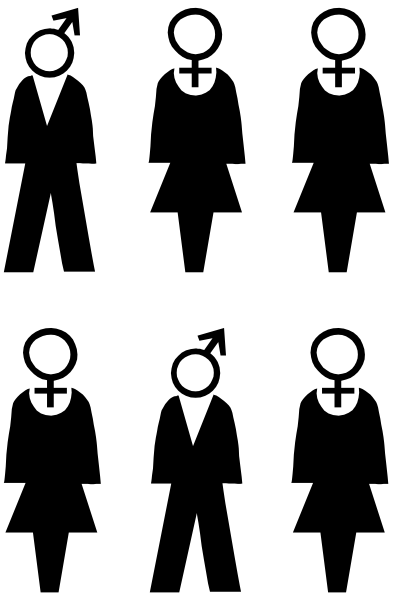
Grunnatriði ályktunartölfræði

Fyrirlestur í Aðferðafræði II

Lýsandi tölfræði

- Lýsum ákveðnum hópi einstaklinga (staka)
  - Mælitölur, t.d. Pearson fylgnistuðull, Kendalls tá, meðaltal, …
  - Myndrit, t.d. súlurit, stöplarit eða önnur myndræn framsetning



Ályktunartölfræði

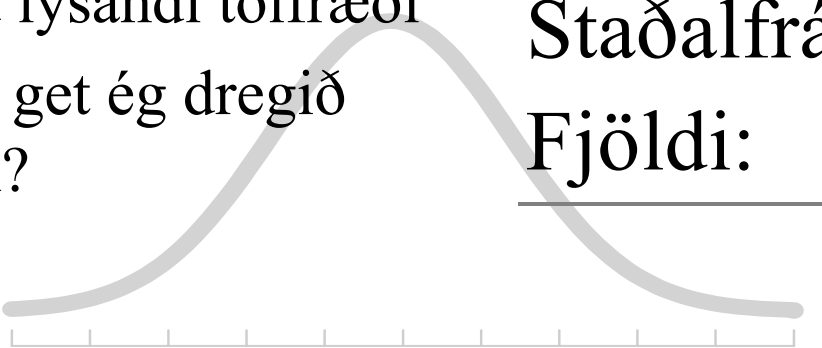
- Aðeins aðgangur að hluta hópsins
  - Ályktað um hópinn allan á grundvelli lýsingar á hluta hans
  - Stuðst við lýsandi tölfræði
  - Spáð um eiginleika alls hópsins
  - Nákvæmni spárinnar metin

A:1.1, bls. 3

Ályktað um hópinn

- Ég vel af tilviljun 5 nemendur
  - Ég reikna helstu lýsandi tölfræði
  - Hvaða ályktanir get ég dregið um allan hópinn?

Meðaltal:	5,50
Staðalfrávik:	1,94
Fjöldi:	5



Fleiri úrtök

- Mikill breytileiki frá einu úrtaki til annars
  - Meðaltalið segir fjarska lítið
  - Meðaltal meðaltala er betra en erfitt að treysta vegna mikillar dreifingar
- Kjarni viðfangsefnis ályktunartölfræði

Úrtak	Meðaltal	Staðal frávik
1	5,50	1,94
2	5,67	1,43
3	5,67	1,33
4	5,50	2,08
5	4,17	1,49
6	6,17	1,35
Í heild	5,44	1,61

Úrtak og þýði

- Þýði (*population*)
  - Skilgreindur hópur sem leitað er upplýsinga um
  - Ýmist endanlegur …
  - …eða óendanlegur
- Úrtak (*sample*)
  - Endanlegur hópur einstaklinga í þýði

Ályktunartölfræði reynir að álykta um þýðið á grunni upplýsinga úr úrtaki

A:1.2, bls. 4–5, 6–7

Þýðis- og úrtakstala

- Þýðistala (*parameter)*
  - Mælitala, t.d. meðaltal eða staðalfrávik, sem lýsir eiginleika þýðis
  - Yfirleitt óþekkt, sérstaklega ef þýðið er stórt eða óaðgengilegt
  - Þýðistölur eru táknaðar með grískum stöfum, t.d. μ, σ og π.
- Úrtakstala (*statistic*)
  - Mælitala sem lýsir eiginleika úrtaks
  - Ef úrtakið er *rétt dregið* getur úrtakstalan gefið upplýsingar um eiginleika þýðisins
  - Úrtakstölur eru táknaðar með rómverskum stöfum, t.d. $\bar{X}$, s og p.

