***COURSERA: STATS W/ R SPECIALIZATION***

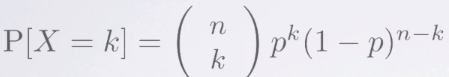
***COURSE 4 – Bayesian Stats***

**WEEK 2 – Bayesian Inference**

* 1. ***Continuous Variables and Eliciting Probability Distributions***

**From the Discrete to the Continuous**

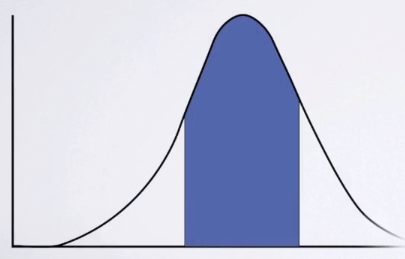
* B**inomial random variable 🡺** # of heads in 10 coin tosses can only take discrete # of values 🡪 0-10
* For a binomial random variable in which probability of success = p + # of trials = n, the probability the random variable takes the value k for k = 0, k = 1, up to k = n 🡺 **n choose k** times p to the power k \* 1-p raised to the power n-k.
* This = **Probability Mass Function (PMF)** for the binomial.



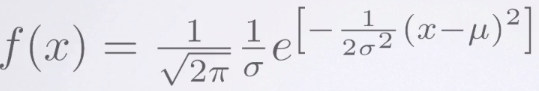
* The PMF can be visualized as a histogram w/ AUC = 1 + the area of each bar = probability of seeing a **binomial random variable** whose value is equal to the x-value at the center of the bar’s base.



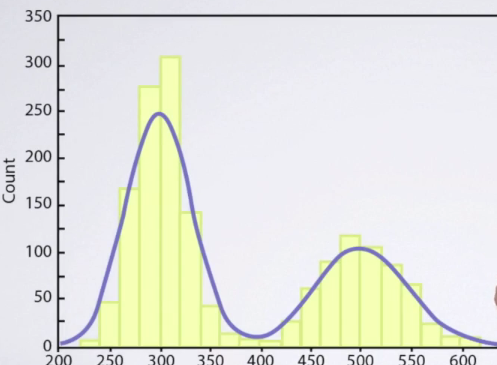
* In contrast, **Normal/Gaussian distribution** can take ANY numerical value between -Infinity + Infinity.
* Since it can take a **continuum** of values, it is a **continuous random variable**.
* *In general, if the set of possible values a random variable can take are separated points, it’s is a* ***discrete random variable****.*
* *But if it can take any value in some possibly infinite interval, it is a* ***continuous random variable.***
* When a random variable is discrete, it has a PMF which tells us the probability the random variable takes each of the possible values.
* But when a random variable is continuous, it has probability = 0 of taking any *single* value + we can only talk about the probability of a continuous random variable lined *within some interval.*
* Suppose heights are approximately normally distributed 🡪 probability of finding someone exactly 6 ft. tall, + 0.0000 inches tall for an infinite #’s of 0s after the decimal point = 0.
* But we can easily calculate probability of finding someone between 5'11" + 6'1" tall.
* Continuous random variables have **Probability Density Functions** (**PDF**) instead of PMFs
* Probability of finding someone whose height lies between 5'11" and 6'1" = AUC for the PDF curve for height between those 2 values.



* Ex: A normal distribution w/ mean = mu + SD = sigma has a PDF curve defined by:



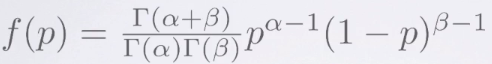
* Here, x = any value the random variable can take.
* Recall a PMF assigns the probability a random variable takes a *specific* value for the *discrete* set of possible values.
* The sum of those probabilities over *all* possible values *must* equal 1.
* Similarly, a PDF = any function of x that is *non-negative* + has AUC = 1.
* The PDF can be thought of as the limit of histograms made from a sample data.

