5.3 Confidence Intervals for Proportions

5.4 Small-Sample Confidence Intervals for a Population Mean

5.3 Confidence Intervals for Proportions

The Old Method

Let

- $p = \frac{\#\text{successes in population}}{\text{population size}}$ = population proportion of successes (a fixed, unknown ______________)
  e.g. (from §5.1 lecture) p = U.S. unemployment rate
- X = #successes in n independent Bernoulli trials with $P(\text{success}) = p$: $X \sim$ ______________
  e.g. X = #unemployed in a random sample of size n ("success" = "unemployed")
- $\hat{p} = \frac{\#\text{successes in sample}}{\text{sample size}}$ = sample proportion of successes (a random ______________)
  So $\hat{p} =$ ——— (= 7.9% in October 3, January, 2013, Gallup poll)

If $np > 10$ and $n(1-p) > 10$, then $X \sim N(np, np(1-p))$ (≈, from §4.8 CLT), so

$$\hat{p} \sim N(\mu_{\hat{p}} = \_\_\_\_\_\_, \sigma^2_{\hat{p}} = \text{———}) = N(\_\_\_\_\_\_, \text{———})$$

.

We could use the §5.2 reasoning to derive a confidence interval for p. Instead, for a different approach, here I'll declare the interval and show that it has the advertised coverage.

Claim: the interval $\hat{p} \pm z_{\alpha/2}\sigma_{\hat{p}}$ contains p for a proportion $1-\alpha$ of random samples.

Proof: $P(p \in \text{ interval}) = P(\hat{p} - z_{\alpha/2}\sigma_{\hat{p}} < p < \hat{p} + z_{\alpha/2}\sigma_{\hat{p}})$

$=$

But we don't know $\sigma_{\hat{p}}$ because we don't know _______. For large n, we can estimate it with _______ to say that an approximate $100\%(1-\alpha)$ confidence interval for p is $\hat{p} \pm z_{\alpha/2}\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}$.

e.g. (p. 193 #3a) Leakage from underground fuel tanks pollutes water. In a random sample of 87 gas stations, 13 had at least one leaking tank. Find a 95% confidence interval for the proportion p of stations with at least one leaking tank.

In many cases, using $p \approx \hat{p}$ (in $\sigma_{\hat{p}}$) makes this interval ____________ to have its claimed confidence.

The New Plus-Four Method

Recent research (1998) describes an improvement: add ______________ observations, ______ successes and _________ failures, to the sample. That is, define $\tilde{n} =$ _________ and $\tilde{p} =$ ______ ("p-tilde") and use the "plus-four" interval

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

(Since $p \in [0, 1]$, _________ the interval if it extends outside $[0, 1]$.)

e.g. (p. 193 #3a) Find the 95% plus-four confidence interval for p in the leaking tank example.

Choosing the Sample Size

For a desired margin of error m, we can find the required sample size:

$m =$ $\qquad\qquad\qquad\qquad \Longrightarrow \tilde{n} =$ $\qquad\qquad\qquad\qquad\qquad \Longrightarrow n =$

(Note: the book forgot ____________ in its derivation on pp. 191-192.)

This relies on an estimate $\tilde{p}$ from ____________. If none is available, use $\tilde{p} =$ ______, which maximizes $\tilde{p}(1-\tilde{p})$, ensuring that n will be large enough to give the desired margin.

e.g. (p. 193 #3c) How many stations must be checked for leaks to get an error margin of .04?

A Pattern to Notice

Many confidence intervals have the form

(point estimate) $\pm$ (margin of error)

=(point estimate) $\pm$ (table value for confidence) $\times$ [(estimated or true) standard deviation of point estimate]

$=\hat{\theta} \pm$ (table value for confidence) $\times\ \sigma_{\hat{\theta}}$

5.4 Small-Sample Confidence Intervals for a Population Mean

In §5.2 we used the CLT to say that, for a large random sample $X_1, \cdots, X_n$ from a population with mean μ and standard deviation σ, $\bar{X} \sim$ ____________. We used this normal distribution to make a confidence interval for μ around $\bar{X}$. But the CLT is no help for a small sample: in this course, we're ________________________.

