Introduction (Ch 1-2)

Yifan Zhu

Iowa State University

Outline

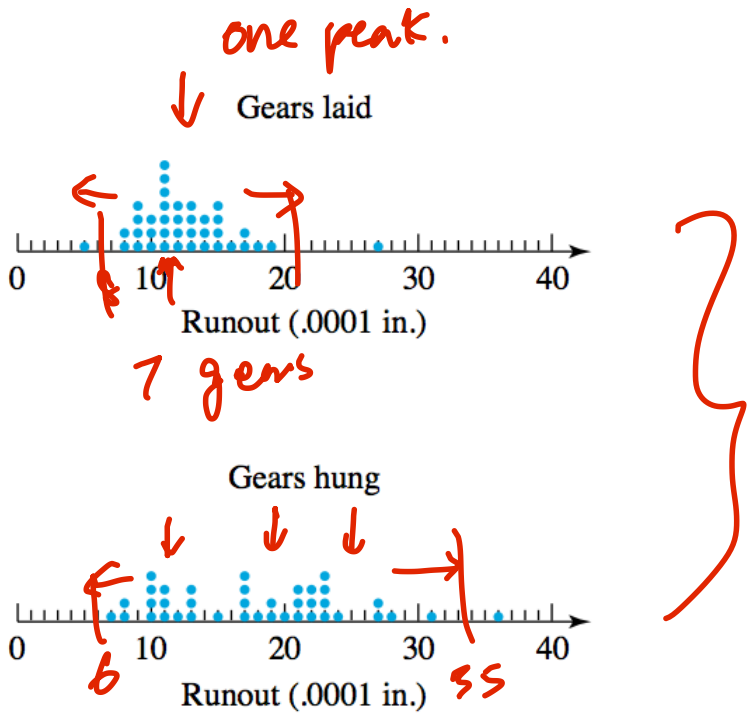
- Statistics is the science of collecting, analyzing, presenting, and making decisions from data.
- Tasks:
  - Summary: describe and display the data.
  - Inference: draw conclusions from the data.
  - Interpretation: explain those conclusions in layman's terms.
- Applications in Engineering
  - Quality control
  - Process control
  - Reliability
  - Risk management
  - System identification
  - Design of experiments

Example: Gears

- Data taken from "Statistical Analysis: Mack Truck Gear Heat Treating Experiments" by P. Brezler (Heat Treating, November, 1986)
- How should gears be loaded into a continuous carburizing furnace in order to minimize distortion during heat treating?
- Options: lay or hang the gears
- Measure of distortion: thrust face runout (0.0001 in)

Gears Laid	Gears Hung
5, 8, 8, 9, 9,	7, 8, 8, 10, 10,
9, 9, 10, 10, 10,	10, 10, 11, 11, 11,
11, 11, 11, 11, 11,	12, 13, 13, 13, 15,
11, 11, 12, 12, 12,	17, 17, 17, 17, 18,
12, 13, 13, 13, 13,	19, 19, 20, 21, 21,
14, 14, 14, 15, 15,	21, 22, 22, 22, 23,
15, 15, 16, 17, 17,	23, 23, 23, 24, 27,
18, 19, 27	27, 28, 31, 36

Summary



Mean laid runout: 12.6×10^{-4} in

Mean hung runout: 17.9×10^{-4} in

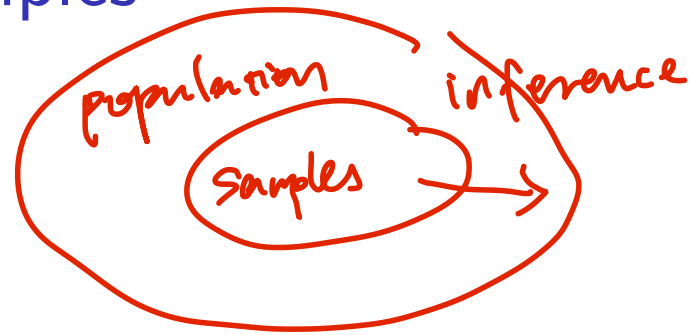
} sample means.

