DATA ANALYSIS

Week 6: Probability

logistics: midterm 1+ problem sets

- conceptual
  - question regrade (income / skewed distribution / median and mode)
- computational
  - scores latest by Friday morning
  - statistics will be posted on Canvas
- problem sets
  - 2nd opt-out deadline: March 4 (Monday)
  - next problem set (PS4) due on March 11

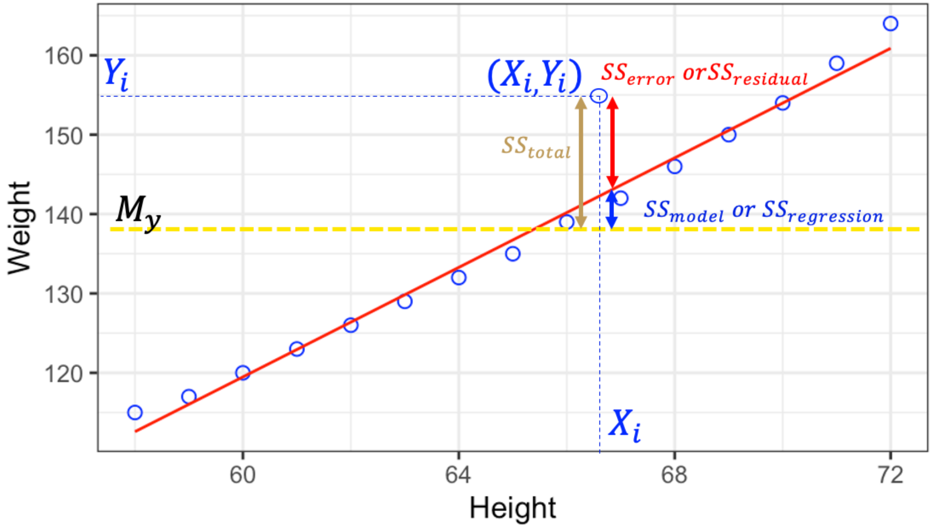
6	W: February 28, 2024	W6: Probability & Sampling
6	F: March 1, 2024	W6 continued…
7	M: March 4, 2024	**Problem Set Opt-out Deadline 2**
7	W: March 6, 2024	W7: Hypothesis Testing
7	F: March 8, 2024	W7 continued…
8	M: March 11, 2024	**Problem Set 4 due**
8	W: March 13, 2024	**Spring Break!**
8	F: March 15, 2024	**Spring Break!**
9	W: March 20, 2024	**Spring Break!**
9	F: March 22, 2024	**Spring Break!**
10	W: March 27, 2024	W10: Modeling Relationships I
10	F: March 29, 2024	W10 continued…

today's agenda

probability and inference

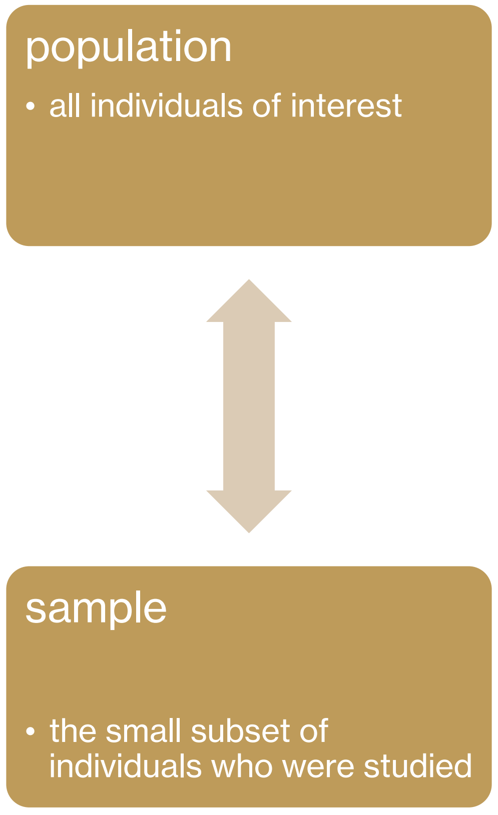
statistical thinking revisit

- data refers to a set of observations, typically on one variable (Y)
- the goal of statistics is to build good and simple models of data (Y)
  - data = model + error
- **level 0**: the simplest models can be built directly from the data
  - mean / median / mode
  - when all we have is Y, its mean is our best model
  - assessing the fit of the mean to the sample: $SS_{total} = \sum(Y - M_y)^2$
- **level 1**: using one more variable to understand Y
  - correlation / regression
  - we can fit a line that tries to explain variation in Y using its relationship to X
  - assessing the fit of the line to the sample: $SS_{error} = \sum(Y - \hat{Y})^2$



from samples to populations

- **level 0**: we obtain a mean for the sample
  - our sample statistic is the mean
  - how can we compare it to the population?
- **level 1**: we obtain a line of best fit for the sample
  - our sample statistic is the correlation (or slope)
  - how can we compare it to the population?
- we can start thinking about what our hypothesis is and what evidence have we collected that supports or contradicts the hypothesis

