

**KEY: MID-TERM BIOL 300: October 2012**

1. (6 points) In 2011, physicists reported measurements of the speed of neutrinos that were (slightly) faster than light, which is believed to be impossible under standard physics theory. The conclusions were based on several measurements from the same lab, and the data from each neutrino measured was very similar to all other neutrinos; that is, there was very little variation among the measurements. However, after five months of checking, it was determined that there was a loose cable and a clock that was slightly too fast, and the combination of the two errors meant that each measurement gave a speed that was slightly too fast relative to the true values. Were the results of the original experiment that erroneously showed too fast speeds for neutrinos a problem of accuracy or precision? Which was more important in this case, bias or sampling error?

There was a systematic error (the clock was too fast) which caused the estimates to be too large *on average*. This is a problem of **accuracy**, and the results showed **bias**.

2. (6 points) If a random number is drawn from a  $\chi^2$  distribution with 12 degrees of freedom, what is the probability that the number is greater than 23.34?

From the stats tables, this probability is 0.025.

If a random number is drawn from a  $\chi^2$  distribution with 12 degrees of freedom, what is the probability that the number is less than 23.34?

$1 - 0.025 = 0.975$ .

3. (20 points) In many competitions, the competitors are judged by subjective criteria one after the other. (Think for example of an audition or musicians in a talent contest.) Does the order in which the competitors appear affect the decisions of the judges? An experiment to look for these order effects (Mantonakis *et al.* 2009) gave volunteers four glasses of wine, and the volunteers were asked to say which of the four was the best wine. In actual fact, all four glasses were poured from the same bottle. This experiment was done on 33 volunteers, of whom 15 preferred the first glass, 5 preferred the second glass, 2 preferred the third glass, and the other 11 preferred the last glass. Is there evidence from these data that the order in which these was presented had an effect on the preference of the volunteers?

$H_0$ : Order doesn't matter for the probability of a glass being chosen. (Each glass is equally likely.)

$H_A$ : Order determines the probability of choice.

Glass	1	2	3	4
Observed	15	5	2	11
Expected	33 (1/4) = 8.25	8.25	8.25	8.25

$$\chi^2 = \frac{(15-8.25)^2}{8.25} + \frac{(5-8.25)^2}{8.25} + \frac{(2-8.25)^2}{8.25} + \frac{(11-8.25)^2}{8.25} = 12.45$$

$$df = 4 - 0 - 1 = 3$$

$\chi^2_{3,0.01(2)} = 11.34$ , and  $\chi^2 = 12.45$  is greater than 11.34, so  $P < 0.01$ .

We reject the null hypothesis. The order of presentation of the wine affects its probability of being chosen.

4. (16 points) In the late eighties in Canada, about 52% of adult men smoked. It was estimated that male smokers had a lifetime probability of 17.2% of developing lung cancer, where as a non-smoker had a 1.3% chance of getting lung cancer during his life (Villeneuve and Mao 1994).

a. What is the conditional probability of a Canadian man getting cancer, conditioned on the fact that he smoked in the late eighties?

0.172

b. What is the probability that a Canadian man from the late eighties both smoked and eventually contracted lung cancer?

$$\begin{aligned} \text{Pr}[\text{smoking and cancer}] &= \text{Pr}[\text{smoking}] \text{Pr}[\text{cancer} | \text{smoking}] \\ &= 0.52 (0.172) \\ &= 0.08944 \end{aligned}$$

c. What is the probability that a Canadian male in the late eighties would eventually develop lung cancer?

$$\begin{aligned} \text{Pr}[\text{cancer}] &= \text{Pr}[\text{cancer} | \text{smoking}] \text{Pr}[\text{smoking}] + \text{Pr}[\text{cancer} | \text{non-smoking}] \text{Pr}[\text{non-smoking}] \\ &= 0.172(0.52) + (0.013) (1-0.52) \\ &= 0.09568 \end{aligned}$$

d. What is the probability that a man from this population smoked given that he eventually developed lung cancer?

$$\begin{aligned}\Pr[\textit{smoking} | \textit{cancer}] &= \frac{\Pr[\textit{cancer} | \textit{smoking}]\Pr[\textit{smoking}]}{\Pr[\textit{cancer}]} \\ &= \frac{(0.172)(0.52)}{0.09568} \\ &= 0.9348\end{aligned}$$

5. (5 points each) For each of the following scenarios, **identify the best statistical test to use and state the null hypothesis**. (Please note, do not give the answer to the specific question, but simply state the best test to use and the null hypothesis for the scenario.)

