STAT UN1201 – Chapter 1

Prof. Joyce Robbins

Waitlist

1. The waitlist moves in order as places open up.

2. Course materials are available here during change of program period: http://github.com/jtr13/1201

3. It is strongly advised to keep up with the material if you are trying to get in the class.

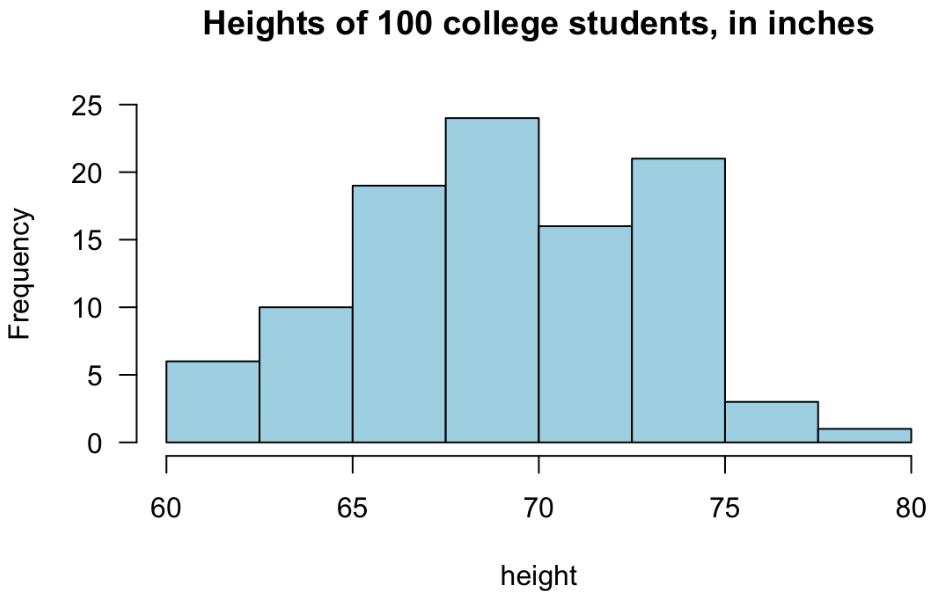
EVERYONE: Once you’ve made a decision not to take the class, please be considerate and drop it from your schedule.

Discrete data

```
## [1] "Heights of 100 college students, in inches"
```

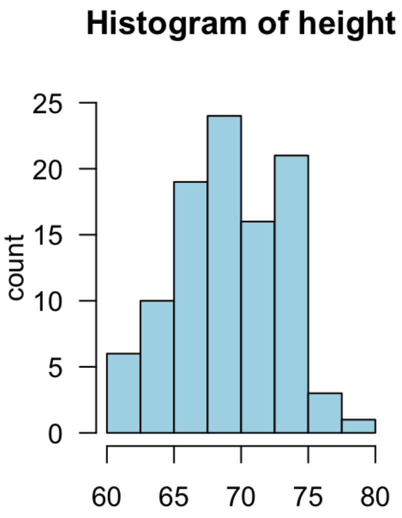
```
##   [1] 60 60 61 61 61 62 63 63 64 64 64 64 65 65 65
##  [16] 65 66 66 66 66 66 66 67 67 67 67 67 67 67 67
##  [31] 67 67 67 67 67 68 68 68 68 68 68 68 69 69 69
##  [46] 69 69 69 69 69 69 70 70 70 70 70 70 70 70 71
##  [61] 71 71 71 71 71 72 72 72 72 72 72 72 72 72 72
##  [76] 73 73 73 74 74 74 74 74 74 74 74 74 74 74 74
##  [91] 74 75 75 75 75 75 76 76 77 79
```

Discrete data

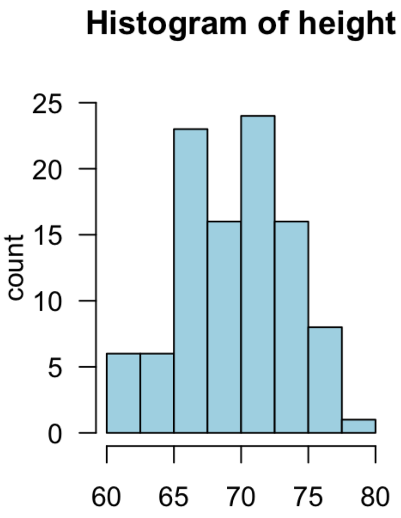


```
##   [1] 60 60 61 61 61 62 63 63 64 64 64 64 65 65 65
##  [16] 65 66 66 66 66 66 66 67 67 67 67 67 67 67 67
##  [31] 67 67 67 67 67 68 68 68 68 68 68 68 69 69 69
##  [46] 69 69 69 69 69 69 70 70 70 70 70 70 70 70 71
##  [61] 71 71 71 71 71 72 72 72 72 72 72 72 72 72 72
##  [76] 73 73 73 74 74 74 74 74 74 74 74 74 74 74 74
##  [91] 74 75 75 75 75 75 76 76 77 79
```