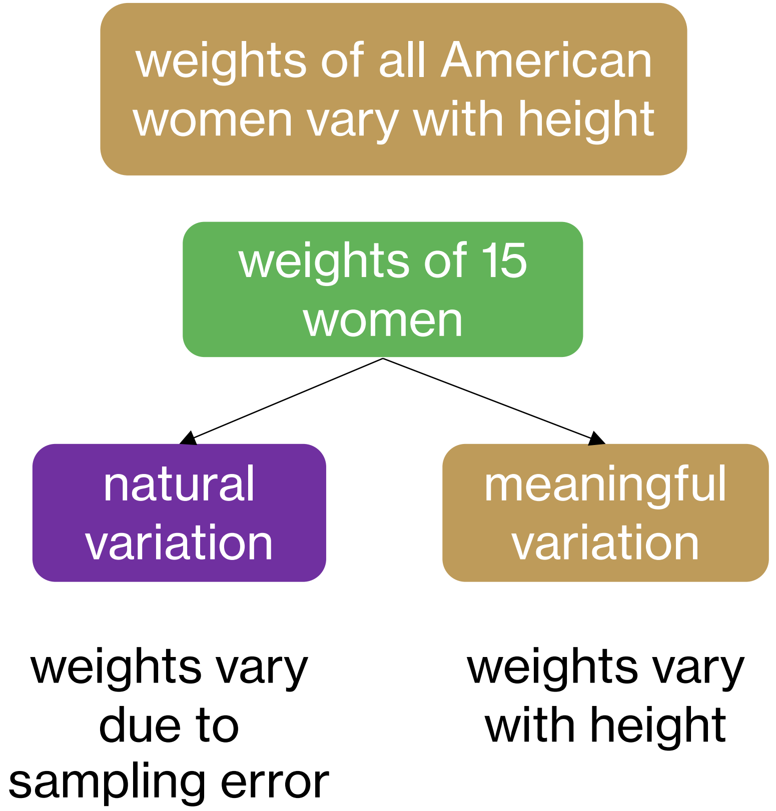


hypothesis testing: fundamentals

- research often begins with a hypothesis about the state of the world, i.e., the population
- we then collect a sample of data that may or may not be consistent with this hypothesis
  - samples differ from populations due to sampling error/natural variation OR meaningful variation that is consistent with the hypothesis
- our goal is to evaluate the likelihood of the hypothesis, given the sample statistic we have obtained, i.e., how likely is my hypothesis?
- P (your hypothesis, given the data sample)

  = P (your hypothesis | sample statistic)

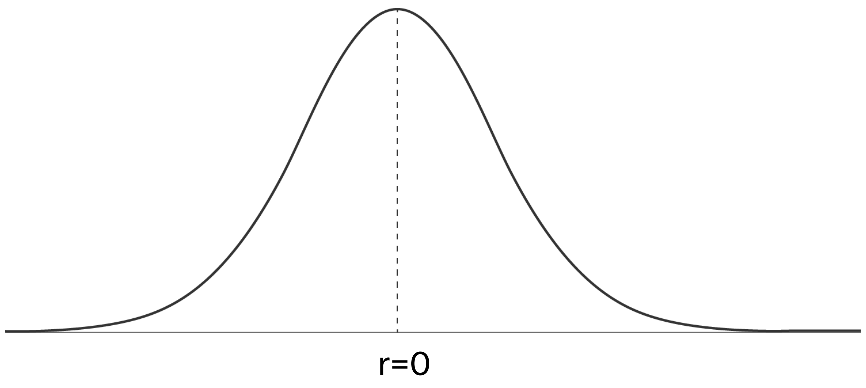
  = P (weights vary with height | r = 0.995)



from samples to populations

- we can start by assuming that our hypothesis is wrong
    - null hypothesis: there is no meaningful relationship between Y (weight) and X (height)
    - what would the true correlation be in this case?
    - population parameter, $\rho = 0$
- if we had a sense of what the sample statistic would look like each time we collected data from a sample of the same size, we could assess where our sample is relative to ALL possible samples from this population
- a sampling distribution: a distribution of the sample statistic for all possible samples of a given size

ALL sample correlations with sample size n when there is no meaningful relationship between height and weight in the population



