MA5701: Statistical Methods

Chapter 2 : Probability and Sampling Distributions

Kui Zhang, Mathematical Sciences

Examples for Probability

- **Example 2.1 (Estimate the cost involved in replacing parts)** – defective screws can be produced at two points in a production line, which must be removed and replaced. We would like to estimate the total cost involved with about 1000 parts manufactured.

Table 2.1 Summary of Defective Screws

Point in the Production Line	Proportion of Parts Having Defective Screws	Cost of Replacement
1	0.008	\$0.23
2	0.004	\$0.69

Examples for Probability

- **Example:** Estimate the number of fishes in a pond (Capturing and Re-capturing method) – interest in total number of fish (N), captured M fishes, mark them and put them back, capture K fishes and find L marked. How to use K, L, and M to estimate N?
- **Example:** Lottery – what is the probability to win lottery if you buy one, or two, or ten tickets?

Population and Sample

- **Definition 1.2 -** A **population** is a data set representing the entire entity of interest.
- **Definition 1.3 -** A **sample** is a data set consisting of a portion of a population. (Obtained in a way to represent the population)

Variables – Texas House Data

Obs	Zip	Age	Bed	Bath	Size	Lot	Exter	garage	fp	Price
1	3	21	3	2	951	64904	Other	0	0	30000
3	4	7	1	1	676	54450	Other	2	0	46500
5	1	51	3	1	1186	10857	Other	1	0	51500
7	3	8	3	2	1368	.	Frame	0	0	56990
9	1	51	2	1	1176	6259	Frame	1	1	65500

Data: 69 families in a midsized city in east Texas. This is only part of it.

Population and Sample

- **Population** – its characteristic is generally described by the parameter.
  - Usually denoted by Greek letters, such as α, β, μ, etc.
  - It is generally unknown.
  - It is primary of interest of statistical inference and is estimated from sample.
- **Sample** – its characteristic is generally described by the statistics.
  - Usually denoted by alphabetic letters, such as x, y, z, p, etc.
  - It can be calculated from the data and is considered as known once the data is collected.
  - It is used to estimate the population parameter.

Parameter and Statistics

- **Definition 2.1** – A **parameter** is a quantity describes a particular characteristic of the distribution of a variable from a population.
- **Definition 2.2** – A **statistic** is a quantity calculated from data that described a particular characteristic of the sample.
- A statistic and a parameter are very similar. They are both descriptions of a characteristic. For example, “In average, more than 5% of MTU graduate students in School of Forestry take MA5701 – Statistical Methods, in last three years”.

Parameter and Statistics

- The difference between them is that a statistic describes a **sample** while a parameter describes an entire **population**.
  - **Parameter** – a numerical value describing some characteristic of a *population.*
  - **Statistic** – a numerical value describing some characteristic of a *sample*.
- For the example on the last slide, is 5% a statistic or parameter?
- The answer of this depends on the context – need to know the population first.

Parameter or Statistic?

Determine whether the underlined value is a parameter or a statistic.

- In a survey conducted in the town of Atherton, **25%** of adult respondents reported that they had been involved in at least one car accident in the past ten years. Here 25% is a _______
  - [✓ Statistic]
  - Parameter

Parameter or Statistic?

Determine whether the underlined value is a parameter or a statistic.

- **27.2%** of the mayors of cities in a certain state are from minority groups. Here 27.2% is a _______
  - Statistic
  - [✓ Parameter]

Variables – Texas House Data

Obs	Zip	Age	Bed	Bath	Size	Lot	Exter	garage	fp	Price
1	3	21	3	2	951	64904	Other	0	0	30000
3	4	7	1	1	676	54450	Other	2	0	46500
5	1	51	3	1	1186	10857	Other	1	0	51500
7	3	8	3	2	1368	.	Frame	0	0	56990
9	1	51	2	1	1176	6259	Frame	1	1	65500

Data: 69 families in a midsized city in east Texas. This is only part of it.

Parameter or Statistic?

Determine whether the underlined value is a parameter or a statistic. From the Texas House data,

- Among 69 houses selected, about **33.3%** of houses have a price from $100000 to $150000. Here the percentage 33.3% is a _______
  - [✓ Statistic]
  - Parameter

Statistical Inference

- **Definition 2.3** - **Statistical inference** is the process of using sample statistics to make decisions about population parameters (probability distribution).
- **Examples of Statistical Inference**
  - The use of sample mean to estimate the population mean. For example, what is the average GRE score of MTU graduate students?
  - Decide if population mean is greater than a certain value (Hypothesis Testing using sample mean and sample variance). For example, is the average GRE score of MTU students greater than the national average?

Sample Space and Event

- **Definition 2.4** – An **experiment** is any process that yields an observation.
- **Definition 2.5** – An **outcome** is a specific result of an experiment.
- Definition – The **sample space** is the combination of all possible outcomes of an experiment, denoted by S.
- **Definition 2.6** – An **event** is a combination of outcomes having some special characteristic of interest. In other words, the event is a subset of the sample space.

Sample Space and Event

- **Example:** Toss a coin.
  - The sample space is $S = \{H, T\}$.
- **Example:** The experiment consists two steps. First a coin is flipped. If tit is a tail, then a die is tossed. If the outcome is head, then the coin is flipped again.
  - The sample space is $S = \{T1, T2, T3, T4, T5, T6, HT, HH\}$
- **Example:** The life of a light bulb.
  - The sample space is $S = \{x: x \geq 0\} = [0, \infty)$
- **Example:** we toss a coin and stop until we get a head. The outcome is the number of tosses.
  - The sample space is $S = \{1,2, \cdots\}$

Sample Space and Event

- **Example:** Toss two coins.
  - The sample space is $S = \{HH, HT, TH, HH\}$.
  - An event can be $A = \{HH, HT, TH\}$ – at least one head.
- **Example:** Toss two coins. The number of heads is the outcome.
  - The sample space is $S = \{0,1,2\}$.
- **Example:** Toss two dice.
  - The sample space is $S = \{(1,1), (1,2), (1,3), \cdots, (6,6)\}$
  - An event can be $A = \{(5,6), (6,5), (6,6)\}$ – sum is at least 11
- **Example:** A bus arrives at between 10:00am to 11:00am.
  - The sample space is $S = (10{:}\,00\text{am}, 11{:}\,00\text{am})$

Probability

- **Definition** – **Probability** is the *measure of chance* (*likelihood*) that an event will occur.
- **Definition of Probability** - A rigorous definition here is difficult, can be considered as a “long-rang relative frequency” or “percentage”.
- **Example**: if we want to know the probability of getting a head when a coin is tossed, we can repeat the experiments many times, record the number and calculate percentage of heads obtained.
- **Example**: we can use the percentage of light blubs last more than 200 hours as probability of that a light bulb lasts more than 200 hours.

Properties of Probability

- **Probability** is the *measure of chance* (*likelihood*) that an event will occur.
  - Can its value be negative? **No**
  - Can its value be equal to 0? **Yes**
  - Can its value be equal to 1? **Yes**
  - Can its value be greater than 1? **No**
- Its value will be in the range from 0 to 1 (including 0 and 1).
- If A is an event, then the probability of A is denoted by $\Pr(A)$ or $P(A)$. And we have $0 \le \Pr(A) \le 1$.

Probability of Equally Likely Sample Space

- If A is an event, then the probability of A is denoted by $\Pr(A)$. We have

$$\Pr(A) = \frac{Size\ of\ the\ event\ A}{Size\ of\ the\ sample\ space\ S}$$

- In this case, size refers to the appropriate measure of chance. Essentially, $\Pr(A)$ represents size of the event A relative to size of sample space S.
- Size can be calculated using counting methods. This formula assumes that all the elements in S have the same probability (equally likely to happen).

A Simple Example – Fair Die

- **Example** – Roll a ***fair*** die and record the number obtained.
- What are the possible outcomes? $\mathbf{1, 2, 3, 4, 5, 6}$
- What is the sample space? $\boldsymbol{S} = \{\mathbf{1, 2, 3, 4, 5, 6})$
- Is $A = \{1,3\}$ an event and what is $\Pr(A)$? **Yes.** $\Pr(\boldsymbol{A})$**=1/3**
- Is $B = \{1, 3, 5\}$ an event and what is $\Pr(B)$? **Yes.** $\Pr(\boldsymbol{B})$**=1/2**
- Is $C = \{1\}$ an event and what is $\Pr(C)$? **Yes.** $\Pr(\boldsymbol{C})$**=1/6**
- Is $D = \{\}$ an event and what is $\Pr(D)$? **Yes.** $\Pr(\boldsymbol{D})$**=0**
- Is $E = \{1,2,3,4,5,6\}$ an event and what is $\Pr(E)$? **Yes.** $\Pr(\boldsymbol{E})$**=1**
- Define an event F = "the number is even" – $\boldsymbol{F} = \{\mathbf{2, 4, 6}\}$. $\Pr(\boldsymbol{F})$**=1/2**
- How many different events do we have for this experiment? **64**$= \mathbf{2^6}$

Example – Maintenance of Spinning Machines

- **Example** – A major manufacturer of textile fibers has several spinning “plants” at a single location. The central maintenance shop provides support for all major repairs and overhauls. Plant 2A has 60 spinning parts while plant 3 has 18 spinning parts. Assume that each part (in plant either 2A or 3) is equally likely to require maintenance (have problem).
- What is the probability that a service request is from Plant 3?

Example – Maintenance of Spinning Machines

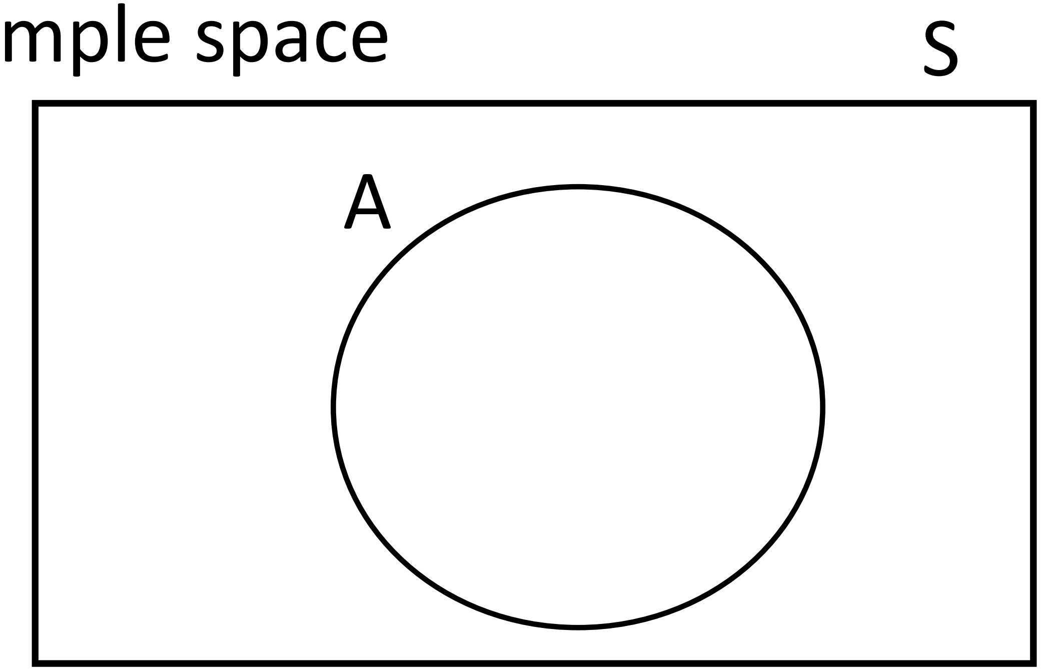
- Sample Space $S = \{1,2,3, \dots, 78\}$
- Define the event A as the spinning parts of Plant 3 need the service. Then $A = \{61, \dots, 78\}$.

$$\Pr(A) = \frac{Size\ of\ A}{Size\ of\ S} = \frac{18}{78} = \frac{3}{13}$$

- How to calculate the probability if we do not think each spinning part is equally likely to require maintenance?
    - Best way is to use basic rules of probability.

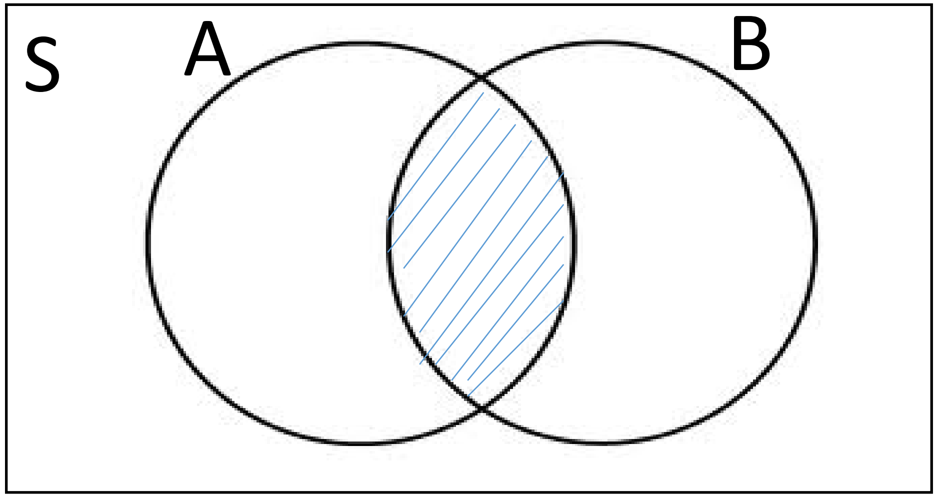
Events and Venn Diagram

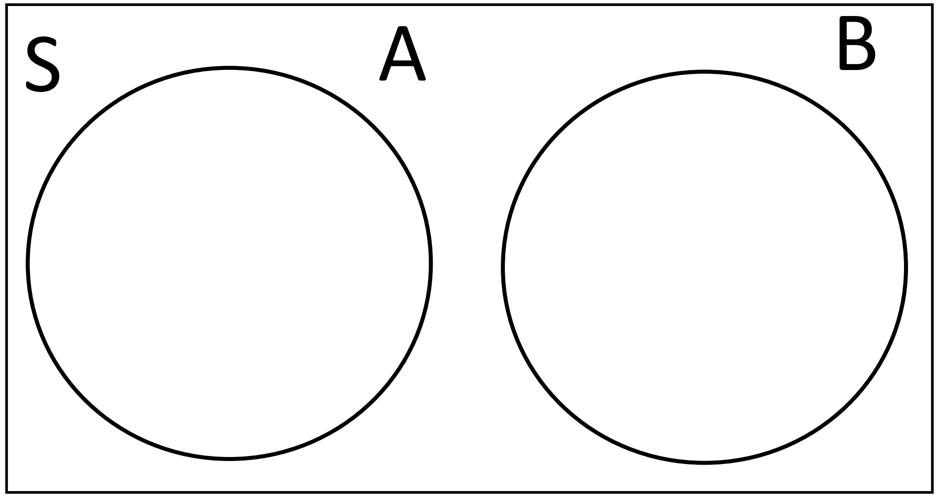
- We can use *Venn diagram* to:
  - Illustrate the relationship between multiple events.
  - Help us to calculate the corresponding probability.
  - Rectangle: represent the sample space
  - Other shapes inside: event
- An example of Venn diagram:



Event Relation: Intersection

- **Intersection**: $A \cap B$ (Both A and B or simply A and B)
- If the intersection of A and B is empty, $A \cap B = \emptyset$, then A and B are called **mutually exclusive (Definition 2.8)**.



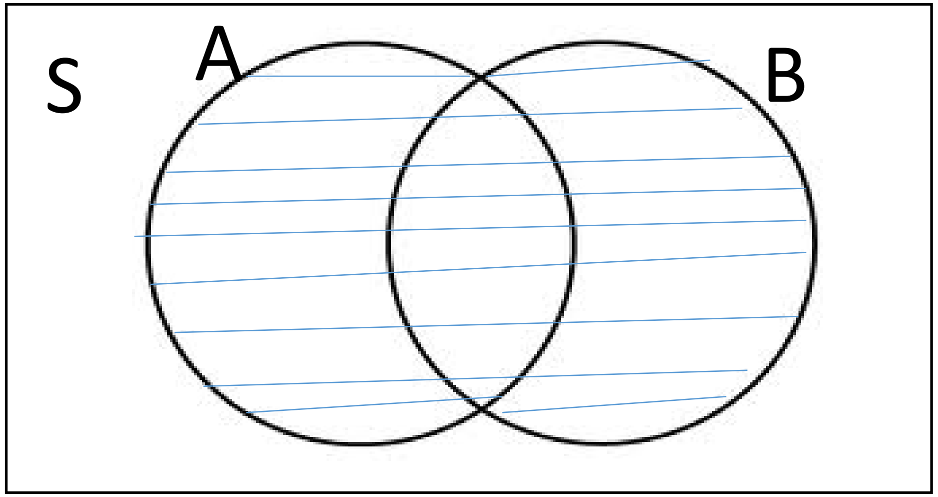


Event Relationship – Mutually Exclusive

- **Definition 2.8 –** If two events cannot occur simultaneously, that is, one "excludes" the other, then the two events are said to be **mutually exclusive** (in other words, if two events are mutually exclusion, then the intersection of them is an empty set).
- **Example:** If we toss a coin, the outcome $\{\text{Head}\}$ and $\{\text{Tail}\}$ are mutually exclusive since you can not get both a head and tail at the same time.
- **Example:** Raining at Chicago today and raining at Houghton today are NOT mutually exclusive.

Event Operation: Union

- **Union**: $A \cup B$ (A or B)
- **Union:** in other words, is *A* alone or *B* alone or *A* and *B* both



Event Operation: Complement

- **Complement** of A: A^c or $\bar{A}$ (Not A)

$$A \cap A^c = \emptyset$$

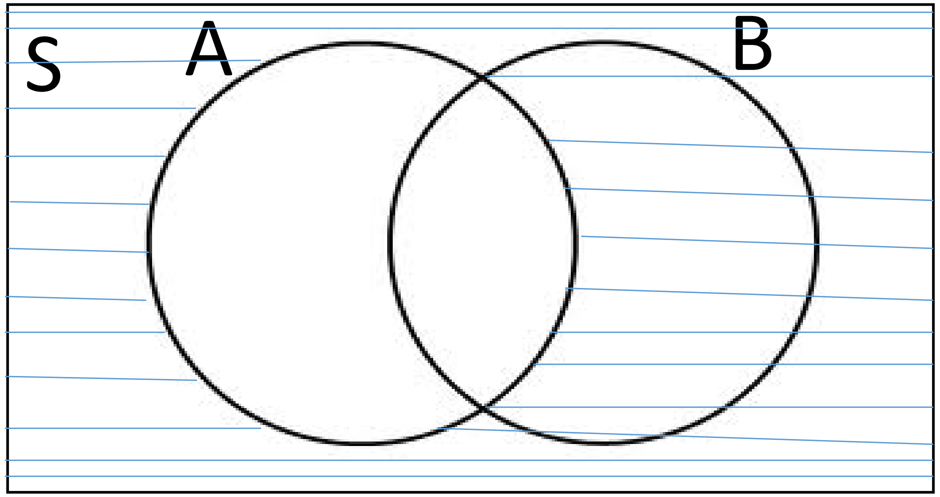
$$A \cup A^c = S$$

- For complement of intersection:

$$(A \cap B)^c = A^c \cup B^c$$

- For complement of union:

$$(A \cup B)^c = A^c \cap B^c$$



Event Relationship – Complement

- **Definition 2.9** – The **complement** of an outcome or event A is the occurrence of any event or outcome that precludes A from happening (also, the union of an event and its complement is an event with all possible outcomes).
- **Example:** There are at least two students in my class have the same birthday. Its **complement** is
  - All students have different birthday.

Event Relationship - Independent

- **Definition 2.10** – Two events A and B are said to be **independent** if the probability of A occurring is in no way affected by event B having occurred or vice versa.
- **Example:** Toss two coins. The outcome from the first toss is independent of the outcome from the second toss.
- **Example:** Are a person's study attitude to study MA5701 and his grade in MA5701 independent?

Rules for Probability Calculation

- **Probability of Empty Set**: $\Pr(\Phi) = 0$
- **Probability of An Event with All Outcomes:** $\Pr(S) = 1$
- **Probability of Intersection of Two Events (Independent):**
$$\Pr(A\ and\ B) = \Pr(A \cap B) = \Pr(A) * \Pr(B)$$
- **Probability of Union of Two Events (General Rule):**
$$\Pr(A\ or\ B) = \Pr(A \cup B) = \Pr(A) + \Pr(B) - \Pr(A \cap B)$$
- **Probability of Union of Two Events (Mutually Exclusive):**
$$\Pr(A\ or\ B) = \Pr(A \cup B) = \Pr(A) + \Pr(B)$$
$$\Pr(A \cup A^c) = 1 = \Pr(A) + \Pr(A^c)$$

A Simple Example – Die Example

- **Example** – Roll a fair die and record the number obtained.
- The sample space is $S = \{1,2,3,4,5,6\}$
- Is $A = \{2, 4, 6\}$ and $\Pr(A) = 1/2$, why?
- Is $B = \{1, 2, 3, 4\}$ and $\Pr(B) = 2/3$, why?
- Actually, A – the event that an outcome is an even number
- Actually, B – the event that the number is at most 4.
- Find the following probability (**Exercise**):

$$\Pr(A \cap B)\,; \Pr(A \cup B)$$

Calculation of Probability – Die Example

- The sample space is equally likely.
- $\Pr(A \cap B) = \Pr(\{4,6\}) = 1/3$
- $\Pr(A \cup B) = \Pr(\{1,2,3,4,6\}) = 5/6$
- $\Pr(A \cup B) = \Pr(A) + \Pr(B) - \Pr(A \cap B) = \frac{1}{2} + \frac{2}{3} - \frac{1}{3} = 5/6$

Calculation of Probability – More Examples

- **Example:** Suppose that 75% all investors invest traditional annuities and 45% of them invest in the stock market. If 85% invest in the stock market and/or traditional annuities, what percent invest in both?
- A = invest in traditional annuities; B = invest in stock market;
- $\Pr(A) = 0.75; \Pr(B) = 0.45; \Pr(A \cup B) = 0.85$
- Invest in both: $\Pr(A \cap B) = \Pr(A) \ + \Pr(B) \ - \Pr(A \cup B) = 0.35$
- Invest only in annuities : $\Pr(A \cap B^c) = \Pr(A) - \Pr(A \cap B) = 0.40$

Calculation of Probability – Exercises

- **Exercises:** Suppose that in a community of 400 adults, 300 bike or swim or do both, 160 swim, and 120 swim and bike, how many of them bike? For an adult selected at random from this community, what is the probability that he/she bikes?
- **Details:** skipped. Discussed in the class.
- **Solution:** 260 of them bike.
- **Solution:** the probability is $260/400 = 0.65$.

