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

your name here

due Feb 6, 2020

Homework 2

Assignment Summary (Goals)

- Practice using R.
- Reflect on how sample proportions can vary from sample to sample.
- Calculate normal probabilities
- continue to use the `infer` syntax. See help sheets here: https://infer-dev.netlify.com/index.html

1. **Side effects of Avandia, Part I** (Problem 2.1, ISRS).

   Rosiglitazone is the active ingredient in the controversial type 2 diabetes medicine Avandia and has been linked to an increased risk of serious cardiovascular problems such as stroke, heart failure, and death. A common alternative treatment is pioglitazone, the active ingredient in a diabetes medicine called Actos. In a nationwide retrospective observational study of 227,571 Medicare beneficiaries aged 65 years or older, it was found that 2,593 of the 67,593 patients using rosiglitazone and 5,386 of the 159,978 using pioglitazone had serious cardiovascular problems. These data are summarized in the contingency table below. [D.J. Graham et al. "Risk of acute myocardial infarction, stroke, heart failure, and death in elderly Medicare patients treated with rosiglitazone or pioglitazone". In: JAMA 304.4 (2010), p. 411. issn: 0098-7484.]

		Cardiovascular	problems	
		yes	no	total
Treatment	Rosiglitazone	2,593	65,000	67,593
	Pioglitazone	5,386	154,592	159,978
	Total	7,979	219,592	227,571

   Determine if each of the following statements is true or false. If false, explain why. Be careful: The reasoning may be wrong even if the statement's conclusion is correct. In such cases, the statement should be considered false.

   (a) Since more patients on pioglitazone had cardiovascular problems (5,386 vs. 2,593), we can conclude that the rate of cardiovascular problems for those on a pioglitazone treatment is higher.

   (b) The data suggest that diabetic patients who are taking rosiglitazone are more likely to have cardiovascular problems since the rate of incidence was (2,593 / 67,593 = 0.038) 3.8% for patients on this treatment, while it was only (5,386 / 159,978 = 0.034) 3.4% for patients on pioglitazone.

   (c) The fact that the rate of incidence is higher for the rosiglitazone group proves that rosiglitazone causes serious cardiovascular problems.

   (d) Based on the information provided so far, we cannot tell if the difference between the rates of incidences is due to a relationship between the two variables or due to chance. (Hint: with such a huge sample size, `infer` is slow, you might try only a few hundred reps.)

2. **Distribution of $\hat{p}$** (Problem 2.11, ISRS). Look at page 115 for necessary images.

   Suppose the true population proportion were p = 0.1. The figure below shows what the distribution of a sample proportion looks like when the sample size is $n = 20$, $n = 100$, and $n = 500$. What does each point (observation) in each of the samples represent?

Describe how the distribution of the sample proportion, $\hat{p}$, changes as n becomes larger.

3. **Area under the curve** (problem 2.15, ISRS)

   What percent of a standard normal distribution N($\mu = 0$; $\sigma = 1$) is found in each region? Be sure to draw a graph (or use a function in R that provides a graph).

(a) Z < -1.35
(b) Z > 1.48
(c) -0.4 < Z < 1.5
(d) |Z| > 2

4. **Scores on the GRE** (Problem 2.17, ISRS).

   A college senior who took the Graduate Record Examination exam scored 620 on the Verbal Reasoning section and 670 on the Quantitative Reasoning section. The mean score for Verbal Reasoning section was 462 with a standard deviation of 119, and the mean score for the Quantitative Reasoning was 584 with a standard deviation of 151. Suppose that both distributions are nearly normal.

(a) Write down the short-hand for these two normal distributions.

(b) What is her Z score on the Verbal Reasoning section? On the Quantitative Reasoning section? Draw a standard normal distribution curve and mark these two Z scores.

(c) What do these Z scores tell you?

(d) Relative to others, which section did she do better on?

(e) Find her percentile scores for the two exams.

(f) What percent of the test takers did better than her on the Verbal Reasoning section? On the Quantitative Reasoning section?

(g) Explain why simply comparing her raw scores from the two sections would lead to the incorrect conclusion that she did better on the Quantitative Reasoning section.

(h) If the distributions of the scores on these exams are not nearly normal, would your answers to parts (b) - (f) change? Explain your reasoning.

5. **Crime concerns in China.** (Problem 2.36, ISRS).

   A 2013 poll found that 24% of Chinese adults see crime as a very big problem, and the standard error for this estimate, which can reasonably be modeled using a normal distribution, is SE = 1.8% (= 0.018)[1]. Suppose an issue will get special attention from the Chinese government if more than 1-in-5 Chinese adults express concern on an issue.

(a) Construct hypotheses regarding whether or not crime should receive special attention by the Chinese government according to the 1-in-5 guideline.

(b) Discuss the appropriateness of using a one-sided or two-sided test for this exercise. Consider: for this decision process, would we care about one or both directions?

(c) Using a **normal** probability, what is the probability that you would get a sample statistic of 0.24 when the true probability is 0.2? [Pay attention: the probability you are calculating is the p-value!]

_____

[1]Environmental Concerns on the Rise in China. September 19, 2013. Pew Research.