Practice Quiz 5: Hypothesis Testing + Multiple Testing

MSE 125 — Lectures 10–11

Use this practice quiz to prepare for Quiz 5 (Wednesday, May 6). The real quiz will have 2 questions in 10 minutes, closed-book. This practice set has 10 questions covering Lectures 10–11: the hypothesis-testing framework (H_0/H_1, choice of test), Type I and Type II errors and the α-β tradeoff, power and effect-size sensitivity, the multiple testing problem, the p-value histogram diagnostic, Bonferroni and Benjamini-Hochberg corrections, confounding and Simpson's paradox, p-hacking and pre-registration, and statistical significance versus practical importance.

Every concept tested on the real quiz appears somewhere on this practice set, with a different scenario.

Question 1. For each scenario, state H_0 and H_1, and choose the appropriate test (**Welch's t-test** or **one-sample z-test for a proportion**). Briefly justify your choice in each case (1 sentence).

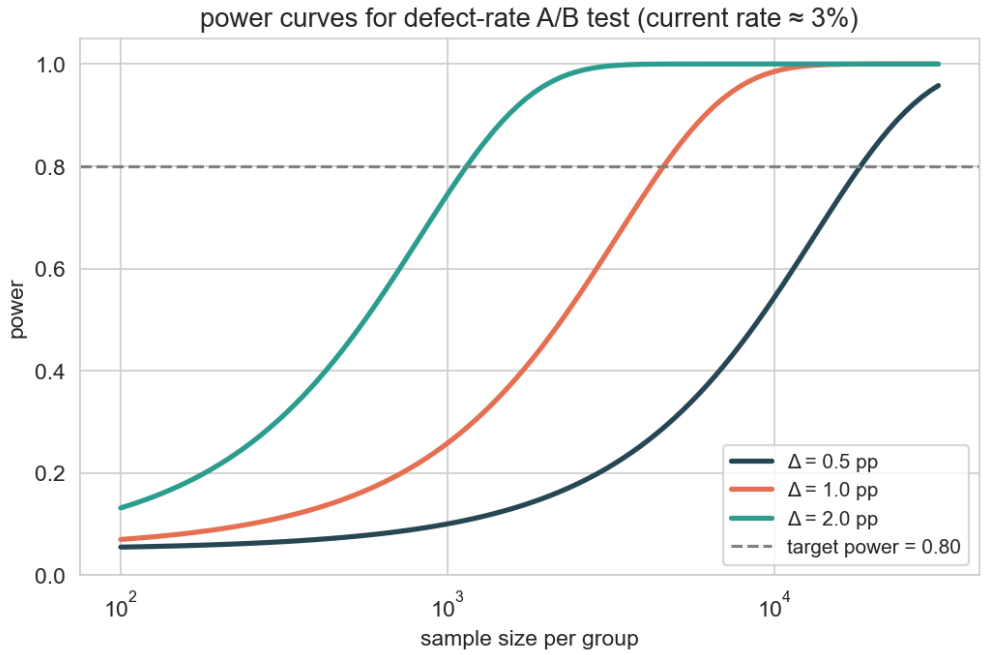
(a) A clinic compares **average wait times (in minutes)** between two staffing schedules to decide whether one schedule is faster.

(b) An election analyst checks whether the **click-through rate** on a get-out-the-vote ad differs from the platform's reported benchmark rate of **4%**.

Solution

(a) $H_0 : \mu_1 - \mu_2 = 0$ (average wait times are equal across schedules); $H_1 : \mu_1 - \mu_2 \neq 0$. Use **Welch's t-test** — the outcome is continuous, the standard error must be estimated from the sample variances, and we are not assuming equal variance.
(b) $H_0 : p = 0.04$ (the ad's click-through rate equals the benchmark); $H_1 : p \neq 0.04$. Use the **one-sample z-test for a proportion** — the outcome is binary, and under H_0 the standard error $\sqrt{p_0(1-p_0)/n}$ is pinned by the null with nothing to estimate from the data.

Question 2. A factory currently runs at a 3% defect rate. Engineers are considering a process change and want to know how large an A/B test to run. They plot power vs sample size per group for three assumed defect-rate reductions Δ.



