Gender Beyond the Binary:

Computationally Mapping Gender to a

Spectrum Using Sex Differences in the Brain

by

Reed Williams

In Partial Fulfillment of the Requirements for the Degree of

MASTER OF SCIENCE

in

The Department of Computer Science

State University of New York
New Paltz, New York 12561

May 2022

GENDER BEYOND THE BINARY: COMPUTATIONALLY MAPPING GENDER TO A SPECTRUM USING SEX DIFFERENCES IN THE BRAIN

Reed Williams

State University of New York at New Paltz

We, the thesis committee for the above candidate for the
Master of Science degree, hereby recommend
acceptance of this thesis.

Anca Ràdulescu, Thesis Advisor
Department of Mathematics, SUNY New Paltz

Chirakkal Easwaran, Thesis Committee Member
Department of Computer Science, SUNY New Paltz

Ashley Suchy, Thesis Committee Member
Department of Computer Science, SUNY New Paltz

Submitted in partial fulfillment
of the requirements for the Master of Science degree
in Computer Science
at the State University of New York at New Paltz

Abstract

Biological sex is far more complex than simply two categories: male and female. The mere existence of transgender and intersex individuals displays this complexity clearly on the surface, while the differences between cisgender people within their own respective categories brings this idea to a deeper level. While sex differences reveal themselves in many different scientific disciplines, this study will focus on findings in the field of neuroscience; specifically, it will narrow in on volumetric measurements of brain regions known to have differing trends across the male and female sexes. The construction of a surrogate data set driven by measurements extracted from existing literature will be used to fit a logistic regression model. The resulting probability function will be used to first create a base *Biological Sex Spectrum*; this refers to a representation of biological sex as a spectrum in the absence of societal influence. This probability function will then be modified to produce a *Societally Influenced Gender Spectrum*; this refers to a spectrum that has been influenced by the concept of the *gender binary* and more closely represents our current world. The comparison of these two spectra will reveal the space for an increase in gender diversity as societal views continue shifting further away from restricting gender stereotypes.

Contents

List of Tables

List of Figures

Chapter 1

Introduction

Gender exists in our society in an overwhelmingly binary way. Before a child is even born, one of the biggest questions people ask is "is it a boy or a girl?". The sex someone is assigned at birth often goes on to dictate their socialization process: how a child is dressed, what toys they are given to play with, how they react to social cues, what jobs they can imagine having when they grow up, etc. These concepts are rooted in the existence of the *gender binary*; a framework for the view that humans are made up of solely two types of beings: male and female [16].

The concept of the *gender binary* assumes that humans are biologically sexually dimorphic. A *dimorphism* is something that exists in solely two distinct forms. A *sexual* dimorhpism deepens this definition to mean that these two distinct forms are separated on the grounds of sex. And, for multiple things to be distinct, they must be completely different with absolutely no overlap between them. The gender binary uses external genitalia as the single deciding biological factor separating the two sexes; however, biological sex is far more complex than this single component. Additionally, *biological sex* and *gender* are different, although intertwined, concepts, revealing further underlying weaknesses in the gender binary.

What is biological sex? This study will use the term *biological sex* to refer to an individual's gender assignment at birth, also known as one's "assigned sex," when it is classifying

individuals represented in data. Conceptually, it will also take into account the web of other underlying biological factors that exemplify sex differences. An important definition to include here is that of the term *intersex*; this is a general term for an individual who is born with external genitalia that do not fit into either male or female categories.

What is gender? This study will use the term *gender* to refer to the societal constructs surrounding male, female, and non-binary individuals. Gender can refer to how individuals self-identity; it is important to note that how someone "identifies" is the same thing as "what someone is." Language is important, and this study does not wish to diminish the merit of how someone views themself; after all, no one can truly understand the lived experiences of someone better than the individual in question. A quick definition to be included here is *transgender*; it is an umbrella term for any individual who identifies as a gender other than what they were assigned to at birth. For example, someone who was assigned female at birth (afab), and is transgender, could identify as male, non-binary, or anything else that is not female. Someone who was assigned male at birth (amab), and is transgender, could identify as female, non-binary, or anything else that is not male. Cisgender is a term for an individual who does identify as their assigned sex. Other related definitions include *gender non-conforming*, a general term someone who does not align with the traditional expectations of their assigned sex, and finally agender, a term for people who identify as neither male, female, or anything in between.

The distinction between gender and biological sex is muddled, as the two can affect one another, and additional literature from the field of gender sciences can be seen to provide a more thorough articulation than what is given here. Only a broad understanding is needed in the scope of this work, so the brief explanations given above will be enough to maintain consistent and meaningful communication within the scope of this study. Throughout this study, note that "male" refers to an individual who was assigned male at birth(amab), and "female" refers to an individual we was assigned female at birth(afab).

Why is all this important? Societal influences play a large role in the formation of

gender identity. Our society's attachment to the gender binary can be detrimental to the transgender, gender non-conforming, and intersex communities. Oftentimes, younger gender non-conforming individuals are unaware of their gender non-conformity due to the lack of representation in media and exposure to individuals like themselves. Or, if they do realize theses things about themselves, their self-expression is often suppressed by people who invalidate non-traditional identities.

Gender identity relies on an important balance between biology and societal attitudes. Over recent years, as gender stereotypes become less harsh and gender diversity visibility increases, there has been an increase in the number of out transgender and non- binary individuals[21]. This study aims to show that the biology to back up this rise in gender diversity has always been present beneath the surface. Gender diverse individuals are not coming out of nowhere; what is changing is societies willingness to let people express themselves. Society is becoming more flexible towards people of varied gender expressions. As a result, there is more representation in the world for younger people to realize that "male" and "female" are not strictly the only options. This is creating a new wave of possibilities and opportunities for people to express themselves. As we begin to learn more about what makes us "male" or "female," we begin to learn how much actually lies in between.

As more academic research emerges, alongside the increase in representation of gender diverse individuals in both everyday life and the media we consume, the concept of the gender binary is being increasingly challenged. This study will contribute to this challenge.

We will first conduct a literature review to explore various biological sex differences. After providing a broad overview of sex differences across a handful of disciplines, we will narrow our scope to focus on the brain. Neuroscience research shows that the brain is not a sexually dimorphic structure; instead, it is a patchwork of features drawn from distributions with male-typical and female-typical trends, but overall, little *internal consistency* (a term that will be defined later). The direction of our literature review will shift towards the exploration of key brain features that have been found to demonstrate sex differences. We

will focus on various regional volumetric measurements of these brain regions from existing literature. Finally, research conducted on different locations and at various scopes regarding the proportion of populations made up by transgender and non-binary individuals will be examined, and key findings will be recorded.

All the data collected about brain features was used to create a surrogate data set where each data point represents a vector of key regions from a brain scan of a hypothetical individual. A surrogate data set is a computationally created data set based on assumptions from existing literature. Both a training and testing data set are generated. Special data points are also handpicked to act as reference points later in the analysis section of the study. Half the data points will represent individuals assigned male at birth, and the remaining data points will represent individuals assigned female at birth.

This training data set is fit to a logistic regression model, where our focus will not be on the classification of points being one or zero, but with the probability function working behind the scenes to dictate these predicted classifications. From here, we have two main goals: (1)Create a Biological Sex Spectrum, which we will also call the Base Logistic Regression, and (2)Create a Societally Influenced Gender Spectrum, which we will also call the skewed logistic regression.

First, we discuss the base logistic regression. Once we have generated the desired probability function, the previously generated testing dataset is inputted; the outputted probability values are recorded, sorted, scatter plotted, and fit to a sigmoid curve. Our focus lies primarily with the scale parameter, s, as it describes the steepness of the curve, which we will use as the base value for our metric of societal influence. We also note the difference between the smallest probability of a female data point and the largest probability of a male data point, as this range represents individuals who exist in the overlap between the two gender extremes.

Next, we discuss the skewed logistic regression. We modify the outputs of the probability function, shifting each value proportionately towards zero for male data points, and similarly

towards one for female data points to reflect the societal pressure to conform to the gender stereotypes that align with an individual's sex assignment at birth. This new, altered set of probability values is then sorted, scatter plotted, and fit to a sigmoid curve. We vary the extent to which we shift these probabilities to produce a curve whose percent of overlap, (recall this is the range of individuals between the exclusive gender extremes,) matches the percentages of transgender and non-binary folks that have been previously extracted from literature. We use the s values computed when we fit the scatter plot to a sigmoid curve to compare the difference between the base logistic regression model results and the results from the different degrees of skewed logistic regression models; this comparison reveals the range of gender identities that are being stifled by gender stereotypes on our society.

Chapter 2

Literature Review

2.1 General Biological Sex Differences

Scientific fields such as biology, endocrinology, physiology, genetics, neuroscience, and reproductive science all have shown that sex exists as a spectrum. To list some, biological markers of sex include chromosomes, gonads, hormones, secondary sex characteristics, external genitalia, internal genitalia, skeletal structure, gene expression, brain structure, and hormone receptor sensitivity. The following non-binary biological markers of sex (meaning there are more than two, consistent, options) are worth highlighting: external genitalia has more than two options, including full-sized penis, small penis, micro-penis, clitoromegaly, enlarged clitoris, and standard sized clitoris; sex chromosome variations include Turner Syndrome, XY mosaicism, XX/XY, Trisomy X, Klinefelter syndrome, XXY with normal phenotype, XXXX, XXXY, XXYY, XXXXY, XXXXX, XX Male Syndrome, XX Gonadal Dysgenesis, and XY Gonadal Dysgenesis; skeletal systems exhibit variation because all men are not taller than all women; secondary sex characteristics are not binary because some women do have dense body hair and some men are unable to grow a full beard [15].

Next we examine a view from the field of behavioral neuroendocrinology, which deals with the bidirectional interactions between hormones and behavior. The gender binary would be

supported by behavioral neuroendocrinology if gonadal hormones were sexually dimorphic and if the levels of these sexually dimorphic hormones were genetically determined and fixed. Some of these proposed hormones include estrogen, progesterone, and testosterone. [16] show that androgens and estrogens are not two sexually dimorphic sets of hormones; they are dynamic hormones, found in everyone, which are subject to influence via environmental and social factors.

Hyde et all review five different ways in which the gender binary is biologically undermined. They first examine neuroscience findings dealing with sex differences in the brain. Their findings tell us that while there are trends that divide males and females on average, the distributions of these brain features for both sexes overlap, and internal consistency across all the features is rare. This means that an individual classified as male may have *many* brain feature measurements that fall into exclusively male sections of the male distributions, but it is rare that within a single brain *every* brain feature does so. For someone to have 'internal consistency,' they would need to have **all** their brain features fall into the distribution that matched their biological sex. For the brain to be a sexually dimorphic structure, **everyone's brain** would need to have internal consistency; which is very far from the case. Most human brains are more of a mosaic of all the different male-typical and female-typical brain features [16].

