Lecture 1
Introduction to Statistical Methods

Answering Questions with data

- What proportion of people in the U.S are biologically female?

  Survey Americans and ask them to report their sex

- Does a low-carb diet result in significant weight loss?

  - Design an experiment to evaluate the effectiveness of the low carb diet on weight loss
  - Record information such as starting and ending weight, calories consumed per day, …

- Are people more likely to stop at Starbucks if they've recently seen a Starbucks TV add?

  - Conduct a marketing survey to record the number of people who have gone to starbucks since add aired
  - Compare between those who saw the add and those who did not

Where do we find Statistics? – Everywhere!

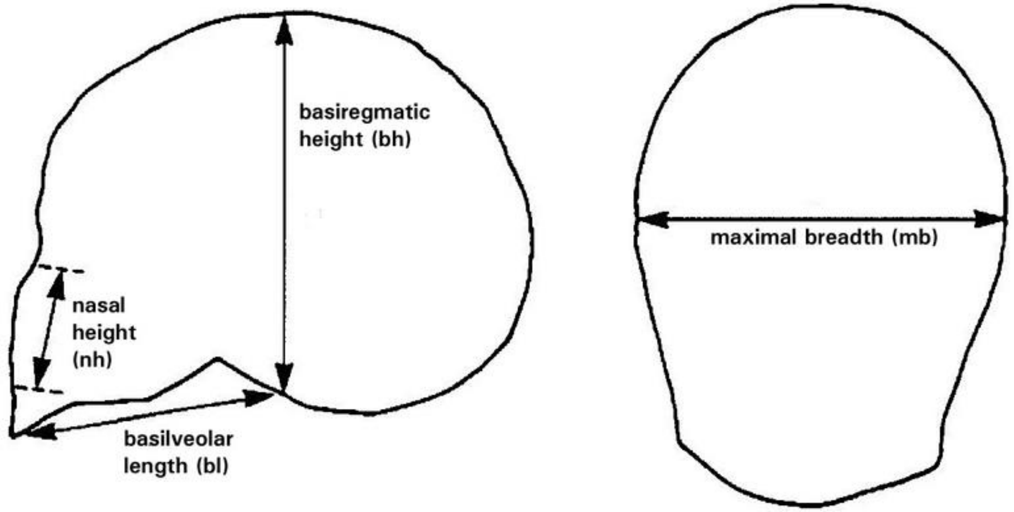
- **Sports** (number of points scored per game by an athlete)
- **Weather** Forecasting (average monthly rainfall)
- **Economy** (median income, unemployment rate)
- **Sales Tracking** (projected sales revenue from iphone 14's sold in a given month)
- **Medicine** (percentage of people who were pain free from using a drug)
- **Manufacturing** (number of cans of coke produced in a given month)

The **science of statistics** deals with the collection, analysis, interpretation, and presentation of data

Data is a collection of observations/measurements

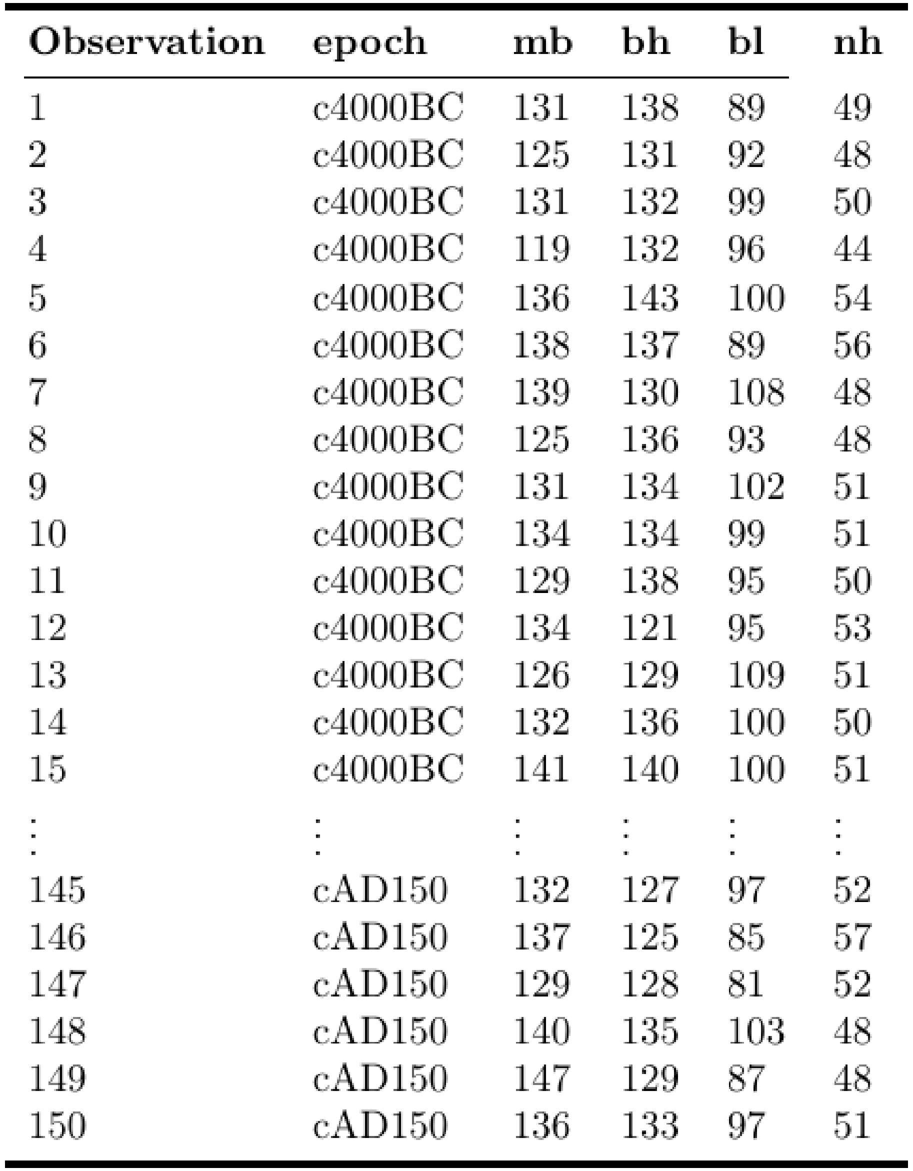
What Do Data Look Like?

The following are data from an analysis of 150 Egyptian skulls from 5 epochs of Egyptian history. For each skull, the epoch, and several measurements characterizing the shape of the skull are recorded



Source: D. J. Hand, F. Daly, A. D. Lunn, K. J. McConway and E. Ostrowski (1994). A Handbook of Small Datasets, Chapman and Hall/CRC, London.

Visualizing Tests for Equality of Covariance Matrices - Scientific Figure on ResearchGate. Available from: https://www.researchgate.net/figure/Diagram-of-the-skull-measurements-for-the-Egyptian-skulls-data-set-Maximal-breadth-and_fig4_325126938 [accessed 21 Aug, 2023]

