Sampling from a population, summarizing data, the normal distribution, t-distribution, and hypothesis testing – Solutions

Introduction to Statistical Thinking and Data Analysis

MSc in Epidemiology and MSc in Health Data Analytics, Imperial College London

10 October 2022; reviewed 17 October 2022

A) Consolidating concepts

A1. What does the standard deviation measure?

a. The spread of the middle 50% of the distribution.

b. The amount of variability in the sample mean.

c. ***The amount of variability in the population.***

d. None of the above.

A2. Which of the following is true of the standard normal distribution?

a. ***It has a mean of 0 and a standard deviation of 1.***

b. It has an area equal to 0.5.

c. It has a mean of 1 and a standard deviation of 0.

d. It cannot be used to approximate any normally distributed variable.

A3. For a normally distributed dataset, we would expect approximately:

a. 70% of observations to lie within 1 standard deviation of the mean.

b. 1 out of 20 observations to lie outside of two standard deviations of the mean.

c. A bell shaped distribution.

d. ***All of the above.***

A4. Given that **X** is a normally distributed variable with a mean of 5 and a standard deviation of 1.2, what is the probability that X is less than 6?

a. ***0.80***

b. 0.65

c. 0.90

d. 0.77

A5. What percentage of females are taller than 170cm in a population with a mean of 165cm and standard deviation of 5.5cm?

a. ***18%***

b. 12%

c. 24%

d. None of the above.

A6. What is the Z-score of the 95th quantile of the standard normal distribution?

a. ***1.64***

b. 1.96

c. 1.28

d. 0.83

A7. Which of the following is not true of the central limit theorem?

a. Provided the sample size is sufficiently large, the distribution of a sample ~~population~~ mean is approximately normal.

b. ***The underlying distribution of the population must be normal.***

c. The sample means will be normally distributed around the population mean.

d. The more you run a random experiment, the more its results will follow a normal distribution.

A8. The normal distribution is also called:

a. Poisson distribution

b. Bernoulli's distribution

c. ***Gaussian distribution***

d. Student's t

A9. The shape of the t distribution:

a. Is the same as the normal distribution.

b. Is skewed.

c. ***Depends on the number of degrees of freedom.***

d. None of the above.

A10. A 95% confidence interval for the mean of a population is such that:

a. The population mean will fall within the confidence interval 95% of the time.

b. ***If we sample the same population 100 times, 95 of the confidence intervals calculated from these random samples will contain the population mean.***

c. It contains 95% of the values in the population.

d. There is a 95% probability that it contains the population mean.

B) Practicing skills

B1. The dataset `perulung_ems.csv` contains data from a study of lung function among a sample of 636 children aged 7 to 10 years living in a deprived suburb of Lima, Peru, introduced on page 27 of Kirkwood and Sterne. FEV1 is the *forced expiratory volume* in 1 second, the maximum amount of air which children could breath out in 1 second measured using a spirometer.

Variable	Description
id	Participant ID number
fev1	Forced Expiratory Volume in 1 second
age	Age in years
height	Height in centimeters
sex	Sex (0 = female, 1 = male)
respsymptoms	Presence of respiratory symptoms (0 = no symptoms; 1 = symptoms)

a. What type of variable is each variable in the dataset?

```
perulung <- read.csv("perulung_ems.csv")
str(perulung)
```

```
## 'data.frame':    636 obs. of  6 variables:
##  $ ï..id       : int  1 2 3 4 5 6 7 8 9 10 ...
##  $ fev1        : num  1.56 1.18 1.87 1.49 1.62 2.11 1.73 1.47 1.83 1.41 ...
##  $ age         : num  9.59 7.49 9.86 8.59 8.97 ...
##  $ height      : num  125 111 136 119 121 ...
##  $ sex         : int  0 1 0 0 1 0 1 0 1 0 ...
##  $ respsymptoms: int  0 0 0 0 0 1 0 1 0 0 ...
```

- `fev1` and `age` are **continuous** numerical variables.
- `sex` and `respsymptoms` are **binary** variables although encoded in the dataframe as integers.

For the categorical variables `sex` and `respsymptoms`, add factor labels to the variables.

```
perulung$sex <- factor(perulung$sex, c(0, 1), c("female", "male"))
perulung$respsymptoms <- factor(perulung$respsymptoms, c(0, 1), c("no symptoms", "symptoms"))

str(perulung)
```

```
## 'data.frame':    636 obs. of  6 variables:
##  $ ï..id       : int  1 2 3 4 5 6 7 8 9 10 ...
##  $ fev1        : num  1.56 1.18 1.87 1.49 1.62 2.11 1.73 1.47 1.83 1.41 ...
##  $ age         : num  9.59 7.49 9.86 8.59 8.97 ...
##  $ height      : num  125 111 136 119 121 ...
##  $ sex         : Factor w/ 2 levels "female","male": 1 2 1 1 2 1 2 1 2 1 ...
##  $ respsymptoms: Factor w/ 2 levels "no symptoms",..: 1 1 1 1 1 2 1 2 1 1 ...
```

b. What shape is the frequency distribution of each of the variables?

- fev1: has a symmetrical distribution
- age: has a bimodal distribution (this is a bit surprising)
- height: has a symmetrical distribution
- sex: binary (response fairly well balanced)
- respsymptoms: binary (response imbalanced - more without symptoms)

```
par(mfrow = c(3, 3),    # panel of plots with 3 rows and 3 columns
    mar = c(2, 3, 2, 1))
hist(perulung$fev1, main = "FEV1", xlab = "")
hist(perulung$age, main = "Age")
hist(perulung$height, main = "Height", xlab = "")
plot(density(perulung$fev1), main = "FEV1", xlab = "")
plot(density(perulung$age), main = "Age", xlab = "")
plot(density(perulung$height), main = "Height", xlab = "")
barplot(table(perulung$sex), main = "Sex", ylab = "Frequency")
barplot(table(perulung$respsymptoms), main = "Respiratory symptoms", ylab = "Frequency")

## Frequency table
table(perulung$sex)
```

```
##
## female   male
##    335    301
```

```
## Frequency proportions
prop.table(table(perulung$sex))
```

