15/11/2018 meeting

Nitay Alon

November 16, 2018

1 UK Biobank data

Relevant data

The relevant data for our analysis is found the following table:

Data set name	Relevant Columns	Comments
FA	44-91	Mean
FA	140-166	
MD	92-139	Mean
MD	176-193	
Volume	194-332	

Noting that some features are normalized to brain size while other aren't. The diagnostic columns are not part of the analysis for now.

Data processing

We've decided on the following data processing schema:

1. Remove missing data
2. Histogram of age for the remaining data
3. Histogram of race for the remaining data
4. KS test for log-normal /normal data
5. Apply log transformation over the data, add the distance between 0 and the closest value $x_{(2)}$ if needed
6. Normalize the data (standardization)

Data analysis

To test our main hypothesis we apply the following routine:

1. Test the null hypothesis of single human distribution vs the composite mixture model
2. If the null hypothesis is rejected test the null hypothesis of pure types vs the composite mixture model
3. After the analysis is done, repeat the process with normalized for brain size data

Enable testing over race - that is, repeat the process for each race separately.

Fitting the log normal model

Following Isaco's comment on the relation between the variance and the distance between the minimum and maximum observation of log normal data: If

$$\mathbf{X} = \exp \mu + z\sigma \implies \sigma^2 = log(\frac{E(X^2)}{E(X)^2})$$

and

$$\frac{max(X)}{min(X)} \sim \exp \sigma\sqrt{8n}$$

now, if the distance between $|mu_1 - mu_2| < \frac{\sigma}{2}$ then the total std exceeded the within std by now more than 3%. The proof is done using the binomial distribution of the between variable - $p(1-p)(\mu_1 - \mu_2)^2 < \frac{\sigma}{16}$ Apply the following procedure over the data to determine if we need to shift the distribution prior to log transformation

1. Select feature with relative high minimum observation
2. Apply log transformation
3. Plot histogram
4. Remove $\frac{9}{10}$ of the distance to zero
5. Repeat 2-3
6. See if the transformed distribution yield "normaler" distribution

Reporting

For each brain feature we report the following:

1. sd of the population, men and women (as part of testing the single distribution hypothesis) and as validation for the log transformation
2. *T-test* and Cohen's-d for mean distance
3. After the EM is done - the mixture parameters
4. Plot the empirical histogram and the densities

In the end report (web app) enable the user to select either log normal or normal distribution for analysis.