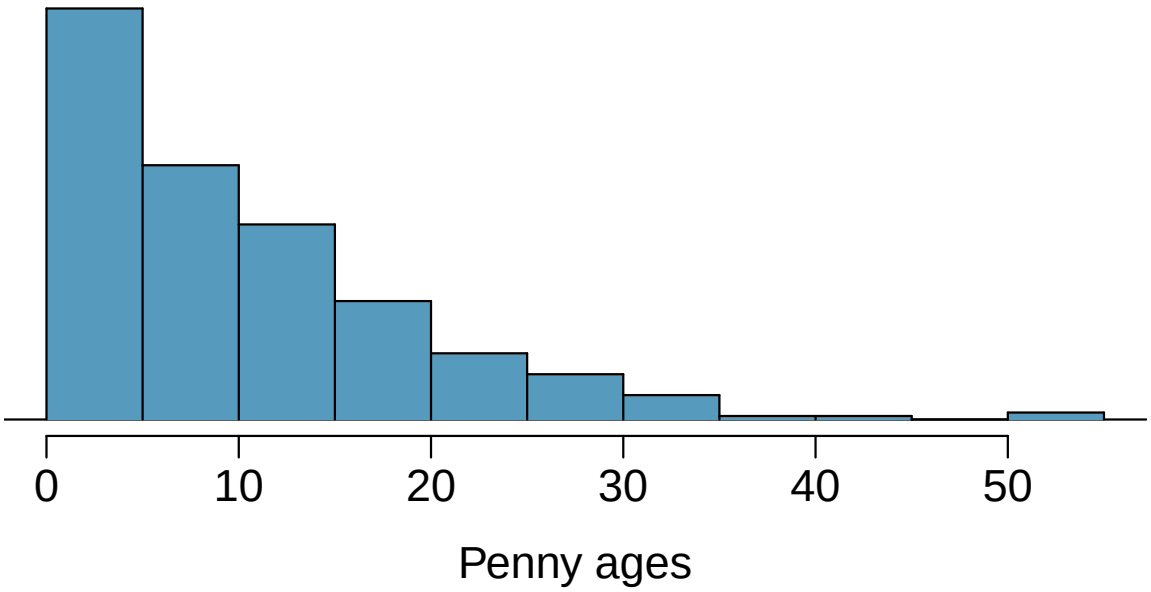
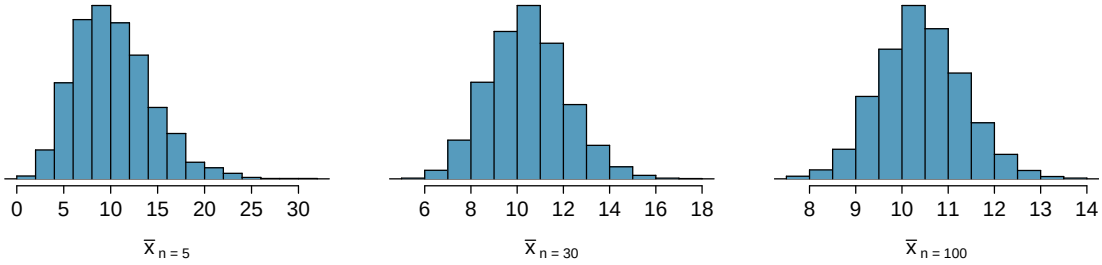
List of exercises

Exercise 1 *Ages of Pennies.* 1
Exercise 2 *Weights of Pennies.* 2
Exercise 3 *Housing Prices.* 2
Exercise 4 *Songs on an iPod.* 3
Exercise 5 *Traffic Lights.* 3
Exercise 6 *Binary Communication Channel.* 3
Exercise 7 *Airplanes.* 4
Exercise 8 *HDL.* 4
Exercise 9 *Rayleigh Distribution.* 5
Exercise 10 *Disease Test.* 5
Exercise 11 *Shear Strength.* 5

** **Ex. 1** — **Ages of Pennies.** The histogram below shows the distribution of ages of pennies at a bank.



1. Describe the distribution.
2. Sampling distributions for means from simple random samples of 5, 30, and 100 pennies is shown in the histograms below. Describe the shapes of these distributions and comment on whether they look like what you would expect to see based on the Central Limit Theorem.



