Metrics

INFO 370

Learning Objectives

Review **probability + statistics** basics from the reading

Discuss the **Central Limit Theorem**

Discuss the purpose and challenge of **developing metrics**

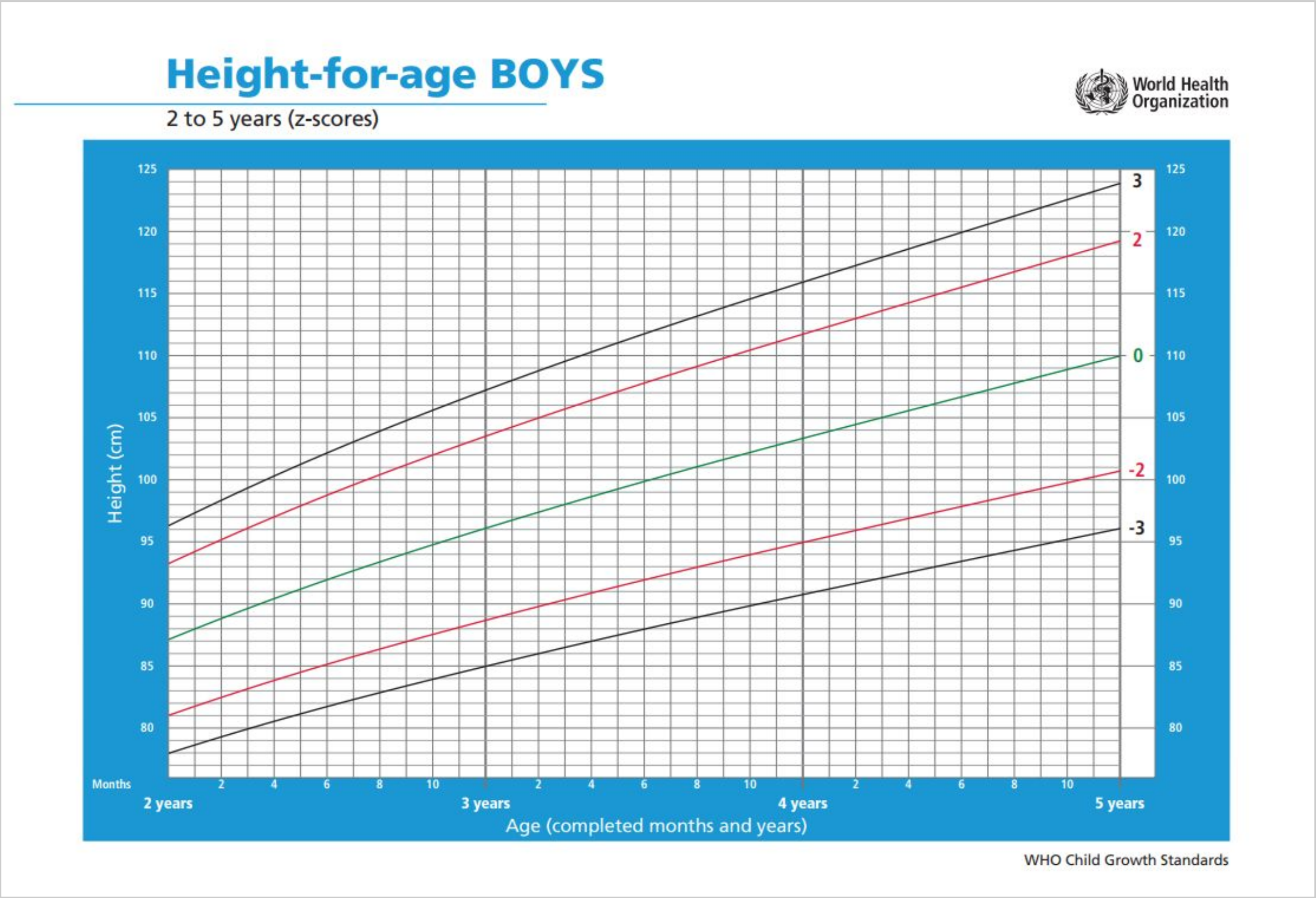
Practice developing metrics for Informatics courses

