1 What is Statistics?

1.1 Introduction

E1.1 Populations of interest and inferential objectives.

(a) Population: U.S. citizens from Generation X. Objective: Estimate proportion of population interested in starting their own business. Method: Randomly sample U.S. citizens from Generation X, e.g. random digit dialing and analyzing only those of age in Generation X.

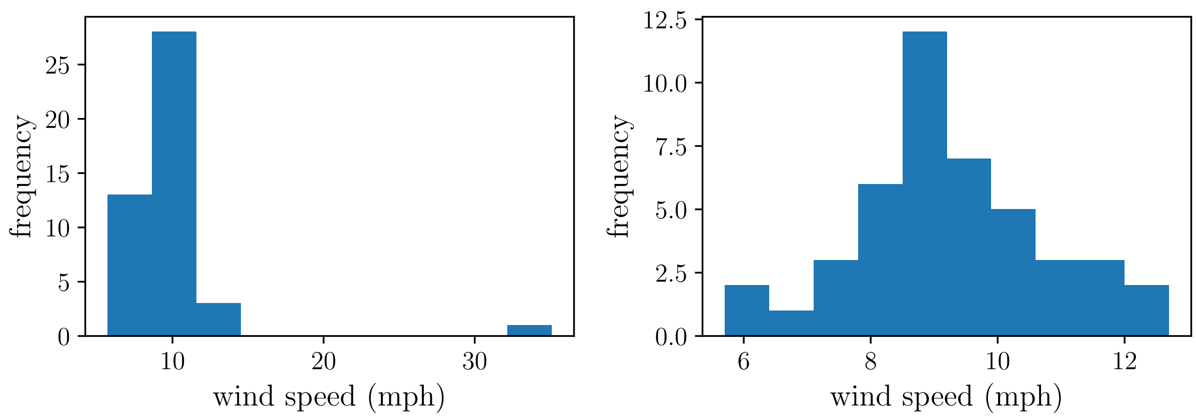
(b) Population: healthy U.S. adults. Objective: Estimate average body temperature of population. Method: RDD, like (a).

(c) Population: Single-family dwelling units in the city. Objective: Estimate average weekly water consumption of population. Method: RDD, like (a).

(d–g) Similar logic as previous parts.

1.2 Characterizing a Set of Measurements: Graphical Methods

E1.2 The Windy City, or is it? (a) Histograms with (left) and without (right) Mt. Washington.



(b) Yes. **(c)** $11/45 = 0.2\overline{4}$. **(d)** Chicago's wind speed is larger than average (and the median), but not unusually large like Mt. Washington.

E1.7 Self-reported heights. The histogram is nearly bimodal. Men and women each have an approximately normal histogram with different means; when pooled, a bimodal distribution appears.

1.3 Characterizing a Set of Measurements: Numerical Methods

E1.10 Internet usage (a) -3 hours. **(b)** $(1-0.68)/2$. **(c)** No. No user spends < 0 hours on the internet.

E1.11 A variance identity. First, expand the sum:

$$\sum_{i=1}^{n}(y_i - \bar{y})^2 = \sum_{i=1}^{n}(y_i^2 + \bar{y}^2 - 2y_i\bar{y}) = \sum_{i=1}^{n} y_i^2 + n\bar{y}^2 - 2\bar{y}\sum_{i=1}^{n} y_i.$$

Recalling $\bar{y} = \frac{1}{n}\sum_{i=1}^{n} y_i$, the last term is $-2n\bar{y}^2$, thus

$$\sum_{i=1}^{n}(y_i - \bar{y})^2 = \sum_{i=1}^{n} y_i^2 - n\bar{y}^2.$$

After substituting $\bar{y} = \frac{1}{n}\sum_{i=1}^{n} y_i$, the identity follows. $\square$

1.4 How Inferences Are Made

1.5 Theory and Reality

1.6 Summary

Supplementary Exercises

E1.22 Zero sum of deviations from their mean.

$$\sum_{i=1}^{n}(y_i - \bar{y}) = \sum_{i=1}^{n} y_i - \sum_{i=1}^{n} \bar{y} = n\bar{y} - n\bar{y} = 0. \qquad \square$$

***E1.32 Tchebysheff's theorem.** Begin with a rearranged definition of the sample variance

$$(n-1)s^2 = \sum_{i=1}^{n}(y_i - \bar{y})^2.$$

Split the sum into points in and out of the interval $I = \bar{y} \pm ks$:

$$\begin{aligned}(n-1)s^2 &= \sum_{x \in I}(y_i - \bar{y})^2 + \sum_{x \notin I}(y_i - \bar{y})^2 \\ &\geq 0 + \sum_{x \notin I}(ks)^2 \\ &= m(ks)^2\end{aligned}$$

for some $k \in \mathbb{R}$ where m is the number of points outside the interval. This can be rearranged to

$$m \leq \frac{n-1}{k^2}.$$

Now, the fraction of points in the interval is

$$\frac{n-m}{n} = 1 - \frac{m}{n} \geq 1 - \frac{n-1}{n}\frac{1}{k^2} \geq 1 - \frac{1}{k^2}. \qquad \square$$

Comment. We made no assumptions on the range of k, but it is only useful for $|k| > 1$.

E1.33 Application of Tchebysheff's theorem. Apply Tchebysheff's theorem with $k = 2$. The interval is $[0.5, 10.5]$.

E1.34 Application of Tchebysheff's theorem 2. $k = 3$ has interval $[-2, 13]$, which contains at least $8/9$ of the days. Thus, at most $1/9$ of the days will have more than 13 absentees.

---

2 Probability

2.1 Introduction

2.2 Probability and Inference

2.3 A Review of Set Notation

E2.2 Representing events. **(a)** $A \cap B$ **(b)** $A \cup B$ **(c)** $\overline{A} \cap \overline{B}$ **(d)** $(A \cap \overline{B}) \cup (\overline{A} \cap B)$

E2.5 Formal set theory proofs. The following proofs all begin with the distributive law.

(a) $(A \cap B) \cup (A \cap \overline{B}) = A \cap (B \cup \overline{B}) = A \cap S = A$. □

(b) $B \cup (A \cap \overline{B}) = (B \cup A) \cap (B \cup \overline{B}) = (B \cup A) \cap S = B \cup A = A$. The last step follows because $B \subset A$. □

(c) $(A \cap B) \cap (A \cap \overline{B}) = A \cap (B \cap \overline{B}) = A \cap \emptyset = \emptyset$. □

(d) $B \cap (A \cap \overline{B}) = (B \cap A) \cap (B \cap \overline{B}) = (B \cap A) \cap \emptyset = \emptyset$. □

E2.8 University housing survey. Represent each student by label s_i, $i = 1, 2, \ldots, 60$. Then the sample space $S = \{s_1, s_2, \ldots, s_{60}\}$. Let $L(\cdot)$ be a function to count elements of a set, e.g. $L(S) = 60$. We are given $L(\{\text{off camp}\}) = 9$, $L(\{\text{ugrad}\}) = 36$, and $L(\{\text{ugrad}\} \cap \{\text{off camp}\}) = 3$.

(a) For any two sets A and B,

$$L(A \cup B) = L(A) + L(B) - L(A \cap B).$$

Thus, $L(\{\text{ugrad}\} \cup \{\text{off camp}\}) = L(\{\text{ugrad}\}) + L(\{\text{off camp}\}) - L(\{\text{ugrad}\} \cap \{\text{off camp}\}) = 42$.

(b) For any two sets A and B,

$$L(A) = L(A \cap B) + L(A \cap \overline{B}).$$

Formally, this can be proven using the property in (a) in conjunction with E2.5(a, c). Thus, $L(\{\text{ugrad}\} \cap \{\text{on camp}\}) = L(\{\text{ugrad}\}) - L(\{\text{ugrad}\} \cap \{\text{off camp}\}) = 33$.

(c) Similarly to (b), $L(\{\text{grad}\} \cap \{\text{on camp}\}) = L(\{\text{on camp}\}) - L(\{\text{ugrad}\} \cap \{\text{on camp}\})$. Now, for any set A,

$$L(\overline{A}) = L(S) - L(A).$$

Thus $L(\{\text{grad}\} \cap \{\text{on camp}\}) = [L(S) - L(\{\text{off camp}\})] - L(\{\text{ugrad}\} \cap \{\text{on camp}\}) = 18$.

2.4 A Probabilistic Model for an Experiment: The Discrete Case

E2.18 A pair of coin tosses.

(a) $S = \{\text{HH}, \text{HT}, \text{TH}, \text{TT}\}$.

(b) Each sample point has equal probability $\frac{1}{4}$.

(c) $A = \{\text{HT}, \text{TH}\}$ and $B = \{\text{HH}, \text{HT}, \text{TH}\}$.

(d) Since each sample point is mutually exclusive, from Axiom 3 of Definition 2.6, we have $P(A) = P(\text{HT}) + P(\text{TH}) = \frac{2}{4}$. Similarly, $P(B) = \frac{3}{4}$. The sets $A \cap B = A$ and $A \cup B = B$ since $A \subset B$, so they have probabilities $P(A)$ and $P(B)$. Finally $\overline{A} \cup B = S$, thus $P(\overline{A} \cup B) = 1$.

***E2.20 The Monty Hall problem.**

(a) If each outcome is equally likely, $P(G) = P(D_1) = P(D_2) = \frac{1}{3}$.

(b) The player should switch. They will only lose if they initially chose G, with $P(G) = \frac{1}{3}$. In the event they initially chose either D_1 or D_2, they will win. This probability is is $P(D_1 \cup D_2) = \frac{2}{3}$.

***E2.21 A probability identity.** From E2.5(a), $A = (A \cap B) \cup (A \cap \overline{B})$ From E2.5(c), $(A \cap B)$ and $(A \cap \overline{B})$ are mutually exclusive. Thus, from Axiom 3 of Definition 2.6, $P(A) = P(A \cap B) + P(A \cap \overline{B})$. □

***E2.22 A probability identity 2.** Assume $B \subset A$. From E2.5(b), $A = B \cup (A \cap \overline{B})$. From E2.5(d), B and $(A \cap \overline{B})$ are mutually exclusive. Thus, from Axiom 3 of Definition 2.6, $P(A) = P(B) + P(A \cap \overline{B})$. □

E2.24 $\boldsymbol{P(B) \leq P(A)}$ **if** $\boldsymbol{B \subset A}$**.** It follows from E2.22 and Axiom 1 of Definition 2.6: $P(A \cap \overline{B}) \geq 0$. $\square$

2.5 Calculating the Probability of an Event: The Sample-Point Method

E2.30 Random wine rankings.

(a) The experiment consists rankings of 3 different wines, which we can represent as sample points like (w_1, w_2, w_3) meaning wine w_1 is best and wine w_3 is worst.

(b) $S = \{w_1w_2w_3, w_1w_3w_2, w_2w_1w_3, w_2w_3w_1, w_3w_1w_2, w_3w_2w_1\}$.

(c) Assume w_1 is best. Then $P = \frac{4}{6}$. This is true for any w_i being the best.

2.6 Tools for Counting Sample Points

Theorem 2.3: Example. Consider the four distinct objects $\{a, b, c, d\}$. First consider partitioning them into two equal sized groups, $k = 2$ and $n_1 = n_2 = 2$. The number of enumerations is given by the binomial coefficient $\binom{4}{2} = 4!/(2!2!) = 6$, each explicitly listed below.

$$\begin{array}{lll} ab|cd & bc|ad & cd|ab \\ ac|bd & bd|ac & \\ ad|bc & & \end{array}$$

Note that $ab|cd$ is the same as $ba|dc$ because in both enumerations, a and b (c and d) appear exactly once in the first (second) group. In other words, order within group does not matter. On the other hand, order of the groups themselves does matter, so $ab|cd$ is distinct from $cd|ab$.

Now consider $k = 3$ and $n_1 = 2$, $n_2 = n_3 = 1$. Theorem 2.3 gives $4!/(2!1!1!) = 12$ enumerations.

$$\begin{array}{lll} ab|c|d & bc|a|d & cd|a|b \\ ab|d|c & bc|d|a & cd|b|a \\ ac|b|d & bd|a|c & \\ ac|d|b & bd|c|a & \\ ad|b|c & & \\ ad|c|b & & \end{array}$$

Each enumeration in the $k = 2$ now has a counterpart because an additional group gives more freedom to arrange the last two letters. The proof relies in this overcounting in all $n!$ possible permutations and divides by the number of ways of overcounting in each group $n_i!$ for every group.

Example 2.13: Alternate solution. This is just a binomial distribution with success probability $p = \frac{1}{M}$, thus

$$\begin{aligned} P(k; n, M) &= \binom{n}{k} \left(\frac{1}{M}\right)^k \left(\frac{M-1}{M}\right)^{n-k} \\ &= \binom{n}{k} \left(\frac{1}{M}\right)^n (M-1)^{n-k}. \end{aligned}$$

E2.37 Random itinerary. **(a)** $P_6^6 = 720$ **(b)** 0.5, by symmetry. For every ordering where Denver is before San Francisco, there is a complementary ordering where Denver is after San Francisco.

E2.38 ***Fixe prix*** **menu possibilities.** $4 \times 3 \times 4 \times 5 = 240$ by Theorem 2.1.

E2.42 New hiree positions. $P_3^{10} = 720$. If the positions were identical, then it would be $C_3^{10} = 120$.

E2.43 Taxis dispatch. $\binom{9}{3\ 5\ 1} = 504$ by Theorem 2.3.

E2.44 Taxis dispatch 2.

(a) There are $N = \binom{9}{3\ 5\ 1}$ sample points. If the broken taxi goes to airport C, there are $n = \binom{8}{3\ 5\ 0}$ arrangements of the remaining 8 taxis fulfilling the dispatch requirements. Thus $p = n/N = \frac{56}{504}$.

(b) There are P_3^3 ways to arrange the three taxis at each of the airports. For each of those arrangements, there are $\binom{6}{2\ 4\ 0}$ ways to dispatch the remaining taxis. Thus $n = P_3^3 \times \binom{6}{2\ 4\ 0}$ and $p = n/N = \frac{90}{504}$.

E2.45 A polynomial expansion. $\binom{17}{2\ 5\ 10} = 408{,}408$

E2.46 Random team assignment. $\binom{10}{2}\binom{8}{2}\binom{6}{2}\binom{4}{2}\binom{2}{2} = 113{,}400$. The first game must choose 2 from 10 teams, ignoring order (e.g. team 1 vs. team 4 is the same as team 4 vs. team 1), with $\binom{10}{2}$ possibilities. The second game must choose 2 from the 8 remaining teams, ignoring order, with $\binom{8}{2}$ possibilities. The logic continues for the remaining games. The total number of ways is the product of all the ways to assign each game.

***E2.47 Random team assignment generalized.** $\binom{2n}{2}\binom{2n-2}{2}\cdots\binom{2}{2} = \frac{2n!}{2^n}$. Same logic as E2.56, simplified using $\binom{n}{k} = \frac{n!}{k!(n-k)!}$.

E2.48 A polynomial expansion 2. $\binom{8}{5} = \binom{8}{3} = 56$.

E2.49 University major possibilities.

(a) $\binom{130}{2} = 8{,}385$. Order does not matter, e.g. (STA, MS) double major is the same as (MS, STA).

(b) $26^2 + 26^3 = 18{,}252$.

(c) $130 + \binom{130}{2} = 8{,}515$.

(d) Yes, if you make no restrictions on the type of code: the result in (b) is larger than (c).

E2.50 Rigged lottery. 2^3 since each digit has 2 possibilities, as opposed to 10^3 in a fair lottery.

E2.51 Raffle probabilities. In all parts, the prizes can be awarded in $N = P_3^{50}$ ways.

(a) $n = \binom{3}{3}P_3^4$, so $p = n/N \approx 2.0 \times 10^{-4}$. There are $\binom{3}{3}$ ways for all three prizes to go to the organizers and P_3^4 ways they can be divided among them. Their product is n, by Theorem 2.1.

(b) $n = \binom{3}{2}P_2^4P_1^{46}$, with $p \approx 0.014$. There are $\binom{3}{2}$ ways for two of the prizes to go to the organizers, and one to the other contestants. There are P_2^4 ways to award the 2 prizes to the 4 organizers, and similarly P_1^{46} for the other contestants. Their product is n, by Theorem 2.1.

(c) $n = \binom{3}{1}P_1^4P_2^{46}$, with $p \approx 0.21$.

(d) $n = \binom{3}{0}P_3^{46}$, with $p \approx 0.77$.

Comments. (1) All the probabilities sum to one, as one of (a)–(d) must happen. (2) Since being awarded a prize is the only thing of interest, we can treat the three prizes as identical (order does not matter) and solely use combinations. For instance, the answer to (b) is also $\binom{4}{2}\binom{46}{1}/\binom{50}{3} \approx 0.014$.

E2.52 Experiment variables. $3 \times 3 \times 2 = 18$

E2.53 Firm contract assignments.

(a) $P_3^5 = 60$. Think of the problem as drawing 3 firms from 5, keeping track of order.

(b) $3P_2^4/P_3^5 = 36/60$. F_3 can be awarded 3 possible contracts. For each case, the two contracts can be assigned to the remaining firms in P_2^4 ways.

E2.54 Filling positions. Order does not matter, so $N = \binom{8}{4}$, $n = \binom{3}{2}\binom{5}{2}$, and $p = n/N = 30/70$.

Comment. Suppose the undergraduates were distinguishable and the question asked for the probability that only the first two undergraduates were chosen. Then there are only $\binom{2}{2}\binom{1}{0}$ ways to choose the two undergraduates, rather than $\binom{3}{2}$, and $p = 10/70$.

E2.55 Nurse sample. **(a)** Order does not matter, so $N = \binom{90}{10}$. **(b)** $n = \binom{20}{4}\binom{70}{6}$ with $p = n/N \approx 0.11$.

Comment. This is the hypergeometric distribution whose pmf

$$P(k; N, K, n) = \binom{K}{k}\binom{N-K}{n-k}\Big/\binom{N}{n},$$

with parameters $N = 90$, $K = 20$, $n = 10$, and $k = 4$

E2.56 Exam problem sample. Order does not matter, so $N = \binom{10}{5}$. The number of ways 5 of the problems were of the 6 she could solve is $n = \binom{6}{5}$, and $p = n/N = 6/252 \approx 0.024$.

E2.57 Card draw probabilities. There are $N = \binom{52}{2}$ possible 2-card draws, ignoring order. There are $\binom{4}{1}$ ways to draw an ace; for each of those ways, there are $\binom{12}{1}$ ways to draw a face card. Thus, $n = \binom{4}{1}\binom{12}{1}$ and $p = n/N \approx 0.036$.

E2.58 Card draw probabilities 2. For both parts, there are $N = \binom{52}{5}$ possible draws, ignoring order.

(a) There are $\binom{4}{3}$ ways to draw 3 aces; for each of those 3-ace draws, there are $\binom{4}{2}$ ways to draw 2 kings. Thus $n = \binom{4}{3}\binom{4}{2}$, and $p = n/N \approx 9.2 \times 10^{-6}$.

(b) First, count the number of ways to draw a triple. There are $\binom{13}{1}$ possible kinds; for each kind, there are $\binom{4}{3}$ ways. For each of those triples, we count the number of ways to draw a double. There are $\binom{12}{1}$ remaining kinds, each with $\binom{4}{2}$ ways. Thus $n = \binom{13}{1}\binom{4}{3}\binom{12}{1}\binom{4}{2}$ and $p \approx 1.4 \times 10^{-3}$.

E2.59 Card draw probabilities 3. For both parts, there are $N = \binom{52}{5}$ possible draws, ignoring order.

(a) $n = \binom{4}{1}^5$ and $p = n/N \approx 3.9 \times 10^{-4}$.

(b) All straights must range from $[\mathrm{A}, 5], [2, 6], \ldots, [10, \mathrm{A}]$ (under high rules). There are 10 such straights, so $n = 10 \times \binom{4}{1}^5$ and $p \approx 3.9 \times 10^{-3}$.

E2.60 Birthday problem. Assume $n \leq 365$. **(a)** There $N = 365^n$ arrangements of all possible birthdays. There are P_n^{365} arrangements for unique birthdays. Thus $p = P_n^{365}/N$. **(b)** $n = 23$

E2.61 Birthday problem 2. In contrast to E2.60, there are no restrictions on n. **(a)** There $N = 365^n$ arrangements of all possible birthdays. There are 364^n arrangements where no one shares my birthday. Thus $p = (364/365)^n$. **(b)** $n = 253$

E2.62 Manufacturer stock division. There are $N = \binom{9}{3\;3\;3}$ possible divisions. If both motors are assigned to the first line, there are $n = \binom{7}{1\;3\;3}$ ways the other motors can be divided. Thus, $p = n/N = 0.08\overline{3}$.

Comment. This can also be done using binomial coefficients only: $p = \binom{2}{2}\binom{7}{1}\binom{6}{3}\binom{3}{3}/\binom{9}{3}\binom{6}{3}\binom{3}{3}$.

E2.63 Potential sex bias. There are $N = \binom{8}{5}$ ways for an 8-member board to vote 5–3 and only $n = \binom{5}{5}\binom{3}{3} = 1$ way for a split vote along sex lines. Thus $p = n/N = 1/56 \approx 0.018$.

E2.64 Dice rolls. There are $N = 6^6$ total possible dice rolls and $n = P_6^6$ orderings of $1, 2, \ldots, 6$, thus $p = n/N \approx 0.015$.

E2.65 Dice rolls 2. There are $N = 6^5$ total possible dice rolls and $n = \binom{5}{2}P_4^4$ orderings of $1, 2, \ldots, 5$, thus $p = n/N \approx 0.031$. The $\binom{5}{2}$ factor, is because the 5 can be chosen in two ways; in each of those ways, there are P_4^4 orderings of the remaining numbers.

Comments. (1) It may be easier to solve this using probabilities. The probability 1 through 5 are observed is $\left(\frac{1}{6}\right)^4\left(\frac{2}{6}\right)$. There are 5! ways 1 through 5 can be observed, yielding $p = 5!\left(\frac{1}{6}\right)^4\left(\frac{2}{6}\right)$. (2) A harder problem is to compute the probability the recorded numbers are 1, 1, 2, 5, 5. The probability those number are observed is $\left(\frac{1}{6}\right)^3\left(\frac{2}{6}\right)^2$, and they can be ordered in $\binom{5}{2}\binom{3}{1}\binom{2}{2} = 30$ ways, so $p = 30 \times \binom{2}{2}\left(\frac{1}{6}\right)^3\left(\frac{2}{6}\right)^2 \approx 0.015$.

E2.66 Labor assignments. **(a)** $\binom{4}{1\;1\;1\;1}\binom{16}{5\;3\;4\;4}/\binom{20}{6\;4\;5\;5} \approx 0.12$ **(b)** The combinatorics simplifies because the question can be rephrased as binary: given 5 jobs and 20 workers, what is the probability no ethnic member is assigned to that job? The answer: $\binom{4}{0}\binom{16}{5}/\binom{20}{5} \approx 0.28$.

E2.67 Supply orders.

(a) $N = 10^7$ and $n_A = P_7^{10}$. Thus $p = n_A/N \approx 0.060$.

***(b)** $n_A = \binom{7}{2}\binom{5}{3}(10-2)^2$, so $p \approx 1.3 \times 10^{-3}$.

***(c)** $n_A = \binom{7}{2}\binom{5}{3}\binom{2}{1}(10-3)^1$, so $p \approx 2.9 \times 10^{-4}$.

***E2.70 Sum of multinomial coefficients.** Substitute $y_1 = y_2 = \cdots = y_k = 1$ into

$$(y_1 + y_2 + \cdots + y_k)^n = \sum_{n_1+n_2+\cdots+n_k=n} \binom{n}{n_1\ n_1\ \cdots\ n_k} y_1^{n_1} y_2^{n_2} \cdots y_k^{n_k}. \qquad \square$$

2.7 Conditional Probability and the Independence of Events

E2.71 Computing conditional probabilities.

(a) $P(A|B) = P(A \cap B)/P(B) = 1/3$.

(b) $P(B|A) = P(A \cap B)/P(A) = 1/5$.

(c) $P(A|A \cup B) = P(A \cap (A \cup B))/P(A \cup B)$. Since $A \subset (A \cup B)$, the numerator is $P(A)$. The denominator is $P(A \cup B) = P(A) + P(B) - P(A \cap B)$. Thus, $P(A|A \cup B) = 5/7$.

(d) $P(A|A \cap B) = P(A \cap (A \cap B))/P(A \cap B)$. Since $(A \cap B) \subset A$, the numerator is $P(A \cap B)$. Thus, $P(A|A \cap B) = 1$.

(e) $P(A \cap B|A \cup B) = P((A \cap B) \cap (A \cup B))/P(A \cup B)$. Since $(A \cap B) \subset (A \cup B)$, the numerator is $P(A \cap B)$. The denominator was computed in (c), thus $P(A \cap B|A \cup B) = 1/7$.

E2.72 Sex and pass rate. **(a)** $P(A) = 60/100$, $P(A|M) = 24/40$. Since $P(A) = P(A|M)$, A and M are independent. **(b)** $P(\overline{A}) = 40/100$, $P(\overline{A}|F) = 24/60$. $\overline{A}$ and F are also independent.

E2.73 Genetic allele probabilities. **(a)** $P(\geq \mathrm{R}) = 3/4$ **(b)** $P(\geq \mathrm{r}) = 3/4$ **(c)** $P(\mathrm{r}|\mathrm{red}) = 2/3$. Formally, this is equal to $P(\mathrm{r\ and\ red})/P(\mathrm{red}) = P(\mathrm{r})/P(\mathrm{red}) = \frac{2/4}{3/4} = 2/3$.

E2.75 Conditional card probabilities.

(a) $P(3\spadesuit|2\spadesuit) = P(3\spadesuit \cap 2\spadesuit)/P(2\spadesuit) = P(5\spadesuit)/P(2\spadesuit) = \dfrac{\binom{13}{5}/\binom{52}{5}}{\binom{13}{2}/\binom{52}{2}} \approx 8.4 \times 10^{-3}$. Alternatively, we could compute the probability of a 3-spade draw from a 50-card deck with 11 spades. Equivalently, $P(3\spadesuit|2\spadesuit) = \binom{11}{3}/\binom{50}{3}$.

(b) $P(2\spadesuit|3\spadesuit) = \binom{10}{2}/\binom{49}{2} \approx 0.038$. Same logic as (a).

(c) $P(1\spadesuit|4\spadesuit) = \binom{9}{1}/\binom{48}{1} = 9/48 \approx 0.19$. We could have skipped the combinatorics altogether by thinking about a single draw from a 48-card deck with 4 fewer spades.

E2.76 Plumbing satisfaction. **(a)** $P(\downarrow|A) = P(\downarrow \cap A)/P(A) = P(A|\downarrow)P(\downarrow)/P(A) = (0.5)(0.1)/0.4 = \frac{1}{8}$.
(b) $P(\downarrow|A) + P(\uparrow|A) = 1$, so $P(\uparrow|A) = \frac{7}{8}$.

Comment. Part (a) used Bayes' theorem.

2.8 Two Laws of Probability

E2.84 3-event probability. $P(\text{at least one } A_i) = P(A_1 \cup A_2 \cup A_3)$. The result follows from the observation under Theorem 2.6.

E2.85 Independence of complements. $P(A|\overline{B}) = P(A \cap \overline{B})/P(\overline{B}) = [P(A) - P(A \cap B)]/[1 - P(B)]$. By independence of A and B, the numerator is $P(A)[1 - P(B)]$. Thus $P(A|\overline{B}) = P(A)$; A and $\overline{B}$ are independent. $\square$

$P(\overline{A}|\overline{B}) = P(\overline{A} \cap \overline{B})/P(\overline{B}) = P(\overline{A \cup B})/[1 - P(B)] = [1 - P(A \cup B)]/[1 - P(B)]$. By independence of A and B, the numerator is $1 - [P(A) + P(B) - P(A)P(B)] = [1 - P(A)][1 - P(B)]$. Thus $P(\overline{A}|\overline{B}) = P(\overline{A})$; $\overline{A}$ and $\overline{B}$ are independent. $\square$

E2.86 Bounds for intersection probabilities.

(a, b) No. $P(A \cup B) = P(A) + P(B) - P(A \cap B) \leq 1$, so $P(A \cap B) \geq P(A) + P(B) - 1 = 0.5$.

(c, d) No. $P(A \cap B) = P(A)P(B|A) \leq P(A)$ since $P(B|A) \leq 1$ and similarly $P(A \cap B) \leq P(B)$. Thus $P(A \cap B) \leq \min(P(A), P(B)) = 0.7$.

E2.90 Skydiving injury probabilities. **(a)** Let A_i denote the event she is injured on the i-th jump. Then we want to compute $P(\text{at least one injury}) = P(A_1 \cup A_2) = P(A_1) + P(A_2) - P(A_1 \cap A_2)$. By independence, $P(A_1 \cup A_2) = \frac{1}{50} + \frac{1}{50} - \left(\frac{1}{50}\right)^2 = 0.0396$. For larger i, it is easier to compute the chance she is not injured on any jump $P(\text{no injuries}) = P(\overline{A}_1 \cap \overline{A}_2 \cap \cdots \cap \overline{A}_n) = P(\overline{A}_1)P(\overline{A}_2)\cdots P(\overline{A}_n)$ by independence, and taking the complement: $P(\text{at least one injury}) = 1 - P(\text{no injuries})$. For $i = 2$, we get $1 - \left(\frac{49}{50}\right)^2 = 0.0396$. **(b)** No. It is $1 - \left(\frac{49}{50}\right)^{50} \approx 0.64$.

E2.92 Lie detector probabilities. **(a)** Since tests are independent, $P(3 \text{ false positive}) = (0.05)^3 = \frac{1}{8{,}000}$. **(b)** $P(\geq 1 \text{ false positive}) = 1 - P(0 \text{ false positive}) = 1 - (1 - 0.05)^3 \approx 0.14$.

E2.93 Alternating hand hits. She wins if either HHH, HHM, or MHH, whose sum of probabilities are $(0.7)(0.4)(0.7) + (0.7)(0.4)(0.3) + (0.3)(0.4)(0.7) = 0.364$.

E2.95 Computing probabilities.
(a) $P(A \cup B) = P(A) + P(B) - P(A \cap B) \Rightarrow P(A \cap B) = 0.1$.
(b) $P(\overline{A} \cup \overline{B}) = P(\overline{A}) + P(\overline{B}) - P(\overline{A} \cap \overline{B}) = 0.8 + 0.7 - [1 - P(A \cup B)] = 0.9$.
(c) $P(\overline{A} \cap \overline{B}) = 1 - P(A \cup B) = 0.6$.
(d) $P(\overline{A}|B) = P(\overline{A} \cap B)/P(B) = [P(A \cup B) - P(A)]/P(B) = \frac{2}{3}$.

E2.97 Circuit relays.
(a) $P(\text{flow}) = 1 - P(\text{all relays fail}) = 1 - (0.1)^3 = 0.999$.
(b) $P(\text{relay pass}|\text{flow}) = P(\text{flow}|\text{relay pass})P(\text{relay pass})/P(\text{flow}) = (1)(0.9)/0.999$.

E2.98 Circuit relays 2. The series system flows if both relays pass: $P(\text{series}) = (0.9)^2 = 0.81$. The parallel system flows if at least one relay passes: $P(\text{parallel}) = 1 - (0.1)^2 = 0.99$.

***E2.100 Theorem 2.6 for conditional probabilities.** $P(A \cup B|C) = P((A \cup B) \cap C)/P(C)$. Applying the distributive law then Theorem 2.6, the numerator becomes

$$\begin{aligned} P((A \cap C) \cup (B \cap C) &= P(A \cap C) + P(B \cap C) - P((A \cap C) \cap (B \cap C)) \\ &= P(A \cap C) + P(B \cap C) - P(A \cap B \cap C). \end{aligned}$$

Dividing all by $P(C)$ yields the three desired conditional probabilities. □

E2.103 Rigged lottery 2. Assume independence of the two lotteries. **(a)** $1/10^3$ **(b)** $(1/2^3)(1/10^3)$

E2.104 Simplified Bonferroni inequality. $P(A \cup B) = P(A) + P(B) - P(A \cap B) \leq 1$. Rearrangement yields $P(A \cap B) \geq P(A) + P(B) - 1 = [1 - P(\overline{A})] + [1 - P(\overline{B})] - 1 = 1 - P(\overline{A}) - P(\overline{B})$. □

E2.105 Parachute jump bounds. Let J_i denote a safe jump event with $P(J_i) = 0.95$. From E2.104, $P(J_1 \cap J_2) \geq 1 - 0.05 - 0.05 = 0.9$.

Comment. The exact probability, assuming independence, is $(0.95)^2 = 0.9025$.

E2.108 Simplified Bonferroni inequality 2. Let $Z = A \cap B$. Then by applying the result of E2.104, $P(A \cap B \cap C) = P(Z \cap C) \geq 1 - P(\overline{Z}) - P(\overline{C})$. Now, $P(\overline{Z}) = P(\overline{A \cap B}) = P(\overline{A} \cup \overline{B})$ by DeMorgan's law. Applying Theorem 2.6, $P(\overline{A} \cup \overline{B}) = P(\overline{A}) + P(\overline{B}) - P(\overline{A} \cap \overline{B})$, and thus substitution yields $P(A \cap B \cap C) \geq 1 - P(\overline{A}) - P(\overline{B}) - P(\overline{C}) + P(\overline{A} \cap \overline{B}) \geq 1 - P(\overline{A}) - P(\overline{B}) - P(\overline{C})$. □

Comment. From the problem statement, a second application of the result of E2.104 yields $P(\overline{Z}) = 1 - P(A \cap B) \leq 1 - [1 - P(\overline{A}) - P(\overline{B})] = P(\overline{A}) + P(\overline{B})$. The same result follows.

2.9 Calculating the Probability of an Event: The Event-Composition Method

Example 2.22(b). $P(\text{exactly 1 match}) = P(A_1 \cap \overline{A}_2 \cap \overline{A}_3) + P(\overline{A}_1 \cap A_2 \cap \overline{A}_3) + P(\overline{A}_1 \cap \overline{A}_2 \cap A_3)$. By symmetry, $P(\text{exactly 1 match}) = 3P(A_1 \cap \overline{A}_2 \cap \overline{A}_3)$, and

$$P(A_1 \cap \overline{A}_2 \cap \overline{A}_3) = P(A_1)P(\overline{A}_2|A_1)P(\overline{A}_3|A_1 \cap \overline{A}_2) = \left(\tfrac{1}{3}\right)\left(\tfrac{1}{2}\right)(1) = \tfrac{1}{6}.$$

Thus, $P(\text{exactly 1 match}) = \frac{3}{6}$.

E2.110 Defective item probability. Let A denote the event an item is not defective. Then, since the lines are mutually exclusive, $P(A) = P(A|\text{I})P(\text{I}) + P(A|\text{II})P(\text{II}) = 0.908$.

E2.111 Product purchase probability. Let A and B denote the events a customer sees an ad and buys a product, respectively. Then $P(B) = P(B|A)P(A) + P(B|\overline{A})P(\overline{A})$. Now $A = M \cup T$, i.e. the union of magazine and television ads, so $P(A) = P(M) + P(T) - P(M \cap T) = 0.21$ and $P(B) = 0.149$.

E2.112 Radar detection. (a) $(0.02)^3 = 8 \times 10^{-6}$ **(b)** $(0.98)^3 \approx 0.94$

E2.113 Radar detection 2. $(0.98)^3(0.02) \approx 0.019$

E2.114 Lie detector accuracy. Let $G_\pm$ and $I_\pm$ denote positive/negative readings for the guilty and innocent persons. **(a)** $P(G_+ \cap I_+) = P(G_+)P(I_+) = 0.095$. **(b)** $P(G_+ \cap I_-) = P(G_+)P(I_-) = 0.855$. **(c)** $P(G_- \cap I_+) = P(G_-)P(I_+) = 0.005$. **(d)** $P(G_+ \cup I_+) = P(G_+) + P(I_+) - P(G_+ \cap I_+) = 0.0955$.

E2.117 Automobile inspection. (a) There are $\binom{4}{3}$ ways three of the four cars will be rejected, each with probability $\left(\frac{1}{2}\right)^3 \left(\frac{1}{2}\right)^1$, so $P = \binom{4}{3}\left(\frac{1}{2}\right)^4 = \frac{1}{4}$. **(b)** $\left(\frac{1}{2}\right)^4 = \frac{1}{16}$

Comment. This is the binomial distribution.

***E2.119 Sum of two dice.** Let S_i denote the sum for the i-th roll. From combinatorics, $P(S_i = 3) = 2/36$, $P(S_i = 4) = 3/36$, and $P(S_i = 7) = 6/36$.

(a) Let A denote the event a sum of 3 is obtained. Let B denote the event neither a 3 or 7 is obtained. From above, $P(A) = 2/36$ and $P(B) = 28/36$. In order to obtain a sum of 3 before 7, we we require sequences of rolls like A, BA, $BBA\ldots$. Thus

$$P(\text{3 before 7}) = \sum_{i=0}^{\infty} P(B)^i P(A) = \frac{2}{36}\sum_{i=0}^{\infty}\left(\frac{28}{36}\right)^i = \frac{1}{4},$$

where the sum was evaluated by the geometric series $\sum_{i=0}^{\infty} r^i = \frac{1}{1-r}$ for $|r| < 1$.

(b) The calculation is the same with $P(A) = 3/36$ and $P(B) = 27/36$. The result is $\frac{1}{3}$.

E2.120 Defective refrigerator inspection. Let D represent a defective refrigerator.

(a) $P = P(D\overline{DD}D\overline{DD}) + P(\overline{D}D\overline{D}D\overline{DD}) + P(\overline{DD}DD\overline{DD}) = \frac{1}{5}$. For example, using conditional probabilities $P(D\overline{DD}D\overline{DD}) = \frac{2}{6}\frac{4}{5}\frac{3}{4}\frac{1}{3}\frac{2}{2}\frac{1}{2} = \frac{1}{15}$.

(b) $P = P((\text{a})) + P(DD\overline{DDDD}) + P(D\overline{D}D\overline{DDD}) + P(\overline{D}DD\overline{DDD}) = \frac{2}{5}$.

(c) The problem can be rephrased as given 1 D and 3 $\overline{D}$, what is the probability we observe $D\overline{DDD}$ or $\overline{D}D\overline{DD}$? Each has probability $\frac{1}{4}$, so the result is $\frac{1}{2}$.

Comment. (b) can be answered using combinatorics. The number of ways to split 6 items into two groups of 4 and 2 is $N = \binom{6}{4\ 2}$. Assuming both defective units are in the group of 4, there are $n = \binom{4}{2\ 2}$ ways to split the others. Thus, $P = n/N = \frac{6}{15}$. This is just like Example 2.10.

E2.121 Guess the password. (a) $\frac{1}{n}$ **(b)** $P(\text{pass 2}) = P(\text{fail 1})P(\text{pass 2}|\text{fail 1}) = \left(1 - \frac{1}{n}\right)\left(\frac{1}{n-1}\right) = \frac{1}{n}$. Similar reasoning leads to $P(\text{pass 3}) = \frac{1}{n}$. **(c)** Let A_i denote the event of passing on the i-th try. Then $p = 1 - P(\overline{A}_1 \cap \overline{A}_2 \cap \overline{A}_3) = 1 - P(\overline{A}_1)P(\overline{A}_2|\overline{A}_1)P(\overline{A}_3|\overline{A}_1 \cap \overline{A}_2) = 1 - \left(1 - \frac{1}{n}\right)\left(1 - \frac{1}{n-1}\right)\left(1 - \frac{1}{n-2}\right)$. After simplification, $p = 1 - \frac{n-3}{n} = \frac{3}{n}$.

Comment. Alternate solution for (c): $p = P(A_1 \cup A_2 \cup A_3) = P(A_1) + P(A_2) + P(A_3) = \frac{3}{n}$ since A_i are mutually exclusive.

2.10 The Law of Total Probability and Bayes' Rule

E2.124 Conditional voter probability. $P(F) = P(F|D)P(D) + P(F|R)P(R) = 0.54$. By Bayes' rule, $P(D|F) = P(F|D)P(D)/P(F) = 0.\overline{7}$.

E2.125 Conditional disease probability. $P(+) = P(+|D)P(D) + P(+|\overline{D})P(\overline{D}) = 0.108$. By Bayes' rule, $P(D|+) = P(+|D)P(D)/P(+) = 0.08\overline{3}$. In words, approximately 8% of those who test positive actually have the disease. This is because the disease is rare and as such may warrant an unreliable label. However in contrast, a more undesirable outcome $P(D|-) \approx 0.0011$ is smaller.

E2.128 Proofs using Theorem 2.8.

(a) B and $\overline{B}$ form a partition of S. Thus, using Theorem 2.8, $P(A) = P(A|B)P(B)+P(A|\overline{B})P(\overline{B})$. Assuming $P(A|\overline{B}) = P(A|B)$, $P(A) = P(A|B)[P(B) + P(\overline{B})] = P(A|B)$, thus A and B are independent. $\square$

(b) $P(A) = P(A|C)P(C) + P(A|\overline{C})P(\overline{C})$ and $P(B) = P(B|C)P(C) + P(B|\overline{C})P(\overline{C})$. The result follows from the assumptions $P(A|C) > P(B|C)$ and $P(A|\overline{C}) > P(B|\overline{C})$. $\square$

E2.129 Male and female reactions. $P(M|-) = P(-|M)P(M)/P(-) = 0.4$.

E2.131 Symmetric difference. Since $A \triangle B = (A \cap \overline{B}) \cup (\overline{A} \cap B)$ and $(A \cap \overline{B}) \cap (\overline{A} \cap B) = \emptyset$, $P(A \triangle B) = P(A \cap \overline{B}) + P(\overline{A} \cap B)$. Since B and $\overline{B}$ partition S, $P(A) = P(A \cap B) + P(A \cap \overline{B})$, or $P(A \cap \overline{B}) = P(A) - P(A \cap B)$. Similarly, $P(\overline{A} \cap B) = P(B) - P(A \cap B)$. The result follows after substitution into $P(A \triangle B) = P(A \cap \overline{B}) + P(\overline{A} \cap B)$. $\square$

E2.132 Missing plane. **(a)** $P(-_1) = \sum_i P(-_1|i)P(i) = (\alpha_1 + 1 + 1)/3 = (\alpha_1 + 2)/3$. By Bayes' rule $P(1|-_1) = P(-_1|1)P(1)/P(-_1) = \alpha_1/(\alpha_1 + 2)$. **(b, c)** $1/(\alpha_1 + 2)$

E2.133 Multiple-choice model. $P(T|+) = P(+|T)P(T)/P(+) \approx 0.94$.

E2.137 Random draws.

(a) $P(2W) = \sum_{i=1}^{5} P(2W|i)P(i) = \frac{1}{5}\sum_{i=1}^{5} P(2W|i) = \frac{1}{5}\sum_{i=1}^{5} \binom{i}{2}/\binom{5}{2} = \frac{2}{5}$.

(b) $P(3|2W) = P(2W|3)P(3)/P(2W) = 0.15$.

***E2.138 Craps win probability.** Let W denote the event the player wins and n the number of rolls. Then $P(W) = P(W|n = 1)P(n = 1) + P(W|n > 1)P(n > 1) = (\frac{8}{12})(\frac{12}{36}) + P(W|n > 1)(\frac{24}{36})$. Define the event $Z = W|n > 1$ and let S_i denote the value of a point sum given $n > 1$, $S_i \in \{4, 5, 6, 8, 9, 10\}$. We can compute $P(Z) = \sum_i P(Z|S_i)P(S_i)$ using geometric series as in E2.119, noting that the probabilities $P(S_i)$ are based off 24, not 36, possible rolls due to the condition $n > 1$. The result is $P(Z) = 0.4\overline{06}$ and $P(W) = 0.4\overline{92}$.

2.11 Numerical Events and Random Variables

E2.142 Random spins. There are $N = 4^2$ possible spins. There are $n_3 = 4$ ways such that $y = 3$ and $n_2 = 4 \times 3$ ways such that $y = 2$. All other values of y are impossible, so $P(Y = i) = n_i/N$, $i = 2, 3$.

2.12 Random Sampling

2.13 Summary

Supplementary Exercises

E2.143 Theorem 2.7 for conditional probabilities. $P(A|B)+P(\overline{A}|B) = [P(A\cap B)+P(\overline{A}\cap B)]/P(B)$. Since A and $\overline{A}$ form a partition of S, the numerator is $P(B)$ and the result follows. $\square$

E2.145 Physical exam sequence. Order matters, so $N = P_{18}^{18} = 18!$ sequences.

E2.146 Card probabilities. Order does not matter, so there are $N = \binom{52}{5}$ possible hands. There are $\binom{4}{1}$ suits to choose from. For a given suit, there are $\binom{13}{5}$ hands. Thus $n = \binom{4}{1}\binom{13}{5}$ and $p = n/N \approx 0.0020$.

E2.147 Card probabilities 2. There are now 47 cards left in the deck. To get a full house, the gambler must be dealt two kings, KK, or one king and one ace, KA or AK. Thus we must compute

$p = P(\mathrm{KK} \cup \mathrm{KA} \cup AK) = P(\mathrm{KK}) + P(\mathrm{KA}) + P(\mathrm{AK})$ since the events are mutually exclusive. Using conditional probabilities, $P(\mathrm{KK}) = P(\mathrm{KA}) = P(\mathrm{AK}) = \frac{3}{47}\frac{2}{46}$ and thus $p \approx 0.0083$.

E2.148 Supplier components. There are $N = \binom{12}{4}$ ways four components can be selected. The event of interest can occur in three ways: $(n_A, n_B, n_C) \in \{(2,1,1),(1,2,1),(1,1,2)\}$. For the first tuple, there are $\binom{3}{2}$ ways supplier A has two components. For each of those ways, there are $\binom{4}{1}$ ways supplier B has one component. Similarly for each of those $\binom{3}{2}\binom{4}{1}$ ways, there are $\binom{5}{1}$ ways supplier C has the last component. Repeating for the other two tuples, $n = \binom{3}{2}\binom{4}{1}\binom{5}{1} + \binom{3}{1}\binom{4}{2}\binom{5}{1} + \binom{3}{1}\binom{4}{1}\binom{5}{2}$ and $p = n/N = 0.\overline{54}$.

E2.149 Disease probabilities. We are given $P(A) = 0.2$, $P(B) = 0.3$ and $P(A \cap B) = 0.1$. **(a)** $P(\overline{A \cup B}) = 1 - P(A \cup B) = 1 - [P(A) + P(B) - P(A \cap B)] = 0.6$. **(b)** $P(A \cup B) = 0.4$. **(c)** $P(A|B) = P(A \cap B)/P(B) = 0.\overline{3}$.

E2.150 Disease probabilities 2. $y \in \{0,1,2\}$. $P(Y=0) = P(\overline{A \cup B}) = 0.6$, $P(Y=1) = P(A \triangle B) = 0.3$, and $P(Y=2) = P(A \cap B) = 0.1$.

Comment. The probabilities are normalized: $\sum_y P(Y=y) = 1$.

***E2.151 A model for the World Series.** $(n_A, n_B) \in \{(4,1),(1,4)\}$ are the possible number of wins for each team. Consider $(n_A, n_B) = (4,1)$. This outcome requires team B to win one of the first four games (and team A to win the other three). This can occur in $\binom{4}{1}$ ways, each with probability $p^3(1-p)$. For each of these ways, team A must win the fifth game, thus $P((4,1)) = \binom{4}{1}p^4(1-p)$. Similarly, $P((1,4)) = \binom{4}{1}(1-p)^4 p$. All together, $P(5\text{ games}) = \binom{4}{1}p^4(1-p) + \binom{4}{1}(1-p)^4 p$.

Comment. We can solve for the probability the World Series lasts 4, 6, or 7 games similarly: $P(4\text{ games}) = \binom{3}{0}p^4 + \binom{3}{0}(1-p)^4$, $P(6\text{ games}) = \binom{5}{2}p^4(1-p)^2 + \binom{5}{2}(1-p)^4p^2$, and $P(7\text{ games}) = \binom{6}{3}p^4(1-p)^3 + \binom{6}{3}(1-p$ Indeed, one can verify $\sum_{i=4}^{7} P(i\text{ games}) = 1$.

E2.152 Lake water nitrates test. **(a)** Denote nitrates in a sample by N, and red color by R. N and $\overline{N}$ form a partition of the sample space, so $P(R) = P(R|N)P(N) + P(R|\overline{N})P(\overline{N}) = 0.355$. **(b)** $P(N|R) = P(R|N)P(N)/P(R) \approx 0.80$.

E2.153 Identical symptoms from different illnesses. $P(I_1|H) = P(H|I_1)P(I_1)/P(H)$ by Bayes' rule, where $P(H) = \sum_j P(H|I_j)P(I_j)$, assuming I_j form a partition, $j = 1,2,3$. Thus, $P(I_1|H) \approx 0.31$.

E2.154 Matching sock pairs.

(a) The person can choose socks in $N = P_4^{10}$ ordered ways. There are $n = 10 \times 8 \times 6 \times 4$ ordered ways there are no matching pairs. Thus $p = n/N \approx 0.38$.

***(b)** $p = [2n \times 2(n-1) \times \cdots \times 2r]/P_{2r}^{n}$.

Comment. (a) Alternatively, we could ignore order. There are $N = \binom{10}{4}$ combinations, with $n = \binom{5}{4}\binom{2}{1}^4$ having no matching pairs [$\binom{5}{4}$ ways to choose the unique pairs, each with 2 choices of socks]. (b) generalizes similarly.

E2.158 Modified bowl probability. Let W_i and B_i denote a white or black ball drawn on the i-th bowl. We wish to calculate $P(W_1|B_2)$. It is easier to compute $P(B_2|W_1) = b/(w+b+n)$ since the second bowl composition is known. Using Bayes' rule, $P(W_1|B_2) = P(B_2|W_1)P(W_1)/P(B_2)$, where

$$P(B_2) = P(B_2|W_1)P(W_1) + P(B_2|B_1)P(B_1) = \frac{b}{w+b+n}\frac{w}{w+b} + \frac{b+n}{w+b+n}\frac{b}{w+b}.$$

Thus

$$P(W_1|B_2) = \frac{bw}{bw + (b+n)b} = \frac{w}{w+b+n}.$$

E2.159 Prove $P(\emptyset) = 0$. $\overline{\emptyset} = S$, thus $P(\emptyset) = 1 - P(\overline{\emptyset}) = 1 - P(S) = 0$. $\square$

E2.160 Random run probabilities. **(a)** Ignoring order, $N = \binom{12}{10}$ and $n = 1$, thus $p = n/N = 1/66$. **(b)** $p = 2/66$, as the only other arrangement is the first two are defective.

E2.161 Random run probabilities 2. $P(R = 0) = P(R = 1) = 0$. From E2.160(b), $P(R = 2) = 2/66$. For $R = 3$, there must be two consecutive defectives not at the end points or one defective at each end point. There are $9 + 1$ such possibilities, so $P(R = 3) = 10/66$ and $P(R \leq 3) = 12/66$.

E2.162 Parking probabilities. There are $N = 9!$ possible ordered arrangements. Ignoring order, there are 7 arrangements of 3 adjacent sports cars. For each of those, there are 3! possible orderings, thus $n = 7 \times 3!$ and $p = n/N \approx 1.2 \times 10^{-4}$.

E2.163 Circuit relays. Let i denote the event a relay works. Then A works if $(1 \cup 2) \cap (3 \cup 4)$ with probability $P(\mathrm{A}) = P(1 \cup 2)P(3 \cup 4) = 0.9801$. Similarly, B works if $(1 \cap 3) \cup (2 \cap 4)$ with probability $P(\mathrm{B}) = P(1 \cap 3) + P(2 \cap 4) - P(1 \cap 2 \cap 3 \cap 4) = 0.9639$. Thus, $P(\mathrm{A}) > P(\mathrm{B})$.

E2.164 Circuit relays 2. $P(1 \cap 4|I) = P(I|1 \cap 4)P(1 \cap 4)/P(I) = (1)(0.9)^2/P(\mathrm{A}) \approx 0.83$.

E2.165 Circuit relays 3. $P(1 \cap 4|I) = P(I|1 \cap 4)P(1 \cap 4)/P(I) = P(2 \cup 3)(0.9)^2/P(\mathrm{B}) \approx 0.83$.

E2.166 Tire rankings. $N = \binom{8}{4}$ and $n = \binom{5}{3}$ [must select rank 3, and can only choose from remaining ranks 4–8], so $p = n/N = \frac{1}{7}$.

E2.167 Tire rankings 2. $y \in \{1, 2, 3, 4, 5\}$. $P(Y = 1) = \binom{7}{3}/N = \frac{35}{70}$, $P(Y = 2) = \binom{6}{3}/N = \frac{20}{70}$, $P(Y = 3) = \binom{5}{3}/N = \frac{10}{70}$, $P(Y = 4) = \binom{4}{3}/N = \frac{4}{70}$, $P(Y = 5) = \binom{3}{3}/N = \frac{1}{70}$.

Comment. The probabilities are normalized: $\sum_y P(Y = y) = 1$.

E2.168 Tire rankings 3.

(a) $n = \binom{3}{2}$ [must select ranks 3 and 7, and can only choose from remaining ranks 4–6], so $p = n/N = \frac{3}{70}$.

(b) There are 4 ranges like (a), so $P(R = 4) = \frac{12}{70}$.

(c) $r \in \{3, 4, 5, 6, 7\}$. $P(R = 3) = 5 \times \binom{2}{2}/N = \frac{5}{70}$, $P(R = 4) = 4 \times \binom{3}{2}/N = \frac{12}{70}$, $P(R = 5) = 3 \times \binom{4}{2}/N = \frac{18}{70}$, $P(R = 6) = 2 \times \binom{5}{2}/N = \frac{20}{70}$, $P(R = 7) = 1 \times \binom{6}{2}/N = \frac{15}{70}$.

***E2.169 Random beer rankings.**

(a) The sample space consists of all rankings of each of the four beers for all three drinkers. For instance, the sample point where all three drinkers rank beers $A > B > C > D$ can be represented as as $[(1, 2, 3, 4), (1, 2, 3, 4), (1, 2, 3, 4)]$. Each drinker can have 4! rankings, so $N = (4!)^3 = 13{,}824$.

(b) Let S_i denote the sum of ranks for beer $i \in \{A, B, C, D\}$. In order for $S_A \leq 4$, the three drinkers must have ranked A with (i) three ones or (ii) two ones and one two. Given option (i), there are $(3!)^3$ ways to rank the remaining beers. Given option (ii), there are $\binom{3}{1} \times (3!^3)$ ways, with the binomial prefactor accounting for the number of ways the rank 2 could be split among the drinkers. Finally, options (i) and (ii) are available for all four beers, so $n = \binom{4}{1} \times [(3!)^3 + \binom{3}{1} \times (3!)^3] = 3{,}456$ and $p = n/N = \frac{1}{4}$.

E2.170 Random name selection. $N = \binom{7}{3}$. Given the first person is selected, there are $n = \binom{6}{2}$ ways to select the remaining two. Thus $p = n/N = \frac{3}{7}$.

E2.172 Conditional identities? **(a)** False **(b)** False **(c)** True, see E2.143.

E2.173 Item inspections. Let D denote a defective item and C an item that was completely inspected. Use Bayes' rule: $P(D|C) = P(C|D)P(D)/P(C) = \frac{1}{4}$.

E2.174 No-pass, no-play. Let D_i denote a disqualification in the i-th period, $i = 2$ for current and $i = 1$ for previous. By Theorem 2.8, $P(D_2) = P(D_2|D_1)P(D_1) + P(D_2|\overline{D}_1)P(\overline{D}_1) = 0.255$.

E2.175 Mutual independence. Intuitively, no; if A and B happen, then C must happen. Formally, the sample space is $\{\mathrm{HH}, \mathrm{HT}, \mathrm{TH}, \mathrm{TT}\}$ and $P(A) = \frac{1}{2}$, $P(B) = \frac{1}{2}$, $P(C) = \frac{1}{2}$, $P(A \cap B) = \frac{1}{4}$,

$P(A \cap C) = \frac{1}{4}$, $P(B \cap C) = \frac{1}{4}$, $P(A \cap B \cap C) = \frac{1}{4}$. Indeed $P(A \cap B \cap C) \neq P(A)P(B)P(C)$.

E2.176 Mutual independence 2.

(a)
$$\begin{aligned} P((A \cup B) \cap C) &= P((A \cap C) \cup (B \cap C)) && \text{(DeMorgan's law)} \\ &= P(A \cap C) + P(B \cap C) - P(A \cap B \cap C) && \text{(Theorem 2.6)} \\ &= [P(A) + P(B) - P(A \cap B)]P(C) && \text{(Mutual independence)} \\ &= P(A \cup B)P(C). && \square \end{aligned}$$

(b) $P(A \cap (B \cap C)) = P(A \cap B \cap C) = P(A)P(B)P(C) = P(A)P(B \cap C)$. $\square$

E2.177 Skydiving injury probabilities 2. Let $p = \frac{1}{50}$ be the chance of injury. **(a)** $(1-p)^{50} \approx 0.36$ **(b)** $1 - (1-p)^{50} \approx 0.64$ **(c)** $n = 25$

***E2.178 Flu epidemic.** Let A (B) represent the event the (non)inoculated person has the flu. Assuming A and B are independent, $P(A \cup B) = P(A) + P(B) - P(A)P(B)$. Let E denote the event a person is exposed to the flu. Then $P(A) = P(A|E)P(E) + P(A|\overline{E})P(\overline{E})$. Since A is inoculated, $P(A|E) = 0.2$, and we can assume $P(A|\overline{E}) = 0$, i.e. if a person is not exposed to the flu, they will not have it. Thus $P(A) = 0.12$. Similarly, $P(B) = 0.54$ and thus $P(A \cup B) = 0.5952$.

***E2.179 Coin toss bets.**

(a) Assume there are no ties. Without loss of generality, we can assume that one player always bets heads and the other tails. There are $N = 2^6 = 64$ possible coin tosses and they break even if there are three heads, thus $n = \binom{6}{3} = 20$ and $p = n/N = 0.3125$.

(b) $N = 2^{10}$. Say Jones calls heads. The end result must have 8 heads and 2 tails. Furthermore, in order for Jones not to win early, at least one of the first six tosses must be tails. This can happen in two mutually exclusive ways: one or two tails in the first six. There are $\binom{6}{1}$ ways one tail can be tossed in the first six; for each of those ways, there are $\binom{2}{1}$ ways a tail can be tossed in the next two tosses (again, so Jones does not win early) and $\binom{2}{0} = 1$ way for the last two tosses to be heads. Similarly, there are $\binom{6}{2}\binom{4}{0}$ ways for two of the first six tosses to be tails. Thus, $n = \binom{6}{2}\binom{4}{0} + \binom{6}{1}\binom{2}{1}\binom{2}{0}$ and $p = n/N = 27/2^{10}$.

***E2.180 Bounded random walk.**

(a) The boundary can only be reached in $n = 4$ ways: 8 steps in the same direction. At each block, there are 4 possible steps, so $N = 4^8$ and $p = n/N = 1/4^7$.

(b) $N = 4^4$. To return to the center, the number of $\uparrow$ and $\downarrow$ steps must be equal, as well as $\leftarrow$ and $\rightarrow$. Let $(n_\uparrow, n_\downarrow, n_\rightarrow, n_\leftarrow)$ denote the a tuple of steps in each direction. The possible configurations are $(0,0,2,2)$, $(1,1,1,1)$, and $(2,2,0,0)$. For the first, there are $\binom{4}{2}$ to choose 2 $\rightarrow$ steps (and $\binom{2}{2}\binom{0}{0}\binom{0}{0} = 1$ way to choose the rest), and similarly for the last. For the second configuration, there are 4! ways (the first step can be in any direction; for each first step, there are three remaining directions; for each first two steps, there are two remaining directions, and so on). This yields $n = 2 \times \binom{4}{2} + 4!$ and $p = 36/256$.

***E2.181 Indistinguishable balls into distinguishable boxes.** The number of sample points is equivalent to finding the number of distinct arrangements of a row of n balls separated by $(N-1)$ lines, representing box boundaries. For instance with $N = n = 2$, there are 3 distinct arrangements.

$$|\bigcirc\bigcirc \qquad \bigcirc|\bigcirc \qquad \bigcirc\bigcirc|$$

There are $n + (N-1) = 3$ possible cells, and $\binom{3}{1}$ ways for one of them to be a line. In the general case, there are $n + (N-1)$ possible cells, and we must choose $(N-1)$ of them to be occupied by lines; thus, there are $\binom{n+N-1}{N-1}$ total arrangements. To complete the problem, we must also count the number of ways no box will be empty. In this case, lines cannot be at either boundary and cannot be in adjacent cells. For example, with $N = 3$ and $n = 4$, there are 3 arrangements.

○ | ○ | ○ ○ ○ | ○ ○ | ○ ○ ○ | ○ | ○

This is equivalent to saying lines must be placed between balls. Since there are $(n-1)$ spaces between balls, we can place the $(N-1)$ bars in $\binom{n-1}{N-1}$ ways. Thus, the probability no box will be empty is $\binom{n-1}{N-1}/\binom{n+N-1}{N-1}$. $\square$

---

3 Discrete Random Variables and Their Probability Distributions

3.1 Basic Definition

3.2 The Probability Distribution for a Discrete Random Variable

E3.1 Water impurities. $P(A \cap B) = P(A) + P(B) - P(A \cup B) = P(A) + P(B) - [1 - P(\overline{A \cup B})] = 0.1$. Possible values are $y \in \{0, 1, 2\}$, with associated probabilities $P(Y = 0) = P(\overline{A \cup B}) = 0.2$, $P(Y = 1) = P(A \triangle B) = 0.7$, and $P(Y = 1) = P(A \cap B) = 0.1$.

