Unit 8: Chi-Square Inference

Merrick Fanning

August 8, 2025

Unit 8 Outline: Inference Using the Chi-Square Distribution

1. The Chi-Square Distribution
2. Chi-Square Goodness of Fit Test
    1. Mendelian Genetics
    2. Teddy Grahams
3. Chi-Square Test of Homogeneity
4. Chi-Square Test of Independence

The χ^2 Distribution

Definition: χ^2_k denotes a χ^2 distribution with k **degrees of freedom**.

- Mean: $E(X) = k$
- Variance: $Var(X) = 2k$
- The shape is **right-skewed**, especially for small k
- Becomes more symmetric (normal) as k increases
- A standard normal variable squared follows a χ^2 distribution with 1 degree of freedom: If $Z \sim N(0, 1)$, then $Z^2 \sim \chi^2_1$.
- The sum of independent χ^2 variables has degrees of freedom equal to the sum of their degrees of freedom:

$$X_i \sim \chi^2_{k_i} \quad \Rightarrow \quad \sum X_i \sim \chi^2_{\sum k_i}$$

Like with the normal distribution, we compute probabilities using technology, tables, or calculators.

Chi-Square Testing in Genetics

- Living organisms inherit **traits** from their parents.
- Traits are determined by **genes**, which are segments of DNA.
- Different **versions of a gene** are called **alleles**.
- When two parents pass down alleles, they form a **genotype**, which determines the observable **phenotype**.

- **Gene:** A segment of DNA that encodes a trait (e.g., eye color).
- **Allele:** A specific version of a gene (e.g., blue or brown).
- **Genotype:** The combination of two alleles inherited from parents.
- **Phenotype:** The observable trait expressed by the genotype.

Punnet Squares: Monohybrid, Dihybrid, and Heterozygotes

- **Monohybrid Cross:** One gene, two alleles.
  - Example: $Pp \times Pp$ (P = purple, p = white)

	P	p
P	PP	Pp
p	Pp	pp

Genotypes: 1 PP, 2 Pp, 1 pp $\Rightarrow$ Phenotypes: 3 purple : 1 white

- **Dihybrid Cross:** Two genes, each with two alleles.
  - Example: $PpTt \times PpTt$ (P = purple, p = white, T = tall, t = short)

	PT	Pt	pT	pt
PT	PPTT	PPTt	PpTT	PpTt
Pt	PPTt	PPtt	PpTt	Pptt
pT	PpTT	PpTt	ppTT	ppTt
pt	PpTt	Pptt	ppTt	pptt

Phenotypes: 9 purple-tall : 3 purple-short : 3 white-tall : 1 white-short

History: Mendel and the Chi-Square Test

- **Gregor Mendel** (1822–1884) — Austrian monk and scientist.
  - Conducted experiments on pea plants.
  - Discovered predictable **inheritance patterns**.
  - Famous ratios: 3:1 (monohybrid), 9:3:3:1 (dihybrid).
- **Karl Pearson** (1857–1936) — English mathematician and statistician.
  - Developed the **chi-square test** in 1900.
  - Purpose: Compare observed data to an expected theoretical model.
- Mendel did **not** use the chi-square test — it was applied to his data later to verify his results.

Mendel's Pea Plant Data: Seed Shape

Monohybrid cross: Round (R) vs Wrinkled (r) seeds

- Cross: $Rr \times Rr$
- Expected phenotype ratio: 3 round : 1 wrinkled

Mendel's observed counts:

Phenotype	Observed (O)	Expected (E)
Round	5474	$7324 \times \frac{3}{4} = 5493$
Wrinkled	1850	$7324 \times \frac{1}{4} = 1831$

Chi-square calculation:

$$\chi^2 = \frac{(5474 - 5493)^2}{5493} + \frac{(1850 - 1831)^2}{1831}$$

$$\chi^2 \approx 0.066 + 0.197 = 0.263$$

Conclusion: With df $= 1$, $\chi^2 = 0.263$ is far below 3.841, so the data is **consistent** with the 3:1 ratio.

