DATA ANALYSIS

Week 4: Sampling

today's agenda

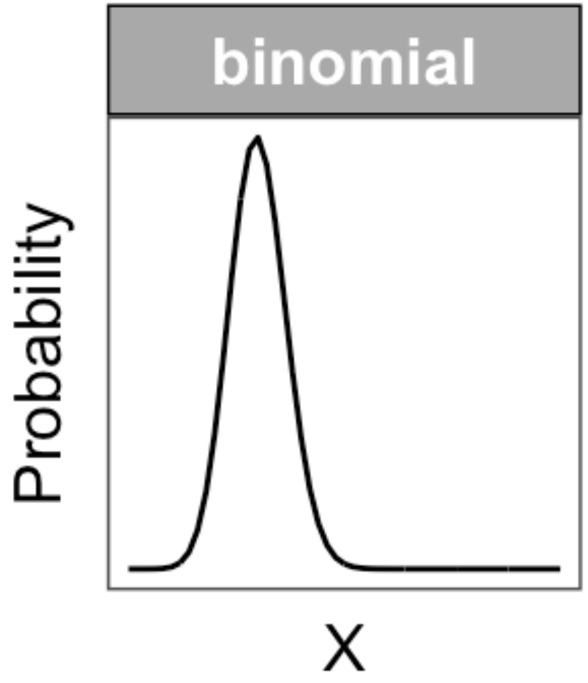
- probability and inference
- sampling
- class survey

recap: 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?

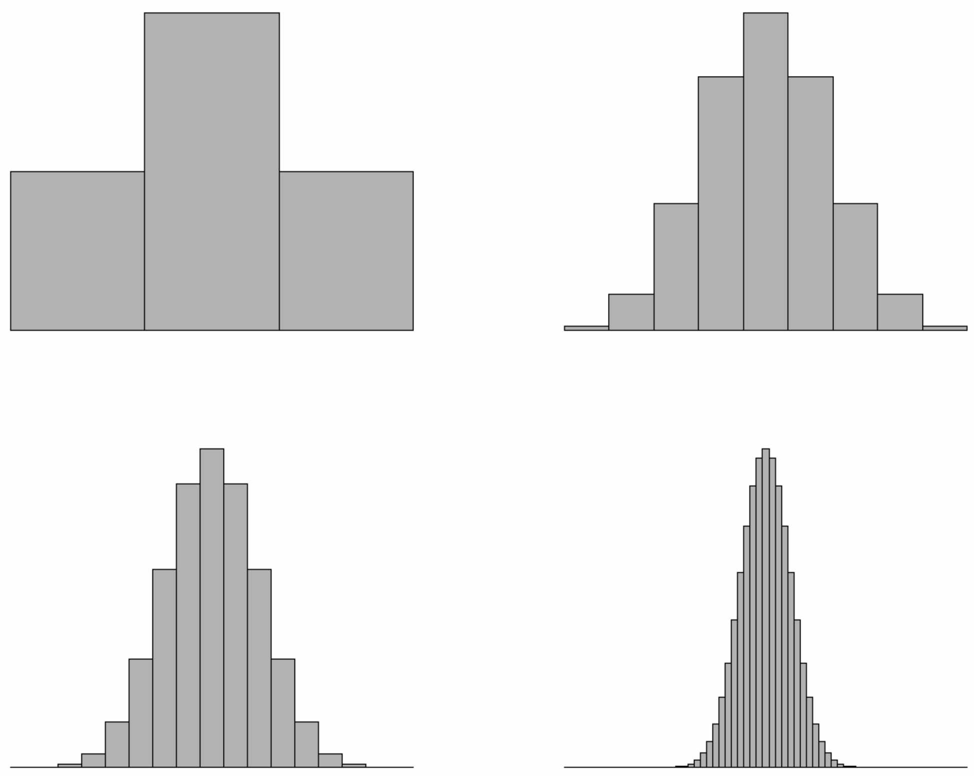
  $p = p(A)\ and\ q = p(B)$

  $p + q = 1$ 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



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 ≥10, the binomial distribution approximates the normal distribution
    - mean: $\mu = pn$
    - standard deviation: $\sigma = \sqrt{npq}$
    - z-score: $z = \frac{X - \mu}{\sigma} = \frac{X - pn}{\sqrt{npq}}$



example 1

- using a balanced coin, what is the probability of obtaining more than 30 heads in 50 tosses?
- balanced coin, i.e., *p* = *p(head)* = 0.5, *q* = p (tail) = 0.5
- n = 50, X = 30
- $\mu = pn = 0.5\ (50) = 25$, $qn = 0.5\ (50) = 25$, i.e., ≥ 10
- $\sigma = \sqrt{npq} = \sqrt{50\ (0.5)(0.5)} = 3.54$
- $\mathrm{z} = \frac{X-\mu}{\sigma} = \frac{30-25}{3.54} = 1.41$
- look up probability in [visual calculator]
- *p* (X > 30) = *more than* .0793