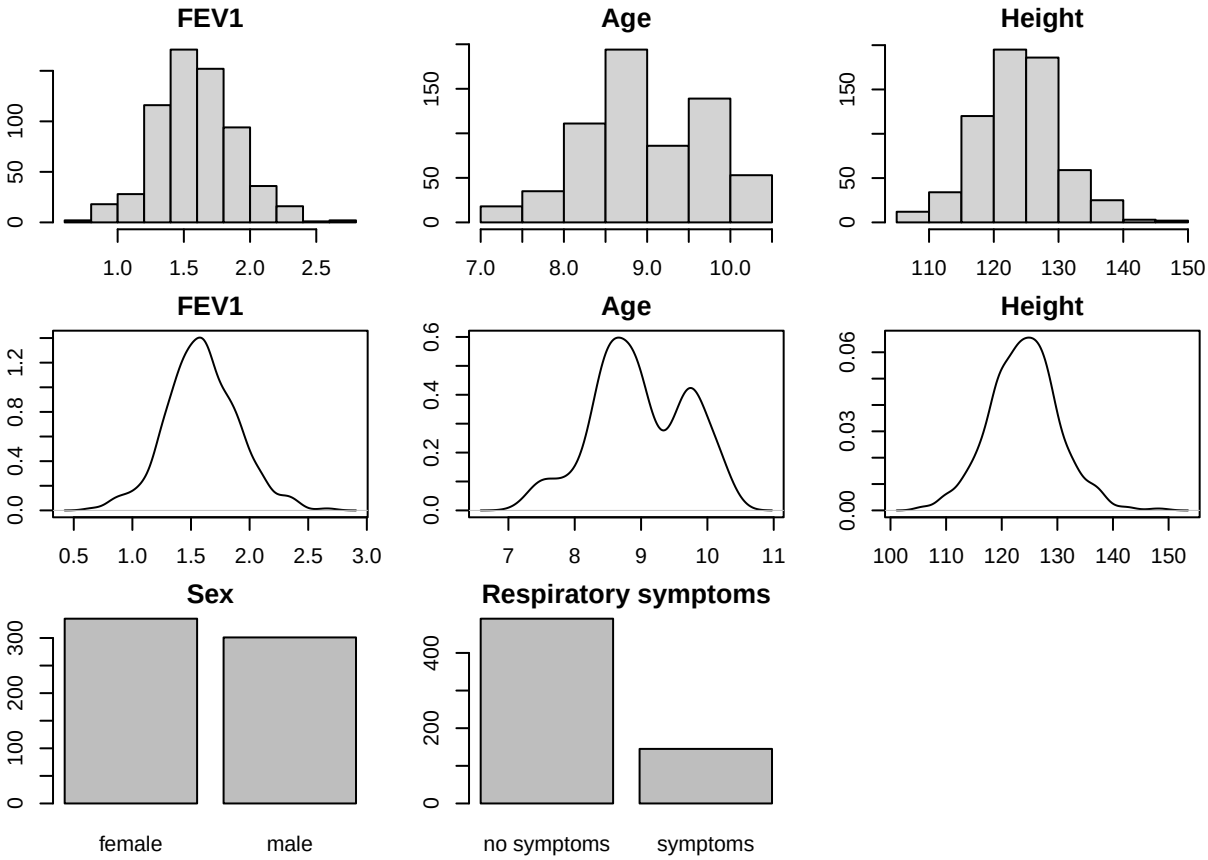
```
##
##    female      male
## 0.5267296 0.4732704
```

```
## Frequency table
table(perulung$respsymptoms)
```

```
##
## no symptoms    symptoms
##         491         145
```

```
## Frequency proportions
prop.table(table(perulung$respsymptoms))
```

```
##
## no symptoms    symptoms
##   0.7720126   0.2279874
```



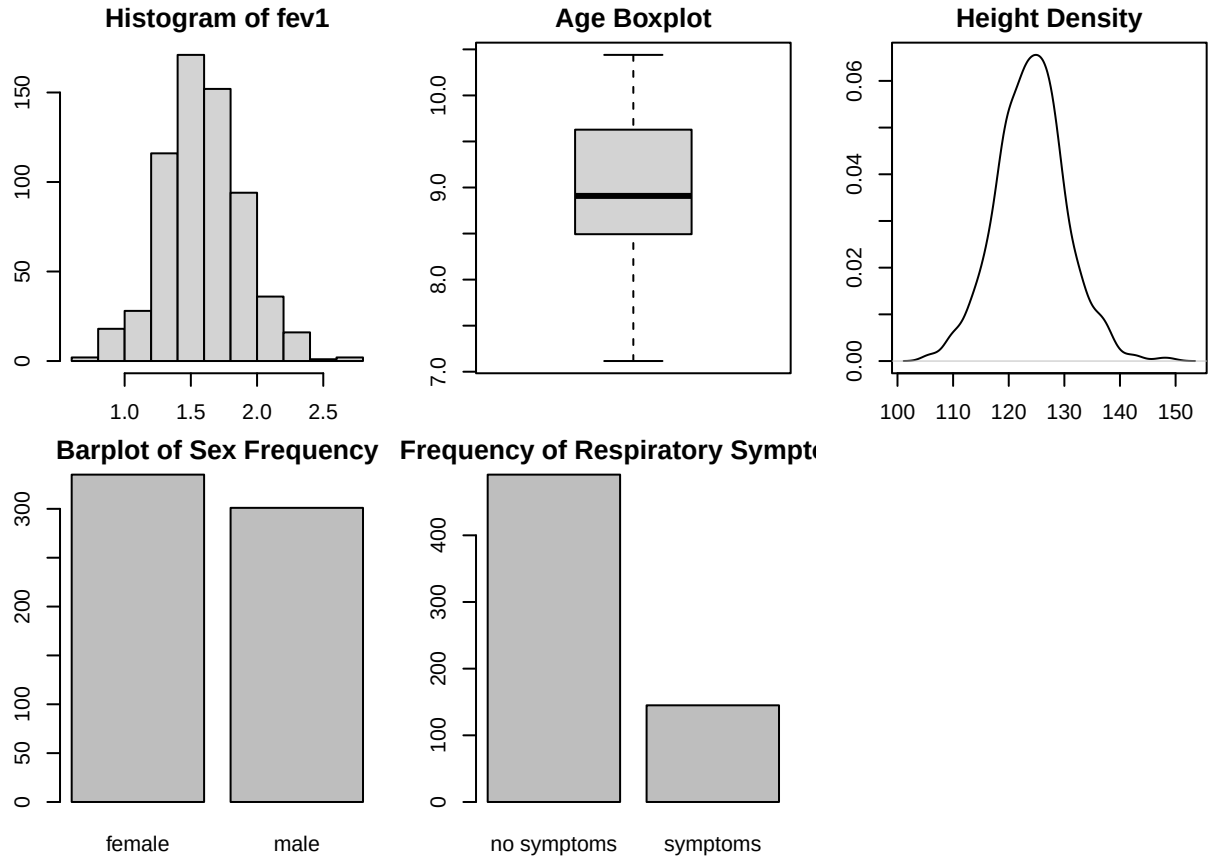
c. What are some research questions which these data could have been collected to address?

- What is the epidemiology of respiratory systems by age and gender?
- Do respiratory symptoms negatively affect pulmonary function measured by forced expiratory volume in 1 second (FEV1)?
- How does pulmonary function develop with age and height?
- What groups of children are most at risk for respiratory illness?

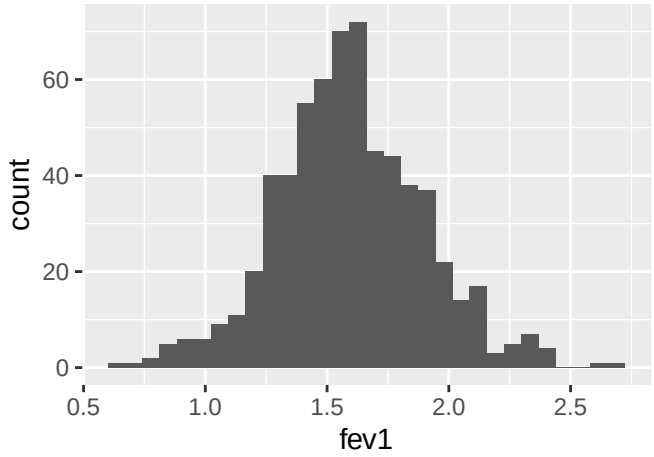
d. Use R to create appropriate univariate graphical summaries of each of the variables.

