Solutions to Exam 2

1. Give thorough answers to the following questions:

   (a) Define a Bernoulli trial.

   Answer: A Bernoulli trial is a random experiment with two possible, mutually exclusive, outcomes. The probability of an outcome (called a "success") is p, for some $0 < p < 1$, and the probability of the other outcome (called a "failure") is $1 - p$. $\square$

   (b) State what a sampling distribution for a statistic is.

   Answer: A sampling distribution of a statistic is the distribution of the set of values of the statistic which results from repeated random sampling. $\square$

   (c) What is a level C confidence interval?

   Answer: A level C confidence interval for a parameter is one computed from sample data by a procedure that produces intervals which capture the parameter with probability C. $\square$

   (d) What does it mean for two random variables, X and Y, to be independent?

   Answer: The discrete random variables X and Y are said to be independent is

   $$P(X = x, Y = y) = P(X = x) \cdot P(Y = y)$$

   for all possible values x and y of the random variables. $\square$

   (e) State the Central Limit Theorem.

   Answer: Let $X_1, X_2, \ldots, X_n, \ldots$ denote independent random variables which have the same distribution with mean μ and variance σ^2. The Central Limit Theorem states that, if n is large, then the distribution of the sample mean,

   $$\overline{X}_n = \frac{X_1 + X_2 + \cdots + X_n}{n},$$

   is approximately normal with mean μ and variance $\dfrac{\sigma^2}{n}$. $\square$

2. For each of the following scenarios, determine whether the binomial distribution is the appropriate distribution for the random variable X. Justify your answer in each case.

(a) X denotes the number of phone calls received in a one-hour period.

> ***Answer:*** X is not binomial since the range of possible values for X is unlimited. □

(b) X is the number of people in a random sample of size 50 from a large population that have type-AB blood.

> ***Answer:*** Yes, X is binomial with parameters $n = 50$ and p being the probability that a person in the population selected at random will have a type-AB blood. □

(c) A hand of 5 cards will be dealt from a standard deck of 52 cards that has been thoroughly shuffled. Let X denote the number of hearts in the hand of 5 cards.

> ***Answer:*** X is not binomial since the selections of the cards to form the hand are not independent trials. □

(d) The digits from $0, 1, , \ldots, 9$ are written on 10 separate cards. The cards are thoroughly shuffled. A card is selected at random, the number noted and the card placed back in the deck. The process is repeated 5 times. Let X denote the number of sevens that are observed in the experiment.

> ***Answer:*** X is binomial with parameters $n = 5$ and $p = 1/10$. □

(e) The digits from $0, 1, , \ldots, 9$ are written on 10 separate cards. The cards are thoroughly shuffled. A card is selected at random, the number noted and the card placed back in the deck. The process is repeated until five sevens are observed. Let X denote the number of trials needed to get the 5 sevens.

> ***Answer:*** X is not binomial since there is not limit to the values that X can take. □

3. A deck consists of five red cards and five black cards thoroughly shuffled.

(a) Six of these cards will be selected at random. Let X denote the number of red cards observed in the set of six selected cards. Which of the following probability distributions is appropriate for modeling the random variable X?

i. The Normal distribution with mean 3 and variance 1.22.

ii. The binomial distribution with parameters $n = 6$ and $p = 0.5$.

iii. The binomial distribution with parameters $n = 10$ and $p = 0.5$.

iv. None of the above.

Answer: iv. None of the above. X actually has a hypergeometric distribution. □

(b) One card is to be selected at random. The color will be observed and the card replaced in the set. The cards are then thoroughly reshuffled. This selection procedure is repeated four times.

Let X denote the number of red cards observed in these four trials. What is the mean of X?

Answer: X has a binomial distribution with parameters $n = 4$ and $p = 1/2$. Hence, the expected value of X is $np = 2$. □

4. The distribution of GPA scores is known to be left-skewed. At a large university, an English professor is interested in learning about the average GPA score of the English majors and minors. A simple random sample of 75 junior and senior English majors and minors results in an average GPA score of 2.97 (on a scale from 0 to 4). Assume that the distribution of GPA scores for all English majors and minors at this university is also left-skewed with standard deviation 0.62.

