Introduction to Statistics and R

Descriptive Statistics - Part II

Eric Stemmler

Khovd University

03.02.2021

Section 1

Recap

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

Section 2

Learning Goals

Learning Goals

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

Section 3

Handedness Data Analysis

Handedness Data Analysis

Table 1: Complete collected data set from the handedness inventory

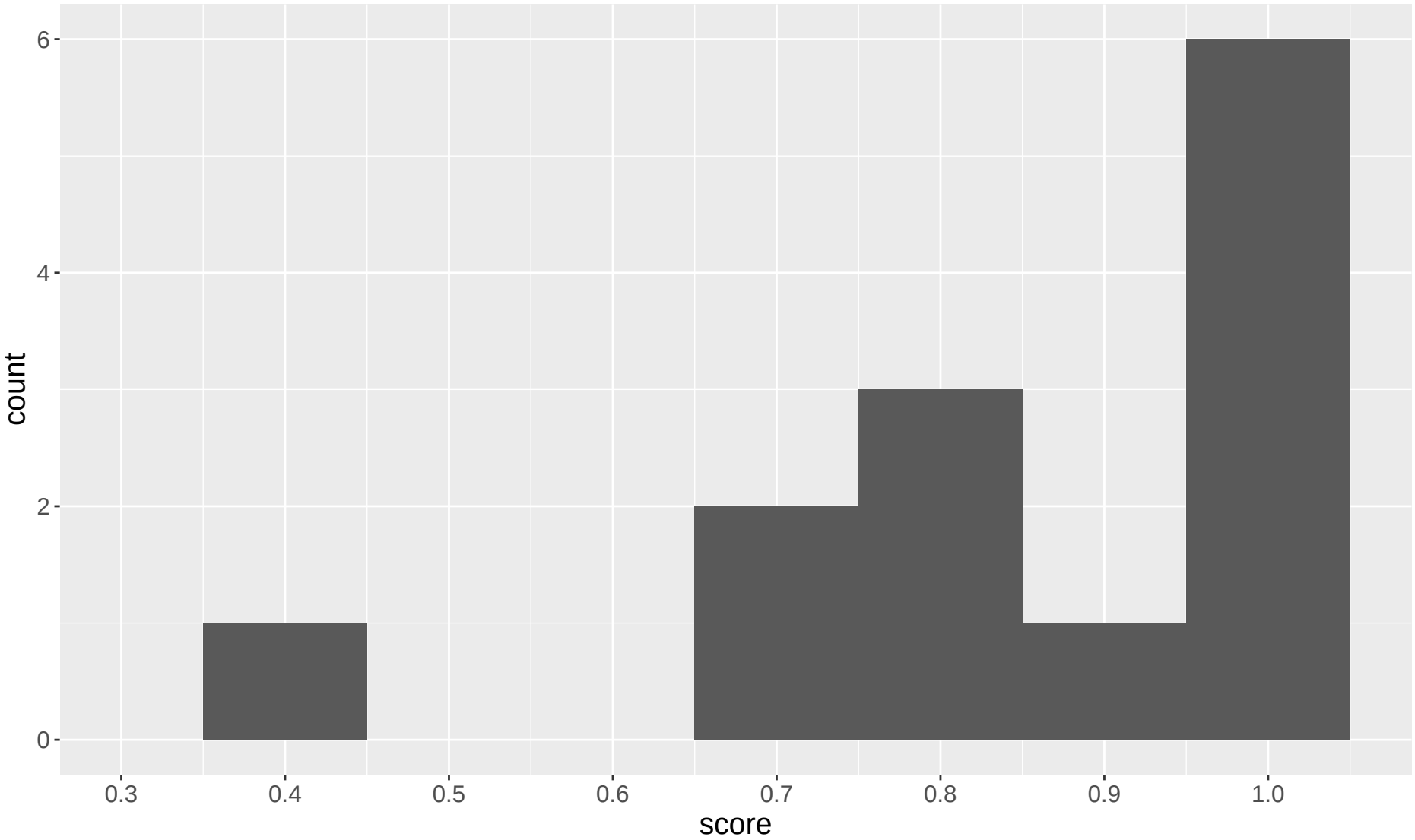
id	writing	drawing	throwing	scissors	toothbrush	knife	spoon	broom	match	box	calc	score	right	left
1	rr	rr	rr	rr	rr	rr	rr	rr	rr	rr	1	1.0000000	20	0
2	r	r	r	r	r	r	r	r	r	r	NA	1.0000000	10	0
3	rr	rr	r	rr	rr	rr	rr	rr	rr	rr	NA	1.0000000	19	0
4	rr	rr	rr	rr	rr	rr	rr	rr	rr	rr	NA	1.0000000	20	0
5	rr	rr	rr	rr	rr	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

Handedness Data Analysis

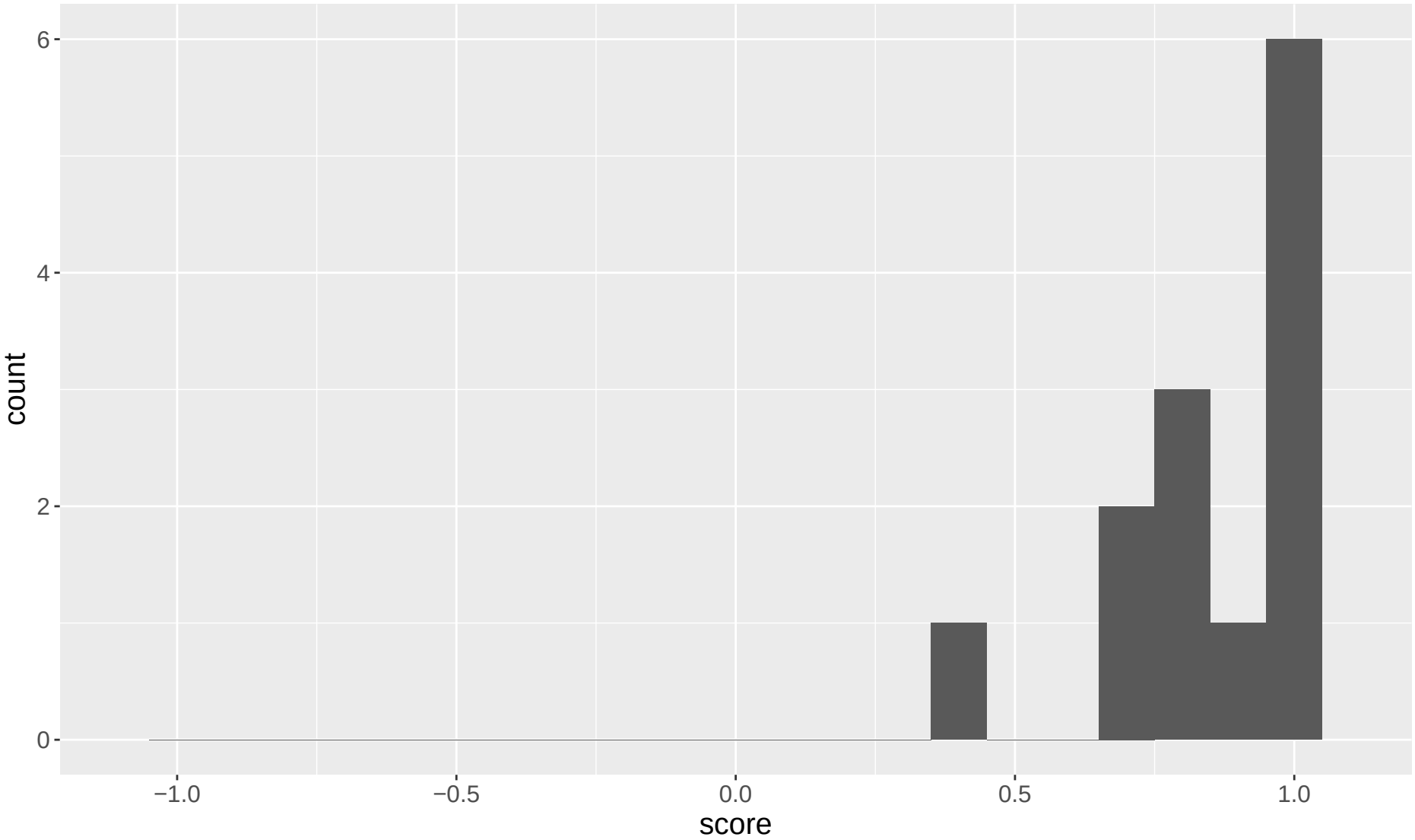
Table 1: Complete collected data set from the handedness inventory