2.2 Sex Differences in the Brain

Many factors go into biological sex, following the broad overview in Section 2.1, this study will narrow the scope to the field of neuroscience in order to focus on sex differences in the brain. First, a long list of findings will be provided, followed by a summary of the brain features to be ultimately highlighted in the computational portion of this study.

Before jumping into the list of highlighted brain features, it is important to note that on average, males have larger intercranial volumes (ICV) [8, 13, 14, 18]. When comparing

brain regions across sexes, many studies will adjust the raw data based on ICV to get a sense of the proportion of the brain that a feature takes up; this aids a more reasonable comparison between the brains of different groups of individuals. For all the data used in the computational portion of this study, values that were already adjusted were sought out. Whenever this information was not available, proportions taking ICV into account were calculated and utilized. Some ICV correction methods that were used throughout the data given in the following section include ICV-proportions, ICV-residuals, and ANCOVA(IVC as a covariate of no interest). These methods were all compared by [24], and ANCOVA and ICV-residuals were found to be the most effective methods.

The data explained in Section 2.2.1 and summarized in Table 3.1 primarily comes from imaging methodologies including structural magnetic resonance imaging(MRI), functional magnetic resonance imaging(fMRI), positron emission tomography(PET), and single photon emission computed tomography(SPECT).

2.2.1 Broad Overview of Studied Sex Differences in the Brain

Now we examine the findings of research on sex differences in the brain. First, we have a volumetric examination of the temporo-limbic and prefrontal structures by [13]. They found that after adjusting for cranial volume, women had larger orbital frontal cortices, an area of the cortex devoted to emotional processing, than men.

An in vivo magnetic resonance imaging evaluation found that sexual dimorphisms in adult brain volumes were more evident in the cortex. After being adjusted for cerebrum size, women were found to have larger frontal and medial paralimbic cortex volumes, while men were found to have volumetrically larger frontomedial cortices, amygdalae, and hypothalamuses [11].

A meta-analysis conducted by [27] found regional sex differences in the volume and tissue density of the amygdala, hippocampus, and insula.

[20] found that women had volumetrically more grey matter in medial and lateral pre-

frontal areas, the superior temporal sulcus, the posterior insula, and the orbitofrontal cortex. They found that men had more grey matter volume dedicated to subcortical temporal structures. Some subcortical temporal structures include the amygdala, hippocampus, temporal pole, fusiform gyrus, visual primary cortex, premotor cortex, putamen, and anterior cerebellum.

A behavior driven neuroscience study found sex differences in tasks that relate to the orbital prefrontal cortex [26].

[28] states that "nonreproductive, cognitive functional differences between sexes might be reflected in higher-order cortical structural dimorphisms." They found that women had larger dorsolateral prefrontal cortices and superior temporal gyri in language related cortical regions, but not in visuospatial cortical regions.

[14] looked at the volumes of frontal and temporal regions, selected limbic structures, and the basal ganglia. They found age-associated changes that differed more or less drastically depending on the individual's sex.

In a study focused on the accessory olfactory system, which is a chemosensory system that detects and responds to pheromones, voxel-based morphology (a neuroimaging technique) found several sexual dimorphisms in several olfactory regions. Women were found to have higher concentrations of grey matter in the bilateral hippocampus and right amygdala, both orbitofrontal regions, and also higher concentrations of gray matter in the left basal insular cortex. Men were found to have higher grey matter concentrations in the left entorhinal cortex, right ventral pallidum, dorsal left insular cortex, and the region Brodmann's area 25 of the orbitofrontal cortex [10].

[29] used tensor-based morphometry and self-reported emotional regulation information to investigate sex differences in brain structure, as well as a possible relationship between orbitofrontal cortex subregions and affective individual differences. They found women to have larger ventromedial prefrontal cortices, right lateral orbitofrontal cortices, cerebellums, and bilateral basal ganglia. They also found that normal variation in the size of one's

ventromedial prefrontal cortex is related to the individual's emotional regulation methods.

The heteromodal association cortex(HASC) has been hypothesized to demonstrate sexual dimorphisms, specifically with respect to asymmetry. Frederiske et al examined MRI scans of the inferior parietal lobule(IPL), a neocortical region that is part of the HASC, and found that males had significantly larger left IPL volumes than females, creating a strong leftward asymmetry for the IPL. Meanwhile, females showed a rightward asymmetry in the IPL [9].

Structural differences in the Hippocampus may be suggested by sex differences that have been frequently observed in episodic and spatial memory. Persson et al studied hippocampal volumes and their structural covariance with the rest of the brain and found that after adjusting for brain size, women had larger posterior hippocampi (pHC). They found that in men the pHC showed a reliable structural covariance with the medial and lateral parietal lobes and the prefrontal cortex. They found that in women the anterior hippocampus (aHC) showed a reliable, bilateral structural covariance with the anterior temporal lobe. These differing findings of hippocampus regions' interaction with other parts of the brain supports a division of labor with regards to episodic and spatial memory [23].

Another quantitative volumetric MRI study with a focus on the underlying structural asymmetries in the human brain found differences related to sex and age. They found males to have larger cerebral, cerebellar, and cerebral cortical volumes (in all lobes except for the left parietal) and they found females to have greater left parietal to left cerebral hemisphere ratios and smaller left temporal to left cerebral hemisphere ratios [5].

[18] conducted a study with a focus on the parietal lobe, a region of the brain which is thought to be connected to spatial ability, specifically with regards to mental rotation. They found women to have proportionally larger gray matter volumes of the parietal lobe which was found to be disadvantageously related to mental rotations test(MRT) performance. Meanwhile, men were found to have a proportionally greater parietal lobe surface area which was found to be advantageously related to MRT performance.

[17] found little sex differences with regards to the volume of the amygdala as a whole;

however, an amygdala subregional shape analysis that controlled for intercranial volume found that men had a larger mean radius than women in the superficial nucleus while women showed a more rapid decline with age in the radius size of the centromedial nucleus.

Another study, [8], dealing with structural sexual dimorphism and asymmetry in the human cerebellum found that males had larger grey matter volumes in the anterior and medial posterior cerebellum, while females had larger grey matter volumes in the lateral posterior cerebellum. With regards to asymmetry, males showed more rightward asymmetry in lobules I_IV, IX, and Crus I along with less leftward asymmetry in lobules Crus II and VIIb.

2.2.2 Summary of Selected Regions

The regions that came up the most often throughout the literature review include all three frontal cortex regions: the prefrontal cortex, the visual primary cortex, and the premotor cortex [20]. Specifically, within the prefrontal cortex, the orbital region was mentioned in [11, 13, 20, 26, 10, 29], and it is one of the regions chosen to be highlighted in this study. The orbitofrontal cortex is thought to be involved "in sensory integration, in representing the affective value of reinforcers, and in decision making and expectation" [19]. The dorsolateral and ventromedial segments of the prefrontal cortex region were also included in several studies, [28, 29], so this study will also highlight the prefrontal region as a whole. The prefrontal cortex "plays a central role in cognitive control functions, and dopamine in the PFC modulates cognitive control, thereby influencing attention, impulse inhibition, prospective memory, and cognitive flexibility" [22].

Next, in regions of the limbic system, which is related to human emotions, learning and memory [1], many regions are emphasized. These regions include: the amygdala [20, 17, 10, 27, 11], a region of the brain though to process fearful and threatening stimuli [3]; the hippocampus [20, 10, 27, 11, 23], a region of the brain associated with episodic and spatial memory [23]; the hypothalamus [11], a region of the brain associated with vegetative system

control, homeostasis of the organism, thermoregulation, and emotional behavior adjustment [25]; And lastly, the basal ganglia [14], specifically the putamen [20], a region of the brain known for coordination and fine motor skills [1]. This study will highlight subregions of the amygdala, the posterior hippocampus (pHC), and the hypothalamus.

Next, commonly mentioned temporal lobe gyri include the fusiform gyrus [20], the superior temporal gyri and sulcus [28, 20], the entorhinal cortex [10], and the temporal pole [20]; the last two regions mentioned have been highlighted in this study. The temporal lobe is associated with understanding language, memory, hearing, and sequencing and organization [1].

The parietal lobe, a region thought to be involved in spatial ability and mental rotation [18] was highlighted by several studies [9, 5, 18]; the surface area of the left inferior parietal lobe will be featured in this study.

Lastly, the cerebellum, which performs functions like muscle coordination, posture maintenance, and balance, was mentioned in several studies [20, 29, 8]. Regions V and VIIIb of the cerebellum will be featured in this study.

2.3 Percentage of Gender Diversity in Society

The portion of this study that reflects gender in society will use statistics dealing with the proportion of the population that is transgender, non-binary, or both; these are listed in the remainder of this section. These statistics will inform the modification of this study's base model, which is constructed solely from the brain region statistical information. [12] conducted a meta-analysis and found that depending on the location and scope of the study, .01-2.7% of the population is transgender or gender non-conforming. [21] researched the percentage of transgender and non-binary individuals in 2017, and created a model to demonstrate how this would increase in the coming years. By following the trend in their model out to the year 2022, they predict 0.6% of the population is transgender and/or non-binary. The

Canadian census of 2021 found that 1% of the Canadian population between the ages of 20 and 24 identify as transgender and/or non-binary [4]. Finally, one study out of Pittsburgh, Pennsylvania found that up to 10% of teens are trans and/or non-binary [2]. These statistics are listed in table 5.2.

It is important to note that many of these statistics are most likely far lower than the true number of trans and non-binary people in the world around us. Fear of discrimination, family pressure, and countless other reasons may keep someone from living as their true self, let alone formally documenting their identity.

Chapter 3

Approach

3.1 Data Creation: Plan

A surrogate data set will be constructed to simulate data collected from the target regions of brain scans for a large number of hypothetical individuals. Average values and standard deviations for each brain region have been extracted from literature and will be used to create normal distributions for both men and women for all the target brain regions. These statistics are summarized in Table 3.1. Each data point representing an individual of the male sex will take a sample from the male distribution of each brain feature, while each data point representing an individual of the female sex will take a sample from the female distribution of each brain feature. Additionally, five "special" data points will be constructed to represent the extremes of each distribution which will act as landmark points during the analysis of our results.

