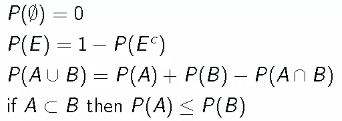
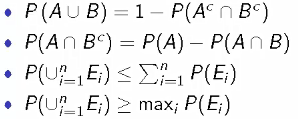
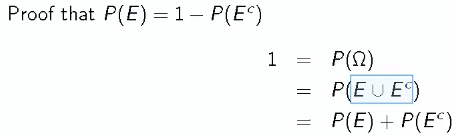
* **Biostatistics** = a theory + methodology for the acquisition + use of quantitative evidence in biomedical research.
* **Biostatisticians** develop innovative designs + analytic methods targeted at increasing available information, improving the relevance + validity of statistical analysis, making 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 simultaneously 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?

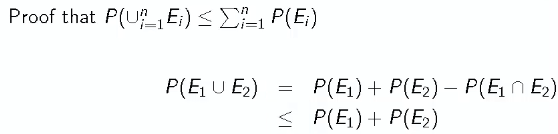
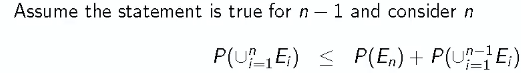
**Probability**

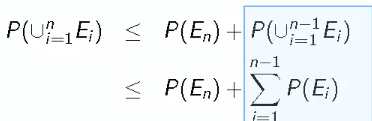
* A **probability measure *P*** = *a real-valued function from the collection of possible events* that's going to govern the rules of probability for us.
* There is basically 3 rules a probability measure has to follow/that must hold:
* For an event ***E*** ⊂ **Ω, 0 <= P(E) <= 1**
* Events E have to be mapped to numbers between 0 + 1
* *Probability is a function that operates on sets.*
* **P(Ω) = 1**
* Sample space must enumerate everything possible that *could* happen
* **If E1 and E2 are ME (no intersection) events, then P(E1 U E2) = P(E1) + P(E2)**
* There is interesting history behind these three rules.
* The Russian mathematician Kolmogorov (generally considered the father of all of the modern probability basically distilled everything we thought of in terms of things that a probability should have to follow down to the minimal set of rules you could possibly have
* If any of these rules do NOT hold, you wind up w/ something that fails in some fundamental way to be probability + if you add any other rules they turn out to be excessive.
* He also found that the theory of exactly what randomness is + what probability measure are very deep problems
* about the third statement because we are giving you an incorrect version of it.
* The third statement says that if two events are mutually exclusive + recall
* Now, rule 3 above is not “complete”.
* It says w/ 2 mutually exclusive events, the probability of the union = sum of the probabilities.
* This easily extends to so-called **finite additivity 🡪 probability of the union of n mutually exclusive events = sum of n probabilities** 🡺 where the {Ai} = mutually exclusive
*  = 
* *This* is then usually extended to **countable additivity** 🡺  🡺 instead of n, go up to infinity
* This requires ideas of limits + other things not covered here
* At any rate, it's the case that ***finite additivity*** *does NOT imply* ***countable additivity***, BUT *countable additivity implies finite additivity*.
* So, in standard or more theoretical probability classes, there is quite a bit of hay in this distinction.
* The general definition gives *countable additivity* rather than finite additivity.
* A more advanced measure theoretic probability class, they will deal w/ this issue at length.
* In general, finite additivity will work just fine for us.
* Recall, our **probability function *P*** operates on **events** (subsets of the **sample space**) + maps them to numbers between 1 + 0, so we need an **appropriate domain *Ϝ***
* ***P*** is defined on ***Ϝ*** a collection of subsets of **Ω**
* A domain ***Ϝ*** is NOT an event, but is a *collection* of events.
* Ex: **Ω = {1,2,3}**, so ***Ϝ =*** *{*∅, {1},{2}, {3}, {1,2}, {1,3}, {2,3}, {1,2,3}}
* The probability function P operates on ALL possible events that are subsets of that sample space
* *Pretty much whenever you have a* ***finite set****, the domain of the probability function will operate on all possible subsets of the sample space.*
* When the sample space is a **continuous set**, it gets a lot harder + you can no longer say things like “the probability operates on the set of all possible subsets of a continuous set”
* This is an incredibly deep mathematical problem.
* The mathematician Cantor thought about measure + sets in a very deep way + came up w/ interesting sets that you can't reasonably include in the definition of a probability.
* For our purposes, when our sample space is a continuous set, we’re mostly going to be concerned w/ things like **intervals** or **unions of intervals,** + in that case, definitions are very easy
* For our definition of **domain** the probability operates on, we just assume that anything we can think of is fine 🡺 **assume *Ϝ is sufficiently rich so that any set we’re interested in will be in it***
* Laundry list of properties a probability function has to have by virtue of its three definitions.

