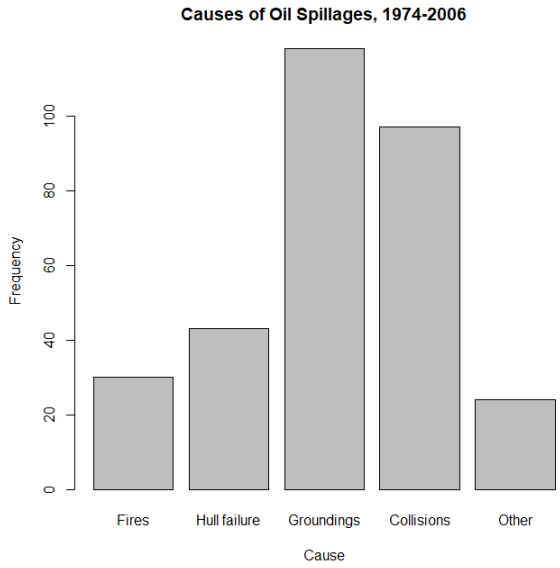
1. The frequencies of causes of large oil tankers spillages, between 1974 and 2006, are shown below:



Would you fit a Normal distribution to these data?

Yes **No**

Explain your choice of answer clearly:
The variable recorded in the data is categorical, being "cause of oil spillage". There is no numerical variable here and as such a Normal distribution is not appropriate.

2. At the time of the 2016 Olympic Games in Brazil, Health Canada were concerned with the spread of the Zika virus and public awareness of the health issue. Suppose Health Canada commissioned a survey with the goal of identifying the proportion of Canadian adults who were aware of the Zika virus. Suppose Health Canada obtained a random sample of 1500 from the target population and obtained information from each person in the sample about whether or not they were aware of the Zika virus.
What is the population parameter of interest in this scenario?

   (a) The number of adult Canadians in 2016 who were aware of the Zika virus.

(b) The 1500 people who were selected for the survey.

(c) The proportion of people surveyed who were aware of the Zika virus.

(d) **The proportion of Canadian adults in 2016 who were aware of the Zika virus.**

(e) Whether or not every Canadian adult in 2016 was aware of the Zika virus.

3. In the scenario in the previous question, what statistic from the sample would be used to estimate the parameter of interest?

(a) The number of people surveyed who were aware of the Zika virus.

(b) 1500

(c) **The number of people surveyed who were aware of the Zika virus, divided by 1500.**

(d) The standard deviation of the number of people surveyed who were aware of the Zika virus.

(e) Whether or not a person taken at random from those surveyed was aware of the Zika virus.

4. In the survey from Q2, suppose 486 responded that they were aware of the Zika virus. Which of the following best describes how the statistic estimating the parameter of interest would vary across repeated surveys of 1500 Canadian adults in 2016?

(a) A binary variable taking the value 0 with probability 1014/1500 and 1 with probability 486/1500.

(b) A Binomial distribution with $n = 1500$ and $p = 486/1500$.

(c) A Binomial distribution with $n = 486$ and $p = 486/1500$.

(d) **A Normal distribution with mean** $486/1500$ **and standard deviation**

$$\sqrt{\frac{(486/1500)\,(1 - 486/1500)}{1500}}.$$

(e) A Normal distribution with mean 486/1500 and standard deviation $\sqrt{(486/1500)\,(1-486/1500)}$.
Explain your answer clearly:
By the CLT, the sample proportion will approximately follow a Normal distribution with mean p and standard deviation

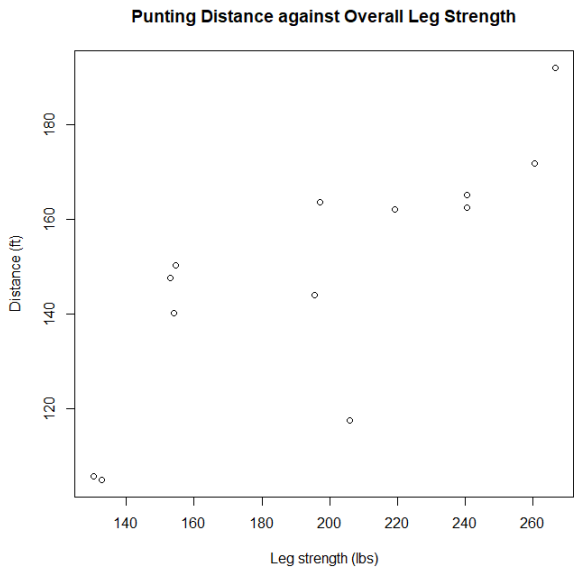
$$\sqrt{\frac{p\,(1-p)}{1500}}$$

where p is the population proportion aware of the virus. Since p is not known and we have a estimate from the sample, we can use the estimate in approximating the sampling distribution of the sample variance.

