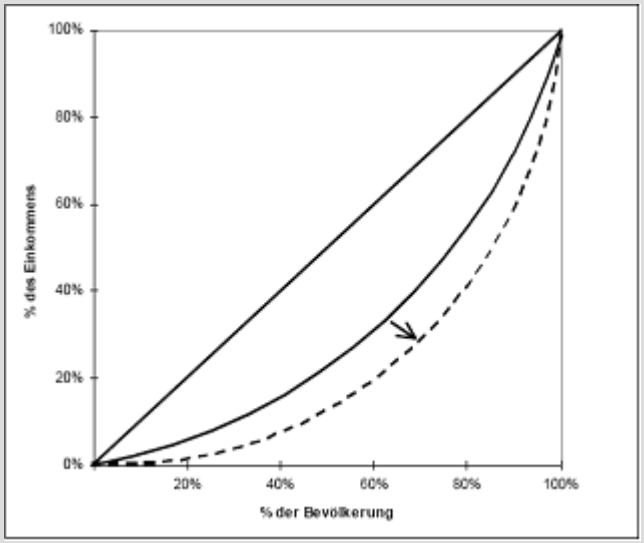
Statistics I

1. Einführung
2. Deskriptive Statistik
3. Korrelationen
4. Lorenzkurve und Konfidenzintervalle

Lorenzkurve

Die Lorenzkurve (benannt nach dem amerikanischen Statistiker und Ökonom Max Otto Lorenz) stellt statistische Verteilungen grafisch dar und veranschaulicht dabei die Ungleichheit. Sie analysiert die Konzentration eines Merkmals.

Je größer die Fläche zwischen der Diagonalen und der Linie ist, desto größer ist das Ungleichgewicht.



So verfügen in dem Beispiel die ärmsten 50% der Haushalte (Merkmalsträger) über 25% des Vermögens (Merkmalssumme), die ärmsten 80% verfügen über 60% des Vermögens.
Bei der gestrichelten Linie ist das Ungleichgewicht noch größer.

Quelle: Wikipedia

Beispiele für Anwendung der Lorenzkurve

Man analysiert, wie sich die Merkmalssumme auf die einzelnen Merkmalsträger der Variablen verteilt. Liegt eine ungleiche Verteilung der Merkmalssumme auf die Merkmalsträger vor, so spricht man von Konzentration des betreffenden Merkmals. Besonders häufig werden solche Konzentrationsphänomene bei Einkommen-, Umsatz- oder Vermögensverteilungen untersucht:

- *Einkommenskonzentration:*
  10 Prozent der Bevölkerung (Merkmalsträger) verdienen 60 Prozent des gesamten Einkommens (Merkmalssumme).
- *Umsatzkonzentration:*
  10 Prozent der umsatzstärksten Unternehmen (Merkmalsträger) erzielen 70 Prozent des Branchenumsatzes (Merkmalssumme).

Quelle: Wewel, Statistik im Bachelor-Studium der BWL und VWL, Pearson, S. 66

Bsp. Betriebsgröße („einfache“ Lorenzkurve)

Beispiel für eine einfache Lorenzkurve (Merkmalssumme bekannt)

Wichtig: Die Betriebe müssen anhand Mitarbeiterzahlen sortiert sein!

Betrieb	*Mitarbeiter = Merkmalssumme (MMSumme)*
1	*100*
2	*125*
3	*150*
4	*175*
5	*250*
6	*450*
7	*1.150*
8	*2.600*
Summe	*5.000*

Quelle: Hafner, Statistik für Sozial- und Wirtschaftswissenschaftler, Springer, Band 1, S. 51

