Solutions to Review Problems for Exam 2

1. For each of the following scenarios, determine whether the binomial distribution is the appropriate distribution for the random variable X. Explain your answer.

   (a) A fair coin is flipped ten times. Let X denote the number of times the coin comes up tails.

   > ***Answer:*** Yes, X has a binomial distribution with parameters $n = 10$ and $p = 1/2$ because each flip is a Bernoulli trial with probability $p = 1/2$ and the flips are independent. □

   (b) A fair coin is flipped multiple times. Let X denote the number of times the coin needs to be flipped until we see ten tails.

   > ***Answer:*** No, X does not have a binomial distribution. First of all, there is no limit to the number of trials needed to see 10 tails. Thus, there is not fixed parameter n. Consequently, the values of X can range from 0, 1, 3, all the way to infinity. □

   (c) A roulette wheel with one ball in it is turned six times. Let X denote the number of times the ball lands on red.

   > ***Answer:*** Yes, X has a binomial distribution with parameters $n = 6$, and p, the probability that the ball will land on a red spot. The outcome of each spin of the roulette is independent from the outcome of any other spin. □

   (d) There are ten people in the room: five men and five women. Three people are to be selected at random to form a committee. Let X denote the number of men on the three-person committee.

   > ***Answer:*** No, X does not have a binomial distribution. X has a hypergeometric distribution with parameters the number of men (5), the number of women (5), and the number of members in each committee. The possible values of X are 0, 1, 2 and 3 and their corresponding probabilities are:

$$P(X = 0) = \frac{\binom{5}{0} \cdot \binom{5}{3}}{\binom{10}{3}} = \frac{1}{12}$$

$$P(X = 1) = \frac{\binom{5}{1} \cdot \binom{5}{2}}{\binom{10}{3}} = \frac{5}{12}$$

$$P(X = 2) = \frac{\binom{5}{2} \cdot \binom{5}{1}}{\binom{10}{3}} = \frac{5}{12}$$

and

$$P(X = 1) = \frac{\binom{5}{3} \cdot \binom{5}{0}}{\binom{10}{3}} = \frac{1}{12}$$

Note that these are different from the ones predicted by the binomial model; namely,

$$P(X = k) = \binom{3}{k} \left(\frac{1}{2}\right)^k \left(\frac{1}{2}\right)^{3-k} \quad \text{for } k = 0, 1, 2, 3.$$

For instance,

$$P(X = 0) = \binom{3}{0} \left(\frac{1}{2}\right)^3 = \frac{1}{8}.$$

□

2. For each of the following scenarios, determine the appropriate distribution for the random variable X.

   (a) A fair die is rolled seven times. Let X denote the number of times we see an even number.

   Answer: X is binomial with parameters $n = 7$ and $p = 1/2$, the probability of a number from 1 to 6 coming our even. □

   (b) A card is selected at random from a standard deck of shuffled cards. The color of the card is determined and the card is returned to the deck. The cards are shuffled again. This selection procedure is repeated sixty times. Let X denote the proportion of times the selected card is red.

   Answer: Let Y denote the number of cards out of the sixty that come out red. Then, $X = \frac{Y}{n}$ is the mean of a random sample of size $n = 60$ drawn from a Bernoulli distribution with parameter $p = 1/2$. By the Central Limit Theorem, the distribution of X is approximately normal with mean $\mu = p = 1/2$ and variance $p(1-p)/n$, or $1/240$; that is,

$$X \text{ is approximately } N\left(\frac{1}{2}, \frac{1}{240}\right).$$

□

(c) The percentage of female students at a large university is known to be 46%. A simple random sample of 100 students is to be taken. Let X denote the number of male students in the sample.

Answer: X is binomial with parameters $n = 100$ and $p = 0.54$. Alternatively, by the Central Limit Theorem, X approximately normal with mean 54 and variance 24.84. □

(d) On any given Saturday during college football season, there are roughly 70 games being played. At each game, a fair coin is flipped to determine which team gets to kick off first. Let X denote the proportion of these coins that land heads.

Answer: By the Central Limit Theorem, X is approximately normal with mean 1/2 and variance 1/280. □

3. A set of ten cards consists of five red cards and five black cards. The cards are shuffled thoroughly.

(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. Describe the probability distribution which appropriately models the random variable X.