- a. Asking whether a new drug will increase the probability that a hepatitis A patient will survive two years after diagnosis (compared to patients not receiving the drug) – Assume large samples.

$\chi^2$  contingency analysis

$H_0$ : The probability of survival is the same for patients on the new drug or controls.

- b. Asking whether the probability of survival for two years of hepatitis patients taking a new drug is greater than 80% (assume sample size is 20).

Binomial test

$H_0$ : The probability of survival on the new drug is 0.80.

- c. Asking whether smokers or non-smokers are more likely to develop throat cancer (assume a study of 15 smokers and 35 non-smokers, with an overall probability of 4% that a study subject gets throat cancer).

Fisher's exact test (The expected values are too small for a  $\chi^2$  test.)

$H_0$ : Smokers and non-smokers are equally likely to get throat cancer.

- d. Asking whether the cases of hepatitis per year in each country is predicted by the population size of the countries.

$\chi^2$  goodness of fit test

$H_0$ : The number of cases of hepatitis is proportional to country population size.

- e. Testing whether car accidents occur independently and with equal probability on all days of the year.

$\chi^2$  goodness of fit test

$H_0$ : The number of car accidents per day fit a Poisson distribution.

6. (18 points) Systolic blood pressure was measured (in units of mm Hg) during preventative health examinations on people in Dallas, Texas. Here are the measurements for a subset of these patients. (For each of the following, remember to give units if appropriate.)

112, 128, 108, 129, 125, 153, 155, 132, 137

a. What is the best estimate of the mean systolic blood pressure in this population?

131 mm Hg

b. What is the best estimate of the standard deviation of blood pressure in this population?

15.95 mm Hg

c. What is the coefficient of variation of systolic blood pressure in this sample?

$$CV = 100\% \frac{s}{\bar{x}} = 100\% \frac{15.95}{131} = 12.2\%$$

d. Give a 95% confidence interval for the mean systolic blood pressure.

$$SE = \frac{15.95}{\sqrt{9}} = 5.32 \quad df = 9 - 1 = 8$$

$$\bar{x} \pm SE t_{df, 0.05(2)} = 131 \pm 5.32(2.31) = 131 \pm 12.3$$

e. What is the standard error of the mean systolic blood pressure?

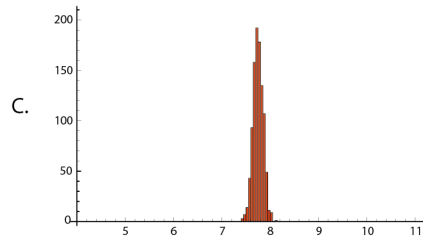
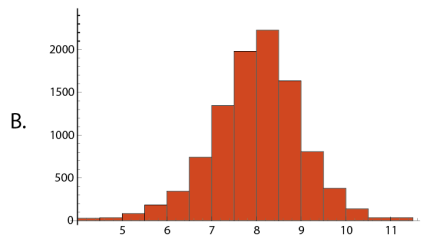
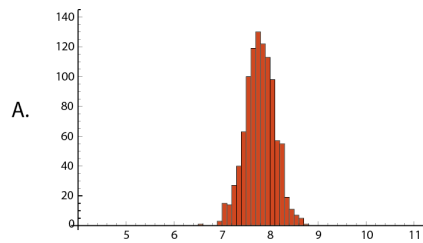
5.32 (See above.)

f. What is the median systolic blood pressure in this sample?

129 mm Hg

7. (9 points) The following three histograms (A, B, and C) plot the following three distributions:

- (1) the distribution of the number of hours of sleep an adult gets per night in Europeans,
  - (2) the distribution of the means of samples of size 10 taken from this population, and
  - (3) the distribution of the means of samples of size 100,
- but not necessarily in that order.



a. Identify which graph goes with which distribution.

- (1) B                      (2) A                      (3) C

b. What features of these distributions allowed you to distinguish which was which?

The variance around the mean should be greatest for the distribution of the individual data points and decrease with sample means taken from samples of larger size.

c. What is the approximate mean of the original distribution? What is the approximate mean of the distributions of sample means?

Approximately 7.8 for both.