WEEK 1 - LECTURE 2

Summarizing Data (cont.)

OUTLINE

- Density Curves
- Empirical Cumulative Distribution Function (ECDF)
- Survival Function
- Quantile-Quantile Plots

Denstiy Curve

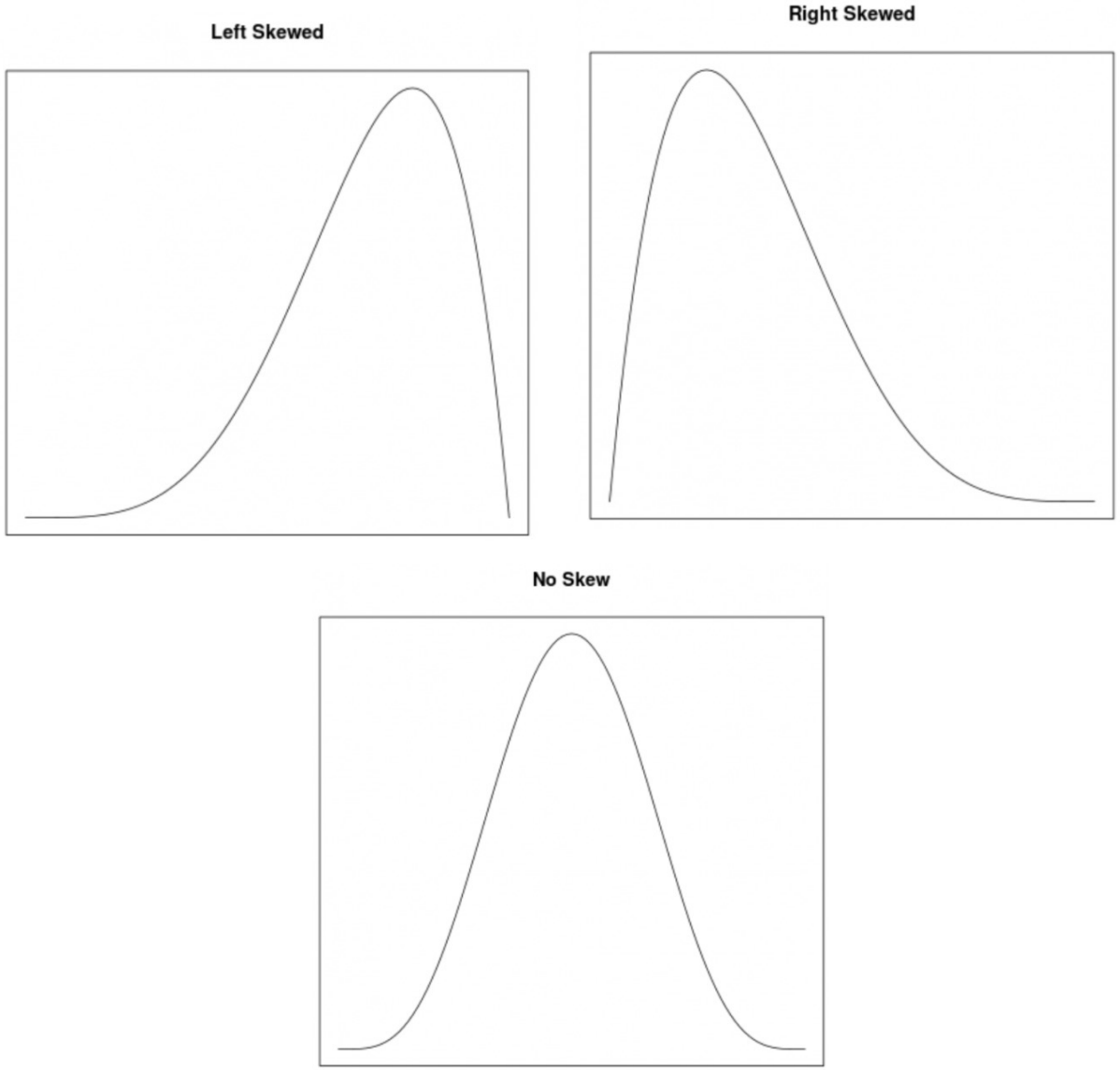
- A density curve is a curve on a graph that represents the distribution of values in a dataset.
- Density curves are useful for three reasons:

