Die quantitative sensorische Weinbewertung: ihre Grenzen und die Möglichkeiten ihrer Optimierung

Quantitative sensory testing of wines: Limitations and possibilities for optimization

Armin Kobler, Versuchszentrum Laimburg, Auer, Italien

Zusammenfassung

Für die sensorische Analyse stehen Dank ihres Stellenwertes in der Weinforschung vielfältige Methoden zur Verfügung. Nach einer Beschreibung der wichtigsten Verfahren wird auf die Verwendung von nichtstrukturierten Skalen und die Kosterprüfung eingegangen. Diese schützt nicht nur vor der Verrechnung unzuverlässiger Daten sondern kann auch der Interpretation widersprüchlicher Endresultate dienen. Diskutiert werden zudem die Probleme, welche sich bei der Abhaltung von Kostwettbewerben ergeben. Eine die Tagesform beurteilende Prüferselektion und ein verstärkter Einsatz der Randomisierung sollen helfen, die Ergebnisse solcher Veranstaltungen nachvollziehbarer zu gestalten.

Summary

Due to its importance in wine research there are various methods available for the sensory evaluation of wines. The most relevant methods are described and the use of non-structurated scales and tasters examination is dealt with in detail. The latter is a useful tool that helps to avoid the calculation with unreliable data as well as the interpretation of contradictory end results. In this article we discuss the problems that can arise during a wine competition. Thus, a selection of the tasters conducted on a daily basis and the application of randomization will help to make the results of such events more understandable.

Einleitung

Die chemische Analyse der Weine hat in den letzten Jahren beachtliche Fortschritte gemacht. Besonders die GC- und HPLC-Methoden haben es erlaubt, sensorisch wichtige Substanzen, wie z.B. jene des Gerbstoff- und Aromakomplexes, detailliert quantitativ zu bestimmen. (LINSKENS und JACKSON 1988). Bezüglich Kosten- und Zeitersparnis bietet die FTIR-Methode interessante Perspektiven (PATZ et al. 2000).

Die Daten der instrumentellen Weinanalytik ersetzen trotz ihres wichtigen Beitrags in der Weinforschung nicht die Ergebnisse der sensorischen Analyse. Kosten bedeutet, die Weine zu "untersuchen, analysieren, beschreiben, definieren und klassifizieren" (RIBEREAU-GAYON et al. 1975) und bleibt nach wie vor das wichtigste Verfahren, um die Qualität zu bestimmen. Dies, weil der Wein aus einer fast unüberschaubaren Vielfalt an Substanzen, welche noch nicht alle identifiziert sind, zusammengesetzt ist und man immer noch zu wenig darüber weiß, wie die Verhältnisse der Inhaltsstoffe untereinander sich auf die Qualität auswirken.

Die Ergebnisse der Sinnesprüfung sind notwendig, um die Bevorzugung der Konsumenten für das eine oder andere Produkt zu untersuchen, zur Qualitätssicherung im Produktionsablauf, zur Verbesserung bereits auf dem Markt befindlicher sowie für die Entwicklung neuer Produkte. Besonders zu letzterem Zweck wurden verschiedene Verfahren entwickelt, welche die Quantifizierung der Sinneseindrücke Aussehen, Geruch und Geschmack aber auch die Messung der trigeminalen und haptischen Wahrnehmungen erlauben. Eine Übersicht bieten AMERINE et al. (1965), DAEPP (1966b), NEUMANN et al. (1983), NOBLE (1988), und FROST und NOBLE (2002).

Die in der önologischen Forschung weitverbreiteten Verfahren, wie der Dreieckstest von BENGTSSON (1943) und der Duo-Trio-Test von PERYAM und SWARTZ (1950) beruhen auf dem Prinzip der Unterscheidbarkeit. Die ebenso oft verwendete Rangsummenmethode von PAUL (1967) bedient sich hingegen der Bevorzugung. Beide Tests sind weit verbreitet, weil sie nicht nur leicht anwendbar und genügend sensibel sind (AMERINE et al. 1965, WEISS et al. 1972, UBIGLI (1990a) sondern weil sie auch geeignet sind, grundlegende Fragen in der Weinforschung zu beantworten: Besteht ein statistisch signifikanter Unterschied zwischen den sensorisch geprüften Produkten und wenn, welcher Wein wird als erster, zweiter, usw. gereiht?

Die Verwendung dieser Verfahren bringt aber auch Einschränkungen psychologischer und statistischer Art mit sich, welche in der Natur der verwendeten Messskalen liegen. KÖHLER et al. (1983) haben die Erfahrung gemacht, dass der Koster oft nur ungern, wie verlangt, verschiedene Ränge vergibt, besonders dann, wenn die Unterschiede zwischen den Proben sehr gering sind oder die Prüfer gar nicht imstande sind, diese zu unterscheiden. Das Vergeben gleicher Ränge in der Form eines Mittelwertes aus den betroffenen Plätzen wird von verschiedenen Autoren (WEISS et al. 1972) als problematisch angesehen, genau so wie das zufällige Vergeben der Ränge an die nicht unterscheidbaren Proben.

Ein weiteres Problem betrifft die Interpretation der Ergebnisse dieser Tests. Wenn der sensorische Vergleich zweier im Keller verschieden behandelter Weine z.B. einen signifikanten Unterschied ergibt, bedeutet dies, dass die Kostkommission eindeutig und reproduzierbar den Wein, der einem bestimmten Verfahren unterworfen wurde, dem anderen bevorzugt hat. Es ist in diesem Fall aber

nicht richtig zu sagen, dass eine önologische Praxis das Produkt signifikant verbessert hat. Um den Begriff der Signifikanz auf die getesteten Verfahren auszudehnen, ist es unbedingt notwendig, mit Wiederholungen aus dem gleichen oder verschiedenen Ausgangsmaterial, je nach Versuchsdesign, zu arbeiten.

