The University of Hong Kong
School of Public Health

CMED6100/MMPH6002/CMED7100
Introduction to Biostatistics (Semester I)
Practical 1 Suggested solution

Descriptive statistics

1. Describe the basic characteristics of the cohort, divided by SGA/AGA status. Fill in the table, and make sure to add a table caption.

We need to convert breastfeeding status variable from numeric variable to factor variable
[Data → Manage variables in active data set → Convert numeric variables to factors [Variables: feeddur; Factor Levels: Use numbers]
[Statistics → Contingency tables → Two-way table [Row variable: feeddur/ sex; Column variable: sga; Statistics: Compute Percentages: Column Percentages]

Characteristics of the "Children of 1997" birth cohort, stratified by SGA/AGA status

Characteristic	AGA (n=600)	SGA (n=300)
Sex		
Female	293 (48.8%)	137 (45.7%)
Breastfeeding status*		
Never breastfed	293 (51.2%)	176 (62.4%)
Partially breastfed	169 (29.5%)	64 (22.7%)
Exclusively breastfed <2m	66 (11.5%)	18 (6.4%)
Exclusively breastfed ≥2m	44 (7.7%)	24 (8.5%)

*Children with missing breastfeeding status were excluded

2. Some children have unknown breastfeeding status. How should this information be represented? Complete the table.

Statistics → Summaries → Count missing observations
There are 46 children with missing breastfeeding status.
Data → Manage variables in active dataset → Recode variables [Variables to recode: feeddur; New Variable name: feeddur.rc; Enter recode directives: NA=4; Make new variable as factor: Yes]
Statistics → Contingency tables → Two-way table [Row variable: feeddur.rc; Column variable: sga]

Characteristics of the "Children of 1997" birth cohort, stratified by SGA/AGA status

Characteristic	AGA (n=600)	SGA (n=300)
Sex		
Female	293 (48.8%)	137 (45.7%)
Breastfeeding status		
Never breastfed	293 (48.8%)	176 (58.7%)
Partially breastfed	169 (28.2%)	64 (21.3%)
Exclusively breastfed <2m	66 (11.0%)	18 (6.0%)
Exclusively breastfed ≥2m	44 (7.3%)	24 (8.0%)
Missing	28 (4.7%)	18 (6.0%)

3. Give 2 possible reasons why these variables might be missing from the dataset.

   If the main caregiver of a child has been changed before the first interview at 3 months of age, he/she may not be able to provide information on breastfeeding status; incomplete interview; coding error.

4. Calculate summary measures of location and dispersion of height, weight and BMI at 7 years of age, and fill in the table by AGA/SGA status.

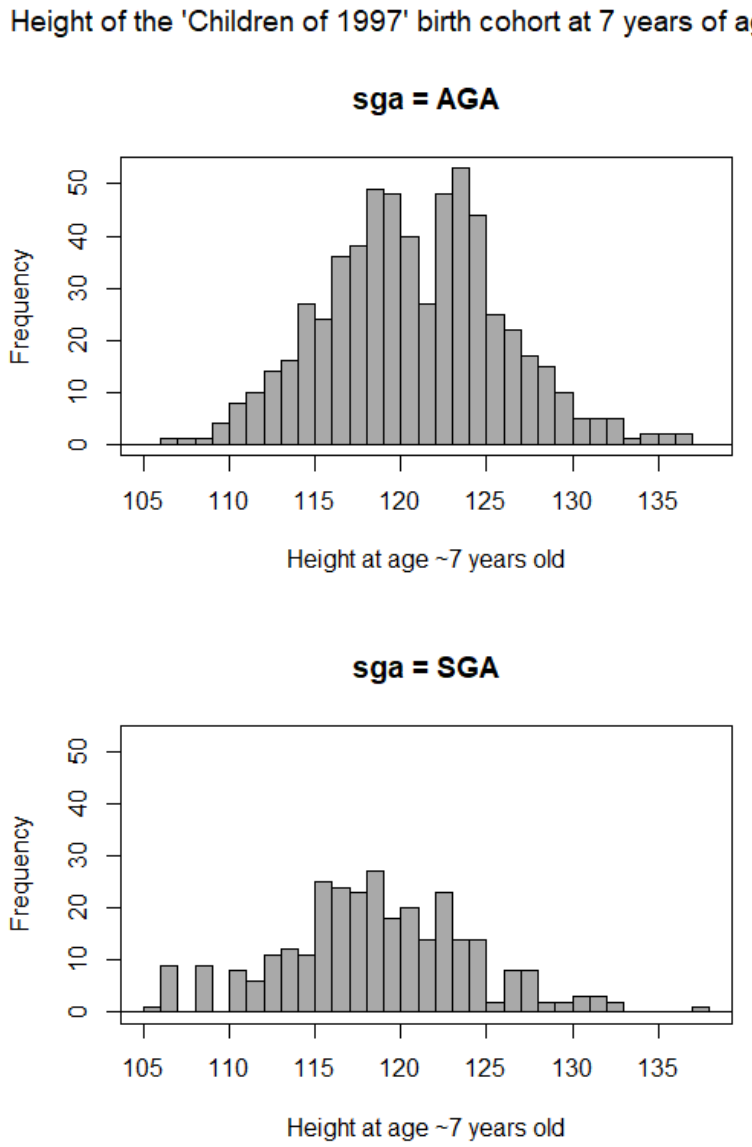
   Statistics → Summaries → Numerical summaries[Variables: hei7, bmi7, wei7; Summarize by; sga; Statistics: Mean, Standard Deviation; Quantiles: 0, .5, 1]
   Range = Max-Min

