Output development using R & R markdown

Ziv Shkedy and Thi Huyen Nguyen
Hasselt University

Foundations for inference
Ha Noi
03/03/25-07/03/25

Visit us on Facebook ER-BioStat

GitHub https://github.com/eR-Biostat

twitter @erbiostat

Case study 3:
The NHANES dataset: number of sleep hours per night

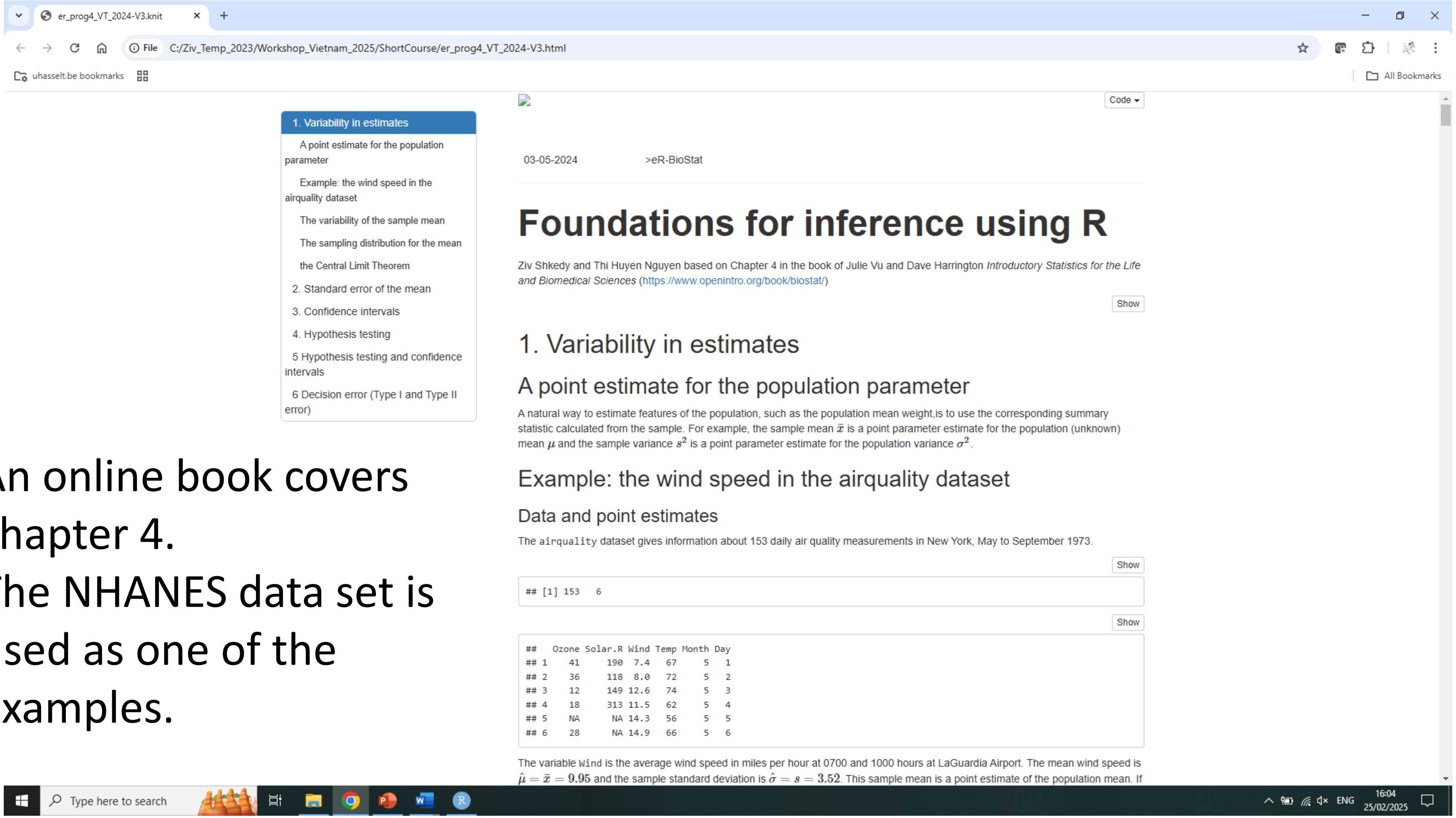
The case study

- How to use the HTML book for a simple analysis in one population:
  - Point estimates.
  - Confidence intervals.
  - Hypothesis testing.
- R code is a part of the book.

The NHANES data set

- The NHANES dataset consists of data from the US National Health and Nutrition Examination Study.
- Information about 76 variables is available for 10000 individuals included in the study.
- The 10000 individuals are considered as the population.

