Stichprobenverteilungen in der Statistik mittels simulierter Wahlumfragen

Marloes Maathuis

ETH Zürich

Lernziele

- Das Verstehen folgender Konzepte der Statistik:
  - Population, Parameter
  - Stichprobe, Parameterschätzung, Stichprobenverteilung

- Wichtig für:
  - Berechnung der Stichprobengrösse
  - Tests
  - Vertrauensintervalle

Anwendung Wahlumfragen

- Warum?
  - Immer aktuelle Beispiele
  - Einfache diskrete Verteilung
  - "Statistical literacy"

- Referenz: A.F. Gourgey (2000), "A Classroom Simulation Based on Political Polling To Help Students Understand Sampling Distributions", *Journal of Statistics Education* v.8, n.3. http://www.amstat.org/publications/JSE/secure/v8n3/gourgey.cfm

Ablauf

- Referendum in Griechenland
- Hands-on Übung + Computersimulation
- Diskussion der Stichprobenverteilung
- Zurück zum Referendum in Griechenland
- Fragen + Diskussion

Wahlumfragen Sparmassnahmen Griechenland

- Vorhersagen am 4./5. Juli:
  - Ja: 48.5% ± 2.5%
  - Ja: 48.5% ± 2.0%
  - Ja: 48.0% ± 2.5%
  - …

- Warum sind die Vorhersagen leicht unterschiedlich?
- Wie interpretieren wir solche Vorhersagen?

Vereinfachungen

- In der Wirklichkeit:
  - wissen wir nicht genau wer abstimmen wird
  - gibt es Personen die untentschieden sind oder ihre Meinung noch ändern werden
  - macht nicht jede Person die angefragt wird mit
  - antworten Personen nicht immer ehrlich

- Vereinfachungen:
  - wir kennen die Population
  - jede Person weiss schon was er/sie stimmen wird
  - jede Person die wir anfragen macht mit
  - jede Person antwortet ehrlich

⇒ Bügelperlen

Bügelperlen als Modell

- Wir haben eine (grosse/unendliche) Population mit p = Prozentsatz blaue Perlen
- Wir möchten den Wert von p wissen.
  Insbesondere: gilt $p > 50\%$ oder $p < 50\%$??

- Simulierte Wahlumfrage:
  - Wir ziehen eine zufällige Stichprobe der Grösse $n = 10$
  - Wir rechnen den Prozentsatz blau ($\hat{p}$) in der Stichprobe.
    $\hat{p}$ ist der Parameterschätzer.
  - Was können wir daraus schliessen?

