Week 3

Probability and Probability Models

Dr. Esam Mahdi

Department of Mathematics, Statistics and Physics

Qatar University

Learning Objectives

- Determine the differences between proportion and probability concepts.
- Compute probability for two events & check for the presence of a statistical relationship.
    - Joint probability.
    - Conditional probability and marginal (univariate) probability.
    - Odds ratio.
- Compute screening tests and predictive values.
- Probability models for continuous data.
    - The normal distribution.
    - The standard normal distribution.
- Probability models for discrete data.
    - Binomial distribution.
    - Poisson distribution.

Probability and Probability Models

- **Proportion and Probability:**
- Consider a population with certain binary characteristic (diseased person and healthy person):
  - What is the chance that a person with the characteristic (diseased person) will be selected?
  - The answer depends on the size of the subpopulation to which he or she belongs (i.e., the proportion). The larger the proportion, the higher the chance (of such a person being selected).
- That chance is measured by the proportion (a number between 0 and 1) called the probability.
  - **Proportion** is a descriptive statistic measures size.
  - **Probability** measures chance.
- When we are concerned about the outcome (uncertain even) with a random selection, a proportion becomes a probability.
- The probability of an event in a target population is defined as the *relative frequency* (i.e., proportion) with which the event occurs in that target population.
- **For example**, suppose that out of $N = 100{,}000$ persons of a certain target population, a total of 5,500 are positive reactors to a certain screening test; then the probability of being positive is

$$\Pr(\text{positive}) = \frac{5{,}500}{100{,}000} = 0.055 \text{ or } 5.5\%$$

Random Sampling

- Let the size of the target population be N (usually, a very large number), a sample is any subset — say, $n < N$ — of the target population.
- Simple random sampling from the target population is sampling so that every possible sample of size n has an equal chance of selection.
- For simple random sampling:

1. Each individual draw is uncertain with respect to any event or characteristic under investigation (e.g., having a disease), but
2. In repeated sampling from the population, the accumulated long-run relative frequency with which the event occurs is the population relative frequency of the event.

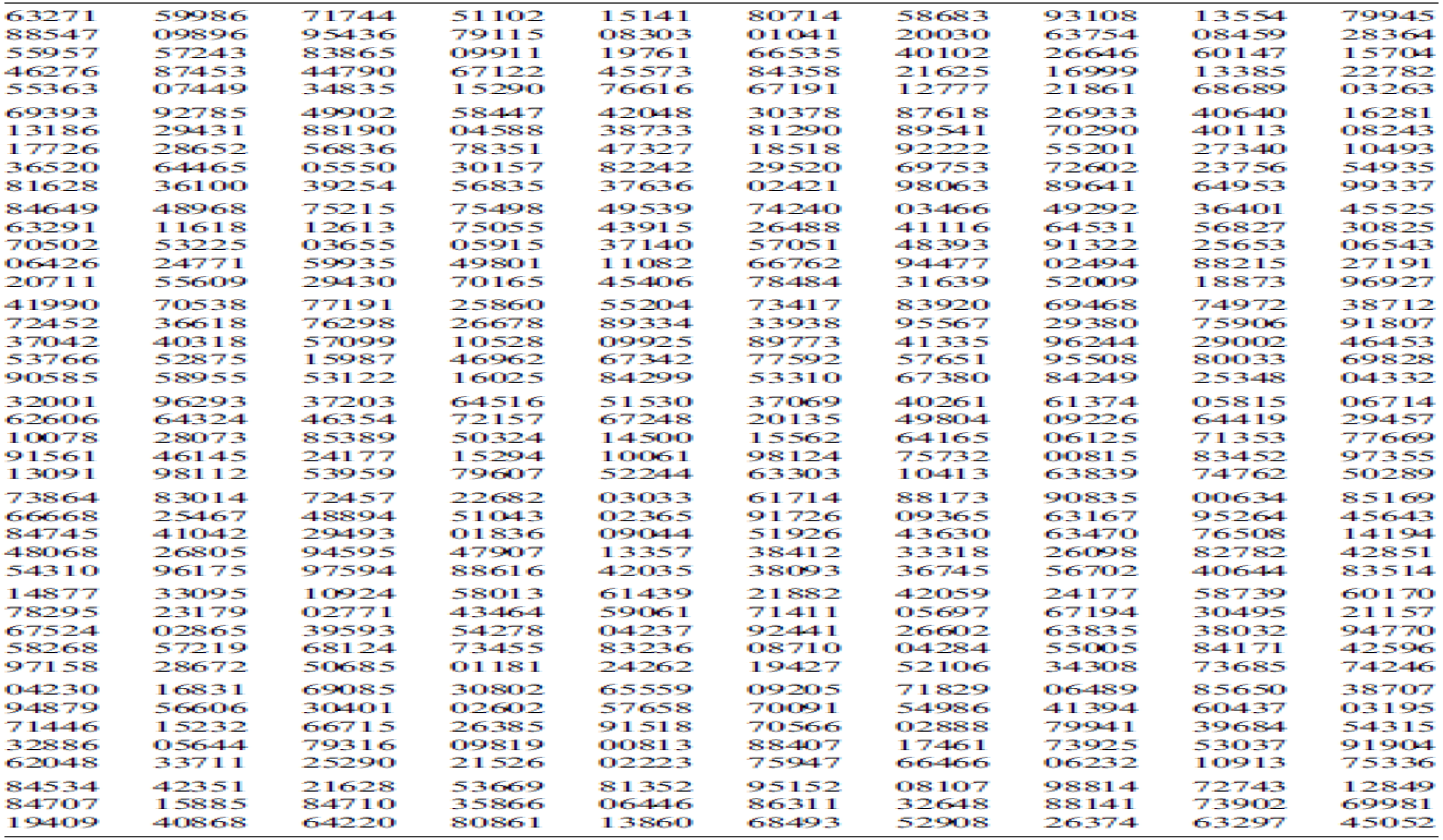
Random Sampling

- The physical process of random sampling can be carried out as follows:

1. A list of all N subjects in the population is obtained. Such a list is termed a frame of the population. The subjects are thus available to an arbitrary numbering (e.g., from $\mathrm{N} = 000\ to\ N = 999$). The frame is often based on a directory (telephone, city, etc.) or on hospital records.
2. A tag is prepared for each subject carrying a number $1, 2, \dots, N$.
3. The tags are placed in a receptacle (e.g., a box) and mixed thoroughly.
4. A tag is drawn blindly. The number on the tag then identifies the subject from the population; this subject becomes a member of the sample.

- Steps 2 to 4 can also be implemented using a table of random numbers (Appendix A). Arbitrarily pick a three-digit column (or four-digit column if the population size is large), and a number selected arbitrarily in that column serves to identify the subject from the population. In practice, this process has been computerized.

Table of Random Numbers



63271	59986	71744	51102	15141	80714	58683	93108	13554	79945
88547	09896	95436	79115	08303	01041	20030	63754	08459	28364
55957	57243	83865	09911	19761	66535	40102	26646	60147	15704
46276	87453	44790	67122	45573	84358	21625	16999	13385	22782
55363	07449	34835	15290	76616	67191	12777	21861	68689	03263
69393	92785	49902	58447	42048	30378	87618	26933	40640	16281
13186	29431	88190	04588	38733	81290	89541	70290	40113	08243
17726	28652	56836	78351	47327	18518	92222	55201	27340	10493
36520	64465	05550	30157	82242	29520	69753	72602	23756	54935
81628	36100	39254	56835	37636	02421	98063	89641	64953	99337
84649	48968	75215	75498	49539	74240	03466	49292	36401	45525
63291	11618	12613	75055	43915	26488	41116	64531	56827	30825
70502	53225	03655	05915	37140	57051	48393	91322	25653	06543
06426	24771	59935	49801	11082	66762	94477	02494	88215	27191
20711	55609	29430	70165	45406	78484	31639	52009	18873	96927
41990	70538	77191	25860	55204	73417	83920	69468	74972	38712
72452	36618	76298	26678	89334	33938	95567	29380	75906	91807
37042	40318	57099	10528	09925	89773	41335	96244	29002	46453
53766	52875	15987	46962	67342	77592	57651	95508	80033	69828
90585	58955	53122	16025	84299	53310	67380	84249	25348	04332
32001	96293	37203	64516	51530	37069	40261	61374	05815	06714
62606	64324	46354	72157	67248	20135	49804	09226	64419	29457
10078	28073	85389	50324	14500	15562	64165	06125	71353	77669
91561	46145	24177	15294	10061	98124	75732	00815	83452	97355
13091	98112	53959	79607	52244	63303	10413	63839	74762	50289
73864	83014	72457	22682	03033	61714	88173	90835	00634	85169
66668	25467	48894	51043	02365	91726	09365	63167	95264	45643
84745	41042	29493	01836	09044	51926	43630	63470	76508	14194
48068	26805	94595	47907	13357	38412	33318	26098	82782	42851
54310	96175	97594	88616	42035	38093	36745	56702	40644	83514
14877	33095	10924	58013	61439	21882	42059	24177	58739	60170
78295	23179	02771	43464	59061	71411	05697	67194	30495	21157
67524	02865	39593	54278	04237	92441	26602	63835	38032	94770
58268	57219	68124	73455	83236	08710	04284	55005	84171	42596
97158	28672	50685	01181	24262	19427	52106	34308	73685	74246
04230	16831	69085	30802	65559	09205	71829	06489	85650	38707
94879	56606	30401	02602	57658	70091	54986	41394	60437	03195
71446	15232	66715	26385	91518	70566	02888	79941	39684	54315
32886	05644	79316	09819	00813	88407	17461	73925	53037	91904
62048	33711	25290	21526	02223	75947	66466	06232	10913	75336
84534	42351	21628	53669	81352	95152	08107	98814	72743	12849
84707	15885	84710	35866	06446	86311	32648	88141	73902	69981
19409	40868	64220	80861	13860	68493	52908	26374	63297	45052

Probability and Random Sampling

- We can link the concepts of probability and random sampling as follows:
- In the example of cancer screening in a community of $N = 100{,}000$ persons, the calculated probability of 0.055 is interpreted as: ‘‘The probability of a randomly drawn person from the target population having a positive test result is 0.055 or 5.5%.’’ The rationale is as follows:
- On an initial draw, the subject chosen may or may not be a positive reactor. However, if this process—of randomly drawing one subject at a time from the population—is repeated over and over again a large number of times, the accumulated long-run relative frequency of positive receptors in the sample will approximate 0.055.

