Corrected: Answers to Odd Exercises

Chapter 1

1. (a) Population: high-school students. Sample: 2000 high-school students. Statistic 47% (b) Population: people living in the U.S. Parameter: 13.9%.
3. (a) Observational study. (b) No, we cannot establish causation. The participants were not randomized into the two groups. (c) No, we cannot generalize to all urban American adolescents.
5. n/N.

Chapter 2

1. $\bar{x} = 12$, $m = 11.5$. $\tilde{x} = 2.26$, $\tilde{m} = 2.39$. No, not the same.
3. (a) No. For example, consider $1 < 3 < 5$ with $f(x) = x^2$. (b) No. For example, consider $1 < 3 < 6$ with $f(x) = (x-3)^2$. (c) f is linear. (d) f is an increasing (or decreasing) function and n is odd, or f is linear.
5. (a) Favor: 899; Oppose: 409. (b) The `summary` command indicates that there were many non-responses. The `table` command does not give any indication that there were non-responses.
   (c)

	Death Penalty	
Own Gun?	Favor	Oppose
No	375	199
Refused	7	2
Yes	243	59

   (d) 80.4% of gun owners favor the death penalty, 65.3% of non-gun owners favor the death penalty.
7. (a)True. $\dfrac{w_1 + w_2 + \cdots + w_n}{n} = \dfrac{n\bar{x} + n\bar{y}}{n} = \bar{x} + \bar{y}$. (b) True.
9. Hint: Show that the function is symmetric about θ: $f(\theta - x) = f(\theta + x)$.
11. $F^{-1}(y) = a\sqrt{y}$, so $a\sqrt{\alpha/2}$ and $a\sqrt{1-\alpha/2}$ are the desired quantiles.
13. No, the discrete nature of this random variable results in $P(X \leq 1) = 0.007$ and $P(X \leq 2) = 0.035$.
15. (a) 3; (b) 4; (c) 6.

Chapter 3

1. (a) $\bar{X}_t - \bar{X}_c = 11 - 7 = 4$. (c) $t = 4$ = observed statistic; P-value = 2/10 = 0.2. (d) 0.2.

3. (a) Observed test statistic 5.886; One simulation gives a P-value $= 0.0003$ so we conclude that the difference in mean delay times between the two carriers is statistically significant. (b) Observed test statistic -5.663; One simulation gives P-value $= 0.0001$, so the difference in mean delay times between May and June is statistically significant.

5. Code provided at `https://sites.google.com/site/ChiharaHesterberg`.

7. (a) The difference in proportions is statistically significant. (P-value = 0.017). (b) The ratio of the variances is not 1 (P-value = 0.374).

9. (b) Mean number of strike-outs in away games: 7.31; mean number of strike-outs in home games: 6.95; (c) Observed test statistic is 0.358. P-value=0.21.

11. $c = 6.6814$, df= 2, P-value $= 0.0354$. Therefore, we conclude there is an association between age and support of marijuana for medicinal purposes.

13. (a) This is a test of homogeneity. Let π_{ij} denote the proportion of fish from region i ($i = 1, 2, 3$) and j rays, ($j =\geq 36, 35, 34, 33, 32, \leq 31$). Then H_0 : $\pi_{1j} = \pi_{2j} = \pi_{3j}$ versus H_a : at least one pair of proportions not the same. (b) $c = 12.803$, df $= 10$, P-value $= 0.23$. We conclude there is not enough evidence to support the hypothesis that fin ray counts differ across the regions.

15. $c = 13.4621$, df $= 1$, P-value $= 0.0002$. The difference between the two airlines is statistically significant.

17. P-value=0.005. Gender and happiness are not independent.

19. (a) Marginal probabilities do not change; degree of freedom will not change either.

21. $c = 8.58$, df $= 5 - 1 = 4$, P-value is 0.074. Yes, it is plausible that these numbers are drawn from a distribution with pdf $f(x) = 2/x^3, x \geq 1$.