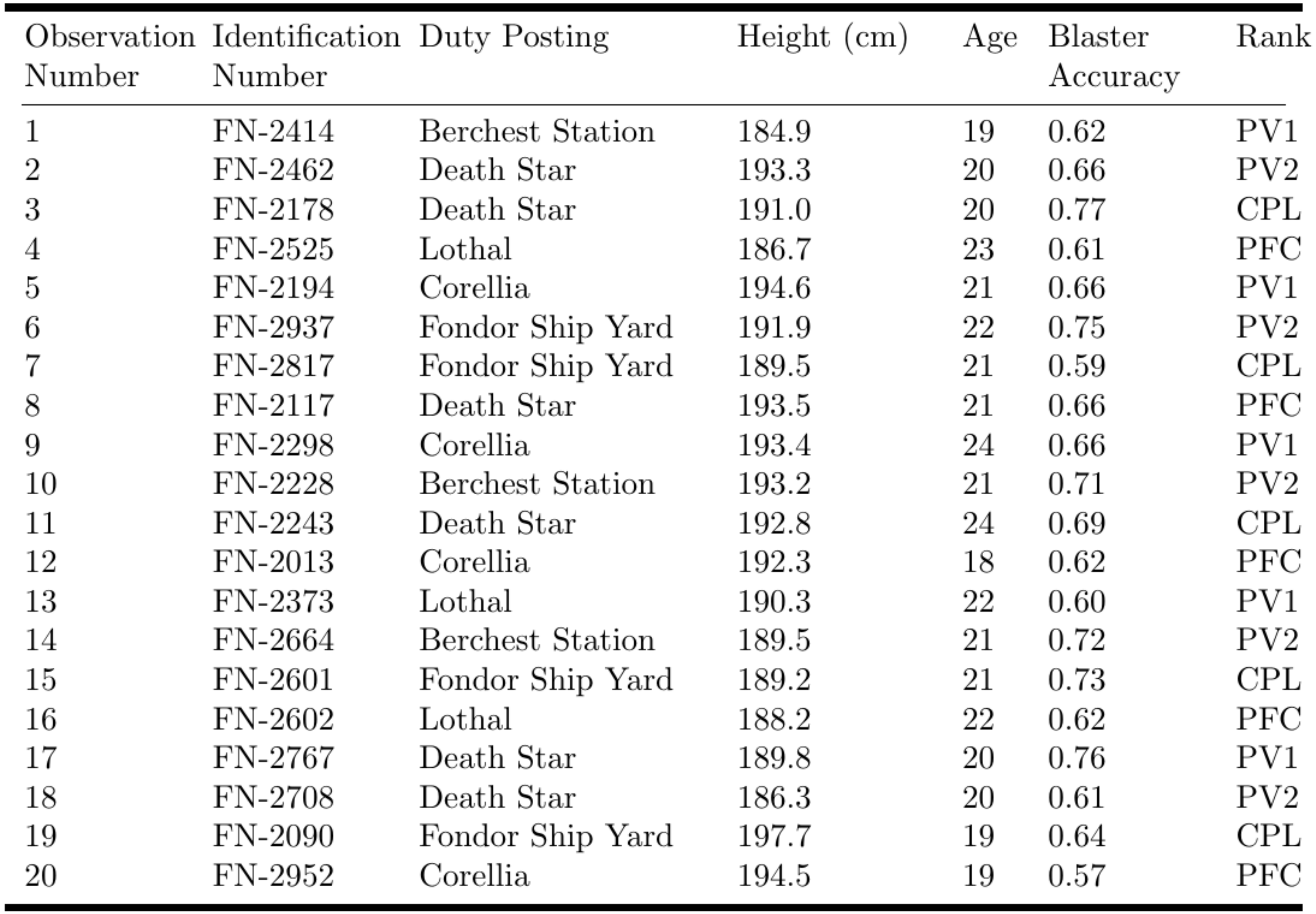


Observation	epoch	mb	bh	bl	nh
1	c4000BC	131	138	89	49
2	c4000BC	125	131	92	48
3	c4000BC	131	132	99	50
4	c4000BC	119	132	96	44
5	c4000BC	136	143	100	54
6	c4000BC	138	137	89	56
7	c4000BC	139	130	108	48
8	c4000BC	125	136	93	48
9	c4000BC	131	134	102	51
10	c4000BC	134	134	99	51
11	c4000BC	129	138	95	50
12	c4000BC	134	121	95	53
13	c4000BC	126	129	109	51
14	c4000BC	132	136	100	50
15	c4000BC	141	140	100	51
⋮	⋮	⋮	⋮	⋮	⋮
145	cAD150	132	127	97	52
146	cAD150	137	125	85	57
147	cAD150	129	128	81	52
148	cAD150	140	135	103	48
149	cAD150	147	129	87	48
150	cAD150	136	133	97	51

Example Data 2

The following table contains fictional data consisting of 20 observations of storm troopers who have just graduated from the Empire's Imperial Academy. The height, age, blaster accuracy, and future duty posting are recorded for each storm trooper

GET OUR TROOPS OFF TATOOINE!
KEEP YOUR EMPIRICAL HANDS OF MY HEALTHCARE
DEATH STAR DESTRUCTION WAS AN INSIDE JOB
OCCUPY DOCKING BAY 94 WE ARE THE 99%
END GALACTIC CORPORATE GREED



Observation Number	Identification Number	Duty Posting	Height (cm)	Age	Blaster Accuracy	Rank
1	FN-2414	Berchest Station	184.9	19	0.62	PV1
2	FN-2462	Death Star	193.3	20	0.66	PV2
3	FN-2178	Death Star	191.0	20	0.77	CPL
4	FN-2525	Lothal	186.7	23	0.61	PFC
5	FN-2194	Corellia	194.6	21	0.66	PV1
6	FN-2937	Fondor Ship Yard	191.9	22	0.75	PV2
7	FN-2817	Fondor Ship Yard	189.5	21	0.59	CPL
8	FN-2117	Death Star	193.5	21	0.66	PFC
9	FN-2298	Corellia	193.4	24	0.66	PV1
10	FN-2228	Berchest Station	193.2	21	0.71	PV2
11	FN-2243	Death Star	192.8	24	0.69	CPL
12	FN-2013	Corellia	192.3	18	0.62	PFC
13	FN-2373	Lothal	190.3	22	0.60	PV1
14	FN-2664	Berchest Station	189.5	21	0.72	PV2
15	FN-2601	Fondor Ship Yard	189.2	21	0.73	CPL
16	FN-2602	Lothal	188.2	22	0.62	PFC
17	FN-2767	Death Star	189.8	20	0.76	PV1
18	FN-2708	Death Star	186.3	20	0.61	PV2
19	FN-2090	Fondor Ship Yard	197.7	19	0.64	CPL
20	FN-2952	Corellia	194.5	19	0.57	PFC

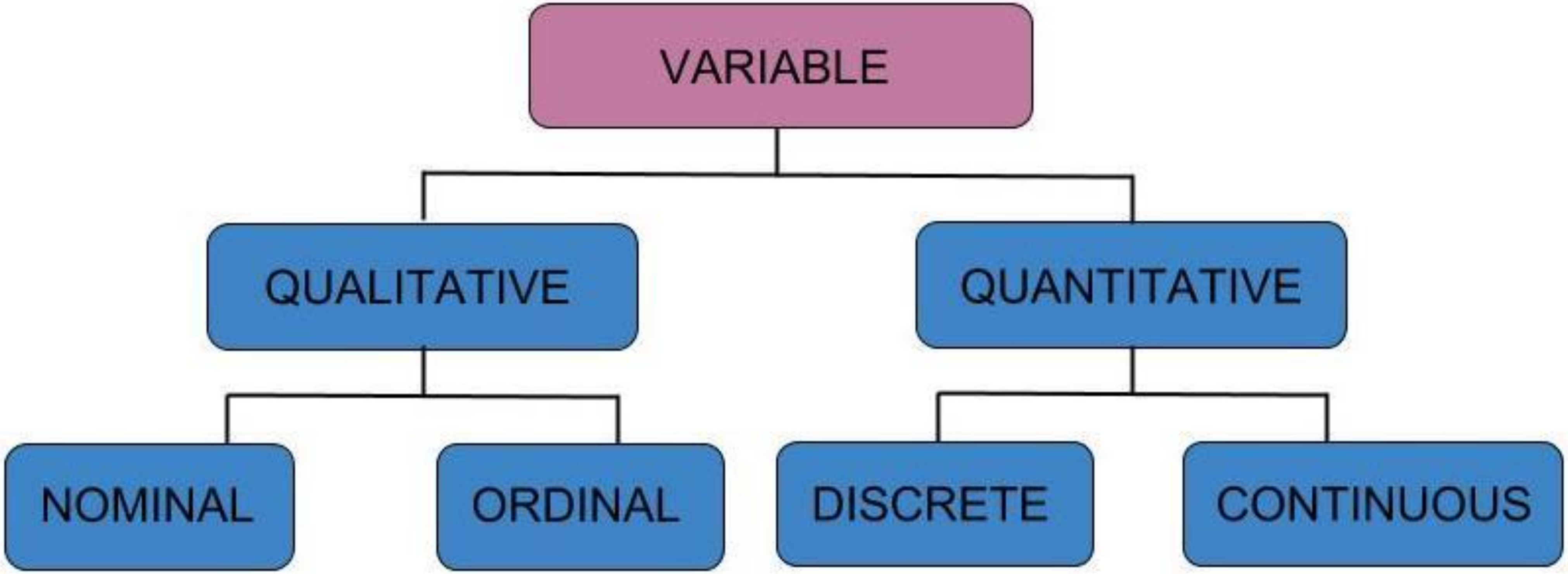
Checkpoint:

Data - a collection of observations/measurements on a set of variables - typically represented as a table.

