Fachhochschule Südwestfalen
University of Applied Sciences

Statistik

CH.10 - Schätzung

SS 2021 || Prof. Dr. Buchwitz, Sommer, Henke

Wir geben Impulse

Outline

Bitte evaluieren Sie den Kurs!

http://evasys.fh-swf.de/evasys/online.php?pswd=64D4W

Outline

Von der Stichprobe zur Grundgesamtheit

- Bei statistischen Untersuchungen ist es häufig nicht möglich (oder zu aufwändig und teuer) Informationen über alle Elemente der Grundgesamtheit zu erheben.
- Daher wwerden die gewünschten Informationen bei einer Teilmenge der Grundgesamtheit, einer **Stichprobe** erholeb.
- **Ziel** ist es dann, mit Hilfe der statistischen Parameter der *Stichprobe* (Mittelwert, Varianz, Verteilung) auf die statistischen Parameter der Grundgesamtheit zu **schließen**.
- Parameter der Grundgesamtheit: μ,σ,π
- Parameter der Stichprobe: $\bar{x}$, s^2, p

Von der Stichprobe zur Grundgesamtheit

GG: μ, σ, π, N **SP:** $\bar{x}$, s^2, p, n

Maßzahlen für Grundgesamtheit und Stichprobe

Erwartungswert

$$\bar{x} = \frac{1}{n} \sum_{i=1}^{n} x_i \quad \text{oder} \quad \mu = \frac{1}{N} \sum_{i=1}^{N} x_i$$

Varianz

$$s^2 = \frac{1}{n-1} \sum_{i=1}^{n} (x_i - \bar{x})^2 \quad \text{oder} \quad \sigma^2 = \frac{1}{N} \sum_{i=1}^{N} (x_i - \bar{x})^2$$

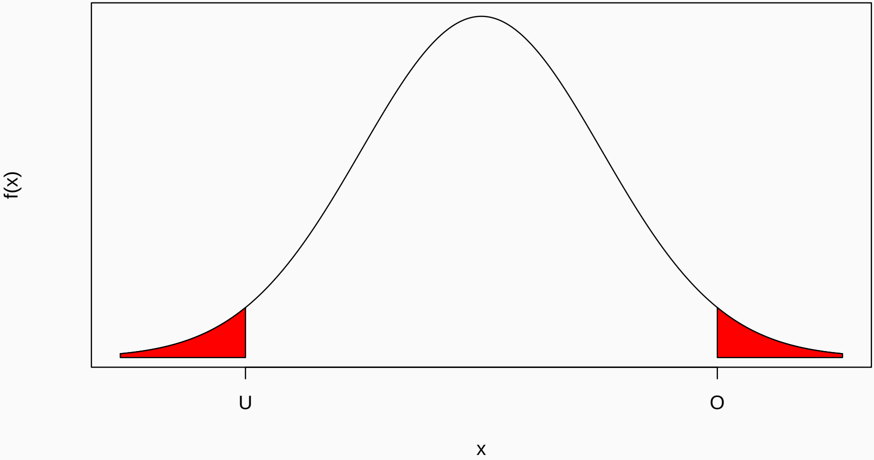
Stichprobenauswahl

- Die korrekte Auswahl einer Stichprobe ist Gegenstand der **Stichprobentheorie**.
- Eine reine Zufallsauswahl liefert eine **repräsentative** Stichprobe.
- **Representativität** kann mit **Strukturgleichheit** übersetzt werden.
- Sofern eine echte Zufallsstichprobe nicht möglich ist wird häufig auf einen Stichprobenplan und eine damit verbundene Quotenauswahl zurückgegegriffen.

Punkt- und Intervallschätzung

- Bei Schätzungen wird zwischen Pukt- und Intervallschätzungen unterschieden. **Punktschätzer** sind ein spezifischer Wert
  - Als Punktschätzer für den Mittelwert der Grundgesamtheit μ wird der Mittelwert der Stichprobe $\bar{x}$ verwendet.
  - Als Punktschätzer für die Varianz der Grundgesamtheit σ^2 wird s^2 der Stichprobe verwendet.
  - Als Punktschätzer für einen Anteilswert π der Grundgesamtheit wird der prozentuale Anteil p in der Stichprobe verwendet.

