Introduction

Bayesian Inference refers to the process of inductive learning via *Baye's Rule*.

More Generally...

Bayesian Methods are data analysis tools derived from the principles of *Bayesian Inference*

Statistical induction is the process of learning about the general charecteristics of a population from a subset of members of that population. Recall also, that population charecteristics can be numerically expressed as a parameter θ of a population (dataset) y. Given a subset of y we can decrease our uncertainty about θ

↳ This is the key idea of *Bayesian Inference*

More formally...

The sample space $\mathcal{Y}$ is the set of all datasets that are possible from which a single dataset y will result.

The parameter space Θ is the set of all values that the parameter could possibly have, from which we hope to identify the value that best describes a population charecteristic. Axiomatically, that is:

1. For each numerical value $\theta \in \Theta$, the **Prior distribution** $p(\theta)$ describes our belief that θ represents the true population charecteristic
2. For each $\theta \in \Theta$ & $y \in \mathcal{Y}$, our **Sampling Model** $p(y|\theta)$ describes our belief that y would be the outcome of our study if we knew θ to be true
3. For each numerical value of $\theta \in \Theta$ our **Posterior distribution** $p(\theta|y)$ describes our belief that θ is the true value given the observed dataset y

Baye's Rule & Example

The posterior distribution can be obtained from the prior distribution & the sampling distribution via

Baye's Rule:

$$P(\theta|y) = \frac{P(y|\theta)P(\theta)}{\int_{\Theta} P(y|\tilde{\theta})P(\tilde{\theta})d\tilde{\theta}}$$

Which provides us with an optimal method of updating beliefs about θ given new information y

Estimating The Probability of a Rare Event:

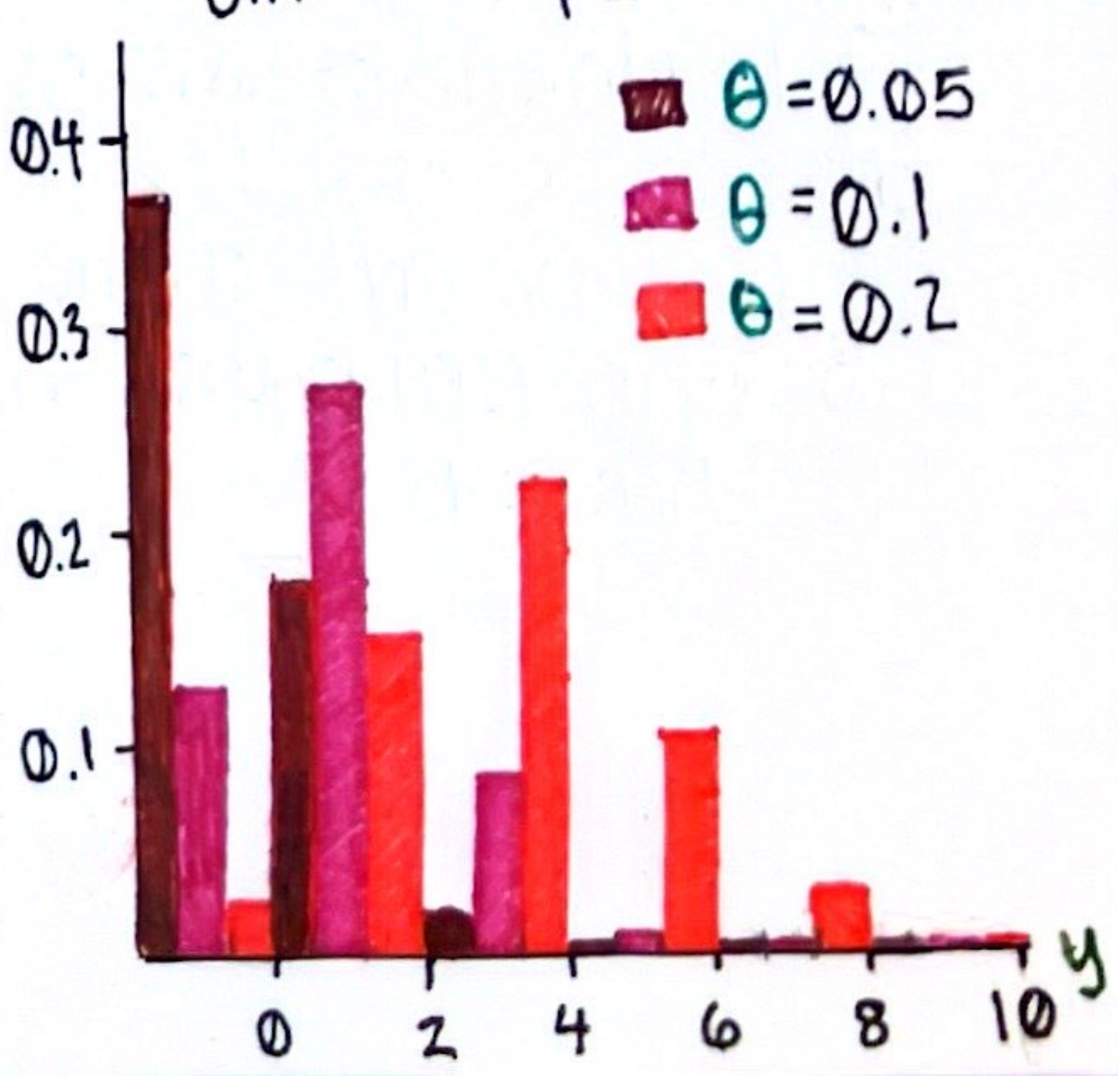
Suppose we are interested in the prevalence of an infectious disease within a small city. A small random sample of 20 individuals will be tested for the disease.

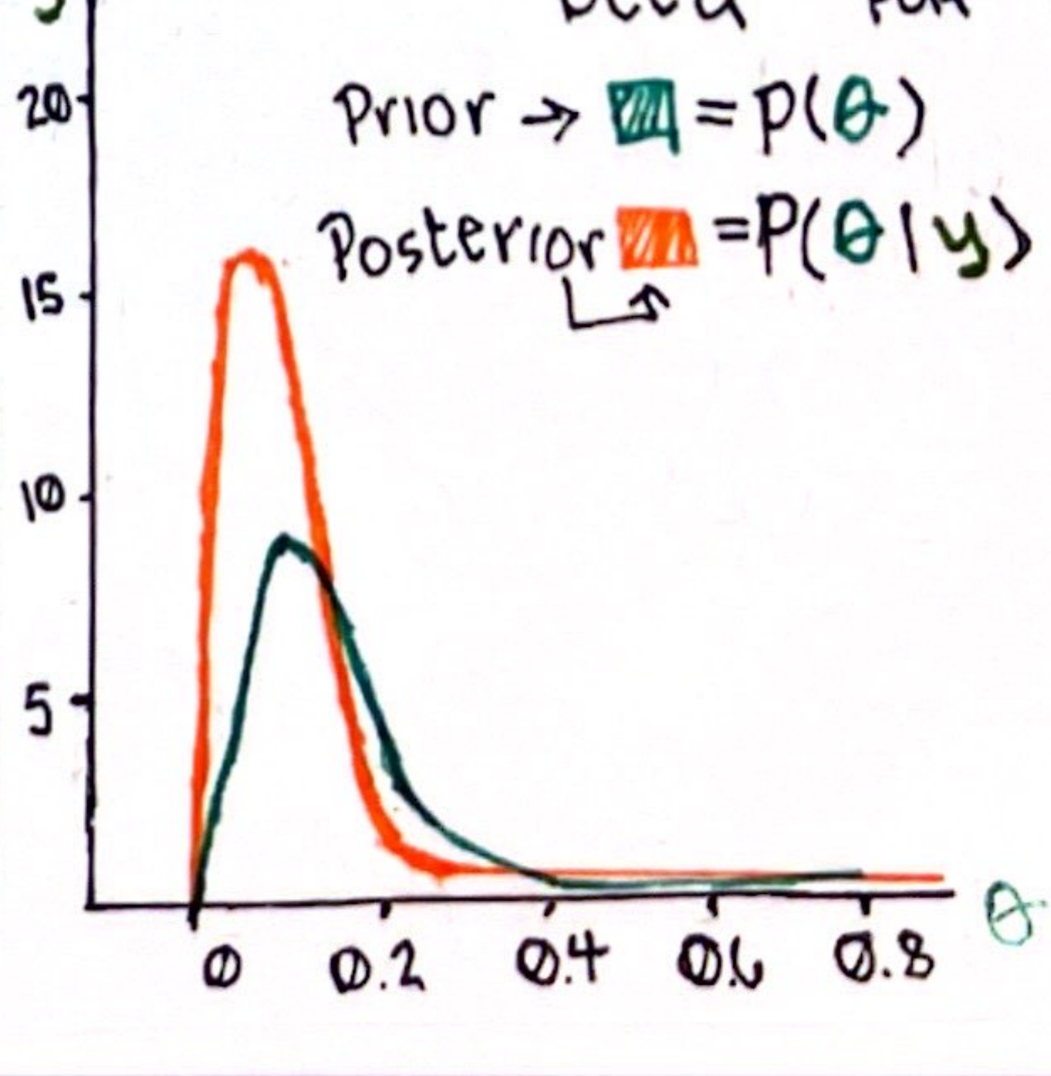
Our parameter of interest is θ = the fraction of individuals in the city infected w/ the disease accordirdingly. The parameter space is given by $\Theta = [0, 1]$. Additionally, the sample statistic y is the number of infected individuals observed.

Thus the sample space is given by $\mathcal{Y} = \{1, 2, 3, \ldots, 20\}$

Before our sample is observed the number of infected individuals present is unknown, let the random variable Y denote this unknown number.

If we did know θ, a reasonable sampling model for Y would be a binomial($n=20$, $p=\theta$) that is $(Y|\theta) \sim \text{binomial}(20, \theta)$





Sampling Model & Prior Distribution

Other studies done on this disease in different locations across the country indicate that infection rate observed in similar cities ranges from 0.05 to 0.20 with an average prevalence of 0.10. Using this *prior* information, we can find a *prior* distribution that is justified by the previous observations. $\theta \sim beta(a, b)$ If ...

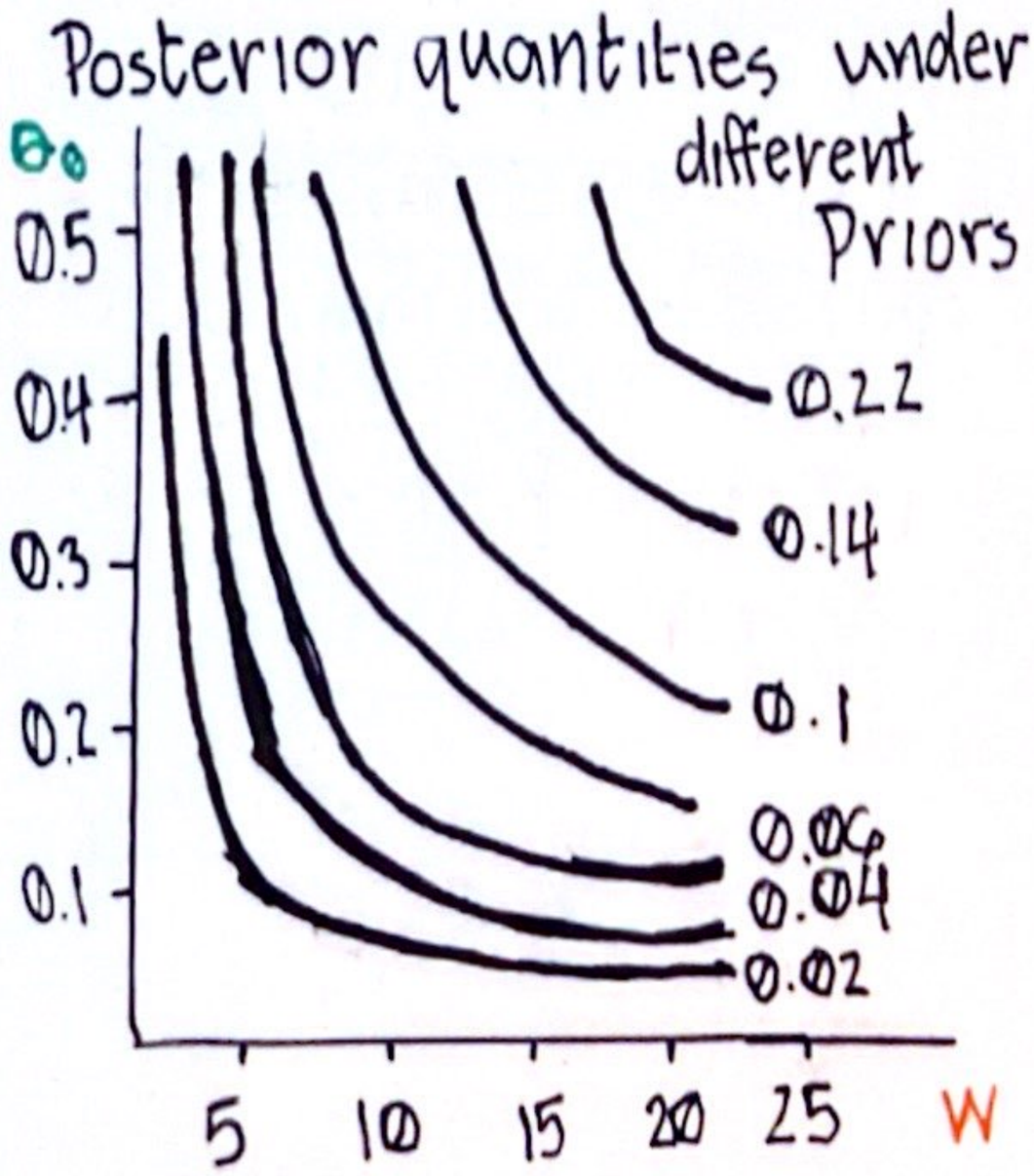
$E(\theta) = \frac{a}{a+b}$ & $\max(P(\theta)) = P\left(\theta = \frac{a-1}{a-1+b-1}\right)$ In other words;

θ has a beta distribution if the mean of θ is equal to $\frac{a}{a+b}$ & the mode of θ is equal to $\frac{a-1}{a-1+b-1}$ thus based on the previous studies we can use $\theta \sim beta(2, 20)$

since $E(\theta) = \frac{2}{22} = \frac{1}{11}$ & $mode(\theta) = \frac{1}{20}$ → $P(\theta < 0.10) = 0.64$

$P(0.05 < \theta < 0.20) = 0.66$

Discussed in more detail in chapter 3, if $(Y|\theta) \sim binomal(n, \theta)$ & $\theta \sim beta(a.b)$ then the posterior distribution is given by $(\theta|y) \sim beta(a+y, b+n-y)$ where y is our observed value of Y & n is our sample size. Suppose that 0 of the 20 individuals in our sample have the disease, meaning $Y = 0$ then the posterior distribution is given by $(\theta|y=0) \sim beta(a=2, b=40)$



each contour on the graph to the left represents different values of $E(\theta|Y=0)$ corresponding to a prior of θ_0 & degree of confidence w. In other words; If someone provides us with a prior guess θ_0 & degree of cofidence w we can approximate their prior beliefs about θ with $beta(a,b)$ where $a = w\theta_0$ & $b = w(1-\theta_0)$

Sensitivity Analysis & Comparison

By computing a range of posterior distributions for different values of θ_0 & w we are performing what's called a **Sensitivity analysis** which is anus investigation of how posterior information is affected bus differences in prior opinion.

Comparison with Frequentist Method

A standard estimate of population proportion θ is the sample mean $\bar{y} = y/n$ i.e: the fraction of infected people observed in the sample. For our sample in which $y=0$ $\bar{y}$ is clearly also equal to zero which is about as helpful in predicting the population parameter as not running a study at all. ← But $\lim_{\varepsilon \to 0}$ calls bull shit on that so lets consider a non-trivial case:

A *95% Confidence Interval* for population proportion can be constructed as $\bar{y} \pm 1.96\sqrt{\bar{y}(1-\bar{y})/n}$ which is known as the *Wald Interval* for θ. This interval has what's known as **correct asymptotic frequentist coverage** which is the graduate level way of saying that as long as n is sufficiently large then with approximately 95% confidence, Y will take on a value y such that the above interval contains θ. However our Bayesian approach does not demand a large sample; It gives a *useful* estimation for <u>all</u> possible sample observations, not asymtotically or sample size dependent. A more thourough comparision of Bayesian & non-bayesian approaches is addressed in chapter 5 (spoiler: it depends... I know, thats shocking.)

Chapter 2: Belief, Probability & Exchangeability

2.1: Belief Functions & Probabilities

Let F, G, & H be three possible overlapping statements about the world. For example: F = {a person votes blue}
& let Be() be a **Belief Function** G = {a person is in poverty}
which is a function that assigns H = {a person lives in a city}
numbers to statements such that the larger the number, the higher the degree of belief

The Axioms of Belief:

Any function that is to numerically represent Beliefs must satisfy the following axioms:

B1: $Be(\text{not } H|H) \leq Be(F|H) \leq Be(H|H)$

B2: $Be(F \text{ or } G|H) \geq \max\{Be(F|H), Be(G|H)\}$

B3: $Be(F \text{ and } G|H)$ can be derived from $Be(G|H)$ & $Be(F|G \text{ and } H)$

↳ Interpretations; Since they are fairly terse statements

B1: "the number we assign to $Be(F|H)$, our conditional belief in F given H, is bounded above & below by the numbers we assign to complete disbelief $Be(\text{not } H|H)$ & complete (certain) belief $Be(H|H)$"

B2: "our belief that the truth lies in a given set of possibilities should not decrease as we add to the set of possibilities ('beliefties' are "independent" of ignorance)"

B3: "If we have to decide between not F & G, knowing that H is true; first deciding if not G is true given H, & if so, then deciding whether or not F is true given G and H"

Recall: The Axioms of Probability

P1: $0 = P(\text{not } H|H) \leq P(F|H) \leq P(H|H) = 1$

P2: $P(F \cup G|H) = P(F|H) + P(G|H)$ if $F \cup G = \varnothing$ ← the empty set, meaning F & G are disjoint events (mutually exclusive)

P3: $P(F \cap G|H) = P(G|H)P(F|G \cap H)$

If a function satisfies the axioms of probability, it also satisfies the axioms of belief

↳ does this mean the two functions are isomorphic?

2.2: Events, Partitions, & Baye's Rule

A **Partition** is a collection of sets $\{H_1, \ldots, H_k\}$ is said to be a partition of another set $\mathcal{H}$ if

1. The events are disjoint $\rightarrow H_i \cap H_j = \varnothing$ for $i \neq j$
2. The union of all sets is $\mathcal{H} \rightarrow \bigcup_{k=1}^{K} H_k = \mathcal{H}$

Suppose $\{H_1, \ldots, H_k\}$ is a partition of $\mathcal{H}$, $P(\mathcal{H}) = 1$ & E is some specific event. Then, the Axioms of probability imply the following:

The Rule of Total Probability: $\sum_{k=1}^{K} P(H_k) = 1$ ← The sample space must sum to one (must integrate to one in the continious case)

The Rule of Marginal Probability: defines the behavior of compound probabilities "E and H_k"

$$P(E) = \sum_{k=1}^{K} P(E \cap H_k) \leftarrow \text{For events that are not mutually exclusive}$$

$$= \sum_{k=1}^{K} P(E \mid H_k) P(H_k)$$

Baye's Rule: As previously mentioned. Baye's Rule relates a posterior probability to a given prior probability

$$P(H_j \mid E) = \frac{P(E \mid H_j) P(H_j)}{P(E)}$$

$$= \frac{P(E \mid H_j) P(H_j)}{\sum_{k=1}^{K} P(E \mid H_k) P(H_k)} \leftarrow \text{See pages 24-25 for various proofs of Baye's Rule.}$$

Implications of Partitions

A subset of the 1996 *General Social Survey* includes data on the education level & Income for a sample of men over age 30. Let $\{H_1, H_2, H_3, H_4\}$ be the events that a random man in this sample is in the 25^{th}, 50^{th}, 75^{th}, & 100^{th} percentile in terms of income, respectively. Bu definition...

$$\{P(H_1), P(H_2), P(H_3), P(H_4)\} = \{0.25, 0.25, 0.25, 0.25\}$$

meaning that the probabilities are disjoint & sum to one which implies that the set $\{H_1, H_2, H_3, H_4\}$ is a partition of the sample space. If we let E be the event that a randomly selected person from the survey has a college education. From the survey data we have...

$$\{P(E|H_1), P(E|H_2), P(E|H_3), P(E|H_4)\} = \{0.11, 0.19, 0.31, 0.53\}$$

These probabilities do <u>not</u> sum to one; they are proportions *within* different subpopulations, meaning they are each in seperate sample spaces. Applying Baye's Rule yields...

$$\{P(H_1|E), P(H_2|E), P(H_3|E), P(H_4|E)\} = \{0.09, 0.17, 0.27, 0.47\}$$

which is much different than the population distribution

In Bayesian Statistics $\{H_1, \ldots, H_K\}$ often refer to disjoint hypotheses & E refers to the outcome of an experiment. To compare hypotheses post-experiment the following ratio often proves useful:

$$\frac{P(H_i|E)}{P(H_j|E)} = \frac{P(E|H_i)\,P(H_i)/P(E)}{P(E|H_j)\,P(H_j)/P(E)} = \frac{P(E|H_i)\,P(H_i)}{P(E|H_j)\,P(H_j)}$$

$$= \frac{P(E|H_i)}{P(E|H_j)} \times \frac{P(H_i)}{P(H_j)} = \text{Baye's Factor} \times \text{Prior Belief}$$

Baye's Rule doesn't tell us *what* to believe, but rather, how our belief should change given a particular observation.