E3.2 Coin toss game. $y \in \{-1, 1, 2\}$ with $P(Y = -1) = -\frac{1}{2}$, $P(Y = 1) = \frac{1}{4}$, and $P(Y = 2) = \frac{1}{4}$.

E3.3 Inspector tests. $y \in \{2, 3, 4\}$. There are $N = \binom{4}{2}$ ways to choose 2 defectives. There is $n_2 = \binom{2}{2} = 1$ way to choose the 2nd defective on try 2, $n_3 = \binom{2}{1} = 2$ ways to choose the 2nd defective on try 3, and $n_4 = \binom{3}{1} = 3$ ways to choose the 2nd defective on try 4. Thus $P(Y = y) = (y - 1)/6$.

E3.4 Water paths. $y \in \{0, 1, 2\}$. Let i be the event valve i works. $P(Y = 0) = P(\overline{1} \cap (\overline{2} \cup \overline{3})) = 0.072$, $P(Y = 1) = P(1 \triangle (2 \cap 3)) = 0.416$, and $P(Y = 2) = P(1 \cap (2 \cap 3)) = 0.512$.

***E3.5 A matching problem.** $y \in \{0, 1, 3\}$. Note that $y \neq 2$ since if 2 are matched correctly, the third must also be. Let i denote the event the i-th picture is matched correctly, $i = 1, 2, 3$. Then

$$\begin{aligned} P(Y = 0) &= P(\overline{1 \cup 2 \cup 3}) \\ &= (3! - |1 \cup 2 \cup 3)/3! \\ &= 1 - (|1| + |2| + |3| - |1 \cap 2| - |1 \cap 3| - |2 \cap 3| + |1 \cap 2 \cap 3|)/3!. \end{aligned}$$

If i is matched, then there are 2! ways to match the remaining pictures, so $|i| = 2$. If i and j are matched, $i \neq j$, there is only one remaining way for the remaining picture to be matched: $|i \cap j| = 1$. Similarly, $|1 \cap 2 \cap 3| = 1/3$. Substitution yields $P(Y = 0) = \frac{2}{6}$.

To compute $P(Y = 1)$, there must be exactly one match. There are $\binom{3}{1}$ ways to choose which picture is matched. For each of those ways, there must be no matches in the remaining two, so there are $\binom{3}{1}|\overline{1' \cup 2'}| = \binom{3}{1}[2! - (|1'| + |2'| - |1' \cap 2'|)] = 3$ ways one picture can be matched, where primes denote the subset of the remaining 2 matches. Thus $P(Y = 1) = \frac{3}{6}$.

The logic is similar for general $P(Y = y)$:

$$P(Y = 2) = \frac{1}{3!}\binom{3}{2}|\overline{1''}| = 0, \qquad P(Y = 3) = \frac{1}{3!}\binom{3}{3} = \frac{1}{6}.$$

Comment. This problem could easily have been solved by listing all 6 outcomes, but the approach used generalizes to the problem of n matches, hence the added asterisk. In general, using the same logic, $P(Y = 0) = 1 - |1 \cup 2 \cup \cdots \cup n|/n!$, where the formula for the cardinality of the unions is given by the inclusion-exclusion principle. The remaining probabilities are

$$P(Y = y) = \frac{1}{n!}\binom{n}{y}[(n - y)! - |1 \cup 2 \cup \cdots \cup (n - y)|]$$

where the cardinality of the unions is computed for the subset with $(n - y)$ matches.

E3.7 Distribution of empty bowls. $y \in \{0, 1, 2\}$ and $N = 3^3$. $P(Y = 0) = 3!/N = \frac{6}{27}$ and $P(Y = 2) = \binom{3}{2}1^3/N = \frac{3}{27}$. $P(Y = 1)$ is a bit trickier to calculate directly. There are $\binom{3}{1}$ ways to choose an empty bowl. For each of those ways, the remaining bowls must have $(2, 1)$ or $(1, 2)$ balls, which can occur in $\binom{3}{2}\binom{1}{1} + \binom{3}{1}\binom{2}{2}$ ways. Thus $P(Y = 1) = \binom{3}{1}[\binom{3}{2}\binom{1}{1} + \binom{3}{1}\binom{2}{2}]/N = \frac{18}{27}$.

3.3 The Expected Value of a Random Variable or a Function of a Random Variable

E3.13 Coin toss game 2. $\mathrm{E}(Y) = \sum_y yp(y) = \frac{1}{4}$. $\mathrm{var}(Y) = \mathrm{E}\,Y^2 - (\mathrm{E}\,Y)^2 = \frac{27}{16}$. You should pay $\mathrm{E}(Y) = \frac{1}{4}$. Your net winnings in one game is $N = c + Y$, where c is the amount you pay such that $\mathrm{E}(N) = c + \mathrm{E}(Y) = 0$.

E3.16 Guess the password 2. For each $y \in \{1, 2, \ldots, n\}$, $P(Y = y) = 1/n$. Thus

$$\mathrm{E}(Y) = \frac{1}{n}\sum_{y=1}^{n} y = \frac{1}{n}\left[(n+1)\frac{n}{2}\right] = \frac{n+1}{2}.$$

$$\mathrm{var}(Y) = \mathrm{E}\,Y^2 - (\mathrm{E}\,Y)^2 = \frac{1}{n}\sum_{y=1}^{n} y^2 - (\mathrm{E}\,Y)^2 = \frac{1}{n}\left[\frac{1}{6}n(n+1)(2n+1)\right] - (\mathrm{E}\,Y)^2 = \frac{(n+1)(n-1)}{12}.$$

***E3.26 Conditioned sales distribution.** The number of possible sales are $x \in \{0, 1, 2\}$, related to the profit by $Y = 50000X$. Let i denote the event i customers are contacted, $i = 1, 2$, and S_j for j successful sales. Then

$$P(X = 0) = P(S_0|1)P(1) + P(S_0|2)P(2) = \binom{1}{0}0.9 \times \frac{1}{3} + \binom{2}{0}0.9^2 \times \frac{2}{3} = 0.84$$

since $P(S_j|i)$ has a binomial distribution of choosing j successful sales from i customers (with success probability $p = 0.1$). Similarly, $P(X = 1) = 0.15\overline{3}$ and $P(X = 2) = 0.00\overline{6}$. Thus $\mathrm{E}(X) = 0.1\overline{6}$ and by Theorem 3.4, $\mathrm{E}(Y) = 50000\,\mathrm{E}(X) = 8333.\overline{3}$. A straightforward calculation yields $\sigma = \sqrt{\mathrm{var}(Y)} = 50000\sqrt{\mathrm{var}(X)} \approx 20000$.

E3.28 Inspector tests. If $Y = y$ occurs, then y devices must be tested and two must be repaired. Thus the total cost is $C = Y \times \$2 + 2 \times \$4$, with $\mathrm{E}(C) = \$2\,\mathrm{E}(Y) + \$8 = \$8 + \$2\sum_{y=2}^{4} y\frac{y-1}{6} = \$14.\overline{6}$.

E3.33 Mean and variance of a linear transformation.

(a)

$$\begin{aligned} \mathrm{E}(aY + b) &= \mathrm{E}(aY) + \mathrm{E}(b) && \text{(Theorem 3.5)} \\ &= a\,\mathrm{E}(Y) + b. \quad \square && \text{(Theorem 3.3 and 3.4)} \end{aligned}$$

(b)

$$\begin{aligned} \mathrm{var}(aY + b) &= \mathrm{E}[(aY + b)^2] - [\mathrm{E}(aY + b)]^2 && \text{(Theorem 3.6)} \\ &= \mathrm{E}[(aY)^2] + \mathrm{E}[2abY] + \mathrm{E}(b^2) - [(a\,\mathrm{E}(Y) + b)^2] && \text{(Part (a))} \\ &= a^2\,\mathrm{E}[Y^2] + 2ab\,\mathrm{E}[Y] + b^2 - [(a\,\mathrm{E}(Y))^2 + 2ab\,\mathrm{E}(Y) + b^2] && \text{(Theorem 3.3 and 3.4)} \\ &= a^2[\mathrm{E}(Y^2) - (\mathrm{E}\,Y)^2] = a^2\,\mathrm{var}(Y). && \square \end{aligned}$$

3.4 The Binomial Probability Distribution

E3.35 Voting sample probabilities. Let A denote the event the first trial was successful. Then $P(B) = P(B|A)P(A) + P(B|\overline{A})P(\overline{A}) = \frac{0.4N-1}{N-1} \times 0.4 + \frac{0.4N}{N-1} \times 0.6 = 0.4$. The conditional probability $P(B|A) \approx 0.3999$ barely differs from the unconditional probability $P(B)$ since N is large.

E3.51 Dice probabilities.
(a) $Y \sim \mathrm{binomial}(n = 4, p = \frac{1}{6})$, so $1 - p(0) \approx 0.518$.
(b) $Y \sim \mathrm{binomial}(n = 24, p = \frac{1}{36})$, so $1 - p(0) \approx 0.491$.

E3.54 Symmetry of the binomial distribution. (a, b) The result follows from substitution of the transformation $Y^* = n - Y$. **(a)** Relabel success with failure, and vice versa.

***E3.64 Binomial maximum likelihood estimate.** We seek to maximize

$$P(Y = y_0) = \binom{n}{y_0} p^{y_0}(1-p)^{n-y_0}$$

with respect to p, which is equivalent to maximizing $f(p) = \log\left[p^{y_0}(1-p)^{n-y_0}\right]$. Simplifying to $f(p) = y_0 \log(p) + (n-y_0)\log(1-p)$ and solving $df/dp = (np - y_0)/[p(p-1)] = 0$ yields $p = y_0/n$. Inspection of the second derivative evaluated at $p^* = y_0/n$ yields $f''(p^*) = n^3/[y_0(y_0-n)] < 0$ for $y_0 < n$, verifying it is a maximum. □

Comment. When $y_0 = n$, inspection of $f(p) = y_0 \log(p)$ shows $p^* = 1$ is still a maximum.

***E3.65 Binomial maximum likelihood estimate 2.**

(a) $\mathrm{E}(Y/n) = \mathrm{E}(Y)/n = (np)/n = p$.

(b) $\mathrm{var}(Y/n) = \mathrm{var}(Y)/n^2 = [np(1-p)]/n^2 = p(1-p)/n$, which tends to zero as $n \to \infty$. The MLE has less variance as n increases.

3.5 The Geometric Probability Distribution

E3.66 Geometric distribution properties.

(a) $\sum_y p(y) = \sum_{y=1}^{\infty}(1-p)^{y-1}p = p\sum_{y=1}^{\infty}(1-p)^{y-1}$. Let $z = y-1$. Summing the resultant geometric series yields $\sum_y p(y) = p\sum_{z=0}^{\infty}(1-p)^z = p\left[\frac{1}{1-(1-p)}\right] = 1$. □

(b) Direct substitution yields the relation. $P(Y=1)$ is most likely.

E3.71 Memoryless property of a geometric random variable.

(a) $P(Y > a) = 1 - \sum_{y=1}^{a} p(y) = 1 - p\sum_{z=0}^{a-1}(1-p)^z$ where $z = y-1$. Summing the geometric series, $P(Y > a) = 1 - p\left[\frac{1-(1-p)^a}{1-(1-p)}\right] = (1-p)^a$. □

(b) $P(Y > a+b | Y > a) = P((Y > a+b) \cap (Y > a))/P(Y > a) = P(Y > a+b)/P(Y > a)$. Each probability can be computed like in (a): $P(Y > a+b|Y > a) = (1-p)^{a+b}/(1-p)^a = (1-p)^b$. The probability does not depend on the previous event $Y > a$. □

(c) The (identical) trials are assumed to be independent.

E3.72 Memoryless property application. If the coin truly is balanced, then the previous ten tosses do not matter, so we can simply compute $P(Y \geq 2) = 1 - P(Y=1) = \frac{1}{2}$.

E3.77 Geometric distribution for odd integers.
$P(Y = \text{odd}) = p(1) + p(3) + p(5) + \cdots = p + q^2p + q^4p + \cdots = p(1 + q^2 + q^4 + \cdots)$. This is a geometric series with $r = q^2$, and thus sums to $p/(1-q^2)$.

E3.79 Mixed survey response. Let T denote those who will answer truthfully and M those who have tried marijuana. The probability a randomly selected person will answer in the affirmative is $P(A) = P(A|M)P(M) + P(A|\overline{M})P(\overline{M}) = P(A|M)P(M)$ since $P(A|\overline{M}) = 0$. Further partitioning, $P(A|M) \equiv P(Z) = P(Z|T)P(T) + P(Z|\overline{T})P(\overline{T}) = P(Z|T)P(T)$, we conclude $P(Y=y)$ is geometric with affirmative probability $P(A) = P(A|M \cap T)P(M)P(T) = 1 \times 0.2 \times 0.3 = 0.06$.

E3.80 Alternating dice rolls. Let X be the event B obtained the first 6 on her second toss and Y be the event B threw the first 6. Then $P(X|Y) = P(Y|X)P(X)/P(Y) = P(X)/P(Y)$. From E3.77, $P(Y) = 1 - p/(1-q^2) = 0.\overline{45}$. Assuming A and B toss identically and independently, $P(X) = \left(\frac{5}{6}\right)^3 \frac{1}{6}$ and thus $P(X|Y) \approx 0.212$.

***E3.85 Geometric variance.** First note that $\mathrm{var}(Y) = \mathrm{E}(Y^2) - (\mathrm{E}\,Y)^2$ and $\mathrm{E}(Y^2) = \mathrm{E}[(Y(Y-1))] + \mathrm{E}(Y)$. Thus we must compute

$$\mathrm{E}[(Y(Y-1))] = \sum_{y=1}^{\infty} y(y-1)q^{y-1}p = pq\sum_{y=1}^{\infty}\frac{d^2}{dq^2}q^y = pq\frac{d^2}{dq^2}\sum_{y=1}^{\infty}q^y = pq\frac{d^2}{dq^2}\sum_{y=2}^{\infty}q^y.$$

The sum is a geometric series $\sum_{y=2}^{\infty} q^y = \frac{1}{1-q} - q - 1$. After differentiating,

$$\mathrm{E}[(Y(Y-1))] = pq\left[\frac{2}{(1-q)^3}\right] = \frac{2(1-p)}{p^2}.$$

Thus $\mathrm{E}(Y^2) = 2(1-p)/p^2 + 1/p$ and $\mathrm{var}(Y) = (1-p)/p^2$.

***E3.86 Geometric MLE.** We seek to maximize $P(Y = y_0) = (1-p)^{y_0-1}p$ with respect to p, which is equivalent to maximizing $f(p) = \log[(1-p)^{y_0-1}p] = (y_0-1)\log(1-p) + \log(p)$. Solving $df/dp = -(y_0-1)/(1-p) + 1/p = 0$ yields $p = 1/y_0$. Inspection of the second derivative evaluated at $p^* = 1/y_0$ yields $f''(p^*) = -y_0^3/(y_0-1) < 0$ for $y_0 > 1$, verifying it is a maximum. □

Comment. When $y_0 = 1$, inspection of $f(p) = \log(p)$ shows $p^* = 1$ is still a maximum.

***E3.87 Geometric MLE 2.**

$$\mathrm{E}(1/Y) = \sum_{y=1}^{\infty} \frac{1}{y} q^{y-1} p = \frac{p}{q} \sum_{y=1}^{\infty} \frac{1}{y} q^y = \frac{p}{q}[-\log(1-q)] = -\frac{p}{1-p}\log(p).$$

***E3.88 Geometric RV counting failures.** $P(Y^* = y) = P(Y = y+1) = q^{(y+1)-1}p = q^y p$. □

***E3.89 Geometric RV counting failures 2.**

(a) $\mathrm{E}(Y^*) = \mathrm{E}(Y-1) = \mathrm{E}(Y) - 1 = 1/p - 1$. $\mathrm{var}(Y^*) = \mathrm{var}(Y-1) = \mathrm{var}(Y) = (1-p)/p^2$.

***(b)** The proof is nearly identical to the Theorem 3.8 proof and E3.85.

3.6 The Negative Binomial Probability Distribution

***E3.99 Trials before the r-th success.** Exactly y trials can occur before the r-th success if $y > r-1$. For each y, $r-1$ successes and $y-(r-1)$ failures must occur, with probability $f(r-1|y,p)$ where $f(\cdot)$ is the pmf of the binomial distribution. Furthermore, the $(y+1)$-th trial must be the r-th success, so

$$P(Y = y) = f(r-1|y,p) \times p = \binom{y}{r-1} p^r (1-p)^{y-r+1} \qquad \text{for integer } y > r-1.$$

Comment. Let $g(\cdot)$ denote the pmf of the negative binomial distribution. The question is equivalent to finding $g(y+1|r,p) = \binom{y}{r-1}p^r(1-p)^{1-r}$, $y+1 = r, r+1, r+2, \ldots$.

***E3.101 Negative binomial MLE. (b)** We seek to maximize $P(Y = y_0) = \binom{y_0-1}{r-1}p^r(1-p)^{y_0-r}$ with respect to p, which is equivalent to maximizing $f(p) = \log[p^r(1-p)^{y_0-r}] = r\log p + (y_0-r)\log(1-p)$. Solving $df/dp = r/p - (y_0-r)/(1-p) = 0$ yields $p = r/y_0$. The second derivative evaluated at $p^* = r/y_0$ yields $f''(p^*) = -y_0^3/[r(y_0-r)] < 0$ for $y_0 > r$, verifying it is a maximum. □

Comment. When $y_0 = r$, inspection of $f(p) = r\log(p)$ shows $p^* = 1$ is still a maximum.

3.7 The Hypergeometric Probability Distribution

E3.104 Cocaine test. Let A denote the event the first 4 all contain contain and B the remaining 2 do not. Then $P(A \cap B) = P(A)P(B|A)$, where each can be computed from the hypergeometric distribution: $P(A) = \binom{15}{4}\binom{5}{0}/\binom{20}{4}$ and $P(B) = \binom{11}{0}\binom{5}{2}/\binom{16}{2}$. Thus $P(A \cap B) \approx 0.0235$.

E3.118 Card probabilities. $P(4\mathrm{A}| \geq 3\mathrm{A}) = P(4\mathrm{A} \cap \geq 3\mathrm{A})/P(\geq 3\mathrm{A}) = P(4\mathrm{A})/[P(3\mathrm{A}) + P(4\mathrm{A})$. We can use the hypergeometric distribution distinguishing ace and not ace: $P(4\mathrm{A}) = \binom{4}{4}\binom{48}{1}/\binom{52}{5}$ and $P(3\mathrm{A}) = \binom{4}{3}\binom{48}{2}/\binom{52}{5}$, thus $P(4\mathrm{A}|3\mathrm{A}) \approx 0.0105$.

E3.119 Card probabilities 2. Let A denote the event the 2nd king is dealt on the 5th card, B the event one king is drawn in the first four cards, and C the event a king is drawn on the 5th card. Then $P(A) = P(B \cap C) = P(B)P(C|B) = P(B) \times \frac{3}{48}$. Using the hypergeometric distribution, $P(B) = \binom{4}{1}\binom{48}{3}/\binom{52}{4}$, thus $P(A) \approx 0.0160$.

***E3.120 Estimating animal population sizes.** The population consists of k tagged and $N-k$ untagged animals. Choosing a sample of $n = 3$ can be seen as draws without replacement from the population of size N. Thus the number of tagged animals in the sample $Y \sim$ hypergeometric$(r = k, n, N)$ and $P(Y = y) = \binom{k}{y}\binom{N-k}{n-y}/\binom{N}{n}$.

Maximizing with respect to N is equivalent to maximizing $f(N) = \binom{N-k}{n-y}/\binom{N}{n}$ which, in terms of factorials,

$$f(N) = \frac{(N-k)!}{(n-y)!(N-k-n+y)!}\frac{n!(N-n)!}{N!}.$$

Removing factors with no N dependence, it is sufficient to maximize

$$g(N) = \frac{(N-k)!}{(N-k-n+y)!}\frac{(N-n)!}{N!}.$$

Note that N is discrete, so use of the derivative test as in E3.64, E3.86, and E3.101 is unwarranted. Instead, we can consider the ratio $x = g(N)/g(N-1)$:

$$x(N) = \frac{(N-k)}{(N-k-n+y)}\frac{(N-n)}{N}.$$

$x < 1$ ($x > 1$) implies N is decreasing (increasing). Consider first when $x < 1$; this occurs if

$$(N-k)(N-n) < N(N-k-n+y).$$

Expanding both sides and canceling terms yields $kn < Ny$ or $N > kn/y$. Similarly, $x > 1$ when $N < kn/y$. We conclude then that $g(N)$ only has one local maximum at $N = kn/y$. Since N is discrete, the first integer estimate before N starts to decrease is the floor $\hat{N} = \lfloor kn/y \rfloor$.

For the particular parameters given, we estimate $N = 12$.

3.8 The Poisson Probability Distribution

E3.123 Manufacturing flaws. $p(y) = \lambda^y e^{-\lambda}/y!$. We are given $p(0) + p(1) = e^{-\lambda}(1+\lambda) = 0.04$, hence $\lambda \approx 5$ and $P(Y > 5) = 1 - \sum_{i=0}^{5} p(i) \approx 0.384$.

E3.128 Car arrivals during a call. For the 1-minute call, $\lambda = 80/60$, so $P = 1 - p(0) \approx 0.736$.

***E3.129 Car arrivals during a call 2.** The probability no cars arrive is $P(0) = e^{-\lambda(t)}$ for some unspecified $\lambda(t) = 80t/60$, where t is the duration of the call in minutes. $P(0) \geq 0.4$ requires $\lambda(t) \leq -\log(0.4)$, or $t \leq -60\log(0.4)/80 \approx 0.687$ minutes.

E3.138 Poisson variance.

$$\mathrm{E}[Y(Y-1)] = \sum_{y=0}^{\infty} y(y-1)\frac{\lambda^y e^{-\lambda}}{y!} = e^{-\lambda}\sum_{y=2}^{\infty} y(y-1)\frac{\lambda^y}{y!} = \lambda^2 e^{-\lambda}\sum_{y=2}^{\infty}\frac{\lambda^{y-2}}{(y-2)!} = \lambda^2 \sum_{z=0}^{\infty}\frac{e^{-\lambda}\lambda^z}{z!}$$

where $z = y - 2$. The sum is $\sum_z p(z) = 1$ where $p(\cdot)$ is the pmf for the Poisson distribution. Thus $\mathrm{E}(Y^2) = \mathrm{E}[Y(Y-1)] + E(Y) = \lambda^2 + \lambda$ and $\mathrm{var}(Y) = \mathrm{E}(Y^2) - (\mathrm{E}\,Y)^2 = \lambda$.

***E3.140 Sales promotion.** Let Y be the number of customers who purchase the item, $Y \sim \text{Poisson}(\lambda = 2)$. The expected cost $C = 100\left(\frac{1}{2}\right)^Y$. It remains to compute $\mathrm{E}(C) = 100\,\mathrm{E}\left(\frac{1}{2}\right)^Y$. From Theorem 3.2,

$$\mathrm{E}\left(\tfrac{1}{2}\right)^Y = \sum_{y=0}^{\infty}\left(\tfrac{1}{2}\right)^y \frac{\lambda^y e^{-\lambda}}{y!} = e^{-\lambda/2}\sum_{y=0}^{\infty}\frac{(\lambda/2)^y e^{-\lambda/2}}{y!} = e^{-\lambda/2},$$

since the sum is the sum of a Poisson pmf with $\lambda^* = \lambda/2$. Thus $\mathrm{E}(C) = 100e^{-1} \approx 36.8$.

***E3.142 Poisson ratio of likelihoods.**

(a) The result follows after direct substitution of $p(y) = \lambda^y e^{-\lambda}/y!$.

(b) $p(y) > p(y-1)$ when $\lambda > y$.

(c) Similarly, $p(y) < p(y-1)$ when $\lambda < y$. Thus, there is one local maximum at $y = \lambda$. Since y is discrete, the first integer estimate before $p(y)$ starts to decrease is $\hat{y} = \lfloor \lambda \rfloor$.

3.9 Moments and Moment-Generating Functions

E3.145 Binomial mgf.

$$m(t) = \mathrm{E}(e^{tY}) = \sum_{y=0}^{n} e^{ty}\binom{n}{y}p^y(1-p)^{n-y} = \sum_{y=0}^{n}\binom{n}{y}(pe^t)^y(1-p)^{n-y}.$$

We recognize the above as the binomial expansion $(a+b)^n = \sum_{i=0}^{n}\binom{n}{i}a^i b^{n-i}$, with $a = pe^t$ and $b = (1-p)$. Thus $m(t) = [pe^t + (1-p)]^n$.

E3.146 Binomial mgf 2. $m'(t) = n(pe^t+q)^{n-1}pe^t$ and $m''(t) = np(pe^2+q)^{n-2}(npe^t+q)e^t$. Evaluated at $t=0$, $\mathrm{E}(Y) = m'(t=0) = n(p+q)^{n-1}p = np$ and $\mathrm{E}(Y^2) = m''(t=0) = np(np+q)$. The variance $\mathrm{var}(Y) = \mathrm{E}(Y^2) - (\mathrm{E}\,Y)^2 = npq$.

E3.147 Geometric mgf.

$$m(t) = \mathrm{E}(e^{tY}) = \sum_{y=1}^{\infty} e^{ty}(1-p)^{y-1}p = \sum_{y=1}^{\infty}[e^t(1-p)]^y(1-p)^{-1}p = \frac{p}{1-p}\sum_{y=1}^{\infty}[e^t(1-p)]^y.$$

For $0 < p < 1$, we can choose some finite t such that $|e^t(1-p)| < 1$ and use a geometric series to evaluate the sum:

$$\sum_{y=1}^{\infty}[e^t(1-p)]^y = \frac{1}{1-e^t(1-p)} - 1 = \frac{e^t(1-p)}{1-e^t(1-p)}.$$

Substitution yields $m(t) = pe^t/[1-e^t(1-p)]$.

E3.155 Unknown mgf. **(a)** $\mathrm{E}(Y) = m'(t=0) = \frac{7}{3}$. **(b)** $\mathrm{var}(Y) = m''(t=0) - (\mathrm{E}\,Y)^2 = 6 - \left(\frac{7}{3}\right)^2$. **(c)** $m(t) = \mathrm{E}(e^{tY}) = \sum_y p(y)e^{ty} = \frac{1}{6}e^t + \frac{2}{6}e^{2t} + \frac{3}{6}e^{3t}$. Matching terms, $y = 1, 2, 3$ with $p(i) = i/6$.

E3.158 Mgf of a linear transformation. $m_W(t) = \mathrm{E}(e^{tW}) = \mathrm{E}[e^{t(aY+b)}] = \mathrm{E}(e^{atY}e^{bt}) = e^{bt}\,\mathrm{E}(e^{atY})$ since the expectation is with respect to y. $\mathrm{E}(e^{atY}) = m_Y(at)$ and the result follows.

E3.159 Mgf of a linear transformation 2. $\mathrm{E}(W) = m_W'(t) = e^{tb}am_Y'(at) + m_Y(at)be^{tb}$ by the product and chain rules. At $t = 0$, $m_w'(t=0) = am_Y'(0) + bm_Y(0) = a\,\mathrm{E}(Y) + b$. A similar procedure yields the variance result.

***E3.162 Cumulant-generating function.** $r'(t) = m'(t)/m(t)$ and $r''(t) = \{m(t)m''(t) - [m'(t)]^2\}/[m(t)]^2$. It follows that $r'(0) = m'(0) = \mathrm{E}\,Y = \mu$ and $r''(0) = m''(0) - [m'(0)]^2 = \mathrm{E}\,Y^2 - (\mathrm{E}\,Y)^2 = \sigma^2$.

Comment. The third cumulant is also the third central moment, but this correspondence does not hold for higher cumulants.

3.10 Probability-Generating Functions

***E3.164 Binomial probability-generating function.**

$$P(t) = \mathrm{E}(t^Y) = \sum_{y=0}^{n} t^y \binom{n}{y} p^y (1-p)^{n-y} = \sum_{y=0}^{n} \binom{n}{y} (pt)^y (1-p)^{n-y}.$$

We recognize the above as the binomial expansion $(a+b)^n = \sum_{i=0}^{n} \binom{n}{i} a^i b^{n-i}$, with $a = pt$ and $b = (1-p)$. Thus $P(t) = [pt + (1-p)]^n$. The mean $\mathrm{E}\,Y = P'(t=1) = np[pt + (1-p)]^{n-1}|_{t=1} = np$.

***E3.165 Poisson probability-generating function.**

$$P(t) = \mathrm{E}(t^Y) = \sum_{y=0}^{\infty} t^y \frac{\lambda^y e^{-\lambda}}{y!} = e^{-\lambda} \sum_{y=0}^{\infty} \frac{(\lambda t)^y}{y!}$$

We recognize the sum as the Taylor series expansion of $e^{\lambda t}$, thus $P(t) = e^{\lambda(t-1)}$. The mean $\mathrm{E}\,Y = P'(t=1) = \lambda e^{\lambda(t-1)}|_{t=1} = \lambda$. The second factorial moment $\mathrm{E}[Y(Y-1)] = P''(t=1) = \lambda^2$, hence $\mathrm{E}\,Y^2 = \mathrm{E}[Y(Y-1)] + \mathrm{E}\,Y = \lambda^2 + \lambda$ and $\mathrm{var}\,Y = \mathrm{E}\,Y^2 - (\mathrm{E}\,Y)^2 = \lambda$.

***E3.166 Poisson distribution third moment.** $\mu_{[3]} = \mathrm{E}[Y(Y-1)(Y-2)] = \mathrm{E}\,Y^3 - 3\,\mathrm{E}\,Y^2 + 2\,\mathrm{E}\,Y$ allows us to compute $\mathrm{E}\,Y^3 = \mu_{[3]} + 3\,\mathrm{E}\,Y^2 - 2\,\mathrm{E}\,Y = \mu_{[3]} + 3\lambda^2 + \lambda$. Using the probability-generating function, $\mu_{[3]} = P^{(3)}(t=1) = \lambda^3$, thus $\mathrm{E}\,Y^3 = \lambda^3 + 3\lambda^2 + \lambda$.

3.11 Tchebysheff's Theorem

E3.168 Random test guesses. (a–c) $Y \sim \text{binomial}(n = 100, p = \frac{1}{5})$, so $\mu = np = 20$ and $\sigma = \sqrt{np(1-p)} = 4$. The intervals follow. **(a)** No. The exact binomial calculation $P = \sum_{y=50}^{100} p(y) \approx 2.14 \times 10^{-11}$. Using Tchebysheff's theorem, $k = (50 - \mu)/\sigma) = 7.5$ and $P \leq 1/k^2 = 0.01\overline{7}$.

Comment. The bound from Tchebysheff's theorem is very generous in this case.

E3.169 Attaining the Tchebysheff bound.

(a) $\mathrm{E}(Y) = \sum_y yp(y) = 0$, $\mathrm{var}(Y) = \mathrm{E}[(Y - \mathrm{E}\,Y)^2] = \sum_y (y-0)^2 p(y) = \frac{1}{9}$.

(b) $\sigma = \sqrt{\mathrm{var}(Y)} = \frac{1}{3}$. Thus $P(|Y - \mu| \geq 3\sigma) = P(|Y| \geq 1) = p(-1) + p(1) = \frac{1}{9}$. Using Tchebysheff's theorem with no distributional assumptions, $P(|Y - \mu| \geq 3\sigma) \geq \frac{1}{3^2} = \frac{1}{9}$.

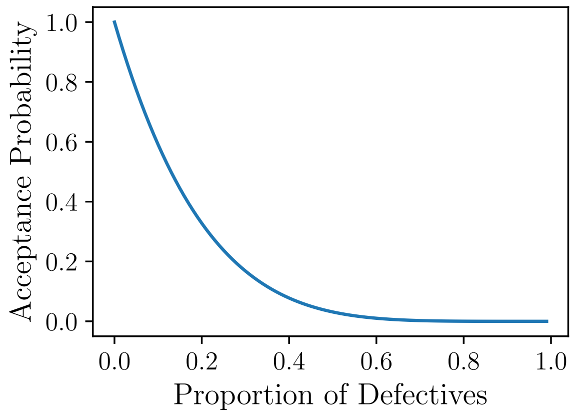
***(c)** See (d), with $k = 2$.

***(d)** Let $w = -1, 0, 1$ and force $p(-1) = p(1)$. Then $\mu_W = 0$ and $|W - \mu_W| = |W|$. It remains to choose $p(\pm 1)$ so that Tchebysheff's bound is attained: $P(|W| \geq k\sigma_W) = 1/k^2$. If $\sigma_W = 1/k$, then $P(|W| \geq 1) = p(-1) + p(1) = 1/k^2$, or $p(\pm 1) = 1/(2k^2)$. It follows that $p(0) = 1 - 1/k^2$.

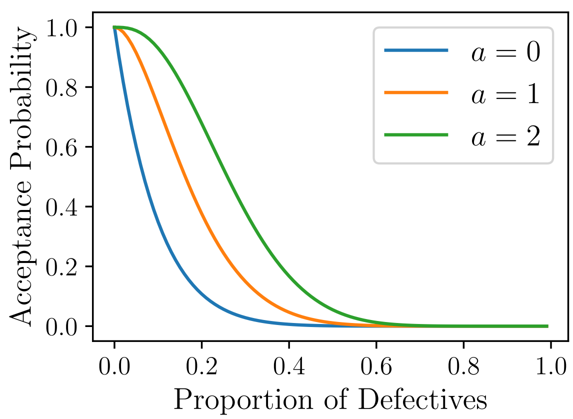
3.12 Summary

Supplementary Exercises

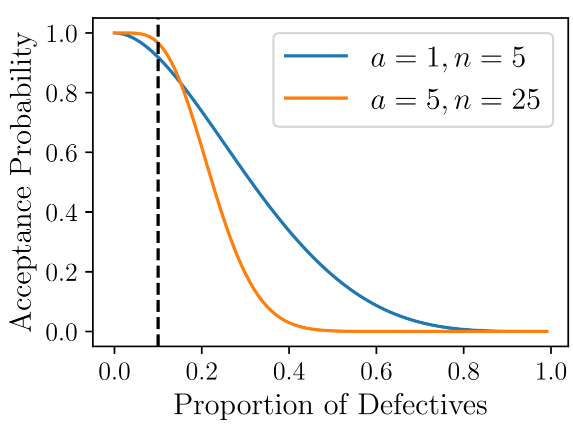
E3.181 Manufacturing sampling plans. Assuming the number of defectives $Y \sim \text{binomial}(n, p)$, the acceptance probability is $P(Y \leq a|n, p) = \sum_{k=0}^{a} \binom{n}{k} p^k (1-p)^{n-k}$, where a is the acceptance number cutoff, n is the sample size, and p is the probability a given item is defective. The operating characteristic curve for $n = 5$, $a = 0$ is below.



E3.182 Manufacturing sampling plans 2. Same logic as E3.181 with $n = 10$.



E3.183 Manufacturing sampling plans 3. The curves are below, with dotted line at $p = 0.1$.



(a) The seller wants to maximize the acceptance probability, thus plan $a = 5, n = 25$.
(b) The buyer wants to minimize the acceptance probability, thus plan $a = 5, n = 25$.

E3.184 City commissioner claim. **(a)** Let Y be the number of residents who prefer a private company. Assuming the commissioner's claim is correct, $Y \sim \text{binomial}(n = 25, p = \frac{4}{5})$, and thus $P(Y \geq 22) = \sum_{y=22}^{n} \binom{n}{22} p^y (1-p)^{n-y} \approx 0.234$. **(b)** $P(Y = 22) \approx 0.136$. **(c)** There is not sufficient evidence to reject the claim.

E3.185 Random integer selection. **(a)** Let Y be the number of students who choose 4, 5, or 6. Assuming their choices and random and independent, the probability of choosing 4, 5, or 6 is $\frac{3}{10}$. Thus $Y \sim \text{binomial}(n = 20, p = \frac{3}{10})$ and $P(Y \geq 8) = \sum_{y=8}^{n} \binom{n}{y} p^y (1-p)^{n-y} \approx 0.277$. **(b)** There is not sufficient evidence to suspect students made non-random, non-independent choices.

E3.187 Dice game. Let Y be the number of rolls the game lasts. Then $Y \sim \text{geometric}(p = \frac{5}{6})$ with pmf $P(Y = y) = (1-p)^{y-1}p$. **(a)** $P(Y = 3) \approx 0.0.231$ **(b)** $\mathrm{E}(Y) = 1/p = \frac{6}{5}$ **(c)** The payout is the random variable $P = 2^{Y-1}$, with $\mathrm{E}(P) = \mathrm{E}(2^{Y-1}) = \sum_{y=1}^{\infty} 2^{y-1} P(Y = y) = p \sum_{z=0}^{\infty} [2(1-p)]^z$, where $z = y - 1$. The sum can be evaluated with a geometric series $\sum_{i=0}^{\infty} r^i = 1/(1-r)$ for $|r| < 1$, yielding $\mathrm{E}(P) = p/[1 - 2(1-p)]$ for $p > \frac{1}{2}$, hence $\mathrm{E}(P) = \frac{5}{4}$.

E3.188 Conditioned binomial probability. $P(Y > 1 | Y \geq 1) = P(Y > 1 \cap Y \geq 1)/P(Y \geq 1)$ by definition. Since $Y > 1 \subset Y \geq 1$ the numerator is $P(Y > 1) = P(Y \geq 1) - P(Y = 1)$. The probabilities can be computed from the binomial pmf $P(Y = y) = \binom{n}{y} p^y (1-p)^{n-y}$. It is simple to compute $P(Y \geq 1) = 1 - P(Y = 0) = 1 - (1-p)^n$ and $P(Y = 1) = np(1-p)^{n-1}$, hence

$$P(Y > 1 | Y \geq 1) = \frac{P(Y \geq 1) - P(Y = 1)}{P(Y \geq 1)} = \frac{1 - (1-p)^n - np(1-p)^{n-1}}{1 - (1-p)^n}.$$

E3.193 Defective items in two lines. Assume each regulator has the same chance of being defective. There are $N = \binom{10}{4}$ orderings of the 4 defective parts. There are $n = \binom{5}{2} \times \binom{5}{2}$ ways two defectives were in line I (and line II), thus $p = n/N \approx 0.476$.

E3.194 Gambling concerns. Let Y be the number of games until her first win: $Y \sim \text{geometric}(p = 0.1)$. She must win in her first 3 plays, thus $P(\text{win once}) = \sum_{y=1}^{3} P(Y = y) = \sum_{y=1}^{3} (1-p)^{y-1} p = 0.271$.

E3.195 Textile weave imperfections. Let Y be the number of imperfections in one square yard. Then $Y \sim \text{Poisson}(\lambda = 4)$, with pmf $P(Y = y) = e^{-\lambda} \lambda^y / y!$. **(a)** $P(Y \geq 1) = 1 - P(Y = 0) \approx 0.982$. **(b)** This is a Poisson process with $\lambda^* = 3\lambda$, hence $P(Y \geq 1) \approx 1.00$.

E3.196 Textile weave imperfections 2. The cost is a random variable equal to $C = 10Y$, where $Y \sim \text{Poisson}(\lambda = 4 \times 8)$. Using $\mathrm{E}\, Y = \operatorname{var} Y = \lambda$, $\mu = 10\, \mathrm{E}\, Y = 320$ and $\sigma = \sqrt{10^2 \operatorname{var} Y} \approx 56.6$.

E3.199 Insulin-dependent diabetes incidence model. **(a)** Yes. We are given a mean incidence rate of $\lambda = 3 \times 10^{-4}$ per year. Let Y be the number of children with IDD: $Y \sim \text{Poisson}(1000\lambda)$. **(b)** $P(Y \geq 2) = 1 - P(Y = 0) - P(Y = 1) \approx 0.0369$.

***E3.202 Hierarchical model for parking.** Let $N \sim \text{Poisson}(\lambda)$ be the number of cars per minute driving past a parking area. Assuming each car parks independently with probability p, the number of cars that park per minute is $(W | N = n) \sim \text{binomial}(n, p)$.

(a) $P(W = 0) = e^{-p\lambda}$ from (b), which considers the general case.

(b) The unconditional probability $P(W = w) = \sum_{n=0}^{\infty} P(W = w | N = n) P(N = n)$ by the total law of probability. $P(N = n) = e^{-\lambda} \lambda^n / n!$ is straightforward. Since $W | N = n \sim \text{binomial}(n, p)$, $P(W = w | N = n) = \binom{n}{w} p^w (1-p)^{n-w}$ and

$$P(W = w) = \sum_{n=0}^{\infty} \binom{n}{w} p^w (1-p)^{n-w} \frac{\lambda^n}{n!} e^{-\lambda}.$$

Pulling terms with no n independence outside the sum and noting that $\binom{n}{w} = 0$ for $n < w$,

$$\begin{aligned}
P(W = w) &= \frac{1}{w!} p^w (1-p)^{-w} e^{-\lambda} \sum_{n=w}^{\infty} \frac{1}{(n-w)!} [(1-p)\lambda]^n \\
&= \frac{1}{w!} p^w (1-p)^{-w} e^{-\lambda} \sum_{z=0}^{\infty} \frac{1}{z!} [(1-p)\lambda]^{z+w} && (z = n - w) \\
&= \frac{1}{w!} (p\lambda)^w e^{-\lambda} \sum_{z=0}^{\infty} \frac{1}{z!} [(1-p)\lambda]^z \\
&= \frac{1}{w!} (p\lambda)^w e^{-\lambda} e^{(1-p)\lambda} && \text{(Series expansion)} \\
&= \frac{(p\lambda)^w}{w!} e^{-p\lambda}.
\end{aligned}$$

Evidently $W \sim \text{Poisson}(p\lambda)$.

Comment. We can view W as a Poisson process where the number of occurrences N is reduced by a constant factor p, hence there is an effective density $\lambda^* = p\lambda$.

E3.204 Turning at a red light. Let Y be the number of drivers who want to turn left, distributed as $Y \sim \text{binomial}(n = 5, p = 0.2)$. We are interested in $P(Y \leq 3) = \sum_{y=0}^{3} \binom{n}{y} p^y (1-p)^{n-y} \approx 0.993$.

E3.205 Dice tosses. We require 2 sixes in the first 8 tosses, with probability $p_1 = \binom{8}{2} \left(\frac{1}{6}\right)^2 \left(\frac{5}{6}\right)^6$ and a six in the last two tosses, with probability $p_2 = \left(\frac{1}{6}\right)^2$. Since the two series of tosses are independent, $p = p_1 p_2 \approx 7.24 \times 10^{-3}$.

Comment. A combinatorics argument is as follows. There are $\binom{8}{2}$ ways to choose 2 sixes in the first 8 tosses; for each of those tosses, there are $\binom{6}{6} \times 5^6$ ways to choose the remaining tosses. For each sequence of 8 tosses, there is one way to choose the last 2 to be sixes. Thus $n = \binom{8}{2}\binom{6}{6} \times 5^6$. The total number of possible tosses is $N = 6^{10}$. Hence $p = n/N \approx 7.24 \times 10^{-3}$.

E3.206 Automobile insurance. Let A be the event an insured driver has an accident and let Y be the proportion the insurance company pays. $\mathrm{E}(Y) = \sum_y yP(Y = y) = \sum_y yP(Y = y|A)P(A) = 0.0468$. The total cost $C = -12000Y$ with $\mathrm{E}(C) = -561.6$, thus the premium must be 561.6 to break even.

***E3.211 Optimal merchant stock.** We are given the distribution of D, the number of items in demand. Let b be the number of items the merchant stocks. The profit is the random variable

$$Y = \begin{cases} 1.2D - b & \text{if } b > D \\ 1.2b - b & \text{if } b \leq D \end{cases}, \qquad \text{for integer } b, D > 0.$$

$b = 2$ and $b = 4$ are simple, since either $b > D$ or $b \leq D$. $\mathrm{E}(Y|b = 2) = 1.2(2) - 2 = 0.4$ and $\mathrm{E}(Y|b = 4) = 1.2\,\mathrm{E}(D|4 \geq D) - 4 = 0.08$. $b = 3$ contains both conditions, so considering all cases manually,

$$\mathrm{E}(Y|b = 3) = \sum_y yP(Y = y|b = 3) = \sum_d yP(Y = y|b = 3) = 0.48$$

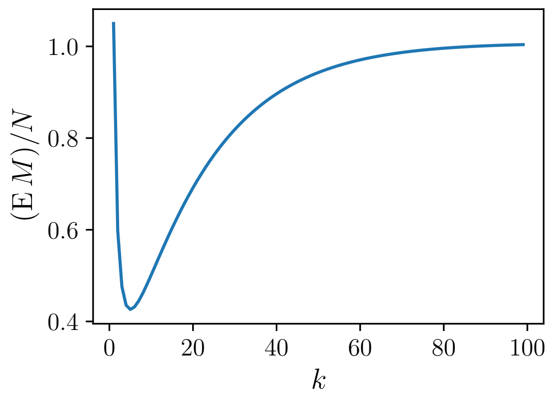
since each d corresponds directly to a y. Thus $b = 3$ maximizes profit.

***E3.214 Insurance model for accidents.** Let A denote the event a new subscriber has an accident during the first year and B the event they have an accident the next year. We are interested in $P(B|A) = P(A \cap B)/P(A)$, since A and B are not independent. Let S denote a safe driver. Then $P(A) = P(A|S)P(S) + P(A|\overline{S})P(\overline{S}) = 0.22$. The numerator can be computed similarly: $P(A \cap B) = P(A \cap B|S)P(S) + P(A \cap B|\overline{S})P(\overline{S}) = 0.082$, yielding $P(B|A) = 0.3\overline{72}$ and a premium of $\$400 \times P(B|A) \approx \$149$.

***E3.215 Pooled blood tests.** Let $p = 0.05$, the proportion of members with disease A.

(a) The probability a pooled test of k people is negative is $P_{k-} = (1-p)^k$. Thus, the number of negative pooled tests $Y \sim \text{binomial}(n, P_{k-})$. The number of total tests needed is $M = Y + (k+1)(n-Y)$, with $\mathrm{E}\, M = \mathrm{E}\, Y + (k+1)(n - \mathrm{E}\, Y) = nP_{k-} + (k+1)(n - nP_{k-})$.

(b) Explicitly substituting $P_{k-} = (1-p)^k$ and $n = N/k$ yields $(\mathrm{E}\, M)/N = \{1 + k[1-(1-p)^k]\}/k$. Solving $d\,\mathrm{E}\, M/dk = 0$ for k cannot be done using analytic functions. Turning to numerics, $(\mathrm{E}\, M)/N$ is plotted as a function of integer k; it is minimized by $k = 5$.



(c) With $k = 5$, $\mathrm{E}\, M \approx 0.426N$ which, on average, saves $(N - \mathrm{E}\, M) \approx 0.574N$ tests.

4 Continuous Variables and Their Probability Distributions

4.1 Introduction

4.2 The Probability Distribution for a Continuous Random Variable

E4.2 Guess the key. **(a)** This is like E2.121, with $p(y) = \frac{1}{5}$ for $y = 1, 2, \ldots, 5$. **(b)** $F(y)$ is a step function with jumps of height $\frac{1}{5}$ at each y. **(c)** $P(Y < 3) = F(2) = \frac{2}{5}$, $P(Y \le 3) = F(3) = \frac{3}{5}$, and $P(Y = 3) = p(3) = \frac{1}{5}$. **(d)** No, here Y is discrete.

E4.3 Bernoulli random variable.
(a)

$$F(y) = \begin{cases} 0 & \text{for } y < 0 \\ 1-p & \text{for } 0 \le y < 1 \\ 1 & \text{for } y \ge 1 \end{cases} .$$

(b) Indeed, $F(y)$ is nondecreasing and tends to 0 and 1 as $y \to -\infty$ and $y \to \infty$, respectively.

E4.4 $n = 1$ binomial random variable. **(a)** $p(y = 0) = 1 - p$ and $p(y = 1) = p$. The probability and thus distribution function is the same as E4.3. **(b)** A single trial binomial r.v. is a Bernoulli r.v.

E4.5 Discrete pmf from the cdf. Since Y is discrete, $p(y) = P(Y \le y) - P(Y \le y-1) = F(y) - F(y-1)$ for $y - 1 = 1, 2, \ldots$ The case $y = 1$ is treated separately: $p(y = 1) = P(Y \le 1) = F(1)$.

E4.6 Cdf of a geometric r.v. **(a)** $F(y) = P(Y \le y) = \sum_{i=1}^{y} p(i) = \sum_{i=1}^{y} q^{i-1}p = p\sum_{z=0}^{y-1} q^z$. The geometric series sums to $(1 - q^y)/(1 - q)$ and since $p = 1 - q$, $F(y) = 1 - q^y$ for $y = 1, 2, \ldots$ Further, $F(y < 1) = 0$. This is the same as the given result, which is presented as a sequence of steps. **(b)** Indeed, $F(y)$ is nondecreasing and tends to 0 and 1 as $y \to -\infty$ and $y \to \infty$, respectively.

E4.15 Distribution of maze completion times.
(a) Assume $b > 0$. Then $f(y) \ge 0$ for all y and $\int_{-\infty}^{\infty} f(y)\,dy = \int_b^{\infty} by^{-2}\,dy = -by^{-1}\big|_b^{\infty} = 1$.
(b) $F(y) = 0$ for $y < b$ and $F(y) = \int_b^y f(x)\,dx = -bx^{-1}\big|_b^y = 1 - b/y$ for $y \ge b$.
(c) $P(Y > b + c) = \int_{b+c}^{\infty} f(y)\,dy$. Perhaps easier, $P(Y > b + c) = 1 - F(b + c) = b/(b + c)$.
(d) $P(Y > b + d | Y > b + c) = P(Y > b + d)/P(Y > b + c)$ since $(Y > b + d) \subset (Y > b + c)$. Thus $P(Y > b + d | Y > b + c) = (b + c)/(b + d)$.

4.3 Expected Values for Continuous Random Variables

***E4.35 Minimizing the expected square distance.**

$$\begin{aligned} \mathrm{E}(Y - a)^2 &= \mathrm{E}[(Y - \mathrm{E}\,Y) + (\mathrm{E}\,Y - a)]^2 \\ &= \mathrm{E}(Y - \mathrm{E}\,Y)^2 + \mathrm{E}(\mathrm{E}\,Y - a)^2 + 2\,\mathrm{E}[(Y - \mathrm{E}\,Y)(\mathrm{E}\,Y - a)] \end{aligned}$$

by linearity of the expectation. Recognize $c = (\mathrm{E}\,Y - a)$ as a constant, thus the last term is zero. Hence minimizing w.r.t. a is the same as minimizing c^2, which is done by setting $c = 0$, or $a = \mathrm{E}\,Y$.

***E4.36 Minimizing the expected square distance 2.** Yes. The discrete expectation is also linear.

***E4.37 Expectation of a symmetric r.v.** $\mathrm{E}\,Y = \int_{-\infty}^{\infty} yf(y)\,dy = \int_{-\infty}^{0} yf(y)\,dy + \int_0^{\infty} yf(y)\,dy \equiv I_1 + I_2$. Let $z = -y$; $I_1 = \int_{\infty}^{0}(-z)f(-z)\,(-dz) = -\int_0^{\infty} zf(z)\,dz = -I_2$ since $f(-z) = f(z)$. Thus $\mathrm{E}\,Y = 0$.

4.4 The Uniform Probability Distribution

E4.38 Uniform cdf. **(a)** For general $Y \sim \text{uniform}(\theta_1, \theta_2)$, $F(y) = \int_{-\infty}^{y} f(y)\,dy = \int_{\theta_1}^{y} \frac{1}{\theta_2 - \theta_1}\,dy = \frac{y - \theta_1}{\theta_2 - \theta_1}$, $\theta_1 < y < \theta_2$. With $(\theta_1, \theta_2) = (0, 1)$, $F(y) = y$. **(b)** $P(a \le Y \le a+b) = F(a+b) - F(a) = a+b-a = b$.

E4.43 Area of a random circle. Denote the circle radius as $R \sim \text{uniform}(0,1)$ and let $g(R) = \pi R^2$. By Theorem 4.4, $\mathrm{E}[g(R)] = \int_{-\infty}^{\infty} g(r) f(r)\, dr = \int_0^1 r^2\, dr = \frac{\pi}{3}$. A similar calculation yields the variance $\mathrm{var}[g(R)] = \mathrm{E}\{[g(R)]^2\} - \{\mathrm{E}[g(R)]\}^2 = \frac{\pi^2}{5} - \frac{\pi^2}{9} = \frac{4\pi^2}{45}$.

4.5 The Normal Probability Distribution

E4.60 Properties of the normal distribution. $e^x \geq 0$ for all x, so in order for $f(y) \geq 0$, $\sigma \geq 0$. Further, $\sigma \neq 0$ since otherwise $f(y)$ would not be normalizable, hence $\sigma > 0$. □

E4.61 Properties of the normal distribution 2. The median $m = \mu$ since the distribution is symmetric about μ.

E4.78 Properties of the normal distribution 3. The argument of the exponential $x \leq 0$, and $e^{-|x|}$ is maximized at $x = 0$ when $y = \mu$, with corresponding value $1/\sigma\sqrt{2\pi}$. □

E4.79 Properties of the normal distribution 4. $f'(y) = -(y-\mu)e^{-(y-\mu)^2/(2\sigma^2)}/(\sigma^3\sqrt{2\pi})$ and $f''(y) = \left[(y-\mu)^2 - \sigma^2\right] e^{-(y-\mu)^2/(2\sigma^2)}/(\sigma^5\sqrt{2\pi})$. $f''(y) = 0$ when the term in brackets is zero, or $y = \mu \pm \sigma$.

E4.80 Folded normal distribution. $A = 3Y^2$, so $\mathrm{E}(A) = 3\,\mathrm{E}(Y^2) = 3[\mathrm{var}\, Y + (\mathrm{E}\, Y)^2] = 3(\sigma^2 + \mu^2)$.

Comment. To compute $\mathrm{E}(L) = \mathrm{E}(|Y|)$, note that $X = |Y|$ has a folded-normal distribution. That is, since $|y|$ maps $\pm y$ to the same point, given normal distribution function $f_Y(y)$

$$f_X(x) = f_Y(x) + f_Y(-x) = \left[e^{-(x-\mu)^2/(2\sigma^2)} + e^{-(x+\mu)^2/(2\sigma^2)}\right] \Big/ (\sigma\sqrt{2\pi})$$

for $x \geq 0$ and $f_X(x) = 0$ for $x < 0$. It follows that $\mathrm{E}(|Y|) = 2\sigma e^{-\mu^2/(2\sigma^2)}/\sqrt{2\pi}$.

4.6 The Gamma Probability Distribution

E4.81 Properties of the gamma function.

(a) $\Gamma(1) = \int_0^\infty e^{-y}\, dy = -e^{-y}\big|_0^\infty = 1$.

***(b)** $\Gamma(\alpha) = -y^{\alpha-1}e^{-y}\big|_0^\infty - \int_0^\infty (\alpha-1)y^{\alpha-2}(-e^{-y})\, dy = (\alpha-1)\Gamma(\alpha-1)$.

E4.95 Exponential and geometric distribution connection.

(a) $P(X=k) = P(k-1 \leq Y < k) = \int_{k-1}^k f(y)\, dy = \int_{k-1}^k \frac{1}{\beta} e^{-y/\beta}\, dy = e^{-(k-1)/\beta} - e^{-k/\beta}$.

(b) Factoring $e^{-(k-1)/\beta}$ yields $P(X=k) = e^{-(k-1)/\beta}(1 - e^{-1/\beta}) = (e^{-1/\beta})^{k-1}(1 - e^{-1/\beta})$.

***E4.100 Gamma and Poisson distribution connection.**

(a) $p(x) = \lambda^x e^{-\lambda}/x!$. Thus $p(x_1) = e^{-\lambda_1} > e^{-\lambda_2} = p(x_2)$ since $\lambda_2 > \lambda_1$.

(b) $P(X_i \leq k) = \sum_{i=0}^k p(x_i) = \sum_{i=0}^k \lambda_i^{x_i} e^{-\lambda_i}/x_i! = P(Y > \lambda_i)$ where $Y \sim \text{gamma}(\alpha = k+1, \beta = 1)$ using the result in the problem statement.

(c) $P(X_1 \leq k) = P(Y \geq \lambda_1) > P(Y \geq \lambda_2) = P(X_2 \leq k)$ since $\lambda_2 > \lambda_1$.

(d) X_1 tends to be smaller than X_2, which agrees with our intuition since $\lambda_1 < \lambda_2$.

E4.111 Moments of the gamma distribution.

(a)

$$\mathrm{E}(Y^a) = \int_0^\infty y^a \frac{y^{\alpha-1}e^{-y/\beta}}{\beta^\alpha \Gamma(\alpha)} = \frac{1}{\beta^\alpha\Gamma(\alpha)} \int_0^\infty y^{\alpha+a-1} e^{-y/\beta} = \frac{\beta^{\alpha+a}\Gamma(\alpha+a)}{\beta^\alpha\Gamma(\alpha)} \int_0^\infty \frac{y^{\alpha+a-1}e^{-y/\beta}}{\beta^{\alpha+a}\Gamma(\alpha+a)}.$$

The integral is over the domain of a $\text{gamma}(\alpha_* = \alpha + a, \beta)$ distribution and thus evaluates to one. Canceling βs yields the result.

(b) We require $\alpha + a > 0$ because the parameter $\alpha_* > 0$, allowing us the evaluate the integral.

(c) $\Gamma(\alpha+1) = \alpha\Gamma(\alpha)$ using the recursion relation proved in E4.81(b) yields the result.

(d, e) $\mathrm{E}(\sqrt{Y}) = \beta^{1/2}\Gamma(\alpha + \frac{1}{2})/\Gamma(\alpha)$, assuming $\alpha + \frac{1}{2} > 0$. The result for other powers is similar.

4.7 The Beta Probability Distribution

E4.127 Uniform distribution from beta distribution. $f(y|\alpha = \beta = 1) = 1/B(1,1) = 1$, $0 \le y \le 1$.

***E4.135 Binomial and beta distribution connection.**

(a) $P(Y_1 = 0) = (1-p_1)^n > (1-p_2)^n = P(Y_2 = 0)$ since $p_1 < p_2$.

(b)

$$
\begin{aligned}
P(Y_1 \le k) &= \sum_{i=0}^{k} \binom{n}{i} p_1^i (1-p_1)^{n-i} = 1 - \sum_{i=k+1}^{n} \binom{n}{i} p_1^i (1-p_1)^{n-i} \\
&= 1 - \int_0^{p_1} \frac{t^k (1-t)^{n-k-1}}{B(k+1, n-k)}\, dt = \int_{p_1}^1 \frac{t^k (1-t)^{n-k-1}}{B(k+1, n-k)}\, dt.
\end{aligned}
$$

(c) From (b), $P(Y_i \le k) = 1 - F(p_i)$, where $F(\cdot)$ is a beta distribution function. Thus we have $P(Y_1 \le k) = 1 - F(p_1) > 1 - F(p_2) = P(Y_2 \le k)$ since $p_1 < p_2$. Y_1 tends to be smaller than Y_2, which agrees with our intuition since $p_1 < p_2$.

4.8 Some General Comments

4.9 Other Expected Values

E4.136 Exponential mgf.

(a)

$$
m(t) = \mathrm{E}(e^{tY}) = \int_0^\infty e^{ty} \left(\frac{1}{\theta} e^{-y/\theta} \right) dy = \frac{\beta}{\theta} \int_0^\infty \frac{1}{\beta} e^{-y/\beta}\, dy
$$

where $\beta = (1/\theta - t)^{-1}$. For $t < 1/\theta$, the integral is over a exponential density evaluating to one. Thus after simplification, $m(t) = (1 - \theta t)^{-1}$.

(b) Differentiating w.r.t. t, $m'(t) = \theta(1-\theta t)^{-2}$, hence $\mathrm{E}(Y) = m'(t=0) = \theta$. Similarly, $m''(t) = 2\theta^2 (1-\theta t)^{-3}$ and $m''(t=0) = \mathrm{E}(Y^2) = 2\theta^2$. Thus $\mathrm{var}(Y) = \mathrm{E}\,Y^2 - (\mathrm{E}\,Y)^2 = \theta^2$.

E4.139 Linear transformation of a normal r.v. $m_X(t) = e^{bt} m_Y(at) = e^{bt} e^{\mu a t + \sigma^2 (at)^2/2} = e^{\mu_* t + \sigma_*^2 t^2/2}$ where $\mu_* = a\mu + b$ and $\sigma_* = a\sigma$. Thus $X \sim \text{normal}(a\mu + b, a\sigma)$. Evidently a change of location and scale of a normal r.v. is still normal.

E4.141 Uniform mgf.

$$
m(t) = \mathrm{E}(e^{tY}) = \int_{\theta_1}^{\theta_2} e^{ty} \left(\frac{1}{\theta_2 - \theta_1} \right) dy = \frac{1}{\theta_2 - \theta_1} \left. \frac{e^{ty}}{t} \right|_{\theta_1}^{\theta_2} = \frac{e^{\theta_2 t} - e^{\theta_1 t}}{(\theta_2 - \theta_1) t}.
$$

E4.142 Linear transformation of a standard uniform r.v.

