1 Sum of Bernoulli Distributions

Here are some basic statistical concepts needed to perform statistical analysis on binary images.

Given a sample space $S$, a random variable $X$ is a rule that assigns a number to each element of $S$. $X$ is a function that maps element $s \in S$ to a real number, i.e.

$$X : S \rightarrow \mathbb{R}.$$ 

Bernoulli random variable $X$ takes values only 0 or 1. The probability distribution $P$ of a random variable $X$ is a function defined for every number $x$ such that

$$P(X = x) = P(s \in S : X(s) = x).$$

The cumulative distribution function (cdf) of $X$ is defined as

$$F(x) = P(X \leq x) = \sum_{y \leq x} P(X = y).$$

The expected value (expectation) of $X$ is a linear operator on $X$ defined as

$$\mathbb{E}X = \sum_{x} xP(X = x)$$

The variance of $X$ is defined as

$$\nabla X = \mathbb{E}(X - \mathbb{E}X)^2$$

$$= \mathbb{E}X^2 - (\mathbb{E}X)^2.$$
Figure 1: The distribution of Binomial random variable with $n = 10$ and $p = 0.5$, and its cumulative distribution function. Plots are generated using the statistical package R.

$X$ is a Bernoulli random variable with parameter $p$ if

$$P(X = 1) = p, \ P(X = 0) = 1 - p$$

and it will be denoted as $X \sim Bernoulli(p)$. The distribution can be compactly written as

$$P(X = x) = (1 - p)^{1-x} p^x.$$  

The expectation and variance of the Bernoulli random variable are trivially

$$\mathbb{E}X = p, \ \forall X = p(1 - p).$$

Consider a Bernoulli experiment that give only two outcomes $H$ or $T$. This can be considered as coin tossing with either obtaining head ($H$) or tail ($T$). Let

$$P(H) = p, \ P(T) = 1 - p.$$  

Now consider $n$ such independent experiments. This is called a Binomial experiment. Let $X$ be the number of total outcomes of $H$. This is called the Binomial random variable with parameter $n$ and $p$ and denoted as

$$X \sim Binomial(n, p).$$
Then obviously
\[ X = X_1 + X_2 + \cdots + X_n, \]
where \( X_i \) is the number of outcome \( H \) in the \( i \)-th experiment. The distribution of \( X \) is then given by
\[ P(X = x) = \binom{n}{x} p^x (1 - p)^{n-x}. \]

The expectation and the variance of the Binomial distribution can be computed from Bernoulli distributions.

\[ \mathbb{E}X = \mathbb{E}X_1 + \cdots + \mathbb{E}X_n = np \]
\[ \mathbb{V}X = \mathbb{V}X_1 + \cdots + \mathbb{V}X_n = np(1 - p). \]

The actual distribution and its cumulative distribution function can be displayed using \( \texttt{R} \) statistical package:

> plot(dbinom(0:10,10,0.5),type="s")
> plot(pbinom(0:10,10,0.5),type="s")

From the central limit theorem, it can be shown that
\[ \frac{X - \mathbb{E}}{\sqrt{\mathbb{V}X}} = \frac{X - np}{\sqrt{np(1 - p)}} \rightarrow N(0, 1), \]
the standard normal distribution.

## 2 Inference on Proportion of Activation

We develop a method for performing statistical tests on binary images using the Binomial random variable approximation to the standard normal distribution. The binary images are obtained from Dong-Eog Kim of Dongguk University Hospital in Korea as a part of image-based stroke registry. See [3] for details on the image. The diffusion-weighted images (DWI) were obtained for stroke patients and stroke lesions were segmented manually and aligned on a MR template (Figure 2). In DWI, each voxel has an image intensity that reflects a single best measurement of the rate of water diffusion at that location. DTI can be used to identify severely ischemic brain regions after stroke onset than more traditional T1 or T2 MRI [4].

The first four columns of binary images are lesions of stroke patients with dysphasia who got better (\( n_1 = 58 \)). The next four columns are lesions of stroke patients with dysphasia who did not get better (\( n_2 = 23 \)). We are interested in determining if there is any pixel showing group difference. This can be done by testing on the equality of proportions.
Figure 2: The first four columns of binary images are lesions of stroke patients with dysphasia who got better \((n_1 = 58)\). The next four columns are lesions of stroke patients with dysphasia who did not get better \((n_2 = 23)\). We are interested in determining if there is any pixel showing group difference.

