Sample Homework

Eugene Katsevich

August 18, 2022

1 Instructions

Setup. Pull the latest version of this assignment from Github and set your working directory to `stat-961-fall-2021/homework/homework-1`. Consult the getting started guide if you need to brush up on `R`, `LaTeX`, or `Git`.

Collaboration. The collaboration policy is as stated on the Syllabus:

> "Students are permitted to work together on homework assignments, but solutions must be written up and submitted individually. Students must disclose any sources of assistance they received; furthermore, they are prohibited from verbatim copying from any source and from consulting solutions to problems that may be available online and/or from past iterations of the course."

In accordance with this policy,

Please list anyone you discussed this homework with: I did not discuss this homework with anyone.

Writeup. Use this document as a starting point for your writeup, adding your solutions between `\begin{sol}` and `\end{sol}`. See the preparing reports guide for guidance on compilation, creation of figures and tables, and presentation quality. Show all the code you wrote to produce your numerical results, and include complete derivations typeset in LaTeX for the mathematical questions.

Programming. The `tidyverse` paradigm for data manipulation (`dplyr`) and plotting (`ggplot2`) are strongly encouraged, but points will not be deducted for using base `R`.

Grading. Each sub-part of each problem will be worth 3 points: 0 points for no solution or completely wrong solution; 1 point for some progress; 2 points for a mostly correct solution; 3 points for a complete and correct solution modulo small flaws. The presentation quality of the solution for each problem (as exemplified by the guidelines in Section 3 of the preparing reports guide) will be evaluated out of an additional 3 points.

Submission. Compile your writeup to PDF and submit to Gradescope.

Problem 1. Quadratic theorem.

Consider a right triangle with legs of lengths a and b and hypotenuse of length c.

(a) State the Pythagorean theorem.

(b) Who initially proposed this theorem?

Solution 1.

(a) $a^2 + b^2 = c^2$.

(b) Pythagoras.

Problem 2. Simulation of normal random variables.

Consider $X_1, \ldots, X_n \overset{\text{i.i.d.}}{\sim} N(0, 1)$.

(a) Draw $n = 1000$ random samples and create a histogram of their distribution.

(b) Comment on the shape of your histogram.

Solution 2.

(a) Figure 1 shows the desired histogram.

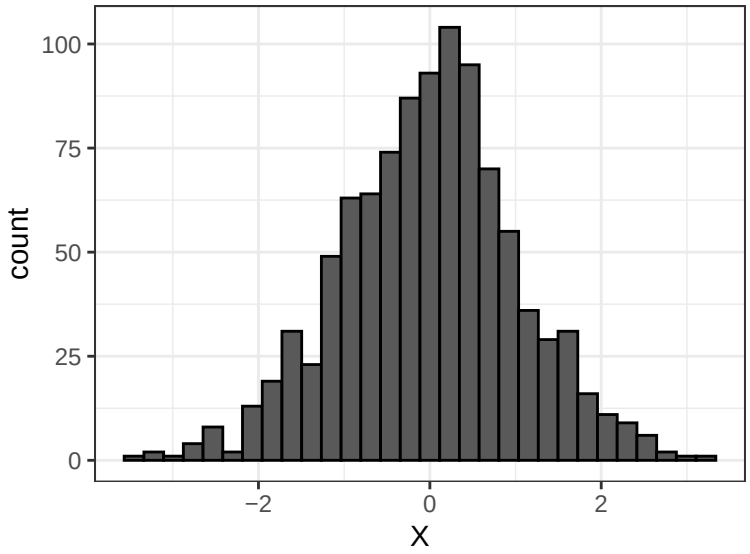

Figure 1: A sample of 1000 standard normal random variables.

(b) The histogram looks bell-shaped, as one would expect.

Problem 3. Data analysis: Diamonds.

Consider the `diamonds` dataset built into `ggplot2`, whose first few rows are shown below.

Table 1: First five rows of `diamonds` data.

carat	cut	color	clarity	depth	table	price	x	y	z
0.23	Ideal	E	SI2	61.5	55	326	3.95	3.98	2.43
0.21	Premium	E	SI1	59.8	61	326	3.89	3.84	2.31
0.23	Good	E	VS1	56.9	65	327	4.05	4.07	2.31
0.29	Premium	I	VS2	62.4	58	334	4.20	4.23	2.63
0.31	Good	J	SI2	63.3	58	335	4.34	4.35	2.75

(a) Create a histogram of the diamond price, and comment on its shape.

(b) Create a table of average price by diamond cut, and comment on any trends.

(c) Run a linear regression of `price` on `carat`, and print a table of the regression summary. Comment on the results.

Solution 3.

(a) Figure 2 shows the distribution of diamond price. We see that the distribution has a long right tail.

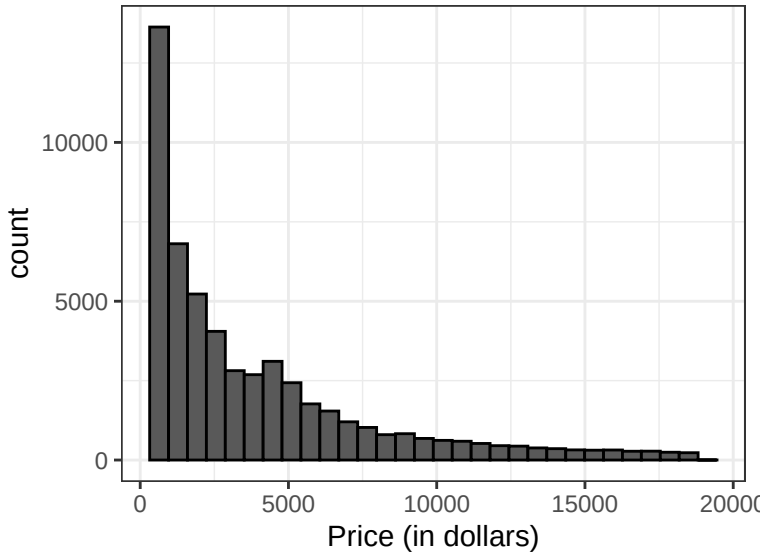

Figure 2: The distribution of `price` in the `diamonds` dataset.

(b) Table 2 shows the mean diamond price by cut. Surprisingly, the mean diamond price appears to *decrease* as cut improves!

(c) Table 3 shows the regression output. It appears that the carat of a diamond has an extremely significant impact on its price.

Table 2: Mean diamond price by cut.

Cut	Mean Price ($)
Fair	4358.76
Good	3928.86
Very Good	3981.76
Premium	4584.26
Ideal	3457.54

Table 3: Results of regressing price on carat.

	Dependent variable:
	price
carat	7,756.426***
	(14.067)
Constant	−2,256.361***
	(13.055)
Observations	53,940
R^2	0.849
Adjusted R^2	0.849
Residual Std. Error	1,548.562 (df = 53938)
F Statistic	304,050.900*** (df = 1; 53938)
Note:	*p<0.1; **p<0.05; ***p<0.01