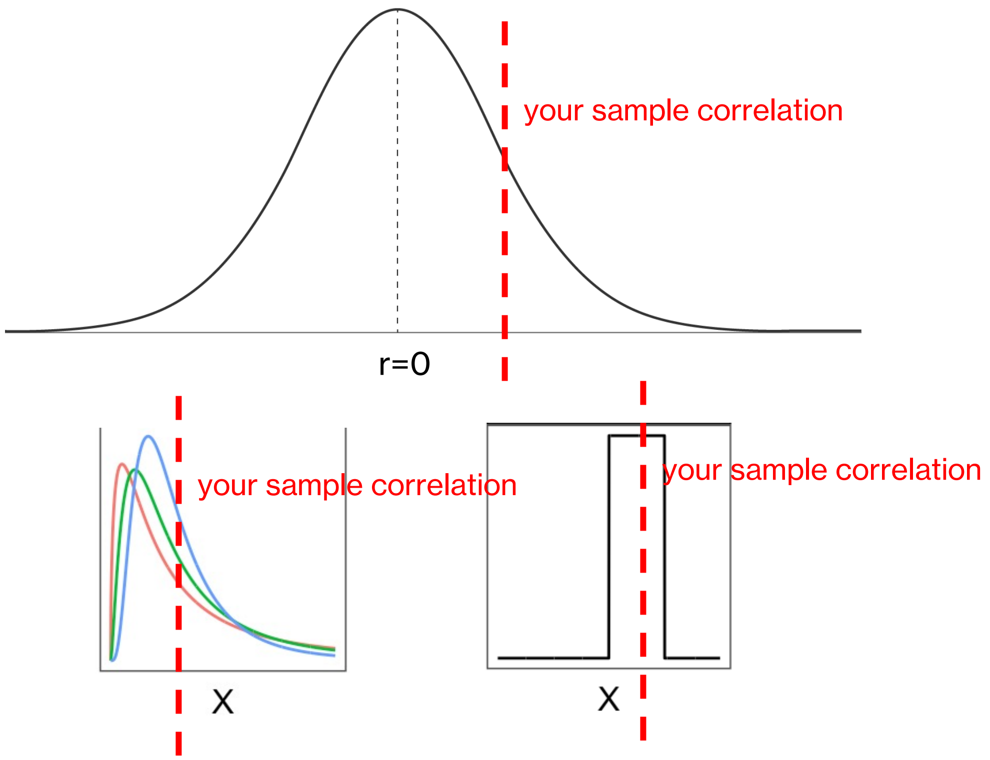
example 2

- a friend bets you that he can draw a king more than 8 times in 20 draws (with replacement) of a fair deck of cards, and he does it. Is this a likely outcome, or should you conclude that the deck is not “fair”?

three outstanding questions

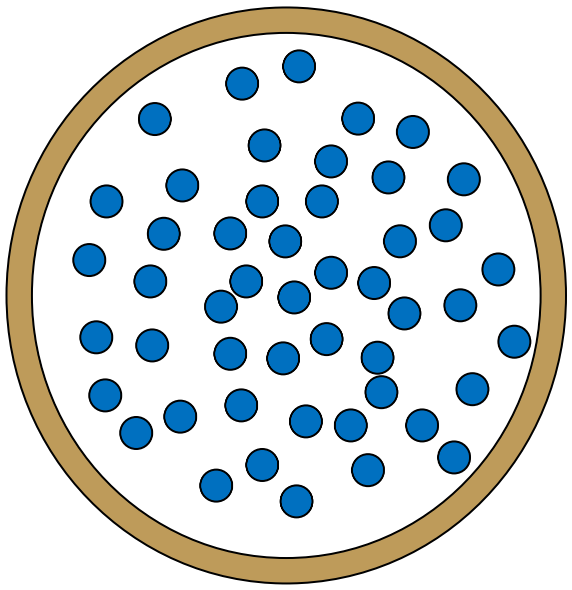
- **<u>question 1</u>**: how do I calculate probabilities if I don't have access to ALL the scores?
- **<u>question 2</u>**: how do we know what the distribution of the null hypothesis looks like? If we don't know the <u>form</u> of the distribution, we cannot calculate probabilities
- **<u>question 3</u>**: 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



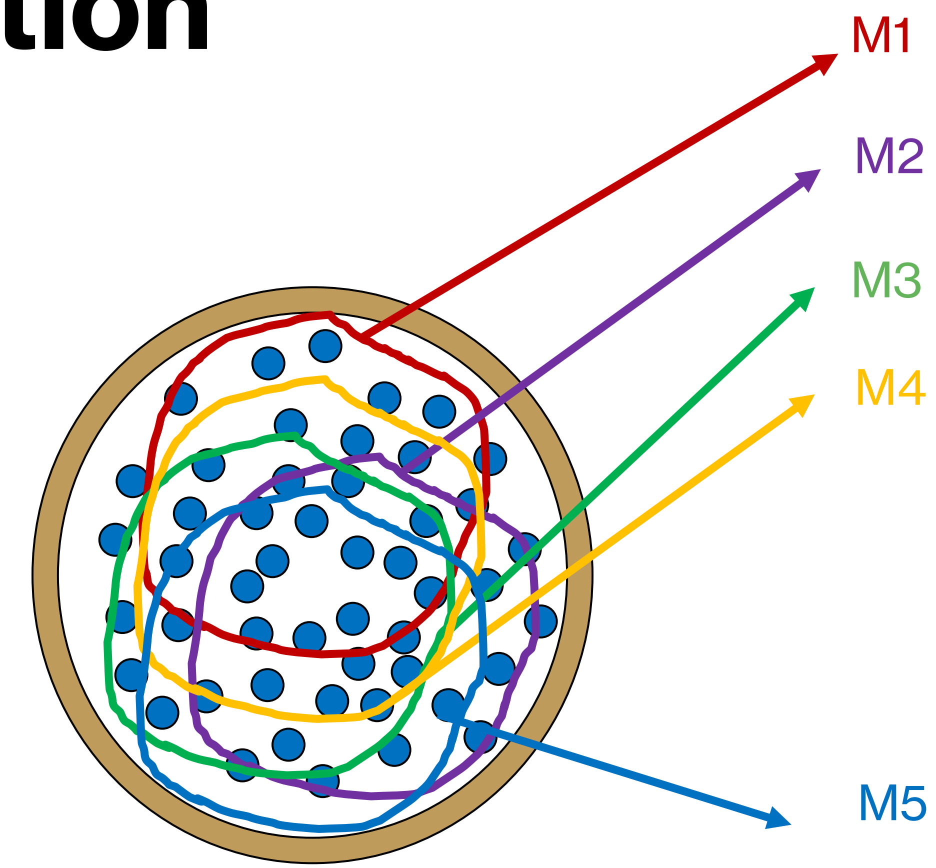
sampling from a population

- sampling error: the discrepancy between the sample statistic and the true population parameter it is estimating
- each time we sample, we compute some type of statistic (e.g., mean, correlation, etc.)
- sampling distribution: distribution of all possible values of the **statistic** obtained from multiple samples of a given size
- distribution of sample means contains all sample means of a size ***n*** that can be obtained from a population



