Lecture 5

Probability

History of Data Science, Winter 2022 @ UC San Diego
Suraj Rampure

Announcements

- Remember: classes are now in-person! Come to class in-person (Center Hall 218) OR via Zoom (link on the course website).
  - Office hours right after lecture will be in the lecture room (Center Hall 218) + Zoom as well.
  - Friday office hours (3:30-4:30PM) will be **remote only**.
- Homework 5 will be released by tomorrow, and will be due **Sunday, February 13th at 11:59PM**.
- **Make sure to read homework solutions (posted on Campuswire)!**

→ up until Homework 4

Agenda

- Brief recap: Galton's development of regression.
- Origins of probability.
- The Lady Tasting Tea: An early hypothesis test by Fisher.
- Gauss' development of the Normal distribution.

Galton

Heights

- One trait Galton was interested in studying was the difference in heights between parents and their children.
- He defined a new quantity, "midparent height", as being the average of a child's mother's and father's heights, after the mother's height was multiplied by 1.08.
  - He also multiplied the heights of daughters by 1.08.
- After collecting data, he estimated that the **correlation** between the **deviations of midparent heights** and the **deviations of child heights** was $\frac{2}{3}$.

Pearson's r = mean of product of x and y, when both are in standard units

TABLE I.

NUMBER OF ADULT CHILDREN OF VARIOUS STATURES BORN OF 205 MID-PARENTS OF VARIOUS STATURES.

(All Female heights have been multiplied by 1·08).

Heights of the Mid-parents in inches.	Heights of the Adult Children.														Total Number of		Medians.
	Below	62·2	63·2	64·2	65·2	66·2	67·2	68·2	69·2	70·2	71·2	72·2	73·2	Above	Adult Children.	Mid-parents.	
Above ..	..	..	..	..	..	..	..	..	..	..	..	1	3	..	4	5	..
72·5	..	..	..	..	..	..	..	1	2	1	2	7	2	4	19	6	72·2
71·5	..	..	..	..	1	3	4	3	5	10	4	9	2	2	43	11	69·9
70·5	1	..	1	..	1	1	3	12	18	14	7	4	3	3	68	22	69·5
69·5	..	..	1	16	4	17	27	20	33	25	20	11	4	5	183	41	68·9
68·5	1	..	7	11	16	25	31	34	48	21	18	4	3	..	219	49	68·2
67·5	..	3	5	14	15	36	38	28	38	19	11	4	..	..	211	33	67·6
66·5	..	3	3	5	2	17	17	14	13	4	..	..	..	..	78	20	67·2
65·5	1	..	9	5	7	11	11	7	7	5	2	1	..	..	66	12	66·7
64·5	1	1	4	4	1	5	5	..	2	..	..	..	..	..	23	5	65·8
Below ..	1	..	2	4	1	2	2	1	1	..	..	..	..	..	14	1	..
Totals ..	5	7	32	59	48	117	138	120	167	99	64	41	17	14	928	205	..
Medians ..	..	..	66·3	67·8	67·9	67·7	67·9	68·3	68·5	69·0	69·0	70·0	..	..	..	..	..

NOTE.—In calculating the Medians, the entries have been taken as referring to the middle of the squares in which they stand. The reason why the headings run 62·2, 63·2, &c., instead of 62·5, 63·5, &c., is that the observations are unequally distributed between 62 and 63, 63 and 64, &c., there being a strong bias in favour of integral inches. After careful consideration, I concluded that the headings, as adopted, best satisfied the conditions. This inequality was not apparent in the case of the Mid-parents.

and breadth (0·45). The concluding passage of the memoir is worth citing for its historical interest:—"The prominent "characteristics of any two correlated variables, so far at least as "I have as yet tested them, are four in number. It is supposed "that their respective measures have been first transmuted into "others of which the unit is in each case equal to the probable "error of a single measure in its own series. Let y = the deviation "of the subject, whichever of the two variables may be taken in "that capacity; and let x_1, x_2, x_3, &c., be the corresponding "deviations of the relative, and let the mean of these be X. Then "we find (1) that $y = r \cdot X$ for all values of y, (2) that r is the same "whichever of the two variables is taken for the subject, (3) that r "is always less than 1, (4) that r measures the closeness of the "co-relation." Galton determined r by a simple graphic method,

$y_{(su)} = r \cdot x_{(su)}$

$y = mx + b$

$m = r \cdot \frac{SD\,y}{SD\,x}$

$b = \bar{y} - m \cdot \bar{x}$

&c = et cetera = - - - - -

TABLE OF DATA FOR CALCULATING TABLES OF DISTRIBUTION OF STATURE AMONG THE KINSMEN OF PERSONS WHOSE STATURE IS KNOWN.

