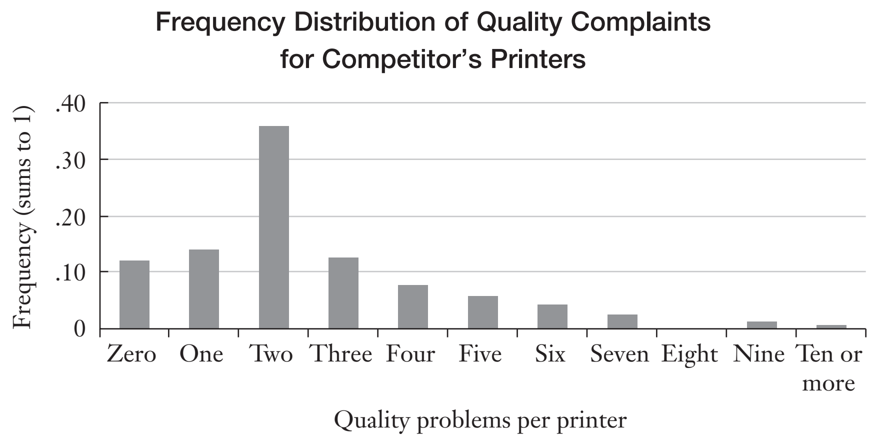
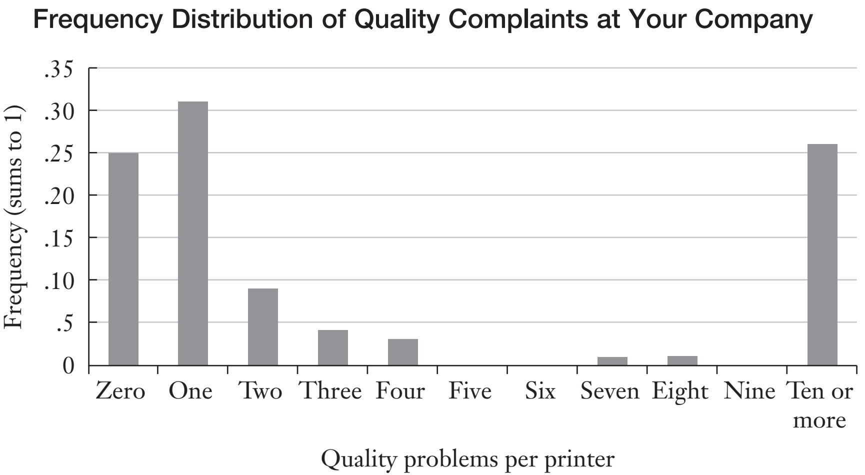
naked statistics

Stripping the Dread from the Data

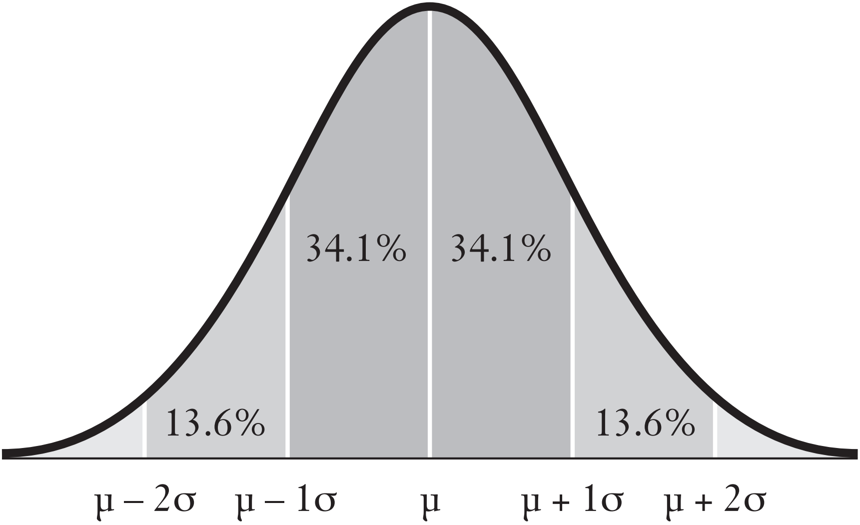
CHARLES WHEELAN

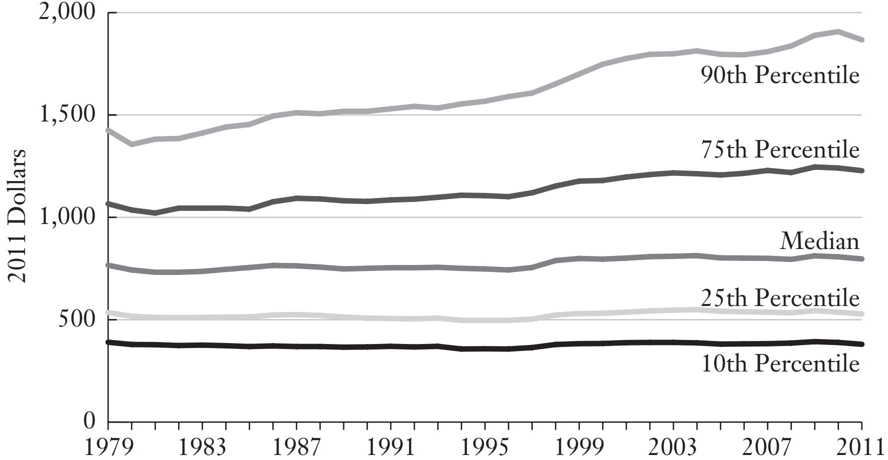
W. W. Norton & Company
New York | London


Frequency Distribution of Quality Complaints
for Competitor's Printers
Frequency (sums to 1)
.40
.30
.20
.10
0
Zero
One
Two
Three
Four
Five
Six
Seven
Eight
Nine
Ten or more
Quality problems per printer


Frequency Distribution of Quality Complaints at Your Company
Frequency (sums to 1)
.35
.30
.25
.20
.15
.10
.5
0
Zero
One
Two
Three
Four
Five
Six
Seven
Eight
Nine
Ten or more
Quality problems per printer

