Introduction to Statistics and R

Descriptive Statistics - Part II

Eric Stemmler

03.02.2021

Contents

1 Recap **1**

2 Learning Goals **1**

3 Handedness Data Analysis **2**

4 Reported Weights **7**

5 Probability Mass/ Density Functions **11**

6 Age guessing error **13**

7 Summary **22**

1 Recap

- data collection process (handedness questionnaire, age guessing demonstration)
- statistical graphics
  - stem-leaf plot: counting data
  - histogram (bin size: shape vs. detail): absolute and relative frequencies
  - scatter plot: relationships
- Numerical statistics: mean, error

Өнгөрсөн долоо хоногт бид хүмүүсийн зургийг хараад насыг нь таах замаар өөрсдөө анкет бөглөж мэдээлэл цуглуулсан. Насыг таах дасгалын үеэр бид мэдээлэл цуглуулах үйл явцын талаар мөн энэ нь өгөгдлийн дүн шинжилгээний үр дүнд хэрхэн нөлөөлж болох талаар ярилцсан. Өгөгдлийн багцыг дүрслэн харуулах 3 боломжийн талаар, тухайлбал өгөгдлийг ангилж, тоолох боломжтой иш-навчны зураглал болон гистограммыг, хоёр өөр хувьсагчийн хоорондын холбоо хамаарлыг визуалчлах /дүрслэх боломжтой тархалтын графикийн талаар сурч мэдсэн. Насыг таамаглах туршилтын үеэр та таасан алдааныхаа дундаж утгыг тооцоолсон бөгөөд үүний тулд та алдааны утгуудын абсолют утгыг тодорхойлох хэрэгтэй бөгөөд тэгэхгүй бол дундаж алдаа нь бодит бус бага байх байсан гэдгийг та бүхэн мэдэж авсан байгаа.

2 Learning Goals

- Numerical summary statistics: variance, precision, standard deviation, quantiles
- Relation between numerical and graphical summaries
- Probability density functions

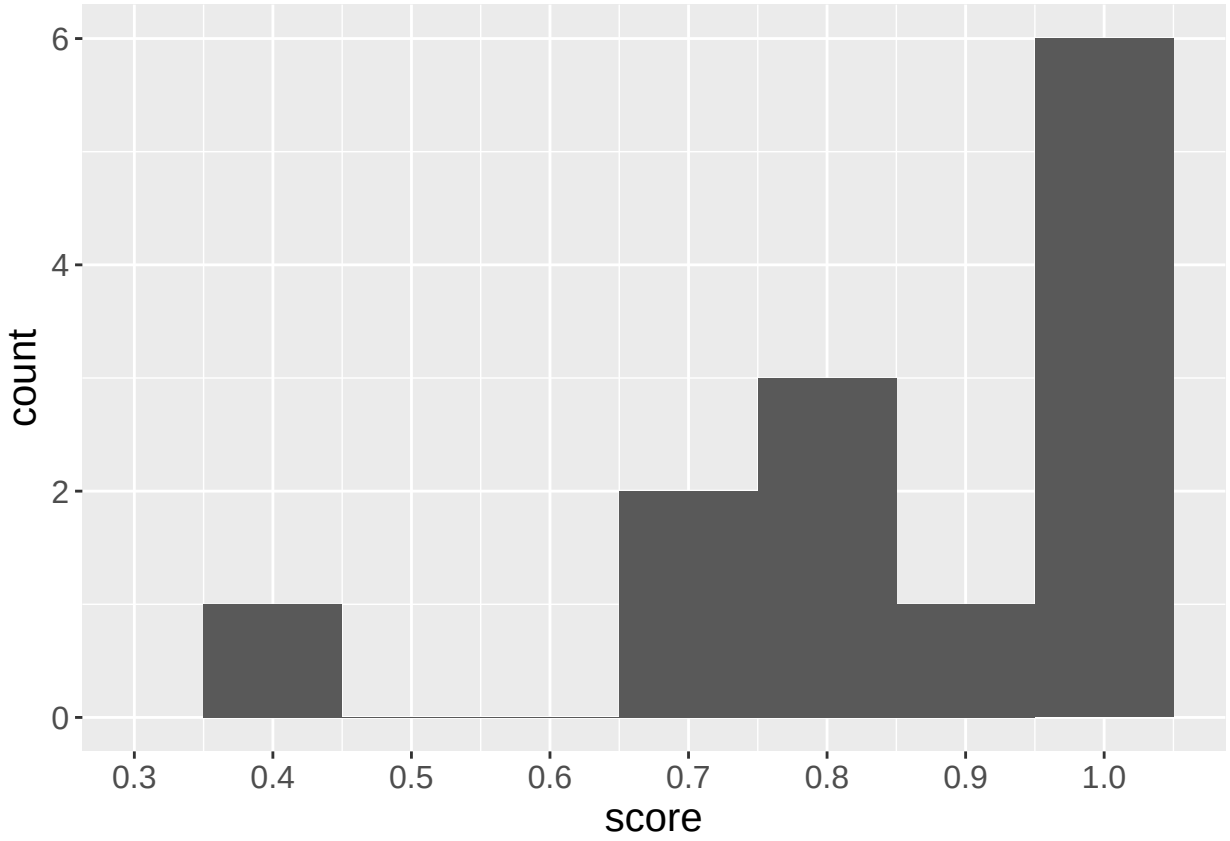
Өнөөдөр бид өгөгдлийн цар хүрээ, тархалтыг тодорхойлоход ашигладаг шинэ тоон статистикийг судлах болно. Эдгээр статистикууд нь бидний аль хэдийн мэддэг болсон статистик графиктай холбож авч үзсэнээр ямар болохыг харцгаая. Бид сурч мэдсэн бүх зүйлээ магадлалын тархалтын маш чухал ойлголттой хэрхэн уялдаж байгааг олж мэдэх үед бидэнд тохиолдож буй бэрхшээлийн хувьд урагш том үсрэлтийг хийх болно.Үнэн хэрэгтээ магадлалын тархалтууд нь статистик таамаглалын тест гүйцэтгэх замаар далд эсэхийг, эсвэл суурь тархалтын статистикийн загварчлал хийх замаар ил тодорхой эсэхийг тодорхойлдог аливаа статистик дүн шинжилгээний бодит үндсэн зорилго юм.

3 Handedness Data Analysis

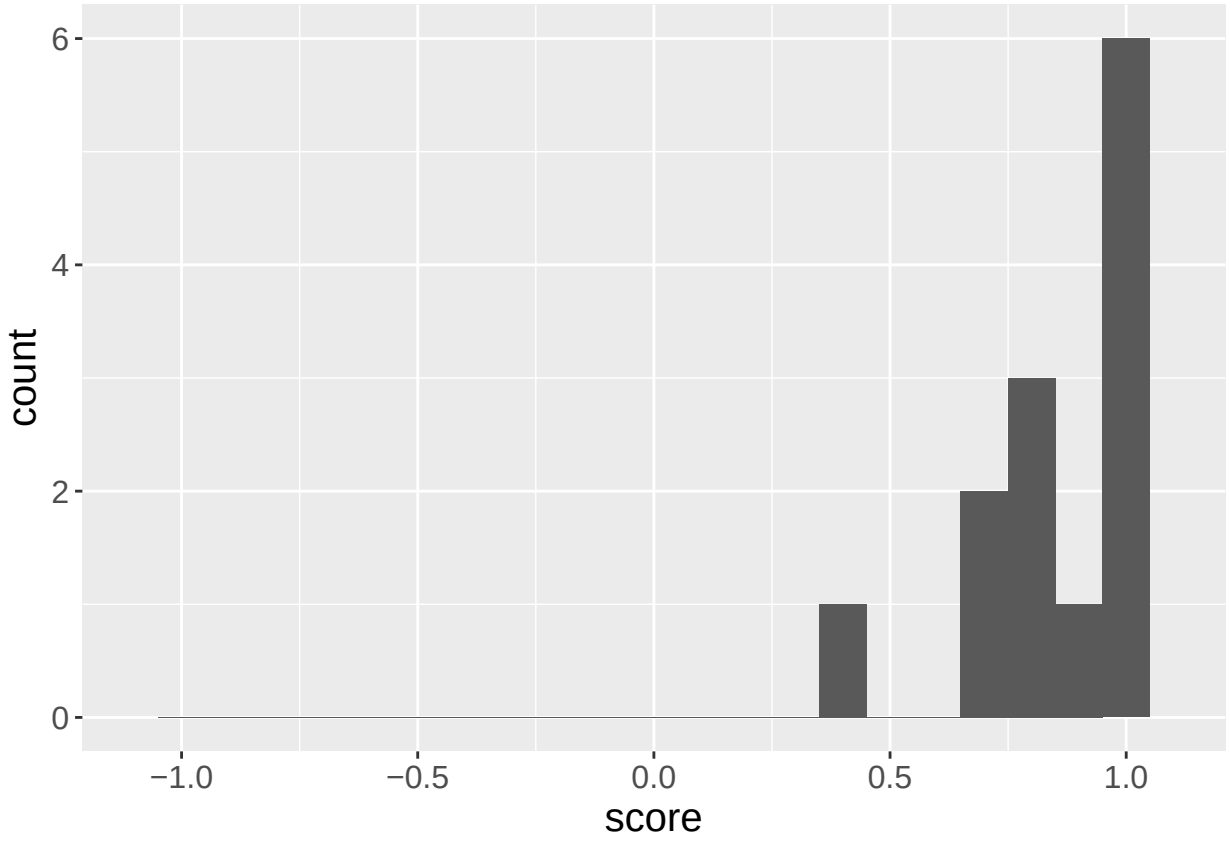
Table 1: Complete collected data set from the handedness inventory

id	writing	drawing	throwing	scissors	toothbrush	knife	spoon	broom	match	box	calc	score	right	left
1	rr	rr	rr	rr	rr	rr	rr	rr	rr	rr	1	1.0000000	20	0
2	r	r	r	r	r	r	r	r	r	r	NA	1.0000000	10	0
3	rr	rr	r	rr	rr	rr	rr	rr	rr	rr	NA	1.0000000	19	0
4	rr	rr	rr	rr	rr	rr	rr	rr	rr	rr	NA	1.0000000	20	0
5	rr	rr	rr	rr	rr	rr	rr	rr	rr	rr	-1	1.0000000	20	0
6	rr	rr	rr	rr	lr	lr	rr	rr	lr	r	0.85 ~ 1	0.6842105	16	3
7	rr	rr	r	rr	rr	r	rr	r	ll	rr	0.8	0.7647059	15	2
8	r	r	r	r	r	r	r	r	lr	r	0.81	0.8181818	10	1
9	rr	rr	rr	rr	rr	rr	rr	NA	lr	rr	NA	0.8888889	17	1
10	r	r	r	r	r	r	r	r	lr	r	0.81	0.8181818	10	1
11	r	r	r	r	r	r	r	r	r	r	1	1.0000000	10	0
12	r	r	lr	r	lr	r	r	lr	lr	r	NA	0.4285714	10	4
13	lr	rr	lr	rr	rr	rr	r	r	r	r	0.75	0.7500000	14	2

