* **Biostatistics** = a theory + methodology for the acquisition + use of quantitative evidence in biomedical research.
* **Biostatisticians** develop innovative designs + analytic methods targeted at increasing available information, improving the relevance + validity of statistical analysis, making the best use of available information, + communicating relevant uncertainties.
* **Experimentation** and **empiricism** are ultimately the language of science.
* **Inductive reasoning** based on empiricism is how the vast majority of science gets conducted   
  + **statistics** = the language of empiricism 🡺 the formal syntax associated with empiricism.
* When studying statistics, you're actually studying how to conduct experiments + quantify information from experimental + empiric knowledge.
* Ex: Existing clinical practice was reversed by empirical evidence 🡺 a clinical trial in 2002 that contradicted prior evidence on the efficacy of hormone replacement therapy.
* Up to that point, HRT had been near universally thought of as a good treatment for post-menopausal women.
* In fact, the trial evidence was exactly the opposite + suggested the possibility of some harm.
* When conducting this experiment, the statisticians + epidemiologists, etc. that designed the trial had to come up w/ ways in which to decide whether or not to stop the trial + how to quantify the evidence.
* In this case they had a statistically based protocol (based on probability) they executed + the study was stopped due to an excess number of negative events
* This is an incredibly important decision, b/c it's unlikely that another trial of this scope would’ve been able to have been executed anytime soon.
* So, cutting it short would‘ve the incredibly negative benefit of not having the full information available but keeping it going based on the evidence of an excess # of negative events could’ve possibly put people at harm.
* Another very good example: **ECMO** (an **extracorporeal membrane oxygenation treatment** for newborn infants) is an incredibly promising treatment + is currently a standard treatment
* There was quite a bit of enthusiasm at the time of creation that it would be revolutionary.
* It was very difficult to consider how you conduct a trial for a treatment that has such obvious potential benefits.
* The ethical considerations in the trial wound up leading to a **statistical design/randomization scheme**, whereby only ONE infant received the control therapy.
* That in turn, introduced sample-sized-based criticisms about the design itself.
* In this particular case, b/c the treatment was so potentially beneficial, it was difficult to figure out how exactly to conduct a trial to evaluate its efficacy + in only one instance were they actually willing to give the control therapy to a child.
* However, immediately the degree of evidence to *actually evaluate* the therapy was very weak
* So, in this case, it wasn't just it wasn't *just statistical* considerations, it was *ethical* + *medical* considerations all bundled together.
* These 2 examples illustrate the central role statistics plays in making public health decisions + making medical decisions + evaluating evidence.
* **Biostatistics** is playing a central role in public health, + in design, analysis, + interpretation of statistical data.
* At JHU Bloomberg School of Public Health, the prevailing philosophy for conducting biostatistics across all of the department includes
* tight coupling of methods used w/ ethical + scientific goals.
* emphasizing scientific interpretation of statistical evidence + trying to use that interpretation to impact public policy.
* Trying to acknowledge assumptions + evaluate robustness of conclusions to these assumptions.
* This is not the only way to think about doing statistics.
* The FDA may not want a tight coupling of statistical methods w/ scientific goals + may actually want statistical procedures to *protect* consumers, regardless of scientific evidence.
* This **loss function** is, in essence, quite different.
* They may want to actually divorce the statistical methods from the science so they can, in the most possibly unbiased way, evaluate a drug.
* Different organizations + entities would think about different statistics + how it applies to scientific problems in different ways.
* In statistics, think of an **experiment** in a much more abstract + general way 🡺 as **any purposeful inquiry**.
* That involves a non-trivial amount of **error** and **randomness** we need to evaluate.
* Consider the **outcome** of an experiment such as a collection of measurement from a sample population or from a lab experiment, or the results of clinical trial or simulated CPU experiment, or values from hospital records samples retrospectively, etc.
* In different kinds of experiments, there ARE different *levels* of evidence that come about, but for all of these, think of them as an experiment that *requires statistical analysis.*
* **Simulation CPU experiment**s are used b/c CPU’s are very valuable tools for performing analysis as well as for methods of actually generating the experiment.
* Ex:Division of Medical Imaging Physics evaluating a new imaging technique
* Before they run very costly human/animal trials, they’ll design a CPU experiment to evaluate image processing/analysis techniques to get highly accurate models of the imaging system they've built w/ added-in randomness + other sources of variation.
* This produces a vast amount of data that requires very sophisticated statistics to actually evaluate
* As a result, they can run quite detailed + quite accurate experiments entirely **in silico** before running expensive human/animal trials.
* **Retrospective sampling experiment** is used b/c in many cases, it's very difficult to follow patients over time + ascertain disease status as it comes about (**prospective studying**).
* Frequently the only possible avenue available = to study things retrospectively from hospital records or however
* Goal = Figure out who has a disease/condition, compare them w/ controls who may be similar ages or in demographics, + evaluate historically what their exposure's for.
* This is frequently done b/c it's the only (convenient) way to study a phenomenon when first looking at it.
* Each of these different kinds of experiments offer different sources of variation, different sources of experimental error, + different kinds of randomness to be evaluated w/ statistical models.
* We're going to connect these numbers resulting from all of these experiments + many others to **probability models**.
* To do that, we must understand the subject of probability at a deeper level w/ the eventual goal to connect these probability models to the actual data.
* Usually when talking about **set notation**, you talk about some uber space that contains everything.
* In statistics, this the **sample space**, omega, **Ω** = the collection of all possible outcomes of an experiment
* Ex: experiment **🡺** Roll a die w/ possible set of outcomes **Ω = {1,2,3,4,5,6}**
* An **event** **E** = any *subset* of the sample space 🡺 even #’s event E = {2,4,6}]
* An **elementary/simple event, lowercase omega ω** = the particular result of an experiment
* If a die roll is a four 🡺 **ω** = 4
* **Null event/empty set** **∅** = the event that nothing occurs.
