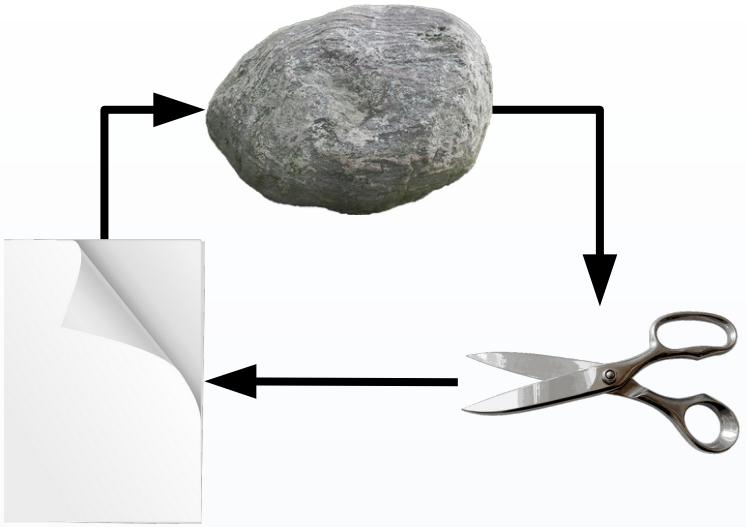
EGM101 – Skills Toolbox

Week 8, Part 6: The Chi-Square Distribution

What about categorical data?

- How would we determine whether players are choosing randomly or not?
- All we have are categories:
  - No ranking
  - No mean/median/variance to calculate
- Additional problem: testing each category one-by-one increases risk of Type I error
- Solution: use a Chi-Square test



rock	paper	scissors
37	26	27

The Chi-square (χ^2) Statistic

- Comparing observed (O) to expected (E) frequencies
- Calculate E using sample size n, expected relative frequency/probability
- Notes:
  - If no difference, $\chi^2 = 0$
  - $\chi^2 \geq 0$
  - Large differences are more heavily weighted (cf. variance, standard deviation)
  - General assumption: each $E_i \geq 5$

	rock	paper	scissors
observed	37	26	27
expected	30	30	30

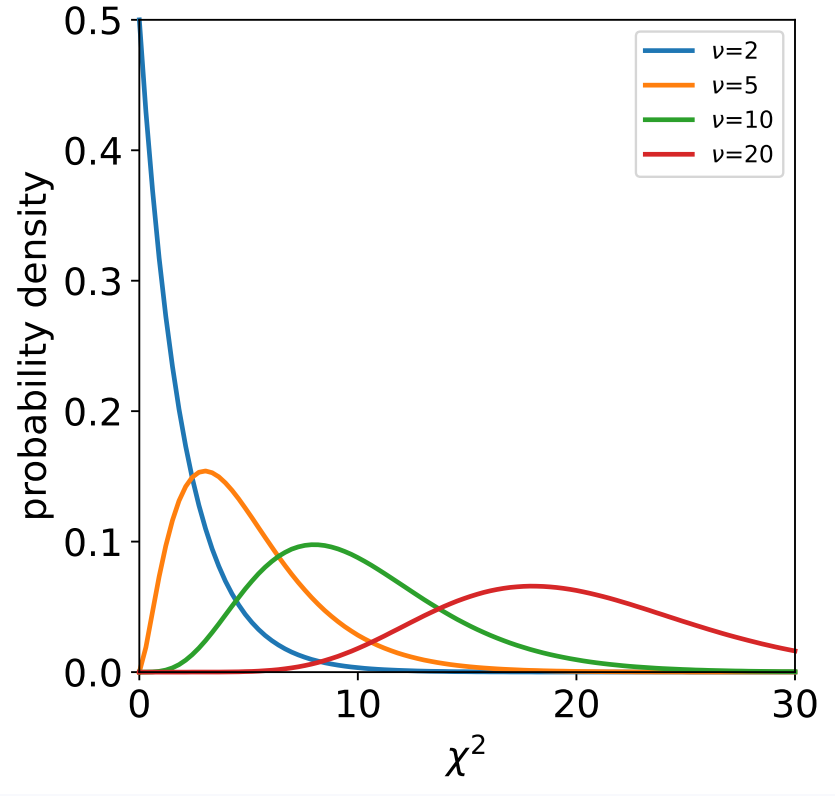
$$\chi^2 = \sum_{i=1}^{c} \frac{(O_i - E_i)^2}{E_i}$$