Inference and Interpretation

- Inference uses probability theory, symbols, and equations to answer questions like:
  - Can we be confident that the "true" mean laid runout is within 2.0×10^{-4} of our observed mean of 12.6×10^{-4} in?
  - Is the mean laid runout significantly lower than the mean hung runout?
- Interpretation explains those answers in layman's terms without all the probability theory, symbols and equations.

Outline

Populations and Samples



- Definitions
  - **Sample**: the collection of objects (most relevant to the central goals of the study) selected for direct measurement.
  - **Population**: the bigger group of things or people from which the sample was taken.
- Gears study
  - Sample: the 77 gears arranged, tested, and measured for distortion
  - Population: All the gears with the same make and model as those included in the experiment.

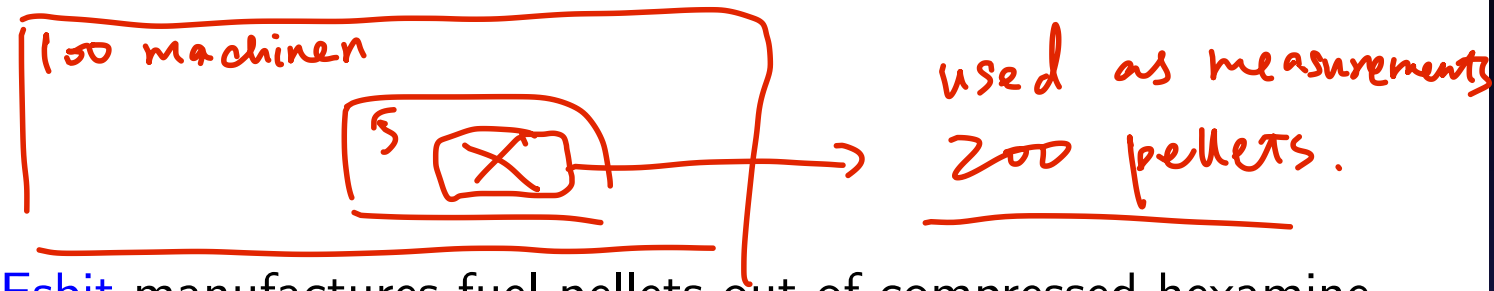
Your turn: state the sample and the population.

- On Dec. 1-2, 2012, the Gallup Poll conducted a study to find out what proportion of Americans prefer a compromise on the Fiscal Cliff issue.
- 1000 adults were randomly selected for telephone interviews. The adults were aged 18 and older and living in any of the 50 U.S. states or the District of Columbia.

Answers

- Sample: the 1000 adults selected.
- Population: All adults aged 18 and older who live in any of the 50 U.S. states or the District of Columbia.

Your turn: state the sample and the population.



- Esbit manufactures fuel pellets out of compressed hexamine powder. Suppose a new shipment of 100 pelletizing machines arrives, and the goal of a new study is to determine the quality of this particular new shipment.
- 5 machines out of the 100 are randomly selected for comprehensive testing in which each produces 200 pellets, and each pellet's mass, volume, flash point, and rate of combustion are measured.

4×200

Answers

- Sample: The 5 pelletizing machines selected for testing.
- Population: The new shipment of 100 pelletizing machines.

cannot extend to other machines.

Outline

Measurement issues

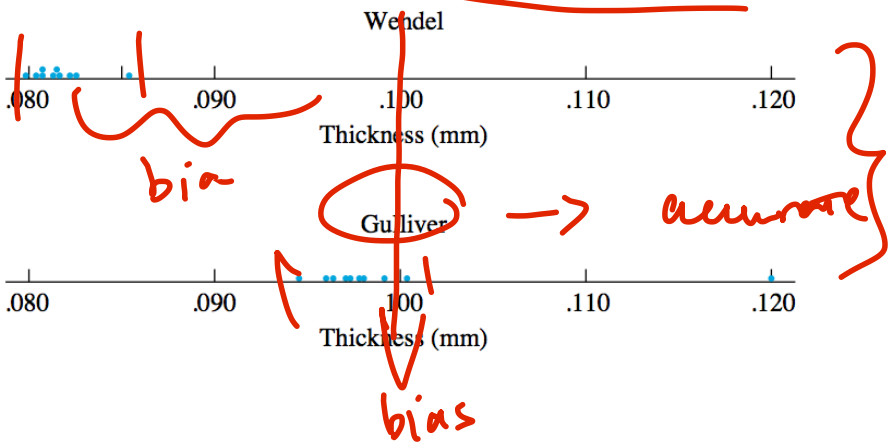
A measurement system is:

- **Valid** if it usefully and appropriately represents the feature of interest.
- **Accurate** if it produces the correct value on average.
- **Precise** if there is little variation in repeated measurements of the same object.