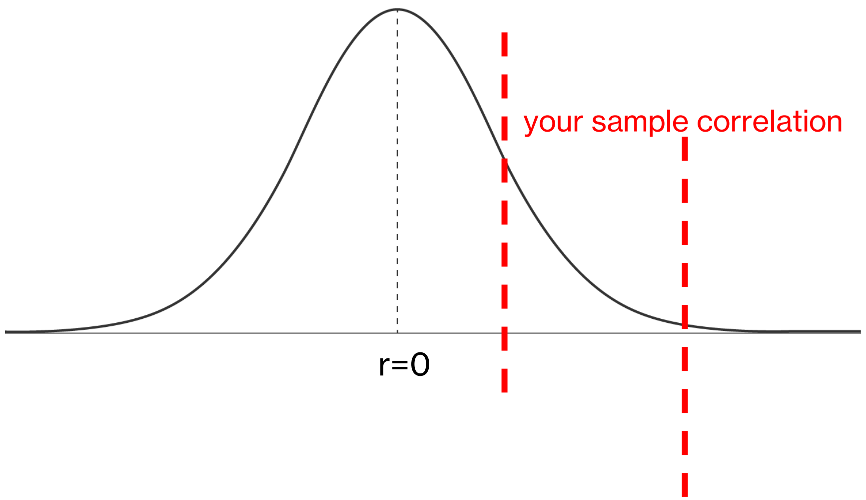
from samples to populations

- once we have a sampling distribution under the null hypothesis, we want to know how likely is the sample statistic you obtained
- P (your sample correlation | true correlation = 0)

  = P (your sample correlation | null hypothesis)

  = P (r= 0.995 | null hypothesis is true)
- if this probability is really low, we can infer that the null hypothesis may not be true, and subsequently infer that your actual hypothesis may be true!

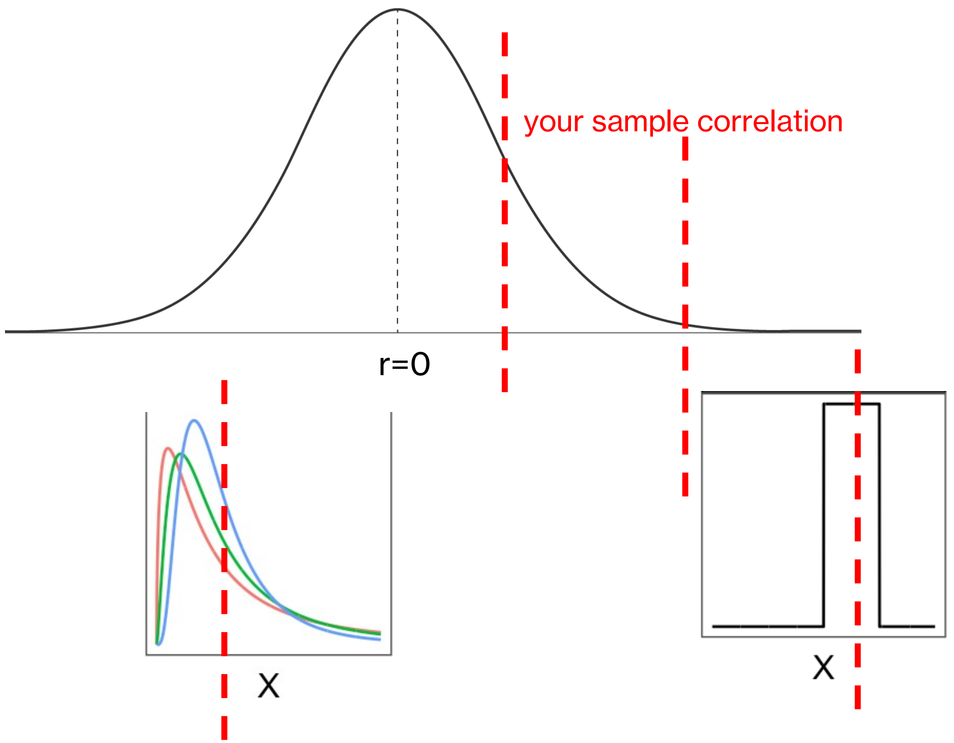
ALL sample correlations with sample size n when there is no meaningful relationship between height and weight in the population



three outstanding questions

- **question 1**: how do I calculate probabilities if I don't have access to ALL the scores?
- **question 2**: how do we know what the distribution of the null hypothesis looks like? If we don't know the form of the distribution, we cannot calculate probabilities
- **question 3**: how do we know whether the probability we obtained, i.e., P(data | null hypothesis) is small enough?

ALL sample correlations with sample size n when there is no meaningful relationship between height and weight in the population



today's agenda

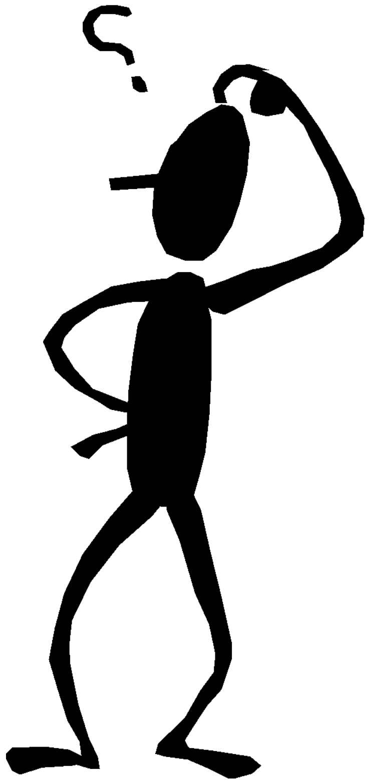
probability and inference

what is probability?

