

## Course Introduction and Terminology

**Pablo E. Gutiérrez-Fonseca, PhD**

University of Vermont

Fall 2024

1

## Outline

- Course introduction
- Terminology

2

## Statistics: What it is?

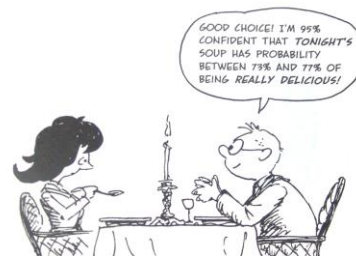
- **Statistics** describes “a set of tools and techniques that are used for describing, organizing, and interpreting data.”
- Without the use of statistics, we muddle through life making choices based on incomplete information.



3

## What does Statistics give us?

- How does the process of statistical analysis allow us to learn from the world around us?
- Transforming uncertainty into Wisdom.
  - What makes statistics unique is its ability to quantify uncertainty.



4

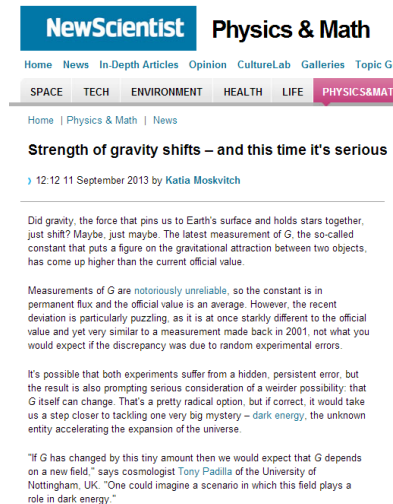
## Transforming Uncertainty into wisdom

- **Nothing can be PROVEN**
  - No matter how strongly a theory is supported by empirical evidence, it is always theoretically conceivable that one day, some data will come in that will force the scientists to modify or even eliminate the theory.

5

## Transforming Uncertainty into wisdom

- Nothing can be PROVEN with 100% certainty.



**NewScientist** Physics & Math

Home News In-Depth Articles Opinion CultureLab Galleries Topic G

SPACE TECH ENVIRONMENT HEALTH LIFE PHYSICS & MATH

Home | Physics & Math | News

### Strength of gravity shifts – and this time it's serious

12:12 11 September 2013 by [Katia Moskvitch](#)

Did gravity, the force that pins us to Earth's surface and holds stars together, just shift? Maybe, just maybe. The latest measurement of  $G$ , the so-called constant that puts a figure on the gravitational attraction between two objects, has come up higher than the current official value.

Measurements of  $G$  are notoriously unreliable, so the constant is in permanent flux and the official value is an average. However, the recent deviation is particularly puzzling, as it is at once starkly different to the official value and yet very similar to a measurement made back in 2001, not what you would expect if the discrepancy was due to random experimental errors.

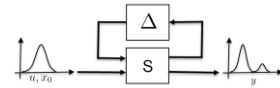
It's possible that both experiments suffer from a hidden, persistent error, but the result is also prompting serious consideration of a weirder possibility: that  $G$  itself can change. That's a pretty radical option, but if correct, it would take us a step closer to tackling one very big mystery – dark energy, the unknown entity accelerating the expansion of the universe.

"If  $G$  has changed by this tiny amount then we would expect that  $G$  depends on a new field," says cosmologist [Tony Padilla](#) of the University of Nottingham, UK. "One could imagine a scenario in which this field plays a role in dark energy."

6

# Transforming Uncertainty into wisdom

- **Our Goal:**
  - Quantify the level of uncertainty and use this to make Informed decisions.



**Laboratory for Uncertainty Quantification**

Aerospace Engineering, Texas A&M University



7

## Outline

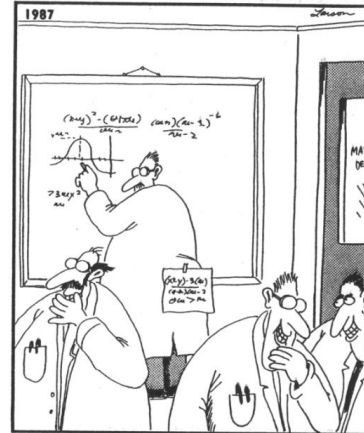
- Course Introduction
- Terminology

8

## Thinking like a statistician

- **Becoming Bilingual:**

- Some key terms you should know and understand so that we can communicate during the rest of the semester.



