7.8 The Normal Distribution

MATHEMATICAL TECHNIQUES

♠ Suppose the random variables $X_1, X_2 \ldots X_n$ are i.i.d. (independent and identically distributed), and consider the sum $S_n = \sum_{i=1}^{n} X_i$. Find the mean and variance of S_n, and write and sketch the p.d.f. of the approximate normal distribution.

- EXERCISE **7.8.1**

  Suppose that $n = 16$, and that X_i takes the value 0 with probability 0.5 and the value 1 with probability 0.5.

- EXERCISE **7.8.2**

  Suppose that $n = 50$, and that X_i takes the value 0 with probability 0.25, the value 1 with probability 0.5, and the value 2 with probability 0.25.

♠ Suppose the random variables $X_1, X_2 \ldots X_n$ are i.i.d. and consider the average $A_n = \frac{1}{n}\sum_{i=1}^{n} X_i$. Find the mean and variance of A_n, and write and sketch the p.d.f. of the approximate normal distribution.

- EXERCISE **7.8.3**

  Suppose that $n = 16$, and that X_i takes the value 0 with probability 0.5 and the value 1 with probability 0.5 (as in exercise 7.8.1).

- EXERCISE **7.8.4**

  Suppose that $n = 50$, and that X_i takes the value 0 with probability 0.25, the value 1 with probability 0.5, and the value 2 with probability 0.25(as in exercise 7.8.2).

♠ Recall the data describing the probabilities of the outcomes of 4 experiments counting the number of mutants in a bacterial culture (exercises 6.7.1–6.7.4).

Number of mutants	Probability			
	Experiment a	Experiment b	Experiment c	Experiment d
0	0.1	0.6	0.3	0.1
1	0.2	0.3	0.2	0.3
2	0.3	0.1	0.2	0.1
3	0.3	0.0	0.2	0.4
4	0.1	0.0	0.1	0.1

Suppose that each experiment is repeated (independently) and the total number of mutants is counted.

- EXERCISE **7.8.5**

  Experiment a is repeated 20 times. Write an integral that estimates the probability that there are between 30 and 40 mutants, and shade the corresponding area on a sketch of the approximate normal p.d.f. Take a guess at the probability based on your sketch. (We found the expectation in exercise 6.8.1 and the variance in exercise 6.10.1.)

- EXERCISE **7.8.6**

  Experiment b is repeated 80 times. Write an integral that estimates the probability that there are between 30 and 40 mutants, and shade the corresponding area on a sketch of the approximate normal p.d.f. Take a guess at the probability based on your sketch. (We found the expectation in exercise 6.8.2 and the variance in exercise 6.10.2.)

- EXERCISE **7.8.7**

  Experiment c is repeated 100 times. Write an integral that estimates the probability that there is an average of less than 1.8 mutants per experiment and shade the corresponding area on a sketch of the approximate normal p.d.f. Take a guess at the probability based on your sketch. (We found the expectation in exercise 6.8.3 and the variance in exercise 6.10.3.)

- EXERCISE **7.8.8**

  Experiment d is repeated 35 times. Write an integral that estimates the probability that there is an average of less than 1.8 mutants per experiment and shade the corresponding area on a sketch of the approximate normal p.d.f. Take a guess at the probability based on your sketch. (We found the expectation in exercise 6.8.4 and the variance in exercise 6.10.4.)

♠ Suppose that X and Y are independent normally distributed random variables with $X \sim N(5.0, 16.0)$ and $Y \sim N(10.0, 9.0)$.

- EXERCISE **7.8.9**
  Find and sketch the p.d.f. of $3X$.
- EXERCISE **7.8.10**
  Find and sketch the p.d.f. of $X + Y$.
- EXERCISE **7.8.11**
  Find and sketch the p.d.f. of the sum of nine independent samples from X.
- EXERCISE **7.8.12**
  Find and sketch the p.d.f. of the mean of nine independent samples from Y.

♠ Show the following facts about the normal distribution.

- EXERCISE **7.8.13**
  The normal p.d.f. with $\mu = 0$ and $\sigma^2 = 1$ takes on its maximum at $x = 0$.
- EXERCISE **7.8.14**
  The normal p.d.f. takes on its maximum at $x = \mu$ for any values of μ and σ.
- EXERCISE **7.8.15**
  The normal p.d.f. with $\mu = 0$ and $\sigma^2 = 1$ has points of inflection at $x = -1$ and $x = 1$.
- EXERCISE **7.8.16**
  The normal p.d.f. has points of inflection at $x = \mu + \sigma$ and $x = \mu - \sigma$ for any values of μ and σ.