From group of persons of the same Stature, to their Kinsmen in various near degrees.	Mean regression=w.	$Q = f = p \times \sqrt{(1 - w^2)}$.
Mid-parents to Sons	2/3	1·27
Brothers to Brothers	2/3	1·27
Fathers or Sons to Sons or Fathers	1/3	1·60
Uncles or Nephews to Nephews or Uncles	2/9	1·66
Grandsons to Grandparents...	1/9	Practically that of Population, or 1·7 inch.
Cousins to Cousins	2/27	

Regression to the mean

$y_{(su)} = r \cdot x_{(su)}$

pretend $x_{(su)} = \frac{1}{2}$ ($\frac{1}{2}$ SD above mean)

predicted $y_{(su)} = r \cdot \frac{1}{2} = \frac{r}{2}$

Since $|r| < 1$,

$|y_{(su)}| \leq |x_{(su)}|$

- The effect that Galton observed was that **children tended to have heights that were closer to average than their parents.**
  - Tall parents tended to have children that were still tall, but closer to the average child's height.
  - Short parents tended to have children that were still short, but closer to the average child's height.
  - The same effect holds true in the opposite direction – remember, the correlation coefficient is symmetric!
- He called this "**reversion** to the mean", and later "**regression** to the mean".
- **The presence of regression to the mean depends on random variability in the distributions from which observations are drawn.**

Galton's successor: Karl Pearson

- Karl Pearson (1857-1936), a British statistician, was one of Galton's disciples.
  - He was also a stauch eugenicist.
  - He further developed the theory of correlation, and defined the correlation coefficient as we know it now.
- He founded the world's first Statistics department, at University College London, in 1911.
  - Started as part of UCL's Eugenics department.
  - Fun fact: UCSD has the world's first Cognitive Science department!

Probability

Reflection

- By now, it should be clear that when it is said that one or two people “invented” a field, they were just among the more prominent individuals in defining the field – they did not create it from scratch.
  - Newton and Leibniz did not create calculus from scratch.
  - Gauss and Legendre did not create least squares from scratch.
- The same idea holds true with probability.
  - Pierre de Fermat and Blaise Pascal are known as the “founders” of probability, but they are far from being the only individuals to contribute to the field.

Mentions of chance in the Bible (Old Testament)

"By chance I happened to be on mount Gilbo'a (2 Samuel 1:6)."

"Now there happened to be there a worthless fellow (2 Samuel 20:1)."

"A certain man drew his bow at a venture and struck the King of Israel (1 Kings 22:34, 2 Chronicles 18:33)."

Source: https://arxiv.org/pdf/1802.09966.pdf

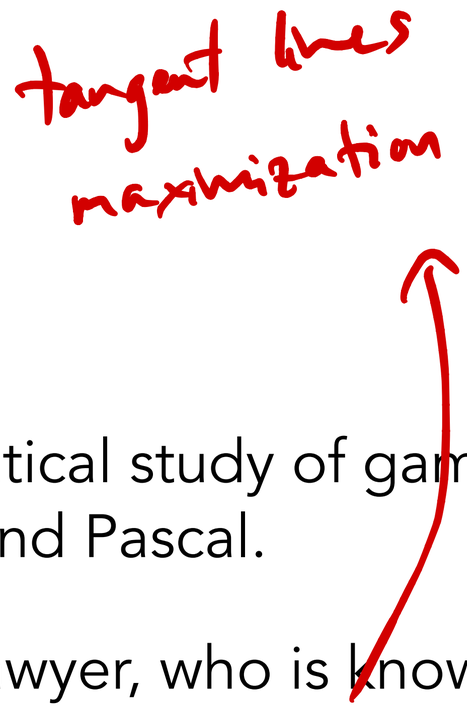
Origins of randomness

- There is evidence that humans relied on randomness thousands of years ago.
  - "Casting lots" refers to the act of generating some random result (e.g. rolling a die) and interpreting the result as being the will of God.
- Gambling as a past-time dates back thousands of years.[1]

