CE 486
Urban Transportation Planning

Lec. 7

Introduction to Statistics and Probability Theory

Dr. Mahmoud Owais

Variables

- A **variable** is a characteristic or condition that can change or take on different values.
- Most research begins with a general question about the relationship between two variables for a specific group of individuals.

Types of Variables

- Variables can be classified as **discrete** or **continuous**.
- **Discrete variables** (such as class size) consist of indivisible categories, and **continuous variables** (such as time or weight) are infinitely divisible into whatever units a researcher may choose. For example, time can be measured to the nearest minute, second, half-second, etc.

Population

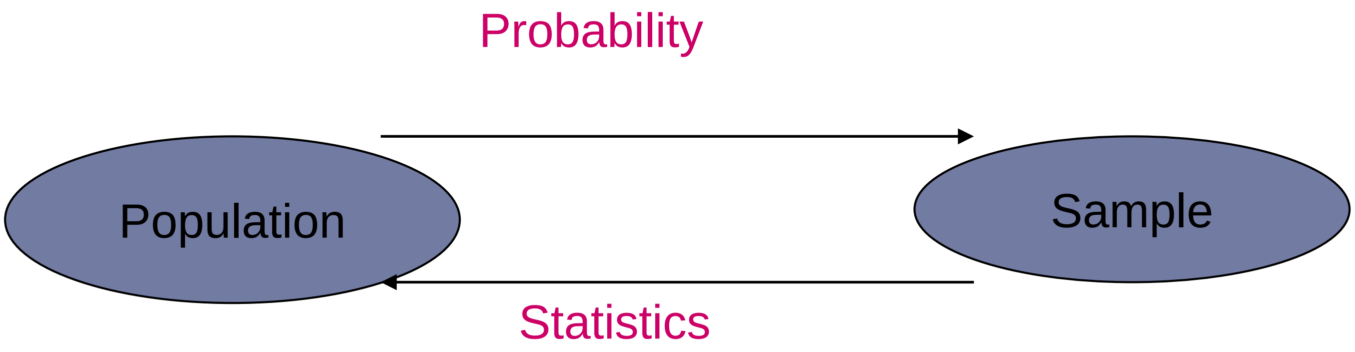
- The entire group of individuals is called the **population**.
- For example, a researcher may be interested in the relation between class size (variable 1) and academic performance (variable 2) for the population of third-grade children.

Sample

- Usually populations are so large that a researcher cannot examine the entire group. Therefore, a **sample** is selected to represent the population in a research study. The goal is to use the results obtained from the sample to help answer questions about the population.

Why to Learn?

- Nothing in life is certain. In everything we do, we gauge the chances of successful outcomes, from business to medicine to the weather
- A probability provides a quantitative description of the chances or likelihoods associated with various outcomes
- To build a bridge between descriptive and inferential statistics



Descriptive Statistics

- **Descriptive statistics** are methods for organizing and summarizing data.
- For example, tables or graphs are used to organize data, and descriptive values such as the average score are used to summarize data.
- A descriptive value for a population is called a **parameter** and a descriptive value for a sample is called a **statistic**.

Inferential Statistics

- **Inferential statistics** are methods for using sample data to make general conclusions (inferences) about populations.
- Because a sample is typically only a part of the whole population, sample data provide only limited information about the population. As a result, sample statistics are generally imperfect representatives of the corresponding population parameters.

What is Probability?

- We measured "how often" using

Relative frequency = *f*/*n*

The possible outcomes of a process are called **events**. (A deterministic process has only one possible outcome.)

The probability of a particular event is the fraction of outcomes in which the event occurs. The probability of event A is **denoted** by P(A).

The sum of probabilities of all mutually exclusive events in a process is 1. For example, if there are n possible mutually exclusive outcomes, then

$$\sum_{i=1}^{n} P(i) = 1$$

Simple probabilities

If A and B are mutually exclusive events, then the probability of either A or B to occur is the union

$$P(A \cup B) = P(A) + P(B)$$

Example: The probability of a hat being red is ¼, the probability of the hat being green is ¼, and the probability of the hat being black is ½. Then, the probability of a hat being red OR black is ¾.

Simple probabilities

If A and B are independent events, then the probability that both A and B occur is the intersection

$$P(A \cap B) = P(A) \times P(B)$$

Conditional probabilities

What is the probability of event *A* to occur given than event *B* did occur. The conditional probability of *A* given *B* is

$$P(A \mid B) = \frac{P(A \cap B)}{P(A)}$$

Problem 1.

Blood Group	Males	Females	Total
O	20	50	70
A	17	0	17
B	8	0	8
AB	5	0	5
Total	50	50	100

Marginal probabilities

Named so because they appear on the “margins” of a probability table. It is probability of single outcome

Example: In problem 1, P(Male), P(Blood group A)

P(Male) = number of males/total

number of subjects

= 50/100

= 0.5

Conditional probabilities

It is the probability of an event on condition that certain criteria is satisfied

Example: If a subject was selected randomly and found to be female what is the probability that she has a blood group O

Here the total possible outcomes constitute a subset (females) of the total number of subjects.

This probability is termed probability of O given F

P(O\F) = 50/50

= 100%

Joint probability

It is the probability of occurrence of two or more events together

Example: Probability of being male & belong to blood group AB

P(M and AB) = P(M∩AB)