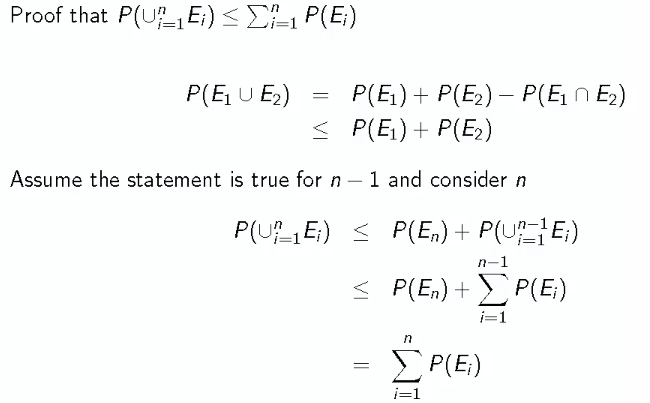
 

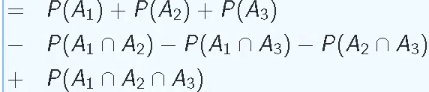
* Our 3 simple definitions imply all these things that we know probabilities have to have.
* 1) P(∅) = 0 🡺 probability of the empty set = zero/the probability nothing happens = 0
* So if you say you're going to roll a die, you actually roll a die + the probability of no roll = 0
* 4) If A is a subset of B, probability of A is less <= probability of B.
* This is analogous to rolling a die + A = get a 1 + B = get a 1 or 2, so probability of getting a 1 is less than the probability of getting a 1 or a 2.
* 5) From DeMorgan's laws we get probability of A union B is one minus the probability of the intersection of A complement and B complement
* 6) Kind of along the lines of subtraction 🡺 (A intersect B) complement = sort of like subtracting B out of A (the component of A that has nothing to do w/ B)
* So the “probability of A removing B” is the probability of A - the probability of A intersect B,
* This works out to be a nice rule 🡺 set level subtraction works out to be equivalent to subtracting the probabilities.
* 7) probability of the union of a collection of events <= sum of the probabilities of the events.
* Again, if the events are mutually exclusive, the probability of the union MUST equal the sum of the probabilities
* This rule doesn't violate the finite additivity rule whatsoever, but *also* accounts for when the events are NOT mutually exclusive.
* 8) Probability the union of events >= probability of the *maximum* of the collection of probabilities.
* *Again, this holds true if the events are mutually exclusive or not*.
* But there's intuition behind this that's very easy.
* The **union** is everything that's in ANY of the events (E1 To En), so the probability of the union MUST be bigger than any of its component events.
* Ex: E1 = roll 1, E2 = roll 2, E3 = roll 3, the probability on the LHS of the equation = probability you get a 1,2, or 3
* On the RHS, for a standard die, probability of each outcome = 1/6 + the LHS = union of the probability of a1 , 2, or 3, = 1/2, which > 1/6
* Example of a **proof** 🡺 P(E) = 1 – P(E) **c**
* Recall P(**Ω**) = 1, but the sample space **Ω** for any event = the union of that event + its complement
* So, **Ω** = E U E **c** , and since **an event is always mutually exclusive w/**
* **its complement**, the probability of their union = sum of the probabilities



* A more complex example of the consequences of the probability rules.
* Recall probability of the union of a collection of events <= sum of the probabilities + that “less than or equal to” is an equality if the events are mutually exclusive.
* Prove this using **mathematical induction**.
* Mathematical induction works b/c you prove it for some small statement, 1 or 2, + then you assume it's true for say n-1, + then prove that it's true for n.
* Consider 2 events, probability of E1 U E2.
* From one of the probability rules we investigated, this equals P(E1) + P(E2) – P(E1 ∩ E2)
* Assume we've proved that one as well.
* We are subtracting off a number that has to be positive (P(E1 ∩ E2)) b/c probabilities have to be between 0-1 🡪 must be non-negative at least.
* So if we throw away that final term, what's left can only get bigger, right?
* If we're subtracting off a positive number, then it's got to get bigger 🡪 establishes the result for the case w/ 2 events
* 
* Now assume the result is true when we have n – 1 events, + consider n events.
* We want to demonstrate the probability of the union of the Ei <= sum of the probabilities
* 
* Write out the probability of the union of the Ei's as En union w/ the union of the rest of them (i from 1 to n-1) in a single set.
* 
* We've already done that 🡺sets En + the union of the remainder n-1 are 2 separate sets, so we can say that the probability of the union E1-En <= probability of EN + probability of the remainder.
* Now, we can say that we’ve only gotten bigger by our induction hypothesis (by assuming that this statement is true for n – 1 events)



* So if we switch from the probability of the union to the sum of the probabilities, we've only made it bigger + we can maintain that inequality.
* Then, collecting the terms, we just have that this is the sum of the probabilities (last line)
* 
* *“equals” on this last line means equals to the previous line, not the first line.*
* You should be able to prove all probability statements outlined above
* Let's take a step back from the mathematics + try to put some of this w/in a context.
* National Sleep Foundation reports ~3% of the American population has Sleep Apnea (sleep disease where the upper airways collapses). They also report ~10% of the North American + European population has restless leg syndrome (for our discussion, assume this is 10% of the *American* population). Similarly, they report 58% of adults in the US experience insomnia.
* A sleep physician wants to know probability a random American has ANY of these 3 disorders
* Can you simply add these probabilities + say “71% of people have at least 1 of these sleep problems”?.
* This question is nothing other than a restatement of the probability relationship we just proved.
* Here, we have 3 events, A
* A1 = {person has sleep apnea}, A2 = {restless leg syndrome}, A3 = {insomnia}
* probability a person has at least 1 of these diseases = A1 union A2 union A3 = 
* This equals to the sum of the probabilities ONLY when A1, A2, + A3 are *mutually exclusive*.
* *Otherwise* it's probability of A1 + probability of A2 + probability of A3, + subtracting out other things (intersections)



* This works out to be .71, but there's all the other stuff that MUST be less than zero when added in (i.e. subtract it) such as A1 intersect A2, A1 intersect A3, A2 intersect A3 + *then* you have to add in the triple intersect
* go through + figure out why exactly it is this formula works out.
* The point is that other stuff is **non-trivial** + it's ALWAYS there unless A1, A2, + A3 are *mutually exclusive*
* In this case, from a scientific perspective, it's probably the case that there's a non-trivial interception of people w/ sleep apnea + restless leg syndrome, a non-trivial interception of people w/ restless leg syndrome + insomnia + so on, such that .71 is not close at all

**Random Variables**

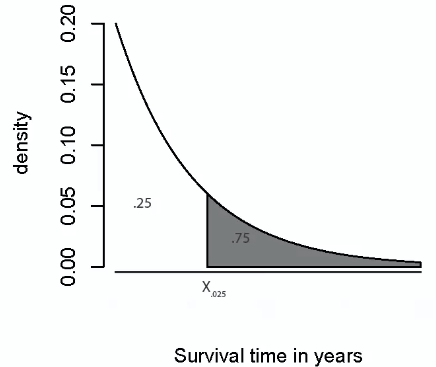
* **Random variable** numerical outcome of an experiment (**discrete**/countable, P(X = K), or **continuous, P(X** ∈ A) that have probability distributions associated with them
* How can we generate a random variable that is both discrete and continuous?
* Suppose we can generate something continuous (any value between 1 and 2 on the number line) and something discrete (die roll).
* Then, flip a coin, and if H, use the continuous number line random variable, and if T, use the die for the discrete random variable
* The resulting random variable could’ve been either, and we describe its behavior as potentially being both discrete AND continuous
* Ex: Imagine we’re looking at insurance expenditures .
* For those who’ve never been sick, we’ve paid out $0
* For everyone else, we pay out some amount we model w/ a continuous variable b/c we must account down to fractions of a penny.
* If the insurance company is modeling the payout distribution behavior, we should model it w/ a random variable that takes a discrete value 0 and all those continuous values <> 0
* In other words, the kinds of random variables we’ll use are **non-exhaustive**
* Examples of variables that could be random variables 🡺 H/T outcome of coin flip, outcome of roll of die, BMI of a subject 4 years after baseline measurement, hypertension status of a subject randomly drawn from a population