Discrete data histogram



RIGHT CLOSED, LEFT OPEN



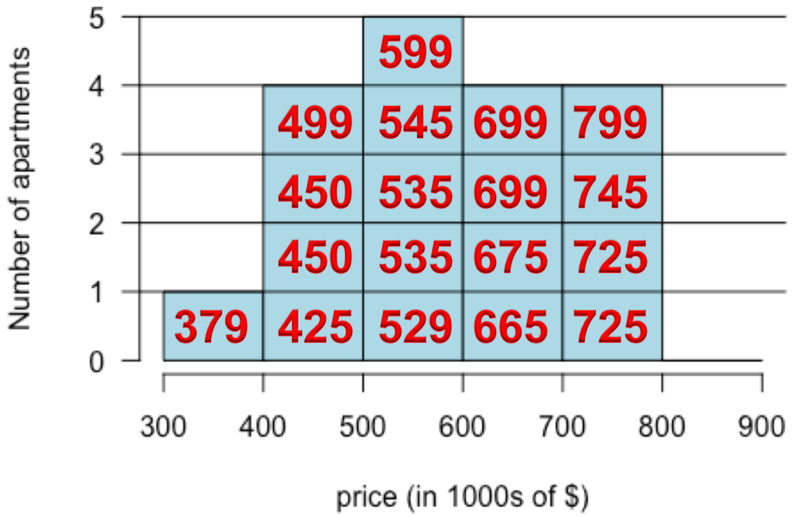
RIGHT OPEN, LEFT CLOSED

```
##   [1] 60 60 61 61 61 62 63 63 64 64 64 64 65 65 65
##  [16] 65 66 66 66 66 66 66 67 67 67 67 67 67 67 67
##  [31] 67 67 67 67 67 68 68 68 68 68 68 68 69 69 69
##  [46] 69 69 69 69 69 69 70 70 70 70 70 70 70 70 71
##  [61] 71 71 71 71 71 72 72 72 72 72 72 72 72 72 72
##  [76] 73 73 73 74 74 74 74 74 74 74 74 74 74 74 74
##  [91] 74 75 75 75 75 75 76 76 77 79
```

EXERCISE

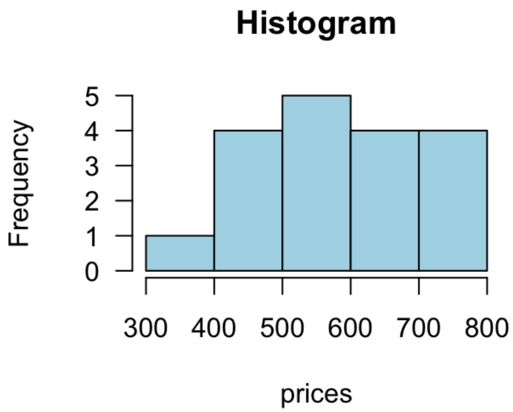
Draw a histogram of the asking prices for one-bedroom apartments in Morningside Heights (prices in thousands of $)
Data source: cityrealty.com, 9/13/2016

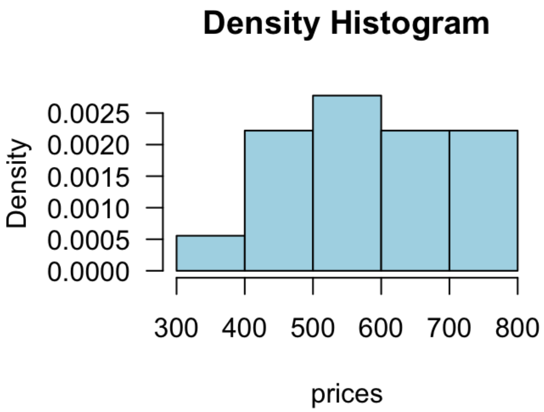
379, 425, 450, 450, 499, 529, 535, 535, 545,
599, 665, 675, 699, 699, 725, 725, 745, 799



