Imperial College London

Week 2 — Problem Set 1 Review; Sampling from a population, summarising data, and the normal distribution

Introduction to Statistical Thinking and Data Analysis
MSc in Epidemiology / Health Data Analytics
Autumn 2022

17 October 2022

Imperial College London

Outline

1. Review problem set 1
   A) Consolidating concepts
   B) Practicing skills
   C) Advanced learning

Imperial College London

Consolidating concepts

A1. What does the standard deviation measure?

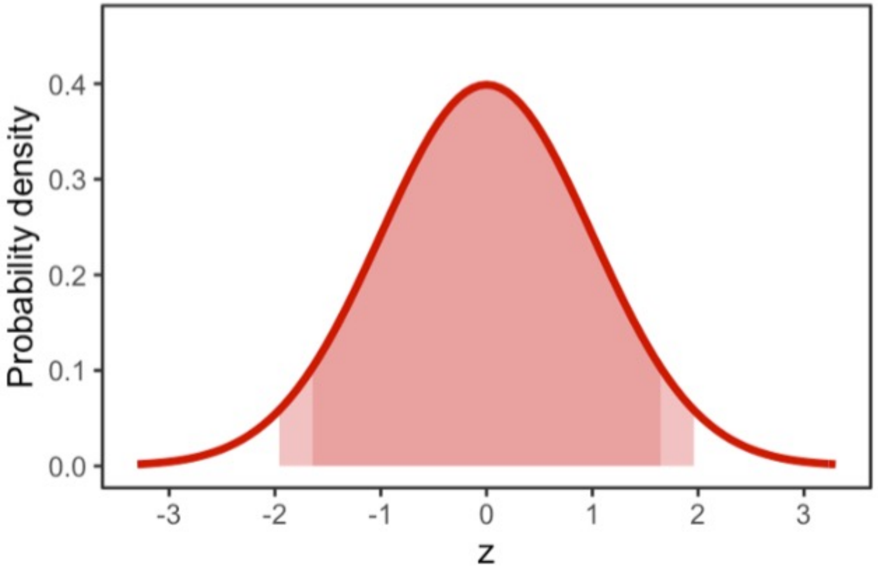
a) The spread of the middle 50% of the distribution.
b) The amount of variability in the sample mean.
c) The amount of variability in the population.
d) None of the above.

Imperial College London

Consolidating concepts

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.

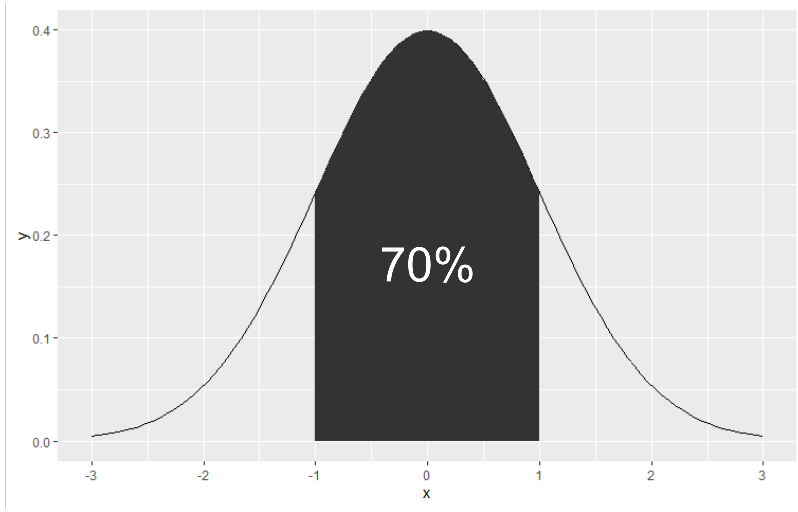


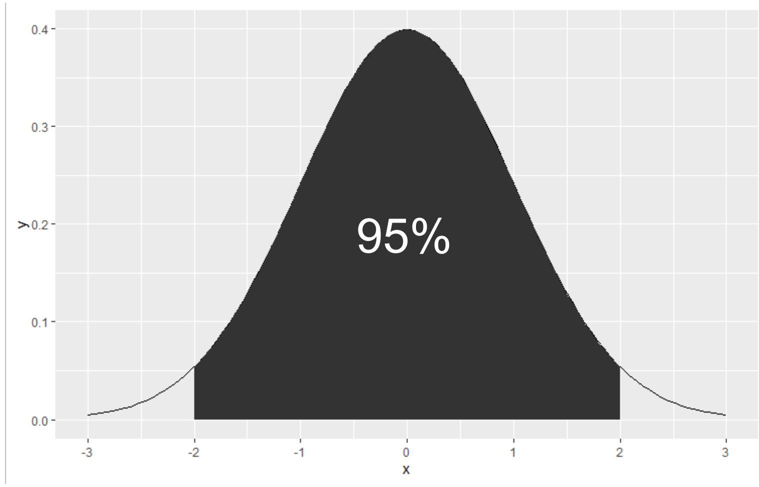
Imperial College London

Consolidating concepts

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.





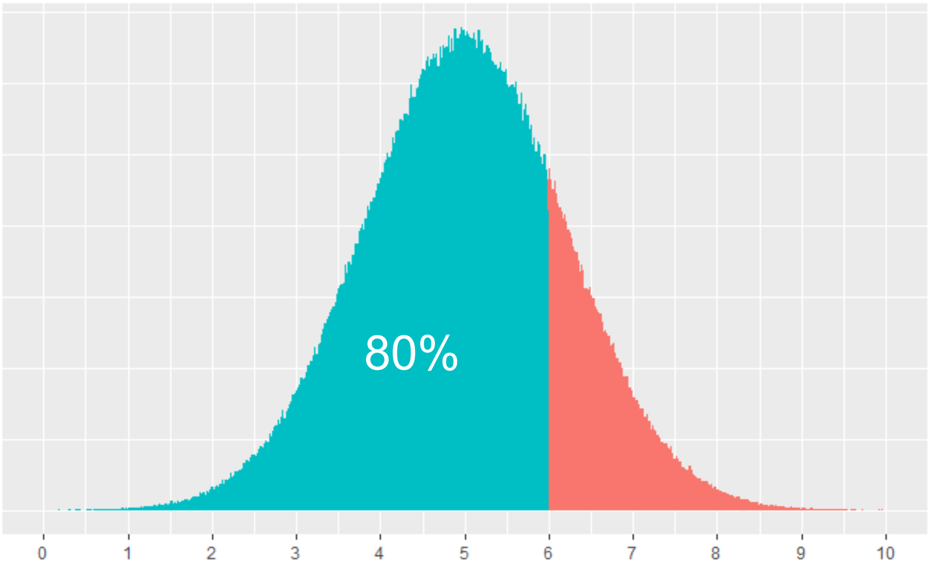
Consolidating concepts

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

```
pnorm(6,5,1.2)
```

```
## [1] 0.7976716
```



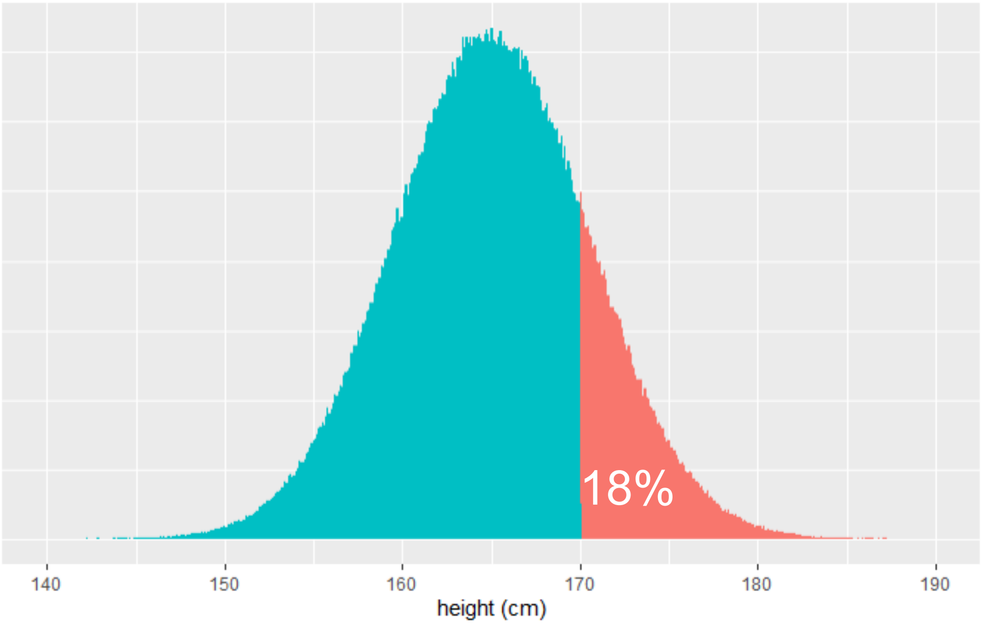
Consolidating concepts

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.

```r
pnorm(170,165,5.5, lower.tail = F)
```

```
## [1] 0.1816511
```



Imperial College London

Consolidating concepts

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

```
qnorm(0.95)
```

```
## [1] 1.644854
```

Consolidating concepts

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 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.

