آمار مقدماتی

امیرحسین زارع مهذبیه

تابستان ۰۰

اهداف

- شانس در زیست‌شناسی
- انواع داده در علوم زیستی
- چگونه داده‌گان را در دو کلام خلاصه کنیم
- استخراج دانش از انبوهی داده

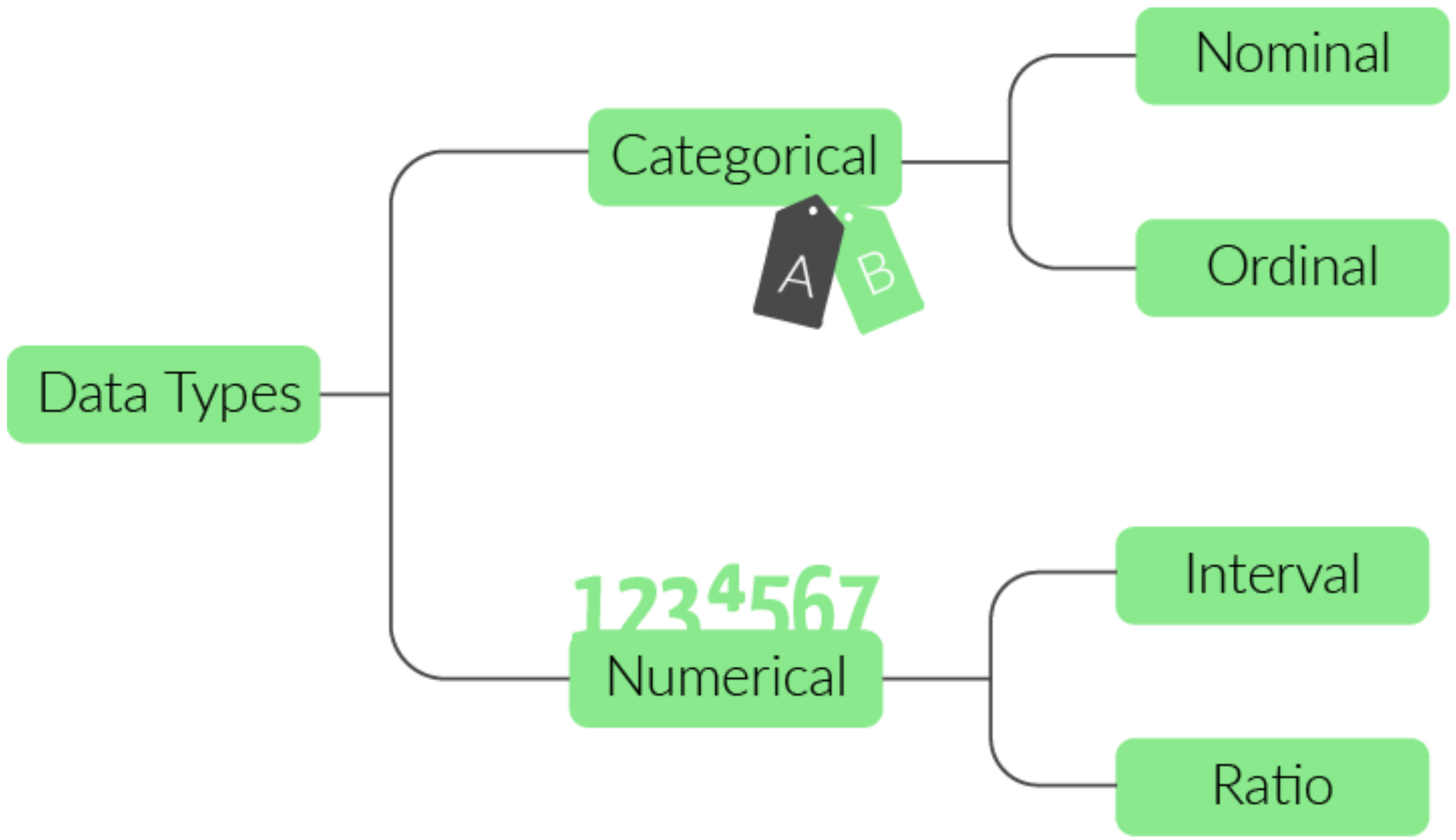
فهرست

۱. مقدمه

…uence	tag	molecular_weight	gravy	aromaticity	instability	isoelectric	helix	turn	sheet	charge	charge_den	aliphatic	boman	hydrophobic	moment_max	moment_mean	frac_a	frac_c	frac_d	frac_e	frac_f	frac_g	frac_h	frac_i	frac_k	frac_l	frac_m	frac_n	frac_p	frac_q	frac_r	frac_s	frac_t	frac_v	frac_w	frac_y
AAAAAAAAGIGKFLHSAKKFGKAPVGEIMNS	train_AMP	3177.676200000000	0.6030303030303030	0.09090909090909090	2.9575757575757560	10.001705741882300	0.2121212121212120	0.2121212121212120	0.4545454545454550	3.096	0.0009739779525378530	80.60606060606060	-0.2563636363636370	0.6060606060606060	0.3222176686166720	0.3222176686166720	0.36363636363636400	0.0	0.0	0.0303030303030300	0.09090909090909090	0.1212121212121210	0.0303030303030300	0.0606060606060606	0.1212121212121210	0.0303030303030300	0.0303030303030300	0.0303030303030300	0.0	0.0	0.0	0.0606060606060606	0.0	0.0303030303030300	0.0	0.0
AAAAAIKMLMDLVNERIMALNKKAKK	train_AMP	2985.6772	0.3035714285714290	0.0	31.117857142857100	10.17177333831790	0.2142857142857140	0.0714285714285714	0.5714285714285710	3.997	0.0013387235068728100	112.14285714285700	0.6964285714285710	0.6428571428571430	0.315482704864734	0.315482704864734	0.3214285714285710	0.0	0.0357142857142857	0.0357142857142857	0.0	0.0	0.0	0.0714285714285714	0.1785714285714290	0.1071428571428570	0.1071428571428570	0.0714285714285714	0.0	0.0	0.0357142857142857	0.0	0.0	0.0357142857142857	0.0	0.0

۲. انواع داده‌گان

۲. انواع داده



Data Types
Categorical
A
B
Nominal
Ordinal
1234567
Numerical
Interval
Ratio

۳. آمار توصیفی

۳. آمار توصیفی

۳.۱ آماره‌های مرکزیت

۳.۲ آماره‌های پراکندگی

۳.۳ آماره‌های شکل

۳.۴ نمودارها

۳.۱. آماره‌های مرکزیت

میانگین

$$\mu = \sum_{i}^{N} x_i \frac{1}{N}$$

۳.۱. آماره‌های مرکزیت

میانگین

DEFINITION

The **sample mean** $\bar{x}$ of observations $x_1, x_2, \ldots, x_n$ is given by

$$\bar{x} = \frac{x_1 + x_2 + \cdots + x_n}{n} = \frac{\sum_{i=1}^{n} x_i}{n}$$

The numerator of $\bar{x}$ can be written more informally as $\sum x_i$ where the summation is over all sample observations.

