MA117 - Worksheet 11
Review
Week 3, Wednesday

Problem 1. Researchers studying the link between autism and prenatal vitamin use in the three months before pregnancy surveyed the mothers of a random sample of children with autism and another random sample of children without autism. The contingency table summarizes the results of these surveys.

	Autism	No Autism	Total
No Prenatal Vitamins	111	70	181
Prenatal Vitamins	143	159	302
Total	254	229	483

(a) Calculate and interpret a 95% confidence interval for the difference in autism rates between mothers who use prenatal vitamins and mothers who don't.

(b) Calculate and interpret a p-value for the hypothesis that there is no difference in autism rates between mothers who use prenatal vitamins and mothers who don't.

Problem 2. According to a report on sleep deprivation by the Centers for Disease Control and Prevention, the proportion of California residents who reported insufficient sleep during each of the preceding 30 days is 8.0%, while this proportion is 8.8% for Oregon residents. This data is based on simple random samples of 11,545 California and 4,691 Oregon residents.

(a) Calculate and interpret a 98% confidence interval for the difference between California and Oregon of the proportions of people who feel they get insufficient sleep.

(b) Calculate and interpret a p-value for the hypothesis that there is no difference between California and Oregon between the proportions of people who feel they get insufficient sleep.

Problem 3. Here's a data set about texbooks:

`https://sagrawalx.github.io/teaching/data/textbooks.csv`

Is there a relationship between the number of pages and the price of a textbook? If you had to use this data to predict the price of a 500 page textbook, what would your estimate be?

Problem 4. A study examined microhabitat factors associated with forage and bed sites of barking deer in Hainan Island, China. In this region, woods make up 4.8% of the land, cultivated grass plot makes up 14.7%, deciduous forests make up 39.6%, and "other" makes up the rest. Of the 426 sites where the deer forage, 4 were categorized as woods, 16 as cultivated grassplot, 61 as deciduous forests, and the rest as "other." Calculate and interpret a p-value for the hypothesis that the barking deer of Hainan Island have no preferences about the habitat in which they forage.

Problem 5. The `tidyverse` packages come equipped with a data frame called `diamonds`. It's a big data frame, recording 10 variables for 53490 different diamonds. If you type `diamonds` into the console, you'll see the beginning of this data frame.

- `price`: price in US dollars
- `carat`: weight of the diamond in carats (1 ct = 200 mg)

- `cut`: quality of the cut (`Fair`, `Good`, `Very Good`, `Premium`, `Ideal`)
- `color`: diamond color (from `D` (best) to `J` (worst))
- `clarity`: a measurement of how clear the diamond is (`I1` (worst), `SI2`, `SI1`, `VS2`, `VS1`, `VVS2`, `VVS1`, `IF` (best))
- `x`: length in mm
- `y`: width in mm
- `z`: depth in mm
- `depth`: total depth percentage (ie, `z / mean(x, y)`)
- `table`: width of top of diamond relative to widest point

(a) Classify each variable as numerical or categorical.

(b) "Standardized price" is price divided by 100 times the weight in carats.

  (i) Calculate and interpret a 95% confidence interval for the difference in average standardized price for 0.99 carat diamonds and 1 carat diamonds.

  (ii) Calculate and interpret a p-values for the data under the hypothesis that there is a no difference between the average standardized prices of 0.99 carat diamonds and 1 carat diamonds.

(c) Choose an appropriate inference task that addresses the following questions, and then conduct the statistical analysis.

  (i) Does the average `depth` of diamonds vary according to `cut`?

  (ii) Does the average weight of a diamond in carats vary according to `clarity`?

  (iii) Does the average standardized price of diamonds vary according to `color`?

  (iv) Is there a difference on average between the length and the width of a diamond?

  (v) Are the `cut` and `clarity` of a diamond associated?

Problem 6. Here's a data set about vocabulary:

`https://sagrawalx.github.io/teaching/data/vocab.csv`

This data matrix contains 30351 observations of 5 variables:

- `X` is an ID of the respondent.
- `year` of the survey.
- `sex` of the respondent.
- `education` in years.
- `vocabulary` test score (out of 10).

(a) During what time period was this data collected?

(b) Is there a *statistically* significant difference between the vocabularies of men and women? If so, is there a *practically* significant difference? Do you clearly understand the difference between these two questions?

(c) Are `education` and `vocabulary` associated? Run an appropriate hypothesis test to answer this question.