Probability of Intersection

- If A and B are independent, then $\Pr(A \cap B) = \Pr(A) \Pr(B)$
- If A and B are not independent, then the calculation of $\Pr(A \cap B)$ can be difficult and tricky – we may need to calculate probability of the event $A \cap B$ directly or rely on $\Pr(A)$, $\Pr(B)$, and $\Pr(A \cup B)$.

Independence – Dice Example

- Roll a fair dice two times, what is the probability getting two 6s?
- **Solution**:

$$\Pr(\text{First Dice is 6 and Second Dice is 6})$$

$$= \Pr(\text{First Dice is 6} \cap \text{Second Dice is 6})$$

$$= \Pr(\text{First Dice is 6}) * \Pr(\text{Second Dice is 6})$$

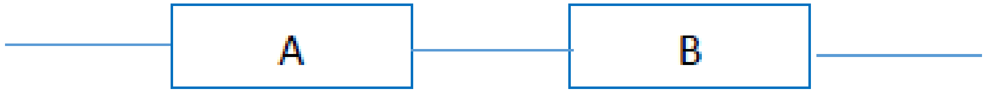
$$= \frac{1}{6} * \frac{1}{6} = \frac{1}{36}$$

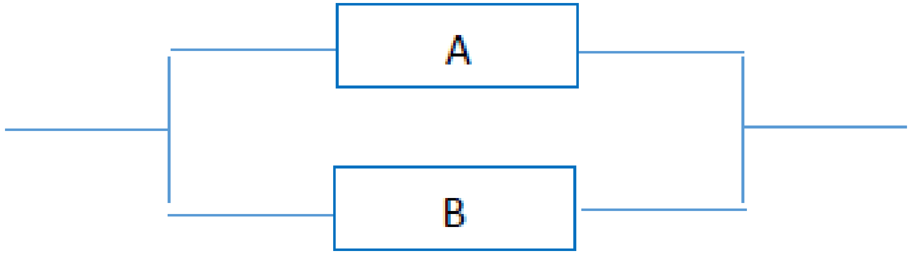
Probability Calculation – System Reliability

- **Reliability of System:** A failing is independent of B failing

  $\Pr(A) = \Pr(A \text{ failing}) = 0.01$ and $\Pr(B) = \Pr(B \text{ failing}) = 0.02$

- **Serial System:** $\Pr = \Pr(A \text{ or } B) = \Pr(A) + \Pr(B) - P(A \text{ and } B) = 0.0298$
- **Parallel System:** $\Pr = \Pr(A \text{ and } B) = \Pr(A) * \Pr(B) = 0.0002$





Exercise - Engineering Design Project Bids

- **Example** – Suppose a project manager of an engineering design firm bids on two projects. The chance that a bid will be accepted is: $\Pr(A) = 0.3; \Pr(B) = 0.8$. We can reasonably assume that bids are *independent*.

1. What is the probability that at least one of bids is accepted?
2. What is the probability that only bid A is accepted?

Solution - Engineering Design Project Bids

1. First, we have $\Pr(\mathrm{A} \cap \mathrm{B}) = \Pr(A) * \Pr(B) = 0.3 * 0.8 = 0.24$
2. What is the probability that at least one of bids is accepted?
   - $\Pr(A \cup B) = \Pr(A) + \Pr(B) - \Pr(A \cap B) = 0.3 + 0.8 - 0.24 = 0.86$
3. What is the probability that only bid A is accepted?
   - $\Pr(A \cap B^c) = \Pr(A) - \Pr(A \cap B) = 0.3 - 0.24 = 0.06$

Exercise – Number of Defects

- **Exercise:** Two factories monitor their production line and find the following distributions of the number of defects:

	0	**1**	**2**	$\geq \mathbf{3}$
Factory 1	0.92	0.04	0.03	0.01
Factory 2	0.95	0.03	0.02	0.00

- We further assume that the number of defects from factory 1 is independent of the number of defects from factory 2.
- A = at most two defects from factory 1; B = 1 defect from factory 2;
- C = total of 4 defects from Factory 1 and Factory 2
- Calculate: $\Pr(A)$; $\Pr(B)$; $\Pr(A \cap B)$; $\Pr(A \cup B)$; $\Pr(A \cup C)$; $\Pr(A \cap C)$

Random Variables

- **Definition 2.11** – A **random variable** is a rule that assigns a numerical value to an outcome of interest (in other words, a function from outcomes to real numbers).
- If Y is a random variable, then the ***cumulative distribution function*** (cdf), denoted by $F(y)$, is give by

$$F(y) = \Pr(Y \leq y)$$

for all real numbers $y(-\infty < y < \infty)$.

- Properties of $F(y)$:

  $0 \leq F(y) \leq 1$; non-decreasing; $F(-\infty) = 0$; $F(\infty) = 1$.

Random Variable - Example

- **Example 2.4 -** Distribution of number of heads of tossing two fair coins.
  - We know the sample space is $\{HH, HT, TH, TT\}$.
  - We assign 0 to $\{TT\}$
  - We assign 1 to $\{HT\}$ or $\{TH\}$
  - We assign 2 to $\{HH\}$.
  - Now the random variable take values 0, 1, 2.
- What is the cdf of this random variable?
  - $\Pr(Y \leq 1.5) = \Pr(Y = 0) + \Pr(Y = 1) = 0.75$
  - $\Pr(Y \leq 1) = \Pr(Y = 0) + \Pr(Y = 1) = 0.75$
  - $\Pr(Y < 1) = \Pr(Y = 0) = 0.25$

Random Variables – Why We Use Them?

- Why we use the random variable instead of the original sample space?
- Answer: data reduction and a better tool to describe the experiments.
- **Example:** suppose that we toss 100 fair coins.
  - The original sample space has 2^{100} elements.
  - If we define the number of heads as the random variable, then we have 101 possible values. Why?
  - Actually, we are more (or only) interested in knowing the number of heads than knowing which tosses result in a head.

Random Variable – Another Example

- **Example:** The random variable is the life of the light bulb in years.
- For this example, the sample space is $S = [0, \infty)$.
- We assign the same value to each value in S.
- There are many possible choices of cumulative probability function (cdf)for this random variable.
- For example, we can use $F(y) = \Pr(Y \leq y) = 1 - \exp(-\frac{y}{3})$.
- Based on this cdf, we can calculate

$$\Pr(Y > 10) = 1 - \Pr(Y \leq 10) = \exp\left(-\frac{10}{3}\right) = 0.037$$

Discrete and Continuous Random Variables

- **Definition 2.5 –** A **discrete random variable** is one that can take on only a countable number of values.
- **Definition 2.6 –** A **continuous random variable** is one that can take on any value in an interval.
- A more formal definition based on the cumulitave probability function $F(y)$: Y is a discreet random variable if $\boldsymbol{F(y)}$ **is a step function** while Y is a continuous random variable if $\boldsymbol{F(y)}$ **is a continuous function**.

Discrete and Continuous Random Variables

- **Example:** Toss two dice and the random variable is the sum of two numbers. This is a discrete random variable.
- **Example:** The lifetime of a light bulb. This is a continuous random variable.
- **Note:** Some random variables can be neither discrete nor continuous.

Exercise – Types of Random Variable

- Y is the aluminum contamination from an area.
  - Continuous random variable.
- Y is the number obtained when we roll a dice.
  - Discrete random variable.
- Y is the number of car accidents in Houghton.
  - Discrete random variable.
- Y is the strength of yarns from a manufacturer.
  - Continuous random variable.

Discrete Random Variable

- The **probability mass function ($\boldsymbol{pmf}$)** for a discrete random variable, denoted by $f(y)$, is defined by $f(y) = \Pr(Y = y)$.
- Two properties of $f(y)$ are:
  1. Non-negative: $0 \le f(y) \le 1$.
  2. $\sum_y f(y) = 1$.
- Is $f(y)$ a monotone function?
  - May not be a monotone function.
- The cdf can be calculated through pmf: $F(y) = \sum_{t \le y} f(t)$

Discrete Random Variable - Example

Example – Roll a fair dice, let Y be the number obtained.

- Is Y a discrete random variable?
- What are the possible values of Y?
- What are the cdf and pmf of Y?
- The pmf of Y: $f(y) = \frac{1}{6}, y = 1, \cdots, 6$
- The cdf is a step function. For example, $F(1) = \Pr(Y = 1) = 1/6$, $F(2.5) = \Pr(Y = 1) + \Pr(Y = 2) = \frac{1}{3}$.
- How about $F(-1)$ and $F(10)$? $F(-1) = 0; F(10) = 1$

Expected Values and Population Mean

- **Definition –** The **expected value** of the discrete random variable is the probability-weighted average of all possible values.
- This value is also called population mean and denoted by μ.
- The mathematical formula is based on the possible values and their probability mass function.

$$\mu = E[Y] = \sum_{y} f(y) * y$$

- Intuitively it is the long-run average value of repetitions of the experiment it represents.

Discrete Random Variable - Example

- **Example 2.4 -** Distribution of number of heads of tossing two fair coins. We know that $f(0) = 0.25; f(1) = 0.50; f(2) = 0.25.$
- The expected value is $\mu = 0.25 * 0 + 0.50 * 1 + 0.25 * 2 = 1.$
- This value can be considered as the long-run average value of repetitions of the experiment it represents, what does it mean?
- That means we conduct such experiment (tossing two fair coins) many times, the average number of heads (the sample mean) is (approximately) 1. If we conduct such experiment infinite times, then the average number of heads is 1.

Population Variance

- **Definition –** The **variance** of the discrete random variable is the probability-weighted average of squared distance of all possible values and its expected value.
- Again, this variance is considered as population variance and can be estimated by sample variance.
- **Population variance** of Y, denoted by σ^2, is

$$\sigma^2 = \text{var}(Y) = \sum_y f(y) * (y - \mu)^2.$$

- **Population standard deviation** of Y, denoted by σ, is $\sigma = \sqrt{\sigma^2}$, the square root of the population variance.

Population Variance - Example

- **Example** – Roll a fair dice. let $Y = 0$ if the number is not greater than 4 and $Y = 1$ otherwise.
- We know that $f(0) = 2/3$ and $f(1) = 1/3$, why?
- Population mean is:

$$\mu = 0 * f(0) + 1 * f(1) = 1/3$$

- Population variance is:

$$\sigma^2 = (0 - \frac{1}{3})^2 * f(0) + \left(1 - \frac{1}{3}\right)^2 * f(1) = \frac{1}{9} * \frac{2}{3} + \frac{4}{9} * \frac{1}{3} = \frac{2}{9}$$

Exercise – Mean and Variance

- **Exercise:** A factory monitor their production line and find the following distributions of the number of defects:

# Defects	**0**	**1**	**2**	**3**
Probability	0.92	0.04	0.03	0.01

- Let Y be the number defects from this factory, then Y is a discrete random variable.
- Find $\Pr(Y \geq 1)$, $\Pr(Y \leq 2)$, and $\Pr(Y \leq 10)$.
- Find the mean and variance of Y.
- **Solution:** in class.

Common Discrete Distributions

What important things do you need to memorize for a discrete distribution (random variable)?

- The **possible values** that the random variable can take.
- The **probability mass function**.
- The **population mean** and **population variance**.

You also want to understand -

- What type of experiments can be described by such a random variable.

Discreet Uniform Distribution

- A random variable Y has a *discrete uniform* $(1, N)$ distribution if $\Pr(Y = k) = \frac{1}{N}, k = 1, \cdots, N$.
- Population mean: $\mu = E[Y] = \frac{N+1}{2}$.
- Population variance: $\sigma^2 = \mathrm{var}(Y) = \frac{(N+1)*(N-1)}{12}$
- Examples: Y = number obtained from rolling a fair dice

$$Y \sim U(1,6), E[Y] = 3.5, \mathrm{var}(Y) = \frac{35}{12} = 2.917$$

Bernoulli Distribution

- A random variable Y has a *Bernoulli*(p) distribution if $f(1) = \Pr(Y = 1) = p$ and $f(0) = \Pr(Y = 0) = 1 - p$.
- Population mean: $\mu = E[Y] = p$.
- Population variance: $\sigma^2 = \mathrm{var}(Y) = p(1 - p)$.
- Bernoulli distribution is used to describe a trial (an experiment) with two outcomes: success and failure. And p is the probability of success.

Bernoulli Distribution - Examples

- Toss a coin, define $Y = 1$ if a head obtained and $Y = 0$ otherwise. Now p is the probability to get a head.
- In an election poll, define $Y = 1$ if candidate A gets a vote and $Y = 0$ otherwise. Now p is the probability that candidate A gets a vote.
- In a manufacturer, define $Y = 1$ if a battery plate meets specifications and $Y = 0$ otherwise. Now p is the probability that a battery plate meets specifications.

Binomial Distribution

- A random variable Y has a *Binomial distribution*, Binomial(n, p), if its pmf is

$$f(y) = \Pr(Y = y) = \binom{n}{y} p^y (1-p)^{n-y}, y = 0, \cdots, n.$$

- If Y is the number of successes from n independent identical Bernoulli trials, then Y has Binomial(n, p).
- If Y~Bernoulli(p) then Y~ Binomial$(1, p)$.
- Population mean: $\mu = E[Y] = np$.
- Population variance: $\sigma^2 = \operatorname{var}(Y) = np(1-p)$.

Binomial Distribution

Consider an experiment satisfies these four conditions:

1. The number of experiments n must be fixed.
2. The trials are independent.
3. Each trial has two mutually exclusive outcomes, success or failure.
4. The probability of success (p) is fixed for each trial of the experiment.

Then Y = the number of successes has Binomial(n, p).

Binomial Distribution - Examples

Determine if the following distributions are binomial:

1. Toss a coin 100 times, Y = number of heads.
2. Roll a fair dice 10 times, Y = the number of rolls getting number greater than 4.
3. Suppose our class has 55 students and 30 of them are female students. I randomly select 5 students. $Y\ =$ number of females.
4. The probability to get a defected bulb is 3%. There are 1000 bulbs produced in one day. $Y\ =$ number of defected bulbs.

Answer: 1, 2, and 4 have a binomial distribution, while 3 does not. The distribution of 3 is hypergeometric distribution.

Example – Nonconforming Brick

- **Example** - Marcucci (1985) reports that a brick manufacturer classifies the product into conforming and nonconforming groups. Historically, nonconforming bricks counts 5% of all bricks. A facility makes 25 bricks per hour. Let $Y \ =$ number of nonconforming bricks. Then Y has Binomial$(25, 0.05)$.
- Probability of two non-conforming bricks is:

$$\Pr(Y = 2) = \binom{25}{2} 0.05^2(1 - 0.05)^{25-2} = 0.2305.$$

- Probability of at least one non-conforming brick is:

$$\Pr(Y \geq 1) = 1 - \Pr(Y = 0) = 1 - \binom{n}{0} 0.05^0(1 - 0.05)^{25-0} = 0.7226.$$

Example – Nonconforming Brick

- Consider the population mean, $\mu = np = 25 * 0.05 = 1.25$.
- Consider the population variance:
$$\sigma^2 = np(1-p) = 25 * 0.05 * 0.95 = 1.1875$$
- This analysis assumes that:
  - Each brick is independent of the other.
  - The historical probability of a brick being defective holds true for the particular hour.

Exercises – Find Distribution and Probability

- **Exercise:** A typesetter, on the average, makes one error in every 500 words typeset. A typical page contains 300 words. What is the probability that there will be no more than two errors in five pages?
- **Solution:** Let Y be the number of errors, then $Y \sim Binomial(1500, 0.002)$.

$$\Pr(Y \leq 2) = \sum_{y=0}^{2} \binom{1500}{y} 0.002^{y} (1 - 0.002)^{1500-y} = 0.4230$$

Exercises – Find Distribution

- A manufacturer receives a lot of 100 parts from a vendor. The probability of the part is defective is 0.01. Let Y be the number of parts that are defective.
  - $Y \sim Binomial(100, 0.01)$
- A manufacturer of water filters for refrigerators monitors the process for defective filters (the filer leaks). Historically, this process averages 5% defective filters. 20 filters are randomly chosen for testing and let Y be the number of defective filters.
  - $Y \sim Binomial(20, 0.05)$

Exercises – Find Distribution and Probability

- A standard drug is known to be effective in 80% of the cases in which it is used. The drug is tested on 100 patients and found to be effective in 85 cases. What distribution can be used to calculate the probability that we can observed at least 85 effective cases among 100 tested patients?
  - $Y \sim Binomial(100, 0.80)$
  - $\Pr(Y \geq 85) = \binom{100}{85} 0.80^{85} 0.20^{100-85} + \cdots + \binom{100}{100} 0.80^{100} 0.20^{100-100}$

$$= \sum_{i=85}^{100} \binom{100}{i} 0.80^{i} 0.20^{100-i} = 0.1285$$

Continuous Random Variable

- A **continuous random variable** is one that can be any possible real value over some interval.
- For any random variable (continuous, discrete, or other types), we have the cumulative probability function $F(y) = \Pr(Y \leq y)$.
- Properties of $F(y)$:

$$0 \leq F(y) \leq 1; \text{ increasing; } F(-\infty) = 0; F(\infty) = 1$$

- A more formal definition based on $F(y)$: Y is a discreet random variable if $\boldsymbol{F(y)}$ **is a step function** while Y is a continuous random variable if $\boldsymbol{F(y)}$ **is a continuous function.**

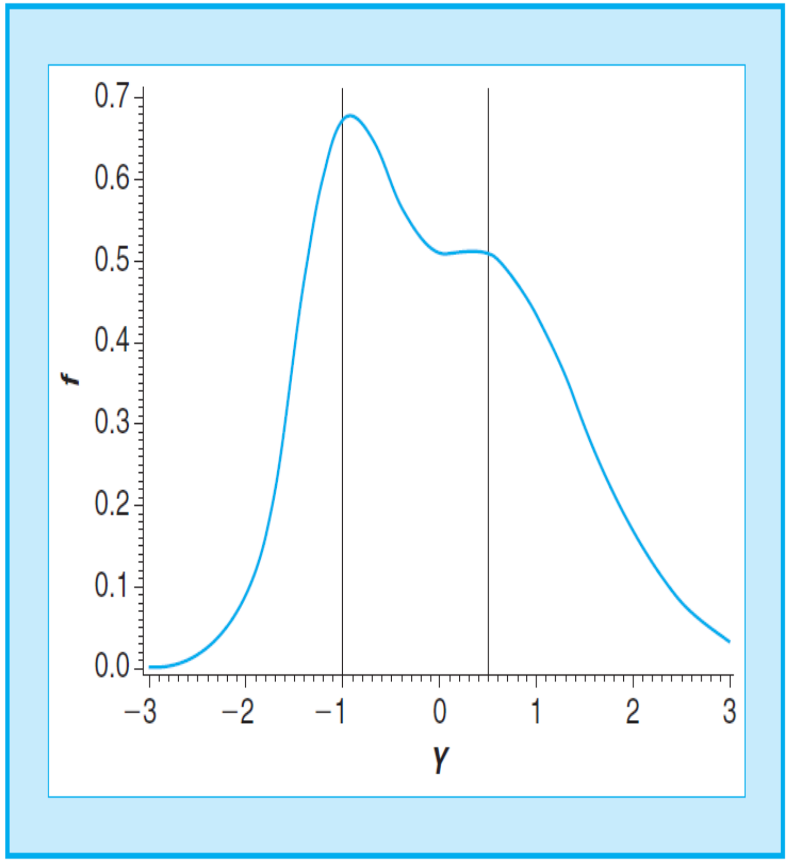
Continuous Random Variables

- Recall for a discrete random variable, we have the **probability mass function ($\boldsymbol{pmf}$)** $f(y) = \Pr(Y = y)$ and $F(y) = \sum_{t \le y} f(t)$.
- For a continuous random variable, we have $\Pr(Y = y) = 0$.
- The **probability density function ($\boldsymbol{pdf}$)** $f(y)$ for a continuous random variable satisfies:

$$F(y) = \Pr(Y \le y) = \int_{-\infty}^{y} f(t)dt$$

Continuous Probability Distribution

- The graph of the distribution is a smooth curve.
- The total area under the curve is 1.
- The area between the curve and horizontal axis from a to b is the probability of the random variable taking a value in the interval (a, b)
  - The probability of taking a specific value is zero and for continues random variables, therefore
    $$\Pr(a \leq X \leq b) = \Pr(a < X < b)$$
- Finding areas under curves (probability) can be difficult – involving the use of calculus. Sometimes, there is no analytical solution.