Observation – the fundamental unit of data – typically observations are rows of a data table

Variables – characteristics of an observation – typically variables are the columns of a data table

Types of Variables: Qualitative vs Quantitative



Types of Variables: Qualitative vs Quantitative

Qualitative (Categorical) variable – non-numeric qualities or characteristics that can be placed in distinct categories

State/city (ID, WA, MT, ...)

Treatment (Drug/Placebo)

Genotype (AA, AT, TT)

Survival (live or die)

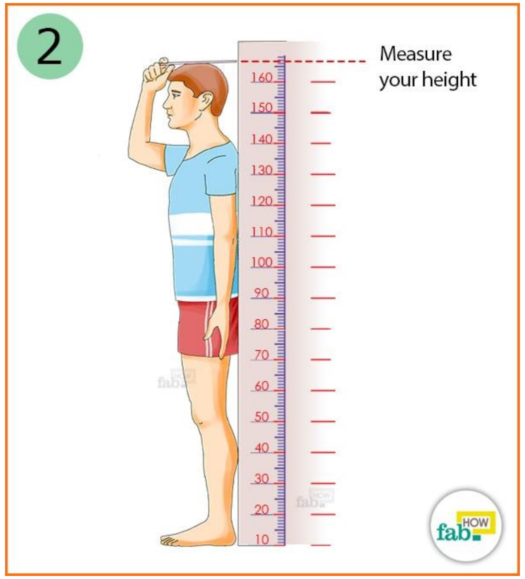
Quantitative (numeric) variable – numerical characteristics - can be ordered or ranked

Height (inches/cm)

Weight (lbs/kg)

Longevity/Age (number of years)

Dose (micrograms per gram)

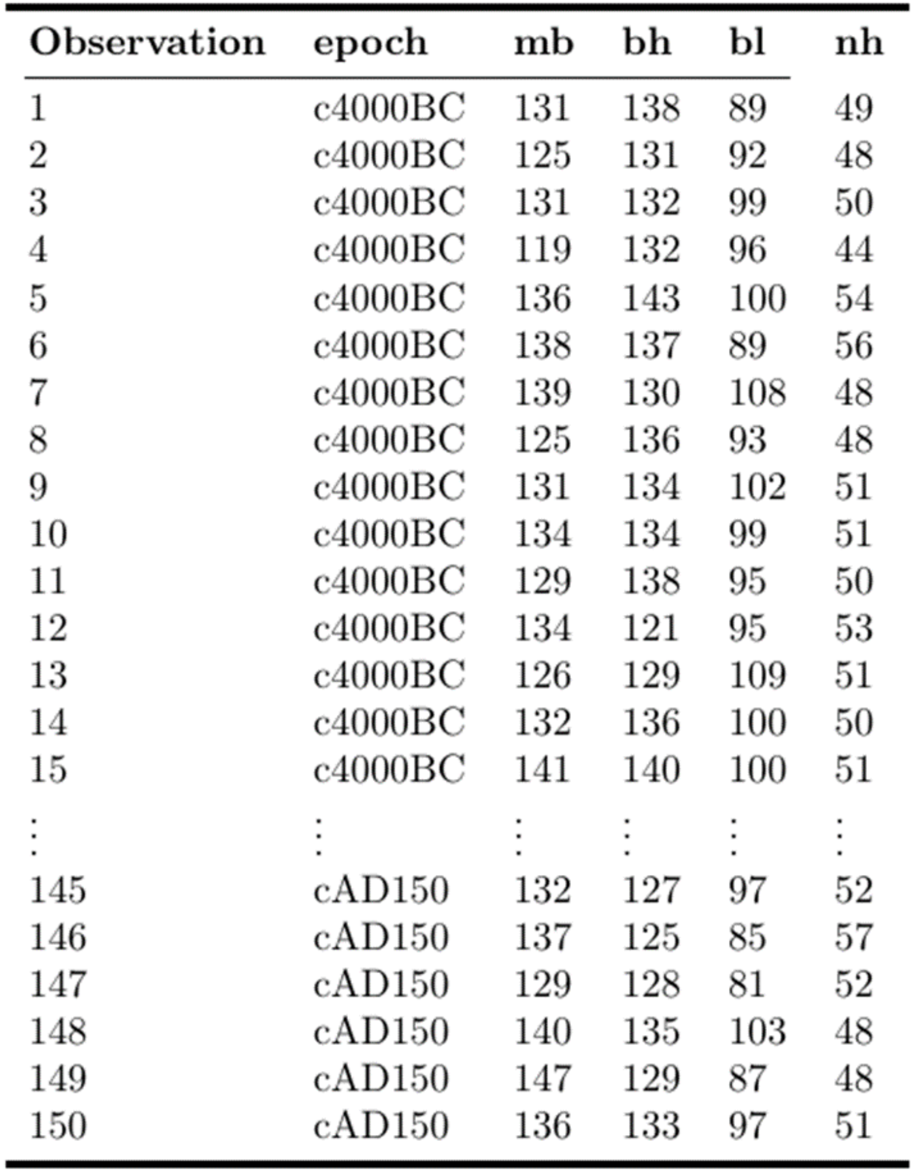


Practice: Qualitative vs Quantitative

From our example data of egyptian skulls, which variables are qualitative and which are quantitative?

