JKU STATISTIK

Mal statistisch gesehen: Was kann man aus Daten lernen?

JOHANNES KEPLER UNIVERSITÄT LINZ

Andreas Quatember

Alternative Befragungsmethoden bei sensitiven Themen

Was ist Statistik (Wissenschaft von den Daten)? –

Intelligente Datenanalyse zum Zweck des Aufdeckens der darin verborgenen Informationen

Die *Surveystatistik* beschäftigt sich mit dem Schlussfolgern von den Resultaten von Stichprobenerhebungen auf interessierende Populationscharakteristika

Gemeint ist die wissenschaftliche Auseinandersetzung mit

- der Auswahltechnik zur Selektion der zu befragenden Erhebungseinheiten,
- der Methode des Rückschlusses von den erhobenen Daten auf die Populationscharakteristika und
- verwandten Themen (wie z.B. die Kompensierung von Nonresponse)

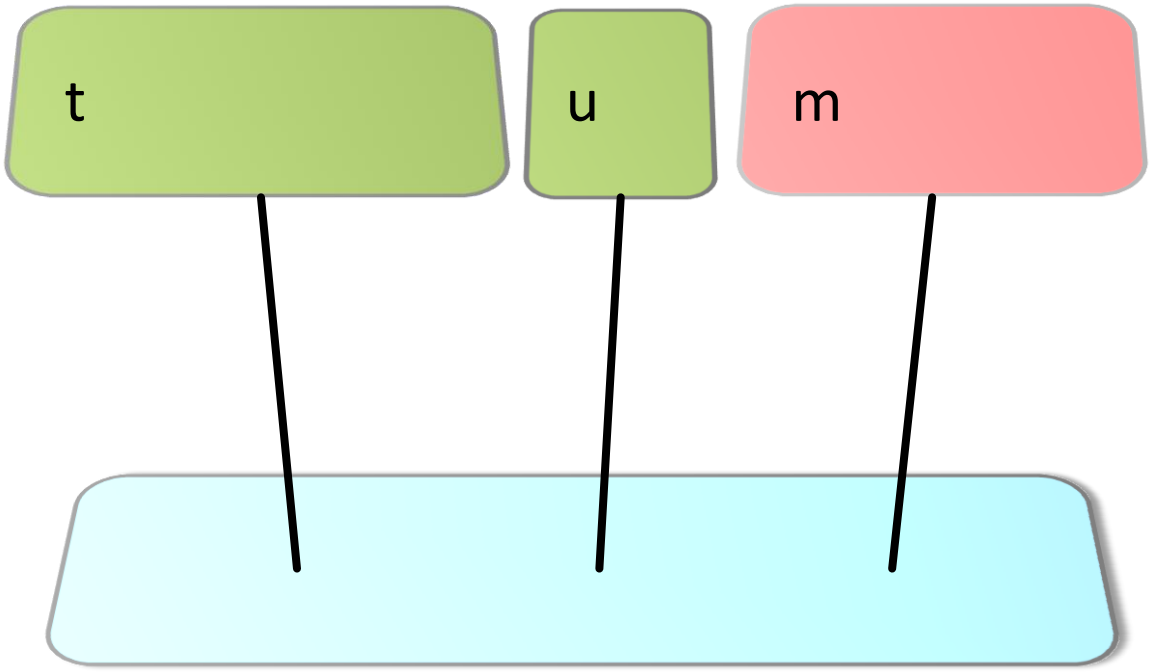
Die Problemstellung am Beispiel der Schätzung eines Anteils:

ZB der Anteil der gegen die Prüfungsregeln verstoßenden Studierenden

In einer einfachen Zufallsstichprobe s wird dieser Anteil durch den Stichprobenanteil geschätzt