3.2 Use of a Logistic Regression

Logistic regression functions map values between zero and L and look like the following:

$$f(x) = \frac{L}{1 + e^{-k(x-x_0)}} \tag{3.1}$$

Region	**Male Avg**	**Male SD**	**Female Avg**	**Female SD**	**Description**	**Source**
frontoorbital cortex	1.2	0.2	1.4	0.2	volume, cm^3	[11]
prefrontal region	13.42	0.9	13.28	1.67	%GM	[13]
amygdala SF	6.77	0.21	6.69	0.2	volume, mm^3	[17]
amygdala CM	8.22	0.35	8.22	0.41	volume, mm^3	[17]
amygdala LB	5.55	0.19	5.5	0.21	volume, mm^3	[17]
pHC	1014	89.5	1058.5	75.5	volume, mm^3	[23]
hypothalamus	0.1	0.01	0.1	0.02	volume, cm^3	[11]
basal ganglia	4.56	0.76	4.44	0.67	%GM	[14]
temporal pole and sup temp	5.77	0.6	5.56	0.5	%GM	[14]
IPL	20880.9	164.3	20292.7	164.3	surface area, mm^2	[18]
Cerebellum region V	8.168	0.918	7.352	0.65	volume, cm^3	[8]
Cerebellum region VII	7.78	1.077	6.517	0.92	volume, cm^3	[8]

Table 3.1: Table of Data for All Brain Region Stats

where L is the maximum value, k is the logistic growth rate or steepness of the curve, and x_0 is the inflection point of the curve. A *logistic regression* is a linear model that aims to classify data with either a zero or a one. Probabilities describing the possible outcomes of a single test data point being inputted into the logistic regression model are projected using a logistic regression function with a maximum of one, $L = 1$. This study plans to use a logistic regression model, where the classification of zero predicts 'male' and the classification of one predicts 'female.' The real interest however, is in the probabilities reflected by the logistic regression function that lie in between the two extremes of zero and one; it is these numbers that will be used to take biological sex from a binary with *only* the options of zero and one to a spectrum with the values of all the infinite possible values in between. This study will refer to the function that generates these probabilities as the *logistic regression's probability function.*

3.3 Varying the Coefficients of the Sigmoid Curve to Reflect Gender in Society

3.3.1 Sigmoid(S) Curve

This study will use a slightly different coefficient arrangement than what is shown in Equation 3.1 to align with existing logistic regression probability function conventions. The remainder of the study will use the following:

$$p(x) = \frac{1}{1 + e^{\frac{-(x-\mu)}{s}}} \tag{3.2}$$

where $p(x)$ is the probability corresponding to a given x, μ is the location parameter or inflection point, and s (which corresponds to $\frac{1}{k}$) is the scale parameter, which dictates the steepness of the curve.

3.3.2 Varying Degrees of the Steepness Coefficient

In order to use the base logistic regression model introduced in the previous section to reflect a pure binary, each data point would need a predicted probability value of either exactly zero or one, with absolutely nothing between the two extremes. This would theoretically correspond to an S-curve that is simply two linear segments, one at y=0 for the first half of the x-range, and one at y=1 for the second half on the x-range. This is the limit case as s approaches zero. In Figure 3.1, in row (1), an s that is close to zero is graphed on the left. Note the steepness of the curve, and the corresponding clustering of the points in the scatter plot to the right. Almost all of the points fall at exactly probability zero or exactly probability one.

The opposite extreme would be a s value of infinity. It would result in a linear S-curve, with an infinite number of different probability values between 0 and 1. In Figure 3.1, a larger s value is used to graph something that approaches this scenario. On the left-hand

side, we see a line that looks nearly linear, and on the right hand side the points are equally spread out, with no clustering occurring on either end.

All the graphs in between, rows 2-4 of Figure 3.1, demonstrate differing stages along potential spreads of gender in our society. We have clusters on either end of the spectrum, with loosely spread points between the two extremes. In Figure 3.2, all these curves are plotted together to aid the ease of visual comparison.

The s value has an inverse relationship with the harshness of the curve. As the curve softens, more room is created for increased gender diversity. As the curve steepens, the realm of different possible gender identities is dampened. This is the core concept that will be used to skew the steepness of the base logistic regression to reflect societal influence.

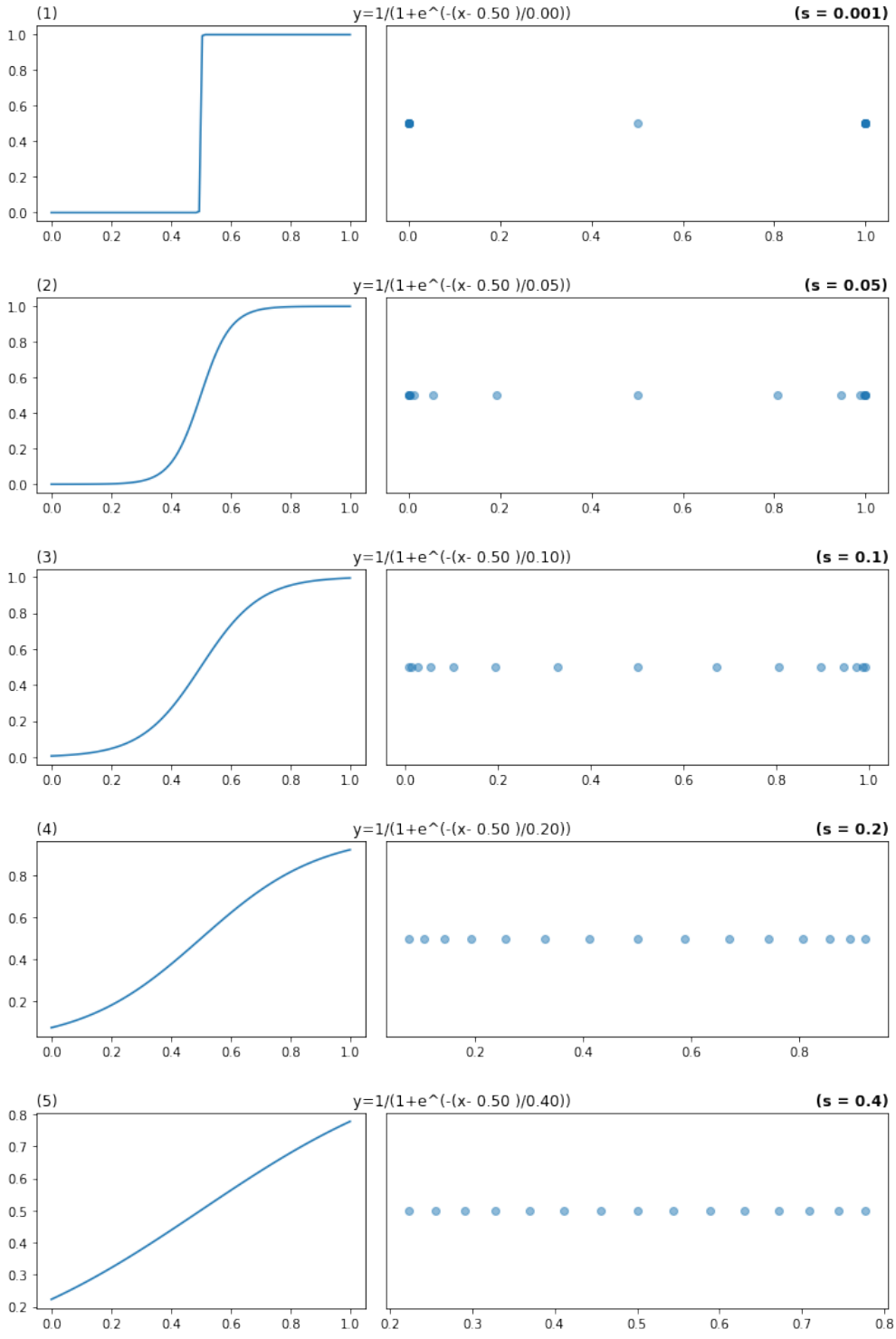

Figure 3.1: Demonstration of the relationship between s and point clustering

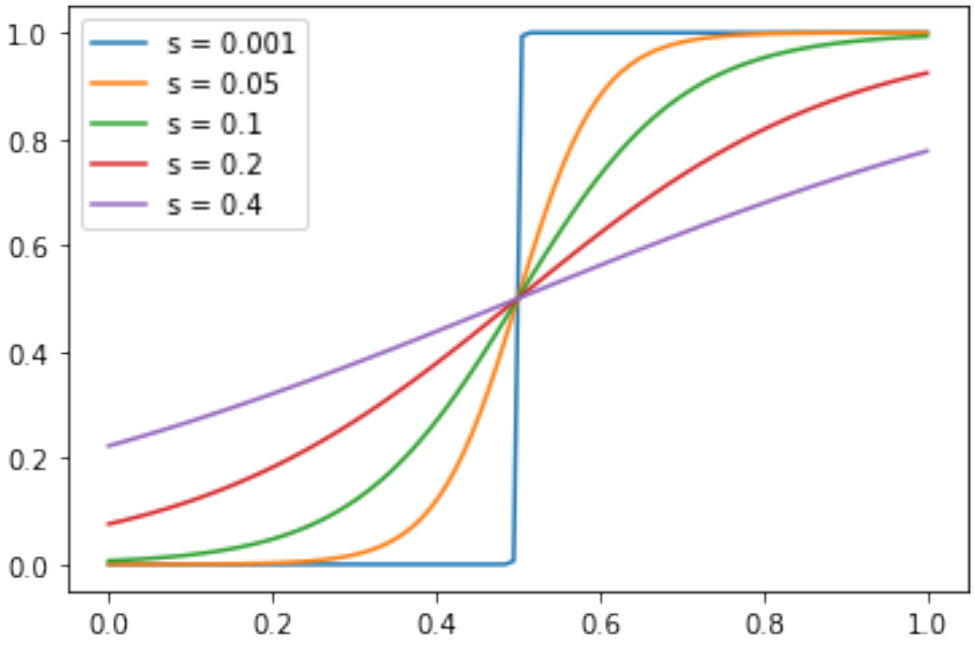

Figure 3.2: Visual comparison of S curves with different s values

3.3.3 Using the Biological Overlap Region to Alter s to Reflect Societal Influence

We rely on the range between two important values to represent the shift from both the gender extremes to the region between them. First we have the minimum probability value that is associated with an afab data point, we call this value min_1. Every data point with a probability lower than $Prob(min_1)$ came from *exclusively* amab data points. Next we have max_0, the maximum probability corresponding to an originally male data point. Every data point with a probability larger than $Prob(max_0)$ came from *exclusively* female data points. The data points corresponding to probabilities between these points, $Prob(range(min_1, max_0))$, represent the data points that fall between the clusters on either end of the gender spectrum. We call this segment the *biological overlap region*.