- fev1: histogram
- age: boxplot
- height: density plot

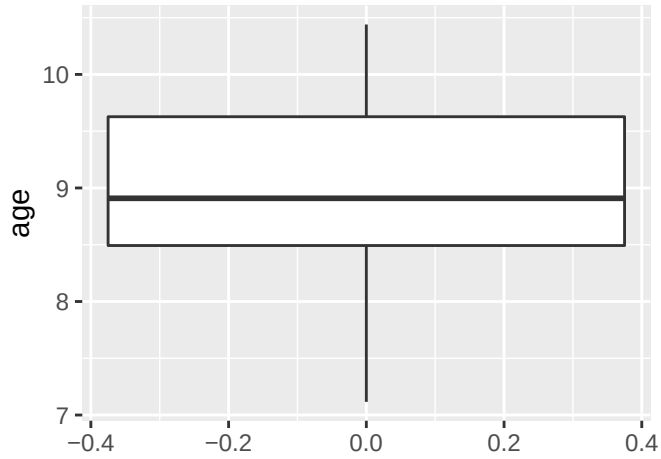
```
par(mfrow = c(2, 3),
    mar = c(2, 3, 2, 1))
hist(perulung$fev1, main="Histogram of fev1", xlab="")
boxplot(perulung$age, main="Age Boxplot")
plot(density(perulung$height), main="Height Density")
barplot(table(perulung$sex), main="Barplot of Sex Frequency")
barplot(table(perulung$respsymptoms), main = "Frequency of Respiratory Symptoms")
```



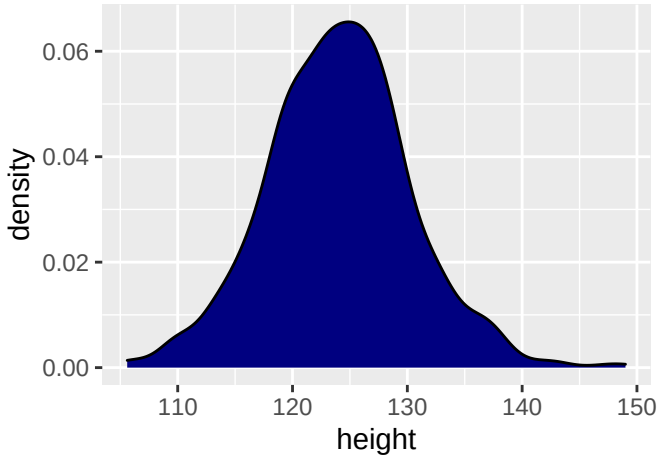

```
## Replicate same figures with ggplot
library(ggplot2)
ggplot(perulung, aes(fev1)) +
  geom_histogram()
```



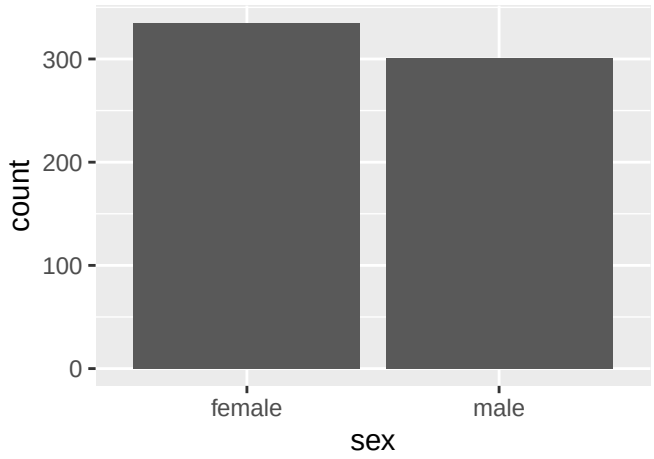

```
ggplot(perulung, aes(y = age)) +
  geom_boxplot()
```



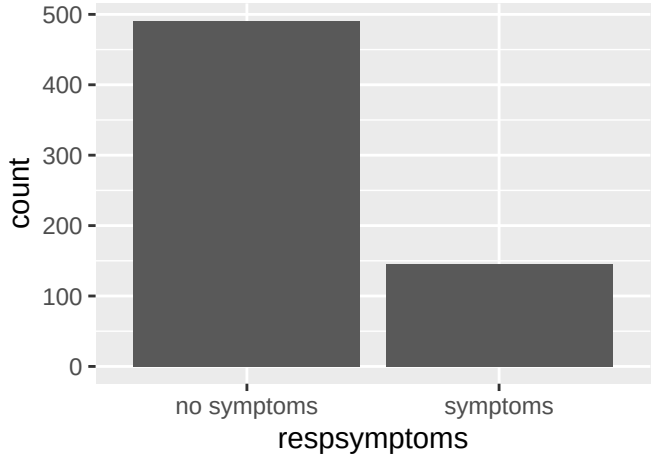

```
ggplot(perulung, aes(height)) +
  geom_density(fill = "navy")
```



```
ggplot(perulung, aes(sex)) +
  geom_bar()
```




```
ggplot(perulung, aes(respsymptoms)) +
  geom_bar()
```



used to describe sample to reader

no inferential statistics here, just descriptive

e. Create a single table summarizing key characteristics of the sample—an appropriate 'Table 1' for a medical or epidemiological paper. (It is probably possible to construct a full table with R commands, but you might find it easier to do calculations of summary statistics with R and copy the R output into a separate table in MS Word, Excel, or similar.)

columns: key exposure or disease

Group 1	No symptoms	Symptoms	Total
N (%)	491 (77%)	145 (23%)	636 (100%)
Sex: male (% of N)	237 (48%)	64 (44%)	301 (47%)
Median age (IQR)	9.0 (8.5–9.7)	8.7 (8.5–9.1)	8.9 (8.5–9.6)
Mean height cm (SD)	124 (6.2)	123 (6.4)	124 (6.2)
Mean FEV1 (SD)	1.63 (0.29)	1.48 (0.33)	1.59 (0.30)

sample info

```
summary(perulung$height)
summary(perulung$fev1)

aggregate(cbind(height, fev1) ~ respsymptoms, perulung, function(x) c(mean = mean(x), sd= sd(x)))
aggregate(fev1 ~ respsymptoms, perulung, mean)

## Calculate with dplyr
library(dplyr)

count(perulung, respsymptoms) %>%
  mutate(p = n/sum(n))
count(perulung)

 perulung %>%
 group_by(respsymptoms) %>%
 summarise(male_freq = sum(sex == "male"),
           male_prop = mean(sex == "male"),
           age_med = median(age),
           age_iqr1 = quantile(age, 0.25),
           age_iqr2 = quantile(age, 0.75),
           height_mean = mean(height),
           height_sd = sd(height),
           fev1_mean = mean(fev1),
           fev1_sd = sd(fev1))

 perulung %>%
 summarise(male_freq = sum(sex == "male"),
           male_prop = mean(sex == "male"),
           age_med = median(age),
           age_iqr1 = quantile(age, 0.25),
           age_iqr2 = quantile(age, 0.75),
           height_mean = mean(height),
           height_sd = sd(height),
           fev1_mean = mean(fev1),
           fev1_sd = sd(fev1))
```

f. In this sample of 636 children, does there appear to be an association between:

(i) sex and height,
(ii) age and height,
(iii) sex and lung function,
(iv) sex and presence of respiratory symptoms,
(v) respiratory symptoms and lung function.

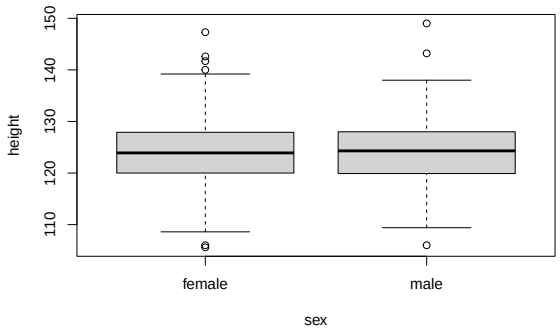
Support your answers with graphical or numerical evidence.