Qualitative: Epoch

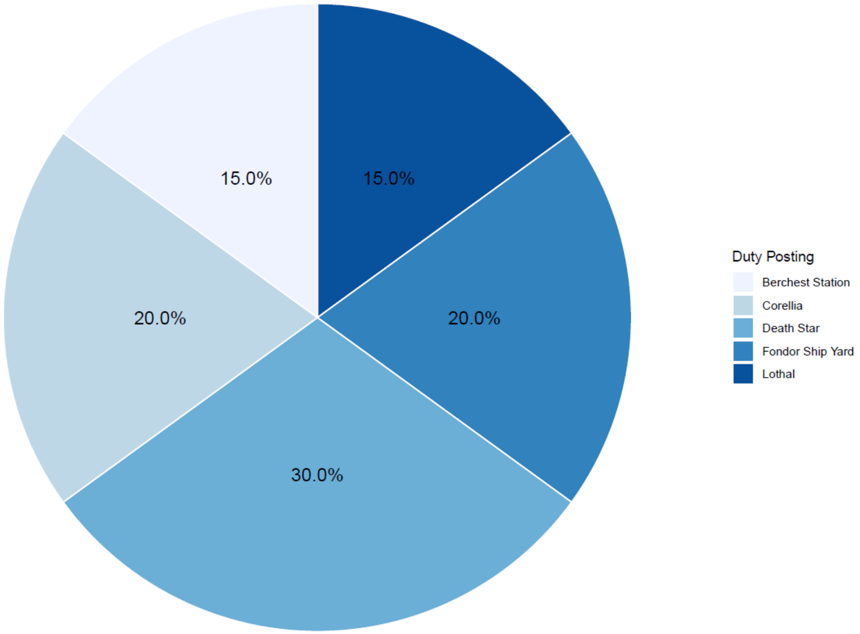
Quantitative: mb, bh, bl, nh



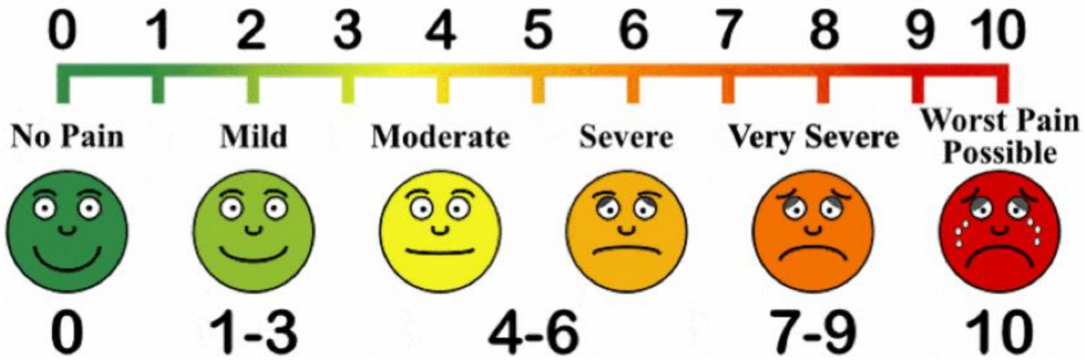
Observation	epoch	mb	bh	bl	nh
1	c4000BC	131	138	89	49
2	c4000BC	125	131	92	48
3	c4000BC	131	132	99	50
4	c4000BC	119	132	96	44
5	c4000BC	136	143	100	54
6	c4000BC	138	137	89	56
7	c4000BC	139	130	108	48
8	c4000BC	125	136	93	48
9	c4000BC	131	134	102	51
10	c4000BC	134	134	99	51
11	c4000BC	129	138	95	50
12	c4000BC	134	121	95	53
13	c4000BC	126	129	109	51
14	c4000BC	132	136	100	50
15	c4000BC	141	140	100	51
⋮	⋮	⋮	⋮	⋮	⋮
145	cAD150	132	127	97	52
146	cAD150	137	125	85	57
147	cAD150	129	128	81	52
148	cAD150	140	135	103	48
149	cAD150	147	129	87	48
150	cAD150	136	133	97	51

Types of Variables: Qualitative vs Quantitative

Qualitative nominal – non-numeric qualities or characteristics that can be placed in distinct categories that do not have a natural ordering (e.g labels, names, colors, etc)



Qualitative ordinal – non-numeric qualities or characteristics that can be placed in distinct categories with an inherent ordering (Likert responses, education level, military ranking)



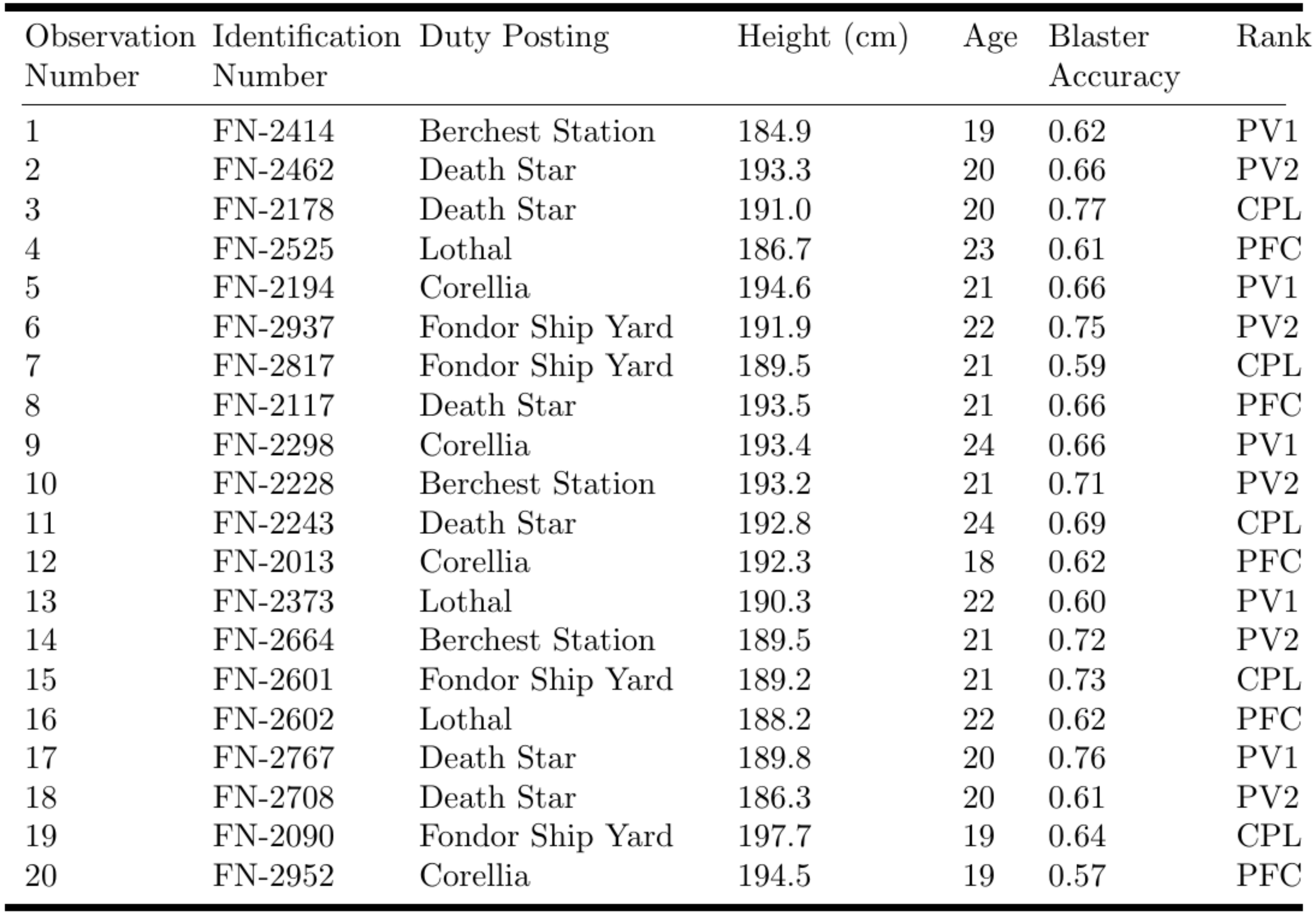
This Photo by Unknown Author is licensed under CC BY-NC-ND

Practice: Qualitative vs Quantitative

In the stormtrooper example data, which variable(s) are qualitative nominal and which are qualitative ordinal?

