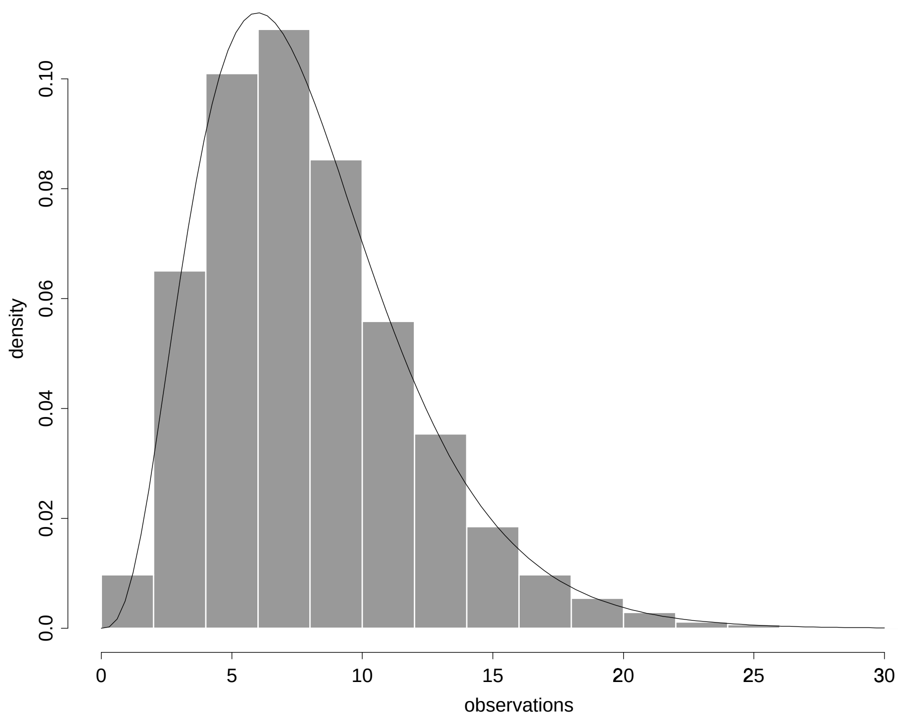
Describing Quantitative Data

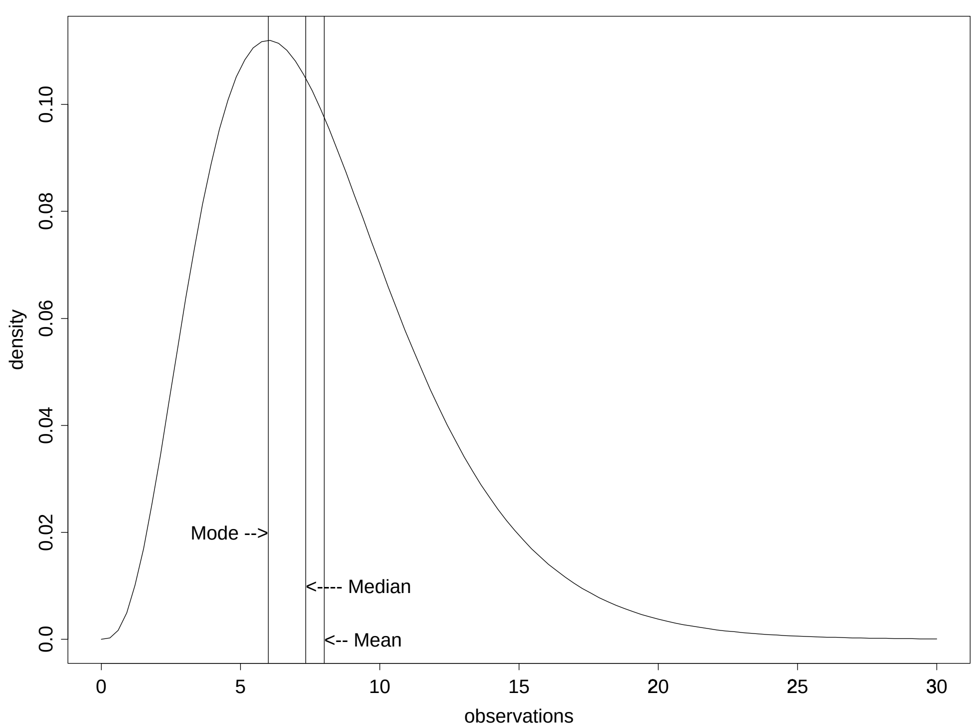
- Symmetry .... (vs: skewed left, skewed right)
- Modes, Gaps .... (how many, describe them)
- Outliers .... (yes/no, describe if present)
- Numerical Summaries .... (choose and report)
  - Mean and SD: symmetric, unimodal, few or no outliers
  - Five Number Summary and IQR: all others!

