Myndræn framsetning

Tölfræði frá grunni

Anna Helga Jónsdóttir og Sigrún Helga Lund

Háskóli Íslands

HÁSKÓLI ÍSLANDS

Helstu atriði:

Myndræn framsetning

- Fyrsta skref í tölfræðiúrvinnslu ætti ætíð að vera að skoða gögnin myndrænt
- Kjarni tölfræðiúrvinnslu er að átta sig sem best á eðli mælinganna sem skoðaðar eru
- Breytileiki gagnanna lykilatriði - hvernig er dreifing gagnanna?
  - Hversu mikinn mun sjáum við á útkomum viðfangsefnanna okkar?
  - Hvernig dreifast útkomurnar?
- Myndræn framsetning er ein besta leiðin til að átta sig á dreifingu mælinganna

Yfirlit

Myndræn framsetning á strjálum breytum

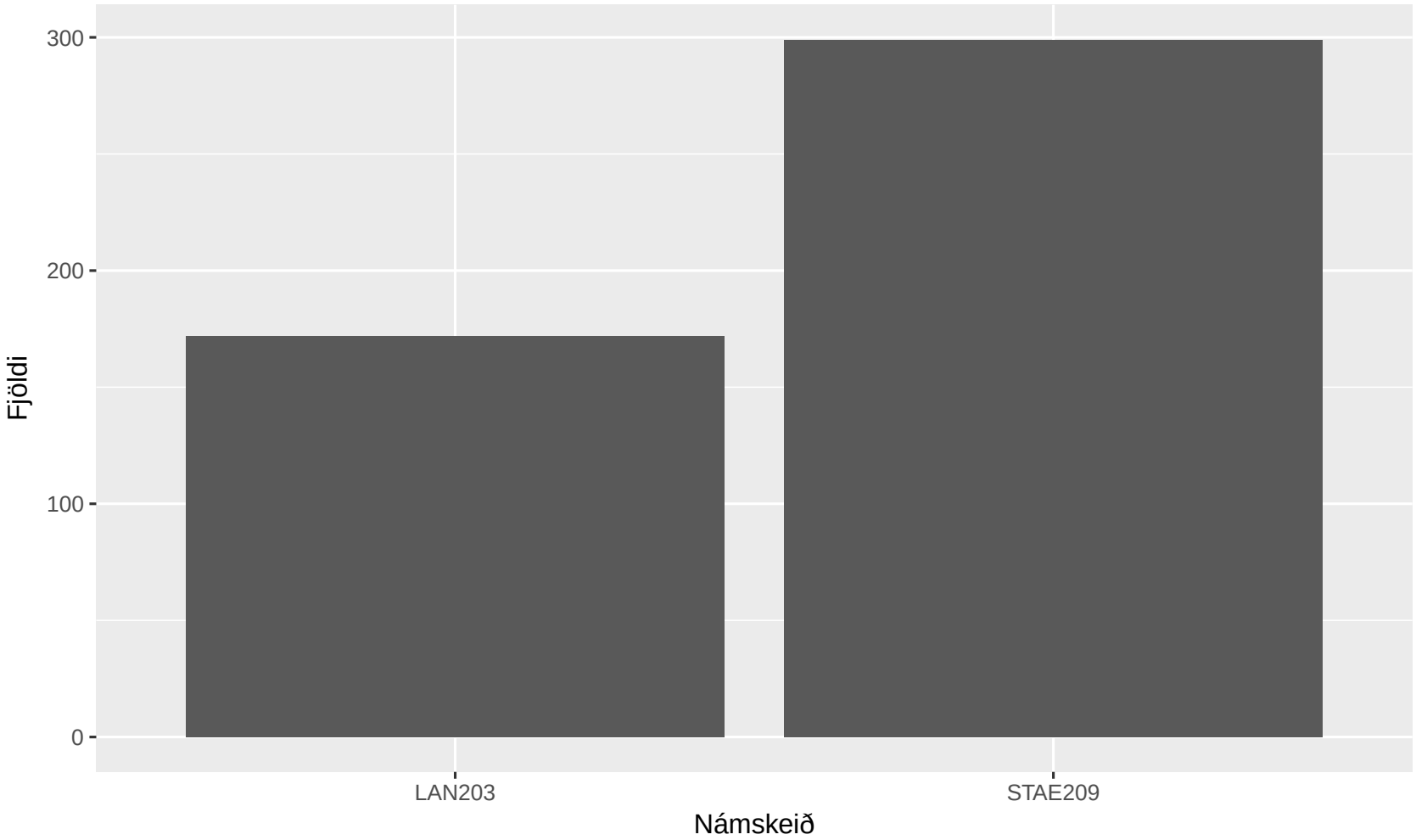
- Algengustu tegundir grafa fyrir strjálar breytur eru **stöplarit** (bar chart) og **kökurit** (pie chart)
- Kökurit eru mikið notuð í viðskiptaheiminum og í fjölmiðlum en eru sjaldséð í tímaritum og bókum um raunvísindi
- Stöplarit má sjá víðsvegar og eru þau í nánast öllum tilfellum betur til þess fallin að sýna gildi strjálla breyta myndrænt en kökuritin