Qualitative nominal: ID Number, Duty Posting

Qualitative ordinal: Rank

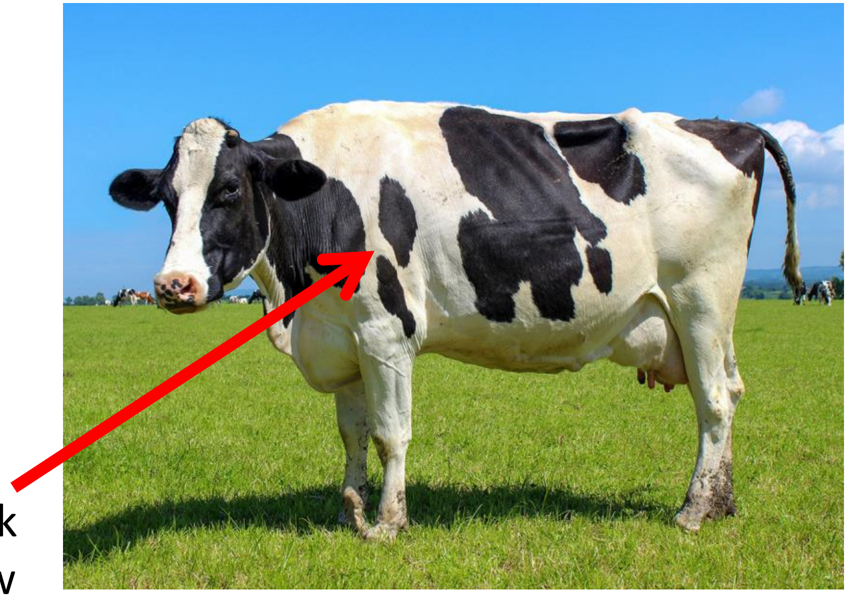


Observation Number	Identification Number	Duty Posting	Height (cm)	Age	Blaster Accuracy	Rank
1	FN-2414	Berchest Station	184.9	19	0.62	PV1
2	FN-2462	Death Star	193.3	20	0.66	PV2
3	FN-2178	Death Star	191.0	20	0.77	CPL
4	FN-2525	Lothal	186.7	23	0.61	PFC
5	FN-2194	Corellia	194.6	21	0.66	PV1
6	FN-2937	Fondor Ship Yard	191.9	22	0.75	PV2
7	FN-2817	Fondor Ship Yard	189.5	21	0.59	CPL
8	FN-2117	Death Star	193.5	21	0.66	PFC
9	FN-2298	Corellia	193.4	24	0.66	PV1
10	FN-2228	Berchest Station	193.2	21	0.71	PV2
11	FN-2243	Death Star	192.8	24	0.69	CPL
12	FN-2013	Corellia	192.3	18	0.62	PFC
13	FN-2373	Lothal	190.3	22	0.60	PV1
14	FN-2664	Berchest Station	189.5	21	0.72	PV2
15	FN-2601	Fondor Ship Yard	189.2	21	0.73	CPL
16	FN-2602	Lothal	188.2	22	0.62	PFC
17	FN-2767	Death Star	189.8	20	0.76	PV1
18	FN-2708	Death Star	186.3	20	0.61	PV2
19	FN-2090	Fondor Ship Yard	197.7	19	0.64	CPL
20	FN-2952	Corellia	194.5	19	0.57	PFC

Quantitative Variables: Discrete vs Continuous

- **Quantitative discrete** – quantitative variables that take on distinct, countable values such 0,1,2,3… (whole numbers or integers)

  - all counts are quantitative discrete variables



e.g. number of black spots on a dairy cow

- **Quantitative continuous** – variables that can take on infinite number of values within an interval of any two specific values (e.g temperature ◦F, height in inches, speed in miles per hour)

Min

Max

40°F

Moscow, ID Daily Surface Temperature (July)

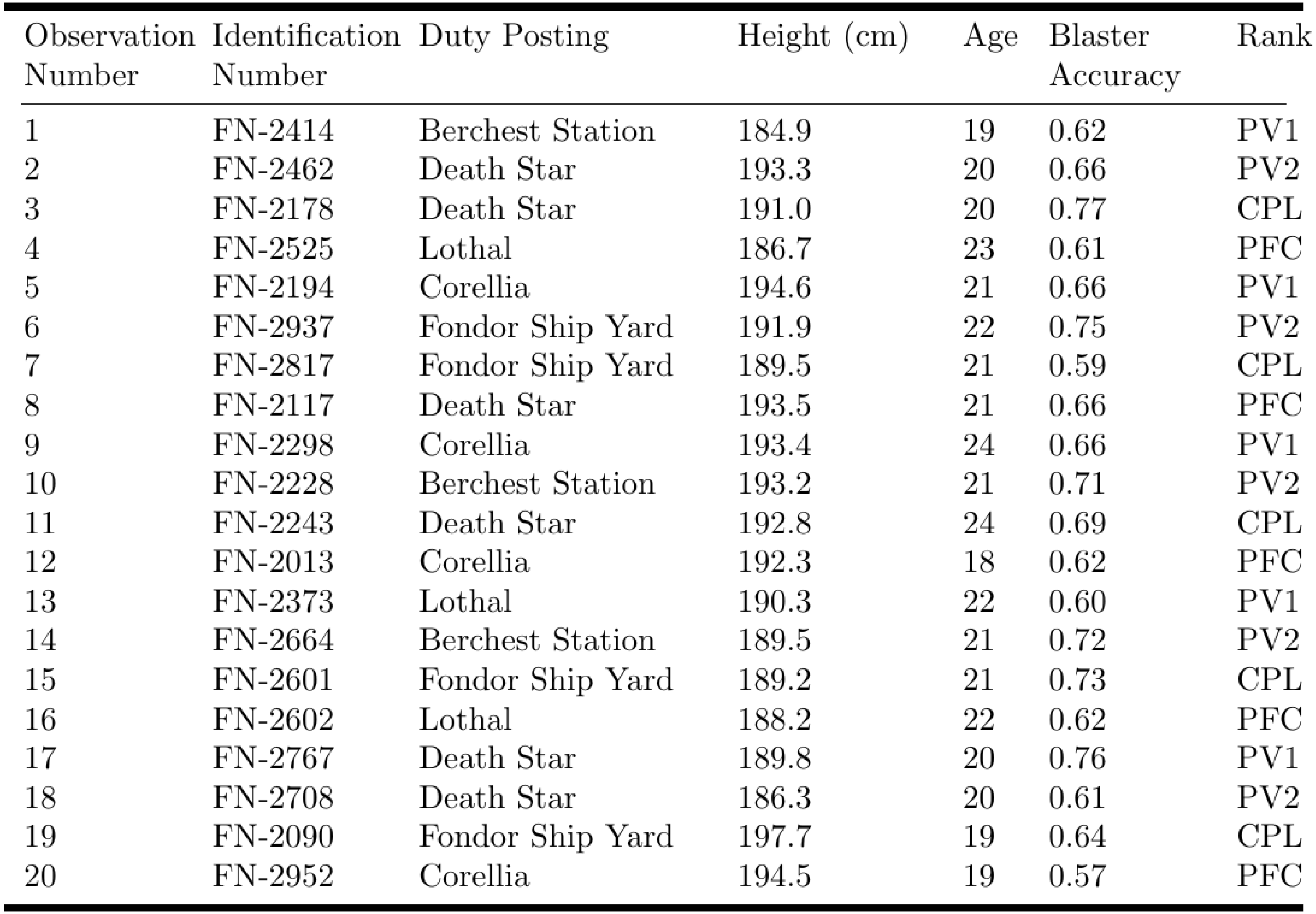
97°F

Practice: Discrete vs Continuous

In the stormtrooper example data, which variable(s) are quantitative discrete and which are quantitative continous?

Quantitative discrete: Age (a count of the number of years)

Quantitative continuous: Height, Blaster Accuracy



Observation Number	Identification Number	Duty Posting	Height (cm)	Age	Blaster Accuracy	Rank
1	FN-2414	Berchest Station	184.9	19	0.62	PV1
2	FN-2462	Death Star	193.3	20	0.66	PV2
3	FN-2178	Death Star	191.0	20	0.77	CPL
4	FN-2525	Lothal	186.7	23	0.61	PFC
5	FN-2194	Corellia	194.6	21	0.66	PV1
6	FN-2937	Fondor Ship Yard	191.9	22	0.75	PV2
7	FN-2817	Fondor Ship Yard	189.5	21	0.59	CPL
8	FN-2117	Death Star	193.5	21	0.66	PFC
9	FN-2298	Corellia	193.4	24	0.66	PV1
10	FN-2228	Berchest Station	193.2	21	0.71	PV2
11	FN-2243	Death Star	192.8	24	0.69	CPL
12	FN-2013	Corellia	192.3	18	0.62	PFC
13	FN-2373	Lothal	190.3	22	0.60	PV1
14	FN-2664	Berchest Station	189.5	21	0.72	PV2
15	FN-2601	Fondor Ship Yard	189.2	21	0.73	CPL
16	FN-2602	Lothal	188.2	22	0.62	PFC
17	FN-2767	Death Star	189.8	20	0.76	PV1
18	FN-2708	Death Star	186.3	20	0.61	PV2
19	FN-2090	Fondor Ship Yard	197.7	19	0.64	CPL
20	FN-2952	Corellia	194.5	19	0.57	PFC

