Homework 7 — Solutions

STAT 515 · Hypothesis Testing for a Mean · 75 points

Instructor copy.

Problem 1: Setting Up Hypotheses (15 points)

(a) Caffeine. Let μ = the true mean caffeine content (mg) per can of the beverage.

$$H_0 : \mu = 80 \text{ mg} \qquad \text{vs.} \qquad H_a : \mu > 80 \text{ mg}.$$

One-sided (upper-tailed). The consumer group's suspicion — that cans contain *more* than claimed — is what they hope to demonstrate, so it goes in H_a. (Writing $H_0 : \mu \leq 80$ is also acceptable; the test is carried out at the boundary $\mu = 80$ either way.)

(b) Hospital stay. Let μ = the true mean length of stay (days) under the new protocol.

$$H_0 : \mu = 4.5 \text{ days} \qquad \text{vs.} \qquad H_a : \mu \neq 4.5 \text{ days}.$$

Two-sided. The administrators ask whether the mean has *changed*, with no direction specified — a shorter stay and a longer stay are both of interest.

(c) Fill machine. Let μ = the true mean fill volume (mL) of bottles from the machine.

$$H_0 : \mu = 500 \text{ mL} \qquad \text{vs.} \qquad H_a : \mu \neq 500 \text{ mL}.$$

Two-sided. Drift in either direction is a calibration failure: underfilling cheats the customer, overfilling costs the plant money.

(d) The null hypothesis must specify a single value of the parameter so that the sampling distribution of the test statistic can actually be computed — "$\mu = 500$" determines a distribution, "$\mu \neq 500$" does not. The researcher's claim goes in H_a because the logic of the test is one of *disproof*: we assume H_0 and look for evidence against it. Rejecting H_0 therefore constitutes positive evidence **for** the research claim, whereas failing to reject is never proof that H_0 is true.

Problem 2: Significance Level and Types of Error (12 points)

(a) α is the probability of rejecting H_0 when H_0 is in fact true — the maximum risk of a false positive the researcher is willing to accept, fixed *before* the data are examined.

(b) Type I error: rejecting H_0 when H_0 is true. In the fill-machine setting: concluding that the machine has drifted off 500 mL when it is in fact correctly calibrated. Consequence — the plant shuts down a working line and pays for an unnecessary recalibration.

(c) Type II error: failing to reject H_0 when H_a is true. Here: concluding there is insufficient evidence of drift when the machine really has drifted. Consequence — the plant keeps shipping mis-filled bottles, risking customer complaints or regulatory trouble.

(d) Either answer earns full credit when the reasoning is consistent.

- *Type II worse:* undetected drift ships defective product to every customer until the next inspection, which can mean regulatory penalties and reputational damage, while an unnecessary recalibration is a bounded, one-time cost.
- *Type I worse:* if recalibration requires halting a high-volume line for a full shift, a false alarm is very expensive, and small drifts would be caught at the next scheduled check anyway.

(e) Lowering α from 0.05 to 0.01 makes the rejection rule stricter: the probability of a **Type I error falls** to 0.01. Because stronger evidence is now demanded before rejecting, genuine departures are missed more often, so the probability of a **Type II error rises** (and the power falls). With n fixed, the two error probabilities move in opposite directions; the only way to reduce both at once is to increase the sample size.

Problem 3: Interpreting a p-Value (12 points)

(a) The p-value is the probability of obtaining a sample result at least as extreme as the one observed — in either direction, since the test is two-sided — **assuming $H_0 : \mu = 100$ is true**. A p-value of 0.068 says that if μ really were 100, samples this far from 100 (or farther) would occur about 6.8% of the time.

(b) $0.068 > 0.05$, so **fail to reject** H_0 at $\alpha = 0.05$. There is not enough evidence at the 5% level to conclude that μ differs from 100.

(c) $0.068 < 0.10$, so **reject** H_0 at $\alpha = 0.10$. There is enough evidence at the 10% level to conclude that μ differs from 100.

(d) The data have not changed — only the threshold has. A p-value of 0.068 is *borderline*: it lies between the two commonly used cutoffs, so the verdict flips depending on which was chosen. This is precisely why the p-value itself should be reported rather than a bare "significant / not significant": the reader can apply whatever threshold suits their own tolerance for a false positive, and can see that this result is marginal rather than decisive. Reporting only "fail to reject at $\alpha = 0.05$" would hide how close the call was.

(e) For a two-sided test the p-value splits evenly between the two tails, so when the sample mean falls in the direction specified by H_a,

$$p_{\text{one-sided}} = \frac{0.068}{2} = 0.034.$$

Note this must be planned in advance — halving a two-sided p-value *after* seeing which direction the data went is not legitimate.

(f) False. The p-value is computed *assuming* H_0 is true, so it cannot be a probability *about* H_0. It is $P(\text{data this extreme} \mid H_0 \text{ true})$, not $P(H_0 \text{ true} \mid \text{data})$. Reversing the conditioning is the most common misreading of a p-value.

Problem 4: One-Sample t-Test (18 points)

$n = 18$, $\bar{x} = 74.2$, $s = 8.6$, $\mu_0 = 78$.

