CHƯƠNG 2

MÔ TẢ CÁC TẬP DỮ LIỆU

Về chương này:

Đôi lúc dữ liệu chúng ta đã thu thập thể hiện một mẫu được chọn từ một tổng thể. Những lúc khác (chẳng hạn như một cuộc tổng điều tra dân số toàn quốc), dữ liệu có thể thể hiện toàn bộ tổng thể. Trong cả hai trường hợp, chúng ta đều cần phải có khả năng mô tả tập dữ liệu đó. Mục tiêu của chương này là trình bày hai loại phương pháp mô tả các tập dữ liệu: (1) các phương pháp mô tả bằng đồ thị và (2) các phương pháp mô tả bằng số. Phương pháp mô tả bằng đồ thị mô tả dữ liệu bằng cách sử dụng biểu đồ và đồ thị. Phương pháp mô tả bằng số sử dụng các con số để giúp chúng ta xây dựng một hình ảnh trong trí óc về dữ liệu.

NGHIÊN CỨU TÌNH HUỐNG

VẬY LÀ ANH/CHỊ MUỐN TRỞ THÀNH NHÀ TRIỆU PHÚ?

Vào thập niên 1980, các luật thuế mới đã dẫn đến việc tạo ra Tài khoản Hưu trí Cá nhân (Individual Retirement Accounts-IRA), đó là những tài khoản tiết kiệm miễn thuế đối với thu nhập hưu trí. Dựa theo nhiều mẫu quảng cáo trên báo chí vào lúc đó, nếu mà chúng ta đầu tư 2.000USD mỗi năm vào một Tài khoản Hưu trí Cá nhân (IRA), thì sau 40 năm tham gia, tiền dự trữ của chúng ta sẽ tăng lên đến trên một triệu đô la. Tất nhiên là kể từ đó, các luật thuế đã được thay đổi vài lần, và với việc xem xét lại thuế gần đây nhất, IRA miễn thuế sẽ không còn có sẵn cho hầu hết chúng ta. Dù vậy, cái nguyên tắc được thể hiện trong những mẩu quảng cáo đó vẫn còn có giá trị. Cách thức tốt nhất để tích lũy một số tiền lớn là tham gia vào một chương trình tiết kiệm và đầu tư có hệ thống và tính lãi kép những số tiền đầu tư qua nhiều năm.

Nếu anh/chị đang để dành tiền cho thời kỳ hưu trí hay nhằm mục đích nào khác, thì số tiền anh/chị tích lũy được sẽ phụ thuộc vào số tiền anh/chị đầu tư mỗi năm, nơi anh/chị đầu tư (tài khoản tiết kiệm tại ngân hàng, quỹ đầu tư thị trường vốn ngắn hạn, hay một trong những quỹ cổ phiếu thường khác nhau), và ai quản lý tài khoản của anh/chị. Về cơ bản, mức tăng trưởng của tài khoản của anh/chị và giá trị cuối cùng của nó sẽ phụ thuộc vào suất sinh lợi hàng năm mà nhà quản lý tài khoản của anh/chị có thể thu nhận được cho anh/chị.

Mặc dù suất sinh lợi từ tiền đầu tư của anh/chị sẽ thay đổi từ ngày này sang ngày khác, nhưng Bảng 2.1 cho anh/chị biết số tiền mình có thể kỳ vọng sẽ tích lũy được sau 40 năm. Những số tiền được trình bày trong bảng này dựa vào giả định rằng anh/chị đầu tư 2.000USD vào đầu mỗi năm trong thời kỳ 40 năm và tiền này được tính kép hàng tháng với lãi suất hàng năm cố định là *I*.

BẢNG **2.1** Số tiền rút ra của một tài khoản sau khi thực hiện đầu tư hàng năm 2.000USD với suất sinh lợi hàng năm cố định *I* (%) trong 40 năm

Lãi suất *I* (%)	**Số tiền trong Tài khoản Sau 40 Năm ($)**
4	197.652
6	328.095
8	559.562
10	973.704
12	1.718.285

Suất sinh lợi hàng tháng thay đổi không chỉ giữa các công cụ đầu tư (ngân hàng, quỹ cổ phiếu thường, quỹ đầu tư thị trường vốn ngắn hạn, v.v.) mà còn trong phạm vi bản thân một công cụ đầu tư. Thí dụ, hãy xét một quỹ đầu tư thị trường vốn ngắn hạn, mà nó đầu tư vào kỳ phiếu thương mại ngắn hạn. Bởi vì những kỳ phiếu này sẽ được thương lượng vào những thời điểm khác nhau và sẽ có khoảng thời gian đến khi đáo hạn khác nhau, nên suất sinh lợi đối với toàn bộ quỹ này tại bất kỳ thời điểm nào đều sẽ phụ thuộc vào kỷ năng của nhà quản lý quỹ trong việc cho vay. Nếu lãi suất tăng trong tương lai, thì nắm giữ các kỳ phiếu có thời gian đến khi đáo hạn trung bình ngắn sẽ có lợi. Nếu lãi suất giảm, thì nắm giữ các kỳ phiếu có thời gian đến khi đáo hạn trung bình dài sẽ có lợi.

BẢNG **2.2** Dữ liệu về 604 quỹ đầu tư thị trường vốn ngắn hạn trong thời kỳ kết thúc vào ngày 13/7/1994†

Quỹ	Thời gian Đáo hạn Trung bình	Lợi suất 7 ngày	Tài sản	Quỹ	Thời gian Đáo hạn Trung bình	Lợi suất 7 ngày	Tài sản	Quỹ	Thời gian Đáo hạn Trung bình	Lợi suất 7 ngày	Tài sản	Quỹ	Thời gian Đáo hạn Trung bình	Lợi suất 7 ngày	Tài sản
AALMny	60	2.91	67	CardGvt	3	3.42	397	Fid FDIT	37	3.96	2090	HanvGov	33	3.87	915
AIM MM C	23	3.46	340	Carnegie	23	3.27	18	FidGvRes	45	3.78	1137	HanvUSTr	42	3.81	1271
AIM MMA	23	3.46	122	CascdCs	21	3.29	100	FidGvPr	47	3.95	159	HanvTreas	41	3.56	952
AVESTA Tr	54	3.82	23	CshActMM	32	3.50	147	FidUS Tr	28	3.84	170	Harbor	68	3.83	67
AccUSGov	28	3.72	19	CshAcctGv	23	3.50	36	FidRetGov	41	3.92	1635	HrtgCsh	32	3.71	964
ActAsGv	41	3.55	495	CashEqv	36	3.65	3408	FidRetMM	41	4.12	2871	HiMrkDv	36	3.51	304
ActAsMny	54	3.89	4392	CshEqGv	27	3.51	1564	FidSpGov	44	3.87	761	HiMrkUS	28	3.43	188
AetnaAdvs	38	4.23	31	CashMgt	21	3.71	2799	FidSpMM	47	4.08	70100	HiMrkUST	33	3.27	201
Aetna Sel	38	4.23	150	CshTrGv	31	3.39	442	FidSpUSTr	70	3.81	1552	HmestdDly	48	3.63	32
AlexBwn	49	3.90	1296	CshTrPr	35	3.46	800	59WallStTreas	56	3.57	143	HorznPr	10	4.04	2812
AlxBTr	46	3.60	551	CshTrTreas	29	3.21	400	59WallStMM	56	3.67	591	HorznTr	45	4.15	2137
AlgerMM	42	4.18	148	CshTrII	29	3.38	235	FnclRsv	54	3.96	387	Hummer	30	3.59	146
AlliaCpRs	51	3.31	2474	CentnGv	35	3.85	640	FstAmerInstit	32	4.04	1082	IAATrMM	27	3.89	37
AliaGvR	49	3.22	2089	Centen	42	3.81	2670	FtBostin	25	4.16	150	IAIMnyMktFd	27	3.63	22
AlliMny	52	3.27	1788	ChchCsh	29	3.80	191	First Muni	38	2.32	79	IDS CshM	30	3.55	1174
AmAAdTrl	1	4.03	78	ColonlMMB	24	2.53	60	FstOmahaGv	56	3.55	78	IDS PLA	21	4.04	29
AmAAdMMI	30	4.25	1280	ColonlMM A	24	3.49	94	FirPrTreasTr	54	3.77	97	IMGLiq	36	3.52	138
AmCRes	17	3.47	438	ColDln f	46	3.77	652	First USGv	51	4.27	71	IndCaGv	41	3.37	246
AmPerCsh	35	3.82	172	ComBkrUSGv	27	4.39	121	FtBost	47	4.29	2149	IndCaMM	41	3.36	313
AmPerTrs	1	3.46	137	ComSens	17	3.34	56	FtInvCs	20	3.70	120	IndOnPr	19	3.75	286
AmSouth Pr	44	3.74	578	CompCs	52	3.93	439	FtPraGv	23	4.05	157	IndOnUS	28	3.60	247
AmSouth US	39	3.51	294	CompUS	48	3.66	279	FtPraMM	34	3.62	146	InfnAlGv	32	3.87	73
AmbMMF	32	3.84	323	CmpCshMM A	40	3.35	121	First USTrs	47	3.81	76	InfnCCR Inst	61	4.07	63
AmbTreas F	46	3.60	208	CnstgCs	13	3.72	124	CalvFtGv	31	3.58	255	InfCCR	61	3.41	333
AmbTreasI	46	3.45	73	CnstgUS	28	3.44	307	FlexFd	51	4.12	160	InstCsh	31	4.49	417
AmbMMI	32	3.69	517	CG CapGov	82	3.68	194	Fortis	20	3.52	109	InstFd	21	3.30	11
Amcore Gv	15	3.67	115	CoreFd	33	3.93	518	ForumDATrs	23	3.74	29	InstGov	53	4.17	195
AmAAdMMM	30	4.00	49	CoreTr a	35	3.72	432	Founders	42	3.63	241	InvCshGv	26	4.00	125
ArchUSTr	44	3.33	2	CortldGn	33	3.21	903	FountSGv	41	3.89	123	InvCshR	37	3.96	779
ArchFd	40	3.66	43	CortldUS	56	3.22	212	FountSCP	28	3.87	221	InvGvtMF	29	3.62	75
ArkMnyMkt	40	4.24	205	CowenStdResv	24	3.72	699	FountSTO	31	3.78	330	Invesco	11	4.13	131
ArkUSGovt	22	4.08	430	CrestCshTr	41	3.48	411	FrnklFT	41	4.07	216	IvyMny	40	3.10	25
ArkUSTrsy	56	3.67	142	CrestUST Tr	49	3.42	311	FrnkLate	4	4.16	53	JPM InstP	34	4.15	305
AMF St Lq	13	3.77	49	DG InvUSGv	54	3.61	166	FrkUSTr	44	4.03	191	JPMInstITrsy	37	4.08	86
AutCsh	36	3.94	1032	DailCsh I	43	3.81	3338	FrklFTGS	15	4.03	215	JanMS Gov	41	3.37	246
AutGvt	32	3.65	2500	DlyPasp	32	3.31	2076	FrkFdl b	12	3.36	210	JanMS MM	41	3.36	313
AuGvSvc	33	3.73	407	DWitrLq	54	3.70	8637	FreeCsh	39	3.67	1098	JHanCshM	52	3.55	217
AutTreasC	37	3.51	172	DWitrUS	47	3.19	784	FreeGv	53	3.47	272	KemperGvt	25	3.83	717
BB&T UST Tr	43	3.46	73	DelaCR f	33	3.33	738	FremntMM	44	4.06	118	KemperM	35	4.02	4182
BNY Hmltn	36	4.02	357	DelaTr	73	3.02	21	FrkMny	48	3.60	1131	KeyLqd	13	3.46	453
BT InstCash	17	4.19	874	DryBasic	59	4.26	1575	FstAminstGv	35	3.99	436	KidPeCsh	39	3.71	1789
BT InstCshRv	17	4.24	786	DryBasGov	46	4.05	248	FstAmerInv	24	3.66	58	KidPeGv	45	3.43	303
BT Inst Trsy	19	3.78	156	Dry100 US	63	3.34	1635	FirPrTreasIn	54	3.37	27	KidPePr	38	3.56	813
BT InvCash	17	3.69	163	DryGvt	42	3.18	480	FirstUnMMI	44	3.86	92	LdmkUSTrs	56	3.32	209
BT InvTrsy	19	3.28	610	DryInG	46	3.28	104	FirstUnTMMT	43	4.00	192	LdmkCashRs	73	3.80	431
BT Pyramid	17	4.07	839	DryInst	46	3.55	332	FirstUnTMMI	43	3.70	451	LdmkPrmLq	73	4.10	207
Babson	40	3.42	43	DryfLA	63	3.43	4851	FdTrMny	74	3.99	84	LdmkInstLiq	73	4.30	208
BartCsR	38	3.65	75	DryMM	52	3.44	173	GEMnyMktC	31	3.88	31	LandlUSTr	56	3.77	115
BayFdMM	37	3.90	144	DryWld	59	3.28	2713	GOC UST	32	3.75	133	LaurlUSTOII	51	3.91	32
BayFdMMInv	37	3.65	55	EatVCsh	34	3.56	101	GT Mny	34	3.17	321	LaurCashInv	31	3.51	199
BedfdGv	51	3.43	176	ElfunMM	32	4.03	84	Gab OC DP	38	3.82	157	LaurGvtInv	49	3.53	49
BdfdMM	43	3.45	721	EnterMM	24	3.43	24	Gab OC UST	36	3.59	143	LaurUSGI	59	4.11	472
BenchmrkCA Mu	37	2.00	91	EvgrnM	51	4.09	281	GabelliUST	48	3.79	205	LaurPRI	38	3.98	681
BenchmrkDivAst	32	4.08	2955	ExcelsrMM	73	4.31	585	GalaxyGv	31	3.75	563	LaurPrTst	52	3.83	115
BenchmrkGovSel	41	4.15	409	ExcelsrTM	56	3.80	59	GalxyMM	29	3.59	836	LaurUSTI	55	4.01	497
BenchmrkGovt	19	3.92	782	FBL	25	2.07	18	GalaxTr	48	3.60	460	LaurUSTTst	39	3.77	190
BenhGvAg	46	3.86	500	FFB Csh	46	3.76	544	GnGvSec	44	3.72	524	LdmkPrUST	56	3.57	226
BenhPrime	68	4.51	484	FFB US Gvt	36	3.64	189	GnMMkt	55	3.47	592	LegMUS	28	3.47	205
BiltPrCshIst	39	4.27	775	FFB US Tr	30	3.70	616	GSILFd	44	4.13	1619	LegMCR f	21	3.58	797
BiltTrea	39	3.89	74	FFBLexicn	21	3.81	117	GSILGv	42	3.82	1394	Lehman100GvA	11	4.11	9
BILTmm	44	3.93	142	FidUSTI	65	4.09	1059	GSILMM	36	4.08	876	Lehman100TrA	68	4.08	65
BiltMMIv	44	3.63	39	FMBCons	12	3.62	3	GSILTrsOblg	44	3.83	942	LehmanGvtOblA	15	4.19	98
WmBlrRdy	40	3.74	481	FMBInst	12	3.62	60	GSILPO	31	3.99	2426	LehmanPrmeVal	41	4.31	1175
BlnchGv	72	3.43	237	FdShtUS	33	3.91	939	GSILTrsinst	43	4.01	472	LehmanPrime	23	4.30	2160
BradGovObl	51	3.43	39	FedMstr	36	3.93	808	GlnmndGvCash	25	4.17	351	LehmanTreasIIA	16	4.12	362
Bradfd	38	3.49	716	FidlnDom	28	4.34	885	GoldenOK-PC	28	3.87	87	Lexingt a	27	3.30	104
BullBDir	69	3.43	74	FidlnGov	50	4.18	3096	GvInvTr	15	3.13	75	LibtyUS	31	3.27	766
CRTGovt	43	3.58	919	FidlnMM	33	4.37	4001	GrdMcDnldGvRv	43	3.45	993	Liq Ins Gv	34	4.00	68
CRTMMkt	35	3.64	200	FidlnTr	29	4.14	1336	GrtHallGv	37	3.49	56	LiqInt	31	4.04	272
CapCash	24	4.00	2	FidlnTrII	31	4.06	3875	GrtHallPr	45	3.56	1002	LiqCapital	26	3.56	287
CapPre II	1	3.56	281	FidCRMM	44	3.56	671	Griffin	54	4.40	37	LiqCshTr f	6	4.23	449
CapPrsv	42	3.72	2803	FidCRGv	29	3.34	331	GrdCsFd	22	3.87	383	LrdAbCR	24	3.33	156
CapitolGovB	74	3.76	143	FidCsRes	43	3.90	13299	GrdCsMg	22	3.56	47	LuthrnBr	32	3.35	273
CapitolMMB	26	3.69	13	FidDMM	39	3.81	1561	HTInsgtCs	39	3.77	428	MASCashRes	18	4.04	33
CapitolTreas	25	3.94	283	FidDUS	29	3.63	2100	HTInsgtGv	45	3.73	243	MFSGovMonA	23	3.26	42
CapitolMM	26	3.74	69	FidDom	24	4.09	374	HanvCsh	64	3.86	659	MFSMonMkA	34	3.51	476