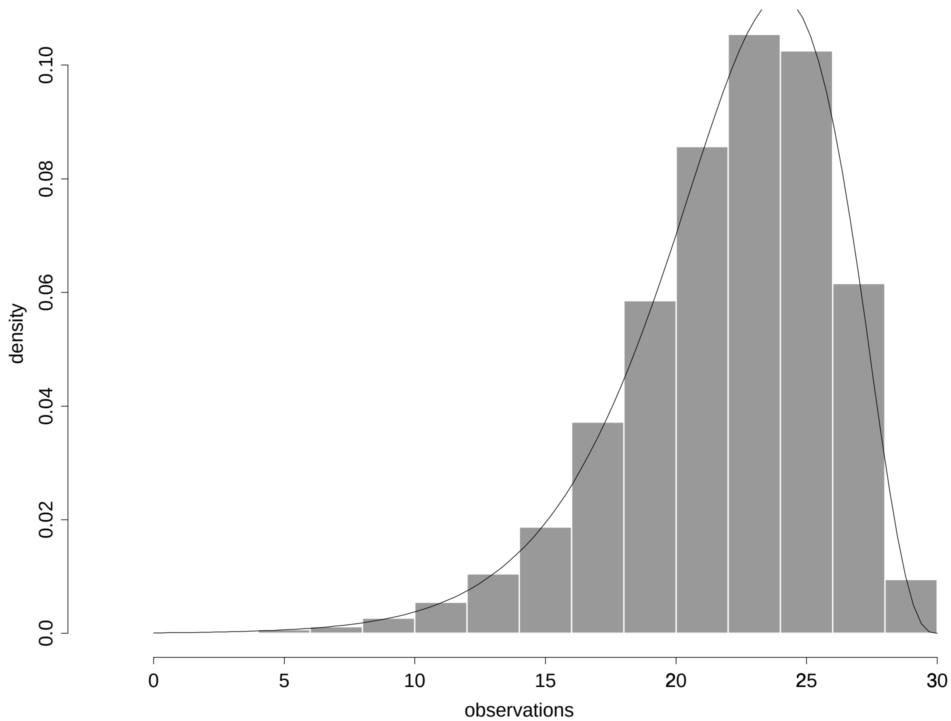
One More Step:

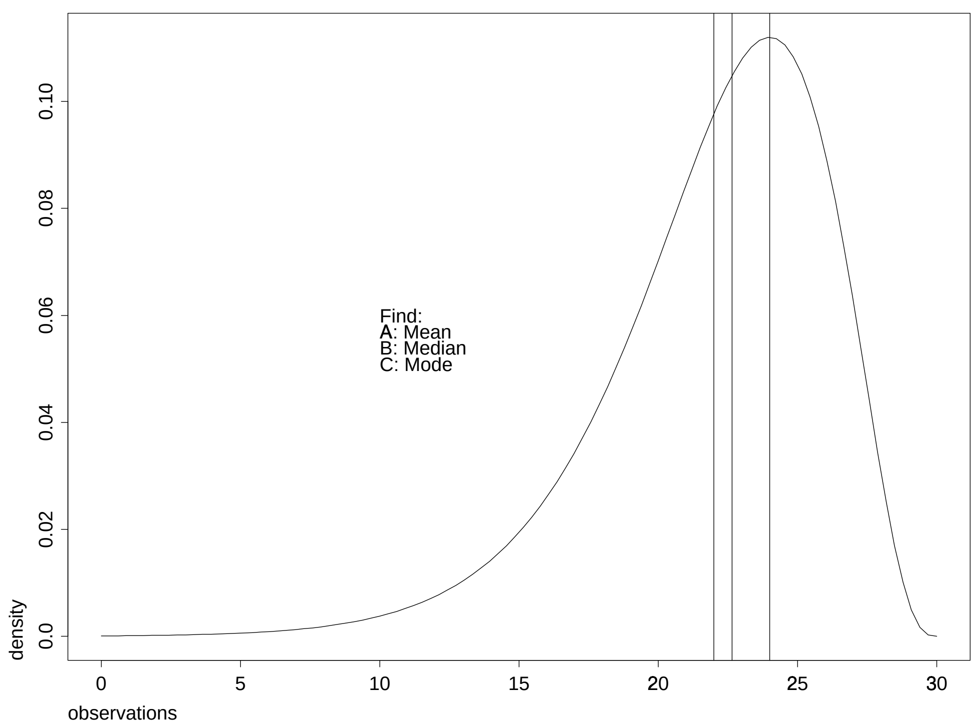
- Sometimes the overall pattern of a large number of observations can be represented by a smooth curve.