Answer: X has a hypergeometric distribution with parameters 5, 5 and 6; that is,

$$P(X = k) = \frac{\binom{5}{k} \cdot \binom{5}{6-k}}{\binom{10}{6}} \quad \text{for} \quad k = 0, 1, 2, 3, 4, 5.$$

□

(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: In this case, X has a binomial distribution with parameters $n = 4$ and $p = 1/2$. It then follows that the expected value of X is 2. □

4. Determine whether each of the following statements is true or false.

(a) The margin of error for a 95% confidence interval for the mean μ increases as the sample size increases.

Answer: FALSE
The margin of error for the mean is

$$z^* \frac{\sigma}{\sqrt{n}}$$

where z^* depends on the confidence level, σ is the standard deviation, and n is the sample size. Hence, as the sample size increases, the margin of error decreases. □

(b) The margin of error for a confidence interval for the mean μ, based on a specified sample size n, increases as the confidence level decreases.

Answer: FALSE
The margin of error for the mean is

$$z^* \frac{\sigma}{\sqrt{n}}$$

where z^* depends on the confidence level, σ is the standard deviation, and n is the sample size. The value z^* increases as the confidence level increases. Hence, the margin of error decreases as the confidence level decreases. □

(c) The margin of error for a 95% confidence interval for the mean μ decreases as the population standard deviation decreases.

Answer: TRUE
The margin of error for the mean is

$$z^* \frac{\sigma}{\sqrt{n}}$$

where z^* depends on the confidence level, σ is the standard deviation, and n is the sample size. Hence, the margin of error decreases as the population standard deviation decreases. □

(d) The sample size required to obtain a confidence interval of specified margin of error μ increases as the confidence level increases.

Answer: TRUE
The margin of error for the mean is

$$z^* \frac{\sigma}{\sqrt{n}}$$

where z^* depends on the confidence level, σ is the standard deviation, and n is the sample size. If we call the margin of error ME, we obtain that

$$\text{ME} = z^* \frac{\sigma}{\sqrt{n}}$$

from which we get that

$$n = \left(z^* \frac{\sigma}{\text{ME}} \right)^2.$$

As the confidence level increases, so does z^*, and consequently n also increases. □

5. Certain Middle School has calculated a 95% confidence interval for the mean height μ of 11–year–old boys at their school and found it to be 56 ± 2 inches.

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

   (a) There is a 95% probability that μ is between 54 and 58.

   Answer: False
   Confidence intervals for the population mean, μ, for the cases in which the population standard deviation, σ, is known, are computed from the sampling distribution of the sample mean, $\overline{X}_n$. For large sample sizes, n, we can use the Central Limit Theorem to approximate

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

   where Z is the standard normal random variable. Thus, the value $P(-z^* < Z < z^*)$ approximates the probability that if we collect a simple random sample of size n from the population, compute the sample mean, $\overline{X}_n$ from that sample, and then form the interval

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

   the interval contains the true population mean.
   Thus, a 95% confidence interval $(54, 58)$ does not mean that there is a probability of 95% that μ will be in that interval. It means, according to Moore, McCabe and Craig (on page 367 of the sixth edition of *Introduction to the Practice of Statistics*), that "the interval was calculated by a method that gives correct results in

95% of all possible samples." It does not say that the probability is 95% that the true mean lies between 54 and 58. "No randomness remains after we draw a particular random sample and compute the interval. The true mean either is or is not between 54 and 58. The probability calculations of standard statistical inference describes how often the *method*, not the particular sample, gives the correct answer." □

(b) There is a 95% probability that the true mean is 56, and there is a 95% chance that the true margin of error is 2.

Answer: False
A 95% confidence interval means that the procedure used to generate the interval will capture the true mean approximately 95% of the time. □

(c) If we took many additional random samples of the same size and from each computed a 95% confidence interval for μ, approximately 95% of these intervals would contain μ.

Answer: True
This is the definition of a 95% confidence interval. □

(d) If we took many additional random samples of the same size and from each computed a 95% confidence interval for μ, approximately 95% of the time μ would fall between 54 and 58.

Answer: False
Each time we collect a sample of size n and compute the sample mean and corresponding interval, we will obtain in general values different from 54 and 58. □

6. A nationally distributed college newspaper conducts a survey among students nationwide every year. This year, responses from a simple random sample of 204 college students to the question "About how many CDs do you own?" resulted in a sample mean $\overline{X}_n = 72.8$. Based on data from previous years, the editors of the newspaper will assume that $\sigma = 7.2$.