† Các bảng yết giá này, do Hiệp hội Quốc gia các Nhà Tự doanh Chứng khoán thu thập, thể hiện trị số trung bình của các lợi suất quy thành năm và các thời gian đáo hạn của danh mục đầu tư bình quân trọng số đô la kết thúc vào ngày 13/7/1994. Lợi suất không bao gồm những khoản lãi hay lỗ về vốn.

BẢNG **2.2** (tiếp theo)

Quỹ	Thời gian Đáo hạn Trung bình	Lợi suất 7 ngày	Tài sản
MIMLIC a	36	3.45	25
MainSty	48	3.87	184
MngdCsh	34	3.89	346
MngdGv	50	3.82	68
Map Gvt	37	3.51	72
MarinCsh	42	4.01	203
MarineGv	47	3.88	148
MarinerUS	45	3.46	135
MrkTwGovtT	20	3.43	148
MarquisTrl	34	3.72	463
MarquisTrR	34	3.63	92
Marshall Tr	37	4.09	1085
ML CBAMon	37	3.67	1281
ML CMAGv	44	3.61	3278
ML CMAMn	40	3.79	26917
MerLyGv	37	4.10	1520
MerLITr	28	3.67	285
MerLyIn	44	4.18	4437
MerLyRdy	42	3.71	6532
MerLyRet	44	3.80	7360
ML CMATr	35	3.42	1099
MerLyUSA	35	3.38	570
ML US Tr	30	3.26	59
MetLfSt E	33	3.71	155
MdlncTrGv	63	3.17	81
MdlnInst	58	3.90	49
MonarchCash	27	4.13	53
MonarchCU	27	4.41	34
MonarchGvt	22	4.16	71
MonarchGU	22	4.42	95
MonarchTrsyInst	4	3.87	37
MonMMgt f	36	3.21	128
MonMkTrst	35	3.92	561
MonitorMMT	40	3.97	327
MonitorUST	40	3.79	248
MonitrMMI	40	3.87	22
MontGovRes	53	3.74	200
MS Mony	48	3.92	688
MunderC	27	3.84	162
NatnsPrTrA	67	4.25	2722
NCC GovtI	33	3.95	759
NCC GovtR	33	3.85	6
NCC MMR	30	3.95	73
NCC MMI	30	4.05	925
NYL Inst	36	3.91	71
NatnsTrinvA	48	3.65	79
NatnsPrinvA	67	3.80	523
NatnsGvMMTrA	51	4.08	393
NatnsTreTrA	48	4.02	2530
NatwMM	43	3.75	451
NeubCsh	40	3.67	313
NeubGvt	44	3.26	204
NewEngCMTUSG	16	3.47	63
NewEngCMT	33	3.48	696
Nicholas	26	3.98	124
Northern	23	3.99	369
NorthernMun	36	1.97	466
NorthernUSGv	37	3.95	111
NorwestCash	46	3.76	1390
NorwestMunill	36	2.21	183
NorwestRdy	38	3.42	166
NorwestTr	50	3.58	554
NorwestUS	47	3.70	1131
OLDE MM	54	3.38	190
OLDE PrPl	50	4.44	262
OLDE Prem	56	3.79	82
OneGrGovt	45	4.20	670
OneGrPrA	52	4.08	1566
OneGrPr B	52	3.83	77
OneGrUSTScA	36	3.82	1329
OneGrUSTScB	36	3.57	56
OneGrTresOn	48	3.98	228
OppMoney	40	3.84	836
Oppenh Csh	38	3.37	94
OvidExTrs	46	3.43	193
OvidExMM	44	3.76	365
PFAMCo MM	28	3.97	7
PNCGovtS	49	3.94	298
PNCMoneyS	58	3.84	539
PNCMoneyI	58	4.14	470
PNCMunil	52	2.25	29
PctCptlCshAs	25	3.80	388
PctCptlUSTrs	41	3.53	87
PcHrzGvHor	29	4.13	531
PacAmerMM	47	3.68	514
PacAmerUST	41	3.82	690
PcHrzGvPH	29	3.81	531
PcHrzPr	10	3.72	2812
PcHrzPrVal	18	4.29	779
PcHrzPValH	18	4.29	779
PcHrzTr	45	3.83	2137
PcHrzTO PH	55	3.45	362
PacificGv	45	3.65	99
Pacifica	50	3.88	156
PWMnyMkA	29	3.48	36
PWMnyMkB	29	3.00	43
PWMnyMkD	29	3.02	44
PW Cash	40	3.77	3527
PW RMA	52	3.70	4413
PW RM US	50	3.48	899
PW Retr	41	3.58	2484
ParagTr	36	3.81	290
ParkTrinvA	36	3.57	58
ParkPrinvA	43	3.64	106
ParkPrinst	43	3.74	572
ParkTrinst	36	3.67	71
ParkUS C	40	3.68	193
ParkUSInvA	40	3.58	186
PeachtGov	55	4.06	57
Peacht Pr	52	4.13	58
PerfrmTrCsr	43	4.02	1
PerfrmTrInst	43	4.27	172
Phonix	54	3.38	206
PierpontMM	34	3.92	2044
PierpontTrsy	37	3.88	100
PillarPrObA	39	3.65	133
PillarUST A	42	3.42	384
PionrCs	46	3.68	174
PionUS	51	3.71	29
PiprInst	59	4.19	35
PiprMM	47	3.46	1239
PiprUS	53	3.52	206
PortgeGv	62	3.49	62
PortInstMn	37	4.04	717
PortMM	41	3.88	150
PortUS	36	3.80	196
PortUSFed	44	3.46	40
PrimeCsh	24	3.77	52
PrinPresCs	38	3.53	52
PrincorCash	36	3.68	292
ProspectHill	38	4.00	52
PIFTrsTr	52	4.01	1325
PIFTpCs	31	4.28	2381
PIFTFd	33	4.05	1088
PIFFed	40	4.15	1133
PIFFedTr	41	4.28	361
PIFTmp	50	4.17	5609
PruCdGvt	37	3.49	326
PruCdMny	37	3.86	2551
PruGvt	39	3.35	677
PruInMM	41	3.98	356
PruMart	36	3.79	6849
PruSMM	31	3.88	463
PruTr	42	3.45	278
PruInstMM	38	3.83	45
PutDDiv	38	3.94	1323
QltyCsh	49	3.46	97
QuestCshGov	36	3.39	122
QuestCshPr	47	3.51	1545
RNC Liq	66	3.87	38
RegisDSI	35	3.75	118
RemTaxTr	22	3.93	483
RemTreasTr	27	3.42	101
RemGovtTr	21	3.79	156
RenaisGvt	1	3.60	46
RenaisMM	31	3.71	322
ResrveFd Gvt	15	3.38	750
ReserveFd	41	3.36	1388
RetirGv	27	3.15	261
RimcoTrs	72	3.67	93
RIMCOPrm	65	3.85	392
RiverUSGv	26	3.72	144
RiversdeCap	34	3.23	137
RdSqMM	19	3.80	842
RdSqUS	18	3.71	399
RshFGI	41	3.45	550
Rshmre	29	3.65	23
RydexUSGv	1	3.23	113
SBSF MM	73	3.72	15
SEICshTrea	21	4.06	64
SEI CsPrB	27	3.88	11
SEI CsPrC	27	3.67	1
SEI CsFd	21	3.44	11
SEI CshGvII	27	4.03	701
SEI CsMM	33	4.35	219
SEI CsPr	27	4.18	2252
SEI LqGv	37	3.80	259
SEI LqPr	26	3.97	1037
SEI LqTr	31	3.73	1344
SEI CsTrllA	52	4.01	324
SEI CsTrllB	52	3.69	24
SITMMkt	21	3.89	15
STIPrQuTr	43	3.84	614
STIUSGvTr	35	3.63	247
STIUSGvIv	35	3.51	37
Safeco f	40	3.56	167
SalomonUST	41	3.68	28
SchbValAdv	50	4.14	1904
SchwabIntl	45	4.03	67
SchwbRetir	43	3.78	23
SchwbGv	40	3.57	1994
SchbMM	52	3.73	1055
Schb UST	51	3.49	559
ScudCshIn	34	3.75	1503
Scud UST	61	3.55	384
SecurityCsh	32	3.39	50
SelectGv	17	3.61	45
SeligCsh Gvt	24	3.29	17
SeligCshPrA	35	3.59	170
SentinelUST	48	3.26	73
SevnSea f	49	3.94	3219
SvnSeaGv	66	3.87	240
1784InstUSTr	36	4.02	180
ShawPrmMMIn	39	3.68	103
ShmtPrTr	39	3.93	443
ShtTrmIncoA	50	3.36	680
ShTrinUS	32	3.34	449
SierraGlbA	65	3.72	53
SierraGovA	30	3.34	31
SigntSIMM	72	3.88	135
SigntSITr	54	3.72	394
SmBarCash	48	3.76	2955
SmBarGvt	44	3.54	625
SmBarRetir	45	3.63	1119
SmBarShrDDiv	49	3.50	15186
SmBarShrGov	44	3.50	2972
SoTrVlcnTres	43	3.75	261
SocietyPrOb	65	3.96	781
SocietyUSGv	15	3.67	426
Stagecoach	41	3.80	869
StagecoachInt	45	3.99	115
StarPrmeObl	49	3.41	77
StarTreas	47	3.48	312
StarbGovT	50	3.44	172
StarbMMT	56	3.66	139
SteinroeCRs	70	3.83	551
SteinroeGvt	38	3.43	111
StepstnInst	38	3.67	489
StepstnInv	38	3.47	83
StepstnTrInst	39	3.74	148
StepstnTrInv	39	3.49	29
Strong	33	3.89	456
StrongUST	30	3.78	97
SumtCsh	33	3.62	135
SunAmMMA	36	3.54	302
TCW MM	37	3.92	83
TFI GovtMMIS	16	3.77	44
TRowPRF	42	3.65	3741
TRowUST	42	3.48	672
TemplInM	67	3.54	223
ThomMB	19	3.67	82
ThmNtl	22	3.77	709
TowerUSTreas	30	3.56	42
TowerCsh	34	3.43	172
TransamCs	29	3.78	382
TransamUS	33	3.94	100
TrstShtGv	34	3.84	1431
TCU MMP	39	4.11	263
TrGvCsh	34	3.94	952
TrstUSTrOb	33	3.76	3527
TwCntPrCapR	17	3.97	45
TwCntCs	18	3.68	1358
231 Prime	33	4.25	1437
231 Trsy	26	4.03	120
UMB Fed	45	3.67	168
UMB Prim	39	3.76	191
US TreSec	18	3.10	163
US Tr Am	39	3.37	198
US Treasry	49	3.62	20
USAA Mutl	52	4.10	980
USAA Treas	53	3.71	38
UST Gvt	14	3.64	647
UST Mny	14	3.91	800
UST Treas	64	3.77	262
USTrCshIns	38	3.81	274
UtdCshM	61	3.51	317
UtdSvsGvt	22	3.93	626
VIMMP	41	4.30	554
VaiLin	37	3.74	356
VnEckUS a	88	4.09	49
VanKmpMM	19	3.28	27
VangAdmUST	41	4.01	1076
VangFdl	41	4.01	2052
VangPr f	42	4.11	13361
VangUST	40	3.84	1983
VictoryPt	39	4.28	462
VictoryUSTrs	33	3.71	145
VisnMM	36	4.05	275
VisnTr	45	3.82	225
VistaFedInst	60	4.28	78
VistaGloPS	79	4.07	494
VistaGloVS	79	3.92	200
VistaGlol	79	4.25	328
VistaGovPS	28	3.80	549
VistaGov VS	28	3.55	335
VistaGovtInt	28	4.08	195
VistaPrInst	25	4.38	65
VistaPrimeP	25	4.20	161
Voyager	33	3.41	34
WPG GovMM	22	3.53	217
WarburgPCR	50	3.91	272
WestcrCs	11	3.81	252
WestcrGv	38	3.62	73
WestcrPr	37	4.06	258
WestcrTr	35	4.05	369
WoodTreas	48	3.78	627
WoodGv	76	3.82	368
WoodMM	62	3.89	1510
WorkAsets	45	3.37	104
ZweigGvt	37	3.59	87

