STAT UN1201 (002)

Prof. Joyce Robbins

January 19, 2017

Exercise

(based on #72, p. 49)

Data on a receptor binding measure:
PTSD: 10, 20, 25, 28, 31, 35, 37, 38, 38, 39, 39, 42, 46
Healthy: 23, 39, 40, 41, 43, 47, 51, 58, 63, 66, 67, 69, 72

PTSD:

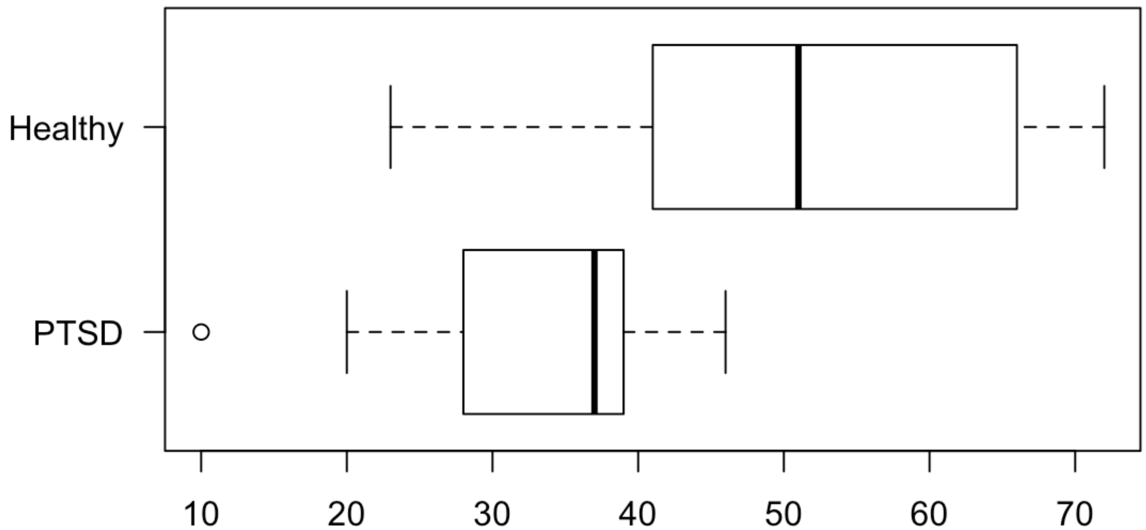
```
##     Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##     10.0    28.0    37.0    32.9    39.0    46.0
```

Healthy:

```
##     Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##     23.0    41.0    51.0    52.2    66.0    72.0
```

Draw a comparative boxplot.