(i) There does not appear to be a strong relationship between sex and height in our dataset. The average height and quartiles are very similar for male and female children.

```
aggregate(height ~ sex, perulung, mean)
```

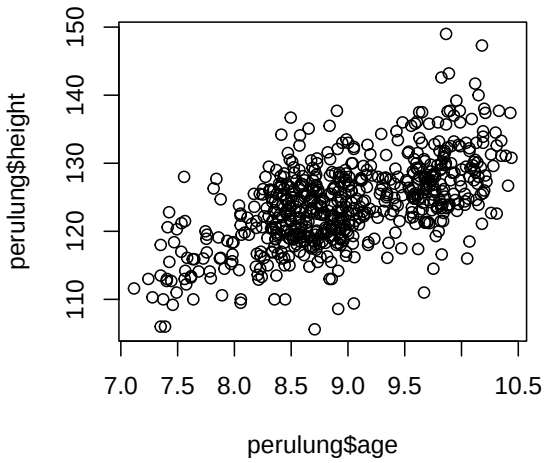
```
##      sex   height
## 1 female 124.0125
## 2   male 124.0980
```

```
boxplot(height ~ sex, perulung)
```



(ii) Height is positively correlated with age, based on a scatterplot of the two variables.

```
plot(perulung$age, perulung$height)
```

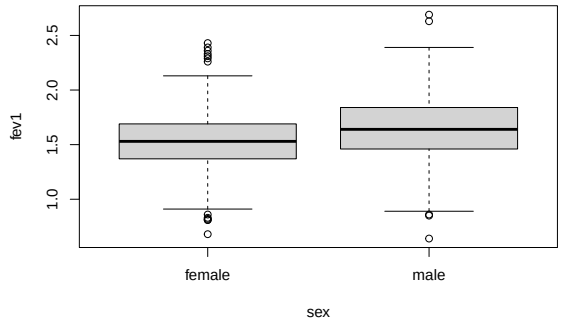


(iii) The mean FEV1 was slighlty higher for male children at 1.66 versus 1.54 for females.

```
aggregate(fev1 ~ sex, perulung, mean)
```

```
##      sex     fev1
## 1 female 1.538388
## 2   male 1.657276
```

```
boxplot(fev1 ~ sex, perulung)
```



(iv) There does not appear to be a strong relationship between sex and respiratory symptoms. 24% of female children had respiratory symptoms compared to 21% of male children.

```
prop.table(table(perulung$sex, perulung$respsymptoms), 1)
```

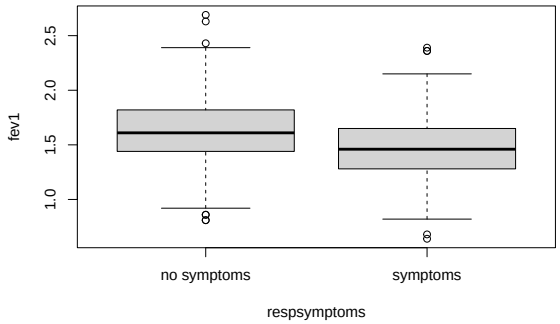
```
##
##          no symptoms  symptoms
##   female   0.7582090 0.2417910
##   male     0.7873754 0.2126246
```

(v) Children with respiratory symptoms had substantially lower FEV1 compared to children with no respiratory symptoms.

```
aggregate(fev1 ~ respsymptoms, perulung, mean)
```

```
##   respsymptoms     fev1
## 1  no symptoms 1.628758
## 2     symptoms 1.479172
```

```
boxplot(fev1 ~ respsymptoms, perulung)
```



g. What is the target population to which your conclusions about these questions might generalize?

- Children from deprived areas in urban Latin America.

h. Calculate an estimate and 95% confidence interval for each of the following quantities:

```
library(tidyverse)
perulung <- read.csv("perulung_ems.csv")
perulung <- perulung %>%
  mutate(sex = factor(sex, c(0, 1), c("female", "male")),
         respsymptoms = factor(respsymptoms, c(0, 1), c("no symptoms", "symptoms")))
```

i. Height in the whole population.
ii. FEV1 in the population.
iii. Height for male and females children separately.
iv. FEV1 for children with respiratory symptoms and those without respiratory symptoms.

For each of these outcomes calculate the large sample CI using the normal distribution and the small-sample CI using the t-distribution. Compare your calculations to the 95% CI estimate output by the `t.test()` function.

```r
x <- perulung$height

## A function to return requested outputs as a vector
calc_ex1a <- function(x, digits = 3) {
  mu <- mean(x)
  se <- sd(x) / sqrt(length(x))

  print_ci <- function(ci) paste0("(", round(ci[1], digits), ", ",
                                  round(ci[2], digits), ")")

  c(mean = round(mu, digits),
    ci_norm = print_ci(mu + c(-1, 1) * qnorm(0.975) * se),
    ci_tdist = print_ci(mu + c(-1, 1) * qt(0.975, length(x)-1) * se),
    ci_t.test = print_ci(t.test(x)$conf.int))
}

rbind(
  height = calc_ex1a(perulung$height),
  fev1 = calc_ex1a(perulung$fev1),
  height_female  = calc_ex1a(perulung$height[perulung$sex == "female"]),
  height_male  = calc_ex1a(perulung$height[perulung$sex == "male"]),
  fev1_nosymp  = calc_ex1a(perulung$fev1[perulung$respsymptoms == "no symptoms"]),
  fev1_symptom  = calc_ex1a(perulung$fev1[perulung$respsymptoms == "symptoms"])
) %>%
knitr::kable(align = "lcccc")
```

	mean	ci_norm	ci_tdist	ci_t.test
height	124.053	(123.568, 124.538)	(123.567, 124.539)	(123.567, 124.539)
fev1	1.595	(1.571, 1.618)	(1.571, 1.618)	(1.571, 1.618)
height_female	124.013	(123.321, 124.704)	(123.319, 124.706)	(123.319, 124.706)
height_male	124.098	(123.42, 124.776)	(123.418, 124.778)	(123.418, 124.778)
fev1_nosymp	1.629	(1.603, 1.654)	(1.603, 1.654)	(1.603, 1.654)
fev1_symptom	1.479	(1.425, 1.533)	(1.425, 1.534)	(1.425, 1.534)

The confidence intervals reported by `t.test()` are exactly the same as the small-sample 95% CI. They are very slightly wider than the large-sample CIs, but both are very close because the sample size is large.

i. For each of the following research questions:

- articulate an appropriate null and alternative hypothesis,
- calculate an appropriate estimate and uncertainty range,
- determine an appropriate statistical test for your hypothesis, and
- report the results of your hypothesis test to answer the question.

i. Is the average height of children aged 7 to 10 years in Lima greater than 124cm?

H0: The average height of children is $\leq$ 124cm.
H1: The average height of children is >124cm.

We used a one-sample t-test with a one-sided alternative hypothesis that the average height is >124cm. We will reject the null hypothesis that if the p-value is less than $\alpha = 0.05$.

The average height among children aged 7 to 10 was 124.05cm with 95% confidence interval (123.57, 124.54), representating a range of values within which the true mean height likely falls. The p-value was 0.415, indicating a reasonably high probability of observing a sample mean of 124.053 under the null hypthosis that height is $\leq$ 124cm. Thus we fail to reject the null hypothesis at the $\alpha = 0.05$ level.

```
        mean(perulung$height)
```

```
## [1] 124.053
```

```
        t.test(perulung$height)$conf.int
```

```
## [1] 123.5673 124.5387
## attr(,"conf.level")
## [1] 0.95
```

```
        t.test(perulung$height, mu = 124, alternative = "greater")
```

```
##
##  One Sample t-test
##
## data:  perulung$height
## t = 0.21422, df = 635, p-value = 0.4152
## alternative hypothesis: true mean is greater than 124
## 95 percent confidence interval:
##  123.6455      Inf
## sample estimates:
## mean of x
##   124.053
```

ii. Is the average height of girls in Lima equal to 123.5cm?

H0: Average height of girls in Lima is equal to 123.5cm.
H1: Average height of girls in Lima not equal to 123.5cm.

Test: One-sample t-test, two-sided alternative hypothesis.

The average height among our sample of 335 girls aged 7 to 10 in Lima is 124.01cm with 95% confidence interval from 123.3cm to 124.7cm. Compared to the reference value of 123.5cm, the t-statistic is 1.45 with 334 degrees of freedom, corresponding to a p-value of 0.147. Based on the 95% confidence range and/or the p-value, we conclude that our data are not inconsistent with the average height of girls aged 7 to 10 in Lima being 123.5cm

Note: we cannot conclude that the average height is 123.5cm (we can never prove the null hypothesis), only conclude that our data are not inconsistent with this value.

```
       x_female <- perulung$height[perulung$sex == "female"]
       length(x_female)
```

```
## [1] 335
```

```
       mean(x_female)
```

```
## [1] 124.0125
```

```
       t.test(x_female, mu = 123.5)
```

```
##
##  One Sample t-test
##
## data:  x_female
## t = 1.4535, df = 334, p-value = 0.147
## alternative hypothesis: true mean is not equal to 123.5
## 95 percent confidence interval:
##  123.3189 124.7062
## sample estimates:
## mean of x
##  124.0125
```

iii. Is there an association between sex and height amongst children in Lima?

