

# Analyse von technischen und wirtschaftlichen Daten

## 1. Erste Vorlesung

In dieser Vorlesung werden die Begriffe arithmetisches Mittel, Median, und korrigierte empirische Standardabweichung erläutert. Der Boxplot und die daraus resultierenden Informationen, sowie die empirische Dichtefunktion werden vorgestellt.

**Beispiel 1:** Es sollen die Studenten, die eine mindestens ausreichende Note in Mathematik A2 erhielten nach ihren Noten klassifiziert:

	Note	Häufigkeit ( $v_i$ )	Relative Häufigkeit ( $v_i/n$ )
$\eta_5$	5	20	$20/500 = 0,04$
$\eta_4$	4	33	$33/500 = 0,066$
$\eta_3$	3	150	$150/500 = 0,3$
$\eta_2$	2	297	$297/500 = 0,594$

In der ersten Spalte der Tabelle sind die Noten, d.h. die möglichen Merkmalausprägungen ( $\eta_i = 5, 4, 3, 2$ ) enthalten. Aus der Summe der dritten Spalte sieht man, dass die Prüfung insgesamt  $n = 500$  Studenten erfolgreich abgelegt haben, diese Zahl ist die **Größe der Stichprobe**, sie wird meistens mit  $n$  oder  $N$  bezeichnet. Die Note des  $i$ -ten Studenten wird mit  $\xi_i^*$  ( $i = 1, \dots, 500$ ) bezeichnet. Das arithmetische Mittel der Noten wird mit der Formel

$$\bar{\xi} = \frac{1}{500} \sum_{i=1}^{500} \xi_i^* \quad (1)$$

berechnet. Es kann, wie folgt umgeschrieben werden:

$$\bar{\xi} = \frac{1}{500} \sum_{j=2}^5 v_j \xi_j = \sum_{j=2}^5 \frac{v_j}{500} \xi_j \quad (2)$$

wo  $v_j$  die **Häufigkeit** der Note  $j$  bezeichnet,  $\xi_j$  ist die  $j$ -te Note. Der Bruch  $v_j/n$  an der rechten Seite der Gleichung heißt **relative Häufigkeit** der Note  $j$ . Die relative Häufigkeit ist also das Verhältnis der interessierenden Ausprägungen zu allen Ausprägungen. Es soll erwähnt werden, daß die Summe der relativen Häufigkeiten immer 1 gibt. Die rechte Seite von Gl. (2) heißt **gewichteter Mittelwert** der Probe.

Die Beobachtungen in diesem Beispiel haben **diskrete** Ausprägungen (diskretes Merkmal), da die Noten nur endliche Zahlenwerte annehmen können (in Ungarn 5, 4, 3, 2, 1 in Deutschland 1; 1,3; 1,7; 2; 2,3; 2,7;...;4,7; 5). **Stetige** Werte würden z.B. die Körpergröße der Studenten (stetiges Merkmal) annehmen.

### 1.1. Boxplot

Das arithmetische Mittel sagt über die Probe nicht viel aus. Um ausreichende Informationen zu gewinnen werden weitere Parameter benötigt. Der erste solche Begriff ist der Median. Der Median ist jenes Element, das in der der Größe nach geordneten stetigen Probe genau in der Mitte liegt, es sind gleichviel Elemente links und rechts von ihm. Die der Größe nach geordnete Probe soll mit der Zahlenfolge  $\xi_1^*, \xi_2^*, \dots, \xi_n^*$  bezeichnet werden. Ist

$n$  eine ungerade Zahl, so ist der Index des mittleren Elementen  $m = \frac{n+1}{2}$ , so ist der Median

$\tilde{\xi} = \xi_{\frac{n+1}{2}}^*$ . Ist die Stichprobengröße eine gerade Zahl, so gibt es kein mittleres Element, nur

zwei Nachbarelemente mit den Indizes  $\frac{n}{2}$  und  $\frac{n+2}{2}$ , in diesem fall ist also der Median

$$\tilde{\xi} = \frac{\xi_{\frac{n}{2}}^* + \xi_{\frac{n+2}{2}}^*}{2}.$$

Es kann auch untersucht werden, was über die Elemente links, bzw. rechts von dem Median gesagt werden kann. Es kann auch der Median dieser Teilmengen, durch den die geordnete Stichprobe „geviertelt“ wird, berechnet werden. Sie werden als **Quartile** (Viertelanteile) bezeichnet. Das **erste Quartil** ist der Median der Teilmenge der Stichprobe links, das **dritte Quartil** ist der Median der Teilmenge rechts vom Median der gesamten Stichprobe. Sie werden jeweils mit  $\tilde{\xi}_{\frac{1}{4}}$ , bzw.  $\tilde{\xi}_{\frac{3}{4}}$  bezeichnet. Das 0-te und das vierte Quartil können auch

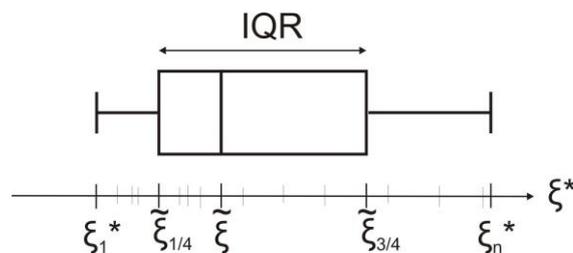
definiert werden, die sind das kleinste und das größte Element ( $\xi_1^*$ , und  $\xi_n^*$ ) in der Stichprobe. Das  $k$ -Quartil ist also jenes Element in der geordneten Stichprobe von welchem  $\frac{k}{4}$  der Elemente kleiner,  $1 - \frac{k}{4}$  größer sind. Das zweite Quartil is natürlich der Median selbst.

Es ist auch üblich **p-Quantile** zu definieren, p-Quantil bezeichnet den Wert der Verteilung für den p% der Beobachtungen kleiner, (1-p)% größer sind.

Es gilt z.B.: Median = 50%-Quantil, oder erstes Quartil = 25%-Quantil.

Der Boxplot der Stichprobe kann mittels der Quartile gezeichnet werden. Er ist im ersten Bild dargestellt. Beispiel für Boxplots:

[http://www.mythologic.hu/digitalcity/servlet/PublishedFileServlet/AAABGTCK/Hircsatorna\\_2008\\_marcus\\_aprilis.pdf](http://www.mythologic.hu/digitalcity/servlet/PublishedFileServlet/AAABGTCK/Hircsatorna_2008_marcus_aprilis.pdf)



**Bild 1** Boxplot

Im Zusammenhang mit dem Boxplot werden weitere Begriffe vorgestellt. Der mittlere Bereich wird als **Inter-Quartil-Region/range** (IQR) bezeichnet. Es ist also

$$IQR = \tilde{\xi}_{\frac{3}{4}} - \tilde{\xi}_{\frac{1}{4}}. \quad (3)$$

**Ausreißer nach unten** heißen die Ausprägungen  $\xi_i$  für die gilt:

$$\xi_i < \tilde{\xi}_{\frac{1}{4}} - 1,5 \cdot IQR. \quad (4)$$

**Extreme Ausreißer nach unten** heißen die Ausprägungen  $\xi_i$  mit:

$$\xi_i < \tilde{\xi}_{\frac{1}{4}} - 3 \cdot \text{IQR} . \quad (5)$$

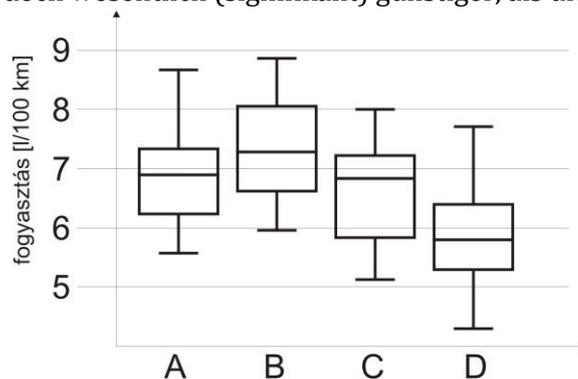
**Ausreißer nach oben** heißen die Ausprägungen  $\xi_i$  für die gilt:

$$\xi_i > \tilde{\xi}_{\frac{3}{4}} + 1,5 \cdot \text{IQR} . \quad (6)$$

**Extreme Ausreißer nach oben** heißen die Ausprägungen  $\xi_i$  mit:

$$\xi_i > \tilde{\xi}_{\frac{3}{4}} + 3 \cdot \text{IQR} . \quad (7)$$

**Beispiel 2:** Man untersuche den Benzinverbrauch von 4 PKW-s im Stadtverkehr. Aus den Meßdaten wurden die Boxplots im Bild 2 konstruiert. So kann der Benzinverbrauch auch visuell verglichen werden. Es ist klar, daß PKW-s A, B, C von einander nicht wesentlich abweichen, Typ D ist jedoch wesentlich (signifikant) günstiger, als die anderen Typen.



**Bild 2** Boxplots vom Verbrauch von vier PKW-s

Eine weitere Auskunft über die Meßdaten gibt ihre **Streuung**. Dazu werden die Differenzen der einzelnen Beobachtungen und ihr Mittelwert  $\xi_i - \bar{\xi}$  benötigt. Es ist klar, daß es immer gilt

$$\frac{1}{n} \sum_{i=1}^n (\xi_i - \bar{\xi}) = 0 \quad (8)$$

So gibt die einfache Differenz keine Auskunft. Entweder der Betrag oder das Quadrat der Differenz soll gemittelt werden:

$$\frac{1}{n} \sum_{i=1}^n |\xi_i - \bar{\xi}| \quad \text{oder} \quad \frac{1}{n} \sum_{i=1}^n (\xi_i - \bar{\xi})^2$$

Die mittlere quadratische Differenz wird als empirische Varianz genannt und mit  $s^2$  bezeichnet.:

$$s^2 = \frac{1}{n} \sum_{i=1}^n (\xi_i - \bar{\xi})^2 . \quad (9)$$

Daraus folgt die **empirische Standardabweichung**

$$s = \sqrt{\frac{1}{n} \sum_{i=1}^n (\xi_i - \bar{\xi})^2} \quad (10)$$

In dieser Formel sind jedoch die Daten von einander nicht unabhängig, da  $\bar{\xi}$  genau das arithmetische Mittel der Meßdaten ist. In der Statistik soll mit unabhängigen Daten operiert werden, so führen wir den Begriff korrigierte empirische Varianz ein:

$$(s^*)^2 = \frac{1}{n-1} \sum_{i=1}^n (\xi_i - \bar{\xi})^2 \quad (11)$$

Dann ist die **korrigierte empirische Standardabweichung**:

$$s^* = \sqrt{\frac{1}{n-1} \sum_{i=1}^n (\xi_i - \bar{\xi})^2} \quad (12)$$

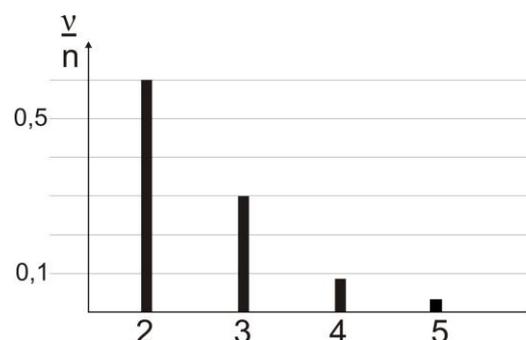
Man fragt sich, was geschieht, wenn  $n \rightarrow \infty$ . Diese Frage wird später beantwortet.

Man soll wahrnehmen, daß die Meßeinheiten des Mittelwertes, des Medianes und der Standardabweichung mit der der einzelnen Meßdaten identisch sind.

## 1.2. Empirische Dichtefunktion

Mit Hilfe der **empirischen Dichtefunktion** oder des **Histogramms** gewinnt man Auskunft über die Verteilung der Beobachtungswerte. Diskrete bzw. stetige Ausprägungen werden bei der Behandlung von einander unterscheidet.

*Im Diskreten Fall* werden auf die Abszisse des Koordinatensystems die Beobachtungswerte, auf die Ordinate die relativen Häufigkeiten aufgetragen. Bei quantitativen Variablen erfolgt eine natürliche, die Größe folgende Skalierung. Bei qualitativen Variablen (wie z.B. Farbe) ist die Reihenfolge uninteressant, wichtig sind nur die Säulenhöhen der einzelnen Häufigkeiten. Bild 3 zeigt das Histogramm, das aus den Daten des 1-ten Beispiels konstruiert wurde.



**Bild 3** Das aus den Daten des Beispiels 1 gezeichnete Histogramm

## 2. Zweite Vorlesung

Die empirische Dichtefunktion wird für den stetigen Fall weiter diskutiert, der Begriff der Verteilungsfunktion wird eingeführt. Die empirische Korrelation sowie ein Spezialfall, die Rangkorrelation werden besprochen.

Für stetige Variablen (z.B. Körpergröße) werden die Meßdaten der Größe nach geordnet, der kleinste und der größte Wert werden gesucht und die Spannweite (der Bereich zwischen diesen Werten) wird in Klassen (Intervalle) aufgeteilt. Es werden über die einzelnen Intervallen Rechtecke gezeichnet, deren Flächen mit der relativen Häufigkeit der Ausprägungen die in diesen Intervallen liegen gleich sind. Die präzise Formulierung ist folgendes. Die Stichprobe sei  $\xi_1, \xi_2, \dots, \xi_j, \dots, \xi_n$ , der kleinste Wert sei  $x_0 = \min \xi_j$ , der größte Wert sei  $x_{ZI} = \max \xi_j$  mit  $ZI =$  Zahl der Intervalle. Für das  $i$ -te Intervall werden folgende Bezeichnungen eingeführt:

- $\Delta x_i$ : Breite des  $i$ -ten Intervalls,
- $\nu_i$ : Zahl der Ausprägungen, die im  $i$ -ten Intervall liegen, d.h. ihre Häufigkeit,
- $f_i = f_n(x_i)$  die Höhe der Rechtecke an der Stelle  $x_i$ .

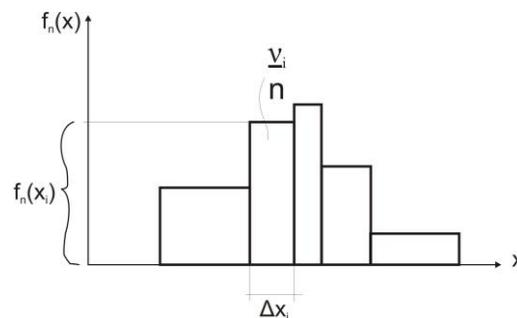
Da die relative Häufigkeit des Ereignisses, daß die Ausprägung im  $i$ -ten Intervall liegt gleich der Fläche des über dieses Intervall gezeichneten Rechtecks ist, gilt:

$$\frac{\nu_i}{n} = \Delta x_i f_n(x_i), \quad (13)$$

und daraus kann die Rechteckhöhe berechnet werden:

$$f_n(x_i) = \frac{\nu_i}{n \Delta x_i}. \quad (14)$$

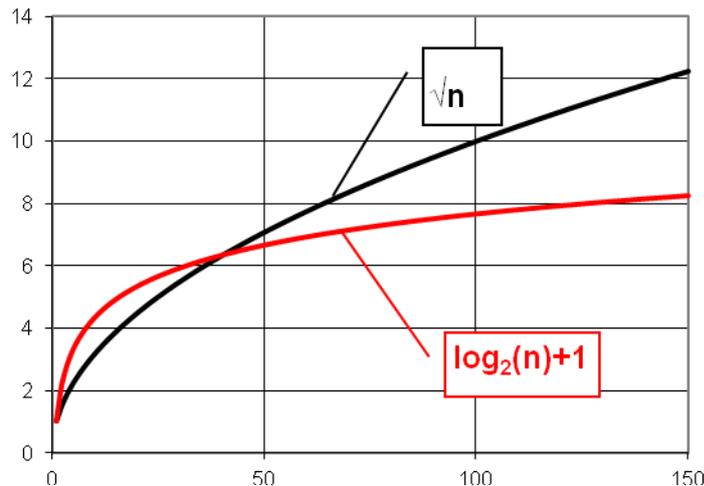
Natürlich gilt  $\sum_i f_n(x_i) \Delta x_i = \frac{1}{n} \sum_i \nu_i = 1$ , d.h. die Gesamtfläche der Dichtefunktion ist 1.



**Bild 4** Empirische Dichtefunktion für stetige Variablen

Zur Festlegung der Zahl der Intervalle werden folgende Regel empfohlen:

- Für  $n < 100$  sei  $ZI \approx \sqrt{n}$ ,
- für  $n > 100$  sei  $ZI \approx \log_2 n + 1$  (Logarithmus  $n$  zur Basis 2 plus 1)



Die Intervallbreiten sollen so bestimmt sein, daß etwa gleiche Zahl von Meßdaten in den einzelnen Intervallen liegen, d.h. die relativen Häufigkeiten etwa gleich seien.

Bei großen Stichproben z.B. bei der Auswertung medizinischer Behandlungen ist es auch üblich gleichbreite Intervalle zu definieren, dabei sind Anzahlen der Ausprägungen, die in die einzelnen Intervallen liegen verschieden groß. Die Säulenhöhen sind dann

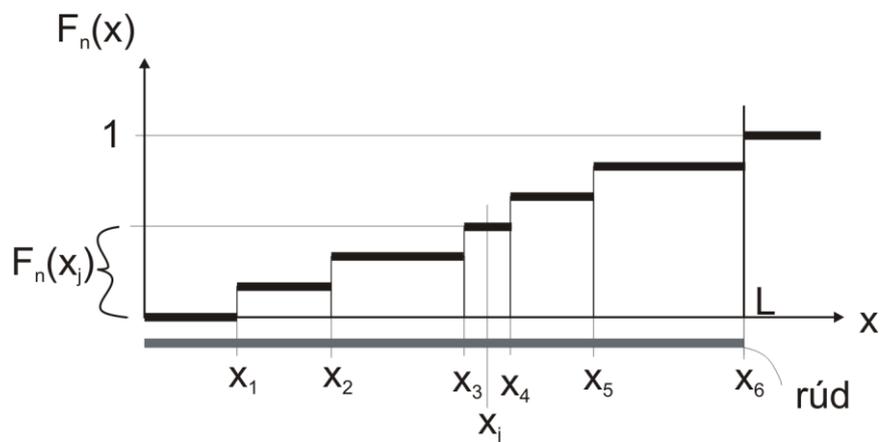
$$f_n(x_i) = \frac{v_i}{n\Delta x}. \text{ Es ist also } \sum_i f_n(x_i)\Delta x = \frac{1}{n} \sum_i v_i = 1.$$

Es gibt symmetrische, aber auch rechtssteile und linkssteile Dichtefunktionen. Ein typisches Beispiel für eine linkssteile Dichtefunktion ist die der Einkommen-Verteilung, viele verdienen wenig Geld und wenige verdienen viel Geld.

## 2.1. Empirische Verteilungsfunktion

Eine alternative Methode zur Darstellung der Beobachtungswerten ist die **empirische Verteilungsfunktion**. Bevor wir sie kennenlernen sollen wir einen Versuch durchdenken. Man betrachte einen homogenen prismatischen Stab. Das eine Ende des Stabes sei im Koordinatenursprung befestigt, der Stab stehe parallel zur x-Achse. Man denke den Stab in horizontaler Richtung gespannt zu sein. Früher oder später reißt der Stab an irgendwelcher Stelle. Der Zugversuch sei vielmal wiederholt. Die Bruchstellen seien auf die x-Achse, die relativen Häufigkeiten der bereits getesteten Stäbe auf die Ordinate aufgetragen. Die Folge der anwachsenden Bruchstellen sei  $x_1, x_2, \dots, x_n$ . So kann das Ergebnis in der im Bild 5 sichtbaren Form als Funktion  $F_n(x_j)$  dargestellt werden.