(a) Use the information given to obtain a 95% confidence interval for the mean number of CDs owned by all college students.

Solution: Use the formula

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

where $z^* = 1.96$ and $n = 204$ to obtain

$$(71.8, 73.8)$$

□

(b) Answer each of the following questions with yes, no, or can't tell.

i. Does the sample mean lie in the 95% confidence interval?

Answer: Yes. □

ii. Does the population mean lie in the 95% confidence interval?

Answer: Can't tell. □

iii. If we were to use a 92% confidence level, would the confidence interval from the same data produce an interval wider than the 95% confidence interval?

Answer: No.
In this case, z^* is about 1.751. So, the interval is shorter. □

iv. With a smaller sample size, all other things being the same, would the 95% confidence interval be wider?

Answer: Yes.
As n decreases, the margin of error, $z^*\dfrac{\sigma}{\sqrt{n}}$, increases; so, the interval will be wider. □

7. A review of voter registration records in a small town yielded the following table of the number of males and females registered as Democrat, Republican, or some other affiliation:

Affiliation \ Gender	Male	Female
Democrat	300	600
Republican	500	300
Other	200	100

Table 1: Data for Problem 7

Suppose we wish to test the null hypothesis that there is no association between party affiliation and gender. Under the null hypothesis, what is the expected number of male Democrats?

Solution: Compute the row totals, column totals and grand total as shown in Table 2.

Affiliation \ Gender	Male	Female	Row Totals
Democrat	300	600	900
Republican	500	300	800
Other	200	100	300
Column Totals	1000	1000	2000

Table 2: Two–way table for Problem 7

The expected values under the assumption of no association between the variables are then computing my multiplying row and column totals and dividing by the grand total. The results are shown in Table 3.

Affiliation \ Gender	Male	Female	Row Totals
Democrat	450	450	900
Republican	400	400	800
Other	150	150	300
Column Totals	1000	1000	2000

Table 3: Expected values for Problem 7

Thus, the expected number of male democrats would be 450. □

8. A specific type of electronic sensor has been experiencing failures. The manufacturer has studied 200 of these devices and has determined the type of failure (A, B, C) that occurred and the location on the sensor where the failure happened (External, Internal). The following table summarizes the findings:

Location \ Type	A	B	C	Total
Internal	50	16	31	97
External	61	26	16	103
Total	111	42	47	200

Table 4: Data for Problem 8

Test the null hypothesis is that there is no association between Type of Failure and Location of Failure.

***Solution**:* We present the solution to this problem using a permutation procedure in R.

First we enter the counts in the two–way table as a 2×3 matrix which we call `Obs`. Type

```
Obs <- matrix(c(50,16,31,61,26,16), 2,3,byrow=T)
```

to get the output

```
> Obs
      [,1] [,2] [,3]
[1,]   50   16   31
[2,]   61   26   16
```

The expected values, assuming that there is no association between the type of failure and location of failure, are obtained by applying the following function which we call `expected()` defined by the following code:

```
# Computing expected values for 2-way tables
# Enter M: the matrix of counts in 2-way table
expected <- function(M)
{
    R <- matrix(rowSums(M))
    C <- matrix(colSums(M),1,dim(M)[2])
    E <-  (R %*% C)/sum(M)
    return(E)
}
```

Typing `Exp <- expected(Obs)`, we obtain

```
> Exp
       [,1]  [,2]   [,3]
[1,] 53.835 20.37 22.795
[2,] 57.165 21.63 24.205
```

The Chi–Squared distance is then obtained from

```
Xsqr <- sum((Obs-Exp)^2/Exp)}.
```

This yields a value of about 8.085554.