23. (a) Using the intervals: $(0, 17], (17, 22], (22, 28], (28, 33], (33, 44]$, then $c = 28.388$, df $= 5 - 0 - 1 = 4$, P-value $= 0$. Test statistics may vary depending on the cut-offs used for the intervals. Conclude that the data do not come from $N(25, 10^2)$.

25. $c = 0.736$, df $= 4$, P-value=0.94. The numbers appear to be drawn randomly. Answers will vary depending on the intervals used.

27. (a) Expected counts: 3.43,11.57, 4.57, 15.42. Two are less than 5. (b) $\binom{20}{2}\binom{15}{6}/\binom{35}{8} = 0.0404$; (c) $\sum_{k=0}^{2}\binom{20}{k}\binom{15}{8-k}/\binom{35}{k} = 0.0461$; (d) If the selection had been completely random, then the chance of obtaining a committee consisting of 2 or fewer seniors is only 4.6%, which gives (mild) evidence that there is cause for suspicion.

Chapter 4

1. There are 20 possible sets of size 3. The mean of the medians is 5.7. The median of the population is 5.5.

3. (a) $\{6, 8, 8, 9, 10, 10, 10, 11, 11, 12, 13, 14\}$. (b) No. (c) $\mathrm{E}[X]+\mathrm{E}[Y] = 10.25$.

5. 0.001

7. 0.332

9. 0.022

11. 91.

13. (a) $W = \bar{X} + \bar{Y} \sim N(36, 3.11^2)$. (c) 0.901

15. See Theorem **??**

17. Simulated answers should be close to: (b) 10, 5; (c) 0.5

19. (a) 13.86

21. (a) $f_{max}(x) = 8(1/x^2 - 1/x^3)$. (b) 1.545

23. $120e^{-12x}(1 - e^{-12x})^9$.

25. $P(X = k) = 30^k e^{-30}/k!$, $k = 0, 1, \ldots$

27. (a) Yes. (c) 4.471

Chapter 5

1. Answers will vary.

3. (a) 3^3, order matters. (b) 10 distinct. (c) 3. (d) No.

5. (a) $\binom{9}{5}$. (b) $\binom{2n-1}{n}$.

7. (a) $N(36, 8^2/200)$. (c) Bootstrap distribution using sample in (b): mean 35.91, se 0.53. (d)

	Mean	Standard Deviation
Population	36	8
Sampling distribution of $\bar{X}$	36	$8/\sqrt{200} = 0.566$
Sample	35.01	7.386
Bootstrap distribution	35.91	0.53

9. For odd n, median will be one of the sample points. Thus for small n, there will be only n possible values for the median, so the sampling distribution is much more "granular" than when n is even. As n increases, this becomes less apparent.

11. About $(13.68, 23.23)$. Bias is about 25%.

13. (a) Bell-shaped, mean at about 5.196, st. dev. at about 1.43. (b) (2.33, 8). (c) For the bootstrap distribution, we sample with replacement from the original sample: we sample from the men, and then from the women. With the permutation distribution, we sample assuming there is no difference in the means; in particular, we sample without replacement from the pooled data.

15. $(3.49, 11.49)$

17. (c) 16.3, se 0.319. (d) $(1.17, 2.22)$.

Chapter 6

1. $\hat{p} = X/n$

3. $\hat{\theta} = 4/(\sqrt{5} + 3 + 3 + \sqrt{10}) = 0.351$.

5. (a) $\hat{\mu} = \bar{X}$.
$\hat{\sigma} = \sqrt{(1/n)\sum_{i=1}^{n}(X_i - \mu)^2}$.

7. Does not exist.

9. $\hat{N} = 9$.

11. $\hat{\lambda} = (n+m)/(\sum_{i=1}^{n} X_i + 2\sum_{j=1}^{m} Y_j)$.