- a branch of mathematics that deals with uncertainty
- informally: a number that describes the likelihood of an event's occurrence
  - what is the probability that I will trip in class today?
  - what is the probability that it will rain today?
  - what is the probability that weight and height have a meaningful relationship?
- some properties:
  - probability of a single event cannot be negative or greater than 1
  - probabilities of all possible outcomes must sum to 1

how do we determine probabilities?

- personal belief: people use some type of estimation / intuition to estimate different types of probabilities
- empirical frequency: repeat the 'experiment' several times and count how often each event happens
  - *law of large numbers*: empirical probability will approach the true probability as the sample size increases
- classical probability: assumes certain rules for events and their occurrence to derive estimates



defining probability

- experiment: any activity that produces an outcome
- sample space: the set of possible outcomes
- event: a subset of the sample space
  - elementary events: single outcomes
  - complex events: one or more possible outcomes
- probability of an outcome A

$$p(A) = \frac{number\ of\ outcomes\ classified\ as\ A}{total\ number\ of\ possible\ outcomes}$$

examples

- coin toss
  - sample space = {heads, tails}
  - probability of getting heads = p(X = heads) = 1 / 2 = 0.5
- dice roll
  - sample space = {1,2,3,4,5,6}
  - probability of getting a 6 = p(X = 6) = 1/ 6 = 0.167
- card deck
  - sample space = {12 face cards (jack, queen, king), 40 other cards}
  - probability of getting a face card = p(X= face card) = 12/52

random sampling

- random sample: each outcome has an equal chance of being selected
- independent random sample: each outcome has an equal chance of being selected AND probability of being selected remains constant if multiple selections are made
- example: probability of drawing a jack of diamonds two times in a row
  - first draw = p (jack of diamonds) = 1/52
  - second draw = p (jack of diamonds) = 1/51 if the first card was NOT a jack of diamonds

    = 0 if the first card was a jack of diamonds
- sampling with replacement is critical here, i.e., putting back the first sample so that the probability of being selected remains constant on the second sample

independent events

- independent events: the occurrence or non-occurrence of one event has no effect on the occurrence or non-occurrence of the other
  - A happening or not doesn't affect B
  - the chance of you getting struck by lightning has no effect on whether or not it is a Monday
- multiplicative law of probability:
  - $p\ (A\ and\ B)$: joint probability of A and B happening
  - If A and B are independent, then $p\ (A\ and\ B) = p(A).\,p(B)$
  - If A and B are not independent, then $p\ (A\ and\ B) = p(A).\,p(B \mid A)$
  - $p(A)$: marginal probability of A happening
  - $p(B \mid A)$: conditional probability of B given A has already happened

independent events: example #1

- what is the probability of drawing an ace and a jack on two successive draws with replacement in a card deck?
- A = drawing an ace, B = drawing a jack
- if draws are with replacement, drawing and ace and a jack are independent events
- $p\ (ace\ and\ jack) = p(ace).\,p(jack)$
- $p\ (ace) = \frac{total\ \#\ of\ ace\ cards}{total\ \#\ of\ cards} = \frac{4}{52}$
- $p\ (jack) = \frac{total\ \#\ of\ jack\ cards}{total\ \#\ of\ cards} = \frac{4}{52}$
- $p\ (ace\ and\ jack) = \frac{4}{52}.\frac{4}{52} = .0059$

independent events: example #2

- what is the probability of drawing an ace and a jack on two successive draws **without** replacement in a card deck?
- A = drawing an ace, B = drawing a jack
- if draws are without replacement, drawing an ace and then a jack are NOT independent events
- $p\ (ace\ and\ jack) = p(ace).\, p(jack \mid ace)$
- $p\ (ace) = \frac{total\ \#\ of\ ace\ cards}{total\ \#\ of\ cards} = \frac{4}{52}$
- $p\ (jack \mid ace) = \frac{total\ \#\ of\ jack\ cards}{total\ \#\ of\ cards\ remaining} = \frac{4}{51}$
- $p\ (ace\ and\ jack) = \frac{4}{52} . \frac{4}{51} = \ .006$

mutually exclusive events

- mutually exclusive events: if the occurrence of one precludes the occurrence of the other
  - if A happened, B cannot have happened
  - if you got a head on a coin flip, you cannot get a tail on the same coin flip
- additive law of probability:
  - $p\ (A\ or\ B)$: A or B happening
  - If A and B are mutually exclusive, then $p\ (A\ or\ B) = \mathrm{p(A)} + \mathrm{p\ (B)}$
  - If A and B are not mutually exclusive, then $p\ (A\ or\ B) = \mathrm{p(A)} + \mathrm{p\ (B)} - \mathrm{p(A\ and\ B)}$

