Robert Davidson

ST1112: Statistics

70% Exam
30% Continuous Assessment (3 parts)

Contents

1 Descriptive Statisitcs

1.1 Sampling the mean

In **probability** we consider the underlying process which has some randomness or uncertainity, 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 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 distrubution, centered on the popular 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 distrubution.
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 Interential Statistics - Interval Estimation

2.1 Confidence Intervals

2.1.1 Confidence Intervals for a mean

Definition Confidence Interval for $n > 30$

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

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

A 95% Confidence Interval, $\alpha = 0.05$, meaning we accept a 5% risk that our interval doesn't contain the true population mean.
This 5% is split into 2.5% in each tail of the distrubution : $Z_{\frac{\alpha}{2}} = Z_{0.025}$.
When using a normal table that shows "area to the left", we need to find the Z value that corresponds to $1 - 0.025 = 0.975$. Thus: $Z_{0.025} = 1.96$.
95% Confidence is most commonly used because increasing the confidence level increases the width of the interval, this may not be useful.

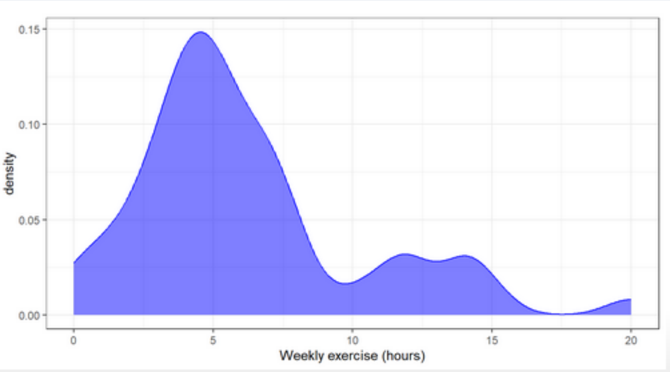
Example The student newspaper wants to know how many students are exercising per week on average

- Take a **sample from this population**
- Esimate the **population paramater** 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:

$$SE = \frac{4.11}{\sqrt{54}} = \approx 0.6$$

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

$$\bar{X} \pm Z_{\frac{\alpha}{2}} \times \frac{s}{\sqrt{n}} = 6.23 \pm 1.96 \times 0.6 = 5, 7.4$$

Example : Confidence intervals in R

```
st1112_data %>% select(exercise) %>% t.test()
##
##  One Sample t-test
##
## data:  .
## t = 11, df = 53, p-value = 2e-15
## alternative hypothesis: true mean is not equal to 0
## 95 percent confidence interval:
##  5.06 7.31
## sample estimates:
## mean of x
##      6.19
```

2.1.2 Confidence Intervals for a small sample size

When σ is known, a 95% Confidence Interval for the population mean is given by:

$$\bar{X} \pm 1.96 \cdot \frac{\sigma}{\sqrt{n}}$$

When σ is unknown, a 95% Confidence Interval for the population mean is given by:

$$\bar{X} \pm 1.96 \cdot \frac{s}{\sqrt{n}}$$

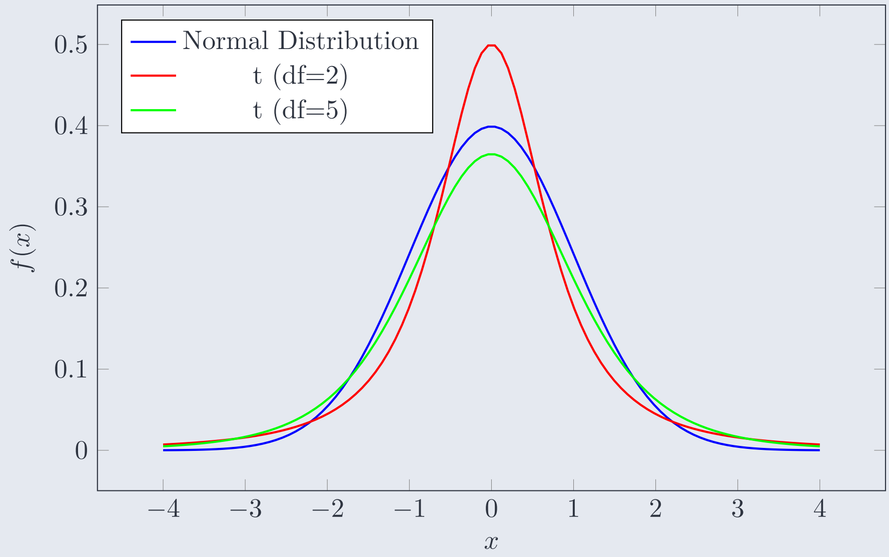
But, what if σ is unknown and $n < 30$?

Definition Confidence Interval for $n < 30$

For a small sample size, $n < 30$, we use the **t-distribution instead of the normal distribution**. The t-distribution has heavier tails than the normal distribution and accounts for the additional uncertainty when estimating σ with s in small samples.
The confidence interval is given by:

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

Where $t_{\frac{\alpha}{2}}$ is the critical value from the t-distribution with $n - 1$ **degrees of freedom**.
The degrees of freedom represent the *number of independent pieces of information used to estimate the standard deviation*. As the degrees of freedom increase, the t-distribution approaches the normal distribution - infinite degrees of freedom is the normal distribution.