۳.۱. آماره‌های مرکزیت

میانگین

Example 1.11 A class was assigned to make wingspan measurements at home. The wingspan is the horizontal measurement from fingertip to fingertip with outstretched arms. Here are the measurements given by 21 of the students.

$x_1 = 60$	$x_2 = 64$	$x_3 = 72$	$x_4 = 63$	$x_5 = 66$	$x_6 = 62$	$x_7 = 75$
$x_8 = 66$	$x_9 = 59$	$x_{10} = 75$	$x_{11} = 69$	$x_{12} = 62$	$x_{13} = 63$	$x_{14} = 61$
$x_{15} = 65$	$x_{16} = 67$	$x_{17} = 65$	$x_{18} = 69$	$x_{19} = 95$	$x_{20} = 60$	$x_{21} = 70$

Figure 1.12 shows a stem-and-leaf display of the data; a wingspan in the 60's appears to be "typical."

۳.۱. آماره‌های مرکزیت

میانگین

Example 1.11 A class was assigned to make wingspan measurements at home. The wingspan is the horizontal measurement from fingertip to fingertip with outstretched arms. Here are the measurements given by 21 of the students.

$x_1 = 60$	$x_2 = 64$	$x_3 = 72$	$x_4 = 63$	$x_5 = 66$	$x_6 = 62$	$x_7 = 75$
$x_8 = 66$	$x_9 = 59$	$x_{10} = 75$	$x_{11} = 69$	$x_{12} = 62$	$x_{13} = 63$	$x_{14} = 61$
$x_{15} = 65$	$x_{16} = 67$	$x_{17} = 65$	$x_{18} = 69$	$x_{19} = 95$	$x_{20} = 60$	$x_{21} = 70$

۳.۱. آماره‌های مرکزیت

میانگین

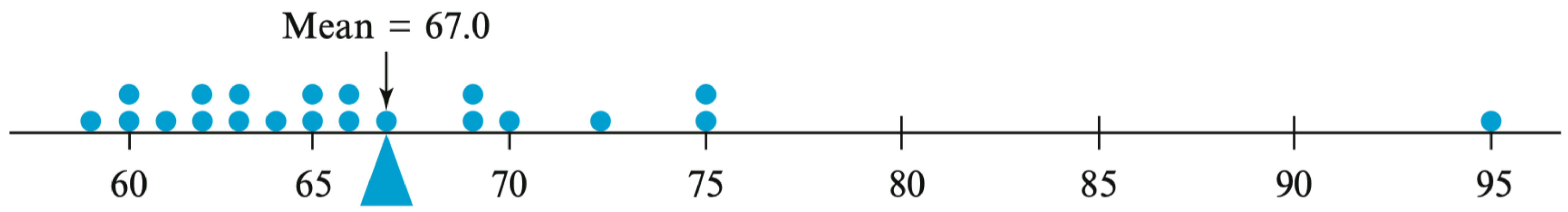


Figure 1.13 The mean as the balance point for a system of weights

۳.۱. آماره‌های مرکزیت

میانه

DEFINITION The **sample median** is obtained by first ordering the n observations from smallest to largest (with any repeated values included so that every sample observation appears in the ordered list). Then,

$$\tilde{x} = \begin{cases} \text{The single middle value if } n \text{ is odd} & = \left(\dfrac{n+1}{2}\right)^{\text{th}} \text{ordered value} \\ \text{The average of the two middle values if } n \text{ is even} & = \text{average of } \left(\dfrac{n}{2}\right)^{\text{th}} \text{ and } \left(\dfrac{n}{2}+1\right)^{\text{th}} \text{ ordered values} \end{cases}$$

۳.۱. آماره‌های مرکزیت

میانه

DEFINITION The **sample median** is obtained by first ordering the n observations from smallest to largest (with any repeated values included so that every sample observation appears in the ordered list). Then,

$$\tilde{x} = \begin{cases} \text{The single middle value if } n \text{ is odd} & = \left(\dfrac{n+1}{2}\right)^{\text{th}} \text{ordered value} \\ \text{The average of the two middle values if } n \text{ is even} & = \text{average of } \left(\dfrac{n}{2}\right)^{\text{th}} \text{and } \left(\dfrac{n}{2}+1\right)^{\text{th}} \text{ordered values} \end{cases}$$

۳.۱. آماره‌های مرکزیت

میانه

Example 1.12 People not familiar with classical music might tend to believe that a composer's instructions for playing a particular piece are so specific that the duration would not depend at all on the performer(s). However, there is typically plenty of room for interpretation, and orchestral conductors and musicians take full advantage of this. We went to the website ArkivMusic.com and selected a sample of 12 recordings of Beethoven's Symphony #9 (the "Choral", a stunningly beautiful work), yielding the following durations (min) listed in increasing order:

62.3 62.8 63.6 65.2 65.7 66.4 67.4 68.4 68.8 70.8 75.7 79.0

۳.۲. آماره‌های پراکندگی

واریانس

$$\sigma^2 = \sum_{i}^{N} (x_i - \mu)^2 \, \frac{1}{N}$$

۳.۲. آماره‌های پراکندگی

واریانس

DEFINITION

The **sample variance**, denoted by s^2, is given by

$$s^2 = \frac{\sum (x_i - \bar{x})^2}{n-1} = \frac{S_{xx}}{n-1}$$

The **sample standard deviation**, denoted by s, is the (positive) square root of the variance:

$$s = \sqrt{s^2}$$

۳.۲. آماره‌های پراکندگی

واریانس

DEFINITION

The **sample variance**, denoted by s^2, is given by

$$s^2 = \frac{\sum (x_i - \bar{x})^2}{n-1} = \frac{S_{xx}}{n-1}$$

The **sample standard deviation**, denoted by s, is the (positive) square root of the variance:

$$s = \sqrt{s^2}$$

۳.۲. آماره‌های پراکندگی

واریانس

Table 1.3 Data for Example 1.14

	x_i	$x_i - \bar{x}$	$(x_i - \bar{x})^2$	
1	27.3	−5.96	35.522	
2	27.9	−5.36	28.730	
3	32.9	−0.36	0.130	
4	35.2	1.94	3.764	
5	44.9	11.64	135.490	
6	39.9	6.64	44.090	
7	30.0	−3.26	10.628	
8	29.7	−3.56	12.674	
9	28.5	−4.76	22.658	
10	32.0	−1.26	1.588	
11	37.6	4.34	18.836	
	$\sum x_i = 365.9$	$\sum (x_i - \bar{x}) = .04$	$\sum (x_i - \bar{x})^2 = 314.110$	$\bar{x} = 33.26$

