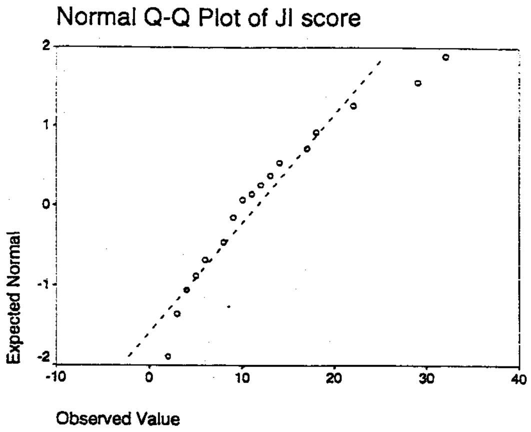
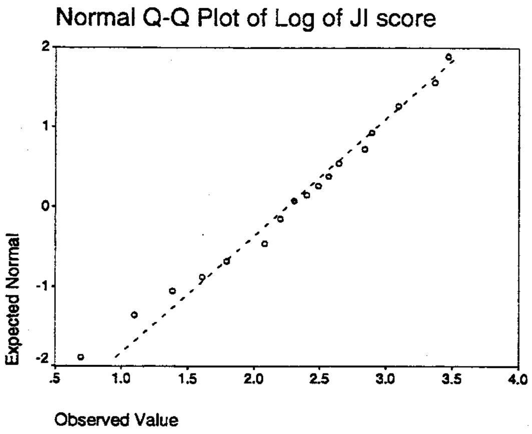
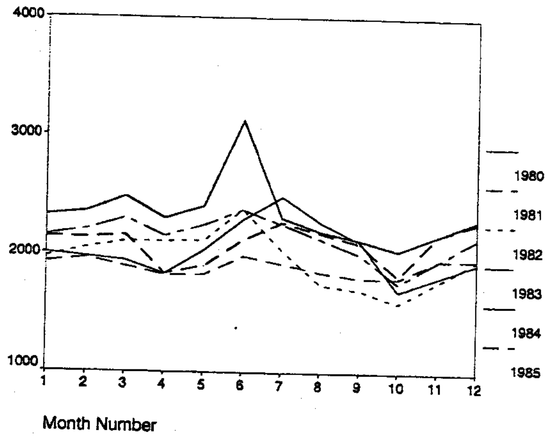


The data appear to be right-skewed.

(b) A log transformation is one way to induce normality. It works here, as can be seen from the normal plot of the transformed data below.



4.27 (a)



There is a dip in sales around October. Otherwise, it stays relatively flat. Using this format makes it easier to detect the cycles.

(d) The trend is not detectable from the histogram. 3126 is an outlier in either case.

(e) The separate run charts is most useful because it allows one to detect the cyclical trends.

Solutions for Section 4.4

4.29 (a) False

(b) False

(c) True

(d) False

4.30 (a) For each type of degree, the proportion of high income earners are the same for men and women (40% for liberal arts, and 60% for engineering).

(b)

Gender	Low Income	High Income
Female	110	90
Male	130	170

The proportions of women high earners is 45%, while for men it is about 57%. This would seem to indicate that men tend to earn more than women. However, this difference is driven by the fact that more men pursue engineering degrees which pay higher salaries. This is an example of Simpson's Paradox.

4.31 (a)

	Black		White	
Gender	Graduated	Did not Graduate	Graduated	Did not graduate
Female	54	89	498	298
Male	197	463	878	747

(b) For blacks, 38% of females graduated compared to 30% of males. For whites, 63% females graduated compared to 54% of males. In both ethnic groups, women had higher graduation rate by about 8-9%.

(c)

Gender	Graduated	Did not Graduate
Female	552	387
Male	1075	1210

The graduation rate for women is 59% compared to 47% for men. This disparity amo graduation rates, similar to that found in (b), indicates that graduation rate is independent of gender.

4.32 (a) For each income, the proportions of drug users who played soccer is the same as proportion of drug users who did not play soccer. This indicates that involvement soccer does not affect drug use.

