Statistics with R

Lecture 2

Logic of quantitative research

Step 1: Develop a research question

Step 2: Operationalize the research question

Step 3A (*in ideal world*): Obtain **data** from the whole **population**

Step 3B (*in reality*): Collect **data** from a **sample** (e.g., corpus / experiment)

Step 4: Describe the data: **descriptive statistics**

Step 5 (*in ideal world, could skip this step*): Make inferences: **inferential statistics**

- use probability theory to make an educated guess to what extent descriptive statistics from the **sample** applies to the **population**

Step 6: Draw a conclusion

Basics of inferential statistics:

(Step 5: use probability theory to make an educated guess to what extent descriptive statistics from the **sample** applies to the **population)**

1. Let's look at our sample
   - Visualizing data (goal: understanding the plots we will look at)
   - Measures of central tendency and variability

     (goal: learn about definitions of important concepts)
2. Normal distribution and probability
3. Sampling distribution of the mean, Central Limit Theorem

(goal: understand logics behind (frequentist) statistics and why the normal distribution is so important)

Plotting data

- Learn about your data!
  - Are there any errors in data collection?
  - What properties does your data have? (-> data distribution)
  - Can you see the effects you expected?

- How is the data distributed – what statistical tests can you use?

- Are there any correlations between variables?

An example dataset: body temperature of two beavers

```
> beaver1$temp
  [1] 36.33 36.34 36.35 36.42 36.55 36.69 36.71 36.75 36.81 36.88 36.89 36.91 36.85
 [14] 36.89 36.89 36.67 36.50 36.74 36.77 36.76 36.78 36.82 36.89 36.99 36.92 36.99
 [27] 36.89 36.94 36.92 36.97 36.91 36.79 36.77 36.69 36.62 36.54 36.55 36.67 36.69
 [40] 36.62 36.64 36.59 36.65 36.75 36.80 36.81 36.87 36.87 36.89 36.94 36.98 36.95
 [53] 37.00 37.07 37.05 37.00 36.95 37.00 36.94 36.88 36.93 36.98 36.97 36.85 36.92
 [66] 36.99 37.01 37.10 37.09 37.02 36.96 36.84 36.87 36.85 36.85 36.87 36.89 36.86
 [79] 36.91 37.53 37.23 37.20 37.25 37.20 37.21 37.24 37.10 37.20 37.18 36.93 36.83
 [92] 36.93 36.83 36.80 36.75 36.71 36.73 36.75 36.72 36.76 36.70 36.82 36.88 36.94
[105] 36.79 36.78 36.80 36.82 36.84 36.86 36.88 36.93 36.97 37.15
> beaver2$temp
  [1] 36.58 36.73 36.93 37.15 37.23 37.24 37.24 36.90 36.95 36.89 36.95 37.00 36.90
 [14] 36.99 36.99 37.01 37.04 37.04 37.14 37.07 36.98 37.01 36.97 36.97 37.12 37.13
 [27] 37.14 37.15 37.17 37.12 37.12 37.17 37.28 37.28 37.44 37.51 37.64 37.51 37.98
 [40] 38.02 38.00 38.24 38.10 38.24 38.11 38.02 38.11 38.01 37.91 37.96 38.03 38.17
 [53] 38.19 38.18 38.15 38.04 37.96 37.84 37.83 37.84 37.74 37.76 37.76 37.64 37.63
 [66] 38.06 38.19 38.35 38.25 37.86 37.95 37.95 37.76 37.60 37.89 37.86 37.71 37.78
 [79] 37.82 37.76 37.81 37.84 38.01 38.10 38.15 37.92 37.64 37.70 37.46 37.41 37.46
 [92] 37.56 37.55 37.75 37.76 37.73 37.77 38.01 38.04 38.07
```

It's very hard to see what's going on!

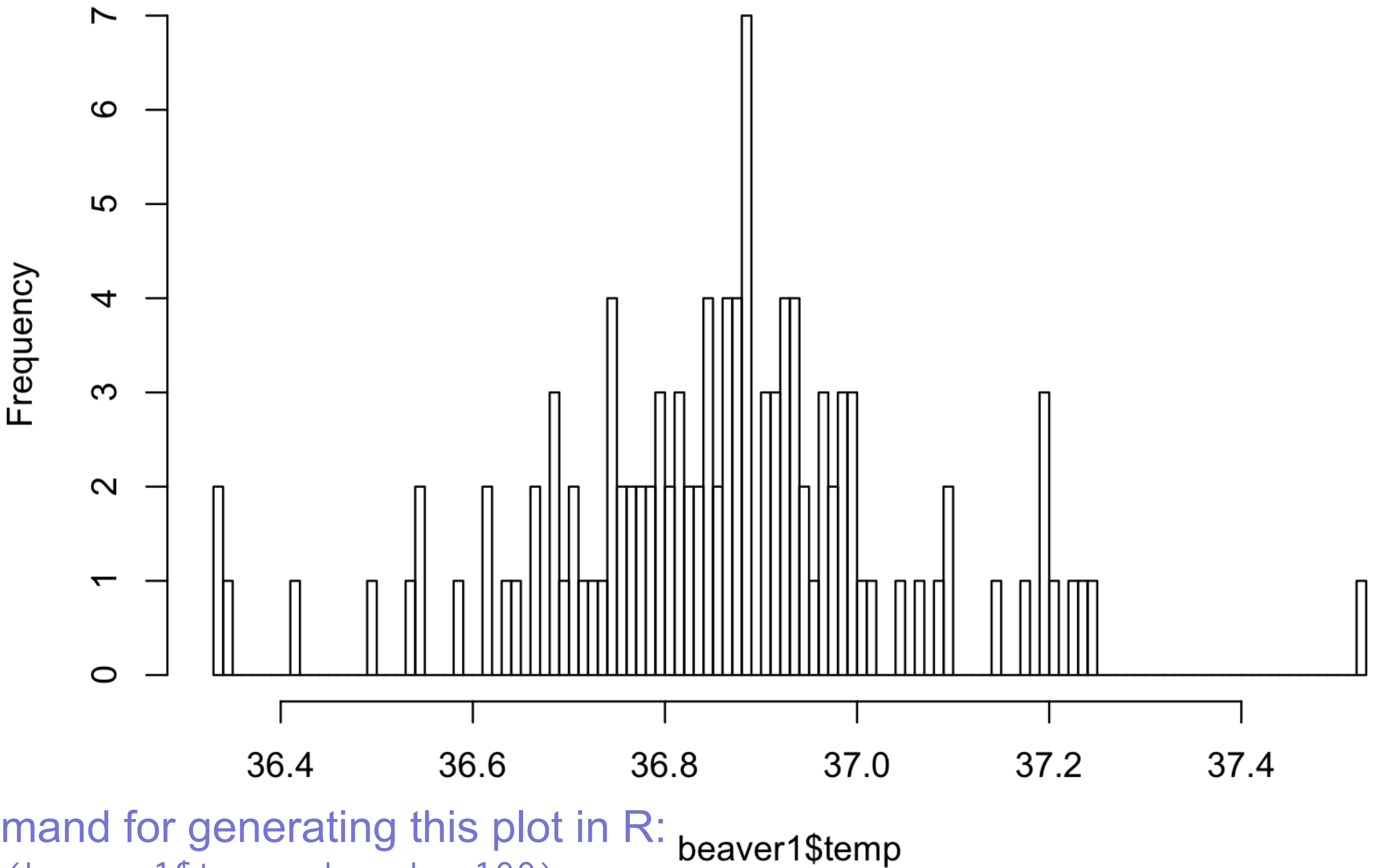
- Do they have similar patterns of body temperature?

- Staring at the numbers makes it hard for humans to say anything about these data

- One useful way of understanding this data better is to plot the data to see how often each body temperature has been observed.

- For example, temperature 36.89 occurred 7 times in the first beaver.

Histogram of beaver1's temperature



command for generating this plot in R:

```
hist(beaver1$temp, breaks=100)
```

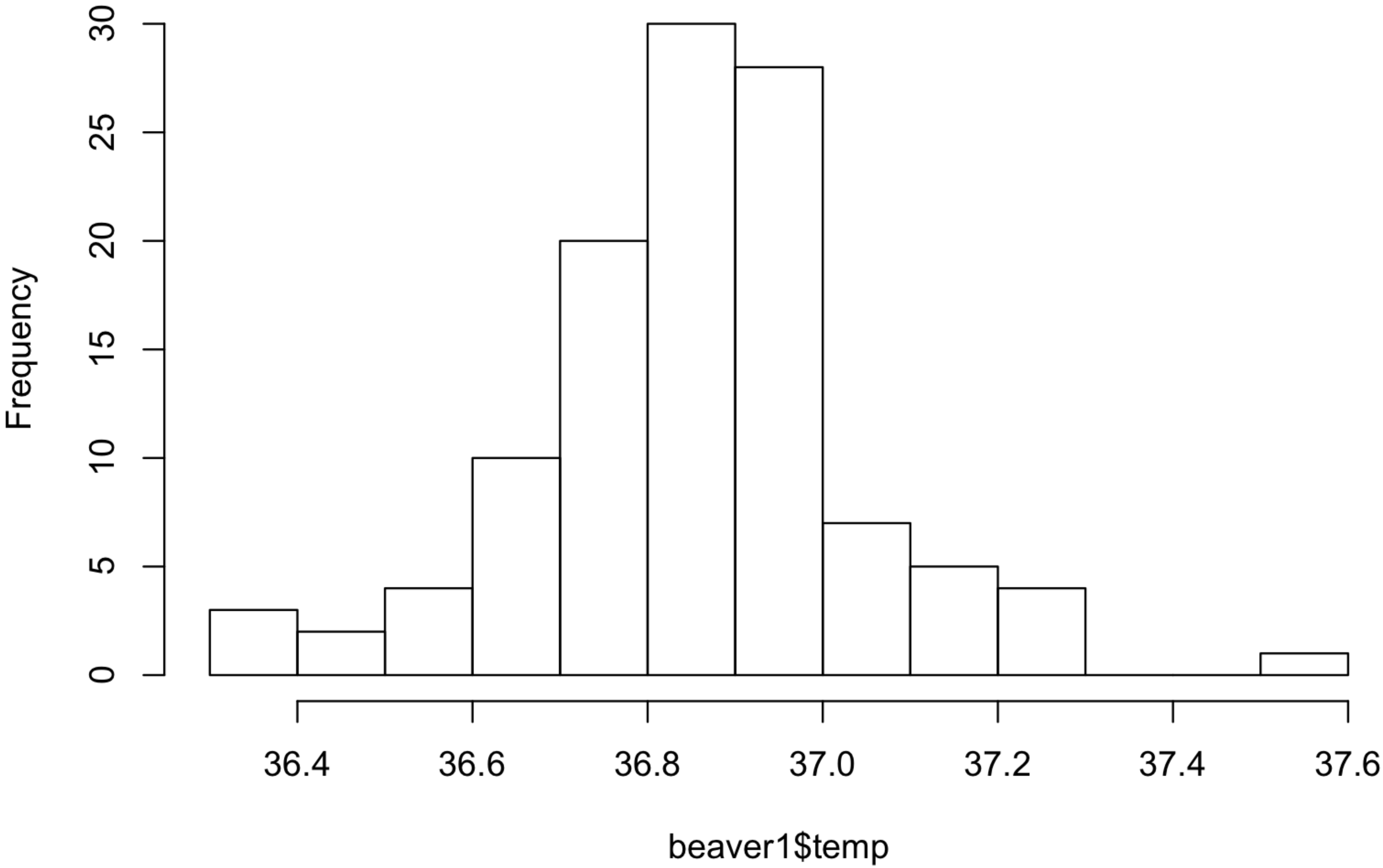
Interpreting a histogram

- For example, temperature 36.89 occurred 7 times in the first beaver.

- Maybe the exact difference between a temperature of 36.89, 36.85 and 36.9 isn't all that important…

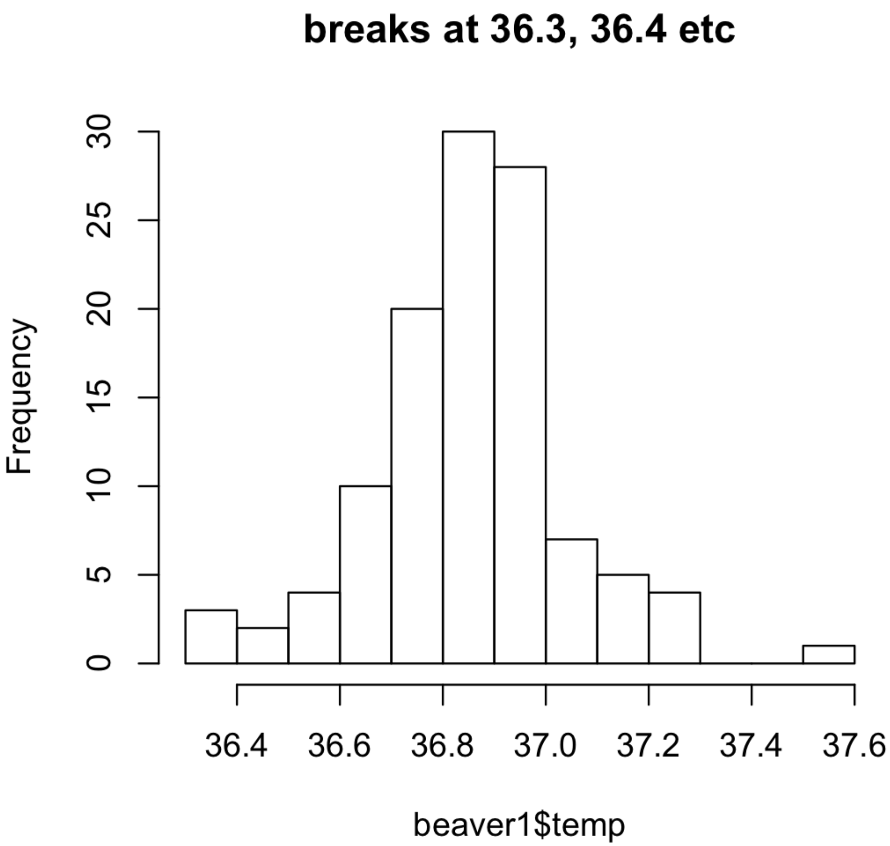
- We can add all the occurrences of temperatures in a specific interval into one "bin".

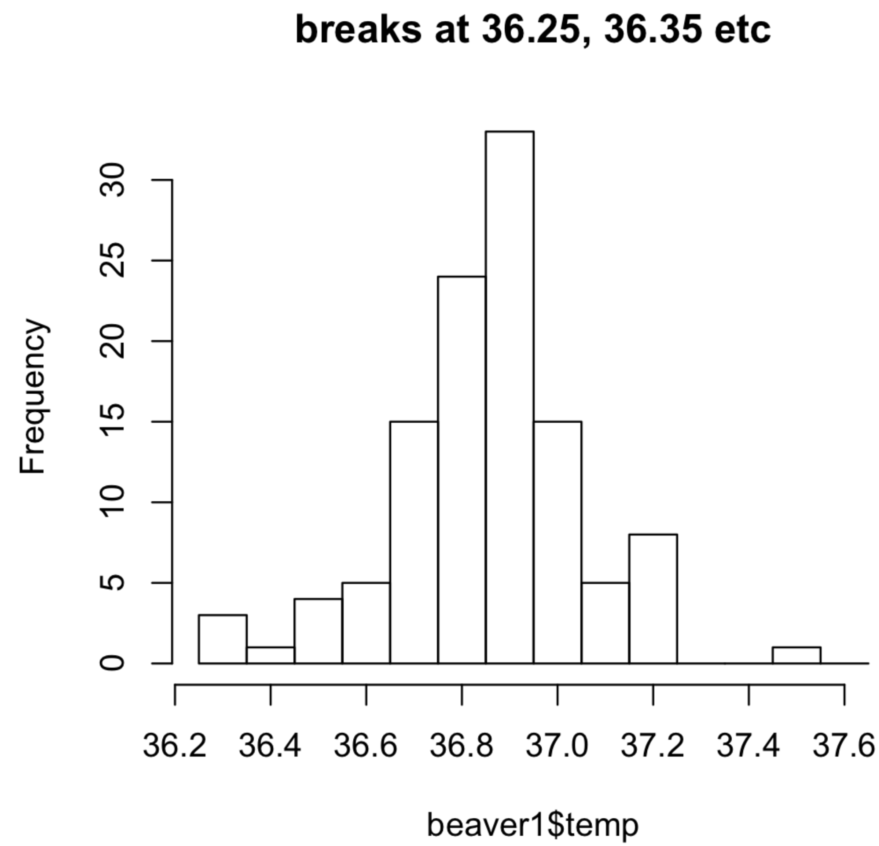
Histogram of beaver1's temperature



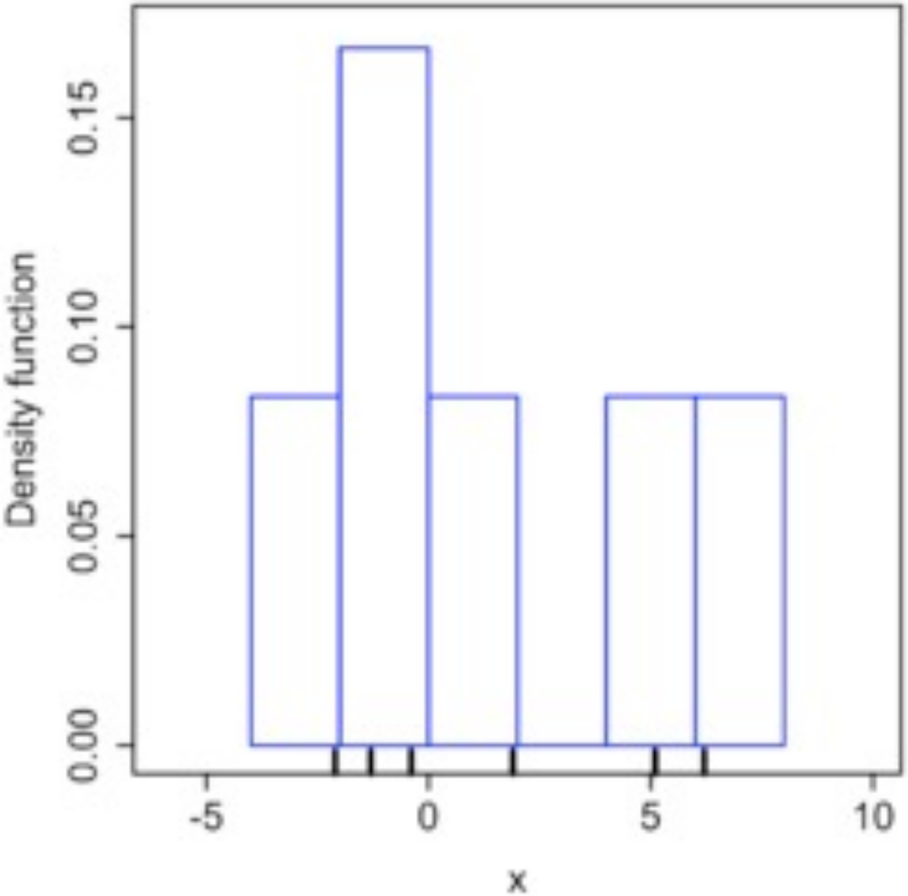
Disadvantages of breaking up into arbitrary intervals