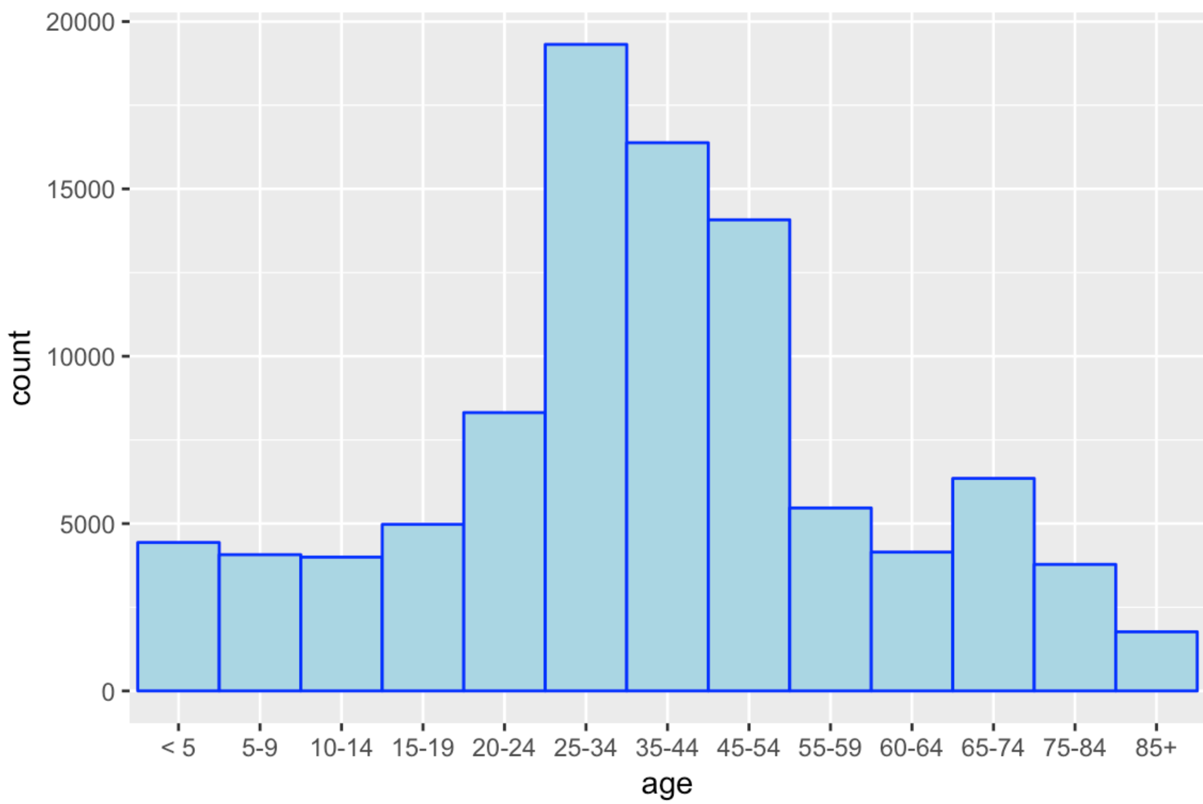
Solution

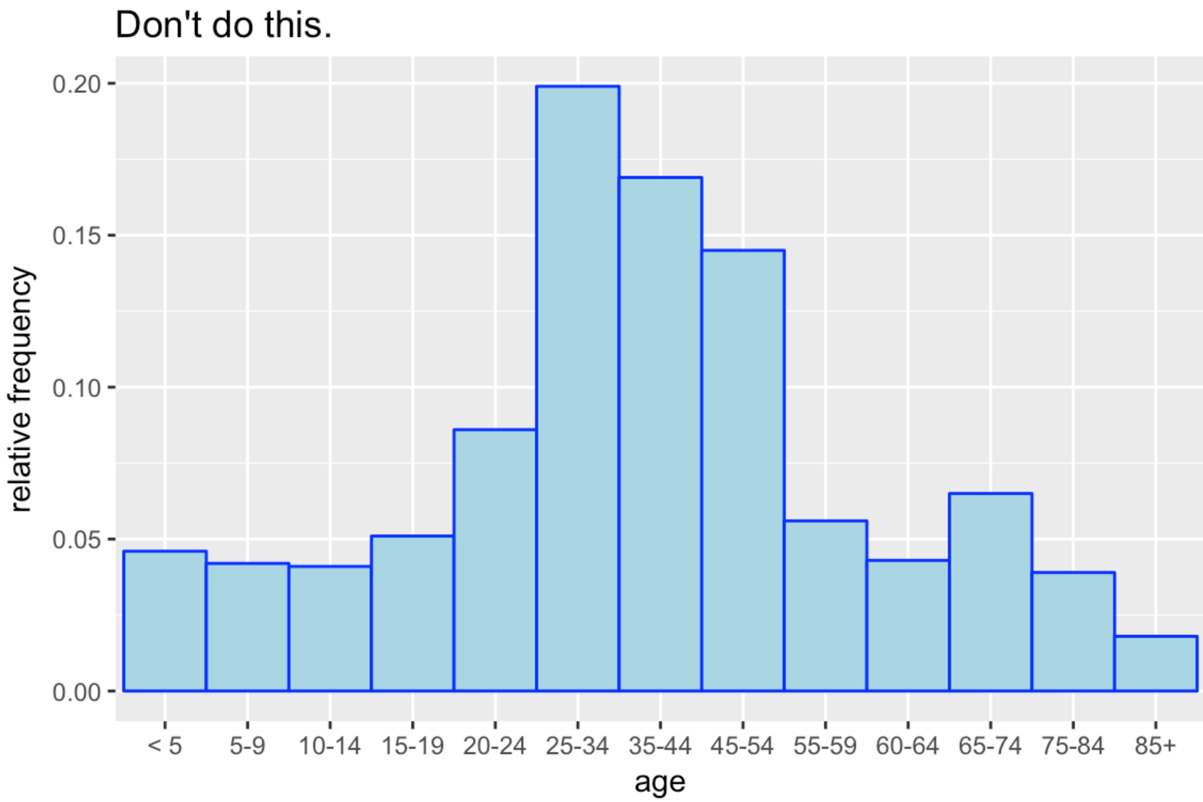


Admin Stuff

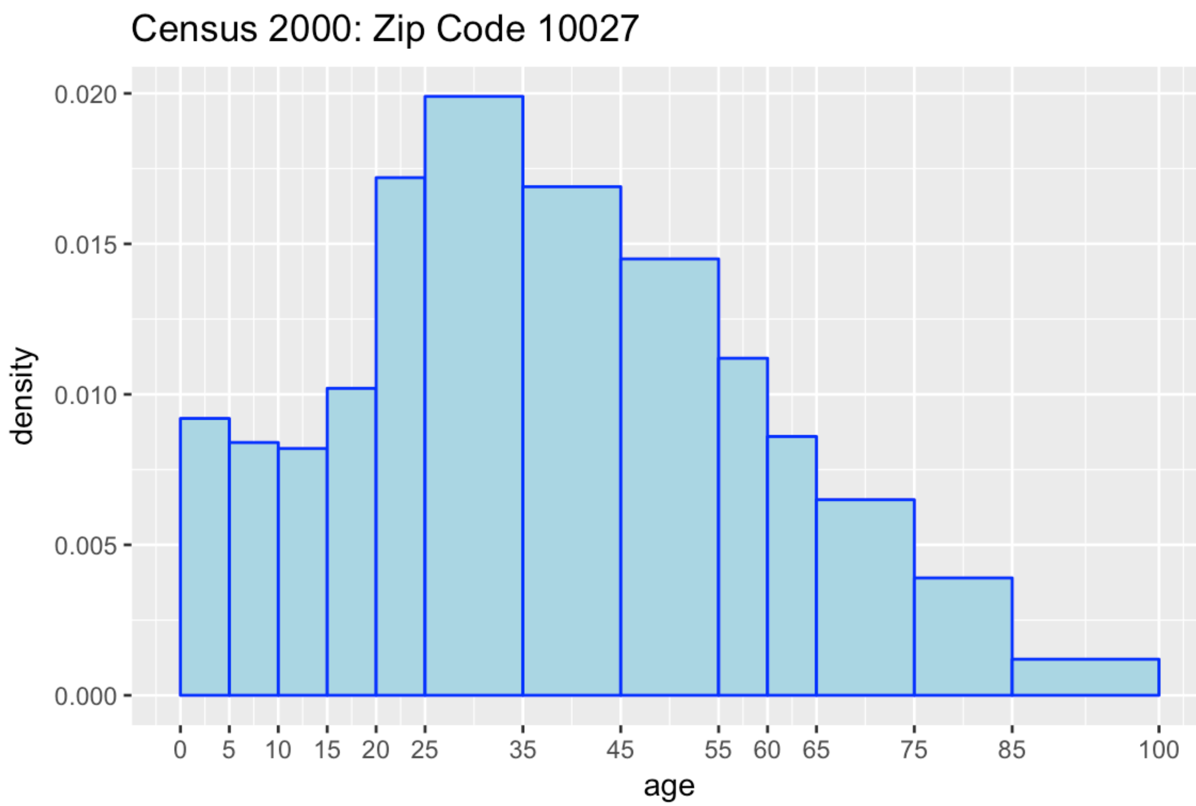
- Textbook
- Piazza
- Canvas app
- Homework / TurboScan
- Help Room
  http://stat.columbia.edu/help-room/ (TBA)

