# Distributions and Inference for Categorical Data

## Categorical Response Data

A `categorical variable` has a measurement scale consisting of a number of categories. For example, breast cancer diagnosis based on a mammogram can be in one of these categories: normal, benign, probably benign, suspicious and malignant.

A `categorical response variable` is a `response` or a `dependent variable` that has categorical measurement scale. For example, if we are interested in studying the relationship between breast cancer diagnosis and a number of risk factors such age, race, ethnicity, and family history, then the diagnosis of breast cancer is a categorical response variable and the set of risk factors are called the explanatory or independent variables.

Categorical response variables can be binary: with two categories, nominal: with more than two categories without any natural ordering, and ordinal: with more than two ordered categories. An example of a binary response variable is the presence or absence of certain disease. An example of a nominal response variable is the types of diseases and conditions that can affect the heart: angina, arrhythmia, congenital heart disease, coronary artery disease, heart attack, etc. Finally, an example of an ordinal response variable is patient conditions: good, fair, serious, and critical.

## Distributions for Categorical Response Variables

In order to make inference or prediction based on health data at hand, it is required to make assumptions about the random mechanism that generated the data. There are two key distributions for binary and nominal or oridnal categorical response variables: i) binomial and ii) multinomial distributions respectively.

### Binomial Distribution

Many health applications refer to a fixed number $n$ of binary observations, $y_1, y_2, \ldots, y_n$, such that $P(Y_i = 1) = \pi$ and $P(Y_i = 0) = 1 - \pi$. The outcome 1 is referred to as success and the outcome 0 is referred as failure. The total number of successes $Y = \sum_{i=1}^{n} Y_i$ has a binomial distribution with probability mass function (pmf):

$$p(y) = \binom{n}{y} \pi^y (1-\pi)^{n-y}, \quad y = 0, 1, 2, \ldots, n$$

Note that the mean and variance of $Y$ are $E(Y) = n\pi$ and $Var(Y) = n\pi(1-\pi)$ respectively.

### Multinomial Distribution

Nominal and ordinal response variables have more than two possible outcomes, and each of $n$ independent and identical trials can have outcome in $1, 2, \ldots, c$ categories. Let $n_j$ denote the number of trials having outcome in category $j$. Then the counts $n_1, n_2, \ldots, n_c$ have the multinomial distribution with pmf:

$$p(n_1, n_2, \ldots, n_c) = \left(\frac{n!}{n_1! n_2! \ldots n_c!}\right) \pi_1^{n_1} \pi_2^{n_2} \ldots \pi_c^{n_c}.$$

The pmf is $c-1$ dimensional since $\sum_j n_j = n$ with $n_c = n - (n_1 + \ldots n_{c-1})$.

For the multinomial distribution,

$$E(n_j) = n\pi_j, \quad Var(n_j) = n\pi_j(1-\pi_j), \quad Cov(n_j, n_k) = -n\pi_j\pi_k$$

## NHANES data

To demonstrate analysis of binary, nominal and ordinal response variables we will consider the `NHANES` data from the R package `NHANES`. First create an analysis dataset by selecting all the variables of interest.

```
library(NHANES)
library(tidyverse)
```

```
## -- Attaching packages ------------------------------------------------ tidyverse 1.3.0 --
## v ggplot2 3.3.3     v purrr   0.3.3
## v tibble  2.1.3     v dplyr   0.8.4
## v tidyr   1.0.2     v stringr 1.4.0
## v readr   1.3.1     v forcats 0.4.0
## -- Conflicts --------------------------------------------------- tidyverse_conflicts() --
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()    masks stats::lag()
```

```
#as_tibble(NHANES)
setwd("~/Box/MyDocs/Teaching/Spring/2021/DSCI 610/LectureMaterials/Week 9/Lecture")

#df_eda <- select(NHANES, ID, SurveyYr, Gender, Age, Race1, Poverty, HomeOwn, #Weight, Height, HeadCirc

#saveRDS(df_eda, file="analysis.rds")
```