Intervallschätzer



Intervallschätzer

- Intervallschätzer geben einen Bereich an in dem der wahre aber unbekannte (und damit zu schätzende) Parameter mit einer bestimmten Wahrscheinlichkeit liegt.
- Die Wahrscheinlichkeit, dass der wahre aber unbekannte Parameter zwischen Untergrenze U und Obergrenze O liegt bezeichnen wir als α, also: $P(U \leq x \leq O) = 1 - \alpha$
- Am rechten und linken Rand der Verteilung (eingezeichneter Bereich) liegt jeweils $\alpha/2$ der Wahrscheinlichkeitsmasse, also $P(X \leq U) = \alpha/2$ und $P(X > O) = \alpha/2$.

Intervallschätzer werden (fast) immer symmetrisch um Punktschätzer konstruiert.

Konfidenzintervalle

Im weiteren Verlauf diskutieren wir die folgenden Konfidenzintervalle, die hinsichtlich der aufgelisteten Fallunterscheidungen differenziert werden können. Vor der Berechnung muss zunächst geklärt werden, welche Art von Konfidenzintervall gefragt ist:

1) Konfidenzintervalle für Mittelwerte
   - Anwendungsfall mit **bekannter Varianz der Grundgesamtheit** (σ^2 bekannt)
   - Anwendungsfall mit **un**bekannter Varianz der Grundgesamtheit (Schätzung mittels s^2)
2) Konfidenzintervalle für Anteilswerte

Konfidenzintervalle für den Mittelwert

- Wenn die Grundgesamtheit einer Normalverteilung folgt, dann sind die Mittelwerte von Stichproben mit Umfang $n << N$ ebenfalls normalverteilt. Es gilt $\bar{x} \sim N(\mu, \frac{\sigma^2}{n})$
- Bei großen Stichproben ($n \geq 30$) sind die Mittelwerte näherungsweise normalverteilt.

Konfidenzintervalle für den Mittelwert bei bekannter Varianz

- **Voraussetzung:** Normalverteilte Grundgesamtheit oder große Stichprobe ($n > 30$).
- **Fall:** Varianz der Grundgesamtheit σ^2 bekannt.
- Das zweiseitige, symmetrische Konfidenzintervall $[\underline{\mu}, \overline{\mu}]$ zum Konfidenzniveau $1 - \alpha$ ist gegeben durch:

$$\underline{\mu}, \overline{\mu} = \bar{x} \pm z_{1-\frac{\alpha}{2}} \cdot \frac{\sigma}{\sqrt{n}}$$

Beispiel

Aufgabe: Das Ergebnis einer Studie mit 100 Ganztagsschülern besagt, dass diese durchschnittlich 52 h/Woche in der Schule verbringen. Aus früheren Untersuchungen ist bekannt, dass die Standardabweichung der wöchentlichen Aufenthaltszeit in der Schule 12 Stunden beträgt.

- Geben Sie ein 95% Konfidenzintervall für die durchschnittliche wöchentliche Aufenthaltszeit in der Ganztagsschule an.
- **Hinweis:** Eine Verteilungstabelle befindet sich auf der nächsten Seite.

Verteilungsfunktion $F(z)$ für $z \sim N(\mu = 0, \sigma^2 = 1)$

Verteilungsfunktion $F(z)$ **der Standardnormalverteilung** $N(0, 1)$
Beispiel: $F(z) = P(z \leq 1.96) = 0.9750$

