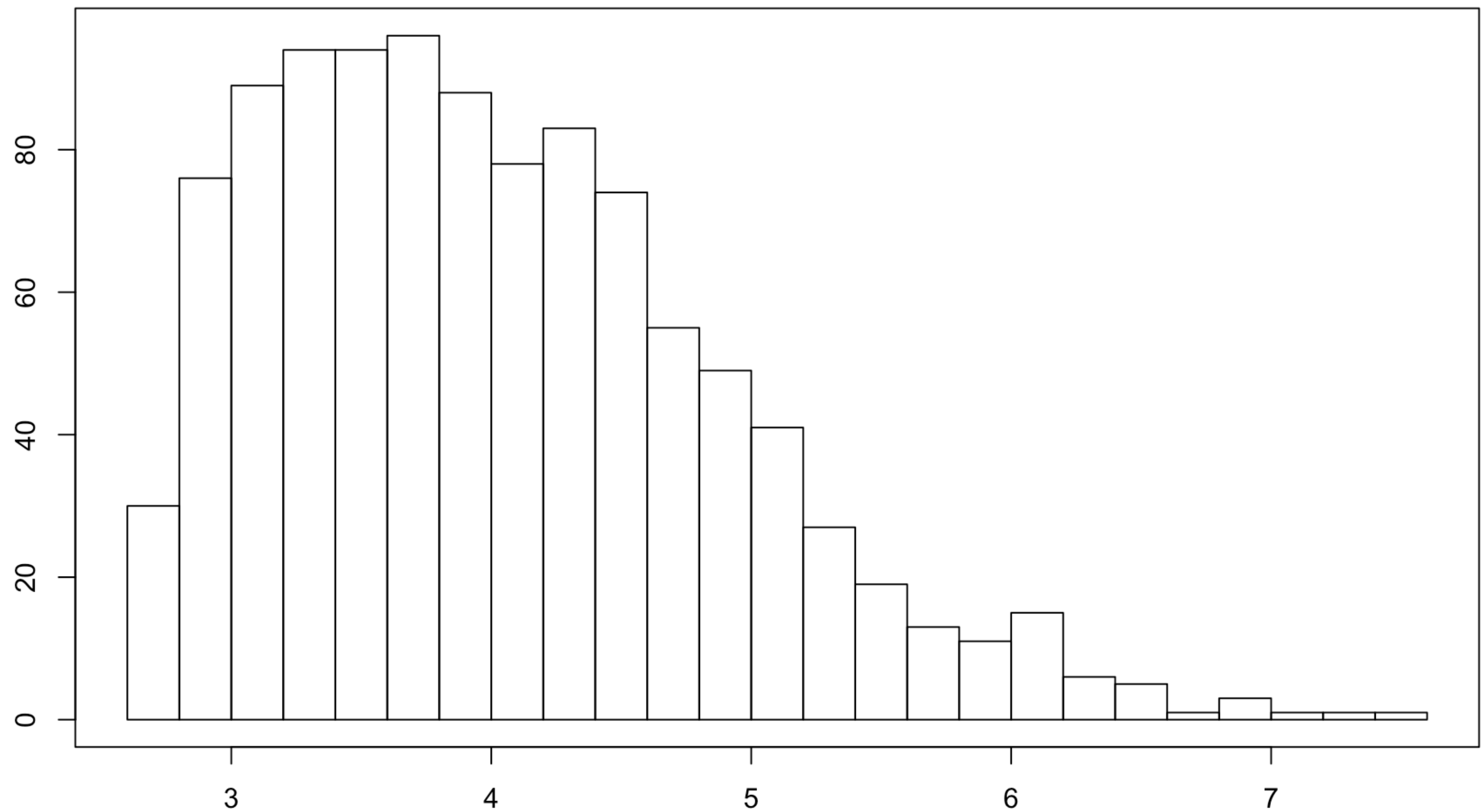
Week 2, January 20th 2017

What is this figure missing?



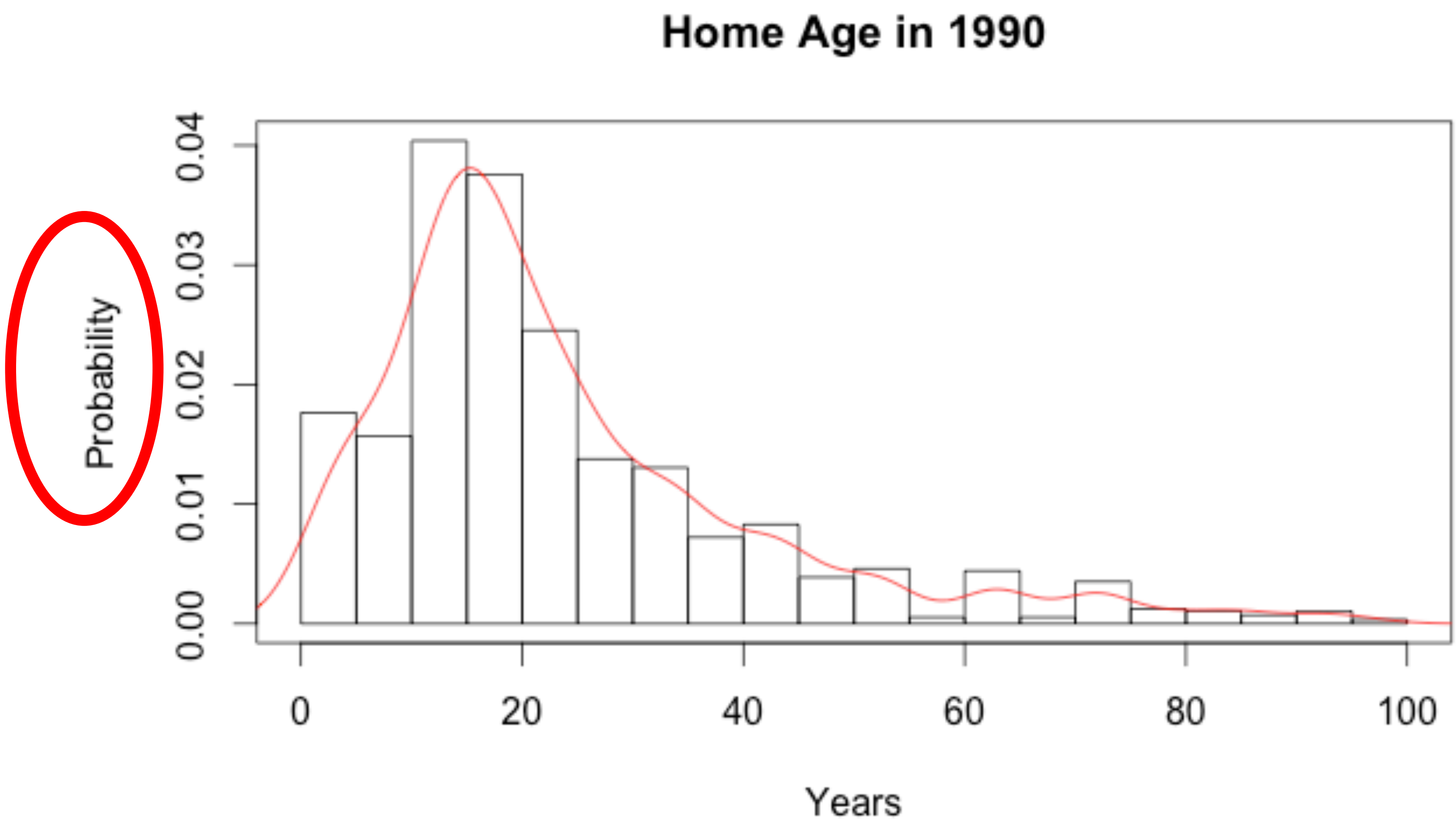
Which is correct?

Option A:
Figure 1 shows the histogram of the log-transformed radon concentrations.

Option B:
The log-transformed radon distributions approximated a normal distribution (Figure 1).

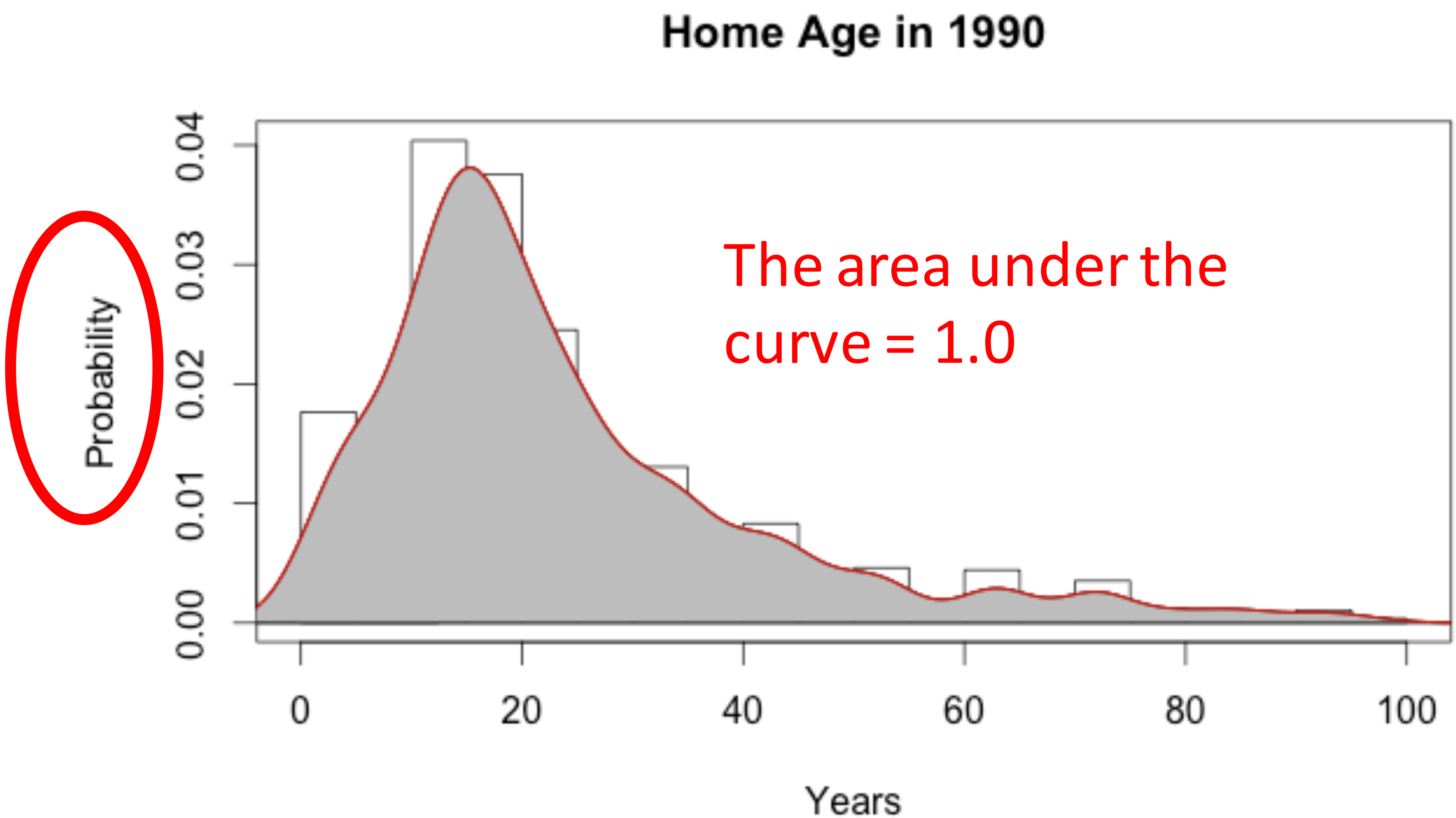
Probability Distributions

- Frequency histograms (vertical bars) give you an general idea of the probability distribution of your data
- The density function (red line) gives the exact shape



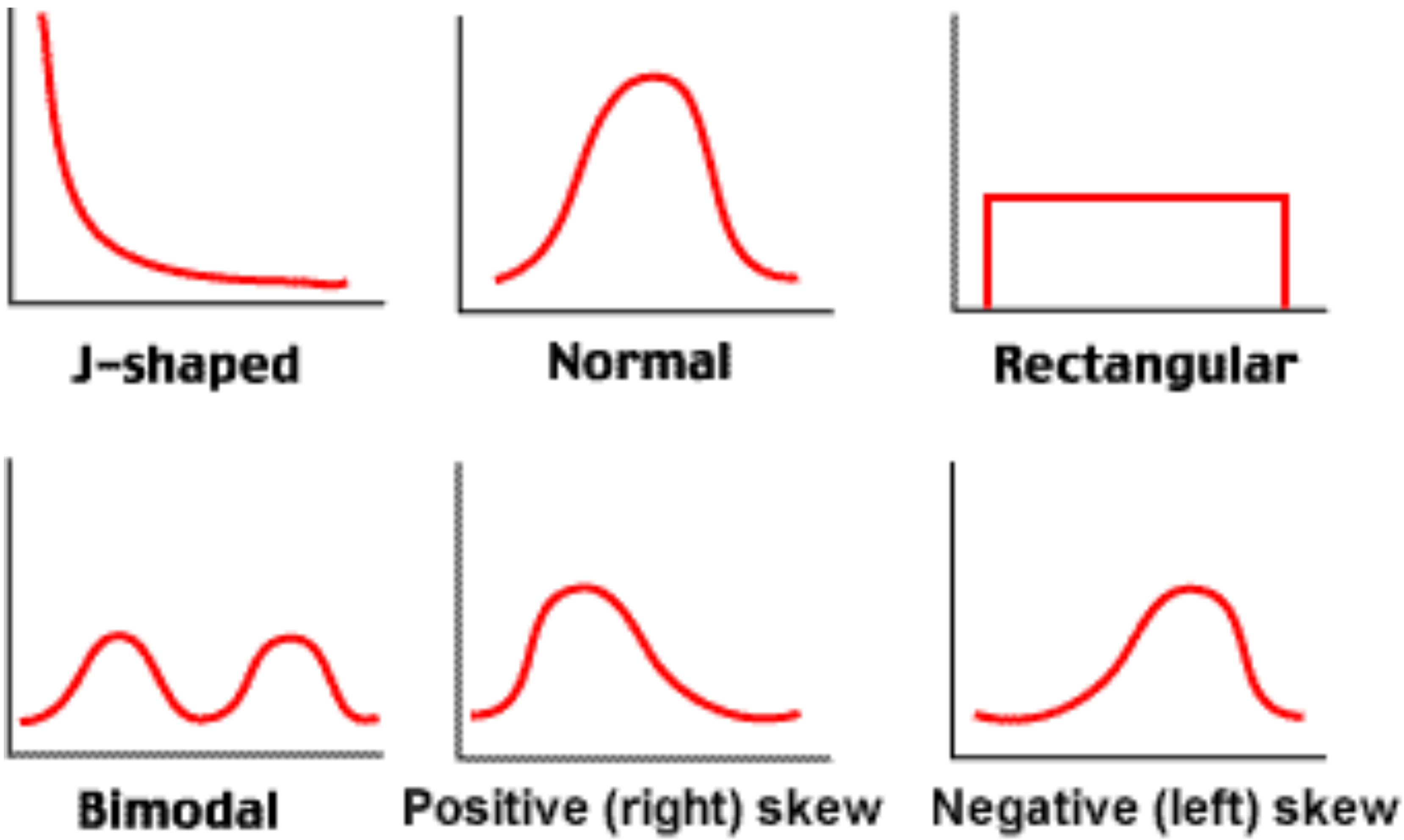
Probability Distributions

- Frequency histograms (vertical bars) give you an general idea of the probability distribution of your data
- The density function (red line) gives the exact shape