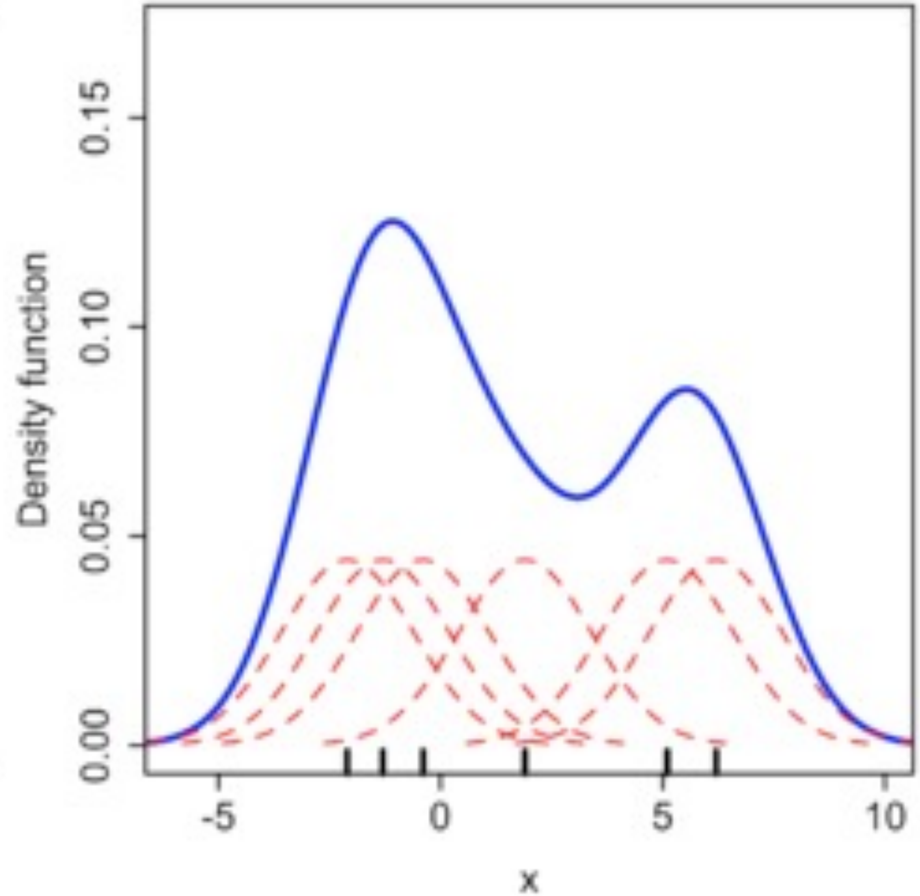
- How to decide when to start a new bin?
- Same data can look different:



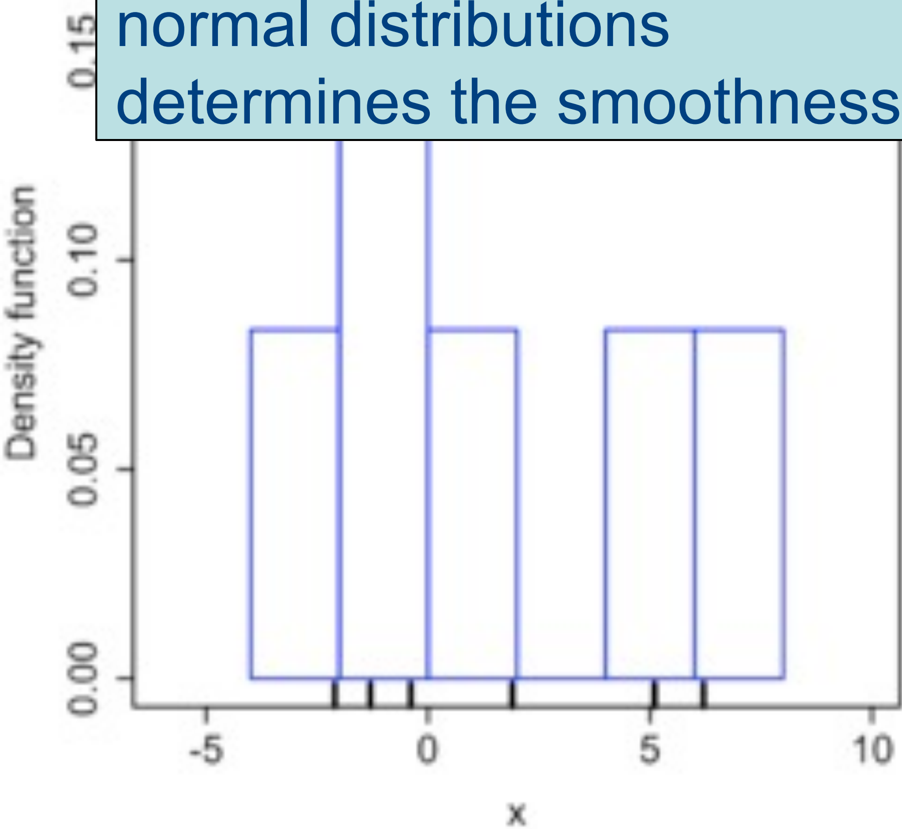


Kernel Density Plot

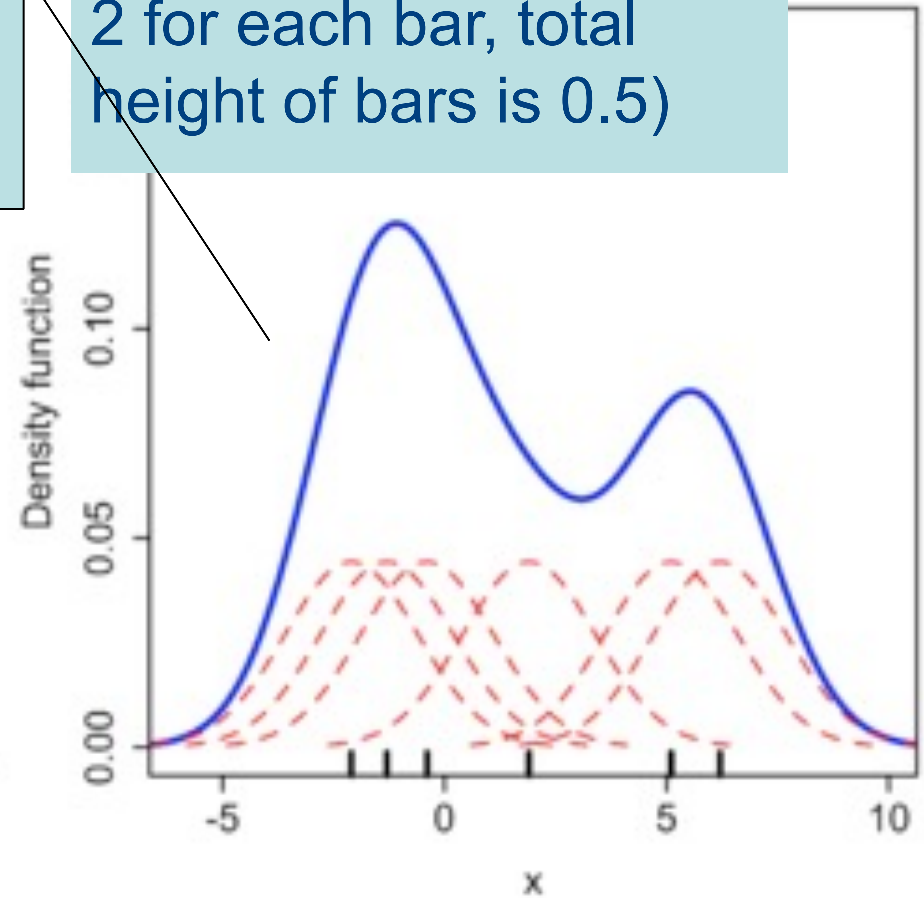




Instead of using small uniform distributions (rectangles as in other plot), normal distributions are used. The width of the normal distributions determines the smoothness.

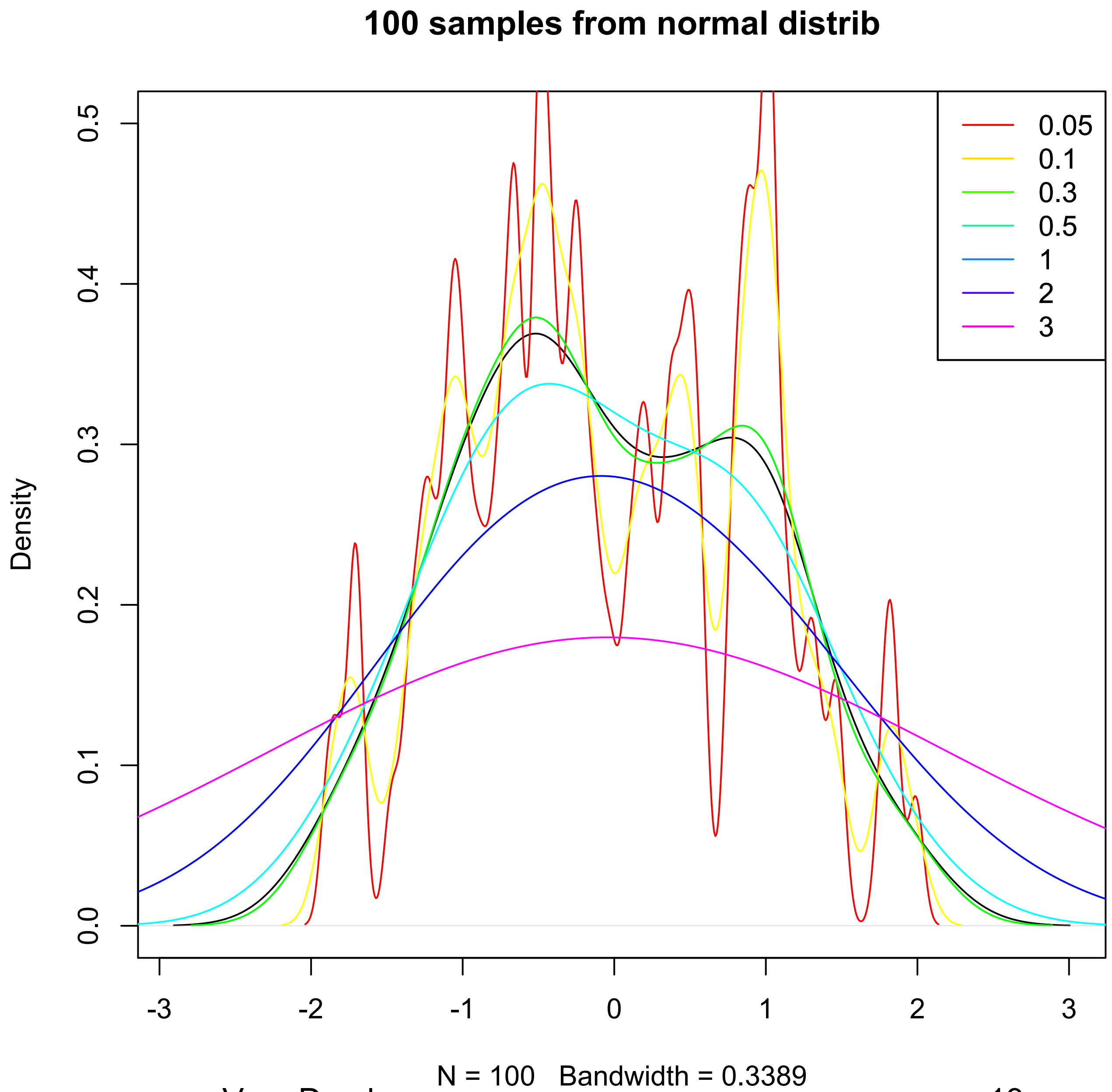


Similar to histogram, but area inside bars equals 1 (here, width of 2 for each bar, total height of bars is 0.5)

