

## Unit 4: Introduction to Inference

Statistics 102 Teaching Team

February 26, 2020

Introduction

Populations vs samples

Confidence intervals

Hypothesis testing

Merging tests and confidence intervals

# Introduction

## GOAL OF THIS MATERIAL

Inference uses the concepts of probability and the tools of data analysis to

- draw inferences about populations via samples taken from a population
- estimate the uncertainty (or level of confidence) associated with the inference

Principles of inference will be illustrated in the setting of

- estimating the average (mean) for a characteristic of a well-defined population
- drawing conclusions about that characteristic

Material drawn from *OI Biostat*, Chapter 4, Sections 4.1 - 4.3.

## Populations vs samples

# YOUTH RISK FACTOR BEHAVIOR SURVEILLANCE SYSTEM (YRBSS)

The YRBSS is a survey conducted by the US CDC to measure health-related activity in high school aged youth.

- 2.6 million high school students participated between 1991 and 2013 in more than 1,100 separate surveys.
- Dataset `yrbss` in the `oibiostat` package contains responses from the 13,583 participants in 2013.
- 13 variables and 13,583 rows
  - each row contains information for a single participant

# YRBSS...

```
#load the data
library(oibistat)
data("yrbss")

#view a subset of the dataframe
yrbss[c(1:3, 13582, 13583), c(1:3, 6:8, 10)]
```

##	age	gender	grade	height	weight	helmet.12m	physically.active.7d
## 1	14	female	9	NA	NA	never	4
## 2	14	female	9	NA	NA	never	2
## 3	15	female	9	1.73	84.37	never	7
## 13582	17	female	12	1.60	77.11	sometimes	5
## 13583	17	female	12	1.57	52.16	did not ride	5

## YRBSS...

Variable	Description
age	Age, in years
gender	Sex, male or female
grade	Grade in high school
height	Height, in meters
weight	Weight, in kilograms
helmet.12m	Frequency that the student wore a helmet while biking in the last 12 months
physically.active.7d	Number of days physically active for 60+ minutes in the last 7 days

There are other variables in the dataset. To view their descriptions, access the dataset documentation with `?yrbss`.



## POPULATION PARAMETERS

The CDC used the responses of 13,572 students to estimate the health behaviors of the *target population*:

- 21.2 million high school aged students in the US in 2013.

The mean weight among the 21.2 million youth is an example of a **population parameter**.

- The symbol  $\mu$  is used to denote a population mean.
- For a variable such as weight, it might be written  $\mu_{\text{weight}}$ .

The mean within a sample, such as mean weight in the 13,572 students in YRBSS, is a **point estimate** of a population parameter.

- The symbol  $\bar{x}$  is used to denote a sample mean.
- Mean weight in a sample can be written as  $\bar{x}_{\text{weight}}$ .

# POPULATION PARAMETERS AND POINT ESTIMATES

Estimating the population mean weight from the sample of 13,572 participants is an example of *statistical inference*.

In nearly all studies, there is one target population and one sample.

Suppose a different random sample of the same size were taken from the same population.

- The sample would consist of different participants, thus. . .
- Numerical values of point estimates could differ from the initial sample

The variability of an estimate from sample to sample is called *sampling variability*.

Understanding the mathematical properties of sampling makes it possible to account for the effect of sampling variability when making an estimate from a sample.

# SAMPLING FROM A POPULATION

Typically, the exact values of population parameters are unknown.<sup>1</sup>

We can observe the effect of sampling variability in an artificial setting where  $\mu$  is known. This artificial setting allows us to compare  $\bar{x}$  and  $\mu$ .

- Suppose our target population consists of the 13,572 individuals in yrbss.
  - Let mean weight in yrbss be the population parameter,  $\mu_{weight}$ .
- Take a sample from yrbss (e.g.,  $n = 10$ ) and calculate,  $\bar{x}_{weight}$ , the mean weight among the sampled individuals.
  - How well does  $\bar{x}_{weight}$  estimate  $\mu_{weight}$ ?
- Take many samples to construct a *sampling distribution* of  $\bar{X}$ .