Probability Density Function

Properties of pdf:

1. $f(y) \geq 0.$
2. $\int_{-\infty}^{\infty} f(t)dt = 1 = F(\infty).$
3. $F(y) = \int_{-\infty}^{y} f(t)dt$

- Actually, any function that satisfies conditions (1) and (2) can be considered as a valid pdf.
- **Question:** do we have $f(y) \leq 1$ for all y?
  - No. For pdf, $f(y)$ can be greater than 1 for some y. For pmf, $f(y) \leq 1$.

Common Continuous Distributions

What important things do you need to memorize for a continuous distribution (random variable)?

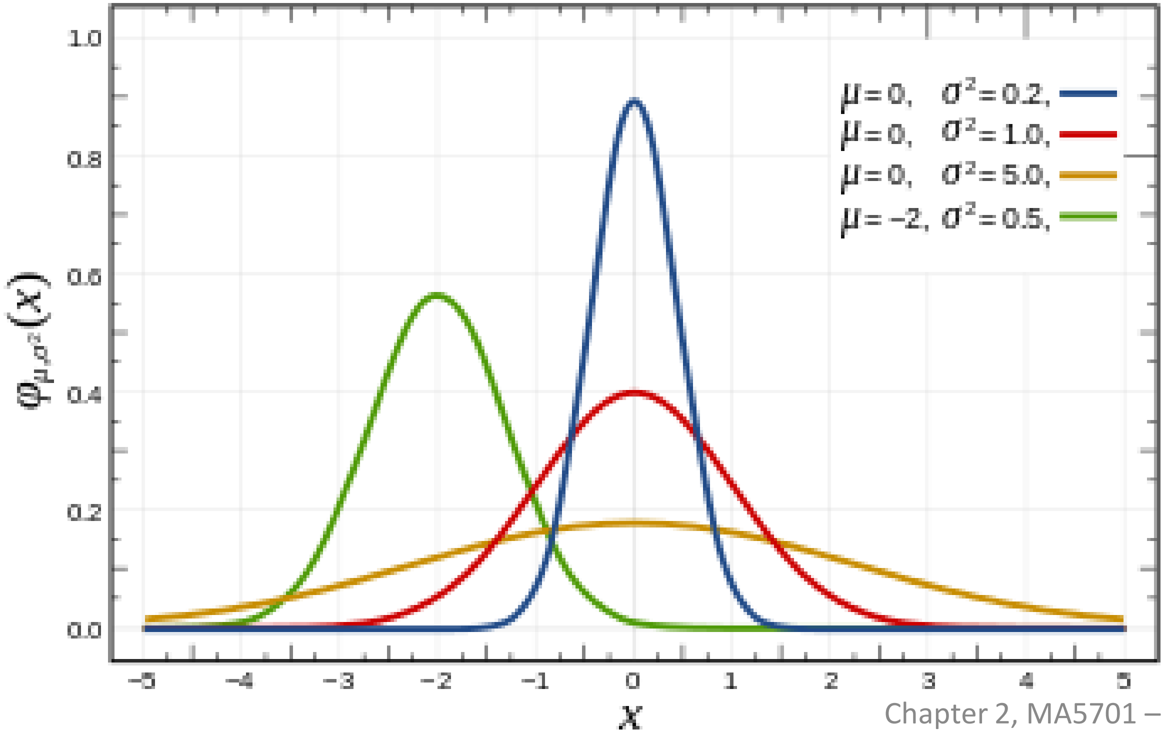
- The **possible values** that the random variable can take.
- The **probability density function**.
- The **population mean** and **population variance**.

You also want to understand -

- What type of experiments can be described by such a random variable.

Normal Distribution

- The most often commonly continuous probability function.
- Probability density function: $f(y) = \frac{1}{\sqrt{2\pi}\sigma} \exp(-\frac{(y-\mu)^2}{2\sigma^2}), -\infty < y < \infty$



Normal Distribution

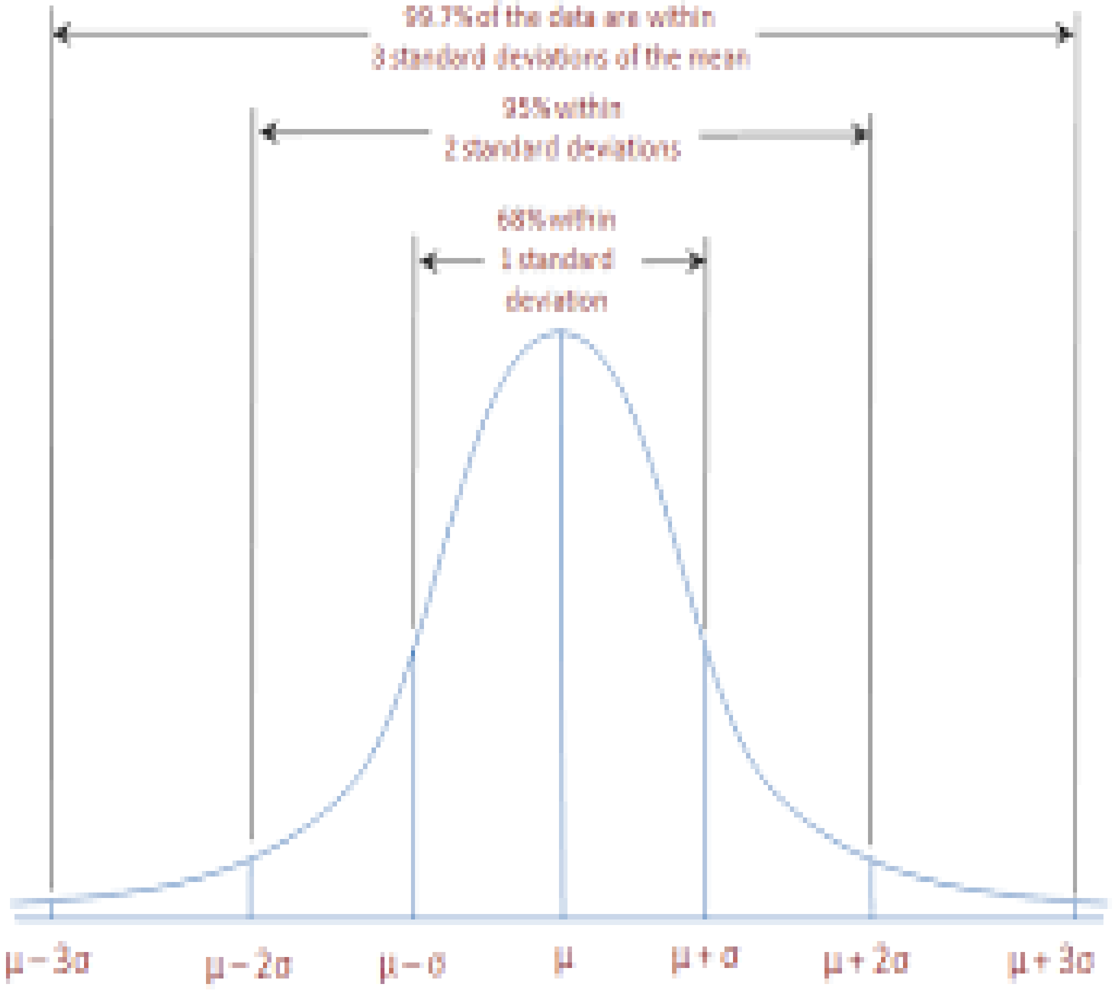
- It is the single most important distribution, which is often described as “bell-shaped” distribution or curve.
- It can be used to model the behavior of many phenomena.
- Under certain conditions, it can be used to model the behavior of averages.
- Many times, we use the simple notation: $Y \sim N(\mu, \sigma^2)$
- Population mean is: μ.
- Population variance and standard deviation are: σ^2 and σ.

Normal Distribution

- It is symmetric at its mean μ.
- Mean = median = mode, this is a single peak and the highest point occurs at $y = \mu$.
- The area under the curve is 1.
- The area under the curve to the left (right) of μ is 0.5.
- The curve never reaches the horizontal axis.

Normal Distribution – Empirical Rule

- Approximately 68% of data fall within $\mu \pm \sigma$
- Approximately 95% of data fall within $\mu \pm 2\sigma$
- Approximately 99.7% of data fall within $\mu \pm 3\sigma$


99.7% of the data are within 3 standard deviations of the mean
95% within 2 standard deviations
68% within 1 standard deviation
$\mu - 3\sigma$
$\mu - 2\sigma$
$\mu - \sigma$
μ
$\mu + \sigma$
$\mu + 2\sigma$
$\mu + 3\sigma$

Standard Normal Distribution

- If $Y{\sim}N(\mu, \sigma^2)$ and $\mu = 0$ and $\sigma^2 = \sigma = 1$, then Y has a standard normal distribution.
- In other words, the standard normal distribution is a normal distribution with mean of 0 and population variance (standard deviation) of 1.
- The cdf for the standard normal distribution can be found in a table.

Standard Normal Distribution

- It is the most important normal distribution. Why?
- If $Y{\sim}N(\mu, \sigma^2)$, then we can not analytically integrate to get $cdf\ F(y)$ - we must rely on numerical integration.
- Fortunately, any normal distribution can be converted to a standard normal distribution.
- So we can summarize the cdf of any normal distribution with the cdf of the standard normal distribution.

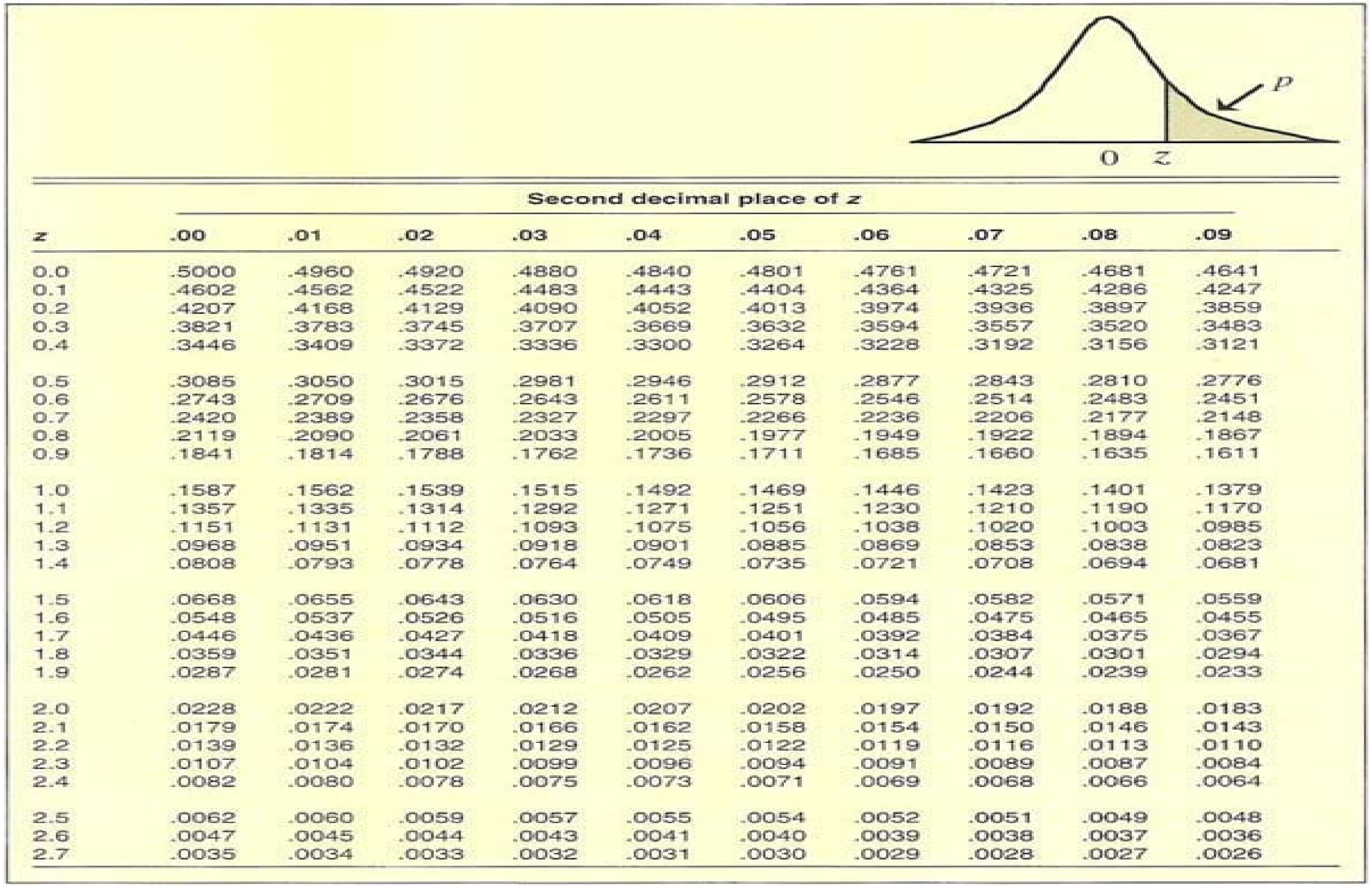
Normal Distribution: Calculations

- We generally use $Z \sim N(0,1)$ to represent the standard normal random variable.

$$\Pr(Z > z) = \Pr(Z < -z) = 1 - \Pr(Z < z)$$
$$\Pr(a < Z < b) = \Pr(Z < b) - \Pr(Z < a)$$
$$\Pr(Z > 0) = \Pr(Z < 0) = 0.5$$

- We rely on normal tables, which has many formats.

Standard Normal Distribution



	Second decimal place of z									
z	.00	.01	.02	.03	.04	.05	.06	.07	.08	.09
0.0	.5000	.4960	.4920	.4880	.4840	.4801	.4761	.4721	.4681	.4641
0.1	.4602	.4562	.4522	.4483	.4443	.4404	.4364	.4325	.4286	.4247
0.2	.4207	.4168	.4129	.4090	.4052	.4013	.3974	.3936	.3897	.3859
0.3	.3821	.3783	.3745	.3707	.3669	.3632	.3594	.3557	.3520	.3483
0.4	.3446	.3409	.3372	.3336	.3300	.3264	.3228	.3192	.3156	.3121
0.5	.3085	.3050	.3015	.2981	.2946	.2912	.2877	.2843	.2810	.2776
0.6	.2743	.2709	.2676	.2643	.2611	.2578	.2546	.2514	.2483	.2451
0.7	.2420	.2389	.2358	.2327	.2297	.2266	.2236	.2206	.2177	.2148
0.8	.2119	.2090	.2061	.2033	.2005	.1977	.1949	.1922	.1894	.1867
0.9	.1841	.1814	.1788	.1762	.1736	.1711	.1685	.1660	.1635	.1611
1.0	.1587	.1562	.1539	.1515	.1492	.1469	.1446	.1423	.1401	.1379
1.1	.1357	.1335	.1314	.1292	.1271	.1251	.1230	.1210	.1190	.1170
1.2	.1151	.1131	.1112	.1093	.1075	.1056	.1038	.1020	.1003	.0985
1.3	.0968	.0951	.0934	.0918	.0901	.0885	.0869	.0853	.0838	.0823
1.4	.0808	.0793	.0778	.0764	.0749	.0735	.0721	.0708	.0694	.0681
1.5	.0668	.0655	.0643	.0630	.0618	.0606	.0594	.0582	.0571	.0559
1.6	.0548	.0537	.0526	.0516	.0505	.0495	.0485	.0475	.0465	.0455
1.7	.0446	.0436	.0427	.0418	.0409	.0401	.0392	.0384	.0375	.0367
1.8	.0359	.0351	.0344	.0336	.0329	.0322	.0314	.0307	.0301	.0294
1.9	.0287	.0281	.0274	.0268	.0262	.0256	.0250	.0244	.0239	.0233
2.0	.0228	.0222	.0217	.0212	.0207	.0202	.0197	.0192	.0188	.0183
2.1	.0179	.0174	.0170	.0166	.0162	.0158	.0154	.0150	.0146	.0143
2.2	.0139	.0136	.0132	.0129	.0125	.0122	.0119	.0116	.0113	.0110
2.3	.0107	.0104	.0102	.0099	.0096	.0094	.0091	.0089	.0087	.0084
2.4	.0082	.0080	.0078	.0075	.0073	.0071	.0069	.0068	.0066	.0064
2.5	.0062	.0060	.0059	.0057	.0055	.0054	.0052	.0051	.0049	.0048
2.6	.0047	.0045	.0044	.0043	.0041	.0040	.0039	.0038	.0037	.0036
2.7	.0035	.0034	.0033	.0032	.0031	.0030	.0029	.0028	.0027	.0026

Standard Normal Distribution

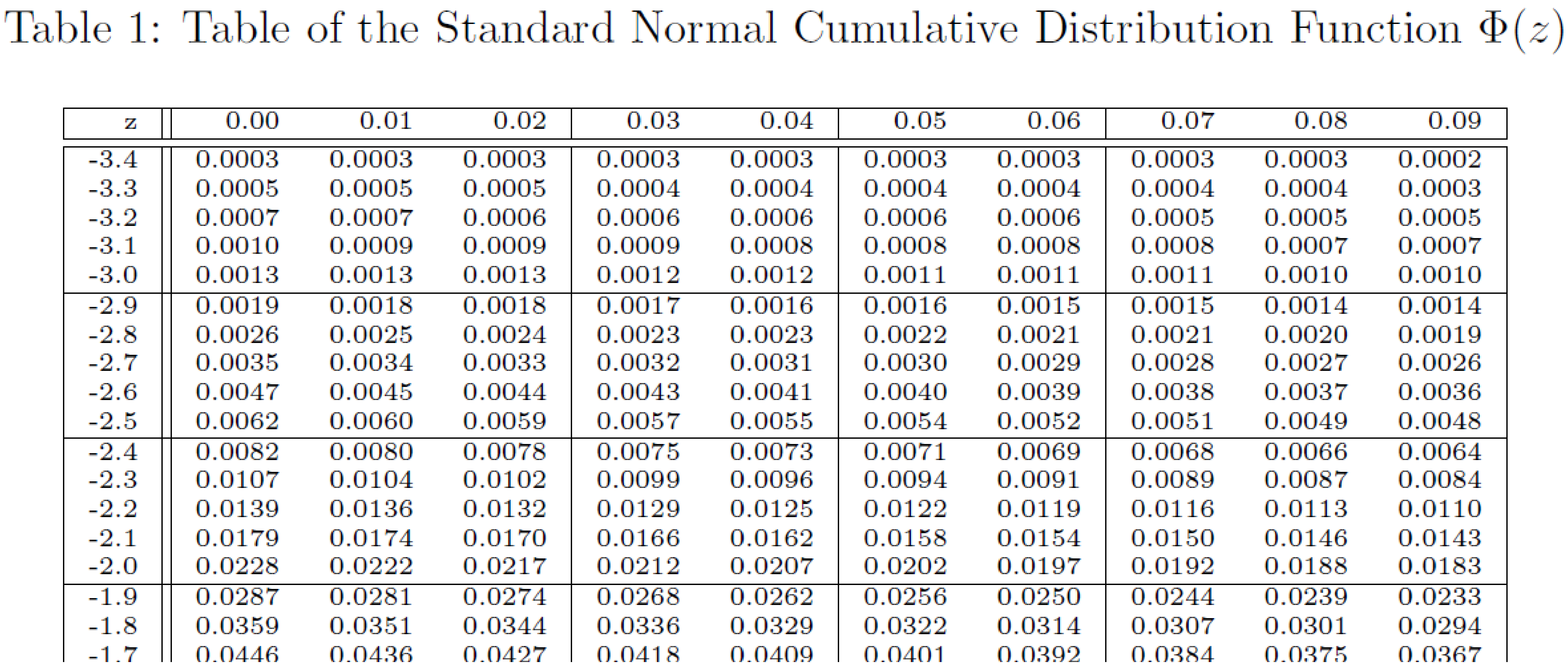


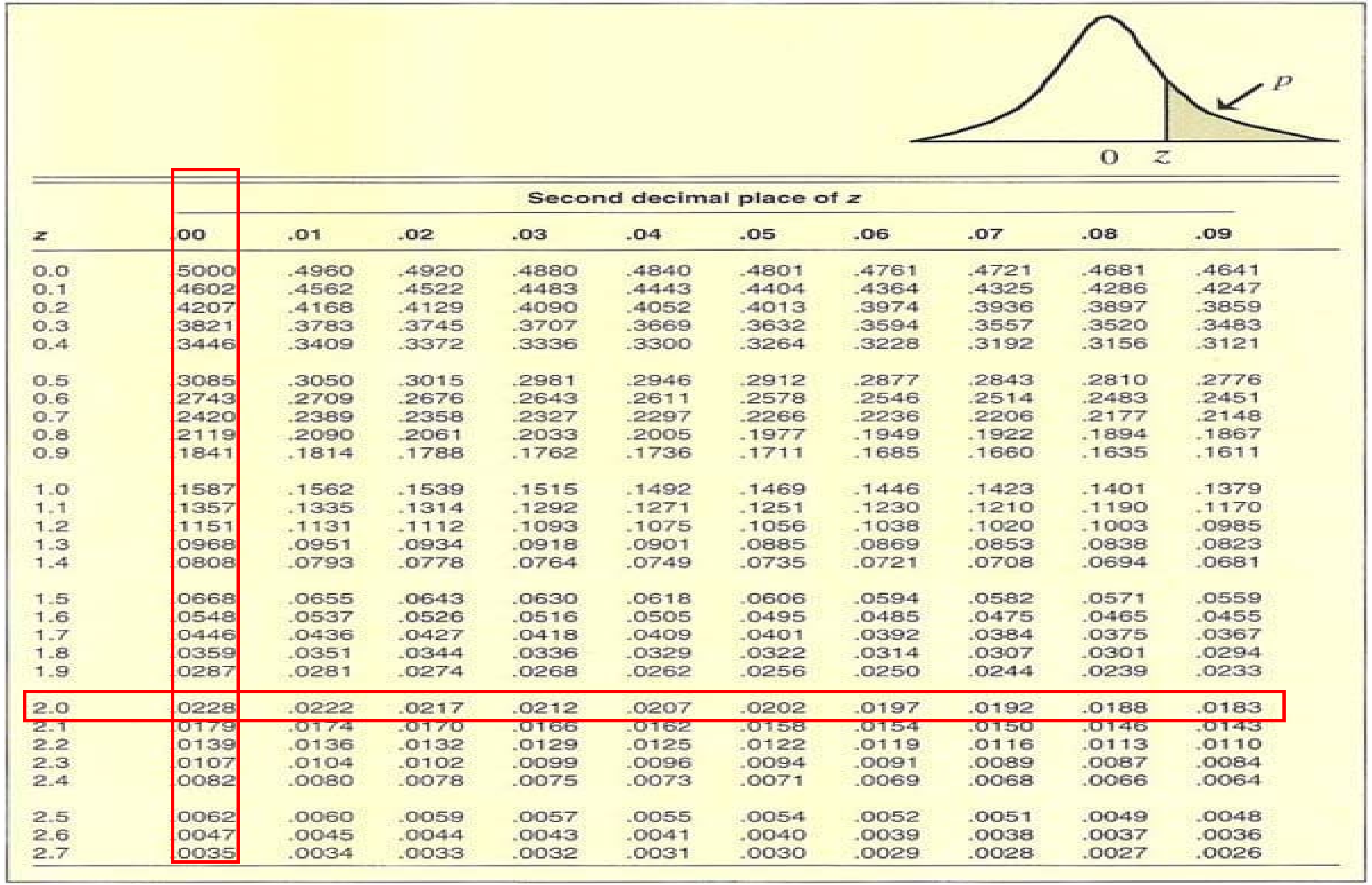
Table 1: Table of the Standard Normal Cumulative Distribution Function $\Phi(z)$