Histogram with Equal Class Widths



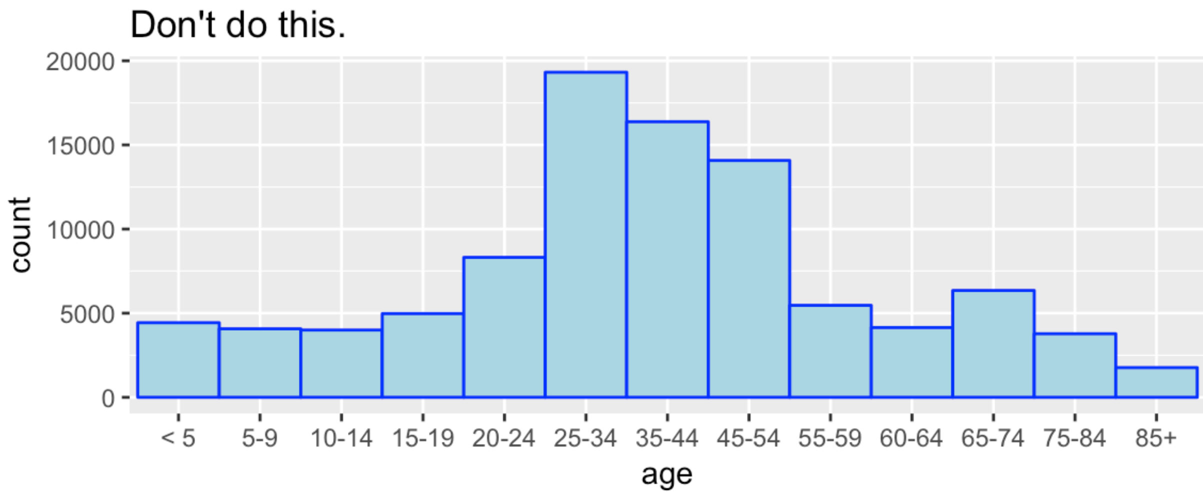

Don't do this.
relative frequency
0.20
0.15
0.10
0.05
0.00
< 5
5-9
10-14
15-19
20-24
25-34
35-44
45-54
55-59
60-64
65-74
75-84
85+
age

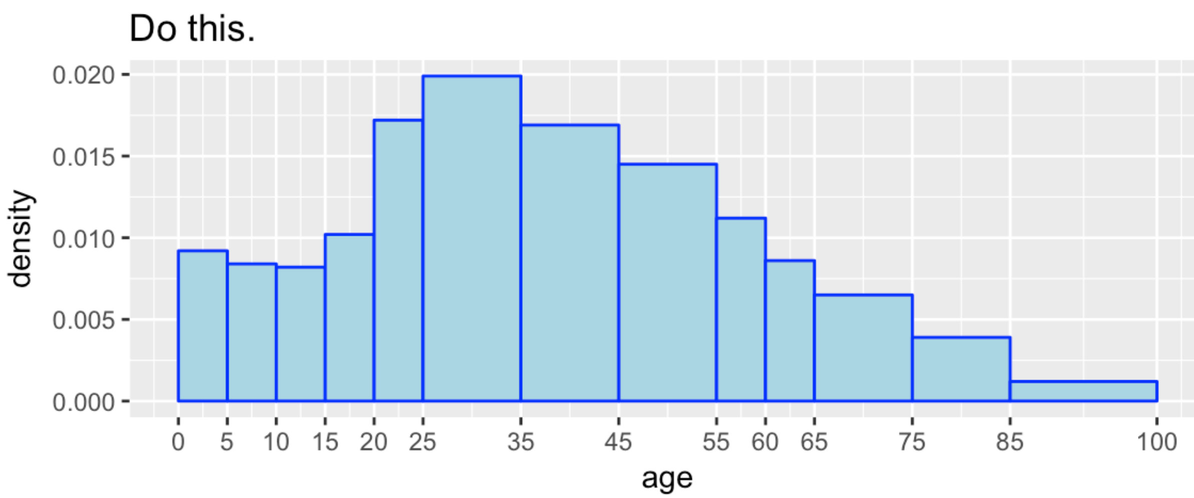
Relative Frequency Histogram with unequal bin (or class) widths



