Grunnhugtök og lýsandi tölfræði

Kennarar diffra 2024

Anna Helga Jónsdóttir

Helstu atriði:

1. Grunnhugtök
2. Úrtakshögun
3. Lögun dreifinga
4. Lýsistærðir

Hvert erum við komin...

1. Grunnhugtök
2. Úrtakshögun
3. Lögun dreifinga
4. Lýsistærðir

Úrtak og þýði

Þýði

Þýði (population) rannsóknar er safn allra viðfangsefna sem draga á ályktanir um.

Úrtak

Úrtak (sample) er safn viðfangsefna sem eru valin úr tilteknu þýði.

- Hvert úrtak getur eingöngu verið úr einu þýði.
- Það má taka mörg úrtök úr sama þýðinu.

Flokkabreytur og talnabreytur

Breyta

Breyta (variable) er ákveðinn eiginleiki sem við skráum niður eða mælum á viðfangsefnunum í úrtakinu okkar.

Flokkabreytur

Flokkabreytur (categorical variables) taka ekki töluleg gildi heldur segja, eins og nafnið gefur til kynna, til um það hvaða flokki viðfangsefnið tilheyrir.

Talnabreytur

Talnabreytur (numerical variables) taka töluleg gildi sem eru mæld í tilteknum einingum.

Raðaðar og óraðaðar flokkabreytur

Röðuð flokkabreyta

Þegar flokkabreyta er *röðuð* (ordinal categorical variable) er flokkum hennar raðað í stærðarröð.

Óröðuð flokkabreyta

Þegar flokkabreyta er *óröðuð* (categorical variable) er flokkum hennar ekki raðað í stærðarröð.

Samfelldar og strjálar breytur.

Samfelldar breytur

Þegar talnabreyta getur tekið hvaða gildi sem er á einhverju bili þá segjum við að hún sé *samfelld* (continuous). Eingöngu talnabreytur geta verið samfelldar.

Strjálar breytur

Ef breytur eru ekki samfelldar segjum við að þær séu *strjálar* (discrete). Allar flokkabreytur eru strjálar og sumar talnabreytur.

Bjagi

Bjagi

Bjagi (bias) verður þegar aðferðirnar gefa markvisst bjagaða mynd af þýðinu sem verið er að skoða.

- Viðfangsefni valin á kerfisbundið bjagaðan hátt: **Úrtaksbjagi**.
- Góð **úrtakshögun** lágmarkar úrtaksbjaga.
- Truflandi áhrif rannsakanda og viðfangsefna: **Rannsakandabjagi** og **lyfleysuáhrif**.
- **Blindni** lágmarkar rannsakandabjaga og lyfleysuáhrif.

Hvert erum við komin...

1 Grunnhugtök

2 Úrtakshögun

3 Lögun dreifinga

4 Lýsistærðir

Slembival

Slembival

Það að velja slembið (random), eða **slembival**, þýðir að velja handahófskennt þannig að öll viðfangsefni eru jafnlíkleg til að vera valin.

Úrtak sem er valið með slembivali kallast **slembiúrtak** (random sample).

Einfalt og lagskipt slembiúrtak

Einfalt slembiúrtak

Þegar við framkvæmum **einfalt slembiúrtak** (simple random sample) veljum við einstaklinga af handahófi úr öllu þýðinu.

Lagskipt slembiúrtak

Þegar við framkvæmum **lagskipt slembiúrtak** (stratified random sample) er þýðinu fyrst skipt niður í nokkur lög eða hópa og síðan eru viðfangsefni valin með einföldu slembiúrtaki úr hverju lagi fyrir sig.

Fjöldi viðfangsefna sem valinn er úr hverju lagi verður að vera ákveðinn fyrirfram en hann má vera mismikill eftir lögum.

Gott að nota þegar fjöldi viðfangsefna í hverju lagi er mjög misstór.

Hvað ef slembiúrtak er ógerlegt?

Stundum valda erfiðleikar í framkvæmd því að við getum ómögulega valið slembiúrtak úr þýði. Þá er farin önnur af tveimur leiðum:

1. Að skilgreina þýðið upp á nýtt þannig að úrtakið verði slembiúrtak.
   - Þá eiga ályktanir okkar eingöngu við um „nýja þýðið". Viljum við það?
2. Að sætta sig við bjagann.
   - Gerum grein fyrir úrtaksbjaganum í umfjöllun okkar.
   - Ræðum ítarlega hvaða afleiðingar hann getur haft í för með sér.
   - Er hægt að gera ráð fyrir að bjaginn sé léttvægur með tilliti til þess sem við erum að rannsaka?

Sjálfboðaliðaúrtök - VARÚÐ!