A:1.2, bls. 6–7

Skekkja og villa

- Skekkja (*bias*)
  - Óskekkt mælitala gefur rétta mynd af þýði þegar til lengdar er litið
  - Í einstökum úrtökum víkur úrtakstalan frá samsvarandi þýðistölu, sbr. *villu*
- Villa (*error*)
  - Villa birtist sem breytileiki úrtakstölu frá einu úrtaki til annars
  - Villan er að jafnaði háð stærð úrtaksins

A:5.1, bls. 122–123; A: 2.3, bls. 22–23

Óskekktar mælitölur

Mælitölur eru ýmist skekktar (*biased*) eða óskekktar.

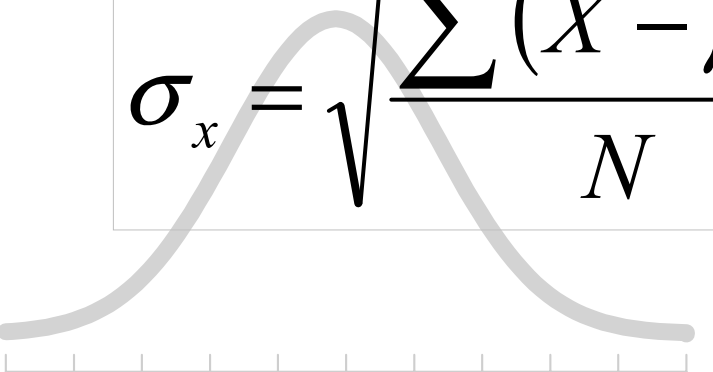
Meðaltal í úrtaki er óskekkt, það er gefur rétta mynd af meðaltali þýðis.

Staðalfrávik eins og það var kennt í Aðferðafræði I er skekkt mælitala og vanmetur að jafnaði breytileika þýðis. Þetta er leiðrétt með því að nota N–1 í stað N í nefnaranum.

Skekkta staðalfrávikið er sjaldan notað, helst þó þegar við þekkjum allt þýðið eða metum úrtakið án þess að vilja spá um þýðið.