The HTML book

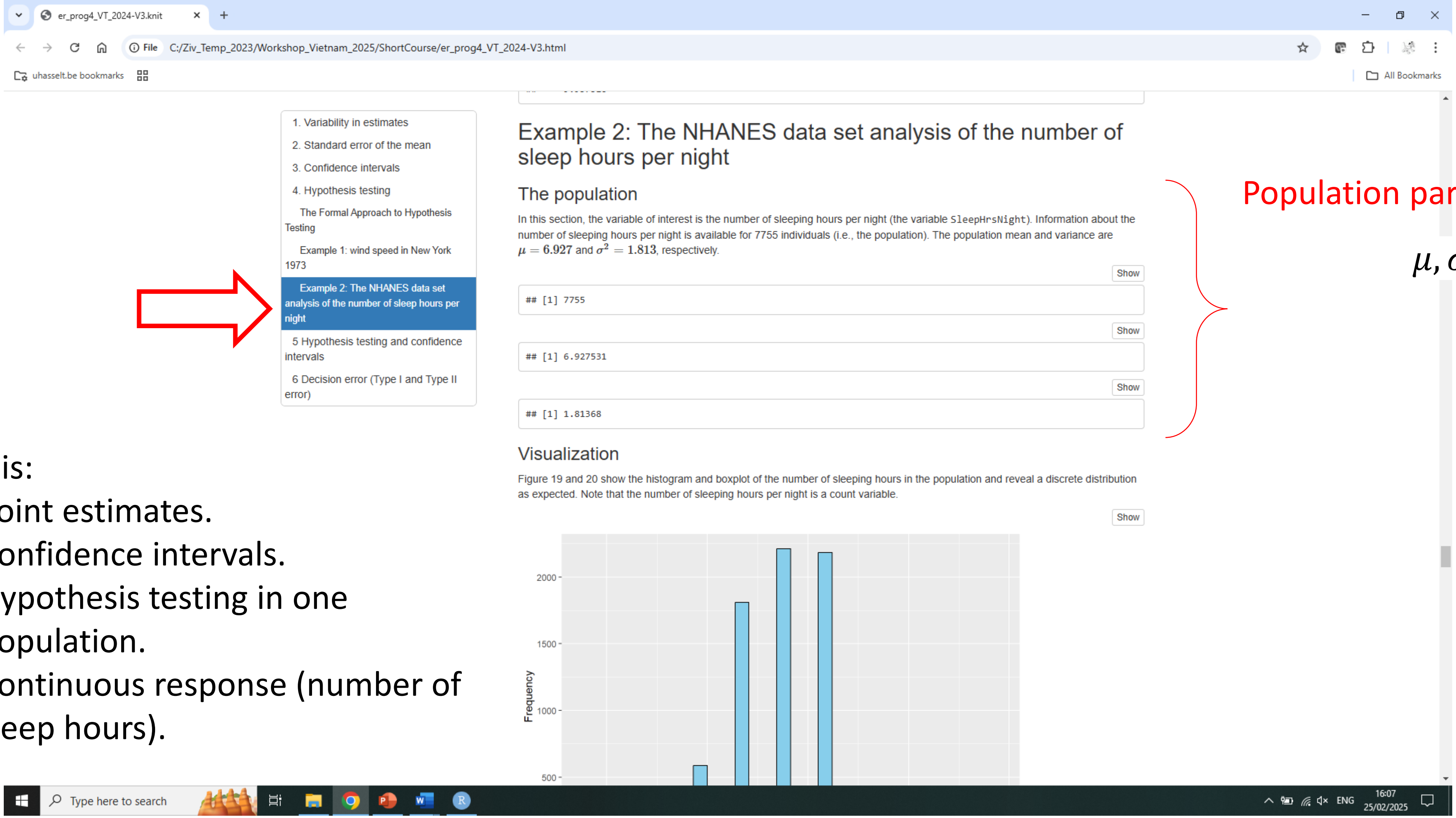


- An online book covers chapter 4.
- The NHANES data set is used as one of the examples.

The NHANES data set: analysis of the number of sleep hours per night

- The variable of interest is the number of sleeping hours per night (the variable `SleepHrsNight`).
- Continuous variable.
- Information about the number of sleeping hours per night is available for 7755 individuals (i.e., the population).

The NHANES data set:
analysis of the number of sleep hours per night



Population parameters:

μ, σ

- Analysis:
  - Point estimates.
  - Confidence intervals.
  - Hypothesis testing in one population.
  - Continuous response (number of sleep hours).

The NHANES data set analysis of the number of sleep hours per night

1. Variability in estimates

2. Standard error of the mean

3. Confidence intervals

4. Hypothesis testing

The Formal Approach to Hypothesis Testing

Example 1: wind speed in New York 1973

Example 2: The NHANES data set analysis of the number of sleep hours per night

5 Hypothesis testing and confidence intervals

6 Decision error (Type I and Type II error)

The population

In this section, the variable of interest is the number of sleeping hours per night (the variable `SleepHrsNight`). Information about the number of sleeping hours per night is available for 7755 individuals (i.e., the population). The population mean and variance are $\mu = 6.927$ and $\sigma^2 = 1.813$, respectively.