9

## Thinking like a statistician



- **Population:** A population is the collection or aggregate of all elements or items of interest in a particular study about which we wish to make an inference.
- **Sample:** a subset of measurements taken from a population ( $n$ ).
  - **Random Sample:** sample drawn in such a way that all observations have an equal chance of being selected
  - *BUT.....be careful.....your sample should represent the complete range of characteristics you expect in your population*
  - *Data is **created**, not found!*
- **Observation:** a recording of some information on a sample unit ( $X$ ).
  - *Observations are sometimes called samples, measurements, values or just plain "n"*

10

## Thinking like a statistician

- **Independence:** selection of any one sample unit (observations) does not affect the chances of any other unit being selected.
  - *Variables can also be independent (co-vary in an unrelated pattern).*
- **Replicate:** If measurements are in some way related, they must be treated as replicates and “nested” within the analysis such that sample size reflects the number of independent observations.
  - *One simple way to work with replicates is to simply average them for each unit of observation.*

11

## Thinking like a statistician

- **Each \*tank\* is a replicate, but within each tank each trout is not independent.**
  - A replicate is "the smallest experimental unit to which a treatment is independently applied."
  - Most models for statistical inference require true replication. *True* replication permits the estimation of *variability within a treatment*.



12

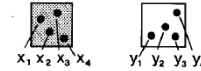
## Thinking like a statistician

- The simplest and most common type of pseudo-replication occurs when there is only one replicate per treatment.
- **Sacrificial pseudo-replication** occurs when data from true replicates are combined before analysis, removing their independence and potentially leading to inaccurate results.
- **Temporal pseudo-replication** is also common in ecological experiments in which a time series of data are accumulated.

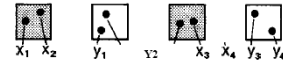
204

STUARTH.

### A SIMPLE PSEUDOREPLICATION



### B. SACRIFICIAL PSEUDOREPLICATION



### C. TEMPORAL PSEUDOREPLICATION

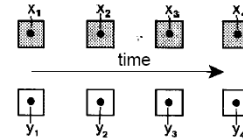
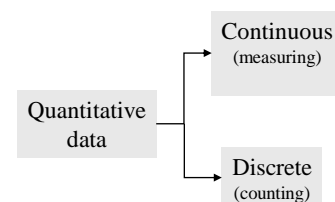


FIG. 5. Schematic representation of the three most common types of pseudoreplication. Shaded and unshaded boxes represent experimental units receiving different treatments. Each dot represents a sample or measurement. Pseudoreplication is a consequence, in each example, of statistically testing for a treatment effect by means of procedures (e.g.,  $t$  test,  $U$  test) which assume, implicitly, that the four data for each treatment have come from four independent experimental units (=treatment replicates).

13

## Thinking like a statistician

- **Variable:** a quantity counted or measured the characteristic that is being observed.
- **Quantitative Variables:** a measurable “amount”.
  - **Continuous variable:** may assume any imaginable value within a certain range.
  - ... can (theoretically) have an infinite number of values.
    - Weights, Heights...
  - **Discrete Variables:** countable as integers (whole numbers).
  - No values between two adjacent values are permissible.
    - Number of bicycles sold in a day.



14

## Thinking like a statistician

- **Variable:** a quantity counted or measured the characteristic that is being observed.
- **Qualitative Variables:** descriptive characteristic assignable to a category
  - **Nominal Variables:** measurements fall into a particular class or category with no order implied.
    - sex (male or female)...
  - **Ordinal Variables:** a ranking scale where order between categories is implied.
  - **Interval (ratio) Variables:** use a quantitative measurement to assign a specific qualitative category (these are still ordinal).

15

## Thinking like a statistician

- **More ways to describe our variables:**
- **Independent:** measurable variables whose value is not dependent upon other measured variables: aka *input* variables
- **Dependent:** variable you want to predict. The outcome of interest: aka the *response* variable
  - We hypothesize that its value is dependent on the value of the other measured variables.