5. Northern hemisphere icebergs are believed to have a mean depth of 270 metres and a standard deviation of 25 meters. If we can assume that the depth of northern hemisphere icebergs is Normally distributed, approximately what proportion will have a depth greater than 295 metres? Show your working clearly.
We know that about 68% of the distribution will fall within one s.d. either side of 270 m. So about 32% lies outside that range. Since 295 m is one s.d. above the mean, about half of the 32% will fall above 295 m since the distribution is symmetric. Hence about 16% of the icebergs have depth at least 295 m.

6. In American football, each team has a designated kicker who is responsible for kicking the ball in certain plays. Researchers studied physical characteristics and ability in thirteen kickers in American football. Each volunteer kicked, or *punted*, a football ten times. The investigators recorded the average distance for the ten punts, in feet. The researchers also recorded a measurement of overall leg strength for each player, a measure of how much weight could be lifted by both legs (in

pounds). A plot of the data is given below:



(a) A linear model is fitted by least squares to the data, the model being

$$y = 63.38 + 0.43x.$$

Provide the units for the estimate 0.43.
The estimate 0.43 is in ft/lb. (1 mark)

(b) In the context of the study, provide a clear interpretation of the estimate of

i. the intercept

The estimate 63.38 ft is the estimated mean distance a player with zero lbs leg strength could kick the ball. This presumably has no physical meaning.

ii. the slope

For a 1 lb increase in leg strength, the model suggests the player will kick the ball on average 0.43 ft further.

(c) Use the model to estimate how far on average a kicker will punt the ball if their overall leg strength measure is 230 pounds.
The estimate is

$$63.38 + 0.43 \times 230 = 162.28$$

(ft).

(d) For the data given from the thirteen players, the correlation between the mean punting distance and the overall leg strength measure is 0.796. Compute the R^2 statistic for the data.
$R^2 = 0.634$

(e) In the context of the data, interpret the R^2 statistic.
About 63.4% of the variation in the punting distances is explained by overall leg strength.

(f) One of the kickers had an average punt of 192.0 feet and had a leg strength measure of 266.56 pounds. Find the residual from the model above for this kicker, showing your working clearly.
The fitted value for this kicker is

$$63.38 + 0.43 \times 266.56 = 178.0$$

feet. Therefore the residual for this kicker is

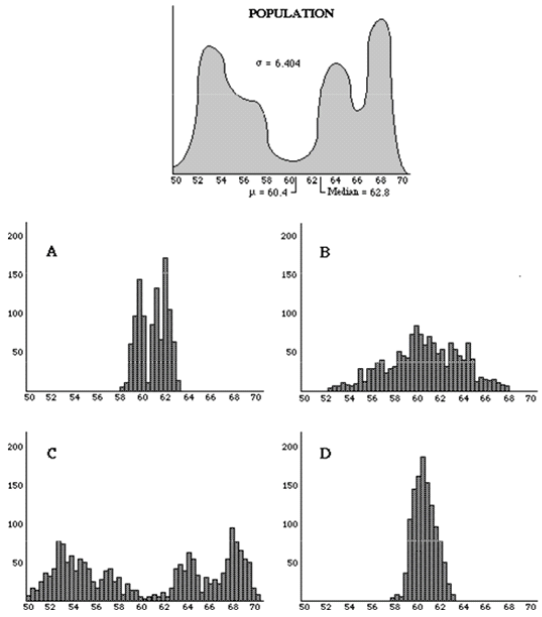
$$192.0 - 178.0 = 14.0$$

feet. Note the observation lies above the fitted line.

(g) Suppose another kicker is included in the study. This kicker has overall leg strength of 270 lbs and punted the ball on average 152 feet. If their data is added to the original data, would the new data point be influential? Explain your thinking clearly.
This observation would be very influential, in that it would have relatively high influence on the model fitted. The data point is on the extreme of the "X range", being the highest leg strength reading, yet the punting distance mean in well below the line fitted for that level of leg strength. Hence the observation would have high "leverage".

7. A hypothetical distribution for a population of test scores is displayed below. The population has a mean of 60.4, a median of 62.8, and a

standard deviation of 6.404. Each of the other four graphs labeled A to D represent possible distributions of sample means for random samples drawn from the population.



(a) Which graph best represents a distribution of sample means for 1000 samples of size 4?

i. A

ii. **B**

iii. C

iv. D

Explain your answer clearly.

