8.3 Estimating the Mean

MATHEMATICAL TECHNIQUES

♠ Consider the following data on 20 plants.

Plant	Weight	Height	Yield	Number of seeds
1	7.3	2.1	0.045	7
2	10.8	3.3	0.132	18
3	12.1	3.5	0.187	19
4	18.6	4.2	0.129	16
5	20.1	5.0	0.201	17
6	32.3	6.1	0.184	14
7	38.4	5.5	0.231	22
8	40.2	7.2	0.298	21
9	42.1	7.4	0.287	20
10	43.8	6.9	0.310	25
Plant	Weight	Height	Yield	Number of seeds
11	43.8	6.7	0.276	24
12	44.9	7.6	0.353	15
13	45.8	5.6	0.462	12
14	46.4	8.0	0.561	23
15	52.7	7.8	0.435	23
16	60.3	9.8	0.598	12
17	98.9	11.2	0.723	25
18	178.3	9.2	0.668	17
19	213.1	13.5	1.022	25
20	298.8	17.3	1.745	21

Find the following for the given measurement.

a. The sample mean.

b. The sample median.

c. The trimmed means $\bar{X}_{tr(5)}$, $\bar{X}_{tr(10)}$, and $\bar{X}_{tr(20)}$.

• EXERCISE **8.3.1**
The weight W.

• EXERCISE **8.3.2**
The height H.

• EXERCISE **8.3.3**
The yield Y.

• EXERCISE **8.3.4**
The seed number S.

♠ Find the sample variance and sample standard deviation for the given measurement.

• EXERCISE **8.3.5**
The weight W in exercise 8.3.1.

• EXERCISE **8.3.6**
The height H in exercise 8.3.2.

• EXERCISE **8.3.7**
The yield Y in exercise 8.3.3.

- EXERCISE **8.3.8**
The seed number S in exercise 8.3.4.

♠ Find how many measurements lie a) less than one sample standard deviation from the sample mean and b)more than two sample standard deviations from the sample mean for the given measurement. Which behave more or less like the normal distribution?

- EXERCISE **8.3.9**
The weight W in exercise 8.3.1.
- EXERCISE **8.3.10**
The height H in exercise 8.3.2.
- EXERCISE **8.3.11**
The yield Y in exercise 8.3.3.
- EXERCISE **8.3.12**
The seed number S in exercise 8.3.4.

♠ Find the 95% confidence limits around the sample mean assuming that the sample variance s is a good estimate of the true variance σ.

- EXERCISE **8.3.13**
The weight W in exercise 8.3.1.
- EXERCISE **8.3.14**
The height H in exercise 8.3.2.
- EXERCISE **8.3.15**
The yield Y in exercise 8.3.3.
- EXERCISE **8.3.16**
The seed number S in exercise 8.3.4.

♠ Find the given confidence limits around the sample mean assuming that the sample variance s is a good estimate of the true variance σ.

- EXERCISE **8.3.17**
98% confidence limits around the weight W in exercise 8.3.1.
- EXERCISE **8.3.18**
90% confidence limits around the height H in exercise 8.3.2.
- EXERCISE **8.3.19**
99.8% confidence limits around the yield Y in exercise 8.3.3.
- EXERCISE **8.3.20**
99.9% confidence limits around the seed number S in exercise 8.3.4.

♠ Find the normal approximation to the following.

- EXERCISE **8.3.21**
The average of 30 numbers chosen from the exponential p.d.f. $g(x) = 2e^{-2x}$.
- EXERCISE **8.3.22**
The average of 30 numbers chosen from the p.d.f. $f(x) = 1$ for $0 \leq x \leq 1$.

♠ Find 95% confidence intervals in the following cases, assuming that the standard deviations are known to match those in the earlier problem. Does it include the true mean?

- EXERCISE **8.3.23**
A sample mean of 0.4 is found in exercise 8.3.21.
- EXERCISE **8.3.24**
A sample mean of 0.7 is found in exercise 8.3.22.

♠ Use the normal approximation to find 95% confidence limits around the estimated proportion $\hat{p}$ in the following cases.