Imperial College London

Consolidating concepts

A8. The normal distribution is also called:

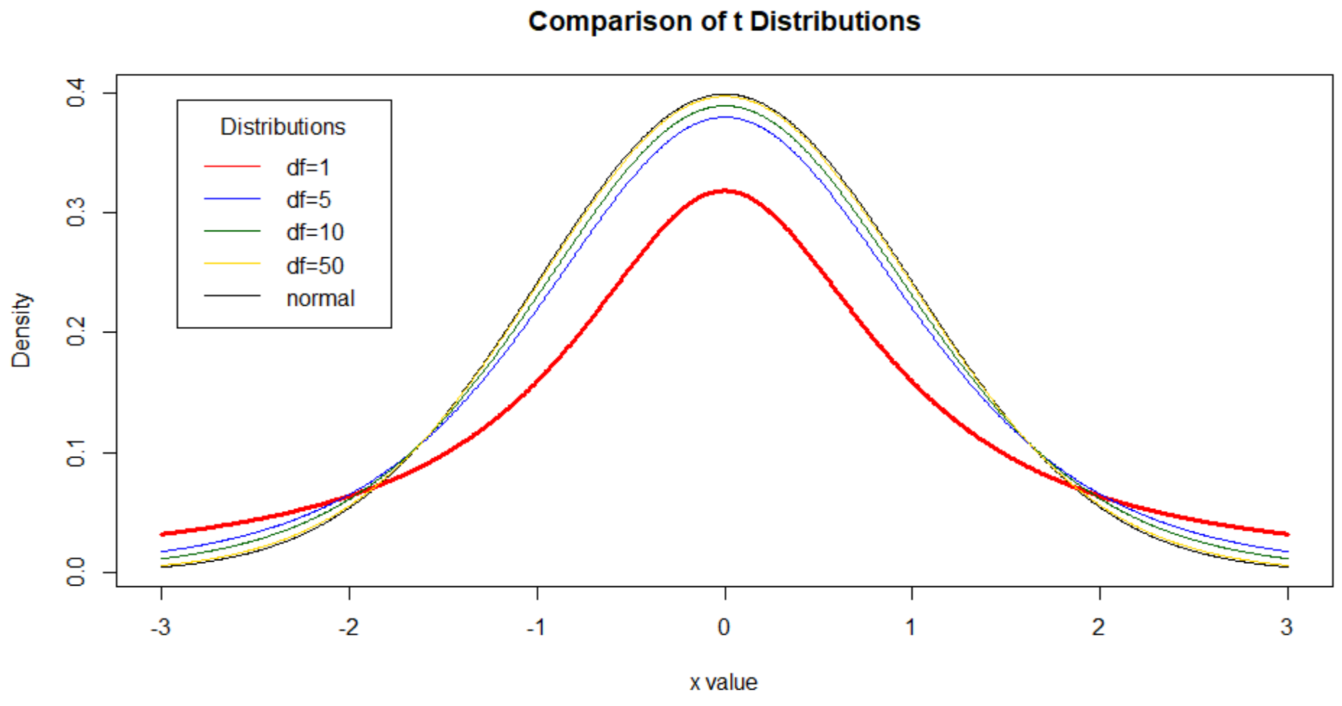
a) Poisson distribution
b) Bernoulli's distribution
c) Gaussian distribution
d) Student's t

Imperial College London

Consolidating concepts

A9. The shape of the t distribution:

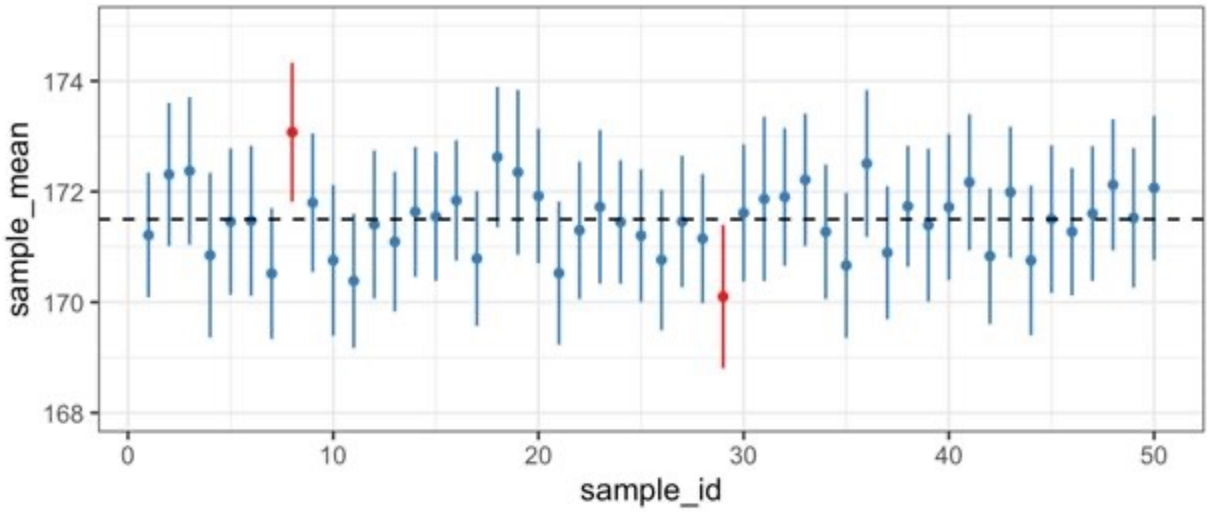
a) Is the same as the normal distribution.
b) Is skewed.
c) Depends on the number of degrees of freedom.
d) None of the above.



Consolidating concepts

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.



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)

Imperial College London

Question B1a,b,d,e

Variable	Type	Distribution shape	Summary statistic	Summary plot
id				
fev1	Continuous numeric	Symmetrical	Mean (SD)	Histogram
age	Continuous numeric	Bimodal (surprising)	Median and IQR	Boxplot
height	Continuous numeric	Symmetrical	Mean (SD)	Density plot
sex	Binary	Binary (well balanced)	Number and percent male	Frequency plot
respsymptoms	Binary	Binary (Imbalanced)	Number and % w/ symptoms	Frequency plot

Imperial College London

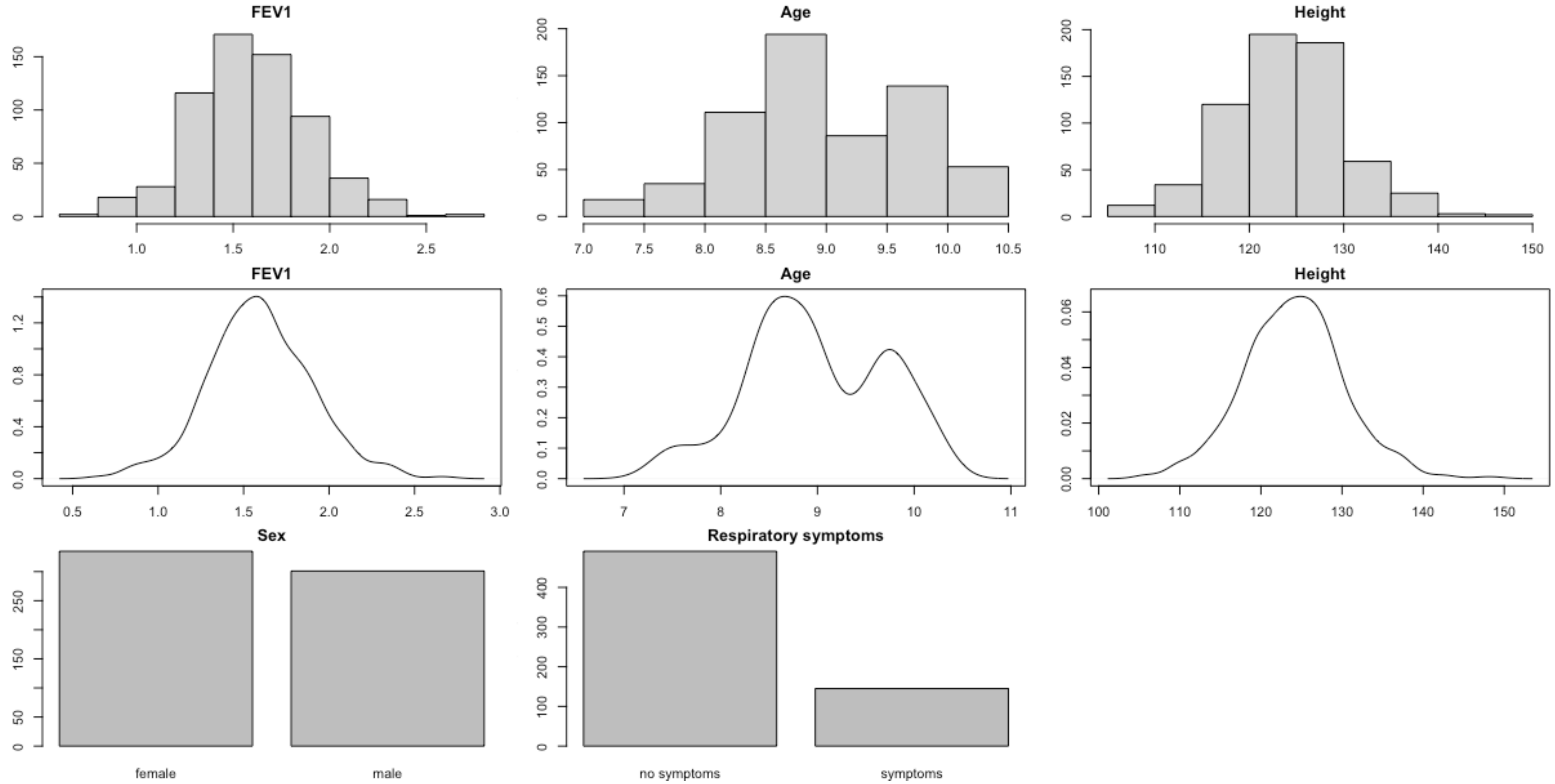
Question B1 c

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

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

Imperial College London

Question B1 d



Imperial College London

Question B1e

Create a single table summarizing key characteristics of the sample ('Table 1').

- Table 1 usually describes the study sample.
- Columns should stratify by the key exposure variable (e.g. RCT) or disease (e.g. case control study).
- Include rows for all variables in the final model.
- No inferential statistics.

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)

Imperial College London

Question B1f

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

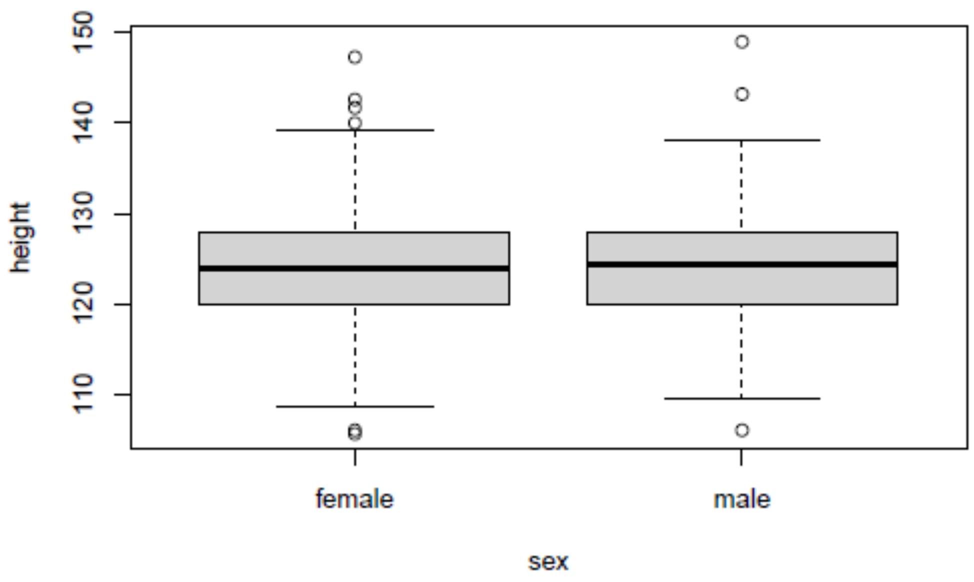
- (i) sex and height
- (ii) age and height
- (iii) sex and lung function
  (iv) sex and presence of respiratory symptoms
- (v) respiratory symptoms and lung function