Stöplarit

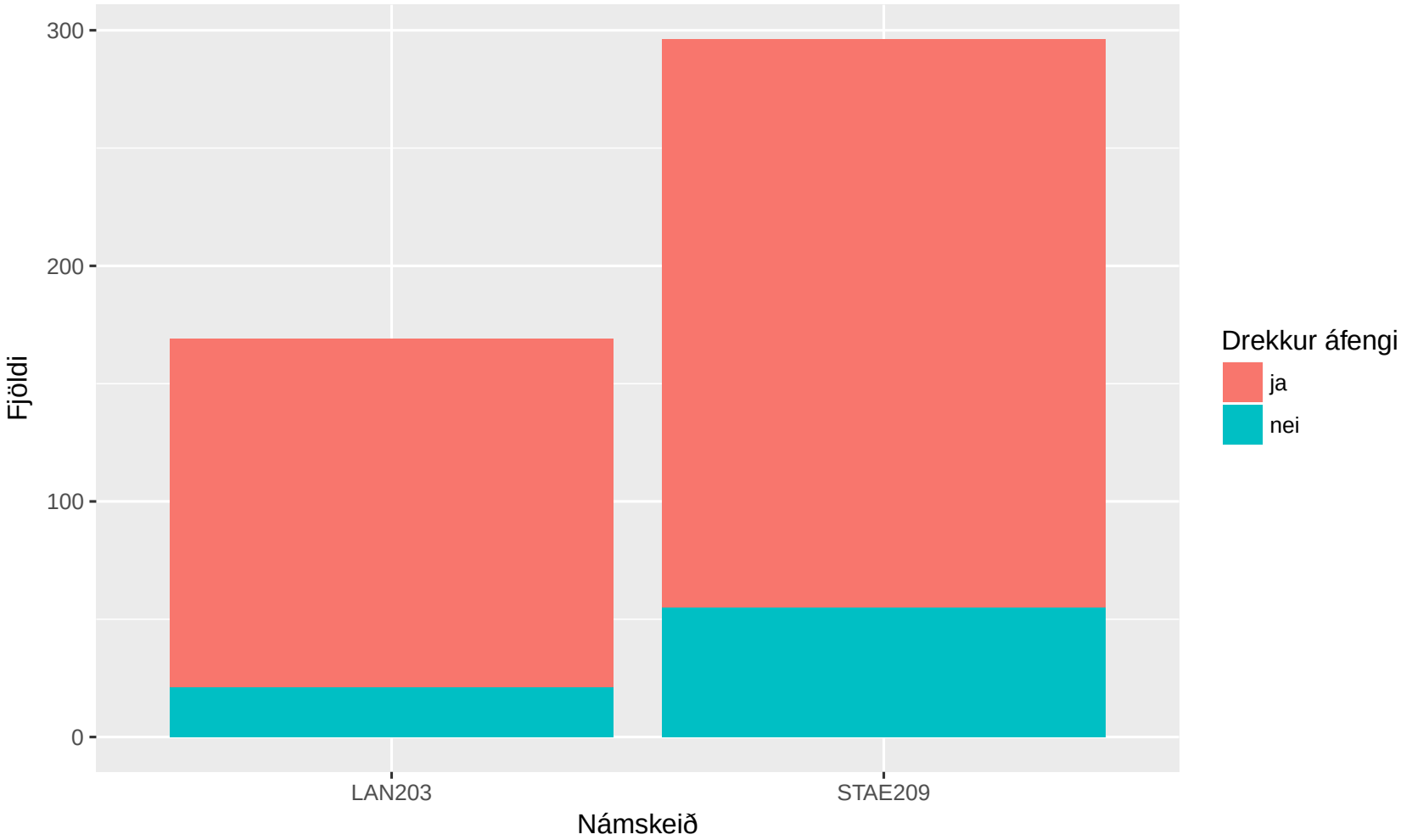
Stöplarit (bar chart)

Stöplarit samanstanda af tveimur eða fleiri súlum. Fjöldi súlna ræðst af fjölda flokka/gilda breytunnar. Hver súla stendur fyrir einn flokk og mega þær ekki liggja hvor að annarri. Hæð súlnanna sýnir tíðni eða hlutfall fyrir viðkomandi flokk. Raða skal súlunum svo auðvelt sé að greina upplýsingarnar, oft er þeim raðað upp eftir stærð. Y-ásinn þarf ávalt að byrja í núlli.