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

Handedness Data Analysis



Handedness Data Analysis



Handedness Data Analysis

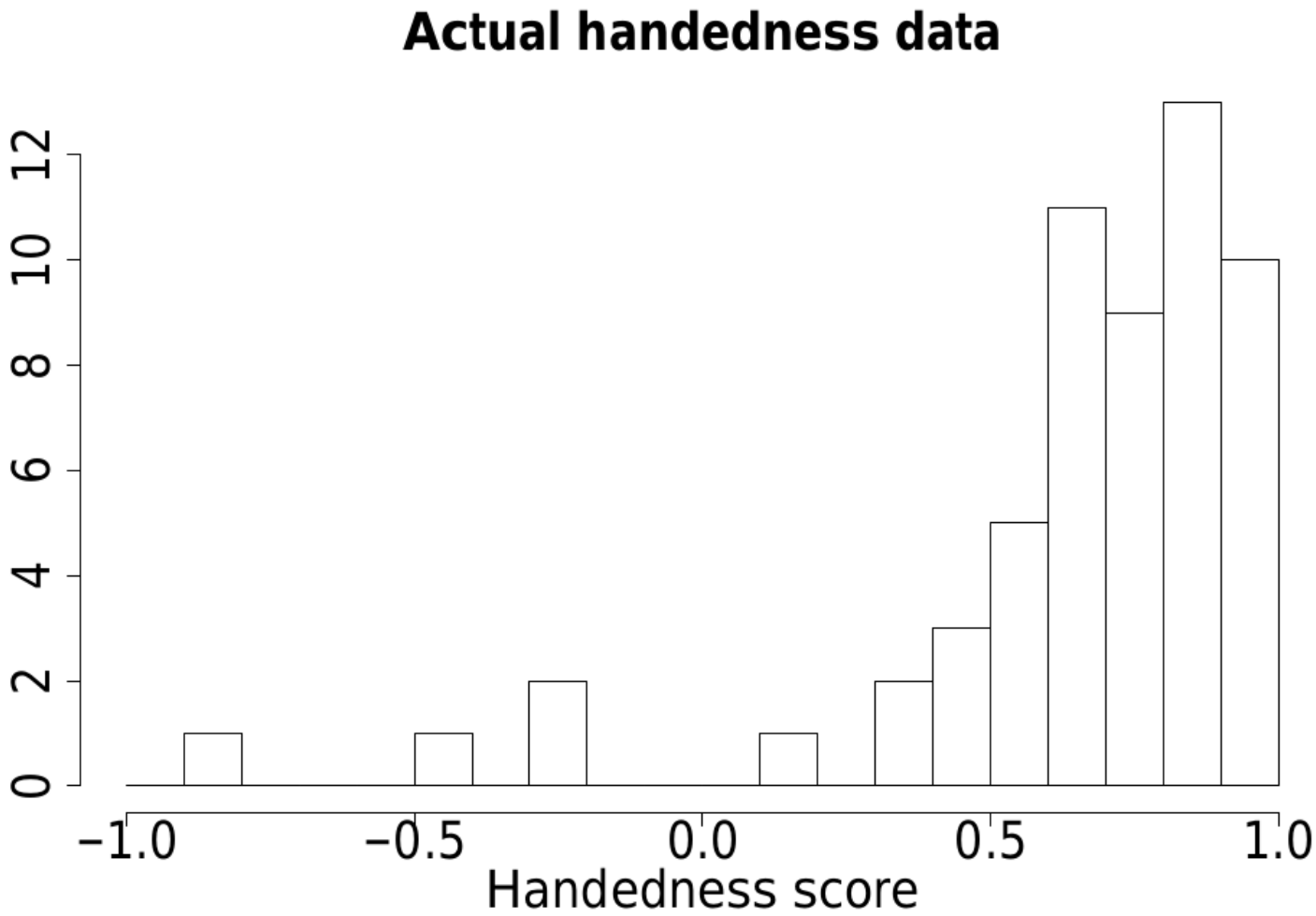
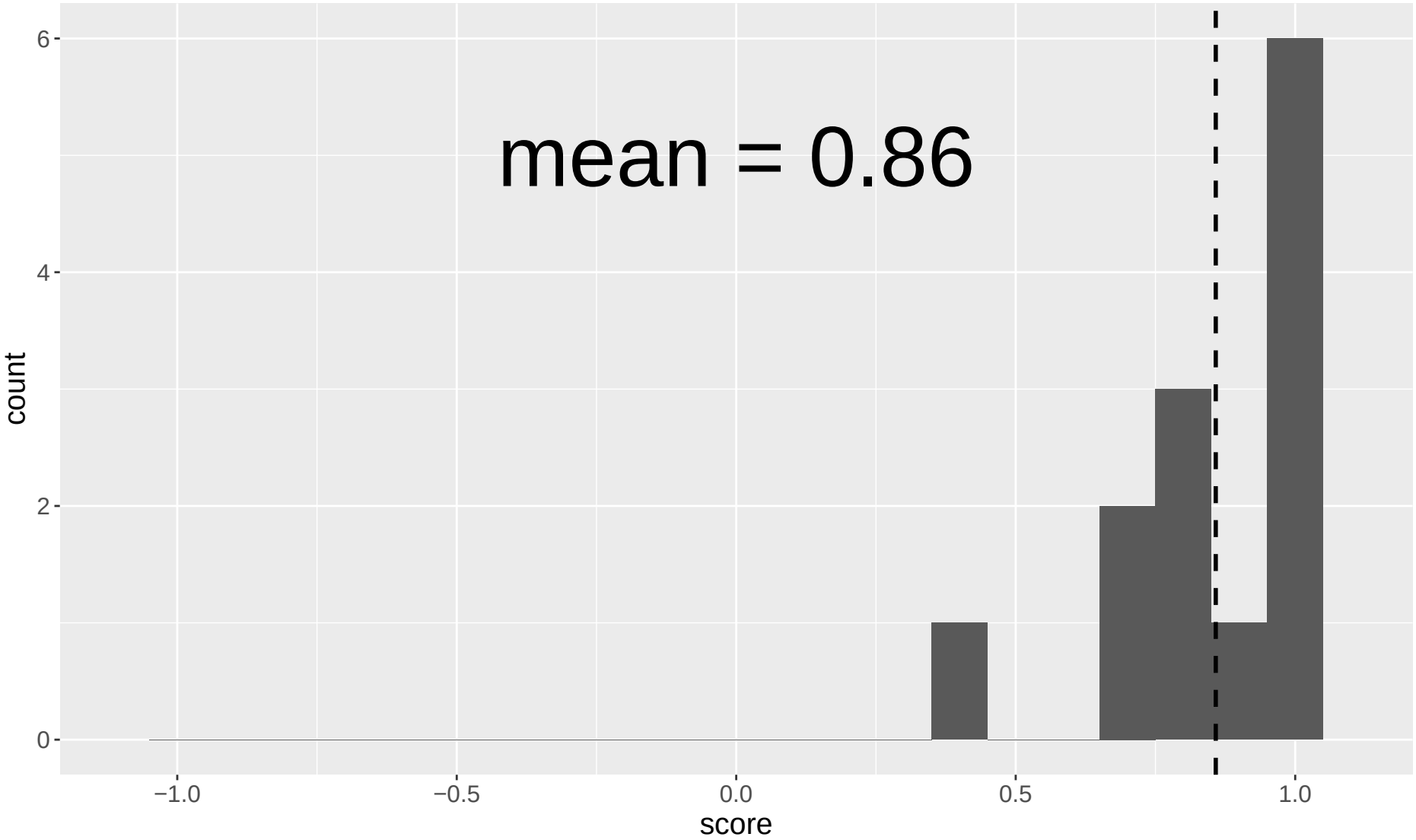
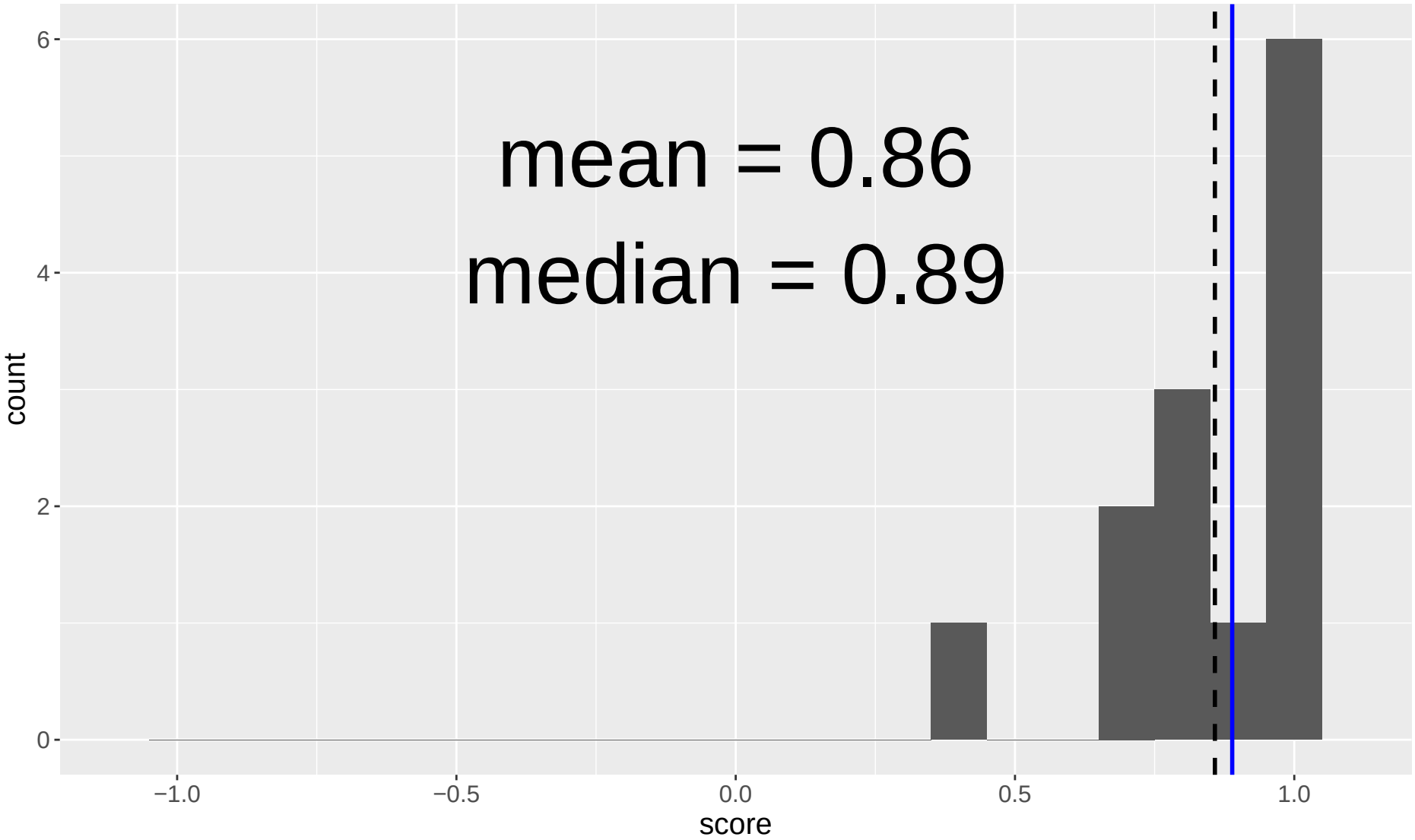


