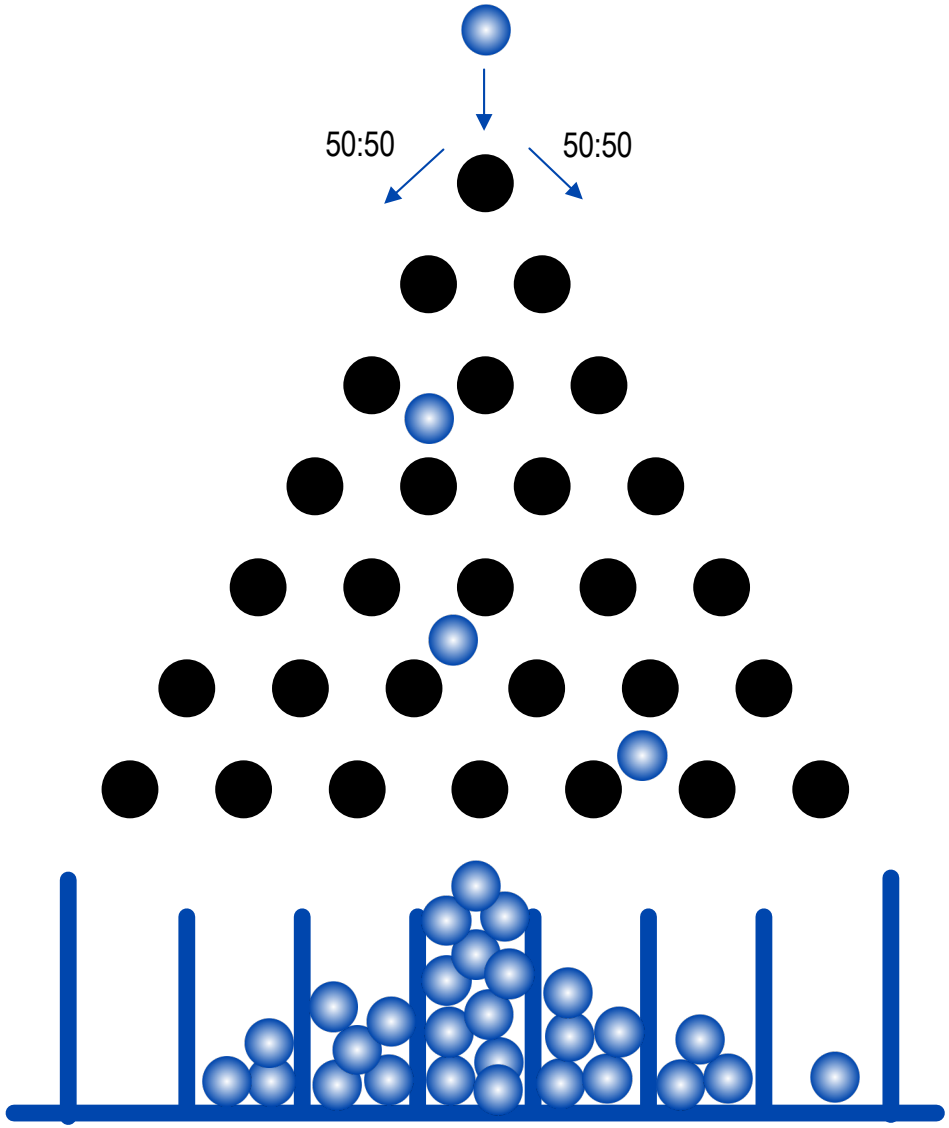
15. Analyse – Statistische Grundlagen

Die Normalverteilung verstehen und

Zusammenhänge zwischen Messgrößen identifizieren

Die Herleitung der Normalverteilung

Das Galton-Brett



Sir Francis Galton
(1822 – 1911)

Die Normalverteilung

(Auch Gaußverteilung oder Glockenkurve)

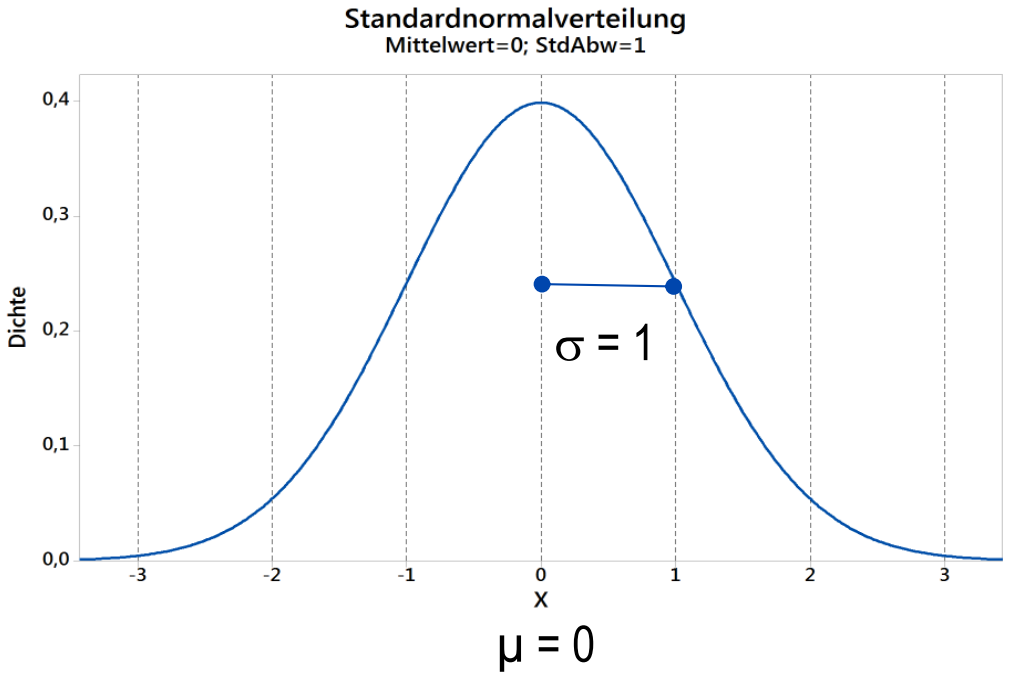
Eigenschaften der Normalverteilung:

Mittelwert = Median

1σ = Wendepunkt

Die Kurve erreicht niemals 0

Fläche unter der Kurve = 100%



Gleichung der Normalverteilungsfunktion („Dichtefunktion“):

$$f(x) = \frac{1}{\sqrt{2\pi \cdot \sigma^2}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}$$