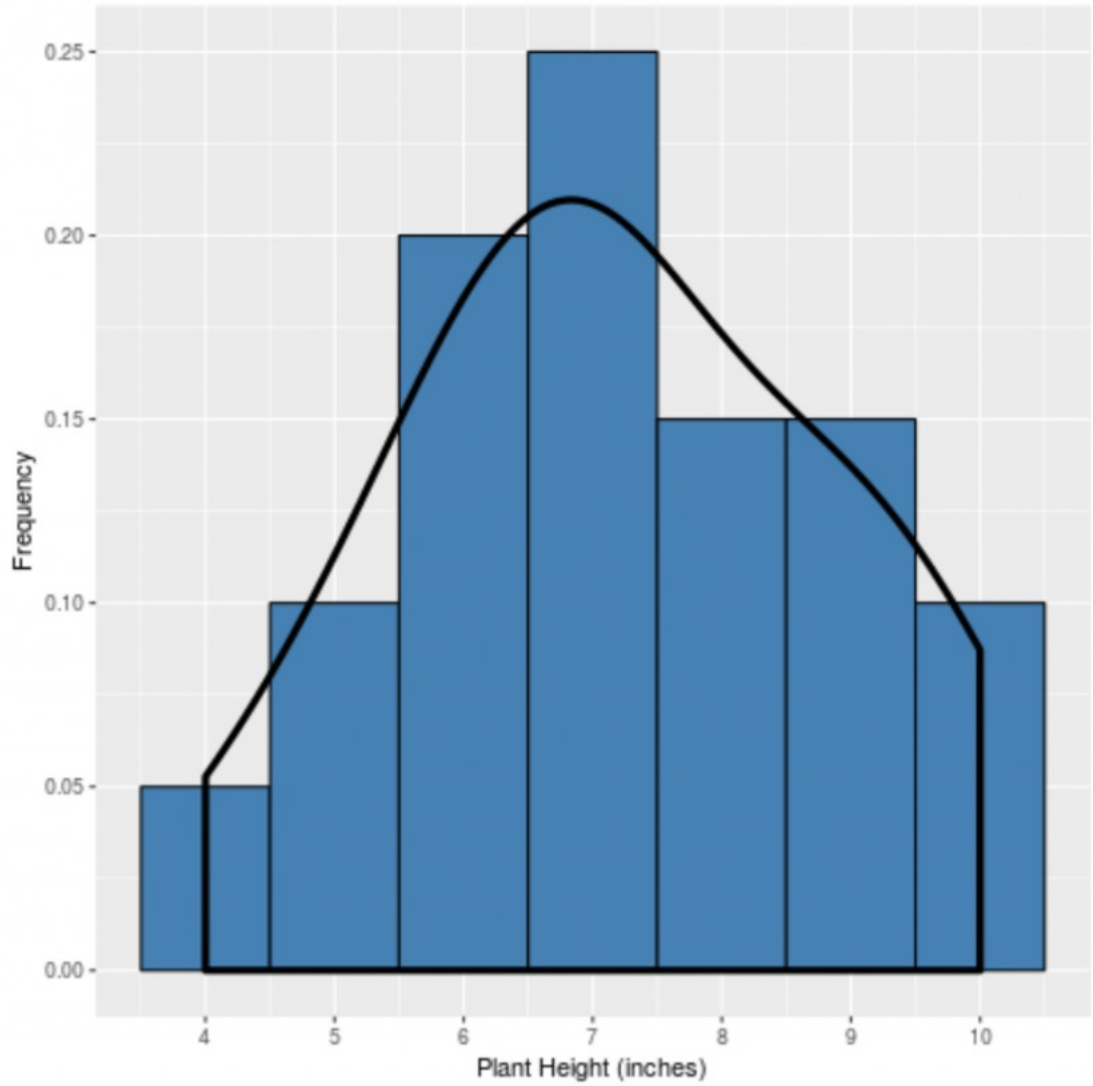
1. A density curve gives us a good idea of the "shape" of a distribution, including whether or not a distribution has one or more "peaks" of frequently occurring values and whether or not the distribution is skewed to the left or the right.

2. A density curve lets us visually see where the mean and the median of a distribution are located.

3. A density curve lets us visually see what percentage of observations in a dataset fall between different values.

Denstiy Curve

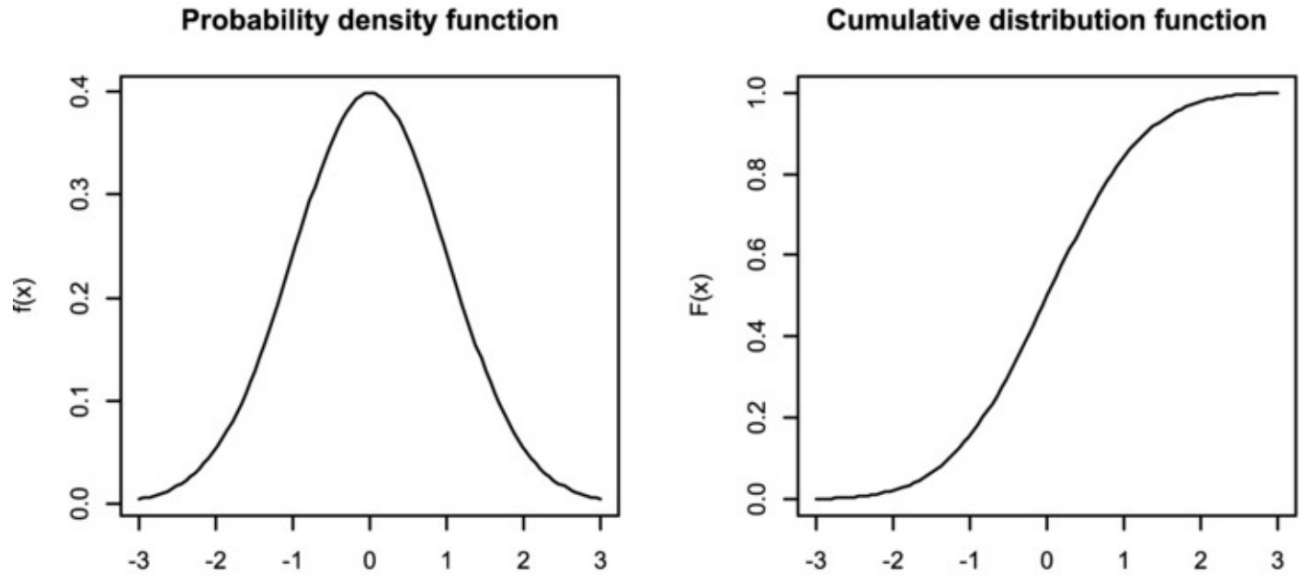




Cumulative Distribution Function

- A cumulative distribution function (CDF) describes the probabilities of a random variable having values less than or equal to x. It is a cumulative function because it sums the total likelihood up to that point. Its output always ranges between 0 and 1.

