Assignment

Bill Ang Li

March 28th, 2019

1. Suppose you have a population of size 5 [i.e. N=5]. You measure some quantity (X) and the corresponding numbers are: 21, 22, 23, 24, 25

a) Calculate the population mean (μ)

```
data = c(21, 22, 23, 24, 25)
mu = mean(data)
mu
```

```
## [1] 23
```

b) Calculate the population variance (σ^2) using the formula $\sigma^2 = \frac{\sum_{j=1}^{N}(X_j - \mu)^2}{N}$

```
var = sum((data - mu) ^ 2) / length(data)
var
```

```
## [1] 2
```

2. Imagine you are taking samples (of size n = 3) from this population with replacement. Recall: "sampling WITH replacement" ensures independence.

a) Write down every possible way that you could have a sample of size 3 with replacement from this population. (hint: there will $555 = 125$ possible combinations)

```
tbl = expand.grid(c(21:25), c(21:25), c(21:25))
tbl
```

```
##     Var1 Var2 Var3
## 1     21   21   21
## 2     22   21   21
## 3     23   21   21
## 4     24   21   21
## 5     25   21   21
## 6     21   22   21
## 7     22   22   21
## 8     23   22   21
## 9     24   22   21
## 10    25   22   21
## 11    21   23   21
## 12    22   23   21
## 13    23   23   21
## 14    24   23   21
## 15    25   23   21
## 16    21   24   21
## 17    22   24   21
## 18    23   24   21
## 19    24   24   21
## 20    25   24   21
## 21    21   25   21
## 22    22   25   21
## 23    23   25   21
## 24    24   25   21
```

```
## 25     25   25   21
## 26     21   21   22
## 27     22   21   22
## 28     23   21   22
## 29     24   21   22
## 30     25   21   22
## 31     21   22   22
## 32     22   22   22
## 33     23   22   22
## 34     24   22   22
## 35     25   22   22
## 36     21   23   22
## 37     22   23   22
## 38     23   23   22
## 39     24   23   22
## 40     25   23   22
## 41     21   24   22
## 42     22   24   22
## 43     23   24   22
## 44     24   24   22
## 45     25   24   22
## 46     21   25   22
## 47     22   25   22
## 48     23   25   22
## 49     24   25   22
## 50     25   25   22
## 51     21   21   23
## 52     22   21   23
## 53     23   21   23
## 54     24   21   23
## 55     25   21   23
## 56     21   22   23
## 57     22   22   23
## 58     23   22   23
## 59     24   22   23
## 60     25   22   23
## 61     21   23   23
## 62     22   23   23
## 63     23   23   23
## 64     24   23   23
## 65     25   23   23
## 66     21   24   23
## 67     22   24   23
## 68     23   24   23
## 69     24   24   23
## 70     25   24   23
## 71     21   25   23
## 72     22   25   23
## 73     23   25   23
## 74     24   25   23
## 75     25   25   23
## 76     21   21   24
## 77     22   21   24
## 78     23   21   24
```

```
## 79    24   21   24
## 80    25   21   24
## 81    21   22   24
## 82    22   22   24
## 83    23   22   24
## 84    24   22   24
## 85    25   22   24
## 86    21   23   24
## 87    22   23   24
## 88    23   23   24
## 89    24   23   24
## 90    25   23   24
## 91    21   24   24
## 92    22   24   24
## 93    23   24   24
## 94    24   24   24
## 95    25   24   24
## 96    21   25   24
## 97    22   25   24
## 98    23   25   24
## 99    24   25   24
## 100   25   25   24
## 101   21   21   25
## 102   22   21   25
## 103   23   21   25
## 104   24   21   25
## 105   25   21   25
## 106   21   22   25
## 107   22   22   25
## 108   23   22   25
## 109   24   22   25
## 110   25   22   25
## 111   21   23   25
## 112   22   23   25
## 113   23   23   25
## 114   24   23   25
## 115   25   23   25
## 116   21   24   25
## 117   22   24   25
## 118   23   24   25
## 119   24   24   25
## 120   25   24   25
## 121   21   25   25
## 122   22   25   25
## 123   23   25   25
## 124   24   25   25
## 125   25   25   25
```

```r
x_bar = rowMeans(tbl)
tbl = cbind(tbl, x_bar)
tbl
```

```
##     Var1 Var2 Var3    x_bar
## 1     21   21   21 21.00000
## 2     22   21   21 21.33333
```

```