(a) From the plot, what's the smallest sample size per group that achieves **80% power** if the team assumes a $\Delta = 1.0$ percentage-point reduction? _____

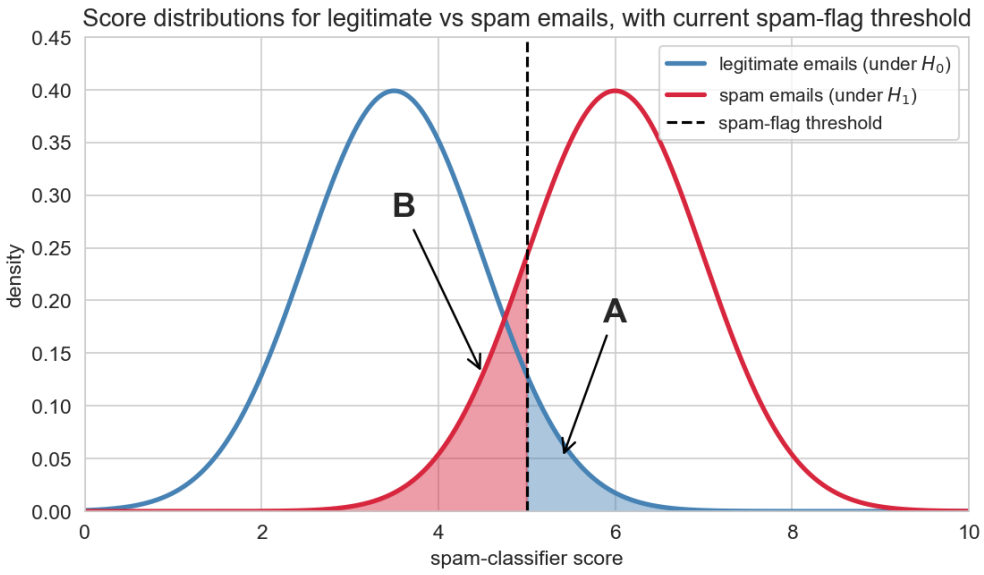
(b) The factory's testing budget allows at most $n = 2{,}000$ units per group. From the plot, what is the **smallest defect-rate reduction** they can reliably detect (power ≥ 0.80) within that budget? Circle one: $\Delta \approx 0.5$ pp / $\Delta \approx 1.0$ pp / $\Delta \approx 2.0$ pp

(c) Why does halving the assumed effect size (from $\Delta = 1$ pp to $\Delta = 0.5$ pp) require **more than twice** the sample size to keep power constant? Answer in 1 sentence.

Solution

(a) Reading the orange ($\Delta = 1.0$ pp) curve where it crosses 0.80: $n \approx$ **4,500** per group (accept any answer in $[3{,}500, 6{,}000]$).
(b) $\Delta \approx 2.0$ **pp.** At $n = 2{,}000$ the green curve sits at power ≈ 1.0, the orange curve sits at power ≈ 0.5 (well below 80%), and the dark blue curve hasn't even started rising. Only $\Delta = 2.0$ pp is reliably detectable at this budget.
(c) Required n scales like $1/\Delta^2$ (because the standard error shrinks like $1/\sqrt{n}$, and power depends on the *standardized* effect Δ/SE). Halving Δ requires roughly **4×** the sample size, not 2×.

Question 3. A spam filter classifies each incoming email by computing a "spam score." If the score exceeds a threshold, the email is flagged as spam; otherwise it goes to the inbox. Frame this as a hypothesis test where H_0: the email is legitimate. The figure below shows the score distributions for legitimate emails (under H_0) and spam emails (under H_1), with the current threshold marked.