Chapter 4

Methods

4.1 Data Creation: Implementation

This study will use Python for the data set construction and the image generation. Imported libraries included are `pandas`, `numpy`, `sklearn`, `scipy`, and `random`. We import the following functions: `pyplot`, `mpatches`, `mlines`, and `colors` from `matplotlib`. From `matplotlib.colors` we import `LinearSegmentedColorMap`. From `sklearn`'s `utils` package `shuffle` is used, `linear_model` is imported, and `preprocessing` is imported. From `scipy`'s `optimize` package, `curve_fit` is imported.

4.1.1 Generalized Surrogate Data Set Creation

Each data point in the data set represents the hypothetical brain scan of a single individual. Each has a 'sex' attribute, with the categorical value '0' representing males, and the categorical value '1' representing females. We will refer to this as the *gender code* in further sections. In total, 1500 data points are created; 750 data points are 'male' and the remaining 750 data points are 'female.' Each data point then has a value corresponding to each brain region being included in the current instance of the simulation; 'male' data points' brain region attribute values are selected from the corresponding male brain region distributions,

while the 'female' data points' values are similarly chosen from the female distributions. The distributions being referred to are normal distributions constructed by the data in Table 3.1.

The `generateData(size, filename)` function is passed a desired size for the data set, in our case 1500, as well as a filename of a comma separated list(.csv) file in the proper format holding the data on each brain region for both sexes. Both the desired format for the data file and the function source code can be found in Section 7.1, and Figure 7.1. Brain region attribute values are selected using the `gauss(mu, sigma)` method in Python's random class, where μ is the average value and σ is the standard deviation; the output is a value from the normal, also known as the gaussian, distribution corresponding to the inputs. In the data file we pass to our `generateData(size, filename)` function, which is also referred to as the *statistics table*, we must have the male average, male standard deviation, female average, and female standard deviation values for each brain region listed.

The dataset produced by the `generateData(size, filename)` function is a `pandas dataFrame`. Each row is a single data point (which represents a single individual), while each column holds the value of a different brain region attribute; except, the leftmost column holds the gender code.

The last step is to scale the data. The different attributes have values that are in different units or at different orders of magnitude depending on what study they come from or how they represent the brain feature. In order to keep one attribute from overpowering all the others, the `MinMaxScaler()` function from the `preprocessing` package of Python's `scikit learn` library was used to transform all the data in the data set to create a scaled data set where all the attributes have values between zero and one.

4.1.2 Special Data Points

An array holding five handpicked, *special data points* was created to provide reference points in graphical representations throughout the study. They are 'male extreme,' 'male moderate,' 'overall average,' 'female moderate,' and 'female extreme.' Using the *statistics table*

corresponding to the data set in question, 'extreme' values were calculated by taking the average and either adding or subtracting the standard deviation to get the value as far away from the opposing gender's average value. 'moderate' values were calculated by taking the average and either adding or subtracting the standard deviation to get the value as close as possible to the opposing gender's average value. Finally, the 'overall average' data point was calculated by averaging the male and female averages. These points were provided as test data for the logistic regression's `predict_proba()` functions that is explained in the next section, and can be seen graphically as the larger black dots in the results. The function used to create this array can be found in Section 7.3.

4.2 Base Logistic Regression: Implementation

The `LogisticRegression()` function in python's `scikit learn` library was used to build the base logistic regression. It was initialized with the *stochastic average gradient descent (sag)* solver (because this solver was found to converge faster for some higher-dimensional data, according to the scikit-learn's documentation), and the *max-iterations* was set to 100. The data explained in section 4.1.2 was split into two segments in order to 'fit' the model: the leftmost column holding the categorical sex code became the target vector component, while the remaining brain region attribute data was the training data.

Once the logistic regression was initialized and split with the training data, new test data was created using the previously explained `dataGeneration()` function. Again, half the 1500 test data points were created using the male distributions and half using the female distributions; the class of each test data point was recorded in order to be color coded, where blue represents amab data and pink represents afab data, in the graphical representations. Each data point in the test set was used as an input for to the logistic regression's `predict_proba` function which returns the probability that the sample is in the 'male' class and the 'female' class. Our logistic regression has only two possible classes (male and fe-

male), so to put together a graphical representation of the model it was only necessary to record the probability, p, of one class (because the remaining class's probability is $1 - p$). The female probability was chosen to be recorded because then a p that is close to 1 implies that the data point is close female, whose code is also '1'.

At this point, we have a set of probabilities, each marked with the color of the sex code of the corresponding data point. We iterate through this array of data points, and record the probabilities of that noteworthy data points min_1 and max_0 in order to calculate the *biological overlap region*. We calculate the percentage of our data set that falls in this region like so:

$$Overlap = \frac{\sum_{i=0}^{len(data)} 1 \quad if Prob(min_1) \leq Prob(x_i) \leq Prob(max_0); \quad else \quad 0}{len(data)} \cdot 100 \qquad (4.1)$$

Next, we create a visual representation of all this information. Since we are dealing with multidimensional data, we are unable to simply graph probability on the y-axis and one independent variable on the x-axis. To get a feel for the shape of the set of probabilities we have created, we sort them from smallest to largest magnitude and graph the points in order. This reveals a scatter plot that looks like a classic sigmoid(S) curve. Note that this is a function of 2D space, and it is separate from the logistic regression's probability function, which exists in 13D space. We will continue to be call this the S curve or the *sigmoid* curve throughout this document.

We use the points in this scatter plot to get an actual equation for the corresponding S curve, equation 3.2, with the `curve_fit()` method from the `optimization` package in Python's `SciPi` library. This function will give us the values of the parameters μ and s, which in turn gives us information about the location of the inflection point and the steepness of the curve.

4.3 Skewed Logistic Regression: Implementation

In efforts to reflect the societal pressure to conform to the gender that an individual is assigned to at birth, the probabilities produced by the logistic regression's `predict_proba()` function were scaled towards the gender matching the individual's sex in the following way:

$$f' = f + \frac{1 - f}{\alpha} \tag{4.2}$$

$$m' = m - \frac{m}{\alpha} \tag{4.3}$$

where f is the original probability of a female data point, alpha is some constant, and f' is the adjusted probability. Similarly, m and m' represent the probabilities of male data points. The further a point is from the extremes of its assigned sex, the further it is pulled towards the extreme. This is to reflect the fact that oftentimes, the more an individual doesn't align to societal constructs, the more of an impact societally influenced gender stereotypes have to pull them back in. As α gets closer to one, the closer we represent a true gender binary with a steeper resulting S curve; as α approaches infinity, the skewed logistic regression approaches the base regression, which approaches a society free from gender identity constraints. The value of α was varied to produce different graphs to search one that appropriately reflects society.

Once we have the new, skewed, sets of probabilities, we follow the exact same process that is outlined above for the base linear regression to produce analogous visual representations for the new sets of probabilities. We vary *alpha*, increasing and decreasing it accordingly, until we have probability sets whose *biological overlap regions* represent the population percentages regarding trans and non-binary folks that were extracted from literature. We fit these scatter plots to S curves, and analyze the resulting steepness coefficients.

Chapter 5

Results

5.1 Surrogate Data Set Construction

The created data frame has the following 13 columns: (1) sex (which is a zero for male data points and a one for female data points), (2) frontoorbital cortex, (3) %grey matter(GM) in prefrontal region, (4) amygdala superficial group (SF), (5) amygdala centromedial group (CM), (6) amygdala laterobasal group (LB), (7) posterior hippocampus (pHC), (8) hypothalamus, (9) %GM basal ganglia, (10) %GM temporal pole and superior temporal gyrus, (11) Surface Area of the inferior parietal lobe (iPL), (12) Cerebellum region V, and (13) Cerebellum region VIII. There are 1500 rows, with 750 sampling from the male distributions, and the remaining 750 sampling from the female distribution. This entire process was done twice to create both a testing set and a training set of data. The .csv file holding the generated values is available upon request, the first and last 5 rows of data can be found in Section 7.4, and a visual representation of color-coded histograms for each column can be found seed in Figures 5.1 and 5.2:

First we have, in figure 5.1, a visual representation of the generated training data:

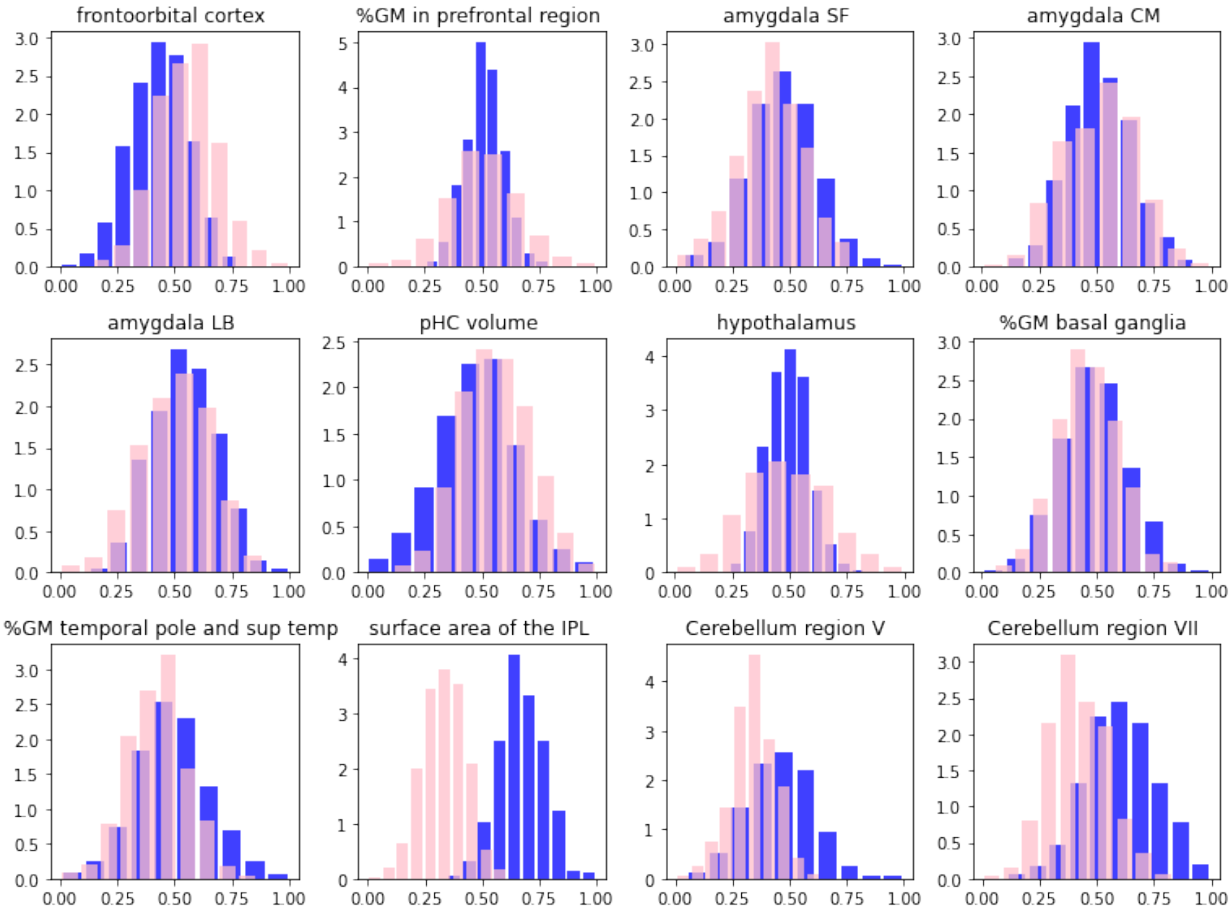

Figure 5.1: **Training Data**, where pink represents afab data and blue represents amab data