The Normal Distribution





Weekly Wages at Selected Percentiles

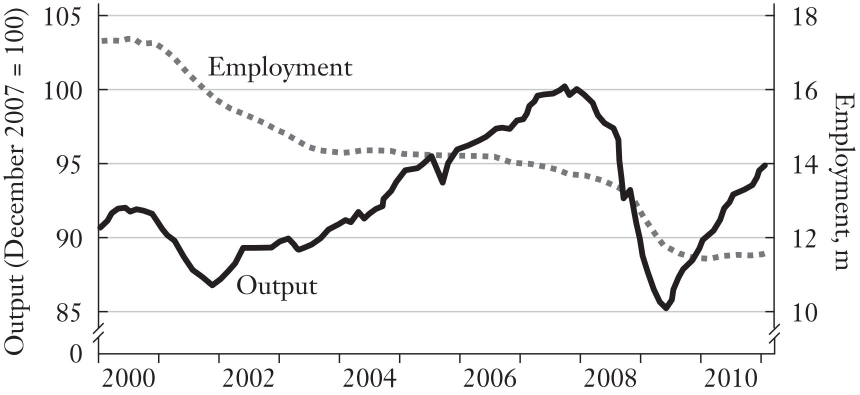
Data for the printer defects graphics

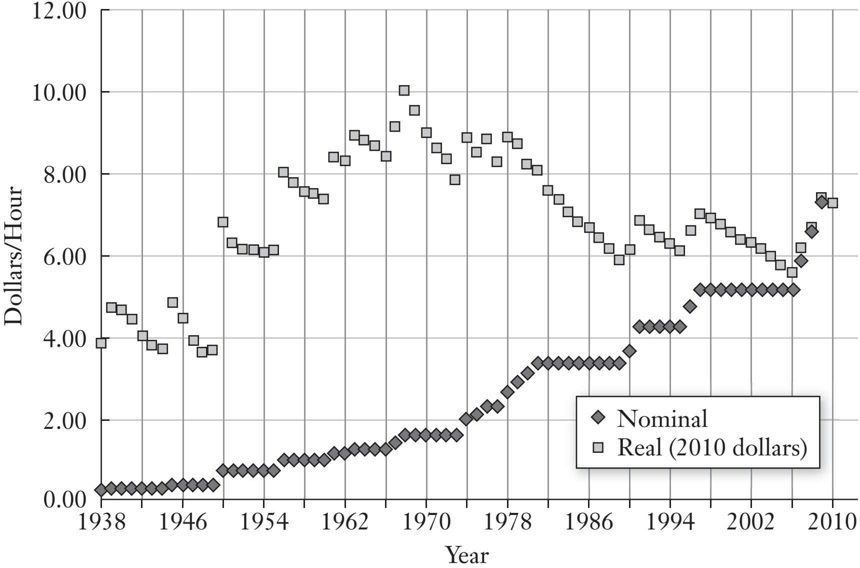
	Zero	One	Two	Three	Four	Five	Six	Seven	Eight	Nine	Ten or more
Frequency of competitor's defects	12	14	36	13	8	6	5	3	0	2	1
	Zero	One	Two	Three	Four	Five	Six	Seven	Eight	Nine	Ten or more
Frequency of your defects	25	31	9	4	3	0	0	1	1	0	26

Group 1	Height ($\mu = 70$ inches)	Distance from the mean = Absolute value of $(x_n - \mu)$*	$(x_n - \mu)^2$	Group 2	Height ($\mu = 70$ inches)	Distance from the mean = Absolute value of $(x_n - \mu)$*	$(x_n - \mu)^2$
Nick	74	4	16	Sahar	65	5	25
Elana	66	4	16	Maggie	68	2	4
Dinah	68	2	4	Faisal	69	1	1
Rebecca	69	1	1	Ted	70	0	0
Ben	73	3	9	Jeff	71	1	1
Charu	70	0	0	Narciso	75	5	25
		Total = 14	Total = 46			Total = 14	Total = 56
			Variance = 46/6 = 7.7				Variance = 56/6 = 9.3
			Standard deviation = $\sqrt{7.7} = 2.8$				Standard deviation = $\sqrt{9.3} = 3$

* Absolute value is the distance between two figures, regardless of direction, so that it is always positive. In this case, it represents the number of inches between the height of the individual and the mean.