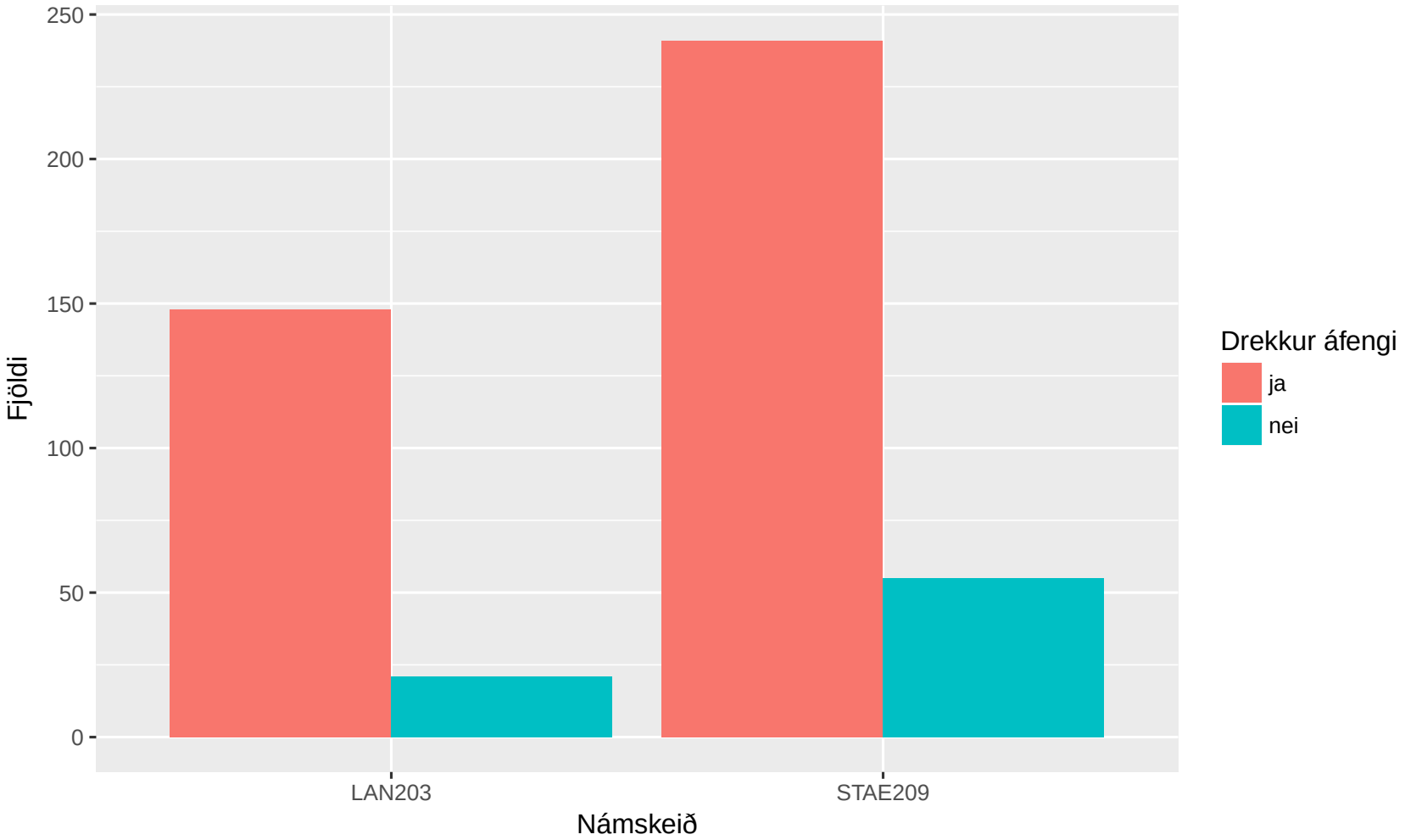
Stöplarit



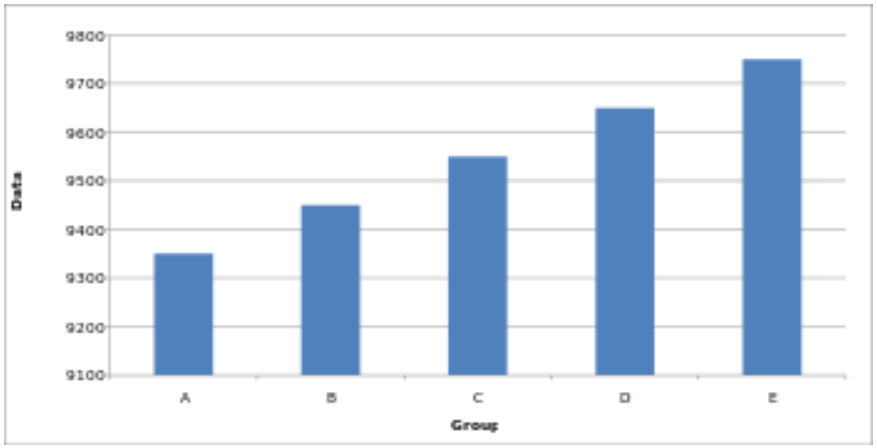
Stöplarit


300
200
100
0
Fjöldi
LAN203
STAE209
Námskeið
Drekkur áfengi
ja
nei

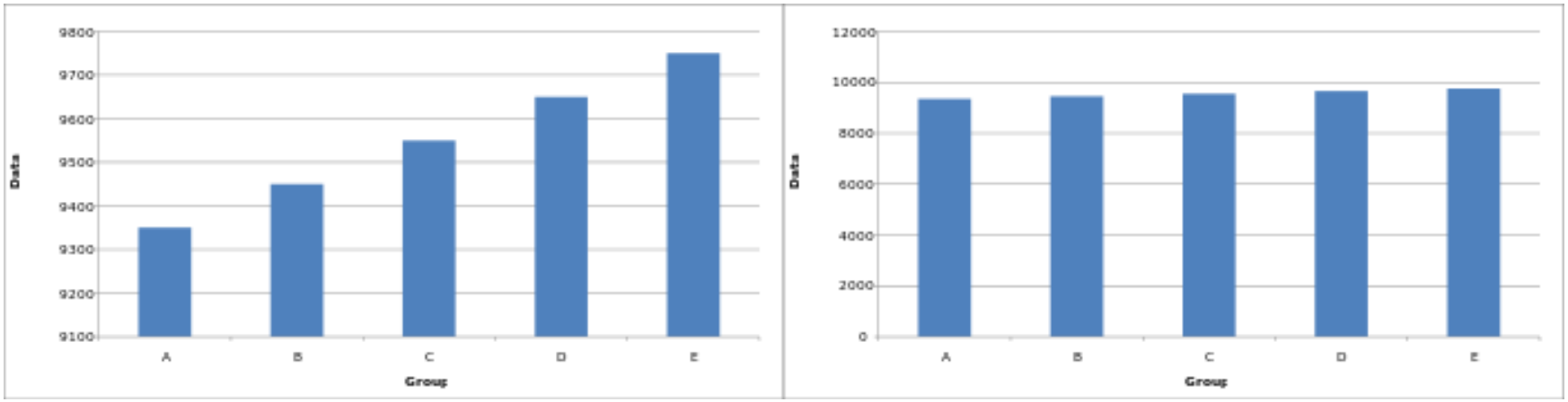
Stöplarit


250
200
150
100
50
0
Fjöldi
LAN203
STAE209
Námskeið
Drekkur áfengi