Imperial College London

Question B1f

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

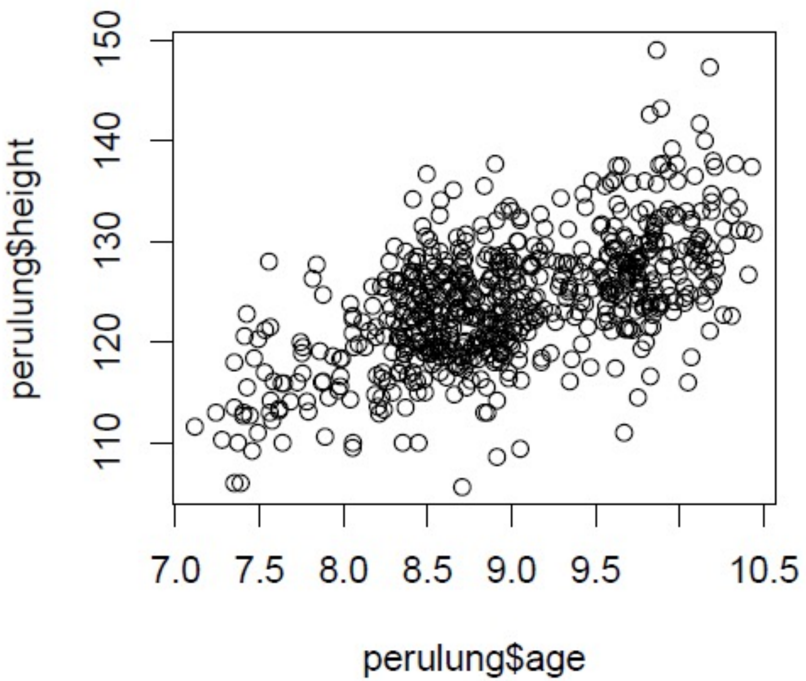
(i) sex and height – **No**



Question B1f

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

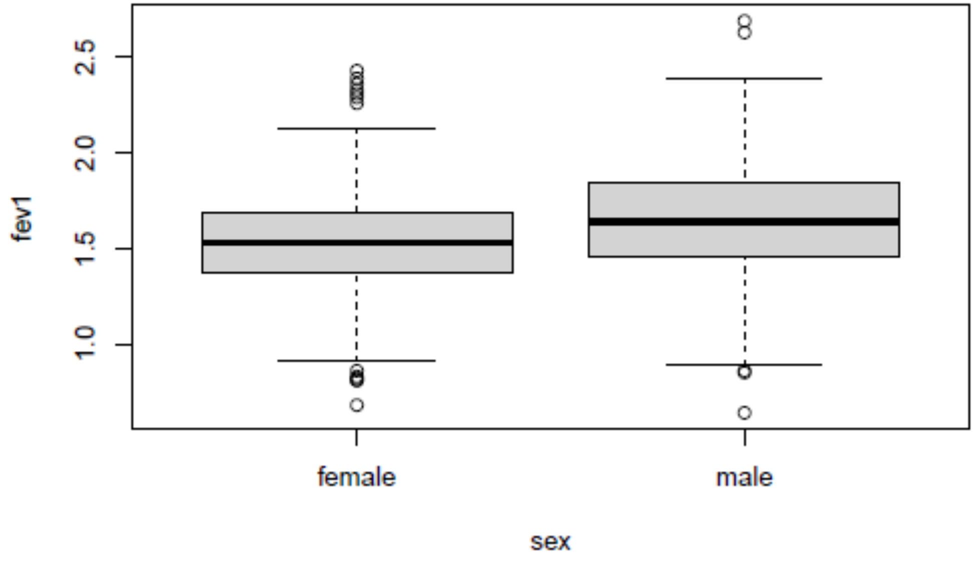
(ii) age and height – **Yes: height positively correlated with age**



Question B1f

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

(iii) sex and lung function – **Mean FEV1 slightly higher for males than females**



Question B1f

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

(iv) sex and presence of respiratory symptoms – **No (24% among female children, 21% among male children)**

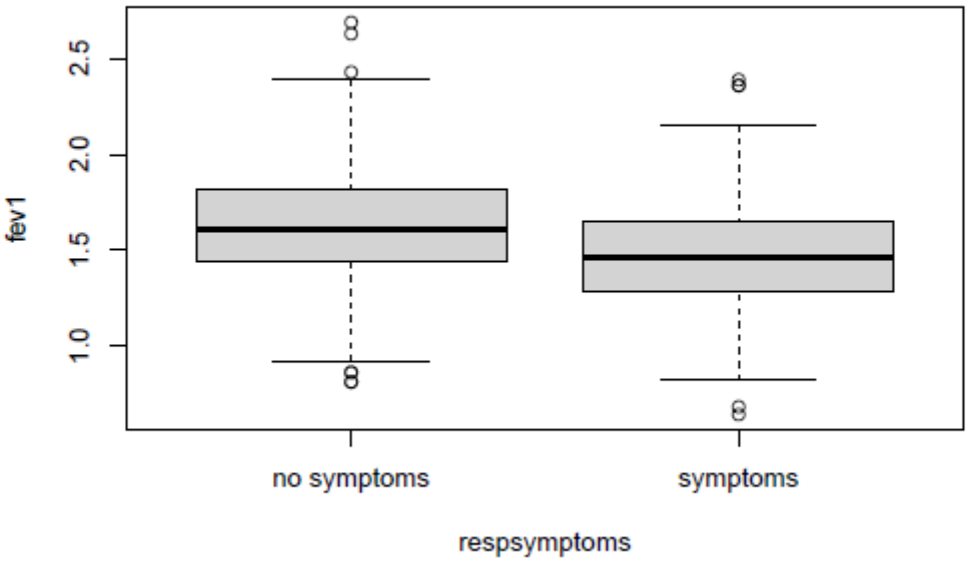
```
##
##             no symptoms  symptoms
##   female      0.7582090 0.2417910
##   male        0.7873754 0.2126246
```

Imperial College London

Question B1f

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

(v) respiratory symptoms and lung function – **Yes (mean FEV1 = 1.48 among with symptoms, mean FEV1 = 1.63 among no symptoms)**



Imperial College London

Question B1g

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

- Children from deprived areas in urban Latin America.

Imperial College London

Question B1h

Question 1a: Calculate an estimate and 95% confidence intervals.

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

- ‘Large-sample CI’ (normal distribution)
- ‘Small-sample CI’ (t-distribution)
- Using R function `t.test(…)`

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)

Imperial College London

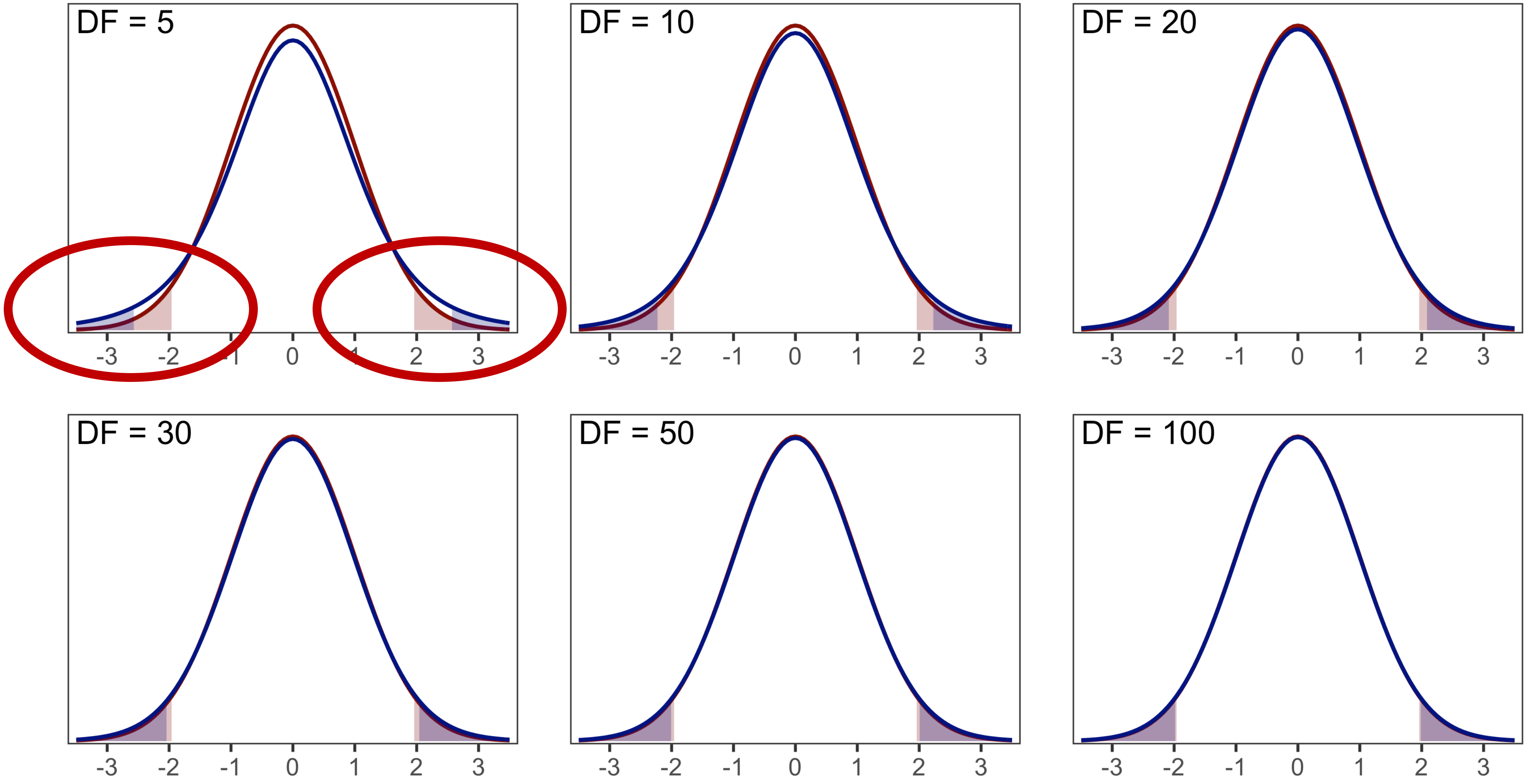
Question B1h

	Mean	SE	DF	95% CI: normal dist.	95% CI: t-distribution	95% CI: `t.test()`
height	124.053	0.247	635	(123.568, 124.538)	(123.567, 124.539)	(123.567, 124.539)
fev1	1.595	0.012	635	(1.571, 1.618)	(1.571, 1.618)	(1.571, 1.618)
height_female	124.013	0.353	334	(123.321, 124.704)	(123.319, 124.706)	(123.319, 124.706)
height_male	124.098	0.346	300	(123.42, 124.776)	(123.418, 124.778)	(123.418, 124.778)
fev1_nosymp	1.629	0.013	490	(1.603, 1.654)	(1.603, 1.654)	(1.603, 1.654)
fev1_symptom	1.479	0.028	144	(1.425, 1.533)	(1.425, 1.534)	(1.425, 1.534)

- DF = Number of observations - 1

- 95% CI with t-distribution (‘small sample CI’) is *slightly* wider than normal distribution (‘large sample CI’).
- 95% CI calculated with t-distribution matches R function `t.test(…)` exactly.