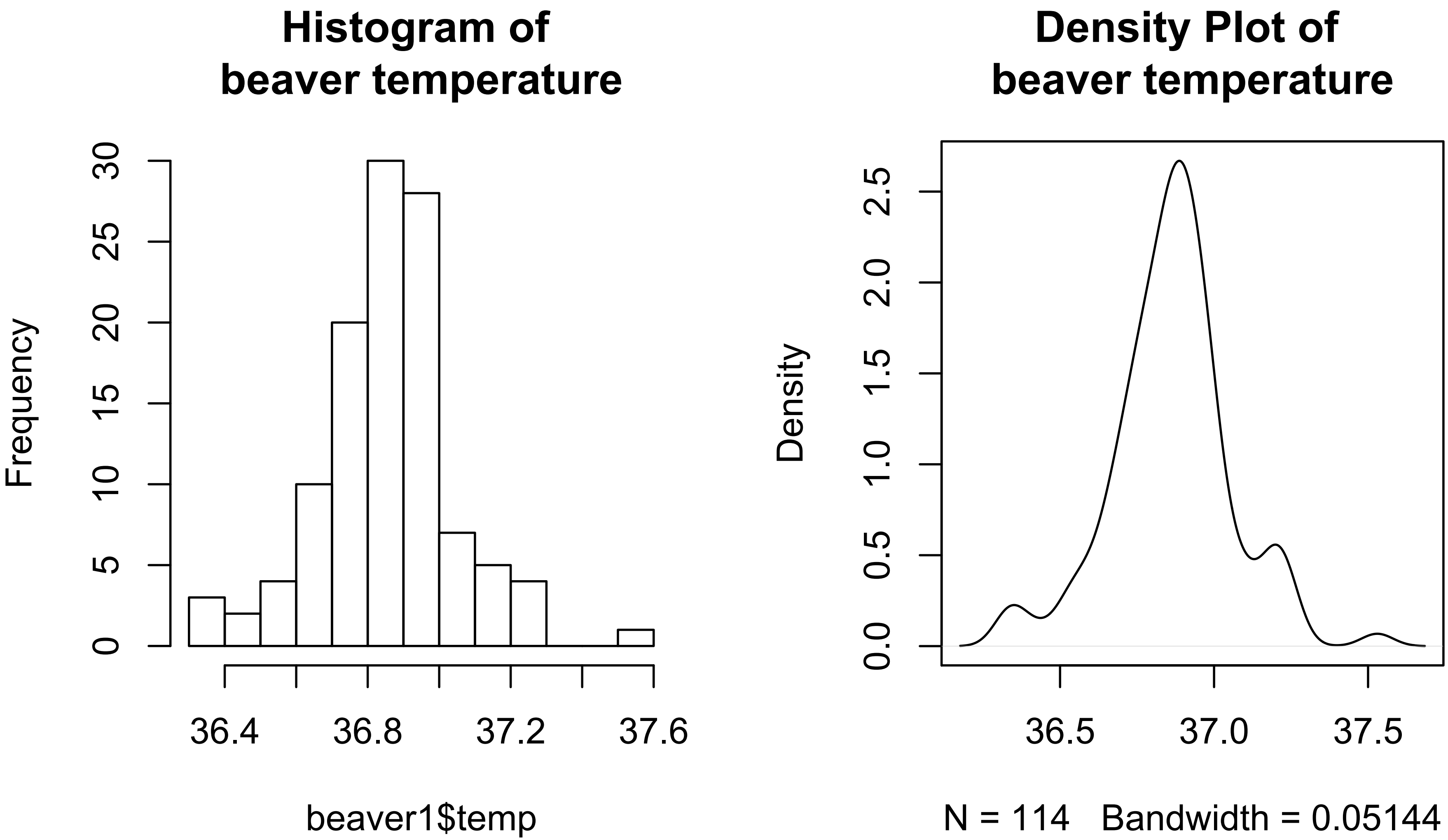


Kernel Density Plot

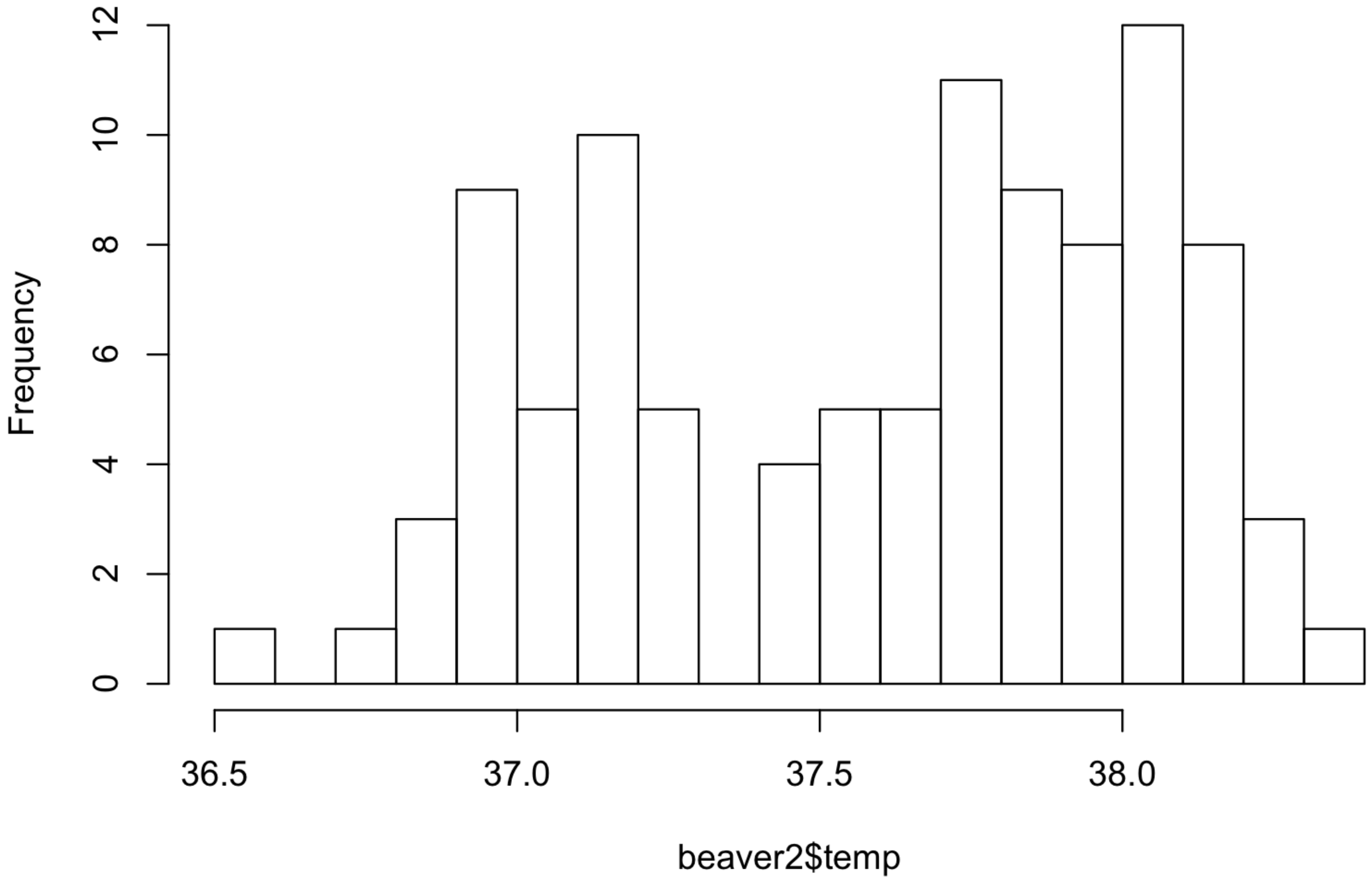
Smoothness



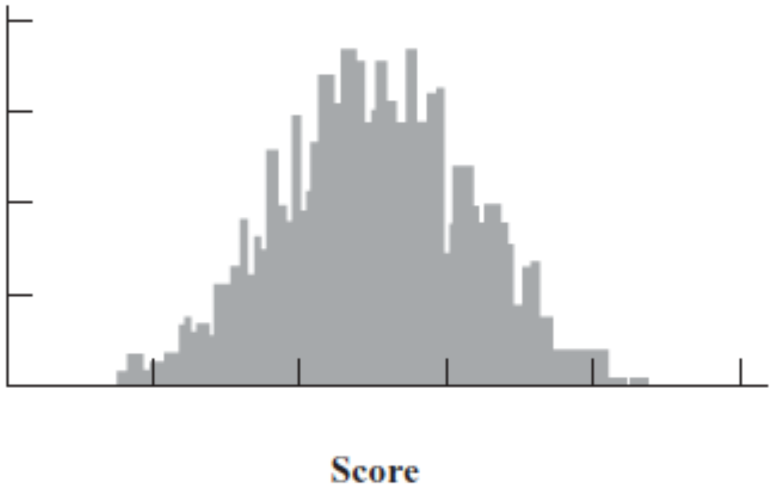
Histogram and Density Plot of same data



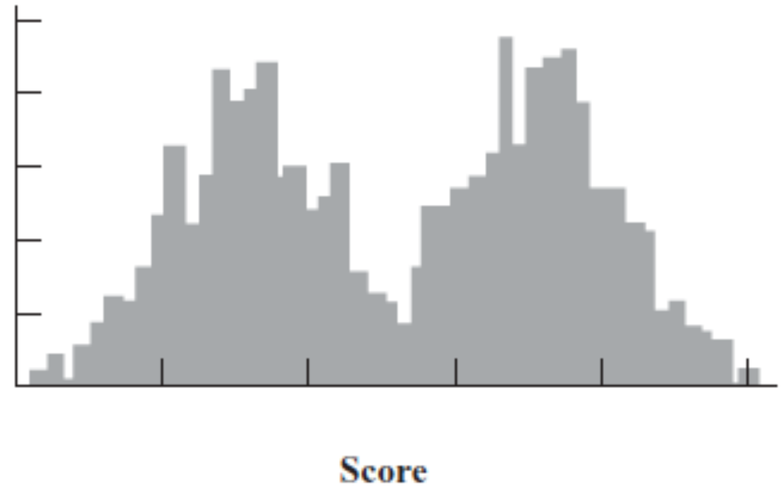
Histogram of beaver2's temperature



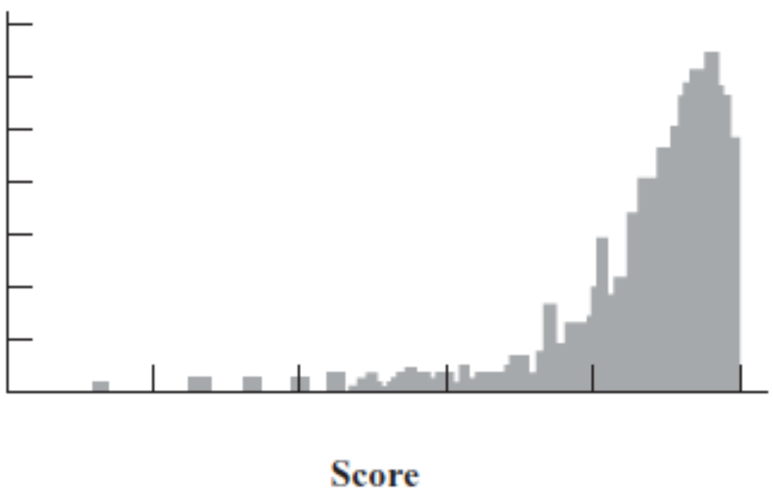
Shapes of distributions



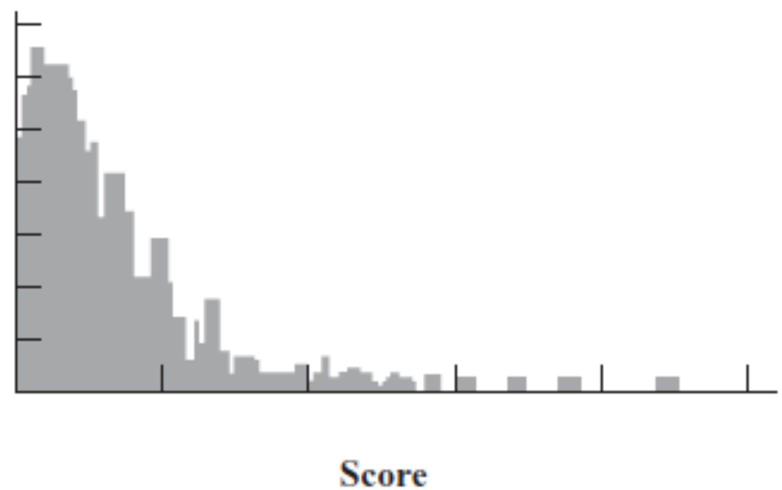
(a) Normal



(b) Bimodal

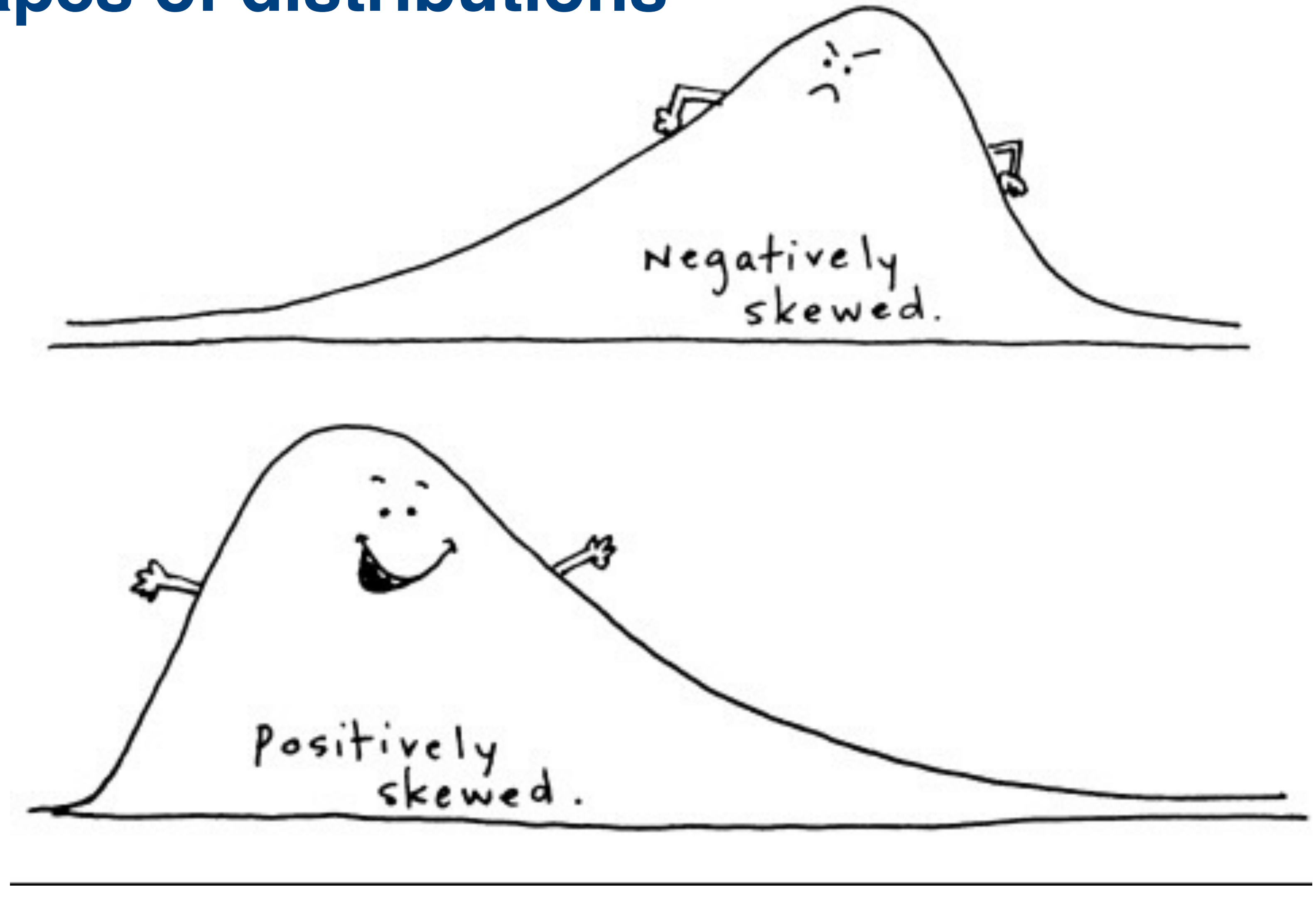


(c) Negatively skewed

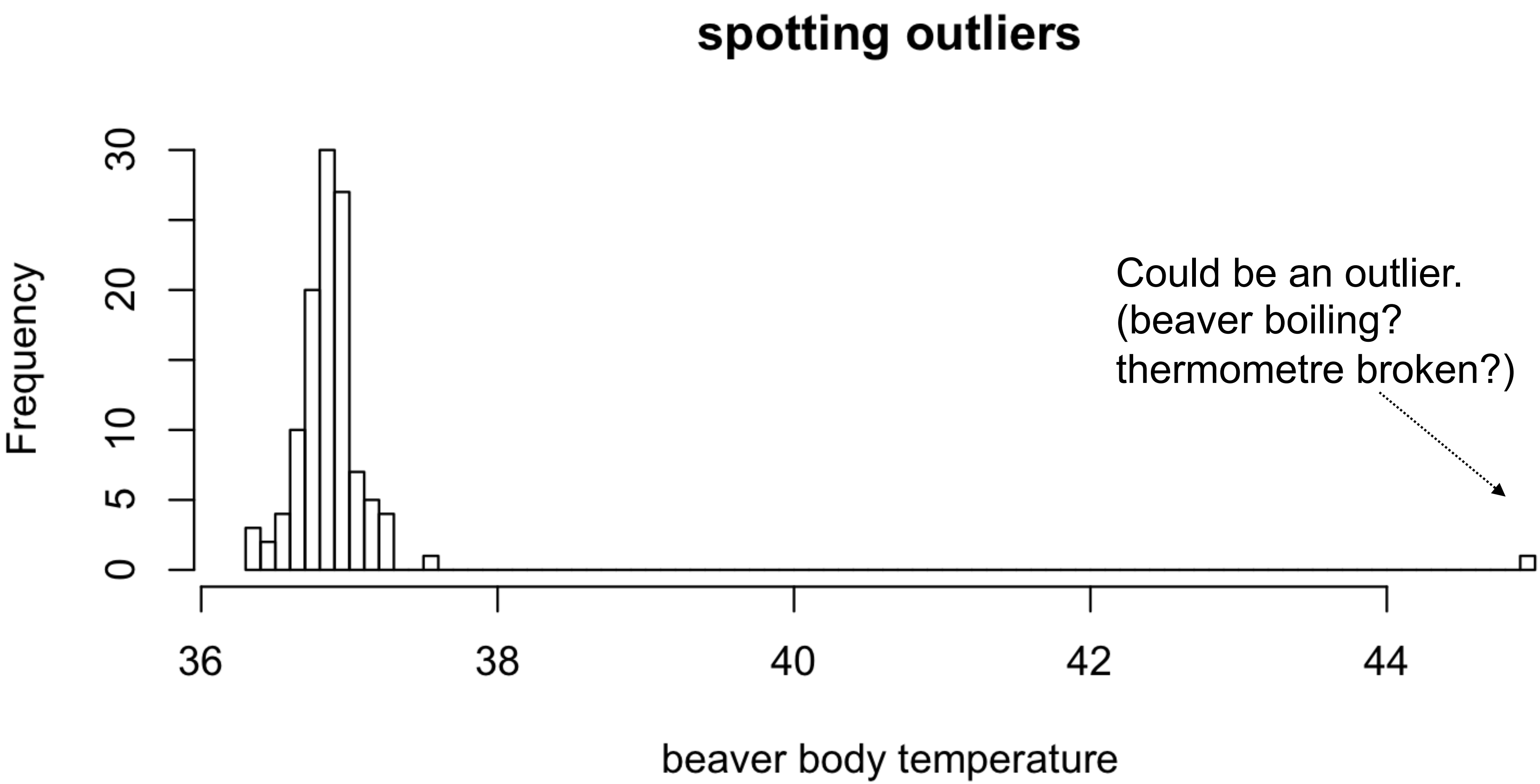


(d) Positively skewed

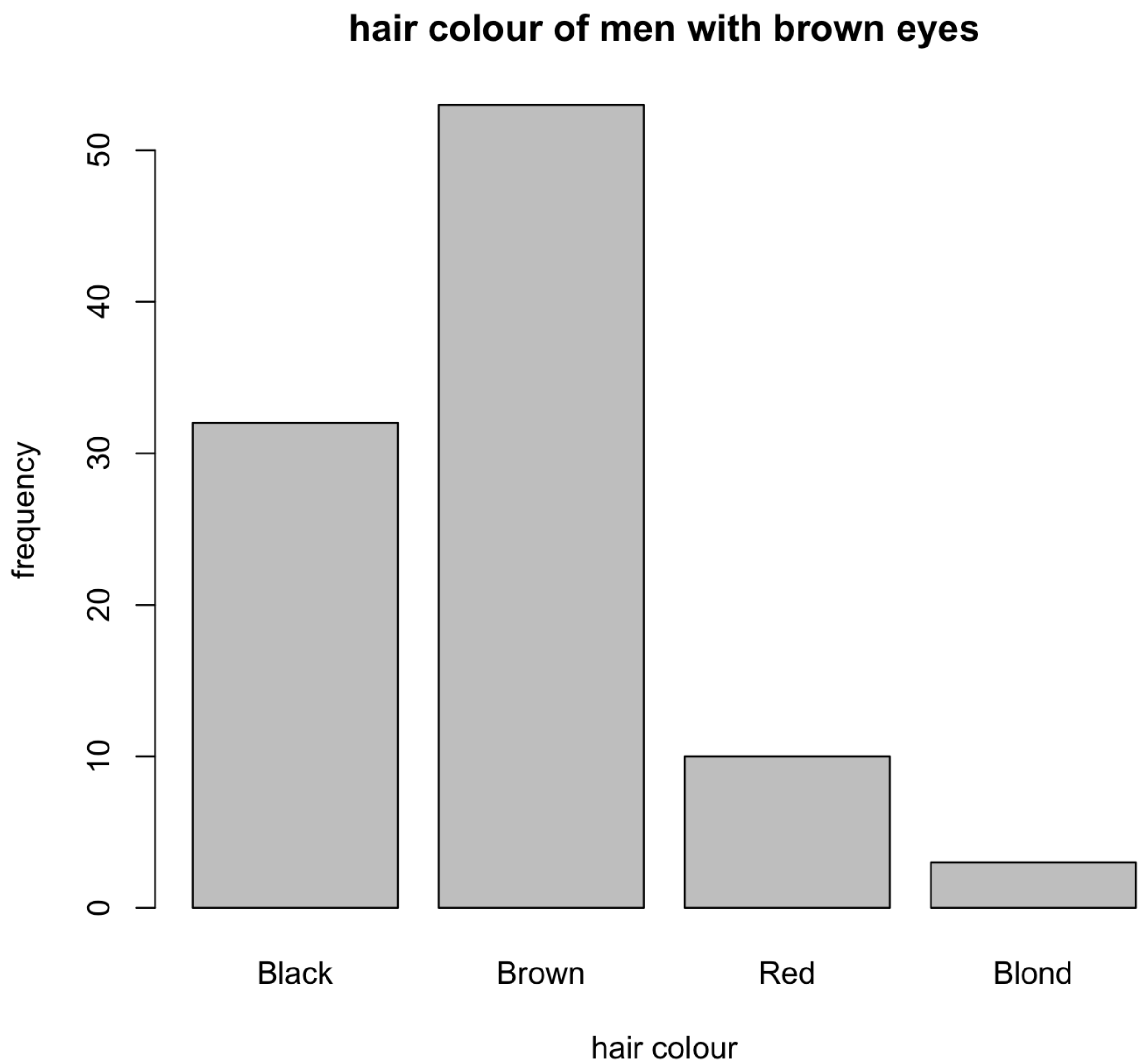
Shapes of distributions



Usefulness of data plotting: identifying outliers



Visualizing discrete data: Bar Chart



Dataset:
"HairEyeColor"