**Probability Mass Functions and Probability Density Functions**

* **Probability Mass Functions (PMF)** maps the rules of probabilities to sets of discrete random variables
* evaluated at a some value, it corresponds to the discrete random variable that takes that value
* ex: die roll 🡺 p(1) = 1/6
* Many functions satisfy the definition of a PMF b/c to be valid, the PMF *p* must satisfy:
* *p*(x) >= 0 for all possible value x of a random variable
*  taken over all possible values x of a random variable
* **Upper case letters X = conceptual (i.e. conceptual die roll)**
* **Lower case letters x = actual value (i.e. actual roll of 4)**
* Constructing a PMF w/ simplest example = a coin flip
* Let X be the result of a coin flip where X = 0 is tails, X = 1 is heads.
* Assuming the coin is fair, we want a PMF that maps 0 to ½ and 1 to ½ (infinitely many ways to write this function)
* We choose , which gives ½ for 0 and for 1
* If we CANNOT assume the die is fair, so we assume it’s bias, let ϴ be the probability of H as some proportion between 0-1 (such as .3, meaning P(T) = .7)
* So we want a PMF that says P(0) = 1 - ϴ and P(1) = ϴ 🡺 
* Greek letters = things we don’t know but would like to know = **parameters**
* In these cases, our PMF’s are the entities that governs the population of coin flips.
* So, to know ϴ, we collect data to estimate + evaluate uncertainty in the estimate via the **probability distributions =** conceptual models of populations + are entities that tie our data to the population
* Consider again the biased coin:
* Our PMF gives p(0) = 1 - ϴ and P(1) = ϴ. Prove this is a PMF
* So, **p(x) > 0** for x = 0,1 🡪 since ϴ is between 0-1 🡪 *satisfies rule 1 of PMF*
* Then, **p(0) + p(1) = ϴ + (1- ϴ) = 1**🡪 *satisfies rule 2 of PMF*
* **Probability Density Functions (PDF)** maps the rules of probabilities to sets of continuous random variables
* AREAS under PDF’s correspond to probabilities for a random variable
* The **Normal Curve/Bell Curve (Normal Density Functions)** is the “king” of PDF’s
* Areas under bell curves correspond to probabilities, so you're modeling something as if the population it belongs to follows a bell curve, we’re saying probabilities associated w/ that random variable are governed by areas under that bell curve.
* PDF’s must follow 2 rules to be valid:
* f(x) >= 0 for all possible values x the random variable can have
* 
* Define PDF’s as if they operate on the entire real line
* Ex: Assume time in years from diagnosis until death of a person w/ a specific cancer follows a density such as  (**exponential density function)** where x = years
* More compactly, ignoring f(x) = 0, we have 
* Prove this is a valid PDF so we could model survival this way
* 1) e raised to any power > 0
* 2) 
* EX: Using PDF’s to assign probabilities:
* Probability a randomly selected person from the population modeled as if it followed the above distribution survived > 6 years?
*  🡪 get the AUC from x > 6 to infinity
* **pexp(q=6, rate=1/5, lower.tail=F) # lower.tail = F means 6+, if T it’d be less than**
* **[1] 0.3011942**
* Remember, specific values for a continuous random variable have probabilities of 0 (i.e. a *line* has an area = 0)

**Cumulative Distribution Function, Survival Functions, Quantiles**

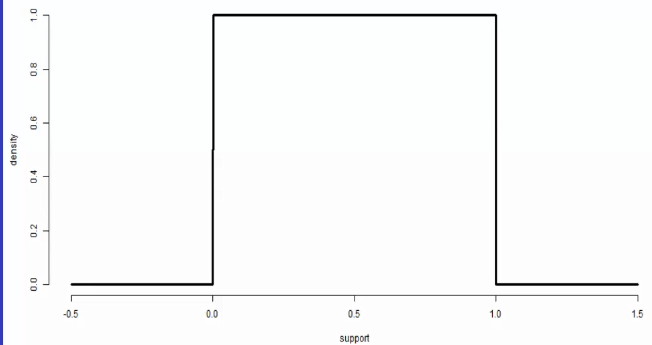
* **Cumulative Distribution Functions (CDF)** = for a value x, it is the function **F(x) = P(X <= x)** which returns the probability that the random variable X is less than or equal to that value
* Applies regardless of whether X is discrete or continuous
* Doesn’t matter if less than or less than or equal to for continuous
* **Survival function** of a random variable x = **S(x) = P(X > x) 🡪** probability the random variable X is greater than or equal to that value
* opposite of CDF/complementary 🡪 so notice that **S(x) = 1 – F(x)**
* \*\*\*For continuous random variables, **PDF = the derivative of the CDF**
* Examples: Remember to calculate, probabilities, calculate AUC for the PDF
*  
* We can recover the PDF via integrating the CDF = 
* **Quantiles α** = properties of distributions/density functions *F* such that they are the points **xα** such that **F(xα) = α** 🡪 probability of being <= xα is *exactly* **α**
* Ex: If **α** = .25, <= xα = the point such that the probability of being <= to it is 25%
* **Percentile** = a quantile **α** expressed as a percent, and **population** **median** = 50th percentile
* To find the 25th percentile of our exponential survival distribution, we want to find the point on the horizontal axis such that the AUC to the left of it = .25



