4 Modeling Data as Random Variables and Populations as Probability Distributions (Cont'd)

MTH 3240 Environmental Statistics

Spring 2020

Objectives

Objectives:

- Recognize normal and lognormal random variables.
- Obtain probabilities from normal distributions.
- Find and interpret percentiles of normal distributions

Continuous Distributions

- Recall that **continuous random variables** can take *any* value over an entire continuum.

Continuous Distributions

- Recall that **continuous random variables** can take *any* value over an entire continuum.
- Their probability distribution is represented by a smooth curve called a ***probability density function*** (or ***curve***).

Continuous Distributions

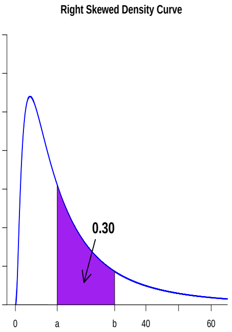
- Recall that **continuous random variables** can take *any* value over an entire continuum.
- Their probability distribution is represented by a smooth curve called a ***probability density function*** (or ***curve***).
- The **density curve** can be thought of as a smooth histogram of a **population**.

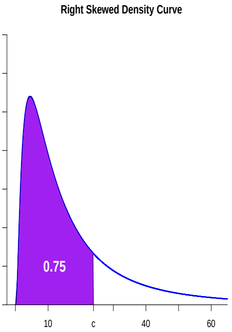
Continuous Distributions

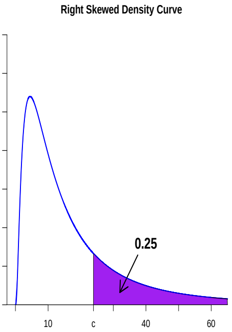
- Recall that **continuous random variables** can take *any* value over an entire continuum.
- Their probability distribution is represented by a smooth curve called a ***probability density function*** (or ***curve***).
- The **density curve** can be thought of as a smooth histogram of a **population**.

  In this case, the random variable is a measurement made on an individual **randomly** selected from the population.

- The **probability** of the random variable falling in any interval on the x-axis is the **area under the curve** over that interval.


Right Skewed Density Curve
0.30
0
a
b
40
60


Right Skewed Density Curve
0.75
10
c
40
60


Right Skewed Density Curve
0.25
10
c
40
60

Mean of a Continuous Probability Distribution

- We measure the **center** of a probability distribution by its ***mean***, denoted μ.

Mean of a Continuous Probability Distribution

- We measure the **center** of a probability distribution by its ***mean***, denoted μ.
- If the total area under a density curve was weight, μ would be the point along the x-axis at which it would balance.

Mean of a Continuous Probability Distribution

- We measure the **center** of a probability distribution by its ***mean***, denoted μ.
- If the total area under a density curve was weight, μ would be the point along the x-axis at which it would balance.
- The value of μ represents the value that the random variable takes **on average** .

- μ can be thought of as the **population mean** if the probability distribution represents a **population.**

Standard Deviation of a Continuous Probability Distribution

- We measure the **spread** in a probability distribution by its ***standard deviation***, denoted σ.

Standard Deviation of a Continuous Probability Distribution

- We measure the **spread** in a probability distribution by its ***standard deviation***, denoted σ.
- A larger value of σ corresponds to a more spread-out probability distribution.

Standard Deviation of a Continuous Probability Distribution

- We measure the **spread** in a probability distribution by its ***standard deviation***, denoted σ.
- A larger value of σ corresponds to a more spread-out probability distribution.
- The value of σ represents a **typical deviation** of the random variable away from μ.