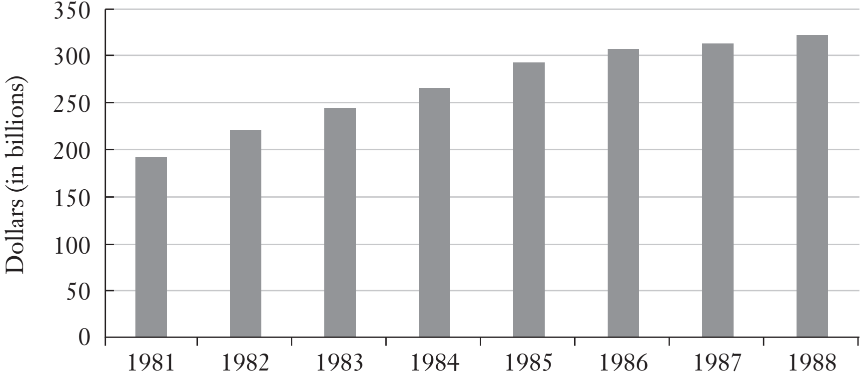
“The Rustbelt Recovery,” March 10, 2011

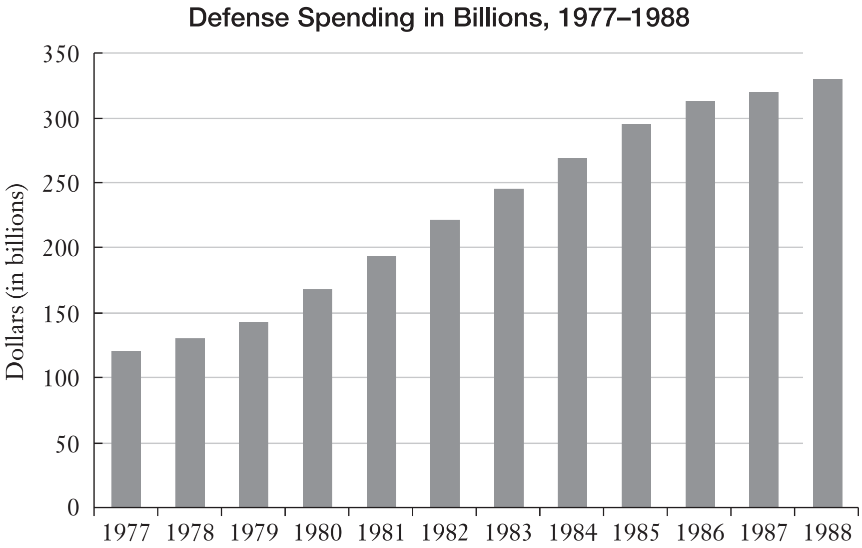




Source: http://oregonstate.edu/instruct/anth484/minwage.html.

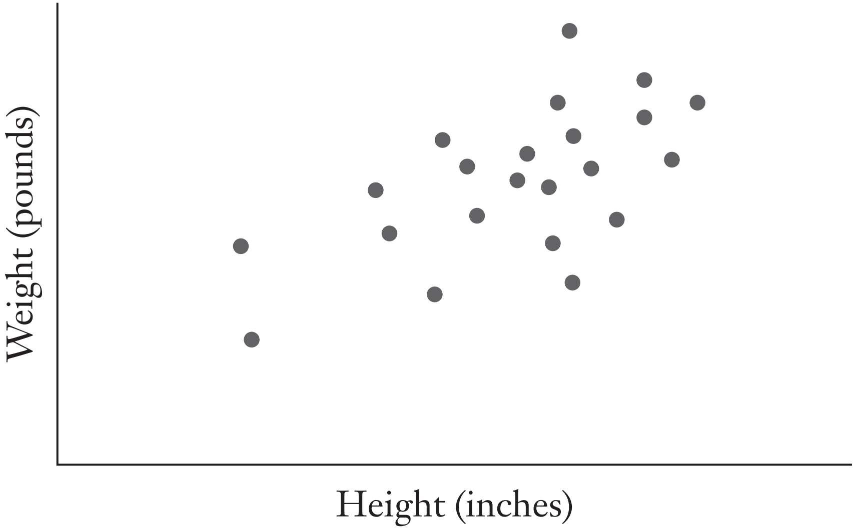
Defense Spending in Billions, 1981–1988





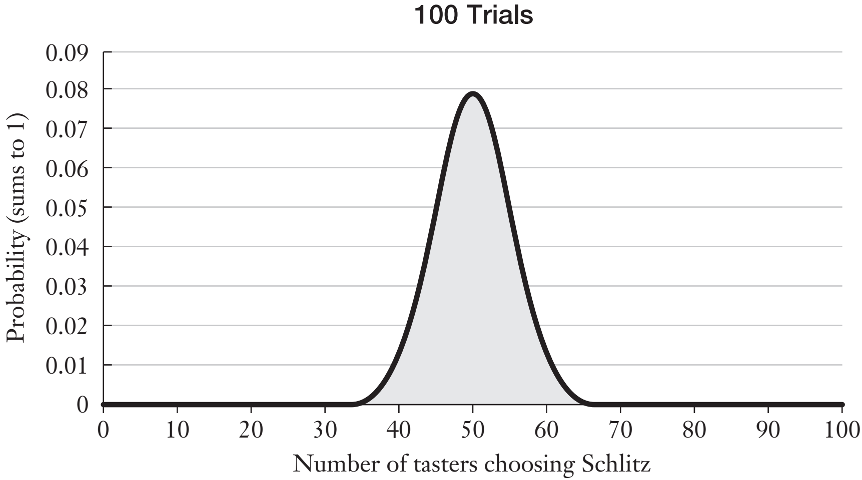
Source: http://www.usgovernmentspending.com/spend.php?span=usgs302&year=1988&view=1&expand=30&expandC=&units=b&fy=fy12&local=s&state=US&pie=#usgs302.