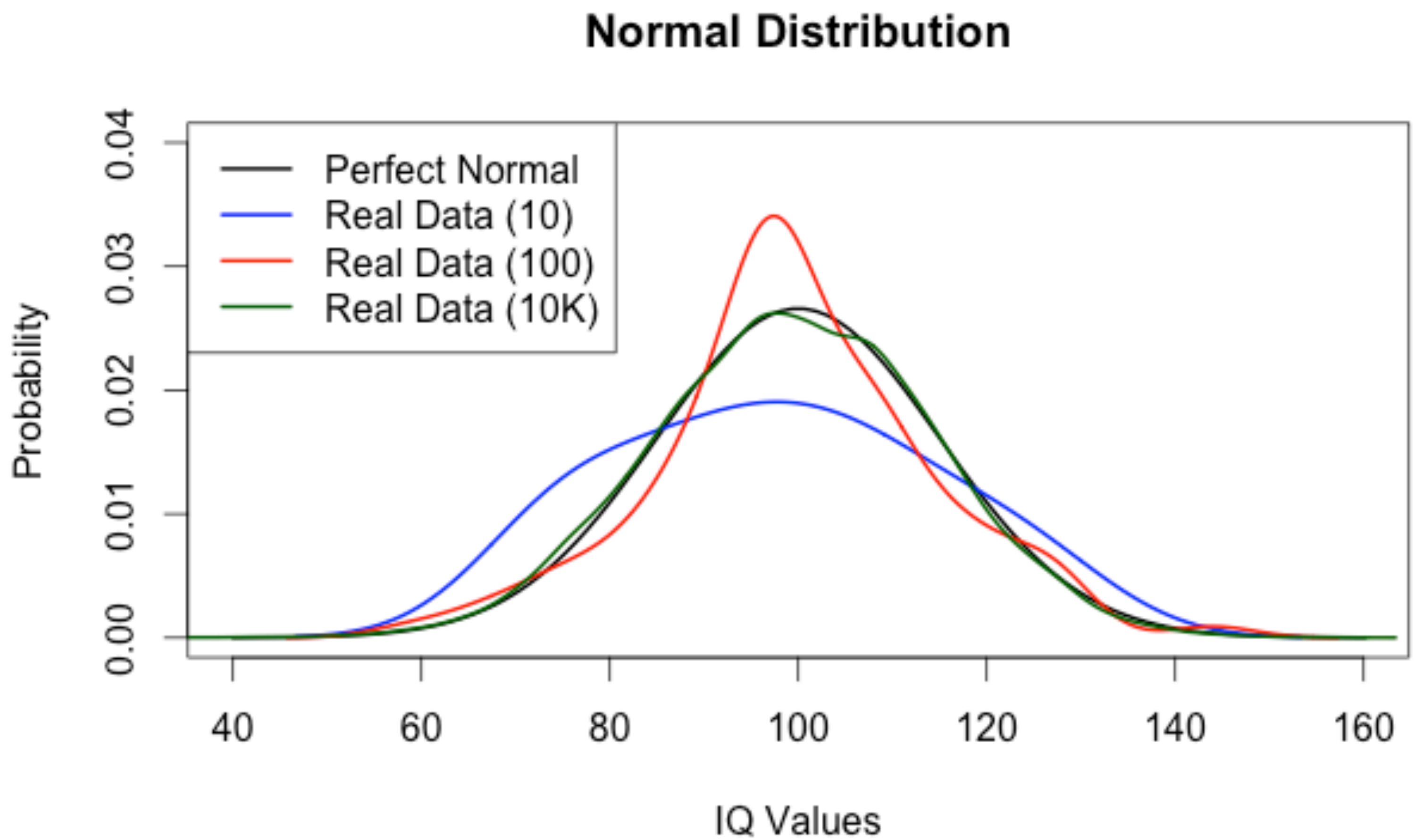
Probability Distributions

- Most continuous data approximate the shape of a STANDARD probability distribution, and this is why we can do statistics
- We will focus on PARAMETRIC methods that assume our data follow some type of normal distribution



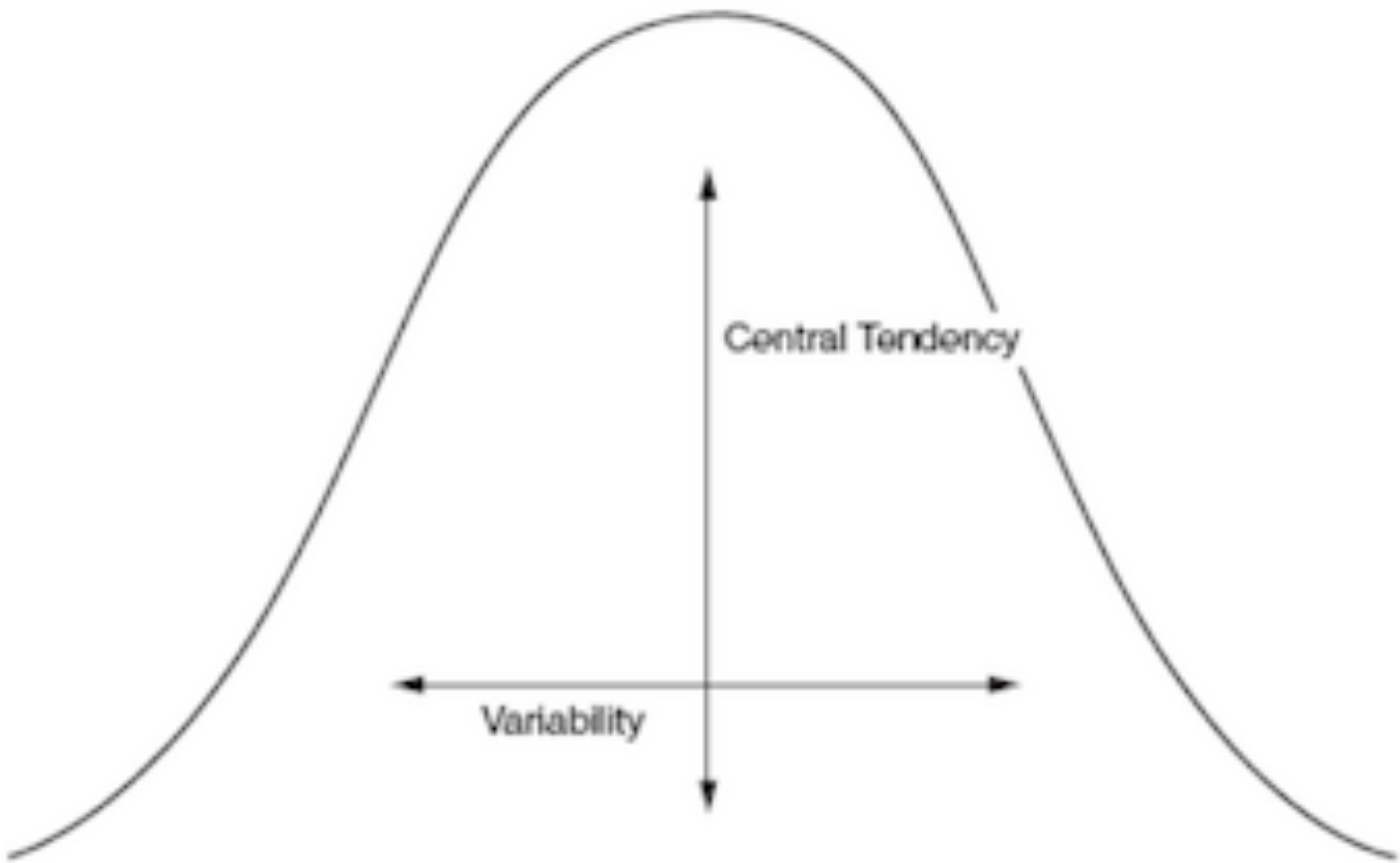
Normal Distribution

- IQ scores of children follow a normal distribution with a mean of 100 and standard deviation of 15
- Real data NEVER follow a hypothetical perfectly normal distribution. The more data you have, the better for characterizing the distribution.



Summary Statistics

- We use these to describe the CENTRAL TENDANCY and VARIABILITY of the data.
- What are the mean, median, and mode? Where are they on a perfectly normal distribution?



Calculate the Mean

$$\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$$

```
[1]  94 115 127 110 102 103  92  82  75  83
```

NOTE: $\bar{x}$ is the SAMPLE MEAN and μ is the POPULATION MEAN

Calculate the Standard Deviation

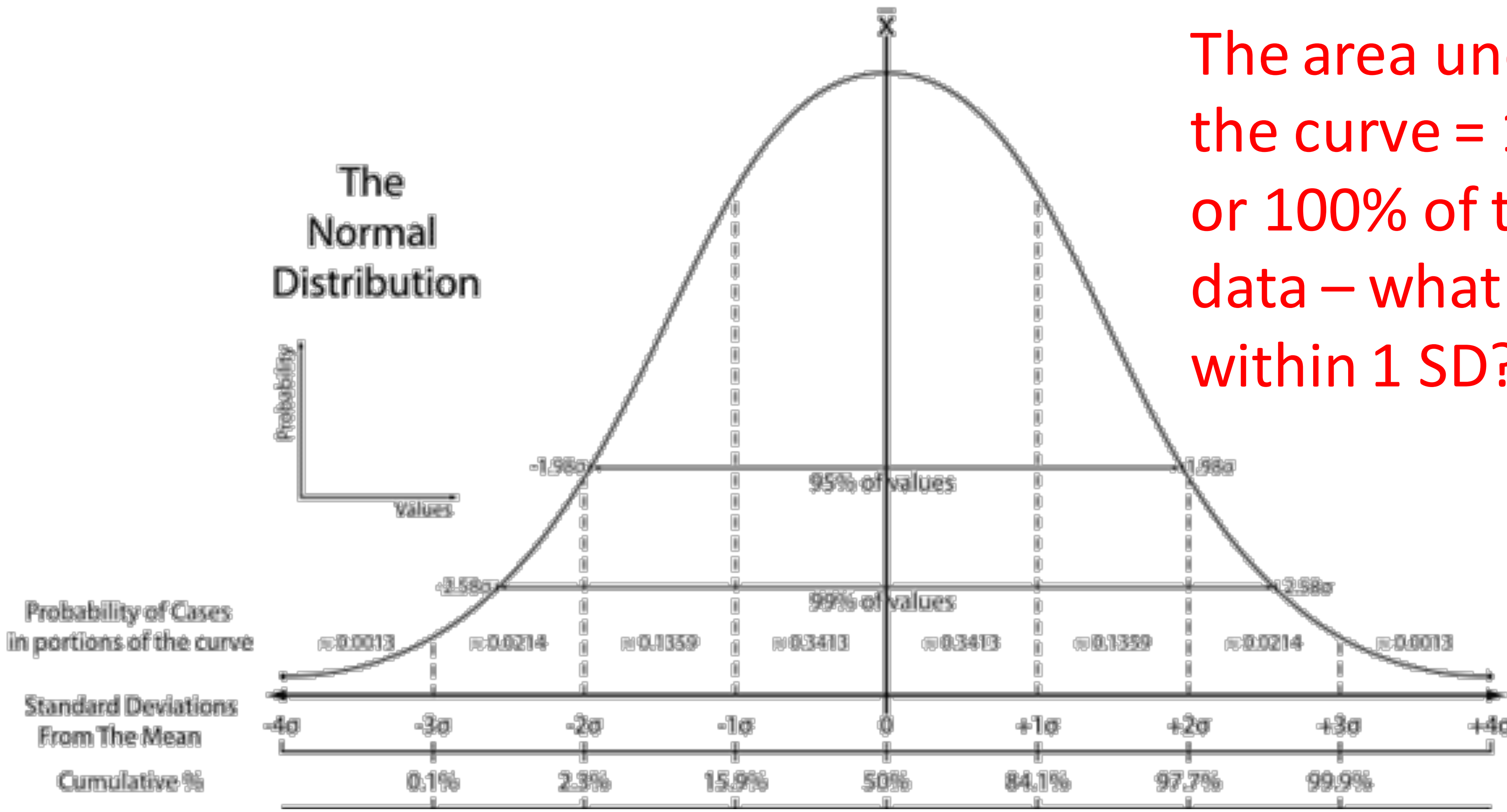
$$s = \sqrt{\frac{1}{N-1} \sum_{i=1}^{N} (x_i - \overline{x})^2}.$$

```
[1]  94 115 127 110 102 103  92  82  75  83
```