sampling from a population

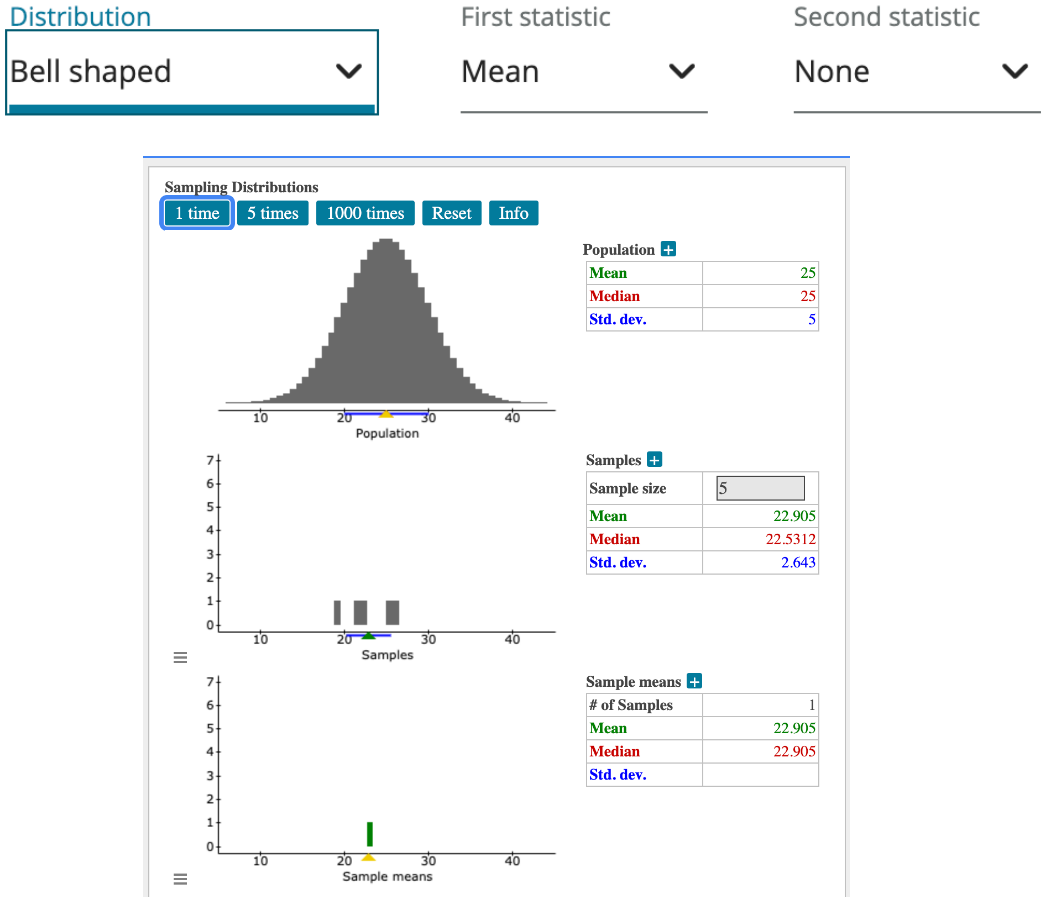
- sampling error: the discrepancy between the sample statistic and the true population parameter it is estimating
- each time we sample, we compute some type of statistic (e.g., mean, correlation, etc.)
- sampling distribution: distribution of all possible values of the **<u>statistic</u>** obtained from multiple samples of a given size
- distribution of sample means contains all sample means of a size ***n*** that can be obtained from a population



what does the sampling distribution of means look like??

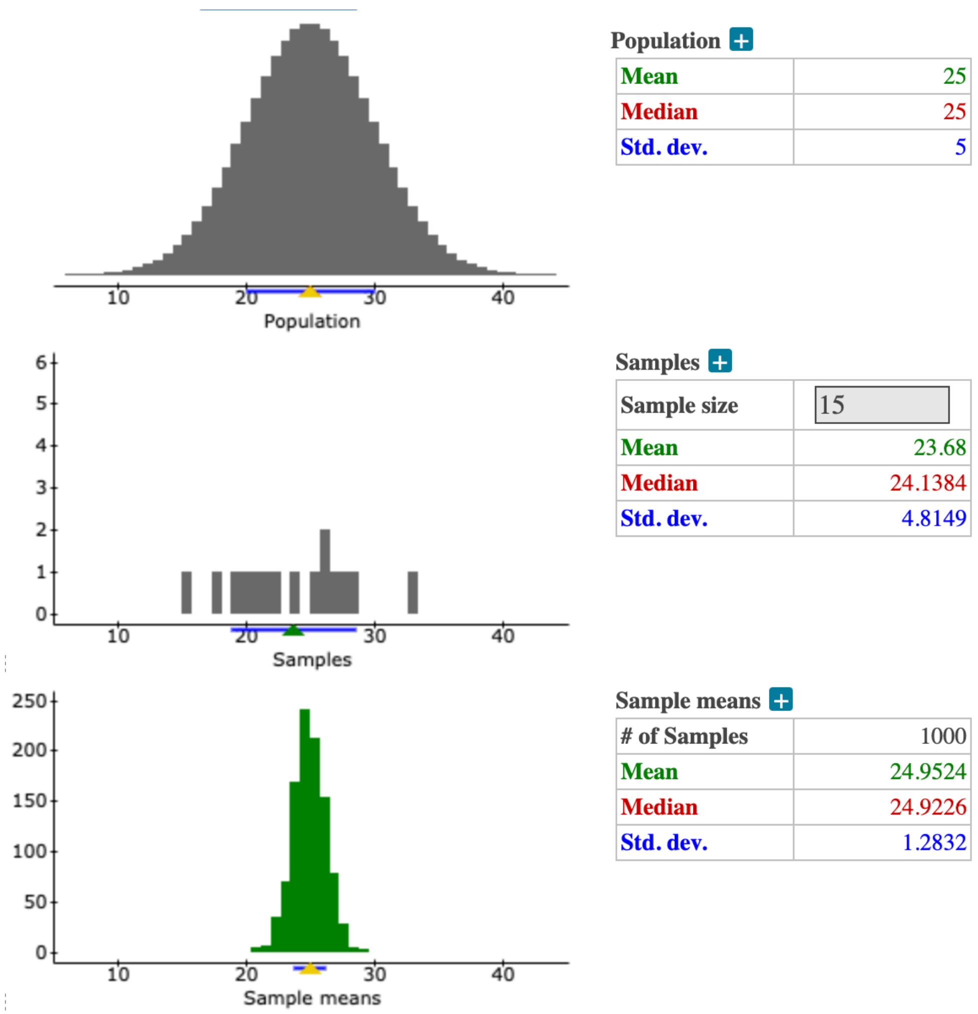
sampling distribution

- simulator
- change the distribution to bell-shaped and make sure the first statistic is the “mean” and the second statistic is “none”
- start with a single sample of size 5 and play it 1 time vs. 5 times vs. 1000 times
- explore what the three graphs are showing



sampling distribution

- three distributions
  - the population distribution
  - the sample distribution
  - the sampling distribution (of all means)
- mean of sample means = population mean (unbiased estimator!)
- the sampling distribution of means approximates the normal distribution as *n* (sample size) increases