Та бүхнээс цуглуулсан аль гараа ашиглах хандлагатай эсэх талаарх асуулгын хуудаснуудаас дүн шинжилгээ хийхэд ашиглаж болох хүснэгтийг файл дотор үүсгэлээ. Тус хүснэгт нь энд харагдаж байна. Эхний баганатай зэрэгцээд тус асуулгаар асуусан үйлдлүүдийн нэрээр нэрлэсэн багануud байна. Эдгээр баганууdын утгыг "r" ба "l" үсгээр кодчилсон ба баруун ба зүүн гарын багана тус бүрт бөглөж бичсэн нэмэх тэмдэгийг орлож байгаа болно. “rr” эсвэл “ll” нь давхар нэмэх тэмдгийг илэрхийлж байгаа болно. Тэдгээр багануud байгаа хэвээрээ байна. Жишээлбэл, ямар ч хариулт байхгүй байгааг “Not available” гэсэн үгийн товчлол болох “NA”-аар кодлосон байгаа . Статистикийн программд ийм кодчилол нь утга байхгүй гэсэн үг юм. Бид өгөгдөл цуглуулахдаа үүнийг өөрчлөх эсвэл тайлбарлахыг хүсдэггүй. Иймд NA нь ашигтай байдаг. Жишээ нь, 9-р хүн шүүрний асуултанд хариулаагүй боловч тухайн хүн "lr" нь хоёр гараа аль алиныг нь хэрэглэдгийг илэрхийлж байгаа гэдгийг ойлгож байгаа. Мөн “calc” баганад оролцогчдын бичсэн үр дүнг байгаагаар нь харуулсан байгаа. Харин "score" гэсэн баганад R програмаар тооцсон үр дүнг харуулсан байгаа. Магадгүй энэ хүн хэзээ ч дэрсэн шүүр хэрэглэдэггүй эсвэл асуултуудыг нь ойлгоогүй байж магадгүй юм. Энэ нь илүү олон удаа тохиолдвол магадгүй бид асуулгын хуудсаа дахин тайлбарлах эсвэл асуултаа солих талаар бодож болох юм. Мэдээллийг байгаа байдлаар нь цуглуулах нь таны өгөгдлийн багц дахь алдааг шалгаж, өгөгдлийг хэрхэн цуглуулсан талаархи мэдээллийг өгөхөд хэрэгтэй байж болох юм. Энэ нь "хүмүүсийн талаархи хүмүүсээс авсан мэдээлэл" тул энэ нь таны шинжлэх ухааны салбарт хэрэглэгдэхгүй хамаагүй зүйл гэж та бодож магадгүй юм. Гэхдээ олон тохиолдолд өгөгдлийг машин биш хүн цуглуулсаар байгаа гэдгийг ойлгох хэрэгтэй. Өгөгдөлд цуглуулсан хүн нь нөлөөлж болно. Экологийн жишээн дээр авч үзвэл, өөр лабораторид химийн шинжилгээ хийлгэхээр илгээсэн ижил сорьц арай өөр үр дүнг гаргаж болох юм. Энэ нь сорьцыг цуглуулж дүн шинжилгээ хийсэн хүний нэрийг оруулаад цаашлаад илүү нарийвчлалтай дүн шинжилгээ хийх боломжийг танд олгож болох юм. Одоо бүгдээрээ өөрсдийн өгөгдлийн багцадаа дүн шинжилгээ хийцгээе: Үүний гистограммын хэлбэрийг та ямар гэж таамаглаж байна вэ?.



Энд бидний цуглуулсан өгөгдлийн гистограмм харагдаж байна. Нэг алхам ухарцгаая: Энэ гистограммын суурь эх олонлог нь хэд вэ?



Энэ бол одоо боломжтой бүх утгыг багтаасан ижил гистограммын дахин хэмжсэн хувилбар юм. Do you notice anything about it? What is your interpretation?

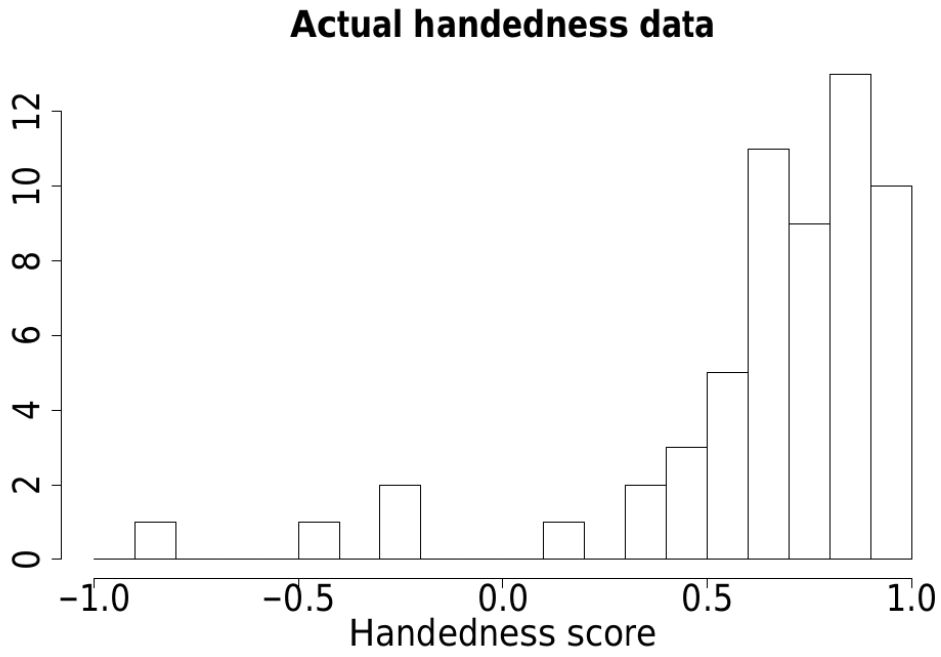
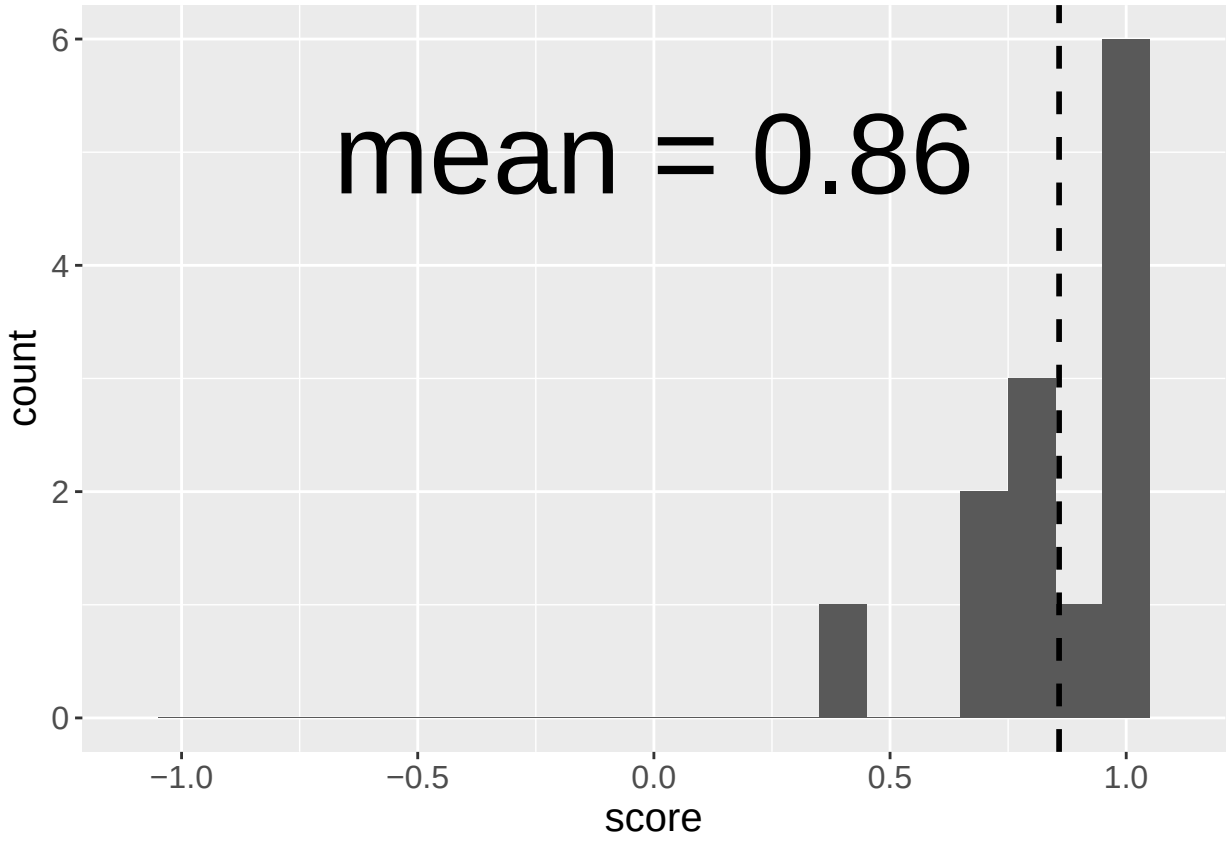
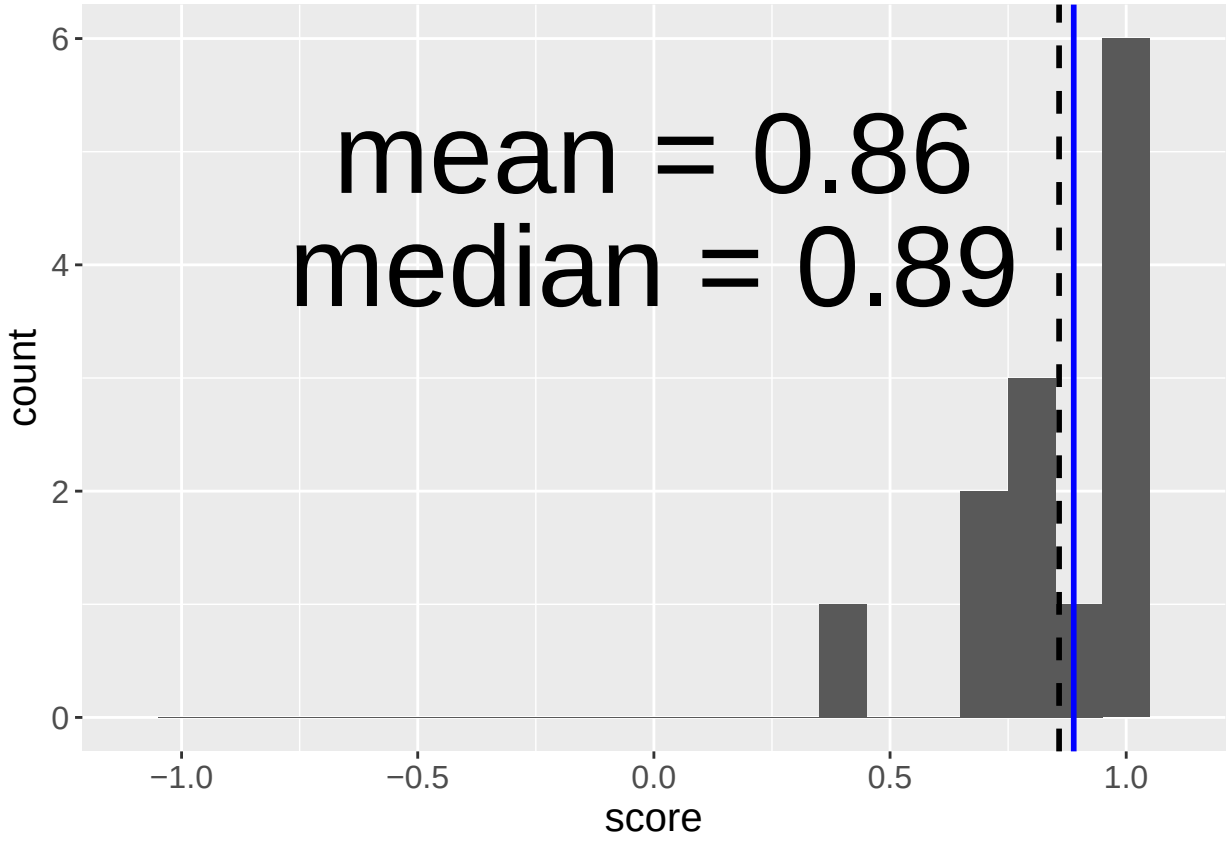