Die empirische Verteilungsfunktion gibt die relative Häufigkeit der Stäbe an, die links von der Stelle  $x_j$  gebrochen worden sind. Anhand dieses Beispiels kann der Begriff empirische Verteilungsfunktion definiert werden. Die empirische Verteilungsfunktion ist eine Funktion, die – für eine Stichprobe von der Größe  $n$  – über die relative Häufigkeit solcher Ereignisse, die kleiner, als der Abszissenwert sind Auskunft gibt. Die empirische Verteilungsfunktion wird mit  $F_n$  bezeichnet.



**Bild 5** Die die Bruchstellen eines homogenen prismatischen Stabes bei einem Zugversuch darstellende empirische Verteilungsfunktion

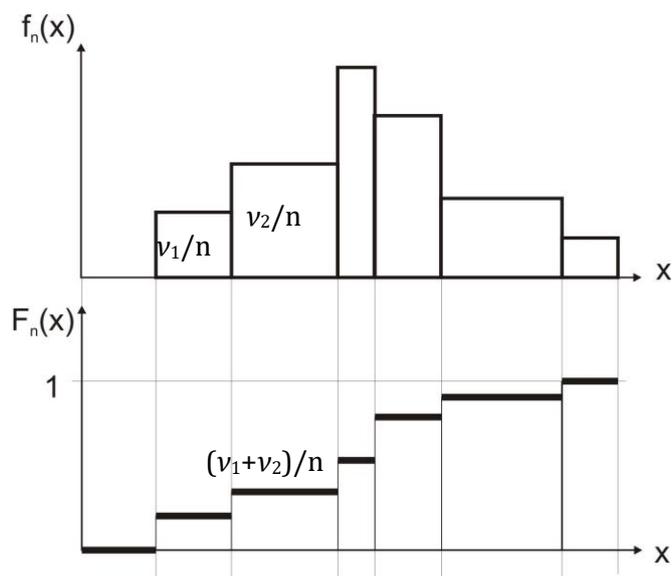
Zwischen der empirischen Dichtefunktion und der empirischen Verteilungsfunktion gilt folgender Zusammenhang. Man zeichne die empirische Dichtefunktion und die empirische Verteilungsfunktion mit gleichen Abszissen für eine Versuchsreihe übereinander (Bild 6). Dann kann folgender Zusammenhang festgestellt werden. Die Höhe der Verteilungsfunktion über dem  $j$ -ten Intervall ist:

$$F_n(x_j) = \frac{\nu_1 + \dots + \nu_j}{n} \quad (15)$$

Wenn man die rechte Seite in einer alternativen Form schreibt, bekommt man:

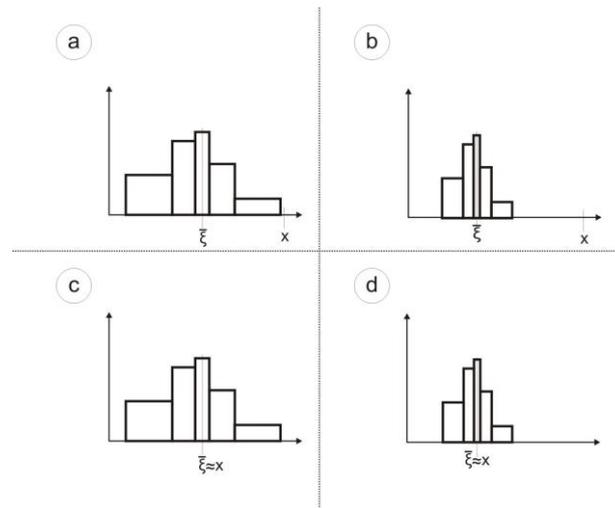
$$F_n(x_j) = f_n(x_1)\Delta x_1 + \dots + f_n(x_j)\Delta x_j = \sum_{i=1}^j f_n(x_i)\Delta x_i \quad (16)$$

Man erkennt, daß die rechte Seite der Gleichung (16) an eine Näherungssumme eines unbestimmten Integrals erinnert. Es ist tatsächlich so, wenn die Intervall-Aufteilung genügend fein ist, dann kann der Wert  $F_n(x_j)$  durch ein Integral definiert werden. Später im Laufe des Semesters wird darauf zurückgekommen.



**Bild 6** Verwandtschaft der empirischen Dichte und Verteilungsfunktion

**Beispiel 3:** Die empirische Dichtefunktion wird in der Praxis auch für Qualitätskontrolle verwendet. Man denke etwa an ein Produkt mit einem Nenn-Maß das als Serienprodukt hergestellt wird. Unser Ziel ist, das Werkzeug und den Fertigungsprozeß genau einzustellen und dadurch Produkte von möglichst hoher Qualität zu erzeugen. In diesem Fall wird das betroffene Maß nach der Fertigung kontrolliert und diese Meßreihe wird dann analysiert. Aus den Meßdaten wird eine empirische Dichtefunktion konstruiert. Es können dabei vier verschiedene Funktionsformen herauskommen, die im Bild 7 dargestellt sind. Der Sollwert des Maßes wird mit  $x$ , der Mittelwert der Meßdaten mit  $\bar{\xi}$  bezeichnet.



**Bild 7** Die aus einer Qualitätskontrolle resultierenden empirischen Dichtefunktionen  
a.) Große Streuung der Fertigung und falsche Maschineneinstellung,  
b.) geringe Streuung der Fertigung jedoch falsche Maschineneinstellung,  
c.) große Streuung aber richtige Maschineneinstellung,  
d.) geringe Streuung und gleichzeitig sorgfältige Maschineneinstellung .

## 2.2. Zusammenhang zwischen zwei Variablen

Es wird in der Physik oder im Ingenieurwesen häufig nach dem Zusammenhang zwischen Variablen gefragt. Man denke z.B. daran, daß der durch ein Ventil fließender Volumenstrom von der Ventilposition abhängt. Natürlich gibt es aber Fälle, wo kein solcher Zusammenhang vorhanden ist. Im folgenden wird die Frage beantwortet ob es ein Zusammenhang zwischen den zwei beobachteten Variablen existiert. Für zwei Variablen werden unsere Beobachtungen mit Wertepaaren erfaßt. Sie seien  $(\xi_1, \eta_1), (\xi_2, \eta_2), \dots, (\xi_n, \eta_n)$ . Diese Wertepaare können in einem kartesischen Koordinatensystem dargestellt werden (siehe Bild 8), wobei auf die Abszisse  $\xi_j - \bar{\xi}$  und auf Ordinate  $\eta_j - \bar{\eta}$  aufgetragen werden. Es bedeutet, daß der Koordinatenursprung bei den Mittelwerten liegt. In diesem Fall können Punkte in allen vier Quadranten der Koordinatenebene liegen. Für alle einzelnen Beobachtungen werden folgende Werte berechnet:

$$c_j = (\xi_j - \bar{\xi})(\eta_j - \bar{\eta}). \quad (17)$$

Das Vorzeichen von  $c_j$  hängt von den Vorzeichen der Multiplikatoren, d.h.:

- $c_j < 0$  im linken oberen und im rechten unteren Quadrant,
- $c_j > 0$  im rechten oberen und im linken unteren Quadrant.

Es bezeichne  $C$  die Summe der  $c_j$  Werte:

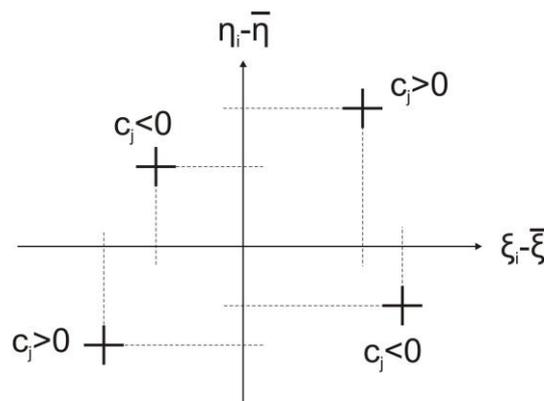
$$C = \sum_{j=1}^n (\xi_j - \bar{\xi})(\eta_j - \bar{\eta}) = \sum_{j=1}^n c_j \quad (18)$$

Das Vorzeichen von  $C$  bestimmt die Lage der Punkte im Koordinatensystem wie es im Bild 9 zu sehen ist. Damit der Wert von  $C$  weder von den einzelnen Punktwerten, noch von ihrer Streuungen abhängt, soll gemittelt und mit den einzelnen Standardabweichungen dividiert werden. Die so erhaltene Zahl wird als **empirischer Korrelationskoeffizient** genannt. Er ist also:

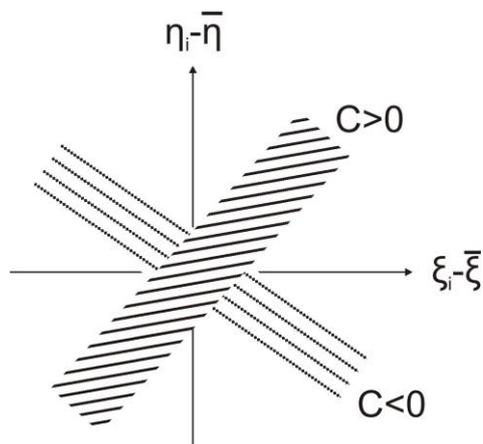
$$\rho(\xi, \eta) = \frac{\frac{1}{n} \sum_{j=1}^n (\xi_j - \bar{\xi})(\eta_j - \bar{\eta})}{s_\xi s_\eta} \quad (19)$$

Für den Korrelationskoeffizient gilt immer  $-1 \leq \rho \leq 1$ . Sein Wert gibt Auskunft darüber ob es ein deterministischer Zusammenhang zwischen den Variablen existiert, leider ist dadurch aber nur der *lineare Zusammenhang* charakterisiert. Und zwar:

- Wenn  $\rho \approx -1$ , dann gibt es zwischen  $\xi$  und  $\eta$  ein entgegengerichteter linearer Zusammenhang und es kann an die Beobachtungswerte eine fallende Gerade angepaßt werden.
- Wenn  $\rho \approx 1$ , dann gibt es zwischen  $\xi$  und  $\eta$  ein gleichgerichteter linearer Zusammenhang und es kann an die Beobachtungswerte eine ansteigende Gerade angepaßt werden.



**Bild 8** Darstellung der Wertepaare und Vorzeichen der  $c_j$  Werte



**Bild 9** Wenn die Punkte entlang einer ansteigenden Gerade liegen, dann ist das Vorzeichen von  $C$  positiv. Wenn aber die Punkte entlang einer fallenden Gerade liegen, dann ist das Vorzeichen von  $C$  negativ

Bei nichtlinearen physikalischen Prozessen ist  $|\rho(\xi, \eta)| \ll 1$ , so z.B. wenn zwischen den Beobachtungswerten ein quadratischer Zusammenhang gibt, wenn also  $\eta = \alpha \cdot \xi^2$ , streuen sich die Punkte um eine Parabel in der oberen Koordinatenebene. Dann ist das Vorzeichen der  $c_j$  Werte in einem Quadrant negativ, im anderen positiv, dadurch ist  $\rho \approx 0$ .

Ein Spezialfall der Korrelation ist die **Rangkorrelation**. Dieser Begriff wird anhand eines Beispiels eingeführt.

**Beispiel 4:** Man denke an einen Weinwettbewerb, wo eine zweiköpfige Jury 4 verschiedene Sorten testet. Die Sorten werden mit den Buchstaben  $A, B, C$  und  $D$ , die Rangzahlen des ersten Begutachters mit  $\xi$ , die des zweiten Begutachters mit  $\eta$  bezeichnet. Die Entscheidung der Jury ist in folgender Tabelle gespeichert. Die Frage ist, ob die Beurteilung der beiden Begutachter im Einklang ist. Die Antwort auf diese Frage gibt der Korrelationskoeffizient zwischen den zwei Ranglisten.

Weinsorten	Punkte des ersten Begutachters ( $\xi$ )	Punkte des ersten Begutachters ( $\eta$ )	$\Sigma$
A	1	1	2
B	2	3	5
C	3	2	5
D	4	4	8
$\Sigma$	10	10	

### 3. Dritte Vorlesung

Wir setzen die Behandlung der Rangkorrelation fort. Zwei Anpassungsmethoden von Kurven an Meßpunkten werden vorgestellt: die Methode der kleinsten Fehlerquadrate und die Methode von Abraham Wald. Auch die Güte der Anpassung, der Determinationskoeffizient wird diskutiert.

Bei der **Rangkorrelation**, auch Spermannsche Korrelation genannt, werden Parameter untersucht, wo kein objektives Bewertungsmaß existiert. Die Beurteilung des Kürlaufs eines Olympiaden Teilnehmers in Eiskunstlauf ist z.B. so ein Parameter.

Zunächst sollen wir an eine Jury mit zwei Begutachtern denken. Nehmen wir an, daß  $NV$  Sportler ( $v_1, v_2, \dots, v_{NV}$ ) am Wettbewerb teilnehmen. Die zwei Jurymitglieder hießen  $Z_1$  und  $Z_2$ . Die vom ersten, bzw. zweiten Jurors gegebene Rangzahlen seien  $r_{i,1}$  und  $r_{i,2}$ . Es bezeichnen die Gesamtrangzahlen der Sportler  $i$  ( $i = 1 \dots NV$ ) die Werte  $c_i = r_{i,1} + r_{i,2}$ . Derjenige Bewerber gewinnt, der die kleinste Gesamtrangzahl erhält. Die Frage ist ob es ein Einklang und in welchem Maße zwischen den Entscheiden der zwei Begutachter gibt. Der Korrelationskoeffizient der zwei Rangzahlenfolgen gibt die Antwort auf die gestellte Frage:

$$\rho(r_1, r_2) = \frac{\frac{1}{n} \sum_{i=1}^{NV} (r_{i,1} - \bar{r}_1)(r_{i,2} - \bar{r}_2)}{s_1 s_2} \quad (20)$$

Es sei die Differenz  $d_i = r_{i,1} - r_{i,2}$  berechnet! Damit lässt sich Gl. (20) in folgende Form schreiben:

$$\rho(r_1, r_2) = 1 - \frac{6 \sum_{i=1}^{NV} d_i^2}{NV(NV^2 - 1)} \quad (21)$$

Wenn die Jury mehrere Mitglieder hat, kann ihre Übereinstimmung mit der **Kendallschen Konkordanz** (Übereinstimmungsmaß) bestimmt werden. Die Mitgliederzahl der Jury sei  $NZ$ . Die Rangzahlen können in eine Tabelle eingetragen werden.

	$Z_1$	$Z_2$	...	$Z_{NZ}$	$\Sigma$
$v_1$	$r_{1,1}$	$r_{1,2}$	...	$r_{1,NZ}$	$c_1$
$v_2$	$r_{2,1}$	$r_{2,2}$		$r_{2,NZ}$	$c_2$
...				...	...
$v_{NV}$	$r_{NV,1}$	$r_{NV,2}$		$r_{NV,NZ}$	$c_{NV}$

In der letzten Spalte stehen die Gesamtrangzahlen der einzelnen Wettbewerber  $i$   $c_i = \sum_{j=1}^{NZ} r_{i,j}$ , d.h. die Summe der Zahlenwerte in der  $i$ -ten Reihe. Es seien diese Summen der Größe nach geordnet:  $c_1^* \leq c_2^* \leq \dots \leq c_{NV}^*$ . Es kann verstanden werden, daß für eine totale Übereinstimmung der Jury, wenn bei allen Jurymitgliedern der selbe Wettbewerber an der  $i$ -ten Stelle steht, ist:

$$c_1^* = NZ, \quad c_2^* = 2NZ, \quad \dots \quad \bar{c}_i^* = NZ \frac{1+NV}{2} \quad (22)$$

$$\Delta c_i^* = NZ, \quad (i = 1, \dots, NV) \quad (23)$$

Wenn es aber absolute Meinungsverschiedenheit vorhanden ist so folgt

$$c_i^* = \bar{c} = \frac{1}{NV} \sum_{i=1}^{NV} c_i \quad (24)$$

$$\Delta c_i^* = 0, \quad (i = 1, \dots, NV) \quad (25)$$

Man berechne folgende an die Varianz erinnernde Summe:

$$S^2 = \sum_{i=1}^{NV} (\bar{c} - c_i^*)^2 \quad (26)$$

Es kann verstanden werden, daß  $S^2$  genau dann seinen maximal erreichbaren Wert erhält, wenn es in der Jury eine vollkommene Übereinstimmung vorhanden ist, d.h. für alle Wettbewerber  $\Delta c_i^* = NZ$ , ist ( $i = 1, \dots, NV$ ).  $S^2$  ist genau dann minimal, wenn es in der Jury absolute Meinungsverschiedenheit existiert, weil dann  $\bar{c} - c_i^* = 0$  und daraus  $S^2 = 0$  folgt. Der **Kendallsche Konkordanzkoeffizient** ist das Verhältnis des aktuellen Werts von  $S^2$  zu seinem maximalen Wert. Dieser Koeffizient wird mit  $W$  bezeichnet:

$$W = \frac{S_{akt}^2}{S_{max}^2} \quad (27)$$

Nach längerer Rechnung erhält man für  $S_{max}^2$ , daß :

$$S_{max}^2 = \sum_{i=1}^{NV} \left( \frac{NZ(1+NV)}{2} - i \cdot NZ \right)^2 = \frac{1}{12} NZ^2 NV (NV^2 - 1). \quad (28)$$

### 3.1. Kurvenanpassung

#### 3.1.1. Methode der kleinsten Fehlerquadrate

Es kann gefragt werden, falls der Zusammenhang zwischen den Variablen nichtlinear ist, wie weit ist er vom linearen, bzw. welcher funktionelle Zusammenhang vorhanden ist und wie die Parameter dieser Funktion gefunden werden können. Die Regressionsanalyse gibt die Antwort auf diese Fragen. Die am meisten verbreitete Methode ist die der **kleinsten Fehlerquadrate**, die vom Friedrich Gauß (1795) erstmals beschrieben wurde. Die Trendlinie (die Anpassungsfunktion von Excel) basiert auch auf diese Methode. Die Güte der Anpassung kann mit dem Determinationskoeffizient gemessen werden.

