# BIOS 6611 Homework 3---SOLUTIONS

Normal Distribution, Functions of Random Variables

## A. Ozone Status

Load **ozone.csv** into R. This is EPA data (`epa.gov/outdoor-air-quality-data`) that tabulates ozone levels from January 2017 to June 2017 for the Denver-Aurora-Lakewood area.

1. Estimate the daily probability of "good" ozone levels.

```
prob_aqi <- sum(aqi$AQI.Category == "Good") / length(aqi$AQI.Category)
prob_aqi
[1] 0.5298013
```

2. Making the assumption that the probability of ozone levels being rated "good" is constant from day-to-day, use the binomial distribution to calculate the "exact" probability that *at least* 5 of the next 7 days will have "good" ozone.

```
1 - pbinom(q = 4, size = 7, prob = prob_aqi) #estimate the probability
[1] 0.2783011
```

3. Recalculate this probability using the normal approximation to the binomial.

**Commented [KAM1]:** 25 points

```
approx_mean <- 7*prob_aqi #estimate mean from binomial: n*p
approx_var <- 7*prob_aqi*(1-prob_aqi) #estimate variance from binomial:
n*p*(1-p)

#Get the probability using the normal with continuity correction
1 - pnorm(q = 4.5, mean = approx_mean, sd = sqrt(approx_var))
[1] 0.2744862
```

4. Do you think it's believable that the total days with "good" ozone status follows a binomial distribution? Justify your answer, either with domain knowledge, statistics intuition, or by using the data itself.

**Commented [KAM2]:** 25 points

(Subjective) No. The binomial distribution assumes independent trials, but there will likely be correlation between ozone levels day-to-day (i.e., we expect a high level Monday makes higher levels on Tuesday more likely). The binomial distribution also assumes *p* is the same for all trials, but it is likely that other factors may lead to a varying probability (e.g., changes in wind patterns which bring more (or less) ozone pollution or don't remove ozone pollution as quickly from the area). Also, there are technically three levels to ozone status, which also violates the binomial distribution assumptions if we did not want to dichotomize.

## B. Estimating Hospital Budget

After receiving one of two medical procedures (coded “1” for standard, “2” for new), patients admitted to a hospital were followed for one month. The total medical costs per patient incurred by the hospital over this month were tabulated in the column “Cost” of **ProcedureCost.csv**, in units of $1000. Note that costs may be zero if no additional medical care was provided to a patient.

### Part 1

**Provide R code that reproducibly creates a table in the following format.** The item in each cell of the table should be the count of patients that match each category. By “reproducible”, we mean that your R code should be able to remake the table if a new “ProcedureCost.csv” file is provided in the same format (but with different data).

Hint: Refer to R Lab 1 Exercise 5: Study Design Sprints.

|  |  | **Cost** |  |
|---|---|---|---|
|  |  | Zero | Non-Zero |
| **Procedure Group** | 1 |  |  |
|  | 2 |  |  |

```
#create new column with indicator variable if cost is non-zero
df$nonzero[df$Cost == 0] <- 0
df$nonzero[df$Cost != 0] <- 1

table(df$Procedure, df$nonzero)

     0  1
  1 48 72
  2 15 65
```

## Part 2

For samples in each procedure group, reproducibly calculate the **proportion of non-zero costs** ($p_1$ and $p_2$), the **mean non-zero cost** ($m_1$ and $m_2$, i.e., cost among subjects that have some positive cost), and the **variance in the non-zero costs** ($v_1$ and $v_2$).

```
p <- aggregate(nonzero ~ Procedure, data = df, FUN = function(x) {sum(
     x != 0) / length(x)})[ , 2]
[1] 0.6000 0.8125

m <- aggregate(Cost ~ Procedure, data=df[df$Cost != 0, ],FUN=mean)[,2]
[1] 2.155417 1.085077

v <- aggregate(Cost ~ Procedure, data=df[df$Cost != 0, ],FUN=var)[,2]
[1] 1.262825 1.583760
```

From our R code above, we see that $p_1$=60%, $p_2$=81.25%, $m_1$=$2,155.42, $m_2$=$1,085.08, $v_1$=1,262.83 dollars$^2$, $v_2$=1,583.76 dollars$^2$.