(b)

Played	Drug Use	
Soccer	Yes	No
Yes	26	274
No	42	258

The proportion of drug users among soccer players is 9%. The proportion of drug us among teenagers who did not play soccer is 14%.

(c) This is an example of Simpson's Paradox. It would be misleading to conclude th involvement in soccer reduces teenage drug use because there is a lurking variable, inco level. Low income families are less likely to involve their children in soccer progra but more likely to have teenage drug users than higher income families.

4.33 (a)

Success Rates	Risk	
Hospital	Low	High
A	80%	20%
B	60%	10%

Hospital A is better.

(b)

Hospital	Success Rate
A	43%
B	46%

Hospital B has the higher success rate.

(c) This is an example of Simpson's Paradox. While Hospital A has a higher success ra for both risk groups, it has a larger percentage of high risk patients than hospital Since the high risk patients have a lower success rate, this discrepancy brings hospit A's overall success rate below hospital B.

4.34

(a) For smokers,

$$\text{Overall Death Rate} = \frac{\text{total number of smokers dead}}{\text{total number of smokers}} = \frac{139}{582} = 0.2388 = 24\%$$

For nonsmokers,

$$\text{Overall Death Rate} = \frac{\text{total number of nonsmokers dead}}{\text{total number of nonsmokers}} = \frac{230}{732} = 0.3142 = 31\%$$

Nonsmokers have the higher death rate.

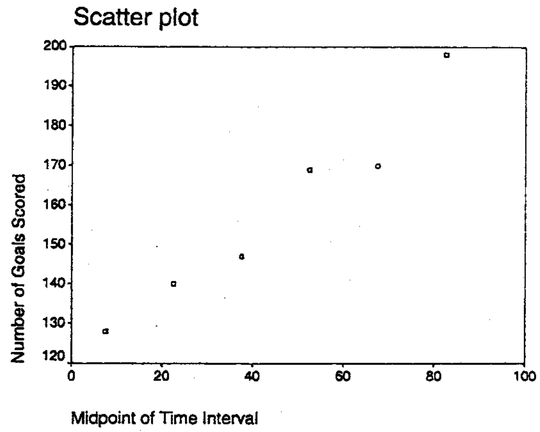
(b) The age-adjusted death rates is the weighted average of the age-specific death rates, weighted by the total number of people in each age group.

Age	Proportion Dead		Total
Group	Smokers	Nonsmokers	People
18–24	0.0364	0.0164	117
25 -34	0.0242	0.0318	281
35 -44	0.1284	0.0579	230
45 -54	0.2077	0.1538	208
55 -64	0.4435	0.3306	236
65 -74	0.8056	0.7829	165
75+	1.000	1.000	77
Wtd Avg	0.3032	0.2590	1314

The age-adjusted death rates (using the total number of people in each age group) are 30.3% for smokers and 25.9% for nonsmokers. Smokers have the higher death rate.

(c) The presence of fewer smokers in higher age groups indicates that smoking shortens lifespans. A higher frequency of smokers died earlier than did nonsmokers.

4.35 (a) The midpoints are 7.5, 22.5, 37.5, 52.5, 67.5, and 82.5.



The scatterplot indicates a positive linear trend in the number of goals scored.

(b) The correlation coefficient is -0.865. Yes, it does match the decreasing but not linear relationship indicated by the scatterplot.

4.38 The student would be expected to score $2 \times r = 1.5$ standard deviations below the mean. So the student's predicted score would be $75 - 1.5 \times 12 = 57$.