Wahrscheinlichkeitsverteilung:

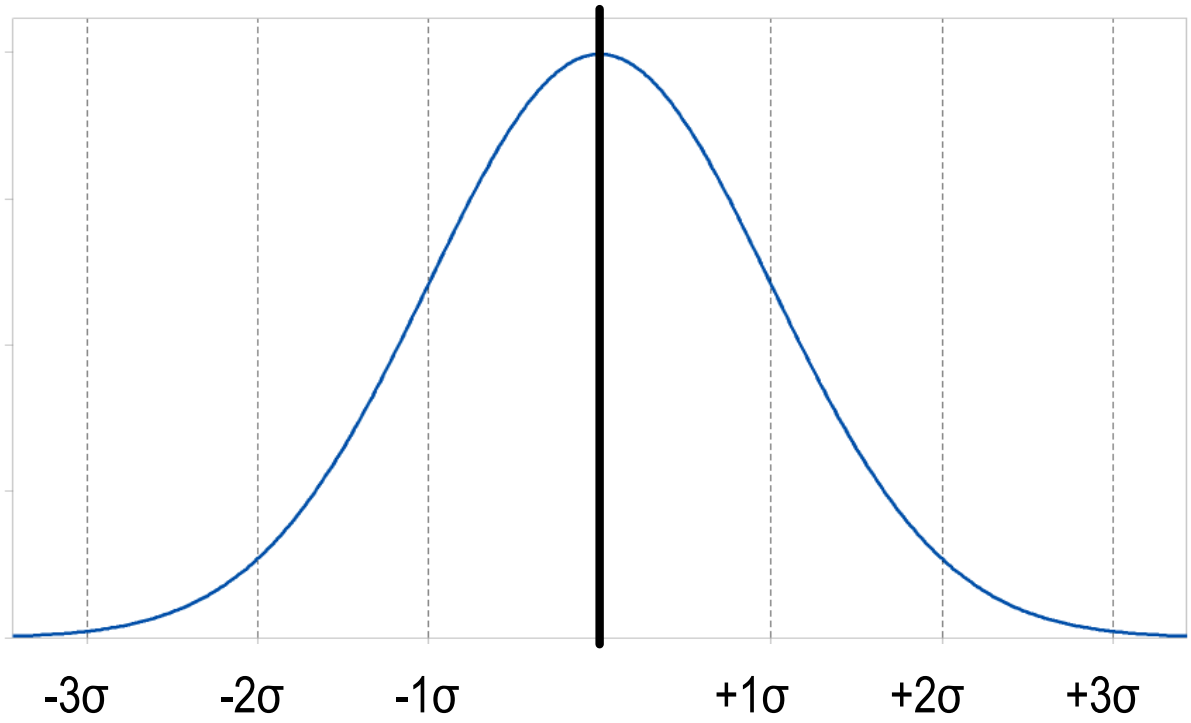
$$f(x) = \frac{1}{\sqrt{2\pi \cdot \sigma^2}} \int_{-\infty}^{x} e^{-\frac{(x-\mu)^2}{2\sigma^2}} \, dx$$

Parameter:

μ: mittlere Lage der Verteilung

σ: Breite der Verteilung (Streuung)

Wahrscheinlichkeit und Sigma



± 1 σ:	68,2%	aller Probenwerte innerhalb	–	31,8% außen
± 2 σ:	95,5%	aller Probenwerte innerhalb	–	4,5% außen
± 3 σ:	99,73%	aller Probenwerte innerhalb	–	0,27% außen

Zusammenhang zwischen Variablen (1)

(kontinuierliche Variablen)

Das **Streudiagramm** visualisiert den Zusammenhang zweier kontinuierlich ausgeprägter Variablen

→ Ergebnis: subjektive Einschätzung des Zusammenhangs

Die **Korrelationsanalyse** prüft, ob es einen linearen statistisch signifikanten Zusammenhang zwischen zwei Variablen gibt

→ Ergebnis 1: Pearson Korrelationskoeffizient r

→ Ergebnis 2: Hypothesentest auf Signifikanz des linearen Zusammenhangs

Die **Regressionsanalyse** beschreibt den Zusammenhang zwischen Variablen anhand einer Regressionsgleichung

→ Ergebnis A: Einfache lineare Regressionsgleichung oder

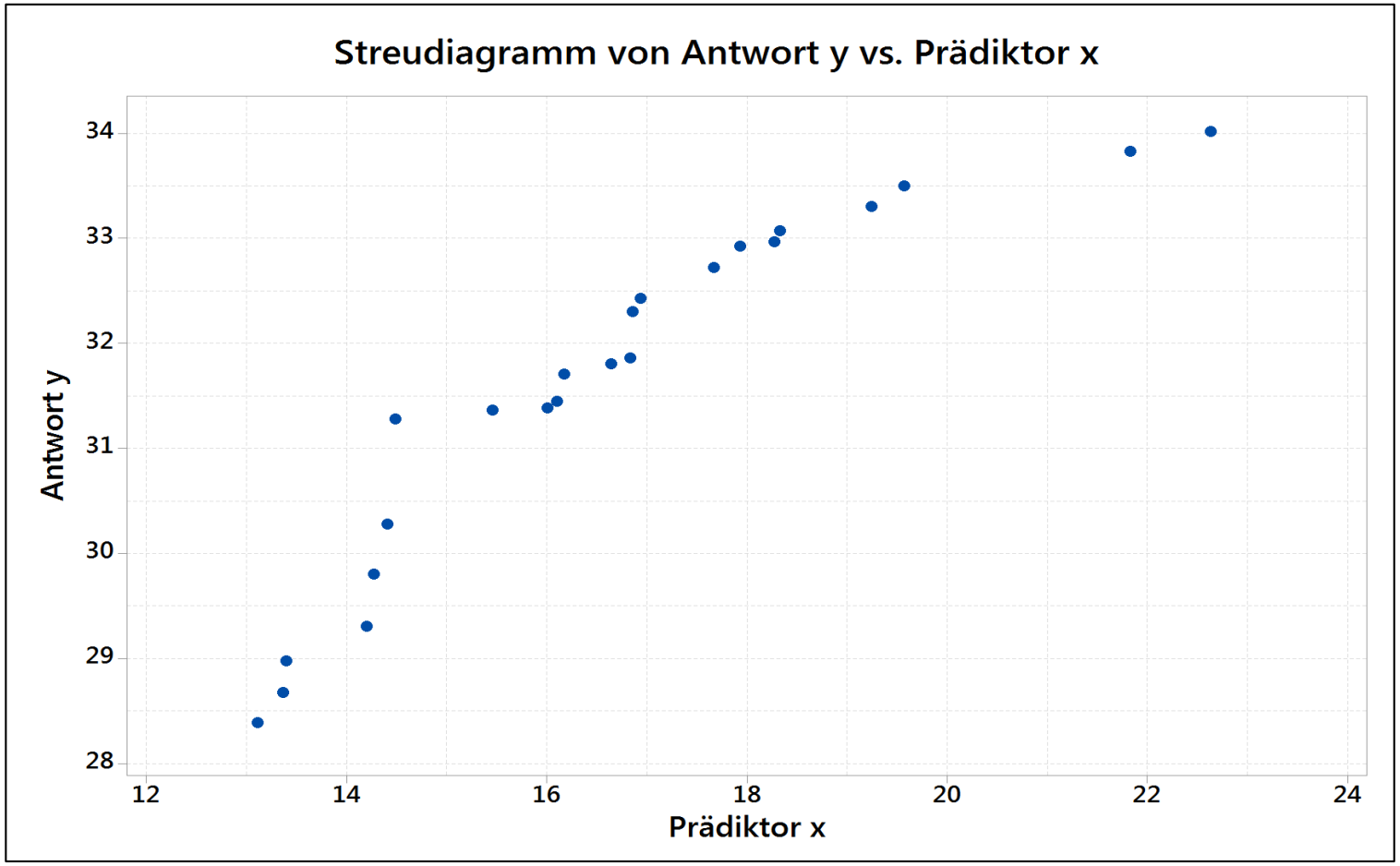
→ Ergebnis B: Einfache nicht-lineare Regressionsgleichung oder