$$CDF(x) = P(X \leq x)$$

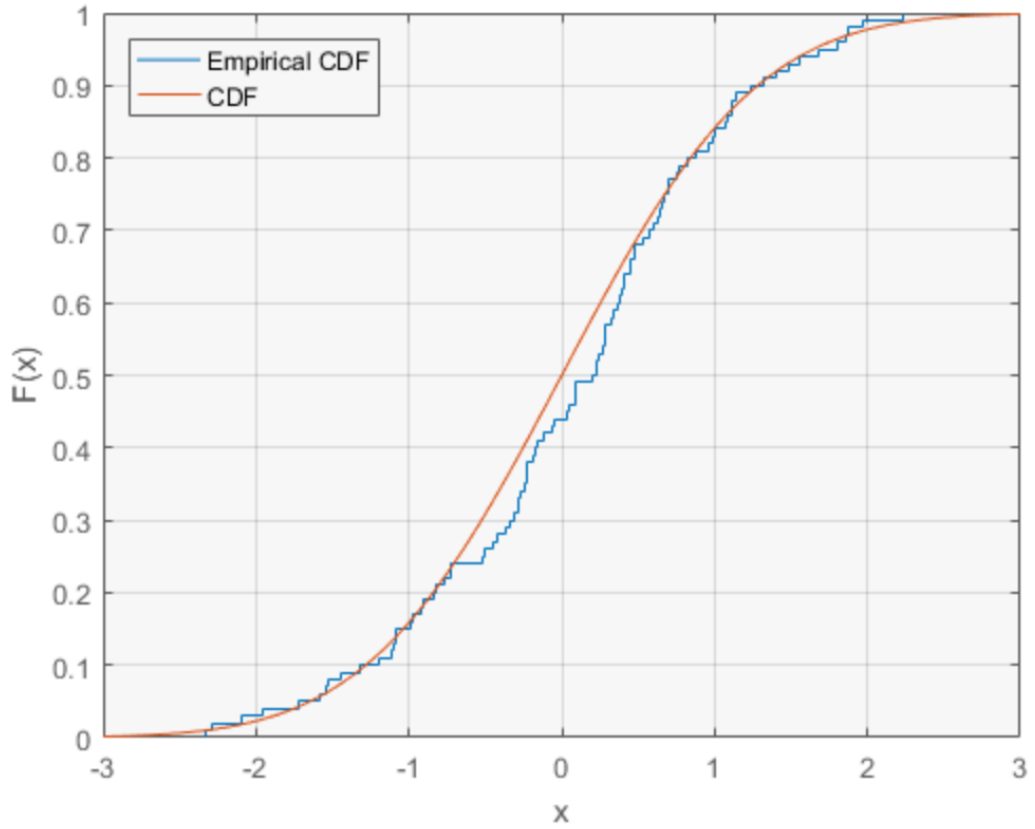


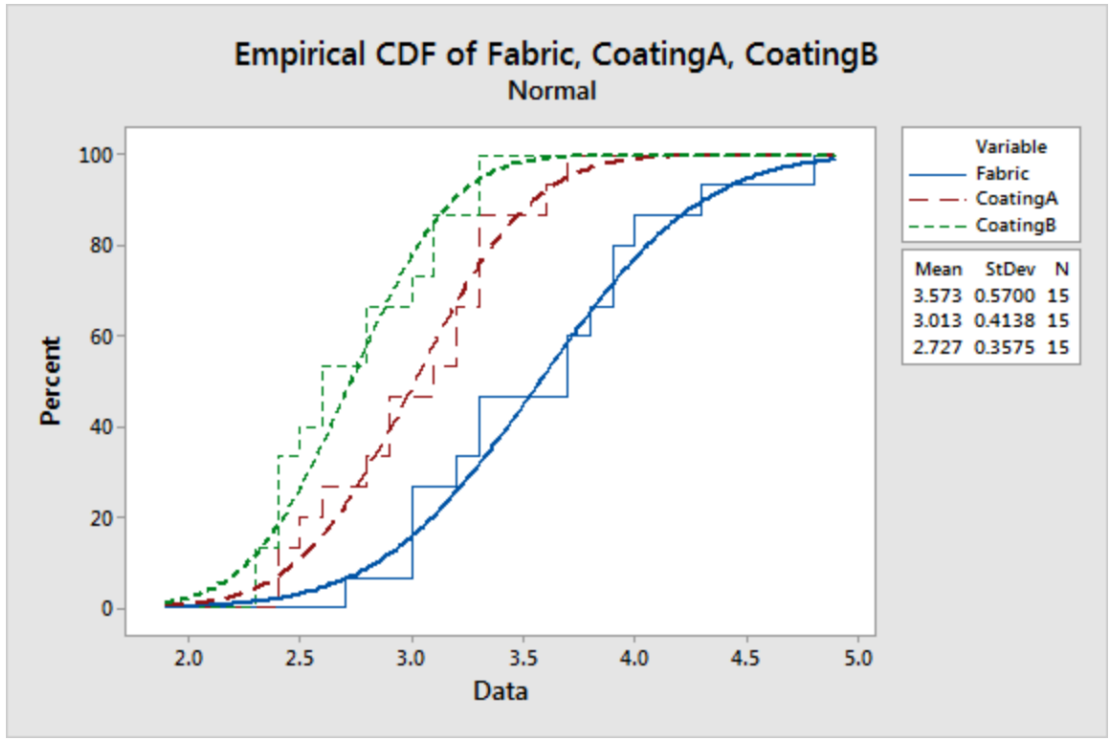
Empirical Cumulative Distribution Function (ECDF)

- An empirical cumulative distribution function plot is used to display the data points in a sample from lowest to highest against their percentiles. **These graphs require continuous variables** and allow to derive percentiles and other distribution properties.
- Use an ECDF plot to assess the following features of your dataset:
  - Percentiles and proportions for data ranges.
  - Identify where most values occur.
  - Assess the range of your data.
  - Compare sample distributions.
  - Determine how well your data follow a fitted distribution.

Empirical Cumulative Distribution Function (cont.)

$$ECDF(x) = \frac{1}{n}(\#x_i \leq x)$$





Empirical Cumulative Distribution Function (cont.)