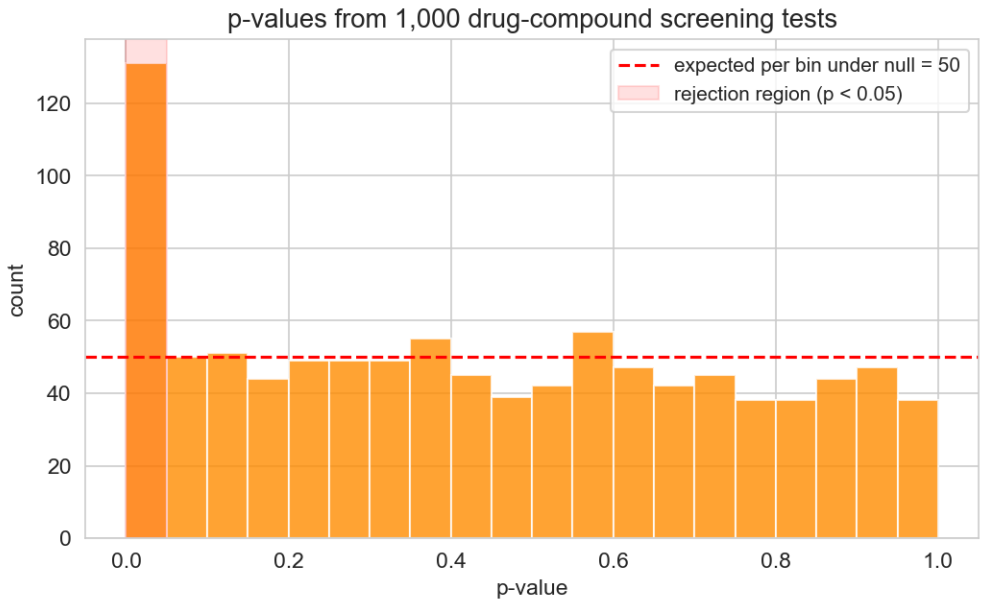
(a) Identify each shaded region. Region **A** is the (circle one): **Type I error rate (α)** / **Type II error rate (β)**. Region **B** is the (circle one): **Type I error rate (α)** / **Type II error rate (β)**. In 1 sentence, state in plain language what the Type I error means here (which kind of email ends up in which folder).

(b) The user complains that too many spam emails reach the inbox. The team proposes **lowering the threshold** (so more emails get flagged as spam) — i.e., moving the dashed line **leftward**. Which shaded region **shrinks**? Which **grows**? In 1 sentence, name a real-world cost of the now-larger error.

Solution

(a) Region **A** = **Type I error rate (α)**: legitimate emails with high enough scores to land past the threshold (right tail of the blue density). Region **B** = **Type II error rate (β)**: spam emails with low enough scores to fall below the threshold (left tail of the red density).
A **Type I error** here means a legitimate email gets flagged as spam (sent to the spam folder by mistake).
(b) Lowering the threshold (moving the dashed line leftward) makes it easier to call something spam — equivalent to easier rejection of H_0. **Region B (Type II) shrinks**: fewer spam emails sneak past the threshold into the inbox. **Region A (Type I) grows**: more legitimate emails get incorrectly flagged.
Real-world cost of the now-larger Type I error: users miss important legitimate emails — a job-offer reply, an urgent message from a doctor, a misflagged invoice. The classic α-β tradeoff: you cannot shrink both error rates at once by moving the threshold alone. To improve both, the team needs a **better-separated classifier** (a different design lever), not just a different threshold.

Question 4. A pharmaceutical company screens 1,000 candidate compounds for binding to a target enzyme. For each compound, they run a hypothesis test against a negative control at $\alpha = 0.05$. The histogram of all 1,000 p-values is below.



(a) If **none** of the 1,000 compounds actually binds (every H_0 is true), how many of the 1,000 tests would you expect to come back significant at $\alpha = 0.05$? _____

(b) The team observed 131 compounds significant at $\alpha = 0.05$. Is there real signal beyond chance? In 1 sentence, point to the feature of the histogram that supports your answer.

(c) Of the 131 observed significant compounds, roughly how many are likely to be **false positives** (chance rejections from null compounds)?

(d) The team plans to advance all 131 reported significant compounds to follow-up testing. Approximately what **fraction** of the 131 are likely to be *real binders* (vs. chance rejections from null compounds)? Use your numbers from (a)–(c). In 1 sentence, what does this fraction tell the team about how much follow-up effort will be wasted on noise?