H0: The average height of female children is equal to the average height of male children. H1: The average height of female children is not equal to the average height of male children.

Hypothesis test: two-sample t-test with equal variance (unequal variance t-test also acceptable); two-sided alternative hypothesis.

In our sample of 301 male children and 335 female children, the male children were 0.09cm taller than the female children (95% CI -0.89–1.06cm). The t-value was 0.172 on 634 degrees of freedom, corresponding to a two-sided p-value of 0.86. Is is very probable that we would have observed a difference this large or greater if there was no true height difference between male and female children. Thus we conclude that our data do not provide evidence to reject the null hypothesis that the average heights of male and female children are equal.

```
       table(perulung$sex)
```

```
##
## female   male
##    335    301
```

```
       x_female <- perulung$height[perulung$sex == "female"]
       x_male <- perulung$height[perulung$sex == "male"]

       mean(x_male) - mean(x_female)
```

```
## [1] 0.08546933
```

```
       t.test(x_male, x_female, var.equal = TRUE)
```

```
##
##  Two Sample t-test
##
## data:  x_male and x_female
## t = 0.17239, df = 634, p-value = 0.8632
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -0.8881117  1.0590504
## sample estimates:
## mean of x mean of y
##  124.0980  124.0125
```

iv. Do children with respiratory symptoms have reduced pulmonary function compared to children with no respiratory symptoms?

H0: FEV1 for children with respitory symptoms is greater than or equal to FEV1 for children with no symptoms.
H1: FEV1 for chidren with respiratory symptoms is less than FEV1 for children with no symptoms.

Hypothesis test: two-sample t-test, unequal variance (equal variance also justifiable), **one-sided** alternative hypothesis.

The FEV1 for children with with respiratory symptoms was 0.15 litres/second lower than children with no respiratory symptoms (95% CI 0.09–0.21 litres/second). The t-statistic was -4.90 on 211.5 degrees of freedom. The corresponding one-sided p-value was <0.001, indicating strong evidence to reject the null hypothesis of no difference and conclude that FEV1 is statistically significantly lower for children with respiratory symptoms compared to those with no symptoms.

```
fev1_symp <- perulung$fev1[perulung$respsymptoms == "symptoms"]
fev1_nosymp <- perulung$fev1[perulung$respsymptoms == "no symptoms"]
mean(fev1_symp) - mean(fev1_nosymp)
```

```
## [1] -0.1495852
```

```
t.test(fev1_symp, fev1_nosymp)
```

```
##
##  Welch Two Sample t-test
##
## data:  fev1_symp and fev1_nosymp
## t = -4.904, df = 211.47, p-value = 1.87e-06
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -0.2097132 -0.0894573
## sample estimates:
## mean of x mean of y
##  1.479172  1.628758
```

```
t.test(fev1_symp, fev1_nosymp, alternative = "less")
```

```
##
```

```
##  Welch Two Sample t-test
##
## data:  fev1_symp and fev1_nosymp
## t = -4.904, df = 211.47, p-value = 9.35e-07
## alternative hypothesis: true difference in means is less than 0
## 95 percent confidence interval:
##        -Inf -0.0991923
## sample estimates:
## mean of x mean of y
##  1.479172  1.628758
```

j. If the null hypothesis were true for all of the above four questions, what is the probability of erroneously rejecting at least one null hypothesis and incorrectly concluding an association exists?

Using a Type I error rate threshold of $\alpha = 0.05$, if the null hypothesis is true, there is a 95% probability of correctly failing to reject the null hypothesis and a 5% probability of erroneously rejecting the null hypothesis and incorrectly concluding that there is an association.

If all four null hypotheses are true (and all of the tests are independent), the probability of correctly failing to reject each of them is $0.95^4 = 0.815$. Thus the probability that we erroneously reject at least one null hypothesis is 1 - 0.815 = 18.5%.

B2. The *National Health and Nutrition Examination Survey* (NHANES) is a nationally representative survey to assess the health and nutrition of adults and children in the United States. The survey was first conducted in the 1960s and has been conducted continuously since 1999 with around 5000 respondents sampled and interviewed in their homes every year. The survey consists of a combination of questionnaire responses and physical and biomarker measurements. Further information about the survey and datasets can be found here: https://www.cdc.gov/nchs/nhanes/index.htm.

The R package `NHANES` contains an extract of 75 variables about 10,000 respondents to NHANES between 2009 and 2012 abstracted for educational purposes. The actual NHANES survey datasets include sampling weights to account for non-equal sampling probability of certain population groups to increase the statistical efficiency of the survey, which is not covered in this course. The dataset of 10,000 respondents in the `NHANES` R package has been constructed such that the dataset can analysed as if it were a simple random sample from the American population. See the package documentation for information and links about accessing and analysing the actual NHANES data for research purposes; there are other R packages available on CRAN to assist with accessing and processing the actual NHANES survey data.

The objective of this exercise is to practice loading large datasets into R, understanding the structure and variables in a dataset, and conducting exploratory analysis.

a. Open and explore the NHANES dataset in R through the following steps:

- Install the `NHANES` R package: `install.packages("NHANES")`
- Load the R package: `library(NHANES)`
- Load the NHANES dataset into your workspace: `data(NHANES)`
- Use the command `?` to access the help page for the dataset: `?NHANES`

  Review the dataset documentation, particularly paying attention to:

  (i) What was the purpose for collecting the data?
  (ii) When and how were the data in the dataset collected?
  (iii) What is the target population of the sample?
  (iv) What is the sample size? Who was eligible to be included in the dataset? Are there different eligibility or inclusion criteria for certain variables?
  (v) What are the areas of information available in the dataset?

```
## install.packages("NHANES")
library(NHANES)
data(NHANES)
?NHANES
```

(i) The data were collected to monitor the health and nutrition of children and adults in the United States.
(ii) The data were collected in two survey rounds between 2009-2012. Data were collected through interviews in the respondent's home and a health examination conducted in a mobile examination centre.
(iii) The target population was the non-institutionalised civilian resident population of the United States.
(iv) The sample size for the analytical dataset is 10,000 adults and children. All non-institutionalised civilian residents of the United States are eligible to be included in the sample. Several variables have different inclusion criteria. For example, educational level and marital status are recorded for participants aged 20 or over only; length only for children under 3; head circumference is measured only for children aged 0-6 months.
(v) Data are available about demographic characteristics, physical health measurements, health biomarkers and reported health state, and lifestyle variables.

In the documentation, note that several of the variables are only collected for respondents of a certain age range or in one of the survey rounds but not the other. This is important to take note because it may affect what questions can be addressed by the data, or result in errant conclusions and incorrect interpretation if eligibility and inclusion criteria are not appropriately considered during analysis. For the remainder of the tutorial, we will only consider the subset of the sample who are adults aged 20 years and older.