16

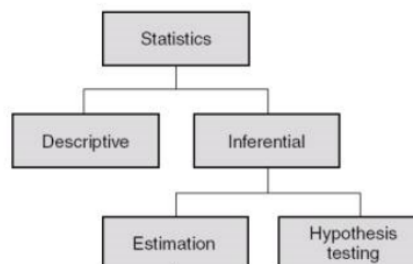


## Thinking like a statistician

- You will find that there are other terms commonly used to refer to variables:
  - Independent Variables can be called:
    - *Factor*
    - *Treatment*
    - *Level*
    - *X*
    - *Input*
  - Dependent Variables can be called:
    - *Response*
    - *Y*
    - *Outcome*

17

## Different Types of Statistics

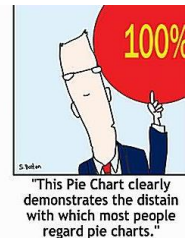


18

## Descriptive statistics

Descriptive

- **Descriptive statistics** are statistics that quantitatively describe or summarize the data.
  - The aim of descriptive statistics is to summarize a sample, *rather* than use the data to learn about the population.
  - Usually, information is displayed visually in tables or figures.



19

## Inferential statistics

- Mathematical methods that use probability theory to infer the properties of a population from the analysis of a sample drawn from it.
- Inferential statistics aim to make generalizations and predictions from the sample to the population.



- Makes assumptions about the entire population based on what you see in your sample.
- Example: quantifying the typical stocking levels for forests across Vermont based on FIA inventory data.
- Allows us to test hypotheses.

20

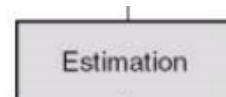
## Why inferential statistics?

- Large populations make investigating each member impractical and expensive.
- Easier and cheaper to take a sample and make estimates about the population from the sample.



21

## Statistics



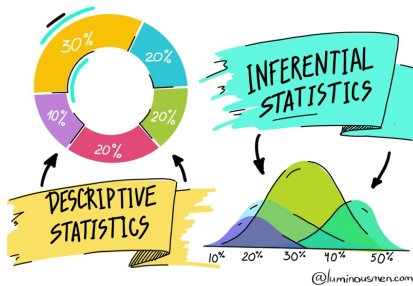
### • Statistical Modeling

- Used to predict future outcomes based on observed patterns.
- Modeling a dependent variable based on one or more independent variables.
  - Example: Predicting future trends in forest stocking based on historical inventories
  - Example: Climate effects based on CO2
    - Can be theoretical (based on established relationships) or Empirical (based on measurements)

22

## One big happy family

- Descriptive and Inferential Statistics: close cousins and both required statistical analyses.



### Descriptive Statistics:

We measured the length of each fish in Shelburne Pond and found that the mean length was 15 inches, the standard deviation was 2 inches, the minimum length was 10 inches, and the maximum length was 20 inches.

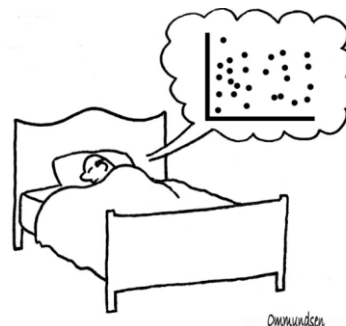
### Inferential Statistics:

We measured the length of each fish in Shelburne Pond and then announced that we are 95% confident that the average length of all fish in the lake is between 14.5 and 15.5 inches.

23

## How to know when you are fluent

- Think like a statistician



24

## Terminology in Action



Moose herds across the Northeast are increasingly under stress from climate change. Specifically, wildlife biologists are concerned that warmer falls and early springs are increasing winter tick numbers that can have detrimental effects on moose. They have tranquilized and sampled 59 moose from across the state of VT to see if the density (mean per sq. inch from 5 patches) of winter ticks is impacting moose vitality (measured as weight).

- **RESEARCH QUESTION**
- **POPULATION**
- **SAMPLE**
- **OBSERVATION**
- **NAME AND DESCRIBE THE VARIABLES**
- **ANALYSIS TYPE**

25

## Terminology in Action

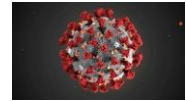


Moose herds across the Northeast are increasingly under stress from climate change. Specifically, wildlife biologists are concerned that warmer falls and early springs are increasing winter tick numbers that can have detrimental effects on moose. They have tranquilized and sampled 59 moose from across the state of VT to see if the density (mean per sq. inch from 5 patches) of winter ticks is impacting moose vitality (measured as weight).