Next we have, in figure 5.2, a visual representation of the generated testing data:

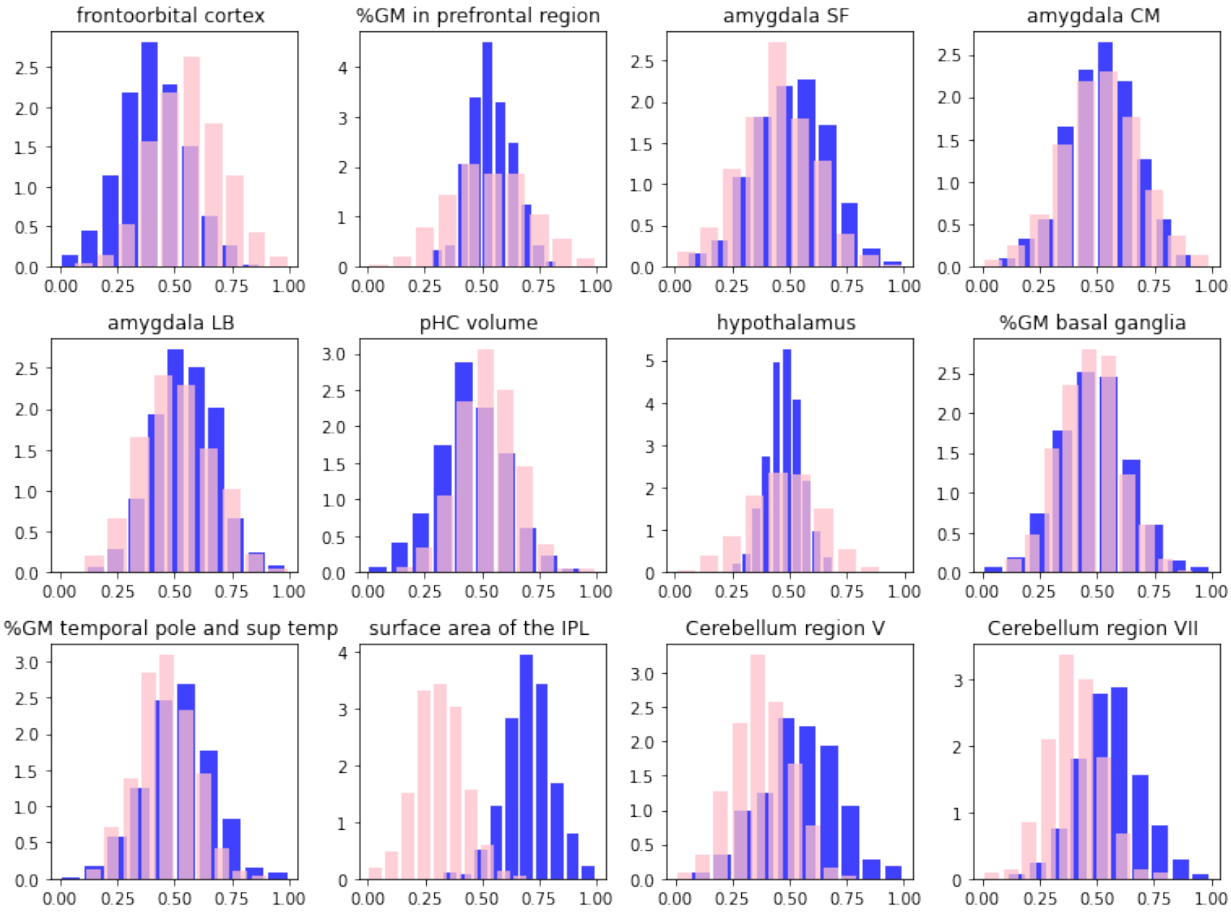

Figure 5.2: **Testing Data**, where pink represents afab data and blue represents amab data

5.2 Base Logistic Regression

5.2.1 Construction

The base logistic regression took 22 iterations to converge, has an intercept of 9.67 in the exponential portion of the logistic regression's probability function's denominator, and the following table, Table 5.1, lists each coefficient next to its corresponding feature, ranked from highest influence on a female outcome to lowest.

The intercept and the coefficients listed in Table 5.1 are used to construct the probability function. The process is shown below, where x_i represents the brain feature value from the

fronto-orbital cortex	3.32772125
pHC	1.76120437
amygdala CM	-0.10100711
prefrontal region	-0.51788266
hypothalamus	-0.56604925
basal ganglia	-0.68760602
amygdala LB	-1.06142803
amygdala SF	-1.29146921
temporal pole and sup temp	-1.76561761
Cerebellum region V	-3.10460268
Cerebellum region VII	-4.01001899
IPL	-12.16977355

Table 5.1: Coefficient for each logistic regression feature

i^{th} column of the data point whose probability is being calculated.

First we have the structure of the linear portion of a logistic regression:

$$z = \theta_0 + x_1 \cdot \theta_1 + x_2 \cdot \theta_2 + ... + x_{12} \cdot \theta_{12} \tag{5.1}$$

Followed by the equation with our model's information plugged in:

$$z = 9.67 + 3.33x_1 - 0.51x_2 - 1.29x_3 - 0.10x_4 - 1.06x_5 + 1.76x_6 - 0.57x_7 - 0.69x_8 - 1.77x_9 - 12.17x_{10} - 3.10x_{11} - 4.01x_{12} \tag{5.2}$$

Next, we show how Equation 5.2 gets used to create the probability function of our logistic regression:

$$e^z = \frac{P}{1-P} \tag{5.3}$$

$$P = \frac{1}{1+e^{-z}} \tag{5.4}$$

Putting all of this information together,the equation for finding the probability based on

brain feature values is as follows:

$$P = \frac{1}{1 + e^{-(9.67+3.33x_1-0.51x_2-1.29x_3-0.10x_4-1.06x_5+1.76x_6-0.57x_7-0.69x_8-1.77x_9-12.17x_{10}-3.10x_{11}-4.01x_{12})}} \tag{5.5}$$

Next, the testing data is given to the logistic regression model. We use the method outlined in Section 4.2 to create a 2D visual representation of our data. The probability that each point is female is recorded in an array, plotted, and fit to an S curve. The equation we are left with is as follows:

$$y = \frac{1}{1 + e^{\frac{-(x-\mu)}{s}}} = \frac{1}{1 + e^{\frac{-(x-771.26)}{72.00}}} \tag{5.6}$$

Since $\mu = 771.26 > 750$, the midpoint of this curve is to the right of center, demonstrating slightly more variability in the 'male' half of the created spectrum. Note the value of $s = 72.00$, the steepness coefficient, as we will want to compare this number to our findings in the next section.

5.2.2 Visual Representation

In Figure 5.3, the top left panel (labeled *Separate View*) shows all the male data points plotted before the female data points in the x-direction. This view more clearly visually represents the y-direction overlap between the blue and pink data points, which represent male and female data points respectively. The top panel of Figure 5.3 shows a plot named *Overlay View*. The data points are plotted with 50 percent transparency; where deeper pinks and blues occur we have a higher concentration of similar data points, and where different shades of purple emerge we have varying degrees of overlap between data points of the opposite sex. This view provides a different demonstration of overlap with respect to color instead of height. Both plots on the first row of this figure conceptually correspond to

the graphs on the left-hand side of F 3.1.

In the second row of Figure 5.3, we have what conceptually corresponds to the right-hand side of Figure 3.1; this portion of the figure demonstrates the degree of clustering of similar points, as well as the overlap of opposite points. As explained in the legend in the center of row one, the black points represent the special data explained in Section 4.1.2. As we would expect, the male extreme, labeled $maleH$, has a probability all the way to the edge at zero, the female extreme, labeled $femaleH$, is assigned to a probability all the way on the other side at one, and the midpoint, labeled mid, falls right around a probability of 0.5. It is interesting to note that the opposite extremes, $maleL$ and $femaleL$, which are date points made up of the *moderate* data values, fall in what appears to be the region of the opposite gender. This model would suggest that data points similar to these could represent transgender individuals; to clarify, data points near $maleL$ would represent transgender females, and data points near $femaleL$ would represent transgender males. Note that this visual comes from the logistic regression's probability functions, not the fitted S curve.

Another feature of this graph is the purple boxed in *biological overlap region*, which is conceptually explained in the second half of Section 4.2. This region spans the distance between the smallest probability of an originally female point, $min1$, and the largest probability of an originally male point, $max0$. The *Range of Overlap* is calculated by dividing the total number of data points by the number of data points between $min1$ and $max0$; this process is outlined in Equation 4.1. In the case of our base logistic regression model, the value for this field is 70.38%. This model suggests that up to 70.38% of individuals could fall into the realm of trans or non-binary individuals solely by the biological make-up of their brains.