Figure 1: From Gelman and Nolan (2017)

Энд та ижил асуулгын хуудсыг ашиглан авсан өөр түүврийн гистограммыг харж байна. Түүврийг АНУ-ын статистикийн ангийн олон тооны оюутнуудаас авсан болно. Зүүн буюу солгой гартай хүмүүс бидний түүвэрт байхгүй байгаа тул бидний түүвэр үнэхээр жижигхэн гэдгийг бид харж байна. Зүүн буюу солгой гартай хүн нийтлэг бус байдаг тул 0 солгой гартай хүмүүсийг түүж авах боломж хамаагүй өндөр байдаг ба иймээс жишээ нь нэг талыг барьсан өрөөсгөл дундаж утгыг барьж авдаг. Гэсэн хэдий ч манай гистограммын хэлбэр нь АНУ-ын түүврийн хэлбэртэй төстэй болохыг анхаарна уу. АНУ-аас авсан түүвэр дээр 40 баруун гартай хүнд ойролцоогоор 1 солгой гартай хүн ногдож байна. Бидний түүврийн хэмжээ 13 тул солгой гартай хүмүүс огт байхгүй байгаа нь гайхах зүйл биш юм.

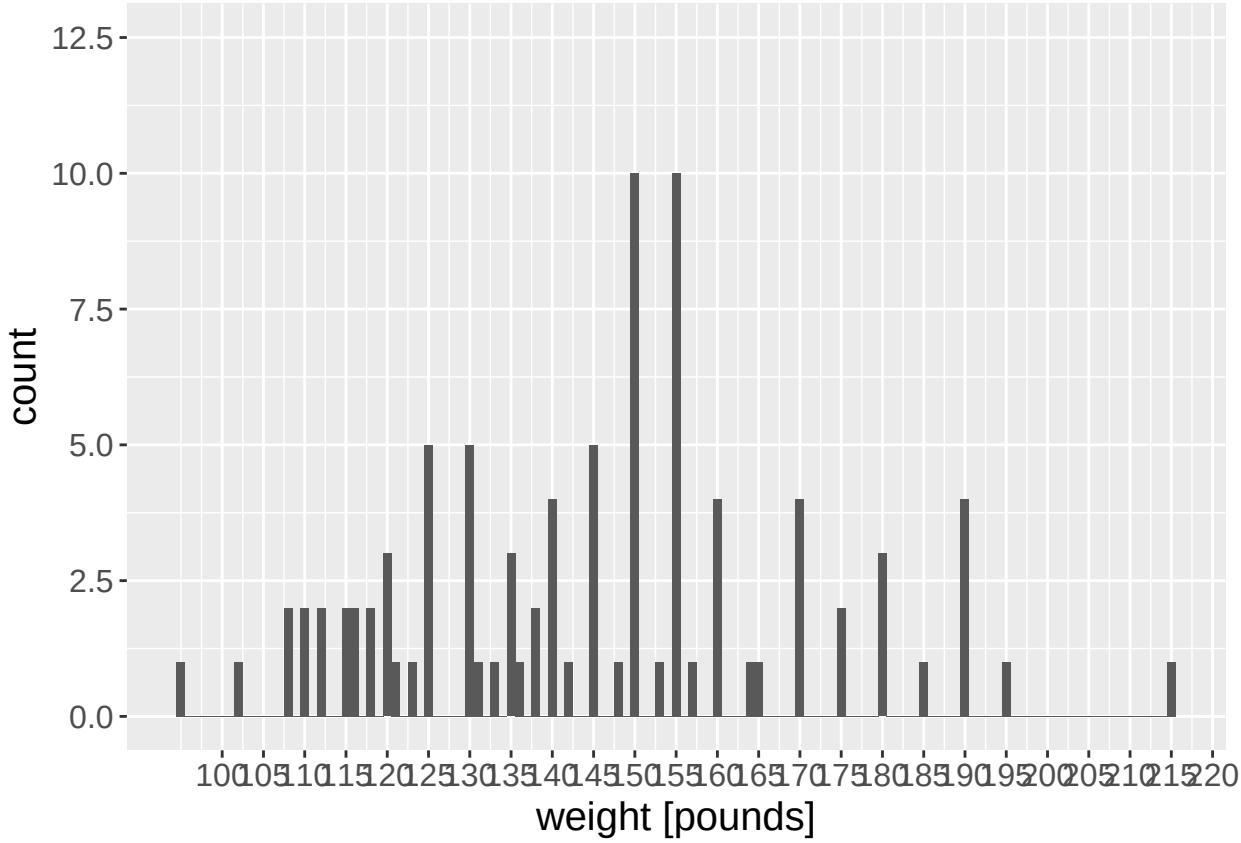


Гистограммын өөр нэг ашигтай тал нь дундаж утгыг ойролцоогоор тооцоолоход хялбар байдаг. Би өгөгдлийн багцын нарийн яг зөв дундаж утгыг тооцоолж, тасархай шугамаар тэмдэглэсэн байгаа. Гистограммын дундажыг ойролцоогоор гаргахын тулд жинлэсэн дундаж утгыг тооцоолж болно. Үүний тулд баар тус бүрийн өндрийг баар тус бүрийн голд байрлах онооны цэгээр үржүүлнэ. Өөрөөр хэлбэл хүний тоог алдааны тоогоор үржүүлнэ. Энэ тохиолдолд $6 \times 1.0 + 1 \times 0.9 + 3 \times 0.8 + 2 \times 0.7 + 1 \times 0.4 = 0.8538462$ болно.

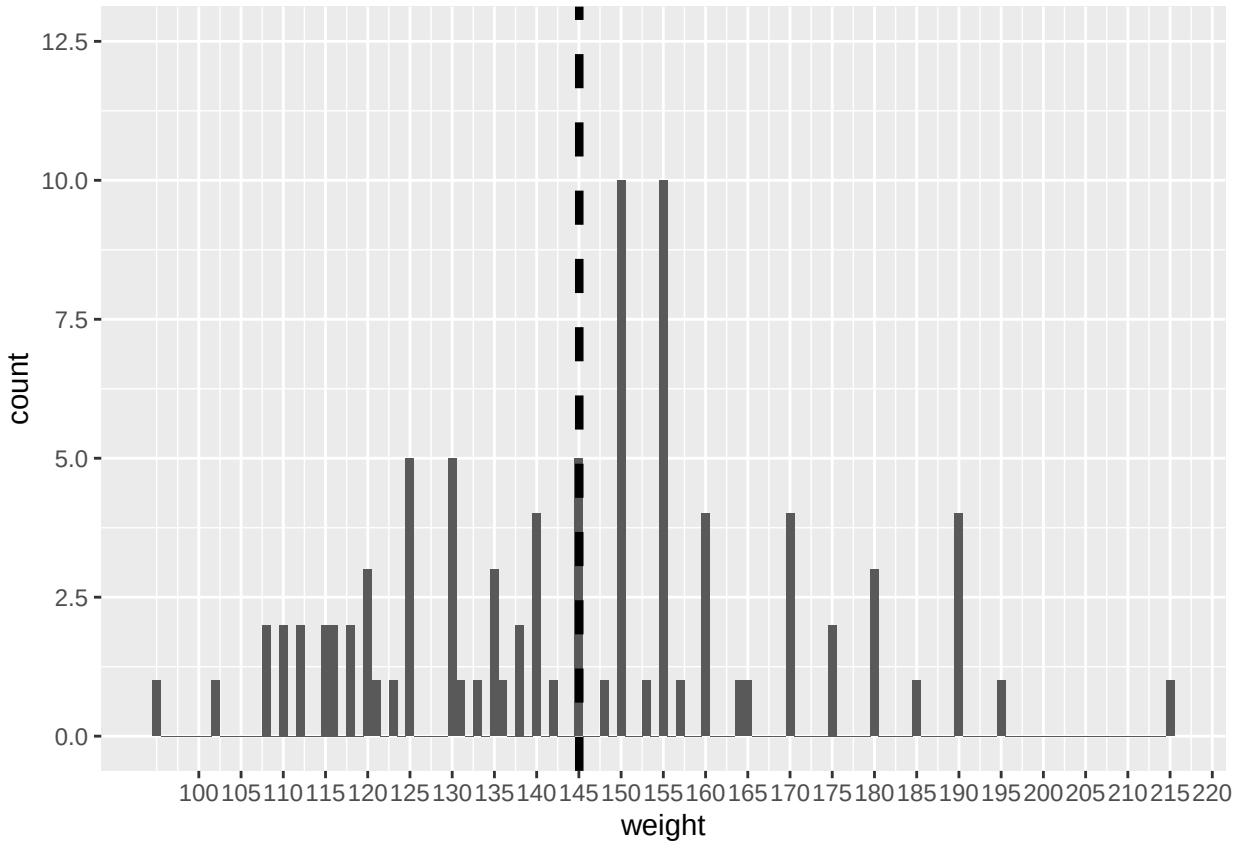


Цэнхэр зураас голчыг харуулж байна. Голч нь ердөө л эрэмбэлэгдсэн онооны дарааллын төв утга юм Ихэвчлэн гистограм нь хазайсан хэлбэртэй байвал голч нь дундаж утгаас их эсвэл бага байдаг. Энэ тохиолдолд өгөгдлийн багцыг 1-ийн утгаар баруун тийш нь хязгаарлах бөгөөд энэ нь ижил нөлөөг үүсгэж байна.