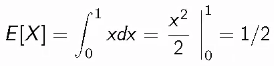
* So we must solve F(x) = .25
* For our problem, we found F(x) =  , so we get 
* **> qexp(p = .25, rate = 1/5)**
* **[1] 1.43841**
* This is interpreted as “About 25% of subjects in this population survive less than 1.44 years after diagnosis.”
* Our median = **population median, b/c we’re dealing w/ population quantities.**
* A probability model connects data to a population using assumptions, so this median = the **estimand**, + the sample median = the **estimator** of the estimand

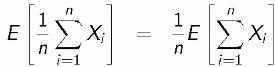
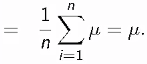
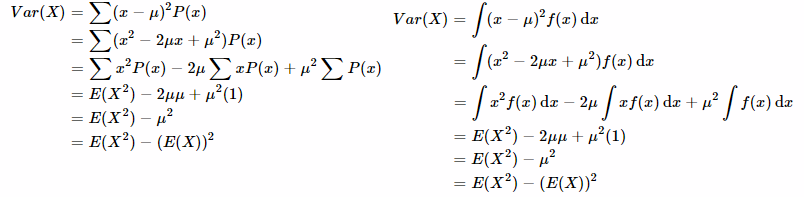
**Expected Values, Variances, Standard Deviation**

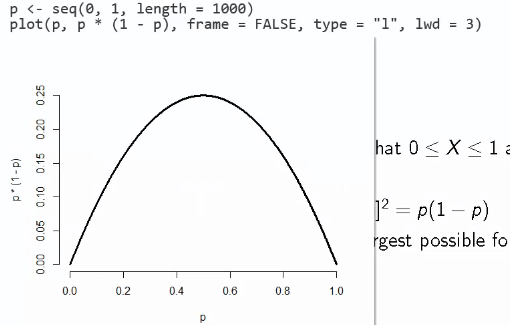
* **Expected values** =ways of describing probability distributions similarly to how quantiles do
* **Expected value/mean** of a random variable = center of random variable’s distribution
* For a **discrete** random variable X with PMF = p(X), the expected value is **E[X] =** over all possible values of x
* The x in front of p(x) makes it such that the sum of p(x) over all x no longer equals one
* This x makes the sum an expected value
* **E[X] =** represents the **center of mass** of a collection of locations and weights, **{x, p(x)}**
* Ex: Coin flip w/ random variable X = 0 or 1, corresponding to H or T, respectively
* E[X] = 0.5 \* 0 + 0.5 \* 1 = 0.5
* Geometrically, this would be obvious since 2 equal weights (0.5) are space at 0 and 1, so the center of mass is at ½
* This demonstrates that expected value E[X] does NOT have to be a value that the random variable can take
* Ex: Rolling die where X = number we roll
* E[X] = 1\*(1/6) + 2\*(1/6) + 3\*(1/6) + 4\*(1/6) + 5\*(1/6) + 6\*(1/6) = 3.5
* Still geometrically makes sense = 6 bars, all equal height, at position 1, 2, 3, …, etc., balanced out
* For a **continuous** random variable X with density (PDF) = ***f***, we have E[X] = 
* The t in front of f(t) makes it such that the AUC of f(t) over all values of t no longer equals one
* This t makes the sum an expected value
* Ex: Consider a density where f(x) = 1 for x between 0 and 1 (0 when x < 0, constant at 1 for ) <= x <= 1, and 0 for x > 1)



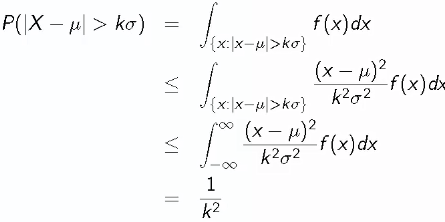
* Verify that this is valid: This is just a square w/ side = 1, so we don’t even need to integrate to get the area = 1
* It is also positive everywhere
* Since X follows this density, it’s expected value is (subbing in 1 for density f(x)):