$$\hat{\sigma}_x = \sqrt{\frac{\sum \left(X - \overline{X}\right)^2}{N-1}}$$

$$\sigma_x = \sqrt{\frac{\sum \left(X - \mu\right)^2}{N}}$$



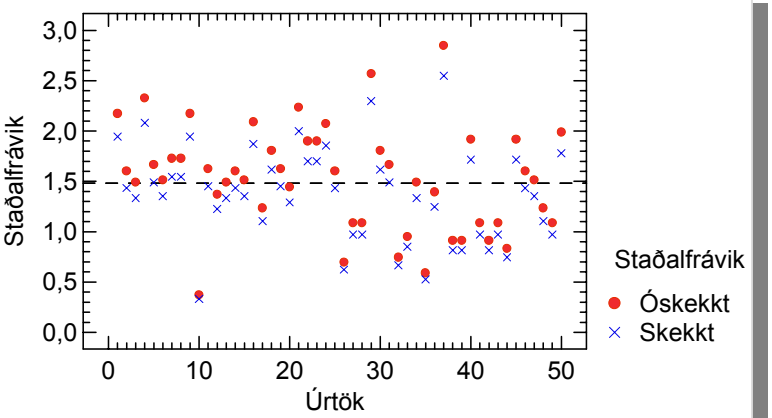
†A:3.4, bls. 59

Samanburður á staðalfrávikum

Skekktu staðalfrávikin vanmeta dreifinguna í þýðinu

- Taflan sýnir fimm nemenda úrtökin sex sem ég dró áðan. Tveir síðustu dálkarnir sýna að skekkta staðalfrávikið er alltaf lægra en hitt.
- Myndin sýnir að það er mikill breytileiki frá einu úrtaki til annars. En í hverju og einu úrtaki er óskekkta mælitalan hærri en sú skekkta.
- Það er því villa í báðum mælitölum sem birtist sem breytileiki milli úrtaka. Óskekkta staðalfrávikið dreifist í kringum þýðistöluna en sú skekkta dreifist í kringum lægra gildi.
- Það er því villa í báðum úrtaksmælitölum en auk þess er önnur þeirra skekkt niður á við.

		Staðalfrávik	
Úrtak	Meðaltal	Skekkt	Óskekkt
1	5,50	1,94	2,17
2	5,67	1,43	1,60
3	5,67	1,33	1,49
4	5,50	2,08	2,33
5	4,17	1,49	1,67
6	6,17	1,35	1,51
Í heild	5,44	1,61	1,80



Spá (*estimation*)

Punktspá (*point estimate*)

Punktspá felst í því að ákvarða sennilegasta gildi þýðistölunnar.

Byggt er á úrtakstölum og þær notaðar sem spátölur (*estimate*). Spátölur eru táknaðar sem gríska stafir með hatti, t.d. $\hat{\sigma}$, $\hat{\mu}$ og $\hat{\pi}$.

Til að vera góðar spátölur þurfa úrtakstölur að vera óskekktar og með sem minnsta villu.

Punktspá er villandi því eitt tiltekið gildi gefur til kynna að spáin sé hárnákvæm.

Bilspá (*interval estimate*)

Bilspá bætir úr annmarka punktspárinnar með því að gefa upp bil í stað eins tiltekins gildi fyrir samsvarandi þýðistölu. .

Bilið miðast við ákveðið öryggi, yfirleitt 95 eða 99%. Ef við viljum vera viss um að bilið innihaldi þýðistöluna veljum við mikið öryggi; það leiðir þó til þess að bilið víkkar svo oft er betra að velja styttra bil og minna öryggi.

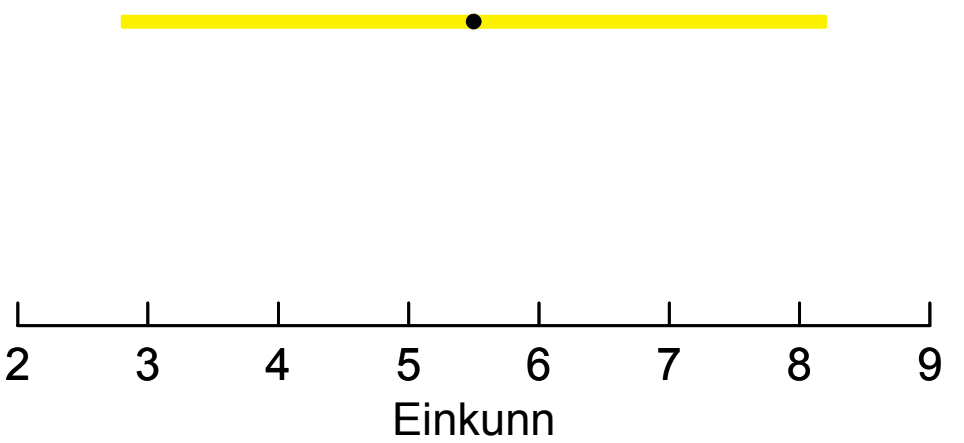
Punktspá: Meðalaldur nemenda í Aðferðafræði er 24,7 ár