Example:

1) Plot the ECDF of this batch of numbers: 1, 14, 10, 9, 11, 9.
2) Plot the ECDF of the following sample.

{−15.4, −8.8, 8.2, 3.4, −7.1, 4.5, −12.7, 5.2, −10.6, −11.2}

Empirical Cumulative Distribution Function

- In the case $X_1, X_2, \ldots, X_n$ is a random sample from a continuous distribution function F, then

$$F_n(x) = \frac{1}{n} \sum_{i=1}^{n} I_{(-\infty,x]}(X_i)$$

where

$$I_{(-\infty,x]}(X_i) = \begin{cases} 1, & \text{if } X_i \leq x \\ 0, & \text{if } X_i > x \end{cases}$$

$$E[F_n(x)] = F(x)$$

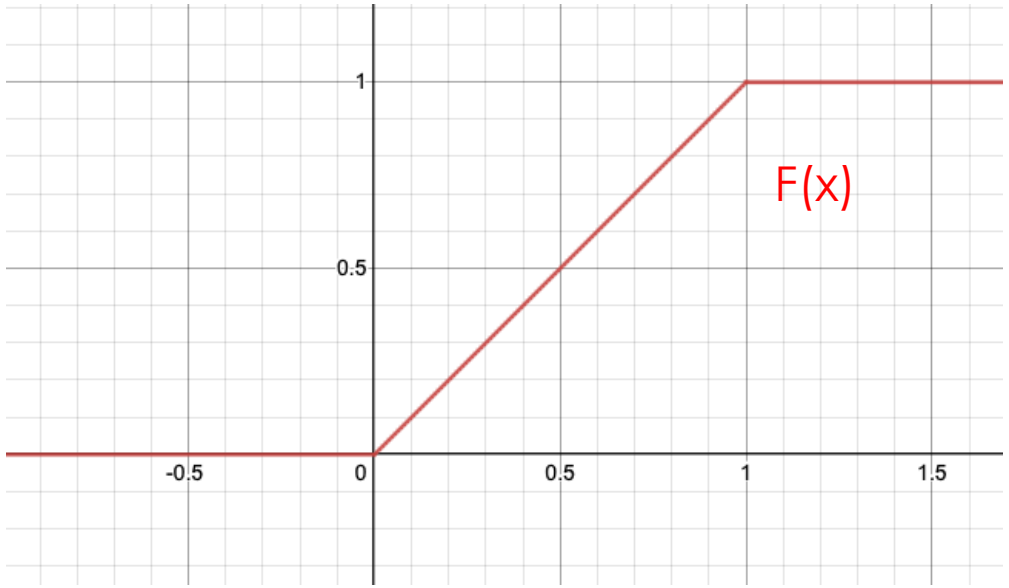
$$\text{Var}[F_n(x)] = \frac{1}{n} F(x)[1 - F(x)]$$

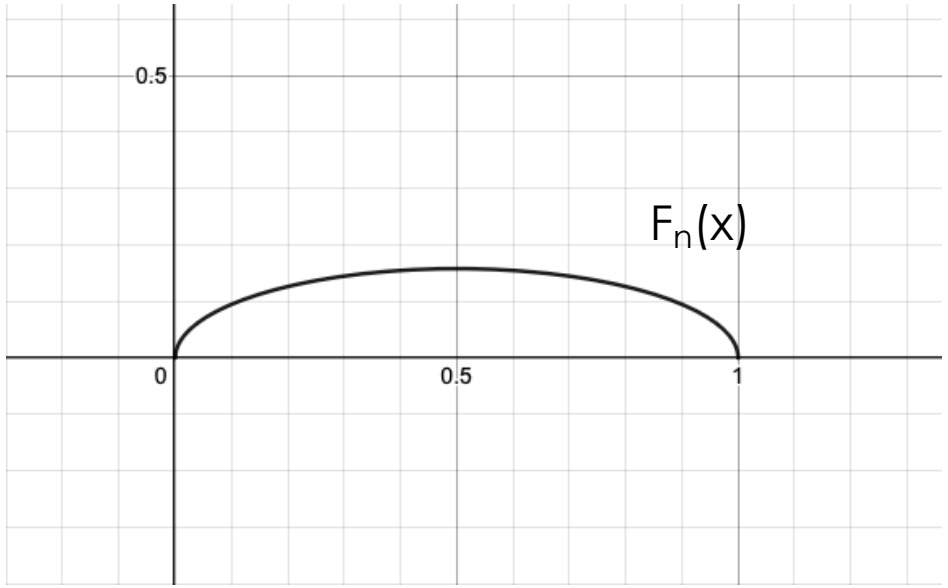
Empirical Cumulative Distribution Function (cont.)

Example:

Suppose that $X_1, X_2, \ldots, X_n$ are independent U[0,1] random variables.

Sketch F(x) and the standard deviation of $F_n(x)$.

