

2.6 Problems

Problem 2.1. Classify each of the following variables as numerical or categorical. Then further classify the numerical variables as continuous or discrete, and the categorical variables as ordinal or nominal.

1. gender
2. race
3. age (in years)
4. age in categories (0–20, 21–35, 36–45, 45–60, 60–85, 85+)
5. zipcode
6. toxicity (mild, moderate, life-threatening, dead)
7. number of hospitalizations in the past year
8. change in HIV-RNA
9. weeks on treatment
10. treatment (placebo vs. estrogen)

Problem 2.2. Generate pseudo-random data from a normal distribution using a computer program or statistics package. In Stata this can be done using the `generate` command and the function `invnorm(uniform())`. Now generate a normal Q-Q plot for these data. Do this for several samples of size 10, 50, and 200. How well do the Q-Q plots approximate straight lines? This is valuable practice for judging how well an actual data set can be expected to approximate a straight line.

Problem 2.3. Generate pseudo-random samples of size 50 from a normal distribution (see Problem 2.2 for how to do this in Stata). Construct histograms of the data using 5, 7, and 15 bins. What do you notice? Do the shapes look like a normal distribution?

Problem 2.4. Warfarin is a drug used to prevent blood clots, for example in patients with irregular heartbeat and after heart surgery. However, too much warfarin can cause unusual bleeding or bruising, so calibration of the dose is important. A study contrasting calibration times (in hours) in two ethnic groups had the following results. For the sample of 18 Caucasians, the times were 2, 4, 6, 7, 8, 9, 10, 10, 12, 14, 16, 19, 21, 24, 26, 30, 35, 44, and 70; for the 18 Asian-Americans, the times were 2, 2, 3, 3, 4, 5, 5, 6, 6, 6, 7, 7, 8, 9, 10, 12, 19, and 32.

1. Display the data numerically to compare the two ethnic groups.
2. Display the data graphically to compare the two ethnic groups.
3. Describe the distribution of the data within ethnic group.
4. Log transform the data and repeat the graphical display. How do the displays with and without log transformation compare?
5. Can you think of other variables you might want to adjust for to help understand the ethnic differences better?

Problem 2.5. The timing of various stages in the contraction of the heart, determined by electro-cardiogram (EKG), can be used to diagnose heart problems. A commonly measured time interval in the contraction of the ventricles is the so-called QRS wave. A study was conducted to see if longer QRS times were related to the ability to induce rapid heart rhythms (called inducible ventricular tachycardia or IVT), which have been associated with adverse outcomes. In a study of 53 subjects, the 18 with IVT had QRS times (in milliseconds) of 70, 75, 86, 90, 96, 102, 110, 114, 116, 117, 120, 130, 136, 142, 145, 152, 170, and 182. The 35 patients without IVT had QRS times of 40, 50, 65, 70, 76, 78, 80, 82, 85, 88, 88, 89, 90, 94, 95, 96, 98, 98, 100, 102, 105, 107, 109, 110, 114, 115, 120, 125, 130, 135, 138, 150, 165, 170, and 180.

1. Display the data numerically to help understand whether QRS time is related to IVT.
2. Display the data graphically to help understand whether QRS time is related to IVT.
3. QRS time is commonly considered as abnormal if the value is greater than 120 msec. Generate a numerical display to help understand if abnormal QRS is related to IVT.
4. What are the advantages and disadvantages of treating QRS as binary (above 120 msec) instead of continuous?

Problem 2.6. Using the WCGS data set, generate a LOWESS (or equivalent) scatterplot smooth of SBP versus weight, comparable to Fig. 2.9. Next try the plot with bandwidths of 0.05, 0.15, and 0.50. How do they compare? Which is most useful for judging the linearity or lack of linearity of the relationship? The WCGS data are available at <http://www.biostat.ucsf.edu/vgsm>.