ja
nei

Algengt "blekkingartrix"stöplarit



Algengt "blekkingartrix"stöplarit



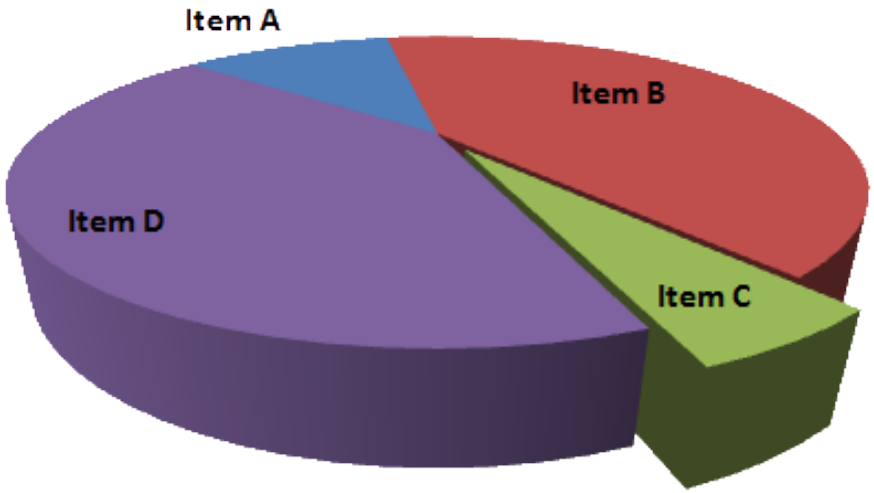
"Truncated Bar Graph"by Smallman12q - Own work. Licensed under CC0 via Wikimedia Commons

Kökurit

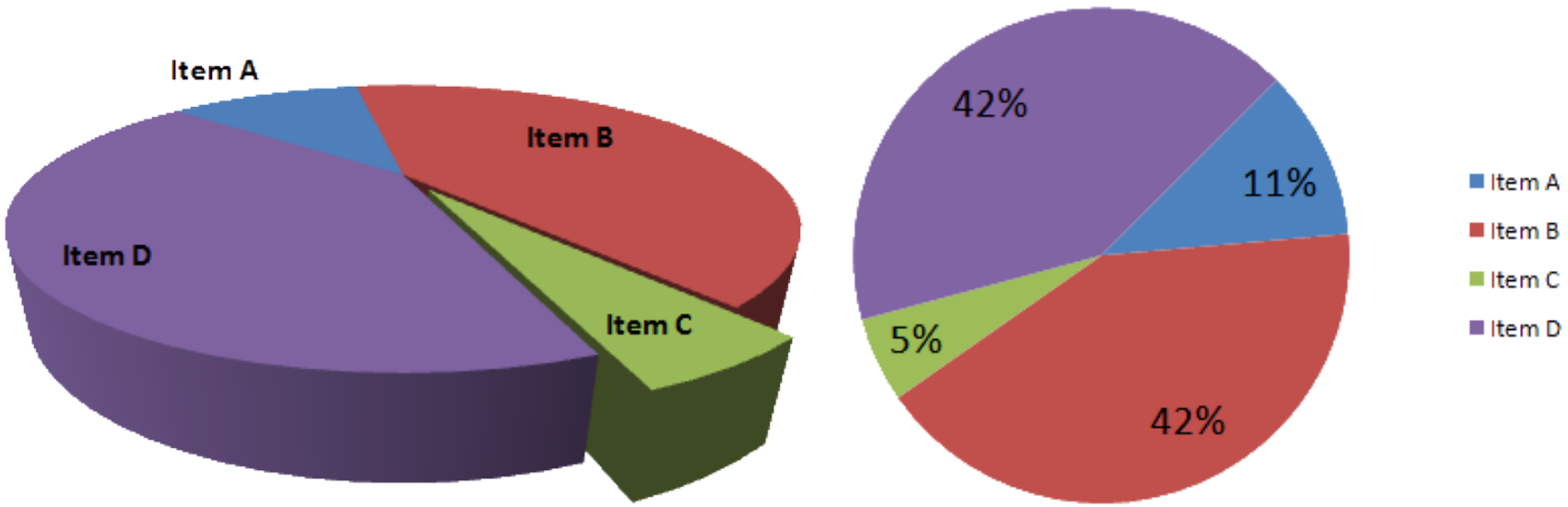
Kökurit (pie chart)