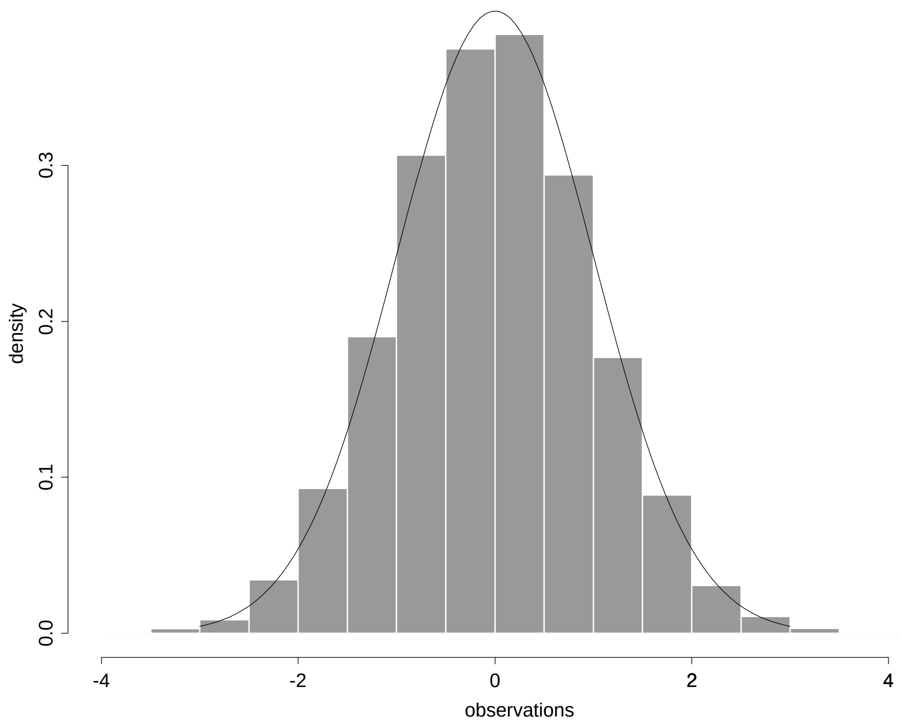
Smooth Curve Densities



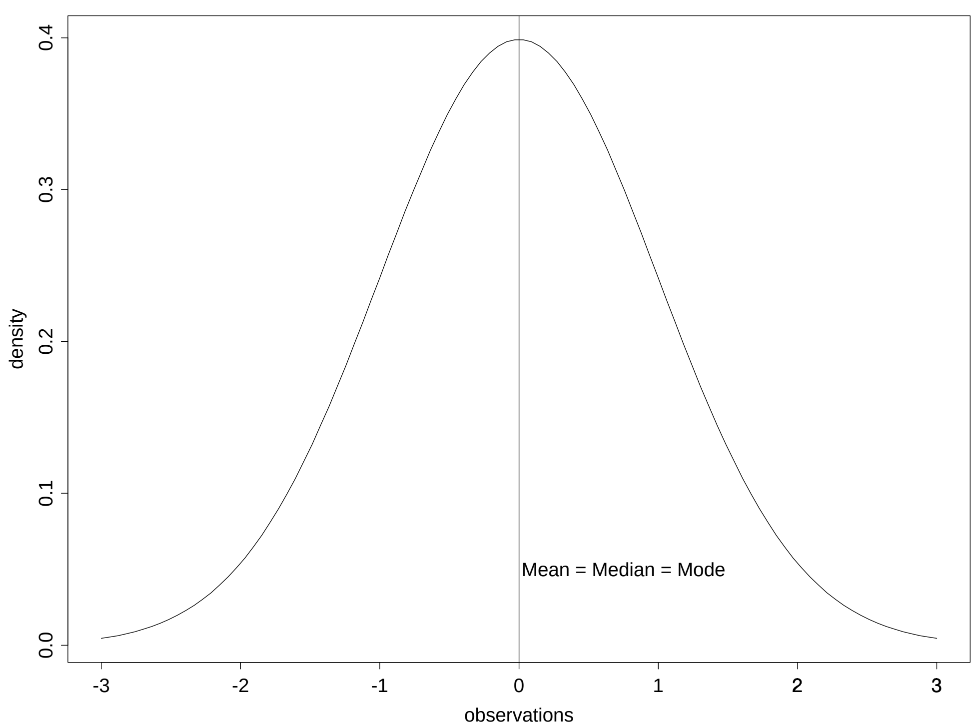



density
0.0
0.02
0.04
0.06
0.08
0.10
0
5
10
15
20
25
30
observations


Find:
A: Mean
B: Median
C: Mode
density
0.0
0.02
0.04
0.06
0.08
0.10
0
5
10
15
20
25
30
observations

