2. REVIEW of PROBABILITY AND STATISTICS
(CHAP 2 - 3)

[1] Important Concepts and Formulas

(1) **Population**: The group of interest.

EX: The heights of Phoenix residents, Household incomes of Phoenix residents, US GNP.

(2) **Probability** (or frequency):

- A small island with 12 households

Income (per day)	# of households	Probability
\$100	2	2/12 = 1/6
\$200	2	2/12 = 1/6
\$300	4	4/12 = 1/3
\$400	4	4/12 = 1/3
	12	1

- Let X = a household's income (X: **random variable**)
- Describe the probability that X takes a specific value by:

  $f(x) = 1/6$, if $x = 100$ or 200; $f(x) = 1/3$ if $x = 300$ or 400. (**probability density function**.)

(3) **Expected value (Population mean) of X:**

- In the above example, the population mean is

  $(100\times2+200\times2+300\times4+400\times4)/12 = 283.3$

- Expected value of X: $E(x) \equiv \mu_x \equiv \Sigma_x xf(x)$:

  $E(x) = 100\times(1/6) + 200\times(1/6) + 300\times(1/3) + 400\times(1/3) = 283.3$.

- Lesson: If you know possible values of x and f(x), can compute the population mean.

(4) Population Variance of X

- Wish to know the dispersion of a population of size B:

  Let x_1, ... , x_B be the members of population. Then, use $var(x) = \frac{1}{B}\Sigma_{i=1}^{B}(x_i - \mu_x)^2$.

- Alternatively, you compute: $\text{var}(x) \equiv \sigma_x^2 = \Sigma_x (x - \mu_x)^2 f(x) = \Sigma_x x^2 f(x) - \mu_x^2$.

- In the above example,

$$\text{var}(x) = \Sigma_{i=1}^{12}(x_i - \mu_x)^2 / 12$$

$$= \{(100\text{-}283.3)^2 + (100\text{-}283.3)^2 + (200\text{-}283.3)^2 + (200\text{-}283.3)^2$$
$$+ (300\text{-}283.3)^2 + (300\text{-}283.3)^2 + (300\text{-}283.3)^2 + (300\text{-}283.3)^2$$
$$+ (400\text{-}283.3)^2 + (400\text{-}283.3)^2 + (400\text{-}283.3)^2 + (400\text{-}283.3)^2\}/12 = 11388.889.$$

$$\text{var(x)} = \Sigma_x(x\text{-}\mu_x)^2 f(x)$$
$$= (100\text{-}283.3)^2*(1/6) + (200\text{-}283.3)^2*(1/6) + (300\text{-}283.3)^2*(1/3) + (400\text{-}283.3)^2*(1/3)$$
$$= 11388.889.$$

(5) Case of Two Random Variables

EX: Income (X) and consumption (Y) of the 12 households.

Y($)\X($)	100	200	300	400		
30	1	1	2	1	:	5
40	1	0	1	1	:	3
50	0	1	1	2	:	4
	2	2	4	4	:	12

1. **Joint Probability Density Function**

Y\X	100	200	300	400		
30	1/12	1/12	2/12	1/12	:	5/12
40	1/12	0	1/12	1/12	:	3/12
50	0	1/12	1/12	2/12	:	4/12
	2/12	2/12	4/12	4/12	:	1

Joint pdf = f(x,y): f(100,50) = 0

2. **Marginal PDFs of X and Y**:

Marginal pdf of X = $f_x(x) = \Sigma_y f(x,y)$ = Pr(X=x) regardless of Y.

Marginal pdf of Y = $f_y(y) = \Sigma_x f(x,y)$ = Pr(Y=y) regardless of X.

Y\X	100	200	300	400		$f_y(y)$
30	1/12	1/12	2/12	1/12	:	5/12
40	1/12	0	1/12	1/12	:	3/12
50	0	1/12	1/12	2/12	:	4/12
$f_x(x)$	2/12	2/12	4/12	4/12	:	1

3. **Conditional pdf**:

f(y|x) = Pr(Y = y, given X = x) = $\dfrac{f(x,y)}{f_x(x)}$; f(x|y) = Pr(X = x, given Y = y) = $\dfrac{f(x,y)}{f_y(y)}$.

- f(y=30|x=100) = $\dfrac{f(100,30)}{f_x(100)} = \dfrac{1/12}{2/12} = \dfrac{1}{2}$; f(y=40|x=300) = $\dfrac{f(300,40)}{f_x(300)} = \dfrac{1/12}{4/12} = \dfrac{1}{4}$.

4. **Population means and variances of X and Y**

$E(x) = \Sigma_x x f_x(x)$; $E(y) = \Sigma_y y f_y(y)$;

$var(x) = \Sigma_x (x-\mu_x)^2 f_x(x)$; $var(y) = \Sigma_y (y-\mu_y)^2 f_y(y)$

5. **Conditional means and conditional variances**

$E(x|y) = \Sigma_x x f(x|y)$; $E(y|x) = \Sigma_y y f(y|x)$;

$var(x|y) = \Sigma_x [x-E(x|y)]^2 f(x|y)$; $var(y|x) = \Sigma_y [y-E(y|x)]^2 f(y|x)$

EX:

$$E(y \mid x = 200) = \Sigma_y y f(y \mid x = 200)$$

$$= 30 \times \frac{f(200,30)}{f_x(200)} + 40 \times \frac{f(200,40)}{f_x(200)} + 50 \times \frac{f(200,50)}{f_x(200)}$$

$$= 30 \times \frac{1/12}{2/12} + 40 \times \frac{0}{2/12} + 50 \times \frac{1/12}{2/12} = 40$$

$$\mathrm{var}(y \mid x = 200) = \Sigma_y (y - E(y \mid x = 200))^2 f(y \mid x = 200)$$

$$= (30-40)^2 \times \frac{1/12}{2/12} + (40-40)^2 \times \frac{0}{2/12} + (50-40)^2 \times \frac{1/12}{2/12} = 100$$

6. **Stochastic Independence:**

X and Y are stochastically independent iff (if and only if) $f(x,y) = f_x(x)f_y(y)$, for all x and y.

- In the above example,

$$f(300,30) = \frac{2}{12} = \frac{24}{144}.$$

$$f_x(300) = \frac{4}{12};\ f_y(30) = \frac{5}{12} \rightarrow f_x(300)f_y(30) = \frac{20}{144}.$$

- If X and Y are stoch. independent, $E(xy) = E(x)E(y)$.

If X and Y are stoch. dependent, $E(xy) \neq E(x)E(y)$, generally.

7. **Covariance:**

$cov(x,y) = E[(x-\mu_x)(y-\mu_y)] = \Sigma_x\Sigma_y(x-\mu_x)(y-\mu_y)f(x,y)$.

which measures how much X and Y are linearly correlated.

$cov(x,y) > (<) 0$ positively (negatively) linearly related.

$cov(x,y) = 0$ no linear relation.

8. **Correlation**

$$corr(x,y) = \frac{\operatorname{cov}(x,y)}{\sqrt{\operatorname{var}(x)\operatorname{var}(y)}} \equiv \frac{\sigma_{xy}}{\sigma_x \sigma_y}.$$

- $-1 \le$ corr(x,y) ≤ 1:
  - corr(x,y) $\rightarrow$ 1: highly positively linearly related.

    corr(x,y) $\rightarrow$ -1: highly negatively linearly related

    corr(x,y) $\rightarrow$ 0: no linear relation.
- If X & Y are stochastically independent, then, corr(x,y) = 0, but not vice versa.

9. **Skewness and Kurtosis**

See book, pp. 26 – 29.

10. **Some useful facts**: (p. 38, 63)

- $E(a+bx+cy) = a + bE(x) + cE(y)$.
- $\text{var}(a+by) = b^2\text{var}(y)$.
- $\text{var}(x) = E(x^2) - [E(x)]^2$
- $\text{cov}(x,y) = E(xy) - E(x)E(y)$.
- $\text{var}(ax+by) = a^2\,\text{var}(x) + 2ab\,\text{cov}(x,y) + b^2\,\text{var}(y)$.

Exercise:

Y\X	100	200	300	400		$f_y(y)$
30	1/12	1/12	2/12	1/12	:	5/12
40	1/12	0	1/12	1/12	:	3/12
50	0	1/12	1/12	2/12	:	4/12
$f_x(x)$	2/12	2/12	4/12	4/12	:	1

- Compute E(x), E(y), var(x), var(y), cov(x,y), E(1+2x+3y), var(2+3x), and var(x+2y).
- Compute $\text{cov}(x,y)$ and $E(xy) - E(x)E(y)$. Are they the same?