Summary statistics for the height, weight and BMI of the "Children of 1997" birth cohort, stratified by SGA/AGA status

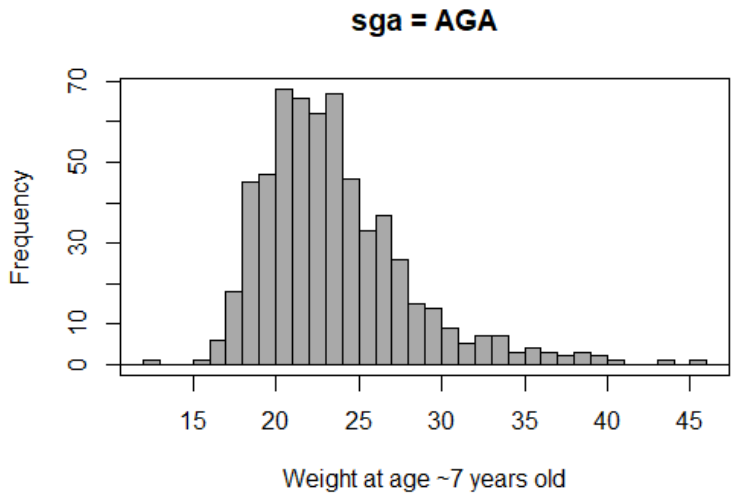
	Height (cm)		Weight (kg)		BMI (kg/m2)	
Summary statistics	**AGA**	**SGA**	**AGA**	**SGA**	**AGA**	**SGA**
Location						
Mean	120.8	118.6	23.6	21.8	16.1	15.4
Median	120.7	118.4	22.8	21.2	15.5	14.9
Dispersion						
Standard deviation	5.1	5.5	4.5	4.1	2.2	2.0
Minimum	106.2	105.8	12.6	14.8	11.0	12.1
Maximum	136.6	137.1	46.0	42.5	27.2	22.9
Range	30.4	31.3	33.4	27.7	16.2	10.8

5. Produce histograms for height, weight and BMI at 7 years of age by AGA/SGA status.

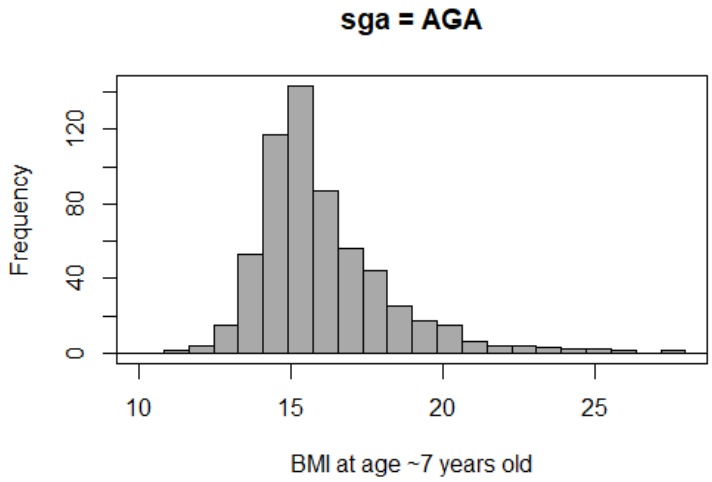
[Graphs → Histogram (Variable: hei7; Plot by: sga); Options: Number of bins: 40; Graph title: Height of the 'Children of 1997' birth cohort at 7 years of age; x-axis: Height at age ~7 years old; y-axis: Frequency]
(Similarly for wei7 and bmi7; Number of bins: wei7: 35, bmi7: 20)

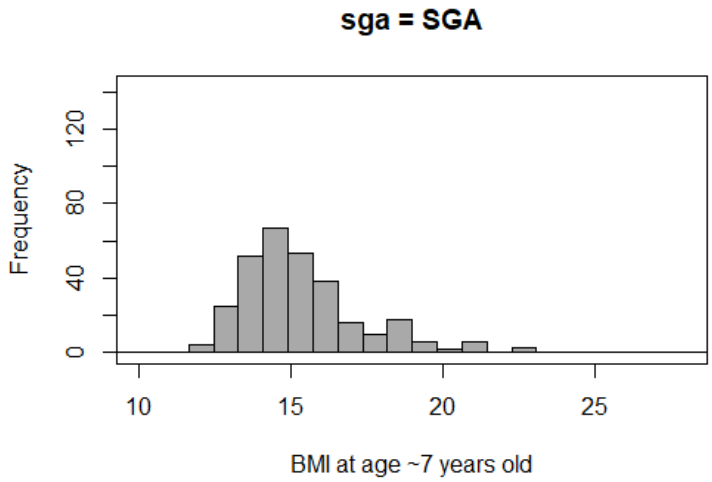


Weight of the 'Children of 1997' birth cohort at 7 years of age



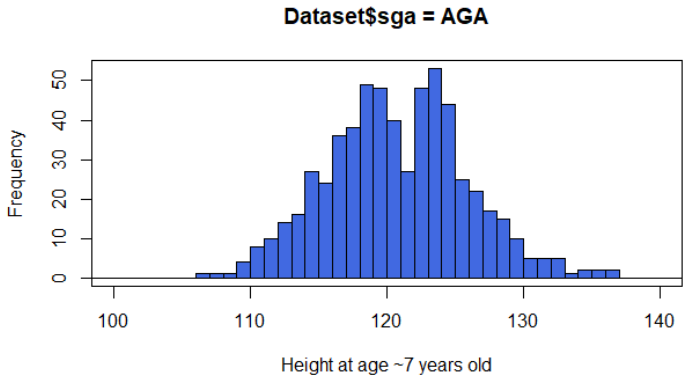
BMI of the 'Children of 1997' birth cohort at 7 years of age