Histogram of Morningside Heights One-Bedroom Apt. Prices

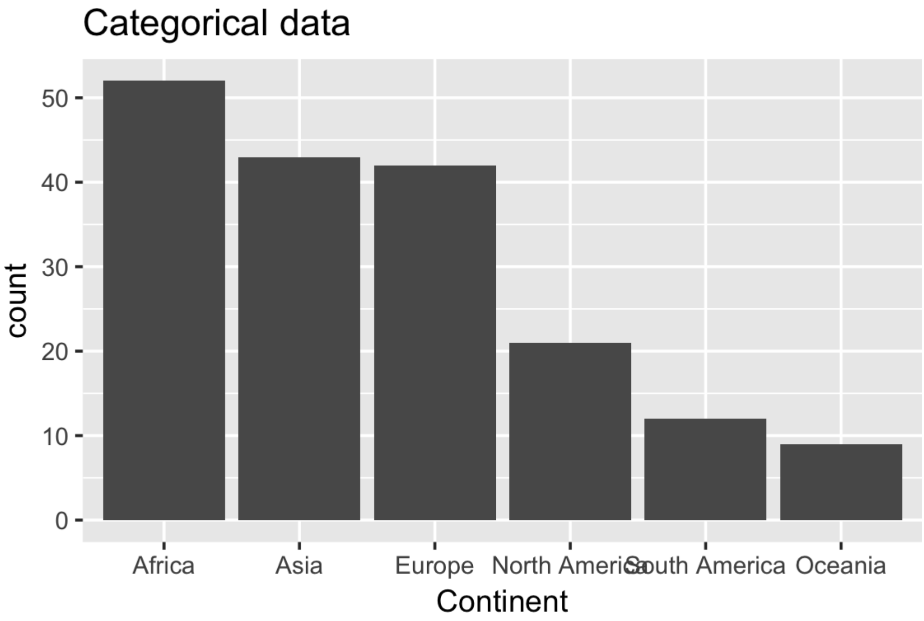
Density histogram



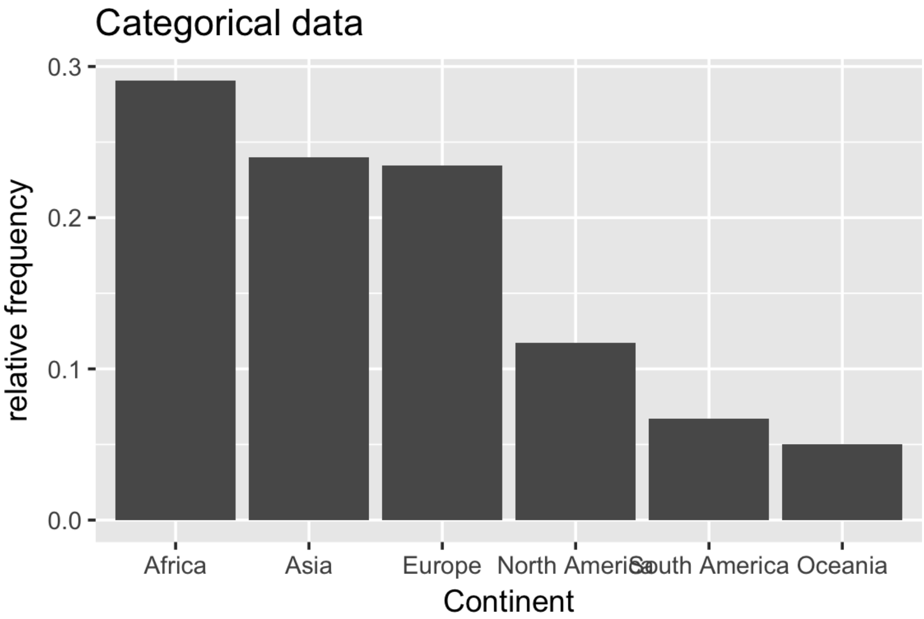


Class	Frequency	Rel. Frequency	Density
(300, 400]	1	.056	.00056
(400, 500]	4	.222	.00222
(500, 600]	5	.278	.00278
(600, 700]	4	.222	.00222
(700, 800]	4	.222	.00222

Frequency bar chart



Relative frequency bar chart



Five number summary

1. min
2. lower fourth
3. median
4. upper fourth
5. max

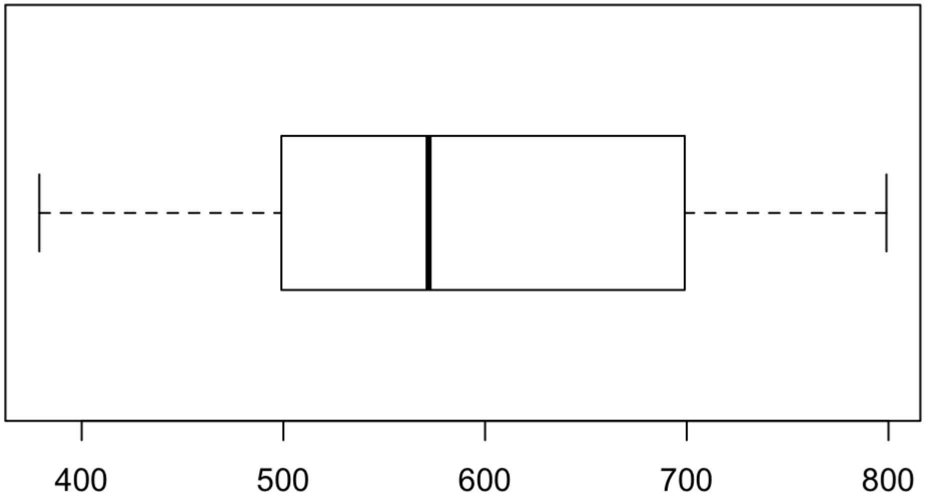
```
fivenum(prices)
```

```
## [1] 379 499 572 699 799
```

