Homework 8 — Solutions

STAT 515 · Inference for Proportions and for a Variance · 70 points

Instructor copy.

Problem 1: Confidence Interval for a Proportion (15 points)

$n = 980$, $x = 637$.

(a)

$$\hat{p} = \frac{637}{980} = 0.650.$$

(b) Large-sample rule: $n\hat{p} = 637 \geq 15$ and $n(1-\hat{p}) = 980 - 637 = 343 \geq 15$. Both counts are far above 15, so the normal approximation to the sampling distribution of $\hat{p}$ is safe.

(c)

$$\text{SE} = \sqrt{\frac{\hat{p}(1-\hat{p})}{n}} = \sqrt{\frac{(0.650)(0.350)}{980}} = \sqrt{0.0002321} = 0.01524,$$

$$0.650 \pm 1.96(0.01524) = 0.650 \pm 0.0299 \implies (0.620,\ 0.680).$$

(d) We are 95% confident that between 62.0% and 68.0% of all students at this university sleep fewer than 7 hours on a typical weeknight.

(e) Two-thirds is 0.667, which **falls inside** the interval (0.620, 0.680). The data are therefore consistent with the true proportion being above *or* below two-thirds, so the interval does **not** support the claim that more than two-thirds are sleep-deprived. The claim is plausible, but so is its opposite; the survey cannot settle the question.

(f) Narrower. A 90% interval uses the smaller multiplier $z^* = 1.645$ instead of 1.96, so the margin of error shrinks (from 0.0299 to 0.0251). Lower confidence buys greater precision.

Problem 2: Wilson's Adjusted (WAC) Interval (12 points)

$n = 20$, $x = 3$.

(a) $\hat{p} = 3/20 = 0.150$. Large-sample check: $n\hat{p} = 3 < 15$. The condition **fails** (the number of successes is far too small), even though $n(1-\hat{p}) = 17 \geq 15$. Both counts must clear the threshold.

(b)

$$\text{SE} = \sqrt{\frac{(0.15)(0.85)}{20}} = \sqrt{0.006375} = 0.0798,$$

$$0.150 \pm 1.96(0.0798) = 0.150 \pm 0.1565 \implies (-0.007,\ 0.307).$$

The lower endpoint is **negative**, which is impossible — a proportion cannot be less than 0. This is the visible symptom of applying a normal approximation when the sampling distribution of $\hat{p}$ is badly skewed, which is what happens when $\hat{p}$ sits near 0 with a small n.

(c)

$$\tilde{p} = \frac{x+2}{n+4} = \frac{5}{24} = 0.2083,$$

$$\text{SE} = \sqrt{\frac{\tilde{p}(1-\tilde{p})}{n+4}} = \sqrt{\frac{(0.2083)(0.7917)}{24}} = \sqrt{0.006873} = 0.0829,$$

$$0.2083 \pm 1.96(0.0829) = 0.2083 \pm 0.1625 \implies (0.046,\ 0.371).$$

(d) The Wald interval $(-0.007,\ 0.307)$ runs off the end of the parameter space; the Wilson interval $(0.046,\ 0.371)$ stays entirely within $[0, 1]$ and is shifted away from the boundary. Adding two successes and two failures pulls the estimate toward $1/2$, away from whichever boundary the data sit near. That small nudge both prevents impossible endpoints and — the real payoff — makes the *actual* coverage much closer to the advertised 95%; the plain Wald interval can cover far less than 95% of the time when $\hat{p}$ is near 0 or 1.

(e) It matters far more for the $n = 20$ survey. The adjustment adds a fixed 4 observations, so its influence shrinks as n grows: with $n = 980$ it would move $\hat{p}$ from 0.6500 to $639/984 = 0.6494$ — a change in the third decimal, invisible in the reported interval. With $n = 20$ it changes the estimate from 0.150 to 0.208 and rescues the interval from impossibility.