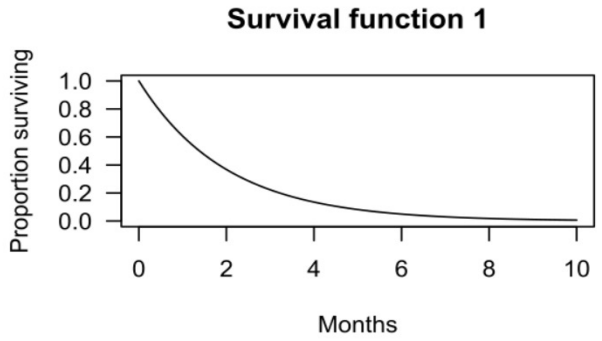


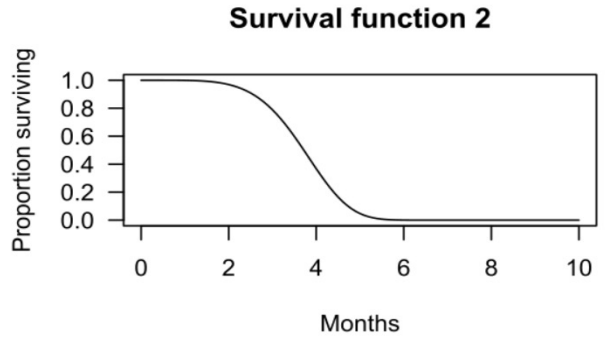


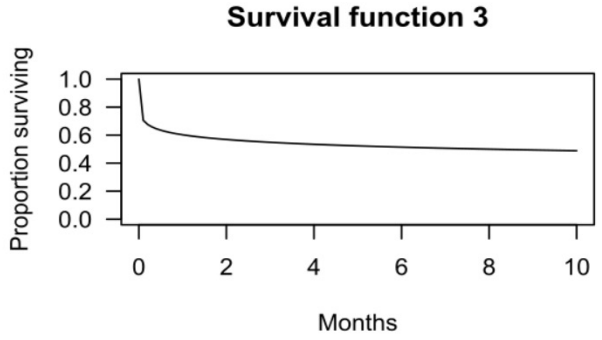
Survival Function

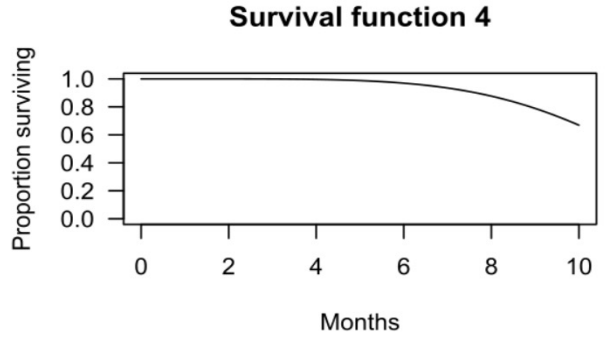
- The survival function is a function that gives the probability that a patient, device, or other object of interest will survive past a certain time. Survival functions are most often used in reliability and related fields.

$$S(t) = P(T > t) = 1 - CDF(t)$$









Hazard Function

- The hazard function is defined as

$$h(t) = \frac{f(t)}{1 - F(t)} = -\frac{d}{dx} \log S(t)$$

where f is the density function and F is the CDF of f.

- The hazard function may be thought of as the instantaneous rate of mortality for an individual alive at time t. If T is the lifetime of a manufactured component, *h(t)* may be thought of as the instantaneous or age-specific failure rate.

Hazard Function (cont.)

Example:

1) Consider a building that has 1200 electric switches

   a) Suppose 4 switches fail in the first three months. What is the hazard rate per month per unit during this period?

   b) In the next two months, 6 more switches fail. What is the hazard rate during this period?

   c) Data are collected till ten months and put in the below table. Find the hazard rate during periods in the table.

Months from start	Number at the start of period	Failed during period	Number at the end of period	Hazard Rate
3	1200	4	1196	0.00111
5	1196	6	1190	0.00251
6	1190	3	1187	0.00252
7	1187	4	1183	0.00337
8	1183	4	1179	0.00338
9	1179	5	1174	0.00424
10	1174	6	1168	0.00511

Hazard Function (cont.)

Example:

2) Calculate the hazard function for

$$F(t) = 1 - e^{-\alpha t^{\beta}}$$

Answer: $h(t) = \alpha\beta t^{\beta-1}$

3) Find the hazard function for exponential distribution.

Answer: $F(t) = 1 - e^{-\lambda t}, t \in [0, \infty)$

$h(t) = \lambda$

Percentile of a Continuous Distribution

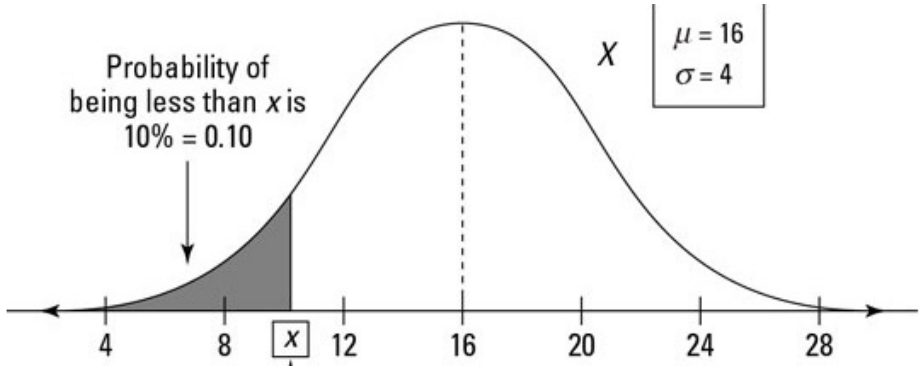
Consider a continuous distribution with cdf $F(x)$ and a number P between 0 and 100.

Definition: The P^{th} percentile of the distribution F is the number x such that $F(x) = \frac{P}{100} = P\%$.

This means that $P\%$ of the population for this distribution is less than the value x.

Example:

- The 5^{th} percentile of the standard normal distribution Φ is -1.645.
- The 35^{th} percentile of the standard normal distribution Φ is -0.385.
- The 80^{th} percentile of the standard normal distribution is 0.842.
- The 70^{th} percentile of the Student's t-distribution of degree 3 is 0.584.



In the left figure, x is the 10^{th} percentile of the distribution.

Sample Percentiles

Example: Given the data set: 5, 3, 9, 1, 0, 7, 8, 5, 9, 4.

Order this set: 0, 1, 3, 4, 5, 5, 7, 8, 9, 9.