z	0.00	0.01	0.02	0.03	0.04	0.05	0.06	0.07	0.08	0.09
-3.4	0.0003	0.0003	0.0003	0.0003	0.0003	0.0003	0.0003	0.0003	0.0003	0.0002
-3.3	0.0005	0.0005	0.0005	0.0004	0.0004	0.0004	0.0004	0.0004	0.0004	0.0003
-3.2	0.0007	0.0007	0.0006	0.0006	0.0006	0.0006	0.0006	0.0005	0.0005	0.0005
-3.1	0.0010	0.0009	0.0009	0.0009	0.0008	0.0008	0.0008	0.0008	0.0007	0.0007
-3.0	0.0013	0.0013	0.0013	0.0012	0.0012	0.0011	0.0011	0.0011	0.0010	0.0010
-2.9	0.0019	0.0018	0.0018	0.0017	0.0016	0.0016	0.0015	0.0015	0.0014	0.0014
-2.8	0.0026	0.0025	0.0024	0.0023	0.0023	0.0022	0.0021	0.0021	0.0020	0.0019
-2.7	0.0035	0.0034	0.0033	0.0032	0.0031	0.0030	0.0029	0.0028	0.0027	0.0026
-2.6	0.0047	0.0045	0.0044	0.0043	0.0041	0.0040	0.0039	0.0038	0.0037	0.0036
-2.5	0.0062	0.0060	0.0059	0.0057	0.0055	0.0054	0.0052	0.0051	0.0049	0.0048
-2.4	0.0082	0.0080	0.0078	0.0075	0.0073	0.0071	0.0069	0.0068	0.0066	0.0064
-2.3	0.0107	0.0104	0.0102	0.0099	0.0096	0.0094	0.0091	0.0089	0.0087	0.0084
-2.2	0.0139	0.0136	0.0132	0.0129	0.0125	0.0122	0.0119	0.0116	0.0113	0.0110
-2.1	0.0179	0.0174	0.0170	0.0166	0.0162	0.0158	0.0154	0.0150	0.0146	0.0143
-2.0	0.0228	0.0222	0.0217	0.0212	0.0207	0.0202	0.0197	0.0192	0.0188	0.0183
-1.9	0.0287	0.0281	0.0274	0.0268	0.0262	0.0256	0.0250	0.0244	0.0239	0.0233
-1.8	0.0359	0.0351	0.0344	0.0336	0.0329	0.0322	0.0314	0.0307	0.0301	0.0294
-1.7	0.0446	0.0436	0.0427	0.0418	0.0409	0.0401	0.0392	0.0384	0.0375	0.0367

Normal Distribution: Calculations

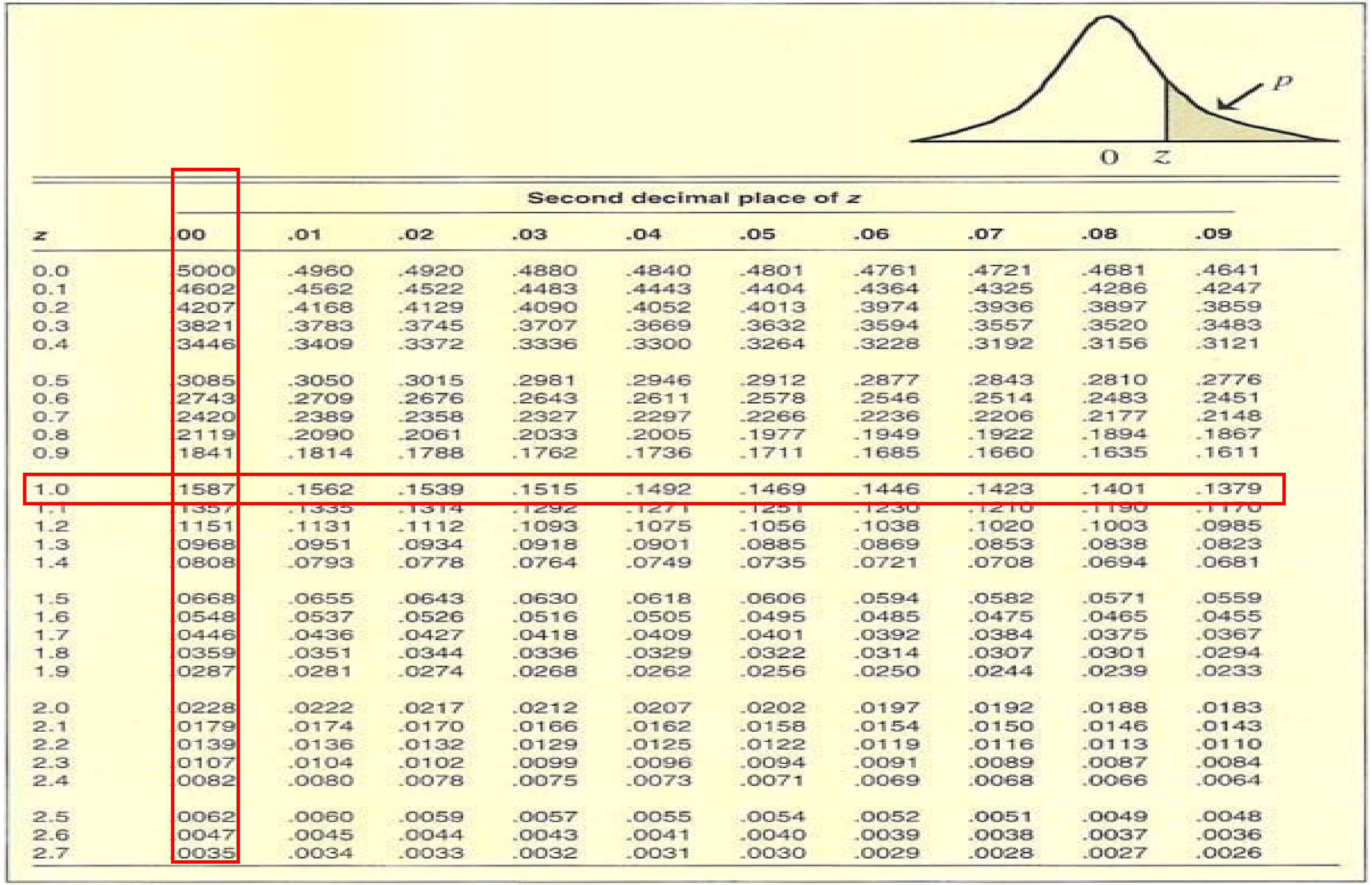
- $\Pr(Z > 2.0) = 0.0228$
- $\Pr(Z < -1) = \Pr(Z > 1) = 0.1587$
- $\Pr(Z < 1.53) = 1 - \Pr(Z > 1.53) = 1 - 0.0630 = 0.937$
- $\Pr(-1 < Z < 1.53) = \Pr(Z < 1.53) - \Pr(Z < -1) = 0.937 - 0.1587 = 0.7783$

Normal Distribution: $\Pr(Z > \mathbf{2.0})$



	Second decimal place of z									
z	.00	.01	.02	.03	.04	.05	.06	.07	.08	.09
0.0	.5000	.4960	.4920	.4880	.4840	.4801	.4761	.4721	.4681	.4641
0.1	.4602	.4562	.4522	.4483	.4443	.4404	.4364	.4325	.4286	.4247
0.2	.4207	.4168	.4129	.4090	.4052	.4013	.3974	.3936	.3897	.3859
0.3	.3821	.3783	.3745	.3707	.3669	.3632	.3594	.3557	.3520	.3483
0.4	.3446	.3409	.3372	.3336	.3300	.3264	.3228	.3192	.3156	.3121
0.5	.3085	.3050	.3015	.2981	.2946	.2912	.2877	.2843	.2810	.2776
0.6	.2743	.2709	.2676	.2643	.2611	.2578	.2546	.2514	.2483	.2451
0.7	.2420	.2389	.2358	.2327	.2297	.2266	.2236	.2206	.2177	.2148
0.8	.2119	.2090	.2061	.2033	.2005	.1977	.1949	.1922	.1894	.1867
0.9	.1841	.1814	.1788	.1762	.1736	.1711	.1685	.1660	.1635	.1611
1.0	.1587	.1562	.1539	.1515	.1492	.1469	.1446	.1423	.1401	.1379
1.1	.1357	.1335	.1314	.1292	.1271	.1251	.1230	.1210	.1190	.1170
1.2	.1151	.1131	.1112	.1093	.1075	.1056	.1038	.1020	.1003	.0985
1.3	.0968	.0951	.0934	.0918	.0901	.0885	.0869	.0853	.0838	.0823
1.4	.0808	.0793	.0778	.0764	.0749	.0735	.0721	.0708	.0694	.0681
1.5	.0668	.0655	.0643	.0630	.0618	.0606	.0594	.0582	.0571	.0559
1.6	.0548	.0537	.0526	.0516	.0505	.0495	.0485	.0475	.0465	.0455
1.7	.0446	.0436	.0427	.0418	.0409	.0401	.0392	.0384	.0375	.0367
1.8	.0359	.0351	.0344	.0336	.0329	.0322	.0314	.0307	.0301	.0294
1.9	.0287	.0281	.0274	.0268	.0262	.0256	.0250	.0244	.0239	.0233
2.0	.0228	.0222	.0217	.0212	.0207	.0202	.0197	.0192	.0188	.0183
2.1	.0179	.0174	.0170	.0166	.0162	.0158	.0154	.0150	.0146	.0143
2.2	.0139	.0136	.0132	.0129	.0125	.0122	.0119	.0116	.0113	.0110
2.3	.0107	.0104	.0102	.0099	.0096	.0094	.0091	.0089	.0087	.0084
2.4	.0082	.0080	.0078	.0075	.0073	.0071	.0069	.0068	.0066	.0064
2.5	.0062	.0060	.0059	.0057	.0055	.0054	.0052	.0051	.0049	.0048
2.6	.0047	.0045	.0044	.0043	.0041	.0040	.0039	.0038	.0037	.0036
2.7	.0035	.0034	.0033	.0032	.0031	.0030	.0029	.0028	.0027	.0026

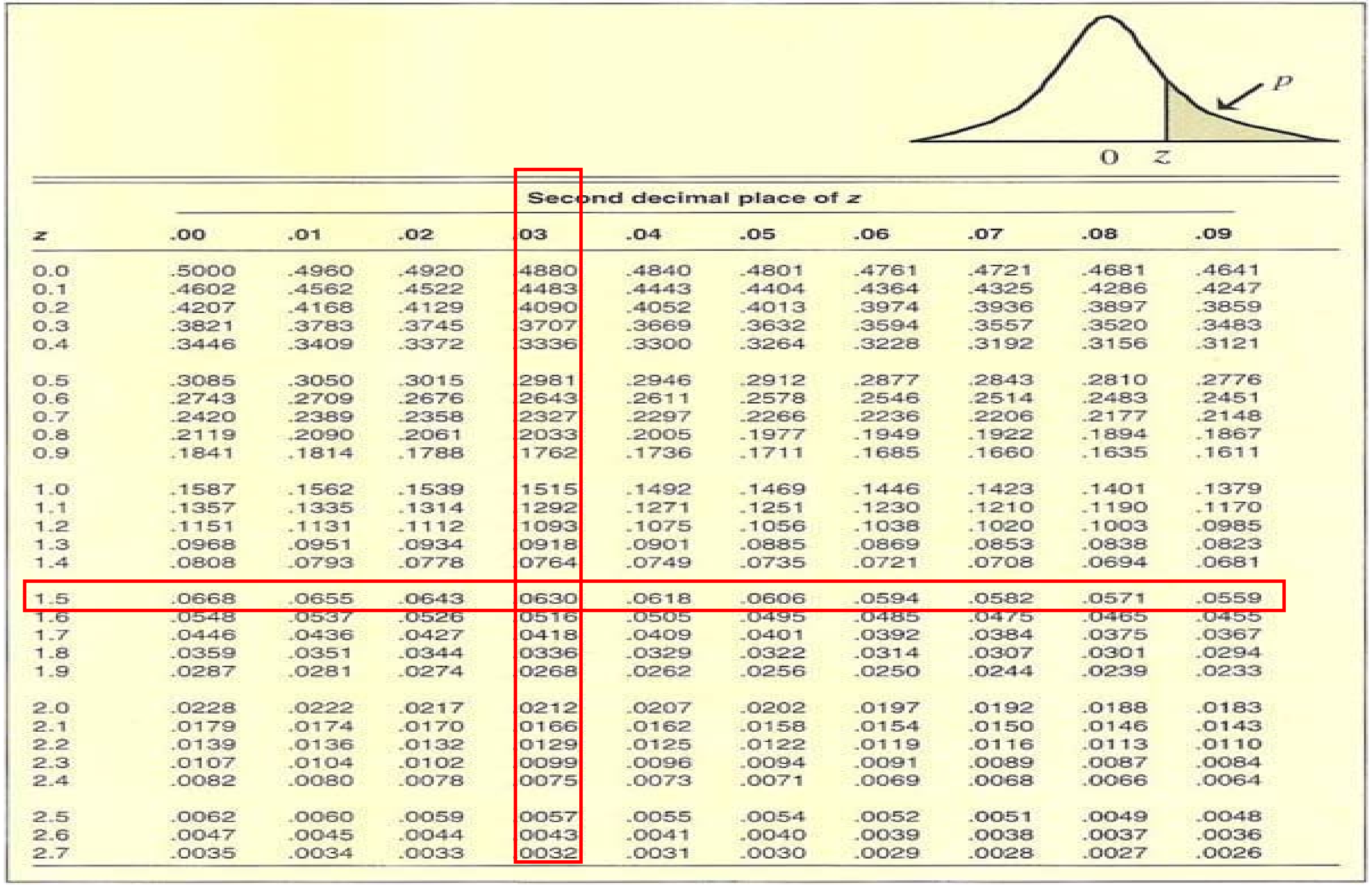
Normal Distribution: $\Pr(Z < -\mathbf{1})$



Second decimal place of z

z	.00	.01	.02	.03	.04	.05	.06	.07	.08	.09
0.0	.5000	.4960	.4920	.4880	.4840	.4801	.4761	.4721	.4681	.4641
0.1	.4602	.4562	.4522	.4483	.4443	.4404	.4364	.4325	.4286	.4247
0.2	.4207	.4168	.4129	.4090	.4052	.4013	.3974	.3936	.3897	.3859
0.3	.3821	.3783	.3745	.3707	.3669	.3632	.3594	.3557	.3520	.3483
0.4	.3446	.3409	.3372	.3336	.3300	.3264	.3228	.3192	.3156	.3121
0.5	.3085	.3050	.3015	.2981	.2946	.2912	.2877	.2843	.2810	.2776
0.6	.2743	.2709	.2676	.2643	.2611	.2578	.2546	.2514	.2483	.2451
0.7	.2420	.2389	.2358	.2327	.2297	.2266	.2236	.2206	.2177	.2148
0.8	.2119	.2090	.2061	.2033	.2005	.1977	.1949	.1922	.1894	.1867
0.9	.1841	.1814	.1788	.1762	.1736	.1711	.1685	.1660	.1635	.1611
1.0	.1587	.1562	.1539	.1515	.1492	.1469	.1446	.1423	.1401	.1379
1.1	.1357	.1335	.1314	.1292	.1271	.1251	.1230	.1210	.1190	.1170
1.2	.1151	.1131	.1112	.1093	.1075	.1056	.1038	.1020	.1003	.0985
1.3	.0968	.0951	.0934	.0918	.0901	.0885	.0869	.0853	.0838	.0823
1.4	.0808	.0793	.0778	.0764	.0749	.0735	.0721	.0708	.0694	.0681
1.5	.0668	.0655	.0643	.0630	.0618	.0606	.0594	.0582	.0571	.0559
1.6	.0548	.0537	.0526	.0516	.0505	.0495	.0485	.0475	.0465	.0455
1.7	.0446	.0436	.0427	.0418	.0409	.0401	.0392	.0384	.0375	.0367
1.8	.0359	.0351	.0344	.0336	.0329	.0322	.0314	.0307	.0301	.0294
1.9	.0287	.0281	.0274	.0268	.0262	.0256	.0250	.0244	.0239	.0233
2.0	.0228	.0222	.0217	.0212	.0207	.0202	.0197	.0192	.0188	.0183
2.1	.0179	.0174	.0170	.0166	.0162	.0158	.0154	.0150	.0146	.0143
2.2	.0139	.0136	.0132	.0129	.0125	.0122	.0119	.0116	.0113	.0110
2.3	.0107	.0104	.0102	.0099	.0096	.0094	.0091	.0089	.0087	.0084
2.4	.0082	.0080	.0078	.0075	.0073	.0071	.0069	.0068	.0066	.0064
2.5	.0062	.0060	.0059	.0057	.0055	.0054	.0052	.0051	.0049	.0048
2.6	.0047	.0045	.0044	.0043	.0041	.0040	.0039	.0038	.0037	.0036
2.7	.0035	.0034	.0033	.0032	.0031	.0030	.0029	.0028	.0027	.0026