(a) Let μ = the true mean resting heart rate (bpm) of cardiac rehab patients.

$$H_0 : \mu = 78 \qquad \text{vs.} \qquad H_a : \mu \neq 78.$$

(b)

$$\text{SE} = \frac{s}{\sqrt{n}} = \frac{8.6}{\sqrt{18}} = \frac{8.6}{4.2426} = 2.027 \text{ bpm},$$

$$t = \frac{\bar{x} - \mu_0}{s/\sqrt{n}} = \frac{74.2 - 78}{2.027} = \frac{-3.8}{2.027} = -1.87, \qquad \text{df} = n - 1 = 17.$$

(c) $p = 0.078 > 0.05$, so **fail to reject** H_0. At the 5% level there is insufficient evidence to conclude that the mean resting heart rate of cardiac rehab patients differs from 78 bpm. The observed difference of 3.8 bpm could reasonably be explained by sampling variability alone.

(d) Yes, the decision changes. $p = 0.078 < 0.10$, so at $\alpha = 0.10$ we **reject** H_0 and conclude that the mean resting heart rate of rehab patients differs from 78 bpm — specifically, the evidence points to a lower mean.

(e) With $t^*_{0.025,\,17} = 2.110$:

$$74.2 \pm 2.110(2.027) = 74.2 \pm 4.277 \implies (69.92,\ 78.48) \text{ bpm}.$$

(f) The hypothesized value $\mu_0 = 78$ **falls inside** the 95% confidence interval (69.92, 78.48). A two-sided test at $\alpha = 0.05$ rejects H_0 exactly when μ_0 falls *outside* the 95% interval, so the interval and the test must agree — and they do: 78 is a plausible value for μ, which is the same statement as "fail to reject at $\alpha = 0.05$." Notice how narrowly: 78 sits just barely inside the upper endpoint 78.48, mirroring the borderline p-value of 0.078.

Problem 5: Conditions and Robustness (10 points)

(a) Conditions for the one-sample t-test:

1. The data are a **random sample** from the population of interest.
2. Observations are **independent** of one another.

3. The population distribution is approximately **normal** — or n is large enough for the CLT to compensate.

With $n = 18$ the normality condition carries real weight, since 18 is below the usual CLT threshold.

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

(c) In practice it means mild to moderate non-normality does not invalidate the t-procedure: a 95% interval still captures μ close to 95% of the time. The protection grows with sample size, because the CLT is doing the work. For large n ($\gtrsim$ 30–40), even clearly skewed populations cause little trouble. For small n like 18, robustness is more limited — the t-test tolerates mild skew and slightly heavy tails, but **strong skew or outliers remain a genuine problem**, since a single extreme value moves both $\bar{x}$ and s.

(d) Plot the 18 observations against their theoretical normal quantiles (`qqnorm(hr); qqline(hr)`) and compare the points to the reference line.

- *Reassuring:* points scattered randomly and closely about the straight line, with no systematic bend and no isolated points far from the rest. Small wiggles are expected with $n = 18$ and should not be over-read.
- *Worrying:* systematic curvature (a bow indicating skew), a pronounced S-shape (heavy or light tails), or one or two points far off the line at an end (outliers). Any of these — especially outliers at this sample size — would make me distrust the t-test and look to a transformation or a nonparametric alternative.

Problem 6: Hypothesis Testing in R (8 points)

(a) Line by line: the first line stores the summary statistics; the second computes the t statistic $(\bar{x} - \mu_0)/(s/\sqrt{n})$; the third computes the two-sided p-value by finding the area below $-|t|$ in a t_{17} distribution and doubling it.

```
> tstat <- (xbar - mu0) / (s / sqrt(n))
> tstat
[1] -1.874652

> 2 * pt(-abs(tstat), df = n - 1)
[1] 0.07811
```

This confirms $t = -1.87$ and $p = 0.078$.

(b)

```
> qt(0.975, df = 17)
[1] 2.109816
```

Since $|t| = 1.87 < 2.110$, we fail to reject at $\alpha = 0.05$ — the same conclusion the p-value gave.

(c)

```
t.test(hr, mu = 78)
```

(The default `alternative = "two.sided"` and `conf.level = 0.95` are what this problem wants, so neither needs to be typed.)

Suggested Point Allocation (75 points)

Problem	Topic	Points	Breakdown
1	Setting up hypotheses	15	(a) 4, (b) 4, (c) 4, (d) 3
2	α, Type I and Type II error	12	(a) 2, (b) 3, (c) 3, (d) 2, (e) 2
3	p-value interpretation	12	(a) 3, (b) 2, (c) 2, (d) 2, (e) 2, (f) 1
4	One-sample t-test	18	(a) 3, (b) 4, (c) 4, (d) 2, (e) 3, (f) 2
5	Conditions and robustness	10	(a) 3, (b) 2, (c) 3, (d) 2
6	R	8	(a) 4, (b) 2, (c) 2

Common errors to watch for. Problem 1: putting the researcher's claim in H_0, or using $\bar{x}$ instead of μ in the hypotheses. Problem 3(f): the inverted-conditioning error — expect this one to be missed often, and it is worth class time. Problem 4(c)–(d): giving a decision with no contextual conclusion. Problem 4(f): students often assert the interval and test agree without naming *why* (μ_0 inside the interval). Problem 5(c): claiming the t-test is robust to *everything*, including outliers.