---

<sup>1</sup>Hence, the need for inference...

## TAKING ONE SAMPLE OF SIZE 10 FROM YRBSS

```
#load the dataset  
library(oibistat)  
data("yrbss")
```

```
#set parameters  
sample.size = 10  
  
#obtain random sample of row numbers  
set.seed(5011)  
sample.rows = sample(1:nrow(yrbss), sample.size)  
  
#calculate point estimates from sampled rows  
mean(yrbss$weight[sample.rows])
```

```
## [1] 63.639
```

```
sd(yrbss$weight[sample.rows])
```

```
## [1] 10.03137
```

## THE SAMPLE MEAN AS A RANDOM VARIABLE

The histograms plotted in the lab provide an approximate of the theoretical sampling distribution of  $\bar{X}$  for a specific sample size (e.g.,  $n = 10$ ).

Any sample statistic is a random variable because each sample drawn from the population is different.

- If the data have not yet been observed, the statistic is simply a function of the same random elements.

In the lab, we observed that

- The sampling distribution of  $\bar{X}$  is centered around  $\mu$ .
- The variability of  $\bar{X}$  becomes smaller with larger sample size,  $n$ .

## Confidence intervals

# CONFIDENCE INTERVALS

A *confidence interval* provides an estimate for a population parameter along with a margin of error that gives a plausible range of values for the population parameter.

A confidence interval for a population mean  $\mu$  has the general form

$$\bar{x} \pm m \rightarrow (\bar{x} - m, \bar{x} + m),$$

where  $m$  is the margin of error.

To calculate  $m$ , we use what is known about the *sampling variability* of  $\bar{X}$ .

We need a bit of notation first, and a new distribution...

## SOME USEFUL NOTATION

- $\mu$ : unknown population mean
- $\sigma$ : unknown population standard deviation
- $n$ : the number of observations in a sample drawn from the population
- $\bar{x}$ : sample mean from a sample taken from the population
- $s$ : calculated sample standard deviation from the same sample used to calculate  $\bar{x}$



# THE $t$ DISTRIBUTION

The  $t$  distribution is symmetric, bell-shaped, and centered at 0.

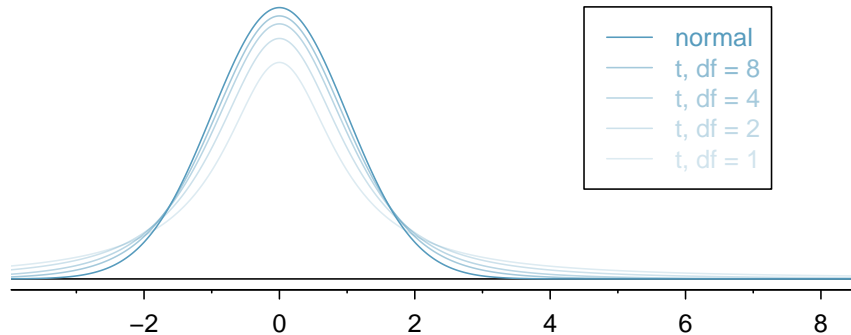
It is very close to the normal distribution, but has one additional parameter called *degrees of freedom* ( $df$ ).<sup>2</sup>

- The tails of a  $t$  distribution are thicker than those in a normal distribution. This adjusts for the variability introduced by using  $s$  as an estimate of  $\sigma$ .
- When  $df$  is large ( $df \geq 30$ ), the  $t$  and  $z$  distributions are virtually identical.
- Degrees of freedom equals  $n - 1$ .

---

<sup>2</sup>Degrees of freedom can be denoted with the symbol  $\nu$ .

## THE $t$ DISTRIBUTION...



## A 95% CONFIDENCE INTERVAL

A 95% confidence interval for a population mean  $\mu$  based on a single sample with mean  $\bar{x}$  is

$$\bar{x} \pm t^* \times \frac{s}{\sqrt{n}} \rightarrow \left( \bar{x} - t^* \frac{s}{\sqrt{n}}, \bar{x} + t^* \frac{s}{\sqrt{n}} \right),$$

where  $t^*$  is the point on a  $t$  distribution with  $n - 1$   $df$  that has 0.975 area to its left (and 0.025 area to its right).