→ Ergebnis C: Multiple lineare Regressionsgleichung oder

→ Ergebnis D: Multiple nicht-lineare Regressionsgleichung

Zusammenhang zwischen Variablen (2)

Das Streudiagramm



Subjektiv betrachtet, gibt es einen positiven Zusammenhang zwischen den beiden Variablen
→ je größer der Prädiktor x, desto größer die Antwort y

Korrelation – Ein Maß für Zusammenhänge

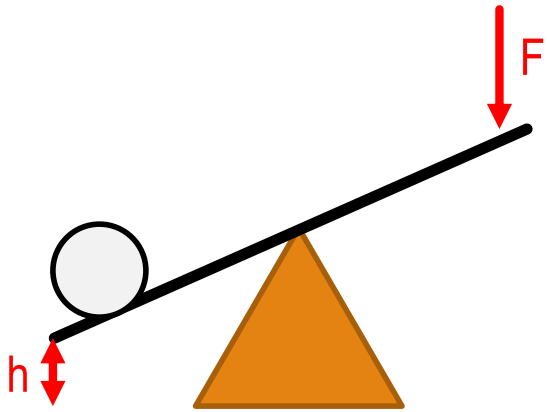
Empirischer Korrelationskoeffizient nach Bravais-Pearson
(kurz: Pearson‘scher Korrelationskoeffizient r)

Es gibt eine mathematische Funktion, die eine Kenngröße r berechnet.
r ist ein Maß über die Stärke des linearen Zusammenhangs zweier Variablen.

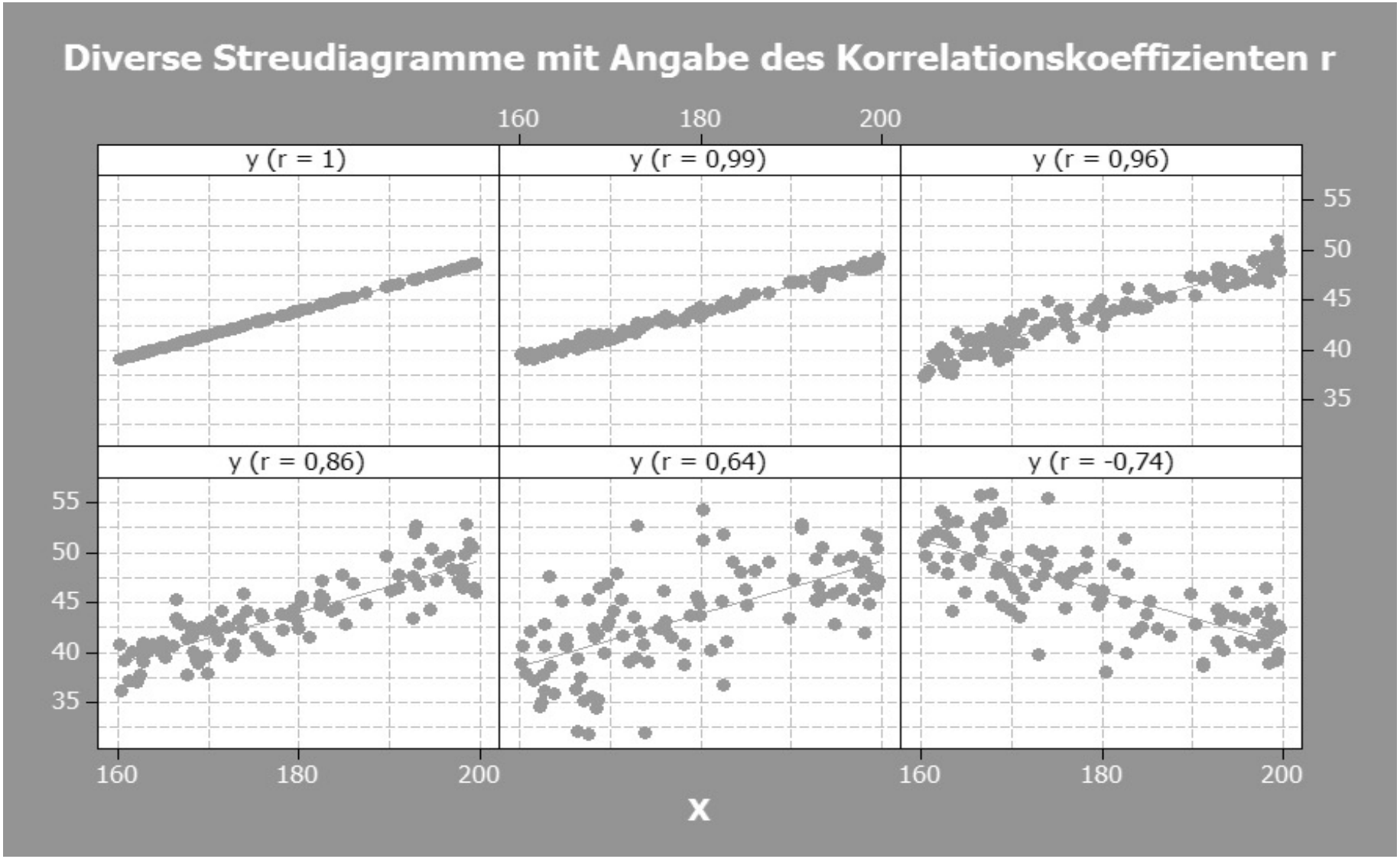
$$r = \frac{\sum (x_i - \bar{x}) * (y_i - \bar{y})}{\sqrt{\sum (x_i - \bar{x})^2 * \sum (y_i - \bar{y})^2}}$$