2.3: Independence of Events

Two events F & G are said to be **Conditionallu Independent** given H if $P(F \cap G|H) = P(F|H)P(G|H)$

↳ meaning that by the third Axiom of probability, P3, the following is also always true:

$P(F \cap G|H) = P(G|H)P(F|H \cap G)$ If F & G are conditionally independent given H then we must have:

$P(G|H)P(F|H) \underset{\text{always}}{=} P(F \cap G|H) \underset{\text{independence}}{=} P(F|H)P(G|H)$

$P(G|H)P(F|H \cap G) = P(F|H)P(G|H)$

$P(F|H \cap G) = P(F|H)$ Conditional Independence

Therefore implies that if we know that H is true & F & G are conditionally independent given H, then knowing G does <u>not</u> change our belief about F

We will use this defintion to build our definition of independence for random variables

Example: Consider the conditional dependence of F & G. when H is assumed to be true in the following two scenarios:

F = {a paitent is a smoker}
G = {a paitent has lung cancer}
H = {smoking causes lung cancer}

F = {you are thinking of a Jack of hearts}
G = {a mind reader claims that you are thinking of a jack of hearts}
H = {the mind reader has ESP}

in both of these situations, H being true implies a relationship between F & G; When H is false then F & G are either related in some other way or they are independent

2.4 Random Variables

In Bayesian Inference a **Random Variable** is defined as an unknown quantity about which we make probability statements. In a more specific case a fixed but unknown population parameter is a random variable

Discrete Random Variables

Let Y be a random variable & let $\mathcal{Y}$ be the set of all possible values of Y. Y is said to be discrete if it has *countably* many possible out comes; That is, if $\mathcal{Y}$ can be expressed as $\mathcal{Y}=\{y_1, y_2, \dots\}$

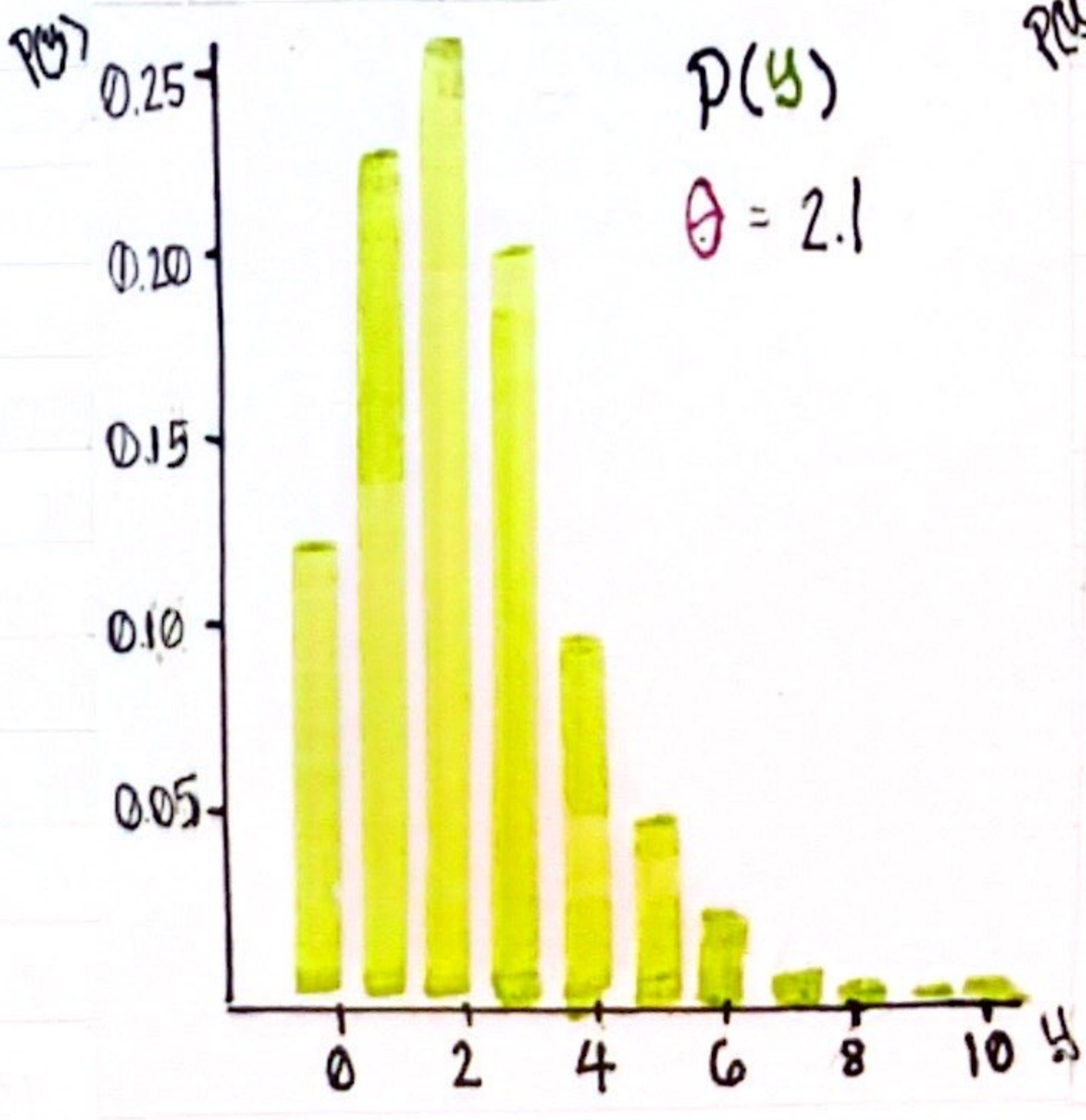
The function $P(Y=y) = p(y)$ is called the *Probability mass function* or pmf & it has the following properties:

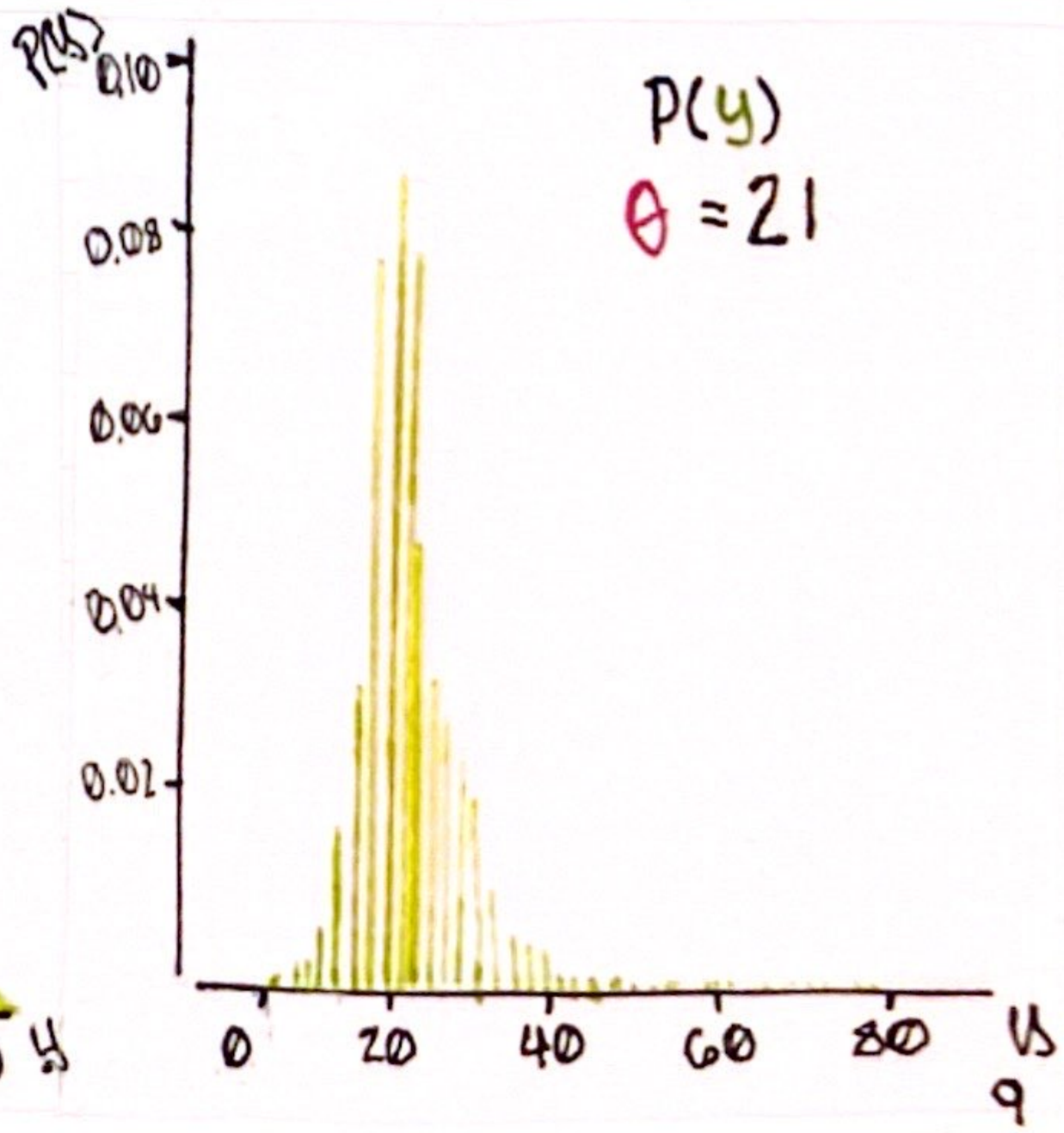
1. $0 \leq p(y) \leq 1$ for all $y \in \mathcal{Y}$
2. $\sum_{y \in \mathcal{Y}} p(y) = 1$

Many useful/general probability statements can be derived from the pmf

Example:

Let $\mathcal{Y}=\{0, 1, 2, \dots, n\}$ for some positive interger n. The uncertainty quality $Y \in \mathcal{Y}$ has a poisson distribution with a mean equal to θ





Random Variables Continued... (Get it?)

Continious Random Variables

Suppose that the sample space $\mathcal{Y}$ is roughly equal to $\mathbb{R}$. We cannot easily express the possible values of Y as a summation; the sample space is *uncountable*
Thus, instead of a pmf we define continious random variables in terms of a piecewise-smooth differentiable function $F(y)$ called the *cumulitive distribution function* or cdf $F(y)=P(Y\leq y)$. Additionally for each cdf there exists an analagous function $f(y)$ called the *probability density function* which is defined as

$$F(y)=\int_{\mathcal{Y}} f(y)\,dy$$ or pdf

The pmf has similar properties to the pdf of a discrete random variable; them being

1. $0\leq f(y)$ for all $y\in\mathcal{Y}$
2. $\int_{y\in\mathbb{R}} f(y)\,dy=1$

Integration can be thought of as a generalization of the summation used to evaluate discrete random variables

Example:

Let $\mathcal{Y}=\mathbb{R}$ & suppose we sample this population which has a known mean of μ & variance of σ^2 the population is said to have the *Normal distribution* if the random variable Y effectively models the population with the following cdf

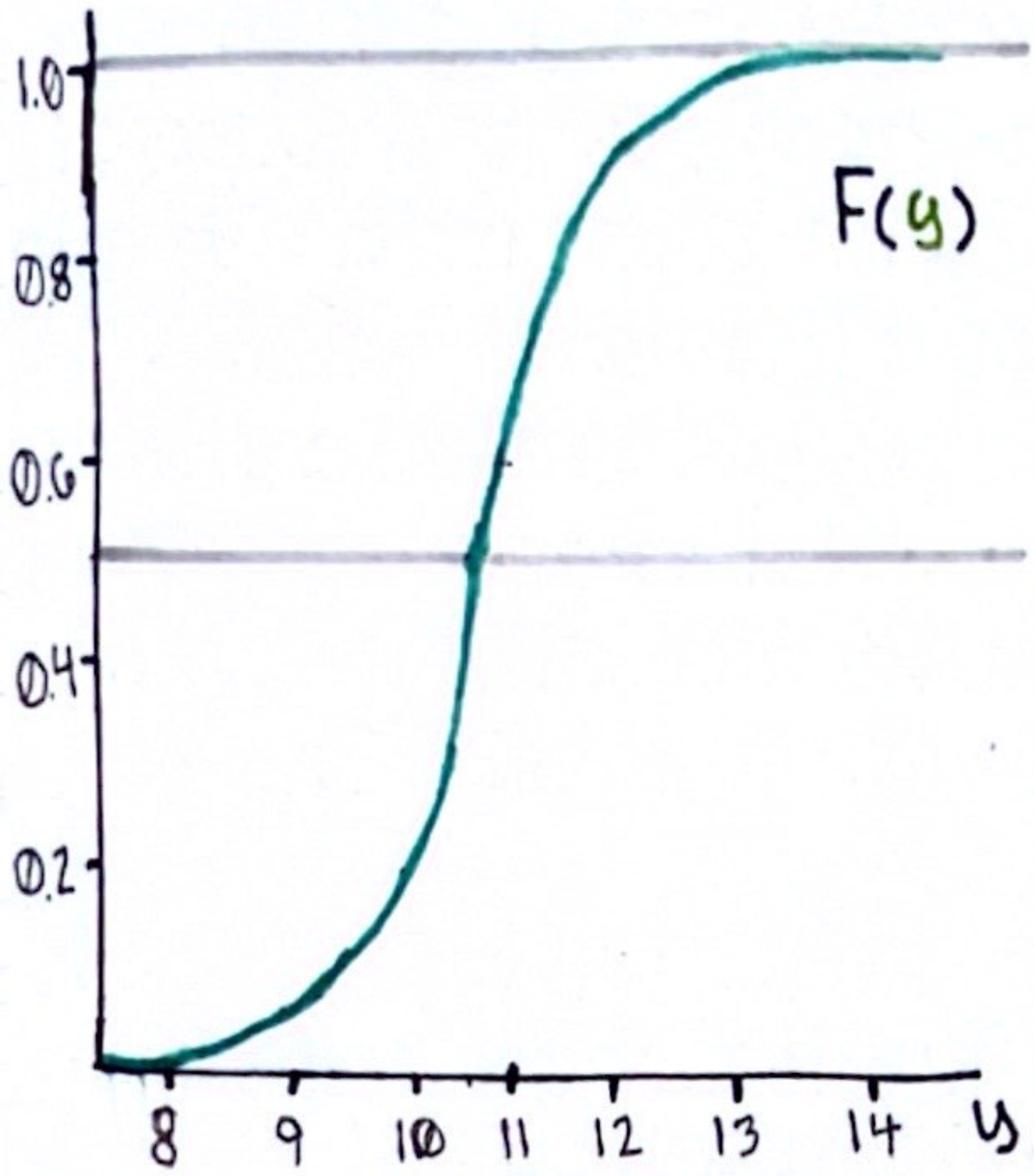
$$P(Y\leq y|\mu,\sigma^2)=F(y)=\int_{-\infty}^{y}\frac{1}{\sqrt{2\pi}\,\sigma}\exp\left\{\frac{-1}{2}\left(\frac{y-\mu}{\sigma}\right)^2\right\}dy$$

If we let $\mu=10.75$ & $\sigma=0.8$ ($\sigma^2=0.64$) then the median value of e^Y equal to about 46,630 which is roughly equal to the median household income in the US in 2005 $P(e^Y>100,000)=P(Y>\ln(100,000))=0.17$ which matches census data from the same year. The cdf & pdf of this distribution are plotted on the next page.

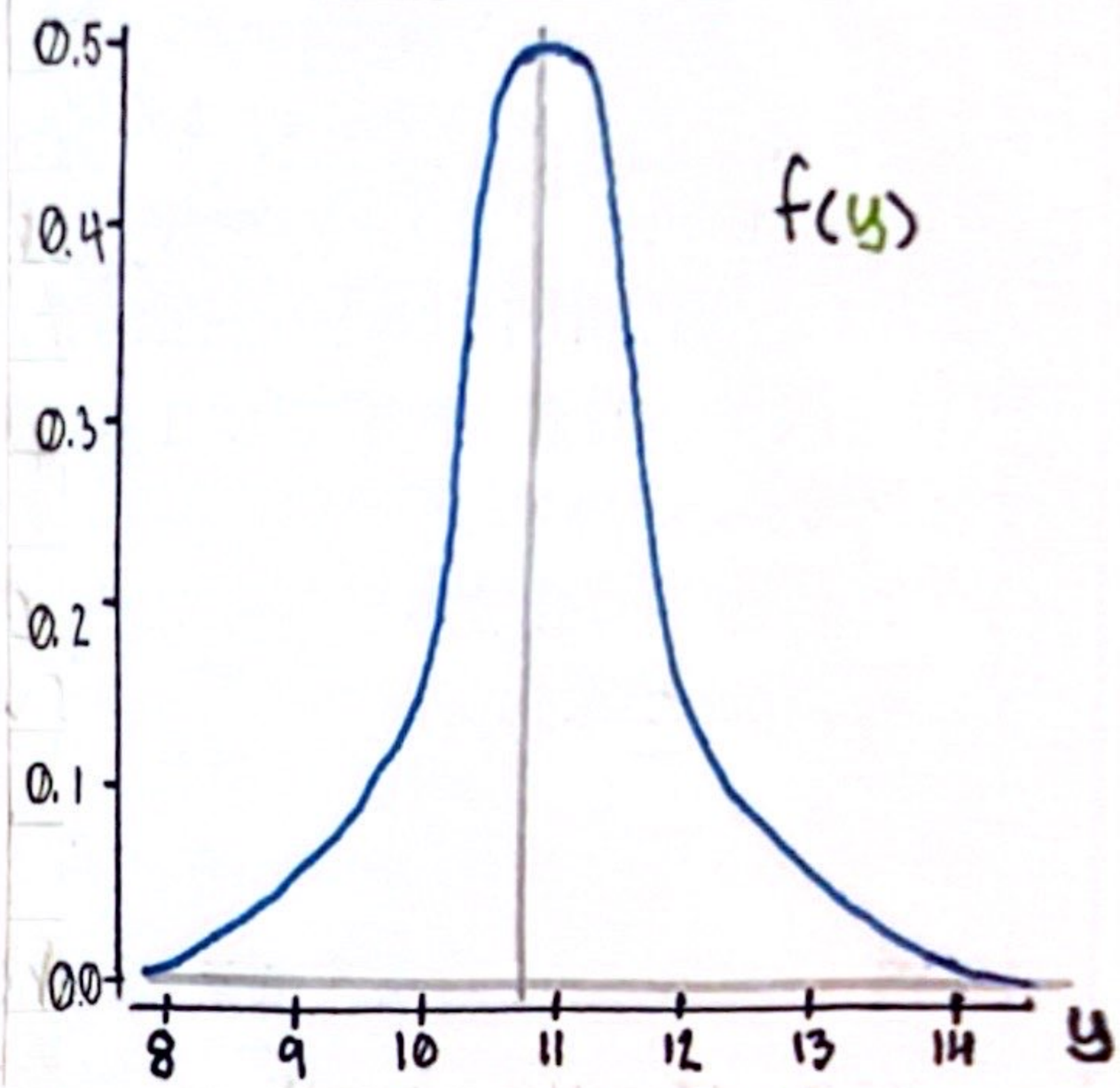
Review: Descriptive Statistics

$Y \sim \text{Normal}(\mu = 10.75, \sigma^2 = 0.64)$

Cumulative density function



Probability density function



Descriptors of Distributions:

mean (expectected value):

$$E[Y] = \sum_{y \in \mathcal{Y}} y\,p(y) \quad \text{or} \quad E[Y] = \int_{y \in \mathcal{Y}} y f(y)\,dy$$

discrete ↑ continious ↑

mode: the most probable value of Y

median: the value with equal probability on either side

→ Key Properties of the Mean:

1. μ is a scaled version of the population total
2. $\bar{y}$ is a least squares estimator of μ

variance: the average squared distance from Y to μ

$$V[Y] = E[(Y - E[Y])^2] = E[Y^2 - 2YE[Y] + E[Y]^2] =$$
$$= E[Y^2] - 2E[Y]^2 + E[Y]^2 = E[Y^2] - E[Y]^2$$

Standard deviation: the square root of $V[Y]$

Since the normal distribution is symetric it median & mode are equal to μ; see the graphs above.

2.5: Joint Probability Distributions

The Discrete Case:

If Y_1 & Y_2 are discrete random variables then *joint beliefs* about Y_1 & Y_2 can be represented as probabilities For example, subsets $A \subset \mathcal{Y}_1$ & $B \subset \mathcal{Y}_2$ the probability $P(\{Y_1 \in A\} \cap \{Y_2 \in B\})$ represents our belief that Y_1 is in A & that Y_2 is in B. The *Joint density function* or *Joint pdf* of Y_1 & Y_2 is defined as $f_{Y_1Y_2}(y_1, y_2) = P(\{Y_1 = y_1\} \cap \{Y_2 = y_2\})$ for $Y_1 \in \mathcal{Y}_1$ & $Y_2 \in \mathcal{Y}_2$

The *marginal density* of Y_1 can be computed from this

$$f_{Y_1}(y_1) \equiv P(Y_1 = y_1) = \sum_{y_2 \in Y_2} P(\{Y_1 = y_1\} \cap \{Y_2 = y_2\}) \equiv \sum_{y_2 \in Y_2} f_{Y_1,Y_2}(y_1, y_2)$$

The *conditional density* of Y_2 given $\{Y_1 = y_1\}$ can be computed from the joint pdf & the marginal density:

$$f_{Y_2|Y_1}(y_2|y_1) = \frac{P(\{Y_1 = y_1\} \cap \{Y_2 = y_2\})}{P(Y_1 = y_1)}$$

Example:

Logan (1983) reports the following joint distribution of occupational categories of fathers & sons

↓fathers sons→	farm	operative	craftsman	sales	Profes.
farm	0.018	0.035	0.031	0.008	0.018
operative	0.002	0.112	0.064	0.032	0.069
craftsman	0.001	0.066	0.094	0.032	0.084
sales	0.001	0.018	0.019	0.010	0.051
Professional	0.001	0.029	0.032	0.043	0.130

Suppose we are to sample a father-son pair from this population. Let Y_1 be the father's occupation & Y_2 be the son's Then.

$$P(Y_2 = \text{professional} \mid Y_1 = \text{farm}) = \frac{P(Y_2 = \text{professional} \cap Y_1 = \text{farm})}{P(Y_1 = \text{farm})}$$

$$= \frac{0.018}{0.018 + 0.35 + 0.031 + 0.008 + 0.018} = 0.164$$

Continious Joint Distributions:

If Y_1 & Y_2 are continious random variables, we start with a cdf. Given a joint cdf $F_{Y_1Y_2}(a,b)$ such that $F_{Y_1Y_2}(a,b) \equiv P(\{Y_1 \leq a\} \cap \{Y_2 \leq b\})$. Then, there exists a function $f_{Y_1Y_2}$ such that

$$F_{Y_1Y_2}(a,b) = \int_{-\infty}^{a}\int_{-\infty}^{b} f_{Y_1Y_2}(y_1, y_2)\, dy_2\, dy_1$$

This function $f_{Y_1Y_2}$ is called the *joint density function*. As with the discrete case, two key properties of the joint density function are:

1. $f_{Y_1}(y_1) = \int_{-\infty}^{\infty} f_{Y_1Y_2}(y_2, y_2)\, dy_2$
2. $f_{Y_2|Y_1}(y_2 \mid y_1) = \dfrac{f_{Y_1Y_2}(y_1, y_2)}{f_{Y_1}(y_1)}$

→ for each value of y_1 This equation gives a corresponding pdf for Y_2

Why not Both? (at the same time?!)

Let Y_1 be discrete & Y_2 be continious. We can define a *marginal* density f_{Y_1} from our beliefs $P(Y_1 = y_1)$ & a *conditional* density $f_{Y_2|Y_1}(y_2|y_1)$ from the following

$P(Y_2 \leq y_2 \mid Y_1 = y_1) \equiv F_{Y_2|Y_1}(y_2|y_1)$ as above.

The joint density of Y_1 & Y_2 is then given by:

$f_{Y_1Y_2}(y_1, y_2) = f_{Y_1}(y_1) \times f_{Y_2|Y_1}(y_2|y_2)$ which has the property that the following is always true:

$$P(Y_1 \in A, Y_2 \in B) = \int_{y_2 \in B} \left\{ f_{Y_1Y_2}(y_1, y_2) \right\} dy_2$$

The subscripts of density functions are often dropped in which case the type of density function is implied by its argument

$f(y_1) \longrightarrow f_{Y_1}(y_1)$

$f(y_1, y_2) \longrightarrow f_{Y_1Y_2}(y_1, y_2)$

$f(y_1|y_2) \longrightarrow f_{Y_1|Y_2}(y_1, y_2)$

Baye's Rule Parameter Estimation:

Let θ = the proportion of people in a large population who have a specific charecteristic

Y = the number of people in a small sample from this population that have the specific charecteristic

We can then treat θ as a continious random variable & Y as a discrete random variable. Bayesian estimation of θ derives from the calculation of $f(\theta|y)$, where y is the observed value of Y. We first need to find a joint pdf which we can build from:

Assumptions

1. $f(\theta)$; beliefs about θ
2. $f(y|\theta)$; beliefs about Y for each θ

Having observed $\{Y=y\}$, we need to compute our beliefs about θ

$$f(\theta|y) = \frac{f(\theta,y)}{f(y)} = \frac{f(\theta)f(y|\theta)}{f(y)}$$

This conditional density is called the *Posterior Density* of θ. Suppose that θ_a & θ_b are two possible values of θ. The posterior density (probability) of θ_a relative to θ_b, conditional on $Y=y$ is given by:

$$\frac{f(\theta_a|y)}{f(\theta_b|y)} = \frac{f(\theta_a)f(y|\theta_a)/f(y)}{f(\theta_b)f(y|\theta_b)/f(y)} = \frac{f(\theta_a)f(y|\theta_a)}{f(\theta_b)f(y|\theta_b)}$$

This means that we need not know the value of $f(y)$ to find the relative posterior probabilities of θ_a & θ_b i.e. as a proportional function ($\propto$) of θ $f(\theta|y) \propto f(\theta)f(y|\theta)$

Where the *constant of proportionality* is given by $\frac{1}{f(y)}$ which can be computed from the following

$$f(y) = \int_{\Theta} f(y,\theta)d\theta = \int_{\Theta} f(y|\theta)f(\theta)d\theta \xrightarrow{\text{which yields}} f(\theta|y) = \frac{f(\theta)f(y|\theta)}{\int_{\Theta} f(\theta)f(y|\theta)d\theta}$$

As will be shown in a later section, the numerator is the critical term of this equation.

2.6: Independent Random Variables

Suppose that $Y_1, \ldots, Y_n$ are random variables & that θ is a parameter describing the conditions under which the random variables are generated. We say that $Y_1 \ldots Y_n$ are *conditionally independent* given θ if for every collection of n sets $\{A_1, \ldots, A_n\}$ we have $P(Y_1 \in A_1, \ldots, Y_n \in A_n)$

$= P(Y_1 \in A_1 | \theta) \times \cdots \times P(Y_n \in A_n | \theta)$

This definition is based on our previously discussed definition of independent events where each $\{Y_j \in A_J\}$ is an event. If independence holds then it follows that $P(Y_i \in A_i | \theta, Y_j \in A_j) = P(Y_i \in A_i | \theta)$.

Thus, conditional independence can be interpreted as meaning that Y_j gives no additional information about Y_i beyond that which is known from θ.

Furthermore, under independence the joint pdf is

$$f(y_1, \ldots, y_n | \theta) = f_{Y_1}(y_1 | \theta) \times \cdots \times f_{Y_n}(y_n | \theta) = \prod_{i=1}^{n} f_{Y_i}(y_i | \theta)$$

which is simply the product of the marginal densities. Suppose that $Y_1, \ldots, Y_n$ are generated in similar ways from a common process. (for example: samples from the same population or rounds of an experiment preformed under similar conditions) This suggests that the marginal densities are all equal to some common densitys

$$f(y_1, \ldots, y_n | \theta) = \prod_{i=1}^{n} f(y_i | \theta)$$

↳ in this case we say that $Y, \ldots, Y_n$ are (deep breath) *conditionally independent & indentically distributed* or, for the sake of brevity, **i.i.d.** Mathematical shorthand for this is $Y_1, \ldots, Y_n | \theta \sim$ i.i.d. $f(y | \theta)$

↳ in the case of samples taken randomly from a large, but finite, population without replacement it is often reasonable to assume that the observations are i.i.d.