mutually exclusive events: example # 1

- what is the probability of drawing a 4 or 10 from a card deck?
- A = drawing a 4, B = drawing a 10
- if you draw a 4, you could not have also drawn a 10
- these events are mutually exclusive
- $p\ (draw\ 4\ or\ 10) = \mathrm{p(draw\ 4)} + \mathrm{p\ (draw\ 10)} = \frac{4}{52} + \frac{4}{52} = \ .154$

mutually exclusive events: example # 2

- what is the probability of drawing a 4 or spade from a card deck?
- A = drawing a 4, B = drawing a spade
- if you draw a 4, you could have ALSO drawn a spade (i.e., a 4 of spades!)
- these events are NOT mutually exclusive
- $p\ (draw\ 4\ or\ 10) = \mathrm{p(draw\ 4)} + \mathrm{p\ (draw\ spade)}\ - p(4\ and\ spade) = \frac{4}{52} + \frac{13}{52} - \frac{1}{52} = .308$
- Thus, it is much more likely to draw a 4 or spade (p = .308) than it is to draw a 4 or 10 (p=.154)

activity

- what is the probability of rolling a prime or an odd number on the same roll using a fair dice?

activity

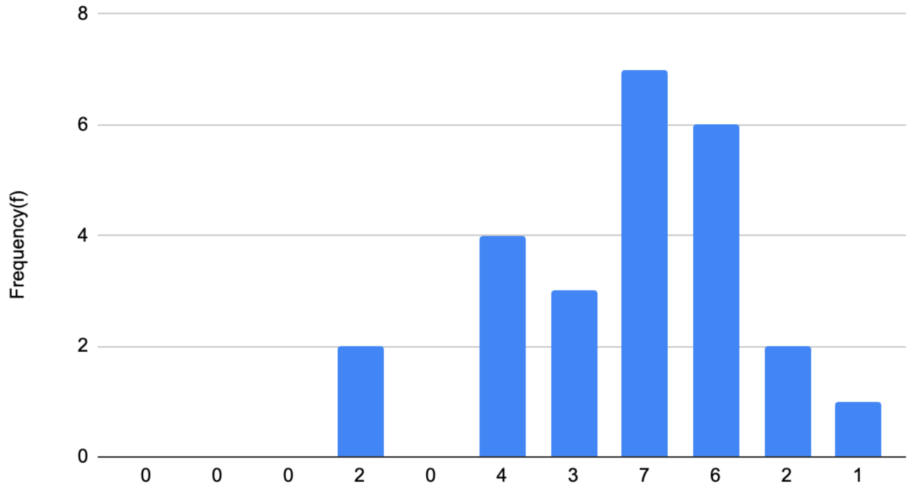
- what is the probability of rolling a prime or an odd number on the same roll using a fair dice?
- 1, 2, 3, 4, 5, 6
- A = rolling a prime number, B = rolling an odd number
- prime numbers: {2, 3, 5}, odd numbers = {1, 3, 5}
- if you roll a prime, you have ALSO rolled an odd number (i.e., 3 is an odd prime number!)
- these events are NOT mutually exclusive
- $p\ (roll\ prime\ or\ odd) = \mathrm{p}(\mathrm{roll}\ prime) + \mathrm{p}\ (\mathrm{roll\ odd})\ - p(prime\ and\ \mathrm{o}dd)$