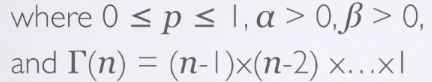


* As the sample size becomes infinitely large, the bin widths shrink to 0.
* There are infinite # of PMF'S + PDF'S.
* Some, such as the binomial + normal, are so important they have been given names.
* 3 more named continuous distributions 🡪 the **uniform, the beta + the gamma distributions**.
* A new discrete distribution = the **Poisson distribution**. Before closing,
* Key ideas
  + Continuous random variables exist + can take any value within some possibly infinite range.
  + The probability a continuous random variable takes a specific value = 0
  + Probabilities from a continuous random variable are determined by the density function that is non-negative + the area beneath it’s curve = 1 is one
  + We can find the probability a random variable lies between 2 values say as the area under the density function that lies between the 2 #’s

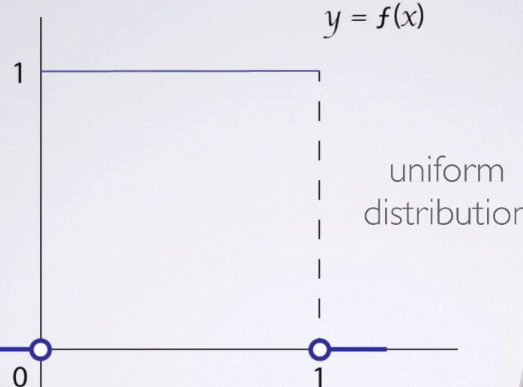
**Elicitation**

* Often, one has a **belief** about the distribution of one's data 🡪 may think your data come from a binomial distribution, + in that case you typically know the n (# of trials) but usually do not know p, the probability of success.
* Or you may think your data come from a normal distribution, but you only know the SD and not the mean, or do not know the mean nor SD
* Besides knowing the distribution of one's data, you may also have beliefs about the **unknown p** in the **binomial** or the **unknown mean** in the **normal**.
* Bayesians express their belief in terms of **personal probabilities** which encapsulate everything a Bayesian knows/believes about the problem.
* *But these beliefs must obey* ***the laws of probability*** *+ be consistent w/ everything else the Bayesian knows*
* You may know nothing at all about the value of p that generated some binomial data, in which case any value between 0-1 is equally likely
* You may want to make an inference on the proportion of people who would buy a new band of toothpaste.
* If you have industry experience, you may have a strong belief about the value of p but if new to the industry, you’d do nothing about p + any value between 0-1 seems equally to you
* This means your personal probability = the uniform distribution whose PDF is flat.
* Often, one knows quite a lot about which values of p are more than others.
* If you we’re tossing a coin, most people believe the probability of H is pretty close to 1/2.
* They know some coin are loaded + some may have two H or T + they probably also know coins aren't perfectly balanced.
* Nonetheless, before they start to collect data by tossing the coin + counting the # of H, their belief is that values of p near 0.5 are very likely + values of p near 0 or 1 are very unlikely.
* So a Bayesian will seek to express their belief about the value of p through a **probability distribution**
* A very flexible family of distributions for this purpose = the **beta family**.
* A member of the beta family is specified by 2 parameters, **α + β** (just as a member of the normal family is specified by the mean + the SD)

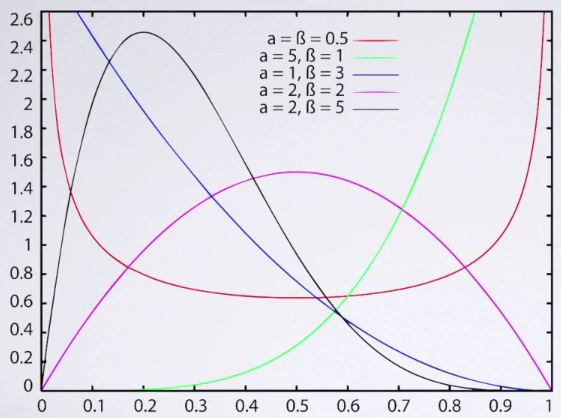




* For the **β**, we shall call these
* two parameters **α** and **β**. In this formula, note the **gamma** (**Ѓ**) functions.
* The **gamma function** is just a factorial at the bottom there 🡪 n-1 times n-2 times n- 3 all the way down to 1.
* When **α** = **β** = 1, one gets the member of the betafamily that is the flat line, which is also the PDF of the uniform distribution.