Þegar búa á til kökurit er mikilvægt að allir flokkar/gildi breytunnar sem verið er að skoða séu með á myndinni. Fjöldi sneiða í kökuritinu ræðst af fjölda flokka/gilda breytunnar. Stærð sneiðarinnar ræðst af hlutfallslegum fjölda í viðkomandi flokki af heildinni. Gætið að hlutföllin séu samanlögð 100%.

Algengt "blekkingartrix"kökurit



Algengt "blekkingartrix"kökurit



"Misleading Pie Chart"by Smallman12q - Own work. Licensed under CC0 via Wikimedia Commons

Yfirlit

Stuðlarit

- Algengasta aðferðin til að skoða samfellda breytu myndrænt er stuðlarit.
- Kassarit eru einnig góð aðferð til að skoða samfelldar breytur og munum við kynnast þeim í kaflanum um lýsandi tölfræði.
- Viljum við skoða samband tveggja talnabreyta notum við punktarit (scatter plot).

Stuðlarit

- Stuðlarit er svipað stöplariti í útliti en helsti munur á útliti þeirra er að ekkert bil er á milli súlnanna í stuðlariti
  - Gott er að hugsa sér stuðlaberg til að muna hvort bil eigi að vera á milli súlnanna í stuðlariti!
- Snúnara að búa til stuðlarit en stöplarit þar sem samfelldar breytur innihalda ekki eiginlega flokka
- Byrja á að mynda flokka áður en talið er hversu margar mælingar falla í hvern flokk

Stuðlarit

Stuðlarit (histogram)

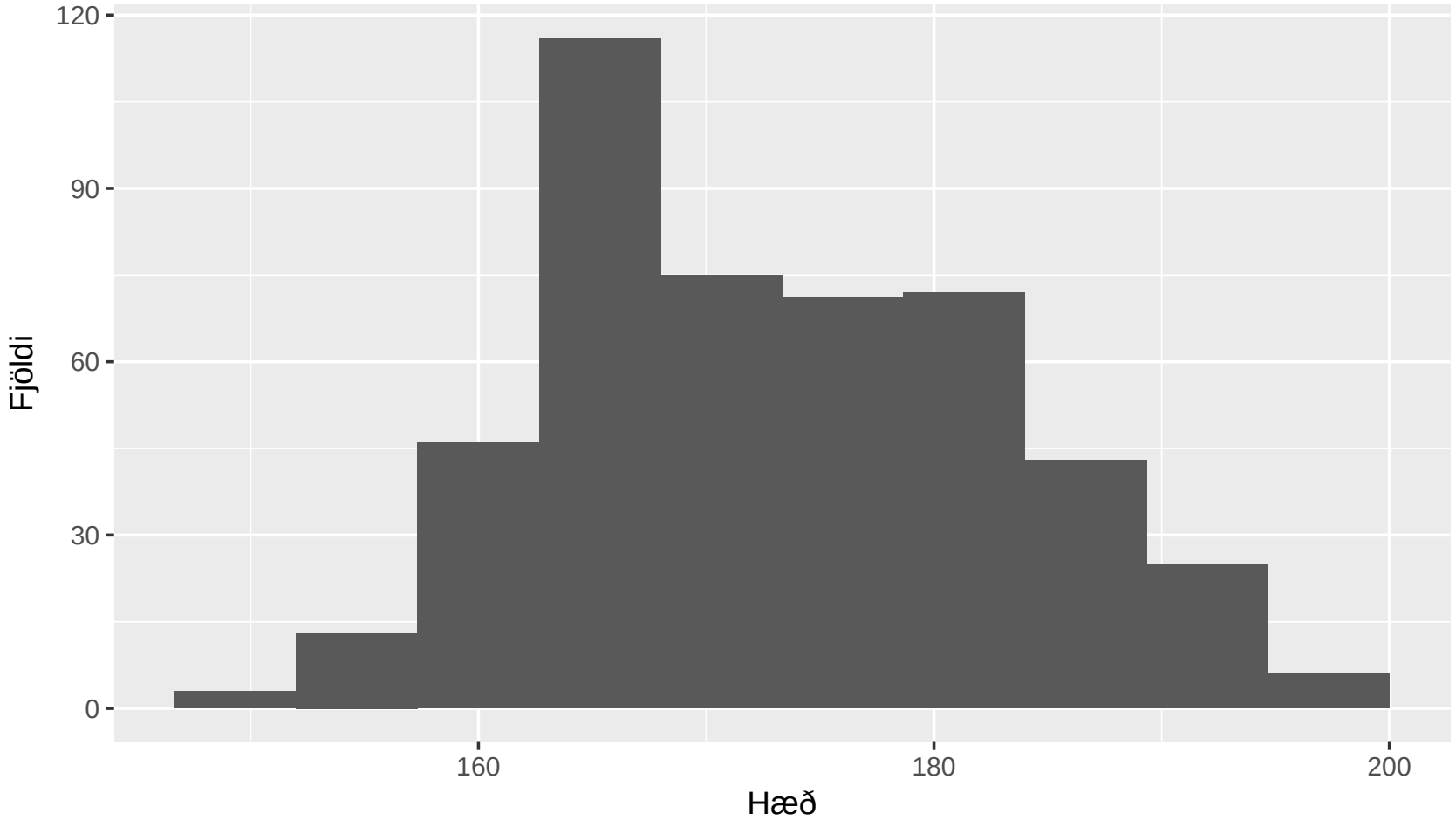
Stuðlarit samanstendur af súlum sem standa hvor upp að annari. Fjöldi súlna ræðst af fjölda flokka sem samfelldu breytunni er skipt upp í. Þegar flokkarnir eru myndaðir er gott að hafa eftirfarandi í huga.