- σ can be thought of as the **population standard deviation** if the probability distribution represents a **population.**

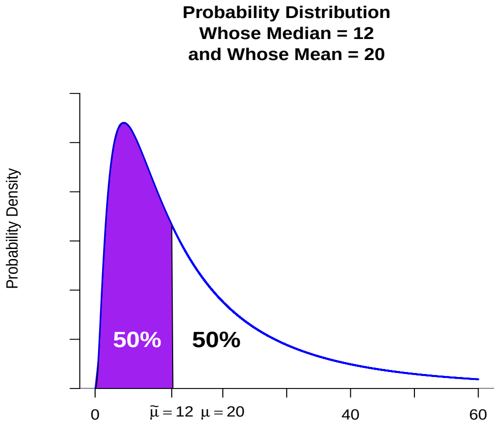
Percentiles of a Continuous Probability Distribution

- The ***median***, or ***50th percentile***, of a continuous distribution, denoted by $\tilde{\mu}$, is the value below which 50% of the population lies (and above which the other 50% lies).

Percentiles of a Continuous Probability Distribution

- The ***median***, or ***50th percentile***, of a continuous distribution, denoted by $\tilde{\mu}$, is the value below which 50% of the population lies (and above which the other 50% lies).
- Thus the variable X has a 50/50 chance of falling above or below $\tilde{\mu}$.

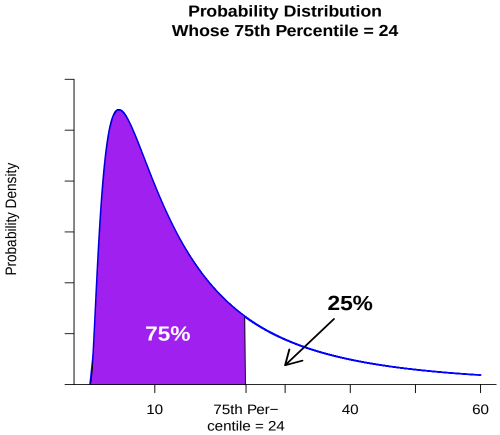
- Whereas the mean μ is the "balancing point" of a distribution, the median $\tilde{\mu}$ is the "equal areas point".


Probability Distribution
Whose Median = 12
and Whose Mean = 20
Probability Density
50%
50%
0
$\tilde{\mu} = 12$
$\mu = 20$
40
60

- In general:
  - For a **symmetric** distribution, mean and median will be the same, i.e. $\mu = \tilde{\mu}$.
  - For a **right skewed** distribution, the mean will be greater than the median, i.e. $\mu > \tilde{\mu}$.

- Other percentiles are defined analogously.

  For example, the 75***th percentile*** is the value below which 75% of the population lies.


Probability Distribution
Whose 75th Percentile = 24
Probability Density
75%
25%
10
75th Per–
centile = 24
40
60

- **Example**: A river's annual peak height is a random variable, X.

- **Example**: A river's annual peak height is a random variable, X.

  The **"100-year flood level"** is the height for which there's only a **1 in 100 chance**, or **0.01 probability**, of being exceeded in any given year.

- **Example**: A river's annual peak height is a random variable, X.

  The **"100-year flood level"** is the height for which there's only a **1 in 100 chance**, or **0.01 probability**, of being exceeded in any given year.

  So the **"100-year flood level"** is the **99th percentile** of the distribution of X.

Theoretical Continuous Probability Distributions

- In the absence of accurate information about the shape of a population's histogram, we have to choose from a set of stock **theoretical density curves** the one that we *think* describes the population.

Theoretical Continuous Probability Distributions

- In the absence of accurate information about the shape of a population's histogram, we have to choose from a set of stock **theoretical density curves** the one that we *think* describes the population.

  We'll look at two:

  1. The **normal** distribution.
  2. The **lognormal** distribution.

The Normal Distribution

- Many variables follow the bell-shaped ***normal distribution***.