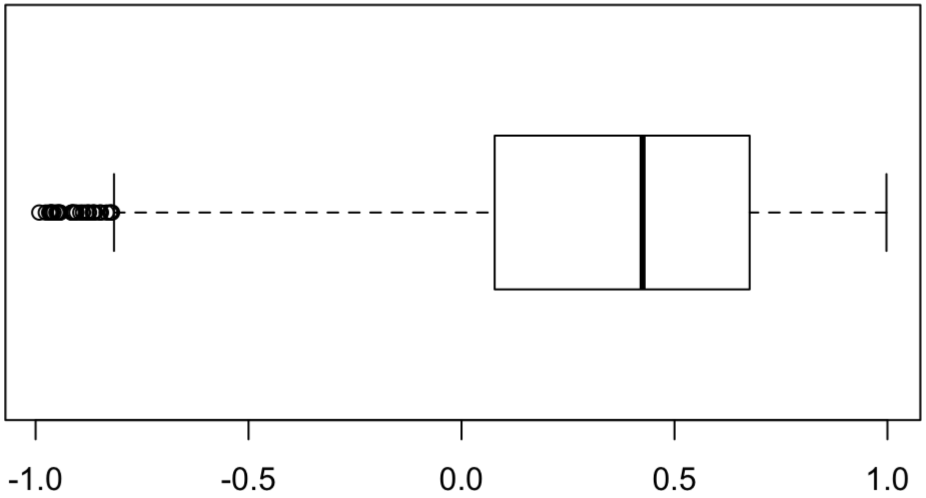
Boxplot

379, 425, 450, 450, 499, 529, 535, 535, 545,
599, 665, 675, 699, 699, 725, 725, 745, 799

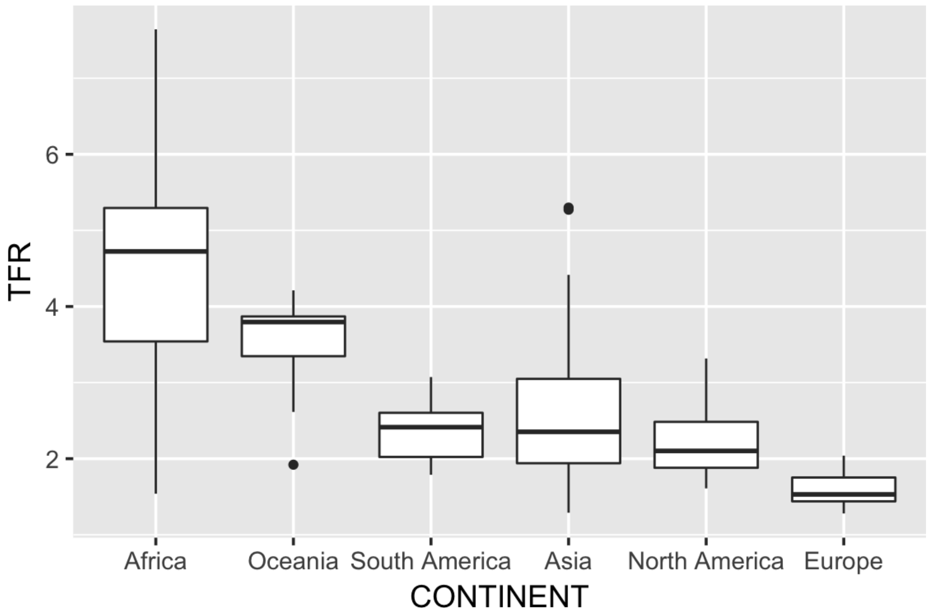
```
## [1] 379 499 572 699 799
```



Boxplot with outliers



Multiple box plots



EXERCISE

(based on #72, p. 49)