Eine Messskala, welche nicht nur eine Klassifizierung, sondern auch die Abstände zwischen den Proben wiedergibt, ist die strukturierte Intervallskala. Die damit ermittelten Werte sind zudem geeignet, allen mathematischen und statistischen Berechnungen zu genügen (WEISS 1981a). Wenn die Skala in genügend viele Intervalle unterteilt wird, nähert sich die Verteilung der Werte einer Gauß'schen Verteilungskurve (LORENZ 1988). Kostschemen mit diesen Maßeinheiten sind sehr verbreitet wegen der Vielzahl an Einsatzmöglichkeiten. Beispiele sind bei AMERINE et al. (1965), DAEPP (1966b), UBIGLI (1990b) sowie FROST und NOBLE (2002) zu finden.

Mit der Verbreitung der beschreibenden Weinsensorik haben sich die nichtstrukturierten Kostschemen durchgesetzt. Sie werden in der weinbaulichen (IACONO et al. 1992) und kellerwirtschaftlichen Forschung (BERTUCCIOLI und ROSI 1992) eingesetzt sowie verwendet, um Weine verschiedener Herkünfte oder Macharten zu charakterisieren (SCHNEIDER und KRECKEL 1995, CLIFF et al. 2002), Tipizitätskriterien von Weinen bestimmter Anbaugebieten zu definieren (UBIGLI 1992a) oder das Konsumentenverhalten zu untersuchen (DELTEIL 2000). Ihre Skalen sind stetig, d.h. sie sind nicht in Klassen unterteilt. Während die numerischen Skalen auf Grund ihrer Unterteilungen nur eine begrenzte Anzahl an Wertausprägungen zulassen, sind bei den nichtstrukturierten Skalen nur die beiden Enden definiert. Dies ermöglicht dem Prüfer, ein sehr differenziertes Urteil abzugeben, auch weil er nicht von den verschiedenen semantischen Bedeutungen der Zahlen beeinflusst wird (WEISS et al. 1972). Auch die graphisch strukturierten Skalen ohne Nummerierung besitzen diese Vorteile, aber die statistische Auswertung mit parametrischen Methoden bringt Schwierigkeiten mit sich, da die Koster bestimmte Teile der Skala bevorzugen und so Assimetrien in der Datenverteilung verursachen (CASTINO 1983). STONE et al. (1974) schlagen nichstrukturierte Kostschemen mit horizontaler Ausrichtung vor, deren Extreme mit "schwach" und "stark" definiert sind und besonders in der beschreibenden Weinsensorik mit Erfolg eingesetzt werden. UBIGLI (1992b) und CASTELLARI et al. (2001) haben hingegen ein unstrukturiertes Schema in der Form eines Rades benutzt, in dem die Radien als Messskalen dienen. WEISS et al. (1972) stellen Skalen vor, welche diagonal in einem Quadrat angeordnet sind, und so den Zusammenhang Intensität/Punktierung graphisch verstärken sollen. Nicht zuletzt wegen ihrer Größe sind diese zur Quantifizierung weniger essentieller Parameter geeigneter (UBIGLI 1992b).

Kosterprüfung

Es ist allgemein akzeptiert, dass die Mitglieder einer Kostkommision gewissen Anforderungen genügen müssen (GIRARDOT et al. 1952, KRUM 1955, AMERINE et al. 1965, DAEPP 1966a, NEUMANN et al. 1983). WEISS (1981b) gibt die Verfügbarkeit, die Motivation, das Interesse und die Fähigkeit zu unterscheiden sowie verlässliche und stabile Urteile abzugeben, an. Die Übung hilft dann, die individuellen Fähigkeiten zu verstärken. Genauso verbreitet ist die Ansicht, dass auch nach bestandener Eignungsprüfung das sensorische Verhalten der Koster beobachtet werden muss, nachdem "die Sensibilität und die Reproduzierbarkeit der Prüfer die Richtung und die Gültigkeit der Resultate beeinflussen" (AMERINE et al. 1965). In bestimmten Situationen, aus physiologischen und psychologischen Gründen, können auch ausgebildete und ausgesuchte Koster Urteile abgeben, die nicht kohärent mit ihrer Schulung und auch nicht reproduzierbar sind. Es folgt daraus, dass die Kohärenz der Koster auch nach der Ausbildung geprüft werden muss, und, im Fall von Bewertungen, die zu sehr kontrastieren, auch ausgeschlossen werden können.

Für SCHRODT und JAKOB (1966) sowie UBIGLI (1986 und 1988) hängt die Objektivität und somit die Eignung eines Kosters, Mitglied eines Prüferpanels zu sein, von der Fähigkeit ab, gruppenkonforme Urteile abzugeben. KÖHLER et al. (1983) zweifeln diese Überlegung an, weil auf diese Weise neben ungeeigneten Kostern auch äußert sensible Prüfer ausgeschlossen werden. Mehr Akzeptanz hat hingegen die Forderung nach Wiederholbarkeit gefunden. NEUMANN et al. (1983) empfehlen die Verwendung eines mittleren Wiederholbarkeitsindexes um die Konstanz der Prüfer zu beurteilen. Dieser Index besteht aus der Standardabweichung +1 zwischen den doppelt gereichten Proben und sollte laut den Autoren bei erfahrenen Kostern den Wert von 1,5 nicht überschreiten.

Bereits 1948 haben OVERMAN und LI einen anderen Lösungsansatz vorgestellt. Dabei wird für jeden Koster eine Varianzanalyse berechnet, wobei die damit erhaltenen F-Werte die Fähigkeit der Koster zur Diskrimination und Wiederholbarkeit ausdrücken. Um berücksichtigt zu werden, müssen die Prüfer F-Werte einer Irrtumswahrscheinlichkeit von ≤1 % aufweisen. WILEY et al. (1957) waren ähnlicher Meinung. In einer ersten Auswahl mussten die Koster P ≤10 % aufweisen, als dauerhafte Kostkommissionsmitglieder ≤5%. Einen Test, der auf den gleichen Grundlagen beruht, wird von KÖHLER et al. (1983) vorgeschlagen. Die Vorgangsweise von OVERMAN e LI (1948) weist aber Unzulänglichkeiten auf, die deren Anwendung erschweren. Wenn nämlich die zu prüfenden Weine geringe Unterschiede aufweisen, werden die F-Werte aller, auch der Koster mit der besten Tagesform, nieder sein. Nur wenige, wenn überhaupt welche, genügen dann der festgelegten Irrtumswahrscheinlichkeit. Die F-Werte der Kommission hängen also nicht nur vom Eignungsniveau deren Mitglieder, sondern auch von den verkosteten Weinen ab. GIRARDOT et al. (1952), ausgehend davon, dass die Kostergruppe an sich genügend verlässlich die Produkte bewertet, vervollständigen ihre Methode mit der Verwendung des Konfidenzintervalls. Koster, welche gegenüber der Gruppe zu

ungenau kosten, d.h. F-Werte ergeben, die sich außerhalb der unteren Vertrauensgrenze (P ≤5 %) befinden, werden in der Auswertung nicht berücksichtigt.