Plot type in R:
"barplot"

Lecture outline

- Exploring data
  - Visualizing data
  - Measures of central tendency and variability
- Normal distribution and probability
- Sampling distribution of the mean and the Central Limit Theorem

Measures of central tendency and variability

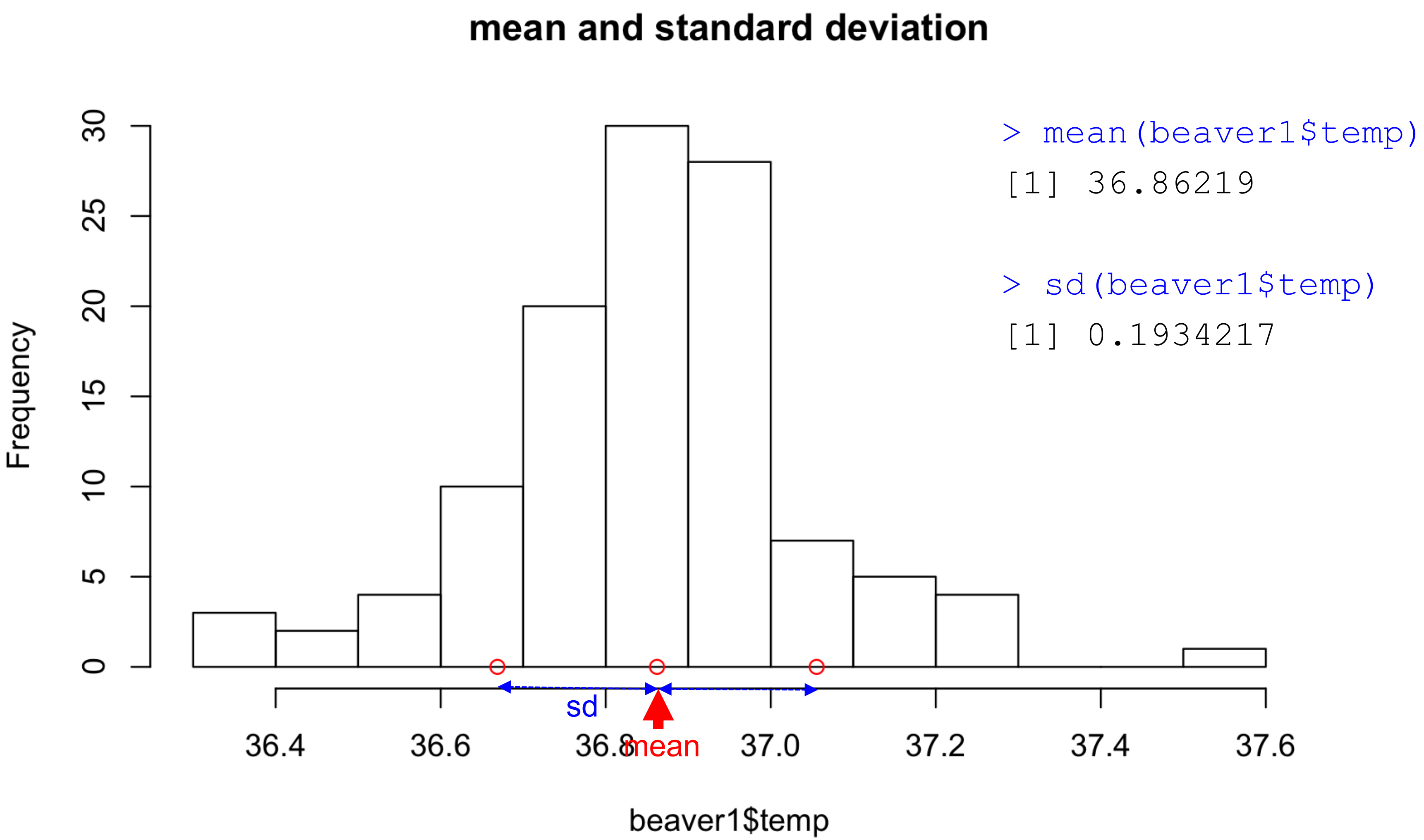
sample mean: $\bar{x} = \frac{\sum x_i}{n}$

variance: $\mathrm{var} = \frac{\sum (x_i - \bar{x})^2}{n-1}$

sample standard deviation: $sd = \sqrt{\frac{\sum (x_i - \bar{x})^2}{n-1}}$

for step-by-step explanation for why the standard deviation has this form, see Section 2.8 in Howell book.

Mean and standard deviation for our beaver1 example

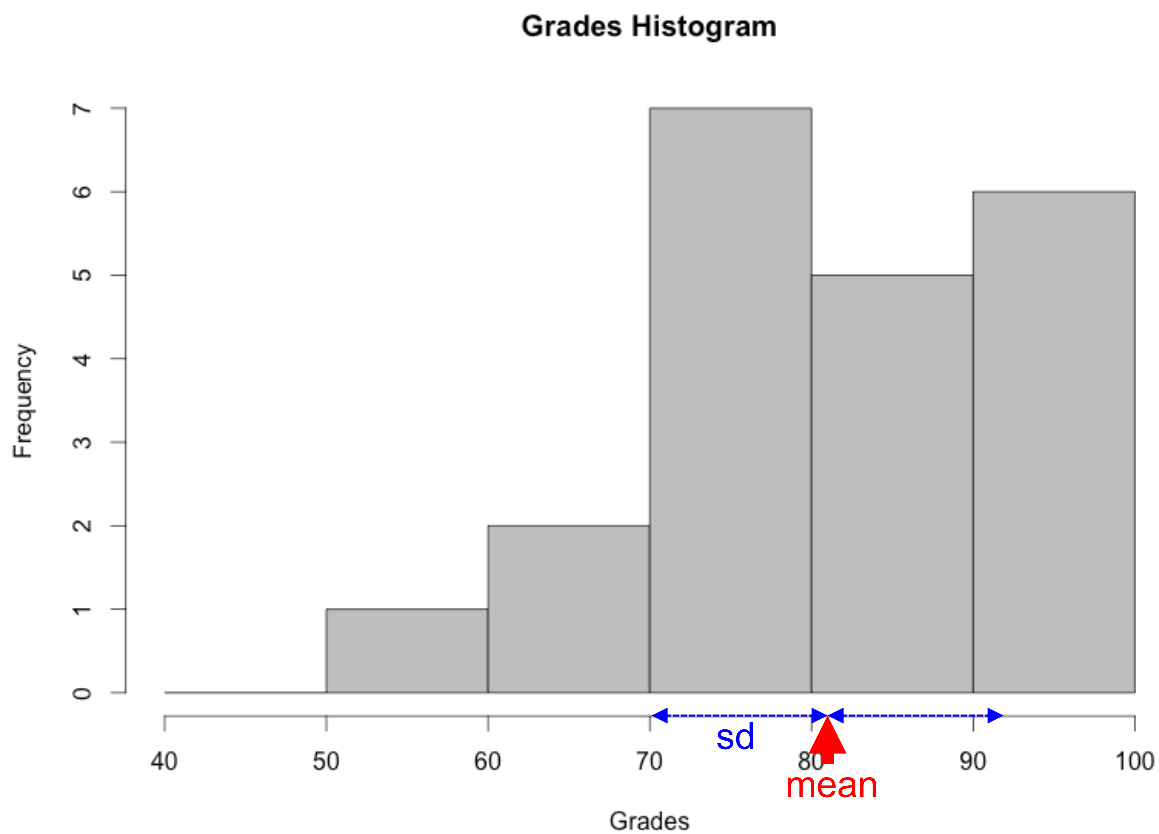


Measures of central tendency and variability

```
> ex.d
 Id Grades
  1     84
  2     69
  3     98
  4     92
  5     94
  6     54
  7     76
  8     91
  9     80
 10     88
 11     79
 12     81
 13     83
 14     79
 15     78
 16     74
 17     83
 18     92
 19     74
 20     98
 21     61
```

```
> mean(ex.d$Grades)
[1] 81.33333

> sd(ex.d$Grades)
[1] 11.32843
```



Median, Quartiles and IQR

Median - the score that divides the set of scores in half; 50% of the scores fall below a median and 50% of the scores fall above a median.

Quartiles:

- **1st quartile** - the score that divides the set of scores in 75% highest and 25% lowest scores
- **3rd quartile** - the score that divides the set of scores in 25% highest and 75% lowest scores

Interquartile range (IQR) - the difference between the 1st and 3rd quartiles.

The median is independent of extreme values, while the mean can be greatly affected by extremes.

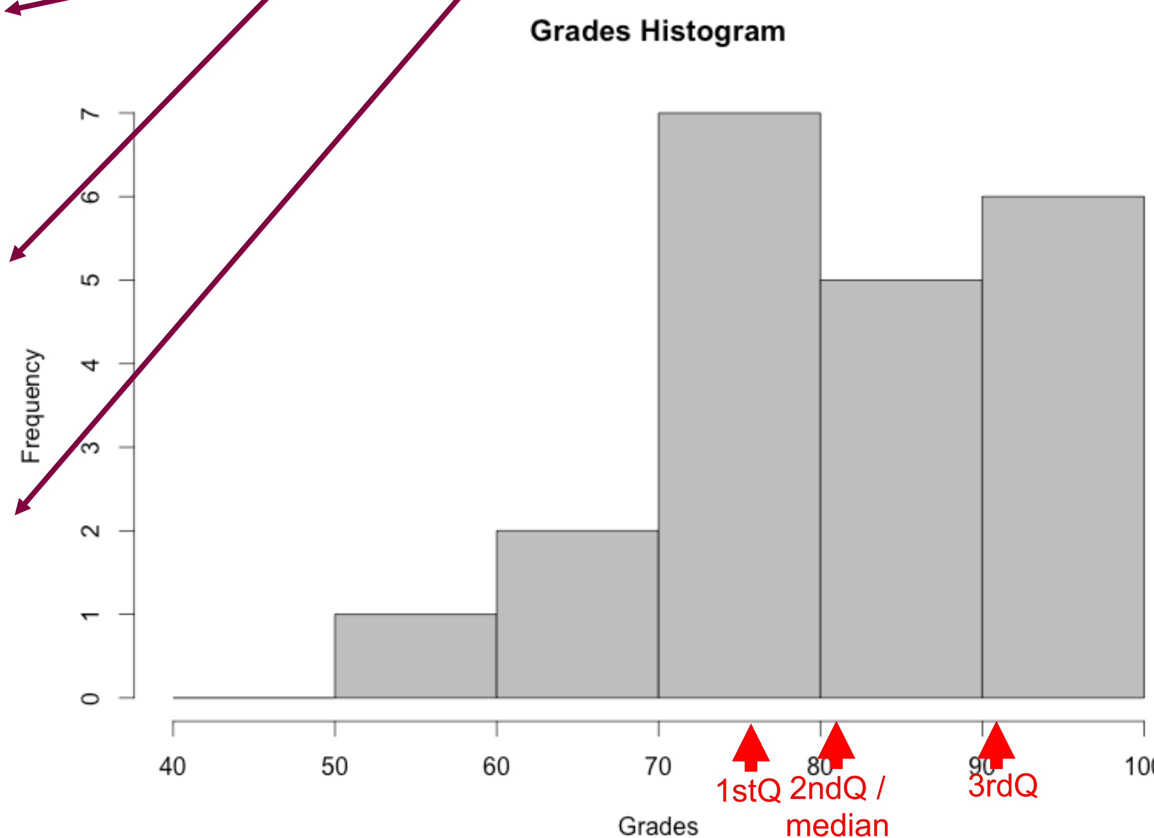
Examples

```
> ex.d
 Id Grades
  1     84
  2     69
  3     98
  4     92
  5     94
  6     54
  7     76
  8     91
  9     80
 10     88
 11     79
 12     81
 13     83
 14     79
 15     78
 16     74
 17     83
 18     92
 19     74
 20     98
 21     61
```

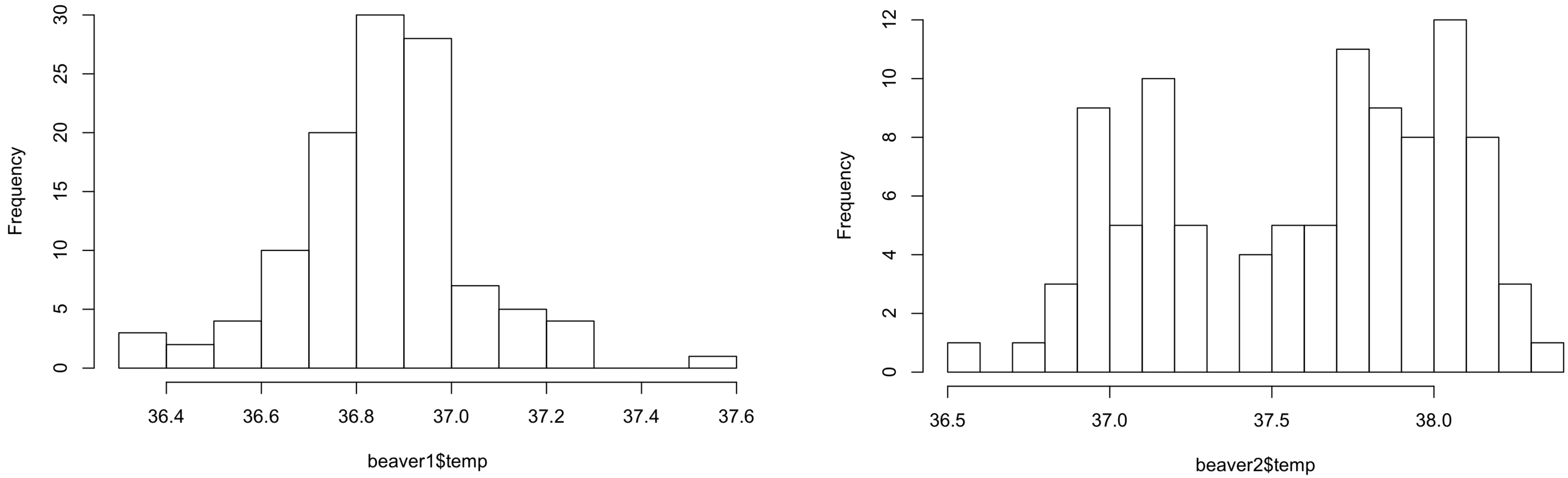
```
> cbind(sort(ex.d$Grades))
      [,1]
 [1,]   54
 [2,]   61
 [3,]   69
 [4,]   74
 [5,]   74
 [6,]   76
 [7,]   78
 [8,]   79
 [9,]   79
[10,]   80
[11,]   81
[12,]   83
[13,]   83
[14,]   84
[15,]   88
[16,]   91
[17,]   92
[18,]   92
[19,]   94
[20,]   98
[21,]   98
```

```
> median(ex.d$Grades)
[1] 81
```

```
> IQR(ex.d$Grades)
[1] 15
```

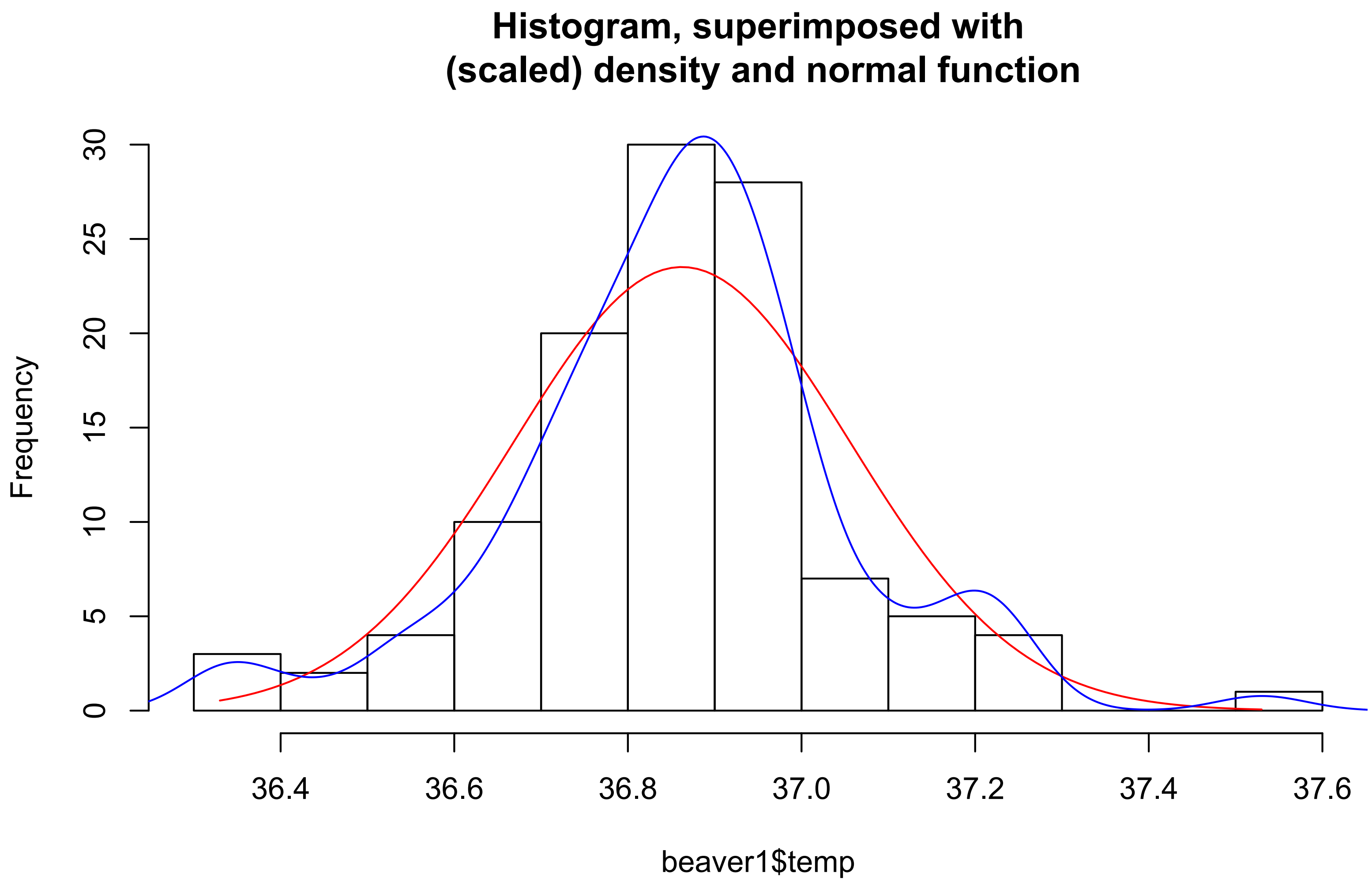


Is it a normal distribution?