■

۳.۲. آماره‌های پراکندگی

انحراف معیار

$$\sigma = \sqrt{\sigma^2}$$

$$s = \sqrt{s^2}$$

۳.۲. آماره‌های پراکندگی

خطای استاندارد

$$SEM = \frac{s}{\sqrt{n}}$$

۳.۲. آماره‌های پراکندگی

ضریب تغییرات

$$CV = \frac{s}{\bar{x}}$$

۳.۳. آماره‌های شکل

چولگی

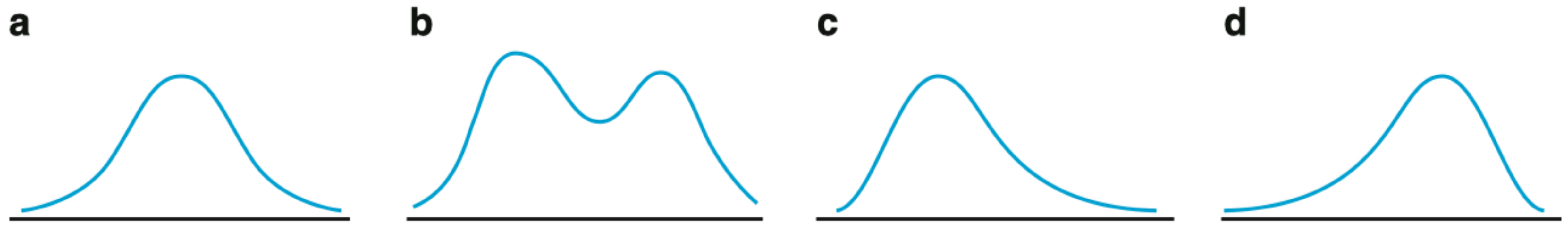


Figure 1.10 Smoothed histograms: (a) symmetric unimodal; (b) bimodal; (c) positively skewed; and (d) negatively skewed