Sampling and Data

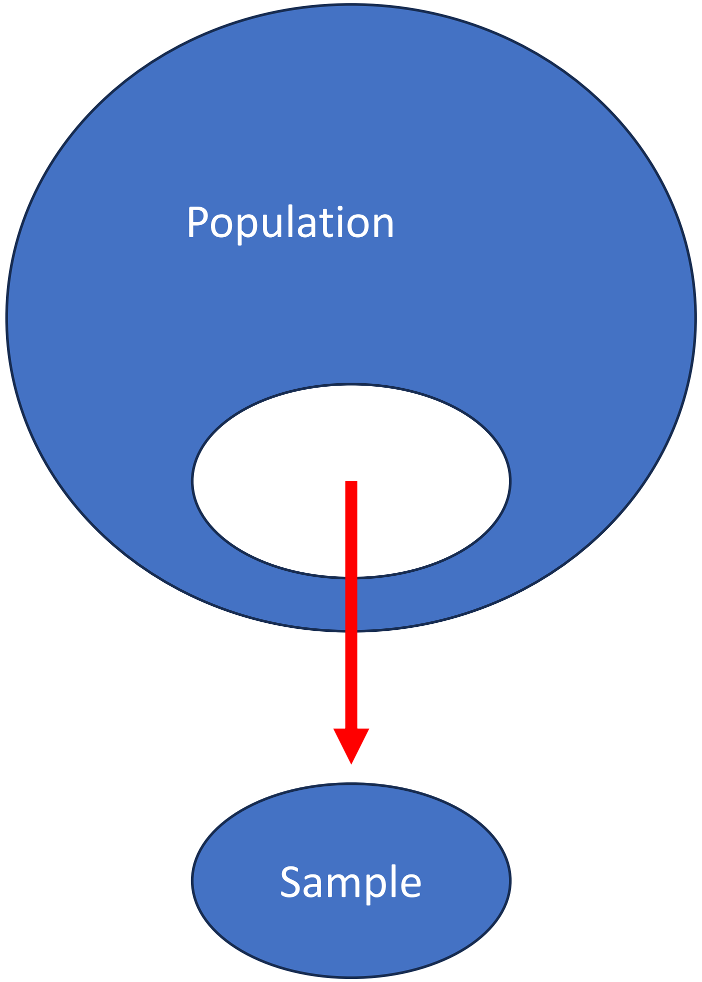
Statistics is generally concerned with studying properties of a **population** – the collection of all possible persons, events, or objects of interest

e.g the set of all possible observations – observed + unobserved

- Populations can be *real* and *finite/countable* (e.g. All employees at a company) or *hypothetical* and potentially *infinite/uncountable* (e.g all possible hands in a game of poker)

A **Sample** is a subset of the population that we actually observe – the observed observations

- The idea of **Sampling** is to select a portion or individuals or objects that are *representative* of the population
- By studying the sample we can gain insights about the population

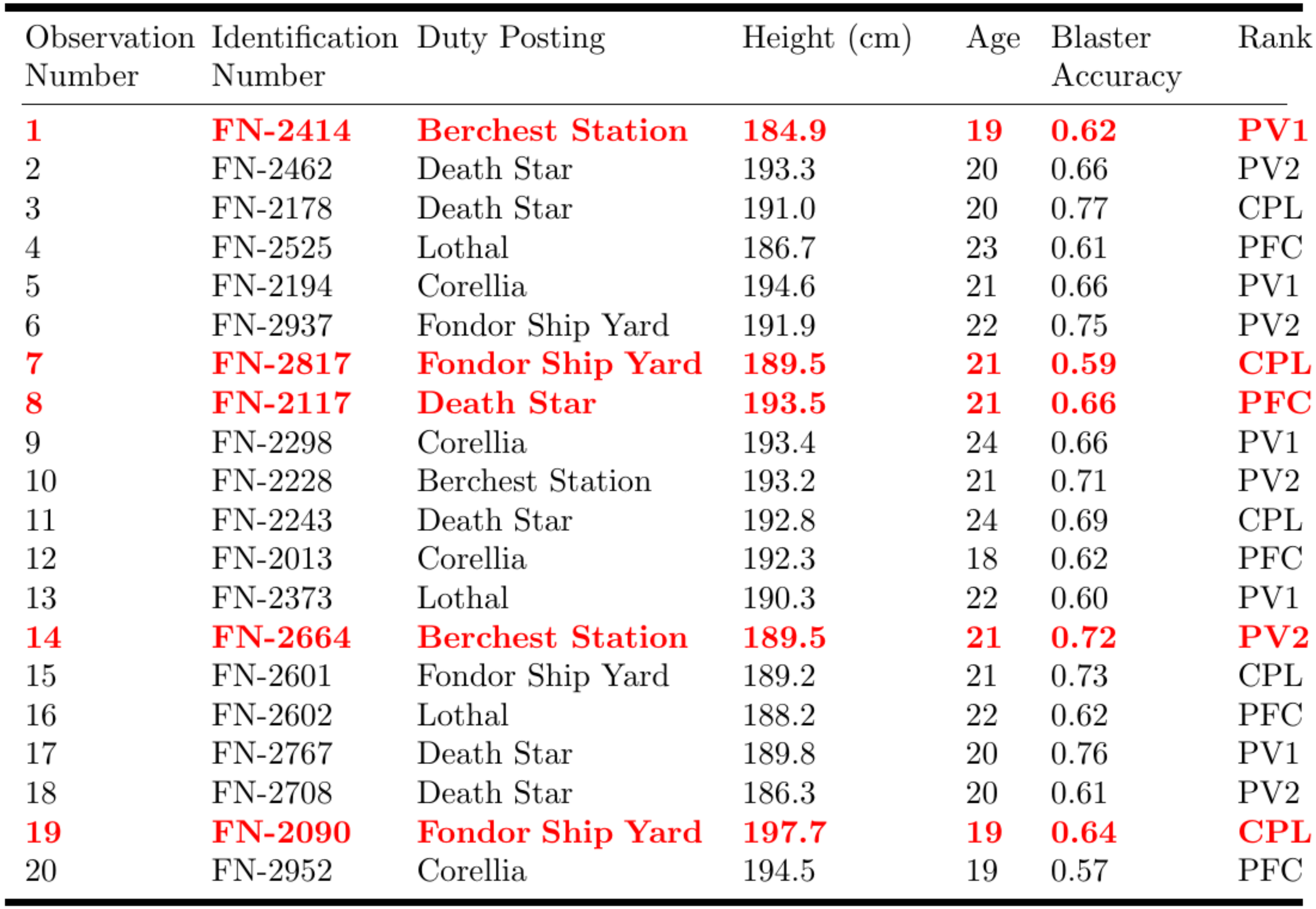


Example

The population is a set of N observations
$\{x_1, x_2, x_3, \dots x_N\}$

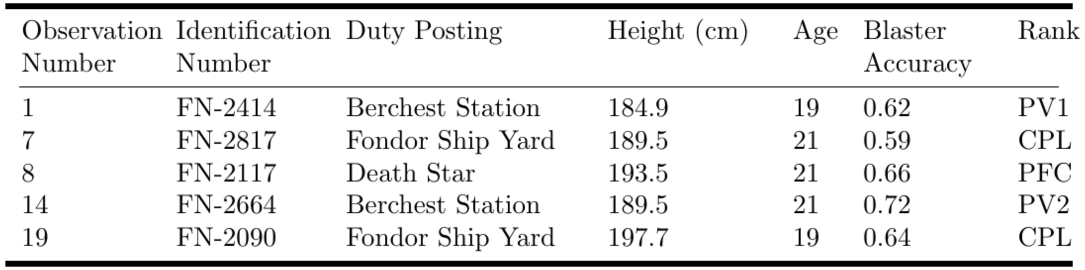
The sample is a set of n observations
$\{x_1, x_7, x_8, \dots x_n\}$