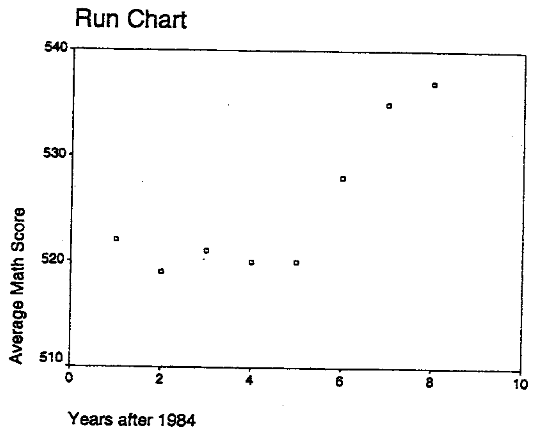
4.39 The person would be predicted to spend $1.5 \times r = -0.9$ standard deviations longer than the average time spent with the family, or 0.9 standard deviations shorter than the average time spent with the family. So the person's predicted family time is $60 - 0.9 \times 20 = 42$ minutes.

4.40 The slope and intercept are $b = r\frac{s_y}{s_x} = 0.8 \times \frac{3.8}{2.5} = 1.216$ and $a = \bar{y} - b\bar{x} = 15.3 - 1.216 \times 8.7 = 4.721$. Then the final equation for the least squares line is

$$y = 4.721 + 1.216x.$$

The estimated access time for 10 simultaneous users is $y = 4.721 + 1.216 \times 10 = 16.881$ milliseconds.

4.41 (a)



The plot indicates an increasing relationship between year and math scores, although it does not look linear.

(b) Relabelling years(x) as 1-8, $\bar{x} = 4.5$, $s_x = 2.449$, $\bar{y} = 525.25$, and $s_y = 7.206$. The correlation coefficient is 0.834. Then the slope and intercept are $b = r\frac{s_y}{s_x} = 0.834 \times \frac{7.206}{2.449} = 2.454$ and $a = \bar{y} - b\bar{x} = 525.25 - 2.454 \times 4.5 = 514.207$. Then the final equation for the least squares line is

$$y = 514.207 + 2.454x.$$

(c)

Chapter 5 Solutions

Solutions for Section 5.1

5.1 (a) The population is the uniform distribution over integers 1 to 5. The mean and variance of the population are

$$\mu = \frac{1}{5}(1+2+\ldots+5) = 3 \text{ and } \sigma^2 = E(X^2) - \mu^2 = \frac{1}{5}(1^2+2^2+\ldots+5^2) - (3)^2 = 2.$$

(b)

(x_1, x_2)	$\bar{x}$	(x_1, x_2)	$\bar{x}$
(1,1)	1.0	(4,1)	2.5
(1,2)	1.5	(4,2)	3.0
(1,3)	2.0	(4,3)	3.5
(1,4)	2.5	(4,4)	4.0
(1,5)	3.0	(4,5)	4.5
(2,1)	1.5	(5,1)	3.0
(2,2)	2.0	(5,2)	3.5
(2,3)	2.5	(5,3)	4.0
(2,4)	3.0	(5,4)	4.5
(2,5)	3.5	(5,5)	5.0
(3,1)	2.0		
(3,2)	2.5		
(3,3)	3.0		
(3,4)	3.5		
(3,5)	4.0		

(c)

$\bar{x}$	1.0	1.5	2.0	2.5	3.0	3.5	4.0	4.5	5.0
$f(\bar{x})$	$\frac{1}{25}$	$\frac{2}{25}$	$\frac{3}{25}$	$\frac{4}{25}$	$\frac{5}{25}$	$\frac{4}{25}$	$\frac{3}{25}$	$\frac{2}{25}$	$\frac{1}{25}$

(d) The mean of $\bar{X}$ is

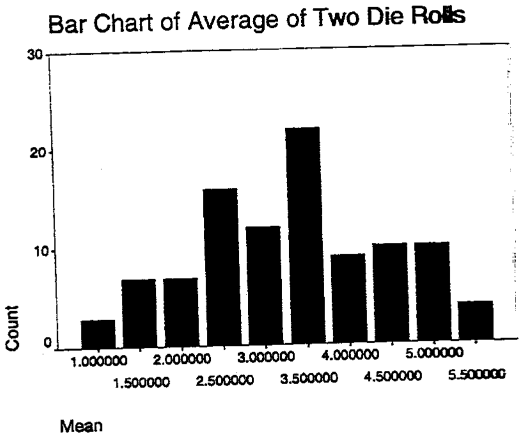
$$E(\bar{X}) = \sum_x x f(x) = 1.0 \times \frac{1}{25} + 1.5 \times \frac{2}{25} + \ldots + 5.0 \times \frac{1}{25} = 3.$$