[2] Examples of pdf's

(1) Normal distribution

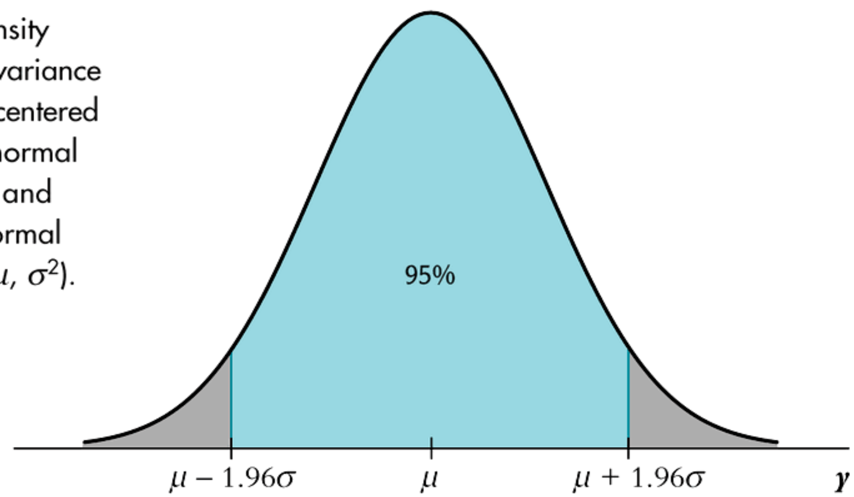
$X \sim N(\mu, \sigma^2)$, where E(x) = μ and var(x) = σ^2

1) pdf: $f(x) = \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{(x-\mu)^2}{2\sigma^2}} = \frac{1}{\sqrt{2\pi}\sigma} \exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right)$, $-\infty < x < \infty$.

2) f(x) is symmetric around x = μ

FIGURE 2.3 The Normal Probability Density

The normal probability density function with mean μ and variance σ^2 is a bell-shaped curve, centered at μ. The area under the normal p.d.f. between $\mu - 1.96\sigma$ and $\mu + 1.96\sigma$ is 0.95. The normal distribution is denoted $N(\mu, \sigma^2)$.



3) Pr(μ - σ < X < μ + σ) ≈ 0.68; Pr(μ - 1.96σ < X < μ + 1.96σ) = 0.95;

Pr(μ - 2.58σ < X < μ + 2.58σ) = 0.99.

4) **Standard Normal Distribution**: Z ~ N(0,1).

Pdf is given:

$$\phi(z) = \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{z^2}{2}\right), \text{-}\infty < z < \infty;\ \Phi(a) = \Pr(Z < a) = \int_{-\infty}^{a} \phi(z)dz\,.$$

5) Some useful facts: (p. 40, pp. 58 -63)

- Pr(a < Z < b) = $\Phi(b) - \Phi(a)$.
- Pr(a < Z) = 1 - $\Phi(a)$.

EX: Find Pr(1 < Z < 2) and Pr(Z > 0.5).

5) If X ~ N(μ,σ^2). Then, $Z = \dfrac{X-\mu}{\sigma}$ ~ N(0,1).

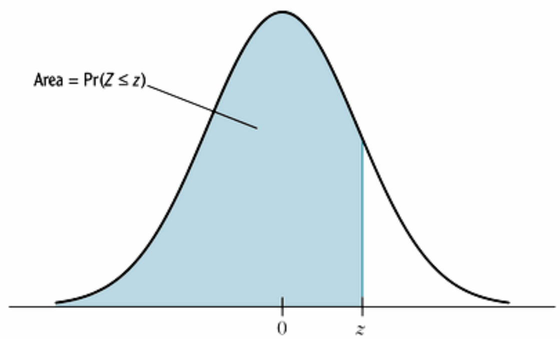
EX: Suppose X ~ N(1,4). Find Pr(X < 3).

SOL: $\Pr(X < 3) = \Pr\left(\frac{X-1}{\sqrt{4}} < \frac{3-1}{\sqrt{4}}\right) = \Pr(Z < 1)$.

6) How to read z-table [Appendix Table 1, pp. 755 - 756]:

The table describes Pr(Z < a) = b: a → b, or b → a.

TABLE 1 **The Cumulative Standard Normal Distribution Function, $\Phi(z)$ = Pr(Z ″ z)**



z	Second Decimal Value of z									
	0	1	2	3	4	5	6	7	8	9
−2.9	0.0019	0.0018	0.0018	0.0017	0.0016	0.0016	0.0015	0.0015	0.0014	0.0014
−2.8	0.0026	0.0025	0.0024	0.0023	0.0023	0.0022	0.0021	0.0021	0.0020	0.0019
−2.7	0.0035	0.0034	0.0033	0.0032	0.0031	0.0030	0.0029	0.0028	0.0027	0.0026
−2.6	0.0047	0.0045	0.0044	0.0043	0.0041	0.0040	0.0039	0.0038	0.0037	0.0036
−2.5	0.0062	0.0060	0.0059	0.0057	0.0055	0.0054	0.0052	0.0051	0.0049	0.0048
−2.4	0.0082	0.0080	0.0078	0.0075	0.0073	0.0071	0.0069	0.0068	0.0066	0.0064
−2.3	0.0107	0.0104	0.0102	0.0099	0.0096	0.0094	0.0091	0.0089	0.0087	0.0084
−2.2	0.0139	0.0136	0.0132	0.0129	0.0125	0.0122	0.0119	0.0116	0.0113	0.0110
−2.1	0.0179	0.0174	0.0170	0.0166	0.0162	0.0158	0.0154	0.0150	0.0146	0.0143
−2.0	0.0228	0.0222	0.0217	0.0212	0.0207	0.0202	0.0197	0.0192	0.0188	0.0183
−1.9	0.0287	0.0281	0.0274	0.0268	0.0262	0.0256	0.0250	0.0244	0.0239	0.0233
−1.8	0.0359	0.0351	0.0344	0.0336	0.0329	0.0322	0.0314	0.0307	0.0301	0.0294
−1.7	0.0446	0.0436	0.0427	0.0418	0.0409	0.0401	0.0392	0.0384	0.0375	0.0367
−1.6	0.0548	0.0537	0.0526	0.0516	0.0505	0.0495	0.0485	0.0475	0.0465	0.0455
−1.5	0.0668	0.0655	0.0643	0.0630	0.0618	0.0606	0.0594	0.0582	0.0571	0.0559
−1.4	0.0808	0.0793	0.0778	0.0764	0.0749	0.0735	0.0721	0.0708	0.0694	0.0681
−1.3	0.0968	0.0951	0.0934	0.0918	0.0901	0.0885	0.0869	0.0853	0.0838	0.0823
−1.2	0.1151	0.1131	0.1112	0.1093	0.1075	0.1056	0.1038	0.1020	0.1003	0.0985
−1.1	0.1357	0.1335	0.1314	0.1292	0.1271	0.1251	0.1230	0.1210	0.1190	0.1170
−1.0	0.1587	0.1562	0.1539	0.1515	0.1492	0.1469	0.1446	0.1423	0.1401	0.1379
−0.9	0.1841	0.1814	0.1788	0.1762	0.1736	0.1711	0.1685	0.1660	0.1635	0.1611

