Assignment 2: Detailed Solutions

Statistical Inference 2026–27

Problem 1(a): Estimation risk versus testing power

Likelihood-ratio test

The likelihood at the observed vector x is

$$L(\theta; x) = (2\pi)^{-p/2} \exp\{-\tfrac{1}{2}\|x - \theta\|^2\}.$$

It is maximized over $\mathbb{R}^p$ at $\widehat{\theta} = x$. Hence the likelihood ratio is

$$\Lambda(x) = \frac{L(0; x)}{\sup_{\theta \in \mathbb{R}^p} L(\theta; x)} = \exp\{-\tfrac{1}{2}\|x\|^2\}.$$

Small values of Λ correspond to large values of $\|x\|^2$. Under H_0, the coordinates are independent standard normals, so

$$Q = \|X\|^2 = \sum_{j=1}^{p} X_j^2 \sim \chi^2_p.$$

Writing $q = \chi^2_{p,1-\alpha}$, the exact level-α test is

$$\boxed{\phi_{\mathrm{LR}}(x) = \mathbf{1}\{\|x\|^2 > q\}.}$$

The null distribution is continuous, so boundary randomization is unnecessary.

Positive-part James–Stein test

Put $k = p - 2 > 0$ and $r = \|x\|$. The proposed statistic is

$$g(r) = \|\widehat{\theta}^{+}_{\mathrm{JS}}(x)\| = \begin{cases} 0, & 0 \le r \le \sqrt{k}, \\ r - k/r, & r > \sqrt{k}. \end{cases}$$

For $r > \sqrt{k}$,

$$g'(r) = 1 + k/r^2 > 0.$$

The assumption $q > k$ ensures that $\sqrt{q}$ lies in this strictly increasing region. Choose the critical value

$$c_+ = g(\sqrt{q}) = \sqrt{q} - \frac{k}{\sqrt{q}} > 0.$$

Then

$$\{g(\|X\|) > c_+\} = \{\|X\| > \sqrt{q}\} = \{\|X\|^2 > q\}.$$

Thus the calibrated positive-part test and the likelihood-ratio test are identical for every observation. They have the same size and the same power at every θ. In particular, since $\|X\|^2 \sim \chi^2_p(\|\theta\|^2)$ under θ,

$$\beta(\theta) = \mathbb{P}_\theta(\|X\|^2 > q) = 1 - F_{\chi^2_p(\|\theta\|^2)}(q)$$

is their common power function.

The student's argument is incorrect. Squared-error risk measures the expected distance of an estimate from the parameter; testing power measures the probability of falling in a calibrated rejection region. Improving the former does not imply improving the latter. Comparing the two statistics at the same numerical cutoff would also fail to compare tests of the same size.

Original James–Stein test

For $r > 0$, its norm is

$$h(r) = \|\widehat{\theta}_{\mathrm{JS}}(x)\| = |r - k/r| = \begin{cases} k/r - r, & 0 < r < \sqrt{k}, \\ r - k/r, & r \geq \sqrt{k}. \end{cases}$$

This decreases from infinity to zero on $(0, \sqrt{k})$ and then increases to infinity. Large values of h therefore arise for both very small and very large radii.

For any positive critical value c, solving $|r - k/r| > c$ gives

$$0 < r < r_-(c) \quad \text{or} \quad r > r_+(c), \qquad r_\pm(c) = \frac{\sqrt{c^2 + 4k} \pm c}{2}.$$

The null law of $h(\|X\|)$ is continuous, and a level-α upper-tail test has a finite positive cutoff when $0 < \alpha < 1$. Its rejection region contains a punctured ball around zero of positive null probability. The LR test contains no such ball. Consequently, the rejection regions do not coincide, even up to null sets. Defining the estimator to be zero at $X = 0$ has no effect because that singleton has probability zero.