### 2.1 One Sample Test

For a collection of binary images, each image has regions of activation. The activated and nonactivated regions are encoded as 1 and 0 respectively, which can be modeled as a Bernoulli distribution. At each fixed voxel, we are interested in testing if the activation is statistically significant. Suppose there are \(n\) binary images and \(p\) proportion of of them are activated at the fixed voxel. Unfortunately, \(p\) is the population level parameter that is unknown. It can be estimated as

\[
\hat{p} = \frac{\# \text{ of activated images}}{n}.
\]

This is referred as the sample proportion and it is a random variable since the total number of activated images is a Binomial distribution \(\text{Binomial}(n, p)\). Then trivially

\[
\mathbb{E}[\hat{p}] = p, \quad \mathbb{V}[\hat{p}] = \frac{p(1-p)}{n}.
\]

We are interested in testing if

\[H_0 : p = p_0 \text{ vs. } H_1 : p > p_0\]

for a specified constant proportion \(p_0\). The often used test statistic is

\[
Z_n = \frac{\hat{p} - p_0 - \mathbb{E}(\hat{p} - p_0)}{\sqrt{\mathbb{V}(\hat{p} - p_0)}}.
\]
Under $H_0$, it becomes

$$Z_n = \frac{\hat{p} - p_0}{\sqrt{p_0(1 - p_0)/n}}.$$  

This is asymptotically $Z_n \to N(0,1)$ for large $n$. If the test statistic is large, it is likely that $H_0$ is not true. So the deviation of $Z_n$ from 0 corresponds to the likelihood of rejecting $H_0$. This is quantitatively determined using the $p$-value:

$$p\text{-value} = P\left( Z > \frac{\hat{p} - p_0}{\sqrt{p_0(1 - p_0)/n}} \right),$$

where $Z \sim N(0,1)$ and $\hat{p}$ is the observed proportion here. For sufficiently large sample size, i.e. $n > 30$, the $p$-value computation is accurate enough.

**Example.** Suppose 55 out of 102 images are activated in a particular voxel. Can you conclude that more than half of images are activated? Note that

$$\hat{p} = \frac{55}{102} = 0.54$$

$$\frac{\hat{p} - 0.5}{\sqrt{0.5 \cdot 0.5/102}} = 0.79$$

Note that $P(Z > 0.79)$ can be computed as

$$1 - \text{normcdf}(0.79) = 0.21$$

in MATLAB. This is not small enough so we cannot conclude more than half of images are activated.

In order for the test procedure to be valid, $p_0$ has to be strictly between 0 and 1. If $p_0$ is 0 or 1, the test statistic breaks down and we can no longer able to compute the $p$-value using the formula (1). Note that Bernoulli models always assume $0 < p < 1$ so testing if the parameter is 0 or 1 is a valid procedure. The $p$-value computation for $H_0 : p = 0$ or $H_0 : p = 1$ is trivial. If all pixels are zero, $p$-value is trivially 0 while at least one pixel is not zero then $p$-value is one. However, it is not useful to have the test procedure to produce only 0 or 1 $p$-value. This problem can be remedied if we do not use the Bernoulli model and model the population proportion to follow Gaussian. So we assume the population proportion $\hat{p}$ to follow

$$\hat{p} \sim N(\mu, \sigma^2)$$

and test for

$$H_0 : \mu = 0 \text{ vs. } H_1 : \mu > 0.$$

5
The test statistic is

Then the p-value is given by

\[ p\text{-value} = P\left( Z > \frac{\hat{p}}{\sqrt{p_0(1-p_0)/n}} \right). \]

This problem can be easily solved if we follow the VBM framework which is developed for handling exactly this type of problems [2]. In VBM, we tradition-ally smooth binary images with Gaussian kernel and model the smoothed binary images as Gaussian getting rid of the need for modeling it as Bernoulli. It is often suggested to apply the logit tranform on the smoothed binary images [1] but once you perform Guassian kernel smoothing, the images become sufficiently Gaussian so the logit transfrom is not really needed.

