

Robert Davidson
ST1112: Statistics

70% Exam
30% Continuous Assessment (3 parts)

Contents

1	Descriptive Statistics	3
1.1	Sampling the mean	3
2	Interential Statistics - Interval Estimation	4
2.1	Confidence Intervals for a mean	4

1 Descriptive Statistics

1.1 Sampling the mean

In **probability** we consider the underlying process which has some randomness or uncertainty, and we try to figure out what happens

In **statistics** we consider the data that we have, and we try to figure out what the underlying process is. The basic aim is to infer the population from the sample.

Example Consider a jar of red and green jelly beans

A probabilist starts by knowing the proportion of red and green jelly beans in the jar, and then tries to figure out the probability of drawing a red jelly bean.

A statistician starts by drawing a sample of jelly beans from the jar, and then tries to figure out the proportion of red and green jelly beans in the jar.

Definition : Central Limit Theorem

Sample means follow a normal distribution, centered on the population mean, with a standard deviation equal to population standard deviation divided by the square root of the sample size.

$$\bar{X} \sim N\left(\mu, \frac{\sigma}{\sqrt{n}}\right)$$

Definition : Standard Error

The standard error is the variability in the sampling distribution.

The standard error describes the typical difference between the sample measurement and the population parameter.

$$SE = \frac{\sigma}{\sqrt{n}}$$

Definition : Estimate σ

Often the value of the population standard deviation is unknown, and hence the standard error of the mean is unknown.

We can estimate the value of the standard error using the sample standard deviation (s) as an unbiased estimator of the population standard deviation (σ).

$$\sigma_{\bar{X}} = \frac{s}{\sqrt{n}}$$

2 Inferential Statistics - Interval Estimation

2.1 Confidence Intervals for a mean

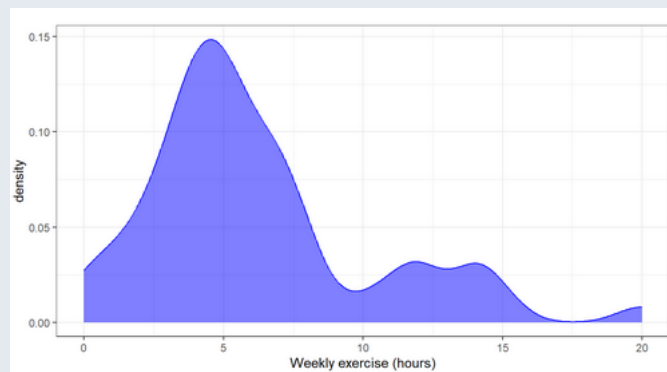
Example The student newspaper wants to know how many students are exercising per week on average

- Take a sample from this population
- Estimate the population parameter using the sample statistic

```
st1112_data %>%
  select(exercise) %>%
  summarise(n = n(),
            mean = mean(exercise, na.rm=TRUE),
            sd = sd(exercise, na.rm=TRUE))

##      n mean  sd
## 1  54 6.19 4.11

st1112_data %>%
  ggplot(aes(x = exercise)) +
  labs(x = "Weekly exercise (hours)") +
  geom_density(colour = "blue",
              fill="blue", alpha=0.5)+theme_bw()
```



But a new survey on another 54 students would lead to a different estimate, so which should we report back to the newspaper? If we sample data from the population, there is uncertainty in our estimate of the population mean. The standard error of the mean is a measure of this uncertainty. In our example, the standard error of the mean is:

$$SE = \frac{4.11}{\sqrt{54}} = 0.56$$

We use the Central Limit Theorem to provide a range of values that will capture 95% of the sample means

Definition Confidence Interval for $n > 30$

For a large sample size, $n > 30$, a Confidence Interval for the population mean is given by:

$$\bar{X} \pm Z_{\frac{\alpha}{2}} \frac{s}{\sqrt{n}}$$

For a 95% Confidence Interval, $\alpha = 0.05$, and $Z_{\frac{\alpha}{2}} = Z_{0.25}$.