Normal distribution vs. t-distribution



Tails showing <2.5% and >97.5%

Imperial College London

Calculating 95% CIs in R

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

x <- perulung$height
n <- length(x)
## 636

sample_mean <- mean(x)
## 124.053
sample_mean_se <- sd(x) / sqrt(n)
## 0.2473493

df <- n - 1
## 635

crit_val_norm <- qnorm(0.975)
## 1.959964
crit_val_t <- qt(0.975, df = df)
## 1.963707

ci_norm <- sample_mean + c(-1, 1) *
           crit_val_norm * sample_mean_se
## 123.5682 124.5378

ci_tdist <- sample_mean + c(-1, 1) *
            crit_val_t * sample_mean_se
## 123.5673 124.5387

t.test(x)
## t = 501.53, df = 635, p-value < 2.2e-16
## alternative hypothesis: true mean != 0
## 95 percent confidence interval:
##  123.5673 124.5387
## sample estimates:
## mean of x
##   124.053
```

Question B1i

- 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.

Imperial College London

Question B1i(i)

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

- H0: The average height of children is ≤ 124cm.
- H1: The average height of children is >124cm.

- The average height among children aged 7 to 10 was 124.05cm with 95% confidence interval (123.57, 124.54).

- **One-sample t-test; one-sided alternative**
  - `t.test(perulung$height, mu = 124, alternative = "greater")`
  - p-value: $p = 0.415$

- 42% probability of observing sample mean of 124.05cm if true average height ≤124cm.
  - **Fail to reject** the null hypothesis that average height of children is ≤ 124cm.

Imperial College London

Question B1i(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.

- The average height among girls aged 7 to 10 was 124.01cm with 95% confidence interval from 123.3cm to 124.7cm.

- **One-sample t-test, two-sided alternative**
  - `t.test(perulung$height[perulung$sex == "female"], mu = 123.5)`
  - t-statistic = 1.45 with 334 degrees of freedom.
  - p-value = 0.147

- Conclusion: our data are not inconsistent with the average height of girls aged 7 to 10 in Lima being 123.5cm.
  - **Fail to reject** the null hypothesis that average height girls is 123.5cm.

Imperial College London

Question B1i(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.

- Average height among 301 boys was 124.10cm.
- Average height among 335 girls was 124.01cm.
- Male children were 0.09cm taller than the female children (95% CI -0.89–1.06cm).

Question B1i(iii)

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

- **Two-sample** t-test with **equal variance** (unequal variance t-test also acceptable)
- **Two-sided** alternative hypothesis

```
x_female <- perulung$height[perulung$sex =="female"]
x_male <- perulung$height[perulung$sex == "male"]
t.test(x_male, x_female, var.equal = TRUE)
```

  - t-statistic = 0.17 on 634 degrees of freedom
  - p-value = 0.8632

- **Fail to reject** the null hypothesis that male height = female height.
- Our **sample** does not provide evidence that height of male children is **different from** female children.

```
##      Two Sample t-test
##
## data:  x_male and x_female
## t = 0.17239, df = 634, p-value = 0.8632
## alternative hypothesis: true difference != 0
## 95 percent confidence interval:
##  -0.8881117  1.0590504
## sample estimates:
## mean of x mean of y
##  124.0980  124.0125
```

Question B1i(iv)

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

- H0: FEV1 for children with respiratory symptoms is ≥ FEV1 for children with no symptoms.
- H1: FEV1 for children with respiratory symptoms is less than FEV1 for children with no symptoms.

- Two-sample t-test, unequal variance (equal variance also justifiable)
- **One-sided** alternative hypothesis.

- 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).
  - t-statistic = -4.90 on 211.5 degrees of freedom; **one-sided** *p* <0.001

- Strong evidence to **reject the null hypothesis** of no difference.
- Conclude that FEV1 is statistically significantly lower for children with respiratory symptoms compared to those with no symptoms.

Question B1j

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?

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

- P(failing to reject each of 4 [true] null hypotheses) = $0.95^4 = 0.815$
- P(erroneously rejecting at least 1) = 1 - 0.815 = 18.5%.

Note this is much larger than $\alpha = 0.05$!

Imperial College London

Practicing skills

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.

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

Imperial College London

Question B2a

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?

Imperial College London

Question 2a

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.

Question B2a

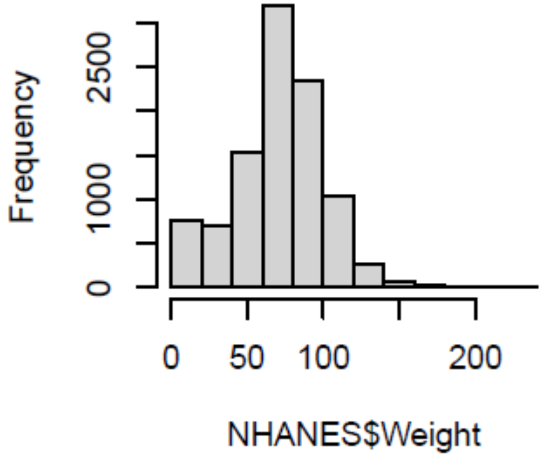
- Subset data to adults aged 20 years and older

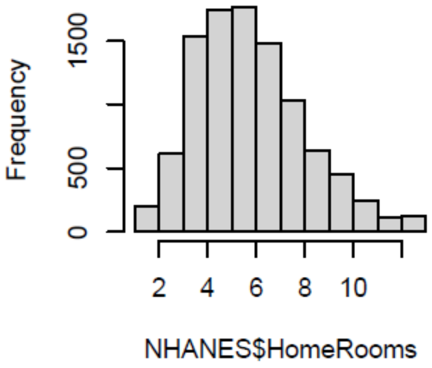
```
nhanes20pl <- NHANES[NHANES$Age >= 20, ]