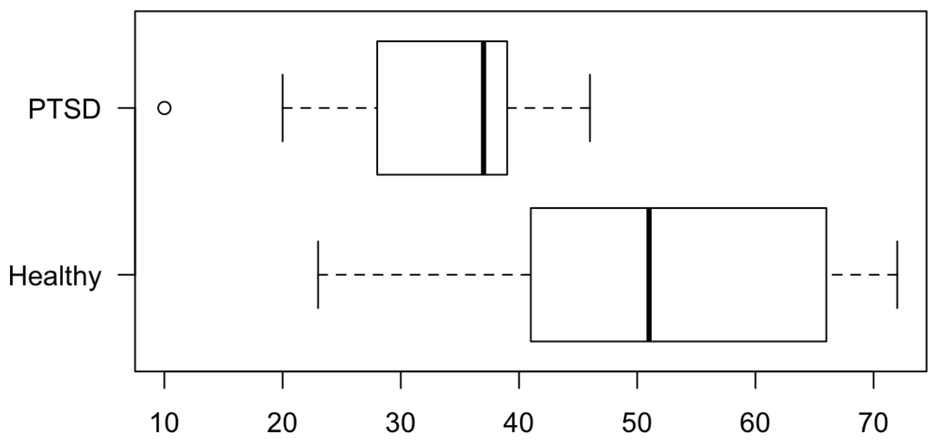
Data on a receptor binding measure:

PTSD: 10, 20, 25, 28, 31, 35, 37, 38, 38, 39, 39, 42, 46

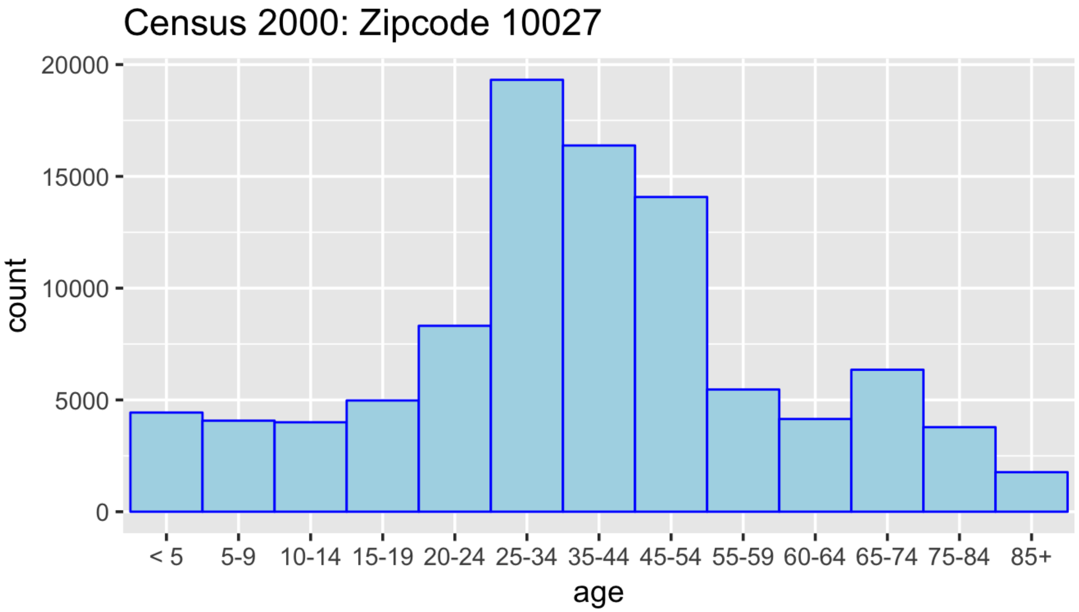
Healthy: 23, 39, 40, 41, 43, 47, 51, 58, 63, 66, 67, 69, 72

Draw a comparative boxplot.