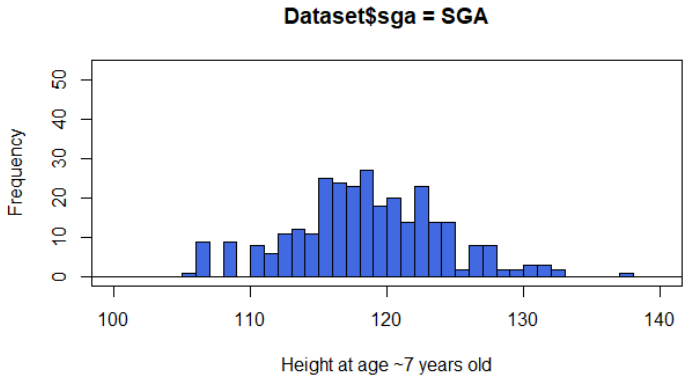




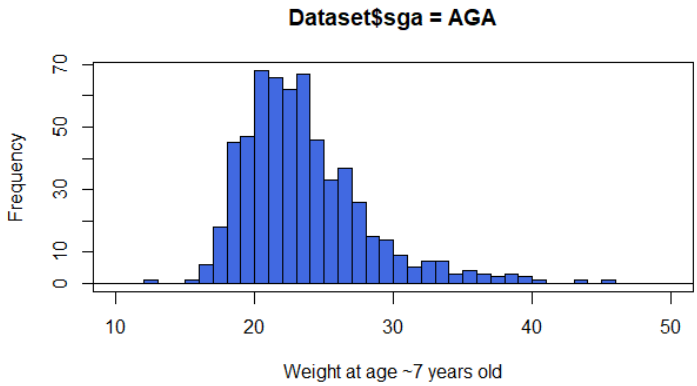
Or you can type the command on the R script panel and click 'Submit' button:
Hist(Dataset$hei7, Dataset$sga, breaks=40, col="royalblue", xlim=c(100,140), ylab="Frequency", xlab="Height at age ~7 years old", main="Height of 'Children of 1997' birth cohort at 7 years old")
Similar for weight and BMI

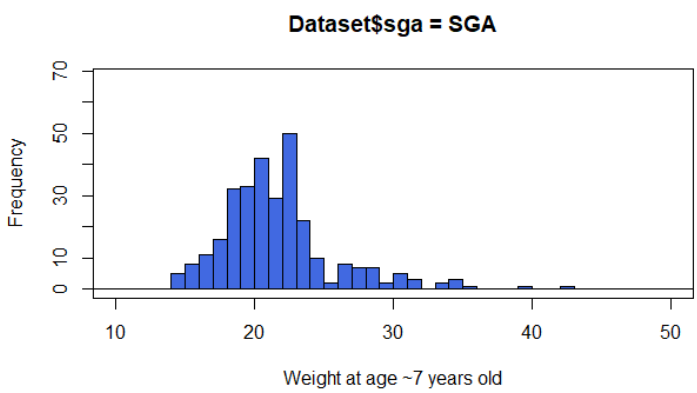
Height of 'Children of 1997' birth cohort at 7 years old