Nguồn: Dữ liệu từ "Tóm lược về Thị trường Vốn ngắn hạn," *Tạp chí Phố Wall*, 14/7/1994, trang C27. In lại với sự cho phép của Tạp chí Phố Wall, © 1984 Dow Jones & Company Inc.

Những đặc điểm của quỹ đầu tư thị trường vốn ngắn hạn như là một công cụ đầu tư được cho thấy trong dữ liệu của Bảng 2.2. Bảng 2.2 trình bày qui mô tài sản (tính bằng triệu đô la), thời gian đáo hạn trung bình (tính bằng ngày) của kỳ phiếu, và lợi suất 7−ngày trung bình (%) trong thời kỳ kết thúc vào ngày 13/7/1994, đối với 604 quỹ đầu tư thị trường vốn ngắn hạn lớn và có sẵn cho các nhà đầu tư. Xem xét Bảng 2.2 thì chúng ta sẽ thấy rõ vấn đề khó khăn về thống kê. Mặc dù có thể có được cảm nhận tổng quát về qui mô tài sản, thời gian đáo hạn trung bình, và suất sinh lợi trung bình qua việc xem xét dữ liệu trong bảng này, nhưng khó mà có được một hình ảnh rõ ràng về những đặc điểm của các tập dữ liệu này bằng cách chỉ xem xét kỹ bảng này. Vấn đề này thúc đẩy chúng ta nghiên cứu đề tài của Chương 2. Trong chương này, chúng ta xem xét những phương pháp mô tả các tập dữ liệu. Sau đó, trong Mục 2.14 (trong nguyên bản tiếng Anh), chúng ta áp dụng những kỹ thuật này vào dữ liệu về quỹ đầu tư thị trường vốn ngắn hạn nói trên và xem thông tin có tính mô tả này phù hợp như thế nào với triển vọng trở thành nhà triệu phú của chúng ta.

2.1 Biến (Variables) và Dữ liệu (Data)

Mục tiêu chủ yếu của chúng ta trong Chương 2 sẽ là trình bày một số kỹ thuật căn bản trong **thống kê mô tả** (**descriptive statistics**)—ngành thống kê liên quan đến việc mô tả những tập hợp các giá trị đo lường, cả **mẫu** (**sample**) và **tổng thể** (**population**). Sau khi chúng ta đã thu thập một tập hợp các giá trị đo lường (measurements), làm sao chúng ta có thể trình bày tập hợp này dưới một hình thức rõ ràng, có thể hiểu được và dễ đọc? Trước tiên, chúng ta phải có thể định nghĩa giá trị đo lường hay dữ liệu là gì và phân loại các loại dữ liệu chúng ta có khả năng gặp phải trong đời sống thực. Chúng ta bắt đầu bằng việc giới thiệu một số định nghĩa, một số thuật ngữ mới trong ngôn ngữ thống kê mà anh/chị cần biết.

ĐỊNH NGHĨA

- **Biến** là một đặc trưng thay đổi hay biến đổi theo thời gian, hay một đặc trưng mà biến đổi giữa các cá nhân hay các đối tượng khác nhau được xem xét tại một thời điểm nhất định •

Thí dụ, giá cổ phiếu là một biến thay đổi theo thời gian trong phạm vi một cổ phiếu đơn lẻ; nó cũng thay đổi từ cổ phiếu này sang cổ phiếu khác tại một thời điểm cho trước. Sự liên kết chính trị, nguồn gốc dân tộc, thu nhập, tuổi, và số con cái đều là biến − đó là những đặc trưng mà khác nhau tùy thuộc vào cá nhân được chọn.

Trong phần giới thiệu, chúng ta đã định nghĩa một **đơn vị thí nghiệm** (**experimental unit**) là đối tượng mà người ta lấy giá trị đo lường. Một cách tương đương, chúng ta có thể định nghĩa một đơn vị thí nghiệm là đối tượng mà trên đó một biến được đo lường. Khi một biến được đo lường thật sự trên một tập hợp các đơn vị thí nghiệm, thì một tập hợp các giá trị đo lường hay **dữ liệu** được tạo ra.

ĐỊNH NGHĨA

- **Một đơn vị thí nghiệm** là cá nhân hay đối tượng mà trên đó một biến được đo lường. Một **giá trị đo lường** đơn lẻ hay một giá trị dữ liệu được tạo ra khi một biến được đo lường thật sự trên một đơn vị thí nghiệm •

Nếu một giá trị đo lường được tạo ra đối với mọi đơn vị thí nghiệm trong toàn bộ tập hợp, thì tập dữ liệu được tạo ra là **tổng thể** được quan tâm. Bất kỳ một tập hợp con nhỏ hơn nào của những giá trị đo lường cũng là một **mẫu.**

THÍ DỤ **2.1** Một tập hợp gồm năm người làm công được chọn từ những người làm công tại một công ty lớn, và những giá trị đo lường sau đây được ghi chép. Hãy thảo luận về các biến được đo đối với năm người làm công này.

Người làm công	Điểm số về thành quả	Giới tính	Số năm phục vụ	Phân loại việc làm	Tiền lương (nghìn đô la)
1	18	Nữ	12	Bán hàng	35
2	15	Nữ	9	Quản lý	55
3	10	Nam	2	Hành chánh	23
4	19	Nam	15	Quản lý	58
5	15	Nữ	13	Bán hàng	36

Lời giải Có một số biến trong thí dụ này. **Đơn vị thí nghiệm** mà trên đó mỗi biến được đo lường là một người làm công nhất định trong công ty. Đối với mỗi người làm công, có năm biến được đo lường: điểm số về thành quả, giới tính, số năm phục vụ, phân loại việc làm, và tiền lương. Mỗi trong những đặc trưng này thay đổi từ người làm công này sang người làm công khác. Nếu chúng ta xem những điểm số về thành quả của tất cả người làm công tại công ty này là tổng thể được quan tâm, thì năm điểm số về thành quả đó thể hiện một **mẫu** từ tổng thể này. Nếu như điểm số về thành quả của mỗi người làm công của công ty này đều được đo lường, thì chúng ta lẽ ra đã tạo ra toàn bộ **tổng thể** các giá trị đo lường cho biến này.

Biến thứ hai được đo lường trên những người làm công này là *giới tính*, mà có thể được xếp vào một trong hai loại − nam hay nữ. Nó không phải là một biến được đánh giá bằng số, và như thế nó có phần khác với *điểm số về thành quả*. Nếu có thể được nêu từng người, thì tổng thể sẽ gồm có một tập hợp những chữ Nam và Nữ, mỗi chữ đại diện cho mỗi người làm công tại công ty này. Tương tự, biến thứ tư, *phân loại việc làm*, tạo ra dữ liệu không phải bằng số, với một loại cho mỗi phân loại việc làm tại công ty này. Các biến thứ ba và thứ năm, *số năm đã làm việc* và *tiền lương*, đều được đánh giá bằng số, chúng ta tạo ra một tập hợp số chứ không phải một tập hợp các loại.

Mặc dù chúng ta đã thảo luận về từng biến một, hãy nhớ rằng chúng ta đã đo lường từng biến trong năm biến này trên năm đơn vị thí nghiệm − đó là năm người làm công. Vì thế, trong thí dụ này, một quan sát trên một cá nhân gồm có năm giá trị đo lường**.** Thí dụ, quan sát được thực hiện trên người làm công 2 mang lại kết quả đo lường sau đây:

(15, Nữ, 9, quản lý, 55) •

Anh/Chị có thể thấy rằng có sự khác biệt giữa một biến *đơn lẻ* được đo lường trên một đơn vị thí nghiệm đơn lẻ và *nhiều* biến được đo lường trên một đơn vị thí nghiệm đơn lẻ. Nếu một biến đơn lẻ được đo lường, thì dữ liệu tạo ra được gọi là **dữ liệu đơn biến**. Nếu hai biến được đo lường trên một đơn vị thí nghiệm đơn lẻ (chẳng hạn như giới tính và tiền lương), thì dữ liệu tạo ra được gọi là **dữ liệu nhị biến**. Nếu nhiều hơn hai biến được đo lường, như trong Thí dụ 2.1, thì dữ liệu được gọi là **dữ liệu đa biến.**

2.2 Các Loại Biến

Thí dụ 2.1 chứng tỏ rằng việc đo lường các biến tạo ra dữ liệu có thể bằng số hoặc không phải bằng số. Các biến mà dẫn đến dữ liệu không phải bằng số, trong đó các quan sát được phân loại dựa theo những điểm tương tự hay những điểm khác biệt về loại, thì được gọi là **biến định tính** (**qualitative variables**). Sự liên kết chính trị, nghề nghiệp, tình trạng gia đình, và số năm học trung học phổ thông đều là những thí dụ về biến định tính, cũng như các biến "giới tính" và "phân loại việc làm" trong Thí dụ 2.1. Các biến được sử dụng để đo lường một đặc điểm mà tạo ra những quan sát bằng số thì được gọi là **biến định lượng (quantitative variables)**. Chỉ số Công nghiệp Dow−Jones, lãi suất cơ bản, số xe taxi không đăng ký ở một thành phố, mức sử dụng điện hàng ngày cho một nhà máy công nghiệp đều là những thí dụ về các biến định lượng, vốn dẫn đến dữ liệu định lượng.

ĐỊNH NGHĨA • **Các biến định lượng** dẫn đến các quan sát bằng số thể hiện một số lượng. **Các biến định tính** dẫn đến các quan sát không phải bằng số mà có thể được phân loại •

Những biến định lượng, mà thường được biểu hiện bằng chữ cái *x*, có thể được phân loại thêm nữa dựa vào miền giá trị bằng số mà một giá trị đo lường có thể có. Các biến, chẳng hạn như số thành viên trong các gia đình ở Arizona, doanh số xe hơi mới tại Trung tâm Mua sắm Xe hơi Riverfront, và số lốp xe có khiếm khuyết được trả lại cho nhà sản xuất để thay thế, có các giá trị tương ứng với một tập hợp con của số đếm 0, 1, 2, …. Cụ thể là các biến này có thể nhận một số có thể đếm được các giá trị và được gọi là **biến rời rạc** (**discrete variables**). Cái tên *rời rạc* phản ánh thực tế là có những khoảng trống rời rạc giữa các giá trị khả dĩ mà dữ liệu có thể có. Mặt khác, những giá trị đo lường trên các biến chẳng hạn như chiều cao, trọng lượng, thời gian, khoảng cách, hay thể tích có thể có những giá trị tương ứng với tất cả các điểm trên một khoảng vạch (line interval). Loại biến này được gọi là **biến liên tục** (**continuous variables**). Giữa bất kỳ hai giá trị nào của một biến liên tục, luôn luôn có thể tìm thấy một giá trị thứ ba.

ĐỊNH NGHĨA • **Biến liên tục** là một biến có thể nhận tất cả giá trị nhiều vô hạn tương ứng với một khoảng vạch. **Biến rời rạc** chỉ có thể nhận một số có thể đếm được các giá trị •

THÍ DỤ **2.2** Hãy xác định mỗi biến trong các biến sau đây là *định tính* hay *định lượng*.

a. Mục đích sử dụng thường xuyên nhất của lò vi ba của Anh/Chị (hâm lại, làm hết đông lạnh, đun nóng, mục đích khác) suốt tháng 12/2000.

b. Số người tiêu dùng từ chối trả lời cuộc điều tra bằng điện thoại.

c. Loại dịch vụ cáp được cung cấp cho nơi cư trú (cáp tiêu chuẩn, cáp cao cấp hay chỉ có anten) ở Atlanta.