TABLE 1 *(continued)*

z	Second Decimal Value of z									
	0	1	2	3	4	5	6	7	8	9
−0.8	0.2119	0.2090	0.2061	0.2033	0.2005	0.1977	0.1949	0.1922	0.1894	0.1867
−0.7	0.2420	0.2389	0.2358	0.2327	0.2296	0.2266	0.2236	0.2206	0.2177	0.2148
−0.6	0.2743	0.2709	0.2676	0.2643	0.2611	0.2578	0.2546	0.2514	0.2483	0.2451
−0.5	0.3085	0.3050	0.3015	0.2981	0.2946	0.2912	0.2877	0.2843	0.2810	0.2776
−0.4	0.3446	0.3409	0.3372	0.3336	0.3300	0.3264	0.3228	0.3192	0.3156	0.3121
−0.3	0.3821	0.3783	0.3745	0.3707	0.3669	0.3632	0.3594	0.3557	0.3520	0.3483
−0.2	0.4207	0.4168	0.4129	0.4090	0.4052	0.4013	0.3974	0.3936	0.3897	0.3859
−0.1	0.4602	0.4562	0.4522	0.4483	0.4443	0.4404	0.4364	0.4325	0.4286	0.4247
−0.0	0.5000	0.4960	0.4920	0.4880	0.4840	0.4801	0.4761	0.4721	0.4681	0.4641
0.0	0.5000	0.5040	0.5080	0.5120	0.5160	0.5199	0.5239	0.5279	0.5319	0.5359
0.1	0.5398	0.5438	0.5478	0.5517	0.5557	0.5596	0.5636	0.5675	0.5714	0.5753
0.2	0.5793	0.5832	0.5871	0.5910	0.5948	0.5987	0.6026	0.6064	0.6103	0.6141
0.3	0.6179	0.6217	0.6255	0.6293	0.6331	0.6368	0.6406	0.6443	0.6480	0.6517
0.4	0.6554	0.6591	0.6628	0.6664	0.6700	0.6736	0.6772	0.6808	0.6844	0.6879
0.5	0.6915	0.6950	0.6985	0.7019	0.7054	0.7088	0.7123	0.7157	0.7190	0.7224
0.6	0.7257	0.7291	0.7324	0.7357	0.7389	0.7422	0.7454	0.7486	0.7517	0.7549
0.7	0.7580	0.7611	0.7642	0.7673	0.7704	0.7734	0.7764	0.7794	0.7823	0.7852
0.8	0.7881	0.7910	0.7939	0.7967	0.7995	0.8023	0.8051	0.8078	0.8106	0.8133
0.9	0.8159	0.8186	0.8212	0.8238	0.8264	0.8289	0.8315	0.8340	0.8365	0.8389
1.0	0.8413	0.8438	0.8461	0.8485	0.8508	0.8531	0.8554	0.8577	0.8599	0.8621
1.1	0.8643	0.8665	0.8686	0.8708	0.8729	0.8749	0.8770	0.8790	0.8810	0.8830
1.2	0.8849	0.8869	0.8888	0.8907	0.8925	0.8944	0.8962	0.8980	0.8997	0.9015
1.3	0.9032	0.9049	0.9066	0.9082	0.9099	0.9115	0.9131	0.9147	0.9162	0.9177
1.4	0.9192	0.9207	0.9222	0.9236	0.9251	0.9265	0.9279	0.9292	0.9306	0.9319
1.5	0.9332	0.9345	0.9357	0.9370	0.9382	0.9394	0.9406	0.9418	0.9429	0.9441
1.6	0.9452	0.9463	0.9474	0.9484	0.9495	0.9505	0.9515	0.9525	0.9535	0.9545
1.7	0.9554	0.9564	0.9573	0.9582	0.9591	0.9599	0.9608	0.9616	0.9625	0.9633
1.8	0.9641	0.9649	0.9656	0.9664	0.9671	0.9678	0.9686	0.9693	0.9699	0.9706
1.9	0.9713	0.9719	0.9726	0.9732	0.9738	0.9744	0.9750	0.9756	0.9761	0.9767
2.0	0.9772	0.9778	0.9783	0.9788	0.9793	0.9798	0.9803	0.9808	0.9812	0.9817
2.1	0.9821	0.9826	0.9830	0.9834	0.9838	0.9842	0.9846	0.9850	0.9854	0.9857
2.2	0.9861	0.9864	0.9868	0.9871	0.9875	0.9878	0.9881	0.9884	0.9887	0.9890
2.3	0.9893	0.9896	0.9898	0.9901	0.9904	0.9906	0.9909	0.9911	0.9913	0.9916
2.4	0.9918	0.9920	0.9922	0.9925	0.9927	0.9929	0.9931	0.9932	0.9934	0.9936
2.5	0.9938	0.9940	0.9941	0.9943	0.9945	0.9946	0.9948	0.9949	0.9951	0.9952
2.6	0.9953	0.9955	0.9956	0.9957	0.9959	0.9960	0.9961	0.9962	0.9963	0.9964
2.7	0.9965	0.9966	0.9967	0.9968	0.9969	0.9970	0.9971	0.9972	0.9973	0.9974
2.8	0.9974	0.9975	0.9976	0.9977	0.9977	0.9978	0.9979	0.9979	0.9980	0.9981
2.9	0.9981	0.9982	0.9982	0.9983	0.9984	0.9984	0.9985	0.9985	0.9986	0.9986

This table can be used to calculate Pr(Z " z) where Z is a standard normal variable. For example, when $z = 1.17$, this probability is 0.8790, which is the table entry for the row labeled 1.1 and the column labeled 7.

EX: $\Pr(Z < 1.96) = ?$

EX: $\Pr(Z < a) = 0.9463$. What is a?

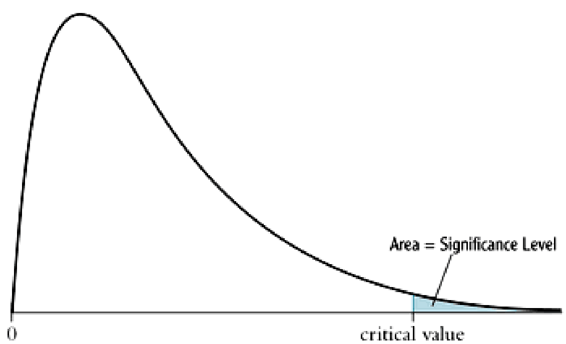
(2) χ^2 (chi-square) distribution

1) Let $Z_1, \ldots, Z_k$ be iid with N(0,1). Define:

$$Y = \Sigma_{i=1}^{k} Z_i^2 = Z_1^2 + Z_2^2 + \ldots + Z_k^2 .$$

Then, $Y \sim \chi^2(k)$, $Y > 0$. Here, k is called degrees of freedom.

2) The pdf is right-skewed except for $k \leq 2$.



3) $E(y) = k$; $var(y) = 2k$.

4) How to read χ^2-table [Appendix Table 3, p. 758].

First, need to know the degrees of freedom (df) for the RV of your interest. Given df, the table describes $\Pr(\chi^2 > a) = b$. Here, a is called "critical value" and b "significance level."

TABLE 3 **Critical Values for the χ^2 Distribution**

	Significance Level		
Degrees of Freedom	**10%**	**5%**	**1%**
1	2.71	3.84	6.63
2	4.61	5.99	9.21
3	6.25	7.81	11.34
4	7.78	9.49	13.28
5	9.24	11.07	15.09
6	10.64	12.59	16.81
7	12.02	14.07	18.48
8	13.36	15.51	20.09
9	14.68	16.92	21.67
10	15.99	18.31	23.21
11	17.28	19.68	24.72
12	18.55	21.03	26.22
13	19.81	22.36	27.69
14	21.06	23.68	29.14
15	22.31	25.00	30.58
16	23.54	26.30	32.00
17	24.77	27.59	33.41
18	25.99	28.87	34.81
19	27.20	30.14	36.19
20	28.41	31.41	37.57
21	29.62	32.67	38.93
22	30.81	33.92	40.29
23	32.01	35.17	41.64
24	33.20	36.41	42.98
25	34.38	37.65	44.31
26	35.56	38.89	45.64
27	36.74	40.11	46.96
28	37.92	41.34	48.28
29	39.09	42.56	49.59
30	40.26	43.77	50.89

This table contains the 90th, 95th, and 99th percentiles of the χ^2 distribution. These serve as critical values for tests with significance levels of 10%, 5%, and 1%.

EX: $\Pr(\chi^2 > a) = 0.1$, and df = 22. Find a.

EX: $\Pr(\chi^2 > 36.41) = b$, and df = 24. Find b.

EX: X and Y are stoch. indep., and N(0,1). Find $\Pr(X^2 + Y^2 < 1)$.

(3) **Student's t distribution**

1) Let Z ~ N(0,1), Y ~ χ^2(k); and let Z and Y be stochastically independent. Define:

$$T = \frac{Z}{\sqrt{Y/k}}.$$

Then, T ~ t(k), where k = df (degrees of freedom).

2) E(t) = 0, k > 1 ; var(t) = k/(k-2), k > 2.

3) As $k \to \infty$, var(t) $\to$ 1: In fact, $T \to Z$.

4) The pdf of T is similar to that of Z, but T has thicker tails.

5) How to read t-table [Appendix Table 2, p. 757]:

The table describes

$\Pr(T > a) = b$ and $\Pr(|T| > a) = b$.

Note that:

$\Pr(|T| > a) = 2 \times \Pr(T > a)$.

Figure 4.2

5% rejection rule for the alternative H_1: $\beta_j > 0$ with 28 *df*.

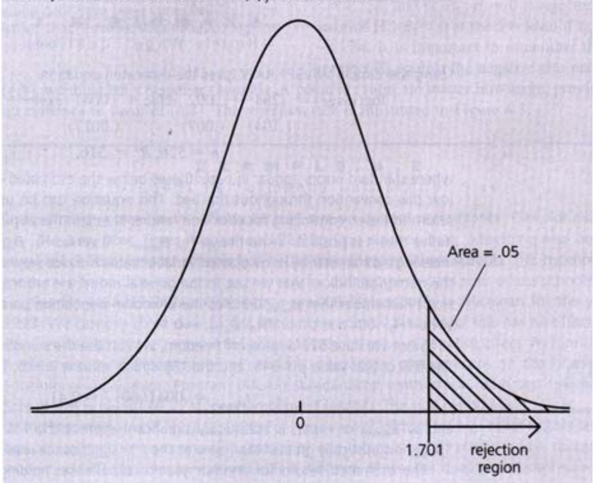


Figure 4.4

5% rejection rule for the alternative H_1: $\beta_j \neq 0$ with 25 *df*.

