# Math 58 / 58B - Introduction to (Bio)Statistics

*don't turn in!!*

*not ever due*

## Homework 7

**Assignment Summary (Goals)**

- chi-square goodness of fit test
- chi-square tests of independence

1. **True or false, Part I**[1] Determine if the statements below are true or false. For each false statement, suggest an alternative wording to make it a true statement.

(a) The chi-square distribution, just like the normal distribution, has two parameters, mean and standard deviation.
(b) The chi-square distribution is always right skewed, regardless of the value of the degrees of freedom parameter.
(c) The chi-square statistic is always positive.
(d) As the degrees of freedom increases, the shape of the chi-square distribution becomes more skewed.

2. **True or false, Part II**[2]

Determine if the statements below are true or false. For each false statement, suggest an alternative wording to make it a true statement.

(a) As the degrees of freedom increases, the mean of the chi-square distribution increases.

(b) If you found $X^2 = 10$ with df $= 5$ you would fail to reject $H_0$ at the 5% significance level.

(c) When finding the p-value of a chi-square test, we always shade the tail areas in both tails.

3. **Jurors**[3]

Consider data from a random sample of 275 jurors in a small county. Jurors identified their racial group, and we would like to determine if these jurors are racially representative of the population. If the jury is representative of the population, then the proportions in the sample should roughly reflect the population of eligible jurors, i.e. registered voters.

| Race | White | Black | Hispanic | Other | Total |
|---|---|---|---|---|---|
| Representation in juries | 205 | 26 | 25 | 19 | 275 |
| Registered voters | 0.72 | 0.07 | 0.12 | 0.09 | 1.00 |

(a) What are the null and alternative hypothesis for the question dealing with these data?

(b) Of the people in the city, 275 served on a jury. If the individuals are randomly selected to serve on a jury, about how many of the 275 people would we expect to be white? How many would we expect to be black?

(c) Calculate the relevant test statistic by hand (use R as a calculator).

(d) Use `xpqchisq()` to fin the p-value.

[1] From OpenIntro Statistics, exercise 3.37
[2] From OpenIntro Statistics, exercise 3.38
[3] From ISRS, Example 3.18

(e) Conclude the study in terms of the problem.

4. **US Volunteerism I**[4] The 2003 study on volunteerism conducted by the Bureau of Labor Statistics reported the sample percentages who performed volunteer work, broken down by many other variables. For example, respondents were categorized by age. The following reports the percentage of sample respondents in each age group who had performed volunteer work in the previous year:

| Age group | 16–24 years | 25–34 years | 35–44 years | 45–54 years | 55–64 years | 65 or more |
|---|---|---|---|---|---|---|
| % volunteer | 21.9% | 24.8% | 34.1% | 31.3% | 27.5% | 22.7% |

(a) Is this information sufficient to construct a segmented bar graph for comparing the proportions of volunteers across the various age categories? If so, do so, and comment on what the graph reveals. If not, explain.

(b) Explain why this information is not sufficient to conduct a chi-square test of whether these sample proportions differ significantly across the age categories.

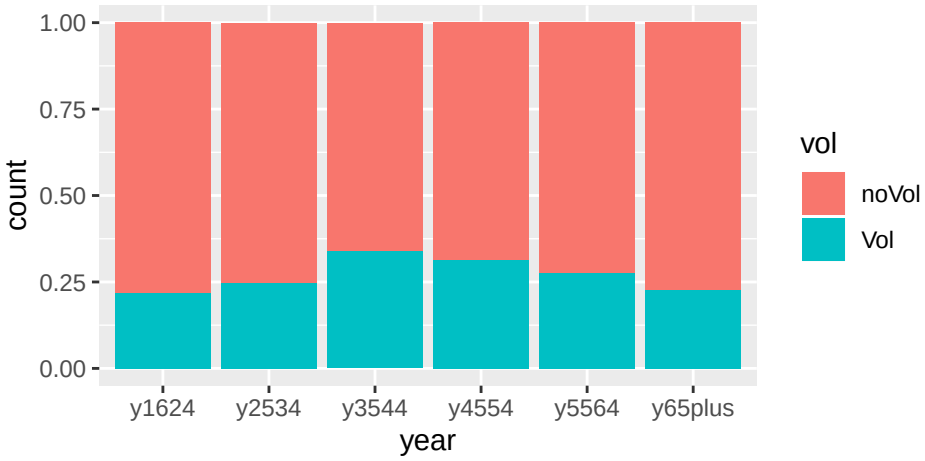
The sample sizes in each age group are not given in the report, but based on other information we can estimate them to be as follows:

| Age group | 16–24 years | 25–34 years | 35–44 years | 45–54 years | 55–64 years | 65 or more |
|---|---|---|---|---|---|---|
| Sample size | 9719 | 10613 | 12070 | 10959 | 7329 | 9310 |

(c) Use this information to produce a table of counts with age groups in columns and volunteer status (yes or no) in rows. [This problem has been done for you.]

```
vol.data <- matrix(c(2128, 2632, 4116, 3430, 2015, 2113, 7591, 7981, 7954, 7529, 5314, 7197),
                   ncol=6, byrow=TRUE)
vol <- rep(rep(c("Vol", "noVol"), 6), times = vol.data)
year <- rep(rep(c("y1624","y2534","y3544","y4554","y5564", "y65plus"), each = 2), times = vol.data)
volunteer <- data.frame(vol, year)

ggplot(volunteer) +
  geom_bar(aes(x = year, fill = vol), position = "fill")
```




(d) Conduct the chi-square test. Report the hypotheses, check of technical conditions, sampling distribution, test statistic, and p-value. (Provide the details of your calculations and/or relevant computer output.)

[4] From ISCAM, HW 5.6

Summarize your conclusion.

(e) Construct a $2 \times 6$ table with the same row and column headings as in (c), but containing only + and – signs indicating whether the observed count is larger (+) or smaller (–) than expected in that cell. Does this table reveal a pattern? Explain what that pattern suggests about the relationship between age group and volunteerism.

note that if we keep the output of `chisq.test`, we can pull out the observed and expected tables from the output. Use `output$expected` and `output$observed` to pull out the relevant information (assuming your output is called `output`).

(f) Use the `infer` syntax to run a randomization test which uses the same test statistic ($X^2$) but does not use the chi-square probability (mathematical) distribution to find a p-value. Instead, the data are permuted to find the sampling distribution assuming $H_0$ is true.

5. **US Volunteerism II**[5] Reconsider the previous question about volunteerism. Suppose that the sample sizes had all been smaller by a factor of 100 (so that the entire study included only about 600 subjects) but that the conditional proportions of volunteerism within each age group had all turned out the same.

(a) How (if at all) would you expect the segmented bar graph to change? Explain.

(b) How (if at all) would you expect the test statistic to change? Explain.

(c) How (if at all) would you expect the p-value to change? Explain.

(d) How (if at all) would you expect your conclusion to change? Explain.

(e) Repeat the chi-square analysis with this greatly reduced sample size (round the observed counts in the new table to the nearest integer). Confirm or correct your answers to (b)–(d) in light of this analysis.

[5]From ISCAM, HW 5.7