= 5/100

= 0.05

∩ = intersection

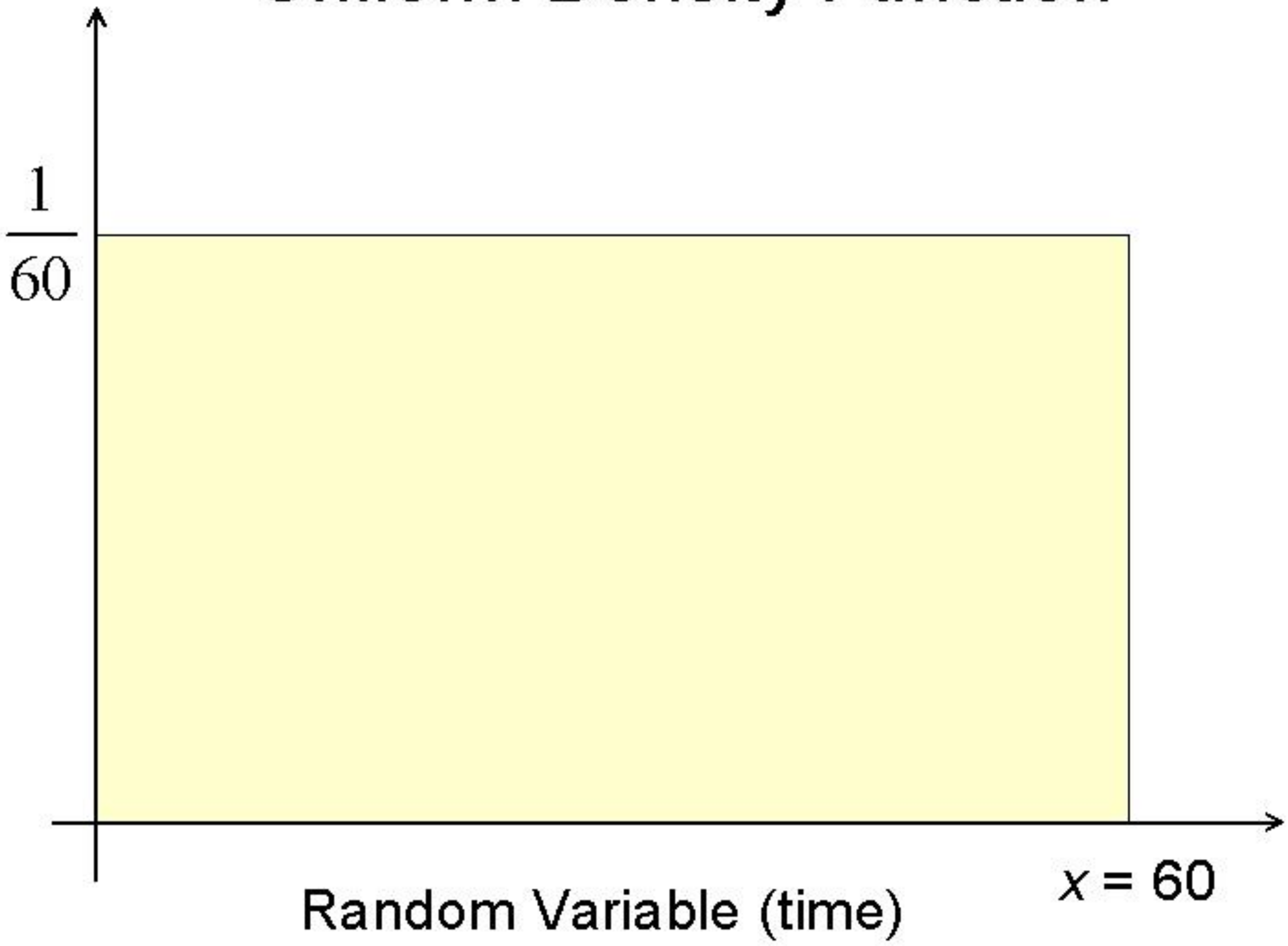
Probability Density Function

A probability density function is an equation that is used to compute probabilities of continuous random variables that must satisfy the following two properties.

1. The area under the graph of the equation over all possible values of the random variable must equal one.

2. The graph of the equation must be greater than or equal to zero for all possible values of the random variable. That is, the graph of the equation must lie on or above the horizontal axis for all possible values of the random variable, i.e. f(x)>=0

Uniform Density Function



Types of Distribution

- Frequency Distribution
- **Normal (Gaussian) Distribution**
- Probability Distribution
- Poisson Distribution
- Binomial Distribution
- Sampling Distribution
- t distribution
- F distribution

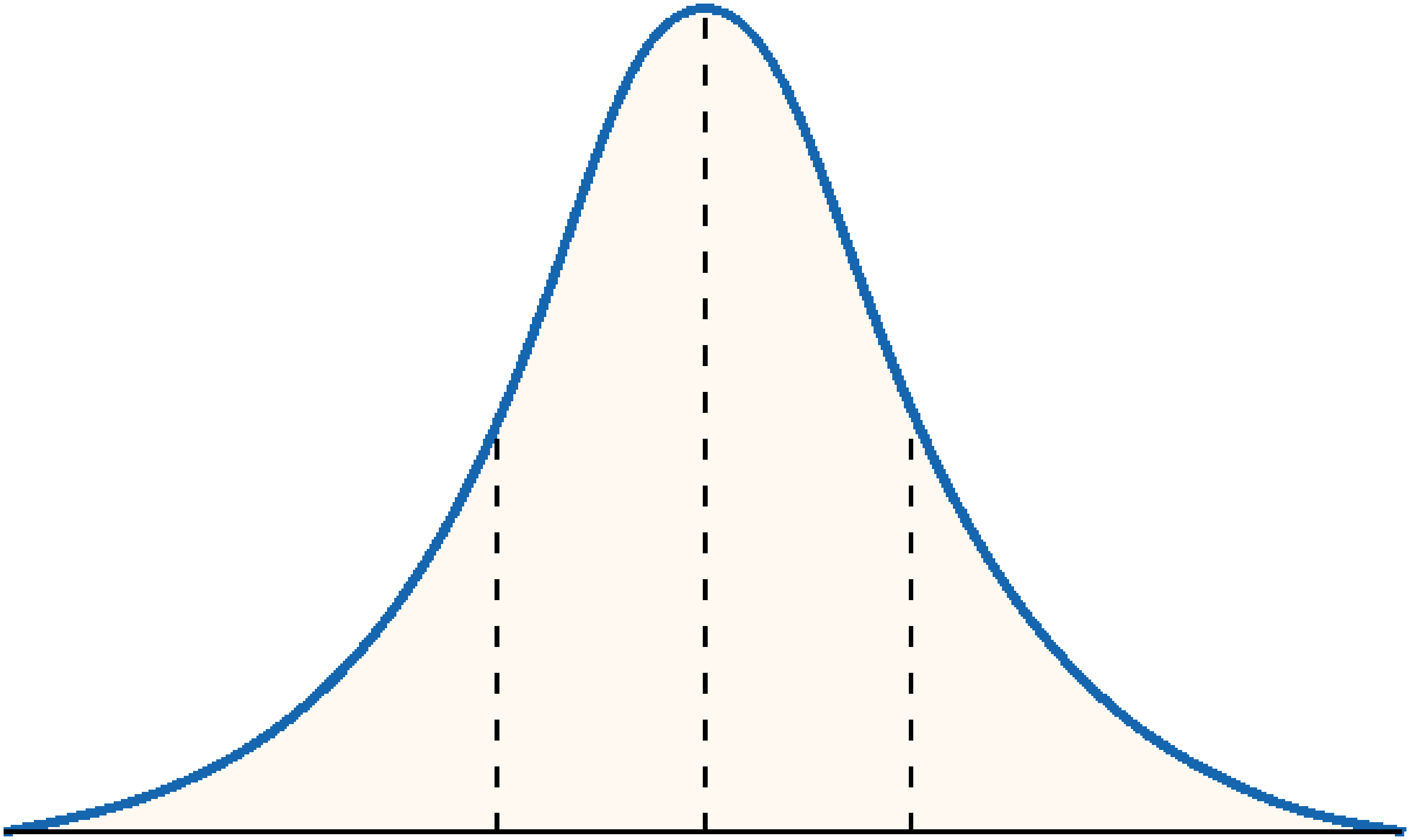
Characteristics of Normal Distribution

- It links frequency distribution to probability distribution

- Has a Bell Shape Curve and is Symmetric

- It is Symmetric around the mean:

  Two halves of the curve are the same (mirror images)



Characteristics of Normal Distribution Cont'd

- Hence Mean = Median
- The total area under the curve is 1 (or 100%)
- Normal Distribution has the same shape as Standard Normal Distribution.