Solution

(a) $1{,}000 \times 0.05 = \mathbf{50}$ false positives expected under the null.
(b) Yes. The first bin sits at ~131, well above the null floor of ~50/bin shown by the red dashed line. The other 19 bins hover near 50 — exactly the flat baseline expected under uniformity. The spike-near-zero shape is the multi-test signature for "real effects exist."
(c) Roughly **50** (or "about 50/131 38% of the observed significant compounds"). Under the null, ~50 false positives would land in the first bin by chance. The remaining ~80 are likely real signals.
(d) Likely real binders: $131 - 50 = 81$. Fraction real: $81/131 \approx \mathbf{62}\%$. So roughly **38% of follow-up effort will be wasted** on chance rejections from inactive compounds. If follow-up is cheap (e.g., a quick repeat assay), this rate is acceptable; if follow-up is expensive (animal studies, weeks of bench work), the team should apply a stricter threshold first to reduce the noise fraction. *(This is the same fraction-of-discoveries-that-are-real reasoning that drives the genome-wide significance threshold of* 5×10^{-8} *— at GWAS scale the prior is so much lower that any "discovery" at* $\alpha = 0.05$ *would be almost entirely noise.)*

Question 5. A marketing team A/B tests **80 different email subject lines** for click-through rate. They want to control the family-wise error rate at $\alpha = 0.05$ — that is, ensure at most a 5% chance of *any* false positive across the 80 tests.

(a) What is the **Bonferroni threshold** the team should use? _____

(b) Without computation: only subject lines with **what kind of effect-size signature** survive this threshold? Circle one and justify in 1 sentence: very strong effects only / moderate-or-larger effects / any effect with $p < 0.05$

Solution

(a) $\alpha_{\text{Bonf}} = \alpha/m = 0.05/80 = \mathbf{6.25 \times 10^{-4}}$ (about 0.000625, or 80× stricter than 0.05).
(b) Very strong effects only. A p-value of 6×10^{-4} corresponds to a t-statistic of about 3.4 — substantially larger than the 1.96 needed at $\alpha = 0.05$. Subject lines with moderate effects (p in the 0.001 to 0.05 range) get filtered out by Bonferroni even when their effects are real. *(That conservative loss is the cost of FWER control; if the team can tolerate a fraction of false discoveries, BH would catch more of the moderate effects.)*

Question 6. Pick **Bonferroni** or **Benjamini-Hochberg (BH)** for each scenario, and justify in 1 sentence.

(a) An aircraft manufacturer tests 20 candidate alloy compositions for a critical wing component. Any approval of an unsafe alloy could cause a fatal accident.

(b) A genomics lab screens 5,000 candidate gene variants for follow-up *experimental* validation in a wet lab. The wet lab can run any number of follow-ups, and the lab leader expects most flagged variants to fail in follow-up.

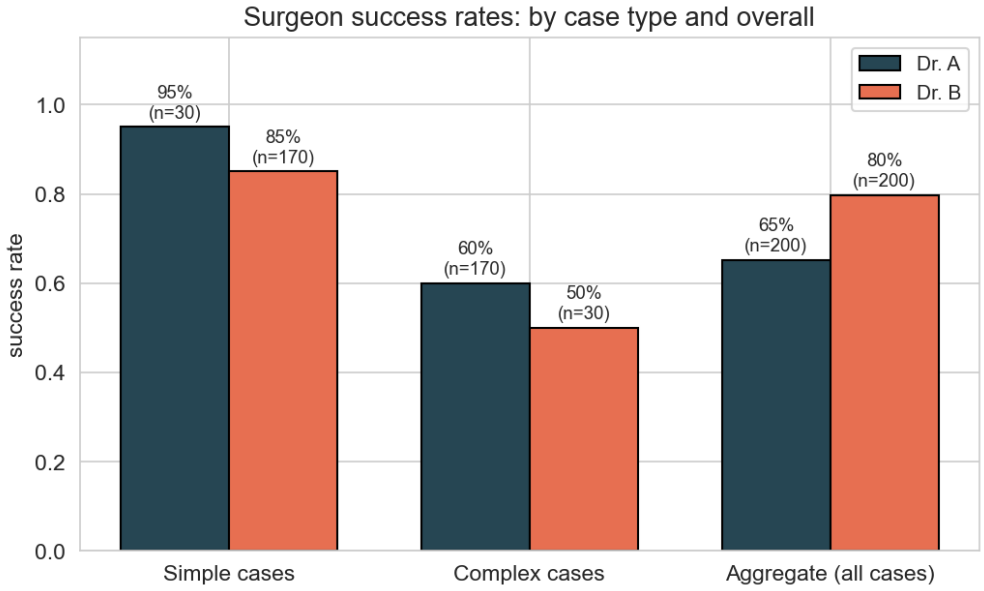
(c) A government agency audits 200 contractors for fraud. A false accusation of fraud creates legal liability and damages the contractor's reputation.