## CALCULATING THE CRITICAL $t$ -VALUE, $t^*$

The function `qt( )` identifies the point on a  $t$  distribution with  $df$  degrees of freedom that has area  $p$  to the left.

- For  $t_{df=n-1}$ , `qt(p, df)` calculates  $t$  such that  $p = P(T \leq t)$ .
- The critical  $t$ -value for a 95% confidence interval where  $n = 10$  is 2.262.

```
qt(0.975, df = 9)
```

```
## [1] 2.262157
```

## CALCULATING A CONFIDENCE INTERVAL BY HAND

The confidence interval for mean weight, from the earlier sample of 10 individuals:

$$\bar{x} \pm t^* \frac{s}{\sqrt{n}}$$

$$63.639 \pm (2.262) \frac{10.031}{\sqrt{10}}$$

$$(56.46, 70.82) \text{ kg}$$

## LETTING R DO THE WORK

```
#set parameters  
sample.size = 10  
  
#obtain random sample of row numbers  
set.seed(5011)  
sample.rows = sample(1:nrow(yrbss), sample.size)  
  
#calculate interval estimate from sampled rows  
t.test(yrbss$weight[sample.rows])$conf.int
```

```
## [1] 56.46299 70.81501  
## attr(,"conf.level")  
## [1] 0.95
```

Full explanation of `t.test()` command coming in the labs.

# GENERAL FORM FOR A CONFIDENCE INTERVAL

A  $(1 - \alpha)(100)\%$  CI for  $\mu$  is given by

$$\bar{x} \pm t^* \times \frac{s}{\sqrt{n}},$$

where  $t^*$ , the critical  $t$ -value, is the point on a  $t$  distribution with degrees of freedom  $n - 1$  that has area  $(1 - \alpha/2)$  to the left (and area  $\alpha/2$  to the right).

```
qt(0.975, df = 9)  #critical t-value for 95% CI, n = 10
```

```
## [1] 2.262157
```

```
qt(0.975, df = 99) #critical t-value for 95% CI, n = 100
```

```
## [1] 1.984217
```

# THE STANDARD ERROR FOR $\bar{X}$

## Theoretical result

If  $\bar{X}$  could be observed in repeated sampling, its standard deviation would be approximately

$$SD_{\bar{X}} = \frac{\sigma_x}{\sqrt{n}}$$

Thus the variability of a sample mean decreases as sample size increases.

The formula for  $SD_{\bar{X}}$  characterizes that behavior more precisely.

- Typically,  $\sigma_x$  is unknown and estimated by  $s_x$ .
- The term  $\frac{s_x}{\sqrt{n}}$  is called the *standard error* of  $\bar{X}$ .



## THE 'CONFIDENCE LEVEL'

The confidence level is also called the *confidence coefficient*.

The correct interpretation:

- The method illustrated for computing a 95% confidence interval will produce an interval that (on average) contains the true population mean 95 times out of 100.
- 5 out of 100 will be incorrect, but, of course, a data analyst does not know whether a particular interval contains the population mean.

## HIDDEN ASSUMPTIONS

1. The data used to calculate the confidence interval are from a random sample taken from the target population.
2. While the population mean from the target population is not known, the target population is well-defined.

Both conditions are true in this classroom example of sampling from yrbss, but may be difficult to verify in practice.

## Hypothesis testing

# INTRODUCTION TO HYPOTHESIS TESTING

Do Americans tend to be overweight?

Body mass index (BMI) is an approximate scale used to assess weight status that adjusts for height.

- When weight is measured in kg and height in meters,

$$\text{BMI} = \frac{\text{weight}}{\text{height}^2}$$

- When weight is measured in lbs and height in inches,

$$\text{BMI} = \left( \frac{\text{weight}}{\text{height}^2} \right) (703)$$

# WHO STANDARDS FOR BMI

Category	BMI range
Underweight	$< 18.50$
Normal (healthy weight)	18.5-24.99
Overweight	$\geq 25$
Obese	$\geq 30$

Table 1: Table 4.12, *Of Biostat*

# THE NATIONAL HEALTH AND NUTRITION SURVEY (NHANES)

The National Health and Nutrition Examination Survey (NHANES) is another survey conducted by the CDC.