Characteristics of Normal Distribution Cont’d

- In a Standard Normal Distribution:

The mean (µ) = 0 and

Standard deviation (σ) =1

Formula for Standard Deviation

$$\sigma = \sqrt{\frac{\sum (X - \mu)^2}{(n-1)}}$$

$\sqrt{}$ =square root

$\sum$=sum (sigma)

X=score for each point in data

μ =mean of scores for the variable

n=sample size (number of observations or cases

Z Score (Standard Score)[3]

- $Z = \dfrac{X - \mu}{\sigma}$
- Z indicates how many standard deviations away from the mean the point x lies.
- Z score is calculated to 2 decimal places.

The Normal Distribution:
as mathematical function (pdf)

$$f(x) = \frac{1}{\boxed{\sigma}\sqrt{2\pi}} \cdot e^{-\frac{1}{2}(\frac{x-\boxed{\mu}}{\sigma})^2}$$

This is a bell shaped curve with different centers and spreads depending on μ and σ

<u>Note constants:</u>

π=3.14159

e=2.71828

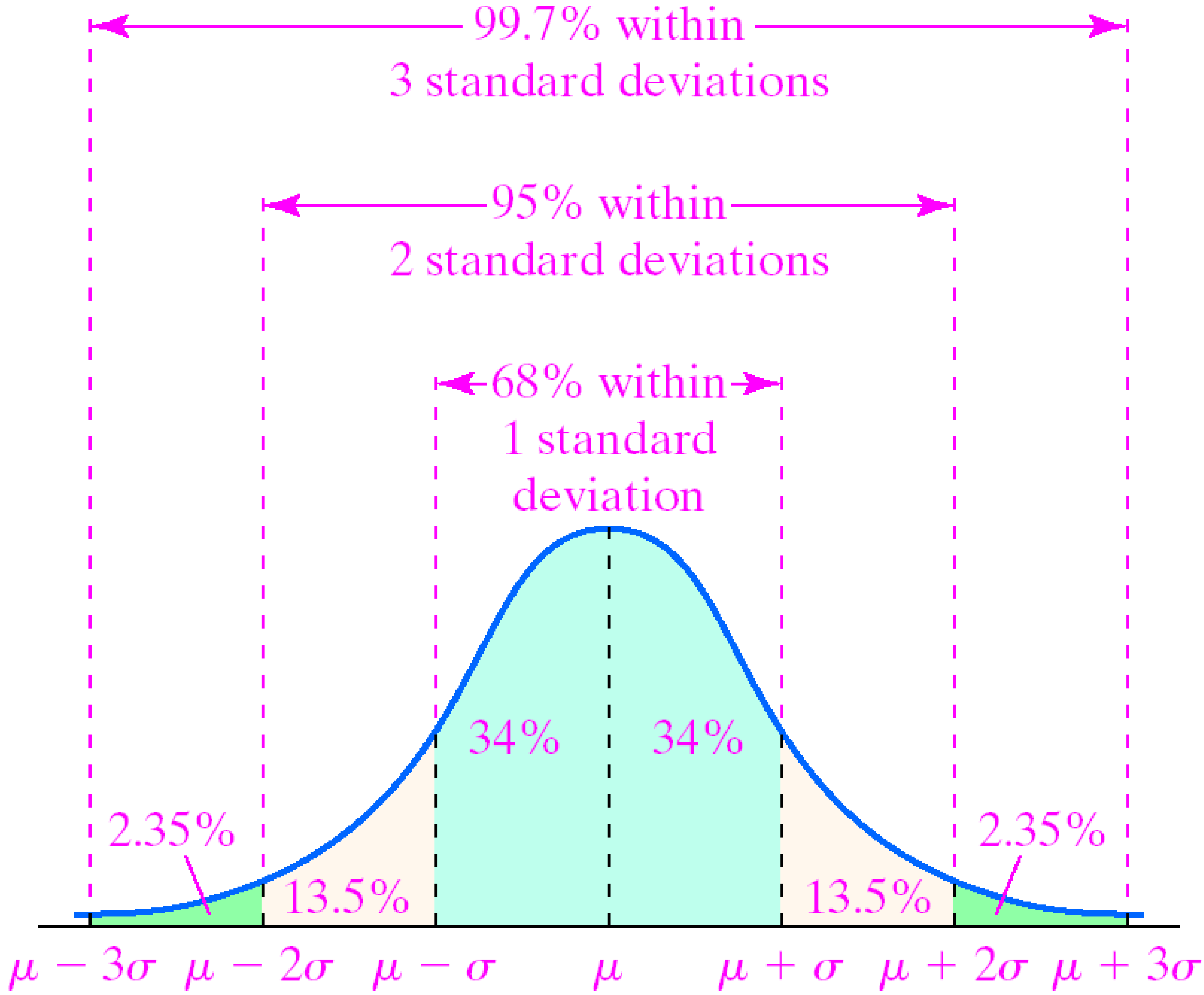
Distinguishing Features

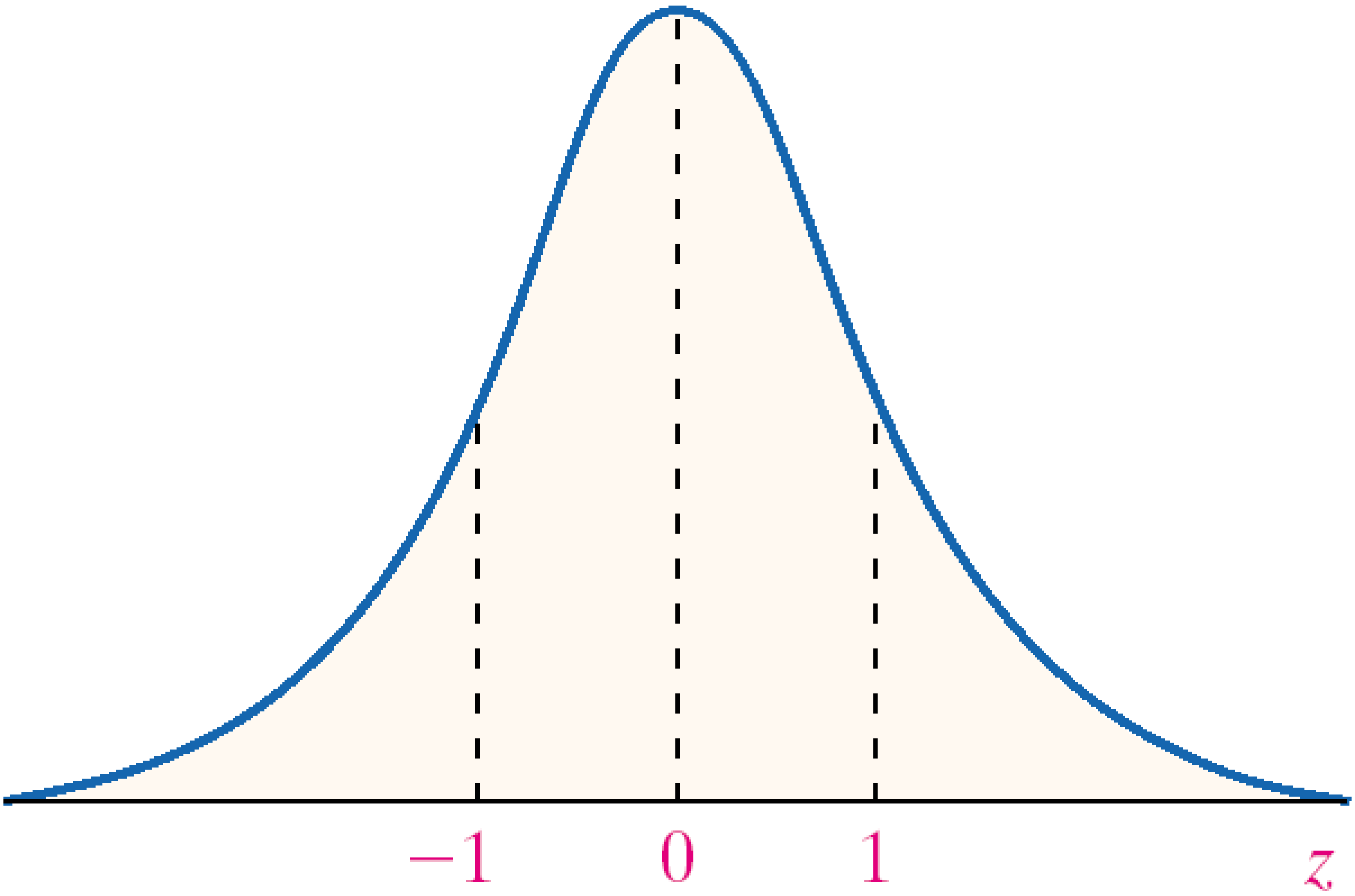
- The mean ± 1 standard deviation covers 66.7% of the area under the curve

- The mean ± 2 standard deviation covers 95% of the area under the curve

- The mean ± 3 standard deviation covers 99.7% of the area under the curve

Normal Distribution





−1
0
1
z

Properties of the Normal Density Curve

1. It is symmetric about its mean, $\mu = 0$.

2. The highest point occurs at $z = 0$

3. It has inflection points at -1 and 1.

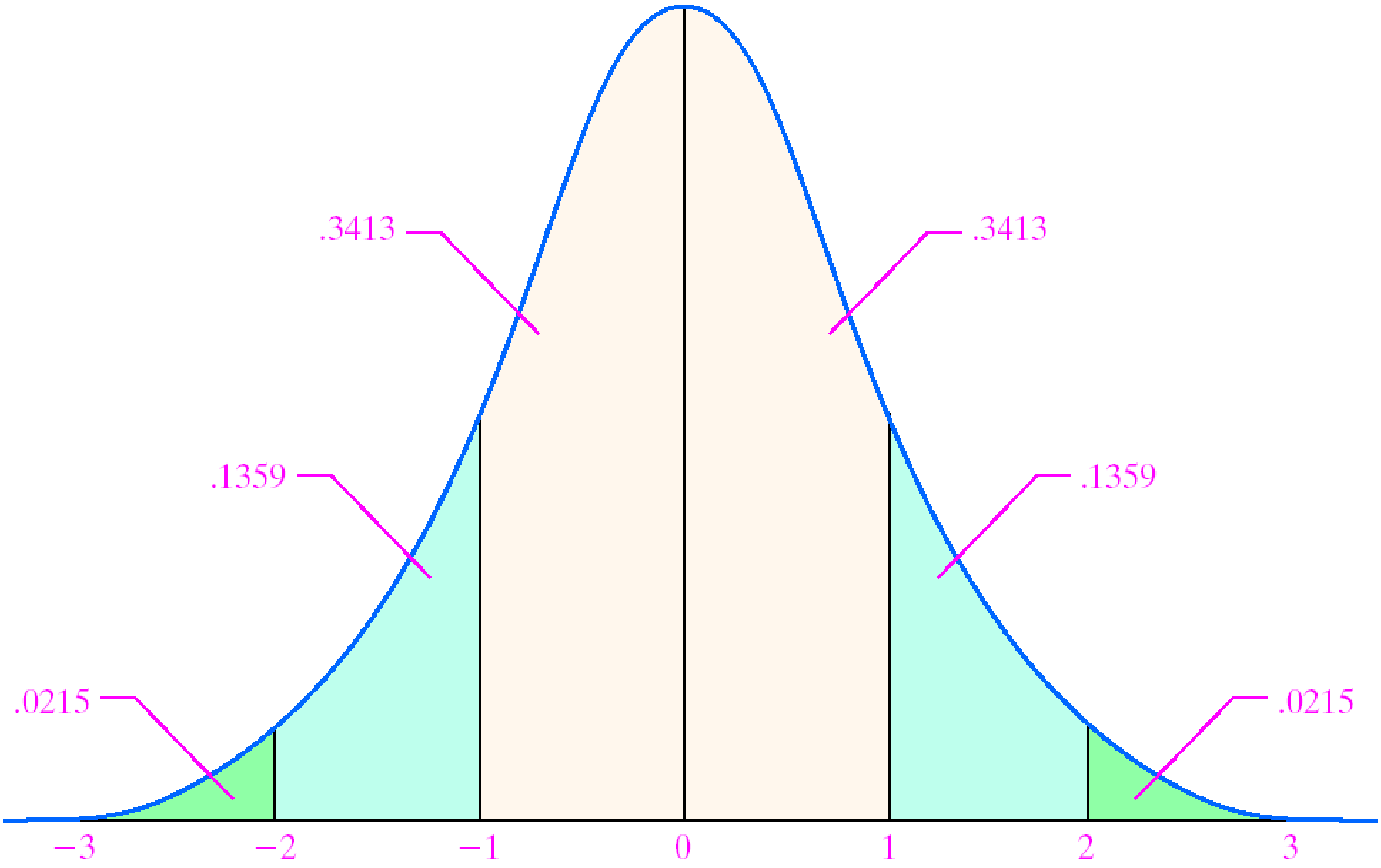
4. The area under the curve is one.