z	0	0.01	0.02	0.03	0.04	0.05	0.06	0.07	0.08	0.09
0.0	0.5000	0.5040	0.5080	0.5120	0.5160	0.5199	0.5239	0.5279	0.5319	0.5359
0.1	0.5398	0.5438	0.5478	0.5517	0.5557	0.5596	0.5636	0.5675	0.5714	0.5753
0.2	0.5793	0.5832	0.5871	0.5910	0.5948	0.5987	0.6026	0.6064	0.6103	0.6141
0.3	0.6179	0.6217	0.6255	0.6293	0.6331	0.6368	0.6406	0.6443	0.6480	0.6517
0.4	0.6554	0.6591	0.6628	0.6664	0.6700	0.6736	0.6772	0.6808	0.6844	0.6879
0.5	0.6915	0.6950	0.6985	0.7019	0.7054	0.7088	0.7123	0.7157	0.7190	0.7224
0.6	0.7257	0.7291	0.7324	0.7357	0.7389	0.7422	0.7454	0.7486	0.7517	0.7549
0.7	0.7580	0.7611	0.7642	0.7673	0.7704	0.7734	0.7764	0.7794	0.7823	0.7852
0.8	0.7881	0.7910	0.7939	0.7967	0.7995	0.8023	0.8051	0.8078	0.8106	0.8133
0.9	0.8159	0.8186	0.8212	0.8238	0.8264	0.8289	0.8315	0.8340	0.8365	0.8389
1.0	0.8413	0.8438	0.8461	0.8485	0.8508	0.8531	0.8554	0.8577	0.8599	0.8621
1.1	0.8643	0.8665	0.8686	0.8708	0.8729	0.8749	0.8770	0.8790	0.8810	0.8830
1.2	0.8849	0.8869	0.8888	0.8907	0.8925	0.8944	0.8962	0.8980	0.8997	0.9015
1.3	0.9032	0.9049	0.9066	0.9082	0.9099	0.9115	0.9131	0.9147	0.9162	0.9177
1.4	0.9192	0.9207	0.9222	0.9236	0.9251	0.9265	0.9279	0.9292	0.9306	0.9319
1.5	0.9332	0.9345	0.9357	0.9370	0.9382	0.9394	0.9406	0.9418	0.9429	0.9441
1.6	0.9452	0.9463	0.9474	0.9484	0.9495	0.9505	0.9515	0.9525	0.9535	0.9545
1.7	0.9554	0.9564	0.9573	0.9582	0.9591	0.9599	0.9608	0.9616	0.9625	0.9633
1.8	0.9641	0.9649	0.9656	0.9664	0.9671	0.9678	0.9686	0.9693	0.9699	0.9706
1.9	0.9713	0.9719	0.9726	0.9732	0.9738	0.9744	0.9750	0.9756	0.9761	0.9767
2.0	0.9772	0.9778	0.9783	0.9788	0.9793	0.9798	0.9803	0.9808	0.9812	0.9817
2.1	0.9821	0.9826	0.9830	0.9834	0.9838	0.9842	0.9846	0.9850	0.9854	0.9857
2.2	0.9861	0.9864	0.9868	0.9871	0.9875	0.9878	0.9881	0.9884	0.9887	0.9890
2.3	0.9893	0.9896	0.9898	0.9901	0.9904	0.9906	0.9909	0.9911	0.9913	0.9916

Konfidenzintervalle für den Mittelwert bei unbekannter Varianz

- **Voraussetzung:** Normalverteilte Grundgesamtheit oder große Stichprobe ($n > 30$).
- **Fall:** Varianz der Grundgesamtheit unbekannt.
- Das zweiseitige, symmetrische Konfidenzintervall $[\underline{\mu}, \overline{\mu}]$ zum Konfidenzniveau $1 - \alpha$ ist gegeben durch:

$$\underline{\mu}, \overline{\mu} = \bar{x} \pm t_{n-1,\, 1-\frac{\alpha}{2}} \cdot \frac{s}{\sqrt{n}}$$

Beispiel

Aufgabe: Nehmen Sie an, dass die Varianz der wöchtenlichen Aufenthaltszeit in der Ganztags der vorherigen Studie ($n = 100, \bar{x} = 52$) auf Basis der Stichprobe mit 144 $(h/Woche)^2$ geschätzt wurde.

- Geben Sie ein 95% Konfidenzintervall für die durchschnittliche wöchentliche Aufenthaltszeit in der Ganztagsschule an.
- **Hinweis:** Eine Verteilungstabelle befindet sich auf der nächsten Seite.