3. The mean age of the pennies from Question 1 is 10.44 years with a standard deviation of 9.2 years. Using the Central Limit Theorem, calculate the means and standard deviations of the distribution of the mean from random samples of size 5, 30, and 100. Comment on whether the sampling distributions shown in Question 2 agree with the values you compute.

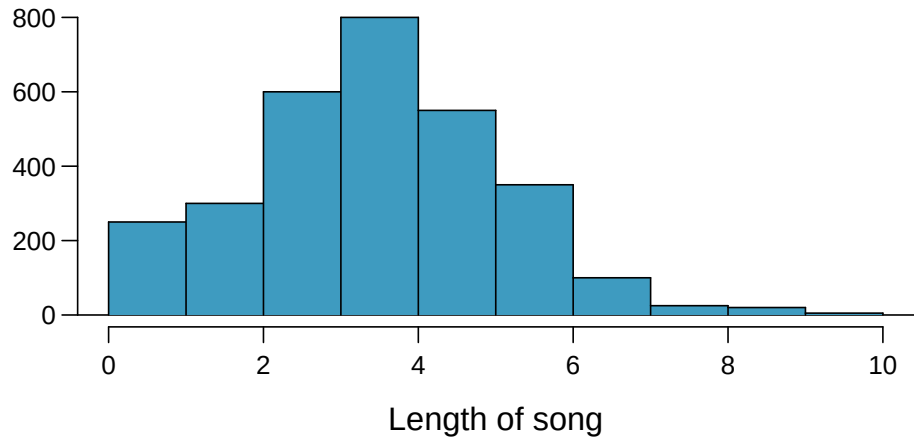
** **Ex. 2** — **Weights of Pennies.** The distribution of weights of US pennies is approximately normal with a mean of 2.5 grams and a standard deviation of 0.03 grams.

1. What is the probability that a randomly chosen penny weighs less than 2.4 grams?
2. Describe the sampling distribution of the mean weight of 10 randomly chosen pennies.
3. What is the probability that the mean weight of 10 pennies is less than 2.4 grams?
4. Sketch the two distributions (population and sampling) on the same scale.
5. Could you estimate the probabilities from 1. and 3. if the weights of pennies had a skewed distribution?

** **Ex. 3** — **Housing Prices.** A housing survey was conducted to determine the price of a typical home in Topanga, CA. The mean price of a house was roughly €1.3 million with a standard deviation of €300,000. There were no houses listed below €600,000 but a few houses above €3 million.

1. Is the distribution of housing prices in Topanga symmetric, right skewed, or left skewed? *Hint:* Sketch the distribution.
2. Would you expect most houses in Topanga to cost more or less than $1.3 million?
3. Can we estimate the probability that a randomly chosen house in Topanga costs more than $1.4 million using the normal distribution?
4. What is the probability that the mean of 60 randomly chosen houses in Topanga is more than $1.4 million?
5. How would doubling the sample size affect the standard error of the mean?

*** **Ex. 4** — **Songs on an iPod.** Suppose an iPod has 3,000 songs. The histogram below shows the distribution of the lengths of these songs. We also know that, for this iPod, the mean length is 3.45 minutes and the standard deviation is 1.63 minutes.



1. Calculate the probability that a randomly selected song lasts more than 5 minutes.
2. You are about to go for an hour run and you make a random playlist of 15 songs. What is the probability that your playlist lasts for the entire duration of your run? *Hint:* If you want the playlist to last 60 minutes, what should be the minimum average length of a song?
3. You are about to take a trip to visit your parents and the drive is 6 hours. You make a random playlist of 100 songs. What is the probability that your playlist lasts the entire drive?