## Part 3

**Commented [KAM3]:** 25 points

Both the (i) estimated total patient cost as well as (ii) the frequency of non-zero costs are important to hospital planning. We can model the cost (Y) for a given patient in a given procedure group by considering Y as the product of two random variables (i.e., Y = RZ) where:

- R = a Bernoulli random variable (binomial with n=1) that takes values of 0 or 1 (for non-zero cost)
- Z = a random variable that takes values between 0 and infinity (for cost)

Assuming that Z and R are independent, **derive mathematical expressions for the expected values and variance of the cost Y for a given subject**, in terms of
p = Pr(R = 1), m = E(Z), and v = Var(Z).

**Answer:**

$E[Y] = E[RZ] = E[R]E[Z] = pm$

$V[Y] = V[RZ]$
$V[Y] = E[R^2Z^2] - (E[RZ])^2$
$V[Y] = E[R^2]E[Z^2] - E[R]^2E[Z]^2$

Using the hint we can re-order our equation for variance ( V[Z] = E[Z²] – E[Z]² ):

$$E[R^2] = V[R] + E[R]^2$$
$$E[Z^2] = V[Z] + E[Z]^2$$

$V[Y] = \{(V[R] + E[R]^2)(V[Z] + E[Z]^2)\} - E[R]^2E[Z]^2$
$V[Y] = V[R]V[Z] + V[R]E[Z]^2 + V[Z]E[R]^2 + E[R]^2E[Z]^2 - E[R]^2E[Z]^2$
$V[Y] = V[R]V[Z] + V[R]E[Z]^2 + V[Z]E[R]^2$

$$E[R] = p \qquad E[Z] = m$$
$$V[R] = p(1-p) \qquad V[Z] = v$$

$V[Y] = p(1-p)v + p(1-p)m^2 + vp^2$
$V[Y] = pv + p(1-p)m^2$

Hints for Part 3:

- When two random variables are independent, the expected value of the product is the product of their expectations: i.e., $E(XY) = E(X)E(Y)$. Note that this does not hold true for the variance.
- Alternatively, those with more statistical background might consider the formal definition of the expectation for a function of two variables: $E(XY) = \sum_{X\in x}\int_y xy\, f(x,y)dy$, where $f(x,y)$ is the joint distribution of $X$ and $Y$.
- The variance of a random variable is $Var(X) = E(X^2) - (EX)^2$.

## Part 4 – Extra Credit

**Commented [KAM4]:** 3 extra credit points

The hospital expects different distributions for Y (cost) between the two procedure groups. Using your estimates for $p_1, m_1, v_1$ and $p_2, m_2, v_2$ based on the sample data (i.e., the values you calculated from Part 2), the hospital is interested in estimating how much they should budget for next year, if: (1) they anticipate treating $n_1 = 120$ and $n_2 = 200$ subjects for each respective group and (2) they want a less than 20% chance of this total expenditure exceeding their budget, i.e. for procedure groups 1 and 2 summed together?

**Formulate this question in terms of random variables using correct notation. Assume that the expenditures in the two procedure groups are independent of each other. Finally, give a numerical recommendation for the budget.**

Hint: Could the Central Limit Theorem apply here? Can we use the results from Part 3? What R function(s) is (are) needed to answer this: dnorm? pnorm? qnorm?

Answer:
Using our results from Part 3 and assuming all observations are independent:
$E(Y_1) = E(Z_1R_1) = p_1m_1$
$Var(Y_1) = Var(Z_1R_1) = p_1v_1 + p_1(1-p_1)m_1^2$
$E(Y_2) = E(Z_2R_2) = p_2m_2$
$Var(Y_2) = Var(Z_2R_2) = p_2v_2 + p_2(1-p_2)m_2^2$

