MATH 10

INTRODUCTORY STATISTICS

Ramesh Yapalparvi

Week 2

- **Chapter 4 – Bivariate Data**
- Data with two/paired variables, Pearson correlation coefficient and its properties, general variance sum law

- **Chapter 6 – Research Design**
- Data collection, sampling bias, causation.

- **Chapter 5 – Probability** ← **today's lecture**

Probability, gambler's fallacy, permutations and combinations, binomial distribution, Bayes' theorem.

Chapter 4 – Bivariate Data, Quick Recap of Key Points

- **Paired Data** : $(\mathrm{X}_1, \mathrm{Y}_1), (\mathrm{X}_2, \mathrm{Y}_2)$,…, $(\mathrm{X}_\mathrm{n}, \mathrm{Y}_\mathrm{n})$, visualize with dot / scatter plot.
- The pairing will be given to you, or is “obvious”.
- (husband, wife), rolling 2 dice at the same time etc.

- Correlation coefficient r : quantifies linear relationships.
- Does not say anything about non-linear relationships.

- Make sure you can calculate r, by doing it at least once.
- Properties: symmetric (in the formula), unaffected by linear transforms.
- Formula: standard deviation cannot be zero for any of the variable.

Chapter 6 – Research Design, Quick Recap of Key Points

- Have hypothesis, collect data, "perform statistics". This tells us if hypothesis is likely to be correct. This might be evidence for (or against) your hypothesis.
- Very important in exam to **NOT SAY** that your hypothesis is "proven" or "true".

- **Sampling Bias** : we want you to recognize obvious sources of bias in the exam, and be able to explain why. Precise technical terms not required.

- E.g. of a 2 points exam question:
- *"You have a set of samples of heights in Hanover NH. You are told that your survey team only measured the heights of their friends, of which 80% are women."*
- *"Would the mean height in this sample be a good estimator of the mean heights in Hanover NH? Why or why not?"*

Chapter 6 – Research Design, Quick Recap of Key Points

- Section 7 – Causation. The 2 key things to know for the exam are…

1) Correlation does not imply causation.

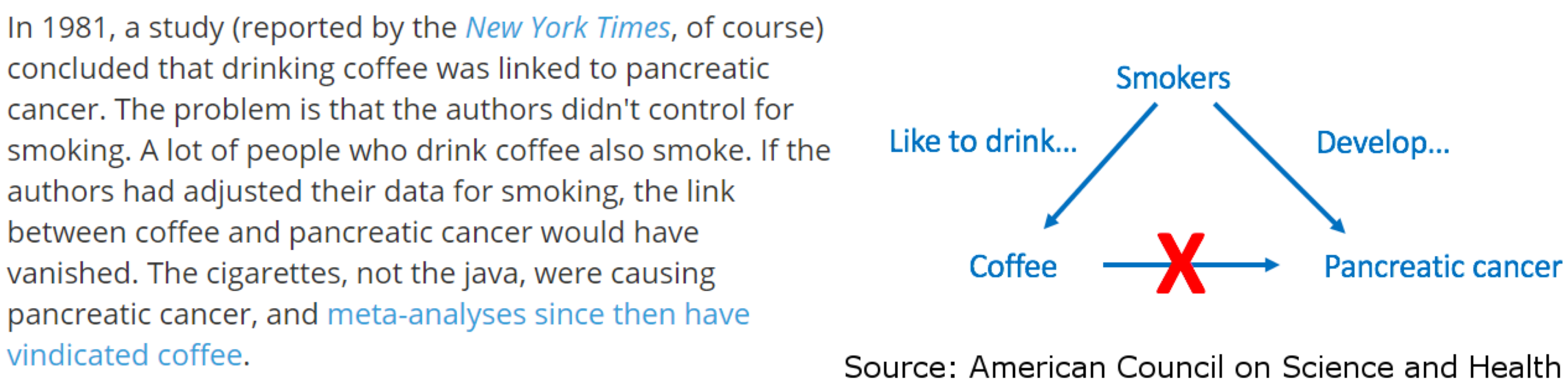
- High correlation coefficient r → potentially related/causal. Evidence that there might be an actual relation.
- Should have physical mechanism or reasonable theory to support this evidence.