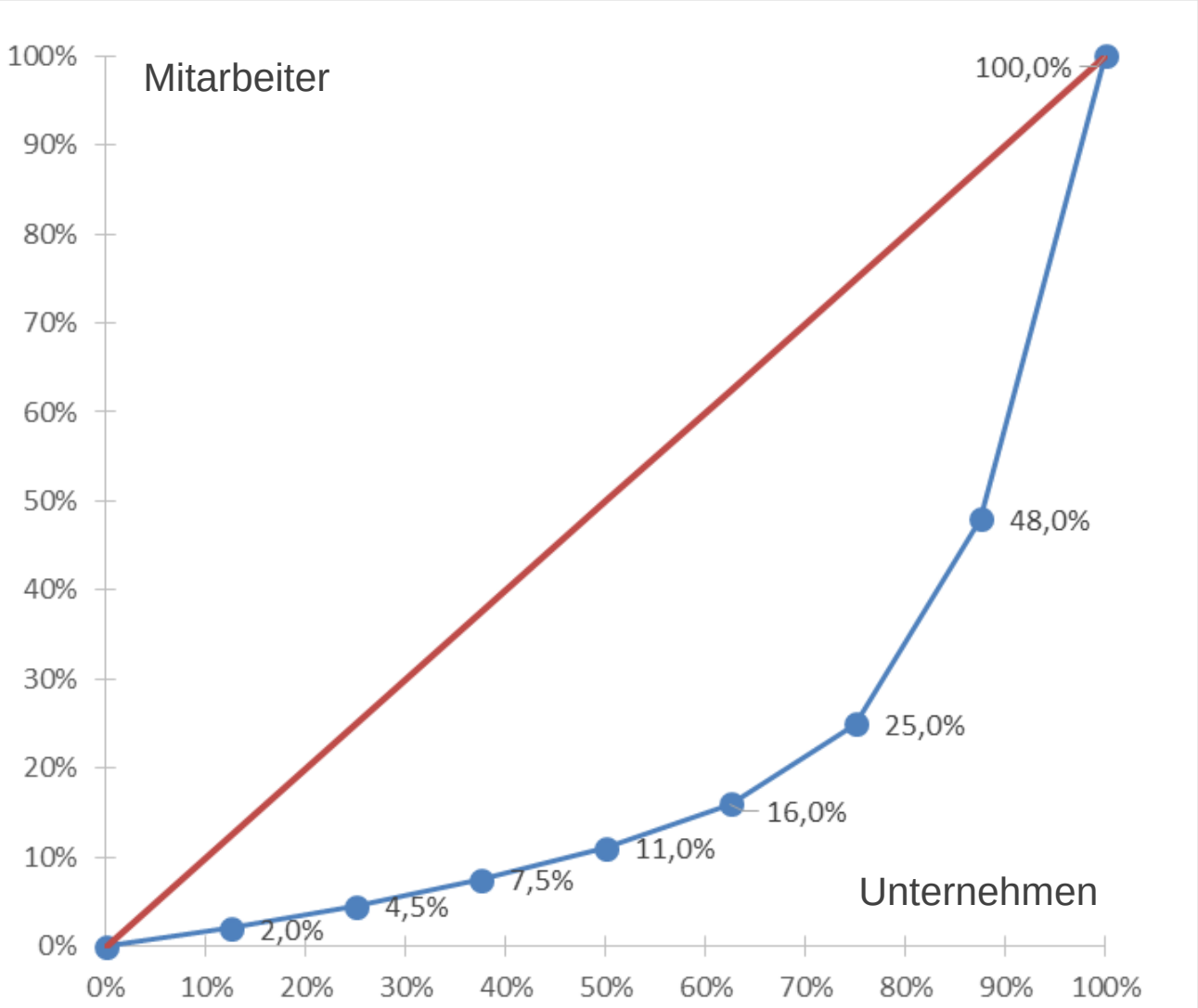
Bsp. Betriebsgröße („einfache“ Lorenzkurve)

Beispiel für eine einfache Lorenzkurve (Merkmalssumme bekannt)

Wichtig: Die Betriebe müssen anhand der Mitarbeiterzahlen sortiert sein!