- The value 4 is the 40th percentile of the set because 40% of the data values are less than or equal to 4. We also say that the percentile rank of 4 is 40.
- The value 5 is the 60th percentile of the set because 60% of the data values are less than or equal to 5. (The percentile rank of 5 is 60.)

- What is the 70th percentile of the set? Answer: The value 7.
- What is the 20th percentile of the set? Answer: The value 2.
- What is the 36th percentile of the set? → There's no such value in the set. Therefore, **multiple ways** were created to deal with this situation. We mention here two ways:

1) The percentile definition in Week 1 – Lecture 1 (please review).
2) For the second way, proceed to the next slide.

Sample Percentiles (2nd way)

Definition: Order the *n* sample observations from smallest to largest. Then the ith smallest observation in the list is taken to be the $[\frac{\mathbf{100(i-0.5)}}{\boldsymbol{n}}]$ th sample percentile. Note that based on this definition, a Pth percentile is the value that separates the bottom $P\%$ of the set and the top (100-P)% of the set.

Example: Given the data set: 5, 3, 9, 1, 0, 7, 8, 5, 9, 4. What percentiles are the values 3, 5 and 9? Use the above definition.

Solution:

Order this set: 0, 1, 3, 4, 5, 5, 7, 8, 9, 9. There are $n = 10$ values in the data set.

- 3 is the 3rd smallest value in the list. Its percentile rank is $\frac{100(3-0.5)}{10} = 25$. Thus, 3 is the 25th percentile (it separates the bottom 25% and the top 75% of the set).
- 5 is the 6th value in the list. Its percentile rank is $\frac{100(6-0.5)}{10} = 55$. Thus, 5 is the 55th percentile (it separates the bottom 55% and the top 45% of the set).
- 9 is the 10th value in the list. Its percentile rank is $\frac{100(10-0.5)}{10} = 95$. Thus, 9 is the 95th percentile (it separates the bottom 95% and the top 5% of the set).

Probability Plot

Definition: Consider a distribution F and a sample data set of size n. On a two-dimensional coordinate system, for all $i = 1, 2, \dots, n$, plot the following pairs

$$\left(\frac{100(i-0.5)}{n}\text{ th percentile of the distribution F, i}^{\text{th}}\text{ smallest value in the set}\right)$$

This plot is called a probability plot. If the standard normal distribution is used, it's called a normal probability plot.

- If data set (approximately) follows the distribution F, the plotted points will then fall close to a 45° line. Substantial deviations of the plotted points from a 45° line cast doubt on the assumption that the distribution under consideration is the correct one.
- Probability plots are used to see if a distribution is plausible for a given data set.

Probability Plot (cont.)

Example:

Consider the data set: $-1.91, -1.25, -0.75, -0.53, 0.20, 0.35, 0.72, 0.87, 1.40, 1.56$. Is it plausible that the data have a standard normal distribution?

Solution:

We will use a probability plot to answer this question.

1) First, we determine what percentiles are the values in the set. Note that data values are already ordered from smallest to largest. We use the formula $\frac{100(i-0.5)}{n}$ to determine percentiles. Below is the result.

Value	-1.91	-1.25	-0.75	-0.53	0.20	0.35	0.72	0.87	1.40	1.56
Percentile	5th	15th	25th	35th	45th	55th	65th	75th	85th	95th

2) Next, determine the value for the percentiles of the standard normal distribution.

Percentile	5th	15th	25th	35th	45th	55th	65th	75th	85th	95th
z-value	-1.645	-1.037	-0.675	-0.385	-0.126	0.126	0.385	0.675	1.037	1.645

3) Lastly, plot the pairs ($\frac{100(i-0.5)}{n}$ th percentile of the distribution F, ith smallest value in the set). (See in the next slide.)

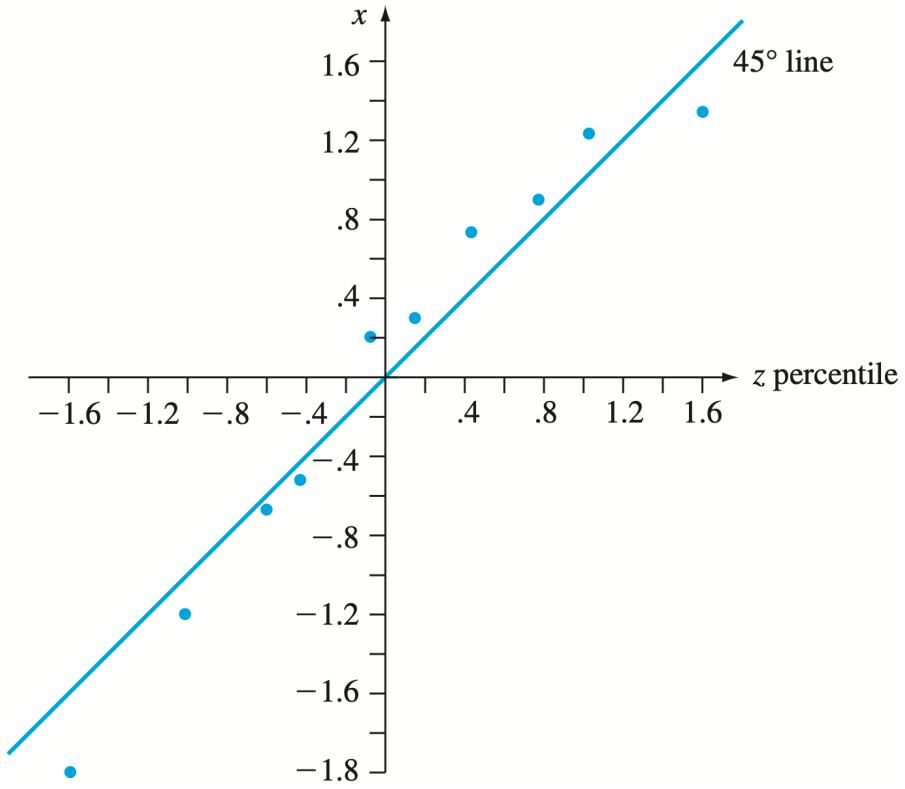
Probability Plot (cont.)

