Regression Analysis with Right-Skewed Data:

Applications for Pre-Clinical Alzheimer's Disease

Mike Malek-Ahmadi, PhD

Bioinformatics Scientist

Banner Alzheimer's Institute

Phoenix, AZ, USA

Disclosures

- Signant Health

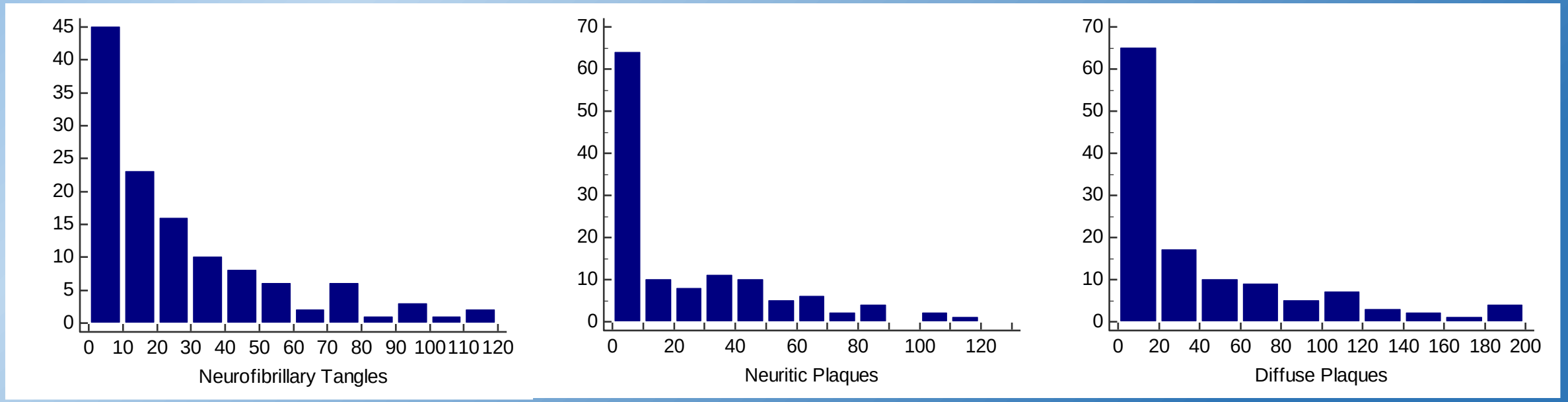
Acknowledgements

- Kewei Chen
- Yi Su
- Ji Luo
- Eric Reiman
- Elliott Mufson
- Sylvia Perez
- Gene Alexander
- Rush Religious Orders Study
- ADNI

Introduction

- It is known that linear regression assumes that the error of the outcome variable follows a normal distribution.
- However, there are many cases where the data are heavily skewed, particularly toward the lower values for a variable.
- This is relevant to pre-clinical AD where lower values on a particular scale or measurement are likely to have higher frequencies.

Right-Skewed Distributions



Data Transformations and Dichotomization

- Log, natural log, and square root are the most common.
- Transformations do not necessarily yield normal distributions, must still perform normality tests to see if transformation was successful.
- Interpretation of transformed values can be difficult. Need to be able to interpret values on their original scale.

- Loss of statistical power
- Clinical/scientific meaningfulness of the cutpoint
- Dichotomization of continuous variables may only be valid if a clinically validated cutpoint is used (e.g., HbA1c > 7.0; AV-45 SUVR > 1.18).

Poisson Regression

- Used to characterize count data where lower values of a variable have the highest frequency.
- Often used to describe the occurrence of events (e.g., number of arrests, number of goals scored in the World Cup).
- In Poisson regression models, one of the major assumptions is that the mean and variance of the outcome variable are equal.
- What happens if variance is high?

Negative Binomial Regression

- The negative binomial (NB) model is similar to the Poisson model, but incorporates an additional term to account for the excess variance.
- Like the Poisson model, the NB model can be used to characterize count data (integers) where the majority of data points are clustered toward lower values of a variable.
- However, the NB model can be used when variance is substantially higher than the mean.

Background of Case Study #1

- In neuropsychology, the term dispersion refers to the degree of variation in performance between different cognitive domains for an individual.