2.7: Exchangeability Example: Happiness

Participants in the 1998 *General Social Survey* were asked whether or not they were generally happy let Y_i be the random variable that models the response to this question. We can define Y_i as the following:

$$Y_i = \begin{cases} 1 & \text{if participant } i \text{ reports general happiness} \\ 0 & \text{otherwise} \end{cases}$$

let's consider the structure of our joint beliefs about $Y_1, \ldots, Y_{10}$; the outcomes of the first ten randomly selected respondants. (Recall: $f(y_1, \ldots, y_{10})$ is shorthand for $P(Y_1 = y_1, \ldots, Y_{10} = y_{10})$, where each y_i is either 0 or 1)

Suppose we want to assign probabilities to 3 unique outcomes, as defined by: $f(1,0,0,1,0,1,1,0,1,1) = ?$

Is there an argument for $f(1,0,1,0,1,1,0,1,1,0) = ?$ assigning all three outcomes $f(1,1,0,0,1,1,0,0,1,1) = ?$ the same probability?... Note that all three outcomes contain six 1s & four 0s

Let $f(y_1, \ldots, y_n)$ be the *joint density* of $Y_1, \ldots, Y_n$. If $f(y_1, \ldots, y_n) = f(y_{\pi_1}, \ldots, y_{\pi_n})$ for all *permutations* π of $\{1, \ldots, n\}$ then $Y_1, \ldots, Y_n$ are said to be **Exchangeable** → Roughly speaking, $Y_1, \ldots, Y_n$ are exchangeable if their subscripts convey no information about the outcomes.

Independence vs. Dependence

Consider the following pair of probability assignments

$P(Y_{10} = 1) = a$

$P(Y_{10} = 1 \mid Y_1 = Y_2 = \cdots = Y_8 = Y_9 = 1) = b$

} if $a \neq b$ then Y_{10} is NOT independent of $Y_1, \ldots, Y_9$. But, if $a = b$, then the probability is the same in both cases, making Y_{10} independent given $Y_1, \ldots, Y_9$.

Conditional Independence Example

Suppose someone told you the numerical value of θ, the rate of happiness among the 1272 respondants. Do the following probability assignments seem reasonable?

$P(Y_{10}=1 \mid \theta) \approx ? \theta$

$P(Y_{10}=1 \mid Y_1=y_1, \dots, Y_9=y_9, \theta) \approx ? \theta$

$P(Y_9=1 \mid Y_1=y_1, \dots, Y_8=y_8, Y_{10}=y_{10}, \theta) \approx ? \theta$ If they are reasonable, then we can consider the Y_i's as conditionally independent & identically distributed given θ (Approximately; $1,272 \gg 10$). Assuming conditional independence yields the following:

$$P(Y_i=y_i \mid \theta, Y_j=y_j, j \neq i) = \theta^{y_i}(1-\theta)^{1-y_i}$$

$$P(Y_1=y_1, \dots, Y_{10}=y_{10} \mid \theta) = \prod_{i=1}^{10} \theta^{y_i}(1-\theta)^{1-y_i} = \theta^{\sum y_i}(1-\theta)^{10-\sum y_i}$$

If θ is uncertain to us, we describe our beliefs about it with $P(\theta)$, a prior distribution. The marginal joint distribution of $Y_1, \dots, Y_{10}$ is then:

$$f(y_1, \dots, y_{10}) = \int_0^1 f(y_1, \dots, y_{10} \mid \theta) P(\theta) \, d\theta = \int_0^1 \theta^{\sum y_i}(1-\theta)^{10-\sum y_i} P(\theta) \, d\theta$$

Now, consider the probabilities of the binary sequences given on the previous page

$f(1,0,0,1,0,1,1,0,1,1)$
$f(1,0,1,0,1,1,0,1,1,0)$
$f(1,1,0,0,1,1,0,0,1,1)$

} $\int_0^1 \theta^6(1-\theta)^4 P(\theta) \, d\theta$ ← This observation is formalized in the following proof

Theorem: If $\theta \sim P(\theta)$ & $Y_1, \dots, Y_n$ are conditionally i.i.d. given θ, then marginally (unconditionally on θ), $Y_1, \dots, Y_n$ are exchangeable

Proof on the following page

Proof: Suppose that $Y_1, \ldots, Y_n$ are conditionally i.i.d. given some unknown parameter θ. Then, for any permutation π of $\{1, \ldots, n\}$ & any set of values $(y_1, \ldots, y_n) \in \mathcal{Y}^n$ we have:

$$f(y_1, \ldots, y_n) = \int f(y_1, \ldots, y_n | \theta) p(\theta) d\theta \quad \text{definition of marginal probability}$$

$$= \int \left\{ \prod_{i=1}^{n} f(y_i | \theta) \right\} p(\theta) d\theta \quad Y_i\text{'s are conditionally i.i.d.}$$

$$= \int \left\{ \prod_{i=1}^{n} f(y_{\pi i} | \theta) \right\} p(\theta) d\theta \quad \text{Product does not depend on order}$$

$$= f(y_{\pi_1}, \ldots, y_{\pi_n}) \quad \text{definition of marginal Probability}$$

Thus, (∴)
If $\theta \sim p(\theta)$ & $Y_1, \ldots, Y_n$ are conditionally i.i.d. given some unknown parameter θ, then marginally (unconditionally on θ), $Y_1, \ldots, Y_n$ are exchangeable.

This theorem lays the groundwork for the first major theorem that we will be studying, *de Finetti's theorem*.

Seperately, it is important that the posterior distribution is *proper*, meaning that it integrates to one. We are allowed to use improper priors but if we do we must take care to ensure that the posterior distribution is proper.

2.8 de Finetti's Theorem

On the previous page we demonstrated that

$$\left.\begin{array}{l} Y_1,\dots,Y_n \mid \theta \text{ i.i.d.} \\ \theta \sim P(\theta) \end{array}\right\} \Rightarrow Y_1,\dots,Y_n \text{ are exchangeable}$$

What about an arrow in the opposite direction? Let $\{Y_1, Y_2, \dots\}$ be a potentially infinite sequence of random variables all having common sample space $\mathcal{Y}$

Theorem 1 (de Finetti): Let $Y_i \in \mathcal{Y}$ for all $i \in \{1,2,\dots\}$ Suppose that for any n, our belief model for $Y_1,\dots,Y_n$ is exchangeable, that is, $f(y_1,\dots,y_n) = f(y_{\pi_1},\dots,y_{\pi_n})$ for all permutation π of $\{1,\dots,n\}$. Then our model can be written as: $f(y_1,\dots,y_n) = \int \left\{ \prod_1^n f(y_i \mid \theta) \right\} p(\theta)\, d\theta$ for some parameter θ, some prior distribution on θ & some sampling model $f(y \mid \theta)$. The prior & sampling model depend on the form of the belief model $f(y_1,\dots,y_n)$

The probability distribution $p(\theta)$ represents our beliefs about the outcomes of $\{Y_1, Y_2, \dots\}$, induced by our belief model $f(y_1, y_2, \dots)$ More precisely:

- ↳ $p(\theta)$ represents our beliefs about $\lim_{n\to\infty} \sum \frac{Y_i}{n}$ in the binary case
- ↳ $p(\theta)$ represents our beliefs about $\lim_{n\to\infty} \sum \frac{Y_i \le c}{n}$ for each c in the general case

The main idea of the last few pages can be summarized as follows:

$$\left.\begin{array}{l} Y_1,\dots,Y_n \mid \theta \text{ are i.i.d} \\ \theta \sim p(\theta) \end{array}\right\} \Leftrightarrow Y_1,\dots,Y_n \text{ are exchangable}$$

So when is this condition ⟶ for all n "reasonable"? We must have exchangeability (labels convey no information) & *repeatability* (finite sample with replacement, large population without replacement (Dianconis & Freedman, 1980), outcomes of reproducible experiment)

Chapter 3: One-Parameter Models

3.1: The Binomial Model

A *one-parmeter model* is a class of sampling distributions that are indexed by a single unknown parameter.

Happiness Data: Each female over age 65 in the 1998 *General Social Survey* was asked whether or not they were generally happy. Let $Y_i = 1$ if respondant i reported yes, & $Y_i = 0$ otherwise. If we lack information distinguishing these $n = 129$ individuals we can treat their responses as exchangeable Since $129 \ll N$ where N is the total number of females over age 65, we can treat the observations as i.i.d. with a joint density given by our beliefs about $\theta = \sum_{i=1}^{N} Y_i / N$ Which implies that our model, conditional on θ, the Y_i's are i.i.d. binary random variables with expectation θ. Thus, the probability for any poetential outcome $\{y_1, \dots, y_{129}\}$, conditional on θ is given by $f(y_1, \dots, y_{129} | \theta) = \theta^{\sum_{i=1}^{129} y_i} (1-\theta)^{129 - \sum_{i=1}^{129} y_i}$ But, we still have yet to define our *prior distribution*. Suppose our parameter θ is uniformly distributed on its domain $[0,1]$, meaning $P(a \le \theta \le b) = P(a+c \le \theta \le b+c)$ for $0 \le a < b < b+c \le 1$ which thus implies that $f(\theta) = 1$ for all $\theta \in [0,1]$ meaning θ has uniform density across all values of θ. For this prior distribution & the above sampling model applying Bayes's rule yields:

$$f(\theta | y_1, \dots, y_{129}) = \frac{f(y_1, \dots, y_{129} | \theta) p(\theta)}{f(y_1, \dots, y_{129})} = f(y_1, \dots, y_{129} | \theta) \times \frac{1}{f(y_1, \dots, y_{129})}$$

In this case $f(\theta | y_1, \dots, y_{129})$ & $f(y_1, \dots, y_{129} | \theta)$ are proportional $\rightarrow \propto f(y_1, \dots, y_{129} | \theta)$ to one another as functions of θ, due to the *posterior distribution* being equal to a number divided by the prior & does not depend on θ; meaning the two functions have the same shape, but not necessarily the same scale.

Suppose the following is observed: of 129 individuals sampled
118 reported being happy (91%)
11 reported being unhappy (9%)
The probability of observing this data for a given value of θ is
$f(y_1,\dots,y_{129}|\theta) = \theta^{"118"}(1-\theta)^{"11"}$ ↰
The plot below on the left. We know the posterior distribution will have the same shape, & as such we know the true value of θ is very likely to be near 0.91, & almost certainly above 0.8. However, we often require more precision than this; we want the shape & scale. From Baye's rule we have

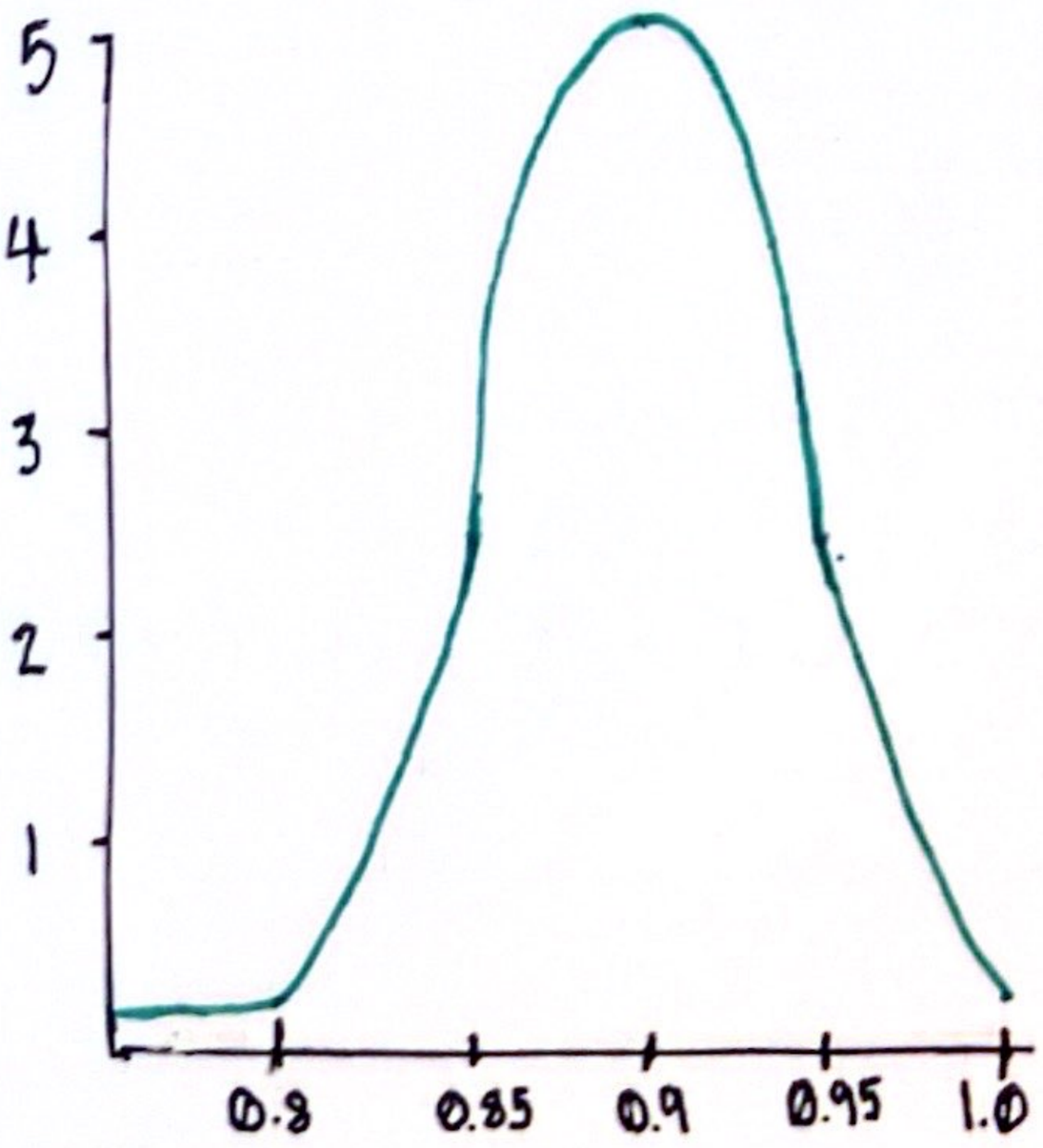
$$f(y_1,\dots,y_{129}) = \theta^{"118"}(1-\theta)^{"11"} \times p(\theta)/f(y_1,\dots,y_{129})$$
$$= \theta^{"118"}(1-\theta)^{"11"} \times 1/f(y_1,\dots,y_{129})$$

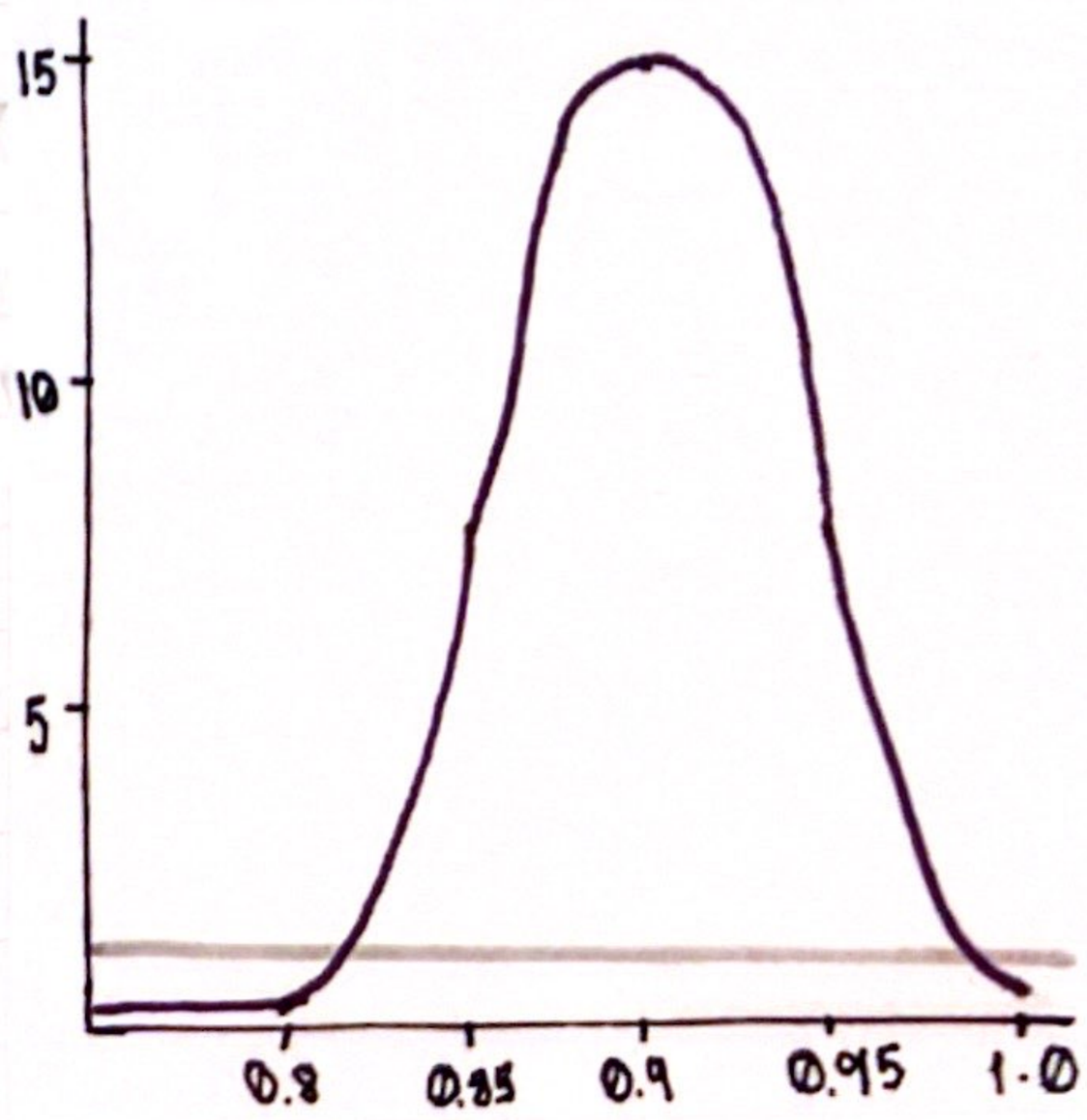
If we calculate the scale or *normalizing constantant* $1/f(y_1,\dots,y_{129})$ using the following from calculus $\int_0^1 \theta^{a-1}(1-\theta)^{b-1}d\theta = \frac{\Gamma(a)\Gamma(b)}{\Gamma(a+b)}$

$1 = \int_0^1 f(\theta|y_1,\dots,y_{129})d\theta$ ↳ Since all probability distributions must integrate/sum to one.

$1 = \int_0^1 \theta^{"118"}(1-\theta)^{"11"}/f(y_1,\dots,y_{129})d\theta$ ↳ From Baye's rule

$$1 = \frac{1}{f(y_1,\dots,y_{129})}\int_0^1 \theta^{"118"}(1-\theta)^{"11"}d\theta = \frac{1}{f(y_1,\dots,y_{129})}\frac{\Gamma(118)\Gamma(11)}{\Gamma(131)} = 1$$





The Beta Distribution

Thus, the previous calculations have shown that

$$f(y_1,\cdots,y_{129}) = \frac{\Gamma(119)\Gamma(12)}{\Gamma(131)}$$

This result holds for any outcome that contains 118 ones & 11 zeros. Putting everything together we have:

$$f(\theta|y_1,\cdots,y_{129}) = \frac{\Gamma(131)}{\Gamma(119)\Gamma(12)}\theta^{118}(1-\theta)^{11}$$

which can be written as

$$= \frac{\Gamma(131)}{\Gamma(119)\Gamma(12)}\theta^{119-1}(1-\theta)^{12-1}$$

This density for θ is called a *beta distribution* with parameters $a=119$ & $b=12$

An uncertainty charecteristic θ between 0 & 1 is said to have a *beta distribution* beta(a,b) if

$$f(\theta) = \text{dbeta}(\theta,a,b) = \frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)}\theta^{a-1}(1-\theta)^{b-1} \quad \text{for } 0\le\theta\le 1$$

additionally,

$$\text{mode}[\theta] = \frac{a-1}{(a-1)+(b-1)} \quad \text{for } a>1 \text{ & } b>1$$

$$E[\theta] = \frac{a}{a+b} \quad \& \quad V[\theta] = \frac{ab}{(a+b+1)(a+b)^2} = E[\theta] \times \frac{E[1-\theta]}{a+b+1}$$

For our data on happiness in which we observed $(Y_1,\cdots,Y_{129})$ $=(y_1,\cdots,y_{129})$ with $\sum_{i=1}^{129} y_i = 118$

$\text{mode}[\theta|y_1,\cdots,y_{129}] = 0.915$

$E[\theta|y_1,\cdots,y_{129}] = 0.908$

$\text{sd}[\theta|y_1,\cdots,y_{129}] = 0.025$

If $Y_1,\cdots,Y_n|\theta$ are i.i.d. binary(θ) we have shown that

$$f(\theta|y_1,\cdots,y_n) = \theta^{\sum y_i}(1-\theta)^{n-\sum y_i} \times \frac{P(\theta)}{f(y_1,\cdots,y_n)}$$

If we compare the relative probabilities of any two θ-values, say θ_a & θ_b, we have

$$\frac{f(\theta_a|y_1,\cdots,y_n)}{f(\theta_b|y_1,\cdots,y_n)} = \frac{\theta_a^{\sum y_i}(1-\theta_a)^{n-\sum y_i} \times P(\theta)/f(y_1,\cdots,y_n)}{\theta_b^{\sum y_i}(1-\theta_b)^{n-\sum y_i} \times P(\theta)/f(y_1,\cdots,y_n)}$$

which shows that the density at θ_a relative to θ_b depends on $y_1,\cdots,y_n$ only through $\sum y_i$. From this, you can show that

$$P(\theta\in A|Y_1=y_1,\cdots,Y_n=y_n) = P\left(\theta\in A|\sum Y_i=\sum y_i\right)$$

Thus, $\sum Y_i$ contains all the information about θ that is available from the data.

The Binomial Distribution

A random variable $Y \in \{0,1,\dots,n\}$ has a binomial(n,θ) distribution if for $y \in \{0,1,\dots,n\}$

$$P(Y=y|\theta) = \binom{n}{y}\theta^y(1-\theta)^{n-y} = \text{dbinom}(y,n,\theta)$$

$E[Y] = n\theta$

$V[Y|\theta] = n\theta(1-\theta)$

Posterior Inference Under a Uniform Prior Distribution: Having observed $Y=y$ our task is to obtain the posterior distribution of θ

$$f(\theta|y) = \frac{f(y|\theta)p(\theta)}{f(y)} = \frac{\binom{n}{y}\theta^y(1-\theta)^{n-y}p(\theta)}{f(y)} = c(y)\theta^y(1-\theta)^{n-y}p(\theta)$$

where $c(y)$ is a function of y, not of θ. For the uniform distribution with $p(\theta)=1$, we can find out what $c(y)$ is

$$1 = \int_0^1 c(y)\theta^y(1-\theta)^{n-y}d\theta = c(y)\int_0^1 \theta^y(1-\theta)^{n-y}d\theta$$

$$= c(y)\frac{\Gamma(y+1)\Gamma(n-y+1)}{\Gamma(n+2)}$$

the *normalizing constant* $c(y)$ is therefore equal to $c(y) = \frac{\Gamma(n+2)}{\Gamma(y+1)\Gamma(n-y+1)}$

$$f(\theta|y) = \frac{\Gamma(n+2)}{\Gamma(y+1)\Gamma(n-y+1)}\theta^y(1-\theta)^{n-y}$$

$$= \frac{\Gamma(n+2)}{\Gamma(y+1)\Gamma(n-y+1)}\theta^{(y+1)-1}(1-\theta)^{(n-y+1)-1} = \text{beta}(y+1, n-y+1)$$

Recall from our happiness example we observed that $Y = \sum y_i = 118$ & $n = 129$. Thus $f(\theta|y_1,\dots,y_2) = f(\theta|y) = \text{beta}(119, 12$

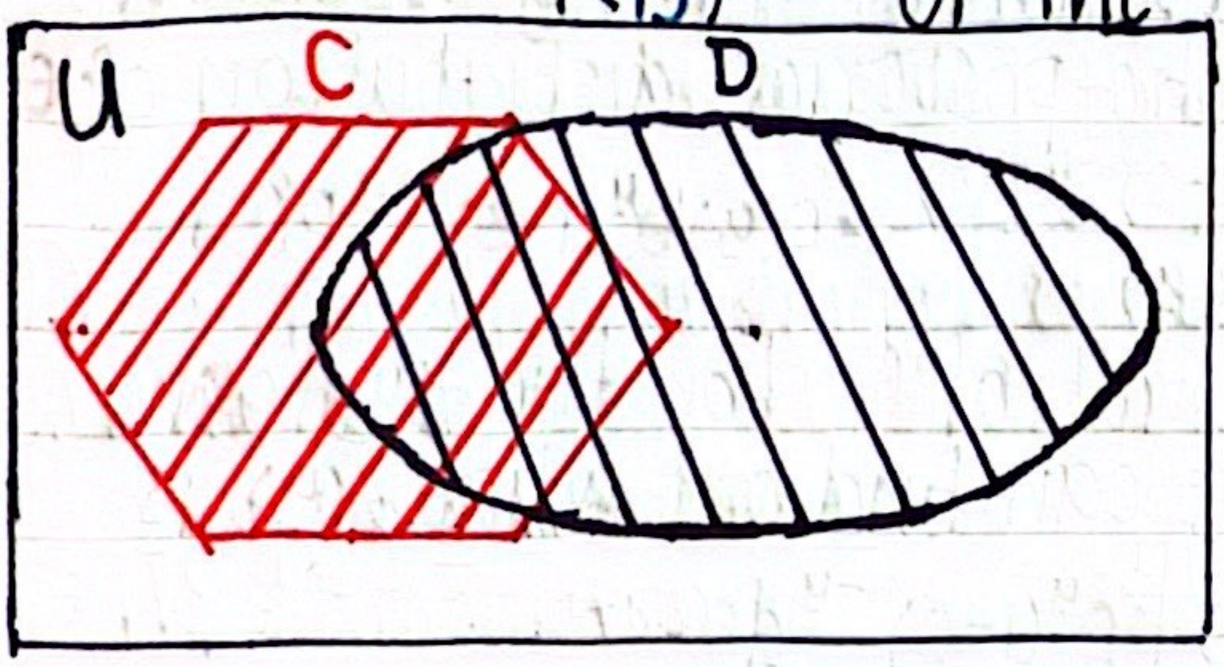
↳ In other words: the information contained in $\{Y_1=y_1,\dots,Y_n=y_n\}$ is the same information contained in $\{Y=y\}$ where $Y = \sum Y_i$ & $y = \sum y_i$

Proofs of Bayes's Rule: 3 Different Ways

$P(A|B) = \frac{P(B|A)P(A)}{P(B)}$ where A & B are events & $P(B) \neq 0$. For Events: *Proof:*

$P(A|B) = \frac{P(A \cap B)}{P(B)}$ if $P(B) \neq 0$ & $P(B|A) = \frac{P(A \cap B)}{P(A)}$ if $P(A) \neq 0$

$P(A|B) = \frac{P(B|A)P(A)}{P(B)}$ if $P(B) \neq 0$ is true by substitution of the above equations.