Normal Distribution: $\Pr(Z < \mathbf{1.53})$



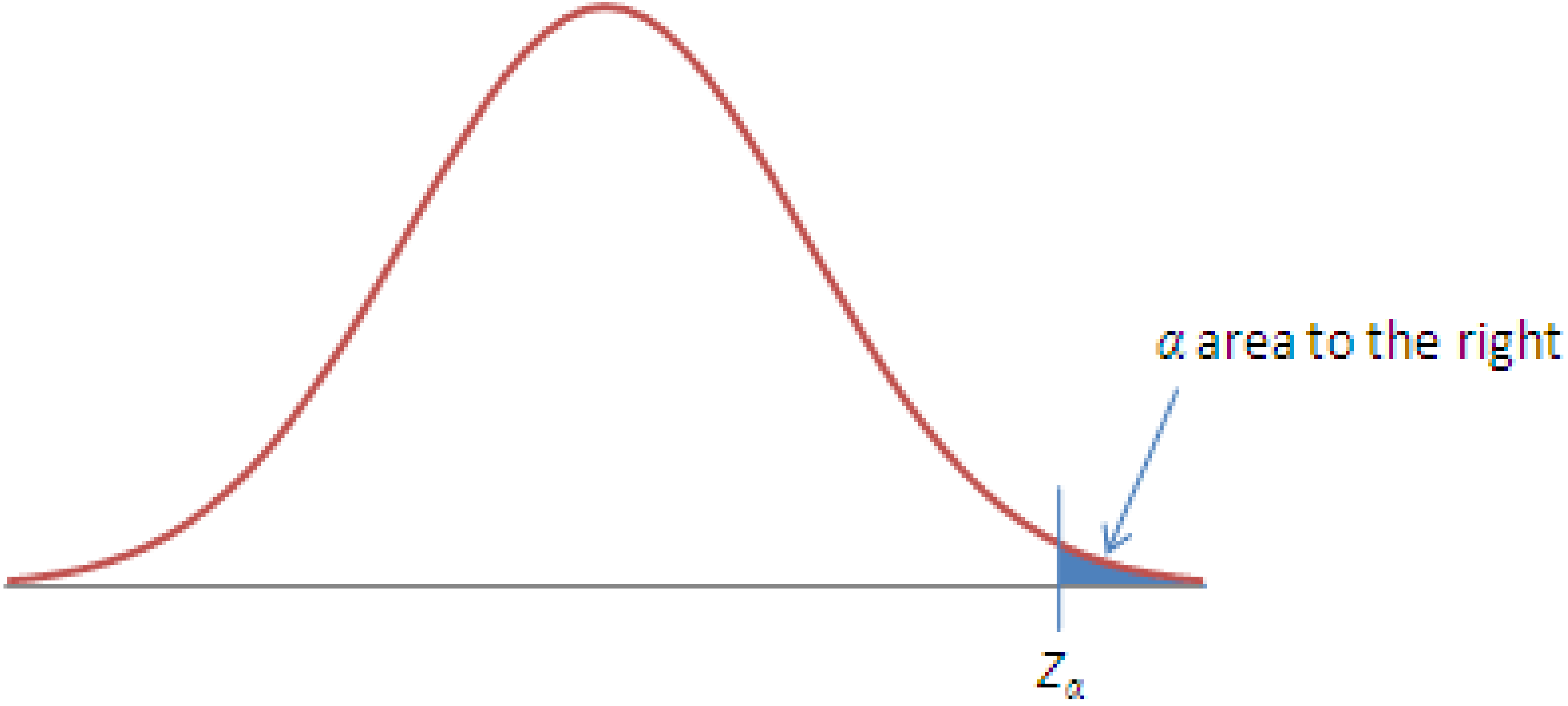
z	Second decimal place of z .00	.01	.02	.03	.04	.05	.06	.07	.08	.09
0.0	.5000	.4960	.4920	.4880	.4840	.4801	.4761	.4721	.4681	.4641
0.1	.4602	.4562	.4522	.4483	.4443	.4404	.4364	.4325	.4286	.4247
0.2	.4207	.4168	.4129	.4090	.4052	.4013	.3974	.3936	.3897	.3859
0.3	.3821	.3783	.3745	.3707	.3669	.3632	.3594	.3557	.3520	.3483
0.4	.3446	.3409	.3372	.3336	.3300	.3264	.3228	.3192	.3156	.3121
0.5	.3085	.3050	.3015	.2981	.2946	.2912	.2877	.2843	.2810	.2776
0.6	.2743	.2709	.2676	.2643	.2611	.2578	.2546	.2514	.2483	.2451
0.7	.2420	.2389	.2358	.2327	.2297	.2266	.2236	.2206	.2177	.2148
0.8	.2119	.2090	.2061	.2033	.2005	.1977	.1949	.1922	.1894	.1867
0.9	.1841	.1814	.1788	.1762	.1736	.1711	.1685	.1660	.1635	.1611
1.0	.1587	.1562	.1539	.1515	.1492	.1469	.1446	.1423	.1401	.1379
1.1	.1357	.1335	.1314	.1292	.1271	.1251	.1230	.1210	.1190	.1170
1.2	.1151	.1131	.1112	.1093	.1075	.1056	.1038	.1020	.1003	.0985
1.3	.0968	.0951	.0934	.0918	.0901	.0885	.0869	.0853	.0838	.0823
1.4	.0808	.0793	.0778	.0764	.0749	.0735	.0721	.0708	.0694	.0681
1.5	.0668	.0655	.0643	.0630	.0618	.0606	.0594	.0582	.0571	.0559
1.6	.0548	.0537	.0526	.0516	.0505	.0495	.0485	.0475	.0465	.0455
1.7	.0446	.0436	.0427	.0418	.0409	.0401	.0392	.0384	.0375	.0367
1.8	.0359	.0351	.0344	.0336	.0329	.0322	.0314	.0307	.0301	.0294
1.9	.0287	.0281	.0274	.0268	.0262	.0256	.0250	.0244	.0239	.0233
2.0	.0228	.0222	.0217	.0212	.0207	.0202	.0197	.0192	.0188	.0183
2.1	.0179	.0174	.0170	.0166	.0162	.0158	.0154	.0150	.0146	.0143
2.2	.0139	.0136	.0132	.0129	.0125	.0122	.0119	.0116	.0113	.0110
2.3	.0107	.0104	.0102	.0099	.0096	.0094	.0091	.0089	.0087	.0084
2.4	.0082	.0080	.0078	.0075	.0073	.0071	.0069	.0068	.0066	.0064
2.5	.0062	.0060	.0059	.0057	.0055	.0054	.0052	.0051	.0049	.0048
2.6	.0047	.0045	.0044	.0043	.0041	.0040	.0039	.0038	.0037	.0036
2.7	.0035	.0034	.0033	.0032	.0031	.0030	.0029	.0028	.0027	.0026

Normal Distribution: *Z*-value

- We often need to use the Z-value associated with specific "tail" areas of the standard normal distribution.
- Let z_α represent the Z-value associated with a right-hand "tail area" of α, such that

$$\Pr(Z > z_\alpha) = \alpha \text{ so } \Pr(Z < z_\alpha) = 1 - \alpha$$

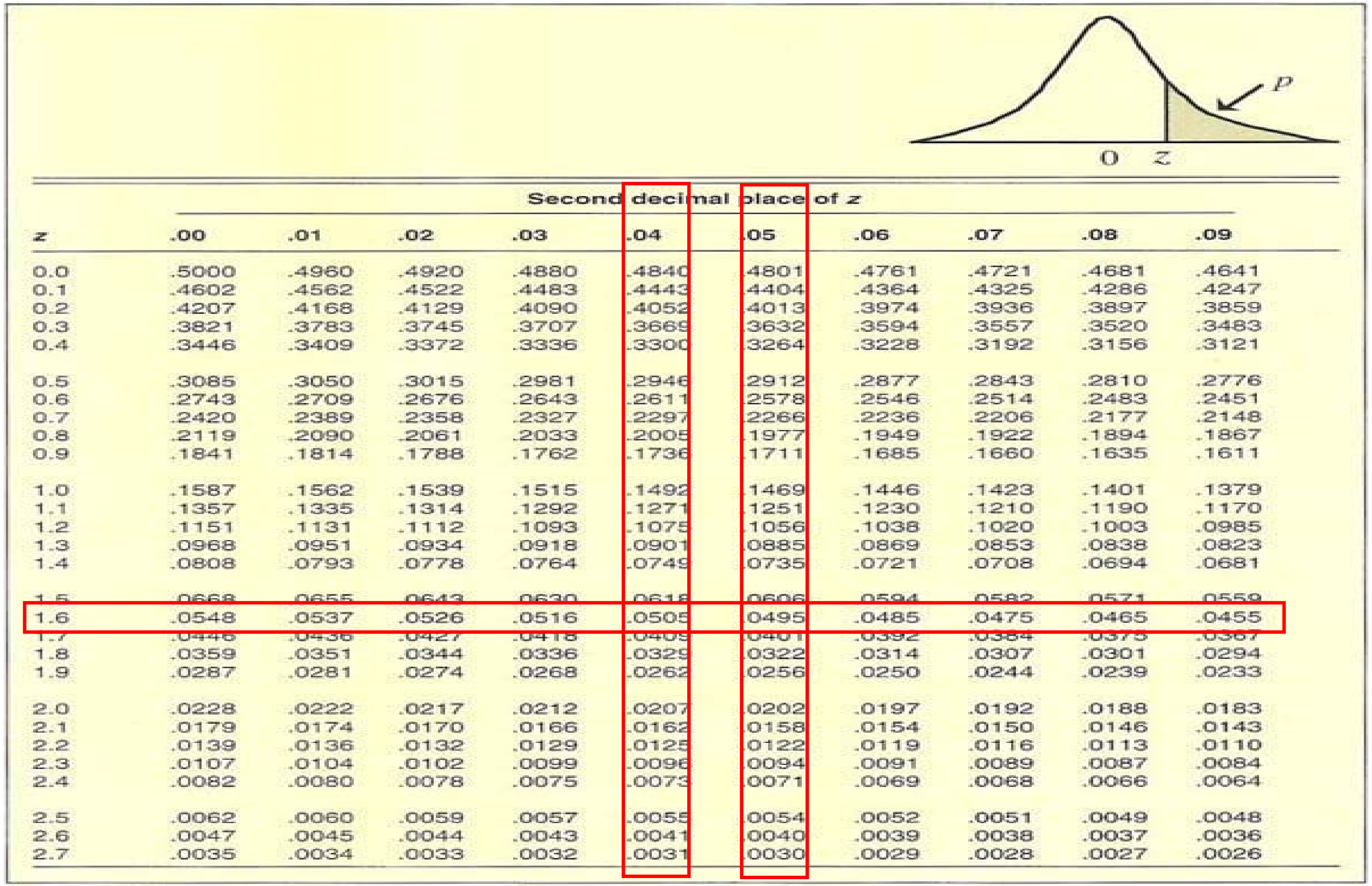
Normal Distribution: *Z*-value



Normal Distribution: Calculations

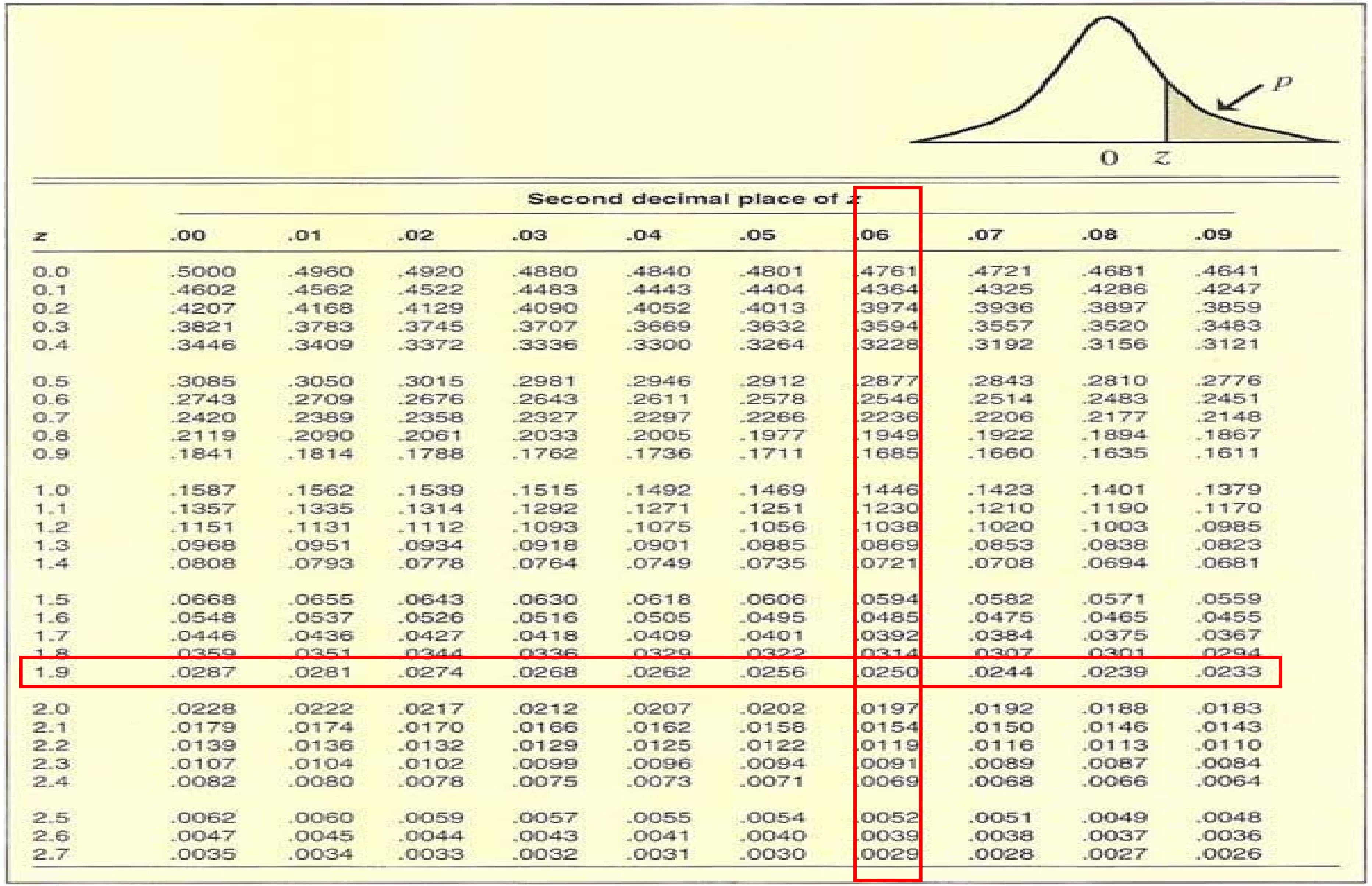
- Find z such that $\Pr(|Z| > z) = 0.10$
- $\Pr(|Z| > z) = 2\Pr(Z > z)$
- $\Pr(Z > z) = 0.05$, so $z = 1.64$ or 1.65 (1.6448)
- Therefore, $z_{0.05} = 1.6448$ or 1.64 or 1.65.
- How about $z_{0.025}$ and $z_{0.085}$?
- From the table $z_{0.025} = 1.96$ and $z_{0.085} = 1.37$

Normal Distribution: $Z_{0.05}$



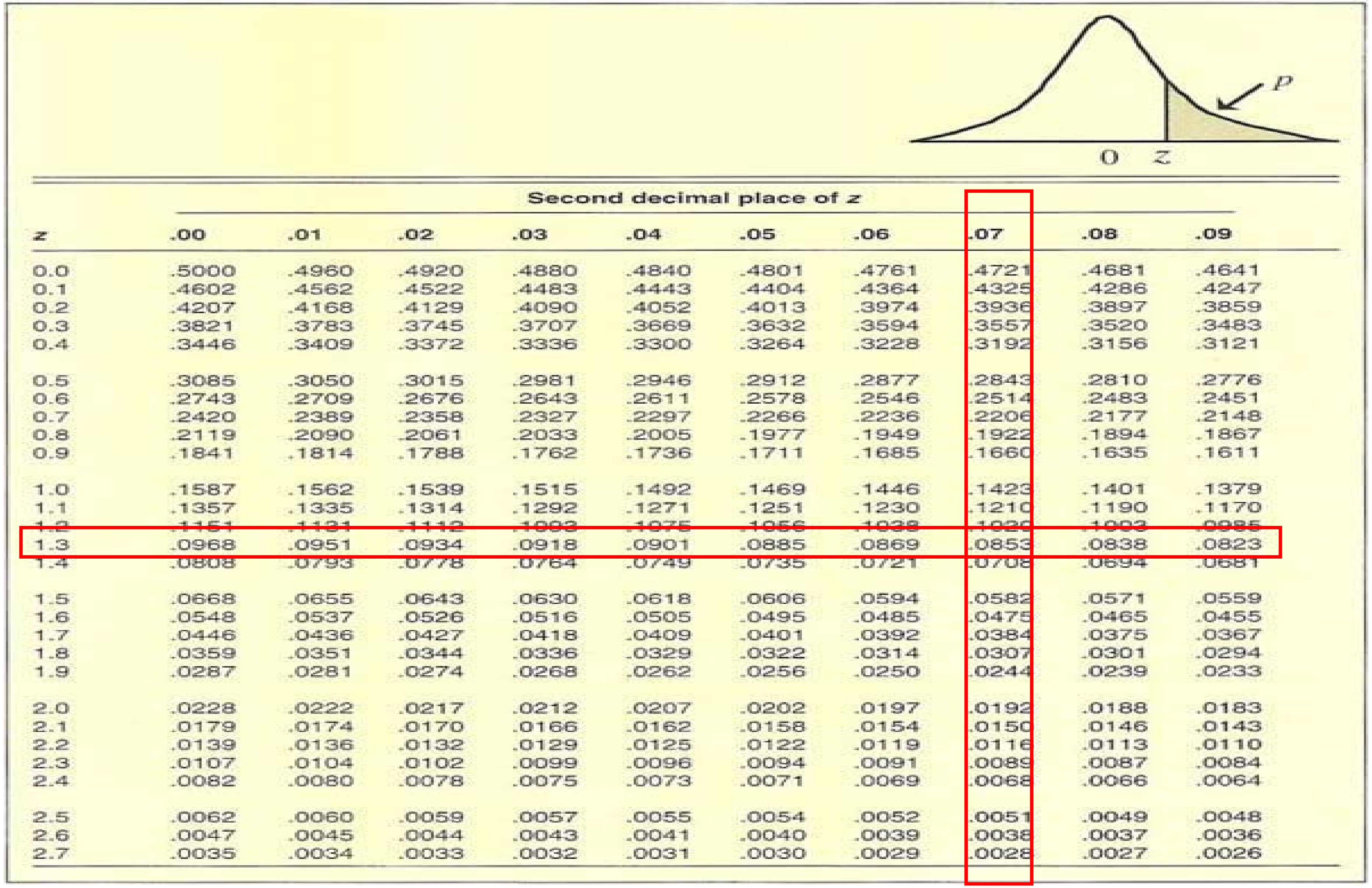
z	.00	.01	.02	.03	.04	.05	.06	.07	.08	.09
	Second decimal place of z									
0.0	.5000	.4960	.4920	.4880	.4840	.4801	.4761	.4721	.4681	.4641
0.1	.4602	.4562	.4522	.4483	.4443	.4404	.4364	.4325	.4286	.4247
0.2	.4207	.4168	.4129	.4090	.4052	.4013	.3974	.3936	.3897	.3859
0.3	.3821	.3783	.3745	.3707	.3669	.3632	.3594	.3557	.3520	.3483
0.4	.3446	.3409	.3372	.3336	.3300	.3264	.3228	.3192	.3156	.3121
0.5	.3085	.3050	.3015	.2981	.2946	.2912	.2877	.2843	.2810	.2776
0.6	.2743	.2709	.2676	.2643	.2611	.2578	.2546	.2514	.2483	.2451
0.7	.2420	.2389	.2358	.2327	.2297	.2266	.2236	.2206	.2177	.2148
0.8	.2119	.2090	.2061	.2033	.2005	.1977	.1949	.1922	.1894	.1867
0.9	.1841	.1814	.1788	.1762	.1736	.1711	.1685	.1660	.1635	.1611
1.0	.1587	.1562	.1539	.1515	.1492	.1469	.1446	.1423	.1401	.1379
1.1	.1357	.1335	.1314	.1292	.1271	.1251	.1230	.1210	.1190	.1170
1.2	.1151	.1131	.1112	.1093	.1075	.1056	.1038	.1020	.1003	.0985
1.3	.0968	.0951	.0934	.0918	.0901	.0885	.0869	.0853	.0838	.0823
1.4	.0808	.0793	.0778	.0764	.0749	.0735	.0721	.0708	.0694	.0681
1.5	.0668	.0655	.0643	.0630	.0618	.0606	.0594	.0582	.0571	.0559
1.6	.0548	.0537	.0526	.0516	.0505	.0495	.0485	.0475	.0465	.0455
1.7	.0446	.0436	.0427	.0418	.0409	.0401	.0392	.0384	.0375	.0367
1.8	.0359	.0351	.0344	.0336	.0329	.0322	.0314	.0307	.0301	.0294
1.9	.0287	.0281	.0274	.0268	.0262	.0256	.0250	.0244	.0239	.0233
2.0	.0228	.0222	.0217	.0212	.0207	.0202	.0197	.0192	.0188	.0183
2.1	.0179	.0174	.0170	.0166	.0162	.0158	.0154	.0150	.0146	.0143
2.2	.0139	.0136	.0132	.0129	.0125	.0122	.0119	.0116	.0113	.0110
2.3	.0107	.0104	.0102	.0099	.0096	.0094	.0091	.0089	.0087	.0084
2.4	.0082	.0080	.0078	.0075	.0073	.0071	.0069	.0068	.0066	.0064
2.5	.0062	.0060	.0059	.0057	.0055	.0054	.0052	.0051	.0049	.0048
2.6	.0047	.0045	.0044	.0043	.0041	.0040	.0039	.0038	.0037	.0036
2.7	.0035	.0034	.0033	.0032	.0031	.0030	.0029	.0028	.0027	.0026