(a) Use E4.141 with $\theta_1 = 0$ and $\theta_2 = 1$: $m(t) = (e^t - 1)/t$.

(b) $m_W(t) = m_Y(at) = (e^{at} - 1)/(at)$ is a uniform r.v. with $\theta_1 = 0$ and $\theta_2 = a$ by inspection.

(c) $m_X(t) = m_Y(-at) = (1 - e^{-at})/(at)$ is a uniform r.v. with $\theta_1 = -a$ and $\theta_2 = 0$ by inspection.

(d) $m_V(t) = e^{bt} m_Y(at) = (e^{(a+b)t} - e^{bt})/(at)$ is a uniform r.v. with $\theta_1 = b$ and $\theta_2 = a + b$.

4.10 Tchebysheff's Theorem

4.11 Expectations of Discontinuous Functions and Mixed Probability Distributions

Mixed distribution. The text states $F(y) = c_1 F_1(y) + c_2 F_2(y)$, where $c_1 + c_2 = 1$, c_i representing the total accumulated probability of being in the i-th mixture. The latter follows from the property

$$
1 = F(y \to \infty) = c_1 F_1(y \to \infty) + c_2 F_2(y \to \infty) = c_1 + c_2.
$$

They can also be viewed as the coefficients $P(Y_i)$ using the total law of probability:

$$F(y) = F(y|Y_1)P(Y_1) + F(y|Y_2)P(Y_2) = c_1F_1(y) + c_2F_2(y).$$

It follows for densities that

$$f(y) = c_1f_1(y) + c_2f_2(y)$$

since $f(y) = dF(y)/dy$ and the derivative is a linear operator. It is then easy to prove Definition 4.15: $\mathrm{E}\, g(Y) = \int_{-\infty}^{\infty} g(y)f(y)\,dy = \int_{-\infty}^{\infty} g(y)[c_1f_1(y) + c_2f_2(y)]\,dy = c_1\,\mathrm{E}\, g(Y_1) + c_2\,\mathrm{E}\, g(Y_2)$ where Y_i is the r.v. of the i-th mixed component with cdf $F_i(\cdot)$.

Mixed distributions can be generalized for n mixtures $Y_1, Y_2, \ldots, Y_n$. The cdf for the mixture Y is the linear combination $F(y) = \sum_{i=1}^{n} c_iF_i(y)$ satisfying $\sum_{i=1}^{n} c_i = 1$.

***E4.155 Cost of supply delay.** $Y \sim$ uniform$(1, 4)$ with $f(y) = \frac{1}{3}$ for $1 \le y \le 4$. The cost is the r.v.

$$C(Y) = \begin{cases} 100, & 1 \le Y \le 2 \\ 100 + 20(Y - 2), & 2 \le Y \le 4 \end{cases}$$

with expected value $\mathrm{E}(C) = \int_{-\infty}^{\infty} C(y)f(y)\,dy = \int_1^2 \frac{100}{3}\,dy + \int_2^4 \frac{100+20(y-2)}{3}\,dy = 113.\overline{3}$.

***E4.156 Duration of timed phone calls.** The duration is a mixture of a discrete r.v. Y_1 with pmf $g_1(y = 2) = \frac{2}{3}$ and $g_1(y = 6) = \frac{1}{3}$ and a continuous r.v. $Y_2 \sim$ gamma$(\alpha = 2, \beta = 2)$ with pdf $g_2(y)$. The mixture Y has pdf $g(y) = c_1g_1(y) + c_2g_2(y)$ where $c_1 = 0.2 + 0.1 = 0.3$ and $c_2 = 1 - c_2 = 0.7$, with mean $\mathrm{E}\, g(Y) = c_1\,\mathrm{E}\, g_1(Y) + c_2\,\mathrm{E}\, g_2(Y) = 4$.

***E4.157 Usage time of maintained electronic components.**

(a) $Y \sim$ exponential$(\beta = 100)$ with pdf $f(y) = e^{-y/\beta}/\beta$, $y \ge 0$. The usage time is the r.v.

$$X(Y) = \begin{cases} Y, & 0 \le Y < 200 \\ 200, & Y \ge 200 \end{cases}.$$

It has the cdf of Y for $0 \le Y < 200$ and a step of size $P(Y \ge 200) = e^{-2}$ at $Y = 200$:

$$F_X(x) = \begin{cases} 1 - e^{-x/\beta}, & 0 \le x < 200 \\ 1, & x \ge 200 \end{cases}.$$

(b) $\mathrm{E}(X) = \int_{\infty}^{\infty} X(y)f(y)\,dy = \int_0^{200} ye^{-y/\beta}/\beta\,dy + \int_{200}^{\infty} 200e^{-y/\beta}/\beta\,dy = 100(1 - e^{-2}) \approx 86.5$.

***E4.158 Kinetic energy of a projectile.** The kinetic energy K is a mixture of two distributions: continuous $Y = mV^2/2$ and discrete Z with $P(Z = 0) = 0.02$. Thus $f_K(k) = c_Yf_Y(k) + c_Zf_Z(k)$ where $c_Y = 0.98$ and $c_Z = 0.02$, and $\mathrm{E}(K) = c_Y\,\mathrm{E}(Y) + c_Z\,\mathrm{E}(Z) = 2.45 \times 10^6 \times m$.

***E4.159 Individual components from a mixture distribution.**

(a, b) Y_1 has two points at $y = 0$ and $y = 0.5$. Y_2 is quadratic for $0 < y < 0.5$ and linear for $0.5 < y < 1$. Thus

$$F_1(y) = 4 \times \begin{cases} 0.1, & 0 \le y < 0.5 \\ (0.1 + 0.15), & 0.5 \le y < 1 \end{cases}, \qquad F_2(y) = \frac{4}{3} \times \begin{cases} y^2, & 0 \le y < 0.5 \\ (y - 0.25), & 0.5 \le y < 1 \end{cases}.$$

The constants are chosen so $F_i(y)$ are valid (single) distribution functions, e.g. $F_i(y)$ tend to 0 and 1 at the endpoints and $F_2(y)$ is continuous everywhere. It follows that $c_1 = \frac{1}{4}$ and $c_2 = \frac{3}{4}$.

(c) Using $p(y_1 = 0) = 0.4$, $p(y_1 = 0.5) = 0.6$, and

$$f(y_2) = \frac{4}{3} \times \begin{cases} 2y, & 0 \le y < 0.5 \\ 1, & 0.5 \le y < 1 \end{cases},$$

we compute $\mathrm{E}(Y) = c_1\,\mathrm{E}(Y_1) + c_2\,\mathrm{E}(Y_2) = 0.5\overline{3}$ and $\mathrm{var}(Y) = c_1\,\mathrm{var}(Y_1) + c_2\,\mathrm{var}(Y_2) = 0.07597\overline{2}$ where the variance calculations assume $\mu = \mathrm{E}(Y)$ for the mixture, not $\mathrm{E}(Y_i)$.

4.12 Summary

Supplementary Exercises

E4.160 A symmetric density.
(a) $F(y) = \int_{-1}^{y} f(t)\,dt = [4\arctan(y) + \pi]/(2\pi)$ for $-1 \le y \le 1$, $F(y < -1) = 0$ and $F(y > 1) = 1$.
(b) $\mathrm{E}(Y) = \int_{-1}^{1} yf(y)\,dy = 0$ since $f(y)$ is even, hence $yf(y)$ is odd.

E4.167 Moments of the beta distribution.

$$\mathrm{E}(Y^k) = \int_0^1 y^k \frac{y^{\alpha-1}(1-y)^{\beta-1}}{B(\alpha,\beta)}\,dy = \frac{B(\alpha+k,\beta)}{B(\alpha,\beta)} \int_0^1 \frac{y^{\alpha+k-1}(1-y)^{\beta-1}}{B(\alpha+k,\beta)}\,dy.$$

The integral is over the domain of a beta($\alpha_* = \alpha + k, \beta$) distribution and thus evaluates to one. In terms of gamma functions,

$$\mathrm{E}(Y^k) = \frac{B(\alpha+k,\beta)}{B(\alpha,\beta)} = \frac{\Gamma(\alpha+k)\Gamma(\beta)\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)\Gamma(\alpha+\beta+k)} = \frac{\Gamma(\alpha+k)\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\alpha+\beta+k)}.$$

E4.168 Counter arrival times. The distribution function is

$$F(t) = P(T \le t) = 1 - P(T > t) = 1 - P(N = 0|t) = 1 - e^{-\lambda t}$$

since $N \sim$ Poisson(λt). Thus the density $f(t) = dF/dt = \lambda e^{-\lambda t}$. Note that $T \sim$ exponential($1/\lambda$).

E4.169 Counter arrival times 2. Let T denote the time between calls, $T \sim$ exponential($1/\lambda$) where $\lambda = 10$. $P(T \ge 0.25) \approx 0.0821$.

***E4.170 Counter arrival times 3.**
(a) See (b) with $k = 2$.
(b) Let K be time until the k-th arrival time, with distribution function

$$F(t) = P(K \le t) = 1 - P(K > t) = 1 - \sum_{n=0}^{k-1} P(N = n) = 1 - \sum_{n=0}^{k-1} \frac{(\lambda t)^n e^{-\lambda t}}{n!}.$$

Using the Poisson-gamma connection presented in E4.100, it is clear that $F(t)$ is the distribution function for a gamma distribution with $\alpha = k$ and $\beta = 1$. Explicitly,

$$f(t) = \frac{dF}{dt} = \sum_{n=0}^{k-1} \frac{\lambda^n e^{-\lambda t}}{n!}\left[\lambda t^n - n t^{n-1}\right] = e^{-\lambda t}\left[\sum_{n=0}^{k-1} \frac{\lambda^{n+1} t^n}{n!} - \sum_{m=0}^{k-2} \frac{\lambda^{m+1} t^m}{m!}\right].$$

where $m = n - 1$ (note the $m = -1$ term is zero). The sums cancel each other except for the $n = k - 1$ term, hence $f(t) = \lambda^k t^{k-1} e^{-\lambda t}/(k-1)! = t^{k-1}e^{-t/\beta}/[\beta^k \Gamma(k)]$ where $\beta = 1/\lambda$ is indeed the beta density with $\alpha = k$.

Comment. This problem can be solved using mgfs. Let T_i be the time of the i-th arrival. $T_i \simeq T \sim$ exponential($1/\lambda$) by the Poisson process assumption and $K = \sum_{i=1}^{K} T_i$. Assuming T_i are independent, $m_K(t) = \sum_{i=1}^{k} m_T(t) = [m_T(t)]^k = (1 - t/\lambda)^{-k}$ is the mgf of a gamma distribution with $\alpha = k$ and $\beta = 1/\lambda$.

E4.171 Counter arrival times 4. (a) We showed in E4.168 that the arrival time $T \sim$ exponential($1/\lambda$), hence $\mu = \sigma = 1/\lambda$. **(b)** $P(N \ge 1|t = 3) = 1 - P(N = 0|t = 3) = 1 - e^{-\lambda t}$.

E4.173 Plant distance distribution. Let $N \sim$ Poisson(λa) denote the number of plants in an area $a = \pi r^2$ and R denote the nearest distance between two plants. This is equivalent to E4.168 with λ representing a spatial instead of temporal density. R has cdf

$$F(r) = P(R \le r) = 1 - P(R > r) = 1 - P(N = 0|r) = 1 - e^{-\lambda\pi r^2},$$

hence density $f(r) = 2\lambda\pi r e^{-\lambda\pi r^2}$.

E4.176 Median of the exponential distribution. $Y \sim$ exponential(β) has cdf $F(y) = 1 - e^{-y/\beta}$. The median m occurs when $F(m) = 1 - e^{-m/\beta} = \frac{1}{2}$, or $m = -\beta \log(\frac{1}{2}) = \beta \log(2)$.

***E4.179 Maximizing retail grocer profit.** Let r_b and r_s be the rates she buys and sells, respectively. Her profit is the r.v.

$$P(Y) = 100 \times \begin{cases} r_s k - r_b k, & Y \geq k \\ r_s Y - r_b k, & Y < k \end{cases},$$

so her expected profit is

$$\mathrm{E}\, P = \int_{-\infty}^{\infty} P(y) f(y)\, dy = 300 \left[\int_0^k (r_s y - r_b k) y^2\, dy + \int_k^1 (r_s k - r_b k) y^2\, dy \right] = 25k[4(r_s - r_b) - r_s k^3]$$

with derivative $d \mathrm{E}\, P/dk = 100[(r_s - r_b) - r_s k^3] = 0$ at $k_* = (1 - r_b/r_s)^{1/3}$ and second derivative $d^2 \mathrm{E}\, P/dk^2|_{k=k_*} = -300 r_s (1 - r_b/r_s)^{2/3} < 0$ for $r_s > r_b > 0$, verifying k_* is a minimum. Substituting $r_b = 0.06$ and $r_s = 0.1$ yields $k_* \approx 0.737$.

Comment. The distribution of P can be computed like in E4.157. $F(p)$ has the cdf of $100(r_s Y - r_b k)$ for $Y < k$ and $F(p) = 1$ for $Y \geq k$, with a step size of $P(Y \geq k) = 1 - k^3$.

E4.184 Laplace distribution.

$$m(t) = \frac{1}{2} \int_{-\infty}^{\infty} e^{ty - |y|}\, dy = \frac{1}{2} \left[\int_{-\infty}^{0} e^{y(t+1)}\, dy + \int_0^{\infty} e^{y(t-1)}\, dy \right] = \frac{1}{2} \left[\frac{1}{t+1} - \frac{1}{t-1} \right] = \frac{1}{1 - t^2}.$$

$m'(t) = 2t/(1-t^2)^2$, thus $\mathrm{E}\, Y = m'(t = 0) = 0$.

Comment. This is a special case of the Laplace distribution $f(y; \mu, b) = e^{-|y-\mu|/b}/(2b)$.

***E4.185 Mixture density.**

(a) $f(y) \geq 0$ for all y since $f_i(y) \geq 0$ and $0 \leq a \leq 1$. Furthermore, $\int_{-\infty}^{\infty} f(y)\, dy = a + (1 - a) = a$ since f_i are densities. $\square$

(b) $\mathrm{E}\, Y = \int_{-\infty}^{\infty} y f(y)\, dy = a \int_{-\infty}^{\infty} y f_1(y)\, dy + (1-a) \int_{-\infty}^{\infty} y f_2(y)\, dy = a\, \mathrm{E}\, Y_1 + (1-a)\, \mathrm{E}\, Y_2$. Similarly, $\mathrm{E}\, Y^2 = a\, \mathrm{E}\, Y_1^2 + (1-a)\, \mathrm{E}\, Y_2^2$ and $\operatorname{var} Y = \mathrm{E}\, Y^2 - (\mathrm{E}\, Y)^2$. Simplification yields the result.

***E4.186 Weibull distribution moments.** Consider the general case. The k-th moment

$$\begin{aligned} \mathrm{E}\, Y^k &= \frac{m}{\alpha} \int_0^{\infty} y^{m+k-1} e^{-y^m/\alpha}\, dy \\ &= \alpha^{k/m} \int_0^{\infty} u^{k/m} e^{-u}\, du && (u = y^m/\alpha) \\ &= \alpha^{k/m} \Gamma(k/m + 1). && (\Gamma(z) = \textstyle\int_0^{\infty} u^{z-1} e^{-u}\, du) \end{aligned}$$

Thus $\mu = \mathrm{E}\, Y = \alpha^{1/m} \Gamma(1/m + 1)$ and $\sigma^2 = \mathrm{E}\, Y^2 - (\mathrm{E}\, Y)^2 = \alpha^{2/m} \{\Gamma(2/m + 1) - [\Gamma(1/m + 1)]^2\}$.

***E4.190 Hazard rate function for Weibull densities.**

(a) $T \sim$ exponential(α) has $f(t) = e^{-t/\alpha}/\alpha$ and $F(t) = 1 - e^{-t/\alpha}$. Substitution yields $r(t) = 1/\alpha$.

(b) $T \sim$ Weibull$(\alpha, m > 1)$ has $f(t) = (m/\alpha) t^{m-1} e^{-t^m/\alpha}$ and

$$F(t) = \frac{m}{\alpha} \int_0^t y^{m-1} e^{-y^m/\alpha}\, dy = \int_0^{t^m/\alpha} e^{-u}\, du = 1 - e^{-t^m/\alpha}$$

where the substitution $u = y^m/\alpha$ was made. Substitution yields $r(t) = (m/\alpha) t^{m-1}$, an increasing function of t since $\alpha > 0$ and $m > 1$.

***E4.191 Conditional survival time.**

(a)

$$P(Y \leq y | Y \geq c) = \frac{P(Y \leq y \cap Y \geq c)}{P(Y \geq c)} = \frac{P(c \leq Y \leq y)}{P(Y \geq c)}. = \frac{F(y) - F(c)}{1 - F(c)}$$

(b) As $y \to \infty$, $F(y) \to 1$ and $P(Y \le y | Y \ge c) \to 1$. In the other limit $y \to c$, $P(Y \le y | Y \ge c) \to 0$. Furthermore $F(y)$ is nondecreasing, so $P(Y \le y | Y \ge c)$ is too.

(c) From E4.190, $F(y) = 1 - e^{-y^m/\alpha}$. Substitution yields $P(Y \le 4 | Y \ge 2) = 1 - e^{-4} \approx 0.982$. Note that this is less than the unconditional probability $P(Y \le 4) = 1 - e^{-16/3} \approx 0.995$ as we might expect because it cannot fail in the first 2 years.

***E4.193 Conditional survival time 2.** $Y \sim \text{exponential}(\beta)$ has $f(y) = e^{-y/\beta}/\beta$ and $F(y) = 1 - e^{-y/\beta}$. Thus

$$E(Y|Y \ge c) = \frac{e^{c/\beta}}{\beta} \int_c^\infty y e^{-y/\beta}\, dy = e^{c/\beta}(c + \beta) e^{-c/\beta} = c + \beta.$$

Note this is equal to $c + \mathrm{E}(Y)$, illustrating the memoryless property of the exponential distribution.

***E4.194 Integrating the standard normal density.** Let $r^2 = x^2 + y^2$ and polar angle $\phi = \arctan(y/x)$ with area element $dA = dr(r\, d\phi)$. Hence the integral becomes

$$I = \frac{1}{2\pi} \int_0^\infty r\, dr \int_0^{2\pi} d\phi\, e^{-ur^2/2} = \int_0^\infty dr\, r e^{-ur^2/2}.$$

Substituting $v = r^2$,

$$I = \frac{1}{2} \int_0^\infty dv\, e^{-uv/2} = \frac{1}{u}, \qquad u > 0.$$

***E4.195 A transformed standard normal r.v.**

(a) Expand $W = Z^4 + 6Z^3 + 9Z^2$, hence $\mathrm{E}\, W = \mathrm{E}\, Z^4 + 6\, \mathrm{E}\, Z^3 + 9\, \mathrm{E}\, Z^2 = 12$.

(b) Using the Markov inequality, $P(W \le w) \ge 1 - \mathrm{E}(W)/w = 0.9$ when $w = 120$.

***E4.198 The Markov inequality.** By definition, $\mathrm{E}(|g(Y)|) = \int_{-\infty}^\infty |g(y)| f(y)\, dy$. Restricting integration to the region $A = \{y \text{ s.t. } |g(y)| \ge k\}$ for $k > 0$, we have the inequality

$$\begin{aligned} \mathrm{E}(|g(Y)|) &\ge \int_A |g(y)| f(y)\, dy && (|g(y)| f(y) \ge 0 \text{ for all } y) \\ &\ge k \int_A f(y)\, dy && (|g(y)| \ge k \text{ in } A) \\ &= kP(A) = kP(|g(Y)| \ge k). && (\text{by definition}) \end{aligned}$$

Thus we conclude $P(|g(Y)| \ge k) \le \mathrm{E}(|g(Y)|/k$, which is equivalent to the stated result using $P(|g(Y)| \ge k) = 1 - P(|g(Y)| \le k)$. □

5 Multivariate Probability Distributions

5.1 Introduction

5.2 Bivariate and Multivariate Probability Distributions

E5.3 Marital status of random promotions. The pair (y_1, y_2) must satisfy $1 \le y_1 + y_2 \le 3$, consisting of possible points $\{(0,1),(0,2),(0,3),(1,0),(1,1),(1,2),(2,0),(2,1),(3,0)\}$. Probabilities of each point can be computed using combinatorics. There are $N = \binom{9}{3}$ ways the executives can be selected and for (y_1, y_2), there are $n(y_1, y_2) = \binom{4}{y_1}\binom{3}{y_2}\binom{2}{3-y_1-y_2}$ ways to choose y_1 married and y_2 never married executives. Thus $p(y_1, y_2) = n(y_1, y_2)/N$.

E5.5 Integrating bivariate densities.

(a) Integration is over a trapezoidal area. Splitting into two regions:

$$F(1/2, 1/3) = \int_0^{1/3}\int_0^{y_1} f(y_1, y_2)\, dy_2\, dy_1 + \int_{1/3}^{1/2}\int_0^{1/3} f(y_1, y_2)\, dy_2\, dy_1 \approx 0.106.$$

(b) $P(Y_2 \le Y_1/2) = \int_0^1 \int_0^{y_1/2} f(y_1, y_2)\, dy_2\, dy_1 = \frac{1}{2}$.

E5.6 Integrating bivariate densities 2.

(a) $P(Y_1 - Y_2 > 0.5) = P(Y_1 > Y_2 + 0.5) = \int_0^{0.5}\int_{y_2+0.5}^1 f(y_1, y_2)\, dy_1\, dy_2 = 0.125$.

(b) $P(Y_1 Y_2 < 0.5) = P(Y_1 < 0.5/Y_2) = \left[\int_0^{0.5}\int_0^1 + \int_{0.5}^1\int_0^{0.5/y_2}\right] f(y_1, y_2)\, dy_1\, dy_2 \approx 0.847$.

E5.7 Integrating bivariate densities 3.

(a) $P(Y_1 < 1, Y_2 > 5) = \int_5^\infty \int_0^1 f(y_1, y_2)\, dy_1\, dy_2 = e^{-5} - e^{-6}$.

(b) $P(Y_1 + Y_2 < 3) = P(Y_1 < 3 - Y_2) = \int_0^3 \int_0^{3-y_2} f(y_1, y_2)\, dy_1\, dy_2 = 1 - 4e^{-3}$.

E5.8 Integrating bivariate densities 4.

(a) $1 = \int_{-\infty}^\infty \int_{-\infty}^\infty f(y_1, y_2)\, dy_1\, dy_2 = \int_0^1\int_0^1 f(y_1, y_2)\, dy_1\, dy_2 = k/4$ implies $k = 4$.

(b) $F(y_1, y_2) = \int_0^{y_2}\int_0^{y_1} f(t_1, t_2)\, dt_1\, dt_2 = y_1^2 y_2^2$ for $0 \le y_1, y_2 \le 1$.

(c) $P(Y_1 \le 1/2, Y_2 \le 3/4) = F(1/2, 3/4) = 9/64$.

5.3 Marginal and Conditional Probability Distributions

E5.23 Practice with conditional densities.

(a) $f_2(y_2) = \int_{-\infty}^\infty f(y_1, y_2)\, dy_1 = \int_{y_2}^1 f(y_1, y_2)\, dy_1 = 3(1 - y_2^2)/2$, $0 \le y_2 \le 1$.

(b) $f(y_1|y_2)$ is defined where $f_2(y_2) > 0$, thus $0 \le y_2 < 1$. Furthermore, $y_2 \le y_1$.

(c) $P(Y_2 > 1/2 | Y_1 = 3/4) = \int_{1/2}^1 f(y_2 | y_1 = 3/4)\, dy_2$. Now $f(y_2|y_1) = f(y_1, y_2)/f_(y_1) = 1/y_1$ for $0 \le y_2 \le y_1 \le 1$, thus $P(Y_2 > 1/2 | Y_1 = 3/4) = \int_{1/2}^{3/4} \frac{1}{3/4}\, dy_2 = \frac{1}{3}$.

E5.27 Practice with conditional densities 2.

(a) $f_1(y_1) = \int_{y_1}^1 f(y_1, y_2)\, dy_2 = 3(y_1 - 1)^2$ for $0 \le y_1 \le 1$.
$f_2(y_2) = \int_0^{y_2} f(y_1, y_2)\, dy_1 = 6y_2(1 - y_2)$ for $0 \le y_2 \le 1$.

(b)
$$P(Y_1 \le 1/2 \mid Y_2 \le 3/4) = \frac{P(Y_1 \le 1/2 \cap Y_2 \le 3/4)}{P(Y_2 \le 3/4)} = \frac{\int_0^{1/2} dy_1 \int_{y_1}^{1/2} dy_2\, f(y_1, y_2)}{\int_0^{3/4} dy_1\, f_1(y_1)} \approx 0.508.$$

(c) $f(y_1|y_2) = f(y_1, y_2)/f(y_2) = 1/y_2$ for $0 \le y_1 \le y_2 \le 1$.

(d) $f(y_2|y_1) = f(y_1, y_2)/f(y_1) = 2(1 - y_2)/(y_1 - 1)^2$ for $0 \le y_1 \le y_2 \le 1$.

(e) $P(Y_2 \ge 3/4 \mid Y_1 = 1/2) = \int_{3/4}^1 f(y_2 \mid y_1 = 1/2)\, dy_2 = \frac{1}{4}$.

E5.38 Sold and stocked item distributions.

(a) We are given $f_1(y_1) = 1$, $0 \le y_1 \le 1$ and $f(y_2|y_1) = 1/y_1$, $0 \le y_2 \le y_1 \le 1$. Thus the joint density $f(y_1, y_2) = f(y_2|y_1) f_1(y_1) = 1/y_1$ for $0 \le y_2 \le y_1 \le 1$.

(b) $P(Y_2 \geq 0.25 \mid Y_1 = 0.5) = \int_{0.25}^{0.5} dy_2\, f(y_2|y_1 = 0.5) = 0.5$

(c) First, we need $f(y_1|y_2) = f(y_1, y_2)/f_2(y_2) = -1/(y_1 \log y_2)$, $0 \leq y_2 \leq y_1 \leq 1$, and thus $P(Y_1 \geq 0.5 \mid Y_2 = 0.25) = \int_{0.5}^{1} dy_1\, f(y_1|y_2 = 0.5) = 0.5$.

***E5.41 Hierarchical quality control plan.** Let Y denote the number of defectives. Given fixed p, the conditional distribution $(Y \mid P = p) \sim \text{binomial}(n, p)$. The unconditional distribution is obtained by integrating over p, i.e. the continuous analog of the total law of probability:

$$P(Y = y) = \int_{-\infty}^{\infty} P(Y = y \mid p) f(p)\, dp,$$

where $f(p) = 1$ for $0 \leq p \leq 1$ since $P \sim \text{uniform}(0, 1)$ and $P(Y = y \mid p) = \binom{n}{y} p^y (1-p)^{n-y}$, $y = 0, 1, \ldots, n$, yielding

$$P(Y = y) = \int_0^1 \binom{n}{y} p^y (1-p)^{n-y}\, dp = \binom{n}{y} \int_0^1 p^{(y+1)-1} (1-p)^{(n-y+1)-1}\, dp.$$

The integral is written in the form of the beta function $B(\alpha, \beta) = \int_0^1 t^{\alpha-1}(1-t)^{\beta-1}\, dt$ with $\alpha = y+1$ and $\beta = (n - y + 1)$. Hence,

$$P(Y = y) = \binom{n}{y} B(y+1, n-y+1) = \binom{n}{y} \frac{\Gamma(y+1)\Gamma(n-y+1)}{\Gamma(n+2)}, \qquad y = 0, 1, \ldots, n.$$

With $n = 3$ and $y = 2$, this evaluates to $P(Y = 2 \mid n = 3) = \binom{3}{2} 2!/4! = \frac{1}{4}$.

Comment. Justifying the first integral another way, note the joint density $f(y, p) = f(y|p) f(p)$. Integrating over p yields the marginal density $f(y)$, which is the discrete $P(Y = y)$ in this case.

***E5.42 Hierarchical model for fabric defects.** The conditional distribution of defects given λ is $(Y \mid \Lambda = \lambda) \sim \text{Poisson}(\lambda)$ where $\Lambda \sim \text{exponential}(\beta = 1)$. Similar to E5.41,

$$P(Y = y) = \int_{-\infty}^{\infty} P(Y = y \mid \lambda) f(\lambda)\, d\lambda = \int_0^{\infty} \frac{\lambda^y e^{-\lambda}}{y!} e^{-\lambda}\, d\lambda = \frac{1}{y!} \int_0^{\infty} \lambda^y e^{-2\lambda}\, d\lambda.$$

Substituting $u = 2\lambda$,

$$P(Y = y) = \frac{1}{2^{y+1} y!} \int_0^{\infty} u^y e^{-u}\, du = \frac{1}{2^{y+1} y!} \int_0^{\infty} u^{(y+1)-1} e^{-u}\, du.$$

The integral is written in the form of the gamma function $\Gamma(\alpha) = \int_0^1 t^{\alpha-1} e^{-t}\, dt$ with $\alpha = y + 1$, hence

$$P(Y = y) = \frac{\Gamma(y+1)}{2^{y+1} y!} = \frac{1}{2^{y+1}}, \qquad y = 0, 1, 2, \ldots$$

5.4 Independent Random Variables

E5.62 Independent coin tosses. Let $A \sim \text{geometric}(p)$ denote the number of tosses until the first heads for person A with pmf $P(A = a) = (1-p)^{a-1} p$, $a = 1, 2, 3, \ldots$, and similarly $B \sim \text{geometric}(p)$. By independence, $P(A = a, B = b) = P(A = a) P(B = b) = (1-p)^{a-1} p (1-p)^{b-1} p$, and the desired probability

$$\sum_{i=1}^{\infty} P(A = i, B = i) = \frac{p^2}{(1-p)^2} \sum_{i=1}^{\infty} [(1-p)^2]^i = \frac{p^2}{(1-p)^2} \left[\frac{1}{1-(1-p)^2} - 1 \right] = \frac{p^2}{1-(1-p)^2}.$$

E5.63 Independent exponential r.v. $P(Y_1 > Y_2 \mid Y_1 < 2Y_2) = P(Y_2 < Y_1 < 2Y_2)/P(Y_1 < 2Y_2)$. The probabilities can be computed using the joint density which, by independence, is $f(y_1, y_2) = e^{-(y_1+y_2)}$, $0 \le y_1, y_2, \le \infty$.

$$P(Y_1 > Y_2 \mid Y_1 < 2Y_2) = \frac{\int_0^\infty dy_2 \int_{y_2}^{2y_2} dy_1\, f(y_1, y_2)}{\int_0^\infty dy_2 \int_0^{2y_2} dy_1\, f(y_1, y_2)} = \frac{1}{4}.$$

E5.64 Independent uniform r.v. $P(Y_1 < 2Y_2 \mid Y_1 < 3Y_2) = Y_1 < 2Y_2)/P(Y_1 < 3Y_2)$. The probabilities can be computed using the joint density which, by independence, is $f(y_1, y_2) = 1$, $0 \le y_1, y_2, \le 1$.

$$P(Y_1 < 2Y_2 \mid Y_1 < 3Y_2) = \frac{\left[\int_0^{1/2} dy_2 \int_0^{2y_2} dy_1 + \int_{1/2}^1 dy_2 \int_0^1 dy_1\right] f(y_1, y_2)}{\left[\int_0^{1/3} dy_2 \int_0^{3y_2} dy_1 + \int_{1/3}^1 dy_2 \int_0^1 dy_1\right] f(y_1, y_2)} = \frac{9}{10}.$$

E5.70 Waiting for a random bus. Let $Y_1 \sim \text{uniform}(0, 1)$ be the time the bus arrives (in hours) and $Y_2 \sim \text{uniform}(0, 1)$ be the time the passenger arrives. The joint distribution is $f(y_1, y_2) = 1$, $0 \le y_1, y_2 \le 1$. We are interested in $P(0 \le Y_1 - Y_2 \le 0.25)$, which is equivalent to

$$P(Y_2 \le Y_1 \le Y_2 + 0.25) = \left[\int_0^{0.75} dy_2 \int_{y_2}^{y_2+0.25} dy_1 + \int_{0.75}^1 dy_2 \int_{y_2}^1 dy_1\right] f(y_1, y_2) = \frac{7}{32}.$$

5.5 The Expected Value of a Function of Random Variables

5.6 Special Theorems

E5.88 Dice tosses until all faces are observed. Per the hint, $\mathrm{E}(Y) = \mathrm{E}(\sum_{i=1}^6 Y_i) = \sum_{i=1}^6 \mathrm{E}\, Y_i = \sum_{i=1}^6 1/p_i = 14.7$.

5.7 The Covariance of Two Random Variables

E5.94 Covariance of sum and difference of uncorrelated r.v.

(a)
$$\begin{aligned}\operatorname{cov}(U_1, U_2) &= \mathrm{E}(U_1U_2) - \mathrm{E}(U_1)\,\mathrm{E}(U_2) = \mathrm{E}[(Y_1 + Y_2)(Y_1 - Y_2)] - \mathrm{E}(Y_1 + Y_2)\,\mathrm{E}(Y_1 - Y_2)\\ &= [\mathrm{E}\, Y_1^2 - \mathrm{E}\, Y_2^2] - [(\mathrm{E}\, Y_1)^2 - (\mathrm{E}\, Y_2)^2] = \operatorname{var} Y_1 - \operatorname{var} Y_2.\end{aligned}$$

(b) $\rho = \operatorname{cov}(U_1, U_2)/\sqrt{\operatorname{var} U_1 \operatorname{var} U_2}$. In general, $\operatorname{var}(X \pm Y) = \operatorname{var} X + \operatorname{var} Y \pm 2\operatorname{cov}(X, Y)$. Since Y_1 and Y_2 are uncorrelated, it follows that $\operatorname{var} U_1 = \operatorname{var} U_2 = \operatorname{var} Y_1 + \operatorname{var} Y_2$, and $\rho = (\operatorname{var} Y_1 - \operatorname{var} Y_2)/(\operatorname{var} Y_1 + \operatorname{var} Y_2)$.

(c) Yes, when $\operatorname{var} Y_1 = \operatorname{var} Y_2$.

E5.95 No correlation does not imply independence. $\operatorname{cov}(Y_1, Y_2) = \mathrm{E}(Y_1Y_2) - \mathrm{E}(Y_1)\,\mathrm{E}(Y_2)$. From the joint pmf, $\mathrm{E}(Y_1Y_2) = 0$. The marginal pmfs are $p_1(y_1) = \frac{1}{3}$ for $y_1 = -1, 0, 1$ and $p_2(y_2 = 0) = \frac{2}{3}$ and $p_2(y_2 = 1) = \frac{1}{3}$. It follows that $\mathrm{E}(Y_1) = 0$ and $\mathrm{E}(Y_2) = \frac{1}{3}$. Thus $\operatorname{cov}(Y_1, Y_2) = 0$. However, Y_1 and Y_2 are not independent. For instance, $\frac{1}{3} = p(-1, 0) \ne p_1(-1)p_2(0) = \frac{2}{9}$.

5.8 The Expected Value and Variance of Linear Functions of Random Variables

E5.115 Dice tosses until all faces are observed 2.

(a) From the memoryless property of the geometric distribution, $P(Y_i = y_i \mid Y_j = y_j) = P(Y_i = y_i)$ for $i \ne j$, i.e. Y_i and Y_j are independent. It follows that $\operatorname{cov}(Y_i, Y_j) = 0$.

(b) By independence, $\operatorname{var}(Y) = \operatorname{var}(\sum_{i=1}^6 Y_i) = \sum_{i=1}^6 \operatorname{var} Y_i = \sum_{i=1}^6 (1 - p_i)/p_i^2 = 38.99$.

(c) By Tchebysheff's theorem, $P(|Y - \mu| \le k\sigma) \le 1 - 1/k^2$. $k = 2$ provides a 3/4 upper bound.

***E5.117 Proportion of sexually mature male and female alligators.**

$$\mathrm{E}\left[\frac{1}{n}(Y_1 - Y_2)\right] = \frac{1}{n}\left(\mathrm{E}\,Y_1 - \mathrm{E}\,Y_2\right) \quad \text{and} \quad \operatorname{var}\left[\frac{1}{n}(Y_1 - Y_2)\right] = \frac{1}{n^2}\left[\operatorname{var} Y_1 + \operatorname{var} Y_2 - 2\operatorname{cov}(Y_1, Y_2)\right].$$

One approach is to compute the joint pmf $P(Y_1 = y_1, Y_2 = y_2)$, sum to obtain the marginal pmfs, and directly compute the expectations and (co)variances. There are $p_1 N$ mature males, $p_2 N$ mature females, and $(1 - p_1 - p_2)N$ other alligators. There are $\binom{N}{n}$ total ways a sample could be drawn. For a sample to have $Y_1 = y_1$ and $Y_2 = y_2$, there are $\binom{p_1N}{y_1}\binom{p_2N}{y_2}\binom{(1-p_1-p_2)N}{n-y_1-y_2}$ ways. Hence,

$$P(Y_1 = y_1, Y_2 = y_2) = \binom{p_1N}{y_1}\binom{p_2N}{y_2}\binom{(1-p_1-p_2)N}{n-y_1-y_2}\bigg/\binom{N}{n}.$$

This is a multivariate hypergeometric distribution. The marginal distributions can be shown to be hypergeometric with $r_i = p_i N$, using the same combinatorial reasoning.

$$P(Y_i = y_i) = \binom{p_iN}{y_i}\binom{(1-p_i)N}{n-y_i}\bigg/\binom{N}{n}.$$

It follows that $\mathrm{E}\,Y_i = np_i$ similarly for $\operatorname{var} Y_i$ (Theorem 3.10). The covariance, however, requires double summation of the joint pmf. A simpler approach is analogous to Example 5.29.

Suppose we observe the set of outcomes $(M_1, F_1), (M_2, F_2), \ldots, (M_n, F_n)$ where

$$M_i = \begin{cases} 1 & \text{if } i\text{-th draw is a mature male} \\ 0 & \text{otherwise} \end{cases}, \qquad F_i = \begin{cases} 1 & \text{if } i\text{-th draw is a mature female} \\ 0 & \text{otherwise} \end{cases}.$$

Note that $Y_1 = \sum_{i=1}^n M_i$ and $Y_2 = \sum_{i=1}^n F_i$. We can then apply Theorem 5.12 using the simpler Bernoulli base events M_i and F_i. In particular, $P(M_i = 1) = p_1$ and $P(F_i = 1) = p_2$ for all $i = 1, 2, \ldots, n$, thus $\mathrm{E}(M_i) = p_1$ and $\mathrm{E}(F_i) = p_2$, and $\operatorname{var}(M_i) = p_1(1-p_1)$ and $\operatorname{var}(F_i) = p_2(1-p_2)$. This allows us to compute $\mathrm{E}\,Y_1 = \sum_{i=1}^n \mathrm{E}\,M_i = np_1$ and similarly $\mathrm{E}\,Y_2 = np_2$. Therefore

$$\mathrm{E}\left[(Y_1 - Y_2)/n\right] = p_1 - p_2.$$

The variance calculation is more involved. First, we need

$$\operatorname{var} Y_1 = \sum_{i=1}^n \operatorname{var} M_i + 2 \sum\sum_{1 \le i < j \le n} \operatorname{cov}(M_i, M_j).$$

The covariance computation requires the joint distribution

$$P(M_i = 1, M_j = 1) = P(M_i = 1 | M_j = 1)P(M_j = 1) = \frac{p_1N - 1}{N - 1} \times p_1, \quad i \ne j.$$

Thus $\operatorname{cov}(M_i, M_j) = \mathrm{E}(M_iM_j) - \mathrm{E}(M_i)\,\mathrm{E}(M_j) = P(M_i = 1, M_j = 1) - p_1^2 = p_1(p_1N - 1)/(N - 1)$, $i \ne j$, and

$$\operatorname{var} Y_1 = n \operatorname{var} M_i + n(n-1)\operatorname{cov}(M_i, M_j) = np_1(1 - p_1)\frac{N - n}{N - 1}.$$

Similarly, $\operatorname{var} Y_2 = np_2(1 - p_2)(N - n)/(N - 1)$. All that remains is

$$\operatorname{cov}(Y_1, Y_2) = \sum_{i=1}^n \sum_{j=1}^n \operatorname{cov}(M_i, F_j).$$

We have

$$P(M_i = 1, F_j = 1) = P(M_i = 1|F_j = 1)P(F_j = 1) = \frac{p_1 N}{N-1} \times p_2, \quad i \neq j,$$

and equal to zero for $i = j$. Therefore $\operatorname{cov}(M_i, F_j) = p_1 p_2 N/(N-1) - p_1 p_2$ for $i \neq j$ and $\operatorname{cov}(M_i, F_j) = 0 - p_1 p_2$ for $i = j$. Substitution yields

$$\operatorname{cov}(Y_1, Y_2) = \sum_{i \neq j} \operatorname{cov}(M_i, F_j) + \sum_{i=j} \operatorname{cov}(M_i, F_j) = n(n-1)\operatorname{cov}(M_k, F_l) + n\operatorname{cov}(M_k, F_k), \quad k \neq l.$$

This simplifies to

$$\operatorname{cov}(Y_1, Y_2) = -np_1p_2\left(\frac{N-n}{N-1}\right).$$

Putting it all together,

$$\begin{aligned}
\operatorname{var}\left[\frac{1}{n}(Y_1 - Y_2)\right] &= \frac{1}{n^2}\left[\operatorname{var} Y_1 + \operatorname{var} Y_2 - 2\operatorname{cov}(Y_1, Y_2)\right] \\
&= \frac{1}{n}\left(\frac{N-n}{N-1}\right)\left[p_1(1-p_1) + p_2(1-p_2) - 2p_1p_2\right].
\end{aligned}$$

5.9 The Multinomial Probability Distribution

5.10 The Bivariate Normal Distribution

***E5.130 Linear combinations of i.i.d. normal r.v.**

(a) Applying Theorem 5.12,

$$\operatorname{cov}(U_1, U_2) = \sum_{i=1}^{n}\sum_{j=1}^{n} a_i b_j \operatorname{cov}(Y_i, Y_j) = \sum_{i=1}^{n} a_i b_i \operatorname{cov}(Y_i, Y_i) + \sum_{i \neq j} a_i b_j \operatorname{cov}(Y_i, Y_j).$$

Since Y_i are i.i.d., $\operatorname{cov}(Y_i, Y_i) = \operatorname{var}(Y_i) = \sigma^2$ for all i and $\operatorname{cov}(Y_i, Y_j) = 0$ for all $i \neq j$. Thus

$$\operatorname{cov}(U_1, U_2) = \sigma^2 \sum_{i=1}^{n} a_i b_i,$$

which is only zero if $\sum_{i=1}^{n} a_i b_i = 0$ since $\sigma^2 > 0$.

(b) Assuming U_1 and U_2 have a bivariate normal distribution, it is sufficient to show independence of the marginals if their covariance $\operatorname{cov}(U_1, U_2) = 0$. The result follows.

***E5.131 Sums of two i.i.d. normal r.v.**

(a) By independence,

$$\begin{aligned}
f(y_1, y_2) = f_1(y_1)f_2(y_2) &= \left[\frac{1}{\sigma\sqrt{2\pi}} e^{-(y_1-\mu_1)^2/(2\sigma^2)}\right]\left[\frac{1}{\sigma\sqrt{2\pi}} e^{-(y_2-\mu_2)^2/(2\sigma^2)}\right] \\
&= \frac{1}{2\pi\sigma^2} e^{-[(y_1-\mu_1)^2 + (y_2-\mu_2)^2]/(2\sigma^2)},
\end{aligned}$$

which is bivariate normal with $\rho = 0$.

(b) We can express U_i as linear combinations of Y_i with coefficients $a_1 = 1$, $a_2 = 1$, $b_1 = 1$, and $b_2 = -1$. From E5.130(b), U_1 and U_2 have a bivariate normal distribution. Furthermore, $\sum_{i=1}^{2} a_i b_i = 0$, so U_1 and U_2 are orthogonal and thus independent.

5.11 Conditional Expectations

E5.135 Hierarchical quality control plan 2. Let Y denote the number of defectives. The hierarchical model is $Y|P \sim \text{binomial}(n, P)$ where $P \sim \text{uniform}(a, b)$.

(a) $\mathrm{E}\,Y = \mathrm{E}[\mathrm{E}(Y|P)] = \mathrm{E}[nP] = n\,\mathrm{E}\,P = n(b-a)/2 = \frac{3}{2}$.

(b) $\operatorname{var} Y = \mathrm{E}[\operatorname{var}(Y|P)] + \operatorname{var}[\mathrm{E}(Y|P)] = \mathrm{E}[nP(1-P)] + \operatorname{var}[nP] = n\,\mathrm{E}\,P - n\,\mathrm{E}\,P^2 + n^2 \operatorname{var} P$. Substituting $\mathrm{E}\,P = (b-a)/2$, $\operatorname{var} P = (b-a)^2/12$ and $\mathrm{E}\,P^2 = \operatorname{var} P + (\mathrm{E}\,P)^2$ yields $\operatorname{var} Y = \frac{5}{4}$.

E5.136 Hierarchical model for fabric defects 2. The hierarchical model is $Y|\Lambda \sim \text{Poisson}(\Lambda)$ where $\Lambda \sim \text{exponential}(\beta)$.

(a) $\mathrm{E}\,Y = \mathrm{E}[\mathrm{E}(Y|\Lambda)] = \mathrm{E}[\Lambda] = \beta = 1$.

(b) $\operatorname{var} Y = \mathrm{E}[\operatorname{var}(Y|\Lambda)] + \operatorname{var}[\mathrm{E}(Y|\Lambda)] = \mathrm{E}[\Lambda] + \operatorname{var}[\Lambda] = \beta + \beta^2 = 2$.

(c) By Tchebysheff's theorem, $P(|Y-\mu| \geq k\sigma) \leq 1/k^2$. $Y = 9$ is $k = (9-1)/\sqrt{2}\,\sigma$ away from the mean. Thus, the probability ≤ 0.03125; it is not likely.

E5.138 Hierarchical model for bacteria. $Y|\Lambda \sim \text{Poisson}(\Lambda)$ where $\Lambda \sim \text{gamma}(\alpha, \beta)$.

(a) $\mathrm{E}\,Y = \mathrm{E}[\mathrm{E}(Y|\Lambda)] = \mathrm{E}[\Lambda] = \alpha\beta$.

(b) $\operatorname{var} Y = \mathrm{E}[\operatorname{var}(Y|\Lambda)] + \operatorname{var}[\mathrm{E}(Y|\Lambda)] = \mathrm{E}[\Lambda] + \operatorname{var}[\Lambda] = \alpha\beta + \alpha\beta^2$. $\sigma = \sqrt{\operatorname{var} Y}$.

E5.139 Model for jobs completion time. $N \sim \text{Poisson}(\lambda)$, $Y_i \sim \text{gamma}(\alpha, \beta)$.

(a) $\mathrm{E}(T \mid N = n) = \mathrm{E} \sum_{i=1}^{n} Y_i = n\alpha\beta$.

(b) $\mathrm{E}\,T = \mathrm{E}[\mathrm{E}(T|N)] = \mathrm{E}[N\alpha\beta] = \alpha\beta\,\mathrm{E}(N) = \lambda\alpha\beta$.

5.12 Summary

Supplementary Exercises

E5.145 Probability of job completion. In minutes from 8 a.m., start time $Y_1 \sim \text{uniform}(0, 15)$. The amount of time the job takes $Y_2 \sim \text{uniform}(20, 30)$. We are interested in $P(Y_1 + Y_2 \leq 30)$. By independence, joint distribution is $f(y_1, y_2) = \frac{1}{150}$ for $0 \leq y_1 \leq 15$ and $20 \leq y_2 \leq 30$, and zero elsewhere. Integrating over the region $y_1 + y_2 \leq 30$ or equivalently $y_1 \leq 30 - y_2$,

$$P(Y_1 + Y_2 \leq 30) = P(Y_1 \leq 30 - Y_2) = \int_{20}^{30} dy_2 \int_{0}^{30-y_2} dy_1\, f(y_1, y_2) = \frac{1}{3}.$$

E5.146 Destruction probability from a bomb. Let R and Θ denote the polar position of the bomb. Then $R \sim \text{uniform}(0, 1)$ and $\Theta \sim \text{uniform}(0, 2\pi)$ are independent, with joint pdf $f(r, \theta) = 1/2\pi$ for $0 \leq r \leq 1$ and $0 \leq \theta \leq 2\pi$. We are interested in

$$P(0 \leq r \leq 0.5, 0 \leq \theta \leq 2\pi) = \int_{0}^{0.5} dr \int_{0}^{2\pi} r d\theta\, f(r, \theta) = \frac{1}{4}.$$

E5.147 Pair meetup probability. Let $Y_i \sim \text{uniform}(0, 60)$ denote the arrival time in minutes for friend i. The two friends meet if $|Y_2 - Y_1| < 10$. Assuming independence, $f(y_1, y_2) = 1/60^2$, $0 \leq y_1, y_2, \leq 60$. Integration can be split into three regions as follows:

$$P(|Y_2 - Y_1| < 10) = \left[\int_{0}^{10} dy_2 \int_{0}^{y_2+10} dy_1 + \int_{10}^{50} dy_2 \int_{y_2-10}^{y_2+10} dy_1 + \int_{50}^{60} dy_2 \int_{y_2-10}^{60} dy_1 \right] f(y_1, y_2) = \frac{11}{36}.$$

E5.148 Random political committee.

(a) Through combinatorial arguments akin to E5.3, $P(Y_1 = y_1, Y_2 = y_2) = \binom{4}{y_1}\binom{3}{y_2}\binom{2}{3-y_1-y_2}/\binom{9}{3}$ where $y_i \in \{0, 1, 2, 3\}$ subject to the constraint $1 \leq y_1 + y_2 \leq 3$.

(b) We could sum the joint pmf to obtain the marginals, but combinatorial arguments are easier: $P(Y_1 = y_1) = \binom{4}{y_1}\binom{5}{3-y_1}/\binom{9}{3}$, $y_1 = 0, 1, 2, 3$, and $P(Y_2 = y_2) = \binom{3}{y_2}\binom{6}{3-y_2}/\binom{9}{3}$, $y_2 = 0, 1, 2, 3$.

(c) $P(Y_1 = 1 \mid Y_2 \geq 1) = P(Y_1 = 1 \cap Y_2 \geq 1)/P(Y_2 \geq 1) = \sum_{i=1}^{2} p(1,i)/\sum_{j=1}^{3} p_2(j) = \frac{9}{16}$.

E5.149 Practice with bivariate densities.

(a) $f_1(y_1) = \int_0^{y_1} dy_2\, f(y_1,y_2) = 3y_1^2$, $0 \leq y_1 \leq 1$.
$f_2(y_2) = \int_{y_2}^{1} dy_1\, f(y_1,y_2) = (3/2)(1-y_2^2)$, $0 \leq y_2 \leq 1$.

(b)

$$P(Y_1 \leq 3/4 \mid Y_2 \leq 1/2) = \frac{P(Y_1 \leq 3/4, Y_2 \leq 1/2)}{P(Y_2 \leq 1/2)} = \frac{\int_0^{1/2} dy_2 \int_{y_2}^{3/4} dy_1\, f(y_1,y_2)}{\int_0^{1/2} dy_2\, f_2(y_2)} = \frac{23}{44}.$$

(c) $f(y_1|y_2) = f(y_1,y_2)/f_2(y_2) = 2y_1/(1-y_2^2)$ for $0 \leq y_2 \leq y_1 \leq 1$.

(d) $P(Y_1 \leq 3/4 \mid Y_2 = 1/2) = \int_{1/2}^{3/4} dy_1\, f(y_1|y_2 = 1/2) = \frac{5}{12}$.

E5.150 Practice with bivariate densities 2.

(a) $f(y_2|y_1) = f(y_1,y_2)/f_2(y_1) = 1/y_1$, $0 \leq y_2 \leq y_1 \leq 1$.
$\mathrm{E}(Y_2 \mid Y_1 = y_1) = \int_0^{y_1} dy_2\, y_2 f(y_2|y_1) = y_1/2$.

(b) $\mathrm{E}\, Y_2 = \mathrm{E}[\mathrm{E}(Y_2|Y_1)] = \mathrm{E}[Y_1/2] = \frac{1}{2}\int_0^1 dy_1\, y_1 f_1(y_1) = \frac{3}{8}$.

(c) $\mathrm{E}\, Y_2 = \int_0^1 dy_2\, y_2 f_2(y_2) = \frac{3}{8}$.

E5.153 Model for egg hatches. Let N be the number of laid eggs, $N \sim \text{Poisson}(\lambda)$. Let Y be the number of hatched eggs. Assuming eggs hatch independently, each with probability p, the conditional distribution $Y|N \sim \text{binomial}(N,p)$. The next two parts apply Theorem 5.14 and 5.15.

(a) $\mathrm{E}\, Y = \mathrm{E}[\mathrm{E}(Y|N)] = \mathrm{E}(Np) = p\,\mathrm{E}\, N = p\lambda$

(b) $\mathrm{var}\, Y = \mathrm{E}[\mathrm{var}(Y|N)] + \mathrm{var}[\mathrm{E}(Y|N)] = \mathrm{E}[Np(1-p)] + \mathrm{var}[Np] = p(1-p)\lambda + p^2\lambda = p\lambda$.

Comment. $Y \sim \text{Poisson}(p\lambda)$. See E3.202 for a derivation of the pmf.

E5.154 Patient response to a drug.

(a) The conditional distribution $(Y \mid P) \sim \text{binomial}(n,P)$, with the distribution of P given. The unconditional pmf is therefore $f_Y(y) = \int_0^1 dp\, f(y|p) f_P(p) = \int_0^1 dp\, \binom{n}{y} p^y (1-p)^{n-y} \times 12p^2(1-p)$. The integral can be evaluated using the beta function as in E5.41. The result:

$$f_Y(y) = 12\binom{n}{y}\frac{\Gamma(y+3)\Gamma(n-y+2)}{\Gamma(n+5)}, \qquad y = 0,1,\ldots,n.$$

(b) We could sum the pmf in (a), but using Theorem 5.14 is much easier in general.
$\mathrm{E}\, Y = \mathrm{E}[\mathrm{E}(Y|P)] = \mathrm{E}[nP] = n\,\mathrm{E}\, P = n\int_0^1 dp\, p f(p) = 3n/5$.

E5.158 Sum of i.i.d. Bernoulli trials. It is easy to show $\mathrm{E}\, X_i = p$ and $\mathrm{var}\, X_i = p(1-p)$. It follows that $\mathrm{E}\, Y = \mathrm{E}\sum_{i=1}^n X_i = \sum_{i=1}^n \mathrm{E}\, X_i = np$ and, by independence, $\mathrm{var}\, Y = \sum_{i=1}^n \mathrm{var}\, X_i = np(1-p)$.

***E5.159 Sum of i.i.d. geometric distributions.** Consider $W_1 = X_1$, i.e. the number of trials until the first success. It follows that $W_1 \sim \text{geometric}(p)$. Next, consider $W_2 = X_2 - X_1$, i.e. the number of trials until the second success, give that X_1 trials have occurred. Because of the memoryless property (e.g. E3.71), $P(Y > X_2|Y > X_1) = P(Y > W_2 + X_1|Y > X_1) = P(Y > W_2)$, it follows that $W_2 \sim \text{geometric}(p)$ independent of W_1. Continuing this reasoning, we deduce $W_i \overset{\text{i.i.d.}}{\sim} \text{geometric}(p)$.

Since $\mathrm{E}\, W_i = 1/p$ and $\mathrm{var}\, W_i = (1-p)/p^2$, analogous to E5.158, it follows that $\mathrm{E}\, Y = r/p$ and $\mathrm{var}\, Y = r(1-p)/p^2$

E5.161 Difference in normal sample means. $\mathrm{E}(\overline{Y}) = \mathrm{E}\left(\frac{1}{n}\sum_{i=1}^n Y_i\right) = \frac{1}{n}\sum_{i=1}^n \mathrm{E}(Y_i) = \mu_Y$ and similarly $\mathrm{E}(\overline{Y}) = \mu_X$. Thus $\mathrm{E}(\overline{Y} - \overline{X}) == \mathrm{E}(\overline{Y}) - \mathrm{E}(\overline{X}) = \mu_Y - \mu_X$. Similarly by independence of samples and within each sample, $\mathrm{var}(\overline{Y} - \overline{X}) = \mathrm{var}(\overline{Y}) + \mathrm{var}(\overline{X}) = \sigma_Y^2/n + \sigma_X^2/m$.

***E5.164 Joint moment-generating function.**

(a, b) $m(t,t,t) = \mathrm{E}\, e^{(X_1+X_2+X_3)t}$, $m(t,t,0) = \mathrm{E}\, e^{(X_1+X_2)t}$.

(c) Let $\mathbf{x} = (x_1, x_2, x_3)$ and $\int d\mathbf{x} = \int_{-\infty}^{\infty} dx_1 \int_{-\infty}^{\infty} dx_2 \int_{-\infty}^{\infty} dx_3$. By definition,

$$m(t_1, t_2, t_3) = \int d\mathbf{x}\, e^{t_1x_1+t_2x_2+t_3x_3} f(\mathbf{x}).$$

Differentiating, assuming we can interchange differentiation with integration,

$$\begin{aligned}\frac{\partial^{k_1+k_2+k_3} m(t_1,t_2,t_3)}{\partial t_1^{k_1}\partial t_2^{k_2}\partial t_3^{k_3}} &= \int d\mathbf{x}\, \frac{\partial^{k_1+k_2+k_3}}{\partial t_1^{k_1}\partial t_2^{k_2}\partial t_3^{k_3}} e^{t_1x_1+t_2x_2+t_3x_3} f(\mathbf{x}) \\ &= \int d\mathbf{x}\, x_1^{k_1}x_2^{k_2}x_3^{k_3} e^{t_1x_1+t_2x_2+t_3x_3} f(\mathbf{x}).\end{aligned}$$

Evaluated at $t_1 = t_2 = t_3 = 0$,

$$\left.\frac{\partial^{k_1+k_2+k_3} m(t_1,t_2,t_3)}{\partial t_1^{k_1}\partial t_2^{k_2}\partial t_3^{k_3}}\right|_{t_1=t_2=t_3=0} = \int d\mathbf{x}\, x_1^{k_1}x_2^{k_2}x_3^{k_3} f(\mathbf{x}) = \mathrm{E}(X_1^{k_1}X_2^{k_2}X_3^{k_3}). \qquad \square$$

Comment. A proof for (c) using Taylor series analogous to Theorem 3.12 is as follows.

$$\begin{aligned} m(t_1,t_2,t_3) &= \int d\mathbf{x}\, e^{t_1x_1+t_2x_2+t_3x_3} f(\mathbf{x}) \\ &= \int d\mathbf{x} \left\{ \sum_{i=0}^{\infty} \frac{1}{i!}(t_1x_1+t_2x_2+t_3x_3)^i \right\} f(\mathbf{x}) && \text{(series expansion)} \\ &= \int d\mathbf{x} \left\{ \sum_{i=0}^{\infty} \frac{1}{i!} \sum_{j_1+j_2+j_3=i} \binom{i}{j_1, j_2, j_3} (t_1x_1)^{j_1}(t_2x_2)^{j_2}(t_3x_3)^{j_3} \right\} f(\mathbf{x}) \\ & && \text{(multinomial expansion)} \\ &= \int d\mathbf{x} \left\{ \sum_{i=0}^{\infty} \sum_{j_1+j_2+j_3=i} \frac{(t_1x_1)^{j_1}(t_2x_2)^{j_2}(t_3x_3)^{j_3}}{j_1!j_2!j_3!} \right\} f(\mathbf{x}). && \text{(simplification)} \end{aligned}$$

Differentiating, assuming we can interchange differentiation with integration/summation,

$$\frac{\partial^{k_1+k_2+k_3} m(t_1,t_2,t_3)}{\partial t_1^{k_1}\partial t_2^{k_2}\partial t_3^{k_3}} = \int d\mathbf{x} \left\{ \sum_{i=0}^{\infty} \sum_{j_1+j_2+j_3=i} \frac{\partial^{k_1+k_2+k_3}}{\partial t_1^{k_1}\partial t_2^{k_2}\partial t_3^{k_3}} \frac{(t_1x_1)^{j_1}(t_2x_2)^{j_2}(t_3x_3)^{j_3}}{j_1!j_2!j_3!} \right\} f(\mathbf{x}).$$

Anticipating substitution of $t_l = 0$, $l = 1, 2, 3$, the only surviving term must have $j_l = k_l$, which occurs in only one term in the double sum over i, j_l. The t_l derivatives evaluate to $k_l! x_l^{k_l}$ and since $j_l = k_l$, the factorials cancel. Thus, we are left with

$$\left.\frac{\partial^{k_1+k_2+k_3} m(t_1,t_2,t_3)}{\partial t_1^{k_1}\partial t_2^{k_2}\partial t_3^{k_3}}\right|_{t_1=t_2=t_3=0} = \int d\mathbf{x}\, x_1^{k_1}x_2^{k_2}x_3^{k_3} f(\mathbf{x}) = \mathrm{E}(X_1^{k_1}X_2^{k_2}X_3^{k_3}). \qquad \square$$

***E5.165 Multinomial joint mgf.**

(a)

$$\begin{aligned} m(t_1,t_2,t_3) &= \sum_{x_1+x_2+x_3=n} e^{t_1x_1+t_2x_2+t_3x_3} \frac{n!}{x_1!x_2!x_3!} p_1^{x_1}p_2^{x_2}p_3^{x_3} && \text{(definition)} \\ &= \sum_{x_1+x_2+x_3=n} \binom{n}{x_1, x_2, x_3} (p_1e^{t_1})^{x_1}(p_2e^{t_2})^{x_2}(p_3e^{t_3})^{x_3} && \text{(simplification)} \\ &= \left(p_1e^{t_1} + p_2e^{t_2} + p_3e^{t_3}\right)^n. && \text{(multinomial expansion)} \end{aligned}$$

(b) $m_{X_1}(t) = m(t_1 = t, t_2 = 0, t_3 = 0) = (p_1 e^t + p_2 + p_3)^n = [p_1 e^t + (1 - p_1)]^n$, which is a binomial($p_1$) mgf (e.g., see E3.145.)