* geometrically makes sense when balancing out between 0-1
* This is a simple, but very important density = **the standard uniform density** = represents that any value between 0-1 is equally likely
* The **expected value operator** is a **linear operator**
* Assume a and b are NOT random (i.e. #’s we can plug in, like 5) and X and Y are 2 *random* variables, so then:
* **E[aX + b] = aE[X] + b E[X + Y] = E[X] + E[Y]**
* The above works out the way it does *because* the expected value is a *linear operator*
* If g is some general function that is NOT linear, the generally E[g(X)] != g(E[X])
* Ex: in general, E[X^2] != E[X]^2
* Could happen w/ some specific values of g and some specific random variables
* Here, X = a random variable, X^2 = the random variable obtained by squaring X
* If X is a die roll, it can take values 1-6, X^2 can take values {1,4,9,16,25,36}, each w/ probability = 1/6
* So, E[X^2] = expected value of the squared random variable
* But, E[X]^2 = what we get if we first calculate E[X] and *then* square that result
* *THIS 2 THINGS ARE NOT EQUAL* (unless the random variable isn’t random
* General rule: If g is NOT a linear function, we can’t just commute expected values outside of g to the inside of g, but you can always do it if g is indeed a linear function
* **Expected value rules hold no matter what constitutes X and Y (discrete, continuous, a mix, etc.)**
* Ex: 1) Flip a coin, X, where 0/1 = H/T and 2) simulate random uniform variable Y between 0-1. Find the expected value of their sum (sum of coin flip and a uniform random variable):
* E[X+Y] = E[X] + E[Y] 🡺 expected value of a coin flip + expected value of a uniform random variable (both of which we’ve done before) 🡺 ½ + ½ = 1
* Ex: Roll a die twice and find the expected value of the average, Y
* X1 and X2 = the results of the 2 rolls so:
* E[Y] = find distribution of the average of the two die rolls 🡺 E[X1+X2/2] = ½(E[X1] + E[X2]) = ½(3.5 + 3.5) = 3.5
* The probability distribution of 1/6 for each die side *implies* a distribution onto the average of 2 die rolls
* Smallest number it can take is (1+1)/2 = 1 and the largest would be (6+6)/2 = 6, but w/ different values in between
* Not all probabilities are the same now, since P(1) and P(6) both are now 1/36, and those in between are higher
* *This rule for expected value of averages of collections of random variables from the same distribution extends to larger collections*
* Ex: Let Xi for i = 1, …, n be a collection of random variables, each from a distribution w/ mean = **μ**
* Expected value of the sample average of the random variables Xi =  🡺 1/n is pulled out b/c its not random 🡺 🡺 expected value commutes across the sum to all random variables Xi 🡺  🡺 expected value of each random variable Xi is, itself, **μ,** so we added up **μ** *n* times and then divided by *n,* giving **μ**
* This means that **it doesn't matter what the distribution of the individual Xi’s are, the distribution of the mean of the Xi’s, has the same mean as the individual means**
* **To summarize, the expected value of the sample mean *is the population mean it’s trying to estimate***
* The population mean of the distribution of the sample mean of n observations is exactly the population mean it’s trying to estimate
* When this happens (the expected value of the **estimator** is indeed the **estimand**), this is good and we say the estimator is **unbiased**
* *Sample means are unbiased estimator of population means (assuming all X’s are from distributions with mean = μ, the value we’re trying to estimate)*
* **Variance** of a random variable is another expected value property of a distribution, and it measures how spread out the distribution is
* For random variable X w/ mean = **μ,** variance of X is **Var(X) = E[(X – μ)^2)] =** expected value (average) of the squared difference between a value of the random variable and the mean
* Densities w/ higher variance = more spread out, and vice versa
* If we were to take a random variable X and figure out its standardized distribution (subtracting the mean from each value of X to get new **μ = 0,** but w/ the same distribution shape w/ shifted values for X**),** a convenient computational form would be **Var(X) = E[X^2] – E[X]^2** where*E[X]^2 is just μ^2*
* *If discrete, just use* *but w/ x^2 multiplied onto p(x) instead of x*
* *If continuous, just use*  *but w/ t^2 multiplied onto f(t) instead of t*
* EXERCISE: Show that **Var(X) = E[(X – μ)^2)] = Var(X) = E[X^2] – E[X]^2** via expansion and the expected value rules:
* **E[(X – μ)^2)] 🡺 E[(X^2) – 2μX + μ^2)] = ∑(x^2) – 2μx + μ^2)p(x) 🡪 (∑(x^2) – ∑(2μx) + ∑(μ^2))p(x) 🡺 (∑(x^2)p(x) – ∑(2μx)p(x) + ∑(μ^2)p(x)) 🡺 (∑(x^2)p(x) – 2μ∑(x)p(x) + μ^2∑(p(x)) 🡺 (∑(x^2)p(x) – 2μ(μ) + μ^2∑(p(x)) 🡺 (∑(x^2)p(x) – 2μ(μ) + μ^2(1) 🡺 (∑(x^2)p(x) – 2μ^2 + μ^2 🡺 (∑(x^2)p(x) –μ^2 🡺 E[X^2] – E[X]^2**
* 
* The variance operator is NOT linear, which is a bit inconvenient 🡪 i.e. if a is constant, we can say Var(aX) = a^2\*Var(X) (if you pull a random variable out of the variance, it gets squared)
* **Standard deviation** = squared root of the variance 🡺 good b/c the SD has the same units as the random variable, making it easier to use to describe the data (more interpretable + applicable)
* **Ex:** Calculate sample variance from a die toss
* **E[X] =** 3.5, so **E[X^2]** = 1^2\*1/6 + 2^2\*1/6 + 3^2\*1/6….. = 15.17
* **Var(X) = E[X^2] – E[X]^2 =** 15.17 – 3.5^2 = 15.17 – 12.25 **– 2.92**
* **Ex:** Calculate sample variance from a coin toss where P(H) = *p*
* **E[X] =** T\*not p + H\*p 🡺 0\*(1-p) + 1\*p = p 🡺 so **E[X^2]** = 0^2 + 1^2 = 1\*p = p so **E[X^2] = E[X] = p**
* **Var(X) = E[X^2] – E[X]^2 =** p – p^2 = **p(1-p)**
* Suppose we have a random variable X such that 0 <= X <= 1 and E[X] = p
* In this case, X^2 <= X, such that E[X^2] <= E[X] = p which means:
* **Var(X) = E[X^2] – E[X]^2 <= E[X] – E[X]^2 = p(1-p)**
* This variance formula is maximized when p = .5 (*the most variable a coin flip can be is a fair coin)*



* So, the most variable a random variable can be, in general, is if we put all of its mass into 2 endpoints and equally distribute it between the endpoints
* Or, the **Bernoulli (binary) variance is largest for random variables bounded between 0 and 1**
* We know “bigger” variances mean a more spread out distribution, but what does “bigger” mean?
* We *could* figure this out in the context of a single, specific distribution (especially w/ Gaussian/bell-shaped density)
* **Chebyshev’s Inequality** = general rule for interpreting variances for *all* distributions
*  == **probability a random variables is k SD’s from its mean or more is less than or equal to 1/k^2**
* 2δ 🡪 25% = probability a random variable is > 2 SD’s away from its mean is 25% or less
* Others: 3δ 🡪 11% 4δ 🡪 6%
* **Note:** This is only a *bound* on the probability statement + not an equality (it is the *worst* it could actually be)and actual probability could be quite a bit smaller
* Chebyshev’s Inequality is also quite easy to prove (in the *continuous* case)

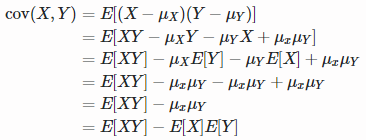
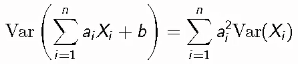
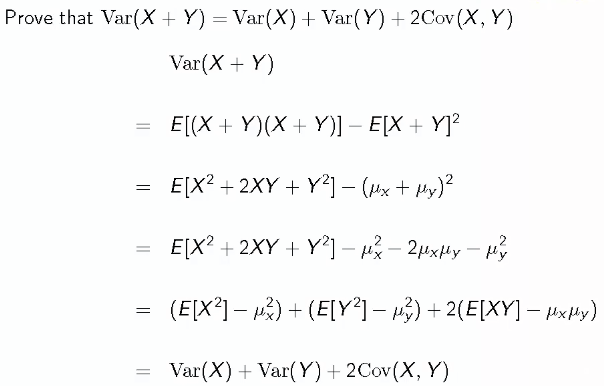