Es seien die beobachteten Punkte:  $(x_j, \eta_j)$ , ( $j = 1, \dots, N$ ). Man sucht eine Funktionskurve, die zwischen den Punkten läuft und die sie am besten annähert. Man nehme zunächst an, daß der Typ der Funktion  $y(x)$  bekannt ist:

$$y(x) = y(x, \alpha_0, \alpha_1, \dots), \quad (29)$$

wo die Parameter  $\alpha_i$  ( $i = 0, 1, \dots$ ) optimal gewählt werden sollen. Es bedeutet, daß man die Differenz der Ordinatenwerte  $y(x_j)$  und  $\eta_j$  in den  $x_j$  Abszissen minimiere. Man braucht also diese Differenzen  $e_j$  für alle  $j$ -Werte:

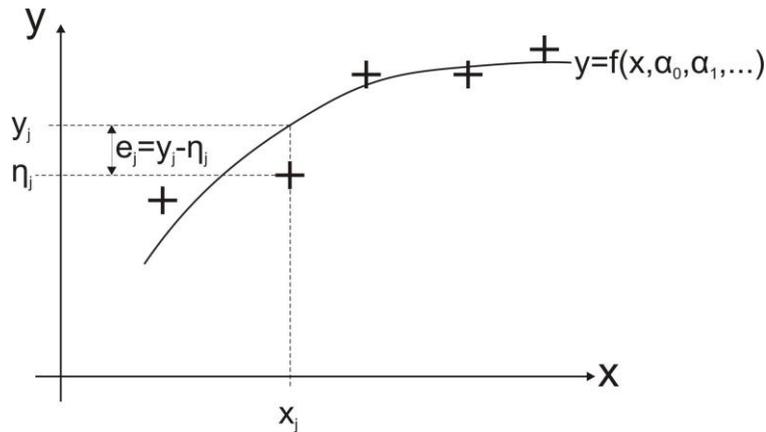
$$e_j = \eta_j - y(x_j, \alpha_0, \alpha_1, \dots), \quad (30)$$

das Vorzeichen der Differenzen ist jedoch uninteressant für uns, so wird quadriert:

$$e_j^2 = (\eta_j - y(x_j, \alpha_0, \alpha_1, \dots))^2. \quad (31)$$

Damit alle Fehler gleichzeitig minimiert werden, müssen die Differenzquadrate summiert werden.

$$D(\alpha_0, \alpha_1, \dots) = \sum_{j=1}^N \left( \eta_j - y(x_j, \alpha_0, \alpha_1, \dots) \right)^2. \quad (32)$$



**Bild 10** Die Differenzen zwischen der mit der Methode des Fehlerquadratminimums angepaßten Kurve  $y$  und dem Meßpunkt

Da  $x_j$  und  $\eta_j$  für alle Indizes  $j = 1, \dots, N$  bekannt sind, hängt die rechte Seite der Gleichung (32) nur von den Parametern  $\alpha_0, \alpha_1, \dots$ . Unsere Aufgabe ist also das Minimum der Differenz  $D(\alpha_0, \alpha_1, \dots)$  als Funktion der Parameter  $\alpha_i$  ( $i = 0, 1, \dots$ ) zu bestimmen. Das Minimum einer Funktion mit mehreren Variablen kann durch partielle Ableitungen bestimmt werden. Mit der Existenz und Eindeutigkeit des Minimums beschäftigen wir uns jetzt nicht. Es wird im Weiteren vorausgesetzt, daß  $y(x)$  eine quadratische Funktion ist:

$$y(x, \alpha_0, \alpha_1, \dots) = \alpha_0 + \alpha_1 x + \alpha_2 x^2. \quad (33)$$

Um das Minimum zu finden muß man folgende Gleichung lösen

$$\frac{\partial}{\partial \alpha_i} \sum_{j=1}^N (\eta_j - \alpha_0 - \alpha_1 x_j - \alpha_2 x_j^2)^2 = 0. \quad (34)$$

Man leitet nach allen  $\alpha_i$  Koeffizienten ab, so bekommt man:

$$\begin{aligned} -2 \sum_{j=1}^N (\eta_j - \alpha_0 - \alpha_1 x_j - \alpha_2 x_j^2) &= 0 \\ -2 \sum_{j=1}^N (\eta_j - \alpha_0 - \alpha_1 x_j - \alpha_2 x_j^2) x_j &= 0 \\ -2 \sum_{j=1}^N (\eta_j - \alpha_0 - \alpha_1 x_j - \alpha_2 x_j^2) x_j^2 &= 0 \end{aligned} \quad (35)$$

Nach Kürzen mit -2, Spalten der Summen und Separation der Glieder bekommt man folgendes lineares algebraisches Gleichungssystem:

$$\begin{aligned} \alpha_0 N + \alpha_1 \sum_{j=1}^N x_j + \alpha_2 \sum_{j=1}^N x_j^2 &= \sum_{j=1}^N \eta_j \\ \alpha_0 \sum_{j=1}^N x_j + \alpha_1 \sum_{j=1}^N x_j^2 + \alpha_2 \sum_{j=1}^N x_j^3 &= \sum_{j=1}^N x_j \eta_j \\ \alpha_0 \sum_{j=1}^N x_j^2 + \alpha_1 \sum_{j=1}^N x_j^3 + \alpha_2 \sum_{j=1}^N x_j^4 &= \sum_{j=1}^N x_j^2 \eta_j \end{aligned} \quad (36)$$

Wir machen darauf aufmerksam, daß im obigen Gleichungssystem die Unbekannten die Parameter  $\alpha_0$ ,  $\alpha_1$  und  $\alpha_2$ , die Koeffizienten aber die Summen, die aus den Messdaten berechnet werden, sind. (Es ist wichtig, dass das Gleichungssystem in den gesuchten Parametern linear sei). Ob dieses System lösbar ist, interessiert uns zurzeit nicht. Wir wissen aus früheren Mathematikstudien, daß obiges System auch in Matrixschreibweise zusammengefaßt sein kann. So ein System ist in Excel leicht lösbar.

### 3.1.2. Determinationskoeffizient / Bestimmtheitsmaß

Nachdem wir das Anpassungspolynom gefunden haben, ist es eine selbstverständliche Frage, wie gut die Anpassung gelungen ist. Die Antwort gibt uns der **Determinationskoeffizient**. Die Meßpunkte seien  $(x_j, \eta_j)$ ,  $j = 1, 2, \dots, N$ , die gewählte Funktion sei  $y(x)$ . Man berechne den Mittelwert der  $\eta_j$  Ordinaten und die auf  $\bar{\eta}$  bezogene Varianz  $S_t^2$ :

$$\bar{\eta} = \frac{1}{N} \sum_{j=1}^N \eta_j \quad (37)$$

$$S_t^2 = \frac{1}{N} \sum (\eta_j - \bar{\eta})^2 = \sigma_\eta^2 \quad (38)$$

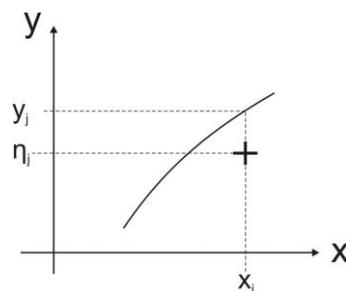
Die Varianz der  $\eta_j$ -Werte kann auch auf die Polynomordinaten bezogen berechnet werden, sie wird mit  $S_m^2$  bezeichnet:

$$S_m^2 = \frac{1}{N} \sum_{j=1}^N (\eta_j - y(x_j))^2 = \frac{1}{N} \sum_{j=1}^N \varepsilon_j^2 = \sigma_\varepsilon^2 \quad (39)$$

$S_m^2$  zeigt also, wie die Substitutionswerte des Polynoms  $y(x_j)$  bei den  $x_j$  Abszissen von den  $\eta_j$  Meßwerten abweichen (Bild11.). Damit ist der Determinationskoeffizient/das Bestimmtheitsmaß:

$$R^2 = \frac{S_t^2 - S_m^2}{S_t^2} \quad (40)$$

Falls  $S_m^2 = 0$ , dann ist  $\eta_j = y(x_j)$ , d.h. die angepaßte Funktion (Polynom) läuft über alle Meßpunkte. Dann ist aber  $R^2 = 1$ . Wenn jedoch  $y = \bar{\eta}$ , dann ist die Näherungsfunktion genau der Mittelwert und  $R^2 = 0$ .



**Bild 11** Bei der Berechnung des Determinationskoeffizienten muß die Standardabweichung der  $\eta_j$  Punkte auf die Anpassungskurve bezogen berechnet werden

$R^2$  gibt also Auskunft darüber, wie Nahe das Polynom an den Meßpunkten läuft. Bei  $R^2 = 1$  liegen die  $N$  Punkte am Polynom, es ist also ein Interpolationspolynom, was keine richtige Wahl ist, da es sogar von der Ordnung  $(N-1)$  sein kann. Die Physik ist höchstwahrscheinlich viel einfacher, als so eine hohe Potenzordnung. So streben wir uns an eher ein Polynom von viel niedrigerer Ordnung zu finden. Der Studierende soll verstehen, daß

$$\rho(y_j, \eta_j) = R^2. \quad (41)$$

### 3.1.3. Methode von Wald

Im Falle eines linearen Zusammenhangs zwischen den Variablen steht die **Methode von Abraham Wald** (Mathematiker in Klausenburg/Kolozsvár, Siebenbürgen) zur Verfügung. Wir werden feststellen, daß diese Methode nur dann anwendbar ist, wenn **beide** Veränderliche durch Fehler behaftet sind. Die Methode basiert auf folgende Idee. Man nehme an, daß die Meßpunkte in zwei Teilpunktgruppen geteilt werden können. Man berechnet die Gewichtspunkte der zwei Teilgruppen und man legt eine Gerade durch diese Gewichtspunkte. Diese Gerade ist unsere Anpassung (Bild 12).

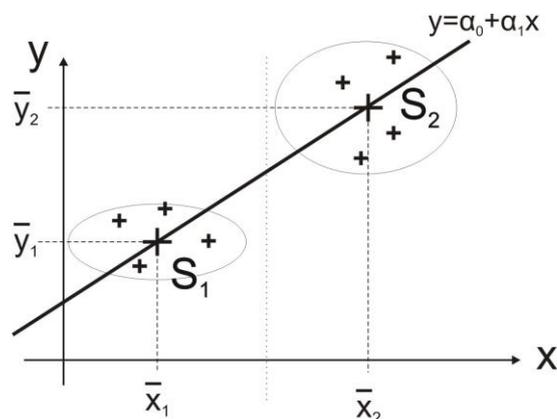
Es seien also die Meßpunkte  $\{\xi_j, \eta_j\}_{j=1}^N$  gegeben, die Gleichung der angepaßten Gerade hat die Form  $y = \alpha_0 x + \alpha_1$ . Die Punkte sollen so aufgeteilt sein, daß die Gewichtspunkte der Teilgruppen möglichst fern von einander liegen. Wenn nach einfacher Betrachtung das Bild der Variablen Ebene mit den dargestellten Punkten keine selbstverständliche Antwort uns gibt, so müssen die Punkte anhand der  $\xi_j$  Koordinaten der Größe nach geordnet werden, man bekommt so die Menge  $\{\xi_j^*, \eta_j^*\}_{j=1}^N$ . Die Trennpunktabzisse ist z.B.  $\xi_r^*$ .

Die Koordinaten der Gewichtspunkte  $S_1$  und  $S_2$  sind durch die Mittelwerte der Koordinaten der Teilpunktgruppen bestimmt:

$$\bar{x}_1 = \frac{1}{r} \sum_{j=1}^r \xi_j^*, \quad \bar{y}_1 = \frac{1}{r} \sum_{j=1}^r \eta_j^*, \quad (42)$$

$$\bar{x}_2 = \frac{1}{N-r} \sum_{j=r+1}^N \xi_j^*, \quad \bar{y}_2 = \frac{1}{N-r} \sum_{j=r+1}^N \eta_j^* \quad (43)$$

Die Gewichtspunkte sind:  $S_1 = (\bar{x}_1, \bar{y}_1), \quad S_2 = (\bar{x}_2, \bar{y}_2) \quad (44)$



**Bild 12** Die mit der Waldschen Methode angepaßte Gerade

Die Richtungstangente der, die zwei Gewichtspunkte verbindenden Gerade ist:

$$\alpha_1 = \frac{\bar{y}_2 - \bar{y}_1}{\bar{x}_2 - \bar{x}_1}, \quad (45)$$

$\alpha_0$  ist dann: 
$$\alpha_0 = \bar{y}_1 - \alpha_1 \bar{x}_1 (= \bar{y}_2 - \alpha_1 \bar{x}_2)$$

## 4. Vierte Vorlesung

Bisher haben wir uns mit der deskriptiven Statistik beschäftigt, wo anhand von endlich vielen, ( $n$ ) Beobachtungen Folgerungen gezogen werden. Es stellt sich die Frage, was geschieht, wenn die Größe der Stichprobe wächst, bzw.  $n \rightarrow \infty$ ? Stabilisiert sich der Mittelwert? Wie ändern sich die früher diskutierten Funktionen?

Es werden in dieser Vorlesung die Grundbegriffe der Wahrscheinlichkeitstheorie erklärt. Wir werden die Begriffe Zufallsvariable und Wahrscheinlichkeit kennenlernen. Die Verteilungsfunktion und die Dichtefunktion werden verallgemeinert und der Erwartungswert wird eingeführt.

### 4.1. Zufallsereignis

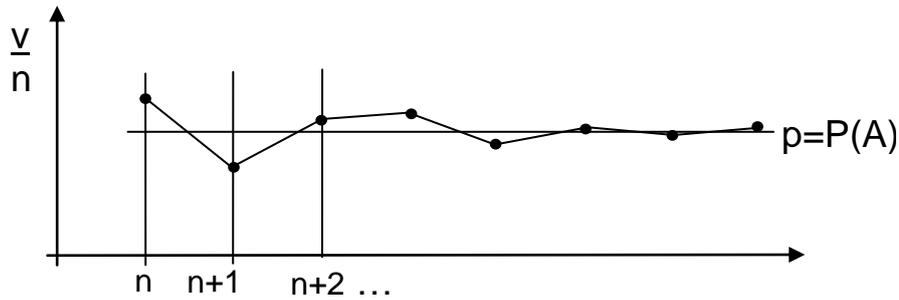
Es gibt in der Praxis zahlreiche Beispiele für Versuche, wo die Ausprägungen von zufälligen Umständen beeinflußt werden. Man nehme z.B. an, daß der Kugeldurchmesser von Kugellagern in einem engen Toleranzbereich gehalten werden muß. Die Erfahrung zeigt, daß es Kugel gibt, deren Durchmesser nicht im vorgeschriebenen Bereich liegen. Dies erklärt sich dadurch, daß die Kugeldurchmesser auch von solchen Umständen beeinflußt werden, worauf wir beim Fertigungsprozeß nicht geachtet haben.

Man kann folgende Definition formulieren. Als **Zufallsereignis** wird jenes Ereignis genannt, dessen Realisation von den berücksichtigten Umständen nicht eindeutig bestimmt ist.

Es ist wichtig zu bemerken, daß in diesem Abschnitt die Anschaulichkeit gegenüber der mathematischen Exaktheit bevorzugt wird. Die exakten Definitionen und die exakten Sätze sind in Fachbüchern der Wahrscheinlichkeitstheorie zu finden.

### 4.2. Relative Häufigkeit, Wahrscheinlichkeit

Man führe einen Versuch durch und man zeichne das Diagramm der relativen Häufigkeit als Funktion der Anzahl der durchgeführten Versuche. Man erfährt, daß sich die relative Häufigkeit schwankt. Wenn man die Zahl der Versuche erhöht, klingt die Schwankung ab und man kann den Mittelwert, der angestrebt wird, eintragen. Dieser Mittelwert wird als die Wahrscheinlichkeit der Realisation des Ereignisses genannt (Bild 13). Selbstverständlich ist diese Definition anschaulich, jedoch ungenau. Die Punktreihe der relativen Häufigkeit strebt eventuell nicht zum Mittelwert, nähert ihn aber immer mehr an.



**Bild 13**

Das untersuchte Ereignis sei  $A$ , dann schreibt sich die Wahrscheinlichkeit der Realisation des Ereignisses  $A$  als:

$$P(A) = p, \quad 0 \leq p < 1. \quad (46)$$

Wenn  $A$  das *sichere Ereignis* ist, dann ist  $P(A) = 1$ , wenn aber  $A$  das *unmögliche Ereignis* ist dann ist  $P(A) = 0$ . Es ist wichtig zu bemerken, daß wenn die Wahrscheinlichkeit eines Ereignisses 0 ist, ist dieses Ereignis nicht unbedingt unmöglich.

Um die Ereignisse mit Zahlenwerten zu bezeichnen, wird der Begriff **Zufallsvariable** eingeführt. Die Realisation solcher Variablen ist nicht eindeutig, d.h. die berücksichtigten Umstände bestimmen den Wert nicht eindeutig. Zufallsvariablen werden häufig mit griechischen Buchstaben bezeichnet. Wir werden demnächst Methoden erläutern, die zur quantitativen Beschreibung von Zufallsvariablen geeignet sind.

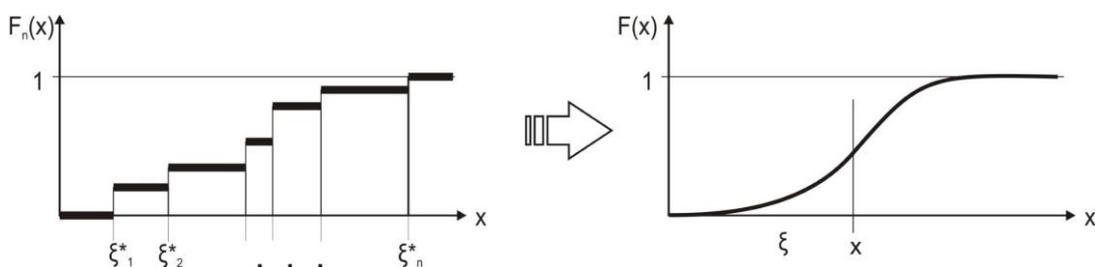
### 4.3. Beschreibung von Zufallsvariablen

Man betrachte die Zufallsvariable  $\xi$ .

Die erste Funktion, womit die Zufallsvariable  $\xi$  beschrieben werden kann ist die **Verteilungsfunktion**. Für  $n \rightarrow \infty$  strebt die empirische Verteilungsfunktion  $F_n(x)$  gegen diese Funktion. Man erinnert sich daran, daß in einer Stichprobe der Größe  $n$  die relative Häufigkeit jener Ausprägungen, die kleiner als  $x$  sind, mit  $F_n(x)$  bezeichnet wurde.

In Analogie dazu behauptet über die Zufallsvariable  $\xi$  der Wert der Verteilungsfunktion  $F$  im Punkt  $x$ , daß die Wahrscheinlichkeit des Ereignisses  $\xi < x$  ist:

$$F(x) = P(\xi < x). \quad (47)$$



**Bild 14** Empirische Verteilungsfunktion  $\leftrightarrow$  Verteilungsfunktion

Für die Verteilungsfunktion  $F$  einer Zufallsvariable  $\xi$  gilt:

- $0 \leq F(x) \leq 1$  für beliebige  $x$  Werte,
- für  $x \rightarrow -\infty$ , strebt  $F(x) \rightarrow 0$ , für  $x \rightarrow \infty$ , strebt  $F(x) \rightarrow 1$ ,
- $F(x)$  ist eine monoton anwachsende Funktion: wenn  $x_1 \leq x_2$  dann  $F(x_1) \leq F(x_2)$
- $P(x_1 \leq \xi \leq x_2) = F(x_2) - F(x_1)$ .

Bemerkung: Es gibt auch unstetige Verteilungsfunktionen, die wir in dieser Vorlesung nicht behandeln.

