
# Spring 2021
# SKKU Biostats and Big data

CHOONG-WAN WOO | COCOAN lab | http://cocoanlab.github.io

# Lecture 21
# Comparing counts (Chi-square test)

CHOONG-WAN WOO | COCOAN lab | http://cocoanlab.github.io

# Review: Key Points

## Chapter 25: Paired Samples and Blocks

- **Paired *t*-test:** use pairwise differences, and then one-sample $t$-test on the pairwise differences
  - $H_0$: $\mu_d = \Delta_0$ (usually, $\Delta_0 = 0$), $t_{n-1} = \frac{\bar{d}-\Delta_0}{SE(\bar{d})}, df = n-1, SE(\bar{d}) = \frac{s_d}{\sqrt{n}}$
- **Confidence interval:** $\bar{d} \pm t^*_{n-1} \times SE(\bar{d})$
- **Nonparametric sign test:**
  - Record 0 for the pairs with negative differences, record 1 for the pairs with positive differences
  - and ignore the pairs with difference = 0
  - Then, test the associated proportion $p = 0.5$ using a $z$-test

# Counts

| Births | Sign |
|---|---|
| 23 | Aries |
| 20 | Taurus |
| 18 | Gemini |
| 23 | Cancer |
| 20 | Leo |
| 19 | Virgo |
| 18 | Libra |
| 21 | Scorpio |
| 19 | Sagittarius |
| 22 | Capricorn |
| 24 | Aquarius |
| 29 | Pisces |

Birth totals by sign for 256 Fortune 400 executives.

Example: zodiac signs of 256 heads of the largest 400 companies

- If the zodiac signs cannot predict the future, we should expect 1/12 counts for each category.
- How closely do the observed numbers of births per sign fit this simple "null" model?
- "Goodness-of-fit" test

# Goodness-of-fit tests

- Procedure:
  - First, observed value minus expected value for each cell: similar to residuals
  - The residual values can be positive and negative, so we need to square them.
  - We divide the residuals by the expected counts.
  - $\sum \frac{(Obs - Exp)^2}{Exp}$
  - How well the theory (expected values) fits the data: **goodness-of-fit**

- It follows the chi-square ($\chi^2$) distribution.
  - $\chi^2 = \sum \frac{(Obs - Exp)^2}{Exp}$
  - This family of models also depends on the degrees of freedom.
  - In the chi-square test, $df = n - 1$, where $n$ is the number of categories, not the sample size.

# Quiz 21-1

https://forms.gle/G7iaMao5mBW9Cqb1A

# Assumptions and Conditions

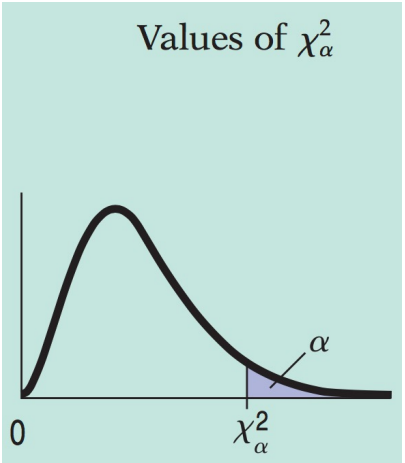
- Counted Data Condition
  - The data must be **counts** for the categories of a categorical variable.

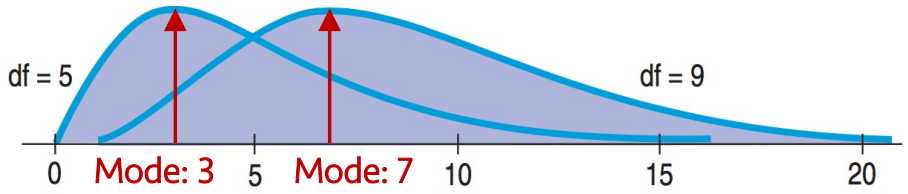
- Independence Assumption
  - The counts in the cells should be independent of each other.

- Sample Size Assumption
  - Expected cell frequency condition:
    - The expected counts for each cell should be at least 5.