Consider how to develop **segregation metrics** for assignment 2

Reading Review

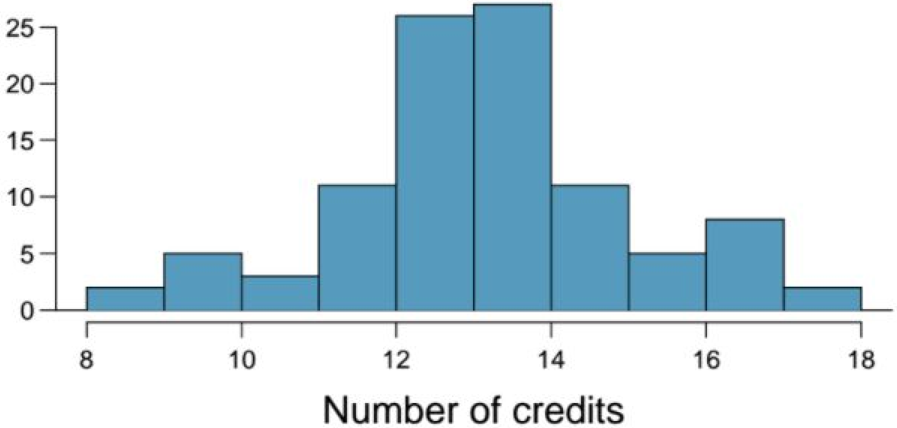
3.3 GRE scores, Part I. Sophia who took the Graduate Record Examination (GRE) scored 160 on the Verbal Reasoning section and 157 on the Quantitative Reasoning section. The mean score for Verbal Reasoning section for all test takers was 151 with a standard deviation of 7, and the mean score for the Quantitative Reasoning was 153 with a standard deviation of 7.67. Suppose that both distributions are nearly normal.

- Why does it matter that both distributions are **<u>nearly normal</u>**?
- Relative to the **population**, did she do better on the verbal or math exam?
- What **percent of the test takers** did better than her on the Verbal Reasoning section?



WHO Cares (source)

4.3 College credits. A college counselor is interested in estimating how many credits a student typically enrolls in each semester. The counselor decides to randomly sample 100 students by using the registrar's database of students. The histogram below shows the distribution of the number of credits taken by these students. Sample statistics for this distribution are also provided.



Min	8
Q1	13
Median	14
Mean	13.65
SD	1.91
Q3	15
Max	18

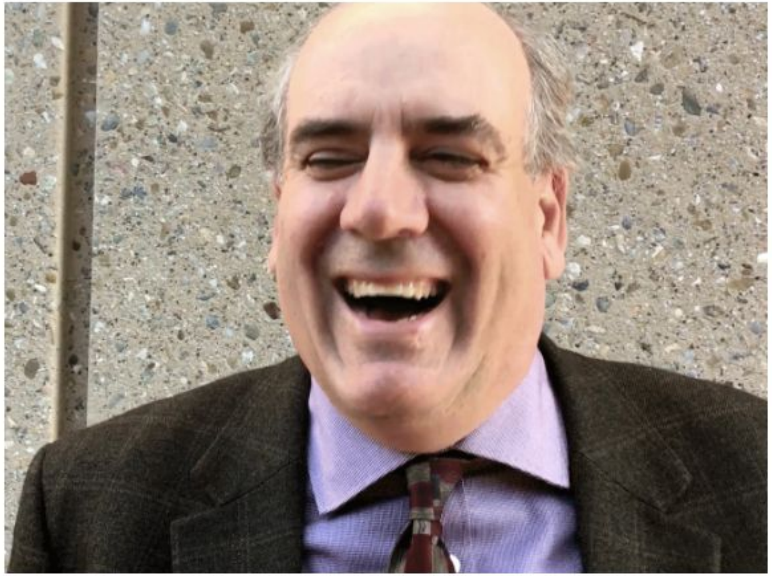
Based on this data, ~⅔ (68%) of students take between X and Y credits. Solve for X and Y.

