Introduction to Statistics and R programming

Descriptive Statistics & Graphics

Eric Stemmler

27.01.2021

Contents

1 Recap

- statistical vocabulary
  - variation: the outcome of a sample varies randomly
  - uncertainty: lack of knowledge of about a true value
  - trial: the realization of an experiment
  - population: all possible events or items
  - population parameter: the true value
- calculating probabilities, probability tree
- experiments with 2 outcomes: binomial distribution

Өмнөх хичээлээр санамсаргүй байдлын хоёр тал болох тодорхойгүй байдал/uncertainty/ болон вариац/variation/-ын талаар авч үзсэн бөгөөд зоос шидэх жишээг ашиглан тодорхойгүй байдлын талаар, савнаас бөмбөлөгийг гаргаж авах жишээг ашиглан вариацын талаар тус тус ярилцсан байгаа. Эдгээр жишээн дээр бид зоосыг тогтсон тоогоор олон удаа давтан хаях, жишээнь 100 удаа хаяхыг нэг trial гэж нэрлэж болох бөгөөд энэ нь туршилтын нэг удаагийн хэрэгжилт гэсэн утгатай үг юм гэдэг талаар бас ярьсан байгаа. Туршилтын trial-уудыг хийснээр бид нэг удаагийн туршилтын trial болон үйл явдлын боломжит бүх үр дүнг багтаасан статистик эх олонлогийг илэрхийлсэн өгөгдлийг цуглуулдаг. Статистик дүн шинжилгээ хийхдээ цуглуулсан өгөгдлөөсөө эх олонлогийн параметрийн талаар олж мэдэх нь бидний зорилго юм. Дан ганц үйл явдлуудын магадлалыг үр дүнгийн хослолтой/үржүүлэх/ хэрхэн нэгтгэх талаар болон жишээ нь 5 зоосыг дараалан хаяхад дор хаяж 3 удаа сүлдээр буух гэсэн

сонирхолтой хэд хэдэн үйл явдлын магадлалыг хэрхэн яаж нэгтгэх талаар сурч мэдсэн. Биномийн тархалтыг зөвхөн 2 үр дүнтэй туршилтын статистик загварчлалд ашиглаж болохыг бид харсан.

2 Handedness Inventory

- Please fill in the questionnaire during **this** lesson
- Return questionnaire at the end
- We will use this data for the next lessons

Ирэх долоо хоногийн хичээл дээр ашиглах энэхүү асуулгын хуудсыг бөглөж өгнө үү. Би энэ асуулгын хуудсыг аль хэдийн та бүхэнд тараасан байгаа бөгөөд хэрэв та хараахан аваагүй байгаа бол наашаа ирээд авна уу. Өнөөдрийн хичээлийн төгсгөлд бөглөөд буцааж өгнө үү. Асуулгын хуудсын талаар танд ямар нэгэн асууж тодруулах зүйл байна уу? Бид энэ өгөгдлийг хүмүүс аль гараа түлхүү ашигладаг эсэх ерөнхий хандлагын талаар дүн шинжилгээ хийхэд болон мөн R-д өгөгдлийн дүн шинжилгээ хийхэд ашиглах болно.

3 Age Guessing

Please indicate which hand you use for each of the following activities by putting a + in the appropriate column, or ++ if you use would never use the other hand for that activity. If in any case you are really indifferent, put + in both columns. Some of the activities require both hands. In these cases the part of the task, or object, for which hand preference is wanted is indicated in parentheses.

Даалгавар (Task)	Зүүн (Left)	Зөв (Right)
Бичих Writing		
Зурах Drawing		
Шидэх Throwing		
Хайч Scissors		
Шүдний сойз Toothbrush		
Сэрээгүй хутга Knife, without fork		
Халбага Spoon		
Дэрс, дээд гар Broom, upper hand		
Хачирхалтай тэмцээн, тэмцээнийг барьж буй гар Striking match (hand that hold the match)		
Нээлтийн хайрцаг (тагийг нь барьдаг гар) Opening box (hand that holds the lid)		
нийт дүн Total		

Right – Left = Right + Left = $\frac{\text{Right - Left}}{\text{Right + Left}}$ =

Create a Left and a Right score by counting the total number of + signs in each column. Your handedness score is (Right − Left)/(Right + Left): thus, a pure right-hander will have a score of score (20 − 0)/(20 + 0) = 1, and a pure left-hander will score (0 − 20)/(0 + 20) = −1.