Statistical Relationship

X = test result

Y = actual case in real life

- Data from the cancer screening test of Example 1.4 are summarized in Table 3.1.
- The probability of a positive test result:

$$\Pr(X = +) = \frac{516}{24{,}103} = 0.021$$

- The probability of a negative test result:

$$\Pr(X = -) = \frac{23{,}587}{24{,}103} = 0.979$$

- The probability of having a disease:

$$\Pr(Y = +) = \frac{379}{24{,}103} = 0.015$$

- The probability of not having a disease:

$$\Pr(Y = -) = \frac{23{,}724}{24{,}103} = 0.985$$

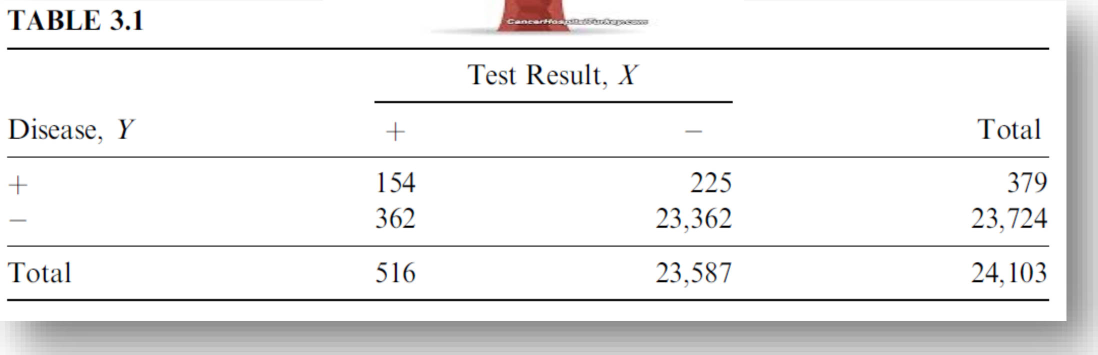
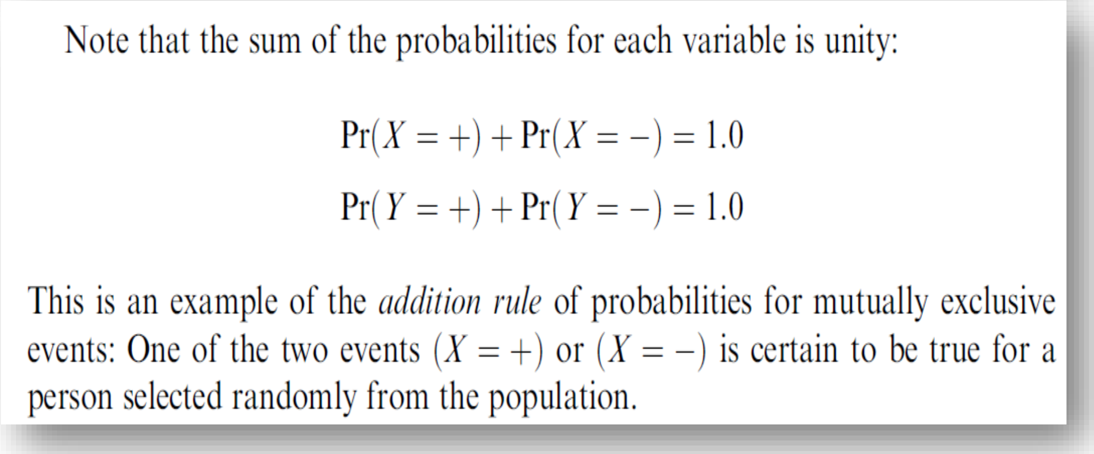


TABLE 3.1

	Test Result, X		
Disease, Y	$+$	$-$	Total
$+$	154	225	379
$-$	362	23,362	23,724
Total	516	23,587	24,103


Note that the sum of the probabilities for each variable is unity:

$$\Pr(X = +) + \Pr(X = -) = 1.0$$

$$\Pr(Y = +) + \Pr(Y = -) = 1.0$$

This is an example of the *addition rule* of probabilities for mutually exclusive events: One of the two events $(X = +)$ or $(X = -)$ is certain to be true for a person selected randomly from the population.

Joint Probability

- **Joint probability** is the probability for two events (such as having the disease and having a positive test result) occurring simultaneously.
- With two variables, X and Y, there are four conditions of outcomes and the associated joint probabilities are:

$$\Pr(X = +, Y = +) = \frac{154}{24{,}103} = 0.006$$

$$\Pr(X = +, Y = -) = \frac{362}{24{,}103} = 0.015$$

$$\Pr(X = -, Y = +) = \frac{225}{24{,}103} = 0.009$$

and

$$\Pr(X = -, Y = -) = \frac{23{,}362}{24{,}103} = 0.970$$

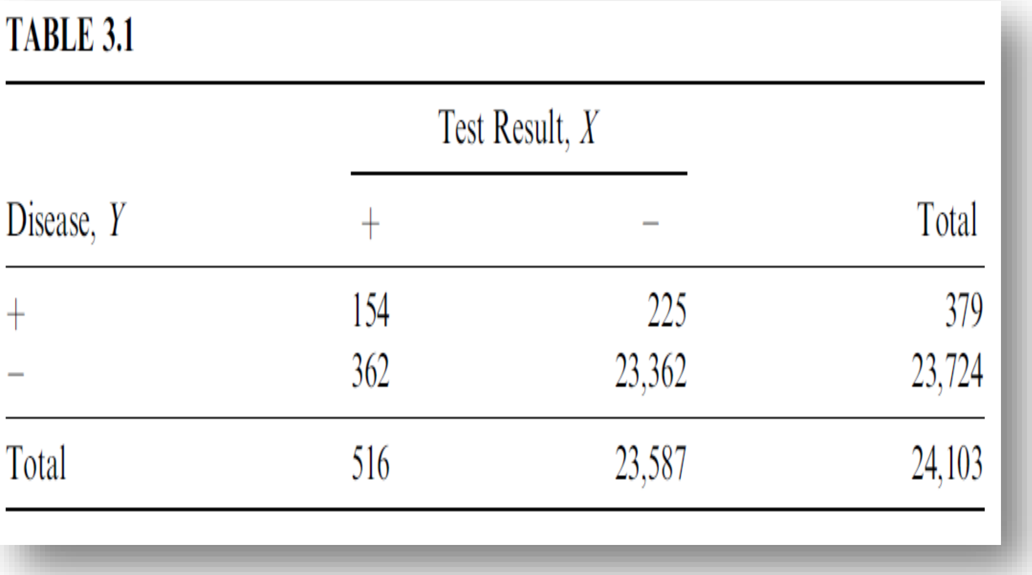


TABLE 3.1

	Test Result, X		
Disease, Y	+	−	Total
+	154	225	379
−	362	23,362	23,724
Total	516	23,587	24,103

Marginal (Univariate) probability

- The joint probabilities in each row (or column) add up to the *marginal or univariate probability* at the margin of that row (or column).

$$\Pr(X=+)=\textstyle\sum_j \Pr(X=+,Y=j)\,,\Pr(X=-)=\sum_j \Pr(X=-,Y=j)$$

$$\Pr(Y=+)=\sum_i \Pr(X=i,Y=+)\,,\Pr(Y=-)=\sum_i \Pr(X=i,Y=-)$$

- For example,

$$\Pr(X=+,Y=+)+\Pr(X=-,Y=+)=\Pr(Y=+)$$
$$=0.015$$

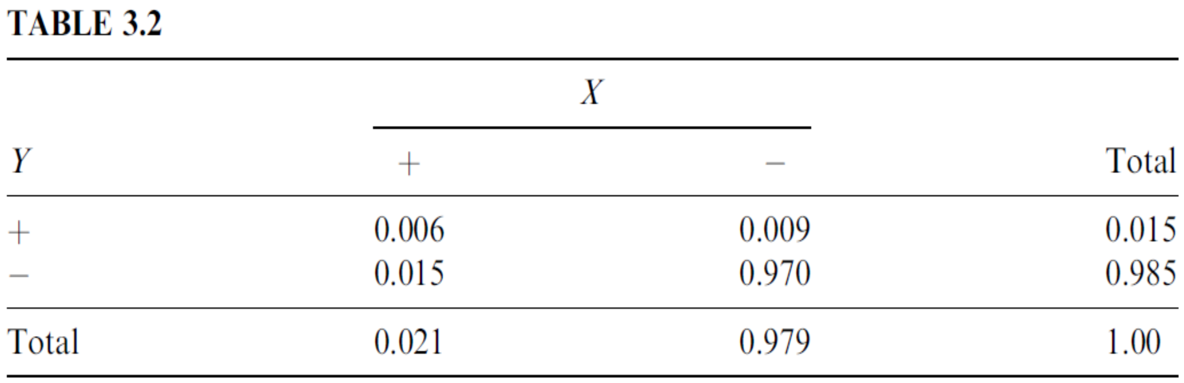


TABLE 3.2

	X		
Y	+	–	Total
+	0.006	0.009	0.015
–	0.015	0.970	0.985
Total	0.021	0.979	1.00

Conditional Probability

- Recall that the probability of an event X occurs ($X = +$) given that another event Y has occurred ($Y = +$) is the conditional probability

$$\Pr(X = +|Y = +) = \frac{\Pr(X = +, Y = +)}{\Pr(Y = +)}, \qquad \Pr(Y = +) \neq 0$$