Hide

```
library(NHANES)
data(NHANES)
#dim(NHANES)
sleep<-na.omit(NHANES$SleepHrsNight)
length(sleep)
```

```
## [1] 7755
```

Hide

```
mean(sleep)
```

Population mean

```
## [1] 6.927531
```

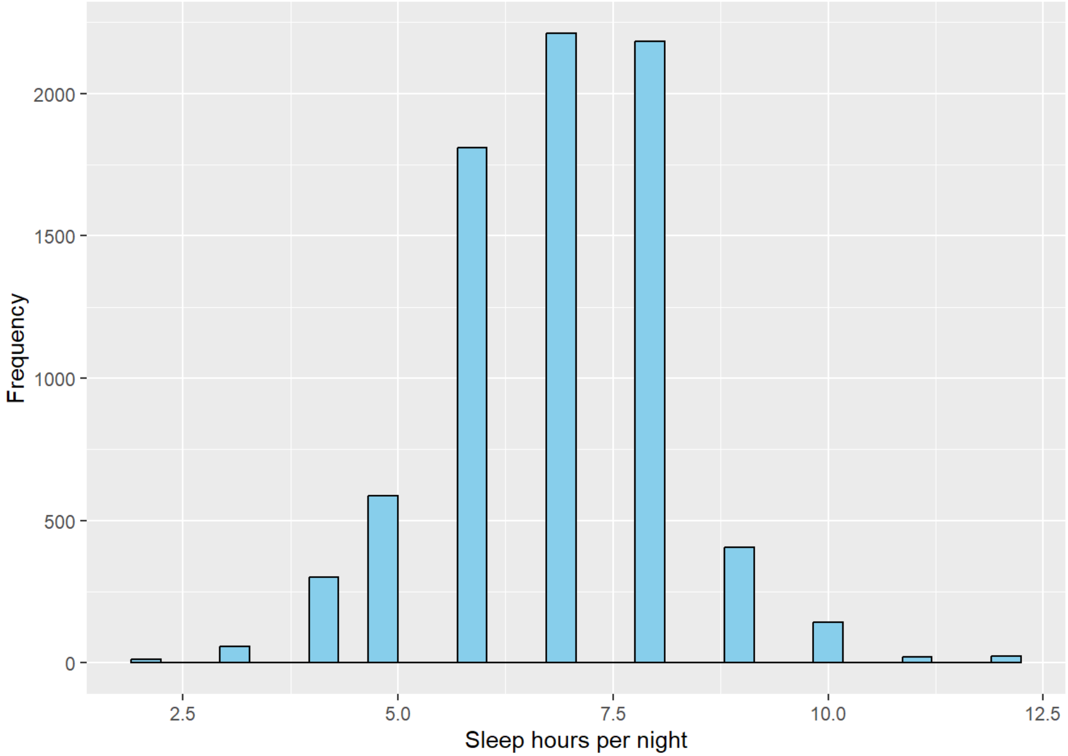
Hide

```
var(sleep)
```

Population variance

```
## [1] 1.81368
```

The number of sleep hours per night in the population



$$n = 7755$$
$$\mu = 6.927$$
$$\sigma = 1.813$$

Visualization

1. Variability in estimates
2. Standard error of the mean
3. Confidence intervals
4. Hypothesis testing

The Formal Approach to Hypothesis Testing

Example 1: wind speed in New York 1973

Example 2: The NHANES data set analysis of the number of sleep hours per night

5 Hypothesis testing and confidence intervals

6 Decision error (Type I and Type II error)

Hide

```
ggplot(NHANES, aes(x = SleepHrsNight)) +
  geom_histogram(fill = "skyblue", color = "black")+
  ylab("Frequency")+
  xlab("Sleep hours per night")
```