*** **Ex. 5** — **Traffic Lights.** There are two traffic lights on a commuter's route to and from work. Let X_1 be the number of lights at which the commuter must stop on his way to work, and X_2 be the number of lights at which he must stop when returning from work. Suppose these two variables are independent, each with pmf given in the accompanying table (so X_1, X_2 is a random sample of size n=2).

X	0	1	2
p(X)	.2	.5	.3

1. Determine the population mean and the population variance.
2. Determine the pmf of $T_0 = X_1 + X_2$
3. Calculate μ_{To}. How does it relate to μ, the population mean?
4. Calculate σ^2_{To}. How does it relate to σ^2, the population variance?

*** **Ex. 6 —** **Binary Communication Channel.** A binary communication channel transmits a sequence of bits (0s and 1s). Supposed that for any particular bit transmitted, there is a 10% chance of a transmission error (a 0 becoming a 1 or a 1 becoming a 0). Assume that bit errors occur independently of one another.

1. Consider transmitting 1000 bits. What is the approximate probability that at most 125 transmission errors occur (use the continuity correction, so 125.5)?
2. Suppose the same 1000-bit message is sent two different times independently of one another. What is the approximate probability that the number of errors in the first transmission is within 50 of the number of errors in the second?

*** **Ex. 7 —** **Airplanes.** Two airplanes are flying in the same direction in adjacent parallel corridors. At time t=0, the first airplane is 10 km ahead of the second one. Suppose the speed of the first plane (km/hr) is normally distributed with mean 520 and standard deviation 10 and the second plane?s speed is also normally distributed with mean and standard deviation 500 and 10, respectively.

1. What is the probability that after 2 hr of flying, the second plane has not caught up to the first plane?

*** **Ex. 8 —** **HDL.** The National Health and Nutrition Examination Survey (NHANES) collects demographic, socioeconomic, dietary, and health related information on an annual basis. Here is a sample of 20 observations on HDL cholesterol level (mg/dl) obtained from the 2009–2010 survey (HDL is 'good' cholesterol; the higher its value, the lower the risk for heart disease): (see Excel file, variable 'HDL')

1. Calculate a point estimate of the population mean HDL cholesterol level.
2. Making no assumptions about the shape of the population distribution, calculate a point estimate of the value that separates the largest 50% of HDL levels from the smallest 50%.
3. Calculate a point estimate of the population standard deviation.
4. An HDL level of at least 60 is considered desirable as it corresponds to a significantly lower risk of heart disease. Making no assumptions about the shape of the population distribution, estimate the proportion p of the population having an HDL level of at least 60.

*** **Ex. 9 —** **Rayleigh Distribution.** Let $X_1, X_2, ..., X_n$ represent a random sample from a Rayleigh distribution with pdf

$$f(x;\theta) = \frac{x}{\theta} e^{\frac{-x^2}{2\theta}}$$

1. It can be shown that $E(X^2) = 2\theta$. Use this fact to construct an unbiased estimator of θ based on $\sum x_i^2$ (and use rules of expected value to show that it is unbiased).
2. Estimate θ from the following n=10 observations on vibratory stress of a turbine blade under specified conditions (see Excel file, variable 'stress')

*** **Ex. 10 —** **Disease Test.** A diagnostic test for a certain disease is applied to n individuals known to not have the disease. Let X = the number among the n test results that are positive (indicating presence of the disease, so X is the number of false positives) and p = the probability that a disease–free individual's test result is positive (i.e., p is the true proportion of test results from disease-free individuals that are positive). Assume that only X is available rather than the actual sequence of test results.

1. Derive the maximum likelihood estimator of p.
2. Is the estimator of part (a) unbiased?
3. If n=20 and x=3, what is the mle of the probability $(1-p)^5$ that none of the next five tests done on disease–free individuals are positive?