Normal Distribution: $Z_{0.025}$



z	Second decimal place of z .00	.01	.02	.03	.04	.05	.06	.07	.08	.09
0.0	.5000	.4960	.4920	.4880	.4840	.4801	.4761	.4721	.4681	.4641
0.1	.4602	.4562	.4522	.4483	.4443	.4404	.4364	.4325	.4286	.4247
0.2	.4207	.4168	.4129	.4090	.4052	.4013	.3974	.3936	.3897	.3859
0.3	.3821	.3783	.3745	.3707	.3669	.3632	.3594	.3557	.3520	.3483
0.4	.3446	.3409	.3372	.3336	.3300	.3264	.3228	.3192	.3156	.3121
0.5	.3085	.3050	.3015	.2981	.2946	.2912	.2877	.2843	.2810	.2776
0.6	.2743	.2709	.2676	.2643	.2611	.2578	.2546	.2514	.2483	.2451
0.7	.2420	.2389	.2358	.2327	.2297	.2266	.2236	.2206	.2177	.2148
0.8	.2119	.2090	.2061	.2033	.2005	.1977	.1949	.1922	.1894	.1867
0.9	.1841	.1814	.1788	.1762	.1736	.1711	.1685	.1660	.1635	.1611
1.0	.1587	.1562	.1539	.1515	.1492	.1469	.1446	.1423	.1401	.1379
1.1	.1357	.1335	.1314	.1292	.1271	.1251	.1230	.1210	.1190	.1170
1.2	.1151	.1131	.1112	.1093	.1075	.1056	.1038	.1020	.1003	.0985
1.3	.0968	.0951	.0934	.0918	.0901	.0885	.0869	.0853	.0838	.0823
1.4	.0808	.0793	.0778	.0764	.0749	.0735	.0721	.0708	.0694	.0681
1.5	.0668	.0655	.0643	.0630	.0618	.0606	.0594	.0582	.0571	.0559
1.6	.0548	.0537	.0526	.0516	.0505	.0495	.0485	.0475	.0465	.0455
1.7	.0446	.0436	.0427	.0418	.0409	.0401	.0392	.0384	.0375	.0367
1.8	.0359	.0351	.0344	.0336	.0329	.0322	.0314	.0307	.0301	.0294
1.9	.0287	.0281	.0274	.0268	.0262	.0256	.0250	.0244	.0239	.0233
2.0	.0228	.0222	.0217	.0212	.0207	.0202	.0197	.0192	.0188	.0183
2.1	.0179	.0174	.0170	.0166	.0162	.0158	.0154	.0150	.0146	.0143
2.2	.0139	.0136	.0132	.0129	.0125	.0122	.0119	.0116	.0113	.0110
2.3	.0107	.0104	.0102	.0099	.0096	.0094	.0091	.0089	.0087	.0084
2.4	.0082	.0080	.0078	.0075	.0073	.0071	.0069	.0068	.0066	.0064
2.5	.0062	.0060	.0059	.0057	.0055	.0054	.0052	.0051	.0049	.0048
2.6	.0047	.0045	.0044	.0043	.0041	.0040	.0039	.0038	.0037	.0036
2.7	.0035	.0034	.0033	.0032	.0031	.0030	.0029	.0028	.0027	.0026

Normal Distribution: $Z_{0.085}$

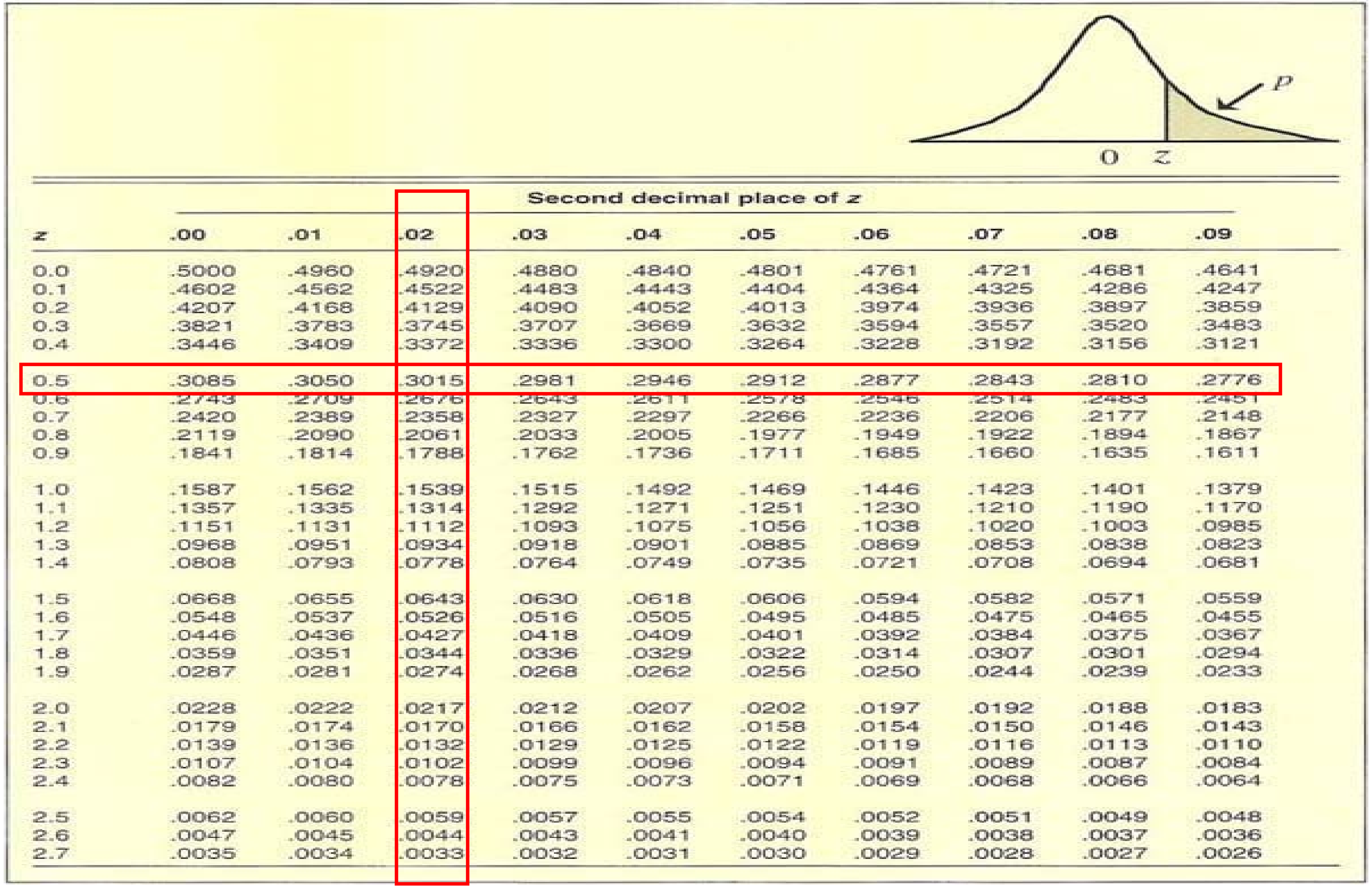


z	Second decimal place of z									
z	.00	.01	.02	.03	.04	.05	.06	.07	.08	.09
0.0	.5000	.4960	.4920	.4880	.4840	.4801	.4761	.4721	.4681	.4641
0.1	.4602	.4562	.4522	.4483	.4443	.4404	.4364	.4325	.4286	.4247
0.2	.4207	.4168	.4129	.4090	.4052	.4013	.3974	.3936	.3897	.3859
0.3	.3821	.3783	.3745	.3707	.3669	.3632	.3594	.3557	.3520	.3483
0.4	.3446	.3409	.3372	.3336	.3300	.3264	.3228	.3192	.3156	.3121
0.5	.3085	.3050	.3015	.2981	.2946	.2912	.2877	.2843	.2810	.2776
0.6	.2743	.2709	.2676	.2643	.2611	.2578	.2546	.2514	.2483	.2451
0.7	.2420	.2389	.2358	.2327	.2297	.2266	.2236	.2206	.2177	.2148
0.8	.2119	.2090	.2061	.2033	.2005	.1977	.1949	.1922	.1894	.1867
0.9	.1841	.1814	.1788	.1762	.1736	.1711	.1685	.1660	.1635	.1611
1.0	.1587	.1562	.1539	.1515	.1492	.1469	.1446	.1423	.1401	.1379
1.1	.1357	.1335	.1314	.1292	.1271	.1251	.1230	.1210	.1190	.1170
1.2	.1151	.1131	.1112	.1093	.1075	.1056	.1038	.1020	.1003	.0985
1.3	.0968	.0951	.0934	.0918	.0901	.0885	.0869	.0853	.0838	.0823
1.4	.0808	.0793	.0778	.0764	.0749	.0735	.0721	.0708	.0694	.0681
1.5	.0668	.0655	.0643	.0630	.0618	.0606	.0594	.0582	.0571	.0559
1.6	.0548	.0537	.0526	.0516	.0505	.0495	.0485	.0475	.0465	.0455
1.7	.0446	.0436	.0427	.0418	.0409	.0401	.0392	.0384	.0375	.0367
1.8	.0359	.0351	.0344	.0336	.0329	.0322	.0314	.0307	.0301	.0294
1.9	.0287	.0281	.0274	.0268	.0262	.0256	.0250	.0244	.0239	.0233
2.0	.0228	.0222	.0217	.0212	.0207	.0202	.0197	.0192	.0188	.0183
2.1	.0179	.0174	.0170	.0166	.0162	.0158	.0154	.0150	.0146	.0143
2.2	.0139	.0136	.0132	.0129	.0125	.0122	.0119	.0116	.0113	.0110
2.3	.0107	.0104	.0102	.0099	.0096	.0094	.0091	.0089	.0087	.0084
2.4	.0082	.0080	.0078	.0075	.0073	.0071	.0069	.0068	.0066	.0064
2.5	.0062	.0060	.0059	.0057	.0055	.0054	.0052	.0051	.0049	.0048
2.6	.0047	.0045	.0044	.0043	.0041	.0040	.0039	.0038	.0037	.0036
2.7	.0035	.0034	.0033	.0032	.0031	.0030	.0029	.0028	.0027	.0026

Normal Distribution: Calculations

- How about $z_{0.70}$?
- We know that $\Pr(Z > z_{0.70}) = 0.70$ so $z_{0.70} < 0.$
- So $\Pr(Z > -z_{0.70}) = 0.30$
- We have $z_{0.70} = -z_{0.30} = -0.52.$

Normal Distribution: $Z_{0.30} = Z_{-0.70}$



z	.00	.01	.02	.03	.04	.05	.06	.07	.08	.09
0.0	.5000	.4960	.4920	.4880	.4840	.4801	.4761	.4721	.4681	.4641
0.1	.4602	.4562	.4522	.4483	.4443	.4404	.4364	.4325	.4286	.4247
0.2	.4207	.4168	.4129	.4090	.4052	.4013	.3974	.3936	.3897	.3859
0.3	.3821	.3783	.3745	.3707	.3669	.3632	.3594	.3557	.3520	.3483
0.4	.3446	.3409	.3372	.3336	.3300	.3264	.3228	.3192	.3156	.3121
0.5	.3085	.3050	.3015	.2981	.2946	.2912	.2877	.2843	.2810	.2776
0.6	.2743	.2709	.2676	.2643	.2611	.2578	.2546	.2514	.2483	.2451
0.7	.2420	.2389	.2358	.2327	.2297	.2266	.2236	.2206	.2177	.2148
0.8	.2119	.2090	.2061	.2033	.2005	.1977	.1949	.1922	.1894	.1867
0.9	.1841	.1814	.1788	.1762	.1736	.1711	.1685	.1660	.1635	.1611
1.0	.1587	.1562	.1539	.1515	.1492	.1469	.1446	.1423	.1401	.1379
1.1	.1357	.1335	.1314	.1292	.1271	.1251	.1230	.1210	.1190	.1170
1.2	.1151	.1131	.1112	.1093	.1075	.1056	.1038	.1020	.1003	.0985
1.3	.0968	.0951	.0934	.0918	.0901	.0885	.0869	.0853	.0838	.0823
1.4	.0808	.0793	.0778	.0764	.0749	.0735	.0721	.0708	.0694	.0681
1.5	.0668	.0655	.0643	.0630	.0618	.0606	.0594	.0582	.0571	.0559
1.6	.0548	.0537	.0526	.0516	.0505	.0495	.0485	.0475	.0465	.0455
1.7	.0446	.0436	.0427	.0418	.0409	.0401	.0392	.0384	.0375	.0367
1.8	.0359	.0351	.0344	.0336	.0329	.0322	.0314	.0307	.0301	.0294
1.9	.0287	.0281	.0274	.0268	.0262	.0256	.0250	.0244	.0239	.0233
2.0	.0228	.0222	.0217	.0212	.0207	.0202	.0197	.0192	.0188	.0183
2.1	.0179	.0174	.0170	.0166	.0162	.0158	.0154	.0150	.0146	.0143
2.2	.0139	.0136	.0132	.0129	.0125	.0122	.0119	.0116	.0113	.0110
2.3	.0107	.0104	.0102	.0099	.0096	.0094	.0091	.0089	.0087	.0084
2.4	.0082	.0080	.0078	.0075	.0073	.0071	.0069	.0068	.0066	.0064
2.5	.0062	.0060	.0059	.0057	.0055	.0054	.0052	.0051	.0049	.0048
2.6	.0047	.0045	.0044	.0043	.0041	.0040	.0039	.0038	.0037	.0036
2.7	.0035	.0034	.0033	.0032	.0031	.0030	.0029	.0028	.0027	.0026

Second decimal place of z

Calculating Probabilities of Normal Distribution

- Characteristics of standard normal distribution. If $Z\sim N(0,1)$, then

$$\Pr(Z > z) = \Pr(Z < -z) = 1 - \Pr(Z < z)$$

$$\Pr(a < Z < b) = \Pr(Z < b) - \Pr(Z < a)$$

- **Exercise** – find the following probability:

$$\Pr(Z > 1.02)$$

$$\Pr(Z < -0.80)$$

$$\Pr(Z < 1.35)$$

$$\Pr(-0.80 < Z < 1.35)$$

Calculating Probabilities of Normal Distribution

- Sometimes you want to find a value $z_\alpha (0 < \alpha < 1)$:

$$\Pr(Z > z_\alpha) = \Pr(Z < -z_\alpha) = \alpha$$

$$\Pr(Z < z_\alpha) = 1 - \alpha$$

$$\Pr\left(-z_{\frac{\alpha}{2}} < Z < z_{\frac{\alpha}{2}}\right) = 1 - \alpha$$

- **Exercise:** find a value z such that

$$\Pr(|Z| > z) = 0.08$$

Normal Distribution: Transformation

- If $Y \sim N(\mu, \sigma^2)$, then $E(Y) = \mu$ and $Var(Y) = \sigma^2$
- Define $Z = \frac{Y-\mu}{\sigma}$ - First re-center Y with population mean μ then rescale it with σ. Then $Z \sim N(0,1)$.
- We have

$$\Pr(Y < y) = \Pr\left(\frac{Y-\mu}{\sigma} < \frac{y-\mu}{\sigma}\right) = \Pr(Z < \frac{y-\mu}{\sigma})$$

Steps for Calculations with Normal Distribution

1. Determine if Y has a normal distribution.
2. Find the population mean and variance of Y.
3. Represent probability in terms of Y.
4. Transfer Y to the standard normal Z.
5. Find the probability with a normal table.

Calculating Probabilities of Normal Distribution

- Suppose that $Y \sim N(10,20)$, find the following probability:

$$\Pr(Y > 15) = \Pr\left(\frac{Y - 10}{\sqrt{20}} > \frac{15 - 10}{\sqrt{20}}\right) = \Pr(\mathrm{Z} > 1.12) = 0.1314$$

- **Exercises:** for $Y \sim N(10,10)$, find:

$$\Pr(Y < 5)$$
$$\Pr(5 < Y < 10)$$

Example – Production at Titanium Dioxide Facility

- **Example** – A major titanium dioxide facility has a designed capacity of 600 tons. Historically, the daily production approximately follows a normal distribution with mean 500 tons and a standard deviation of 50 tons.

Example – Production at Titanium Dioxide Facility

- The company can sell anything this facility can make. So management really would like to know the probability that it can make more than 600 tons (exceeding the capacity) of product.
- Let Y be the total production, then $Y \sim N(500,50)$.

$$\Pr(Y > 600) = \Pr\left(\frac{Y - 500}{50} > \frac{600 - 500}{50}\right)$$

$$= \Pr(Z > 2) = 0.0228.$$

- So the facility should exceed the capacity only 2% of time.

Example – Setting Mean for Dairy Packing

Example – 8 oz of milk should weigh 245 grams. Federal inspectors require that the mean amount of milk packaged must be significantly greater than 245 grams in order to minimize any underage. Plant manager would like the mean amount to be no more than necessary to meet standards and argues that a reasonable mean weight should produce less than 1% of cartons underweight. Historically, the weight approximately follows a normal distribution with a standard deviation of 1.65 grams.

Example – Setting Mean for Dairy Packing

- Let Y be the milk weight and $Y{\sim}N(\mu, 1.65^2)$

$$\Pr(Y < 245) = \Pr\left(\frac{Y-\mu}{1.65} < \frac{245-\mu}{1.65}\right)$$

$$= \Pr\left(Z < \frac{245-\mu}{1.65}\right) < 0.01$$

- So $\frac{245-\mu}{1.65} < z_{0.99}$ and

$\mu > 245 -1.65{*}z_{0.99} = 245 - 1.65 * (-2.33) = 248.83$

- So we need an average of 248.83 grams.

Rational Behind Boxplot

- Recall the boxplot from Chapter 1, we call an observation y:
  - An outlier, if $y >$ Upper Inner fence or $y <$ Lower Inner Fence
  - An extreme outlier, if $y >$ Upper Outer fence or $y <$ Lower Outer Fence
- Interquartile Range (IQR)$= Q_3 - Q_1$ and Step $= 1.5 * IQR$
- Upper Inner Fence $= Q_3 + Step$
- Lower Inner Fence $= Q_1 - Step$
- Upper Outer Fence $= Q_3 + 2 * Step$;
- Lower Outer Fence $= Q_1 - 2 * Step$;

Exercise - Rational Behind Boxplot

Suppose that the data is from the standard normal. Find

- $Q_3;\ Q_1$
- $IQR\ = Q_3 - Q_1$
- $Step\ =\ 2\ *\ IQR$
- $UIF = Q_3 + Step; LIF = Q_3 - Step$
- $\Pr(Z > UIF)$ and $\Pr(Z < LIF)$
- $UOF = Q_3 + 2 * Step; LOF = Q_1 - 2 * Step$
- $\Pr(Z > UOF)$ and $\Pr(Z < LOF)$

Solution - Rational Behind Boxplot

Suppose that the data is from the standard normal. Find

- $Q_3 = z_{0.25} = 0.6749;\ Q_1 = -z_{0.25} = -0.6749$
- $IQR\ = Q_3 - Q_1 = 2 * 0.6749 = 1.3490$
- $Step\ =\ 2\ *\ IQR\ =\ 1.5 *\ 1.3490\ = 2.0234$
- $UIF = Q_3 + Step = 2.6980; LIF = -UIF =- -2.6980$
- $\Pr(Z > UIF) = \Pr(Z < LIF) = 0.0035$
- $UOF = Q_3 + 2 * Step = 4.7214; LOF = -UOF =- -4.7214$
- $\Pr(Z > UOF) = \Pr(Z < LOF) \approx 0$

Sampling Distribution

- **Statistical Inference -** making inferences on population parameters using sample statistic.
- **Definition 2.15 -** The **sampling distribution** of a statistic is the probability distribution of that statistic.
- Statistic is a random variable.

Sample Mean

- Sample mean – let $y_1, \cdots, y_n$ denote a sample of interest. The sample mean, denoted by $\bar{y}$, is given by

$$\bar{y} = \frac{1}{n}(y_1 + \cdots + y_n) = \frac{1}{n}\sum_{i=1}^{n} y_i.$$

- Sample mean is a random variable!

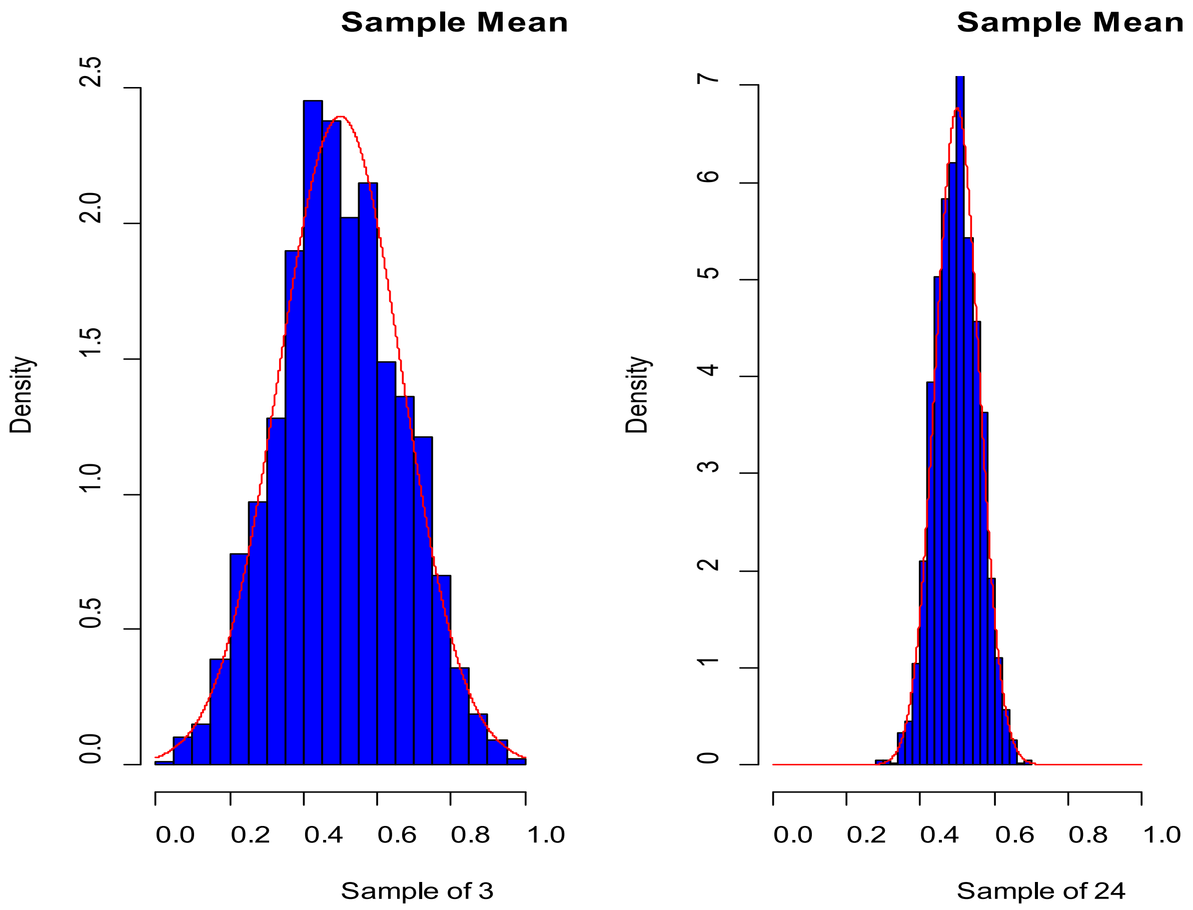
Sampling Distribution

- **Theorem 2.5.1 Sampling distribution of the mean -** The sampling distribution of the mean from a random sample of size n from a population with mean μ and variance σ^2 will have mean μ and variance σ^2 / n.
- This is true for a random sample from any distribution.