Hands-on Übung

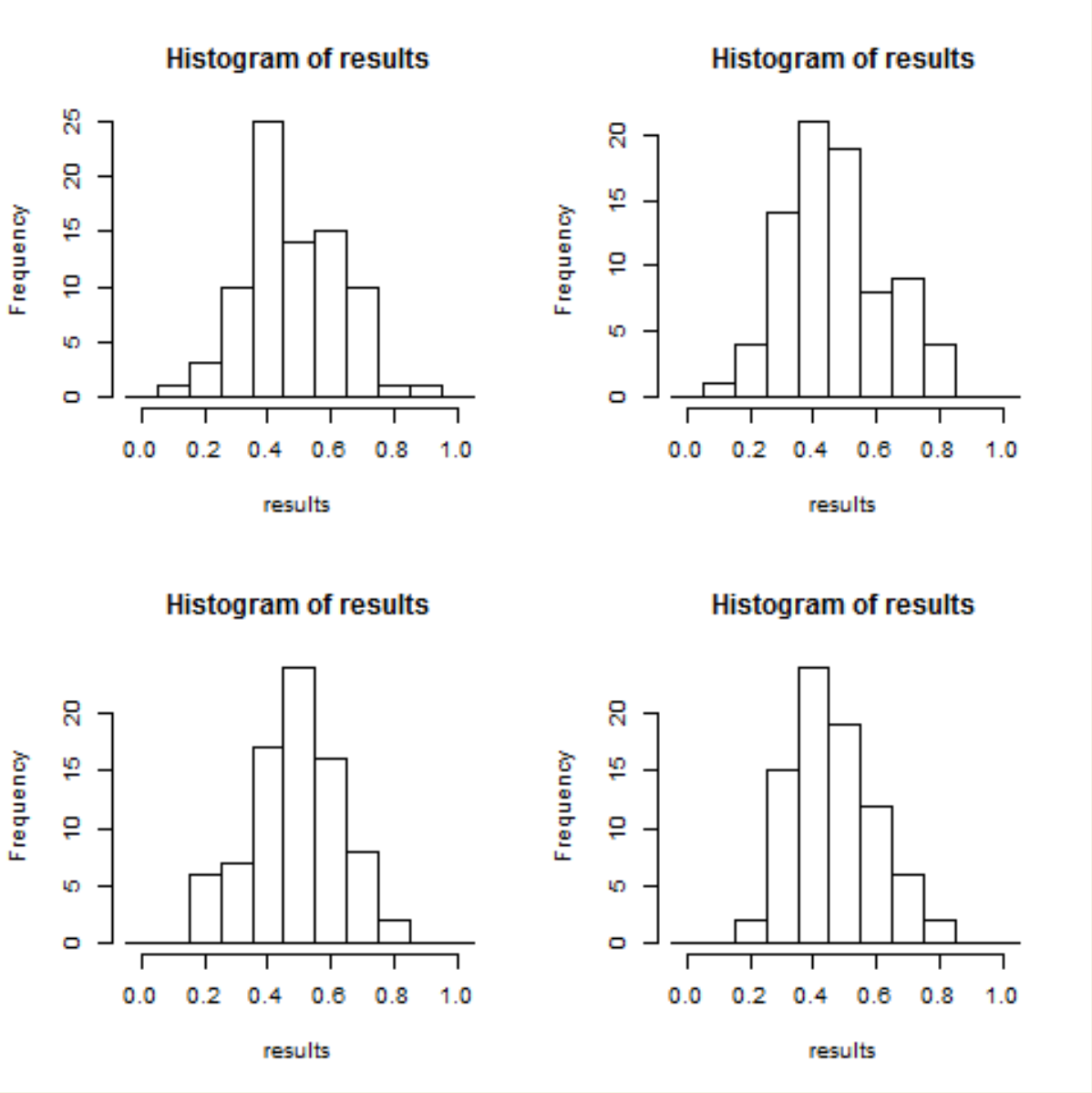
- 4 Gruppen
- Jede Gruppe hat eine Kopie der Population
- Vorgehen (etwa 20 mal):
  1. Ziehen Sie eine zufällige Stichprobe der Grösse $n = 10$ und notieren Sie den Prozentsatz blau ($\hat{p}$)
  2. Legen Sie die Perlen zurück, mischen Sie sie zusammen und wiederholen Sie Punkt 1

- Wir visualisieren die Resultate

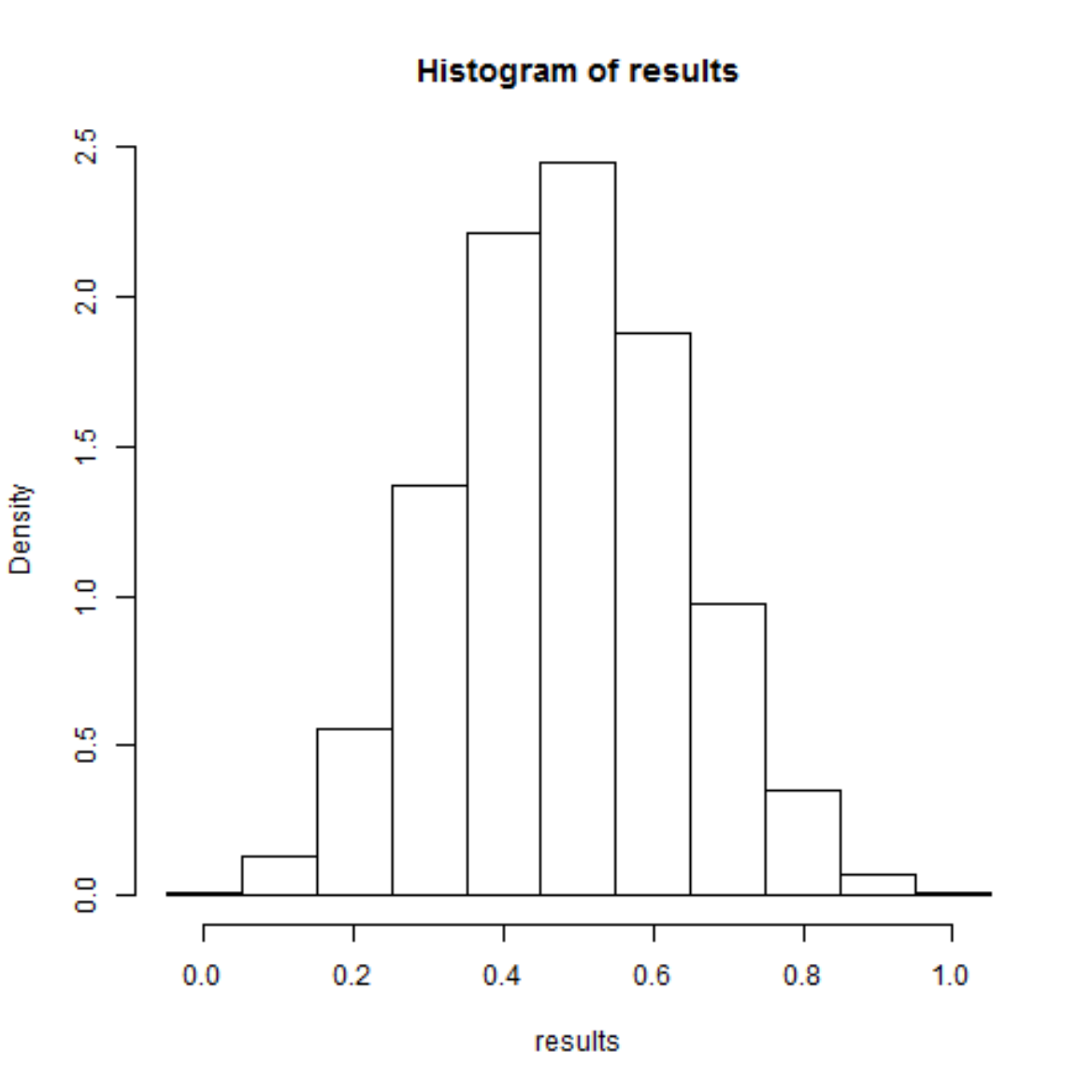
Computersimulationen mit R

```
# Parameters für die Simulation
npop    <- 2000    # Populationsumfang
p       <- 0.48    # Populutionsparameter (% 1s)
nsimul  <- 80      # Anzahl Stichproben
nsample <- 10      # Stichprobenumfang

# Die Population
population <- c(rep(1,npop*p),rep(0,npop*(1-p)))

# Simulation
results <- rep(NA, nsimul)
for (i in 1:nsimul){
   sample <- sample(population, nsample, replace=F)
   results[i] <- sum(sample)/nsample
}
```