# Chi-Square P-values

- The chi-square should be used *only* for testing hypotheses, *not* for constructing confidence intervals.
- We can do only *one-sided* test (by squaring the differences, we made all the deviations positive).
- There's *no* direction to the rejection of the null model. All we know is that it doesn't fit.
- It is testing all of the cells together. There are many ways the null hypothesis can be wrong (*many-sided* in some sense)
- Chi-square models are skewed.
  - The mode is at $\chi^2 = df - 2$, and its mean is at $df$.



# The example of zodiac sign

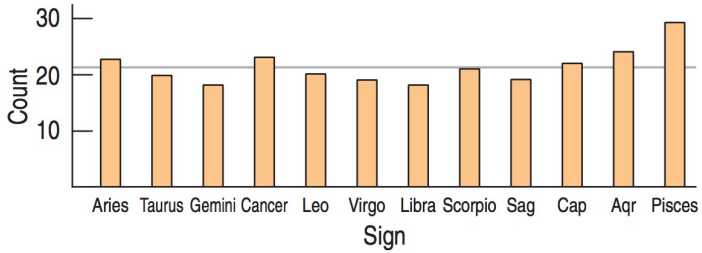
| Births | Sign |
|---|---|
| 23 | Aries |
| 20 | Taurus |
| 18 | Gemini |
| 23 | Cancer |
| 20 | Leo |
| 19 | Virgo |
| 18 | Libra |
| 21 | Scorpio |
| 19 | Sagittarius |
| 22 | Capricorn |
| 24 | Aquarius |
| 29 | Pisces |

Birth totals by sign for 256 Fortune 400 executives.

$H_0$: Births are uniformly distributed over zodiac signs.[2]

$H_A$: Births are not uniformly distributed over zodiac signs.

The conditions are satisfied, so I'll use a $\chi^2$ model with $12 - 1 = 11$ degrees of freedom and do a **chi-square goodness-of-fit test.**




The bar chart shows some variation from sign to sign, and Pisces is the most frequent. But it is hard to tell whether the variation is more than I'd expect from random variation.

- ✔ **Counted Data Condition:** I have counts of the number of executives in 12 categories.
- ✔ **Independence Assumption:** The birth dates of executives should be independent of each other.
- ✔ **Randomization Condition:** This is a convenience sample of executives, but there's no reason to suspect bias.
- ✔ **Expected Cell Frequency Condition:** The null hypothesis expects that 1/12 of the 256 births, or 21.333, should occur in each sign. These expected values are all at least 5, so the condition is satisfied.

# The example of zodiac sign

| Births | Sign |
|---|---|
| 23 | Aries |
| 20 | Taurus |
| 18 | Gemini |
| 23 | Cancer |
| 20 | Leo |
| 19 | Virgo |
| 18 | Libra |
| 21 | Scorpio |
| 19 | Sagittarius |
| 22 | Capricorn |
| 24 | Aquarius |
| 29 | Pisces |

Birth totals by sign for 256 Fortune 400 executives.

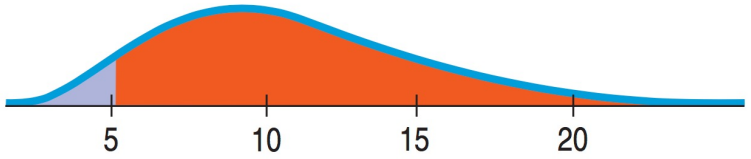
$H_0$: Births are uniformly distributed over zodiac signs.[2]

$H_A$: Births are not uniformly distributed over zodiac signs.

The expected value for each zodiac sign is 21.333.

$$\chi^2 = \sum \frac{(Obs - Exp)^2}{Exp} = \frac{(23 - 21.333)^2}{21.333} + \frac{(20 - 21.333)^2}{21.333} + \ldots$$

$$= 5.094 \text{ for all 12 signs.}$$




$$P\text{-value} = P(\chi^2 > 5.094) = 0.926$$

The conditions are satisfied, so I'll use a $\chi^2$ model with $12 - 1 = 11$ degrees of freedom and do a **chi-square goodness-of-fit test.**

# Trouble with Goodness-of-fit tests

- Goodness-of-fit: **How well does the theory fit the data?**
- **The only null hypothesis available** ($H_0$: the theory is true)
  - We can only reject or fail to reject the null hypothesis.
  - We can never confirm the theory is true.
- **It is also difficult to know what is the alternative.**
  - The theory can be wrong in many ways.
- **Thus, there is no way to prove that a favored model is true, with goodness-of-fit tests.**
  - Alternative: model comparison

