Biomedical Data Analysis Master of BIOINFORMATICS FOR HEALTH SCIENCES

Session 2 (revision of Descriptive statistics, probability concepts, normal and Binomial distribution)

Observational error (or measurement error) is the difference between a measured value of a quantity and its true value. In statistics, an error is not a "mistake." Variability is an inherent part of the results of measurements and of the measurement process.

Measurement errors can be divided into two components: random error and systematic error.

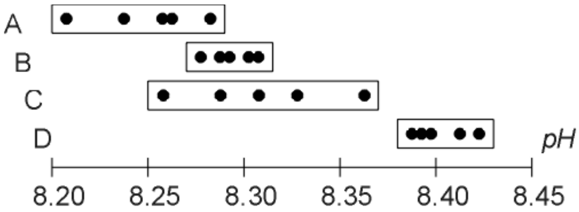
Systematic errors are errors that are not determined by chance but are introduced by an inaccuracy (involving either the observation or measurement process) inherent to the system. Random errors are related to the sampling. Each subsequent measurement has a random error, leading to imprecision in the estimation. A measurement with a low random error is said to be precise.

In systematic error each subsequent measurement has the same recurring error do to a bias.

Exercise 1 Four analysts, A, B, C and D, each prepared five replicate samples to measure the pH of a specific sample of soil. Results are following

						Mean	SD
A	8,208	8,239	8,258	8,264	8,283	8,250	0,028
B	8,278	8,288	8,293	8,304	8,308	8,294	0,012
C	8,259	8,289	8,308	8,329	8,363	8,310	0,039
D	8,389	8,393	8,399	8,413	8,423	8,403	0,014

The four sets of results are shown diagrammatically below:



Which set has less random error? which has the highest random error?

Which set is more precise? which set is less precise?

Set B and D give divergent results, could you state there is a systematic error between sets B and D? Which one is more likely to be biased?

If the true value were known to be 8.31, which set would be more accurate?

Exercise 2. In a certain college, 55% of the students are women. Suppose we take a sample of two students. Use a probability tree to find the probability

(a) that both chosen students are women. (0.3025)

(b) that at least one of the two students is a woman. (0.495)

Exercise 3. Suppose that a student who is about to take a multiple choice test has only learned 40% of the material covered by the exam. Thus, there is a 40% chance that she will know the answer to a question. However, even if she does not know the answer to a question, she still has a 20% chance of getting the right answer by guessing. If we choose a question at random from the exam, what is the probability that she will get it right? (0.52)

Exercise 4. If a woman takes an early pregnancy test, she will either test positive, meaning that the test says she is pregnant, or test negative, meaning that the test says she is not pregnant. Suppose that if a woman really is pregnant, there is a 98% chance that she will test positive. Also, suppose that if a woman really is not pregnant, there is a 99% chance that she will test negative.

(a) Suppose that 1,000 women take early pregnancy tests and that 100 of them really are pregnant. What is the probability that a randomly chosen woman from this group will test positive? (0.107)

(b) Suppose that 1,000 women take early pregnancy tests and that 50 of them really are pregnant. What is the probability that a randomly chosen woman from this group will test positive? (0.0585)

(c) Consider the setting of part (a). Suppose that a woman tests positive. What is the probability that she really is pregnant? (0.916)

(d) Consider the setting of part (b). Suppose that a woman tests positive. What is the probability that she really is pregnant? (0.838)

Exercise 5. In a study of the relationship between health risk and income, a large group of people living in Massachusetts were asked a series of questions. Some of the results are shown in the following table.

		Income			Total
		Low	Medium	High	
Stress	Stressed	526	274	216	1016
	Not stressed	1954	1680	1899	5533
	Total	2480	1954	2115	6549

(a) What is the probability that someone in this study is stressed? (0.155)

(b) Given that someone in this study is from the high income group, what is the probability that the person is stressed? (0.102)

(c) Compare your answers to parts (a) and (b). Is being stressed independent of having high income? Why or why not? (No)

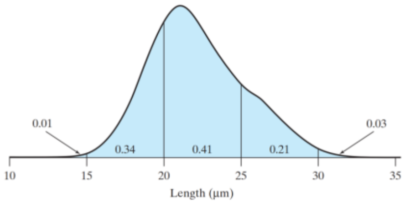
(d) What is the probability that someone in this study has low income? (0.379)

(e) What is the probability that someone in this study either is stressed or has low income (or both)? (0.454)

(f) What is the probability that someone in this study either is stressed and has low income? (0.080)

Exercise 6. Suppose that in a certain population of married couples 30% of the husbands smoke, 20% of the wives smoke, and in 8% of the couples both the husband and the wife smoke. Is the smoking status (smoker or nonsmoker) of the husband independent of that of the wife? Why or why not? (0.06 different to 0.08)

Exercise 7. In a certain population of the parasite Trypanosoma, the lengths of individuals are distributed as indicated by the density curve shown here. Areas under the curve are shown in the figure.