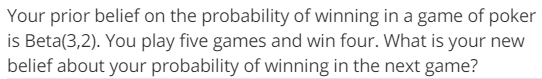


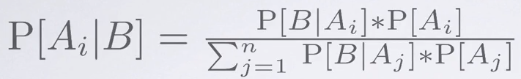
* If we take **α** = **β**, one gets a PDF that is symmetrical around 1/2



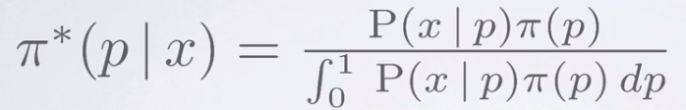
* + For large but equal values of **α** and **β**, the area under the **beta** density near ½ is very large.
  + These kinds of priors are probably appropriate If you want to make inference on the probability of getting H in a coin toss.
  + 
    - 
    - **LARGE**
* The betafamily also includes skewed densities, which are appropriate if one thinks p is likely to be nearer to 0 or nearer to 1.
* As we know **Bayes' rule** is a machine for turning once prior beliefs into **posterior beliefs**.
* W/ binomial data you start w/ whatever beliefs you may have about p, then observe data in the form of the # of H in say 20 tosses of a coin
* Bayes' rule tells you how that data should change your opinion about p.
* The same principle applies to all other inferences 🡪 start w/ your prior probability distribution over some parameter, then use data to update that distribution to become the posterior distribution that expresses your new belief.
* Bayes' rule ensures a change in distributions from prior to posterior is the uniquely rational solution.
* So long as you begin w/ a prior distribution that reflects your TRUE opinion, you can hardly go wrong
* **But, expressing that prior can be difficult**.
* There are proofs + methods whereby a rational + coherent thinker can self-illicit their true prior distribution, but these are impractical + people are rarely rational + coherent.
* The good news is that w/ the few simple conditions, no matter what prior distribution you choose, if you observe *enough data*, you will converge to an accurate posterior distribution.
* So, 2 Bayesians can start w/ different priors but observe the same data.
* As the amount of data increases, they will converge to the same posterior distribution.
* Key Ideas
  + Bayesians express their uncertainty through probability distributions.
  + One can think about a situation + self-elicit a probability distribution that approximately reflects their personal probability.
  + One's personal probability should change according Bayes' rule, as new data are observed
  + The **β** family of distribution can describe a wide range of prior beliefs.

**Conjugacy**

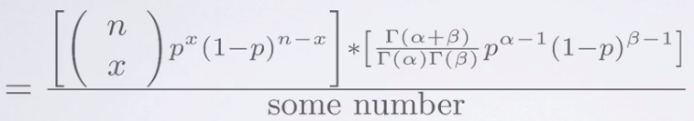
* Suppose you believe your data come from a binomial distribution w/ a known n, but an unknown p.
* Suppose your prior belief about p has the beta PDF w/ parameters (α, β)
* If you observe x successes in n trials, it turns out **Bayes' rule** implies your new belief about the density of p is ***also β,*** but now w/ parameters α + x and β + n – x = **beta(α + x, β + n – x)**
* This is an example of **conjugacy**, which occurs when your *new belief/posterior distribution is in the same family of PDF as your prior belief, but w/ new parameter values updated to reflect what you learned from the data.*
*  *=* 
* Why are the beta-binomials families conjugate?
* Recall the form of Bayes' rule used for *discrete* random variables.



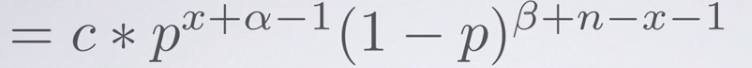
* This formula cannot apply to *continuous* random variables, such as the p w/ the beta prior, since the denominator sums over ALL possible values of the random variable, but the p can take any value between 0-1
* Need the version of Bayes' rule that applies to *continuous* random variables (analogous to the discrete form)



* The integral in the denominator is like a sum = some constant that ensures the total AUC curve/the PDF = 1.
* Also, note the 1st term in the numerator = probability of observing your data, given a specific value of p, + the 2nd term = density for your prior belief about p.
* In the beta-binomial case:

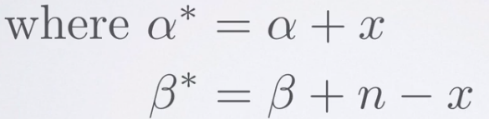


* The probability of observing x heads + n tosses, when the probability of success = p, is given by the 1st term in the numerator.
* If you have a beta prior w/ parameters α + β, the density for p = the 2nd term in the numerator.
* So, if we use the continuous version of Bayes' rule, we find the posterior distribution is given by the formula in the last line.
* In the numerator, we can collect the terms that involve p



* Everything else are just constants + must take the unique value needed to ensure the AUC between 0 and 1 is equal to 1





* With the above, we can find the answer w/out doing the integral
* Without **conjugacy**, one has to do the integral, which is often impossible to evaluate.
* That obstacle is the primary reason most statistical theory in the 20th century was not Bayesian.
* The situation didn't change until modern computing allowed researchers to compute integrals numerically.
* Key Ideas: There are some families w/ densities that are **conjugate pairs =** if the data come from the 1st of those families + your belief about the unknown parameter has a distribution from the 2nd family, then after observing your data, your new belief/posterior density is *also a member of the 2nd family*, but w/ *different parameter values*.
  1. ***Three Conjugate Families***

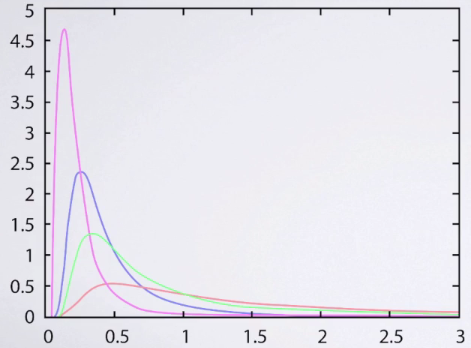
**Inference on a Binomial Proportion**

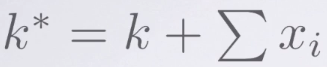
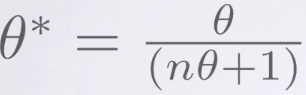
* Motivating example: Slightly simplified version of a real clinical trial in Scotland on if RU-486, a morning after pill being studied, was effective at preventing unwanted pregnancies.
* Had 800 women, each of whom had intercourse no more than 72 hours before reporting to a family planning clinic to seek contraception.
* 1/2 were randomly assigned to the standard contraceptive (large dose of estrogen + progesterone) + 1/2 were assigned RU-486.
* Among the RU-486 group = no pregnancies + among standard therapy, 4 became pregnant
* Statistically, one can model these data as coming from a binomial distribution.
* Ex: Coin w/ 1 side labeled standard therapy + the other RU-486.
* Coin was tossed 4 times + each time landed on standard therapy.
* A **frequentist** would set up a null w/ “RU-486 side has p >= 1/2” + the alternative w/ p < 1/2
* **Significance probability** = chance of obtaining no RU-486 outcomes when p = 1/2 = ½^4 = .0625
* Since the significance probability .0625 > 0.05, the frequentist would conclude there was no reason to reject the null + would conclude RU-486 is not superior to standard therapy.
* A **Bayesian** may self-illicit their beliefs about the drug + decide they have no prior knowledge about efficacy of RU-486 at all.
* This would be reasonable if, for example, it were the 1st clinical trial of the drug.
* In that case, they’d be using the uniform distribution on the interval from 0-1, which corresponds to the **beta(1,1) density**.
* From **conjugacy**, we know that since there were 4 “failures” for RU-486 + no successes, **posterior probability** of RU-486 child = beta w/ parameters (1 + success) + (1 + failures) = (1+0) and (1+4)
* This is a beta that has much more area near p = 0.
* The mean of a beta w/ parameters alpha + beta, **beta(α, β),** is **α** / (**α + β),** so the Bayesian now believes the unknown p (probability of an RU-468 child) is about = 1 / (1+5) = 1 / 6.
* Before she saw the data, the Bayesian's uncertainty expressed by their SD was 0.71.
* After seeing the data, it was much reduced 🡺 posterior SD = just 0.13.
* W/ calculus, we find that this Bayesian now believes their posterior probability that p < 1/2 = .96875
* Before seeing data, they thought there was a 50-50 chance RU-486 is better, but now they think there's about a 97% chance RU-486 is better.
* Suppose a 5th child were born, also to a mother who received standard chip therapy.
* Now the Bayesian's prior = **beta(1,5)** + the additional DP further updates a new posterior **beta(1,6)**
* As data comes in, the Bayesian's previous posterior becomes their new prior, so learning is self-consistent.
* 
* beta(3, 2) **🡺**  (3 + success) + (2 + failures) = (3+4) and (2+1) = **beta(7,3) = .7**