Simulierte Stichprobenverteilungen, nsimul=80, nsample=10



Simulierte Stichprobenverteilung, nsimul=100000, nsample=10



Was sehen wir?

- $p \neq \hat{p}$
- p ist unbekannt, aber fix
- $\hat{p}$ kann man berechnen, ist aber zufällig:
  es ändert sich von Stichprobe zu Stichprobe

$\hat{p}$ hat eine Verteilung.

Können wir die Verteilung analysieren?

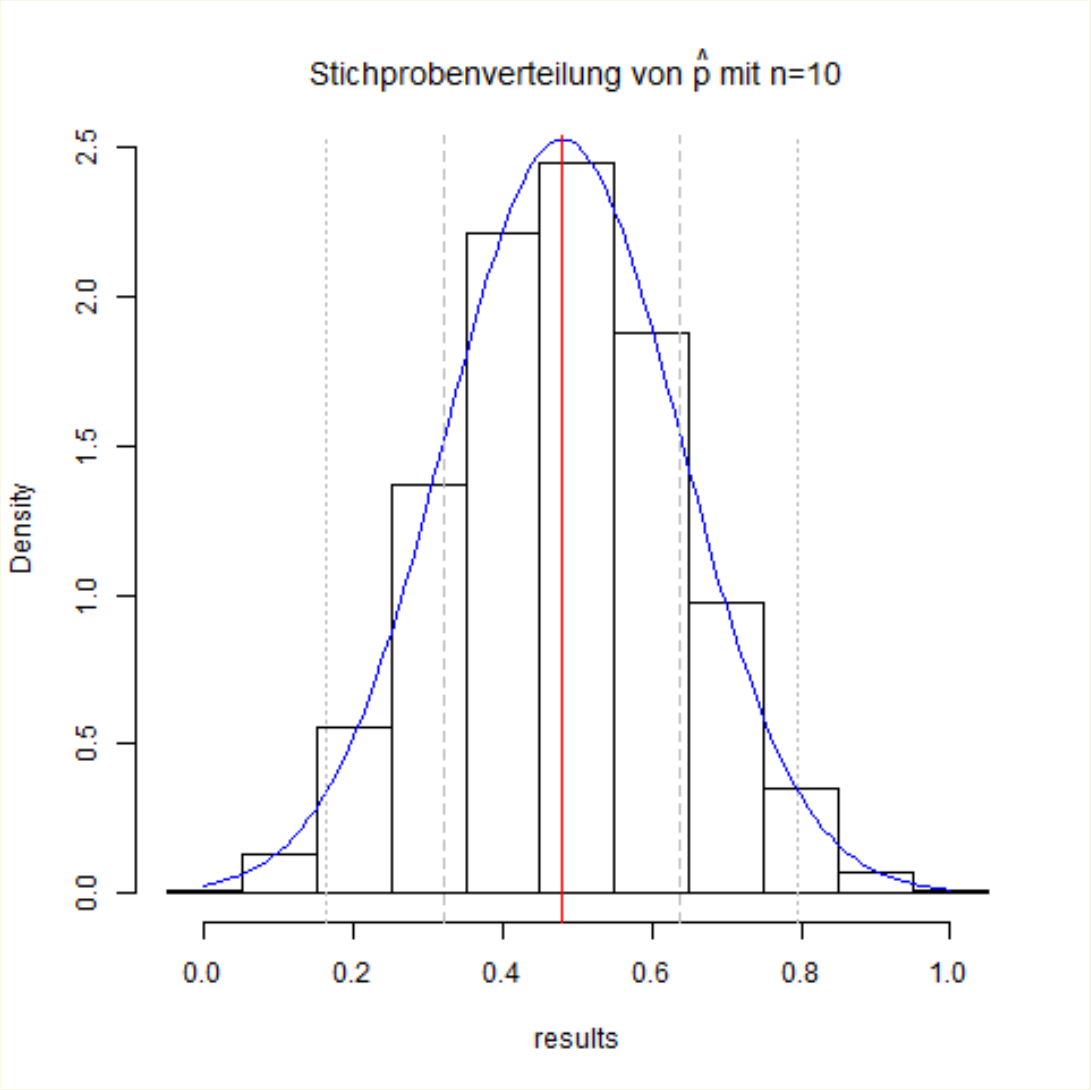
Analyse der Stichprobenverteilung

- $X =$ #blau in der Stichprobe, $\hat{p} = X/n$
- $X \sim Binom(n, p)$
- $E(X) = np$, $\mathrm{Var}(X) = np(1-p)$