Bilspá: Meðalaldur nemenda er á bilinu 21,8 – 27,6 ár miðað við 95% öryggi

A:5, bls. 121, 122, 125

Bilspá fyrir nemendurna fimm

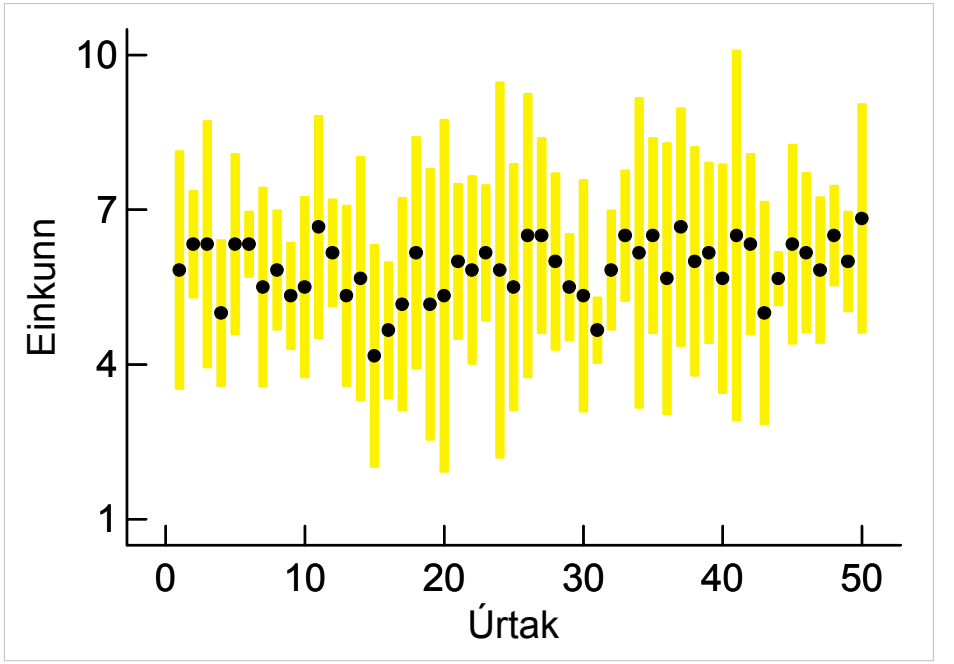
- Bilspá gefur til kynna að …
  - þýðismeðaltalið liggi á bilinu 2,8 – 8,2
  - þetta sé vitað með 95% öryggi



A:5.2, bls. 125–126

Túlkun öryggisbila

- Að jafnaði munu 95% þessara öryggisbila innihalda þýðismeðaltalið
  - Að jafnaði munu aðeins 2–3 öryggisbila *ekki* innihalda þýðismeðaltalið
  - Við höfum enga möguleika til að vita *hver* þeirra eða nákvæmlega *hve mörg* innihalda þýðismeðaltalið



A:5.2, bls. 129–130

Athugun á heimilisofbeldi

Í rannsókn á heimilisofbeldi frá 1997 kom fram a 0,8% karla og 1,3% kvenna urðu fyrir ofbeldi af hendi maka á síðustu 12 mánuðum.

Miðað við punktspá samsvarar þetta því að um 1% þjóðarinnar eða 1.750 manns verði fyrir heimilisofbeldi; 650 karlar og 1.100 konur.

Bilspá gefur aðra sýn. Miðað við 95% öryggi eru það 0,7–1,5% af þjóðinni sem verða fyrir ofbeldi af hendi maka eða 986–2.285 manns.

$N = 3.000$

$p = 0{,}01$

Það samsvarar 1.750 einstaklingum

á aldrinum $18 - 65$ ára

Með 95% öryggi:

$0{,}007 \leq \pi \leq 0{,}015$

Eða á bilinu $986 - 2.285$ einstaklingar

miðað við þjóðina alla

Tilgátuprófun

Í stað þess að (a) lýsa úrtaki eða þýði með mælitölu eða (b) álykta um þýði með punktspá eða bilspá viljum við stundum (c) fá svör við tilteknum spurningum.