However, for the special case that *the population is* ____________, so that $X_i \sim N(\mu, \sigma^2)$, we saw in §4.3 that $\bar{X} \sim N(\mu, \frac{\sigma^2}{n})$ (____________), even for small n.

Standardizing gives $Z = \frac{\bar{X}-\mu}{\sigma/\sqrt{n}}$, but we don't know ______. For large n, we used ____________, but this approximation ____________ for small n. A new distribution solves this problem.

The Student's t Distribution

Define the random variable $T = \frac{\bar{X}-\mu}{s/\sqrt{n}}$. T's distribution isn't normal; it's the *Student's t distribution with* $n-1$ *degrees of freedom*, denoted t_{n-1}. ("Student" is a pseudonym for William Gosset, a statistician at ________________.) Here are some of its properties:

- T is a sample version of a ____________, estimating how far $\bar{X}$ is from __________________, in ________________________
- t_{n-1} looks like $N(0, 1)$: symmetric about ____, ____________-peaked, and ________-shaped
- T's variance is ____________ than Z's because estimating σ (_________) by s (____________) gives T more variation than Z: t_{n-1} is shorter with thicker tails (draw $N(0, 1)$ and t_{6-1})

- As n increases, t_{n-1} gets closer to ____________ (s becomes a __________________ of σ); in the limit as $n \to \infty$, they're ____________

Let $t_{n-1,\alpha}$ = the critical value t cutting off a __________________ area of α from t_{n-1} (draw). Table A.3 (p. 523) gives ____________ tail probabilities, using ν ("nu") for $n-1$.

e.g. Use Table A.3 to find the critical value t

- cutting off a right tail area of .05 from the t_{6-1} distribution: $t_{5,.05} =$ ____________
- such that the area under the t_{22-1} curve between $-t$ and t is 98%
- such that the area under the t_{25-1} curve left of t is .025
- such that the area under the t_{25-1} curve left of t is .70

Confidence Intervals Using the Student's t Distribution

We can work from $T = \frac{\bar{X}-\mu}{s/\sqrt{n}} \sim t_{n-1}$ to a $100\%(1-\alpha)$ confidence interval for μ.

Start with $P(-t_{n-1,\alpha/2} < T < t_{n-1,\alpha/2}) = 1-\alpha$ (draw). It implies

$P(-t_{n-1,\alpha/2} < \frac{\bar{X}-\mu}{s/\sqrt{n}} < t_{n-1,\alpha/2}) = 1-\alpha$, which we solve in two ways:

- for $\bar{X}$ in the middle: $P(\mu - t_{n-1,\alpha/2}\frac{s}{\sqrt{n}} < \bar{X} < \mu + t_{n-1,\alpha/2}\frac{s}{\sqrt{n}}) = 1-\alpha$ (draw)
- for μ: $P(\bar{X} - t_{n-1,\alpha/2}\frac{s}{\sqrt{n}} < \mu < \bar{X} + t_{n-1,\alpha/2}\frac{s}{\sqrt{n}}) = 1-\alpha$ (add to drawing for a typical $\bar{X}$)

That is, $\boxed{\bar{X} \pm t_{n-1,\alpha/2}\frac{s}{\sqrt{n}}}$ contains μ for a proportion $1-\alpha$ of random samples. It's the $100\%(1-\alpha)$ confidence interval for μ, for a small random sample from a __________ population with mean μ.

Example

e.g. We can study air bubbles in amber (fossilized tree resin) to learn what the atmosphere was like long ago. Amber bubbles from 85 million years ago (when _____________ were around) have these percentagess of nitrogen (N):

63.4 65.0 64.4 63.3 54.8 64.5 60.8 49.1 51.0

Assume this is a SRS from the atmosphere (some experts disagree). Find a 90% confidence interval for the mean percent of N in ancient air.

(Wiki says N is 78.1% today. Of which movie does this remind you? ____________________)

Cautions

- The data must be (reasonably regarded as) a _____________ from the population
- _____________ are uncommon in data from normal distributions (§4.3), so don't use this interval with data containing an _____________
- Use t interval if data appear reasonably normal (roughly symmetric, single _____, no _________)