- E.g. 2 points exam question:
- "*Your data shows that the annual per capita consumption of mozzarella cheese and the annual number of civil engineering PhD awarded has a correlation coefficient of r = 0.9586.*"
- *"Would you say that the high r indicates that mozzarella cheese consumption is related to the award of civil engineering PhD? Explain."*

Chapter 6 – Research Design, Quick Recap of Key Points

- **2) Confounding or "Third" Variable**
- Two specific scenarios: with respect to correlation coefficient r, and with respect to clinical trials or treatment/control groups.

- **With respect to the correlation coefficient r.**



In 1981, a study (reported by the *New York Times*, of course) concluded that drinking coffee was linked to pancreatic cancer. The problem is that the authors didn't control for smoking. A lot of people who drink coffee also smoke. If the authors had adjusted their data for smoking, the link between coffee and pancreatic cancer would have vanished. The cigarettes, not the java, were causing pancreatic cancer, and meta-analyses since then have vindicated coffee.

Source: American Council on Science and Health

Chapter 6 – Research Design, Quick Recap of Key Points

- **With respect to the correlation coefficient r.**

- Another example: miners exposed to asbestos tend to smoke and later develop lung cancer.

- E.g. exam question for 2 points.
- *"Your data showed that amount of ice cream sold and number of outdoor accidents that happened each day are highly correlated."*
- *" You later discover that in your data, ice cream sales and frequency of outdoor sports are higher on warmer days than colder days."*
- *"Is it correct to conclude that ice cream causes outdoor accidents? Explain."*

Chapter 6 – Research Design, Quick Recap of Key Points

- **With respect to clinical trial or treatment/control group case.**
- Get a group of healthy people (sample A), feed them garlic, observe that most of them got lung cancer. Turns out sample A are all heavy smokers.
- Get a group of people with disease (sample B), feed them green tea, observe that most of them got cured of their disease. You later find out that sample B are all on a medication for the disease.

- In both cases, the fix is to randomly assign them to a treatment and a control group, so that you can isolate the effect of what you're feeding them.
- E.g. 4 points exam question: "*Could you conclude that green tea can cure the disease? Why or why not? Without increasing the number of people in sample B, what is one change you can make to improve this experiment?*"

Chapter 5, Section 2 – Introduction to Probability

- **Descriptive statistics** : describes the data, sometimes visually.
- **Inferential statistics** : make intelligent guesses using data (and mathematics).

A key idea in statistics is forming a testable hypothesis/conjecture. Collecting data to test your hypothesis. Then, calculating the **chance or probability** of your hypothesis being true given this data.

How do we define chance or probability? What does it mean for an event A to have "higher chance" of occurring than a event B?

This is a very complex mathematical and philosophical issue.

Chapter 5, Section 3 – Basic Concepts in Probability

- Thankfully, for this course, we are going to stick to a simple model of chance.

- **Frequentist Probability**
- Define a sample space = set of possible outcomes of an experiment.
- Event = subset of the sample space.
- Probability of the event happening can be approximated by repeating the experiment a large n number of times, counting the number of occurrences x of the event, and calculating the proportion $\frac{x}{n}$.

- E.g. *Rolling a die, sample space = {1,2,3,4,5,6}, event = {1,6}. Roll the die n = 6,000 times. Get a 1 or 6 in 2,052 of those rolls. Probability is approximately* $\frac{2052}{6000} = 0.342$*.*

Chapter 5, Section 3 – Basic Concepts in Probability

- In the frequentist model of probability, we assume that if we can repeat the experiment an infinite number of times, we will get the "true" probability of the event happening.
- The more we repeat the experiments (the larger n is), the more confident we are that $\frac{x}{n}$ is "close" to the "true" probability.
- E.g. *If we flip a fair coin 1,000 million times, we are more confident that 50% of those flips will be heads, than if we flip it just 10 times.*
- It is important to note that this applies to the proportion of heads and not the frequency of heads.
- In fact, as the number of coin flips n becomes large, the difference (number of heads minus number of tails) will often be large.

Chapter 5, Section 2 – The Idea of Symmetrical Outcomes.