Problem 3: Determining the Sample Size for a Proportion (10 points)

(a) Use $p = 0.5$. The product $p(1-p)$ appears in the numerator and is maximized at $p = 0.5$, where $p(1-p) = 0.25$; any other value makes the product smaller. Choosing 0.5 therefore gives the **largest** required n, which guarantees the margin of error is met no matter what the true p turns out to be — the conservative, safe choice when nothing is known.

(b) 90% confidence, so $z^* = 1.645$, with $E = 0.04$:

$$n = \frac{(1.645)^2(0.5)(0.5)}{(0.04)^2} = \frac{(2.706)(0.25)}{0.0016} = \frac{0.6765}{0.0016} = 422.8 \implies n = 423.$$

(c) With $p = 0.25$, so $p(1-p) = 0.1875$:

$$n = \frac{(2.706)(0.1875)}{0.0016} = \frac{0.5074}{0.0016} = 317.1 \implies n = 318.$$

(d) A p farther from 0.5 means less variability in $\hat{p}$, so fewer observations are needed to pin it down to the same precision — here a saving of 105 respondents, about 25%. The risk is that the prior estimate is wrong: if the true p is nearer 0.5 than 0.25, the realized margin of error will **exceed** the ± 0.04 target and the poll will be less precise than promised.

Problem 4: Normality, Robustness, and Choosing a Procedure (10 points)

(a) A procedure is **robust** if its true performance — the actual coverage of a confidence interval, or the actual Type I error rate of a test — remains close to the nominal level even when one of its assumptions is moderately violated.

(b) For the t interval, the normality assumption is a convenience that the CLT increasingly makes unnecessary as n grows: moderate skew costs little, and a 95% interval still covers near 95% of the time. The analyst must therefore check normality before reporting a t interval for a mean.

(c) Points curving sharply upward at **both** ends, away from the line, indicate **heavy tails** — more extreme values in both directions than a normal distribution produces. I would **not** report the interval from 4(b). Heavy tails are exactly the violation the χ^2 interval is most sensitive to, and because extreme observations inflate s^2, the resulting interval would be both misplaced and far from its advertised 95% coverage.

(d) Random, close scatter about a straight line is consistent with a normal population, with no systematic departure to explain. All of the procedures in this assignment are then appropriate: the χ^2 interval for σ^2 and σ in Problem 4, and a t interval or t-test for the mean fill volume. (The proportion intervals in Problems 1 and 2 never required normality of the underlying measurements — they rely instead on the large-sample counts checked there.)

Problem 5: Proportions and Variances in R (8 points)

(a) The lines store the counts, compute $\hat{p} = x/n$, compute the standard error $\sqrt{\hat{p}(1-\hat{p})/n}$, and then build $\hat{p} \pm z^*\text{SE}$ using $z^* = \texttt{qnorm(0.975)} = 1.96$.

```
> phat
[1] 0.65
> se
[1] 0.01523624
> phat + c(-1, 1) * qnorm(0.975) * se
[1] 0.6201375 0.6798625
```

(b)

```
> x <- 3; n <- 20
> ptilde <- (x + 2) / (n + 4)
> se <- sqrt(ptilde * (1 - ptilde) / (n + 4))
> ptilde + c(-1, 1) * qnorm(0.975) * se
[1] 0.04585591 0.37081076
```

(c) The blanks are `0.975` and `0.025`, in that order — `qchisq(0.975, 19)` is the large critical value 32.852, which belongs in the denominator of the *lower* endpoint:

```
> n <- 20; s <- 4.6
> (n - 1) * s^2 / qchisq(c(0.975, 0.025), df = n - 1)
[1] 12.23779 45.13998
```

Common errors to watch for. Problem 1(e): concluding the claim is *supported* because $\hat{p} = 0.65$ is close to 0.667 — the point is that 0.667 lies inside the interval, so nothing is settled. Problem 2(a): checking only $n(1-\hat{p})$ and declaring the condition met. Problem 3(b): using $z^* = 1.96$ for 90% confidence. Problem 4(b): swapping the two χ^2 critical values, which inverts the interval — a quick sanity check is that $s^2 = 21.16$ must fall inside.