Auf der Basis der angeführten Betrachtungen wird eine Vorgangsweise zur sensorischen Beurteilung von Versuchsweinen vorgeschlagen, welche auf den Einsatz nichtstrukturierter Skalen beruht und eine Prüfung der Koster beinhaltet.

Sensorische Bewertung von Versuchsweinen

Als Beispiel wird die Durchführung eines technologischen Kellerversuchs angeführt, in dem vier verschiedene Maischegärverfahren auf ihr Eignung getestet wurden.

Die Maische wurde unter ständigem Rühren auf acht Behälter aufgeteilt. Jede Variante wurde somit in zweifacher Wiederholung umgesetzt. Die acht Weine wurden nach der Gärung den üblichen Ausbaumethoden unterworfen, gefüllt, und einige Monate später einem Kosterpanel der Sektion Kellerwirtschaft im Verssuchszentrum Laimburg zur Beurteilung gereicht. Die Kommission setzte sich aus erfahrenen externen Weinkostern aber z.T. auch aus neuen Mitarbeitern zusammen. Alle waren aber mit den verwendeten Kostschemen vertraut. Die Parameter, nach denen gefragt wurde, sowie die End- und Eckpunkte ihrer Skalen, sind in Tabelle 1 angeführt.

Tab. 1: Abgefragte Parameter und verwendete Kostschemen.

Parameter	Kostschema	Skalenbeginn	Eckpunkt	Skalenende
Reintönigkeit	Abbildung 1	sehr unsauber		ganz reintönig
Geruch	Abbildung 1	einfach		vielfältig
Geruchsrichtung	Abbildung 1	frisch fruchtig		"marmeladig"
Typizität	Abbildung 1	sehr untypisch		sehr typisch
Genussreife	Abbildung 2	zu jung	optimal	zu alt
Gerbstoffgehalt	Abbildung 2	zu wenig	optimal	zu viel
Gerbstoffqualität	Abbildung 1	hart, grob, bitter		weich, voll
Gesamtqualität	Abbildung 1	schlecht		ausgezeichnet

Bei Parametern, wo das Optimum einer Eigenschaft mit dem Maximum zusammenfällt, wurden die Skalen von WEISS et al. (1972), verändert, wie in Abbildung 1 dargestellt, übernommen. Für die sensorischen Eigenschaften, wo die bestmögliche Ausprägung vor dem Maximum erfolgt, wurde ein Schema verwendet, dass noch stärker vom Original abweicht (Abb. 2).

Die Pegelung erfolgte mit zwei zufällig ausgesuchten Weinen der Versuchsserie. Diese wurden von den beteiligten Personen an Hand des verwendeten Kostschemas angesprochen und deren Eigenschaften zum Teil auch gegensätzlich diskutiert. Es wurde nicht auf eine einstimmiges Urteil gedrängt.

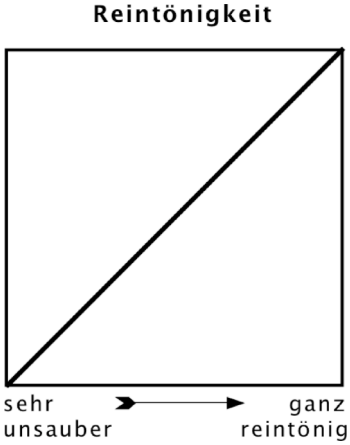

Abb. 1: Unstrukturierte Skala nach WEISS et al. (1972), verändert. Der Koster markiert auf der diagonalen Linie die Intensität der sensorische Empfindung (verkleinerte Darstellung).

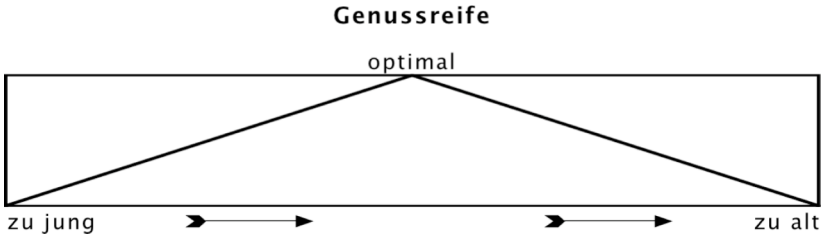

Abb. 2: Unstrukturierte Skala für Parameter, bei denen die optimale Ausprägung nicht mit dem Maximum übereinstimmt. Der Koster markiert auf der aufsteigenden oder absteigenden diagonalen Linie die Intensität der sensorische Empfindung (verkleinerte Darstellung).

Im Anschluss darauf wurden alle acht Versuchsweine in zweifacher Wiederholung den Kostern einzeln gereicht. Eine Hälfte der Koster bekam die 16 Proben in einer zufälligen Reihenfolge, der anderen Hälfte der Kommission wurden die Weine in der exakt umgekehrten Sequenz eingeschenkt. Dies soll Fehlbeurteilungen, welche auf der Reihenfolge der Proben beruhen, auf ein Minimum beschränken.