- Using the idea of symmetrical outcomes, we can argue what the “true” probability must be.
- Assume the number of experiments n is large. To approximate the probability of an event occurring, we count the number of occurrences x.
- Suppose the experiment is rolling a fair die with six sides {1,2,3,4,5,6}.
- We can argue that because the die is fair, all six sides are “symmetric” and hence should have roughly the same number of occurrences for a large enough n.
- Hence, the true probability of getting one particular number is $\frac{1}{6}$.

- $\text{Probability} = \frac{Number\ of\ "favorable"\ outcome}{Numer\ of\ possible\ equally-likely\ outcomes}$

Chapter 5, Section 3 – Basic Concepts

- We write $P(A)$ for the (true) probability of event A happening.

- Since we defined probability in terms of proportions, $0 \leq P(A) \leq 1.$
- In an experiment/trial, either A happens or A does not happen,

  $P(A) = 1 - P(A).$

- $P(A\ or\ B) = P(A) + P(B) - P(A\ and\ B)$
- Event $A\ or\ B$ includes : "A occurs, B does not", "A does not, B occurs", "both occurs".
- This assumes A and B can both occur in one trial. i.e. They are not *mutually exclusive*.

- If A and B are independent, then $P(A\ and\ B) = P(A)P(B)$.

Chapter 5, Section 3 – Conditional Probability

- If A and B are independent, then $P(A\ and\ B) = P(A)P(B)$.

- If A and B are **<u>not</u>** independent, then $P(A\ and\ B) = P(A)P(A|B)$.

- $P(A|B) =$ probability of A given that B has occurred.
- "Probability of A conditional on B occurring".
- The fact that B has occurred changes the probability since not independent.
- E.g. *52 card deck, P(ace on 2nd draw | ace on 1st draw)?*

BREAK TIME!   \o/

- 10 minutes break starts after I handed out the exercise and left the room.

- Questions 1 and 7 are not the kind of questions you will expect in a exam.

- In a job interview…maybe. (I am not even joking)

Chapter 5, Section 3 – Birthday Problem

- A common classroom example of how probability can differ from our intuition.
- If there are 25 people in a room, what is the probability that at least two of them share the same birthday?
- Important exam trick : if $P(A)$ is hard to calculate, instead calculate $1 - P(A)$.
- Here, $P(at\ least\ two\ same\ birthday)$ is going to be hard to calculate.
- So, calculate:

$$1 - P(at\ least\ two\ same\ birthday) = P(no\ one\ has\ the\ same\ birthday).$$

- $P(no\ one\ has\ the\ same\ birthday) = \frac{365}{365} \cdot \frac{364}{365} \cdot \frac{363}{365} \cdot \frac{362}{365} \cdots \frac{342}{365} \cdot \frac{341}{365} = 0.4313.$
- $P(at\ least\ two\ same\ birthday) = 1 - 0.4313 = 0.5687.$

Chapter 5, Section 3 – Gambler’s Fallacy

- Flip a fair coin 5 times.
- What is the probability of getting heads for the 5th flip given that you flipped tails 4 times in a row?

The hot hand fallacy, and the hot hand fallacy-fallacy.

- Flip a fair coin 4 times in a row. Repeat experiment many times and record the results.
- Pick a heads, that is not in the 4th place, randomly from the results.
- What is the probability that the following flip was also a heads?

"FALLACY"

How Researchers Discovered the Basketball ‘Hot Hand’

By Jesse Singal

It should come as no surprise that the U.S. men’s Olympic basketball team has some very talented shooters on it. But the best of them is probably Klay Thompson, the Golden State Warriors shooting guard known for suddenly catching fire — *swish*, *swish*, *swish* — and becoming seemingly unable to miss. The most spectacular manifestation of Thompson’s shooting abilities came two Januaries ago, in a game against the Sacramento Kings. That was when he dropped 37 points in a single quarter, smashing the previous record of

https://www.thecut.com/2016/08/how-researchers-discovered-the-basketball-hot-hand.html