♠ Suppose T_1, T_2, $T_3 \ldots$ are i.i.d. exponential random variables with $\lambda = 1.0$, and that $S_n = \sum_{i=1}^{n} T_i$ (Figure 7.40).

- EXERCISE **7.8.17**
  Use the law of total probability to show that

$$\Pr(S_2 = t) = \int_{s=0}^{t} \Pr(T_1 = s) \Pr(T_2 = t - s) ds.$$

- EXERCISE **7.8.18**
  Evaluate the integral to find the p.d.f. for S_2.
- EXERCISE **7.8.19**
  Use the same trick to find the p.d.f. for S_3.
- EXERCISE **7.8.20**
  Can you guess the pattern? Why does the answer look so much like the Poisson distribution?

♠ The central limit theorem does not work if random variables are not independent. The simplest case to compute is where they are perfectly correlated with each other. In particular, suppose the random variables $X_1, X_2 \ldots X_n$ are all equal, and consider the sum $S_n = \sum_{i=1}^{n} X_i$. Find the mean and variance of S_n, and sketch its probability distribution.

- EXERCISE **7.8.21**
  Suppose that $n = 16$, and that X_i takes the value 0 with probability 0.5 and the value 1 with probability 0.5. Compare with the results in exercise 7.8.1 (you can overlay the graphs).
- EXERCISE **7.8.22**
  Suppose that $n = 50$, and that X_i takes the value 0 with probability 0.25, the value 1 with probability 0.5, and the value 2 with probability 0.25. Compare with the results in exercise 7.8.2.

APPLICATIONS

♠ Based on the probabilities in Figure 7.34b, we can find some of the probability that a plant gains 5 cm in height from 10 genes ($\Pr(H = 5)$) directly.

- EXERCISE **7.8.23**
  Find the two ways to add up 0's, 1's, and 2.5's to get 5.
- EXERCISE **7.8.24**
  Find the number of ways each could occur (use binomial coefficients).
- EXERCISE **7.8.25**
  Find the probability associated with each way.

- EXERCISE **7.8.26**
  Add them up to find the total probability.
- EXERCISE **7.8.27**
  Find the normal distribution approximating added height, and find the value of the normal p.d.f. at 5.0.
- EXERCISE **7.8.28**
  Use the result of the previous problem to approximate the probability that the height is in the interval between 4.5 and 5.5. Compare with the result in exercise 7.8.26.

♠ Scientists develop a sophisticated new model of human IQ that includes three independent factors: genes of large effect, genes of small effect, and environmental effects. All genes are assumed to be dominant. There are 10 smart genes of large effect, each of which adds 2.5 IQ points. There are 20 smart genes of small effect, each of which adds 0.6 IQ points. There are 30 environmental factors, with favorable ones adding 0.9 IQ points. People have a baseline IQ of 80 if they have no favorable effects. Finally, suppose that the probability of getting each smart gene is 0.75, and the probability of getting each favorable environmental effect is 0.5.

- EXERCISE **7.8.29**
  Find the normal approximation for IQ based only on genes of large effect.
- EXERCISE **7.8.30**
  Find the normal approximation for IQ based only on genes of small effect.
- EXERCISE **7.8.31**
  Find the normal approximation for IQ based only on environmental effects.
- EXERCISE **7.8.32**
  Find the normal approximation for IQ with both genetic and environmental effects. What is the maximum possible IQ with the model?

♠ Suppose immigration and emigration change the sizes of four populations with the following probabilities (from 6.8.29–6.8.32).