Die Dichtefunktion läßt sich aus der empirischen Dichtefunktion  $f_n(x)$  für  $n \rightarrow \infty$  ableiten. Wenn die Häufigkeit des Ereignisses, daß eine Ausprägung im Intervall  $j$  lag mit  $v_j$  bezeichnet war, so hieß es  $f_n(x_j) = \frac{v_j}{n\Delta x_j}$ . Man erhöhe die Stichprobengröße  $n$ , dann wächst

auch die Häufigkeit, dass ein Wert in dem Intervall  $j$  liegt und für  $n \rightarrow \infty$  werden die Intervalle immer schmaler, die empirische Dichtefunktion nähert eine stetige Funktion, die **Dichtefunktion**  $f(x)$  an. Im Bild 15 sind die Verteilungsfunktion  $F(x)$  und die Dichtefunktion  $f(x)$  einer Zufallsvariable  $\xi$  dargestellt. Man erkennt, daß die Wahrscheinlichkeit, daß die Zufallsvariable im Intervall  $\Delta x$  liegt, ist  $\Delta F$ , d.h.  $\Delta F = P(\xi \in \Delta x)$ . Die Dichtefunktion wurde so definiert, daß die Wahrscheinlichkeit, daß die Ausprägung im Intervall  $\Delta x$  liegt, gleich dem Gebiet des Ausschnittes unter der Dichtefunktion über diesem Intervall sei, also  $P(\xi \in \Delta x) \approx f(x)\Delta x$ . Damit wird

$$\Delta F \approx f(x)\Delta x. \quad (48)$$

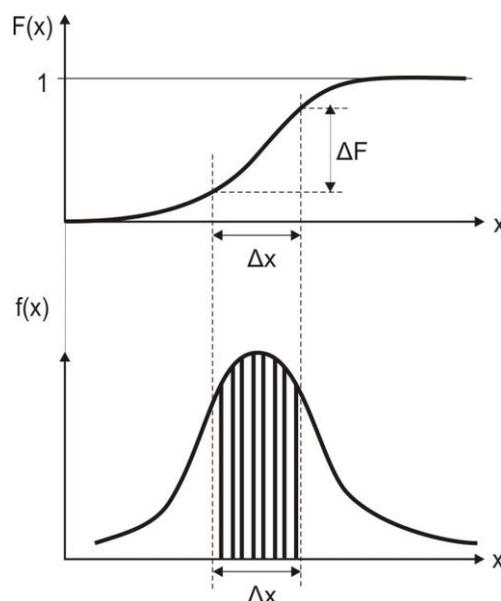
Daraus folgt für die Dichtefunktion:

$$\frac{\Delta F}{\Delta x} \cong f(x). \quad (49)$$

Nach Grenzübergang resultiert für die Koppelung beider Funktionen:

$$f(x) = \frac{dF}{dx}. \quad (50)$$

$F(x)$  ist die Stammfunktion von  $f(x)$ .



### Bild 15 Zusammenhang zwischen Verteilung- und Dichtefunktion

Die Dichtefunktion hat folgende Eigenschaften:

- $f(x) \geq 0$  für beliebige  $x$ -Werte,
- $F(x_2) - F(x_1) = \int_{x_1}^{x_2} f(x) dx$
- für  $x \rightarrow \pm\infty$  gilt  $f(x) \rightarrow 0$ ,
- $\int_{-\infty}^{+\infty} f(x) dx = 1$ .

Der Erwartungswert der Zufallsvariable  $\xi$  ist eine Konstante um die der Wert von  $\xi$  schwankt. Um ihn zu definieren, sei zunächst der Mittelwert der Ausprägungen berechnet:

$$\bar{\xi} = \frac{1}{n} \sum_{i=1}^n \xi_i \quad (51)$$

Der Mittelwert kann anhand des Zusammenhangs zwischen der relativen Häufigkeit und der empirischen Dichtefunktion angenähert werden:

$$\bar{\xi} \cong \frac{1}{n} \sum_{j=1}^J \nu_j x_j = \sum_{j=1}^J \frac{\nu_j}{n \Delta x_j} x_j \Delta x_j \quad (52)$$

hier bezeichnen  $\Delta x_j$  die Breite des Intervalls  $j$  und  $\nu_j$  die Häufigkeit, daß der aktuelle Wert im Intervall  $j$  liegt. Damit ist die Säulenhöhe über dem Intervall  $j$ :

$$f_n(x_j) = \frac{\nu_j}{n \Delta x_j}$$

und damit ist

$$\sum_{j=1}^J \frac{\nu_j}{n \Delta x_j} x_j \Delta x_j = \sum_{j=1}^J f_n(x_j) x_j \Delta x_j \quad (53)$$

Für  $n \rightarrow \infty$  ist die obige Summe eine Näherung des Integrals:

$$\sum_{j=1}^J f_n(x_j) x_j \Delta x_j \rightarrow \int_{-\infty}^{+\infty} x f(x) dx. \quad (54)$$

Dieses uneigentliche Integral ist der Erwartungswert  $M(\xi)$  der Zufallsvariable  $\xi$ :

$$M(\xi) = \int_{-\infty}^{+\infty} x f(x) dx \bigg/ \int_{-\infty}^{+\infty} f(x) dx = \int_{-\infty}^{+\infty} x f(x) dx \bigg/ 1 = \text{Lage des Schwerpunktes} \quad (55)$$

Obwohl der Mittelwert vom Zufall abhängig ist, ist  $M(\xi)$  keine Zufallsvariable, sondern eine, die Zufallsvariable der Stichprobe charakterisierende konstante Größe. Die Eigenschaften des Erwartungswertes sind:

- $M(\xi + a) = M(\xi) + a$ .
- $M(a\xi) = aM(\xi)$ .
- $M(\xi + \eta) = M(\xi) + M(\eta)$ .

## 5. Fünfte Vorlesung

In dieser Vorlesung wird die Varianz einer Zufallsvariable definiert. Wir zeigen, wie dieser Wert standardisiert wird. Der Satz der großen Zahlen wird formuliert. Zwei wichtige Verteilungen, die gleichmäßige Verteilung und die Normalverteilung werden kennengelernt. Der zentrale Satz der Grenzverteilung wird auch erwähnt.

Die **Varianz** der Zufallsvariable  $\xi$  zeigt den Maß der Schwankung von  $\xi$  um ihren Erwartungswert. Die Varianz wird mit  $D^2(\xi) = \sigma_{\xi}^2$  bezeichnet, sie ist

$$D^2(\xi) = M[(M(\xi) - \xi)^2] \quad (56)$$

Die Standardabweichung ist die positive Quadratwurzel der Varianz:  $\sigma = +\sqrt{\sigma^2}$ , mit folgenden Eigenschaften:

- $0 \leq \sigma$ , und ist 0, wenn  $\xi$  eine deterministische Variable ist
- Wenn  $\zeta = a \cdot \xi$  mit  $a = \text{konstant}$ , so folgt  $D^2(\zeta) = a^2 D^2(\xi)$
- Wenn  $\zeta = a + \xi$  mit  $a = \text{konstant}$ , so folgt  $D^2(\zeta) = D^2(\xi)$
- Wenn  $\zeta = a\xi + b$  mit  $a = \text{konstant}$  und  $b = \text{konstant}$ , so folgt  $D^2(\zeta) = a^2 D^2(\xi)$
- Wenn  $\zeta = \xi_1 + \xi_2$ , und  $\xi_1, \xi_2$  unabhängig sind, d.h. die Realisation eines Ereignisses ist unabhängig von dem anderen, sie beeinflussen einander nicht, so ist  $D^2(\zeta) = D^2(\xi_1) + D^2(\xi_2)$

### 5.1. Anwendung

Im Weiteren wird über den Erwartungswert des Mittelwertes der Ausprägungen einer Stichprobe gesprochen. Man betrachte die Ausprägungen  $\xi_1, \xi_2, \dots, \xi_n$ , den Mittelwert bezeichne  $\bar{\xi}$ . Nehmen wir an, daß die Ausprägungen  $\xi_i$  unabhängig und gleichförmig verteilt sind. Wenn die Beobachtungen unabhängig sind und gleiche Verteilung haben heißt die Stichprobe **repräsentativ**.

Da der Mittelwert selbst auch vom Zufall beeinflusst wird, ist für uns sein Erwartungswert interessant. Er kann anhand der Definition des Mittelwerts und der Eigenschaften des Erwartungswerts wie folgt berechnet werden:

$$M(\bar{\xi}) = M\left(\frac{1}{n} \sum_{i=1}^n \xi_i\right) = \frac{1}{n} M\left(\sum_{i=1}^n \xi_i\right) = \frac{1}{n} \sum_{i=1}^n M(\xi_i) \quad (57)$$

Da die einzelnen Beobachtungen (Ausprägungen) gleichförmig verteilt sind, besitzen sie den selben Erwartungswert: Dadurch ist:

$$M(\bar{\xi}) = \frac{1}{n} \sum_{i=1}^n M(\xi_i) = \frac{1}{n} n M(\xi_1) = M(\xi_1) \quad (58)$$

So ist der **Erwartungswert des Mittelwertes** identisch mit dem Erwartungswert der einzelnen Ausprägungen, er schwankt um den gleichen Wert, wie die einzelnen Beobachtungen. Dadurch haben wir erkannt, daß der Mittelwert eine Abschätzung (Näherung) des Erwartungswertes ist.

Es stellt sich die Frage, wie ähnliche Abschätzungen kategorisiert werden können. Wir führen den Begriff, **unverzerrte Abschätzung** ein: wenn der Erwartungswert einer Abschätzung einer Größe mit dieser Größe gleich ist, so spricht man über eine unverzerrte Abschätzung. Dementsprechend ist der Mittelwert eine unverzerrte Abschätzung des Erwartungswertes.

Auch die Varianz des Mittelwertes kann berechnet werden:

$$D^2(\bar{\xi}) = D^2\left(\frac{1}{n} \sum_{i=1}^n \xi_i\right) = \frac{1}{n^2} D^2\left(\sum_{i=1}^n \xi_i\right) \quad (59)$$

Da die Ausprägungen  $\xi_i$  unabhängig sind und die selbe Standardabweichung  $\sigma$  haben, ist:

$$D^2(\bar{\xi}) = \frac{1}{n^2} D^2\left(\sum_{i=1}^n \xi_i\right) = \frac{1}{n^2} \sum_{i=1}^n D^2 \xi_i = \frac{1}{n^2} \sum_{i=1}^n \sigma^2 = \frac{\sigma^2}{n} \quad (60)$$

Daraus ist schon klar, daß es sich lohnt, den Mittelwert zu berechnen, weil er weniger, als die einzelnen Ausprägungen um den Mittelwert schwankt. Es ist auch klar, daß für  $n \rightarrow \infty$  die Varianz des Mittelwertes gegen 0 strebt.

## 5.2. Standardisierung

Es sei  $\xi$  eine Zufallsvariable mit dem Erwartungswert  $M(\xi) = m$  und der Varianz  $D^2(\xi) = \sigma^2$ . So kann sie zur folgenden Zufallsvariable  $\eta$  transformiert werden:

$$\eta = \frac{\xi - m}{\sigma} \quad (61)$$

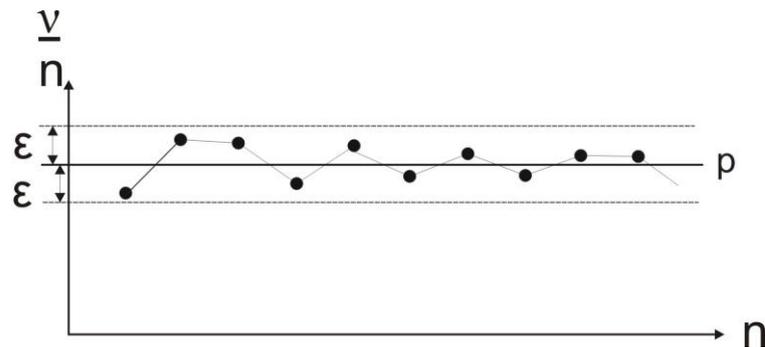
Es ist leicht einzusehen, daß ihr Erwartungswert 0 und ihre Varianz 1 ist:

$$M(\eta) = M\left(\frac{\xi - m}{\sigma}\right) = \frac{1}{\sigma} (M(\xi) - M(m)) = \frac{1}{\sigma} (m - m) = 0 \quad (62)$$

$$D^2(\eta) = D^2\left(\frac{\xi - m}{\sigma}\right) = \frac{1}{\sigma^2} (D^2(\xi) + D^2(m)) = \frac{1}{\sigma^2} (\sigma^2 + 0) = 1 \quad (63)$$

Diese Transformation heißt **Standardisierung**.

### 5.3. Der Satz der großen Zahlen



**Bild 16** Wenn ein Ereignis oftmals wiederholt realisiert wird, so bleibt seine relative Häufigkeit innerhalb der Umgebung mit dem Radius  $\epsilon$  um den Erwartungswert

Der **Satz der großen Zahlen** besagt, daß für  $n \rightarrow \infty$  die relative Häufigkeit des Ereignisses um die Wahrscheinlichkeit schwankt und bleibt in der Umgebung mit dem Radius  $\epsilon$  um die Wahrscheinlichkeit. Formal:

$$\lim_{n \rightarrow \infty} P\left(\left|\frac{v}{n} - p\right| > \epsilon\right) = 0 \quad (64)$$

Obige **stochastische Konvergenz** wird wie folgt bezeichnet:

$$\frac{v}{n} \xrightarrow{s} p, \quad \text{für } n \rightarrow \infty. \quad (65)$$

In diesem Sinn konvergiert der Mittelwert stochastisch zum Erwartungswert:  $\bar{\xi} \xrightarrow{s} M(\xi)$ .

### 5.4. Einige namhafte Verteilungen

#### 5.4.1 Binomialverteilung

Man betrachte einen Produktsatz von  $n$  Elementen, worin ein Element entweder ein Fehlprodukt oder eine gute Ware ist. Die Wahrscheinlichkeit des Ereignisses, daß ein zufällig ausgewähltes Produkt ein Fehlprodukt ist, sei  $p$ . Wie groß ist die Wahrscheinlichkeit, dass im untersuchten Satz genau  $k$  Fehlprodukte gibt?

Um diese Frage zu beantworten definieren wir die Zufallsvariable  $\xi$ , die es uns zeigt, mit welcher Wahrscheinlichkeit wir genau  $k$  Fehlprodukte finden, in diesem Fall ist die Dichtefunktion der binomialverteilten Zufallsvariable  $\xi$  mit folgender Formel beschrieben.

(Es gibt  $\binom{n}{k}$  verschiedene Möglichkeiten um aus  $n$  Elementen  $k$  Elemente auszuwählen,

wenn die Reihenfolge der Auswahl außer Acht gelassen wird. Wenn es unter den  $n$  Elementen des Satzes genau  $k$  Fehlprodukte gibt, dann gibt es  $(n-k)$  gute Elemente. Die Wahrscheinlichkeit einer solchen Ausprägung ist also  $p^k \cdot (1-p)^{(n-k)}$ .)

$$P(\xi = k) = \binom{n}{k} p^k (1-p)^{(n-k)}.$$

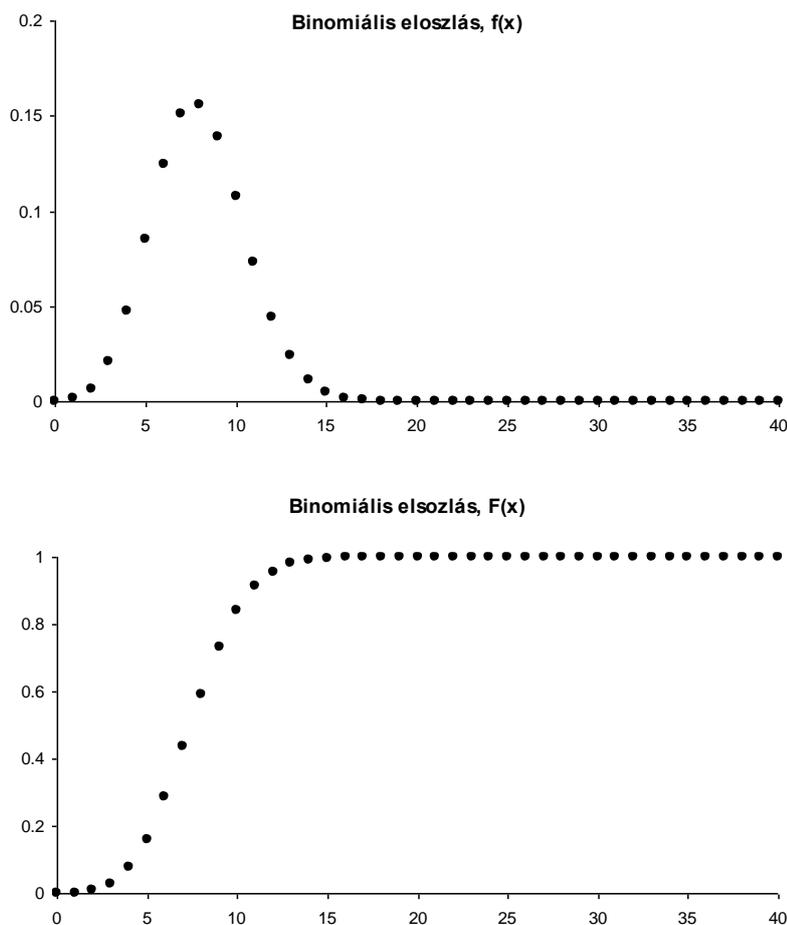
Die Verteilungsfunktion der Verteilung ist

$$P(\xi < k) = \sum_{i=0}^{k-1} P(\xi = i) = \sum_{i=0}^{k-1} \binom{n}{i} p^i (1-p)^{(n-i)}.$$

Erwartungswert und Varianz der Binomialverteilung sind

$$M(\xi) = np \quad \text{und} \quad D^2(\xi) = np(1-p)$$

Die Zufallsvariable  $\xi$  ist eine diskrete Variable, da sie nur Werte von ganzen Zahlen zwischen 0 und  $n$  haben kann, das ist der Grund dafür, dass die Verteilungsfunktion nicht in Integralform definiert ist, sie ist durch zwei Parameter,  $n$  und  $p$  eindeutig bestimmt.



**Bild 17** Wahrscheinlichkeitsdichte und Verteilung der Binomialverteilung mit den Parametern  $n = 40$  und  $p = 0,2$

### 5.4.2 Poissonverteilung

Die Anzahl von Zufallsereignissen, die während einer bestimmten Zeitspanne stattfinden, oder die Anzahl der zufällig verteilten Punkte, die in ein bestimmtes Intervall liegen folgen die Poissonverteilung. Diese Zufallsvariable  $\xi$  ist auch diskret. Die Wahrscheinlichkeit des Ereignisses, dass sie genau den Wert  $k$  besitzt, ist

$$P(\xi = k) = \frac{\lambda^k}{k!} e^{-\lambda},$$

wobei das Parameter  $\lambda$  der Poissonverteilung einen festgelegten Wert hat. Die Wahrscheinlichkeit  $P(\xi = k)$  hängt nur von der Länge des Intervalls ab.