- Previous studies have found that individuals with higher cognitive domain dispersion are more likely to develop Alzheimer's disease (AD)*.

- No studies linking cognitive dispersion to pathological findings of AD.

*Kalin et al, *Frontiers in Aging and Neuroscience* 2014;6:147.
*Vaughan et al, *Current Gerontology and Geriatrics Research* 2013;495793.

Journal of Alzheimer's Disease 58 (2017) 575–583
DOI 10.3233/JAD-161233
IOS Press

Cognitive Domain Dispersion Association with Alzheimer's Disease Pathology

Michael Malek-Ahmadi[a], Sophie Lu[b], YanYan Chan[c], Sylvia E. Perez[d], Kewei Chen[a] and Elliott J. Mufson[d,*]

[a]*Banner Alzheimer's Institute, Phoenix, AZ, USA*
[b]*Williams College, Williamstown, MA, USA*
[c]*Arizona State University, Tempe, AZ, USA*
[d]*Department of Neurobiology and Neurology, Barrow Neurological Institute, Phoenix, AZ, USA*

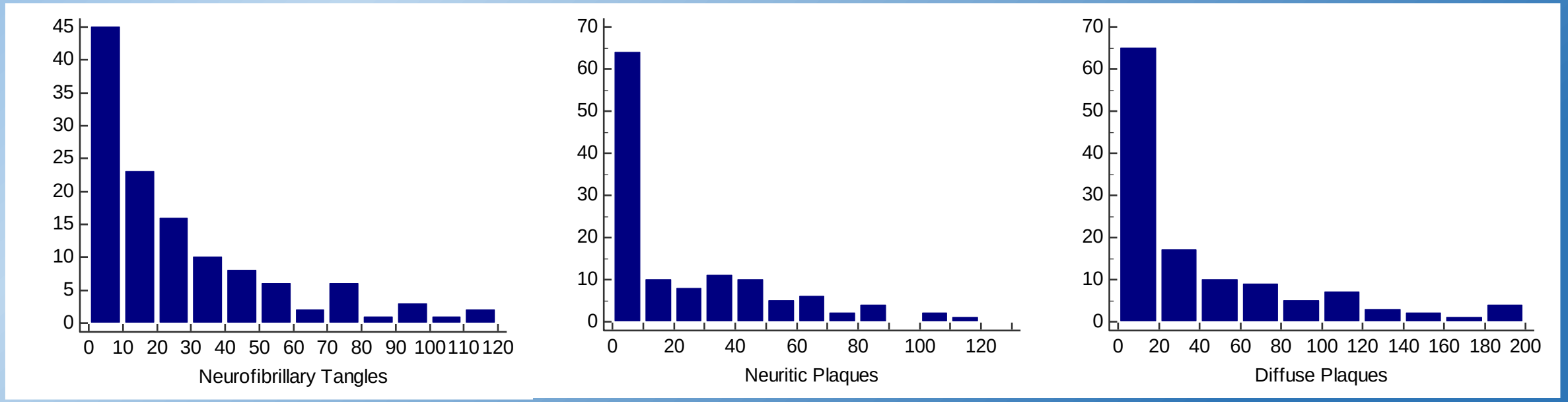
Case Study #1

- Rush Religious Order Study
- Data from 123 cognitively unimpaired (CU) older adults who had cognitive and neuropathological data.
- Five cognitive domains: Episodic Memory, Semantic Memory, Working Memory, Perceptual Speed, Visuospatial
- Three measures of AD pathology: neurofibrillary tangles (NFTs), diffuse plaques (DPs), and neuritic plaques (NPs). Each measure represents a summation of counts from 5 cortical regions.
- Research question: Is within-subject cognitive domain dispersion associated with AD pathology?

Methodology

- Used the Intraindividual Standard Deviation (ISD) as the measure of dispersion.
- ISD = standard deviation of cognitive domain z-scores.
- Used Poisson and Negative Binomial regression models to assess the association between ISD and measures of AD pathology (NP, DP, NFT) while adjusting age at death, sex, education, and APOE e4 carrier status.