$$= \frac{3}{6} + \frac{3}{6} - \frac{2}{6} = \frac{4}{6} = .667$$

probabilities from frequency tables

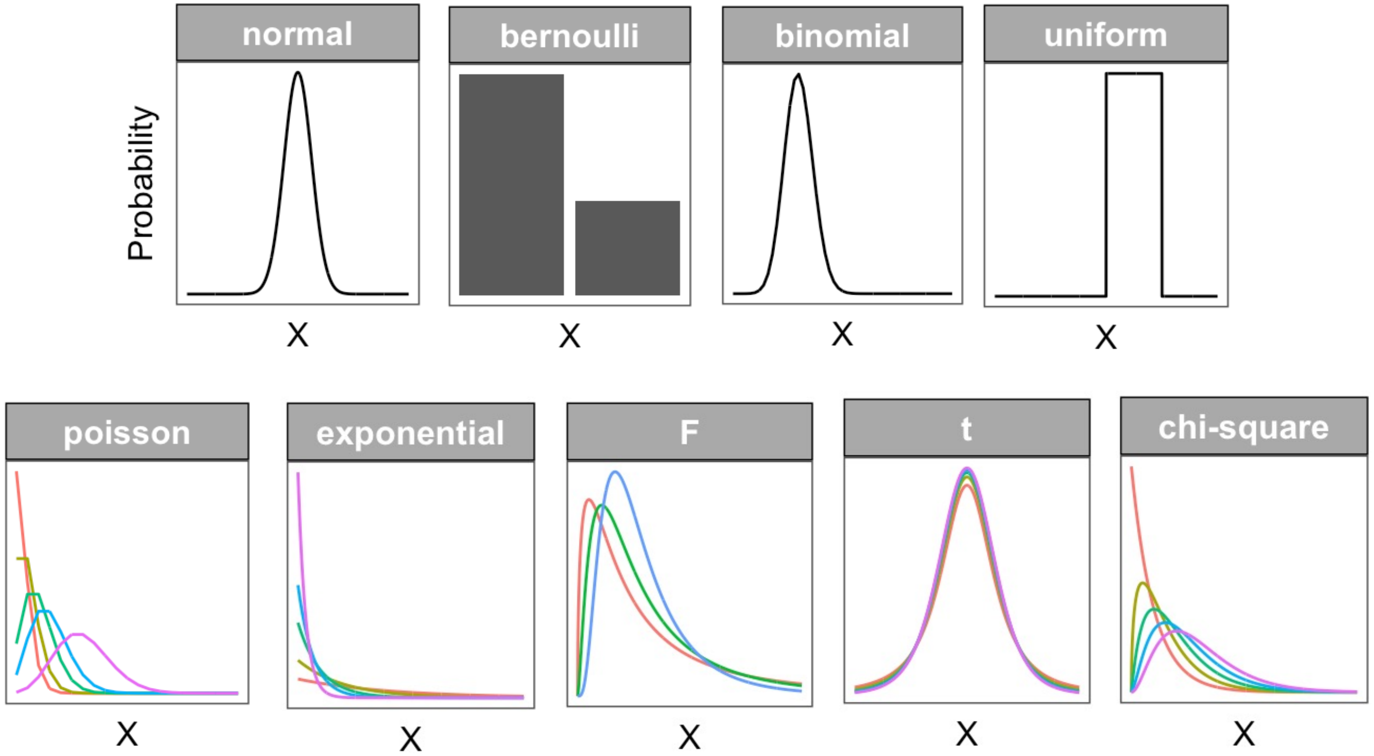
- probabilities can be obtained from frequency tables
- p = f / N = proportion
- probabilities and proportions are equivalent!

X	Frequency(f)	fX	proportion	percentage
0	0	0	0	0
1	0	0	0	0
2	0	0	0	0
3	2	6	0.08	8
4	0	0	0	0
5	4	20	0.16	16
6	3	18	0.12	12
7	7	49	0.28	28
8	6	48	0.24	24
9	2	18	0.08	8
10	1	10	0.04	4



probability distributions

- data come in many forms and distributions
- a probability distribution describes the probability of all of the possible outcomes in an experiment.
- which distributions have we seen already?

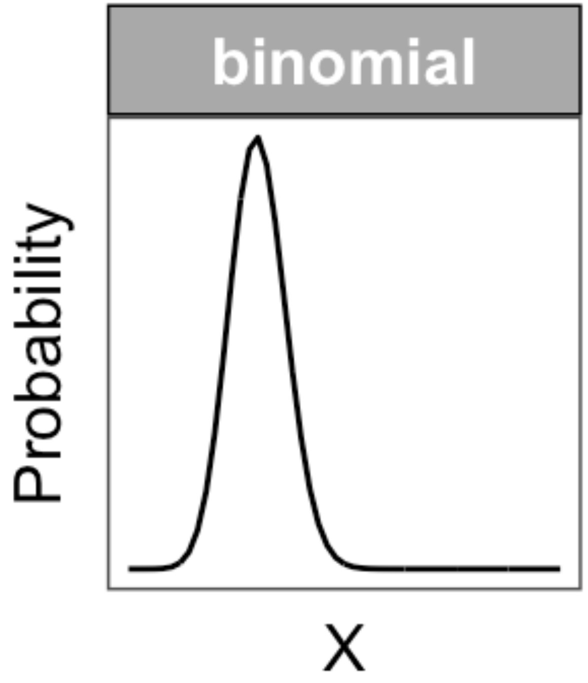


binomial distribution

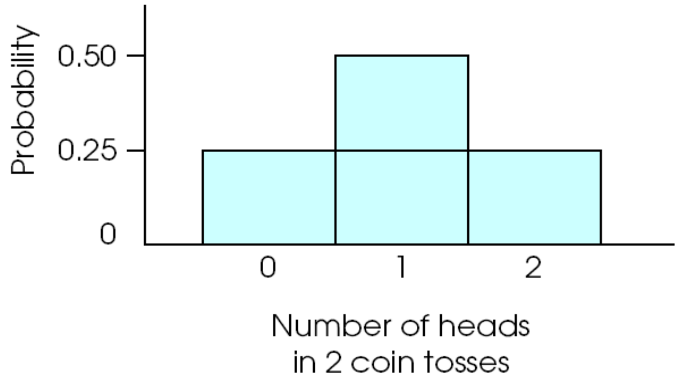
- data can only take two possible values (bi = two, nomial = names)
- a sequence of "bernoulli trials" (with only 2 possible outcomes)
- question of interest: how often does an outcome (A or B) occur in a sample of observations?