Example: paper

Two students measured the thickness of paper in the same book. Each student took 10 measurements.

Wendel:	.0807, .0826, .0854, .0817, .0824, .0799, .0812, .0807, .0816, .0804
Gulliver:	.0972, .0964, .0978, .0971, .0960, .0947, .1200, .0991, .0980, .1033



If the true thickness of the paper is 0.1 mm, then:

1. Who is more accurate?
2. Who is more precise?

Example: paper

1. Gulliver is more accurate because his measurements are closer to 0.1 mm on average. bias lower
2. Wendel is more precise because her measurement s are less varied. variance lower.

Types of data

- **Categorical (qualitative)**: Each measurement is a non-numerical value (male or female, operating condition A or B, green or blue).
- **Numerical (quantitative)**: Each measurement is a number (thrust face runout of gears, mass of fuel pellets)
  - **Discrete**: measurements are separated points (number of pages in a book, number of shark attacks, CPU FLOPs per day, credit card numbers)
  - **Continuous**: measurements lie on a continuum (mass of fuel pellets, mpg of a car, thrust face runout of gears)
- **Univariate**: one measurement per sample unit
- **Multivariate**: multiple measurements per sample unit (2 measurements = bivariate data)
- **Paired Data**: bivariate data where both variables are attempting to quantify the "same thing" - often, before + after (e.g. like metal specimen hardness before and after heat treating) or measurements of the same quantity made with different instrucments/systems

Gears runout data: univariate or bivariate?

Gears Laid	Gears Hung
5, 8, 8, 9, 9,	7, 8, 8, 10, 10,
9, 9, 10, 10, 10,	10, 10, 11, 11, 11,
11, 11, 11, 11, 11,	12, 13, 13, 13, 15,
11, 11, 12, 12, 12,	17, 17, 17, 17, 18,
12, 13, 13, 13, 13,	19, 19, 20, 21, 21,
14, 14, 14, 15, 15,	21, 22, 22, 22, 23,
15, 15, 16, 17, 17,	23, 23, 23, 24, 27,
18, 19, 27	27, 28, 31, 36

Answer: bivariate

Arrange the data in a table, where:

- Each row is a **sample unit**, or thing that you measure (gear, in this case).
- Each column is a **variable**, or characteristic that you control or measure.

categorical numerical

Arrangement	Runout
laid	5
laid	8
laid	8
laid	9
laid	9
⋮	⋮
hung	31
hung	36

# of col = # of row

This "sample unit $\times$ variable" arrangement is the standard way to display data, and it helps account for all the variables.

Outline

Variables

- **Variable**: a characteristic that you control or measure
- **Level**: a possible value of a variable

Types of variables:

- **Treatment variable**: a variable that the experimenter sets by acting on the sample (gear arrangement: laid or hung).
  - This action on the sample is called a **treatment** (laying or hanging the gears).
  - Treatment levels divide the sample into **treatment groups** (laid gears and hung gears).
- **Concomitant variable**: a passively measured variable (i.e., runout)
- **Factor**: any discrete or categorical variable with a finite set of possible levels (operating condition A, B, or C for a chemical process)
- **Response variable**: the outcome or result of a study (i.e., runout)
- **Predictor variable (predictor, covariate)**: any variable that is not the response

Blocking variables

Blocking variable: a discrete concomitant variable that describes some innate characteristic of the sample before treatment (gear size: big or small).