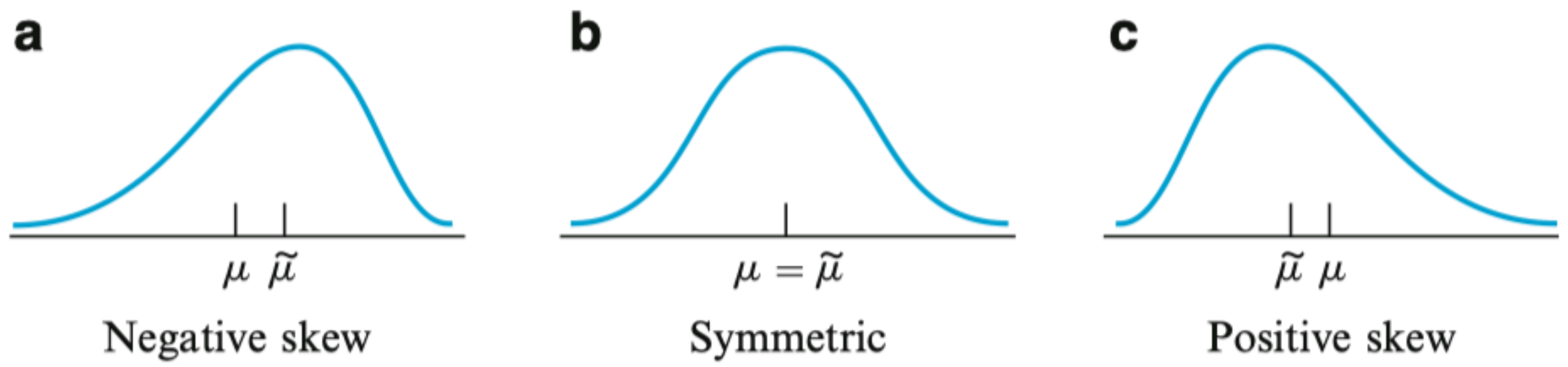


Figure 1.14 Three different shapes for a population distribution

۳.۴. نمودارها

جعبه و سیبیل

40 52 55 60 70 75 85 85 90 90 92 94 94 95 98 100 115 125 125

۳.۴. نمودارها

جعبه و سیبیل

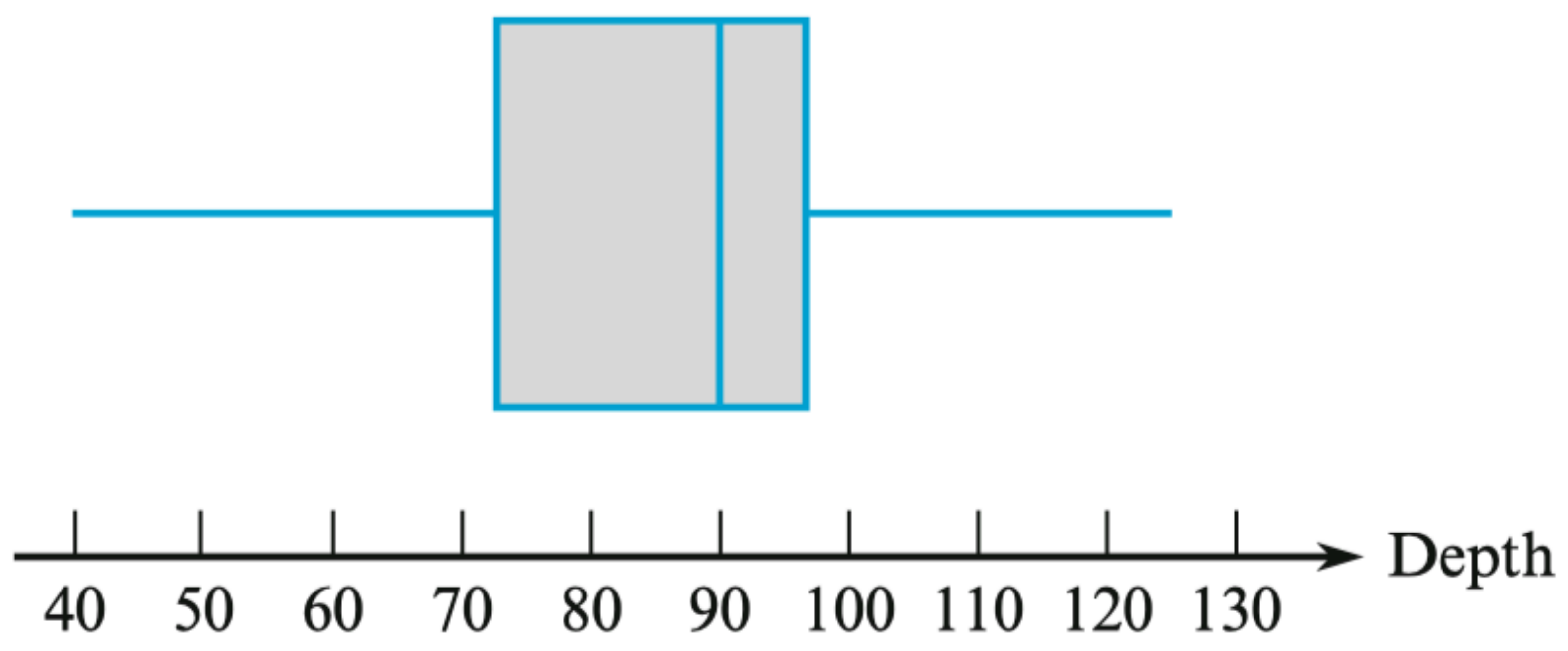


Figure 1.17 A boxplot of the corrosion data

۴. آمار استنباطی

۴. آمار استنباطی

۴.۱ تابع توزیع احتمال

۴.۲ تست آماری

۴.۳ برای تصمیم‌گیری در مورد یک متغیر

۴.۴ برای تصمیم‌گیری در مورد دو متغیر

۴.۱ تابع توزیع احتمال

متغیر تصادفی

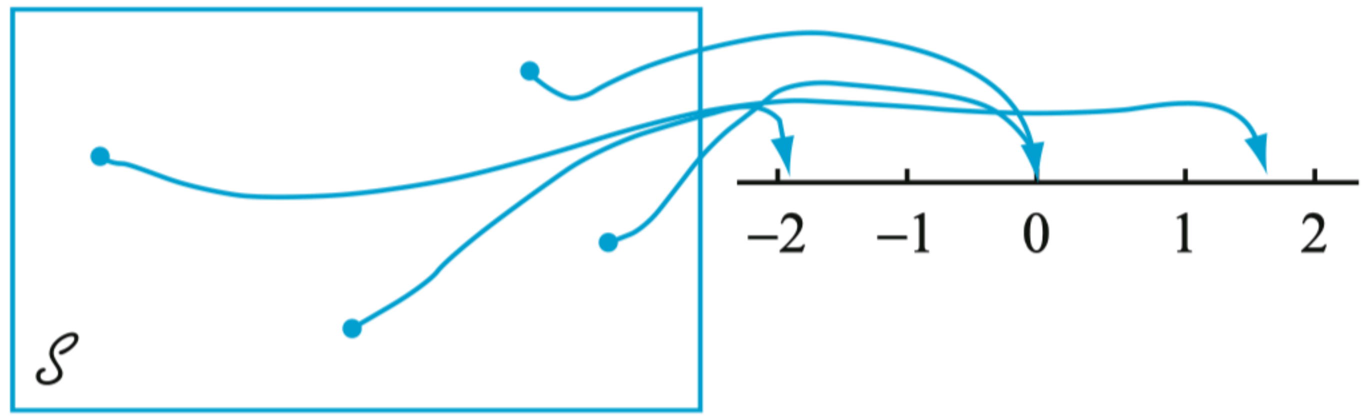


Figure 3.1 A random variable

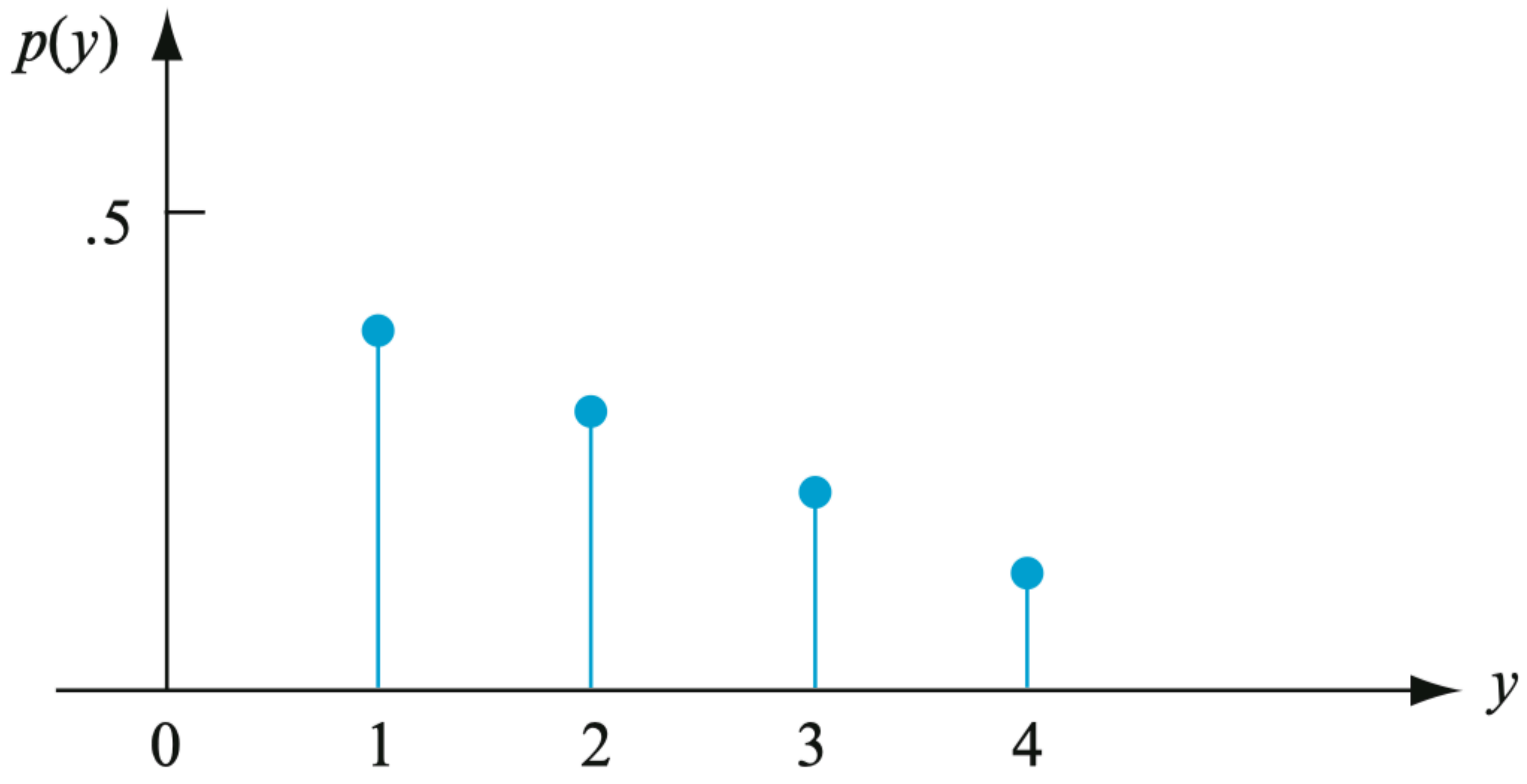
۴.۱ تابع توزیع احتمال

تابع توزیع احتمال

Example 3.9 Consider a group of five potential blood donors—A, B, C, D, and E—of whom only A and B have type O+ blood. Five blood samples, one from each individual, will be typed in random order until an O+ individual is identified. Let the rv Y = the number of typings necessary to identify an O+ individual. Then the pmf of Y is

