Unit 8 Review: Chi-Square Inference (AP Statistics)

Goodness-of-Fit • Independence • Homogeneity

The χ^2 Distribution

Notation: $X \sim \chi^2_k$ (k = degrees of freedom).
Facts:

- Mean $E(X) = k$; Variance $\mathrm{Var}(X) = 2k$.
- Right-skewed for small k; more symmetric as k grows.
- If $Z \sim N(0,1)$ then $Z^2 \sim \chi^2_1$.
- If $X_i \sim \chi^2_{k_i}$ are independent, then $\sum X_i \sim \chi^2_{\sum k_i}$.

Use tech/tables to find p-values: $p = P(\chi^2_k \geq \chi^2_{\text{obs}})$ (always right-tail).

Core Ingredients

Expected count formula (all tests):

$$E_{ij} = \frac{(\text{row total})(\text{column total})}{\text{grand total}}$$

Test statistic:

$$\chi^2 = \sum \frac{(O-E)^2}{E}, \qquad \text{df depends on test (below).}$$

Conditions (all χ^2 tests)

- **Randomness:** Data from random sample(s) or randomized experiment.
- **Independence of observations:** Each individual contributes to exactly one cell. If sampling w/o replacement, check 10% condition.
- **All expected counts ≥ 5:** Ensures χ^2 approximation is valid.

Which Test Do I Use?

- **Goodness-of-Fit (GOF):** One categorical variable; compare sample distribution to a *claimed/model* distribution.
- **Independence:** *One* random sample; classify each individual on *two* categorical variables; ask if variables are associated.
- **Homogeneity:** *Two or more* independent random samples (or treatments); compare the *same* categorical variable's distribution across groups.

Degrees of Freedom (df)

- **GOF:** df $= k - 1$ (where k = # categories).
- **Independence/Homogeneity:** df $= (r-1)(c-1)$ (rows × columns).

4-Step Workflow (all tests)

1) State: Context; H_0 and H_a.

- GOF: H_0: distribution equals the stated model; H_a: not that model.
- Independence: H_0: variables are independent; H_a: associated.
- Homogeneity: H_0: all groups share the same distribution; H_a: at least one differs.

2) Plan/Check: Conditions as above.
3) Do: Compute E_{ij}, χ^2, df, and p-value.
4) Conclude: Compare p to α; write a *contextual* conclusion.

Interpretation Templates

P-value (AP style):
"Assuming H_0 is true, there is about a $\underline{p \times 100\%}$ chance of getting a χ^2 statistic as large as or larger than the observed value purely by random sampling variation."
Decision/Context:
If $p < \alpha$: Reject H_0; there is evidence of *(association / difference from model)*.
If $p \geq \alpha$: Fail to reject H_0; data are *consistent* with *(independence / the model / equal distributions)*.

Mini Examples (at a glance)

GOF (Mendel): Compare pea phenotypes to 3:1 model; df $= 2 - 1 = 1$.
GOF (Teddy Grahams): Up/Down vs 1:1 claim; df = 1.
Independence (Star Trek): Shirt color (3) vs Status (2); df $= (3-1)(2-1) = 2$.
Homogeneity (Sports): 3 continents ($r = 3$) × 3 sports ($c = 3$); df $= (3-1)(3-1) = 4$.

Common Pitfalls

- Using observed counts < 5 (combine categories or collect more data).
- Treating percentages as inputs—*always* use counts for χ^2.
- Forgetting that χ^2 tests are **right-tailed only**.
- Writing non-contextual conclusions (always tie back to the story).

Quick Tech Notes

Most calculators/software report χ^2, df, and p directly given the contingency table. For GOF, supply observed counts and the expected % model.

TI-84 Commands

GOF Test: 1. Enter observed counts in L1, expected counts in L2 2. STAT → TESTS → χ^2GOF-Test 3. Set df = categories −1 → Calculate

Independence/Homogeneity: 1. 2nd MATRIX → EDIT → Enter observed counts in [A] (no totals) 2. STAT → TESTS → χ^2-Test 3. Observed = [A], Expected = [B] → Calculate 4. View expected counts: 2nd MATRIX → NAMES → [B]