$P(C) = \frac{C}{U}$ $P(D) = \frac{D}{U}$

$P(D|C) = \frac{P(C \cap D)}{P(C)}$

$P(C|D) = \frac{P(C \cap D)}{P(D)}$

$P(C) \cdot P(D|C) = \frac{C}{U} \times \frac{P(C \cap D)}{C} = \frac{P(C \cap D)}{U}$

$P(D) \cdot P(C|D) = \frac{D}{U} \times \frac{P(C \cap D)}{D} = \frac{P(C \cap D)}{U}$

$\left.\begin{array}{l} P(C|D) = \frac{P(C)P(D|C)}{P(D)} \\ P(D|C) = \frac{P(D)P(C|D)}{P(C)} \end{array}\right\}$

Continious Case:

$f_{X|Y=y}(x) = \frac{f_{XY}(x,y)}{f_Y(y)}$ & $f_{Y|X=x}(y) = \frac{f_{XY}(x,y)}{f_X(x)}$ Therefore,

$f_{X|Y=y}(x) = \frac{f_{Y|X=x}(y) f_X(x)}{f_Y(y)}$

Formal Proof of Baye's Rule

Recall: Our Axioms of Probability: ~~-Aiw~~
↳ (Kolmogorov Axioms adapted conditionally)

P1 (non-negativity):
$0 = P(\neg H|H) \leq P(F|H) \leq P(H|H) = 1$

P2 (addativity of disjoint events)
$P(F \cup G) = P(F|H) + P(G|H)$ if $F \cap G = \emptyset$

P3 (chain rule for conditional probabilities)
$P(F \cap G|H) = P(G|H)P(F|G \cap H)$

QED: $$P(H_j|E) = \frac{P(E|H_j)P(H_j)}{\sum_{k=1}^{k} P(E|H_k)P(H_k)}$$

1. $P(H_j\|E) = \frac{P(H_j \cap E)}{P(E)}$	*definition of conditional probability*
2. $P(H_j \cap E) = P(E \cap H_j\|\Omega) = P(E\|H_j)P(H_j)$	**P3** Chain Rule
3. $P(E) = \sum_{k=1}^{k} P(E \cap H_k)$	*the law of total probability*
4. $P(E \cap H_k) = P(E\|H_k)P(H_k)$ Thus, $P(E) = \sum_{k=1}^{k} P(E\|H_k)P(H_k)$	**P3** Chain Rule
5. $P(H_j\|E) = \frac{P(E\|H_j)P(H_j)}{\sum_{k=1}^{k} P(E\|H_k)P(H_k)}$ ■	Substitute 3. into 2.

Posterior Distributions under a Beta Prior

The uniform prior distribution has $P(\theta)=1$ for all $\theta \in [0,1]$ which is actuallu just a special case of the beta distribution when $a=1$ & $b=1$, that is,

$$P(\theta)\frac{\Gamma(2)}{\Gamma(1)\Gamma(1)}\theta^{1-1}(1-\theta)^{1-1} = \frac{1}{1\times 1}1\times 1 = 1$$

since $\Gamma(1)=1$ this can be generalized bus the following statement:

$$\text{If } \left\{\begin{array}{l}\theta \sim \text{beta}(1,1)\ (\text{uniform}) \\ Y \sim \text{binomial}(n,\theta)\end{array}\right\} \text{ then, } \{\theta | Y=y\} \sim \text{beta}(1+y, 1+n-y)$$

Does this hold if our prior is any arbitrary beta distribution? Suppose $\theta \sim \text{beta}(a,b)$ & $Y|\theta \sim \text{binomial}(n,\theta)$

$$f(\theta|y) = \frac{f(y|\theta)P(\theta)}{f(y)}$$

$$= \frac{1}{f(y)} \times \frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)}\theta^{a-1}(1-\theta)^{b-1} \times \binom{n}{y}\theta^{y}(1-\theta)^{n-y}$$

$$= c(n,y,a,b) \times \theta^{a+y-1}(1-\theta)^{b+n-y-1}$$

$$= \text{dbeta}(\theta,\ a+y,\ b+n-y)$$

This line shows $f(\theta|y)$ as a function of θ is proportional to $\theta^{a+y-1}(1-\theta)^{b+n-y-1}$ meaning it has the same *shape* as the beta density function. This in addition to the law of total probability demanding that both $f(\theta|y)$ & the beta density must integrate to one (meaning the two densities have the same *scale*) give us enough proof to conclude that the two densities are actually given by the same function.

Conjugacy:

A class $\mathcal{P}$ of prior distributions for θ is called a **Conjugate** for a sampling model $f(y|\theta)$ if

$$P(\theta) \in \mathcal{P} \Rightarrow f(\theta|y) \in \mathcal{P}$$

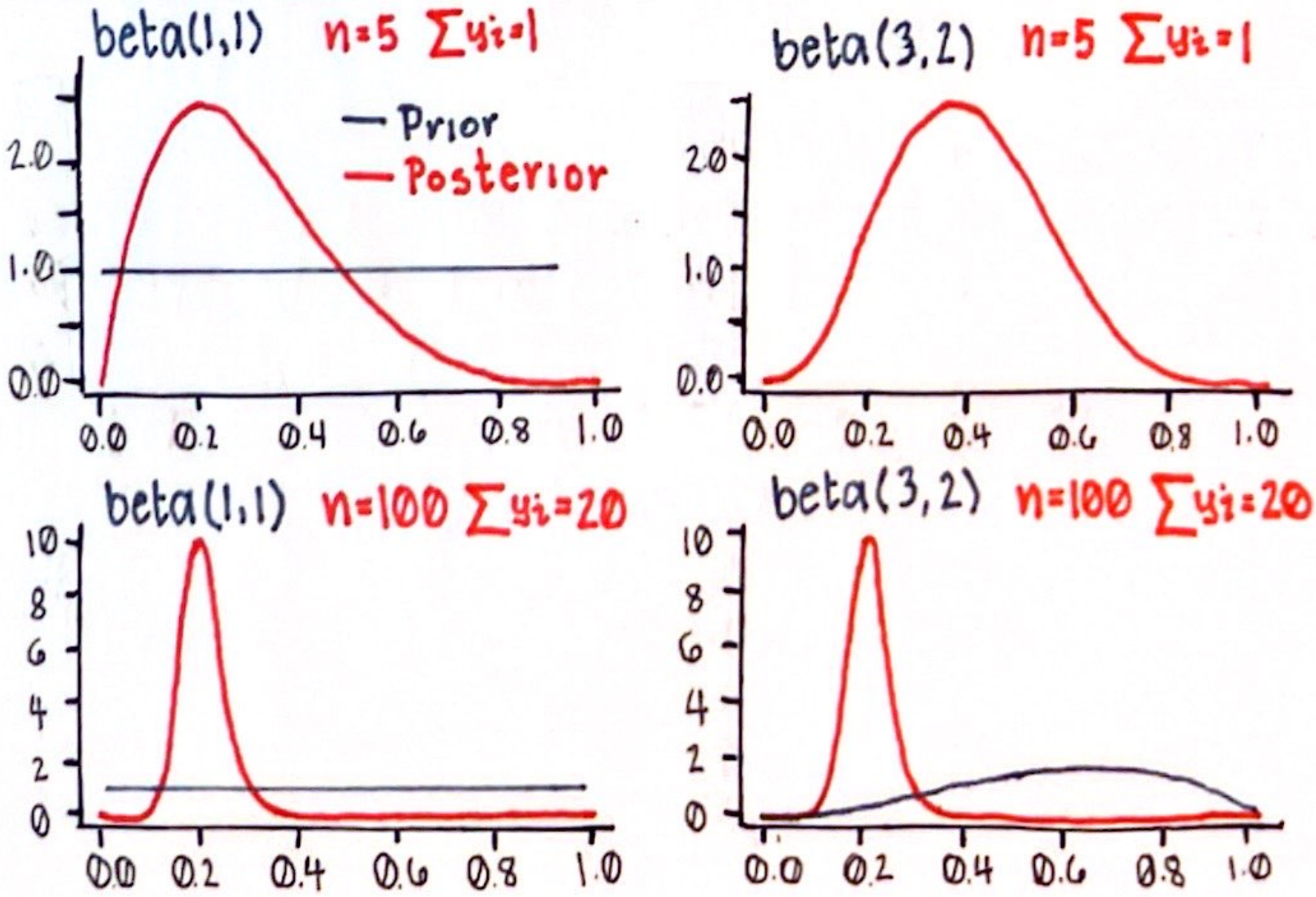
Conjugate priors make calculating the posterior much easier, but they might not actually represent our prior information. Luckily, we can mix distributions to fit our data making conjugate priors an incredibly flexible method of calculating posterior distributions.

If $\theta|\{Y=y\} \sim \text{beta}(a+y, b+n-y)$

Then $E[\theta|y] = \frac{a+y}{a+b+n}$ $\quad$ $\text{mode}[\theta|y] = \frac{a+y-1}{a+b+n-2}$

& $V[\theta|y] = \frac{E[\theta|y]E[1-\theta|y]}{a+b+n+1}$

notice how the *posterior expectation* $E[\theta|y]$ is a combination of information from the prior & information from the observed data.



depicted above is the plots of two different sample sizes & two different prior distributions.

Under the previously discussed model & prior distribution, the *posterior expectation* is a weighted average of the *prior expectation* & the observed sample mean

$$E[\theta|y] = \frac{a+y}{a+b+n} = \frac{a+b}{a+b+n}\frac{a}{a+b} + \frac{n}{a+b+n}\frac{y}{n}$$

$$= \frac{a+b}{a+b+n} \times \text{Prior expectation} + \frac{n}{a+b+n} \text{ data mean}$$

Where the weights are proportional to $a+b$ & n respectively
This leads to the interpretation of a & b as "prior data"

$a \approx$ "prior number of 1's"
$b \approx$ "prior number of 0's"
$a+b \approx$ "prior sample size" ← If our sample size n is larger than our prior sample size, then a majority of our information about the posterior distribution of θ should be coming from our data as opposed to our prior distribution.

↳ In the case that $n \gg a+b$, then, $\frac{a+b}{a+b+n} \approx 0$, $E[\theta|y] \approx \frac{y}{n}$, & $V[\theta|y] \approx \frac{1}{n}\frac{y}{n}(1-\frac{y}{n})$

A Note on Bayesian Prediction:

1. The predictive distribution does <u>not</u> depend on any unknown quantities; if it did it wouldn't be predictive
2. The predictive <u>does</u> depend on our observed data that is, $\tilde{Y}$ is <u>not</u> independent of $Y_1, \ldots, Y_n$. This is due to the fact that observing $Y_1, \ldots, Y_n$ gives information about θ which in turn provides information about $\tilde{Y}$ → if $\tilde{Y}$ were independent of $Y_1, \ldots, Y_n$ our whole inferential process would be in vain

↳ see page 61 for more thourogh treatment of prediction under the Binomial Model

Confidence Regions:

We will often want to identify regions of the parameter space that are likely to contain the true value of the parameter. To do so, after observing data ($Y=y$) we can construct an inteval $[l(y), u(y)]$ such that the probability that $l(y) < \theta < u(y)$ is sufficiently large. An Interval $[l(y), u(y)]$ based on the observed data $Y=y$ is said to have 95% **Bayesian Coverage** for θ if

$$P(l(y) < \theta < u(y) \mid Y=y) = 0.95$$

To interpret this interval we would say something like "We are 95% confident, *after observing our sample data*, that the true value of θ is between $l(y)$ & $u(y)$" Whereas, the traditional definition of frequentist coverage covers the true value *before* the sample is observed A random interval $[l(Y), u(Y)]$ has 95% **Frequentist Coverage** for θ if *before* the data is gathered

$$P(l(y) < \theta < u(y) \mid \theta) = 0.95$$

Can a confidence interval have the same Bayesian & Frequentist Coverage? Hartigan(1966) showed that for the intervals we will be studying, if an interval has 95% Bayesian Coverage then the following holds

$$P(l(Y) < \theta < u(Y) \mid \theta) = 0.95 + \varepsilon_n \quad \text{where } |\varepsilon_n| < \frac{a}{n} \text{ for some constant } a$$

This means that a confidence interval with 95% Bayesian coverage will have approximately 95% Frequentist coverage as well (at least asymptotically), but the two coverages of a single interval will usually be fairly similar, nonasymptonically. This means that mathematically speaking: $\lim_{n\to\infty} P(l(Y) < \theta < u(Y)) = 0.95$ given the interval described above.

Quantile-Based Intervals:

Perhaps the easiest way to find a confidence Interval is to use the posterior quantiles. To make a $100 \times (1-\alpha)\%$ quantile-based confidence interval, find numbers $\theta_{\alpha/2}, \theta_{1-\alpha/2}$ such that

$$\left.\begin{aligned} P(\theta < \theta_{\alpha/2} \mid Y=y) &= \alpha/2 \\ P(\theta > \theta_{1-\alpha/2} \mid Y=y) &= \alpha/2 \end{aligned}\right\} \text{ \& The numbers } \theta_{\alpha/2}, \theta_{1-\alpha/2} \text{ are the } \alpha/2 \text{ \& } 1-\alpha/2 \text{ posterior quantiles}$$

Of θ, & as such it follows that

$$\begin{aligned} P(\theta \in [\theta_{\alpha/2}, \theta_{1-\alpha/2}] \mid Y=y) &= 1 - P(\theta \notin [\theta_{\alpha/2}, \theta_{1-\alpha/2}] \mid Y=y) \\ &= 1 - [P(\theta < \theta_{\alpha/2} \mid Y=y) + P(\theta > \theta_{1-\alpha/2} \mid Y=y)] \\ &= 1-\alpha \end{aligned}$$

Example:
Binomial Sampling & Uniform Prior:

Suppose, that out of $n=10$ conditionally independent draws of a binary random variable we observe $Y=2$ ones. Using a Uniform prior distribution for θ, the posterior distribution is $\theta \mid \{Y=2\} \sim \text{beta}(1+2, 1+8)$. A 95% posterior confidence interval can be obtained from the 0.025 & 0.975 quantiles of this beta distribution. These quantiles are 0.06 & 0.52 respectively, & so the posterior probability that $\theta \in [0.06, 0.52]$ is 95%

```
> a <-1; b <-1 #Prior
> n <-10; y <-2 #data
> qbeta(c(0.025, 0.975), a+y, b+n+y)
```

```
[1] 0.06021773 0.51775585
```

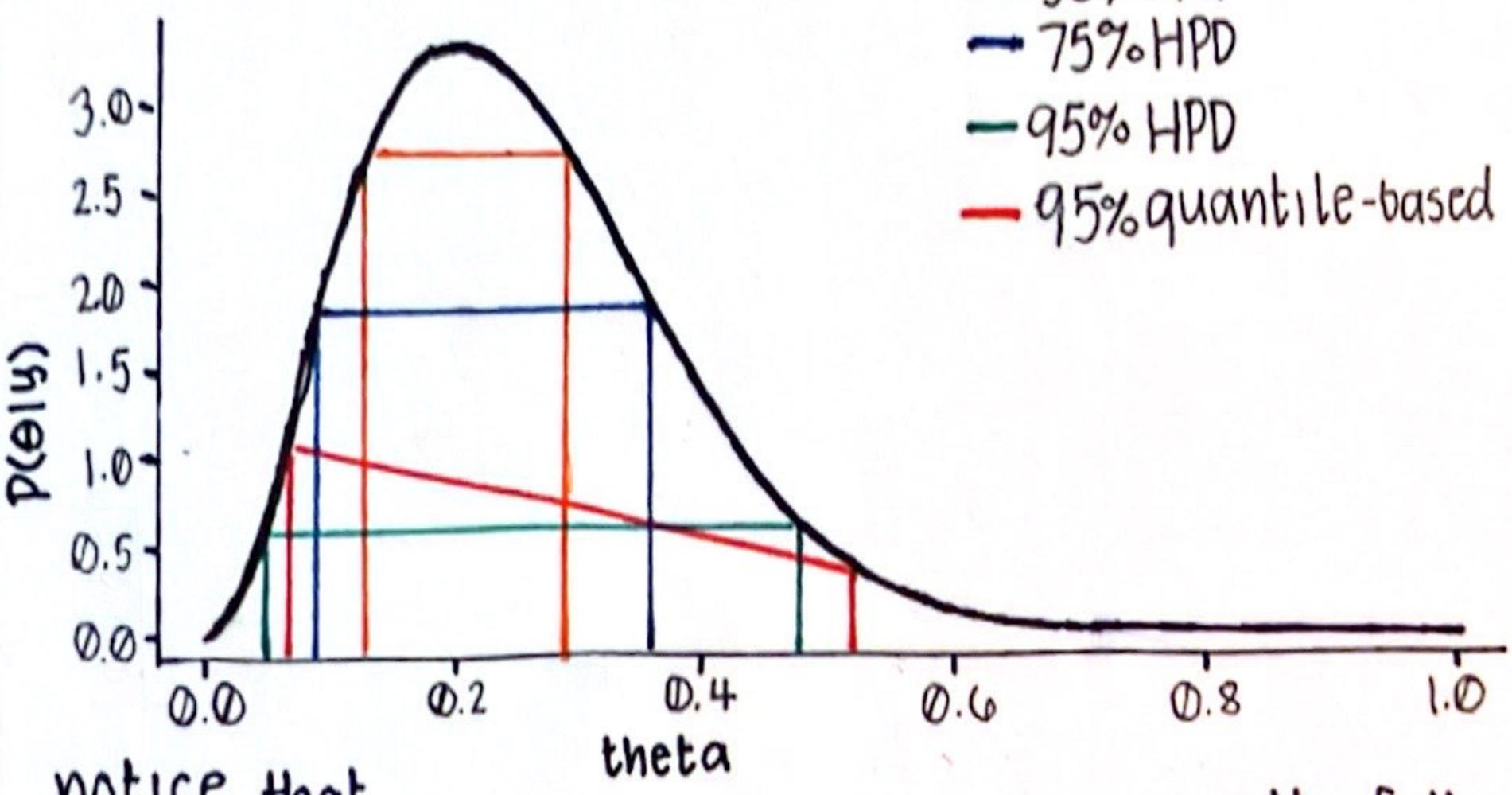
Above is the snippet of R-code that was used to preform our calculation.

Highest Posterior Density Regions

A $100 \times (1-\alpha)\%$ **Highest Posterior Density (HPD) Region** consists of a subset of the parameter space $S(y) \subset \Theta$ such that

1. $P(\theta \in S(y) | Y=y) = 1-\alpha$;
2. If $\theta_a \in S(y)$ & $\theta_b \notin S(y)$, then $f(\theta_a | Y=y) > f(\theta_b | Y=y)$

All points in a HPD region have a higher posterior density than points outside the region. However, in the case of multimodal data the HPD region may not be a single continious interval. Gradually move a horizontal line down the density plot until the posterior probabilities of the θ-values in the region is equal to $(1-\alpha)$. The plot below visualizes the construction of a HPD region for different values of α



notice that the only non-perpendicular line is a result of the singular quantile-based region.

3.2: The Poisson Model

A random variable Y is said to have a *Poisson Distribution* with mean θ if

$$P(Y=y|\theta) = dpois(y,\theta) = \theta^y e^{-\theta}/y! \quad \text{for } y \in \{1,2,\dots\}$$

meaning that,

$$E[Y|\theta] = V[Y|\theta] = \theta \longrightarrow$$

The possion distribution is good at modeling infrequent discrete events in a fixed period of time (such as hourly arrival of customers to a store).

Posterior Inference

If we model $Y_1,\dots,Y_n$ as i.i.d. poisson distributed with mean θ, then, the joint density of the sample is:

$$P(Y_1=y_1,\dots,Y_n=y_n|\theta) = \prod_{i=1}^{n} f(y_i|\theta) = \prod_{i=1}^{n} \frac{1}{y_i!}\theta^{y_i}e^{-\theta}$$

$$= c(y_1,\dots,y_n)\theta^{\Sigma y_i}e^{-n\theta}$$

comparing two values of θ *a posteriori*, we have:

$$\frac{f(\theta_a|y_1,\dots,y_n)}{f(\theta_b|y_1,\dots,y_n)} = \frac{c(y_1,\dots,y_n)}{c(y_1,\dots,y_n)} \frac{e^{-n\theta_a}}{e^{-n\theta_b}} \frac{\theta_a^{\Sigma y_i}}{\theta_b^{\Sigma y_i}} \frac{p(\theta_a)}{p(\theta_b)}$$

$$= \frac{e^{-n\theta_a}}{e^{-n\theta_b}} \frac{\theta_a^{\Sigma y_i}}{\theta_b^{\Sigma y_i}} \frac{p(\theta_a)}{p(\theta_b)}$$

As in the case of i.i.d. binary models, $\sum_{i=1}^{n} Y_i$ contains the same information about θ that is available from the data, meaning that $\sum_{i=1}^{n} Y_i$ is a sufficient statistic

A Further Note on Conjugate Priors: For the Poisson sampling model, our posterior distribution is of the form

$$f(\theta|y_1,\dots,y_n) \propto p(\theta) \times f(y_1,\dots,y_n|\theta) = p(\theta) \times \theta^{\Sigma y_i} e^{-n\theta}$$

This is a special case of a foundational class of probability distributions: The Gamma family. Let us next meet the matriarch herself.