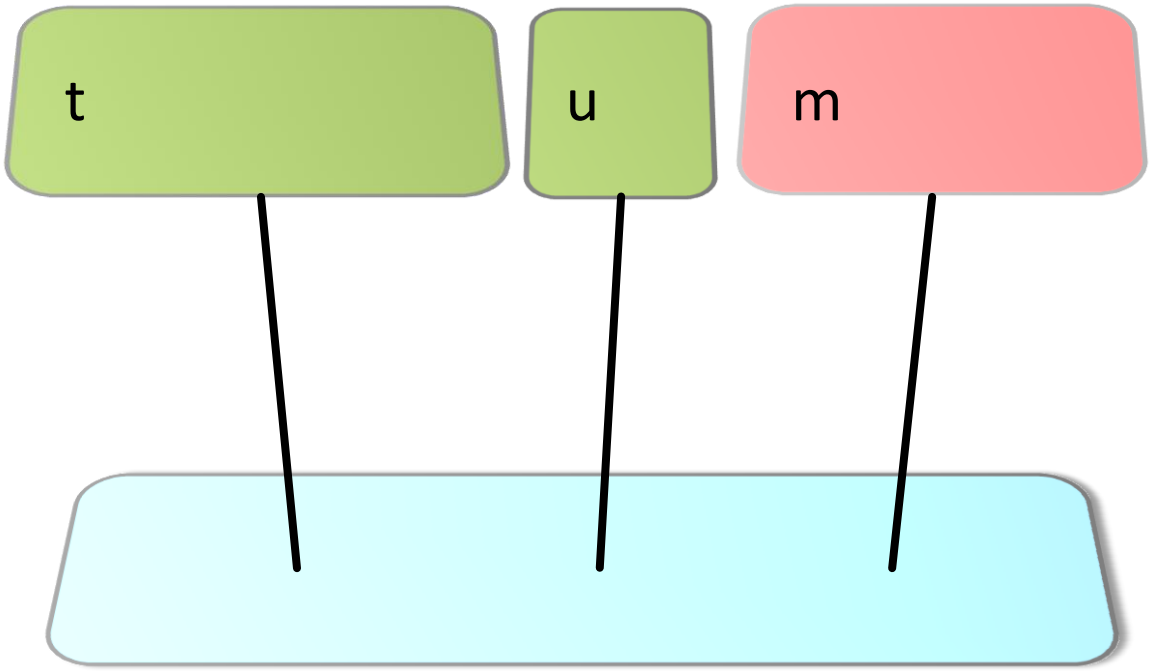
Unit-, Item-Nonresponse und unwahre Antworten zerlegen s hinsichtlich des interessierenden Merkmals in $s = t \cup u \cup m$

Folge: Verzerrung unbekannten Ausmaßes (Verringerung der Datenqualität)



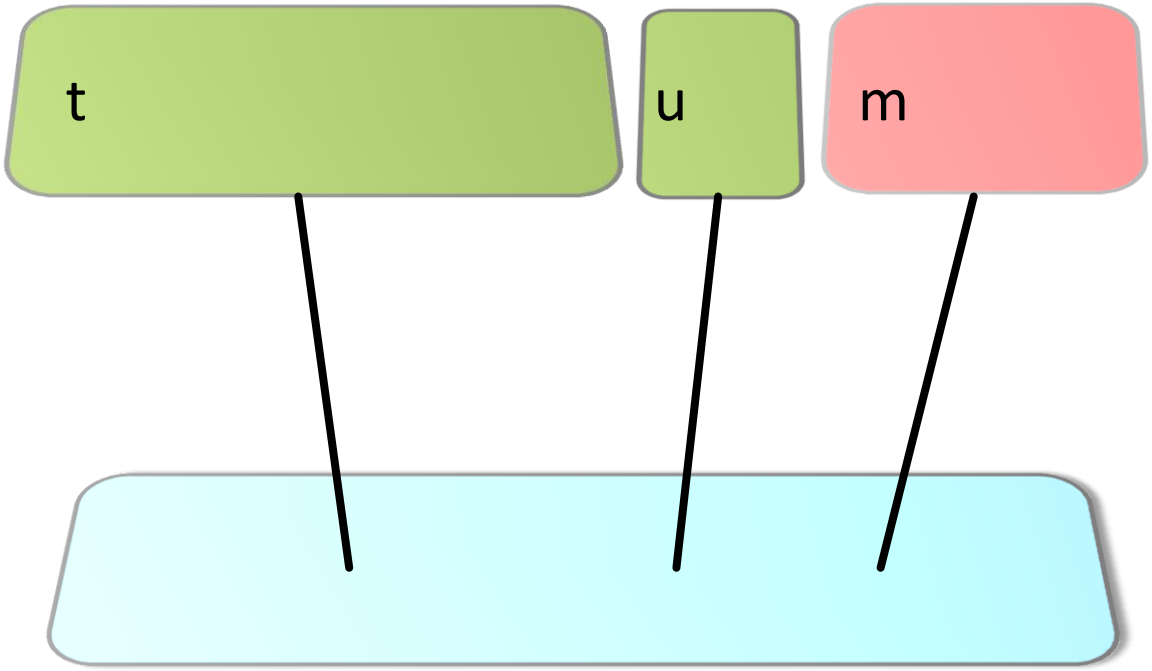
Methoden der Surveyforschung (*vor* der Datenerhebung)

Kontaktversuche, Anreize, Motivationsschreiben, Datenerhebungstechnik etc. verringern sowohl Nonresponse als auch unwahre Antworten