The variance of $\bar{X}$ is

$$\text{Var}(\bar{X}) = E(\bar{X}^2) - \mu^2 = \left[1.0^2 \times \frac{1}{25} + 1.5^2 \times \frac{2}{25} + \ldots + 5.0^2 \times \frac{1}{25}\right] - 3^2 = 1.$$

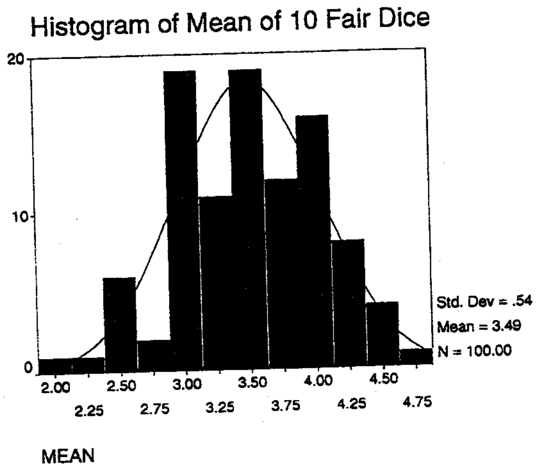
These equal μ and $\sigma^2/2$, respectively.

5.2 (a)



(b) From the simulation, the mean is 3.335 and the variance is 1.308. These are reasonably close to the true values, except for some sampling error.

5.3 (a)



(b) From the simulation, the mean is 3.49 and the variance is $(0.54)^2 = 0.2916$. These are reasonably close to the true values, except for some sampling error.

(c) The bar chart for the average of 10 dice looks more normal, and has a smaller variance, than the one for the average of two dice.

5.4 (a)

$$P(X \leq 355) \approx P\left(Z = \frac{X - 355.2}{0.5} \leq \frac{355 - 355.2}{0.5}\right) = \Phi(-0.4) = 0.3446.$$

(b) $\bar{X}$ is also normally distributed, with mean $\mu = 355.2$ and standard deviation $\sigma/\sqrt{n} = 0.5/\sqrt{6} = 0.204$. Then

$$P(\bar{X} \le 355) \approx P\left(Z = \frac{\bar{X} - 355.2}{0.204} \le \frac{355 - 355.2}{0.204}\right) = \Phi(-0.98) = 0.1635.$$

5.5 (a) U is approximately normal with mean $\mu = 40$ and SD $= \sigma/\sqrt{n} = 15/\sqrt{50} = 2.121$. V is approximately normal with mean $\mu = 40$ and SD $= \sigma/\sqrt{n} = 15/\sqrt{100} = 1.5$.

(b) Since V has a smaller standard deviation, more of the probability is clustered close to the mean of 40, so we would expect $P(35 \le V \le 40)$ to be larger.

(c)

$$\begin{aligned}
P(35 \le U \le 45) &\approx P\left(\frac{35-40}{2.121} \le Z = \frac{U-40}{2.121} \le \frac{45-40}{2.121}\right) \\
&= \Phi(2.357) - \Phi(-2.357) = 0.9909 - 0.0091 = 0.9818. \\
P(35 \le V \le 45) &\approx P\left(\frac{35-40}{1.5} \le Z = \frac{V-40}{1.5} \le \frac{45-40}{1.5}\right) \\
&= \Phi(3.333) - \Phi(-3.333) = 0.9996 - 0.0004 = 0.9992.
\end{aligned}$$

5.6 All 50 boxes can be sent in one shipment if the total weight of all 50 boxes is less than 4000. The total weight of all 50 boxes is normally distributed with mean $\mu = 78 \times 50 = 3900$ and SD $= \sigma\sqrt{n} = 12\sqrt{50} = 84.853$. Then the probability of sending all 50 boxes at once is

$$P\left(\sum X_i \le 4000\right) = P\left(Z = \frac{\sum X_i - 3900}{84.853} \le \frac{4000 - 3900}{84.853}\right) = \Phi(1.179) = 0.8810.$$

If the weights are not normally distributed, the answer will still be approximately correct. According to the Central Limit Theorem, the sample mean is approximately normal for large n regardless of the original distribution of the data.