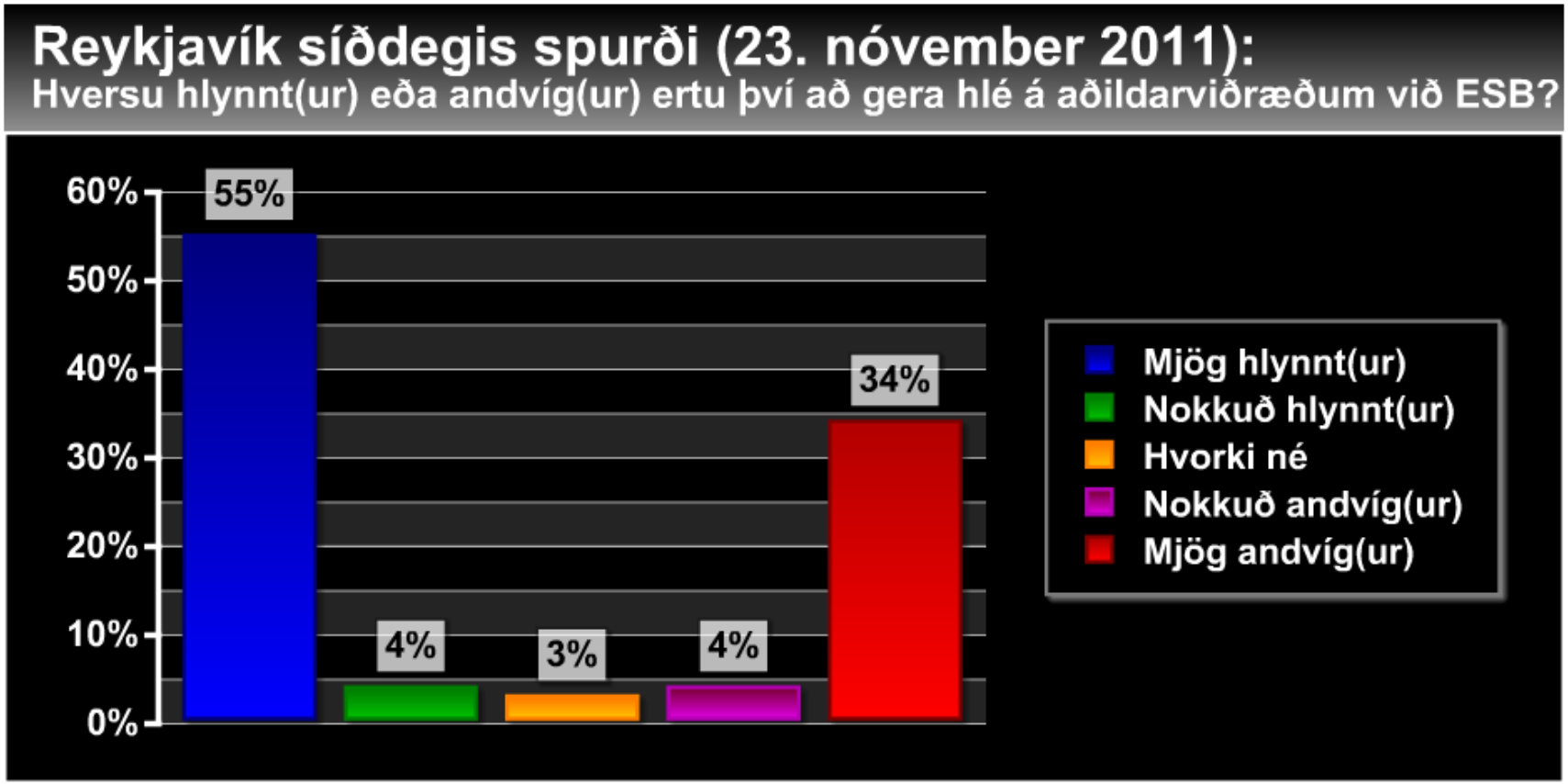
Sjálfboðaliðaúrtök

Sjálboðaliðaúrtök eiga við þegar viðfangsefnin eru fólk og þá eru eingöngu framkvæmdar mælingar á þeim sem bjóða sig fram til þess.