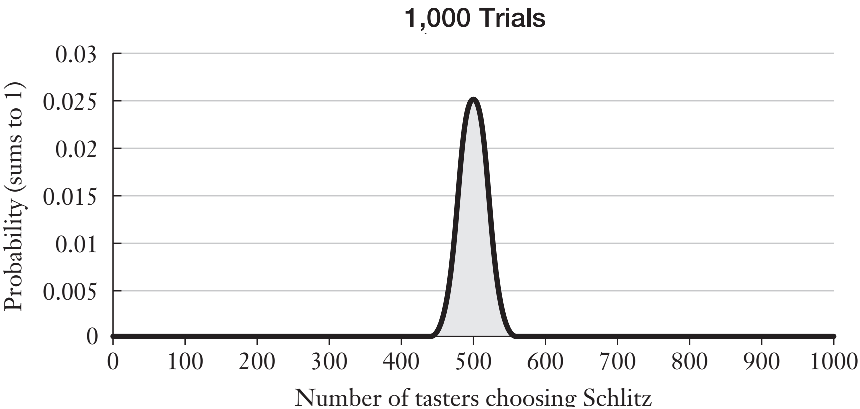
Scatter Plot for Height and Weight



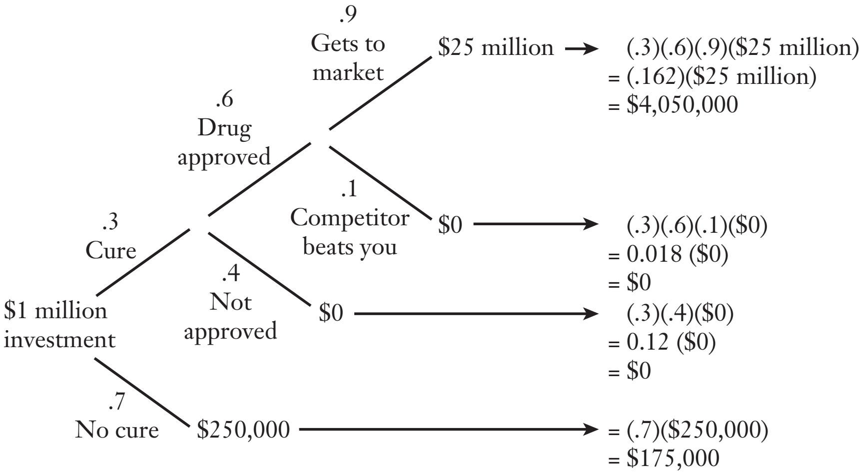
A	B	C	D	E	F
Student	Height	Weight	Height in standard units	Weight in standard units	(Weight in standard units) × (Height in standard units)
Nick	74	193	1.21	0.99	1.19
Elana	66	133	–0.63	–0.67	0.42
Dinah	68	155	–0.17	–0.06	0.01
Rebecca	69	147	0.06	–0.29	–0.02
Ben	73	175	0.98	0.49	0.48
Charu	70	128	0.29	–0.81	–0.24
Sahar	60	100	–2.00	–1.59	3.18
Maggie	63	128	–1.32	–0.81	1.07
Faisal	67	170	–0.40	0.35	–0.14
Ted	70	182	0.29	0.68	0.20
Narciso	70	178	0.29	0.57	0.17
Katrina	70	118	0.29	–1.09	–0.32
CJ	75	227	1.44	1.93	2.77
Sophia	62	115	–1.54	–1.17	1.81
Will	74	211	1.21	1.49	1.80
Mean	68.73	157.33			Total = 12.39
Standard Deviation	4.36	36.12		Correlation coefficient = Total/n = 12.39/15 = 0.83	

10 Trials
Probability (sums to 1)
0.3
0.25
0.2
0.15
0.1
0.05
0
0
1
2
3
4
5
6
7
8
9
10
Number of tasters choosing Schlitz


100 Trials
Probability (sums to 1)
0.09
0.08
0.07
0.06
0.05
0.04
0.03
0.02
0.01
0
0
10
20
30
40
50
60
70
80
90
100
Number of tasters choosing Schlitz

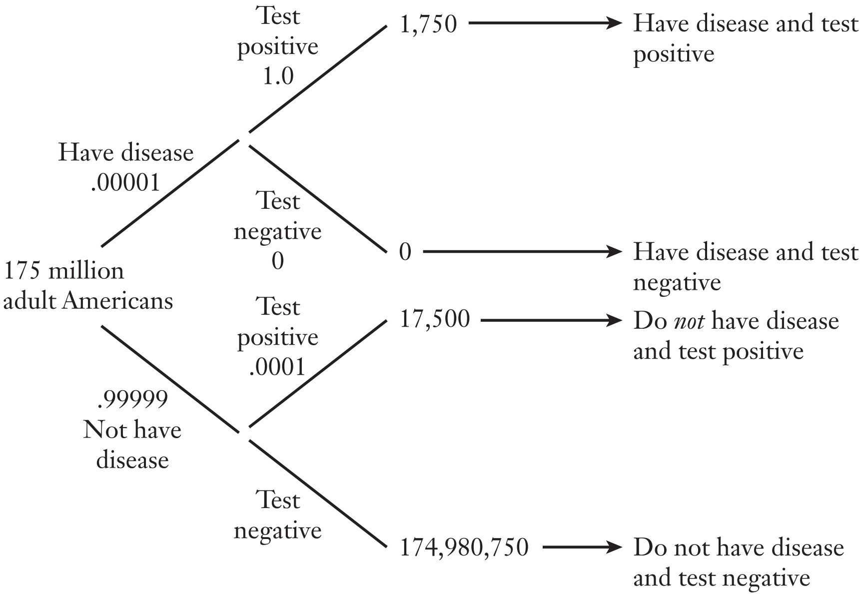

1,000 Trials
Probability (sums to 1)
0.03
0.025
0.02
0.015
0.01
0.005
0
0
100
200
300
400
500
600
700
800
900
1000
Number of tasters choosing Schlitz