Conditions for the Chi-Square Goodness-of-Fit Test

- **1. Random Sampling**
  - Data should come from a random sample or a randomized experiment.
  - Ensures results can be generalized to the population.
- **2. All Expected Counts** ≥ 5
  - Expected count in each category should be at least 5.
  - Prevents large sampling variability that would make the χ^2 approximation inaccurate.
- **3. Independent Observations**
  - Each individual belongs to exactly one category.
  - One observation does not influence another.
  - For sampling without replacement, population size should be at least 10 times the sample size.

Why These Matter

Meeting these conditions ensures that the **sampling distribution of** χ^2 follows the chi-square model closely, making *p*-values and conclusions reliable.

Example: Teddy Grahams and Arm Positions

Imagine you open a box of Teddy Grahams and notice some bears have **arms up** while others have **arms down**.
According to the manufacturer, these two arm positions occur in a **1:1 ratio**.
To test this claim, you:

- Randomly select $n = 100$ Teddy Grahams from the box.
- Count:
  - Arms up: 54
  - Arms down: 46

Question: Is there evidence, at the $\alpha = 0.05$ level, that the arm positions are *not* in a 1:1 ratio?

1) State the Problem & Information

We want to test the manufacturer's claim that Teddy Graham arm positions occur in a **1:1 ratio**.

- Sample size: $n = 100$
- Observed counts: Arms up $= 54$, Arms down $= 46$
- Significance level: $\alpha = 0.05$

Hypotheses (Goodness-of-Fit):

$$H_0: \text{ The distribution is as claimed (1:1 ratio)} \iff p_{\text{up}} = p_{\text{down}} = 0.5$$

$$H_a: \text{ The distribution is not 1:1 } (H_0 \text{ false})$$

Question: Is there evidence, at $\alpha = 0.05$, that the arm-position distribution differs from 1:1?

2) Check Conditions

- **Random sampling:** Cookies selected at random from the box. ✓
- **Expected counts** (under 1:1):

$$E_{\text{up}} = 100 \cdot \frac{1}{2} = 50, \quad E_{\text{down}} = 100 \cdot \frac{1}{2} = 50 \ \ (\geq 5) \checkmark$$

- **Independence:** The sample of cookies is clearly less than 10% of the population of all cookies ✓

3) Compute Test Statistic and p-Value

$$\chi^2 = \sum \frac{(O-E)^2}{E}$$

Category	O	E	$(O-E)^2/E$
Arms up	54	50	$\frac{(4)^2}{50} = 0.32$
Arms down	46	50	$\frac{(-4)^2}{50} = 0.32$
		Total χ^2	$0.32 + 0.32 = \mathbf{0.64}$

Degrees of freedom: $df = k - 1 = 2 - 1 = 1$.

p-value: $P(\chi^2_{(1)} \geq 0.64) \approx \mathbf{0.424}$.

4) Conclusion (with Interpretation)

Since $p = 0.424 > \alpha = 0.05$, we **fail to reject** H_0.

Interpretation: Assuming the true distribution of Teddy Graham arm positions is 1:1, there is about a **42.4%** chance of observing a chi-square test statistic *as large as or larger* than the one we found in this sample, purely by random chance in future samples.

This is not strong (condemning) evidence against H_0, so the sample data are consistent with the manufacturer's 1:1 claim.

University Application: Fruit Fly Genetics

In many introductory biology labs, students test Mendelian inheritance using *Drosophila melanogaster* (fruit flies).

Typical procedure:

1. **Select parent flies** with known traits (e.g., red-eyed female $\times$ white-eyed male).
2. **Set up a mating vial** with a small group of flies and nutrient medium.
3. **Allow mating and egg-laying**, then remove parents to avoid confusion.
4. **Wait $\sim$10 days** for eggs $\rightarrow$ larvae $\rightarrow$ pupae $\rightarrow$ adult flies.
5. **Count offspring phenotypes** under a microscope (e.g., red eyes vs white eyes, normal wings vs vestigial wings).
6. **Compare observed counts to Mendelian ratios** (e.g., 3:1 or 9:3:3:1) using a chi-square goodness-of-fit test.