Die ausgefüllten Kostschemen wurden grafisch ausgewertet, die Werte dann mittels des Tabellenkalkulationsprogramms Microsoft Excel 2004 für Mac (Microsoft Corporation, Redmond) und dem Statistikprogramm SPSS 10.0 Mac OS Version (SPSS Inc., Chicago) verarbeitet. Zum Zweck der Kosterprüfung wurden für jeden Koster und alle Parameter einzelne Varianzanalysen berechnet, in der die acht gereichten Weine als Faktoren eingingen. Die so errechneten F-Werte sind ein gutes Maß für die Tagesform des Prüfers, sind sie doch der Quozient der durchschnittlichen Abweichungsquadratsummen zwischen den verschiedenen Weinen (Fähigkeit des Kosters, zwischen den Proben zu unterscheiden und Auspunktungsbereitschaft) und der Fehlervarianz (Abweichungen bei der Beurteilung der gleichen Weine).

In Abbildung 3 ist das Kosterverhalten einiger Prüfer am Beispiel des Parameters "Gesamtqualität" grafisch dargestellt. Die Enden der einzelnen Balken geben die Bewertung bei der zweimaligen Kost

der gleichen Proben wider. Je niedriger der Balken ist, desto besser ist die Reproduzierbarkeit. Je entfernter die Balken zueinander stehen, desto unterschiedlicher hat der Prüfer die verschiedenen Proben bewertet. Der Koster 8 zeichnet sich durch eine mittlere Auspunktungsbereitschaft und Reproduzierbarkeit aus, was sich in einem F-Wert von 3,08 niederschlägt. Ungenügende Kostperfomance legt der Prüfer 10 an den Tag: er differenziert kaum und reproduziert im Verhältnis sehr schlecht, was in diesem Fall einen F-Wert von 0,34 ergibt. Unwesentlich besser (0,66) bewertet der Koster 13 die Weine. Seine auf den ersten Blick gute Wiederholbarkeit ist auf die äußerst geringe Auspunktungsbereitschaft von nicht einmal 1,5 Beurteilungseinheiten zurückzuführen. Prüfer 12 nutzt die Skala sehr gut aus und kann seine Urteile auch sehr gut reproduzieren. Sein F-Wert von 21,15 ist erfahrungsgemäß als sehr hoch zu betrachten.

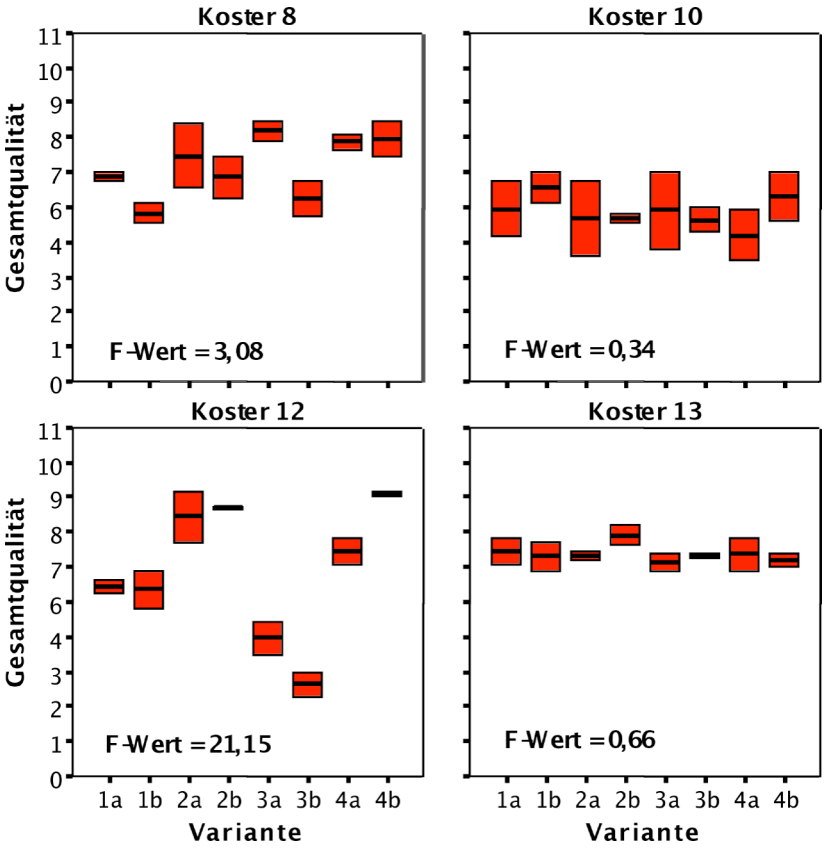

Abb. 3: Das Kostverhalten von vier Prüfern und ihre F-Werte. Die oberen und unteren Enden der Balken stellen die beiden Benotungen beim wiederholten Beurteilen der gleichen Weine dar.

Um die Gesamttagesleistung der Koster zu bestimmen, wird aus den einzelnen Parameter-F-Werten, in diesem Fall aus deren acht, ein Median errechnet. Dieser mittlere Wert ist im Gegensatz zum arithmetischen Mittel unempfindlicher gegen Ausreißer und spiegelt somit besser das Gesamt-Kostvermögen der einzelnen Koster über alle Parameter wider.

In der Folge wurde das Konfidenzintervall errechnet. Alle Koster, welche einen F-Wert oberhalb der unteren Konfidenzgrenze aufweisen, werden für die Endauswertung der Weinkost herangezogen. In

der Abbildung 4 sind die der Größe nach gereihten mittleren F-Werte der Koster sowie die untere Konfidenzgrenze, welche in unserem Fall den Wert 0,95 beträgt, veranschaulicht. In diesem einem Fall wurden 7 von 14 Kostern, meistens sind es aber zwischen zwei Drittel und drei Viertel der Teilnehmer, berücksichtigt. Ihre Werte wurden, da es sich um Messwiederholungen handelt, gemittelt und gingen in die varianzanalytische Endauswertung, mit den vier Maischegärverfahren als Faktoren, ein.