* 1st, integrate f(x) over set of x wherein x is more than k SD’s from the mean
* We note  so  so when we multiply f(x) by this, we make the integral bigger, so is not larger than 
* Then we make it even bigger by integrating over all real numbers rather than our bounded x values like in the 1st step (since (x – μ)^2 is strictly positive)
* Then notice, we can factor our a scalar 🡺 , and this resulting integral is *exactly the definition of the variance* 🡺 , and the δ^2 value cancels the Var(X), leaving us with 1/k^2 in that final step
* Ex: IQ’s are often said to be distributed w/ μ = 100 and δ = 15.
* Find P(>160 or < 40) 🡺 this is P(> 4δ in either direction)
* Using Chebyshev’s Inequality, we know this P is <= 1/4^2 🡺 <= 1/16, so P is no larger than 6%
* But, we know that IQ distributions are often cited as bell-curved/ Gaussian, so this calculated bound is very conservative
* The probability of drawing a random variable that’s more than 4δ from the mean from a Gaussian distribution ends up being on the order 10^-5 (1 thousandth of a percent), which does not violate Chebyshev’s Inequality, just shows that its conservative
* Ex: Buzzphrase is industrial quality control = Motorola’s “Six Sigma”, whereby businesses are suggested to control extreme events/rare defective parts
* Chebyshev’s Inequality states the probability of a Six Sigma event is < 1/6^2, or about 3% (over ALL distributions)
* In actuality, assuming a Gaussian distribution, the true probability of a Six Sigma event is on the order of 10^-9 (one ten millionth of a percent)

**Random Vectors and Independence**

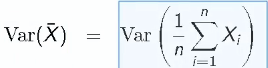
* **Random vector =** ordinary vector containing random variables (a collection of random variables)
* Ex: (X,Y) is a random vector containing random variables X and Y
* Just like how individual random variables have density, mass, + distribution functions that govern their probabilistic behavior, random vectors have **joint** density, mass, + distribution functions that govern *their* probabilistic behavior
* **Joint density/Joint PDF = *f*(x, y)** satisfies ***f* > 0** (on 2D plane = height is above horizontal plane) + the integral over the xy-plane **∫∫f(x, y) dxdy = 1** (extension of 1D random variable PDF)
* *Works just like univariate PDF 🡪 AUC (volumes/integrals) correspond to probabilities*
* For *discrete* random variables, *f* is a **joint PMF** *that maps possible combined values of x and y to probabilities*
* To satisfy definition of being a join PMF, **f > 0** for all possible combos of x and y, and then **∑∑ f(x, y) = 1** (sum over all possible combos of x and y)
* *Works just like univariate PMF in that the sums of collections of possible values of x and y correspond to probabilities for that collection*
* We’re focusing on 1 kind of joint densities = **independent random variables** where f(x, y) = f(x) \* g(y)
* *Events* A and B = **independent** if **P(A ∩ B) = P(A)P(B)**
* If A is independent (unrelated) of B, then
* Acomp = independent of B
* A = independent of Bcomp
* Acomp = independent of Bcomp
* *Random Variables X and Y* are independent if, for any 2 sets A and B,:
* **P([X ε A] ∩ P[X ε B]) = P(X ε A)P(X ε B) 🡺** *intersection of the probability that X lies in A and the probability that X lies in B = their product*
* Ex: P(HH) in 2 consecutive coin flips?
* A = {head on flip 1} 🡺 P(A) = ½ B = {head on flip 2} 🡺 P(B) = ½
* (A ∩ B) = {heads on flips 1 *AND* 2)  **P(A ∩ B) = P(A)P(B) 🡺** ½\*½ = ¼
* Ex: Vol 309 of *Science* reports on a physician on trial for expert testimony in a criminal trial
* Based on estimated prevalence of SIDS = 1/8543, Dr. Meadow testified that the probability of a mother having 2 children w/ SIDS = (1/8543)^2, so a mother on trial was convicted of murder
* The principal mistake here == assuming probabilities of having SIDS w/in the same family are independent
* P(A1 ∩ A2) is not necessarily = P(A1)P(A2) in this case
* Biological processes that have a believed genetic/familiar environmental component tend to be **dependent** w/in families
* And, the estimated prevalence used above was obtained from an *unpublished* report on *single cases* (i.e. having no info on recurrence of SIDS w/in families)
* \*\*\*\*\**If a collection of random variables X1, X2, …, Xn are independent, then their* ***joint distribution/density function*** *is equal to the product of their individual* ***densities/mass functions***
* *If fi if the density for random variable Xi, then we have* 
* Most common model used = X1, X2, …, Xn, where they are all from the same distribution
* In the instance that f1 = f2 = , …., = fn (same distributions), then the Xi are **iid** = independent + **identically distributed**
* **iid variables** are the default model for random samples, and many important theories of stats are founded on assuming the variables are iid
* Ex: each roll of a die is a draw from the uniform distribution of #’s 1-6
* When we model a process as if its iid, we are saying it’s like rolling a die for each variable we’re modeling from some population-level distribution
* But this pretty much never actually the case (can rarely guarantee our sample are random draws from some population distribution over and over)
* So, it’s just a statistical model used to simplify calculations + discussion
* But one must be aware of the fact that it is an enormously simplifying assumption
* Ex: Flip a biased coin w/ probability of success = *p* a total of *n* times, find the join density of the collection of outcomes
* Recall each coin flip = Bernoulli random variables w/ P(success) = p, + these random variables are iid w/ densities = 
* Therefore the joint PMF =  = p to the number of heads times 1 minus p to the number of tails
* Ex: 3 flips, (x1,x2,x3,x4) = (1,0,1,1) 🡺 *order doesn’t matter*
* 