Population: $N = 20$



Observation Number	Identification Number	Duty Posting	Height (cm)	Age	Blaster Accuracy	Rank
1	**FN-2414**	**Berchest Station**	**184.9**	**19**	**0.62**	**PV1**
2	FN-2462	Death Star	193.3	20	0.66	PV2
3	FN-2178	Death Star	191.0	20	0.77	CPL
4	FN-2525	Lothal	186.7	23	0.61	PFC
5	FN-2194	Corellia	194.6	21	0.66	PV1
6	FN-2937	Fondor Ship Yard	191.9	22	0.75	PV2
7	**FN-2817**	**Fondor Ship Yard**	**189.5**	**21**	**0.59**	**CPL**
8	**FN-2117**	**Death Star**	**193.5**	**21**	**0.66**	**PFC**
9	FN-2298	Corellia	193.4	24	0.66	PV1
10	FN-2228	Berchest Station	193.2	21	0.71	PV2
11	FN-2243	Death Star	192.8	24	0.69	CPL
12	FN-2013	Corellia	192.3	18	0.62	PFC
13	FN-2373	Lothal	190.3	22	0.60	PV1
14	**FN-2664**	**Berchest Station**	**189.5**	**21**	**0.72**	**PV2**
15	FN-2601	Fondor Ship Yard	189.2	21	0.73	CPL
16	FN-2602	Lothal	188.2	22	0.62	PFC
17	FN-2767	Death Star	189.8	20	0.76	PV1
18	FN-2708	Death Star	186.3	20	0.61	PV2
19	**FN-2090**	**Fondor Ship Yard**	**197.7**	**19**	**0.64**	**CPL**
20	FN-2952	Corellia	194.5	19	0.57	PFC

Sample: n = 5



Observation Number	Identification Number	Duty Posting	Height (cm)	Age	Blaster Accuracy	Rank
1	FN-2414	Berchest Station	184.9	19	0.62	PV1
7	FN-2817	Fondor Ship Yard	189.5	21	0.59	CPL
8	FN-2117	Death Star	193.5	21	0.66	PFC
14	FN-2664	Berchest Station	189.5	21	0.72	PV2
19	FN-2090	Fondor Ship Yard	197.7	19	0.64	CPL

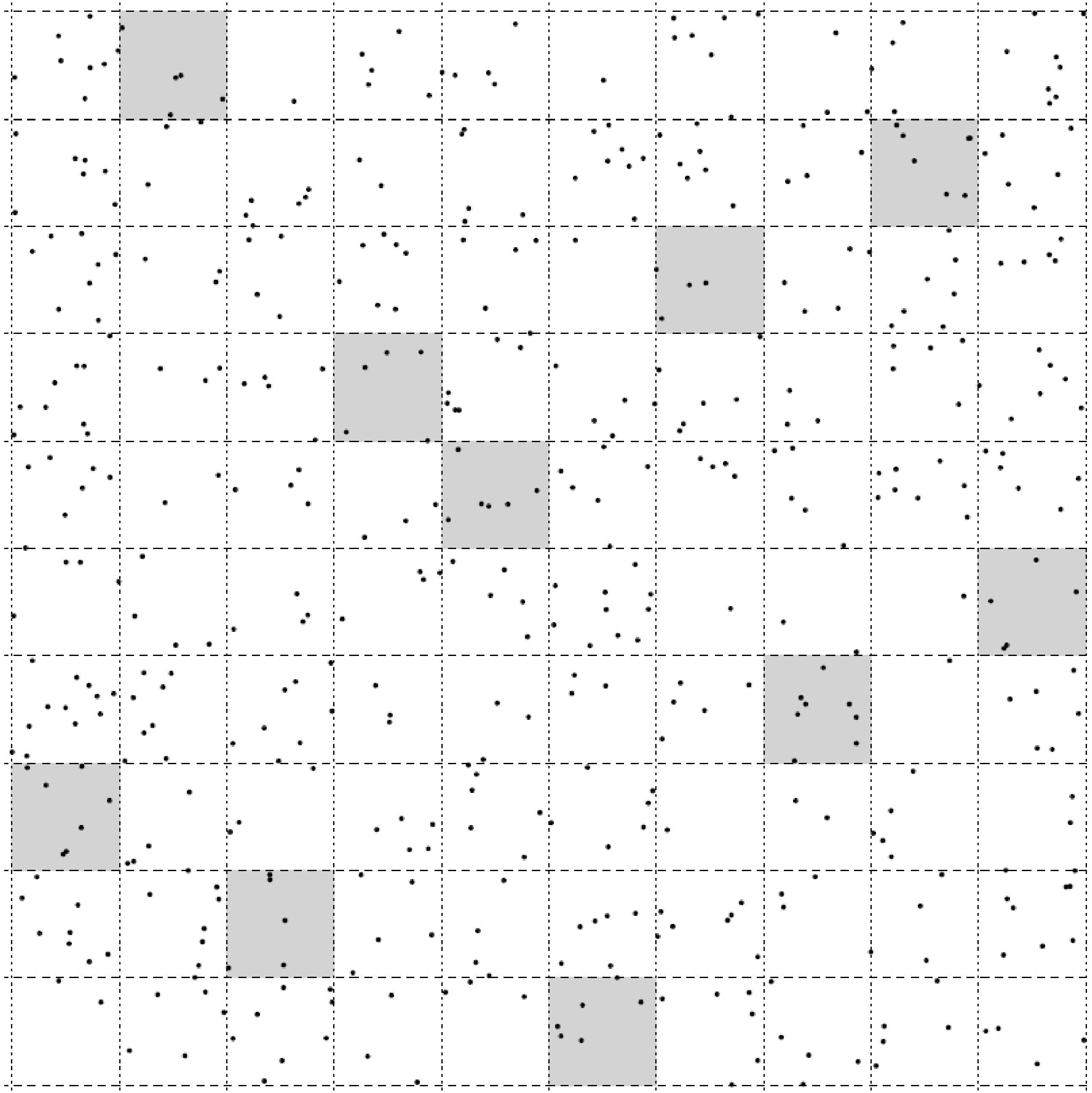
Example 2

Consider a rectangular-shaped piece of land that has been divided into 100 smaller rectangular units, each containing something of interest (e.g., trees, burrows, archaeological artifacts).

A subset of 10 of those smaller units was selected and the number of objects in each of these units was counted.

Population Size: $N = 100$

Sample Size: $n = 10$



Why is statistics so valuable?

- Most of the time, we can’t measure everyone or every unit in the population and therefore must limit our measurements to a sample.

- Statistics primarily deals with **estimation** – the process of inferring an unknown quantity about a population using set of sample data

- The tools for estimation allow us to approximate almost everything about populations **using only samples**.

Where we can go with estimates

- With **estimates** we can
  - Assess differences among groups and relationships between variables.
  - Describe populations. Examples of estimates include averages, proportions, measures of variation, and measures of relationship.
  - Then we can ask and answer questions or formally, test and evaluate hypotheses.

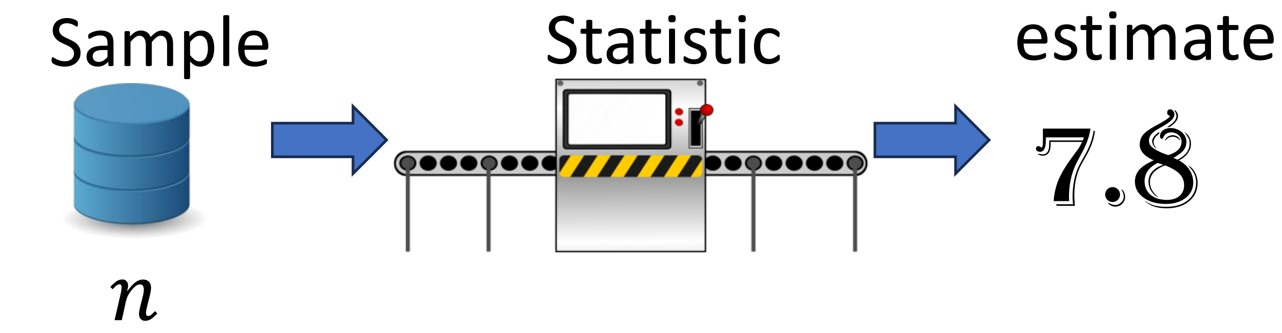
Statistics Vs Parameters

A **statistic** is a numerical characteristic of a **sample** that estimates a population parameter.