* Sets in probability theory follow all the same rules as ordinary set notation just w/ different interpretations
* **ω** ∈ **E 🡺** “elementary event is an element of an event E” **🡺** implies E occurs when ω occurs
* ω = 4, then E = even
* **ω** ~~∈~~ **E** **🡺** “elementary event is NOT an element of an event E” **🡺** implies E does NOT occur when ω occurs
* ω = 5, then E = even is not true
* E ⊂ F implies the occurrence of E implies the occurrence of F
* E = {2,4,6} F = {2,4,5,6} 🡺 if you roll an element of E, you’ve also rolled an element of F
* **E** ∩ **F (intersect)** implies the event that both E and F occur
* E evens = {2,4,6} F prime # = {1,2,3,5} 🡺 only element 2 works (even AND prime)
* **E** U **F** (**union**) implies the event that *at least one of* E or F occurs
* E evens = {2,4,6} F prime # = {1,2,3,5} 🡺 get an even or a prime or both (2)
* **E** ∩ **E = ∅** means E and F are **mutually exclusive** (cannot simultaneal occur)
* E = {2,4,6} F = {1,3,5} 🡺 can’t roll an even and an odd
* **Ec** (**complement**) **or –E** = the event that E does NOT occur
* E = {2,4,6} **Ec** = {1,3,5} 🡺 can’t roll an even and an odd 🡺 E and **Ec =** always mutually exclusive
* Standard set theory facts
* **DeMorgan’s Laws:** 
* **A intersect B complement** = **A complement union B complement**
* Imagine the complement symbol distributes across the parentheses to A and to B + flips the cap into a cup
* **A union B complement** = **A complement intersect B complement**
* C distributes itself across the parentheses + the cup turns into a cap.
* DeMorgan's law basically says if you complement across either an intersection or union, the *complement distributes itself, but it flips everything.*
* Ex: event A = you're an alligator, event B = you're a turtle, so event A U B = event you’re EITHER a turtle or an alligator (at least 1)
* Complementing that w/ **(A U B) c** (if you are NOT an alligator or turtle), DeMorgan's Law says we have complement A intersect complement B **(Ac** ∩ **B c)**🡺 you are not an alligator + you are not a turtle 🡺 “if an alligator or a turtle you are not, you are not an alligator + also not a turtle”
* Ex: event A = your car is hybrid + event B = your car is diesel, so event that your car is not *both* hybrid and diesel **(A** ∩ **B) c,** complement the inner sections.
* This gives **(A c U B c)**🡺 if your car is not *both* hybrid + diesel, your car is either not hybrid or not diesel = complement A union complement B
* **(B c) c = B** (if you do “not, not” get an even number, you get an even number
* **(A U B) ∩ C = (A U C) U (B ∩ C)**
* Think of the **union** as sort of plus + the **intersection** is multiplication.
* This rule looks like the distributive property 🡺 C sort of gets multiplied by A + by B +
* Now to use set notation to develop probability as a modeling tool to analyze data in a strategy that underlies much of science
* For a given experiment
* **attribute *everything know*n to a systematic model** (like lines, planes, + hyperplanes) where we presume an outcome (like hypertension) depends on a lot of predictors in a linear fashion.
* This is either “known” or theorized or assumed for sake of convenience, but it relates known predictors to the known outcome + then **attributes everything else to randomness**.
* This is a very difficult bullet to swallow for many people b/c in nearly all applications of probability, the word “**random**” is very difficult to tie down
* If you were to model the outcome of whether or not a person had a disease as predicted by medical history via some form of retrospective sampling, it's not exactly clear where the randomness is coming from or even what randomness *means* in this context.
* Even if that's the case, we still often use probability to evaluate the collection of unknown things in an experiment, treating them as if they *were* random
* Must be careful in HOW we interpret our probability statement in that context, relative to what the word “random” is meaning in that case.
* In some other settings, people have *very specific definitions* of what random means.
* Sometimes people analyze clinical trials using **randomization** used to assign patients to “treatment” or “control” *as the actual probability that they're modeling in their mathematical models*.
* There, they can point very directly to what randomness they're modeling (this has its own problems as well)
* The process of **using probability to quantify uncertainty in conclusions** **to model this randomness** is actually a very delicate subject.
* All three of the above bullet points come w/ quite a bit of baggage in terms of assumptions + things you cannot evaluate at all
* So, we'd like to **check how sensitive our conclusions are to assumptions in these models**.
* In some cases, we can directly verify them + check whether the relationship between the response + the predictors actually looks like a line, so we'd be okay if modeling it as a line.
* In other cases, they involve assumptions that we can't possibly check 🡺 variables we did not collect or variables that we do not even know.
* In this case, we have to *evaluate our sensitivity to our model in terms of unknowns* + evaluate how **robust** our approach is to unknowns via the study of how the data was collected, how the statistics were used, and what exactly is probability actually modeling?
* These are very difficult topics that many people struggle w/ if thought of with sufficient depth
* If you just did *one thing* when thinking about probability in your data that you're analyzing is when you say, “I have a 95% CI” or “my p-value is ….”, etc., where you actually use probability in your actual data analysis, **go through the exercise of:**
* **trying to think of *what is it* that you're modeling as random.**
* What is the **sources of this randomness** (where does it arise from)
* How good of a job do you think your probability statements do at characterizing this randomness?
* Where did systemic model components arise from?
* How did observational units come to be in the study (+ is there importance to the missing DP’s?)
* Do results generalize beyond the study?
* Were important variables unaccounted for in the model?
* How drastically would inferences change depending on answers to the above questions?