Purpose: to assess the health and nutritional status of adults and children in the United States

- The NHANES dataset in the NHANES package contains responses from 10,000 participants.
- The `nhanes.samp.adult` dataset in the `oibiostat` package contains responses from a random sample from participants who were age 21 or older.

We will treat `nhanes.samp.adult` as our sample and think of the adult participants in the NHANES dataset as the population.

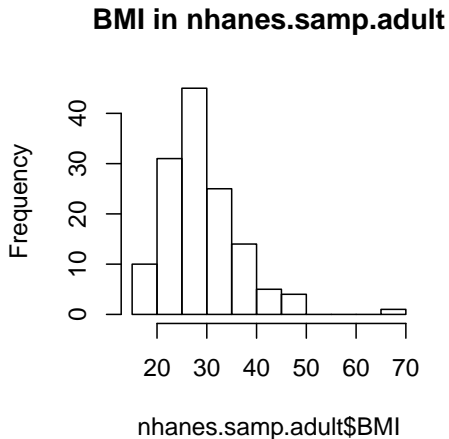
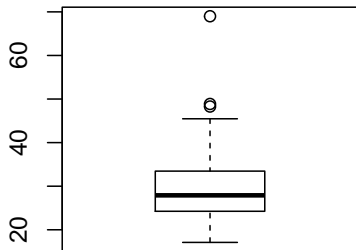
# THE NHANES SAMPLE

```
#load the dataset  
library(oibiostat)  
data("nhanes.samp.adult")  
  
#calculate summary statistics for BMI  
summary(nhanes.samp.adult$BMI)
```

##	Min.	1st Qu.	Median	Mean	3rd Qu.	Max.
##	17.10	24.25	27.90	29.10	33.46	69.00

## THE NHANES SAMPLE...

```
par(mfrow = c(1, 2))  
boxplot(nhanes.samp.adult$BMI)  
hist(nhanes.samp.adult$BMI, main = "BMI in nhanes.samp.adult")
```





## TWO APPROACHES TO THE PROBLEM

- I. Calculate a confidence interval for the population mean BMI.
2. Use the formal logic of hypothesis testing.

# I. CALCULATING A 95% CONFIDENCE INTERVAL

```
t.test(nhanes.samp.adult$BMI, conf.level=0.95)$conf.int
```

```
## [1] 27.81388 30.38524  
## attr(,"conf.level")  
## [1] 0.95
```

Confidence interval suggests that population average BMI is well outside the range defined as normal, 18.5 - 24.99.

## II. FORMAL APPROACH TO HYPOTHESIS TESTING

Observations come from either of two competing population distributions:

- The *null* distribution: a usual distribution that has been true in the past
- The *alternative* distribution: new distribution induced by an intervention or a changing condition

We conclude that observations come from the null distribution, unless...

- the value of an observed statistic (e.g.,  $\bar{x}$ ) is so extreme that it would be unlikely to occur under the null distribution

## FORMAL APPROACH (*Ol Biostat* SECTION 4.3.1)

Steps in hypothesis testing. Details coming in subsequent slides.

1. Formulate null and alternative hypotheses
2. Specify a significance level,  $\alpha$
3. Calculate a test statistic
4. Calculate a  $p$ -value
5. State a conclusion in the context of the original problem

# 1. NULL AND ALTERNATIVE HYPOTHESES

The *null hypothesis* ( $H_0$ ) posits a distribution for the population that reflects no change from the past.

- $H_0$  can be thought of as representing the status quo.

The *alternative hypothesis* ( $H_A$ ) claims a 'real' difference between the distribution of the observed data and the null-hypothesized distribution.<sup>3</sup>

- $H_A$  is an alternative claim under consideration and is often represented by a range of possible parameter values.

---

<sup>3</sup>That is, the discrepancy between  $\bar{x}$  and  $\mu$  is large enough that it seems unlikely to occur from sampling variability.