Recap: The "Redshirt" Myth

- In **Star Trek** fandom, there's a long-running joke:

  "If you beam down to a planet wearing a red shirt, you probably won't make it back."
- The idea comes from many episodes where security and engineering crew (red shirts) meet unfortunate ends on away missions.
- But is it **really** true in the original series? Or just selective memory?
- Using crew roster and status data from the original series (1966–1969), we can test if survival **independent** of shirt color?

Chi-Square Test of Independence: Star Trek Crew Survival

Question: Is **shirt color** independent of **status** (alive/dead) among Enterprise crew members?
Source: Matthew Barsalou, "Keep Your Redshirt On: A Bayesian Exploration," *Significance Magazine*. Data compiled from **Memory Alpha** (fan-curated Star Trek wiki). Article link

	Alive	Dead	Total
Blue	129	7	136
Yellow	46	9	55
Red	215	24	239
Total	390	40	430

Step 1: Hypotheses & Important Information

Null hypothesis (H_0): Shirt color and crew status are **independent** (no association).
Alternative hypothesis (H_A): Shirt color and crew status are **not independent** (associated).

Test: Chi-square test of independence on a 3×2 table.
Significance level: $\alpha = 0.05$.

Step 2: Check Conditions

- **Randomness:** Assume the Enterprise crew is a random sample from all Starfleet personnel.
- **10% Condition:** $n = 430 < 0.10 \times N_{\text{Starfleet}}$, so sampling without replacement is fine.
- **Expected Counts:** All $E_{ij} \geq 5$ (verified on next slide).

Step 3: Expected Counts & Test Statistic

Expected counts: $E_{ij} = \frac{\text{(row total)(col total)}}{\text{grand total}}$

Example: $E_{\text{Blue, Alive}} = \frac{136 \times 390}{430} \approx 123.3$ $E_{\text{Blue, Dead}} = 136 - 123.3 = 12.7$

	Alive (E)	Dead (E)
Blue	123.3	12.7
Yellow	49.9	5.1
Red	216.8	22.2

$$\chi^2 - \text{stat} \approx \frac{(129 - 123.3)^2}{123.3} + \frac{(7 - 12.7)^2}{12.7} + \cdots + \frac{(24 - 22.2)^2}{22.2} \approx 6.61$$

Degrees of freedom: $(3-1)(2-1) = 2$, $P(\chi^2_{(2)} \geq 6.61) \approx 0.037$.

Step 4: Conclusion (with Interpretation)

Since $p \approx 0.037 < \alpha = 0.05$, we **reject** H_0.

Interpretation (AP style): Assuming shirt color and crew status are truly independent, there is about a **3.7%** chance of observing a chi-square statistic as large as or larger than 6.61 purely by random variation in future rosters. This is sufficiently unlikely, so we conclude there is evidence of an **association**: shirt color appears linked to survival among Enterprise crew.

Where Will You See Chi-Square Tests of Independence?

- **Biology/Genetics:** Testing whether *eye color* is associated with *gene variant* in a sample of fruit flies.
- **Medicine:** Determining if *treatment type* is associated with *recovery rate* in a randomized controlled trial.
- **Public Health:** Investigating if *smoking status* is associated with *lung disease prevalence* in a community health survey.
- **Business/Marketing:** Exploring whether *purchase preference* is related to *age group* in a consumer sample.
- **Political Science:** Examining if *voting preference* is associated with *education level* in a poll.

Key idea: Chi-square independence tests appear wherever you have *two categorical variables* and want to know if they are related.

Chi-Square Test of Homogeneity: Sports Preference by Continent

Scenario: A sports marketing firm surveys random samples of adults from three continents to learn about their favorite sport to watch. Each participant picks one of: **Soccer**, **Basketball**, or **Cricket**.

	Soccer	Basketball	Cricket	Total
Europe	45	35	20	100
Asia	30	25	45	100
North Am.	25	60	15	100
Total	100	120	80	300

Question: Is the distribution of favorite sports the same across continents?