- Neðri og efri mörk eiga að vera einföld og auðskilin
- Bilin mega ekki skarast og verða að ná yfir allar mælingar
- Bilin eiga að vera jafn breið
- Flokkarnir eiga að vera hæfilega margir. Engin ein rétt lausn er til en ágætt er að nota þumalputtaregluna að fjöldi flokka á að vera u.þ.b 5 sinnum logaritminn af fjölda mælinga,

$$\text{fjöldi flokka} = 5 \cdot \log(\text{fjöldi mælinga})$$

Þegar flokkarnir hafa verið myndaðir er teiknuð ein súla fyrir hvern flokk og ræðst hæð súlunnar af fjölda (eða hlutfalli) mælinga í þeim flokki.

Stuðlarit



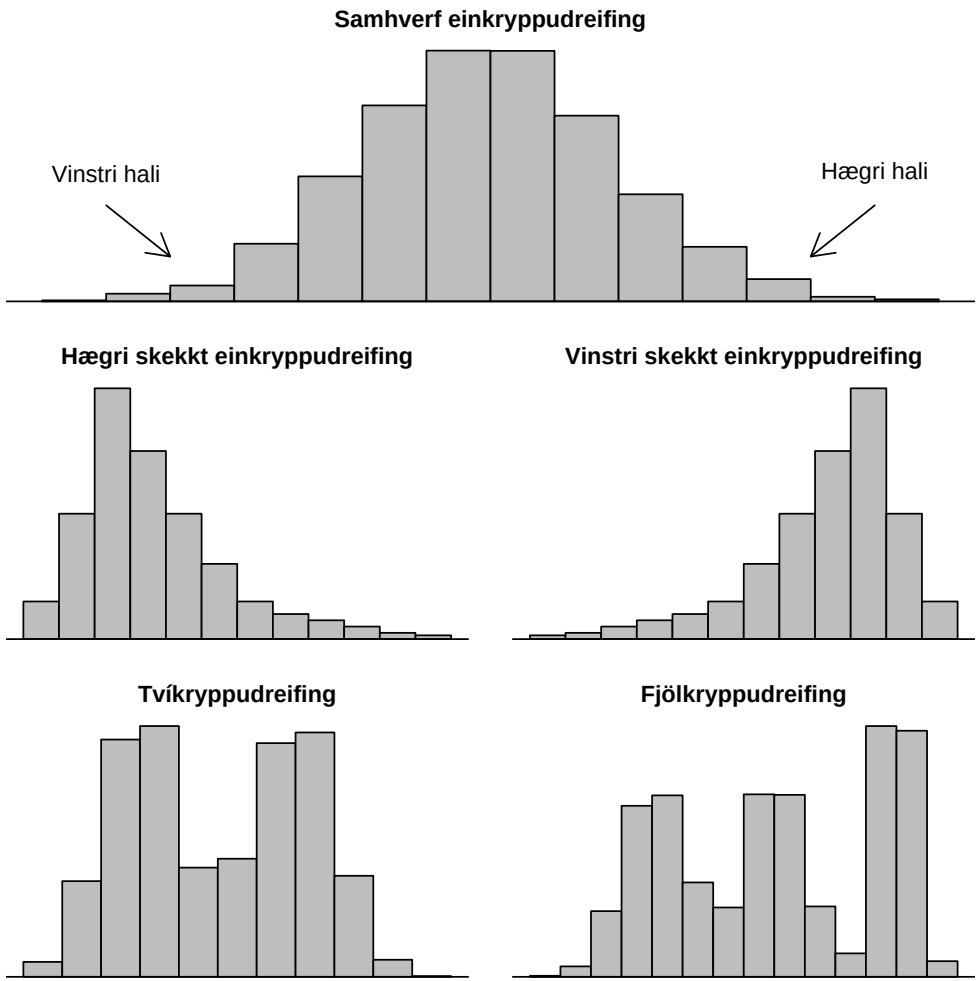
Lögun dreifinga

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).