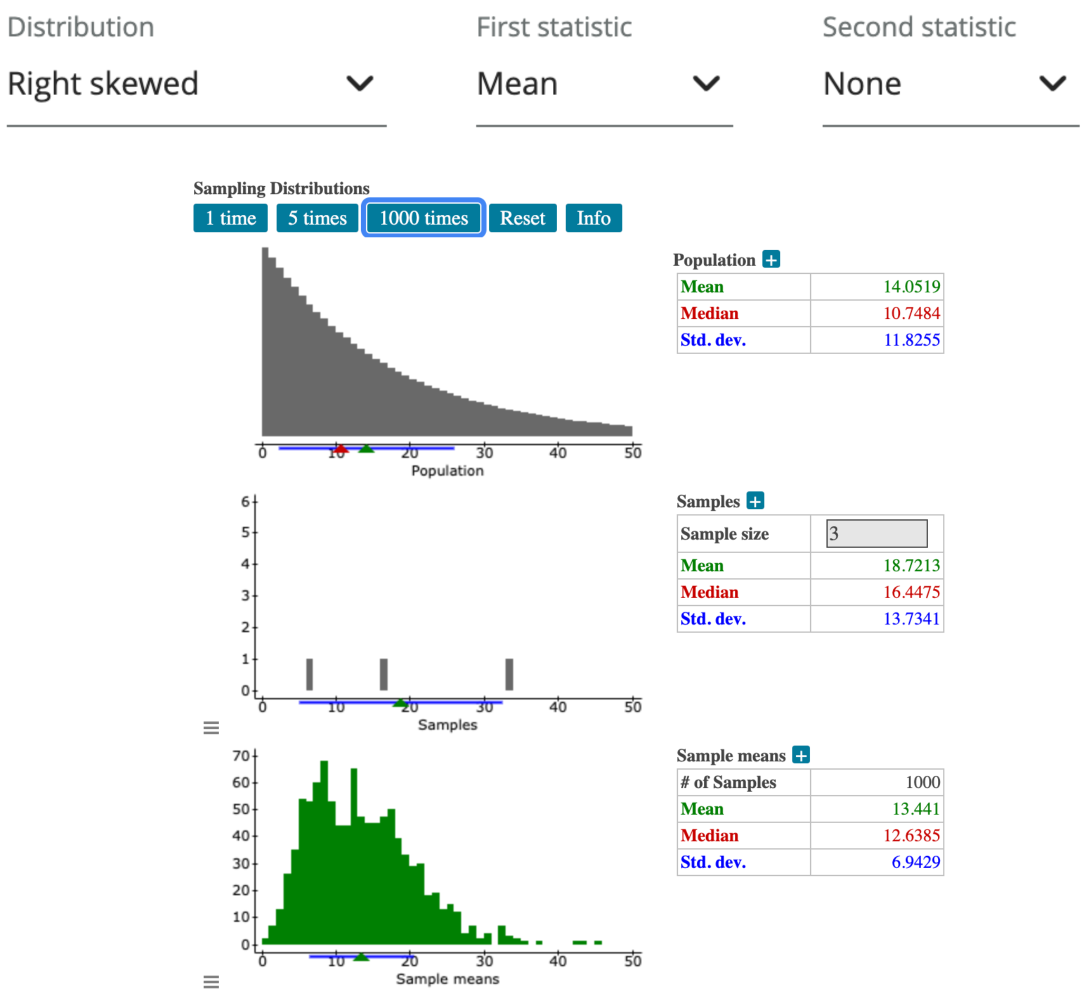


from all samples to few samples

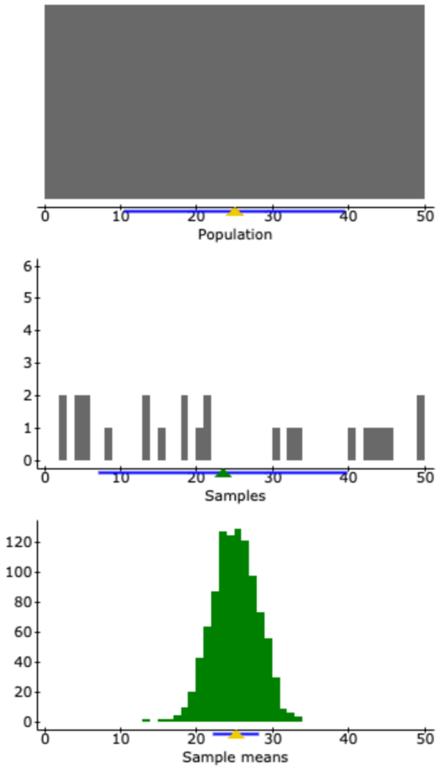
- in practice, we cannot compute <u>all</u> possible samples of size n
- the central limit theorem states that for <u>**any**</u> population with mean μ and standard deviation σ , the distribution of sample means for sample size *n* will have:
  - a mean of $\mu_M = \mu$ = expected value of M
  - a standard deviation of $\sigma_M = \frac{\sigma}{\sqrt{n}}$ = standard error of the mean or M
  - will approach a normal distribution as n approaches infinity
  - distribution of sample means will be normally distributed <u>**even if the population was not normally distributed (if n is large enough!)**</u>
  - typically n (number of scores in a sample) around 30 yields a reasonably normal distribution

any distribution?

- [simulator]
- change the population distribution to any non-normal distribution
- make sure the first statistic is the “mean” and the second statistic is “none”
- explore what the sampling distribution looks like for small and large samples