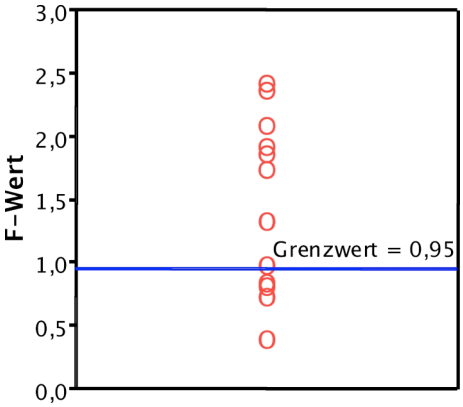

Abb. 4: Die nach Größe gereihten Koster-F-Werte und die untere Konfidenzgrenze, welche die berücksichtigten (darüber) von den unberücksichtigten Prüfer trennt.

Die durchschnittlichen F-Werte einer Kommission dienen nicht nur der notwendigen Kosterprüfung. Hohe F-Werte zeigen, dass die Kommission insgesamt sich leicht tat, die Weine differenziert und wiederholbar zu bewerten. Dies, weil das Niveau der beteiligten Koster hoch war und/oder die Unterschiede zwischen den sensorischen Eigenschaften der Weine beträchtlich waren. Ebenso kann man mittels der F-Werte nachvollziehen, bei welchen Kenngrößen sich die Koster leichter oder schwerer taten bzw. in welchen Eigenschaften sich die Proben mehr oder weniger unterscheiden.

Aus der Tabelle 2 geht hervor, dass die getesteten Maischegärverfahren keine signifikanten Unterschiede bezüglich der sensorischen Eigenschaften der Weine bewirkten. Einzig die Differenzen im empfundenen Gerbstoffgehalt näherten sich der Signifikanz. Ohne Kosterprüfverfahren wäre hier die Auswertung und Interpretation des Versuchs im wesentlichen beendet.

Tab. 2: F-Werte aus der Varianzanalyse (VA) der Kostergruppe, F- und p-Werte der Wein-Endauswertung, nach Parametern getrennt.

Parameter / F/p-Werte	Reintönigkeit	Geruch	Geruchsrichtung	Typizität	Genussreife	Gerbst.-Gehalt	Gerbst.-Qualität	Gesamtqualität
Mittelw. VA Koster Sign. >3,50	1,56	2,45	2,06	6,26	2,58	2,31	2,06	4,62
VA Weine Sign. >6,59	2,32	1,58	1,31	1,43	0,22	6,07	0,14	4,19
p-Werte VA Weine	0,22	0,33	0,39	0,36	0,88	0,06	0,93	0,10

Die Deutung der durchschnittlichen Koster-F-Werte kann aber wichtige zusätzliche Anhaltspunkte liefern. Ist der Einfluss eines Faktors auf einen bestimmten Parameter wirklich nur zufällig (Annahme

der Null-Hypothese in der Endauswertung) oder wirken sich die Techniken zwar auf die Sensorik aus, ihr Einfluss wird aber von den Kostern verschieden bewertet? Speziell bei hedeonistischen-bewertenden Fragestellungen kann dies der Fall sein. Besonders deutlich wird dies in unserem Fall am Beispiel der Parameter "Typizität" und "Gesamtqualität", für welche keine statistisch absicherbaren Unterschiede errechnet werden konnten. Die Koster selbst unterschieden relativ deutlich und reproduzierbar bezüglich diesen Parametern, was die an sich schon signfikanten F-Werte von 6,26 und 4,62 widerspiegeln, aber sie waren verschiedener Meinung bezüglich der Präferenz. Die Abbildung 5 verdeutlicht prägnant diese Wechselwirkung. Beide angeführten Prüfer bewerteten überdurchschnittlich differenzierend und verlässlich die Proben (F-Werte von 15,90 und 7,65), beurteilten sie aber sehr konträr hinsichtlich der angepeilten Typizität. Mittelt man solche Einzelurteile, dann heben sich erwartungsgemäß die vorher so deutlich ausgedrückten Unterschiede wieder auf.

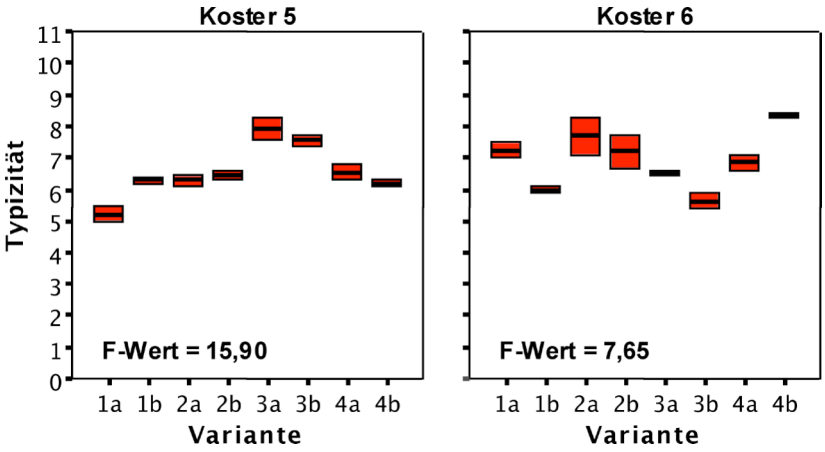

Abb. 5: Das gegensätzliche Kostverhalten zweier berücksichtigter Prüfer.

Die Aussagen, die nach der Betrachtung der Koster-F-Werte getroffen werden können, sind somit wesentlich aufschlussreicher: Die Koster unterscheiden deutlich zwischen den verschiedenen Maischegärverfahren bezüglich den Parametern "Typizität" und "Gesamtqualität", ihre Uneinigkeit verhindert aber eine allgemeingültige, statistisch signifikante Aussage. Dem potentiellen Anwender wird bewusst, dass die geprüften Gärtechniken sich unterschiedlich auf das sensorische Bild der Weine auswirken, er muss aber selbst in Erfahrung bringen, ob das vorhandene Differenzierungspotential in sein Produktionskonzept passt.