- Note that $\Pr(X = +, Y = +) = \Pr(X = +|Y = +)\Pr(Y = +)$.
- If the conditional probability equals the marginal probability, $\Pr(\mathrm{X} = \mathrm{i}|\mathrm{Y} = \mathrm{j}) = \Pr(\mathrm{X} = \mathrm{i})$, then the two events $\Pr(X = i)$ and $\Pr(Y = j)$ are said to be *independent (not* statistically associated*).*
- In general X and Y are independent if $\Pr(\mathrm{X} = \mathrm{i}, \mathrm{Y} = \mathrm{j}) = \Pr(\mathrm{X} = \mathrm{i})\Pr(\mathrm{Y} = \mathrm{j})$.
- If the two events are not independent, they have a statistical relationship or we say that they are statistically associated.
- For the screening example,

$$\Pr(X = +) = 0.021$$

$$\Pr(X = + \mid Y = +) = 0.406$$

clearly indicating a strong statistical relationship [because $\Pr(\mathrm{X} = +|\mathrm{Y} = +) \neq \Pr(\mathrm{X} = +)$]

Odds Ratio

- There are several different ways to check for the presence of a statistical relationship.

1. Calculation of the odds ratio.

$$\text{odds ratio} = \frac{\Pr(X = + \mid Y = +)/(\Pr(X = - \mid Y = +))}{\Pr(X = + \mid Y = -)/(\Pr(X = - \mid Y = -))} = \frac{\Pr(X = +, Y = +)\ \Pr(X = -, Y = -)}{\Pr(X = +, Y = -)\ \Pr(X = -, Y = +)}$$

   - The example above yields $\mathrm{OR} = \frac{(0.006)(0.970)}{(0.015)(0.009)} = 43.11$ clearly indicating a statistical relationship.
   - **Note:** when X and Y are independent, the odds ratio equals 1.

2. Comparison of conditional probability and unconditional (or marginal) probability:

   For example, $\Pr(X = +|Y = +)$ versus $\Pr(X = +)$.

3. Comparison of conditional probabilities: for example, $\Pr(X = +|Y = +)$ versus $\Pr(X = +|Y = -)$

The screening example above yields

$$\Pr(X = + \mid Y = +) = 0.406$$

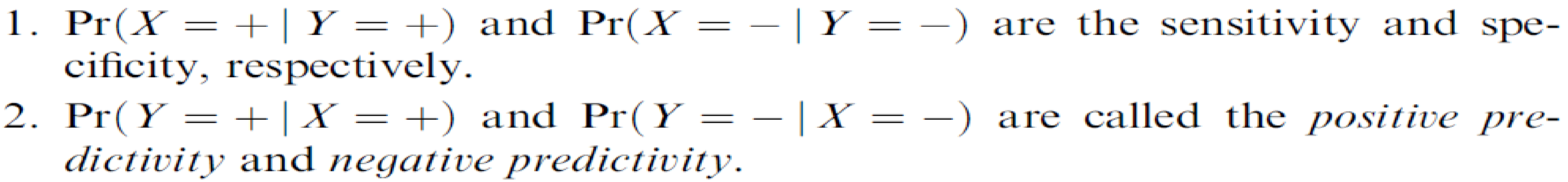
whereas

$$\Pr(X = + \mid Y = -) = \frac{362}{23{,}724} = 0.015$$

again clearly indicating a statistical relationship.

Using Screening Tests

X = test result
Y = actual case in real life



1. $\Pr(X = + \mid Y = +)$ and $\Pr(X = - \mid Y = -)$ are the sensitivity and specificity, respectively.
2. $\Pr(Y = + \mid X = +)$ and $\Pr(Y = - \mid X = -)$ are called the *positive predictivity* and *negative predictivity*.

- With positive predictivity (or positive predictive value), the question is: Given that the test *X* suggests cancer, what is the probability that, in fact, cancer is present?

- Table 3.3 shows that unlike sensitivity and specificity, the positive and negative predictive values depend not only on the efficiency of the test but also on the disease prevalence of the target population.
- In both cases, the test is 90% sensitive and 90% specific. However:

1. Population A has a prevalence of 50%, leading to a positive predictive value of 90%.
2. Population B has a prevalence of 10%, leading to a positive predictive value of 50%.

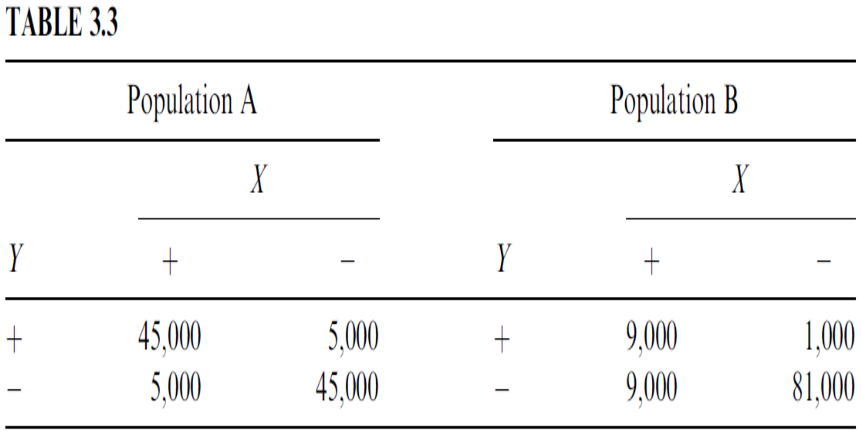


TABLE 3.3

Population A			Population B		
	X			*X*	
Y	+	–	*Y*	+	–
+	45,000	5,000	+	9,000	1,000
–	5,000	45,000	–	9,000	81,000

The conclusion: If a test—even a highly sensitive and highly specific one—is applied to a target population in which the disease prevalence is low, the positive predictive value is low.

Predictive Values

$$\begin{array}{r}\text{positive}\\ \text{predictivity}\end{array} = \frac{(\text{prevalence})(\text{sensitivity})}{(\text{prevalence})(\text{sensitivity}) + (1 - \text{prevalence})(1 - \text{specificity})}$$

and

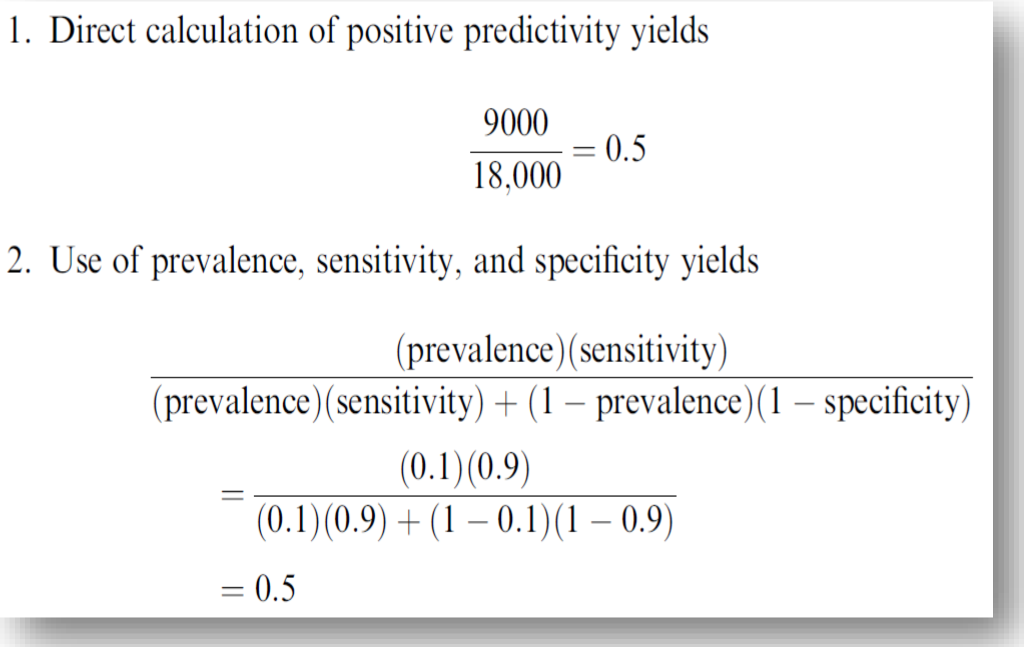
$$\begin{array}{r}\text{negative}\\ \text{predictivity}\end{array} = \frac{(1 - \text{prevalence})(\text{specificity})}{(1 - \text{prevalence})(\text{specificity}) + (\text{prevalence})(1 - \text{sensitivity})}$$

$$\Pr(Y=+\,|\,X=+) = \frac{\Pr(X=+, Y=+)}{\Pr(X=+)}$$

$$= \frac{\Pr(X=+, Y=+)}{\Pr(X=+, Y=+) + \Pr(X=+, Y=-)}$$

$$= \frac{\Pr(Y=+)\Pr(X=+\,|\,Y=+)}{\Pr(Y=+)\Pr(X=+\,|\,Y=+) + \Pr(Y=-)\Pr(X=+\,|\,Y=-)}$$

$$= \frac{\Pr(Y=+)\Pr(X=+\,|\,Y=+)}{\Pr(Y=+)\Pr(X=+\,|\,Y=+) + [1-\Pr(Y=+)][1-\Pr(X=-\,|\,Y=-)]}$$