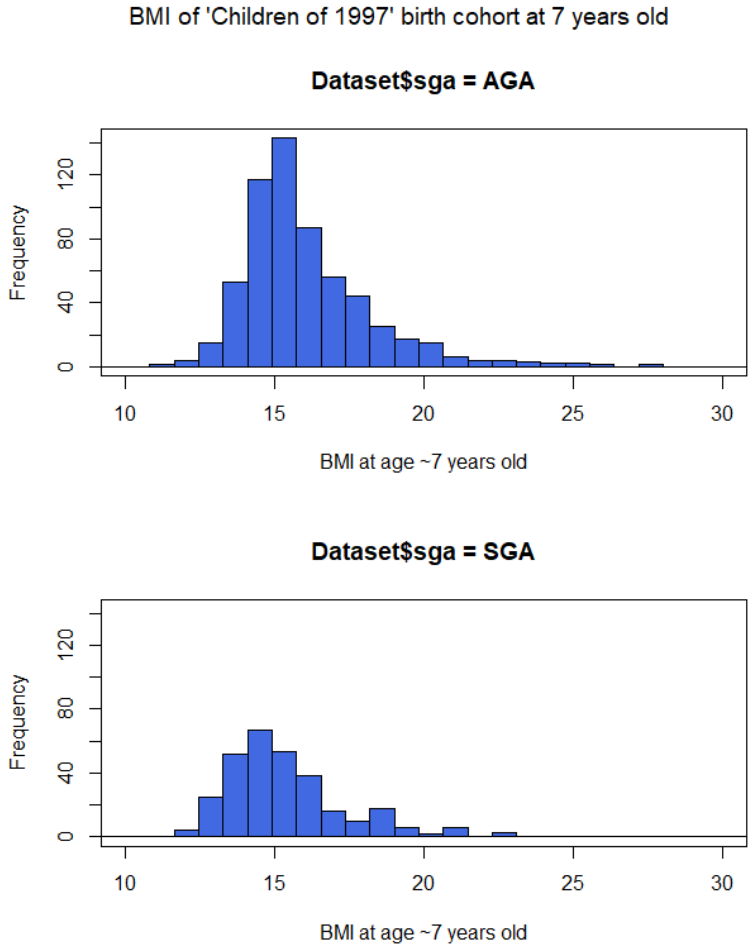




Weight of 'Children of 1997' birth cohort at 7 years old



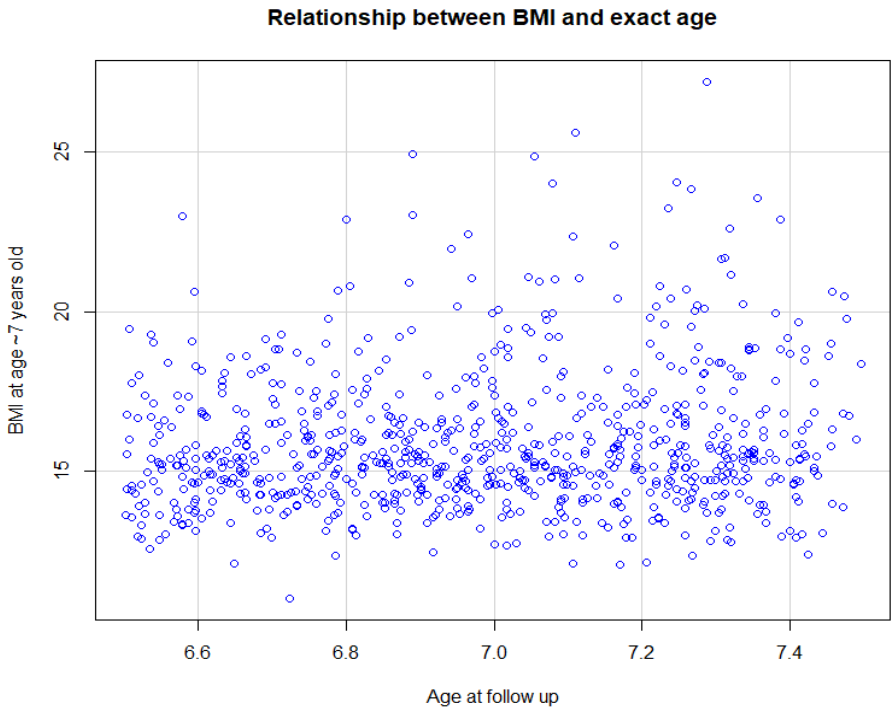




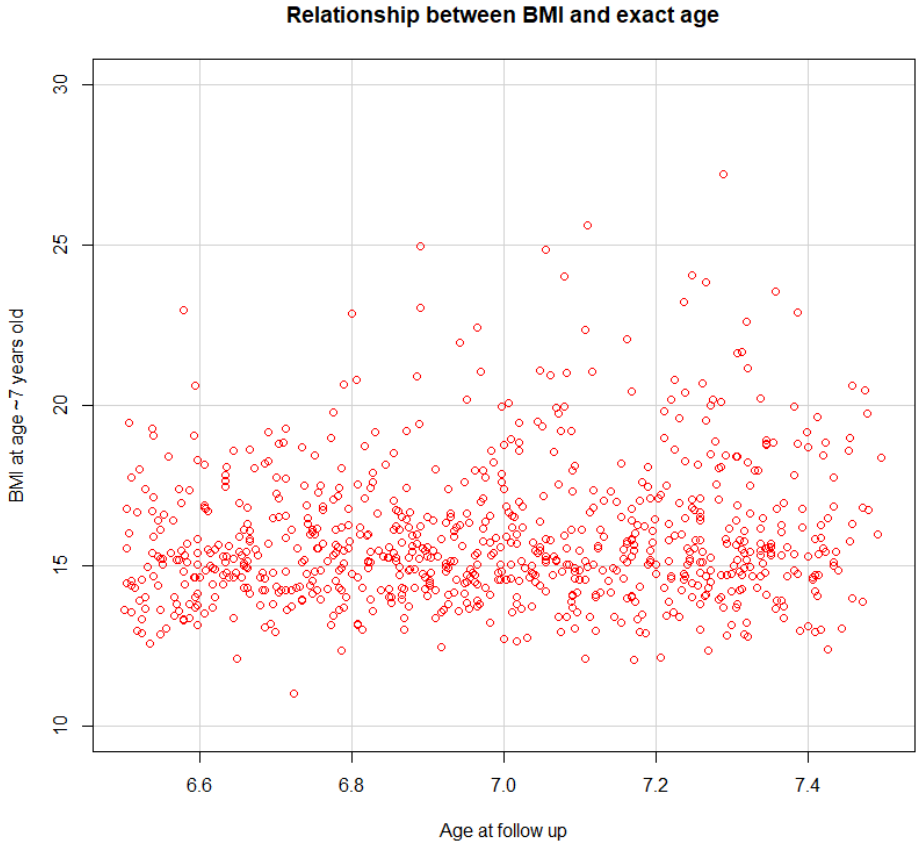
Age and childhood obesity

6. Describe the association between exact age at follow-up and BMI via the following steps.
   a. Create a scatter plot for exact age at follow-up (x-axis) versus BMI (y-axis).