$$= \frac{(37-30)^2}{30} + \frac{(26-30)^2}{30} + \frac{(27-30)^2}{30}$$

$$= \frac{49+16+9}{30} = 2.467$$

The Chi-Square (χ^2) Distribution

- Right-skewed distribution
- Shape of curve depends on degrees of freedom, $\nu = c - 1$
  - Note: this is the number of categories, not the number of observations!
  - $\mu = \nu$
  - $\sigma = \sqrt{2\nu}$
- As ν increases, peak spreads out, becomes more symmetrical

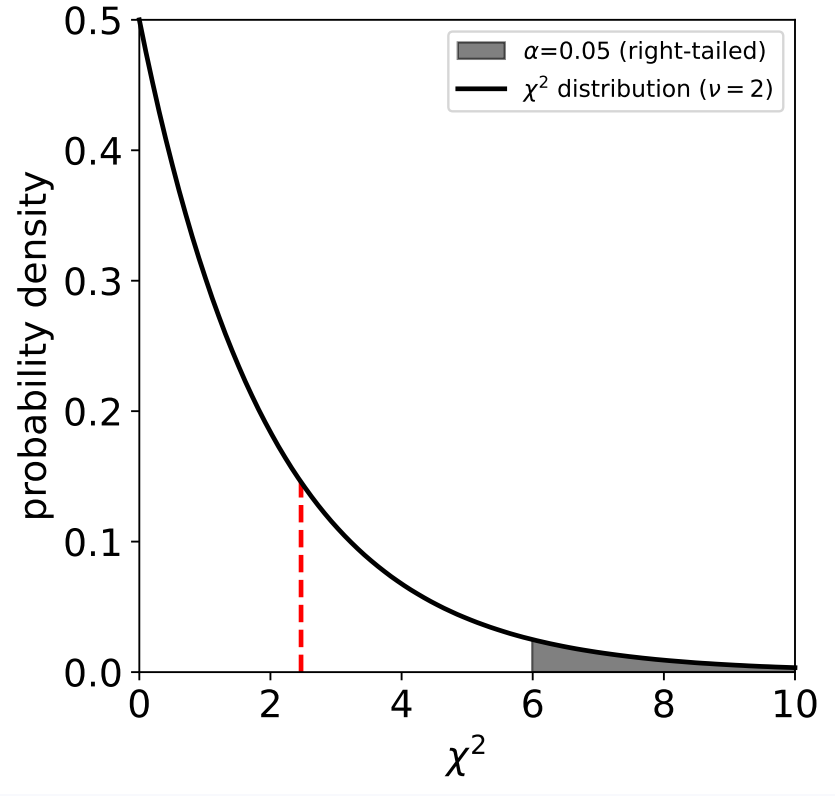


Applications of the Chi-square Distribution

- (Pearson's) Goodness-of-fit test
  - Do observations match a particular frequency distribution?
- Test of independence
  - Is there a relationship between different factors?
- Test of homogeneity
  - Are different populations distributed the same?
- Test of a single variance
  - Does the population have the given variance?

Goodness-of-fit test

- Hypotheses:
  - H_0: The players' choice of rock/paper/scissors is random
  - H_a: The players' choice of rock/paper/scissors is not random
- Note: goodness-of-fit test is always right-tailed
- From before, $\chi^2 = 2.467$
- For $\nu = c - 1 = 2$, $\alpha = 0.05$, critical value is 5.991
- Because $\chi^2 < \chi_c^2$, we **do not reject** H_0



Contingency tables and χ^2

- Recall: contingency table (W7, P3):
  - Group outcomes/frequency based on multiple variables
  - Help to determine conditional probabilities
- Example: pet species and fish preference
- Expected frequency $E = RC / n$, where:
  - R: row total
  - C: column total
  - n: sample size