NOTE: *s* is the SAMPLE SD and σ is the POPULATION SD

Percentiles / Quantiles

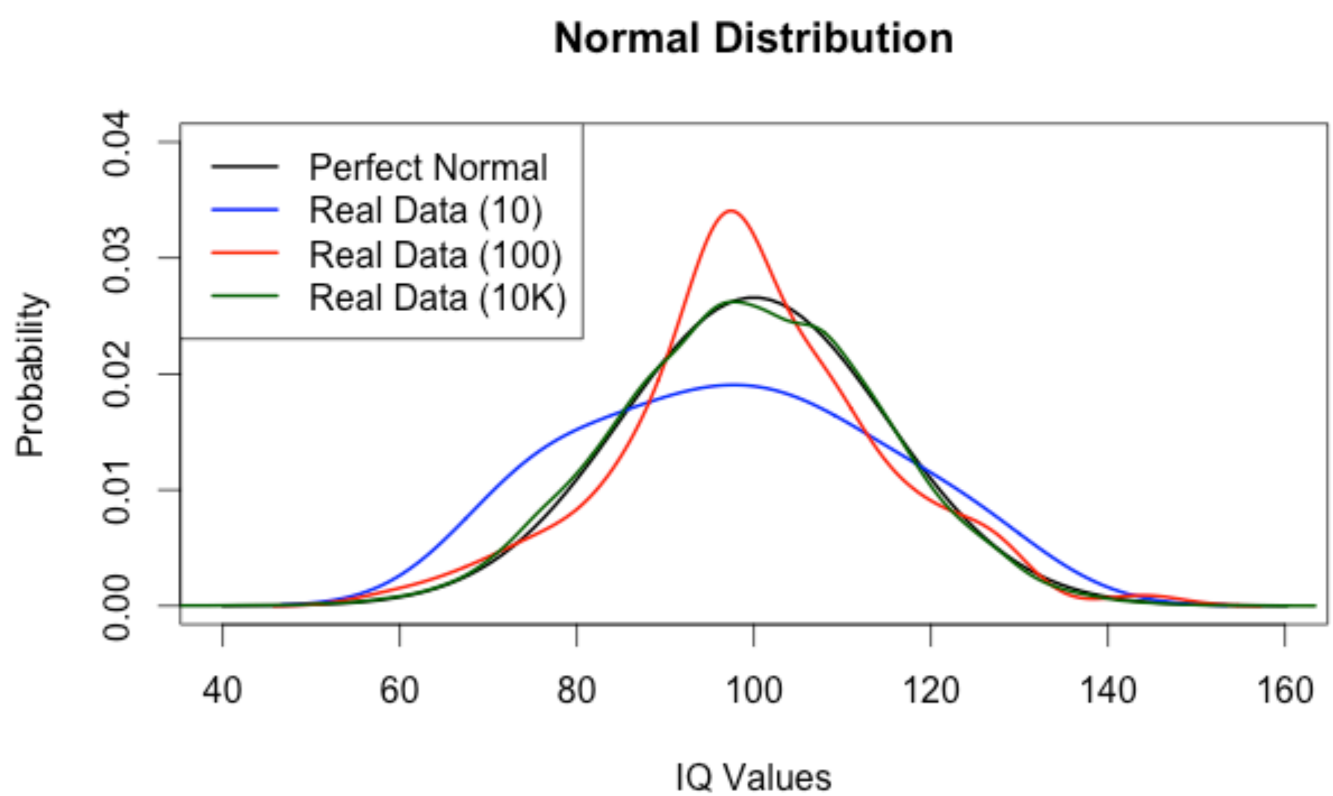
- Percentiles are the values below which a percentage of the data are distributed.
- The MEDIAN is also the 50th percentile
- The 75th percentile is the value below which 75% of the data can be found
- The INTERQUARTILE RANGE (IQR) is the range between the 25th and 75th percentiles



The area under the curve = 1.0 or 100% of the data – what % within 1 SD?

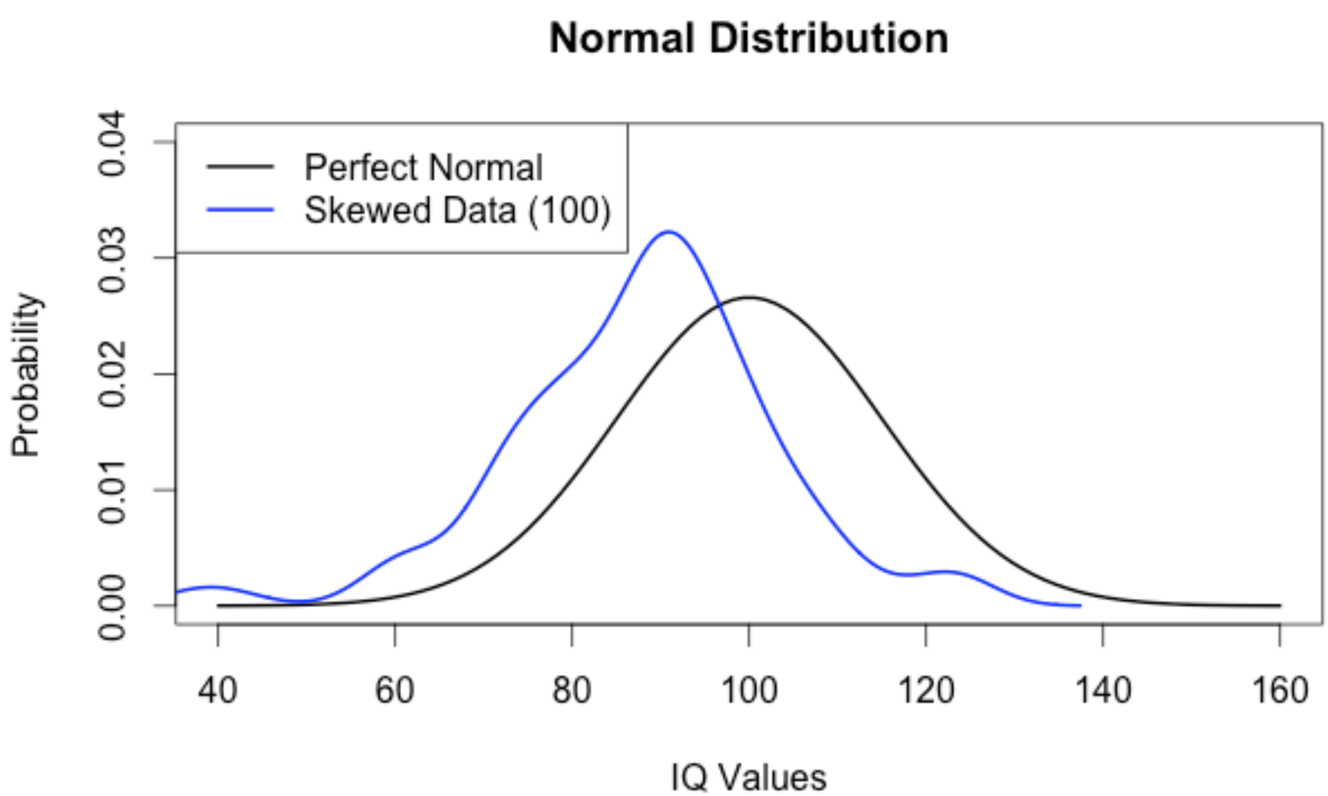
Are These Data Normal?

- We have different tools available to address this question:
  - Visual inspection of the density function of the sample compared with a standard normal curve. How do its skewness and kurtosis compare?
  - Visual inspection of the quantile-quantile plot (QQ plot)
  - Statistical tests for goodness-of-fit with data that follow a perfectly normal distribution, like the Shapiro-Wilk test



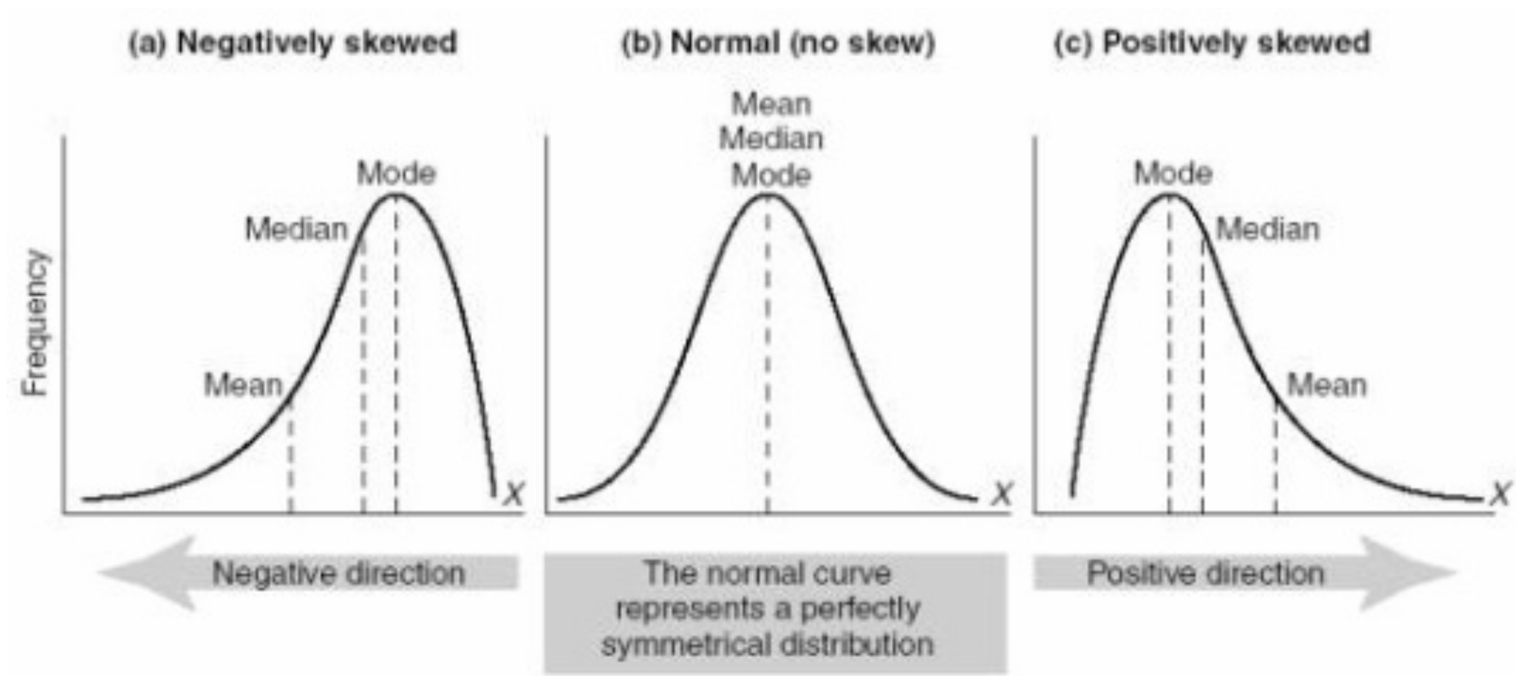
Are These Data Normal?

- We have different tools available to address this question:
  - Visual inspection of the density function of the sample compared with a standard normal curve. How do its skewness and kurtosis compare?
  - Visual inspection of the quantile-quantile plot (QQ plot)
  - Statistical tests for goodness-of-fit with data that follow a perfectly normal distribution, like the Shapiro-Wilk test



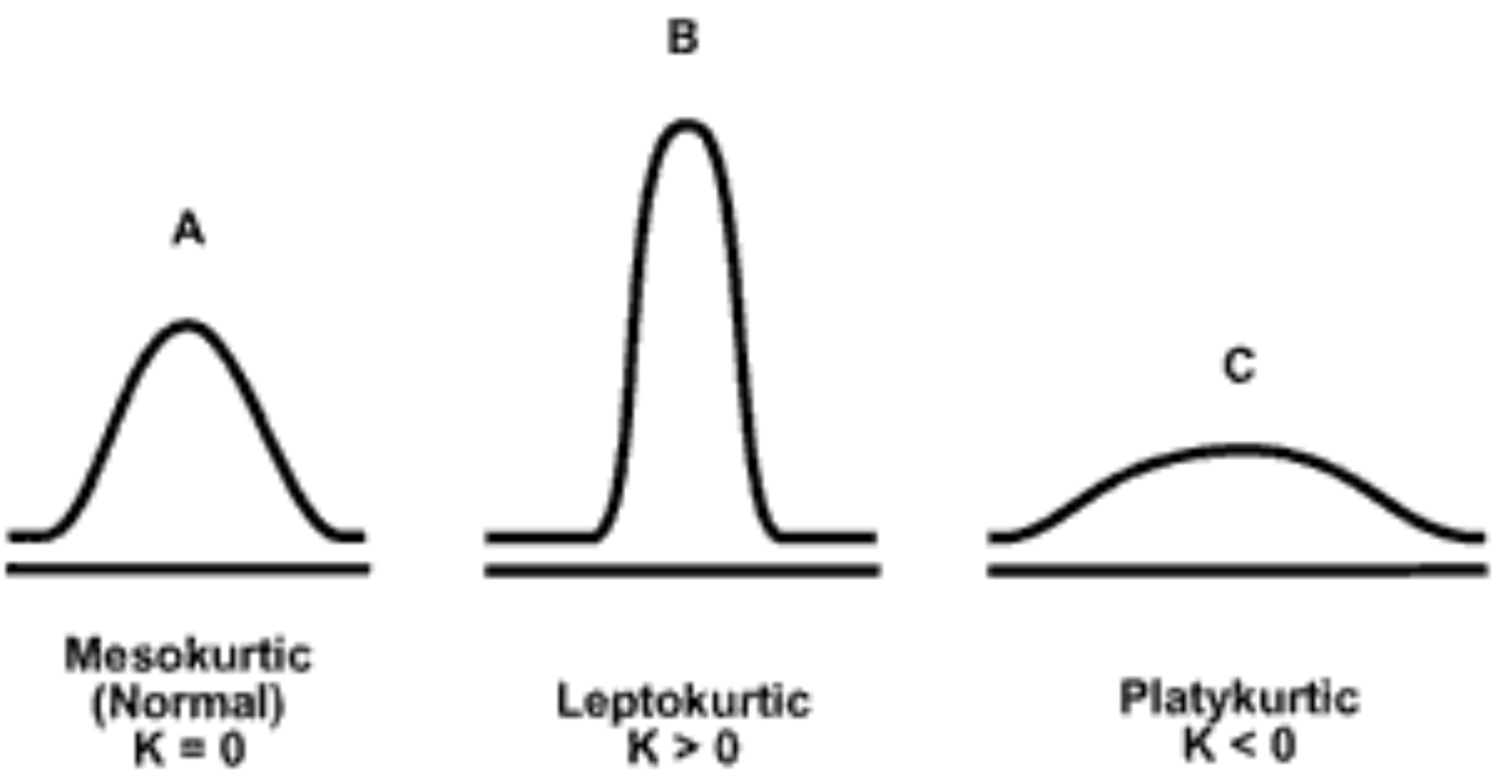
Skewness

- Skewness describes the shape of the distribution on the x-axis relative to the hypothetical normal
- A perfectly normal distribution will have the same MEAN and MEDIAN value
- If the distribution is POSITIVELY or RIGHT-SKEWED (longer right tail) the mean is higher than the median
- If the distribution is NEGATIVE or LEFT-SKEWED (longer left tail) the mean is lower than the median