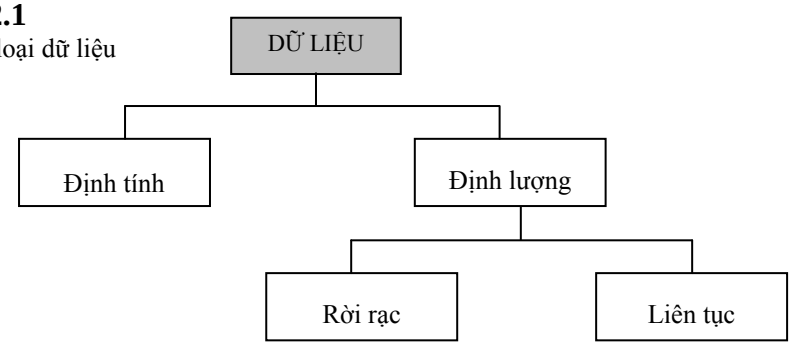
d. Thời gian hoàn tất đối với một nhiệm vụ nào đó được thực hiện bởi một chương trình phần mềm máy tính.

e. Số cổ phiếu ở Sở Giao dịch Chứng khoán New York cho thấy có tăng giá từ 1/3/2000 đến 1/7/2000.

Lời giải Các biến (a) và (c) đều là biến định tính, bởi vì chỉ có một đặc điểm được đo lường trên mỗi đơn vị thí nghiệm. Các loại đối với hai biến này được trình bày trong các ngoặc đơn. Ba biến còn lại là biến định lượng. Số người tiêu dùng là biến *rời rạc*; nó có thể nhận bất kỳ giá trị nào trong các giá trị 0, 1, 2, …, với giá trị tối đa phụ thuộc vào số người tiêu dùng được gọi điện thoại phỏng vấn. Tương tự, số cổ phiếu cho thấy có tăng giá có thể nhận bất kỳ giá trị nào trong các giá trị 0, 1, 2, …, với giá trị tối đa phụ thuộc vào số cổ phiếu ở Sở Giao dịch Chứng khoán New York. Biến (d), thời gian hoàn tất đối với một nhiệm vụ nào đó, là biến *liên tục* duy nhất trong danh sách ở trên. Thời gian hoàn tất có thể là 121 giây, 121,25 giây, hay một giá trị nằm giữa hai giá trị bất kỳ được liệt kê. •

Tại sao chúng ta phải quan tâm đến các loại khác nhau của biến và dữ liệu chúng tạo ra? Các kỹ thuật được sử dụng để tổng hợp (summarizing) và mô tả các tập dữ liệu phụ thuộc vào loại dữ liệu được thu thập. Dữ liệu định tính thường được tổng hợp bằng cách xác định số lượng hay tỷ lệ những quan sát trong mỗi một trong một số loại. Sau đó các kết quả được biểu hiện bằng cách sử dụng bảng và đồ thị. Những biểu hiện bằng đồ thị có phần khác nhau đối với các biến định lượng rời rạc và liên tục, nhưng nhìn chung chúng tập trung vào những đồ thị trong đó số quan sát trong một lớp hay loại được vẽ theo các lớp hay các loại. Đối với mỗi tập dữ liệu Anh/Chị gặp phải, thì kỹ xảo sẽ là xác định loại dữ liệu nào liên quan và làm sao anh/chị có thể biểu hiện nó theo một cách thức rõ ràng và có thể hiểu được đối với cử tọa của mình (xem Hình 2.1)

HÌNH **2.1**
Các loại dữ liệu



2.3 Các Phương pháp Bằng số để Mô tả Một Tập Dữ liệu

Các phương pháp bằng đồ thị hết sức hữu ích trong việc biểu hiện dữ liệu và trong việc truyền tải sự mô tả tổng quát và nhanh chóng về dữ liệu được thu thập. Điều này chứng minh, trong nhiều khía cạnh, cho câu tục ngữ một bức họa đáng giá cả ngàn từ. Tuy nhiên, có những hạn chế đối với việc sử dụng kỹ thuật bằng đồ thị để mô tả và phân tích dữ liệu. Ví dụ như, giả sử chúng ta muốn thảo luận về dữ liệu của mình trước một nhóm người và không có sẵn máy chiếu phóng đại! Chúng ta sẽ buộc phải sử dụng những thước đo mô tả khác mà sẽ truyền tải cho người nghe một hình ảnh trong trí óc về biểu đồ tần suất. Một hạn chế thứ hai và không thật là hiển nhiên của biểu đồ tần suất và các kỹ thuật bằng đồ thị khác, đó là chúng khó sử dụng nhằm những mục đích về suy luận thống kê

(statistical inference). Giả sử chúng ta sử dụng biểu đồ tần suất của mẫu để đưa ra những suy luận về hình dạng và vị trí của biểu đồ tần suất của tổng thể, dùng để mô tả tổng thể này và chúng ta chưa biết. Sự suy luận của chúng ta dựa vào giả định đúng, đó là một mức độ tương tự nào đó sẽ tồn tại giữa hai biểu đồ tần suất này, nhưng rồi chúng ta phải đối mặt với vấn đề đo lường mức độ tương tự này. Chúng ta biết rõ khi hai hình vẽ giống hệt nhau, nhưng tình hình này sẽ không có khả năng xảy ra trong thực tiễn. Nếu chúng giống hệt nhau, chúng ta có thể nói "Chúng giống nhau." Nhưng, nếu chúng khác nhau, thì khó mà mô tả được "mức độ khác biệt."

Những hạn chế của phương pháp mô tả dữ liệu bằng đồ thị có thể được khắc phục bằng việc sử dụng những **thước đo mô tả bằng số**. Thước đo mô tả bằng số dành cho một tổng thể được gọi là **tham số**. Thước đo mô tả bằng số tương ứng được tính toán từ một mẫu thì được gọi là **trị thống kê**. Như thế, chúng ta muốn sử dụng dữ liệu của mẫu để tính toán một tập hợp các con số, các trị thống kê, mà sẽ truyền tải một hình ảnh trong trí óc thật tốt về phân phối tần suất tương đối của mẫu và sẽ hữu ích trong việc đưa ra những suy luận về phân phối tần suất tương đối của tổng thể.

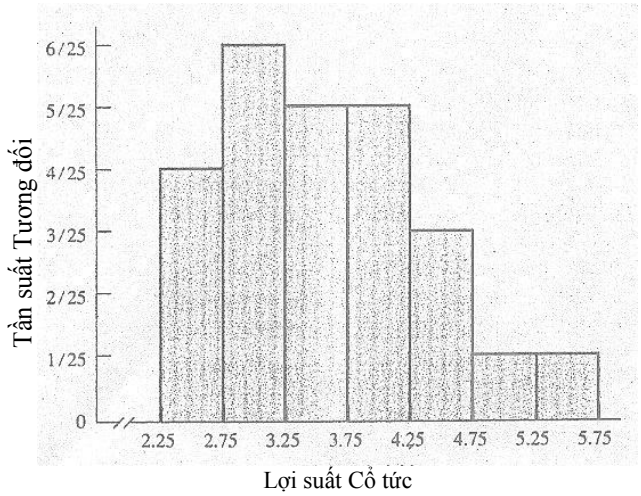
ĐỊNH NGHĨA • Các thước đo mô tả bằng số được tính từ những giá trị đo lường của tổng thể được gọi là **tham số** •

ĐỊNH NGHĨA • Các thước đo mô tả bằng số được tính từ những giá trị đo lường của mẫu được gọi là **trị thống kê** •

2.4 Các Thước đo Hướng Tâm

Trong việc xây dựng một hình ảnh trong trí óc về phân phối tần suất cho một tập hợp giá trị đo lường trên một biến định lượng, *x*, chúng ta rất có thể hình dung ra một biểu đồ tần suất tương tự với biểu đồ được trình bày trong Hình 2.2, đối với dữ liệu về lợi suất cổ tức của cổ phiếu ngân hàng. Một trong những thước đo mô tả đầu tiên được quan tâm là **thước đo hướng tâm** (**measure of central tendency**), đó là một thước đo, chẳng hạn như một số trung bình, xác định vị trí trung tâm của phân phối. Chúng ta lưu ý rằng lợi suất cổ tức thay đổi trong khoảng từ mức thấp là 2,3 lên mức cao là 5,3, với trung tâm của biểu đồ tần suất nằm gần 3,6. Bây giờ chúng ta hãy xem xét một số quy tắc rõ ràng để xác định vị trí trung tâm của một phân phối dữ liệu.

HÌNH 2.2
Biểu đồ Tần suất Tương đối



Một trong những thước đo hướng tâm hữu ích và thông dụng nhất, đó là trị số trung bình số học của một tập hợp các giá trị đo lường. Trị số này thường cũng được gọi là **trung bình số học (arithmetic mean**), hay chỉ đơn giản là **trung bình (mean**), của một tập hợp các giá trị đo lường. Bởi vì chúng ta sẽ muốn phân biệt giữa trung bình của một mẫu và trung bình của một tổng thể, nên chúng ta sẽ sử dụng ký hiệu $\bar{x}$ (x gạch ngang trên đầu) để biểu hiện trung bình của mẫu và μ (chữ muy thường của Hy Lạp) để biểu hiện trung bình của tổng thể.

ĐỊNH NGHĨA • **Trung bình số học** của một tập hợp các giá trị đo lường bằng tổng số các giá trị đo lường này chia cho số lượng giá trị đo lường •

Những quy trình tính toán trung bình mẫu và nhiều trị thống kê khác được thể hiện một cách thuận lợi thành các công thức. Do vậy, chúng ta sẽ cần một ký hiệu để biểu hiện quy trình tính tổng số. Nếu chúng ta biểu thị n số lượng phải được tính tổng số là x_1, x_2, …, x_n, thì tổng số của chúng được biểu thị bằng ký hiệu

$$\sum_{i=1}^{n} x_i$$

Chữ sigma viết hoa của Hy Lạp (Σ) là chỉ dẫn *cộng lại*. Số lượng x_i ở bên phải của Σ là phần tử tiêu biểu sẽ được cộng lại. Những ký hiệu $i = 1$ ở dưới và n ở bên trên chữ Σ chỉ ra rằng i là biến của phép tính tổng số và bắt đầu bằng trị số 1, tăng dần thêm 1, và kết thúc bằng trị số n. Thí dụ,

$$\sum_{i=1}^{3} x_i = x_1 + x_2 + x_3$$

Sử dụng ký hiệu này, chúng ta có thể biểu hiện các công thức cho trung bình mẫu và trung bình tổng thể như sau:

Các Công thức Tính Trị số Trung bình

Trung bình mẫu: $\bar{x} = \dfrac{\sum_{i=1}^{n} x_i}{n}$

Trung bình tổng thể: $\mu = \dfrac{\sum_{i=1}^{N} x_i}{N}$

THÍ DỤ **2.3** Tìm trung bình của tập hợp các giá trị đo lường 2, 9, 11, 5, 6.

Lời giải

$$\bar{x} = \frac{\sum_{i=1}^{n} x_i}{n} = \frac{2+9+11+5+6}{5} = 6{,}6$$

Thậm chí quan trọng hơn việc xác định vị trí trung tâm của một tập hợp các giá trị đo lường của mẫu, $\bar{x}$ sẽ được sử dụng làm một hàm ước lượng (hàm tiên đoán) về giá trị của trung bình μ chưa biết của tổng thể. Thí dụ, trung bình của dữ liệu trong Bảng 2.3 bằng

BẢNG **2.3**

Lợi suất cổ tức	3,1	4,2	2,3	3,3	2,8
(%) đối với	5,3	3,5	3,1	2,6	3,3
25 cổ phiếu	4,7	3,7	3,0	2,6	4,0
thường của	3,8	4,4	3,2	3,2	3,8
ngân hàng	5,1	3,7	2,3	4,3	3,9

$$\bar{x} = \frac{\sum_{i=1}^{n} x_i}{n} = \frac{89{,}2}{25} = 3{,}568$$

Hãy lưu ý giá trị này xấp xỉ rơi vào trung tâm của tập hợp các giá trị đo lường. Trung bình của toàn bộ tổng thể lợi suất cổ tức, μ, thì chúng ta chưa biết; nhưng nếu như chúng ta ước lượng giá trị của nó, thì giá trị ước lượng của chúng ta về μ sẽ là 3,586.

Một thước đo hướng tâm thứ hai là **trung vị**.

ĐỊNH NGHĨA • **Trung vị** *m* của một tập hợp *n* giá trị đo lường x_1, x_2, x_3,…, x_n là giá trị của *x* mà nằm ở giữa khi các giá trị đo lường này được xếp theo thứ tự từ nhỏ nhất đến lớn nhất •

Nếu các giá trị đo lường trong một tập dữ liệu được xếp từ nhỏ nhất đến lớn nhất, thì trung vị sẽ là giá trị của *x* nằm ở giữa. Nếu số *n* giá trị đo lường là lẻ, thì số trung vị sẽ là giá trị đo lường có thứ hạng bằng $(n + 1)/2$. Nếu số *n* giá trị đo lường là chẵn, thì số trung vị được chọn là giá trị của *x* nằm ở điểm giữa hai giá trị đo lường ở khoảng giữa − đó là ở điểm giữa giá trị đo lường có thứ hạn $n/2$ và giá trị đo lường có thứ hạng $(n / 2) + 1$. Quy tắc tính toán số trung vị được trình bày trong hộp sau đây:

Quy tắc Tính toán Số Trung vị

Xếp hạng *n* giá trị đo lường từ nhỏ nhất đến lớn nhất

1. Nếu *n* lẻ, số trung vị *m* là giá trị đo lường có thứ hạng $(n + 1)/2$
2. Nếu *n* chẵn, số trung vị *m* là giá trị của *x* nằm ở điểm giữa giá trị đo lường có thứ hạng $n/2$ và giá trị đo lường có thứ hạng $(n/2) + 1$.