3.1 Data Collection

Guessing ages. For each card your group is given, estimate the age of the person on the card and write your guess in the table below in the row corresponding to that numbered card. Later, you will be told the true ages and you can compute your errors. The error is defined as estimated minus actual age.	Team-ID:
	No. team members:

Card	Estimated Age	Actual Age	Error
1			
2			
3			
4			
5			
6			
7			
8			
9			
10			

Тойм статистик ба графикийн талаархи өнөөдрийн хичээлийг би та бүхэн хүмүүсийн насыг хэр сайн тааж чаддаг болохыг харах бяцхан туршилтаар эхлүүлэхийг хүсч байна. Эхлээд та бүхэн 10 багт хуваагдана уу.Дараа нь баг тус бүр багийн гишүүдийн тоогоо бичнэ үү. Одоо би баг тус бүрт нэг нэг хүний зургийг тарааж өгөх болно. Баг бүр зураг дээрх хүний насыг ярилцаж, таана уу. Баг бүр нэг зураг дээр нэг л таамаглал дэвшүүлэх ёстой. Гэрэл зураг бүр дугаартай байгаа. Зургийн дугаартай таарч буй мөрөнд тухайн зургийн талаархи таамагласан насаа бичнэ үү. " Estimated Age " гэсэн баганад таамагласан насаа бичнэ. Дараа нь “А” баг гэрэл зургаа “Б” багт, “Б” баг “С” багт дамжуулах гэх мэтээр зургаа багууд хоорондоо дамжуулна. “J” баг гэрэл зургаа “А” багт дамжуулна. Дараа нь дараагийн зургийнхаа насыг тааж, зургийн дугаарыг шалгаад, таамагласан насаа хүснэгтэнд тохирох газар нь оруулж бичнэ үү. Баг тус бүр 10 зургийг бүгдийг нь харж дуусах хүртлээ үүнийг давтан хийнэ. 20 минутын дараа дуусгана уу.

Би үр дүнгүүдийг хураангуйлах хүснэгтийг самбар дээр зурсан байгаа. Би одоо гэрэл зургууд дээрх хүмүүсийн жинхэнэ насыг хэлж өгье. Жинхэнэ насыг өгөгдлийн хуудсандаа хуулж бичээд, тус бүрийн алдааг тооцоолно уу. Алдааг тооцоолж дуусаад алдааныхаа дунджийг тооцоолно уу. Үүний дараа гарч ирээд самбар дээр байгаа хураангуйлах хүснэгтэнд үр дүнгээ хуулж бичнэ үү. Одоо харж байгаа үр дүнгүүдийнхээ талаар ярилцана уу. Бидэнд ялагч байна уу? Хүснэгтийн үр дүнгүүд нь таны хүлээж байсан үр дүнтэй нийцэж байна уу? Мэдээллийн талаар таны ажигласан өөр зүйл байна уу? Өгөгдөлийг цуглуулсан арга нь үр дүнд нөлөөлж болох уу? Өгөгдлийг санамсаргүй түүврийн хэлбэрээр цуглуулсан уу?

4 Stem-Leaf Plot

Consider the following data points:

20, 14, 16, 18, 22, 38, 61, 52, 52, 55, 76, 84, 79, 81, 82

```
##
##   The decimal point is 1 digit(s) to the right of the |
##
##   1 | 468
##   2 | 02
##   3 | 8
##   4 |
##   5 | 225
##   6 | 1
##   7 | 69
##   8 | 124
```

- Create a steam-leaf plot of the averaged errors

Guessing ages. For each card your group is given, estimate the age of the person on the card and write your guess in the table below in the row corresponding to that numbered card. Later, you will be told the true ages and you can compute your errors. The error is defined as estimated minus actual age.	Team-ID:
	No. team members:

Card	Estimated Age	Actual Age	Error
1			
2			
3			
4			
5			
6			
7			
8			
9			
10			

Одоо би та бүхэнд өгөгдлийн багцыг хэрхэн графикаар нэгтгэн дүгнэх зарим боломжуудын талаар танилцуулья. Дүн шинжилгээ хийхийг хүсч буй аливаа өгөгдлийн багцыгаа визуалчлах явдал нь үр-

гэлж маш чухал байдаг. Өгөгдлийг визуалчлах нь таны өгөгдлийн багцыг ойлгох, өгөгдлийн цар хүрээ, өгөгдлийн хэлбэрийг хурдан харах, өгөгдөл цуглуулахад ямар нэг алдаа гарсан эсэхийг мэдэхэд тусалдаг. Тоон өгөгдлийг визуалчлах маш энгийн арга бол "иш-навчны график" юм. Иш-навчны график нь 1-рт тоонуудыг бүлэгт ангилах ба 2-рт бүлэг тус бүрт хамаарах өгөгдлийн цэгүүдийн тоог тооцож гаргадаг график юм. Бүлэг бүрийг тухайн тооны эхний цифр/тооны орон/-ээр тодорхойлно. Өгөгдлийн цэг эсвэл тоо нь ижилхэн бүлэглэх дугаараар эхэлсэн бол тодорхой нэг бүлэгт хамаарна.Ерөнхийдөө бүлэглэх дугаарыг эхний нэг цифр/орон эсвэл эхний хоёр, гурван цифр/орон байхаар сонгож болдог. Энэ тохиолдолд эхний цифр/орон-ийг сонгосон бөгөөд эхний мөр нь бид 14, 16, 18 гэсэн өгөгдлийн цэгүүдтэй байгааг хэлж өгч байна.

Асуух зүйл байна уу? Одоо хураангуйлсан хүснэгтэн дэх алдааны дундаж утгуудын иш-навчны графикийг үүсгэнэ үү. Энэхүү өгөгдлийн хураангуйг бид хэрхэн тайлбарлаж болох вэ?

5 Histogram

5.1 Introduction

- stem-leaf plot only works for numbers, not for categories
- stem-leaf plots are always based on digits
- count data in terms of intervals/ bins (e.g. width=5): (0,5], (5,10], (10,15] ... (80,85], (85,90], (90,95], (95,100]

Example: 20, 14, 16, 18, 22, 38, 61, 52, 52, 55, 76, 84, 79, 81, 82

20	(15,20]	38	(35,40]	76	(75,80]
14	(10,15]	61	(60,65]	84	(80,85]
16	(15,20]	52	(50,55]	79	(75,80]
18	(15,20]	52	(50,55]	81	(80,85]
22	(20,25]	55	(50,55]	82	(80,85]

interval	(10,15]	(15,20]	(20,25]	(35,40]	(50,55]	(60,65]	(75,80]	(80,85]
count	1	3	1	1	3	1	2	3

Мэдээллийн багцыг визуалчлах бас нэг илүү уян хатан боломж бол гистограм зурах явдал юм. Гистограм нь статистик шинжилгээний супер чухал хэрэгсэл бөгөөд үүнийг сайн ойлгож авах шаардлагатай. Гэхдээ энэ нь бас өгөгдлийг бүлэглэж, дараа нь бүлэг тус бүрийн өгөгдлийн цэгүүдийн тоог тооцоолж гаргадаг учраас иш-навчны графиктай төстэй юм. Гистограм дахь бүлгүүдийг “bins” гэж нэрлэдэг. “bin” нь математикийн хувьд хагас хаалттай хагас нээлттэй интервалыг хэлнэ. “bins” буюу интервалууд нь тоон өгөгдлийн аль цэгийг бүлэглэхийг тодорхойлдог. “bins”/интервалууд-ын хэмжээг “bin”/интервал-ын өргөн гэж нэрлэдэг. Гистограм зурахаас өмнө “bin”/интервал-ынхаа өргөнийг шийдэх хэрэгтэй. Өмнөхтэй ижил тооны дарааллыг дахин авч үзье.