Chapter 5, Section 6 – Permutations and Combinations

- Factorial

$$n! = 1 \times 2 \times \cdots \times (n-1) \times n$$
$$0! = 1$$

- Permutation

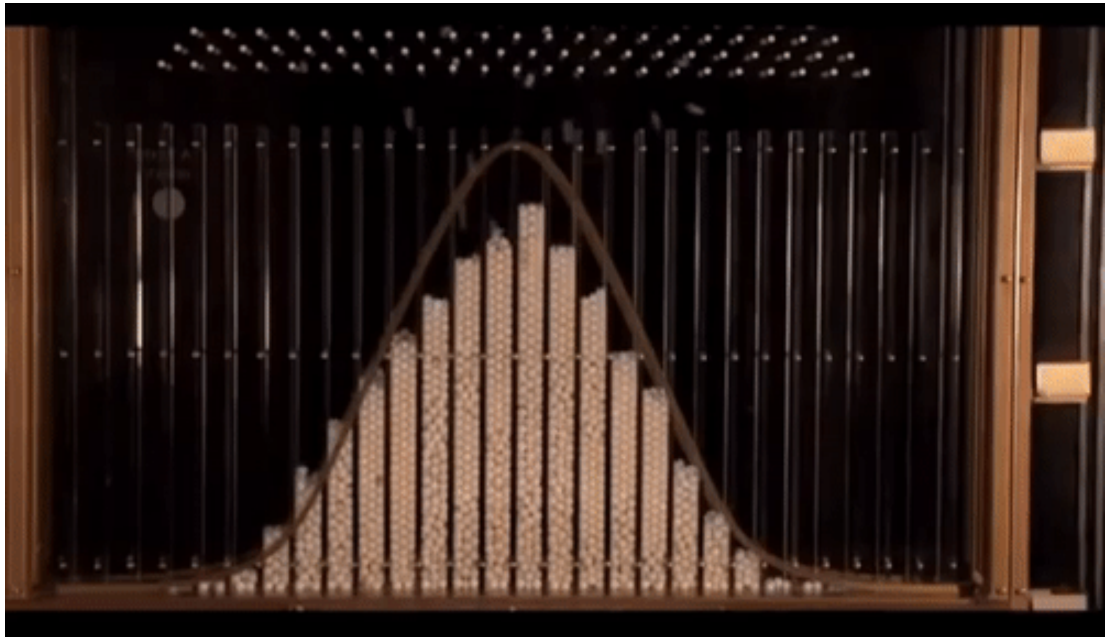
$$\mathrm{nPr} = \frac{n!}{(n-r)!}$$

- Combination

$$\mathrm{nCr} = \frac{n!}{r!\,(n-r)!}$$

Chapter 5, Section 8 – The Binomial Distribution

- n trials, with probability p of success. → we say n, p are parameters
- If you repeat the experiment: “n trials” many times, count the number of successes each time and plot a histogram of the results, you get a **binomial distribution**.



Galton board and video by Index Fund Advisors, Inc.

- P(k success in those n trials) =

$$\frac{n!}{k!(n-k)!}p^k(1-p)^{n-k}$$

- Combination: $nCk = \frac{n!}{k!(n-k)!}$

Chapter 5, Section 8 – The Binomial Distribution

- Some math associated with the binomial distribution.

- Cumulative probabilities.
- Textbook's example: toss a fair coin 12 times. What is the probability that we get between 0 to 3 heads?
- Answer: P(0 heads) + P(1 heads) + P(2 heads) + P(3 heads)
- You can add them up because these events are mutually exclusive!