r kann dabei Werte von -1 bis +1 annehmen.

- +1 bedeutet einen vollständigen positiven Zusammenhang
- -1 bedeutet einen vollständigen negativen Zusammenhang
- 0 bedeutet keinen Zusammenhang



Beispiele für Streudiagramme und Korrelationskoeffizient r



Die Interpretation des Korrelationskoeffizienten r

- Je näher sich r dem Wert 1 bzw. – 1 nährt, desto stärker ist die lineare Beziehung zwischen den beiden Variablen.
- In der Literatur finden sich häufig Schwellen wie
  | r | > 0,75 = gesicherter Zusammenhang

→ Dies ist nicht korrekt!

- Die Signifikanzschwelle (ab welcher man von einem signifikanten Zusammenhang sprechen kann) hängt von der Anzahl der Wertepaare ab.
- In der rechten Tabelle sind die Schwellen angegeben, ab welcher der Korrelations-koeffizient statistisch signifikant ist (bei 95% Vertrauensniveau).
- MINITAB berechnet hierfür den p-Wert.
  Ein p-Wert < 0,05 entspricht derselben bestätigten statistischen Signifikanz (wie in der Tabelle).

Anzahl Wertepaare	min r-Wert der Stichprobe
6	0,77
7	0,68
8	0,60
9	0,58
10	0,55
15	0,44
20	0,38
25	0,34

(bei 95% Vertrauensniveau)

Einfache lineare Regression – Einleitung

- Die einfache lineare Regression ist ein statistisches Modell zur quantitativen Beschreibung eines Zusammenhangs von Eingangs- und Ausgangsgrößen eines Systems / Prozesses.
- Ableitung einer mathematischen Gleichung = Regressionsfunktion
  - Die Berechnung erfolgt i.d.R. nach der Methode der kleinsten Quadrate
- Eingangsgrößen = Unabhängige Variable = X (Prädiktor, Regressor)
- Ausgangsgröße = Abhängige Variable = Y (Antwort, Regressand)

→ Die Regressionsgerade (basierend auf der Regressionsfunktion) ist folglich den einzelnen x-y Wertpaaren optimal angepasst.

Wichtige Unterscheidung:

- Die **Korrelationsanalyse** beantwortet, „ob“ zwischen zwei Variablen ein statistisch signifikanter Zusammenhang besteht (anhand r und p).
- Die **Regressionsanalyse** beschreibt den linearen Zusammenhang zwischen zwei Variablen mathematisch: y = f(x)

Einfache und multiple Regression

Regressionstypen:

einfache lineare Regression:

$y = a_0 + a_1 * x$

multiple lineare Regression:

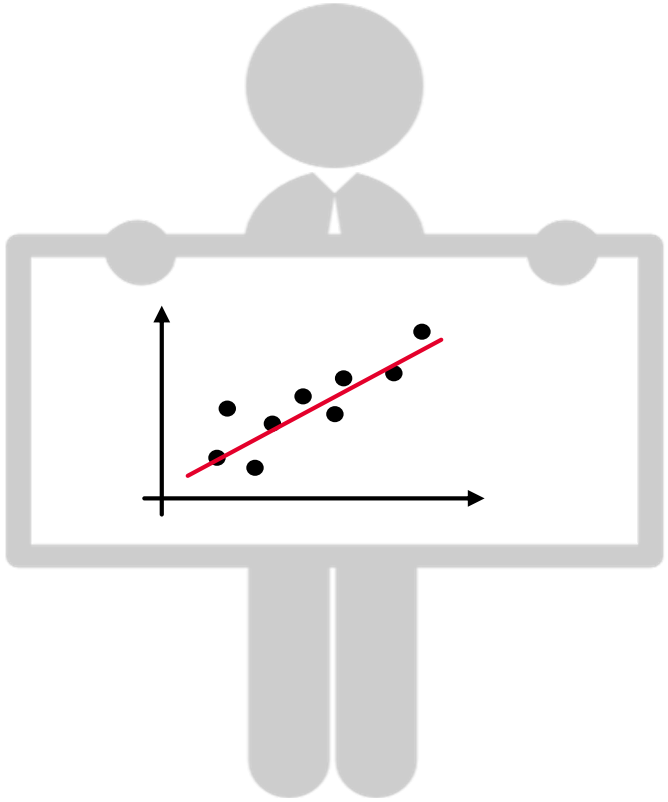
$y = a_0 + a_1 * x_1 + a_2 * x_2 + .. a_k * x_k$

nicht lineare Regression:

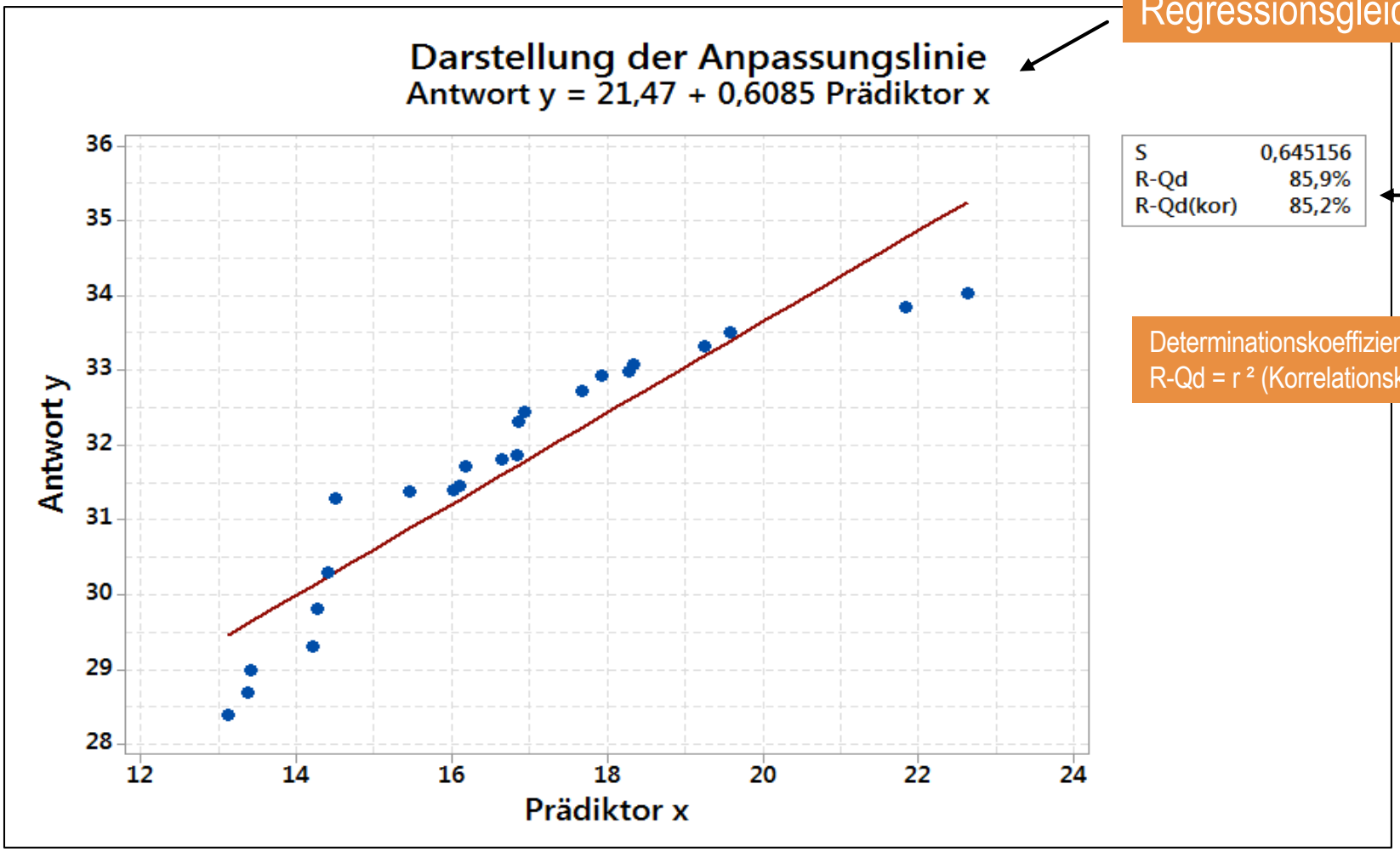
Polynomial, Exponentiell, Logarithmisch ...

Binäre logistische Regression:

Y ist diskret (gut/schlecht) und x_i sind stetig!



Einfache lineare Regression – Anpassungslinie



Regressionsgleichung

S	0,645156
R-Qd	85,9%
R-Qd(kor)	85,2%

Determinationskoeffizient (Bestimmtheitsmaß)
R-Qd = r ² (Korrelationskoeffizient ²)

16. Analyse – Grundlagen der Hypothesentests

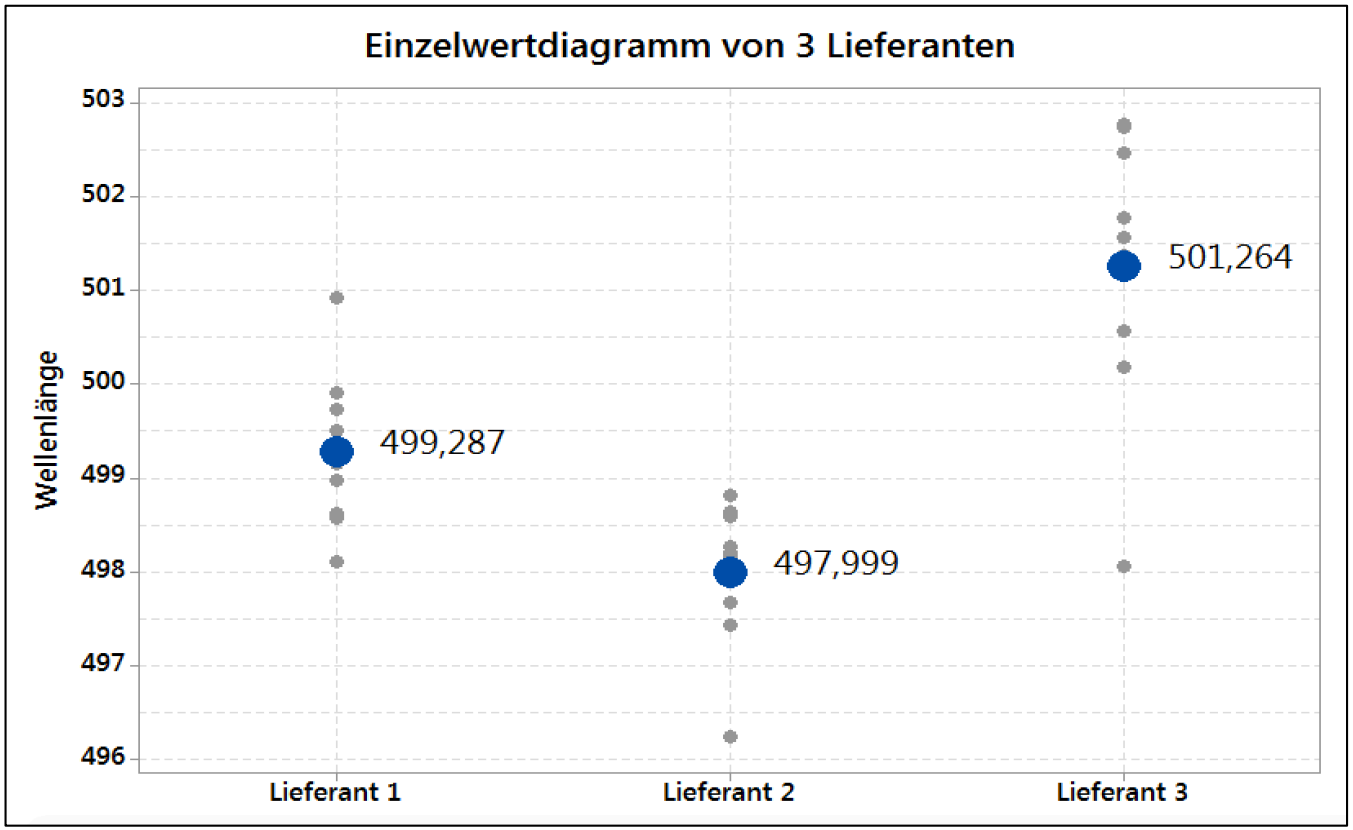
Signifikante Unterschiede erkennen und darstellen