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
```

Imperial College London

Question B2b

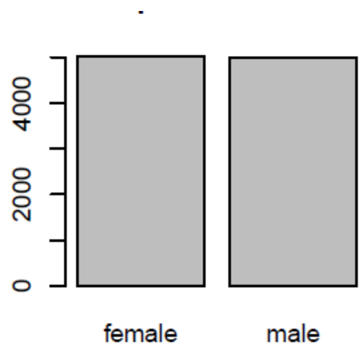
- Types of variables: identify at least one variable of each of the types of variables: continuous, discrete numeric, binary, categorical, and ordered categorical.

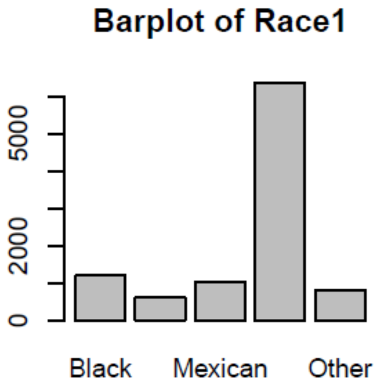


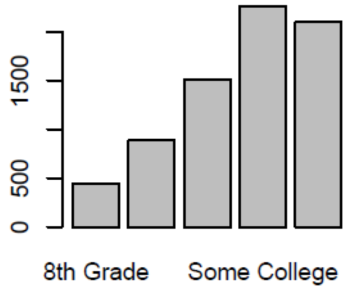


- Continuous variable: Weight — mean and standard deviation
- Discrete numeric variable: HomeRooms — median and interquartile range

  due to positive skew (mean & SD probably also reasonable)
- Binary variable: Gender — frequency table, frequency proportions
- Categorical variable: Race1 — frequency table, frequency proportions
- Ordered categorical: Education — frequency table, frequency proportions





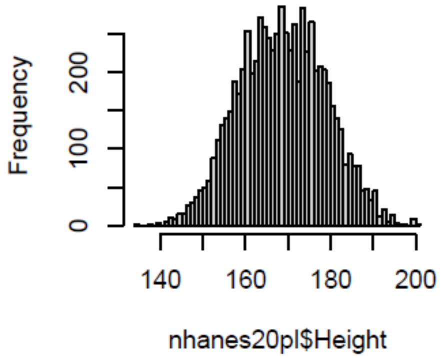


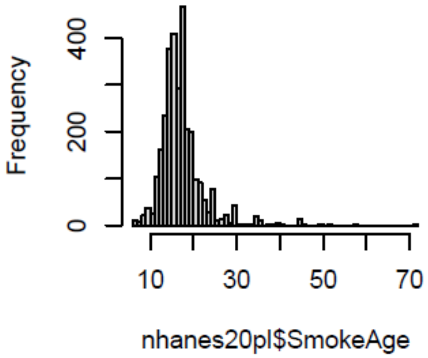
Imperial College London

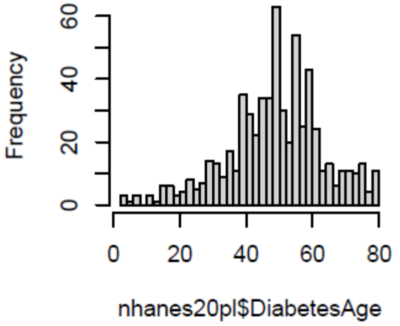
Question 2B2c

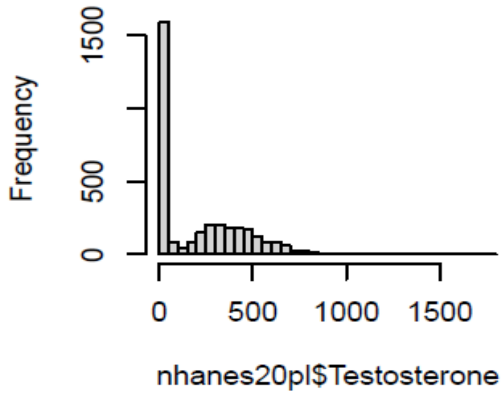
Shapes of frequency distributions

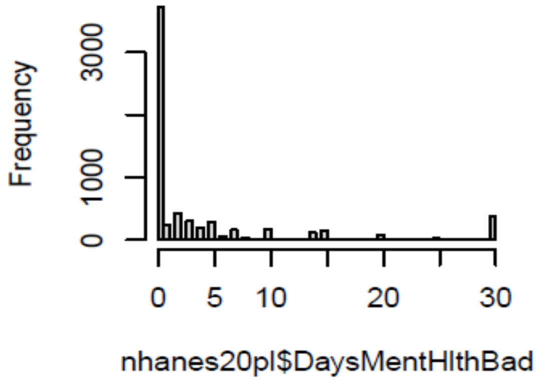
- Symmetric: Height
- Positive skew: SmokeAge
- Negative skew: DiabetesAge
- Bimodal: Testosterone
- Reverse-J: DaysMentHlthBad (tenuous example)
- Uniform: *no good example*











Imperial College London

Practicing skills

B3: Statistical modelling: using the normal distribution to estimate population distributions.

Practice applying the normal distribution to estimate the distribution of an outcome in a population using data from a sample drawn from the population.

Use dataset `nhanes20pl:` adult respondents aged 20 plus.

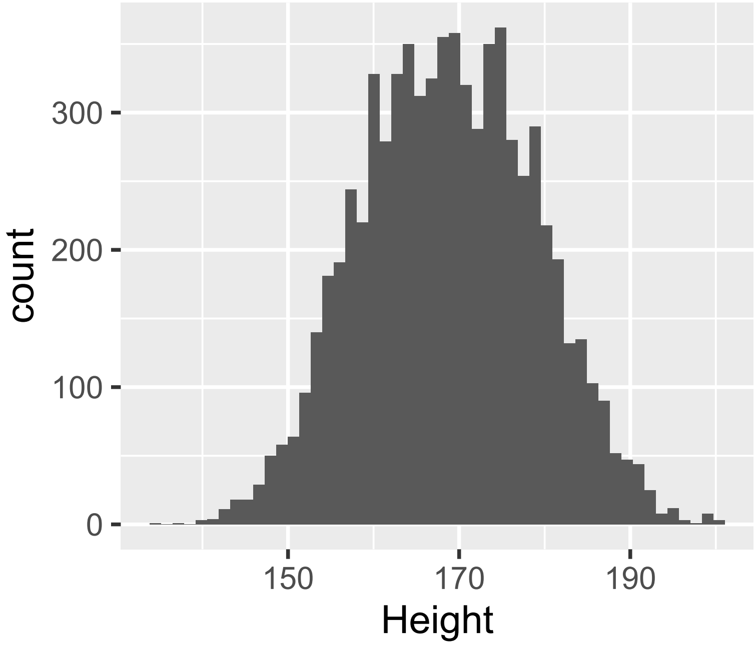
```
nhanes20pl <- NHANES[NHANES$Age >= 20, ]
```

Consider three continuous variables:

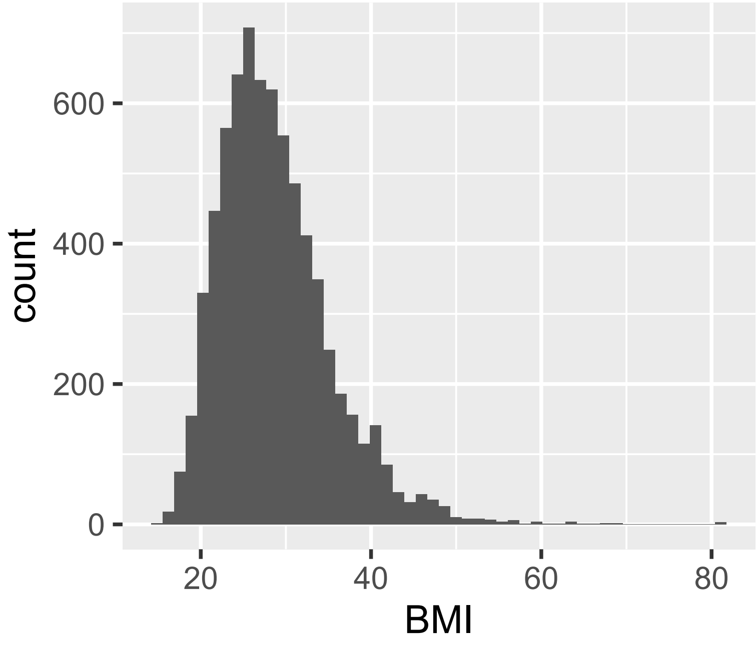
- `Height`: Standing height in centimeters.
- `BMI`: Body Mass Index
- `AlcoholYear`: Number of days over the past year that participant drank alcoholic beverages.

Question B3a

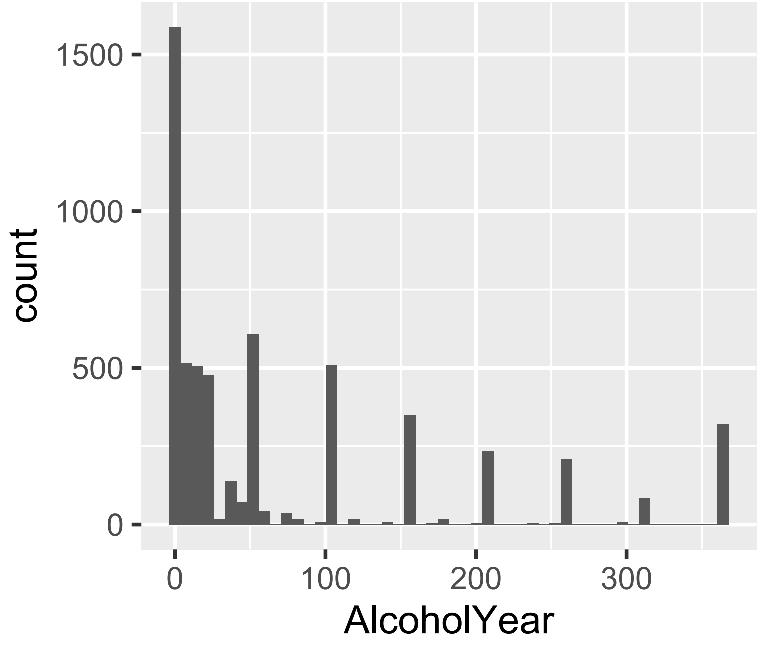
Does each outcome appear consistent with arising from a normal distribution?



Approximately normal



Slight positive skew



Severe positive skew

Question B3b

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

	Sample mean	Sample standard deviation
Height	168.8	10.1
BMI	28.8	6.7
AlcoholYear	75.7	103.6