Problem 1(b): Nonexistence of a UMPU test

The coordinate tests are unbiased

Let $z = z_{1-\alpha/2} > 0$. For either $j = 1, 2$, the size is

$$\mathbb{E}_0 \phi_j(X) = \mathbb{P}_0(|X_j| > z) = \alpha.$$

At a general parameter θ, the power depends only on $t = \theta_j$:

$$b(t) = \mathbb{P}\{|N(t, 1)| > z\} = 1 - \Phi(z - t) + \Phi(-z - t).$$

This is an even function. With φ denoting the standard normal density,

$$b'(t) = \varphi(z - t) - \varphi(z + t).$$

For $t > 0$, $(z+t)^2 - (z-t)^2 = 4zt > 0$, so $\varphi(z-t) > \varphi(z+t)$ and $b'(t) > 0$. By symmetry, b is minimized at $t = 0$, where $b(0) = \alpha$. Therefore

$$\mathbb{E}_\theta \phi_j(X) = b(\theta_j) \geq \alpha \qquad \text{for every } \theta \neq 0.$$

Both tests are unbiased. Equality is allowed: if $\theta_j = 0$ but another coordinate is nonzero, the power is exactly α.

Mixture alternatives and uniqueness

Fix $a > 0$ and define the density

$$g_j(x) = \tfrac{1}{2} f_{ae_j}(x) + \tfrac{1}{2} f_{-ae_j}(x).$$

Its likelihood ratio against f_0 is

$$\ell_j(x) = g_j(x)/f_0(x) = e^{-a^2/2} \cosh(ax_j),$$

which is strictly increasing in $|x_j|$. The Neyman–Pearson test of f_0 against g_j rejects when $|x_j| > z$, so it is precisely ϕ_j.

Moreover, this most powerful test is unique up to null sets. Its likelihood-ratio cutoff is $K = e^{-a^2/2} \cosh(az) > 0$, and

$$\mathbb{P}_0\{\ell_j(X) = K\} = \mathbb{P}_0\{|X_j| = z\} = 0.$$

For completeness, let ψ be any possibly randomized test with $0 \leq \psi \leq 1$ and $\mathbb{E}_0 \psi \leq \alpha$. Then

$$\mathbb{E}_{g_j} \phi_j - \mathbb{E}_{g_j} \psi = \mathbb{E}_0[(\phi_j - \psi)(\ell_j - K)] + K(\alpha - \mathbb{E}_0 \psi) \geq 0.$$

Both terms are nonnegative. Away from the boundary, the first integrand is positive wherever $\psi \neq \phi_j$. Thus equality of mixture powers forces $\psi = \phi_j$ almost surely under P_0 (and forces full size). This is the precise uniqueness statement used below.

Contradiction for a putative UMPU test

Suppose an unbiased level-α UMPU test ψ exists. Since ϕ_j is unbiased, UMPU optimality implies

$$\mathbb{E}_{ae_j} \psi \geq \mathbb{E}_{ae_j} \phi_j, \qquad \mathbb{E}_{-ae_j} \psi \geq \mathbb{E}_{-ae_j} \phi_j.$$

Averaging gives $\mathbb{E}_{g_j} \psi \geq \mathbb{E}_{g_j} \phi_j$. The Neyman–Pearson bound gives the reverse inequality because ψ has level at most α. Hence equality holds. By uniqueness,

$$\psi = \phi_j \quad P_0\text{-almost surely.}$$

Applying this argument for $j = 1$ and $j = 2$ would require $\phi_1 = \phi_2$ almost surely. But independence under P_0 gives

$$\mathbb{P}_0(\phi_1 \neq \phi_2) = 2\alpha(1 - \alpha) > 0.$$

This contradiction proves that **no UMPU test exists**. Notice that $p \geq 2$ suffices for this nonexistence argument; $p \geq 3$ is needed for the stated James–Stein improvement.

There is no conflict with James–Stein theory. That result gives an estimator with smaller squared-error risk than a particular competitor. It does not assert the existence of a test maximizing power at every alternative. Also, estimator unbiasedness ($\mathbb{E}_\theta T = \theta$) and test unbiasedness (power at least size throughout the alternative) are different conditions.

Problem 2(a): Question, exclusions, and distributions

Parameter. Restrict the target to the specified age range. With Y denoting systolic blood pressure and A reported recreational activity, define

$$\Delta = \mathbb{E}[Y \mid A = \text{Yes},\ 20 \le \text{Age} \le 69] - \mathbb{E}[Y \mid A = \text{No},\ 20 \le \text{Age} \le 69].$$

The two-sided hypotheses are $H_0 : \Delta = 0$ and $H_1 : \Delta \neq 0$. A negative value means lower mean BP among active adults. This is an associational parameter, not an intervention effect.