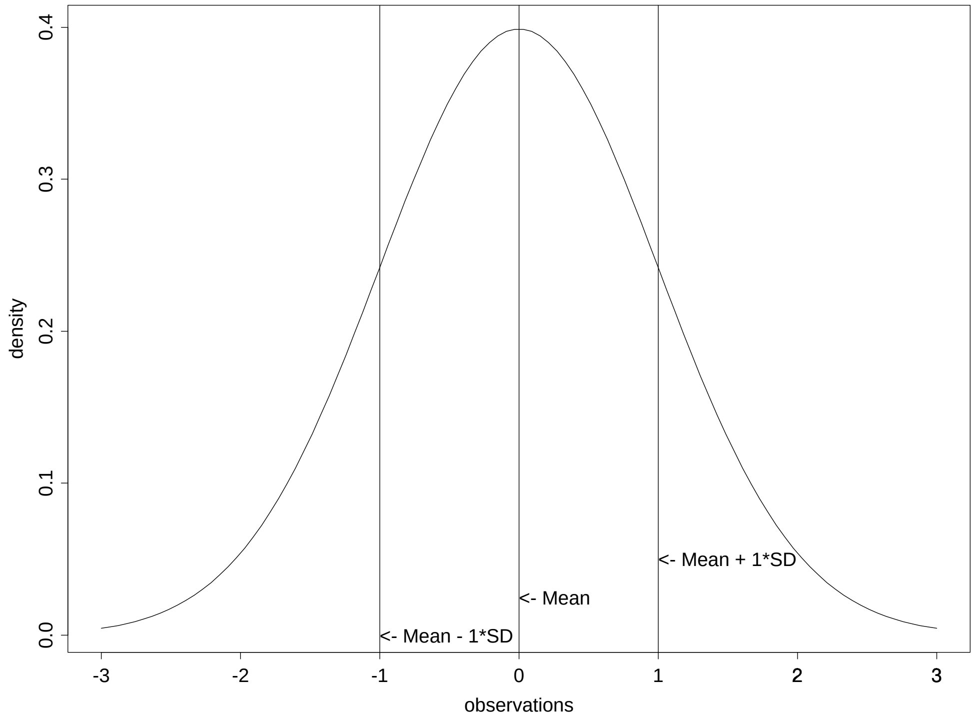


density
0.0
0.1
0.2
0.3
-4
-2
0
2
4
observations

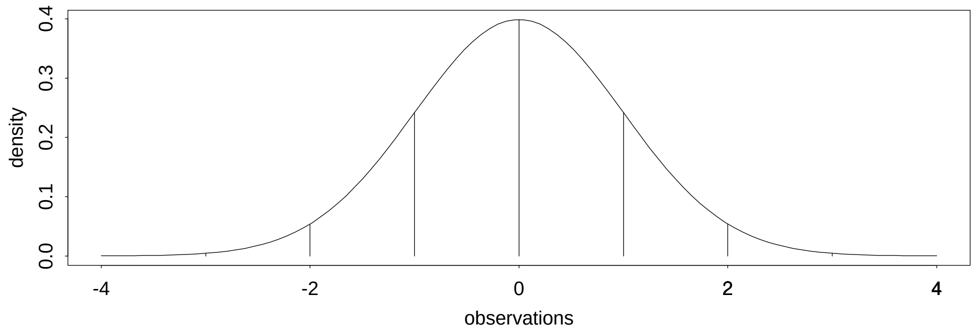


density
0.0
0.1
0.2
0.3
0.4
Mean = Median = Mode
-3
-2
-1
0
1
2
3
observations

Normal Distribution



- Description
  - Symmetric
  - Unimodal
  - Outliers: Expect about 1 or 2 per 100 observations
- Center, spread
  - Mean = Median = Mode
  - SD = distance to "inflection points" (change from cupped-down to cupped-up)

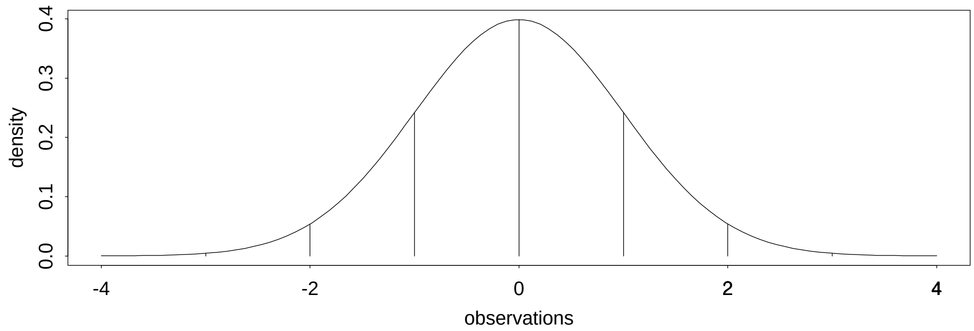


- Area under curve = percent of observations
  - Like a density histogram!
- "Reading" the Normal Curve

The 68%–95%–99.7% Rule

68%	between	Mean $-$ 1$\times$SD	and	Mean $+$ 1$\times$SD
95%	between	Mean $-$ 2$\times$SD	and	Mean $+$ 2$\times$SD
99.7%	between	Mean $-$ 3$\times$SD	and	Mean $+$ 3$\times$SD

Interpolate or *guess* intermediate percentages (there exist tables if you really need accurate answers…)