- EXERCISE **8.3.25**
A coin is flipped 100 times and comes out heads 44 times.
- EXERCISE **8.3.26**
1000 people are polled and 320 favor the use of mathematics in biology.

♠ In simple cases, we can see why using a denominator of n in the equation for the sample variance produces a biased estimate. Suppose a population consists of half 0's and half 1's.

- EXERCISE **8.3.27**

  Use the variance of a Bernoulli distribution with $p = 0.5$ to find the exact variance of each measurement.
- EXERCISE **8.3.28**

  Find all possible samples of size 2 and their associated probabilities.
- EXERCISE **8.3.29**

  Find the mean squared deviation from the mean for each and average them to find the expected sample variance.
- EXERCISE **8.3.30**

  Compare with the true answer and show that using a denominator of $n-1$ rather than n would give the right answer. Try to explain the bias.

♠ One measurement of 10.0 is made from a normal distribution with known variance $\sigma^2 = 1$.

- EXERCISE **8.3.31**

  Write the likelihood function for the unknown parameter μ.
- EXERCISE **8.3.32**

  Write the support function.
- EXERCISE **8.3.33**

  Find the maximum likelihood estimator of μ.
- EXERCISE **8.3.34**

  Use the method of support to estimate the 95% confidence interval. and compare with the usual method in the text.

APPLICATIONS

♠ Consider the following data on immigration into 4 populations over 20 years. (based on the probabilities in exercises 7.8.33-7.8.36.) For each, find the sample mean and the sample standard deviation. Compare them with the mathematical mean and standard deviation found in the earlier problem.

Year	Population a	Population b	Population c	Population d
1	0	-1	1	-10
2	-1	0	1	1
3	1	0	1	0
4	1	0	1	1
5	-1	0	-1	-10
6	-1	0	-1	0
7	1	0	0	1
8	2	-1	-1	-10
9	0	2	-1	1
10	0	-1	-1	-10
Year	Population a	Population b	Population c	Population d
11	-1	1	0	1
12	-1	0	100	1
13	-1	2	1	-10
14	-1	2	1	1
15	-1	0	0	1
16	0	0	-1	-10
17	-1	0	-1	-10
18	1	0	1	1
19	-1	0	1	-10
20	0	2	1	-10

- EXERCISE **8.3.35**

  Population a (see exercise 7.8.33).
- EXERCISE **8.3.36**

  Population b (see exercise 7.8.34).

- EXERCISE **8.3.37**

  Population c (see exercise 7.8.35). Why are the mean and variance about half of the mathematical expectations?
- EXERCISE **8.3.38**

  Population d (see exercise 7.8.36).

♠ Find the 95% confidence intervals around the mean number of immigrants using both the true variance and the sample variance. Does the true mean lie within the confidence limits?

- EXERCISE **8.3.39**

  Population a.
- EXERCISE **8.3.40**

  Population b.
- EXERCISE **8.3.41**

  Population c.
- EXERCISE **8.3.42**

  Population d.

♠ Several plants are crossed, producing the following proportions. Find 99% confidence limits around the fraction of tall plants in each case.

- EXERCISE **8.3.43**

  50 plants are crossed, and 35 are tall.
- EXERCISE **8.3.44**

  500 plants are crossed, and 350 are tall.
- EXERCISE **8.3.45**

  100 plants are crossed, and 52 are tall.
- EXERCISE **8.3.46**

  200 plants are crossed, and 13 are tall.

♠ Mutations are counted in a large section of the genome. 14 mutations are found in one set of one million base pairs. and then 30 different sets of one million base pairs are measured, and the average number of mutations per million is found to be 13.5.

- EXERCISE **8.3.47**

  Use the normal approximation to the Poisson distribution to estimate 95% confidence limits around the true mean number of mutations per million base pairs and compare with exercise 8.2.41.
- EXERCISE **8.3.48**

  30 different sets of one million base pairs are measured, and the average number of mutations per million is found to be 13.5. Estimate the standard deviation, and find the standard error and the 99% confidence limits.