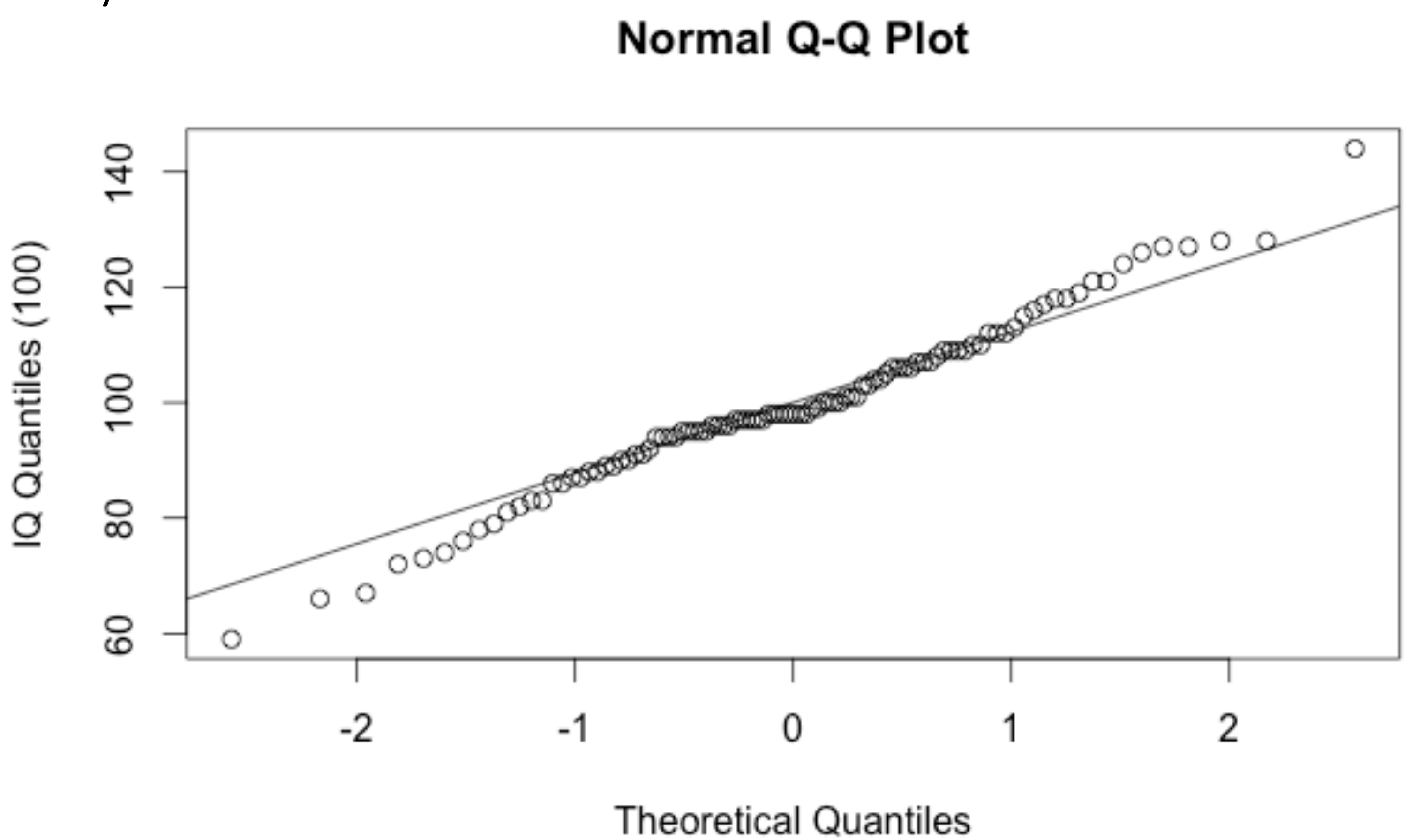
Kurtosis

- Kurtosis describes the shape of the distribution on the y-axis relative to the hypothetical normal
- A perfectly normal distribution has a bell shaped
- A distribution with positive kurtosis is relatively taller and skinnier, a value >3 indicates critically non-normal kurtosis
- A distribution with negative kurtosis is relatively shorter and flatter, a value <-3 indicates critically non-normal kurtosis



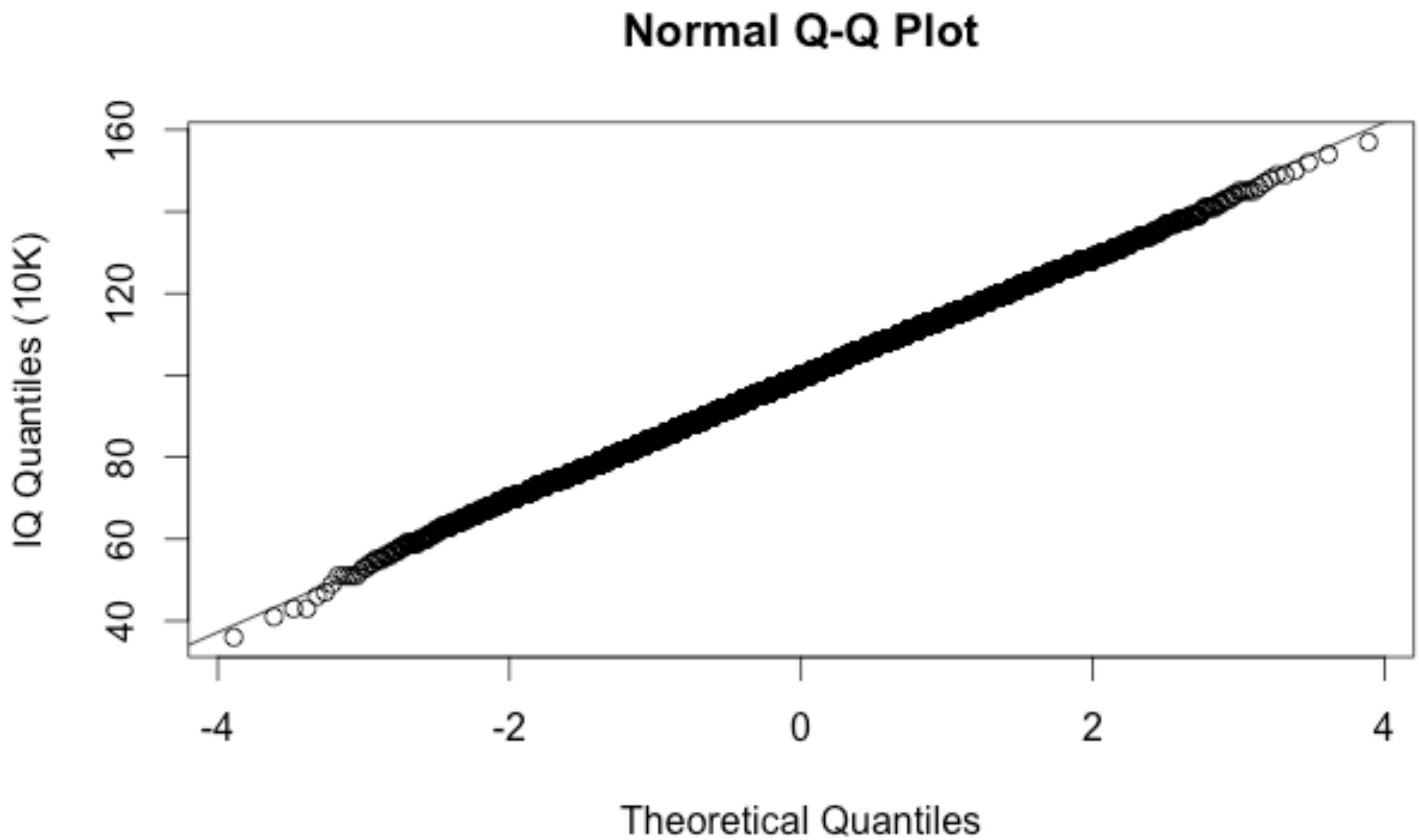
Quantile-Quantile Plots

- QQ plots show the quantiles of the sample compared with the quantiles of a standard normal distribution
- This is basically a scatter plot of the 1st, 2nd, 3rd…98th, 99th, and 100th percentiles of the sample and the standard normal
- If the scattered of dots falls along a straight line, it is evidence that the data are normally distributed



Quantile-Quantile Plots

- QQ plots show the quantiles of the sample compared with the quantiles of a standard normal distribution
- This is basically a scatter plot of the 1st, 2nd, 3rd…98th, 99th, and 100th percentiles of the sample and the standard normal
- If the scattered of dots falls along a straight line, it is evidence that the data are normally distributed



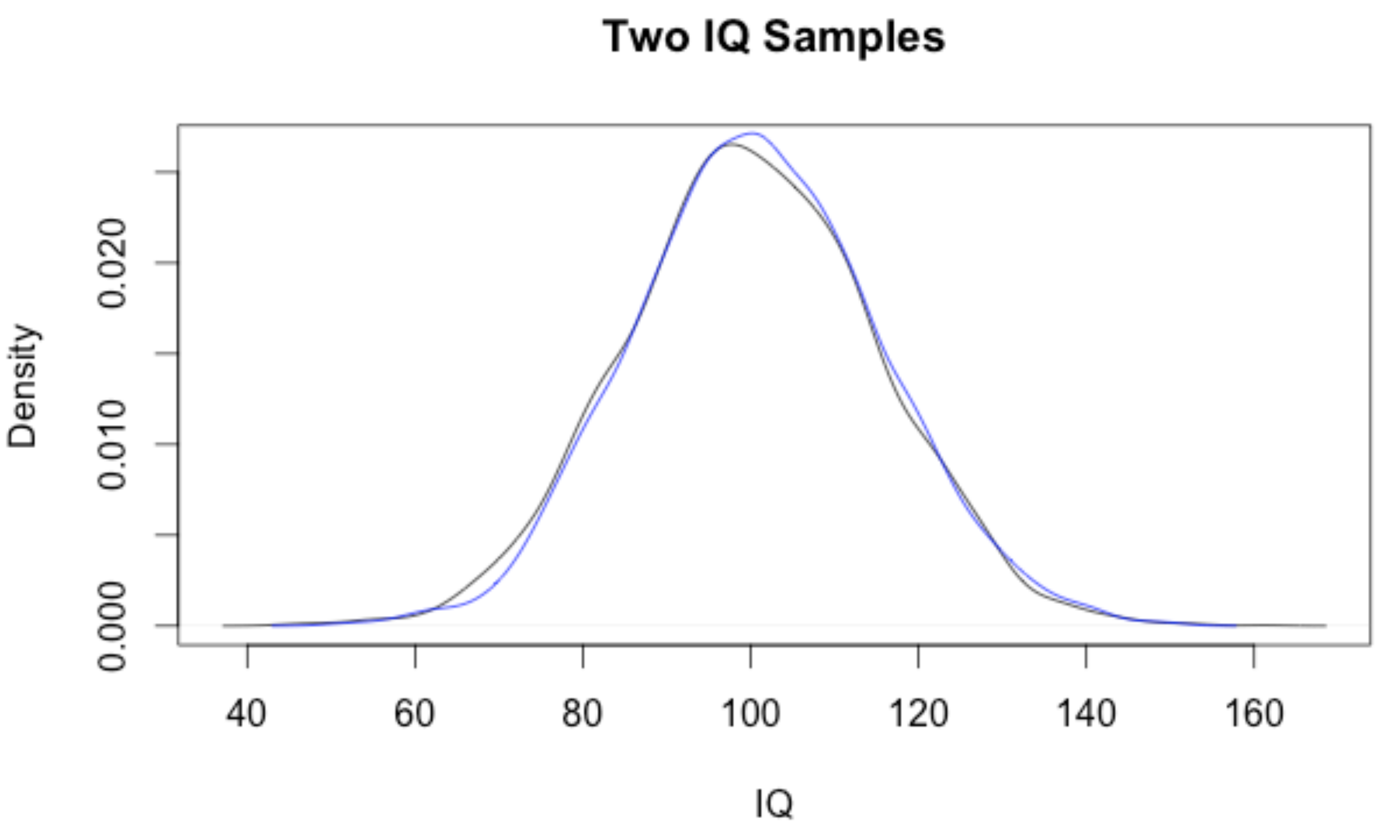
Statistical Hypothesis Testing

- You are always testing a NULL HYPOTHESIS (H_0)
- If the test passes, the H_0 is accepted
- If the test fails, the ALTERNATE HYPOTHESIS (H_1) is accepted
- We use the p-value to evaluate whether the test passed or failed
- Typically we use 0.05 as the critical value for p, which means we are willing to wrongly reject H_0 in 5% of cases.
- So, if the value of $p < 0.05$, the test fails and we:
    1. Reject the null hypothesis
    2. Accept the alternate hypothesis
    3. Have a statistically significant result!
- **A SIGNIFICANT result does not imply a MEANINGFUL result**

Significant vs. Meaningful

```
x = rnorm(5000,100,15)
y = rnorm(5000,101,15)
```

```
        Welch Two Sample t-test

data:  x and y
t = -2.4605, df = 9992.2, p-value = 0.01389
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 -1.3171309 -0.1490637
sample estimates:
mean of x mean of y
 100.1030  100.8361
```



What is the null hypothesis?

Do we reject it?