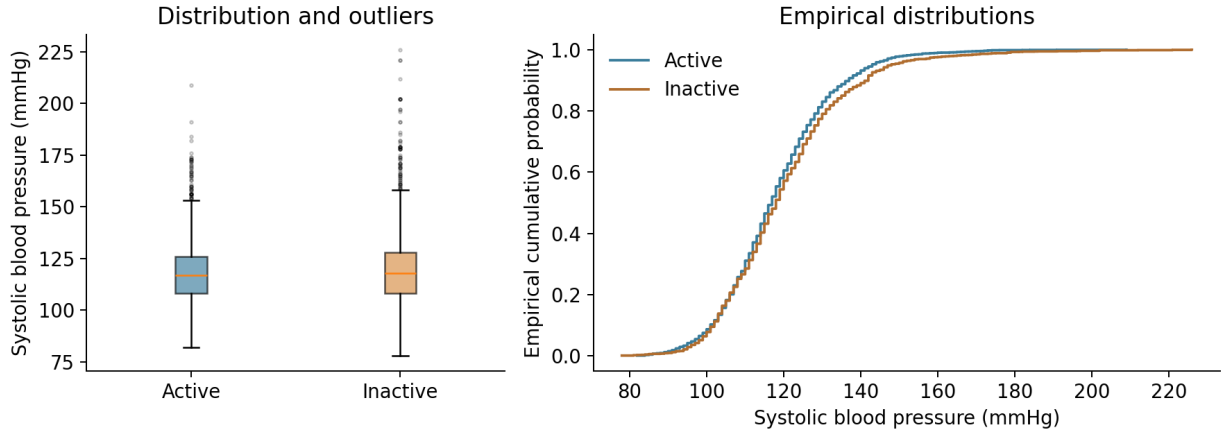
Complete-case estimation directly describes records with observed activity and BP. Extending it to all adults in the age range requires an assumption that missingness does not alter the relevant group means; the assignment's sampling convention does not itself supply that assumption.

Filtering step	Records
Original teaching dataset	10,000
Excluded: age outside 20–69	3,685
Age-eligible records	6,315
Excluded: missing activity or BP among age-eligible	231
Final analysis sample	6,084

There are no missing ages. Among age-eligible records, none have missing `PhysActive`; all 231 complete-case exclusions have missing `BPSysAve`. No additional trimming or removal of repeated IDs is performed.

Activity	n	Mean	SD	Median	Q1	Q3
Yes	3,381	118.09	14.36	117	108	126
No	2,703	120.13	16.76	118	108	128

All BP summaries are in mmHg. The IQRs are 18 and 20 mmHg. Both distributions have upper tails and unusually high readings; their ranges are 82–209 and 78–226 mmHg, respectively. They overlap substantially. The inactive group is more variable, with estimated variance ratio $s^2_{\text{No}}/s^2_{\text{Yes}} = 1.3623$. The large samples support mean-based asymptotic inference despite nonnormality; unequal variances motivate avoiding a pooled-variance assumption.



Data interpretation. The package documentation describes 10,000 records resampled for teaching and permits an educational simple-random-sample approximation. We use that convention here, not a claim of design-valid U.S. population inference. The groups are unpaired: no one-to-one matching of active and inactive records is specified. Resampling duplicates are retained as part of the stipulated teaching-data analysis.

Problem 2(b): Comparing mean blood pressure

Welch comparison

Let group 1 be active and group 0 inactive. Use Welch's test because it targets the difference in means without requiring equal variances. Define

$$\widehat{\Delta} = \bar{Y}_1 - \bar{Y}_0, \qquad \widehat{\mathrm{se}} = \sqrt{\frac{s_1^2}{n_1} + \frac{s_0^2}{n_0}}, \qquad T = \frac{\widehat{\Delta}}{\widehat{\mathrm{se}}}.$$

The Welch–Satterthwaite degrees of freedom are

$$\nu = \frac{(s_1^2/n_1 + s_0^2/n_0)^2}{(s_1^2/n_1)^2/(n_1-1) + (s_0^2/n_0)^2/(n_0-1)}.$$

Reject if $|T| > t_{\nu,0.975}$; the two-sided p-value is $2\mathbb{P}(t_\nu \geq |T|)$. A 95% confidence interval is $\widehat{\Delta} \pm t_{\nu,0.975}\widehat{\mathrm{se}}$.