Population a		Population b		Population c		Population d	
Number	Probability	Number	Probability	Number	Probability	Number	Probability
-1	0.4	-1	0.1	-1	0.4	-10	0.4
0	0.2	0	0.3	0	0.2	0	0.2
1	0.3	1	0.2	1	0.3	1	0.3
2	0.1	2	0.4	100	0.1	2	0.1

- EXERCISE **7.8.33**
  Suppose immigrants arrive into population a for 20 years. Find the p.d.f. of the normal approximation for the average number of immigrants, sketch a graph, and shade and estimate the area corresponding to an increase in the population. What is the maximum possible average change in the population?
- EXERCISE **7.8.34**
  Suppose immigrants arrive into population b for 20 years. Find the p.d.f. of the normal approximation for the average number of immigrants, sketch a graph, and shade and estimate the area corresponding to an increase in the population. What is the maximum possible average change in the population?
- EXERCISE **7.8.35**
  Suppose immigrants arrive into population c for 10 years. Find the p.d.f. of the normal approximation for the average number of immigrants, sketch a graph, and shade and estimate the area corresponding to an increase in the population. What is maximum possible average decrease in the population? How accurate do you think the normal approximation is?
- EXERCISE **7.8.36**
  Suppose immigrants arrive into population d for 10 years. Find the p.d.f. of the normal approximation for the average number of immigrants, sketch a graph, and shade and estimate the area corresponding to an increase in the population. What is maximum possible average increase in the population? How accurate do you think the normal approximation is?

♠ Although the central limit theorem applies to the **sums** of independent and identically distributed random variables, we can use logarithms to analyze **products** of independent and identically distributed random variables. Consider populations growing for the given number of years with the given distribution for the random variable R giving per capita reproduction. Find the normal distribution that approximates the logarithm of the population size P_t assuming that $P_0 = 100$.

• EXERCISE **7.8.37**

$R_i = 4$ with probability 0.5, $R_i = 0.25$ with probability 0.5 (as in exercise 6.9.35). Find the approximate normal distribution for $\ln(P_{50})$. Use the rule of thumb that most populations end up within two standard deviations from the mean to give a range of probable population sizes.

• EXERCISE **7.8.38**

$R_i = 4$ with probability 0.25, $R_i = 0.25$ with probability 0.75 (as in exercise 6.9.36). Find the approximate normal distribution for $\ln(P_{25})$. Use the rule of thumb that most populations end up within two standard deviations from the mean to give a range of probable population sizes.

♠ As in exercises 7.8.37 and 7.8.38, the Central Limit Theorem for sums can be used to approximate the logarithm of a product even when the random variables multiplied together are continuous random variables. Find the normal distribution that approximates the logarithm of the population size P_{50} assuming that $P_0 = 1$. You will need the indefinite integrals

$$\begin{aligned}\int \ln(x)dx &= x\ln(x) - x\\ \int \ln(x)^2 dx &= x\ln(x)^2 - 2x\ln(x) + 2x\end{aligned}$$

to evaluate the expectation and variance of $\ln(R)$.

• EXERCISE **7.8.39**

Let R be a random variable giving the per capita reproduction in a population with p.d.f. $g(x) = 5.0$ for $1.0 \le x \le 1.2$ (the values used to generate Figure 6.2b. Use the rule of thumb that most populations end up within two standard deviations from the mean to give a range of probable population sizes. Are the simulations in Figure 6.2b within this range?

• EXERCISE **7.8.40**

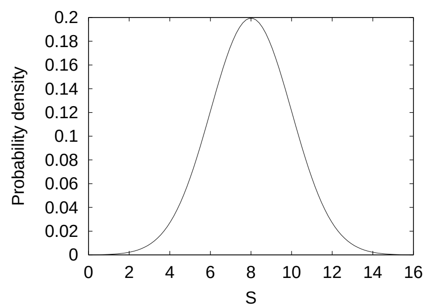
Let R be a random variable giving the per capita reproduction in a population with p.d.f. $g(x) = 1.25$ for $0.7 \le x \le 1.5$ (the values used to generate Figure 6.2c. Use the rule of thumb that most populations end up within two standard deviations from the mean to give a range of probable population sizes. Are the simulations in Figure 6.2c within this range?

Chapter 8

Answers

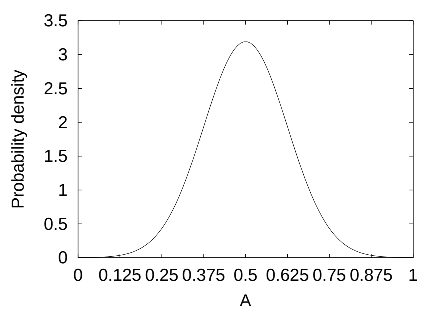
7.8.1. For these Bernoulli random variables, $E(X_i) = 0.5$ and $\text{Var}(X_i) = 0.25$. Then $E(S_n) = 16 \cdot 0.5 = 8$, and $\sigma^2 = \text{Var}(S_n) = 16 \cdot 0.25 = 4$. Therefore, $\sigma = 2$, and the p.d.f. is