(c) $\operatorname{cov}(X_1, X_2) = \mathrm{E}(X_2 X_1) - (\mathrm{E}\, X_2)(\mathrm{E}\, X_1)$. Since $X_i \sim$ binomial(p_i), $\mathrm{E}\, X_i = np_i$. Using the joint mgf with $k_1 = k_2 = 1$ and $k_3 = 0$, to obtain $\mathrm{E}(X_2 X_1)$,

$$\frac{\partial^2 m(t_1, t_2, t_3)}{\partial t_1 \partial t_2} = n(n-1)p_1 p_2 \left(p_1 e^{t_1} + p_2 e^{t_2} + p_3 e^{t_3}\right)^{n-2} e^{t_1 + t_2}.$$

Evaluated at $t_1 = t_2 = t_3 = 0$ yields $\mathrm{E}(X_2 X_1) = n(n-1)p_1 p_2$. Thus $\operatorname{cov}(X_1, X_2) = -np_1 p_2$.

***E5.166 Multivariate hypergeometric distribution.** This problem is analogous to E5.117, where we found $\operatorname{cov}(Y_1, Y_2) = -np_1 p_2 (N - n)/(N - 1)$ where $p_i = N_i / N$. We also showed the marginal distributions are hypergeometric with $\operatorname{var} Y_i = np_i(1 - p_i)(N - n)/(N - 1)$. Thus, the correlation $\rho = \operatorname{cov}(Y_1, Y_2)/(\sigma_1 \sigma_2) = -\sqrt{p_1 p_2 / [(1 - p_1)(1 - p_2)]}$.

---

6 Functions of Random Variables

6.1 Introduction

6.2 Finding the Probability Distribution of a Function of Random Variables

6.3 The Method of Distribution Functions

E6.5 Cost of a random delay. $Y \sim$ uniform$(1,5)$ with pdf $f_Y(y) = \frac{1}{4}$, $1 \le y \le 5$. Note $5 \le u \le 53$. $P(U \le u) = P(2Y^2 + 3 \le u) = P(-\sqrt{(u-3)/2} \le Y \le \sqrt{(u-3)/2}) = P(1 \le Y \le \sqrt{(u-3)/2})$ where the last step follows as $u \ge 5$. Integrating, $F_U(u) = \int_1^{\sqrt{(u-3)/2}} f_Y(y)\,dy = \frac{1}{4}[\sqrt{(u-3)/2} - 1]$ and differentiating, $f_U(u) = 1/[16\sqrt{(u-3)/2}]$, $5 \le u \le 53$.

E6.7 Square of a standard normal r.v. $f_Z(z) = e^{-z^2/2}/\sqrt{2\pi}$, $-\infty < z < \infty$. U has support $0 \le u < \infty$. $F_U(u) = P(U \le u) = P(Z^2 \le u) = P(-\sqrt{u} \le Z \le \sqrt{u}) = F_Y(\sqrt{u}) - F_Y(-\sqrt{u})$, thus $f_U(u) = [f_Y(\sqrt{u}) - f_Y(-\sqrt{u})]/(2\sqrt{u}) = e^{-u/2}/\sqrt{2\pi u}$, $u > 0$. Thus $U \sim$ gamma$(\alpha = \frac{1}{2}, \beta = 2)$, i.e. a χ^2 distribution with 1 d.f.

E6.8 A related beta density. $Y \sim$ beta(α,β) has density $f_Y(y) = y^{\alpha-1}(1-y)^{\beta-1}/B(\alpha,\beta)$, $0 \le y \le 1$. U has support $0 \le u \le 1$. $F_U(u) = P(U \le u) = P(1 - Y \le u) = P(Y \ge 1 - u) = 1 - F_Y(1-u)$, thus $f_U(u) = f_Y(1-u)$, $0 \le u \le 1$, which is beta(β,α) density (i.e. parameters swapped).

E6.12 A related gamma density. $Y \sim$ gamma(α,β) has density $f_Y(y) = y^{\alpha-1}e^{-y/\beta}/[\beta^\alpha\Gamma(\alpha,\beta)]$, $y \ge 0$. U has support $u \ge 0$. $F_U(u) = P(U \le u) = P(cY \le u) = P(Y \le u/c) = F_Y(u/c)$, thus $f_U(u) = f_Y(u/c)/c$, $u \ge 0$, which is gamma$(\alpha, c\beta)$ density.

E6.13 Sum of i.i.d. exponential r.v. By independence, $f(y_1,y_2) = e^{-(y_1+y_2)/\beta}/\beta^2$, $y_1, y_2 \ge 0$. U has support $u \ge 0$. $F_U(u) = P(U \le u) = P(Y_1 + Y_2 \le u) = P(Y_1 \le u_2 - Y_2) = \int_0^u dy_2 \int_0^{u-y_2} dy_1\, f(y_1,y_2)$. Integration yields $F_U(u) = 1 - e^{-u/\beta} - ue^{-u/\beta}/\beta$, thus $f_U(u) = ue^{-u/\beta}/\beta^2$, $u \ge 0$.

Comment. $U \sim$ gamma$(2,\beta)$.

E6.14 Product distribution. By independence, $f(y_1,y_2) = 18y_1(1-y_1)y_2^2$, $0 \le y_1, y_2 \le 1$. U has support $0 \le u \le 1$. $F_U(u) = P(Y_1Y_2 \le u) = P(Y_1 \le u/Y_2) = [\int_0^u dy_2 \int_0^1 dy_1 + \int_u^1 dy_2 \int_0^{u/y_2} dy_1] f(y_1,y_2)$. Integration yields $F_U(u) = 6u^2 \log u - 8u^3 + 9u^2$, thus $f_U(u) = 18u[u(\log u - 1) + 1]$, $0 \le u \le 1$.

E6.17 Power family of distributions.

(a) $f_Y(y) = \alpha y^{\alpha-1}/\theta^\alpha$, $0 \le y \le \theta$, and zero elsewhere.

(b) $U = F_Y(Y)$ has a uniform distribution with inverse $G(U) = F_Y^{-1}(U) = \theta U^{1/\alpha}$.

(c) Denote the uniform values as u_i, $i = 1, 2, \ldots, 5$. The power distribution values are $G(u_i)$.

E6.18 Pareto family of distributions.

(a) $f_Y(y) = \alpha(\beta/y)^\alpha/y$, $y \ge \beta$, and zero elsewhere.

(b) $U = F_Y(Y)$ has a uniform distribution with inverse $G(U) = F_Y^{-1}(U) = \beta(1-u)^{-1/\alpha}$.

(c) Denote the uniform values as u_i, $i = 1, 2, \ldots, 5$. The Pareto distribution values are $G(u_i)$.

E6.19 Pareto and power distribution relation. X has support $0 \le x \le 1/\beta$ and distribution function $F_X(x) = P(X \le x) = P(1/Y \le x) = P(Y \ge 1/x) = 1 - P(Y \le 1/x) = 1 - F_Y(1/x) = (\beta x)^\alpha$, which is power distributed with $\theta = \beta^{-1}$.

E6.20 Transformed uniform variables.

(a) W has support $0 \le w \le 1$ and cdf $F_W(w) = P(W \le w) = P(Y^2 \le w) = P(-\sqrt{w} \le Y \le \sqrt{w})$, equivalent to $P(0 \le Y \le \sqrt{w}) = \int_0^{\sqrt{w}} dy\, f_Y(y) = \sqrt{w}$. Thus $f_W(w) = \frac{1}{2}w^{-1/2}$, $0 \le w \le 1$.

(b) W has support $0 \le w \le 1$ and cdf $F_W(w) = P(W \le w) = P(\sqrt{Y} \le w) = P(Y \le w^2)$ Integrating, $F_W(w) = \int_0^{w^2} dy\, f_Y(y) = w^2$. Thus $f_W(w) = 2w$, $0 \le w \le 1$.

6.4 The Method of Transformations

E6.26 Weibull transformations.

(a) The transformation $u = h(y) = y^m$ is monotone for $y > 0$ with inverse $h^{-1}(u) = u^{1/m}$, hence $f_U(u) = f_Y(h^{-1}(u))|\frac{dh^{-1}(u)}{du}| = (m/\alpha)u^{1-1/m}e^{-u/\alpha}|u^{1/m-1}/m| = e^{-u/\alpha}/\alpha$, $u^{1/m} > 0 \Rightarrow u > 0$. Evidently $U \sim$ exponential(α).

(b) $Y^k = U^{k/m}$, so $\mathrm{E}\,Y^k = \mathrm{E}\,U^{k/m} = \int_0^\infty du\, u^{k/m} f_U(u)$. See E4.186 for a more direct computation.

E6.27 Weibull exponential relation.

(a) $f_Y(y) = \frac{1}{\beta}e^{-y/\beta}$, $y > 0$. The transformation $w = h(y) = \sqrt{y}$ is monotone for $y > 0$, with inverse $h^{-1}(w) = w^2$, hence $f_W(w) = f_Y(h^{-1}(w))|\frac{dh^{-1}(w)}{dw}| = \frac{1}{\beta}e^{-w^2/\beta}|2w| = \frac{2}{\beta}we^{-w^2/\beta}$, $w^2 > 0 \Rightarrow w > 0$. Evidently $W \sim$ Weibull$(\alpha = \beta, m = 2)$.

(b) $\mathrm{E}\,Y^{k/2} = \mathrm{E}\,W^k$.

E6.28 Uniform exponential relation. $Y \sim$ uniform$(0,1)$ has $f_Y(y) = 1$, $0 \le y \le 1$. Consider the transformation $U = h(Y) = -c\ln Y$, $c > 0$. $h(y)$ is monotone over $0 \le y \le 1$, with inverse $h^{-1}(u) = e^{-u/c}$. Thus $f_U(u) = f_Y(h^{-1}(u))|\frac{dh^{-1}(u)}{du}| = \frac{1}{c}e^{-u/c}$, for $0 \le e^{-u/c} \le 1 \Rightarrow u > 0$. Evidently $U \sim$ exponential(c).

E6.32 Cost of a random delay 2. $Y \sim$ uniform$(1,5)$ has $f_Y(y) = \frac{1}{4}$, $1 \le y \le 5$. Consider the transformation $U = h(Y) = aY^2 + b$, $a, b > 0$. $h(y)$ is monotone over $1 \le y \le 5$, with inverse $h^{-1}(u) = \sqrt{(u-b)/a}$. Thus $f_U(u) = f_Y(h^{-1}(u))|\frac{dh^{-1}(u)}{du}| = 1/[8a\sqrt{(u-b)/a}]$, $1 \le \sqrt{(u-b)/a} \le 5$, implying $a + b \le u \le 25a + b$. The result agrees with E6.5.

E6.34 Rayleigh density. The transformation $U = h(Y) = Y^2$ is monotone over $y > 0$, with inverse $h^{-1}(u) = \sqrt{u}$. Thus $f_U(u) = f_Y(h^{-1}(u))|\frac{dh^{-1}(u)}{du}| = \frac{1}{\theta}e^{-u/\theta}$, for $\sqrt{u} > 0 \Rightarrow u > 0$. Evidently $U \sim$ exponential(θ). It follows $\mathrm{E}\,Y = \mathrm{E}\,U^{1/2} = \frac{1}{\theta}\int_0^\infty du\, u^{1/2}e^{-u/\theta}$ and $\mathrm{var}\,Y = \mathrm{E}\,U - (\mathrm{E}\,Y)^2$.

E6.35 Product of uniform densities. The joint density $f(y_1, y_2) = 1$ with support $0 \le y_1, y_2 \le 1$. For fixed y_1, the transformation $U = h(Y_2) = y_1Y_2$ is monotone over the support with inverse $h^{-1}(U) = U/y_1$. Thus $f(y_1, u) = f(y_1, h^{-1}(u))|\frac{dh^{-1}(u)}{du}| = 1/y_2$ for $0 \le y_1 \le 1$ and $0 \le u/y_1 \le 1$, implying $0 \le u \le y_1 \le 1$. Integrating over y_1 to obtain the marginal density, $f_U(u) = \int_u^1 dy_1\, f(y_1, u) = -\log(u)$, $0 \le u \le 1$.

E6.36 Rayleigh density 2. The joint density $f(y_1, y_2) = 4y_1y_2e^{-(y_1^2+y_2^2)/\theta}/\theta^2$ with support $y_1, y_2 > 0$. For fixed y_1, the transformation $U = h(Y_2) = y_1^2 + Y_2^2$ is monotone over the support with inverse $h^{-1}(U) = \sqrt{U - y_1^2}$. Thus $f(y_1, u) = f(y_1, h^{-1}(u))|\frac{dh^{-1}(u)}{du}| = 2y_1e^{-u/\theta}/\theta^2$ for $y_1 > 0$ and $\sqrt{u - y_1^2} > 0$, implying $u > y_1^2$. Integrating over y_1 to obtain the marginal density, we have $f_U(u) = \int_0^{\sqrt{u}} dy_1\, f(y_1, u) = ue^{-u/\theta}/\theta^2$, $u > 0$. Evidently $U \sim$ gamma$(\alpha = 2, \beta = \theta)$.

6.5 The Method of Moment-Generating Functions

E6.37 Sum of independent Bernoulli r.v. (a) By definition, $m_{Y_i}(t) = \mathrm{E}(e^{tY_i}) = pe^t + (1-p)$. **(b)** By independence, $m_W(t) = \prod_{i=1}^n m_i(t) = [pe^t + (1-p)]^n$, the mgf of a binomial r.v. **(c)** By uniqueness of mgfs, $W \sim$ binomial(n, p).

E6.38 Mgf of a linear combination. By definition,

$$M_U(t) = \mathrm{E}[e^{t(a_1Y_1+a_2Y_2)}] = \mathrm{E}[e^{(a_1t)Y_1}e^{(a_2t)Y_2}] = \mathrm{E}[e^{(a_1t)Y_1}]\,\mathrm{E}[e^{(a_2t)Y_2}],$$

where the last step follows by independence (Theorem 5.9). This is just $M_U(t) = M_{Y_1}(a_1t)M_{Y_2}(a_2t)$.

E6.43 Density of a normal sample mean. This is a special case of Theorem 6.3. $\overline{Y} \sim$ normal$(\mu, \sigma^2/n)$.

***E6.44 Packing watermelons.** Let $X_i \sim \text{normal}(\mu, \sigma^2)$ be the weight of a watermelon, $\mu = 15$ and $\sigma^2 = 4$. We are interested in the sum of weights $Y_n = \sum_{i=1}^n X_i$ for undetermined n. From Theorem 6.3, we know $Y_n \sim \text{normal}(n\mu, n\sigma^2)$, and we want to determine n such that $P(Y_n \geq 140) \leq 0.05$. Since $f_{Y_n}(y) = e^{-(y-n\mu)^2/(2n\sigma^2)}/\sqrt{2\pi n\sigma^2}$, this is equivalent to finding the largest n such that

$$P(Y_n \geq 140) = \int_{140}^{\infty} dy \, \frac{1}{\sqrt{2\pi n\sigma^2}} e^{-(y-n\mu)^2/(2n\sigma^2)} \leq 0.05.$$

$n = 8$ yields $P(Y_8 \geq 140) \approx 2.03 \times 10^{-4}$ while $n = 9$ yields $P(Y_9 \geq 140) \approx 0.202$.

E6.49 Sum of two independent binomial r.v. with same p. The mgf of $X \sim \text{binomial}(n, p)$ is $m_X(t) = (pe^t + q)^n$. By independence,

$$M_{Y_1+Y_2}(t) = M_{Y_1}(t)M_{Y_2}(t) = (pe^t + q)^{n_1}(pe^t + q)^{n_2} = (pe^t + q)^{n_1+n_2}.$$

By uniqueness of mgfs, $(Y_1 + Y_2) \sim \text{binomial}(n_1 + n_2, p)$.

E6.50 A related binomial distribution. Let $Z = n - Y$. By definition,

$$m_Z(t) = \mathrm{E}(e^{tZ}) = \mathrm{E}(e^{t(n-Y)}) = e^{tn}\,\mathrm{E}(e^{(-t)Y}) = e^{tn}M_Y(-t) = e^{tn}(pe^{-t} + q)^n = (p + qe^t)^n,$$

which is the mgf of a binomial(n, q) r.v.

E6.51 Sum of two independent binomial r.v. with $p_1 = 1 - p_2$. From E6.50, the distribution of $(n_2 - Y_2) \sim \text{binomial}(n_2, q_2) \simeq \text{binomial}(n_2, p_1)$. From E6.49, $Y_1 + (n_2 - Y_2) \sim \text{binomial}(n_1 + n_2, p_1)$.

E6.52 Sum of two independent Poisson r.v.
$X \sim \text{Poisson}(\lambda)$ has mgf $m_X(t) = e^{\lambda(e^t-1)}$ and pmf $p(y) = \lambda^y e^{-\lambda}/y!$.

(a) By independence, $M_{Y_1+Y_2}(t) = M_{Y_1}(t)M_{Y_2}(t) = e^{\lambda_1(e^t-1)}e^{\lambda_2(e^t-1)} = e^{(\lambda_1+\lambda_2)(e^t-1)}$. By uniqueness of mgfs, $(Y_1 + Y_2) \sim \text{Poisson}(\lambda_1 + \lambda_2)$.

(b) $P(Y_1 = y_1 | Y_1 + Y_2 = m) = P(Y_1 = y_1, Y_1 + Y_2 = m)/P(Y_1 + Y_2 = m)$ by definition of conditional probability. Using the result from (a), letting $Z = Y_1 + Y_2$, the denominator is obtained from

$$P(Z = z) = (\lambda_1 + \lambda_2)^z e^{-(\lambda_1+\lambda_2)}/z!, \qquad z = 0, 1, 2, \ldots.$$

The numerator is equivalent to $P(Y_1 = y_1, Y_2 = m - y_1)$ and can be obtained from the joint pmf $P(y_1, y_2) = \lambda_1^{y_1}\lambda_2^{y_2}e^{-(\lambda_1+\lambda_2)}/(y_1!y_2!)$, with support $y_1 = 0, 1, 2, \ldots$ and $y_2 = 0, 1, 2, \ldots$

$$P(Y_1 = y_1, Y_2 = m - y_1) = \lambda_1^{y_1}\lambda_2^{m-y_1}e^{-(\lambda_1+\lambda_2)}/(y_1!(m - y_1)!), \qquad y_1 = m, m - 1, \ldots, 0.$$

Thus,

$$P(Y_1 = y_1 | Y_1 + Y_2 = m) = \frac{m!}{y_1!(m - y_1)!}\frac{\lambda_1^{y_1}\lambda_2^{m-y_1}}{(\lambda_1 + \lambda_2)^m} = \binom{m}{y_1}p^{y_1}(1 - p)^{m-y_1},$$

$y_1 = 0, 1, \ldots, m$, is binomial distributed with m trials and $p = \lambda_1/(\lambda_1 + \lambda_2)$.

E6.53 Sum of independent binomial r.v. Let $Z = \sum_{i=1}^n Y_i$ and $N = \sum_{i=1}^n n_i$.

(a, b) These are extensions of E6.49. $Z \sim \text{binomial}(N, p)$.

(c) $P(Y_1 = y_1 | Z = m) = P(Y_1 = y_1, Z = m)/P(Z = m)$. From (b), the denominator is obtained from

$$P(Z = z) = \binom{N}{z}p^z(1 - p)^{N-z}, \qquad z = 0, 1, \ldots, N.$$

The numerator is equal to $P\left(Y_1 = y_1, y_1 + \sum_{i=2}^n Y_i = m\right) = P\left(Y_1 = y_1, W = m - y_1\right)$ where $W \sim$ binomial$(k = N - n_1, p)$. The joint pmf of Y_1 and W is

$$p(y_1, w) = \binom{n_1}{y_1} p^{y_1}(1-p)^{n_1 - y_1} \binom{k}{w} p^w (1-p)^{k-w}, \qquad y_1 = 0, 1, \ldots n_1,\ w = 0, 1, \ldots, k.$$

Thus

$$P\left(Y_1 = y_1, W = m - y_1\right) = \binom{n_1}{y_1} p^{y_1}(1-p)^{n_1 - y_1} \binom{k}{m - y_1} p^{m - y_1}(1-p)^{k - m + y_1}$$

for $y_1 = 0, 1, \ldots, n_1$ and $m - y_1 = 0, 1, \ldots, k$, i.e. $y_1 = 0, 1, \ldots, n_1$. Dividing, we obtain the hypergeometric distribution

$$P(Y_1 = y_1 | Z = m) = \binom{n_1}{y_1}\binom{N - n_1}{m - y_1} \Big/ \binom{N}{m}, \qquad y_1 = 0, 1, \ldots, n_1.$$

(d) $P(Y_1 + Y_2 = u | Z = m) = P(Y_1 + Y_2 = u, Z = m)/P(Z = m)$. The denominator is obtained just like in (c). The numerator is $P(Y_1 + Y_2 = u, u + \sum_{i=3}^n Y_i = m) = P(Y_1 + Y_2 = u, V = m - u)$ where $V \sim$ binomial$(l = N - n_1 - n_2, p)$. Noting that $U = Y_1 + Y_2 \sim$ binomial$(n_1 + n_2, p)$,

$$p(u, v) = \binom{n_1 + n_2}{u} p^i (1-p)^{n_1 + n_2 - u} \binom{l}{v} p^v (1-p)^{l - v}, \qquad y = 0, 1, \ldots n_1 + n_2,\ v = 0, 1, \ldots, l.$$

From this we obtain $P(Y_1 + Y_2 = u, V = m - u)$. The final result is still hypergeometric:

$$P(Y_1 + Y_2 = u | Z = m) = \binom{n_1 + n_2}{u}\binom{N - n_1 - n_2}{m - u} \Big/ \binom{N}{m}, \qquad u = 0, 1, \ldots, n_1 + n_2.$$

(e) No. The exponents $a^n b^m$ only add when $a = b$.

E6.54 Sum of independent Poisson r.v. Let $Z = \sum_{i=1}^n Y_i$ and $\Lambda = \sum_{i=1}^n \lambda_i$.

(a) This is an extension of E6.52. $Z \sim$ Poisson(Λ).

(b) Analogous to E6.53(c), $P(Y_1 = y_1 | Z = m) = P(Y_1 = y_1, W = m - y_1)/P(Z = m)$ where $W = \sum_{i=2}^n Y_i$. In general from (a), $P(Z = z) = \Lambda^z e^{-\Lambda}/z!$, $z = 0, 1, 2, \ldots$, allowing us to obtain the denominator. To obtain the numerator, note $W \sim$ Poisson$(\Lambda - \lambda_1)$, so by independence, joint density $p(y_1, w) = [\lambda_1^{y_1} e^{-\lambda_1}/y_1!] \times [(\Lambda - \lambda_1)^{y_1} e^{-(\Lambda - \lambda_1)}/w!]$ for $y_1 = 0, 1, 2, \ldots$ and $w = 0, 1, 2, \ldots$. Thus $P(Y_1 = y_1, W = m - y_1) = [\lambda_1^{y_1} e^{-\lambda_1}/y_1!] \times [(\Lambda - \lambda_1)^{m - y_1} e^{-(\Lambda - \lambda_1)}/(m - y_1)!]$ for $y_1 = 0, 1, \ldots, m$. Substituting and simplifying we obtain a binomial distribution

$$P(Y_1 = y_1 | Z = m) = \binom{m}{y_1} p^{y_1}(1-p)^{m - y_1}, \qquad y_1 = 0, 1, \ldots, m, \qquad \text{where } p = \lambda_1/\Lambda.$$

(c) Since $Y_1 + Y_2 \sim$ Poisson$(\lambda_1 + \lambda_2)$ and $V = \sum_{i=3}^n \sim$ Poisson$(\Lambda - \lambda_1 - \lambda_2)$, from (b) we conclude $(Y_1 + Y_2 \mid Z = m) \sim$ binomial$(m, (\lambda_1 + \lambda_2)/\Lambda)$.

E6.55 Customer arrival. This is a Poisson process with $\lambda^* = 2\lambda$. Let Y denote the number of customer arrivals with $p(Y) = (2\lambda)^y e^{-2\lambda}/y!$. We are interested in $1 - \sum_{i=0}^{19} p(i)$.

Comment. The Poisson process assumption can be justified as a sum of r.v. $X_i \sim$Poisson(λ_i), whose sum $\sum_{i=1}^n X_i \sim$ Poisson$(\sum_{i=1}^n \lambda_i)$ from E6.54. In this case, $n = 2$ and $\lambda_1 = \lambda_2$.

E6.57 Sum of independent gamma r.v. Y_i has mgf $m_{Y_i}(t) = (1 - \beta t)^{-\alpha_i}$, thus by independence, $Z = \sum_{i=1}^n Y_i$ has mgf $m_Z(t) = (1 - \beta t)^{-A}$, $A = \sum_{i=1}^n \alpha_i$. By uniqueness of mgfs, $Z \sim$ gamma(A, β).

6.6 Multivariable Transformations Using Jacobians

***E6.66 Density of a sum.**

(a) Consider the one-to-one transformation $(u_1, u_2) = (y_1 + y_2, y_2)$ over $(y_1, y_2) \in \mathbb{R}^2$, with inverse $(y_1, y_2) = (u_1 - u_2, u_2)$. The Jacobian of the transformation $J = \frac{\partial y_1}{\partial u_1}\frac{\partial y_2}{\partial u_2} - \frac{\partial y_1}{\partial u_2}\frac{\partial y_2}{\partial u_1} = 1$, thus the density $f_{U_1,U_2}(u_1, u_2) = f_{Y_1,Y_2}(u_1 - u_2, u_2)$, $(u_1 - u_2) \in \mathbb{R}$ and $u_2 \in \mathbb{R}$, i.e. $(u_1, u_2) \in \mathbb{R}^2$.

(b, c) By definition, $f_{U_1}(u_1) = \int_{-\infty}^{\infty} du_2\, f_{Y_1,Y_2}(u_1 - u_2, u_2)$. For Y_1 independent of Y_2, the joint density factors into $f_{Y_1,Y_2}(u_1 - u_2, u_2) = f_{Y_1}(u_1 - u_2) f_{Y_2}(u_2)$.

***E6.67 Density of a quotient.**

(a) Consider the one-to-one transformation $(u_1, u_2) = (y_1/y_2, y_2)$ over $(y_1, y_2) \in \mathbb{R}^2$, with inverse $(y_1, y_2) = (u_1 u_2, u_2)$. The Jacobian of the transformation $J = \frac{\partial y_1}{\partial u_1}\frac{\partial y_2}{\partial u_2} - \frac{\partial y_1}{\partial u_2}\frac{\partial y_2}{\partial u_1} = u_2$, thus the density $f_{U_1,U_2}(u_1, u_2) = f_{Y_1,Y_2}(u_1 - u_2, u_2)|u_2|$, $u_1 u_2 \in \mathbb{R}$ and $u_2 \in \mathbb{R}$, i.e. $(u_1, u_2) \in \mathbb{R}^2$.

(b, c) By definition, $f_{U_1}(u_1) = \int_{-\infty}^{\infty} du_2\, f_{Y_1,Y_2}(u_1 - u_2, u_2)|u_2|$. For Y_1 independent of Y_2, the joint density factors into $f_{Y_1,Y_2}(u_1 - u_2, u_2) = f_{Y_1}(u_1 - u_2) f_{Y_2}(u_2)|u_2|$.

***E6.70 Sum and difference of independent standard uniform r.v.**

(a) Consider the one-to-one transformation $(u_1, u_2) = (y_1 + y_2, y_1 - y_2)$ over $(y_1, y_2) \in [0,1] \times [0,1]$, with inverse $(y_1, y_2) = ((u_1 + u_2)/2, (u_1 - u_2)/2)$. The Jacobian of the transformation

$$J = \frac{\partial y_1}{\partial u_1}\frac{\partial y_2}{\partial u_2} - \frac{\partial y_1}{\partial u_2}\frac{\partial y_2}{\partial u_1} = -\frac{1}{2},$$

thus the density $f_{U_1,U_2}(u_1, u_2) = |J| f_{Y_1,Y_2}((u_1 + u_2)/2, (u_1 - u_2)/2) = \frac{1}{2}$, $0 \le (u_1 + u_2)/2 \le 1$ and $0 \le (u_1 - u_2)/2 \le 1$. The support is equivalent to $-u_2 \le u_1 \le 2 - u_2$ and $u_2 \le u_1 \le 2 + u_2$, which is the region of a square with corners at $(u_2, u_1) \in \{(0,0), (1,1), (0,2), (-1,1)\}$.

Comment. To prove the transformation $(u_1, u_2) = \mathbf{g}(y_1, y_2)$ is one-to-one, we must show that $\mathbf{g}(y_1, y_2) = \mathbf{g}(x_1, x_2)$ implies $(y_1, y_2) = (x_1, x_2)$. Equating $\mathbf{g}(y_1, y_2) = \mathbf{g}(x_1, x_2)$ yields

$$y_1 + y_2 = x_1 + x_2, \qquad y_1 - y_2 = x_1 - x_2.$$

Adding and subtracting yields $2y_1 = 2x_1$ and $2y_2 = 2x_2$, so indeed $(y_1, y_2) = (x_1, x_2)$.

(c) $f_{U_1}(u_1 \in [1,2]) = \int_{u_1-2}^{2-u_1} du_2\, f_{U_1,U_2}(u_1, u_2) = 2 - u_1$.
$f_{U_1}(u_1 \in [0,1]) = \int_{-u_1}^{u_1} du_2\, f_{U_1,U_2}(u_1, u_2) = u_1$.

(d) $f_{U_2}(u_2 \in [0,1]) = \int_{u_2}^{2-u_2} du_1\, f_{U_1,U_2}(u_1, u_2) = 1 - u_2$.
$f_{U_2}(u_2 \in [-1,0]) = \int_{-u_2}^{2+u_2} du_1\, f_{U_1,U_2}(u_1, u_2) = 1 + u_2$.

(e) No, $f_{U_1,U_2}(u_1, u_2) \ne f_{U_1}(u_1) f_{U_2}(u_2)$ for all (u_1, u_2).

***E6.71 Sum and quotient of independent exponential r.v.**

$Y_i \sim$ exponential(β) has $f_{Y_i}(y_i) = \frac{1}{\beta} e^{-y_i/\beta}$, $y_i \ge 0$. By independence $f_{Y_1,Y_2}(y_1, y_2) = \frac{1}{\beta^2} e^{-(y_1+y_2)/\beta}$.

(a) Consider the one-to-one transformation $(u_1, u_2) = (y_1 + y_2, y_1/y_2)$ over $(y_1, y_2) \in [0,\infty) \times [0,\infty)$, with inverse $(y_1, y_2) = (u_1 u_2/(u_2 + 1), u_1/(u_2 + 1))$. The Jacobian of the transformation

$$J = \frac{\partial y_1}{\partial u_1}\frac{\partial y_2}{\partial u_2} - \frac{\partial y_1}{\partial u_2}\frac{\partial y_2}{\partial u_1} = -\frac{u_1}{(u_2 + 1)^2},$$

thus

$$f_{U_1,U_2}(u_1, u_2) = |J| f_{Y_1,Y_2}\left(\frac{u_1 u_2}{u_2 + 1}, \frac{u_1}{u_2 + 1}\right) = \frac{u_1}{(u_2 + 1)^2} \frac{1}{\beta^2} e^{-u_1/\beta},$$

$u_1 u_2/(u_2 + 1) \ge 0$ and $u_1/(u_2 + 1) \ge 0$. Clearly $u_1 = 0$ is allowed. For $u_1 \ne 0$, the latter implies u_1 and $(u_2 + 1)$ have the same sign; the former then implies $u_2 \ge 0 \Rightarrow (u_2 + 1) > 0 \Rightarrow u_1 \ge 0$. Therefore the support is also $(u_1, u_2) \in [0,\infty) \times [0,\infty)$.

Comment. To prove the transformation $(u_1, u_2) = \mathbf{g}(y_1, y_2)$ is one-to-one, we must show that $\mathbf{g}(y_1, y_2) = \mathbf{g}(x_1, x_2)$ implies $(y_1, y_2) = (x_1, x_2)$. Equating $\mathbf{g}(y_1, y_2) = \mathbf{g}(x_1, x_2)$ yields

$$y_1 + y_2 = x_1 + x_2, \qquad y_1/y_2 = x_1/x_2,$$

with unique solution $(y_1, y_2) = (x_1, x_2)$.

(b) Yes. The joint pdf factors into $h(u_1)h(u_2)$.

6.7 Order Statistics

E6.72 Minimum of two uniform r.v.
$Y \sim \text{uniform}(0, 1)$ has $f_Y(y) = 1$, $0 \le y \le 1$ and $F_Y(y) = y$, $0 \le y \le 1$.
(a) $U = \min(Y_1, Y_2)$ has $f_U(u) = 2[1 - F_Y(u)]f_Y(u) = 2(1 - u)$, $0 \le u \le 1$.
(b) $\mathrm{E}\,U = \int_0^1 du\, u f_U(u) = \frac{1}{3}$. Similarly, $\mathrm{var}\, U = \mathrm{E}\,U^2 - (\mathrm{E}\,U)^2 = \frac{1}{18}$.

E6.73 Maximum of two uniform r.v.
(a) $V = \max(Y_1, Y_2)$ has $f_V(v) = 2F_Y(v)f_Y(v) = 2v$, $0 \le v \le 1$.
(b) $\mathrm{E}\,V = \int_0^1 dv\, v f_V(v) = \frac{2}{3}$. Similarly, $\mathrm{var}\, V = \mathrm{E}\,V^2 - (\mathrm{E}\,V)^2 = \frac{1}{18}$.

E6.74 Minimum of uniform random sample.
$Y \sim \text{uniform}(0, \theta)$ has $f_Y(y) = 1/\theta$, $0 \le y \le \theta$ and $F_Y(y) = y/\theta$, $0 \le y \le \theta$.
(a) $F_{(n)}(y_n) = [F(y_n)]^n = (y_n/\theta)^n$, $0 \le y_n \le \theta$.
(b) $f_{(n)}(y_n) = n y_n^{n-1}/\theta^n$, $0 \le y_n \le \theta$.
(c) $\mathrm{E}\,Y_{(n)} = \int_0^\theta dy_n\, y_n f_{(n)}(y_n) = n\theta/(n+1)$. Similarly, $\mathrm{E}\,Y_{(n)}^2 = n\theta^2/(n+2)$ from which we obtain $\mathrm{var}\, Y_{(n)} = \mathrm{E}\,Y_{(n)}^2 - (\mathrm{E}\,Y_{(n)})^2$.

E6.75 Minimum of uniform random sample 2. $P(Y_{(5)} \le 10 \mid \theta = 15) = F_{(5)}(10; \theta = 15) \approx 0.132$.

***E6.76 Order statistics of uniform random sample.**
$Y \sim \text{uniform}(0, \theta)$ has $f_Y(y) = 1/\theta$, $0 \le y \le \theta$ and $F_Y(y) = y/\theta$, $0 \le y \le \theta$.
(a)

$$f_{(k)}(y_k) = \frac{n!}{(k-1)!(n-k)!}\left(\frac{y_k}{\theta}\right)^{k-1}\left(1 - \frac{y_k}{\theta}\right)^{n-k}\frac{1}{\theta}, \qquad 0 \le y_k \le \theta. \qquad \text{(Theorem 6.5)}$$

(b)

$$\begin{aligned}\mathrm{E}\,Y_{(k)} &= \frac{n!}{(k-1)!(n-k)!}\frac{1}{\theta}\int_0^\theta dy_k\, y_k \left(\frac{y_k}{\theta}\right)^{k-1}\left(1 - \frac{y_k}{\theta}\right)^{n-k} && \text{(by definition)} \\ &= \frac{n!}{(k-1)!(n-k)!}\theta\int_0^1 du\, u^k\,(1-u)^{n-k}. && (u = y_k/\theta)\end{aligned}$$

The integral is related to the beta density with $\alpha = k+1$ and $\beta = n-k+1$, hence evaluates to $\Gamma(\alpha)\Gamma(\beta)/\Gamma(\alpha+\beta) = k!(n-k)!/(n+1)!$, recalling $\Gamma(n) = (n-1)!$ for integer n. Thus $\mathrm{E}\,Y_{(k)} = k\theta/(n+1)$.
(c) Similar to (b), $\mathrm{E}\,Y_{(k)}^2 = k(k+1)\theta^2/[(n+1)(n+2)]$ and $\mathrm{var}\, Y_{(n)} = \mathrm{E}\,Y_{(n)}^2 - (\mathrm{E}\,Y_{(n)})^2$.
(d) $\mathrm{E}(Y_{(k)} - Y_{(k-1)}) = \mathrm{E}\,Y_{(k)} - \mathrm{E}\,Y_{(k-1)} = \theta/(n+1)$ is constant. The expected order statistics are spaced uniformly over $[0, \theta]$.

***E6.77 Order statistics of uniform random sample 2.**
(a) Using Theorem 6.5, for $0 \le y_j \le y_k \le \theta$,

$$f_{(j)(k)}(y_j, y_k) = \frac{n!}{(j-1)!(k-1-j)!(n-k)!}\left(\frac{y_j}{\theta}\right)^{j-1}\left(\frac{y_k - y_j}{\theta}\right)^{k-1-j}\left(1 - \frac{y_k}{\theta}\right)^{n-k}\frac{1}{\theta^2},$$

(b) $\text{cov}(Y_{(j)}, Y_{(k)}) = \text{E}(Y_{(j)}Y_{(k)}) - \text{E}\,Y_{(j)}\,\text{E}\,Y_{(k)}$. The second term can be obtained from E6.76(b). The first term can be obtained from (a), letting C denote the constant prefactor, as follows:

$$\begin{aligned}
\text{E}(Y_{(j)}Y_{(k)}) &= \int_0^\theta dy_j \int_{y_j}^\theta dy_k\, y_j y_k f_{(j)(k)}(y_j, y_k) \\
&= C\theta^2 \int_0^1 du\, u^j \int_u^1 dv\, v(v-u)^{k-1-j}(1-v)^{n-k} && \left(u = \frac{y_j}{\theta}, v = \frac{y_k}{\theta}\right) \\
&= C\theta^2 \Bigg[\int_0^1 du\, u^j (1-u)^{n-j+1} \int_0^1 dw\, w^{k-j}(1-w)^{n-k} && \left(w = \frac{v-u}{1-u}\right) \\
&\qquad + \int_0^1 du\, u^{j+1}(1-u)^{n-j} \int_0^1 dw\, w^{k-1-j}(1-w)^{n-k} \Bigg] \\
&= C\theta^2 \left[\frac{j!(k-j)!(n-k)!}{(n+2)!} + \frac{(j+1)!(k-1-j)!(n-k)!}{(n+2)!} \right] && \text{(related beta densities)} \\
&= \theta^2 \frac{j(k+1)}{(n+2)(n+1)}. && \text{(simplification)}
\end{aligned}$$

Thus, after some simplification, $\text{cov}(Y_{(j)}, Y_{(k)}) = \theta^2 j(n-k+1)/[(n+2)(n+1)^2]$.

(c) $\text{var}(Y_{(k)} - Y_{(j)}) = \text{var}\,Y_{(k)} + Y_{(j)} - 2\,\text{cov}(Y_{(j)}, Y_{(k)})$, where the covariance was found in (b) and the variance in E6.76(c).

E6.78 Order statistics of uniform random sample 3. Letting $\theta = 1$ in E6.76(a) yields the result.

E6.79 Order statistics of uniform random sample 4. Consider the one-to-one transformation $(u, v) = (y_{(1)}/y_{(n)}, y_{(n)})$ over $0 \le y_{(1)} \le y_{(n)} \le \theta$, with inverse $(y_{(1)}, y_{(n)}) = (uv, v)$. The Jacobian

$$J = \frac{\partial y_{(1)}}{\partial u}\frac{\partial y_{(n)}}{\partial v} - \frac{\partial y_{(1)}}{\partial v}\frac{\partial y_{(n)}}{\partial u} = v,$$

thus using E6.77(a),

$$f_{U,V}(u, v) = |J| f_{Y_{(1)},Y_{(n)}}(uv, v) = \frac{n!}{(n-2)!}\frac{1}{\theta^2} v \left[\frac{v(1-u)}{\theta}\right]^{n-2}, \qquad 0 \le uv \le v \le \theta,$$

hence the support $0 \le u \le 1$ and $0 \le v \le \theta$. The pdf can be factored into $f_{U,V}(u, v) = g(u)h(v)$, so U and V are independent.

E6.82 Exponential random sample. $Y \sim \text{exponential}(\beta)$ has $F_Y(y) = 1 - e^{-y/\beta}$, with median such that $\frac{1}{2} = F_Y(m) = 1 - e^{-m/\beta}$, implying $m = -\beta \log(\frac{1}{2})$. The cdf for the maximum $F_{Y_{(n)}}(y_n) = [F_Y(y_n)]^n = (1 - e^{-y/\beta})^n$, allowing us to obtain $P(Y_{(n)} > m) = 1 - F_{Y_{(n)}}(m) = 1 - (\frac{1}{2})^n$.

E6.83 The max of a random sample in relation to its mean.
Without loss of generality, $P(Y_{(n)} > m) = 1 - F_{Y_{(n)}}(m) = 1 - [F_Y(m)]^n$. By definition of the median, $F_Y(m) = \frac{1}{2}$ for any continuous r.v. Y, hence in general $P(Y_{(n)} > m) = 1 - (\frac{1}{2})^n$.

***E6.89 Range distribution of a uniform random sample.** Consider the general case $Y \sim \text{uniform}(a, b)$ with $f_Y(y) = 1/(b-a)$, $a \le y \le b$ and $F_Y(y) = y/(b-a)$, $a \le y \le b$. The joint density for the min/max, from Theorem 6.5, is

$$f_{(1),(n)}(y_1, y_n) = \frac{n!}{(n-2)!}\left(\frac{y_n - y_1}{b-a}\right)^{n-2} \frac{1}{(b-a)^2}, \qquad a \le y_1 \le y_n \le b.$$

Consider the one-to-one transformation $(U, R) = (Y_{(1)}, Y_{(n)} - Y_{(1)})$ over $a \le y_1 \le y_n \le b$ with inverse $(Y_{(1)}, Y_{(n)}) = (U, R + U)$. The Jacobian

$$J = \frac{\partial y_{(1)}}{\partial u}\frac{\partial y_{(n)}}{\partial r} - \frac{\partial y_{(1)}}{\partial r}\frac{\partial y_{(n)}}{\partial u} = 1,$$

thus the transformed density

$$f_{U,R}(u,r) = |J| f_{(1),(n)}(u, r+u) = \frac{n!}{(n-2)!}\left(\frac{r}{b-a}\right)^{n-2}\frac{1}{(b-a)^2}, \qquad a \le u \le r+u \le b.$$

The support is the triangular region $a \le u \le b$ and $0 \le r \le b-u$. The marginal density is obtained by integrating $f_{U,R}(u,r)$ over u:

$$\begin{aligned} f_R(r) &= \int_a^{b-r} du\, f_{U,R}(u,r) = \frac{n!}{(n-2)!}\frac{1}{(b-a)^2}\left(\frac{r}{b-a}\right)^{n-2}\int_a^{b-r} du \\ &= n(n-1)\frac{b-a-r}{(b-a)^n} r^{n-2}, \qquad 0 \le r \le b-a. \end{aligned}$$

For the standard uniform distribution, $(a,b) = (0,1)$, $R \sim \text{beta}(\alpha = n-1, \beta = 2)$.

***E6.91 Total observed life of independent exponentially distributed lifetimes.**
$Y \sim \text{exponential}(\theta)$ has $f_Y(y) = \frac{1}{\theta}e^{-y/\theta}$ and $F_Y(y) = 1 - e^{-y/\theta}$, with support $y \ge 0$.

(a) The components are an i.i.d. sample $Y_1, Y_2, \ldots, Y_n$ of lifetimes, each $Y_i \sim \text{exponential}(\theta)$. The lifetime W_j is the j-th order statistic. Consider $T_1 = W_1 = \min(Y_1, Y_2, \ldots, Y_n) = Y_{(1)}$; using Theorem 6.5,

$$f_{(1)}(w_1) = n\left(e^{-w_1/\theta}\right)^{n-1}\frac{1}{\theta}e^{-w_1/\theta} = \frac{1}{\theta/n}e^{-w_1/(\theta/n)}, \qquad w_1 \ge 0,$$

is the pdf for an exponential(θ/n) r.v. For $j \ge 2$, we are interested in the difference of two order statistics $T_j = W_j - W_{j-1}$. We can determine the distribution using the joint pdf

$$f_{(j-1),(j)}(w_{j-1}, w_j) = \frac{n!}{(j-2)!(0)!(n-j)!}\left[1 - e^{-w_{j-1}/\theta}\right]^{j-2}\left[e^{-w_j/\theta}\right]^{n-j}\left[\frac{1}{\theta^2}e^{-(w_j+w_{j-1})/\theta}\right],$$

$0 \le w_{j-1} \le w_j < \infty$, and the one-to-one transformation $(U_j, T_j) = (W_j, W_j - W_{j-1})$ with inverse $(W_j, W_{j-1}) = (U_j, U_j - T_j)$. The Jacobian of the transformation $J = -1$, so the transformed density

$$\begin{aligned} f_{U_j,T_j}(u_j, t_j) &= \frac{n!}{(j-2)!(n-j)!}\left[1 - e^{-(u_j-t_j)/\theta}\right]^{j-2}\left[e^{-u_j/\theta}\right]^{n-j}\left[\frac{1}{\theta^2}e^{-(2u_j-t_j)/\theta}\right] \\ &= \frac{n!}{(j-2)!(n-j)!}\frac{1}{\theta^2}e^{t_j/\theta}\left[1 - e^{-(u_j-t_j)/\theta}\right]^{j-2}\left[e^{-u_j(n-j+2)/\theta}\right], \end{aligned}$$

$0 \le u_j - t_j \le u_j < \infty$. $u_j - t_j \le u_j < \infty$ implies $t \ge 0$ and $0 \le u_j - t_j$ implies $t_j \le u_j$. Hence the support is $0 \le t_j \le u_j < \infty$. Integrating over u_j, the marginal density

$$\begin{aligned} f_{T_j}(t_j) &= \int_{t_j}^{\infty} du_j\, f_{U_j,T_j}(u_j, t_j) \\ &= \frac{n!}{(j-2)!(n-j)!}\frac{e^{t_j/\theta}}{\theta}\int_1^0 \left[-dv_j\, v_j^{-1}\right][1 - v_j]^{j-2}\left[v_j e^{-t_j/\theta}\right]^{n-j+2} && (v_j = e^{-(u_j-t_j)/\theta}) \\ &= \frac{n!}{(j-2)!(n-j)!}\frac{e^{-t_j(n-j+1)/\theta}}{\theta}\int_0^1 dv_j\ (1-v_j)^{j-2}\, v_j^{n-j+1} && \text{(simplification)} \\ &= \frac{n!}{(j-2)!(n-j)!}\frac{e^{-t_j(n-j+1)/\theta}}{\theta}\frac{(j-2)!(n-j+1)!}{n!} && \text{(related beta density)} \\ &= \frac{n-j+1}{\theta}e^{-t_j(n-j+1)/\theta}, \end{aligned}$$

which an exponential$(\theta/(n-j+1))$ pdf. Thus $T_j \sim \text{exponential}(\theta/(n-j+1))$, $j = 1, 2, \ldots, r$.

(b) Observe $\sum_{j=1}^r T_j = W_1 + \sum_{j=2}^r (W_j - W_{j-1}) = W_r$, hence $(n+1)\sum_{j=1}^r T_j = (n+1)W_r$. Similarly, $\sum_{j=1}^r jT_j = W_1 + \sum_{j=2}^r j(W_j - W_{j-1}) = -\sum_{j=1}^{r-1} W_j + rW_r = -\sum_{j=1}^r W_j + (r+1)W_r$. Thus $U_r = \sum_{j=1}^r (n-j+1)T_j = \sum_{j=1}^r W_j + (n-r)W_r$. From (a), $T_j \sim$ exponential$(\theta/(n-j+1))$, thus $\mathrm{E}\, U_r = \sum_{j=1}^r (n-j+1)\,\mathrm{E}\, T_j = r\theta$.

6.8 Summary

Supplementary Exercises

E6.92 A linear combination of normal r.v. Consider a linear combination $U = aY_1 + bY_1$. Using the method of moment-generating functions, i.e. Theorem 6.3, $U \sim \text{normal}(a\mu_1 + b\mu_2, a^2\sigma_1^2 + b^2\sigma_2^2)$.

E6.93 Distribution of power. $I \sim \text{uniform}(0,1)$ has pdf $f_I(i) = 1$, $0 \le i \le 1$. Assuming independence, the joint distribution of current and resistance is $f_{I,R}(i,r) = 2r$, $0 \le i, r \le 1$.

Now consider the power generated $P = I^2 R$. We will use the method of (bivariate) transformations. The transformation $(p,q) = (i^2 r, r)$ is one-to-one over $0 \le i, r \le 1$, with inverse $(i,r) = (\sqrt{p/q}, q)$ such that $0 \le \sqrt{p/q}, q \le 1$. $0 \le \sqrt{p/q} \le 1$ implies $0 \le p \le q$, hence the transformed support is $0 \le p \le q \le 1$. The Jacobian of the transformation

$$J = \frac{\partial i}{\partial p}\frac{\partial q}{\partial r} - \frac{\partial i}{\partial r}\frac{\partial q}{\partial i} = \frac{1}{2}\sqrt{\frac{1}{pq}},$$

hence the transformed pdf

$$f_{P,Q}(p,q) = |J| f_{I,R}(\sqrt{p/q}, q) = \frac{1}{2}\sqrt{\frac{1}{pq}}(2q) = \sqrt{\frac{q}{p}}, \qquad 0 \le p \le q \le 1.$$

The marginal density of P is obtained by integrating over Q. The support is a triangular region. For a given $p \in [0,1]$, q ranges from p to 1, thus

$$f_P(p) = \int_p^1 dq\, f_{P,Q}(p,q) = \int_p^1 dq \sqrt{\frac{q}{p}} = \frac{2}{3}\left(p^{-1/2} - p\right), \qquad 0 \le p \le 1.$$

E6.95 Bivariate uniform transformations. $Y \sim \text{uniform}(0,1)$ has pdf $f_Y(y) = 1$, $0 \le y \le 1$. In all of the following, assuming independence, the joint distribution $f_{Y_1,Y_2}(y_1,y_2) = 1$, $0 \le y_1, y_2 \le 1$. We will use the following one-to-one transformations to obtain the transformed pdf $f_{U_i,V_i}(u_i,v_i) = |J_i|$.

(a)	$(u_1,v_1) = (y_1/y_2, y_2) \Leftrightarrow (y_1,y_2) = (u_1v_1, v_1)$	$\Rightarrow$	$J_1 = v_1$.
(b)	$(u_2,v_2) = (-\log(y_1y_2), y_2) \Leftrightarrow (y_1,y_2) = (e^{-u_2}/v_2, v_2)$	$\Rightarrow$	$J_2 = -e^{-u_2}/v_2$.
(c)	$(u_3,v_3) = (y_1y_2, y_2) \Leftrightarrow (y_1,y_2) = (u_3/v_3, v_3)$	$\Rightarrow$	$J_3 = 1/v_3$.

Finally, integration of v_i over the transformed support will yield the desired pdf.

(a) $f_{U_1,V_1}(u_1,v_1) = v_1$ over support $0 \le u_1v_1, v_1 \le 1$, or $0 \le v_1 \le 1$ and $0 \le u_1 \le 1/v_1 < \infty$. Divide into two regions: $0 \le u_1 \le 1$ and $u > 1$. $f_{U_1}(0 \le u_1 \le 1) = \int_0^1 dv_1\, f_{U_1,V_1}(u_1,v_1) = \frac{1}{2}$ and $f_{U_1}(u_1 > 1) = \int_0^{1/u_1} dv_1\, f_{U_1,V_1}(u_1,v_1) = \frac{1}{2}u^{-2}$.

(b) $f_{U_2,V_2}(u_2,v_2) = -e^{-u_2}/v_2$ over support $0 \le e^{-u_2}/v_2, v_2 \le 1$, equivalent to $0 \le v_2 \le 1$ and $u_2 \ge -\log v_2 \ge 0$. $f_{U_2}(u_2) = \int_{e^{-u_2}}^1 dv_2\, f_{U_2,V_2}(u_2,v_2) = u_2e^{-u_2}$, $u_2 \ge 0$.

(c) $f_{U_3,V_3}(u_3,v_3) = 1/v_3$ over support $0 \le u_3/v_3, v_3 \le 1$, equivalent to $0 \le u_3 \le v_3 \le 1$. $f_{U_3}(u_3) = \int_{u_3}^1 dv_3\, f_{U_3,V_3}(u_3,v_3) = -\log u_3$, $0 \le u_3 \le 1$.

E6.97 Sum of independent binomial r.v. 2. Consider the general case $Y_1 \sim \text{binomial}(n_1, p_1)$ and $Y_2 \sim \text{binomial}(n_2, p_2)$. By independence, their joint pmf

$$P(Y_1 = y_1, Y_2 = y_2) = \binom{n_1}{y_1} p_1^{n_1}(1-p_1)^{n_1 - y_1} \binom{n_2}{y_2} p_2^{n_2}(1-p_2)^{n_2-y_2}, \qquad y_i = 0, 1, \ldots, n_i.$$

Consider the transformation $(w, v) = (y_1 + y_2, y_2)$ with inverse $(y_1, y_2) = (w - v, v)$ such that $w - v = 0, 1, \ldots, n_1$ and $v = 0, 1, \ldots, n_2$. Each (y_1, y_2) is uniquely mapped to a (w, v). No Jacobian is needed in the discrete case, thus

$$f_{W,V}(w, v) = \binom{n_1}{w-v} p_1^{n_1}(1-p_1)^{n_1-(w-v)} \binom{n_2}{v} p_2^{n_2}(1-p_2)^{n_2-v}.$$

Summing over v yields the marginal pmf. To my knowledge, there is no closed form solution for the sum. A numerical approach is to sum $v = 0, 1, \ldots, n_2$ for a given w, noting that some $\binom{n_1}{w-v}$ coefficients may be zero.

***E6.98 Machine operation model.** The joint density is $f_{Y_1,Y_2}(y_1, y_2) = e^{-(y_1+y_2)}$, $y_i > 0$. The transformation $(u, v) = (y_1/(y_1 + y_2), y_1 + y_2)$ is one-to-one with inverse $(y_1, y_2) = (uv, (1-u)v)$ such that $uv > 0$ and $(1-u)v > 0$. Since v is the sum of two positive numbers, $v > 0$. The first condition then implies $u > 0$ while the second implies $(1-u) > 0$ or $u < 1$. Thus the transformed support is $0 < u < 1$ and $v > 0$. The Jacobian of the transformation is $J = uv + v(1-u) = v$. Thus

$$f_{U,V}(u, v) = |J| f_{Y_1,Y_2}(uv, (1-u)v) = ve^{-v}, \qquad 0 < u < 1, v > 0.$$

Integrating over v yields the marginal density

$$f_U(u) = \int_0^\infty dv\, f_{U,V}(u, v) = 1, \qquad 0 < u < 1.$$

Evidently $U \sim \text{uniform}(0, 1)$.

***E6.99 Machine operation model 2.** Inspection of $f_{U,V}(u, v)$ in E6.98 shows that it can be factored into $g(u)h(v)$ and has rectangular support, therefore U and V are independent.

E6.100 Random sample of electronic devices. $Y_1, Y_2, \ldots, Y_5$ each have pdf $f_Y(y) = \frac{1}{\beta}e^{-y/\beta}$ and cdf $F_Y(y) = 1 - e^{-y/\beta}$, $y > 0$. The probabilities of interested can be computed using the pdf of the first order statistic $g_{(1)}(y_{(1)}) = \frac{n!}{(n-1)!}[e^{-y_{(1)}/\beta}]^{n-1}\frac{1}{\beta}e^{-y_{(1)}/\beta}$, $y_{(1)} > 0$ with $n = 5$ and $\beta = 15$. **(a)** $P(Y_{(1)} > 9) = \int_9^\infty dy_{(1)}\, g_{(1)}(y_{(1)})$. **(b)** $P(Y_{(1)} < 12) = \int_0^{12} dy_{(1)}\, g_{(1)}(y_{(1)})$.

***E6.101 Parachutist landing.** Let Y denote her landing location, distributed as $Y \sim \text{uniform}(a, b)$ with pdf $f_Y(y) = 1/(b-a)$, $a \le y \le b$. We are interested in the distance $D = |Y - t|$, where $t = (b + a)/2$. We view the distance as a mixture of distributions dividing into $Y > t$ and $Y < t$, where the distance functions become $D_+ = Y - t$ and $D_- = t - Y$, respectively. By symmetry, $P(Y > t) = P(Y < t) = \frac{1}{2}$. In each half, she may land uniformly; $(Y \mid Y > t) \sim \text{uniform}(t, b)$ and $(Y \mid Y < t) \sim \text{uniform}(a, t)$. Given $Y > t$ and making the transformation $D_+ = Y - t$,

$$f_{D_+}(d) = f_{Y|Y>t}(d + t) = 1/(b-t), \qquad 0 \le d \le b - t.$$

Similarly,

$$f_{D_-}(d) = f_{Y|Y<t}(t - d) = 1/(t-a), \qquad t - a \ge d \ge 0.$$

Therefore the mixture density $f_D(d) = \frac{1}{2}[f_{D_+}(d) + f_{D_-}(d)]$. Note that $t - a = b - t = (b-a)/2$. Therefore $f_{D_+}(d) = f_{D_-}(d)$ and the density simplifies to $f_D(d) = 2/(b-a)$, $0 \le d \le (b-a)/2$. Evidently, $D \sim \text{uniform}(0, (b-a)/2)$.

Comment. The method of distribution functions is simpler.

$$P(D \le d) = P(-d \le Y - t \le d) = P(t - d \le Y \le t + d) = \int_{t-d}^{t+d} dy\, f_Y(y) = 2d/(b-a)$$

for $0 \le d \le b - t$, since otherwise $f_Y(y) = 0$. As before, this is the cdf of a uniform$(0,(b-a)/2)$ r.v.

E6.102 Sentry patrol. Let $X \sim$ uniform$(0,1)$ and $Y \sim$ uniform$(0,1)$ be independent r.v. representing the sentry positions with joint pdf $f_{X,Y}(x,y) = 1$, $0 \le x, y \le 1$. We are interested in the distance $D = |X - Y|$ and in particular the probability $P(D \le \frac{1}{2})$. Using the method of distribution functions,

$$P(D \le d) = P(-d \le X - Y \le d) = P(Y - d \le X \le Y + d) = 1 - P(Y - d \ge X, X \ge Y + d).$$

Probabilities are nonzero for $0 \le d \le 1$. Integration yields

$$P(D \le d) = 1 - \left[\int_d^1 dy \int_0^{y-d} dx + \int_0^{1-d} dy \int_{y+d}^1 dx\right] f_{X,Y}(x,y) = d(2-d), \qquad 0 \le d \le 1.$$

It follows that $P(D \le \frac{1}{2}) = \frac{3}{4}$.

***E6.103 Ratio of independent standard normal r.v.** $Y \sim$ normal$(0,1)$ has pdf $f_Y(y) = e^{-y^2/2}/\sqrt{2\pi}$, $y \in \mathbb{R}$. Assuming independence, the joint pdf $f_{Y_1,Y_2}(y_1,y_2) = e^{-(y_1^2+y_2^2)/2}/(2\pi)$, $(y_1,y_2) \in \mathbb{R}^2$. The transformation $(u,v) = (y_1/y_2, y_2)$ is one-to-one with inverse $(y_1,y_2) = (uv, v)$ such that $(u,v) \in \mathbb{R}^2$. The Jacobian of the transformation $J = v$, hence the transformed pdf

$$f_{U,V}(u,v) = |J| f_{Y_1,Y_2}(uv, v) = \frac{1}{2\pi}|v|e^{-(u^2+1)v^2/2} \qquad (u,v) \in \mathbb{R}^2.$$

Integration over v yields the marginal density of U.

$$\begin{aligned}
f_U(u) &= \int_{-\infty}^{\infty} dv\, f_{U,V}(u,v) = \frac{1}{2\pi}\int_{-\infty}^{\infty} dv\, |v|e^{-(u^2+1)v^2/2} \\
&= \frac{1}{\pi}\int_0^{\infty} dx\, ve^{-(u^2+1)v^2/2} && \text{(symmetric around } v = 0) \\
&= \frac{1}{2\pi}\int_0^{\infty} dx\, e^{-(u^2+1)x/2} && (x = v^2) \\
&= \frac{1}{\pi}\frac{1}{u^2+1}, \qquad u \in \mathbb{R}.
\end{aligned}$$

Comment. This is a standard Cauchy distribution. The general case allows for arbitrary σ_i. The means must be $\mu_i = 0$; otherwise, the symmetry simplification cannot be made.