  $p = p(A)\ and\ q = p(B)$

  $p + q = 1$ i.e., $q = 1 - p(A)\ and\ p = 1 - p(B)$
- n : number of observations/individuals in the sample
- X: number of times that A occurs in the sample
  - X ranges between 0 and n
- the binomial distribution shows the probability associated with each X value from X=0 to X=n



example



- for two coin tosses, n = 2
- there are 4 possible outcomes (HH, HT, TH, TT)
- X = the number of times heads occurs
- X ranges from 0 to 2 (0 heads, 1 head, 2 heads)
- $p\ (X = 2\ heads) = \frac{1}{4} = 0.25$
- $p\ (X = 1\ head) = \frac{2}{4} = 0.50$
- $p\ (X = 0\ heads) = \frac{1}{4} = 0.25$
- $p\ (X = 0) + p\ (X = 1) + p\ (X = 2) = 1$

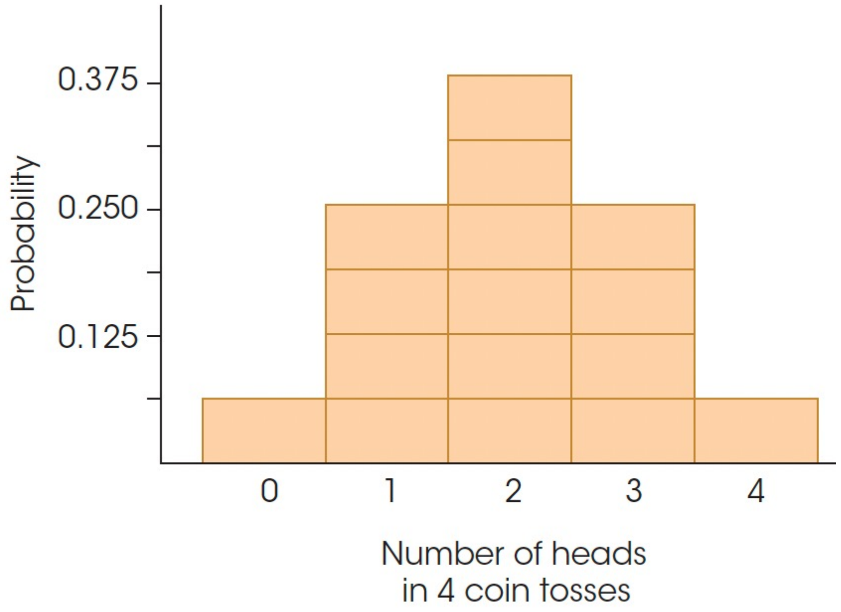
outcome	toss 1	toss 2
1	heads	heads
2	heads	tails
3	tails	heads
4	tails	tails

activity: 4 coin tosses

- what is the probability of obtaining 2 heads in 4 coin tosses?