Betrieb	MMTräger in %	MMTräger kumuliert	Mitarbeiter= MMSumme	MMSumme in %	MMSumme kumuliert
1	12,5	12,5	100	2,0	2,0
2	12,5	25,0	125	2,5	4,5
3	12,5	37,5	150	3,0	7,5
4	12,5	50,0	175	3,5	11,0
5	12,5	62,5	250	5,0	16,0
6	12,5	75,0	450	9,0	25,0
7	12,5	87,5	1.150	23,0	48,0
8	12,5	100,0	2.600	52,0	100,0
Summe	100,0		5.000	100,0	

Bsp. Betriebsgröße (Lorenzkurve)



Je größer die Fläche zwischen roter und blauer Linie, desto größer ist das Ungleichgewicht.

Auf die kleinsten 4 Betriebe (50% der Firmen) entfallen lediglich 11% der gesamten Mitarbeiter.

Bsp. Bruttogehalt (Lorenzkurve)

Beispiel für eine Lorenzkurve mit unbekannter Merkmalssumme

Bruttogehalt	*Anteil der Mitarbeiter = Merkmalsträger (MMTräger)*
[0;3.000[	*12%*
[3.000;5.000[	*12%*
[5.000;7.000[	*20%*
[7.000;10.000[	*24%*
[10.000;20.000[	*32%*

Quelle: Wewel, Statistik im Bachelor-Studium der BWL und VWL, Pearson, S. 67

Bsp. Bruttogehalt (Lorenzkurve)

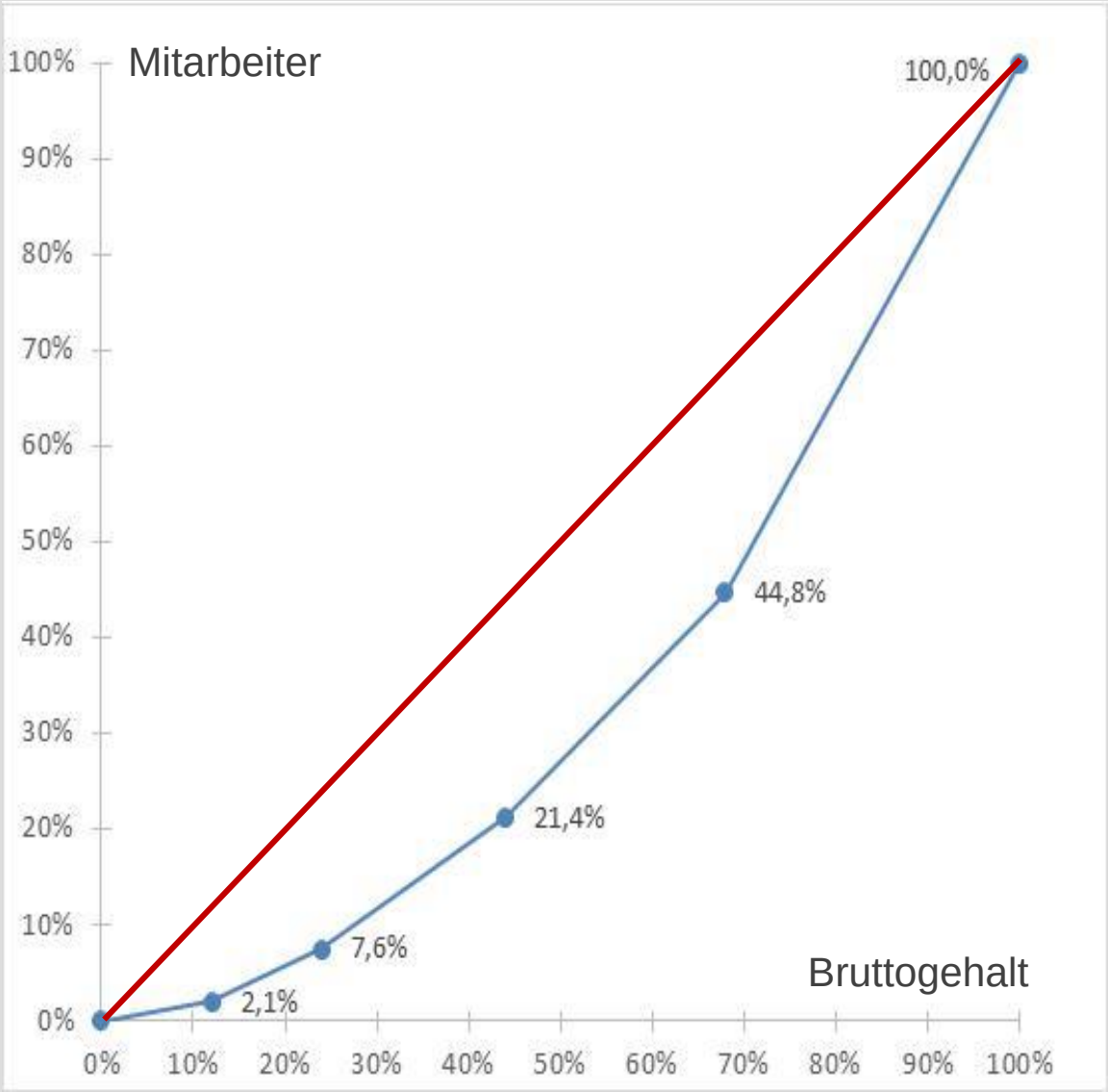
Beispiel für eine Lorenzkurve mit unbekannter Merkmalssumme