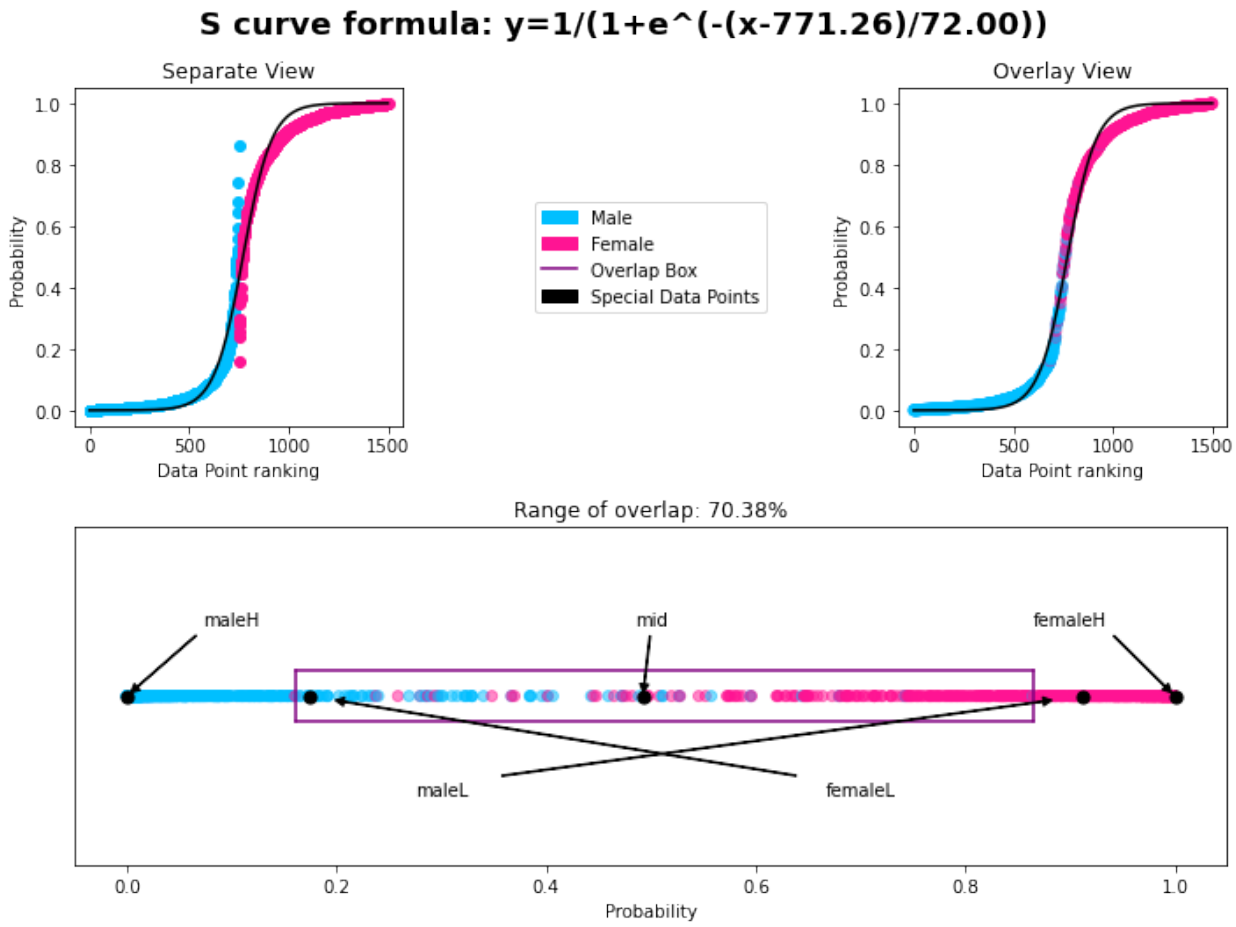

Figure 5.3: **Base Logistic Regression**, note that the black curves in both panels of row one are identical and are not included in the legend

5.3 Skewed Logistic Regression Skewed

Next, we aim to model the societal influence on gender by matching *biological overlap ranges* to actual statistics collected on the percentage of transgender and non-binary individuals in our world. Recall the following statistics regarding the numbers of transgender and non-binary individuals:

Throughout the six graphs in Figures 5.4 and 5.5, the necessary α was selected to produce a plot with the following desired overlap regions from Table 5.2: (1) $\alpha = 1.5$ gives a *Range of Overlap* $= 0.00\%$ to reflect a strict binary, (2) $\alpha = 2.4213$ gives a *Range of Overlap* 0.01%, to reflect the lower end of [12]'s range, (3) $\alpha = 2.423$ gives a *Range of Overlap* of 0.06%, to

Percentage	Description	Source
0.01-2.7	Meta analysis across many locations and scopes	[12]
0.6	Projected number of Americans	[21]
1.0	Canadians between the ages of 20 and 24 in the 2021 census	[4]
10.0	Teens in Pittsburgh	[2]

Table 5.2: Proportion of transgender and non-binary individuals

reflect [21]'s metric, (4)$\alpha = 2.4556$ gives a *Range of Overlap* of 1.00%, to reflect [4]'s metric, (5)$\alpha = 2.1573$ gives a *Range of Overlap* of 2.70% to reflect the higher end of [12]'s range, and finally (6)$\alpha = 2.822$ gives a *Range of Overlap* of 10% to reflect [2]'s metric.

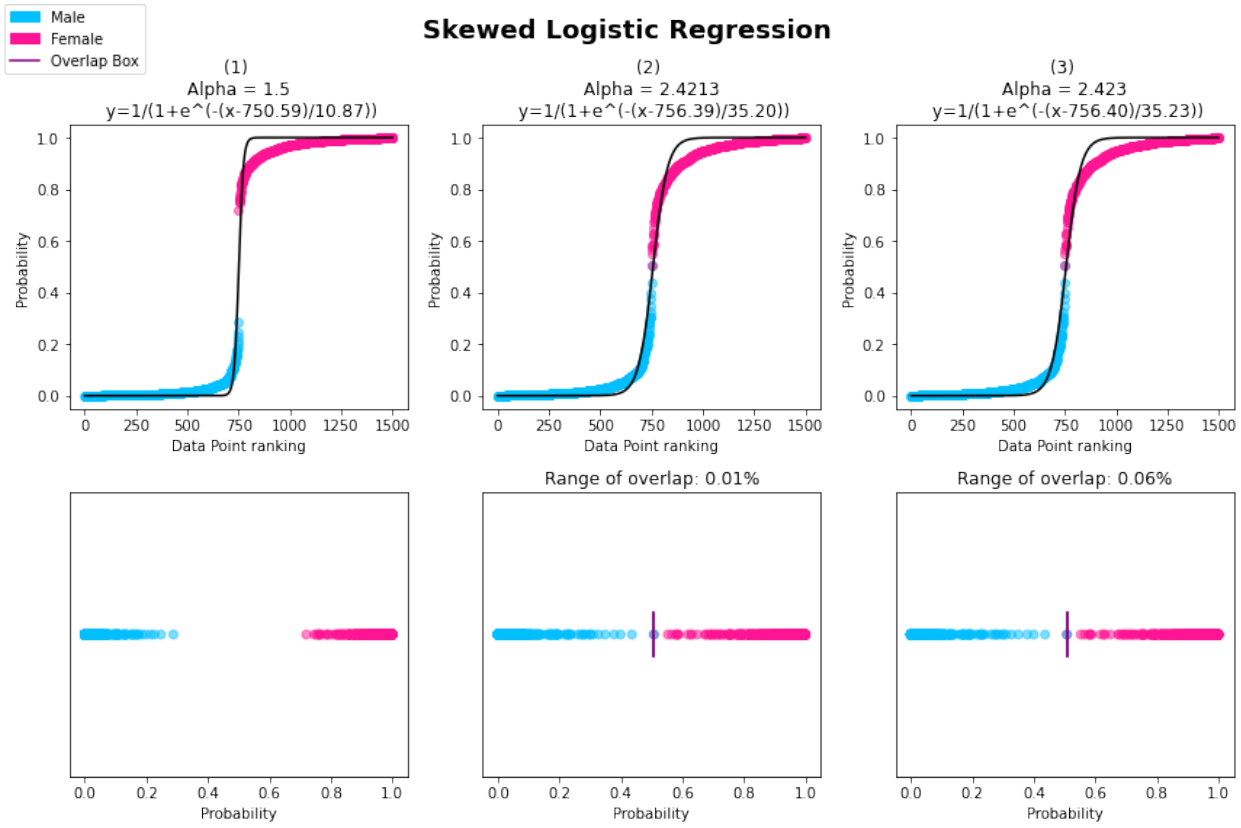

Figure 5.4: **Skewed Logistic Regression**, the S curve is shown in black in row 1

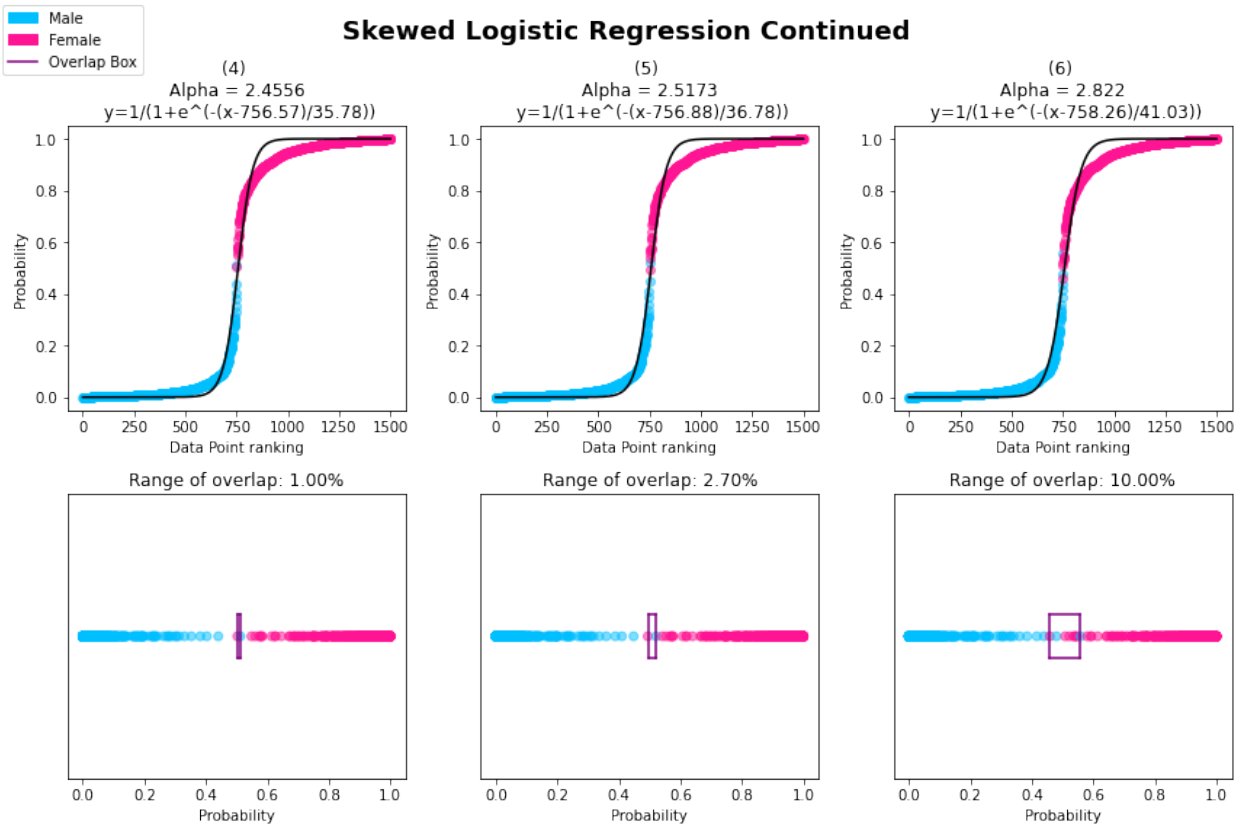

Figure 5.5: **Skewed Logistic Regression Continued**, the S curve is shown in black in row 1

Chapter 6

Conclusion

6.1 Discussion

The base logistic growth model suggests that only 15% of the population falls into an exclusively male region, and only 15% of the population falls into an exclusively female region. The remaining 70% fall somewhere between the two extremes, with some brain region measurements drawing values from accepted ranges of the opposite biological sex. These percentages are extremely different than those found representing proportions of transgender and non-binary individuals extracted from literature; meaning that the societal influence on individuals with the biological potential to be gender diverse is significantly stifled by the societal construct of the gender binary.