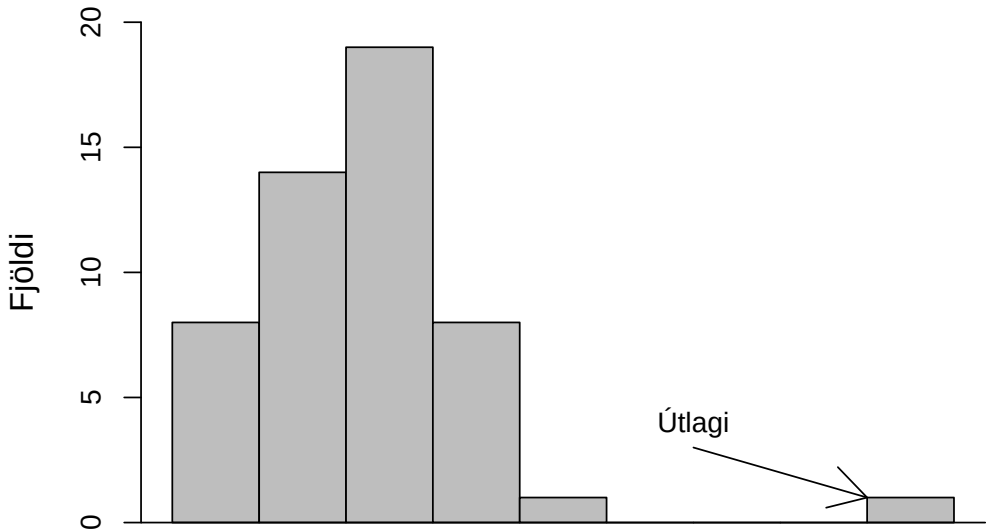
Lögun dreifinga



Útlagar

Útlagar (Outliers)

Útlagar eru mæligildi sem eru mjög ólík öðrum mæligildum í sama gagnasafni. Ýmsar ástæður geta verið fyrir útlögum og er mjög mikilvægt að skoða þá sérstaklega og hugleiða ástæðu þeirra.



Yfirlit

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.

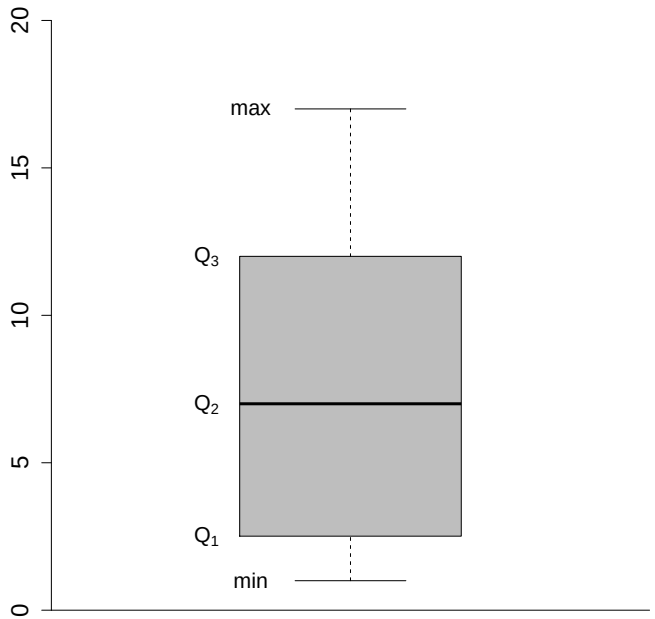
Kassarit

Kassarit (Box-plot)

- Kassarit samanstendur af kassa og tveimur línum sem ganga út frá endum kassans. Þessar línur eru oft kallaðar skegg (whiskers).
- Kassinn má liggja (láréttur) eða standa (lóðréttur). Þá skal y-ásinn hafa gildi sem nær frá neðsta gildi gagnasafnsins (eða rétt þar fyrir neðan) og upp í hæsta gildi gagnasafnsins (eða rétt þar fyrir ofan).
- Neðri endi kassans skal standa í Q_1 og efri hluti kassans í Q_3. Draga skal línu í gegnum kassann í Q_2.

Kassarit

- **Kassarit** er notað til að skoða miðju og breytileika mælinga.
- Endurspegla vel gögnin og sýna glöggt hvort dreifingin er samhverf eða skekkt.
- Til eru nokkrar útfærslur. Útgáfan sem við skoðum hér er sú einfaldasta



Kassarit

Kassarit (Box-plot)

- Kassarit samanstendur af kassa og tveimur línum sem ganga út frá endum kassans. Þessar línur eru oft kallaðar skegg (whiskers).
- Kassinn má liggja (láréttur) eða standa (lóðréttur), við látum kassann standa. Þá skal y-ásinn hafa gildi sem nær frá neðsta gildi gagnasafnsins (eða rétt þar fyrir neðan) og upp í hæsta gildi gagnasafnsins (eða rétt þar fyrir ofan).
- Neðri endi kassans skal standa í Q_1 og efri hluti kassans í Q_3. Draga skal línu í gegnum kassann í Q_2.
- Neðra skeggið skal ná í minnsta mæligildið (min) og efra skeggið skal ná í það hæsta (max).

1.5 · IQR reglan fyrir útlaga

- Útlagar (outliers) eru mæligildi sem eru mjög ólík öðrum mæligildum og því er mikilvægt að finna þá.
- Ein leið til að átta sig á hvort um útlaga sé að ræða er að bera saman fjarlægð frá gildinu sem sker sig úr og í næsta fjórðungamark (Q_1 eða Q_3).