Shapiro-Wilk Test

- H_0: the sample was drawn from a normal distribution
- H_1: the sample was not drawn from a normal distribution
- Note that H_1 does not tell us ANYTHING about the distribution the sample was drawn from if H_0 is rejected. We would have to try another test for another distribution.

```
        Shapiro-Wilk normality test

data:  iq1
W = 0.96356, p-value = 0.8256
```

Real Data, 10 values

```
        Shapiro-Wilk normality test

data:  iq2
W = 0.98524, p-value = 0.3305
```

Real Data, 100 values

```
        Shapiro-Wilk normality test

data:  iq4
W = 0.97049, p-value = 0.02412
```

Skewed Data, 100 values

Test Statistics

- Whenever you run a statistical test, the foundation of that test is called its **STATISTIC**
- The test statistic is calculated by the computer, but you should understand the concept
- The values of the test statistic, themselves, expected to follow some probability distribution
- The p-value represents the probability of observing the calculated test statistic **BY CHANCE ALONE**

$$W = \frac{\left(\sum_{i=1}^{n} a_i x_{(i)}\right)^2}{\sum_{i=1}^{n}(x_i - \overline{x})^2}$$

For Your Assignment / Always

Untransformed values =
use the continuous data as measured

Log-transformed values =
use the natural logarithm of the untransformed values

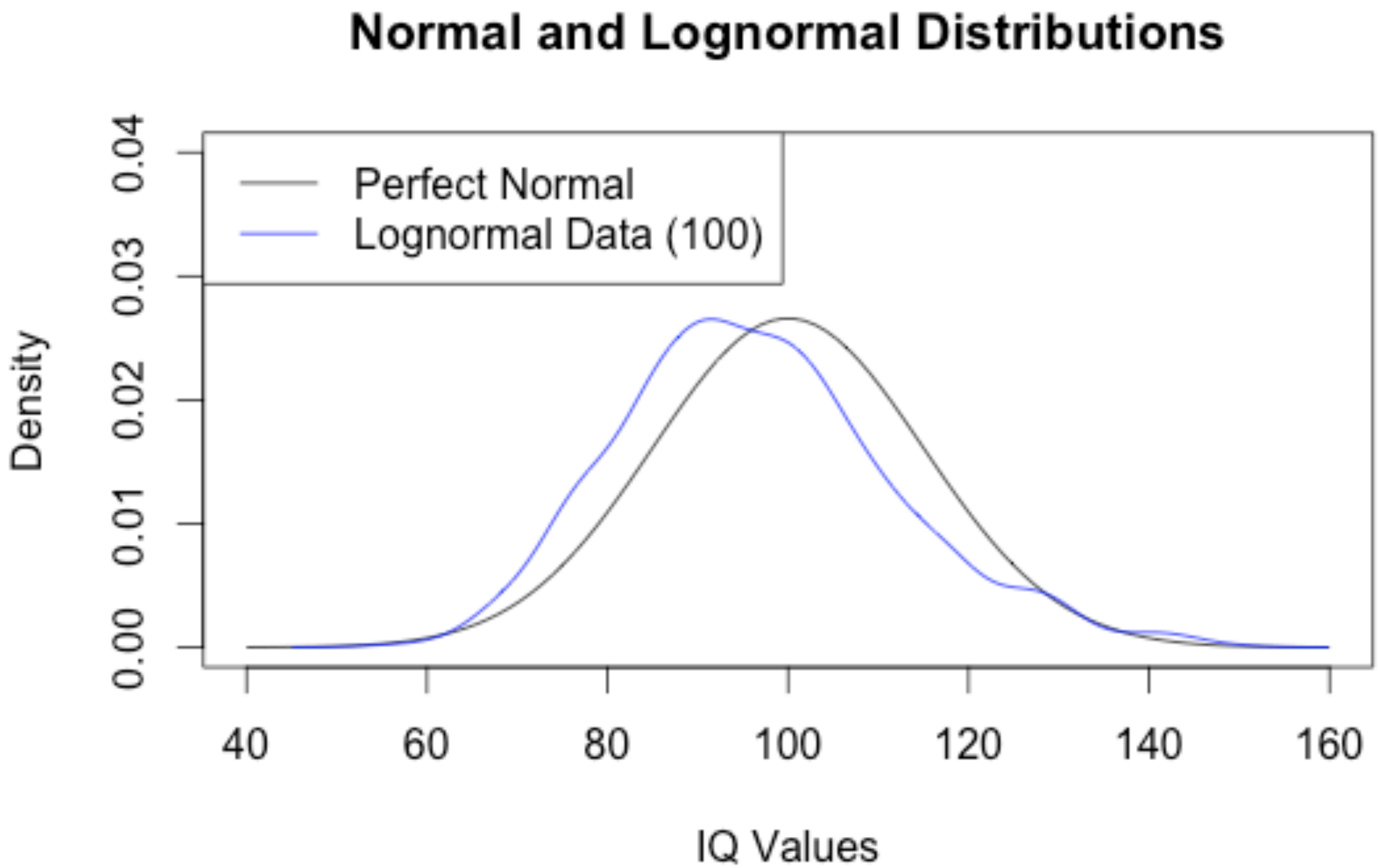
Arithmetic mean = mean of untransformed values
Arithmetic SD = SD of the untransformed values

Log mean = mean of the log-transformed values
Log SD = SD of the log-transformed values

Geometric mean = $e^{(\text{mean of the log-transformed values})}$
Geometric SD = $e^{(\text{SD of the log-transformed values})}$

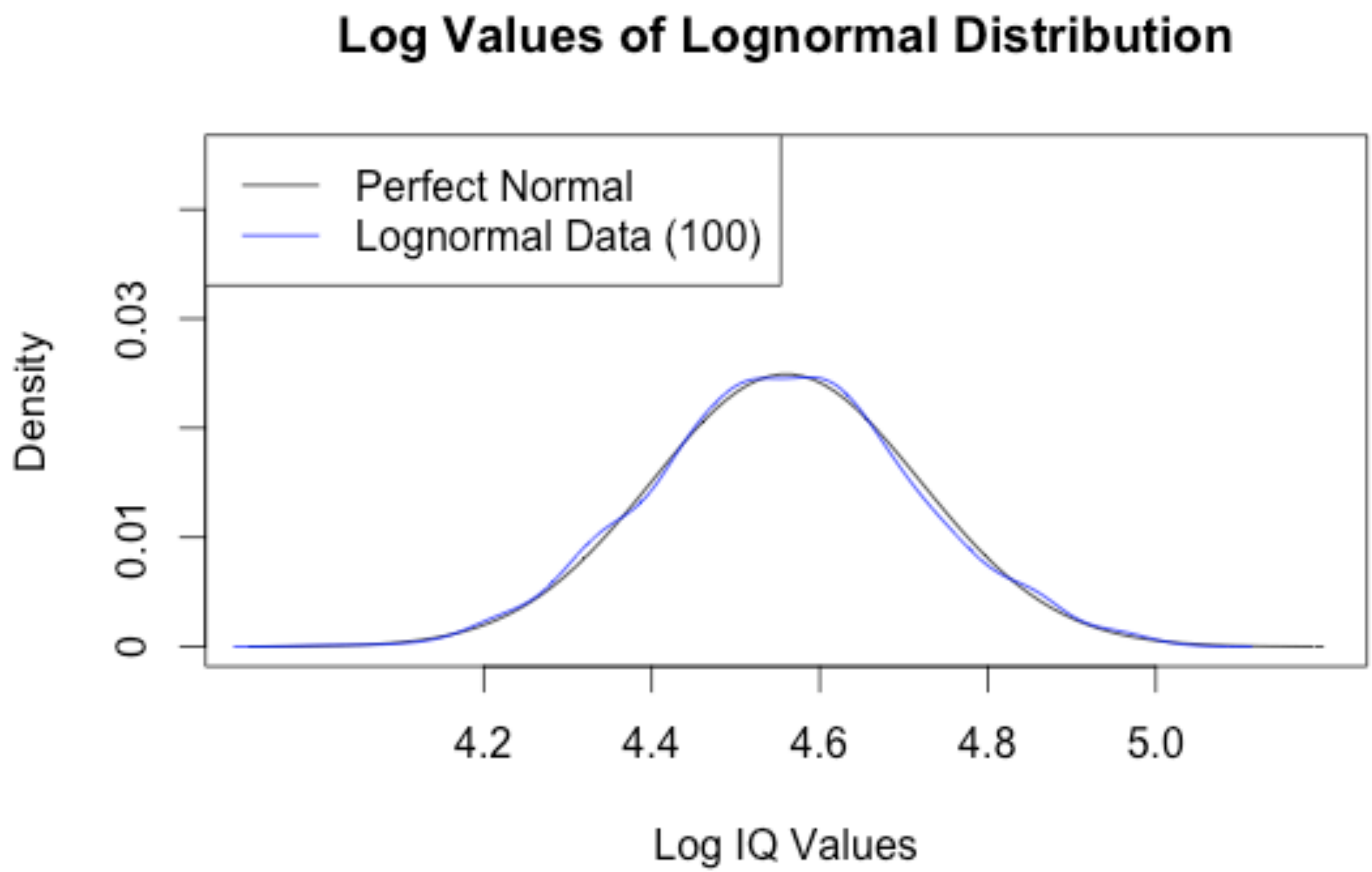
Log-Normal Distribution

- Distribution of the sample Is right-skewed
- Taking the natural logarithm of the sample yields a normal distribution



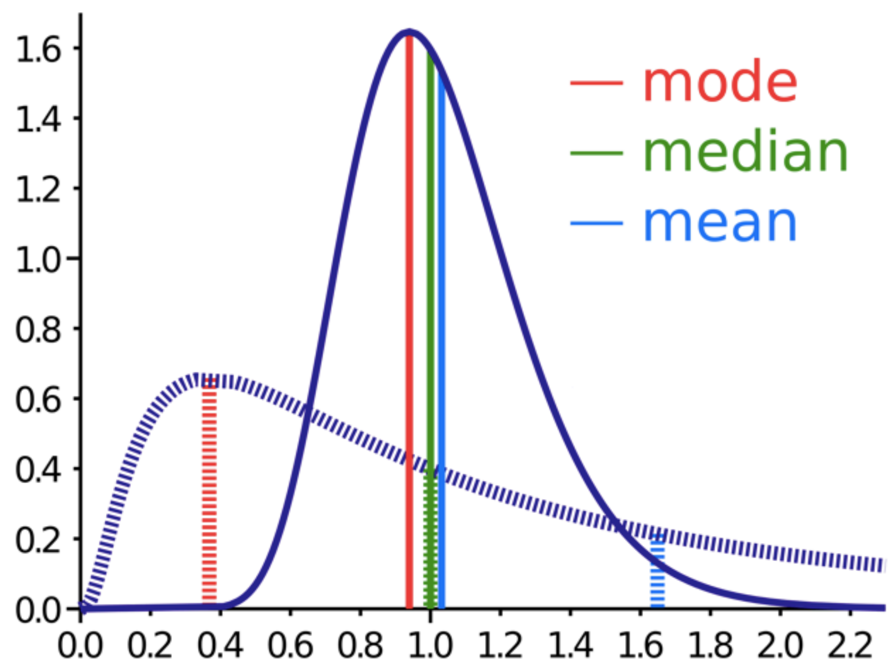
Log-Normal Distribution

- Distribution of the sample Is right-skewed
- Taking the natural logarithm of the sample (i.e. log-transforming the values) yields a normal distribution



Summary Statistics

- We still use these to describe the CENTRAL TENDANCY and VARIABILITY of the data.
- What are the mean, median, and mode? Where are they on a lognormal distribution?
- The central tendency of lognormal distributions is better described by the GEOMETRIC mean than the ARITHMETIC mean
- The spread of lognormal distributions is better described by the GEOMETRIC standard deviation than the ARITHMETIC standard deviation



Calculate the Geometric Mean

- If you say “mean” people (including me) will assume that you are talking about the ARITHMETIC mean
- If you are reporting the GEOMETRIC mean, you must specify
- The GEOMETRIC mean will always be less than the ARITHMETIC mean

$$\overline{x}_g = \exp\left[\frac{1}{n}\sum_{i=1}^{n}\log X_i\right]$$

In other words, the antilog of the arithmetic mean of the log-transformed values

```
[1]  94 115 127 110 102 103  92  82  75  83
```

Calculate the Geometric SD

- The GEOMETRIC standard deviation is analogous to the ARITHMETIC standard deviation in the same way that the geometric mean is analogous to the arithmetic mean

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

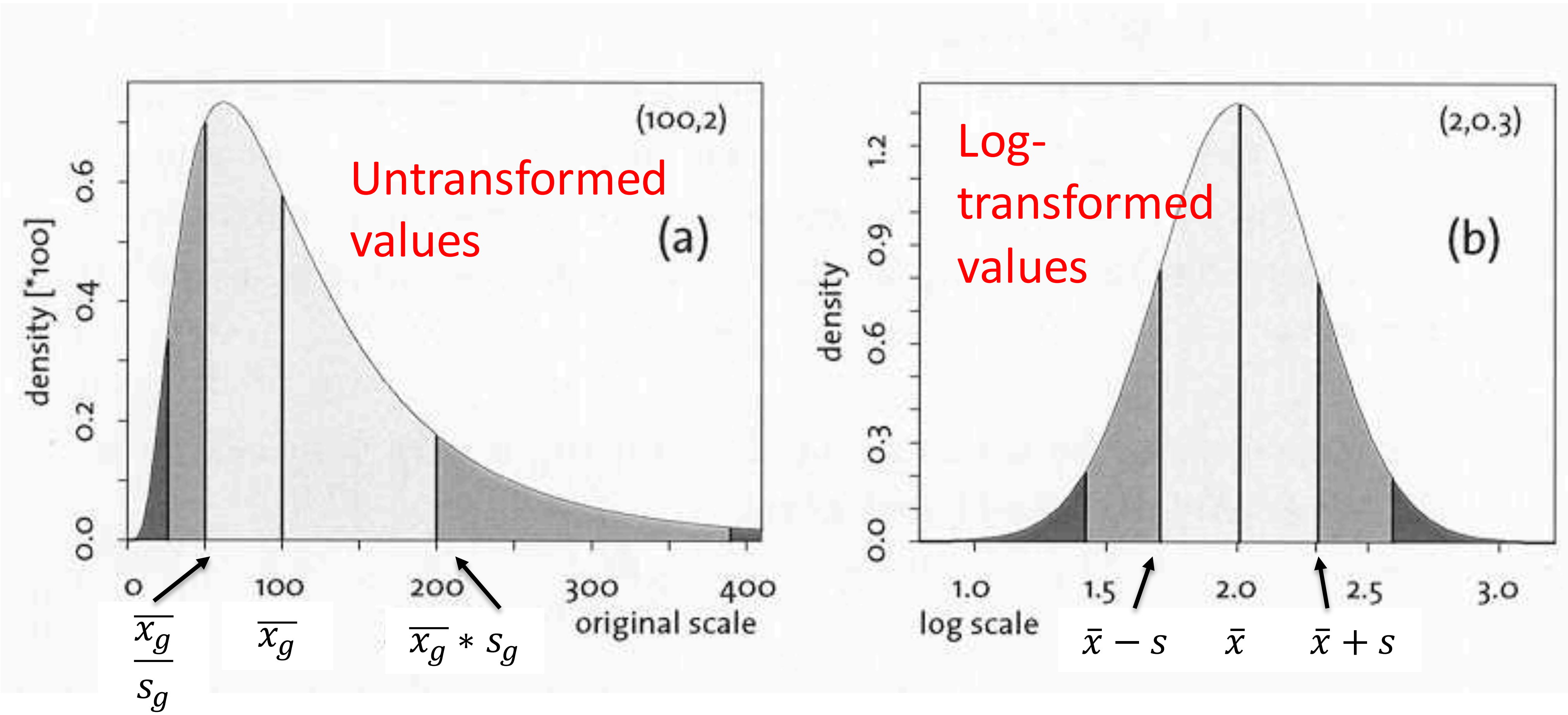
$$s_g = exp\left(\sqrt{\frac{1}{N-1}\sum_{i=1}^{N}(log x_i - log \bar{x}_i)^2}\right)$$

In other words, the antilog of the arithmetic standard deviation of the log-transformed values

```
[1]  94 115 127 110 102 103  92  82  75  83
```