Next we create a smaller analysis data set `df_CD` with all categorical variables.

```
df_analysis <-readRDS("analysis.rds")

df_CD<-select(df_analysis, Gender, Race1,  HomeOwn, Diabetes, SmokeNow, HealthGen,Depressed, Marijuana,

#df_CD_nomis <- filter(df_CD, Diabetes !="NA", SmokeNow !="NA", HealthGen #!="NA",Depressed !="NA", Mar
```

To summairize these variables we can tabulate their categroies as follows:

```
summary(df_CD)
```

```
##     Gender          Race1       HomeOwn     Diabetes    SmokeNow
##  female:5020   Black   :1197   Own  :6425   No  :9098   No  :1745
##  male  :4980   Hispanic: 610   Rent :3287   Yes : 760   Yes :1466
```

```
##                Mexican :1015   Other: 225   NA's: 142   NA's:6789
##                White   :6372   NA's :  63
##                Other   : 806
##
##      HealthGen      Depressed    Marijuana    PregnantNow
##  Excellent: 878   None   :5246   No  :2049   Yes    :  72
##  Vgood    :2508   Several:1009   Yes :2892   No     :1573
##  Good     :2956   Most   : 418   NA's:5059   Unknown:  51
##  Fair     :1010   NA's   :3327               NA's   :8304
##  Poor     : 187
##  NA's     :2461
```

Note that `Gender`, `Diabetes`, `SmokeNow`, `Marijuana`, and `PregnantNow` are binary variables with two categories. `Race1` and `HomeOwn` are nominal variables with five and three categories respectively. Finally, `HealthGen` and `Depressed` are ordinal variables with five and three categories respectively.

Depending on your research question, `Diabetes`, `HealthGen`, and `Depressed` may be regarded as dependent variables and rest of the variables in the `analysis` data set may be considered as explanatory or independent variables.

## Statistical Inference for Binomial Parameter

Note the binomial parameter is $\pi$, the probability of success. Statistical inference for $\pi$ is made in terms of tests and confidence intervals. Both of these depend on an optimum estimator of $\pi$, which is the maximum likelihood estimator (MLE) $\hat{\pi} = \frac{y}{n}$. Here $y$ is the number of successes in $n$ independent trials.

$$E(\hat{\pi}) = E(\frac{Y}{n}) = \frac{E(Y)}{n} = \pi, \quad Var(\hat{\pi}) = \frac{\pi(1-\pi)}{n}.$$

### Test about $\pi$

A typical test on $\pi$ is as follows:

$$H_0 : \pi = \pi_0.$$

The Wald statistic for testing $H_0 : \pi = \pi_0$ is:

$$z_W = \frac{\hat{\pi} - \pi_0}{S.E.(\hat{\pi})} = \frac{\hat{\pi} - \pi_0}{\sqrt{\hat{\pi}(1-\hat{\pi})/n}}.$$

The score statistic for testing $H_0 : \pi = \pi_0$ can be written as:

$$z_S = \frac{\hat{\pi} - \pi_0}{\sqrt{\pi_0(1-\pi_0)/n}}.$$

### Confidence Interval for $\pi$

A confidence interval for $\pi_0$ can be obtained by inverting the Wald test statistic for which $z_W < z_{\alpha/2}$ as:

$$\hat{\pi} \pm z_{\alpha/2}\sqrt{\frac{\hat{\pi}(1-\hat{\pi})}{n}},$$

where $\alpha$ is the smallest Type-I error which is also known as the level of significance.

### Example 1 : Inference about $\pi$

Let us consider the hypothesis test on the proportion of diabetic individuals in the NHANES data. We are interested to test if 10% of the participants are diabetic. Then the null and alternative hypotheses can be set up as:

$$H_0 : \pi = 0.1, \quad \text{versus} \quad H_a : \pi \neq 0.1$$

The function `prop.test` gives the score test and score confidence interval for $\pi$.

```
df_CD1 <- filter(df_CD, Diabetes !="NA")
summary(df_CD1$Diabetes)
```

```
##   No  Yes
## 9098  760
```

```
prop.test(760, 9858, p=.1, correct=FALSE)
```

```
##
##  1-sample proportions test without continuity correction
##
## data:  760 out of 9858, null probability 0.1
## X-squared = 57.467, df = 1, p-value = 3.437e-14
## alternative hypothesis: true p is not equal to 0.1
## 95 percent confidence interval:
##  0.07199237 0.08252659
## sample estimates:
##          p
## 0.07709475
```

Note the p-value of the test is essentially zero, the confidence interval for $\pi$ does not include 0.1, and lies entirely to the left of the null hypothesis value 0.1. Thus, there is strong evidence against the null hypothesis from the sample data that leads us to reject $H_0$. In addition, based on the confidence interval of $\pi$, we can safely infer that the actual proportion of diabetic individuals is less than 10%.