Энд харагдаж байгаа хүснэгтэнд аль тоо аль интервалд хамаарахыг харуулав. Гистограм үүсгэхийн тулд интервал бүрт хамаарах өгөгдлийн цэгүүдийн тоог тооцоолно. Оролцогчдоос асуух асуулт: Бидэнд 50-55 хүртэлх интервалд хэдэн өгөгдлийн цэг байна вэ?

Хоёр дахь хүснэгтээс бид үр дүнг харж болно. Жишээлбэл 15-аас 20 хүртэлх интервалд 14, 16, 18 гэсэн 3 өгөгдлийн цэг байна.

Одоо бид энэ сүүлчийн хүснэгтээс гистограм үүсгэж болно. Гистограм нь тэгш өнцөгтийн график юм. Тэгш өнцөгтийн өргөн нь bins/интервалуудын хэмжээ буюу өргөнтэй ижил утгатай байна. X тэнхлэг дээрх тэгш өнцөгтийн байрлал нь тодорхой интервалтай тохирч байна. Тэгш өнцөгтийн өндөр нь тухайн интервалд тооцсон өгөгдлийн цэгүүдийн тоотой тохирч байна. Хэрэв гистограмыг тооцсон бодит тоогоор зурсан бол үүнийг абсолют давтамжийн график гэж нэрлэдэг. Ихэвчлэн абсолют тоо нь

тийм ч сонирхолтой байдаггүй. Бид ихэвчлэн тодорхой bin/интервал-д байх өгөгдлийн эзлэх хувийг илүү сонирхдог.

Энэ зорилгоор бодит тооны оронд өгөгдлийн цэгүүдийн хувийг тодорхой интервалд байх нийт өгөгдлийн цэгүүдийн тооноос хялбархан тооцоолох боломжтой. Тэгш өнцөгтийн өндөр нь өгөгдлийн цэгүүдийн бутархай тоотой/ өчүүхэн хэсэгтэй тохирч байгаа гистограмыг харьцангуй давтамжийн график гэж бас нэрлэдэг. Тэгш өнцөгт бүрийн өндрийг тогтмол тоогоор тодорхойлсон учраас тэгш өнцөгтийн харьцангуй өндөр өөрчлөгдөөгүй тул гистограммын бодит хэлбэр өөрчлөгдөөгүй болохыг анхаарна уу.

5.2 Exercise

Create your own histogram

1. Plot the error values as a histogram
2. Choose a bin width/ size
3. Draw the histogram on the blank plotting sheet
4. Present your histogram

Try to answer and discuss the following questions:

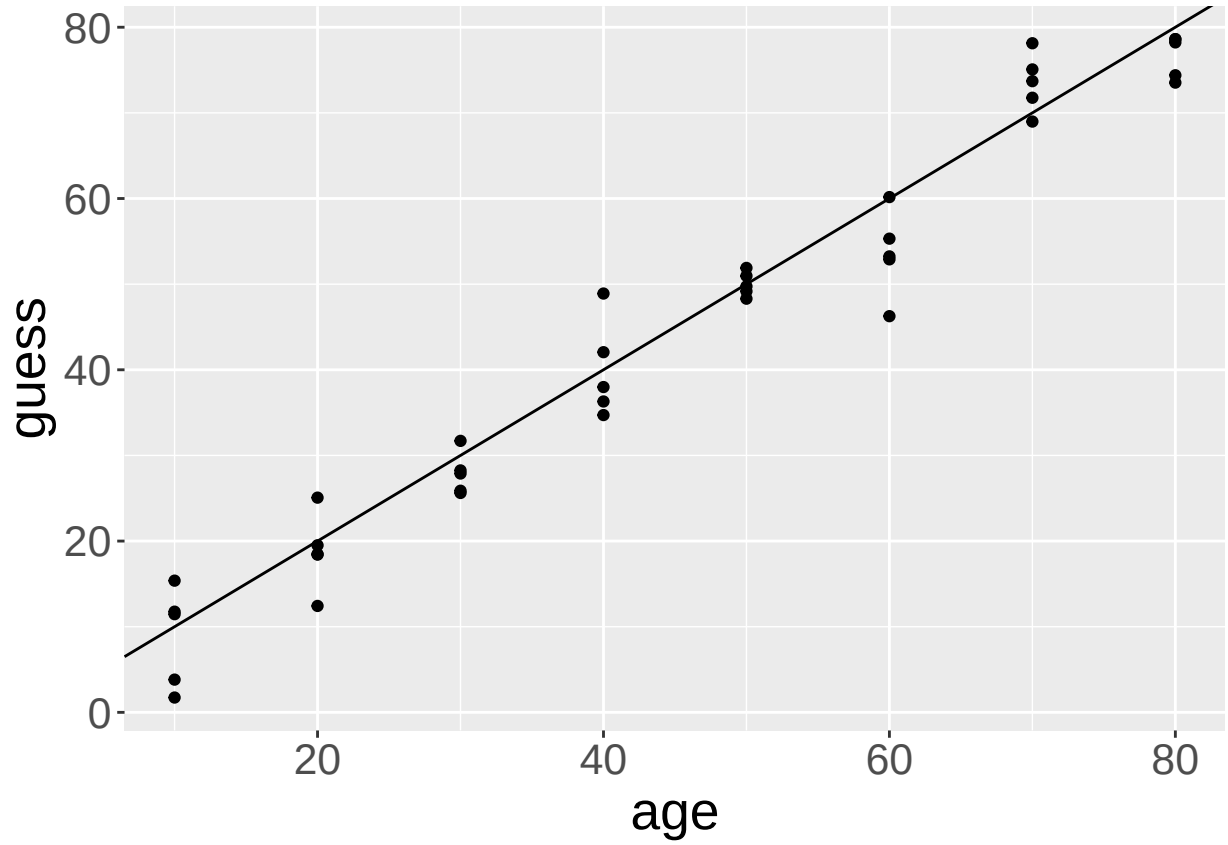
1. How does the histogram change with different bin widths?
2. How accurate is age guessing?
3. Does guessing accuarry vary across age?
4. What is the variance/ spread of the error?

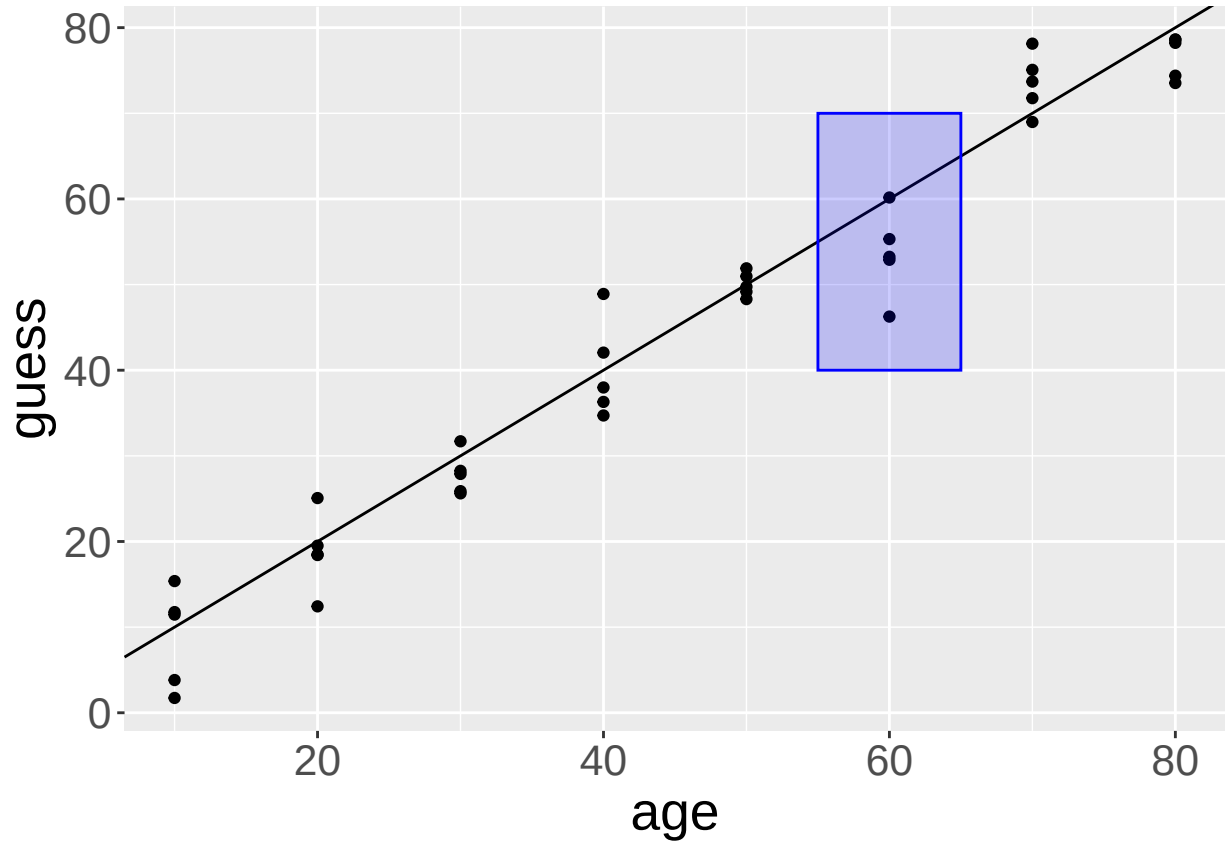
Одоо таны ээлж боллоо. График зурах хоосон цаас аваад гистограмаа зураарай. Самбар дээр байгаа бодит алдааны утгуудын өгөгдлийг визуалчлана уу. Дундаж утгыг биш харин бодит алдааны утгуудыг ашиглаарай!