Here,

$$\widehat{\Delta} = -2.03351, \quad \widehat{\mathrm{se}} = 0.40601, \quad \nu = 5335.10, \quad T = -5.00854,$$

$$\boxed{95\%\ \text{CI} = (-2.82945, -1.23757)\ \text{mmHg}, \qquad p = 5.6610 \times 10^{-7}.}$$

Under the stated sampling approximation, there is strong evidence of a nonzero pooled mean difference. Active adults have mean BP about 2.03 mmHg lower. Statistical significance alone does not establish clinical importance or causality. Welch calibration is approximate; its justification here relies on independent observations, finite variances, and large group sizes rather than exact normal sampling.

Bootstrap robustness analysis

Use $B = 10{,}000$ replicates and seed 20261028. Resample independently within each activity group, keeping the respective group sizes. Pooling would discard their distinct distributions.

Confidence interval. In replicate b, draw samples of sizes n_1 and n_0 with replacement from their respective groups and compute

$$D_b^* = \bar{Y}_{1,b}^* - \bar{Y}_{0,b}^*.$$

The empirical 2.5th and 97.5th percentiles give

$$\boxed{\text{Percentile bootstrap } 95\%\ \text{CI} = (-2.83420, -1.22476)\ \text{mmHg}.}$$

This is close to the Welch interval and excludes zero.

Null bootstrap test. An ordinary bootstrap distribution is centered near $\widehat{\Delta}$, not zero, so it must not be treated directly as a null distribution. Center each group's data at its own sample mean, imposing a common mean of zero while retaining the group's empirical shape and spread. Independently resample those centered values and compute

$$T_b^{*,0} = \frac{\bar{Y}_{1,b}^{*,0} - \bar{Y}_{0,b}^{*,0}}{\sqrt{(s_{1,b}^*)^2/n_1 + (s_{0,b}^*)^2/n_0}}.$$

Equivalently, using the uncentered bootstrap draws above,

$$T_b^{*,0} = \frac{D_b^* - \widehat{\Delta}}{\sqrt{(s_{1,b}^*)^2/n_1 + (s_{0,b}^*)^2/n_0}}.$$

Count $K = \sum_{b=1}^{B} \mathbf{1}\{|T_b^{*,0}| \geq |T|\}$ and report the simulation estimate $(K+1)/(B+1)$. In this run, $K = 0$, giving $1/10001 \approx 0.000100$.

This value is the reporting floor at this simulation size, not evidence that the true tail probability is precisely 10^{-4}. With zero exceedances, an approximate one-sided 95% upper bound on the conditional bootstrap tail probability is $3/B = 0.0003$. Ten thousand replicates cannot resolve a tail probability on the scale of 10^{-7}. The plus-one convention avoids reporting zero; it does not make this bootstrap procedure an exact test.

Both Welch and the null-centered bootstrap target $H_0 : \Delta = 0$ while allowing unequal variances and shapes. Their agreement supports the numerical robustness of the pooled association to these two calibrations. It is not independent confirmation from new data and does not eliminate confounding, missing-data selection, or survey-design limitations. Bootstrap validity also depends on the independent-sampling convention and suitable moment/regularity conditions; it is an approximation, not a distribution-free finite-sample guarantee.

Problem 2(c): Age as an alternative explanation

The age summaries, using the same complete-case sample, are:

Activity	Mean age	SD	Median	Q1	Q3
Yes	41.65	13.75	41	30	52
No	44.75	13.48	45	33	56

The inactive group is about 3.10 years older on average. The proportions aged 20–39 are 46.82% among active adults and 36.77% among inactive adults; the proportions aged 60–69 are 12.84% and 16.87%, respectively. Thus, the age compositions differ meaningfully.

Apply the same Welch comparison independently within each prespecified band. All differences below are active minus inactive, in mmHg; intervals are individual, unadjusted 95% intervals.