Let the total cost be: $T = \sum_{i=1}^{n_1} Y_{1,i} + \sum_{j=1}^{n_2} Y_{2,j}$

| | |
|---|---|
| $\begin{aligned} E(T) &= E\left(\sum_{i=1}^{n_1} Y_{1,i} + \sum_{j=1}^{n_2} Y_{2,j}\right) \\ \square &= E\left(\sum_{i=1}^{n_1} Y_{1,i}\right) + E\left(\sum_{j=1}^{n_2} Y_{2,j}\right) \\ \square &= \sum\nolimits_{i=1}^{n_1} E(Y_{1,i}) + \sum\nolimits_{j=1}^{n_2} E(Y_{2,j}) \\ \square &= n_1E(Y_1) + n_2E(Y_2) \end{aligned}$ | $\begin{aligned} V(T) &= V\left(\sum_{i=1}^{n_1} Y_{1,i} + \sum_{j=1}^{n_2} Y_{2,j}\right) \\ \square &= V\left(\sum_{i=1}^{n_1} Y_{1,i}\right) + V\left(\sum_{j=1}^{n_2} Y_{2,j}\right) \\ \square &= \sum\nolimits_{i=1}^{n_1} V(Y_{1,i}) + \sum\nolimits_{j=1}^{n_2} V(Y_{2,j}) \\ \square &= n_1V(Y_1) + n_2V(Y_2) \end{aligned}$ |

```
n1 <- 120
n2 <- 200

#Values from part 2
p1 <- p[1] #0.6
p2 <- p[2] #0.8125
m1 <- m[1] #2.155
m2 <- m[2] #1.085
v1 <- v[1] #1.263
v2 <- v[2] #1.584

EY1 <- p1*m1 #1.293
VY1 <- p1*v1 + p1*(1-p1)*m1^2 #1.872692
EY2 <- p2*m2 #0.8816
VY2 <- p2*v2 + p2*(1-p2)*m2^2 #1.466173
ETot <- n1*EY1 + n2*EY2 #331.515
VTot <- n1*VY1 + n2*VY2 #517.9577

qnorm( 0.80, mean=ETot, sd=sqrt(VTot) )
[1] 350.6692
```

The hospital should budget approximately $351,000.

## Part 5

**Commented [KAM5]:** 25 points

**Carry out the budget calculation in 4 using simulation.** Assume that the Z value (the cost per patient) for each procedure group is a *gamma* distributed random variable with E(Z) = m and Var(Z) = v, specific to each procedure group. The random variable R is as above – Bernoulli with p = Pr(R = 1) as estimated per group in Part 2.

Write R code to simulate 10,000 sets of data to answer the budget question in Part 4, and give a numerical answer for the budget.

Hint: *n* gamma distributed observations can be simulated in R using the following code.

```
        Shape (beta) <- m^2/v
        Scale (alpha) <- v/m
        Z <- rgamma(n, shape = shape, scale = scale)
```

```
Answer:
> n.iter <- 100000
> Tot <- rep(NA, n.iter)
> set.seed(515)
>
> for( i in 1:n.iter ){
+   alpha1 <- m1^2/v1
+   beta1  <- v1/m1
+   Z1 <- rgamma(n1, shape=alpha1, scale=beta1)
+   R1 <- rbinom(n1, 1, p1)
+   Y1 <- R1*Z1
+   alpha2 <- m2^2/v2
+   beta2  <- v2/m2
+   Z2 <- rgamma( n2, shape=alpha2, scale=beta2 )
+   R2 <- rbinom(n2, 1, p2)
+   Y2 <- R2*Z2
+   Tot[i] <- sum(Y1) + sum(Y2)
+ }
> quantile(Tot, 0.8)
     80%
350.6266
```

The hospital should budget approximately $351,000.

## Part 6 – Extra Credit

**Commented [KAM6]:** 2 points, 1 for part a and 1 for part b

**Compare your results in Part 4 and Part 5.**

a. In general, under what circumstances will these values be similar?

   These values will be similar when sample size is large (gamma ~ normal with sample size large enough because CLT)

**b.** What level of accuracy do you estimate can be achieved for each of the groups using an approximation via simulation instead of deriving the theoretical properties of the random variable?

   Based on the above problem, the two values only differed by 0.0066, which is 0.00188% (0.0066/350.6692) percent different, so I would say there is a very high level of accuracy using an approximation.