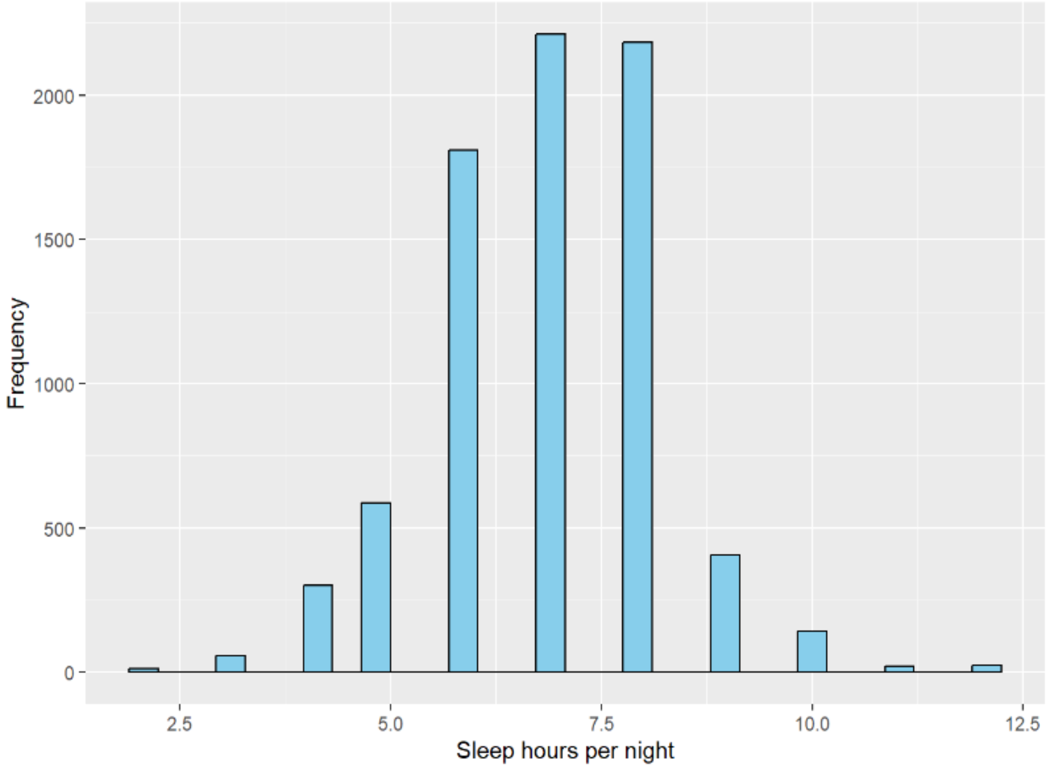


Figure 19: Histogram of sleep hours per night.

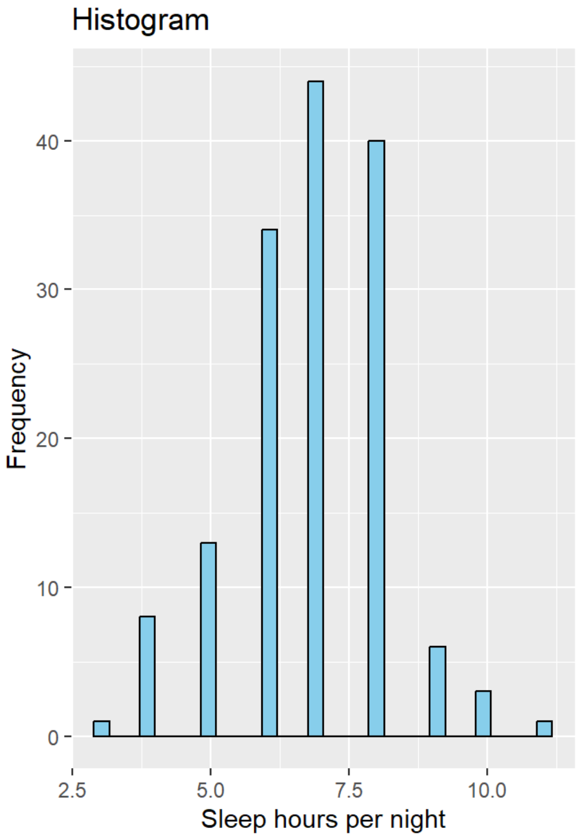
Case study 3:

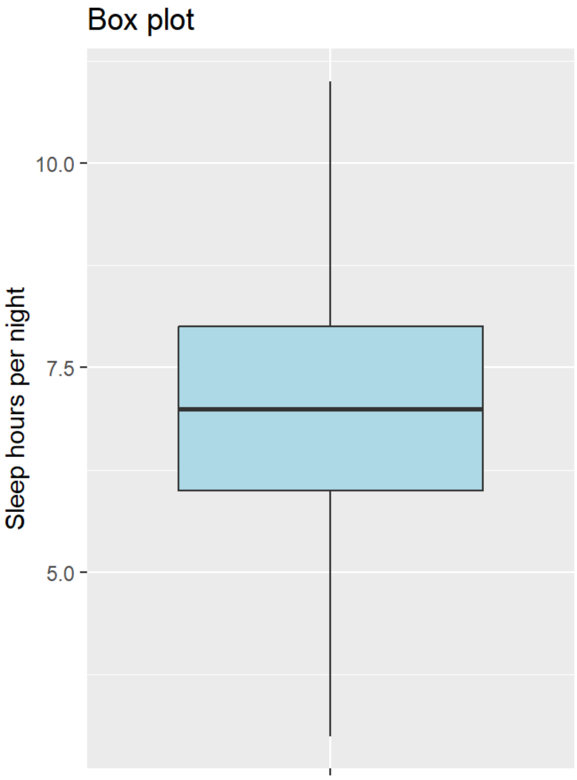
Point estimates

A random sample from the population

- Population size: 7755.
- We draw a random sample from the population.
- Sample size: 150.