Construct this dataset with the R command: `nhanes20pl <- NHANES[NHANES$Age >= 20, ]`

```
nhanes20pl <- NHANES[NHANES$Age >= 20, ]
```

(vi) Confirm that your new dataset has 7235 respondents remaining. Use at least one different R command to acheive construction of the same subsetted dataset. Confirm that your alternate command has the same number of rows and columns.

```
nrow(nhanes20pl)
```

```
## [1] 7235
```

```
## Three ways construct the subsetted dataset
nhanes20pl_a <- subset(NHANES, Age >= 20)
nhanes20pl_b <- NHANES[which(NHANES$Age >= 20), ]
nhanes20pl_c <- filter(NHANES, Age >= 20)          # using dplyr

nrow(nhanes20pl_a)
```

```
## [1] 7235
```

```
nrow(nhanes20pl_b)
```

```
## [1] 7235
```

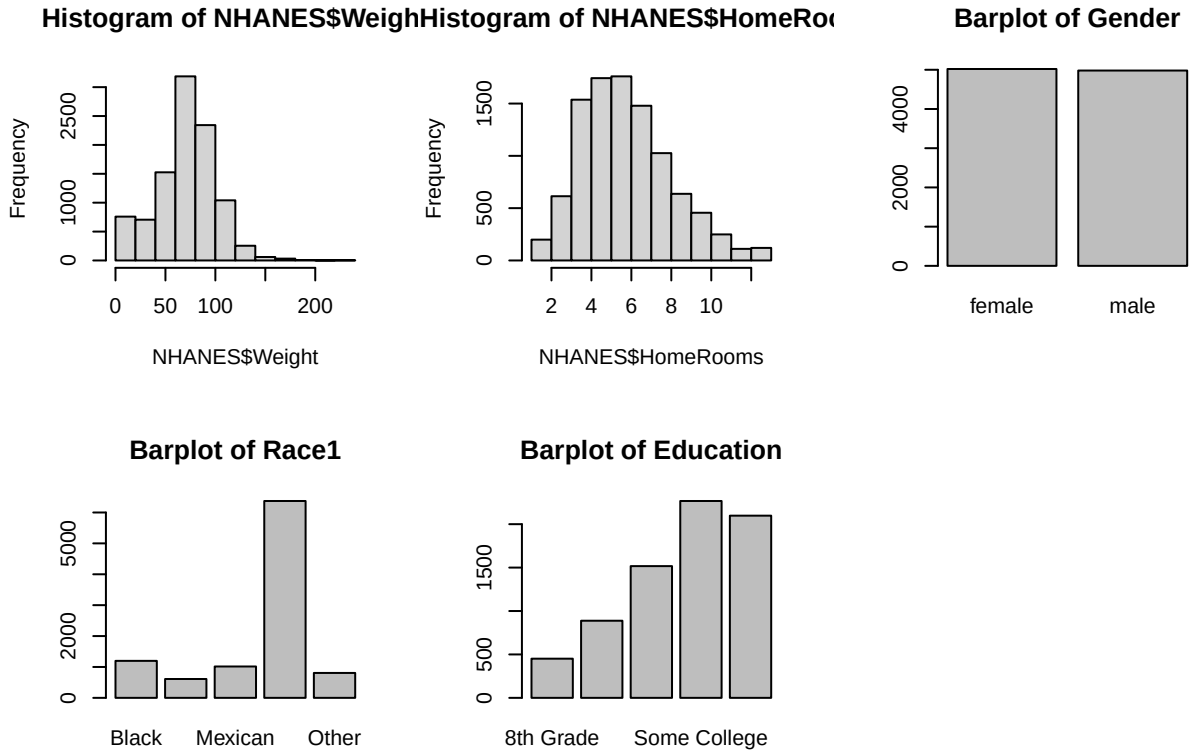
```
nrow(nhanes20pl_c)
```

```
## [1] 7235
```

b. *Types of variables.* Using the subsetted `nhanes20pl` dataset, identify at least one variable of each of the types of variables: continuous, discrete numeric, binary, categorical, and ordered categorical. For an identified variable of each type, create an appropriate summary of the frequency distribution and calculate an appropriate measure of central tendency and variation.

- Continous variable: Weight — mean and standard deviation
- Discrete numeric variable: HomeRooms — median and interquartile range due to positive skew (mean & SD probably also reasonble)
- Binary variable: Gender — frequency table, frequency proportions
- Categorical variable: Race1 — frequency table, frequency proportions
- Ordered categorical: Education — frequency table, frequency proportions

```
par(mfrow = c(2, 3))
hist(NHANES$Weight)
hist(NHANES$HomeRooms)
barplot(table(NHANES$Gender), main = "Barplot of Gender")
barplot(table(NHANES$Race1), main = "Barplot of Race1")
barplot(table(NHANES$Education), main = "Barplot of Education")
```




```
mean(NHANES$Weight, na.rm=TRUE)
```

```
## [1] 70.9818
```

```
sd(NHANES$Weight, na.rm=TRUE)
```

```
## [1] 29.12536
```

```
median(NHANES$HomeRooms, na.rm=TRUE)
```

```
## [1] 6
```

```
quantile(NHANES$HomeRooms, c(0.25, 0.75), na.rm=TRUE)
```

```
## 25% 75%
##   5   8
```

```
   table(NHANES$Gender)
```

```
##
## female   male
##   5020   4980
```

```
   prop.table(table(NHANES$Gender))
```

```
##
## female   male
##  0.502  0.498
```

```
   table(NHANES$Race1)
```

```
##
##    Black Hispanic  Mexican    White    Other
##     1197      610     1015     6372      806
```

```
   prop.table(table(NHANES$Race1))
```

```
##
##    Black Hispanic  Mexican    White    Other
##   0.1197   0.0610   0.1015   0.6372   0.0806
```

```
   table(NHANES$Education)
```

```
##
##      8th Grade 9 - 11th Grade    High School   Some College   College Grad
##            451            888           1517           2267           2098
```

```
   prop.table(table(NHANES$Education))
```

```
##
##      8th Grade 9 - 11th Grade    High School   Some College   College Grad
##     0.06245672     0.12297466     0.21008171     0.31394544     0.29054148
```

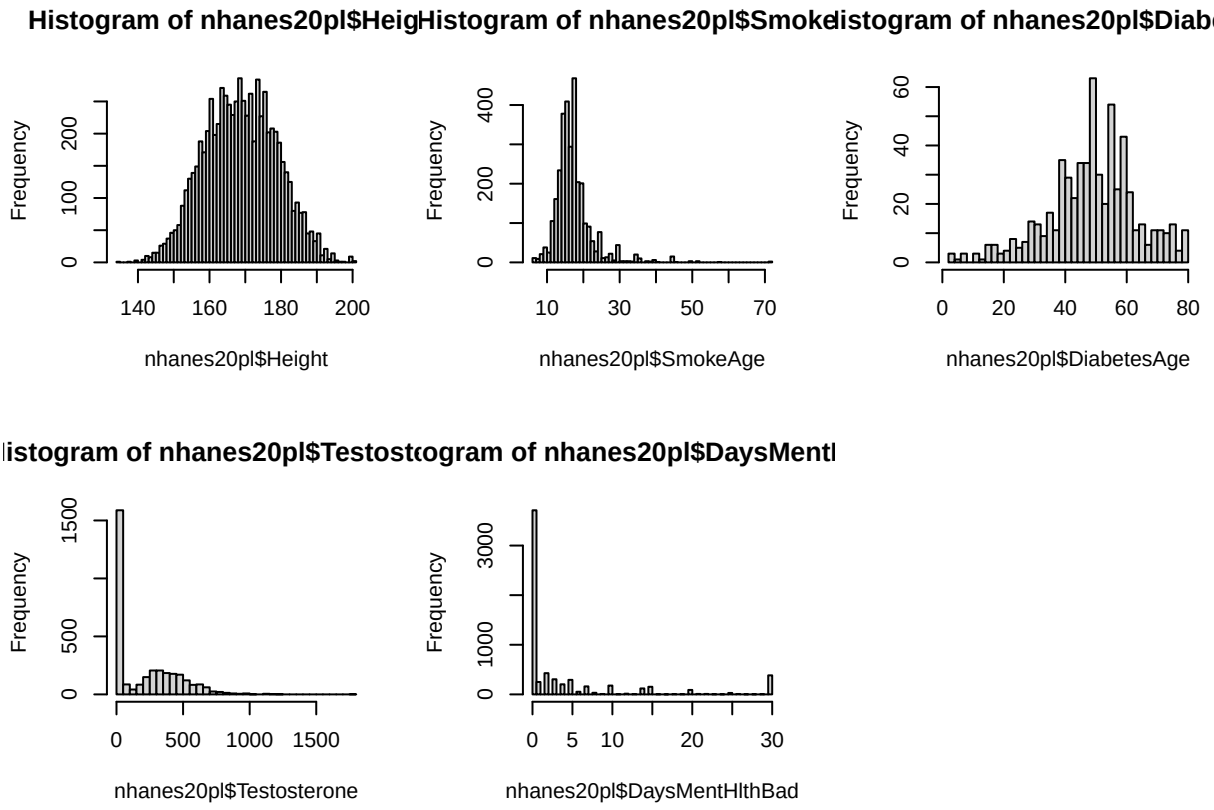
c. *Frequency distributions.* Identify at least one variable in the dataset that has a frequency distribution matching each of the shapes described in Kirkman and Sterne Figures 3.5 and 3.6 (pages 20-21). For each of the identified variables, calculated the mean, median, mode, variance, standard devation, range, and interquartile range.