***E6.104 Difference of geometric r.v.** $Y \sim$ geometric(p) has pmf $f_Y(y) = p(1-p)^{y-1}$, $y = 1,2,3,\ldots$. Assuming independence, the joint pmf $f_{Y_1,Y_2}(y_1,y_2) = p^2(1-p)^{y_1+y_2-2}$, $y_i = 1,2,3,\ldots$. The transformation $(u,v) = (y_1 - y_2, y_2)$ is one-to-one with inverse $(y_1,y_2) = (u+v, v)$ such that $v = 1,2,3,\ldots$ and $u + v = 1,2,3,\ldots$. Since each (y_1,y_2) maps uniquely to a (u,v), no Jacobian is needed. The transformed pmf

$$f_{U,V}(u,v) = f_{Y_1,Y_2}(u+v, v) = p^2(1-p)^{u+2v-2}, \qquad v = 1,2,3,\ldots \text{ and } u+v = 1,2,3,\ldots.$$

Note that $u \in \mathbb{Z}$. For a given $u \ge 0$, $v = 1,2,3,\ldots$ while for $u < 0$, $v = 1-u, 2-u, 3-u\ldots$. Summing over v for $u \ge 0$,

$$f_U(u \ge 0) = \sum_{v=1}^{\infty} f_{U,V}(u,v) = p^2(1-p)^{u-2}\sum_{v=1}^{\infty}[(1-p)^2]^v = p^2(1-p)^{u-2}\frac{(1-p)^2}{1-(1-p)^2} = \frac{p(1-p)^u}{2-p}.$$

Similarly,

$$f_U(u<0) = \sum_{v=1-u}^{\infty} f_{U,V}(u,v) = p^2(1-p)^{u-2}\frac{(1-p)^{2(1-u)}}{1-(1-p)^2} = \frac{p(1-p)^{-u}}{2-p}.$$

E6.105 Beta distribution of the second kind. The transformation $U = 1/(1+Y)$ is one-to-one with inverse $Y = 1/U - 1$ such that $1/u - 1 > 0$ or $0 < u < 1$. Furthermore $J = dy/du = -u^{-2}$, hence

$$f_U(u) = |J| f_Y(1/u-1) = \frac{1}{B(\alpha,\beta)} u^{\alpha+\beta-2}(1/u-1)^{\alpha-1} = \frac{1}{B(\alpha,\beta)} u^{\beta-1}(1-u)^{\alpha-1}, \qquad 0 \le u \le 1$$

is a beta(α,β) density.

E6.106 Probability integral transformation. The transformation $U = F(Y)$ has support $0 \le u \le 1$ by properties of $F(\cdot)$. Furthermore, $P(U \le u) = P(F(Y) \le u) = P(Y \le F^{-1}(u)) = F(F^{-1}(u)) = u$. Evidently $U \sim \text{uniform}(0,1)$.

E6.108 Reliability of a circuit.

E6.109 Profit distribution. The profit is the r.v.

$$X = \begin{cases} C_1 - C_3 & \text{if } 1/3 < y < 2/3 \\ C_2 - C_3 & \text{if } 0 < y < 1/3 \text{ or } 2/3 < y < 1 \end{cases}.$$

$1/3 < y < 2/3$ occurs with probability $p_1 = \int_{1/3}^{2/3} dy\, f(y) = \frac{101}{243}$. A similar calculation shows $(0 < y < 1/3$ or $2/3 < y < 1)$ occurs with probability $p_2 = \frac{142}{243}$. Thus the distribution of profit is $X(x = C_1 - C_3) = p_1$ and $X(x = C_2 - C_3) = p_2$.

7 Sampling Distributions and the Central Limit Theorem

7.1 Introduction

7.2 Sampling Distributions Related to the Normal Distribution

Summary of main results. Consider a random sample $Y_1, Y_2, \ldots, Y_n$ drawn from a normal distribution: $Y_i \sim \text{normal}(\mu, \sigma^2)$. We are often interested in making inferences about the population mean μ. This can be done using the sample mean statistic $\overline{Y} = \frac{1}{n}\sum_{i=1}^n Y_i$. If σ is known, we can use Theorem 7.1, $\overline{Y} \sim \text{normal}(\mu, \sigma^2/n)$, to make inferences (e.g. Examples 7.2, 7.3). It is common, however, that σ is unknown and must be estimated by the sample standard deviation $S = \sqrt{S^2}$ where $S^2 = \frac{1}{n-1}\sum_{i=1}^n (Y_i - \overline{Y})^2$. By Theorem 7.3, $(n-1)S^2/\sigma^2 \sim \chi^2(\text{d.f.} = n-1)$. This motivates the definition of the t-statistic $T = (\overline{Y} - \mu)/(S/\sqrt{n})$ which is distributed as the t-distribution with $(n-1)$ d.f., allowing for inference of μ with unknown σ (in a normal random sample).

To make inferences on ratios of sample variance, or equivalently σ_1/σ_2, between two normal random samples from populations $\text{normal}(\mu_1, \sigma_1^2)$ and $\text{normal}(\mu_2, \sigma_2^2)$, we use the F-statistic $F = (S_1^2/\sigma_1^2)/(S_2^2/\sigma_2^2)$, distributed as the F-distribution.

E7.19 Ammeter specifications. $P(S^2 \geq 0.65) = P[(n-1)S^2/\sigma^2 \geq (n-1)(0.65)/\sigma^2]$. Since σ and n are known and $(n-1)S^2/\sigma^2 \sim \chi^2(\text{d.f.} = n-1)$, using the χ^2 distribution yields $P(S^2 \geq 0.65) \approx 0.102$. Assuming $\sigma = 0.2$, a value of $s^2 \geq 0.65$ would be observed $\approx 10.2\%$ of the time by chance.

E7.29 Reciprocal of *F*-distribution. $F = (W_1/\nu_1)/(W_2/\nu_2)$ is a ratio of W_i/ν_i, $W_i \sim \chi^2(\nu_i)$ and $i = 1, 2$. $1/F$ is the same ratio with labels swapped: $1 \leftrightarrow 2$. Thus $1/F$ is also F-distributed with numerator and denominator d.f. swapped.

E7.33 *T*- and *F*-distribution relation.
$U = T^2 = (\overline{Y} - \mu)^2/(S^2/n) = [(\overline{Y} - \mu)/(\sigma/\sqrt{n})]^2/[(n-1)S^2/\sigma^2/(n-1)] = Z^2/[W/(n-1)]$, where $Z \sim \text{normal}(0,1)$ and $W \sim \chi^2(n-1)$ by Theorems 7.1 and 7.3, respectively. By Theorem 7.2, $Z^2 \sim \chi^2(1)$, hence U is F-distributed with 1 numerator d.f. and $(n-1)$ denominator d.f.

***E7.36 *F*-distribution probabilities.**

(a) The F-statistic is $F = (S_1^2/\sigma_1^2)/(S_2^2/\sigma_2^2)$. Assuming $\sigma_1^2 = 2\sigma_2^2$, $F = (S_1^2/2)/S_2^2$ follows an F-distribution with $(n_1 - 1)$ numerator d.f. and $(n_2 - 1)$ denominator d.f., hence the probability $P(S_1^2/S_2^2 \leq b) = P[(S_1^2/2)/S_2^2 \leq b/2] = 0.95$ when $b/2 \approx 3.68$.

(b) $P(a \leq S_1^2/S_2^2) = 1 - P(S_1^2/S_2^2 \leq a) = 1 - P[(S_1^2/2)/S_2^2 \leq a/2] = 0.95$ when $a/2 \approx 0.304$.

(c) $P(a \leq S_1^2/S_2^2 \leq b) = P[a/2 \leq (S_1^2/2)/S_2^2 \leq b/2] = 0.9$.

***E7.39 Independent normal samples with identical variance.**

(a) $\overline{X}_i$ are a linear combination of n_i r.v. from $\mathcal{N}(\mu_i, \sigma^2)$ with identical coefficients $1/n_i$, so $\overline{X}_i \sim \mathcal{N}(\mu_i, \sigma^2/n_i)$. $\hat{\theta}$ is a linear combination of k r.v. from $\mathcal{N}(\mu_i, \sigma^2/n_i)$ with coefficients c_i, $i = 1, 2, \ldots, k$, so $\hat{\theta} \sim \mathcal{N}(\sum_{i=1}^k c_i\mu_i, \sigma^2\sum_{i=1}^k c_i^2/n_i)$.

Comment. $\mathrm{E}\,\hat{\theta} = \theta$, so $\hat{\theta}$ is an unbiased estimator of θ.

(b) $W \equiv \text{SSE}/\sigma^2 = \sum_{i=1}^k (n_i - 1)S_i^2/\sigma^2 \equiv \sum_{i=1}^k W_i$ where by Theorem 7.3, $W_i \sim \chi^2(n_i - 1)$. It follows by Theorem 7.2 that $W_i \simeq \sum_{j=1}^{n_i - 1} Z_j^2$, where $Z_j \sim \mathcal{N}(0,1)$ are independent, hence

$$W = \sum_{i=1}^k W_i \simeq \sum_{i=1}^k \sum_{j=1}^{n_i - 1} Z_j^2 = \sum_{j=1}^{\sum_{i=1}^k (n_i - 1)} Z_j^2 \sim \chi^2\left(\sum_{i=1}^k n_i - k\right).$$

(c) Noting θ and σ are constants, $\hat{\theta}$ can be standardized by $Z = (\hat{\theta} - \theta)/\sqrt{\sigma^2\sum_{i=1}^k c_i^2/n_i}$. The quantity of interest can be expressed as $T = Z/\sqrt{\text{MSE}/\sigma^2} = Z/\sqrt{(\sum_{i=1}^k n_i - k)](\text{SSE}/\sigma^2)}$. The denominator is $\sqrt{W/\nu}$, $\nu = (\sum_{i=1}^k n_i - k)$, thus $T = Z/\sqrt{W/\nu}$ is t-distributed with ν d.f.

Comment. The way MSE defined is not truly a mean due to the $-k$ term. However, it is an unbiased estimator of σ^2: $\mathrm{E(MSE)} = \mathrm{E(SSE)}/\nu = \sigma^2$.

7.3 The Central Limit Theorem

CLT with unknown variance. Consider a random sample $Y_1, Y_2, \ldots, Y_n$ from some distribution Y. Often times μ and σ are both unknown. The CLT states $\overline{Y} \approx \mathcal{N}(\mu, \sigma^2/n)$ for large n, but with σ unknown, inference using the sample mean cannot be made. However, Slutsky's theorem shows that the sample standard deviation S converges to σ (in probability), hence a further approximation $\overline{Y} \approx \mathcal{N}(\mu, S^2/n)$ allows inference to be made by estimating σ with S. See E7.48 for an example.

E7.48 Estimated savings. **(a)** $P(|\overline{Y} - \mu| \leq 1) \approx P(-\sqrt{n/s^2} \leq Z \leq \sqrt{n/s^2}) \approx 0.378$. The second step follows since, per the note above, $\overline{Y} \approx \mathcal{N}(\mu, S^2/n)$ for large n. **(b)** No. The estimates could be wrong.

7.4 A Proof of the Central Limit Theorem

7.5 The Normal Approximation to the Binomial Distribution

Supplementary Exercises

***E7.97 Sum of χ^2 r.v.**

(a) $\mathrm{E}\,X_i = 1$ and $\mathrm{var}\,X_i = 2$. By the CLT, $\overline{X} = Y/n \sim \mathcal{N}[\mathrm{E}\,X_i, \mathrm{var}(X_i)/n] \simeq \mathcal{N}(1, 2/n)$, or $Y \sim \mathcal{N}(n, 2n)$. It follows that $Z = (Y - n)/\sqrt{2n} \sim \mathcal{N}(0, 1)$.

(b) Let $Y_1, Y_2, \ldots, Y_n$ be a random sample of size $n = 50$ from $\mathcal{N}(\mu = 6, \sigma^2 = 0.2)$ with associated cost $C_i = 4(Y_i - \mu)^2$. We are interested in the probability $P(\sum_{i=1}^n C_i \geq 48)$. Express $C_i = 4\sigma^2[(Y_i - \mu)/\sigma]^2 = 4\sigma^2 U$ where $U \sim \chi^2(1)$ since $Z = (Y_i - \mu)/\sigma \sim \mathcal{N}(0, 1)$. It follows that $\sum_{i=1}^n C_i = 4\sigma^2 W$ where $W \sim \chi^2(n)$. We can then use the $\chi^2(n)$ distribution to compute $P(\sum_{i=1}^n C_i \geq 48) = P[W \geq 48/(4\sigma^2)] \approx 0.157$.

***E7.100 Normal approximation to the Poisson distribution ($\boldsymbol{\lambda \gg 1}$).**

(a) $m_Y(t) = \mathrm{E}\,e^{tY} = \mathrm{E}\,e^{t(X-\lambda)/\sqrt{\lambda}} = e^{-t\sqrt{\lambda}}\,\mathrm{E}\,e^{(t/\sqrt{\lambda})X} = e^{-t\sqrt{\lambda}} M_X(t/\sqrt{\lambda}) = e^{-t\sqrt{\lambda}} e^{\lambda(e^{t/\sqrt{\lambda}}-1)}$. Simplifying yields the given result $m_Y(t) = \exp(\lambda e^{t/\sqrt{\lambda}} - \sqrt{\lambda}\,t - \lambda)$.

(b) $e^{t/\sqrt{\lambda}} = 1 + t/\sqrt{\lambda} + t^2/(2\lambda) + \mathcal{O}(\lambda^{-3/2})$, hence $\lambda e^{t/\sqrt{\lambda}} = \lambda + \sqrt{\lambda}\,t + t^2/2 + \mathcal{O}(\lambda^{-1/2})$. It follows that $M_Y(t) = \exp(t^2/2 + \mathcal{O}(\lambda^{-1/2}))$ and $\lim_{\lambda\to\infty} M_Y(t) = \exp(t^2/2)$.

(c) $\lim_{\lambda\to\infty} Y$ has a standard normal mgf and thus a standard normal distribution (Theorem 7.5).

***E7.101 Application of normal-Poisson approximation.** $P(X \leq x)$ can be computed exactly as the sum $P(X \leq x) = \sum_{i=0}^{x} \lambda^i e^{-\lambda}/i! \approx 0.853$ with $\lambda = 100$ and $x = 110$. Using the normal approximation, $P(X \leq x) = P[(X - \lambda)/\sqrt{\lambda} \leq (x - \lambda)/\sqrt{\lambda}] \approx P[Z \leq (x - \lambda)/\sqrt{\lambda}] \approx 0.841$.

***E7.104 Binomial approximation to the Poisson distribution ($\boldsymbol{n \gg 1, np = \lambda}$).** $M_{Y_n}(t) = [pe^t + (1-p)]^n = [1 + p(e^t - 1)]^n = [1 + (\lambda/n)(e^t - 1)]^n$. Using $\lim_{n\to\infty}(1 + x/n)^n = e^x$, $\lim_{n\to\infty} M_{Y_n}(t) = \exp[\lambda(e^t - 1)]$, which is the mgf of a Poisson(λ) r.v. Thus by Theorem 7.5, $\lim_{n\to\infty} Y_n \sim \text{Poisson}(\lambda)$.

***E7.105 Application of binomial-Poisson approximation.** Let Y denote the number of people who suffer adverse reactions; $Y \sim \text{binomial}(n, p)$. The exact probability can be computed using the binomial pmf: $P(Y \geq y) = 1 - P(Y \leq y - 1) = 1 - \sum_{i=0}^{y-1} \binom{n}{i} p^i (1-p)^{n-i} = 0.26424108....$ However in general, large binomial coefficients are difficult to compute (even numerically). Using the Poisson approximation, $Y \approx \text{Poisson}(\lambda = np)$ with $y = 2$, $P(Y \geq y) \approx 1 - \sum_{i=0}^{y-1} \lambda^i e^{-\lambda}/i = 0.26424111....$ The approximation is very good.

Comment. We could use the normal approximation to the binomial distribution, approximating $Y \approx \mathcal{N}(np, np(1-p))$. The result (with continuity correction) is $P(Y \geq 2) \approx 0.308$. Evidently in this case, the Poisson approximation is much better.

---

8 Estimation

8.1 Introduction

8.2 The Bias and Mean Square Error of Point Estimators

E8.1 MSE in terms of bias and variance. $\operatorname{var}\hat\theta = \mathrm{E}[(\hat\theta - \mathrm{E}\,\hat\theta)^2]$ motivates manipulation of

$$\mathrm{MSE}(\hat\theta) = \mathrm{E}[(\hat\theta-\theta)^2] = \mathrm{E}[(\hat\theta - \mathrm{E}\,\hat\theta + \mathrm{E}\,\hat\theta - \theta)^2] = \mathrm{E}[(\hat\theta - \mathrm{E}\,\hat\theta)^2 + (\mathrm{E}\,\hat\theta-\theta)^2 + 2(\hat\theta - \mathrm{E}\,\hat\theta)(\mathrm{E}\,\hat\theta-\theta)].$$

The first term is $\operatorname{var}\hat\theta$ and, since θ is a constant, the second term is $(\operatorname{bias}\hat\theta)^2$. The last term is zero since since $(\mathrm{E}\,\hat\theta-\theta)$ is a constant and $\mathrm{E}[(\hat\theta - \mathrm{E}\,\hat\theta)] = 0$. Thus, $\mathrm{MSE}(\hat\theta) = \operatorname{var}\hat\theta + (\operatorname{bias}\hat\theta)^2$.

E8.3 Linearly biased estimator. **(a)** $\operatorname{bias}\hat\theta = \mathrm{E}\,\hat\theta - \theta = (a-1)\theta + b$. **(b)** $\operatorname{bias}\hat\theta^* = \mathrm{E}\,\hat\theta^* - \theta = 0$ requires $\mathrm{E}\,\hat\theta^* = \theta$, i.e. the inverse linear transformation $\hat\theta^* = (\hat\theta - b)/a$.

E8.5 Linearly biased estimator 2. **(a)** $\mathrm{MSE}(\hat\theta^*) = \operatorname{var}\hat\theta^* = (\operatorname{var}\hat\theta)/a^2$ since $\operatorname{bias}\hat\theta^* = 0$ (see E8.1).

E8.6 Constrained sum of estimators. **(a)** $\mathrm{E}\,\hat\theta_3 = a\,\mathrm{E}\,\hat\theta_1 + (1-a)\,\mathrm{E}\,\hat\theta_2 = \theta$. **(b)** By independence, $\operatorname{var}\hat\theta_3 = a^2\operatorname{var}\hat\theta_1 + (1-a)^2\operatorname{var}\hat\theta_2 = a^2\sigma_1^2 + (1-a)^2\sigma_2^2$, whose derivative $\frac{d\operatorname{var}\hat\theta_3}{da} = 0$ at $a = \sigma_2/(\sigma_1+\sigma_2)$, corresponding to a minimum since $\frac{d^2\operatorname{var}\hat\theta_3}{da^2} = 2(\sigma_1+\sigma_2) > 0$.

E8.7 Constrained sum of estimators 2. $\operatorname{var}\hat\theta_3 = a^2\operatorname{var}\hat\theta_1 + (1-a)^2\operatorname{var}\hat\theta_2 + 2a(1-a)\operatorname{cov}(\hat\theta_1,\hat\theta_2)$. $\frac{d\operatorname{var}\hat\theta_3}{da} = 0$ at $a = (\sigma_2^2 - c)/(\sigma_1^2+\sigma_2^2-2c)$, corresponding to a minimum if $\frac{d^2\operatorname{var}\hat\theta_3}{da^2} = 2(\sigma_1^2+\sigma_2^2-2c) > 0$.

E8.8 Estimators of exponential density mean.

(a) $\mathrm{E}\,\theta_i = \theta$ allows for straightforward computation of $\mathrm{E}\,\hat\theta_i = \theta$, $i = 1,2,3,5$, showing they are unbiased. The pdf of $X = \hat\theta_4$ can be obtained using order statistics: $f_X(x) = (3/\theta)e^{-3x/\theta}$, $x > 0$ (Theorem 6.5), which is exponential with mean $\theta/3$, hence $\mathrm{E}\,\hat\theta_4 = \theta/3$ is biased.

(b) $\operatorname{var}\theta_i = \theta^2$ and independence of Y_j, $j = 1,2,3$, allow for straightforward computation of $\operatorname{var}\hat\theta_i$. The sample mean has the smallest variance $\operatorname{var}\hat\theta_5 = \theta^2/3$.

E8.9 Unbiased estimator of a related exponential mean. $Y \sim \text{exponential}(\theta{+}1)$, hence $\mathrm{E}\,\overline{Y} = \theta{+}1$ is a biased estimator of θ. It is easy to see that $(\overline{Y}-1)$ is unbiased.

E8.10 Poisson sample estimation. **(a)** $\overline{Y}$. **(b)** $\mathrm{E}\,C = 3\,\mathrm{E}\,Y + \mathrm{E}\,Y^2 = 3\,\mathrm{E}\,Y + [\operatorname{var}Y + (\mathrm{E}\,Y)^2] = 4\lambda + \lambda^2$. **(c)** We want to find $\hat\theta$ such that $\mathrm{E}\,\hat\theta = \mathrm{E}\,C$. $\mathrm{E}\,\overline{Y} = \lambda$ and $\mathrm{E}\,\overline{Y}^2 = \operatorname{var}\overline{Y} + \mathrm{E}\,\overline{Y}^2 = \lambda/n + \lambda^2$, so we can construct an unbiased estimator $\hat\theta = (\overline{Y}^2 - \overline{Y}/n) + 4\overline{Y}$.

E8.11 Third central moment estimator. We want to find $\hat\theta$ such that $\mathrm{E}\,\hat\theta = \mathrm{E}[(Y - \mathrm{E}\,Y)^3]$. Expanding $\mathrm{E}[(Y-\mathrm{E}\,Y)^3] = \mathrm{E}\,Y^3 - 3(\mathrm{E}\,Y)(\mathrm{E}\,Y^2) + 3(\mathrm{E}\,Y)^2(\mathrm{E}\,Y) - (\mathrm{E}\,Y)^3 = \mathrm{E}\,Y^3 - 3(\mathrm{E}\,Y)(\mathrm{E}\,Y^2) + 2(\mathrm{E}\,Y)^3$. Assuming $\hat\theta_i$ are unbiased estimators of $\mathrm{E}(Y^i)$, $i = 2,3$, then $\mathrm{E}\,\hat\theta_i = \mathrm{E}(Y^i)$. It follows that $\hat\theta = \hat\theta_3 - 3(\mathrm{E}\,Y)\hat\theta_2 + 2(\mathrm{E}\,Y)^3$.

E8.12 Uniform sample estimation.

(a) $Y_i \sim \text{uniform}(\theta, \theta+1)$ has $\mathrm{E}\,Y_i = (2\theta+1)/2$, therefore $\mathrm{E}\,\overline{Y} = (2\theta+1)/2$ is biased.

(b) $\operatorname{bias}\overline{Y} = \frac{1}{2}$, so $\hat\theta = \overline{Y} - \operatorname{bias}\overline{Y} = \overline{Y} - \frac{1}{2}$. is unbiased.

(c) $\mathrm{MSE}(\overline{Y}) = \operatorname{var}\overline{Y} + (\operatorname{bias}\overline{Y})^2 = 1/(12n) + (1/2)^2$.

E8.13 Binomial variance estimator. Let $\hat\theta = n(Y/n)(1 - Y/n)$.

(a) $\mathrm{E}\,\hat\theta = \mathrm{E}(Y - Y^2/n) = \mathrm{E}\,Y - [\operatorname{var}Y + (\mathrm{E}\,Y)^2]/n = np - [np(1-p) + (np)^2]/n$. Simplification yields $\mathrm{E}\,\hat\theta = (1-1/n)\operatorname{var}Y$. Evidently $\operatorname{bias}(\hat\theta) = -(\operatorname{var}Y)/n$.

(b) $\hat\theta^* = \hat\theta - \operatorname{bias}(\hat\theta) = \hat\theta + (\operatorname{var}Y)/n$ is unbiased, but not particularly helpful because we are estimating $\operatorname{var}Y$. It is straightforward to show that $\hat\theta^* = \hat\theta/[1/(1-1/n)]$ is unbiased.

E8.14 Power family sample estimation.

(a) The pdf of $X = \hat\theta$ can be obtained using order statistics: $f_X(x) = \alpha n x^{n\alpha-1}/\theta^{n\alpha}$, $0 \le x \le \theta$

(Theorem 6.5), is power distributed with $\alpha^* = n\alpha$. $\mathrm{E}\,X = (\alpha^*\theta)/(\alpha^*+1)$, hence X is a biased estimator of θ with $\mathrm{bias}(X) = \mathrm{E}\,X - \theta = \theta[1+\alpha^*/(\alpha^*+1)]$.

(b) It is easily verified that $\hat\theta^* = \hat\theta/[\alpha^*/(\alpha^*+1)]$ is unbiased.

(c) $\mathrm{MSE}(\hat\theta) = \mathrm{E}[(X-\theta)^2] = \int_0^\theta dx\, f_X(x)(x-\theta)^2 = 2\theta^2/[(\alpha^*)^2 + 3\alpha^* + 2)$.

***E8.16 Normal sample variance estimation.**

(a) $S = \sqrt{S^2} = \sqrt{(n-1)S^2/\sigma^2}/\sqrt{(n-1)/\sigma^2} = \sigma\sqrt{W}/\sqrt{(n-1)}$ where $W = (n-1)S^2/\sigma^2$. By Theorem 7.3, $W \sim \chi^2(k = n-1)$ with pdf $f_W(w) = w^{k/2-1}e^{-w/2}/[2^{k/2}\Gamma(k/2)]$, $w > 0$. We can compute

$$\mathrm{E}\,\sqrt{W} = \int_0^\infty dw\, f_W(w)\sqrt{w} = \frac{1}{2^{k/2}\Gamma(k/2)}\int_0^\infty dw\, w^{(k+1)/2-1}e^{-w/2}.$$

The integral is related to a gamma($\alpha = (k+1)/2, \beta = 2$) density, hence

$$\mathrm{E}\,\sqrt{W} = \frac{2^{(k+1)/2}\Gamma[(k+1)/2]}{2^{k/2}\Gamma(k/2)} = \frac{\sqrt{2}\Gamma[(k+1)/2]}{\Gamma(k/2)}.$$

It follows that $\mathrm{E}\,S = \sigma\,\mathrm{E}(\sqrt{W})/\sqrt{n-1} = \sigma\sqrt{2/(n-1)}\,\Gamma(n/2)/\Gamma[(n-1)/2] = c\sigma$, where $c = \sqrt{2/(n-1)}\,\Gamma(n/2)/\Gamma[(n-1)/2]$ is a constant.

(b) Since c is just a constant, $\hat\sigma = S/c$ is unbiased.

(c) We seek an estimator $\hat\theta$ such that $\mathrm{E}\,\hat\theta = \mu - z_\alpha\sigma$. Since an unbiased estimator of the mean is $\hat\mu = \overline{Y}$, an unbiased estimator of $\hat\theta$ is $\hat\theta = \hat\mu - \hat\sigma/z_\alpha = \overline{Y} - S/(cz_\alpha)$.

E8.17 Binomial probability estimator.

(a) $\mathrm{bias}(\hat p_2) = \mathrm{E}(\hat p_2) - p = (np+1)/(n+2) - p = (1-2p)/(n+2)$.

(b) $\mathrm{MSE}(\hat p_1) = \mathrm{var}(\hat p_1) + (\mathrm{bias}\,\hat p_1)^2 = \mathrm{var}(Y/n) = p(1-p)/n$.
$\mathrm{MSE}(\hat p_2) = \mathrm{var}(\hat p_2) + (\mathrm{bias}\,\hat p_2)^2 = np(1-p)/(n+2)^2 + [(1-2p)/(n+2)]^2$.

(c) Some algebra shows $\Delta = \mathrm{MSE}(\hat p_2) - \mathrm{MSE}(\hat p_1) = c[p^2 - p + n/(8n+4)]$ for constant $c > 0$. This is quadratic in p, so $\Delta > 0$ when $p < \frac{1}{2}[1 - \sqrt{1 - n/(2n+1)}]$ or $p > \frac{1}{2}[1 + \sqrt{1-n/(2n+1)}]$.

Comment. This makes sense because $\mathrm{var}(\hat p_2) < \mathrm{var}(\hat p_1)$ and when $p = \frac{1}{2}$, $\mathrm{bias}(\hat p_2) = 0$.

E8.18 Uniform sample estimation using minimum observed value. $f(y) = 1/\theta$ and $F(y) = y/\theta$, $0 \le y \le \theta$. Let $X = Y_{(1)}$. By Theorem 6.5, $f_X(x) = (n/\theta)(1 - x/\theta)^{n-1}$, $0 \le x \le \theta$, hence $\mathrm{E}\,X = \int_0^\theta dx\, x f_X(x) = n\theta\int_0^1 dw\, w^{2-1}(1-w)^{n-1} = nB(2,n)\theta = \theta/(n+1)$. ($w = x/\theta$ and we used the beta density to evaluate the integral.) It follows that $\hat\theta = (n+1)Y_{(1)}$ is unbiased.

***E8.20 Exponential sample estimators of transformations.**

(a) $f_{Y_1,Y_2}(y_1,y_2) = (1/\theta^2)e^{-(y_1+y_2)/\theta}$. By direct integration,

$$\mathrm{E}\,X = \int_0^\infty dy_1\int_0^\infty dy_2\,\sqrt{y_1y_2}\,f_{Y_1,Y_2}(y_1,y_2) = \frac{1}{\theta^2}\int_0^\infty dy_1\,\sqrt{y_1}\,e^{-y_1/\theta}\int_0^\infty dy_2\,\sqrt{y_2}\,e^{-y_2/\theta}.$$

The integrals are related to the $\Gamma(\alpha = 3/2, \beta = \theta)$ density, hence each evaluate to $\beta^\alpha\Gamma(\alpha)$. The result: $\mathrm{E}\,X = \Gamma^2(3/2)\theta = (\pi/4)\theta$ and $\hat\theta = 4\sqrt{Y_1Y_2}/\pi$ is an unbiased estimator of θ.

(b) Similar to (a) except there are four independent integrals: $\hat\theta^2 = 16\sqrt{Y_1Y_2Y_3Y_4}/\pi^2$.

8.3 Some Common Unbiased Point Estimators

8.4 Evaluating the Goodness of a Point Estimator

E8.21 Internet user sample. $\overline{Y}$ is an unbiased estimator of μ, hence $\hat\mu = \overline{Y}$. Its standard error $\sigma_{\hat\mu} = \sigma/\sqrt{n}$ can be approximated using the sample standard deviation: $\sigma_{\hat\mu} \approx S/\sqrt{n}$. By

Tchebysheff's theorem, $P(|\hat{\mu}-\mu| < 2\sigma_{\hat{\mu}}) = P(\mu - 2\sigma_{\hat{\mu}} < \hat{\mu} < \mu + 2\sigma_{\hat{\mu}}) \geq 0.75$. Noting that $P(\mu - 2\sigma_{\hat{\mu}} < \hat{\mu} < \mu + 2\sigma_{\hat{\mu}}) = P(\hat{\mu} - 2\sigma_{\hat{\mu}} < \mu < \hat{\mu} + 2\sigma_{\hat{\mu}})$, the interval $[\hat{\mu} - 2\sigma_{\hat{\mu}}, \hat{\mu} + 2\sigma_{\hat{\mu}}]$ will contain the true mean with probability ≥ 0.75. Note that the interval has random variable endpoints. Our specific realization of this random interval $[\overline{y} - 2s/\sqrt{n}, \overline{y} + 2s/\sqrt{n}] \approx [10.0, 12.0]$ either contains or does not contain μ, but we can attach an at least 0.75 confidence level.

Comment. Invoking the central limit theorem, the sample mean $\overline{Y} \approx \mathcal{N}(\mu, \sigma^2/n) \approx \mathcal{N}(\mu, \hat{\sigma}^2_{\hat{\mu}})$, therefore $P(|\hat{\mu}-\mu| < 2\sigma_{\hat{\mu}}) = P(\mu - 2\sigma_{\hat{\mu}} < \hat{\mu} < \mu + 2\sigma_{\hat{\mu}}) \approx 0.95 = P(\hat{\mu} - 2\sigma_{\hat{\mu}} < \mu < \hat{\mu} + 2\sigma_{\hat{\mu}})$. Assuming the CLT is a good approximation, our interval $[10.0, 12.0]$ has a ≈ 0.95 confidence level.

E8.23 Possible effects of trace elements in drinking water on kidney-stone disease.

(a) Similar to E8.21, $P(\hat{\mu} - 2\sigma_{\hat{\mu}} < \mu < \hat{\mu} + 2\sigma_{\hat{\mu}}) \geq 0.75$. The corresponding interval is $[9.76, 12.8]$.

(b) Let $\theta = \mu_1 - \mu_2$. $\hat{\theta} = \overline{Y}_1 - \overline{Y}_2$ with $\operatorname{var}\hat{\theta} = \operatorname{var}(\overline{Y}_1 - \overline{Y}_2) = \operatorname{var}\overline{Y}_1 + \operatorname{var}\overline{Y}_2 = \sigma_1^2/n_1 + \sigma_2^2/n_2$, assuming independence of Y_i, and thus $\sigma_{\hat{\theta}} = \sqrt{\sigma_1^2/n_1 + \sigma_2^2/n_2} \approx \sqrt{S_1^2/n_1 + S_2^2/n_2}$. Similar to E8.21, by Tchebysheff's theorem, $P(\hat{\theta} - 2\sigma_{\hat{\theta}} < \theta < \hat{\theta} + 2\sigma_{\hat{\theta}}) \geq 0.75$. Letting $1 \leftrightarrow$ Carolinas and $2 \leftrightarrow$ Rockies, the corresponding interval is $[-3.00, 0.404]$.

(c) Let $\theta = p_1 - p_2$. $\hat{\theta} = Y_1/n_1 - Y_2/n_2 = \hat{p}_1 - \hat{p}_2$, where each $Y_i \sim \text{binomial}(n_i, p_i)$, with $\operatorname{var}\hat{\theta} = \operatorname{var}(Y_1/n_1 - Y_2/n_2) = \operatorname{var}(Y_1)/n_1^2 + \operatorname{var}(Y_2)/n_2^2 = p_1(1-p_1)/n_1 + p_2(1-p_2)/n_2$, assuming independence of Y_i, and thus $\sigma_{\hat{\theta}} \approx \sqrt{\hat{p}_1(1-\hat{p}_1)/n_1 + \hat{p}_2(1-\hat{p}_2)/n_2}$. Similar to E8.21, by Tchebysheff's theorem, $P(\hat{\theta} - 2\sigma_{\hat{\theta}} < \theta < \hat{\theta} + 2\sigma_{\hat{\theta}}) \geq 0.75$. Letting $1 \leftrightarrow$ Carolinas and $2 \leftrightarrow$ Rockies, the corresponding interval is $[0.0897, 0.250]$.

E8.29 Baseball fan poll. We want to estimate $\theta = p_1 - p_2$; let $1 \leftrightarrow$ November and $2 \leftrightarrow$ March. From E8.23(c), assuming the two polls constitute independent samples, $\hat{\theta} = 6\%$ and $\sigma_{\hat{\theta}} \approx 2.23\%$. $\hat{\theta}$ is $\approx 2.69\sigma_{\hat{\theta}}$ away from zero, so there is only mild evidence fan support is greater in November.

E8.34 Poisson sample estimator. $\hat{\lambda} = \overline{Y}$ is an unbiased estimator: $\operatorname{E}\hat{\lambda} = \lambda$, with $\operatorname{var}\lambda = \operatorname{var}\overline{Y} = \lambda/n$, hence standard error $\sigma_{\hat{\lambda}} = \sqrt{\lambda/n}$. However, λ is unknown (we are trying to estimate it after all), but we can estimate it as $\sigma_{\hat{\lambda}} \approx \sqrt{\hat{\lambda}/n}$.

E8.36 Exponential sample estimator. Similar to E8.34, $\hat{\theta} = \overline{Y}$ with $\sigma_{\hat{\theta}} = \theta/\sqrt{n} \approx \hat{\theta}/\sqrt{n}$.

E8.38 Geometric sample estimator.

(a) We seek an estimator $\hat{\sigma}^2(Y)$ such that $\operatorname{E}\hat{\sigma}^2 = \operatorname{var}Y = (1-p)/p^2 = 1/p^2 - 1/p$. Since $\operatorname{E}Y^2 = \operatorname{var}Y + (\operatorname{E}Y)^2 = 2/p^2 - 1/p$, the estimator $\hat{\sigma}^2 = Y^2/2 - Y/2$ is unbiased.

(b) $\hat{\mu} = Y$ is an unbiased estimator of $\mu = 1/p$ with $\operatorname{var}\hat{\mu} = \operatorname{var}Y$, hence standard error $\sigma_{\hat{\mu}} \approx \hat{\sigma}$.

8.5 Confidence Intervals

E8.40 Normal distribution confidence intervals. Consider the general case of an observation from $Y \sim \mathcal{N}(\mu, \sigma^2)$ with only μ unknown. The transformation $Z = (Y - \mu)/\sigma \sim \mathcal{N}(0,1)$ and is therefore a pivotal quantity, allowing us to use the pivotal method.

(a) We seek $\hat{\mu}_i$ such that $P(\hat{\mu}_L < \mu < \hat{\mu}_U) = 0.95$. Using Z, we seek constants a, b such that $P(a \leq Z \leq b) = 0.95$. Once a, b are determined, a confidence interval for μ is constructed by $P(a \leq Z \leq b) = P(a \leq (Y-\mu)/\sigma \leq b) = P(Y - a\sigma \geq \mu \geq Y - b\sigma) = 0.95$; evidently $\hat{\mu}_L = Y - b\sigma$ and $\hat{\mu}_U = Y - a\sigma$. One such pair of constants satisfies $P(Z < a) = P(Z > b) = 0.025$ resulting in $(a, b) \approx (-2, 2)$, hence the (random) interval $I = Y \pm 2\sigma$ where $\mu \in I$ with probability 0.95.

(b) Similar to (a), we seek $P(\mu < \hat{\mu}_U) = 0.95$, or constant c such that $P(Z < c) = 0.05$, satisfied with $c \approx -1.64$, so that $0.05 = P((Y-\mu)/\sigma < c) = P(\mu > Y - c\sigma) = 1 - P(\mu < Y - c\sigma)$, or $P(\mu < Y - c\sigma) = 0.95$ as desired. Thus, the corresponding interval is $(-\infty, Y - c\sigma]$.

(c) Similar to (b), we seek $P(\hat{\mu}_L \leq \mu) = 0.95$. The corresponding interval is $[Y + c\sigma, \infty)$.

E8.43 Uniform sample confidence interval. $Y \sim \text{uniform}(0,\theta)$ has pdf $f_Y(y) = 1/\theta$ and cdf $F_Y(y) = y/\theta$, $0 \le y \le \theta$. By Theorem 6.5, $f_{(n)}(y) = ny^{n-1}/\theta^n$, $0 \le y \le \theta$.

(a) The transformation $U = Y_{(n)}/\theta$ is one-to-one over $0 \le y \le \theta$ with inverse $Y_{(n)} = \theta U$, hence $f_U(u) = f_Y(\theta u)|dy_{(n)}/du| = nu^{n-1}$, $0 \le \theta u \le \theta$ or $0 \le u \le 1$, with $F_U(u) = \int_0^u dx\, f_U(x) = u^n$, $0 \le u \le 1$ and equal to 0 (1) for $u \le 1$ ($u \ge 1$).

(b) We seek a lower bound $\hat{\theta}_L$ such that the interval $P(\hat{\theta}_L \le \theta) = 0.95$. Since U is a pivotal quantity, for some constant c, $P(c \le U) = P(c \le Y_{(n)}/\theta) = P(\theta \le Y_{(n)}/c) = 1 - P(Y_{(n)}/c \le \theta)$. Therefore $\hat{\theta}_L = Y_{(n)}/c$ and $P(\hat{\theta}_L \le \theta) = 1 - P(c \le U) = 0.95$ for $P(c \le U) = 1 - F_U(c) = 0.05$, or $c = 0.95^{1/n}$. The corresponding interval is $[Y_{(n)}/c, \infty)$

Comment. Note $Y_{(n)}/c > Y_{(n)}$, which is intuitive because $P(\theta \ge Y_{(n)}) = 1$.

E8.46 Exponential confidence interval using a single observation.
$Y \sim \text{exponential}(\theta)$ has mgf $m_Y(t) = 1/(1-\theta t)$.

(a) Let $U = 2Y/\theta$. $m_U(t) = \mathrm{E}\, e^{tU} = \mathrm{E}\, e^{(2t/\theta)Y} = m_Y(2t/\theta) = 1/(1-\beta t)$, which is the mgf of a $\chi^2(\nu = 2)$ distribution. By uniqueness of mgfs, $U \sim \chi^2(2)$. U is a function of Y and independent of θ, thus is a pivotal quantity.

(b) We seek $\hat{\theta}_i$ such that $P(\hat{\theta}_L < \theta < \hat{\theta}_U) = 0.90$. We can relate this to U for constants $a, b > 0$ as $P(a \le U \le b) = P(a \le 2Y/\theta \le b) = P(2Y/a \ge \theta \ge 2Y/b)$. Thus $\hat{\theta}_L = 2Y/b$ and $\hat{\theta}_U = 2Y/a$, where the constants can be determined by setting $P(a \le U \le b) = 0.9$. One such pair of constants satisfies $P(U \le a) = P(U \ge b) = 0.05$. The result: $a \approx 0.103$ and $b \approx 5.99$. The corresponding interval is approximately $[0.334Y, 19.4Y]$, the same as Example 8.4.

E8.47 Exponential confidence interval using a random sample.
$Y \sim \text{exponential}(\theta)$ has mgf $m_Y(t) = 1/(1-\theta t)$.

(a) Let $U = (2/\theta)\sum_{i=1}^n Y_i$. $m_U(t) = \mathrm{E}\, e^{tU} = \mathrm{E}\, e^{(2t/\theta)\sum_{i=1}^n Y_i} = \mathrm{E}\, e^{(2t/\theta)\sum_{i=1}^n Y_i}$. By independence of Y_i, $m_U(t) = [m_Y(2t/\theta)]^n = 1/(1-\beta t)^n$, which is the mgf of a $\chi^2(\nu = 2n)$ distribution. By uniqueness of mgfs, $U \sim \chi^2(2n)$. U is a function of Y_i and independent of θ, thus is a pivotal quantity.

(b) We seek $\hat{\theta}_i$ such that $P(\hat{\theta}_L < \theta < \hat{\theta}_U) = 0.95$. We can relate this to U for constants $a, b > 0$ as $P(a \le U \le b) = P(a \le 2n\overline{Y}/\theta \le b) = P(2n\overline{Y}/a \ge \theta \ge 2n\overline{Y}/b)$. Thus $\hat{\theta}_L = 2n\overline{Y}/b$ and $\hat{\theta}_U = 2n\overline{Y}/a$, where the constants can be determined by setting $P(a \le U \le b) = 0.95$. One such pair of constants satisfies $P(U \le a) = P(U \ge b) = 0.025$.

(c) $(\hat{\theta}_L, \hat{\theta}_U) \approx (2.56, 11.9)$

8.6 Large-Sample Confidence Intervals

E8.57 Baseball fan poll 2. Let Y be the number of adults who are fans in November 2003 and model $Y \sim \text{binomial}(n, p)$ with $n = 1001$ and p unknown. To estimate p, we will use estimator $\hat{p} = Y/n$, which is unbiased, i.e. $\mathrm{E}\,\hat{p} = (\mathrm{E}\, Y)/n = p$, with variance $\mathrm{var}\,\hat{p} = (\mathrm{var}\, Y)/n^2 = p(1-p)/n$, hence standard error $\sigma_{\hat{p}} = \sqrt{p(1-p)/n}$.

To construct a $(1-\alpha)$ confidence interval for p, we must find endpoints $\hat{p}_L$ and $\hat{p}_U$ such that $P(\hat{p}_L \le p \le \hat{p}_U) = 1 - \alpha$. We will do this approximately using the pivotal method. Since $Y = \sum_{i=1}^n X_i$, $X_i \sim \text{Bernoulli}(p)$, and n is large, we can invoke the CLT and approximate $Y \approx \mathcal{N}(n\,\mathrm{E}\, X, n \,\mathrm{var}\, X) = \mathcal{N}[np, np(1-p)]$, hence $\hat{p} = Y/n \approx \mathcal{N}[p, p(1-p)/n] = \mathcal{N}(p, \sigma_{\hat{p}}^2)$. It follows that $Z = (\hat{p} - p)/\sigma_{\hat{p}} \sim \mathcal{N}(0,1)$ is a pivotal quantity if we approximate $\sigma_{\hat{p}} \approx \sqrt{\hat{p}(1-\hat{p})/n}$ (since p is unknown). For constants a, b, we can then construct the interval of interest $[\hat{p}_L, \hat{p}_U]$ using $P(a \le Z \le b) = P[a \le (\hat{p}-p)/\sigma_{\hat{p}} \le b] = P[\hat{p} - a\sigma_{\hat{p}} \ge p \ge \hat{p} - b\sigma_{\hat{p}}] = 1 - \alpha$, identifying $\hat{p}_L = \hat{p} - b\sigma_{\hat{p}}$ and $\hat{p}_U = \hat{p} - a\sigma_{\hat{p}}$. All that remains is to determine a, b such that $P(a \le Z \le b)$. One construction requires $P(Z \le a) = P(Z \ge b) = \alpha/2$. For $\alpha = 0.01$, $a \approx -2.56$ and $b \approx 2.56$.

Therefore our 0.99 confidence interval is $[\hat{p}_L, \hat{p}_U] \approx [0.469, 0.551]$. Our realization of the interval, $[0.469, 0.551]$, either contains p or does not. However, the process used to generate the (random) interval $[\hat{p}_L, \hat{p}_U]$ if repeated on many samples will result in $\approx 0.99\%$ intervals that contain p. Thus, we can be reasonably confident p is contained in our interval, but precise probability statements can no longer be made.

E8.60 Confidence interval of body temperatures.

(a) Let $Y_1, Y_2, \ldots, Y_n$ denote body temperature measurements for each of the n humans. Assuming Y_i constitute an i.i.d. sample from some distribution with unknown mean μ and variance σ^2, we are interested in estimating μ. The sample mean is an unbiased estimator, $\hat{\mu} = \overline{Y}$, with variance $\operatorname{var}\hat{\mu} = \operatorname{var}\overline{Y} = \operatorname{var} Y/n$, hence standard error $\sigma_{\hat{\mu}} = \sigma/\sqrt{n}$.

To construct a confidence interval, we must know something about the sampling distribution of $\hat{\mu}$ so we can make a probabilistic statement over some interval $P(\hat{\mu}_L \leq \mu \leq \hat{\mu}_U) = 1 - \alpha$ for to be determined $\hat{\mu}_i$. Invoking the CLT, $\hat{\mu} \approx \mathcal{N}(\mu, \sigma^2_{\hat{\mu}})$. Further approximating $\sigma^2_{\hat{\mu}} \approx S^2/n$ with the sample variance S^2, we can apply the pivotal method using quantity $Z = (\hat{\mu} - \mu)/\sigma_{\hat{\mu}}$ to construct a confidence interval as follows. Given constants a, b such that $P(a \leq Z \leq b) = 1 - \alpha$, we can manipulate $P(a \leq Z \leq b) = P[a \leq (\hat{\mu} - \mu)/\sigma_{\hat{\mu}} \leq b] = P(\hat{\mu} - a\sigma_{\hat{\mu}} \geq \mu \geq \hat{\mu} - b\sigma_{\hat{\mu}})$ and identify $[\hat{\mu}_L, \hat{\mu}_U] = [\hat{\mu} - b\sigma_{\hat{\mu}}, \hat{\mu} - a\sigma_{\hat{\mu}}]$ as a $(1 - \alpha)$ confidence interval. One way to determine a, b is to require $P(Z \leq a) = P(Z \geq b) = \alpha/2$. For $\alpha = 0.01$, $a \approx -2.56$ and $b \approx 2.56$. The associated confidence interval is approximately $[98.1, 98.4]$.

(b) Our interval does not contain the accepted value 98.6. We cannot conclude $\mu \neq 98.6$ because either μ is in our interval is not. However, if the assumptions we used (i.i.d. sample, CLT, sample variance approximation) are valid, then we can be fairly confident our interval contains μ. Ideally another group would repeat the experiment and construct a confidence interval in the same way. In repeated sampling, 99% of intervals constructed this way will contain μ.

***E8.68 Multinomial confidence intervals.** Each Y_i can be written as a sum of Bernoulli r.v., hence the CLT may apply for large n_i. Linear combinations of normal r.v. are also normal, so linear combinations of Y_i are approximately normal.

(a) $\operatorname{var}(Y_1 - Y_2) = \operatorname{var} Y_1 + \operatorname{var} Y_2 - 2\operatorname{cov}(Y_1, Y_2) = np_1(1 - p_1) + np_2(1 - p_2) + 2np_1p_2$.

(b) $\hat{\theta} = (Y_1 - Y_2)/n = \hat{p}_1 - \hat{p}_2$ with $\operatorname{var}\hat{\theta} = \operatorname{var}[(Y_1 - Y_2)/n)] = (p_1(1 - p_1) + p_2(1 - p_2) + 2p_1p_2)/n$, hence s.e. $\sigma_{\hat{\theta}} = \sqrt{(p_1(1 - p_1) + p_2(1 - p_2) + 2p_1p_2)/n} \approx (\hat{p}_1(1 - \hat{p}_1) + \hat{p}_2(1 - \hat{p}_2) + 2\hat{p}_1\hat{p}_2)/n$. Invoking the CLT, $\hat{\theta} \approx \mathcal{N}(p_1 - p_2, \sigma^2_{\hat{\theta}})$ and $Z = [\hat{\theta} - (p_1 - p_2)]/\sigma_{\hat{\theta}} \approx \mathcal{N}(0, 1)$. Using the pivotal method, $P(\hat{\theta} - z_{\alpha/2}\sigma_{\hat{\theta}} \leq p_1 - p_2 \leq \hat{\theta} + z_{\alpha/2}\sigma_{\hat{\theta}}) = 1 - \alpha$, hence a $(1 - \alpha)$ two-sided confidence interval is $[\hat{\theta} - z_{\alpha/2}\sigma_{\hat{\theta}}, \hat{\theta} + z_{\alpha/2}\sigma_{\hat{\theta}}]$. With $\alpha = 0.05$ and given values, we have $[-0.140, -0.0598]$.

***E8.69 Independent binomial confidence intervals.** We want to estimate $\theta = (p_3 - p_1) - (p_4 - p_2)$. $\hat{\theta} = (Y_3/n_3 - Y_1/n_1) - (Y_4/n_4 - Y_2/n_2) = (\hat{p}_3 - \hat{p}_1) - (\hat{p}_4 - \hat{p}_2)$. By independence of Y_i, the s.e. $\sigma_{\hat{\theta}} = \sqrt{[p_3(1 - p_3)/n_3 + p_1(1 - p_1)/n_1] + [p_4(1 - p_4)/n_4 + p_2(1 - p_2)/n_2]}$. Under assumptions where each n_i is large enough so the CLT applies, $Y_i/n_i \approx \mathcal{N}[p_i, p_i(1 - p_i)/n_i]$, hence $\hat{\theta} \approx \mathcal{N}(\theta, \sigma^2_{\hat{\theta}})$. Using the pivotal method as usual, a two-sided $(1 - \alpha)$ confidence interval is $[\hat{\theta} - z_{\alpha/2}\sigma_{\hat{\theta}}, \hat{\theta} + z_{\alpha/2}\sigma_{\hat{\theta}}]$.

8.7 Selecting the Sample Size

E8.70 Binomial sample size. We want our estimate $\hat{p} = Y/n$ to satisfy $P(\hat{p} - 0.05 \leq p \leq \hat{p} + 0.05) \geq 0.95$. This is equivalent to a two-sided confidence interval $\hat{p} \pm 0.05$ at the $(1 - \alpha) \geq 0.95$ confidence level, or $\alpha \leq 0.05$. Under the large-sample approximation (e.g. invoking the CLT as in Section 8.6 for the distribution of $\hat{p}$), the confidence interval is $\hat{p} \pm z_{\alpha/2}\sigma_{\hat{p}}$ where $\sigma_{\hat{p}} = \sqrt{p(1 - p)/n}$. Often we further approximate $\sigma_{\hat{p}} \approx \sqrt{\hat{p}(1 - \hat{p})/n}$, thus we want to choose n such that $z_{0.05/2}\sqrt{\hat{p}(1 - \hat{p})/n} \leq 0.05$, or

$n \geq (z_{0.05/2})^2 \hat{p}(1-\hat{p})/(0.05)^2$.
(a) Using the exact $p = 0.9$ yields $n \geq 138.3$.
(b) Using $p = 0.5$ yields $n \geq 384.1$.

E8.71 Sample size for a mean. We want our estimate $\hat{\mu} = \overline{Y}$ to satisfy $P(\hat{\mu}-2 \leq \mu \leq \hat{\mu}+2) \geq (1-\alpha)$. This is equivalent to a two-sided confidence interval $\hat{\mu} \pm 2$ at the $(1-\alpha)$ confidence level or greater. Under the large-sample approximation, the confidence interval is $\hat{\mu} \pm z_{\alpha/2}\sigma_{\hat{\mu}}$ where $\sigma_{\hat{\mu}} = \sigma/\sqrt{n}$. Thus we want to choose n such that $z_{\alpha/2}\sigma/\sqrt{n} \leq 2$, or $n \geq (z_{\alpha/2})^2\sigma^2/4$.

8.8 Small-Sample Confidence Intervals for μ and $\mu_1 - \mu_2$

***E8.93 Confidence interval for linear combination of means from independent normal samples.** Consider the general case of estimating $\theta = a\mu_1 + b\mu_2$ where $X \sim \mathcal{N}(\mu_1, \sigma_1^2)$ and $Y \sim \mathcal{N}(\mu_2, \sigma_2^2)$. An unbiased estimator is $\hat{\theta} = a\overline{X} + b\overline{Y}$. By the normality assumption, $\overline{X} \sim \mathcal{N}(\mu_1, \sigma_1^2/n)$ and $\overline{Y} \sim \mathcal{N}(\mu_2, \sigma_2^2/m)$, hence from Theorem 6.3, $\hat{\theta} \sim \mathcal{N}(a\mu_1 + b\mu_2, a^2\sigma_1^2/n + b^2\sigma_2^2/m)$.

(a) With variances known, $Z = (\hat{\theta} - \theta)/\sigma_{\hat{\theta}}$ is a pivotal quantity. A two-sided $(1-\alpha)$ confidence interval is therefore $[\hat{\theta} + z_{1-\alpha/2}\sigma_{\hat{\theta}}, \hat{\theta} + z_{\alpha/2}\sigma_{\hat{\theta}}]$, or equivalently $\hat{\theta} \pm z_{\alpha/2}\sigma_{\hat{\theta}}$.

(b) With unknown but equal variances, $T = (\hat{\theta} - \theta)/(S_p\sqrt{a^2/n + b^2/m})$ is a pivotal quantity where S_p is the pooled sample variance. A two-sided $(1-\alpha)$ confidence interval is therefore $\hat{\theta} \pm t_{\alpha/2} S_p \sqrt{a^2/n + b^2/m}$.

In the general case of possibly unequal variance, $T = (\hat{\theta} - \theta)/(\sqrt{a^2S_1^2/n + b^2S_2^2/m})$ is approximately T-distributed with

$$\nu \approx \frac{(a^2S_1^2/n_1 + b^2S_2^2/n_2)^2}{(a^2S_1^2/n_1)^2/(an_1 - 1) + (bS_2^2/n_2)^2/(bn_2 - 1)}$$

degrees of freedom (**Welch–Satterthwaite equation**), allowing us to construct an approximate $(1-\alpha)$ confidence interval as $\hat{\theta} \pm t_{\alpha/2}\sqrt{a^2S_1^2/n + b^2S_2^2/m}$.

Comment. This argument generalizes for linear combinations of means from $l > 2$ normal samples, using the Welch–Satterthwaite equation for l samples.

E8.94 Difference in means upper bounded confidence interval. We seek an interval of the form $P(\theta \leq \hat{\theta}_U) = 1 - \alpha$. Using pivotal quantity $T = (\hat{\theta} - \theta)/S_p\sqrt{1/n_1 + 1/n_2}$, for constant c we have $P(T \geq c) = P[(\hat{\theta} - \theta)/S_p\sqrt{1/n_1 + 1/n_2} \geq c] = P[\theta \leq \hat{\theta} - cS_p\sqrt{1/n_1 + 1/n_2}]$. Evidently $\hat{\theta}_U = \hat{\theta} - cS_p\sqrt{1/n_1 + 1/n_2}$ with constant $c = -t_\alpha$ such that $P(T \geq c) = 1 - \alpha$.

8.9 Confidence Intervals for σ^2

8.10 Summary

Confidence intervals for the population mean μ. In the following cases, we will consider an i.i.d. random sample $Y_1, Y_2, \ldots, Y_n$ from some distribution Y with unknown mean μ and possibly known standard deviation σ. Any other assumptions on Y will be explicitly stated.

- **Large sample, μ unknown and σ known.** The sample mean $\hat{\mu} = \overline{Y}$ is an unbiased estimator of μ with standard error $\sigma_{\hat{\mu}} = \sigma/\sqrt{n}$. Invoking the CLT, $\overline{Y} \approx \mathcal{N}(\mu, \sigma_{\hat{\mu}}^2)$ thus $Z = (\overline{Y} - \mu)/\sigma_{\hat{\mu}} \approx \mathcal{N}(0,1)$ is a (approximately) pivotal quantity. Using the pivotal method, a $(1-\alpha)$ confidence interval for μ can be constructed by algebraic manipulation of $P(a \leq Z \leq b)$ for constants a, b. A typical two-sided interval is $\hat{\mu} \pm z_{\alpha/2}\sigma_{\hat{\mu}}$ where the constant $z_{\alpha/2}$ satisfies $P(Z \leq z_{\alpha/2}) = 1 - \alpha/2$.

- **Large sample, μ and σ unknown.** If σ is unknown, $Z = (\overline{Y} - \mu)/\sigma_{\hat{\mu}}$ is no longer pivotal. However by Slutsky's theorem, the sample standard deviation S converges to σ, hence a typical two-sided interval is again $\hat{\mu} \pm z_{\alpha/2}\sigma_{\hat{\mu}}$ where $\sigma_{\hat{\mu}}$ is approximated by $\sigma_{\hat{\mu}} \approx S/\sqrt{n}$.

- **Small sample, $Y \sim \mathcal{N}(\mu, \sigma^2)$, μ unknown and σ known.** In a small sample we cannot reasonably invoke the CLT and must have simplifying distributional assumptions on Y. Normality of Y implies $\overline{Y} \sim \mathcal{N}(\mu, \sigma^2_{\hat{\mu}})$ (exactly), hence we can use pivotal quantity $Z = (\overline{Y} - \mu)/\sigma_{\hat{\mu}} \sim \mathcal{N}(0,1)$, resulting in (exact) confidence interval $\hat{\mu} \pm z_{\alpha/2}\sigma_{\hat{\mu}}$.

- **Small sample, $Y \sim \mathcal{N}(\mu, \sigma^2)$, μ and σ unknown.** If σ is unknown, $Z = (\overline{Y} - \mu)/\sigma_{\hat{\mu}}$ is no longer pivotal. However, $T = (\overline{Y} - \mu)/(S/\sqrt{n})$ is t-distributed with $(n-1)$ d.f. and thus a pivotal quantity. The pivotal method results in a typical two-sided confidence interval $\overline{Y} \pm t_{\alpha/2}S/\sqrt{n}$ where the constant $t_{\alpha/2}$ satisfies $P(T \leq t_{\alpha/2}) = 1 - \alpha/2$.

  Comment. The text states that the interval $\overline{Y} \pm t_{\alpha/2}S/\sqrt{n}$ is robust even under moderate departures of Y from normality.

Confidence intervals for difference in population means $\theta = \mu_X - \mu_Y$. In the following cases, we will consider two independent i.i.d. random samples $X_1, X_2, \ldots, X_{n_X}$ and $Y_1, Y_2, \ldots, Y_{n_Y}$ from some distributions X and Y with unknown means and standard deviations. Distributional assumptions are needed for the small sample limit.