Quantile der $t_{n;\,\gamma}$ Verteilung

Quantile $t_{n;\,\gamma}$ der t_n-Verteilung

Beispiel: $P(t_{10} \leq 0.6998) = 0.750$

$n \setminus \gamma$	0.995	0.99	0.975	0.95	0.9	0.75	0.5
1	63.6567	31.8205	12.7062	6.3138	3.0777	1.0000	0
2	9.9248	6.9646	4.3027	2.9200	1.8856	0.8165	0
3	5.8409	4.5407	3.1824	2.3534	1.6377	0.7649	0
4	4.6041	3.7469	2.7764	2.1318	1.5332	0.7407	0
5	4.0321	3.3649	2.5706	2.0150	1.4759	0.7267	0
6	3.7074	3.1427	2.4469	1.9432	1.4398	0.7176	0
7	3.4995	2.9980	2.3646	1.8946	1.4149	0.7111	0
8	3.3554	2.8965	2.3060	1.8595	1.3968	0.7064	0
9	3.2498	2.8214	2.2622	1.8331	1.3830	0.7027	0
10	3.1693	2.7638	2.2281	1.8125	1.3722	0.6998	0
20	2.8453	2.5280	2.0860	1.7247	1.3253	0.6870	0
30	2.7500	2.4573	2.0423	1.6973	1.3104	0.6828	0
40	2.7045	2.4233	2.0211	1.6839	1.3031	0.6807	0
50	2.6778	2.4033	2.0086	1.6759	1.2987	0.6794	0
60	2.6603	2.3901	2.0003	1.6706	1.2958	0.6786	0
70	2.6479	2.3808	1.9944	1.6669	1.2938	0.6780	0
80	2.6387	2.3739	1.9901	1.6641	1.2922	0.6776	0
90	2.6316	2.3685	1.9867	1.6620	1.2910	0.6772	0
100	2.6259	2.3642	1.9840	1.6602	1.2901	0.6770	0
150	2.6090	2.3515	1.9759	1.6551	1.2872	0.6761	0
200	2.6006	2.3451	1.9719	1.6525	1.2858	0.6757	0
250	2.5956	2.3414	1.9695	1.6510	1.2849	0.6755	0
300	2.5923	2.3388	1.9679	1.6499	1.2844	0.6753	0
400	2.5882	2.3357	1.9659	1.6487	1.2837	0.6751	0

Konfidenzintervalle für Anteilswerte

- **Voraussetzung:** $n \cdot p \geq 5$ und $n \cdot (1 - p) \geq 5$
- Das zweiseitige, symmetrische Konfidenzintervall $[\underline{\pi}, \overline{\pi}]$ zum Konfidenzniveau $1 - \alpha$ ist gegeben durch:

$$\underline{\pi}, \overline{\pi} = p \pm z_{1-\frac{\alpha}{2}} \cdot \sqrt{\frac{p \cdot (1 - p)}{n}}$$

Interpretation von Konfidenzintervallen

- Konfidenzintervalle geben die Überdeckungswahrscheinlichkeit mit dem wahren aber unbekannten Parameter der Grundgesamtheit an.

Outline

Beispiel: Eine Reifenfirma wirbt damit, dass die durchschnittliche Lebensdauer einer angebotenen Reifenmarke 100 000 km übersteigt.

Wie prüft man das?

Beispiel: Eine Reifenfirma wirbt damit, dass die durchschnittliche Lebensdauer einer angebotenen Reifenmarke 100 000 km übersteigt.

Wie prüft man das?

Überlegungen zu Ergebnissen aus einer Stichprobe:

- $\bar{x} \geq 100000km$ $\rightarrow$ Der Werbung kann man glauben.
- $\bar{x} = 99500km$ $\rightarrow$ Nah genug um der Werbung noch zu glauben.
- $\bar{x} = 60000km$ $\rightarrow$ Zweifel an der Aussage.

Das Testen von Hypothesen ist eine Systematisierung des diskutierten Vorgehens.

Aufabu von Hypothesentests

Beim Testen von Hypothesen werden **zwei** Hypothesen aufgestellt.