1. Direct calculation of positive predictivity yields

$$\frac{9000}{18{,}000} = 0.5$$

2. Use of prevalence, sensitivity, and specificity yields

$$\frac{(\text{prevalence})(\text{sensitivity})}{(\text{prevalence})(\text{sensitivity}) + (1 - \text{prevalence})(1 - \text{specificity})}$$

$$= \frac{(0.1)(0.9)}{(0.1)(0.9) + (1-0.1)(1-0.9)}$$

$$= 0.5$$

The Normal Distribution

- The Normal Curve

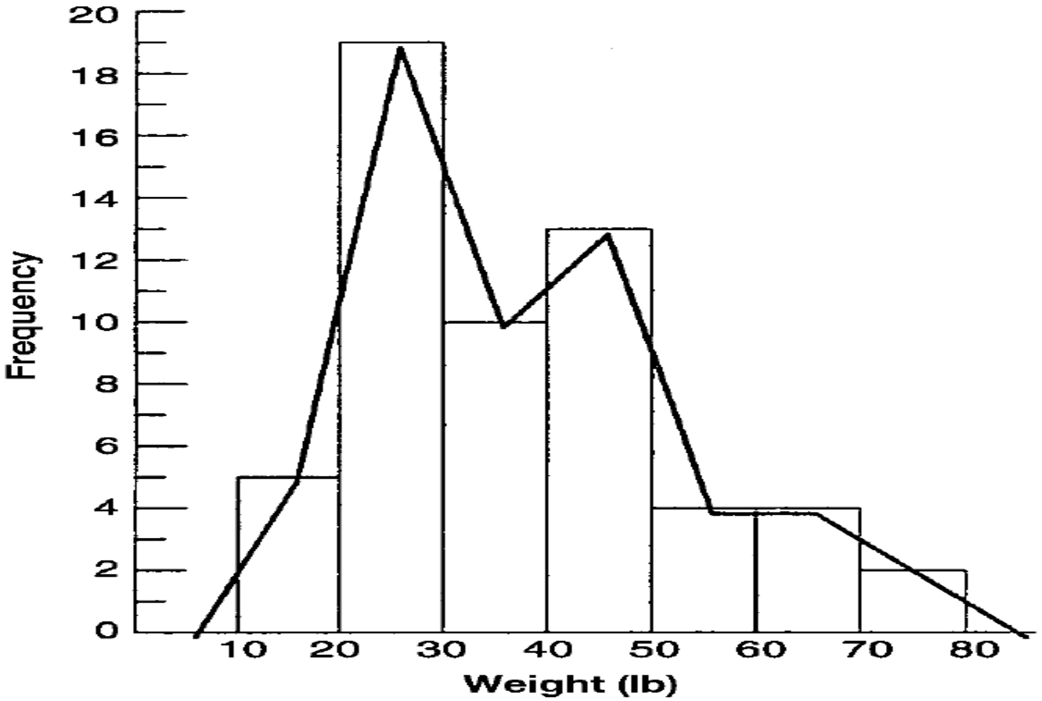


Figure 3.1 Distribution of weights of 57 children.

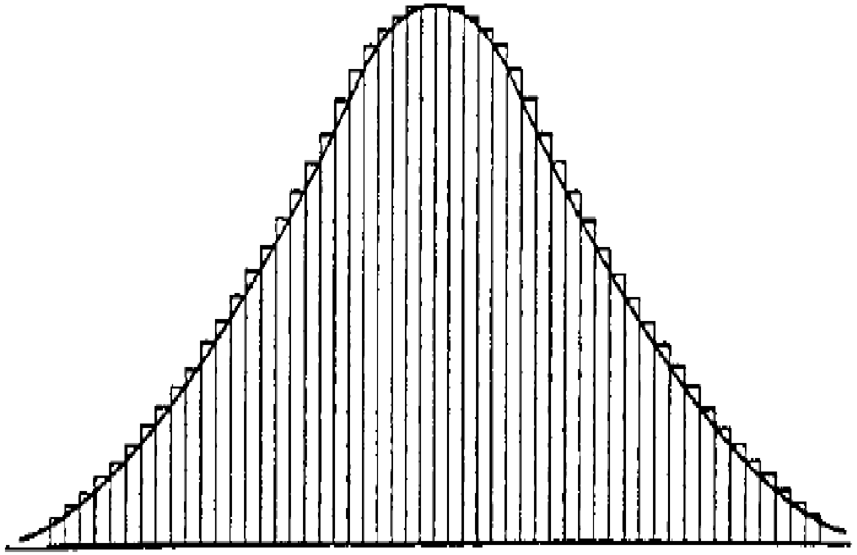


Figure 3.2 Histogram based on a large data set of weights.

Central Limit Theorem

- The term normal curve, refers not to one curve but to a family of curves, each characterized by a mean μ and a variance σ^2. In the special case where $\mu = 0$ and $\sigma^2 = 1$, we have the standard normal curve is commonly designated by the letter Z.
- **Central Limit Theorem:** For samples that are ‘‘big enough,’’ values of their sample means, $\bar{x}'s$ (including sample proportions as a special case), are approximately distributed as normal, even if the samples are taken from really strangely shaped distributions.

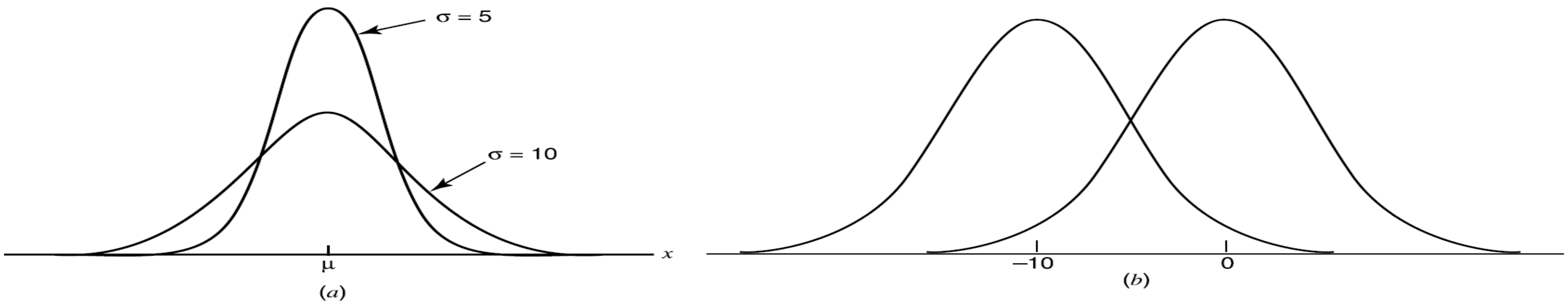


Figure 3.3 Family of normal curves: (*a*) two normal distributions with the same mean but different variances; (*b*) two normal distributions with the same variance but different means.

Areas under the Standard Normal Curve

- About 68% of the area is contained within ∓1:

  $\Pr(-1 < z < 1) = 0.6826$

- About 95% of the area is contained within ∓2:

  $\Pr(-2 < z < 2) = 0.9545$

- More areas under the standard normal curve have been computed and are available in tables, one of which is Appendix B.

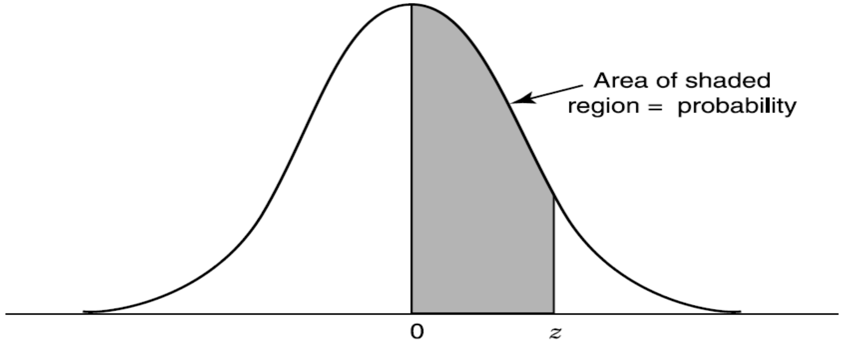


Figure 3.5 Area under the standard normal curve as in Appendix B.