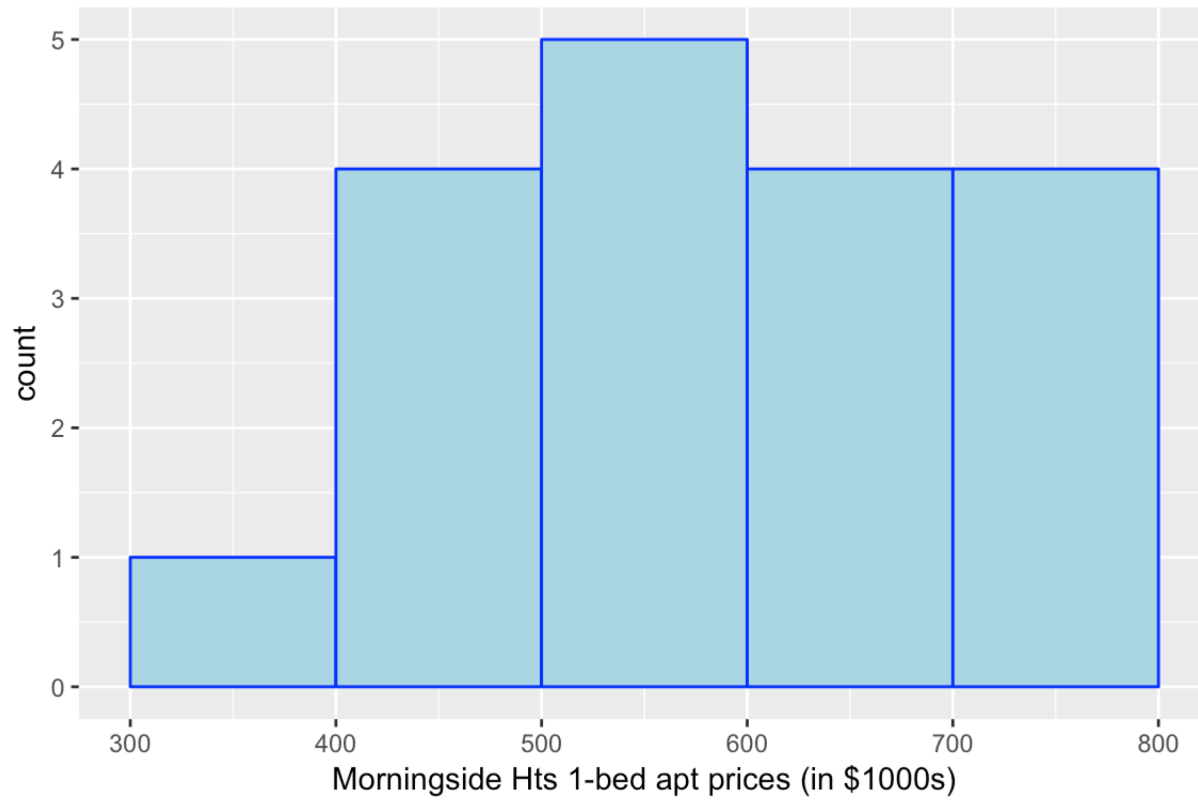
Creating a histogram with unequal class widths

Class	Frequency	RelFreq	ClassWidth	Density
< 5	4435	0.046	5	0.009
5-9	4072	0.042	5	0.008
10-14	3999	0.041	5	0.008
15-19	4977	0.051	5	0.010
20-24	8316	0.086	5	0.017
25-34	19317	0.199	10	0.020
35-44	16380	0.169	10	0.017
45-54	14077	0.145	10	0.014
55-59	5467	0.056	5	0.011
60-64	4148	0.043	5	0.009
65-74	6350	0.065	10	0.007
75-84	3781	0.039	10	0.004
85+	1767	0.018	15	0.001