Statistische vs. praktische Signifikanz

- Die Statistik beantwortet mit einer selbst zu wählenden Unsicherheit, ob Rückschlüsse (die man bspw. über technisch-physikalische Zusammenhänge vermutet) zulässig oder zu verwerfen sind.

Variable	N	Mittelwert
Lieferant 1	10	499,29
Lieferant 2	10	498,00
Lieferant 3	10	501,26

Produziert Lieferant 2 wirklich im Mittel die kürzesten Wellen und Lieferant 3 die längsten?



Parameter des Signifikanztests verstehen

Formulierung der statistischen Hypothesen

- Die statistische Nullhypothese H0 ist (fast immer) das Gegenteil der Arbeitshypothese.

Beispiel: Vergleich von Maschinen

- **Arbeitshypothese**:
  Maschine A hat eine andere Produktivität als Maschine B.
- **Statistische Nullhypothese H0**:
  Produktivität von Maschine A = Produktivität von Maschine B.

In der statistischen Gegenhypothese, Alternativhypothese oder kurz Alternative H1 steht das, was eigentlich gezeigt werden soll:

- **Statistitische Alternative H1:**
  Produktivität von Maschine A ≠ Produktivität von Maschine B

Fehler 1. und 2. Art: α und β Fehler

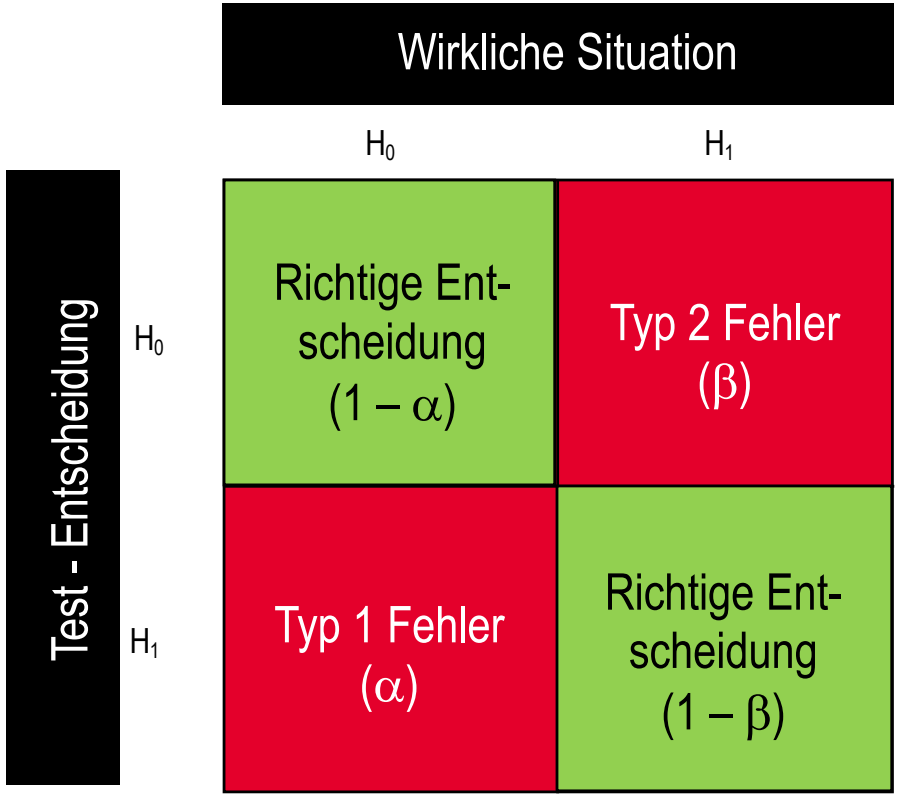
Es existieren zwei Möglichkeiten eine falsche Entscheidung beim Hypothesentest zu ziehen:

Typ 1 Fehler: (α)

- Die Nullhypothese wird verworfen, obwohl sie in Wirklichkeit zutrifft.

Typ 2 Fehler: (β)

- Die Nullhypothese wird nicht verworfen, obwohl sie in Wirklichkeit nicht zutrifft.



α Risiko – Bezeichnungen:
- Herstellerrisiko
- Fehler 1. Art
- Signifikanzniveau
- Irrtumswahrscheinlichkeit