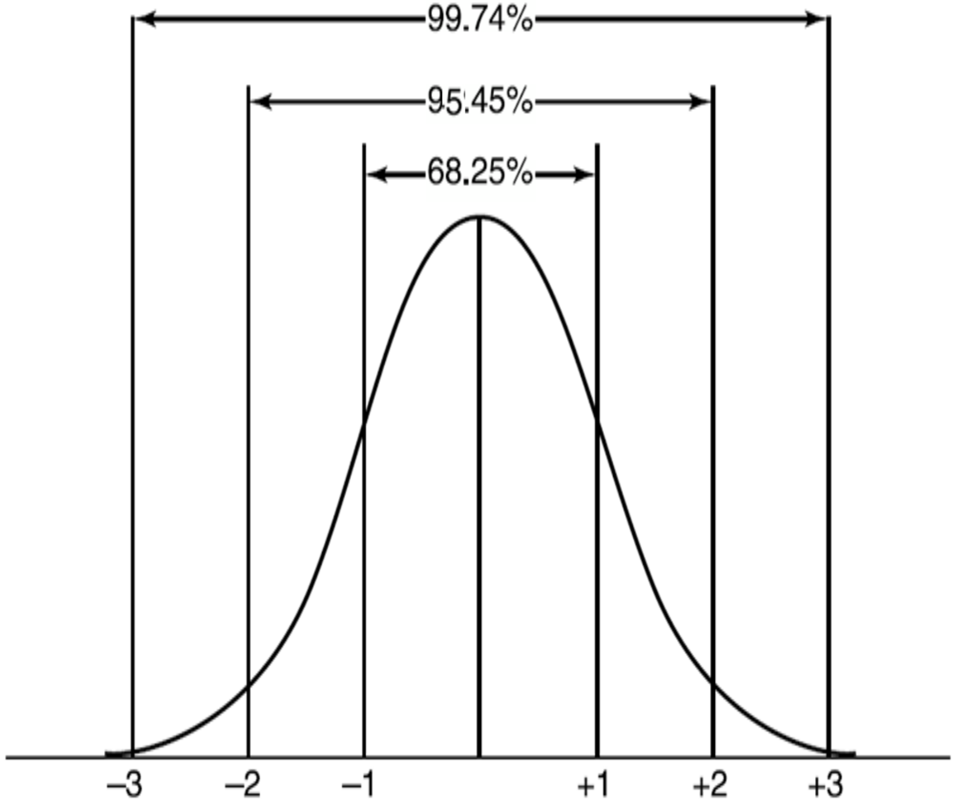
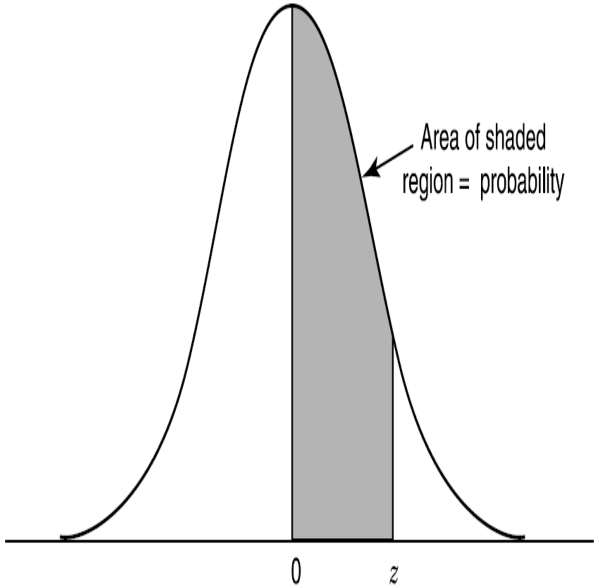
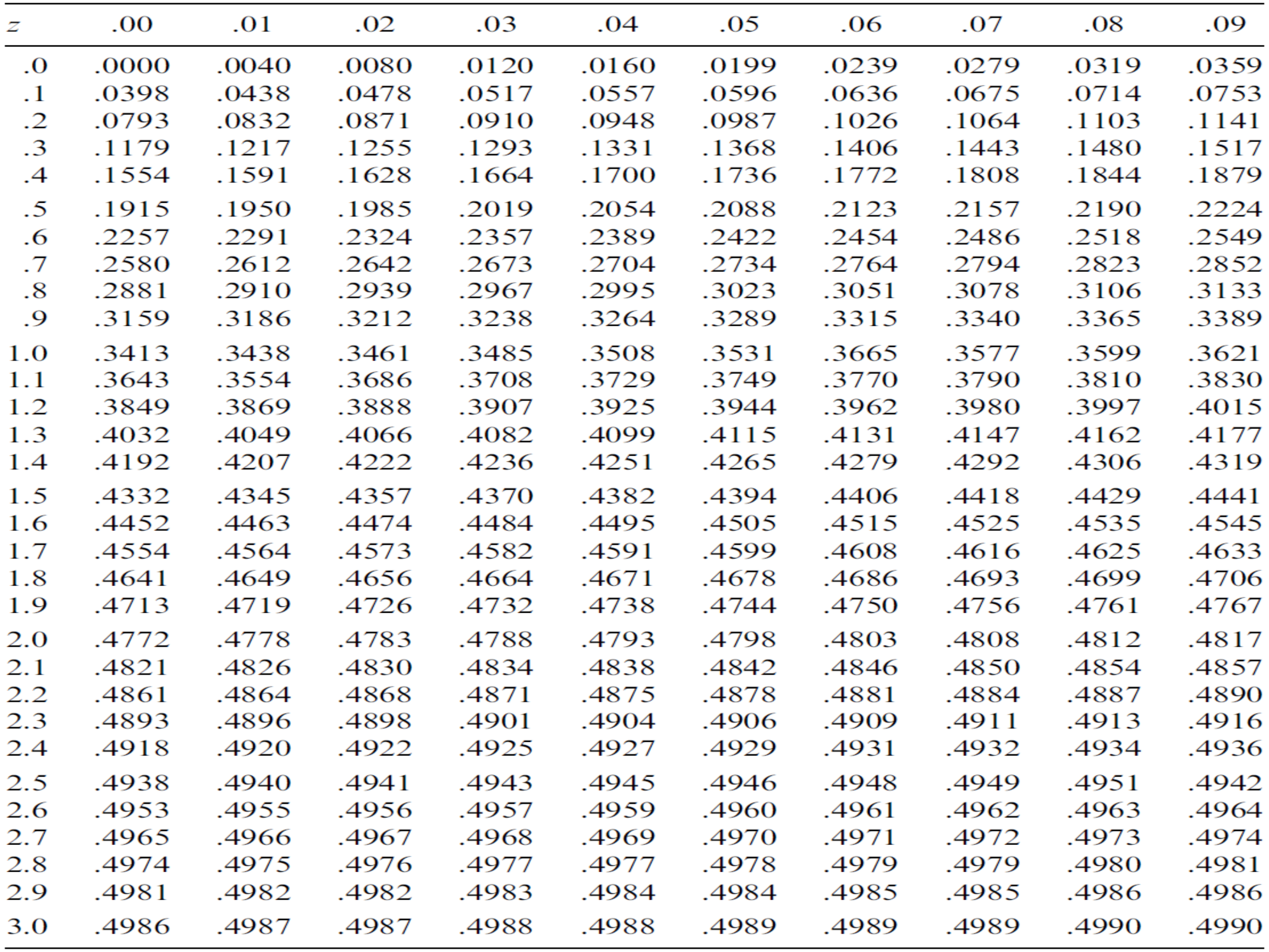


Figure 3.4 Standard normal curve and some important divisions.

Appendix B: Area Under The Standard Normal Curve

- Entries in the table give the area under the curve between the mean and Z standard deviations above the mean.