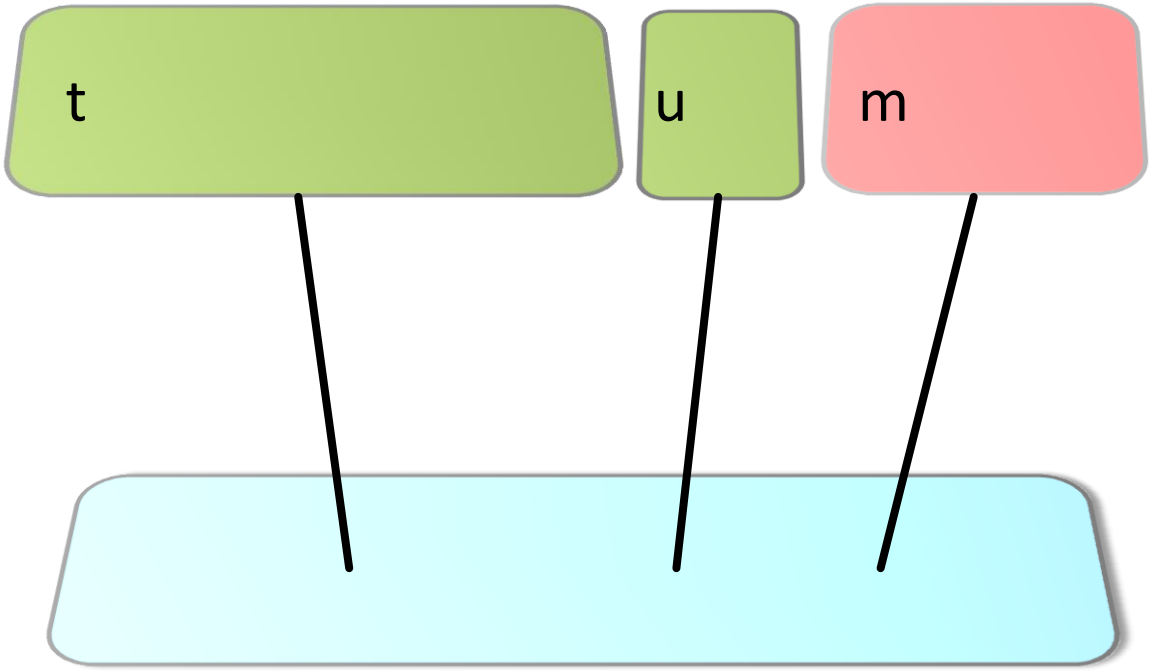
Methoden der Surveyforschung (*vor* der Datenerhebung)

Kontaktversuche, Anreize, Motivationsschreiben, Datenerhebungstechnik etc. verringern sowohl Nonresponse als auch unwahre Antworten



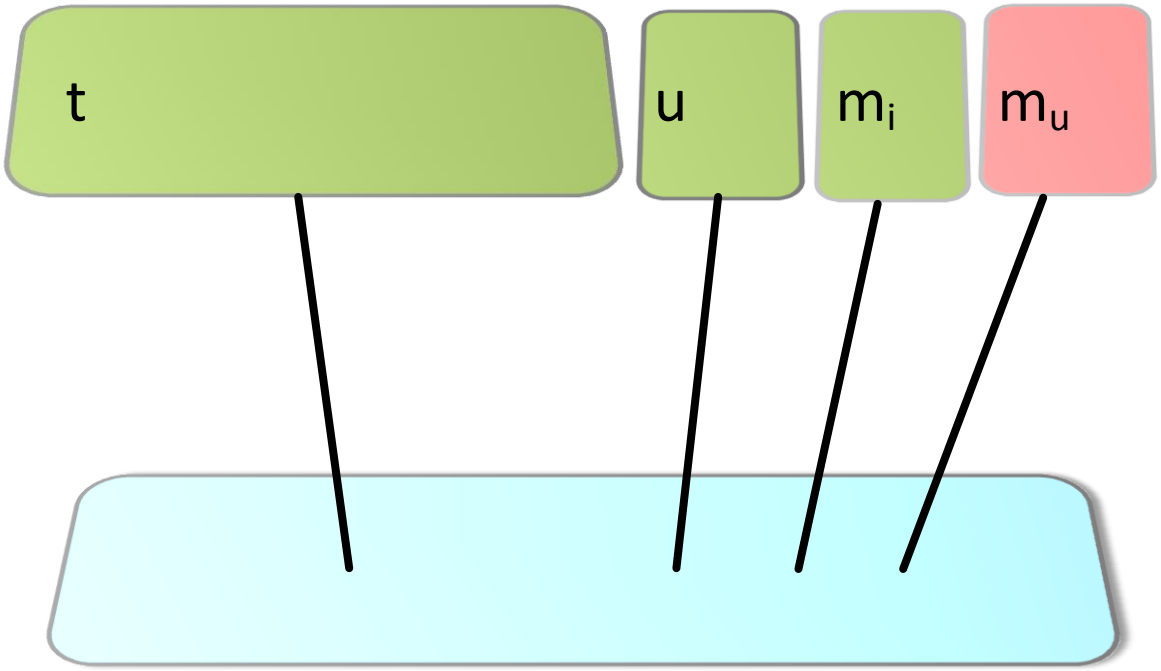
Datenimputation (*nach* der Datenerhebung)

Kompensiert Item-Nonresponse, aber nicht unwahre Antworten (Qualität vom Zutreffen des Nonresponsemodells abhängig): $m = m_i \cup m_u$