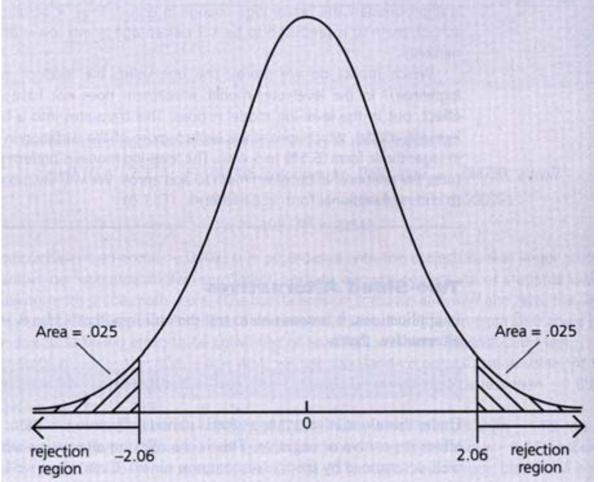


TABLE 2 Critical Values for 2-Sided and 1-Sided Tests Using the Student t Distribution

Degrees of Freedom	Significance Level				
	20% (2-Sided) 10% (1-Sided)	10% (2-Sided) 5% (1-Sided)	5% (2-Sided) 2.5% (1-Sided)	2% (2-Sided) 1% (1-Sided)	1% (2-Sided) 0.5% (1-Sided)
1	3.08	6.31	12.71	31.82	63.66
2	1.89	2.92	4.30	6.96	9.92
3	1.64	2.35	3.18	4.54	5.84
4	1.53	2.13	2.78	3.75	4.60
5	1.48	2.02	2.57	3.36	4.03
6	1.44	1.94	2.45	3.14	3.71
7	1.41	1.89	2.36	3.00	3.50
8	1.40	1.86	2.31	2.90	3.36
9	1.38	1.83	2.26	2.82	3.25
10	1.37	1.81	2.23	2.76	3.17
11	1.36	1.80	2.20	2.72	3.11
12	1.36	1.78	2.18	2.68	3.05
13	1.35	1.77	2.16	2.65	3.01
14	1.35	1.76	2.14	2.62	2.98
15	1.34	1.75	2.13	2.60	2.95
16	1.34	1.75	2.12	2.58	2.92
17	1.33	1.74	2.11	2.57	2.90
18	1.33	1.73	2.10	2.55	2.88
19	1.33	1.73	2.09	2.54	2.86
20	1.33	1.72	2.09	2.53	2.85
21	1.32	1.72	2.08	2.52	2.83
22	1.32	1.72	2.07	2.51	2.82
23	1.32	1.71	2.07	2.50	2.81
24	1.32	1.71	2.06	2.49	2.80
25	1.32	1.71	2.06	2.49	2.79
26	1.32	1.71	2.06	2.48	2.78
27	1.31	1.70	2.05	2.47	2.77
28	1.31	1.70	2.05	2.47	2.76
29	1.31	1.70	2.05	2.46	2.76
30	1.31	1.70	2.04	2.46	2.75
60	1.30	1.67	2.00	2.39	2.66
90	1.29	1.66	1.99	2.37	2.63
120	1.29	1.66	1.98	2.36	2.62
∞	**1.28**	**1.64**	**1.96**	**2.33**	**2.58**

Values are shown for the critical values for 2-sided ($\neq$) and 1-sided (>) alternative hypotheses. The critical value for the 1-sided (<) test is the negative of the 1-sided (>) critical value shown in the table. For example, 2.13 is the critical value for a 2-sided test with a significance level of 5% using the Student t distribution with 15 degrees of freedom.

EX: Pr(T > 1.72) = b and df = 21. Find b.

EX: Pr(|T| > 1.70) = b and df =30. Find b.

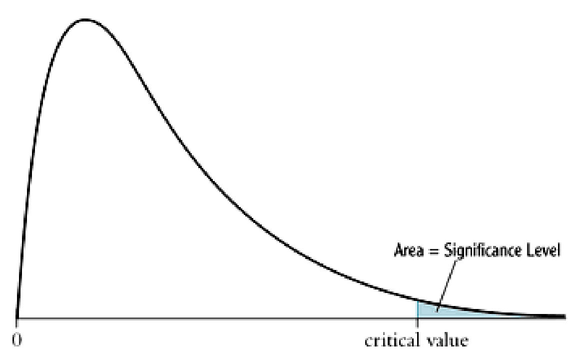
EX: Suppose X ~ N(0.1) and Y ~ $\chi^2(4)$ are stochastically independent. Find $\Pr(X < 1.5\sqrt{Y/4})$.

(4) **F (Fisher's) distribution**

1) Let $Y_1 \sim \chi^2(k_1)$ and $Y_2 \sim \chi^2(k_2)$ be stoch. indep.. Then, $M = \dfrac{Y_1/k_1}{Y_2/k_2} \sim F(k_1, k_2)$.

2) The pdf of F is right-skewed. F has a thicker tail than χ^2.

3) $F(k_1,\infty) = \chi^2(k_1)/k_1$.



4) How to read F-table [Appendix Table 5A-5C, pp. 760 – 762]: The table shows Pr(F > a) = b.

TABLE 5A Critical Values for the F_{n_1,n_2} Distribution—10% Significance Level

Denominator Degrees of Freedom (n_2)	Numerator Degrees of Freedom (n_1)									
	1	2	3	4	5	6	7	8	9	10
1	39.86	49.50	53.59	55.83	57.24	58.20	58.90	59.44	59.86	60.20
2	8.53	9.00	9.16	9.24	9.29	9.33	9.35	9.37	9.38	9.39
3	5.54	5.46	5.39	5.34	5.31	5.28	5.27	5.25	5.24	5.23
4	4.54	4.32	4.19	4.11	4.05	4.01	3.98	3.95	3.94	3.92
5	4.06	3.78	3.62	3.52	3.45	3.40	3.37	3.34	3.32	3.30
6	3.78	3.46	3.29	3.18	3.11	3.05	3.01	2.98	2.96	2.94
7	3.59	3.26	3.07	2.96	2.88	2.83	2.78	2.75	2.72	2.70
8	3.46	3.11	2.92	2.81	2.73	2.67	2.62	2.59	2.56	2.54
9	3.36	3.01	2.81	2.69	2.61	2.55	2.51	2.47	2.44	2.42
10	3.29	2.92	2.73	2.61	2.52	2.46	2.41	2.38	2.35	2.32
11	3.23	2.86	2.66	2.54	2.45	2.39	2.34	2.30	2.27	2.25
12	3.18	2.81	2.61	2.48	2.39	2.33	2.28	2.24	2.21	2.19
13	3.14	2.76	2.56	2.43	2.35	2.28	2.23	2.20	2.16	2.14
14	3.10	2.73	2.52	2.39	2.31	2.24	2.19	2.15	2.12	2.10
15	3.07	2.70	2.49	2.36	2.27	2.21	2.16	2.12	2.09	2.06
16	3.05	2.67	2.46	2.33	2.24	2.18	2.13	2.09	2.06	2.03
17	3.03	2.64	2.44	2.31	2.22	2.15	2.10	2.06	2.03	2.00
18	3.01	2.62	2.42	2.29	2.20	2.13	2.08	2.04	2.00	1.98
19	2.99	2.61	2.40	2.27	2.18	2.11	2.06	2.02	1.98	1.96
20	2.97	2.59	2.38	2.25	2.16	2.09	2.04	2.00	1.96	1.94
21	2.96	2.57	2.36	2.23	2.14	2.08	2.02	1.98	1.95	1.92
22	2.95	2.56	2.35	2.22	2.13	2.06	2.01	1.97	1.93	1.90
23	2.94	2.55	2.34	2.21	2.11	2.05	1.99	1.95	1.92	1.89
24	2.93	2.54	2.33	2.19	2.10	2.04	1.98	1.94	1.91	1.88
25	2.92	2.53	2.32	2.18	2.09	2.02	1.97	1.93	1.89	1.87
26	2.91	2.52	2.31	2.17	2.08	2.01	1.96	1.92	1.88	1.86
27	2.90	2.51	2.30	2.17	2.07	2.00	1.95	1.91	1.87	1.85
28	2.89	2.50	2.29	2.16	2.06	2.00	1.94	1.90	1.87	1.84
29	2.89	2.50	2.28	2.15	2.06	1.99	1.93	1.89	1.86	1.83
30	2.88	2.49	2.28	2.14	2.05	1.98	1.93	1.88	1.85	1.82
60	2.79	2.39	2.18	2.04	1.95	1.87	1.82	1.77	1.74	1.71
90	2.76	2.36	2.15	2.01	1.91	1.84	1.78	1.74	1.70	1.67
120	2.75	2.35	2.13	1.99	1.90	1.82	1.77	1.72	1.68	1.65

This table contains the 90th percentile of the F_{n_1,n_2} distribution, which serves as the critical values for a test with a 10% significance level.