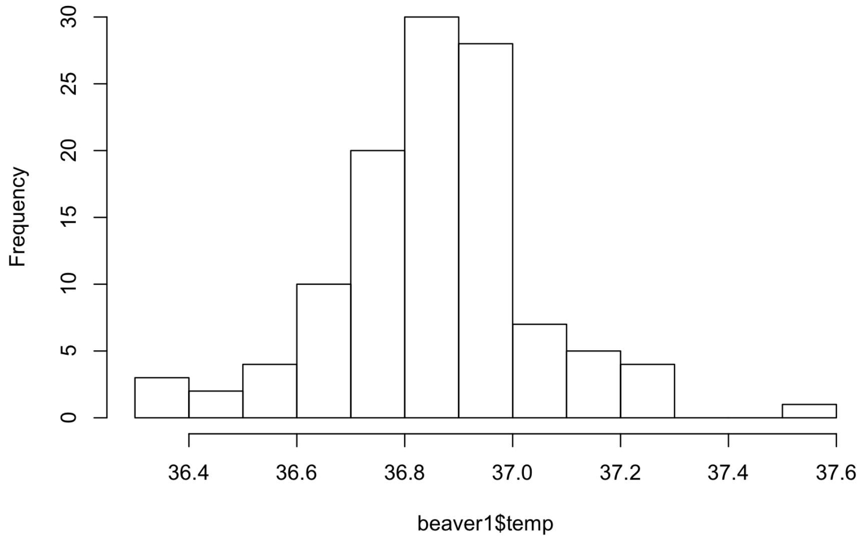


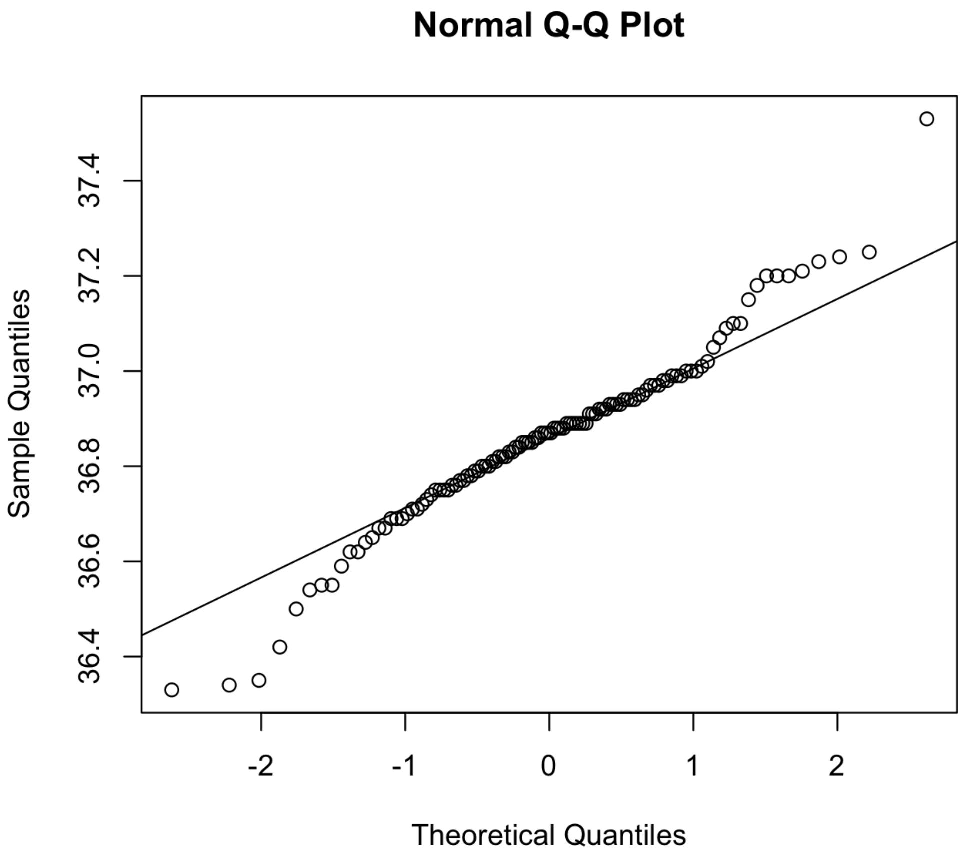
Beaver1’s temperature looks more like a normal distribution than beaver2’s temperature – but can we quantify this?

Superimposed normal curve (red) and kernel density function (blue)



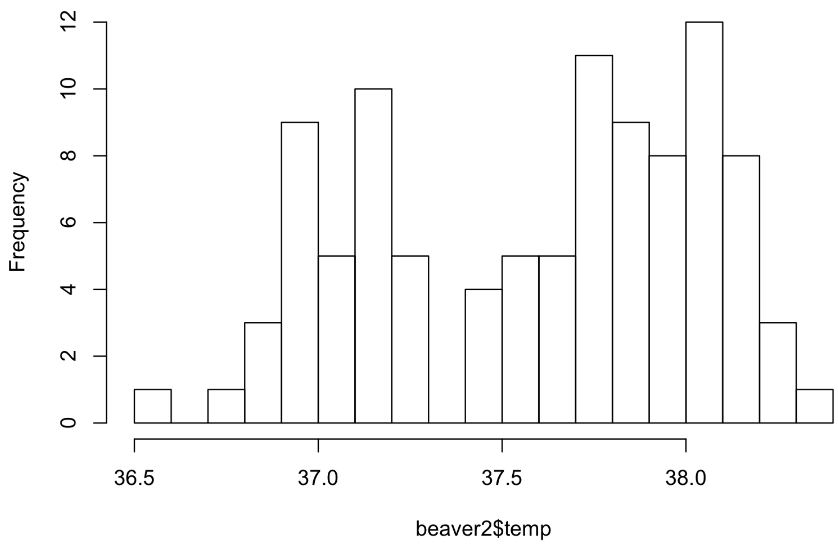
Quantile-quantile plots (Q-Q plots)

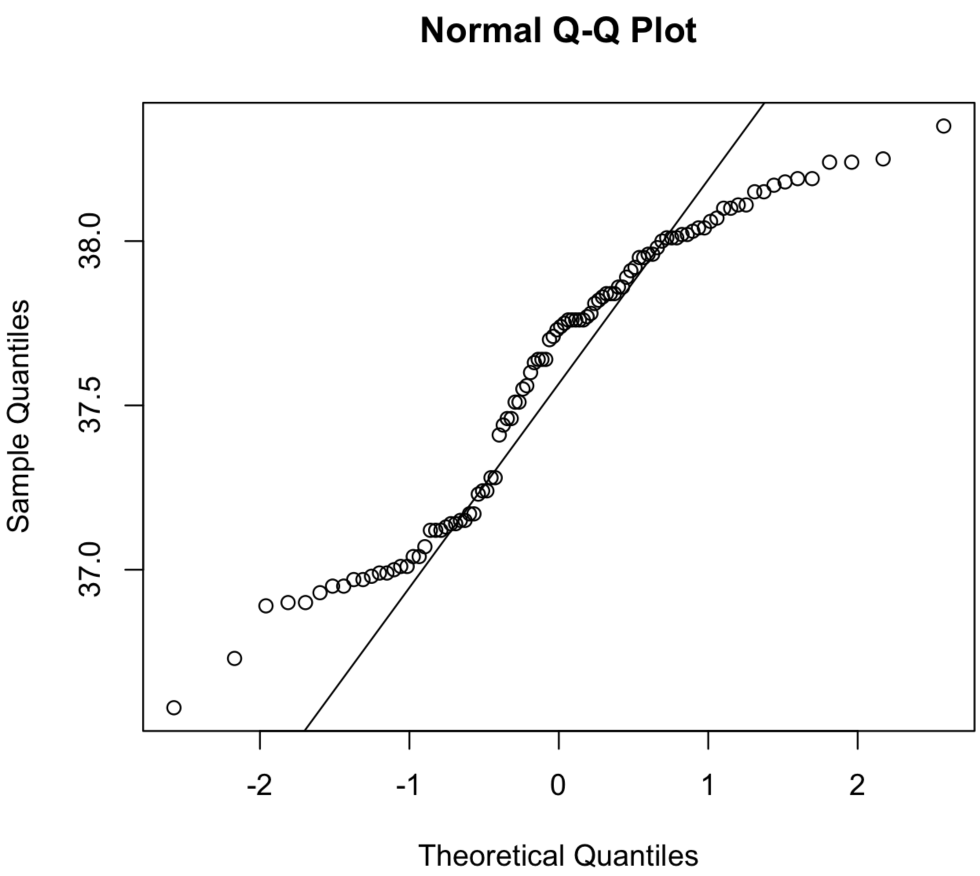




For a perfectly normal distribution, all datapoints would lie on the line. Here, we have temp that's too high, and a few measurements that are lower than expected.

Is my distribution normal?





No. (beaver with fever?)

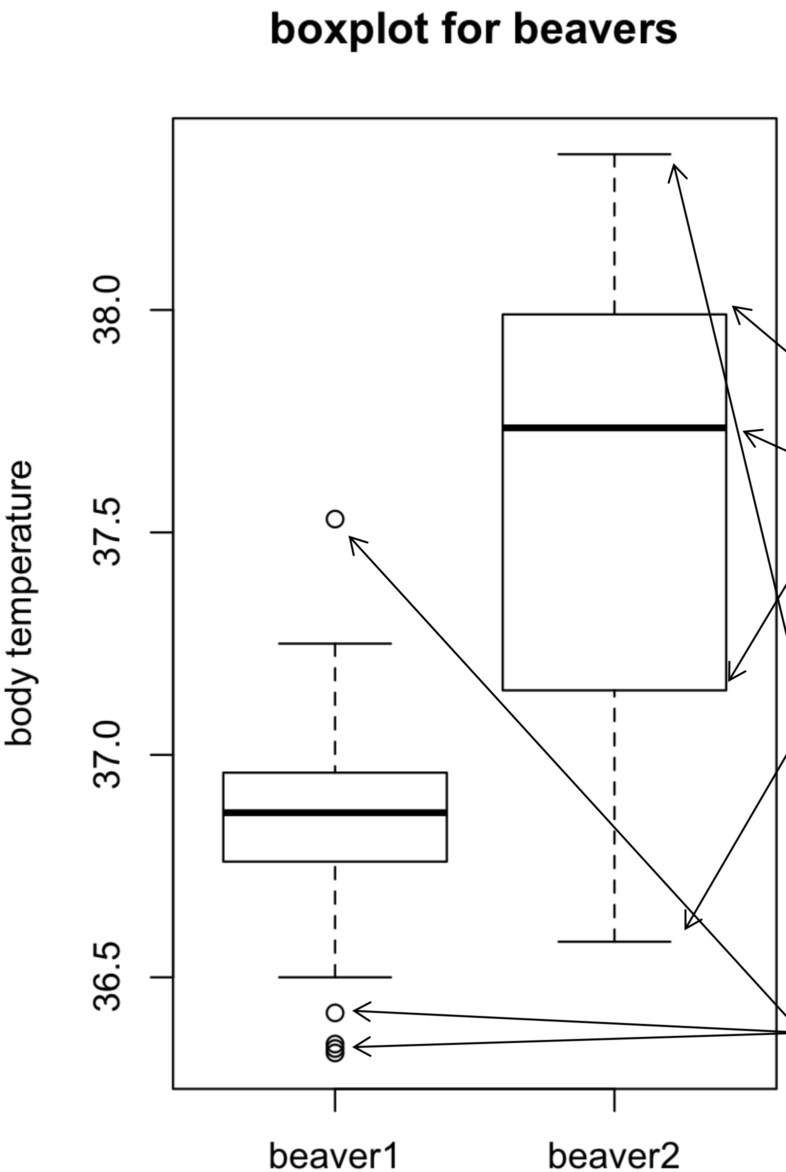
Percentiles

10% Percentile: the score that divides the set of scores into 10% lowest and 90% highest.

114/10=11.4; average scores between 11th and 12th score: 36.63

```
> sort(beaver1$temp)
  [1] 36.33 36.34 36.35 36.42 36.50 36.54 36.55 36.55 36.59 36.62
 [11] 36.62 36.64 36.65 36.67 36.67 36.69 36.69 36.69 36.70 36.71
 [21] 36.71 36.72 36.73 36.74 36.75 36.75 36.75 36.75 36.76 36.76
 [31] 36.77 36.77 36.78 36.78 36.79 36.79 36.80 36.80 36.80 36.81
 [41] 36.81 36.82 36.82 36.82 36.83 36.83 36.84 36.84 36.85 36.85
 [51] 36.85 36.85 36.86 36.86 36.87 36.87 36.87 36.87 36.88 36.88
 [61] 36.88 36.88 36.89 36.89 36.89 36.89 36.89 36.89 36.89 36.91
 [71] 36.91 36.91 36.92 36.92 36.92 36.93 36.93 36.93 36.93 36.94
 [81] 36.94 36.94 36.94 36.95 36.95 36.96 36.97 36.97 36.97 36.98
 [91] 36.98 36.99 36.99 36.99 37.00 37.00 37.00 37.01 37.02 37.05
[101] 37.07 37.09 37.10 37.10 37.15 37.18 37.20 37.20 37.20 37.21
[111] 37.23 37.24 37.25 37.53
```

Interpreting a Boxplot with whiskers



third quartile

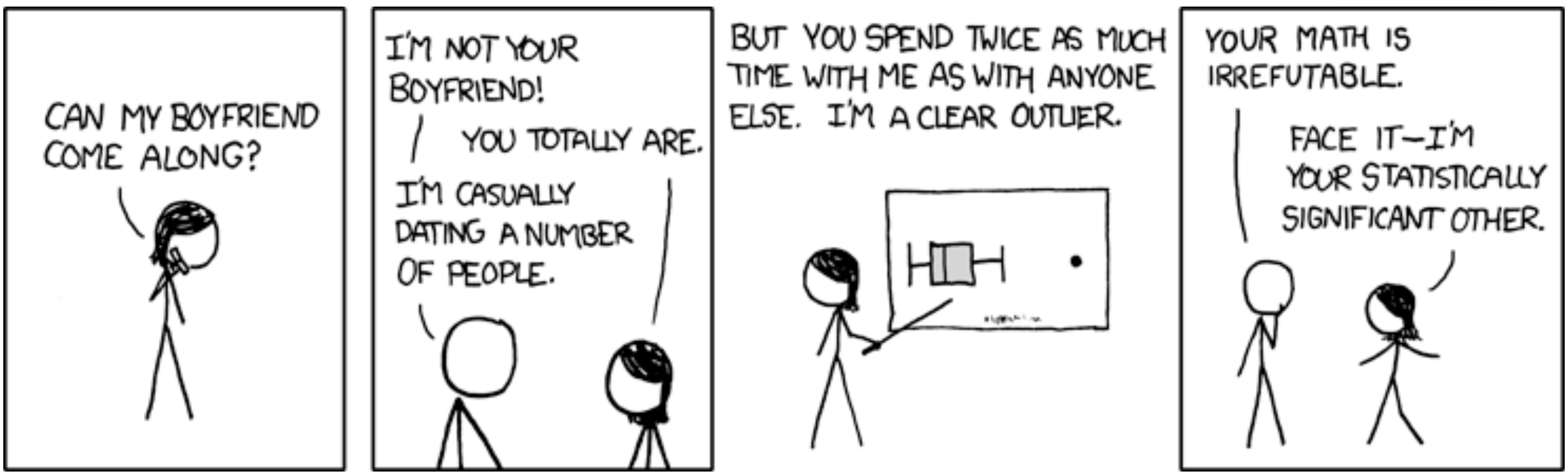
median

first quartile

most extreme points *inside* 1.5 times inter-quartile range

datapoints outside range

boxplot with whiskers



Two more measures of central tendency and variability

Mode and range:

Mode – the most frequent score.

Range – the smallest and the biggest scores.

Measures of central tendency and variability

```
> ex.d
 Id Grades
  1     84
  2     69
  3     98
  4     92
  5     94
  6     54
  7     76
  8     91
  9     80
 10     88
 11     79
 12     81
 13     83
 14     79
 15     78
 16     74
 17     83
 18     92
 19     74
 20     98
 21     61
```

```
> cbind(sort(ex.d$Grades))
      [,1]
 [1,]   54
 [2,]   61
 [3,]   69
 [4,]   74
 [5,]   74
 [6,]   76
 [7,]   78
 [8,]   79
 [9,]   79
[10,]   80
[11,]   81
[12,]   83
[13,]   83
[14,]   84
[15,]   88
[16,]   91
[17,]   92
[18,]   92
[19,]   94
[20,]   98
[21,]   98
```

```
> library(prettyR)
> Mode(ex.d$Grades)
[1] ">1 mode"

> range(ex.d$Grades)
[1] 54 98
```

Measures of central tendency and variability for categorial data

```
 [1] "Black" "Black" "Black" "Black" "Black" "Black" "Black" "Black" "Black" "Black"
[11] "Black" "Black" "Black" "Black" "Black" "Black" "Black" "Black" "Black" "Black"
[21] "Black" "Black" "Black" "Black" "Black" "Black" "Black" "Black" "Black" "Black"
[31] "Black" "Black" "Brown" "Brown" "Brown" "Brown" "Brown" "Brown" "Brown" "Brown"
[41] "Brown" "Brown" "Brown" "Brown" "Brown" "Brown" "Brown" "Brown" "Brown" "Brown"
[51] "Brown" "Brown" "Brown" "Brown" "Brown" "Brown" "Brown" "Brown" "Brown" "Brown"
[61] "Brown" "Brown" "Brown" "Brown" "Brown" "Brown" "Brown" "Brown" "Brown" "Brown"
[71] "Brown" "Brown" "Brown" "Brown" "Brown" "Brown" "Brown" "Brown" "Brown" "Brown"
[81] "Brown" "Brown" "Brown" "Brown" "Brown" "Red"   "Red"   "Red"   "Red"   "Red"
[91] "Red"   "Red"   "Red"   "Red"   "Red"   "Blond" "Blond" "Blond”
```

```
> library(prettyR)
> Mode(brownMaleHair)
[1] ”Brown"
```

Measures of central tendency and variability

Continuous (quantitative):

- Mean and standard deviation
- Median and interquartile range
- Mode and range

Categorical (discrete):

- Mode

Lecture outline

- Exploring data
  - Visualizing data
  - Measures of central tendency and variability
- Normal distribution and probability
- Sampling distribution of the mean and the Central Limit Theorem

Let’s get back to our beavers...