- Symmetric: Heigh
- Positive skew: SmokeAge
- Negative skew: DiabetesAge
- Bimodal: Testosterone
- Reverse-J: DaysMentHlthBad (tenuous example)
- Uniform: *no good example*

```
    summary(nhanes20pl[c("Height", "SmokeAge", "DiabetesAge", "Testosterone", "DaysMentHlthBad")])
```

```
##      Height         SmokeAge      DiabetesAge     Testosterone
##  Min.   :134.5   Min.   : 6.00   Min.   : 3.00   Min.   :   0.25
##  1st Qu.:161.4   1st Qu.:15.00   1st Qu.:41.00   1st Qu.:  19.88
##  Median :168.7   Median :17.00   Median :50.00   Median :  91.38
##  Mean   :168.8   Mean   :17.83   Mean   :49.27   Mean   : 217.08
##  3rd Qu.:176.0   3rd Qu.:19.00   3rd Qu.:59.00   3rd Qu.: 385.00
##  Max.   :200.4   Max.   :72.00   Max.   :80.00   Max.   :1795.60
##  NA's   :53      NA's   :4155    NA's   :6618    NA's   :3903
##  DaysMentHlthBad
##  Min.   : 0.00
##  1st Qu.: 0.00
##  Median : 0.00
##  Mean   : 4.22
##  3rd Qu.: 4.00
##  Max.   :30.00
##  NA's   :762
```

```
    par(mfrow = c(2,3))
    hist(nhanes20pl$Height, breaks = 50)
    hist(nhanes20pl$SmokeAge, breaks = 50)
    hist(nhanes20pl$DiabetesAge, breaks = 50)
    hist(nhanes20pl$Testosterone, breaks = 50)
    hist(nhanes20pl$DaysMentHlthBad, breaks = 50)
```



Note: here's a bit of R code to select all of the numeric or integer variables and plot a histogram of each.

```r
df <- nhanes20pl[sapply(nhanes20pl, class) %in% c("integer", "numeric")]
df <- df[!apply(is.na(df), 2, all)]

par(mfrow = c(6, 8))
Map(hist, df, main = names(df), breaks = 100)
```

B3. *Statistical modelling: using the normal distribution to estimate population distributions.* The purpose of this exercise is to practice applying the normal distribution to estimate the distribution of an outcome in a population using data from a sample drawn from the population. The version of the NHANES dataset we are using is constructed to approximate a simple random sample from the American population. Thus we can use the observations in our sample to make inference about the American population.

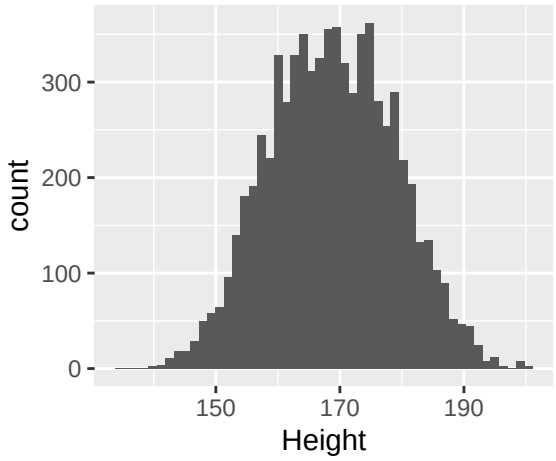
For each of the questions below, consider three continuous variables:

- `Height`: Standing height in centimeters.
- `BMI`: Body Mass Index
- `AlcoholYear`: Number of days over the past year that participant drank alcoholic beverages.

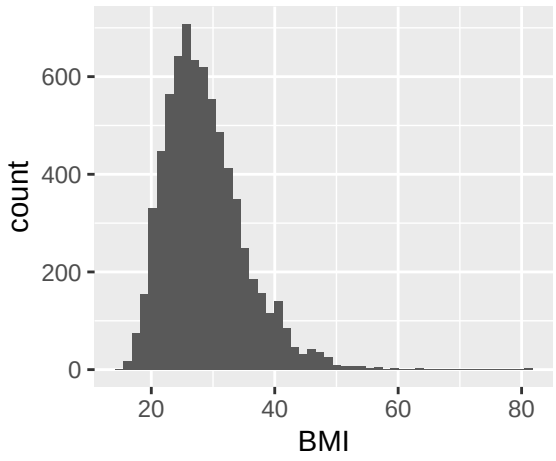
a. Create a graph of the frequency distribution for each of the variables, either using histograms or density plots. Does each outcome appear consistent with arising from a normal distribution?

Height appears normally distributed. BMI appears normally distributed perhaps with a slight positive skew. Alcohol days per year is severely positively skewed.

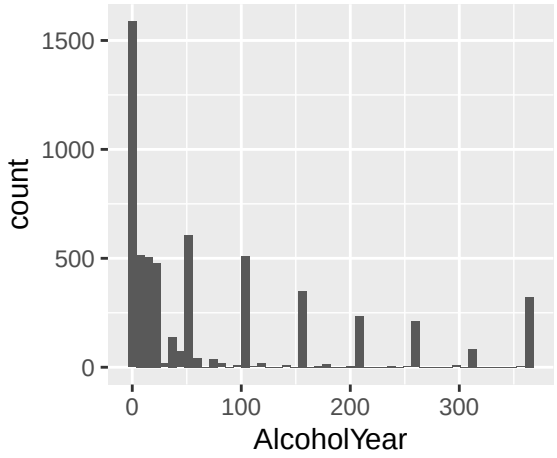
```
ggplot(nhanes20pl, aes(Height)) +
geom_histogram(bins = 50)
```




```
ggplot(nhanes20pl, aes(BMI)) +
geom_histogram(bins = 50)
```



```
ggplot(nhanes20pl, aes(AlcoholYear)) +
geom_histogram(bins = 50)
```



b. Calculate and report the sample mean and standard deviation for each of the outcomes.

```
height_mean <- mean(nhanes20pl$Height, na.rm = TRUE)
height_sd <- sd(nhanes20pl$Height, na.rm = TRUE)
bmi_mean <- mean(nhanes20pl$BMI, na.rm = TRUE)
bmi_sd <- sd(nhanes20pl$BMI, na.rm = TRUE)
alc_mean <- mean(nhanes20pl$AlcoholYear, na.rm = TRUE)
alc_sd <- sd(nhanes20pl$AlcoholYear, na.rm = TRUE)

data.frame(mean = c(height_mean, bmi_mean, alc_mean),
           sd = c(height_sd, bmi_sd, alc_sd),
           row.names = c("height", "bmi", "alcohol_year")) %>%
knitr::kable(digits = 1, align = "cc")
```

	mean	sd
height	168.8	10.1
bmi	28.8	6.7
alcohol_year	75.7	103.6

c. Use the normal distribution as a model to estimate the following:

(i) The proportion of adults who are above 165cm tall.
(ii) The proportion of adults between 153cm and 160cm tall.
(iii) The height of a door in order that 90% of adults can walk under without ducking.
(iv) The proportion of adults who are obese, defined as BMI above 30.
(v) The proportion of adults who are overweight, defined as BMI between 25 and 30.
(vi) The BMI threshold at which 25% of adults are below.
(vii) The proportion adults who drink alcohol on more than 100 days per year.
(viii) The proportion of adults who drink alcohol on fewer than 10 days per year.
(ix) The interquartile range for the number of days per year that American adults drink alcohol.