♠ Consider measuring n plants with a known standard deviation of 3.2 cm (as in the text). How many plants would have to be measured to achieve the following?

- EXERCISE **8.3.49**

  95% confidence limits 2.0 cm wide.
- EXERCISE **8.3.50**

  99% confidence limits 2.0 cm wide.
- EXERCISE **8.3.51**

  95% confidence limits 0.5 cm wide.
- EXERCISE **8.3.52**

  99% confidence limits 0.5 cm wide.

Chapter 9

Answers

8.3.1. $\bar{W} = 67.43$, $\tilde{W} = 43.8$, $\bar{W}_{tr(5)} = 57.92$, $\bar{W}_{tr(10)} = 51.17$, $\bar{W}_{tr(20)} = 42.57$.
8.3.3. $\bar{Y} = 0.442$, $\tilde{Y} = 0.304$, $\bar{Y}_{tr(5)} = 0.392$, $\bar{Y}_{tr(10)} = 0.369$, $\bar{Y}_{tr(20)} = 0.360$.
8.3.5. $s^2 = 5726.6$ and $s = 75.7$.
8.3.7. $s^2 = 0.152$ and $s = 0.390$.
8.3.9. 17 of the values lie in the range from $\bar{W} - s = -8.24$ to $\bar{W} + s = 143.1$. This is more than the 68% (about 13 or 14) we would expect with a normal distribution. Only the last value, 298.8, lies outside the range from $\bar{W} - 2s = -83.91$ to $\bar{W} + 2s = 218.78$. This is what we would expect with the normal distribution, even though these values are far from normal.
8.3.11. 17 of the values lie in the range from $\bar{Y} - s = 0.053$ to $\bar{Y} + s = 0.832$. This is more than we expect with a normal distribution. Only the last value, 1.745, lies outside the range from $\bar{Y} - 2s = -0.337$ to $\bar{Y} + 2s = 1.222$. This is what we would expect with the normal distribution.
8.3.13. $\bar{W} - 1.96s/\sqrt{20} = 34.27$ and $\bar{W} + 1.96s/\sqrt{20} = 100.6$.
8.3.15. $\bar{Y} - 1.96s/\sqrt{20} = 0.27$ and $\bar{Y} + 1.96s/\sqrt{20} = 0.61$.
8.3.17. To find 98% confidence limits, we need to know where the standard normal c.d.f. crosses 0.01 and 0.99. From a table or computer, we find these to be at -2.326 and 2.326. Then the limits are from $\bar{W} - 2.326s/\sqrt{20} = 28.08$ to $\bar{W} + 2.326s/\sqrt{20} = 106.8$.
8.3.19. To find 99.8% confidence limits, we need to know where the standard normal c.d.f. crosses 0.001 and 0.999. From a table or computer, we find these to be at -3.090 and 3.090. Then the limits are from $\bar{Y} - 3.09s/\sqrt{20} = 0.173$ to $\bar{Y} + 3.09s/\sqrt{20} = 0.712$.
8.3.21. The mean and variance of an exponentially distributed number with $\lambda = 2.0$ are 0.5 and 0.25. The average of 30 numbers will be approximately normal with mean 0.5 and variance $\frac{0.25}{30} = 0.0083$, or $N(0.5, 0.0083)$.
8.3.23. Lower limit is $0.4 - 1.96 \cdot \sqrt{0.0083} = 0.221$. Upper limit is $0.4 + 1.96 \cdot \sqrt{0.0083} = 0.579$. These include the true mean of 0.5.
8.3.25. $\hat{p} = 0.44$ and $s^2 = 0.44 \cdot 0.56 = 0.246$. Then

$$\begin{aligned} p_l &= 0.44 - 1.96\frac{\sqrt{0.246}}{\sqrt{100}} = 0.44 - 0.097 = 0.537 \\ p_h &= 0.44 + 1.96\frac{\sqrt{0.246}}{\sqrt{100}} = 0.44 + 0.097 = 0.343. \end{aligned}$$

8.3.27. The variance of each measurement is 0.25.
8.3.29. The variance of the first and last is 0. The mean squared deviation from the mean (0.5) of the others is

$$\frac{(0 - 0.5)^2 + (1 - 0.5)^2}{2} = 0.25.$$

The average is $0 \cdot 0.25 + 0.25 \cdot 0.25 + 0.25 \cdot 0.25 + 0 \cdot 0.25 = 0.125$.