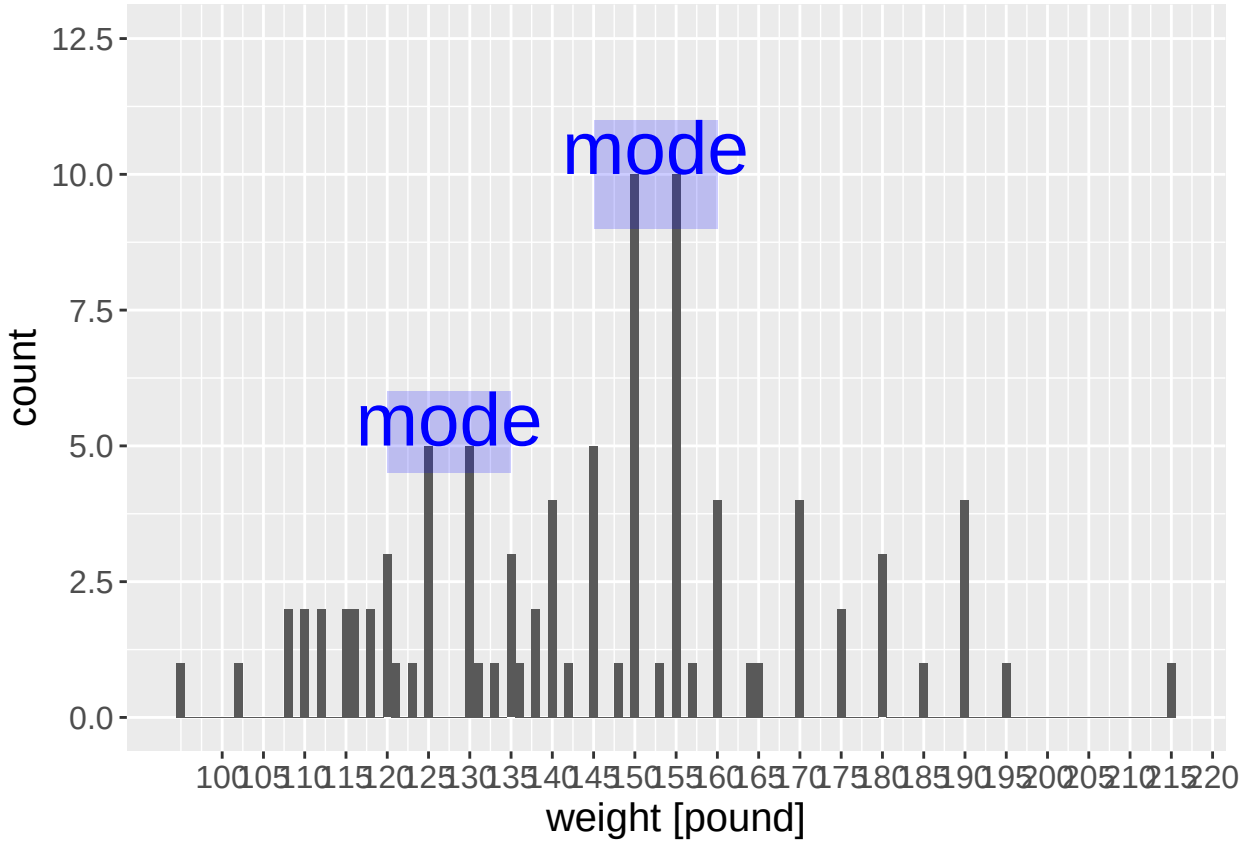
4 Reported Weights



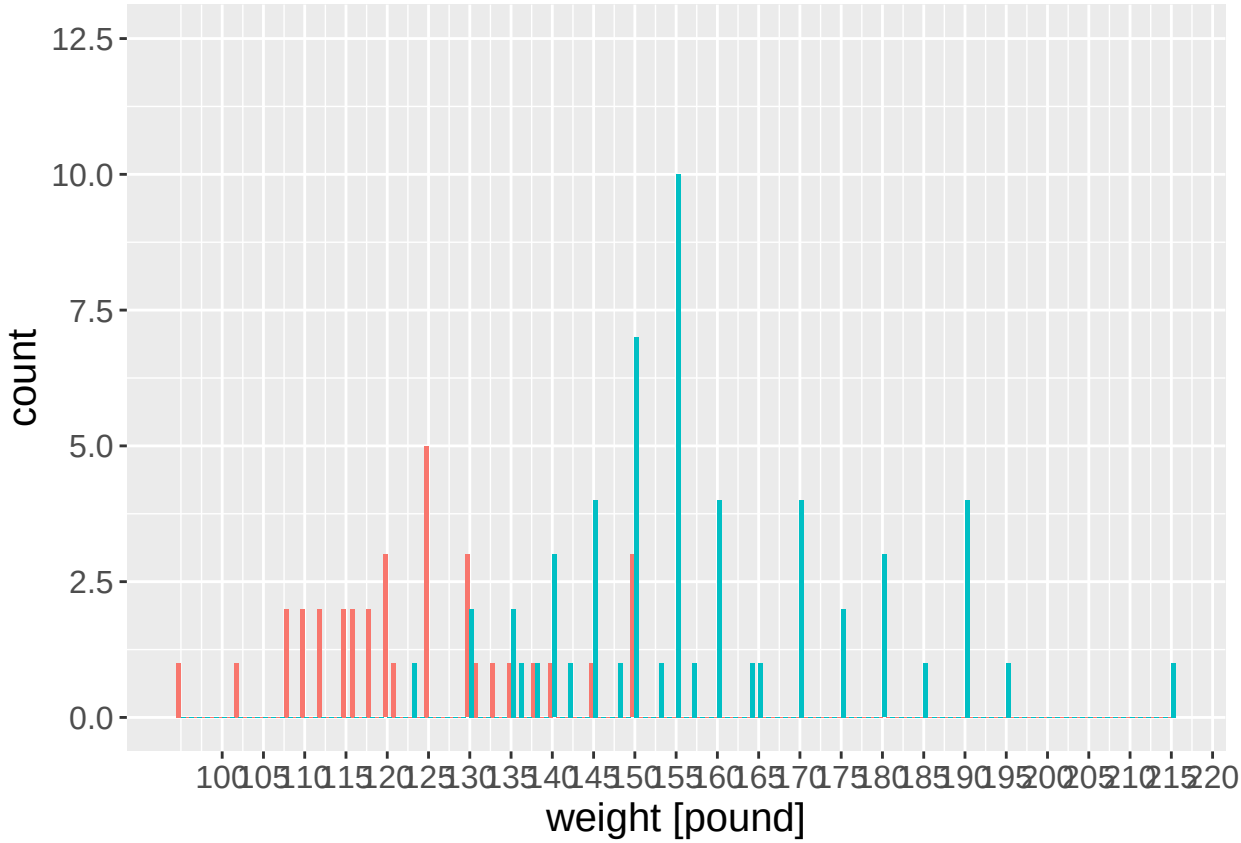
Одоо шинэ өгөгдлийн багцыг авч үзье. Энд та амаар мэдээлсэн биеийн жингийн гистограммыг харж байна. Өөрөөр хэлбэл коллежийн оюутнуудаас санамсаргүй түүврээр тэдний биеийн жинг асуусан болно. Энэхүү өгөгдлийн багцыг та өөрийн үгээр хэрхэн тодорхойлох вэ? Та ямар онцлог шинж чанарыг анзаарч байна вэ? Та тэдгээрийг хэрхэн тайлбарлаж байна вэ? Танд ямар асуулт байна вэ?



Энд би дахиад дундаж утгыг дүрслэсэн байгаа. Энэ нь 150 ба 155 гэсэн хамгийн их давтамжтай утгад байрлахгүй байгааг ажиглана уу. Энэ нь юу гэсэн үг вэ? Энэ гистограм нь маш их хоосон зайтай байгаа бөгөөд хүмүүс жингээ 5 фунтын шатлалаар хэлэх хандлагатай байдаг гэдгийг бас анхааараарай.



Энэхүү гистограм нь 2 моодтой болохыг анхаарна уу. Моодууд нь local maxima юм. Энэ нь юу гэсэн үг вэ?



Энэ мэдээллийн багцад хоёр бүлэг хүмүүс давамгайлж байгаа нь харагдаж байна: Та ямар хүмүүс гэдгийг таамаглаж байна уу? Сонирхолтой нь эдгээр 2 бүлгийн хооронд өөр ялгаа бий. Хоёр бүлгийн баарны хоорондох зай нь өөр өөр байгааг анхаарна уу. Энэ нь 110-180 фунт жингийн хооронд хамгийн тод илэрч байна. Хоёр өөр бүлэг нь үнэндээ эмэгтэй, эрэгтэй хүмүүс юм. Эмэгтэйчүүд эрэгтэйчүүдээс дунджаар хөнгөн байдаг бөгөөд жингийнхээ талаар илүү нарийвчлалтай эсвэл илүү их мэддэг болохыг нь баарны янз бүрийн зайнуудаас харж болж байна.

```
##
##   The decimal point is 1 digit(s) to the right of the |
##
##    9 | 5
##   10 | 288
##   11 | 0022556688
##   12 | 0001355555
##   13 | 0000013555688
##   14 | 00002555558
##   15 | 0000000000355555555557
##   16 | 000045
##   17 | 000055
##   18 | 0005
##   19 | 00005
##   20 |
##   21 | 5
```

Энэхүү өгөгдлийн багцын иш-навчны зургийг харахад сүүлийн санаа дэмжигдэж байна. Олон тэг, тавын тоонууд нь илүү бага биеийн жинд хэрхэн бага давамгайлж байгааг анзаараарай. Тиймээс санаарай: Хэрэв бид сайн харах юм бол гистограм нь маш их мэдээлэл өгөх боломжтой. Статистик график дээр харж байгаа зүйлдээ шүүмжлэлтэй хандахыг хичээ. Ихэнх тохиолдолд бүх зүйлс харагдаж байгаа шигээ энгийн байдаггүй. Дахиад энэ тохиолдолд өгөгдөл цуглуулах нь үр дүнд чухал үүрэг гүйцэтгэж байна.Учир нь бид жинг нь бодитоор хэмжихийн оронд хүмүүсээс асууж цуглуулсан

өгөгдлийг харж байна.

5 Probability Mass/ Density Functions

- Continuous random variables can take any value with arbitrary precision (e.g. weight in kg)
- The probability of getting a very precise and specific value is very small
- Why? Because the interval for this value would be very small/ close to zero
- In the limit of infinitely small intervalls, the probability becomes zero
- This means the area under a point is zero, however, this doesn't mean that the point has zero value
- The distribution of continuous variables is therefore described by probability **density** functions

Бидний сүүлчийн жишээнүүдэд тасралтгүй хувьсагчийн өгөгдөл багтсан болно. Тасралтгүй хувьсагчдын хувьд бид магадлалын нягтын функцийг одоо танилцуулах болно. Эхлээд үүнийг энгийнээр тайлбарлаж , дараа нь математикийн тодорхойлолтыг нь өгөх болно. Тасралтгүй хувьсагч нь бодит тооны утгыг авдаг тул нарийвчлалаар хязгаарлагдахгүй. Дээрх амаар мэдээлсэн жингийн жишээнд бид 5-н алхамаар бүхэл тоонд цуварсан утгуудыг авч үзсэн. Гэсэн хэдий ч жин нь өөрөө тасралтгүй хувьсагч бөгөөд хүний бодит жинг зарчмын хувьд хязгааргүй нарийвчлалтай хэмжиж болох юм, ж.нь. 60.345215675 кг. Хэрэв бид хүмүүсээс хэт нарийн хэмжээсээр хэмжсэн бодит жингийн маш том дээж цуглуулах боломжтой байсан бол бид маш нарийн интервал бүхий гистограм үүсгэж болох бөгөөд ингэсэн тохиолдолд интервал бүрт нь гистограмыг хэлбэржүүлэх хангалттай өгөгдлийн цэгүүд байсаар л байх байсан. Хэрэв бид үүнийг хязгааргүй нарийвчлал болон түүврийн хэмжээ хязгааргүй гэж үзвэл хязгааргүй жижиг интервалтай байж болох юм. Хэтэрхий олон интервал бүхий гистограммын хэлбэр нь баар шиг шатлалын оронд тасралтгүй шугам шиг харагдана. Гэсэн хэдий ч хэрэв интервал нь хэтэрхий нарийхан байвал энэ интервалын утга бүхий ямар ч тасралтгүй хувьсагчийн магадлал нь тэг болно. Үүнийг хялбарчлах юм бол: "Дэлхий дээр хэн ч яг 60.34525675 кг жинтэй байдаггүй. Гэсэн хэдий ч жин нь 60.34 - 60.35 кг хооронд байдаг хүн байдаг нь гарцаагүй "гэж хэлж болно. Хувьсагчийн магадлалыг олж авахын тулд бид өмнөх шигээ эдгээр утгуудын интервалыг тодорхой заах хэрэгтэй. Энэ магадлал нь энэхүү хязгааргүй жижиг интервал бүхий "гистограм" -ын доорх талбай юм. Үүнтэй адилаар бид өмнө нь баарны талбайг хэмжиж байсан: өндрийг хэмжиж, өргөнөөр нь үржүүлсэн. Гэсэн хэдий ч баарны өргөн нь бүх бааранд тэнцүү байсан бөгөөд нэгжийн утгатай өөрөөр хэлбэл е. 1 гэж тодорхойлж болно. Хязгааргүй жижиг интервалтай, хязгааргүй том түүвэр бүхий гистограмыг бид магадлалын нягтын тархалт гэж нэрлэдэг. Сая тайлбарласан зүйлийг илүү онолын үүднээс тайлбарлаж өгье.