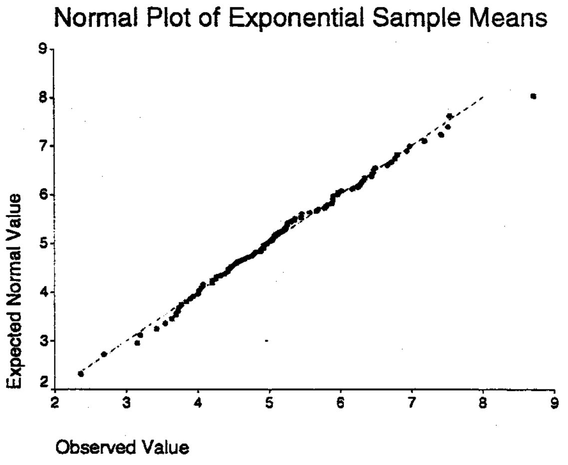
5.7 (a) Since X_i is exponential with $\lambda = 1/5$, $Y_i = X_i/25$ is exponential with $\lambda = 25 \times 1/5 = 5$. Then $\bar{X} = \sum X_i/25 = \sum Y_i$, which is Gamma (5, 25). From this Gamma distribution,

$$E(\bar{X}) = \frac{r}{\lambda} = \frac{25}{5} = 5$$

and

$$\mathrm{Var}(\bar{X}) = \frac{r}{\lambda^2} = \frac{25}{5^2} = 1.$$

(b)



This straight line pattern suggests that the sample means are approximately **normal**.

(c)

$$P(4.5 \leq \bar{X} \leq 5.5) = P\left(\frac{4.5-5}{1} \leq Z = \frac{\bar{X}-5}{1} \leq \frac{5.5-5}{1}\right)$$
$$= \Phi(0.5) - \Phi(-0.5) = 0.6915 - 0.3085 = 0.3830.$$

5.8 (a) $\bar{X}$ is normal with mean $\mu = 50000$ and SD $= \sigma/\sqrt{n} = 1000$.

(b) This does not require the use of the Central Limit Theorem. The sample mean of **n i.i.d** $N(\mu, \sigma^2)$ random variables is $N(\mu, \sigma^2/n)$ distributed.

(c)

$$P(\bar{X} \leq 47000) = P\left(Z = \frac{\bar{X}-50000}{1000} \leq \frac{47000-50000}{1000}\right) = \Phi(-3) = 0.0013.$$

5.9 (a)

$$P(X \leq 10) = 0.586$$

(b)

$$P(X \leq 10) \approx P\left(Z = \frac{X-np}{\sqrt{np(1-p)}} \leq \frac{10-10}{2.449}\right) = \Phi(0) = 0.5.$$

(c)

$$P(X \leq 10.5) \approx P\left(Z = \frac{X-np}{\sqrt{np(1-p)}} \leq \frac{10.5-10}{2.449}\right) = \Phi(0.204) = 0.5793.$$

5.10 (a)

$$P(X \leq 5) = 0.617$$

(b)

s^2	0.0	0.5	2.0	4.5	8.0
$f(s^2)$	$\frac{5}{25}$	$\frac{8}{25}$	$\frac{6}{25}$	$\frac{4}{25}$	$\frac{2}{25}$

(c)

$$E(S^2) = \sum_{s^2} s^2 f(s^2) = 0.0 \times \frac{5}{25} + 0.5 \times \frac{8}{25} + \ldots + 8.0 \times \frac{2}{25} = 2.$$

This equals $\sigma^2 = 2$ calculated in exercise 5.1.