Figure 1: From Gelman and Nolan (2017)

Handedness Data Analysis



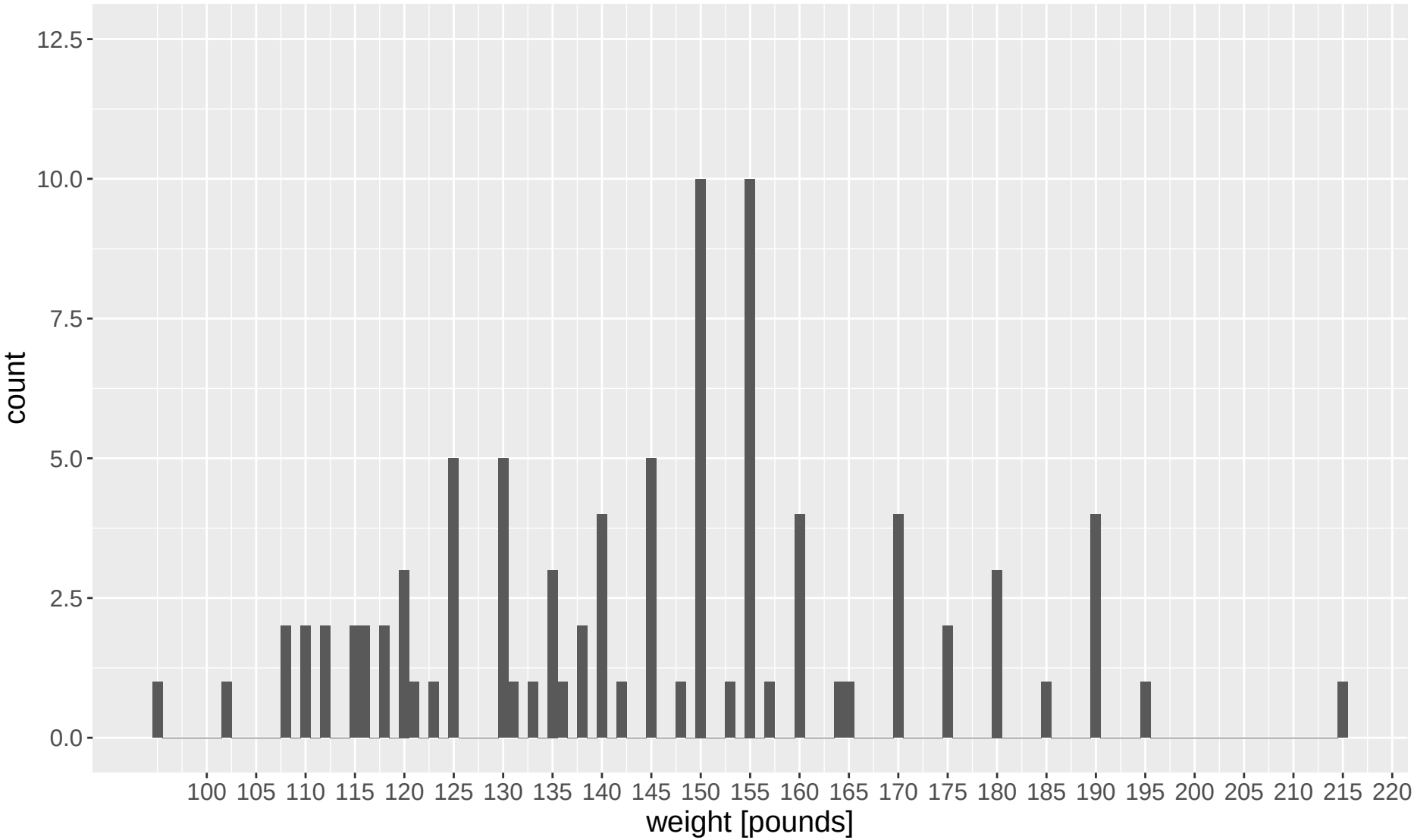
Handedness Data Analysis



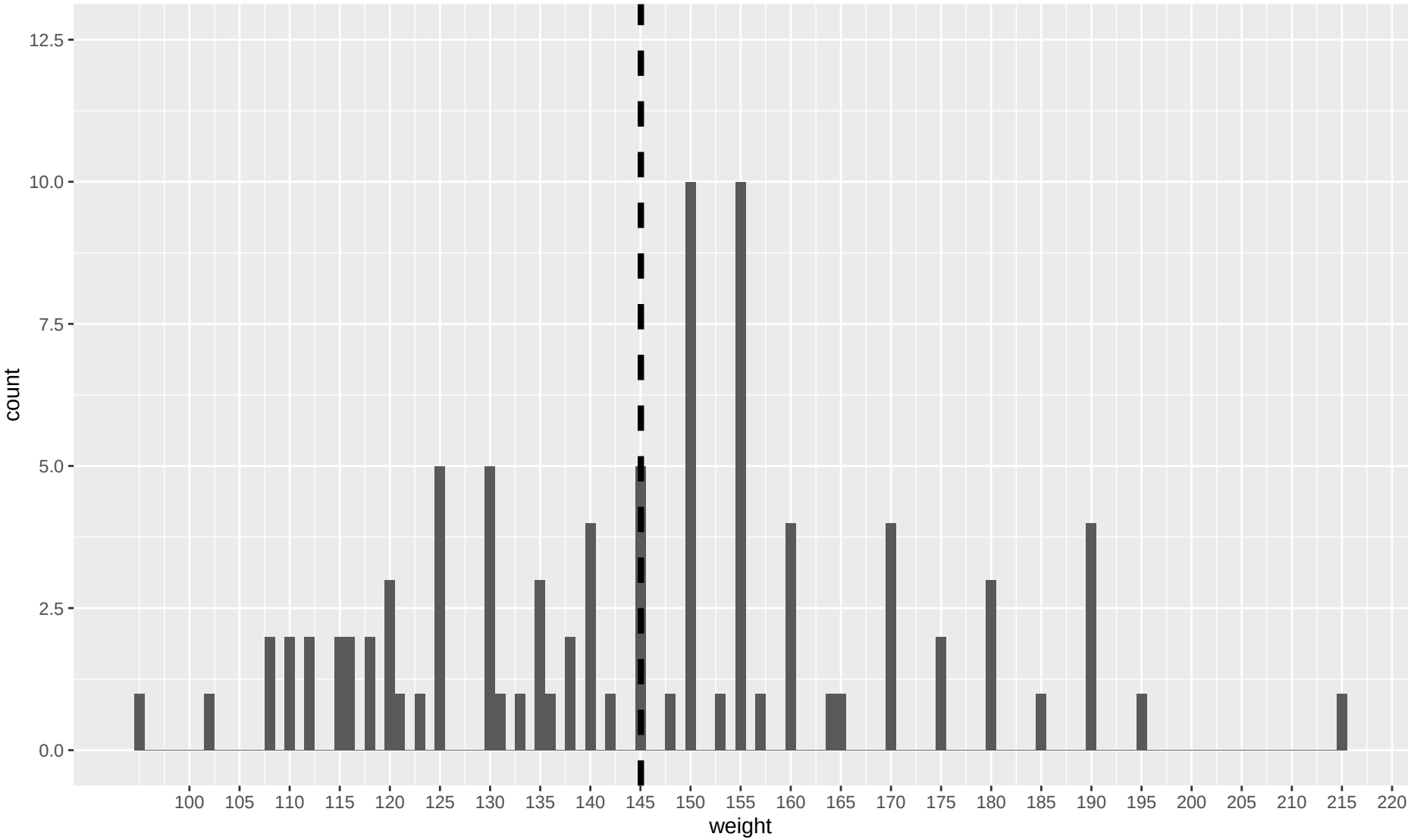
Section 4

Reported Weights

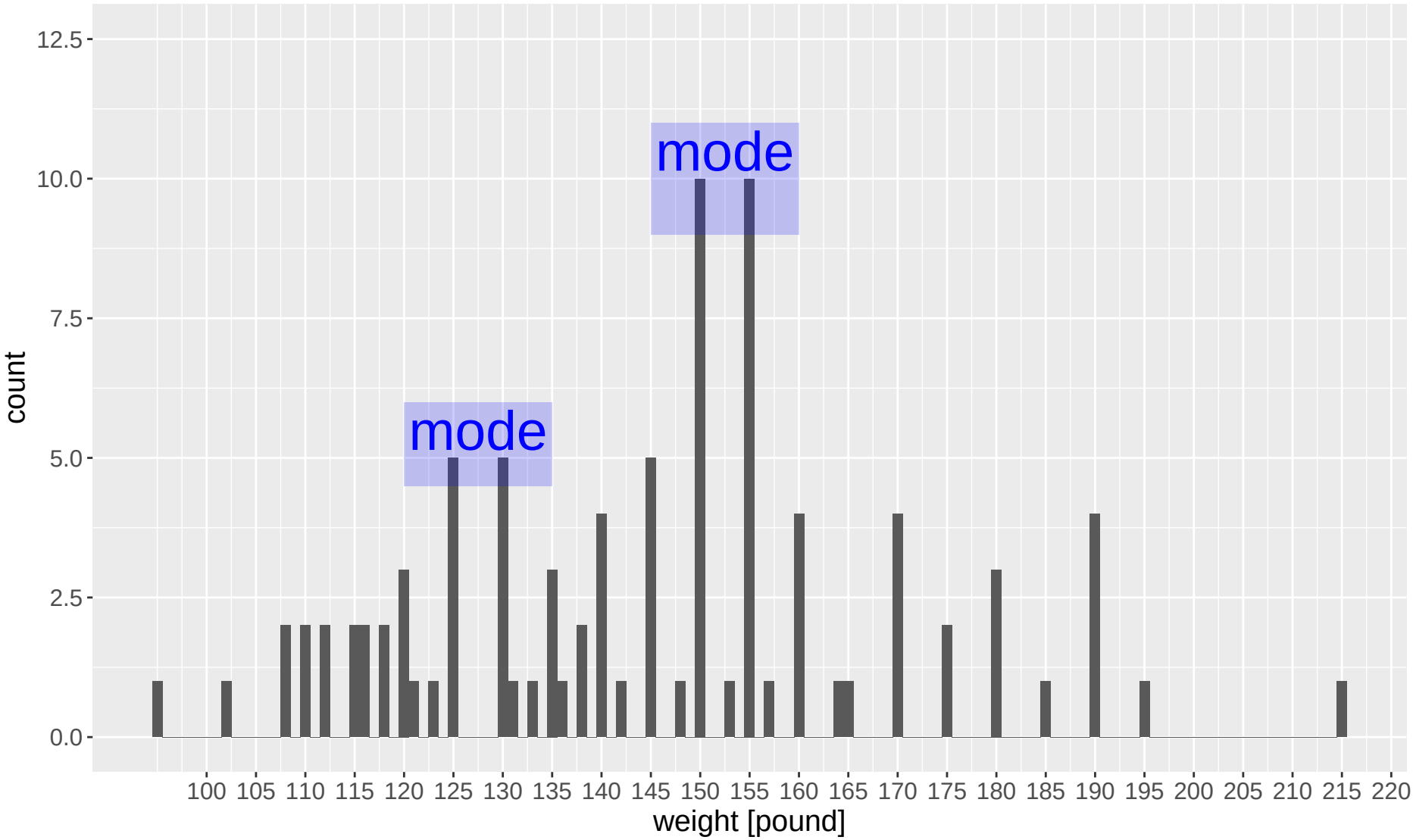
Reported Weights



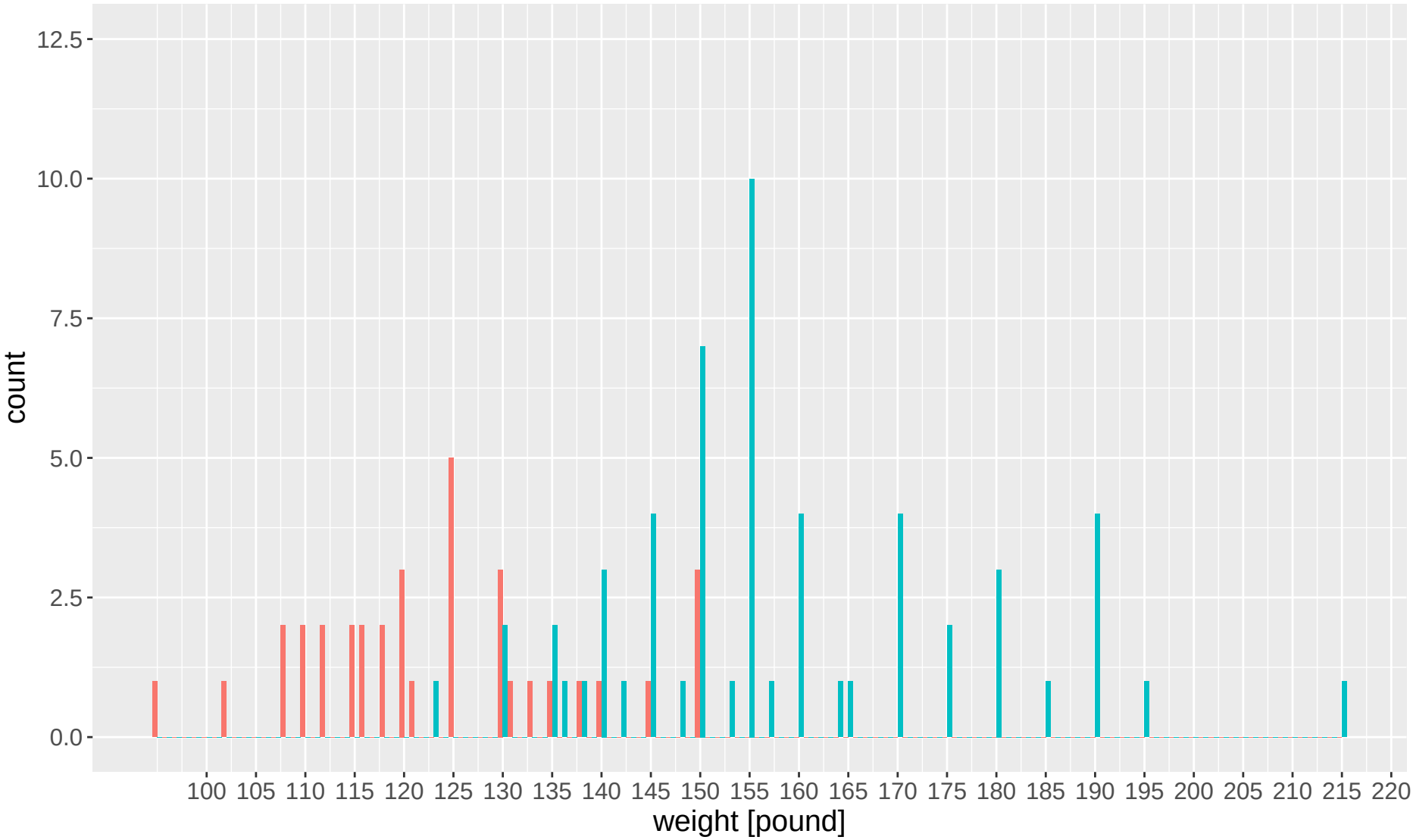
Reported Weights



Reported Weights



Reported Weights



Reported Weights

```
##
##   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 | 000000000035555555557
##   16 | 000045
##   17 | 000055
##   18 | 0005
##   19 | 00005
##   20 |
##   21 | 5
```

Section 5

Probability Mass/ Density Functions

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