Brutto-gehalt	MMTräger in %	MMTräger kumuliert	Merkmals-summe *)	MMSumme in %	MMSumme kumuliert
[0;3.000[	12	12	180	2,1	2,1
[3.000;5.000[	12	24	480	5,5	7,6
[5.000;7.000[	20	44	1.200	13,8	21,4
[7.000;10.000[	24	68	2.040	23,4	44,8
[10.000;20.000[	32	100	4.800	55,2	100,0
Summe	100		8.700	100,0	

*) Merkmalssumme = Klassenmitte * Häufigkeit (z.B. 1.500 * 0,12 = 180)

Quelle: Wewel, Statistik im Bachelor-Studium der BWL und VWL, Pearson, S. 67

Bsp. Bruttogehalt (Lorenzkurve)



12% der MA verdienen nur 2,1% der Gehaltssumme bzw. 44% verdienen nur 21,4%.

Umgekehrt verdienen die restlichen 56% 78,4%.

Konfidenzintervalle

Konfidenzintervalle zählen zur analytischen Statistik, um von einer Stichprobe auf die Grundgesamtheit schließen zu können.

Analytische Statistik, auch schließende Statistik oder Interferenzstatistik beschäftigt sich damit, von Ergebnissen, die anhand von Daten gewonnen wurden, allgemein gültige Aussagen abzuleiten.

Schluss von Stichprobe auf Grundgesamtheit

Beispiel Körpergröße von Männern in einer Stadt.

Grundgesamtheit N = 12.193 erwachsene Männer, $\bar{x}$ = 175,61 cm

Stichprobengröße n	*Mittelwert $\bar{x}$*
100	*177,3 cm*
500	*175,2 cm*
1.000	*175,4 cm*
1.500	*175,5 cm*
2.000	*175,6 cm*

D.h. mit steigender Stichprobengröße nähert man sich dem Mittelwert der Grundgesamtheit.

Schluss von Stichprobe auf Grundgesamtheit

Um vom Mittelwert der Stichprobe auf die Grundgesamtheit schließen zu können, gibt man ein Konfidenzintervall (auch Vertrauensintervall genannt) an, innerhalb dessen sich der Mittelwert der Grundgesamtheit mit einer vorgegebenen Wahrscheinlichkeit bewegt.

Dabei können Konfidenzintervalle nicht nur für den Mittelwert, sondern auch für Standardabweichung und Anteile berechnet werden.

Man spricht bei Konfidenzintervallen auch von einem Bereichschätzer (im Gegensatz zum Punktschätzer). Es wird ein Bereich definiert, der mit einer Sicherheit von (1 – α) überdeckt wird.