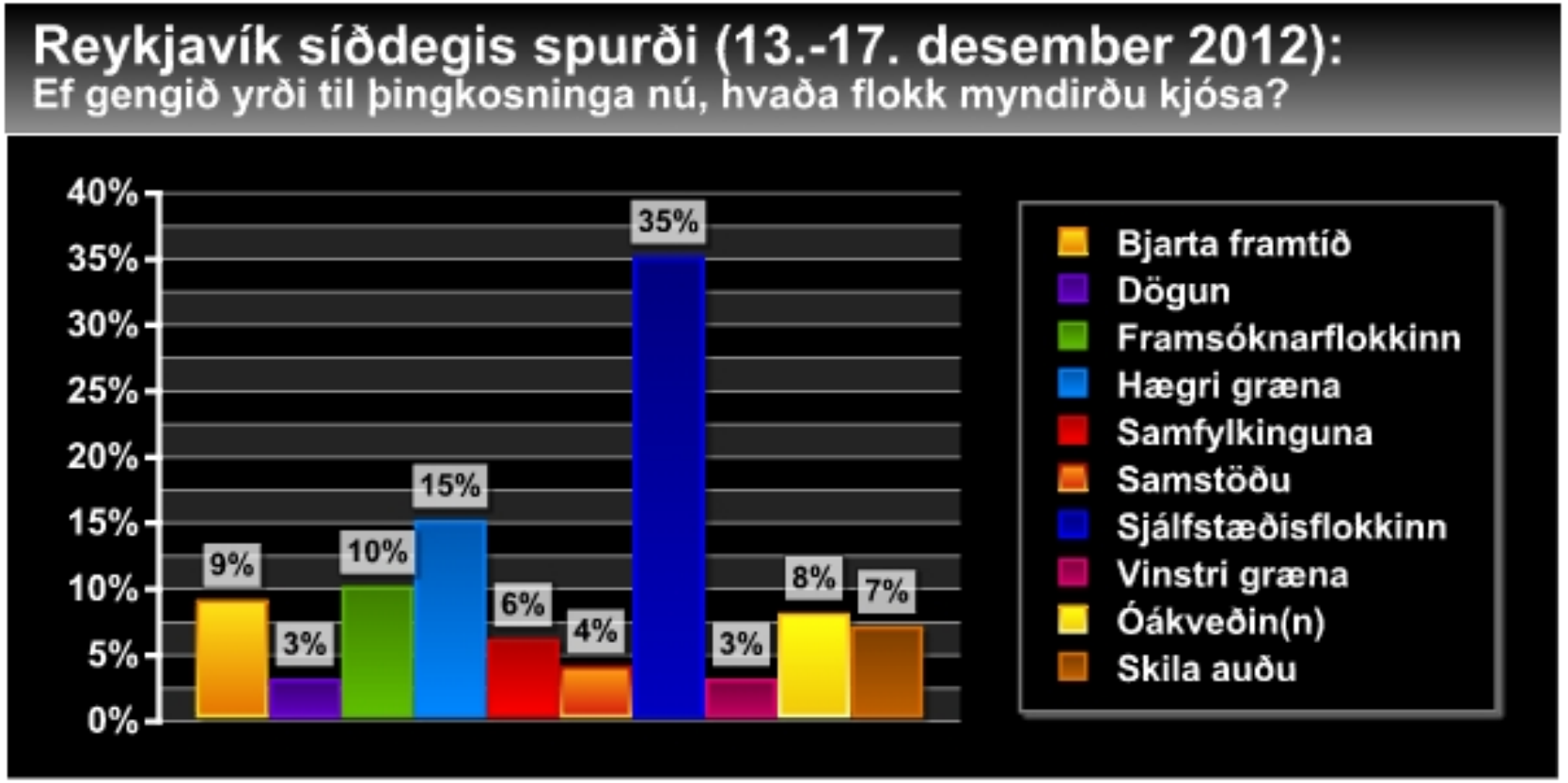
Hér verður úrtaksbjagi vegna þess að ákveðin viðfangsefni geta verið líklegri til að bjóða sig fram en önnur.

Oft getur sá bjagi orðið svo mikill að lítið er hægt að álykta út frá þeim mælingum sem fengnar eru.