For discrete variables, the probability of X can be determined by summation over the probability **mass** function of x:

$$P\left(a \leq x \leq b\right) = \sum_{x:a \leq x \leq b} p\left(x\right)$$

For continuous variables, where intervals are infinitely small, the summation becomes an integral over the probability **density** function of x:

$$P\left(a \leq x \leq b\right) = \int_{a}^{b} p\left(x\right)$$

Properties of a probability density function (pdf) $p\left(x\right)$ are :

$$p\left(x\right) \geq 0$$

$$\int_{-\infty}^{\infty} p\left(x\right) dx = 1$$

$$P\left(x = a\right) = \int_{a}^{a} p\left(x\right) dx = 0$$

Probabilities are defined over intervalls

$$P\left(a \leq x \leq a + \delta\right)$$

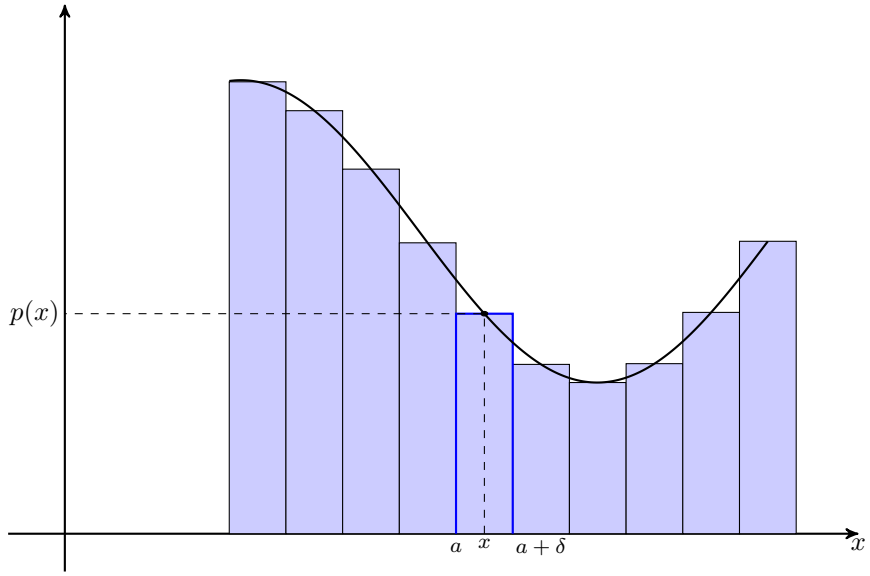
where we define an interval $[a, a + \delta]$ of length δ and let $\delta \geq 0$ and *small*, then can approximate the probability as

$$P\left(a \leq x \leq a + \delta\right) \approx p(a)\delta$$

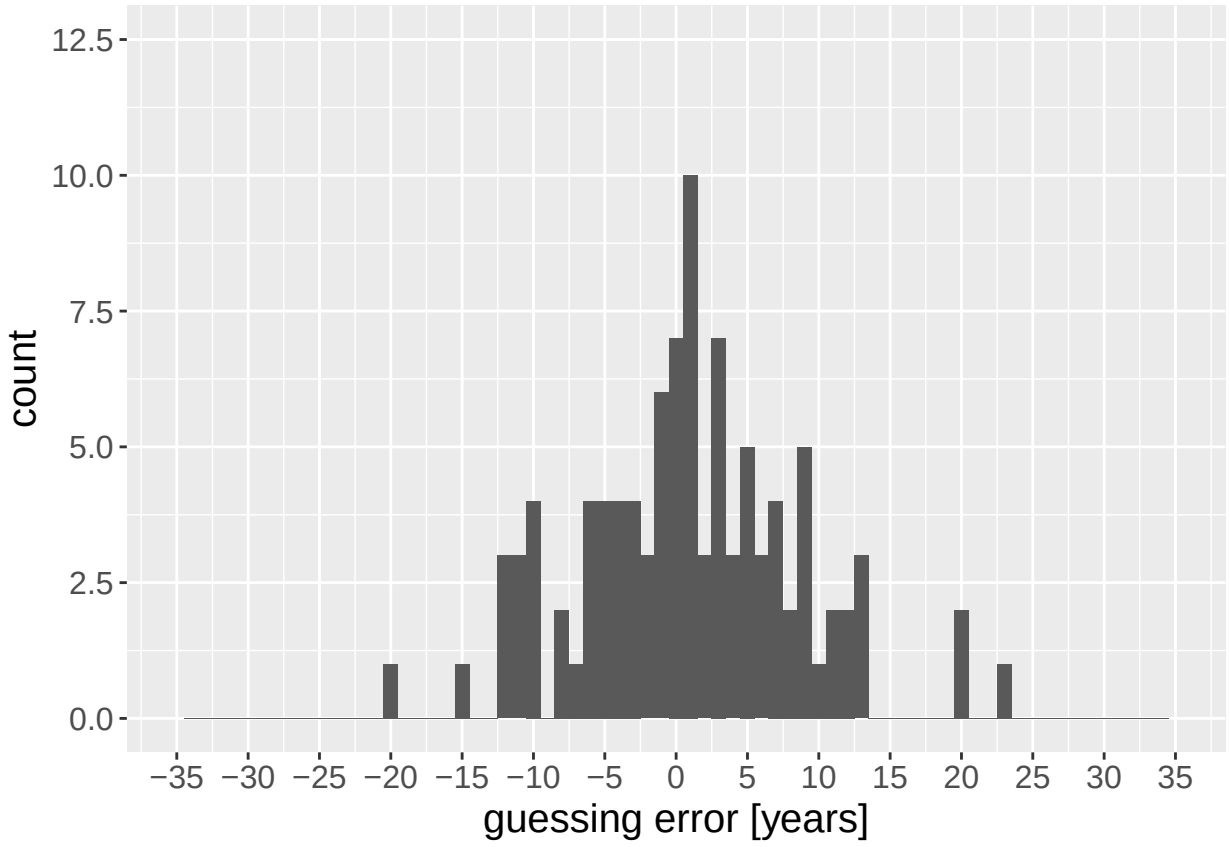
and isolate $p\left(x\right)$ on the right-hand side

$$p\left(x\right) = P\left(a \leq x \leq a + \delta\right)/\delta$$

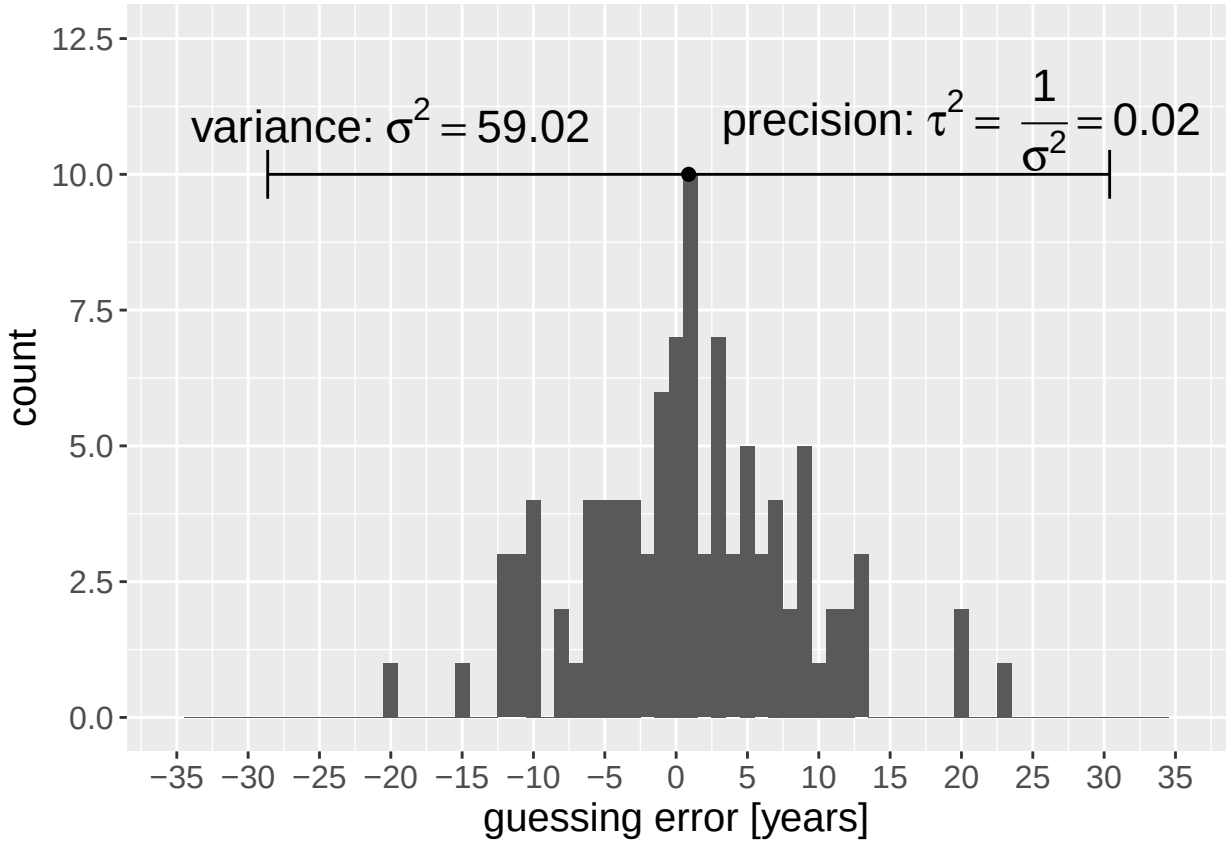
$p\left(x\right)$ is therefore called probability density: a probability per unit length