distribution of sample means



Population

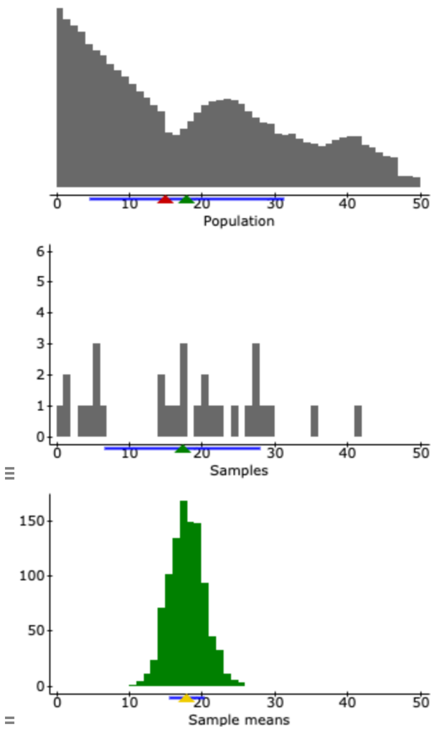
Mean	25
Median	25
Std. dev.	14.4309

Samples

Sample size	25
Mean	23.5236
Median	20.6888
Std. dev.	16.3464

Sample means

# of Samples	1000
Mean	25.1766
Median	25.172
Std. dev.	2.929



Population

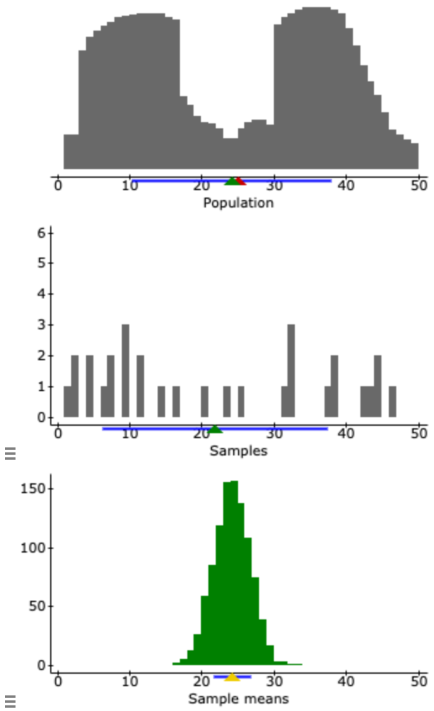
Mean	17.8409
Median	15.0101
Std. dev.	13.3769

Samples

Sample size	30
Mean	17.3445
Median	17.694
Std. dev.	10.6863

Sample means

# of Samples	1000
Mean	17.9594
Median	17.8971
Std. dev.	2.3847



Population

Mean	24.157
Median	25.0403
Std. dev.	13.7207

Samples

Sample size	30
Mean	21.8028
Median	18.6057
Std. dev.	15.5371

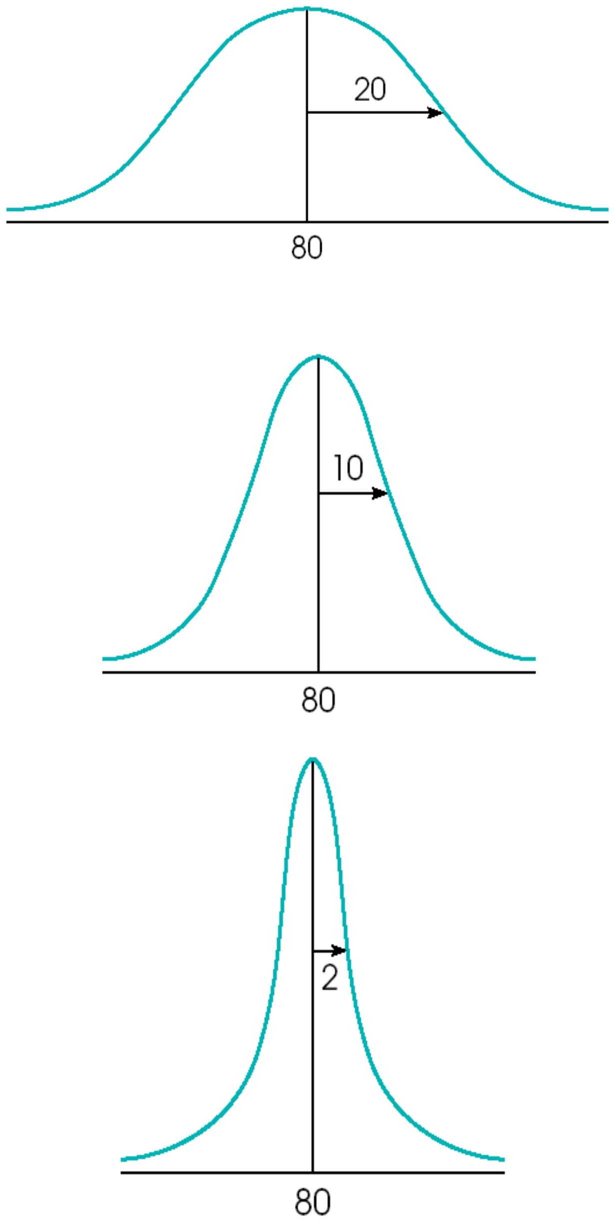
Sample means

# of Samples	1000
Mean	24.2364
Median	24.2623
Std. dev.	2.5189

for a large n, distribution of sample means will be normally distributed
even if the population was not normally distributed!

standard error of the mean

- standard error of the mean: $\sigma_M = \sqrt{\frac{\sigma^2}{n}} = \frac{\sigma}{\sqrt{n}}$
- how different a mean from one sample could be from another on "average"
- also measures reliability: how well an individual sample's mean represents the entire distribution of sample means
- law of large numbers: the larger the sample size (n), the more likely that the sample mean is closer to the population mean, and smaller the σ_M
- insight: we cannot control the population standard deviation but we can control the sample size!
- if we want our standard error of the mean to be low, we can use larger samples



example

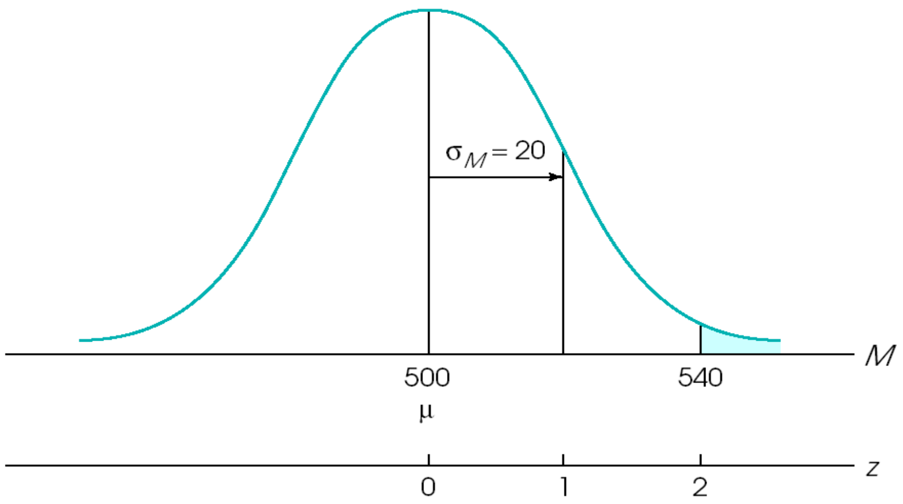
- SAT-scores population (μ =500, σ =100). If you take a random sample of *n* = 16 students, what is the **probability that the sample mean will be greater than *M* = 540**?
- we are talking about the sample mean, not an individual score anymore! i.e., we use the distribution of sample means (i.e., the sampling distribution) which approaches the normal distribution for large n
- represent the problem graphically
- calculate σ_M and z