۴.۱ تابع توزیع احتمال

تابع توزیع احتمال



۴.۱ تابع توزیع احتمال

تابع توزیع احتمال

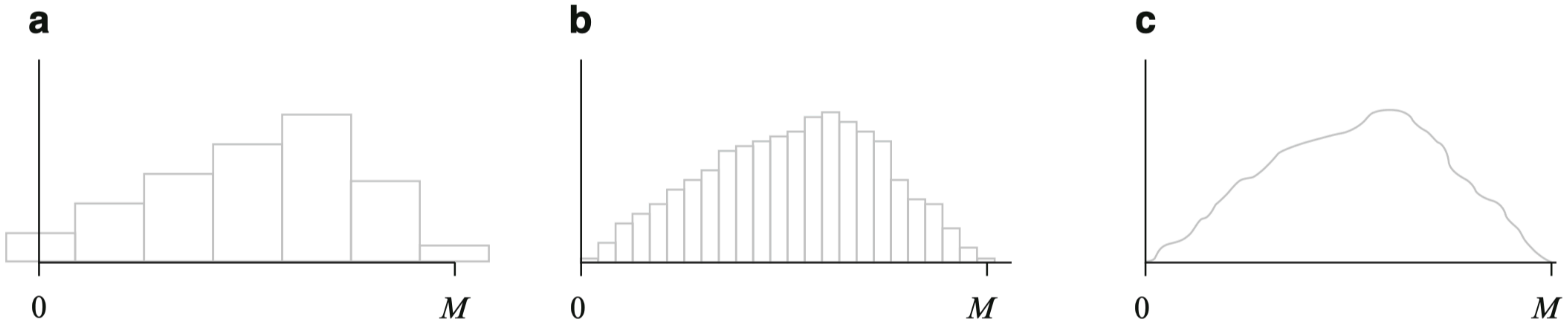


Figure 4.1 (a) Probability histogram of depth measured to the nearest meter; (b) probability histogram of depth measured to the nearest centimeter; (c) a limit of a sequence of discrete histograms

۴.۱ تابع توزیع احتمال

تابع توزیع احتمال

DEFINITION

Let X be a continuous rv. Then a **probability distribution** or **probability density function** (pdf) of X is a function $f(x)$ such that for any two numbers a and b with $a \leq b$,

$$P(a \leq X \leq b) = \int_a^b f(x)dx$$

That is, the probability that X takes on a value in the interval $[a, b]$ is the area above this interval and under the graph of the density function, as illustrated in Figure 4.2. The graph of $f(x)$ is often referred to as the *density curve*.

۴.۱ تابع توزیع احتمال

تابع توزیع احتمال

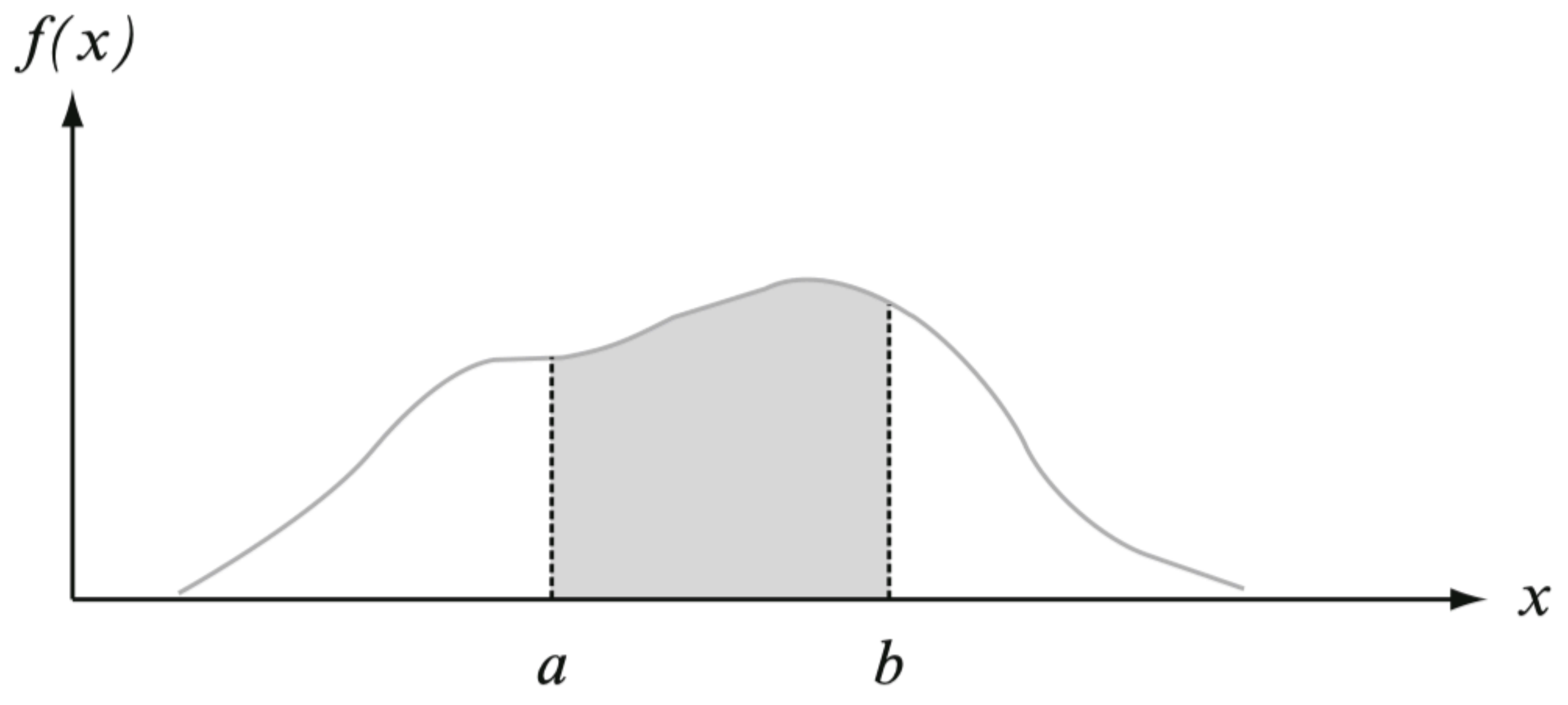


Figure 4.2 $P(a \leq X \leq b)$ = the area under the density curve between a and b

۴.۱ تابع توزیع احتمال

توزیع نرمال

DEFINITION A continuous rv X is said to have a **normal distribution** with parameters μ and σ (or μ and σ^2), where $-\infty < \mu < \infty$ and $0 < \sigma$, if the pdf of X is

$$f(x; \mu, \sigma) = \frac{1}{\sqrt{2\pi}\sigma} e^{-(x-\mu)^2/(2\sigma^2)} \qquad -\infty < x < \infty \tag{4.3}$$