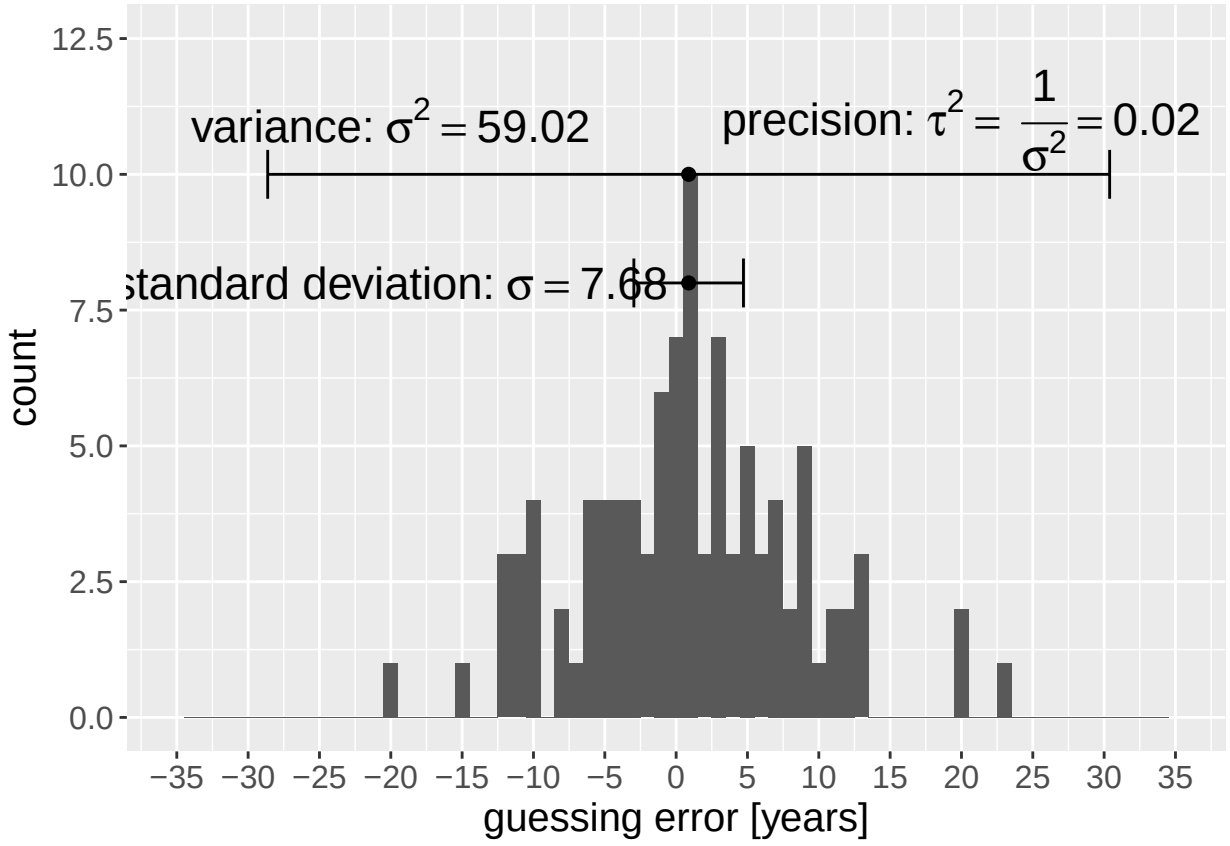
6 Age guessing error



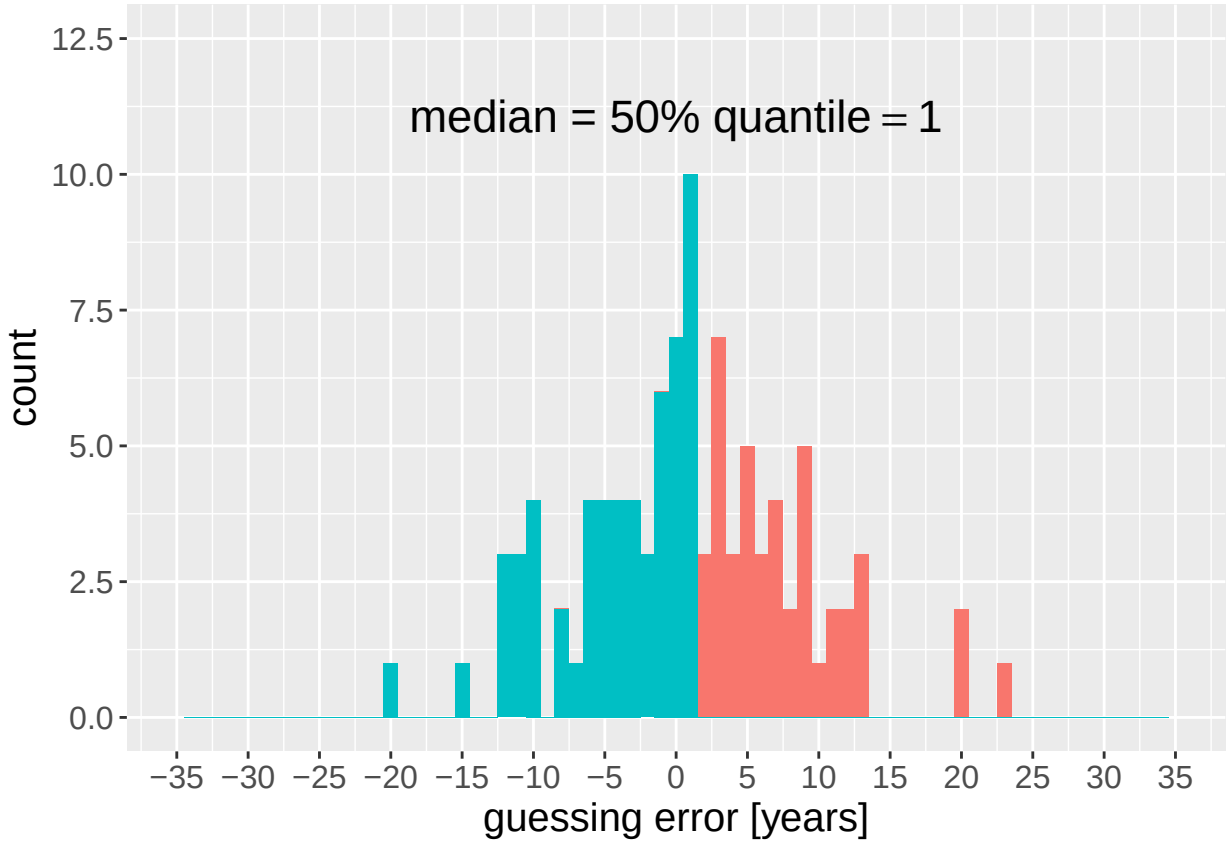
Өнгөрсөн долоо хоногт бид нас таах туршилтыг хийж, нас таах алдааны талаарх мэдээллийг цуглуулсан. Энэ бол насыг таамагласан алдааны гистограм бөгөөд бид одоо цааш нь нарийвчлан задлан шинжилж болно. Хэрэв та санаж байгаа бол туршилт эхлэхээс өмнө би “Та хүний насыг таахдаа хэр сайн бэ?” гэж асуусан.



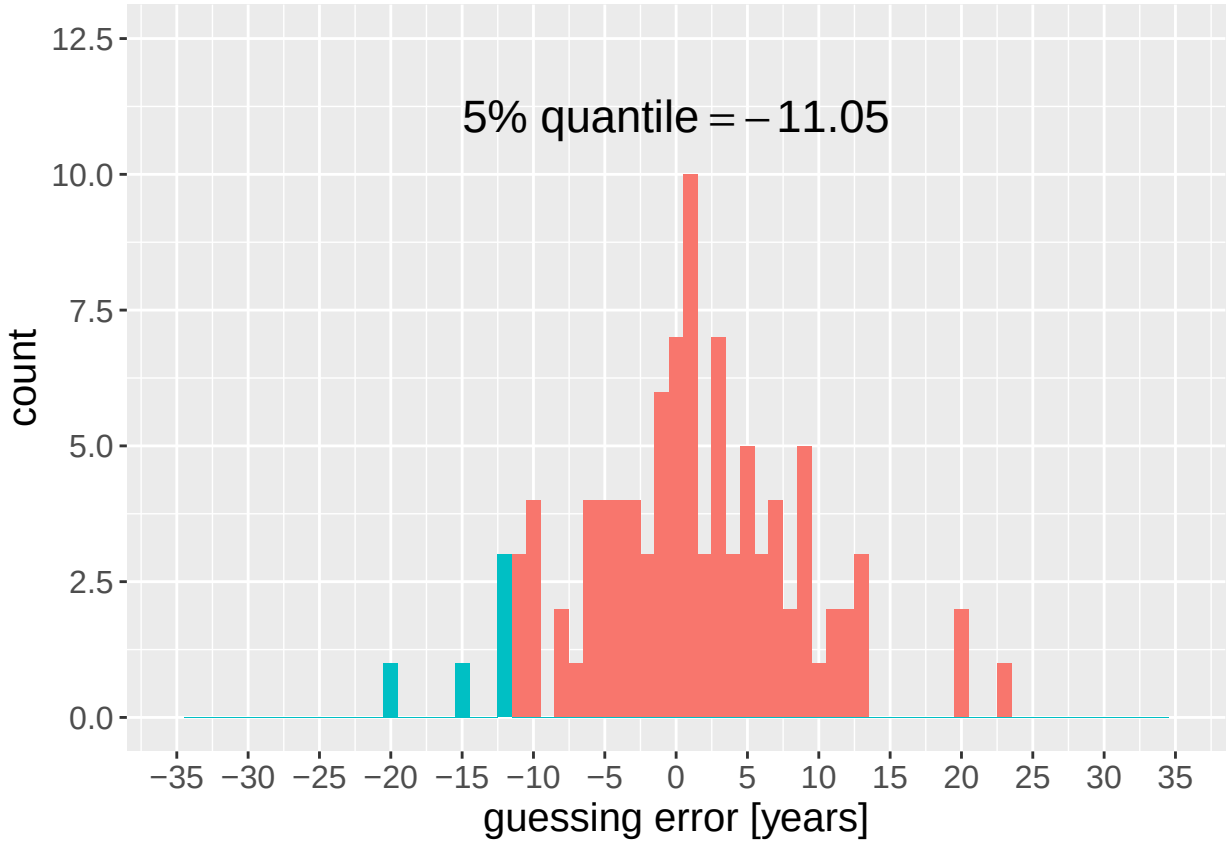
Энэ төрлийн асуултанд хариултаа илэрхийлэх нэг боломж бол дисперс юм. Дисперс нь дундаж квадрат алдаа гэж тодорхойлогддог (самбар дээр бичих). Гистограмын орой дээрх алдааны баар нь дисперстэй тэнцүү урттай байна. Төвд нь байгаа цэг нь дундаж утга юм. Энэ тохиолдолд дисперс нь өгөгдлийн далайцаасаа их болохыг анхаарна уу. Учир нь энэ нь квадрат дундаж алдаа юм. Тиймээс энэ нь мэдээллийн тархацтай бараг хамаагүй юм. Гэсэн хэдий ч дисперс нь зарим математикийн шинж чанаруудаас шалтгаалан мөн дисперсийг олон төрлийн магадлалын тархалтын параметр болгон байнга ашигладаг тул маш чухал статистик юм. Дисперсийн урвууг нарийвчлал гэнэ. Үүнийг бас заримдаа тархалтыг дүрслэхэд ашиглахад бодитой байдаг. Нэршлээс нь аль хэдийн мэдсэнчлэн, дисперс нь бага байхад нарийвчлал нь их утгатай байдаг ба эсрэгээрээ дисперс нь их байхад нарийвчлал нь бага утгатай байдаг. Нарийвчлал нь санамсаргүй хувьсагчийг хэр нарийвчлалтайгаар урьдчилан таамаглаж болохыг харуулдаг.