8.3.31.

$$L(\mu) = \frac{1}{\sqrt{2\pi}} e^{-\frac{(10.0-\mu)^2}{2}}.$$

8.3.33. $S'(\mu) = (10.0 - \mu)$, which is 0 at $\mu = 10.0$. This is the maximum of this quadratic function.
8.3.35. The sample mean is -0.2, the sample variance is 0.905, and the sample standard deviation is 0.951. These are reasonably close to the true mean of 0.1, variance of 1.09, and standard deviation of 1.044.
8.3.37. The sample mean is 5.1, the sample variance is 499.8, and the sample standard deviation is 22.3. These are about half of the true mean of 9.9 and variance of 902.7 because we only got one year with 100 and would have expected 2.
8.3.39. The sample mean is -0.2 and the true standard deviation is 1.044. These give 95% confidence limits of

$$\mu_l = -0.2 - 1.96 \cdot 1.044/\sqrt{20} = -0.657, \mu_h = -0.2 + 1.96 \cdot 1.044/\sqrt{20} = 0.257.$$

The sample standard deviation is 0.951. These give 95% confidence limits of

$$\mu_l = -0.2 - 1.96 \cdot 0.951/\sqrt{20} = -0.617, \mu_h = -0.2 + 1.96 \cdot 0.951/\sqrt{20} = 0.217.$$

The true mean lies within the confidence limits in both cases.
8.3.41. The sample mean is 5.1 and the true standard deviation is 30.0. These give 95% confidence limits of

$$\mu_l = 5.1 - 1.96 \cdot 30.0/\sqrt{20} = -8.048, \mu_h = 5.1 + 1.96 \cdot 30.0/\sqrt{20} = 18.25.$$

The sample standard deviation is 22.3. These give 95% confidence limits of

$$\mu_l = 5.1 - 1.96 \cdot 22.3/\sqrt{20} = -4.67, \mu_h = 5.1 + 1.96 \cdot 22.3/\sqrt{20} = 14.87.$$

The true mean lies within the confidence limits in both cases.
8.3.43. $\hat{p} = 0.7$ and $s^2 = 0.7 \cdot 0.3 = 0.21$. Then

$$\begin{aligned} p_l &= 0.7 - 2.575\frac{\sqrt{0.21}}{\sqrt{50}} = 0.623 \\ p_h &= 0.7 + 2.575\frac{\sqrt{0.21}}{\sqrt{50}} = 0.776. \end{aligned}$$

8.3.45. $\hat{p} = 0.52$ and $s^2 = 0.52 \cdot 0.48 = 0.2496$. Then

$$\begin{aligned} p_l &= 0.52 - 2.576\frac{\sqrt{0.2496}}{\sqrt{100}} = 0.456 \\ p_h &= 0.52 + 2.576\frac{\sqrt{0.2496}}{\sqrt{100}} = 0.584. \end{aligned}$$

8.3.47.

$$\Lambda_l \approx 14 - 1.96 \cdot \sqrt{14} = 6.7$$
$$\Lambda_h \approx 14 + 1.96 \cdot \sqrt{14} = 21.3.$$

These are reasonably close to the values 7.65 and 23.5 found earlier.
8.3.49. The 95% confidence limits are 1.96 standard errors above and below the sample mean, for a total width of 3.92 standard errors. Therefore, we require that standard error $= \frac{2.0}{3.92} = 0.51$. But

$$\text{standard error} = \frac{3.2}{\sqrt{n}}.$$

so $n = (3.2/0.51)^2 = 39$ and 39 plants are required.
8.3.51. Improving by a factor of 4 takes 16 times more plants, or about 630.