- (Population) Mean and Variance

$$\mu = np \qquad \sigma^2 = np(1-p)$$

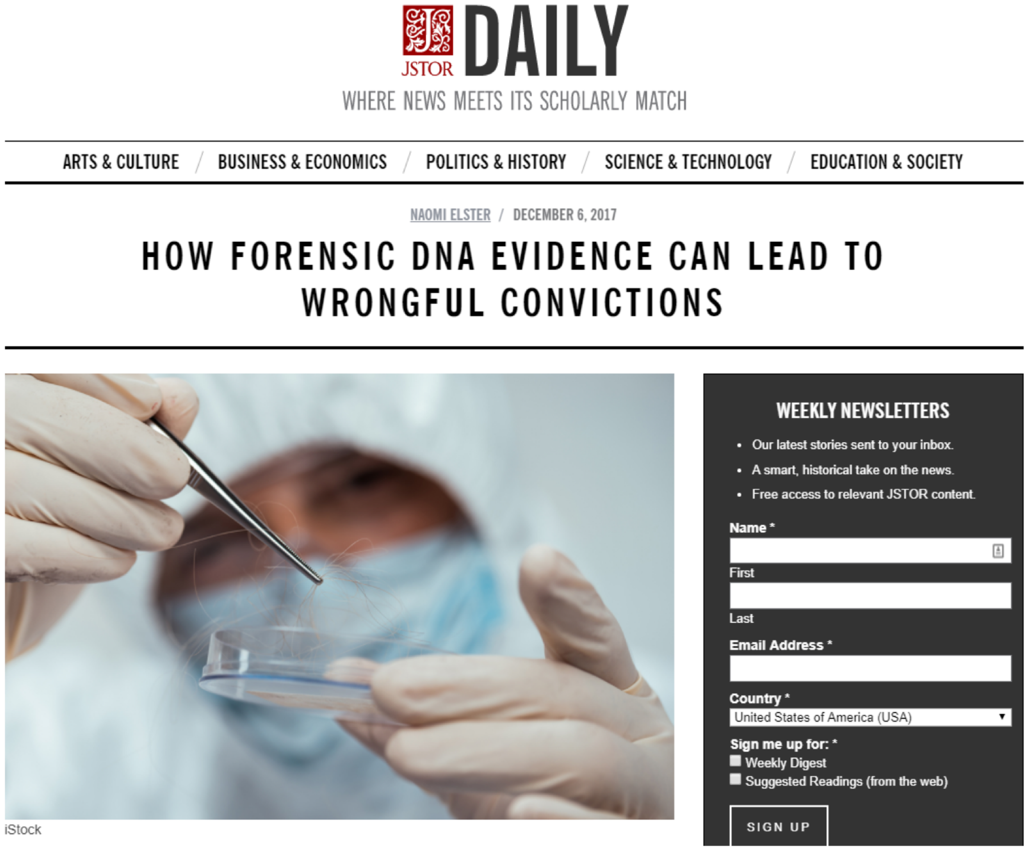
Public Service Announcement

- We are skipping these sections in Chapter 5.

- Section 10 – Poisson Distribution
- Section 11 – Multinomial Distribution
- Section 12 – Hypergeometric Distribution

Chapter 5, Section 13 – Base Rates

- Base rate = true proportion of a population having some condition, attribute or disease.
- The probability of positives that are false in tests depends heavily on the base rate.
- **Example**: a test for a disease is 99% accurate. You took the test and the result is positive. What is the probability that you actually have the disease?



JSTOR DAILY

WHERE NEWS MEETS ITS SCHOLARLY MATCH

ARTS & CULTURE / BUSINESS & ECONOMICS / POLITICS & HISTORY / SCIENCE & TECHNOLOGY / EDUCATION & SOCIETY

NAOMI ELSTER / DECEMBER 6, 2017

HOW FORENSIC DNA EVIDENCE CAN LEAD TO WRONGFUL CONVICTIONS

iStock

WEEKLY NEWSLETTERS

- Our latest stories sent to your inbox.
- A smart, historical take on the news.
- Free access to relevant JSTOR content.

Name *

First

Last

Email Address *

Country *

United States of America (USA)

Sign me up for: *

□ Weekly Digest

□ Suggested Readings (from the web)

SIGN UP

Chapter 5, Section 13 – Base Rates

- **Example**: a test for a disease is 99% accurate. You took the test and the result is positive. What is the probability that you actually have the disease?

- The answer depends on the *base rate*, which is the proportion of people having the disease.
- Suppose 1 mil people are tested. 1% or 10,000 of them actually has the disease.
- Out of the 990,000 disease-free people, the test would produce 9,900 positives!
- Out of the 10,000 diseased people, the test would also produce 9,900 positives.

- Chance that a positive test result is correct = 50%!

Chapter 5, Section 13 – Base Rates

- **False Positive** : tests result positive but you don’t actually have the disease.

- **Bayes Theorem**

$$P(D|T) = \frac{P(T|D)P(D)}{P(T)} = \frac{P(T|D)P(D)}{P(T|D)P(D) + P(T|ND)P(ND)}$$

- D = have disease. ND = no disease. T = tested positive for disease.