To estimate the p–value for the test, we run the `IndepTest()` function defined by the following code:

```
# Permutation Test of Independence for 2-way tables
#
# Enter a matrix, M, of observed counts in 2-way table
# and the number of repetitions, N
# This code assumes that the expected() function has
# been defined.
#
IndepTest <- function(M,N)
    {
    Exp <- expected(M) # computes expected values
    CS <- array(dim=N) # sets up array dimension N
    RowGroups <- rep(1:dim(M)[1],rowSums(M))
        # sets up row groups
    L <- length(RowGroups)
    ColGroups <- rep(1:dim(M)[2],colSums(M))
        # sets up column groups
    for (k in 1:N)
    {
        perms <- sample(RowGroups,L)
        Sim <- table(perms,ColGroups)
        CS[k] <- sum((Sim-Exp)^2/Exp)
}
return(CS)
}
```

Typing

```
ChiSqr <- IndepTest(Obs,10000)
```

yields the array `ChiSqr` of $10,00$ simulated Chi–Squared distances. The p–value can then be estimated by

$$p\text{–value} \approx \texttt{pHat(ChiSqr,Xsqr)} \approx 0.0185.$$

Hence, the p–value is about 1.85%. Thus, we can reject the null hypothesis at the 5% significance level. We can then say that there is strong evidence to suggest that there is an association between the type of failure and the location. □

Alternate Solution: We can also estimate the p–value in the previous solution by using the χ^2_{df} distribution with

$$\text{df} = (r-1)(c-1)$$

degrees of freedom. In this case $r = 2$ and $c = 3$, so that $\text{df} = 2$. The estimate for the p value is then

$$p\text{–value} \approx P(\chi^2_2 \geqslant \texttt{Xsqr}).$$

In R, we may estimate this further as

$$p\text{–value} \approx \texttt{1-pchisq(Xsqr,2)} \approx 0.01754867.$$

We deduce the same conclusion as before. □

9. A researcher is interested in determining if the model used for the distribution of main economic concerns in the year 2003 for residents in a certain county can still be used in the year 2004. A sample of 370 residents from that county was surveyed in 2004. The following table displays the model for the distribution of economic concerns for the year 2003 and the observed number of sampled respondents in the survey for the same economic concerns for the year 2004:

	Jobs	Medical Care	Higher Education	Housing
2003 Model	5%	36%	27%	32%
2004 Counts	25	138	116	91

Table 5: Data for Problem 9

Perform an appropriate significance test.

Solution: We perform a goodness of fit test with the null hypothesis that the observations come from a population whose probability distribution is given by the proportions in the 2003 Model:

$$\text{H}_o\colon\ p_1 = 0.05,\ p_2 = 0.36,\ p_3 = 0.27,\ p_4 = 0.32.$$

We enter these proportions in an array called `Prop` in R:

```
Prop <-  c(0.05, 0.36, 0.27, 0.32)
```

The observed counts are entered in a vector called `Obs`:

```
Obs <-  c(25, 138, 116, 91)
```

The total number of observations, or the sample size, is then

```
n <- sum(Obs)
```

which yields $n = 370$.
The expected values are then

```
Exp <- n*Prop
```

This yields

```
> Exp
[1]  18.5 133.2  99.9 118.4
```

The Chi–Squared distance is then given by

```
Xsqr <- sum((Obs-Exp)^2/Exp)
```

which yields

```
> Xsqr
[1] 11.39233
```

At this point we may perform a randomization test to simulate taking samples of size n from a population with a distribution given by the null hypothesis. We present here a code that can be used for all goodness of fit tests like this one. It takes the vector of observed counts, the vector of postulated proportions, and the desired number of repetitions.

```
# Randomization Test for Goodness of Fit
# Author: Adolfo J. Rumbos
# Date: December 3, 2008
#
# Enter an array of observed counts, Obs, and array of postulated
# proportions, P, and the number of repetitions, N
#
GoodFitTest <- function(Obs,P,N)
```

```
{
    k <- length(Obs)  # number of categories
    n <- sum(Obs)   # sample size
    Exp <- n*P   # computes expected values
    CS <- array(dim=N) # sets up array dimension N
    CatCounts <- round(100*Prop) # Expected counts per categories in 100
    Categories <- rep(1:k,CatCounts)
    for (i in 1:N)
    {
        s <- sample(Categories,n,replace=T) # simulated samples
        Sim <- table(s)
        CS[i] <- sum((Sim-Exp)^2/Exp)
    }
return(CS)
}
```

Typing `ChiSqr <- GoodFitTest(Obs,Prop,10000)` returns an array of $10,000$ simulated values for the Chi–Squared distance. The histogram of this particular run is shown in Figure 1.