NOV. 24, 2015 AT 12:12 PM

Not Even Scientists Can Easily Explain P-values

By Christie Aschwanden

Filed under Scientific Method



P Values (link)

"the probability of getting results at least as extreme as the ones you observed, given that the null hypothesis is correct"

Assuming the null hypothesis is true, what is the probability of observing my data?

The probability of observing the data (7.42 hrs/night) is only 1% (p-value = .011) if the null hypothesis it true.
We should **reject the null hypothesis** in favor of the alternative.

$$Z = \frac{X - null}{SE}$$

$$Z = \frac{7.42 - 7}{(1.75)/\sqrt{110}}$$

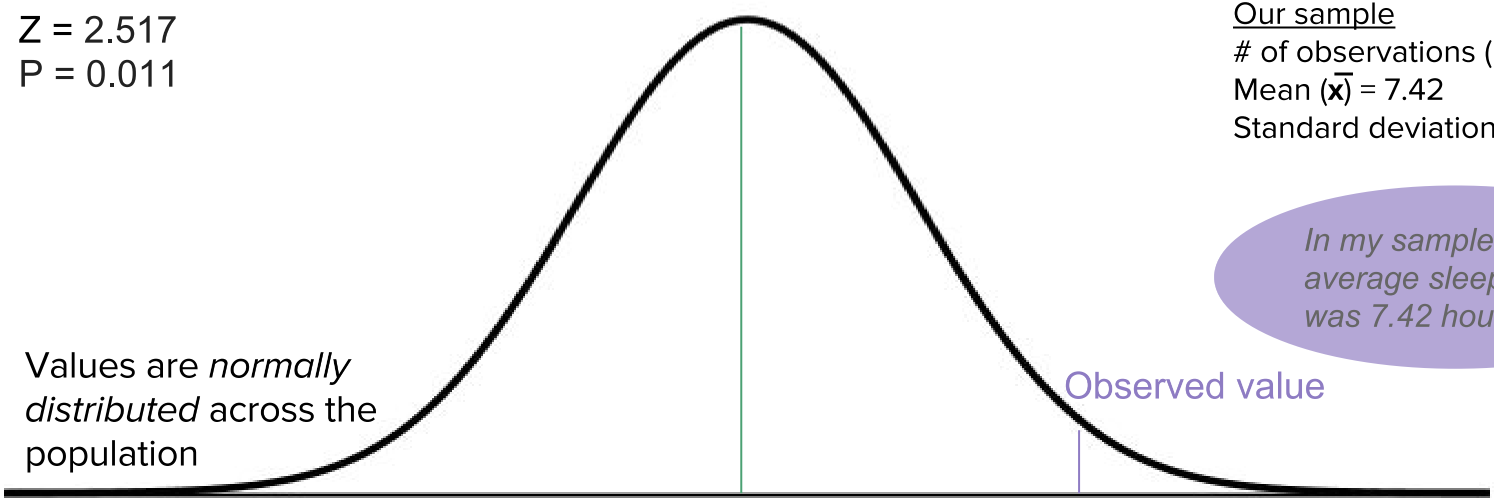
Z = 2.517
P = 0.011

On average, people sleep 7 hours / night

Null Hypothesis
(assume this is true)

Our sample
\# of observations (**n**) = 110
Mean (**$\bar{x}$**) = 7.42
Standard deviation (**s**)= 1.75

In my sample, the average sleep/night was 7.42 hours.