Example (cont.):

Consider the data set: $-1.91, -1.25, -0.75, -0.53, 0.20, 0.35, 0.72, 0.87, 1.40, 1.56$. Is it plausible that the data have a standard normal distribution?

Solution:

The probability plot shows that data fit the standard normal distribution well.



Probability Plot (cont.)

- It is frequently the case that just one probability plot will suffice for assessing the plausibility of an entire family. (This means that if your normal probability plot gives you a straight-line tendency, but it is not a 45° line, then your data follows a normal distribution, but not the standard normal distribution.) If the plot deviates substantially from a straight line, no member of the family is plausible. When the plot is quite straight, further work is necessary to estimate values of the parameters that yield the most reasonable distribution of the specified type.
- If the sample observations are in fact drawn from a normal distribution with mean value μ and standard deviation σ, the points should fall close to a straight line with slope μ and intercept σ. Thus, a normal probability plot for which the points fall close to some straight line suggests that the assumption of a normal population distribution is plausible.

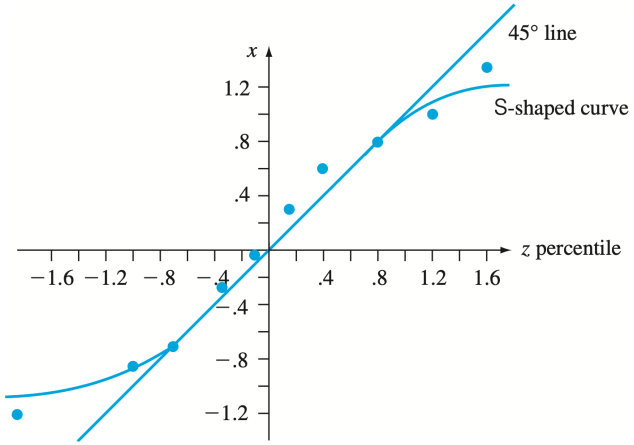
Probability Plot (cont.)

- A nonnormal population distribution can often be placed in one of the following three categories:

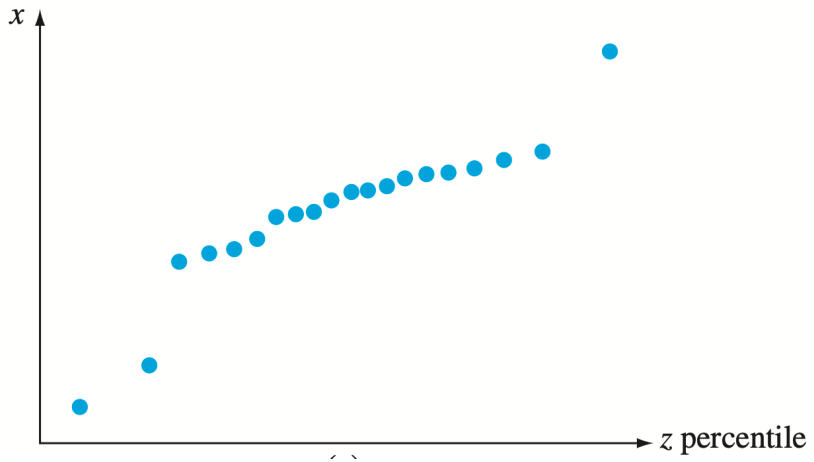
1. It is symmetric and has “lighter tails” than does a normal distribution; that is, the density curve declines more rapidly out in the tails than does a normal curve.

2. It is symmetric and heavy-tailed compared to a normal distribution.

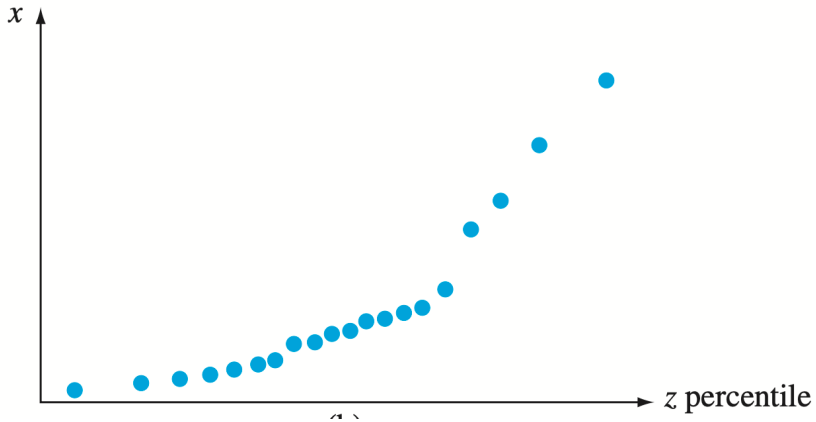
3. It is skewed.



Light-tailed distribution



Heavy-tailed distribution



Skewed distribution

Quantiles of a Continuous Distribution

Definition: Quantiles are the same as percentiles, but are given as decimal values, values such as 0.95, 0.4, and 0.27. The 0.95 quantile point is exactly the same as the 95^{th} percentile point.

Example:

- The 0.05 quantile of the standard normal distribution Φ is -1.645.
- The 0.35 quantile of the standard normal distribution Φ is -0.385.
- The 0.8 quantile of the standard normal distribution is 0.842.
- The 0.7 of the Student's t-distribution of degree 3 is 0.584.

Sample Quantiles

- Just like percentiles, there are different ways people use to compute quantiles. These ways produce slightly different quantiles, but that usually does not affect the result of data analysis.
- One commonly used formula to determine the quantile of the k^{th} smallest value in a sample data set is $\frac{k}{n+1}$, where n is the size of the sample.