TABLE 5B Critical Values for the F_{n_1,n_2} Distribution—5% Significance Level

Denominator Degrees of Freedom (n_2)	Numerator Degrees of Freedom (n_1)									
	1	2	3	4	5	6	7	8	9	10
1	161.40	199.50	215.70	224.60	230.20	234.00	236.80	238.90	240.50	241.90
2	18.51	19.00	19.16	19.25	19.30	19.33	19.35	19.37	19.39	19.40
3	10.13	9.55	9.28	9.12	9.01	8.94	8.89	8.85	8.81	8.79
4	7.71	6.94	6.59	6.39	6.26	6.16	6.09	6.04	6.00	5.96
5	6.61	5.79	5.41	5.19	5.05	4.95	4.88	4.82	4.77	4.74
6	5.99	5.14	4.76	4.53	4.39	4.28	4.21	4.15	4.10	4.06
7	5.59	4.74	4.35	4.12	3.97	3.87	3.79	3.73	3.68	3.64
8	5.32	4.46	4.07	3.84	3.69	3.58	3.50	3.44	3.39	3.35
9	5.12	4.26	3.86	3.63	3.48	3.37	3.29	3.23	3.18	3.14
10	4.96	4.10	3.71	3.48	3.33	3.22	3.14	3.07	3.02	2.98
11	4.84	3.98	3.59	3.36	3.20	3.09	3.01	2.95	2.90	2.85
12	4.75	3.89	3.49	3.26	3.11	3.00	2.91	2.85	2.80	2.75
13	4.67	3.81	3.41	3.18	3.03	2.92	2.83	2.77	2.71	2.67
14	4.60	3.74	3.34	3.11	2.96	2.85	2.76	2.70	2.65	2.60
15	4.54	3.68	3.29	3.06	2.90	2.79	2.71	2.64	2.59	2.54
16	4.49	3.63	3.24	3.01	2.85	2.74	2.66	2.59	2.54	2.49
17	4.45	3.59	3.20	2.96	2.81	2.70	2.61	2.55	2.49	2.45
18	4.41	3.55	3.16	2.93	2.77	2.66	2.58	2.51	2.46	2.41
19	4.38	3.52	3.13	2.90	2.74	2.63	2.54	2.48	2.42	2.38
20	4.35	3.49	3.10	2.87	2.71	2.60	2.51	2.45	2.39	2.35
21	4.32	3.47	3.07	2.84	2.68	2.57	2.49	2.42	2.37	2.32
22	4.30	3.44	3.05	2.82	2.66	2.55	2.46	2.40	2.34	2.30
23	4.28	3.42	3.03	2.80	2.64	2.53	2.44	2.37	2.32	2.27
24	4.26	3.40	3.01	2.78	2.62	2.51	2.42	2.36	2.30	2.25
25	4.24	3.39	2.99	2.76	2.60	2.49	2.40	2.34	2.28	2.24
26	4.23	3.37	2.98	2.74	2.59	2.47	2.39	2.32	2.27	2.22
27	4.21	3.35	2.96	2.73	2.57	2.46	2.37	2.31	2.25	2.20
28	4.20	3.34	2.95	2.71	2.56	2.45	2.36	2.29	2.24	2.19
29	4.18	3.33	2.93	2.70	2.55	2.43	2.35	2.28	2.22	2.18
30	4.17	3.32	2.92	2.69	2.53	2.42	2.33	2.27	2.21	2.16
60	4.00	3.15	2.76	2.53	2.37	2.25	2.17	2.10	2.04	1.99
90	3.95	3.10	2.71	2.47	2.32	2.20	2.11	2.04	1.99	1.94
120	3.92	3.07	2.68	2.45	2.29	2.18	2.09	2.02	1.96	1.91

This table contains the 95th percentile of the F_{n_1,n_2} distribution, which serves as the critical values for a test with a 5% significance level.

TABLE 5C Critical Values for the F_{n_1,n_2} Distribution—1% Significance Level

Denominator Degrees of Freedom (n_2)	Numerator Degrees of Freedom (n_1)									
	1	2	3	4	5	6	7	8	9	10
1	4052.00	4999.00	5403.00	5624.00	5763.00	5859.00	5928.00	5981.00	6022.00	6055.00
2	98.50	99.00	99.17	99.25	99.30	99.33	99.36	99.37	99.39	99.40
3	34.12	30.82	29.46	28.71	28.24	27.91	27.67	27.49	27.35	27.23
4	21.20	18.00	16.69	15.98	15.52	15.21	14.98	14.80	14.66	14.55
5	16.26	13.27	12.06	11.39	10.97	10.67	10.46	10.29	10.16	10.05
6	13.75	10.92	9.78	9.15	8.75	8.47	8.26	8.10	7.98	7.87
7	12.25	9.55	8.45	7.85	7.46	7.19	6.99	6.84	6.72	6.62
8	11.26	8.65	7.59	7.01	6.63	6.37	6.18	6.03	5.91	5.81
9	10.56	8.02	6.99	6.42	6.06	5.80	5.61	5.47	5.35	5.26
10	10.04	7.56	6.55	5.99	5.64	5.39	5.20	5.06	4.94	4.85
11	9.65	7.21	6.22	5.67	5.32	5.07	4.89	4.74	4.63	4.54
12	9.33	6.93	5.95	5.41	5.06	4.82	4.64	4.50	4.39	4.30
13	9.07	6.70	5.74	5.21	4.86	4.62	4.44	4.30	4.19	4.10
14	8.86	6.51	5.56	5.04	4.69	4.46	4.28	4.14	4.03	3.94
15	8.68	6.36	5.42	4.89	4.56	4.32	4.14	4.00	3.89	3.80
16	8.53	6.23	5.29	4.77	4.44	4.20	4.03	3.89	3.78	3.69
17	8.40	6.11	5.18	4.67	4.34	4.10	3.93	3.79	3.68	3.59
18	8.29	6.01	5.09	4.58	4.25	4.01	3.84	3.71	3.60	3.51
19	8.18	5.93	5.01	4.50	4.17	3.94	3.77	3.63	3.52	3.43
20	8.10	5.85	4.94	4.43	4.10	3.87	3.70	3.56	3.46	3.37
21	8.02	5.78	4.87	4.37	4.04	3.81	3.64	3.51	3.40	3.31
22	7.95	5.72	4.82	4.31	3.99	3.76	3.59	3.45	3.35	3.26
23	7.88	5.66	4.76	4.26	3.94	3.71	3.54	3.41	3.30	3.21
24	7.82	5.61	4.72	4.22	3.90	3.67	3.50	3.36	3.26	3.17
25	7.77	5.57	4.68	4.18	3.85	3.63	3.46	3.32	3.22	3.13
26	7.72	5.53	4.64	4.14	3.82	3.59	3.42	3.29	3.18	3.09
27	7.68	5.49	4.60	4.11	3.78	3.56	3.39	3.26	3.15	3.06
28	7.64	5.45	4.57	4.07	3.75	3.53	3.36	3.23	3.12	3.03
29	7.60	5.42	4.54	4.04	3.73	3.50	3.33	3.20	3.09	3.00
30	7.56	5.39	4.51	4.02	3.70	3.47	3.30	3.17	3.07	2.98
60	7.08	4.98	4.13	3.65	3.34	3.12	2.95	2.82	2.72	2.63
90	6.93	4.85	4.01	3.53	3.23	3.01	2.84	2.72	2.61	2.52
120	6.85	4.79	3.95	3.48	3.17	2.96	2.79	2.66	2.56	2.47

This table contains the 99th percentile of the F_{n_1,n_2} distribution, which serves as the critical values for a test with a 1% significance level.

EX: $\Pr(F > 3.52) = b$ with $k_1 = 4$ and $k_2 = 5$. Find b.

EX: $\Pr(F > a) = 0.01$ with $k_1 = 7$ and $k_2 = 9$. Find a.

EX: Suppose $X \sim \chi^2(3)$ and $Y \sim \chi^2(5)$ are stochastically independent. Find $\Pr\left(\frac{X}{3} < 2 \times \frac{Y}{5}\right)$.

[3] Statistical Inference

(1) Point Estimation

- Wish to know the population mean and variance of college graduates' hourly earnings (Y).
- Do not know pdf of Y
- **Unknown parameters**: The things a researcher wishes to estimate (such as population mean or population variance).
- Need **Sample of data** to estimate unknown parameters

1) Random sampling

- **Random sample** is a sample in which n objects are drawn at random from a population and each object is equally likely to be drawn.
  - Let Y_i be the value of the i'th randomly drawn object. Then, the random variables Y_1, ... , Y_n are said to be independent and identically distributed.
  - A random sample is a sample that can represent the population well.