Example : Find t-value from tables For a 95% Confidence Interval

We have:

$$a = 1 - 0.95 = 0.05 \Rightarrow \frac{a}{2} = 0.025$$

We also have:

$$n = 13 \Rightarrow df = 13 - 1$$

We want to find the $t_{12,0.025}$ value from the t-distribution table, which is 2.179.

Example : Finding the t-value using R

```
qt(1 - 0.025, df = 12)
## [1] 2.179
```

2.2 Bootstrap, proportions and counts

What do we do when our data is not normally distributed? We have two options:

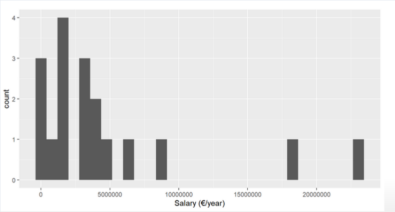
- Transform the data to make it normally distributed (e.g. log transformation, square root transformation)
- Use a non-parametric method (Bootstrap or CI for the population median)

2.2.1 Log Transform of the Data

Example : NBA Salaries

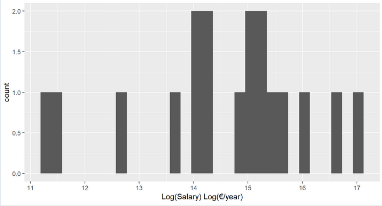
```
sals <- read.csv("data/NBA_season1718_salary.csv")

sals %>%
    filter(Tm=="DAL", season17_18>50000) %>%
        ggplot(aes(x=season17_18)) +
        geom_histogram()  + xlab("Salary␣($/year)")
```



The data is right-skewed, so we can apply a log transformation to make it more normally distributed.

```
sals %>%
filter(Tm=="DAL", season17_18>50000) %>%
    ggplot(aes(x=log(season17_18))) +
    geom_histogram() + xlab("log(Salary)")
```



We can now calculate the mean and standard deviation of the log-transformed data.

```
sals %>% filter(Tm=="DAL", season17_18>50000) %>%
    mutate(log_sal = log(season17_18)) %>%
        select(log_sal) %>% summarise(n=n(),
                        mean=mean(log_sal),
                        sd=sd(log_sal))

##     n mean   sd
## 1 18 14.5 1.57

qt([p = 1 - 0.025, df = 18-1)])
## [1] 2.11
```

Produce an interval estimate for the populat mean

```
14.5 - 2.11 * (1.57/sqrt(18))
14.5 + 2.11 * (1.57/sqrt(18))
## [1] 13.7 ## [1] 15.3
```

Take exponentials of the resulting interval to get back to original scale

```
exp(14.5 - 2.11 * (1.57/sqrt(18)))
exp(14.5 + 2.11 * (1.57/sqrt(18)))
## [1] 908172 ## [1] 4238843
```

2.2.2 Bootstrap

We can quantify the uncertainity in our estimate of the population mean by using the Central Limit Theorem or simulatiuon via the Bootstrap method, as follows:

1. Take a bootstrap sample - random sample taken with replacment from the original sample (same size as original sample)
2. Calculate the bootstrap statistic - such as mean, median, proportion, etc.
3. Repeat steps 1 and 2 many times
4. Calculate the bounds of the 95% Confidence interval as the middle 95% of the bootstrap distrubution

Example : Simple bootstrap

```
my_date <- (1,1,1,2,3,4,5)
mean(my_data)
## [1] 2.43

sample(my_data, replace = FALSE)
## [1] 1 1 3 2 1 5 4

sample(my_data, replace = TRUE)
## [1] 4 1 1 1 3 3 1

replicate(1000, mean(sample(my_data, replace = TRUE)))
## [1]
```

3 Inferential Statistics - Hypothesis Testing

3.1 Hypothesis Testing

A hypothesis test is intended to assess whether a population parameter of interest is equal to some specified value of direct interest to the researcher
Hypothesis tests are structured in a very specific and, what may seem initially, peculiar manner.
The p-value is central to the notion of a hypothesis test.
The Central Limit Theorem (CLT) and t-distribution provide the framework for assessing how different the sample statistic is from the proposed parameter value.

3.1.1 Stages in Hypothesis Testing

1. State the null (H_0) and alternative (H_1) hypotheses
2. Take a random sample from the populat of interest and calculate a suitable test statistic (T_0) under the assumed model (H_0)
3. Write down the distrubution that the test statistic follows
4. Investigate how likely the value of the test statistic is if the null hypothesis is true
5. Make a decision to reject or fail to reject the null hypothesis (using 3 and 4)
6. Write down the conclusion

The **Null Hypothesis** is the hypotheses that the population statistic is equal to some claimed value.
The **Alternative Hypothesis** is the hypotheses that the population statistic is not equal to the claimed value. It must be true if the null hypothesis is false.
We asses through a test statistic, how probable (p-value), it would be to observe data as or more extreme than the data we have, if the null hypothesis is true.