The Gamma Distribution:

An uncertain positive quantity θ has a gamma(a,b) if

$$p(\theta)=\text{dgamma}(\theta,a,b)=\frac{b^a}{\Gamma(a)}\theta^{a-1}e^{-b\theta} \quad \text{for } \theta,a,b>0$$

$$E[\theta]=\frac{a}{b} \qquad V[\theta]=\frac{a}{b^2} \qquad \text{mode}[\theta]=\begin{cases}(a-1)/b & \text{if } a>1\\ 0 & \text{if } a\le 1\end{cases}$$

Posterior Distribution of θ: Suppose $Y_1,\ldots,Y_n|\theta \sim$ i.i.d. poisson(θ) & $p(\theta)=\text{dgamma}(\theta,a,b)$. Then,

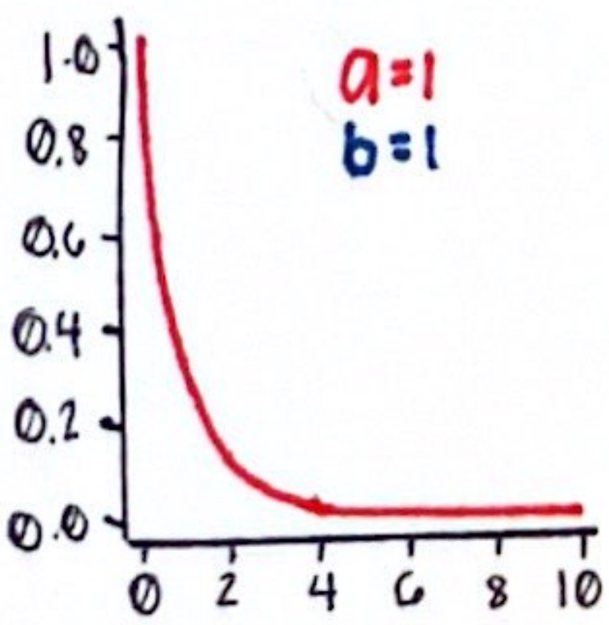
$$f(\theta|y_1,\ldots,y_n)=\frac{p(\theta)\times f(y_1,\ldots,y_n|\theta)}{f(y_1,\ldots,y_n)}$$

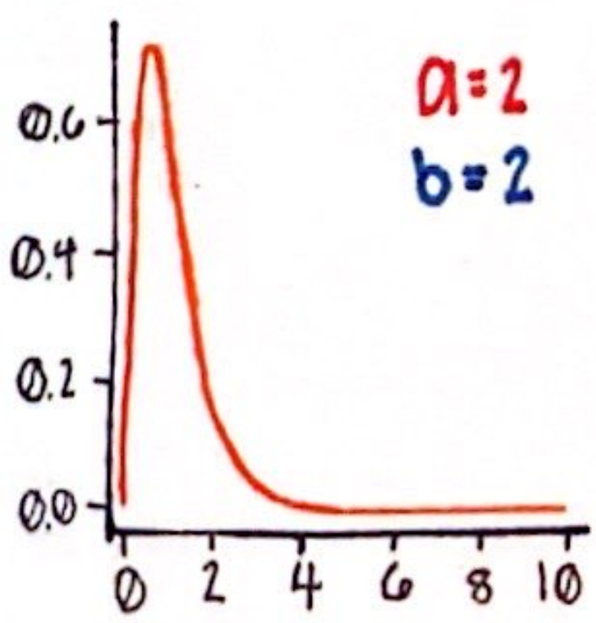
$$=\left\{\theta^{a-1}e^{-b\theta}\right\}\times\left\{\theta^{\sum y_i}e^{-n\theta}\right\}\times c(y_1,\ldots,y_n,a,b)$$

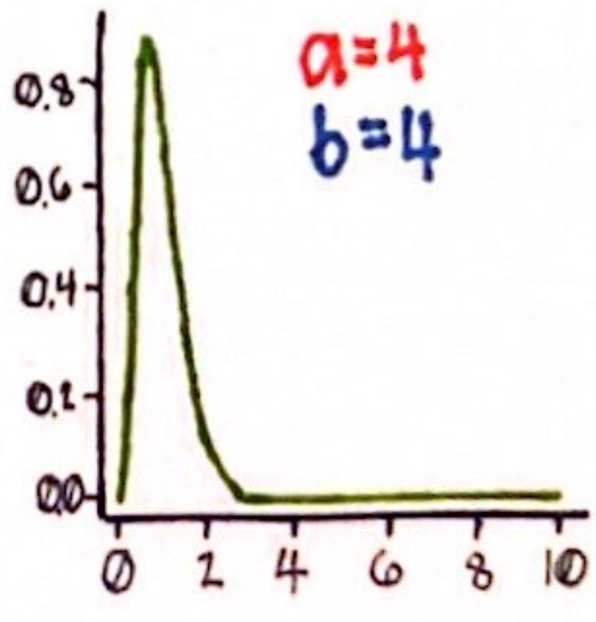
This → $$=\left\{\theta^{a+\sum y_i-1}e^{-(b+n)\theta}\right\}\times c(y_1,\ldots,y_n,a,b)$$

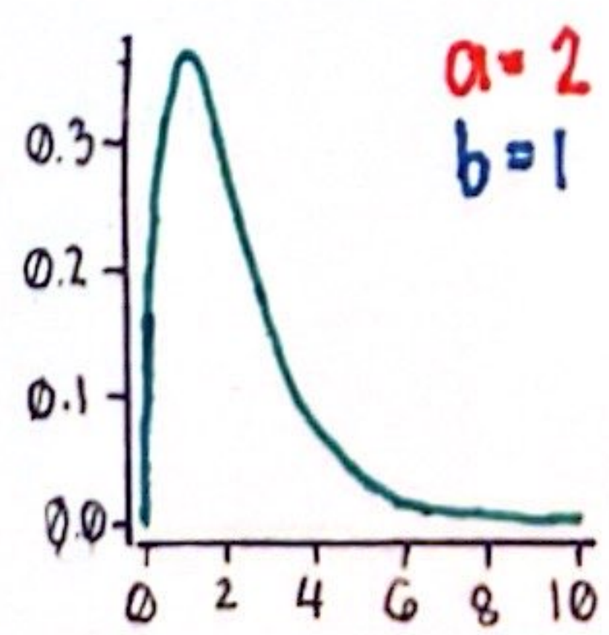
confirms the conjugacy of the gamma distribution for the Poisson sampling model:

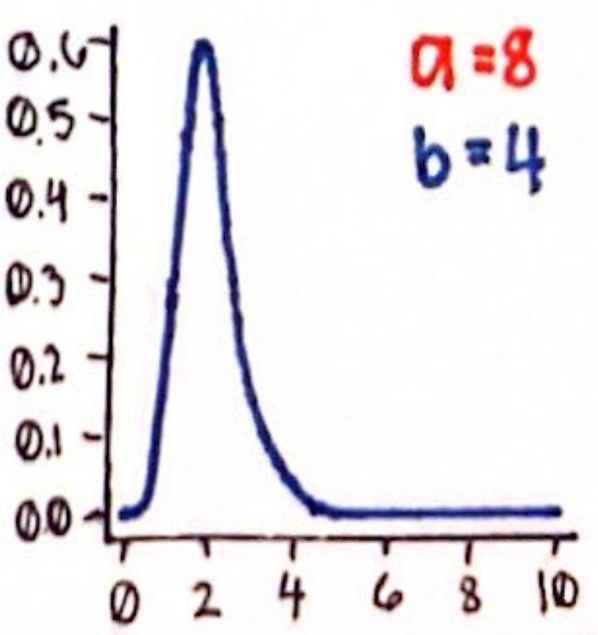
$$\left.\begin{aligned}\theta &\sim \text{gamma}(a,b)\\ Y_1,\ldots,Y_n|\theta &\sim \text{poisson}(\theta)\end{aligned}\right\} \Rightarrow \{\theta|Y_1,\ldots,Y_n\}\sim\text{gamma}\left(a+\sum_{i=1}^n Y_i,\ b+n\right)$$

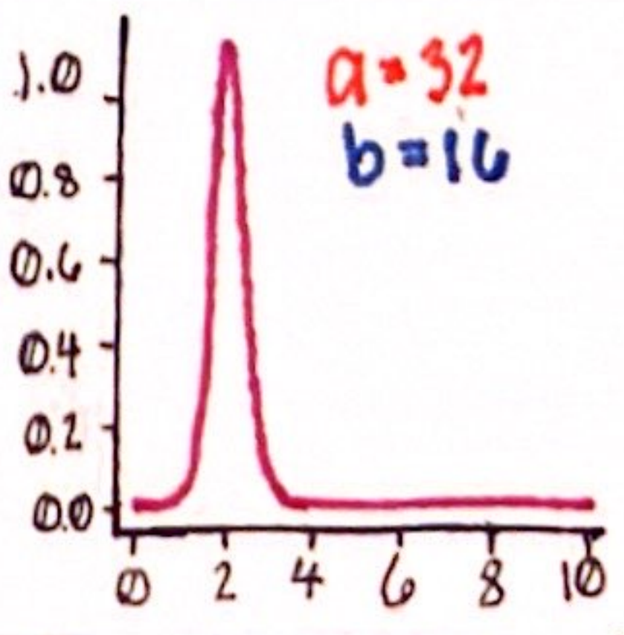












Shown above is the shape of the gamma distribution for various values of a & b

Estimation in the Poisson sampling model procedes similarly as it does under the Binomial model. The posterior expectation of θ is a convex combination of the prior expectation & the sample mean.

$$E[\theta|y_1,\dots,y_n] = \frac{a+\sum y_i}{b+n} = \frac{b}{b+n}\frac{a}{b} + \frac{n}{b+n}\frac{\sum y_i}{n}$$

b is interpreted as the number of prior observations
a is interpreted as the sum of the counts from b prior observations.

↳ As a result, for very large n, the sample data dominates the prior belief; mathematically that is:

$$n \gg b \Rightarrow E[\theta|y_1,\dots,y_n] \approx \tilde{y}, \; V[\theta|y_1,\dots,y_n] \approx \tilde{y}/n$$

Example: Birthrates: Over the course of the 1990s the *General Social Survey* gathered data on the education & number of children of 155 women who were 40 years old at the time. Fertility rates in the 1970s were historically low; we will compare the number of children had by women who hold college degrees to those who do not hold college degrees. Let $Y_{11},\dots,Y_{n_11}$ denote the number of children for the n_1 women without degrees & $Y_{12},\dots,Y_{n_22}$ denote the number of children for the n_2 women who have college degrees. We will use the following models

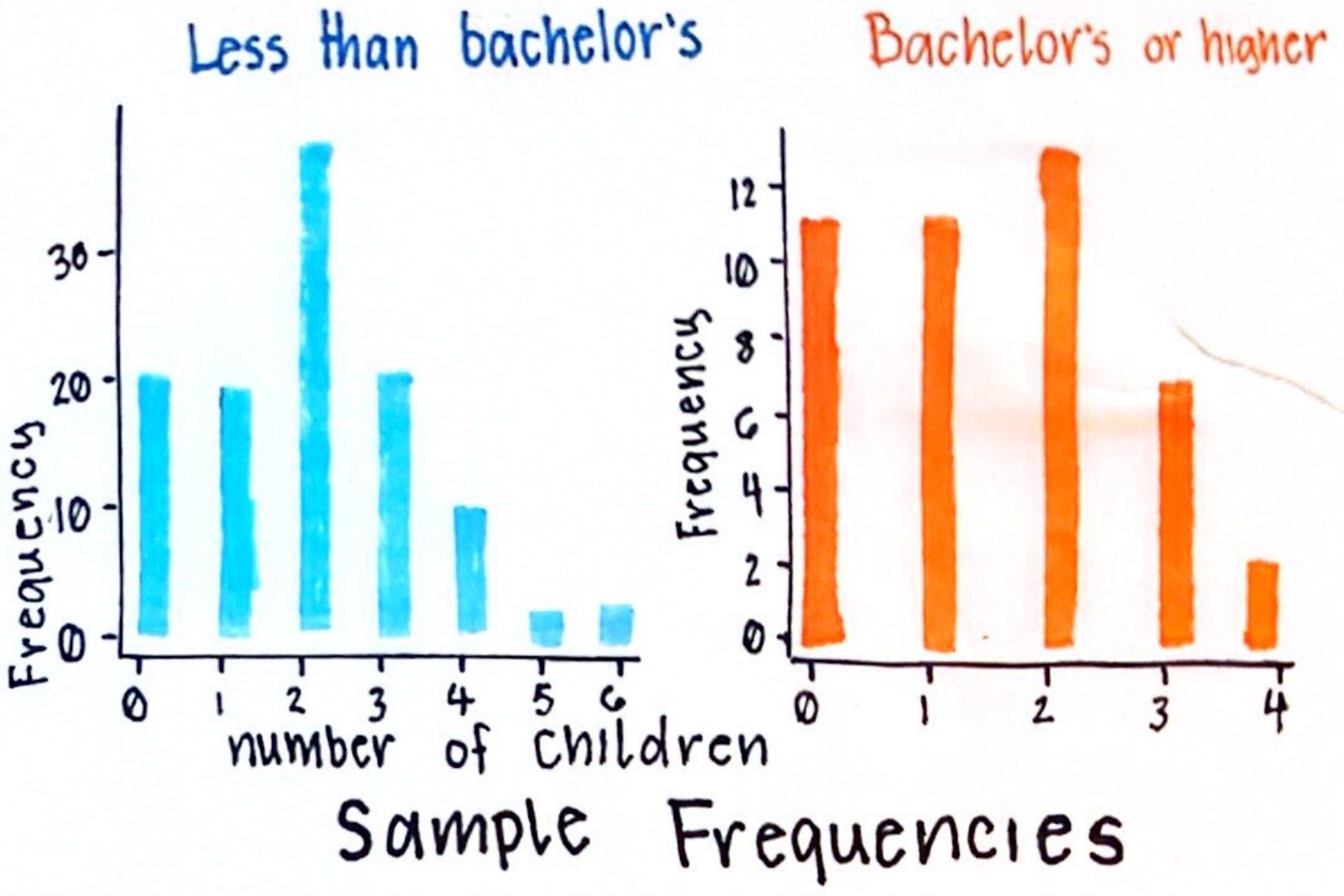
$$Y_{11},\dots,Y_{n_11}|\theta_1 \sim \text{i.i.d. Poisson}(\theta_1)$$
$$Y_{11},\dots,Y_{n_22}|\theta_2 \sim \text{i.i.d. Poisson}(\theta_2)$$

According to our sample suppose we observed

no degree held: $\sum_{i=1}^{n} Y_{i1} = 217$ $n_1 = 111$, $\tilde{Y}_1 = 1.95$

degree held: $\sum_{i=1}^{n_2} Y_{i2} = 66$ $n_2 = 44$, $\tilde{Y}_2 = 1.50$

Example: Birthrates 2/3



Sample Frequencies

In the case that $\{\theta_1, \theta_2\} \sim$ i.i.d. gamma($a=2, b=1$) we have the following posterior distributions:

$\theta_1 | \{n_1 = 111, \sum Y_{i1} = 217\} \sim \text{gamma}(2+217, 1+111) = \text{gamma}(219, 212)$

$\theta_2 | \{n_2 = 44, \sum Y_{i2} = 66\} \sim \text{gamma}(2+66, 1+44) = \text{gamma}(68, 45)$

Posterior means, modes, & 95% quantile-based intervals for θ_1 & θ_2 can be obtained using these posterior gamma distributions respectively.

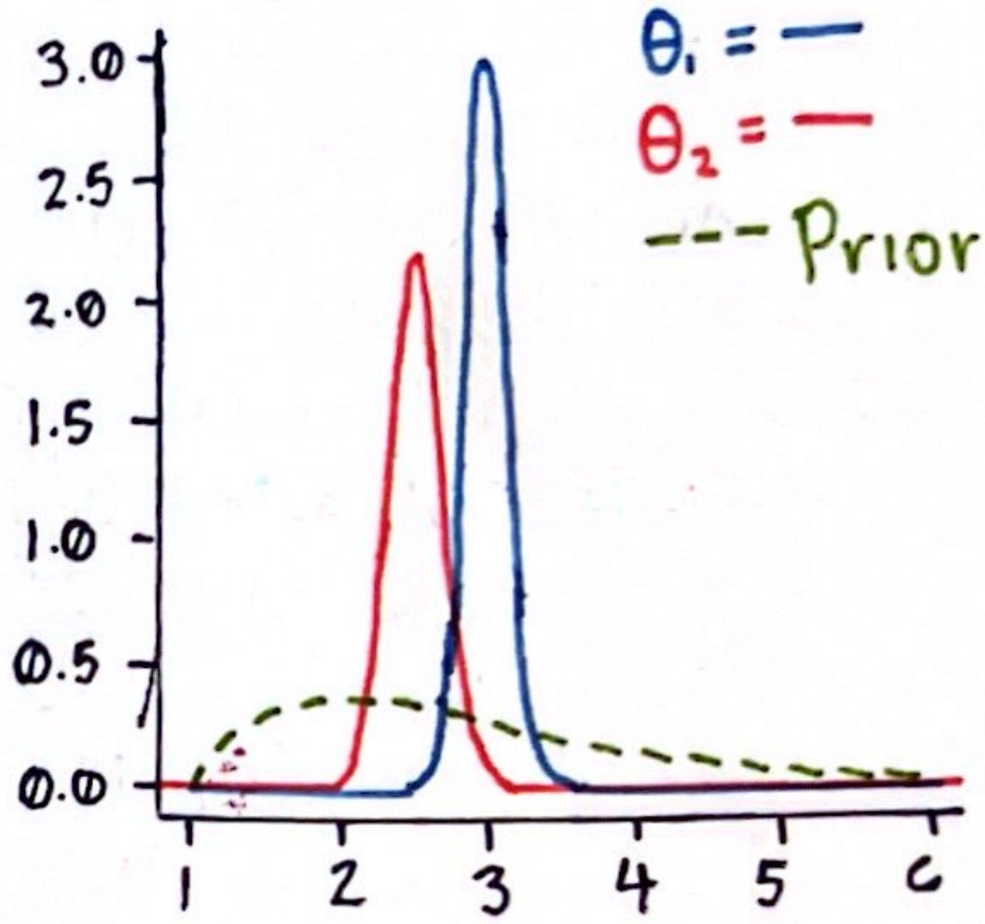
```
> a <- 2;  b <- 1   #Prior Parameters
> n1 <- 111;  sy1 <- 217   # No degree data
> n2 <- 44;  sy2 <- 66   # Degree data

> Posterior_Mean <- (a+sy)/(b+n)
> Posterior_Mode <- (a+sy-1)/(b+n)
> Posterior_CI <- qgamma(c(0.025, 0.975), a+sy, b+n)
```

no degree → mean = 1.955 mode = 1.947 CI = (1.705, 2.223)
degree holder → mean = 1.511 mode = 1.489 CI = (1.173, 1.891)

Example: Birthrates 3/3

The posterior densities for the sample mean are plotted below. These two densities are indicative of substantial evidence that $\theta_1 > \theta_2$; that is:
$P(\theta_1 > \theta_2 \mid \Sigma y_{i1} = 217, \Sigma Y_{i2} = 66) = 0.97$



Now, consider two randomly selected individuals, one from each populations. To what extent do we expect the person without a degree to have more children than the other person?

↳ The predictive distribution for a Poisson model for $\tilde{y} \in \{0,1,2,\ldots\}$ is negative binomially distributed with parameters $(a+\Sigma y_i,\ b+n)$ with $E[\tilde{Y} \mid y_1,\ldots,y_n] = \frac{a+\Sigma y_i}{b+n} = E[\theta \mid y_1,\ldots,y_n]$ &

$V[\tilde{Y} \mid y_1,\ldots,y_n] = \frac{a+\Sigma y_i}{b+n}\,\frac{b+n+1}{b+n} = V[\theta \mid y_1,\ldots,y_n] \times (b+n+1)$
$= E[\theta \mid y_1,\ldots,y_n] \times \frac{b+n+1}{b+n}$

If we were to plot these predictive distributions on the same graph we would notice much more overlap than there is in the plot of the posterior distributions (pictured above). Which is indicative of the extreme importance of the difference between $\{\theta_1 > \theta_2\}$ & $\{\tilde{Y}_1 > \tilde{Y}_2\}$; Strong evidence of a difference between two popoulations is not itself evidence that the present difference is large.

3.3 Exponential Families & Conjugate Priors

To further generalize, the Binomial & Poisson models previously discussed are one parameter instances of what are called *exponential family Models*. Which are any model whose density can be expressed in the form

$$f(y|\phi) = h(y)c(\phi)e^{\phi t(y_i)}$$

where ϕ is the unknown parameter & $t(y)$ is the sufficient statistic. Diaconis & Ylvisaker (1979) study conjugate prior distributions for the general exponential family models, & in particular prior distributions of the form:

$$f(\phi|n_0,t_0) = k(n_0,t_0)c(\phi)^{n_0}e^{n_0t_0\phi}$$

Combining such a prior with information from observing $Y_1,\dots,Y_n \sim$ i.i.d. $f(y|\theta)$ yields the following posterior distribution

$$f(\phi|y_1,\dots,y_n) \propto p(\phi)f(y_1,\dots,y_n|\phi)$$

$$\propto c(\phi)^{n_0+n}\exp\left\{\phi \times \left[n_0t_0 + \sum_{i=1}^{n} t(y_i)\right]\right\}$$

$$\propto f(\phi|n_0+n, n_0t_0 + n\bar{t}(y))$$

where $\bar{t}(y) = \frac{\sum t(y_i)}{n}$

This formulation of the posterior distribution allows us to interpret n_0 as the "prior sample size" & t_0 as the "prior guess" of $t(Y)$. Diaconis & Ylvisaker (1979) make this interpretation more precise by showing that

$$E[t(Y)] = E[E[t(Y)|\phi]] = E[-c'(\phi)/c(\phi)] = t_0$$

In other words, t_0 is the prior expected value of $t(Y)$, & n_0 is a measure of how informative the prior is. As a function of ϕ, $f(\phi|n_0,t_0)$ has the same shape as a likelyhood $f(\tilde{y}_1,\dots,\tilde{y}_{n_0}|\phi)$ based on n_0 "prior observations" $\tilde{y}_1,\dots,\tilde{y}_{n_0}$ for which $\sum t(\tilde{y}_i)/n_0 = t_0$. Which means the prior distribution $f(\phi|n_0,t_0)$ contains the same amount of information as would be contained in n_0 independent observations randomly sampled from the population.

Exponential Family Representation of the Binomial

The exponential family representation of the Binomial(θ) model can be obtained from the density function for a single binary random variable as follows:

$$f(y|\theta) = \theta^y(1-\theta)^{1-y} = \left(\frac{\theta}{1-\theta}\right)^y(1-\theta) = e^{\phi y}(1+e^{\phi})^{-1}$$

Where $\phi = \ln\left(\frac{\theta}{1-\theta}\right)$ is the log-odds. The conjugate prior of ϕ is thus given by the following $f(\phi|n_0, t_0) \propto (1+e^{\phi})^{-n_0} e^{n_0 t_0 \phi}$ where t_0 is the prior expectation of $t(y)=y$ (i.e. t_0 is our prior probability that $Y=1$). Using the change of variable technique (shudder) the prior distribution of θ becomes such that:

$$f(\theta|n_0, t_0) \propto \theta^{n_0 t_0 - 1}(1-\theta)^{n_0(1-t_0)-1}$$

which is a beta($n_0 t_0$, $n_0(1-t_0)$) distribution. Finally, a weakly informative prior can be obtained by setting t_0 equal to our prior expectation of Y & $n_0 = 1$ yielding a beta(1/2, 1/2) distribution if our expectation is 1/2

↳ This is called *Jeffrey's Prior*

Exponential Family Representation of the Poisson

The Poisson(θ) can be shown to be a member of the exponential family of models with $t(y)=y$, $\phi = \ln(\theta)$, & $c(\phi) = \exp(e^{-\phi})$. Which makes the conjugate prior distribution for ϕ equal to $f(\phi|n_0, t_0) = \exp(n_0 e^{-\phi}) e^{n_0 t_0 y}$ where t_0 is the prior expectation of Y. This translates into a prior density for θ of the form:

$$f(\theta|n_0, t_0) \propto \theta^{n_0 t_0 - 1} e^{-n_0 \theta}$$

which is a gamma($n_0 t_0$, n_0) distribution. A weakly informative prior can be obtained by setting $n_0 = 1$ which yields a gamma(t_0, 1) prior distribution. The posterior of such a prior is given by $\{\theta|y_1, \ldots, y_n\} \sim$ gamma($t_0 + \sum y_i$, $1+n$)

Chapter 4: Monte Carlo Approximation

4.1: The Monte Carlo Method

Obtaining exact values for posterior quantities is difficult, or often impossible. However, if we can generate random sample values of the parameters from their posterior distribution, then we can approximate these quantities to an arbitrary degree of precision using the Monte Carlo method.