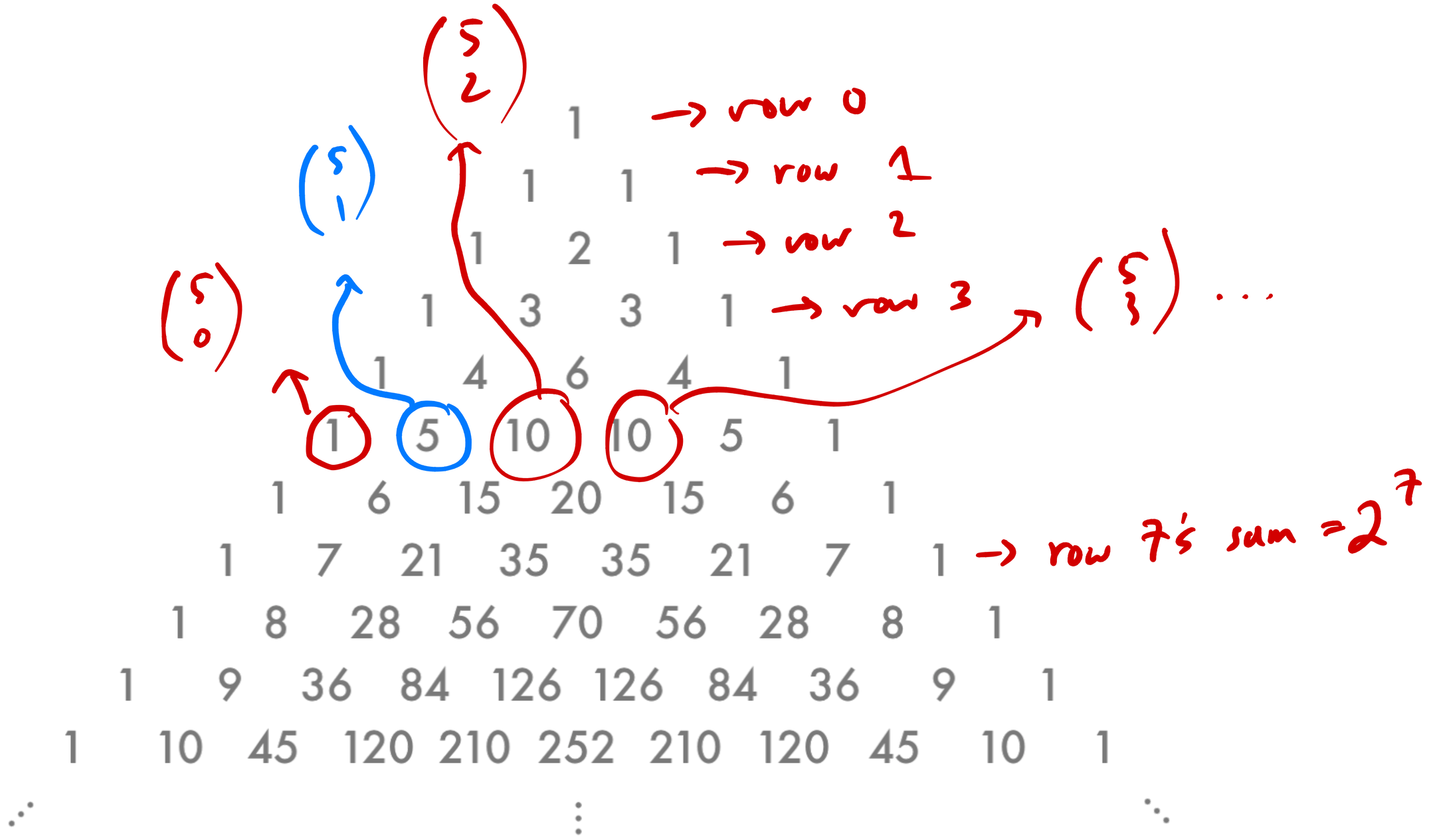
Early dice made of animal bones. (source)