۴.۱ تابع توزیع احتمال

توزیع نرمال

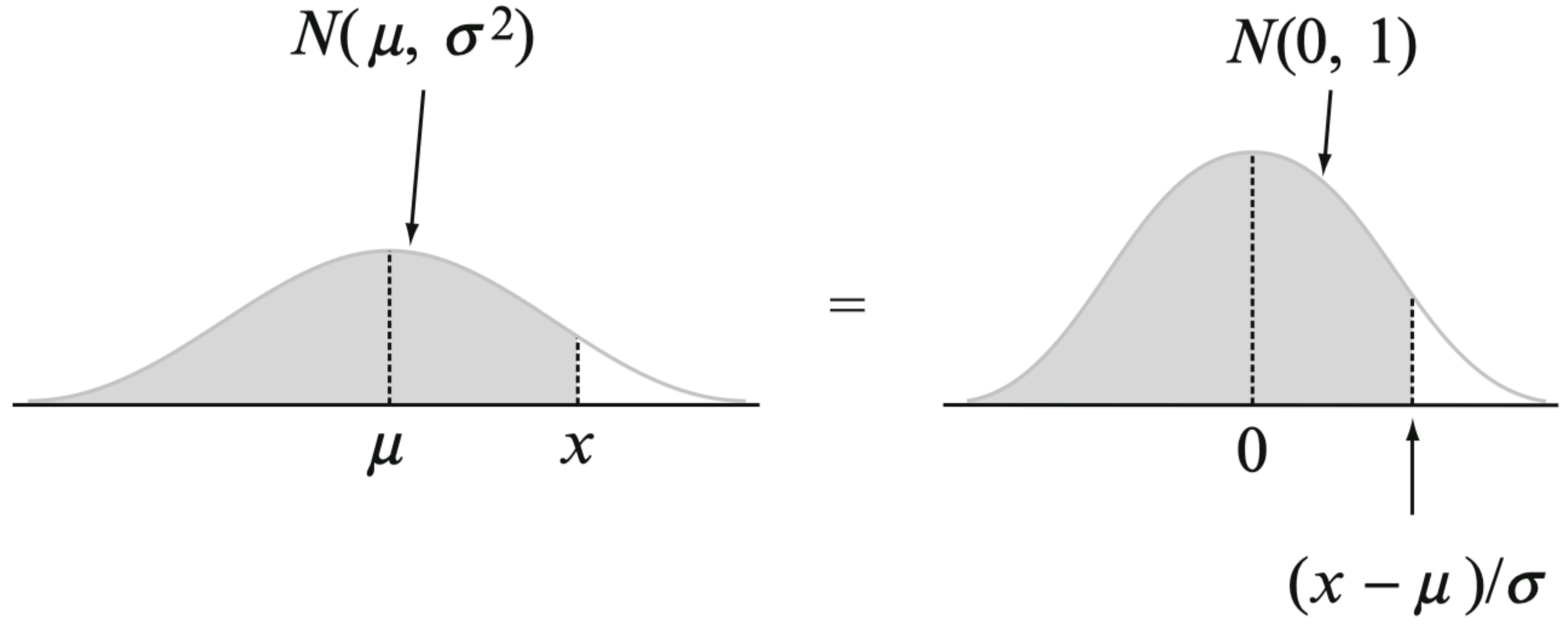


Figure 4.21 Equality of nonstandard and standard normal curve areas

۴.۱ تابع توزیع احتمال

مثال: احتمال مشاهده یک پیشامد در پدیده‌ای با توزیع نرمال

Example 4.22 The time that it takes a driver to react to the brake lights on a decelerating vehicle is critical in avoiding rear-end collisions. The article "Fast-Rise Brake Lamp as a Collision-Prevention Device" (*Ergonomics*, 1993: 391–395) suggests that reaction time for an in-traffic response to a brake signal from standard brake lights can be modeled with a normal distribution having mean value 1.25 s and standard deviation of .46 s. What is the probability that reaction time is between 1.00 and 1.75 s?

۴.۱ تابع توزیع احتمال

راه‌حل: احتمال مشاهده یک پیشامد در پدیده‌ای با توزیع نرمال

If we let X denote reaction time, then standardizing gives

$$1.00 \le X \le 1.75$$

if and only if

$$\frac{1.00 - 1.25}{.46} \le \frac{x - 1.25}{.46} \le \frac{1.75 - 1.25}{.46}$$

Thus

$$\begin{aligned} P(1.00 \le X \le 1.75) &= P\left(\frac{1.00 - 1.25}{.46} \le Z \le \frac{1.75 - 1.25}{.46}\right) \\ &= P(-.54 \le Z \le 1.09) = \Phi(1.09) - \Phi(-.54) \\ &= .8621 - .2946 = .5675 \end{aligned}$$

This is illustrated in Figure 4.22. Similarly, if we view 2 s as a critically long-reaction time, the probability that actual reaction time will exceed this value is

$$P(X > 2) = P\left(Z > \frac{2 - 1.25}{.46}\right) = P(Z > 1.63) = 1 - \Phi(1.63) = .0516$$

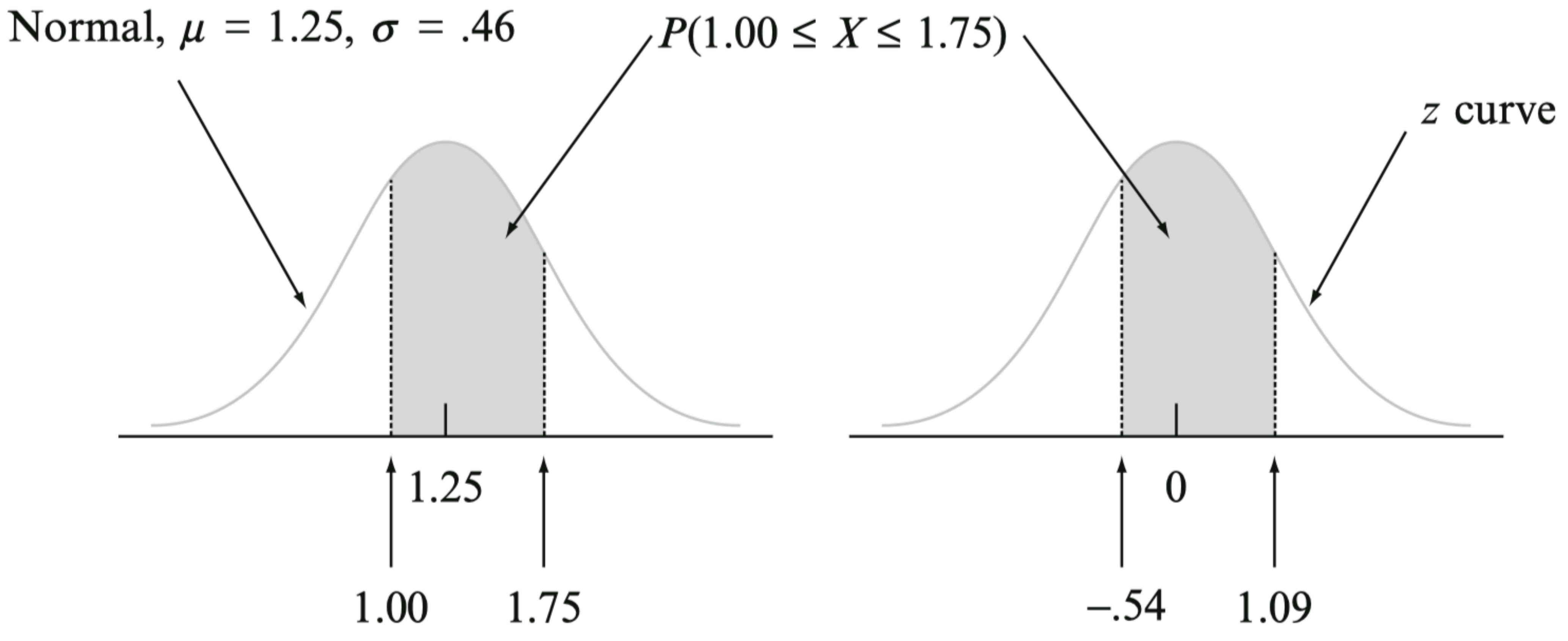


Figure 4.22 Normal curves for Example 4.22 ■

۴.۲ تست آماری

فنجان‌های چای

۴.۲ تست آماری

فرض صفر، فرض جایگزین، پی ولیو

DEFINITION

The **null hypothesis**, denoted by H_0, is the claim that is initially assumed to be true (the "prior belief" claim). The **alternative hypothesis**, denoted by H_a, is the assertion that is contradictory to H_0.