## 3      23   21   21 21.66667
## 4      24   21   21 22.00000
## 5      25   21   21 22.33333
## 6      21   22   21 21.33333
## 7      22   22   21 21.66667
## 8      23   22   21 22.00000
## 9      24   22   21 22.33333
## 10     25   22   21 22.66667
## 11     21   23   21 21.66667
## 12     22   23   21 22.00000
## 13     23   23   21 22.33333
## 14     24   23   21 22.66667
## 15     25   23   21 23.00000
## 16     21   24   21 22.00000
## 17     22   24   21 22.33333
## 18     23   24   21 22.66667
## 19     24   24   21 23.00000
## 20     25   24   21 23.33333
## 21     21   25   21 22.33333
## 22     22   25   21 22.66667
## 23     23   25   21 23.00000
## 24     24   25   21 23.33333
## 25     25   25   21 23.66667
## 26     21   21   22 21.33333
## 27     22   21   22 21.66667
## 28     23   21   22 22.00000
## 29     24   21   22 22.33333
## 30     25   21   22 22.66667
## 31     21   22   22 21.66667
## 32     22   22   22 22.00000
## 33     23   22   22 22.33333
## 34     24   22   22 22.66667
## 35     25   22   22 23.00000
## 36     21   23   22 22.00000
## 37     22   23   22 22.33333
## 38     23   23   22 22.66667
## 39     24   23   22 23.00000
## 40     25   23   22 23.33333
## 41     21   24   22 22.33333
## 42     22   24   22 22.66667
## 43     23   24   22 23.00000
## 44     24   24   22 23.33333
## 45     25   24   22 23.66667
## 46     21   25   22 22.66667
## 47     22   25   22 23.00000
## 48     23   25   22 23.33333
## 49     24   25   22 23.66667
## 50     25   25   22 24.00000
## 51     21   21   23 21.66667
## 52     22   21   23 22.00000
## 53     23   21   23 22.33333
## 54     24   21   23 22.66667
## 55     25   21   23 23.00000
## 56     21   22   23 22.00000
```

```
## 57     22   22   23 22.33333
## 58     23   22   23 22.66667
## 59     24   22   23 23.00000
## 60     25   22   23 23.33333
## 61     21   23   23 22.33333
## 62     22   23   23 22.66667
## 63     23   23   23 23.00000
## 64     24   23   23 23.33333
## 65     25   23   23 23.66667
## 66     21   24   23 22.66667
## 67     22   24   23 23.00000
## 68     23   24   23 23.33333
## 69     24   24   23 23.66667
## 70     25   24   23 24.00000
## 71     21   25   23 23.00000
## 72     22   25   23 23.33333
## 73     23   25   23 23.66667
## 74     24   25   23 24.00000
## 75     25   25   23 24.33333
## 76     21   21   24 22.00000
## 77     22   21   24 22.33333
## 78     23   21   24 22.66667
## 79     24   21   24 23.00000
## 80     25   21   24 23.33333
## 81     21   22   24 22.33333
## 82     22   22   24 22.66667
## 83     23   22   24 23.00000
## 84     24   22   24 23.33333
## 85     25   22   24 23.66667
## 86     21   23   24 22.66667
## 87     22   23   24 23.00000
## 88     23   23   24 23.33333
## 89     24   23   24 23.66667
## 90     25   23   24 24.00000
## 91     21   24   24 23.00000
## 92     22   24   24 23.33333
## 93     23   24   24 23.66667
## 94     24   24   24 24.00000
## 95     25   24   24 24.33333
## 96     21   25   24 23.33333
## 97     22   25   24 23.66667
## 98     23   25   24 24.00000
## 99     24   25   24 24.33333
## 100    25   25   24 24.66667
## 101    21   21   25 22.33333
## 102    22   21   25 22.66667
## 103    23   21   25 23.00000
## 104    24   21   25 23.33333
## 105    25   21   25 23.66667
## 106    21   22   25 22.66667
## 107    22   22   25 23.00000
## 108    23   22   25 23.33333
## 109    24   22   25 23.66667
## 110    25   22   25 24.00000
```

```
## 111   21   23   25 23.00000
## 112   22   23   25 23.33333
## 113   23   23   25 23.66667
## 114   24   23   25 24.00000
## 115   25   23   25 24.33333
## 116   21   24   25 23.33333
## 117   22   24   25 23.66667
## 118   23   24   25 24.00000
## 119   24   24   25 24.33333
## 120   25   24   25 24.66667
## 121   21   25   25 23.66667
## 122   22   25   25 24.00000
## 123   23   25   25 24.33333
## 124   24   25   25 24.66667
## 125   25   25   25 25.00000
```

