Math 218, Mathematical Statistics
D Joyce, Spring 2016

From chapter 9, page 330, exercises 1–3, 6.

1. A business journal publisher plans to survey a sample of the subscribers to estimate the proportion p with annual household incomes over \$100,000.

a. How many subscribers must be surveyed to obtain a 99% CI for p with a margin of error no greater than 0.05? Assume that no prior estimate of p is available.

The sample size for a CI with $\alpha = 0.01$ and $E \leq 0.05$ when no prior estimate is assumed is

$$\begin{aligned} n &= \left(\frac{z_{\alpha/2}}{E}\right)^2 \frac{1}{4} \\ &= \left(\frac{z_{.005}}{0.05}\right)^2 \frac{1}{4} \\ &= \left(\frac{2.575}{0.05}\right)^2 \frac{1}{4} = 663.06 \end{aligned}$$

So, $n = 664$ is sufficient.

b. The marketing department thinks that $p = 0.30$ would be a reasonable guess. What is the corresponding sample size?

Assuming marketing's estimate is accurate, then, with $p^* = 0.30$ and $q^* = 0.70$ we have

$$\begin{aligned} n &= \left(\frac{z_{\alpha/2}}{E}\right)^2 p^* q^* \\ &= \left(\frac{z_{.005}}{0.05}\right)^2 0.30 \cdot 0.70 \\ &= \left(\frac{2.575}{0.05}\right)^2 0.21 = 556.9 \end{aligned}$$

So, $n = 557$ is sufficient.

But if marketing is off a bit, say $\hat{p}$ comes out to be 0.35, then $n = 557$ would not justify the conclusions. We would need $n = \left(\frac{2.575}{0.05}\right)^2 0.35 \cdot 0.65 \approx 603$. Thus, we need to be conservative about what you think p might be, and overestimate to be safe.

c. Refer to the sample size obtained in part b. If a 40% nonresponse rate is anticipated, how many surveys need to be mailed? How may such a high nonresponse rate cause bias in the estimate?

Suppose we need $n = 557$ responses, but 40% of those asked don't respond. Let N be the number we need to ask. Then $0.40N = n = 557$, so $N \approx 1393$.

The interval estimate is only valid when we have a random sample of size 557. We can do everything possible to make sure it's random, but when only 40% respond (which is fairly high), the sample is no longer random. Those who are predisposed to answer questionnaires are are likely to respond to the questions on the questionnaire differently than those who don't respond.

2. While imprisoned by the Germans during World War II, the English mathematician John Kerrich tossed a coin 10,000 times and obtained 5067 heads. Let p bet the probability of a head on a single toss. We wish to check if the data are consistent with the hypothesis that the coin was fair.

a. Set up the hypotheses. Why should the alternative by two-sided?

Naturally, H_0 should be $p = \frac{1}{2}$ and H_1 should be $p \neq \frac{1}{2}$. The test is two-sided because we will reject the hypothesis of a fair coin either if very many or very few heads come up.

b. Calculate the P-value. Can you reject H_0 at the .05 level?

The P-value is

$$\begin{aligned} P\text{-value} &= P(\text{Test rejects } H_0 | p = p_0) \\ &= 2(1 - \Phi(|z|) \\ &= 2\left(1 - \Phi\left(\left|\frac{\hat{p} - p}{\sqrt{pq/n}}\right|\right)\right) \end{aligned}$$

Since

$$\frac{\hat{p} - p}{\sqrt{pq/n}} = \frac{\frac{5067}{10000} - \frac{1}{2}}{\sqrt{\frac{1}{2} \cdot \frac{1}{2}/10000}} = 0.0067 \cdot 200 = 1.34$$

therefore the P-value is

$$2(1 - \Phi(1.34)) = 2(1 - 0.9099) = 0.1802.$$

No, H_0 can't be rejected at the .05 level.

c. Find a 95% CI for the proportion of heads for Kerrich's coin.

The confidence interval has endpoints

$$\begin{aligned} \hat{p} \pm z_{\alpha/2}\sqrt{\frac{\hat{p}\hat{q}}{n}} &= 0.5067 \pm 1.96\sqrt{\frac{0.5067 \cdot 0.4933}{10000}} \\ &= 0.5067 \pm 0.010 \end{aligned}$$

So the confidence interval is $[0.4967, 0.5167]$.