Solution

(a) Bonferroni. Any false positive (approving an unsafe alloy) is catastrophic — we want to control the probability of *any* error, even at the cost of missing some safe alloys. FWER control is the right framing.
(b) BH. The lab is willing to follow up on candidates and weed out false positives experimentally, so a controlled fraction of false discoveries is acceptable. BH catches more real signals than Bonferroni when individual p-values are moderate (Hedenfalk-style).
(c) Bonferroni is the safer default. Each false accusation has a real, severe cost (legal liability, reputation damage). Treat any false positive as costly — control FWER. *(This is the "drug safety / criminal trial" pattern from the chapter.)*
Acceptable alternative for partial credit: a student who argues for **BH** on a "screen-then-investigate" framing — flag a candidate set with controlled FDR, then internally investigate each before issuing any public accusation — has identified a defensible workflow. The key judgment is that *the public accusation* must be FWER-controlled, not necessarily the *initial*

flag; arguing this distinction explicitly earns full credit too.

Question 7. Two surgeons report their success rates by case difficulty. The figure below shows simple cases, complex cases, and the aggregate (all cases) for both surgeons; sample sizes are marked.



(a) If you needed a **complex** procedure, which surgeon would you choose? Justify in 1 sentence.

(b) The hospital's annual report shows only the **aggregate** success rates. Which surgeon would the report flag as the "better" surgeon? In 1–2 sentences, name the statistical phenomenon at play and explain what's driving the aggregate-vs-subgroup reversal.

Solution

(a) Dr. A. Dr. A's complex-case success rate is 60% (vs 50% for Dr. B). For a patient choosing a surgeon for a complex procedure, the within-subgroup rate is the relevant number — that's the population they belong to.
(b) The aggregate report would flag **Dr. B** as the better surgeon (80% overall vs 65% for Dr. A). This is **Simpson's paradox**: aggregate rates can reverse the within-subgroup story. The driver is the **case mix** — Dr. B handles mostly simple cases (170 of 200), where success rates are naturally higher, while Dr. A is given mostly complex cases (170 of 200). Dr. B's higher aggregate just reflects easier cases, not better skill. Case difficulty is a **confounder**: it drives both the input (which surgeon gets the case) and the outcome (success rate), so the aggregate comparison answers the wrong question.

Question 8. A health blog reports: *"Counties with more Whole Foods locations per capita have 18% lower rates of heart disease ($r = -0.42$). Conclusion: shopping at Whole Foods prevents heart disease."*

(a) Name **two distinct confounders** that could produce this correlation without organic groceries causing anything.

(b) The correlation is computed at the **county** level (one point per county), not the individual level. Why might the *individual-level* correlation between organic shopping and heart disease be much weaker than $r = -0.42$? Name the term for this phenomenon (1 sentence).

(c) The blog later updates: *"We tested 50 different supermarket chains; only Whole Foods showed a significant correlation with reduced heart disease. The result survives Bonferroni correction at $\alpha = 0.05$. This proves the effect is real."* Does Bonferroni-survival establish that Whole Foods *causes* lower heart disease? Answer in 1–2 sentences explaining what Bonferroni does and does not control.

Solution

(a) Any two of:

- **Income / wealth.** Whole Foods locates in high-income areas; high-income individuals have better healthcare access, healthier diets, more exercise time, lower smoking rates. Income drives both Whole Foods presence and heart disease.
- **Education level.** Higher education predicts both organic-food preference and heart-protective behaviors.
- **Age distribution.** Whole Foods may avoid retirement-heavy areas; younger populations have lower baseline heart disease.
- **Urban vs rural.** Whole Foods locates in dense urban areas; urban populations differ from rural ones on many health-relevant dimensions (air quality, walkability, healthcare access).

(b) Ecological correlation fallacy. Correlations on group averages (one point per county) can be much stronger than correlations on individuals — averaging within counties hides within-county variation, so the strong county-level r may correspond to a much weaker (or null) person-level correlation.

