(b) The paired analysis would be more powerful because it would eliminate sources of variability (e.g., time, depth) unrelated to the source of variation in question: differences between the protected and nonprotected areas.

CHAPTER 9

Categorical Data: One-Sample Distributions

9.1.1 The possible values of $\tilde{P}$ are 2/7, 3/7, 4/7, and 5/7. These correspond to getting 0 persons with lung cancer, 1 with lung cancer, 2 with lung cancer, and all 3 with lung cancer.

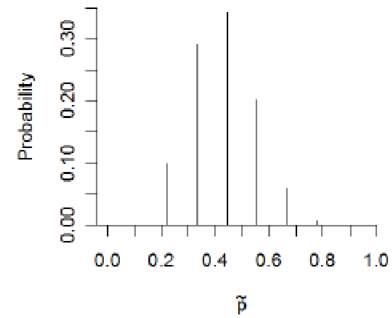
• **9.1.2 (a)** Pr{ $\tilde{P}$ = 2/7} = Pr{no mutants} = Pr{all are non-mutants} = $(1 - 0.37)^3 = 0.250$.

(b) Pr{ $\tilde{P}$ = 3/7} = Pr{1 mutant} = $_3C_1 p^1 (1 - p)^2$, where p = 0.37. This is = $(3)(0.37^1)(0.63^2) = 0.441$.

The smallest possible number of mutants is zero, for which $\tilde{P}$ = 2/7. Thus, it is not possible that $\tilde{P}$ will equal zero.

9.1.3 (a) (i) 0.10; (ii) 0.29; (iii) 0.34; (iv) 0.20; (v) 0.06; (vi) 0.01

(b)



• **9.1.4** We are concerned with the sampling distribution of $\tilde{P}$, which is governed by a binomial distribution. Letting "success" = "responder," we have p = 0.2 and 1 - p = 0.8. The number of trials is n = 15.

(a) The event $\tilde{P}$ = 5/19 occurs if there are 3 successes in the 15 trials (because (3+2)/(15+4) = 5/19). Thus, to find the probability that $\tilde{P}$ = 5/19, we can use the binomial formula $_nC_j p^j (1 - p)^{n-j}$ with j = 3, so n - j = 12:

Pr{ $\tilde{P}$ = 5/19 } = $_{15}C_3 p^3 (1 - p)^{12} = (455)(0.2^3)(0.8^{12}) = 0.2501$.

(b) The event $\tilde{P}$ = 2/19 occurs if there are 0 successes in the 15 trials. Thus, to find the probability that $\tilde{P}$ = 2/19, we can use the binomial formula with j = 0, so n - j = 15:

Pr{ $\tilde{P}$ = 2/19} = $_{15}C_0 p^0 (1 - p)^{15} = (1)(1)(0.8^{15}) = 0.0352$.

• **9.1.5 (a)** Letting "success" = "infected," we have p = 0.25 and 1 - p = 0.75. The number of trials is n = 4. We then use the binomial formula $_nC_j p^j (1 - p)^{n-j}$ with n = 4 and p = 0.25. The values of $\tilde{P}$ correspond to numbers of successes and failures as follows:

$\tilde{P}$	Number of successes (j)	Number of failures (n - j)
2/8	0	4
3/8	1	3
4/8	2	2
5/8	3	1
6/8	4	0

Thus, we find

(i) $\Pr\{\tilde{P} = 2/8\} = {}_4C_0 p^0(1-p)^4 = (1)(1)(0.75^4) = 0.3164$

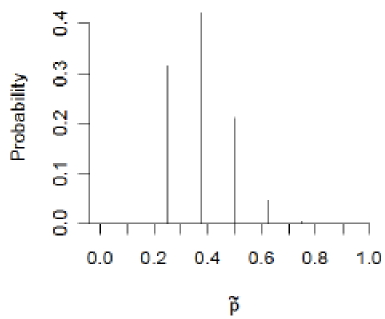
(ii) $\Pr\{\tilde{P} = 3/8\} = {}_4C_1 p^1(1-p)^3 = (4)(0.25)(0.75^3) = 0.4219$

(iii) $\Pr\{\tilde{P} = 4/8\} = {}_4C_2 p^2(1-p)^2 = (6)(0.25^2)(0.75^2) = 0.2109$