Konfidenzintervall für Mittelwerte

Konfidenz-, Vertrauens- oder Sicherheitsintervall für Mittelwerte $\bar{x}$ zur Sicherheit $(1 - \alpha)$

$$[\ \underline{\mu},\ \bar{\mu}\ ] = \bar{x} \pm TINV(\alpha;\ n\text{-}1) * s / \sqrt{n}$$

$s_m = s / \sqrt{n}$ *Standardfehler des Mittelwerts*

EXCEL: TINV(α; n-1) = (1 – α/2)-Quantil der Student-Verteilung mit (n – 1) Freiheitsgraden = $t_{n-1;\ 1-\alpha/2}$

D.h. mit einer Sicherheit von (1 – α) Prozent überdeckt das Konfidenzintervall den Mittelwert der Grundgesamtheit.

Anders ausgedrückt: Mit einer Sicherheit von (1 – α) Prozent liegt der tatsächliche Mittelwert (der Grundgesamtheit) in diesem Intervall.

Beispiele: Konfidenzintervall für Mittelwerte

Konfidenz-, Vertrauens- oder Sicherheitsintervall für Mittelwerte $\bar{x}$ zur Sicherheit $(1 - \alpha)$

$$[\ \underline{\mu}, \overline{\mu}\ ] = \bar{x} \pm TINV(\alpha; n-1) * s / \sqrt{n}$$

Eine Umfrage unter 200 Personen ergab einen durchschnittlichen Intelligenzquotienten von $\bar{x} = 98{,}5$ mit einer Standardabweichung von $s = 17{,}1$.
In welchem Intervall rund um den ermittelten Mittelwert liegt der tatsächliche IQ in der Grundgesamtheit mit einer Sicherheit von 95%?

Untere Grenze $\underline{\mu} = 98{,}5 - TINV(0{,}05; 199) * 17{,}1/\sqrt{200} = 98{,}5 - 1{,}972 * 1{,}209 = 96{,}1$

Obere Grenze $\overline{\mu} = 98{,}5 + TINV(0{,}05; 199) * 17{,}1/\sqrt{200} = 98{,}5 + 1{,}972 * 1{,}209 = 100{,}9$

d.h. das Konfidenzintervall ist [96,1 ; 100,9]

Beispiel: Konfidenzintervall für Mittelwerte

Wie ändert sich das Konfidenzintervall, wenn man eine höhere Sicherheit von 99% erreichen möchte?

*Untere Grenze = 98,5 – TINV(0,01; 199) * 17,1/ √200 = 98,5 – 2,601 * 1,209 = 95,4*
*Obere Grenze = 98,5 + TINV(0,01; 199) * 17,1/ √200 = 98,5 + 2,601 * 1,209 = 101,6*
d.h. das Konfidenzintervall wird größer (breiter)

Wie ändert sich das Konfidenzintervall, wenn die Stichprobe von 200 auf 500 steigt (bei gleicher Sicherheit von 99% und gleicher Standardabweichung)?

*Untere Grenze = 98,5 – TINV(0,01; 499) * 17,1/ √500 = 98,5 – 2,576 * 0,765 = 96,5*
*Obere Grenze = 98,5 + TINV(0,01; 499) * 17,1/ √500 = 98,5 + 2,576 * 0,765 = 100,5*
d.h. das Konfidenzintervall wird kleiner

Konfidenzintervall für Anteile

Konfidenzintervall für Anteile p zur Sicherheit (1 – α):

$$[\underline{p}, \overline{p}] = p \pm NORMINV(1-\alpha/2; 0; 1) * \sqrt{p(1-p)/n}$$

EXCEL: NORMINV(1-α/2; 0; 1) = (1 – α/2)-Quantil der Standardnormalverteilung = $z_{1-\alpha/2}$

D.h. mit einer Sicherheit von (1 – α) Prozent überdeckt das Konfidenzintervall den festgestellten Anteil in der Grundgesamtheit.

