Vorlesung 7b

Mittelwerte

1. Populationsmittelwert und Stichprobenmittelwert

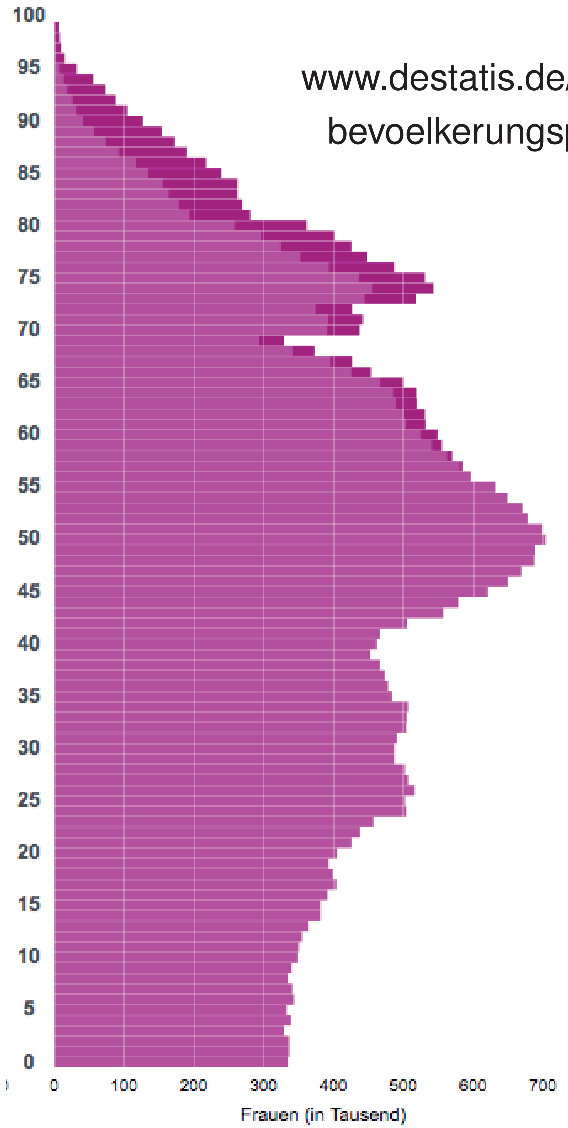
Denken wir an eine Liste (eine “Population”)

von reellen Daten

$$w_1, \ldots, w_g$$

z. B. die Lebensalter aller Frauen

in der deutschen Bevölkerung 2014



www.destatis.de/
bevoelkerungspyramide/

Angenommen man möchte den Populationsmittelwert

$$\mu := \frac{1}{g} \sum_{j=1}^{g} w_j.$$

schätzen,

und zwar aus den Werten einer

aus der Population gezogenen Stichprobe

$$x_1, \ldots, x_n$$

(sagen wir für $n = 100$).

Als *Schätzwert* für μ bietet sich an:

$$m := \frac{1}{n}(x_1 + \ldots + x_n)$$

Wie zuverlässig ist diese Schätzung?

Goldene Idee der Statistik:

Man fasst $x_1, \ldots, x_n$ auf als Ergebnis eines rein zufälligen Ziehens aus der Population:

$$X_1 := w_{J_1},\ X_2 := w_{J_2}, \ldots$$

mit $J_1, J_2, \ldots$ rein zufällige Wahl aus $\{1, \ldots, g\}$ (“Ziehen mit Zurücklegen”).

Wir setzen hier g als (sehr) groß gegenüber n voraus,

damit entstehen *auch*

beim n-maligen Ziehen *mit* Zurücklegen

Kollisionen nur mit (verschwindend) kleiner W'keit.

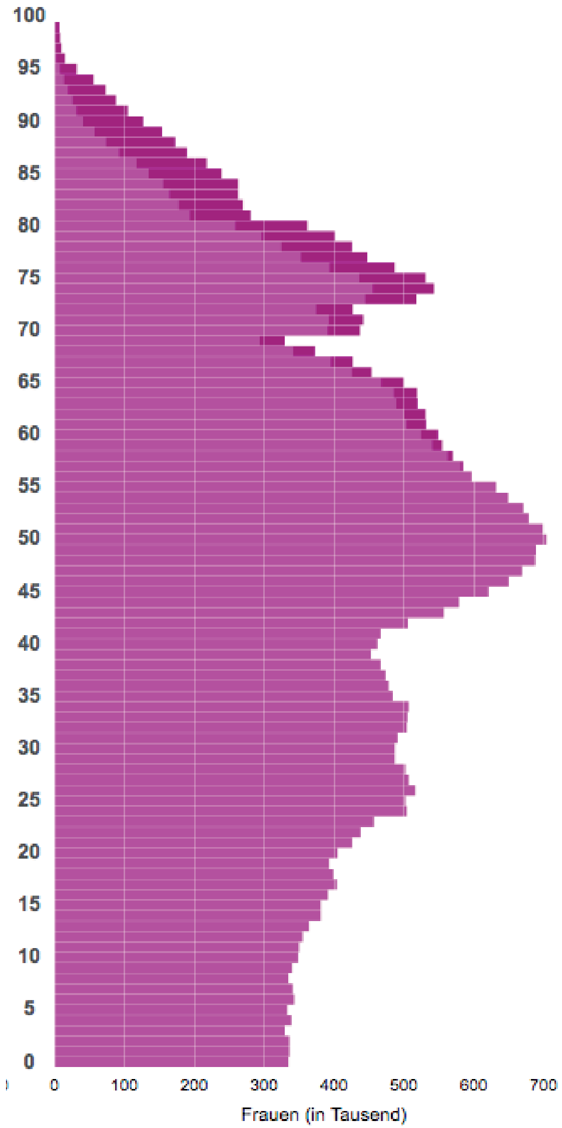
Das führt auf die Vorstellung:

$x_1, \ldots, x_n$ sind entstanden

durch n-maliges unabhängiges Ziehen $X_1, \ldots, X_n$

aus der Verteilung ρ auf $\mathbb{R}$

mit $\rho([c, d]) := \frac{1}{g}\#\{i : w_i \in [c, d]\}$


100
95
90
85
80
75
70
65
60
55
50
45
40
35
30
25
20
15
10
5
0
0
100
200
300
400
500
600
700
Frauen (in Tausend)

$$m = \frac{1}{n}(x_1 + \ldots + x_n)$$

fasst man also auf

als *eine* Realisierung (*einen* Ausgang)

der **Zufallsvariable**

$$M := \bar{X} := \frac{1}{n}(X_1 + \ldots + X_n)$$

(Mittelwert der zufälligen Stichprobe $(X_1, \ldots, X_n)$)

Es gilt jedenfalls:

$$\mathbf{E}[M] = \mu$$

Der Erwartungswert des Stichprobenmittelwertes

ist gleich dem Populationsmittelwert.

2. Populationsvarianz
und Varianz des Stichprobenmittelwertes

Der Populationsmittelwert war gleich $\mathbf{E}[X_1] = \mu$.

Wir haben

$$\mathbf{Var}[X_1] = \frac{1}{g}\sum_{j=1}^{g}(w_j - \mu)^2 =: \sigma^2$$

Diese Zahl σ^2 heißt auch die **Populationsvarianz**.

$$M = \frac{1}{n}(X_1 + \ldots + X_n)$$

$$\mathbf{E}[M] = \mu$$

$$M = \frac{1}{n}(X_1 + \ldots + X_n)$$

$$\mathbf{Var}[M] = ?$$

Wird mit Zurücklegen gezogen, dann sind die X_i unabhängig, und es ergibt sich

$$\mathbf{Var}[M] = \frac{\sigma^2}{n}$$

Wird ohne Zurücklegen gezogen
und ist die Populationsgröße g nicht sehr groß
gegenüber der Stichprobengröße n, dann hat es Sinn,
die *Korrektur für endliche Populationen* zu berücksichtigen
(vgl. Aufgabe 23):

$$\mathbf{Var}[M] = \frac{\sigma^2}{n} \cdot \frac{g-n}{g-1}$$

Diese Korrektur werden wir
für den Rest dieser Vorlesung vernachlässigen (wir denken
an großes g, bzw. – wie schon gesagt – an ein wiederholtes
unabhängiges Ziehen aus einer Verteilung).

Dann gilt:

$$\mathbf{Var}[M] = \frac{\sigma^2}{n};$$

die Standardabweichung des Stichprobenmittelwertes M ist also $\frac{\sigma}{\sqrt{n}}$.

3. Approximative Verteilung des Stichprobenmittelwertes

Wie ist (für nicht zu kleines n)

der Stichprobenmittelwert M verteilt?

Der Zentrale Grenzwertsatz gibt eine Antwort:

In der oben beschriebenen Situation gilt

M ist approximativ $\mathrm{N}(\mu, \frac{\sigma^2}{n})$-verteilt.

4. Die Stichprobenvarianz als Schätzung für die Populationsvarianz

Ein Problem in der Praxis: Im Allgem. kennt man σ^2 nicht.

Auch σ^2 muss man dann schätzen.

Zwei Vorschläge für die

(aus der Stichprobe) **geschätzte** (Populations-)**Varianz**:

(i) die *Stichprobenvarianz*

$$\hat{\sigma}^2 := \frac{1}{n}\sum_{i=1}^{n}(x_i - m)^2.$$

(ii) die *modifizierte Stichprobenvarianz*

$$s^2 := \frac{1}{n-1}\sum_{i=1}^{n}(x_i - m)^2$$

Es gibt theoretische Begründung für beide Vorschläge

(vgl. Buch S. 124, S. 138).

Wir kommen darauf später zurück

und halten uns erst einmal an den Vorschlag (ii):

$$s^2 = \frac{1}{n-1} \sum_{i=1}^{n} (x_i - m)^2$$

Die Standardabweichung des Stichprobenmittelwertes $\bar{X}$ ist

$$\frac{\sigma}{\sqrt{n}}.$$

Die geschätzte Standardabweichung

des Stichprobenmittelwertes $\bar{X}$ ist

$$\boxed{s/\sqrt{n} =: f}$$

Diese Größe nennen wir auch den *Standardfehler*.

M ist approximativ $N(\mu, \frac{\sigma^2}{n})$-verteilt.

Und (gut für die Praxis):

M ist approximativ $N(\mu, f^2)$-verteilt.