### 2.2 Two Sample Test

This type of test procedure can be extended in testing if there is difference between the two population proportions. We will assume there are \( p_i \) proportion of images activated out of \( n_i \) images in the \( i \)-th group. The total number of activated images are modeled as \( \text{Binomial}(n_i, p_i) \). The null and alternate hypotheses for the two-sample test is then

\[ H_0 : p_1 = p_2 \text{ vs. } H_1 : p_1 > p_2. \]
Note that

$$\mathbb{E}(\hat{p}_1 - \hat{p}_2) = \frac{p_1}{n_1} - \frac{p_2}{n_2}.$$ 

$$\nabla \hat{p} = \frac{p_1(1 - p_1)}{n_1} + \frac{p_2(1 - p_2)}{n_2}$$

Under null, the test statistic is similarly given by the $Z$-statistic as

$$Z_n = \frac{\hat{p}_1 - \hat{p}_2}{\sqrt{\frac{p_1(1-p_1)}{n_1} + \frac{p_2(1-p_2)}{n_2}}}.$$ 

The $p$-value is then given by

$$p\text{-value} = P(Z > \frac{\hat{p}_1 - \hat{p}_2}{\sqrt{\frac{p_1(1-p_1)}{n_1} + \frac{p_2(1-p_2)}{n_2}}} ),$$

where $p_1$ and $p_2$ are observed proportions. There are bunch of many different test procedures for testing the equality of proportions. This is probably the most simple procedure among them.

### 3 MATLAB Implementation

Images are stored in the bitmap format and files have different length of file names. The first and the second group images are respectively stored in the directories `/stroke/0` and `/stroke/1`. We read the sequence of file names into structured arrays $g_0$ and $g_1$. The first file name in the first group is then given by $g_0(1).name$. The following code read all the bitmap images in the directory.

```matlab
% MATLAB code for reading bitmap images from directories

g0 = dir('*.bmp');
n0 = size(g0,1); % number of images in group 0

% Initialize groups

for i = 1:n0
    fname = g0(i).name;
    f = imread(fname,'bmp');
    group0(:,:,i) = double(f(:,:,1)/255);
end;
```
Figure 4: Two sample test result showing almost no region of group difference although there seems like a region of group difference where $Z$-values are larger than 3. This high $Z$-statistic value is most likely to be caused by the discretization error.

```matlab
g1 = dir('*.*');
n1 = size(g1,1); % number of images in group 1
group1 = zeros(370,301, n1);
for i = 1:n1
    fname = g0(i).name;
    f = imread(fname,'bmp');
    group1(:,:,i) = double(f(:,:,1)/255);
end;
```

$n0$ and $n1$ are the sample sizes. Each image slice of dimension $370 \times 301$ is loaded into the 3D matrices `group0` and `group1` for the two groups. The
sample proportions are computed as
p0=mean(group0, 3);
p1=mean(group1,3);

The $Z$-statistic for one sample is computed as
$$Z\text{value} = \frac{(p0-pc)}{\sqrt{pc.*(1-pc)/nc}}$$
for some specific value $pc$. On the other hand, the $Z$-statistic for the two samples
is computed as
$$Z\text{value} = \frac{(p0-p1)}{\sqrt{p0.*(1-p0)/n0 + p1.*(1-p1)}}.$$  

The $p$-value is then computed using `normcdf`. The whole routine is written as
a function below. If the second argument is empty, i.e. [], the function performs
the one sample test.

```
[pvalue Zvalue] = stat_proportion(group0, []);
[pvalue Zvalue] = stat_proportion(group0, group1);
```

The resulting outputs are Figure 3 for the one sample test and Figure 4 for the two
sample test. One sample test result shows many regions of statistical significance
while the two sample test result shows almost no region of significance at 0.05.

References


matter concentration in autism: 2D voxel-based morphometry. *NeuroImage*,

Jeong, J.S. Lee, and Choi H.-K. A new image-based stroke registry containing
quantitative magnetic resonance imaging data. *Cerebrovascular Diseases*,

U. Mödder, and H.J. Freund. Diffusion-and perfusion-weighted mri: the