13. (a) $\hat{\alpha} = n/(\sum_{i=1}^{n} x_i^{\beta})$.
(b)

$$n/\alpha = \sum_{i=1}^{n} x_i^{\beta}$$

$$\alpha \sum_{i=1}^{n} X_i^{\beta} \ln(x_i) - n/\beta = n \sum (\ln(x_i))$$

15. $\hat{r} = 22.88, \hat{\lambda} = 3.138$.

17. Shape=0.917, scale= 17.344. c=14.217, P-value = 0.047, so there is marginal evidence that times between successive earthquakes does not follow a Weibull distribution.

19. (a) $\bar{X}/(\bar{X} - 2)$.

21. p

23. $\theta + 1$, so $\bar{X}$ is biased.

25. $\sum_{i=1}^{n} a_i = 1$.

27. (a) $-\sigma^2/n$. (b) $(\sigma^4/n^2) \cdot 2(n-1)$. (c) $2(n-1)\sigma^4/n^2 + (\sigma^2/n)^2$

29. $\mathrm{Var}[\hat{\theta_1}] = \theta^2$, $\mathrm{Var}[\hat{\theta_2}] = (1/4)(2\theta^2) = (1/2)\theta^2$ and $\mathrm{Var}[\hat{\theta_3}] = (1/9)(\theta^2 + 2^2\theta^2) = (5/9)\theta^2$, so $\hat{\theta_2}$ is the most efficient and $\hat{\theta_1}$ is the least efficient.

31. The two curves approach each other.

33. (a) $2/(3\theta)$; (b) Bias: $-17/(27\theta)$; MSE $589/(1458\theta^2)$. (c) 27 T/10, where $T = X_1/9 + X_2/9 + X_3/3$.

35. (a) $f_{max}(x) = (n\alpha x^{\alpha n-1})/(\beta^{\alpha n})$. (b) $(n\alpha\beta)/(\alpha n + 1)$. (c) Bias: $-(\beta)/(\alpha n+1)$. (d) MSE: $(2\beta^2)/((\alpha n+1)(\alpha n+2))$

37. (d) $(\pi - 4)/(4\lambda)$.

Chapter 7

1. (a) Population mean here is not random. The probability that it is contained in the interval is 0 or 1. (c) Not 95% of the time, but 95% of the confidence intervals generated by each sample will contain true mean. (e) Each sample gives rise to a different confidence interval and 95% of these intervals will contain the true mean.
3. (a) (201.8, 218.2); (b) 97; (c) 166.
5. 4n
7. 118.01
9. (9.19, 13.39) cm.
11. (b)(-533.61, -83.294)
13. (b) $(11.47, \infty)$. We are 95% confident that on average, seedlings grown in fertilized plots grow at least 11.5cm more than seedlings grown in non-fertilized plots.
15. (b) $(59.451, 70.266)$; (c) Yes, $(59.1, 66.8)$. The upper end point decreased 3 inches.
17. Unbalanced sample sizes, pooled CI's are not capturing true mean difference at 95%. Balanced case, pooled CI does better.
19. $[5.14, \infty)$.
21. (a) 1064; (b) 968
23. (a) $(0.07, 0.13)$; (c) Yes, the intervals overlap; thus, we cannot make the firm conclusion that taking the drug makes a difference.
25. (b) $(52.87, 103.29)$; (c) $(54.78, 105.01)$. Bootstrap t: $(56.84, 112.15)$. Bootstrap t. Intervals will vary slightly.
27. (a) 1497 non-smokers, 90 smokers; (c) $(-27.68, 190.69)$. This interval contains 0. Percentile: $(-24.79, 187.13)$. Bootstrap t: $(-27.46, 190.27)$. They are all roughly the same. Report formula t or bootstrap t.
29. $(95.76, 155.51)$.

Chapter 8

1. P-value=0.237, so we conclude that mean calcium levels are the same.
3. P-value=0.007, so we conclude that on average, body temperatures are higher for children in Sodor.
5. P-value=0.0002 so we conclude that mean pH levels are different at the two locations.