A random sample from the population





- A random sample from the population:

$$n = 150$$
$$\bar{x} = 6.846$$
$$s^2 = 1.862$$

A random sample from the population

1. Variability in estimates

2. Standard error of the mean

3. Confidence intervals

4. Hypothesis testing

The Formal Approach to Hypothesis Testing

Example 1: wind speed in New York 1973

Example 2: The NHANES data set analysis of the number of sleep hours per night

5 Hypothesis testing and confidence intervals

6 Decision error (Type I and Type II error)

A random sample of size 150 from the population

We draw a sample of 150 indivduals from the population ($n = 150$). The point estimates for the sample are $\bar{x} = 6.8466$ and $\sigma^2 = 1.8622$.

Hide

```
set.seed(456789)
x.sleep<-sample(na.omit(NHANES$SleepHrsNight),size=150,replace=FALSE)
length(x.sleep)
```

```
## [1] 150
```

Show

```
## [1] 6.846667
```

$$\bar{X} = \frac{1}{n}\sum_{i=1}^{n} X_i$$

Hide

```
var(x.sleep)
```

```
## [1] 1.862237
```

$$S^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \bar{X})^2$$

Sample mean and variance.

A random sample from the population

1. Variability in estimates
2. Standard error of the mean
3. Confidence intervals
4. Hypothesis testing

5 Hypothesis testing and confidence intervals

Two sided alternatives and confidence intervals

Example: The NHANES dataset - analysis of the total cholestrol level

6 Decision error (Type I and Type II error)

Hide

```
box_sleep = ggplot(data.frame(SleepHrsNight = x.sleep), aes(x = "", y = SleepHrsNight)) +
  geom_boxplot(fill = "lightblue")+
  xlab("")+
  ylab("Sleep hours per night")+
  ggtitle("Box plot")
```

Hide

```
grid.arrange(hist_sleep, box_sleep, ncol = 2)
```

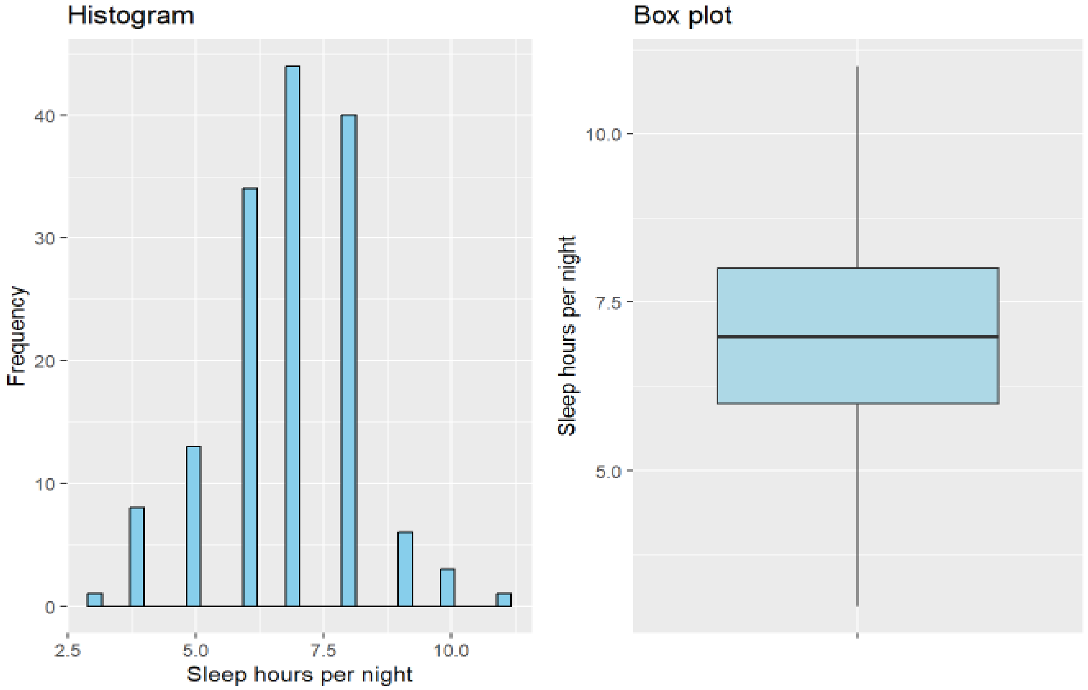


Figure 21: Histogram and box plot of sleep hours per night in the sample.

Case study 3:

Confidence interval for the population mean

Confidence interval for the population mean (Case 2)

If $X \sim F$

Then: $\bar{X} \sim N(\mu, \frac{S^2}{n})$

and $T_{\bar{X}} = \frac{\bar{X} - \mu}{\sqrt{\frac{S^2}{n}}} \sim N(0,1)$

3. X has an unknown distribution, but we have a large sample (n> 30)

$$E(X) = \mu$$

$$Var(X) = \sigma^2$$

The same as case 1 but we replace σ^2 by S^2.