   [Graphs → Scatterplot → x-variable: agefu; y-variable: bmi7; Options: x-axis: Age at follow up; y-axis: BMI at age ~7 years old; Graph title: Relationship between BMI and exact age]



Or you can type the command on the R script panel and click 'Submit' button:
scatterplot(Dataset$bmi7~Dataset$agefu, regLine=FALSE, smooth=FALSE, boxplots=FALSE, col="red", xlab= "Age at follow up", ylab= "BMI at age ~7 years old", main="Relationship between BMI and exact age")



b. Use linear regression to predict BMI from age at follow-up. Write down the equation for the regression line (Hint: the equation takes the form $y = \alpha + \beta x$ where α is the intercept and β is the slope).

[Statistics → Fit Models → Linear Regression (Response variable: bmi7; Explanatory variable: agefu)]

Coefficients	Estimates
Constant	7.654
Age at follow up	1.171

BMI = 7.654 + 1.171*Age at follow up

c. What is the mean BMI for children aged 6.5 and aged 7.5 years old (Hint: calculate by using the regression equation in (b)).

Mean BMI for children aged 6.5 years old:

$BMI_{6.5} = 7.654 + 1.171*6.5 = 15.3$

Mean BMI for children aged 7.5 years old:

$BMI_{7.5} = 7.654 + 1.171*7.5 = 16.4$

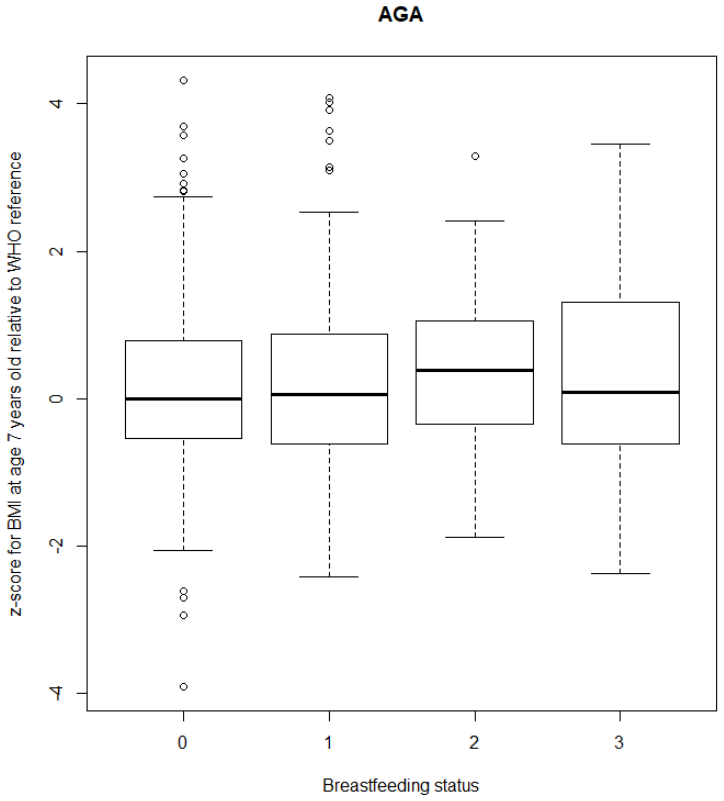
There is a positive relation between age and raw data on BMI at age 7. The effect of breastfeeding on BMI may vary in different breastfeeding status if the age distributions among the four breastfeeding status are different.

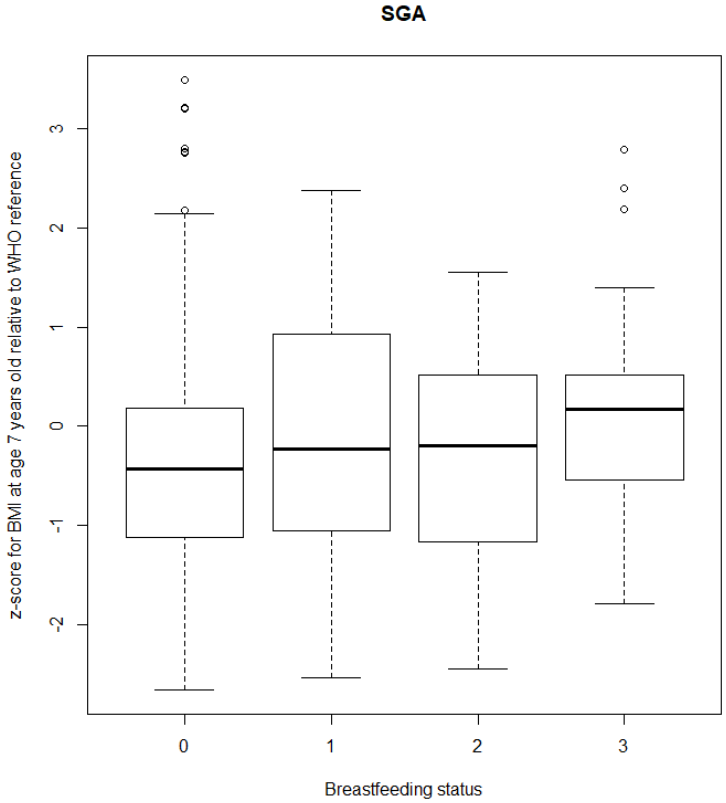
A solution to the problem is to use z-scores instead of raw BMI. You do not need to calculate z-scores for height, weight and BMI by yourself. The z-scores were included in the dataset, and were estimated to the nearest day relative to the WHO growth charts for healthy infants.