- **Large samples.** $\hat{\theta} = \overline{X} - \overline{Y}$ is an unbiased estimator. In the large n_X and n_Y limit, $\overline{X} \approx \mathcal{N}(\mu_X, \sigma^2_X/n_X)$ and $\overline{Y} \approx \mathcal{N}(\mu_Y, \sigma^2_Y/n_Y)$, hence $\hat{\theta} \approx \mathcal{N}(\theta, \sigma^2_X/n_X + \sigma^2_Y/n_Y)$ by independence of samples. It follows that $Z = (\hat{\theta} - \theta)/\sqrt{\sigma^2_X/n_X + \sigma^2_Y/n_Y}$ is a pivotal quantity and a typical two-sided $(1-\alpha)$ confidence interval for θ is $(\overline{X} - \overline{Y}) \pm z_{\alpha/2}\sqrt{\sigma^2_X/n_X + \sigma^2_Y/n_Y}$. By Slutsky's theorem, this interval can be approximated by replacing population variances with sample variances.

- **Small samples, $X \sim \mathcal{N}(\mu_X, \sigma^2_X)$ and $Y \sim \mathcal{N}(\mu_Y, \sigma^2_Y)$.** Due to the normality assumption, $\hat{\theta} \sim \mathcal{N}(\theta, \sigma^2_X/n_X + \sigma^2_Y/n_Y)$. However in the small sample limit, Slutsky's theorem does not apply. As in the one sample case, we replace population variances with sample variances and examine the distribution of $T = (\hat{\theta} - \theta)/\sqrt{S^2_X/n_X + S^2_Y/n_Y}$. It turns out this quantity is approximately t-distributed with ν degrees of freedom,

$$\nu = \frac{(\sigma^2_X/n_X + \sigma^2_Y/n_Y)^2}{(\sigma^2_X/n_X)^2/(n_X-1) + (\sigma^2_Y/n_Y)^2/(n_Y-1)} \approx \frac{(S^2_X/n_X + S^2_Y/n_Y)^2}{(S^2_X/n_X)^2/(n_X-1) + (S^2_Y/n_Y)^2/(n_Y-1)}.$$

  (For a proof, see: Lloyd, Michael. "2-Sample t-distribution approximation." *Academic Forum.* Vol. 31. 2013.) Therefore T can be used as a pivotal quantity, and we obtain two-sided $(1-\alpha)$ confidence interval for θ as $(\overline{X} - \overline{Y}) \pm t_{\alpha/2}\sqrt{S^2_X/n_X + S^2_Y/n_Y}$. This is related to **Welch's unequal variances *t*-test** and is a generalization of the equal population variance case presented in the text (Section 8.8).

  Comment. The text states that the confidence interval for $\theta = \mu_X - \mu_Y$ is robust even under moderate departures of X and Y from normality.

Confidence interval for the population variance σ^2 of a normal population. Consider an i.i.d. random sample $Y_1, Y_2, \ldots, Y_n$ from distribution $Y \sim \mathcal{N}(\mu, \sigma^2)$, μ and σ unknown. $\hat{\sigma}^2 = S^2$ is an unbiased estimator of σ^2. Furthermore, by Theorem 7.3, $W = (n-1)S^2/\sigma^2 \sim \chi^2(n-1)$ is a pivotal quantity. Using the pivotal method, a typical two-sided $(1-\alpha)$ confidence interval for σ^2 is $[(n-1)S^2/\chi^2_{\alpha/2}, (n-1)S^2/\chi^2_{1-\alpha/2}]$ where the constant χ_c satisfies $P(W \le \chi_c^2) = 1-c$.

Comment. The text states that the confidence interval for σ^2 is not robust under departures of Y from normality.

Confidence interval for ratio of population variances $\boldsymbol{\theta = \sigma_X^2/\sigma_Y^2}$ of normal populations. Consider two independent i.i.d. normal random samples $X_1, X_2, \ldots, X_{n_X}$ and $Y_1, Y_2, \ldots, Y_{n_Y}$ with unknown means and variances. Since $\hat\sigma_X^2 = S_X^2$ and $\hat\sigma_Y^2 = S_Y^2$ are unbiased estimators and are independent from each other, by Theorem 5.9, $\hat\theta = S_X^2/S_Y^2$ is an unbiased estimator of θ. The quantity $F = (S_X^2/S_Y^2)/(\sigma_X^2/\sigma_Y^2) = \hat\theta/\theta$ is F-distributed with (n_X-1) numerator d.f. and (n_Y-1) denominator d.f., hence is a pivotal quantity. A typical two-sided $(1-\alpha)$ confidence interval is then $[\hat\theta/f_{\alpha/2}, \hat\theta/f_{1-\alpha/2}]$ where the constant f_c satisfies $P(F \le f_c) = 1-c$.

Supplemental Exercises

E8.105 Confidence coefficient from a confidence interval. The confidence interval is of the form $\hat\mu \pm z_{\alpha/2}\sigma/\sqrt{n}$. Equating ranges, $(7.37-5.37) = 2z_{\alpha/2}\sigma/\sqrt{n}$, we deduce $z_{\alpha/2} = \sqrt{n}/\sigma$. Since $P(Z \le z_{\alpha/2}) = 1-\alpha/2$, $\alpha = 2[1-P(Z \le z_{\alpha/2})] \approx 0.0412$ and the corresponding confidence coefficient $(1-\alpha) \approx 0.959$.

E8.106 Controlled pollination study survival proportions. Using the large-sample approximation, two-sided $(1-\alpha)$ confidence intervals for $\theta = p_1 - p_2$ are of the form

$$\hat\theta \pm z_{\alpha/2}\sigma_{\hat\theta} \approx (\hat p_1 - \hat p_2) \pm z_{\alpha/2}\sqrt{\hat p_1(1-\hat p_1)/n_1 + \hat p_2(1-\hat p_2)/n_2}$$

where $\hat p_i = Y_i/n_i$. **(a)** $[-0.0403, 0.0508]$ **(b)** $[-0.0557, 0.0344]$

E8.107 Controlled pollination study survival proportions 2. We want an interval such that the probability $P(|\theta - \hat\theta| \le 0.03) = P(\hat\theta - 0.03 \le \theta \le \hat\theta + 0.03) \ge 0.95$. Evidently we want to choose n_i so that $z_{0.05/2}\sigma_{\hat\theta} \le 0.03$, or $\sqrt{\hat p_1(1-\hat p_1)/n_1 + \hat p_2(1-\hat p_2)/n_2} \le 0.0153$. Since $\hat p_i$ are unknown apriori, to be conservative we let $\hat p_i = \frac{1}{2}$ so that $\sigma_{\hat\theta}$ is maximized w.r.t. $\hat p_i$. Setting $n_1 = n_2 = n$, this simplifies to $\sqrt{1/(2n)} \le 0.0153$, or $n \ge 2134.1$.

Comment. If we used $\hat p_i$ in E8.106, the required sample size $n \ge 764.8$ is reduced dramatically.

***E8.127 Gamma sample confidence interval.** Let $Y_1, Y_2, \ldots, Y_n$ denote an i.i.d. random sample from $Y \sim \text{gamma}(\alpha, \beta)$. Recall $\mathrm{E}\,Y = \alpha\beta$ and $\operatorname{var} Y = \alpha\beta^2$. With known α, $\hat\beta = \overline{Y}/\alpha$ is an unbiased estimator of β. Invoking the CLT, $\hat\beta \approx \mathcal{N}[\beta, \beta^2/(n\alpha)]$, hence $U = (\hat\beta - \beta)/(\hat\beta/\sqrt{n\alpha}) \approx \mathcal{N}(0,1)$ is an approximately pivotal quantity. The resultant two-sided $(1-\alpha)$ confidence interval is $\hat\beta(1 \pm z_{\alpha/2}/\sqrt{n\alpha})$.

***E8.129 Variance estimator for a normal random sample.**
By Theorem 7.3, $W = (n-1)S^2/\sigma^2 \sim \chi^2(n-1)$, thus $\operatorname{var} S^2 = [\sigma^2/(n-1)]^2 \operatorname{var} W = 2\sigma^4/(n-1)$. Since $S'^2 = [(n-1)/n]S^2$, $\operatorname{var} S'^2 = [(n-1)/n]^2 \operatorname{var} S^2 < \operatorname{var} S^2$. Explicitly, $S'^2 = 2\sigma^4(n-1)/n^2$.

***E8.130 Variance estimator for a normal random sample 2.** Recall $\mathrm{MSE}(\hat\theta) = \operatorname{var}\hat\theta + (\operatorname{bias}\hat\theta)^2$. Example 8.1 showed S^2 is unbiased, hence $\mathrm{E}\,S'^2 = [(n-1)/n]\sigma^2$ has $\operatorname{bias} S'^2 = \mathrm{E}\,S'^2 - \sigma^2 = -\sigma^2/n$ and $\mathrm{MSE}(S'^2) = \sigma^4[(2n-1)/n^2]$. In comparison, $\mathrm{MSE}(S^2) = 2\sigma^4/(n-1)$. Their difference

$$\mathrm{MSE}(S^2) - \mathrm{MSE}(S'^2) = \left(\frac{2}{n-1} - \frac{2n-1}{n^2}\right)\sigma^4 = \frac{3n-1}{n^2(n-1)}\sigma^4 > 0$$

for $n > 1$, hence $\mathrm{MSE}(S^2) > \mathrm{MSE}(S'^2)$. S^2 has the smaller MSE, despite being biased.

***E8.131 Variance estimator for a normal random sample 3.** Denote $S_*^2 = c\sum_{i=1}^n (Y_i - \overline{Y})^2 = c(n-1)S^2$ for undetermined constant c. It follows that $\mathrm{E}\,S_*^2 = c(n-1)\sigma^2$, $\operatorname{bias} S_*^2 = [c(n-1)-1]\sigma^2$,

$\operatorname{var} S_*^2 = 2c^2(n-1)\sigma^4$, and $\operatorname{MSE}(S_*^2) = \{2c^2(n-1) + [c(n-1)-1]^2\}\sigma^4$. Ordinary calculus shows that $c = 1/(n+1)$ minimizes $\operatorname{MSE}(S_*^2)$.

***E8.133 Pooled variance estimator.**

(a) $\operatorname{E} S_i^2 = \sigma^2$, therefore it follows $\operatorname{E} S_p^2 = \sigma^2$.

(b) $\operatorname{var} S_i^2 = 2\sigma^4/(n_i - 1)$. By independence of the two samples, the covariance is zero and direct calculation yields $\operatorname{var} S_p^2 = 2\sigma^4/(n_1 + n_2 - 2)$.

Comment. The pooled estimator is unbiased and has a smaller variance than either S_i^2 alone (with effective sample size $n_1 + n_2 + 1$).

***E8.134 Expected width of small normal sample confidence interval for $\boldsymbol{\mu}$.** With σ unknown, the small-sample interval is $\overline{Y} \pm t_{\alpha/2} S/\sqrt{n}$ with width $W = 2t_{\alpha/2} S/\sqrt{n}$. Since S is a r.v., we need to compute $\operatorname{E} S$. From Theorem 7.3, $U = (n-1)S^2/\sigma^2 \sim \chi^2(n-1)$. Transforming again, $V = \sqrt{U} = (\sqrt{n-1}/\sigma)S$. Evidently $\operatorname{E} S = (\sigma/\sqrt{n-1}) \operatorname{E} V$. Finally, note $\operatorname{E} V = \operatorname{E} \sqrt{U}$, hence using the $\chi^2(k = n-1)$ pdf,

$$\operatorname{E} V = \frac{1}{2^{k/2}\Gamma(k/2)} \int_0^\infty du\, \sqrt{u} u^{k/2-1} e^{-u/2} = \frac{1}{2^{k/2}\Gamma(k/2)} \int_0^\infty du\, u^{(k+1)/2-1} e^{-u/2}.$$

The integral is related to a gamma density with $\alpha = (k+1)/2$ and $\beta = 2$ hence evaluates to $\Gamma(\alpha)\beta^\alpha$. Simplifying yields $\operatorname{E} V = \sqrt{2}\,\Gamma(n/2)/\Gamma[(n-1)/2]$, thus $\operatorname{E} W = [2t_{\alpha/2}\sigma/\sqrt{n(n-1)}] \operatorname{E} V$.

***E8.135 Biased variance confidence interval of a normal sample.**
The interval is $[(n-1)S^2/\chi^2_{\alpha/2}, (n-1)S^2/\chi^2_{1-\alpha/2}]$ with midpoint $M = \frac{1}{2}(n-1)S^2(1/\chi^2_{\alpha/2} + 1/\chi^2_{\alpha/2})$ and expected value $\operatorname{E} M = \frac{1}{2}(n-1)(1/\chi^2_{\alpha/2} + 1/\chi^2_{\alpha/2})\sigma^2$, which in general does not equal σ^2.

***E8.136 Prediction interval for a normal population.** We are interested in an interval of the form $P(\hat{Y}_L \le Y_p \le \hat{Y}_U) = 1 - \alpha$ where, for a normal population, $Y_p \sim \mathcal{N}(\mu, \sigma^2)$. We know the sample mean $\overline{Y} \sim \mathcal{N}(\mu, \sigma^2/n)$. Therefore assuming independence, $(Y_p - \overline{Y}) \sim \mathcal{N}[0, \sigma^2(1 + 1/n)]$. It follows that

$$T = \frac{Y_p - \overline{Y}}{S\sqrt{1 + 1/n}} = \frac{(Y_p - \overline{Y})/(\sigma\sqrt{1 + 1/n})}{\sqrt{[(n-1)S^2/\sigma^2]/(n-1)}} \simeq \frac{Z}{\sqrt{W/\nu}}$$

is t-distributed with $\nu = (n-1)$ d.f., since $Z \sim \mathcal{N}(0,1)$ and $W \sim \chi^2(\nu)$ (Definition 7.2). Therefore T is a pivotal quantity for Y_p. Applying the pivotal method, a two-sided $(1-\alpha)$ prediction interval is $\overline{Y} \pm t_{\alpha/2} S\sqrt{1 + 1/n}$ where $t_{\alpha/2}$ satisfies $P(T \le t_{\alpha/2}) = 1 - \alpha/2$.

Comment. Recall the confidence interval for μ is $\mathrm{CI} = \overline{Y} \pm t_{\alpha/2} S/\sqrt{n}$ satisfying $P(\mu \in \mathrm{CI}) = 1 - \alpha$. Rearranging the prediction interval, $\mathrm{PI} = \overline{Y} \pm t_{\alpha/2} S\sqrt{(n+1)/n}$ satisfying $P(Y_p \in \mathrm{PI}) = 1 - \alpha$, we see that for the same $(1-\alpha)$ level, the PI is wider by a factor of $\sqrt{n+1}$.

---

9 Properties of Point Estimators and Methods of Estimation

9.1 Introduction

9.2 Relative Efficiency

***E9.8 Efficient estimators using the Cramer–Rao inequality.**

(a) $f(Y) = e^{(Y-\mu)^2/(2\sigma^2)}/\sqrt{2\pi\sigma^2}$. $\log f(Y) = (Y-\mu)^2/(2\sigma^2) + g(\sigma)$, thus $\partial^2 \log f(Y)/\partial\mu^2 = 1/\sigma^2$ is a constant and $I(\mu) = \sigma^2/n$. The (unbiased) estimator $\overline{Y}$ has $\operatorname{var}\overline{Y} = \sigma^2$, thus $\overline{Y}$ is efficient.

(b) $f(Y) = \lambda^Y e^{-\lambda}/Y!$. $\log f(Y) = Y\log\lambda - \lambda - g(Y)$, thus $\partial^2 \log f(Y)/\partial\lambda^2 = -Y/\lambda^2$ and $I(\lambda) = \lambda/n$. The (unbiased) estimator $\overline{Y}$ has $\operatorname{var}\overline{Y} = \lambda/n$, thus $\overline{Y}$ is efficient.

9.3 Consistency

E9.17 Consistent estimator for $\boldsymbol{\theta = \mu_1 - \mu_2}$. $\hat\theta = \overline{X} - \overline{Y}$ is unbiased. $\operatorname{var}\hat\theta = \operatorname{var}\overline{X} + \operatorname{var}\overline{Y}$ by independence, thus $\operatorname{var}\hat\theta = (\sigma_1^2 + \sigma_2^2)/n$ tends to zero as $n \to \infty$. By Theorem 9.1, $\hat\theta$ is consistent.

E9.18 Consistent estimator for pooled sample variance. We can express the given quantity as $S_p^2 = [(n-1)S_1^2 + (n-1)S_2^2]/(2n-2)$ where S_i^2 are unbiased estimators of their sample variances. It follows that S_p^2 is unbiased, i.e. $\operatorname{E} S_p^2 = \sigma^2$. Furthermore from E8.113, $\operatorname{var} S_p^2 = \sigma^2/(n-1) \to 0$ as $n \to \infty$, hence S_p^2 is consistent.

E9.20 Consistent estimator for binomial probability. $\hat p = Y/n$ has $\operatorname{E}\hat p = p$ and $\operatorname{var}\hat p = p(1-p)/n$. It is unbiased and $\lim_{n\to\infty} \operatorname{var}\hat p = 0$. By Theorem 9.1, $\hat p$ is consistent.

E9.23 A variance estimator.

(a) Using linearity of the expectation and independence of Y_i,

$$\operatorname{E}\hat\sigma^2 = \frac{1}{2k}\sum_{i=1}^{k} \operatorname{E}(Y_{2i}^2 + Y_{2i-1}^2 - 2Y_{2i}Y_{2i-1}) = E(Y_{2i}^2) - (\operatorname{E} Y_{2i})^2 = \operatorname{var} Y = \sigma^2.$$

(b) The quantities $U_i = (Y_{2i} - Y_{2i-1})^2$ can be viewed as independent draws from some distribution U with mean $\operatorname{E} U = 2\sigma^2$ [from (a)] and finite variance $\operatorname{var} U$. It follows that their mean $\overline{U}_k = (1/k)\sum_{i=1}^{k} U_i$ converges in probability to $\operatorname{E} U = 2\sigma^2$ as $k \to \infty$ (e.g., Example 9.2). Since $\hat\sigma^2 = \overline{U}_k/2$, $\hat\sigma^2$ converges in probability to σ^2 and thus is consistent.

(c) The fourth moment must be finite so that $\operatorname{var} U$ is finite.

E9.24 Sum of independent standard normal r.v. $Z = \sum_{i=1}^{n} Y_i^2 \sim \chi^2(n)$. W_n is an average of Y_i^2 r.v., hence converges in probability to $\operatorname{E} Y_i^2 = 1$.

***E9.26 Establishing consistency from $\boldsymbol{\lim_{n\to\infty} P(|\hat\theta_n - \theta| \le \epsilon) = 1}$.**
$\mathcal{P}(\epsilon) = P(|Y_{(n)} - \theta| \le \epsilon) = P(\theta - \epsilon \le Y_{(n)} \le \theta + \epsilon) = F_{(n)}(\theta+\epsilon) - F_{(n)}(\theta-\epsilon)$. For $\epsilon > \theta$, $\mathcal{P}(\epsilon) = 1 - 0 = 1$. For $0 < \epsilon < \theta$, $\mathcal{P}(\epsilon) = 1 - [(\theta-\epsilon)/\theta]^n \to 1$ as $n \to \infty$. Therefore $\lim_{n\to\infty}\mathcal{P}(\epsilon) = 1$ for all $\epsilon > 0$ and we conclude $Y_{(n)}$ is a consistent estimator of θ.

Comment. Direct integration of the density function for $Y_{(n)}$ yields $\operatorname{E} Y_{(n)} = [n/(n+1)]\theta$ and $\operatorname{var} Y_{(n)} = [n/(n^3 + 4n^2 + 5n + 2)]\theta^2$. Indeed $\operatorname{E} Y_{(n)} \to \theta$ and $\operatorname{var} Y_{(n)} \to 0$ as $n \to \infty$.

***E9.27 Establishing consistency from $\boldsymbol{\lim_{n\to\infty} P(|\hat\theta_n - \theta| \le \epsilon) = 1}$.**
Like E9.26, $\mathcal{P}(\epsilon) = P(|Y_{(1)} - \theta| \le \epsilon) = F_{(1)}(\theta+\epsilon) - F_{(1)}(\theta-\epsilon)$. For $\epsilon > \theta$, $\mathcal{P}(\epsilon) = 1 - 0 = 1$. For $0 < \epsilon < \theta$, $\mathcal{P}(\epsilon) = 1 - \{1 - [1 - (\theta-\epsilon)/\theta]^n\} = (\epsilon/\theta)^n \to 0$ as $n \to \infty$. Therefore $Y_{(1)}$ is not a consistent estimator of θ.

Comment. Direct integration of the density function for $Y_{(1)}$ yields $\operatorname{E} Y_{(1)} = [1/(n+1)]\theta$ and $\operatorname{var} Y_{(1)} = \operatorname{var} Y_{(n)}$ (computed in E9.26). $Y_{(1)}$ is not a consistent estimator because it is biased; it tends to zero as $n \to \infty$.

***E9.28 Establishing consistency from $\lim_{n\to\infty} P(|\hat{\theta}_n - \theta| \le \epsilon) = 1$.**
Like E9.26, $\mathcal{P}(\epsilon) = P(|Y_{(1)}-\beta| \le \epsilon) = F_{(1)}(\beta+\epsilon)-F_{(1)}(\beta-\epsilon)$. For $\epsilon > 0$, $\mathcal{P}(\epsilon) = 1-[\beta/(\beta+\epsilon)]^{\alpha n} \to 1$ as $n \to \infty$. Therefore $Y_{(1)}$ is a consistent estimator of β.

E9.31 Gamma sample mean convergence. By LLN, $\overline{Y}$ converges to $\mathrm{E}\, Y = \alpha\beta$.

E9.32 Sample mean convergence of a inverse power law density. LLN does not apply because $\mathrm{E}\, Y^2$, and therefore $\operatorname{var} Y$, is not finite.

E9.33 Proportion of bacteria estimator. By LLN, $\hat{\lambda}_1 = \overline{X}$ and $\hat{\lambda}_2 = \overline{Y}$ are consistent estimators of λ_i. By Theorem 9.2, $(\overline{X}+\overline{Y})$ is a consistent estimator of $(\lambda_1+\lambda_2)$ and furthermore, $\hat{\theta} = \overline{X}/(\overline{X}+\overline{Y})$ is a consistent estimator of $\lambda_1/(\lambda_1+\lambda_2)$.

E9.34 Rayleigh density estimator. By LLN, W_n converges in probability to $\mathrm{E}\, Y^2 = \theta$. □

E9.35 Sequence of r.v. with equal mean but unequal variance. **(a)** $\mathrm{E}\,\overline{Y}_i = \mu$. **(b)** Assuming independence, $\operatorname{var} \overline{Y}_i = (1/n^2)\sum_{i=1}^n \sigma_i^2$. **(c)** $\sum_{i=1}^n \sigma_i^2$ must be finite.

E9.36 Justifying $\sigma_{\hat{p}} = \sqrt{p(1-p)/n} \approx \sqrt{\hat{p}(1-\hat{p})/n}$ in large-sample confidence intervals.
Recall Y can be expressed as a sum of Bernoulli trials $Y_n = \sum_{i=1}^n X_i$ with mean $\mathrm{E}\, X_i = p$ and $\operatorname{var} X_i = p(1-p)$. In the large n limit, by the CLT, $\overline{Y}_n = (1/n)\sum_{i=1}^n X_i \approx \mathcal{N}[p, p(1-p)/n]$. Let $\sigma_{\hat{p}} = \sqrt{p(1-p)/n}$; then the distribution of $Z_n = (\overline{Y}_n - p)/\sigma_{\hat{p}} \approx \mathcal{N}(0,1)$ with exact convergence in the limit $n \to \infty$.

Now consider $W_n = \sqrt{\hat{p}(1-\hat{p})/n}/\sigma_{\hat{p}}$. By LLN, $\hat{p} = \overline{Y}_n$ is a consistent estimator of p. By Theorem 9.2, $\hat{p}(1-\hat{p})$ converges to $p(1-p)$ and thus W_n converges to 1. It follows from Theorem 9.3 that $Z_n/W_n = (\overline{Y}_n - p)/\sqrt{\hat{p}(1-\hat{p})/n}$ converges to a standard normal distribution. □

9.4 Sufficiency

Example of an insufficient statistic. Consider a random sample of Bernoulli r.v. $X_1, X_2, \ldots, X_n$. We saw that $Y = \sum_{i=1}^n X_i$ is a sufficient statistic via $P(\mathbf{X} = \mathbf{x}|Y = y) = 1/\binom{n}{y}$ (and factorization in E9.37). Consider the statistic $T(\mathbf{X}) = X_1$. Intuitively, as it is a function of only one sample point, we would not expect it to contain all the information in the sample about p. Indeed

$$P(\mathbf{X} = \mathbf{x} \mid X_1 = x_1) = \frac{p^{\sum_{i=1}^n x_i}(1-p)^{n-\sum_{i=1}^n x_i}}{p^{x_1}(1-p)^{1-x_1}} = p^{\sum_{i=2}^n x_i}(1-p)^{n-\sum_{2=1}^n x_i}$$

is still a function of p, therefore X_1 is not sufficient.

Can a sufficient statistic be inconsistent? Yes. Continuing with the Bernoulli random sample, $\mathrm{E}\, Y = np$ and $\operatorname{var} Y = np(1-p)$ clearly proves Y is not consistent.

Can a consistent statistic be insufficient? Also yes. Consider the Bernoulli random sample with statistic $T(\mathbf{X}) = \overline{Y} - X_1/n = (1/n)\sum_{i=2}^n X_i$. It is straightforward to show $\mathrm{E}\, T = [(n-1)/n]p$ and $\operatorname{var} T = [(n-1)/n^2]p(1-p)$. Furthermore $\lim_{n\to\infty} \mathrm{E}\, T = p$ and $\lim_{n\to\infty} \operatorname{var} T = 0$, therefore T is consistent for p. Intuitively, T makes no use of the first observation X_1 so we expect it to be insufficient. Since any one-to-one function of a sufficient statistic is sufficient, sufficiency of T can be shown by checking the simpler r.v. $U = nT$. Since $U = \sum_{i=2}^n X_i \sim \text{binomial}(n-1, p)$,

$$P(\mathbf{X} = \mathbf{x} \mid U = u) = \frac{p^{\sum_{i=1}^n x_i}(1-p)^{n-\sum_{i=1}^n x_i}}{\binom{n-1}{u}p^u(1-p)^{n-1-u}} = \frac{p^{x_1}(1-p)^{1-x_1}}{\binom{n-1}{\sum_{i=2}^n x_i}}$$

is still a function of p, hence U and thus T is not sufficient even though T is consistent.

E9.37 Sufficiency of a Bernoulli sum. Let $Y = \sum_{i=1}^n X_i$. By independence of $\mathbf{X} \equiv X_1, X_2, \ldots, X_n$, the likelihood $L(\mathbf{x} \mid p) = p^y(1-p)^{n-y}$. Defining nonnegative functions $g(y,p) = p^y(1-p)^{n-y}$ and $h(\mathbf{x}) = 1$, $L(\mathbf{x} \mid p) = g(y,p)h(\mathbf{x})$ and by the factorization theorem, Y is sufficient.

E9.38 Sufficient statistics for a normal random sample.
$Y \sim \mathcal{N}(\mu, \sigma^2)$ has pdf $f(y) = e^{-(y-\mu)^2/(2\sigma^2)}/\sqrt{2\pi\sigma^2}$, hence the likelihood

$$L(\mathbf{y} \mid \mu, \sigma) = \frac{1}{(2\pi\sigma^2)^{n/2}} \prod_{i=1}^{n} e^{-(y_i-\mu)^2/(2\sigma^2)} = \frac{1}{(2\pi\sigma^2)^{n/2}} \exp\left[-\frac{1}{2\sigma^2}\sum_{i=1}^{n}(y_i-\mu)^2\right].$$

(a) It is straightforward to show

$$L(\mathbf{y} \mid \mu, \sigma) = \frac{1}{(2\pi\sigma^2)^{n/2}} \exp\left[-\frac{1}{2\sigma^2}\left(\sum_{i=1}^{n} y_i^2 + n\mu^2 - 2n\mu\overline{y}\right)\right].$$

If σ^2 is known, we can treat it as a constant. The likelihood factorizes as a product of $g(\overline{y}, \mu) = \exp\left[-\frac{1}{2\sigma^2}\left(n\mu^2 - 2n\mu\overline{y}\right)\right]$ and $h(\mathbf{y}) = \exp\left[-\frac{1}{2\sigma^2}\sum_{i=1}^{n} y_i^2\right]/(2\pi\sigma^2)^{n/2}$. Therefore by the factorization theorem, $\overline{Y}$ is sufficient for μ.

(b) Let $U = \sum_{i=1}^{n}(Y_i-\mu)^2$. Then we can express $L(\mathbf{y} \mid \mu, \sigma) = e^{-u/(2\sigma^2)}/(2\pi\sigma^2)^{n/2}$, which factorizes as a product of $g(u, \sigma) = L(\mathbf{y} \mid \mu, \sigma)$ and $h(\mathbf{y}) = 1$. Therefore U is sufficient for σ^2.

(c) Let $V = \sum_{i=1}^{n} Y_i^2$. From (a),

$$L(\mathbf{y} \mid \mu, \sigma) = \frac{1}{(2\pi\sigma^2)^{n/2}} \exp\left[-\frac{1}{2\sigma^2}\left(v + n\mu^2 - 2\mu n\overline{y}\right)\right],$$

which factorizes into $g(n\overline{y}, v, \mu, \sigma) = L(\mathbf{y} \mid \mu, \sigma)$ and $h(\mathbf{y}) = 1$. Thus $n\overline{y}$ and v are jointly sufficient for μ and σ^2. Realizing substitutions $\overline{Y} = n\overline{Y}/n$ and $S^2 = (V - n\overline{Y}^2)/(n-1)$ only introduce constants, it follows that $\overline{Y}$ and S^2 are jointly sufficient for μ and σ^2.

E9.39 Sufficiency of a Poisson sum. Let $Z = \sum_{i=1}^{n} Y_i$. By E6.54, $Z \sim \text{Poisson}(n\lambda)$. Therefore

$$P(\mathbf{Y} = \mathbf{y} \mid Z = z) = \frac{\prod_{i=1}^{n} \lambda^{y_i} e^{-\lambda}/y_i!}{(n\lambda)^z e^{-n\lambda}/z!} = \frac{\prod_{i=1}^{n} /y_i!}{n^z/z!}$$

which has no dependence on λ, hence Z is sufficient for λ.

Alternate solution. Using the factorization theorem, the likelihood factorizes

$$L(\mathbf{y} \mid \lambda) = \prod_{i=1}^{n} \lambda^{y_i} e^{-\lambda}/y_i! = \underbrace{\lambda^z e^{-n\lambda}}_{g(z,\lambda)} \times \underbrace{\left(1\Big/\prod_{i=1}^{n} y_i!\right)}_{h(\mathbf{y})}.$$

E9.42 Sufficiency of a geometric sample mean. Since one-to-one functions of sufficient statistics are sufficient, it is enough to prove $Z = n\overline{Y} = \sum_{i=1}^{n} Y_i$ is sufficient. Note Z is negative binomial distributed with parameters (n, p). Therefore

$$P(\mathbf{Y} = \mathbf{y} \mid Z = z) = \frac{\prod_{i=1}^{n} p(1-p)^{y_i-1}}{\binom{z-1}{n-1} p^n (1-p)^{z-n}} = \frac{1}{\binom{z-1}{n-1}}$$

which has no dependence on p, hence Z and thus $\overline{Y}$ is sufficient for p.

Alternate solution. Using the factorization theorem, the likelihood factorizes

$$L(\mathbf{y} \mid p) = \prod_{i=1}^{n} p(1-p)^{y_i-1} = \underbrace{p^n(1-p)^{z-n}}_{g(z,p)} \times \underbrace{1}_{h(\mathbf{y})}.$$

***E9.49 Sufficiency of the maximum of a uniform sample.** $Y \sim \text{uniform}(0, \theta)$ has pdf $f_Y(y) = 1/\theta$ for $0 \le y \le \theta$. Note that the support of y itself depends on θ. Therefore the support itself must be taken into consideration when establishing a quantity is independent of θ. To do this, define indicator variables

$$I_A(x) \equiv \begin{cases} 1 & \text{if } x \in A \\ 0 & \text{if } x \notin A \end{cases}.$$

Then we can express $f_Y(y) = (1/\theta) \times I_{[0,\theta]}(y)$. From E6.76, $f_{(n)}(y) = (ny^{n-1}/\theta^n) I_{[0,\theta]}(y)$. Thus

$$f(\mathbf{Y} = \mathbf{y} \mid Y_{(n)} = y_{(n)}) = \frac{\prod_{i=1}^n I_{[0,\theta]}(y_i)/\theta}{(ny_{(n)}^{n-1}/\theta^n) I_{[0,\theta]}(y_{(n)})} = \frac{\prod_{i=1}^n I_{[0,\theta]}(y_i)}{(ny_{(n)}^{n-1}) I_{[0,\theta]}(y_{(n)})}.$$

In this form, it appears (through indicator variables) there is dependence on θ. However, note that $\prod_{i=1}^n I_{[0,\theta]}(y_i) = 1$ only if $y_i \in [0, \theta]$ for all $i = 1, 2, \ldots, n$. This occurs if and only if $y_{(1)} \ge 0$ and $y_{(n)} \le \theta$. Therefore $\prod_{i=1}^n I_{[0,\theta]}(y_i) = I_{[0,\infty)}(y_{(1)}) I_{(-\infty,\theta]}(y_{(n)})$. Finally noting that $y_{(n)} \ge 0$, $I_{(-\infty,\theta]}(y_{(n)}) = I_{[0,\theta]}(y_{(n)})$. Substituting into the conditional density yields

$$f(\mathbf{Y} = \mathbf{y} \mid Y_{(n)} = y_{(n)}) = \frac{I_{[0,\infty)}(y_{(1)})}{ny_{(n)}^{n-1}},$$

which no longer has θ dependence. Thus, $Y_{(n)}$ is sufficient for θ.

Alternate solution. Using the factorization theorem,

$$L(\mathbf{y} \mid \theta) = \prod_{i=1}^n I_{[0,\theta]}(y_i)/\theta = \frac{1}{\theta^n} \prod_{i=1}^n I_{[0,\theta]}(y_i) = \underbrace{\frac{1}{\theta^n} I_{(-\infty,\theta]}(y_{(n)})}_{g(y_{(n)},\theta)} \times \underbrace{I_{[0,\infty)}(y_{(1)})}_{h(\mathbf{y})}.$$

***E9.50 Joint sufficiency of a uniform sample.** Using indicator variables as in E9.49, $Y \sim \text{uniform}(\theta_1, \theta_2)$ has pdf $f_Y(y) = I_{[\theta_1,\theta_2]}(y)/(\theta_2 - \theta_1)$. Using the factorization theorem,

$$L(\mathbf{y} \mid \theta_1, \theta_2) = \prod_{i=1}^n I_{[\theta_1,\theta_2]}(y_i)/(\theta_2 - \theta_1) = \frac{1}{(\theta_2 - \theta_1)^n} \prod_{i=1}^n I_{[\theta_1,\theta_2]}(y_i).$$

Note that $y_i \in [\theta_1, \theta_2]$ for all $i = 1, 2, \ldots, n$ if and only if $y_{(1)} \ge \theta_1$ and $y_{(n)} \le \theta_2$. Thus $\prod_{i=1}^n I_{[\theta_1,\theta_2]}(y_i) = I_{[\theta_1,\infty)}(y_{(1)}) I_{(-\infty,\theta_2]}(y_{(n)})$. Substitution into the likelihood yields a function $g = g(y_{(1)}, y_{(n)}, \theta)$, hence factorizes with $h(\mathbf{y}) = 1$. Thus $Y_{(1)}$ and $Y_{(n)}$ are jointly sufficient for θ_1 and θ_2.

9.5 The Rao–Blackwell Theorem and Minimum-Variance Unbiased Estimation

More on Rao–Blackwell.

- From Theorem 9.5, it seems natural to find an unbiased estimator $\hat{\theta}$ and a sufficient statistic for θ, then compute the improved estimator $\hat{\theta}^* = \mathrm{E}(\hat{\theta} \mid U)$. However, as the text states, we can generally restrict our search to estimators that are functions of sufficient statistics, which by the factorization theorem is often easier to do. This is because the improved estimator $\hat{\theta}^* = \mathrm{E}(\hat{\theta} \mid U)$ is a function of U, say $h(U)$, and by Theorem 9.5 cannot be improved upon.

- Rao–Blackwell applies to finding estimators for functions of parameters, say $f(\theta)$. Suppose $\hat{f}$ is an unbiased estimator for $f(\theta)$, then consider $\hat{f}^* = \mathrm{E}(\hat{f} \mid U)$, now a function of U, say $g(U)$. As U is sufficient for θ, $g(U)$ does not depend on θ and the proof of Rao–Blackwell proceeds similarly; we conclude that $g(U)$ is an improved estimator. Again, we can restrict our search to estimators that are functions of U (e.g., see E9.65).

- If U is sufficient and **complete** and $\hat{\theta}$ is unbiased, then by the Lehmann–Scheffé theorem, the Rao–Blackwell estimator is the unique minimum variance unbiased estimator (MVUE).
- The Rao–Blackwell Theorem holds for biased estimators $\hat{\phi}$ in the sense that $\hat{\phi}^* = \mathrm{E}(\hat{\phi} \mid U)$ has $\mathrm{E}(\hat{\phi}^*) = \mathrm{E}(\hat{\phi})$ and $\mathrm{var}(\hat{\phi}^*) \le \mathrm{var}(\hat{\phi})$.

E9.56 MVUE for normal random sample with known $\boldsymbol{\mu}$. $U = \sum_{i=1}^n (Y_i - \mu)^2$ is sufficient for σ^2. We need a function $h(U)$ such that $\mathrm{E}[h(U)] = \sigma^2$. Noting $\mathrm{E}\,U = n\,\mathrm{E}[(Y-\mu)^2] = n\sigma^2$, it follows that $\hat{\sigma}^2 = U/n$ is a MVUE.

E9.59 MVUE for squared Poisson r.v. We want to estimate $\mathrm{E}\,C = 3[\mathrm{var}\,Y + (\mathrm{E}\,Y)^2] = 3(\lambda + \lambda^2)$. We saw in E9.39 that $U = \sum_{i=1}^n Y_i$ is a sufficient statistic for λ. Since $\mathrm{E}\,U = n\lambda$, a MVUE estimator for λ is $\hat{\lambda} = U/n = \overline{Y}$. Since we also need an estimator for λ^2, we can try $\overline{Y}^2$. However, $\mathrm{E}\,\overline{Y}^2 = \mathrm{var}\,\overline{Y} + (\mathrm{E}\,\overline{Y})^2 = \lambda/n + \lambda^2$ is biased. Evidently $\widehat{\lambda^2} = \overline{Y}^2 - \overline{Y}/n$ is unbiased (and a function of sufficient statistic $\overline{Y}$). Therefore $\widehat{C} = 3(\hat{\lambda} + \widehat{\lambda^2}) = 3(\overline{Y} + \overline{Y}^2 - \overline{Y}/n)$ is an MVUE for C.

E9.60 Exponential family density.

(a)
$$L(\mathbf{y} \mid \theta) = \prod_{i=1}^n \theta y_i^{\theta-1} = \theta^n \exp\left[\log\left(\prod_{i=1}^n y_i^{\theta-1}\right)\right] = \theta^n \exp\left[(\theta-1)\sum_{i=1}^n \log y_i\right].$$

Thus the likelihood factorizes with $h(\mathbf{y}) = 1$ and $U = \sum_{i=1}^n (-\log y_i)$ is sufficient for θ.

(b) $f_W(w) = f_Y(e^{-w})|de^{-w}/dw| = \theta e^{-w\theta}$, $0 < e^{-w} < 1$ or $w > 0$. Indeed $W \sim$ exponential$(1/\theta)$.

(c) $m_W(t) = (1-t/\theta)^{-1}$. By independence, $U = \sum_{i=1}^n (W_i)$ has $m_U(t) = (1-t/\theta)^{-n}$. Furthermore, $V = 2\theta U$ has $m_V(t) = m_U(2\theta t) = (1-2t)^{-n}$, which is the mgf of a $\chi^2(\nu = 2n)$ r.v.

(d) By direct integration, $\mathrm{E}(1/V) = \int_0^\infty dv\,(1/v) f_V(v) = 1/[2(n-1)]$.

(e) U is a sufficient statistic for θ. From (d), $\mathrm{E}(1/U) = \mathrm{E}(2\theta/V) = 2\theta\,\mathrm{E}(1/V) = \theta/(n-1)$. Evidently $\hat{\theta} = (n-1)/U$ is a MVUE for θ.

E9.61 MVUE for uniform random sample. $\mathrm{E}\,Y_{(n)} = [n/(n+1)]\theta$, thus $\hat{\theta} = [(n+1)/n]Y_{(n)}$.

E9.64 MVUE for normal random sample with known $\boldsymbol{\sigma^2}$.

(a) Refering to E9.38(a), $\overline{Y}$ is unbiased and sufficient for μ. $\mathrm{E}\,\overline{Y}^2 = \mathrm{var}\,\overline{Y} + (\mathrm{E}\,\overline{Y})^2 = \sigma^2/n + \mu^2$ is biased for μ^2, but with σ^2 known it is easy to see $\widehat{\mu^2} = \overline{Y}^2 - \sigma^2/n$ is unbiased and a function of sufficient statistics, hence is the MVUE of μ^2.

(b) $\mathrm{var}\,\widehat{\mu^2} = \mathrm{var}\,\overline{Y}^2 = \mathrm{E}\,\overline{Y}^4 - (\mathrm{E}\,\overline{Y}^2)^2$ where $\overline{Y} \sim \mathcal{N}(\mu, \sigma^2)$.

***E9.65 Bernoulli variance MVUE.**

(a) $\mathrm{E}\,T = P(Y_1 = 1 \cap Y_2 = 0) = p(1-p)$ by independence of Y_i. Evidently T is an unbiased estimator of $p(1-p)$. However, it is a poor estimator since it is inconsistent (constant variance).

(b)
$$P(T = 1 \mid W = w) = \frac{P(T = 1 \cap W = w)}{P(W = w)} = \frac{p(1-p)\binom{n-2}{w-1}p^{w-1}(1-p)^{(n-2)-(w-1)}}{\binom{n}{w}p^w(1-p)^{n-w}} = \frac{\binom{n-2}{w-1}}{\binom{n}{w}}.$$

The second equality follows by requiring $Y_1 = 1$, $Y_2 = 0$, and the remaining $(n-2)$ Y_i such that they sum to $(w-1)$. Substitution of factorials yields $P(T = 1 \mid W = w) = [w(n-w)]/[n(n-1)]$.

(c) Since W is sufficient for p (and hence functions of p), by Rao–Blackwell, an improved estimator of $p(1-p)$ is $\mathrm{E}(T \mid W) = P(T = 1 \mid W) = W(n-W)/[n(n-1)] = [n/(n-1)]\overline{Y}(1-\overline{Y})$.

Alternate solution. $p(1-p) = p - p^2$, thus we must find estimators $\hat{p}$ and $\widehat{p^2}$. $W = \sum_{i=1}^n Y_i$ is sufficient for p with $\mathrm{E}\,W = np$, therefore $\hat{p} = W/n = \overline{Y}$. We may be tempted to estimate p^2 by $(\hat{p})^2$, but $\mathrm{E}(\hat{p})^2 = \mathrm{E}\,\overline{Y}^2 = \mathrm{var}\,\overline{Y} + (\mathrm{E}\,\overline{Y})^2 = p(1-p)/n + p^2 = p/n + p^2(1-1/n)$ is biased for p^2. However, from $\mathrm{E}\,\overline{Y}^2$ we can construct an unbiased function of sufficient statistic $\overline{Y}$ as $\widehat{p^2} = [\overline{Y}^2 - \overline{Y}/n]/(1-1/n)$. Simplification yields MVUE $\hat{p} - \widehat{p^2} = [n/(n-1)]\overline{Y}(1-\overline{Y})$.

9.6 The Method of Moments

Sample moment definition. For a discrete r.v., the k-th moment $\mu'_k = \mathrm{E}(Y^k) = \sum_y y^k P(Y = y)$. In a random sample $Y_1, Y_2, \ldots, Y_n$, we assume each point is equally probable, hence $P(Y_i = y_i) = 1/n$ and $m'_k = \sum_{i=1}^n y_i^k P(Y_i = y_i) = (1/n)\sum_{i=1}^n y_i^k$.

E9.69 Method of moments estimate. The first population moment is simply the mean; through direct integration, $\mu'_1 = \mathrm{E}\,Y = \int_0^1 dy\, y f(y) = (\theta+1)/(\theta+2) = (\theta+1)/(\theta+2)$. Equating to the first sample moment $m'_1 = \mu'_1$, noting $m'_1 = \overline{Y}$ by definition, implies $\overline{Y} = (\theta+1)/(\theta+2)$. Solving yields method of moments estimate $\hat{\theta} = (1-2\overline{Y})/(\overline{Y}-1)$. $\overline{Y}$ converges to μ as $n \to \infty$ so, by Theorem 9.2, $\hat{\theta}$ converges to $(1-2\mu)/(\mu-1)$ which simplifies to θ; therefore $\hat{\theta}$ is consistent. Since it is not a function of the sufficient statistic obtained by the factorization theorem (for details of factorization, see E9.60), it is not a MVUE.

E9.72 Method of moments estimation of normal mean and variance. $\mu'_1 = \mathrm{E}\,Y = \mu$ and $\mu'_2 = \mathrm{E}\,Y^2 = \operatorname{var} Y + (\mathrm{E}\,Y)^2 = \sigma^2 + \mu^2$. Equating to sample moments yields $\overline{Y} = \mu$ and

$$\frac{1}{n}\sum_{i=1}^n Y_i^2 = \sigma^2 + \mu^2.$$

Evidently $\hat{\mu} = \overline{Y}$ and

$$\widehat{\sigma^2} = \frac{1}{n}\sum_{i=1}^n Y_i^2 - (\hat{\mu})^2 = \frac{1}{n}\left(\sum_{i=1}^n Y_i^2 - n\overline{Y}^2\right) = \frac{1}{n}\left(\sum_{i=1}^n (Y_i - \overline{Y})^2\right),$$

the usual sample variance without Bessel's correction.

E9.73 Method of moments estimate of random draws without replacement. $Y \sim$ hypergeometric(N, n, θ). Therefore $\mu'_1 = \theta(n/N)$. Equating to $\overline{Y}$ yields $\hat{\theta} = \overline{Y}(N/n)$. Note that with a single sample $\overline{Y} = Y$.

E9.76 Method of moments estimate of a single binomial observation. $Y \sim$ geometric(p). Therefore $\mu'_1 = 1/p$. Equating to $\overline{Y}$ yields $\hat{p} = 1/\overline{Y}$. Note that with a single sample $\overline{Y} = Y$.

9.7 The Method of Maximum Likelihood

E9.80 Poisson sample MLE.

(a) $L(\lambda) = \prod_{i=1}^n \lambda^{y_i} e^{-\lambda}/y_i! \propto \lambda^{\sum_{i=1}^n y_i} e^{-n\lambda}$ where the proportional terms are independent of λ. It is sufficient to maximize $\log L(\lambda) = \sum_{i=1}^n y_i \log\lambda - n\lambda + f(\mathbf{y})$. Setting $d \log L(\lambda)/d\lambda = 0$ yields MLE $\hat{\lambda} = (1/n)\sum_{i=1}^n Y_i = \overline{Y}$. (One can verify $d^2 \log L(\hat{\lambda})/d\lambda^2 = -n/\overline{Y} < 0$ since $y_i \geq 0$. In the case all $y_i = 0$, $L(\lambda) = e^{-\lambda}$ is still maximized at $\hat{\lambda} = \overline{Y} = 0$.)

(b) $\mathrm{E}\,\hat{\lambda} = \mathrm{E}\,Y = \lambda$ and $\operatorname{var}\hat{\lambda} = (\operatorname{var} Y)/n = \lambda/n$.

(c) From (b) and Theorem 9.1, it follows $\hat{\lambda}$ is consistent.

(d) By the invariance property, $\widehat{e^{-\lambda}} = e^{-\hat{\lambda}} = e^{-\overline{Y}}$.

E9.81 Exponential sample MLE. $L(\theta) = \prod_{i=1}^n e^{-y_i/\theta}/\theta = e^{-n\overline{y}/\theta}/\theta^n$. Setting $d \log L(\theta)/d\theta = 0$ yields MLE $\hat{\theta} = \overline{Y}$. (One can verify $d^2 \log L(\hat{\theta})/d\theta^2 = -nY^{-(n+2)}e^{-n} < 0$ since $y > 0$.) By the invariance property, $\widehat{\theta^2} = (\hat{\theta})^2 = \overline{Y}^2$.

E9.82 Exponential family density estimator.

(a) $L(\theta) = \prod_{i=1}^n (r/\theta) y_i^{r-1} e^{-y_i^r/\theta} = (r/\theta)^n e^{-\sum_{i=1}^n y_i^r/\theta} h(\mathbf{y})$ where $h(\mathbf{y}) = \prod_{i=1}^n y_i^{r-1}$. The remaining factor is a function $g(z = \sum_{i=1}^n y_i^r, \theta)$, hence by the factorization theorem Z is sufficient for θ.

(b) It is sufficient to maximize $g(z,\theta)$ or $\log g(z,\theta) = n\log(r/\theta) - z/\theta$. Setting $d \log g(z,\theta)/d\theta = 0$ yields MLE $\hat{\theta} = Z/n$. (One can verify $d^2 \log g(z, \hat{\theta})/d\theta^2 = -n^3/z^2 < 0$.)

(c) $\hat{\theta}$ is a MVUE for θ as it is a function of sufficient statistic Z.

E9.83 Uniform sample MLE. I assume $\theta > -\frac{1}{2}$.

(a) $L(\theta) = \prod_{i=1}^n I_{[0,2\theta+1]}(y_i)/(2\theta+1) = [1/(2\theta+1)^n]\prod_{i=1}^n I_{[0,2\theta+1]}(y_i)$. $1/(2\theta+1)^n$ is a positive decreasing function of θ for $\theta > -\frac{1}{2}$ and the indicators simplify in terms of order statistics: $\prod_{i=1}^n I_{[0,2\theta+1]}(y_i) = I_{[0,\infty)}(y_{(1)})I_{(-\infty,2\theta+1]}(y_{(n)})$. Evidently $L(\theta) = 0$ when $y_{(n)} \geq 2\theta+1$ or $\theta \leq (y_{(n)}-1)/2$. Thus the smallest value of θ such that $L(\theta) > 0$ is the MLE $\hat{\theta} = (Y_{(n)}-1)/2$.

(b) $\operatorname{var} Y = \frac{1}{12}(2\theta+1)^2$. By the invariance property, $\widehat{\operatorname{var} Y} = \frac{1}{12}(2\hat{\theta}+1)^2 = \frac{1}{12}Y_{(n)}^2$.

E9.84 Gamma sample MLE, $\boldsymbol{\alpha}$ known.

(a) $L(\theta) = \prod_{i=1}^n y_i^{\alpha-1}e^{-y_i/\theta}/[\Gamma(\alpha)\theta^\alpha] \propto e^{-z/\theta}/\theta^{n\alpha}$ where $Z = \sum_{i=1}^n Y_i$ and the proportional terms are independent of θ. Setting $d\log L(\theta)/d\theta = 0$ yields MLE $\hat{\theta} = Z/(n\alpha) = \overline{Y}/\alpha$.

(b) $\operatorname{E}\hat{\theta} = \theta$ and $\operatorname{var}\hat{\theta} = \theta^2/(n\alpha)$.

(c) Because $\hat{\theta}$ is unbiased, we can apply Tchebysheff's inequality $P(|\hat{\theta}-\theta| \leq k\sigma_\theta) \leq 1-1/k^2$ for any $k > 0$. Estimating $\sigma_\theta \approx \sqrt{\operatorname{var}\hat{\theta}} \approx \hat{\theta}\sqrt{n\alpha}$ and setting $k = 2$, we have approximately $P(|\hat{\theta}-\theta| \leq 2\hat{\theta}/\sqrt{n\alpha}) \leq 0.75$.

(d) $\operatorname{var} Y = \alpha\theta^2$. By the invariance property, $\widehat{\operatorname{var} Y} = \alpha\hat{\theta}^2 = \overline{Y}^2/\alpha$.

E9.85 Gamma sample MLE, $\boldsymbol{\alpha}$ known 2. (a–d) See E9.84. By Theorem 9.1, $\hat{\theta}$ is consistent for θ. Furthermore by the factorization theorem, $Z = \sum_{i=1}^n Y_i$ is a sufficient statistic.

(e) By E6.57, $Z \sim \text{gamma}(n\alpha,\theta)$. From E6.12, $U = (2/\theta)Z \sim \text{gamma}(n\alpha,2) \simeq \chi^2(2n\alpha)$ and is therefore a pivotal quantity. We seek an interval such that $P(\hat{\theta}_L \leq \theta \leq \hat{\theta}_U) = 1-\alpha$. Applying the pivotal method, one such interval is $(2Z/\chi^2_{1-\alpha/2}, 2Z/\chi^2_{\alpha/2})$ where $P(U \leq \chi^2_c) = 1-c$. With the given values, this interval is $(0.0638Z, 0.184Z)$.

E9.86 Normal samples MLE, shared $\boldsymbol{\sigma^2}$.

$$\begin{aligned} L(\sigma^2) &= \prod_{i=1}^m \frac{1}{(2\pi\sigma^2)^{1/2}}e^{-(x_i-\mu_1)^2/(2\sigma^2)} \times \prod_{j=1}^n \frac{1}{(2\pi\sigma^2)^{1/2}}e^{-(y_i-\mu_2)^2/(2\sigma^2)} \\ &= \frac{1}{(2\pi\sigma^2)^{(m+n)/2}}\exp\left[-\frac{1}{2\sigma^2}\left(mS_1^2+nS_2^2\right)\right] \end{aligned}$$

where $S_1^2 = \sum_{i=1}^m(x_i-\mu_1)^2$ and $S_2^2 = \sum_{j=1}^n(y_j-\mu_2)^2$. It follows

$$\log L(\sigma^2) = -\frac{1}{2}(m+n)\log\left(2\pi\sigma^2\right) - \frac{1}{2\sigma^2}\left(mS_1^2+nS_2^2\right).$$

Solving $d\log L(\sigma)/d\sigma = 0$ yields MLE $\widehat{\sigma^2} = (mS_1^2+nS_2^2)/(m+n)$. (Indeed, one can verify that $d^2\log L(\sigma = \sqrt{\widehat{\sigma^2}})/d\sigma^2 = -2(m+n)^2/(mS_1^2+nS_2^2) < 0$.)

E9.87 Multinomial MLE. The sample constitutes a multinomial experiment where y_i votes are observed, $i = 1,2,3$, for each of the candidates. Noting $p_3 = 1-p_1-p_2$,

$$L(p_1,p_2) = \binom{n}{y_1,y_2,y_3}p_1^{y_1}p_2^{y_2}p_3^{y_3} \propto p_1^{y_1}p_2^{y_2}(1-p_1-p_2)^{y_3}.$$

Thus it is sufficient to maximize $f(p_1,p_2) = y_1\log p_1 + y_2\log p_2 + y_3\log(1-p_1-p_2)$. Solving simultaneously $\partial f(p_1,p_2)/\partial p_j$, $j = 1,2$, yields MLE $\hat{p}_j = Y_j/n$ where $n = \sum_{i=1}^3 Y_i$. (One can verify the Hessian determinant evaluated at the critical point $D = n^5/(y_1y_2y_3) > 0$ and $\partial^2 f(p_1,p_2)/\partial p_1^2 = -(y_1+y_3)n^2/(y_1y_3) < 0$.) By the invariance property, $\hat{p}_3 = 1-\hat{p}_1-\hat{p}_2 = Y_3/n$.

Also by the invariance property, $\Delta p = p_1 - p_2$ is estimated by $\widehat{\Delta p} = \hat{p}_1 - \hat{p}_2$ with variance $\operatorname{var}\widehat{\Delta p} = \operatorname{var}(Y_1-Y_2)/n^2 = [\operatorname{var} Y_1 + \operatorname{var} Y_2 - 2\operatorname{cov}(Y_1,Y_2)]/n^2 = [(p_1(1-p_1)+p_2(1-p_2)+2p_1p_2]/n$, hence $\sigma_{\widehat{\Delta p}} = \sqrt{\operatorname{var}\widehat{\Delta p}}$, which can be approximated by replacing p_i with $\hat{p}_i$. Invoking the CLT, $\widehat{\Delta p} \approx \mathcal{N}(\Delta p, \sigma^2_{\widehat{\Delta p}})$ (since a multinomial distribution is still a sum of r.v., see E8.68 for further justification). Therefore an approximate $(1-\alpha)$ confidence interval is $\widehat{\Delta p} \pm z_{\alpha/2}\sigma_{\widehat{\Delta p}}$. Explicitly with $\alpha = 0.05$, the interval is $[-0.241, 0.0809]$.

E9.90 Independent binomial sample MLE with common p.
Let $X \sim \text{binomial}(m,p)$ and $Y \sim \text{binomial}(n,p)$ be the number of men and women in favor, respectively. $L(p) = \binom{m}{x}p^x(1-p)^{m-x}\binom{n}{y}p^y(1-p)^{n-y} \propto p^{x+y}(1-p)^{m+n-x-y}$ is maximized at $\hat{p} = (X+Y)/(m+n)$.

***E9.95 Odds ratio MLE.** Let $Y \sim \text{binomial}(n,p)$ be the number of defective items. We want a MLE for $\theta = p/(1-p)$. By the invariance property, $\hat{\theta} = \hat{p}/(1-\hat{p})$. We have seen seen that $\hat{p} = Y/n$, hence $\hat{\theta} = (Y/n)/(1-Y/n)$.

E9.97 Geometric sample estimators.
(a) $\mu_1' = m_1'$ implies $1/p = \overline{Y}$, hence $\hat{p} = 1/\overline{Y}$.
(b) $L(p) = p^n(1-p)^{n(\overline{Y}-1)}$ is also maximized at $\hat{p} = 1/\overline{Y}$.

9.8 Some Large-Sample Properties of Maximum-Likelihood Estimators

MLE for functions of parameters. The invariance property implies that if $\hat{\theta}$ is a MLE for θ, then $t(\hat{\theta})$ is a MLE for $t(\theta)$. Variances of estimators can be difficult to compute directly. In the large n limit, the asymptotic normality approximation given in the text implies that

$$\operatorname{var} t(\hat{\theta}) = \left[\frac{\partial t(\theta)}{\partial \theta}\right]^2 \bigg/ \left\{ n \operatorname{E}\left[-\frac{\partial^2 \log f(Y \mid \theta)}{\partial \theta^2}\right]\right\}.$$

In particular,

$$\operatorname{var}\hat{\theta} = 1 \bigg/ \left\{ n \operatorname{E}\left[-\frac{\partial^2 \log f(Y \mid \theta)}{\partial \theta^2}\right]\right\}.$$

9.9 Summary

Supplemental Exercises

E9.104 Comparison of MOM and MLE estimators.
(a) $\mu_1' = m_1'$ implies $\theta + 1 = \overline{Y}$, hence the MOM estimator $\hat{\theta}_1 = \overline{Y} - 1$.
(b) $L(\theta) = \prod_{i=1}^n e^{-(y_i-\theta)} I_{[\theta,\infty)}(y_i) = e^{-n(\overline{Y}-\theta)} I_{[\theta,\infty)}(y_{(1)})$. $e^{-n(\overline{Y}-\theta)}$ is an increasing function of θ, hence we want θ to be as small as possible subject to the constraint posed by the indicator $y_{(1)} \geq \theta$ (so that $L(\theta)$ is nonzero). Thus the MLE $\hat{\theta}_2 = Y_{(1)}$.
(c) $\operatorname{E}\hat{\theta}_1 = \theta$ is unbiased with $\operatorname{var}\hat{\theta}_1 = \operatorname{var}\overline{Y} = (\operatorname{var} Y)/n = 1/n$. From Theorem 6.5, $Y_{(1)}$ is distributed as $f_{Y_{(1)}}(y) = ne^{-n(y-\theta)} I_{[\theta,\infty)}(y)$, hence $\operatorname{E}\hat{\theta}_2 = \theta + 1/n$ is biased. It follows $\hat{\theta}_2^* = Y_{(1)} - 1/n$ is unbiased with variance $\operatorname{var}\hat{\theta}_2^* = 1/n^2$. The efficiency of $\hat{\theta}_1$ relative to $\hat{\theta}_2^*$ is $\operatorname{eff}(\hat{\theta}_1, \hat{\theta}_2^*) = \operatorname{var}\hat{\theta}_2^* / \operatorname{var}\hat{\theta}_1 = 1/n$. The MLE estimate is more efficient.

E9.106 Poisson MVUE. $L(\lambda) = \prod_{i=1}^n \lambda^{y_i} e^{-\lambda}/y_i = \lambda^z e^{-\lambda} \times h(\mathbf{y})$ where $U = \sum_{i=1}^n Y_i$. By the factorization theorem, U is a sufficient statistic for λ (and thus functions of λ). We seek an unbiased estimator of $\theta = e^{-\lambda}$, which can then be improved upon by Rao–Blackwell. Consider the simple estimator

$$T = \begin{cases} 1 & \text{if } Y_1 = 0 \\ 0 & \text{otherwise} \end{cases}.$$

$\mathrm{E}\, T = P(Y_1 = 0) = e^{-\lambda}$ is unbiased. Applying Rao–Blackwell, $\hat{\theta} = \mathrm{E}(T \mid U) = P(T = 1 \mid U)$ is a MVUE. Noting $U \sim \text{Poisson}(n\lambda)$,

$$P(T = 1 \mid U = u) = \frac{P(T = 1 \cap U = u)}{P(U = u)} = \frac{P(Y_1 = 0 \cap \sum_{i=2}^{n} Y_i = U)}{(n\lambda)^u e^{-n\lambda}/u!} = \frac{e^{-\lambda}[(n-1)\lambda]^u e^{-(n+1)\lambda}/u!}{(n\lambda)^u e^{-n\lambda}/u!},$$

which simplifies to $P(T = 1 \mid U = u) = [(n-1)/n]^u$. Evidently $\hat{\theta}^* = [(n-1)/n]^U$.