(iv) $\Pr\{\tilde{P} = 5/8\} = {}_4C_3 p^3(1-p)^1 = (4)(0.25^3)(0.75) = 0.0469$

(v) $\Pr\{\tilde{P} = 6/8\} = {}_4C_4 p^4(1-p)^0 = (1)(0.25^4)(1) = 0.0039$

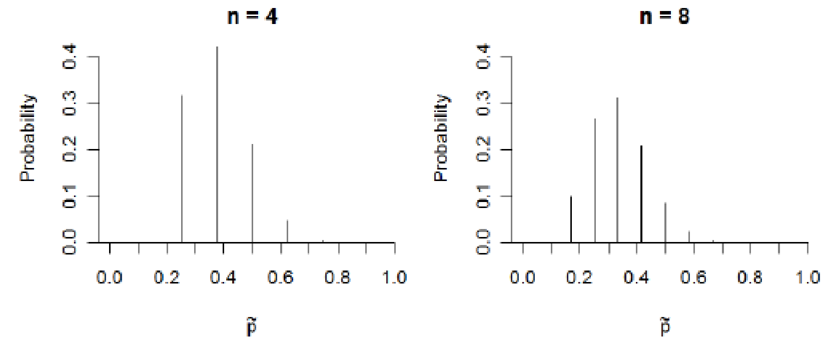
(b) The distribution is displayed in the following histogram:



9.1.6 (a)

$\tilde{P}$	Probability		
2/12	0.75^8	=	0.1001
3/12	$(8)(0.25)(0.75^7)$	=	0.2670
4/12	$(28)(0.25^2)(0.75^6)$	=	0.3115
5/12	$(56)(0.25^3)(0.75^5)$	=	0.2076
6/12	$(70)(0.25^4)(0.75^4)$	=	0.0865
7/12	$(56)(0.25^5)(0.75^3)$	=	0.0231
8/12	$(28)(0.25^6)(0.75^2)$	=	0.0039
9/12	$(8)(0.25^7)(0.75)$	=	0.0004
10/12	0.25^8	=	0.0000

(b)



The distribution for n = 8 is somewhat narrower than the distribution for n = 4.

9.1.7 (a) $(20)(0.6^3)(0.4^3) = 0.2765$

(b) $(15)(0.6^4)(0.4^2) = 0.3110$

(c) $(6)(0.6^5)(0.4^1) = 0.1866$

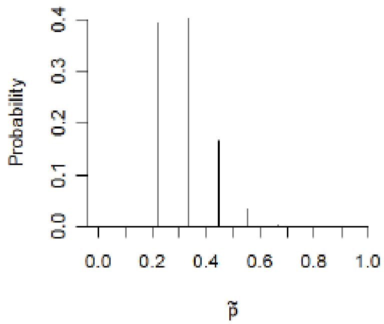
(d) $0.2765 + 0.3110 + 0.1866 = 0.7741$

(e) 0.7741 (from part (d))

9.1.8 (a)

Number contaminated	$\tilde{P}$	Probability		
0	2/9	0.83^5	=	0.3939
1	3/9	$(5)(0.17)(0.83^4)$	=	0.4034
2	4/9	$(10)(0.17^2)(0.83^3)$	=	0.1652
3	5/9	$(10)(0.17^3)(0.83^2)$	=	0.0338
4	6/9	$(5)(0.17^4)(0.83)$	=	0.0035
5	7/9	0.17^5	=	0.0001

(b) The sampling distribution for n = 5 shown below is more spread out and less symmetric than the sampling distributions for larger samples shown in Figure 9.1.3.



• **9.1.9** Because p = 0.25, the event E occurs if $\tilde{P}$ is within ± 0.05 of 0.25; this happens if there are 3, 4, or 5 successes, as follows:

Number of successes (j)	$\tilde{p}$
3	5/24 = 0.208
4	6/24 = 0.250
5	7/24 = 0.292

We can calculate the probabilities of these outcomes using the binomial formula with n = 20 and p = 0.25:

$\Pr\{\tilde{P} = 0.208\} = {}_{20}C_3 p^3(1 - p)^{17} = (1{,}140)(0.25^3)(0.75^{17}) = 0.1339$
$\Pr\{\tilde{P} = 0.250\} = {}_{20}C_4 p^4(1 - p)^{16} = (4{,}845)(0.25^4)(0.75^{16}) = 0.1897$
$\Pr\{\tilde{P} = 0.292\} = {}_{20}C_5 p^5(1 - p)^{15} = (15{,}504)(0.25^5)(0.75^{15}) = 0.2023$

Finally, we calculate Pr{E} by adding these results:
Pr{E} = 0.1339 + 0.1897 + 0.2023 = 0.5259.

9.1.10 The sample percentage, $\hat{P}$, of students who smoke varies from one sample to the next. The sampling distribution of the sample percentage is the distribution of $\hat{P}$ -- the proportion of smokers in a sample -- across repeated samples. That is, the sampling distribution of the sample percentage is the distribution of sample percentages of smokers in samples of size 10.