*** **Ex. 11 —** **Shear Strength.** The shear strength of each of ten test spot welds is determined, yielding the following data (see Excel file, variable 'strength')

1. Assuming that shear strength is normally distributed, estimate the true average shear strength and standard deviation of shear strength using the method of maximum likelihood.
2. Again assuming a normal distribution, estimate the strength value below which 95% of all welds will have their strengths. (Hint: What is the 95th percentile in terms of μ and σ? Now use the invariance principle.)
3. Suppose we decide to examine another test spot weld. Let X = shear strength of the weld. Use the given data to obtain the mle of P(X$\leq$ 400)

Answer of exercise 1

1. The distribution is unimodal and strongly right skewed with a median between 5 and 10 years old. Ages range from 0 to slightly over 50 years old, and the middle 50% of the distribution is roughly between 5 and 15 years old. There are potential outliers on the higher end.
2. When the sample size is small, the sampling distribution is right skewed, just like the population distribution. As the sample size increases, the sampling distribution gets more unimodal, symmetric, and approaches normality. The variability also decreases. This is consistent with the Central Limit Theorem.
3. According to the CLT, we expect here to obtain sampling distributions with smaller sampling error (since the sampling error, *i.e.*, the standard deviation of the sampling distribution, is inversely proportional to $\sqrt{n}$). Of course, the mean of the three distributions will remain the same.
   Let us summarize the information as follows:

$$\begin{array}{lcll} n = 5 & \rightarrow & \mu_{\bar{x}} = 10.44 & \sigma_{\bar{x}} = 9.2/\sqrt{5} = 4.11 \\ n = 30 & \rightarrow & \mu_{\bar{x}} = 10.44 & \sigma_{\bar{x}} = 9.2/\sqrt{30} = 1.68 \\ n = 100 & \rightarrow & \mu_{\bar{x}} = 10.44 & \sigma_{\bar{x}} = 9.2/\sqrt{100} = 0.92 \end{array}$$

   The centers of the sampling distributions shown in the pictures appear to be around 10. It is difficult to estimate the standard deviation for the sampling distribution when n = 5 from the histogram (since the distribution is somewhat skewed). If 1.68 is a plausible estimate for the standard deviation of the sampling distribution when n = 30, then using the 68-95-99.7% Rule, we would expect the values to range roughly between $10.44 \pm 3 \times 1.68 = (5.4, 15.48)$, which seems to be the case. Similarly, when n = 100, we would expect the values to range roughly between $10.44 \pm 3 \times 0.92 = (7.68, 13.2)$, which also seems to be the case.

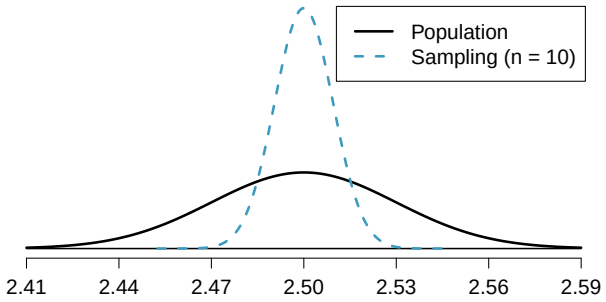
Answer of exercise 2

1. Assuming a normal distribution with mean $\mu = 2.5$ grams and standard deviation $\sigma = 0.03$ grams, we first compute the z score as:

$$z = \frac{x - \mu}{\sigma} = \frac{2.4 - 2.5}{0.03} = -3.33$$

   Now, from the table/Statistical software we derive that $p(z \leq -3.33) = 0.0004$.
2. The population SD is known and the data are nearly normal, so the sample mean will be nearly normal with distribution $N(\mu, \sigma/\sqrt{n})$, i.e. $N(2.5, 0.0095)$.
3. $Z = -10.53 \rightarrow \approx 0$.

4. Consider the picture below, where we observe both the population distribution and the distribution of the sampling process:



5. We could not estimate the answer to question 1 without a nearly normal population distribution. We also could not estimate the answer to question 3, since the sample size is not sufficient to yield a nearly normal sampling distribution if the population distribution is not nearly normal.