SGA, Breastfeeding, childhood obesity

7. Produce box plots of BMI versus breastfeeding status for those children classified as AGA (there should be four boxes, one for each breastfeeding category). Repeat your analysis for those classified as SGA. Describe your results.

Data → Active data set → Subset active data set [Subset expression: sga=="AGA"; Name for new data set: aga]
Graphs → Boxplot [Variable: bz7; Plot by: feeddur; Identify Outliers: No; x-axis label: Breastfeeding status; y-axis label: z-score for BMI at age 7 years old relative to WHO reference; Graph title: AGA]
Data → Active data set → Select Data Set [Data Sets: Dataset] or just click Data set at the top left of the window to select "Dataset"
Repeat for SGA


AGA
z-score for BMI at age 7 years old relative to WHO reference
4
2
0
-2
-4
0
1
2
3
Breastfeeding status

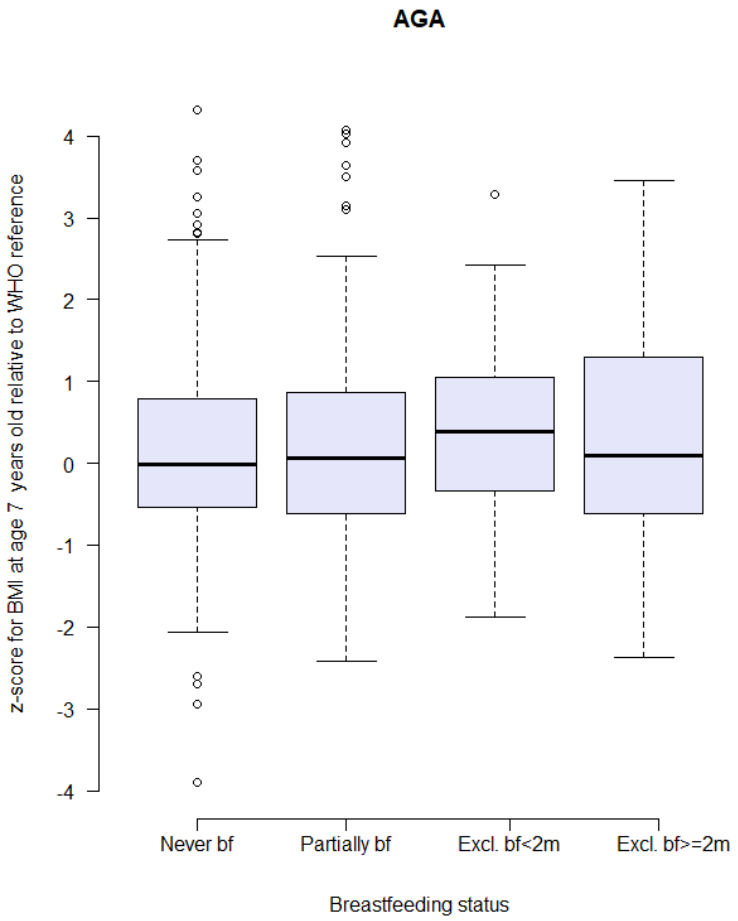

SGA
z-score for BMI at age 7 years old relative to WHO reference
3
2
1
0
-1
-2
0
1
2
3
Breastfeeding status

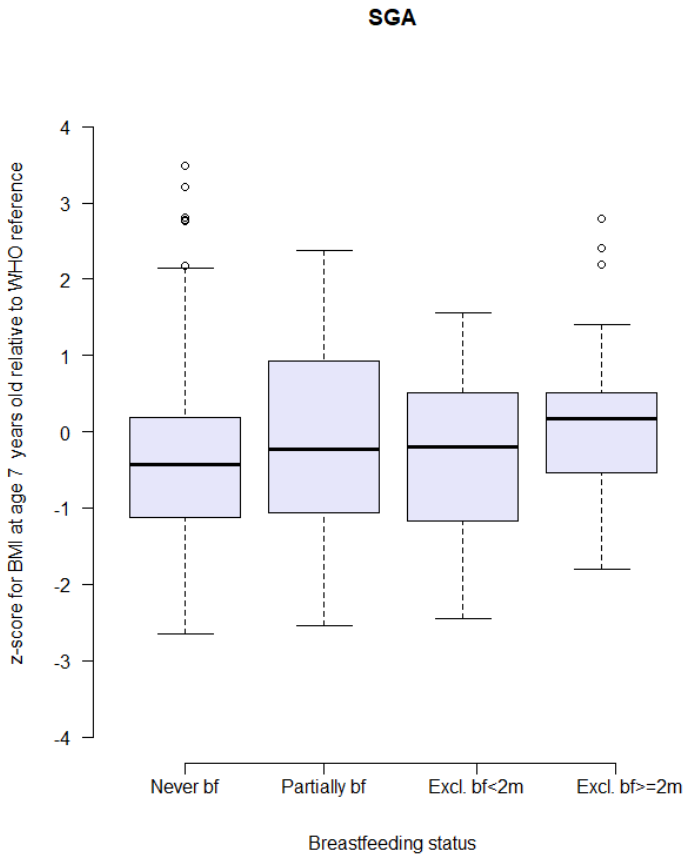
Or you can write R code on the R script panel and click 'Submit' button:

```
boxplot(Dataset$bz7[Dataset$sga==”AGA”]~ Dataset$feeddur[Dataset$sga==”AGA”], col="lavender", xlim=c(0.5,4.5), ylim=c(-4,4.5), axes=FALSE, ylab="z-score for BMI at age 7  years old relative to WHO reference", xlab="Breastfeeding status", main="AGA")
axis(1, at=1:4, labels= c("Never bf","Partially bf","Excl. bf<2m","Excl. bf>=2m"), las=1)
axis(2, at=-4:4, labels=-4:4, las=1)
```