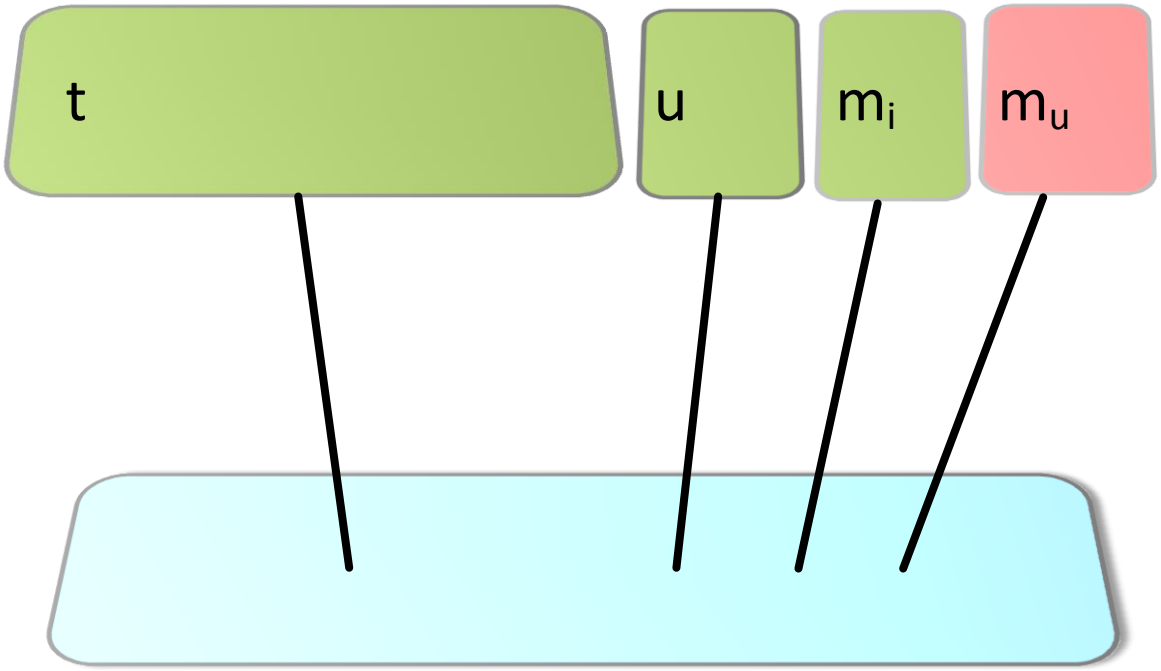
Datenimputation (*nach* der Datenerhebung)

Kompensiert Item-Nonresponse, aber nicht unwahre Antworten (Qualität vom Zutreffen des Nonresponsemodells abhängig): $m = m_i \cup m_u$



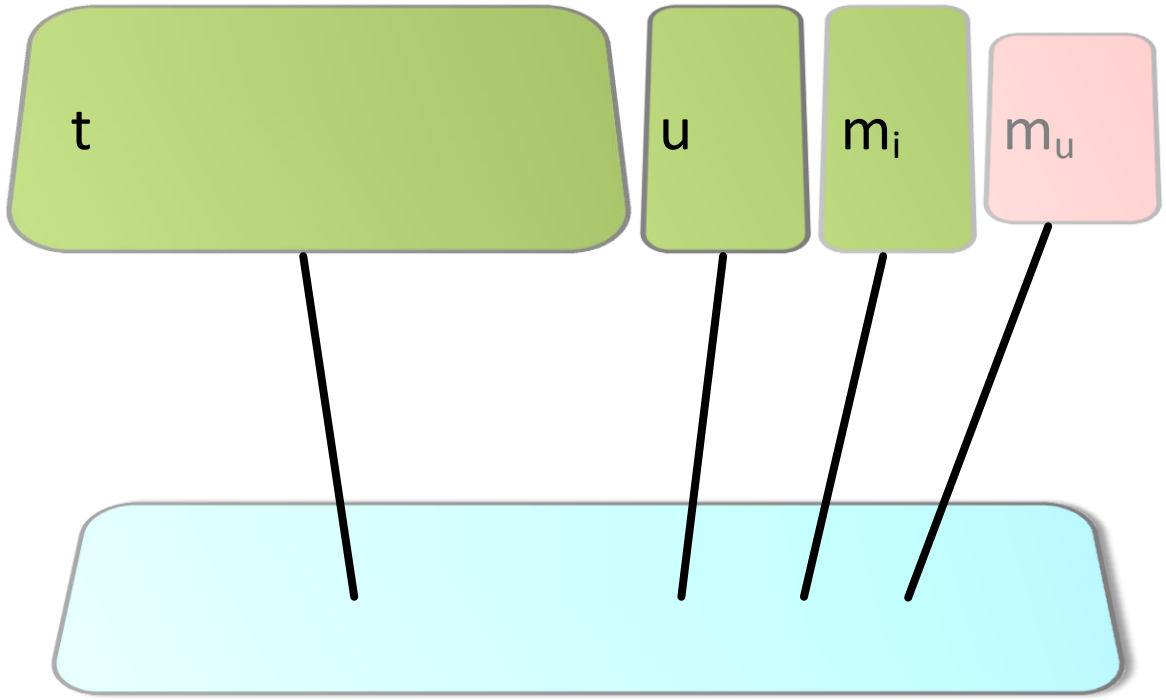
Gewichtungsanpassung (*nach* der Datenerhebung)