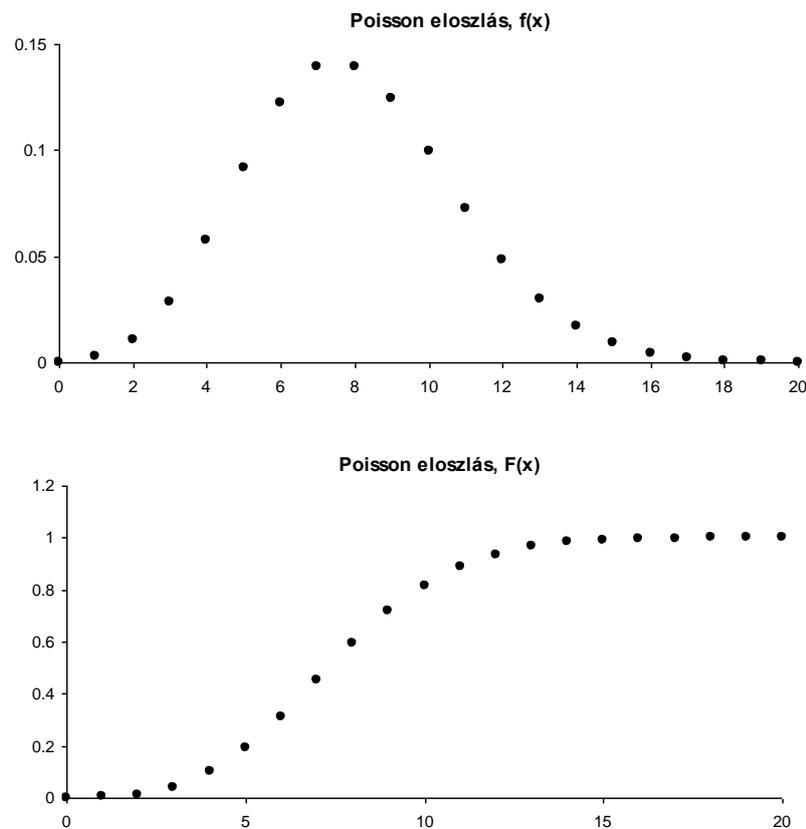
Erwartungswert und Varianz der Poissonverteilung sind

$$M(\xi) = \lambda \quad \text{und} \quad D^2(\xi) = \lambda.$$

Es wird später vom Vorteil sein, dass unter gewissen Umständen die Poissonverteilung eine gute Näherung der Binomialverteilung ist. Wenn nämlich  $n$  groß gegen  $k$  ist, und

gleichzeitig  $p$  klein ist, dann gilt folgende Näherung  $\binom{n}{k} p^k (1-p)^{(n-k)} \approx \frac{(np)^k}{k!} e^{-np}$ ,

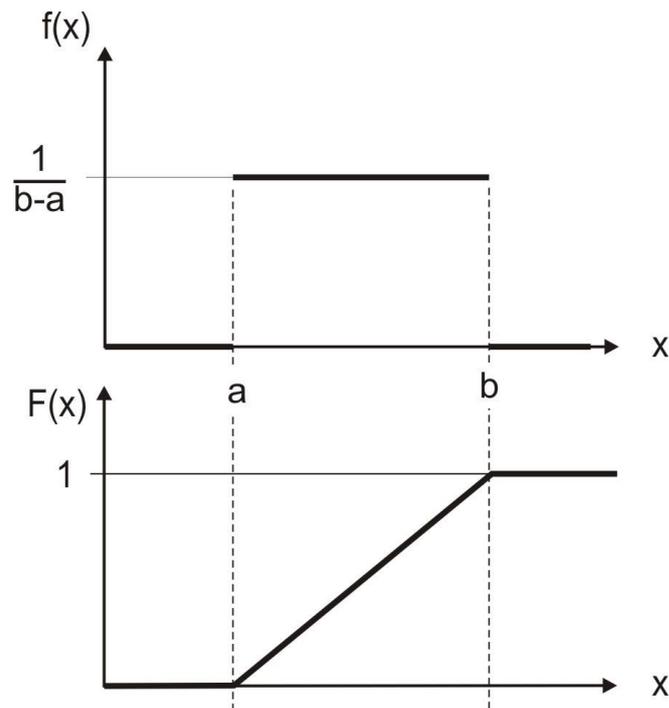
d.h. für  $\lambda = np$ .



**Bild 18** Wahrscheinlichkeitsdichte und Verteilung der Poissonverteilung mit dem Parameter  $\lambda = 8$

### 5.4.3 Gleichmäßig verteilte Zufallsvariable

Zufallsvariablen können durch ihre Dichte-, bzw. Verteilungsfunktion charakterisiert werden. Die einfachste Zufallsvariable hat eine **gleichmäßige Verteilung**. Es kann so aufgefasst werden, dass wenn ein Intervall  $[a,b]$  definiert ist, fällt der Wert der Zufallsvariable  $\xi$  mit gleicher Wahrscheinlichkeit auf beliebige Strecken dieses Intervalls. Die Dichtefunktion und die Verteilungsfunktion sehen im Bild und mit Formeln folgendermaßen aus:



**Bild 19** Die Dichte- und die Verteilungsfunktion einer im Intervall  $[a, b]$  gleichmäßig verteilten Zufallsvariable

$$f(x) = \begin{cases} 0 & \text{ha } x < a \\ \frac{1}{b-a} & \text{ha } a \leq x \leq b; \\ 0 & \text{ha } b < x \end{cases} \quad F(x) = \begin{cases} 0 & \text{ha } x < a \\ \frac{x-a}{b-a} & \text{ha } a \leq x \leq b. \\ 1 & \text{ha } b < x \end{cases}$$

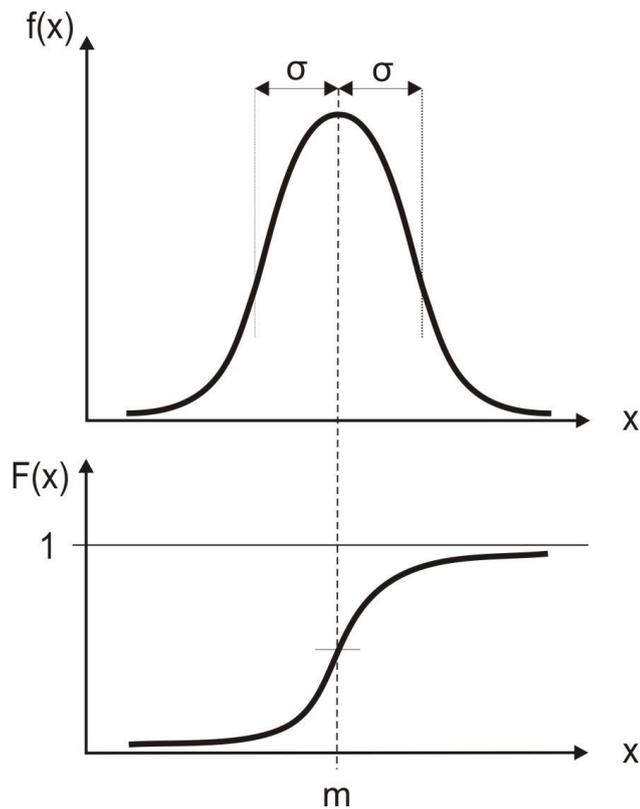
$$m = \frac{a+b}{2}; \quad \sigma^2 = \frac{(b-a)^2}{12}$$

#### 5.4.4 Normalverteilte Zufallsvariable

Eine zweite namhafte Verteilung ist die **Normalverteilung**. Es wird mit ihrer Verteilungsdichte repräsentiert (Bild 20):

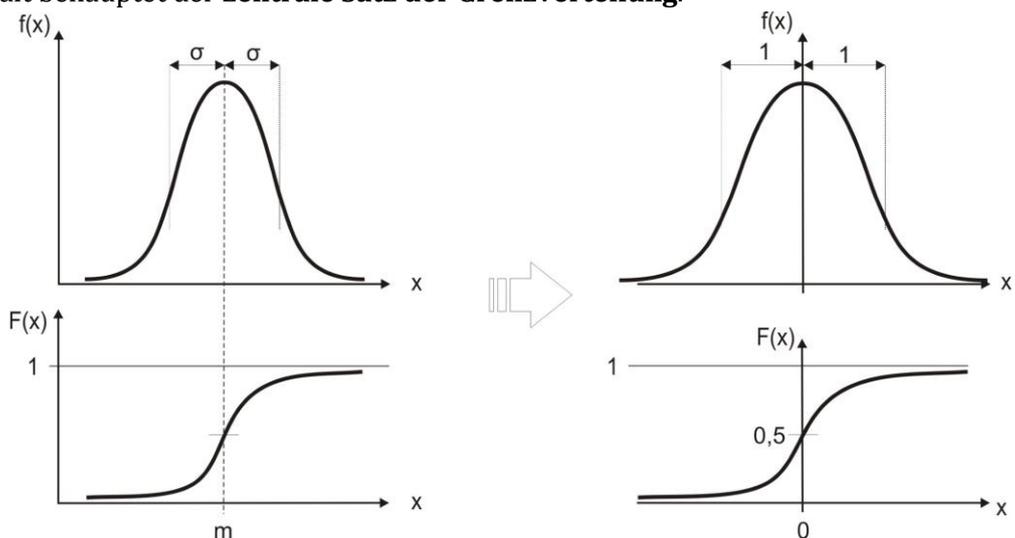
$$f(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(x-m)^2}{2\sigma^2}}, \quad (66)$$

Wo  $m$  der Erwartungswert,  $\sigma$  die Standardabweichung bezeichnen. Eine beliebige normalverteilte Zufallsvariable kann in eine sogenannte **standardnormalverteilte** Zufallsvariable mit dem Erwartungswert 0 und Standardabweichung 1 transformiert werden. Sie ist im Bild 21 dargestellt.



**Bild 20** Die Dichte- und Verteilungsfunktion einer normalverteilten Zufallsvariable mit Erwartungswert  $m$  und Standardabweichung  $\sigma$

Es kann gezeigt werden, dass wenn eine Zufallsvariable dadurch entsteht, dass sehr viele unabhängige gleicherweise verteilte (**iid = independent identically distributed**) Zufallsvariablen addiert werden, dann ist die Summe meistens normalverteilt. Diesen Sachverhalt behauptet der **zentrale Satz der Grenzverteilung**.



**Bild 21** Rechts im Bild sind die Dichte- und Verteilungsfunktion der standardisierten Normalverteilung dargestellt zu sehen

## 6. Sechste Vorlesung

Wir haben die Grundbegriffe der Wahrscheinlichkeitsrechnung kennengelernt. Wo kann der Ingenieurstudent diese Kenntnisse brauchen? Als Ingenieur mißt man oft, aber diese Meßdaten müssen auch ausgewertet werden und danach müssen auf die Meßgenauigkeit begründete Folgerungen gezogen werden. Wir werden jetzt die Auswertungsmethoden und die dazu gehörenden statistischen Verfahren kennenlernen. Die Begriffe systematischer Fehler und Zufallsfehler, sowie direkte und indirekte Messungen werden erläutert.

### 6.1. Meßtätigkeit

Eine physikalische Größe wird mit einem Zahlenwert (Meßzahl) und der dazu gehörenden Meßeinheit eindeutig gegeben. Zum Beispiel ein Wellendurchmesser ist 20mm, die Entfernung Budapest-Nyíregyháza ist 245km. Das einheitliche Meßsystem in Europa ist das SI.

Bei Messungen wird eine physikalische Größe Mittels Experimenten bestimmt.

Die Geschichte der Meßtätigkeit geht auf die Ägypter zurück. In Kairo waren Steinstäbe zur Längenmessung, Steinblöcke zur Gewichtmessung im Gebrauch. Die Zeit wurde mit einer Wasseruhr gemessen, sie bestand aus einer Schale mit einer kleinen Öffnung am Boden wodurch Wasser ausfloß und die Zeitdauer zeigten Rillen an der Innenwand der Schale an.

Heutzutage werden **Meßmodelle** angewendet. Wenn z.B. der Kugeldurchmesser von Kugellagern bestimmt werden muß, nimmt man an, daß die Produkte kugelsymmetrisch sind. Das ist selbstverständlich nicht wahr, auf der Oberfläche der einzelnen Kugeln sind kleine Kerben, andere Formfehler. Wir nehmen also beim Meßverfahren einen idealisierten Zustand an. Für die Durchführung von Messungen werden dann Meßinstrumente gebraucht, ihre Hauptteile sind der **Sensor**, der **Umformer** und die **Anzeige**.

In der Tat sind alle drei Komponente mit Fehlern behaftet, wir sagen, daß die Meßwerte mit einem Meßgeräusch belastet sind.

Die Meßeinheiten werden durch Etalons definiert. Solche Etalons sind der Meter, das Lichtjahr, usw. Es gibt natürlich auch Meßgrößen für die wir nicht über Etalons verfügen, ein Beispiel dafür ist der Wirkungsgrad.

Für die Bestimmung des Zahlenwertes gibt es vier verschiedene Methoden: **Vergleich** mit dem Etalon (Längenmessung mit dem Zollstock), **Ausschlag** eines Zeigers z.B. unter der Wirkung einer Kraft gegen eine Federkraft, die im Instrument eingebaut ist. **Ausgleichen** geschieht z.B. mit der auf der Markt gebrauchten Wage mit zwei Meßschalen, in eine Schale werden Lebensmittel, in die andere Meßgewicht gelegt bis der Zeiger den Wert Null zeigt. Wenn die Differenz zwischen Etalon und physikalische Größe bestimmt wird, spricht man über **Differenzmeßverfahren**.

Der Vergleich mit dem Etalon ist ein **direktes Meßverfahren**. Wenn dies nicht möglich ist, muß ein **indirektes Meßverfahren** angewendet werden. Am Quecksilber Thermometer wird nicht die Temperatur, sondern die dazu proportionale Höhe der Quecksilbersäule an einer Skale abgelesen.

### 6.2. Direkte Meßverfahren

Der Nennwert eines Maschinenelements ist gegeben. Wegen Fertigungsfehler kann dieser Wert nicht verwirklicht werden. Das realisierte Maß kann wegen Meßfehler auch nicht

genau bestimmt werden. Trotzdem sollen wir über die Güte des Produktes Folgerungen ziehen.

Der Nennwert, den wir genau kennen sei mit  $x$ , der Meßwert, die eine Zufallsvariable ist mit  $\xi$  bezeichnet. Wir nehmen an, daß

$$\xi = x + \varepsilon, \quad (67)$$

wo  $\varepsilon$  der Meß- oder Fertigungsfehler ist. Nehmen wir weiter an, daß  $\varepsilon$  eine normalverteilte Zufallsvariable mit dem Erwartungswert  $\Delta x$  und der Standardabweichung  $\sigma$  ist, d.h.:  $\varepsilon \in N(\Delta x, \sigma)$ . Diese Annahme ist mit dem zentralen Satz der Grenzverteilung begründet. Es werden zwei Fälle bezüglich des Erwartungswertes von  $\varepsilon$  unterschiedet. Wenn  $\Delta x \neq 0$ , dann ist  $\varepsilon$  ein **systematischer Fehler**. Wenn aber  $\Delta x = 0$ , dann ist  $\varepsilon$  ein **Zufallsfehler**.

### 6.2.1. Systematischer Fehler

Wenn  $M(\varepsilon) = \Delta x \neq 0$ , dann belastet dieser Fehler alle Meßwerte gleichermaßen. So werden wir über das Vorhandensein solcher Fehler nur dann informiert, wenn wir unsere Messung mit einer anderen Messung – wo ein genaueres Meßinstrument gebraucht wurde – vergleichen können. Wenn unsere Schublehre z.B. schon abgenutzt ist, belastet alle Meßergebnisse der gleiche Fehler. Der systematische Fehler ist bekannt, wenn sowohl sein Wert als auch sein Vorzeichen bekannt ist. Um diesen Fehler zu eliminieren wird kalibriert, danach ist dann schon  $\Delta x = 0$ . Mit der Kalibrierung werden wir uns nicht weiter beschäftigen.

### 6.2.2. Zufallsfehler

Wenn  $\varepsilon$  ein Zufallsfehler ist, kann der Erwartungswert von  $\xi$  berechnet werden. Dazu werden die Eigenschaften des Erwartungswertes gebraucht:

$$M(\xi) = M(x + \varepsilon) = M(x) + M(\varepsilon) = x + 0 = x. \quad (68)$$

So kann man sagen daß für den Erwartungswert die Auswertung der Meßreihe eine gute Abschätzung bietet.

## 6.3. Indirekte Messungen

Bei der Auswertung von indirekten Messungen soll die Fortpflanzung der Fehler auch behandelt werden. Wir kennen den funktionalen Zusammenhang zwischen den Veränderlichen, woraus wir die uns interessierenden Größen berechnen. Das ist z.B. der Fall bei der Wirkungsgradmessung oder bei Bestimmung des Volumenstroms. Es seien die direkt meßbare Variablen  $x_1, x_2, \dots, x_n$ ;  $y$  läßt sich aus diesen Variablen berechnen:

$$y = f(x_1, x_2, \dots, x_n). \quad (69)$$

Der Fehler von  $y$  wird gesucht. Unsere Meßwertreihen seien  $\xi_1, \xi_2, \dots, \xi_n$ . Wir nehmen an, daß sie alle fehlerbehaftet sind. Der Fehler von  $y$  hängt von diesen Meßfehlern ab.

### 6.3.1. Der systematische Fehler von $y$

Die systematischen Fehler der einzelnen Komponenten  $\Delta x_1, \Delta x_2, \dots, \Delta x_n$  seien bekannt. Es ist wichtig zu betonen, daß  $\Delta x_i$  keine Fehlerschranke ist. Der systematische Fehler von  $y$  sei  $\Delta y$ :

$$\Delta y = f(x_1 + \Delta x_1, x_2 + \Delta x_2, \dots, x_n + \Delta x_n) - f(x_1, x_2, \dots, x_n). \quad (70)$$

Falls  $f$  in Taylorsche Reihe entwickelt werden kann und die einzelnen Fehler  $\Delta x_i$  genügend klein sind, so ist der lineare Teil der Taylor-Reihe des ersten Gliedes von Gl. 70:

$$f(x_1 + \Delta x_1, x_2 + \Delta x_2, \dots, x_n + \Delta x_n) \cong f(x_1, x_2, \dots, x_n) + \sum_{i=1}^n \frac{\partial f}{\partial x_i} \Delta x_i \quad (71)$$

Dann läßt sich (70) so schreiben:

$$\Delta y = \sum_{i=1}^n \frac{\partial f}{\partial x_i} \Delta x_i \quad (72)$$

Man sieht, daß der Fehler von  $y$  sogar auch dann 0 sein kann, wenn die Komponenten fehlerbehaftet sind. Der Koeffizient von  $\Delta x_i$  in der Entwicklung ist,  $\frac{\partial f}{\partial x_i}$  und er heißt

**Empfindlichkeitskoeffizient.** Wenn er für einen Index groß ist, dann beeinflusst der entsprechende Eingangsfehler  $y$  mit einem größeren Gewicht. Wenn der Koeffizient kleiner, als 1 ist, so ist der Parameter von einem geringeren Einfluß auf  $y$ . In Spezialfällen können die partiellen Ableitungen einfach berechnet werden.  $y$  habe folgende Form:

$$y = C x_1^{k_1} \dots x_n^{k_n} \quad (73)$$

Dann ist:

$$\frac{\partial y}{\partial x_i} = C x_1^{k_1} \dots k_i x_i^{k_i-1} \dots x_n^{k_n} = k_i \frac{y}{x_i} \quad (74)$$

Damit ist der Fehler von  $y$ :

$$\Delta y = \sum_{i=1}^n k_i \frac{y}{x_i} \Delta x_i \quad (75)$$

Und der relative Fehler ist:

$$\frac{\Delta y}{y} = \sum_{i=1}^n k_i \frac{\Delta x_i}{x_i} \quad (76)$$

### 6.3.2. Zufallsfehler von $y$

Nehmen wir an, daß die Meßdaten  $\xi_i$  mit den Fehlern  $\varepsilon_i$  ( $i = 1, 2, \dots, n$ ) belastet sind:

$$\xi_i = x_i + \varepsilon_i \quad (77)$$

Da  $\varepsilon_i$  ( $i = 1, 2, \dots, n$ ) ein Zufallsfehler ist, so ist  $\varepsilon_i \in N(0, \sigma_i)$ . Die Zufallsvariablen  $\xi_i$  seien weiterhin unabhängig. Der Fehler von  $y$  sei  $\varepsilon_y$ :

$$\varepsilon_y = f(\xi_1, \xi_2, \dots, \xi_n) - f(x_1, x_2, \dots, x_n) \quad (78)$$