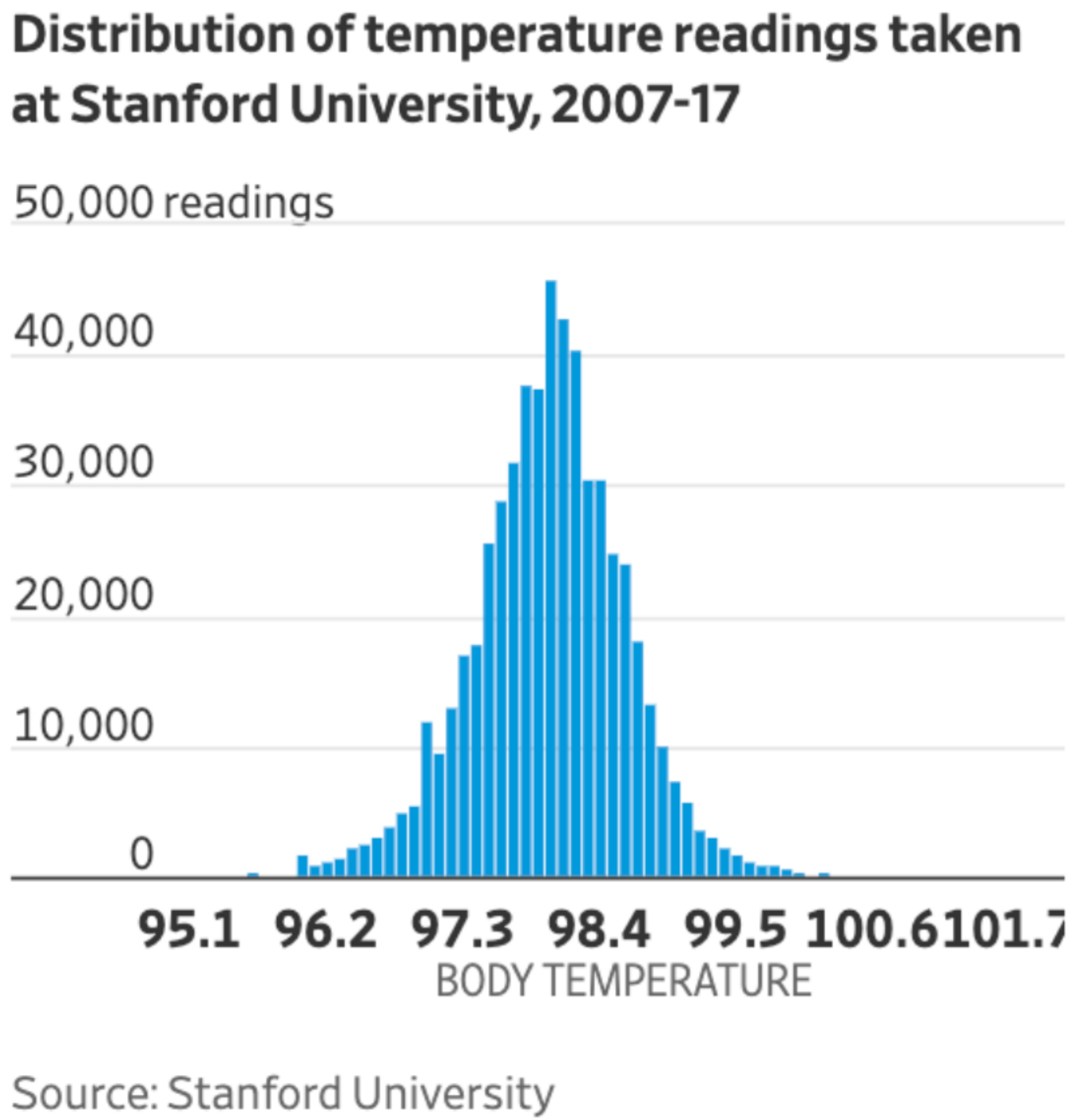
Let’s assume we have beaver body temperature measurements from 1000 beavers, from a whole year. This gives us a pretty good idea of “normal” beaver body temperatures.

Q: Is a body temperature of a beaver on a specific day in the normal range?

We’ll use this example to go through the logics of hypothesis testing.

Body temperatures usually follow a "normal distribution".

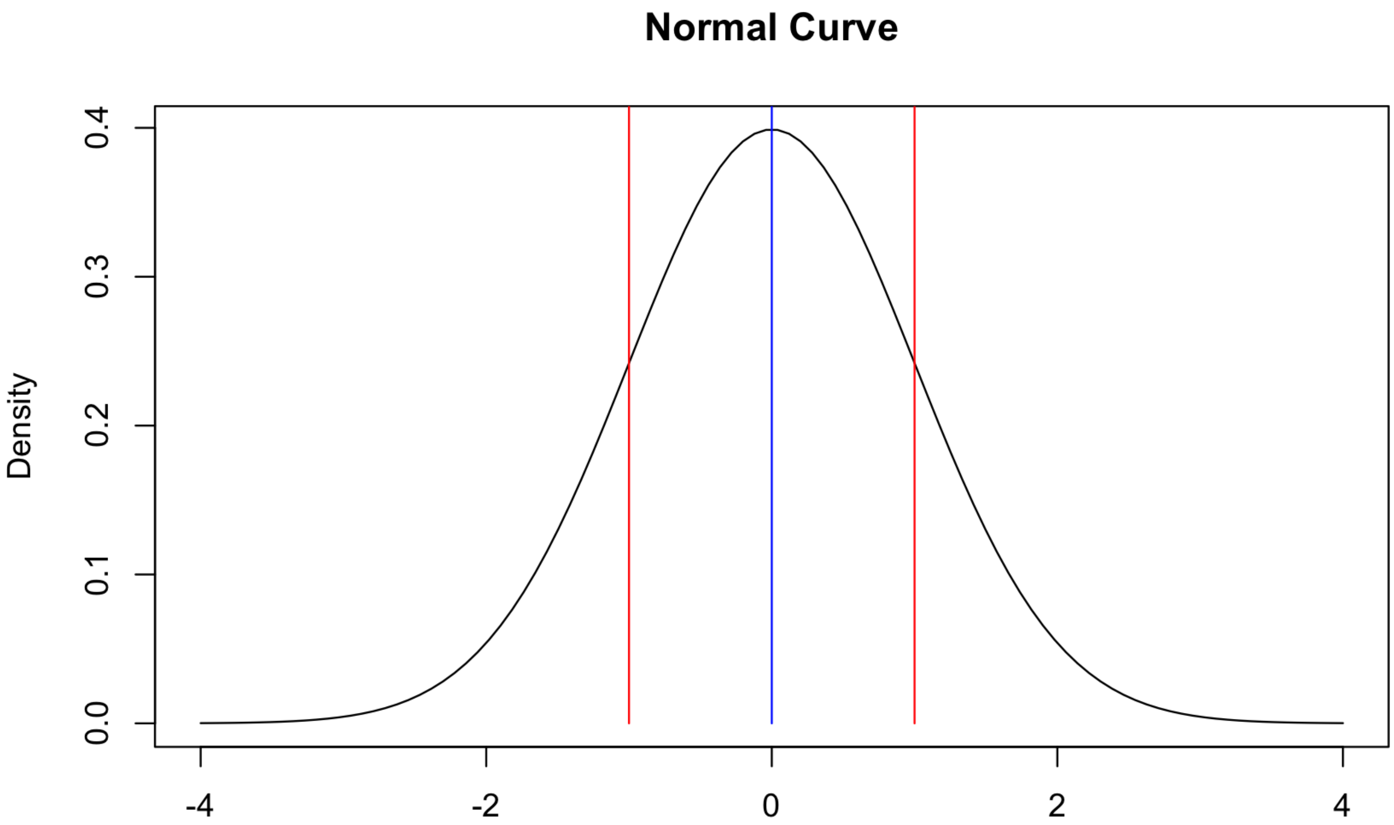
Q: Is a body temperature of a beaver on a specific day in the normal range?



(We assume that beaver body temps are normally distributed:)

Normal distribution

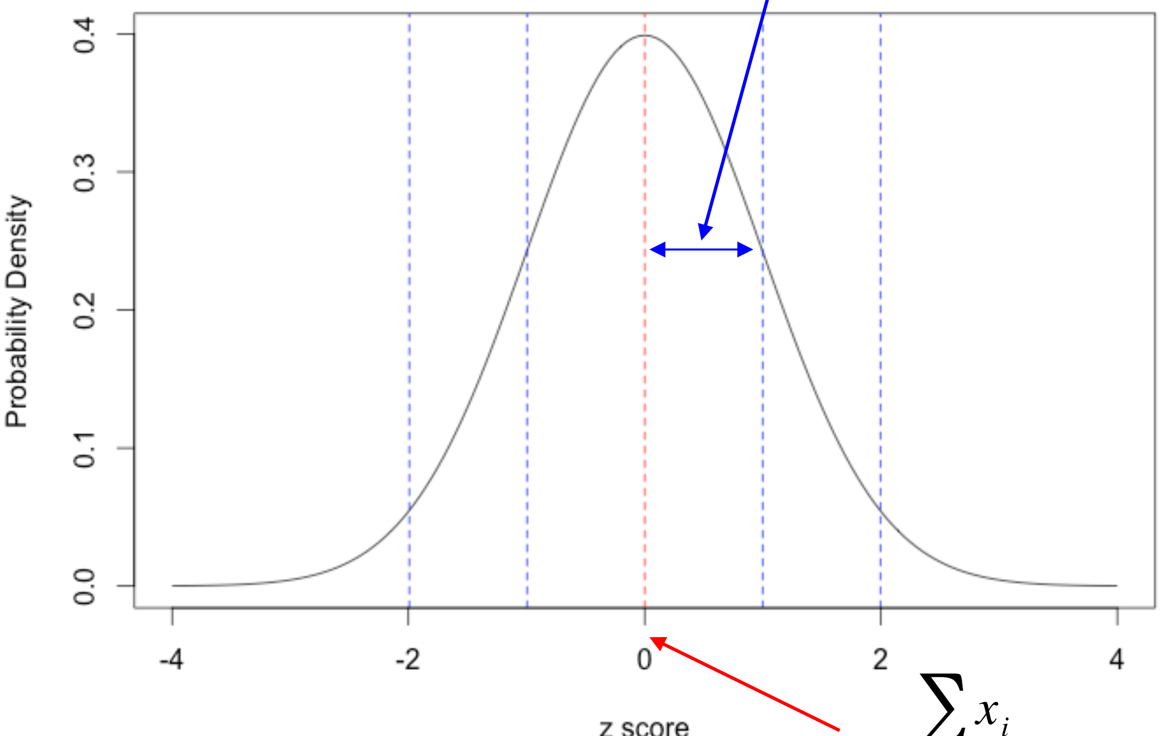
the backbone of traditional parametric statistics



Gauss

Standard normal (z) distribution
z scores

$$\sigma = \sqrt{\frac{\sum (x_i - \mu)^2}{N}}$$



$$\mu = \frac{\sum x_i}{N}$$

$$z = \frac{x - \mu}{\sigma}$$

x – *individual score*

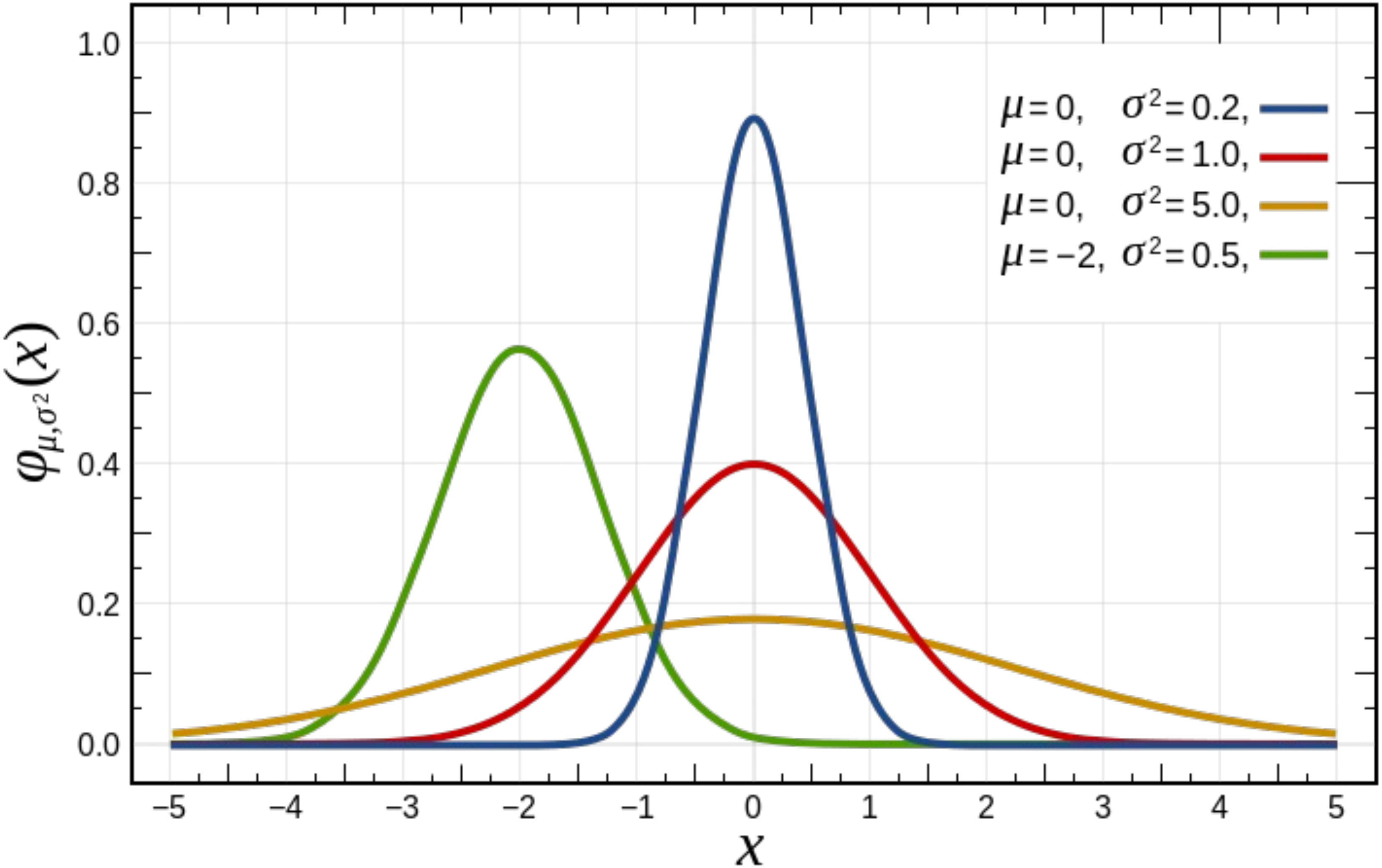
μ – *mean in a population*

σ – *standard deviation in a population*

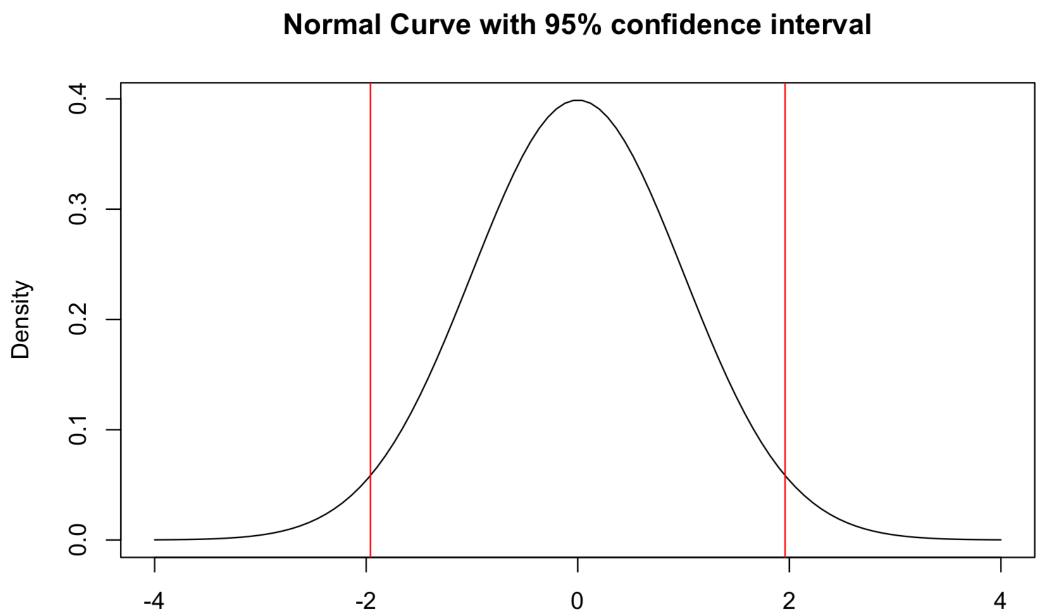
N – *population size*

z value of a score represents how much that score is different from the mean in terms of standard deviations estimated from a population