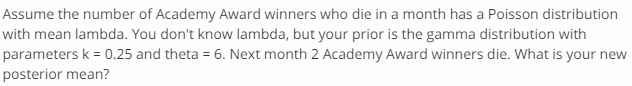
**The Gamma-Poisson Conjugate Families**

* A 2nd important case = **Gamma-Poisson conjugate families** where the data come from a **Poisson distribution**, + the prior + posterior are both **gamma distributions**
* The **Poisson random variable** can take any *non-negative* integer value up to infinity + is used in describing count data (# of independent events that occur in a fixed amount of time, a fixed area, or a fixed volume)
* The Poisson distribution has been used to describe the # of phone calls one receives in an hour, # of pediatric cancer cases in the city to see if pollution has elevated cancer rate above that of in previous years or for similar cities.
* It is also used in medical screening for diseases, such as HIV, where one can count # of T-cells in the tissue sample.
* The Poisson distribution has a single parameter, **lambda**, which is both the mean + the variance of the Poisson random variable.
* The PMF for the Poisson:



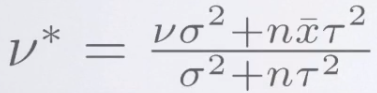
* This gives the probability of observing a random variable = K
* Obviously, lambda must be > 0 since it is both the variance, + also the average of the count.
* 
  + **1-(((1.5^1)/1!)\*E^(-1.5) + ((1.5^0)/0!)\*E^(-1.5)) = 0.44217459962892547**
  + 
* von Bortkiewicz used the Poisson distribution to study the # of Prussian cavalrymen who were kicked to death by a horse each year.
* Count data over a year, + events are probably independent, so Poisson model makes sense
* He had data on 15 cavalry units for 20 years between 1875 + 1894, inclusive w/ total # of cavalrymen killed by horse kick = 200.
* One can imagine a Prussian general might want to estimate lambda = average # per year, per unit in order to see whether some educational campaign about best practices for equine safety would make a difference.
* Suppose the general is a Bayesian whose introspective elicitation leads him to think **λ**= ~.75 + his uncertainty in this belief is expressed as a SD = 1.
* General will need to express his prior as a *member of a family* ***conjugate*** *to the Poisson*.
* It turns out that this family consists of the **Gamma distributions** = describe continuous non-negative random variables.
  + We know value of **λ** in the Poisson can take any non-negative value so this fits
* The **gamma family** is pretty flexible w/ a wide range of gamma shapes.
  + - * + 
* The PDF for the gamma is indexed by 2 parameters, K + theta (some books parameterize it in a slightly different way)
* For our parameterization, the mean of the gamma = **K\*ϴ** + the SD = **ϴ\*Sqrt(K)**
* So, the general's prior is that there is a gamma such that **K\*ϴ** = 0.75 + **ϴ\*Sqrt(K) = 1**
* Solving these simultaneous equations shows the general's priors 🡺 K = 9/16 + the his **ϴ** = 4/3
* For the **Gamma-Poisson conjugate family**, suppose one observes data that’re Poisson distributed w/ values X1, X2, … Xn
* Then, in the same way we recognize the kernel of the beta distribution In the integral form for
* Bayes' rule for the beta-binomial family, *we’d recognize the kernel of the gamma when using the gamma-Poisson family*.
* It turns out that the posterior gamma has updated parameters K\*+ **ϴ\*** where K\* = K + Sum of the observed values, + **ϴ\* = ϴ / (n\*ϴ** + 1)