$$\sigma_M = \frac{\sigma}{\sqrt{n}} = \frac{100}{\sqrt{16}} = \frac{100}{4} = 20$$

$$z = \frac{M - \mu}{\sigma_M} = \frac{540 - 500}{20} = 2$$

- visual calculator
- p (M > 540) = .0228

distribution of sample means



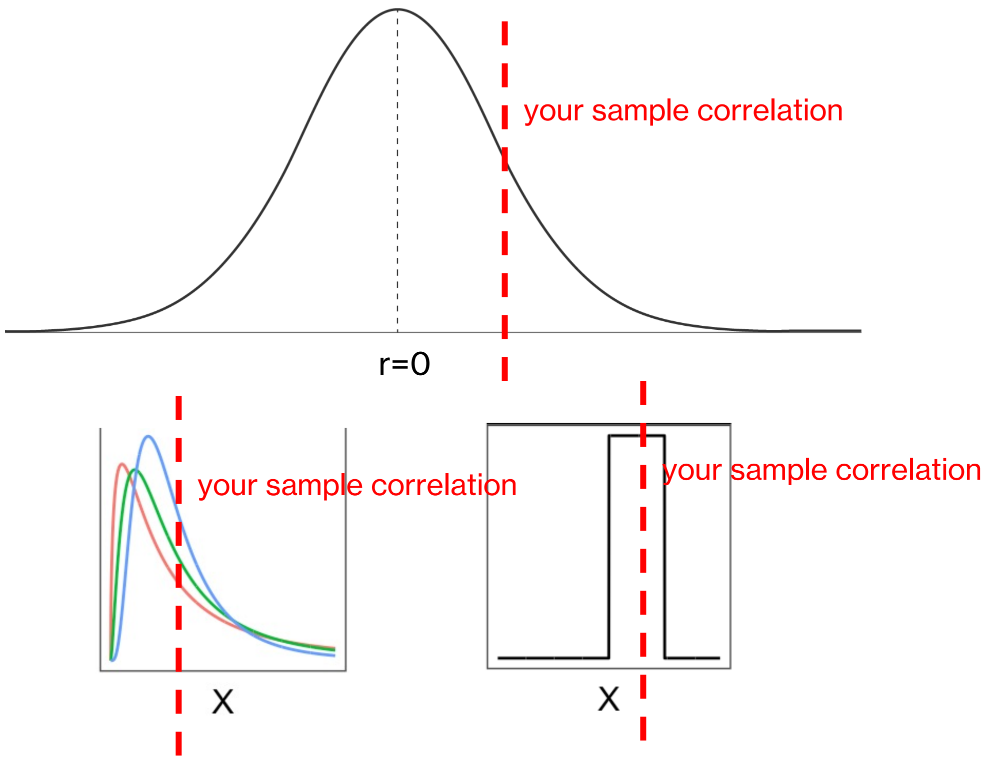
activity

- Jumbo shrimp are those that require 10–15 shrimp to make a pound. Suppose that the number of jumbo shrimp in a 1-pound bag averages $\mu = 12.5$ with a standard deviation of $\sigma = 1$, and forms a normal distribution. What is the probability of randomly picking a sample of $n = 25$ 1-pound bags that average more than $M = 13$ shrimp per bag?

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?

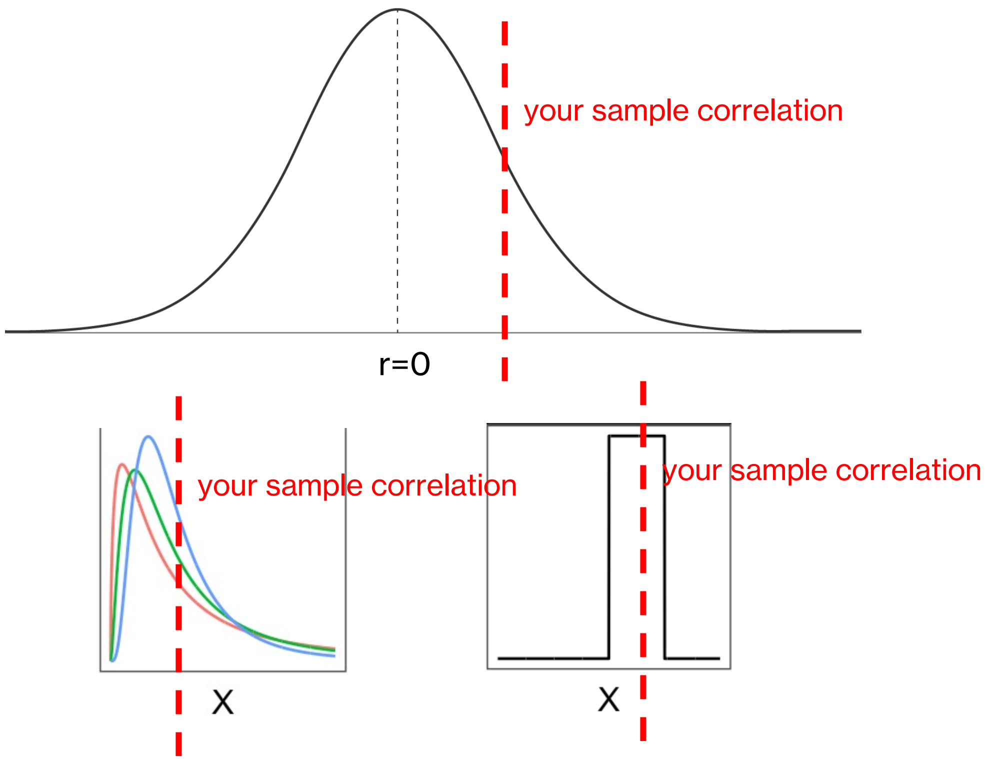
ALL sample correlations with sample size n when there is no meaningful relationship between height and weight in the population



