**IST-772 Quantitative Reasoning in Data science**

Week3/HW-3: Sampling Distributions

**Probabilities in the long run (Page-50, 51: Problems:2-7)**

Sharat Sripada ([vssripad@syr.edu](mailto:vssripad@syr.edu))

1. Use R’s built-in dataset ChickWeight. Capture and explain output of summary() and shape() functions

> cw <- ChickWeight

> summary(cw)

weight Time Chick Diet

Min. : 35.0 Min. : 0.00 13 : 12 1:220

1st Qu.: 63.0 1st Qu.: 4.00 9 : 12 2:120

Median :103.0 Median :10.00 20 : 12 3:120

Mean :121.8 Mean :10.72 10 : 12 4:118

3rd Qu.:163.8 3rd Qu.:16.00 17 : 12

Max. :373.0 Max. :21.00 19 : 12

(Other):506

The summary function above shows the min, 1st quartile, Median, Mean, 3rd quartile and max values for each variable or column viz. weight, Time, Chick and Diet:

* Weight (in grams): Min/Max = 35/373, 1st/3rd quartile = 63/163.8, Median/Mean = 103/121.8
* Time (in days): Min/Max = 0/21, 1st/3rd quartile = 4/16, Median/Mean = 10/10.72
* Chick variable or column is type factor
* Diet variable or column is type factor comprising 4-levels viz. 1-4

> dim(cw)

[1] 578 4

The output of dim(cw) shows 578 rows and 4 columns.

1. Run the following commands. Report and explain the output of each:

summary(ChickWeight$weight)

head(ChickWeight$weight)

mean(ChickWeight$weight)

myChkWts <- ChickWeight$weight

quantile(myChkWts, 0.5)

> summary(cw)

weight Time Chick Diet

Min. : 35.0 Min. : 0.00 13 : 12 1:220

1st Qu.: 63.0 1st Qu.: 4.00 9 : 12 2:120

Median :103.0 Median :10.00 20 : 12 3:120

Mean :121.8 Mean :10.72 10 : 12 4:118

3rd Qu.:163.8 3rd Qu.:16.00 17 : 12

Max. :373.0 Max. :21.00 19 : 12

(Other):506

*Command interpretation:* Displays the min, max, central tendencies of mean/median and quartile ranges for the dataset.

> head(cw)

weight Time Chick Diet

1 42 0 1 1

2 51 2 1 1

3 59 4 1 1

4 64 6 1 1

5 76 8 1 1

6 93 10 1 1

*Command interpretation:* Displays the fist few rows of the data-frame, cw (ChickWeight dataset).

> mean(cw$weight)

[1] 121.8183

*Command interpretation:* Calculates the mean of the weight variable or column in the ChickWeight dataset.

This value matches the Mean value for weight from the summary(cw) output.

> myChkWts <- cw$weight

*Command interpretation:* Using R’s $, access a specific variable or column and assign those values to a variable myChkWts

> quantile(myChkWts, 0.5)

50%

103

*Command interpretation:* The quantile function takes two arguments:

* Arg-1: Vector or data – here, chick weights
* Arg-2: Cut-point over the data

The output shows 50% (equivalent to 0.5) as 103. This value matches the median for weight from the summary(cw) output.

1. Create a histogram of the variable myChkWts and then write R code to display the 2.5% and 97.5% quantiles. Further explain the mean, median and shape of distribution. Describe clearly what the 2.5% and 97.5% quantiles signify

> hist(myChkWts)

Chart, histogram

Description automatically generated

The shape of the curve is not a regular bell-shape or normal distribution. Instead, it seems to be right-skewed or positive-skewed since there is a long tail in the positive direction of the number line.

> # Display quantile of 2.5% or 0.025 and 97.5% or 0.975

> quantile(myChkWts, c(0.025, 0.975))

2.5% 97.5%

41.000 294.575

> mean(myChkWts)

[1] 121.8183

> median(myChkWts)

[1] 103

The mean or average weight of chicks across 21 days is 121.82g. The median or middle point of data is 103g. Quantile output shows two values - 2.5% and 97.5% cut-points as 41g and 294.56g respectively. This means that:

* 2.5% of all values or observations lie at or below 41
* 2.5% of all values or observations lie at or above 294.56

1. Write R code that will construct sampling distribution of means from weight data. Use a sample size of 11 with replacement. Show a histogram of the distribution of sample means. Also display the 2.5% and 97.5% quantiles of the sampling distribution on the histogram with vertical line.

We can use the following R-code to achieve this:

# Sample the weights

cwweights\_mean\_sample <- replicate(1000, mean(sample(myChkWts, size=11, replace = T)))

hist(cwweights\_mean\_sample, xlim=c(50, 200))

# Using abline indicate the 2.5% and 97.5% values/cut-points on the histogram

abline(v=quantile(cwweights\_mean\_sample, c(0.025, 0.975)))

Chart, histogram

Description automatically generated

1. Briefly describe what the difference between a distribution of raw data and a distribution of sample means is. Comment why the 2.5% and 97.5% quantiles are so different.

The distribution of raw data showed a right-skewed distribution whereas, the distribution of sampling means was a normal distribution.

The reason distribution of sampling means tends to be a bell-shaped curve or normal distribution is because of the **law of large numbers** and **central limit theorem**:

* when you run a statistical process like sampling many times, it will generally converge on a particular result
* the distribution of sampling means starts to show a bell-shaped or normal distribution

Finally, when comparing the 2.5% and 97.5% quantiles we see 41, 294.56 respectively for raw data vs 83.33, 167.1 respectively for distribution of sampling means. Since the data tends to shift or converge towards the mean/median when sampling (with means), the quantiles tend to shift right correspondingly.

1. Redo the sampling with a larger size = 100. Explain the difference is quantiles for 2.5% and 97.5%.

Using n=100 for sampling size, we see the following histogram:

Chart, histogram

Description automatically generated

The corresponding quantiles at 2.5% and 97.5% are:

> quantile(cwweights\_mean\_sample, c(0.025, 0.975))

2.5% 97.5%

108.3190 136.2745

Comparing the quantiles for 2.5% and 95.5% for sample size=11 vs 100 (83.33, 167.1 vs 108.31, 136.27) we can see a further right shift towards the central mean/median - recap that the mean/median for the weights column as initially recorded via summary() was 21/103 respectively. This aligns with our understanding of central limit theorem.