Similar for SGA

	Explanation
xlim=, ylim=	Specifies the lower and upper limits of x-axis/ y-axis
axes=FALSE	Not showing the axes on the plot
at=	Indicating position of the labels
las=	Axes labels are parallel (=0), horizontally oriented (=1) or perpendicular (=2) to axis



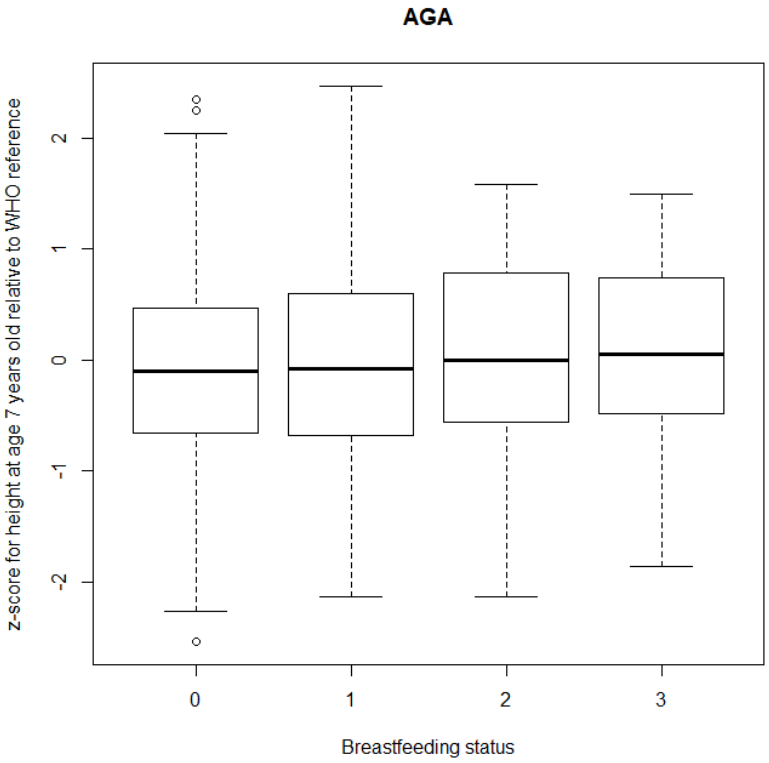


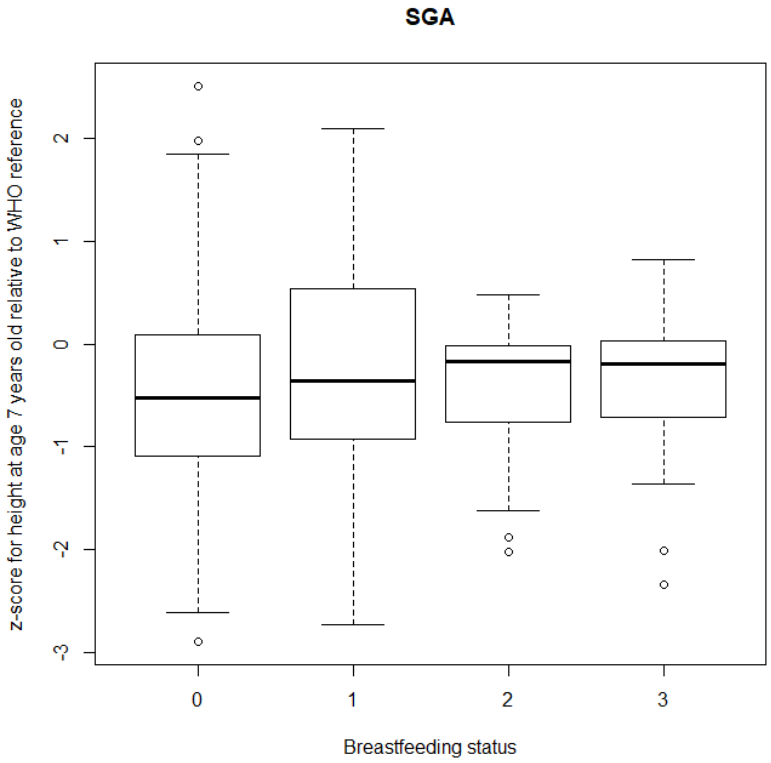
For AGA, children who are never breastfed tend to have more extreme BMI. The median BMI were similar among different breastfeeding status.