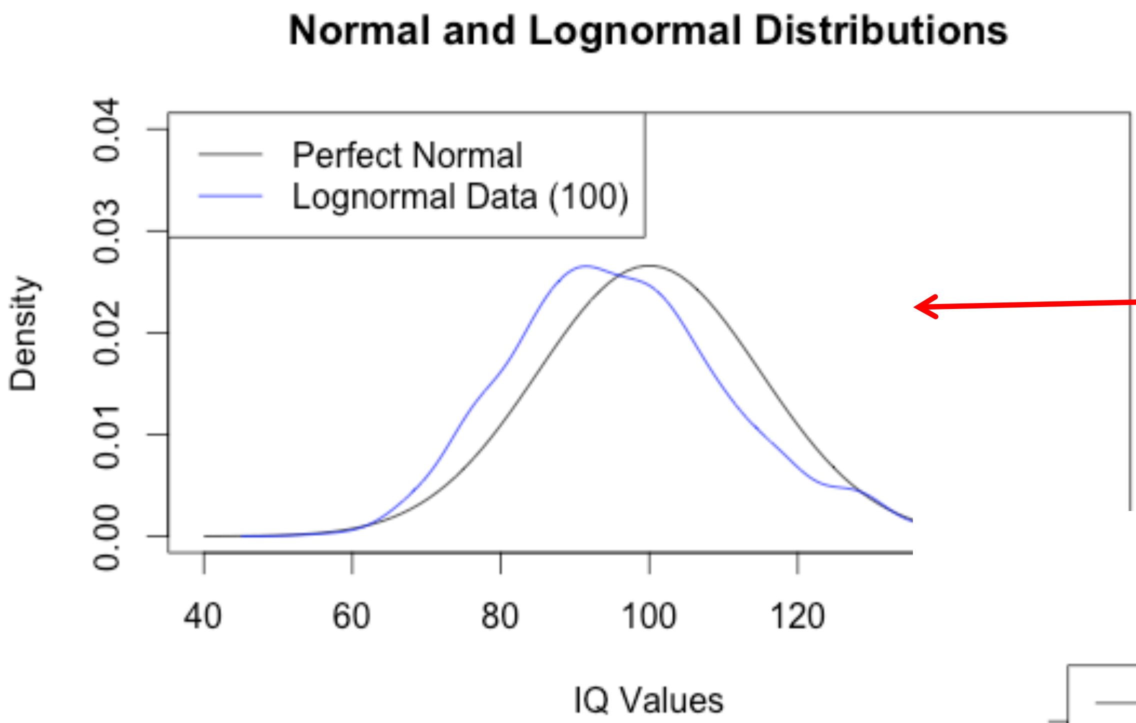
Lognormal Curves

- In a perfect lognormal distribution the MEDIAN and the GEOMETRIC MEAN are exactly the same!
- Instead of adding the GEOMETRIC standard deviation to the geometric mean, you multiply to the right of the median and divide to the left

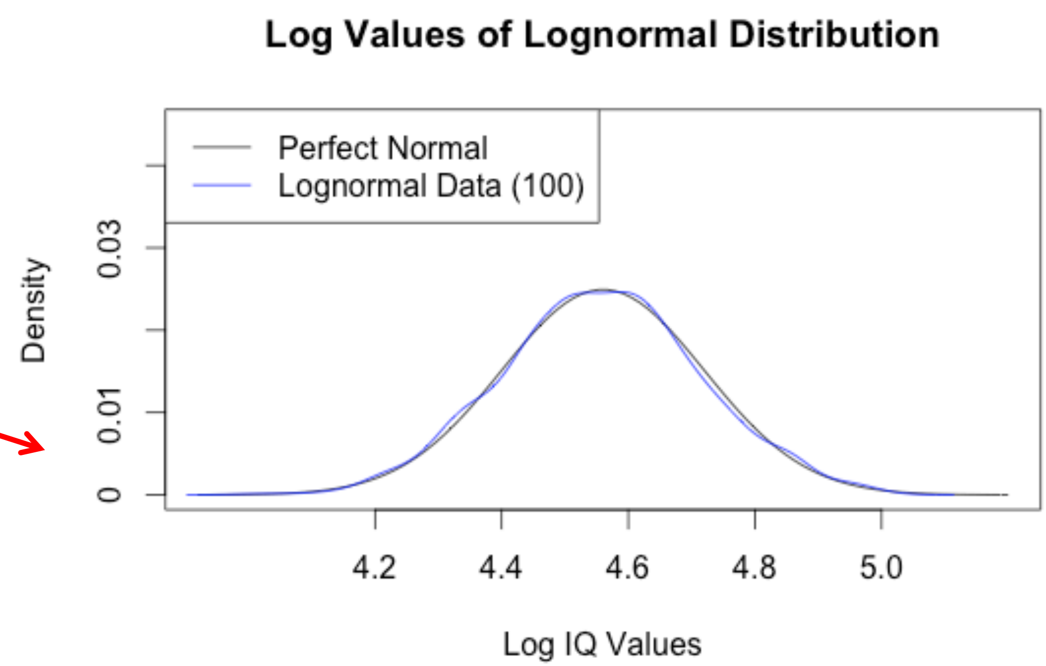


Question: how do we test for log normality?

Plot Distributions

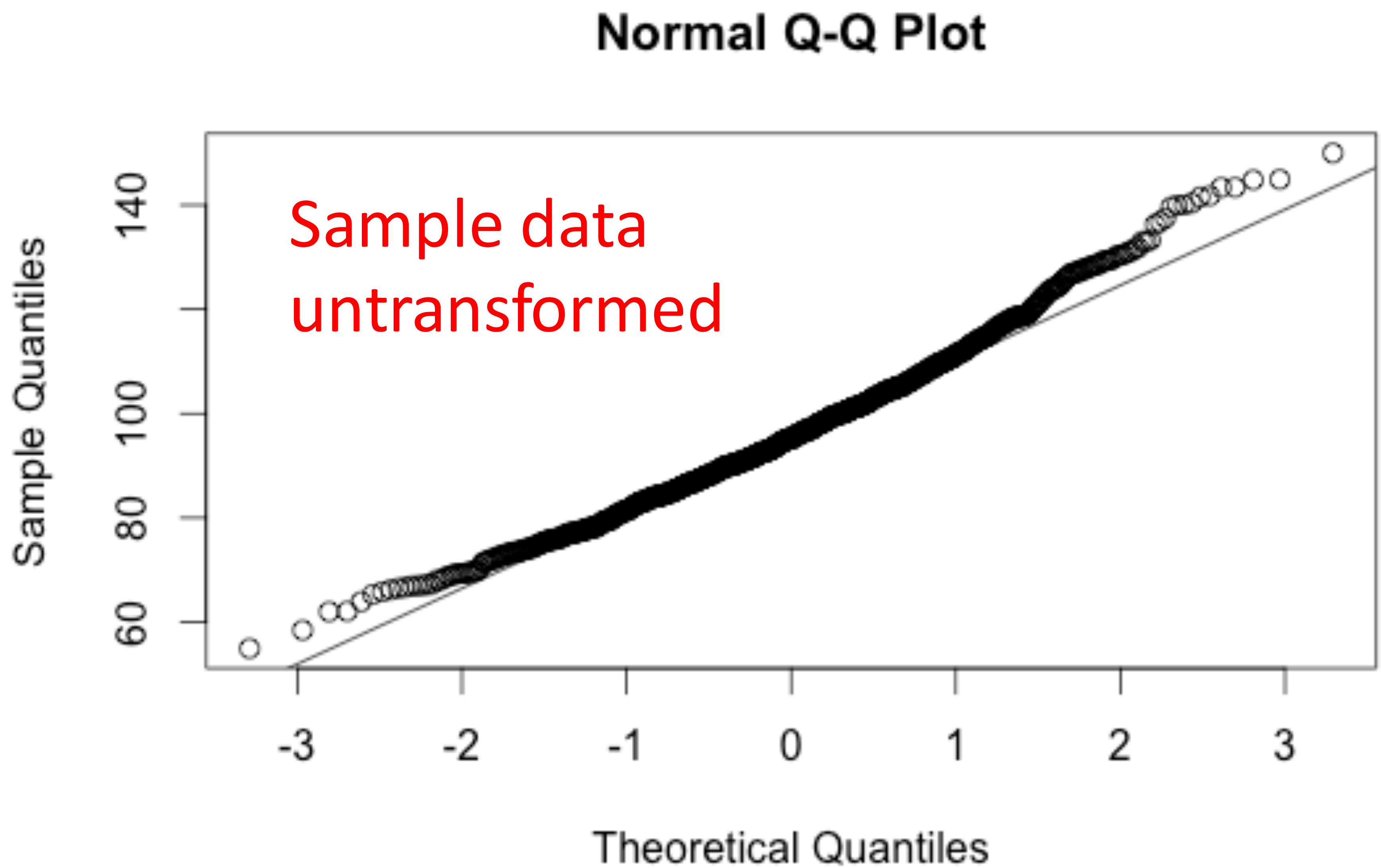


Sample data untransformed

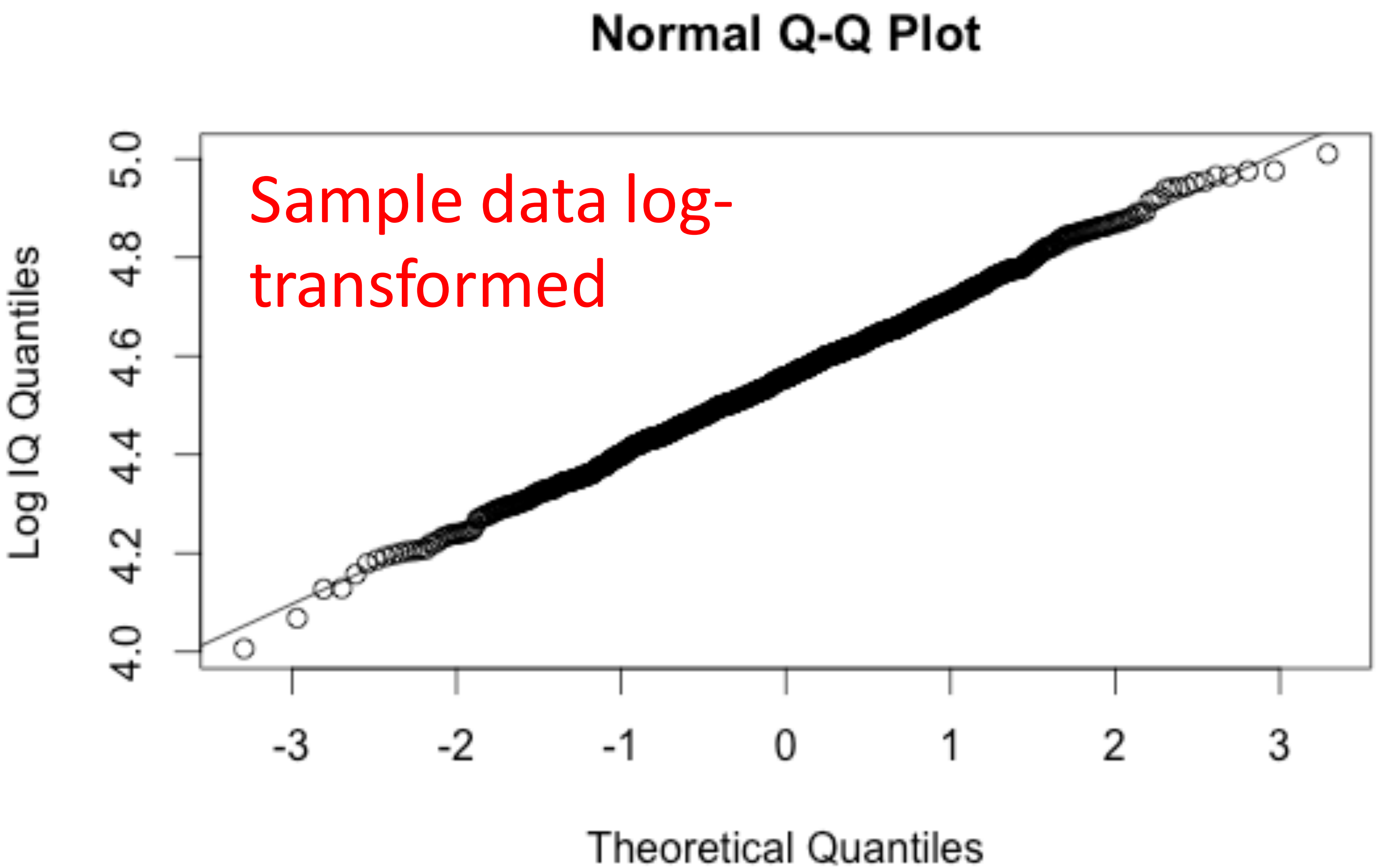


Sample data log-transformed

QQ Plots


Normal Q-Q Plot
Sample data
untransformed
Sample Quantiles
60
80
100
140
-3
-2
-1
0
1
2
3
Theoretical Quantiles

QQ Plots



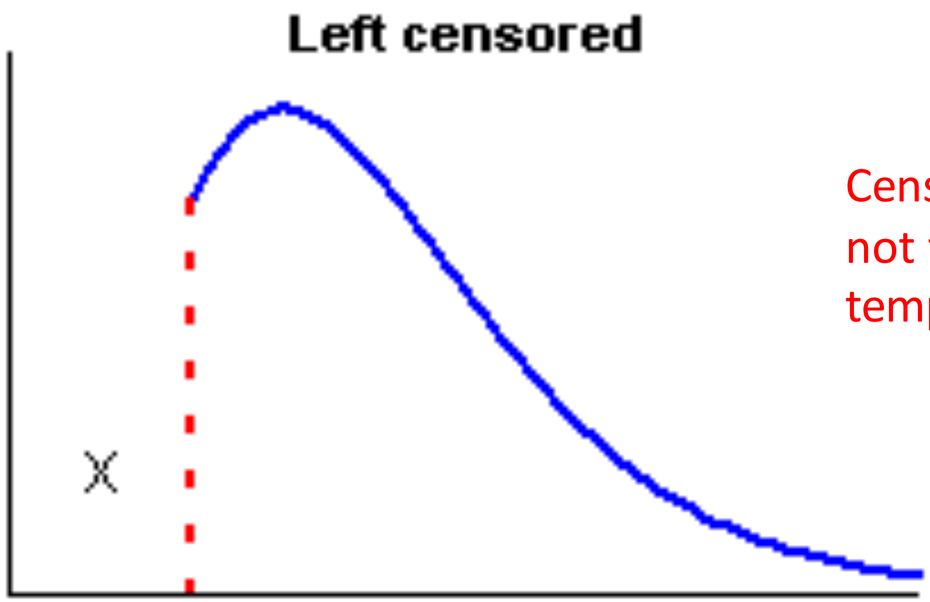
Normal Q-Q Plot
Sample data log-transformed
Log IQ Quantiles
4.0
4.2
4.4
4.6
4.8
5.0
-3
-2
-1
0
1
2
3
Theoretical Quantiles