Гистограмаа бусад оролцогчдод танилцуулахад бэлдэнэ үү. Нас таалт нь хэр оновчтой, хэр их ялгаатай байна гэсэн асуултанд хариулахыг хичээгээрэй. bin/интервал-ынхаа хэмжээг яагаад сонгохоор болсноо тайлбарлана уу.

6 Multiple Variables

6.1 Scatter plot





Иш-навчны графикууд ба гистограмууд нь нэг хувьсагчийн өгөгдлийг харуулах боломжууд юм. Хэрэв бид 2 хувьсагчийг визуалчлахыг хүсвэл нийтлэг графикийн төрөл нь scatter plot / тархалтын график юм. Зоос шидэх туршилтын талаар ярилцсан 1-р хичээл дээр бид аль хэдийн тархалтын графикийг үзсэн. Тухайн хүний гэрэл зургийн нэг тооцоолсон насны утгад хамаарах цэг бүрийг харуулсан энэ тархалтын график дээр та цэгүүдийг харж байна. Энэхүү график дээрх одоогийн өгөгдлийг зохиомлоор үүсгэсэн болно. Цэг тус бүрийн хувьд x тэнхлэг дээрх байрлал нь зураг авахуулсан хүний жинхэнэ настай тохирч байна. y тэнхлэг дээрх байршил нь тооцоолсон настай тохирч байна. Таны харж байгаачлан жинхэнэ насны утга бүрт үргэлж олон тооны тооцоолсон утга байна. Хэрэв тооцоолсон утга нь төгс бөгөөд жинхэнэ утгуудтай тэнцүү байсан бол бүх цэгүүд тархалтын графикт үзүүлсэн шиг налуу 1-тэй шулуун шугам дээр байрлах байсан.

Шугамын доогуур эсвэл дээгүүр байгаа цэгүүдийн тайлбар юу байж болох вэ? Хэрэв цэгүүд шугамын доогуур эсвэл дээгүүр нь системтэйгээр байрлаж байгаа бол статистикийн хэллэгээр үүнийг bias/хэвийх байдал гэнэ. Энэ тохиолдолд 60 нас хүрсэн хүмүүс наснаасаа залуу харагдах хандлагатай гэсэн үг юм.

7 Summary

- ***Always*** visualize your data set before analysing it!
- data collection & experiment: randomization, random sample
- descriptive statistics: mean, error, standard deviation, variance, bias
- statistical graphics: stem-leaf plot, histogram, scatter plot
- bins:
  - too narrow: loss of shape
  - too wide: missing the details