The Investment Decision

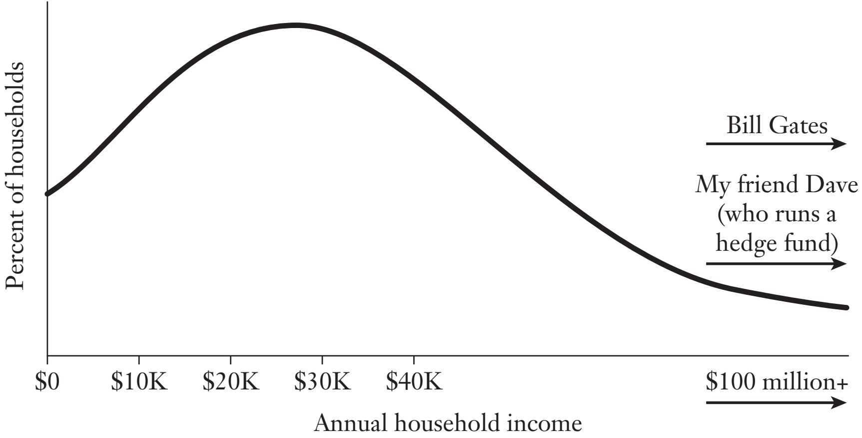


Expected payoff = \$4,050,000 + \$0 + \$0 + \$175,000
= \$4,225,000

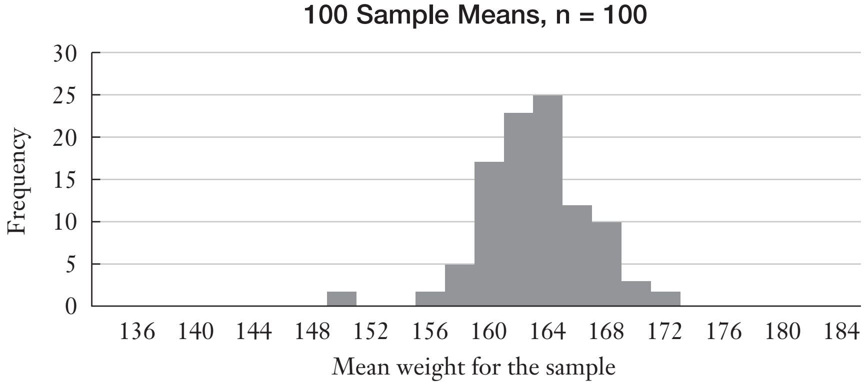
Widespread Screening for a Rare Disease



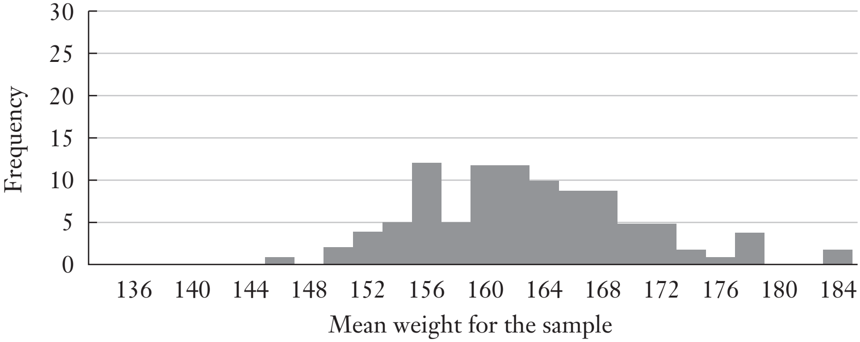
$$\frac{\text{People with disease}}{\text{Those told they have the disease}} = \frac{1{,}750}{1{,}750 + 17{,}500} = \frac{1{,}750}{19{,}250} = .09 = 9\%$$



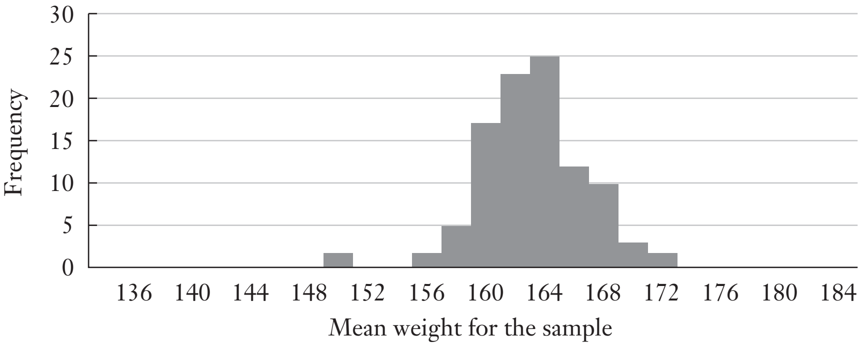
Percent of households
Bill Gates
My friend Dave (who runs a hedge fund)
$0
$10K
$20K
$30K
$40K
$100 million+
Annual household income


100 Sample Means, n = 100
Frequency
30
25
20
15
10
5
0
136
140
144
148
152
156
160
164
168
172
176
180
184
Mean weight for the sample