The null hypothesis will be rejected in favor of the alternative hypothesis only if sample evidence suggests that H_0 is false. If the sample does not strongly contradict H_0, we will continue to believe in the plausibility of the null hypothesis. The two possible conclusions from a hypothesis-testing analysis are then *reject* H_0 or *fail to reject* H_0.

۴.۲ تست آماری

در یک نگاه

۴.۲ برای یک متغیر (کمّی)

مقایسه‌ی میانگین نمونه با جمعیتی که واریانس آن را می‌دانیم

$$z = \frac{\bar{x} - \mu}{\sigma/\sqrt{n}} \sim N(0,1)$$

۴.۲ برای یک متغیر (کمّی)

مثال: Z-test

Example 9.6 A manufacturer of sprinkler systems used for fire protection in office buildings claims that the true average system-activation temperature is 130°. A sample of $n = 9$ systems, when tested, yields a sample average activation temperature of 131.08°F. If the distribution of activation times is normal with standard deviation 1.5°F, does the data contradict the manufacturer's claim at significance level $\alpha = .01$?

۴.۲ برای یک متغیر (کمّی)

راه‌حل: Z-test

1. Parameter of interest: μ = true average activation temperature.

2. Null hypothesis: H_0: $\mu = 130$ (null value $= \mu_0 = 130$).

3. Alternative hypothesis: H_a: $\mu \neq 130$ (a departure from the claimed value in *either* direction is of concern).

4. Test statistic value:

$$z = \frac{\bar{x} - \mu_0}{\sigma/\sqrt{n}} = \frac{\bar{x} - 130}{1.5/\sqrt{n}}$$

5. Rejection region: The form of H_a implies use of a two-tailed test with rejection region *either* $z \geq z_{.005}$ *or* $z \leq -z_{.005}$. From Section 4.3 or Appendix Table A.3, $z_{.005} = 2.58$, so we reject H_0 if either $z \geq 2.58$ or $z \leq -2.58$.

6. Substituting $n = 9$ and $\bar{x} = 131.08$,

$$z = \frac{131.08 - 130}{1.5/\sqrt{9}} = \frac{1.08}{.5} = 2.16$$

   That is, the observed sample mean is a bit more than 2 standard deviations above what would have been expected were H_0 true.

7. The computed value $z = 2.16$ does not fall in the rejection region ($-2.58 < 2.16 < 2.58$), so H_0 cannot be rejected at significance level .01. The data does not give strong support to the claim that the true average differs from the design value of 130. ■

۴.۲ برای یک متغیر (کمّی)

مقایسه‌ی میانگین نمونه با جمعیتی که واریانس آن را نـــــــمی‌دانیم

$$t = \frac{\bar{x} - \mu}{s/\sqrt{n}} \sim T(df = n - 1)$$

۴.۲ برای یک متغیر (کمّی)

مثال: One sample T-test

Example 9.9 A well-designed and safe workplace can contribute greatly to increased productivity. It is especially important that workers not be asked to perform tasks, such as lifting, that exceed their capabilities. The accompanying data on maximum weight of lift (MAWL, in kg) for a frequency of four lifts/min was reported in the article "The Effects of Speed, Frequency, and Load on Measured Hand Forces for a Floor-to-Knuckle Lifting Task" (*Ergonomics*, 1992: 833–843); subjects were randomly selected from the population of healthy males age 18–30. Assuming that MAWL is normally distributed, does the following data suggest that the population mean MAWL exceeds 25?

25.8	36.6	26.3	21.8	27.2

۴.۲ برای یک متغیر (کمّی)

راه‌حل: One sample T-test

Let's carry out a test using a significance level of .05.

1. $\mu =$ population mean MAWL
2. H_0: $\mu = 25$
3. H_a: $\mu > 25$
4. $t = \dfrac{\bar{x} - 25}{s/\sqrt{n}}$
5. Reject H_0 if $t \geq t_{\alpha,\, n-1} = t_{.05,4} = 2.132$.
6. $\Sigma x_i = 137.7$ and $\Sigma x_i^2 = 3911.97$, from which $\bar{x} = 27.54$, $s = 5.47$, and

$$t = \frac{27.54 - 25}{5.47/\sqrt{5}} = \frac{2.54}{2.45} = 1.04$$

   The accompanying MINITAB output from a request for a one-sample t test has the same calculated values (the P-value is discussed in Section 9.4).

   Test of mu = 25.00 vs mu > 25.00

Variable	N	Mean	StDev	SE Mean	T	P-Value
mawl	5	27.54	5.47	2.45	1.04	0.18

7. Since 1.04 does not fall in the rejection region ($1.04 < 2.132$), H_0 cannot be rejected at significance level .05. It is still plausible that μ is (at most) 25. ■

۴.۳ برای دو متغیر (کیفی-کمّی)

مقایسه‌ی میانگین دو نمونه با فرض واریانس برابر

$$t = \frac{\bar{x}_1 - \bar{x}_2}{\sqrt{\frac{(n_1 - 1)s_1^2 + (n_2 - 1)s_2^2}{n_1 + n_2 - 2}} \cdot \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}} \sim T\,(df = n_1 + n_2 - 2)$$

۴.۳ برای دو متغیر (کیفی-کمّی)

مقایسه‌ی میانگین دو نمونه با فرض واریانس نـــــــابرابر

$$t = \frac{\bar{x}_1 - \bar{x}_2}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}} \sim T\left(df = \frac{\left(\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}\right)^2}{\frac{\left(s_1^2/n_1\right)^2}{n_1 - 1} + \frac{\left(s_2^2/n_2\right)^2}{n_2 - 1}}\right)$$

۴.۳ برای دو متغیر (کیفی-کمّی)

مثال: Independant sample T-test

Example 10.4 A study was carried out in an attempt to improve student performance in a low-level university mathematics course. Experience had shown that many students had fallen by the wayside, meaning that they had dropped out or completed the course with minimal effort and low grades. The study involved assigning the students to sections based on odd or even Social Security number. It is important that the assignment to sections not be on the basis of student choice, because then the differences in performance might be attributable to differences in student attitude or ability. Half of the sections were taught traditionally, whereas the other half were taught in a way that hopefully would keep the students involved. They were given frequent assignments that were collected and graded, they had frequent quizzes, and they were allowed retakes on exams. Lotus Hershberger conducted the experiment and he supplied the data. Here are the final exam scores for the 79 students taught traditionally (the control group) and for the 85 students taught with more involvement (the experimental group):

