

R Exercise 2



Problem 1

Suppose you are responsible for a clinical trial with 75 participants. Each participant in the trial has a condition, and the drug being tested is estimated to have a 45% success rate for curing the condition. Let X be the number of trial participants who are cleared of their condition by the drug.

- What is the expected value and standard deviation of the number of cured patients in the trial.
Hint: Assume that X is a binomially-distributed random variable. What are n and p ?
- Find the probability of all possible values of X both numerically and graphically.
- As stated above, the assumed efficacy of the drug is 45%. The pharmaceutical company who makes the drug will scrap further R&D into the drug if they find evidence that the effectiveness is at or below 30%. What is the probability of such a finding in this clinical trial, assuming that the true effectiveness of the drug is indeed 45%?
- Now let's assume that the pharmaceutical company has sufficient money to run as many trials as they please. In fact, let's simulate 10,000 trials as described above. You can do so using `rbinom(10000, n, p)`. Estimate the mean and standard deviation of this using the built-in `mean()` and `sd()` functions in R. Contrast these results to those from part a. How and why are they different?

Problem 2

You run a successful fast food chain. On an average day at your restaurant, you see around 600 patrons. Let X be the number of patrons who enter your restaurant on a given day.

- What is the expected value and standard deviation of the number of patrons in your restaurant on a given day?
Hint: Assume that X is a Poisson-distributed random variable. What is λ ?
- Find the possibility of all possible values of X graphically and the collective probability of each grouping of one hundred patrons (e.g. $P(0 \leq X < 100)$, $P(100 \leq X < 200)$ etc.) numerically.
- Your cooks are getting cranky and have told you that they will quit if there are more than 675 patrons ever. What is the probability of such an event?
- Simulate your restaurant workload for the next year (take a look at the function `rpois()`). Compare your results against those from part b.

Problem 3

Read in the following data which describes various health insurance policy holders as well as their policy rates.

```
insurance <- read.csv("https://raw.githubusercontent.com/EricBrownTTU/ISQS5346/main/insurance.csv")
```

- a. The body mass index (BMI) of a policy holder is given in the variable "bmi." Construct a histogram of the data. Does it look normally distributed to you? Why or why not?
- b. Simulate five sets (using `rnorm()`) of normally-distributed data with the same sample size, mean, and standard deviation as the BMI data. Construct a histogram for each.
- c. The data simulated in part (b) is normal. How many of your truly normal simulated datasets are similar to your actual data? Consider the shape of the distribution, skewness, kurtosis, etc.
- d. Construct the qq plot of the BMI data. Interpret the plot and communicate your findings.
- e. Construct qq plots for each of your five simulated datasets. How many of these qq plots resemble that of your BMI data? Compare with your results from (c) and decide how many of the simulated datasets are similar to your BMI data.
- f. What we've done in this problem is an implicit version of hypothesis testing (which we will learn about soon). We assume (H_0) that our data are normally-distributed. Consider your answer to part (e). You can use it to estimate the probability that your data reflects the null hypothesis (how many normal datasets out of five reflected your data?), or p-value. What is your p-value and what can you conclude from it (If $p < 0.05$, we have sufficient evidence to reject the null hypothesis that our data is not normally distributed).