$$f(x) = \frac{1}{2\sqrt{2\pi}} e^{-\frac{(x-8)^2}{8}}.$$



7.8.3. We have $E(X_i) = 0.5$ and $\text{Var}(X_i) = 0.25$. Then $E(A_n) = E(X_i) = 0.5$ and $\sigma^2 = \text{Var}(A_n) = \frac{0.25}{16} = 0.0156$. Therefore, $\sigma = 0.125$, and the p.d.f. is

$$f(x) = \frac{1}{0.125\sqrt{2\pi}} e^{-\frac{(x-0.5)^2}{0.5}}.$$



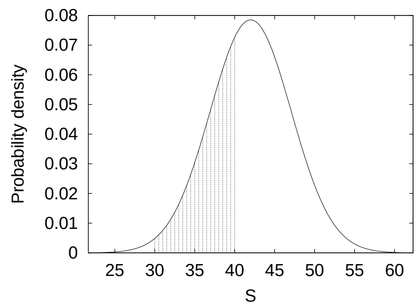
7.8.5. Let M_i be a random variable giving the number of mutants in the ith experiment. We found that $E(M_i) = 2.1$ and $\text{Var}(M_i) = 1.29$. Then $S_{20} = \sum_{i=0}^{i=20} M_i$ has mean $E(S_n) = 20 \cdot 2.1 = 42$ and variance $\sigma^2 = 20 \cdot 1.29 = 25.8$. Then $\sigma = 5.08$, and the p.d.f. approximating S_n is

$$f(x) = \frac{1}{5.08\sqrt{2\pi}} e^{-\frac{(x-42)^2}{51.6}}.$$

The probability that the number is between 30 and 40 is approximately

$$\Pr(30 \le S_n \le 40) = \int_{x=30}^{x=40} f(x)dx.$$

The area looks like about 0.4 of the total to me.



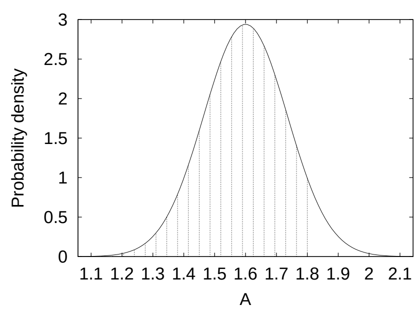
7.8.7. Let M_i be a random variable giving the number of mutants in the ith experiment. We found that $\mathrm{E}(M_i) = 1.6$ and $\mathrm{Var}(M_i) = 1.84$. Then $A_{100} = \frac{1}{100}\sum_{i=0}^{i=100} M_i$ has mean $\mathrm{E}(A_n) = 1.6$ and variance $\sigma^2 = \frac{0.45}{100} = 0.0184$. Then $\sigma = 0.136$, and the p.d.f. approximating S_n is

$$f(x) = \frac{1}{0.136\sqrt{2\pi}} e^{-\frac{(x-1.6)^2}{0.0368}}.$$

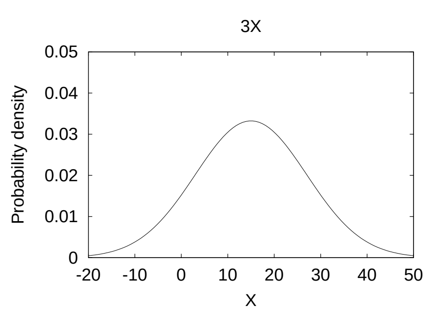
The probability that the average is less than 1.8 is

$$\Pr(A_n \le 1.8) = \int_{x=-\infty}^{x=1.8} f(x)dx.$$

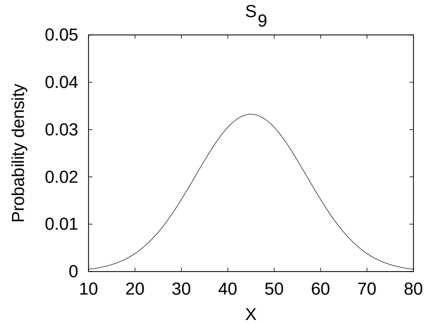
The area looks like about 0.9 of the total to me.



7.8.9. $3X \sim N(15.0, 144.0)$. The variance is multiplied by 9.