- $E(\hat{p}) = E\left(\frac{X}{n}\right) = \frac{1}{n}E(X) = p$
- $Var(\hat{p}) = Var\left(\frac{X}{n}\right) = \frac{1}{n^2}Var(X) = \frac{p(1-p)}{n}$
- Standardfehler $\sigma_{\hat{p}} = \sqrt{\frac{p(1-p)}{n}}$; $\hat{\sigma}_{\hat{p}} = \sqrt{\frac{\hat{p}(1-\hat{p})}{n}}$

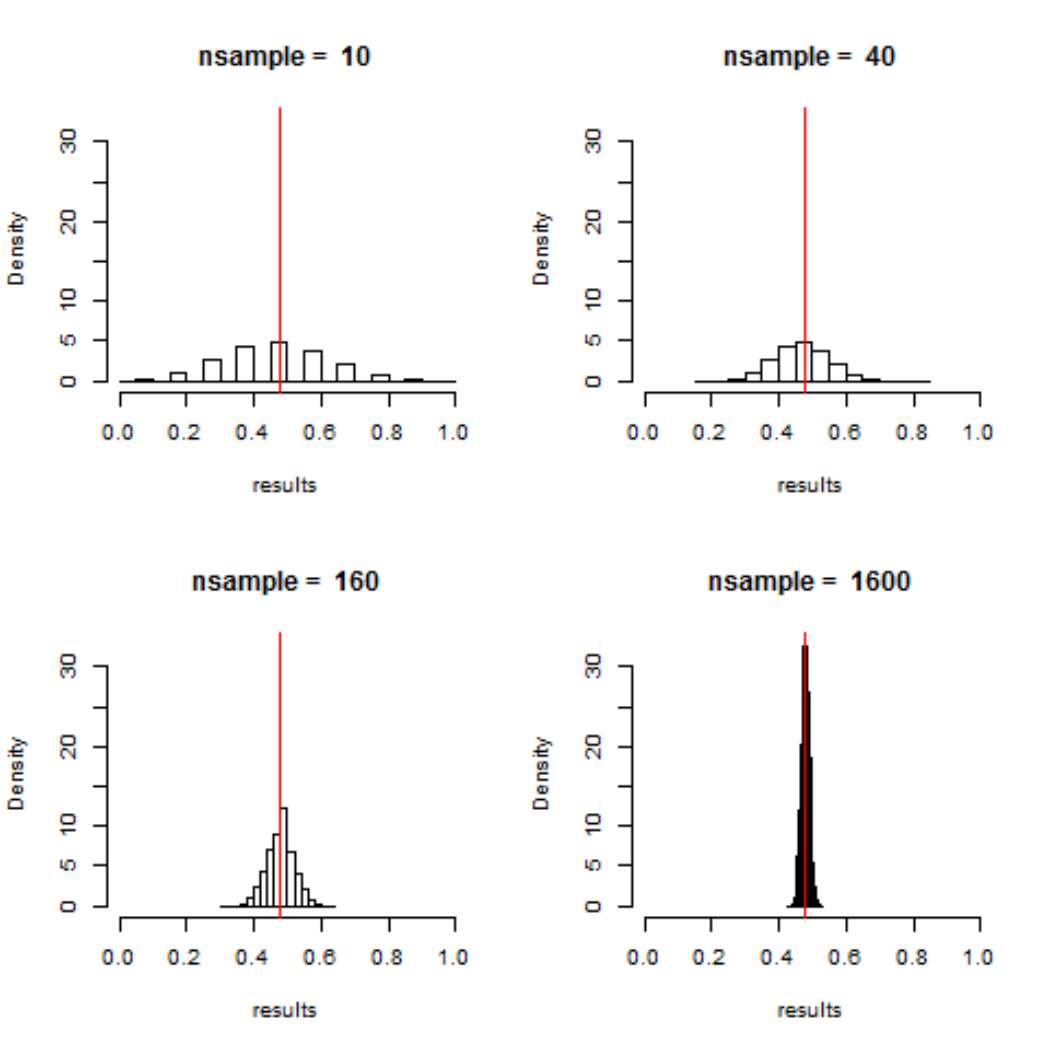
- Die verteilung von $\hat{p}$ ist approximativ Normal (ZGWS)
- Bei etwa 66% der Stichproben ist $\hat{p}$ im Intervall $p \pm \sigma_{\hat{p}}$
- Bei etwa 95% der Stichproben ist $\hat{p}$ im Intervall $p \pm 2\sigma_{\hat{p}}$