Shapiro-Wilk Test

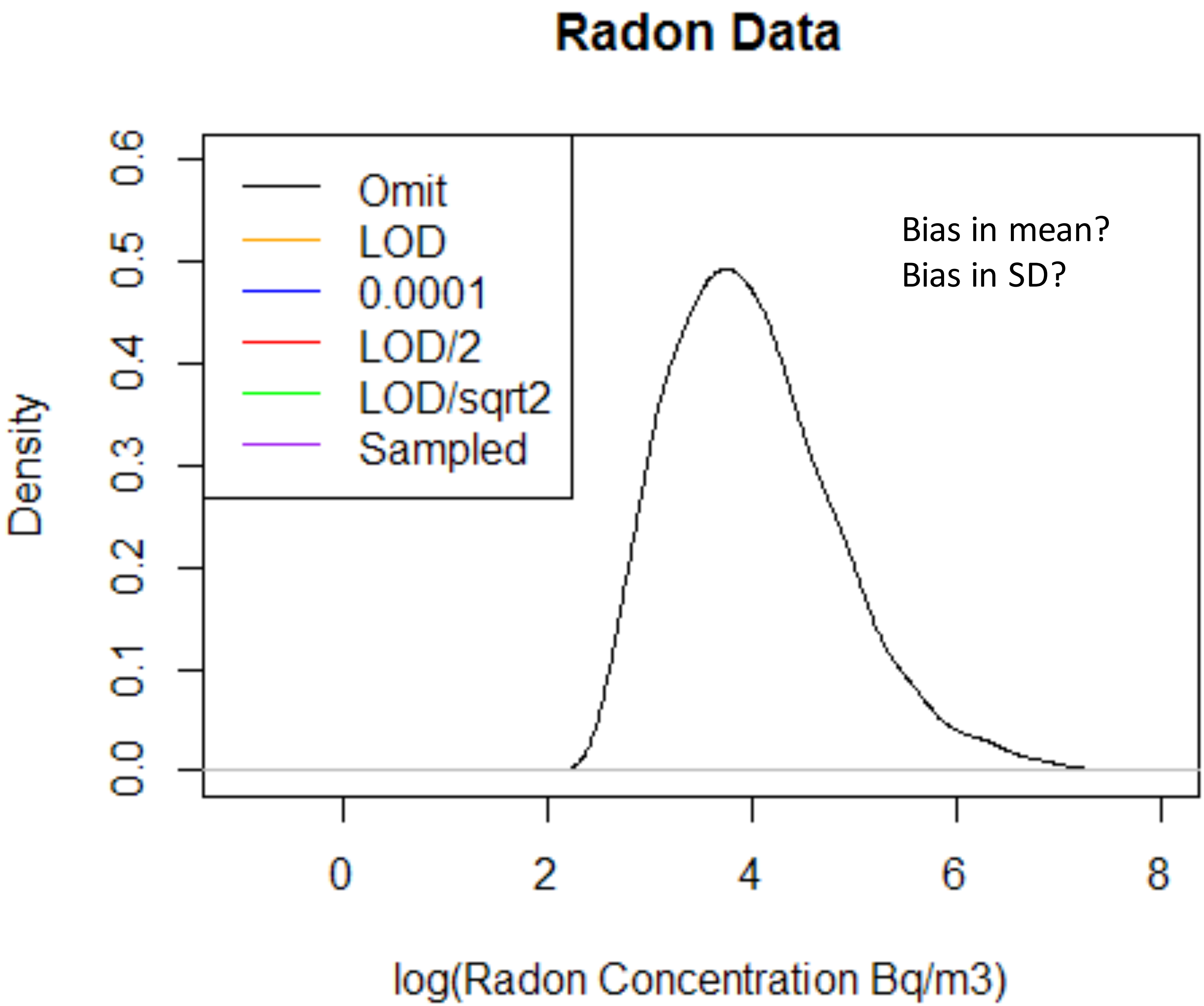
```
> shapiro.test(iq4)

        Shapiro-Wilk normality test

data:  iq4
W = 0.9867, p-value = 6.962e-08

> shapiro.test(log(iq4))

        Shapiro-Wilk normality test

data:  log(iq4)
W = 0.9988, p-value = 0.7738
```

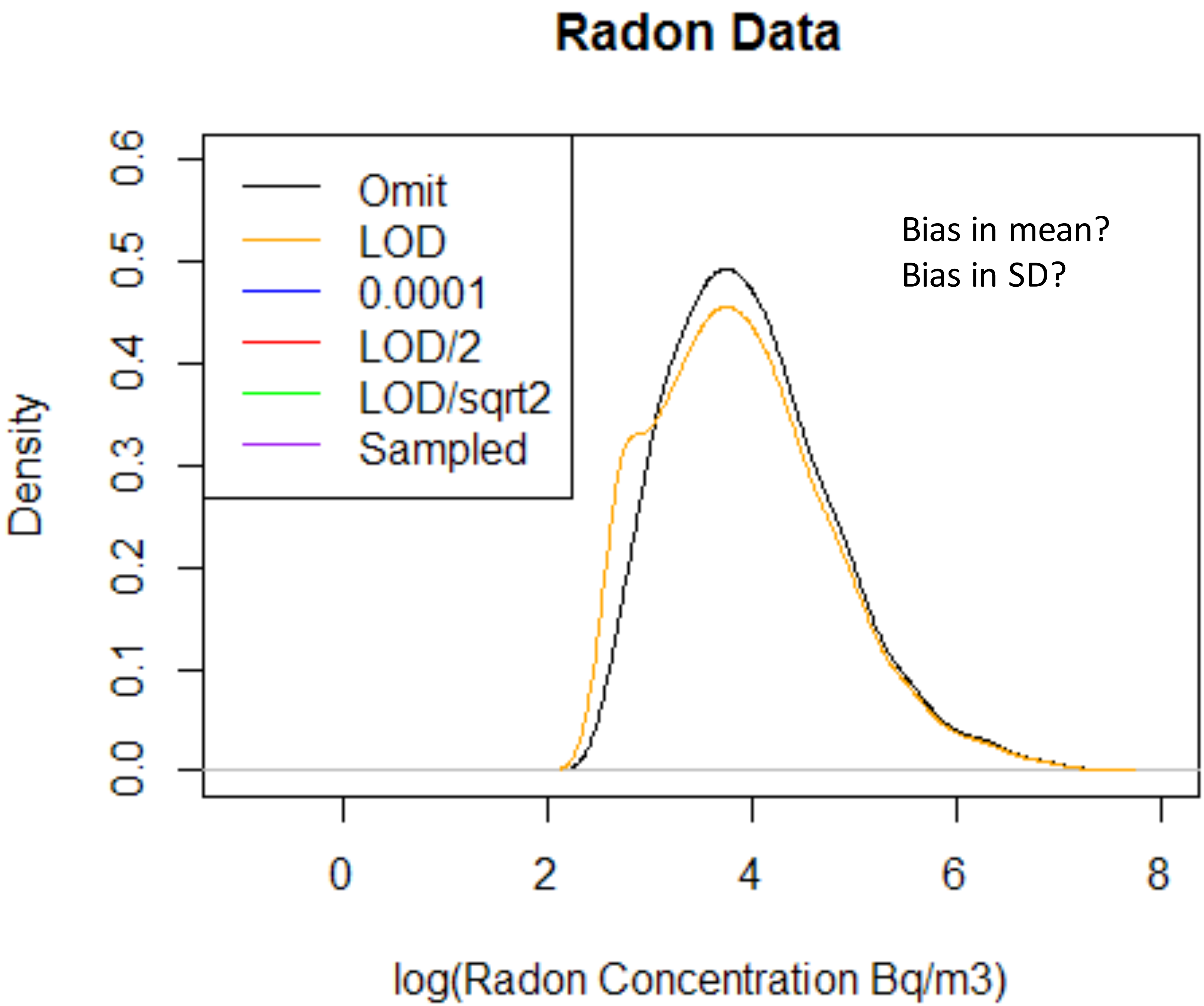
Values Below the LOD

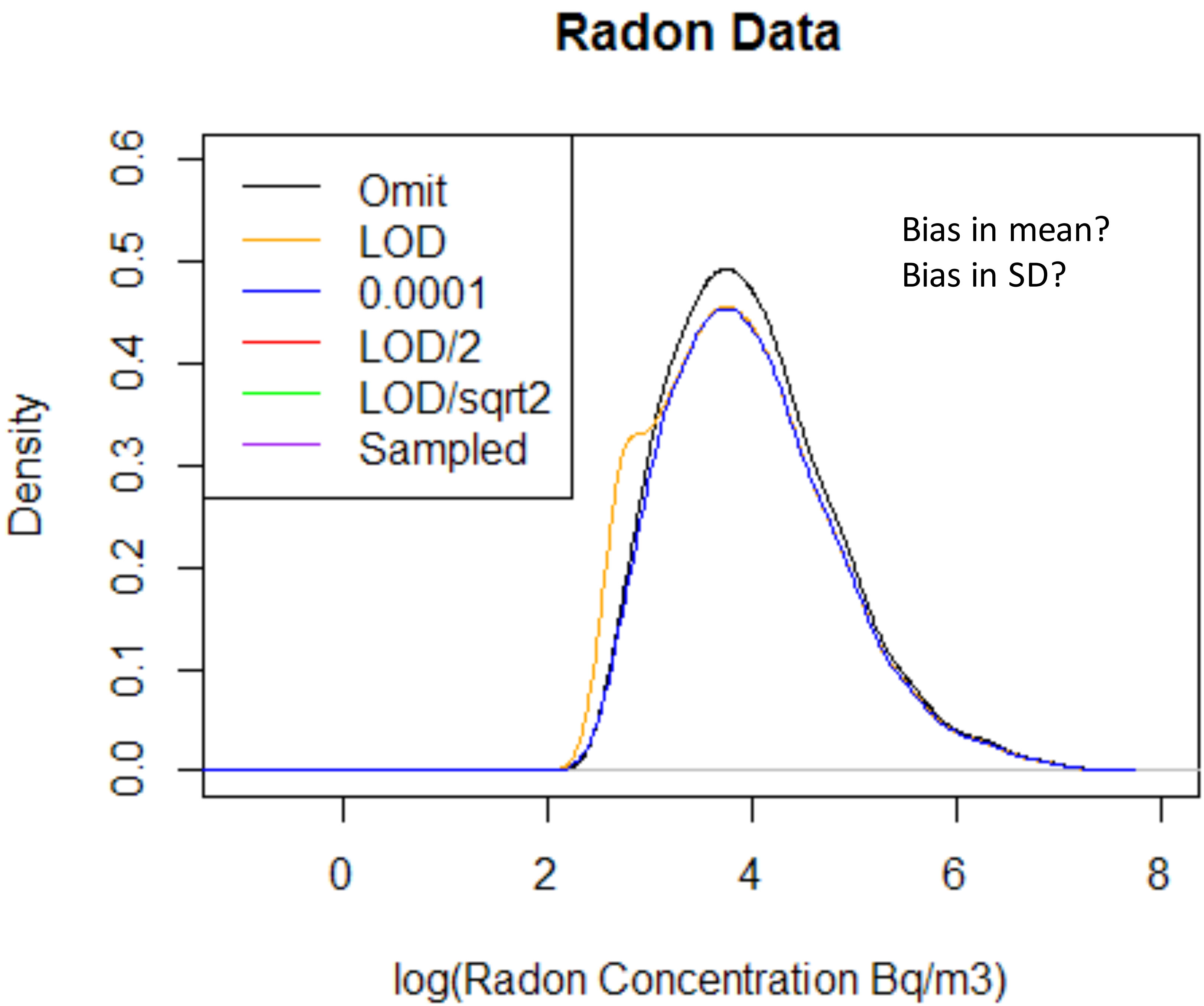
- What happens if we simply delete them from the dataset?
- What are the alternative strategies for dealing with them?
    - Use the measured values despite their lack of reliability
    - Use the LOD
    - Set them to 0 or something negligibly different from 0
    - Set them to the LOD/2
    - Set them to the LOD/sqrt(2)
    - Use some probabilistic technique based on the distribution of the values >LOD



Censored distribution not to be confused with temporal censoring!