Probability Mass/ Density Functions

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

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

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

$$P(a \leq x \leq b) = \int_a^b p(x)$$

Probability Mass/ Density Functions

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

$$p(x) \geq 0$$

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

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

Probability Mass/ Density Functions

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

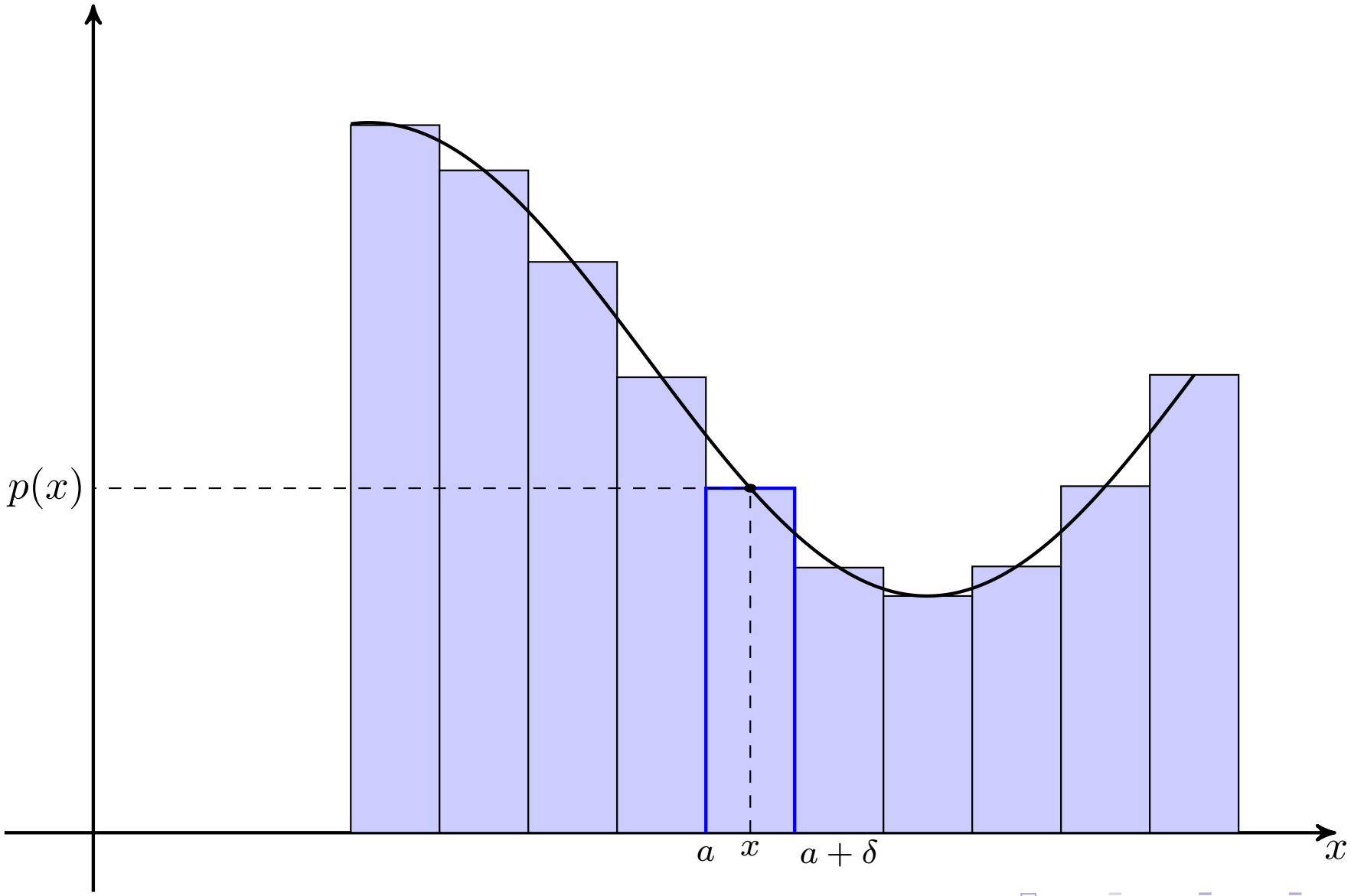
$$P\left(a \leq x \leq a+\delta\right) \approx p(a)\delta$$