7. P-value=0.001; With 95% confidence, we conclude that, on average, ales have 7.1 more calories than lagers.

9. P-value=0.0187. Evidence supports the hypothesis that the proportion of sex workers in Bamako who are HIV positive is greater than 0.391.

11. P-value=0.01. Evidence supports the claim that the proportions in the two regions are not the same.

13. Hint: Use the `rbinom` command to draw random samples. For example, `rbinom(1000, 10, .5)` will result in 1000 samples of size 10 with $p = 0.5$.

15. (a) Type I error: we conclude that mean arsenic levels in the community are higher than 10 ppb even though it is in fact equal than or equal to 10 ppb. The community may take unnecessary and expensive measures to lower the arsenic levels. Type II error: we conclude that mean arsenic levels in the community are less than or equal to 10 ppb even though in reality they are actually higher than 10 ppb. Community is drinking unsafe water and not taking any action. (b) Not necessarily, since the t-distribution has fatter tails.

17. 8

19. $n \geq 15$.

21. $X \leq 22$ or $X \geq 38$.

23. (a) 0.04; (b) 0.473.

25. (a) 0.0497; (b) 0.548

27. $1 - (3/4)^{\theta+1}$

29. $1 - \beta = P\left(Z < q_1 + \dfrac{\mu_0 - \mu_1}{\sigma/\sqrt{n}}\right) + P\left(Z < q_2 - \dfrac{\mu_1 - \mu_0}{\sigma/\sqrt{n}}\right)$

31. 0.0003

33. (b) 0.224

Chapter 9

1. -1/100

3. 133

5. 1

7. (a) 0.4996; (c) 0.992

9. When solving for a and b that minimizes $g(a, b)$, we take the partial derivative: $\dfrac{\partial g}{\partial a} = 2\sum_{i=1}^{n}(y_i - (a + bx_i))(-1) = 0$ Thus, $2\sum_{i=1}^{n}(y_i - (\hat{a} + \hat{b}x_i))(-1) = 0$. But $\hat{a} + \hat{b}x_i = \hat{y}_i$. So $\sum_{i=1}^{n}(y_i - \hat{y}_i) = 0$.

11. (a) $\texttt{weight} = 20.078 + 1.607\texttt{height}$; (b) 116.498; (c) $R^2 = 0.5625$; 56% of the variability in weight is explained by the model.

13. (a) Increasing variance; (c) indicates that a linear model is appropriate;

15. (b) $\texttt{Kills} = 1.736 + 0.947\texttt{Assts}$. For every additional assist per set, there is associated a nearly one additional kill. About 93.7% of the variability in kills per set is explained by this model. (c) Yes, a straight-line model is appropriate.

17. (a) 0.349; (b) $\texttt{Weight} = -2379.69 + 148.995 \cdot \texttt{Gestation}$; (c) $R^2 = 0.122$; (d) The problem is that σ is not constant for different gestation lengths (in part because there are so few data points at 42 weeks).

19. $(0.6994, 1.0345)$

21. (b) $\texttt{Level} = -3280 + 1.83\texttt{Year}$; (c) No, there appears to be serial correlation.

23. 0.4397; $(0.206, 0.663)$

25. Proof omitted. 27. Proof omitted. 29. Proof omitted.

31. (a) $\ln(\hat{p}/(1-\hat{p})) = 2.905 - 0.037\texttt{Temperature}$. (b) $e^{-0.037(-5)} = 1.203$. A 5 degree decrease in temperature increases the odds of an O-ring incident by a (multiplicative) factor of 1.203. Or, there is a 20% increase in the odds of an O-ring incident for every 5 degree drop in temperature. (d) The range of temperatures in the data set is from 53 to 81 degrees. We are extrapolating quite a bit at temperature 33 degrees.

33. (a) $\ln(\hat{p}/(1-\hat{p})) = -2.23 + 0.29\texttt{Hits}$. (b) $e^{0.29} = 1.03$. Each additional hit increases the odds of winning by 3%. (d) $(0.86, 0.98)$. Answers will vary.