- An example of nonrandom sampling:
  - Suppose you wish to estimate the % of supporters of the Republican Party in the Phoenix metropolitan area.
  - Choose people living in the street corners.
  - If you do, your sample is not random. Because rich people are likely to live in corner houses!

2) Definition of Sample Mean and Sample Variance

- Let $\{Y_1, \ldots, Y_n\}$ be a random sample ($Y_1, \ldots, Y_n$ are i.i.d.) from a population with population mean μ_Y and population variance σ_Y^2 Define sample mean and variance by

$$\bar{Y} = \frac{1}{n}(Y_1 + \ldots + Y_n) = \frac{1}{n}\Sigma_{i=1}^{n} Y_i ; s_Y^2 = \frac{1}{n-1}\Sigma_{i=1}^{n}\left(Y_i - \bar{Y}\right)^2 .$$

- The formulas $\bar{Y}$ and s_Y^2 are called "estimators". But the actual values of $\bar{Y}$ and s_Y^2 computed from actual data are called "estimates".

→ Estimator is an random variable in the sense that its outcome can change depending on sample chosen. Estimate is a nonrandom variable.

3) Sampling distributions of sample mean and sample variance

- Let $\{Y_1, \ldots , Y_n\}$ be a random sample ($Y_1, \ldots , Y_n$ are i.i.d.) from a population with population mean μ_Y and population variance σ_Y^2
- Consider the set of all possible random samples of size n:

			Estimates
Sample 1:	$\{Y_1^{[1]}, Y_2^{[1]}, \ldots , Y_n^{[1]}\}$	$\rightarrow$	$\bar{Y}^{[1]},\ s_Y^{2[1]}$
Sample 2:	$\{Y_1^{[2]}, Y_2^{[2]}, \ldots , Y_n^{[2]}\}$	$\rightarrow$	$\bar{Y}^{[2]},\ s_Y^{2[2]}$
Sample 3:	$\{Y_1^{[3]}, Y_2^{[3]}, \ldots , Y_n^{[3]}\}$	$\rightarrow$	$\bar{Y}^{[3]},\ s_Y^{2[3]}$
	:		:
Sample b:	$\{Y_1^{[b]}, Y_2^{[b]}, \ldots , Y_n^{[b]}\}$	$\rightarrow$	$\bar{Y}^{[b]},\ s_Y^{2[b]}$

- Consider the population of $\{\bar{Y}^{[1]}, \ldots, \bar{Y}^{[b]}\}$:
  - $E(\bar{Y}) = \mu_Y; E(s_Y^2) = \sigma_Y^2$.

    → We say that $\bar{Y}$ (s_Y^2) is an **unbiased estimator** of μ_Y (σ_Y^2).
  - If $\bar{Y}$ is computed from a nonrandom sample, it could be biased.
  - $\text{var}(\bar{Y}) = \dfrac{\sigma_Y^2}{n}$; $\text{var}(s_Y^2) = \dfrac{2\sigma_Y^4}{n-1}$.
  - If the population is normally distributed, so is the sampling distribution of the sample mean $\bar{Y}$: $\bar{Y} \sim N\left(\mu_Y, \dfrac{\sigma_Y^2}{n}\right)$.
  - What would be the sampling distribution of the sample mean if the population is not normal?

- **Efficiency** of $\overline{Y}$
  - Let $\tilde{Y}$ be an estimator of μ_Y other than $\overline{Y}$. Can we find $\tilde{Y}$ such that $\text{var}(\tilde{Y}) < \text{var}(\overline{Y})$? [If $\text{var}(\tilde{Y}) < \text{var}(\overline{Y})$, $\tilde{Y}$ must be more reliable (efficient) estimator than $\overline{Y}$.]
  - Consider $\tilde{Y} = a_1Y_1 + a_2Y_2 + ... + a_nY_n$ for some nonnegative a_1, ... , a_n such that $a_1 + ... + a_n = 1$. The estimators of this form are called "linear unbiased estimators."
  - It can be shown that $\text{var}(\tilde{Y})$ is minimized when $a_1 = ... = a_n = 1/n$.

    → This means that $\overline{Y}$ is the minimum-variance estimator among linear unbiased estimators.

    → $\overline{Y}$ is called the "best linear unbiased estimator" (BLUE).

    → If the population is normally distributed, $\overline{Y}$ and s_Y^2 are the minimum-variance unbiased estimators.
  - Least Square Estimator: $\overline{Y}$ = value of m which min. $\Sigma_{i=1}^{n}(Y_i - m)^2$

[Some Math Exercises]

If random sample is used, $E(\overline{Y}) = \mu_Y$; $E(s_Y^2) = \sigma_Y^2$

<Proof>

$$E(\overline{Y}) = E\left(\frac{1}{n}Y_1 + \frac{1}{n}Y_2 + \ldots + \frac{1}{n}Y_n\right) = \frac{1}{n}E(Y_1) + \ldots + \frac{1}{n}E(Y_n) = \frac{1}{n}\mu_Y + \ldots + \frac{1}{n}\mu_Y = \frac{1}{n}n\mu_Y = \mu_Y .$$

$$Var(\overline{X}) = Var\left(\frac{1}{n}Y_1 + \frac{1}{n}Y_2 + \ldots + \frac{1}{n}Y_n\right) = \left(\frac{1}{n}\right)^2 Var(Y_1) + \ldots + \left(\frac{1}{n}\right)^2 Var(Y_n) = \left(\frac{1}{n}\right)^2 n\sigma_Y^2 = \frac{\sigma_Y^2}{n}$$

Observe that:

$$\Sigma_{i=1}^n (Y_i - \overline{Y})^2 = \Sigma_{i=1}^n [(Y_i - \mu_Y) - (\overline{Y} - \mu_Y)]^2 = \Sigma_{i=1}^n (Y_i - \mu_Y)^2 - n(\overline{Y} - \mu_Y)^2$$

Thus,

$$\begin{aligned} E(\Sigma_{i=1}^n (Y_i - \overline{Y})^2) &= E(\Sigma_{i=1}^n (Y_i - \mu_Y)^2 - n(\overline{Y} - \mu_Y)^2) = \Sigma_{i=1}^n E[(Y_i - \mu_Y)^2] - nE[(\overline{Y} - \mu_Y)^2] \\ &= \Sigma_{i=1}^n \sigma_Y^2 - n(\sigma_Y^2 / n) = (n-1)\sigma_Y^2 . \end{aligned}$$

Thus,

$$E(s_Y^2) = E\left(\frac{1}{n-1}\Sigma_{i=1}^n (Y_i - \overline{Y})^2\right) = \sigma_Y^2 .$$

4) Sampling distribution of $\overline{Y}$ and s_Y^2 when n is large.

- **The Law of Large Numbers:**

  Under some general conditions, $\lim_{n\to\infty} Pr[\overline{Y} \textit{ is in the very close neighbor of } \mu_Y] = 1$.

  → We say that $\overline{Y}$ is a consistent estimator of μ_Y.

  → The sample variance, s_Y^2 is also a consistent estimator of σ_Y^2.

[Intuition for the consistency of $\overline{Y}$]

- $\mathrm{var}(\overline{Y})$ measures the average deviation of $\overline{Y}$ from μ_Y.
- Observe that $\mathrm{var}(\overline{Y}) = \dfrac{\sigma_Y^2}{n} \to 0$, as n → ∞.

- **The Central Limit Theorem:**

  Under some general conditions,

$$\frac{\overline{Y} - \mu_Y}{\sqrt{\mathrm{var}(\overline{Y})}} = \frac{\overline{Y} - \mu_Y}{\sqrt{\sigma_Y^2 / n}} \to N(0,1), \text{ as n} \to \infty.$$

→ Even if the population is not normally distributed, the sampling distribution of $\overline{Y}$ is roughly normal, when n is large.

→ We say that $\overline{Y}$ is **asymptotically normally distributed**.

(2) **Confidence Interval for Population Mean**

- What would be the possible range for the true value of μ_Y?
- Confidence interval: the range between lower and upper bounds that can contain the true μ_Y.
- Confidence level (1-α): Prespecified probability that the confidence interval contains the true μ_Y (90%, 95%, 99%).

1) Standard Error of $\bar{Y}$:

$$SE(\bar{Y}) = \sqrt{Estimated\ \text{var}(\bar{Y})} = \sqrt{\frac{s_Y^2}{n}} = \frac{s_Y}{\sqrt{n}}$$, where s_Y is called sample standard deviation.

2) The Central Limit Theorem

When n is large, $\dfrac{\bar{Y} - \mu_Y}{SE(\bar{Y})} \approx N(0,1)$.