Slík tilgátuprófun krefst þess að ég setji fram tilgátu (*hypothesis*) sem ég prófa á formlegan hátt.

Tilgátuprófun felur í sér ályktun um þýðið en ekki spá. Hún útilokar þó ekki punkt- eða bilspá.

Kyn	Fjöldi	%
Karlar	18	36,0
Konur	32	64,0
Samtals	50	100,0

Ég vil vita hvort annað hvort kynið sé fjölmennara en hitt í námskeiðinu. Ég dreg 50 nemenda úrtak og leita svars við spurningunni.

A:6, bls. 154–155

Tilgátur

- Upplýsingar um 50 nemenda úrtak segir mér ekki eitt og sér hvort kynjahlutfallið sé jafnt í *þýðinu*
  - Til að svara því þarf ég að setja fram *tilgátu* (hypothesis)
  - Aðaltilgátan setur fram þá spurningu sem ég vil leita svara við

Aðaltilgáta (*alternative hypothesis*)

$$H_1 : \pi_{Karlar} \neq \pi_{Konur}$$

A:6.1, bls. 155–156

Núlltilgáta

- Aðaltilgátuna get ég ekki prófað beint
  - Því set ég einnig fram núlltilgátu (*null hypothesis*), andhverfu aðaltilgátunnar
  - Núlltilgátan staðhæfir að enginn munur sé á hlutföllunum í þýðinu

Núlltilgáta (*null hypothesis*)

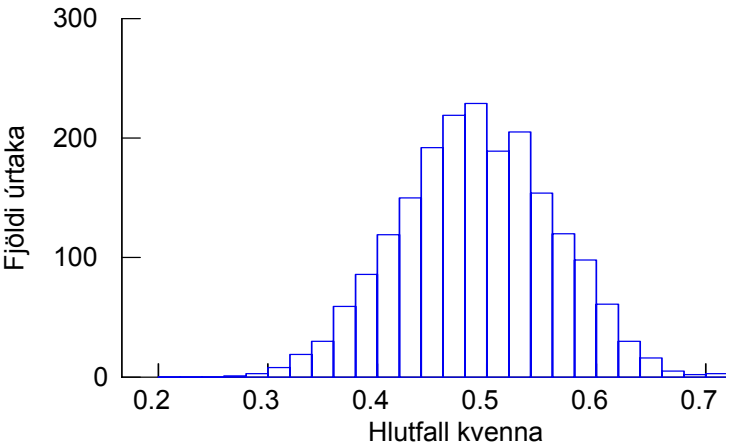
$$H_0 : \pi_{Karlar} = \pi_{Konur}$$

A:6.1, bls. 156–157

Núlltilgátuprófun

Tilgátuprófun (*hypothesis testing*) krefst þess að við vitum dreifingu úrtakstalna þegar núlltilgátan er rétt.

Úrtakstala sem er ólíkleg ef núlltilgátan er rétt er vísbending um að hún sé röng og aðaltilgátan rétt.



Myndin sýnir dreifingu 2.000 úrtaka þegar H_0 er rétt, þ.e. $\pi_{kvk} = \pi_{kk}$. Við getum reiknað bæði meðaltal og staðalfrávik yfir úrtökin 2.000 og þannig metið hversu líklegt úrtakshlutfallið (p=0,64) er undir H_0.

Fjöldi: 2.000 úrtök
Hágildi: 0,74
Lággildi: 0,28
Meðaltal: 0,501
p_{kvk}= 0,64

A:6.1, bls. 157–159

Útreikningur

$$N = 50$$

$$\pi = 0{,}5$$

$$p_{kvk} = 0{,}64$$

$$\hat{\sigma}_p = 0{,}0707$$

$$z = \frac{p_{kvk} - \pi}{\hat{\sigma}_p}$$

$$= \frac{0{,}64 - 0{,}5}{0{,}0707} = \frac{0{,}14}{0{,}0707}$$

$$= 1{,}98,\ p \approx 0{,}05$$