Weinkostwettbewerbe

Weinkostwettbewerbe sind Veranstaltungen, die dem Zweck dienen, qualitativ herausragende Produkte zu prämieren. Die Ergebnisse dienen den Konsumenten als Entscheidungshilfe und können das Ansehen positiv abschneidender Betriebe merkbar verbessern. Es ist dementsprechend im Sinne der Veranstalter, dass sich möglichst viele Produzenten am Wettbewerb beteiligen. Der großen Menge

an Proben steht eine zumeist begrenzte Anzahl an Juroren gegenüber. Um die Ergebnisse auf eine breitere Basis zu stellen, ist die Expertengruppe bezüglich beruflichen Hintergrund und geographischer Herkunft heterogen zusammengestellt. Dementsprechend differieren Kostvermögen und Präferenzen.

Nachdem nicht alle Koster die Gesamtzahl der eingereichten Produkte in der zur Verfügung stehenden Zeit prüfen können, werden Kostkommissionen gebildet. Deren Mitglieder bewerten alle ihnen zugeteilten Proben, sofern einzeln gereicht, in der gleichen Reihenfolge. Als Kostvermögen wird, falls überhaupt, einzig die Fähigkeit bewertet, in Einklang mit den anderen Kommissionsmitgliedern die Weine zu bewerten.

Dieser sehr verbreitete Ablauf bringt mehrere Probleme mit sich: Die Bewertung der Weine ist stark abhängig von den Ansprüchen der Kommission, welche sie bewertet, vom qualitativen Niveau der Serie, in der sie gereicht werden, und von der Reihenfolge innerhalb der Serien.

Die Reihenfolge der Proben ist insofern von Bedeutung, dass sie für zwei der wichtigsten psychologischen Einflussfaktoren wichtig ist (AMERINE et al. 1965): Beim häufig beobachteten Kontrastfehler wird eine Probe auf Grund der Qualität der Vorprobe anders bewertet. Ein Wein mittlerer Qualität z.B. wird, wenn er nach einem minderer Güte gereicht wird, tendenziell besser benotet werden als wenn er nach einem sehr guten Wein ausgegeben wird. Der Positionsfehler hingegen besteht darin, dass die erste gereichte Probe großzügiger bepunktet wird als die folgenden.

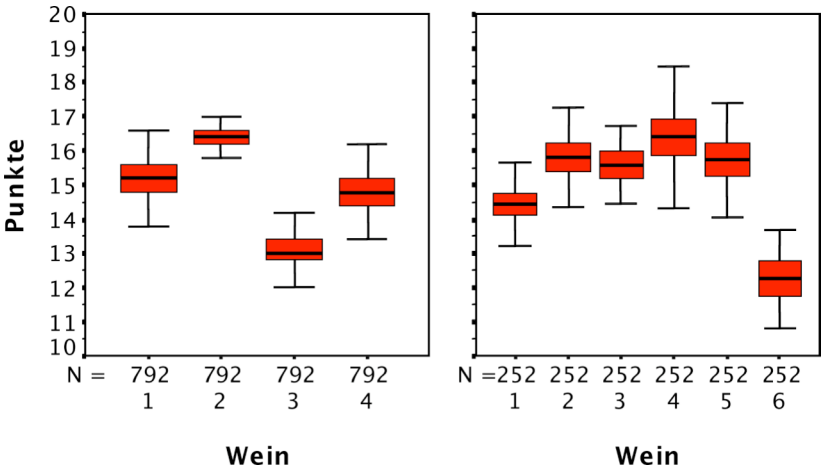

Abb. 6: Aus der Beurteilung von vier Gewürztraminer- (links) und sechs Blauburgunder-Weinen (rechts) errechnete Kommissionsmittelwerte.

Dass die Heterogenität bezüglich der Qualitätsansprüche der Kommissionsmitglieder entscheidend für den Ausgang einer Prämierung sein können, erläutern die Beispiele, welche in der Abbildung 6 dargestellt sind. Im ersten beurteilten zwölf berücksichtigte Koster mit Hilfe des 20-Punkte-Schemas in zweifacher Wiederholung vier Weine der Sorte Gewürztraminer des gleichen Weinbaugebietes und Jahrgangs hinsichtlich ihrer Gesamtqualität. Aus den zwölf Kostern wurden danach die maximale Anzahl an verschiedenen Kommissionen zu fünf Mitgliedern - diese ist eine häufige Anzahl bei

Weinprämierungen - gebildet. Auf diese Art konnten für jeden Wein 792 Kommissionsmittelwerte errechnet werden. Analog dazu wurde im Beispiel des rechts angeführten Box-Plot-Diagramms verfahren. Dort sind die Ergebnisse angeführt, die zehn berücksichtigte Koster bei der Beurteilung von sechs Blauburgunder-Weinen gebildet haben.

Die Bewertungen der Gewürztraminer-Proben schwankten, je nachdem von welcher Kommission sie beurteilt wurden, im Schnitt um 2,4 Punkte. Auch wenn die mittleren Quartile, welche definitionsgemäß 50 % der Beurteilungen abdecken, einen Bereich von nur 0,7 Punkte abdecken, wird der Kommissionseinfluss deutlich. Abweichungen von 1,2 bis 3,3 Punkte sind in der international üblichen Bepunktungspraxis gravierend. Zudem fällt auf, dass hinsichtlich einer Prämierung mit Ausnahme des Weines 2 alle Box-Plots sich überschneiden, d.h. dass alle Produkte untereinander austauschbar sind.

Noch offensichtlicher wird dieser Sachverhalt im Blauburgunder-Beispiel. Die Kommissionsmittelwerte schwankten im Schnitt um bis zu 3,0 Punkte, die mittleren Quartile durchschnittlich um bis zu 0,9 Punkte. Auch hier kann bezüglich einer Rangordnung fast jeder Wein durch jeden anderen ersetzt werden. Einzig der Wein 6 entzieht sich im negativen Sinn dem Vergleich. Das weiteste ermittelte Bewertungsspektrum von 14,3 bis 18,5 Punkte im Fall des Blauburgunders 4 ist vom kommerziellen Gesichtspunkt aus gesehen eklatant.