β Risiko – Bezeichnungen:
- Abnehmerrisiko
- Fehler 2. Art

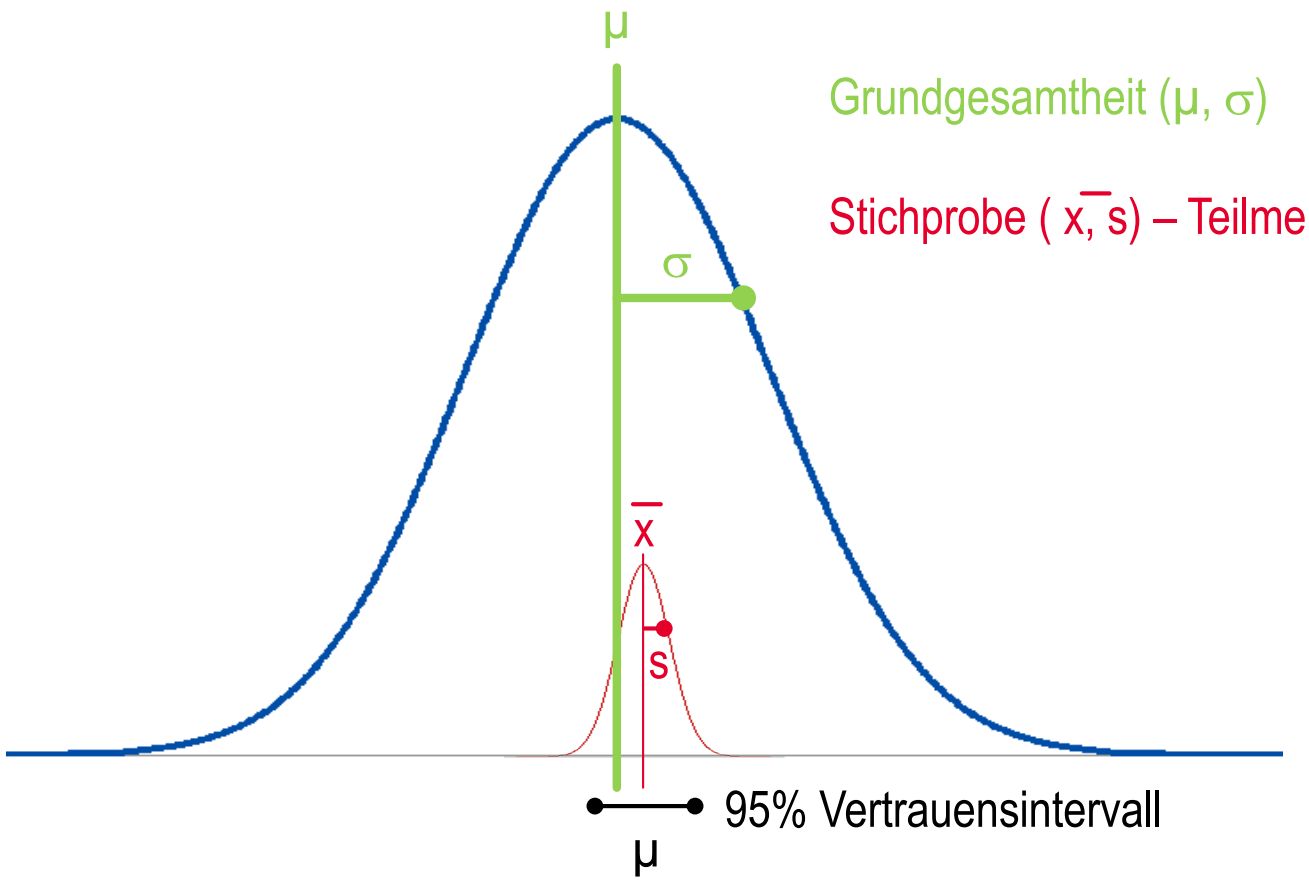
Beispiel: Fehler 1. und 2. Art: α und β Fehler

Gerichtsurteil:	**wahr** (Angeklagter ist wirklich unschuldig.)	**falsch** (Angeklagter ist wirklich schuldig.)
Angenommen (Das Gericht glaubt, der Angeklagte ist unschuldig.)	Freispruch gerechtfertigt (1-α)	Krimineller frei! falsche Entscheidung (β)
Abgelehnt (Das Gericht spricht den Angeklagten schuldig.)	Unschuldiger verurteilt **Absolut falsche Entscheidung** (α)	Schuldige korrekt bestraft (1-β)

α Risiko – Bezeichnungen:
- Herstellerrisiko
- Fehler 1. Art
- Signifikanzniveau
- Irrtumswahrscheinlichkeit

β Risiko – Bezeichnungen:
- Abnehmerrisiko
- Fehler 2. Art

Rückschlüsse aus einer Stichprobe ziehen



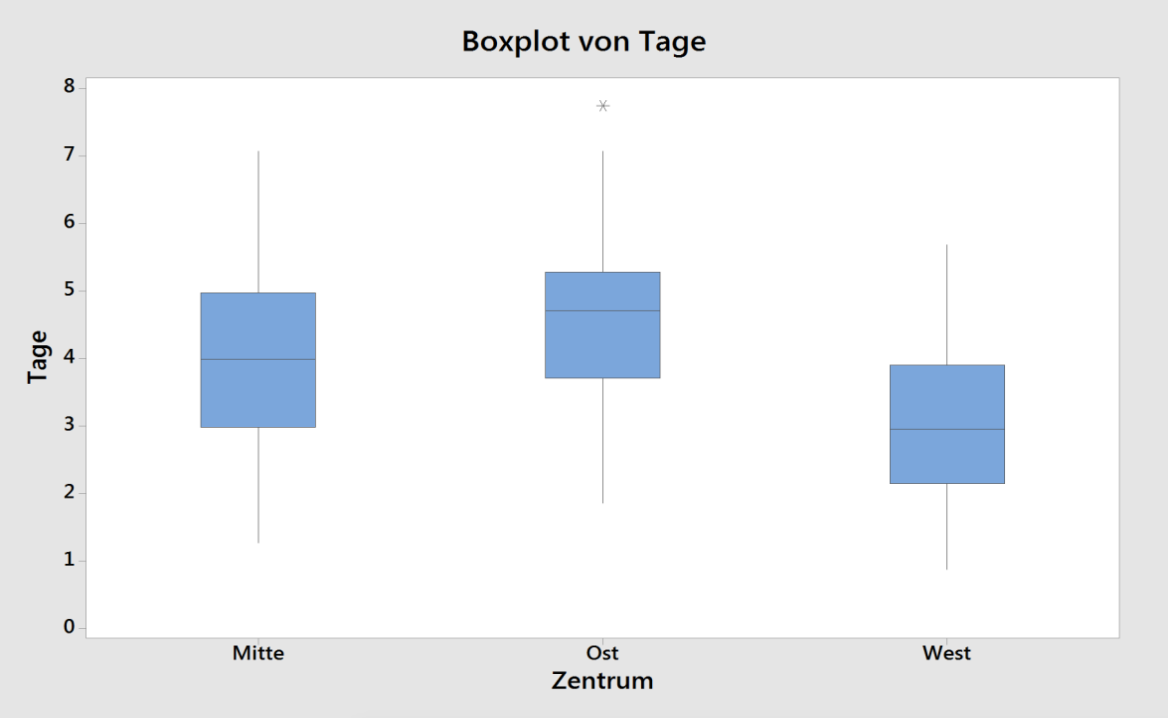
Grundgesamtheit (μ, σ)