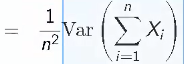
**Correlation**

* For 2 random variables, the **covariance** is **Cov(X,Y) = E[(X – μx)(Y – μy)] = E[XY] – E[X]E[Y]**
* **PROOF** 🡺 
* **Cov(X,Y) = Cov(Y,X)**
* **Cov(X,Y)** can be negative or positive
* **|Cov(X,Y)|**  **(**thanks to an application **Jensen’s’ inequality**)
* 
* **Correlation** between X and Y is Cor(X,Y) =  = covariance divided by the product of the standard deviations of X and Y
* This normalizes the covariance such that correlation is **-1 <= Cor(X,Y) <= 1**
* ***Cor(X,Y) = +/- 1 if and only if X = a + bY for some constants a and b (only if random variables X and Y are linearly related)***
* **Cor(X,Y) is unit-less 🡪** covariance has units X times units Y, standard deviations have unit X and units Y, so they cancel in the correlation equation
* X and Y are **uncorrelated** if Cor(X,Y) = 0
* Converse if NOT true 🡪 things can be uncorrelated + NOT be independent
* If independent, then they’re uncorrelated, but if they’re uncorrelated, they’re not necessarily independent
* *But, these statements always agree when using Gaussian random variables*
* X and Y are more **positively** **correlated** = Cor(X,Y) is closer to 1, X and Y are more **negatively** **correlated** = Cor(X,Y) is closer to -1,
* These are all descriptions of a POPULATION QUANTITY, *not* a sample quantity
* Correlation = summary how of related joint random variables are from a distribution = summary of a population quantity
* Since this gives a population quantity summary, we want a sample version that we use to estimate the population quantity = the **sample correlation**
* We want to try to estimate the population correlation via the sample correlation + a probability model (or any sample + population values)
* Now, if we have a collection of random variables X1 to Xn = , *when the X’s  are uncorrelated, then* 
* *“*the variance of the sums of the X’s times some constant a plus some constant b = the constant a^2 times the variance of the individual X’s
* *If b = 0, the variance of the sum of independent X’s equals the sum of the individual variances*
* We know constants get pulled out of variances and then squared (a^2) and that if we have a random variable + we shift it by constant b, variance doesn’t change, the density is just moved
* If the **’s ARE correlated, then the variance depends on the covariance such that  
* 
* 
* SUMMARY
* If a collection of random variables are uncorrelated, then the sum of the variances = the variances of the sum
* i.e. sums of variances tend to be useful, rather than sums of SD’s (SD of the sums of individual random variables = square root of the sum of the variances, NOT the sum of the SD’s)

**Variance Properties and Sample Variance**

* \*\*\*\***Variance of the sample mean of a collection of iid random variables = δ^2 / n**
* Ex: collection Xi are iid w/ variance of the distribution they’re drawn from is = **δ^2:**

 🡺 variance of sample mean = variance of sum of Xi values divided by n

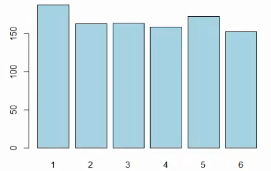
 🡺 pull out constant 🡪 which must then be squared

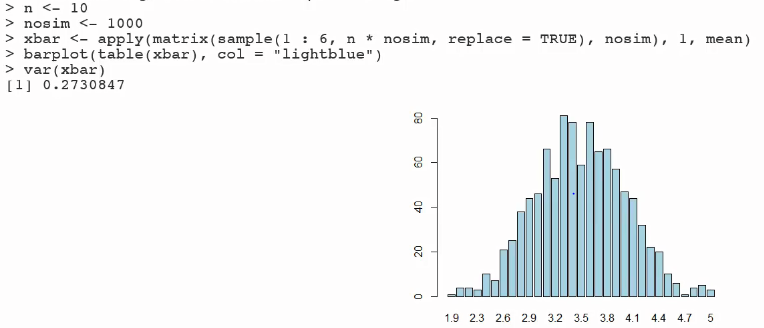
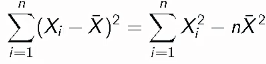
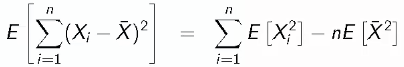
🡺 variance of the sum = sum of the variances b/c Xi = independent (iid) + hence uncorrelated

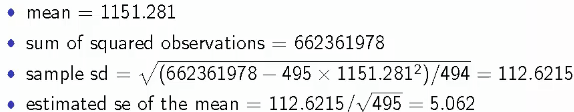
 🡺 b/c Xi’s are iid, they variances are all the same (all from the same distribution) and are = **δ^2,** and this is done *n* times

🡺 multiply and simplify

* This proof says that, to know the variance of the distribution of averages of say, 10 random variables from a distribution, *we don’t need to know what the distribution of averages actually is*
* All we have to know is what the variance was from the original distribution that the individual observations were drawn from was (**δ^2**), which we then just divide by *n*
* To get SD, just take the original population’s SD + divide by the square root of *n* (**standard error**)
* This is important b/c eventually we want to tie back all these population model ideas to data
* If we have a bunch of things we’re willing to model as iid, then we get multiple draws from the distribution of individual observations (all Xi’s are separate draws from the original distribution)
* We can then estimate things like **δ^2**
* But w/ only 1 sample mean, say from a sample of 100, so we don’t have anything to empirically estimate the variance of sample means of 100 variables (don’t have repeated samples of 100 variables)
* This proof tells us that *we don’t need repeated samples*, just the original population’s distribution’s variance (which we *can* estimate)and divide it by n
* Ex: The variance of a die roll (values 1-6, each w/ equal likelihood = 1/6) was 2.92
* **E[X] =** 3.5, so **E[X^2]** = 1^2\*1/6 + 2^2\*1/6 + 3^2\*1/6….. = 15.17
* **Var(X) = E[X^2] – E[X]^2 =** 15.17 – 3.5^2 = 15.17 – 12.25 **– 2.92**
* So if we roll a die over and over and over, we get the following distribution:
* 