Hypothesentests sind fast immer *konservativ* aufgebaut:

- Die **Nullhypothese** H_0, ist die Hypothese die man (üblicherweise) verwerfen möchte.
- Die **Alternativhypothese** H_A (oder H_1), bildet das Gegenteil der Nullhypothese ab und beinhaltet die Aussage, die man zeigen möchte.

Beispiele:

a) $H_0 : \mu = 25$ vs. $H_A : \mu \neq 25$
b) $H_0 : \mu \leq 25$ vs. $H_1 : \mu > 25$
c) $H_0 : \mu \geq 25$ vs. $H_1 : \mu < 25$
d) H_0 : Medikament wirkt gegen Krankheit vs. H_1 : Medikament wird nicht gegen Krankheit

Fehler 1. Art (α) und Fehler 2. Art (β)

	H_0 beibehalten	H_0 verwerfen
H_0 wahr	korrekt	Fehler 1. Art
	$P = 1 - \alpha$	$P = \alpha$
H_A wahr	Fehler 2. Art	korrekt
	$P = \beta$	$P(1 - \beta)$

Zusammenhang Fehlerarten

	Angeklagten freisprechen	Angeklagten verurteilen
Angeklagter unschuldig	korrekt $P = 1 - \alpha$	Fehler 1. Art $P = \alpha$
Angeklagter schuldig	Fehler 2. Art $P = \beta$	korrekt $P(1 - \beta)$

- Wie können Sie den Fehler erster Art α minimieren, sodass $\alpha = 0$?
- Was passiert mit dem Fehler zweiter Art β?

Vorgehen

1) Formulieren Sie Nullhypothese H_0 und Alternativhypothese H_A.
2) Legen Sie ein Signifikanzniveau α fest.
3) Wählen Sie die passende Teststatistik.
4) Bestimmen Sie den Wert der Teststatistik, ab dem die Nullhypothese verworfen werden muss.
5) Bestimmen Sie den Vergleichswert aus den Stichprobendaten.
6) Entscheiden Sie durch Vergleich der Werte aus (iv) und (v), ob Sie die Nullhypothese verwerfen können.

Beispiel

- Ein Unternehmen erwägt die Einführung eines neuen Zahlungssystems.
- Der Business Case für das Projekt hat ergeben, dass sich die Einführung des Zahlungssystems nur dann lohnt, wenn der durchschnittliche Monatsumsatz pro Kunde mehr als 70 Euro beträgt. Eine Stichprobe von 200 Kunden ergab einen durchschnittlichen Umsatz von 74 Euro pro Kunde.
- Kann man unter der Annahme einer Normalverteilung mit $\sigma = 30$ Euro mit 95%-iger Sicherheit davon ausgehen, dass sich die Einführung des Zahlungssystems für das Unternehmen lohnt?

Hypothesentest für Mittelwerte bei bekannter Varianz

$H_0: \quad \mu \leq \mu_0$ vs. $H_A: \quad \mu > \mu_0$

Testvorschrift:

Verwerfe H_0, wenn gilt:

$$\bar{x} \geq \mu_0 + z_{1-\alpha} \cdot \frac{\sigma}{\sqrt{n}}$$

Hypothesentest für Mittelwerte bei bekannter Varianz

$H_0: \quad \mu \leq \mu_0$ vs. $H_A: \quad \mu > \mu_0$

Testvorschrift:

Verwerfe H_0, wenn gilt:

$$\bar{x} \geq \mu_0 + z_{1-\alpha} \cdot \frac{\sigma}{\sqrt{n}}$$

- Alternativ können Sie die Formel auch direkt umstellen:

$$\frac{\bar{x} - \mu_0}{\sigma/\sqrt{n}} \geq z_{1-\alpha}$$

Beispiel

Beispiel

Beispiel

Verständnisfragen

1. Beschreiben Sie den Unterschied zwischen Stichprobe und Grundgesamtheit? Wozu dienen Stichproben?
2. Erläutern Sie in eigenen Worten was ein Konfidenzintervall ist und wie diese interpretiert werden können.
3. Wovon ist die Breite eines Konfidenzintervalls abhängig?