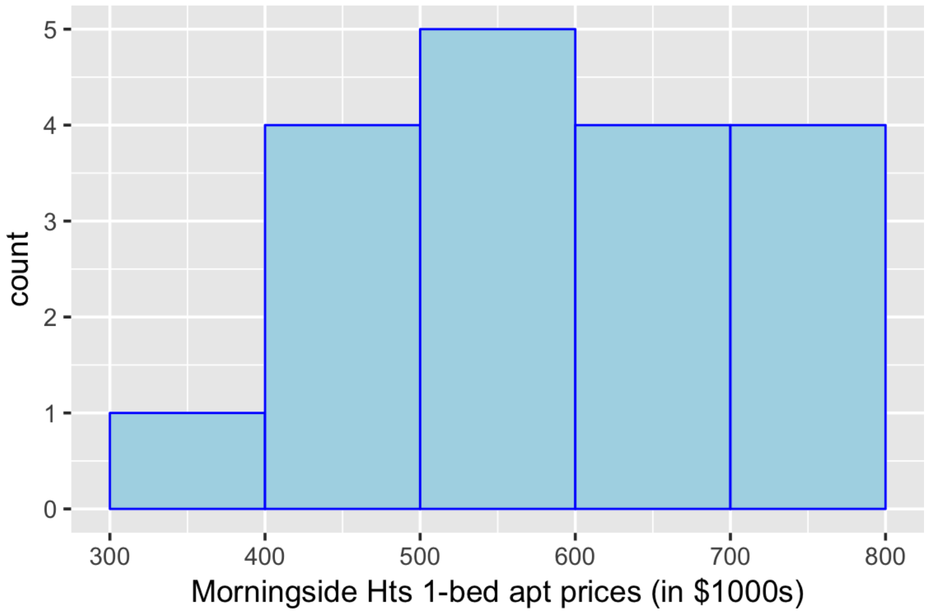
Comparative boxplot



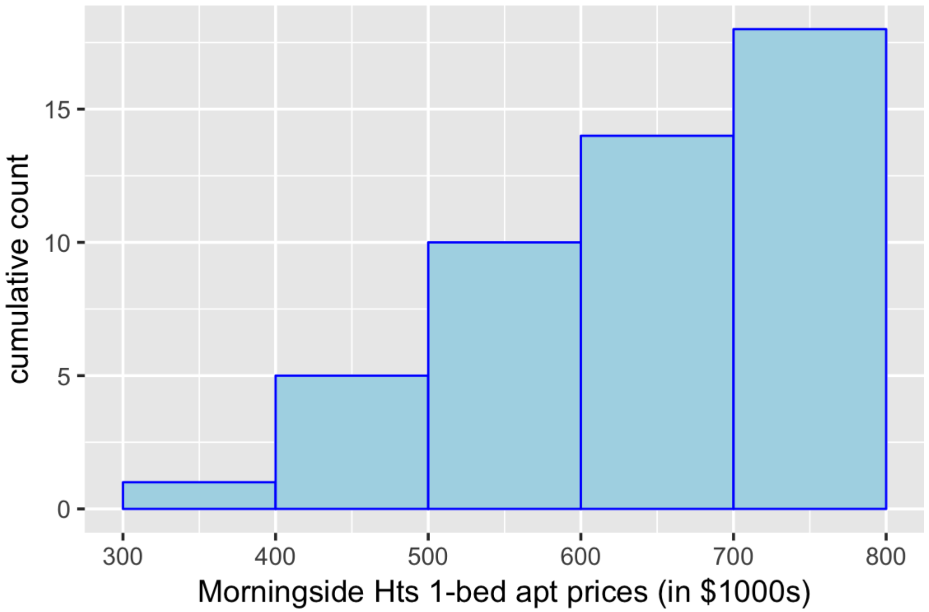
Histogram: what's wrong?