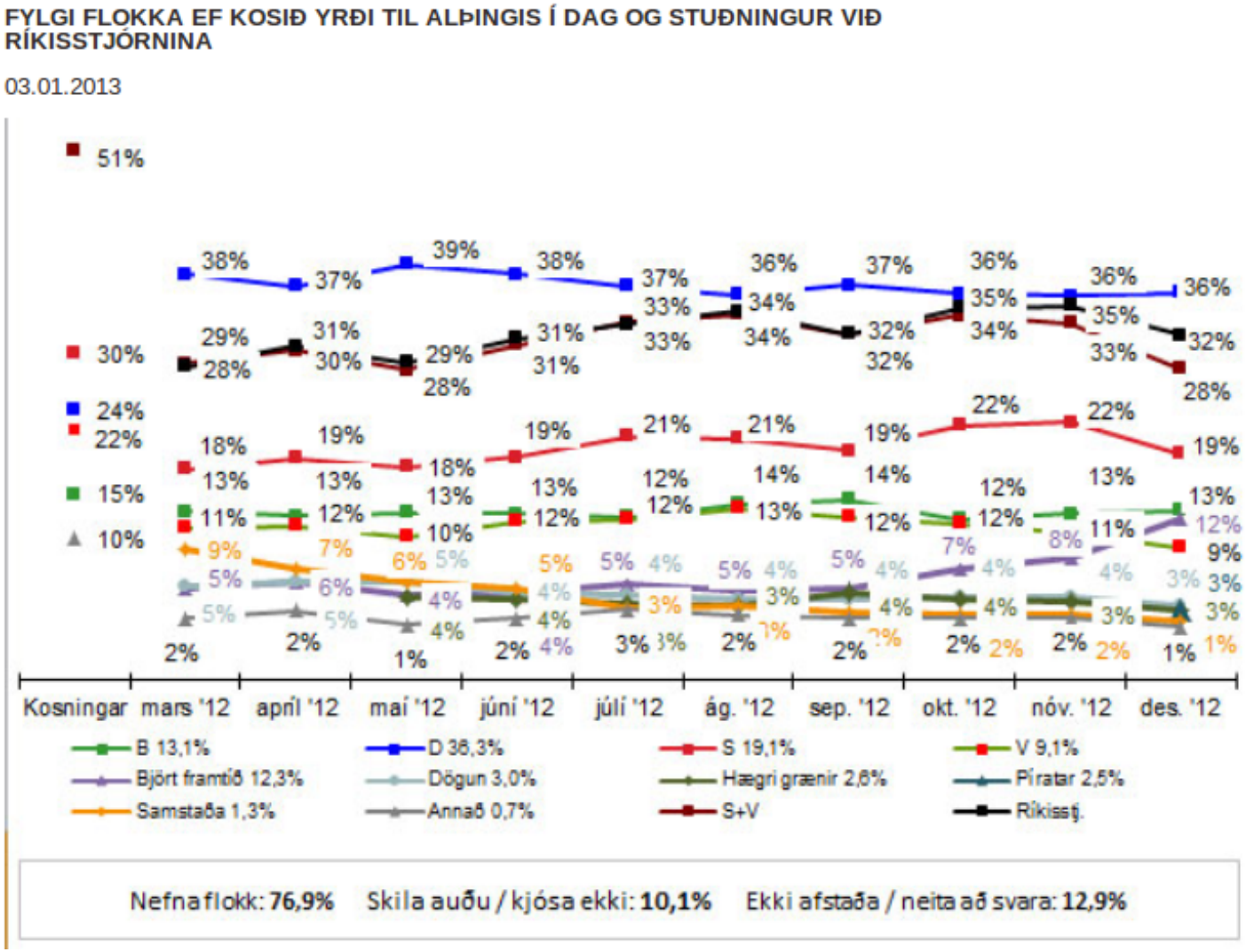
Sjálfboðaliðaúrtak



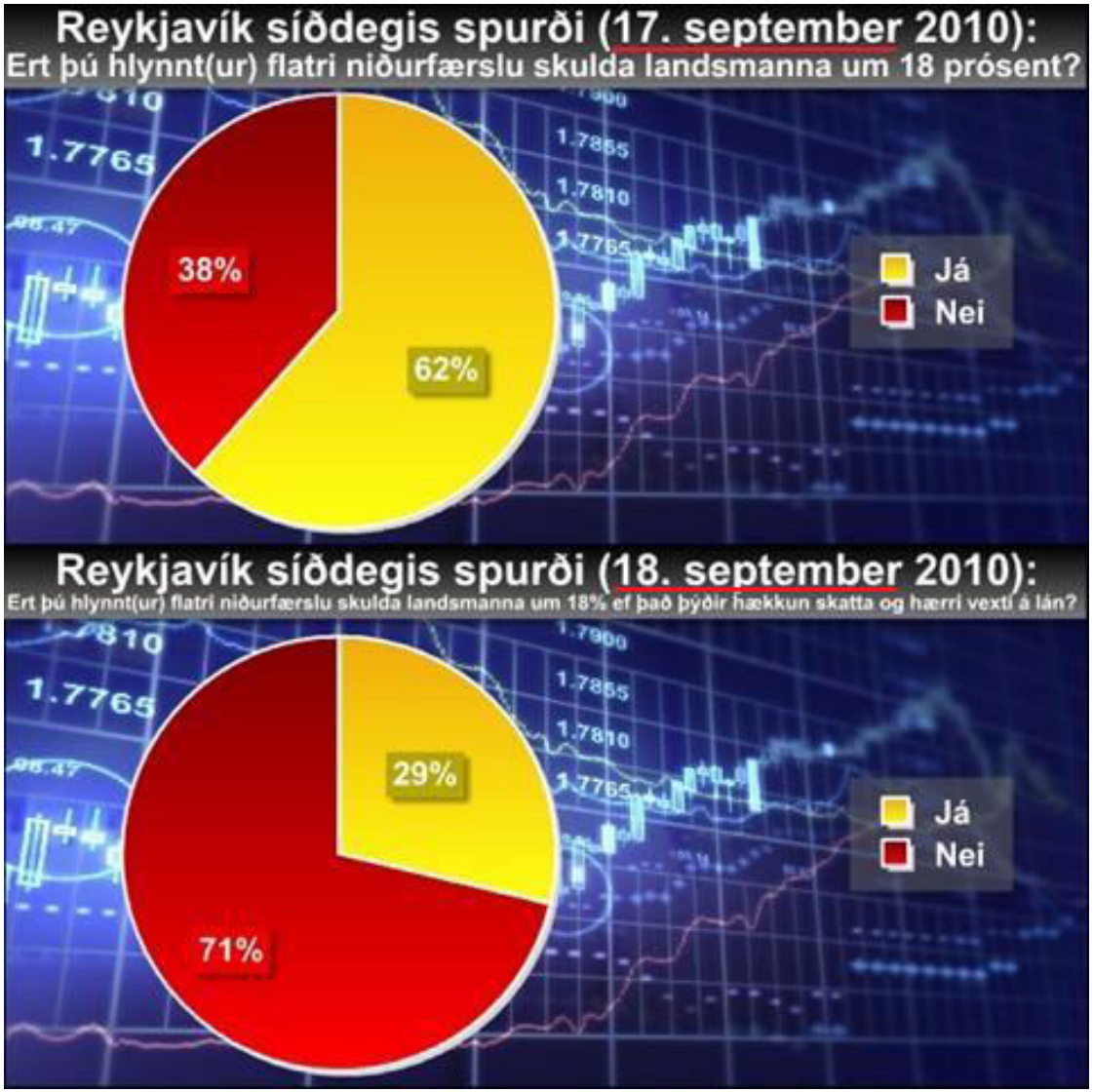
Sjálfboðaliðaúrtak - slembiúrtak



Sjálfboðaliðaúrtak - slembiúrtak



Orðalag


Reykjavík síðdegis spurði (17. september 2010):
Ert þú hlynnt(ur) flatri niðurfærslu skulda landsmanna um 18 prósent?
38%
62%
Já
Nei
Reykjavík síðdegis spurði (18. september 2010):
Ert þú hlynnt(ur) flatri niðurfærslu skulda landsmanna um 18% ef það þýðir hækkun skatta og hærri vexti á lán?
29%
71%
Já
Nei

Hvert erum við komin...