Answer of exercise 3

1. Right skewed. There is a long tail on the higher end of the distribution but a much shorter tail on the lower end.
2. Less than, as the median would be less than the mean in a right skewed distribution.
3. We should not.
4. Even though the population distribution is not normal, the conditions for inference are reasonably satisfied, with the possible exception of skew. If the skew isn't very strong (we should ask to see the data), then we can use the Central Limit Theorem to estimate this probability. For now, we'll assume the skew isn't very strong, though the description suggests it is at least moderate to strong. Use $N(1.3, SE_{\bar{x}} = 0.3/\sqrt{60})$: $Z = 2.58 \rightarrow 0.0049$.
5. It would decrease it by a factor of $1/\sqrt{2}$.

Answer of exercise 4

1. We cannot use the normal model for this calculation, but we can use the histogram. About 500 songs are shown to be longer than 5 minutes, so the probability is about $500/3000 = 0.167$.
2. Two different answers are reasonable:

- *Option 1.* Since the population distribution is only slightly skewed to the right, even a small sample size will yield a nearly normal sampling distribution. We also know that the songs are sampled randomly and the sample size is less than 10% of the population, so the length of one song in the sample is independent of another. We are looking for the probability that the total length of 15 songs is more than 60 minutes, which means that the average song should last at least $60/15 = 4$ minutes. Using $SE = 1.63/\sqrt{15}$, $Z = 1.31 \rightarrow 0.0951$.
- *Option 2.* Since the population distribution is not normal, a small sample size may not be sufficient to yield a nearly normal sampling distribution. Therefore, we cannot estimate the probability using the tools we have learned so far.

3. Since the sample size is large, we can now be confident that the conditions are satisfied. $Z = 0.92 \rightarrow 0.1788$.

Answer of exercise 5

a. Since each X is 0 or 1 or 2, the possible values of T_o are 0, 1, 2, 3, 4.
$P(T_o = 0) = P(X_1 = 0 \text{ and } X_2 = 0) = (.2)(.2) = .04$ since X_1 and X_2 are independent.
$P(T_o = 1) = P(X_1 = 1 \text{ and } X_2 = 0, \text{ or } X_1 = 0 \text{ and } X_2 = 1) = (.5)(.2) + (.2)(.5) = .20$.
Similarly, $P(T_o = 2) = .37$, $P(T_o = 3) = .30$, and $P(T_o = 4) = .09$. These values are displayed in the pmf table below.

t_o	0	1	2	3	4
$p(t_o)$	.04	.20	.37	.30	.09

b. $E(T_o) = 0(.04) + 1(.20) + 2(.37) + 3(.30) + 4(.09) = 2.2$. This is exactly twice the population mean: $E(T_o) = 2\mu$.

c. First, $E(T_o^2) = 0^2(.04) + 1^2(.20) + 2^2(.37) + 3^2(.30) + 4^2(.09) = 5.82$. Then $V(T_o) = 5.82 - (2.2)^2 = .98$. This is exactly twice the population variance: $V(T_o) = 2\sigma^2$.

Answer of exercise 6

a. Let X = the number of erroneous bits out of 1000, so $X \sim \text{Bin}(1000, .10)$. If we approximate X by a normal rv with $\mu = np = 100$ and $\sigma^2 = npq = 90$, then with a continuity correction $P(X \le 125) = P(X \le 125.5) \approx P\left(Z \le \frac{125.5 - 100}{\sqrt{90}}\right) = P(Z \le 2.69) = \Phi(2.69) = .9964$.

b. Let Y = the number of errors in the second transmission, so $Y \sim \text{Bin}(1000, .10)$ and is independent of X. To find $P(|X - Y| \le 50)$, use the facts that $E[X - Y] = 100 - 100 = 0$ and $V(X - Y) = V(X) + V(Y) = 90 + 90 = 180$. So, using a normal approximation to both binomial rvs, $P(|X - Y| \le 50) \approx P\left(|Z| \le \frac{50}{\sqrt{180}}\right) = P(|Z| \le 3.73) \approx 1$.

