4.1 Exact Binomial Test and Confidence Interval for Prevalence Proportion

Suppose that in a population of size N, X have contracted a certain disease. We defined the prevalence proportion of the disease as $p = X/N$. Suppose now that in a sample of size n, the disease is recorded for x individuals. The point estimator for p is $\hat{p} = x/n$. If we want to test $H_0 : p = p_0$ vs $H_1 : p \geq p_0$ (upper-tailed) or $p \leq p_0$ (lower-tailed) or $p \neq p_0$ (two-tailed), we can conduct an approximate z-test with the test statistic $z = \dfrac{\hat{p} - p_0}{\sqrt{p_0(1-p_0)/n}}$, which under H_0 has a standard normal distribution. Also, an approximate $100 \cdot (1-\alpha)\%$ confidence interval for p is $\left[\hat{p} - z_{1-\alpha/2}\sqrt{\dfrac{\hat{p}(1-\hat{p})}{n}},\ \hat{p} + z_{1-\alpha/2}\sqrt{\dfrac{\hat{p}(1-\hat{p})}{n}}\right]$.

However, it is better to conduct not an approximate but an exact test based on a binomial distribution. This test is called the **exact binomial test**. The distribution of X is binomial with parameters N and p. Since we draw a *simple random sample* (for which every individual in the population is equally likely to be chosen), it is reasonable to assume that x is also binomially distributed with parameters n and p. Denote by B a binomial random variable with parameters n and p_0. The p-value for the test depends on the form of the alternative hypothesis and is computed as follows:

- For $H_1 : p \geq p_0$, p-value$= \mathbb{P}(B \geq x) = \displaystyle\sum_{k=x}^{n} \binom{n}{k} p_0^k\,(1-p_0)^{n-k}$.
- For $H_1 : p \leq p_0$, p-value$= \mathbb{P}(B \leq x) = \displaystyle\sum_{k=0}^{x} \binom{n}{k} p_0^k\,(1-p_0)^{n-k}$.
- For $H_1 : p \neq p_0$, if $x \geq n/2$,

$$p\text{-value} = \mathbb{P}(B \geq x \text{ or } B \leq n-x) = \sum_{k=x}^{n} \binom{n}{k} p_0^k\,(1-p_0)^{n-k}$$

$$+ \sum_{k=0}^{n-x} \binom{n}{k} p_0^k\,(1-p_0)^{n-k} = \sum_{k=x}^{n} \binom{n}{k} \left[p_0^k\,(1-p_0)^{n-k} + p_0^{n-k}(1-p_0)^k\right].$$

- For $H_1 : p \neq p_0$, if $x \leq n/2$,

$$p\text{-value} = \mathbb{P}(B \leq x \text{ or } B \geq n-x) = \sum_{k=0}^{x} \binom{n}{k} p_0^k\,(1-p_0)^{n-k}$$

$$+ \sum_{k=n-x}^{n} \binom{n}{k} p_0^k\,(1-p_0)^{n-k} = \sum_{k=0}^{x} \binom{n}{k} \left[p_0^k\,(1-p_0)^{n-k} + p_0^{n-k}(1-p_0)^k\right].$$

A $100 \cdot (1-\alpha)\%$ **exact binomial confidence interval** for p is chosen so that the probability of being below the interval is the same as being above it. Such intervals are called *equal-tailed.* The probability of each tail is $\alpha/2$. The CI has the form $[p_L, p_U]$ where p_L, the lower confidence limit, and p_U, the upper confidence limit, solve:

$$\sum_{k=0}^{x} \binom{n}{k} p_U^k \, (1-p_U)^{n-k} \; = \; \alpha/2,$$

and

$$\sum_{k=x}^{n} \binom{n}{k} p_L^k \, (1-p_L)^{n-k} \; = \; \alpha/2.$$

Proof: The values of p_0 that are covered by the confidence interval will be those for which the null hypothesis will not be rejected in favor of a two-sided alternative, at the $\alpha\%$ significance level. It means that the lower confidence limit p_L is the smallest p_0 that satisfies $\mathbb{P}(B \geq x) \geq \alpha/2$, and the upper confidence level p_U is the largest p_0 for which $\mathbb{P}(B \leq x) \geq \alpha/2$. $\square$

Example. A cohort of 10,000 individuals was followed for the duration of an influenza season. In this study, 725 individuals contracted the flu. The estimated prevalence proportion for this sample is $\hat{p} = 725/10000 = 0.0725$. We would like to test at the 5% significance level whether the true population prevalence proportion is larger than 7%, and construct an exact 90% confidence interval for the population parameter based on the binomial distribution, and an approximate 90% CI based on a normal distribution. We run the following SAS and R codes that compute the p-value for the exact binomial test as well as the confidence interval based on the binomial distribution. An approximate confidence interval based on z-distribution is also given in the output.

In SAS:

```
data flu_freq;
do i=1 to 725;
   flu="yes";
   output;
end;
```

```
  do i=726 to 10000;
   flu="no";
   output;
end;
run;

proc freq data=flu_freq;
table flu/binomial (p=.07 level="yes") alpha=0.1;
run;

/*or */
data flu_freq2;
input flu$ freq;
cards;
yes 725
no 9275
;

proc freq data=flu_freq2;
table flu/binomial (p=0.07 level="yes") alpha=0.1;
weight freq;
run;
```

The FREQ Procedure

flu	Frequency	Percent	Cumulative Frequency	Cumulative Percent
no	9275	92.75	9275	92.75
yes	725	7.25	10000	100.00

Binomial Proportion	
flu = yes	
Proportion	0.0725
ASE	0.0026
90% Lower Conf Limit	0.0682
90% Upper Conf Limit	0.0768
Exact Conf Limits	
90% Lower Conf Limit	0.0683
90% Upper Conf Limit	0.0769

Test of H0: Proportion = 0.07	
ASE under H0	0.0026
Z	0.9798
One-sided Pr > Z	0.1636
Two-sided Pr > |Z|	0.3272

In R:

```
#exact binomial test
binom.test(725, 10000, p=0.07, alternative="greater")
#options are alternative=c("two.sided", "greater", "less")
```

```
Exact binomial test

data:  725 and 10000
number of successes = 725, number of trials = 10000, p-value = 0.1683
alternative hypothesis: true probability of success is greater than 0.07
95 percent confidence interval:
 0.06827716 1.00000000
sample estimates:
probability of success
                0.0725
```

```
#exact binomial confidence interval
binom.test(725, 10000, conf.level=0.9)
```

```
Exact binomial test

data:  725 and 10000
number of successes = 725, number of trials = 10000, p-value < 2.2e-16
alternative hypothesis: true probability of success is not equal to 0.5
90 percent confidence interval:
 0.06827716 0.07690823
sample estimates:
probability of success
                0.0725
```

#approximate test
prop.test(725, 10000, conf.level=0.9)

```
1-sample proportions test with continuity correction

data:  725 out of 10000, null probability 0.5
X-squared = 7308.5, df = 1, p-value < 2.2e-16
alternative hypothesis: true p is not equal to 0.5
90 percent confidence interval:
 0.06830067 0.07693330
sample estimates:
     p
0.0725
```

From the above outputs, in SAS, the one-sided p-value is 0.1636, the exact 90% CI is $[0.0683, 0.0769]$, and the approximate 90% CI is [0.0682, 0.0768]. In R, the p-value= 0.1683, the exact 90% CI is $[0.06827716, 0.07690823]$, and the approximate CI is $[0.06830067, 0.07693330]$. The conclusion is that the true population prevalence proportion is not larger than 7%. $\square$