```
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)
```

Imperial College London

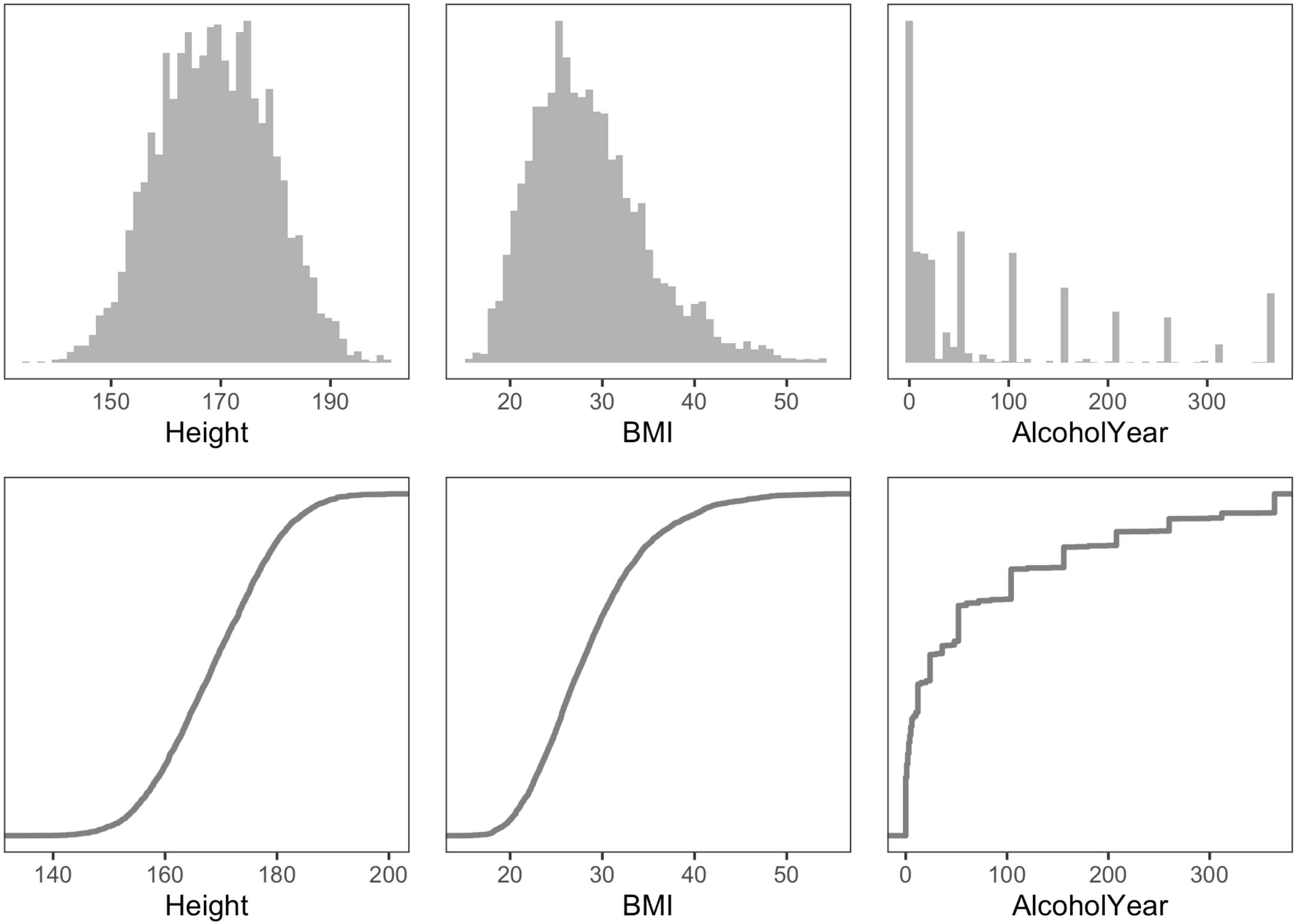
Question B3c

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

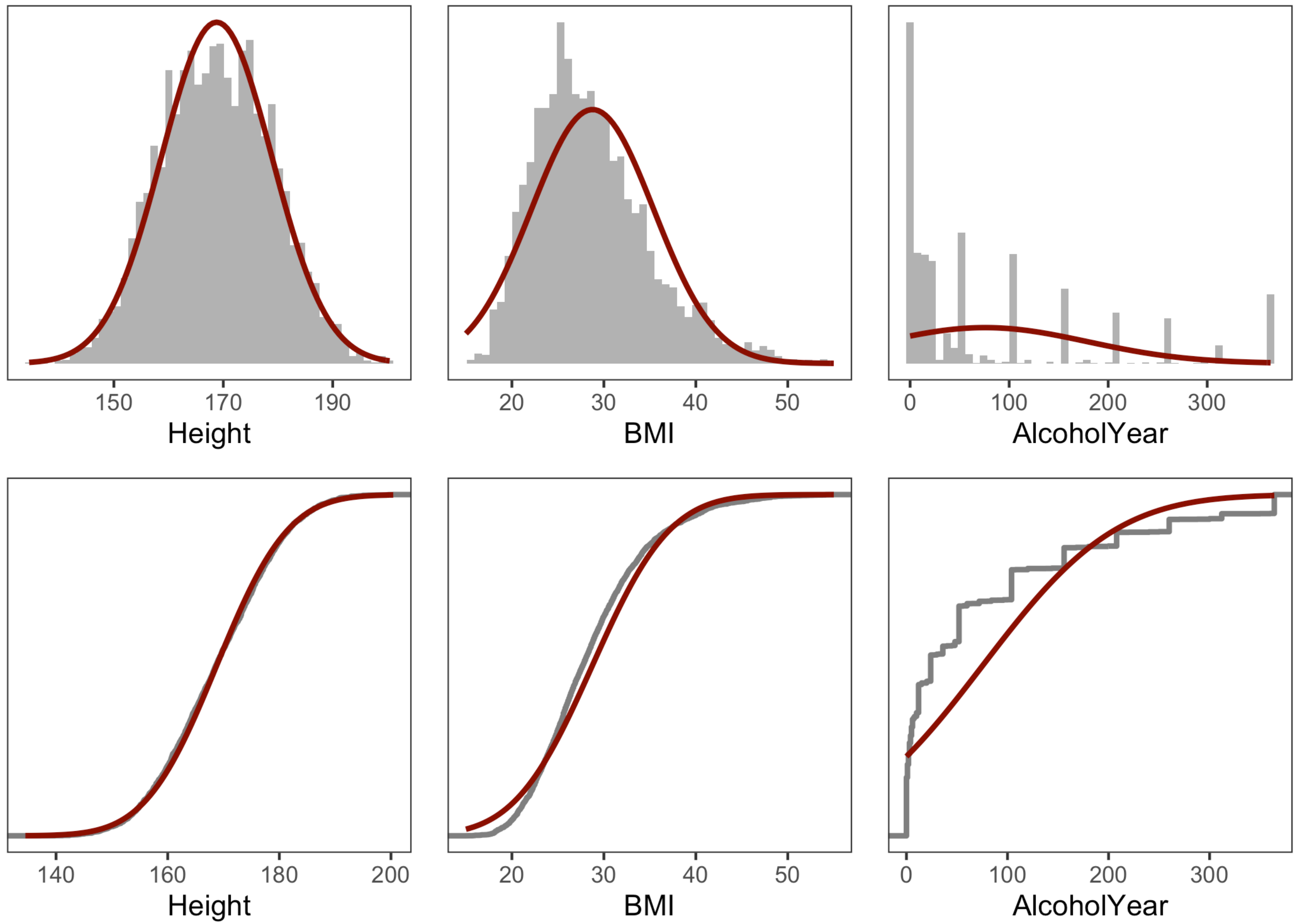
i. The proportion of adults who are above 165cm tall.
ii. The proportion of adults between 153cm and 160cm tall.
iii. The height of a door in order that 90% of adults can walk under without ducking.
iv. The proportion of adults who are obese, defined as BMI above 30.
v. The proportion of adults who are 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.

Imperial College London

Question B3c

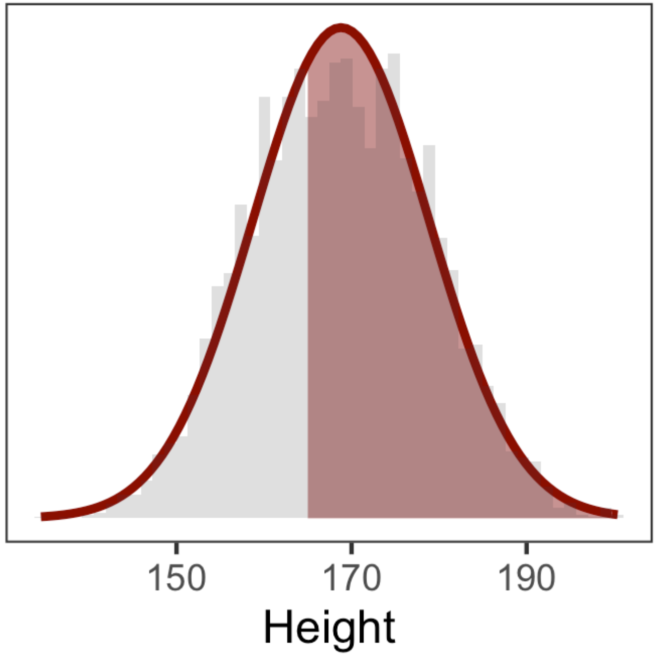


Question B3c

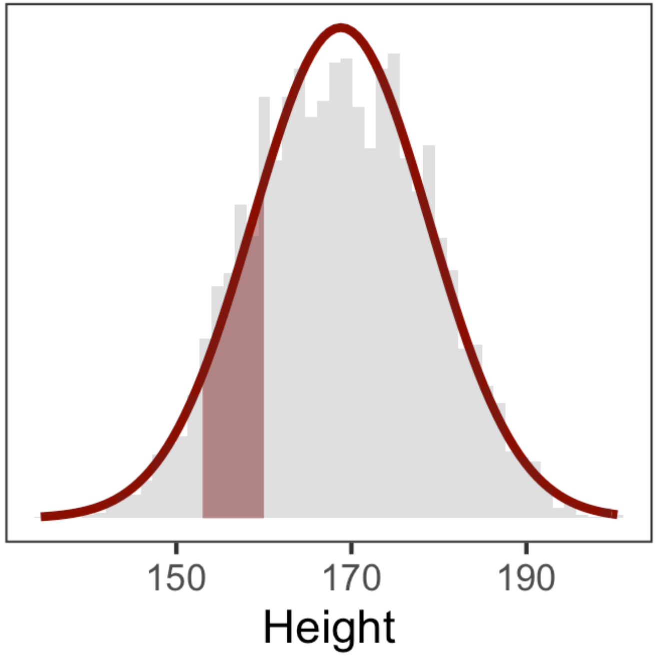


Question B3c(i-iii)

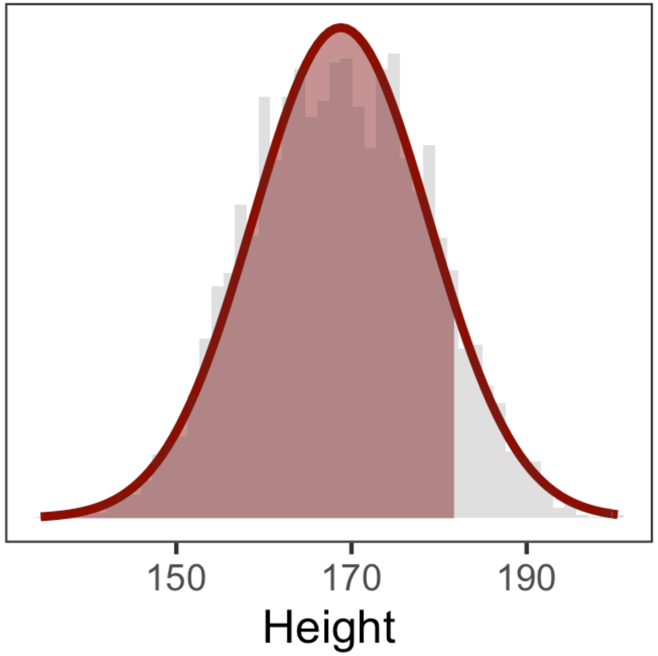
P(Height > 165cm)



P(153 < Height < 160cm)



P(Height < X) = 0.9



65% have height >165cm

```
1 - pnorm(165, height_mean, height_sd)
```

90% have height < **181.7cm**

```
qnorm(0.9, height_mean, height_sd)
```

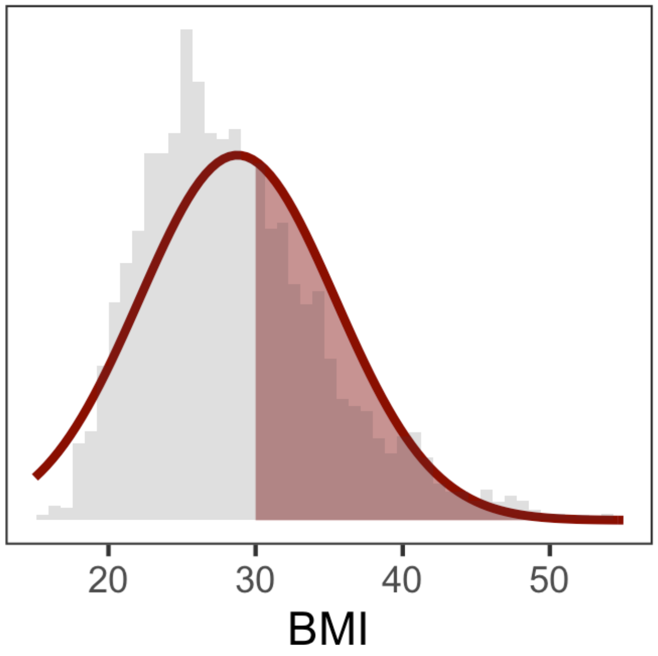
13% have height between 153 and 160cm