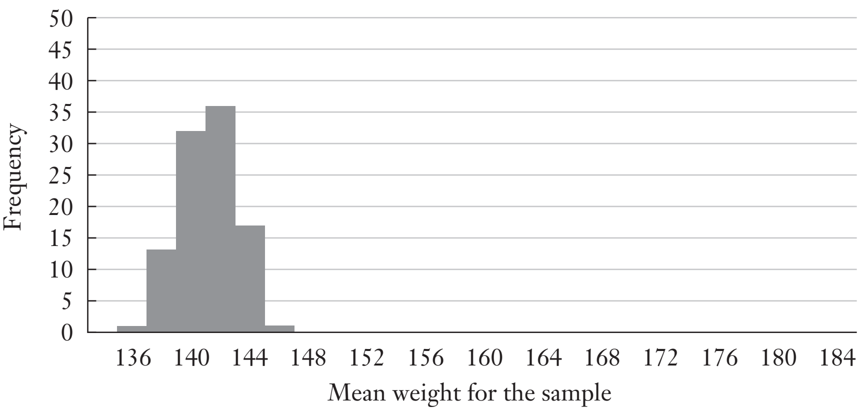
100 Sample Means, n = 20



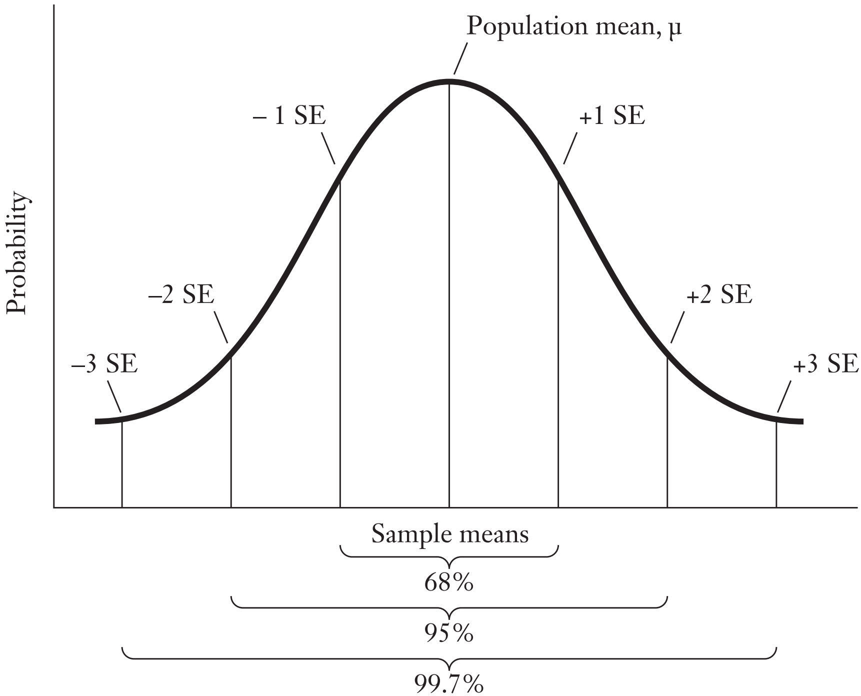
100 Sample Means, n = 100



Female Population Only/100 Sample Means, n = 100

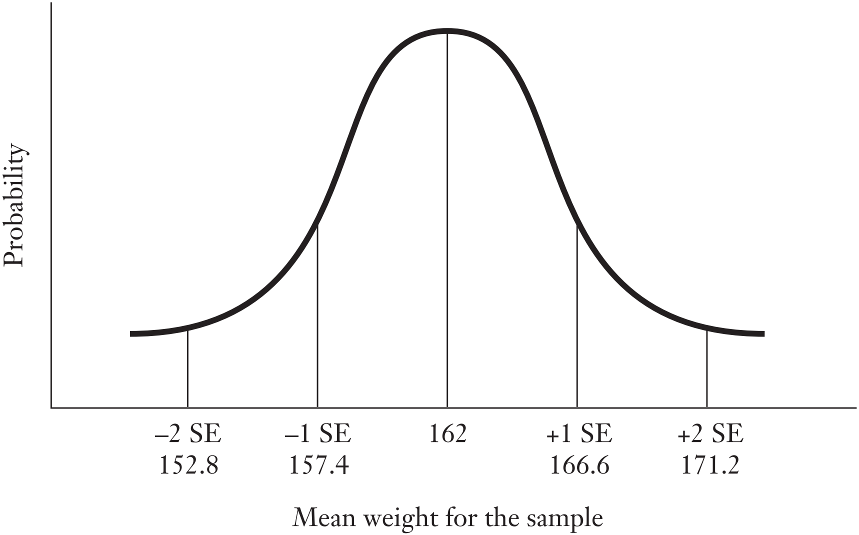


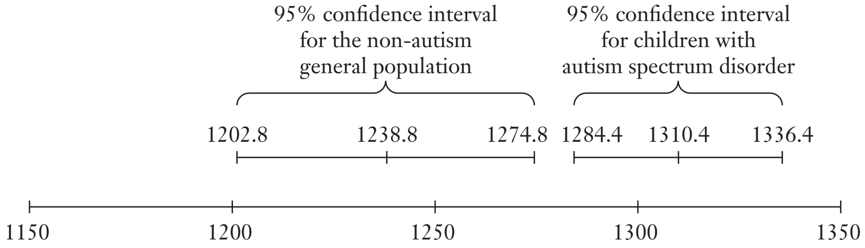
Frequency Distribution of Sample Means



* When the standard deviation for the population is calculated from a smaller sample, the formula is tweaked slightly: $SE = s/\sqrt{n-1}$. This helps to account for the fact that the dispersion in a small sample may understate the dispersion of the full population. This is not highly relevant to the bigger points in this chapter.

Distribution of Sample Means




95% confidence interval
for the non-autism
general population
95% confidence interval
for children with
autism spectrum disorder
1202.8
1238.8
1274.8
1284.4
1310.4
1336.4
1150
1200
1250
1300
1350

Formula for comparing two means

$$\frac{\bar{x} - \bar{y}}{\sqrt{\frac{s_x^2}{n_x} + \frac{s_y^2}{n_y}}}$$

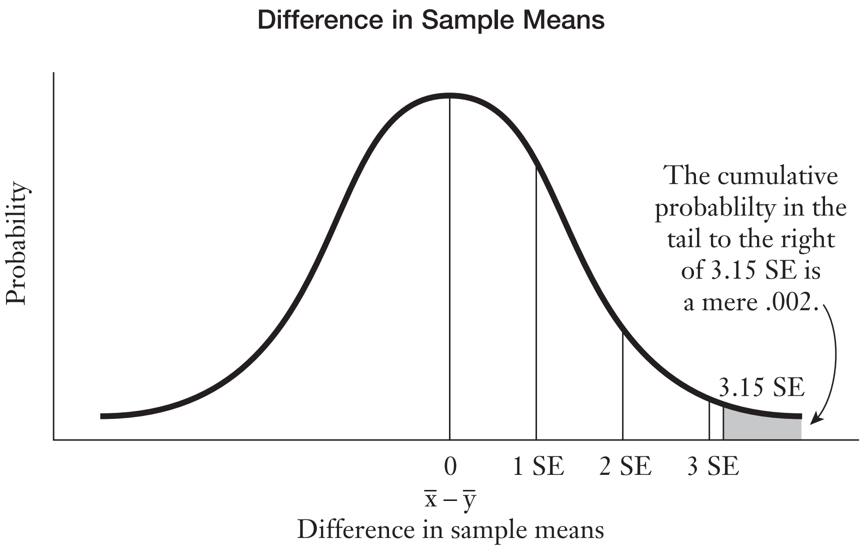
$\longrightarrow$ numerator yields the size of the difference in means

$\longrightarrow$ denominator yields the standard error for a difference in mean between two samples