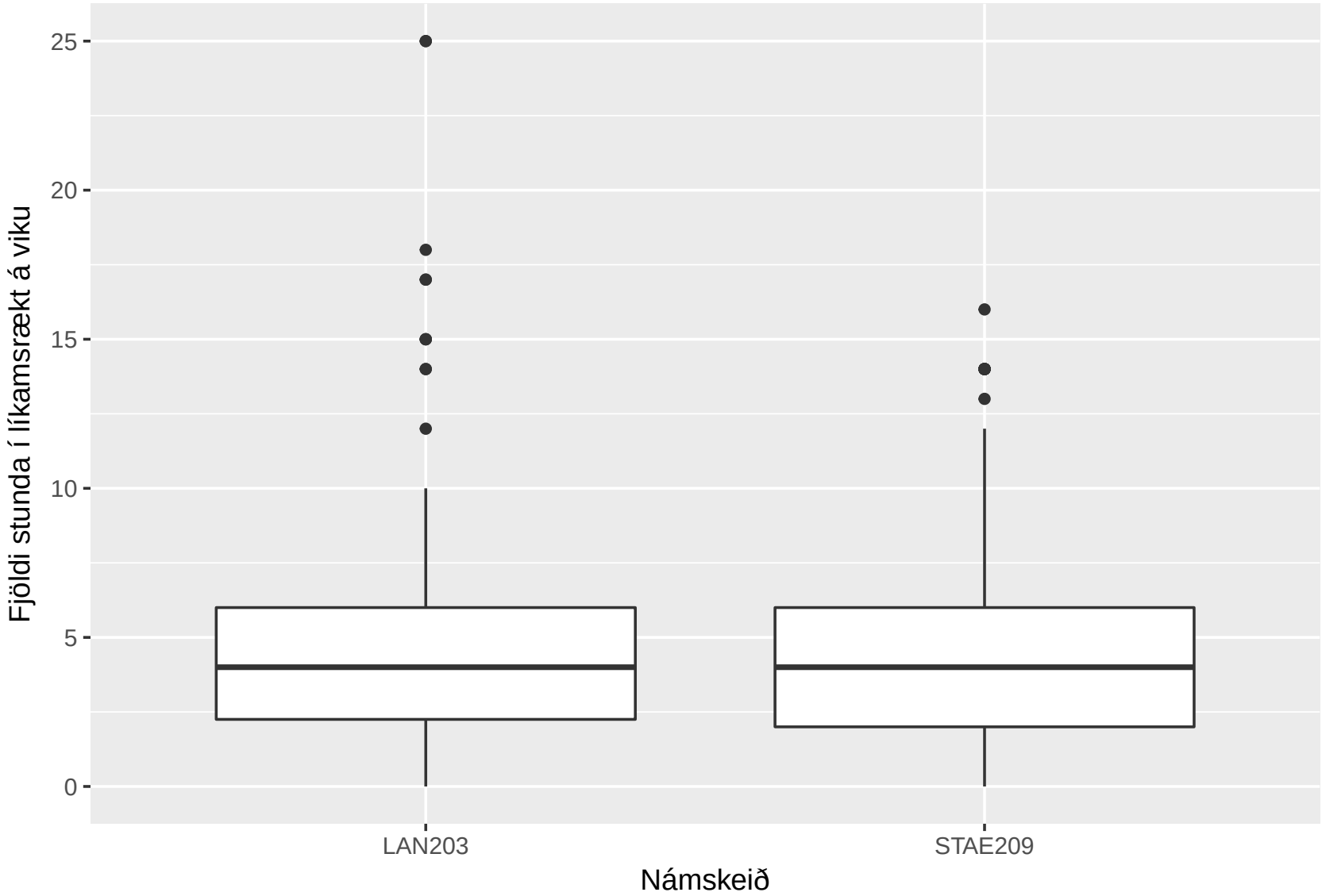
1,5 · IQR reglan fyrir útlaga

- Byrjum á að reikna út fjarlægð mælingunnar sem sker sig úr frá næsta fjórðungamarki (Q_1 eða Q_3).
- Þessi fjarlægð er síðan borin saman við fjórðungaspönnina. Ef fjarlægð mæligildisins frá næsta fjórðungamarki er meiri en $1.5 \cdot IQR$ er litið á mælinguna sem útlaga.

1.5 · IQR reglan fyrir útlaga

- Mörg tölfræðiforrit nota 1.5 · IQR regluna þegar teiknuð eru kassarit og eru þau kassarit oft kölluð **breytt kassarit** (modified boxplot).
- Línurnar sem ganga út frá kassanum, skeggið, eru þá látnar ná allt að einni og hálfri kassalengd frá brúnum kassans en ekki að hæsta og lægsta gildinu eins og gert er í einföldustu útgáfunni.
- Mæligildi sem eru utan við skeggið eru útlagar og merktir inn á ritið með hring.

Breytt kassarit



Yfirlit

Punktarit

Punktarit

Við notum **punktarit** (scatter plot) til að skoða samband milli tveggja talnabreyta.

Gildi annarrar breyturnar eru á y-ásnum (lóðréttur) og hinnar á x-ásnum (láréttur).

Þegar önnur breytan er skýribreyta og hin er svarbreyta er svarbreytan alltaf á y-ásnum og skýribreytan á x-ásnum.

Punktarit