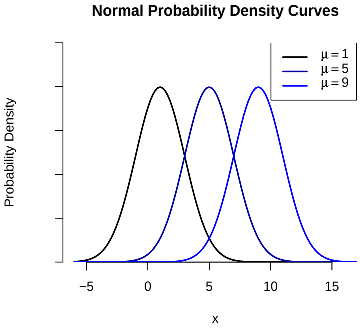
The Normal Distribution

- Many variables follow the bell-shaped ***normal distribution***.
- Its **mean** μ and **standard deviation** σ determine, respectively, the center and spread of the distribution.

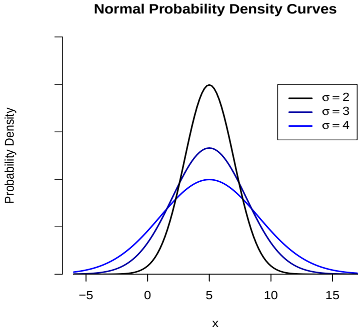
The Normal Distribution

- Many variables follow the bell-shaped ***normal distribution***.
- Its **mean** μ and **standard deviation** σ determine, respectively, the center and spread of the distribution.

  They're referred to as the ***parameters*** of the distribution.



Normal Probability Density Curves
Probability Density
x
−5
0
5
10
15
μ = 1
μ = 5
μ = 9



Normal Probability Density Curves
Probability Density
x
−5
0
5
10
15
σ = 2
σ = 3
σ = 4

- We'll use the notation

$$\boldsymbol{X} \sim \textbf{N}(\boldsymbol{\mu},\, \boldsymbol{\sigma})$$

to mean that the random variable X follows a normal distribution with mean μ and standard deviation σ.

- We'll use the notation

$$\boldsymbol{X} \sim \mathbf{N}(\boldsymbol{\mu},\, \boldsymbol{\sigma})$$

to mean that the random variable X follows a normal distribution with mean μ and standard deviation σ.

- Because the distribution is symmetric, the **median** of the **normal distribution** is also $\boldsymbol{\mu}$.

- Normal distribution probabilities (areas under the curve) $P(a < X < b)$ can be obtained using either of the following:
  - A table (the so-called Z table)
  - Statistical software

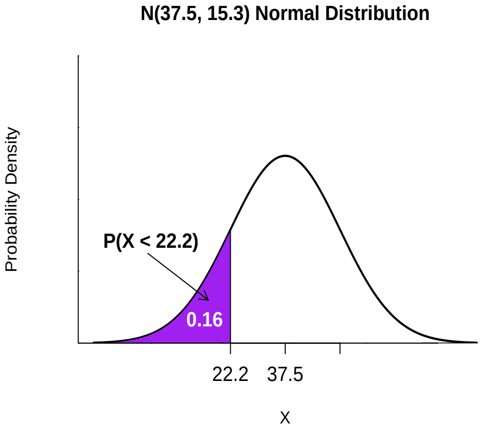
Example

A study suggests that **blood glucose levels** in *johnny darter* fish follow a **normal** distribution with mean **37.5** mg/100 ml and standard deviation **15.3** mg/100 ml.

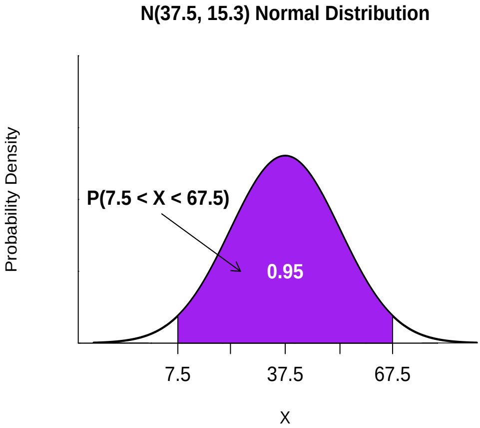
Example

A study suggests that **blood glucose levels** in *johnny darter* fish follow a **normal** distribution with mean **37.5** mg/100 ml and standard deviation **15.3** mg/100 ml.