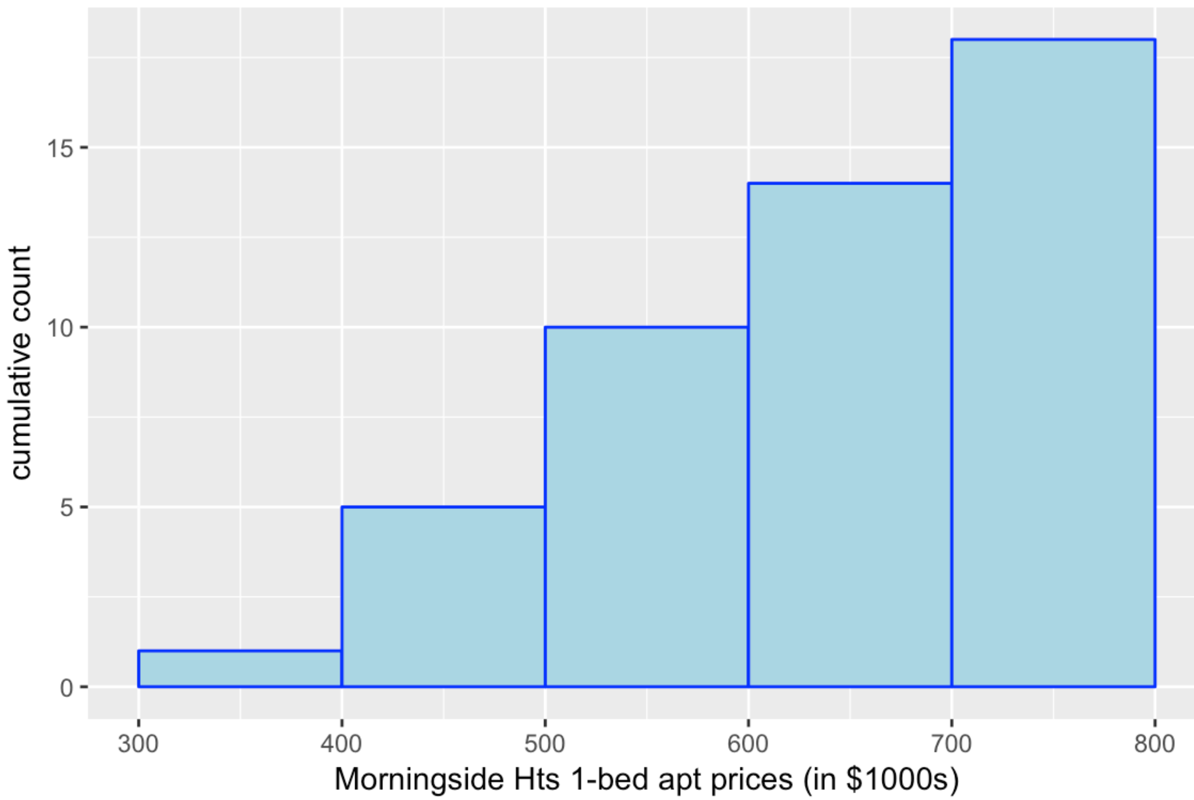

Don't do this.
20000
15000
10000
5000
0
count
< 5
5-9
10-14
15-19
20-24
25-34
35-44
45-54
55-59
60-64
65-74
75-84
85+
age


Do this.
0.020
0.015
0.010
0.005
0.000
density
0
5
10
15
20
25
35
45
55
60
65
75
85
100
age

Frequency Histogram



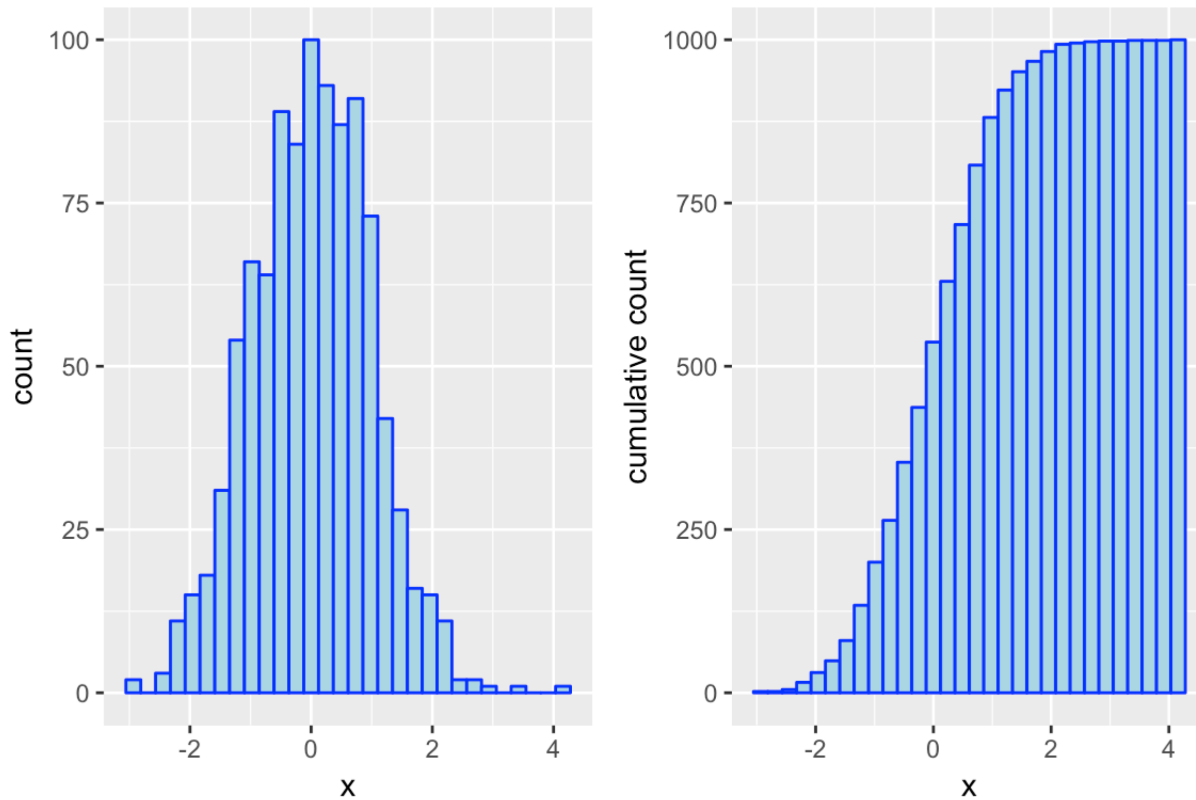
Cumulative Frequency Histogram



Drawing a Cumulative Frequency Histogram

Class	Freq	CumulativeFreq
300-400	1	1
400-500	4	5
500-600	5	10
600-700	4	14
700-800	4	18