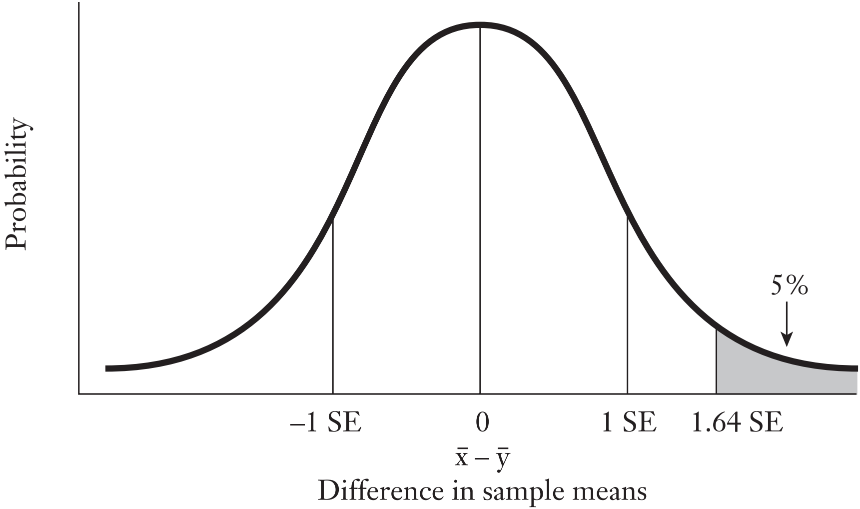
where $\bar{x}$ = mean for sample x

$\bar{y}$ = mean for sample y

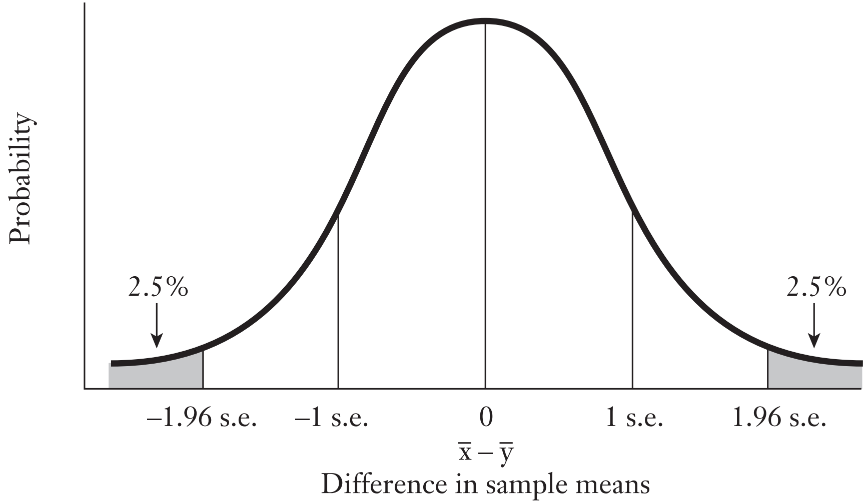
s_x = standard deviation for sample x
s_y = standard deviation for sample y
n_x = number of observations in sample x
n_y = number of observations in sample y


Difference in Sample Means
Probability
The cumulative probablilty in the tail to the right of 3.15 SE is a mere .002.
3.15 SE
0
1 SE
2 SE
3 SE
$\bar{x} - \bar{y}$
Difference in sample means

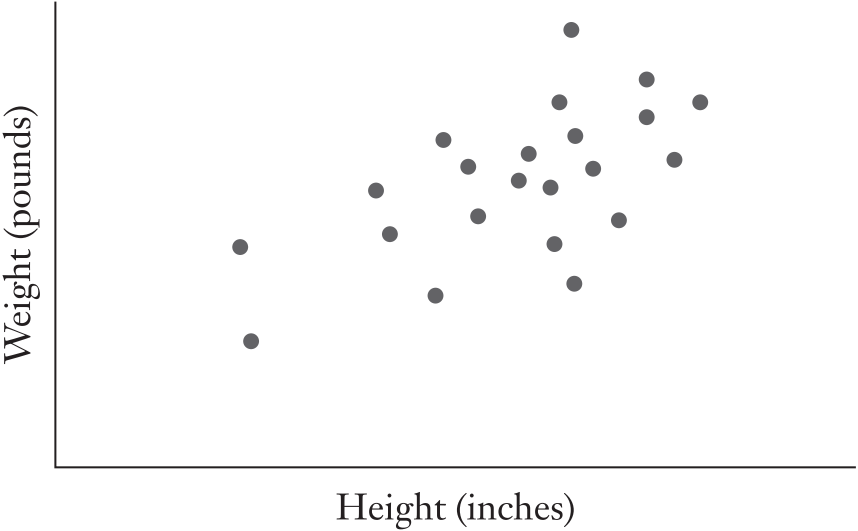
Difference in Sample Means
(Measured in Standard Errors)

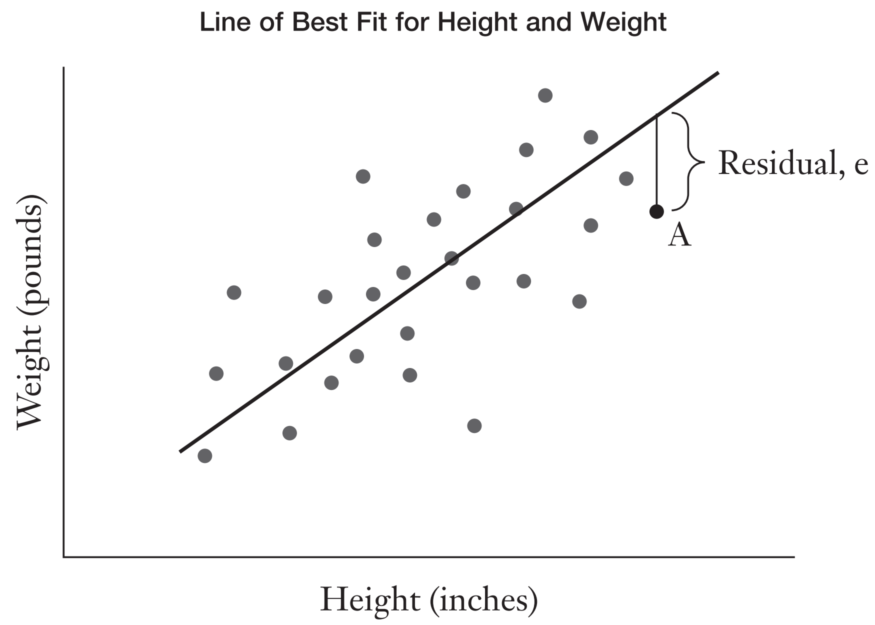


Difference in Sample Means (Measured in Standard Errors)