In the skewed regression portion of the results section, we can see that the further α gets from 1, the larger the societally perceived biological overlap section begins to grow. If we take the $\lim_{\alpha \to \infty} P'(x)$, where $P'(x)$ is the skewed probability function, societal influence would be theoretically removed, and $P'(x)$ would approach the original probability function constructed by the base logistic regression model.

When researching the proportion of transgender and non-binary individuals in society, it is clear that the values of these statistics are on the rise[21]. Significantly more young

people are reporting that they fall into gender diverse categories; as these young people grow up and this trend continues, the overall percentages are bound to grow. This increase of visible transgender and non-binary individuals also increases the representation of trans and non-binary individuals within society, both in the media we consume and on a personal level in everyday life; this visibility diminishes the hold the concept of the gender binary has on society, which will in turn also serve to increase the percentage of transgender and non-binary individuals. This combination of an increase in transgender and non-binary individuals and an increase of visible transgender and non-binary individuals is a self-perpetuating cycle. Over time, this study predicts that the α values in the skewed regression portion will have to grow along with societal changes to continue to provide an accurate representation of our world, with an upperbound of a spread of gender based solely on biological factors.

6.2 Limitations

First, the fact that we are using a surrogate data set is a limitation. Sampling values from a normal distribution when constructing this data set is simply a starting point, rather than a scientifically educated choice. Finally, each brain region value is independent from one another, which is also a limitation.

It is important to note that grouping transgender and non-binary individuals together is an oversimplification; more research specifically on the neuroscience of trans individuals would have been necessary to create more informed assumptions. This is especially relevant in the field of neuroendocrinology for trans and non-binary folks who choose to seek out certain gender confirmation medical treatments, as the human brain changes and adapts to the introduction of new hormones and environments. Some transgender individuals do identify with the binary opposite to what they were assigned at birth, and in some ways it is a disservice to lump them in with non-binary individuals. Clearer articulation between "biological sex" and "gender" could more respectfully represent certain trans folks than what

appears in this study.

It is also important to note that the statistics used to govern the *Societally Influenced Gender Spectrum* may be extremely inaccurate. Unfortunately, many individuals do not feel safe, comfortable, willing, or able to document their gender identities for a myriad of reasons from fear of gender identity based discrimination to a lack of love and support from family and friends.

Another limitation of this study is on the inclusion of intersex individuals. Less data is available on these individuals, and further knowledge on the topic would be needed to create accurate assumptions to govern the appropriate data creation. This study also does not take individuals who identify as agender into account.

Finally, as explained at the start of the literature review, the brain is only one element of what can be used to demonstrate biological sex differences. At that, volumetric measurements are also only one element of biological sex differences in the field of neuroscience. Sex and gender are extremely complicated topics with many different layers of detail; more can always be included, studied, and used to display the main idea of this study. In the field of modeling, the balance between too much information and not enough to produce the best possible results can always be re-evaluated, and that is the case with this study. Sex differences from different fields of study could be included to increase the complexity of the data being inputted, while less important sex differences in the brain that were included could be removed. I truly hope that this study can be expanded into something impactful; I thank you for reading!

Chapter 7

Appendixes

7.1 Data Generation Function

```
def generateData(size, filename):
    stats= pd.read_csv (filename).iloc[:,1:]
    n=len(stats.columns)
    tempm = np.zeros((1,n))
    tempf = np.zeros((1,n))

    #create first two rows of data
    tempm[0,0] = 0
    tempf[0,0] = 1
    for i in range(1,n):
        tempm[0,i] = random.gauss(stats.iloc[0,i],stats.iloc[1,i])
        tempf[0,i] = random.gauss(stats.iloc[2,i],stats.iloc[3,i])
    data = pd.DataFrame(tempm,columns=stats.columns)
    data = data.append(pd.DataFrame(tempf,columns=stats.columns))
```

```
#create rest of data
for x in range(0,int(size/2)-1):
    tempm[0,0] = 0
    tempf[0,0] = 1
    for i in range(1,n):
        tempm[0,i] = random.gauss(stats.iloc[0,i],stats.iloc[1,i])
        tempf[0,i] = random.gauss(stats.iloc[2,i],stats.iloc[3,i])
    data = data.append(pd.DataFrame(tempm,columns=stats.columns))
    data = data.append(pd.DataFrame(tempf,columns=stats.columns))

data = data.sample(frac=1).reset_index(drop=True)

# Apply MinMax Scaler
scaler = preprocessing.MinMaxScaler()
scaled_data = scaler.fit_transform(data)
scaled_data = pd.DataFrame(scaled_data, columns = data.columns)

return scaled_data
```

7.2 Desired Form of Statistics File

		Brain Region Name 1	Brain Region Name 2	...	Brain Region Name N
male avg	0				
male std	0				
female avg	1				
female std	1				

Figure 7.1: Format for file given to generateData

7.3 Special Data Points Generation Code

```
def specialData(filename):
    stats= pd.read_csv (filename).iloc[:,2:]
    df = pd.DataFrame(columns = stats.columns, index = ['maleH', 'maleL',␣
 ↪'mid','femaleL', 'femaleH' ])
    for i in range(len(stats.columns)):
        if(stats.iloc[0,i]>stats.iloc[2,i]):
            df.iloc[0,i]=stats.iloc[0,i] + stats.iloc[1,i]
            df.iloc[1,i]=stats.iloc[0,i] - stats.iloc[1,i]
            df.iloc[3,i]=stats.iloc[2,i] + stats.iloc[3,i]
            df.iloc[4,i]=stats.iloc[2,i] - stats.iloc[3,i]
        else:
            df.iloc[0,i]=stats.iloc[0,i] - stats.iloc[1,i]
            df.iloc[1,i]=stats.iloc[0,i] + stats.iloc[1,i]
            df.iloc[3,i]=stats.iloc[2,i] - stats.iloc[3,i]
            df.iloc[4,i]=stats.iloc[2,i] + stats.iloc[3,i]
        df.iloc[2,i]=(stats.iloc[0,i] + stats.iloc[2,i])/2
    # Apply MinMax Scaler
    scaler = preprocessing.MinMaxScaler()
    scaled_data = scaler.fit_transform(df)
    scaled_data = pd.DataFrame(scaled_data, columns = df.columns, index =␣
 ↪df.index)


    return scaled_data
```

7.4 First and Last 5 Rows

7.4.1 Training Set

	sex	frontoorbital cortex {goldstein2001normal}	\
0	1.0	0.551636	
1	1.0	0.374266	
2	0.0	0.435482	
3	1.0	0.413086	
4	1.0	0.609731	
...	...	...	
1495	1.0	0.762958	
1496	0.0	0.521304	
1497	1.0	0.362092	
1498	1.0	0.530653	
1499	0.0	0.363570	

	%GM in prefrontal region {gur2002sex}	amygdala SF {kim2012sex}	\
0	0.864451	0.293584	
1	0.547353	0.637666	
2	0.386575	0.500192	
3	0.434491	0.442195	
4	0.536055	0.655391	
...	...	...	
1495	0.295136	0.569869	
1496	0.495561	0.436405	
1497	0.742950	0.579531	
1498	0.423399	0.343295	

```
1499                                0.504423                  0.646232


      amygdala CM {kim2012sex}  amygdala LB {kim2012sex}  \
0                     0.316698                  0.490001
1                     0.610590                  0.265531
2                     0.734007                  0.388647
3                     0.445442                  0.696581
4                     0.358952                  0.424306
...                        ...                       ...
1495                  0.433925                  0.632484
1496                  0.479979                  0.598765
1497                  0.522710                  0.381699
1498                  0.516279                  0.313194
1499                  0.595225                  0.770857


      pHC volume {perrson2014sex}  hypothalamus {goldstein2001normal}  \
0                        0.658779                              0.697941
1                        0.846820                              0.340922
2                        0.318658                              0.475348
3                        0.497127                              0.545029
4                        0.490658                              0.500138
...                           ...                                   ...
1495                     0.524280                              0.359250
1496                     0.583763                              0.519773
1497                     0.871072                              0.597367
1498                     0.535472                              0.670598
```

```
1499                    0.326658                                0.540320


      %GM basal ganglia {gur2002brain}  \
0                              0.408029
1                              0.367913
2                              0.230517
3                              0.247754
4                              0.509555
...                                 ...
1495                           0.582879
1496                           0.398196
1497                           0.423545
1498                           0.646237
1499                           0.484167


      %GM temporal pole and sup temp {gur2002brain}  \
0                                          0.661881
1                                          0.238319
2                                          0.783647
3                                          0.364632
4                                          0.301474
...                                             ...
1495                                       0.180291
1496                                       0.368335
1497                                       0.356362
1498                                       0.481232
```

```
1499                                    0.571547


      surface area of the IPL {koscik2009sex}  \
0                                      0.180149
1                                      0.406519
2                                      0.671412
3                                      0.163833
4                                      0.327474
...                                         ...
1495                                   0.417442
1496                                   0.545874
1497                                   0.424478
1498                                   0.419125
1499                                   0.764180


      Cerebellum region V {fan2010sexual  Cerebellum region VII␣
 ↪{fan2010sexual
0                           0.329887                           0.520849
1                           0.446902                           0.428153
2                           0.685311                           0.795049
3                           0.446985                           0.452570
4                           0.507938                           0.238303
...                              ...                                ␣
 ↪...
1495                        0.321445                           0.279423
1496                        0.355813                           0.193440
```

```
1497                          0.333031                                  0.412350
1498                          0.265067                                  0.302823
1499                          0.549494                                  0.653400

[1500 rows x 13 columns]
```