(a) Calculate a 95% confidence interval for the mean GPA of the junior and senior English majors and minors.

Solution: Use the formula

$$\left(\overline{X}_n - z^*\frac{\sigma}{\sqrt{n}}, \overline{X}_n + z^*\frac{\sigma}{\sqrt{n}}\right),$$

with $n = 75$, $\overline{X}_n = 2.97$, $\sigma = 0.62$, and $z^* = 1.96$, to get the interval

$$(2.97 - 0.14, 2.97 + 0.14).$$

□

(b) Determine whether each of the following statements is true or false.

i. If many samples of 75 English students were taken and many 95% confidence intervals calculated, only 5% of the time the sample mean would not fall in one of those confidence intervals.

Answer: False. □

ii. If many samples of 75 English students were taken and many 95% confidence intervals calculated, only 5% of the time the population mean would not fall in one of those confidence intervals.

Answer: True. □

iii. If many samples of 75 English students were taken and many 95% confidence intervals calculated, only 5% of the time the sample mean would not fall between the bounds of the confidence interval calculated in the previous question.

Answer: False. □

iv. The probability that the population mean falls between the bounds of the confidence interval calculated in the previous question equals 0.95.

Answer: False. □

5. The following table provides the results of a study in a major hospital concerning patients and their supplemental health coverage. A random sample of 95 surgical patients showed that 36 had supplemental health coverage; in a second random sample of 125 medical patients 56 had coverage:

	Medical Patients	Surgical Patients
Supplemental Health	56	36
No Supplemental Health	69	59

Table 1: Data for Problem 5

(a) Compute the proportion of the two types of patients that have supplemental health coverage. What do the data suggest?

Solution: Table 2 shows the row and column totals for the two–way table in Table 1. We can use them to compute various proportions. For instance, the proportion of medical patients with supplemental insurance is

	Medical Patients	Surgical Patients	Row Totals
Supplemental Health	56	36	92
No Supplemental Health	69	59	128
Column Totals	125	95	220

Table 2: Two–Way Table for 5

$$p_{_{MS}} = \frac{56}{125} \approx 0.45;$$

the proportion of surgical patients with supplemental insurance is

$$p_{SS} = \frac{36}{95} \approx 0.38;$$

and the proportion of the two types of patients that have supplemental health coverage is

$$p_{S} = \frac{92}{220} \approx 0.42.$$

Thus, medical patients are more likely to have supplemental insurance than surgical patients are. Is the difference between the proportions statistically significant? □

(b) State an appropriate null hypothesis for the data in Table 1.

Solution: A possible null model is that there is no association between the type of patient and their insurance status.

Alternatively, we could state the null hypothesis in terms of the proportions p_{MS} and p_{SS} by saying that they must be equal to that for the two types of patients altogether. □

(c) Compute the table of expected values assuming that the null hypothesis stated in the previous part is true.

Solution: Assuming that there is no association between the type of patient and their insurance status, we obtain the expected values shown in Table 3.

	Medical Patients	Surgical Patients	Row Totals
Supplemental Health	52.27273	39.72727	92
No Supplemental Health	72.72727	55.27273	128
Column Totals	125	95	220

Table 3: Expected Values for Data in Problem 5

The values in Table 3 are obtained by multiplying the row and column totals and dividing by the grand total. □

(d) Compute the Chi–Squared distance between counts in Table 1 and the expected values predicted by the null model.

Answer: We compute the Chi–Squared distance by using the formula

$$X^2 = \sum \frac{(\text{observed} - \text{expected})^2}{\text{expected}}.$$

We obtain $X^2 \approx 1.057838$. □

(e) The frequency histogram in Figure 1 on page 7 of this test shows the distribution of Chi–Squared distances resulting from 100 simulations of sampling done under the assumption that the null hypothesis is true. Use the histogram to estimate the p–value for the test. What do you conclude?