5. The area under the curve to the right of $\mu = 0$ equals the area under the curve to the left of $\mu = 0$ equals ½.

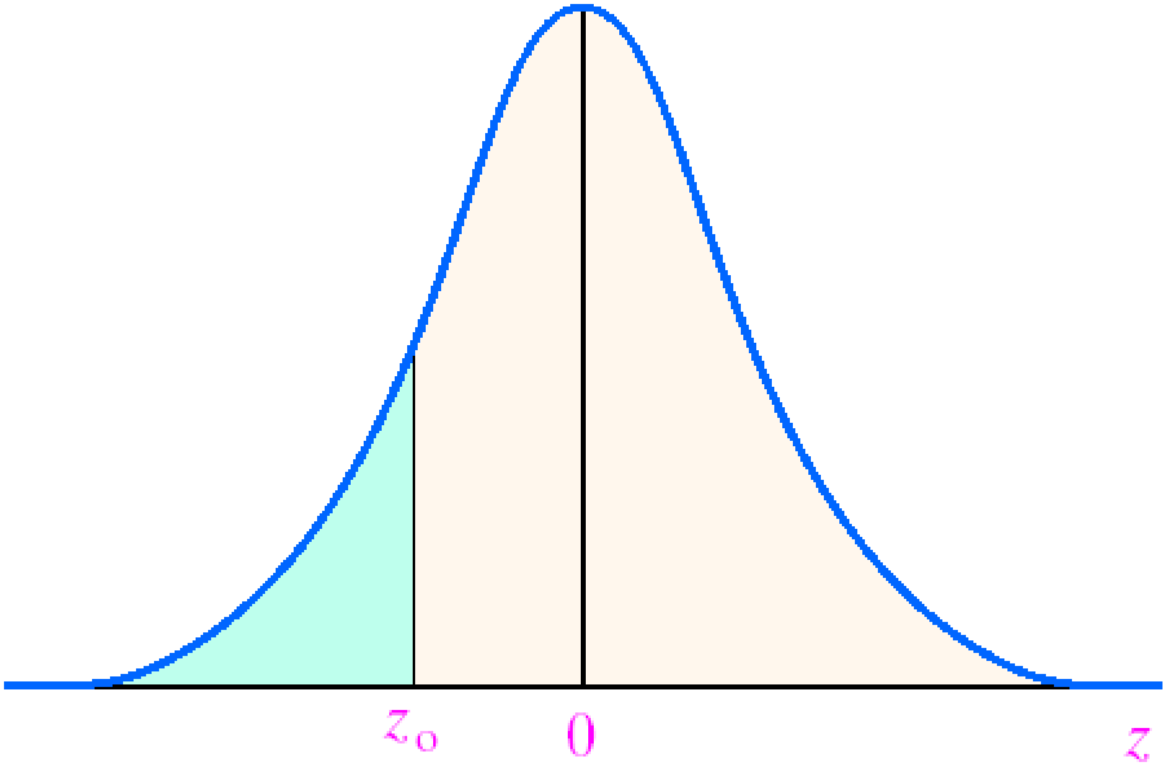
6. As z increases without bound, the graph approaches, but never equals, zero. As z decreases without bound the graph approaches, but never equals, zero.

Properties of the Normal Density Curve

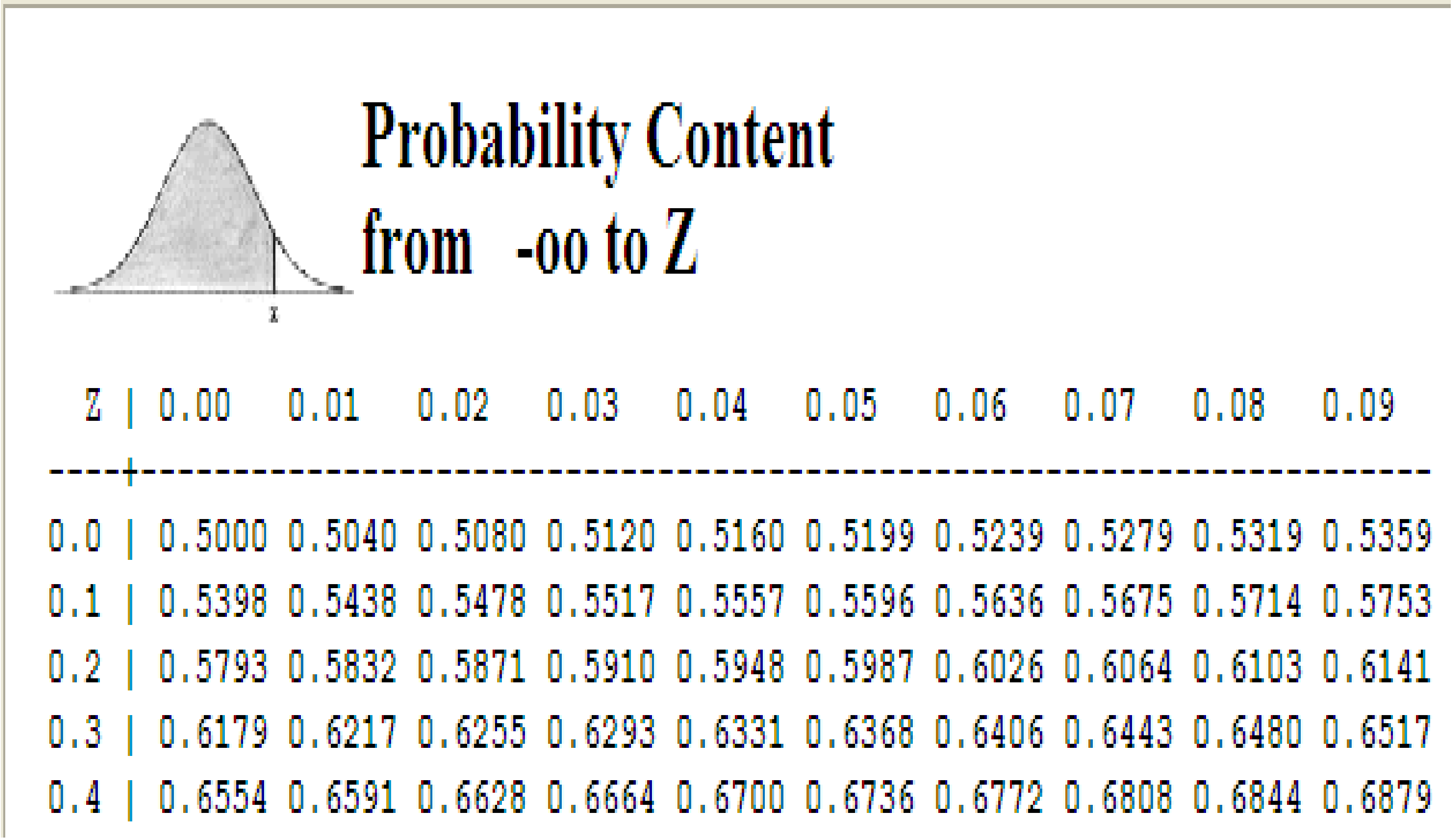
7. The Empirical Rule: About 68% of the area under the graph is between -1 and 1; about 95% of the area under the graph is between -2 and 2; about 99.7% of the area under the graph is between -3 and 3.