1 Grunnhugtök

2 Úrtakshögun

3 Lögun dreifinga

4 Lýsistærðir

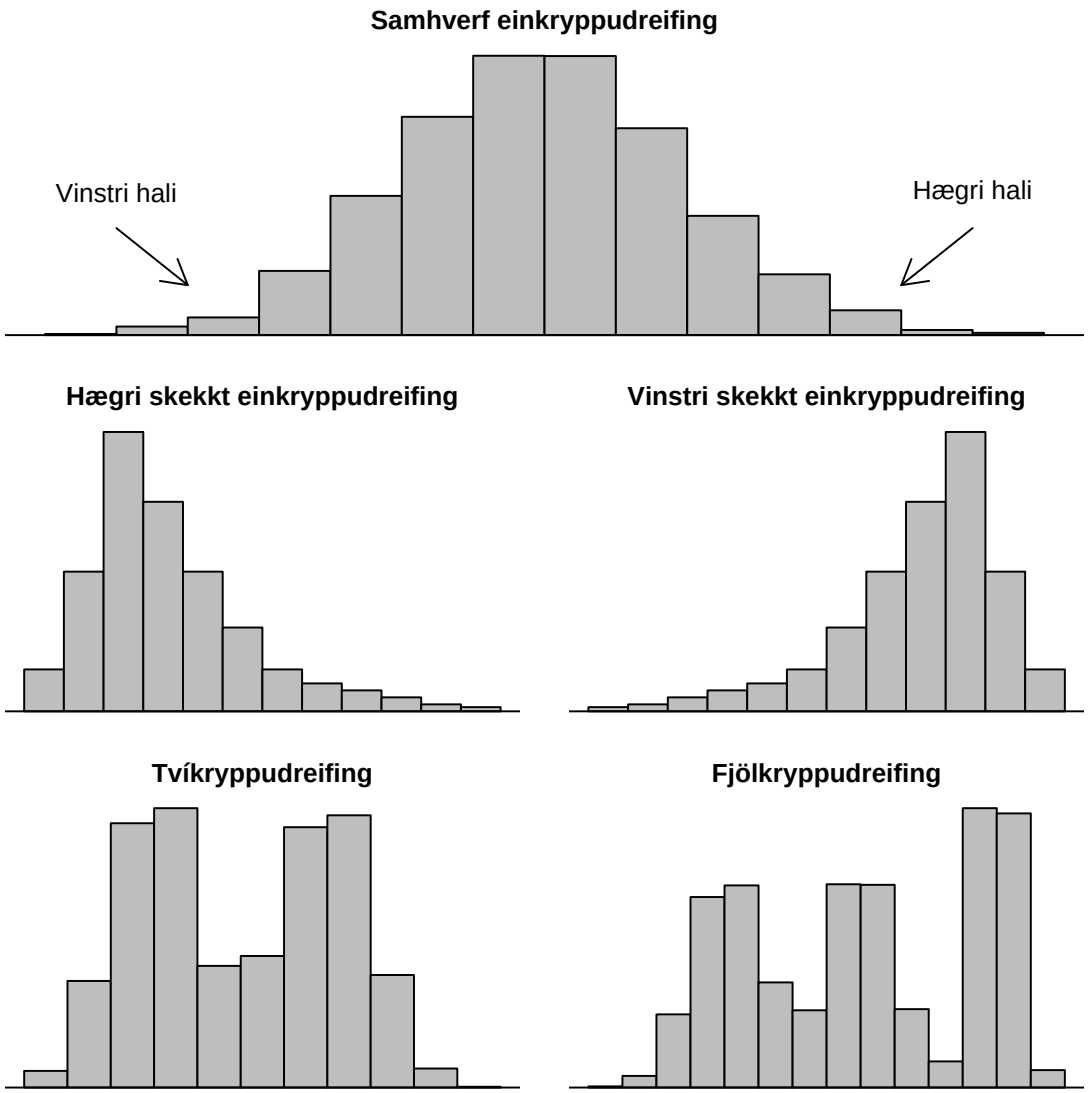
Lögun dreifinga

Lögun dreifinga (Shape of distributions)

Eftirfarandi hugtök eru oft notuð til að lýsa dreifingum mælinga.

- Dreifingu minnstu mælinganna köllum við *vinstri hala* (left-tail) dreifingarinnar. Dreifingu stærstu mælinganna köllum við *hægri hala* (right-tail) dreifingarinnar.
- Dreifing er *samhverf* (symmetric) ef hægri hlið hennar dreifist eins og spegilmynd vinstri hliðarinnar.
- Dreifing sem ekki er samhverf er *skekkt* (skewed). Dreifing er *skekkt til hægri* (skewed to the right) ef hægri hali hennar en lengri en sá vinstri og *skekkt til vinstri* (skewed to the left) ef sá vinstri er lengri en sá hægri.
- Ef dreifingin hefur einn topp er talað um *einkryppudreifingu* (unimodal).
- Ef dreifingin hefur tvo toppa er talað um *tvíkryppudreifingu* (bimodal).
- Ef dreifing hefur fleiri en tvo toppa er talað um *fjölkryppudreifingu* (multimodal).

Lögun dreifinga



Hvert erum við komin...

1 Grunnhugtök

2 Úrtakshögun

3 Lögun dreifinga

4 Lýsistærðir

Lýsistærðir

- Við notum lýsandi tölfræði til að lýsa tilteknum eiginleikum mælinganna okkar.
- Góð aðferð til þess er að nota **lýsistærðir** en þær eru tölur sem lýsa tilteknum eiginleikum mælinga

Lýsistærð (statistic)

Lýsistærð er tala sem er reiknuð með einhverjum ákveðnum hætti út frá mælingunum okkar

Lýsistærðir

- Til eru margar gerðir af lýsistærðum, sem lýsa ólíkum eiginleikum mælinga
- Skoðum nú tvær tegundir lýsistærða fyrir talnabreytur:
  - Lýsistærðir sem lýsa **miðju** (center) gagna
  - Lýsistærðir sem lýsa **breytileika** (spread) gagna
- Algengastar eru meðaltal og staðalfrávik

Lýsistærðir fyrir miðju

Skoðum þrjár mismunandi lýsistærðir sem allar lýsa miðju:

1. Tíðasta gildi (mode)
2. Miðgildi (median)
3. Meðaltal (mean, arithmetic mean)

Tíðasta gildi

Tíðasta gildi (mode)

Gerum ráð fyrir að við höfum n mælingar $x_1, x_2, ...x_n$. **Tíðasta gildið** er sú útkoma sem oftast kemur fyrir í mælingunum okkar. Það er sú eina af lýsistærðunum fyrir miðju sem við fjöllum um sem er hægt er að nota til að lýsa flokkabreytum. Hins vegar er ekki við hæfi að reikna tíðasta gildið þegar mældar eru samfelldar talnabreytur.

Miðgildi

Miðgildi (median)

Gerum ráð fyrir að við höfum n mælingar $x_1, x_2, ...x_n$. Byrjum á að raða þessum mælingum upp í stærðarröð, frá minnst gildi upp í stærsta gildi. Reiknum svo

$$\text{Sæti í röð} = 0.5 \cdot (n+1).$$

Miðgildi er oft táknað með M. Það fer eftir því hvort n sé oddatala eða slétt tala hvernig við reiknum út miðgildið.