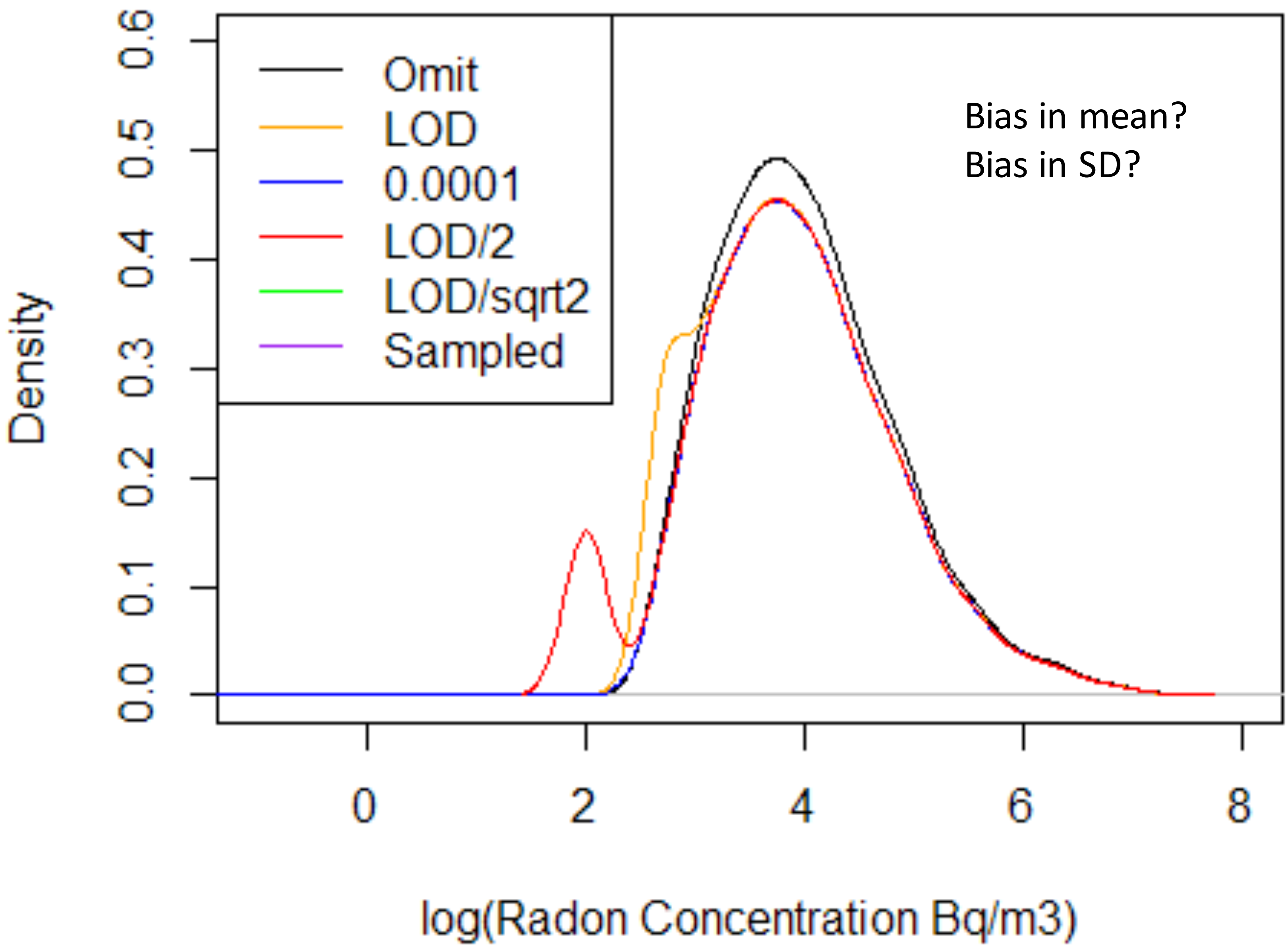

Radon Data
Omit
LOD
0.0001
LOD/2
LOD/sqrt2
Sampled
Bias in mean?
Bias in SD?
Density
0.0
0.1
0.2
0.3
0.4
0.5
0.6
0
2
4
6
8
log(Radon Concentration Bq/m3)

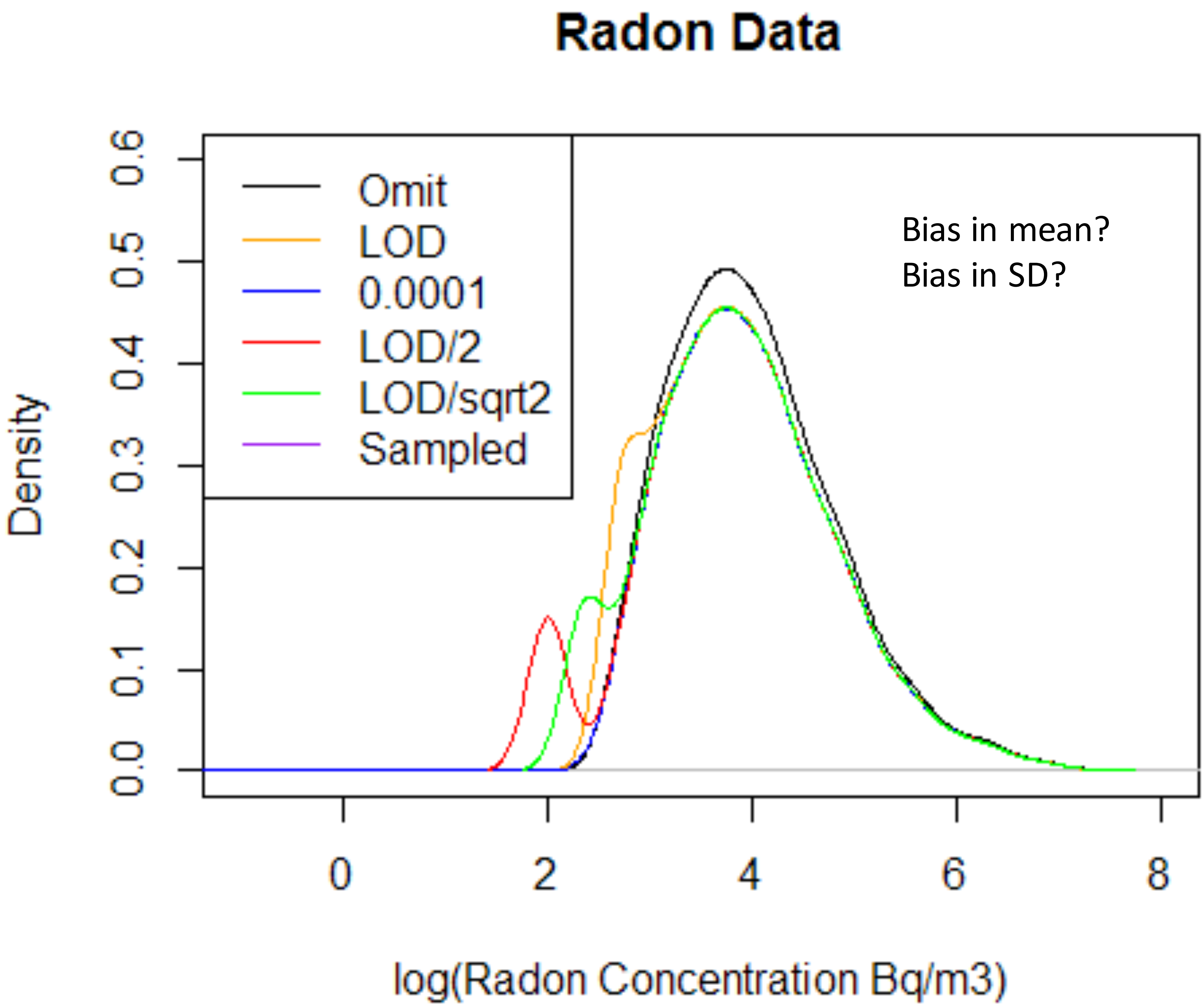

Radon Data
Omit
LOD
0.0001
LOD/2
LOD/sqrt2
Sampled
Bias in mean?
Bias in SD?
Density
0.0
0.1
0.2
0.3
0.4
0.5
0.6
0
2
4
6
8
log(Radon Concentration Bq/m3)

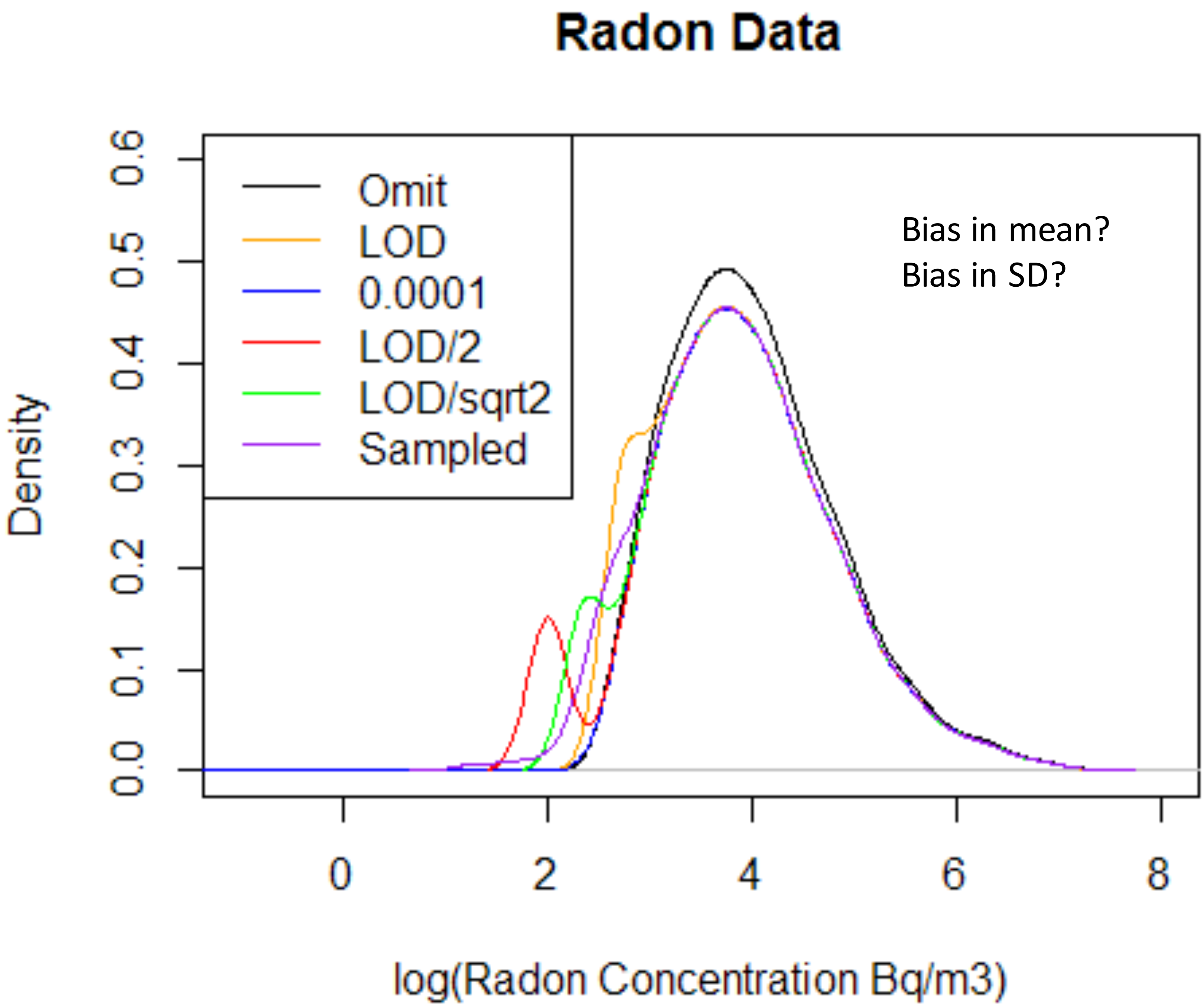


Radon Data
Omit
LOD
0.0001
LOD/2
LOD/sqrt2
Sampled
Bias in mean?
Bias in SD?
Density
0.0
0.1
0.2
0.3
0.4
0.5
0.6
0
2
4
6
8
log(Radon Concentration Bq/m3)

Radon Data


Omit
LOD
0.0001
LOD/2
LOD/sqrt2
Sampled
Bias in mean?
Bias in SD?
Density
0.0
0.1
0.2
0.3
0.4
0.5
0.6
0
2
4
6
8
log(Radon Concentration Bq/m3)


Radon Data
Omit
LOD
0.0001
LOD/2
LOD/sqrt2
Sampled
Bias in mean?
Bias in SD?
Density
0.0
0.1
0.2
0.3
0.4
0.5
0.6
0
2
4
6
8
log(Radon Concentration Bq/m3)



Radon Data
Omit
LOD
0.0001
LOD/2
LOD/sqrt2
Sampled
Bias in mean?
Bias in SD?
Density
0.0
0.1
0.2
0.3
0.4
0.5
0.6
0
2
4
6
8
log(Radon Concentration Bq/m3)

Maximum Likelihood Estimates

- An interactive statistical technique that attempts to estimate values below the LOD using what is know about values above the LOD
- Fits the most likely population mean and variance based on the observed data
- Uses that most likely distribution to estimate the unknown values
- Requires some very heavy mathematics that your computer can do for you!

What to Enter in Exposure Data Column	Exposure Data	Log Likelihood of Observation, given estimated mean & SD
Enter Observed Data:		=LN((1/((2*PI())^0.5*F$17))*EXP(-(1/2)*((LN(B17)-F$16)/F$17)^2))
Enter Observed Data:		=LN((1/((2*PI())^0.5*F$17))*EXP(-(1/2)*((LN(B18)-F$16)/F$17)^2))
Enter Observed Data:		=LN((1/((2*PI())^0.5*F$17))*EXP(-(1/2)*((LN(B19)-F$16)/F$17)^2))
Enter Observed Data:		=LN((1/((2*PI())^0.5*F$17))*EXP(-(1/2)*((LN(B20)-F$16)/F$17)^2))
Enter Observed Data:		=LN((1/((2*PI())^0.5*F$17))*EXP(-(1/2)*((LN(B21)-F$16)/F$17)^2))
Enter Observed Data:		=LN((1/((2*PI())^0.5*F$17))*EXP(-(1/2)*((LN(B22)-F$16)/F$17)^2))
Enter Observed Data:		=LN((1/((2*PI())^0.5*F$17))*EXP(-(1/2)*((LN(B23)-F$16)/F$17)^2))
Enter Observed Data:		=LN((1/((2*PI())^0.5*F$17))*EXP(-(1/2)*((LN(B24)-F$16)/F$17)^2))
Enter Observed Data:		=LN((1/((2*PI())^0.5*F$17))*EXP(-(1/2)*((LN(B25)-F$16)/F$17)^2))
Enter Observed Data:		=LN((1/((2*PI())^0.5*F$17))*EXP(-(1/2)*((LN(B26)-F$16)/F$17)^2))
Enter Observed Data:		=LN((1/((2*PI())^0.5*F$17))*EXP(-(1/2)*((LN(B27)-F$16)/F$17)^2))
Enter Observed Data:		=LN((1/((2*PI())^0.5*F$17))*EXP(-(1/2)*((LN(B28)-F$16)/F$17)^2))
Enter Observed Data:		=LN((1/((2*PI())^0.5*F$17))*EXP(-(1/2)*((LN(B29)-F$16)/F$17)^2))
Enter Observed Data:		=LN((1/((2*PI())^0.5*F$17))*EXP(-(1/2)*((LN(B30)-F$16)/F$17)^2))
Enter Observed Data:		=LN((1/((2*PI())^0.5*F$17))*EXP(-(1/2)*((LN(B31)-F$16)/F$17)^2))
Enter Observed Data:		=LN((1/((2*PI())^0.5*F$17))*EXP(-(1/2)*((LN(B32)-F$16)/F$17)^2))
Data Below LOD, Enter Detection Limit:		=LN(NORMDIST(LN(B33),F$16,F$17,TRUE))
Data Below LOD, Enter Detection Limit:		=LN(NORMDIST(LN(B34),F$16,F$17,TRUE))
Data Below LOD, Enter Detection Limit:		=LN(NORMDIST(LN(B35),F$16,F$17,TRUE))

Next Week

- Assessing the relationship between in dichotomous and continuous variable
- Box plots to visualize
- T-tests to test for differences in the means
- Hypothesis generation
- Simple linear regression
- Standard reporting