Control

37	22	29	29	33	22	32	36	29	06	04	37	00	36	00	32
27	07	19	35	26	22	28	28	32	35	28	33	35	24	21	00
32	28	27	08	30	37	09	33	30	36	28	03	08	31	29	09
00	00	35	25	29	03	33	33	28	32	39	20	32	22	24	20
32	07	08	33	29	09	00	30	26	25	32	38	22	29	29	

Experimental

34	27	26	33	23	37	24	34	22	23	32	05	30	35	28	25
37	28	26	29	22	33	31	23	37	29	00	30	34	26	28	27
32	29	31	33	28	21	34	29	33	06	08	29	36	07	21	30
28	34	28	35	30	34	09	38	09	27	25	33	09	23	32	25
37	28	23	26	34	32	34	00	24	30	36	28	38	35	16	37
25	34	38	34	31											

۴.۳ برای دو متغیر (کیفی-کمّی)

راه‌حل: Independant sample T-test

Table 10.1 Summary results for Example 10.4

Group	Sample Size	Sample Mean	Sample SD
Control	79	23.87	11.60
Experimental	85	27.34	8.85

Let μ_1 and μ_2 denote the true mean scores for the control condition and the experimental condition, respectively. The two hypotheses are H_0: $\mu_1 - \mu_2 = 0$ versus H_a: $\mu_1 - \mu_2 < 0$. H_0 will be rejected if $z \le -z_{.05} = -1.645$. Then

$$z = \frac{23.87 - 27.34}{\sqrt{\frac{11.60^2}{79} + \frac{8.85^2}{85}}} = \frac{-3.47}{1.620} = -2.14$$

Since $-2.14 \le -1.645$, H_0 is rejected at significance level .05. Alternatively, the P-value for a lower-tailed z test is

$$P\text{-value} = \Phi(z) = \Phi(-2.14) = .016$$

which implies rejection at significance level .05. Also, if the test had been two-tailed, then the P-value would be $2(.016) = .032$, so the two-tailed test would reject H_0 at the .05 level.

۴.۳ برای دو (ولی یک) متغیر (کیفی)

مقایسه‌ی فراوانی یک نمونه با یک توزیع خاص

$$\chi^2 = \sum_{i=1}^{k} \frac{(observed - expected)^2}{expected} \sim \chi^2 \; (df = k - 1)$$

۴.۳ برای دو (ولی یک) متغیر (کیفی)

مثال: تست مربع کای برای نیکویی برازش

Example 13.1 If we focus on two different characteristics of an organism, each controlled by a single gene, and cross a pure strain having genotype AABB with a pure strain having genotype aabb (capital letters denoting dominant alleles and small letters recessive alleles), the resulting genotype will be AaBb. If these first-generation organisms are then crossed among themselves (a dihybrid cross), there will be four phenotypes depending on whether a dominant allele of either type is present. Mendel's laws of inheritance imply that these four phenotypes should have probabilities 9/16, 3/16, 3/16, and 1/16 of arising in any given dihybrid cross.

The article "Linkage Studies of the Tomato" (*Trans. Royal Canad. Institut.*, 1931: 1–19) reports the following data on phenotypes from a dihybrid cross of tall cut-leaf tomatoes with dwarf potato-leaf tomatoes. There are $k = 4$ categories corresponding to the four possible phenotypes, with the null hypothesis being

$$H_0 : p_1 = \frac{9}{16},\ p_2 = \frac{3}{16},\ p_3 = \frac{3}{16},\ p_4 = \frac{1}{16}$$

The expected cell counts are $9n/16$, $3n/16$, $3n/16$, and $n/16$, and the test is based on $k - 1 = 3$ df. The total sample size was $n = 1611$. Observed and expected counts are given in Table 13.2.

Table 13.2 Observed and expected cell counts for Example 13.1

	$i = 1$ **Tall, Cut-Leaf**	$i = 2$ **Tall, Potato-Leaf**	$i = 3$ **Dwarf, Cut-Leaf**	$i = 4$ **Dwarf, Potato-Leaf**
n_i	926	288	293	104
np_{i0}	906.2	302.1	302.1	100.7

۴.۳ برای دو (ولی یک متغیر (کیفی)

راه‌حل: تست مربع کای برای نیکویی برازش

The contribution to χ^2 from the first cell is

$$\frac{(n_1 - np_{10})^2}{np_{10}} = \frac{(926 - 906.2)^2}{906.2} = .433$$

Cells 2, 3, and 4 contribute .658, .274, and .108, respectively, so $\chi^2 = .433 + .658 + .274 + .108 = 1.473$. A test with significance level .10 requires $\chi^2_{.10,3}$, the number in the 3 df row and .10 column of Appendix Table A.6. This critical value is 6.251. Since 1.473 is not at least 6.251, H_0 cannot be rejected even at this rather large level of significance. The data is quite consistent with Mendel's laws. ■

۴.۳ برای دو متغیر (کیفی-کیفی)

بررسی وابستگی دو متغیر

$$\chi^2 = \sum_{i=1}^{I} \sum_{j=1}^{J} \frac{(observed - expected)^2}{expected} \sim \chi^2 \ (df = (I-1)(J-1))$$

۴.۳ برای دو متغیر (کیفی-کیفی)

مثال: تست مربع کای برای بررسی وابستگی

Example 13.14 A study of the relationship between facility conditions at gasoline stations and aggressiveness in the pricing of gasoline ("An Analysis of Price Aggressiveness in Gasoline Marketing," *J. Market. Res.*, 1970: 36–42) reports the accompanying data based on a sample of $n = 441$ stations. At level .01, does the data suggest that facility conditions and pricing policy are independent of one another? Observed and estimated expected counts are given in Table 13.10.

Table 13.10 Observed and estimated expected counts for Example 13.14

		Observed Pricing Policy				Expected Pricing Policy			
		Aggressive	Neutral	Nonaggressive	$n_{i.}$				
	Substandard	24	15	17	56	17.02	22.10	16.89	56
Condition	**Standard**	52	73	80	205	62.29	80.88	61.83	205
	Modern	58	86	36	180	54.69	71.02	54.29	180
	$n_{.j}$	134	174	133	441	134	174	133	441

۴.۳ برای دو متغیر (کیفی-کیفی)

راه‌حل: تست مربع کای برای بررسی وابستگی

Thus

$$\chi^2 = \frac{(24 - 17.02)^2}{17.02} + \cdots + \frac{(36 - 54.29)^2}{54.29} = 22.47$$

and because $\chi^2_{.01,4} = 13.277$, the hypothesis of independence is rejected.

We conclude that knowledge of a station's pricing policy does give information about the condition of facilities at the station. In particular, stations with an aggressive pricing policy appear more likely to have substandard facilities than stations with a neutral or nonaggressive policy. ■