## 1. NULL AND ALTERNATIVE...

Several possible choices for  $H_0$  and  $H_A$  for our BMI question. Let's choose

- $H_0 : \mu_{\text{bmi}} = 21.7 = \mu_0$ , the midpoint of the normal range, and
- $H_A : \mu_{\text{bmi}} > 21.7$

This form of  $H_A$  is called a one-sided alternative.

- $H_A : \mu_{\text{bmi}} \neq 21.7$  would be a two-sided alternative.

The choice of one- or two-sided alternative is context-dependent.

## 2. SPECIFYING A SIGNIFICANCE LEVEL $\alpha$

The significance level  $\alpha$  can be thought of as the value that quantifies how rare or unlikely an event must be in order to represent sufficient evidence against  $H_0$ .

Typically,  $\alpha$  is chosen to be 0.05, 0.01, or some other small value.

In the context of decision errors,  $\alpha$  is the probability of making a Type I error.

- Type I error refers to incorrectly rejecting the null hypothesis.
- More on this definition in Lab 4. . .

### 3. CALCULATE A TEST STATISTIC

The test statistic measures the discrepancy between the observed data and what would be expected if the null hypothesis were true.

- Specifically, how many standard deviations is the observed sample value from the population value under the null hypothesis?

When testing hypotheses about a mean, the test statistic is

$$T = \frac{\bar{X} - \mu_0}{s/\sqrt{n}},$$

where the test statistic  $T$  follows a  $t$  distribution with  $n - 1$  degrees of freedom.



## 4. CALCULATE A $p$ -VALUE

What is the probability that we would observe a result as or more extreme than the observed sample value, if the null hypothesis is true?

- The  $p$ -value represents this probability.<sup>4</sup>

Calculate the  $p$ -value associated with the test statistic and compare it to the significance level  $\alpha$ .

- A result is considered unusual if its  $p$ -value is less than  $\alpha$ .

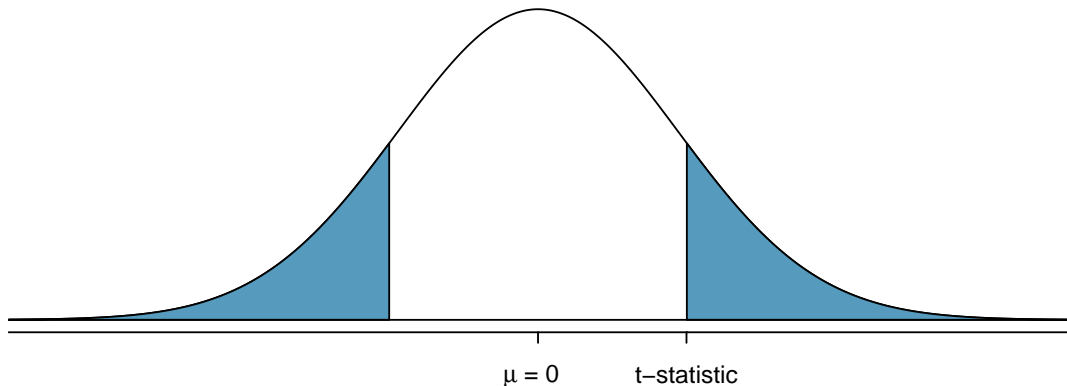
---

<sup>4</sup>The  $p$ -value is *not* the probability that  $H_0$  is true (nor is it the probability that  $H_A$  is false).

#### 4. THE $p$ -VALUE...

For a two-sided alternative,  $\mu \neq \mu_0$ , the  $p$ -value is the total area from both tails of the  $t$  distribution that are beyond the absolute value of the observed statistic.

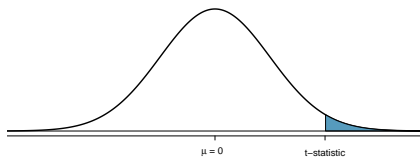
- $p = 2P(T \geq |t|) = P(T \leq -|t|) + P(T \geq |t|)$



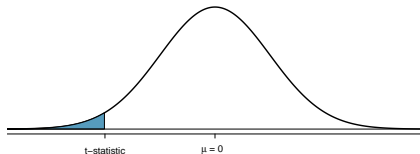
## 4. THE $p$ -VALUE...