Values are *normally distributed* across the population

P Values

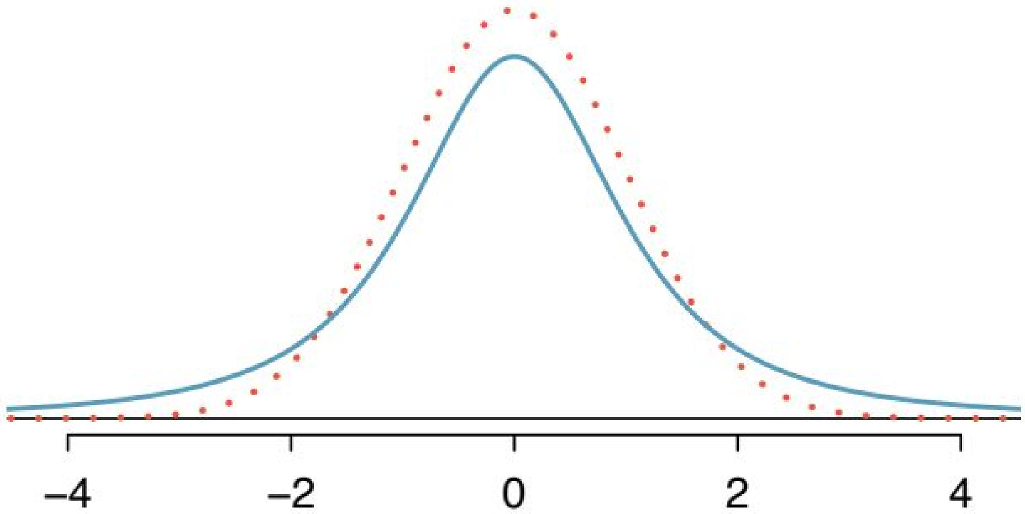


Figure 5.1: Comparison of a t-distribution (solid line) and a normal distribution (dotted line).

T-distribution v.s. Normal Distribution

5.7 Sleep habits of New Yorkers. New York is known as "the city that never sleeps". A random sample of 25 New Yorkers were asked how much sleep they get per night. Statistical summaries of these data are shown below. Do these data provide strong evidence that New Yorkers sleep less than 8 hours a night on average?

n	$\bar{x}$	s	min	max
25	7.73	0.77	6.17	9.78

Using the values above, compute the t-score for the data, and accept or reject the null hypothesis.

Basic Probability Applications

Design an airplane layout to "comfortably seat" **99%** of the population.

Understand **relative performance** of students across courses/exams/instructors.

Confirm or deny statements about **averages** (*the average person…*)

Understand the **probability** of something based on the **distribution** of the event

Central Limit Theorem

Consider Multiple Samples

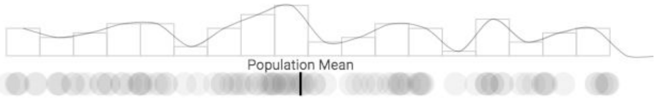
So let's say, hypothetically, that we could take multiple samples from our population. While this is something that may happen (especially with political polls), we'll use this more as an explanatory tool (there are other considerations that come into play when you actually sample repeated times that we won't consider). For each sample, let's keep track of how the **sample mean** compares to our **population mean** each time we draw a sample. Once we repeat the process multiple times, we will have a **distribution of sample means**, often referred to as the **sampling distribution of the mean** or (more simply) the **sampling distribution**. We'll display the distribution of sample means in a density plot below our population's density plot:

TAKE A SAMPLE

As we repeat this process, you may notice something interesting happen: the difference between our true population mean and the mean of our sample means begins to shrink. This makes sense, as it feels similar to simply drawing a larger sample (with replacement) from our population. Unfortunately, this doesn't solve the problem of limited resources for understanding a population's position on a particular issue. In order to understand the quality of a single estimate, we first need to understand what the **distribution of sample means** looks like.

Thinking about Distributions

The current visualization of our population's opinions as a density plot is a bit hard to decipher, so let's add a bit more information with a histogram of the opinions. This will better allow us to see the shape of the distribution:



Central Limit Theorem (link)

Metrics

What's a metric?

Daily Active Users

Metrics

Reduction of a complex idea/data structure to a **single measure**

Allows the **comparison** of multiple elements

How can they be developed?