Comment. We can express $\hat{\theta}^* = (1 - 1/n)^{n\overline{Y}}$. As $n \to \infty$, $(1 - 1/n)^n \to e^{-1}$, hence $\hat{\theta}^* \to e^{-\overline{Y}}$, which is the MLE estimate (E9.80).

E9.107 Reliability MLE. We want to estimate $g(\theta) = e^{-t/\theta}$. By the invariance property, $\widehat{g(\theta)} = g(\hat{\theta})$. From E9.81, $\hat{\theta} = \overline{Y}$ hence $\widehat{g(\theta)} = e^{-t/\overline{Y}}$.

***E9.108 Reliability MVUE.**

(a) $\mathrm{E}\, V = P(Y_1 > t) = \int_t^\infty dy_1\, e^{-y_1/\theta}/\theta = e^{-t/\theta}$.

(b) $U = \sum_{i=1}^{n} Y_i \sim \text{gamma}(n, \theta)$ (e.g. E6.13). Therefore

$$\begin{aligned} f_{Y_1|U}(y_1 \mid u) &= \frac{f_{Y_1,U}(y_1, \sum_{i=2}^{n} y_i = u - y_1)}{f_U(u)} \\ &= \frac{\{e^{-y_1/\theta}/\theta\}\{(u-y_1)^{(n-2)} e^{-(u-y_1)/\theta}/[\Gamma(n-2)\theta^{n-2}]\}}{u^{(n-1)} e^{-u/\theta}/[\Gamma(n-1)\theta^{n-1}]} \\ &= (n-1)(u-y_1)^{n-2}/u^{n-1}, \qquad 0 \le y_1 \le u. \end{aligned}$$

(c) $\mathrm{E}(V \mid U) = P(V = 1 \mid U) = P(Y_1 > t \mid U) = \int_t^u dy_1\, f_{Y_1|U}(y_1 \mid u) = (1 - t/u)^{n-1}$.

Comment. We can express $\widehat{g(\theta)} = [1 - (t/\overline{Y})/n]^{n-1} \to e^{-t/\overline{Y}}$ as $n \to \infty$, agreeing with the MLE in E9.107.

***E9.109 Discrete uniform MOM.** Let Y be the random integer drawn, with distribution $P(Y = y) = 1/N$ over support $y = 1, 2, \ldots, N$. It follows $\mu = \mathrm{E}\, Y = \sum_{y=1}^{N} yP(Y = y) = (N+1)/2$. In this problem, we consider a random sample from this distribution $Y_1, Y_2, \ldots, Y_n$.

(a) Equating μ with the sample moment $\overline{Y}$ yields $\hat{N}_1 = 2\overline{Y} - 1$.

(b) $\mathrm{E}\, \hat{N}_1 = 2\mu - 1 = N$ is unbiased with $\operatorname{var} \hat{N}_1 = 4(\operatorname{var} Y)/n = (N^2 - 1)/(3n)$.

***E9.110 Discrete uniform MLE.**

(a) $L(N) = \prod_{i=1}^{n} (1/N) I_{\{1,2,\ldots,N\}}(y_i) = (1/N)^n I_{\{1,2,\ldots\}}(y_{(1)}) I_{\{\ldots,N-1,N\}}(y_{(n)})$. We want to minimize N subject to the indicator constraint integer $y_{(n)} \le N$. Thus $\hat{N}_2 = Y_{(n)}$.

(b) We need the distribution of $Y_{(n)}$. Its cdf $F(k) = P(Y_{(n)} \le k) = \prod_{i=1}^{n} P(Y_i \le k) = (k/N)^n$ for $k = 1, 2, \ldots, N$. Thus $P(Y_{(n)} = k) = F(k) - F(k-1) = (k/N)^n - [(k-1)/N]^n$ and

$$\mathrm{E}\, Y_{(n)} = \sum_{k=1}^{N} k \left[\left(\frac{k}{N}\right)^n - \left(\frac{k-1}{N}\right)^n \right] \equiv \sum_{k=1}^{N} g(k).$$

This sum does not have an analytic solution. However, similar to the normal approximation to the binomial, we can extend $g(k)$ to the continuous interval $0 \le k \le N$ and approximate the sum as an integral:

$$\begin{aligned} \mathrm{E}\, Y_{(n)} &\approx \int_0^N dk\, g(k) = \frac{(n+1)N^2 + (1+1/N)^n[-(n+1)N^2 + nN + 1] - (-1)^n/N^n}{(n+2)(n+1)} \\ &\approx \frac{(n+1)N^2 + (1-n/N)[-(n+1)N^2 + nN + 1]}{(n+2)(n+1)} = \frac{n(n+1)N + nN - n^2 + 1 - n/N}{(n+2)(n+1)} \\ &\approx \frac{n(n+1)N + nN}{(n+2)(n+1)} = \frac{nN}{n+1}. \end{aligned}$$

Further approximations were made in the large N limit (where the integral approximation is better), keeping $\mathcal{O}(N)$ or higher terms. It follows $\hat{N}_3 = [(n+1)/n]Y_{(n)}$ is approximately unbiased, the approximation improving for large N.

(c) A similar approximation yields

$$\mathrm{E}\,Y_{(n)}^2 = \sum_{k=1}^{N} k^2 \left[\left(\frac{k}{N} \right)^n - \left(\frac{k-1}{N} \right)^n \right] \equiv \sum_{k=1}^{N} h(k) \approx \int_0^N dk\, h(k) \approx \frac{n}{n+2} N^2$$

(binomial expanding to second order) in the large N limit, hence

$$\operatorname{var} Y_{(n)} = \mathrm{E}\,Y_{(n)}^2 - (\mathrm{E}\,Y_{(n)})^2 \approx \frac{n}{(n+1)^2(n+2)} N^2.$$

Thus

$$\operatorname{var} \hat{N}_3 = \left(\frac{n+1}{n} \right)^2 \operatorname{var} Y_{(n)} \approx \frac{N^2}{n(n+2)}.$$

(d) In the large N limit,

$$\Delta \equiv \operatorname{var} \hat{N}_3 - \operatorname{var} \hat{N}_1 \approx \frac{N^2}{n(n+2)} - \frac{N^2}{3n} = \frac{N^2(1-n)}{3n(n+2)}.$$

Evidently for $n > 1$, $\Delta < 0$ or $\operatorname{var} \hat{N}_3 < \operatorname{var} \hat{N}_1$.

***E9.111** **Tank estimation problem.** Using $\hat{N}_3$ from E9.110, $\hat{N} \approx [(n+1)/n]Y_{(n)} = (6/5)210 = 252$ with $\operatorname{var} \hat{N} \approx N/[n(n+2)] \approx \hat{N}/[n(n+2)]$, hence $2\sigma_{\hat{N}} \approx 85.2$.

10 Hypothesis Testing

10.1 Introduction

10.2 Elements of a Statistical Test

E10.1 Type I and II errors. $\alpha = P(\text{reject } H_0 \mid H_0 \text{ is true})$. $\beta = P(\text{accept } H_0 \mid H_a \text{ is true})$.

E10.2 Drug effectiveness hypothesis.
(a) Rejecting the drug claims even though the claims are true.
(b) Assuming H_0 to be true, we can model $Y \sim \mathcal{B}(n = 20, p = 0.8)$. We reject H_0 if $y \leq 12$ with probability $\alpha = \sum_{i=0}^{12} \binom{20}{i} (0.8)^i (0.2)^{20-i} \approx 0.0321$.
(c) Accepting the drug claims even though the claims are false.
(d) Assuming H_a to be true, we can model $Y \sim \mathcal{B}(n = 20, p = 0.6)$. We accept (fail to reject) H_0 if $y > 12$ with probability $\beta = \sum_{i=13}^{20} \binom{20}{i} (0.6)^i (0.4)^{20-i} \approx 0.416$.
(e) $\beta = \sum_{i=13}^{20} \binom{20}{i} (0.4)^i (0.6)^{20-i} \approx 0.021$.

E10.3 Drug effectiveness hypothesis 2. **(a)** $\alpha(c)$ is an increasing function of c for $c > 0$ with $\alpha(0) = 0$. Computing $\alpha(c)$ for the largest integer c such that $\alpha < 0.01$ yields $c = 11$. **(b)** $\beta \approx 0.596$. **(c)** $\beta \approx 0.0565$. *Comment.* A decrease in α corresponds to an increase in β.

E10.4 Two-stage testing procedure. Assuming H_0 to be true, let $p = p_0 = 0.05$. Let A denote the event the first two sheets are error free. Then

$$\begin{aligned} \alpha = P(\text{reject } H_0 \mid H_0) &= P(\text{reject } H_0 \mid H_0, A) P(A \mid H_0) + P(\text{reject } H_0 \mid H_0, \overline{A}) P(\overline{A} \mid H_0) \\ &= (1)(1 - p_0)^2 + (1 - p_0)[1 - (1 - p_0)^2] \approx 0.995. \end{aligned}$$

Assuming H_a to be true, let $p = p_a$. Then

$$\begin{aligned} \beta = P(\text{accept } H_0 \mid H_a) &= P(\text{accept } H_0 \mid H_a, A) P(A \mid H_a) + P(\text{accept } H_0 \mid H_0, \overline{A}) P(\overline{A} \mid H_a) \\ &= (0)(1 - p_a)^2 + (p_a)[1 - (1 - p_a)^2] = p_a[1 - (1 - p_a)^2]. \end{aligned}$$

E10.5 Uniform sample hypothesis test. $Y \sim \mathcal{U}(0, 1)$ has pdf $f_Y(y) = 1$, $0 \leq y \leq 1$. By independence, the joint distribution of $f_{Y_1,Y_2}(y_1, y_2) = 1$, $0 \leq y_i \leq 1$. It follows that

$$\begin{aligned} \alpha_1 &= P(Y_1 > 0.95 \mid H_0 : \theta = 0) = \int_{.95}^{1} dy\, f_Y(y) = 0.05, \\ \alpha_2 &= P(Y_1 + Y_2 > c \mid H_0 : \theta = 0) = \begin{cases} \left[\int_0^c dy_1 \int_{c-y_1}^1 dy_2 + \int_c^1 dy_1 \int_0^1 dy_2\right] f_{Y_1,Y_2}(y_1, y_2), & 0 < c < 1 \\ \int_{c-1}^1 dy_1 \int_{c-y_1}^1 dy_2\, f_{Y_1,Y_2}(y_1, y_2), & 1 < c < 2 \end{cases} \\ &= \begin{cases} 1 - c^2/2, & 0 < c < 1 \\ c^2/2 - 2c + 2, & 1 < c < 2 \end{cases}. \end{aligned}$$

Setting $\alpha_1 = \alpha_2$ yields $c \approx 1.68$.

E10.6 Balanced coin hypothesis.
$P(|Y - 18| \geq 4) = P(Y - 18 \geq 4 \cup Y - 18 \leq -4) = P(Y \geq 22 \cup Y \leq 14) = P(Y \geq 22) + P(Y \leq 14)$.
$\alpha = P(|Y - 18| \geq 4 \mid H_0 : p = 0.5) = \sum_{i=0}^{14} \binom{36}{i} (0.5)^i (0.5)^{36-i} + \sum_{i=22}^{36} \binom{36}{i} (0.5)^i (0.5)^{36-i} \approx 0.243$.
$\beta = P(|Y - 18| < 4 \mid H_a : p = 0.7) = \sum_{15}^{21} \binom{36}{i} (0.7)^i (0.3)^{36-i} \approx 0.0916$.

E10.7 Balanced coin hypothesis 2. **(a)** False, it is computed assuming H_0 is true. **(b)** False, it is computed assuming H_a is true. **(c)** True, since $H_a \neq H_0$. **(d)** True, since H_a is closer to H_0. **(e)** False, it is α. **(f)** (i) True, since the new RR is bigger. (ii) True, since the new RR is bigger. (iii) False, larger α implies smaller β.

***E10.8 Two-stage testing procedure.** Let A denote the event the study is terminated after the first stage, $H_0 : p = p_0 = 0.1$, and Y_i be the number of responders at each stage, $i = 1, 2$. Then

$$\begin{aligned}\alpha &= P(\text{reject } H_0 \mid H_0)\\ &= P(\text{reject } H_0, A \mid H_0) + P(\text{reject } H_0, \overline{A} \mid H_0)\\ &= P(\text{reject } H_0 \mid H_0, A)P(A \mid H_0) + P(Y_1 + Y_2 \geq 6, Y_1 \leq 3 \mid H_0)\\ &= (1)P(A \mid H_0) + P(Y_1 + Y_2 \geq 6, Y_1 \leq 3 \mid H_0).\end{aligned}$$

It is straightforward to show $P(A \mid H_0) = \sum_{i=4}^{15} \binom{15}{i} p_0^i (1-p_0)^{15-i} \approx 0.0556$. For the remaining probability, let $B \equiv (Y_1 + Y_2 \geq 6 \cap Y_1 \leq 3)$. Note $C_j \equiv (Y_1 = j)$, $j = 0, 1, 2, 3$ forms a partition of the sample space, hence by the total law of probability

$$P(B \mid H_0) = \sum_{j=0}^{3} P(B \mid H_0, C_j)P(C_j \mid H_0) = \sum_{j=0}^{3} P(Y_1{+}Y_2 \geq 6 \mid H_0, Y_1 = j)P(Y_1 = j \mid H_0) \approx 0.0433.$$

It follows that $\alpha \approx 0.0988$. β is computed similarly: assuming H_a to be true, let $p = 0.3$. Then

$$\begin{aligned}\beta &= P(\text{accept } H_0 \mid H_a)\\ &= P(\text{accept } H_0 \mid H_a, A)P(A \mid H_a) + P(\text{accept } H_0, \overline{A} \mid H_a)\\ &= (0)P(A \mid H_a) + P(Y_1 + Y_2 < 6, Y_1 \leq 3 \mid H_a)\\ &= \sum_{j=0}^{3} P(Y_1 + Y_2 < 6 \mid H_a, Y_1 = j)P(Y_1 = j \mid H_a) \approx 0.0679.\end{aligned}$$

10.3 Common Large-Sample Tests

$\boldsymbol{\alpha}$ significance level interpretation. In hypothesis testing, sometimes we will reject H_0 even if H_0 is true, i.e. commit type I errors. We want a procedure that is unlikely to commit such errors, which we can quantify by computing the probability α a type I error will occur and setting a significance level, e.g. $\alpha = 0.05$. As α is probabilistic in nature, we cannot make strong statements about a single observed sample and corresponding test statistic; a type I error was either made or not made. However in n repeated samples, we would expect αn type I errors. Thus if α is small and we observe an extreme test statistic, we can be reasonably confident that a type I error was not committed and instead attribute it as an outside effect H_a is based upon. Note the similarity to confidence levels in confidence intervals that do not make probabilistic statements about the observed interval, but provide a vague confidence level.

E10.17 Swim practice distances between two groups.

(a) $H_0 : \mu_1 = \mu_2$, $H_a : \mu_1 > \mu_2$.

(b) Invoking the CLT, $\overline{Y}_i \approx \mathcal{N}(\mu_i, S_i^2/n_i)$. Therefore $\Delta \equiv \overline{Y}_1 - \overline{Y}_2 \approx \mathcal{N}(\mu_1 - \mu_2, S_1^2/n_1 + S_2^2/n_2)$. and $Z = [\Delta - (\mu_1 - \mu_2)]/S \approx \mathcal{N}(0, 1)$ where $S^2 = S_1^2/n_1 + S_2^2/n_2$. Assuming H_0, $Z = \Delta/S$. We reject H_0 if z is large. Specifically, at level $\alpha = P(\text{reject } H_0 \mid H_0) = P(Z \geq z) = 1 - P(Z \leq z)$, or $z = z_\alpha$. The rejection region corresponds to values more extreme: $\{z > z_\alpha\}$.

(c) $z = [(\overline{y}_1 - \overline{y}_2) - 0]/s \approx 4.76$.

(d) z is much larger than $z_{0.01} \approx 2.33$, therefore we reject H_0 at the $\alpha = 0.01$ level.

(e) Medley swimmers cannot devote all their practice time to a single stroke.

10.4 Calculating Type II Error Probabilities and Finding the Sample Size for Z Tests

E10.43 Does sports participation increase dexterity scores?

(a) Let index 1 correspond to the sports group. We test $H_0 : \mu_1 = \mu_2$ with $H_a : \mu_1 > \mu_2$. Invoking the CLT, $\overline{Y}_i \approx \mathcal{N}(\mu_i, S_i^2/n_i)$, hence $Z = [(\overline{Y}_1 - \overline{Y}_2) - (\mu_1 - \mu_2)]/S \approx \mathcal{N}(0,1)$ where $S^2 = S_1^2/n_1 + S_2^2/n_2$. Assuming H_0, $Z = (\overline{Y}_1 - \overline{Y}_2)/S \approx \mathcal{N}(0,1)$. We reject H_0 if $z > z_\alpha$ at level $\alpha = P(\text{reject } H_0 \mid H_0) = 1 - P(Z \le z \mid H_0)$ since by definition $P(Z \le z_\alpha) = 1 - \alpha$. The observed value $z \approx 0.493$ is not so unusual, so we do not have sufficient evidence to reject H_0.

(b) $\beta = P(\text{accept } H_0 \mid H_a : \mu_1 - \mu_2 = 3)$. For fixed α, the rejection region is $(\overline{Y}_1 - \overline{Y}_2)/S > z_\alpha$. Assuming H_a, $Z = [(\overline{Y}_1 - \overline{Y}_2) - 3]/S \approx \mathcal{N}(0,1)$. Manipulating β to contain Z,

$$\beta = P\{(\overline{Y}_1 - \overline{Y}_2)/S \le z_\alpha \mid H_a\} = P\{Z \le z_\alpha - 3/S \mid H_a\} \approx 0.105.$$

E10.44 Does sports participation increase dexterity scores? 2. From E10.43(b), $z_{1-\beta} = z_\alpha - 3/S$ since by definition $P(Z \le z_{1-\beta}) = \beta$. Solving for $S(n_1, n_2)$ yields $S = 3/(z_\alpha - z_{1-\beta})$. Thus n_1 or n_2 must be chosen such that this equality holds. In the case $n = n_1 = n_2$, it can be inverted to obtain $n = (z_\alpha - z_{1-\beta})^2(S_1^2 + S_2^2)/9 \approx 47.7$, so we require $n = 48$ in each group.

10.5 Relationships Between Hypothesis-Testing Procedures and Confidence Intervals

10.6 Another Way to Report the Results of a Statistical Test: Attained Significance Levels, or p-Values

10.7 Some Comments on the Theory of Hypothesis Testing

10.8 Small-Sample Hypothesis Testing for μ and $\mu_1 - \mu_2$

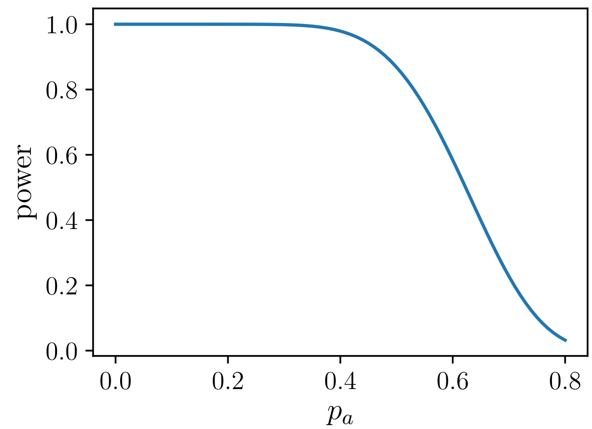
10.9 Testing Hypotheses Concerning Variances

Two-tailed p-values for asymmetric distributions. Suppose we have a random sample with test statistic t. Two-tailed p-values can be calculated as the probability T is more extreme than t assuming H_0. For symmetric distributions T, this amounts to $2P(T \ge |t| \mid H_0)$. The factor of 2 is clear because an extreme value of t could easily be the other extreme $-t$, by symmetry. For asymmetric distributions, e.g. χ^2, the 'extremely small' counterpart to an 'extremely large' region is unclear, e.g. see Fig. 10.10(c). It turns out it is common practice to still double a one-sided p-value. To see why this makes sense, prespecify a two-tailed rejection region at level α. A sensible construction distributes probability mass evenly between the tails [as in Fig. 10.10(c)], marking regions at both extremes with equal probability. For any value of t, we can modify α to just barely contain t on one tail; the region on the other tail has equal probability, by construction. The one-tailed p-value is easy to compute, and since the other extreme tail has the same probability, it is sufficient to double the one-tail p-values.

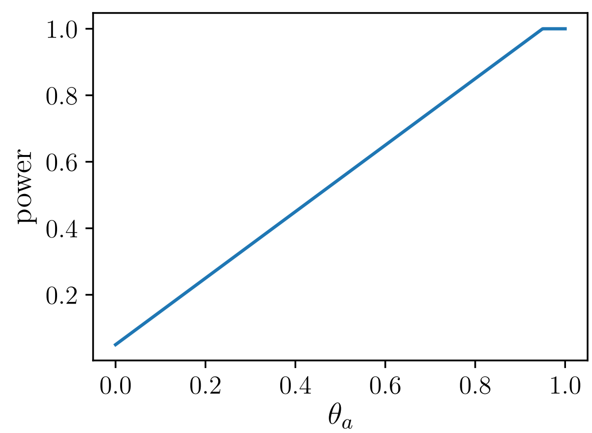
An alternate method defines extreme values in terms of likelihoods. In the symmetric T case, an extreme value t has a small likelihood. The probability T is more extreme than t then is the sum (integral) of likelihoods for the set of points $\{X\}$ with likelihoods $L(x) < L(t)$. For asymmetric distributions, the approach is analogous but the p-value is, in general, not doubled. For a more detailed discussion, see the following links: Stack Exchange 1, Stack Exchange 2, arxiv.

10.10 Power of Tests and the Neyman–Pearson Lemma

E10.88 Drug effectiveness hypothesis: power calculation.
In general, $\text{power}(p_a) = 1 - \beta(p_a) = P(\text{reject } H_0 \mid p_a) = P(Y \le 12 \mid p_a) = \sum_{i=1}^{12} p_a^i (1-p_a)^{12-k}$, plotted below. It tends to one as $p_a \to 0$ and to α as $p_a \to p_0$.



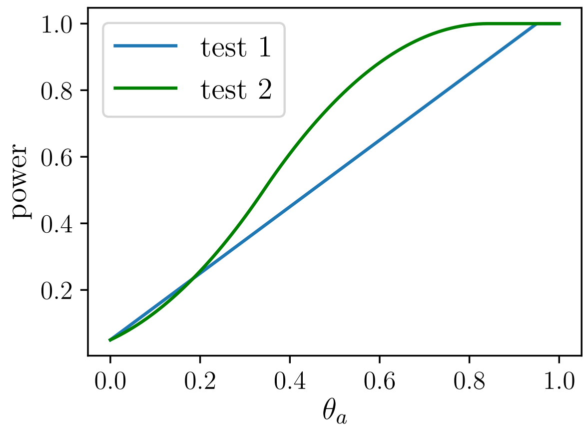
E10.89 Uniform sample hypothesis test: power calculation 1.
$\text{power}(\theta_a) = 1 - \beta(\theta_a) = P(\text{reject } H_0 \mid \theta_a) = P(Y_1 \ge 0.95 \mid \theta_a) = \int_{.95}^{\theta_a+1} dy = \theta_a + 0.05$, for $0 < \theta_a \le 0.95$, and is equal to 1 for $\theta_a > 0.95$.



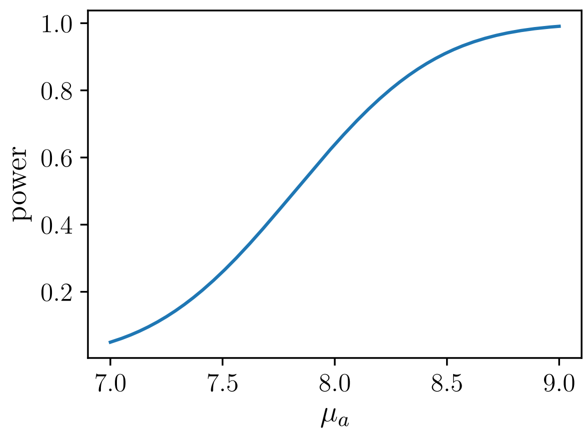
***E10.90 Uniform sample hypothesis test: power calculation 2.** From E10.5, use $c \approx 1.68$ so the α of the test is equal to E10.89. Also note the joint pdf $f_{Y_1,Y_2}(y_1, y_2 \mid \theta_a) = 1$ for $\theta_a \le y_i \le \theta_a + 1$.

$$
\begin{aligned}
\text{power}(\theta_a) &= P(\text{reject } H_0 \mid \theta_a) = P(Y_1 + Y_2 \ge c \mid \theta_a) \\
&= \begin{cases} \int_{c-1-\theta_a}^{\theta_a+1} dy_1 \int_{c-y_1}^{\theta_a+1} dy_2, & 0 < \theta_a < (c-1)/2 \\ \int_{\theta_a}^{c-\theta_a} dy_1 \int_{c-y_1}^{\theta_a+1} dy_2 + \int_{c-\theta_a}^{\theta_a+1} dy_1 \int_{\theta_a}^{\theta_a+1} dy_2, & (c-1/2) \le \theta_a \le c/2 \\ 1, & \theta_a > c/2 \end{cases} \\
&= \begin{cases} (\theta_a+1)^2/2 + (\theta_a+1)(-c+\theta_a+1) + (-c+\theta_a+1)^2/2, & 0 < \theta_a < (c-1)/2 \\ -c^2/2 + 2c\theta_a - 2\theta_a^2 + 1, & (c-1/2) \le \theta_a \le c/2 \\ 1, & \theta_a > c/2 \end{cases} .
\end{aligned}
$$

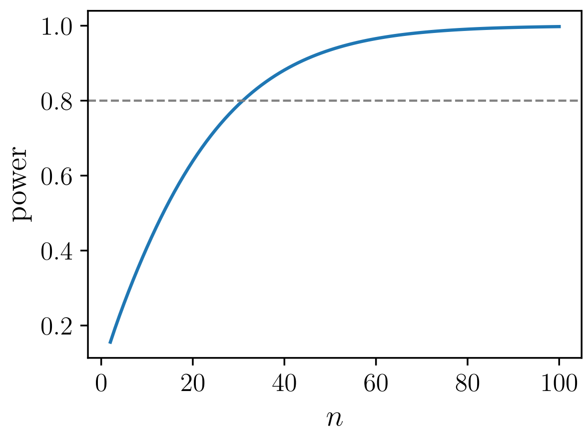
The power is overlaid in green with the E10.89 plot. Comparing the test powers, $\text{power}_2 \geq \text{power}_1$ for $\theta_a > c/2 + \sqrt{100c - 165}/20 - 3/4 \approx 0.184$. For smaller θ_a, test 1 has slightly more power.



E10.91 Normal sample hypothesis test: power calculation. Following Example 10.23, the most powerful test is a z-test based on statistic $\overline{Y} \sim \mathcal{N}(\mu, \sigma^2/n)$ with rejection region $\overline{y} > \mu_0 + z_\alpha \sigma/\sqrt{n} \equiv k$. $\text{power}(\mu_a) = P(\text{reject } H_0 \mid \mu_a) = P(\overline{Y} > k \mid \mu_a)$, which is obtained from the cdf of $\overline{Y} \sim \mathcal{N}(\mu_a, \sigma^2/n)$.

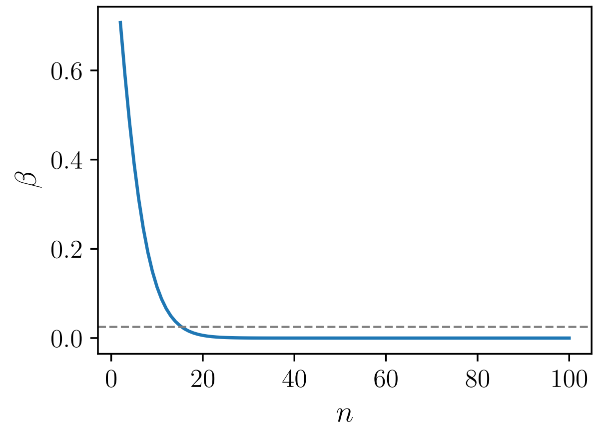


E10.92 Normal sample hypothesis test: power calculation 2.



As n changes, so does the rejection region and hence power. Numerically, we find $n = 31$.

E10.93 Normal sample hypothesis test: power calculation 3. Similar to E10.91, the most powerful test is a z-test based on statistic $\overline{Y} \sim \mathcal{N}(\mu, \sigma^2/n)$ with rejection region $\bar{y} < \mu_0 - z_\alpha \sigma/\sqrt{n} \equiv k$. $\beta(\mu_a) = P(\text{accept } H_0 \mid \mu_a) = P(\overline{Y} > k \mid \mu_a)$, which is obtained from the cdf of $\overline{Y} \sim \mathcal{N}(\mu_a, \sigma^2/n)$. Note the rejection region, and hence β, is a function of n. Numerically, we find $n = 16$.



***E10.97 Most powerful test for a particular discrete distribution.**

(a) $L(\theta \mid \mathbf{y}) = \prod_{i=1}^n p(y_i \mid \theta) = [p(1 \mid \theta)]^{N_1} [p(2 \mid \theta)]^{N_2} [p(3 \mid \theta)]^{N_3} = [\theta^2]^{N_1} [2\theta(1-\theta)]^{N_2} [(1-\theta)^2]^{N_3}$ such that $N_1 + N_2 + N_3 = n$. The likelihood simplifies to $L(\theta \mid \mathbf{y}) = 2^{N_2} \theta^{2N_1 + N_2} (1-\theta)^{N_2 + 2N_3}$.

(b) Applying Neyman–Pearson, the most powerful test has a rejection region determined by

$$\frac{L(\theta_0)}{L(\theta_a)} = \left(\frac{\theta_0}{\theta_a}\right)^{2N_1+N_2} \left(\frac{1-\theta_0}{1-\theta_a}\right)^{N_2+2N_3} < k$$

for constant k. Taking logarithms and substituting $N_3 = n - N_1 - N_2$,

$$(2N_1 + N_2) \log\left(\frac{\theta_0}{\theta_a}\right) + [2n - (2N_1 + N_2)] \log\left(\frac{1-\theta_0}{1-\theta_a}\right) < \log k.$$

Factoring $T = 2N_1 + N_2$ yields

$$T > \frac{\log k - 2nc_2}{c_1 - c_2} \equiv k'$$

where k', $c_1 = \log(\theta_0/\theta_a)$, and $c_2 = \log[(1-\theta_0)/(1-\theta_a)]$ are constants since n, θ_0, θ_a are known (the inequality flips because $c_1 - c_2 < 0$). Thus the rejection region is based on statistic $T = 2N_1 + N_2$. Tuning k' adjusts the significance level α.

(c) $\alpha = P(\text{reject } H_0 \mid H_0) = P(T < k' \mid H_0)$. For a given $H_0 : \theta = \theta_0$, we know the distribution of the sample, and hence (in theory) distributions of functions of the sample like T. Given the distribution of T, we can then choose constant k' so that $P(T < k' \mid \theta = \theta_0) = \alpha$.

(d) T is independent of θ_a and k' is a constant. Therefore for any θ_a, we would obtain the same rejection criteria, hence the test is uniformly most powerful.

E10.99 Most powerful test for $\boldsymbol{\lambda}$ in a Poisson random sample.

(a) The most powerful test has a rejection region determined by $L(\lambda_0)/L(\lambda_a) < k$. The likelihood function $L(\lambda) = \prod_{i=1}^n \lambda^{y_i} e^{-\lambda}/y_i! = \lambda^z e^{-n\lambda}/\prod_{i=1}^n y_i!$ where $z = \sum_{i=1}^n y_i$. It follows that $L(\lambda_0)/L(\lambda_a) = (\lambda_0/\lambda_a)^z e^{-n(\lambda_0 - \lambda_a)} < k$, or $z > [n(\lambda_0 - \lambda_a) + \log k] \log(\lambda_0/\lambda_a)$ since $\log(\lambda_0/\lambda_a) < 0$ for $\lambda_a > \lambda_0$. Since n, λ_0, λ_a are known, the r.h.s. is just a constant, say $z > k'$.

(b) Since $Z \sim \text{Poisson}(n\lambda)$, we can compute $\alpha = P(Z > k' \mid H_0 : \lambda = \lambda_0)$ and invert to obtain the rejection region boundary k' for desired α level.
(c) Yes, because Z and k' are independent of λ_a. Therefore the result holds for any λ_a.
(d) $z < k'$ since $\log(\lambda_0/\lambda_a) > 0$ for $\lambda_a > \lambda_0$.

E10.100 Most powerful test with two Poisson random samples.
Let $Z_1 = \sum_{i=1}^n Y_i$ and $Z_2 = \sum_{i=1}^m X_i$. It is straightforward to show $L(\lambda_1, \lambda_2) \propto \lambda_1^{z_1} e^{-n\lambda_1} \lambda_2^{z_2} e^{-m\lambda_2}$, hence $L(2,2)/L(1/2,3) = (4)^{z_1} e^{-3n/2} (2/3)^{z_2} e^m < k$, or $z_1 \log 4 + z_2 \log(2/3) < ke^{3n/2}e^{-m} \equiv k'$. The constant k' is chosen such that $\alpha = P(Z_1 \log 4 + Z_2 \log(2/3) < k' \mid H_0)$.

E10.101 Most powerful test for $\boldsymbol{\theta}$ in an exponential random sample.
(a) The most powerful test has a rejection region determined by $L(\theta_0)/L(\theta_a) < k$. The likelihood function $L(\theta) = \prod_{i=1}^n e^{y_i/\theta}/\theta = e^{z/\theta}/\theta^n$ where $z = \sum_{i=1}^n y_i$. It follows that the likelihood ratio $L(\theta_0)/L(\theta_a) = e^{z(1/\theta_0 - 1/\theta_a)}(\theta_a/\theta_0)^n < k$, or $z < k'$ for $\theta_a < \theta_0$ where the constant $k' = [\log k + n \log(\theta_0/\theta_a)]/(1/\theta_0 - 1/\theta_a)$. The constant k' is chosen at desired level $\alpha = P(Z < k' \mid H_0)$.
(b) Yes. Z is independent of θ_a and k' is a constant.

***E10.103 Most powerful test for $\boldsymbol{\theta}$ in a uniform random sample.** Using indicator variables and order statistics (e.g. E9.49), $L(\theta) = \theta^{-n} \prod_{i=1}^n I_{[0,\theta]}(y_i) = L(\theta) = \theta^{-n} I_{[0,\infty)}(y_{(1)}) I_{(-\infty,\theta]}(y_{(n)})$. Thus, the ratio $L(\theta_0)/L(\theta_a) = (\theta_a/\theta_0)^n I_{(-\infty,\theta_0]}(y_{(n)})/I_{(-\infty,\theta_a]}(y_{(n)}) < k$, or $I_{(-\infty,\theta_0)}(y_{(n)})/I_{(-\infty,\theta_a)}(y_{(n)}) < k'$ where $k' = k(\theta_0/\theta_a)^n$. The ratio only depends on the maximum $Y_{(n)}$ and since $\theta_a < \theta_0$, the l.h.s. is smaller for small $y_{(n)}$. Since the result is obtained for any $\theta_a < \theta_0$, the test is UMP.

***E10.104 Most powerful test for $\boldsymbol{\theta}$ in a uniform random sample 2.** As in E10.103, we arrive at $I_{(-\infty,\theta_0)}(y_{(n)})/I_{(-\infty,\theta_a)}(y_{(n)}) < k'$. Now since $\theta_a > \theta_0$, the l.h.s. is smaller for large $y_{(n)}$. Again, the test is UMP.

10.11 Likelihood Ratio Tests

Review of likelihood ratio tests. Likelihood ratio tests require a model of the underlying population. Typically this involves a random sample $X_1, X_2, \ldots, X_n$ for X_i generated from some distribution with pdf $f(x)$. Then likelihoods on parameters $\boldsymbol{\theta}$ given the data, $L(\boldsymbol{\theta} \mid \mathbf{x})$, can be computed for inference the parameters. In particular, we are interested in comparing two likelihoods: the maximum likelihood in a restricted model given by H_0, forcing parameters $\boldsymbol{\theta} \in \Omega_0$, against the maximum likelihood in an unrestricted model given by H_a, allowing parameters $\boldsymbol{\theta} \in (\Omega_0 \cup \Omega_a) \equiv \Omega$. This results in the likelihood ratio statistic $\lambda = \sup_{\boldsymbol{\theta} \in \Omega_0} L(\boldsymbol{\theta}) / \sup_{\boldsymbol{\theta} \in \Omega} L(\boldsymbol{\theta})$.

As $\Omega_0 \subset \Omega$, it must be that $0 \le \lambda \le 1$ with smaller values favoring rejection of H_0. We define a rejection region based on this intuition: $\lambda < k$ for some cutoff $k \in [0,1]$. We determine how small k must be by setting a significance level $\alpha = P(\text{reject } H_0 \mid H_0)$. For simple (non-composite) H_0, we know in theory the distribution of X_i and hence functions of X_i. As λ is a function of the data $\mathbf{X}$, we also know the distribution of λ, say $g(\lambda)$ defined over support $0 \le \lambda \le 1$. k is then determined by $\alpha = \int_0^k d\lambda\, g(\lambda)$. In practice, the distribution of λ is difficult to determine analytically. It can be approximated by Monte Carlo simulation as follows: simulate random samples assuming the data is distributed with constraints imposed by H_0 and compute λ for each sample. Alternatively for large samples, we can appeal to an asymptomatic result: for large n, $-2\log\lambda \approx \chi^2(\nu)$, where ν is the difference in dimension of the unrestricted and restricted parameter spaces, Ω and Ω_0. The rejection region can then be obtained from the χ^2 distribution.

E10.105 Equivalence of likelihood ratio test of variance to $\boldsymbol{\chi^2}$ test. Let $\boldsymbol{\theta} = (\mu, \sigma^2)$. The parameter space of interest is the set $\Omega = \{\boldsymbol{\theta} : \mu \in \mathbb{R}, \sigma^2 \ge \sigma_0^2\}$. $H_0 : \sigma^2 = \sigma_0^2$ restricts $\boldsymbol{\theta}$ to the subset $\Omega_0 = \{\boldsymbol{\theta} : \mu \in \mathbb{R}, \sigma^2 = \sigma_0^2\}$. We are interested in test statistic $\lambda = \sup_{\boldsymbol{\theta} \in \Omega_0} L(\boldsymbol{\theta}) / \sup_{\boldsymbol{\theta} \in \Omega} L(\boldsymbol{\theta})$. The likelihood function of a normal random sample is

$$L(\boldsymbol{\theta}) = \frac{1}{(2\pi\sigma^2)^{n/2}} e^{-\sum_{i=1}^n (y_i-\mu)^2/(2\sigma^2)}.$$

As $\log(\cdot)$ is a monotonic function, it is equivalent and often easier to maximize the log-likelihood

$$l(\boldsymbol{\theta}) \equiv \log L(\boldsymbol{\theta}) = -\frac{n}{2}\log(2\pi\sigma^2) - \frac{1}{2\sigma^2}\sum_{i=1}^n (y_i-\mu)^2.$$

Restricting to Ω_0, there is only one free parameter μ:

$$l(\boldsymbol{\theta} \mid \Omega_0) = l(\mu) = -\frac{n}{2}\log(2\pi\sigma_0^2) - \frac{1}{2\sigma_0^2}\sum_{i=1}^n (y_i-\mu)^2.$$

Differentiating,

$$\frac{\partial l}{\partial \mu} = \frac{1}{\sigma_0^2}\sum_{i=1}^n (y_i-\mu) = \frac{n}{\sigma_0^2}(\overline{y}-\mu) = 0$$

when $\mu = \overline{y}$. This indeed is a maximum since $\partial^2 l(\overline{y})/\partial\mu^2 = -n/\sigma_0^2 < 0$, so

$$\sup_{\boldsymbol{\theta}\in\Omega_0} L(\boldsymbol{\theta}) = \frac{1}{(2\pi\sigma_0^2)^{n/2}} e^{-\sum_{i=1}^n (y_i-\overline{y})^2/(2\sigma_0^2)}.$$

Considering now Ω, there are two free parameter, μ and σ^2, and we must set $\partial l(\boldsymbol{\theta} \mid \Omega)/\partial\mu = 0$ and $\partial l(\boldsymbol{\theta} \mid \Omega)/\partial\sigma^2 = 0$. Differentiating yields the set of equations

$$\frac{n}{\sigma^2}(\overline{y}-\mu) = 0, \qquad -\frac{n}{2\sigma^2} + \frac{1}{2\sigma^4}\sum_{i=1}^n (y_i-\mu)^2 = 0.$$

We conclude $\mu = \overline{y}$ from the first. Then the second yields $\sigma^2 = \sum_{i=1}^n (y_i - \bar{y})^2/n$ for $\sigma^2 \geq \sigma_0^2$. Indeed the Hessian determinant can verify this is a maximum, thus

$$\sup_{\boldsymbol{\theta}\in\Omega} L(\boldsymbol{\theta}) = \frac{1}{[2\pi\sum_{i=1}^n (y_i-\bar{y})^2/n]^{n/2}} e^{-n/2}.$$

Inserting $s^2 = \sum_{i=1}^n (y_i - \bar{y})^2/(n-1)$, our test statistic becomes

$$\lambda = \frac{\sup_{\boldsymbol{\theta}\in\Omega_0} L(\boldsymbol{\theta})}{\sup_{\boldsymbol{\theta}\in\Omega} L(\boldsymbol{\theta})} = \left[\frac{(n-1)s^2}{n\sigma_0^2}\right]^{n/2} \exp\left\{\left[\frac{n}{2} - \frac{(n-1)s^2}{2\sigma_0^2}\right]\right\} = \left[\frac{w_0}{n}\right]^{n/2} \exp\left\{\left[\frac{n}{2} - \frac{w_0}{2}\right]\right\}$$

where $w_0 \equiv (n-1)s^2/\sigma_0^2$. (In the event $\sigma^2 < \sigma_0^2$, due to the one-sided nature of H_a, we set $\sigma^2 = \sigma_0^2$. Then $\lambda = 1$ and we do not reject H_0.) The rejection region is determined by $\lambda \leq k$ for some constant k. This is equivalent to $-2\log\lambda \geq -2\log k$, or

$$w_0 - n\log(w_0/n) - n \geq -2\log k.$$

The r.h.s. is a constant, say k', and the l.h.s. is a function $f = f(w_0)$. $f(w_0)$ is large for $w_0 \to 0$ and $w_0 \to \infty$, or equivalently for $s^2/\sigma_0^2 \to 0$ and $s^2/\sigma_0^2 \to \infty$. Since we are conducting $H_a : \sigma^2 > \sigma_0^2$, assuming H_a to be true, we would expect s^2/σ_0^2 to be large. Therefore the likelihood test rejection region $f(w_0) \geq k'$ for sufficiently large k' is equivalent to $w_0 \geq k''$ for sufficiently large k''; the latter is the usual χ^2 test. The constants are determined by setting a level $\alpha = P(\text{reject } H_0 \mid H_0)$.

Comments. (1) The argument is similar for lower- and two-tailed hypotheses. For $H_a : \sigma^2 < \sigma_0^2$, $f(w_0)$ is large for small w_0, hence the equivalent tests are $f(w_0) \geq k'$ and $w_0 \leq k''$. $H_a : \sigma^2 \neq \sigma_0^2$ has a rejection region at each tail since $f(w_0)$ is large whenever w_0 is small or large.

(2) As $n \to \infty$, $f(w_0) \approx \chi^2(1)$. This allows us to determine the constant k'. For example, testing $H_a : \sigma^2 \neq \sigma_0^2$ has rejection region $f(w_0) \geq \chi^2_\alpha$. This is equivalent to a two-sided χ^2 test with $(n-1)$ d.f. where $W_0 = (n-1)S^2/\sigma_0^2$ with rejection region $w_0 < \chi^2_{1-\alpha/2}$ or $w_0 > \chi^2_{\alpha/2}$. (The likelihood ratio test appears to have a one-sided rejection region, but that is because $f(w_0)$ is a complicated function of w_0. Small and large values of w_0 are mapped to large values of $f(w_0)$, so the test still rejects H_0 for large or small w_0 with the same significance level as the χ^2 test.)

E10.106 Independent binomial samples hypothesis. Let $Y_i \sim \text{binom}(n_i, p_i)$ represent the number who favor A in each ward, $i = 1, 2, 3, 4$. We want to test $H_0 : p_1 = p_2 = p_3 = p_4$. Let $\boldsymbol{\theta} = (p_1, p_2, p_3, p_4)$. The parameter space of interest is the set $\Omega = \{\boldsymbol{\theta} : 0 < p_i < 1\}$. H_0 restricts $\boldsymbol{\theta}$ to the subset $\Omega_0 = \{\boldsymbol{\theta} : p_i = p, 0 < p < 1\}$. We are interested in test statistic $\lambda = \sup_{\boldsymbol{\theta}\in\Omega_0} L(\boldsymbol{\theta}) / \sup_{\boldsymbol{\theta}\in\Omega} L(\boldsymbol{\theta})$ and in particular $T = -2\log\lambda = 2[\sup_{\boldsymbol{\theta}\in\Omega} l(\boldsymbol{\theta}) - \sup_{\boldsymbol{\theta}\in\Omega_0} l(\boldsymbol{\theta})]$ where $l(\boldsymbol{\theta}) = \log L(\boldsymbol{\theta})$ since $T \approx \chi^2(3)$. The likelihood and log-likelihood are

$$L(\boldsymbol{\theta}) = \prod_{i=1}^4 p_i^{y_i}(1-p_i)^{n_i - y_i} \qquad \Rightarrow \qquad l(\boldsymbol{\theta}) = \sum_{i=1}^4 \left[y_i \log p_i + (n_i - y_i)\log(1-p_i)\right].$$

Restricting to Ω_0, there is only one free parameter p:

$$l(\boldsymbol{\theta} \mid \Omega_0) = \sum_{i=1}^4 [y_i \log p + (n_i - y_i)\log(1-p)] = y\log p + (N-y)\log(1-p)$$

where $y = \sum_{i=1}^4 y_i$ and $N = \sum_{i=1}^4 n_i$. Setting $\partial l(\boldsymbol{\theta} \mid \Omega_0)/\partial p = 0$ yields $p = y/N$, which can be verified to be the maximum. Thus

$$\sup_{\boldsymbol{\theta}\in\Omega_0} l(\boldsymbol{\theta}) = y\log(y/N) + (N-y)\log(1-y/N).$$

Considering all of Ω, now with four parameters, we must solve the four equations $\partial l(\boldsymbol{\theta} \mid \Omega)/\partial p_i = 0$. The resultant maximum is $p_i = y_i/n_i$, hence

$$\sup_{\boldsymbol{\theta}\in\Omega} l(\boldsymbol{\theta}) = \sum_{i=1}^4 [y_i \log(y_i/n_i) + (n_i - y_i)\log(1 - y_i/n_i)]$$

Since $T \approx \chi^2(3)$, an appropriate bound $T \geq k$ is $k = \chi^2_\alpha \approx 7.81$. We observe $t \approx 10.5$ and thus reject the null. This corresponds to p-value ≈ 0.0145.

Comment. This test is sometimes referred to as the G-test. We could also try Pearson's χ^2 test, which compares observed with expected frequencies assuming H_0, yielding a similar result: p-value ≈ 0.0133. Asymptotically, the two tests are the same.

***E10.109 Independent exponential samples hypothesis.**

(a) Let $\boldsymbol{\theta} = (\theta_1, \theta_2)$. The parameter space of interest is the set $\Omega = \{\boldsymbol{\theta} : \theta_1 > 0, \theta_2 > 0\}$. H_0 restricts $\boldsymbol{\theta}$ to the subset $\Omega_0 = \{\boldsymbol{\theta} : \theta_1 = \theta_2 = \theta, \theta > 0\}$. We are interested in test statistic $\lambda = \sup_{\boldsymbol{\theta}\in\Omega_0} L(\boldsymbol{\theta}) / \sup_{\boldsymbol{\theta}\in\Omega} L(\boldsymbol{\theta})$. The likelihood function

$$L(\boldsymbol{\theta}) = \prod_{i=1}^m \frac{1}{\theta_1} e^{-x_i/\theta_1} \prod_{j=1}^n \frac{1}{\theta_2} e^{-y_j/\theta_2} = \theta_1^{-m}\theta_2^{-n} e^{-(u/\theta_1 + v/\theta_2)}$$

where $u = \sum_{i=1}^m x_i$ and $v = \sum_{i=j}^n y_j$. It is easier to maximize the log-likelihood

$$l(\boldsymbol{\theta}) = -(u/\theta_1 + v/\theta_2) - m\log\theta_1 - n\log\theta_2.$$

Maximizing over Ω, there are two free parameters θ_i, $l = 1, 2$. Setting $\partial l(\boldsymbol{\theta})/\partial\theta_i = 0$ yields MLEs $\hat{\theta}_1 = u/m = \overline{x}$ and $\hat{\theta}_2 = v/n = \overline{y}$, hence

$$\sup_{\boldsymbol{\theta}\in\Omega} L(\boldsymbol{\theta}) = (\overline{x})^{-m}(\overline{y})^{-n}e^{-(m+n)}.$$

Restricting to Ω_0, there is only one free parameter $\theta = \theta_1 = \theta_2$ with MLE $\hat{\theta} = (u+v)/(m+n) \equiv \overline{z}$, hence

$$\sup_{\boldsymbol{\theta}\in\Omega_0} L(\boldsymbol{\theta}) = (\overline{z})^{-(m+n)}e^{-(m+n)}.$$

It follows that

$$\lambda = \overline{x}^m\overline{y}^n/\overline{z}^{m+n}.$$

Notice that $\lambda = 1$ under $H_0 : \theta_1 = \theta_2$ and $\lambda < 1$ for $\theta_1 \neq \theta_2$. Thus, we reject H_0 for sufficiently small $\lambda < k$ with k determined such that $\alpha = P(\text{reject } H_0 \mid H_0)$.

(b) The sum of i.i.d. exponential r.v. is a gamma r.v. Specifically, $U \sim \text{gamma}(m, \theta_1)$ and $V \sim \text{gamma}(n, \theta_2)$. From E6.12, $2U/\theta_1 \sim \text{gamma}(m, 2) \simeq \chi^2(2m)$. Similarly, $2V/\theta_2 \sim \chi^2(2n)$. It follows that

$$F = \frac{(2U/\theta_1)/(2m)}{(2V/\theta_2)/(2n)} = \frac{\overline{X}/\theta_1}{\overline{Y}/\theta_2}$$

is F-distributed. Further assuming H_0, $F_0 = \overline{X}/\overline{Y}$ is F-distributed, and an F-test rejects H_0 at extreme values of F_0. Expressing the likelihood ratio statistic in terms of F_0,

$$\Lambda = (m+n)^{m+n}\left[\frac{\overline{X}}{m\overline{X}+n\overline{Y}}\right]^m\left[\frac{\overline{Y}}{m\overline{X}+n\overline{Y}}\right]^n = (m+n)^{m+n}\left[\frac{1}{m+nF_0^{-1}}\right]^m\left[\frac{1}{mF_0+n}\right]^n.$$

At extreme values of F_0, Λ is small. Recall for small Λ, the likelihood ratio test rejects H_0. Hence, the F-test and likelihood ratio test are equivalent.

E10.111 Neyman–Pearson lemma as a special case of a likelihood ratio test.

(a)
$$\lambda = \frac{\sup_{\theta\in\Omega_0} L(\theta)}{\sup_{\theta\in\Omega} L(\theta)} = \frac{L(\theta_0)}{\sup\{L(\theta_0), L(\theta_a)\}} = \frac{1}{\sup\{1, L(\theta_a)/L(\theta_0)\}} = \inf\left\{1, \frac{L(\theta_0)}{L(\theta_a)}\right\}.$$

(b) The LR test rejection region is $\lambda < k$ for some constant $k \in [0,1]$. Since $L(\theta_0) \leq L(\theta_a)$, $\inf\{1, L(\theta_0)/L(\theta_a)\} = L(\theta_0)/L(\theta_a)$, hence we can instead consider $L(\theta_0)/L(\theta_a) \leq k'$, $k' \in [0,1]$.

(c) The likelihood ratio test reduces to the Neyman–Pearson lemma.

E10.112 Equivalence of likelihood ratio test and two-sample t-test, 1-sided. Let the samples be $X_1, X_2, \ldots, X_{n_1}$ and $Y_1, Y_2, \ldots, Y_{n_2}$. We want to test $H_0 : \mu_1 - \mu_2 = \Delta$ with $H_a : \mu_1 - \mu_2 > \Delta$. Let $\boldsymbol{\theta} = (\mu_1, \mu_2, \sigma^2)$; the full parameter space of interest is the set $\Omega = \{\boldsymbol{\theta} : \mu_1 - \mu_2 \geq \Delta, \mu_2 \in \mathbb{R}, \sigma^2 > 0\}$. H_0 restricts the space from 3- to 2-dimensional: $\Omega_0 = \{\boldsymbol{\theta} : \mu_1 = \mu_2 + \Delta, \mu_2 \in \mathbb{R}, \sigma^2 > 0\}$. The likelihood ratio is $\lambda = \sup_{\boldsymbol{\theta}\in\Omega_0} L(\boldsymbol{\theta})/\sup_{\boldsymbol{\theta}\in\Omega} L(\boldsymbol{\theta})$ with rejection region $\lambda < k$ for some constant $k \in [0,1]$ set by the significance level.

The likelihood function is

$$\begin{aligned} L(\boldsymbol{\theta}) &= \prod_{i=1}^{n_1}\frac{1}{(2\pi\sigma^2)^{1/2}}e^{-(x_i-\mu_1)^2/(2\sigma^2)}\prod_{j=1}^{n_2}\frac{1}{(2\pi\sigma^2)^{1/2}}e^{-(y_j-\mu_2)^2/(2\sigma^2)} \\ &= \frac{1}{(2\pi\sigma^2)^{N/2}}\exp\left[-\frac{1}{2\sigma^2}\left(\sum_{i=1}^{n_1}(x_i-\mu_1)^2 + \sum_{j=1}^{n_2}(y_j-\mu_2)^2\right)\right] \end{aligned}$$

where $N = n_1 + n_2$. It is easier to maximize the log-likelihood

$$l(\boldsymbol{\theta}) \equiv \log L(\boldsymbol{\theta}) = -\frac{N}{2}\log(2\pi\sigma^2) - \frac{1}{2\sigma^2}\left(\sum_{i=1}^{n_1}(x_i-\mu_1)^2 + \sum_{j=1}^{n_2}(y_j-\mu_2)^2\right).$$

Maximizing in Ω, we simultaneously solve $\partial l(\boldsymbol{\theta})/\partial\mu_i = 0$, $i = 1, 2$, and $\partial l(\boldsymbol{\theta})/\partial\sigma^2 = 0$. This results in MLEs $\hat{\mu}_1 = \overline{x}$, $\hat{\mu}_2 = \overline{y}$, and $\widehat{\sigma^2} = \left(\sum_{i=1}^{n_1}(x_i-\overline{x})^2 + \sum_{j=1}^{n_2}(y_j-\overline{y})^2\right)/N$ for $\overline{x}_1 - \overline{x}_2 \geq \Delta$. In the event $\overline{x} - \overline{y} < \Delta$, $\sup_{\boldsymbol{\theta}\in\Omega} L(\boldsymbol{\theta}) = \sup_{\boldsymbol{\theta}\in\Omega_0} L(\boldsymbol{\theta})$ and $\lambda = 1$, i.e. we do not reject H_0. We restrict the rest of the analysis to $\overline{x}_1 - \overline{x}_2 \geq \Delta$. Inserting the sample variances, it follows

$$\sup_{\boldsymbol{\theta}\in\Omega} L(\boldsymbol{\theta}) = \frac{1}{\{2\pi[(n_1-1)s_1^2 + (n_2-1)s_2^2]/N\}^{N/2}} e^{-N/2}.$$

Maximizing in Ω_0,

$$l(\boldsymbol{\theta} \mid \Omega_0) = -\frac{N}{2}\log(2\pi\sigma^2) - \frac{1}{2\sigma^2}\left(\sum_{i=1}^{n_1}(x_i-\mu_2-\Delta)^2 + \sum_{j=1}^{n_2}(y_j-\mu_2)^2\right),$$

with MLEs $\hat{\mu}_{2,0} = (n_1\overline{x} + n_2\overline{y} - n_1\Delta)/N$ and $\widehat{\sigma_0^2} = \left(\sum_{i=1}^{n_1}(x_i-\hat{\mu}_{2,0}-\Delta)^2 + \sum_{j=1}^{n_2}(y_j-\hat{\mu}_{2,0})^2\right)/N$. The sums simplify as follows:

$$\begin{aligned}
\sum_{i=1}^{n_1}(x_i-\hat{\mu}_{2,0}-\Delta)^2 &= \sum_{i=1}^{n_1}(x_i-\overline{x}+\overline{x}-\hat{\mu}_{2,0}-\Delta)^2 \\
&= \sum_{i=1}^{n_1}(x_i-\overline{x})^2 + \sum_{i=1}^{n_1}(\overline{x}-\hat{\mu}_{2,0}-\Delta)^2 + 2(\overline{x}-\hat{\mu}_{2,0}-\Delta)\sum_{i=1}^{n_1}(x_i-\overline{x}) \\
&= (n_1-1)s_1^2 + n_1(\overline{x}-\hat{\mu}_{2,0}-\Delta)^2 \\
&= (n_1-1)s_1^2 + (n_1n_2^2/N^2)(\overline{x}-\overline{y}-\Delta)^2.
\end{aligned}$$

Similarly,

$$\sum_{j=1}^{n_2}(y_i-\hat{\mu}_{2,0})^2 = (n_2-1)s_2^2 + (n_1^2n_2/N^2)(\overline{x}-\overline{y}-\Delta)^2.$$

Thus

$$\widehat{\sigma_0^2} = [(n_1-1)s_1^2 + (n_2-1)s_2^2 + (n_1n_2/N)(\overline{x}-\overline{y}-\Delta)^2]/N$$

and

$$\sup_{\boldsymbol{\theta}\in\Omega_0} L(\boldsymbol{\theta}) = \frac{1}{\left\{2\pi[(n_1-1)s_1^2 + (n_2-1)s_2^2 + (n_1n_2/N)(\overline{x}-\overline{y}-\Delta)^2]/N\right\}^{N/2}} e^{-N/2}.$$

The likelihood ratio is then

$$\lambda = \left[\frac{(n_1-1)s_1^2 + (n_2-1)s_2^2}{(n_1-1)s_1^2 + (n_2-1)s_2^2 + (n_1n_2/N)(\overline{x}-\overline{y}-\Delta)^2}\right]^{N/2}.$$

The rejection region $\lambda < k$ is equivalent to $\lambda^{-2/N} > k^{-2/N}$, or

$$\frac{(n_1-1)s_1^2 + (n_2-1)s_2^2 + (n_1n_2/N)(\overline{x}-\overline{y}-\Delta)^2}{(n_1-1)s_1^2 + (n_2-1)s_2^2} = 1 + \frac{(n_1n_2/N)(\overline{x}-\overline{y}-\Delta)^2}{(n_1-1)s_1^2 + (n_2-1)s_2^2} > k^{-2/N}.$$

Defining $s_p^2 = [(n_1-1)s_1^2+(n_2-1)s_2^2]/(N-2)$ and $t = (\overline{x}-\overline{y}-\Delta)/(s_p\sqrt{1/n_1+1/n_2})$, this becomes

$$t^2 > (k^{-2/N}-1)/(N-2) \qquad \Rightarrow \qquad |t| > \sqrt{(k^{-2/N}-1)/(N-2)}.$$

The l.h.s. contains the usual two-sample t-statistic and the r.h.s. is just a constant $k' \in (0,\infty)$. Furthermore in testing $H_a : \mu_1 - \mu_2 > \Delta$, the above analysis holds only for $\overline{x}-\overline{y} \geq \Delta$ (otherwise, we fail to reject H_0 as discussed earlier). Then $t \geq 0$ and the test is just the one-sided $t > k'$, i.e. we reject H_0 for large t. Thus, this is exactly equivalent to the two-sample t-test. The constant k' is set by the significance level $\alpha = P(\text{reject } H_0 \mid H_0)$. Under H_0, the statistic $T = [\overline{X}-\overline{Y}-(\mu_1-\mu_2)]/(S_p\sqrt{1/n_1+1/n_2})$ follows a t distribution with $(N-2)$ d.f., hence under H_0, $T = (\overline{X}-\overline{Y}-\Delta)/(S_p\sqrt{1/n_1+1/n_2})$ is t-distributed. This allows us to determine $k' = t_\alpha$.