(c) No. Bonferroni controls the chance of finding *false* correlations among many tests — a guarantee against false positives produced by chance alone. It does **not** address **confounding**: the income/education/urban-rural variables flagged in (a) can produce a correlation between Whole Foods and heart disease that survives any multiple-testing correction, because the correlation is real (not chance), but its *cause* is the confounders, not the groceries. Multiple-testing corrections protect against discovering effects that aren't there; they cannot protect against discovering effects that are there but reflect the wrong mechanism.

Question 9. An education researcher **pre-registers** 12th-grade math test scores as the primary outcome of a school-wide reading-curriculum trial. After the trial, the result for the primary outcome is $p = 0.31$. She then explores 50 secondary analyses (different subgroups, subjects, and outcome measures) and reports: *"The new curriculum significantly improved 8th-grade reading scores in low-income districts ($p = 0.04$). The curriculum works."*

(a) The press release for the trial reads: *"Curriculum failed: no benefit found."* Why is this an overstatement of the primary result? Reference the difference between **"fail to reject"** and **"accept."** (1–2 sentences)

(b) Why is the secondary finding ($p = 0.04$ in low-income 8th-grade reading) misleading? Name the practice and explain in 1–2 sentences what would have prevented it.

Solution

(a) Failing to reject H_0 at $p = 0.31$ does **not** mean "the curriculum has no effect." It means the data we collected are *compatible* with H_0, but they are also compatible with many small (or even moderate) real effects we lacked the power to detect. **Absence of evidence is not evidence of absence.** A more accurate press release: "the trial did not find evidence of an effect on the primary outcome at the planned sample size."
(b) This is **p-hacking** (also called **data dredging** or exploiting **researcher degrees of freedom**) — running many tests and reporting only the survivor that crossed $p < 0.05$. With 50 secondary tests at $\alpha = 0.05$, we'd expect ~ 2.5 false positives by chance even if the curriculum has no effect anywhere. The reported $p = 0.04$ is the survivor of hidden multiplicity, not a confirmed finding. **Pre-registration of the secondary outcomes** would have prevented it: any secondary outcome the researcher cared about should have been listed up-front, with a multiple-testing correction applied. Findings from the unplanned 50-outcome exploration should be labeled **exploratory** and treated as hypotheses for a future confirmatory trial — not as discoveries.

Question 10. A mobile app team runs an A/B test for a new UI tweak with $n = 500{,}000$ users per arm. Results:

- Control mean session duration: 8.20 min
- Treatment mean session duration: 8.22 min
- 95% CI for the difference: **[+0.01, +0.03] minutes**
- p-value: **< 0.001**

The product manager declares: *"Statistically significant — ship it!"*

(a) Without computing anything, what does the 95% CI tell you about the verdict of a two-sided hypothesis test of $H_0 : \mu_T - \mu_C = 0$ at $\alpha = 0.05$? Justify in 1 sentence using the CI/test duality.

(b) Despite the tiny p-value, give **one reason** you might recommend NOT shipping this UI change. What is the term for the failure mode the PM is exhibiting?

Solution

(a) The 95% CI **excludes 0**, so by **CI/test duality** the two-sided test rejects H_0 at $\alpha = 0.05$. (Equivalently: the duality says the 95% CI is exactly the set of null values that the test would *fail* to reject; since 0 is not in that set, the test rejects.)
(b) Reasons (any one earns full credit):

- **The effect is tiny** — 0.02 minutes is 1.2 seconds on an 8-minute session, about 0.25%. Users will not notice this. The cost of shipping (engineering, maintenance, A/B-test pipeline real estate) likely outweighs a 1.2-second benefit.
- **With $n = 500{,}000$ per arm, the test has so much power that any nonzero effect — including effects too small to matter — becomes "significant."** A small

p-value here is consistent with a clinically meaningless effect.

- **Opportunity cost** — the engineering effort could be spent on a change that produces a meaningful effect.

Failure mode: **confusing statistical significance with practical importance.** A small p-value tells you the effect is unlikely to be exactly zero. It does *not* tell you the effect is large enough to matter. **Always report the effect size and CI alongside the p-value** so the reader can judge practical importance — not just whether the effect is distinguishable from zero.