THÍ DỤ **2.4** Hãy tìm số trung vị của tập hợp năm giá trị đo lường sau đây.

9, 2, 7, 11, 14

Lời giải Trước tiên, chúng ta xếp hạng $n = 5$ giá trị đo lường từ nhỏ nhất đến lớn nhất, 2, 7, 9, 11, 14. Như thế, vì $n = 5$ là số lẻ, nên chúng ta chọn 9 là số trung vị. Giá trị này là giá trị đo lường có thứ hạng là $(n + 1)/2 = (5 + 1)/2 = 3$ •

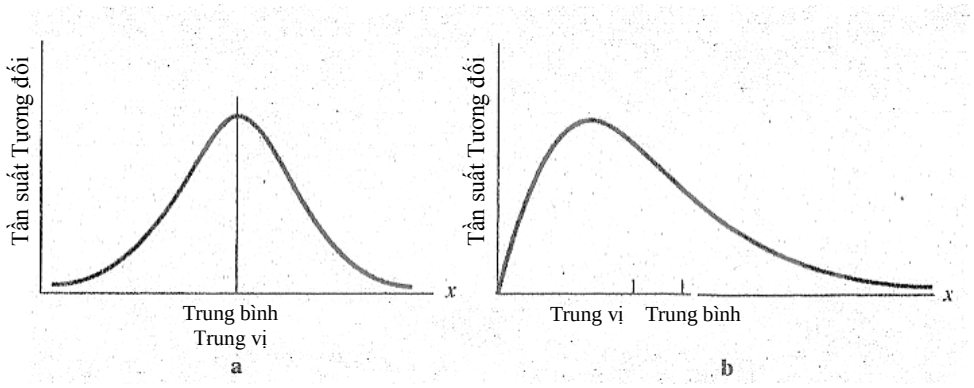
THÍ DỤ **2.5** Hãy tìm số trung vị của tập hợp các giá trị đo lường sau đây.

9, 2, 7, 11, 14, 6

Lời giải Vì $n = 6$ là số chẵn, nên chúng ta xếp hạng các giá trị đo lường thành 2, 6, 7, 9, 11, 14 và chọn số trung vị là điểm giữa của hai giá trị đo lường ở khoảng giữa, 7 và 9. Vì thế, số trung vị bằng 8 •

Mặc dù cả trung bình lẫn trung vị đều là hai thước đo tốt đối với trung tâm của một phân phối các giá trị đo lường, nhưng trung vị kém nhạy cảm với những giá trị thái cực (cực trị). Thí dụ, nếu phân phối này đối xứng qua trung bình của nó − nghĩa là hai nửa bên trái và bên phải của phân phối này là những hình ảnh phản chiếu − thì số trung bình và số trung vị bằng nhau [xem Hình 2.3 (a)]. Nếu một phân phối không đối xứng và có những quan sát thái cực nằm ở đuôi bên phải của phân phối này, thì phân phối này được gọi là bị *lệch xiên về bên phải* [xem Hình 2.3(b)]. Bởi vì những giá trị thái cực lớn ở đuôi trên của phân phối này làm tăng tổng số của các giá trị đo lường, nên số trung bình chuyển dịch sang phải. Số trung vị không bị ảnh hưởng bởi những giá trị thái cực này, bởi vì giá trị bằng số của các giá trị đo lường không được sử dụng trong việc tính toán số trung vị. Cuối cùng, nếu một phân phối bị *lệch xiên về bên trái*, thì số trung bình chuyển dịch sang trái.

HÌNH **2.3**
Các phân phối tần suất tương đối cho thấy tác động của các giá trị thái cực đối với trung bình và trung vị



Một thước đo hướng tâm khác là **yếu vị** (**cao tần**), được định nghĩa là giá trị quan sát xảy ra thường nhất trong một tập dữ liệu.

ĐỊNH NGHĨA • **Số Yếu vị** của một tập hợp n giá trị đo lường $x_1, x_2, x_3, \ldots, x_n$ là giá trị của x xảy ra với tần suất lớn nhất •

Khi các giá trị đo lường được phân nhóm trong một biểu đồ tần suất tương đối, thì lớp có tần suất tương đối lớn nhất được gọi là **lớp yếu vị**, và điểm giữa của lớp yếu vị được lấy làm giá trị của yếu vị

THÍ DỤ **2.6** Cho trước những giá trị đo lường của mẫu

5, 5, 7, 7, 7, 10, 15

giá trị 7 xảy ra ba lần, giá trị năm xảy ra hai lần, và các giá trị 10 và 15 thì mỗi số xảy ra một lần. Vì thế, số yếu vị của những giá trị đo lường của mẫu này là 7.

Đối với những phân phối đối xứng, thì các số trung bình, trung vị, và yếu vị đều bằng nhau. Trong những phân phối bị lệch xiên về bên phải, số yếu vị nằm bên trái số trung vị và số trung bình. Xem các Hình 2.3(a) và 2.3(b). Nếu phân phối bị lệch về bên trái, thì vị trí của ba thước đo này được đảo ngược, số yếu vị nằm bên phải số trung bình và số trung vị.

Một phân phối các giá trị đo lường có thể có nhiều hơn 1 số yếu vị. Thí dụ, việc phân phối tiền lương đối với một nhóm nhiều người làm công có thể tạo ra một *phân phối có hai yếu vị*, có thể phản ánh một hỗn hợp các giá trị đo lường được lấy trên những người làm công cổ xanh và cổ trắng.

Bài tập

Các Kỹ thuật Căn bản

2.1 Hãy xét $n = 5$ giá trị đo lường, 0, 5, 1, 1, 3.

a. Hãy vẽ một đồ thị phân tán cho dữ liệu này. [*Gợi ý*: Nếu hai giá trị đo lường giống nhau, hãy đặt chấm này ở trên chấm kia]. Hãy phỏng đoán "trung tâm" xấp xỉ.
b. Hãy tìm số trung bình, số trung vị, và số yếu vị.
c. Hãy xác định vị trí của ba thước đo vừa tìm ra trong phần (b) trên đồ thị phân tán trong phần (a). Dựa trên các vị trí tương đối của số trung bình và số trung vị, Anh/Chị cho là những giá trị đo lường này đối xứng hay bị lệch xiên?

2.2 Hãy xét $n = 8$ giá trị đo lường, 3, 1, 5, 4, 4, 3, 5.

a. Tìm $\bar{x}$
b. Tìm m
c. Dựa trên kết quả của các phần (a) và (b), những giá trị đo lường này bị lệch xiên hay đối xứng? Hãy vẽ đồ thị phân tán để xác nhận câu trả lời của anh/chị

2.3 Cho trước $n = 10$ giá trị đo lường, 3, 5, 4, 6, 10, 5, 6, 9, 2, 8, hãy tìm:

a. $\bar{x}$
b. m
c. số yếu vị

Ứng dụng

2.4 Nhiều người mua máy tính đã phát hiện ra rằng họ có thể tiết kiệm được một số tiền đáng kể bằng việc mua máy tính cá nhân từ một công ty nhận đặt và giao hàng qua đường bưu điện − trung bình là 900USD theo giá trị ước lượng của họ. ("Who's Tops," 1992). Điểm xếp hạng về sự thỏa mãn của khách hàng (trên thang đo từ 1 đến 9) đối với bảy công ty như thế, dựa trên cuộc điều tra 4.000 người mua, được trình bày dưới đây.

Công ty	Xếp hạng	Công ty	Xếp hạng
CompuAdd	7,5	Insight	7,8
Dell	7,9	Northgate	7,7
FastMicro	7,4	Zeos	8,0
Gateway	8,2		

a. Điểm xếp hạng trung bình về sự thỏa mãn của khách hàng đối với bảy công ty này là bao nhiêu?
b. Hãy cho biết số trung vị của những điểm xếp hạng về sự thỏa mãn của khách hàng.
c. Nếu anh/chị là một người mua máy tính, anh/chị có quan tâm đến điểm xếp hạng trung bình về sự thỏa mãn của khách hàng hay không? Nếu không, thước đo nào anh/chị quan tâm? Hãy giải thích.

2.5 Thu nhập bình quân mỗi cổ phiếu trong quý hai, năm 1994, đối với một mẫu gồm 20 công ty được trình bày dưới đây:

$ 0,72	0,56	0,21	0,54	0,32
1,28	0,10	1,64	0,29	0,33
0,29	0,73	0,29	0,33	0,43
0,56	0,89	0,84	0,62	0,44

Nguồn: Dữ liệu trích từ *Press−Enterprise*, Riverside, Calif, 20 tháng 7, 1994

a Anh/Chị cho rằng phân phối về thu nhập bình quân mỗi cổ phiếu là đối xứng hay bị lệch xiên?
b Hãy tính số trung bình, số trung vị và số yếu vị cho những giá trị ước lượng này.
c Hãy vẽ một biểu đố tần suất tương đối cho tập dữ liệu này. Hãy xác định vị trí của số trung bình, số trung vị và số yếu vị dọc theo trục hoành. Câu trả lời của anh/chị đối với phần (a) có đúng hay không?

2.6 Tạp chí *PC World* cung cấp một nguồn thông tin tuyệt vời cho những người sử dụng máy tính muốn nâng cấp hệ điều hành hiện tại của họ hay mua những hệ điều hành mới. Số gần đây của tạp chí *PC World* ("Top 10," 1994) đã liệt kê mười bộ tăng tốc dựa trên Windows hàng đầu, cùng với điểm xếp hạng giá trị toàn bộ và giá ngoài đường ước lượng, như được trình bày trong bảng sau đây:

Bộ tăng tốc	**Điểm Xếp hạng Giá trị Toàn bộ**	**Giá Ngoài đường Ước lượng**
Diamond Stealth	87	$249
Number Nine	86	275
Genoa Phantom	85	245
Hercules Dynamite Pro	82	210
miroCrystal8S	82	195
Orchid Kelvin	75	275
Hercules Graphite	73	335
Matrox MGA	73	475
Hercules Dynamite Power	72	237
Paradis Ports o'Call	72	235

a Điểm xếp hạng giá trị toàn bộ trung bình cho mười sản phẩm này là bao nhiêu?
b Giá ngoài đường ước lượng trung bình là bao nhiêu?
c Nếu anh/chị sắp mua một bộ tăng tốc, thì những số trung bình này có quan trọng đối với anh/chị hay không? Hãy giải thích.

2.7 Trong một bài báo với nhan đề "Bạn không phải là người bị hoang tưởng nếu bạn nghĩ ai đó quan sát mọi hành động của bạn", *Tạp chí Phố Wall* (19 tháng 3, 1985) lưu ý rằng doanh nghiệp lớn thu thập số liệu thống kê chi tiết về hành vi của bạn. Công ty Jockey International biết bạn có bao nhiêu quần lót; công ty Frito−Lay, Inc., biết bạn ăn cái nào trước − những cái

bánh quy bể trong gói bánh hay những cái còn nguyên; và bắt tay vào những điều cụ thể, công ty Coca−Cola biết rằng bạn bỏ vào ly 3,2 cục nước đá. Bạn đã bao giờ bỏ 3,2 cục nước đá vào cái ly chưa? Bài báo của *Tạp chí Phố Wall* muốn nói gì qua lời phát biểu đó?

2.8 Bảng sau đây trình bày nợ bình quân đầu người đối với từng bang trong 50 bang trong năm tài chính 1992.

Bang	Nợ bình quân đầu người	Thuế bình quân đầu người	Bang	Nợ bình quân đầu người	Thuế bình quân đầu người
AL	998	1019	MT	2266	1153
AK	8418	2730	NE	1092	1176
AZ	743	1259	NV	1457	1369
AR	809	1145	NH	3882	770
CA	1225	1495	NJ	2540	1643
CO	857	1018	NM	1015	1415
CT	3644	1846	NY	3083	1661
DE	5140	1944	NC	558	1316
FL	911	1068	ND	1615	1186
GA	662	1076	OH	1106	1099
HI	4040	2335	OK	1138	1206
ID	1210	1303	OR	2114	1113
IL	1611	1157	PA	1079	1354
IN	913	1143	RI	5125	1270
IA	669	1280	SC	1300	1092
KS	192	1110	SD	2657	794
KY	1762	1353	TN	558	900
LA	2331	991	TX	453	964
ME	2135	1347	UT	1187	1096
MD	1698	1324	VT	2706	1339
MA	4002	1651	VA	1160	1101
MI	1097	1195	WA	1400	1648
MN	924	1662	WV	1431	1297
MS	621	954	WI	1457	1380
MO	1213	988	WY	1920	1386

Nguồn: Dữ liệu từ Bộ Thương mại Hoa Kỳ, Cục Điều tra Dân số, *The World Almanac and Book of Facts*, ấn bản 1994, trang 105

a Hãy tìm số nợ bình quân đầu người trung bình cho 50 bang.
b Hãy tìm số nợ bình quân đầu người trung vị cho 50 bang này và so sánh nó với số trung vị đã tính trong phần (a)
c Dựa trên sự so sánh của anh/chị trong phần (b), anh/chị có kết luận rằng phân phối của nợ bình quân đầu người bị lệch xiên? Hãy giải thích.