Sampling Distribution: An example

- For a uniform distribution from 0 to 1
  - Population mean and variance are 0.5 and 0.08333
- Consider the sample size of 3
  - Mean and variance of sample mean is 0.5 and 0.08333/3=0.0278
- The empirical estimate from 1000 samples
  - Mean and variance of sample mean is 0.5 and 0.08333/3=0.0278
- How about the sample size of 24?
  - Mean and variance of sample mean is 0.5 and 0.08333/24=0.00347

Sampling Distribution: An example



Usefulness of Sampling Distribution

- Mean of sampling distribution of the sample mean is the population mean - implies that “on the average” the sample mean is the same as the population mean. We therefore say that the sample mean is an **unbiased estimate** of the population mean.
- Variance of the distribution of the sample mean is σ^2/n .This implies the variability of sample mean decreases with the increasing sample size.

Sampling Distribution of Sample Mean

- But what is the exact distribution of sample mean, $\bar{y}$ ($\bar{Y}$)?
- Its distribution depends on
  - The distribution of sample.
  - The samples themselves.
- In this book, we focus on the ***random sample*** – A random sample is one in which all the observations are statistically independent and follow exactly the same distribution.

Sampling Distribution of Sample Mean

- If the sample is a random sample from the normal distribution $N(\mu, \sigma^2)$, then sample mean $\sim N(\mu, \frac{\sigma^2}{n})$
- If the sample is a random sample from another distribution, then the sampling distribution of sample mean is generally unknown. However, the normal distribution can be used as an approximation.

Central Limit Theorem

- **Central Limit Theorem** - If random samples of size n are taken from any distribution with mean μ and variance σ^2, sample mean will have a distribution approximately normal with mean μ and variance σ^2/n.
- How large n should be for the Central Limit Theorem:

$$n > 30 \text{ in general.}$$

- Much small n is fine if data is approximately normal.

Central Limit Theorem - Example

- **Example -** An aptitude test for high school students is designed so that scores on the test have μ = 90 and σ = 20. In a section of 100 students the mean score is 86. What is the probability of getting a mean of 86 or lower on test?
- We have $\bar{Y} \sim N(90, \frac{20^2}{100})$ and

$$\Pr(\bar{Y} < 86) = \Pr\left(\frac{\bar{Y} - 90}{2} < \frac{86 - 90}{2}\right)$$

$$= Pr(Z < -2) = 0.0228$$

CLT – Thickness of Silicon Wafers

- **Example –** Hurwitz and Spagon (1993) analyzed the performance of a planarization device that polishes silicon wafers to a high degree of smoothness. Historically, the thickness of the wafer has a mean of 3200 angstroms with a standard deviation 80 angstroms. The following thicknesses are from 23 wafers.

CLT – Thickness of Silicon Wafers

3240	3200	3220	3210	3250	3220
3190	3190	3150	3160	3270	3180
3200	3270	3180	3300	3250	3330
3300	3280	3270	3270	3200	

- Sample mean is $\bar{y} = 3232$, which is larger than 3200.
- Production management would like to know whether there is evidence to suggest that this lot is thicker than usual.

Exercise - Thickness of Silicon Wafers

- How to formulate “whether there is evidence to suggest that this lot is thicker than usual”?
- If the probability of the mean thickness of 23 wafers greater than the observed mean (3232) is small, we have an evidence of “unusual”.
- We need to calculate:

$$\Pr(\bar{Y} > 3232)$$

Solution - Thickness of Silicon Wafers

- We have

$$\Pr(\bar{Y} > 3232) = \Pr\left(\frac{\bar{Y} - 3200}{\frac{80}{\sqrt{23}}} > \frac{3232 - 3200}{\frac{80}{\sqrt{23}}}\right)$$

$$= \Pr(Z > 1.92) = 0.0274$$

- Yes. We have some evidence to suggest that this lot is thicker than usual.

Central Limit Theorem

- When do we use $N(\mu, \sigma^2)$ and when do we use $N(\mu, \sigma^2/n)$?
- One is for an individual observation and one is for the sample mean.
- Example (Thickness of Silicon Wafer) – What is the probability of the mean thickness of 100 wafers great than 3232? $N(\mu, \sigma^2/n)$
- Example (Thickness of Silicon Wafer) – What is the probability of thickness of a random selected wafer great than 3232? $N(\mu, \sigma^2)$

Normal Approximation to Binomial

- If $Y \sim Binomial(n, p)$, then $Y = \sum_{i=1}^{n} Y_i$ and $Y_i (i = 1, \cdots, n)$ is a random sample from Bernoulli distribution. So $\frac{Y}{n}$ is considered as a sample mean.
- By Central Limit Theorem (CLT):

$$\frac{\frac{Y}{n} - p}{\sqrt{p(1-p)/n}} = \frac{Y - np}{\sqrt{np(1-p)}}$$

can be approximated by $N(0,1)$ for large n.

Normal Approximation to Binomial

- By Central Limit Theorem (CLT): $\frac{\frac{Y}{n}-p}{\sqrt{p(1-p)/n}} = \frac{Y-np}{\sqrt{np(1-p)}}$ can be approximated by $N(0,1)$ for large n.
- If we want to use normal distribution to approximate the sample proportion: use $N(p, \frac{p(1-p)}{n})$.
- If we want to use normal to approximate the number of success (which has a binomial distribution): use $N(np, np(1-p))$.

Normal Approximation to Binomial

- In other words, $\frac{\frac{Y}{n}-p}{\sqrt{p(1-p)/n}}$ or $\frac{Y-np}{\sqrt{np(1-p)}}$ can be approximated by $N(0,1)$ for sufficiently large n.
- We consider n to be sufficiently large if

$$np \geq 5 \text{ and } n(1-p) \geq 5.$$

- The approximation works even better if

$$np \geq 10 \text{ and } n(1-p) \geq 10.$$

Example 2.15 - Election

- **Example 2.15:** Suppose a random sample of 100 voters show 61 with a preference for candidate A. If the election were in fact a toss-up (that is, $p = 0.5$) what is the probability of obtaining that (or a more extreme value)?
- Let be the number of voters with a preference for candidate A, then $Y \sim Binomial(100, 0.5)$ and we want to calculate $\Pr(Y \geq 61)$
- Can we use normal approximation?
- Check: $100 * 0.5 = 50$. Yes. We can.

Exercise – Normal Approximation to Binomial

- **Exercise:** Based on data from 2007 National Health Interview Survey, it is estimated that “10% of adults experienced feelings of sadness for all, most, or some of the time” during the 30 days prior to the interview. You interview a random sample of 68 people who have recently filed for unemployment benefits in your county, and ask this same question in your survey. If the proportion of your population with these feelings is the same as the 10% nationally, what is the probability that your sample will have 12 or more people with these feelings?

Exercise – Normal Approximation to Binomial

- **Exercise:** An insurance company wishes to keep the error rates in medical claims at or below 10%. If there is evidence of an error rate greater than this, they will need to introduce new quality procedures. They randomly select 60 independent claims and audit them for errors and use the following rule: *Decide error rate is acceptable if there are eight or fewer errors in the sample of 60*.
- **Question:** If the probability of error is truly 10%, what is the chance they will decide their error rate is acceptable?

Sample Variance

- To use the central limit theorem, we need to know the population variance. If we want to use data to make inference about the population, we generally do not know the population variance.
- In this situation, we can use sample variance in our calculation.

Descriptive Statistics – Sample Variance

- **Definition 1.14 –** The **sample variance**, denoted by s^2 is defined by

$$s^2 = \frac{1}{n-1}\left[(y_1 - \bar{y})^2 + (y_2 - \bar{y})^2 + \cdots + (y_n - \bar{y})^2\right]$$

- It looks like an "average" of the squared deviations from the sample mean.
- Note that we use $n - 1$ instead of n.
- Sample variance is a measure of **dispersion** which is the extent to which a distribution is stretched or squeezed.

Descriptive Statistics – Sample Variance

- Another formula for sample variance, s^2, is

$$s^2 = \frac{1}{n-1}\sum_{i=1}^{n}(y_i - \bar{y})^2$$

$$= \frac{1}{n-1}\left(\sum_{i=1}^{n} y_i^2 - n\bar{y}^2\right).$$

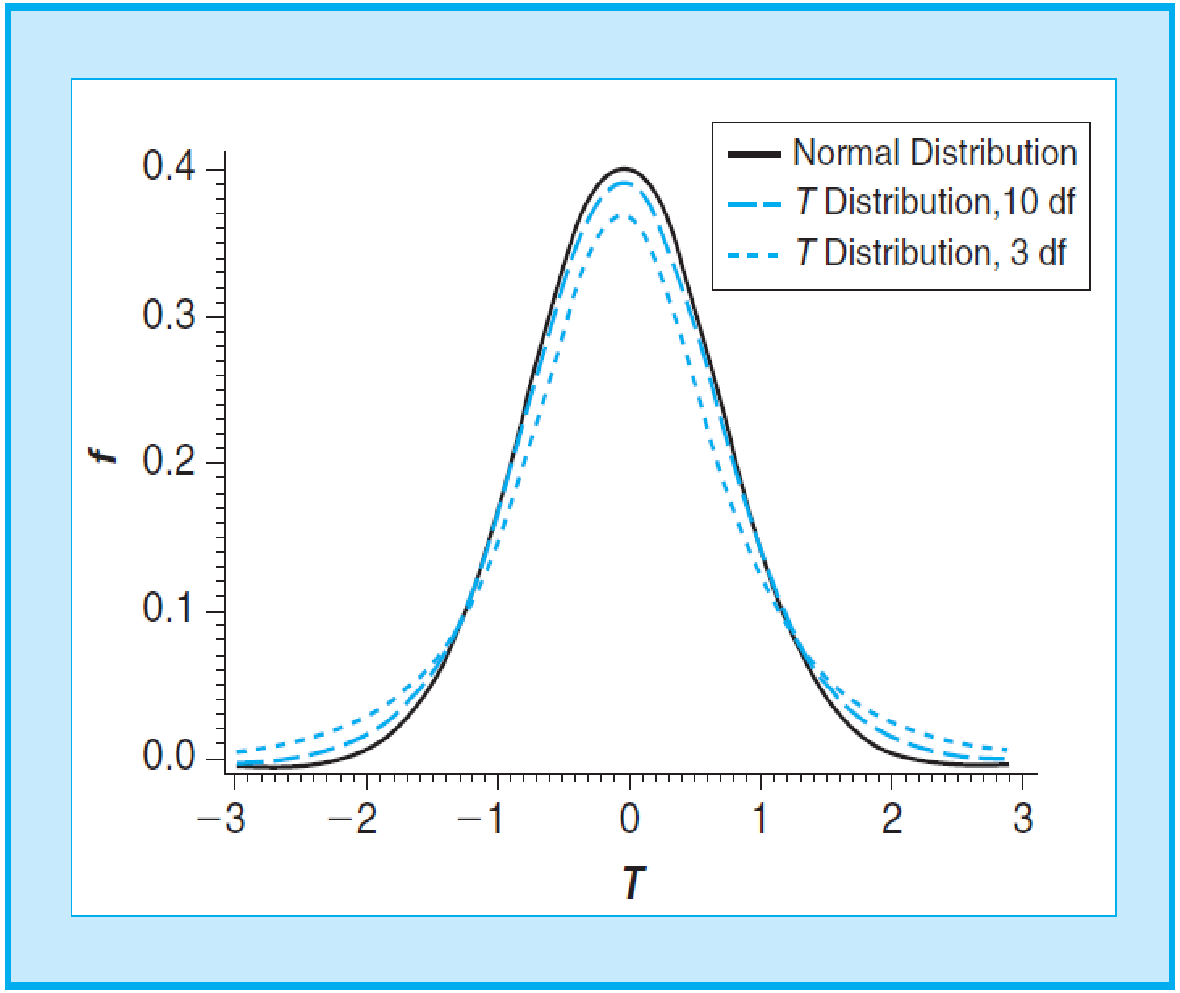
- **Definition 1.15 –** The **standard deviation** of a set of observed values is defined to be the positive square root of the variance. In other words, the sample standard deviation is: $s = \sqrt{s^2}$.

Random Behavior of Means with Unknown Variance

- For a random sample size of n from $N(\mu, \sigma^2)$, we can
- Use $Z = \frac{\bar{Y}-\mu}{\sigma/\sqrt{n}} \sim N(0,1)$ if σ is known.
- Use $t = \frac{\bar{Y}-\mu}{s/\sqrt{n}} \sim t_{n-1}$ - t-distribution with $n-1$ degrees of freedom if σ is unknown.
- In general, degrees of freedom = number of observations – number of parameters estimated

t-distribution

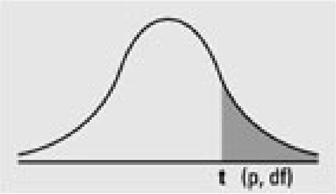
- t-distribution has a similar shape with standard normal but with a “fatter” tail.
- When $n \to \infty$, t_{n-1} becomes $N(0, 1)$ since the sample variance converges to population variance.

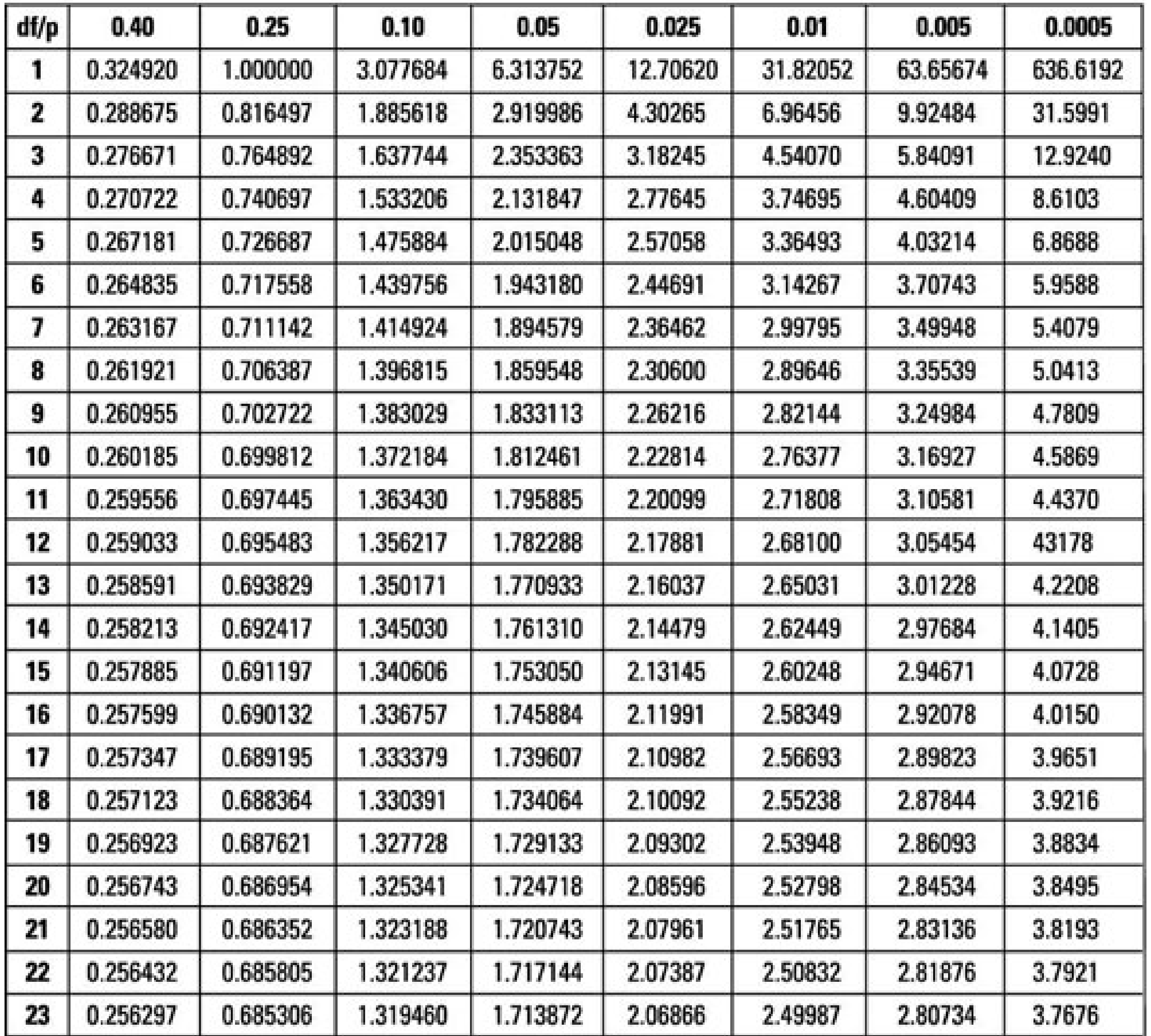


t-distribution

- Similarly with standard normal, we can use t-table to calculate the probability.
- May need to use symmetric property of t-distribution.
- We can find $t_{n,\alpha}$ such that $\Pr\left(T_n > t_{n,\alpha}\right) = \alpha$.
- **Exercise:** Find $\Pr(T_{10} > 0.5)$.
- **Example:** Find $t_{10,0.05}$.
- **Example:** Find $t_{26,0.05}$.

Numbers in each row of the table are values on a *t*-distribution with (*df*) degrees of freedom for selected right-tail (greater-than) probabilities (*p*).





df/p	0.40	0.25	0.10	0.05	0.025	0.01	0.005	0.0005
1	0.324920	1.000000	3.077684	6.313752	12.70620	31.82052	63.65674	636.6192
2	0.288675	0.816497	1.885618	2.919986	4.30265	6.96456	9.92484	31.5991
3	0.276671	0.764892	1.637744	2.353363	3.18245	4.54070	5.84091	12.9240
4	0.270722	0.740697	1.533206	2.131847	2.77645	3.74695	4.60409	8.6103
5	0.267181	0.726687	1.475884	2.015048	2.57058	3.36493	4.03214	6.8688
6	0.264835	0.717558	1.439756	1.943180	2.44691	3.14267	3.70743	5.9588
7	0.263167	0.711142	1.414924	1.894579	2.36462	2.99795	3.49948	5.4079
8	0.261921	0.706387	1.396815	1.859548	2.30600	2.89646	3.35539	5.0413
9	0.260955	0.702722	1.383029	1.833113	2.26216	2.82144	3.24984	4.7809
10	0.260185	0.699812	1.372184	1.812461	2.22814	2.76377	3.16927	4.5869
11	0.259556	0.697445	1.363430	1.795885	2.20099	2.71808	3.10581	4.4370
12	0.259033	0.695483	1.356217	1.782288	2.17881	2.68100	3.05454	43178
13	0.258591	0.693829	1.350171	1.770933	2.16037	2.65031	3.01228	4.2208
14	0.258213	0.692417	1.345030	1.761310	2.14479	2.62449	2.97684	4.1405
15	0.257885	0.691197	1.340606	1.753050	2.13145	2.60248	2.94671	4.0728
16	0.257599	0.690132	1.336757	1.745884	2.11991	2.58349	2.92078	4.0150
17	0.257347	0.689195	1.333379	1.739607	2.10982	2.56693	2.89823	3.9651
18	0.257123	0.688364	1.330391	1.734064	2.10092	2.55238	2.87844	3.9216
19	0.256923	0.687621	1.327728	1.729133	2.09302	2.53948	2.86093	3.8834
20	0.256743	0.686954	1.325341	1.724718	2.08596	2.52798	2.84534	3.8495
21	0.256580	0.686352	1.323188	1.720743	2.07961	2.51765	2.83136	3.8193
22	0.256432	0.685805	1.321237	1.717144	2.07387	2.50832	2.81876	3.7921
23	0.256297	0.685306	1.319460	1.713872	2.06866	2.49987	2.80734	3.7676

Example 2.18 – Grade Point Ratio Study

- **Example 2.18:** Grade point ratios (GPRs) have been recorded for a random sample of 16 from the entering freshman class at a major university. It can be assumed that the distribution of GPR values is approximately normal. The sample yielded a mean, $\bar{y}$ = 3.1, and standard deviation, $s = 0.8$. The nationwide mean GPR of entering freshmen is μ = 2.7. We want to know the probability of getting this sample mean (or higher) if the mean GPR of this university is the same as the nationwide population of students.
- **What do we want to know?** We want the probability of getting a $\bar{y}$ that is greater than or equal to 3.1 from a population whose mean is 2.7.