Populationsverteilung versus Stichprobenverteilung

```
> table(population)
population
   0    1
1040  960

> p
[1] 0.48
```



Rolle des Stichprobenumfangs



$npop = 20000$
$nsample = 10; 40; 160; 1600$
$nsimul = 100000$

$$\sigma_{\hat{p}} = \sqrt{\frac{p(1-p)}{n}}$$

n vier mal so gross ⇒
$\hat{p}$ zwei mal so genau

Zurück zur Abstimmung in Griechenland

- Die Vorhersagen sind gegeben als $\hat{p} \pm 2\,\hat{\sigma}_{\hat{p}}$.
  Diese sind 95% Vertrauensintervalle für p.

- Wir können n bestimmen:

$$\hat{\sigma}_{\hat{p}} \approx 1.25\% \Rightarrow \sqrt{\frac{0.5 * 0.5}{n}} \approx 1.25\% \quad \Rightarrow \quad n \approx 1600$$

Resultate Griechenland

- Die Resultate…

61.31%
38.69%

- In der Wirklichkeit:
  - wissen wir nicht genau wer abstimmen wird
  - gibt es Personen die untentschieden sind oder ihre Meinung noch ändern werden
  - macht nicht jede Person die angefragt wird mit
  - antworten Personen nicht immer ehrlich

Warum hands-on Übung?

- Hätten wir auch direkt Computersimulationen benutzen können?
- Ja, aber ich finde es schön zuerst die hands-on Übung zu machen:
  - Es ist konkreter. Man sieht wirklich, dass $\hat{p}$ zufällig ist.
  - Man merkt es sich besser (?)
  - Man kann die Computersimulation nachher hoffentlich besser verstehen

Mehr Material (I)

- The power of random sampling:
  - Chapter 19 of "Basic Statistics" by Freedman, Pisani and Purves

- Hands-on Experience with Sampling Distributions of the Sample Mean and Sample Proportion
  - Election polls: http://www.amstat.org/publications/JSE/secure/v8n3/gourgey.cfm
  - Dice and pennies: https://www.causeweb.org/repository/StarLibrary/activities/andrews_2003/
  - M&Ms: http://www.amstat.org/education/stew/pdfs/populationparameterswithmms.pdf

Mehr Material (II)

- Applet:
  - http://onlinestatbook.com/stat_sim/sampling_dist/index.html

- A story-based simulation for teaching sampling distributions:
  - Paper: http://onlinelibrary.wiley.com/doi/10.1111/test.12067/full
  - Video: https://www.youtube.com/channel/UC-E1qTmBIpOoHXJcbuhzqdw

- Open source statistical software R:
  - https://cran.r-project.org/

Fragen und Diskussion