Confidence interval for a binomial proportion

In 1994 CNBC reported the results of a survey of top business executives. Of the 100 executives surveyed, 93 stated that they believed that the salaries of top management should be based on corporate performance. By contrast, only 63 stated that they believed that their company follows that policy.

What can we say about the opinions of top business executives in general on these questions? That is, what can we say about the true proportion of business executives who feel that salaries of top management should be based on corporate performance, for example? Call that unknown parameter p; can we construct a confidence interval for p? The answer is yes, by analogy to the construction for the interval for μ. The number of executives X who stated that they believe that salaries should be linked to corporate performance is binomially distributed, so if

$$\overline{p} \equiv \frac{X}{n}$$

then $E(\overline{p}) = p$ and $V(\overline{p}) = p(1-p)/n$. Thus, a $100 \times (1-\alpha)\%$ confidence interval for p is

$$\overline{p} \pm z_{\alpha/2}\sqrt{\overline{p}(1-\overline{p})/n}$$

(you might have expected that the interval would be based on a t–distribution, since we are estimating $V(\overline{p})$, but the theory of Gosset doesn't apply here, so we appeal to the Central Limit Theorem instead). So, for the above data, 95% confidence intervals for the two proportions are

$$.93 \pm (1.96)\sqrt{(.93)(.07)/100} = .93 \pm .05 = (.88, .98)$$

for the former question, and

$$.63 \pm (1.96)\sqrt{(.63)(.37)/100} = .63 \pm .095 = (.535, .725)$$

for the latter question, respectively. This interval is commonly called a *Wald* interval, since it is based on a construction originally proposed (in a wider context) by Abraham Wald.

These types of intervals are probably the ones most commonly seen in the popular media (in political polls, for example). You might also sometimes hear the "margin of error" of the estimate mentioned. What does that mean? In common usage, it refers to one–half the width of a 95% confidence interval, or in the above cases 5 percentage points and 9.5 percentage points, respectively. The unique nature of the binomial interval makes it possible to put an upper bound on this value **before** any sampling is done for a given sample size. The reason is that the width of the interval is maximized if $\overline{p} = .5$, implying that the maximum margin of error of the estimate is

$$(1.96)\sqrt{(.5)(.5)/n},$$

or roughly $1/\sqrt{n}$. So, for example, for $n = 100$ the maximum margin of error is $1/10 = .1$, or about 10 percentage points.

This interval is an approximate one, being based on the Central Limit Theorem (or, more precisely, the normal approximation to the binomial). A better interval would be one that is not approximate at all, but is based on the actual (exact) distribution of $\overline{p}$. Unfortunately, such an interval (which would be based on the inherently binomial distribution of the number of successes in the data) is not amenable to hand calculation. `Minitab` does provide this interval as the default choice, and this is the interval that should be used if the computer is available. For the data above, the exact 95% confidence intervals are $(.861, .971)$ and $(.528, .724)$. The benefits of using the exact interval are most apparent for small samples, and when $\overline{p}$ is close to zero or one. For example, say that 98 of the 100 business executives had stated that salaries should be based on corporate performance. The exact 95% confidence interval in this case is $(.930, .998)$, while the approximate interval is $(.953, 1.007)$; with an upper limit greater than one, we know that the latter interval can't be right.

There have been several suggestions made in the literature to improve the normal–based approximation to the exact binomial confidence interval. Details can be found in the Appendix.

Minitab commands

To obtain confidence intervals for a Binomial proportion, click on `Stat` → `Basic Statistics` → `1 Proportion`. Enter in the number of trials (n) and the number of successes (X) under `Summarized data`. The interval that comes out is the exact version (based on the Binomial distribution). To get the standard Central Limit Theorem–based interval, click on `Options`, and then mark the box next to `Use test and interval based on normal distribution`.

Appendix: Improvements to the normal–based interval

"Fake data" interval

A simple modification of the standard interval that seems to help coverage for small samples is to use the usual formulas, but to change the estimate of p to

$$\tilde{p} = \frac{X+2}{n+4}.$$

Note that this effectively adds two "fake" observations to the number of successes, and two "fake" observations to the number of failures (thus, we could call it the "+2" interval; it has recently been given the name *Agresti–Coull* interval, based on a 1998 paper by those authors investigating its properties). So, for the above examples, the confidence intervals are

$$.913 \pm [(1.96)\sqrt{(.913)(.087)/100}] = .913 \pm .055 = (.858, .968)$$

for the first question, and

$$.625 \pm [(1.96)\sqrt{(.625)(.375)/100}] = .625 \pm .095 = (.53, .72)$$

for the second question.