outstanding question #1

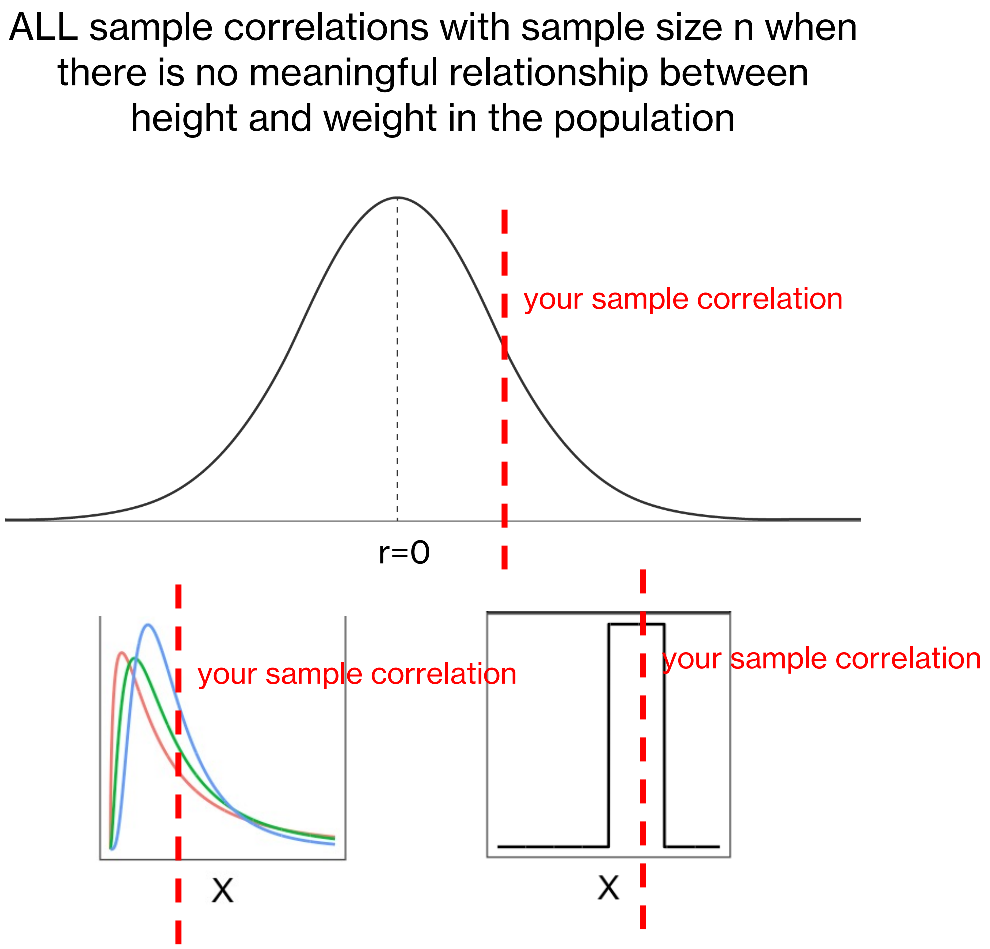
- **<u>question 1</u>**: how do I calculate probabilities if I don't have access to ALL the scores?
- if I know that the distribution of the sample statistic is normal, or approaches normal, then I can calculate probabilities!

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



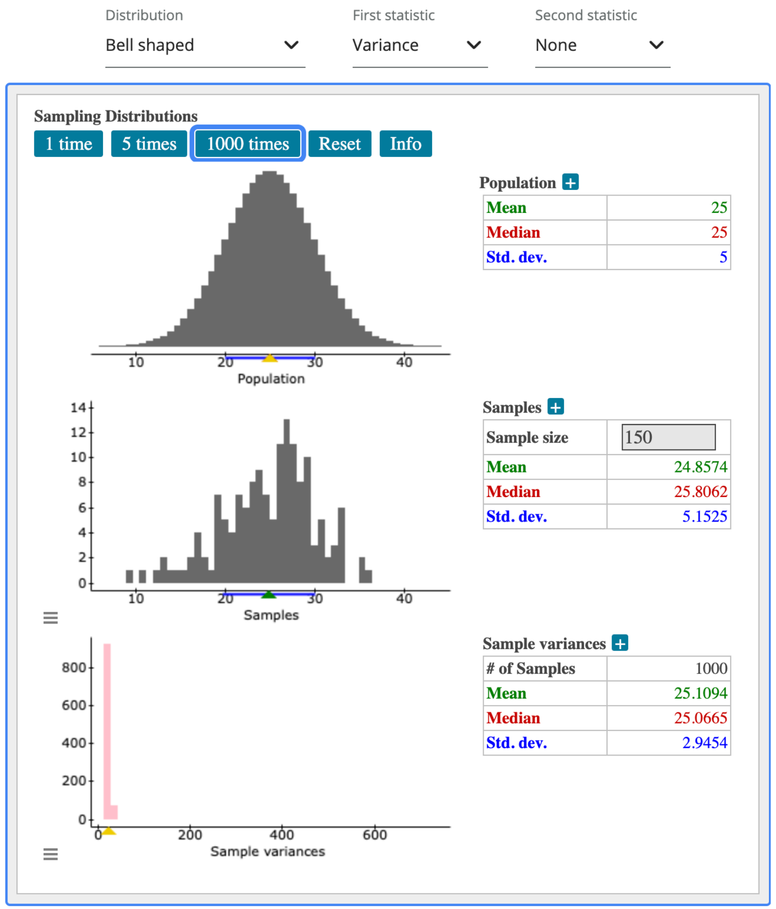
outstanding question #2

- **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
- the central limit theorem states that for **any** population, the distribution of **sample means** will be normal for large sample (n > 30)
- caveat: CLT only applies to sample means, NOT other sample statistics!