3. You should have noticed that the values in the "X bar" column are repetitive. For example, 21.3333333 will show up 3 times.

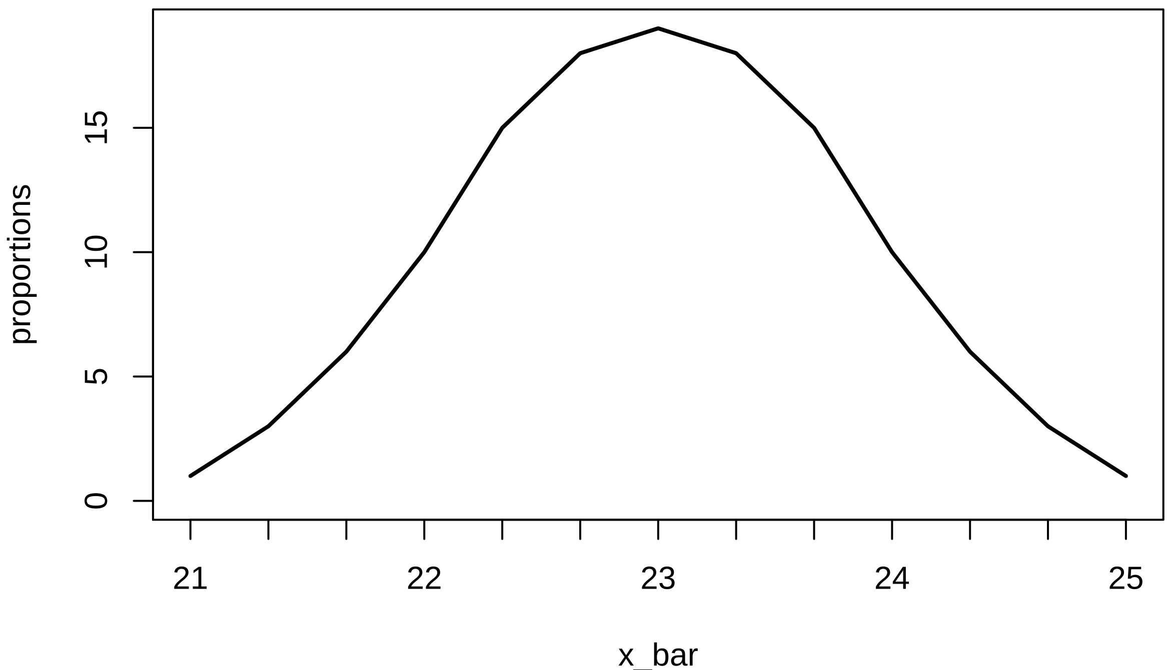
a) Construct a frequency table based on the column "X bar". [i.e. write down which values showed up how many times]. Now using the frequencies (also known as counts) calculate proportion of each of those repeated values. [For example: proportion of 21.3333333 will be 3/125]

```r
proportions = table(x_bar)
proportions
```

```
## x_bar
##              21 21.3333333333333 21.6666666666667               22
##               1                3                6               10
## 22.3333333333333 22.6666666666667               23 23.3333333333333
##              15               18               19               18
## 23.6666666666667               24 24.3333333333333 24.6666666666667
##              15               10                6                3
##              25
##               1
```

b) Plot these proportions against the values and connect the points using a non-linear line. (it will look like a density plot). Does the shape of this plot look like any known distribution?

```r
plot(proportions, type="l")
```



