Hypothesis Testing: The Lady Tasting Tea

AP Statistics · Mr. Merrick · February 3, 2026

A lady claims she can tell whether milk was poured into tea first or tea into milk first. Is she truly better than random guessing, or did she just get lucky? The *parameter* we aim to study is

$$p = \text{the probability she correctly identifies a cup.}$$

"Is this result surprising under the status quo?"

- A hypothesis test starts with a status quo assumption (the null hypothesis).
- We ask: If the null were true, how likely is our data (or something even more extreme)?
- That likelihood is the p-value. Small p-value $\Rightarrow$ data is surprising under the null.

Think of H_0 as "nothing special is happening." If the data looks too unusual for H_0, we lean toward the alternative explanation.

Set up hypotheses

- Null hypothesis H_0: the status quo; what we assume for the sake of argument.
- Alternative hypothesis H_a: what we want evidence for.

Tea-tasting hypotheses (one-sided):

$$H_0 : p = 0.5 \quad \text{(guessing)}$$

$$H_a : p > 0.5 \quad \text{(better than guessing)}$$

Some Assumptions We Are Making

- Independent trials: random order; each guess doesn't change the next.
- Constant probability (under H_0): $p = 0.5$ each time if guessing.

Collect data (the experiment)

- Prepare $n = 8$ cups total. 4 cups are milk-first, and 4 cups are tea-first.
- Randomize the order of the 8 cups.
- She labels each cup as milk-first or tea-first.

Observed result: She gets $x = 7$ out of $n = 8$ correct.

Test statistic: We can use either

$$x \quad \text{or} \quad \hat{p} = \frac{x}{n} = \frac{7}{8} = 0.875.$$

Compute the p-value (binomial model)

If H_0 is true and she is guessing, then each cup is correct with probability 0.5. So the number correct follows a binomial model:

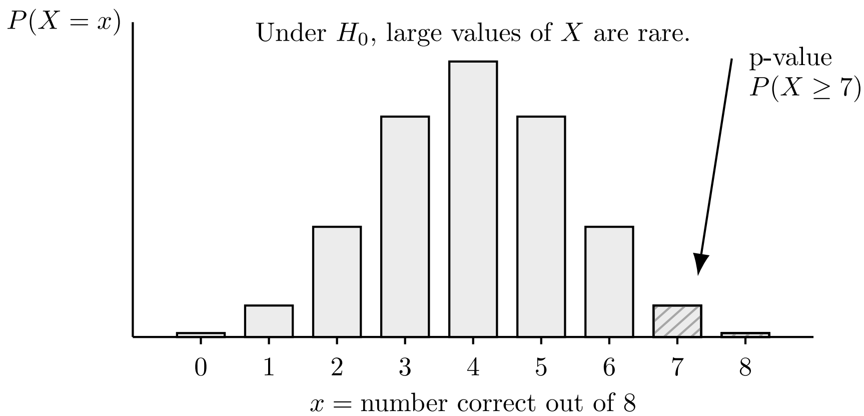
$$X \sim \text{Bin}(n = 8,\ p = 0.5) \quad \text{(under } H_0\text{)}.$$

Because $H_a : p > 0.5$, "as extreme or more extreme" means 7 or 8 correct. So the p-value is:

$$\text{p-value} = P(X \geq 7) = P(X = 7) + P(X = 8).$$

$$P(X \geq 7) = \binom{8}{7}(0.5)^8 + \binom{8}{8}(0.5)^8 = \frac{8+1}{256} = \frac{9}{256} \approx 0.0352.$$

Assuming she is only guessing ($p = 0.5$), the probability of getting 7 or more correct out of 8 is about 0.035.



Make a decision

Choose a significance level (a cutoff for "rare"):

$$\alpha = 0.05.$$

Decision rule:

$$\text{If p-value} < \alpha,\ \text{reject } H_0. \quad \text{Otherwise, fail to reject } H_0.$$

Here, p-value $\approx 0.0352 < 0.05$, so we reject H_0.

Conclusion in context

Assuming that the lady is randomly guessing ($p = 0.5$), there is a roughly 3.5% chance of observation a sample proportion of $\hat{p} = \frac{7}{8}$ or greater in future experiments. Because the p-value ≈ 0.035, which is less than $\alpha = 0.05$, we have convincing evidence that the lady can identify the pouring order better than random guessing.

What you actually need for AP Stats: one-sample z test for p

The binomial model above is a great way to understand the logic of hypothesis testing. However, on the AP Statistics exam, you are typically expected to use the *one-sample z test for a proportion* when the sample size is large.