Example: Determine the quantiles of the values in the set {1, 6, 3, 5, 10}.

Value	Calculation	Quantile
1	1/(5+1) = 0.167	0.167 quantile
3	2/(5+1) = 0.333	0.333 quantile
5	3/(5+1) = 0.5	0.5 quantile
6	4/(5+1) = 0.667	0.666 quantile
10	5/(5+1) = 0.833	0.833 quantile

Q-Q Plot

Definition: Consider a distribution F and a sample data set of size n. On a two-dimensional coordinate system, for all $i = 1, 2, \ldots, n$, plot the following pairs

$$\left(\frac{k}{n+1} \text{ quantile of the distribution, } k^{th} \text{ smallest sample observation}\right)$$

This plot is called a Q-Q plot. If the standard normal distribution is used, it's called a normal Q-Q plot.

- Just like probability plots, Q-Q plots are used to see if a distribution is plausible for a given data set. If data set (approximately) follows the distribution F, the plotted points will then fall close to a 45° line. Otherwise, it does not follow the distribution F.
- We can also construct a Q-Q plot for two data sets to determine if they come from populations with a common distribution. In this case, the Q-Q plot is plotted for the quantiles of the first data set against the quantiles of the second data set.

Q-Q Plot (cont.)

Example:

Do the following values come from a normal distribution?

7.19, 6.31, 5.89, 4.5, 3.77, 4.25, 5.19, 5.79, 6.79

Solution:

We will use a Q-Q plot to answer this question.

1) First, we determine what quantiles are the values in the set. We use the formula $\frac{k}{n+1}$ to determine quantiles.

Value	3.77	4.25	4.5	5.19	5.79	5.89	6.31	6.79	7.19
Quantile	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9

2) Next, determine the value for the quantiles of the standard normal distribution.

Quantile	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9
z-value	-1.28	-0.84	-0.52	-0.25	0	0.25	0.52	0.84	1.28

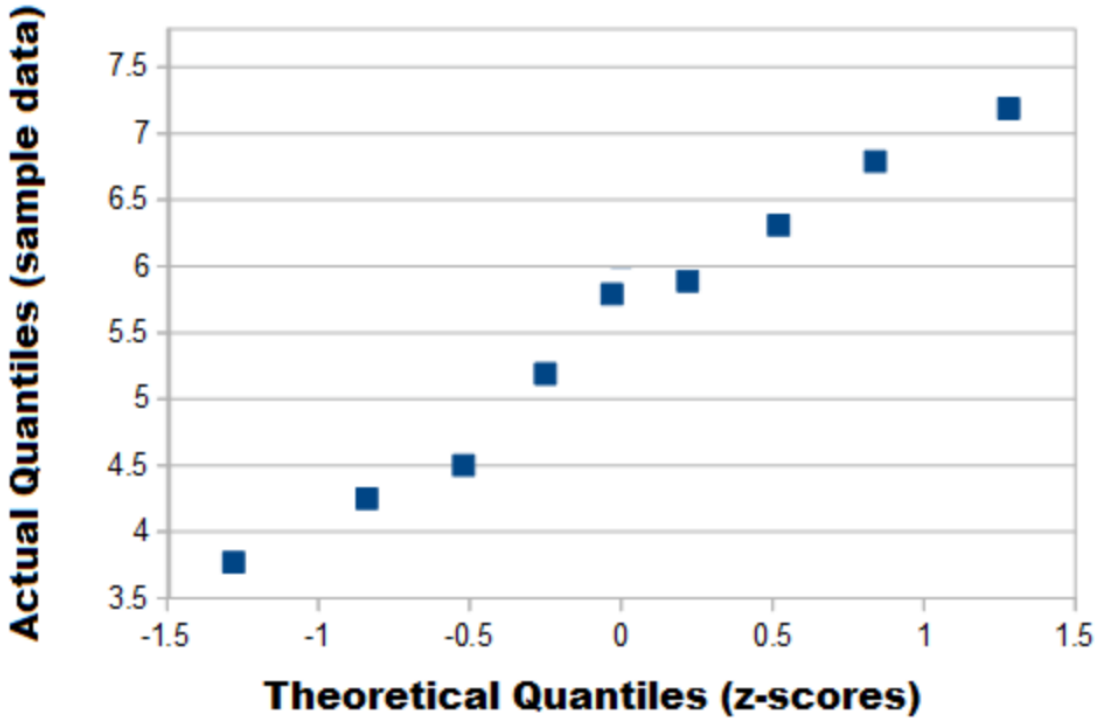
3) Lastly, plot the pairs (-1.28, 3.77), (-0.84, 4.25), (-0.52, 4.5), (-0.25, 5.19), (0, 5.79), (0.25, 5.89), (0.52, 6.31), (0.84, 6.79), (1.28, 7.19) on the coordinate system. (See in the next slide.)

Q-Q plot (cont.)

Example:

Do the following values come from a normal distribution?

7.19, 6.31, 5.89, 4.5, 3.77, 4.25, 5.19, 5.79, 6.79



Q-Q Plot (cont.)

Example:

Cleveland et al. (1974) used Q-Q plots in a study of air pollution. They plotted the quantiles of distributions of the values of various variables on Sunday against the quantiles for weekdays (see figure below).

- The Q-Q plot of the ozone maxima shows that the very highest quantiles occur on weekdays but that all the other quantiles are larger on Sundays.
- For carbon monoxide, nitrogen oxide, and aerosols, the differences in the quantiles increase with increasing concentration.
- The very high and very low quantiles of solar radiation are about the same on Sundays and weekdays (presumably corresponding to very clear days and days with heavy cloud cover), but for intermediate quantiles, the Sunday quantiles are larger.