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

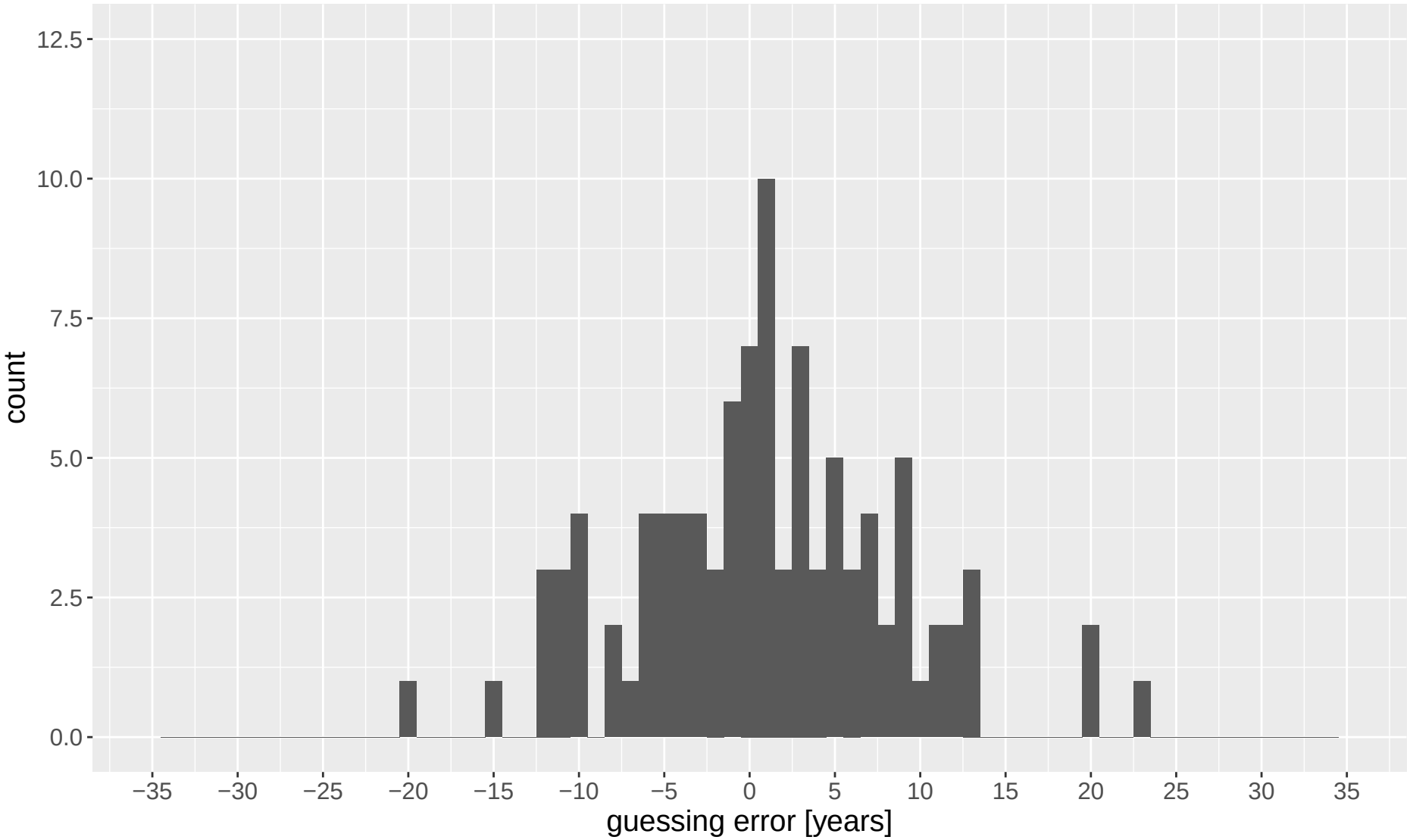
Probability Mass/ Density Functions



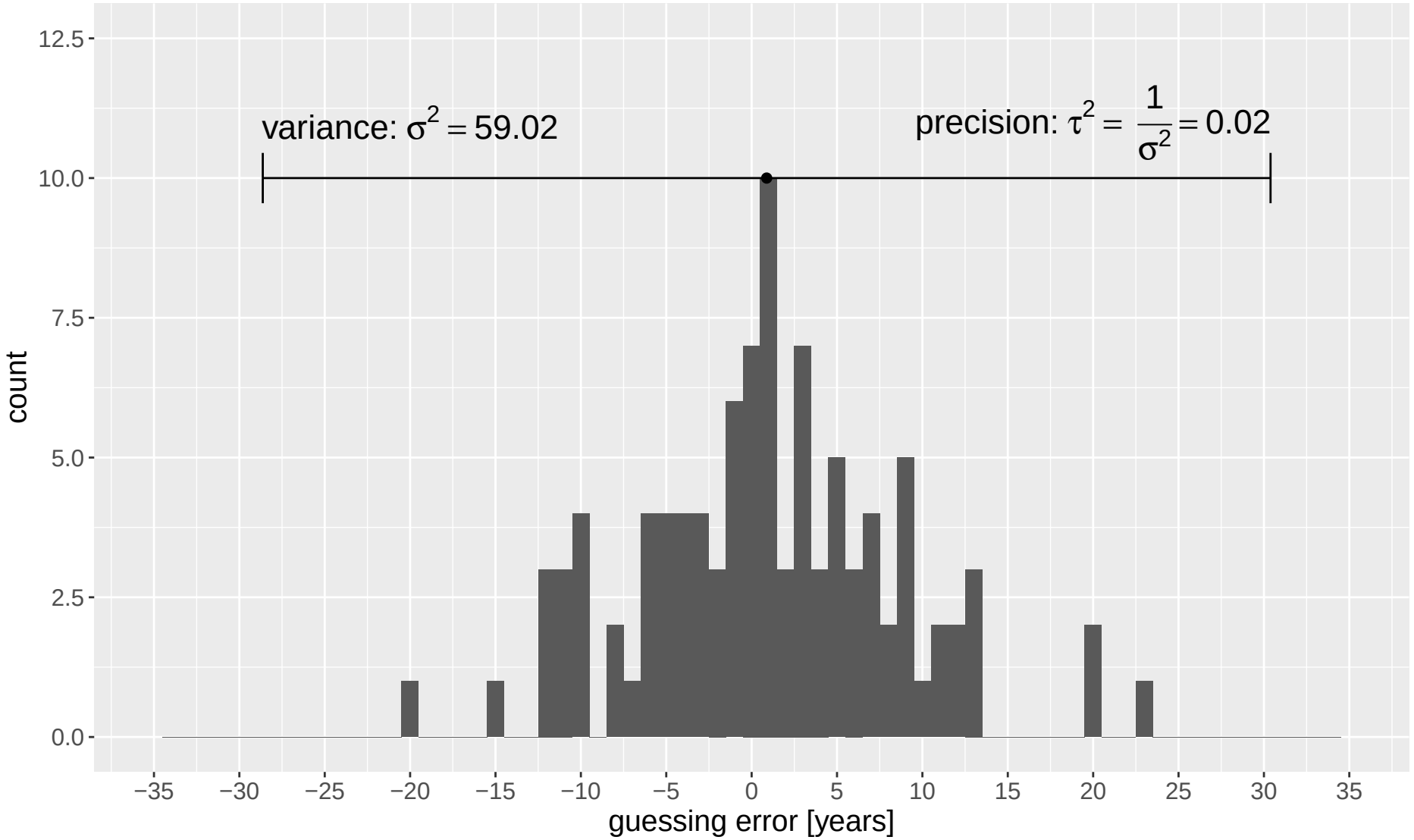
Section 6

Age guessing error