For a one-sided alternative, the  $p$ -value is the area in the tail of the  $t$  distribution that matches the direction of the alternative.

For  $H_A : \mu > \mu_0$ :



For  $H_A : \mu < \mu_0$ :



## 4. THE $p$ -VALUE...

The smaller the  $p$ -value, the stronger the evidence against the null hypothesis.

- If the  $p$ -value is as small or smaller than  $\alpha$ , we *reject* the null hypothesis. The result is statistically significant at level  $\alpha$ .
- If the  $p$ -value is larger than  $\alpha$ , we *fail to reject* the null hypothesis. The result is not statistically significant at level  $\alpha$ . In other words, the evidence does not contradict the null hypothesis.

A subtle but important point: not rejecting  $H_0$  is not the same as proving that  $H_0$  is true. We simply do not have sufficient evidence that  $H_0$  is not true!

## 5. DRAW A CONCLUSION

State the conclusion in the context of the original problem, using the language and units of that problem.

This is the part most often omitted by students, but it is the most important!

## BMI IN NHANES...

```
#use r as a calculator
x.bar = mean(nhanes.samp.adult$BMI)
mu.0 = 21.7
s = sd(nhanes.samp.adult$BMI)
n = length(nhanes.samp.adult$BMI)

t = (x.bar - mu.0)/(s/sqrt(n))
t
```

```
## [1] 11.38311
```

```
#calculate the p-value
pt(t, df = n - 1, lower.tail = FALSE)
```

```
## [1] 1.006759e-21
```

# BMI IN NHANES...

```
#let r do the work...  
t.test(nhanes.samp.adult$BMI, mu = 21.7, alternative = "greater")
```

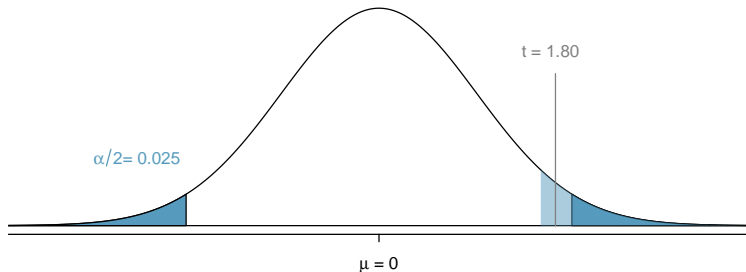
```
##  
## One Sample t-test  
##  
## data:  nhanes.samp.adult$BMI  
## t = 11.383, df = 134, p-value < 2.2e-16  
## alternative hypothesis: true mean is greater than 21.7  
## 95 percent confidence interval:  
##  28.02288      Inf  
## sample estimates:  
## mean of x  
##  29.09956
```

## TWO-SIDED VS. ONE-SIDED HYPOTHESIS TESTS

Two-sided tests are more 'conservative', as the  $p$ -value for a two-sided test is twice that of a one-sided test.

- More about choosing between one-sided and two-sided tests in Lab 4...

In practice, two-sided tests are usually used, and two-sided tests are expected by most journals and regulatory authorities.





## Merging tests and confidence intervals

## TWO-SIDED TESTS AND CONFIDENCE INTERVALS

The relationship between a hypothesis test and the corresponding confidence interval is defined by  $\alpha$ .

- Hypothesis test: is  $\bar{x}$  far enough away from  $\mu_0$  to be considered extreme?
- Confidence interval: is  $\mu_0$  close enough to  $\bar{x}$  to be plausible?

In both cases, “far enough” and “close enough” are defined by  $\alpha$ .

- More on this idea in Lab 4. . .

## IN PRACTICE...

If a 95% confidence interval for a population mean does not contain a hypothesized value  $\mu_0$ , then:

- the data contradict the the null hypothesis  $H_0 : \mu = \mu_0$  at significance level  $\alpha = 0.05$
- the implied two-sided alternative hypothesis is  $H_A : \mu \neq \mu_0$