1. https://www.britannica.com/topic/gambling/History

Fermat and Pascal



- While gambling is thousands of years old, the mathematical study of gambling is more recent, and catalyzed by a discussion by Fermat and Pascal.
  - Recall, **Pierre de Fermat** (1601-1665) was a French lawyer, who is known for several results in number theory but also developed the **method of adequality**, a precursor to calculus.
  - **Blaise Pascal** (1623-1662) was a French philosopher, physicist, and mathematician.
    - Among other things, he is known for Pascal's triangle (though he was not the first to discover it).


(5 2)
(5 1)
(5 0)
(5 3) ...
→ row 0
→ row 1
→ row 2
→ row 3
1
1 1
1 2 1
1 3 3 1
1 4 6 4 1
1 5 10 10 5 1
1 6 15 20 15 6 1
1 7 21 35 35 21 7 1
1 8 28 56 70 56 28 8 1
1 9 36 84 126 126 84 36 9 1
1 10 45 120 210 252 210 120 45 10 1
→ row 7's sum = 2^7

Source

Problem of Points

- Consider the following scheme, discussed by several Italian (Pacioli, Cardano, Tartaglia) and French (de Mere) mathematicians:
  - Player A and Player B decide to flip a fair coin repeatedly.
    - If heads, Player A earns a point. If tails, Player B earns a point.
    - **The first to earn 10 points wins a prize of $100.**
  - Now, suppose the players are forced to stop playing when Player A has 8 points and Player B has 7 points.
    - **Question: How should the prize money be divided amongst the two players?** 🤔

Problem of Points: possible solutions

doesn't make sense when score is 1-0

- Possible solutions:
  - The player who is closer to winning should take all of the money.
  - (Pacioli) The money should be allocated according to the **proportion of points won per player**.[1]
    - e.g. If A won 8 points and B won 7 points, A should take $\frac{8}{15}$ of the prize money.
  - (Tartaglia) Let r be the ratio of the two quantities defined below. The player closer to winning takes $\left(\frac{1}{2} + r\right)$ of the prize money, and the losing player takes the rest.
    - Numerator: The difference between the winning player's points and the losing player's points, and
    - Denominator: The target score (10 in our example).

$\frac{8-7}{10} = \frac{1}{10}$

① $\frac{1}{2} + \frac{8}{10} > 1$

② treats 2-1 as equal to 9-8

1. http://math.ucdenver.edu/~wcherowi/courses/history2/PrblmOfPoints.pdf

Fermat and Pascal's correspondence

to get to 9–9, A needs 1, B needs 2, then add 1 to win = 4

- In a series of letters to one another, Fermat and Pascal thought of a new way of splitting up the prize money.
  - Suppose again that Player A has 8 points, Player B has 7 points, and they are playing until 10 points.
  - In this scenario, **no matter what**, the game would end after 4 more rounds if they were allowed to continue playing (it may have even ended sooner!).
  - **Fermat's Idea:** consider all possible outcomes for the remaining 4 rounds, and count the proportion of them in which Player A wins.
    - That is the chance that Player A would win if the game were to continue, and that is the proportion of the prize money Player A should take.

Player A: 8 points – **needs 2 points (heads) to win**

Player B: 7 points – **needs 3 points (tails) to win**

Target: 10 points

2 or more heads = A

1 or fewer heads = B

$\binom{4}{2} + \binom{4}{3} + \binom{4}{4} = 6 + 4 + 1 = 11$

HHHH	HTHH	THHH	TTHH
HHHT	HTHT	THHT	TTHT
HHTH	HTTH	THTH	TTTH
HHTT	HTTT	THTT	TTTT

$\frac{11}{16}$

$P(A \text{ wins}) = \frac{11}{16}$

$2^4 = 16$ possible outcomes

$P(B \text{ wins}) = \frac{5}{16}$

The strategy

$(a-1)+(b-1)+1$

$= a+b-1$

- If Player A is a points from winning and Player B is b points from winning, then at most, the game will last $a + b - 1$ more rounds.
- There are 2^{a+b-1} possible outcomes.
- How many result in a win for Player A?