outstanding question #2

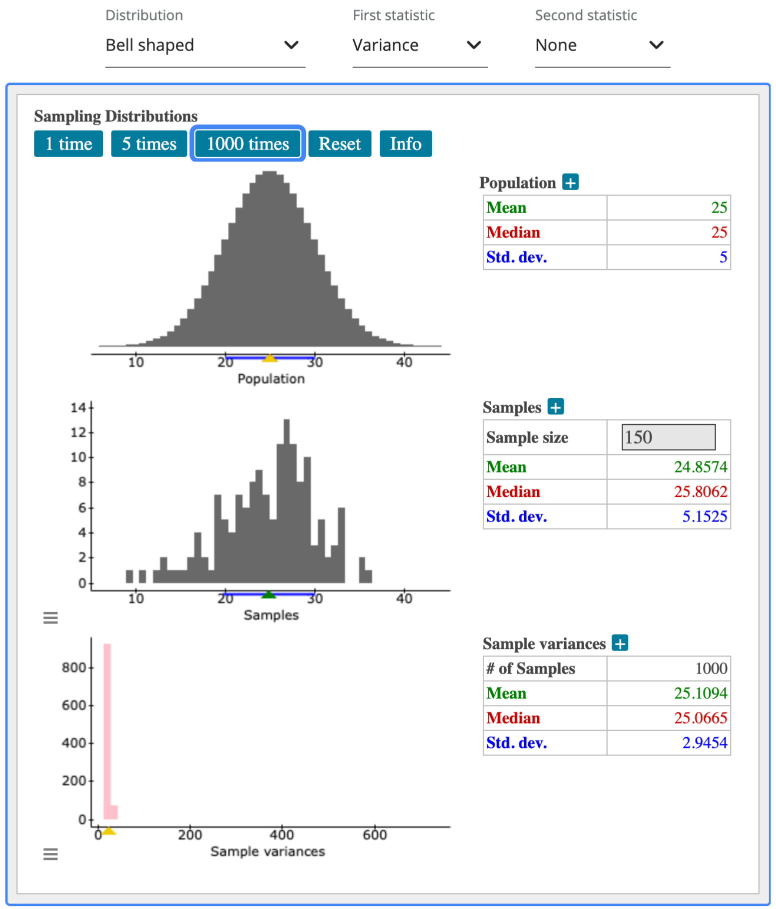
- simulator
- change the distribution to bell-shaped and make sure the first statistic is the “**variance**” and the second statistic is “none”
- start with a single sample of size 5 and play it 1 time vs. 5 times vs. 1000 times
- explore what the three graphs are showing
- the sampling distribution of variances is NOT normally distributed
- sampling distribution of several other statistics (e.g., correlation) may also NOT be normally distributed!



outstanding question #2

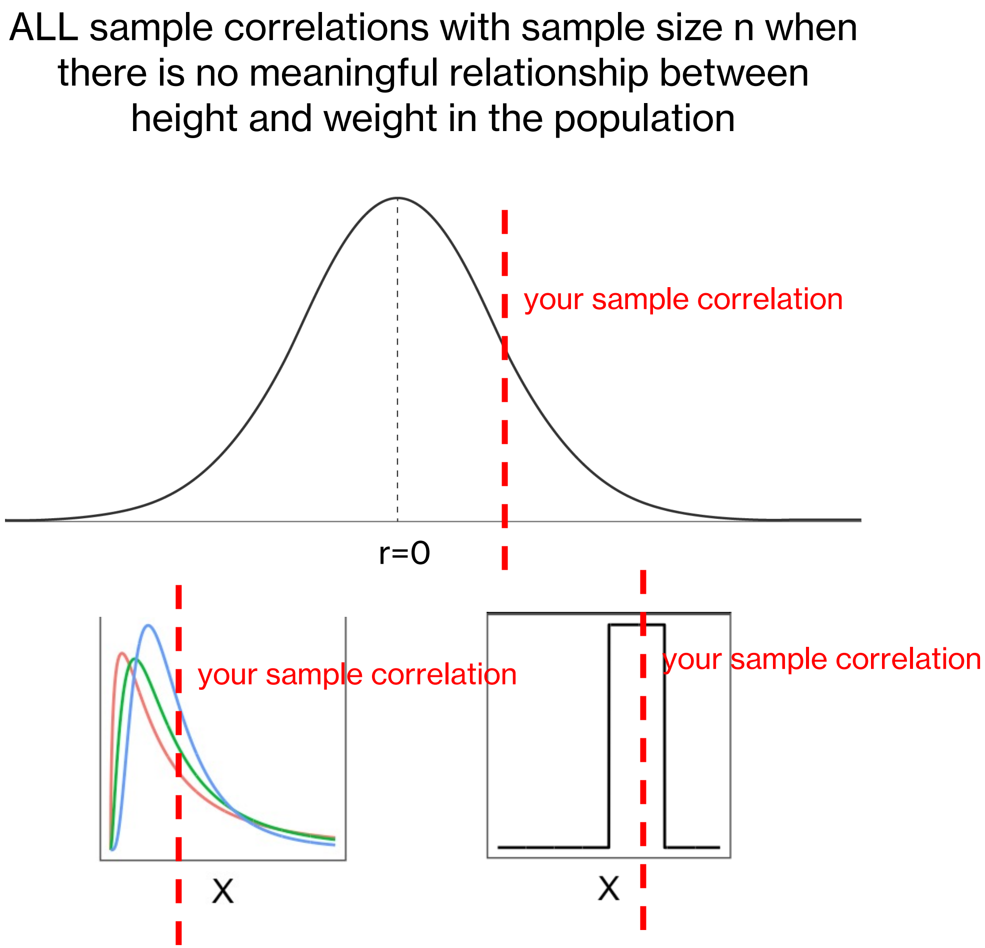
- **<u>question 2</u>**: how do we know what the distribution of the null hypothesis looks like? If we don't know the <u>form</u> of the distribution, we cannot calculate probabilities

- if we can figure out the sampling distribution of the sample statistic (e.g., means, variances, correlations, etc.), and we know the mathematical form of these distributions, we can find probabilities



outstanding question #3

- **<u>question 3</u>**: how do we know whether the probability we obtained, i.e., P(data | null hypothesis) is small enough?
- we need to set thresholds in place BEFORE we look at our data (no peeking!)
- all researchers/scientists must follow the same framework when testing hypotheses
- enter: null hypothesis significance testing (NHST)



today's agenda

probability and inference

sampling

class survey

some changes

- problem sets now due Tuesday night
  - PS4 is due March 11
  - PS4 revision is due March 27
- quizzes will still be due Monday night
- pace will be slower
- more practice problems using Sheets (but, we have limited time)
- PLEASE watch the videos (when they are listed on the website!)
- rethinking office hour times (TBD)
- thanks for your feedback!

next time

- **before** class
  - *prep*: textbook readings
  - *try*: week 6 quiz
  - *apply*: PS4 problems (chapters 6 and 7)
  - *apply*: optional meme
- **during** class
  - hypothesis testing