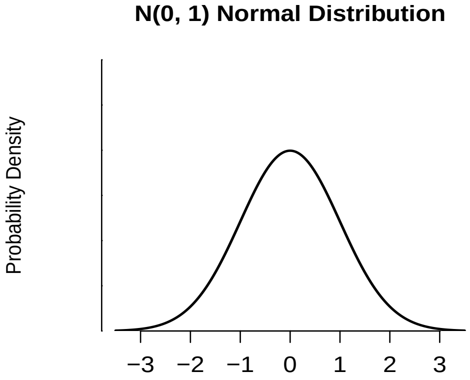
The **probability** that the glucose level in a randomly selected fish will be **below 22.2** mg/100 ml is **0.16** (obtained using software and depicted on the next slide).


N(37.5, 15.3) Normal Distribution
Probability Density
P(X < 22.2)
0.16
22.2
37.5
X

The **probability** that the glucose level will be **within 30** of the **mean** is **0.95** (obtained using software and depicted on the next slide).


N(37.5, 15.3) Normal Distribution
Probability Density
P(7.5 < X < 67.5)
0.95
7.5
37.5
67.5
X

- The normal distribution with $\mu = 0$ and $\sigma = 1$ is called the ***standard normal distribution*** and denoted $\mathbf{N}(0,\ 1)$.


N(0, 1) Normal Distribution
Probability Density
−3 −2 −1 0 1 2 3

- We can convert **any normal random variable** to a **standard normal** one using the following fact.

- We can convert **any normal random variable** to a **standard normal** one using the following fact.

Fact: If $X \sim \mathrm{N}(\mu, \sigma)$, and we convert X to a variable Z via

$$Z = \frac{X - \mu}{\sigma},$$

then $Z \sim \mathrm{N}(0, 1)$.

- When we convert a value X to a value Z, we say that X has been ***standardized***, or converted to a ***z-score***.

- When we convert a value X to a value Z, we say that X has been ***standardized***, or converted to a ***z-score***.

  A **standardized value**, or z**-score**, is measured in *standard deviations away from the mean*, or ***standard units***.

- When we convert a value X to a value Z, we say that X has been ***standardized***, or converted to a ***z-score***.

  A **standardized value**, or z**-score**, is measured in *standard deviations away from the mean*, or ***standard units***.

  It will be **positive** or **negative** depending on whether $\boldsymbol{X}$ is **above** or **below** the **mean.**

Example

Recall that **blood glucose levels** in *johnny darter* fish follow a **normal** distribution with mean **37.5** mg/100 ml and standard deviation **15.3** mg/100 ml.

Example

Recall that **blood glucose levels** in *johnny darter* fish follow a **normal** distribution with mean **37.5** mg/100 ml and standard deviation **15.3** mg/100 ml.

If one of these fish has a glucose level of $\boldsymbol{X = 60.4}$, its z**-score** is

$$Z = \frac{60.4 - 37.5}{15.3} = \mathbf{1.5},$$

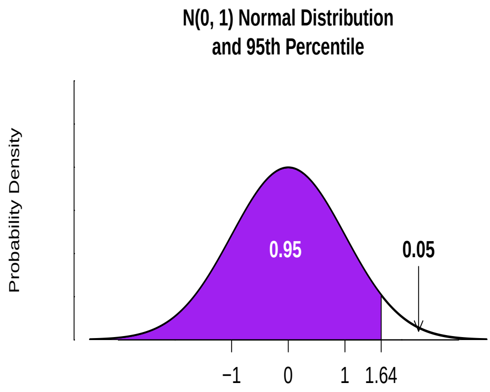
so the fish is a one and a half standard deviations above the mean.