Kompensiert Unit-Nonresponse, aber nicht unwahre Antworten (Qualität vom Zutreffen des Nonresponsemodells abhängig)



Gewichtungsanpassung (*nach* der Datenerhebung)

Kompensiert Unit-Nonresponse, aber nicht unwahre Antworten (Qualität vom Zutreffen des Nonresponsemodells abhängig)



Fragen zu sensitiven Themen wie

- Einkommen
- Parteipräferenz
- Korruption
- häusliche Gewalt
- Steuerhinterziehung
- Fremdenfeindlichkeit
- …
- Wahlbeteiligung
- Sexuelle Orientierung
- Drogenkonsum (Doping)
- Mobbing am Arbeitsplatz
- Schwarzarbeit
- Antisemitismus

führen trotz aller Bemühungen zu hohen Anteilen an *Item-Nonresponse* **und** *unwahren Angaben*

Alternative Befragungsdesigns

Setzen an direkt *bei* der Datenerhebung

Schützen die Privatsphäre der Befragten in der Interviewsituation im Gegensatz zur direkten Befragung (Individualebene)

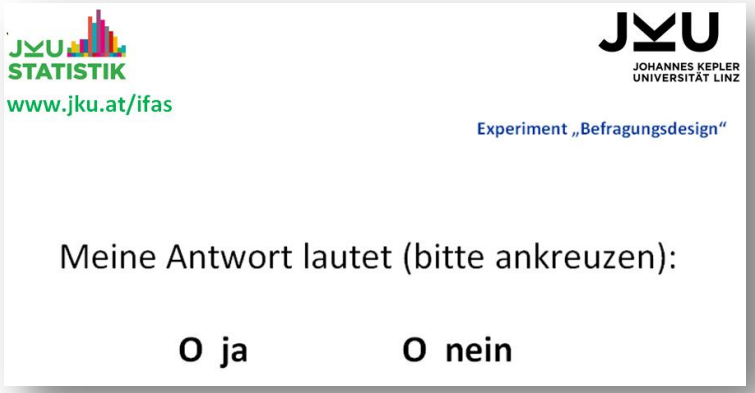
Zielen damit ab auf eine Verringerung von Item-Nonresponse **und** unwahren Antworten

Die Schätzung der interessierenden Populationscharakteristika ist wegen der Kenntnis der Designparameter dennoch möglich (Aggregatsebene)

Ein Beispiel eines solchen indirekten Befragungsdesigns ist folgendes Experiment mit einer „Randomized Response“-Technik („Maskieren“ der Antworten):

Sensitives Merkmal ist der Verstoß gegen die Prüfungsregeln in einem Statistikkurs („Schummeln“)

Frage: **Haben Sie wissentlich gegen die Prüfungsregeln verstoßen?**



Bei direkter Befragung ist mit hohem Item-Nonresponse („Das sag‘ ich nicht!“) und unwahren Antworten („nein“) zu rechnen!

Instruktionen der Randomized Response-Technik:

Folgen Sie vor dem Ankreuzen Ihrer Antwort bitte exakt den folgenden Anweisungen, damit Ihre Privatsphäre geschützt bleibt:

- *Denken Sie an eine Person (sie selbst, Mutter, Freund, …), von der Sie wissen, wann diese Geburtstag hat*
- *Merken Sie sich dieses ausgewählte Geburtsdatum und bleiben Sie dabei!*

Wenn der ausgewählte Geburtstag **im Jänner oder Februar** liegt, dann kreuzen Sie am Fragebogen „**ja**“ an!

Wenn der ausgewählte Geburtstag **im März oder April** liegt, dann kreuzen Sie am Fragebogen „**nein**“ an!

Wenn der ausgewählte Geburtstag **von Mai bis Dezember** liegt, dann beantworten Sie am Fragebogen nachfolgende Frage wahrheitsgetreu mit „**ja**“ **oder** „**nein**“:

Haben Sie wissentlich gegen die Prüfungsregeln verstoßen?

Beantworten Sie nachfolgende Frage wahrheitsgetreu mit „**ja**“ **oder** „**nein**“:

Haben Sie wissentlich gegen die Prüfungsregeln verstoßen?

Wenn der ausgewählte Geburtstag **im Jänner oder Februar** liegt, dann kreuzen Sie am Fragebogen „**ja**“ an!