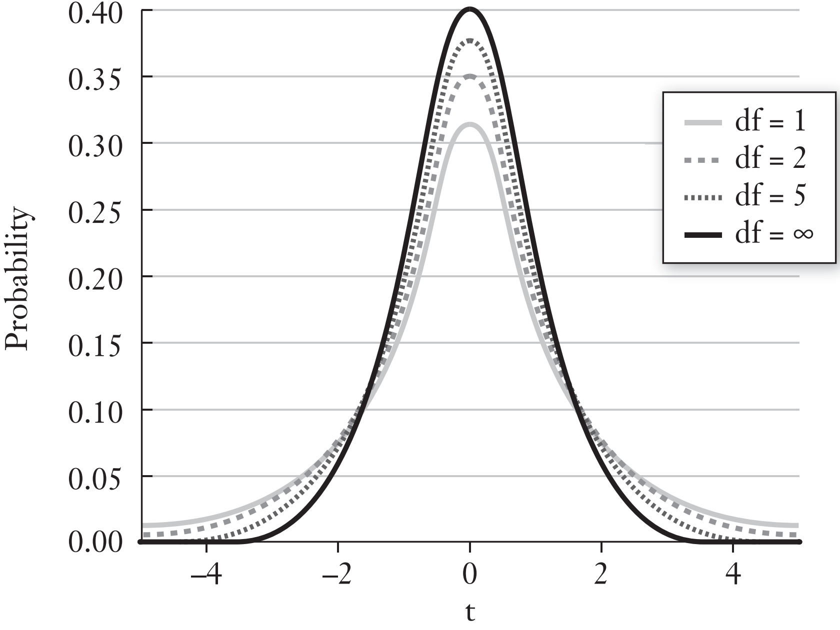
Scatter Plot for Height and Weight




Line of Best Fit for Height and Weight
Residual, e
A
Weight (pounds)
Height (inches)

$$\text{WEIGHT} = -145 + 4.6 \times (\text{HEIGHT IN INCHES}) + .1 \times (\text{AGE IN YEARS})$$

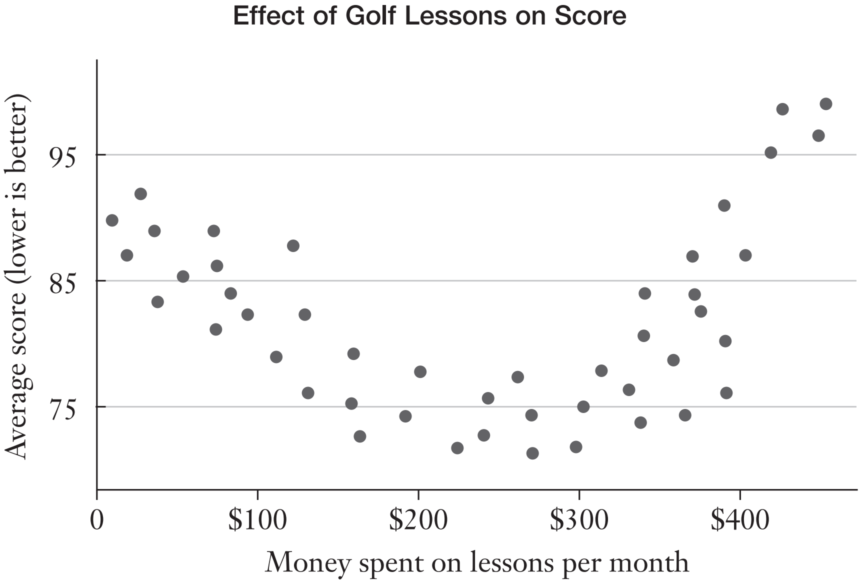
$$\text{WEIGHT} = -118 + 4.3 \times (\text{HEIGHT IN INCHES}) + .12\ (\text{AGE IN YEARS}) - 4.8\ (\text{IF SEX IS FEMALE})$$



0.40
0.35
0.30
0.25
0.20
0.15
0.10
0.05
0.00
Probability
–4
–2
0
2
4
t
df = 1
df = 2
df = 5
df = ∞

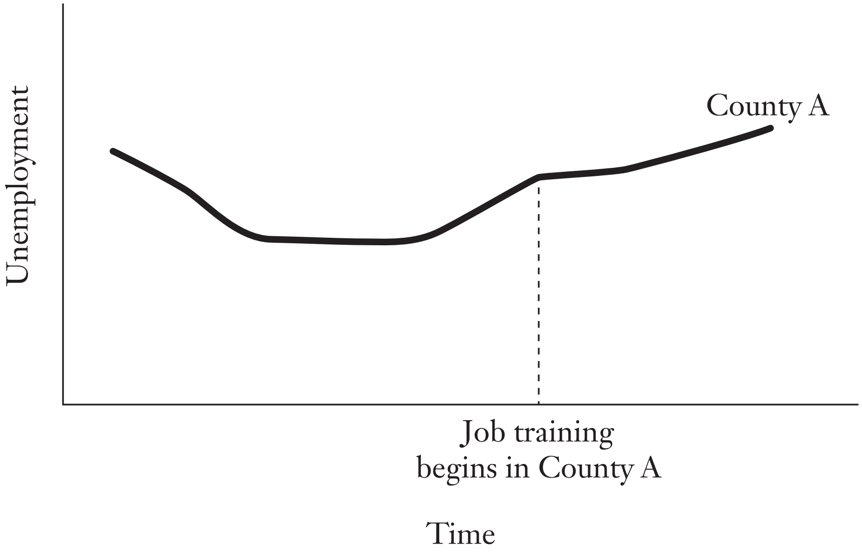
Regression Equation for Weight

Variable	Coefficient	Standard Error	t-statistic	p-value (two-tailed test)	95% Confidence Interval
Height	4.4	.2	21.4	.000	4.0 to 4.8
Age	.08	.03	2.2	.026	.01 to .2
Sex	–5.7	1.7	–3.4	.001	–9.0 to –2.4
Years of Educational Attainment	–.7	.2	–3.5	.000	–1.1 to –.3
Bottom Quintile of Physical Activity	3.7	1.4	2.6	.009	.9 to 6.5
Dummy for Receiving Food Stamps	5.6	2.1	2.7	.007	1.5 to 9.7
Non-Hispanic Black	9.7	1.3	7.2	.000	7.0 to 12.3
Intercept	–117				



* There are more sophisticated methods that can be used to adapt regression analysis for use with nonlinear data. Before using those tools, however, you need to appreciate why using the standard ordinary least squares approach with nonlinear data will give you a meaningless result.

Effect of Job Training on Unemployment in County A



Effect of Job Training on Unemployment in County A, with County B as a Comparison