.3413
.3413
.1359
.1359
.0215
.0215
−3
−2
−1
0
1
2
3

The table gives the area under the standard normal curve for values to the left of a specified Z-score, z_o, as shown in the figure.



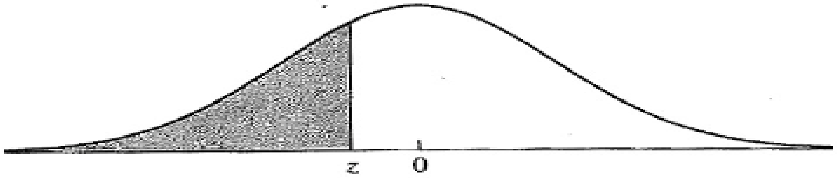
Tables of the Normal Distribution



Probability Content from -oo to Z

Z	0.00	0.01	0.02	0.03	0.04	0.05	0.06	0.07	0.08	0.09
0.0	0.5000	0.5040	0.5080	0.5120	0.5160	0.5199	0.5239	0.5279	0.5319	0.5359
0.1	0.5398	0.5438	0.5478	0.5517	0.5557	0.5596	0.5636	0.5675	0.5714	0.5753
0.2	0.5793	0.5832	0.5871	0.5910	0.5948	0.5987	0.6026	0.6064	0.6103	0.6141
0.3	0.6179	0.6217	0.6255	0.6293	0.6331	0.6368	0.6406	0.6443	0.6480	0.6517
0.4	0.6554	0.6591	0.6628	0.6664	0.6700	0.6736	0.6772	0.6808	0.6844	0.6879

Cumulative normal distribution (*z* table)



z	0.00	0.01	0.02	0.03	0.04	0.05	0.06	0.07	0.08	0.09
−3.6	.0002	.0002	.0001	.0001	.0001	.0001	.0001	.0001	.0001	.0001
−3.5	.0002	.0002	.0002	.0002	.0002	.0002	.0002	.0002	.0002	.0002
−3.4	.0003	.0003	.0003	.0003	.0003	.0003	.0003	.0003	.0003	.0002
−3.3	.0005	.0005	.0005	.0004	.0004	.0004	.0004	.0004	.0004	.0003
−3.2	.0007	.0007	.0006	.0006	.0006	.0006	.0006	.0005	.0005	.0005
−3.1	.0010	.0009	.0009	.0009	.0008	.0008	.0008	.0008	.0007	.0007
−3.0	.0013	.0013	.0013	.0012	.0012	.0011	.0011	.0011	.0010	.0010
−2.9	.0019	.0018	.0018	.0017	.0016	.0016	.0015	.0015	.0014	.0014
−2.8	.0026	.0025	.0024	.0023	.0023	.0022	.0021	.0021	.0020	.0019
−2.7	.0035	.0034	.0033	.0032	.0031	.0030	.0029	.0028	.0027	.0026
−2.6	.0047	.0045	.0044	.0043	.0041	.0040	.0039	.0038	.0037	.0036
−2.5	.0062	.0060	.0059	.0057	.0055	.0054	.0052	.0051	.0049	.0048
−2.4	.0082	.0080	.0078	.0075	.0073	.0071	.0069	.0068	.0066	.0064
−2.3	.0107	.0104	.0102	.0099	.0096	.0094	.0091	.0089	.0087	.0084
−2.2	.0139	.0136	.0132	.0129	.0125	.0122	.0119	.0116	.0113	.0110
−2.1	.0179	.0174	.0170	.0166	.0162	.0158	.0154	.0150	.0146	.0143
−2.0	.0228	.0222	.0217	.0212	.0207	.0202	.0197	.0192	.0188	.0183
−1.9	.0287	.0281	.0274	.0268	.0262	.0256	.0250	.0244	.0239	.0233
−1.8	.0359	.0351	.0344	.0336	.0329	.0322	.0314	.0307	.0301	.0294
−1.7	.0446	.0436	.0427	.0418	.0409	.0401	.0392	.0384	.0375	.0367
−1.6	.0548	.0537	.0526	.0516	.0505	.0495	.0485	.0475	.0465	.0455
−1.5	.0668	.0655	.0643	.0630	.0618	.0606	.0594	.0582	.0571	.0559
−1.4	.0808	.0793	.0778	.0764	.0749	.0735	.0721	.0708	.0694	.0681
−1.3	.0968	.0951	.0934	.0918	.0901	.0885	.0869	.0853	.0838	.0823
−1.2	.1151	.1131	.1112	.1093	.1075	.1056	.1038	.1020	.1003	.0985
−1.1	.1357	.1335	.1314	.1292	.1271	.1251	.1230	.1210	.1190	.1170
−1.0	.1587	.1562	.1539	.1515	.1492	.1469	.1446	.1423	.1401	.1379
−0.9	.1841	.1814	.1788	.1762	.1736	.1711	.1685	.1660	.1635	.1611
−0.8	.2119	.2090	.2061	.2033	.2005	.1977	.1949	.1922	.1894	.1867
−0.7	.2420	.2389	.2358	.2327	.2296	.2266	.2236	.2206	.2177	.2148
−0.6	.2743	.2709	.2676	.2643	.2611	.2578	.2546	.2514	.2483	.2451
−0.5	.3085	.3050	.3015	.2981	.2946	.2912	.2877	.2843	.2810	.2776
−0.4	.3446	.3409	.3372	.3336	.3300	.3264	.3228	.3192	.3156	.3121
−0.3	.3821	.3783	.3745	.3707	.3669	.3632	.3594	.3557	.3520	.3483
−0.2	.4207	.4168	.4129	.4090	.4052	.4013	.3974	.3936	.3897	.3859

Notation for the Probability of a Standard Normal Random Variable

$P(a < Z < b)$ represents the probability a standard normal random variable is between *a* and *b*

$P(Z > a)$ represents the probability a standard normal random variable is greater than *a*.

$P(Z < a)$ represents the probability a standard normal random variable is less than *a*.

Area under the normal curve to the right of z_o

$$= 1 - \text{Area to the left of } z_o$$

EXAMPLE *Finding the Area Under the Standard Normal Curve*

Find the area under the standard normal curve to the right of *Z* = 1.25.