On the preceding pages we obtained the following posterior distributions for birthrates of women who do not have a degree and those do, respectively:

$$f(\theta_1 | \sum_{i=1}^{111} Y_{i1} = 217) = \text{dgamma}(\theta_1, 219, 112)$$

$$f(\theta_2 | \sum_{i=1}^{44} Y_{i2} = 66) = \text{dgamma}(\theta_2, 68, 45)$$

We modeled θ_1 & θ_2 as conditionally independent given the data.

we found $P(\theta_1 > \theta_2 | \sum Y_{i1} = 217, \sum Y_{i2} = 66) = 0.97$. What does the work for that number look like? $P(\theta_1 > \theta_2 | y_{11}, \dots, y_{n_2 2}) =$

It looks like this & my time is just more valuable than that. Thank god for the diva named Monte Carlo.

$$= \int_0^\infty \int_0^{\theta_1} f(\theta_1, \theta_2 | y_{11}, \dots, y_{n_2 2}) \, d\theta_2 \, d\theta_1$$

$$= \int_0^\infty \int_0^{\theta_1} \text{dgamma}(\theta_1, 219, 112) \times \text{dgamma}(\theta_2, 68, 45)$$

$$\hookrightarrow = \frac{112^{219} 45^{68}}{\Gamma(219)\Gamma(68)} \int_0^\infty \int_0^{\theta_1} \theta_1^{218} \theta_2^{67} e^{-112\theta_1} e^{-45\theta_2} \, d\theta_2 \, d\theta_1$$

Let θ be a parameter of interest & let $y_1, \dots, y_n$ be the numerical values of a sample from the distribution $f(y_1, \dots, y_n | \theta)$. Suppose we could sample some number S of of independent, random θ-values from the posterior distribution $f(\theta | y_1, \dots, y_n)$: $\theta^{(1)}, \dots, \theta^{(S)} \sim$ i.i.d. $f(\theta | y_1, \dots, y_n)$

Then, the empirical distribution of $\{\theta^{(1)}, \dots, \theta^{(S)}\}$ would approximate $f(\theta | y_1, \dots, y_n)$ with the approximation improving with increasing S

The empirical distribution of $\{\theta^{(1)},\dots,\theta^{(S)}\}$ is known as a **Monte Carlo Approximation** to $f(\theta|y_1,\dots,y_n)$. The figure below shows sucessive Monte Carlo approximations to the density of gamma(68, 45). Additionally, let $g(\theta)$ be (just about) any function; then, the *Law of Large Numbers* states that if $\theta^{(1)},\dots,\theta^{(S)}$ are i.i.d. samples from $f(\theta|y_1,\dots,y_n)$ then,

$$\frac{1}{S}\sum_{s=1}^{S} g(\theta^{(s)}) \to E[g(\theta|y_1,\dots,y_n)] = \int g(\theta) f(\theta|y_1,\dots,y_n)\,d\theta \text{ as } S\to\infty$$

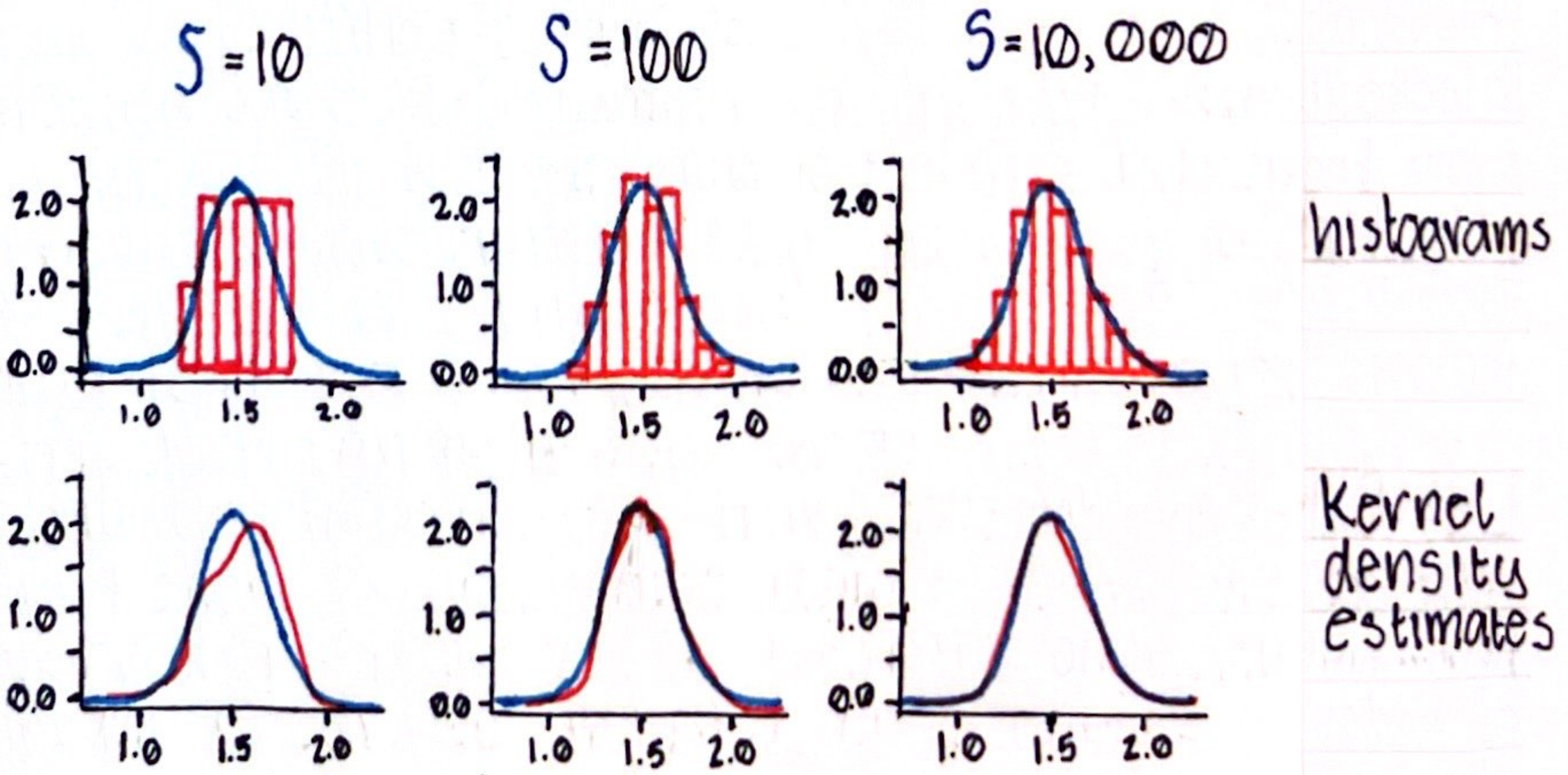
which implies that as $S\to\infty$

$\#(\theta^{(s)}\le c)/S \to P(\theta\le c|y_1,\dots,y_n)$,
the median of $\{\theta^{(1)},\dots,\theta^{(S)}\}\to\theta_{1/2}$,
the α-% of $\{\theta^{(1)},\dots,\theta^{(S)}\}\to\theta_\alpha$

$\bar{\theta} = \sum_{s=1}^{S}\theta^{(s)}\times\frac{1}{S}\to E[\theta|y_1,\dots,y_n]$ &

$\sum_{s=1}^{S}(\theta^{(s)}-\bar{\theta})^2\times\frac{1}{S-1}\to V[\theta|y_1,\dots,y_n]$

as well as



Monte Carlo Approximation of gamma(68,45)

Just about any aspect of a posterior distribution can be approximaty arbitrarily exact with a large enough Monte Carlo sample.

4.2: Posterior Inference for Arbitrary Functions

Suppose we are interested in the posterior distribution of some computable function $g(\theta)$ of θ. In the Binomial model, for example, we are sometimes interested in the log-odds: $\log \text{odds} = \log \frac{\theta}{1-\theta} = \gamma$ The *Law of Large Numbers* state that if we generate a sequence $\{\theta^{(1)}, \theta^{(2)}, \ldots\}$ from the posterior distribution of θ, then, the average value of $\log \frac{\theta}{1-\theta^{(s)}}$ converges to $E[\log \frac{\theta}{1-\theta} | y_1, \ldots, y_n]$ However, we may also be interested in other aspects of the posterior distribution of $\gamma = \log \frac{\theta}{1-\theta}$. Monte Carlo say hi.

Sample $\theta^{(1)} \sim f(\theta | y_1, \ldots, y_n)$, compute $\gamma^{(1)} = g(\theta^{(1)})$
Sample $\theta^{(2)} \sim f(\theta | y_1, \ldots, y_n)$, compute $\gamma^{(2)} = g(\theta^{(2)})$
⋮ ⋮ ⋮
sample $\theta^{(S)} \sim f(\theta | y_1, \ldots, y_n)$, compute $\gamma^{(S)} = g(\theta^{(S)})$
} independently

The sequence $\{\gamma^{(1)}, \ldots, \gamma^{(S)}\}$ constitutes S independent samples from $f(\gamma | y_1, \ldots, y_n)$, and so as $S \to \infty$

$$\bar{\gamma} = \sum_{s=1}^{S} \gamma^{(s)}/S \to E[\gamma | y_1, \ldots, y_n] \quad \& \quad \sum_{s=1}^{S} \frac{(\gamma^{(s)} - \bar{\gamma})^2}{S-1} \to V[\gamma | y_1, \ldots, y_n] \&$$

The empirical distribution of $\{\gamma^{(1)}, \ldots, \gamma^{(S)}\} \to f(\gamma | y_1, \ldots, y_n)$

Example: Log-Odds: Of the $n=860$ individuals in a sample $y=441$ (51%) had the trait; whereas of the $m=1011$ individuals in a sample from a similar population $y=353$ (35%). Let θ be the population proportion. Using a Binomial sampling model & a uniform prior yields a posterior distribution of θ of beta(442, 420). Using the Monte Carlo algorithm described above, we can obtain samples of the log-odds $\gamma = \log[\theta/(1-\theta)]$ from both the prior & posterior distributions of γ with a few lines of R code.

4.3: Predictive Sampling

The predictive distribution for a random variable $\tilde{Y}$ is a probability distribution for $\tilde{Y}$ such that:

1. Known quantities have been conditioned on
2. Unknown quantities have been integrated out

For example, let $\tilde{Y}$ be the number of children of a person who is sampled from the population of women over 40 who hold degrees. If we knew the mean birthrate θ of this population, we could describe our uncertainty with a poisson(θ) distribution.

We cannot use this model to make predictions, however, as we do not actually know θ. If we did not have any sample data from the population; we obtain our predictive distribution by integrating out θ. If $\theta \sim \text{gamma}(a,b)$ then we have shown that the predictive distribution is negative binomial(a, b)

Sampling Model:

$$P(\tilde{Y}=\tilde{y}|\theta)=f(\tilde{y}|\theta)=\theta^{\tilde{y}}e^{-\theta}/\tilde{y}!$$

Predictive Model:

$$P(\tilde{Y}=\tilde{y})=\int f(\tilde{y}|\theta)p(\theta)d\theta$$

A predictive distribution that integrates over unknown parameters, but is not conditional on observed data is called a **Prior Predictive Distribution**. Such distribution can be useful in evaluating if a prior distribution for θ actually translates into reasonable prior beliefs for observable data $\tilde{Y}$. After we have observed a sample $Y_1, \ldots, Y_n$ from the population, the relevant predictive distribution becomes:

$$P(\tilde{Y}=\tilde{y}|Y_1=y_1,\ldots,Y_n=y_n)=\int f(\tilde{y}|\theta,y_1,\ldots,y_n)f(\theta|y_1,\ldots,y_n)d\theta$$

$$=\int f(\tilde{y}|\theta)f(\theta|y_1,\ldots,y_n)d\theta$$

This is called the **Posterior Predictive Distribution**, because it conditions on an observed sample dataset.

4.3 Posterior Predictive Models Continued

In the case of a Poisson model with a gamma prior distribution, the posterior predictive distribution is given by negative binomial($a+\Sigma y_i$, $b+n$). Often, the distribution $f(\tilde{y}|y_1,\dots,y_n)$ will become too nasty to compute directly, in these cases we can sample from the posterior predictive distribution indirectly using Monte Carlo approximation. Since we can see that $f(\tilde{y}|y_1,\dots,y_n) = \int f(\tilde{y}|\theta) f(\theta|y_1,\dots,y_n) d\theta$ $f(\tilde{y}|y_1,\dots,y_n)$ is the posterior expectation of $f(\tilde{y}|\theta)$. To obtain the posterior predictive probabilites that $\tilde{Y}$ is equal to some specific value $\tilde{y}$, we just preform the Monte Carlo procedure, Sample: $\theta^{(1)},\dots,\theta^{(S)} \sim$ i.i.d. $f(\theta|y_1,\dots,y_n)$, & then approximate $f(\tilde{y}|y_1,\dots,y_n)$ with $\sum_{s=1}^{S} f(\tilde{y}|\theta^{(s)})/S$. This strategy works well if $f(y|\theta)$ is discrete & the quantities we are interested in are easily computable from $f(y|\theta)$. However, if we want a set of samples of $\tilde{Y}$ from its posterior predictive distribution we need

sample $\theta^{(1)} \sim f(\theta|y_1,\dots,y_n)$, sample $\tilde{y}^{(1)} \sim f(\tilde{y}|\theta^{(1)})$
sample $\theta^{(2)} \sim f(\theta|y_1,\dots,y_n)$, sample $\tilde{y}^{(2)} \sim f(\tilde{y}|\theta^{(2)})$
⋮ ⋮ ⋮
sample $\theta^{(S)} \sim f(\theta|y_1,\dots,y_n)$, sample $\tilde{y}^{(S)} \sim f(\tilde{y}|\theta^{(S)})$

↳ This sequence constitutes S independent samples from the joint posterior distribution of $(\theta, \tilde{Y})$.
↳ $\{(\theta,\tilde{y})^{(1)},\dots,(\theta,\tilde{y})^{(S)}\}$ &
The sequence $\{\tilde{y}^{(1)},\dots,\tilde{y}^{(S)}\}$ constitutes S independent samples from the *marginal* posterior distribution of $\tilde{Y}$, which is the posterior predictive distribution.

Example: Poisson Model Monte Carlo

The predictive probability that a woman who is 40 without a college degree will have more children than a 40 year old woman with a degree is 0.48. To compute this number exactly we would need to solve the following double infinite sum:

$$P(\tilde{Y}_1 > \tilde{Y}_2 \mid \sum Y_{i1} = 217, \sum Y_{i2} = 66) =$$

$$\sum_{\tilde{y}_2=0}^{\infty} \sum_{\tilde{y}_1=\tilde{y}_2+1}^{\infty} \text{dnbinom}(\tilde{y}_1, 219, 112) \times \text{dnbinom}(\tilde{y}_2, 68, 45)$$

Alternatively, this sum can be approximated using Monte Carlo sampling. Since $\tilde{Y}_1$ & $\tilde{Y}_2$ are *a posteriori* independent, samples from their joint posterior distribution can be created by sampling each variable seperately. Posterior predictive samples from the conjugate Poisson model can be generated as follows:

sample $\theta^{(1)} \sim$ gamma$(a + \sum y_i, b+n)$, sample $\tilde{y}^{(1)} \sim$ poisson$(\theta^{(1)})$

sample $\theta^{(2)} \sim$ gamma$(a + \sum y_i, b+n)$, sample $\tilde{y}^{(2)} \sim$ poisson$(\theta^{(2)})$

$\vdots$ $\vdots$ $\vdots$ $\vdots$ $\vdots$

sample $\theta^{(s)} \sim$ gamma$(a + \sum y_i, b+n)$, sample $\tilde{y}^{(s)} \sim$ poisson$(\theta^{(s)})$

In practice, these samples can be generated with a few lines of R code.

4.4: Posterior Predictive Model Checking

The empirical distribution of the number of childen for our sample of 40-year-old women without a degree reveals that out of $n=111$ women, the number of women with exactly two children is 38, which is twice the number of women in the sample with one child. In contrast, the groups posterior predictive distribution suggests that the probability of sampling a woman with two children is slightly less probable than sampling a woman with one. This apparent contradiction has many possible explanations:

↳ Sampling variability: The empirical distribution of sampled data does not generally match the population distribution, & may differ wildly at small sample sizes.

↳ Inapropriate model: The population may in actuality have a sharp peak at two children, however, there is no parameter of a posson distribution that makes it effectively model this population.

Monte Carlo approximation allows us to asses these explanations of the discrepency numerically:
For every vector y of length $n=111$, let $t(y)$ be the ratio of the number of 2's in y to the number of 1's, so for our observed data y_{obs}, $t(y_{obs})=2$. Now, suppose we were to sample a different set of 111 women recording their number of children, obtaining a data vector $\tilde{Y}$ of length 111. What value do we expect to result from $t(\tilde{Y})$? We can use the Monte Carlo procedure to obtain samples from the posterior predictive distribution of $t(\tilde{Y})$

To obtain the aforementioned samples from the posterior predictive distribution of $t(\tilde{Y})$ we will implement the following procedure for each $s \in \{1, \ldots, S\}$:

1. sample $\theta^{(s)} \sim f(\theta | Y = y_{obs})$
2. sample $\tilde{Y}^{(s)} = (\tilde{y}_1^{(s)}, \ldots, \tilde{y}_n^{(s)}) \sim$ i.i.d. $f(y | \theta^{(s)})$
3. compute $t^{(s)} = t(\tilde{Y}^{(s)})$

In R that is,

```
a ← 2;  b ← 1
t.mc ← NULL
for(s in 1:10000){
   theta1 ← rgamma(1, a+sy1, b+n1)
   y1.mc ← rpois(n1, theta1)
   t.mc  ← c(t.mc, sum(y1.mc==2)/sum(y1.mc==1)
   }
```

In this Monte Carlo sampling scheme:

$\{\theta^{(1)}, \ldots, \theta^{(S)}\}$ are samples from the posterior distribution of θ

$\{\tilde{Y}^{(1)}, \ldots, \tilde{Y}^{(S)}\}$ are posterior predictive datasets of size n

$\{t^{(1)}, \ldots, t^{(S)}\}$ are samples from the posterior predictive distribution of $t(\tilde{Y})$

Out of 10,000 Monte Carlo datasets, only half of a percent had values of $t(y)$ that equaled or exceeded $t(y_{obs})$. This indicates that our Poisson model does not effectively model the population. In general, we should ensure that our predictive datasets $\tilde{Y}$ resemble the observed dataset in features of interest. If this condition does not hold it may be time for a more complex model. However, White 1982, Klesin 2006, & Bunke and Milhaud 1998 have shown that an incorrect model can provide correct inference under certain circumstances.

Chapter 5: The Normal Model

A random variable Y is said to be *Normally Distributed* with mean θ & variance $\sigma^2 > 0$ if its density is given by:

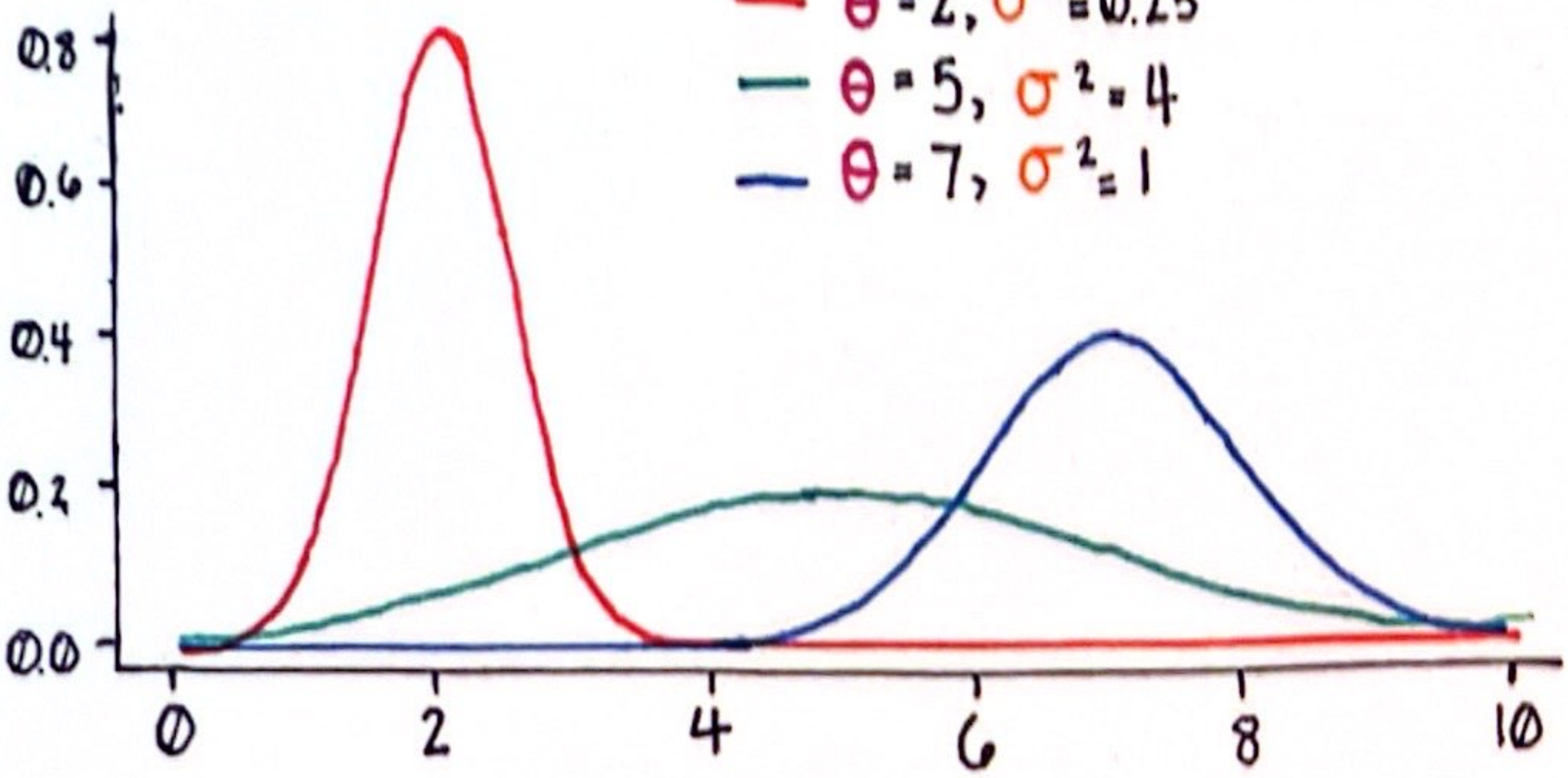
$$f(y|\theta,\sigma^2) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{1}{2}\left(\frac{y-\theta}{\sigma}\right)^2} \quad \text{for} \quad -\infty < y < \infty$$

Some key properties of the normal distribution include:

- The distribution is symetric about θ; the mean, median, & mode are all equal to θ
- About 95% of the data is within 2 standard deviations of the mean (1.96 to be exact)
- If $X \sim \text{norm}(\mu, \tau^2)$ & $Y \sim \text{norm}(\theta, \sigma^2)$, & X & Y are independent, then $aX + bY \sim \text{norm}(a\mu + b\theta, a^2\tau^2 + b^2\sigma^2)$

The importance of the normal distribution comes from the central limit theorem which states that the mean (or sum) of a set of random variables is approximately normally distributed; which, as it name implies, is *central* to the field of inferential statistics.

Note: The commands dnorm, rnorm, pnorm, & qnorm, *in R take the standard deviation* σ *rather than the variance* σ^2, *be mindful as to not accidently use the variance.*

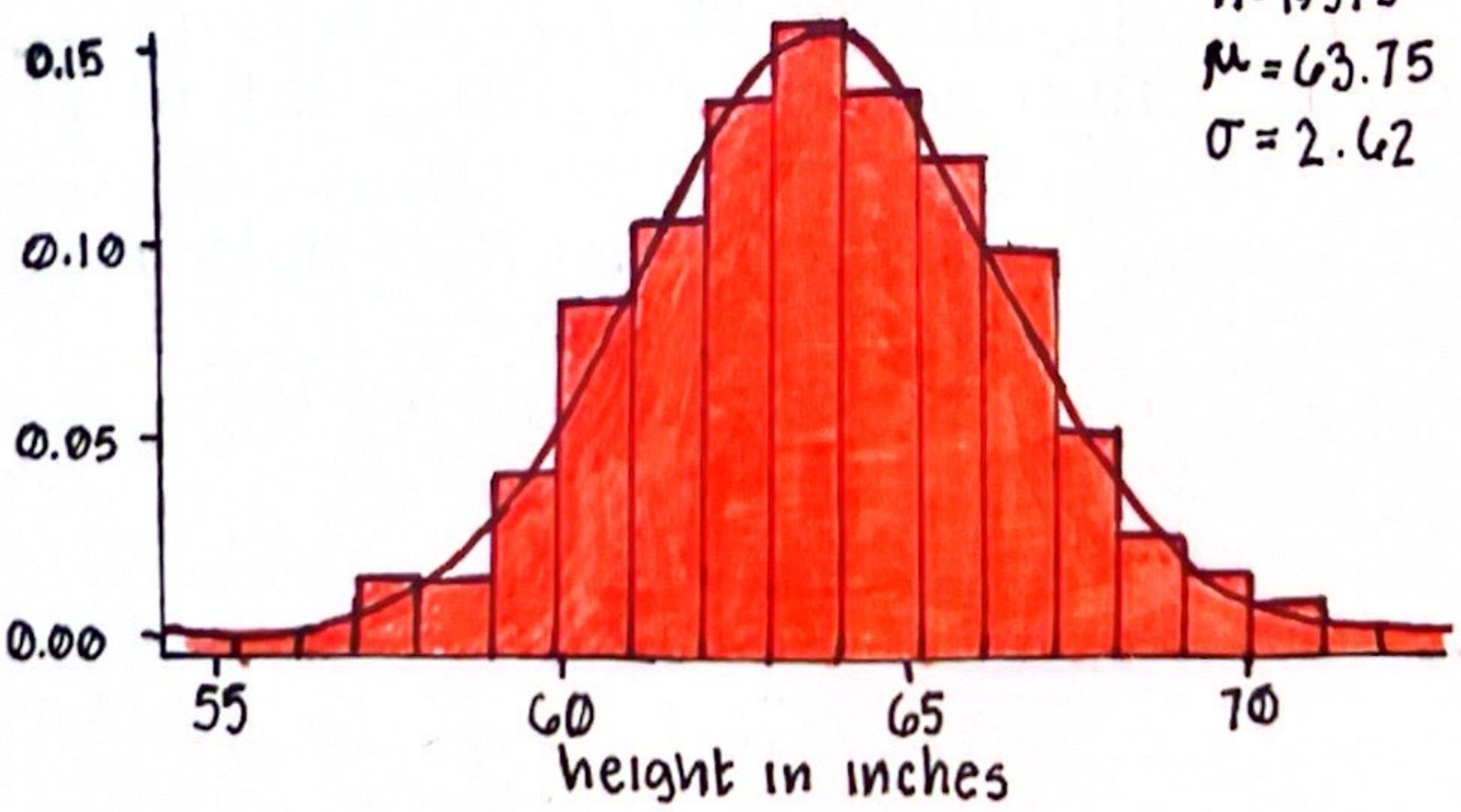


Normal Densities

Example: Normal Distibution (women's height)

A study of 1,100 English families from 1893 to 1898 gathered height data on $n = 1,375$ adult women. The sample had a mean of $\bar{y} = 63.75$ & sample standard deviation of $s = 2.62$. The sample data & the densitys curve of norm(63.75, 2.62)

The *Central Limit Theorem* says that if each observation in a sample can be expressed as a linear combination of a large number of unknown *additive* effect factors then the empirical distribution of such a sample will be approximately normally distributed. Which is a possible explanation for the extremely aparent normalitu of this sample.



Women's Height data & a Normal Densitu

5.2: Inference for the Mean, Conditional on the Var.

Suppose our model is $\{Y_1,\dots,Y_n \mid \theta,\sigma^2\} \sim$ i.i.d. norm(θ,σ^2)
The joint sampling density is then given by

$$f(y_1,\dots,y_n \mid \theta,\sigma^2) = \prod_{i=1}^{n} f(y_i \mid \theta,\sigma^2)$$

$$= \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{1}{2}\left(\frac{y_i-\theta}{\sigma}\right)^2} = (2\pi\sigma^2)^{-\frac{n}{2}} \exp\left\{-\frac{1}{2}\sum\left(\frac{y_i-\theta}{\sigma}\right)^2\right\}$$

Expanding the quadratic term in the exponential we see that $f(y_1,\dots,y_n \mid \theta,\sigma^2)$ depends on $y_1,\dots,y_n$ through

$$\sum_{i=1}^{n}\left(\frac{y_i-\theta}{\sigma}\right)^2 = \frac{1}{\sigma^2}\sum y_i^2 - 2\frac{\theta}{\sigma^2}\sum y_i + n\frac{\theta^2}{\sigma^2}$$

From this, it can be shown that: $\{\Sigma y_i^2, \Sigma y_i\}$ constitutes a two dimensional sufficient statistic; knowing these values is equivalent to knowing the values of $\bar{y} = \frac{\Sigma y_i}{n}$ & $s^2 = \frac{\Sigma(y_i-\bar{y})^2}{n-1}$ meaning $\{\bar{y}, s^2\}$ is also a sufficient statistic.