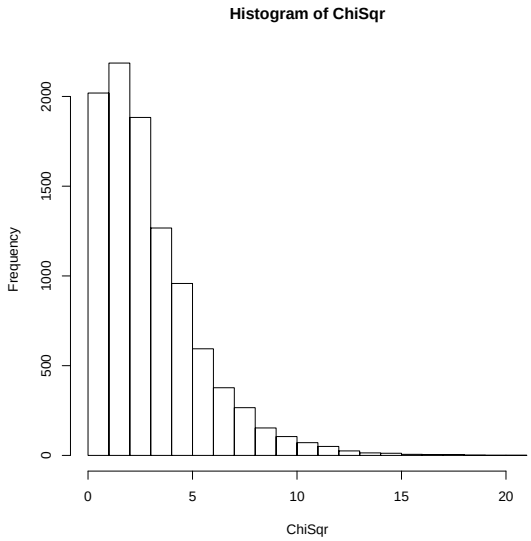

Figure 1: Histogram of `ChiSqr`

We now estmate the p–value as usual, applying the `pHat()` function with the observed Chi–Squared distance, `Xsqr`, as input to obtain a

p–value of about 0.0103. Thus, at the 5% level of significance, we can reject the null hypothesis. □

Alternate Solution: As was done in the previous problem, we can also estimate the p–value by using the χ^2_{df} distribution with

$$\text{df} = k - 1 = 3$$

degrees of freedom. The estimate for the p value is then

$$p\text{–value} \approx P(\chi^2_3 \geqslant \texttt{Xsqr}).$$

In R, we may estimate this further as

$$p\text{–value} \approx \texttt{1-pchisq(Xsqr,3)} \approx 0.009782996.$$

We deduce the same conclusion as before. Perhaps, we could claim a 1% significance level; but then again, this is another approximation to the p–value. □

10. There have been many studies that looked at the incidence of heart attack on the different days of the week. Studies in Japan and Scotland seemed to find that there was a substantial "spike" in heart attack on Mondays, perhaps as many as 20% more. A researcher studied a random selection of 200 heart attack victims and recorded the day of the week that their attack occurred. The following table summarizes the results:

Day	Sunday	Monday	Tuesday	Wednesday	Thursday	Friday	Saturday
Counts	24	36	27	26	32	26	29

Table 6: Data for Problem 10

The researcher was interested in whether the distribution of heart attacks was the same or different across the days of the week. Perform an appropriate significance test to answer this question.

Solution: We perform a goodness of fit test. In this case,

```
Obs <- c(24, 36, 27, 26, 32, 26, 29)


n <- sum(Obs)
```

```
k <- length(Obs)

Prop <- rep(1/k,k)

Exp <- n*Prop

Xsqr <- sum((Obs - Exp)^2/Exp)
```

The last line yields a value for `Xsqr` of 3.63. We estimate the p–value by performing a randomization test using the `GoodFitTest()` function defined in the previous problem:

```
ChiSqr <- GoodFitTest(Obs,Prop,10000)
```

The density histogram of `ChiSqr` is shown in Figure 2. In this case the

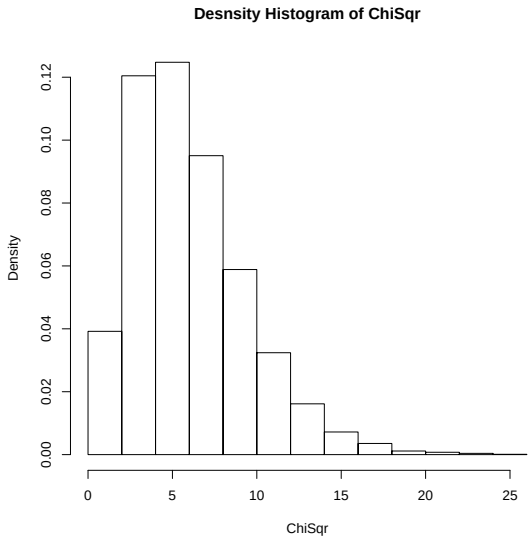

Figure 2: Density Histogram of `ChiSqr`

p–value is about 0.7317, and so we cannot reject the null hypothesis. Thus, the data do not support the assertion that distribution of heart attacks was different across the days of the week. Thus, the day of

the week does not have any bearing of the number of heart attacks.
□