Look in the Normal distribution table:

P(x>=1.25) =1-P(x<1.25)=1-0.8944=0.1056

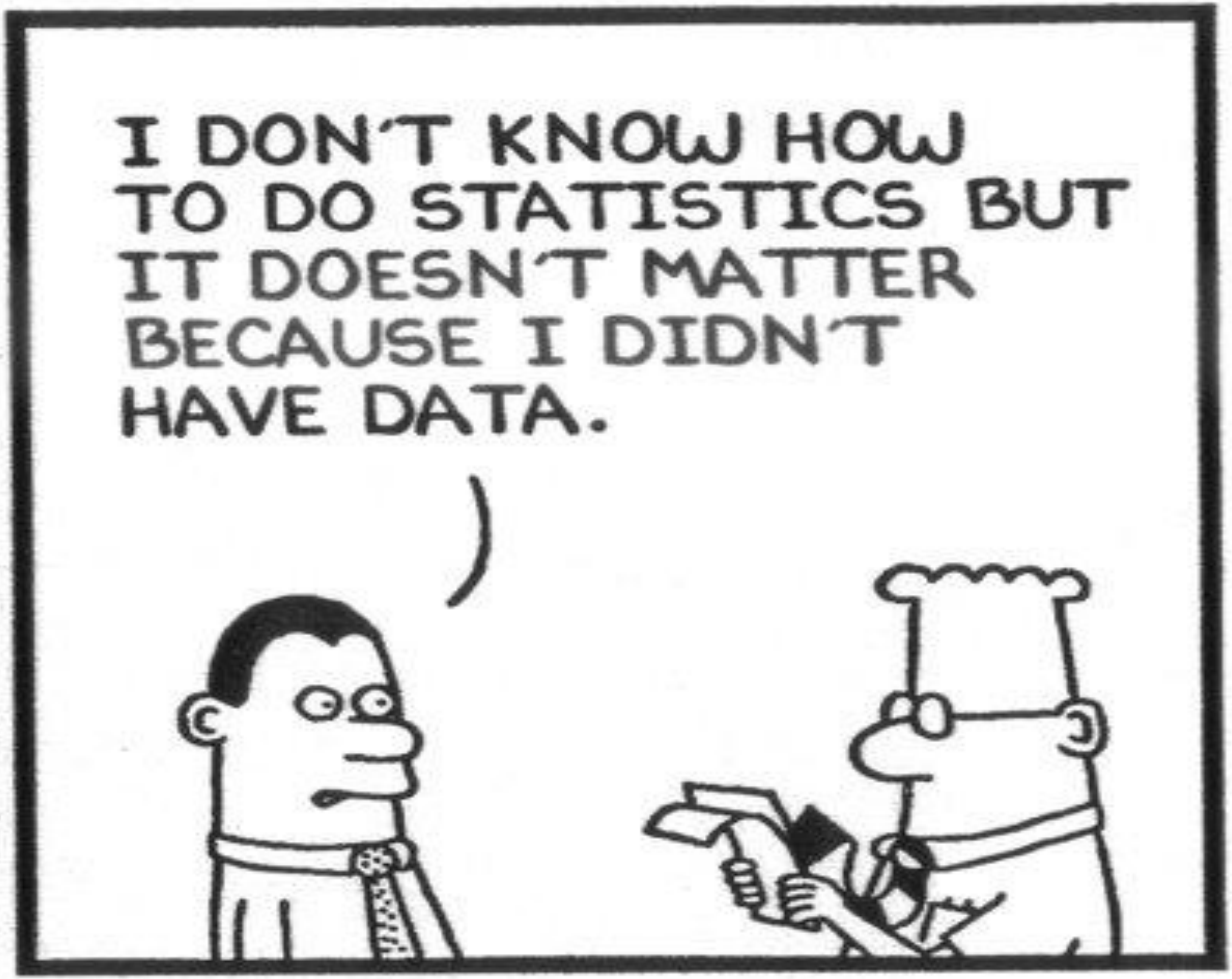
EXAMPLE *Finding the Area Under the Standard Normal Curve*

Find the area under the standard normal curve between *Z* = -1.02 and *Z* = 2.94.

P(-1.02<x<2.94)=P(x<2.94)-p(x<-1.02)

=0.9984-0.1539

=0.8445


I DON'T KNOW HOW TO DO STATISTICS BUT IT DOESN'T MATTER BECAUSE I DIDN'T HAVE DATA.

Frequency table

- **absolute frequency “n_i”** (Data Tab→Data Analysis→Histogram)
- **relative frequency “f_i”**

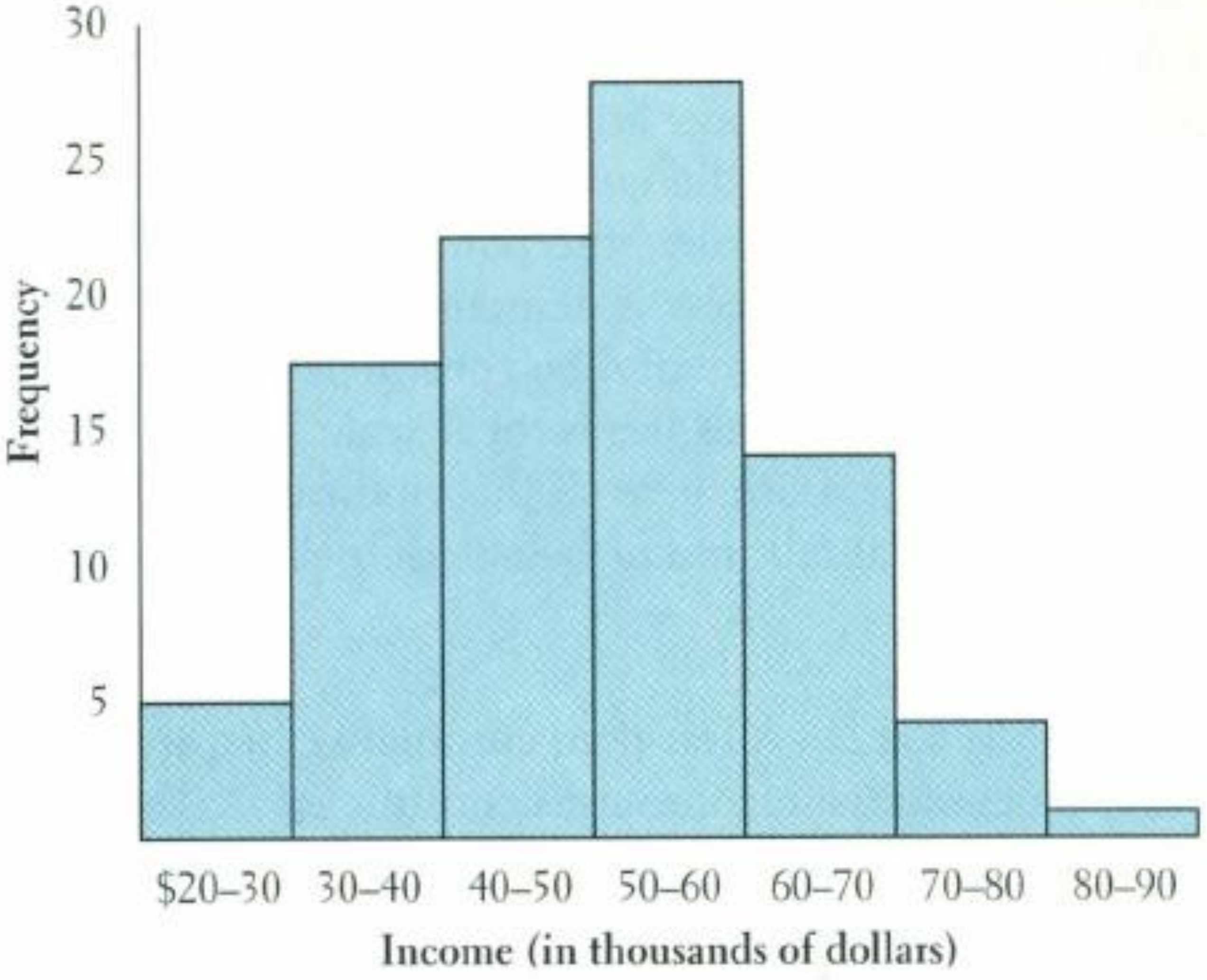
Cumulative frequency distribution shows the total number of occurrences that lie above or below certain key values.