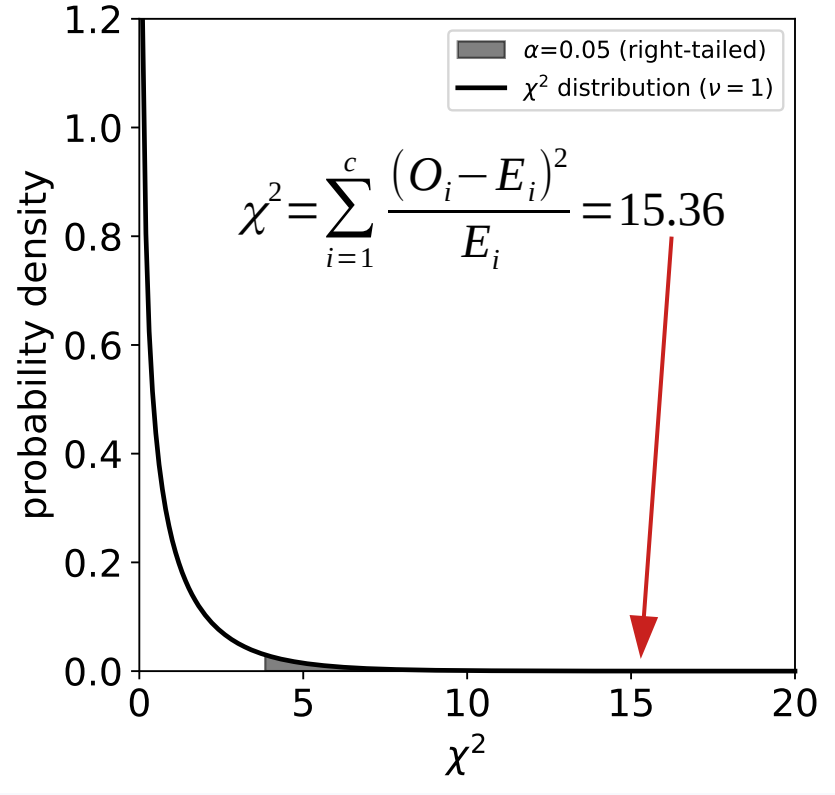
	Salmon	Tuna	Total
Dog	39	10	49
Cat	21	30	51
Total	60	40	100

	Salmon	Tuna
Dog	29.4	19.6
Cat	30.6	20.4

$$E_{\text{dog, salmon}} = \frac{R_{\text{dog}} * C_{\text{salmon}}}{n_{\text{total}}} = \frac{49 * 60}{100} = 29.4$$

Independence Test

- Hypotheses:
  - H_0: the two variables (species, fish preference) are independent
  - H_a: the two variables are associated
- $v = (r - 1)(c - 1) = 1$
  - For $\alpha = 0.05$, $\chi_c^2 = 3.84$
- Because $\chi^2 > \chi_c^2$, we reject H_0
- Note:
  - Test is right-tailed
  - Requires "sufficient" E in each cell (typically at least 5)



Test for homogeneity

- Tests whether two populations have the same (unknown) distribution
- Hypotheses:
  - H_0: the distributions of the two populations are the same
  - H_a: the distributions of the two populations are not the same
- Works the same as the goodness-of-fit test:
  - Calculate χ^2 statistic using the two distributions instead of observed, expected
  - $v = c - 1$ (number of categories)
  - Calculate χ_c^2 for v, α
  - If $\chi^2 > \chi_c^2$, reject null hypothesis (right-tailed test)

Test of a single variance

- Assumes that the population is normally distributed
  - A sneaky parametric test
- Tests whether or not population variance σ^2 equals some value, σ_0^2
  - cf. tests of the population mean
- Hypotheses:
  - H_0: $\sigma^2 = \sigma_0^2$
  - H_a: $\sigma^2 \neq \sigma_0^2$
- Can also be left-tailed ($\sigma^2 < \sigma_0^2$) or right-tailed ($\sigma^2 > \sigma_0^2$)
- $\nu = n - 1$ degrees of freedom

$$\chi^2 = \frac{(n-1)s^2}{\sigma_0^2}$$

Summary

- Chi-square tests help compare observed, expected *distributions*
  - No assumptions about distributions: non-parametric test
- Multiple tests with different applications
- Can be used with categorical as well as quantitative data

Additional Resources

- Illowsky and Dean, Chapter 11
- Caswell, Chapter 15.5
- Weiss, Chapter 12
- Chi-square Test of Independence [Jim Frost]
- Chi-square distribution introduction [Khan Academy]
- Pearson's chi-square test [Khan Academy]