A larger-sample example (what AP questions look like)

Suppose we repeat the tasting experiment many times and record whether she is correct each time. For example, she tries $n = 100$ cups and gets $x = 62$ correct.

Step 1 — Define the target (parameter + hypotheses + α).
Let p be the true proportion of cups the lady correctly identifies. Use $\alpha = 0.05$.

$$H_0 : p = 0.50 \qquad H_a : p > 0.50$$

Step 2 — Justify the method (test + conditions).
We will perform a one-sample z test for a proportion.

Check conditions:

- Random: the cups are presented in a randomized order across trials.
- Independence: one trial does not affect another (reasonable for repeated tastings).
- Large Counts (use $p_0 = 0.50$): $np_0 = 100(0.50) = 50 \geq 10$ and $n(1 - p_0) = 100(0.50) = 50 \geq 10$.

Since conditions are met, this test is appropriate.

Step 3 — Carry out the procedure (test statistic + p-value).

$$\hat{p} = \frac{x}{n} = \frac{62}{100} = 0.62$$

$$z = \frac{\hat{p} - p_0}{\sqrt{\frac{p_0(1 - p_0)}{n}}} = \frac{0.62 - 0.50}{\sqrt{\frac{0.50(0.50)}{100}}} = \frac{0.12}{0.05} = 2.4$$

Right-tailed p-value:

$$\text{p-value} = P(Z \geq 2.4) \approx 0.0082$$

Step 4 — Interpret the result (decision + context).
Assuming H_0 is true (she is randomly guessing$\Rightarrow p = 0.5$), there is a roughly 0.8% probability of observing $\hat{p} = 0.62$ or higher in a sample of 100 cups due to random chance alone. Since $0.0082 < 0.05$, we reject H_0. There is sufficient evidence at the $\alpha = 0.05$ level to conclude that the lady identifies cups correctly more than half the time (she is better than guessing).

A brief historical note
This scenario is inspired by a real 1920s experiment involving Muriel Bristol and the statistician Ronald Fisher. Bristol claimed she could tell whether milk or tea had been poured first, and Fisher set out to determine whether her performance could reasonably be explained by random chance.

The experimental design Fisher used was slightly different from the methods presented here, but the statistical problem was the same: assume a null hypothesis (guessing) and ask how surprising the observed result would be if that assumption were true. This idea became the foundation of modern hypothesis testing.

When the sample size is large, AP Statistics expects you to use the *one-sample z test for a proportion.*

One-sample z test for a population proportion

Test statistic (use the null value p_0 in the standard error):

$$z = \frac{\hat{p} - p_0}{\sqrt{\frac{p_0(1 - p_0)}{n}}}$$

P-value:

$$\text{p-value} = P(\text{observing a } z\text{-statistic at least as extreme as the one computed} \mid H_0)$$

When can we use the one-sample z test for p?

Before using this test, verify the following conditions:

- Random: data come from a random sample or a randomized experiment.
- Independence: if sampling without replacement, the sample size satisfies $n \leq 0.10N$.
- Large Counts (check using p_0): $np_0 \geq 10$ and $n(1 - p_0) \geq 10$.

When these conditions are met, the sampling distribution of $\hat{p}$ is approximately Normal under H_0.

Common misconceptions about hypothesis tests

Understanding the p-value

- The p-value is *not* the probability that H_0 is true.
- It is *not* the probability that the alternative hypothesis is correct.
- It *is* the probability of observing a result at least as extreme as the one obtained, assuming the null hypothesis is true.

Interpreting the decision

- Rejecting H_0 does *not* prove the alternative hypothesis. It means the data would be unusually rare if H_0 were true.
- Failing to reject H_0 does *not* mean H_0 is true. It means the data is reasonably consistent with the null model.

Final thought: Hypothesis testing is a tool for making decisions in the presence of randomness. It does not establish certainty—it tells us when the data is too surprising to attribute to chance alone.

Example 1: World of Tanks Win Rate

A student claims that they win 80% of matches in the video game World of Tanks. To investigate this claim, the outcomes of the student's most recent 200 matches are recorded. The student wins 148 of the 200 games. Is there convincing statistical evidence, at the $\alpha = 0.05$ significance level, that the student's true win rate is less than 80%?

Example 2: Medical Treatment Effectiveness

A pharmaceutical company claims that a new medication is effective for 60% of patients who take it. In a clinical trial, 300 patients receive the medication, and 171 of them experience the intended improvement. At the $\alpha = 0.05$ significance level, is there convincing statistical evidence that the true effectiveness rate of the medication differs from 60%?