Distribution of Neuropathology Measures



Descriptive Statistics for Neuropathology Measures

	Mean	**Variance**	**Std. Dev.**
NFTs	26.36	724.69	26.92
NPs	21.93	480.92	27.98
DPs	38.11	2513.02	50.13

Results from Poisson and NB Models

Poisson	**Coefficient**	**Std. Error**	**Residual Deviance**	**p-value**
ISD and NFTs	-0.23	0.12	2374.3	0.04
ISD and NPs	1.10	0.12	3781.2	<0.001
ISD and DPs	0.34	0.10	6574.4	<0.001

Negative Binomial	**Coefficient**	**Std. Error**	**Residual Deviance**	**p-value**
ISD and NFTs	-0.23	0.53	135.83	0.66
ISD and NPs	1.06	1.08	130.79	0.33
ISD and DPs	0.44	1.44	133.07	0.70

All models adjusted for age, sex, education, and APOE e4 status; df = 116

Which Model is Correct?

- Model fit can be assessed using the residual deviance values and the df in the Chi-square distribution.
- For all models, df = 116 with an expected Chi-square value 142.14.
- Residual deviance values lower than 142.14 indicate a good fit while higher values indicate a lack of fit.

Results from Poisson and NB Models

	Poisson Residual Deviance	**NB Residual Deviance**
ISD and NFTs	2374.3	135.83
ISD and NPs	3781.2	130.79
ISD and DPs	6574.4	133.07

Chi-square critical value (df 116) = 142.14.

Additional Model Fit Methods

- In R, residual deviance value can be treated as a chi-square value. Can obtain a p-value, given the degrees of freedom.

1 - pchisq (residual deviance, df)

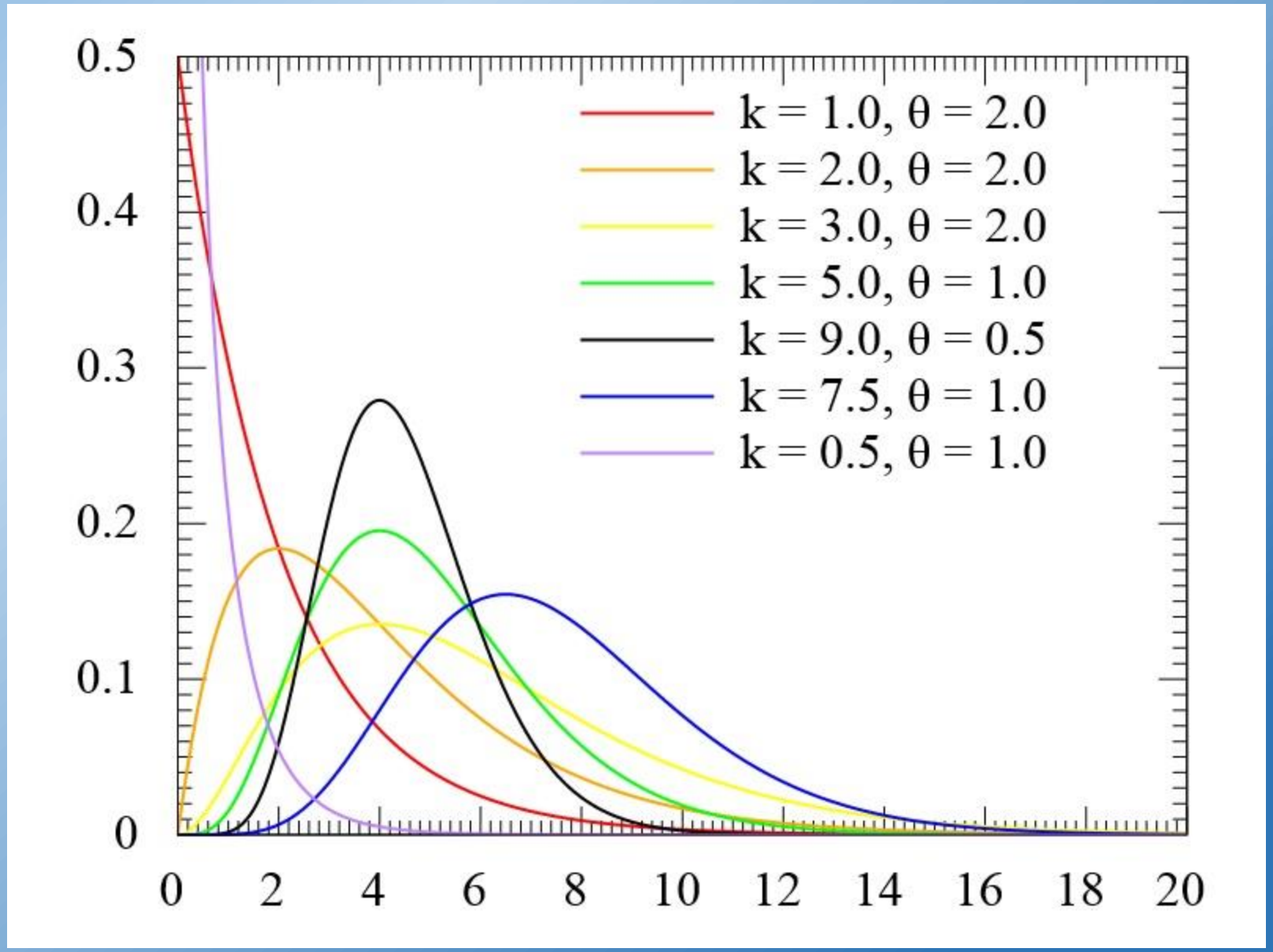
```
➢1-pchisq(135.83, 116)
➢[1] 0.1007073
```