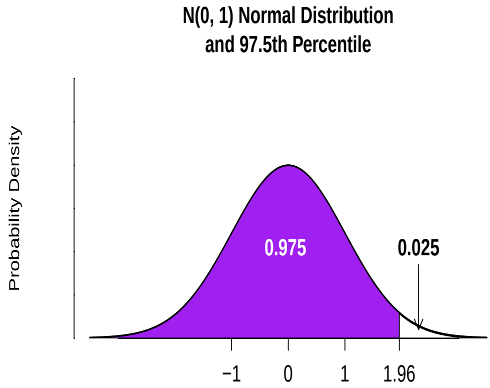
- Some **percentiles** of the $\mathbf{N}(0,\ 1)$ distribution are shown below.

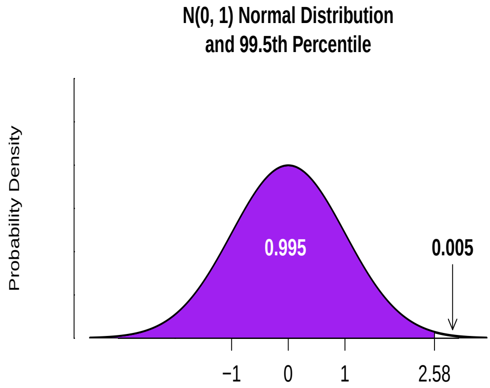
- Some **percentiles** of the **N**$(0, 1)$ distribution are shown below.

N$(0, 1)$ **Percentiles**

50th	0.00
95th	1.64
97.5th	1.96
99.5th	2.58


N(0, 1) Normal Distribution
and 95th Percentile
Probability Density
0.95
0.05
−1
0
1
1.64


N(0, 1) Normal Distribution
and 97.5th Percentile
Probability Density
0.975
0.025
−1
0
1
1.96

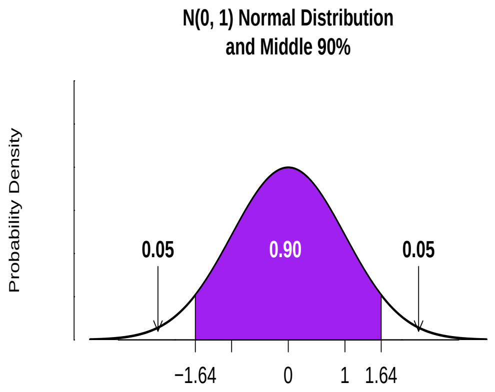

N(0, 1) Normal Distribution
and 99.5th Percentile
Probability Density
0.995
0.005
−1
0
1
2.58

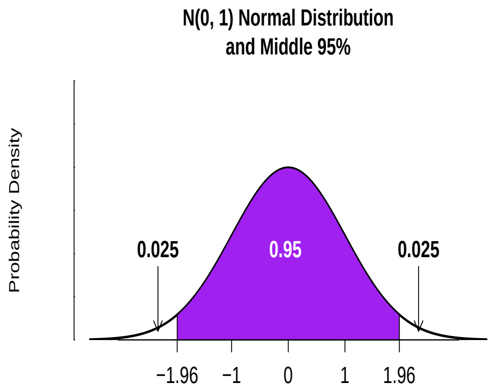
- We can use **percentiles** to characterize **middle percentages** of the $\mathbf{N}(0,\ 1)$ distribution.

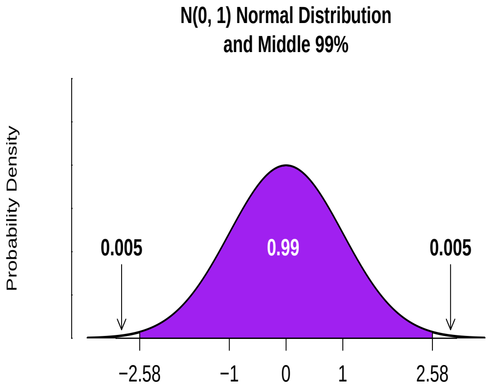
- We can use **percentiles** to characterize **middle percentages** of the **N**$(0, 1)$ distribution.