C.I. for case 2

Step 1: example, choose 1-α = 0.95

Step 2: case 2, so :

$$\boxed{\frac{\bar{X}-\mu}{\sqrt{\frac{\sigma^2}{n}}} \sim N(0,1)} \quad \text{or} \quad \boxed{\frac{\bar{X}-\mu}{\sqrt{\frac{S^2}{n}}} \sim N(0,1)}$$

Step 3: critical points: -1.96 and 1.96

(the same as in Case 1, since we are still using the standard normal distribution function)

Step 4: Calculate the point estimator (s) $\bar{\chi}$ (and possibly s^2)

C.I. for case 2

Step 5: In the same manner as in Case 1:

The (1-α) CI for μ is :

$$\left[\bar{x} - z\sqrt{\frac{\sigma^2}{n}}, \bar{x} + z\sqrt{\frac{\sigma^2}{n}}\right] \quad \text{or} \quad \left[\bar{\boldsymbol{x}} - \boldsymbol{z}\sqrt{\frac{\boldsymbol{s}^2}{\boldsymbol{n}}}, \bar{\boldsymbol{x}} + \boldsymbol{z}\sqrt{\frac{\boldsymbol{s}^2}{\boldsymbol{n}}}\right]$$

Example for case 2

- Suppose X = number of sleep hours per night.
- X has an unknown distribution with unknown variance.
- But large sample (n = 150 >> 30).

The 95% CI for μ : the mean number of sleep hours per night in the population.

Step 1: choose confidence level 1-α = 0.95

Step 2: case 2, so :

$$\boxed{\frac{\bar{X}-\mu}{\sqrt{\frac{S^2}{n}}} \sim N(0,1)}$$

Step 3: critical points: -1.96 and 1.96

(the same as in Case 1, since we are still using the standard normal distribution function)

Example for case 2

Step 4 : Calculate the point estimators:

$$\bar{x} = 6.8466 \text{ and } s^2 = 1.8622$$

Step 5 : In the same manner as in Case 1:

The (1-α) CI for μ is :

$$\Rightarrow \quad \left[\bar{x} - z\sqrt{\frac{s^2}{n}}, \bar{x} + z\sqrt{\frac{s^2}{n}}\right]$$

$$\Rightarrow \quad \left[6.8466 - 1.96\sqrt{\frac{1.8622}{150}}, 6.8466 + 1.96\sqrt{\frac{1.8622}{150}}\right]$$

$$\Rightarrow \quad [6.6283, 7.0650]$$

Example for case 2

- A 95% CI for the population mean μ of the number of sleep hours per night [6.6283, 7.0650]

- **Interpretations:**
  - Based on our sample, we are 95% confident that the true mean of number of sleep hours per night lie between 6.6283 and 7.0650.

A 95% C.I. for the mean sleep hours per night

1. Variability in estimates
2. Standard error of the mean
3. Confidence intervals
4. Hypothesis testing

5 Hypothesis testing and confidence intervals

Two sided alternatives and confidence intervals

Example: The NHANES dataset - analysis of the total cholestrol level

6 Decision error (Type I and Type II error)

A 95% C.I for the mean sleep hours per night

The sample standard deveation and the standard error of the sample mean are equal to 1.3646 and 0.1114, respectivly.

Hide

```
n<-length(x.sleep)
SD.x<-sqrt(var(x.sleep))
SD.x
```

```
## [1] 1.364638
```

Standard deviation

Hide

```
SE<-SD.x/sqrt(n)
SE
```

$$SE = \sqrt{\frac{s^2}{n}}$$

```
## [1] 0.1114222
```

Standard error of the sample mean

For the sample, the error margin for a 95% confidence interval is $m = 1.96 \times SE = 1.96 \times 0.1114$ and the confidence interval is given by

$$\bar{x} \pm m = 6.8466 \pm 0.2183 = (6.628279, 7.065054).$$

Hide

```
LL<-mean(x.sleep)-1.96*SE
UL<-mean(x.sleep)+1.96*SE
c(LL,UL)
```

```
## [1] 6.628279 7.065054
```

A 95% C.I. for the mean sleep hours per night

- A 95% Confidence interval for the population mean using the R function `z.test()`.