Cumulative Frequency Histogram



Exercise 1

(based on #17, p. 26)
Construction industry data:

bidders	contracts
2	7
3	20
4	26
5	16
6	11
7	9
8	6
9	8
10	3
11	2

1. What proportion of the contracts involved at most five bidders?

2. What proportion of the contracts involved between five and ten bidders, inclusive?

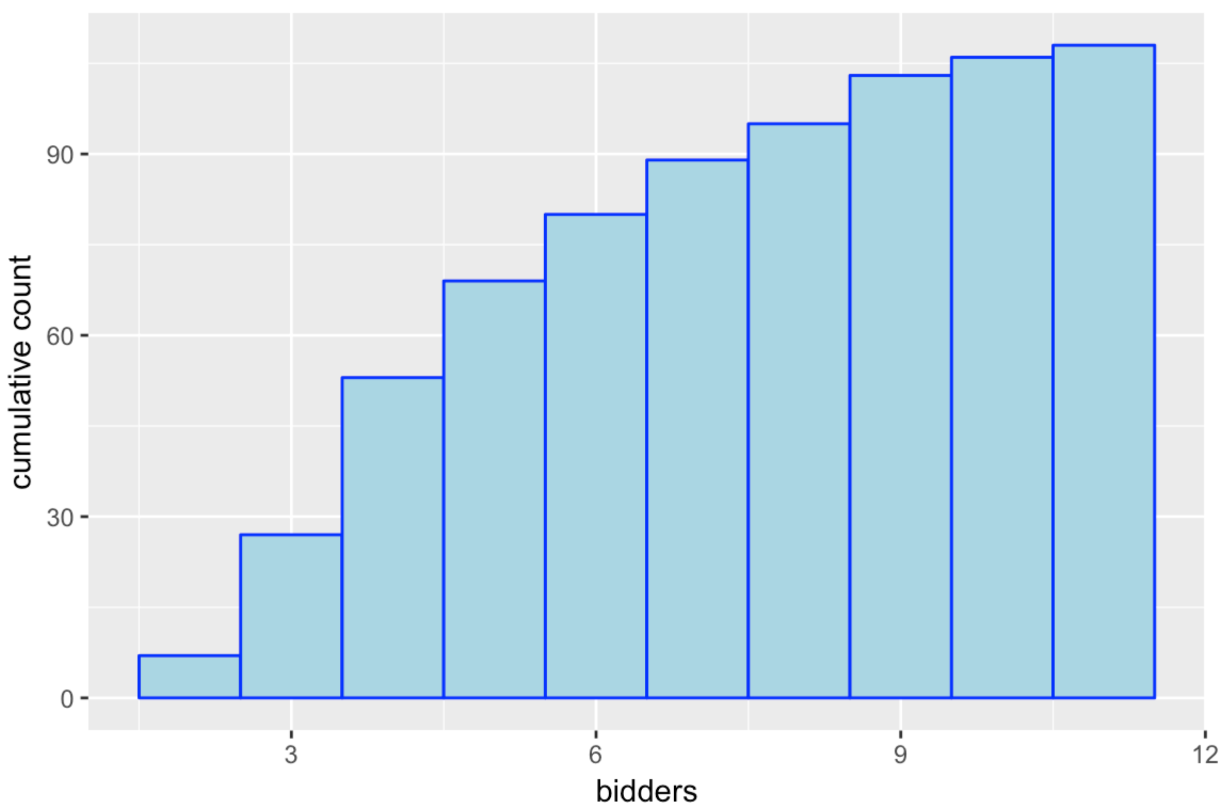
3. Draw a cumulative frequency histogram.