Normal distribution can fit to our data by changing mean and standard deviation



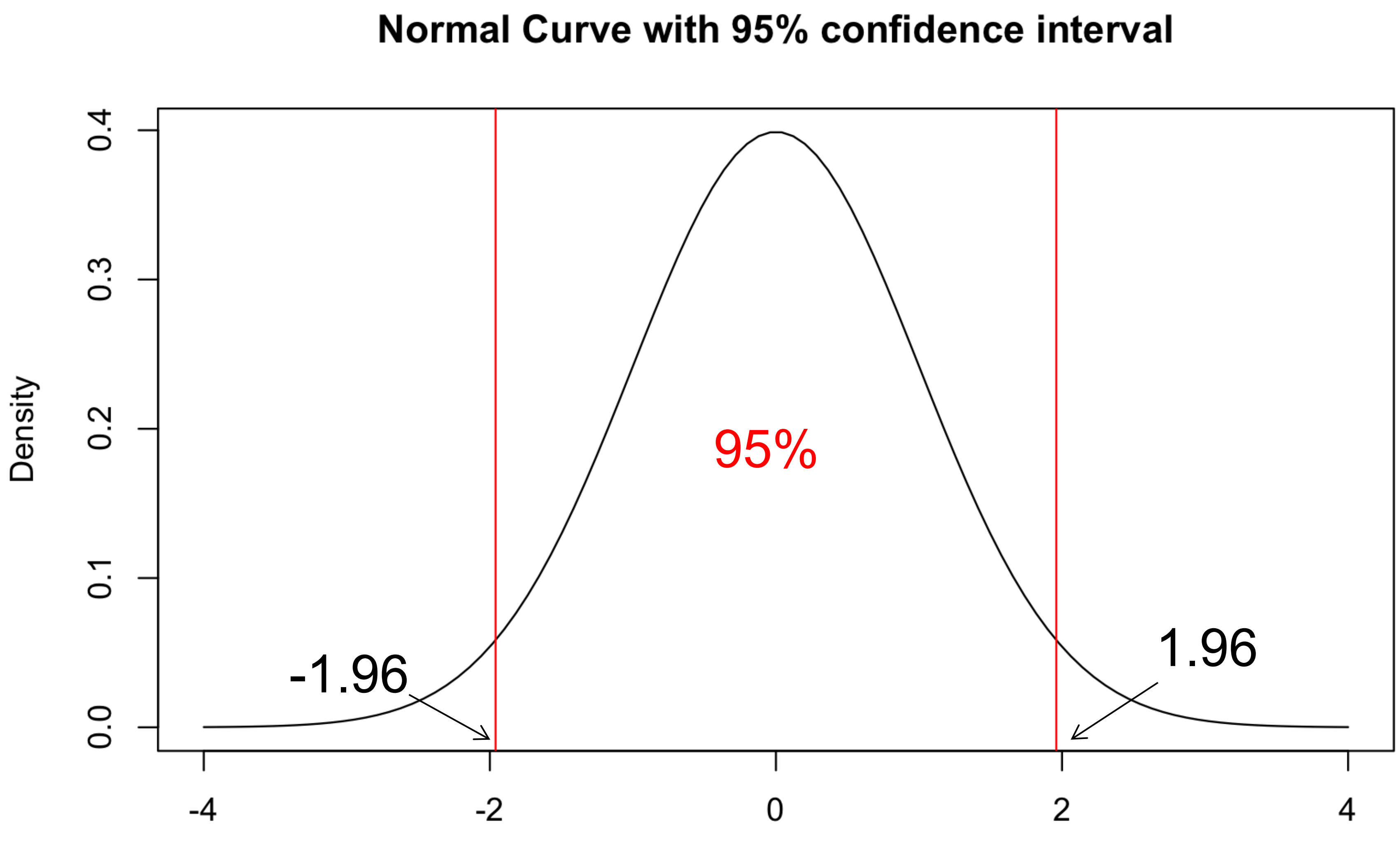
Normal distribution



Need to calculate the area under the curve to find out for any values of x what probabilities they correspond to.

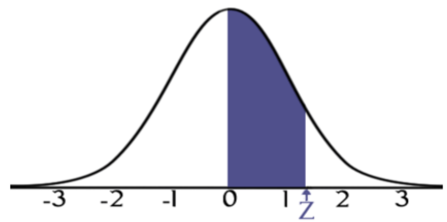
We won't do this by hand – either have stats program do it for you or look it up in a z-score table.

Normal distribution



Normal distribution

lookup table



Standard Normal Table (z)

Entries in the table give the area under the curve between the mean and z standard deviations above the mean. For example, for $z = 1.25$ the area under the curve between the mean (0) and z is 0.3944.

z	0.00	0.01	0.02	0.03	0.04	0.05	0.06	0.07	0.08	0.09
0.0	0.0000	0.0040	0.0080	0.0120	0.0160	0.0190	0.0239	0.0279	0.0319	0.0359
0.1	0.0398	0.0438	0.0478	0.0517	0.0557	0.0596	0.0636	0.0675	0.0714	0.0753
0.2	0.0793	0.0832	0.0871	0.0910	0.0948	0.0987	0.1026	0.1064	0.1103	0.1141
0.3	0.1179	0.1217	0.1255	0.1293	0.1331	0.1368	0.1406	0.1443	0.1480	0.1517
0.4	0.1554	0.1591	0.1628	0.1664	0.1700	0.1736	0.1772	0.1808	0.1844	0.1879
0.5	0.1915	0.1950	0.1985	0.2019	0.2054	0.2088	0.2123	0.2157	0.2190	0.2224
0.6	0.2257	0.2291	0.2324	0.2357	0.2389	0.2422	0.2454	0.2486	0.2517	0.2549
0.7	0.2580	0.2611	0.2642	0.2673	0.2704	0.2734	0.2764	0.2794	0.2823	0.2852
0.8	0.2881	0.2910	0.2939	0.2969	0.2995	0.3023	0.3051	0.3078	0.3106	0.3133
0.9	0.3159	0.3186	0.3212	0.3238	0.3264	0.3289	0.3315	0.3340	0.3365	0.3389
1.0	0.3413	0.3438	0.3461	0.3485	0.3508	0.3513	0.3554	0.3577	0.3529	0.3621
1.1	0.3643	0.3665	0.3686	0.3708	0.3729	0.3749	0.3770	0.3790	0.3810	0.3830
1.2	0.3849	0.3869	0.3888	0.3907	0.3925	0.3944	0.3962	0.3980	0.3997	0.4015
1.3	0.4032	0.4049	0.4066	0.4082	0.4099	0.4115	0.4131	0.4147	0.4162	0.4177
1.4	0.4192	0.4207	0.4222	0.4236	0.4251	0.4265	0.4279	0.4292	0.4306	0.4319
1.5	0.4332	0.4345	0.4357	0.4370	0.4382	0.4394	0.4406	0.4418	0.4429	0.4441
1.6	0.4452	0.4463	0.4474	0.4484	0.4495	0.4505	0.4515	0.4525	0.4535	0.4545
1.7	0.4554	0.4564	0.4573	0.4582	0.4591	0.4599	0.4608	0.4616	0.4625	0.4633
1.8	0.4641	0.4649	0.4656	0.4664	0.4671	0.4678	0.4686	0.4693	0.4699	0.4706
1.9	0.4713	0.4719	0.4726	0.4732	0.4738	0.4744	0.4750	0.4756	0.4761	0.4767
2.0	0.4772	0.4778	0.4783	0.4788	0.4793	0.4798	0.4803	0.4808	0.4812	0.4817
2.1	0.4821	0.4826	0.4830	0.4834	0.4838	0.4842	0.4846	0.4850	0.4854	0.4857
2.2	0.4861	0.4864	0.4868	0.4871	0.4875	0.4878	0.4881	0.4884	0.4887	0.4890
2.3	0.4893	0.4896	0.4898	0.4901	0.4904	0.4906	0.4909	0.4911	0.4913	0.4916
2.4	0.4918	0.4920	0.4922	0.4925	0.4927	0.4929	0.4931	0.4932	0.4934	0.4936
2.5	0.4938	0.4940	0.4941	0.4943	0.4945	0.4946	0.4948	0.4949	0.4951	0.4952
2.6	0.4953	0.4955	0.4956	0.4957	0.4959	0.4960	0.4961	0.4962	0.4963	0.4964
2.7	0.4965	0.4966	0.4967	0.4968	0.4969	0.4970	0.4971	0.4972	0.4973	0.4974
2.8	0.4974	0.4975	0.4976	0.4977	0.4977	0.4978	0.4979	0.4979	0.4980	0.4981
2.9	0.4981	0.4982	0.4982	0.4983	0.4984	0.4984	0.4985	0.4985	0.4986	0.4986
3.0	0.4987	0.4987	0.4987	0.4988	0.4988	0.4989	0.4989	0.4989	0.4990	0.4990
3.1	0.4990	0.4991	0.4991	0.4991	0.4992	0.4992	0.4992	0.4992	0.4993	0.4993
3.2	0.4993	0.4993	0.4994	0.4994	0.4994	0.4994	0.4994	0.4995	0.4995	0.4995
3.3	0.4995	0.4995	0.4995	0.4996	0.4996	0.4996	0.4996	0.4996	0.4996	0.4997
3.4	0.4997	0.4997	0.4997	0.4997	0.4997	0.4997	0.4997	0.4997	0.4997	0.4998

What is the probability that a random student from a large class with mean grade 81.3 and standard deviation 11.3 will get a grade equal to or above 95?

=

What is the area under the normal curve for z values greater than (95-81.3)/11.3?

Normal distribution lookup table

… or look up

z = (95-81.3)/11.3

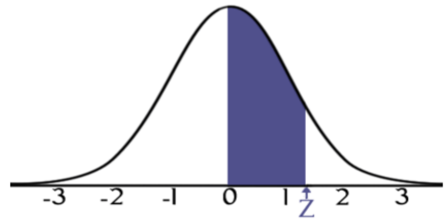
= 1.212389 in table

⇒0.3869

+ 0.5

= 0.8869

1 - 0.8869 = 0.1131



STANDARD NORMAL TABLE (Z)

Entries in the table give the area under the curve between the mean and z standard deviations above the mean. For example, for $z = 1.25$ the area under the curve between the mean (0) and z is 0.3944.

z	0.00	0.01	0.02	0.03	0.04	0.05	0.06	0.07	0.08	0.09
0.0	0.0000	0.0040	0.0080	0.0120	0.0160	0.0190	0.0239	0.0279	0.0319	0.0359
0.1	0.0398	0.0438	0.0478	0.0517	0.0557	0.0596	0.0636	0.0675	0.0714	0.0753
0.2	0.0793	0.0832	0.0871	0.0910	0.0948	0.0987	0.1026	0.1064	0.1103	0.1141
0.3	0.1179	0.1217	0.1255	0.1293	0.1331	0.1368	0.1406	0.1443	0.1480	0.1517
0.4	0.1554	0.1591	0.1628	0.1664	0.1700	0.1736	0.1772	0.1808	0.1844	0.1879
0.5	0.1915	0.1950	0.1985	0.2019	0.2054	0.2088	0.2123	0.2157	0.2190	0.2224
0.6	0.2257	0.2291	0.2324	0.2357	0.2389	0.2422	0.2454	0.2486	0.2517	0.2549
0.7	0.2580	0.2611	0.2642	0.2673	0.2704	0.2734	0.2764	0.2794	0.2823	0.2852
0.8	0.2881	0.2910	0.2939	0.2969	0.2995	0.3023	0.3051	0.3078	0.3106	0.3133
0.9	0.3159	0.3186	0.3212	0.3238	0.3264	0.3289	0.3315	0.3340	0.3365	0.3389
1.0	0.3413	0.3438	0.3461	0.3485	0.3508	0.3513	0.3554	0.3577	0.3529	0.3621
1.1	0.3643	0.3665	0.3686	0.3708	0.3729	0.3749	0.3770	0.3790	0.3810	0.3830
1.2	0.3849	0.3869	0.3888	0.3907	0.3925	0.3944	0.3962	0.3980	0.3997	0.4015
1.3	0.4032	0.4049	0.4066	0.4082	0.4099	0.4115	0.4131	0.4147	0.4162	0.4177
1.4	0.4192	0.4207	0.4222	0.4236	0.4251	0.4265	0.4279	0.4292	0.4306	0.4319
1.5	0.4332	0.4345	0.4357	0.4370	0.4382	0.4394	0.4406	0.4418	0.4429	0.4441
1.6	0.4452	0.4463	0.4474	0.4484	0.4495	0.4505	0.4515	0.4525	0.4535	0.4545
1.7	0.4554	0.4564	0.4573	0.4582	0.4591	0.4599	0.4608	0.4616	0.4625	0.4633
1.8	0.4641	0.4649	0.4656	0.4664	0.4671	0.4678	0.4686	0.4693	0.4699	0.4706
1.9	0.4713	0.4719	0.4726	0.4732	0.4738	0.4744	0.4750	0.4756	0.4761	0.4767
2.0	0.4772	0.4778	0.4783	0.4788	0.4793	0.4798	0.4803	0.4808	0.4812	0.4817
2.1	0.4821	0.4826	0.4830	0.4834	0.4838	0.4842	0.4846	0.4850	0.4854	0.4857
2.2	0.4861	0.4864	0.4868	0.4871	0.4875	0.4878	0.4881	0.4884	0.4887	0.4890
2.3	0.4893	0.4896	0.4898	0.4901	0.4904	0.4906	0.4909	0.4911	0.4913	0.4916
2.4	0.4918	0.4920	0.4922	0.4925	0.4927	0.4929	0.4931	0.4932	0.4934	0.4936
2.5	0.4938	0.4940	0.4941	0.4943	0.4945	0.4946	0.4948	0.4949	0.4951	0.4952
2.6	0.4953	0.4955	0.4956	0.4957	0.4959	0.4960	0.4961	0.4962	0.4963	0.4964
2.7	0.4965	0.4966	0.4967	0.4968	0.4969	0.4970	0.4971	0.4972	0.4973	0.4974
2.8	0.4974	0.4975	0.4976	0.4977	0.4977	0.4978	0.4979	0.4979	0.4980	0.4981
2.9	0.4981	0.4982	0.4982	0.4983	0.4984	0.4984	0.4985	0.4985	0.4986	0.4986
3.0	0.4987	0.4987	0.4987	0.4988	0.4988	0.4989	0.4989	0.4989	0.4990	0.4990
3.1	0.4990	0.4991	0.4991	0.4991	0.4992	0.4992	0.4992	0.4992	0.4993	0.4993
3.2	0.4993	0.4993	0.4994	0.4994	0.4994	0.4994	0.4994	0.4995	0.4995	0.4995
3.3	0.4995	0.4995	0.4995	0.4996	0.4996	0.4996	0.4996	0.4996	0.4996	0.4997
3.4	0.4997	0.4997	0.4997	0.4997	0.4997	0.4997	0.4997	0.4997	0.4997	0.4998

Normal distribution in R

The Normal Distribution

Description

Density, distribution function, quantile function and random generation for the normal distribution with mean equal to `mean` and standard deviation equal to `sd`.

Usage

```
dnorm(x, mean = 0, sd = 1, log = FALSE)
pnorm(q, mean = 0, sd = 1, lower.tail = TRUE, log.p = FALSE)
qnorm(p, mean = 0, sd = 1, lower.tail = TRUE, log.p = FALSE)
rnorm(n, mean = 0, sd = 1)
```

Arguments

`x, q`
: vector of quantiles.

`p`
: vector of probabilities.

`n`
: number of observations. If `length(n) > 1`, the length is taken to be the number required.

`mean`
: vector of means.

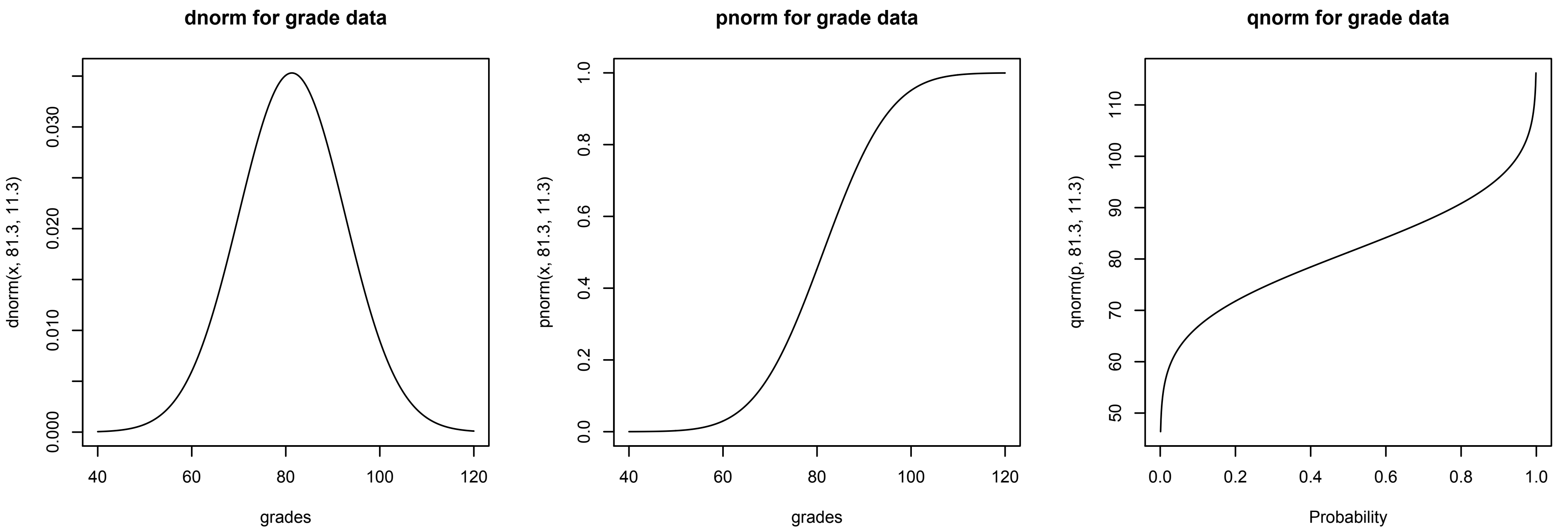
`sd`
: vector of standard deviations.

Normal distribution in R

Usage

```
dnorm(x, mean = 0, sd = 1, log = FALSE)
pnorm(q, mean = 0, sd = 1, lower.tail = TRUE, log.p = FALSE)
qnorm(p, mean = 0, sd = 1, lower.tail = TRUE, log.p = FALSE)
rnorm(n, mean = 0, sd = 1)
```

`dnorm` gives the density, `pnorm` gives the distribution function, `qnorm` gives the quantile function, and `rnorm` generates random deviates.