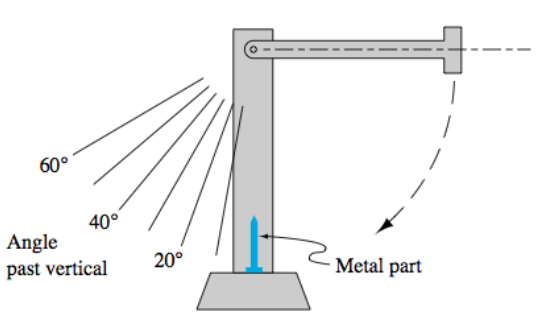
- Blocking variables divide the sample into smaller samples called **blocks** (big gears and small gears).
- In practice, each block is collected as a separate sample. (Take a sample of big gears from a population of big gears, and then take a sample of small gears from a population of small gears.)

Outline

- **Experimental study (experiment)**: a study with at least one treatment variable: i.e., one in which the investigator acts on the sample in some way.
- **Observational study**: a study with no treatment variables: all variables are concomitant, and all phenomena are passively observed.

Example: metallurgy

- A senior design class in metallurgical engineering took on the project of helping a manufacturer determine the brittleness of a spike-shaped metal part. The manufacturer wants these parts to bend with impact rather than shatter.
- Each spike was hit with a swinging arm, and the response (bend or shatter) was recorded. The angle past vertical of the swinging arm varied among spikes.
- Some spikes were made of metal A, others were made of metal B.
- The experimental conditions — i.e., the temperature of the room, the force holding the spikes in place, etc. — were held constant across levels of the arm.



Is this an experiment or an observational study? Identify and classify all the variables.

Example: metallurgy

- The study is an experiment.
- Variables:
  - Treatment variable: swinging arm angle.
  - Blocking variable: material of the spike. metal A/B
  - Response variable: post-impact status of the spike: bent or shattered.
- Suppose we find:
  - The higher the swinging arm, the more often the spikes shattered.
  - Fewer metal A spikes shattered than metal B spikes.
- Answer the following:
  1. Does raising the swinging arm higher CAUSE more parts to be shattered?
  2. Is metal A better than metal B for making minimally brittle spikes?

Example: metallurgy

1. Yes: the experimenter controlled the swinging arm, and the level of the arm was not correlated with any experimental conditions.
2. Not in this case:
   - Metal A (titanium) is more brittle than metal B (steel).
   - The titanium spikes were made from a good manufacturer, and the steel spikes were made from a terrible manufacturer.

Notice:

1. one manufacture 2. collect randomly from different batches

- Metal type is not a treatment variable.
- Metal type was correlated with a **lurking variable**: the identity of the manufacturer.

Causality in studies

You can only infer causality when:

1. The prospective cause is a treatment variable.
2. If all possible predictors are uncorrelated with the treatment variable. (i.e., if you can't predict the treatment variable based on experimental conditions.)

That means:

- You cannot infer causality in an observational study.
- In an experiment, you can infer causality between the response and a treatment if:
  - All the experimental conditions are controlled, or:
  - The investigator randomly assigns sample units to treatment levels in a way that does not depend on experimental conditions.

Example: sales data

An analyst at Kmart looked at the last five years of sales data. He discovered that sales were high on days when the displays were mostly red and low on days when the displays were mostly blue.

1. Is this an experimental or observational study?
2. Does replacing blue with red in the displays cause sales to improve?

Example: sales data

1. This is an observational study.
2. No. Displays were red around Christmas time and blue otherwise. The spike in sales was caused by the holiday season, not the display color.

The display color:

- Was not a treatment variable.
- Was correlated with the schedule of holidays.

Example: rat data

A biologist studied the effect of a growth hormone on weight gain in rats. 600 rats were selected for the study. The rats varied greatly in:

- Age (Younger ones gain weight faster)
- Breed (Breed A grows faster than breed B.)
- origin (Either biology lab or the wilderness. Let's say lab rats grow faster.)

However, none of these factors were taken into account in the study. Instead, 300 rats were randomly selected from the original 600 to receive the hormone. The others received a placebo. At the end of three months, the rats with the hormone gained more weight than the rats without the hormone on average.

1. Is this an experiment or an observational study?
2. Does the hormone cause weight gain in rats?

Answers

1. Experimental study.
2. The hormone *does* cause weight gain in rats.
   - The hormone was a treatment.
   - Age, breed, and origin, and all other experimental conditions are *constant on average* with hormone level. That's because we randomly selected rats to receive the hormone
   - In addition, since the treatment groups are large, the conditions are *approximately constant* with hormone level.