9.2.1 (a) y = (50)(0.84) = 42; $\tilde{p}$ = (42+2)/(50+4) = 0.815.

$$SE = \sqrt{\frac{0.815(1-0.815)}{50+4}} = 0.053.$$

(b) y = (200)(0.84) = 168; $\tilde{p}$ = (168+2)/(200+4) = 0.833.

$$SE = \sqrt{\frac{0.833(1-0.833)}{200+4}} = 0.026.$$

• **9.2.2** (a) The number of mutants in the sample is y = (100)(0.20) = 20. Thus, $\tilde{p}$ = (20+2)/(100+4) = 0.212.
The standard error is

$$SE = \sqrt{\frac{\tilde{p}(1-\tilde{p})}{n+4}} = \sqrt{\frac{0.212(1-0.212)}{100+4}} = 0.040.$$

(b) The number of mutants in the sample is y = (400)(0.20) = 80. Thus, $\tilde{p}$ = (80+2)/(400+4) = 0.203.
The standard error is

$$SE = \sqrt{\frac{\tilde{p}(1-\tilde{p})}{n+4}} = \sqrt{\frac{0.203(1-0.203)}{400+4}} = 0.020.$$

• **9.2.3** (a) The 95% confidence interval is
$\tilde{p} \pm 1.96\,SE_{\tilde{p}}$

0.212 ± (1.96)(0.040)

0.212 ± 0.078

(0.134,0.290) or $0.134 < p < 0.290$.

(b) The 95% confidence interval is
$\tilde{p} \pm 1.96\,SE_{\tilde{p}}$

0.203 ± (1.96)(0.020)

0.203 ± 0.039

(0.164,0.242) or $0.164 < p < 0.242$.

9.2.4 $\tilde{p}$ = (28 + 2)/(580 + 4) = 0.051; $SE = \sqrt{\frac{0.51(1-0.51)}{584}} = 0.009$.
The 95% confidence interval is 0.051 ± (1.96)(0.009) or (0.033,0.069) or $0.033 < p < 0.069$.

• **9.2.5** (a) $\tilde{p}$ = (69+2)/(339+4) = 0.207.
The standard error is

$$SE = \sqrt{\frac{\tilde{p}(1-\tilde{p})}{n+4}} = \sqrt{\frac{0.207(1-0.207)}{339+4}} = 0.022.$$

The 95% confidence interval is
$\tilde{p} \pm 1.96\,SE_{\tilde{p}}$

0.207 ± (1.96)(0.022)

0.207 ± 0.043

(0.164,0.250) or $0.164 < p < 0.250$.

(b) We are 95% confident that the probability of adverse reaction in infants who receive their first injection of vaccine is between 0.164 and 0.250.

(c) Yes; the confidence interval does not include 0.25 so we have evidence that p is less than 0.25.

(d) 97.5%: A one-sided 97.5% confident interval would have an upper limit of 0.25.

9.2.6 $\tilde{p} = (14+2)/(71+4) = 0.213$; SE = $\sqrt{\frac{0.213(1-0.213)}{71+4}} = 0.047$.
95% confidence interval: 0.213 ± (1.96)(0.047) or (0.121,0.305) or $0.121 < p < 0.305$.

• **9.2.7** The required n must satisfy the inequality

$$\sqrt{\frac{(\text{Guessed } \tilde{p})(1 - \text{Guessed } \tilde{p})}{n+4}} \le \text{Desired SE}$$

or

$$\sqrt{\frac{0.6(0.4)}{n+4}} \le 0.04.$$

It follows that $\frac{\sqrt{0.6(0.4)}}{0.04} \le \sqrt{n+4}$

or $\frac{(0.6)(0.4)}{0.04^2} \le n+4$ or $150 \le n+4$, so $n \ge 146$.

9.2.8 $\sqrt{\tilde{p}(1-\tilde{p})}$ is largest when $\tilde{p} = 0.5$ as it is here. The required n must satisfy the inequality

$$\sqrt{\frac{(0.5)(0.5)}{n+4}} \le .04.$$

It follows that $\frac{\sqrt{(0.5)(0.5)}}{0.04} \le \sqrt{n+4}$

or $\frac{(0.5)(0.5)}{0.04^2} \le n+4$ or $156.25 \le n+4$, so $n \ge 152.25$; thus, $n \ge 153$.

9.2.9 Desired SE = 0.01; guessed $\tilde{p} = 0.7$. The required n must satisfy the inequality

$$\sqrt{\frac{0.7(0.3)}{n+4}} \le 0.01.$$

It follows that $\frac{\sqrt{0.7(0.3)}}{0.01} \le \sqrt{n+4}$

or $\frac{0.7(0.3)}{0.01^2} \le n+4$ or $2100 \le n+4$, so $n \ge 2096$.