```
pnorm(160, height_mean, height_sd) -
pnorm(153, height_mean, height_sd)
```

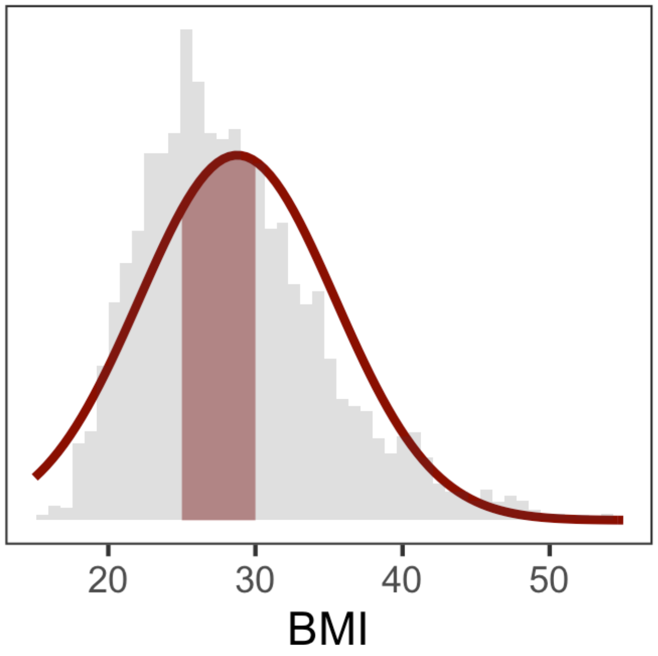
Imperial College London

Question B3c(iv-vi)

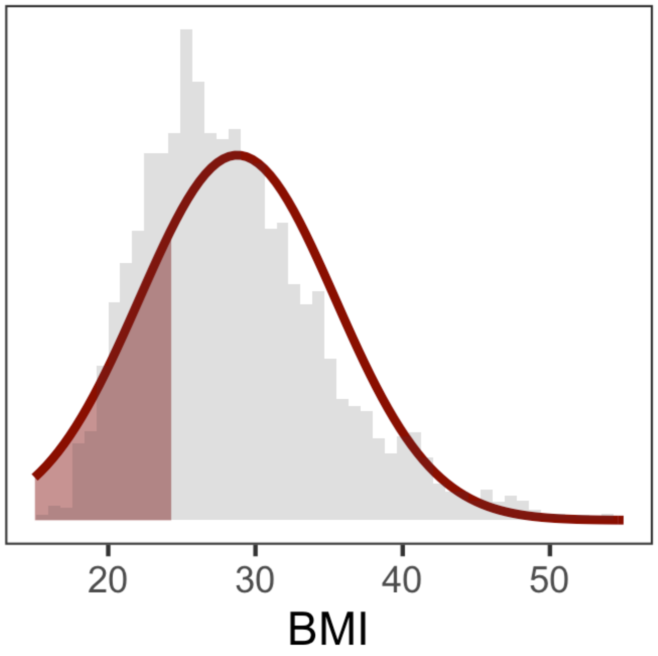
P(BMI > 30)



P(25 < BMI < 30)



P(BMI < X) = 0.25



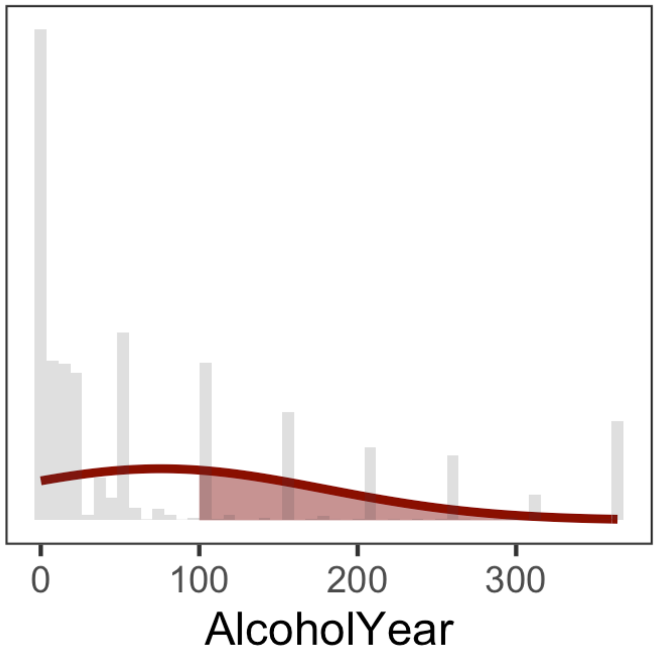
<u>**43%**</u> have BMI > 30

25% have BMI < <u>**24.3**</u>

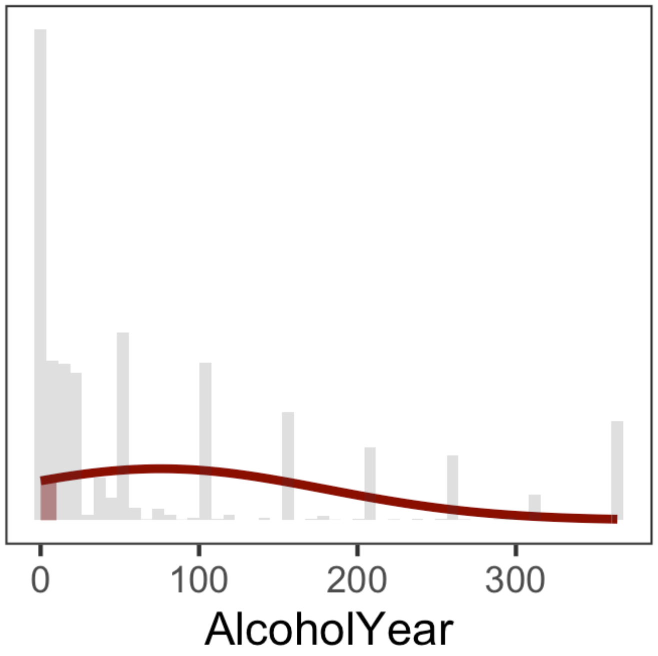
<u>**29%**</u> have height between 25 and 30

Question B3c(vii-ix)

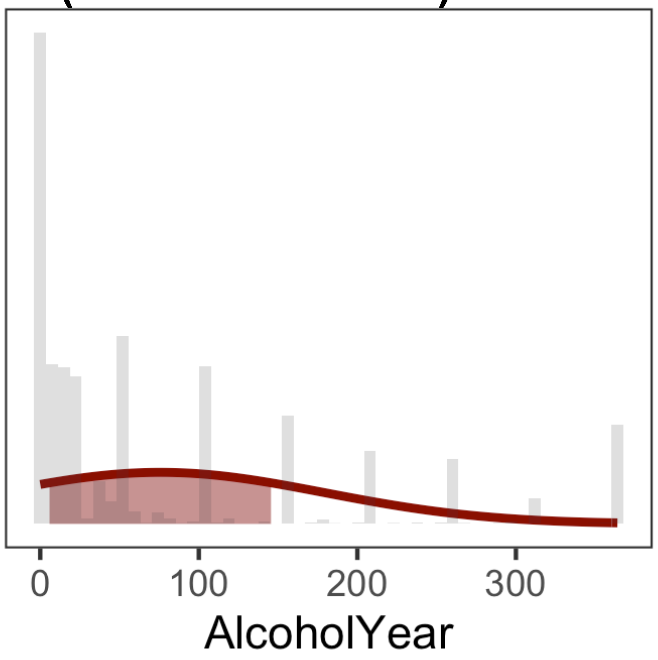
P(AlcoholYear > 100)



P(AlcoholYear ≤ 10)



P(AlcoholYear < X) = 0.25
P(AlcoholYear < Y) = 0.75



<u>41%</u> drink alcohol >100 days

IQR for days drinking alcohol **<u>139.7 days</u>**

<u>26%</u> drink alcohol 10 days or fewer

Imperial College London

Question B3d

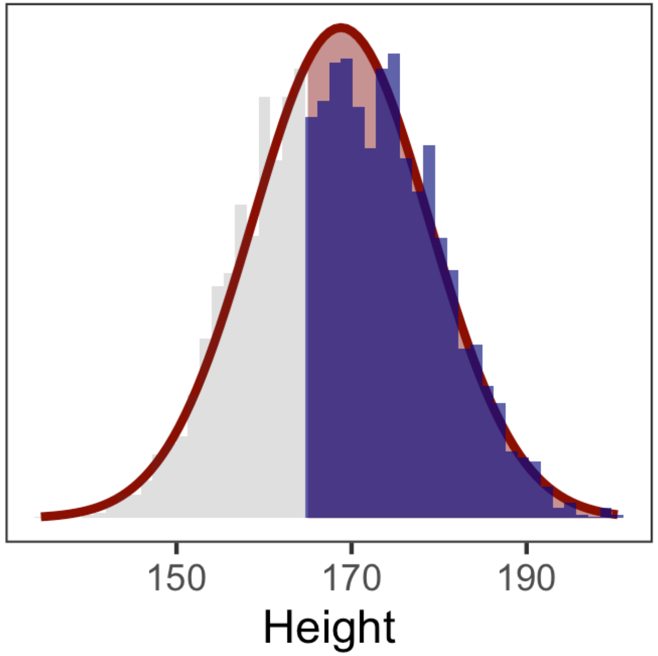
Check your estimates based on the normal distribution by directly calculating each of the above proportions amongst the observed sample.

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?

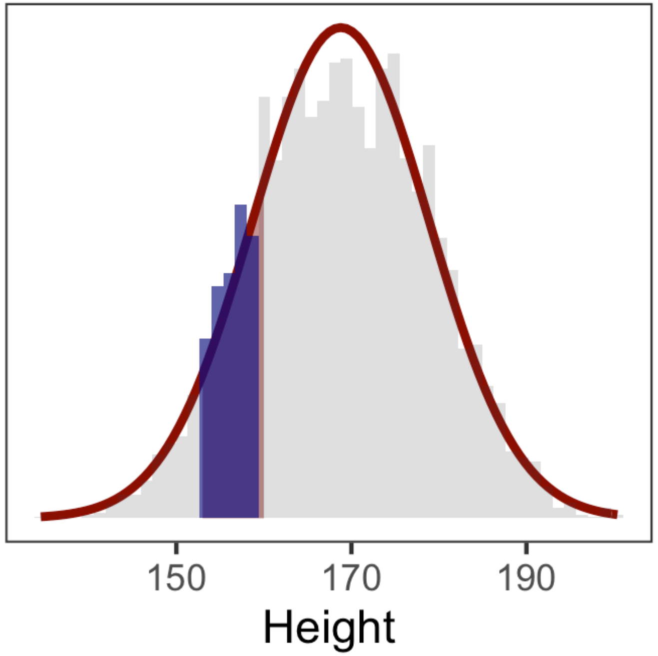
Imperial College London

Question B3d(i-iii)

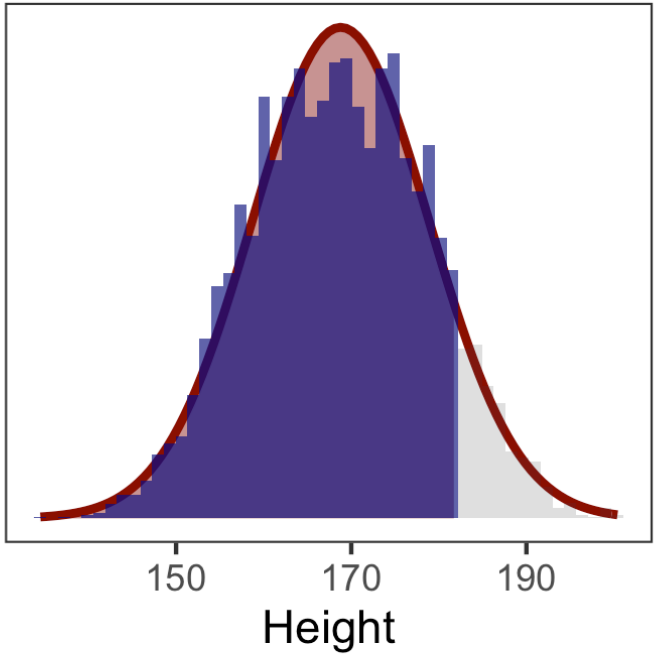
P(Height > 165cm)



P(153 < Height < 160cm)



P(Height < X) = 0.9



65% have height >165cm

<u>63%</u> of sample have height >165cm