- **cumulative frequency “N_i”**
- **cumulative relative frequency “F_i”**

Histogram

- Frequently used to graphically present interval and ratio data
- Is often used for interval and ratio data
- The adjacent bars indicate that a numerical range is being summarized by indicating the frequencies in arbitrarily chosen classes

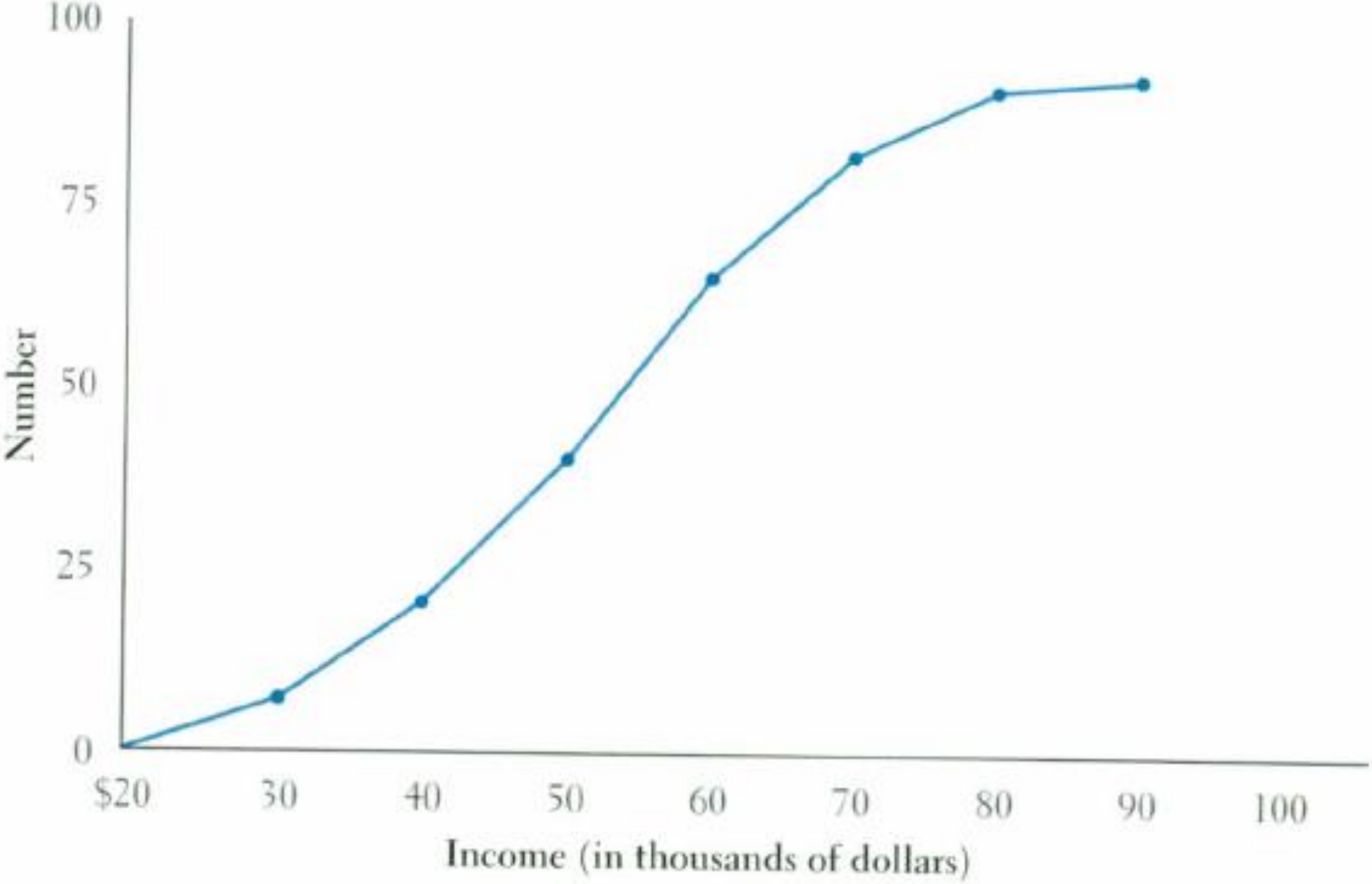
Histogram—Executive Incomes for the Sunrunner Corporation



Ogive

- A graph of a cumulative frequency distribution
- Ogive is used when one wants to determine how many observations lie above or below a certain value in a distribution.
- First cumulative frequency distribution is constructed
- Cumulative frequencies are plotted at the upper class limit of each category
- Ogive can also be constructed for a relative frequency distribution.

Ogive—Executive Incomes (frequencies)



Frequency Distribution

Consider a data set of 26 children of ages 1-6 years. Then the frequency distribution of variable ‘age’ can be tabulated as follows:

Frequency Distribution of Age

Age	1	2	3	4	5	6
Frequency	5	3	7	5	4	2

Grouped Frequency Distribution of Age:

Age Group	1-2	3-4	5-6
Frequency	8	12	6

Cumulative Frequency

Cumulative frequency of data in previous page

Age	1	2	3	4	5	6
Frequency	5	3	7	5	4	2
Cumulative Frequency	5	8	15	20	24	26

Age limit	2	4	6
Cumulative Frequency	8	20	26