9.2.10 $\sqrt{\tilde{p}(1-\tilde{p})}$ is largest when $\tilde{p} = 0.5$. Thus, the required n must satisfy the inequality

$$\sqrt{\frac{(0.5)(0.5)}{n+4}} \le 0.01.$$

It follows that $\frac{\sqrt{(0.5)(0.5)}}{0.01} \le \sqrt{n+4}$

or $\frac{(0.5)(0.5)}{0.01^2} \le n+4$ or $2500 \le n+4$, so $n \ge 2496$.

9.2.11 The required sample size n would be approximately reduced by a factor of 4 because n is inversely proportional to the square of the SE requirement (when n >> 4).

9.2.12 It is necessary to use $\tilde{p} = 0.5$ because the proportion of progeny that are resistant is unknown in advance. Thus, the required n must satisfy the inequality

$$\sqrt{\frac{(0.5)(0.5)}{n+4}} \le 0.05.$$

It follows that $n \ge 96$.

9.2.13 Yes. The two mechanisms give p = 0.5 and p = 0.75. The agronomist can be sure that the confidence interval will not contain both 0.5 and 0.75 because the confidence interval will be no larger than $\tilde{p}$ ± (1.96)(0.05), which is $\tilde{p}$ ± 0.098. The width of the interval is no more than .196, so the interval cannot cover both 0.5 and 0.75.

9.2.14 $\tilde{p} = \frac{60+2}{264+4} = 0.231$. The SE is $\sqrt{\frac{0.231(0.769)}{268}} = 0.0257$. The 95% confidence interval is 0.231 ± 1.96*0.0257 or 0.231 ± 0.050 or (0.181,0.281).

9.2.15 **(a)** From 9.2.13 $\tilde{p} = 62/268 = 0.231$. Then $\sqrt{\frac{0.231(0.769)}{n+4}} \le 0.04$ implies that n is at least 108.

(b) $\sqrt{\frac{0.5(0.5)}{n+4}} \le 0.04$ implies that n is at least 153.

9.2.16 **(a)** The confidence interval estimates the proportion of all trees in the arboretum that are native species to the area.

(b) We are 95% confident that the proportion of all trees in the arboretum that are native species to the area in the population is between 71% and 89%.

9.2.17 **(a)** Yes. The interval is entirely below 0.50. We are 95% confident that hands are washed between 38.1 and 48.1% of the time.

(b) Since there were 375 hand washing opportunities in this study and fewer than 40 staff likely observed, the researchers necessarily observed some subjects multiple times. The observations, therefore, are not independent, thus casting doubt about the validity of the interval as an estimate of the proportion of workers who wash their hands.

9.3.1 $\hat{p}$=0.037. Thus y ≈ (848)(0.037) = 31.376, so y must be 31.

Thus, $\tilde{p} = \frac{31+0.5(1.645^2)}{848+1.645^2} = 0.0380$ and SE = $\sqrt{\frac{0.038(1-0.038)}{848+1.645^2}} = 0.0066$.

90% confidence interval: 0.0380 ± (1.645)(0.0066) or (0.0271,0.0489) or $0.0271 < p < 0.0489$.

9.3.2 **(a)** y = (959)(0.157) = 150.56, so y must be 151.

Thus, $\tilde{p} = \frac{151+0.5(1.645^2)}{959+1.645^2} = .158$ and SE = $\sqrt{\frac{0.158(1-0.158)}{959+1.645^2}} = 0.012$.

90% confidence interval: 0.158 ± (1.645)(0.012) or (0.138,0.178) or $0.138 < p < 0.178$.

(b) The confidence interval from part (a) is a confidence interval for the probability of interference with the pacemaker for that type of cellular telephone.

9.3.3 n = 180; y = 23; $\tilde{p} = \frac{y+0.5(1.645^2)}{n+1.645^2} = 0.133$.

SE = $\sqrt{\frac{\tilde{p}(1-\tilde{p})}{n+1.645^2}} = 0.025$.

The 90% confidence interval is 0.133 ± (2.576)(0.025) or 0.133 ± 0.064 or (0.069,0.197).

• **9.3.4** $\tilde{p} = \frac{40+0.5(1.645^2)}{53+1.645^2} = 0.7423$ and SE = $\sqrt{\frac{0.7423(1-0.7423)}{53+1.645^2}} = 0.0586$.

90% confidence interval: 0.7423 ± (1.645)(0.0586) or (0.646,0.839) or $0.646 < p < 0.839$.

Note: In hypothesis testing problems involving the χ^2 statistic, expected frequencies are shown in parentheses.