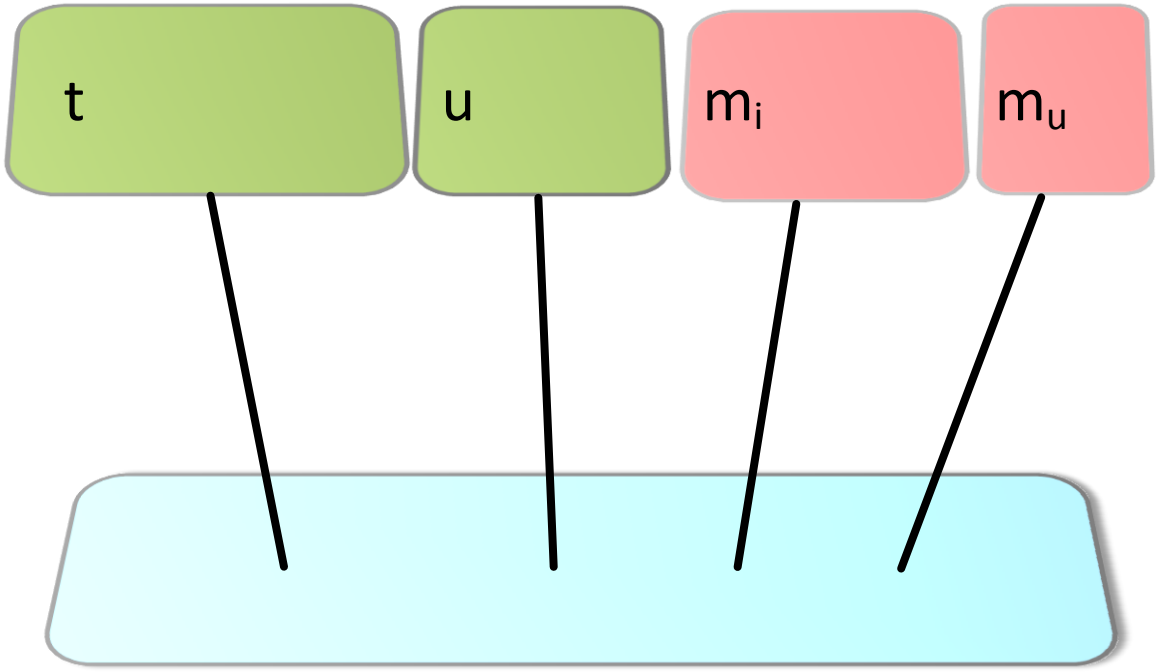
Wenn der ausgewählte Geburtstag **im März oder April** liegt, dann kreuzen Sie am Fragebogen „**nein**“ an!

Wenn der ausgewählte Geburtstag **von Mai bis Dezember** liegt, dann beantworten Sie am Fragebogen nachfolgende Frage wahrheitsgetreu mit „**ja**“ **oder** „**nein**“:

Haben Sie wissentlich gegen die Prüfungsregeln verstoßen?

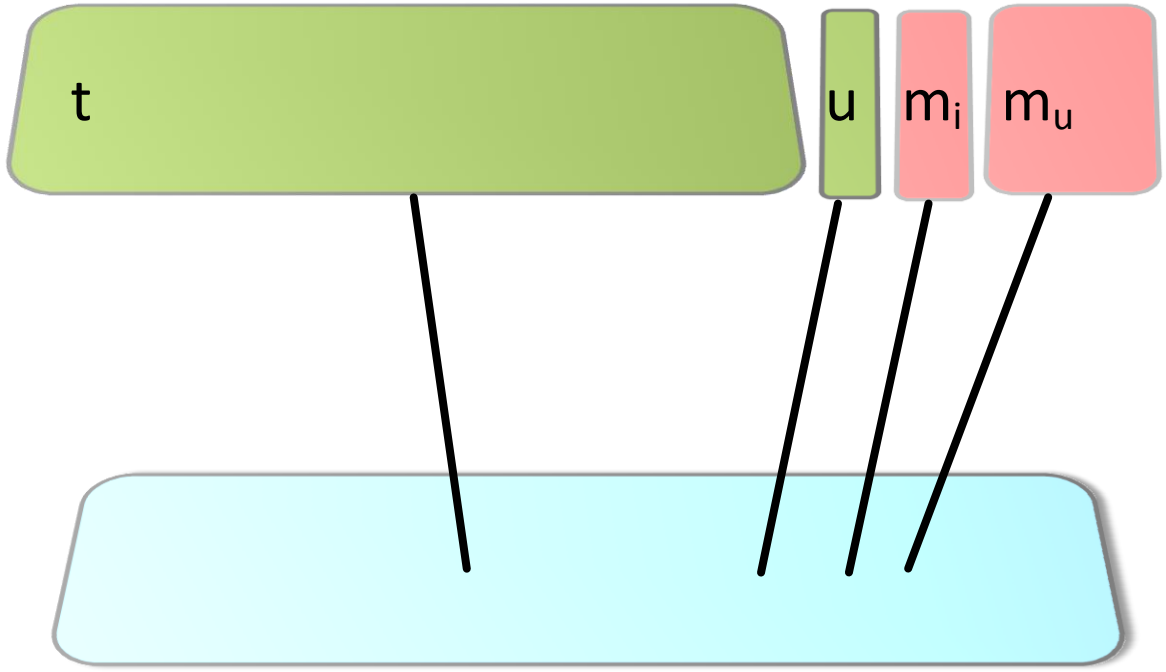
So kann die gegebene „ja“/„nein“-Antwort nicht der heiklen Frage zugeordnet werden