1. Variability in estimates
2. Standard error of the mean
3. Confidence intervals
4. Hypothesis testing

5 Hypothesis testing and confidence intervals

Two sided alternatives and confidence intervals

Example: The NHANES dataset - analysis of the total cholestrol level

6 Decision error (Type I and Type II

Hide

```
z.test(x.sleep,sd=SD.x)
```

```
##
##  One Sample z-test
##
## data:  x.sleep
## z = 61.448, n = 150.00000, Std. Dev. = 1.36464, Std. Dev. of the sample
## mean = 0.11142, p-value < 2.2e-16
## alternative hypothesis: true mean is not equal to 0
## 95 percent confidence interval:
##  6.628283 7.065050
## sample estimates:
## mean of x.sleep
##        6.846667
```

Case study 3:

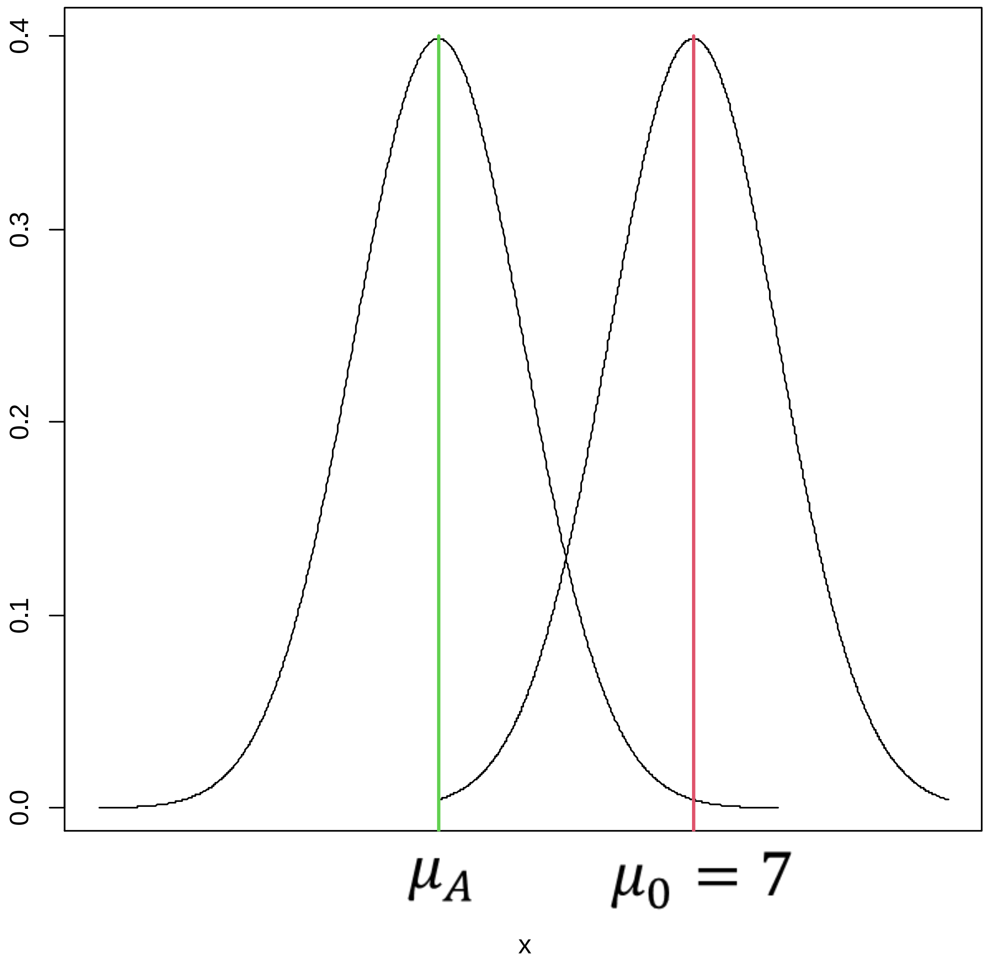
Hypotheses testing

Test of hypothesis: a one sided test

$$H_0\colon \mu = 7$$
$$H_A\colon \mu < 7$$

- We test the null hypothesis versus a one sided alternative.

- In our case, under the alternative the mean is smaller than 7 (but not specified).



The null hypothesis

Test statistic

$$t = \frac{\bar{x} - \mu_0}{\sqrt{\frac{1.3646^2}{150}}} = -1.3761$$

The population variance σ^2 is unknown but…n=150.

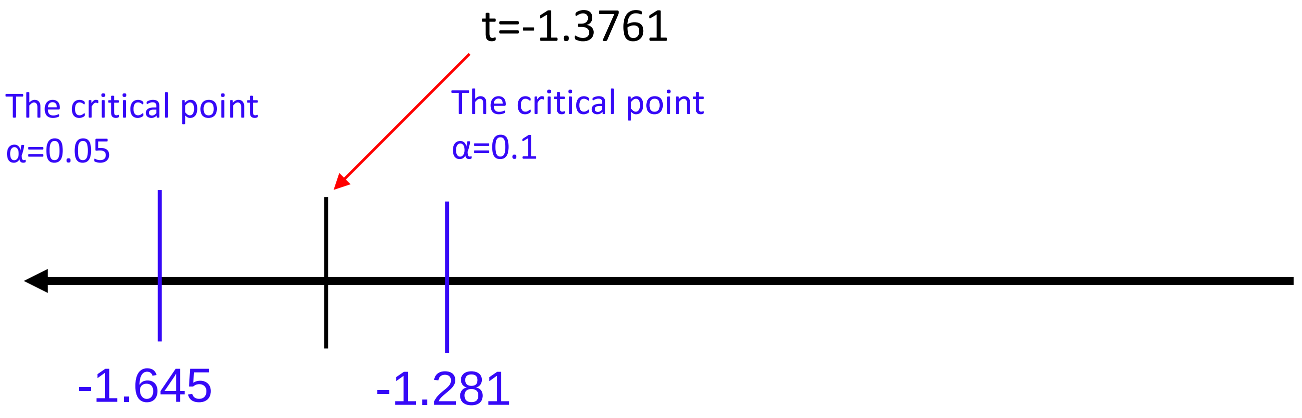
$$\Rightarrow \boxed{\frac{\bar{X} - \mu}{\sqrt{\frac{S^2}{n}}} \sim N(0,1)}$$