N$(0, 1)$ **Percentiles (Cont'd)**

Middle 90%	Between $\pm$1.64
Middle 95%	Between $\pm$1.96
Middle 99%	Between $\pm$2.58


N(0, 1) Normal Distribution
and Middle 90%
Probability Density
0.05
0.90
0.05
−1.64
0
1
1.64


N(0, 1) Normal Distribution
and Middle 95%
Probability Density
0.025
0.95
0.025
−1.96
−1
0
1
1.96


N(0, 1) Normal Distribution
and Middle 99%
Probability Density
0.005
0.99
0.005
−2.58
−1
0
1
2.58

- A **percentile** of a **N**$(\mu,\ \sigma)$ distribution is obtained by **"unstandardizing"** the corresponding percentile of the **N**$(0,\ 1)$ distribution using the following.

- A **percentile** of a $\mathbf{N}(\boldsymbol{\mu},\ \boldsymbol{\sigma})$ distribution is obtained by **"unstandardizing"** the corresponding percentile of the $\mathbf{N}(\mathbf{0},\ \mathbf{1})$ distribution using the following.

Percentiles of a Normal Distribution: A percentile x of a N$(\mu,\ \sigma)$ distribution is

$$x \;=\; \mu + z\sigma,$$

where z is the corresponding percentile of the N$(0,\ 1)$ distribution.

Exercise

A study suggests that **blood glucose levels** in *johnny darter* fish follow a **normal** distribution with mean **37.5** mg/100 ml and standard deviation **15.3** mg/100 ml.

Exercise

A study suggests that **blood glucose levels** in *johnny darter* fish follow a **normal** distribution with mean **37.5** mg/100 ml and standard deviation **15.3** mg/100 ml.

Solve the following problems by ”unstandardizing” appropriate $N(0, 1)$ percentiles.

Exercise

A study suggests that **blood glucose levels** in *johnny darter* fish follow a **normal** distribution with mean **37.5** mg/100 ml and standard deviation **15.3** mg/100 ml.

Solve the following problems by ”unstandardizing” appropriate N$(0,\ 1)$ percentiles.

a) Find the glucose level below which **97.5%** of glucose levels fall (that is, the **97.5th percentile** of the distribution).

Exercise

A study suggests that **blood glucose levels** in *johnny darter* fish follow a **normal** distribution with mean **37.5** mg/100 ml and standard deviation **15.3** mg/100 ml.

Solve the following problems by ”unstandardizing” appropriate N$(0,\ 1)$ percentiles.

a) Find the glucose level below which **97.5%** of glucose levels fall (that is, the **97.5th percentile** of the distribution).

b) Find the **two** glucose levels **between** which the **middle 95%** of glucose levels fall.

The Lognormal Distribution

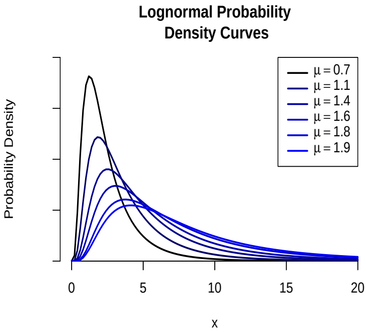
- Environmental quantities such as pollutant concentrations often follow **right skewed** distributions.

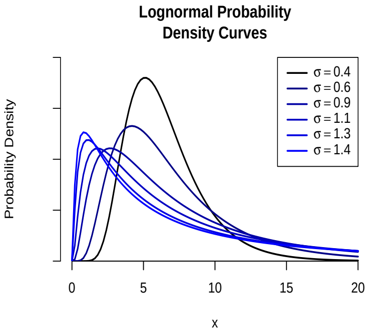
The Lognormal Distribution

- Environmental quantities such as pollutant concentrations often follow **right skewed** distributions.
- A useful *theoretical density curve* for **right skewed** populations is the ***lognormal distribution***.