2.9 Việc định giá đơn vị đã trở thành một tiêu chuẩn toàn ngành trong hoạt động kinh doanh tạp hóa. Công việc của người tiêu dùng là cân nhắc chất lượng của sản phẩm so với giá đơn vị để cố gắng xác định "món hời nhất". Những giá trị đo lường (measurements) sau đây là giá mỗi túi nhựa lót thùng rác, được ghi nhận đối với 10 nhãn hiệu khác nhau của túi nhựa lót thùng rác 13−gallon và cao (*Báo cáo Người tiêu dùng*, Tháng 2, 1994).

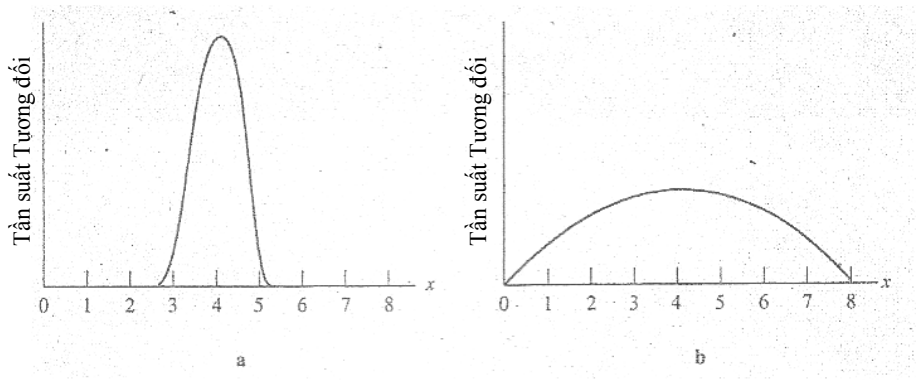
10	9	13	8	9
10	10	6	5	11

a Hãy tìm giá trung bình mỗi túi nhựa lót
b Hãy tìm giá trung vị mỗi túi nhựa lót
c Nếu anh/chị đang viết báo cáo để mô tả những dữ liệu này, thước đo hướng tâm nào anh/chị sẽ sử dụng? Hãy giải thích.

2.5 Những Thước đo về Độ Biến thiên

Một khi chúng ta đã xác định trung tâm của một phân phối dữ liệu, bước tiếp theo là cung cấp một thước đo về **độ biến thiên** (**variability**), hay **độ phân tán** (**dispersion)**, của dữ liệu này. Hãy xét hai phân phối được trình bay trong Hình 2.4. Cả hai phân phối đều được đặt ở vị trí có trung tâm tại $x = 4$, nhưng có sự khác biệt lớn về độ biến thiên của những giá trị đo lường xung quanh số trung bình đối với hai phân phối này. Các giá trị đo lường trong Hình 2.4(a) thay đổi xấp xỉ từ 3 đến 5; trong Hình 2.4(b), các giá trị đo lường thay đổi từ 0 đến 8.

HÌNH **2.4**
Độ biến thiên hay độ phân tán của dữ liệu



Sự biến thiên là một đặc trưng quan trọng của dữ liệu. Thí dụ, nếu chúng ta đang chế tạo bu lông, thì sự biến thiên quá mức trong đường kính của bu lông sẽ kéo theo một tỷ lệ phần trăm cao của sản phẩm có khiếm khuyết. Mặt khác, khi chúng ta sử dụng một bài kiểm tra để phân biệt giữa những kế toán viên giỏi và kém, thì chúng ta sẽ không vui nhất nếu bài kiểm tra này lúc nào cũng mang lại những điểm kiểm tra với ít biến thiên, bởi vì điều này sẽ làm cho việc phân biệt trở nên rất khó khăn.

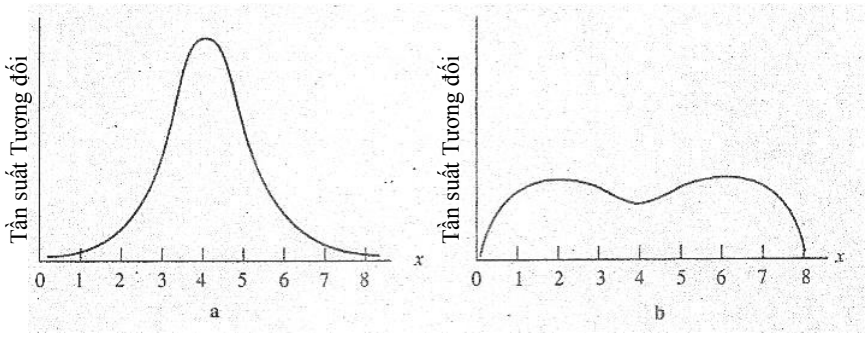
Ngoài tầm quan trọng trên thực tế của sự biến thiên trong dữ liệu, một thước đo về đặc trưng này còn cần thiết cho việc xây dựng một hình ảnh trong trí óc về phân phối tần suất. Chúng ta sẽ chỉ thảo luận về vài trong số nhiều thước đo về sự biến thiên.

Thước đo đơn giản nhất về sự biến thiên là **khoảng biến thiên** (**miền**).

ĐỊNH NGHĨA • **Khoảng biến thiên** (**range**) của một tập hợp n giá trị đo lường x_1, x_2, x_3,…, x_n được định nghĩa là chênh lệch giữa giá trị đo lường lớn nhất và giá trị đo lường nhỏ nhất •

Dữ liệu về lợi suất cổ tức thay đổi từ 2,3 đến 5,3. Như thế, khoảng biến thiên là (5,3−2,3) = 3,0. Khoảng biến thiên dễ tính toán, dễ diễn giải, và hoàn toàn thỏa đáng trong vai trò một thước đo về sự biến thiên cho những tập dữ liệu nhỏ. Nhưng đối với những tập dữ liệu lớn thì khoảng biến thiên không phải là một thước đo thỏa đáng về độ biến thiên. Thí dụ, hai phân phối tần suất tương đối trong Hình 2.5 có cùng khoảng biến thiên nhưng lại có hình dạng và độ biến thiên rất khác nhau.

HÌNH **2.5**
Những phân phối có khoảng biến thiên bằng nhau và độ biến thiên khác nhau



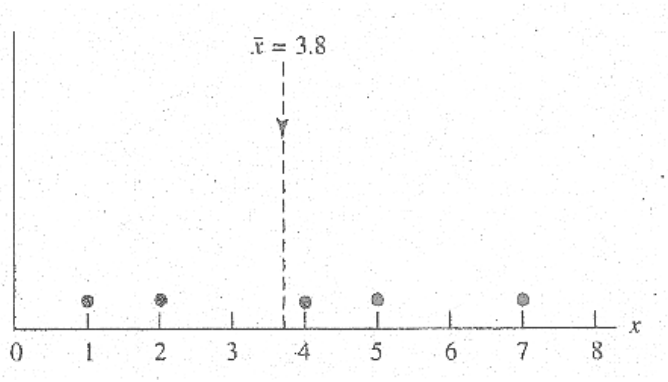
Chúng ta có thể tìm ra một thước đo về độ biến thiên nhạy cảm hơn khoảng biến thiên hay không? Lấy thí dụ, hãy xét các giá trị đo lường của mẫu 5, 7, 1, 2, 4, được biểu hiện thành đồ thị phân tán trong Hình 2.6. Số trung bình của năm giá trị đo lường này là

$$\bar{x} = \frac{\sum_{i=1}^{n} x_i}{n} = \frac{19}{5} = 3{,}8$$

như được chỉ ra trong đồ thị phân tán này

Bây giờ chúng ta có thể xem xét độ biến thiên theo khoảng cách giữa mỗi điểm (giá trị đo lường) và trung bình $\bar{x}$. Nếu những khoảng cách này lớn thì chúng ta có thể nói rằng dữ liệu biến thiên nhiều hơn so với khi những khoảng cách này nhỏ. Nói rõ hơn, chúng ta định nghĩa **độ lệch** của một giá trị đo lường khỏi số trung bình của nó là lượng $(x_i - \bar{x})$. Những giá trị đo lường nằm bên phải của số trung bình tạo ra độ lệch dương, và những giá trị đo lường nằm bên trái tạo ra độ lệch âm. Đối với thí dụ của chúng ta, các giá trị của *x* và các độ lệch được trình bày trong cột thứ nhất và cột thứ hai của Bảng 2.4.

HÌNH **2.6**
Đồ thị phân tán



BẢNG **2.4** Những phép tính liên quan đến độ lệch của mẫu

x_i	$(x_i - \bar{x})$	$(x_i - \bar{x})^2$	$\lvert x_i - \bar{x} \rvert$
5	1,2	1,44	1,2
7	3,2	10,24	3,2
1	−2,8	7,84	2,8
2	−1,8	3,24	1,8
4	0,2	0,04	0,2
19	0,0	22,80	9,2

Nếu chúng ta đồng ý rằng các độ lệch chứa đựng thông tin về sự biến thiên, thì bước tiếp theo của chúng ta là xây dựng một thước đo về sự biến thiên dựa trên các độ lệch xung quanh số trung bình. Khả năng đầu tiên là chúng ta có thể chọn trung bình của các độ lệch. Đáng tiếc là trung bình này sẽ không có tác dụng, bởi vì một số độ lệch thì dương, một số thì âm, và tổng số luôn luôn bằng không (trừ khi những sai số làm tròn số đã được đưa vào các phép tính). Hãy lưu ý rằng các độ lệch trên cột thứ hai của Bảng 2.4 có tổng bằng không.

Có hai cách để tránh được vấn đề này. Tại sao không tính số trung bình của giá trị tuyệt đối của các độ lệch? Thước đo này được gọi là **độ lệch tuyệt đối trung bình** (**mean absolute deviation**, MAD).

ĐỊNH NGHĨA • **Độ lệch tuyệt đối trung bình** của một tập hợp n giá trị đo lường $x_1, x_2, \ldots, x_n$ là số trung bình của giá trị tuyệt đối của các độ lệch xung quanh trung bình mẫu và được cho bởi công thức

$$\text{MAD} = \frac{\sum_{i=1}^{n} \lvert x_i - \bar{x} \rvert}{n} \quad \bullet$$

Các độ lệch tuyệt đối của tập hợp $n = 5$ giá trị quan sát (observations) của chúng ta cùng với tổng số của chúng được trình bày trong Bảng 2.4. Vì thế cho nên,

$$\text{MAD} = \frac{\sum_{i=1}^{n} \lvert x_i - \bar{x} \rvert}{n} = \frac{9,2}{5} = 1,84$$

Mặc dù MAD đôi khi được sử dụng làm thước đo về độ biến thiên cho một tập dữ liệu, nhưng nó chủ yếu được dùng trong việc đánh giá độ chính xác của tiên đoán.

Cách thứ hai để sử dụng độ lệch là làm việc với tổng các bình phương của độ lệch. Bằng việc sử dụng tổng của các độ lệch bình phương, chúng ta tính toán một thước đo đơn lẻ gọi là **phương sai** (**variance**) của một tập hợp các giá trị đo lường. Để phân biệt giữa phương sai của *mẫu* và phương sai của *tổng thể*, chúng ta sử dụng ký hiệu s^2 để biểu hiện phương sai mẫu và σ^2 (chữ sigma thường của Hy Lạp) để biểu hiện phương sai tổng thể. *Thước đo này sẽ tương đối lớn đối với dữ liệu biến thiên nhiều và tương đối nhỏ đối với dữ liệu biến thiên ít.*

ĐỊNH NGHĨA • **Phương sai của tổng thể** gồm N giá trị đo lường $x_1, x_2, \ldots, x_N$ được định nghĩa là trị trung bình của các bình phương của độ lệch của các giá trị đo lường xung quanh số trung bình μ của chúng. Phương sai của tổng thể (phương sai tổng thể) được cho bởi công thức

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

Rất thường là anh/chị sẽ không có sẵn tất cả các giá trị đo lường của tổng thể, mà sẽ cần tính toán *phương sai của một mẫu* gồm n giá trị đo lường.

ĐỊNH NGHĨA • **Phương sai của mẫu** gồm n giá trị đo lường $x_1, x_2, \ldots, x_n$ được định nghĩa là tổng các độ lệch bình phương của các giá trị đo lường này xung quanh số trung bình $\bar{x}$ của chúng, chia cho $(n - 1)$. Phương sai mẫu được ký hiệu bằng chữ s^2 và được cho bởi công thức

$$s^2 = \frac{\sum_{i=1}^{n}(x_i - \bar{x})^2}{n-1} \quad \bullet$$

Thí dụ, chúng ta có thể tính phương sai cho tập hợp gồm $n = 5$ giá trị đo lường của mẫu, được trình bày trong Bảng 2.4. Bình phương của độ lệch của mỗi giá trị đo lường được ghi trên cột thứ ba của Bảng 2.4. Cộng lại, chúng ta thu được

$$\sum_{i=1}^{5}(x_i - \bar{x})^2 = 22{,}80$$

Phương sai mẫu là

$$s^2 = \frac{\sum_{i=1}^{n}(x_i - \bar{x})^2}{n-1} = \frac{22{,}80}{4} = 5{,}70$$

Phương sai được đo theo bình phương của các đơn vị đo lường ban đầu. Nếu các giá trị đo lường ban đầu được tính bằng inch, thì phương sai được biểu hiện bằng inch bình phương. Lấy căn bậc hai của phương sai này, thì chúng ta có **độ lệch chuẩn** (**standard deviation**), mà chuyển thước đo độ biến thiên trở lại các đơn vị đo lường ban đầu.