7.4.2 Testing Set

[6]:
```
      sex  frontoorbital cortex {goldstein2001normal}  \
0     0.0                                     0.266815
1     0.0                                     0.256059
2     1.0                                     0.631510
3     0.0                                     0.278588
4     0.0                                     0.324913
...   ...                                          ...
1495  1.0                                     0.539391
1496  0.0                                     0.453590
1497  1.0                                     0.613846
1498  1.0                                     0.536739
1499  1.0                                     0.632863

      %GM in prefrontal region {gur2002sex}  amygdala SF {kim2012sex}  \
0                                  0.598055                  0.357589
1                                  0.387410                  0.646699
2                                  0.555595                  0.298508
3                                  0.404813                  0.372706
4                                  0.584001                  0.566669
```

```
...                                    ...                          ...
1495                                0.346684                     0.441049
1496                                0.521619                     0.584113
1497                                0.439358                     0.580036
1498                                0.358377                     0.549461
1499                                0.572318                     0.433431

      amygdala CM {kim2012sex}  amygdala LB {kim2012sex}  \
0                     0.588851                  0.557197
1                     0.575747                  0.569671
2                     0.431289                  0.682171
3                     0.067165                  0.566593
4                     0.532372                  0.442702
...                        ...                       ...
1495                  0.500460                  0.646385
1496                  0.527380                  0.431580
1497                  0.654162                  0.379306
1498                  0.500772                  0.558339
1499                  0.951164                  0.563251

      pHC volume {perrson2014sex}  hypothalamus {goldstein2001normal}  \
0                        0.261031                             0.462057
1                        0.442622                             0.347077
2                        0.577369                             0.332801
3                        0.606951                             0.345035
4                        0.240114                             0.482480
```

```
...                          ...                                ...
1495                    0.748724                           0.590033
1496                    0.464654                           0.514735
1497                    0.692764                           0.537320
1498                    0.395355                           0.449880
1499                    0.388773                           0.354566

      %GM basal ganglia {gur2002brain}  \
0                           0.767862
1                           0.448437
2                           0.505446
3                           0.598563
4                           0.733157
...                              ...
1495                        0.571109
1496                        0.545421
1497                        0.726344
1498                        0.769284
1499                        0.512463

      %GM temporal pole and sup temp {gur2002brain}  \
0                                          0.313136
1                                          0.506778
2                                          0.620524
3                                          0.506574
4                                          0.384754
```

```
...                                                 ...
1495                                           0.314736
1496                                           0.540982
1497                                           0.387938
1498                                           0.534381
1499                                           0.523165


      surface area of the IPL {koscik2009sex}  \
0                                         0.672912
1                                         0.628915
2                                         0.446408
3                                         0.563174
4                                         0.716573
...                                            ...
1495                                      0.413391
1496                                      0.603372
1497                                      0.295698
1498                                      0.270021
1499                                      0.421558


      Cerebellum region V {fan2010sexual  Cerebellum region VII␣
↪{fan2010sexual
0                                0.382551                                0.345238
1                                0.513751                                0.753752
2                                0.414815                                0.428085
3                                0.821899                                0.408632
```

```
4                          0.623266                          0.467282
...                                ...                                ⌴
 ↪...
1495                       0.224793                          0.420960
1496                       0.363925                          0.457728
1497                       0.369830                          0.361627
1498                       0.629090                          0.112054
1499                       0.259582                          0.439813

[1500 rows x 13 columns]
```

Bibliography

[1] Anatomy of the brain.

[2] Dan Avery. Nearly 1 in 10 teens identify as gender-diverse in pittsburgh study. *NBC Out*, May 2021.

[3] Mark G. Baxter and Paula L. Croxson. Facing the role of the amygdala in emotional information processing. *Proceedings of the National Academy of Sciences*, 109(52):21180–21181, 2012.

[4] Miriam Berger. 1 in 300 canadians 15 and up identify as transgender or nonbinary. *The Washington Post*, Apr 2022.

[5] Ross P Carne, Simon Vogrin, Lucas Litewka, and Mark J Cook. Cerebral cortex: an mri-based study of volume and variance with age and sex. *Journal of Clinical Neuroscience*, 13(1):60–72, 2006.

[6] Henry Vandyke Carter. *Anatomy of the Human Body*.

[7] Kelly P Cosgrove, Carolyn M Mazure, and Julie K Staley. Evolving knowledge of sex differences in brain structure, function, and chemistry. *Biological psychiatry*, 62(8):847–855, 2007.

[8] Lingzhong Fan, Yuchun Tang, Bo Sun, Gaolang Gong, Zhang J Chen, Xiangtao Lin, Taifei Yu, Zhenping Li, Alan C Evans, and Shuwei Liu. Sexual dimorphism and asymme-

try in human cerebellum: an mri-based morphometric study. *Brain research*, 1353:60–73, 2010.

[9] Melissa E Frederikse, Angela Lu, Elizabeth Aylward, Patrick Barta, and Godfrey Pearlson. Sex differences in the inferior parietal lobule. *Cerebral Cortex*, 9(8):896–901, 1999.

[10] Alicia Garcia-Falgueras, Carme Junque, Mónica Giménez, Xavier Caldú, Santiago Segovia, and Antonio Guillamon. Sex differences in the human olfactory system. *Brain Research*, 1116(1):103–111, 2006.

[11] Jill M Goldstein, Larry J Seidman, Nicholas J Horton, Nikos Makris, David N Kennedy, Verne S Caviness Jr, Stephen V Faraone, and Ming T Tsuang. Normal sexual dimorphism of the adult human brain assessed by in vivo magnetic resonance imaging. *Cerebral cortex*, 11(6):490–497, 2001.

[12] Michael Goodman, Noah Adams, Trevor Corneil, Baudewijntje Kreukels, Joz Motmans, and Eli Coleman. Size and distribution of transgender and gender nonconforming populations: a narrative review. *Endocrinology and Metabolism Clinics*, 48(2):303–321, 2019.

[13] Ruben C Gur, Faith Gunning-Dixon, Warren B Bilker, and Raquel E Gur. Sex differences in temporo-limbic and frontal brain volumes of healthy adults. *Cerebral cortex*, 12(9):998–1003, 2002.

[14] Ruben C Gur, Faith M Gunning-Dixon, Bruce I Turetsky, Warren B Bilker, and Raquel E Gur. Brain region and sex differences in age association with brain volume: a quantitative mri study of healthy young adults. *The American journal of geriatric psychiatry*, 10(1):72–80, 2002.

[15] Cade Hildreth. Gender spectrum: A scientist explains why gender isn't binary, Feb 2022.

[16] Janet Shibley Hyde, Rebecca S Bigler, Daphna Joel, Charlotte Chucky Tate, and Sari M van Anders. The future of sex and gender in psychology: Five challenges to the gender binary. *American Psychologist*, 74(2):171, 2019.

[17] Hengjun J Kim, Namkug Kim, Sehyun Kim, Seokjun Hong, Kyungmo Park, Sabina Lim, Jung-Mi Park, Byungjo Na, Younbyoung Chae, Jeongchan Lee, et al. Sex differences in amygdala subregions: evidence from subregional shape analysis. *Neuroimage*, 60(4):2054–2061, 2012.

[18] Tim Koscik, Dan O'Leary, David J Moser, Nancy C Andreasen, and Peg Nopoulos. Sex differences in parietal lobe morphology: relationship to mental rotation performance. *Brain and cognition*, 69(3):451–459, 2009.

[19] Morten L Kringelbach. The human orbitofrontal cortex: linking reward to hedonic experience. *Nature reviews neuroscience*, 6(9):691–702, 2005.

[20] Martin Lotze, Martin Domin, Florian H Gerlach, Christian Gaser, Eileen Lueders, Carsten O Schmidt, and Nicola Neumann. Novel findings from 2,838 adult brains on sex differences in gray matter brain volume. *Scientific reports*, 9(1):1–7, 2019.

[21] Esther L Meerwijk and Jae M Sevelius. Transgender population size in the united states: a meta-regression of population-based probability samples. *American journal of public health*, 107(2):e1–e8, 2017.

[22] Michael T. Murray and John Nowicki. 82 - ginkgo biloba (ginkgo tree). In Joseph E. Pizzorno and Michael T. Murray, editors, *Textbook of Natural Medicine (Fifth Edition)*, pages 620–628.e2. Churchill Livingstone, St. Louis (MO), fifth edition edition, 2020.

[23] Jonas Persson, R Nathan Spreng, Gary Turner, Agneta Herlitz, Arvid Morell, Eva Stening, Lars-Olof Wahlund, Johan Wikström, and Hedvig Söderlund. Sex differences in volume and structural covariance of the anterior and posterior hippocampus. *Neuroimage*, 99:215–225, 2014.

[24] Carl Wolfgang S Pintzka, Tor Ivar Hansen, Hallvard R Evensmoen, and Asta Kristine Håberg. Marked effects of intracranial volume correction methods on sex differences in neuroanatomical structures: a hunt mri study. *Frontiers in neuroscience*, 9:238, 2015.

[25] Miana Gabriela Pop, Carmen Crivii, and Iulian Opincariu. Anatomy and function of the hypothalamus. In *Hypothalamus in health and diseases*. IntechOpen, 2018.

[26] Rebecca Reavis and William H Overman. Adult sex differences on a decision-making task previously shown to depend on the orbital prefrontal cortex. *Behavioral neuroscience*, 115(1):196, 2001.

[27] Amber NV Ruigrok, Gholamreza Salimi-Khorshidi, Meng-Chuan Lai, Simon Baron-Cohen, Michael V Lombardo, Roger J Tait, and John Suckling. A meta-analysis of sex differences in human brain structure. *Neuroscience & Biobehavioral Reviews*, 39:34–50, 2014.

[28] Thomas E Schlaepfer, Gordon J Harris, Allen Y Tien, Luon Peng, Seong Lee, and Godfrey D Pearlson. Structural differences in the cerebral cortex of healthy female and male subjects: a magnetic resonance imaging study. *Psychiatry Research: Neuroimaging*, 61(3):129–135, 1995.

[29] B Locke Welborn, Xenophon Papademetris, Deidre L Reis, Nallakkandi Rajeevan, Suzanne M Bloise, and Jeremy R Gray. Variation in orbitofrontal cortex volume: relation to sex, emotion regulation and affect. *Social cognitive and affective neuroscience*, 4(4):328–339, 2009.

[30] Jordan E Wong, Jinyan Cao, David M Dorris, and John Meitzen. Genetic sex and the volumes of the caudate-putamen, nucleus accumbens core and shell: original data and a review. *Brain Structure and Function*, 221(8):4257–4267, 2016.