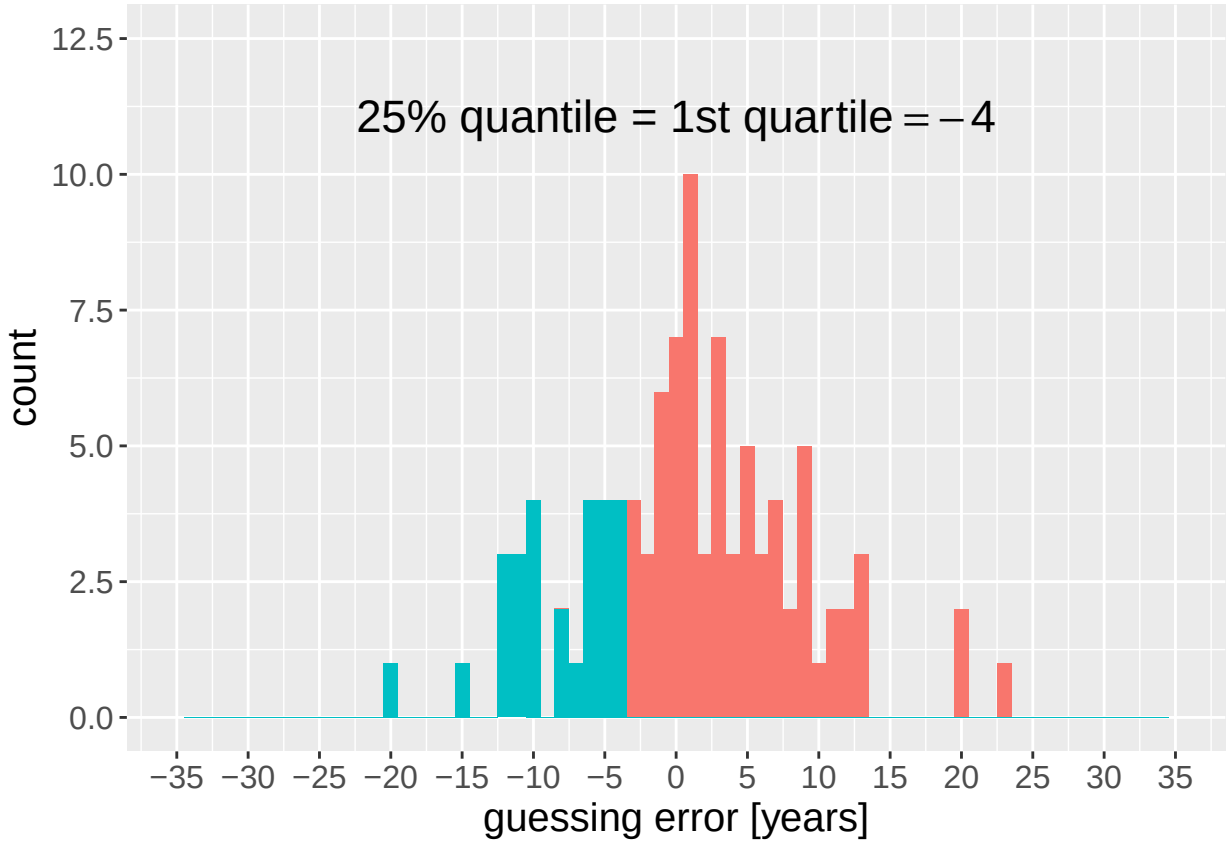
Нүдэн баримжааны хувьд илүү мэдээлэгч хэмжүүр бол стандарт хазайлт юм. Энэ нь дисперсийн квадрат язгууpаар тодорхойлогдоно. Энэ тохиолдолд стандарт хазайлт нь хамгийн олон давтамжтай үр дүнгийн талбарыг хамардаг бөгөөд бидний анхны асуултанд хариулт болж өгөх болно: Бидний таамагласан алдаа нь -2.96-4.72 хооронд хэлбэлздэг.



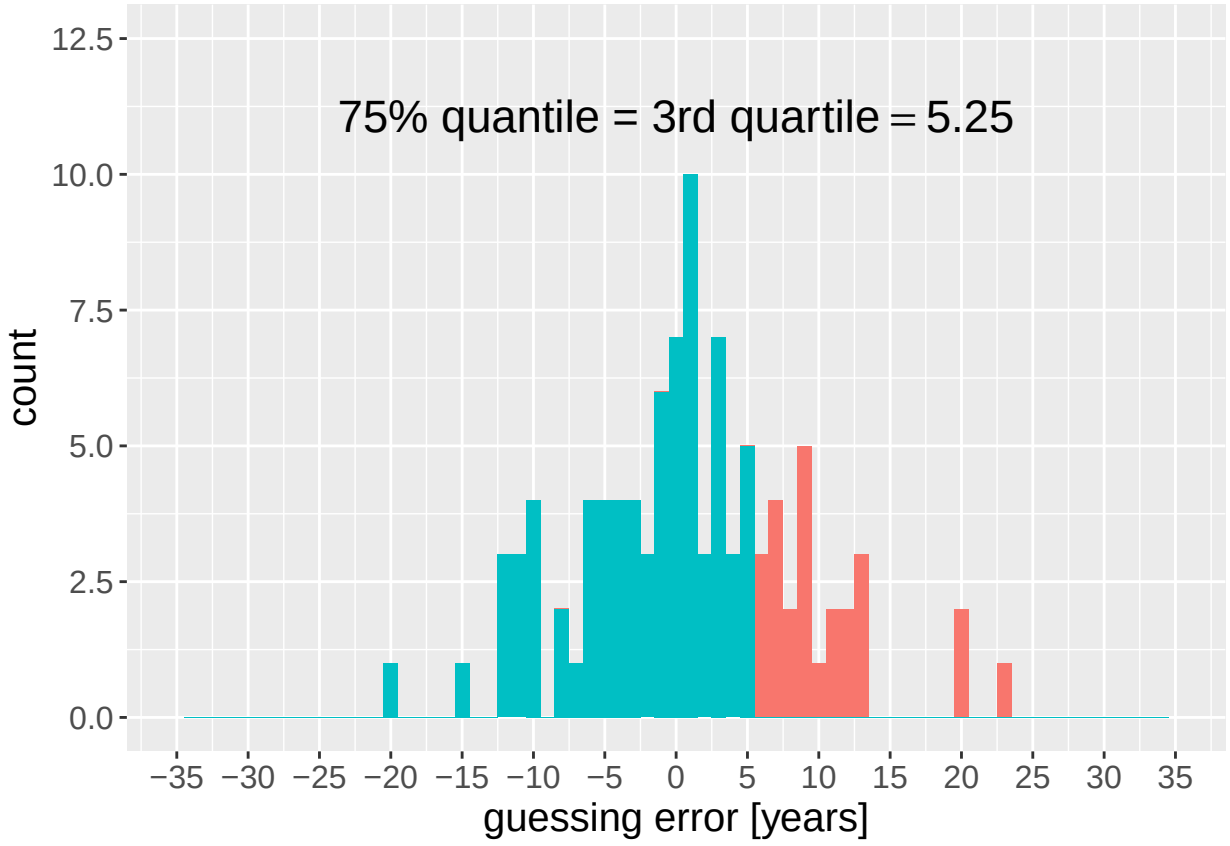
Өгөгдлийн багцыг нэгтгэн дүгнэх илүү олон хувилбар байдаг. Байршлын хэмжүүрийн хувьд бид өгөгдлийн утгын эрэмбэлэгдсэн дарааллын төв цэг гэж тодорхойлсон median/голч-ыг тооцоолж болно (жишээг самбарт бичих). Байршлын хэмжүүрийн хувьд бид өгөгдлийн утгын эрэмбэлэгдсэн дарааллын төв цэг гэж тодорхойлсон median/голч-ыг тооцоолж болно (жишээг самбарт бичих). Median/Голч-ыг заримдаа 50% -ийн квантил гэж нэрлэдэг.



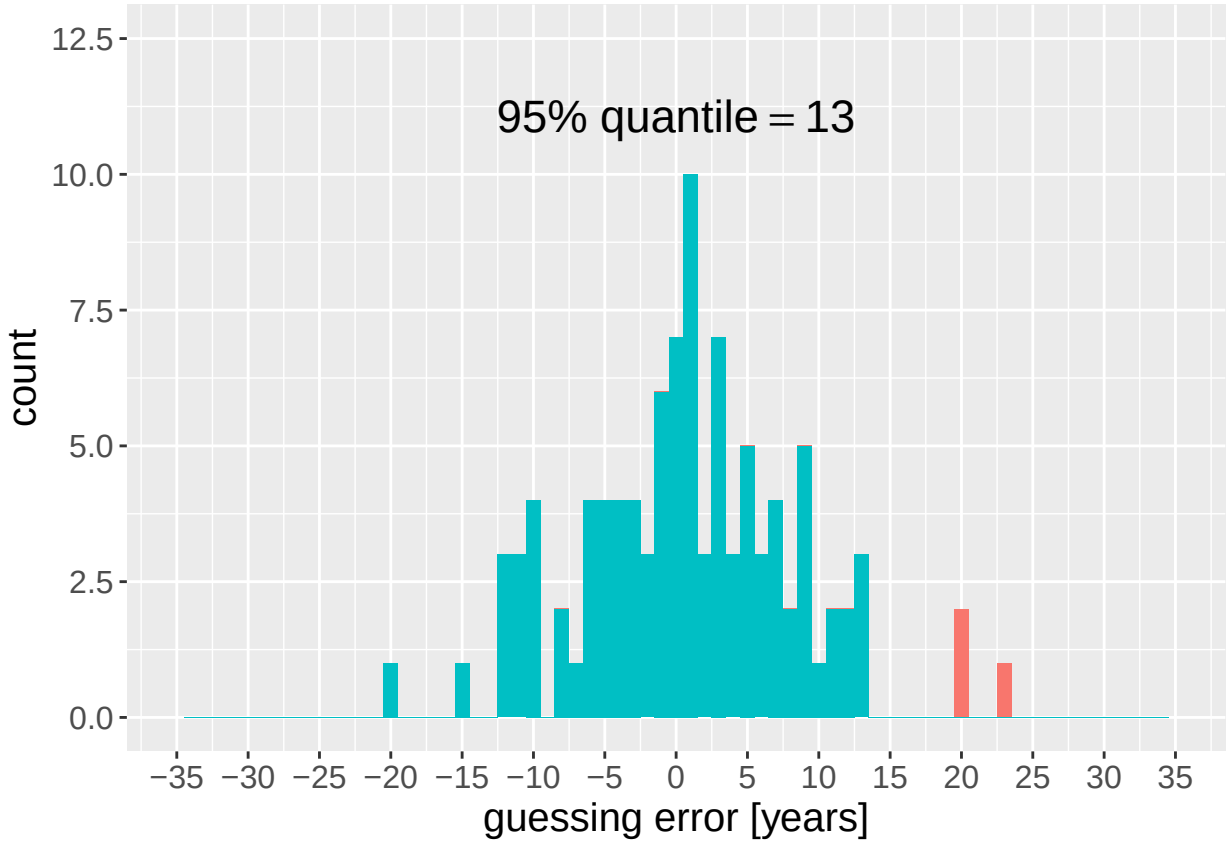
Та янз бүрийн квантилийг тооцоолж болно, ж.нь. 5% -ийн квантил, өгөгдлийн 5% нь тэнцүү эсвэл бага байх цэгийг хэлнэ. Үүнийг график дээр цэнхэр баараар тэмдэглэв.