Chi-Square Test of Homogeneity: Sports Preference by Continent

Scenario: A sports marketing firm surveys random samples of adults from three continents to learn about their favorite sport to watch. Each participant picks one of: **Soccer**, **Basketball**, or **Cricket**.

	Soccer	Basketball	Cricket	Total
Europe	45	35	20	100
Asia	30	25	45	100
North Am.	25	60	15	100
Total	100	120	80	300

Question: Is the distribution of favorite sports the same across continents?

Step 1: Hypotheses and Important Information

Null hypothesis (H_0): The distribution of favorite sports is the **same** for all continents.
Alternative hypothesis (H_A): The distribution of favorite sports **differs** for at least one continent.

Test: Chi-square test of homogeneity on a 3×3 table.
Significance level: $\alpha = 0.05$.

Step 2: Check Conditions

- **Randomness:** Each continent's sample is a simple random sample of adults from that continent.
- **10% Condition:** Each sample size $n = 100$ is less than 10% of the adult population on that continent.
- **Expected Counts:** All expected cell counts $E_{ij} \geq 5$ (verified on next slide).

Step 3: Expected Counts & Test Statistic

Expected counts: $E_{ij} = \frac{\text{(row total)(col total)}}{\text{grand total}}$

Example: $E_{\text{Europe, Soccer}} = \frac{100 \times 100}{300} = 33.33$

	Soccer (E)	Basketball (E)	Cricket (E)
Europe	33.33	40	26.67
Asia	33.33	40	26.67
North Am.	33.33	40	26.67

$$\chi^2 = \sum \frac{(O-E)^2}{E} \approx 39.41$$

Degrees of freedom: $(3-1)(3-1) = 4$

$p = P(\chi^2_{(4)} \geq 39.41) \approx 2.5 \times 10^{-8}$

Step 4: Conclusion with Interpretation

Since $p \approx 2.5 \times 10^{-8} < \alpha = 0.05$, we **reject** H_0.

Interpretation (AP style): Assuming the distribution of favorite sports is truly the same across continents, there is an approximately 0.0000025% chance of observing a chi-square statistic as large as 39.41 purely by random variation in future samples. This is extremely unlikely, so we conclude there is strong evidence that sports preferences differ by continent.

Chi-Square Independence vs. Homogeneity

Chi-Square Test of Independence	Chi-Square Test of Homogeneity
Purpose: Determine if two categorical variables are associated in *one population*.	**Purpose:** Compare the distribution of a categorical variable across *two or more populations*.
Data collection: One random sample, each individual classified on both variables.	**Data collection:** Two or more independent random samples, each classified on the same categorical variable.
Star Trek Example: Sample = Enterprise crew; Variables = *Shirt color & Status (alive/dead)*.	**Sports Example:** Samples = 100 adults from each of 3 continents; Variable = *Favorite sport*.
Key question: Is there an association between the variables?	**Key question:** Do all populations have the same distribution of the variable?

Tip to Discriminate: Ask: "Was the data from *one sample classified twice* (→ independence) or *multiple samples classified once* (→ homogeneity)?"

Practice: Which Chi-Square Test?

For each scenario, decide if you would use a **Chi-Square Goodness-of-Fit**, **Chi-Square Independence**, or **Chi-Square Homogeneity** test.

1. A poll of 500 city residents asks for their favorite type of cuisine (Italian, Chinese, Mexican). You want to see if the distribution matches last year's city report.
2. One random sample of university students is classified by *year in school* (Freshman, Sophomore, etc.) and *whether they have a campus meal plan* (Yes/No).
3. Separate random samples from three grocery stores record whether each shopper purchased produce, meat, or bakery items.
4. A marketing survey asks 1,000 randomly selected consumers to pick their favorite soda brand. You want to test if the proportions match the company's claimed distribution.
5. Random samples of patients from four hospitals record whether their treatment outcome was "Full Recovery," "Partial Recovery," or "No Recovery."