* See our 1000-sample variance is almost 2.92, because the sample distribution’s variance of many, many die rolls estimated the population distribution (equal likelihoods for 1-6)
* Now, suppose we roll 10 dice, take their average value, and repeat this process 1000 times.
* Now our distribution is no longer uniform for values 1-6, but would have min = 1 (average of ten 1’s = 1), max = 6 (average of ten 6’s = 6)
* 
* What’s the variance of *this* distribution?
* Seems hard 🡪 need the distribution of the average of 10 die rolls.
* But the calculation above (R code) tells us *we don’t need that*
* We know the variance of the distribution of individual die rolls (2.92), so the variance of the distribution of averages of 10 die rolls = **2.92 / 10 = .292**
* For our sample, it’s **.273**
* So, when the collection of Xi’s are independent w/ a common variance , then:
* **Population δ** = talks about how variable the population is
*  = **standard error** of the sample mean
* *How variable the population of averages of size n from the original population is*
* Ex:δ = how variable IQ scores are, SE for n = 10 🡺 how variable averages of 10 IQ scores are
* *SE of the sample mean = SD of the distribution of the sample mean*
* δ = SD of the distribution of a *single* observation
* Easy way to remember = **sample mean must be less variable then a single observation, therefore it’s SD is divided by the square root of n**
* Sample variance =  🡺 an estimator of **δ^2 (**population variance)
* Since we’re dividing by n-1 and not n, it’s *nearly* the mean of the squared deviations of sample values from the sample mean
* A quicker version (for calculations) of the numerator = 
* Why divide by n-1 and not n (for large samples, it’s actually irrelevant)?
* Recall property of **unbiasedness** 🡪 *the sample variance is unbiased* = the expected value of some statistic is equal to the quantity it’s estimating
* Sample variance = function of our observed data/random variables, so it is a statistic/random variable itself, so it has a distribution
* This distribution has a variance + a mean, and the mean of this distribution ends up being **δ^2,** if we have do the calculation where we divide by n-1
* Prove that the expected value of the numerator of the sample variance = (n-1)\***δ^2** (same as showing the sample variance is unbiased since we divide both sides by n-1)
*  🡺 use shortcut formula from above
*  🡺 recall the fact that the variance is the expected value of a squared random variable - minsut the squared expected value
* 
* *Shift this formula around to say the expected value squared = (variance + mean)^2*
* 
* variance of a collection Xi = δ^2of the collection
* variance of a the mean of the collection = δ^2 / n, since the mean itself is another random variable
*  🡺 δ + μ is just a constant, and we get n of those
* 
* This proof says that **the expected value of the sample variance is in fact the quantity it’s trying to estimate if in fact we divide by n-1 instead of n (this is why we divide by n-1 and not n)**
* **We don’t know the population mean** μ, and if we knew it, instead of plugging in X\_bar into the sample variance we’d plug in μ + get the deviation of observed observation around the population mean rather than from the sample mean
* The idea **= we lose a degree of freedom by using X\_bar instead of** μ
* Bias-Variance Trade-off 🡪 we have an unbiased estimator by dividing by n-1, not n
* If we divide by n, + calculate the expected value of the sample variance if it was calculated w/ n in the denominator instead of n-1.
* i.e. what is the expected value of (n-1/n)\*δ^2
* You can calculate that very easily, it is not S^2, but it is quite close to it, so it's a biased estimator
* Which of the 2 estimators, S^2 calculated w/ n-1 in the denominator the variance w/ n in the denominator, has a lower variance?
* Remember sample variance = a random variable, so it has a distribution, which itself has a variance.
* Which of the 2 calculations (dividing by n or by n-1) results in a smaller variance of that distribution? = How precise is your estimate of the variance?
* **Sample variance divided by n has a slightly lower variance than the sample variance divided by n-1**
* *In this case, we divide by n-1 b/c we want unbiasedness*, but we get p w/ a slightly greater variance
* If we divide by n, we get w/ a slightly lower sample variance but it's slightly biased
* Some extremely well-established statisticians would prefer lower variance, but pretty much every introductory statistics textbook divides by n-1.
* So, divide by n-1 when calculating the sample variance
* People have a tendency to confuse this n w/ the n we divide by for **standard error of the mean**.
* Suppose you have a bunch of observations you're willing to model as **IID** w/ population mean μ + population variance δ^2
* The **sample variance, S^2, estimates the population variance, δ^2**.
* Calculation of S^2 involves dividing by n-1 + standard error of the mean = δ / square root of n.
* So, **S / square root n** estimates standard error of the mean 🡺 
* We've already divided S^2 by n-1, + then square rooted, + then divided by an *additional* square root of n to get SE of the mean.
*  = sample standard error of the mean (an estimator of the population standard error of the mean
* Ex: Study of a subset of 465 organo-lead workers w/ the following summaries for total brain volume (TBV) = measure of brain volume inside the skull in cm cubed:



* If willing to assume these workers are an **IID** draw of organo-lead workers from a population we're interested in, 1151 = an estimate of that population mean.
* The sum of the squared observations 662361978, so we use it to get sample SD, which itself it used to get the sample SE of the mean
* **SD = 112 describes the variance of the population of brain volumes of organo-lead workers =** a direct estimate of my sample variation
* Also, if we view the data as a sample from a population of organo-lead workers, this attempts to estimate the population SD of that distribution.
* We can use Chebyshev’s Rule to interpret what the combination of the mean + SD say about brain volumes of lead workers in the population
*  == **probability a random variables is k SD’s from its mean or more is less than or equal to 1/k^2**
* SE = 5 is NOT longer talking about variation in TBV in the population, but instead talks about variation in *averages* of the sample of 495 workers.
* If we're willing to model our 495 workers as a draw from a population of workers, SE = 5 estimates the *distribution of averages of 495 draws of workers from that population.*
* SE = 5 talks about how variable *averages* of 495 TBV’s are.
* SD = 112 🡺 how variable TBV’s are of workers in sample, estimating for the population
* SE = 5 🡺 estimate of the population SD of *averages of 495 workers*.
* These sample quantities (sample mean, SD, SE of the mean) tell us things about out observed data.
* Then, there's the assumptions (ex: they're IID) that help to try + connect these values to a population to possibly generalize results from this data to a population of organo-lead workers
* Ex: Wanted to use this data to inform policy + these #’s would be estimates of population quantities
* Dividing the SD by the square root 495 tells us things about how variable this mean is relative to the variability in the population.