E10.113 Equivalence of likelihood ratio test and two-sample *t*-test, 2-sided. The derivation is almost identical to E10.112. In the 2-sided case, $H_0 : \mu_1 - \mu_2 = \Delta$ with $H_a : \mu_1 - \mu_2 \neq \Delta$, hence $\Omega = \{\boldsymbol{\theta} : \mu_1 \in \mathbb{R}, \mu_2 \in \mathbb{R}, \sigma^2 > 0\}$. t is not restricted to the nonnegatives, hence we have rejection region $|t| > k'$. Weighting each side equally, we reject H_0 if $|t| > t_{\alpha/2}$.

***E10.114 Equivalence of *F*-test and likelihood ratio test testing $\boldsymbol{\mu_1 = \mu_2 = \mu_3}$.** Let the samples be $X_{i1}, X_{i2}, \ldots, X_{in_i}$, $i = 1,2,3$. $Y_1, Y_2, \ldots, Y_{n_2}$. We want to test $H_0 : \mu_1 = \mu_2 = \mu_3$ with H_a : at least one inequality. Let $\boldsymbol{\theta} = (\mu_1, \mu_2, \mu_2, \sigma^2)$; the full parameter space of interest is the set $\Omega = \{\boldsymbol{\theta} : \mu_i \in \mathbb{R}, \sigma^2 > 0\}$. H_0 restricts the space from 4- to 2-dimensional: $\Omega_0 = \{\boldsymbol{\theta} : \mu_1 = \mu_2 = \mu_3 \equiv \mu, \mu \in \mathbb{R}, \sigma^2 > 0\}$. The likelihood ratio is $\lambda = \sup_{\boldsymbol{\theta}\in\Omega_0} L(\boldsymbol{\theta})/\sup_{\boldsymbol{\theta}\in\Omega} L(\boldsymbol{\theta})$ with rejection region $\lambda < k$ for some constant $k \in [0,1]$ set by the significance level.

The likelihood function is

$$L(\boldsymbol{\theta}) = \prod_{i=1}^{3}\prod_{j=1}^{n_i} \frac{1}{(2\pi\sigma^2)^{1/2}} e^{-(x_{ij}-\mu_i)^2/(2\sigma^2)} = \frac{1}{(2\pi\sigma^2)^{N/2}} \exp\left[-\frac{1}{2\sigma^2}\sum_{i=1}^{3}\sum_{j=1}^{n_i}(x_{ij}-\mu_i)^2\right]$$

where $N = n_1 + n_2 + n_3$. It is easier to maximize the log-likelihood

$$l(\boldsymbol{\theta}) \equiv \log L(\boldsymbol{\theta}) = -\frac{N}{2}\log(2\pi\sigma^2) - \frac{1}{2\sigma^2}\sum_{i=1}^{3}\sum_{j=1}^{n_i}(x_{ij}-\mu_i)^2.$$

The resultant MLEs are $\hat{\mu}_i = \overline{x}_i$ and $\widehat{\sigma^2} = \sum_{i=1}^3\sum_{j=1}^{n_i}(x_{ij}-\overline{x}_i)^2/N = \sum_{i=1}^3(n_i-1)s_i^2/N$. Restricting to Ω_0,

$$l(\boldsymbol{\theta} \mid \Omega_0) = -\frac{N}{2}\log(2\pi\sigma^2) - \frac{1}{2\sigma^2}\sum_{i=1}^{3}\sum_{j=1}^{n_i}(x_{ij}-\mu)^2.$$

The resultant MLEs are $\hat{\mu}_0 = \sum_{i=1}^3 n_i\overline{x}_i/N$ and $\widehat{\sigma_0^2} = \sum_{i=1}^3\sum_{j=1}^{n_i}(x_{ij}-\hat{\mu}_0)^2/N$. The latter simplifies to

$$\widehat{\sigma_0^2} = \sum_{i=1}^{3}(n_i-1)s_i^2/N + \sum_{i=1}^{3} n_i(\overline{x}_i - \hat{\mu}_0)^2/N.$$

We can then compute the likelihood ratio $\lambda = \left(\widehat{\sigma^2}/\widehat{\sigma_0^2}\right)^{N/2}$ with rejection region $\lambda < k$. This is equivalent to $\lambda^{-2/N} > k^{-2/N}$, or after some algebra,

$$\frac{\sum_{i=1}^3 n_i(\overline{x}_i-\hat{\mu}_0)^2}{\sum_{i=1}^3(n_i-1)s_i^2} > k^{-2/N} - 1.$$

The l.h.s. is F distributed with 2 numerator d.f. and $(N-3)$ denominator d.f. (see E13.6 for a proof), hence the likelihood ratio test is equivalent to an exact F-test.

10.12 Summary

Overview of hypothesis testing. All the small-sample tests in Sections 10.8 and 10.9 are in fact likelihood-ratio (LR) tests, the t-test being the most well known. LR tests rely on a model of the population, e.g. normally-distributed in the t-test with unknown parameters μ, σ^2, but tend to have good properties (e.g. Neyman–Pearson lemma). The distribution of the LR is often unknown for finite samples as it can be a complicated function of the data, but the LR is asymptotically χ^2 distributed and hence a rejection region at (approximate) significance level α can be specified Specifically, $-2\log(\text{LR}) \approx \chi^2(\nu)$ where ν is the difference in dimensions between the restricted and unrestricted model, with rejection region $-2\log(\text{LR}) \geq \chi^2_\alpha$.

There are other likelihood based tests, namely the Wald and score test. As they are based on likelihoods, they rely on a model of the population. In fact, we've already encountered Wald tests; they have a test statistic of the form

$$W = \frac{\hat{\theta} - \theta_0}{\sigma_{\hat{\theta}}}$$

where $\hat{\theta}$ is an MLE of the unrestricted model. It can be viewed as the distance between the best estimate and the estimate posed by H_0, weighted by the standard error of the best estimate. A rejection region is constructed by extreme values of W. In finite samples, the distribution of W is often unknown, but many estimates have known asymptotic distributions. For instance, if we want to estimate the mean μ of a population with finite variance σ^2, the sample mean $\overline{X}$ is often an MLE with $\overline{X} \approx \mathcal{N}(\mu, \sigma^2/n)$ for large samples by the CLT. Then the Wald statistic has the form

$$W = \frac{\overline{X} - \mu_0}{\sigma_{\overline{X}}} \approx \mathcal{N}(0,1) \qquad \text{assuming } H_0 : \mu = \mu_0,$$

and we can construct an approximate rejection region using the normal distribution. This is precisely the z-test. More generally, under regularity assumptions, the MLE converges to a normal distribution (e.g., see Section 9.8) in which case $W \approx \mathcal{N}(0,1)$ assuming H_0.

It can be shown that all three likelihood-based tests are asymptotically similar. LR tests tend to have the best statistical properties, while the other tests may be computationally less expensive. There are also tests not based on likelihood, e.g. permutation tests (like Fisher exact test) and many goodness of fit tests (like Kolmogorov–Smirnov test). These tests do not need to assume a parametric form for the underlying model.

Supplemental Exercises

E10.115 Hypothesis testing: true or false. **(a)** True. **(b)** False. α is the probability the null is rejected assuming the null is true. **(c)** False. Statistical significance does not say anything about the effect size. **(d)** True. **(e)** False. It always assumes H_a is true. **(f)** False. We require $p < \alpha$. **(g)** False, by definition of UMP. **(h)** False, $\Omega_0 \subset \Omega$. **(i)** True.

E10.117 Difference in chemical and atmospheric nitrogen densities.

(a) Assuming the masses are normally distributed (or close to it), we can apply a two-sided two-sample t-test. Formally, we test $H_0 : \mu_1 = \mu_2$ with $H_a : \mu_1 \neq \mu_2$. Assuming H_0 to be true, the test statistic $T = (\overline{Y}_1 - \overline{Y}_2)/(S_p/\sqrt{1/n_1 + 1/n_2})$ is t-distributed with $(n_1 + n_2 - 2)$ d.f., where the pooled sample variance $S_p^2 = [(n_1 - 1)S_1^2 + (n_2 - 1)S_2^2]/(n_1 + n_2 - 2)$. The resultant p-value of the test is $p = P(T \leq -|t|) + P(T \geq |t|) \approx 2.15 \times 10^{-4}$. The small p-value can be interpreted as follows: the probably of observing the data or more extreme assuming H_0 to be true is small, so we reject H_0.

(b) A two-sided confidence interval is $(\overline{y}_1 - \overline{y}_2) \pm t_{\alpha/2} s_p \sqrt{1/n_1 + 1/n_2} \approx [-0.0115, -0.00951]$.
(c) There is sufficient evidence because $H_0 : \mu_1 - \mu_2 = 0$ is not contained in the interval.
(d) There is no conflict due to the duality between confidence intervals and hypothesis tests.

E10.119 Testing effectiveness of advertising on customers. **(a)** $H_0 : p = 0.2$, $H_a : p > 0.2$
(b) $\alpha = P(\text{reject } H_0 \mid H_0) = \sum_{i=92}^{400} \binom{400}{i} (0.2)^i (0.8)^{400-i}$.

E10.126 Linear function of normal sample means. Consider the general case of k samples $X_{i1}, X_{i2}, \ldots, X_{in_i}$, $i = 1, 2, \ldots, k$. By the invariance property, $\hat{\theta} = \sum_{i=1}^{k} a_i \overline{X}_i$.
(a) $\operatorname{var} \hat{\theta} = \sum_{i=1}^{k} a_i^2 \operatorname{var} \overline{X}_i = \sum_{i=1}^{k} a_i^2 \sigma^2 / n_i$, and $\operatorname{se} \hat{\theta} = \sqrt{\operatorname{var} \hat{\theta}}$.
(b) A linear combination of independent normal r.v. is normal: $\hat{\theta} \sim \mathcal{N}(\theta, \operatorname{var} \hat{\theta})$. It follows that $(\hat{\theta} - \theta)/\operatorname{se} \hat{\theta} \sim \mathcal{N}(0, 1)$.
(c) By Theorem 7.3, $(n_i - 1) S_i^2 / \sigma^2 \sim \chi^2(n_i - 1)$. The sum of independent χ^2 is χ^2 with additive d.f., hence defining $n = \sum_{i=1}^{k} n_i$ and $S_p^2 = \sum_{i=1}^{k} (n_i - 1) S_i^2 / (n - k)$, it follows $(n-k) S_p^2 / \sigma^2 \sim \chi^2(n-k)$. Comparing

$$T = \frac{(\hat{\theta} - \theta)/\operatorname{se} \hat{\theta}}{\sqrt{[(n-k) S_p^2 / \sigma^2]/(n-k)}} = \frac{\hat{\theta} - \theta}{S_p \sqrt{\sum_{i=1}^{k} a_i^2 / n_i}}$$

to Definition 7.2, it follows $T \sim \mathcal{T}(n - k)$.
(d) T is a pivotal quantity, so a two-sided $(1 - \alpha)$ CI can be constructed using the methods of Section 8.5. The result: $\hat{\theta} \pm t_{\alpha/2} S_p \sqrt{\sum_{i=1}^{k} a_i^2 / n_i}$.
(e) Assuming $H_0 : \theta = \theta_0$, $T \sim \mathcal{T}(n - k)$, hence we can compute p-value $= 2P(T \geq |t|)$. Alternatively, use the CI in (d) and reject H_0 if $\theta_0 \notin$ CI.

11 Linear Models and Estimation by Least Squares

11.1 Introduction

Benefits of regression modeling. Suppose we measure the heights and weights of n randomly selected people. One way to model the heights, and the way we have learned up until now, is to consider independent and identical random variables $Y_1, Y_2, \ldots, Y_n$ where each $Y_i \sim \mathcal{N}(\mu, \sigma^2)$. We can then draw inference on, for example, the population parameter μ by examining the sample mean $\overline{Y} = (1/n)\sum_{i=1}^n Y_i$. $\overline{Y}$ is a good estimator because $\mathrm{E}\,\overline{Y} = (1/n)\sum_{i=1}^n \mathrm{E}\,Y_i = \mu$ (and it is consistent). In particular, we assumed $\mathrm{E}\,Y_i = \mu$ for each person. This last assumption turns out to be limiting because it assumes $\mathrm{E}\,Y_i$ is a constant for each person, independent of other variables.

A better model accounts for the relation between variables. For instance, height Y is related to weight X. For a given $X = x$, a simple linear regression models the relationship as

$$\mathrm{E}(Y|x) = \beta_0 + \beta_1 x,$$

where β_0 and β_1 are assumed constant parameters. For each weight x, the average height is equal to $\beta_0 + \beta_1 x$. The model can be recast as

$$Y|x = \beta_0 + \beta_1 x + \varepsilon$$

where the random error ε has mean zero and constant variance σ^2. Then the regression can be interpreted as follows: for each x, there exists a distribution of Y with mean $\mathrm{E}(Y|x) = \beta_0 + \beta_1 x$ and $\mathrm{var}(Y|x) = \mathrm{var}\,\varepsilon = \sigma^2$. This is illustrated by Figure 11.2. If we assume $\varepsilon \sim \mathcal{N}(0, \sigma^2)$, then $Y|x \sim \mathcal{N}(\beta_0 + \beta_1 x, \sigma^2)$. Estimation of β_0, β_1, and σ^2 allows us to make inference on $Y|x$. For instance, the least squares line

$$\hat{Y}|x = \hat{\beta}_0 + \hat{\beta}_1 x$$

is an estimator for $\mathrm{E}(Y|x)$ at every x.

Often times the constant variance of errors is unrealistic and depends on the size of x (e.g. tall adults have more variation in weight than short babies), but there are extensions of simple linear regression that address this. There are several other extensions, perhaps the most common using more variables in the model (multiple linear regression) and more complicated functions of those variables (e.g. splines), including interactions.

11.2 Linear Statistical Models

11.3 The Method of Least Squares

11.4 Properties of the Least-Squares Estimators: Simple Linear Regression

E11.20 MLE equivalence to least-squares estimators under normality. We want to estimate the parameters $\boldsymbol{\theta} = (\beta_0, \beta_1, \sigma^2)$. Let $\mu_i = \beta_0 + \beta_1 x_i$. The likelihood function

$$L(\boldsymbol{\theta} \mid \mathbf{Y}) = \prod_{i=1}^n \frac{1}{(2\pi\sigma^2)^{1/2}} e^{-(Y_i - \mu_i)^2/(2\sigma^2)} = \frac{1}{(2\pi\sigma^2)^{n/2}} \exp\left[-\frac{1}{2\sigma^2}\sum_{i=1}^n (Y_i - \mu_i)^2\right],$$

or

$$l(\boldsymbol{\theta} \mid \mathbf{Y}) \equiv \log L(\boldsymbol{\theta} \mid \mathbf{Y}) = -\frac{n}{2}\log(2\pi\sigma^2) - \frac{1}{2\sigma^2}\sum_{i=1}^n (Y_i - \mu_i)^2.$$

The MLES satisfy $\partial l/\partial\beta_j = 0$, $j = 0, 1$, and $\partial l/\partial\sigma^2 = 0$. $\partial l/\partial\beta_j = 0$ is equivalent to the least squares criterion.

E11.22 MLE equivalence to least-squares estimators under normality 2. $\partial l/\partial\sigma^2 = 0$ implies MLE $\widehat{\sigma^2} = \sum_{i=1}^n (Y_i - \hat\beta_0 - \hat\beta_1 x_i)^2/n = \sum_{i=1}^n (Y_i - \hat Y_i)^2/n = \mathrm{SSE}/n$. Note that $\widehat{\sigma^2}$ is a biased estimator of σ^2 whereas $S^2 = [n/(n-2)]\widehat{\sigma^2}$ is unbiased.

11.5 Inferences Concerning the Parameters $\boldsymbol{\beta_i}$

E11.28 Likelihood ratio test equivalence to *t*-test on regression coefficient. Let $\boldsymbol\theta = (\beta_0, \beta_1, \sigma^2)$. The full parameter space of interest $\Omega = \{\boldsymbol\theta : \beta_i \in \mathbb{R}, \sigma^2 > 0\}$, $i = 0, 1$. $H_0 : \beta_1 = 0$ restricts the space from 3- to 2-dimensional: $\Omega_0 = \{\boldsymbol\theta : \beta_1 = 0, \beta_0 \in \mathbb{R}, \sigma^2 > 0\}$. The likelihood ratio is $\lambda = \sup_{\boldsymbol\theta\in\Omega_0} L(\boldsymbol\theta)/\sup_{\boldsymbol\theta\in\Omega} L(\boldsymbol\theta)$ with rejection region $\lambda < k$ for some constant $k \in [0,1]$ set by the significance level. Defining $\mu_i = \beta_0 + \beta_1 x_i$, the likelihood function is

$$L(\boldsymbol\theta \mid \mathbf{Y}) = \prod_{i=1}^n \frac{1}{(2\pi\sigma^2)^{1/2}} e^{-(Y_i-\mu_i)^2/(2\sigma^2)} = \frac{1}{(2\pi\sigma^2)^{n/2}} \exp\left[-\frac{1}{2\sigma^2}\sum_{i=1}^n (Y_i - \mu_i)^2\right],$$

or

$$l(\boldsymbol\theta \mid \mathbf{Y}) \equiv \log L(\boldsymbol\theta \mid \mathbf{Y}) = -\frac{n}{2}\log(2\pi\sigma^2) - \frac{1}{2\sigma^2}\sum_{i=1}^n (Y_i - \mu_i)^2.$$

The MLEs in Ω were found in E11.20 and E11.22. They are $\hat\beta_1 = S_{xy}/S_{xx}$, $\hat\beta_0 = \overline{Y} - \hat\beta_1\overline{X}$, and $\widehat{\sigma^2} = \mathrm{SSE}/n$ where $\mathrm{SSE} = \sum_{i=1}^n (Y_i - \hat\beta_0 - \hat\beta_1 x_i)^2$. Maximizing in Ω_0,

$$l(\boldsymbol\theta \mid \mathbf{Y}, H_0) = -\frac{n}{2}\log(2\pi\sigma^2) - \frac{1}{2\sigma^2}\sum_{i=1}^n (Y_i - \beta_0)^2,$$

with MLEs $\hat\beta_{00} = \overline{Y}$ and $\widehat{\sigma_0^2} = \mathrm{SSE}_0/n$ where $\mathrm{SSE}_0 = \sum_{i=1}^n (Y_i - \hat\beta_{00})^2$. It follows that

$$\lambda = \left(\frac{\mathrm{SSE}}{\mathrm{SSE}_0}\right)^{n/2}.$$

The inverse ratio of SSE simplifies to

$$\frac{\mathrm{SSE}_0}{\mathrm{SSE}} = \frac{\mathrm{SSE} + \hat\beta_1 S_{xy}}{\mathrm{SSE}} = 1 + \frac{\hat\beta_1^2 S_{xx}}{\mathrm{SSE}}. = 1 + \frac{\hat\beta_1^2 S_{xx}/\sigma^2}{(n-2)S^2/\sigma^2}.$$

Under $H_0 : \beta_1 = 0$, the second term is distributed as T^2. Hence the rejection region $\lambda < k$ is equivalent to

$$T^2 > k^{-2/n} - 1 \qquad \Rightarrow \qquad |T| > \sqrt{k^{-2/n} - 1},$$

which is just a two-sided t-test with constant r.h.s. set by the significance level.

11.6 Inferences Concerning Linear Functions of the Model Parameters: Simple Linear Regression

11.7 Predicting a Particular Value of $\boldsymbol{Y}$ by Using Simple Linear Regression

Confidence and prediction interval widths. CIs are intervals for the constant parameter $\mathrm{E}(Y|x) = \beta_0 + \beta_1 x$ while PIs are intervals for the random variable $Y = \beta_0 + \beta_1 x + \varepsilon$. PIs are necessarily wider as random variation from ε must be accounted for. Both intervals are narrowest at $\overline{x}$ and wider at the tails due to variation associated with estimating $\hat\beta_1$. Intuitively, β_1 controls the slope of the line. Uncertainty in this slope does not effect the line at $\overline{x}$, but cause large deviations far from $\overline{x}$. On the contrary, estimation of $\hat\beta_0$ and $\widehat{\sigma^2}$ result in uniform uncertainty for all x.

11.8 Correlation

11.9 Some Practical Examples

Example 11.10. Regression output is below. Note the p-value (attained significance level) is two-sided. In either case, it is clear we reject $H_0 : \beta_1 = 0$.

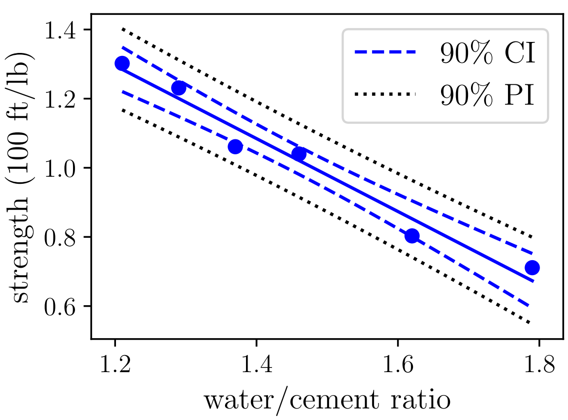
Dep. Variable:	y	**R-squared:**	0.968
Model:	OLS	**Adj. R-squared:**	0.960
Method:	Least Squares	**F-statistic:**	122.5
Date:	Wed, 26 Feb 2020	**Prob (F-statistic):**	0.000379
Time:	10:52:57	**Log-Likelihood:**	11.167
No. Observations:	6	**AIC:**	-18.33
Df Residuals:	4	**BIC:**	-18.75
Df Model:	1		

	coef	**std err**	**t**	**P> |t|**	**[0.025**	**0.975]**
Intercept	2.5606	0.140	18.283	0.000	2.172	2.949
x	-1.0544	0.095	-11.067	0.000	-1.319	-0.790

Omnibus:	nan	**Durbin-Watson:**	2.973
Prob(Omnibus):	nan	**Jarque-Bera (JB):**	0.912
Skew:	-0.605	**Prob(JB):**	0.634
Kurtosis:	1.522	**Cond. No.**	15.9

Warnings:
[1] Standard Errors assume that the covariance matrix of the errors is correctly specified.

The data, fit, confidence intervals (CIs) for the fit, and prediction intervals (PIs) for a future observation are plotted. For ratios outside the range of the data, the intervals will be much wider due to the quadratic variance associated with variation in β_1.



More on linearization. Consider the suggested nonlinear model

$$\mathrm{E}\,Y = \alpha_0 e^{\alpha_1 t}.$$

Strictly speaking, taking the logarithm of both sides,

$$\ln(\mathrm{E}\,Y) = \ln \alpha_0 + \alpha_1 t,$$

which is not the same as $\mathrm{E}(\ln Y) = \ln \alpha_0 + \alpha_1 t$ as suggested by the presented linearized model

$$\ln Y = \ln \alpha_0 + \alpha_1 t + \varepsilon.$$

Therefore linearization is not identical to the nonlinear regression solution. In particular, the presented additive error linearized model corresponds to multiplicative errors in the original nonlinear problem. To see this, consider nonlinear model

$$Y = \alpha_0 e^{\alpha_1 t} \epsilon$$

where the error ϵ has mean one. Indeed $\mathrm{E}\, Y = \alpha_0 e^{\alpha_1 t}$. Log-transforming yields the additive error linearized model

$$\ln Y = \ln \alpha_0 + \alpha_1 t + \ln \epsilon.$$

Transforming the data therefore changes the underlying error structure. Fits from solving the nonlinear problem and the linearized problem will be different.

***E11.65** **Proportional growth model.** Consider the linearized transformation

$$-\log(1 - \mathrm{E}\, Y) = \beta t.$$

This suggests linear model

$$-\log(1 - Y) = \beta t + \varepsilon,$$

which can be fit via least squares. Assuming $\varepsilon \sim \mathcal{N}(0, \sigma^2)$, $\hat{\beta}$ is normally distributed, hence a confidence interval for β can be constructed as $\hat{\beta} \pm t_{\alpha/2} \sigma_{\hat{\beta}}$.

11.10 Fitting the Linear Model by Using Matrices

11.11 Linear Functions of the Model Parameters: Multiple Linear Regression

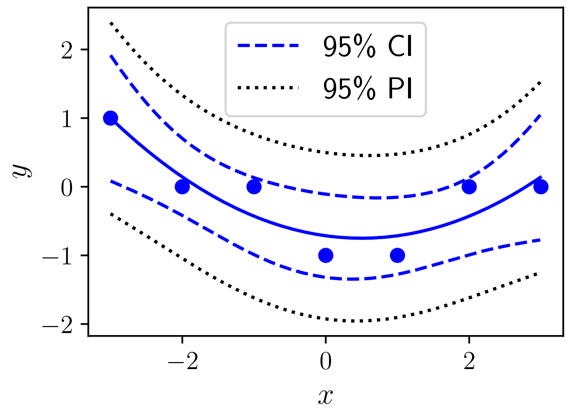
11.12 Inferences Concerning Linear Functions of the Model Parameters: Multiple Linear Regression

E11.73 **Hypothesis test for minimum of a quadratic model.** The minimum of $\mathrm{E}\, Y = \beta_0 + \beta_1 x + \beta_2 x^2$ occurs when $\beta_1 + 2\beta_2 x = 0$. The experimenter's test amounts to testing $H_0 : \beta_1 + 2\beta_2 = 0$ with $H_a : \beta_1 + 2\beta_2 \neq 0$. In matrix form, $H_0 : \mathbf{a}'\boldsymbol{\beta} = 0$ where $\mathbf{a}' = [0, 1, 2]$ and $\boldsymbol{\beta}' = [\beta_0, \beta_1, \beta_2]$. The test statistic is $t = (\mathbf{a}'\hat{\boldsymbol{\beta}} - 0)/(S\sqrt{\mathbf{a}'(\mathbf{X}'\mathbf{X})^{-1}\mathbf{a}})$, where $S^2 = \mathrm{SSE}/[n - (2+1)]$, $\mathrm{SSE} = \mathbf{y}'\mathbf{y} - \hat{\boldsymbol{\beta}}'\mathbf{X}'\mathbf{y}$, and $\mathbf{X}$ is the design matrix. Regression output and a fit plot are below.

Dep. Variable:	y	**R-squared:**	0.800
Model:	OLS	**Adj. R-squared:**	0.700
Method:	Least Squares	**F-statistic:**	8.000
Date:	Sat, 29 Feb 2020	**Prob (F-statistic):**	0.0400
Time:	15:44:10	**Log-Likelihood:**	-1.1632
No. Observations:	7	**AIC:**	8.326
Df Residuals:	4	**BIC:**	8.164
Df Model:	2		

	coef	**std err**	**t**	**P> |t|**	**[0.025**	**0.975]**
Intercept	-0.7143	0.218	-3.273	0.031	-1.320	-0.108
x	-0.1429	0.071	-2.000	0.116	-0.341	0.055
np.power(x, 2)	0.1429	0.041	3.464	0.026	0.028	0.257

Omnibus:	nan	**Durbin-Watson:**	2.964
Prob(Omnibus):	nan	**Jarque-Bera (JB):**	0.938
Skew:	0.643	**Prob(JB):**	0.626
Kurtosis:	1.750	**Cond. No.**	8.25



We find $t \approx 1.31$ with associated p-value ≈ 0.261. We fail to reject the experimenter's hypothesis. Indeed from the above plot, it is plausible the minimum is near $x = 1$. Had the experimenter claimed the minimum occurs at $x = -1$, the associated p-value ≈ 0.0171; this result is more unlikely providing some evidence to reject that claim.

11.13 Predicting a Particular Value of Y by Using Multiple Regression

11.14 A Test for $H_0 : \beta_{g+1} = \beta_{g+2} = \cdots = \beta_k = 0$

Example 11.19 using ANOVA. Composite tests are implemented in software using ANOVA tests comparing two nested models. (For details, see note in Section 13.3.) The resultant ANOVA table is below, yielding the same results as the example worked by hand in the text.

	df_resid	ssr	df_diff	ss_diff	F	Pr(>F)
0	8.0	326.622995	0.0	NaN	NaN	NaN
1	5.0	77.948010	3.0	248.674984	5.317112	0.051599

E11.87 Evaluation of R^2 metric.

(a) There are k numerator d.f. and $[n - (k+1)]$ denominator d.f. We find $F = 4.5$ with p-value $= 0.19$, so we cannot reject H_0.

(b) k is large relative to n. The factor $([n-(k+1)]/k$ is therefore small, making F (relatively) small and the p-value large despite the large R^2.

(c) $F \approx 2.35$ with p-value ≈ 0.0866, so we reject H_0.

(d) k is small relative to n. The factor $([n-(k+1)]/k$ is therefore large, making F (relatively) large and the p-value small despite the small R^2.

E11.88 Interpreting three different models. **(a)** False. The estimate is based on $[n-(k+1)] = 15$ d.f. **(b)** False. We perform an F-test between two nested models. **(c)** True. SSE is a nondecreasing function of k. **(d)** False, since we must also consider the ratio of d.f. A specific counter-example occurs when $\text{SSE}_\text{I} = \text{SSE}_\text{II}$ **(e)** True. **(f)** False. The models are not nested.

11.15 Summary and Concluding Remarks

Supplemental Exercises

---

12 Considerations in Designing Experiments

12.1 The Elements Affecting the Information in a Sample

12.2 Designing Experiments to Increase Accuracy

12.3 The Matched-Pairs Experiment

E12.17 Non-normal r.v.s whose difference is normal.

(a) $\mathrm{E}(Y_{ij}) = \mu_i + \mathrm{E}(U_j) + \mathrm{E}(\varepsilon_{ij}) = \mu_i$.

(b) The sum of a uniform and normal r.v. is not normally distributed.

(c) $\mathrm{cov}(Y_{1j}, Y_{2j}) = \mathrm{cov}(\mu_1 + U_j + \varepsilon_{1j}, \mu_2 + U_j + \varepsilon_{2j})$. Distributing, noting the covariance with a constant is zero, $\mathrm{cov}(Y_{1j}, Y_{2j}) = \mathrm{cov}(U_j, U_j) + \mathrm{cov}(U_j, \varepsilon_{2j}) + \mathrm{cov}(\varepsilon_{1j}, U_j) + \mathrm{cov}(\varepsilon_{1j}, \varepsilon_{2j})$. Assuming independence, the only nonzero contribution is $\mathrm{cov}(U_j, U_j) = \mathrm{var}(U_j) = \frac{1}{3}$.

(d) $D_j = Y_{1j} - Y_{2j} = (\mu_1 - \mu_2) + (\varepsilon_{1j} - \varepsilon_{2j})$, which is the sum of normals and constants, hence is normal. Furthermore, the ε_{ij} are all independent, hence all D_j are independent.

(e) We could have assumed any distribution for U_j since after differencing between matched pairs, the U_j cancel.

12.4 Some Elementary Experimental Designs

12.5 Summary

Supplemental Exercises

E12.35 Comparing two methods of ground temperature measurement. This is a matched pairs experiment. We conduct a paired t-test on the difference between measurements $\Delta = T_g - T_a$ with $H_0 : \Delta = 0$ and $H_a : \Delta \neq 0$. A $(1-\alpha)$ confidence interval is constructed as $\overline{\Delta} \pm t_{\alpha/2} S_\Delta/\sqrt{n}$, where S_Δ^2 is the sample variance of the differences. This procedure assumes normality of Δ.

(a) $p \approx 0.0124$, so the data present sufficient evidence at the $\alpha = 0.05$ level to claim a difference.

(b) $[-2.59, -0.566]$

(c) Requiring $t_{\alpha/2} S_\Delta/\sqrt{n} < 0.2$ and estimating S_Δ from the five measurements, $n \geq 47$.

Comment. Recall that $t_{\alpha/2}$ is a function of n.

---

13 The Analysis of Variance

13.1 Introduction

13.2 The Analysis of Variance Procedure

SST formula. $2n\overline{Y} = n\overline{Y}_1 + n\overline{Y}_2$, hence $\overline{Y} = (\overline{Y}_1 + \overline{Y}_2)/2$. It follows that $(\overline{Y}_1 - \overline{Y}) = (\overline{Y}_1 - \overline{Y}_2)/2$ and $(\overline{Y}_2 - \overline{Y}) = (\overline{Y}_2 - \overline{Y}_1)/2$. Thus $\text{SST} = n_1 \sum_{i=1}^{2} (\overline{Y}_i - \overline{Y})^2 = 2n_1(\overline{Y}_1 - \overline{Y}_2)^2/4 = n_1(\overline{Y}_1 - \overline{Y}_2)^2/2$.

E13.1 **Equivalence between ANOVA and t-test for two treatments.** Using ANOVA, $F \approx 2.93$. Using a 2-sample t-test, $t \approx -1.71$. Indeed $F = t^2$, so both yield p-value ≈ 0.109. Both tests assume the two samples are independent and generated from normal distributions with equal variance.

Comment. Welch's t-test for possibly unequal variance yields a similar p-value, equal to 3 significance figures.

13.3 Comparison of More Than Two Means: Analysis of Variance for a One-Way Layout

ANOVA equivalence to composite coefficient test in regression. In ANOVA, we partition the total sum of squares, TSS, into two sums of squares

$$\text{TSS} = \text{SST} + (\text{TSS} - \text{SST}) \equiv \text{SST} + \text{SSE}.$$

We then compute F-statistic

$$F = \frac{\text{SST}/(k-1)}{\text{SSE}/(n-k)}.$$

It is an analysis of *variance* because the test statistic is based on a ratio of sums of squares. A much larger variance among treatment groups compared to errors within groups is evidence for a difference in treatment effects. This analysis is analogous to testing to testing $\beta_{g+1} = \beta_{g+2} = \cdots = \beta_k = 0$ considered in the regression setting of Section 11.14. There we considered two nested models and partitioned the larger variance of the reduced model SSE_R into two sums of squares. Since the complete model sum $\text{SSE}_C \le \text{SSE}_R$, we partitioned as

$$\text{SSE}_R = \text{SSE}_C + (\text{SSE}_R - \text{SSE}_C)$$

and then computed

$$F = \frac{(\text{SSR}_R - \text{SSE}_C)/(k-g)}{\text{SSE}_C/(n-k-1)}.$$

Thus, composite tests of coefficients for regression models can be done via ANOVA.

13.4 An Analysis of Variance Table for a One-Way Layout

E13.8 **Professor salaries at different types of institutions.** This is a one-way layout. The dependent variable is salary (continuous), with one independent variable: institution type (categorical factor). The levels are public, private, and church institutions. The populations of interest are male assistant professors at each level. A one-way ANOVA yields p-value ≈ 0.00913, so there is sufficient evidence to conclude salary depends on institution for male assistant professors.

Comment. A two-way ANOVA could compare male/female assistant professors at each institution type. A three-way ANOVA could in addition compare type of professorship.

13.5 A Statistical Model for the One-Way Layout

13.6 Proof of Additivity of the Sums of Squares and E(MST) for a One-Way Layout

13.7 Estimation in the One-Way Layout

E13.24 Professor salaries at different types of institutions 2. Regression output modeling salary as a function of categorical institution type with an intercept is below. x_1, x_2, and x_3 correspond to public, private, and church, respectively.

Dep. Variable:	y	**R-squared:**	0.543
Model:	OLS	**Adj. R-squared:**	0.467
Method:	Least Squares	**F-statistic:**	7.123
Date:	Sat, 07 Mar 2020	**Prob (F-statistic):**	0.00913
Time:	15:39:52	**Log-Likelihood:**	-50.142
No. Observations:	15	**AIC:**	106.3
Df Residuals:	12	**BIC:**	108.4
Df Model:	2		

	coef	**std err**	**t**	**P>|t|**	**[0.025**	**0.975]**
const	45.2200	1.482	30.503	0.000	41.990	48.450
x1	8.8200	2.839	3.107	0.009	2.635	15.005
x2	25.5600	2.839	9.004	0.000	19.375	31.745
x3	10.8400	2.839	3.819	0.002	4.655	17.025

Omnibus:	2.407	**Durbin-Watson:**	1.953
Prob(Omnibus):	0.300	**Jarque-Bera (JB):**	1.612
Skew:	0.784	**Prob(JB):**	0.447
Kurtosis:	2.656	**Cond. No.**	8.23e+15

We estimate $\theta = \mu_2 - \mu_1 = \beta_2 - \beta_1$ with $\hat\theta = \hat\beta_2 - \hat\beta_1$ and $\operatorname{var}\hat\theta = \operatorname{var}\hat\beta_2 + \operatorname{var}\hat\beta_1 - 2\operatorname{cov}(\hat\beta_1, \hat\beta_2)$. The resultant interval is $[3.76, 29.7]$.

Comment. To estimate μ_i, $i = 1, 2, 3$, we use $\hat\theta = \hat\beta_0 + \hat\beta_i$ with $\operatorname{var}\hat\theta = \operatorname{var}\hat\beta_0 + \operatorname{var}\hat\beta_i + 2\operatorname{cov}(\hat\beta_0, \hat\beta_1)$. These estimators are equivalent to those used in the text. For justification on the equivalence between ANOVA and regression, see note in Section 13.13.

13.8 A Statistical Model for the Randomized Block Design

Viewing a randomized block design as a two-way ANOVA with no interactions. In a one-way ANOVA, we are interested in the response variable as a function of a factor with several levels. The statistical model for the i-th observation from level j is

$$Y_{ij} = \mu + \tau_j + \varepsilon_{ij}.$$

In a two-way ANOVA, we are interested in the response as a function of two factors, with possibly an interaction:

$$Y_{ijk} = \mu + \tau_j + \beta_k + \tau_j\beta_k + \varepsilon_{ijk}.$$

The model for the randomized block design is the two-way ANOVA without the interaction term. Qualitatively, the block can be viewed as another factor that affects the response variable.

13.9 The Analysis of Variance for a Randomized Block Design

E13.44 **Automobile insurance costs across insurance companies.** This is a randomized block design. We are primarily interested in policy costs for single men with basic coverage licensed for 6–8 years with no violations or accidents who drive 12600–15000 mpy. Despite the specific criteria, we still must account for location, hence the blocking. The resultant ANOVA table is below.

	df	sum_sq	mean_sq	F	PR(>F)
C(company)	4.0	731309.00	182827.250000	12.204172	0.000343
C(location)	3.0	1176270.15	392090.050000	26.172983	0.000015
Residual	12.0	179768.60	14980.716667	NaN	NaN

There is sufficient evidence to conclude that costs differ by company, adjusting for location, and by location, adjusting by company.

13.10 Estimation in the Randomized Block Design

13.11 Selecting the Sample Size

13.12 Simultaneous Confidence Intervals for More Than One Parameter

13.13 Analysis of Variance Using Linear Models

Equivalence between one-way ANOVA and linear regression with one categorical factor. Consider a factor with k levels. Given n observations, the one-way ANOVA model for the i-th observation from level j is

$$Y_{ij} = \mu + \tau_j + \varepsilon_{ij}, \qquad i = 1, 2, \ldots, n, \quad j = 1, 2, \ldots, k.$$

The typical linear regression model is expressed differently. The i-th observation can be expressed

$$Y_i = \beta_0 + \beta_1 x_{i1} + \cdots + \beta_k x_{ik} + \varepsilon_i$$

where x_{ij} are indicator variables

$$x_{ij} = \begin{cases} 1 & \text{if observation } i \text{ is in level } j \\ 0 & \text{otherwise} \end{cases}.$$

These two models are simply two ways of expressing the same model. To see why, suppose the i-th observation is in level j. Then only the $\beta_j x_{ij} = \beta_j$ term is nonzero and

$$Y_i = \beta_0 + \beta_j + \varepsilon_i,$$

which is precisely the ANOVA model with $\mu \leftrightarrow \beta_0, \tau_j \leftrightarrow \beta_j$, and $\varepsilon_{ij} \leftrightarrow \varepsilon_i$. It follows that testing all $\tau_1 = \tau_2 = \cdots = \tau_k = 0$ is the same as testing $\beta_1 = \beta_2 = \cdots = \beta_k = 0$.

Comments. (1) Typically in linear regression only $(k-1)$ dummy variables are needed, with one level being absorbed in the intercept β_0. This changes the interpretation of β_0 as a baseline level, rather than the mean of all observations μ, and makes the connection to ANOVA less clear. Formally, ANOVA constrains $\sum_{j=1}^{k} \tau_j = 0$, so only $(k-1)$ parameters need to be estimated.

(2) We considered one-way ANOVA, but the equivalence exists for n-way ANOVA (with interactions) to linear regression with n categorical variables (with interactions). The logic is the same; for every term in the ANOVA model, there is an equivalent term in the regression model selected by the dummy variables.

13.14 Summary

ANOVA is a broad term. In its simplest form, a one-way ANOVA is an extension of the t-test comparing means for more than 2 samples, each with different levels of a single factor (e.g. treatment type). More complicated designs (randomized block, n-way ANOVA) consider multiple factors and the effect on the mean of each factor, accounting for variation due to the remaining factors.

n-way ANOVA can be viewed a special case of a regression model with categorical independent variables, hence can be used for composite regression coefficient tests between nested models considered in Section 11.14. The two models can have continuous independent variables; either way, ANOVA assesses the importance of the variables in the full model not present in the restricted one.

Supplemental Exercises

E13.78 Drug study design. **(a)** This is a balanced two-way ANOVA. The people in the study are not blocks. **(b)** This is still a balanced two-way ANOVA.

E13.80 Study of car mileage varying model and gasoline brand.
A two-way ANOVA table (without interaction) is below.

	df	sum_sq	mean_sq	F	PR(>F)
C(model)	2.0	15.468889	7.734444	6.198575	0.059509
C(brand)	2.0	1.342222	0.671111	0.537845	0.621055
Residual	4.0	4.991111	1.247778	NaN	NaN

At the $\alpha = 0.05$ level, we cannot conclude either factor affects mileage.

E13.81 Study of car mileage varying model and gasoline brand 2.
A one-way ANOVA table is below.

	df	sum_sq	mean_sq	F	PR(>F)
C(model)	2.0	15.468889	7.734444	7.327368	0.024513
Residual	6.0	6.333333	1.055556	NaN	NaN

(a) Now at the $\alpha = 0.05$ level, we can conclude car brand affects mileage.
(b) The total sum of squares always remains the same. In practice, blocking will always explain some of the variance (even if it is unrelated), hence the SSE will be lower. More importantly, some d.f. are lost even if blocking explains no variance. If the blocking truly is unrelated, then it is more difficult to detect a difference in the main factor of interest.
(c) Gasoline brand may have an effect, even if it is small. Furthermore, testing the gasoline effect as in E13.80, then using its result to test model effect is similar to a multiple comparisons problem if the nominal α is unadjusted.

14 Analysis of Categorical Data

14.1 A Description of the Experiment

14.2 The Chi-Square Test

14.3 A Test of a Hypothesis Concerning Specified Cell Probabilities: A Goodness-of-Fit Test

E14.4 Heart attack dependence on day of the week. We perform a goodness of fit test testing $H_0 : p_1 = p_2 = \cdots = p_7 = \frac{1}{7}$, i.e. a heart attack is equally probable to occur on any day. The resultant is p-value ≈ 0.727. There is not enough evidence to suggest heart attacks occur on different days of the week.

E14.5 Heart attack dependence on day of the week 2.

(a) Our best estimate of p, the probability a heart attack victim suffers a heart attack on Monday, is $\hat{p} = Y/n$ where $Y \sim \text{binom}(n, p)$. For large n, $Y \approx \mathcal{N}(np, np(1-p))$, therefore statistic $Z = (\hat{p} - p)/\sqrt{p(1-p)/n} \approx \mathcal{N}(0, 1)$. Evidently we can conduct a z-test with $H_0 : p = \frac{1}{7}$ and $H_a : p > \frac{1}{7}$. The result is p-value ≈ 0.0667, so we fail to reject H_0.

(b) We should formulate our hypothesis before inspecting the data. We should not inspect the data, then test which hypothesis is most likely to be rejected, i.e. testing the day of the week with the most observed counts.

(c) Perhaps heart attacks are strongly related to one's prior day sleep and it is hypothesized people sleep less on Sunday before work, hence are more likely to have heart attacks.

14.4 Contingency Tables

E14.13 Study of party affiliation and JFK assassination opinion. A χ^2 test of independence yields $X^2 \approx 18.7$ which, under H_0 : party affiliation is independent from opinion, is approximately distributed as $\chi^2(\text{d.f.} = 4)$, hence $p \approx 8.96 \times 10^{-4}$. We therefore reject H_0. The p-value is approximate because the χ^2 test is an asymptotic (large sample) test, not an exact test.

14.5 $r \times c$ Tables with Fixed Row or Column Totals

Difference between χ^2 test of homogeneity and χ^2 test of independence. Procedurally, the tests are conducted identically. Expected cell counts in a $r \times c$ table are computed as $\widehat{\mathrm{E}(n_{ij})} = r_i c_j / n$ and compared to the observed counts using a χ^2 test with $(r-1)(c-1)$ d.f. in both cases. Despite the mathematical equivalence, the tests are designed differently and thus have a different interpretation.

Test of independence. A single sample on two variables is collected, e.g. voter preference and ward number. Given r preferences and c wards, we can model all $r \times c$ probabilities as a multinomial distribution. We are testing if the two variables are independent.

Test of homogeneity. Multiple independent samples for a single variable are collected, e.g. for each of multiple wards, we collect voter preference data. Each of the c total samples is modeled by a multinomial distribution with r probabilities. We are testing if all samples come from the same distribution.

E14.22 Equivalence of χ^2 homogeneity test and two-sided z-test for 2×2 table. (a) $X^2 \approx 18.5$. **(b)** Let p_1 (p_2) denote the portion that improved in the treated (untreated) group, estimated by $\hat{p}_i = Y_i/n_i$. In the large sample approximation, $\hat{p}_i \approx \mathcal{N}(p_i, p_i(1-p_i)/n_i)$. By independence of samples, under $H_0 : p_1 = p_2$ the quantity $T = (\hat{p}_1 - \hat{p}_2)/\sqrt{\hat{p}(1-\hat{p})(1/n_1 + 1/n_2)} \approx \mathcal{N}(0, 1)$ where the pooled estimator $\hat{p} = (Y_1 + Y_2)/(n_1 + n_2)$. (Pooling is justified since under H_0, we assume

the two samples come from the same distribution.) This yields $t \approx 4.30$ which is equal to $\sqrt{X^2}$.
(c) Since the tests are identical (X^2 has a χ^2 distribution with 1 d.f., T has a $\mathcal{N}(0,1)$ distribution whose square is also $\chi^2(1)$), both tests yield p-value $\approx 1.67 \times 10^{-5}$.

14.6 Other Applications

14.7 Summary and Concluding Remarks

Supplemental Exercises

E14.37 Binomial goodness of fit test. Let Y be the number of successes. We want to test the hypothesis $H_0 : Y \sim \text{binom}(n = 4, p)$ using a χ^2 goodness of fit test. Under H_0, the expected cell probabilities are

$$P(Y = y) = \binom{n}{y} p^y (1-p)^{n-y}, \qquad y = 0, 1, \ldots, n,$$

hence expected cell frequencies can be computed as $\widehat{\mathrm{E}(n_y)} = P(Y = y) \times N$, where $N = 100$ is the number of repeated experiments. Evidently we must estimate p. We proceed with maximum likelihood estimation. The likelihood function

$$L(p \mid \mathbf{y}) = \prod_{i=1}^{N} P(Y_i = y_i) = \prod_{i=1}^{N} \binom{n}{y_i} p^{y_i}(1-p)^{n-y_i} \propto p^z (1-p)^{Nn-z}$$

where $z = \sum_{i=1}^{N} y_i$. It is straightforward to show that the likelihood is maximized at $\hat{p} = z/(Nn)$. Indeed, this is the total number of successes divided by the total number of (independent) Bernoulli experiments, which we could have reasoned from the beginning. The expected cell counts are then $\widehat{\mathrm{E}(n_y)} = N\binom{n}{y}\hat{p}^y(1-\hat{p})^{n-y}$ for $y = 0, 1, \ldots, n$, and the statistic $X^2 \sim \chi^2(\text{d.f.} = n-2)$ as one parameter was estimated and the cell probabilities are constrained to sum to one. The resultant test yields $X^2 = 8.56$ with p-value ≈ 0.0358; we can reject H_0 at the $\alpha = 0.05$ level and conclude the data is not binomial distributed.

***E14.40 Genetic model for offspring proportion.**

(a) Let n_i denote the observed frequencies for the respective $k = 3$ classes, $i = 1, 2, 3$, which sum to n. The expected counts are $\widehat{\mathrm{E}(n_1)} = np^2$, $\widehat{\mathrm{E}(n_2)} = 2np(1-p)$, and $\widehat{\mathrm{E}(n_3)} = n(1-p)^2$. One parameter p must be estimated to conduct a χ^2 test for goodness of fit with $(k-2)$ d.f. The likelihood of the data

$$L(p) = (p^2)^{n_1}[2p(1-p)]^{n_2}[(1-p)^2]^{n_3}$$

is maximized at $\hat{p} = (n_1 + n_2/2)/n$. The resultant test statistic is $X^2 = 4$ with p-value ≈ 0.0455, and we reject the model at the $\alpha = 0.05$ significance level.

(b) The MLE in (a) was $\hat{p} = 0.5$, so again $X^2 = 4$. However, we do not need to estimate p if we test $H_0 : p = \frac{1}{2}$, so the test has $(k-1)$ d.f. with p-value ≈ 0.135. Now we cannot reject H_0 at the same significance level.

Comment. This problem highlights the importance of planning for an experiment. More flexible models, e.g. (a), may require more data. Furthermore different hypotheses can lead to different results, so they should be pre-specified and supported by existing evidence (before data is analyzed).

***E14.43 Linear model for probability of insect survival.** Denote n_i and m_i as the observed counts for survivors and non-survivors, respectively, with expected counts $\hat{n}_i$ and $\hat{m}_i$, $i = 1, 2, \ldots, k$ where k is the number of groups. Finally, denote N_i as the number of insects in each group. Modeling $p_i = 1 + \beta D_i$, the expected counts are $\hat{n}_i = p_i N_i$ and $\hat{m}_i = (1-p_i)N_i$. β and k probabilities must be estimated, hence the χ^2 test for goodness of fit has $2k-1-k=k-1$ d.f. We estimate β as follows: the likelihood of the data assuming our model to be true is

$$L(\beta)=\prod_{i=1}^{k} p_i^{n_i}(1-p_i)^{N_i-n_i}=\prod_{i=1}^{k}(1+\beta D_i)^{n_i}(-\beta D_i)^{N_i-n_i}$$

with log-likelihood

$$l(\beta)=\sum_{i=1}^{k}\left[n_i\log(1+\beta D_i)+(N_i-n_i)\log(-\beta D_i)\right].$$

The MLE occurs at

$$\frac{\partial l(\beta)}{\partial\beta}=\sum_{i=1}^{k}\left[\frac{n_iD_i}{1+\beta D_i}+\frac{(N_i-n_i)}{\beta}\right]=0.$$

This is a complicated function of β, but can be solved numerically via the bisection method since we know $0<p<1$ or $-\frac{1}{4}<\beta<0$. This yields $\hat{\beta}\approx-0.232$. We can then estimate p_i and hence cell counts. The χ^2 test yields p-value $\approx 3.86\times10^{-16}$. We strongly reject the linear model.

---

15 Nonparametric Statistics

15.1 Introduction

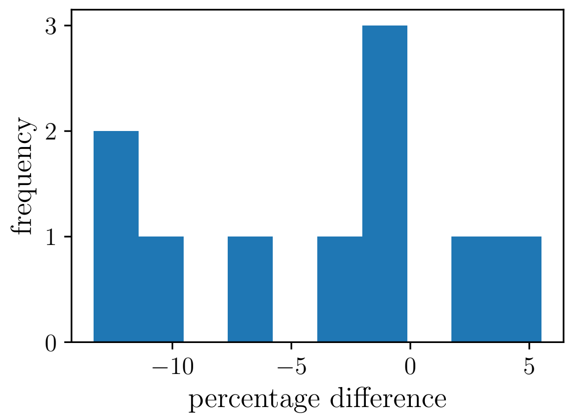
15.2 A General Two-Sample Shift Model

15.3 The Sign Test for a Matched-Pairs Experiment

E15.3 Clinical effectiveness of two drugs.

(a) Using a sign test, two-sided p-value ≈ 0.109. There is little evidence to indicate a difference.

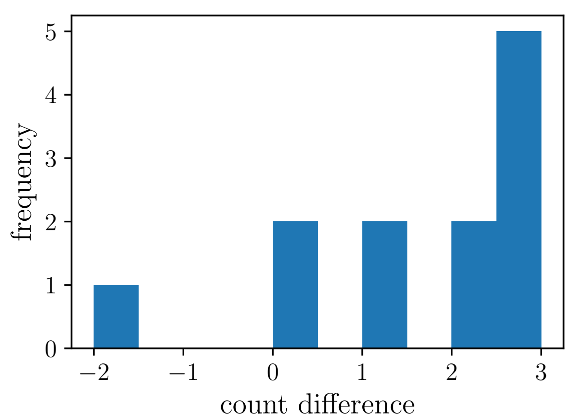
(b) The normality assumption is poor, evidenced by the histogram of the observed differences. (The t-test yields p-value ≈ 0.0685.)



E15.9 Effectiveness of an intervention for accidents.

(a) Using a sign test, two-sided p-value ≈ 0.0215. There is evidence to indicate a difference.

(b) The data are not continuous, which for small counts makes the the normality assumption poor. In addition, the histogram of observed differences suggest the data is skewed. (The t-test yields p-value ≈ 0.00611, suggesting much stronger evidence than attained using the sign test.)



15.4 The Wilcoxon Signed-Rank Test for a Matched-Pairs Experiment

E15.19 Nonparametric tests for the median.

(a) Under $H_0 : \xi = \xi_0$, we would expect $P(Y > \xi_0) = P(Y < \xi_0) = \frac{1}{2}$. Therefore we should consider the differences $D_i = Y_i - \xi_0$, $i = 1, 2, \ldots, n$, satisfying $P(D_i > 0) = P(D_i < 0) = \frac{1}{2}$.

The sign tests counts the number of positive differences $D_i > 0$ which is binomially distributed. Specifically, the test statistic is the number of positive differences $X \sim \text{binom}(n, \frac{1}{2})$. Extreme values of X indicate rejection of H_0.

(b) Similar to (a), we again consider differences $D_i = Y_i - \xi_0$. Under $H_0 : \xi = \xi_0$, we would expect a similar distribution of ranks of absolute differences $|D_i|$ for positive and negative differences. For a two-sided test, we consider the minimum of the sum of ranks for positive and negative differences, $T = \min(T_+, T_-)$. Small values of T indicate rejection of H_0.

15.5 Using Ranks for Comparing Two Population Distributions: Independent Random Samples

15.6 The Mann–Whitney *U* Test: Independent Random Samples

15.7 The Kruskal–Wallis Test for the One-Way Layout

15.8 The Friedman Test for Randomized Block Designs

Generalizing Mann–Whitney, Kruskal–Wallis, and Friedman test. The Mann–Whitney U test is a nonparametric alternative to a two independent sample t-test. The Kruskal–Wallis generalizes the Mann–Whitney test to allow for n groups, i.e. the nonparametric alternative to a one-way ANOVA. The Friedman test further generalizes to allow for blocking, i.e. the nonparametric alternative to a two-way ANOVA with no interactions. There is no further generalization of a Kruskal–Wallis like test to an n-way ANOVA (with or without interactions). However, the Kruskal–Wallis (thus also Mann–Whitney) test is a special case of the proportional odds ordinal logistic model: see Frank Harrell's answer here. Such a model naturally accommodates multiple factors with interactions and adjustment for other variables of interest.

15.9 The Runs Test: A Test for Randomness

15.10 Rank Correlation Coefficient

15.11 Some General Comments on Nonparametric Statistical Tests

Supplemental Exercises

---

16 Introduction to Bayesian Methods for Inference

16.1 Introduction

16.2 Bayesian Priors, Posteriors, and Estimators

16.3 Bayesian Credible Intervals

16.4 Bayesian Tests of Hypotheses

16.5 Summary and Additional Comments

Comparing frequentist and Bayesian estimation methods for a Bernoulli sample. Suppose we are interested in estimating the success probability p from a Bernoulli random sample $Y_1, Y_2, \ldots, Y_n$, $Y_i \sim \text{Bernoulli}(p)$, given observed data $y_1, y_2, \ldots, y_n$. We will proceed with the frequentist (Chapter 8–9) and Bayesian (Chapter 16) estimation methods. Both approaches seek to estimate the true unknown parameter p. The frequentist approach relies solely on the data, while the Bayesian approach additionally incorporates prior beliefs modeling the distribution of p.

Frequentist. We treat the parameter p as an unknown constant which we wish to estimate. A good estimator $\hat{p}$ can be obtained via maximum likelihood estimation (MLE). The likelihood function

$$L(p \mid \mathbf{y}) = \prod_{i=1}^{n} P(Y_i = y_i) = p^u (1-p)^{n-u}$$

where $u = \sum_{i=1}^{n} y_i$. It is straightforward to show the likelihood is maximized at $u/n = \overline{y}$, so the MLE is $\hat{p} = \overline{Y}$. It is an unbiased estimator with $\operatorname{var} \hat{p} = p(1-p)/n$.

The MLE is a random variable–a different sample would yield a different estimate. To quantify the sampling uncertainty, we construct a γ confidence interval such that given N samples, approximately γN of these intervals would contain the true p. Exact intervals can be computed for the binomial case (Clopper–Pearson intervals), but for simplicity we consider large sample intervals. Invoking the CLT, for large n, $\overline{Y} \approx \mathcal{N}(p, p(1-p)/n) \approx \mathcal{N}(p, \overline{Y}(1-\overline{Y})/n)$. It follows that $Z = (\overline{Y} - p)/\sqrt{\overline{Y}(1-\overline{Y})/n}$ is approximately pivotal, yielding γ confidence interval

$$\left[\overline{Y} - z_{\alpha/2}\sqrt{\overline{Y}(1-\overline{Y})/n}, \overline{Y} + z_{\alpha/2}\sqrt{\overline{Y}(1-\overline{Y})/n}\right].$$

This interval has random variable endpoints. For a specific sample $y_1, y_2, \ldots, y_n$, the confidence interval now has fixed endpoints $\overline{y} \pm z_{\alpha/2}\sqrt{\overline{y}(1-\overline{y})/n}$ and probabilistic statements about this specific interval cannot be made. From a strictly frequentist view, p is either contained or not contained. Even so, a narrow interval implies the sampling uncertainty is small.

Bayesian. We no longer consider p as a constant, but rather model the unknown parameter as a probability distribution. We have a prior belief of what value we think p takes and quantify it as a prior probability distribution. For instance, we may suspect p is near $\frac{1}{2}$ and assume our prior is normally distributed with mean at $\frac{1}{2}$. Or, following Example 16.1, we specify the prior $p \sim \text{beta}(\alpha, \beta)$. The posterior distribution is then computed using the prior, the data, and Bayes' theorem. For the given beta prior, the posterior distribution is $g(p|\mathbf{y}, \alpha, \beta) \sim \text{beta}(\alpha+u, \beta-u+n)$. The posterior distribution $g(\cdot)$ is key to inference on p. For instance, a sensible estimator of p is

$$\hat{p} = \mathrm{E}(p|\mathbf{Y}, \alpha, \beta) = \int_0^1 dp\, p \times g(p|\mathbf{Y}, \alpha, \beta) = \frac{\alpha + n\overline{Y}}{\alpha + \beta + n}.$$

$\hat{p}$ is again a random variable. It is straightforward to show that $\hat{p}$ is biased:

$$\mathrm{E}\,\hat{p} = \frac{\alpha + np}{\alpha + \beta + n},$$

with

$$\operatorname{var}\hat{p} = \frac{np(1-p)}{(\alpha+\beta+n)^2}.$$

An interval estimate is also easily computed from the posterior distribution. In fact since we are modeling the distribution of p, we can construct a probabilistic γ credible interval $[a, b]$ by integration such that

$$\gamma = P(a < p < b) = \int_a^b dp\, g(p|\mathbf{Y}, \alpha, \beta).$$

This results in an interval with the following interpretation: the probability $p \in [a, b]$ is γ. This holds true even for a credible interval computed from a specific sample $y_1, y_2, \ldots, y_n$.

Comparison of estimators. Consider the mean square error of the frequentist and Bayesian point estimates, $\hat{p}_F = \overline{Y}$ and $\hat{p}_B = (\alpha + n\overline{Y})/(\alpha+\beta+n)$. Using $\mathrm{MSE}(\hat{\theta}) = \operatorname{var}\hat{\theta} + (\operatorname{bias}\hat{\theta})^2$,

$$\mathrm{MSE}(\hat{p}_F) = \frac{p(1-p)}{n}, \qquad \mathrm{MSE}(\hat{p}_B) = \frac{np\,(1-p) + \left[p\,(\alpha+\beta+n) - \alpha - np\right]^2}{(\alpha+\beta+n)^2}.$$

Even though the Bayesian estimator is biased, in some situations it has lower MSE. For instance, in the case $\alpha = \beta = 1$,

$$\mathrm{MSE}(\hat{p}_B) < \mathrm{MSE}(\hat{p}_F) \qquad \text{when} \qquad \frac{1}{2}\left[1 - \sqrt{1 - n/(2n+1)}\right] < p < \frac{1}{2}\left[1 + \sqrt{1 - n/(2n+1)}\right].$$

(For details of the calculation, see E8.17.)