Answer of exercise 7

Let X_1 and X_2 denote the (constant) speeds of the two planes.

a. After two hours, the planes have traveled $2X_1$ km and $2X_2$ km, respectively, so the second will not have caught the first if $2X_1 + 10 > 2X_2$, i.e. if $X_2 - X_1 < 5$.
$X_2 - X_1$ has a mean $500 - 520 = -20$, variance $100 + 100 = 200$, and standard deviation 14.14. Thus,
$P(X_2 - X_1 < 5) = P\left(Z < \frac{5 - (-20)}{14.14}\right) = P(Z < 1.77) = .9616$.

Answer of exercise 8

a. A sensible point estimate of the population mean μ is the sample mean, $\bar{x} = 49.95$ mg/dl.

b. Of interest is the population median, $\tilde{\mu}$. The logical point estimate is the sample median, $\tilde{x}$, which is the average of the 10^{th} and 11^{th} ordered values: $\tilde{x} = 47.5$ mg/dl.

c. The point estimate of the population sd, σ, is the sample standard deviation, $s = 16.81$ mg/dl.

d. The natural estimate of $p = P(X \ge 60)$ is the sample proportion of HDL observations that are at least 60. In this sample of $n = 20$ observations, 4 are 60 or higher, so the point estimate is $\hat{p} = 4/20 = .2$.

Answer of exercise 9

a. $E(X^2) = 2\theta$ implies that $E\left(\frac{X^2}{2}\right) = \theta$. Consider $\hat{\theta} = \frac{\sum X_i^2}{2n}$. Then

$E\left(\hat{\theta}\right) = E\left(\frac{\sum X_i^2}{2n}\right) = \frac{\sum E\left(X_i^2\right)}{2n} = \frac{\sum 2\theta}{2n} = \frac{2n\theta}{2n} = \theta$, implying that $\hat{\theta}$ is an unbiased estimator for θ.

b. $\sum x_i^2 = 1490.1058$, so $\hat{\theta} = \frac{1490.1058}{20} = 74.505$.

Answer of exercise 10

a. To find the mle of p, we'll take the derivative of the log-likelihood function

$\ell(p) = \ln\left[\binom{n}{x} p^x (1-p)^{n-x}\right] = \ln\binom{n}{x} + x\ln(p) + (n-x)\ln(1-p)$, set it equal to zero, and solve for p.

$\ell'(p) = \frac{d}{dp}\left[\ln\binom{n}{x} + x\ln(p) + (n-x)\ln(1-p)\right] = \frac{x}{p} - \frac{n-x}{1-p} = 0 \Rightarrow x(1-p) = p(n-x) \Rightarrow p = x/n$, so the mle of p is $\hat{p} = \frac{x}{n}$, which is simply the sample proportion of successes. For $n = 20$ and $x = 3$, $\hat{p} = \frac{3}{20} =$ .15.

b. Since X is binomial, $E(X) = np$, from which $E(\hat{p}) = E\left(\frac{X}{n}\right) = \frac{1}{n}E(X) = \frac{1}{n}(np) = p$; thus, $\hat{p}$ is an unbiased estimator of p.

c. By the invariance principle, the mle of $(1-p)^5$ is just $(1-\hat{p})^5$. For $n = 20$ and $x = 3$, we have $(1 - .15)^5 = .4437$.

Answer of exercise 11

a. $\hat{\mu} = \bar{x} = 384.4; s^2 = 395.16$, so $\frac{1}{n}\sum(x_i - \bar{x})^2 = \hat{\sigma}^2 = \frac{9}{10}(395.16) = 355.64$ and $\hat{\sigma} = \sqrt{355.64} = 18.86$ (this is not s).

b. The 95th percentile is $\mu + 1.645\sigma$, so the mle of this is (by the invariance principle) $\hat{\mu} + 1.645\hat{\sigma} = 415.42$.

c. The mle of $P(X \le 400)$ is, by the invariance principle, $\Phi\left(\frac{400 - \hat{\mu}}{\hat{\sigma}}\right) = \Phi\left(\frac{400 - 384.4}{18.86}\right) = \Phi(0.83) =$ .7967.