Hypothesis Tests

PSCI 1801 · Class Handout

Dr. Marc Trussler

This handout is designed as an in-class supplement to the textbook. The material contained on this handout does not represent the full set of information you need to know for the exams. Exclusively studying off of this handout will not be sufficient for the exams.

The Logic of Hypothesis Testing

The core question of statistics: "How likely is my sample result under a specific hypothesis about the population?"

If a sample result would be very unusual under the hypothesis, we reject the hypothesis. If not, we retain it.

What counts as "very unusual" is our chosen significance level α *— the false positive rate we're willing to tolerate. Standard is* $\alpha = 0.05$.

The Five Steps of a Hypothesis Test

1. Specify null (H_0) and alternative (H_a) hypotheses.

The null is what we're trying to disprove. Most often "no effect" or "equal to some baseline." One-tailed vs two-tailed:

- One-tailed: $H_a : \mu > \mu_0$ or $H_a : \mu < \mu_0$
- Two-tailed: $H_a : \mu \neq \mu_0$ (the default in social science — it's more conservative)

2. Choose a test statistic and the level of test α.

The test statistic is what we compute from the sample (usually $\bar{X}$*). α is conventionally 0.05.*

3. Derive the sampling distribution and center it on the null hypothesis.

By CLT, this is $N(\mu_0, S^2/n)$ — or a t-distribution for small samples.

4. Compute the p-value.

The p-value is the probability of observing a result at least as extreme as our test statistic if H_0 is true.

- One-tailed: area in the single tail beyond the test statistic
- Two-tailed: doubled — area in both tails

5. Decision rule.

- If $p \leq \alpha$: **reject** H_0
- If $p > \alpha$: **fail to reject** H_0 (not "accept"!)

The T-Statistic

Because we're using S (the sample SD) as a stand-in for σ, we technically use the t-distribution, not the normal:

$$t = \frac{\bar{X} - \mu_0}{S/\sqrt{n}}$$

This is "how many standard errors is our sample mean from the null?" For large n, the t-distribution is indistinguishable from the normal.

Look up the p-value with `pt(t, df = n - 1)`*. Or use* `t.test()`*, which does everything.*

Worked Example: Kelly in Arizona

2013 poll of 2613 Arizonans: Kelly's two-party support = 52.4%. Is this significantly different from a tied race?

Steps 1–2: $H_0 : p = 0.5$, $H_a : p \neq 0.5$, $\alpha = 0.05$.

Step 3: Under H_0, sampling distribution is $N(0.5, \sigma^2/n)$.

Step 4: Compute the t-statistic and the two-sided p-value:

```
t = 2.408  p = 0.0161
```

Step 5: $p < 0.05$, so we reject the null. There's strong evidence Kelly is not tied.

Or just use `t.test()`

```
    One Sample t-test

data:  sm.az$DemSenateVote
t = 2.4084, df = 2612, p-value = 0.01609
alternative hypothesis: true mean is not equal to 0.5
95 percent confidence interval:
 0.5043737 0.5426986
sample estimates:
mean of x
0.5235362
```

CIs and Hypothesis Tests Are Equivalent

A two-tailed hypothesis test at α rejects H_0 if and only if a $(1-\alpha)\cdot 100\%$ CI fails to contain μ_0.

So: if a 95% CI does not include the null value, the two-tailed test at $\alpha = 0.05$ rejects.

Four Possible Outcomes

	H_0 true	H_0 false
Reject H_0	False positive (α)	True positive
Fail to reject	True negative	False negative (β)

- α = probability of a false positive (we choose this)
- β = probability of a false negative (depends on truth, sample size, effect size)
- **Power** $= 1 - \beta$ = probability of correctly rejecting a false H_0

What Affects Power?

Given a fixed α, three things drive power:

1. **Effect size.** The bigger the gap between μ_0 and the true μ, the less overlap between the null and true sampling distributions $\rightarrow$ higher power.
2. **Sample size n.** Bigger n shrinks both sampling distributions $\rightarrow$ less overlap $\rightarrow$ higher power. Power is why we care about n.
3. **Population variance σ^2.** More variance = wider sampling distributions = lower power.

You cannot reduce both α and β arbitrarily without one of the above three lever changes. That's the fundamental tradeoff.

Example: our Kelly analysis has $\beta \approx 0.32$ — a 32% false-negative rate even though we correctly rejected the null. That's a lot! If the race had been truly 51%/49% we'd have called it tied most of the time.

On the NBC decision desk we operate at $\alpha \approx 0.005$ to control false positives across ~600 race calls per election. A 5% false-positive rate would mean 30 wrong calls per election.