ĐỊNH NGHĨA • **Độ lệch chuẩn** của một tập hợp các giá trị đo lường bằng căn bậc hai dương của phương sai. •

Hệ thống ký hiệu

n: số giá trị đo lường trong mẫu	N : số giá trị đo lường trong tổng thể
s^2 : phương sai mẫu	σ^2 : phương sai tổng thể
$s = \sqrt{s^2}$: độ lệch chuẩn của mẫu	$\sigma = \sqrt{\sigma^2}$: độ lệch chuẩn của tổng thể

Đối với tập hợp $n = 5$ giá trị đo lường của mẫu trong Bảng 2.4, phương sai mẫu là $s^2 = 5{,}70$, do đó độ lệch chuẩn của mẫu là $s = \sqrt{s^2} = \sqrt{5{,}70} = 2{,}39$. Tập dữ liệu càng biến thiên, thì giá trị của s sẽ càng lớn

Đối với tập hợp nhỏ của các giá trị đo lường chúng ta đã sử dụng, thì việc tính toán phương sai không quá khó. Tuy nhiên, đối với một tập hợp lớn hơn, những tính toán có thể trở nên rất nhàm chán. Hầu hết máy tính cầm tay có khả năng thống kê đều có các chương trình cài sẵn mà sẽ tính $\bar{x}$ và s hay μ và σ, do đó công việc tính toán của anh/chị sẽ được giảm đến mức thấp nhất. Phím trung bình của mẫu hay tổng thể thường được đánh dấu bằng chữ $\bar{x}$. Phím độ lệch chuẩn của mẫu thường được đánh dấu bằng chữ s hay σ_{n-1}, và phím độ lệch chuẩn của tổng thể thường được đánh dấu bằng chữ σ hay σ_N. Khi sử dụng bất kỳ máy tính cầm tay nào có những phím chức năng cài sẵn này, hãy nắm chắc rằng anh/chị biết phép tính toán nào đang được thực hiện bởi mỗi phím!

Nếu anh/chị cần tính s^2 và s bằng tay, thì sẽ dễ dàng hơn nhiều nếu sử dụng công thức tính thay thế được cho dưới đây. Hình thức tính toán này đôi khi được gọi là phương pháp đi tắt để tính toán s^2.

Công thức tính toán đối với s^2

$$s^2 = \frac{\sum_{i=1}^{n} x_i^2 - \frac{\left(\sum_{i=1}^{n} x_i\right)^2}{n}}{n-1}$$

trong đó

$\sum_{i=1}^{n} x_i^2 =$ tổng các bình phương của những giá trị quan sát đơn lẻ

$\left(\sum_{i=1}^{n} x_i\right)^2$ = bình phương của tổng các giá trị quan sát đơn lẻ

Độ lệch chuẩn của mẫu, s, là căn bậc hai dương của s^2.

THÍ DỤ **2.7** Hãy tính phương sai và độ lệch chuẩn cho năm giá trị đo lường trong Bảng 2.4 mà được cho trước là 5, 7, 1, 2, và 4. Hãy sử dụng công thức tính toán đối với s^2 và so sánh các kết quả của anh/chị với các kết quả thu được bằng cách sử dụng định nghĩa nguyên thủy của s^2.

BẢNG **2.5** Bảng dành để tính toán s^2 và s theo cách đã đơn giản hóa

x_i	x_i^2
5	25
7	49
1	1
2	4
4	16
19	95

Lời giải Những số ghi trong Bảng 2.5 là các giá trị đo lường đơn lẻ, x_i, và bình phương của chúng, x_i^2, cùng với tổng của chúng. Bằng việc sử dụng công thức tính toán đối với s^2, chúng ta có

$$s^2 = \frac{\sum_{i=1}^{n} x_i^2 - \frac{\left(\sum_{i=1}^{n} x_i\right)^2}{n}}{n-1} = \frac{95 - \frac{(19)^2}{5}}{4} = \frac{22,80}{4} = 5,70$$

và $s = \sqrt{s^2} = \sqrt{5,70} = 2,39,$ như đã tính trước đây •

THÍ DỤ **2.8** Hãy tính phương sai mẫu và độ lệch chuẩn cho $n = 25$ lợi suất trong Bảng 2.3

BẢNG **2.3**

Lợi suất cổ tức	3,1	4,2	2,3	3,3	2,8
(%) đối với	5,3	3,5	3,1	2,6	3,3
25 cổ phiếu	4,7	3,7	3,0	2,6	4,0
thường của	3,8	4,4	3,2	3,2	3,8
ngân hàng	5,1	3,7	2,3	4,3	3,9

Lời giải Bằng việc sử dụng một máy tính cầm tay có các chức năng thống kê cài sẵn, anh/chị có thể kiểm tra những kết quả sau đây:

$$\sum_{i=1}^{n} x_i = 89,2$$

$$\sum_{i=1}^{n} x_i^2 = 333,82$$

Sử dụng công thức tính toán

$$s^2 = \frac{\sum_{i=1}^{n} (x_i - \bar{x})^2}{n-1} = \frac{\sum_{i=1}^{n} x_i^2 - \frac{\left(\sum_{i=1}^{n} x_i\right)^2}{n}}{n-1} = \frac{333,82 - \frac{(89,2)^2}{25}}{24} = \frac{15,5544}{24} = 0,6481$$

và $s = \sqrt{s^2} = \sqrt{0,6481} = 0,81$ •

Anh/Chị có thể tự hỏi tại sao chúng ta chia cho $n - 1$ chứ không phải n, khi chúng ta tính toán phương sai mẫu. Trung bình mẫu $\bar{x}$ được sử dụng như là một hàm ước lượng của trung bình tổng thể, bởi vì nó cung cấp một giá trị ước lượng tốt về μ. Nếu chúng ta muốn sử dụng phương sai mẫu như là hàm ước lượng của phương sai tổng thể σ^2, thì phương sai mẫu s^2 với $n = 1$ ở mẫu số sẽ cho ra những giá trị ước lượng về σ^2 tốt hơn so với một hàm ước lượng được tính với n ở mẫu số. **Vì lý do này, chúng ta sẽ luôn luôn chia cho $n - 1$ khi tính toán phương sai mẫu s^2 và độ lệch chuẩn của mẫu s.**

Vào lúc này, anh/chị đã biết cách thức tính toán phương sai và độ lệch chuẩn của một tập hợp các giá trị đo lường. Hãy nhớ những điểm sau đây:

- Giá trị của s^2 hay s càng lớn, thì độ biến thiên của tập dữ liệu càng lớn
- Nếu s^2 hay s bằng số không, thì tất cả các giá trị đo lường phải có cùng giá trị
- Độ lệch chuẩn s được tính toán để có một thước đo về độ biến thiên mà được đo lường bằng cùng đơn vị như các giá trị quan sát.

Thông tin này cho phép chúng ta so sánh vài tập dữ liệu xét theo vị trí và độ biến thiên của chúng. Chúng ta có thể sử dụng những thước đo này như thế nào để nói điều gì đó cụ thể hơn về một tập dữ liệu duy nhất? Định lý và quy tắc được trình bày trong phần sau sẽ giúp chúng ta trả lời câu hỏi này.

2.6 Các Thước đo về Vị trí Tương đối (Measures of Relative Standing)

Đôi khi chúng ta muốn biết vị trí của một giá trị quan sát so với những giá trị quan sát khác trong một tập dữ liệu. Thí dụ, nếu anh/chị dự một kỳ thi tìm việc làm và đạt số điểm là 640, anh/chị có thể muốn biết tỷ lệ phần trăm những người tham dự đạt số điểm thấp hơn 640. Một **thước đo về vị trí tương đối** như thế của một giá trị quan sát trong một tập dữ liệu được gọi là **phân vị**.

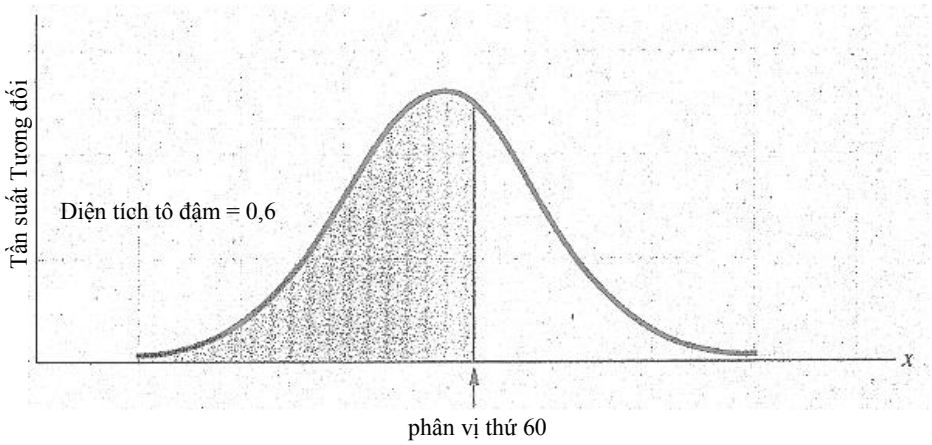
ĐỊNH NGHĨA
- Cho $x_1, x_2, \ldots, x_n$ là một tập hợp n giá trị đo lường được sắp xếp theo thứ tự tăng dần. **Phân vị** thứ p là giá trị của x sao cho nhiều nhất là p phần trăm các giá trị đo lường là thấp hơn giá trị đó của x và nhiều nhất là $(100-p)$ phần trăm là lớn hơn. •

THÍ DỤ **2.9** Trước khi được nhận vào học một chương trình thạc sĩ quản trị kinh doanh (MBA) tại một trường đại học, anh/chị đã được thông báo rằng số điểm của anh/chị là 610 trong Kỳ Kiểm tra Miệng về Thành tích của Người Tốt nghiệp Đại học đã đặt anh/chị tại phân vị thứ 60 trong phân phối của những số điểm. Số điểm 610 của anh/chị đứng ở đâu so với những số điểm của những người khác cùng dự kỳ thi kiểm tra với anh/chị?

Lời giải Đạt số điểm tại phân vị thứ 60 có nghĩa là 60% những số điểm kiểm tra khác là thấp hơn số điểm của anh/chị và 40% là cao hơn.

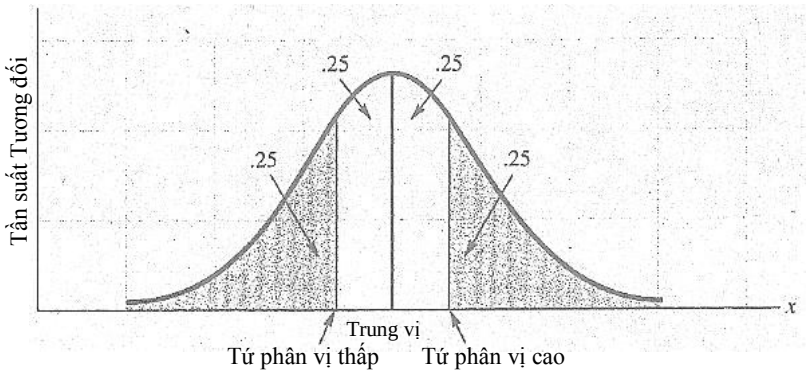
Xem xét theo đồ thị thì một phân vị nào đó, ví dụ phân vị thứ 60, là một điểm trên trục hoành *x* và nằm ở vị trí sao cho 60% diện tích bên dưới biểu đồ tần suất tương đối của dữ liệu nằm bên trái phân vị thứ 60 (xem Hình 2.7) và 40% diện tích này nằm bên phải. Như thế, theo định nghĩa, trung vị của một tập dữ liệu là phân vị thứ 50, bởi vì một nửa của các giá trị đo lường trong tập dữ liệu là nhỏ hơn số trung vị và một nửa là lớn hơn.

HÌNH **2.7**
Phân vị thứ 60 được trình bày trên biểu đồ tần suất tương đối của một tập dữ liệu



Phân vị thứ 25 và phân vị thứ 75, được gọi là **tứ phân vị thấp** và **tứ phân vị cao** (**lower and upper quartiles**), cùng với trung vị (phân vị thứ 50), xác định vị trí những điểm mà chia dữ liệu thành bốn tập hợp có cỡ bằng nhau. Hai mươi lăm phần trăm các giá trị đo lường sẽ thấp hơn tứ phân vị thấp (đầu tiên), 50% sẽ thấp hơn trung vị (tứ phân vị thứ hai), và 75% các giá trị đo lường sẽ thấp hơn tứ phân vị cao (thứ ba). Như thế, trung vị và các tứ phân vị cao và thấp nằm tại những điểm trên trục *x* sao cho diện tích bên dưới biểu đồ tần suất tương đối của dữ liệu được phân chia thành bốn diện tích bằng nhau, như được cho thấy trong Hình 2.8. Anh/Chị có thể nhận thấy (trong Hình 2.8) rằng ¼ diện tích bên dưới biểu đồ nằm bên trái của tứ phân vị thấp và ¾ nằm bên phải. Tứ phân vị cao là giá trị của *x* sao cho ¾ của diện tích này nằm bên trái và ¼ nằm bên phải.

HÌNH **2.8**
Vị trí của các tứ phân vị



Còn có **Giá trị *z*** (***z*−score**) là một thước đo khác về vị trí tương đối; nó sử dụng cả trung bình và độ lệch chuẩn của tập dữ liệu.