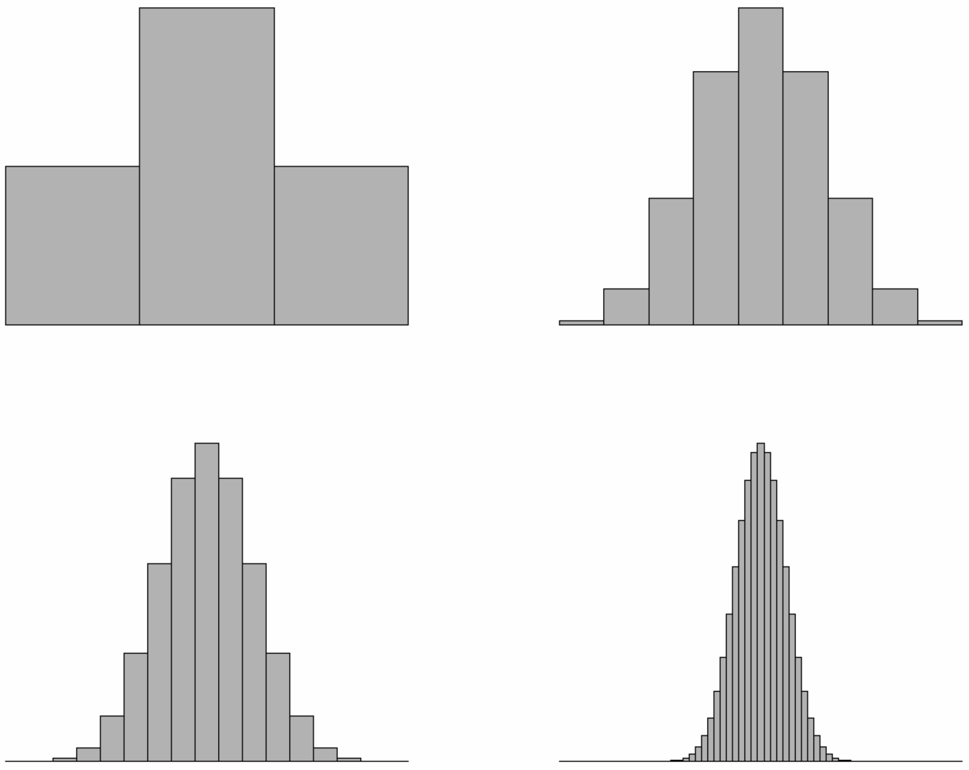
activity: 4 coin tosses

- what is the probability of obtaining 2 heads in 4 coin tosses?
- $p\ (X = 2\ heads)$
- 16 possible outcomes: HHHH, HHHT, HHTH, HTHH, THHH, HHTT, HTHT, HTTH, THTH, TTHH, THHT, HTTT, THTT, TTHT, TTTH, TTTT
- X ranges from 0 (no heads) to 4 (four heads)
- $p\ (X = 2\ heads) = \frac{6}{16} = 0.375$



increasing n...

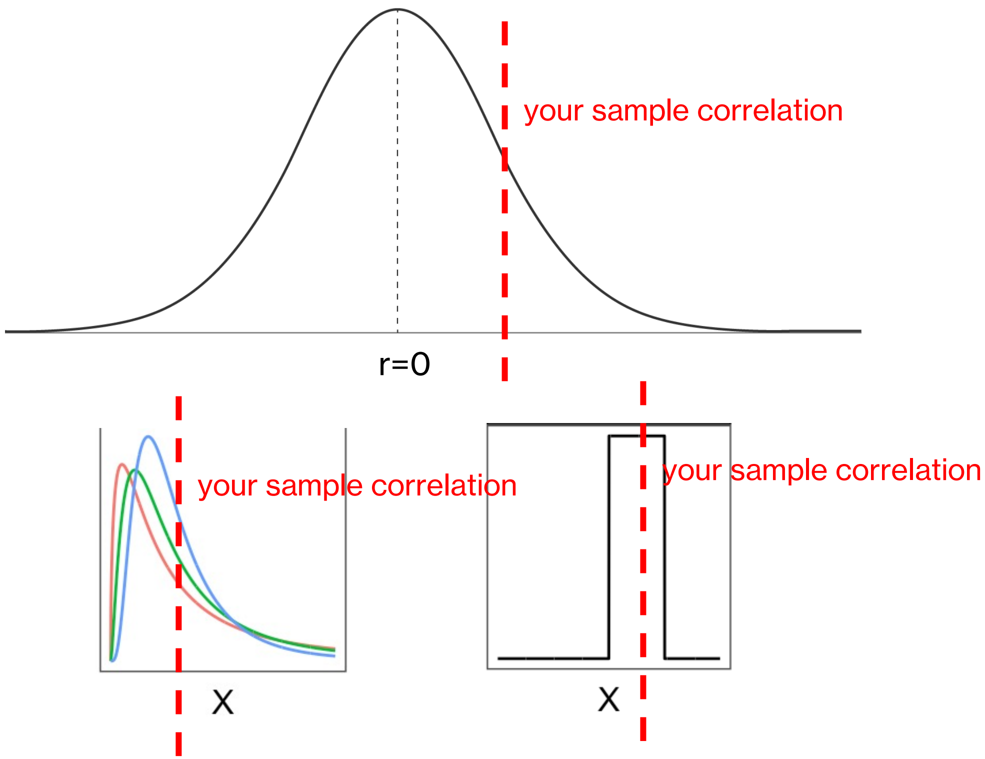
- play with the coin toss simulator
  - increase number of coin tosses (n)
  - simulate flips!
- as the number of coin tosses (n) increases, the distribution starts to resemble a normal distribution!
- rule of thumb: when pn and $qn \geq 10$, the binomial distribution approximates the normal distribution
  - mean: $\mu = pn$
  - standard deviation: $\sigma = \sqrt{npq}$
  - z-score: $z = \frac{X - \mu}{\sigma} = \frac{X - pn}{\sqrt{npq}}$



three outstanding questions

- **question 1**: once we have a sample, we can obtain probabilities, i.e., P (data | null hypothesis)
- **question 2**: how do we know what the distribution of the null hypothesis looks like? If we don't know the <u>form</u> of the distribution, we cannot calculate probabilities
- **question 3**: how do we know whether the probability we obtained, i.e., P(data | null hypothesis) is small enough?

ALL sample correlations with sample size n when there is no meaningful relationship between height and weight in the population



next time

- **before** class
  - *prep*: chapters 6 and 7 (specific sections)
  - *try*: PS4 (chapter 6 problems)
  - *apply*: optional meme
- **during** class
  - class survey discussion
  - distributions of sample statistics