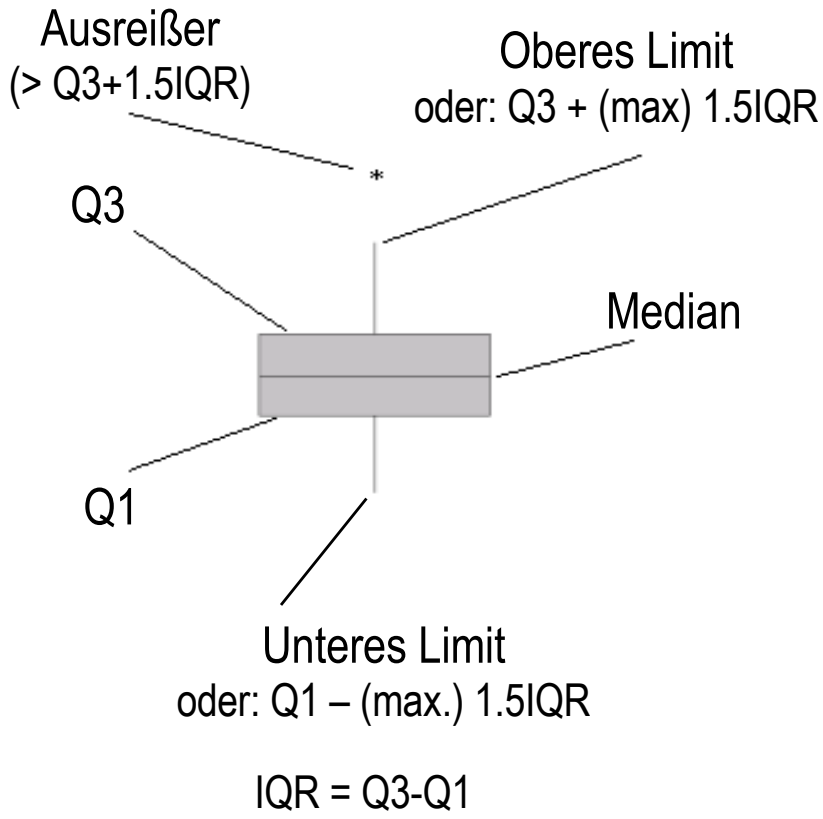
Stichprobe ($\bar{x}$, s) – Teilmenge der Grundgesamtheit

Werden Zufallsproben (Stichproben) aus einer Grundgesamtheit entnommen, so besitzen diese eine (berechenbare) Unschärfe (→ Vertrauensintervall).

- Sind einige Informationen der Grundgesamtheit bekannt, so reduziert sich diese Unschärfe. (z.B.: Verteilung, σ, ...)

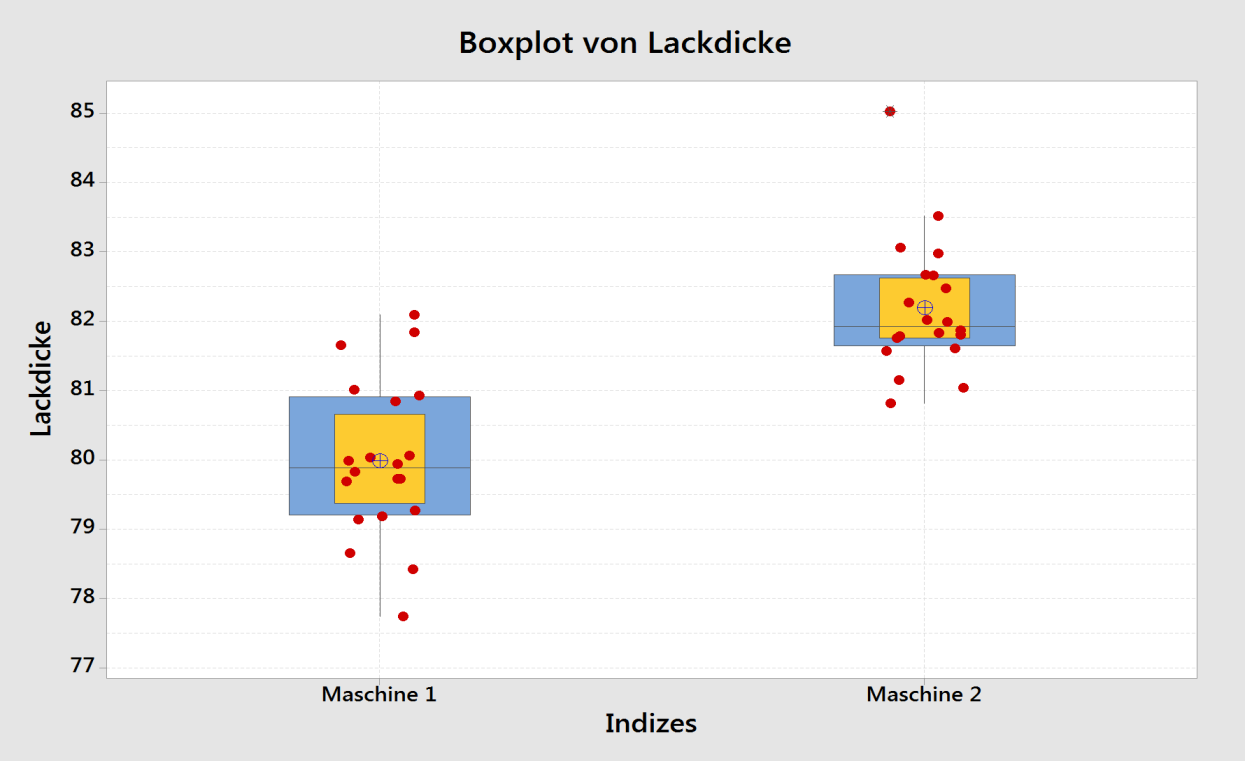
Boxplot (Box Whisker Plot)





- Der Boxplot zeigt (ideal bei mehreren Gruppen) die mittlere Lage und die Streuung an.
- Die mittlere Linie ist der Median, die Begrenzungen der Box das 1. bzw. 3 Quartil.

Boxplot mit Signifikanztest



Der oben dargestellte Boxplot beinhaltet die 95%-Konfidenzintervalle der beiden Mediane.

=> Die gelben 95% Konfidenzintervalle überlappen nicht, somit sind die Mediane der beiden Stichproben signifikant verschieden.