ĐỊNH NGHĨA • **Giá trị *z* của mẫu** tương ứng với một giá trị quan sát x là một thước đo về vị trí tương đối và được định nghĩa bằng công thức

$$\text{giá trị } z = \frac{x - \bar{x}}{s} \quad \bullet$$

Một giá trị *z* đo lường số lượng độ lệch chuẩn giữa một giá trị quan sát và trung bình của tập dữ liệu. Giả sử chúng ta biết rằng trung bình và độ lệch chuẩn của một tập hợp các số điểm kiểm tra, dựa trên một tổng số là 100 điểm, là $\bar{x} = 74$ và $s = 8$. Giá trị *z* đối với điểm kiểm tra 92 của anh/chị được tính là

$$\text{giá trị } z = \frac{x - \bar{x}}{s} = \frac{92 - 74}{8} = 2{,}25$$

Vì thế số điểm của anh/chị nằm cao hơn 2,25 độ lệch chuẩn so với trung bình; đó là, 92 = 74 + 2,25(8).

Bản thân các giá trị *z* chỉ đơn thuần cho thấy số điểm kiểm tra cao hơn hay thấp hơn trung bình bao nhiêu độ lệch chuẩn. Tuy nhiên, khi giá trị *z* được sử dụng cùng với Định lý Tchebysheff, thì có thể đưa ra một số lời phát biểu thận trọng về vị trí tương đối của một giá trị quan sát. Hơn nữa, nếu dữ liệu có hình dạng cái gò, thì Quy tắc Thực nghiệm có thể được dùng để đưa ra những lời phát biểu mạnh hơn về vị trí tương đối của một giá trị quan sát xét theo giá trị *z* của nó. Bởi vì ít nhất là 75%, và rất có thể là 95%, các giá trị quan sát trong một tập dữ liệu nằm trong phạm vi hai độ lệch chuẩn so với trung bình, nên các giá trị *z* trong khoảng từ −2 đến +2 là rất có khả năng xảy ra, và như thế không phải là không bình thường. Tuy nhiên, ít nhất là 8/9, hay rất có thể là tất cả, các giá trị quan sát nằm trong phạm vi ba độ lệch chuẩn so với trung bình. Vì thế, các giá trị *z* trong khoảng từ 2 đến 3, tính theo giá trị tuyệt đối, ít có khả năng xảy ra hơn nhiều, trong khi đó các giá trị *z* cao hơn 3, tính theo giá trị tuyệt đối, rất không có khả năng xảy ra và phải được xem xét cẩn thận. Một điểm kiểm tra có giá trị *z* cao hơn 3 là xuất sắc, trong khi đó một cổ phiếu mà tỷ số giá trên thu nhập của nó (giá của cổ phiếu chia cho thu nhập bình quân mỗi cổ phiếu hàng năm) có giá trị *z* là −3 sẽ được xem là một cuộc đầu tư có tiềm năng thu nhập tốt.

Giá trị *z* cực kỳ lớn và giá trị *z* cực kỳ nhỏ nêu lên câu hỏi về hiệu lực (validity) của một giá trị quan sát. Có thể giá trị quan sát này chỉ là hết sức lớn hoặc hết sức nhỏ so với những giá trị quan sát khác. Tuy nhiên, giá trị quan sát này có thể đã được ghi nhận không đúng, hoặc vì lý do nào đó, nó có thể không thuộc về tổng thể mà chúng ta đã mong muốn lấy mẫu. Những giá trị quan sát với các giá trị *z* hết sức lớn hoặc nhỏ thường được gọi là **giá trị dị biệt** bởi vì chúng nằm cách xa trung tâm của tập dữ liệu. Những giá trị quan sát nằm cao hơn hay thấp hơn trung bình trong khoảng từ hai đến ba độ lệch chuẩn là những giá trị dị biệt có thể có, trong khi đó những giá trị quan sát nằm cao hơn hay thấp hơn trung bình nhiều hơn ba độ lệch chuẩn thì được xem là những giá trị dị biệt rõ ràng.

THÍ DỤ **2.10** Hãy xét một mẫu gồm $n = 10$ giá trị đo lường:

3, 2, 0, 15, 2, 3, 4, 0, 1, 3

Thoạt nhìn anh/chị có thể thấy giá trị đo lường $x = 15$ dường như là một giá trị dị biệt. Hãy tính giá trị z cho giá trị quan sát này, và hãy trình bày các kết luận của anh/chị.

Lời giải Đối với mẫu này, chúng ta có những phép tính toán sau đây:

$$\sum_{i=1}^{10} x_i = 3{,}3 \quad \text{và} \quad \sum_{i=1}^{10} x_i^2 = 277$$

Như thế

$$\bar{x} = \frac{\sum_{i=1}^{10} x_i}{} = \frac{33}{10} = 3{,}3$$

$$s^2 = \frac{\sum_{i=1}^{n} x_i^2 - \frac{\left(\sum_{i=1}^{n} x_i\right)^2}{n}}{n-1}$$

$$= \frac{277 - \frac{(33)^2}{10}}{9} = \frac{168{,}1}{9} = 18{,}6778$$

$$s = 4{,}32$$

Bằng việc sử dụng những số lượng này để tính giá trị z cho giá trị dị biệt nghi ngờ $x = 15$, chúng ta tìm ra

$$\text{giá trị } z = \frac{x - \bar{x}}{s} = \frac{15 - 3{,}3}{4{,}32} = 2{,}71$$

Như thế giá trị đo lường $x = 15$ nằm cách trung bình mẫu $\bar{x} = 3{,}3$ một khoảng cách là 2,71 độ lệch chuẩn mẫu. Bởi vì giá trị z này cao hơn 2, nên chúng ta xác định $x = 15$ là một giá trị dị biệt có thể có. Chúng ta phải xem xét thủ tục lấy mẫu của mình để xem liệu có bằng chứng cho thấy $x = 15$ là một giá trị quan sát bị sai hay không •

Anh/Chị có thể sử dụng Minitab hay Excel để tạo ra nhiều trong số những thước đo mô tả bằng số mà chúng ta đã thảo luận. Trong Minitab, hãy dùng **Stat → Basic Statistics → Display Descriptive Statistics**, và chọn các biến thích hợp để mô tả. Trong Excel, hãy dùng **Tool → Data Analysis → Descriptive Statistics**, chọn dãy những ô có chứa dữ liệu, nhấn vào hộp có đánh dấu "Số liệu Thống kê Tổng hợp" ("Summary Statistics"), và chọn một ô trong đó thể hiện số liệu ra. Anh/Chị sẽ nhận ra trung bình, độ lệch chuẩn, trung vị, và những giá trị quan sát tối thiểu và tối đa trong cả hai bản in kết quả ra. Excel cũng tính số yếu vị, phương sai, và khoảng biến thiên (miền), trong khi đó Minitab bao gồm cả các tứ phân vị thấp và cao. Cả hai bản in kết quả ra đều có một số trị thống kê khác mà chúng ta chưa thảo luận.

Bản in kết quả ra của Minitab, được trình bày trong Hình 2.9, tổng hợp (summarize) các lợi suất cổ tức của Thí dụ 2.8 (dữ liệu được cho trong Bảng 2.3) và những giá trị quan sát trong Thí dụ 2.10. Anh/Chị có thể so sánh các giá trị của những trị thống kê được tính trong các thí dụ đó với các giá trị được trình bày trong bản in kết quả ra.

HÌNH **2.9**
Bản in kết quả ra của Minitab sử dụng lệnh DESCRIBE (MÔ TẢ) cho dữ liệu trong Thí dụ 2.8 (C1) và dữ liệu của Thí dụ 2.10 (C2).

	N	MEAN	MEDIAN	TRMEAN	STDEV	SEMEAN
C1	25	3,568	3,500	3,548	0,805	0,161
C2	10	3,30	2,50	2,25	4,32	1,37

	MIN	MAX	Q1	Q3
C1	2,300	5,300	3,050	4,100
C2	0,00	15,00	0,75	3,25

BÀI TẬP

Các Kỹ thuật Căn bản

2.10 Hãy sử dụng tập dữ liệu sau đây:

3, 9, 6, 5, 5, 4, 7, 6, 8, 2, 6, 7, 3

a Hãy tính $\bar{x}$ và s
b Hãy tính giá trị z cho những giá trị quan sát nhỏ nhất và lớn nhất. Có giá trị nào trong hai giá trị quan sát này lớn hay nhỏ bất thường không?

2.11 Hãy tìm giá trị z cho giá trị quan sát lớn nhất trong tập dữ liệu sau đây:

19, 12, 16, 0, 14, 9, 6, 1, 12, 13, 10, 19, 7, 5, 8

2.12 Nếu Anh/Chị đạt số điểm trong phân vị thứ 90 trong kỳ kiểm tra tuyển sinh vào lớp cao học, số điểm của anh/chị đứng ở vị trí ra sao so với những người khác cùng dự kỳ kiểm tra?

Ứng dụng

2.13 Tham khảo dữ liệu về số nợ bình quân đầu người trong Bài tập 2.8.

a Hãy tìm trung bình và độ lệch chuẩn của các số nợ bình quân đầu người này.
b Tham khảo bài tập 2.8 để tìm số nợ bình quân đầu người ở bang của anh/chị trong năm 1992. Hãy sử dụng một giá trị z để mô tả số nợ bình quân đầu người ở bang của anh/chị so sánh như thế nào với những số nợ tương ứng ở những bang khác.

2.14 Một bài báo trong Tạp chí *American Demographics* (*Nhân Khẩu học Hoa Kỳ*) (Kirchner, R., và Thomas, R., "Những Thị trường Mới đối với Bảo hiểm Y tế," tháng 12, 1990, trang 40) cung cấp một số số liệu thống kê thú vị về số dân Mỹ không có loại bảo hiểm y tế nào trong năm 1988. Trong nhiều trường hợp, những người Mỹ không được bảo hiểm này ít nhất cũng có đủ nguồn lực để trả cho bảo hiểm y tế. Theo bài báo này, "gần 40 phần trăm

số người không được bảo hiểm có thu nhập là 20.000US hay nhiều hơn; 22 phần trăm có thu nhập là 30.000USD hay nhiều hơn; và 13 phần trăm, hay trên 4 triệu, sống trong những hộ gia đình có thu nhập là 40.000USD hay nhiều hơn.” Hãy nhận dạng những phân vị nào có thể được xác định từ thông tin này.

2.15 Theo *Consumer Reports* (*Báo cáo Người Tiêu dùng*) (Tháng 3/1994), giá trung bình của một Sony SLV-700HF stereo VCR là 410USD, với độ lệch chuẩn là 14USD. Nếu anh/chị mua loại VCR này với giá 430USD, hãy tính giá trị *z* đối với giá mua của anh/chị. Giá này có cao bất thường không?

2.7 Tóm tắt

Những phương pháp mô tả tập hợp các giá trị đo lường có thể chia thành hai loại, đó là phương pháp bằng đồ thị và phương pháp bằng số. Biểu đồ tần suất tương đối là một phương pháp bằng đồ thị cực kỳ hữu ích để biểu thị đặc trưng một tập hợp các giá trị đo lường. Các thước đo mô tả bằng số là các con số mà cố gắng tạo ra một hình ảnh trong trí óc về biểu đồ tần suất (hay phân phối tần suất). Chúng ta đã hạn chế nội dung thảo luận trong các thước đo hướng tâm và sự biến thiên, mà hữu ích nhất trong các thước đo này là trung bình và độ lệch chuẩn. Mặc dù trung bình có ý nghĩa mô tả theo trực giác, nhưng độ lệch chuẩn chỉ có ý nghĩa khi được sử dụng cùng với Định lý Tchebysheff và Quy tắc Thực nghiệm. Mục tiêu của việc lấy mẫu là mô tả (đưa ra những suy luận về) về tổng thể từ đó mẫu này đã được lấy ra. Mục tiêu này được hoàn thành bằng việc sử dụng trung bình mẫu $\bar{x}$ và số lượng s^2 như là các hàm ước lượng về trung bình tổng thể μ và phương sai σ^2. Khi dữ liệu gồm có những cặp giá trị quan sát, thì đồ thị nhị biến được dùng để đánh giá bằng hình ảnh cách thức *x* thay đổi theo *y*, trong khi đó hệ số tương quan được dùng để xác định sức mạnh của mối quan hệ tuyến tính giữa *x* và *y*. Các thước đo khác, chẳng hạn như phân vị hay giá trị *z*, được dùng để xác định vị trí tương đối của một quan sát trong tổng thể hay trong một mẫu. Các đồ thị hộp là những tóm lược dữ liệu bằng hình ảnh và chúng hữu ích trong việc phát hiện các giá trị dị biệt.

Các phương pháp mô tả và các thước đo bằng số đã được trình bày chỉ là một số nhỏ trong những phương pháp và thước đo lẽ ra có thể được thảo luận. Ngoài ra, ở đây cũng đã bỏ qua nhiều kỹ thuật tính toán đặc biệt thường được tìm thấy trong những cuốn sách giáo khoa sơ cấp. Sự bỏ qua này bắt buộc phải có do thời gian hạn chế trong bất kỳ khóa học sơ cấp nào. Ngoài ra, việc sử dụng phổ biến máy tính cầm tay và máy tính đã tối thiểu hóa tầm quan trọng của những công thức tính toán đặc biệt. Nhưng quan trọng hơn, việc bao gồm vào những kỹ thuật như thế sẽ thường hay làm lu mờ và gây khó hiểu cho mục tiêu chính của thống kê hiện đại và cuốn sách này: đó là suy luận thống kê.