Es muss klargestellt werden, dass alle Koster aus einem engen geografischen und beruflichen Umfeld stammen und profunde Kenner der gekosteten Sorten sind. Im Fall von international besetzten Jurorengruppen und Kostwettbewerben, wo zudem eine Kategorisierung nach Sorten nicht durchgeführt wird, werden diese Schwankungen eher größer als kleiner sein.

Das Streichen der Resultate des strengsten und des großzügigsten Kosters kann nicht als akzeptable Lösung angesehen werden, bringt das doch einen unnötigen Informationsverlust mit sich. Die Prüfer mit guten Kostfähigkeiten, welche deshalb auch klar differenzieren und dementsprechend auspunkten, sind obendrein erfahrungsgemäß von dieser Regelung am häufigsten betroffen.

Viele der obgenannten Unzulänglichkeiten sind aus organisatorischen Gründen nicht behebbar. Dass alle Koster die Gesamtzahl der eingereichten Weine beurteilen, wird weiterhin Wunschdenken bleiben. Die Mehrzahl der Fehler sind aber im Sprachgebrauch der analytischen Statistik systematische Abweichungen bzw. systematische Fehler. Die Methode, sie zu vermeiden, oder zumindest zu minimieren, beruht darauf, das Zufallsprinzip so oft wie möglich anzuwenden, d.h. alle bekannten Faktoren möglichst vollständig zu randomisieren: Jeder Wein sollte in jeder erdenklichen Kombination aus Koster, Serie und Reihenfolge an der Bewertung teilnehmen können. Je größer die Anzahl an möglichen Kombinationen ist, desto geringer ist die Wahrscheinlichkeit, daß ein Wein nur deshalb schlechter bzw. besser beurteilt wird, weil seine Kommission strenger bzw. milder ist, die Serie ein

höheres bzw. ein tieferes Niveau aufweist und der Wein, der vorher gereicht wird, eine bessere bzw. schlechtere Qualität aufweist.

Mit dem Ziel, diesen Idealvorstellungen möglichst nahe zu kommen, wurden bei den zuletzt von der Sektion Kellerwirtschaft organisierten Kostwettbewerben, welche immer nur eine Sorte zum Inhalt hatten, ein eigens entwickeltes System angewandt. Mittels dem Zufallsgenerator werden aus dem Pool der gesamten zu verkostenden Proben Koster für Koster die ihnen zustehende Anzahl von Weinen zufällig zugewiesen. So hat jeder Koster mit einer sehr großen Wahrscheinlichkeit eine Auswahl an Weinen und eine Reihenfolge, wie sie kein anderer Koster hat. Jeder Koster ist also seine eigene Kostkommission, wodurch die Anzahl der ermöglichten Kombinationen maximiert wird. Die Weine werden einzeln gereicht und bewertet.

Um die Zuverlässigkeit der Kosturteile zu prüfen, wird, wie vorher erklärt, eine gewisse Anzahl von Weinen doppelt gereicht. Die mehrfach gereichten Weine sind für alle Teilnehmer gleich und werden zufällig unter die anderen Proben gemischt.

Bei der letzten Veranstaltung wurden 68 Weine einer Beurteilung unterzogen, wobei 19 Juroren zur Verfügung standen. Jeder Wein wurde von zumindest 13 Personen verkostet. 4 Koster konnten nicht berücksichtigt werden, weswegen die Mindestanzahl berücksichtigter Prüfer pro Wein 9 betrug.

Die Mediane der berücksichtigten Koster bilden das Endergebnis der Kost. Wichtig ist, dass die Endergebnisse mit so vielen Stellen ausgegeben werden, wie den Kostern bei der Bepunktung zur Verfügung standen. Arbeitet man z.B. mit dem 20-Punkte-System und lässt halbe Punkte zu, müssen die Ergebnisse auf ganze oder halbe Punktezahlen ab- bzw. aufgerundet werden. Ansonsten wird eine überhöhte Genauigkeit vorgetäuscht und Differenzen aufgezeigt, die im Moment der sensorischen Beurteilung gar nicht vermerkt werden konnten.

Abhängig vom Ziel der Kost kann eine Rangordnung z.B. der besten zehn gebildet werden. Es bietet sich auch an, jene Weine hervorzuheben, welche sich im positiven oder negativen Sinn abheben. Zu diesem Zweck kann wieder das Vertrauens- oder Konfidenzintervall angewandt werden. Abhängig vom ermittelten Gesamtmittelwert, der Streuung der Einzelwerte rund um diesen und der festgesetzten Irrtumswahrscheinlichkeit errechnen Statistikprogramme jene Grenzen, außerhalb denen ein Wein im vorteilhaften oder abträglichen Sinn als nicht mehr zur Gruppe gehörig erklärt werden kann.

Schlussfolgerungen

Die sensorische Bewertung von Weinen im Sinne der Weinforschung und zum Zweck von Prämierungen sind erheblichen Unzulänglichkeiten ausgesetzt. Sofern sie als solche erkannt werden, ist es möglich, die Qualität der Verkostungen deutlich zu steigern. Voraussetzung dafür ist auch, dass eingefahrene Organisations- und Ablaufformen besonders bei Weinkostwettbewerben hinterfragt und eventuell geändert werden. Eine Überprüfung der Tagesform der Koster sowie die Verringerung systematischer Fehler durch eine konsequente Anwendung der Randomisierung sind erste Schritte, um eine nachvollziehbarere Beurteilung der Weine als vielfach üblich zu gewährleisten.

Literatur

Amerine M.A., Pangborn R.M., Roessler E.B. (1965). Principles of Sensory Evaluation of Food. Academic Press, New York, San Francisco and London.

Bengtsson K. (1943). Provsmakning som analysmetod. Statistik behandling av resultaten. Svenska Bryggareforen Månadsblad, (58): 59 71 102 111 149 157 in: Amerine M.A., Pangborn R.M., Roessler E.B. (1965). Principles of Sensory Evaluation of Food. Academic Press, New York, San Francisco and London.