Aside: combinatorics

Recall from Homework 4, $\binom{n}{k}$ represents the number of ways of selecting k "successes" from n trials.

"n choose k"

4 coin flips → how many of them have exactly 2 heads?

$$\binom{4}{2} = \frac{4!}{2!2!} = 6$$

HHTT
HTHT
THTH
TTHH
HTTH
THHT

at least 2 heads

$$\binom{4}{2} + \binom{4}{3} + \binom{4}{4}$$

2h 3h 4h

$$= 2^4 - \binom{4}{0} - \binom{4}{1}$$

Back to the strategy

Player A needs a or more heads to win

Player B needs a-1 or fewer heads to win

Total flips: a+b-1

# of flips where Player A wins = (# w/ a heads) + (# w/ a+1 heads) + (# w/ a+2 heads) + ... + (# w/ a+b-1 heads)

flips

heads

$$= \binom{a+b-1}{a} + \binom{a+b-1}{a+1} + \binom{a+b-1}{a+2} + \cdots + \binom{a+b-1}{a+b-1}$$

The result

- We've now determined, as Fermat and Pascal did, that if Player A needs a points to win and Player B needs b points to win, that the **probability** that Player A would win if the game continued is

$$P(A \text{ wins}) = \frac{\sum_{k=a}^{a+b-1} \binom{a+b-1}{k}}{2^{a+b-1}}$$

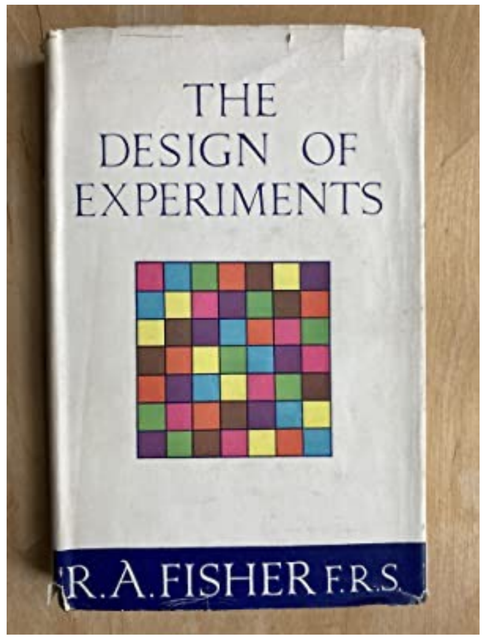
  - What made this result revolutionary is that it **inferred about what was likely to happen in the future**, rather than depending solely on the past.

  - **Let's implement this rule – and approximate it using a simulation – in a Jupyter Notebook.**

The Lady Tasting Tea

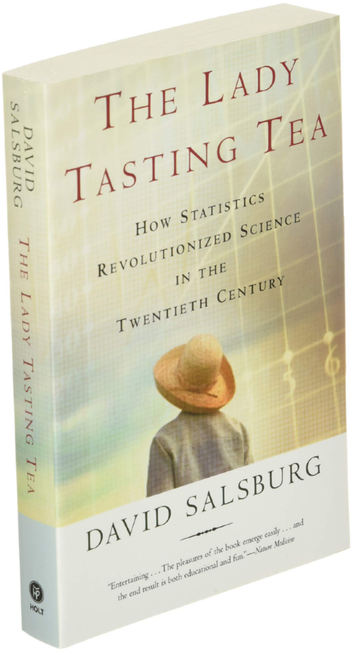
Fisher

- Sir Ronald Fisher (1890-1962) was a British statistician.
  - Was the "Galton Chair of Eugenics" at UCL, and was an editor of *The Annals of Eugenics*.
- He is credited for popularizing the idea of a **p-value**, and in particular the **5% cutoff for statistical significance.**
- One of the earliest documented examples of the use of a null hypothesis and a p-value comes in a story by Fisher known as "The Lady Tasting Tea", which he published as part of his book "The Design of Experiments."