Hint: use the R functions ‘pnorm(...)‘ and ‘qnorm(...)‘.

```
ans2c <- c(1-pnorm(165, height_mean, height_sd),
           pnorm(160, height_mean, height_sd) - pnorm(153, height_mean, height_sd),
           qnorm(0.9, height_mean, height_sd),
           1-pnorm(30, bmi_mean, bmi_sd),
           pnorm(30, bmi_mean, bmi_sd) - pnorm(25, bmi_mean, bmi_sd),
           qnorm(0.25, bmi_mean, bmi_sd),
           1-pnorm(100, alc_mean, alc_sd),
           pnorm(10, alc_mean, alc_sd),
           qnorm(0.75, alc_mean, alc_sd) - qnorm(0.25, alc_mean, alc_sd))
```

d. Check your estimates based on the normal distribution by directly calculating each of the above proportions amongst the observed sample. Based on this comparison, do you think that the normal distribution is a good model for the population distribution of each outcome in the population?

```
ans2d <- c(mean(nhanes20pl$Height > 165, na.rm=TRUE),
           mean(nhanes20pl$Height > 153 & nhanes20pl$Height < 160, na.rm=TRUE),
           quantile(nhanes20pl$Height, 0.9, na.rm=TRUE),
           mean(nhanes20pl$BMI > 30, na.rm=TRUE),
           mean(nhanes20pl$BMI > 25 & nhanes20pl$BMI <= 30, na.rm=TRUE),
           quantile(nhanes20pl$BMI, 0.25, na.rm=TRUE),
           mean(nhanes20pl$AlcoholYear > 100, na.rm=TRUE),
           mean(nhanes20pl$AlcoholYear <= 10, na.rm=TRUE),
           diff(quantile(nhanes20pl$AlcoholYear, c(0.25, 0.75), na.rm=TRUE)))

labels <- c("Height above 165cm", "Height between 153-160cm", "Height 90th percentile",
            "BMI above 30", "BMI between 25-30", "BMI 25th percentile",
            "Alcohol days above 100", "Alcohol days below 10", "Alcohol days IQR")
data.frame(normal_approx = ans2c,
           empirical_dist = ans2d,
           row.names = labels) %>%
           knitr::kable(digits = 2)
```

	normal_approx	empirical_dist
Height above 165cm	0.65	0.63
Height between 153-160cm	0.13	0.15
Height 90th percentile	181.72	181.80
BMI above 30	0.43	0.36
BMI between 25-30	0.29	0.33
BMI 25th percentile	24.28	24.10
Alcohol days above 100	0.41	0.31
Alcohol days below 10	0.26	0.36
Alcohol days IQR	139.70	101.00

C) Advanced learning

C1. *Missing data.* Using the `nhanes20pl` dataset, review the amount of missing data for each of the variables pertaining to *Demographic Variables* and *Physical Measurements (Hint: the R function 'is.na' may be useful).* (i) Among each grouping (*Demographic Variables* and *Physical Measurements*) identify the one variable with the highest proportion of missing cases. (Do not consider variables that were available for only one of the survey rounds or not recorded for this age range when making your assessment.)

- Demographic Variables go from Gender to HomeOwn. Race3 has 3648 missing values because it was only asked for the second survey round 2011-2012. HHIncome and HHIncomeMid have the greatest number of NAs (missing for 603 observations)
- Physical measurements go from Weight to Testosterone. Excluding Testosterone which was only measured in 2011-2012 and variables measured only for children, BPSys1 and BPDia1 have the largest number of missing observations (missing for 519 cases).

```
## Demographic variables are Gender to HomeOwn: columns 3:15
names(nhanes20pl[3:15])
```

```
##  [1] "Gender"        "Age"           "AgeDecade"     "AgeMonths"
##  [5] "Race1"         "Race3"         "Education"     "MaritalStatus"
##  [9] "HHIncome"      "HHIncomeMid"   "Poverty"       "HomeRooms"
## [13] "HomeOwn"
```

```
colSums(is.na(nhanes20pl[3:15]))  # number missing values for each column
```

```
##        Gender           Age     AgeDecade     AgeMonths         Race1
##             0             0           333          3743             0
##         Race3     Education MaritalStatus      HHIncome   HHIncomeMid
##          3648            14             4           603           603
##       Poverty     HomeRooms       HomeOwn
##           537            55            49
```

```
## Physical Measurements go from Weight to Testosterone (columns 17:33) columns 3:15
names(nhanes20pl[17:33])
```

```
##  [1] "Weight"          "Length"          "HeadCirc"        "Height"
##  [5] "BMI"             "BMICatUnder20yrs" "BMI_WHO"        "Pulse"
##  [9] "BPSysAve"        "BPDiaAve"        "BPSys1"          "BPDia1"
## [13] "BPSys2"          "BPDia2"          "BPSys3"          "BPDia3"
## [17] "Testosterone"
```

```
colSums(is.na(nhanes20pl[17:33]))
```

```
##          Weight          Length        HeadCirc          Height
##              57            7235            7235              53
##             BMI BMICatUnder20yrs         BMI_WHO           Pulse
##              63            7235              93             254
##         BPSysAve         BPDiaAve          BPSys1          BPDia1
##             264             264             519             519
##          BPSys2          BPDia2          BPSys3          BPDia3
##             414             414             394             394
##     Testosterone
##             3903
```

(ii) For each of these two variables, in this sample, is there any relationship between `Gender` or `Race1` and the probability that data on the outcome is missing?

- HHIncome is slightly more likely to be missing for female respondents (9.0%) than male respondents (7.6%). HHIncom is much less likely to be missing for white respondents (5.9%) compared to other groups for whom it is missing between 12.2% and 14.8%.
- BPSys1 is more likely to be missing for females (8.6%) than for males (5.7%). BPSys1 is more likely to be missing for Black and Other race groups.

```
prop.table(table(nhanes20pl$Race1, is.na(nhanes20pl$HHIncome)), 1)
```

```
##
##                 FALSE       TRUE
##   Black    0.87454324 0.12545676
##   Hispanic 0.86538462 0.13461538
##   Mexican  0.85191348 0.14808652
##   White    0.94060631 0.05939369
##   Other    0.87773723 0.12226277
```

```
prop.table(table(nhanes20pl$Gender, is.na(nhanes20pl$HHIncome)), 1)
```

```
##
##               FALSE       TRUE
##   female 0.90958458 0.09041542
##   male   0.92398649 0.07601351
```

```
prop.table(table(nhanes20pl$Gender, is.na(nhanes20pl$BPSys1)), 1)
```

```
##
##               FALSE       TRUE
##   female 0.91420038 0.08579962
##   male   0.94284910 0.05715090
```

```
prop.table(table(nhanes20pl$Race1, is.na(nhanes20pl$BPSys1)), 1)
```

```
##
##                 FALSE       TRUE
##   Black    0.91108404 0.08891596
##   Hispanic 0.92788462 0.07211538
##   Mexican  0.94009983 0.05990017
##   White    0.93235719 0.06764281
##   Other    0.90510949 0.09489051
```