Solution: We estimate the p–value by computing the proportion of Chi–Squared distances stored in the array `ChiSqr`, and whose frequency distribution in pictured in Figure 1, that are equal to or larger than the X^2 valued computed in the previous part. According to the histogram, there are about 62 values that are below 1, it then follows that the p–value is about $\dfrac{38}{100}$ or 38%. This value os too big for the data in 1 to be statistically significant. It then follows that we cannot reject the null hypothesis and we conclude that there is no association between the type of patient and their insurance status. □

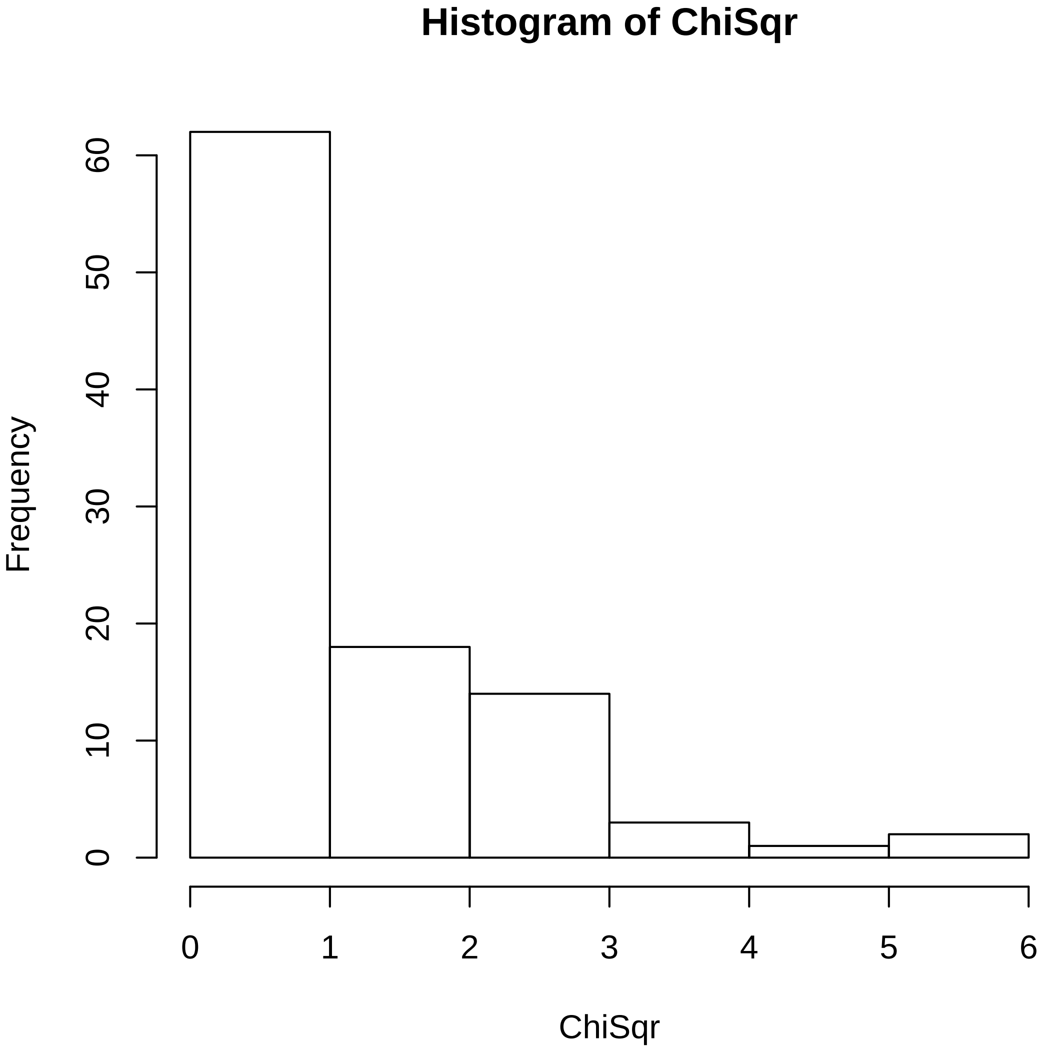

Figure 1: Histogram of `ChiSqr`