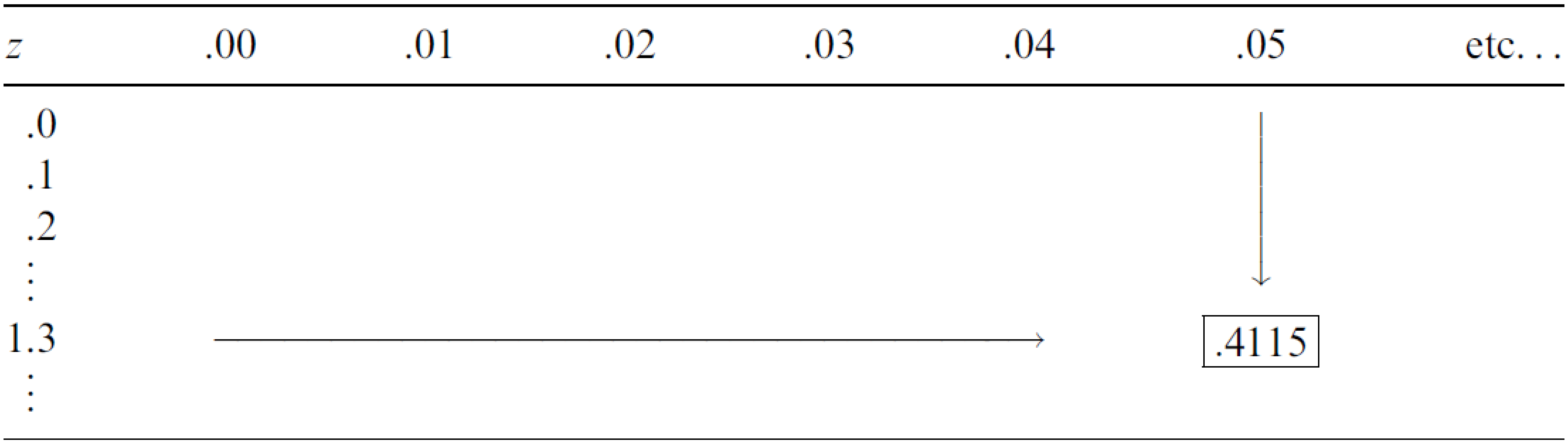


z	.00	.01	.02	.03	.04	.05	.06	.07	.08	.09
.0	.0000	.0040	.0080	.0120	.0160	.0199	.0239	.0279	.0319	.0359
.1	.0398	.0438	.0478	.0517	.0557	.0596	.0636	.0675	.0714	.0753
.2	.0793	.0832	.0871	.0910	.0948	.0987	.1026	.1064	.1103	.1141
.3	.1179	.1217	.1255	.1293	.1331	.1368	.1406	.1443	.1480	.1517
.4	.1554	.1591	.1628	.1664	.1700	.1736	.1772	.1808	.1844	.1879
.5	.1915	.1950	.1985	.2019	.2054	.2088	.2123	.2157	.2190	.2224
.6	.2257	.2291	.2324	.2357	.2389	.2422	.2454	.2486	.2518	.2549
.7	.2580	.2612	.2642	.2673	.2704	.2734	.2764	.2794	.2823	.2852
.8	.2881	.2910	.2939	.2967	.2995	.3023	.3051	.3078	.3106	.3133
.9	.3159	.3186	.3212	.3238	.3264	.3289	.3315	.3340	.3365	.3389
1.0	.3413	.3438	.3461	.3485	.3508	.3531	.3665	.3577	.3599	.3621
1.1	.3643	.3554	.3686	.3708	.3729	.3749	.3770	.3790	.3810	.3830
1.2	.3849	.3869	.3888	.3907	.3925	.3944	.3962	.3980	.3997	.4015
1.3	.4032	.4049	.4066	.4082	.4099	.4115	.4131	.4147	.4162	.4177
1.4	.4192	.4207	.4222	.4236	.4251	.4265	.4279	.4292	.4306	.4319
1.5	.4332	.4345	.4357	.4370	.4382	.4394	.4406	.4418	.4429	.4441
1.6	.4452	.4463	.4474	.4484	.4495	.4505	.4515	.4525	.4535	.4545
1.7	.4554	.4564	.4573	.4582	.4591	.4599	.4608	.4616	.4625	.4633
1.8	.4641	.4649	.4656	.4664	.4671	.4678	.4686	.4693	.4699	.4706
1.9	.4713	.4719	.4726	.4732	.4738	.4744	.4750	.4756	.4761	.4767
2.0	.4772	.4778	.4783	.4788	.4793	.4798	.4803	.4808	.4812	.4817
2.1	.4821	.4826	.4830	.4834	.4838	.4842	.4846	.4850	.4854	.4857
2.2	.4861	.4864	.4868	.4871	.4875	.4878	.4881	.4884	.4887	.4890
2.3	.4893	.4896	.4898	.4901	.4904	.4906	.4909	.4911	.4913	.4916
2.4	.4918	.4920	.4922	.4925	.4927	.4929	.4931	.4932	.4934	.4936
2.5	.4938	.4940	.4941	.4943	.4945	.4946	.4948	.4949	.4951	.4942
2.6	.4953	.4955	.4956	.4957	.4959	.4960	.4961	.4962	.4963	.4964
2.7	.4965	.4966	.4967	.4968	.4969	.4970	.4971	.4972	.4973	.4974
2.8	.4974	.4975	.4976	.4977	.4977	.4978	.4979	.4979	.4980	.4981
2.9	.4981	.4982	.4982	.4983	.4984	.4984	.4985	.4985	.4986	.4986
3.0	.4986	.4987	.4987	.4988	.4988	.4989	.4989	.4989	.4990	.4990

How to Read the Table in Appendix B

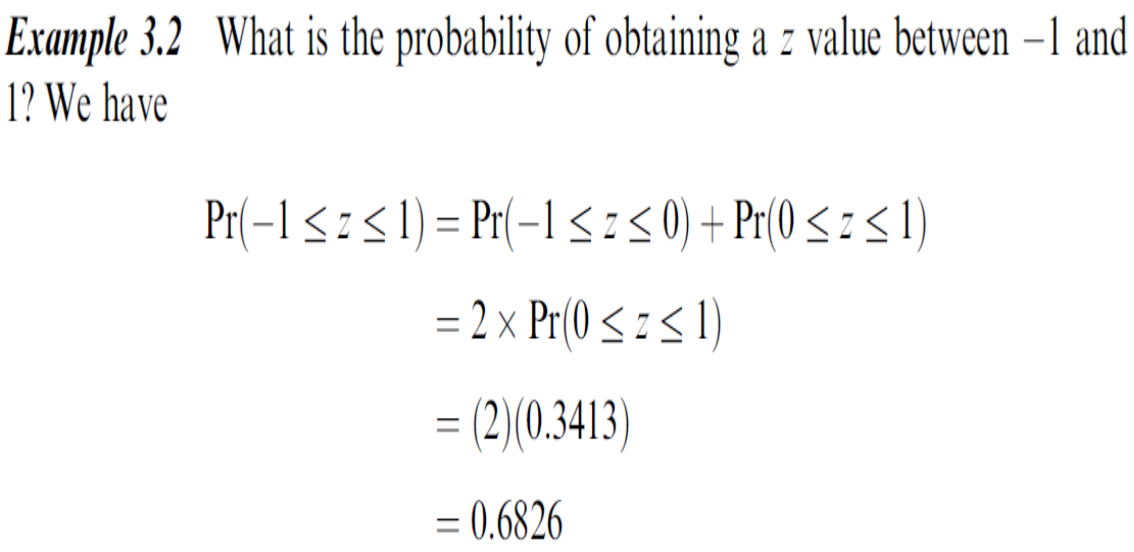
- Suppose that we are interested in the area between $z = 0$ and $z = 1.35$

TABLE 3.8



z	.00	.01	.02	.03	.04	.05	etc. . .
.0						↓	
.1							
.2							
⋮							
1.3	→				→	$\boxed{.4115}$	
⋮							

Example 3.2



Example 3.2 What is the probability of obtaining a z value between −1 and 1? We have

$$\begin{aligned}\Pr(-1 \le z \le 1) &= \Pr(-1 \le z \le 0) + \Pr(0 \le z \le 1)\\ &= 2 \times \Pr(0 \le z \le 1)\\ &= (2)(0.3413)\\ &= 0.6826\end{aligned}$$

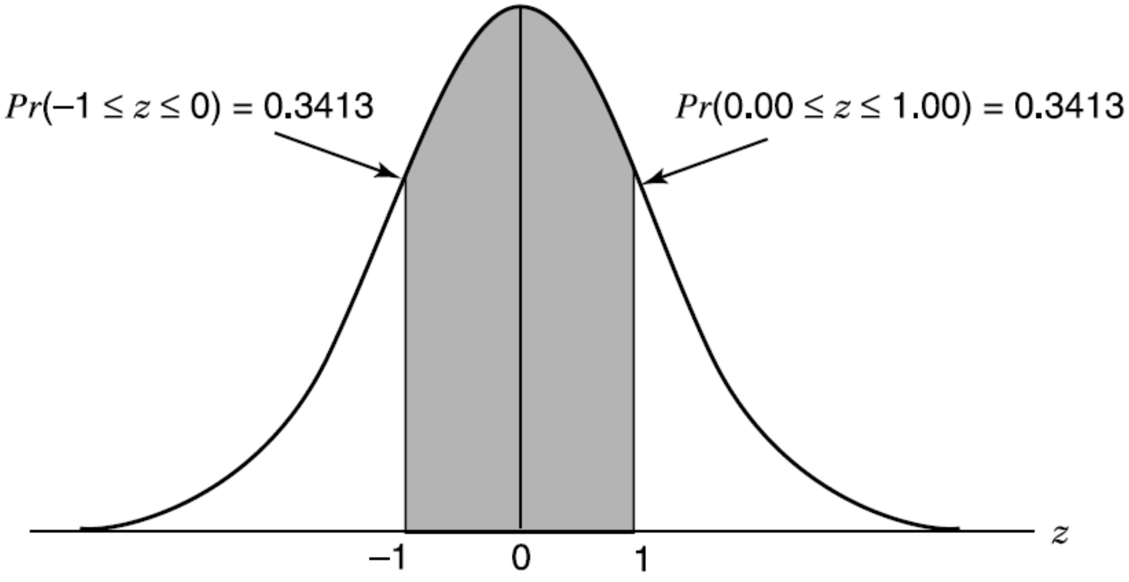


Figure 3.6 Graphical display for Example 3.2.

Examples 3.3 & 3.4

Example 3.3 What is the probability of obtaining a z value of at least 1.58? We have

$$\begin{aligned}\Pr(z \geq 1.58) &= 0.5 - \Pr(0 \leq z \leq 1.58)\\ &= 0.5 - 0.4429\\ &= 0.0571\end{aligned}$$

and this probability is shown in Figure 3.7.

Example 3.4 What is the probability of obtaining a z value of -0.5 or larger? We have

$$\begin{aligned}\Pr(z \geq -0.5) &= \Pr(-0.5 \leq z \leq 0) + \Pr(0 \leq z)\\ &= \Pr(0 \leq z \leq 0.5) + \Pr(0 \leq z)\\ &= 0.1915 + 0.5\\ &= 0.6915\end{aligned}$$

and this probability is shown in Figure 3.8.

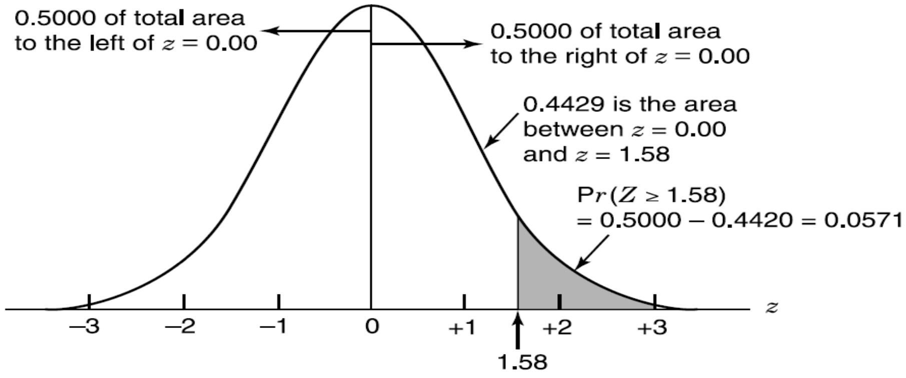


Figure 3.7 Graphical display for Example 3.3.