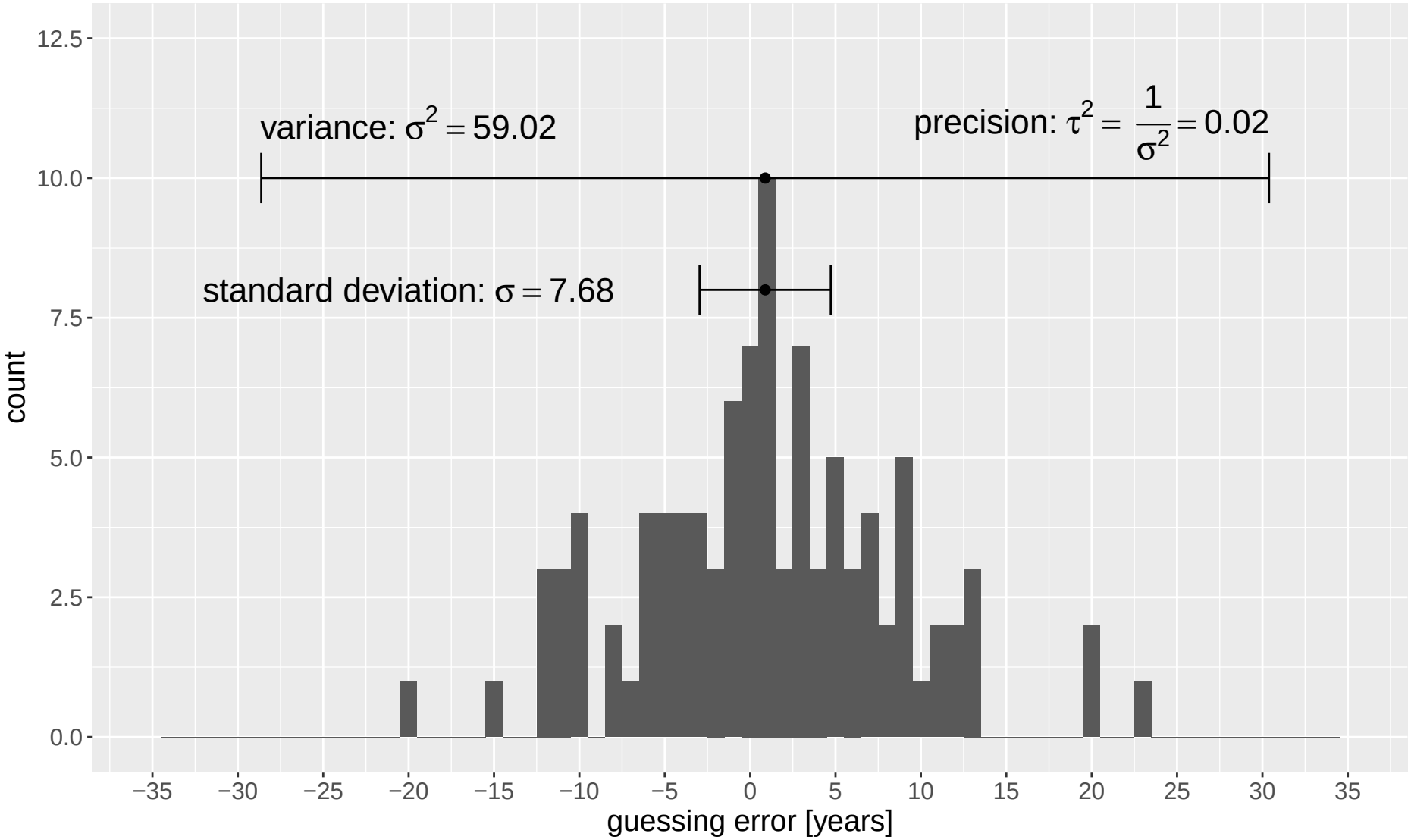
Age guessing error



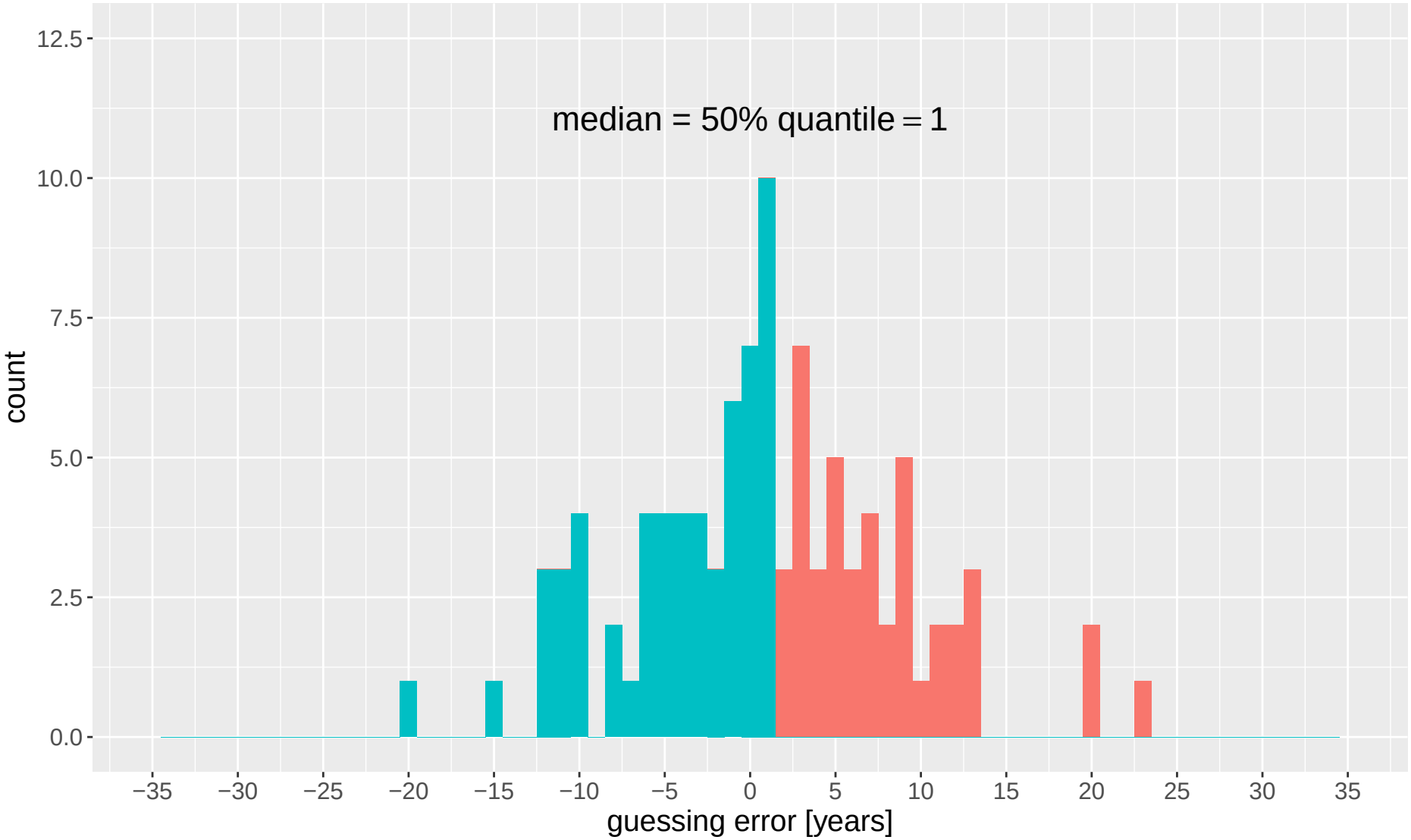
Age guessing error



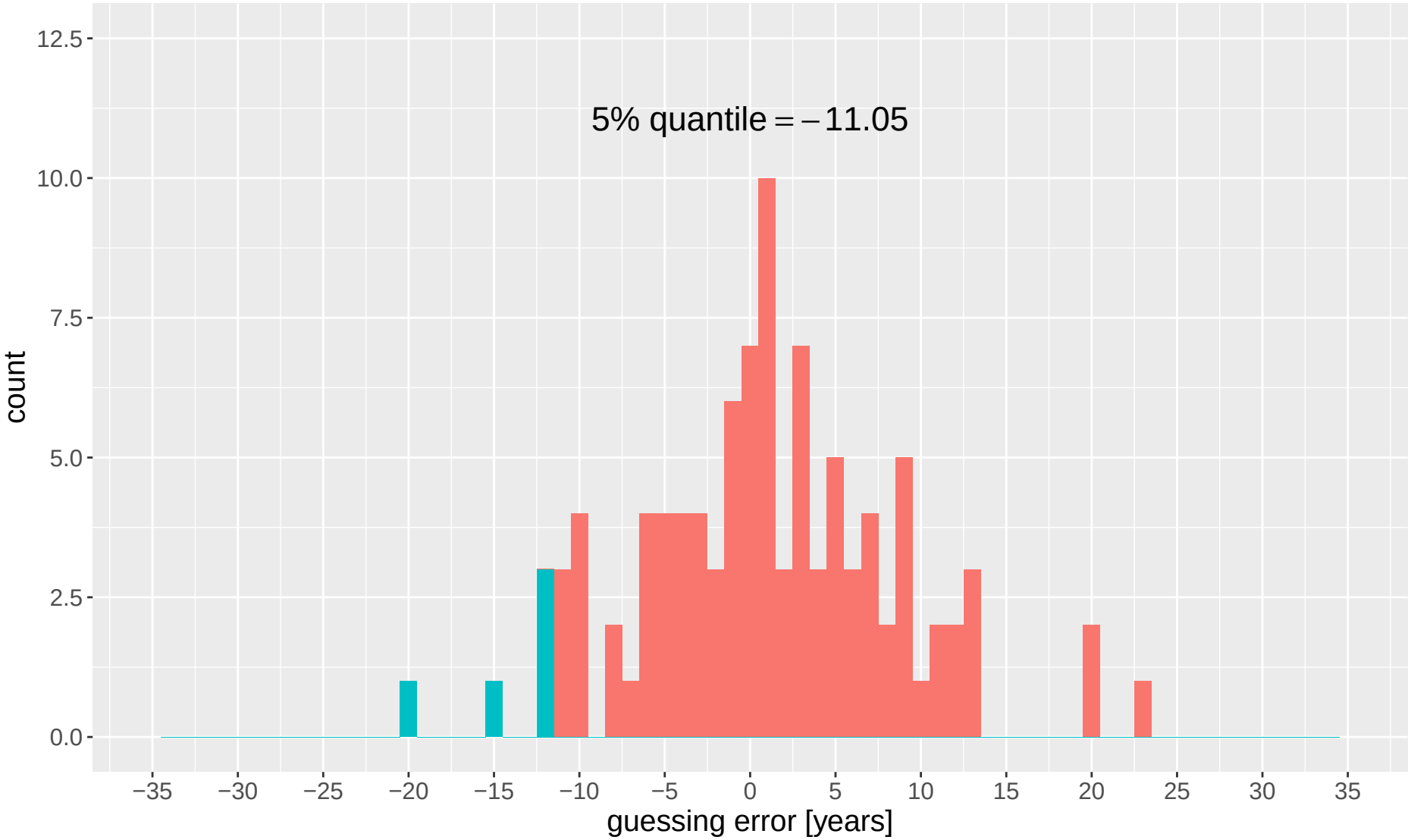
Age guessing error



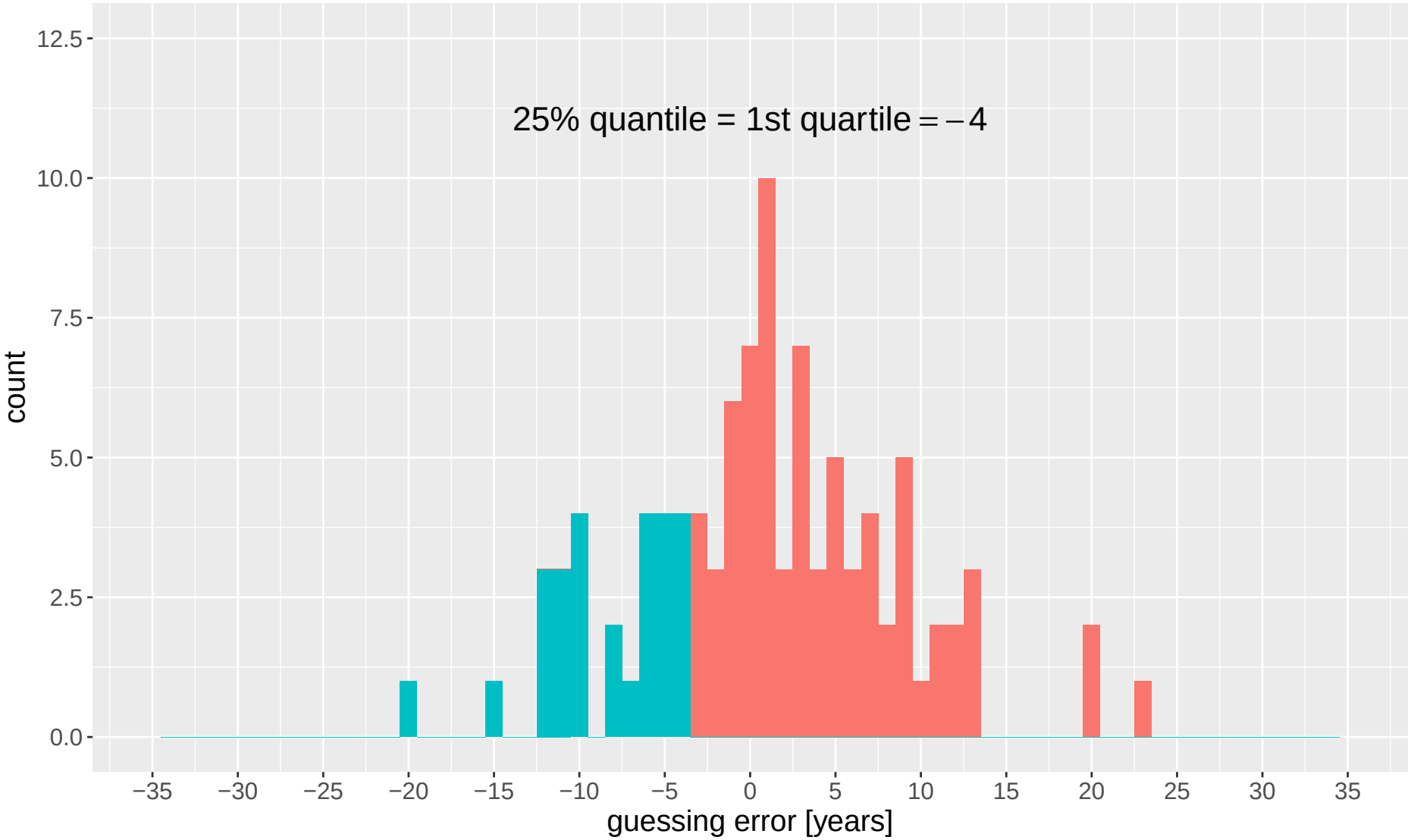
Age guessing error