It looks like a normal distribution.

c) Using the table of proportions or otherwise, calculate the mean of these 125 numbers and compare it to your answer of 1(a).

```
mean(x_bar)
```

```
## [1] 23
```

It is the same as 1(a), so it shows sample mean is a good estimator of population mean.

d) Using the table of proportions or otherwise, calculate the variance of these 125 numbers. Use thepopulation variance formula (i.e. divide by 125 not 124). What is the relationship of this answer to your answer of 1(b)?

```
var(x_bar)
```

```
## [1] 0.672043
```

This is the sample variance, the relationship is $s^2 = \frac{\sigma^2}{n}$, where $n = 3$ is the sample size.

e) Which theorem did you demonstrate empirically in part b, c and d?

Central Limit Theorem.

For each of these sample of size 3, calculate the sample variance using the following two formulas

$$S^2 = \frac{1}{n-1}\sum(X_i - \bar{X})^2$$

and

$$\hat{\sigma}^2 = \frac{1}{n}\sum(X_i - \bar{X})^2$$

Assume the population variance, $\sigma^2 = 2$.

a) By calculating (numerically using the 125 different values) Bias[S^2] and Bias[σ^2] check the unbiasedness of these two estimators.

```
s_2 = ((tbl[1] - tbl[4]) ^ 2 + (tbl[2] - tbl[4]) ^ 2 + (tbl[3] - tbl[4]) ^ 2) / 2
var_2 = ((tbl[1] - tbl[4]) ^ 2 + (tbl[2] - tbl[4]) ^ 2 + (tbl[3] - tbl[4]) ^ 2) / 3
sprintf("Bias[s^2]: %s, Bias[var^2]: %s", sum(s_2) / 125 - 2, sum(var_2) / 125 - 2)
```

```
## [1] "Bias[s^2]: 0, Bias[var^2]: -0.666666666666667"
```

b) By calculating all three components separately check the following identity

$$MSE[\hat{\sigma}^2] = var[\hat{\sigma}^2] + (Bias[\hat{\sigma}^2])^2$$

```
mse = sum((var_2 - 2) ^ 2) / 125
var = var(var_2)
bias_squared = (sum(var_2) / 125 - 2) ^ 2
mse - var - bias_squared
```

```
##              Var1
## Var1 -0.008124253
```

It is pretty much zero.

5. Even though we need sample size n to be large to apply central limit theorem, but let's apply it anyway. Suppose you know that the population variance, $\sigma^2 = 2$.

a) For each of these 125 cases, calculate a 95% confidence interval and finally calculate the proportion of the intervals that includes $\mu = 23$.

```
me = qnorm(0.975) * sqrt(2 / 3)
lower = tbl[4] - me
upper = tbl[4] + me
sum(lower <= 23 & upper >= 23) / 125
```

```
## [1] 0.936
```

b) Suppose someone observes only one of these 125 combinations (23,24,25). If that person is testing the null hypothesis $H_0 : \mu = 23$, based on this observed sample calculate the p-value that the person will get using central limit theorem.

```
data = c(23, 24, 25)
test_statistic = (mean(data) - 23) / sqrt(2 /3)
p = 2 * (1 - pnorm(test_statistic))
p
```

```
## [1] 0.2206714
```

c) Calculate the p-value numerically using the 125 $\bar{X}$ values that you calculated in part 2(b) (do not use CLT here).

```
2 * sum(x_bar >= mean(data)) / 125
```

```
## [1] 0.32
```

Quiz Questions

```
quantile(x_bar)
```

```
##       0%      25%      50%      75%     100%
## 21.00000 22.33333 23.00000 23.66667 25.00000
```