Note here in the population $\mu = 60.4$, $\sigma = 6.4$, and the sample size is $n = 4$. So we expect most of the distribution for the sample mean to fall within 3×3.2 of 60.4, and for the distribution to be approximately Normal. Hence the distribution in B is similar to what we would expect.

(b) What do you expect for the shape of the (sampling) distribution of sample means for many samples of size $n = 4$?

i. **Shaped more like a Normal distribution than like the population distribution.**

ii. Shaped like a Uniform distribution.

iii. Shaped like a bimodal distribution.

iv. Shaped more like the population distribution than like a Normal distribution.

v. Shaped like neither the population nor the normal distribution.

(c) What do you expect for the variability (i.e., spread) of the (sampling) distribution of the sample mean from samples of size $n = 4$?

i. Same as for the population.

ii. **Less variability than the population (a less dispersed distribution).**

iii. More variability than the population (a more dispersed distribution).

(d) What do you expect for the shape of the distribution of a sample of size $n = 50$ from the distribution?

i. Shaped more like a Normal distribution than like the population distribution.

ii. Shaped like a Uniform distribution.

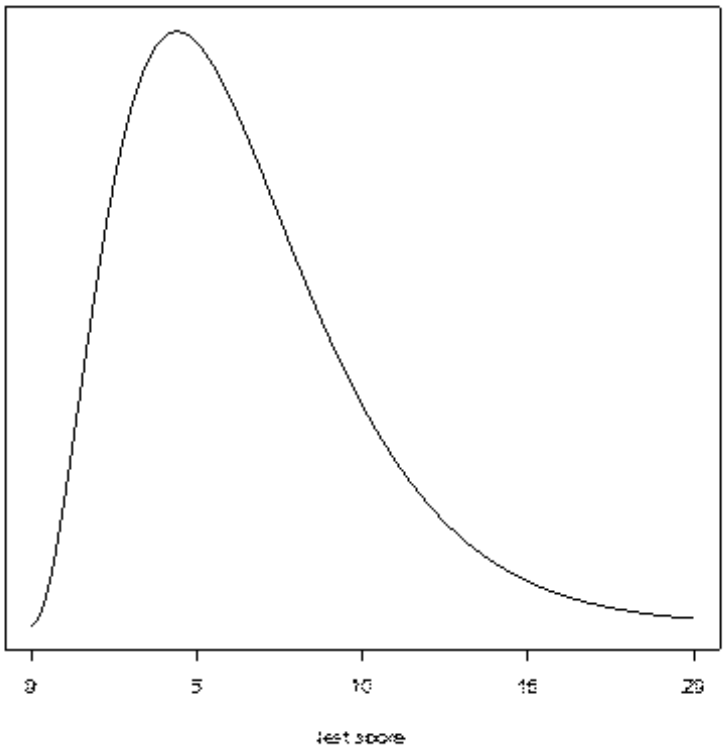
iii. Shaped like a unimodal distribution.

iv. **Shaped more like the population distribution than like a Normal distribution.**

v. Shaped like neither the population nor the normal distribution.

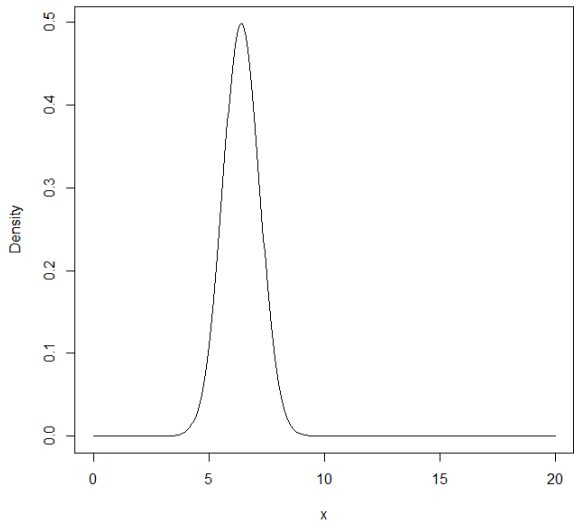
8. The graph below displays a distribution for a population of test scores. The mean score is 6.4 and the standard deviation is 3.6. Imagine we take a random sample of 20 scores from this distribution and calculate a mean. Suppose we repeat this process many times. Draw what you think the histogram of the sample means would look like. Sketch it

relative to the distribution.



Here $\mu = 6.4$, $\sigma = 3.6$, $n = 20$, *and so by the CLT the sample mean*

follows the Normal distribution with mean 6.4 *and s.d.* $3.6/\sqrt{20} = 0.80$.




```
curve(dnorm(x, mean=6.4, sd=0.8), from=0, to=20, n=500)
```