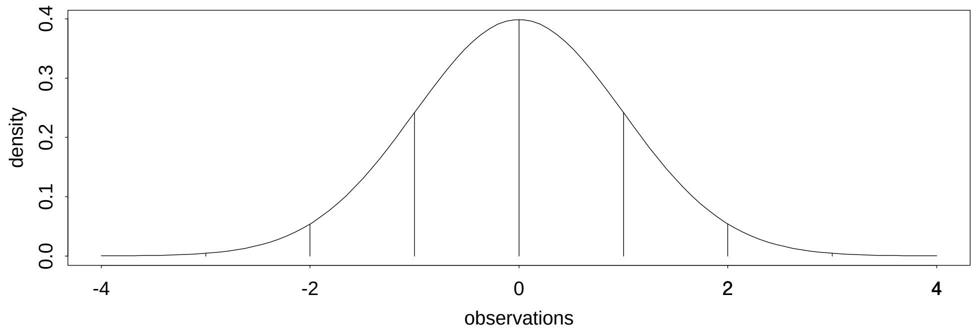


Example: Length of Pregnacies

The length of human pregnancies from conception to birth varies according to a distribution that is approximately normal with mean 266 days and standard deviation 16 days. Use the 68–95–99.7 rule to answer the following questions.

(a) Between what values do the lengths of the middle 95% of all pregnancies fall?

(b) How short are the shortest 2.5% of all pregnancies?



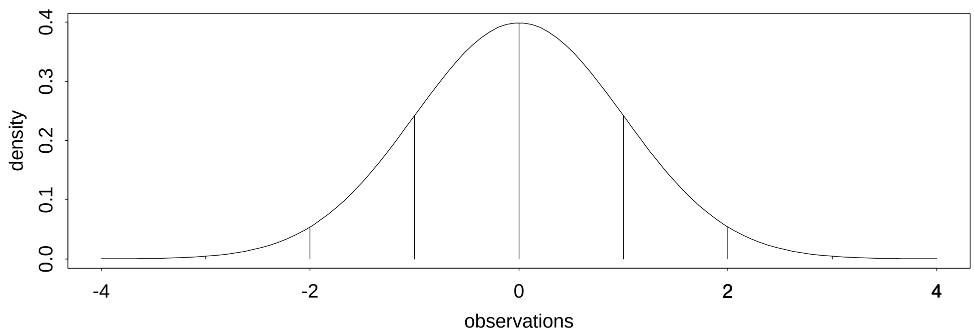
Example: SAT Scores

SAT scores are approximately normal with mean 500 and standard deviation 100. Scores of 800 or higher are reported as 800, so you do not need a perfect paper to score 800 on the SAT.

(a) Mary got a score of 650. About what percent of students scored lower than Mary? (this is Mary's *percentile*.)

(b) What percent of students who take the SAT score 800?

Z-scores ("Standard Scores")



- To use the 68%–95%–99.7% Rule, you really only need to know *"How many SD's from the mean?"*

$$Z\text{-score} = \frac{X - Mean}{SD}$$

- *Pregnancies:* Mean=266, SD=16.

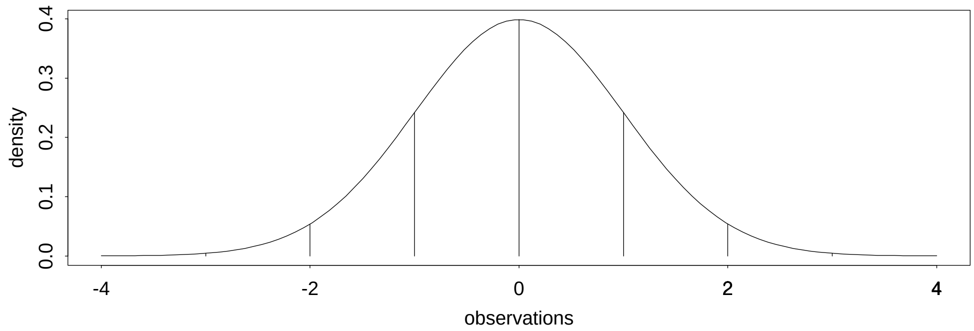
  298 days: Z-score $= \frac{298-266}{16} = \frac{32}{16} = +2.$

  250 days: Z-score $= \frac{250-266}{16} = \frac{-16}{16} = -1.$

- *SAT scores:* Mean=500, SD=100.