Example 2.18 – Grade Point Ratio Study

- **Statistically,** we want to know $\Pr(\bar{Y} \geq 3.1)$, where $\bar{Y}$ is the sample mean from 16 samples with $N(2.7, \sigma^2)$. If we know σ^2, we can use central limit theorem to calculate it. Unfortunately, we do not know the population variance σ^2.
- **Solution:** we can use the t-distribution.

$$\Pr(\bar{Y} \geq 3.1) = \Pr\left(\frac{\bar{Y} - 2.7}{\frac{s}{\sqrt{16}}} \geq \frac{3.1 - 2.1}{\frac{s}{\sqrt{16}}}\right) = \Pr\left(T_{15} \geq \frac{3.1 - 2.1}{\frac{0.8}{\sqrt{16}}}\right) = \Pr(T_{15} \geq 2.0)$$

- From Appendix Table on Slide 133, we can find the range of this probability, which is between 0.025 and 0.05. Therefore, we can say that the probability of obtaining a sample mean this large or larger is between 0.025 and 0.05.
- The exact probability from SAS is 0.032.

Exercise – Filling Bottles

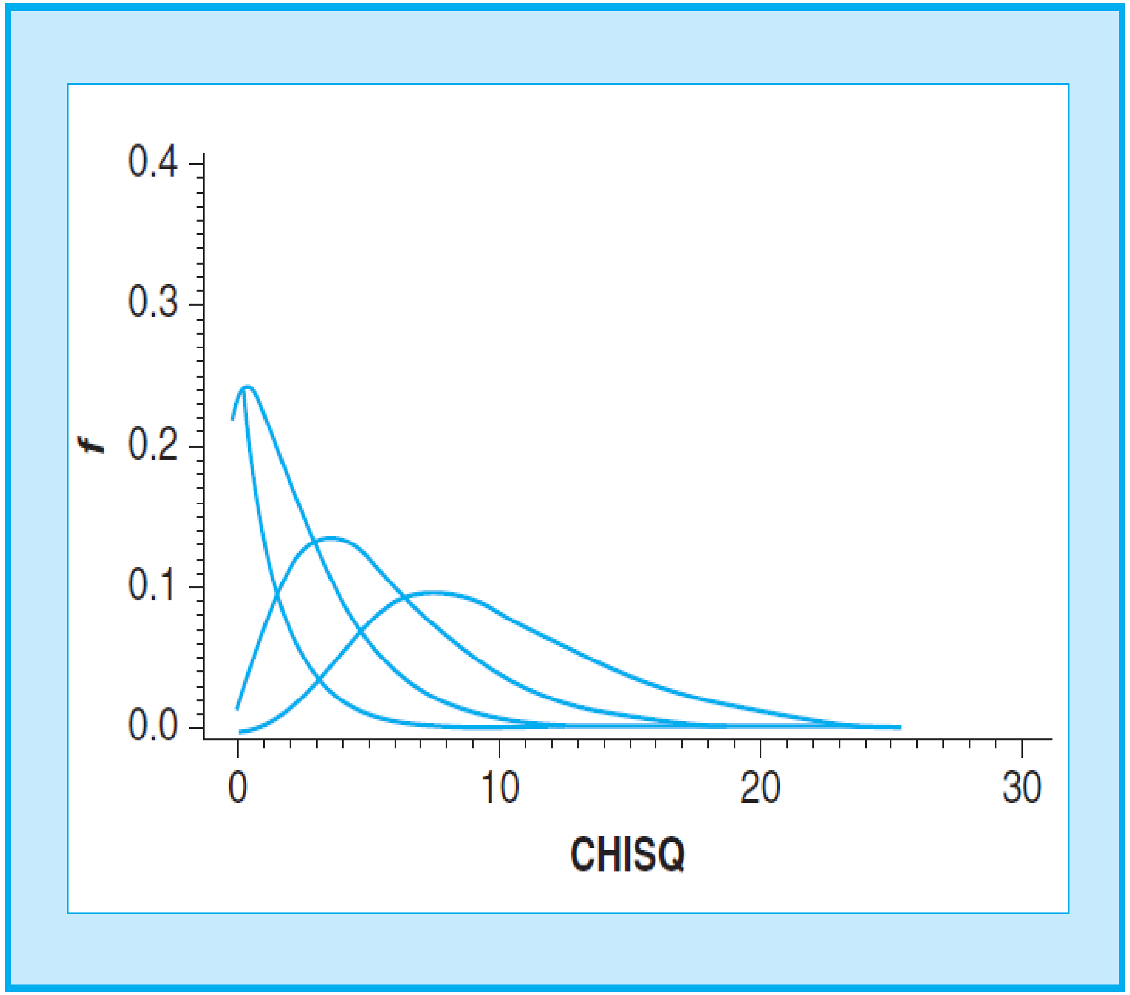
- Consider the filling operation for 20-oz bottles of a popular soft drink. Historically, this operation average 20.2 oz. A recent random sample of 12 bottles yielded these volumes:

20.1	**20.1**	**20.0**	**19.9**	**20.5**	**20.9**
20.1	**20.4**	**20.2**	**19.1**	**20.1**	**20.0**

- From this data, we have $\bar{y} = 20.12; s^2 = 0.1779$
- **Exercise:** assume that the historical average is true, find the probability that you can observe a sample mean from 12 random samples not greater than 20.12? **Why we want to do this?**

Other Distributions

- Chi-square distribution describes the distribution of sample variance.
- We use χ_n to represent a random variable with n degrees of freedom.
- *F* distribution describes the distribution of the ratio of two estimates of population variance.



How to Check Normality

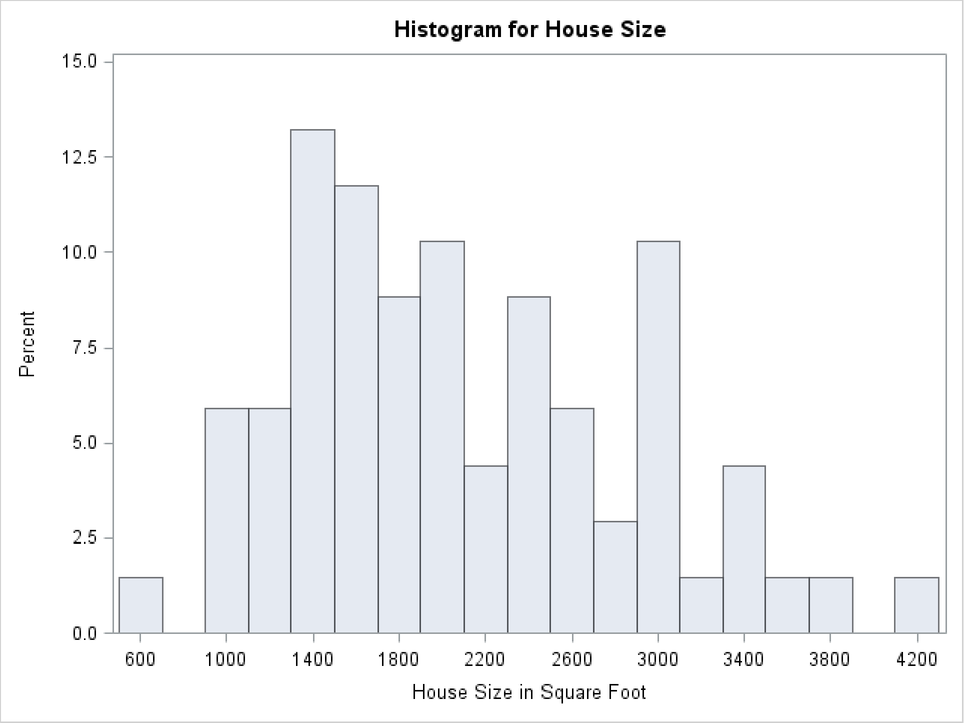
- We can see that we require that the data have a normal distribution in many of our calculations and models.
- How do I know if the data have a normal distribution?
- Construct either a histogram or stem-and-leaf display for the data and check the shape of the graph. If the data are approximately normal, the shape of the histogram or stem-and-leaf display should be bell-shaped.

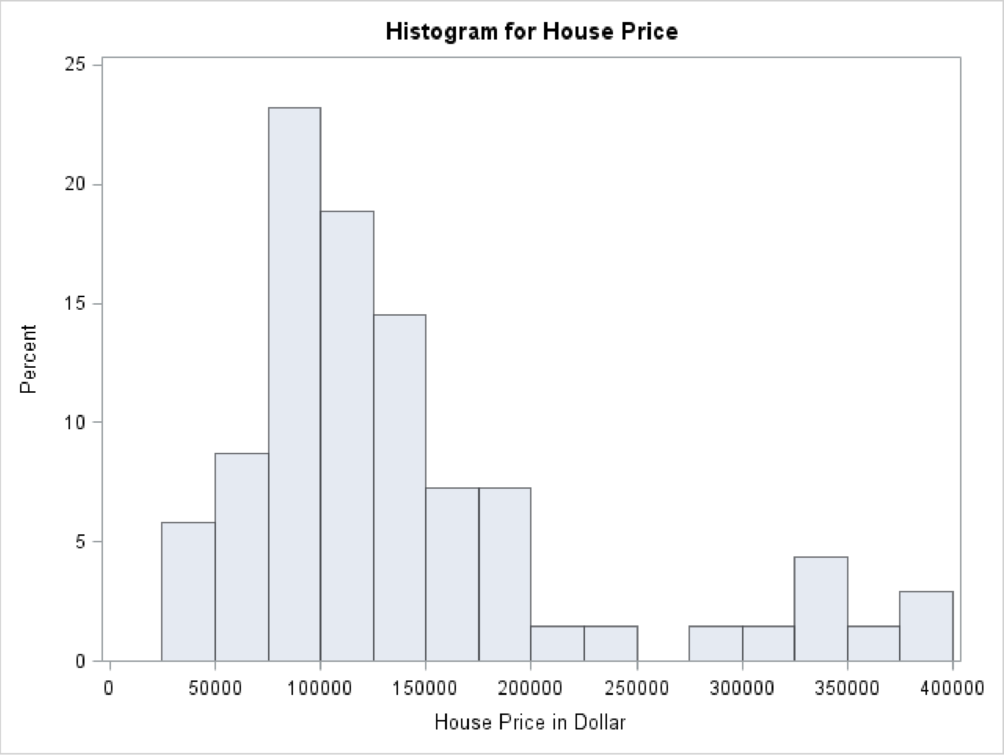
Data – Texas House Data

Obs	Zip	Age	Bed	Bath	Size	Lot	Exter	garage	fp	Price
1	3	21	3	2	951	64904	Other	0	0	30000
3	4	7	1	1	676	54450	Other	2	0	46500
5	1	51	3	1	1186	10857	Other	1	0	51500
7	3	8	3	2	1368	.	Frame	0	0	56990
9	1	51	2	1	1176	6259	Frame	1	1	65500

Data: 69 families in a midsized city in east Texas. This is only part of it.

Histogram - Texas House Data





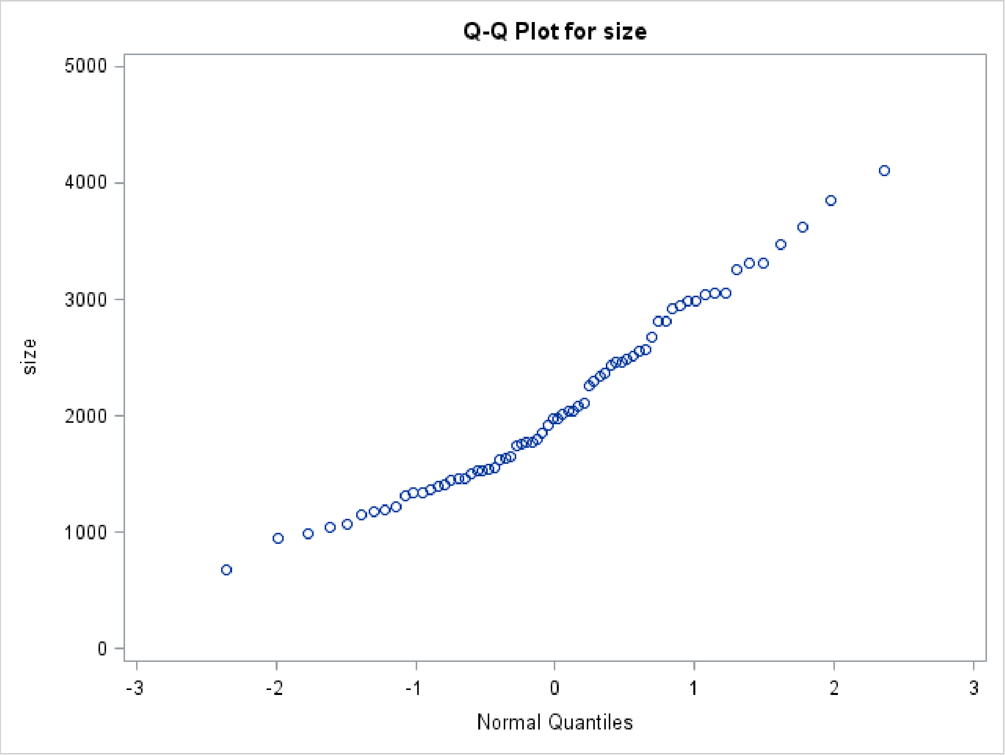
How to Check Normality

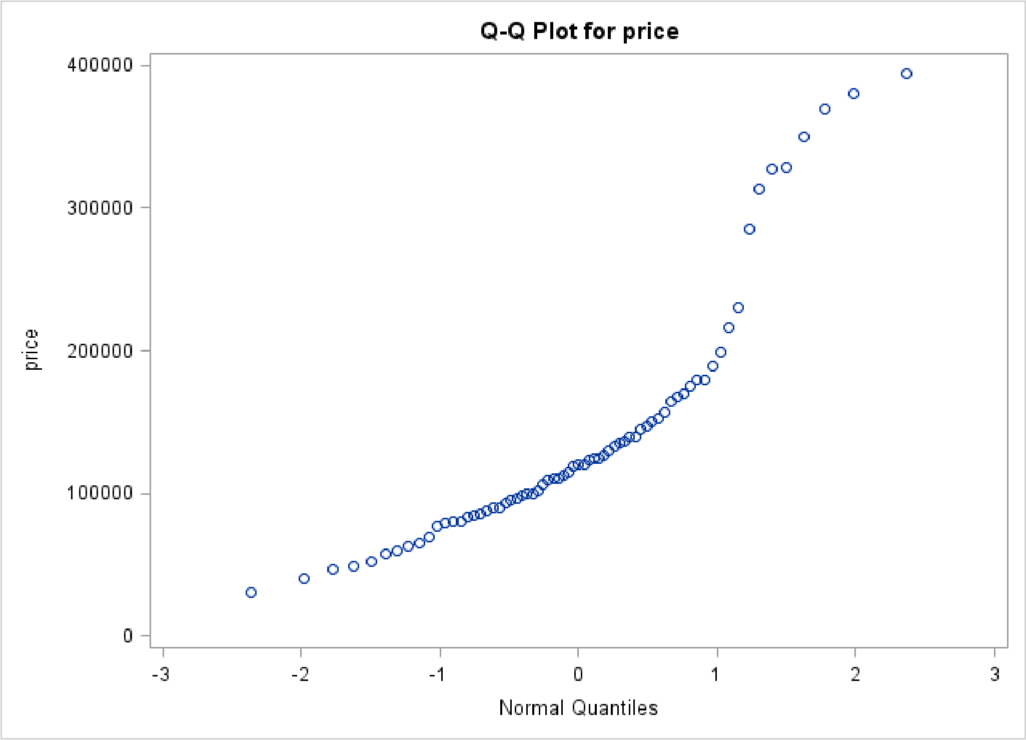
- We can use Quantile-Quantile (Q-Q) plot to check the normality.
- The Q-Q plot is a graphical technique for determining if a set of data plausibly came from some theoretical distribution such as a **Normal** or other distribution.
- A Q-Q plot is a scatterplot created by plotting two sets of quantiles against one another. If both sets of quantiles came from the same distribution, we should see the points forming a line that's roughly straight.

Normal Q-Q Plot

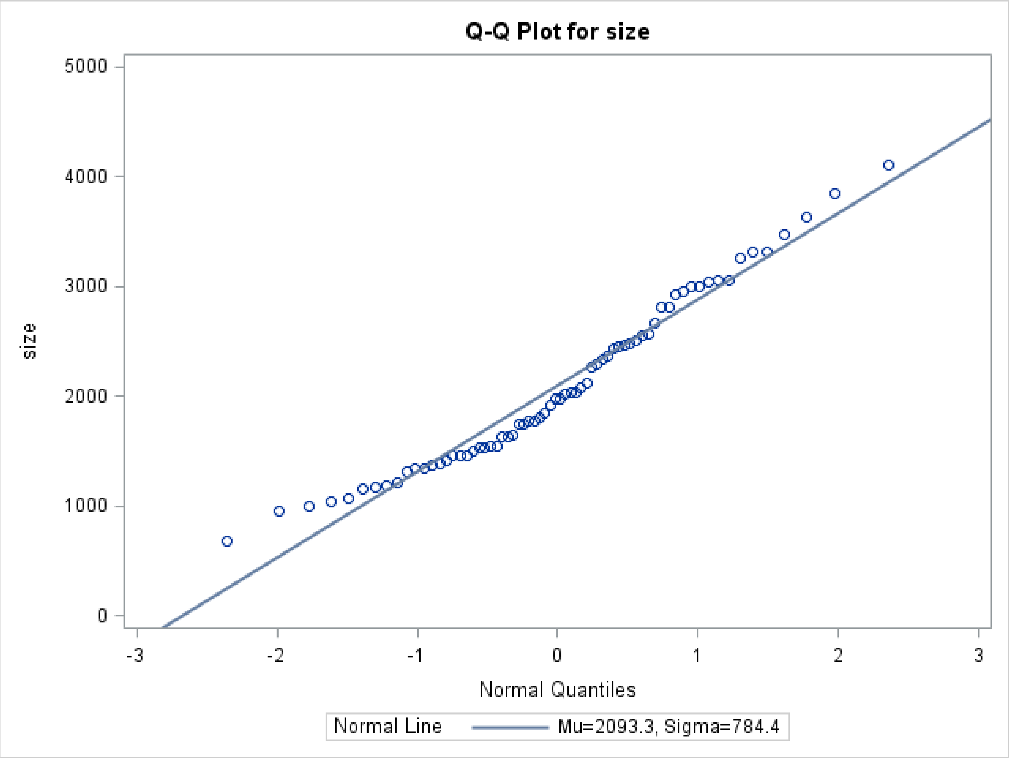
- The “quantile” are often referred to as “percentiles”.
- Q-Q plots take your sample data, sort it in ascending order, and then plot them versus quantiles calculated from a theoretical distribution.
- The number of quantiles is selected to match the size of your sample data.
- For normal Q-Q plot, the theoretical distribution is normal.
- For example, you can calculate Q_1, Q_2, Q_3 from the data, then correspond quantiles from the standard normal distribution are $z_{0.75} = -0.6745$, $z_{0.5} = 0$, and $z_{0.25} = 0.6745$, respectively.

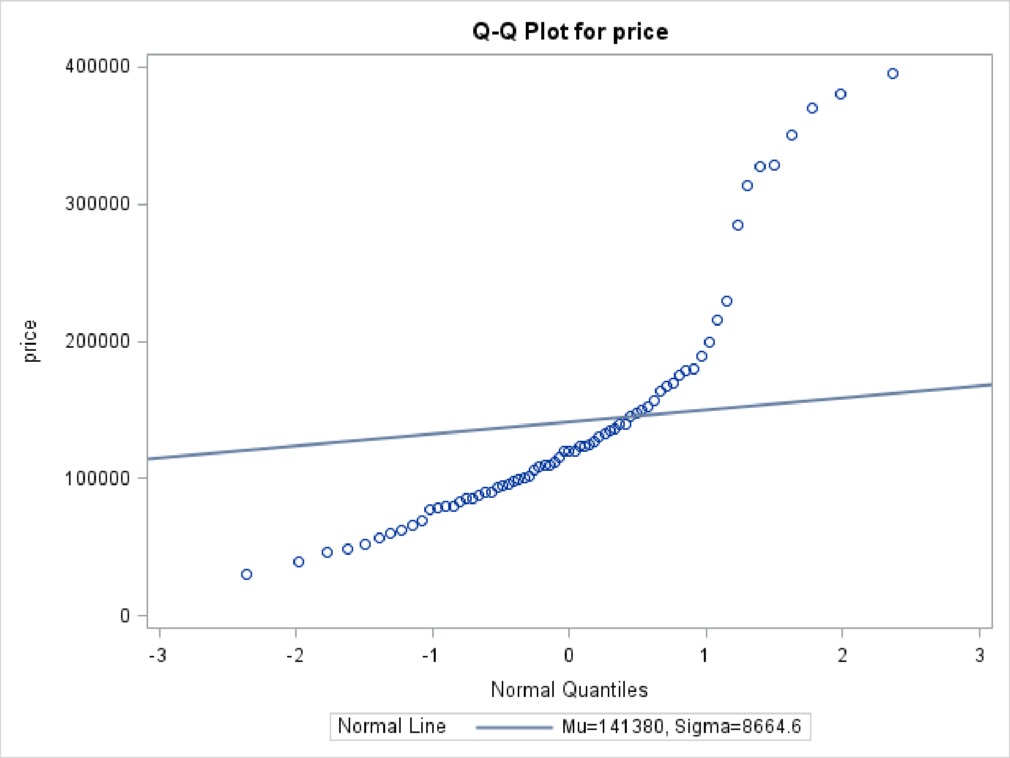
Q-Q Plot - Texas House Data





Q-Q Plot - Texas House Data


Q-Q Plot for size
size
Normal Quantiles
Normal Line
Mu=2093.3, Sigma=784.4


Q-Q Plot for price
price
Normal Quantiles
Normal Line
Mu=141380, Sigma=8664.6