For SGA, the median BMI were higher for children who were exclusively breastfed, their BMI were also less variable when comparing with children who were never breastfed.

8. Produce box plots of height versus breastfeeding status for those classified as AGA. Repeat your analysis for those classified as SGA. Describe your results and how they relate to the findings in (7).

   Data → Active dataset → Select Data Set [Data Sets: aga]
   Graphs → Boxplot [Variable: hz7; Plot by: feeddur; Identify Outliers: No; x-axis label: Breastfeeding status; y-axis label: z-score for bmi at age 7 years old relative to WHO reference; Graph title: AGA]
   Similarly for SGA

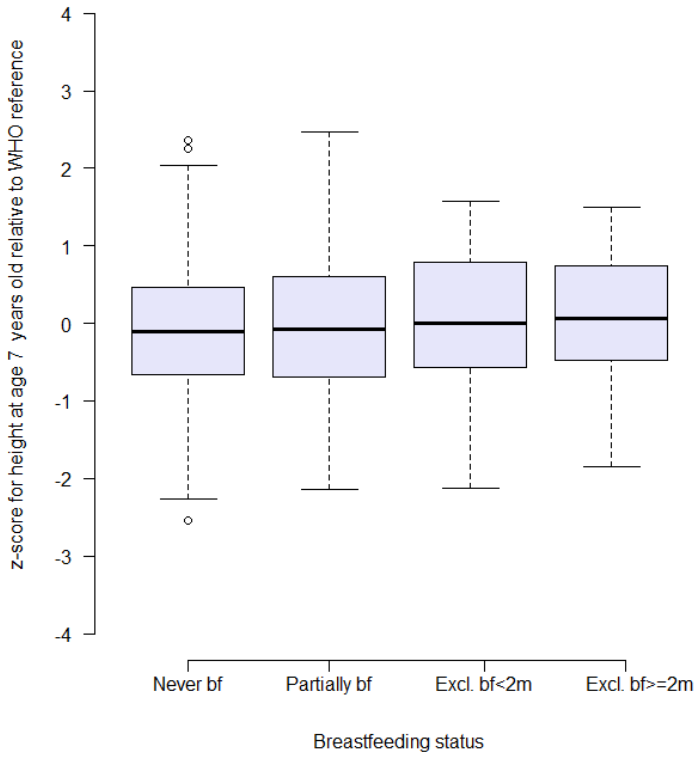




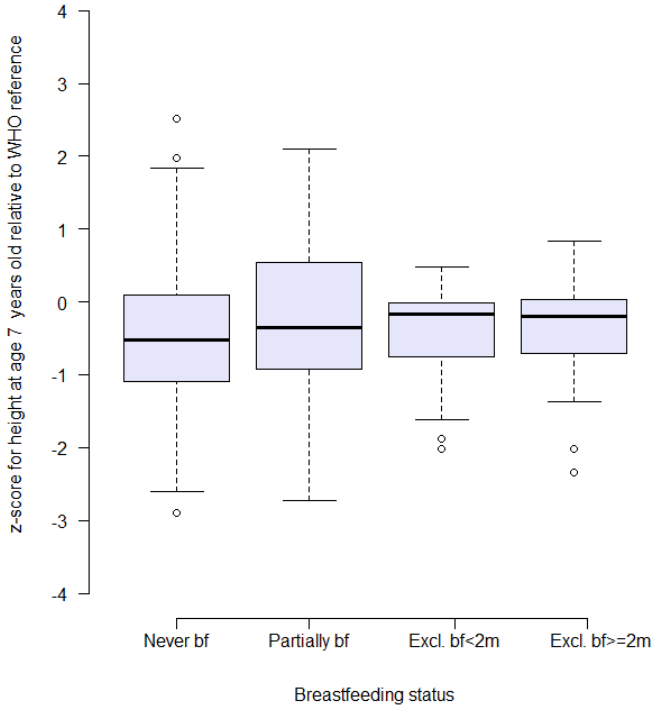
It appears height is increasing based on breastfeeding status for the AGA group while in the AGA group BMI does not seem to be related to breastfeeding status.

Or you can write R code on the R script panel and click ‘Submit’ button:
boxplot(Dataset$hz7[Dataset$sga==”AGA”]~Dataset$feeddur[Dataset$sga==”AGA”], col="lavender", xlim=c(0.5,4.5), ylim=c(-4,4.5), axes=FALSE, ylab="z-score for height at age 7 years old relative to WHO reference", xlab="Breastfeeding status", main="AGA")
axis(1, at=1:4, labels=c("Never bf","Partially bf","Excl. bf<2m","Excl. bf>=2m"), las=1)
axis(2, at=-4:4, labels=-4:4, las=1)

AGA


z-score for height at age 7 years old relative to WHO reference
4
3
2
1
0
-1
-2
-3
-4
Never bf
Partially bf
Excl. bf<2m
Excl. bf>=2m
Breastfeeding status

SGA


z-score for height at age 7 years old relative to WHO reference
4
3
2
1
0
-1
-2
-3
-4
Never bf
Partially bf
Excl. bf<2m
Excl. bf>=2m
Breastfeeding status