Inference in this two-parameter model can be broken down into two one parameter problems We turn our attention first to the case that σ^2 is known and we would like to estimate θ, using a conjugate prior distribution for θ. For any prior distribution $f(\theta \mid \sigma^2)$, we know the posterior distribution will satisfy:

$$f(\theta \mid y_1,\dots,y_n,\sigma^2) \propto f(\theta \mid \sigma^2) \times e^{\frac{-1}{2\sigma^2}\sum(y_i-\theta)^2}$$

$$\propto f(\theta \mid \sigma^2) \times e^{c_1(\theta-c_2)^2}$$

Since this is a *conjugate* prior distribution by definition it must produce a posterior distribution of the same class as itself; thus the resulting distribution contains a similar quadratic term to the one that appears in the density of a normal distribution. Densities with the form above are members of what's called the *Normal Family of Densities* which are the simplest class of probability densities on $\mathbb{R}$

Claim:
If $f(\theta|\sigma^2)$ is normal & $y_1 \dots y_n$ are i.i.d. $norm(\theta, \sigma^2)$ then $f(\theta|y_1,\dots,y_n,\sigma^2$ is also a normal densitus.

Proof: Let $\theta \sim norm(\mu_0, \tau_0^2)$, then

$$f(\theta|y_1,\dots,y_n,\sigma^2) = \frac{f(\theta|\sigma^2)f(y_1,\dots,y_n|\theta,\sigma^2)}{f(y_1,\dots,y_n|\sigma^2)} \propto f(\theta|\sigma^2)f(y_1,\dots,y_n|\theta,\sigma^2)$$

$$\propto \exp\left\{\frac{-1}{2\tau_0^2}(\theta-\mu_0)^2\right\}\exp\left\{\frac{-1}{2\sigma^2}\sum(y_i-\theta)^2\right\}$$

Adding the terms in the exponents & ignoring the $^{-1}/_2$ yields

$$\frac{1}{\tau_0^2}(\theta^2-2\theta\mu_0+\mu_0^2)+\frac{1}{\sigma^2}(\sum y_i^2-2\theta\sum y_i+n\theta^2)=a\theta^2-2b\theta+c$$

where $a=\frac{1}{\tau_0^2}+\frac{n}{\sigma^2}$, $b=\frac{\mu_0}{\tau_0^2}+\frac{\sum y_i}{\sigma^2}$, & $c=c(\mu_0,\tau_0^2,\sigma^2,y_1,\dots,y_n)$

plugging this back in:

$$f(\theta|\sigma^2,y_1,\dots,y_n) \propto \exp\left\{\tfrac{-1}{2}(a\theta^2-2b\theta)\right\}$$

$$= \exp\left\{\frac{-1}{2}a\left(\theta^2-2b\frac{\theta}{a}+\frac{b^2}{a^2}\right)+\frac{1}{2}\frac{b^2}{a}\right\}$$

$$\propto \exp\left\{\frac{-1}{2}a\left(\theta-\frac{b}{a}\right)^2\right\} = \exp\left\{\frac{-1}{2}\left(\frac{\theta-b/a}{1/\sqrt{a}}\right)^2\right\}$$

This function has the exact shape of a normal curve with $1/\sqrt{a}$ acting as the standard deviation & b/a acting as the mean. Thus, $f(\theta|\sigma^2;y_1,\dots,y_n)$ is normallus distributed with mean μ_n & variance τ_n^2 where:

$$\mu_n = \frac{b}{a} = \frac{\frac{1}{\tau_0^2}\mu_0+\frac{n}{\sigma^2}\bar{y}}{\frac{1}{\tau_0^2}+\frac{n}{\sigma^2}} \quad \& \quad \tau_n^2 = \frac{1}{\frac{1}{\tau_0^2}+\frac{n}{\sigma^2}}$$

Normal Model: Posterior Analysis

The (conditional) posterior parameters τ_n^2 & μ_n combine the prior parameters τ_0^2 & μ_0 with terms from the data.

The formula for $1/\tau_n^2$ is $\frac{1}{\tau_n^2} = \frac{1}{\tau_0^2} + \frac{n}{\sigma^2}$ This combination of the *inverse* of the prior variance & the *inverse* of the sample variance is called the **Precision**

For the normal model let $\tilde{\sigma} = \frac{1}{\sigma^2}$ = sampling precision i.e. how close the y_is to θ

It is often helpful to think about the precision as the quantity of information on an additive scale

$\tilde{\tau}_0^2 = \frac{1}{\tau_0^2}$ = prior precision

$\tilde{\tau}_n^2 = \frac{1}{\tau_n^2}$ = posterior precision

$\tilde{\tau}_n^2 = \tilde{\tau}_0^2 + n\tilde{\sigma}^2$

posterior information = prior information + data information

Posterior Mean:

$$\mu_n = \frac{\tilde{\tau}_0^2}{\tilde{\tau}_0^2 + n\tilde{\sigma}^2}\mu_0 + \frac{n\tilde{\sigma}^2}{\tilde{\tau}_0^2 + n\tilde{\sigma}^2}\bar{y}$$

This implies that the posterior mean is a weighted average of the prior mean & the sample mean. The weight on the sample mean is n/σ^2, the sampling precision of the sample mean. The weight on the prior mean is $1/\tau_0^2$, the prior precision. If the prior mean were based on κ_0 prior observations from the same (or similar) population as $Y_1, \ldots, Y_n$, then we might want to set $\tau_0^2 = \sigma^2/\kappa_0$, the variance of the mean of the prior observations in which case the formula for the posterior mean reduces to:

$$\mu_n = \frac{\kappa_0}{\kappa_0 + n}\mu_0 + \frac{n}{\kappa_0 + n}\bar{y}$$

Prediction Under the Normal Model

Consider predicting a new observation $\tilde{Y}$ from the population after having observed $(Y_1 = y_1, \ldots, Y_n = y_n)$. To find the predictive distribution we will use the following fact:

$$\{\tilde{Y} \mid \theta, \sigma^2\} \sim \text{norm}(\theta, \sigma^2) \iff \tilde{Y} = \theta + \tilde{\epsilon},\ \{\tilde{\epsilon} \mid \theta, \sigma^2\} \sim \text{norm}(0, \sigma^2)$$

In other words, saying that $\tilde{Y}$ is normally distributed with mean θ is the same as saying $\tilde{Y}$ is equal to θ plus a mean-zero normally distributed error term. Using this we can calculate the posterior mean & variance of $\tilde{Y}$

$$\begin{aligned} E[\tilde{Y} \mid y_1, \ldots, y_n, \sigma^2] &= E[\theta + \tilde{\epsilon} \mid y_1, \ldots, y_n, \sigma^2] \\ &= E[\theta \mid y_1, \ldots, y_n, \sigma^2] + E[\tilde{\epsilon} \mid y_1, \ldots, y_n, \sigma^2] \\ &= \mu_n + 0 = \mu_n \end{aligned}$$

$$\begin{aligned} V[\tilde{Y} \mid y_1, \ldots, y_n, \sigma^2] &= V[\theta + \tilde{\epsilon} \mid y_1, \ldots, y_n, \sigma^2] \\ &= V[\theta \mid y_1, \ldots, y_n, \sigma^2] + V[\tilde{\epsilon} \mid y_1, \ldots, y_n, \sigma^2] \\ &= \tau_n^2 + \sigma^2 \end{aligned}$$

We have previously shown that the sum of independent normal random variables is, itself, a normal random variable. Therefore, since both θ & $\tilde{\epsilon}$, conditional on $y_1, \ldots, y_n$ & σ^2, are normally distributed, so is $\tilde{Y} = \theta + \tilde{\epsilon}$. The predictive distribution is therefore:

$$\tilde{Y} \mid \sigma^2, y_1, \ldots, y_n \sim \text{norm}(\mu_n, \tau_n^2 + \sigma^2)$$

In general, our uncertainty about a new sample $\tilde{Y}$ is a function of our uncertainty about the center of the population (τ_n^2) as well as how variable the population is (σ^2). As $n \to \infty$ we become more certain about where θ is, and the posterior variance τ_n^2 of θ goes to zero. But, certainty about θ does not reduce the sampling variability σ^2, as such our uncertainty about $\tilde{Y}$ is never able to drop below σ^2.

Normal Model Prediction Example: Midge Wing Length

Grogan & Wirth (1981) provide data on the wing leng, in mm, of 9 members of midge flies. From these nine measurements we wish to infer the population mean θ. Studies from other species suggest that wing lengths are typically around 1.9mm, so we set $\mu_0 = 1.9$. We also know length must be positive which implies that $\theta > 0$. Therefore, ideally we would use a prior distribution for θ that only has mass on $\theta > 0$. We can approximate this restriction with a normal prior distribution for θ as follows: since for any normal distribution the vast majority of the data lies within two standard deviations of the mean, we chose τ_0^2 so that $\mu_0 - 2\tau_0 > 0$ or equivalently, $\tau_0 < 1.9/2 = 0.95$. For now, we take $\tau_0 = 0.95$, which somewhat overstates our prior uncertainty about θ. The observations in increasing order:

(1.64, 1.70, 1.72, 1.74, 1.82, 1.82, 1.82, 1.90, 2.08)

which yields $\bar{y} = 1.804$. Using the formula above for μ_n & τ_n^2 we have $\{\theta | y_1, \ldots, y_9, \sigma^2\} \sim \text{norm}(\mu_n, \tau_n^2)$, where

$$\mu_n = \frac{\frac{1}{\tau_0^2}\mu_0 + \frac{n}{\sigma^2}\bar{y}}{\frac{1}{\tau_0^2} + \frac{n}{\sigma^2}} = \frac{1.11 \times 1.9 + \frac{9}{\sigma^2}1.804}{1.11 + \frac{9}{\sigma^2}}$$

$$\tau_n^2 = \frac{1}{\frac{1}{\tau_0^2} + \frac{n}{\sigma^2}} = \frac{1}{1.11 + \frac{9}{\sigma^2}}$$

If $\sigma^2 = s^2 = 0.017$, then,
$\{\theta | y_1, \ldots, y_9, \sigma^2 = 0.017\} \sim \text{norm}(1.805, 0.002)$
A 95% quantile-based confidence interval for θ based on this distribution is (1.72, 1.89). However, this interval assumes we are certain that $\sigma^2 = s^2$, when in reality s^2 is only a rough estimate of σ^2 based on 9 observations. To get a more accurate representation of our information we need to account for the fact that σ^2 is not known.

5.3 Joint Inference for Mean & Variance

Bayesian inference for two or more unknown parameters is conceptually the same as the one-parameter case. For any joint prior distribution $f(\theta,\sigma^2)$ for θ & σ^2, posterior inference procedes with Baye's rule:

$$f(\theta,\sigma^2|y_1,\cdots,y_n) = \frac{f(y_1,\cdots,y_n|\theta,\sigma^2)\,f(\theta,\sigma^2)}{f(y_1,\cdots,y_n)}$$

As we did previously, we begin by developing a simple class of conjugate prior distributions which make posterior calculation easier. Recall from our second axiom of probability that a joint distribution for two quantities can be expressed as the product of a conditional probability & a marginal probability: $f(\theta,\sigma^2) = f(\theta|\sigma^2)f(\sigma^2)$. A few pages ago we saw that if σ^2 were known, then a conjugate prior distribution for θ was $\text{norm}(\mu_0,\tau_0^2)$. Let's consider a particular case where $\tau_0^2 = \sigma^2/\kappa_0$:

$$f(\theta,\sigma^2) = f(\theta|\sigma^2)f(\sigma^2) = \text{dnorm}(\theta,\mu_0,\tau_0=\sigma/\sqrt{\kappa_0}) \times f(\sigma^2)$$

in this case the parameters μ_0 & κ_0 can be interpreted as the "mean" & "sample size" from a set of prior observations. For σ^2 we need a family of prior distributions that has support on $(0,\infty)$. The gamma distribution is defined on the proper interval, but the gamma family is unfortunately not conjugate with the normal variance. However, all is not lost! The gamma family is conjugate with $1/\sigma^2$ (aka the precision). When utilizing such a prior we say that σ^2 has the *inverse-gamma* distribution

$$\text{precision} = \frac{1}{\sigma^2} \sim \text{gamma}(a,b) \;\&\; \text{variance} = \sigma^2 \sim \text{inverse-gamma}(a,b)$$

For interpretability we will reparameterize this prior distribution as $\frac{1}{\sigma^2} \sim \text{gamma}\left(\frac{\nu_0}{2},\frac{\nu_0}{2}\sigma_0^2\right)$

It then follows that if $\frac{1}{\sigma^2} \sim \text{gamma}(\frac{\nu_0}{2}, \frac{\nu_0}{2}\sigma_0^2)$ then,

$E[\sigma^2] = \sigma_0^2 \frac{\nu_0/2}{\nu_0/2-1}$, $\text{mode}[\sigma^2] = \sigma_0^2 \frac{\nu_0/2}{\nu_0/2+1} \rightarrow \text{mode}[\sigma^2] < \sigma_0^2 < E[\sigma^2]$

$V[\sigma^2]$ is decreasing in ν_0

Posterior Inference:

Suppose our prior distribution & sampling model are as follows: $\frac{1}{\sigma^2} \sim \text{gamma}(\frac{\nu_0}{2}, \frac{\nu_0\sigma_0^2}{2})$, $\theta|\sigma^2 \sim \text{norm}(\mu_0, \sigma^2/\kappa_0)$, & $Y_1,\dots,Y_n|\theta,\sigma^2 \sim \text{i.i.d. norm}(\theta,\sigma^2)$

Just as the prior distribution for θ & σ^2 can be decomposed as $f(\theta,\sigma^2) = f(\theta|\sigma^2)f(\sigma^2)$, the posterior distribution can be similarly decomposed:

$$f(\theta,\sigma^2|y_1,\dots,y_n) = f(\theta|\sigma^2,y_1,\dots,y_n)f(\sigma^2|y_1,\dots,y_n)$$

↑ The conditional distribution of θ given the data & σ^2 can be obtained by plugging σ^2/κ_0 for τ_0^2 yields $\{\theta|y_1,\dots,y_n,\sigma^2\} \sim \text{norm}(\mu_n, \sigma^2/\kappa_n)$, where $\kappa_n = \kappa_0 + n$ &

$$\mu_n = \frac{(\kappa_0/\sigma^2)\mu_0 + (n/\sigma^2)\bar{y}}{\kappa_0/\sigma^2 + n/\sigma^2} = \frac{\kappa_0\mu_0 + n\bar{y}}{\kappa_n}$$

Therefore, if μ_0 is the mean of κ_0 prior observations, then $E[\theta|y_1,\dots,y_n,\sigma^2]$ is the sample mean of the current & prior observations & $V[\theta|y_1,\dots,y_n,\sigma^2]$ is σ^2 divided by the total number of observations, both prior & current. The posterior distribution of σ^2 can be obtained by integrating over the unknown values of θ: $f(\sigma^2|y_1,\dots,y_n) \propto f(\sigma^2)f(y_1,\dots,y_n|\sigma^2)$

Solving this integral yields, ← $= f(\sigma^2)\int f(y_1,\dots,y_n|\theta,\sigma^2)f(\theta|\sigma^2)d\theta$

$\{1/\sigma^2|y_1,\dots,y_n\} \sim \text{gamma}(\frac{\nu_n}{2}, \frac{\nu_n\sigma_n^2}{2})$

where

$\nu_n = \nu_0 + n$

$$\sigma_n^2 = \frac{1}{\nu_n}[\nu_0\sigma_0^2 + (n-1)s^2 + \frac{\kappa_0 n}{\kappa_n}(\bar{y}-\mu_0)^2]$$

ν_0 can be interpreted as the "prior sample size" from which σ_0^2, the prior sample variance has been calculated

similarly $\nu_0\sigma_0^2$ & $\nu_n\sigma_n^2$ are prior & posterior sums of squares.

Joint Normal Model: Example

Returning to our midge data, studies suggest that the true mean & standard deviation of our population should not be too far from 1.9mm & 0.1mm respectively, suggesting that $\mu_0 = 1.9$ & $\sigma_0^2 = 0.01$. However, this population may be different from the others in terms of wing length, so we choose $k_0 = \nu_0 = 1$ so that our prior distributions are only weakly centered on the estimates from other populations. The sample mean & variance of our observed data are $\bar{y} = 1.804$ & $s^2 = 0.01609$ ($s = 0.130$). From these values we can compute μ_n & σ_n^2

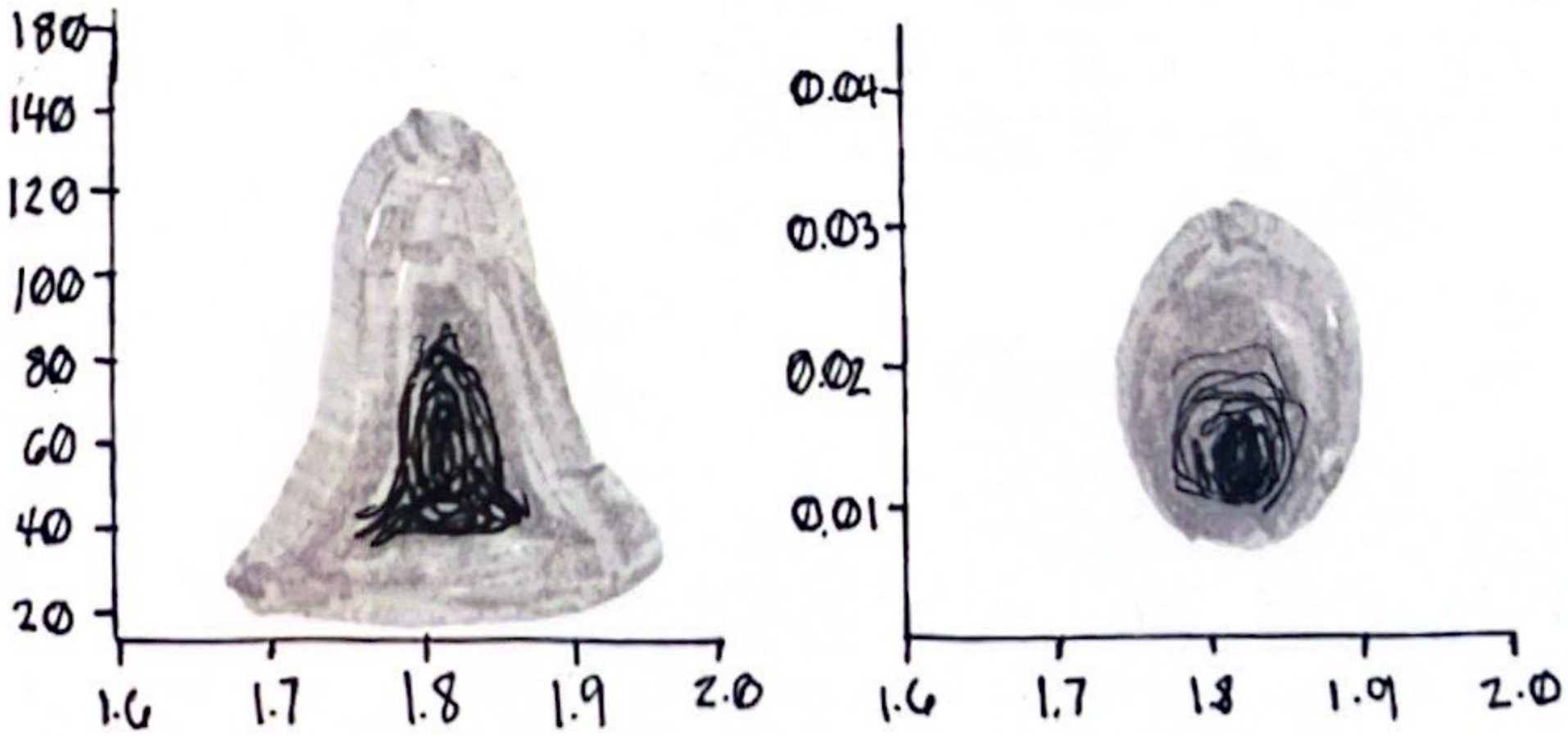
$$\mu_n = \frac{k_0\mu_0 + n\bar{y}}{k_n} = \frac{1.9 + 9 \times 1.804}{1+9} = 1.814$$

$$\sigma_n^2 = \frac{1}{\nu_n}\left[\nu_0\sigma_0^2 + (n-1)s^2 + \frac{k_0 n}{k_n}(\bar{y} - \mu_0)^2\right] = \frac{0.01 + 0.135 + 0.008}{10} = 0.015$$

Our joint posterior distribution is completely determined by the values $\mu_n = 1.814$, $k_n = 10$, $\sigma_n^2 = 0.015$, $\nu_n = 10$ which can be expressed as

$$\{\theta \mid y_1, \ldots, y_n, \sigma^2\} \sim \text{norm}(1.814, \sigma^2/10)$$

$$\{1/\sigma^2 \mid y_1, \ldots, y_n\} \sim \text{gamma}(10/2, 10 \cdot 0.015/2)$$



Joint Normal Model Monte Carlo Sampling

For many data analyses, interest primarily lies in estimating the mean of the population θ, & so we would like to calculate quantities like $E[\theta|y_1,\cdots,y_n]$, $sd[\theta|y_1,\cdots,y_n]$, $P(\theta_1<\theta_2|y_{11},\cdots,y_{n_2 2})$, etc... These quantities are all determined by the *marginal* posterior distribution of θ given the data. But all we know is the *conditional* distribution of θ given the data, and σ^2 is normal, & that σ^2 given the data is inverse-gamma distributed. If we could generate *marginal* samples of θ, from $f(\theta|y_1,\cdots,y_n)$, then we could use the Monte Carlo method to approximate the aforementioned quantities of interest. We can accomplish this by generating samples of θ & σ^2 from their joint posterior distribution. Consider simulating parameter values using the following Monte Carlo procedure:

$\sigma^{2(1)} \sim$ inverse-gamma$(\nu_n/2, \sigma_n^2\nu_n/2)$, $\theta^{(1)} \sim$ norm$(\mu_n, \sigma^{2(1)}/\kappa_n)$

$\vdots$ $\qquad\qquad$ $\vdots$

$\sigma^{2(S)} \sim$ inverse-gamma$(\nu_n/2, \sigma_n^2\nu_n/2)$, $\theta^{(S)} \sim$ norm$(\mu_n, \sigma^{2(S)}/\kappa_n)$

Note that each $\theta^{(s)}$ is sampled from its conditional distribution given the dat & $\sigma^2=\sigma^{2(s)}$

This Monte Carlo procedure can be implemented in two lines of R code

```
s2.postsample  <- 1/rgamma(10000, nun/2, s2n*nun/2)
theta.post sample <- rnorm(10000, mun, sqrt(s2.postsample/k
```

A sequence of pairs $\{(\sigma^{2(1)},\theta^{(1)}),\ldots,(\sigma^{2(s)},\theta^{(s)})\}$ simulated using this procedure are independent samples from the joint posterior distribution of $f(\theta,\sigma^2|y_1,\cdots,y_n)$

Improper Priors

What happens as our prior sample size goes to zero?
Recall we have previosly shown that

$$\mu_n = \frac{k_0\mu_0 + n\bar{y}}{k_0 + n}$$

$$\sigma_n^2 = \frac{1}{\nu_0+n}\left[\nu_0\sigma_0^2 + (n-1)s^2 + \frac{k_0 n}{k_0+n}(\bar{y}-\mu_0)^2\right]$$ Thus, as $k_0, \nu_0 \to 0$

$$\mu_n \to \bar{y}, \text{ \& } \sigma_n^2 \to \frac{n-1}{n}s^2 = \frac{1}{n}\sum(y_i-\bar{y})^2$$

This has lead some to suggest the following "posterior distribution"

$$\left\{\frac{1}{\sigma^2}\middle| y_1,\dots,y_n\right\} \sim \text{gamma}\left(\frac{n}{2}, \frac{n}{2}\frac{1}{n}\sum(y_i-\bar{y})^2\right)$$

$$\{\theta|\sigma^2, y_1,\dots,y_n\} \sim \text{norm}(\bar{y}, \sigma^2/n)$$

5.4 Bias, Variance, & Mean Squared Error

A *point estimator* of an unknown parameter θ is a function that converts our data into a single element of the parameter space Θ. For example, in the case of a normal sampling model & a conjugate prior distribution of the last section, the posterior mean estimator of θ is

$$\hat{\theta}_b(y_1,\dots,y_n) = E[\theta|y_1,\dots,y_n] = \frac{n}{\kappa_0+n}\bar{y} + \frac{\kappa_0}{\kappa_0+n}\mu_0 = w\bar{y} + (1-w)\mu_0$$

The *sampling properties* of an estimator such as $\hat{\theta}_b$ refer to its behavior under hypothetical repeatable surveys or experiments. Let us compare the sampling properties of $\hat{\theta}_b$ to $\hat{\theta}_e(y_1,\dots,y_n) = \bar{y}$, the sample mean, when the true value of the population mean is θ_0:

$E[\hat{\theta}_e|\theta=\theta_0] = \theta_0$, & we say that $\hat{\theta}_e$ is "*unbiased*"

$E[\hat{\theta}_b|\theta=\theta_0] = w\theta_0 + (1-w)\mu_0$, and if $\mu_0 \neq \theta_0$ we say that $\hat{\theta}_e$ is "*biased*"

Bias refers to how close the center of mass of the sampling distribution of an estimator is to the true value. An unbiased estimator is one with zero bias. However, bias doesn't really tell us how close the estimator's value is to the true value of the parameter we are estimating. To evaluate how close an estimator $\hat{\theta}$ is likely to be to the true value θ_0, we could use the *mean squared error* (MSE). Letting $m = E[\hat{\theta}|\theta_0]$, the MSE is

$$\begin{aligned} MSE[\hat{\theta}|\theta_0] &= E[(\hat{\theta}-\theta_0)^2|\theta_0] \\ &= E[(\hat{\theta}-m+m-\theta_0)^2|\theta_0] \\ &= E[(\hat{\theta}-m)^2|\theta_0] + 2E[(\hat{\theta}-m)(m-\theta_0)|\theta_0] + E[(m-\theta_0)^2|\theta_0] \end{aligned}$$

Since $m = E[\hat{\theta}|\theta_0]$ it follows that $E[(\hat{\theta}-m)|\theta_0] = 0$ & so the second term is zero. The first term is the variance of $\hat{\theta}$ & the third term is the square of the bias & so

$$MSE[\hat{\theta}|\theta_0] = V[\hat{\theta}|\theta_0] + Bias^2[\hat{\theta}|\theta_0]$$

Before any data is gathered, the expected distance from the estimator to the true value depends on how close θ_0 is to the center of the distribution of $\hat{\theta}$ (the bias), as well as how spread out the distribution is (the variance); allow me to restate the derivation of MSE, given that $m = E[\hat{\theta}|\theta_0]$

$$\begin{aligned} MSE(\hat{\theta}|\theta_0) &= E[(\hat{\theta}-\theta_0)^2|\theta_0] \\ &= E[(\hat{\theta}-m+m-\theta_0)^2|\theta_0] \\ &= E[(\hat{\theta}-m)^2|\theta_0] + 2E[(\hat{\theta}-m)(m-\theta_0)|\theta_0] + E[(m-\theta_0)^2|\theta_0] \\ &= V[\hat{\theta}|\theta_0] + Bias^2[\hat{\theta}|\theta_0] \end{aligned}$$

Returning to our comparison of $\hat{\theta}_e$ & $\hat{\theta}_b$, the bias of $\hat{\theta}_e$ may be zero, however

Prediction with the Binomial Model

I Neglected to mention an important feature of Bayesian Inference is the existance of a predictive distribution for *new* observations. Let $y_1, \ldots, y_n$ be the outcomes from a sample of n binary random variables, & let $\tilde{Y} \in \{0,1\}$ be an additional outcome from the same population, *that has yet to be observed*. Then, the **Predictive distribution** of $\tilde{Y}$ is the conditional distribution of $\tilde{Y}$ given that $\{Y_1 = y_1, \ldots, Y_n = y_n\}$. For conditionally i.i.d. binary variables this distribution can be derived from the distribution of $\tilde{Y}$ given θ & the posterior distribution of θ

$$P(\tilde{Y}=1 \mid y_1,\ldots,y_n) = \int P(\tilde{Y}=1, \theta \mid y_1,\ldots,y_n)\, d\theta$$

$$= \int P(\tilde{Y}=1 \mid \theta, y_1,\ldots,y_n)\, f(\theta \mid y_1,\ldots,y_n)\, d\theta$$

$$= \int \theta\, f(\theta \mid y_1,\ldots,y_n)\, d\theta$$

$$= E[\theta \mid y_1,\ldots,y_n] = \frac{a + \sum y_i}{a+b+n}$$

Alternatively,

$$P(\tilde{Y}=0 \mid y_1,\ldots,y_n) = 1 - E[\theta \mid y_1,\ldots,y_n] = \frac{b + \sum (1-y_i)}{a+b+n}$$

The uniform prior distribution, or beta(1,1), can be thought of as equivalent to the information in a prior dataset consisting of a single "1" & a single "0". Under this prior distribution $E[\theta \mid Y=y] = \frac{2}{2+n}\frac{1}{2} + \frac{n}{2+n}\frac{y}{n}$, & $\text{mode}(\theta \mid Y=y) = \frac{y}{n}$ where $Y = \sum_{i=1}^{n} Y_i$

Example: Inference About Genetic Status

Human males have an X & a Y chromosome & Females have two X chromosomes. Hemophilia is a disease that exhibits X-chromosome-linked-recessive inheritance, meaning that a male who inherits the gene that causes the disease on the X chromosome is affected, whereas a female carrying it on only one chromosome is not affected. The disease is generally fatal for women who inherit two such genes which occurs infrequently.

Prior Distribution:

Consider a woman who's brother is affected, which implies that her mother carries the disease on one of her chromosomes. We are also told that her father is *not* affected; thus the woman herself has a fifty-fifty chance of carrying the gene. Thus, our unknown quantity of interest has only two values: the woman either carries the gene ($\theta=1$) or she does not ($\theta=0$) & we know that $P(\theta=1) = P(\theta=0) = 1/2$

Data Model & Liklihood

The data used to update the prior information consist of the affection status of the woman's sons. Suppose she has two sons, neither of which are affected. Let $y_i=1$ or 0 denote an affected or unaffected son respectively. The outcomes of the two sons are exchangable & conditionally independent given θ. This produces the following likelyhood function. $P(y_1=0, y_2=0|\theta=1) = (0.5)(0.5) = 0.25$

$P(y_1=0, y_2=0|\theta=0) = (1)(1) = 1$

These expressions follows from the fact that if the woman carries the gene then each son has a fifty-fifty chance of being affected. Whereas if she does not have the gene her sons also will not

Posterior Distribution:

Baye's rule can now be used to combine the information in the data with the prior probabilities; in particular, the posterior probability that the woman carries the gene that causes the disease. Using y to denote the joint data (y_1, y_2) yields

$$P(\theta=1|y) = \frac{P(y|\theta=1)P(\theta=1)}{P(y|\theta=1)P(\theta=1)+P(y|\theta=0)P(\theta=0)}$$

$$= \frac{(0.25)(0.5)}{(0.25)(0.5)+(1.0)(0.5)} = \frac{0.125}{0.625} = 0.20$$

Intuitively it is clear that if a woman has two unaffected sons, it is less probable that she carries the gene herself; Bayes rule provides a formal mechanism for determining the extent of the correction.

Adding More Data:

A key aspect of Bayesian analysis is the case in which sequential analyses can be performed. For example, suppose that the woman has a secret third son who is also unaffected. We can use the posterior distribution calculated above as our new prior which yields

$$P(\theta=1|y_1,y_2,y_3) = \frac{(0.5)(0.2)}{(0.5)(0.2)+(1)(0.8)} = 0.111$$

Whereas if the third son is affected the probability that she is a carrier again becomes one.

Historical Note: Bayes & Laplace

Jacob Bernoulli stated that if $y \sim Bin(n,\theta)$, then $P(|y/n - \theta| > \varepsilon | \theta) \to 0$ as $n \to \infty$ for any θ & fixed value of $\varepsilon > 0$. The Reverend Thosmas Bayes & Pierre Simon Laplace both are independently credited with being the first to *invert* the probability statement & speak in terms of θ given the observed value(s) y.

In his famous 1763 paper, Bayes sought the probability $P(\theta \in (\theta_1, \theta_2) | y)$; his solution is based on a geometric analogy of a probability space to a rectangular plane of finite size where:

1. *Prior Distribution*: "A ball W is randomly thrown. The horizontal position of the ball on the plane is θ, expressed as a fraction of the plane's width."
2. *Likelihood*: "A ball O is randomly thrown n times. The value of y is the number of times O lands to the right of W"

Thus, θ is assumed to have a uniform prior distribution on $[0,1]$. He then obtains:

$$P(\theta \in (\theta_1, \theta_2) | y) = \frac{P(\theta \in (\theta_1, \theta_2), y)}{P(y)} = \frac{\int_{\theta_1}^{\theta_2} P(y|\theta) P(\theta) d\theta}{P(y)}$$

$$= \frac{\int_{\theta_1}^{\theta_2} \binom{n}{y} \theta^y (1-\theta)^{n-y} d\theta}{P(y)}$$

Bayes succeeded in showing that the denominator is equal to $P(y) = \int_0^1 \binom{n}{y} \theta^y (1-\theta)^{n-y} d\theta$; this shows that all possible values of y are equally likely *a priori*

$$= \frac{1}{n+1} \quad \text{for } y = 0, 1, 2, \cdots, n$$

Laplace wrote more rigorously & discovered beta integration & binomial normal approximation

Example: Estimating the Probability of a Female Birth

Let our parameter of interest θ be the proportion of female births. Let y be the number of girls recorded out of n births. By applying the binomial model we are assuming that the n births are independent given θ. To begin, we must specify a prior distribution for θ; we can begin by assuming that θ is uniformly distributed on $[0,1]$. By applying Baye's rule we can easily obtain a posterior distribution of

$$p(\theta|y) \propto \theta^y (1-\theta)^{n-y}$$

with fixed values of n & y the binomial coefficient $\binom{n}{y}$ does not vary with θ & as such can merely be treated as a constant in our posterior distribution. In this case, $\theta|y \sim \text{Beta}(y+1, n-y+1)$. Furthermore, if we let $\tilde{y}$ denote the outcome of a new trial the predictive distribution for $\tilde{y}$ is given by:

$$P(\tilde{y}=1|y) = \int_0^1 P(\tilde{y}=1|\theta,y)\,P(\theta|y)\,d\theta$$

$$= \int_0^1 \theta\, P(\theta|y)\,d\theta = E[\theta|y] = \frac{y+1}{n+2}$$

Derivation of the beta posterior from Binomial & Uniform Prior

If we have $y|\theta \sim \text{binomial}(n, \theta)$ & $\theta \sim \text{uniform}(0, 1)$

By Baye's rule we have: $P(\theta|y) = \frac{P(y|\theta)P(\theta)}{P(y)}$

$$P(y) = \int P(y, \theta) d\theta = \int \binom{n}{y} \theta^y (1-\theta)^{n-y} d\theta = \binom{n}{y} \int \theta^{(y+1)-1} (1-\theta)^{n-y+1-1} d\theta$$

$$= \binom{n}{y} B(y+1, n-y+1) = \binom{n}{y} \frac{\Gamma(y+1)\Gamma(n-y+1)}{\Gamma(n+2)} = P(y)$$

$$P(\theta|y) = \frac{P(y|\theta)P(\theta)}{P(y)} = \frac{\binom{n}{y}\theta^y(1-\theta)^{n-y}}{\binom{n}{y}\frac{\Gamma(y+1)\Gamma(n-y+1)}{\Gamma(n+2)}}$$

$$= \frac{\Gamma(n+2)}{\Gamma(y+1)\Gamma(n-y+1)} \theta^{(y+1)-1}(1-\theta)^{(n-y)+1}$$

$$= \text{beta}(y+1, n-y+1)$$

Derivation of Posterior Predictive Distribution Binomial Model with a Uniform Prior

Suppose that $y|\theta \sim \text{Binomial}(n,\theta)$ & $\theta \sim \text{Beta}(1,1)$
To predict the probability of a single future observation being a success we will find $P(\tilde{y}=1|y)$

$$P(\tilde{y}=1|y) = \int_0^1 P(\tilde{y}=1,\theta|y)\,d\theta = \int_0^1 P(\tilde{y}=1|\theta,y)\,P(\theta|y)\,d\theta$$

$$= \int_0^1 P(\tilde{y}=1|\theta)\,P(\theta|y)\,d\theta = \int_0^1 \text{Binomial}(1,\theta)\,\text{Beta}(y+\alpha, n-y+\beta)\,d\theta$$

(Likelihood: $P(\tilde{y}=1|\theta)$; Prior: $P(\theta|y)$)

$$= \int_0^1 \cancel{\binom{1}{1}}\,\theta^1\,\cancel{(1-\theta)^{1-1}}\,\frac{\Gamma(n+\alpha+\beta)}{\Gamma(y+\alpha)\Gamma(n-y+\beta)}\,\theta^{y+\alpha-1}(1-\theta)^{n-y+\beta-1}\,d\theta$$

$$= \frac{\Gamma(n+\alpha+\beta)}{\Gamma(y+\alpha)\Gamma(n-y+\beta)} \int_0^1 \theta^{y+\alpha+1-1}(1-\theta)^{n-y+\beta-1}\,d\theta$$

$$= \frac{\Gamma(n+\alpha+\beta)}{\Gamma(y+\alpha)\Gamma(n-y+\beta)}\,\frac{\Gamma(y+\alpha+1)\Gamma(n-y+\beta)}{\Gamma(n+\alpha+\beta+1)} \int_0^1 \frac{\Gamma(n+\alpha+\beta+1)}{\Gamma(y+\alpha+1)\Gamma(n-y+\beta)}\,\theta^{y+\alpha}(1-\theta)^{n+y+\beta-1}\,d\theta$$

$$= \frac{\Gamma(n+\alpha+\beta)}{\Gamma(y+\alpha)\cancel{\Gamma(n-y+\beta)}}\,\frac{\Gamma(y+\alpha+1)\cancel{\Gamma(n-y+\beta)}}{\Gamma(n+\alpha+\beta+1)}$$

↳ This is a beta distribution & thus it must integrate to one.

Recall: $\Gamma(\alpha) = (\alpha-1)!$

$$= \frac{(n+\alpha+\beta-1)!\;(y+\alpha)!}{(y+\alpha-1)!\;(n+\alpha+\beta)!}$$

$$= \frac{\cancel{(n+\alpha+\beta-1)!}\;(y+\alpha)\cancel{(y+\alpha-1)!}}{\cancel{(y+\alpha-1)!}\;(n+\alpha+\beta)\cancel{(n+\alpha+\beta-1)!}}$$

$$= \frac{y+\alpha}{n+\alpha+\beta} = \frac{\text{total successes}}{\text{total sample size}} = E[\theta|y]$$

(Single Observation)

Posterior Derivation: Normal w/ known Variance

That is, let $P(y_i|\theta) \sim \mathcal{N}(\theta, \sigma^2)$ & $P(\theta) \sim \mathcal{N}(\theta_0, \tau^2)$

$$P(\theta|y) \propto P(y|\theta)\,P(\theta) \rightarrow \mathcal{N}(\theta, \sigma^2)\,\mathcal{N}(\theta_0, \tau^2)$$

$$\propto \exp\left\{\frac{-(y-\theta)^2}{2\sigma^2}\right\} \times \exp\left\{\frac{-(\theta-\theta_0)^2}{2\tau^2}\right\}$$

$$\propto \exp\left\{\frac{-(y-\theta)^2}{2\sigma^2} - \frac{-(\theta-\theta_0)^2}{2\tau^2}\right\}$$

$$\propto \exp\left\{\frac{-(y-\theta)^2\tau^2 - (\theta-\theta_0)^2\sigma^2}{2\sigma^2\tau^2}\right\} \rightarrow \text{Complete the squares}$$

$$\propto \exp\left\{\frac{-(y^2\tau^2 - 2\theta y\tau^2 + \theta^2\tau^2 + \theta^2\sigma^2 - 2\theta\theta_0\sigma^2 + \theta_0^2\sigma^2)}{2\sigma^2\tau^2}\right\}$$

$$\propto \exp\left\{\frac{-(\theta^2(\tau^2+\sigma^2) - 2\theta(y\tau^2+\theta_0\sigma^2) + y^2+\tau^2+\theta_0^2\sigma^2)}{2\sigma^2\tau^2}\right\}$$

$$\propto \exp\left\{\frac{-(\tau^2+\sigma^2)}{2\sigma^2\tau^2}\left(\theta^2 - 2\theta\left(\frac{y\tau^2+\theta_0\sigma^2}{\tau^2+\sigma^2}\right) + \frac{y^2\tau^2+\theta_0^2\sigma^2}{\tau^2+\sigma^2}\right)\right\}$$

Since we are using proportionality, this can be treated as a constant

$$\propto \exp\left\{\frac{-(\tau^2+\sigma^2)}{2\sigma^2\tau^2}\left(\theta - \left(\frac{y\tau^2+\theta_0\sigma^2}{\tau^2+\sigma^2}\right)\right)^2\right\}$$

$$\propto \exp\left\{\frac{-1}{2\frac{\sigma^2\tau^2}{(\sigma^2+\tau^2)}}\left(\theta - \left(\frac{y\tau^2+\theta_0\sigma^2}{\tau^2+\sigma^2}\right)\right)^2\right\}$$

$$\hookrightarrow \mathcal{N}\left(\frac{\tau^2 y + \sigma^2\theta_0}{\tau^2+\sigma^2},\ \frac{\sigma^2\tau^2}{\tau^2+\sigma^2}\right)$$

Fisher Information & Jeffrey's Prior

Consider the likelihood $P(y|\theta)$; its **Fisher Information** is given by: $I(\theta) = -\mathbb{E}^{Y|\theta}\left(\frac{\partial^2 \ln(P(y|\theta))}{\partial\theta}\right)$ which measures the sensitivity of an estimator in the neighborhood(?) on maximum likelihood. It is proportional to the expected curvature of the log-likelihood function at the maximum likelihood estimator: a higher Fisher Information indicates greater sensitivity. Jeffreys uses this proportionality to suggest the following as an invariant uniformative prior:

$$\pi(\phi) = \pi(g(\phi))\left|\frac{dg(\phi)}{d\phi}\right| = \pi(\theta)\left|\frac{d\theta}{d\phi}\right|$$

But...

$$I(\phi) = -\mathbb{E}^{Y|\phi}\left(\frac{\partial^2 \ln(P(y|\phi))}{\partial\phi^2}\right) = -\mathbb{E}^{Y|\theta}\left(\frac{\partial^2 \ln(P(y|\theta))}{\partial\theta^2} \cdot \left|\frac{d\theta}{d\phi}\right|^2\right)$$

$$= I(\theta)\left|\frac{d\theta}{d\phi}\right|^2 \quad \text{Thus, } \sqrt{I(\phi)} = \sqrt{I(\theta)}\left|\frac{d\theta}{d\phi}\right|$$

Jeffrey's Prior, then, is defined by $\pi(\theta)$

Example:

Let $Y_1, \ldots, Y_n$ be random sampled observations from a normal distribution with known variance σ^2:

Likelihood: $p(\bar{y}|\theta) \propto \exp\left\{\frac{-n(\bar{y}-\theta)^2}{2\sigma^2}\right\}$

$$\frac{\partial^2 \ln(P(\bar{y}|\theta))}{\partial\theta^2} = \frac{\partial^2}{\partial\theta^2}\left(\frac{-n(\bar{y}-\theta)^2}{2\sigma^2}\right) = \frac{\partial}{\partial\theta}\left(\frac{2n(\bar{y}-\theta)}{2\sigma^2}\right) = \frac{-n}{\sigma^2}$$

which means that making this $\pi(\theta) \propto \sqrt{\frac{n}{\sigma^2}} \propto 1$

Jeffereu's Prior for the normal likelihood with known variance

Single-Parameter Models Table

Likelihood	Conjugate	Posterior	Predictive
Binomial (n, θ)	Beta (α, β)	Beta $(y+\alpha, n-y+\beta)$	Beta Binomial
Poisson (θ)	Gamma (α, β)	Gamma $(\alpha+n\bar{y}, \beta+n)$	Negative Binomial $(\alpha+n\bar{y}, \beta+n)$
Normal (known σ^2) (θ, σ^2)	Normal (θ_0, τ^2)	Normal $\left(\frac{\frac{n\bar{y}}{\sigma^2}+\frac{\theta_0}{\tau^2}}{\frac{n}{\sigma^2}+\frac{1}{\tau^2}}, \frac{1}{\frac{n}{\sigma^2}+\frac{1}{\tau^2}}\right)$	Normal
Normal (known μ) (μ, σ^2)	Inverse Gamma (α, β)	Inverse Gamma $\left(\frac{n+2\alpha}{2}, \frac{n\nu+2\beta}{2}\right)$	$\int p(\tilde{y}\mid\theta)p(\theta\mid y)d\theta$

Multivariate Normal Model: known variance

Multivariate Normal Likelihood:

The basic model to be discussed concerns an observable vector y of d components; with the multivariate normal distribution. $y|\mu,\Sigma \sim \mathcal{N}(\mu,\Sigma)$ where μ is a (column) vector of length d & Σ is a $d \times d$ variance matrix which is positive definite & symetric. The likelihood function for a single observation is:

$$P(y|\mu,\Sigma) \propto |\Sigma|^{-1/2} \exp\left\{\frac{-1}{2}(y-\mu)^T\Sigma^{-1}(y-\mu)\right\}$$

Conjugate Analysis: As with in the univariate case we first consider the case where Σ is known. In such a case the log-likelihood is a quadritic function of μ. Therefore, the conjugate prior distribution of μ is the multivariate normal model which we will denote as $\mu \sim \mathcal{N}(\mu_0, \Lambda_0)$. The resultina posterior is then

$$P(\mu|y,\Sigma) \propto \exp\left\{\frac{-1}{2}\left((\mu-\mu_0)^T\Lambda_0^{-1}(\mu-\mu_0) + \sum_{i=1}^{n}(y_i-\mu)^T\Sigma^{-1}(y_i-\mu)\right)\right\}$$

which is an exponental of this quadratic form. Completing the square & pulling out the constant factors yields the following simplified form.

$$P(\mu|y,\Sigma) \propto \exp\left\{\frac{-1}{2}(\mu-\mu_n)^T\Lambda_n(\mu-\mu_n)\right\} = \mathcal{N}(\mu|\mu_n,\Lambda_n)$$

Where: $\mu_n = (\Lambda_0^{-1}+n\Sigma^{-1})^{-1}(\Lambda_0^{-1}\mu_0 + n\Sigma^{-1}\bar{y})$ &
$\Lambda_n = \Lambda_0^{-1} + n\Sigma^{-1}$

As we have seen in similar cases these values are weighted aveages of the prior information & the information contained in the sample data. (the variance is the sum of the variances)

Multiparameter Normal Example

In 1882 Simon Newcomb conducted an experiment to measure the speed of light by timing how long it takes light to travel 7442 meters; he collected 66 observations in total.

Likelihood: we will (inapropriately) assume these observations are exchangable & normally distributed, giving us the following likelihood:

$$y_i | \mu, \sigma^2 \sim \mathcal{N}(\mu, \sigma^2)$$

Prior: For the sake of simplicity, we will use an uninformative prior which we know from the single parameter case must be an inverse gamma/inverse χ^2 distribution that is proportional to the following:

$$P(\mu, \sigma^2) \propto \frac{1}{\sigma^2} \Longrightarrow \text{Inv-Gamma}(0, 0)$$

Posterior: According to Newcomb the observed sample mean was $\bar{y} = 26.2$ & the sample standard deviation was $s = 10.8$. With this in mind we first need to find the marginal posterior of σ^2 which is given by the following:

$$\sigma^2 | y \sim \text{Inv-}\chi^2(n-1, s^2) = (66-1, 10.8^2) = (65, 116.64)$$

Using this marginal posterior we can calculate the posterior distribution of μ:

$$\mu | \sigma^2, y \sim \mathcal{N}(\bar{y}, \sigma^2/n) = (26.2, \frac{\sigma^2}{66})$$ where σ^2 is a sampled value from the marginal posterior distribution This can be expressed analytically as follows

$$\mu | y \sim t_{n-1}(\bar{y}, \frac{s}{\sqrt{n}}) = t_{65}(20.6, \frac{10.8}{\sqrt{66}})$$

Integrating out Nuissanse Parameters

We can skip right to solving for the parameters of interest's posterior distribution given some data by integrating over all values of the necessary nuisance parameters. The resulting distribution is equivalent to the distribution that results from successively solving the conditional distributions.

Pros:
- less steps
- produces a single analytic form of the posterior
- don't need to handle nuissance parameters directly

Cons:
- Can produce some pretty dark sided distributions
- **ANALYTIC INTEGRALS!**

In the case of our multiparameter normal model we can integrate $P(\mu, \sigma^2|y)$ across all possible values of σ^2 & the result is $P(\mu|y)$

$$\int P(\mu, \sigma^2|y)\,d\sigma^2 \propto \int (\sigma^2)^{-(\frac{n}{2}+1)} \exp\left\{\frac{-1}{2\sigma^2}\left((n-1)s^2 + n(\bar{y}-\mu)^2\right)\right\} d\sigma^2$$

This is the kernel of an Inv-Gamma $\left(\frac{n}{2}, \frac{(n-1)s^2+n(\bar{y}-\mu)^2}{2}\right)$ distribution meaning that when multiplied by a normalizing constant it must equal one meaning we can rewrite this integral as the reciprocal of its normalizing constant.

$$\frac{\Gamma(\frac{n}{2})}{\left(\frac{(n-1)s^2+n(\bar{y}-\mu)^2}{2}\right)^{n/2}} \propto \Gamma\left(\frac{n}{2}\right)\left((n-1)s^2\right)^{\frac{-n}{2}}\left(1+\frac{1}{n-1}\left(\frac{(\bar{y}-\mu)^2}{s^2/n}\right)\right)^{\frac{-n}{2}}$$

↑ This is the PDF of a t_{n-1} distribution! which makes sense after all this is plain normal sample data.