Frequency histogram



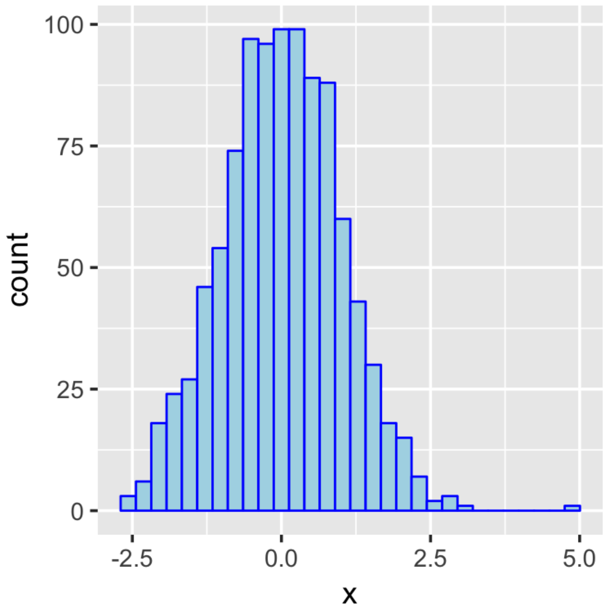
Cumulative frequency histogram

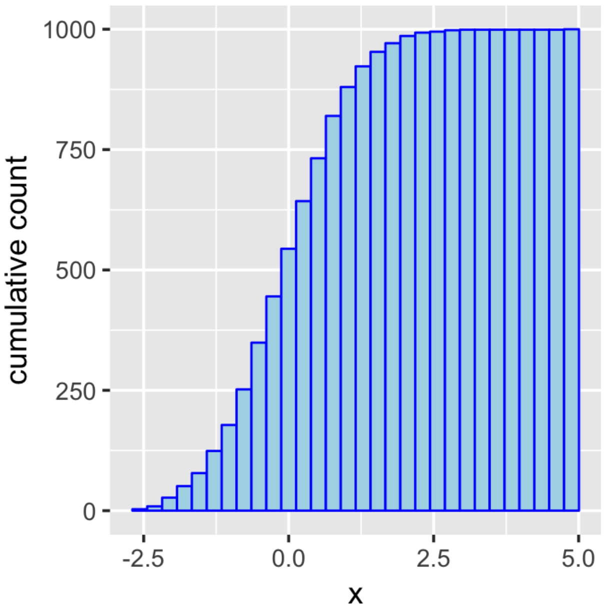


Cumulative frequency histogram

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

(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

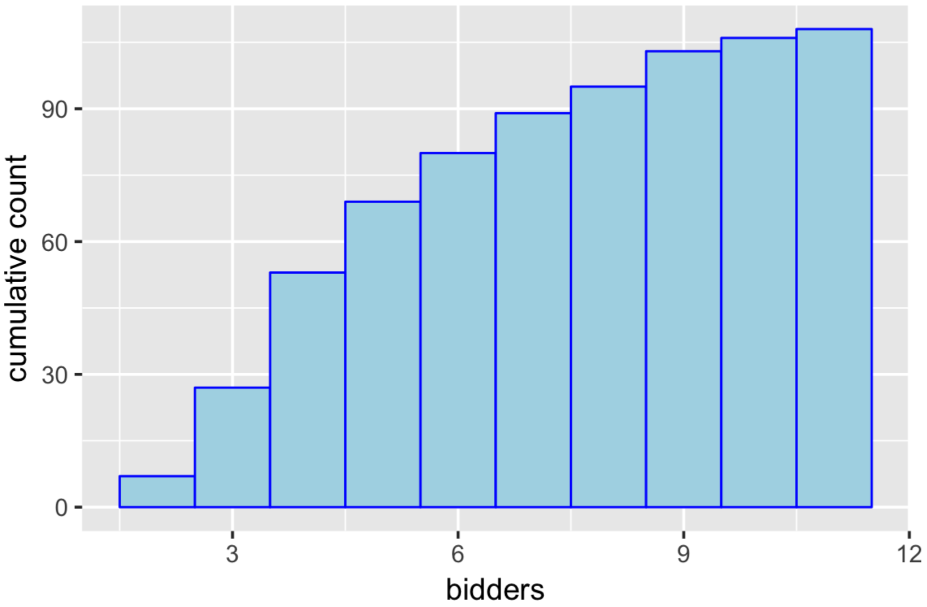
a) What proportion of the contracts involved at most five bidders?

b) What proportion of the contracts involved between five and ten bidders, inclusive?

c) Draw a cumulative frequency histogram.

