Lýsandi tölfræði

(STATS205.1: Grunnhugtök, tilraunahögun og lýsandi tölfræði)

Anna Helga Jonsdottir og Sigrun Helga Lund

January 7, 2013

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:
  - Lýsistærðir sem lýsa **miðju** (center) gagna
  - Lýsistærðir sem lýsa **dreifð** (spread) gagna
- Algengastar eru meðaltal og staðalfrávik

Lýsistærðir fyrir miðju

Skoðum fimm mismunandi lýsistærðir sem allar lýsa miðju mælinga

1. Miðja spannar (mid range)
2. Tíðasta gildi (mode)
3. Miðgildi (median)
4. Meðaltal (mean, arithmetic mean)
5. Vegið meðaltal (weighted mean)

Miðja spannar

Midja spannar (mid range)

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

$$\text{Miðja spannar} = \frac{x_{\text{min}} + x_{\text{max}}}{2}.$$

Tíðasta gildi

Tidasta 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 í með fjölda mælinga.

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

Vegið meðaltal

Þegar meðaltal er reiknað eins og við gerðum hér að framan fá allar mælingarnar sama vægi. Í sumum tilfellum viljum við gefa mælingunum misjafnt vægi, þá er talað um **vegið meðaltal**.

Vegið meðaltal (weighted mean)

Gerum ráð fyrir að við höfum n mælingar $x_1, x_2, \ldots x_n$ og vægi þeira $w_1, w_2, \ldots w_n$. Vegið meðaltal er reiknað sem

$$\bar{x}_w = \frac{w_1 \cdot x_1 + w_2 \cdot x_2 + \ldots w_n \cdot x_n}{w_1 + w_2 + \ldots + w_n} = \frac{\sum_{i=1}^{n} w_i \cdot x_i}{\sum_{i=1}^{n} w_i}$$

Samanburður á lýsistærðum fyrir miðju

- Allar þær lýsistærðir sem við höfum skoðað eru áhugaverðar og gott að reikna út þegar við fáum ný gögn í hendurnar
- 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.

Lýsistærðir fyrir dreifð

Dreifð mælinga (spread)

Dreifð mælinga er aðferð sem lýsir því hversu dreift mælingarnar liggja.

Við munum fjalla um sex lýsistærðir sem allar lýsa dreifð mælinga

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

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.

$= 0.25 \cdot (n+1)$

Q_2 - sæti í röð: $= 0.50 \cdot (n+1)$

Q_3 - 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.

Prosentumork (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.

Frávikshlutfall

- Varast skal að bera saman staðalfrávik gagna þegar mælingarnar eru í misjöfnum mælieiningum eða meðaltal þeirra mjög frábrugðið.
- Í þeim tilvikum reiknum við frávikshlutfall til að bera saman dreifð tveggja eða fleiri hópa. Það er táknað með CV.

Frávikshlutfall (coefficient of variation)

Frávikshlutfall reiknum við með

$$CV = \frac{s}{\bar{x}}$$

Eftir því sem CV er hærra því dreifðari eru gögnin.

Samanburður á lýsistærðum fyrir dreifð

- Allar þær lýsistærðir sem við höfum skoðað sem lýsa dreifð 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 og á aðeins að nota þegar meðaltal er notað sem mælikvarði á miðju
- 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 dreifð gagnanna

Fimm tölu samantekt

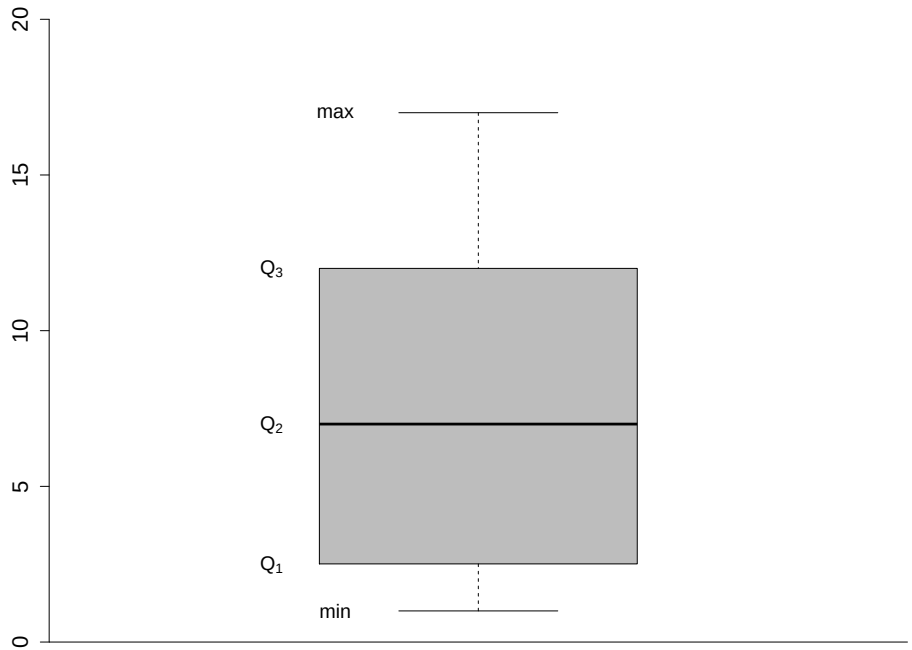
Fimm tölu samantekt (five-number summary)

Fimm-tölu samantekt samanstendur af minnsta gildi (min), fjórðungamörkunum og stærsta gildi (max), þ.e.a.s

$$\text{min},\ Q_1,\ Q_2,\ Q_3,\ \text{max}$$

Kassarit

- **Kassarit** er notað til að skoða miðju og dreifð mælinga.
- Endurspegla vel gögnin og sýna glöggt hvort dreifingin er samhverf eða skekkt.
- Ef dreifingin er skekkt, tvíkryppu - eða fjölkryppu, er fimm tölu samantekt og kassarit besta leiðin til að lýsa miðsækni og dreifð.
- 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.