwerte jeweils nur von ihren jeweiligen Werten abhängen. Intuitiv lässt sich dies dadurch erklären, dass die drei Größen zwar nichts miteinander zu tun haben, aber eine schlechte Schätzung für einen der Mittelwerte durch die anderen Variablen ausgeglichen werden soll. Schauen Sie sich die Verlustfunktion noch einmal an. Für die James-Stein-Entscheidungsfunktion lautet der zugehörige Verlust

xxx

Daher hängt jeder quadrierte Term von den Informationen aus allen drei Variablen ab. In gewisser Weise wollen wir, wie oben erwähnt, gleichzeitig eine gute Schätzung für alle drei Mittelwerte.

### Zusammenfassung

Diese Lektion dient als allgemeine Einführung in die Kernkonzepte der statistischen Entscheidungstheorie. Die wichtigsten Elemente der statistischen Entscheidungstheorie wurden in Abschnitt 5.1 vorgestellt. Die wichtigsten Erkenntnisse aus diesem Abschnitt waren die relevanten Definitionen des Zustandsraums, der Entscheidungsfunktionen, der Verlustfunktionen und schließlich der Risikofunktion. Dies sind die Elemente, in Bezug auf die Probleme der statistischen Entscheidungstheorie formuliert, analysiert und bewertet werden.

Die Konsolidierung einer Risikofunktion zu einer einzigen Zahl ist keine sehr einfache Aufgabe. Es ist jedoch ein Ideal, das wir anstreben, um die beste Entscheidungsfunktion für ein bestimmtes Problem zu wählen. Zu diesem Zweck haben wir das Minimax- und das Bayes-Risiko und die damit verbundenen Entscheidungsfunktionen besprochen.

Schließlich haben wir den Begriff der Zulässigkeit erörtert. Dieser Begriff ist eine Eigenschaft einer Entscheidungsfunktion, die wünschenswert ist. Das Konzept der Zulässigkeit wurde anhand einiger Beispiele erörtert. Das Stein-Paradoxon war das letzte Thema des Abschnitts 5.3. Wir gaben ein Beispiel für einen James-Stein-Schätzer (Entscheidungsfunktion), der zulässig ist.