Solution

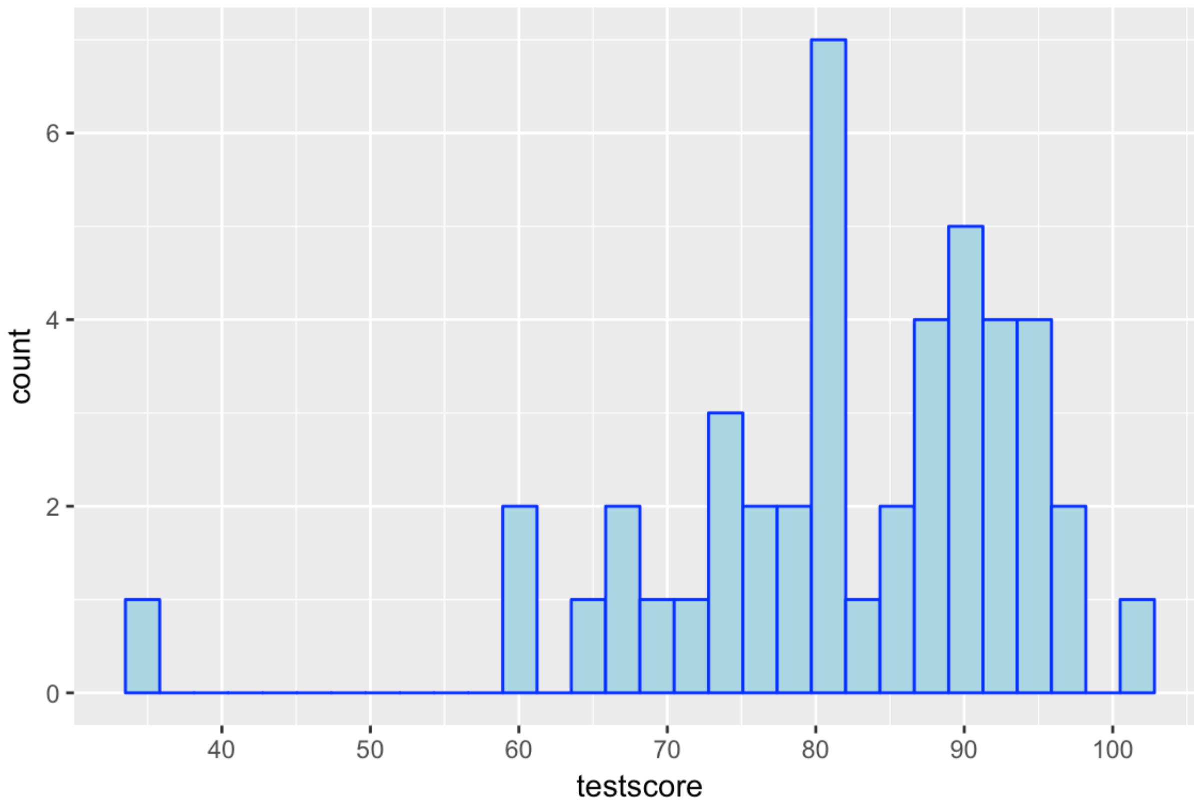
1. 0.639

2. 0.491

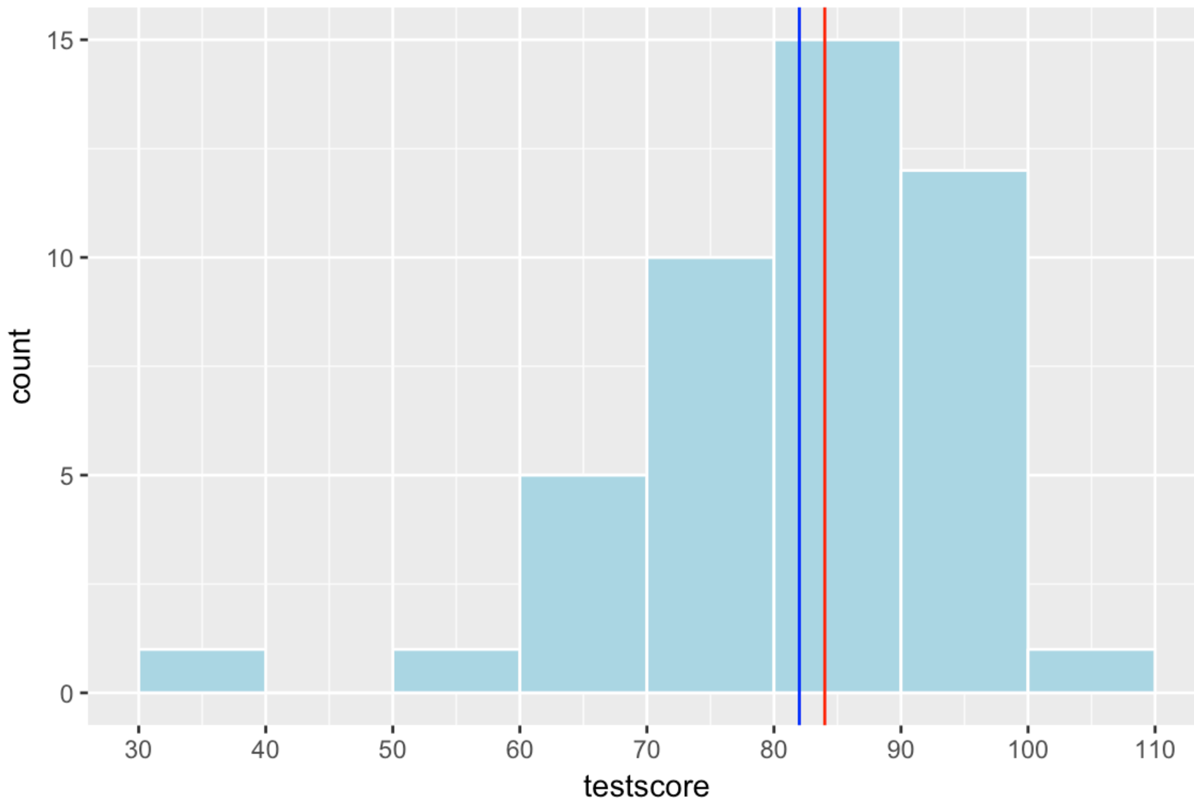
3.