bidders	contracts
11	2

Cumulative frequency histogram



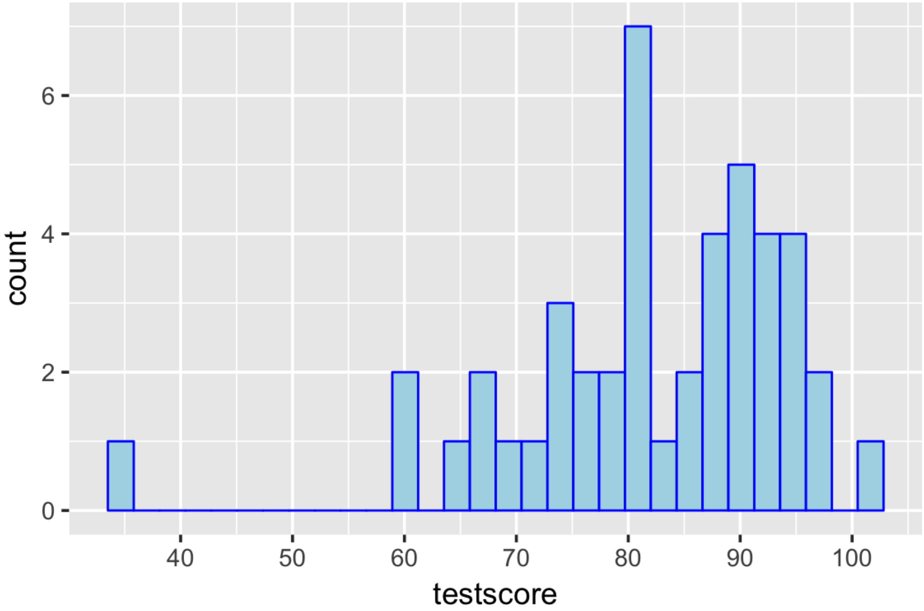
Five number summary

1. min
2. lower fourth
3. median
4. upper fourth
5. max

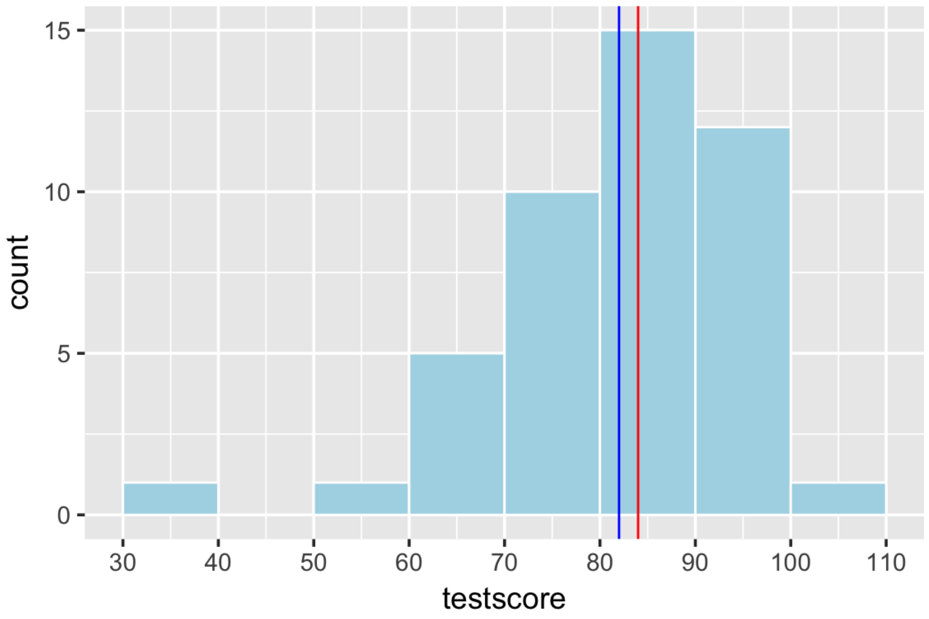
```r
summary(prices)
```

```
##     Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##      379     506     572     593     699     799
```

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 15%**.

.15 x 45 = 6.75 values

Trim 6:

$\frac{6}{45}$ = 0.133

$\bar{x}_{tr(13.33)}$ = 83.667

Trim 7:

$\frac{7}{45}$ = 0.156

$\bar{x}_{tr(15.56)}$ = 83.774

Interpolate:

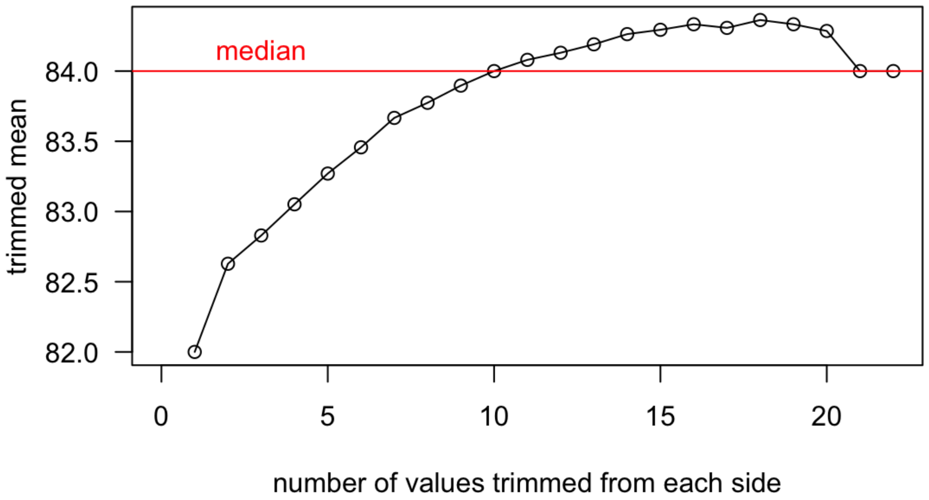
83.667 + .75 * (difference) =

83.667 + .75 * (83.774 - 83.667) =

83.667 + .75 * (.107) =

83.747

Median vs. trimmed mean



Sample and population means

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

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

population median: $\widetilde{\mu}$

sample median: $\widetilde{x}$

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

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

Why n-1?

Short answer: using **n** would result in an underestimation, since the values in the sample are closer to the sample mean than to the true population mean (which we don't know)

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 (p. 47, #62)

Consider the following information on ultimate tensile strength (lb/in^2) for a sample of n = 4 hard zirconium copper wire specimens:

$\bar{x}$ = 76,831
s = 180
smallest x_i = 76,683
largest x_i = 77,048

Set up equations to determine the values of the two middle sample observations. *Do not solve.*

EXERCISE: sd for n = 3

Find the sample mean, variance, and standard deviation:

X1	X2	X3	mean	var	sd
1	2	3			
2	4	6			
0	5	10			
99	100	101			
-8	-5	-2			