- Additional method of Poisson and NB model fitting proposes that when the ratio of residual deviance to the df for a model is equal to or approximately 1.00 then the model fit is acceptable.*

*Allison PD, Waterman RP (2002) Fixed effects negative binomial regression models. *Sociol Methodol* **32**, 247-265.

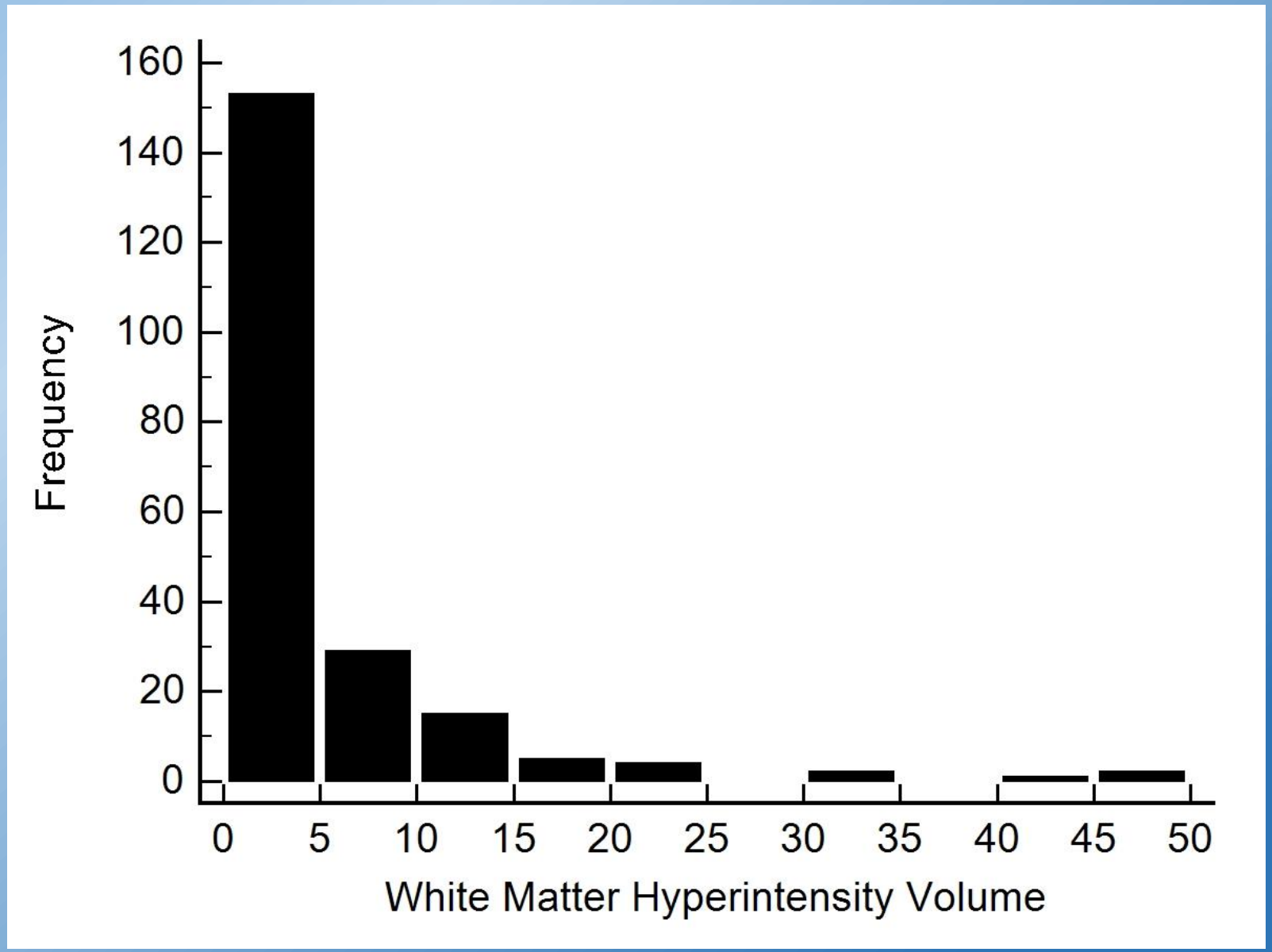
The Gamma Distribution

- What if the outcome variable is not an integer?
- Under the generalized linear model (GLM) family, the gamma distribution can be specified.
- However, the dependent variable cannot have zeroes or negative numbers.


0.5
0.4
0.3
0.2
0.1
0
0
2
4
6
8
10
12
14
16
18
20
k = 1.0, θ = 2.0
k = 2.0, θ = 2.0
k = 3.0, θ = 2.0
k = 5.0, θ = 1.0
k = 9.0, θ = 0.5
k = 7.5, θ = 1.0
k = 0.5, θ = 1.0

Case Study #2 - Gamma GLM Example - ADNI

- White matter hyperintensity volume (WMHV) - FLAIR
- 297 CU ADNI Cases
- Research question – Is AV-45 SUVR associated with greater WMHV?


Frequency
160
140
120
100
80
60
40
20
0
0
5
10
15
20
25
30
35
40
45
50
White Matter Hyperintensity Volume

Case Study #2 - Gamma GLM Example - ADNI

- WMHV was the dependent variable with AV-45 as the independent variable.
- Covariates included age, sex, education, Hachinski Ischemic Scale (vascular risk), and APOE e4 carrier status.