If Y ~ $N(\mu_Y, \sigma_Y^2)$, $$\frac{\bar{Y} - \mu_Y}{\sqrt{s_Y^2 / n}} = \frac{(\bar{Y} - \mu_Y)/\sqrt{\sigma_Y^2 / n}}{\sqrt{s_Y^2 / n} / \sqrt{\sigma_Y^2 / n}} = \frac{\dfrac{\bar{Y} - \mu_Y}{\sqrt{\sigma_Y^2 / n}}}{\sqrt{\dfrac{s_Y^2}{\sigma_Y^2}}} = \frac{N(0,1)}{\sqrt{\chi^2(n-1)/(n-1)}} \sim t(n-1)$$

<Insert a standard normal pdf graph>

$$\rightarrow \quad \Pr\left(-1.96 < \frac{\bar{Y} - \mu_Y}{SE(\bar{Y})} < 1.96\right) = 0.95. \rightarrow \Pr\left(-1.96 < \frac{\mu_Y - \bar{Y}}{SE(\bar{Y})} < 1.96\right) = 0.95$$

$$\rightarrow \quad \Pr\left(-1.96SE(\bar{Y}) < \mu_Y - \bar{Y} < 1.96SE(\bar{Y})\right) = 0.95.$$

$$\rightarrow \quad \Pr\left(\bar{Y} - 1.96SE(\bar{Y}) < \mu_Y < \bar{Y} + 1.96SE(\bar{Y})\right) = 0.95.$$

(3) **Testing hypothesis regarding the population mean**

- Let Y be a random variable denoting a college graduate's hourly earnings. Someone claims that E(Y) = \$20. How can I test for this hypothesis?
  - Null hypothesis: H_o: E(Y) = 20.

    Alternative hypothesis: H_1: E(Y) $\neq$ 20 (two-sided alternative).

- Test procedure for H_o: E(Y) = $\mu_{Y,0}$ against H_1: E(Y) $\neq$ $\mu_{Y,0}$

  STEP 1: Determine the significance level (α) (Usually, 5 or 1%)

  STEP 2: From the z-table, find the critical value (c).

  c = 1.96 if α = 5%.

  STEP 3: Compute the t-statistic:

$$t = \frac{\bar{Y} - \mu_{Y,0}}{SE(\bar{Y})}.$$

  STEP 4: If |t| > c, reject H_o in favor of H_1. If |t| < c, do not reject H_o.

  Why? < Insert a standard normal pdf graph>

- Why the above statistic is called a "t-statistic"?
  - If the population is normally distributed, the sampling distribution of the t-statistic follows a t-distribution with degrees of freedom equal to (n-1).

EXAMPLE:

Y ~ N(μ_Y, σ_Y^2). From a sample of size n = 21, you obtained $\Sigma_{i=1}^{n} Y_i = 21$ and $\Sigma_{i=1}^{n}(Y_i - \overline{Y})^2 = 420$. Test H_o: $\mu_Y = 4$ against H_1: $\mu_Y \neq 4$ with the significance level of 5%.

[Solution]

STEP 1: α = 5%.

STEP 2: From the z-table, c = 1.96.

STEP 3: $\overline{Y} = \frac{1}{n}\Sigma_{i=1}^{n} Y_i = \frac{21}{21} = 1$; $s_Y^2 = \frac{1}{n-1}\Sigma_{i=1}^{n}(Y_i - \overline{Y})^2 = \frac{420}{20} = 21$.

$$t = \frac{\overline{Y} - \mu_{Y,0}}{SE(\overline{Y})} = \frac{\overline{Y} - \mu_{Y,0}}{\sqrt{s_Y^2 / n}} = \frac{1-4}{\sqrt{21/21}} = -3.$$

STEP 4: Since |t| = 3 > 1.96, reject H_o.

EXAMPLE:

Y ~ N(μ_Y, σ_Y^2). From a sample of size n = 100, you obtained $\Sigma_{i=1}^n Y_i = 590$ and $\Sigma_{i=1}^n \left(Y_i - \overline{Y}\right)^2 = 420$.

Test H_o: $\mu_Y = 6$ against $\mu_Y \neq 6$ with the 5% significance level.

[Solution]

STEP 1: $\alpha = 5\%$.

STEP 2: From the z-table, c = 1.96.

STEP 3: $\overline{Y} = \dfrac{590}{100} = 5.9$; $s_Y^2 = \dfrac{420}{99} = 4.24$;

$$t = \frac{5.9 - 6}{\sqrt{4.24/100}} = -0.49.$$

STEP 4: Since $|t| < 1.96$, do not reject H_o.

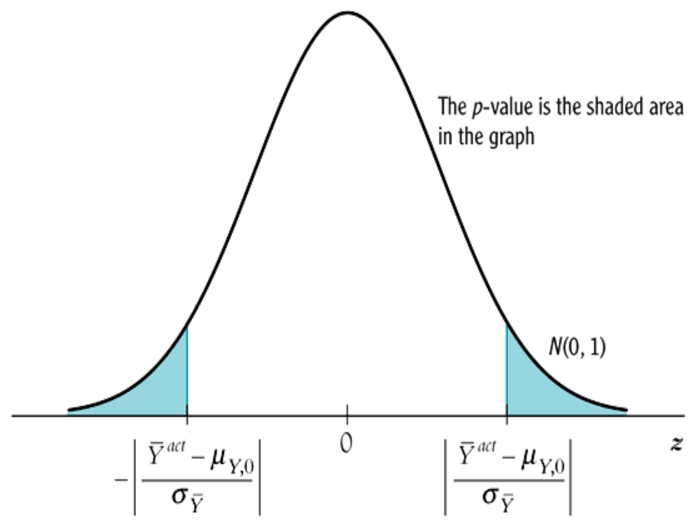
- P-value:
  - The minimum significance level at which the null hypothesis can be rejected.

    → If $p > \alpha$, do not reject H_o. If $p < \alpha$, reject H_o.

    → p-value = $2\times\Pr(Z > |t|) = 2\times(1-\Pr(Z < |t|) = 2\times(1-\Phi(|t|))$.

FIGURE 3.1 Calculating a *p*-value

The *p*-value is the probability of drawing a value of $\bar{Y}$ that differs from $\mu_{Y,0}$ by at least as much as $\bar{Y}^{act}$. In large samples, $\bar{Y}$ is distributed $N(\mu_{Y,0}, \sigma^2_{\bar{Y}})$ under the null hypothesis, so $(\bar{Y} - \mu_{Y,0})/\sigma_{\bar{Y}}$ is distributed $N(0, 1)$. Thus the *p*-value is the shaded standard normal tail probability outside $\pm|(\bar{Y}^{act} - \mu_{Y,0})/\sigma_{\bar{Y}}|$.



- Test procedure for H_o: $E(Y) = \mu_{Y,0}$ against H_1: $E(Y) > \mu_{Y,0}$ (One tail)

STEP 1: Determine the significance level (α) (Usually, 5 or 1%)

STEP 2: From the z-table, find the critical value (c).

$c = 1.645$ if $\alpha = 5\%$.

STEP 3: Compute the t-statistic:

$$t = \frac{\bar{Y} - \mu_{Y,0}}{SE(\bar{Y})}.$$

STEP 4: If $t > c$, reject H_o in favor of H_1. If $t < c$, do not reject H_o.

< Insert a standard normal pdf graph>

→ p-value = Pr(Z > t-statistic) = 1- Φ(t).

EXAMPLE:

Y ~ N(μ_Y, σ_Y^2). From a sample of size n = 21, you obtained $\Sigma_{i=1}^{n} Y_i = 21$ and $\Sigma_{i=1}^{n}(Y_i - \overline{Y})^2 = 420$. Test H_o: $\mu_Y = 0$ against H_1: $\mu_Y > 0$ at 5% of significance level.

[Solution]

STEP 1: $\alpha = 5\%$.

STEP 2: From the z-table, c = 1.645.

STEP 3: $\overline{Y} = \frac{1}{n}\Sigma_{i=1}^{n} Y_i = \frac{21}{21} = 1$; $s_Y^2 = \frac{1}{n-1}\Sigma_{i=1}^{n}(Y_i - \overline{Y})^2 = \frac{420}{20} = 21$.

$$t = \frac{\overline{Y} - \mu_{Y,0}}{SE(\overline{Y})} = \frac{\overline{Y} - \mu_{Y,0}}{\sqrt{s_Y^2 / n}} = \frac{1-0}{\sqrt{21/21}} = 1.$$

STEP 4: Since t < 1.645, do not reject H_o.

- Test procedure for H_o: E(Y) = $\mu_{Y,0}$ against H_1: E(Y) < $\mu_{Y,0}$ (One tail)

  STEP 1: Determine the significance level (α) (Usually, 5 or 1%)

  STEP 2: From the z-table, find the critical value (c): c = 1.645 if α = 5%.

  STEP 3: Compute the t-statistic: $t = \frac{\overline{Y} - \mu_{Y,0}}{SE(\overline{Y})}$.