- Devise an equation that captures desired components
- Weight different aspects appropriately
- Intuitive v.s. data-derived

Some desired attributes:

- Insensitive to outliers(?)
- Normalized to a certain scale(?)

The four separate calculations can be expressed in the following equations:

$$a = \left(\frac{\mathrm{COMP}}{\mathrm{ATT}} - .3\right) \times 5$$

$$b = \left(\frac{\mathrm{YDS}}{\mathrm{ATT}} - 3\right) \times .25$$

$$c = \left(\frac{\mathrm{TD}}{\mathrm{ATT}}\right) \times 20$$

$$d = 2.375 - \left(\frac{\mathrm{INT}}{\mathrm{ATT}} \times 25\right)$$

where

> **ATT** = Number of passing attempts
>
> **COMP** = Number of completions
>
> **YDS** = Passing yards
>
> **TD** = Touchdown passes
>
> **INT** = Interceptions

If the result of any calculation is greater than 2.375, it is set to 2.375. If the result is a negative number, it is set to zero.

Then, the above calculations are used to complete the passer rating:

$$\text{Passer Rating} = \left(\frac{(a+b+c+d)}{6}\right) \times 100$$

NFL Passer Rating (source)

Course Metrics

How *good* is the ______ Informatics course?

An aside: Data and Operation Abstractions

Nominal (categorical): Labels, such as Fruits: apples, bananas, oranges, etc.

- Operations: $=$, $\neq$

Ordinal (categorical): Ordered, such as Grade of meat: Grade A, Grade AA, etc.

- Operations: $=$, $\neq$, $\leq$, $\geq$, $<$, $>$

Interval (continuous): Arbitrary 0, such as Dates: 05/15/2012, 04/17/2015, etc.

- Operations: $=$, $\neq$ $\leq$, $\geq$, $<$, $>$, $\pm$

Ratio (continuous): Zero fixed, such as Length: 1 in, 1.5 in, 2 in, etc.

- Operations: $=$, $\neq$ $\leq$, $\geq$, $<$, $>$, $\pm$, $\div$

wiki

Activity

Collaboratively develop a short survey regarding informatics courses

- Continuous questions only (scale of 1 - 5)

In groups, develop proposed metrics that reduce the results to a **single value**

Everyone takes the survey

Compute resulting metrics and see scores (together)

Segregation

How can we measure how segregated a city is?

	GEO.display.label	pop	pop.white	pop.not.white	pct.white	pct.not.white
1	Census Tract 101, Baltimore city, Maryland	1378	1276	102	0.9259796...	0.07402032
2	Census Tract 102, Baltimore city, Maryland	1262	1082	180	0.8573692...	0.14263074
3	Census Tract 103, Baltimore city, Maryland	965	847	118	0.8777202...	0.12227979
4	Census Tract 104, Baltimore city, Maryland	1196	1146	50	0.9581939...	0.04180602
5	Census Tract 105, Baltimore city, Maryland	951	759	192	0.7981072...	0.20189274
6	Census Tract 201, Baltimore city, Maryland	871	697	174	0.8002296...	0.19977038
7	Census Tract 202, Baltimore city, Maryland	791	617	174	0.7800252...	0.21997472
8	Census Tract 203, Baltimore city, Maryland	1646	1427	219	0.8669501...	0.13304982
9	Census Tract 301, Baltimore city, Maryland	879	269	610	0.3060295...	0.69397042
10	Census Tract 302, Baltimore city, Maryland	887	533	354	0.6009019...	0.39909808
11	Census Tract 401, Baltimore city, Maryland	2046	1363	683	0.6661779...	0.33382209
12	Census Tract 402, Baltimore city, Maryland	574	262	312	0.4564459...	0.54355401

Census tract data snapshot

Visualizing the Racial Divide (link)

measures-of-segregation

Upcoming...

a2-data-wrangling due ***next Tuesday before class***

Notebook 3 due **Friday night** (notably shorter, more time in class/lab)

This week:

- Metrics
- Basic statistical tests (t-tests)