Problem Identification Examples

April 25, 2016

Identification of Problem Types

Recall the following notation:

- K: categorical variable with 2 groups
- G: categorical variable with 3+ groups
- H: continuous variable

For each of the following problems,

A) identify the model type (e.g., $K_1 \sim K_2$),
B) determine which type of problem it is
    - One Proportion
    - Two Proportions
    - Multiple Proportions (Goodness of Fit)
    - Multiple Proportions (Test of Independence)
    - One Mean
    - Two Means (Independent)
    - Two Means (Paired)
    - Multiple Means
    - Linear Regression
    - Logistic Regression
C) draw a sketch of an effective visualization of the sample data and give the name of that type of plot,
D) write (in symbols) the parameter(s) and point estimate(s),
E) write the null and alternative hypothesis,
F) identify what conditions need to be met in order to use a named distribution/theoretical approach,
G) if relevant, provide a formula for the confidence interval of the parameter of interest,
H) give a formula for how to calculate the P-value based on simulation/randomization AND via the theoretical approach (when applicable), and
I) assuming conditions are met, provide the named distribution (e.g., $t(df = 22)$) for the null distribution.

You can assume that the significance level is 5% for all problems here.

1. The National Survey of Family Growth conducted by the Centers for Disease Control gathers information on family life, marriage and divorce, pregnancy, infertility, use of contraception, and men's and women's health. One of the variables collected on this survey is the age at first marriage. 5,534 randomly sampled US women between 2006 and 2010 completed the survey. The women sampled here had been married at least once. Do we have evidence that the mean age that all US women from 2006 to 2010 had an average age of first marriage of greater than 23 years?

```
ageAtMar <- read.csv("data/ageAtMar.csv")
```

2. Average income varies from one region of the country to another, and it often reflects both lifestyles and regional living expenses. Suppose a new graduate is considering a job in two locations, Cleveland, OH and Sacramento, CA, and he wants to see whether the average income in one of these cities is higher than the other. He would like to conduct a hypothesis test based on two small samples from the 2000 Census.

```
cleSac <- read.delim("data/cleSac.txt", header = TRUE)
```

3. Trace metals in drinking water affect the flavor and an unusually high concentration can pose a health hazard. Ten pairs of data were taken measuring zinc concentration in bottom water and surface water at 10 randomly selected locations on a stretch of river. Do the data suggest that the true average concentration in the bottom water exceeds that of surface water?

```
zinc_conc <- read.delim("data/zinc_conc.txt", header = TRUE)
```

9. The Child Health and Development Studies investigate a range of topics. One study considered a random sample of pregnancies between 1960 and 1967 among women in the Kaiser Foundation Health Plan in the San Francisco East Bay area with the focus on understanding what variables tend to influence the baby's `weight`. The variable `smoke` is coded 1 if the mother is a smoker, and 0 if not. Another variable considered is `parity`, which is 0 if the child is the first born, and 1 otherwise. Which one(s) of these variables is/are good predictor(s) of the response variable?

```
babies <- read.csv("data/babies.csv")
```

4. A 2010 survey asked 827 randomly sampled registered voters in California "Do you support? Or do you oppose? Drilling for oil and natural gas off the Coast of California? Or do you not know enough to say?" Conduct a hypothesis test to determine if the data provide strong evidence that the proportion of college graduates who do not have an opinion on this issue is different than that of non-college graduates.

```
offshore <- read.delim("data/offshore_drilling.txt")
```

5. A random sample of 500 U.S. adults were questioned regarding their political affiliation (`democrat` or `republican`) and opinion on a tax reform bill (`favor`, `indifferent`, `opposed`). Based on this sample, do we have reason to believe that political party and opinion on the bill are related?

```
tax_opinion <- read.csv("data/party_tax.csv")
```

6. A particular brand of candy-coated chocolate comes in five different colors: brown, yellow, orange, green, and coffee. The manufacturer of the candy says the candies are distributed in the following proportions: brown - 40%, yellow - 20%, orange = 20%, and the remaining are split evenly between green and coffee. A random sample of 580 pieces of this candy are collected. Does this random sample provide evidence against the manufacturer's claim?

```
candies <- read.csv("data/candies.csv")
```

10. On January 28, 1986, a routine launch was anticipated for the Challenger space shuttle. Seventy-three seconds into the flight, disaster happened: the shuttle broke apart, killing all seven crew members on board. An investigation into the cause of the disaster focused on a critical seal called an O-ring, and it is believed that damage to these O-rings during a shuttle launch may be related to the ambient temperature during the launch. Observational data on O-rings for 23 randomly selected shuttle missions was collected, where the mission order is based on the temperature at the time of the launch. `temp` gives the temperature in Fahrenheit and `damaged` is 1 when O-ring failed and 0 when it was undamaged.

```
orings <- read.delim("data/orings.txt", header = TRUE)
```

7. The CEO of a large electric utility claims that 80 percent of his 1,000,000 customers are satisfied with the service they receive. To test this claim, the local newspaper surveyed 100 customers, using simple random sampling. Based on these findings from the sample, can we reject the CEO's hypothesis that 80% of the customers are satisfied?

```
elec_survey <- c(rep("satisfied", 73), rep("unsatisfied", 27))
```

8. Chicken farming is a multi-billion dollar industry, and any methods that increase the growth rate of young chicks can reduce consumer costs while increasing company profits, possibly by millions of dollars. An experiment was conducted to measure and compare the effectiveness of various feed supplements on the growth rate of chickens. Newly hatched chicks were randomly allocated into six groups, and each group was given a different feed supplement. Do these data provide strong evidence that the average weights of chickens that were fed linseed and horsebean are different?

```
chick_wts <- read.csv("data/chickwts.csv")
```