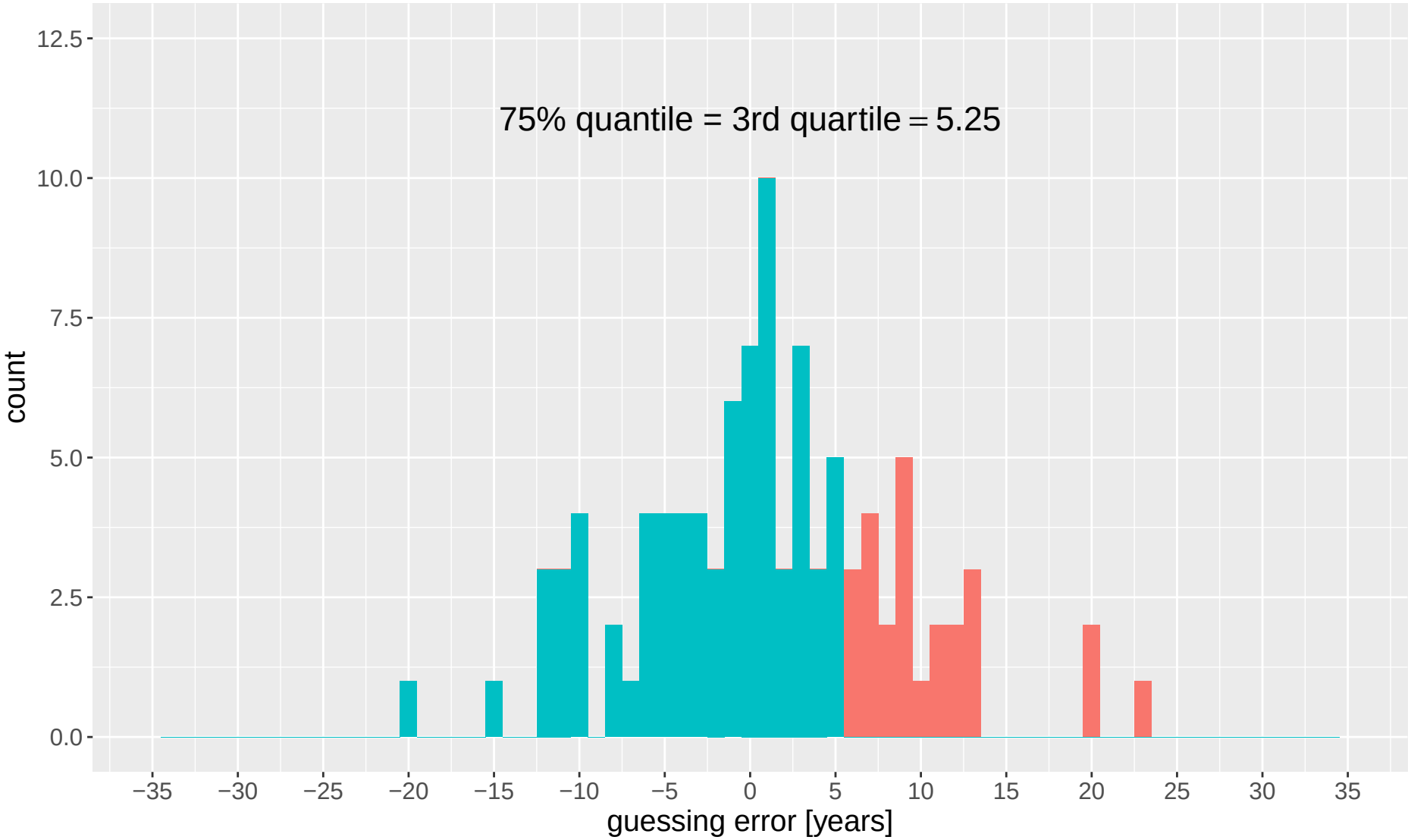
Age guessing error



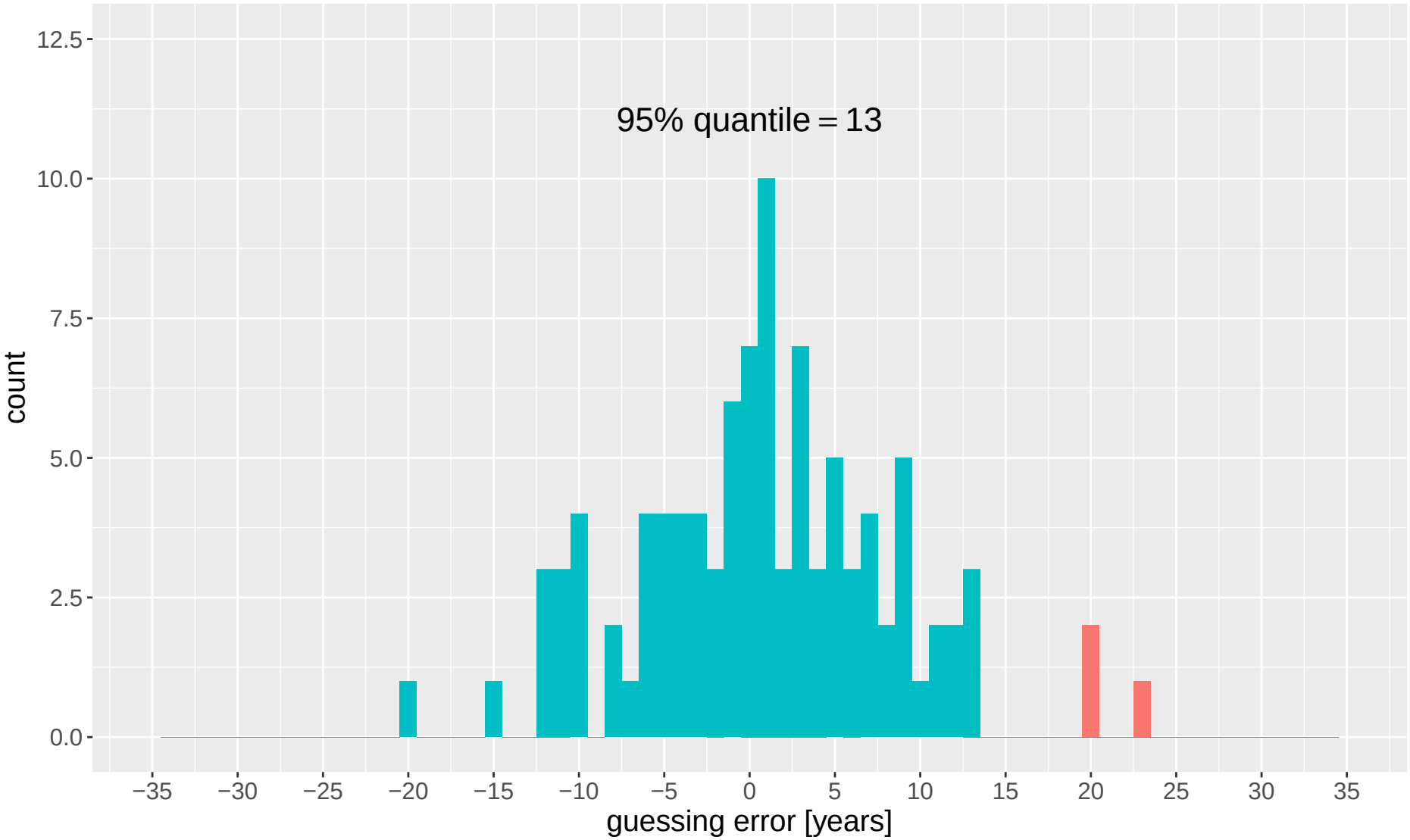
Age guessing error



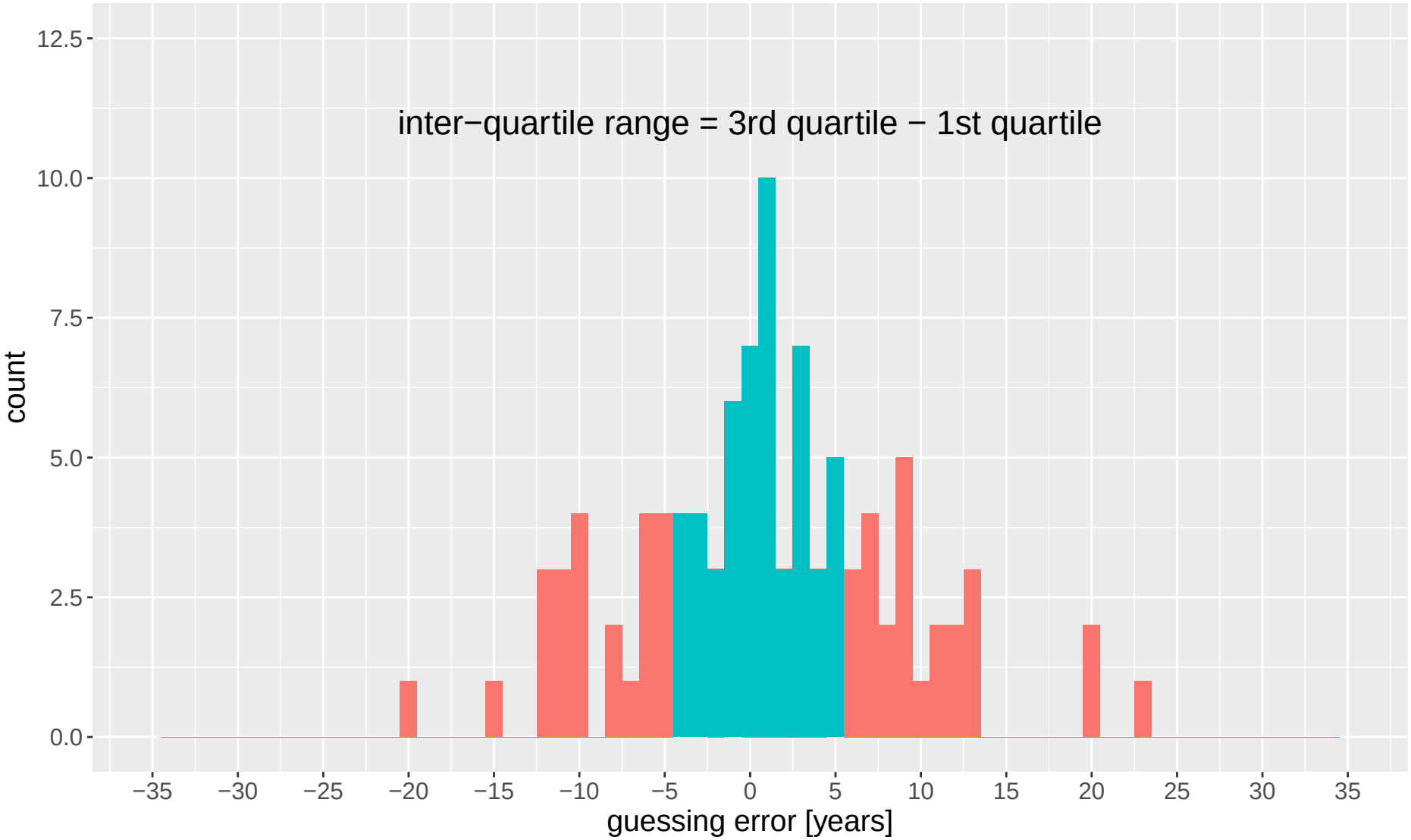
Age guessing error



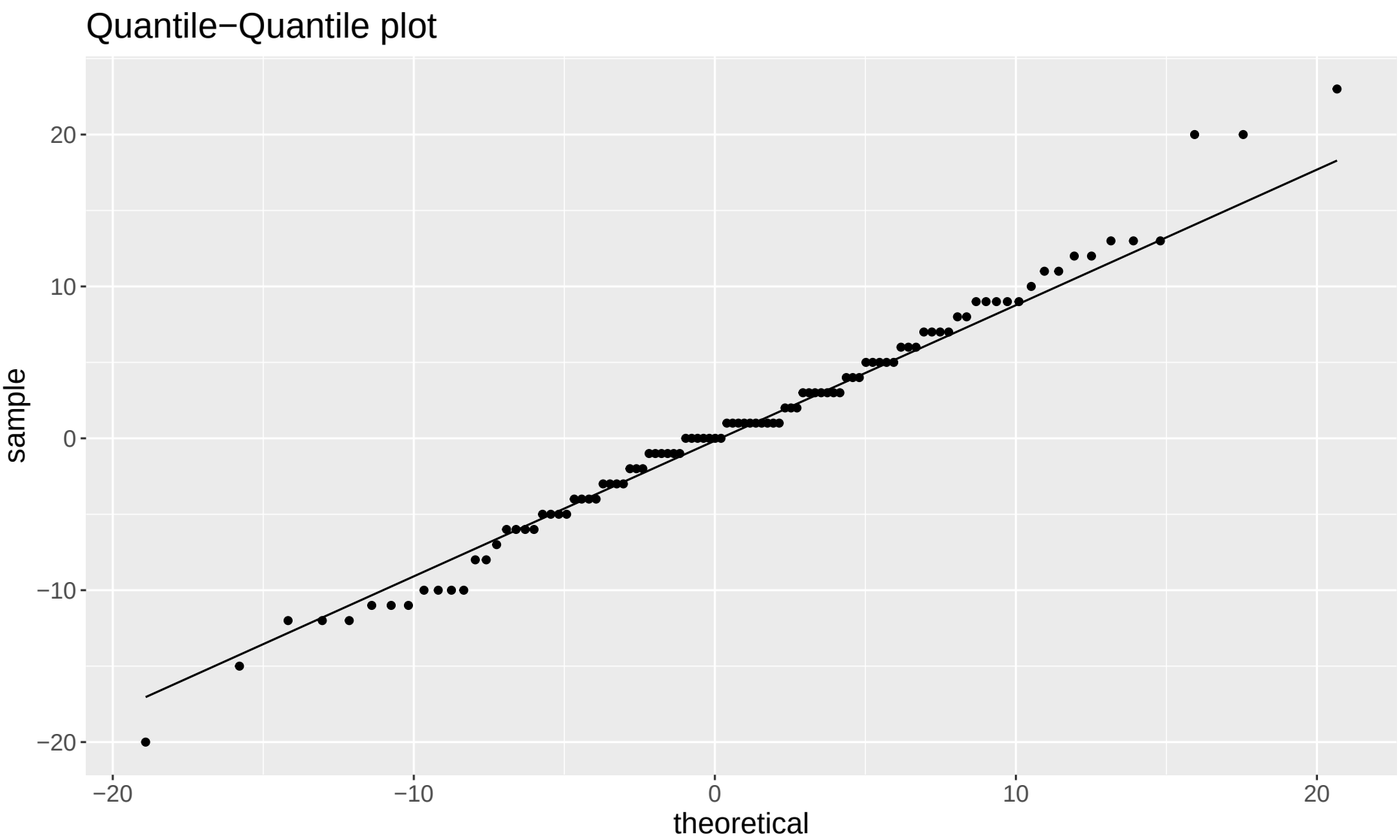
Age guessing error



Age guessing error



Age guessing error



Section 7

Summary

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: $P(x = a) = 0$

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