3. Calls to technical support service of a software company are monitored on a sampling basis for quality assurance. Each monitored call is classified as satisfactory or unsatisfactory by the supervisor in terms of the quality of help offered. A random sample of 100 calls was monitored over one month for a new trainee; 8 calls were classified as unsatisfactory.

a. Calculate a 95% CI for the actual proportion of unsatisfactory calls during the month. Use both formulas 9.1 and 9.3 and compare the results.

The first formula is the simpler one. It gives endpoints

$$\begin{aligned} \hat{p} \pm z_{\alpha/2}\sqrt{\frac{\hat{p}\hat{q}}{n}} &= 0.08 \pm 1.96\sqrt{0.08 \cdot 0.92/100} \\ &= 0.08 \pm 0.053 \end{aligned}$$

so that the interval is $[0.027, 0.133]$.

The second formula is more accurate, but more complicated. It gives the following endpoints where $z = z_{\alpha/2} = 1.96$

$$
\begin{aligned}
& \frac{\hat{p} + \dfrac{z^2}{2n} \pm \sqrt{\dfrac{\hat{p}\hat{q}z^2}{n} + \dfrac{z^4}{4n^2}}}{\left(1 + \dfrac{z^2}{n}\right)} \\
= & \frac{0.08 + \dfrac{1.96^2}{200} \pm \sqrt{\dfrac{0.08 \cdot 0.92 \cdot 1.96^2}{100} + \dfrac{1.96^4}{4(100)^2}}}{\left(1 + \dfrac{1.96^2}{100}\right)} \\
= & \frac{0.08 + 0.0192 \pm \sqrt{0.00283 + 0.00037}}{1.038} \\
= & \frac{0.0992 \pm 0.0566}{1.038} = 0.096 \pm 0.055
\end{aligned}
$$

so that the interval is $[0.041, 0.151]$.

The intervals are about the same length, but they're not centered at the same number.

b. This CI is used to test $H_0 : p = 0.10$ vs. $H_1 : p \neq 0.10$. If H_0 is not rejected, then monitoring of the trainee is continued at the same frequency; if H_0 is rejected in the lower tail, then monitoring frequency is reduced; and if H_0 is rejected in the upper tail, then the trainee is provided additional training. Based on the CI calculated in part a, what action should be taken on this trainee?

We'll reject H_0if $\hat{p}$ doesn't lie in the interval. Since $\hat{p} = 0.08$ does lie in the interval, don't reject H_0, so keep monitoring the trainee at the same frequency.

6. A blood test intended to identify patients at "high risk" of cardiac disease gave positive results on 80 out of 100 known cardiac patients, but also on 16 out of 200 known normal patients.

a. Find a 90% CI for the sensitivity of the test, which is defined as the probability that a cardiac patient is correctly identified.

Our population here consists of cardiac patients. There is some probability p that a cardiac patient will be identified as being high risk. The sample of size $n = 100$ gave an estimate of $\hat{p} = 0.80$. Therefore, a confidence interval has endpoints

$$
\begin{aligned}
\hat{p} \pm z_{\alpha/2}\sqrt{\frac{\hat{p}\hat{q}}{n}} &= 0.80 \pm 1.645\sqrt{0.80 \cdot 0.20/100} \\
&= 0.80 \pm 0.0658
\end{aligned}
$$

So the confidence interval is $[0.73, 0.87]$.

b. Find a 90% CI for the specificity of the test, which is defined as the probability that a normal patient is correctly identified.

Now our population consists of normal patients. There is some probability p (this is a different p than in part a) that a normal patient will be identified as not being high risk. The sample of size $n = 200$ gave an estimate of $\hat{p} = \frac{184}{200} = 0.92$. Therefore, a confidence interval has endpoints

$$
\begin{aligned}
\hat{p} \pm z_{\alpha/2}\sqrt{\frac{\hat{p}\hat{q}}{n}} &= 0.92 \pm 1.645\sqrt{0.92 \cdot 0.08/200} \\
&= 0.92 \pm 0.0316
\end{aligned}
$$

So the confidence interval is $[0.89, 0.95]$.

Math 218 Home Page at `http://math.clarku.edu/~djoyce/ma218/`