- Ef n er oddatala þá er miðgildið staðsett í sæti $0.5 \cdot (n+1)$ í röðinni.
- Ef n er slétt tala þá er miðgildið meðaltalið af þeim tveimur mælingum sem standa við sæti $0.5 \cdot (n+1)$ í röðinni.

VARÚÐ: $0.5 \cdot (n+1)$ er númerið á sætinu í röðinni, ekki miðgildið sjálft!

Meðaltal

Meðaltal (mean, arithmetic mean)

Gerum ráð fyrir að við höfum n mælingar $x_1, x_2, ...x_n$. **Meðaltalið** fæst með að leggja mælingarnar saman og deila með fjölda mælinga.

$$\bar{x} = \frac{x_1 + x_2 + ... + x_n}{n} = \frac{\sum_{i=1}^{n} x_i}{n}$$

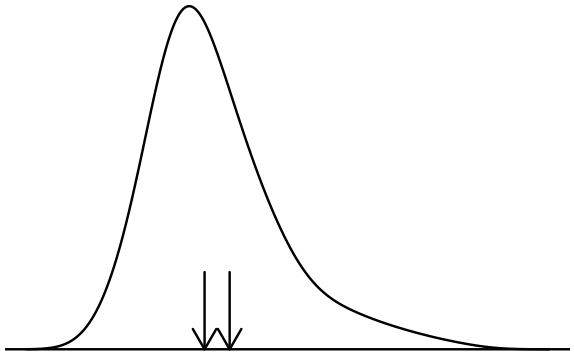
Samanburður á lýsistærðum fyrir miðju

- Áður en ákvörðun er tekin um hvaða lýsistærð skal nota til að lýsa miðju gagnanna er gott að skoða gögnin myndrænt til að átta sig á dreifingu gagnanna
- Sé dreifingin skekkt, tvíkryppu- eða fjölkryppu dreifing skal nota miðgildið fram yfir meðaltalið
- Miðgildi er betri mælikvarði á miðju gagna ef útlagar eru í gagnasafninu

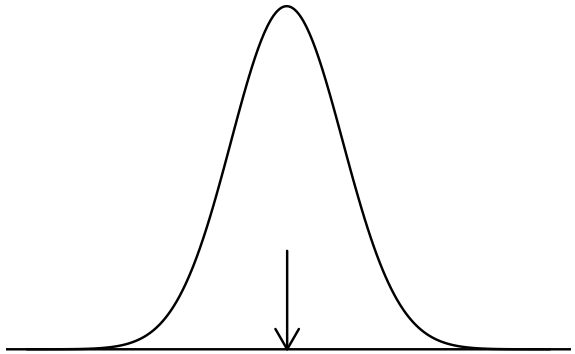
Samanburður á meðaltali og miðgildi

- Ef dreifingin er skekkt til hægri er meðaltalið hærra en miðgildið.
- Ef dreifingin er samhverf er meðaltalið og miðgildið það sama.
- Ef dreifingin er skekkt til vinstri er meðaltalið lægra en miðgildið.

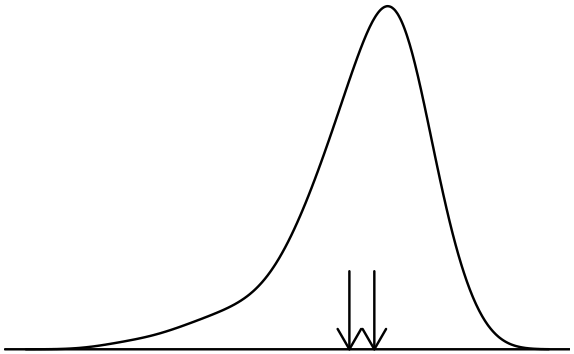
Hægri skekkt dreifing



Samhverf dreifing



Vinstri skekkt dreifing



Lýsistærðir fyrir breytileika

Skoðum sex lýsistærðir sem allar lýsa breytileika

1. Spönn (range)
2. Fjórðungamörk (quartiles)
3. Fjórðungaspönn (interquartile range)
4. Prósentumörk (percentiles)
5. Dreifni/fervik (variance)
6. Staðalfrávik (standard deviation)

Spönn/dreifisvið

Spönn (range)

Gerum ráð fyrir að við höfum n mælingar $x_1, x_2, ...x_n$ og látum x_{min} tákna þá minnstu og x_{max} þá stærstu. **Spönn** gagnanna er reiknuð með

$$\text{Spönn} = x_{\text{max}} - x_{\text{min}}$$

Fjórðungamörk

Fjórðungamörkin eru þrjú og er algengt að kalla þau, Q_1, Q_2 og Q_3. Í sumum kennslubókum og ritum eru fjórðungamörkin kölluð $Q_{25\%}, Q_{50\%}$ og $Q_{75\%}$. Við munum halda okkur við fyrri ritháttinn í þessari bók.