7.8.11. Let S_9 be the sum of nine independent samples from X. Then $S_9 \sim N(45.0, 144.0)$.



7.8.13. The p.d.f. is

$$f(x) = \frac{1}{\sqrt{2\pi}} e^{-\frac{x^2}{2}}.$$

Taking the derivative

$$\frac{df}{dx} = -\frac{x}{\sqrt{2\pi}} e^{-\frac{x^2}{2}}$$

which is 0 only at $x = 0$. Furthermore, $f(x)$ approaches 0 at $x = -\infty$ and $x = \infty$ because the power in the exponent approaches negative infinity. Therefore, $x = 0$ must be a maximum.

7.8.15. The second derivative is

$$\begin{aligned}\frac{d^2 f}{dx^2} &= \frac{-1}{\sqrt{2\pi}} e^{-\frac{x^2}{2}} + \frac{x^2}{\sqrt{2\pi}} e^{-\frac{x^2}{2}} \\ &= (x^2 - 1)\frac{1}{\sqrt{2\pi}} e^{-\frac{x^2}{2}}\end{aligned}$$

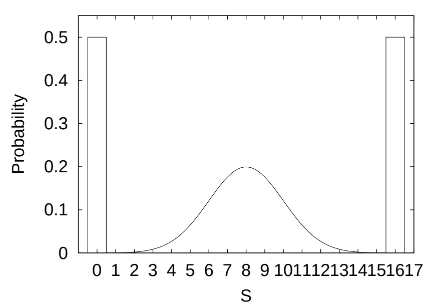
which is 0 where $x^2 = 1$ or $x = -1$ and $x = 1$.

7.8.17. If the second event occurs at time t, the first had to occur at some time s before that, and the second wait had to make up the difference of $t - s$ exactly.

7.8.19.

$$\begin{aligned}\Pr(S_3 = t) &= \int_{s=0}^{t} \Pr(S_2 = s)\Pr(T_3 = t - s) ds = \int_{s=0}^{t} s e^{-s} e^{-t+s} ds \\ &= \int_{s=0}^{t} s e^{-t} ds = \frac{s^2}{2} e^{-t}\Big|_0^t = \frac{t^2}{2} e^{-t}.\end{aligned}$$

7.8.21. For these Bernoulli random variables, $\mathrm{E}(X_i) = 0.5$ and $\mathrm{Var}(X_i) = 0.25$. But if all the random variables take on the same value (so $X_2 = X_1$, $X_3 = X_1$, and so forth), $S_n = 16X_1$. Then $\mathrm{E}(S_n) = 16 \cdot \mathrm{E}(X_1) = 16 \cdot 0.5 = 8$, and $\mathrm{Var}(S_n) = \mathrm{Var}(16X_1) = 256 \cdot 0.25 = 64$. More explicitly, $S_n = 0$ with probability 0.5 and $S_n = 16$ with probability 0.5. The result has a much higher variance, and bears no resemblance to a normal distribution.



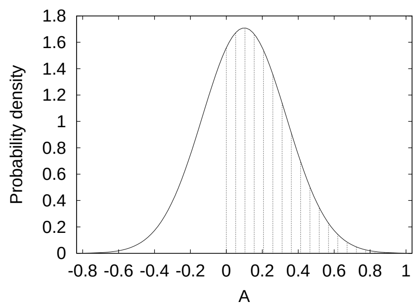
7.8.23. Five 1's and five 0's, or two 2.5's and eight 0's.

7.8.25. The probability of the five 1's and five 0's is $252 \cdot 0.5^5 0.25^5 = 0.0077$. The probability of two 2.5's and eight 0's is $45 \cdot 0.25^2 0.25^8 = 0.0004$.