Consider the length of an individual trypanosome chosen at random from the population. Find

(a) Pr{20 < length < 30} (0.62)

(b) Pr{length > 20} (0.65)

(c) Pr{length < 20} (0.35)

Suppose we take a sample of two trypanosomes. What is the probability that

(d) both trypanosomes will be shorter than 20? (0.1225)

(e) the first trypanosome will be shorter than 20 and the second trypanosome will be longer than 25? (0.084)

(f) exactly one of the trypanosomes will be shorter than 20 and one trypanosome will be longer than 25? (0.168)

Exercise 8. In a certain population of the freshwater sculpin, Cottus rotheus, the distribution of the number of tail vertebrae, Y, is as shown in next Table

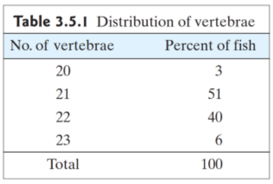


Table 3.5.1 Distribution of vertebrae

No. of vertebrae	Percent of fish
20	3
21	51
22	40
23	6
Total	100

Calculate the mean of Y and the variance of Y

The mean of a discrete random variable Y is defined as $\mu_Y = \sum y_i \Pr(Y = y_i)$

The variance of a discrete random variable Y is defined as $\sigma_Y^2 = \sum (y_i - \mu_Y)^2 \Pr(Y = y_i)$

The mean of Y is 21.49 and the variance of Y is 0.4299

Exercise 9. Consider rolling a die that is perfectly balanced so that each of the six faces is equally likely to come up and let the random variable Y represent the number of spots showing.

Calculate the expected value, or mean, of Y and the variance of Y

$E(Y) = \mu_Y = 3.5 \quad \sigma_Y^2 = 2.92$

Exercise 10. The seeds of the garden pea (Pisum sativum) are either yellow or green. A certain cross between pea plants produces progeny in the ratio 3 yellow 1 green. If four randomly chosen progeny of such a cross are examined, what is the probability that

(a) three are yellow and one is green? (0.4219)

(b) all four are yellow? (0.3164)

(c) all four are the same color? (0.3203)

Exercise 11. In the United States, 42% of the population has type A blood. Consider taking a sample of size 4. Let Y denote the number of persons in the sample with type A blood. Find

(a) Pr{Y = 0} (0.113)

(b) Pr{Y = 1} (0.328)

(c) Pr{Y = 2} (0.356)

(d) $\Pr\{0 \le Y \le 2\}$ (0.797)

Exercise 12. A certain drug treatment cures 90% of cases of hookworm in children. Suppose that 20 children suffering from hookworm are to be treated, and that the children can be regarded as a random sample from the population. Find the probability that

(a) all 20 will be cured. (0.1216)

(b) all but 1 will be cured. (0.2702)

(c) exactly 18 will be cured. (0.2852)

(d) exactly 90% will be cured. (0.2852)

Exercise 13. The shell of the land snail Limocolaria martensiana has two possible color forms: streaked and pallid. In a certain population of these snails, 60% of the individuals have streaked shells. Suppose that a random sample of 10 snails is to be chosen from this population. Find the probability that the percentage of streaked-shelled snails in the sample will be

(a) 50%. (0.2007)
(b) 60%. (0.2508)
(c) 70%. (0.2150)

Exercise 14. Following with the same sample of size 10 from the snail population

(a) What is the mean number of streaked-shelled snails? (6)

(b) What is the standard deviation of the number of streaked-shelled snails? (1.55)

Exercise 15. Students in a large botany class conducted an experiment on the germination of seeds of the Saguaro cactus. As part of the experiment, each student planted five seeds in a small cup, kept the cup near a window, and checked every day for germination (sprouting).The class results on the seventh day after planting were as displayed in the table.

NUMBER OF SEEDS		Number of students
Germinated	Not germinated	
0	5	17
1	4	53
2	3	94
3	2	79
4	1	33
5	0	4

a). Calculate the number of expected frequencies of each group of germinated seed (0, 1, 2, 3, 4 and 5)

b). Two students, Fran and Bob, were talking before class. All of Fran's seeds had germinated by the seventh day, whereas none of Bob's had. Bob wondered whether he had done something wrong. With the perspective gained from seeing all 280 students' results, what would you say to Bob? (Hint: Can the variation among the students be explained by the hypothesis that some of the seeds were good and some were poor, with each student receiving a randomly chosen five seeds?)