The Lady Tasting Tea

- The story goes as follows:
  - One afternoon (in Britain), a woman claimed to be able to **distinguish** between
    - Tea where the **milk was poured before** the tea (milk-first), and
    - Tea where the **tea was poured before** the milk (tea-first)
  - Fisher thought of an experiment, where he would
    - Prepare 8 cups of tea, 4 milk-first and 4 tea-first, and
    - Ask the woman to pick out the 4 that were milk-first.
  - **Issue:** Even if the woman is guessing, she may correctly identify some cups due to random chance.



The setup

- **Null Hypothesis:** The woman is guessing at random and can't actually tell the difference between milk-first and tea-first tea.
- There are 70 possible ways in which we can choose 4 cups of 8 to be milk-first.
- Using combinatorics, we can determine the probability that the woman correctly identifies 0, 1, 2, 3, or 4 milk-first teas, under the assumption that she is guessing at random.
- If we were to perform the experiment, we could then compute a **p-value**, and compare it to the cutoff 0.05.

M = milk first, T = tea first

Relevant probabilities

Where does 70 come from? 70 = # possible outcomes

$\binom{8}{4}$ → MTMTMMTT
TTMTMMTM
⋮

prob (all correct) $= \frac{1}{\binom{8}{4}} = \frac{1}{70} \approx 0.015$

prob (3 correct) $= \frac{\binom{4}{3}\binom{4}{3}}{70} = \frac{16}{70}$

pick 3 of 4 to be M

pick 3 of 4 to be T

e.g.	M	M	T	M	T	T	M	T
order of cups →	M	M	M	M	T	T	T	T
her guesses →	M1	M2	M3	M4	T1	T2	T3	T4

Relevant probabilities

$$p(\text{all correct}) = \frac{1}{70}$$

$$p(3 \text{ correct}) = \frac{16}{70}$$

$$p(2 \text{ correct}) = \frac{36}{70} = \frac{\binom{4}{2}\binom{4}{2}}{70}$$

$$p(1 \text{ correct}) = \frac{16}{70}$$

$$p(0 \text{ correct}) = \frac{1}{70}$$

$$p(\text{at least 3 correct}) = \frac{1+16}{70}$$

$$= \frac{17}{70} > 0.05$$

Conclusion

- If the woman is guessing at random, there is only a $\frac{1}{70} \approx 0.0143$ chance that she correctly identifies all 4 milk-first cups. This is below the 0.05 cutoff, so in this case we'd reject the null hypothesis.
- Similarly, if she is guessing at random, there is a $\frac{1+16}{70} = \frac{17}{70} \approx 0.2429$ chance that she correctly identifies 3 or more milk-first teas. This is above any reasonable cutoff, so we'd fail to reject the null.
  - The small sample size makes it hard to base conclusions on these results. (What if she can actually tell the difference, but just happened to make a mistake?)
- **The legend says that she correctly identified all 8 cups!** ☕

The “normal” distribution

Gauss and least squares

- **Recall from Lecture 4:** one of the key differences between the approaches to least squares by Gauss and Legendre was that Gauss linked the theory of least squares to probability theory.
- Specifically, he posed the least squares **model** where

$$y_i = a + bx_i + \epsilon_i$$

where ϵ_i is a **random variable** that follows the following **error distribution**:

$$\phi(x \mid \mu, \sigma) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}$$

- Where did this distribution come from? 🤔

↓ probability density function

Criteria for the Gaussian distribution

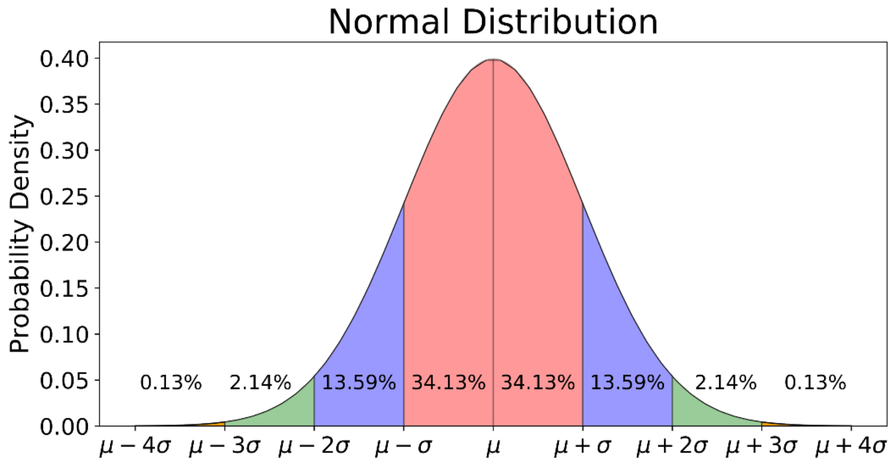
- Gauss described that the distribution of errors should satisfy three criteria:

    1. Small errors are more likely than large errors.

    2. For any real number ϵ the likelihood of errors of magnitude ϵ and $-\epsilon$ are equal.

    3. In the presence of several measurements of the same quantity, the most likely value of the quantity being measured is their average.

- We will now derive the Gaussian distribution using just these three criteria. (Buckle up!)



$\phi(x)$ "phi of x"

1. $\phi(x)$ is maximized when $x = 0$, and
2. $\phi(x) = \phi(-x)$

Aside: new function $f(x) = \dfrac{\phi'(x)}{\phi(x)}$

Property of f: $f(-x) = -f(x)$

use chain rule on numerator

$\phi'(-x) = -\phi'(x)$

(3) Suppose p is some fixed unknown number, and $M_1, M_2, \ldots, M_n$ are estimates of p.

Error: $M_1 - p,\ M_2 - p,\ \ldots,\ M_n - p$

$\frac{d}{dx} f(x)g(x)h(x)$

$= f'(x)\,g(x)\,h(x)$

$+ f(x)\,g'(x)\,h(x) + f(x)g(x)h'(x)$

Likelihood $L(p) = \phi(M_1 - p)\,\phi(M_2 - p)\,\phi(M_3 - p) \cdots \phi(M_n - p)$

Told: average of Ms maximizes $L(p)$

$\frac{d}{dp} L(p) = 0$ satisfied when $p = \overline{M} = \frac{M_1 + M_2 + \cdots + M_n}{n}$

$\frac{d}{dp} L(p) = -\phi'(M_1 - p)\,\phi(M_2 - p)\,\phi(M_3 - p) \cdots \phi(M_n - p)$

$- \phi(M_1 - p)\,\phi'(M_2 - p)\,\phi(M_3 - p) \cdots \phi(M_n - p)$

$- \cdots$

$$\frac{d}{dp} L(p) = \sum_{i=1}^{n} \frac{\phi'(M_i - p)}{\phi(M_i - p)} \cdot L(p)$$

$$= L(p) \sum_{i=1}^{n} f(M_i - p) = 0$$

satisfied when $p = \bar{M}$

$$\Rightarrow \sum_{i=1}^{n} f(M_i - \bar{M}) = 0$$

$$\vdots \quad \text{(see reading!)}$$

$$\Rightarrow f(x) = cx$$

$$a(x) = f(x)\, g(x)$$

$$a'(x) = f'(x) g(x) + f(x) g'(x)$$

$$= \frac{f'(x)}{f(x)} a(x) + \frac{g'(x)}{g(x)} a(x)$$

$$f(nx) = n f(x)$$

$$f(x) = cx$$

$$\frac{\phi'(x)}{\phi(x)} = cx$$

integrate both sides

$$\int \frac{\phi'(x)}{\phi(x)}\,dx = \int cx\,dx$$

$$\ln \phi(x) = \frac{c}{2}x^2 + D$$

$$e^{\ln \phi(x)} = e^{\frac{c}{2}x^2 + D}$$

$$\phi(x) = e^{\frac{c}{2}x^2 + D}$$

$$= A e^{\frac{c}{2}x^2}$$

① Note that $\frac{c}{2}$ must be negative

$$\int_{-\infty}^{\infty} e^{-x^2}\,dx = \sqrt{\pi}$$

Summary, next time

Summary, next time

- Fermat and Pascal's discussion regarding the Problem of Points helped establish the field of probability.
- Fisher's "The Lady Tasting Tea" is an early example of hypothesis testing and, in particular, calculating a p-value.
- Gauss derived the Gaussian distribution (later referred to by Quetelet as the Normal distribution) by using just three key properties.
- **Next three classes:**
  - Data visualization.
  - Computers.
  - Machine learning.