The proportion package contains a great variety of confidence intervals for a binomial parameter $\pi$, including Wald, likelihood-ratio, and score intervals. The likelihood-ratio-based confidence interval is more complex analytically and its algebraic formulation is omitted here.

```
library(proportion)
ciAllx(760,9858,0.05)
```

```
##       method   x LowerLimit UpperLimit LowerAbb UpperAbb ZWI
## 1       Wald 760 0.07182918 0.08236031       NO       NO  NO
## 2    ArcSine 760 0.07191192 0.08244236       NO       NO  NO
## 3 Likelihood 760 0.99986779 0.99995914       NO       NO  NO
## 4      Score 760 0.07199237 0.08252659       NO       NO  NO
## 5 Logit-Wald 760 0.07199124 0.08252787       NO       NO  NO
## 6     Wald-T 760 0.07182593 0.08236356       NO       NO  NO
```

Note the confidence intervals from all five methods except the likelihood-based method are pretty similar. When the success rate is less than 20% the likelihood-based method does not work.

## Statistical Inference for Multinomial Parameters

The MLEs of the multinomial parameters $\{\pi_j\}, j = 1, 2, \ldots c$ are as follows:

$$\hat{\pi}_j = \frac{n_j}{n}, \;\; j = 1, 2, \ldots, c.$$

### Test about $\pi_j$

The hypotheses of interest are as follows:

$$H_0 : \pi_j = \pi_{0j} \;\; j = 1, 2, \ldots, c,$$

where $\sum_j \pi_{0j} = 1$. When $H_0$ is true, the expected values of $n_j$, the expected frequencies are $\mu_j = n\pi_{0j}, j = 1, 2, \ldots, c$. The test is known as the goodness-of-fit of a set of multinomial probabilities, and is based on what is known as the Pearson $\chi^2$ statistic defined as follows:

$$\chi^2 = \sum_{j=1}^{c} \frac{(n_j - \mu_j)^2}{\mu_j}$$

Greater the differences $|n_j - \mu_j|$, greater the value for $\chi^2$. For large samples, $\chi^2$ has approximately a chi-squared distribution with degrees of freedom $df = c - 1$. If we denote $\chi^2_o$ as the observed value of $\chi^2$, then the p-value for the test is approximated by $P(\chi^2_{c-1} \geq \chi^2_o)$, where $\chi^2_{c-1}$ denotes a chi-squared random variable with $df = c - 1$.

### Example 2: Goodness-of-fit test for a set of multinomial proportions

Let us assume that the actual proportion of US population with general health condition is as follows. 10% of the population have excellent health, 35% have very good health, 40% have good health, 12% have fair and 3% have poor health respectively. We have observed proportions from the NHANES data as follows:

```
df_CD2 <- filter(df_CD, HealthGen !="NA")
summary(df_CD2$HealthGen)
```

```
## Excellent     Vgood      Good      Fair      Poor
##       878      2508      2956      1010       187
```

The function `chisq.test` can perform the Pearson chi-squared test of goodness-of-fit of a set of multinomial probabilities.

```
chisq.test(table(df_CD2$HealthGen), p=c(.1,.35,.4,.12,.03))
```

```
##
##  Chi-squared test for given probabilities
##
## data:  table(df_CD2$HealthGen)
## X-squared = 47.12, df = 4, p-value = 1.44e-09
```

The observed $\chi_o^2 = 47.653$ and the p-value is close to zero. Thus, there is strong evidence against the null hypothesis based on sample data and we reject $H_0$.

## References


1. Chapter 1: Introduction: Distributions and Inference for Categorical Data. Alan Agresti (2013). Categorical Data Analysis, John Wiley and Sons.