- **RESEARCH QUESTION**  
Is winter tick density related to moose weight?
- **POPULATION**  
All moose in VT
- **SAMPLE**  
The 59 moose they were able to catch
- **OBSERVATION**  
A moose (the 5 patches on each moose are replicates)
- **NAME AND DESCRIBE THE VARIABLES**  
Dependent = Moose Weight (continuous)  
Independent = Tick density (continuous)
- **ANALYSIS TYPE**  
Inferential - Correlation

26

## Terminology in Action



- We examine the relationship between cryptocurrencies and COVID-19 cases/deaths. This will help explore whether cryptocurrencies can serve as a hedge against COVID-19. The wavelet coherence analysis indicates that there is initially a negative relationship between Bitcoin and the number of reported cases and deaths; however, the relationship becomes positive during the later period. The findings for Ethereum and Ripple are also similar but with weaker interactions. This supports the hedging role of cryptocurrencies against the uncertainty raised by COVID-19.

- **RESEARCH QUESTION**  
Is there a relationship between cryptocurrencies and deaths due to covid19?
- **POPULATION**  
Worldwide COVID-19 cases and deaths
- **SAMPLE**  
Worldwide COVID-19 cases and deaths
- **OBSERVATION**  
Cases and deaths
- **NAME AND DESCRIBE THE VARIABLES**  
Dependent: Covid cases and death  
Independent: US\$ prices of Bitcoin (BTC), Ethereum (ETH), and Ripple (XRP)
- **ANALYSIS TYPE**  
Inferential – Correlation

27

## Terminology in Action

**Abstract** River fragmentation and alterations in flow and thermal regimes are the main stressors affecting migrating fish, which could be aggravated by climate change and increasing water demand. To assess these impacts and define mitigation measures, it is vital to understand fish movement patterns and the environmental variables affecting them. This study presents a long-term (1995–2019) analysis of upstream migration patterns of anadromous and potamodromous brown trout in the lower River Bidasoa (Spain). For this, captures in a monitoring station were analyzed using Survival Analysis and Random Forest techniques. Results showed that most upstream movements of potamodromous trout

occurred in October–December, whereas in June–July for anadromous trout, although with differences regarding sex and size. Both, fish numbers and dates varied over time and were related to the environmental conditions, with different influence on each ecotype. The information provided from comparative studies can be used as a basis to develop adaptive management strategies to ensure freshwater species conservation. Moreover, studies in the southern distribution range can be crucial under climate warming scenarios, where species are expected to shift coldwards.

- **RESEARCH QUESTION**  
Is there a significant trend in trout migration?
- **POPULATION**  
All Trout in the lower River Bidasoa
- **SAMPLE**  
Trout in the lower River Bidasoa (Spain)
- **OBSERVATION**  
Abundance each year
- **NAME AND DESCRIBE THE VARIABLES**  
Dependent: Number of fish  
Independent: Year (ordinal), Stream Discharge
- **ANALYSIS TYPE**  
Inferential – Random forest regression

28

## Terminology in Action



*The Vermont DEC is monitoring twelve “sentinel” streams in Vermont. These reference streams are widely variable in terms of development and agricultural density. Scientists are using species richness of sensitive macroinvertebrates at three drift net locations along each stream as an indicator of overall water quality to see if the density of agriculture and development (categorized as minimal, mixed or moderate) within watersheds significantly impairs water quality.*

- **RESEARCH QUESTION**  
Does the density of agriculture and development impact macroinvertebrate species richness?
- **POPULATION**  
All streams in VT
- **SAMPLE**  
12 Sentinel streams in VT
- **OBSERVATION**  
A stream (drift net replicates at each stream)
- **NAME AND DESCRIBE THE VARIABLES**  
Dependent: species richness (continuous)  
Independent: Ag density (ordinal), Dev density (ordinal)
- **ANALYSIS TYPE**  
Inferential: Factorial ANOVA