Bzw. mit einer Sicherheit von (1 – α) Prozent liegt der tatsächliche Anteil (der Grundgesamtheit) in diesem Intervall.

Beispiel: Konfidenzintervall für Anteile

Konfidenzintervall für Anteile p zur Sicherheit (1 – α):

$$[\underline{p}, \overline{p}] = p \pm NORMINV(1-\alpha/2; 0; 1) * \sqrt{p(1-p)/n}$$

EXCEL: NORMINV(1-α/2; 0; 1) = (1 – α/2)-Quantil der Standardnormalverteilung = $z_{1-\alpha/2}$

Das Ergebnis einer Umfrage mit 2.000 Personen ergab, dass 910, also 45,5% die Partei XPÖ wählen wollen. In welchem Intervall liegt der tatsächliche Anteil in der Grundgesamtheit mit einer Sicherheit von 95 Prozent?

$$\text{Untere Grenze } \underline{p} = 0{,}455 - NORMINV(0{,}975; 0; 1) * \sqrt{0{,}455 * (1 - 0{,}455) / 2000} =$$
$$0{,}455 - 1{,}960 * 0{,}011 = 0{,}433$$

$$\text{Obere Grenze } \overline{p} = 0{,}455 + NORMINV(0{,}975; 0; 1) * \sqrt{0{,}455*(1 - 0{,}455) / 2000} =$$
$$0{,}455 + 1{,}960 * 0{,}011 = 0{,}477$$

Beispiel: Konfidenzintervall für Anteile

Wie ändert sich das Konfidenzintervall, wenn man eine höhere Sicherheit von 99% erreichen möchte?

$$\text{Untere Grenze} = 0{,}455 - \text{NORMINV}(0{,}995;\ 0;\ 1) * \sqrt{0{,}455 * (1 - 0{,}455) / 2000} = 0{,}455 - 2{,}576 * 0{,}011 = 0{,}426$$

$$\text{Obere Grenze} = 0{,}455 + \text{NORMINV}(0{,}995;\ 0;\ 1) * \sqrt{0{,}455 * (1 - 0{,}455) / 2000} = 0{,}455 + 2{,}576 * 0{,}011 = 0{,}484$$

d.h. das Konfidenzintervall wird größer (breiter)

Wie ändert sich das Konfidenzintervall, wenn die wenn die Stichprobe von 2.000 auf 500 sinkt (bei gleicher Sicherheit von 99%)?

$$\text{Untere Grenze} = 0{,}455 - \text{NORMINV}(0{,}995;\ 0;\ 1) * \sqrt{0{,}455 * (1 - 0{,}455) / 500} = 0{,}455 - 2{,}576 * 0{,}022 = 0{,}398$$

$$\text{Obere Grenze} = 0{,}455 + \text{NORMINV}(0{,}995;\ 0;\ 1) * \sqrt{0{,}455 * (1 - 0{,}455) / 500} = 0{,}455 + 2{,}576 * 0{,}022 = 0{,}512$$

d.h. das Konfidenzintervall wird breiter

Ü Akademikeranteil in Österreich

Nach den Ergebnissen des Mikrozensus (n = 35.200) aus dem Jahr 2003 haben 10,2% der Österreicher über 15 Jahre einen Hochschulabschluss.

Berechnen Sie das 99%-Konfidenzintervall für den tatsächlichen Anteil in der Gesamtbevölkerung.

Ü Wählerbefragung

Sie lesen in der Zeitung, dass bei einer Umfrage von 500 Personen die Partei XPÖ von 38 Prozent der Wähler gewählt werden wird, die Oppositionspartei aber nur von 31 Prozent.

Was halten Sie von derartigen Aussagen?

Ü Kinderzahl

Eine Umfrage von 1.000 repräsentativ ausgewählten Frauen hat ergeben, dass diese im Durchschnitt 1,35 Kinder geboren haben (Standardabweichung = 1,0).

Wie groß ist die durchschnittliche Kinderzahl in der Grundgesamtheit in einem Sicherheitsintervall von 95 Prozent?