```
mean(nhanes20pl$Height > 165, na.rm=TRUE)
```

90% have height < 181.7cm

90% of sample have height < **181.8cm**

```
quantile(nhanes20pl$Height, 0.9, na.rm=TRUE)
```

13% have height between 153 and 160cm

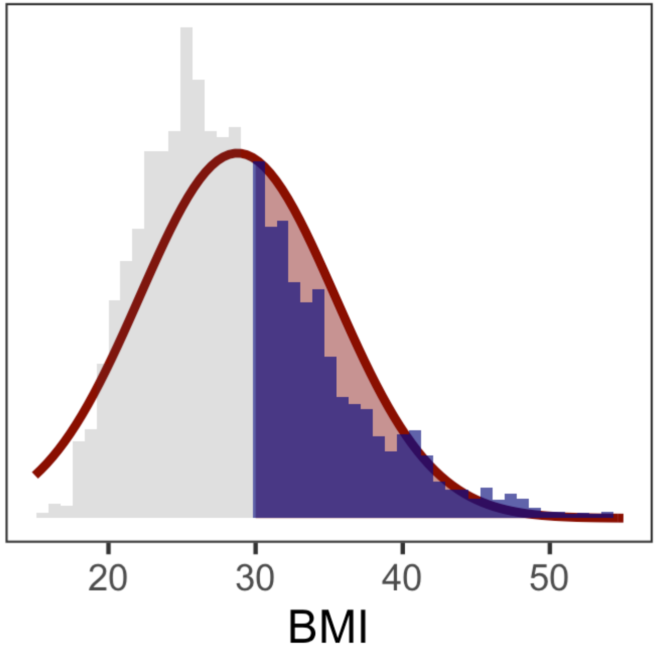
<u>15% of sample</u> have height between 153 and 160cm

```
mean(nhanes20pl$Height > 153 &
     nhanes20pl$Height < 160, na.rm=TRUE)
```

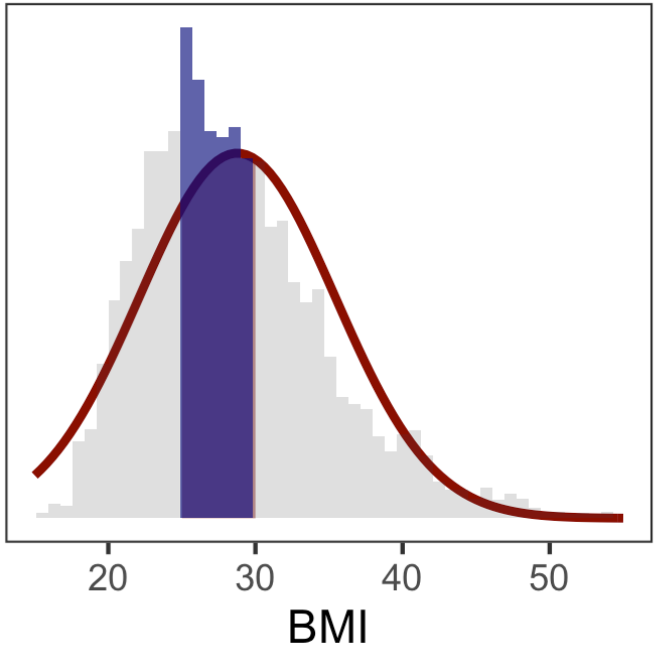
Imperial College London

Question B3d(iv-vi)

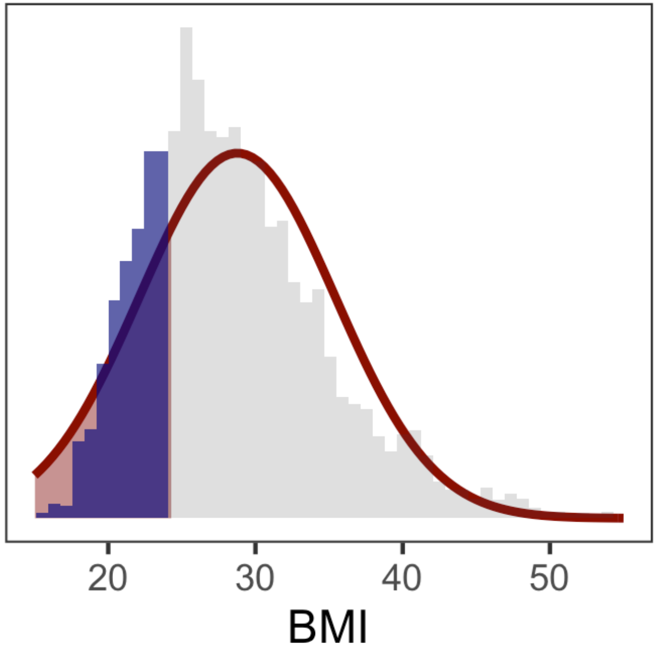
P(BMI > 30)



P(25 < BMI < 30)



P(BMI < X) = 0.25



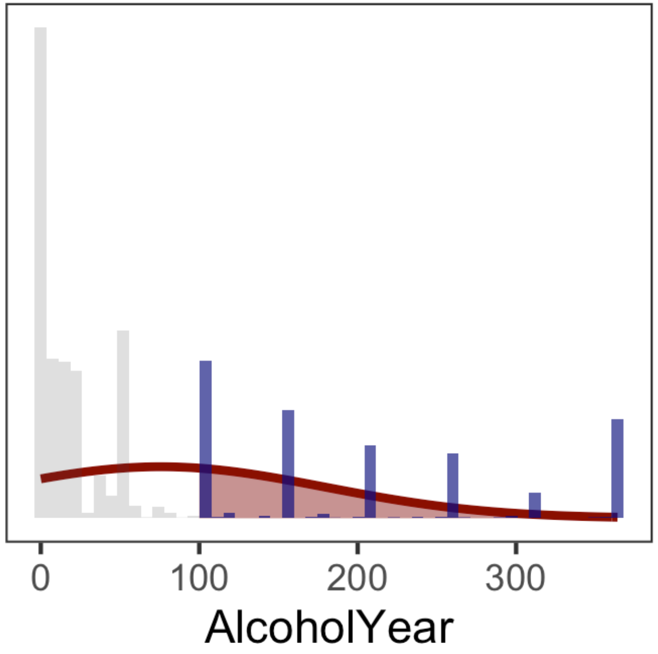
43% have BMI > 30
<u>36%</u> of sample have BMI > 30

25% have BMI < 24.3
25% of sample have BMI < **24.1**

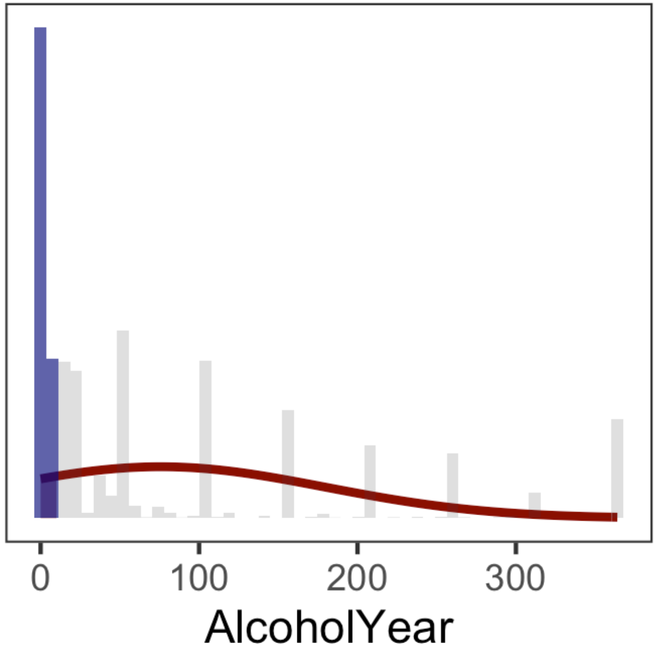
29% have BMI between 25 and 30
<u>33% of sample</u> have BMI between 25 and 30

Question B3d(vii-ix)

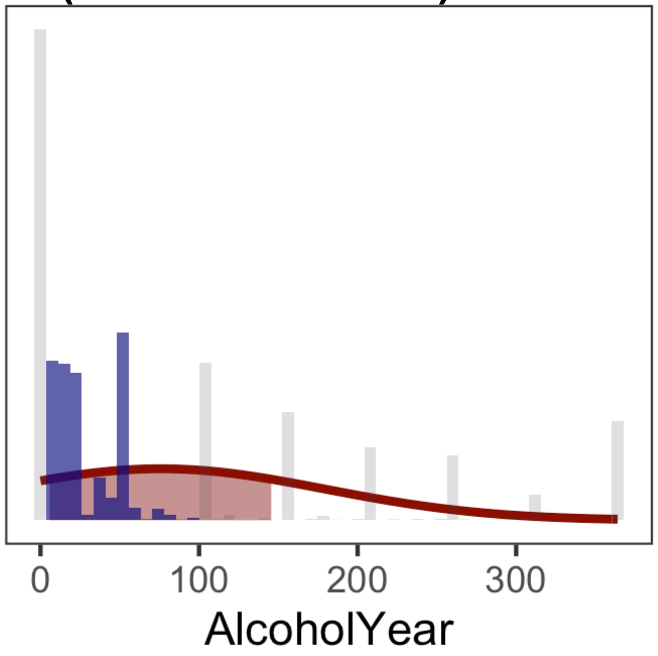
P(AlcoholYear > 100)



P(AlcoholYear ≤ 10)



P(AlcoholYear < X) = 0.25
P(AlcoholYear < Y) = 0.75



41% drink alcohol >100 days
<u>31% of sample</u> drink alcohol >100 days

IQR for days drinking alcohol 139.7 days
Sample IQR for days drinking alcohol **101 days**

26% drink alcohol 10 days or fewer
<u>36% of sample</u> drink alcohol 10 days or fewer

Imperial College London

Advanced learning

C1) Missing data

- Demographic variables:
  - 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:
  - 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).

Imperial College London

- HHIncome is slightly more likely to be missing for female respondents (9.0%) than male respondents (7.6%). HHIncome 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.

Imperial College London

Any questions?