- model<-glm(wmhv~age+sex+educ+apoe+hachinski+suvr, family=Gamma(link = "log"), data=dataset)

Gamma GLM Example - ADNI

```
Coefficients:
                          Estimate Std. Error t value Pr(>|t|)
(Intercept)               -5.39831    1.31487  -4.106 5.25e-05 ***
age                        0.06416    0.01414   4.537 8.35e-06 ***
gender                     0.02873    0.18049   0.159  0.87363
educ                       0.02692    0.03446   0.781  0.43533
apoe_carrierNon-Carrier    0.04765    0.19849   0.240  0.81047
hachinski                  0.21840    0.13028   1.676  0.09474 .
suvr                       1.36929    0.52107   2.628  0.00905 **
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for Gamma family taken to be 2.134672)

    Null deviance: 406.28  on 296  degrees of freedom
Residual deviance: 328.41  on 290  degrees of freedom
AIC: 1481
```

```
> 1-pchisq(328.41, 290)
[1] 0.05975271
> 
```

Discussion

- When data are right-skewed, Poisson, NB, and Gamma regression models can be used.

- No need to utilize transformations which can make interpretation problematic. In addition, data may still not be normally distributed after a transformation is applied.

- These regression methods may be especially helpful in pre-clinical AD studies where the distribution of imaging and biomarker data are likely to be right-skewed.

Questions?