Erhöhter Schutz der Privatsphäre ➔ verringerte Falschantwort- und Item-Nonresponse-Rate



So kann die gegebene „ja“/„nein“-Antwort nicht der heiklen Frage zugeordnet werden

Erhöhter Schutz der Privatsphäre ➔ verringerte Falschantwort- und Item-Nonresponse-Rate



Vorhandene Informationen aus dem durchgeführten Experiment:

- 64 TeilnehmerInnen
- kein Nonresponse
- 20 x „ja“ angekreuzt
- 44 x „nein“

Designwahrscheinlichkeiten:

- $Pr(\text{Jänner/Februar}) \approx 59/365 \approx 1/6$
- $Pr(\text{März/April}) \approx 61/365 \approx 1/6$
- $Pr(\text{Mai bis Dezember}) \approx 245/365 \approx 4/6$

Wie lässt sich mit Hilfe statistischer Methodik aus diesen maskierten Informationen der interessierende Anteil an „Schwindlern“ unter den 64 Personen schätzen?

Der Anteil an „ja“-Antworten ist 20/64
Der erwartete Anteil an Jänner/Februar-Geburtstagen ist 1/6
Der erwartete Anteil an Mai bis Dezember-Geburtstagen ist 4/6

➔ Der Schätzer $\hat{\pi}_{Sch.}$ für den Anteil an Schummlern ist:

$$\hat{\pi}_{Sch.} = \frac{20/64 - 1/6}{4/6}$$

Wie lässt sich mit Hilfe statistischer Methodik aus diesen maskierten Informationen der interessierende Anteil an „Schummlern“ unter den 64 Personen schätzen?

Der Anteil an „ja“-Antworten ist 20/64
Der erwartete Anteil an Jänner/Februar-Geburtstagen ist 1/6
Der erwartete Anteil an Mai bis Dezember-Geburtstagen ist 4/6

➔ Der Schätzer $\hat{\pi}_{Sch.}$ für den Anteil an Schummlern ist:

$$\hat{\pi}_{Sch.} = \frac{20/64 - 1/6}{4/6} = 0{,}22$$

Geschätzte 22 Prozent der 64 Studierenden haben wissentlich gegen die Prüfungsregeln verstoßen (oder 14 Personen)

Die Formalisierung

Anwendungsvoraussetzung ist neben einer möglichst einfachen Vorgehensweise eine allgemein einsetzbare Theorie

Einheitliche statistische Theorie für

- verschiedene Merkmalstypen
- möglichst viele verschiedene Befragungsdesigns
- alle Formen der zufälligen Ziehung von Stichproben
- individuell steuerbaren Schutz der Privatsphäre
- Einbeziehung wahrer Antworten

Für verschiedene Randomized-Response-Techniken ist eine einheitliche Darstellung des unverzerrten Anteilschätzers gegeben durch

$$\hat{\pi}_{RR} = \frac{1}{N} \cdot \sum_s \frac{z_k - u_k}{v_k} \cdot d_k = \frac{1}{N} \cdot \sum_s y_k^i \cdot d_k$$

mit $u_k \equiv p_{2k} + p_{3k} \cdot \pi_B + p_{4k}$ und $v_k \equiv p_{1k} - p_{2k}$ und ein unverzerrter Varianzschätzer durch

$$\hat{V}(\hat{\pi}_{RR}) = \frac{1}{N^2} \cdot \left(\sum\sum_s \frac{\pi_{kl} - \pi_k \cdot \pi_l}{\pi_{kl}} \cdot \frac{y_k^i}{\pi_k} \cdot \frac{y_l^i}{\pi_l} + \sum_{s_{RR}} \frac{u_k(1-u_k)}{v_k^2} \cdot d_k + \sum_{s_{RR}} y_l^i \cdot \frac{1 - v_k - 2u_k}{v_k} \cdot d_k \right)$$

(Quatember 2016).

Für das durchgeführte Experiment gilt

$$\hat{\pi}_{Sch.} = \frac{\overline{z}_s - p_4}{p_1} \approx \frac{20/64 - 1/6}{4/6} = 0{,}22$$

und

$$\hat{V}(\hat{\pi}_{Sch.}) = \frac{1}{Np_1} \cdot \left(\frac{p_4(1-p_4)}{p_1} + (1 - p_1 - 2p_4) \cdot \hat{\pi}_{Sch.} \right)$$

$$\approx \frac{1}{64\frac{4}{6}} \cdot \left(\frac{\frac{1}{6}\left(1 - \frac{1}{6}\right)}{\frac{4}{6}} + \left(1 - \frac{4}{6} - 2\frac{1}{6}\right) \cdot 0{,}22 \right)$$

$$= 0{,}0049$$

Literaturhinweise:

Indirekte Befragungsdesigns:

Chaudhuri A. und Christofides T.C. (2013), Indirect questioning in sample surveys, Springer, Heidelberg.