The Lognormal Distribution

- Environmental quantities such as pollutant concentrations often follow **right skewed** distributions.
- A useful *theoretical density curve* for **right skewed** populations is the ***lognormal distribution***.
- **Lognormal distributions** are **right skewed** and lie entirely to the **right of zero**.


Lognormal Probability
Density Curves
Probability Density
x
0
5
10
15
20
μ = 0.7
μ = 1.1
μ = 1.4
μ = 1.6
μ = 1.8
μ = 1.9


Lognormal Probability
Density Curves
Probability Density
x
0
5
10
15
20
σ = 0.4
σ = 0.6
σ = 0.9
σ = 1.1
σ = 1.3
σ = 1.4

- We write

$$\boldsymbol{X} \sim \textbf{LN}(\boldsymbol{\mu}, \boldsymbol{\sigma})$$

to mean that X is a random variable that follows a lognormal distribution with ***parameters*** $\boldsymbol{\mu}$ and $\boldsymbol{\sigma}$

- We write

$$\boldsymbol{X} \sim \textbf{LN}(\boldsymbol{\mu},\, \boldsymbol{\sigma})$$

  to mean that X is a random variable that follows a lognormal distribution with ***parameters*** $\boldsymbol{\mu}$ and $\boldsymbol{\sigma}$

- The following fact explains how the lognormal distribution gets its name.

Fact: If $X \sim$ LN(μ, σ), and we make the (natural) log transformation

$$Y = \log(X),$$

then Y is a new random variable, and

$$Y \sim \mathsf{N}(\mu, \sigma).$$

Fact: If $X \sim$ LN(μ, σ), and we make the (natural) log transformation

$$Y = \log(X),$$

then Y is a new random variable, and

$$Y \sim \mathsf{N}(\mu, \sigma).$$

- In words, if X is ***lognormal***, then ***its log is normal***.

Fact: If $X \sim \text{LN}(\mu, \sigma)$, and we make the (natural) log transformation

$$Y = \log(X),$$

then Y is a new random variable, and

$$Y \sim \text{N}(\mu, \sigma).$$

- In words, if X is ***lognormal***, then ***its log is normal***.

  Thus **we can convert** a ***lognormal*** variable to a ***normal*** one by taking it's **log.**

Example

To illustrate the effect of the making the **log transformation** on **right skewed**, **lognormal data**, the following $n = 50$ observations were obtained from a **LN**$(5, 1)$ distribution using a computer random number generator.

Example

To illustrate the effect of the making the **log transformation** on **right skewed**, **lognormal data**, the following $n = 50$ observations were obtained from a **LN**$(5, 1)$ distribution using a computer random number generator.

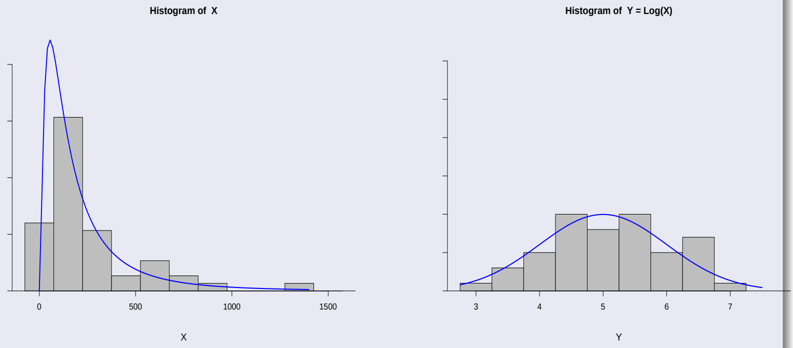
202.7	347.2	300.5	812.3	38.6	83.9	157.5	35.3	180.6	152.4
90.4	95.5	234.7	618.9	149.2	169.6	427.6	89.1	204.3	90.9
681.5	55.4	625.5	45.7	68.9	828.4	21.3	561.4	315.8	97.4
95.6	69.5	650.0	77.1	367.1	49.2	478.9	182.3	273.8	33.2
313.9	107.9	86.4	287.3	194.3	203.0	164.9	1307.0	209.4	164.7

Example

To illustrate the effect of the making the **log transformation** on **right skewed**, **lognormal data**, the following $n = 50$ observations were obtained from a **LN**$(5, 1)$ distribution using a computer random number generator.