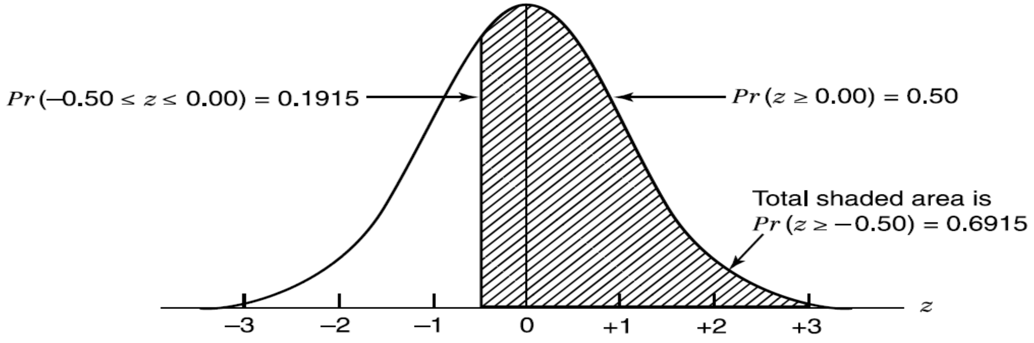
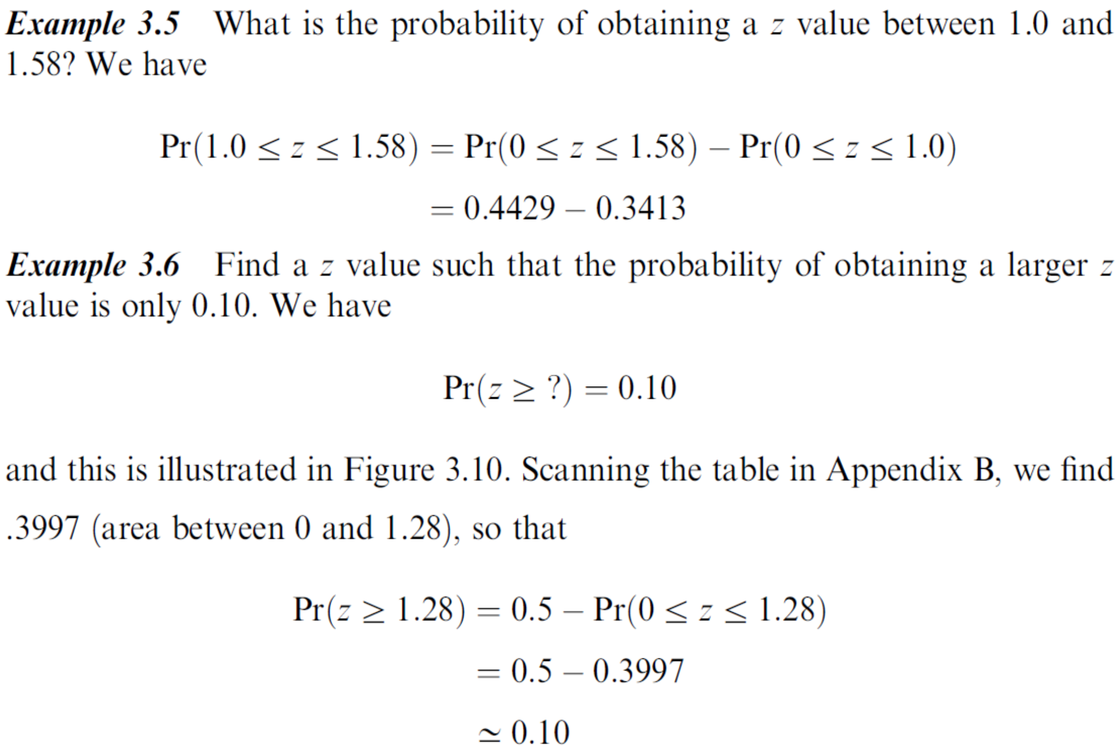


Figure 3.8 Graphical display for Example 3.4.

Examples 3.5 & 3.6



Example 3.5 What is the probability of obtaining a z value between 1.0 and 1.58? We have

$$\begin{aligned}\Pr(1.0 \leq z \leq 1.58) &= \Pr(0 \leq z \leq 1.58) - \Pr(0 \leq z \leq 1.0)\\ &= 0.4429 - 0.3413\end{aligned}$$

Example 3.6 Find a z value such that the probability of obtaining a larger z value is only 0.10. We have

$$\Pr(z \geq ?) = 0.10$$

and this is illustrated in Figure 3.10. Scanning the table in Appendix B, we find .3997 (area between 0 and 1.28), so that

$$\begin{aligned}\Pr(z \geq 1.28) &= 0.5 - \Pr(0 \leq z \leq 1.28)\\ &= 0.5 - 0.3997\\ &\simeq 0.10\end{aligned}$$

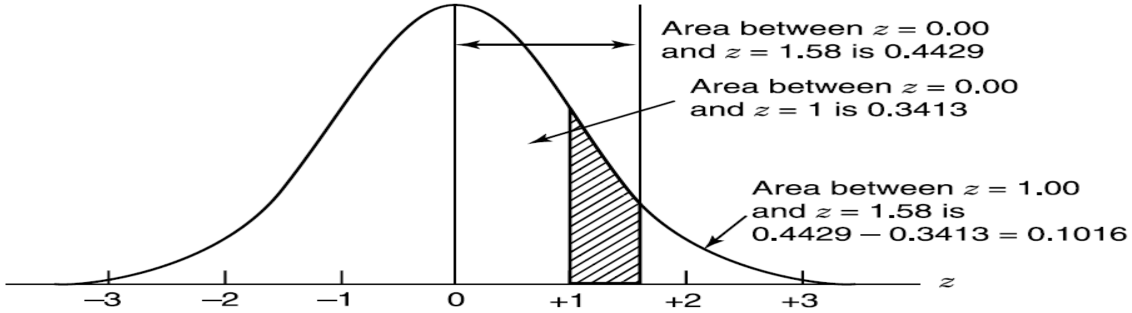

Area between $z = 0.00$ and $z = 1.58$ is 0.4429
Area between $z = 0.00$ and $z = 1$ is 0.3413
Area between $z = 1.00$ and $z = 1.58$ is $0.4429 - 0.3413 = 0.1016$
−3
−2
−1
0
+1
+2
+3
z

Figure 3.9 Graphical display for Example 3.5.

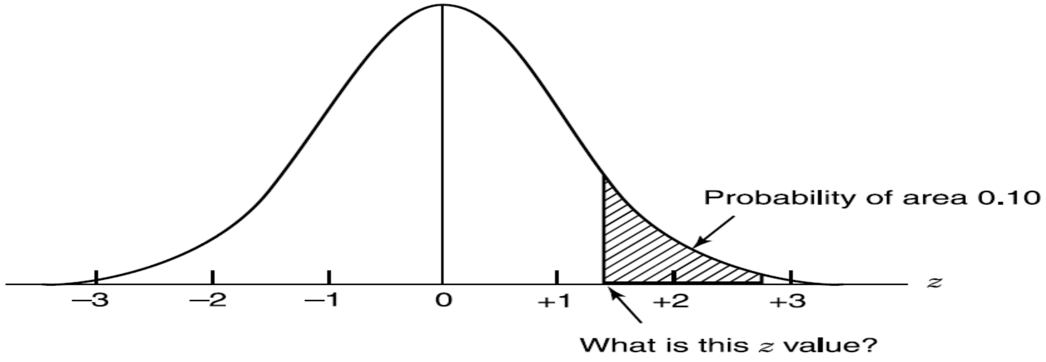


Figure 3.10 Graphical display for Example 3.6.

Standard Normal Distribution

- If $X \sim N(\mu, \sigma^2)$ then z score $Z = \frac{X-\mu}{\sigma} \sim N(0,1)$
- $\Pr(a < X) = \Pr(a \leq X) = \Pr(\frac{a-\mu}{\sigma} < Z) = \Pr(\frac{a-\mu}{\sigma} \leq Z)$
- $\Pr(b > X) = \Pr(b \geq X) = \Pr\left(\frac{b-\mu}{\sigma} > Z\right) = \Pr(\frac{b-\mu}{\sigma} \geq Z)$
- $\Pr(a < X < b) = \Pr(a \leq X \leq b) = \Pr(\frac{a-\mu}{\sigma} < Z < \frac{b-\mu}{\sigma}) = \Pr(\frac{a-\mu}{\sigma} \leq Z \leq \frac{b-\mu}{\sigma})$

Example 3.7

- If the total cholesterol values for a certain target population are approximately normally distributed with a mean of 200 (mg/100 mL) and a standard deviation of 20 (mg/100 mL), the probability that a person picked at random from this population will have a cholesterol value greater than 240 (mg/100 mL) is

$$\begin{aligned}\Pr(x \geq 240) &= \Pr\left(\frac{x-200}{20} \geq \frac{240-200}{20}\right)\\ &= \Pr(z \geq 2.0)\\ &= 0.5 - \Pr(z \leq 2.0)\\ &= 0.5 - 0.4772\\ &= 0.0228 \quad \text{or} \quad 2.28\%\end{aligned}$$

Example 3.8

- Figure 3.11 is a model for hypertension and hypotension, presented here as a simple illustration on the use of the normal distribution; acceptance of the model itself is not universal.
- Data from a population of males were collected by age as shown in Table 3.9. From this table, using Appendix B, systolic blood pressure limits for each group can be calculated (Table 3.10).
- For example, the highest healthy limit for the 20–24 age group is obtained as follows:

$$\Pr(x \geq ?) = 0.10$$
$$= \Pr\left(\frac{x - 123.9}{13.74} \geq \frac{? - 123.9}{13.74}\right)$$