A **parameter** is a numerical characteristic of a **population** that can be estimated by a statistic.

Put another way…

A **statistic** is a function of the observations of in a sample while a **parameter** is a function of all observations in the population.



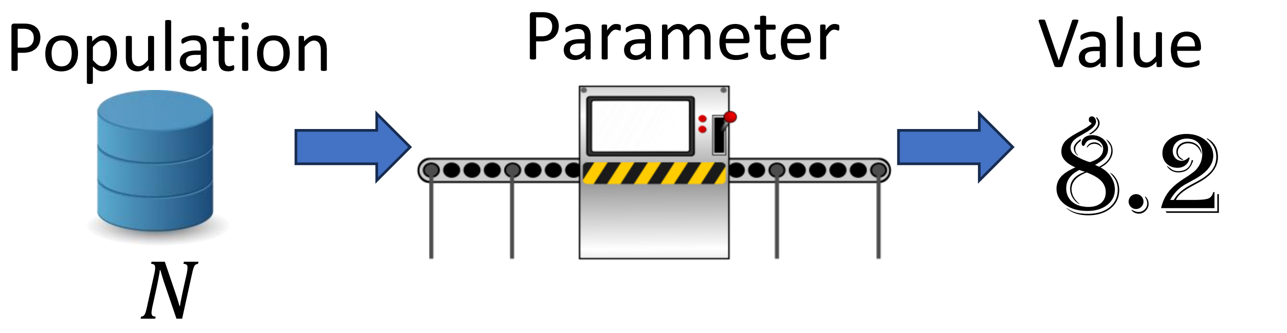
Sample

Statistic

estimate

7.8

n



Population

Parameter

Value

8.2

N

Mean and Proportion

A **proportion** describes the fraction of a whole that represent some property or category. Usually, it is expressed a percentage.

Notation:

$\hat{p}$ - denotes the sample proportion

p - denotes the population proportion

The arithmetic **mean** is the center of a set of data (we often use the words mean and average interchangeably)

Notation:

$\bar{x}$ - denotes the mean of a sample

μ – denotes the mean of a population (i.e the population parameter)

Parameter

$$\mu = \frac{1}{N}\sum_{i=1}^{N} x_i$$

$$p = \frac{\text{Number of objects in category}}{N}$$

Statistic

$$\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$$

$$\hat{p} = \frac{\text{Number of objects in category}}{n}$$

Example: Gallup Pool

- On April 20, 2010, one of the worst environmental disasters took place in the Gulf of Mexico when the Deepwater Horizon offshore oil rig exploded.
- In response to the spill, many activists called for an end to offshore drilling for oil.
- Almost nine months later, turbulence in the middle east caused the price of oil to surge to an all-time high.
- In March 2011, Gallup conducted a survey and found that 60% of Americans favored offshore drilling as means to reduce U.S dependence on foreign oil.
- The poll was based on interviews with 1,021 adults aged 18 and older, living in the continental U.S, and selected using random digit dialing.
- **What is the population under study, and what is the population parameter being estimated? What is the sample statistic ?**

Descriptive Vs. Inferential Statistics

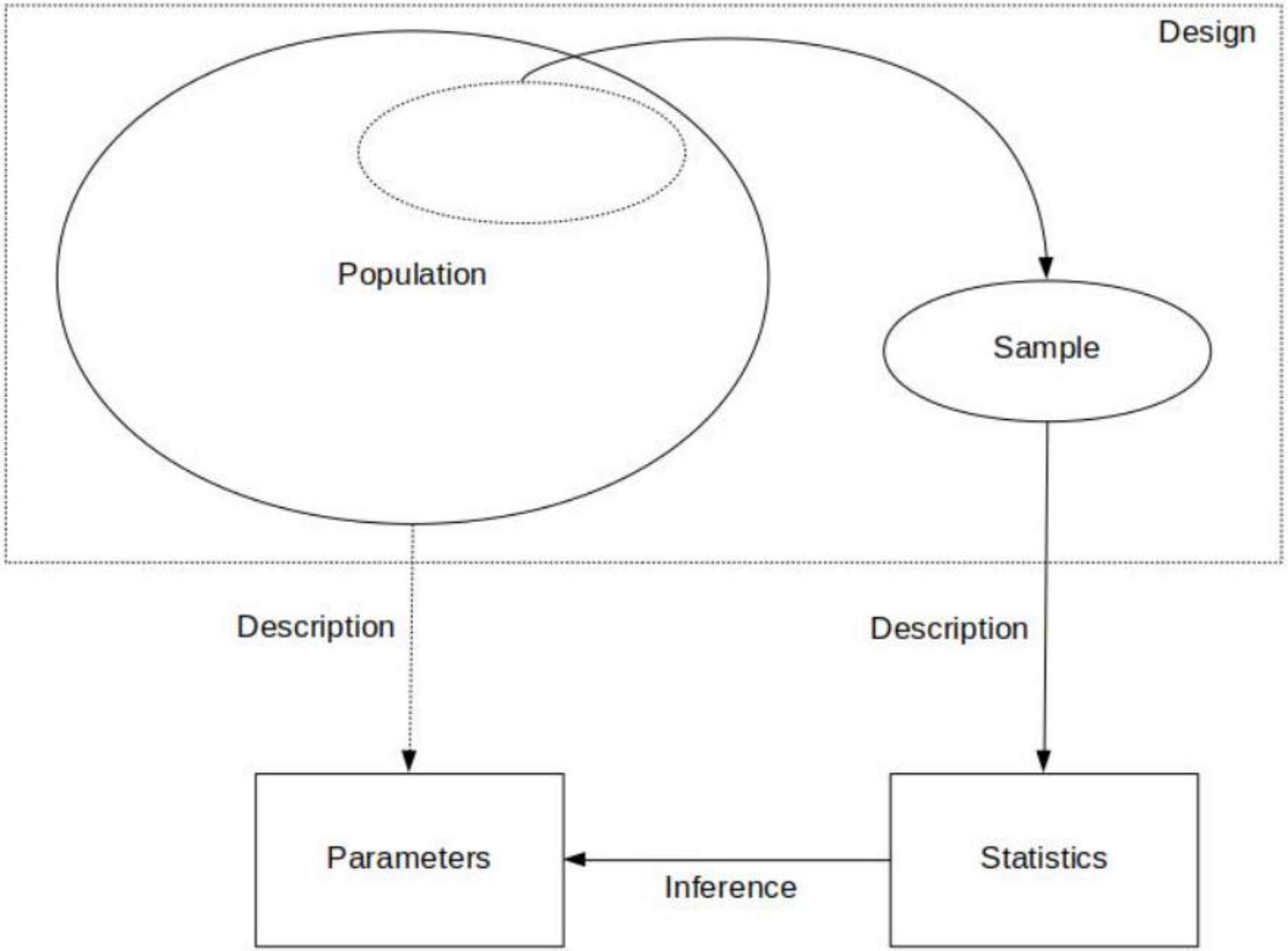
1. **Design** – The process/method in which we plan to collect data to answer our statistical question

2. **Descriptive Statistics –** refers to describing the observations in a sample using statistics or a population using parameters
    1. - collection, organization, summarization and visualization of data

3. **Inferential Statistics** (or statistical inference) – refers to using a sample (usually a statistic) to answer a question about a population (such as estimating the value of a parameter)

    - estimation, hypothesis testing, determining relationships among variables, prediction

The Big Picture


Design
Population
Sample
Description
Description
Parameters
Inference
Statistics