7.8.27. $N(11.25, 7.97)$. The value of the normal is $f(5.0; 11.25, 7.97) = 0.012$.
7.8.29. Each locus gives a mean of 1.875 points, with variance of 1.172. The normal approximation is $IQ \sim N(98.75, 11.72)$.
7.8.31. Each factor gives a mean of 0.45 points, with variance of 0.202. The normal approximation is $IQ \sim N(93.5, 6.06)$.
7.8.33. Let I_i be a random variable giving the number of immigrants in the ith year. We found that $\mathrm{E}(I_i) = 0.1$, and can compute that $\mathrm{Var}(I_i) = 1.09$. Then $A_{20} = \frac{1}{20}\sum_{i=0}^{i=20} I_i$ has mean $\mathrm{E}(A_{20}) = 0.1$ and variance $\sigma^2 = \frac{1.09}{20} = 0.0545$. Then $\sigma = 0.233$, and the p.d.f. approximating A_{20} is

$$f(x) = \frac{1}{0.233\sqrt{2\pi}} e^{-\frac{(x-0.1)^2}{1.09}}.$$

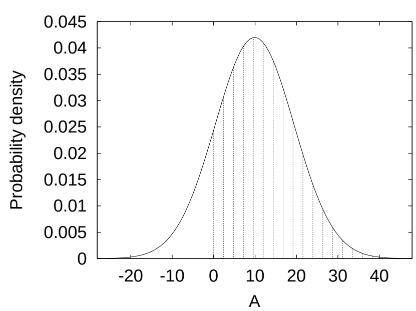
The maximum possible average change is 2. The area looks like about 0.6 of the total to me.



7.8.35. Let I_i be a random variable giving the number of immigrants in the ith year. We found that $\mathrm{E}(I_i) = 9.9$, and can compute that $\mathrm{Var}(I_i) = 902.69$. Then $A_{10} = \frac{1}{10}\sum_{i=0}^{i=10} I_i$ has mean $\mathrm{E}(A_{10}) = 9.9$ and variance $\sigma^2 = \frac{902.69}{10} = 90.269$. Then $\sigma = 9.501$, and the p.d.f. approximating A_{10} is

$$f(x) = \frac{1}{9.501\sqrt{2\pi}} e^{-\frac{(x-9.9)^2}{180.538}}.$$

The area looks like about 0.8. The maximum possible average decrease is -1, but this is only about 1 standard deviation below the mean. At least this part of the normal approximation is not very accurate.



7.8.37. We have that $P_{50} = R_1 \cdot R_2 \cdots R_{50} P_0$, so

$$\ln(P_{50}) = \ln(100) + \sum_{i=1}^{50} \ln(R_i).$$

By the Central Limit Theorem for sums, we can find the normal distribution that approximates $\ln(P_{50})$ from the mean and variance of $\ln(R_i)$. But $\mathrm{E}(\ln(R_i)) = \ln(4)\cdot 0.5 + \ln(0.25)\cdot 0.5 = 0$ and $\mathrm{Var}(\ln(R_i)) = \ln(4)^2 \cdot 0.5 + \ln(0.25)^2 \cdot 0.5 = 1.92$. After adding on the $\ln(100) = 4.6$ to the mean, we get that $\ln(P_{50}) \sim N(4.6 + 50\cdot 0, 1.92\cdot 50) = N(4.6, 96)$. The standard deviation is 9.8, so $\ln(P_{50})$ will lie between $4.6 + 2\cdot 9.8 = 24.2$ and

$4.6-2\cdot9.8=-15.0$ about 95% of the time. Exponentiating, the population itself will lie between $e^{24.2}=3.24\times10^{10}$ and $e^{-15.0}=3.06\times10^{-7}$ about 95% of the time. Even though it doesn't change on average, there is a huge range of possible populations.

7.8.39. The log population is approximately

$$\log(P_{50})\sim N(50\mathrm{E}(\ln(R)),50\mathrm{Var}(\ln(R)))$$

But

$$\begin{aligned}\mathrm{E}(\ln(R)) &= \int_{1.0}^{1.2}5.0\ln(x)dx=5.0(x\ln(x)-x)|_{1.0}^{1.2}=0.0939\\ \mathrm{Var}(\ln(R)) &= \int_{1.0}^{1.2}5.0\ln(x)^2dx-0.0939^2=5.0(x\ln(x)^2-2x\ln(x)+2x)|_{1.0}^{1.2}-0.0939^2=0.0027.\end{aligned}$$

Therefore,

$$\ln(P_{50}(t))\sim N(50\cdot0.0939,50\cdot0.0027)=N(4.695,0.135)$$

The standard deviation is $\sqrt{0.135}=0.37$, so the log population size are likely to lie between $4.695+2\cdot0.37=5.44$ and $4.695-2\cdot0.37=3.95$, and the actual population size to lie between $e^{5.44}=230$ and $e^{3.95}=52$. The simulations in Figure 6.2b end up at about 90 and 190, right in this range.