# Chi-square test of Homogeneity

- Testing whether the proportions are same across multiple groups
- two-way table

Post-graduation activities of the class of 2006 for several colleges of a large university

| | Agriculture | Arts & Sciences | Engineering | Social Science | Total |
|---|---|---|---|---|---|
| Employed | 379 | 305 | 243 | 125 | **1052** |
| Grad School | 186 | 238 | 202 | 96 | **722** |
| Other | 104 | 123 | 37 | 58 | **322** |
| **Total** | **669** | **666** | **482** | **279** | **2096** |

Percentage →

| | Agriculture | Arts & Sciences | Engineering | Social Science | Total |
|---|---|---|---|---|---|
| Employed | 56.7% | 45.8% | 50.4% | 44.8% | **50.2** |
| Grad School | 27.8 | 35.7 | 41.9 | 34.4 | **34.4** |
| Other | 15.5 | 18.5 | 7.7 | 20.8 | **15.4** |
| **Total** | **100** | **100** | **100** | **100** | **100** |

Expected values for the '06 graduates

| | Agriculture | Arts & Sciences | Engineering | Social Science | Total |
|---|---|---|---|---|---|
| Employed | 335.777 | 334.271 | 241.920 | 140.032 | **1052** |
| Grad School | 230.448 | 229.414 | 166.032 | 96.106 | **722** |
| Other | 102.776 | 102.315 | 74.048 | 42.862 | **322** |
| **Total** | **669** | **666** | **482** | **279** | **2096** |

# Chi-square test of Homogeneity

- $\chi^2 = \sum \frac{(Obs - Exp)^2}{Exp}$
- The example of the agriculture school $\frac{(Obs - Exp)^2}{Exp} = \frac{(379 - 335.777)^2}{335.777} = 5.564$

- And summing these across all the schools,
  - $\chi^2 = \sum \frac{(Obs - Exp)^2}{Exp} = 54.51$

- Degrees of freedom:
  - $(R-1)(C-1)$, where R: the number of rows, C: the number of columns

# Examining the Residuals

- Which cell? How far from the expected values?

- **Standardized residual:**
  - $c = \frac{(Obs - Exp)}{\sqrt{Exp}}$
  - square roots of the components we calculated for each cell
  - Their sign indicates whether we observed more or fewer cases than we expected.

| | Ag | A&S | Eng | Soc Sci |
|---|---|---|---|---|
| Employed | 2.359 | −1.601 | 0.069 | −1.270 |
| Grad School | −2.928 | 0.567 | 2.791 | −0.011 |
| Other | 0.121 | 2.045 | −4.305 | 2.312 |

# Chi-square test of independence

Race effects on police vehicle search

| | | Race | | | |
|---|---|---|---|---|---|
| | | Black | White | Other | Total |
| Search | No | 787 | 594 | 27 | **1408** |
| | Yes | 813 | 293 | 19 | **1125** |
| | **Total** | **1600** | **887** | **46** | **2533** |

- Are police search and race independent? or have relationship?
- Contingency table
- From L09:
  - "**Independence:** the occurrence of A does not change the probability of B, $P(\mathbf{B}|\mathbf{A}) = P(\mathbf{B})$"

- The calculation is identical to the homogeneity test.

- What's different?
  - Independence test: Two categorical variables measured on a single population
    - Homogeneity test: a single categorical variable independently measured on two or more populations
  - Independence test's question: "Are the variables independent?"
    - Homogeneity test: "Are the groups homogeneous?"

# Quiz 21-2

https://forms.gle/ibrzmieShNgGdK1X7

# Review: Key Points

## Chapter 26: Comparing Counts

- Goodness-of-fit tests: $\chi^2 = \sum \frac{(Obs - Exp)^2}{Exp}$
- Assumption and conditions:
  - counted data condition, independence assumption, expected cell frequency condition
- Chi-square distribution: only positive, right skewed, mode: df-2, mean: df
- Chi-square test for a one-way count table
- Chi-square test for a two-way table: Chi-square test of homogeneity
- Chi-square test for a contingency table: Chi-square test of independence