Bertuccioli M., Rosi I. (1992). Esperienze di valutazione sensoriale dei vini a fini tecnologici. Vini d'Italia, (34, 2): 43 48

Castino M. (1983). La valutazione organolettica dei vini con una scala non strutturata. Vignevini, (10, 10): 53 61.

Castellari L., Ubigli M., Cravero M.C., Bosso A., Guaita M. (2001). Il Gutturnio: i risultati delle analisi sensoriali - 2. parte. Vignevini, (28, 10): 43 49

Cliff M., Yuksel D., Girard B., King M. (2002). Characterization of Canadian Ice Wines by Sensory and Compositional Analyses. Am. J. Enol. Vitic., (53): 46 53

Daepp H.U. (1966a). Die Grundlagen der Sinnenprüfung. Schw. Zeitschrift Obstbau Weinbau, (102): 611 618

Daepp H.U. (1966b). Die Methoden der Sinnenprüfung. Schw. Zeitschrift Obstbau Weinbau, (102): 695 706

Delteil D. (2000). Positionnement d'un vin par test consommateur et analyse sensorielle descriptive quantifiée - L'exemple de la Cartographie des Préférences. Revue française d´Œnologie, (40, 182): 31 35

Frost M.B., Noble A.C. (2002). Preliminary Study of the Effect of Knowledge and Sensory Expertis on Liking for Red Wines. Am. J. Enol. Vitic., (53): 275 284

Girardot N.F., Peryam D.R., Shapiro R. (1952). Selection of Sensory Testing Panels. Food Technology, (6): 140 143

Iacono F., Campostrini F., De Micheli L., Falcetti M. (1992). Esperienze di analisi sensoriale dei vini quale strumento di valutazione dei risultati di ricerche viticole. Vini d'Italia, (34, 2): 59 68

Köhler H., Curschmann K., Günther P. (1983). Zur Prüfung von Einzelurteilen auf ihren Wert für die Verrechnung. Die Weinwirtschaft Markt, (119, 12): 393 396

Krum J.K. (1955). Sensory Panel Testing. Food Engineering, (27): 74 83

Linskens H.F., Jackson J.F. (1988), Wine Analysis. Springer-Verlag, Berlin, Heidelberg, New York, London, Paris, Tokyo.

Lorenz R.J. (1988). Grundbegriffe der Biometrie. Gustav Fischer Verlag, Stuttgart.

Neumann R., Molnár P., Arnold S. (1983). Sensorische Lebensmitteluntersuchung. VEB Fachbuchverlag, Leipzig.

Noble A.C. (1988). In Linskens H.F., Jackson J.F. Wine Analysis. Springer-Verlag, Berlin, Heidelberg, New York, London, Paris, Tokyo.

Overman A., Li J.C.R. (1948). Dependability of Food Judges as Indicated by an Analysis of Scores of a Food-Tasting Panel. Food Research, (13): 441 449

Patz C.D., Giehl A., Dietrich H. (2000). Eine revolutionäre Methode für die Qualitätskontrolle. Der Deutsche Weinbau, (16-17): 30 33

Paul F. (1967). "Die Rangziffern-Methode", eine einfache Möglichkeit für den organoleptischen Vergleich zweier oder mehrerer Proben. Mitt. Rebe und Wein, (17): 280 288

Peryam D.R., Swartz V.W. (1950). Measurement of sensory differences. Food Techn., (4): 390 395 in: Amerine M.A., Pangborn R.M., Roessler E.B. (1965). Principles of Sensory Evaluation of Food. Academic Press, New York, San Francisco and London.

Ribéreau-Gayon J., Peynaud E., Ribéreau-Gayon P., Sudraud P. (1975). Sciences et techniques du vin - Tome 2. Dunod, Paris.

Schrodt W., Jakob L. (1966). Die statistisch erfaßbaren Wechselwirkungen bei der technischen Weinprobe. Mitt. Rebe und Wein, (16): 357 369

Schneider V., Kreckel R. (1995). Beschreiben statt bewerten. Das deutsche Weinmagazin, (2, 9): 16 24

Stone H., Sidel J., Oliver S., Woolsey A., Singleton R. (1974). Sensory evalation by quantitative descriptive analysis. Food technology, (28, 11): 24 34 in: Ubigli M. (1992b). La valutazione qualitativa dei vini mediante scheda astrutturata. Vini d'Italia, (34, 2): 29 42

Ubigli M. (1986). Analisi sensoriale - Studio del comportamento di un gruppo di assaggiatori. Vini d'Italia, (28, 2): 11 26

Ubigli M. (1988). L'attendibilità dei test organolettici effettuati dagli esperti. Vini d'Italia, (30, 6): 21 30

Ubigli M. (1990a). Test discriminatori nell'assaggio delle bevande. Vignevini, (17, 4): 23 29

Ubigli M. (1990b). Analisi sensoriale: un esempio di elaborazione dati. Vignevini, (17, 5): 29 36

Ubigli M. (1992a). Un approccio sensoriale per la definizione dei caratteri di tipicità di un vino a DOC. Vini d'Italia, (34, 1): 49 64

Ubigli M. (1992b). La valutazione qualitativa dei vini mediante scheda astrutturata. Vini d'Italia, (34, 2): 29 42

Weiss J. (1981a). Rating scales in the sensory analysis of foodstuffs. II. Paradigmatic application of the rating method with unstructured scale. Acta Alimentaria, (10, 4): 395 405

Weiss J. (1981b). Selection of Sensory Judges. Journal of Food Quality, (4): 55 63

Weiss J., Willisch E., Knorr D., Schaller A. (1972). Ergebnisse von Untersuchungen bezüglich der differenzierten Wirkung einer sensorischen bewertenden Prüfmethode gegenüber einer sensorischen Rangordnungs-Prüfmethode am Beispiel von Apfelsaft und Birnennektar. Confructa, (17, 4/5): 237 250

Wiley R.C., Briant A.M., Fagerson I.S., Sabry J.H., Murphy E.F. (1957). Evaluation of Flavor Changes Due to Pesticides - A Regional Approach. Food Research, (22): 192 205

Tabellen- und Abbildungsverzeichnis