Chapter 10

1. (a) Posterior column: 0.954; 0.5828; 0.3221; (b) After observing 4 wins out of 5, you now believe that there is a 32.2% chance that your long term probability of winning is 0.6. (c) 0.5; 0.5227.

3. (b) After observing 4 wins out of 7, you now believe that there is a 2.3% chance that your probability of winning is 0.2. (c) 0.5, 0.55.

5. (a) (0.2048, 0.2424); (b) (0.228, 0.261); (c) 0.0693

7. 0.55; 0.0012 9. $\mu\,|\,\mathbf{x} \sim N(0.829, 0.0323^2)$

11. (a) $\mu\,|\,\mathbf{x} = 19.53$, $\sigma\,|\,\mathbf{x} = 1.4797$, precision 0.457. (c) $\mu\,|\,\mathbf{x} = 19.14$, $\sigma\,|\,\mathbf{x} = 1.531$, precision 0.4267.

13. (a) not conjugate, so computations are difficult, (b) is non-zero for $\theta < 0$ and $\theta > 1$, values which are impossible.

15. (a) $f(\theta) = 1/\theta^n$ (with $\theta > \max\{X_1, X_2, \ldots, X_N\}$; (b) Pareto distribution with parameters $\alpha + N, \beta$, (with $\theta > \max\{\beta, X_1, X_2, \ldots, X_N\}$); (c) 0.0089.

17. (a) $\theta^n \exp -\theta \sum_{i=1}^{n} X_i$; (d) Gamma with parameters $14, 17$.

Chapter 11

1. (a) The female distribution is centered about 9.3 and male distribution about 14.5, the means of the original data. The male distribution has a larger standard deviation, 1.16 vs .92; these numbers match $s/\sqrt(n)$ for each dataset. Both smoothed bootstrap distributions are very close to normal. (d) Both are very close to normal, with the same mean -5.2 (for female minus male); the smoothed bootstrap standard error is larger, 1.48 compared to 1.44; for comparison the usual formula standard error is 1.48.

3. The distribution is close to normal, with slight positive skewness. The mean is 7.30, and standard deviation 6.9. A 95% bootstrap percentile interval is $(6.01, 8.70)$.

5. (a) $r/(n\lambda^2)$ (b) $f(x) = x^2$, so $f'(\mu) = 2\mu$. $\mathrm{Var}[\bar{X}^2] \approx (2\mu)^2 r/(n\lambda^2)$. (c) $f(x) = \log(x)$, so $f'(\mu) = 1/\mu$. $\mathrm{Var}[\bar{X}^2] \approx (1/\mu)^2 r/(n\lambda^2)$. (d) $f(x) = \sqrt(x)$, so $f'(\mu) = 1/(2\sqrt{\mu}$. $\mathrm{Var}[\bar{X}^2] \approx (1/(4\mu))r/(n\lambda^2)$. (e) Pick sample size n and parameters r and λ. Generate N samples, say $N = 10^4$; for each compute $\bar{x}$, $\bar{x}^2$, $\log \bar{x}$, and $\sqrt{\bar{x}}$. Compute the variance of each set of N statistics, and compare. Double-check by repeating with other r, λ, and n.

7. (a) 0.9046; a confidence interval based on 10^6 replications. would be $0.9046 \pm 2 \cdot 0.000125$. (b) 0.9045242 with absolute error ¡ 1e-14

9. If $f(x) = \exp(-x)$ for $x > 0$ and $h(x) = \sin(x)$, then compare `integrate(function(x) sin(x) * exp(-x), 0, Inf)` and `N <- 1000000; x <- rexp(N); mean(sin(x))`.

11. Using $N = 10^6$ gives approximately: (a) $0.597 \pm 2 \cdot 0.00037$ (b) $0.597 \pm 2 \cdot 0.00025$ (d) The second one gives more accurate estimates. The corresponding graph is flatter, the h values do not vary as much.

13. Proof omitted.

Appendix B

1. (a) 0.201; (b) 9

3. (a) 0.00138; (b) 16.65