5.16 $\chi^2_{5,0.01} = 15.085$, $\chi^2_{10,0.05} = 18.307$, $\chi^2_{10,0.95} = 3.940$, and $\chi^2_{10,0.75} = 6.737$.

5.17 (a) $E(\chi^2_8) = 8$ and $\text{Var}(\chi^2_8) = 2 \times 8 = 16$.

(b) $a = 15.507$, $b = 1.646$, $c = 13.362$, $d = 2.180$, $e = 17.534$.

(c) $a = \chi^2_{8,0.05}$, $b = \chi^2_{8,0.99}$, $c = \chi^2_{8,0.10}$, $d = \chi^2_{8,0.975}$, and $e = \chi^2_{8,0.025}$.

5.18 (a) $E(\chi^2_{14}) = 14$ and $V(\chi^2_{14}) = 2 \times 14 = 28$.

(b) $a = 23.685$, $b = 4.660$, $c = 21.064$, $d = 5.629$, and $e = 26.119$.

(c) $a = \chi^2_{14,0.05}$, $b = \chi^2_{14,0.99}$, $c = \chi^2_{14,0.10}$, $d = \chi^2_{14,0.975}$, and $e = \chi^2_{14,0.025}$.

5.19 Since

$$P(-z_{\alpha/2} \le Z \le z_{\alpha/2}) = P(Z^2 \le z^2_{\alpha/2}) = 1 - \alpha$$

and

$$Z^2 \sim \chi^2_1,$$

the $1-\alpha$ critical point of the χ^2_1 distribution is equal to $z^2_{\alpha/2}$. Hence, $\chi^2_{1,\alpha} = z^2_{\alpha/2}$.

5.20

$$\begin{aligned} P\left(S^2 > 2\sigma^2\right) &= P\left(\frac{(n-1)S^2}{\sigma^2} > 2(n-1)\right) \\ &= P\left(\chi^2_{n-1} > 2(n-1)\right) \end{aligned}$$

For $n = 8$, $P\left(\chi^2_7 > 14\right) \approx 0.05$. For $n = 17$, $P\left(\chi^2_{16} > 32\right) \approx 0.01$. For $n = 21$, $P\left(\chi^2_{20} > 40\right) \approx 0.005$. The probability that S^2 will exceed the true variance by more than a factor of two decreases as you increase the sample size. This is because our estimate of σ^2 improves with a larger and larger sample size.

5.21 (a) 100 random samples were generated.

(b) From the simulation, $\chi^2_{4,0.25} = 1.676$, $\chi^2_{4,0.5} = 3.187$, and $\chi^2_{4,0.90} = 6.734$. The exact values are $\chi^2_{4,0.25} = 1.923$, $\chi^2_{4,0.50} = 3.357$, and $\chi^2_{4,0.90} = 7.779$.

5.22 (a) 100 random samples were generated.

(b) From the simulation, $\chi^2_{4,0.25} = 1.993$, $\chi^2_{4,0.5} = 3.118$, and $\chi^2_{4,0.90} = 7.393$. The exact values are $\chi^2_{4,0.25} = 1.923$, $\chi^2_{4,0.50} = 3.357$, and $\chi^2_{4,0.90} = 7.779$.

5.23

(a) c satisfies

$$P\left(S > 5c|\sigma = 5\right) = 0.1$$

or

$$P\left(\chi^2_{19} = \frac{(n-1)S^2}{\sigma^2} > \frac{19 \times (5c)^2}{5^2}\right) = 0.1$$

or

$$P\left(\chi^2_{19} = 19c^2\right) = 0.1$$

or

$$19c^2 = \chi^2_{19,0.1} = 27.203,$$

so

$$c = \sqrt{\frac{27.203}{19}} = 1.197.$$

(b) Since $s = 7.5 > 5c = 5.983$, the engineer would conclude that $\sigma > 5$.