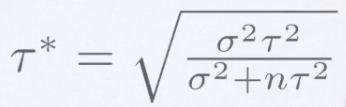
* For this dataset, n = 15\*20 = 300 observations, + the total # of cavalrymen who died = 200.
* Therefore, average # of Prussian cavalry officers killed by hoofs of their horses follows a *gamma distribution w/ parameter K\* = (9/16) + 200 = 200.5625 +* **ϴ\* =** *(4/3)/((300\*4/3) + 1) = 0.0033.*
* Before he saw the data, general believed **λ** was about .75, Now, he believes it's about **200.5625\*0.0033 = 0.67** + his uncertainty about **λ**, expressed as a SD, was = 1 + is now **0.0033\*sqrt(200.5625) = uncertainty has shrunk to 0.047**
* 
  + 1 month, 2 deaths, so n = 1
  + *n*= # months in new data, + ∑*xi* = # deaths in that period
    - If k Academy Award winners died in the next n months, these would be the new k, n
    - If told k winners died the following year, then you’d have n=12.
  + Average # of academy award winner deaths follows a *gamma distribution w/ parameters of K\* = (.25) + 2 deaths = 2.25 +* **ϴ\* =** *(6)/((1 month\*6) + 1) = 6/7 = .8571*
  + New posterior= **2.25\**.8571* = 1.928** + his uncertainty about **λ**, expressed as a SD is now ***.8571*\*sqrt(2.25)= uncertainty has shrunk to 1.286**

**The Normal-Normal Conjugate Families**

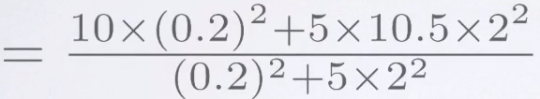
* There are other congregate families such as the **normal-normal pair**.
* If data come from a normal distribution w/ known SD (σ) but unknown mean (μ) + if your prior on μ has a normal distribution w/ **self-elicited mean, nu (ν)** + **self-elicited standard deviation, tau (τ),** then your *posterior* density for μ, after seeing a sample of size **n** w/ sample mean x\_, is ALSO normal.
* As a practical matter, one often does not know σ (SD of the normal from which the data come).
* In that case, you could use a more advanced conjugate family, but there are cases in which it is reasonable to treat σ as known.
* 1 example: Analytical chemist whose balance produces measurements that’re normally distributed w/ μ = to the true mass of the sample + σ estimated by the manufacturer balance + confirmed against calibration standards provided by the National Institute of Standards + Technology
* For the normal-normal conjugate families, assume the prior on the unknown μ is normal w/ mean = **nu (ν)** + SD = **τ**.
* Also assume the data X1, X2... Xn are independent + come from a normal w/ SD = σ.
* Then, the posterior distribution from μ, *after seeing the data*, is normal w/ mean = a weighted average of the prior mean + the sample mean.



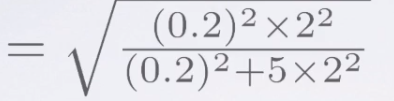
* The posterior SD = the square root of σ squared times **τ** squared all over σ squared + n \* τ squared.



* Suppose an analytical chemist wants to measure the mass of a sample of ammonium nitrate w/ a balance w/ a known SD = 0.2 milligrams.
* By looking at the sample, she thinks this mass is ~10 mg + based on her previous experience in estimating masses by eye, her uncertainty in that guess (the SD of her prior) = 2.
* So she decides her prior for the mass of the sample is a normal distribution w/, μ = 10 mg + σ = 2mg, or **N(10,2)**
* She collects 5 measurements on the sample + finds the average of those = 10.5 mg.
* By **conjugacy** of the normal-normal family, our posterior belief about the mass of the sample has the normal distribution + the new mean of that posterior normal is found by plugging into the formula.

 = 10.499.

* Similarly, her posterior SD also changes by plugging into the formula for the posterior SD

 = reduces from 2 mg to 0.089 mg.

* So the Bayesian analytical chemist has learned a lot in her analysis of the ammonium nitrate, as the posterior mean has shifted quite a bit + uncertainty has dropped by a lot, exactly what an analytical chemist wants.
* This is the last of 3 examples of conjugate families.
* There are many more, but they do not suffice for every situation one might have.

***2.3 Credible Intervals + Predictive Inference***