Wir werden auch jetzt das erste Glied in (78) um  $(x_1, x_2, \dots, x_n)$  in Taylorsche Reihe entwickeln und die nichtlinearen Glieder gleich vernachlässigen:

$$\varepsilon_y \cong \sum_{i=1}^n \frac{\partial f}{\partial x_i} \Bigg|_{\xi_1, \xi_2, \dots, \xi_n} \varepsilon_i \quad (79)$$

Der Ausdruck  $\frac{\partial f}{\partial x_i} \Bigg|_{\xi_1, \xi_2, \dots, \xi_n} = \overset{\text{bezeichn.}}{\frac{\partial f}{\partial x_i}} \Bigg|_{\xi_i}$  hinter der Summa ist auch eine Zufallsvariable, er wird jedoch als eine Konstante betrachtet. So ist  $\varepsilon_y$  eine lineare

Kombination von den Zufallsfehlern  $\varepsilon_i$ . Da  $\varepsilon_i \in N(0, \sigma_i)$ , so ist auch die lineare Kombination von solchen Größen normalverteilt. Anhand der Eigenschaften des Erwartungswerts kann  $M(\varepsilon_y)$  berechnet werden:

$$M(\varepsilon_y) = M\left(\sum_{i=1}^n \frac{\partial f}{\partial x_i} \Big|_{\xi_i} \varepsilon_i\right) = \sum_{i=1}^n \frac{\partial f}{\partial x_i} \Big|_{\xi_i} M(\varepsilon_i) = 0 \quad (80)$$

Die zweite Gleichheit ist richtig, da wir die Ableitungen als Konstanten betrachten. Die Linearität des Operators Erwartungswert ist vorausgesetzt. Die Varianz von  $\varepsilon_y$  kann jetzt berechnet werden:

$$D^2(\varepsilon_y) = D^2\left(\sum_{i=1}^n \frac{\partial f}{\partial x_i} \Big|_{\xi_i} \varepsilon_i\right) = \sum_{i=1}^n D^2\left(\frac{\partial f}{\partial x_i} \Big|_{\xi_i} \varepsilon_i\right) = \sum_{i=1}^n \left(\frac{\partial f}{\partial x_i} \Big|_{\xi_i}\right)^2 D^2(\varepsilon_i) = \sum_{i=1}^n \left(\frac{\partial f}{\partial x_i} \Big|_{\xi_i}\right)^2 \sigma_i^2 \quad (81)$$

Der zweite Schritt folgt aus der Unabhängigkeit der Zufallsvariablen  $\xi_i$ .  $D^2(\varepsilon_y)$  heißt auch **resultierende Varianz**. Da über die Zufallsvariablen nicht vorausgesetzt wurde, daß ihre Varianz 0 ist, ist auch die resultierende Varianz nicht 0. Um  $D^2(\varepsilon_y)$  zu vermindern, müssen vor allem die mit größerem Einfluß verbundenen Varianzen vermindert werden. Wenn  $y$  ein Produkt von Potenzen ist,  $y = Cx_1^{k_1} \dots x_n^{k_n}$ , dann ist:

$$D^2\left(\frac{\varepsilon_y}{y}\right) = \sum_{i=1}^n k_i^2 \left(\frac{\sigma_i}{x_i}\right)^2 \quad (82)$$

*Beispiel 5:* Das Volumen eines Zylinders ist:

$$V = \frac{d^2 \pi}{4} h \quad (83)$$

$V$  läßt sich als eine Funktion mit zwei Variablen, dem Durchmesser des Grundkreises ( $d$ ) und der Höhe des Zylinders ( $h$ ) aufschreiben. Der Durchmesser sei mit  $\pm 2\%$ , die Höhe mit  $\pm 3\%$  Relativfehler gegeben. Die relative Varianz des Volumens ist damit:

$$D^2\left(\frac{\varepsilon_V}{V}\right) = 2^2 (2\%)^2 + 1^2 (3\%)^2 = 25\% \quad (84)$$

So ist die relative Standardabweichung

$$D\left(\frac{\varepsilon_V}{V}\right) = 5\% \quad (85)$$

Wenn dieser Wert kleiner werden soll, muß der Fehler des Durchmessers vermindert werden, weil er mit dem Quadrat in der Formel erscheint.

#### 6.4. Auswertung von direkten Messungen

Wie früher schon erwähnt, wird der Erwartungswert mit dem mittleren Fehler der Meßreihe abgeschätzt.

Der Meßwert als Zufallsvariable sei  $\xi$ , die einzelnen Meßwerte seien  $\xi_1, \xi_2, \dots, \xi_n$ . Es ist also  $\xi = x + \varepsilon$ , wo  $x$  die gesuchte physikalische Größe ist,  $M(\varepsilon) = 0$  und der Fehler ist normalverteilt. Dann ist  $M(\xi_i) = x$ , ( $i = 1, 2, \dots, n$ ). Weiterhin sei vorausgesetzt, daß auch  $\xi$  normalverteilt ist. Der Mittelwert der Meßwerte  $\xi_i$  ist:

$$\bar{\xi} = \frac{1}{n} \sum_{i=1}^n \xi_i \quad (86)$$

Wie früher gelernt, ist  $M(\bar{\xi}) = x$ , aber im Allgemeinen ist  $\bar{\xi} \neq x$ . Es ist von Interesse, in welcher Umgebung des Mittelwertes der Erwartungswert mit hoher Wahrscheinlichkeit liegt. Nehmen wir ein Intervall mit dem Radius  $a$  an. Wie groß soll  $a$  sein, damit der Erwartungswert im Intervall liegt, d.h.

$$\bar{\xi} - a \leq M(\xi) \leq \bar{\xi} + a \quad (87)$$

Das so definierte Intervall heißt **Konfidenzintervall** und die Zahl  $a$  heißt **Radius des Konfidenzintervalls**. Da auch  $\bar{\xi}$  eine Zufallsvariable ist, ist dieser Radius nicht eindeutig gegeben. Wir wünschen aber, daß die Ungleichheit (87) mit hoher Wahrscheinlichkeit  $p$  gilt:

$$P(\bar{\xi} - a \leq M(\xi) \leq \bar{\xi} + a) = p. \quad (88)$$

$p$  wird **Signifikanzniveau** genannt, diesen Wert geben wir vor. Wir suchen also bei gegebenem Signifikanzniveau den Radius des Konfidenzintervalls.

Zunächst sei vorausgesetzt, daß die Standardabweichung  $\sigma$  der zufallsvariable  $\xi$  bekannt ist. Dann kann folgende Transformation durchgeführt werden:

$$P\left(-\frac{a}{\sigma}\sqrt{n} \leq \frac{\bar{\xi} - M(\xi)}{\sigma}\sqrt{n} \leq \frac{a}{\sigma}\sqrt{n}\right) = p. \quad (89)$$

Es sei

$$\eta = \frac{\bar{\xi} - M(\xi)}{\sigma}\sqrt{n} \quad \text{und} \quad \lambda = \frac{a}{\sigma}\sqrt{n} \quad (90)$$

Man sieht, daß  $\eta$  durch Standardisierung von  $\bar{\xi}$  entstand, damit sind  $M(\eta) = 0$  und  $D^2(\eta) = 1$ .

Da  $\xi$  normalverteilt ist, ist auch  $\bar{\xi}$  von diesem Typ und daraus folgt, daß  $\eta$  standardnormal verteilt ist,  $\eta \in N(0,1)$ . (89) schreibt sich also:

$$P(-\lambda \leq \eta \leq \lambda) = p. \quad (91)$$

Die Verteilungsfunktion der Standardnormalverteilung ist  $\Phi$  (siehe Bild 19, rechts). Die Wahrscheinlichkeit, daß  $\eta$ , im Intervall  $[-\lambda, \lambda]$  liegt ist:

$$P(-\lambda \leq \eta \leq \lambda) = \Phi(\lambda) - \Phi(-\lambda) = p. \quad (92)$$

Da  $\Phi$  zum Wert  $x = 0; y = 0,5$  punktsymmetrisch ist und für  $-\infty$  zu 0, für  $+\infty$  zu 1 geht (siehe Bild 21 rechts unten), ist

$\Phi(\lambda) + \Phi(-\lambda) = 1$ . Dadurch schreibt sich (92) so:

$$2\Phi(\lambda) - 1 = p. \quad (93)$$

Wenn wir  $p$  vorgeben, ist

$$\Phi(\lambda) = \frac{p+1}{2}, \quad (94)$$

und aus einer Tabelle der Verteilungsfunktion  $\Phi(\lambda)$  kann  $\lambda$  gefunden werden. Somit ist der Konfidenzintervall Radius:

$$a = \frac{\lambda\sigma}{\sqrt{n}}. \quad (95)$$

## 7. Siebte Vorlesung

Das Konfidenzintervall wird weiter diskutiert. Der Satz von Gauß - Markow wird formuliert und die Gaußschen Normalgleichungen werden aufgeschrieben.

Die Varianz der Zufallsvariable  $\xi$  ist im Allgemeinen unbekannt. In solchen Fällen wird sie mit der empirischen Varianz ersetzt:

$$s^{*2} = \frac{1}{n-1} \sum_{i=1}^n (\bar{\xi} - \xi_i)^2. \quad (96)$$

Damit wird  $\eta$  eine differente Form haben:

$$\eta_1 = \frac{\bar{\xi} - M(\xi)}{s^*} \sqrt{n}. \quad (97)$$

Jetzt wird auch  $s^*$  vom Zufall beeinflusst, dadurch ist  $\eta_1$  nicht normalverteilt. Daraus ist klar, daß der Radius des Konfidenzintervalls nicht mit Hilfe der  $\Phi$ -Verteilung bestimmt werden kann. Man kann aber beweisen, daß  $\eta_1$  jetzt die sogenannte **Studentsche** Verteilung folgt. Diese Verteilungsfunktion ist mit zwei Parametern, Signifikanz-Niveau  $p$  und Stichprobengröße  $n$  in tabellarischer Form gegeben. Damit wird der Radius des Konfidenzintervalls:

$$a = \frac{\lambda_{st} s^*}{\sqrt{n}}, \quad (98)$$

wo  $\lambda_{st} = \lambda(p, n-1)$  aus der erwähnten Tabelle zu entnehmen ist. Die Zahl  $n-1$  heißt Freiheitsgrad.

Das Ergebnis ist:

$$\bar{\xi} \pm a \quad (p). \quad (99)$$

Damit die obige Methode brauchbar wird, soll gesichert werden, daß die **Meßwerte nur mit Zufallsfehlern belastet seien**.

Mit anderen Wörtern heißt es, daß wenn die Meßdaten nur mit Zufallsfehlern belastet sind, ist die Fehlerschranke der direkten Messung der Konfidenzintervall Radius. Man muß betonen, daß der Fehler nicht eliminiert, nur zwischen Schranken gehalten werden kann.

Im Allgemeinen gibt man bei der Auswertung von Meßdaten das Verhältnis  $a/\bar{\xi}$  an. Wann wir uns mit diesem Wert zufrieden sein können, hängt vom Anwendungsgebiet an. Wenn wir diesen Wert nicht akzeptieren können, muß irgendein Parameter geändert werden.

Der Tabellenwert  $\lambda_{st}$  hängt von zwei Parametern,  $n$  und  $p$  ab. Wenn die Stichprobengröße  $n$  erhöht wird, nimmt der Radius  $a$  ab. Parallel damit nähert sich der  $\lambda_{st}$  Wert zum theoretischen  $\lambda$  Wert der Standard Normalverteilung. Ein erhöhter  $p$  Wert bedeutet, daß aus derselben Stichprobe mit einer größeren Sicherheit eine Aussage getroffen werden soll. Dann wird aber der Konfidenzintervall Radius und damit auch  $\lambda_{st}$  größer.

### 7.1. Bewertung von Abschätzungen

Im Kapitel 5.1 haben wir schon gelernt, daß der Mittelwert eine Abschätzung des Erwartungswerts ist. Wir sahen, daß  $M(\bar{\xi}) = M(\xi)$ ,  $\bar{\xi}$  schwankt sich um den physikalischen Wert (dem Erwartungswert). Diese Eigenschaft nannten wir als **Unverzerrtheit der**

**Abschätzung.** Die allgemeingültige Definition der unverzerrten Abschätzung ist folgendes. Wenn die Abschätzung der Größe  $a$  ist und  $M(\alpha) = a$ , dann heißt es,  $\alpha$  ist eine unverzerrte Abschätzung von  $a$ .

Demnächst werden wir allgemeingültige Bedingungen für die Anwendbarkeit der Methode der kleinsten Fehlerquadrate aussagen. Wir werden die Gaußschen Normalgleichungen zur Bestimmung der Polynom-Koeffizienten aufschreiben.

### 7.1.1. Gauß – Markow Satz

Es seien  $x$  und  $y$  Zufallsvariablen. Nehmen wir an, daß folgende Bedingungen erfüllt sind:

1.  $y = \sum_{i=0}^n a_i x^i$ , wo  $n$  eine gegebene natürliche Zahl ist,  $a_i$  sind reelle Zahlen für alle Indizes  $i = 0, 1, \dots, n$ . Das heißt, es gibt ein polynomialer Zusammenhang zwischen  $x$  und  $y$ .
2. Die Messung (Einstellung) von  $x$  ist nicht mit Fehlern behaftet, sie ist eine deterministische Variable.  
 $y$  ist mit Zufallfehlern behaftet ( $y$  kann nie genau gemessen werden). Es gibt eine Messreihe  $\{x_j, \eta_j\}_{j=1}^N$ . Das  $j$ -te Meßergebnis ist  $\eta_j = y_j + \varepsilon_j$ , und es gilt  $M(\eta_j) = y_j$ .
3. Die Meßergebnisse  $\eta_j$  sind von einander unabhängig.

Die Werte  $a_i$ , ( $i = 0, 1, \dots, n$ ) bezeichnen die mit Hilfe der Methode der kleinsten Fehlerquadrate berechneten Koeffizienten. Dann ist  $M(\alpha_i) = a_i$ , das bedeutet, daß  $\alpha_i$  eine unverzerrte Abschätzung von  $a_i$  ist.

### 7.1.2. Anwendung des Gauß-Markowschen Satzes in der Praxis

Über die Bedingungen des Gauß-Markowschen Satzes kann der Ingenieur folgendes sagen.

1. In der Praxis ist es fast nie der Fall, daß zwischen den zwei Variablen ein polynomialer Zusammenhang existiert.
2. Die zweite Bedingung ist problemspezifisch, in manchen Fällen ist sie erfüllt. Wenn die unabhängige Veränderliche z.B. die Zeit oder der Weg ist, können die genau gemessen werden.
3. Die Unabhängigkeit der Veränderlichen ist im Allgemeinen erfüllt.

Die Methode der kleinsten Fehlerquadrate soll im Sinn des Obigen angewendet werden. Der Zusammenhang zwischen den Veränderlichen sei in Form  $y = f(x)$  geschrieben, wir nähern aber die Funktion  $f(x)$  mit der Funktion  $g(x)$  an und wir werden ihre Näherungsfunktion  $g_n(x)$  mittels der Meßergebnisse bestimmen. So kennen wir zwischen den Funktionen  $g$  und  $f$  keinen Zusammenhang. Wir minimieren den Abstand der Funktion  $g_n(x)$  und der Meßpunkte. Dadurch können wir sichern, daß die Funktion  $g_n(x)$  zwischen den Meßpunkten läuft, wir können aber mathematisch keine exakte Behauptung formulieren.

### 7.1.3. Gaußsche Normalgleichungen

Die Methode der kleinsten Fehlerquadrate für die Meßdaten  $\{x_j, \eta_j\}_{j=1}^N$  heißt:

$$F(\alpha_0, \alpha_1, \dots, \alpha_n) = \sum_{j=1}^N \left( \eta_j - \sum_{i=0}^n \alpha_i x_j^i \right)^2 \quad (100)$$

Diese Summe soll als Funktion der Parameter  $\alpha_0, \alpha_1, \dots, \alpha_n$  minimiert werden. Es ist wichtig zu betonen, daß der Index  $j$  die Meßpunkte,  $i$  die Ordnung der Glieder der Näherungsfunktion bezeichnen. Die notwendige Bedingung des Minimums ist, daß die partiellen Ableitungen der Summe nach  $\alpha_k$  ( $k = 0, 1, \dots, n$ ) alle gleich 0 sind:

$$\frac{\partial}{\partial \alpha_k} \left[ \sum_{j=1}^N \left( \eta_j - \sum_{i=0}^n \alpha_i x_j^i \right)^2 \right] = 0 \quad (101)$$

Nach Durchführung der Ableitung wird:

$$-2 \sum_{j=1}^N \left( \eta_j - \sum_{i=0}^n \alpha_i x_j^i \right) x_j^k = 0. \quad (102)$$

Nach Dividieren mit 2 werden die Klammerausdrücke umformt, damit die Koeffizienten  $\alpha_i$  bestimmt werden können:

$$\sum_{j=1}^N \sum_{i=0}^n \alpha_i x_j^{i+k} = \sum_{j=1}^N \eta_j x_j^k \quad (k = 0, 1, \dots, n) \quad (103)$$

Da die Reihenfolge der endlichen Summen getauscht werden darf, folgt

$$\sum_{i=0}^n \alpha_i \sum_{j=1}^N x_j^{i+k} = \sum_{j=1}^N \eta_j x_j^k \quad (104)$$

Obiges Gleichungssystem kann auch in Matrizenform geschrieben werden. Die Reihen der Matrix sind die  $k$  Werte der Gleichungen, die Spalten die  $i$  Werte der Summen:

$$\begin{pmatrix} N & \sum_{j=1}^N x_j & \sum_{j=1}^N x_j^2 & \sum_{j=1}^N x_j^3 & \dots \\ \sum_{j=1}^N x_j & \sum_{j=1}^N x_j^2 & \sum_{j=1}^N x_j^3 & \dots & \\ \vdots & & & & \\ \sum_{j=1}^N x_j^n & \sum_{j=1}^N x_j^{n+1} & \dots & & \end{pmatrix} \begin{pmatrix} \alpha_0 \\ \alpha_1 \\ \vdots \\ \alpha_n \end{pmatrix} = \begin{pmatrix} \sum_{j=1}^N \eta_j \\ \sum_{j=1}^N \eta_j x_j \\ \vdots \\ \sum_{j=1}^N \eta_j x_j^n \end{pmatrix} \quad (105)$$

So erhielten wir ein algebraisches Gleichungssystem für die Koeffizienten  $\alpha_k$ .

Wenn z.B. eine Annäherung von kubischer Ordnung vorausgesetzt wird, soll dieses Gleichungssystem für  $k = 3$  gelöst werden. Das Excel-Programm wendet auch diese Methode an.

## 8. Achte Vorlesung

In dieser Vorlesung werden die statistischen Tests vorgestellt.