  800 points: Z-score $= \frac{800-500}{100} = \frac{300}{100} = +3.$

  650 points: Z-score $= \frac{650-500}{100} = \frac{150}{100} = +1.5.$



Example: SAT vs ACT

Jim scores 700 on the mathematics part of the SAT. Scores on the SAT follow the normal distribution with mean 500 and standard deviation 100. Julie takes the ACT test of mathematical ability, which has mean 18 and standard deviation 6. She scores 24. Use the 68–95–99.7 rule as needed to answer the following.

(a) What is Jim's Z-score? What is his percentile?

(b) What is Sally's Z-score? What is her percentile?

(c) If both tests measure the same kind of ability, who has the higher score?

Transformations, Part I

Do the units of measurement "matter"?

Adding or subtracting a constant:

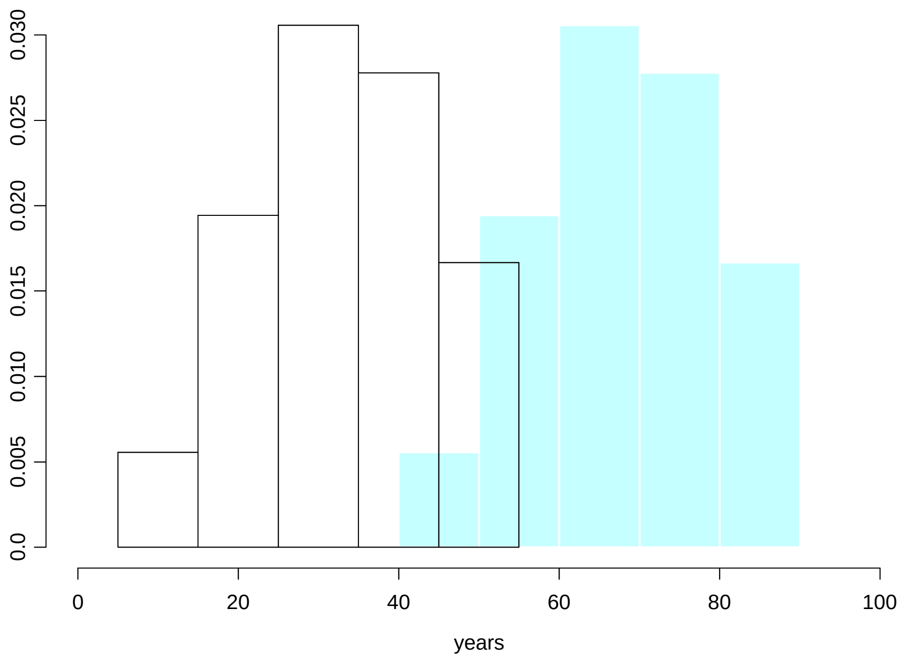
Age at Death of US Presidents:

TABLE 4-3 Age at death of U.S. presidents

Washington	67	Fillmore	74	Roosevelt	60
Adams	90	Pierce	64	Taft	72
Jefferson	83	Buchanan	77	Wilson	67
Madison	85	Lincoln	56	Harding	57
Monroe	73	Johnson	66	Coolidge	60
Adams	80	Grant	63	Hoover	90
Jackson	78	Hayes	70	Roosevelt	63
Van Buren	79	Garfield	49	Truman	88
Harrison	68	Arthur	56	Eisenhower	78
Tyler	71	Cleveland	71	Kennedy	46
Polk	53	Harrison	67	Johnson	64
Taylor	65	McKinley	58	Nixon	81

```
AGE AT DEATH
67 74 60 90 64 72 83 77 67 85 56 57
73 66 60 80 63 90 78 70 63 79 49 88
68 56 78 71 71 46 53 67 64 65 58 81

YEARS SINCE AGE 35 (ELEGIBILITY FOR OFFICE)
32 39 25 55 29 37 48 42 32 50 21 22
38 31 25 45 28 55 43 35 28 44 14 53
33 21 43 36 36 11 18 32 29 30 23 46
```