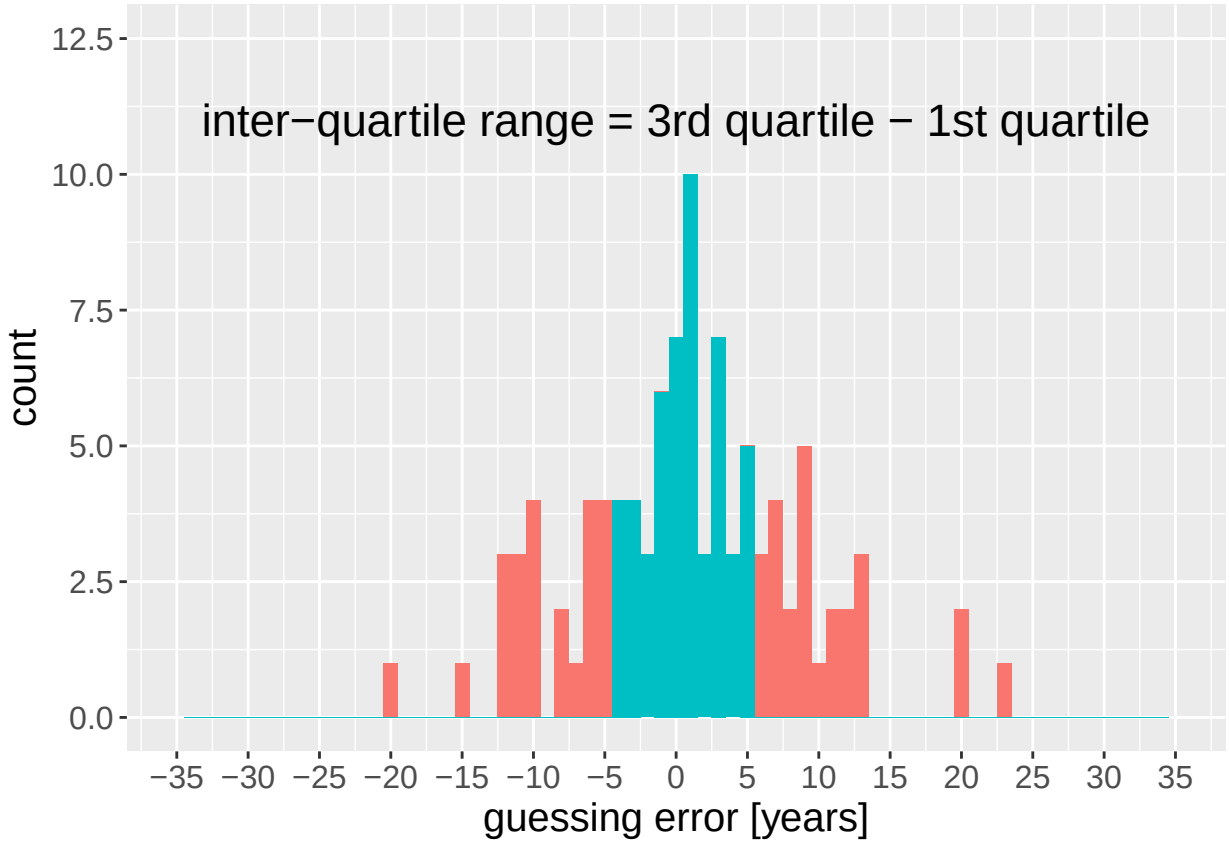
Эсвэл үүнтэй адил өгөгдлийн дөрөвний нэг нь тэнцүү эсвэл бага байх цэг болох 25% квантил юм. 25% -ийн квантил нь онцгой зүйл юм. Учир нь үүнийг 1-р квартиль гэж нэрлэдэг тул та өгөгдлийн 4-ний нэг буюу 1-р хагасийг энэ цэгээс доогуур байна гэж хэлж болно.



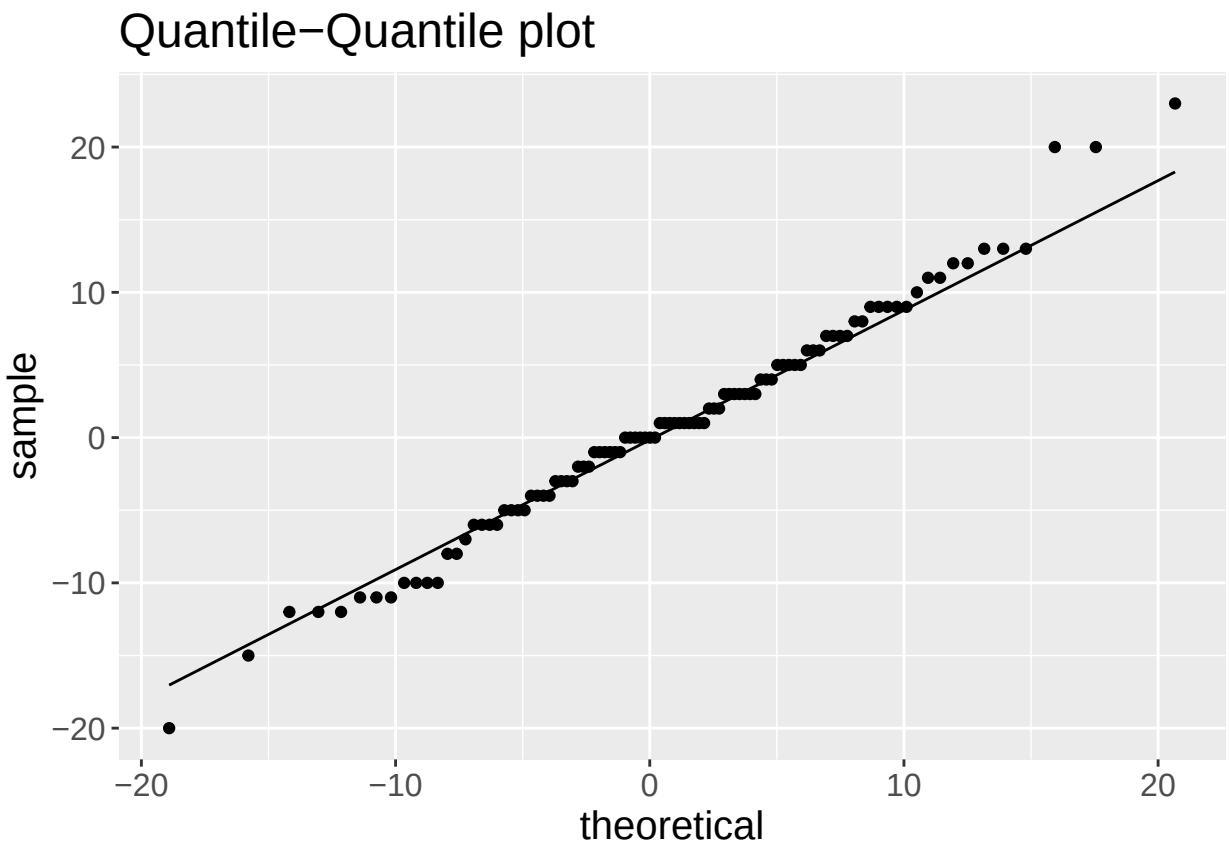
Үүнтэй адилаар бид 3-р квартиль гэж нэрлэгддэг 75% -ийн квантилийг тодорхойлж болно, учир нь та өгөгдлийн дөрөвний гурвыг энэ цэгээс доогуур гэж хэлж болно.



Өөр нэг түгээмэл хэрэглэгддэг квантиль бол 95% -ийн квантил бөгөөд энэ нь аяндаа бараг бүх өгөгдлийг хамардаг. Хэрэв бид параметрийн тархалтын талаар ярьж байгаа тохиолдолд таамаглалын тест гүйцэтгэхэд ашиглаж болно. Жишээ нь 13 ба түүнээс дээш утгыг авах боломж нь 5%-ийн ач холбогдлын түвшинд статистик ач холбогдолтой гэж томъёолж болно.



Квантилийг мөн түүнчлэн тодорхой магадлалын массыг өгөгдсөн өгөгдлийн тодорхой далайцыг тодорхойлоход ашиглаж болно. Үүний нэг онцгой тохиолдол бол "квартил хоорондын далайц" гэж нэрлэгддэг зүйл юм. Энэ бол зүгээр л 3-р квантилаас 1-р квартилийг хассан утга юм. Тиймээс энэ нь голчоос гадагш чиглэсэн өгөгдлийн 50%-ийг хамардаг.



Өнгөрсөн долоо хоногт та бүхний нэг нь өгөгдөлийг "хэвийн харагдаж байна" гэж хэлж байсан. Тэгж хэлсэнд баярлалаа. Энэ үнэн эсэхийг шалгахыг санал болгож байна. Өнөөдрийн энэ сүүлчийн график нь квантил - квантил график гэж нэрлэгддэг графикийг харуулсан байгаа. Та ердийн тархалт гэх мэт онолын тархалттай тохирч байгаа эсэхийг шалгахын тулд энэ төрлийн графикийг ашиглаж болно. Графикийг ийм байдлаар үүсгэсэн болно: Дээжээс бид тодорхой тооны квантилийг тодорхойлно, ж.нь. 1%, 2%, 3% -иас 99%-ийн квантил хүртэл. Бид түүврийн дундажтай тэнцүү дундаж болон түүврийн дисперстэй тэнцүү дисперсийг тохируулах онолын хэвийн тархалтын хувьд ижил зүйлийг хийж болно. Дараа нь бид цэг тус бүрийн түүвэр ба онолын квантил тус бүр дээр тааруулж цэг тус бүрийн х бүрэлдэхүүн хэсэг нь онолын квантил, у бүрэлдэхүүн хэсэг нь түүврийн квантил байхаар тархалтын график зургийг зурна. Одоо түүврийг ердийн тархалтаар тодорхойлж болохоор бол цэг шулуун шугамын эргэн тойронд байрлана. Нэмж дурдахад бид тэр шугамыг нэгтэй тэнцүү налуугаар зурж болно. Та юу гэмээр байна? Бидний түүвэр хэвийн харагдаж байна уу?

7 Summary

- Data coding and formatting
- Numerical statistics:
  - scale/ spread:
    - variance/ precision (parameter)
    - standard deviation used as data summary
  - location:
    - mean (parameter and data summary)
    - median (data summary, more stable, less susceptible to extrem values)
    - mode (property of a distribution)
  - measures of distribution: quantiles/ percentiles/ quartiles
- Probability distribution:

- probability **mass** distribtuion: discrete variables
- probability **density** distribtuion: continuous distribution
- properties of pmfs/ pdfs:
  - total area/ integral over domain: $\int_{-\infty}^{\infty} p(x)\, dx = 1$
  - pdfs: $P(x = a) = 0$

References

Andrew Gelman and Deborah Nolan. *Teaching statistics: A bag of tricks*. Oxford University Press, 2017.