### 8.1. Parametertests

*Beispiel 6:* Man betrachte die 1kg Zucker enthaltende Säcke auf dem Regal eines Lebensmittelgeschäftes. Der Kunde kann sich fragen, ob die Säcke tatsächlich 1kg Zucker enthalten. Nehmen wir an, dass wir das aktuelle Gewicht der Säcke mit der Kontrollwaage wiegen können. Anhand unserer Beobachtungen möchten wir über das Gewicht der Zuckersäcke statistisch begründete Aussagen formulieren. Die Frage kann mathematisch folgendermaßen formuliert werden.

Es sei unsere Meßreihe  $\xi_1, \xi_2, \dots, \xi_n$ . Man kann um den Mittelwert  $\bar{\xi}$  dieser Meßdaten ein Konfidenzintervall mit dem Radius  $a$  bestimmen das sowohl von der Größe  $n$  der Meßreihe als auch von dem von uns gewähltem Signifikanzniveau  $p$  abhängt. Es sei  $m_0 = 1kg$ . Es gibt zwei Möglichkeiten.

- Wenn  $m_0 \in [\bar{\xi} - a; \bar{\xi} + a]$ , dann ist
  1. entweder die Masse des Zuckersacks tatsächlich  $m_0 = 1kg$ ,
  2. oder  $m_0 \neq 1kg$ , aber aus der Meßreihe scheint es doch der Fall zu sein.
- Wenn  $m_0 \notin [\bar{\xi} - a; \bar{\xi} + a]$ , dann ist
  1. entweder  $m_0 \neq 1kg$
  2. oder  $m_0 = 1kg$ , aber aus der Meßreihe scheint es nicht der Fall zu sein.

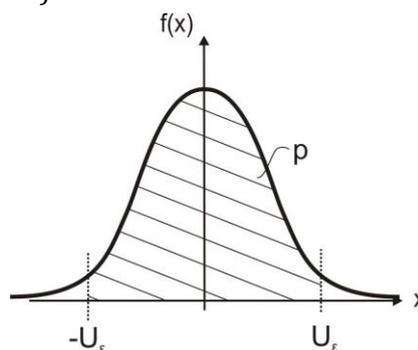
Die Parametertests geben eine mathematische Antwort auf die Frage, welche der oben aufgezählten Fälle in der konkreten Untersuchung auftritt. Die allgemeine Methode heißt **Gauß-Test** oder **U-Test**.

### 8.1.1. Der U-Test

Die Beobachtungen der Zufallsvariable  $\xi$  sind  $\xi_1, \xi_2, \dots, \xi_n$ . Nehmen wir an, daß die Standardabweichung  $\sigma$  der normalverteilten Zufallsvariable  $\xi$  bekannt ist. Dann kann die Annahme gemacht werden, daß der Erwartungswert von  $\xi$  eine gegebene Zahl  $a_0$  ist. Diese Annahme heißt **Nullhypothese**, die als  $H_0$  bezeichnet ist. Wir führen eine Größe  $U_{akt}$  ein, sie ist der aktuelle Wert des Tests:

$$U_{akt} = \frac{\bar{\xi} - a_0}{\sigma} \sqrt{n}. \quad (106)$$

Wenn der Erwartungswert von  $\xi$   $a_0$  ist, dann sind  $M(U_{akt}) = 0$  und  $D^2(U_{akt}) = 1$ , d.h.  $U_{akt} \in N(0,1)$ , da (106) ja gerade die standardisierte Form des Mittelwertes ist. Wir kennen die Dichte der Standardnormalverteilung, so können wir für ein gegebenes Signifikanzniveau  $p$  ein Intervall bestimmen (Bild 22).



**Bild 22** Die Fläche  $p$  unter der Dichtefunktion gibt dasjenige Intervall an, welches die Zufallsvariable mit der Wahrscheinlichkeit  $p$  enthält.

Wenn unsere Hypothese  $H_0$  richtig ist, dann liegt  $U_{akt}$  in diesem Intervall mit der Wahrscheinlichkeit  $p$ . Umgekehrt kann folgendes behauptet werden:

- Ist  $U_{akt} \in [-U_\epsilon, U_\epsilon]$ , so sprechen die statistischen Daten unserer Hypothese  $H_0$  nicht wider. Dann sagen wir, dass wir  $H_0$  **annehmen**.

- Ist  $U_{akt} \notin [-U_\varepsilon, U_\varepsilon]$ , so widersprechen die statistischen Daten der Hypothese  $H_0$ . Dann sagen wir, wir **lehnen**  $H_0$  **ab**.

## 8.2. Die allgemeine Vorgehensweise der statistischen Tests

Ziel eines statistischen Tests ist auf eine gestellte Frage statistisch begründete Antwort zu geben, Entscheidung zu treffen. Wenn z.B. eine Länge mit zwei Schublehren gemessen wird, kann gefragt werden, ob die Schublehren vertauschbar sind. Um die Frage zu beantworten, müssen Versuche durchgeführt und die Nullhypothese formuliert werden. Zunächst soll die Frage für die Statistik formuliert und die Verteilung der Statistik gefunden werden. Wir können ein Signifikanzniveau  $p$  wählen und aus der Verteilung können die Intervallgrenzen  $U_\varepsilon$  und  $-U_\varepsilon$  berechnet werden. Danach soll, der aktuelle Wert der Statistik bestimmt werden, damit wir über die Nullhypothese entscheiden können. Die Schritte der statistischen Tests sind:

1. Formulierung der Frage.
2. Durchführung von Experimenten.
3. Aufstellung der Nullhypothese.
4. Ausarbeitung der Statistik, womit auf die gestellte Frage geantwortet werden kann. Bestimmung der Verteilung der Statistik.
5. Berechnung der Intervallgrenzen  $-U_\varepsilon$  und  $U_\varepsilon$  beim Signifikanzniveau  $p$ .
6. Berechnung des aktuellen Wertes  $U_{akt}$ .
7. Feststellung, ob  $U_{akt} \in [-U_\varepsilon, U_\varepsilon]$ . Wenn ja, dann wird  $H_0$  angenommen, wenn nicht, dann wird sie abgelehnt.

Die Schritte 4 und 5 sind nicht Aufgaben des Ingenieurs, er wendet diese Schritte nur an. Deswegen werden in dieser Vorlesung die Tests nicht ausführlich besprochen, nur aufgezählt.

Die möglichen Fehler unserer Entscheidungen, als wir die Nullhypothese  $H_0$  annehmen oder ablehnen, können klassifiziert werden. **Wir wissen** in der Tat **nie, ob die Nullhypothese richtig oder falsch ist**. Deswegen müssen wir darüber im Klaren sein, wie groß die Wahrscheinlichkeit der falschen Entscheidung ist.

Wenn die Hypothese wahr ist und trotzdem abgelehnt wird, spricht man über Fehler **erster Art**, diese Möglichkeit hat die Wahrscheinlichkeit  $1-p$ . Wenn die falsche Hypothese angenommen wird, spricht man über Fehler **zweiter Art**. Die Wahrscheinlichkeit dieser Möglichkeit kann nicht gegeben werden, da wir in diesem Fall die richtige Verteilung nicht kennen. Die Wahrscheinlichkeit des Fehlers zweiter Art hängt auch von der Güte unserer Nullhypothese ab.

	$H_0$ ist wahr	$H_0$ ist falsch
und wird angenommen	X	Fehler zweiter Art
und wird abgelehnt	Fehler erster Art	X

*Beispiel 7:* Die Aufgabe des Beispiels 6 soll weiter diskutiert werden. Die Anzahl der durchgeführten Massenmessungen sei  $n = 25$ . Der Mittelwert sei  $\bar{\xi} = 0,98\text{kg}$ , die Standardabweichung sei bekannt,  $\sigma = 0,05\text{kg}$ . Unsere Nullhypothese  $H_0$  ist, daß der Erwartungswert des Gewichts der Zuckersäcke  $m_0 = 1\text{kg}$  ist. Zunächst wird der aktuelle Wert berechnet.

$$U_{akt} = \frac{\bar{\xi} - m_0}{\sigma} \sqrt{n} = \frac{0,98 - 1}{0,05} \sqrt{25} = -2$$

Für  $p = 95\%$ , ist  $U_\varepsilon = 1,96$ , und  $U_{akt} \notin [-U_\varepsilon, U_\varepsilon]$ , wir lehnen also unsere Nullhypothese ab. Wenn aber  $p = 99\%$ , dann ist  $U_\varepsilon = 2,58$ , und  $U_{akt} \in [-U_\varepsilon, U_\varepsilon]$ , unsere Nullhypothese wird also angenommen. Es ist nicht der richtige Weg, das Signifikanzniveau zu erhöhen, wenn wir eine sichere Antwort geben wollen, Es ist richtiger, die Anzahl der Experimente zu erhöhen und den aktuellen Wert erneut zu bestimmen.

### 8.3. t-Test

Wenn die Standardabweichung  $\sigma$  nicht bekannt ist, wenden wir nicht den U-Test, sondern den **t-Test** an. Es sei  $\xi \in N(a_0, \sigma)$ , und unsere Beobachtungen seien  $\xi_1, \xi_2, \dots, \xi_n$ . Da der korrigierte empirische Standardabweichung  $s^*$  für die Standardabweichung  $\sigma$  eine unverzerrte Abschätzung ist, so werden wir  $\sigma$  mit  $s^*$  abschätzen. Unsere Nullhypothese ist jetzt, dass der Erwartungswert des Mittelwertes  $\bar{\xi}$  unserer Stichprobe eine gegebene Zahl  $a_0$  ist. Der aktuelle Wert der Statistik wird so berechnet:

$$t_{akt} = \frac{\bar{\xi} - a_0}{s^*} \sqrt{n} \tag{107}$$

Damit folgt  $t_{akt}$  die Studentsche-Verteilung. Jetzt können für eine gewählte Signifikanz  $p$  die Schranken  $-t_\varepsilon$  und  $t_\varepsilon$  gefunden werden. Der Wert von  $t_\varepsilon$  hängt sowohl von  $p$ , als auch von  $n$  ab, er ist aus Tabellen zu entnehmen, oder mit dem Rechner zu bestimmen. Die Nullhypothese  $H_0$  wird angenommen, wenn  $t_{akt} \in [-t_\varepsilon, t_\varepsilon]$ .

### 8.4. Gekoppelter t-Test

Nehmen wir an, wir messen die Länge eines Werkstücks mit zwei Meßinstrumenten. Die Frage ist, ob die Instrumente einheitlich funktionieren. Die Meßergebnisse mit dem ersten Instrument seien  $\xi_1, \xi_2, \dots, \xi_n$ , die mit dem zweiten  $\eta_1, \eta_2, \dots, \eta_n$ . Die Nullhypothese ist:  $M(\xi) = M(\eta)$ . Da aber die zwei Variablen nicht unabhängig sind, können wir nicht feststellen, ob die Erwartungswerte gleich sind. Statt dessen betrachten wir die Differenzen der Meßergebnisse  $\delta_i = \xi_i - \eta_i$ , ( $i = 1, 2, \dots, n$ ). Dann kann unsere neue Nullhypothese  $M(\delta) = 0$  sein. Jetzt kann schon der t-Test angewendet werden. Der aktuelle Wert ist

$$t_{akt} = \frac{\bar{\delta} - 0}{s_\delta^*} \sqrt{n} \tag{108}$$

Es ist von Interesse, daß in der pharmazeutischen Industrie auch dieser gekoppelte t-Test angewendet wird, wenn darüber zu entscheiden ist, ob die Wirkungen von zwei Medikamenten verschieden sind.

### 8.5. Zweistichprobentests U-, bzw. t

Zweistichprobentests U-, bzw. t werden gebraucht, wenn die zwei Stichproben von einander unabhängig sind und zu entscheiden ist, ob die Erwartungswerte der Stichproben gleich sind. Diese Probe wird nicht in der Vorlesung, sondern in der Seminarübung kennengelernt.

### 8.6. Grubbs Test

Eine neue Frage ist, ob der kleinste oder der größte Element einer Stichprobe zur selben Verteilung, wie die anderen Elemente gehört. Die gemessenen Ausprägungen sind wieder  $\xi_1, \xi_2, \dots, \xi_n$  und  $\xi \in N(a_0, \sigma)$ . Es sei  $\xi_{min}$  das kleinste,  $\xi_{max}$  das größte Element. Man kann  $\bar{\xi}$  und  $s^*$  bestimmen, danach wird das Verhältnis der Differenz des kleinsten (größten) Elementes und des Mittelwertes zur korrigierten empirischen Standardabweichung berechnet. Der aktuelle Wert ist also:

$$g_{akt} = \frac{\bar{\xi} - \xi_{min}}{s^*} . \tag{109}$$

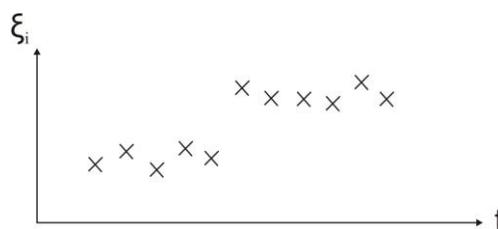
Der kritische Wert  $g_{krit}$  kann anhand des Signifikanzniveaus  $p$  und der Stichprobengröße  $n$  aus Tabellen entnommen werden. Wenn  $g_{akt} < g_{krit}$ , dann wird die Nullhypothese  $\xi_{min} \in N(a_0, \sigma)$  angenommen. Analogerweise für den maximalen Wert ist der aktuelle Wert der Statistik:

$$g_{akt} = \frac{\xi_{max} - \bar{\xi}}{s^*} . \tag{110}$$

Die Entscheidung ist ähnlich, wie vorher.

### 8.7. Abbe Test

Das Meßinstrument kann während der Messung beschädigt werden, oder die Nebenbedingungen der Messung können sich geändert haben. Die Frage ist, ob alle Ausprägungen des Versuchs zur selben Verteilung gehören.



**Bild 23** Die zum Zeitpunkt  $t$  gemessenen Ergebnisse  $\xi_i$

Nehmen wir an, daß  $\xi \in N(a_0, \sigma)$ , die Meßdaten sind  $\xi_1, \xi_2, \dots, \xi_n$ . Unsere Nullhypothese ist, daß alle Meßdaten den selben Erwartungswert haben, d.h.  $M(\xi_i) = a_0$ , ( $i = 1, 2, \dots, n$ ). Wir untersuchen, ob sich der Erwartungswert während des Versuchs geändert hat. Nicht nur die korrigierte empirische Varianz  $s^{*2}$ , sondern auch folgende Summe der Differenzen der Meßwertenpaare werden jetzt berechnet:

$$q^2 = \frac{1}{2(n-1)} \sum_{i=1}^{n-1} (\xi_{i+1} - \xi_i)^2 . \tag{111}$$

Wir betonen, daß jetzt die Reihenfolge der Meßwerte eine Rolle spielt. Der aktuelle Wert der Statistik sei das Verhältnis von  $q^2$  zu  $s^{*2}$ :

$$r_{akt} = \frac{q^2}{s^{*2}} \quad (112)$$

Wenn sich der Erwartungswert nicht geändert hat, so weichen die Größen  $q^2$  und  $s^{*2}$  von einander nicht sehr ab. Wenn aber in dem Erwartungswert ein Sprung entsteht, dann ändert sich nur ein einziges Glied der Summe in  $q^2$  aber alle Glieder von  $s^{*2}$  ändern sich. Dadurch nimmt der Wert von  $r_{akt}$  ab. Die kritischen Werte dieses Tests können auch jetzt aus Tabellen entnommen werden. Die Nullhypothese wird angenommen, wenn  $r_{akt} > r_{krit}$ . Der kritische Wert hängt vom Freiheitsgrad ab.

### 8.8. F-Test

Mit dem F oder Fischer Test werden die Varianzen von zwei Beobachtungen mit einander verglichen. Eine der Stichproben sei  $\xi_1, \xi_2, \dots, \xi_n$  mit der unbekanntem Varianz  $\sigma_\xi^2$ . Die andere Stichprobe sei  $\eta_1, \eta_2, \dots, \eta_m$  mit der unbekanntem Varianz  $\sigma_\eta^2$ . Nehmen wir weiterhin an, daß  $\xi$  und  $\eta$  normalverteilt sind. Unsere Nullhypothese ist, daß die Varianzen gleich sind,  $\sigma_\xi^2 = \sigma_\eta^2$ . Für die Nullhypothese sollen die korrigierten empirischen Varianzen berechnet werden:

$$s_\xi^{*2} = \frac{1}{n-1} \sum_{i=1}^n (\bar{\xi} - \xi_i)^2 \quad \text{und} \quad s_\eta^{*2} = \frac{1}{m-1} \sum_{j=1}^m (\bar{\eta} - \eta_j)^2 \quad (113)$$

Der aktuelle Wert ist:

$$F_{akt} = \max \left\{ \frac{s_\xi^{*2}}{s_\eta^{*2}}, \frac{s_\eta^{*2}}{s_\xi^{*2}} \right\} \quad (114)$$

Der kritische Wert befindet sich zum Signifikanzniveau  $p$  und zu den Freiheitsgraden  $n-1$  und  $m-1$  in Tabellen. Die Nullhypothese wird angenommen, wenn der aktuelle Wert kleiner, als der kritische Wert ist:  $H_0$  ist war, wenn  $F_{akt} < F_{krit}(p, m-1, n-1)$ .

## 9. Neunte Vorlesung

Nach Darstellung der Binomialverteilung werden die Grundbegriffe und Methoden der Qualitätskontrolle besprochen Die Dias sind auf folgende Webseite zu finden:

<http://www.hds.bme.hu/letoltesek/targyak/BMEGEVGAG14/minell.pdf>

## 10. Zehnte Vorlesung

Bisher haben wir solche Tests behandelt, wo die Nullhypothese mit einem Parameter verknüpft war. Wir werden nun solche Tests untersuchen, wo die Verteilung im Fokus steht. In solchen Fällen gibt es drei Haupttypen der Tests, der sogenannten **Nichtparameter Tests**. Der erste Test ist der **Anpassungstest** wo die empirische Verteilung der Stichprobe mit einer theoretischen Verteilung verglichen wird, es wird untersucht, wie gut die beiden Verteilungen an einander passen. Der zweite ist der **Homogenitätstest**, es wird untersucht ob zwei Stichproben oder Meßreihen zu gleichen Verteilungen gehören. Der dritte Typ ist der **Unabhängigkeitstest**, dabei interessiert uns ob zwei Veränderlichen unabhängig sind.

In allen diesen Fällen wird hauptsächlich der  $\chi^2$ -Test angewendet, aber bei kleineren Stichprobengrößen ist auch die Anwendung des **Ryan-Joiner Tests** im Gebrauch.

### 10.1. Anpassungsuntersuchung mit $\chi^2$ -Test

Der Kern dieses Tests kann anhand eines Beispiels vorgestellt werden.

*Beispiel 8:* Man betrachte einen Würfel mit sechs Seitenflächen, es soll überprüft werden, ob unser Würfel regelmäßig ist. Wenn ja, dann erwartet man, daß wenn man oft genug würfelt, ergeben sich alle sechs Zahlen ( $r=6$ ) mit näherungsweise gleicher Häufigkeit. Man betrachte folgendes Würfelspielergebnis. Die Größe der Stichprobe ist  $N = 840$ . Die Häufigkeiten sind in einer Tabelle angegeben. Man trägt in die Tabelle auch ein, was das theoretische Ergebnis ist, wenn ein regelmäßiger Würfel vorausgesetzt wird.