Chaudhuri A., Christofides T.C. und Rao C.R. (eds.) (2016), Handbook of Statistics (Volume 34): Data Gathering, Analysis and Protection of Privacy through Randomized Response Techniques: Qualitative and Quantitative Human Traits, Elsevier, Amsterdam.

Statistische Theorie zu Randomized Response-Techniken:

Quatember A. (2014), A randomized response design for a polychotomous sensitive population and its application to opinion polls, Model Assisted Statistics and Applications, 9, pp. 11-23.

Quatember A. (2015), Pseudo-Populations - A Basic Concept in Statistical Surveys, Springer, Cham, ch. 6.

Quatember A. (2016), A Mixture of True and Randomized Responses in the Estimation of the Number of People Having a Certain Attribute, In: Chaudhuri A. and Christofides T.C. and Rao C.R. (eds.) (2016), Handbook of Statistics (Volume 34): Data Gathering, Analysis and Protection of Privacy through Randomized Response Techniques: Qualitative and Quantitative Human Traits, Elsevier, Amsterdam, pp. 91-103.

Quatember A. (2018), A discussion of the two different aspects of privacy protection in indirect questioning designs, Quality & Quantity, accepted.

Praktische Anwendungen:

Corbacho A., *Gingerich* D.W., *Oliveros V. und Ruiz-Vega M.* (2016), **Corruption** as a Self-Fulfilling Prophecy: Evidence from a Survey Experiment in Costa Rica, in: American Journal of Political Science, Vol. 60(*4*), 1077-1092.

Gonzalez-Ocantos E., *Kiewiet de Jonge* C., *Melendez C., Osorio J. und Nickerson D.W.* (2012), **Vote Buying** and Social Desirability Bias: Experimental Evidence from Nicaragua, in: American Journal of Political Science, Vol. 56(*1*), 202-217.

Kirchner A., Krumpal I., Trappmann M. und von Hermanni H. (2013), Messung und Erklärung von **Schwarzarbeit** in Deutschland - Eine empirische Befragungsstudie unter besonderer Berücksichtigung des Problems der sozialen Erwünschtheit (Measuring and Explaining Undeclared Work in Germany – An Empirical Survey with a Special Focus on Social Desirability Bias), Zeitschrift für Soziologie, Vol. 42(4), pp.291-314.

Krumpal I. (2012), Estimating the Prevalence of **Xenophobia** and **Anti-Semitism** in Germany: A Comparison of Randomized Response and Direct Questioning, Social Science Research, 41(6), pp. 1387-1403.

Qualitätsüberprüfung:

Jann B., *Jerke J. und Krumpal I.* (2012), Asking Sensitive Questions Using the Crosswise Model: An Experimental Survey Measuring Plagiarism, in: The Public Opinion Quarterly 76(*1*), 32-49.

Lensvelt-Mulders G.J.L.M., van der Heijden P.G.M., Laudy O. und van Gils G. (2006), A validation of a computer-assisted randomized response survey to estimate the prevalence of fraud in social security, Journal of the Royal Statistical Society. A, 169(2), pp. 305-318.

Rosenfeld B., *Imai K. und, Shapiro J.N.* (2016), An Empirical Validation Study of Popular Survey Methodologies for Sensitive Questions, in: American Journal of Political Science, Vol. 60(*3*), 783-802.

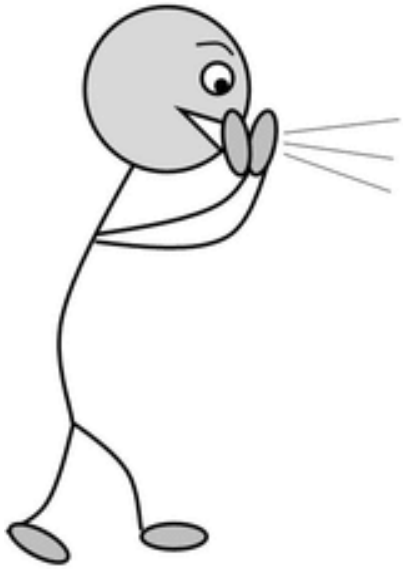
Ist der Vortrag gelungen?

Falls Sie selbst **im Jänner oder Februar** Geburtstag haben, dann sagen Sie gleich laut „**ja**“!

Falls Sie selbst **im März oder April** Geburtstag haben, dann sagen Sie gleich laut „**nein**“!

Falls Ihr Geburtstag **von Mai bis Dezember** liegt, dann beantworten Sie nachfolgende Frage wahrheitsgetreu mit „**ja**“ **oder** „**nein**“:

Ist der Vortrag gelungen?



Antworten Sie

JETZT!