Q_1: Um fyrsta fjórðungamarkið gildir að 25% af mælingunum eru lægri en Q_1. Q_1 er því miðgildi neðri helmings mælinganna, að undanskildu miðgildinu.

Q_2: Um annað fjórðungamarkið gildir að 50% af mælingunum eru lægri en Q_2. Q_2 er því miðgildið, $Q_2 = M$.

Q_3: Um þriðja fjórðungamarkið gildir að 75% af mælingunum eru lægri en Q_3. Q_3 er því miðgildi efri helmings mælinganna, að undanskildu miðgildinu.

Fjórðungamörk

Fjórðungamörk (quartiles)

Gerum ráð fyrir að við höfum n mælingar $x_1, x_2, ...x_n$. Byrjum á að raða mælingum upp í stærðarröð, frá lægsta gildinu upp í hæsta gildið. Reiknum svo

$$Q_1 \text{ - sæti í röð: } = 0.25 \cdot (n+1)$$
$$Q_2 \text{ - sæti í röð: } = 0.50 \cdot (n+1)$$
$$Q_3 \text{ - sæti í röð: } = 0.75 \cdot (n+1)$$

- Q_1 er mælingin sem stendur í sæti $0.25 \cdot (n+1)$ eða meðaltalið af þeim tveimur mælingum sem standa við sæti $0.25 \cdot (n+1)$ í röðinni.
- Q_2 er mælingin sem stendur í sæti $0.50 \cdot (n+1)$ eða meðaltalið af þeim tveimur mælingum sem standa við sæti $0.50 \cdot (n+1)$ í röðinni.
- Q_3 er mælingin sem stendur í sæti $0.75 \cdot (n+1)$ eða meðaltalið af þeim tveimur mælingum sem standa við sæti $0.75 \cdot (n+1)$ í röðinni.

Fjórðungaspönn

Fjórðungaspönn (interquartile range)

Gerum ráð fyrir að við höfum n mælingar $x_1, x_2, ...x_n$ og látum Q_1 tákna fyrsta fjórðungamark og Q_3 þriðja fjórðungamark. **Fjórðungaspönn** gagnanna er táknuð með IQR og reiknuð með

$$IQR = Q_3 - Q_1.$$

Prósentumörk

Hugmyndin að baki *prósentumörkum* (percentiles) er svipuð og sú að baki fjórðungamörkum nema í stað þess að skoða eingöngu mörkin við 25 %, 50 % eða 75 % mælinganna getum við leyft hvaða hlutfall sem er.

Prósentumörk (percentiles)

Með $a\%$ prósentumörkum er átt við þá tölu sem er þannig að $a\%$ mælinganna hafa lægra gildi en sú tala.

Líkt og með fjórðungamörkin eru nokkrar ólíkar leiðir til þess að reikna prósentumörk og er það nær aldrei gert „ í höndunum" heldur er notast við tölfræðihugbúnað.

Dreifni

Dreifni (variance)

Gerum ráð fyrir að við höfum n mælingar $x_1, x_2, ...x_n$. Dreifni mælinga er táknuð s^2 og er reiknuð með

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

$s^2 = 0$ þá og því aðeins að allar mælingarnar séu jafnar, annars er s^2 ávallt stærra en 0. Því lengra sem mælingarnar liggja frá meðaltalinu því hærra verður s^2.

Staðalfrávik

Staðalfrávik (standard deviation)

Gerum ráð fyrir að við höfum n mælingar $x_1, x_2, ...x_n$. Staðalfrávik mælinga er táknað með s og er reiknað með

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

$s = 0$ þá og því aðeins að allar mælingarnar eru jafnar, annars er s ávallt stærra en 0. Því lengra sem mælingarnar liggja frá meðaltalinu því hærra verður s.

Samanburður á lýsistærðum fyrir breytileika

- Allar þær lýsistærðir sem við höfum skoðað sem lýsa breytileika eru áhugaverðar og gott að reikna út þegar við fáum ný gögn í hendurnar
- Dreifni og staðalfrávik eru notuð til að lýsa breytileika mælinga umhverfis meðaltalið
- Staðalfrávik er yfirleitt notað fram yfir dreifni þar sem mælieiningin á staðalfrávikinu er sú sama og á mælingunum
- Staðalfrávik er viðkvæmt fyrir skekkingu og útlögum. Aðeins fáir útlagar geta gert staðalfrávikið mjög hátt
- Séu mælingarnar skekktar eða ef útlagar eru til staðar er fimm tölu samantekt og kassarit besti mælikvarðinn á breytileika gagnanna