The critical points and the test statistic

For one sided test and $\alpha=0.05$, $Z=-1.645$.

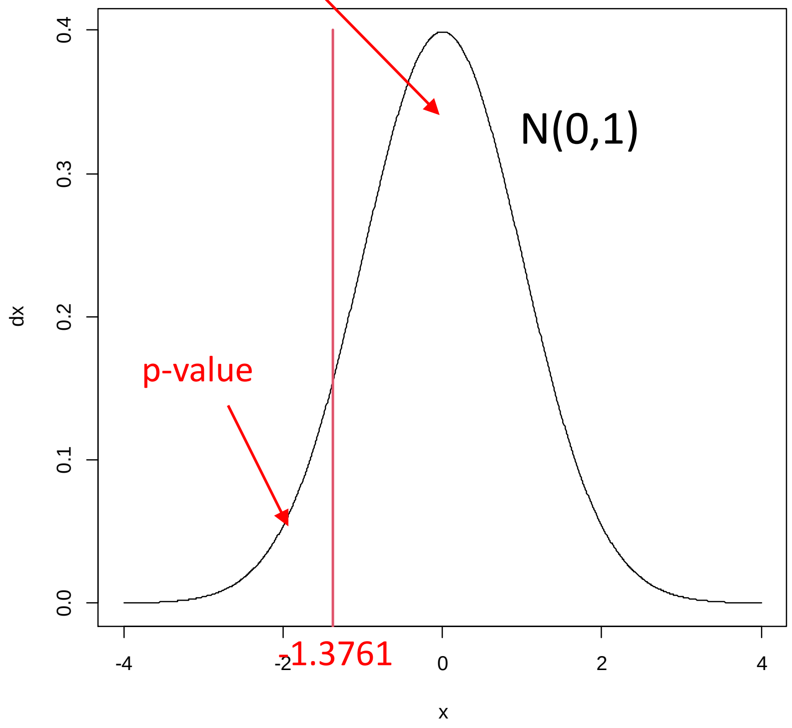
For one sided test and $\alpha=0.1$, $Z=-1.281$.

For $\alpha=0.1$ We reject H_0 : $-1.3761 < -1.281$.



p-value

$$\frac{\overline{X}-\mu}{\sqrt{\frac{S^2}{n}}} \sim N(0,1)$$



$H_0\!: \mu = 7$
$H_A\!: \mu < 7$

$P(Z < -1.3761)$=0.08439

- For α=0.05, we DO NOT reject the null hypothesis.
- For α=0.1, we reject the null hypothesis.

Hypothesis testing

1. Variability in estimates
2. Standard error of the mean
3. Confidence intervals
4. Hypothesis testing

5 Hypothesis testing and confidence intervals

Two sided alternatives and confidence intervals

Example: The NHANES dataset - analysis of the total cholestrol level

6 Decision error (Type I and Type II error)

Hypothesis testing

We wish to test the null hypothesis $\mu = 7$ aginst a one sided alternative $H_1 : \mu < 7$. This can be done using the argument `alternative = 'less'` in the function `z.test`. Note that we assume that in the population, $\sigma = 1.3646$. As can be seen in the panel below, for the sample, the mean number of sleeping hours is equal to $\bar{x} = 6.8466$ and the test statistic is equal to -1.3761. The p-=0.08439 > 0.05$. We cannot reject the null hypothesis and conclude that $\mu = 7$.

Hide

```
mean(x.sleep)
```

```
## [1] 6.846667
```

Hide

```
sqrt(var(x.sleep))
```

```
## [1] 1.364638
```

Hide

```
z.test(x.sleep,mu=7, 1.364638, alternative = 'less')
```

```
##
##  One Sample z-test
##
## data:  x.sleep
## z = -1.3761, n = 150.00000, Std. Dev. = 1.36464, Std. Dev. of the
## sample mean = 0.11142, p-value = 0.08439
## alternative hypothesis: true mean is less than 7
## 95 percent confidence interval:
##     -Inf 7.02994
## sample estimates:
## mean of x.sleep
##        6.846667
```