What is the probability that a random student from a large class with mean grade 81.3 and standard deviation 11.3 will get a grade equal to or above 95? = What is the area under the normal curve for z values greater than (95-81.3)/11.3?

➢ `1 - pnorm((95-81.3)/11.3)`

```
[1] 0.1126817
```

(difference to result with table is due to rounding / limited granularity of table)

Lecture outline

- Exploring data
  - Visualizing data
  - Measures of central tendency and variability
- Normal distribution and probability
- Sampling distribution of the mean and the Central Limit Theorem: let's get a better understanding of sampling.

Sampling Variability

even perfectly **random** (unbiased) samples are subject to sampling variability

If we repeat our experiment

- same conditions
- different sample of same size from same population

We will find a different mean, different standard deviation.
JUST DUE TO CHANCE DURING SAMPLING!

Sampling – Law of large numbers (more is always better)

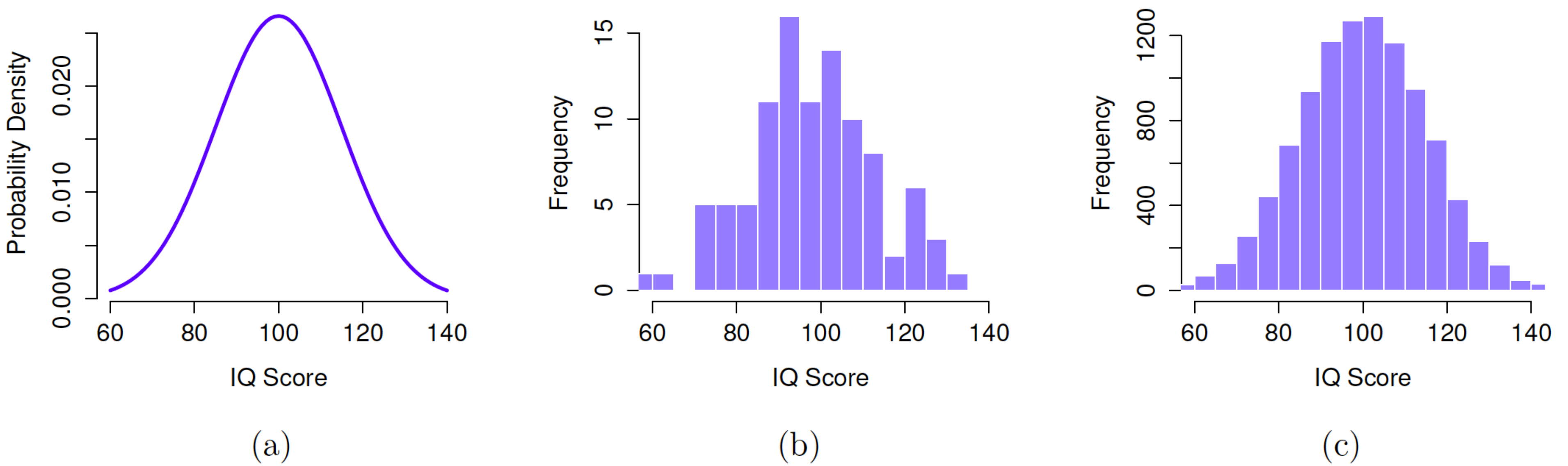


Figure 10.4: The population distribution of IQ scores (panel a) and two samples drawn randomly from it. In panel b we have a sample of 100 observations, and panel c we have a sample of 10,000 observations.

Sampling – Law of large numbers (more is always better)

BUT…

It is not enough to know that we will eventually arrive at the right answer when calculating the sample mean.

Knowing that an infinitely large data set will tell me the exact value of the population mean is **cold comfort when my actual data set has a sample size of N = 100**.

In real life, then, we must know something about the **behaviour of the sample mean** when it is calculated from a more modest data set!

Sampling distributions

Idea:

- We simulate how much variation in the mean we would get for a dataset of a certain size (e.g. N=5)

Table 10.1: Ten replications of the IQ experiment, each with a sample size of $N = 5$.

	Person 1	Person 2	Person 3	Person 4	Person 5	Sample Mean
Replication 1	90	82	94	99	110	95.0
Replication 2	78	88	111	111	117	101.0
Replication 3	111	122	91	98	86	101.6
Replication 4	98	96	119	99	107	103.8
Replication 5	105	113	103	103	98	104.4
Replication 6	81	89	93	85	114	92.4
Replication 7	100	93	108	98	133	106.4
Replication 8	107	100	105	117	85	102.8
Replication 9	86	119	108	73	116	100.4
Replication 10	95	126	112	120	76	105.8

Sampling distribution of the mean.

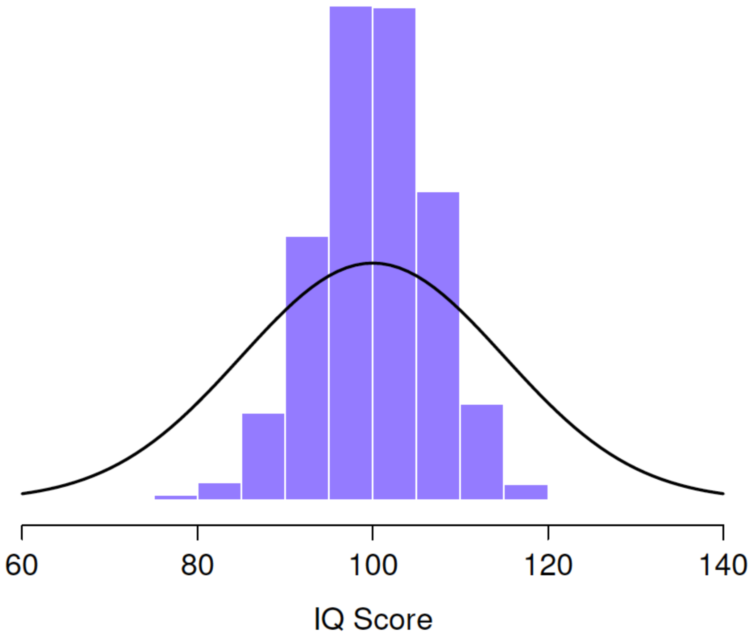
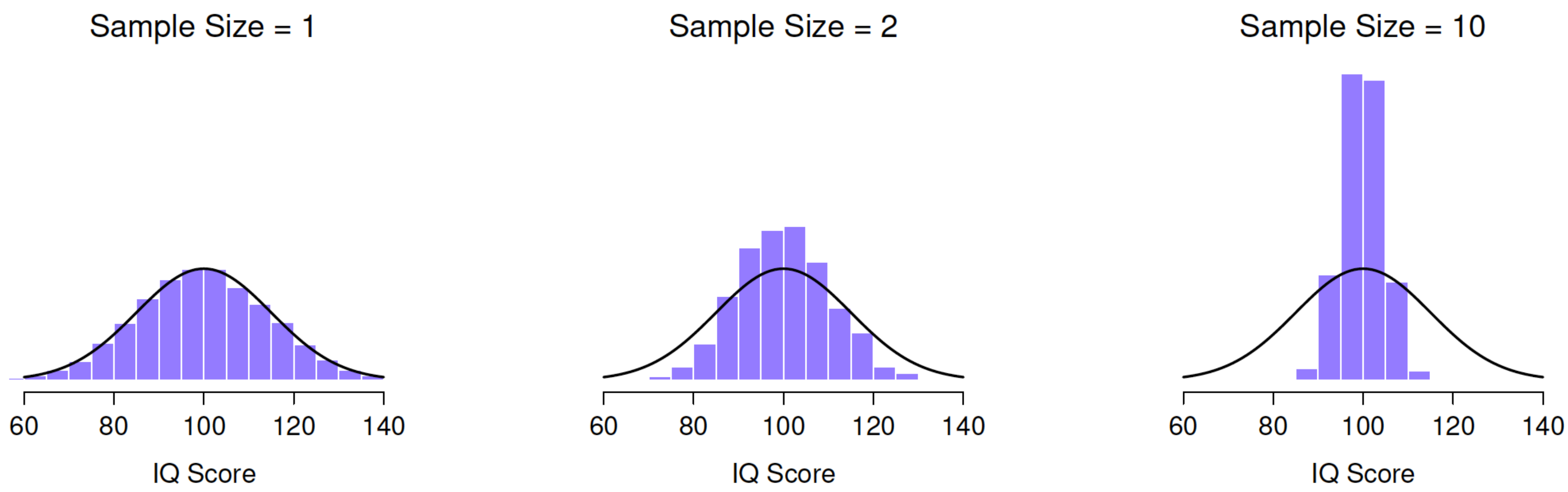


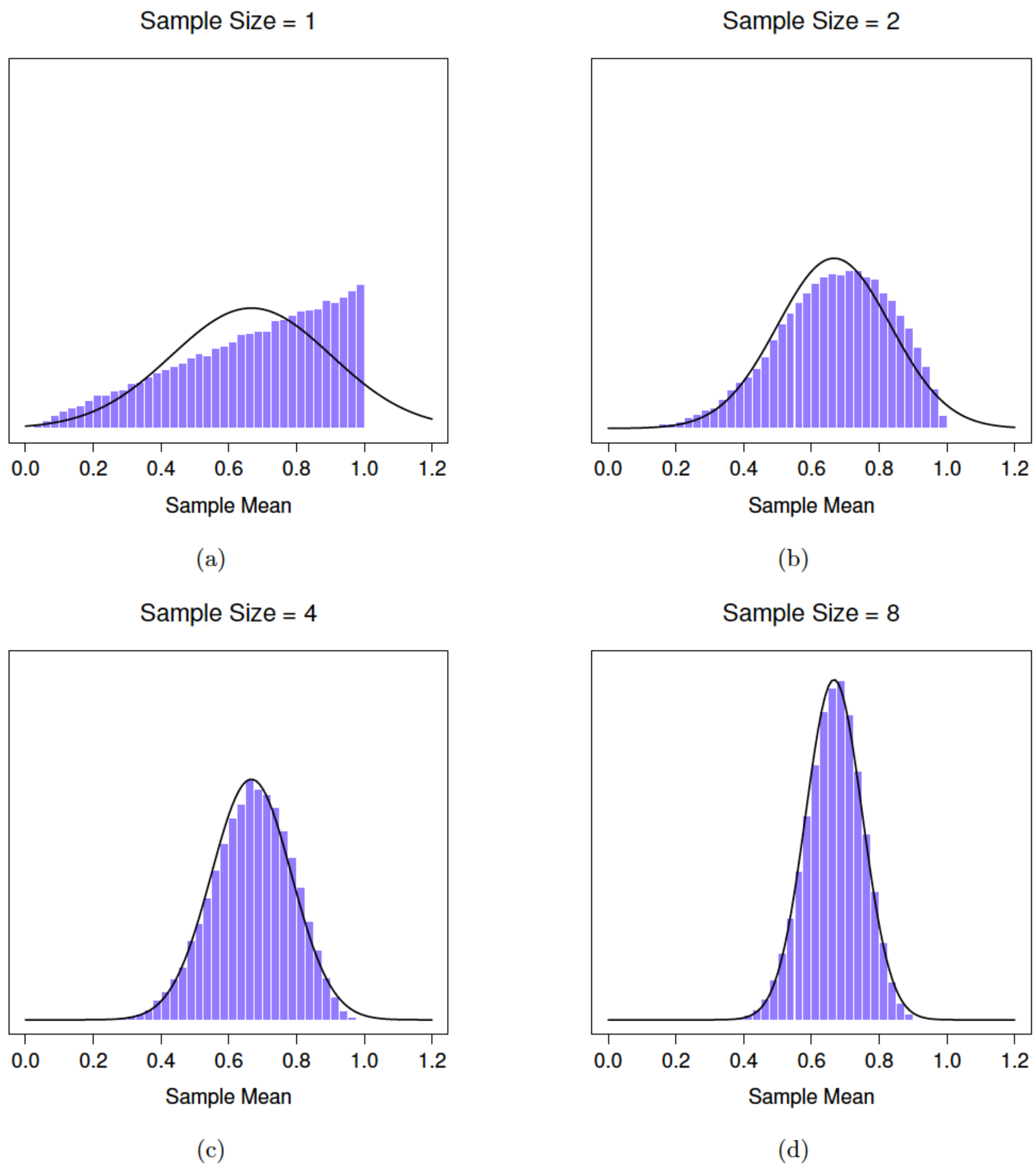
Figure 10.5: The sampling distribution of the mean for the "five IQ scores experiment". If you sample 5 people at random and calculate their *average* IQ, you'll almost certainly get a number between 80 and 120, even though there are quite a lot of individuals who have IQs above 120 or below 80. For comparison, the black line plots the population distribution of IQ scores.

We get less variance with larger samples. The sampling distribution is always normal.



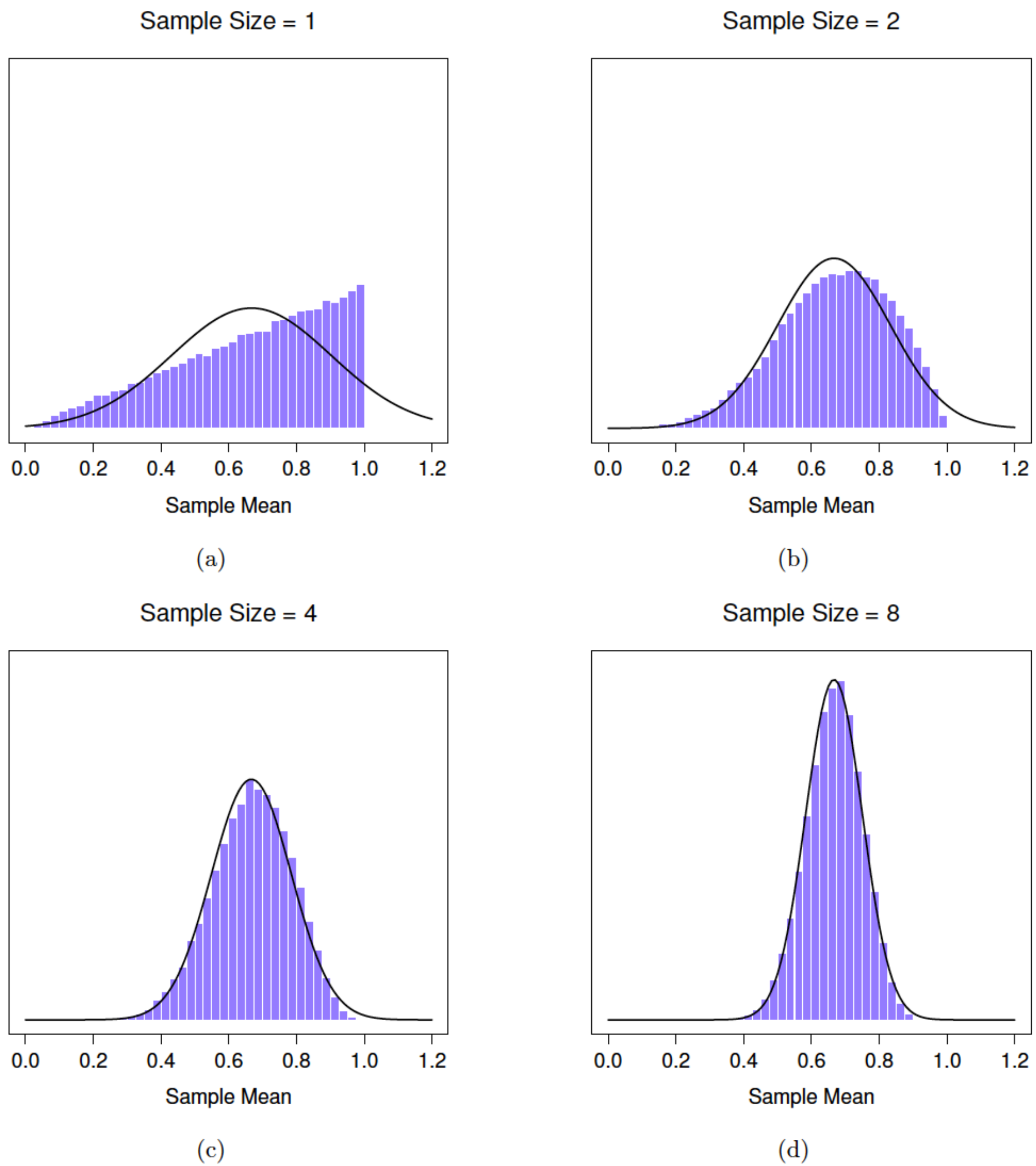
For each plot 10,000 samples of IQ data were generated. Then calculate the mean IQ observed within each of these data sets. The histograms in these plots show the distribution of these means (i.e., the sampling distribution of the mean).

We can also do this for distributions that are themselves not normally distributed.



GOOD NEWS!

As long as your sample size isn’t tiny, the sampling distribution of the mean will be approximately normal, no matter what your population distribution looks like!



Central Limit Theorem

- mean(sampling distribution) = mean(population)

- The **standard deviation** of the sampling distribution (i.e., the *standard error*) gets smaller as the sample size increases.

$$\mathrm{SEM} = \frac{\sigma}{\sqrt{N}}$$

- The shape of the sampling distribution becomes normal as the sample size increases.

Standard error vs. standard deviation

The **standard deviation** describes how much the data points in a sample or population differ from one another.

The **standard error** describes how unsure we are about a parameter (here: the mean).

The standard deviation is used to estimate the standard error of the mean. See also:

https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3148365/

Central Limit Theorem

What has the CLT ever done for us?

- This tells us *why* large experiments are more reliable than small ones.
- it gives us an explicit formula for the standard error, so we can calculate *how much* more reliable a large experiment is.

$$\mathrm{SEM} = \frac{\sigma}{\sqrt{N}}$$

Summary

- Plotting data helps us to see what's going on
  - Histograms or kernel density plots for continuous data
  - Shapes of distributions and identifying outliers
  - Bar charts for visualizing discrete data
- Measures of central tendency and variability
  - Mean and standard deviation
  - Median, quartiles, percentiles; boxplot with whiskers
  - Mode and range
- Normal distribution and probability
  - z scores, Assessing normality
- Central limit theorem:
  - We can sample distributions to find out how certain we can be about our sample mean!

Materials to read:

Materials

- Howell chapter 2
- parts of Navarro chapter 9