and from Example 3.5 we have

$$1.28 = \frac{? - 123.9}{13.74}$$

leading to

$$? = 123.9 + (1.28)(13.74)$$
$$= 141.49$$

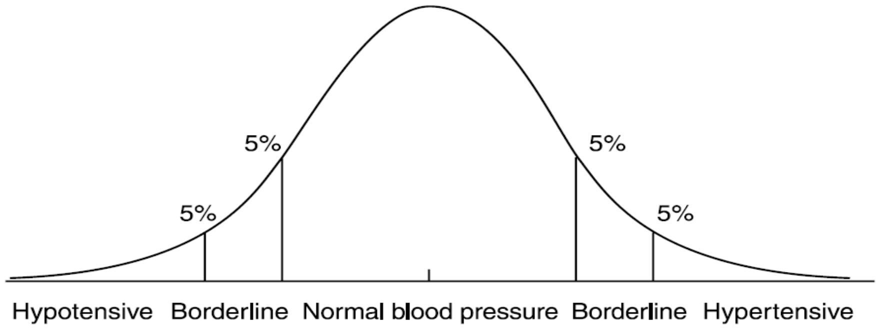
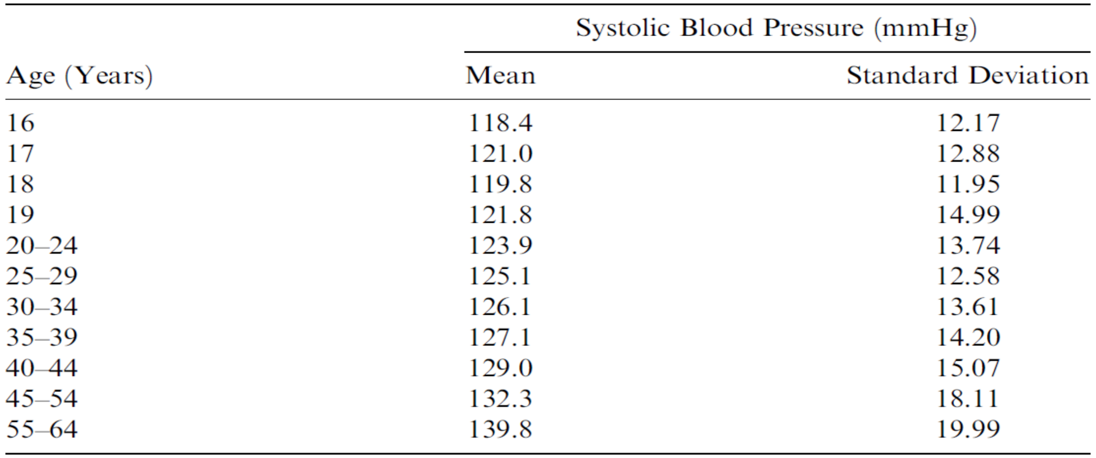


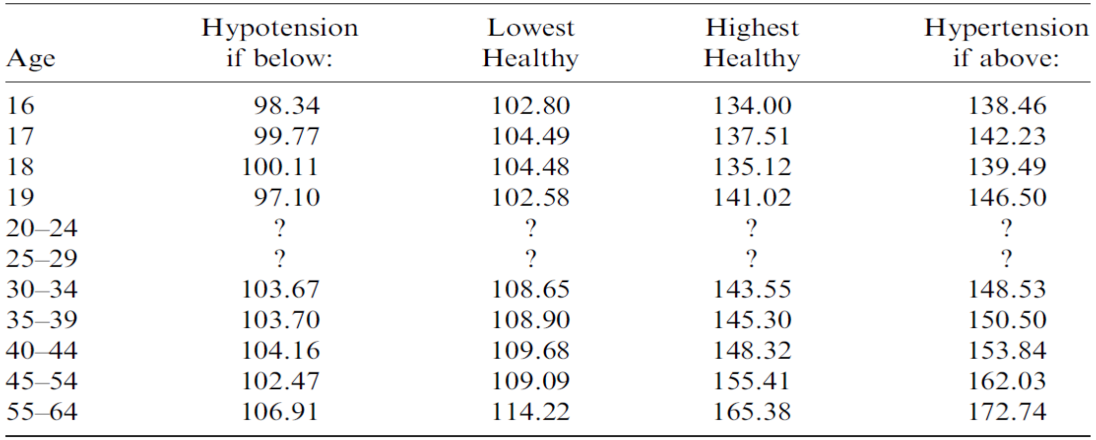
Figure 3.11 Graphical display of a hypertension model.

TABLE 3.9



Age (Years)	Systolic Blood Pressure (mmHg) Mean	Standard Deviation
16	118.4	12.17
17	121.0	12.88
18	119.8	11.95
19	121.8	14.99
20–24	123.9	13.74
25–29	125.1	12.58
30–34	126.1	13.61
35–39	127.1	14.20
40–44	129.0	15.07
45–54	132.3	18.11
55–64	139.8	19.99

TABLE 3.10



Age	Hypotension if below:	Lowest Healthy	Highest Healthy	Hypertension if above:
16	98.34	102.80	134.00	138.46
17	99.77	104.49	137.51	142.23
18	100.11	104.48	135.12	139.49
19	97.10	102.58	141.02	146.50
20–24	?	?	?	?
25–29	?	?	?	?
30–34	103.67	108.65	143.55	148.53
35–39	103.70	108.90	145.30	150.50
40–44	104.16	109.68	148.32	153.84
45–54	102.47	109.09	155.41	162.03
55–64	106.91	114.22	165.38	172.74

Probability Models for Continuous Data

- The normal distribution plays an important role in statistical inference because:

1. Many real-life distributions are approximately normal.
2. Many other distributions can be almost normalized by appropriate data transformations (e.g., taking the log). When $\log X$ has a normal distribution, X is said to have a lognormal distribution.
3. As a sample size increases, the means of samples drawn from a population of any distribution will approach the normal distribution. This theorem, when stated rigorously, is known as the central limit theorem.

- In addition to the normal distribution (Appendix B), topics introduced in subsequent chapters involve three other continuous distributions:
  - The t distribution (Appendix C).
  - The chi-square distribution (Appendix D).
  - The F distribution (Appendix E).

Probability Models for Discrete Data

- **Binomial Distribution**
- Dichotomous outcomes examples: male–female, survived–not survived, infected–not infected, white–nonwhite, or simply positive–negative.
- Such dichotomous data can be summarized into proportions, rates, and ratios.
- In this section we are concerned with the probability of a compound event: the occurrence of X (positive) outcomes ($0 \leq X \leq n$) in n trials, called a binomial probability.

Binomial Distribution

- In general, the binomial model applies when each trial of an experiment has two possible outcomes (often referred to as ‘‘failure’’ and ‘‘success’’ or ‘‘negative’’ and ‘‘positive’’; one has a success when the primary outcome is observed).
- Let the probabilities of failure and success be, respectively, $1-\pi$ and π, and we ‘‘code’’ these two outcomes as 0 (zero successes) and 1 (one success).
- The experiment consists of n repeated trials satisfying these assumptions:
    1. The n trials are all independent.
    2. The parameter π is the same for each trial.
- The model is concerned with the total number of successes in n trials as a random variable, denoted by X. Its probability density function is given by

$$\Pr(X=x)=\binom{n}{x}\pi^x(1-\pi)^{n-x}, \qquad \text{for } x=0,1,2,\dots,n$$

where $\binom{n}{x}=\frac{n!}{x!(n-x)!}$ and $n!$ is the product of the first n integers.

- The mean and variance of the binomial distribution are $\mu = n\pi$ and $\sigma^2 = n\pi(1-\pi)$

Example

- If a certain drug is known to cause a side effect 10% of the time and if five patients are given this drug:

1. What is the probability that none of them will experience the side effect?

2. What is the probability that four or more experience the side effect?

3. What is the probability that at most five will experience the side effect?

Approximate the Binomial Distribution by a Normal Distribution

- When the number of trials n is from moderate to large ($n \geq 25$, say), we approximate the binomial distribution by a normal distribution and answer probability questions by first converting to a standard normal score:

$$z = \frac{x - n\pi}{\sqrt{n\pi(1-\pi)}}$$

where π is the probability of having a positive outcome from a single trial.

- For example, for $\pi = 0.1$ and $n = 30$, we have

$$\mu = (30)(0.1) = 3$$
$$\sigma^2 = (30)(0.1)(0.9) = 2.7$$

so that

$$\Pr(x \geq 7) \simeq \Pr\left(z \geq \frac{7-3}{\sqrt{2.7}}\right)$$
$$= \Pr(z \geq 2.43) \ = 0.0075$$

- In other words, if the true probability for having the side effect is 10%, the probability of having seven or more of 30 patients with the side effect is less than 1% ($=\ 0.0075$).

Poisson Distribution

- Poisson distribution has been used extensively in health science to model the distribution of the number of occurrences x of some random event in an interval of time or space, or some volume of matter.
- The Poisson distribution is characterized by its probability density function:

$$\Pr(X = x) = \frac{\theta^x e^{-\theta}}{x!}, \qquad \text{for } x = 0,1,2,\ldots$$

- It turns out, interestingly enough, that for a Poisson distribution the variance is equal to the mean, the parameter θ above.
- We can approximate a Poisson distribution by a normal distribution with mean θ if $\theta \geq 10$.

$$z = \frac{x - \theta}{\sqrt{\theta}}$$

Example

- A hospital administrator has been studying daily emergency admissions over a period of several months and has found that admissions have averaged three per day.
- He or she is then interested in finding the probability that no emergency admissions will occur on a particular day.
- **Solution:**

Infant Mortality Rate as Poisson distribution

- The infant mortality rate (IMR) is defined as

$$\mathrm{IMR} = \frac{d}{N}$$

- for a certain target population during a given year, where d is the number of deaths during the first year of life and N is the total number of live births.
- In the studies of IMRs, N is conventionally assumed to be fixed and d to follow a Poisson distribution.

Example 3.9

- For the year 1981 we have the following data for the New England states

$$d = 1585, \quad N = 164{,}200$$

- For the same year, the national infant mortality rate was 11.9 (per 1000 live births). If we apply the national IMR to the New England states, we would have

$$\theta = (11.9)(164.2) \simeq 1954 \text{ infant deaths}$$

- Then the event of having as few as 1585 infant deaths would occur with a probability

$$\Pr(d \leq 1585) = \Pr\left(z \leq \frac{1585 - 1954}{\sqrt{1954}}\right) = \Pr(z \leq -8.35) \simeq 0$$

- The conclusion: Either we observed an extremely improbable event, or infant mortality in the New England states is lower than the national average.
- The rate observed for the New England states was 9.7 deaths per 1000 live births.