Rohit Wason
Spring 2021

Math 560 Homework (#5, Binomial)

Problem 1. The probability distribution of a random variable X is given, and X_1 and X_2 are random samples of size 2. Assuming X_1 and X_2 are independent:

Solution. a) Since the variables are independent, the 9 possible pairs (X_1, X_2) and their probabilities are:

1. $P(0,0) = P(X_1 = 0 \wedge X_2 = 0) = P(X_1 = 0)P(X_2 = 0) = 0.6^2 = 0.36$
2. Similarly $P(0,1) = 0.6 * 0.3 = 0.18$
3. $P(0,2) = 0.6 * 0.1 = 0.06$
4. $P(1,0) = P(0,1) = 0.18$
5. $P(1,1) = 0.3^2 = 0.09$
6. $P(1,2) = 0.3 * 0.1 = 0.03$
7. $P(2,0) = P(0,2) = 0.06$
8. $P(2,1) = P(1,2) = 0.03$, and
9. $P(2,2) = 0.1^2 = 0.01$

b) As for the random variable $Y = \frac{X_1+X_2}{2}$ the different possible values are:

Y	0	1	2
0	0	0.5	1
1	0.5	1	1.5
2	1	1.5	2

Therefore the distribution for Y, obtained by adding the (X_1, X_2) probabilities from above, is:

Y	0	0.5	1	1.5	2
$P(Y)$	0.36	0.36	0.21	0.06	0.01

□

Problem 2. Population data on StatVillage (a hypothetical 128-block village in Canada) is given in the tab-delimited data file `StatVillage.txt`. The variables are listed in the first line of the data file, and information about the variables included in the file is given in the file `codesForStatVillage.txt`. Use `R` or other computer software to answer the following questions:

```
population = read.table(file ="StatVillage.txt",
  header = TRUE)
```

Solution. (a) The variable labeled `TOTINCH` gives the total household income. Determine the proportion of households in this population with a total household income greater than $100,000$.

```
> nrow( population [ population$TOTINCH>100000 ,])
  /nrow( population )
[1] 0.1113281
```

$\therefore \boxed{p \approx 0.1113}$

(b) If 100 households are selected at random with replacement from this population, what is the probability that at least 10 of the households in the sample will have a total household income greater than $100,000$? Compute the exact answer, rounded to at least 4 decimal places.

Adding all probabilities from $X = 10, 11, \ldots, 100$ in R

```
prob=0;
for (n in c(10:100))
  prob = prob + dbinom(n, size=100, prob=0.1113);
prob;
```

we get $\boxed{P(X \geq 10) = 0.6868}$

(c) If 100 households are selected at random with replacement from this population, then let X be the number of households in the sample with income above $100,000$. What is the mean and the standard deviation of the sampling distribution of X?

X follows $\approx B(n,p) = B(100, 0.11)$ from above. According to CLT, this also $\approx N(np, \sqrt{np(1-p)}) = N(11.13, \sqrt{11.13(0.8887)}) = N(11.13, 3.1450)$

Or $\boxed{\mu = 11.13, \sigma = 3.1450}$

(d) Use the central limit theorem with a continuity correction to approximate the probability computed in part (b).

We need to approximate $P(X \geq 10)$ which, with continuity correction really is $P(X \geq 9.5)$ since each bar in the histogram is centered on an integer and actually spans from the $\frac{1}{2}$ or the preceeding integer.

Using $\mu = 11.13, \sigma = 3.1450$ from above, for $Z = \frac{X-11.13}{3.1450}$, the required probability,

$$\begin{aligned} P(X \geq 9.5) &= P(\frac{X - 11.13}{3.1450} \geq \frac{9.5 - 11.13}{3.1450}) \\ &= P(Z \geq \frac{-1.63}{3.1450} = -0.5183) \\ &\approx 1 - P(Z < -0.52) \\ &= 1 - 0.3015 \\ &= \boxed{0.6985} \end{aligned}$$

□

Problem 3. Your friend generated 5 observations from a normal distribution using a random number generator in R. Your friend remembers that the standard deviation was $\sigma = 2$ but forgets what was used for the mean μ. Your friend generated and summarized the data using the following commands:

```
> set.seed(62341)
> mu=####
> sigma=2
> x=rnorm(5,mean=mu,sd=sigma)
> x
[1] 8.742144 10.503939 7.301674 6.524349 8.542242
```

```
> mean(x)
[1] 8.32287
> sd(x)
[1] 1.521389
```

Solution. Given:
$n = 5$
$\mu_{\bar{X}} = 8.32287$
$\sigma_{\bar{X}} = 1.521389$
$\sigma = 2$

(a) The probability for producing the interval, $C = 0.90$
Looking up the Normal distribution table for $0.90 + \frac{1-0.90}{2} = 0.95$, we get
$C = 0.90 \implies z^* = 1.645$
Margin of error, $(z^*\sigma/\sqrt{n}) = 1.645(2)/\sqrt{5} = 1.4713$

$\therefore$ the 90% two-sided confidence interval for μ, $(\mu_{\bar{X}} \pm 1.4713) = \boxed{8.32287 \pm 1.4713}$

(b) The probability for producing the interval, $C = 0.995$
Looking up the Normal distribution table for $0.995 + \frac{1-0.995}{2} = 0.9975$, we get
$C = 0.995 \implies z^* = 2.81$
Margin of error $= 2.81(2)/\sqrt{5} = 2.5133$

$\therefore$ the 99.5% two-sided confidence interval for $\mu = \boxed{8.32287 \pm 2.5133}$

(c) By increasing the confidence level from 90% to 99.5%, we saw that the interval for estimating μ expanded. It seems that a 90% confidence interval is a range of values that we can be 90% certain contains the true mean of the population. As the confidence level shrinks, we are less and less confident that the population mean could lie in the interval.

It's easy to see that the two-sided 100% confidence interval is the whole normal distribution (μ must lie somewhere on the graph!). □