Test Score Data



Fewer bins



Test score dataset

Original data set of scores:
35, 59, 61, 64, 66, 66, 70, 72, 73, 74, 75, 76, 76, 78, 79, 80, 80, 81, 81, 82, 82, 82, 84
86, 86, 88, 88, 88, 88, 89, 89, 90, 91, 91, 92, 92, 92, 92, 94, 94, 94, 94, 96, 98, 102

Mean: 82

Median: 84

Trimmed dataset (min and max removed):
59, 61, 64, 66, 66, 70, 72, 73, 74, 75, 76, 76, 78, 79, 80, 80, 81, 81, 82, 82, 82, 84
86, 86, 88, 88, 88, 88, 89, 89, 90, 91, 91, 92, 92, 92, 92, 94, 94, 94, 94, 96, 98

Mean: 82.63

Median: 84

How much was trimmed? $\frac{1}{45}$ = 2.22%

Trimmed means

Suppose we want to trim 10%.

.1 * 45 = 4.5 values

Trim 4:

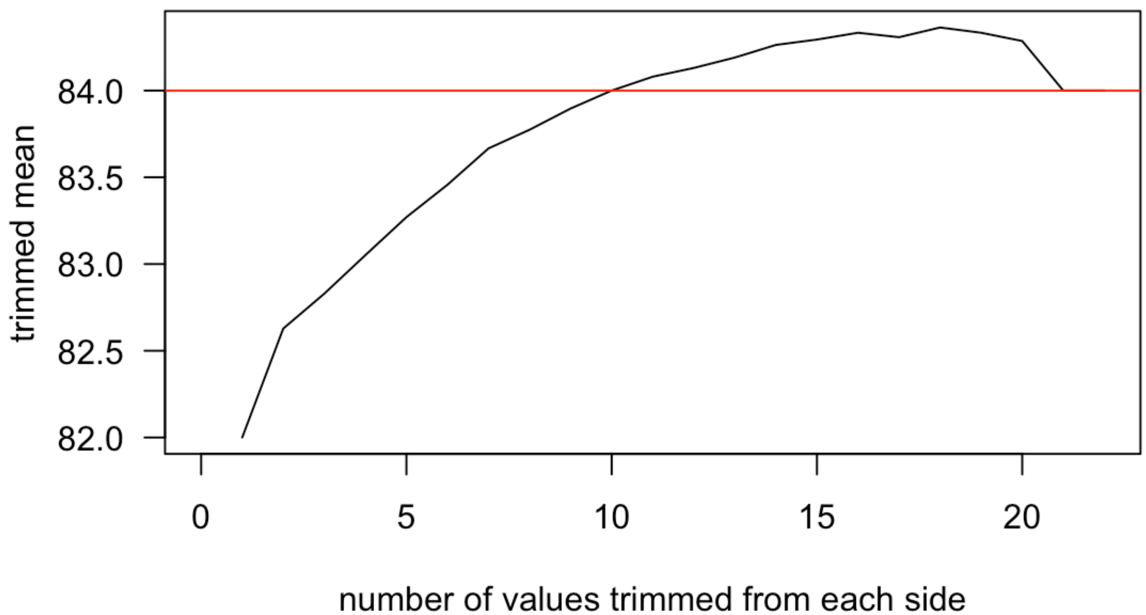
$\frac{4}{45}$ = 8.89%

$\bar{x}_{tr(8.89)}$ = 83.27

$\frac{5}{45}$ = 11.11%

$\bar{x}_{tr(11.11)}$ = 83.457

Median vs. Trimmed Mean



Sample and Population Means

population mean: μ = sum of N population values / N

sample mean: $\bar{x} = \frac{x_1+x_2+\ldots+x_n}{n} = \frac{\sum_{i=1}^{n} x_i}{n}$

population median: $\tilde{\mu}$

sample median: $\tilde{x}$

Measures of Variability

deviations from the mean

$x_1 - \bar{x},\ x_2 - \bar{x},\ etc.$

Data: 3, 8, 11, 14
Mean: 9

value	*deviation*	*deviation*^{2}
3	-6	36
8	-1	1
11	2	4
14	5	25

Sum of squared deviations S_{xx}: 36 + 1 + 4 + 25 = 66

Population variance σ^2 = 66/4 = 16.5

$\sigma^2 = \sum_{i=1}^{N} (x_i - \mu)^2 / N$

Sample Variance

Sum of squared deviations S_{xx}: 36 + 1 + 4 + 25 = 66

Sample variance: s^2 = 66 / **3** = 22

$$s^2 = \frac{\sum_{i=1}^{n}(x_i - \bar{x})^2}{n-1}$$

Why n - 1?

- We don’t have the true population mean, our estimate would be too high if we divided by n instead of n-1

Standard Deviation

Square root of variance

Population s.d. = $\sqrt{\sigma^2}$

Sample s.d. = $\sqrt{s^2}$

- same units as original values

Variance of test scores: 156.636

Standard deviation of test scores: 12.515

Exercise 2

(p. 35, #38)

Blood pressure values are often reported to the nearest 5 mmHg (100, 105, 110, etc.). Suppose the actual blood pressure values for nine randomly selected individuals are:

118.6 127.4 138.4 130.0 113.7 122.0 108.3 131.5 133.2

1. What is the median of the *reported* blood pressure values?
2. Suppose the blood pressure of the second individual is 127.6 rather than 127.4 (a small change in a single value). How does this affect the median of the reported values?

Solution

1. 125
2. 130