  STEP 4: If t < -c, reject H_o in favor of H_1. If t > -c, do not reject H_o.

< Insert a standard normal pdf graph>

→ P-value = Pr(Z < t-statistic) = Φ(t).

EXAMPLE:

Y ~ N(μ_Y, σ_Y^2). From a sample of size n = 21, you obtained $\Sigma_{i=1}^n Y_i = 21$ and $\Sigma_{i=1}^n (Y_i - \overline{Y})^2 = 420$. Test H_o: μ_Y = 4 against H_1: μ_Y < 4 with the significance level of 5%.

[Solution]

STEP 1: α = 5%.

STEP 2: From the z-table, c = 1.645.

STEP 3: $\overline{Y} = \frac{1}{n}\Sigma_{i=1}^n Y_i = \frac{21}{21} = 1$; $s_Y^2 = \frac{1}{n-1}\Sigma_{i=1}^n (Y_i - \overline{Y})^2 = \frac{420}{20} = 21$;

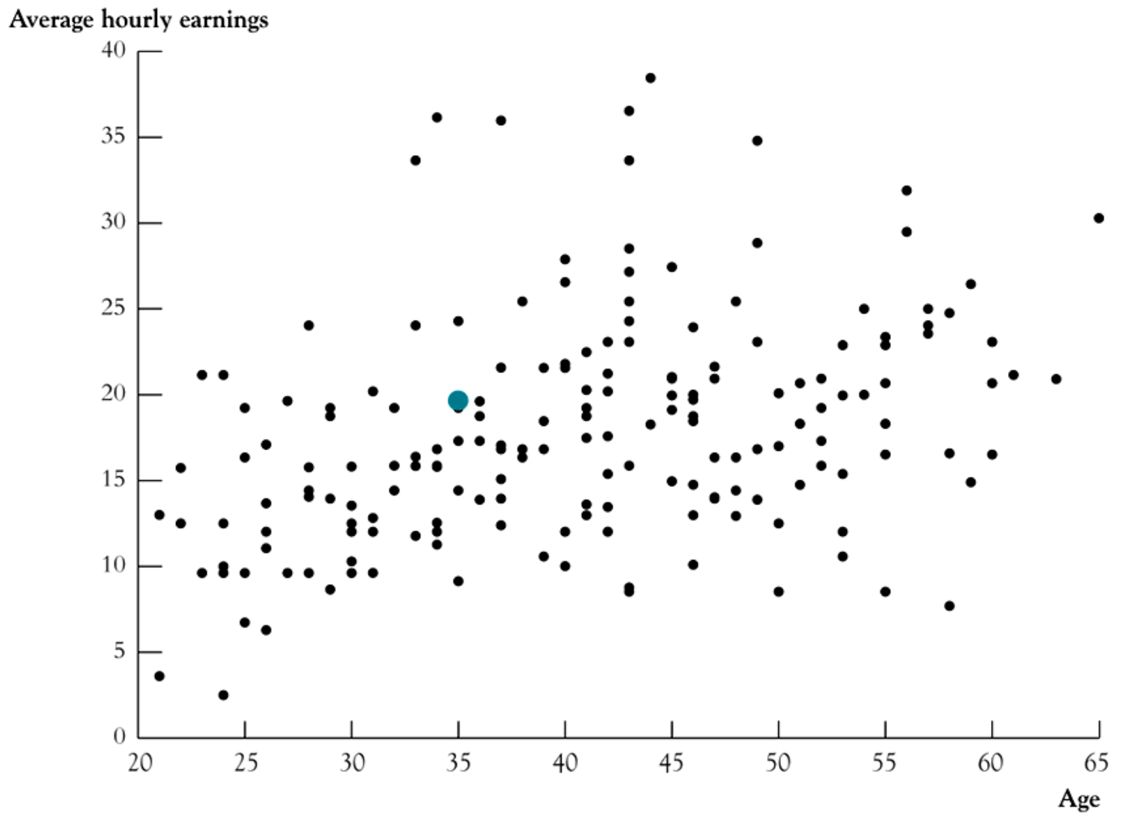
$$t = \frac{\overline{Y} - \mu_{Y,0}}{SE(\overline{Y})} = \frac{\overline{Y} - \mu_{Y,0}}{\sqrt{s_Y^2 / n}} = \frac{1-4}{\sqrt{21/21}} = -3.$$

STEP 4: Since t < - 1.645, reject H_o.

(4) Scatterplots, Sample Covariance and Sample Correlation

- Scatterplots

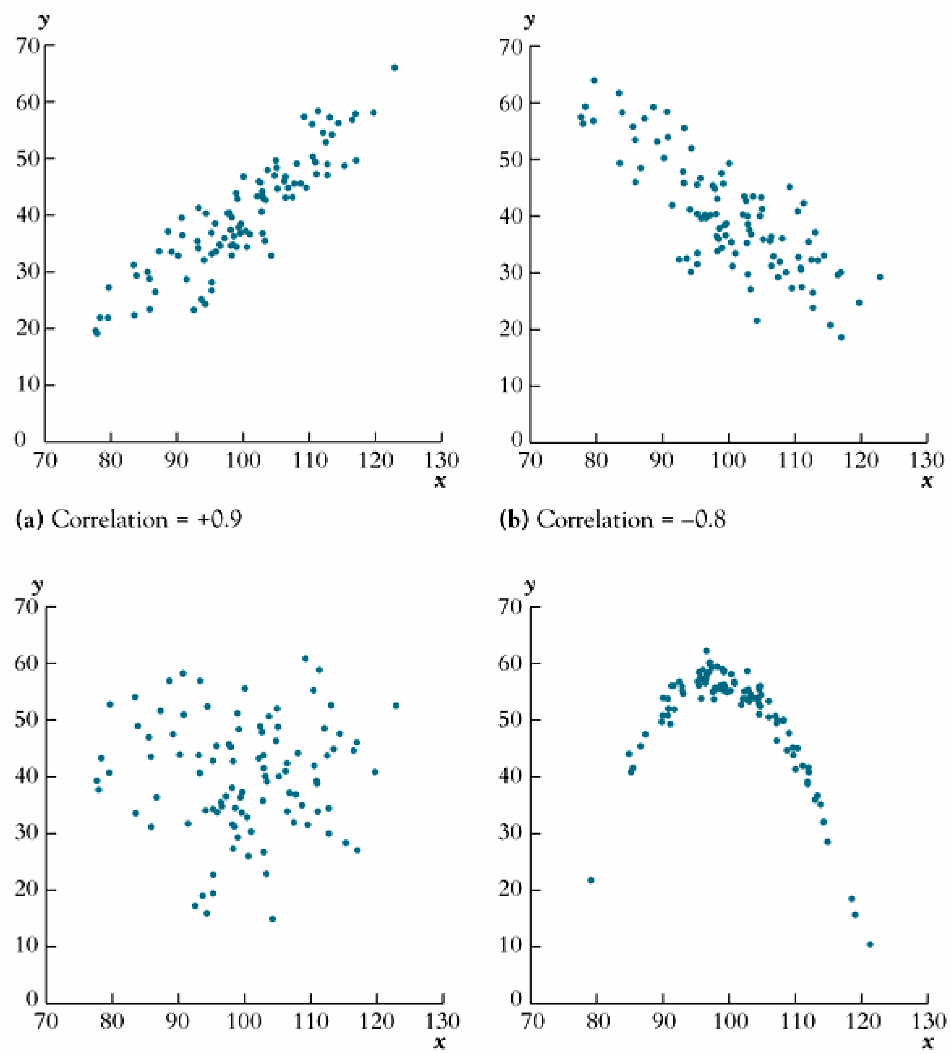
FIGURE 3.2 Scatterplot of Average Hourly Earnings vs. Age



Each point in the plot represents the age and average earnings of one of the 184 workers in the sample. The colored dot corresponds to a 35-year-old worker who earns $19.61 per hour. The data are for technicians in the communications industry without college degrees from the March 1999 CPS.

- Sample covariance:
  - Data: (X_1,Y_1), ... , (X_n,Y_n).
  - sample covariance: $s_{XY} = \frac{1}{n-1}\Sigma_{i=1}^{n}(X_i - \bar{X})(Y_i - \bar{Y})$.
- sample correlation: $r_{XY} = \frac{s_{XY}}{s_X s_Y} \rightarrow -1 \le r_{XY} \le 1$.

FIGURE 3.3 Scatterplots for Four Hypothetical Data Sets



(a) Correlation = +0.9

(b) Correlation = −0.8

(c) Correlation = 0.0

(d) Correlation = 0.0 (quadratic)

The scatterplots in Figures 3.3a and 3.3b show strong linear relationships between X and Y. In Figure 3.3c, X is independent of Y and the two variables are uncorrelated. In Figure 3.3d, the two variables also are uncorrelated even though they are related nonlinearly.