202.7	347.2	300.5	812.3	38.6	83.9	157.5	35.3	180.6	152.4
90.4	95.5	234.7	618.9	149.2	169.6	427.6	89.1	204.3	90.9
681.5	55.4	625.5	45.7	68.9	828.4	21.3	561.4	315.8	97.4
95.6	69.5	650.0	77.1	367.1	49.2	478.9	182.3	273.8	33.2
313.9	107.9	86.4	287.3	194.3	203.0	164.9	1307.0	209.4	164.7

A histogram of these observations is shown below on the left along with the **LN**$(5, 1)$ density curve.


Histogram of X
0
500
1000
1500
X
Histogram of Y = Log(X)
3
4
5
6
7
Y

After making the log transformation of each of the 50 observations, the data values (now on the log scale) are:

After making the log transformation of each of the 50 observations, the data values (now on the log scale) are:

5.31	5.85	5.71	6.70	3.65	4.43	5.06	3.57	5.20	5.03
4.50	4.56	5.46	6.43	5.01	5.13	6.06	4.49	5.32	4.51
6.52	4.01	6.44	3.82	4.23	6.72	3.06	6.33	5.76	4.58
4.56	4.24	6.48	4.35	5.91	3.90	6.17	5.21	5.61	3.50
5.75	4.68	4.46	5.66	5.27	5.31	5.11	7.18	5.34	5.10

After making the log transformation of each of the 50 observations, the data values (now on the log scale) are:

5.31	5.85	5.71	6.70	3.65	4.43	5.06	3.57	5.20	5.03
4.50	4.56	5.46	6.43	5.01	5.13	6.06	4.49	5.32	4.51
6.52	4.01	6.44	3.82	4.23	6.72	3.06	6.33	5.76	4.58
4.56	4.24	6.48	4.35	5.91	3.90	6.17	5.21	5.61	3.50
5.75	4.68	4.46	5.66	5.27	5.31	5.11	7.18	5.34	5.10

A histogram of these log-transformed values along with the $\mathbf{N}(5,\ 1)$ curve is shown on the right in previous slide.

After making the log transformation of each of the 50 observations, the data values (now on the log scale) are:

5.31	5.85	5.71	6.70	3.65	4.43	5.06	3.57	5.20	5.03
4.50	4.56	5.46	6.43	5.01	5.13	6.06	4.49	5.32	4.51
6.52	4.01	6.44	3.82	4.23	6.72	3.06	6.33	5.76	4.58
4.56	4.24	6.48	4.35	5.91	3.90	6.17	5.21	5.61	3.50
5.75	4.68	4.46	5.66	5.27	5.31	5.11	7.18	5.34	5.10

A histogram of these log-transformed values along with the **N**$(5,\ 1)$ curve is shown on the right in previous slide.

The **log-transformed data** can be treated as a random sample from a **N**$(5,\ 1)$ distribution.

- Note that the **parameters** μ and σ of the **LN**(μ, σ) distribution refer to the **mean** and **standard deviation** of the **N**(μ, σ) distribution that results **after** making the log-transformation.

- Note that the **parameters** μ and σ of the **LN**$(\boldsymbol{\mu}, \boldsymbol{\sigma})$ distribution refer to the **mean** and **standard deviation** of the **N**$(\boldsymbol{\mu}, \boldsymbol{\sigma})$ distribution that results **after** making the log-transformation.

  In other words, μ and σ *aren't* the mean and standard deviation of the original **LN**$(\boldsymbol{\mu}, \boldsymbol{\sigma})$ distribution.