Punktzahl Im Spiel ( $i$ )	Häufigkeit ( $v_i$ )	theoretische Wahrscheinlichkeit ( $p_i$ )	theoretische Häufigkeit ( $Np_i$ )
1	124	1/6	140
2	152	1/6	140
3	130	1/6	140
4	148	1/6	140
5	152	1/6	140
6	134	1/6	140

Man betrachte folgende Summe, die als aktueller Wert des  $\chi^2$ -Tests genannt wird:

$$\chi^2_{akt} = \sum_{i=1}^r \frac{(\text{erwartet} - \text{gemessen})^2}{\text{erwartet}} = \sum_{i=1}^r \frac{(Np_i - v_i)^2}{Np_i} \quad (115)$$

Wenn der Würfel regelmäßig ist, ist  $Np_i \approx v_i$ , und  $\chi^2_{akt}$  ist klein.

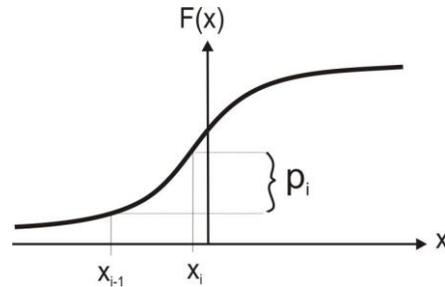
Der kritische Wert dieses Tests kann für ein Signifikanzniveau  $1-p$  und Klasse  $r$  aus der  $\chi^2$  Verteilung bestimmt werden:  $\chi^2_{krit}(1-p; r-1)$ . Im Beispiel ist  $r = 6$ , da der Würfel sechs Seiten hat. Die Nullhypothese wird beibehalten, wenn  $\chi^2_{akt} < \chi^2_{krit}$ . Im Beispiel ist mit  $p = 95\%$ , d.h. 5% Signifikanzniveau  $\chi^2_{akt} = 5,29$ , der kritische Wert ist  $\chi^2_{krit} = 11,1$ , d.h. die Nullhypothese, daß der Würfel regelmäßig ist, wird beibehalten. Der Wert  $1-p$  ist die Wahrscheinlichkeit des Ereignisses, daß die Nullhypothese abgelehnt wird, obwohl sie wahr ist, d.h. Signifikanzniveau  $1-p$  ist die Wahrscheinlichkeit des Fehlers erster Art.

#### 10.1.1. Anwendung: Normalitätstest

Es wird oft gefragt, ob eine Stichprobe zu einer Normalverteilung gehört. Die Beobachtungen seien  $\xi_1, \xi_2, \dots, \xi_N$ . Sie werden zunächst in  $r$  Klassen (Intervallen) geteilt und die Häufigkeiten der Ausprägungen der einzelnen Klassen wird zusammengezählt. Danach wird aus der Normalverteilung bestimmt, mit welcher Wahrscheinlichkeit eine aus der Normalverteilung resultierende Beobachtung in die einzelnen Klassen fallen würde. Wenn die Verteilung  $F(x)$  bekannt ist, zeigt sie an der Stelle  $x$  die Wahrscheinlichkeit des Ereignisses, daß die Variable

kleiner als  $x$  ist. So ist die Wahrscheinlichkeit  $p_i$ , daß ein Wert im Intervall  $[x_{i-1};x_i]$  liegt, die Differenz der Funktionswerte (siehe Bild 24):

$$p_i = F(x_i) - F(x_{i-1}). \quad (116)$$



**Bild 24** Die Wahrscheinlichkeit, daß eine Ausprägung im Intervall  $[x_{i-1},x_i]$  liegt, kann aus der Verteilungsfunktion berechnet werden

Es lohnt sich, eine Tabelle zusammenzustellen.

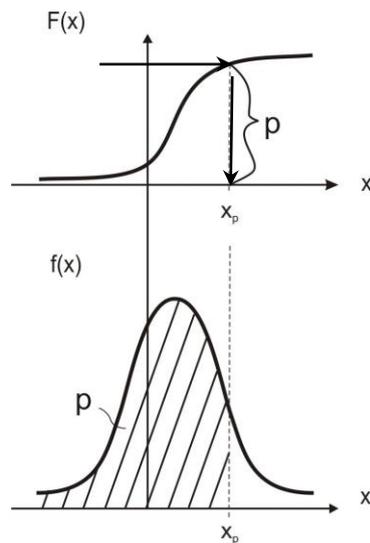
Intervall	Häufigkeit	Verteilungsfunktion	$p_i$	$\frac{(Np_i - \nu_i)^2}{Np_i}$
$-\infty \rightarrow x_1$	$\nu_1$	$F(x_1)$	$F(x_1)$	...
$x_1 \rightarrow x_2$	$\nu_2$	$F(x_2)$	$F(x_2) - F(x_1)$	
$x_2 \rightarrow x_3$	$\nu_3$	$F(x_3)$	$F(x_3) - F(x_2)$	
...	...	...	...	
$x_{r-1} \rightarrow +\infty$	$\nu_r$	$F(x_r=+\infty)=1$	$1 - F(x_{r-1})$	...
	$\Sigma=N$			$\Sigma$

Wenn die Wahrscheinlichkeit des Ereignisses, eine Ausprägung liegt in einem definierten Intervall, bestimmt wird, kann der aktuelle Wert des  $\chi^2$  Tests berechnet werden. Berechnen wir die Summanden der Formel (115) für alle Intervalle und summieren wir sie. Es ist sinnvoll, die Intervallgrenzen so zu wählen, daß die Häufigkeiten der in den einzelnen Intervallen liegenden Werte etwa gleich seien. Wir erinnern uns daran, daß wir bei der Histogramm Konstruktion auch diesen Weg gingen. Die  $F(x_i)$  Werte können in Excel problemlos gefunden werden. Der kritische Wert der Probe wird zu einem Signifikanzniveau  $1-p$  und zu der Anzahl der Klassen  $r$  bestimmt:  $\chi^2_{krit}(1-p; r-1)$ .

## 10.2. Anpassungsuntersuchung mit Ryan-Joiner Test

Für Anpassungstests bei kleineren Stichprobengrößen wird auch der Ryan-Joiner Test angewendet. Im wesentlichen wird das aus der Stichprobe konstruierte Histogramm mit der theoretischen Verteilungsfunktion verglichen. Wenn die beiden Funktionen ausreichend gut übereinstimmen, wird die Nullhypothese beibehalten. Bevor wir diesen Test näher darstellen, brauchen wir einen neuen Begriff kennenzulernen. Betrachten wir die Zufallsvariable  $\xi$  und ihre Verteilungsfunktion  $F(x)$ .

Das  $p$ - **Perzentil** einer Zufallsvariable mit Verteilungsfunktion  $F(x)$  ist die Zahl  $x_p$  wofür  $p = F(x_p)$ , also der Umkehrwert der Verteilungsfunktion. Es ist im Bild 25 dargestellt.

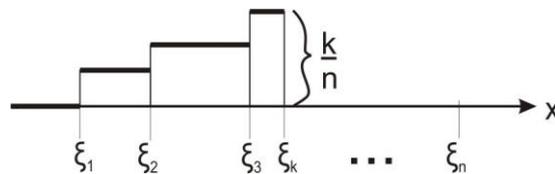


**Bild 25** Zusammenhang zwischen Verteilungs- und Dichtefunktion  
Veranschaulichung des  $p$ -Perzentils

Das  $p$  Perzentil kann auch anders veranschaulicht werden. Die Wahrscheinlichkeit des Ereignisses, daß die Zufallsvariable  $\xi$  Werte annimmt, die kleiner als  $x_p$  sind ist  $p$ . Diese Tatsache kann auch anhand der Dichtefunktion formuliert werden, da die Fläche unter der Dichtefunktion gerade diese Wahrscheinlichkeit  $p$  ist:

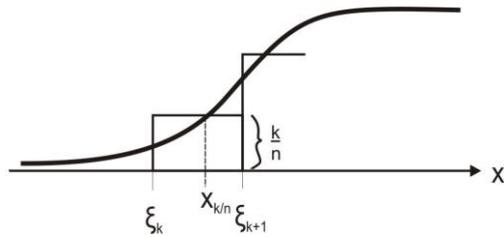
$$p = F(x_p) = \int_{-\infty}^{x_p} f(x) dx. \quad (118)$$

Für die Durchführung des Tests brauchen wir auch die empirische Verteilungsfunktion.



**Bild 26** Empirische Verteilungsfunktion

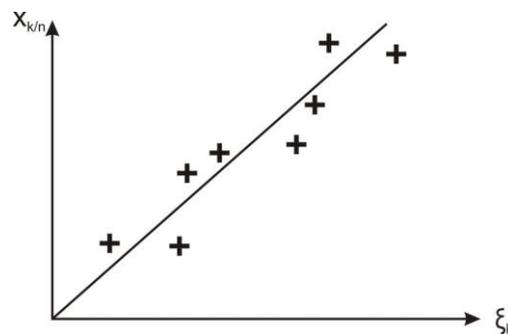
Wir konstruieren diese Funktion aus der der Größe nach geordneter Stichprobe als eine Treppenfunktion, deren Treppen  $1/n$  hoch sind, und die Treppe über der  $k$ -ten Ausprägung genau  $k/n$  hoch ist (Bild 26). Betrachten wir eine Treppe der empirischen Verteilungsfunktion und stellen wir auch die theoretische Verteilungsfunktion mit einer durchgezogenen Linie dar! Es ist im Bild 27 zu sehen. Das  $k/n$  Perzentil ist die Abszisse  $x_{k/n}$  des Schnittpunktes der  $k$ -ten Treppe mit der theoretischen Verteilungsfunktion. Auf dieser Weise können mittels der theoretischen Verteilungsfunktion zu allen diskreten Treppen die Perzentile bestimmt werden. Dadurch kann eine Datenreihe generiert werden, die aus der theoretischen Verteilungsfunktion resultiert und selbe Größe, wie die Stichprobe hat.



**Bild 27** Theoretische Verteilungsfunktion und  $k$ -te Treppe der empirischen Verteilungsfunktion. Das  $k/n$  Perzentil  $x_{k/n}$  gehört zum Schnittpunkt beider Kurven

Auskunft über die Ähnlichkeit dieser generierten Datenreihe und der Stichprobe gibt der Korrelationskoeffizient. Wenn diese Punktpaare in einem Diagramm aufgetragen werden, streuen sie sich um eine Gerade, falls der Korrelationskoeffizient genügend groß ist (Bild 28).

Statistiker haben bewiesen, daß es günstiger ist, das Perzentil an den Stellen  $\frac{k-0,375}{n+0,25}$  anstatt  $k/n$  zu bestimmen.



**Bild 28** Die mit der Ryan-Joiner Methode generierte Datenreihe als Funktion der gemessenen Werte der Stichprobe.

Das aktuelle Wert des Tests ist also der Korrelationskoeffizient  $Korr(x_{pk}, \xi_k)$ , er ist immer positiv, da die Punkte gleichgerichtet sind. Die kritischen Werte können aus einer Tabelle entnommen werden. Die Hypothese wird dann beibehalten, wenn der aktuelle Wert größer als der kritische Wert ist. Dieser Test kann nicht nur für normalverteilte Variablen angewendet werden.

### 10.3. Homogenitätsuntersuchung mit $\chi^2$ -Test

Bei einer Homogenitätsuntersuchung stellt sich die Frage, ob zwei Versuchsreihen gleich verteilt sind? Ein Produkt kann z.B. in der Nachtschicht und am Tag hergestellt werden. Sind die Leistungen der beiden Schichten gleich? Die zwei Zufallsvariablen seien  $\xi$  und  $\eta$  die Verteilungsfunktionen seien  $F(x)$  und  $G(x)$ . Frage: stimmen  $F(x)$  und  $G(x)$  überein? Die Nullhypothese ist also, daß  $F(x) = G(x)$ . Die Ausprägungen der Versuche sind mit  $\xi_1, \xi_2, \dots, \xi_n$  und  $\eta_1, \eta_2, \dots, \eta_m$  bezeichnet. Wir betonen, daß die Stichprobengrößen nicht gleich sein müssen.

Bei der Anwendung des  $\chi^2$  Tests Werden die Variablen in Klassen geteilt und die relativen Häufigkeiten der in den einzelnen Klassen liegenden Ausprägungen zusammengezählt. Wenn sie etwa gleich sind, nehmen wir unsere Nullhypothese an. An einem Beispiel wird der Vorgang dargestellt.

*Beispiel 9:* Man untersucht die Bruchfestigkeit eines Kristalls beim Sonnenlicht und beim Mondlicht. Die gemessenen Werte sind in der Tabelle eingetragen. In unserem Fall sind die

Stichprobengrößen  $n = 120, m = 100$ . Der aktuelle Wert des Homogenitätstests  $\chi^2_{akt}$  wird mit folgender Formel berechnet:

$$\chi^2_{akt} = nm \sum_{i=1}^r \frac{\left(\frac{\nu_i}{n} - \frac{\mu_i}{m}\right)^2}{\frac{\nu_i}{n} + \frac{\mu_i}{m}} \quad (117)$$

Im Beispiel ist die Anzahl der Intervalle  $r = 6, \chi^2_{akt} = 1,19, \chi^2_{krit}(1-p=0,95, r-1=1) = 11,07$ . Da  $\chi^2_{akt} = 1,19 < \chi^2_{krit} = 11,07$ , nehmen wir die Nullhypothese an.

$\sigma_t 10^3 \frac{N}{m^2}$	$\nu_i(\xi)$	$\mu_i(\eta)$
10-15	36	28
16-18	41	36
19-21	28	22
22-25	11	8
26-29	3	4
30-	1	2
$\Sigma$	$n = 120$	$m = 100$

#### 10.4 Unabhängigkeitstest

Die Frage ist ob zwei Zufallsvariablen die gleiche Verteilung haben. Als Beispiel denken wir an zwei geometrische Abmessungen eines Maschinenelements, etwa die Länge und den Durchmesser. Die Stichproben sind mit Wertepaaren gegeben:  $\{\xi_1, \eta_1\} \{\xi_2, \eta_2\} \dots \{\xi_N, \eta_N\}$ . Die Unabhängigkeit der Variablen soll untersucht werden.

Die Nullhypothese ist:  $H_0: P(\xi < x \text{ und } \eta < y) = P(\xi < x) \cdot P(\eta < y)$ , da die Unabhängigkeit genau so definiert wurde.

Als erster Schritt werden beide Stichproben  $\xi$  und  $\eta$  in Klassen aufgeteilt:

$$(-\infty =) x_0 < x_1 < x_2 < \dots < x_r (= \infty) \quad (-\infty =) y_0 < y_1 < y_2 < \dots < y_s (= \infty).$$

Die Anzahl der Klassen kann verschieden sein.

Es werden Ereignisse A und B definiert.

Ereignis A:  $x_{i-1} < \xi < x_i$ .

Ereignis B:  $y_{j-1} < \eta < y_j$ .

Es wird eine Kontingenztabelle aufgestellt, die in den Zellen die gemeinsame Häufigkeiten der Ereignisse A und B enthält. In der rechten Spalte sind die Häufigkeiten des Ereignisses A, in der untersten Reihe der Tabelle, die Häufigkeiten des Ereignisses B angegeben:

	B <sub>1</sub>	B <sub>2</sub>	B <sub>j</sub>	B <sub>s</sub>	
A <sub>1</sub>	$\nu_{11}$	$\nu_{12}$	$\nu_{1j}$	$\nu_{1s}$	$\nu_{1*}$
A <sub>2</sub>	$\nu_{21}$	$\nu_{22}$	$\nu_{2j}$	$\nu_{2s}$	$\nu_{2*}$
A <sub>i</sub>	$\nu_{i1}$	$\nu_{i2}$	$\nu_{ij}$	$\nu_{is}$	$\nu_{i*}$
A <sub>r</sub>	$\nu_{r1}$	$\nu_{r2}$	$\nu_{rj}$	$\nu_{rs}$	$\nu_{r*}$
	$\nu_{*1}$	$\nu_{*2}$	$\nu_{*j}$	$\nu_{*s}$	N

Die Wahrscheinlichkeiten der Ereignisse  $A_i$  und  $B_j$  können mit den relativen Häufigkeiten folgendermaßen abgeschätzt werden:  $P(A_i) \approx \frac{v_{i*}}{N}$  und  $P(B_j) \approx \frac{v_{*j}}{N}$ . Die Statistik

$$\text{ist } \chi_{akt}^2 = N \sum_{j=1}^s \sum_{i=1}^r \frac{\left( v_{ij} - \frac{v_{i*} v_{*j}}{N} \right)^2}{v_{i*} v_{*j}}.$$

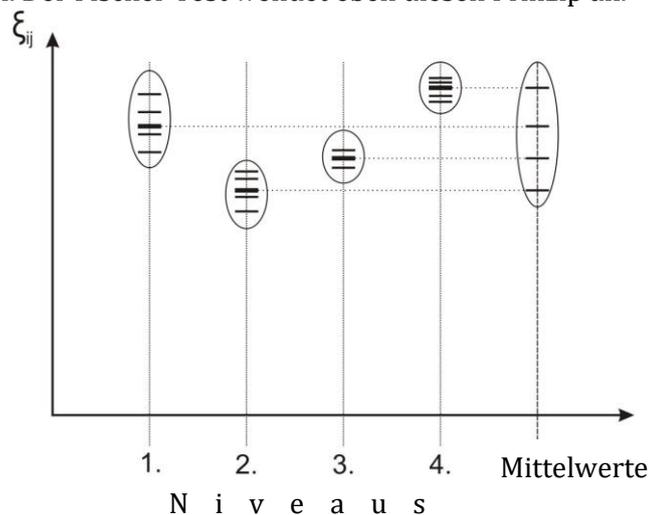
Der Freiheitsgrad der  $\chi^2$  Verteilung ist jetzt  $(r-1)(s-1)$ . Mit der Wahrscheinlichkeit  $p$ , bzw. Signifikanz  $1-p$  kann der kritische Wert  $\chi_{krit}^2$  in Tabellen gefunden werden. Die Nullhypothese wird beibehalten, wenn  $\chi_{akt}^2 \leq \chi_{krit}^2$ , sonst wird sie abgelehnt.

## 10.5 Varianzanalyse

Es soll häufig mittels objektiven Maßnahmen entschieden werden, ob sich verschiedene Typen eines Produktes tatsächlich unterscheiden. Es soll z.B. entschieden werden, ob der Benzinverbrauch von verschiedenen PKW Typen maßgeblich verschieden ist.

Bei der Varianzanalyse haben wir nicht mit quantitativen sondern mit qualitativen Variablen zu tun. Die unabhängigen Variablen werden **Faktoren** genannt. Im obigen Beispiel ist der Autotyp der Faktor. Die Typen der PKW-s sind die **Niveaus** des Faktors. Die abhängige Variable ist der **Versuchsergebnis** oder **Zellenwert**. Die Frage ist, ob die verschiedenen Niveaus auf das Ergebnis eine Auswirkung haben.

Im Bild 29 ist das Ergebnis eines Einfaktorversuchs zu sehen. Es wurden Beobachtungen auf vier Niveaus durchgeführt, die Ergebnisse sind mit dünnen waagerechten Linien im Bild dargestellt. Bei allen Niveaus sind auch die Mittelwerte berechnet, die sind im rechten Teil der Abbildung aufgetragen. Man überprüft, ob die Streuung der Mittelwerte größer, als die der einzelnen Niveaus ist. Wenn ja, dann spielen die Faktoren eine wesentliche Rolle. Wenn die Streuung der Mittelwerte klein ist, dann kann diese Tatsache mit der Varianz der Faktoren erklärt werden. Der Fischer Test wendet eben diesen Prinzip an.



**Bild 29** Einfaktor-Versuch