Data	Mean	SD	Median	Q1	Q3	IQR
Age at Death	69.14	11.28	67.5	61.5	78	16.5
Yrs Since 35	34.14		32.5	26.5	43	

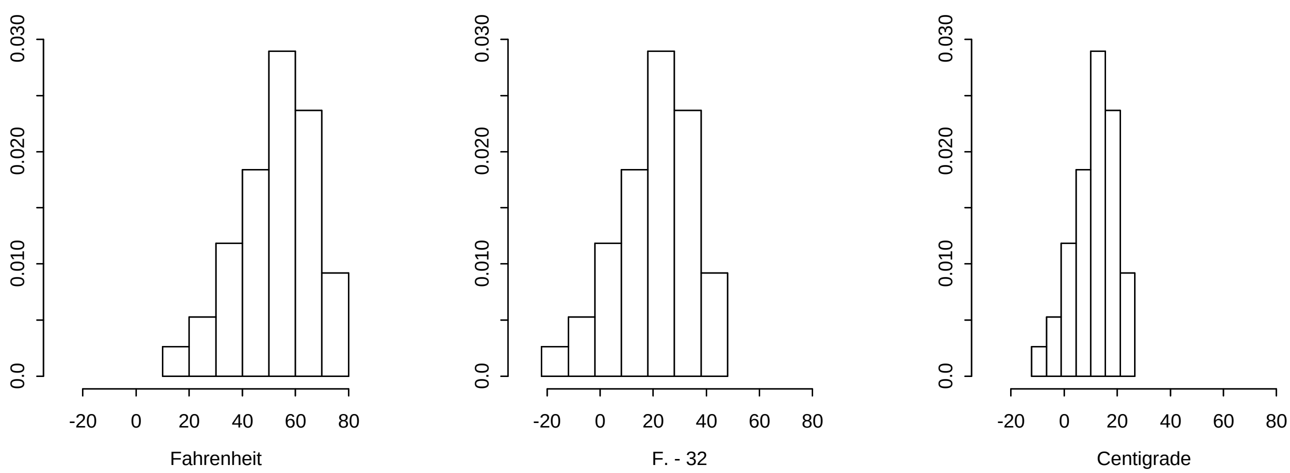
Kennedy: Died age 46; Z-score $= \frac{46-69.14}{11.28} = -2.05$.
Died ll years after elegibility; Z-score $= \frac{11-34.14}{} =$

Nixon: Died age 81; Z-score $= \frac{81-69.14}{11.28} = +1.05$.
Died 46 years after elegibility; Z-score $= \frac{46-34.14}{} =$

Multiplying or dividing by a constant:
Predicted highs in 76 U.S. Cities, Jan 24, 1999.

```
51 58 27 58 56 63 56 11 59 57 45 48 67 48 63 36
43 38 65 43 70 41 46 36 38 72 14 63 52 70 72 43
64 34 52 57 61 60 46 56 80 34 27 53 63 60 66 70
40 68 58 73 27 44 58 60 34 62 50 51 45 74 61 52
80 40 27 46 68 57 72 68 58 59 49 65
```

Original readings in Fahrenheit; what happens if we switch to Centigrade? [$C = (F - 32) \times 5/9$]



Data	Mean	SD	Median	Q1	Q3	IQR
F	53.13	14.63	56.50	43.50	63.00	19.50
$F-32$	21.13		24.50	11.50	31.00	
$(F-32)\times 5/9$	11.74		13.61	6.39	17.22	

Lowest high: Billings Montana

- 11^o Fahrenheit: Z-score $= \frac{11-53.13}{14.63} = -2.88.$
- $^oF - 32 = -21$: Z-score $= \frac{-21-21.13}{14.63} = -2.88.$
- 11.67^o Centigrade: Z-score $= \frac{-11.67-11.74}{8.13} = -2.88$