Technically, adding two successes and two failures is appropriate for a 95% confidence interval, but not in general. The formulation for general α is to add $z^2_{\alpha/2}/2$ successes and failures to the sample; when $\alpha = .05$, $z^2_{.025}/2 = 1.96^2/2 = 1.92 \approx 2$. So, for example, a 99% Agresti–Coull interval would add $z^2_{.005}/2 = 2.58^2/2 = 3.3$ fake successes and failures, or roughly 3 successes and 3 failures.

Score (q–)interval

T.J. Santner, in a 1998 article in *Teaching Statistics* ("Teaching large–sample binomial confidence intervals," **20**, 20–23), described an interesting (and apparently more effective) way to use the Central Limit Theorem to construct a confidence interval for p (which he termed a q–interval, but is usually called the score interval). The Central Limit Theorem implies that for large n

$$P\left(\left|\frac{\overline{p}-p}{\sqrt{p(1-p)/n}}\right| \le z_{\alpha/2}\right) \approx 1-\alpha.$$

Squaring both sides of the inequality in the parentheses gives the following:

$$P\left(\left[\frac{\overline{p}-p}{\sqrt{p(1-p)/n}}\right]^2 \le z^2_{\alpha/2}\right) = P\Big([\overline{p}-p]^2 \le z^2_{\alpha/2}p(1-p)/n\Big)$$
$$\approx 1-\alpha.$$

The zeroes of the quadratic equation $[\overline{p} - p]^2 = z^2_{\alpha/2} p(1-p)/n$ then define the endpoints of the confidence interval for p. Note that the only difference between this interval and the usual Gaussian–based one is that the standard error of $\overline{p}$ is estimated in the usual interval by $\sqrt{\overline{p}(1-\overline{p})/n}$, while the exact value $\sqrt{p(1-p)/n}$ is used for this interval. Some algebra ultimately gives the form of the score interval:

$$\left[\overline{p}\left(\frac{n}{n+z^2_{\alpha/2}}\right) + \frac{1}{2}\left(\frac{z^2_{\alpha/2}}{n+z^2_{\alpha/2}}\right)\right] \pm \sqrt{\left[\frac{\overline{p}(1-\overline{p})}{n}\right]\left[\frac{n^2 z^2_{\alpha/2}}{(n+z^2_{\alpha/2})^2}\right] + \frac{1}{4}\left[\frac{z^4_{\alpha/2}}{(n+z^2_{\alpha/2})^2}\right]}.$$

This looks quite daunting, but what's actually going on is some "fudging" with the estimates of both p and the standard error of the estimate of p. The first part of the interval shows that the estimate of p is "shrunk" from $\overline{p}$ towards $1/2$, the center of the possible values for p (this is the principle behind the "fake data" interval also). The second part of the interval shows that the estimated standard error is slightly adjusted, with the largest effect coming when $\overline{p}$ is close to 0 or 1. Santner showed that these score intervals achieve their nominal coverage levels of $1-\alpha$ for n as small as 5 for all p, while the usual and continuity–corrected intervals can have actual coverage less than $1-\alpha$ for $n \leq 20$ if p is close to 0 or 1. This interval is also called the Wilson interval.

The score intervals for the two previous examples are then

$$.914 \pm .052 = (.862, .966)$$

and

$$.625 \pm .093 = (.532, .718),$$

respectively. These are similar to the fake data intervals, which is often the case.

General population(s)

$\Downarrow$ sample (n)

samples (n_1, n_2)

Binomial population(s)

trials (n)

$\Downarrow$ 2 sets of trials (n_1, n_2)

ESTIMATION

μ with $\bar{x}$

σ with s

p with $\bar{p}$

HOW WELL DO THESE ESTIMATORS WORK?

$s.e.(\bar{x}) = \sigma/\sqrt{n}$

$s.e.(\bar{p}) = \sqrt{\frac{p(1-p)}{n}}$

$\bar{x} \sim N(\mu, \sigma^2/n) \longleftarrow$ C.L.T. $\longrightarrow \bar{p} \sim N\left(p, \frac{p(1-p)}{n}\right)$

INTERVAL ESTIMATION

$\bar{x} \pm z_{\alpha/2}\,\sigma/\sqrt{n} \longleftarrow$ C.L.T. $\longrightarrow \bar{p} \pm z_{\alpha/2}\sqrt{\frac{\bar{p}(1-\bar{p})}{n}}$

$\bar{x} \pm t_{\alpha/2}^{(n-1)}\, s/\sqrt{n}$ $\leftarrow$ Gosset, $\leftarrow$ normality } Normality less important for large n

$\bar{x} \pm t_{\alpha/2}^{(n-1)}\, s\sqrt{1+\frac{1}{n}}$ $\leftarrow$ Gosset prediction interval, $\leftarrow$ normality