Age	n_1	n_0	Mean_1	Mean_0	Difference	95% CI
20–39	1,583	994	114.76	113.38	1.38	$(0.38, 2.37)$
40–59	1,364	1,253	119.35	122.71	-3.37	$(-4.60, -2.13)$
60–69	434	456	126.32	127.72	-1.40	$(-3.68, 0.87)$
Pooled	3,381	2,703	118.09	120.13	-2.03	$(-2.83, -1.24)$

The pooled finding does not describe every age band. In ages 20–39 the association reverses: active adults have a higher mean. Ages 40–59 show a lower mean among active adults. For ages 60–69, the interval includes zero and is compatible with differences in either direction; this is not proof of equality.

BP means increase across these age bands in both groups. Because active adults are younger on average, age composition can contribute to their lower pooled mean. In particular, if μ_{ag} is a band-specific mean and w_{ag} the proportion in band g within activity group a, then

$$\mu_1 - \mu_0 = \sum_g w_{1g}\mu_{1g} - \sum_g w_{0g}\mu_{0g}.$$

The weights differ between groups, so the pooled comparison is not simply a common-weight average of within-band differences.

These results are consistent with age confounding and suggest that a single pooled association conceals heterogeneity. They do not establish causal effect modification, nor do broad bands remove residual age confounding. Other factors, such as sex, health status, medication, and activity changes following illness, remain possible explanations. These are exploratory subgroup results; differences in significance alone are not a test that subgroup effects differ. The three intervals are not a simultaneous 95% confidence statement.

Problem 2(d): Stress-testing unequal variances

Data-generating mechanism

Choose a normal simulation to isolate unequal variances from skewness and heavy tails. Use the observed group sizes and standard deviations as fixed population parameters:

$$Y_{1i} \overset{\text{iid}}{\sim} N(0, 14.3566698^2), \quad i = 1, \ldots, 3381,$$

$$Y_{0i} \overset{\text{iid}}{\sim} N(0, 16.7566864^2), \quad i = 1, \ldots, 2703,$$

with independence within and between groups. Both population means are exactly zero and the variance ratio is 1.3622878. Choosing a common mean of zero is harmless because both tests are invariant to adding the same constant to all observations.

Generate $M = 10{,}000$ independent datasets using NumPy seed 20261029. On each dataset, apply two two-sided 5% tests:

1. Welch's test, using its replicate-specific standard error and degrees of freedom.

2. The pooled-variance t test, using

$$S_P^2 = \frac{(n_1 - 1)S_1^2 + (n_0 - 1)S_0^2}{n_1 + n_0 - 2}, \qquad T_P = \frac{\bar{Y}_1 - \bar{Y}_0}{S_P\sqrt{1/n_1 + 1/n_0}},$$

and reference degrees of freedom $n_1 + n_0 - 2$.

The pooled test's exact normal-theory justification requires a common variance, which this simulation intentionally violates. Welch's procedure accommodates distinct variances, with approximate t calibration.

Results and Monte Carlo uncertainty

For each test, let R_b indicate rejection in replicate b. Estimate its null rejection probability and Monte Carlo standard error by

$$\widehat{r} = \frac{1}{M}\sum_{b=1}^{M} R_b, \qquad \widehat{\mathrm{MCSE}} = \sqrt{\frac{\widehat{r}(1 - \widehat{r})}{M}}.$$

Test	Rejections	Rejection rate	MCSE
Welch	515	0.0515	0.00221
Pooled variance	552	0.0552	0.00228

Thus Welch rejects in 5.15% of datasets and the pooled test in 5.52%, with Monte Carlo uncertainty of about 0.22–0.23 percentage points. These are rejection rates under a true null, not estimates of power against a nonzero mean difference.

The pooled procedure mildly overrejects in this configuration. The larger variance belongs to the smaller group, and pooling understates the variance of the mean difference. Indeed, with the simulated population variances,

$$\frac{\mathbb{E}(S_P^2)(1/n_1 + 1/n_0)}{\sigma_1^2/n_1 + \sigma_0^2/n_0} = 0.96639.$$

Its estimated variance is therefore about 3.36% too small on average. The distortion is mild because the variance ratio and sample-size imbalance are modest. An assumption violation need not create a dramatic failure in every configuration, and different simulation seeds give slightly different rates.

The simulation supports choosing Welch's procedure over imposing equal variances here. It does not establish that actual NHANES observations are normal, independent, or missing at random, and it says nothing causal about activity. It also does not establish universal superiority of Welch's test under every distribution. Pooling observed groups to generate the simulated null would erase their unequal variances and fail to test the feature at issue.