Part III

BINOMIAL DISTRIBUTION

This part deals with tests on proportions. After the Gaussian distribution the binomial distribution is probably the next most famous distribution. Binomial samples are very common and more intuitive than a Gaussian distribution. *Ill* and *healthy*, *success* and *failure*, *poor* and *rich* are well known binomial outcomes. The binomial distribution is linked to the normal distribution via large sample approximation and the normal distribution is the square root of the χ^2-distribution. More importantly, the binomial distribution is a special case of the multinomial distribution. This distribution plays a crucial role in the analysis of contingency tables and the tests in Chapter 4 can also be described in a contingency table set-up. However, this topic in general is covered in Chapter 14. Here we deal with well known special cases. Often the question occurs, if a proportion is the same as a predefined value or if two (or more) proportions differ significantly from each other.

Statistical Hypothesis Testing with SAS and R, First Edition. Dirk Taeger and Sonja Kuhnt.

4

Tests on proportions

In this chapter we present tests for the parameter of a binomial distribution. We first treat a test on the population proportion in the one-sample case. We further cover tests for the difference of two proportions using the pooled as well as the unpooled variances. The last test in this chapter deals with the equality of proportions for the multi-sample case. Not all tests are covered by a SAS procedure or R function. We give the appropriate sample code to perform all discussed tests.

4.1 One-sample tests

In this section we deal with the question, if a population proportion differs from a predefined value between 0 and 1.

4.1.1 Binomial test

Description: Tests if a population proportion p differs from a value p_0.

Assumptions:
- Data are randomly sampled from a large population with two possible outcomes.
- Let $X = 1$ be denoted as "success" and $X = 0$ as "failure".
- The parameter p of interest is given by the proportion of successes in the population.
- The number of successes $\sum_{i=1}^{n} X_i$ in a random sample of size n follows a binomial distribution $B(n, p)$.

Hypothesis:
(A) $H_0 : p = p_0$ vs $H_1 : p \neq p_0$
(B) $H_0 : p \leq p_0$ vs $H_1 : p > p_0$
(C) $H_0 : p \geq p_0$ vs $H_1 : p < p_0$

Test statistic:
$$Z = \frac{\sum\limits_{i=1}^{n} X_i - np_0}{\sqrt{np_0(1 - p_0)}}$$

Statistical Hypothesis Testing with SAS and R, First Edition. Dirk Taeger and Sonja Kuhnt.

Test decision: Reject H_0 if for the observed value z of Z
(A) $z < z_{\alpha/2}$ or $z > z_{1-\alpha/2}$
(B) $z > z_{1-\alpha}$
(C) $z < z_{\alpha}$

p-value: (A) $p = 2\Phi(-|z|)$
(B) $p = 1 - \Phi(z)$
(C) $p = \Phi(z)$

Annotation:

- This is the large sample test. If the sample size is large [rule of thumb: $np(1-p) \geq 9$] the test statistic Z is approximately a standard normal distribution.
- For small samples an exact test with $Y = \sum_{i=1}^{n} X_i$ as test statistic and critical regions based on the binomial distribution are used.

Example: To test the hypothesis that the proportion of defective workpieces of a machine equals 50%. The available dataset contains 40 observations (dataset in Table A.4).

SAS code

```
*** Version 1 ***;
* Only for hypothesis (A) and (C);

proc freq data=malfunction;
  tables malfunction / binomial(level=’1’ p=.5 correct);
  exact binomial;
run;

*** Version 2 ***;
* For hypothesis (A), (B), and (C);

* Calculate the numbers of successes and failures;
proc sort data=malfunction;
 by malfunction;
run;

proc summary data=malfunction n;
 var malfunction;
 by malfunction;
 output out=ptest01  n=n;
run;

* Retrieve the number of successe and failures;
data ptest02 ptest03;;
 set ptest01;
 if malfunction=0 then output ptest02;
 if malfunction=1 then output ptest03;
run;
```

```
* Rename number of failures;
data ptest02;
 set ptest02;
 rename n=failures;
 drop malfunction _TYPE_ _FREQ_;
run;

* Rename number of successes;
data ptest03;
 set ptest03;
 rename n=successes;
 drop malfunction _TYPE_ _FREQ_;
run;

* Calculate test statistic and p-values;
data ptest04;
  merge ptest02 ptest03;
  format test $20.;

  n=successes+failures;
 * Estimated Proportion;
  p_estimate=successes/n;
 * Proportion to test;
  p0=0.5;

 * Perform exact test;
  test="Exact";
  p_value_B=probbnml(p0,n,failures);
  p_value_C=probbnml(p0,n,successes);
  p_value_A=2*min(p_value_B,p_value_c);
  output;

 * Perform asymptotic test;
   test="Asymptotic";
   Z=(successes-n*p0)/sqrt((n*p0*(1-p0)));
   p_value_A=2*probnorm(-abs(Z));
   p_value_B=1-probnorm(-abs(Z));
   p_value_C=probnorm(-abs(Z));
   output;

* Perform asymptotic test with continuity correction;
   test="Asymptotic with correction";
   Z=(abs(successes-n*p0)-0.5)/sqrt((n*p0*(1-p0)));
   p_value_A=2*probnorm(-abs(Z));
   p_value_B=1-probnorm(-abs(Z));
   p_value_C=probnorm(-abs(Z));
  output;
run;

* Output results;
proc print;
  var test Z p_estimate p0 p_value_A p_value_B p_value_C;
run;
```

SAS output

```
Version 1

    Test of H0: Proportion = 0.5

ASE under H0                 0.0791
Z                           -1.4230
One-sided Pr <  Z            0.0774
Two-sided Pr > |Z|           0.1547

Exact Test
One-sided Pr <=  P           0.0769
Two-sided = 2 * One-sided    0.1539

The asymptotic confidence limits and test
    include a continuity correction.

Version 2

test                   p_value_A   p_value_B  p_value_C
Exact                  0.15386     0.95965    0.076930
Asymptotic             0.11385     0.94308    0.056923
Asymptotic with corr   0.15473     0.92264    0.077364
```

Remarks:

- PROC FREQ is the easiest way to perform the binomial test, but the procedure calculates p-values only for hypotheses (A) and (C).
- *level*= indicates the variable level for successes.
- *p*= specifies p_0. The default is 0.5.
- *correct* requests the asymptotic test with continuity correction. This yields a better approximation in some cases by subtracting 0.5 in the numerator if $\sum_{i=1}^{n} X_i - np_0 \geq 0$ and adding 0.5 otherwise. Omitting this option will result in a test without continuity correction.
- *exact binomial* forces SAS to perform the exact test as well.

R code

```
# Number of observations
n<-length(malfunction$malfunction)
# Number of successes
d<-length(malfunction$malfunction
                [malfunction$malfunction==1])
# Proportion to test
p0<-0.5

# Exact test
binom.test(d,n,p0,alternative="two.sided")
```

```
# Asymptotic test
prop.test(d,n,p0,alternative="two.sided",correct=TRUE)
```

R output

```
Exact binomial test
number of successes = 15, number of trials = 40,
                                  p-value = 0.1539

1-sample proportions test with continuity correction
X-squared = 2.025, df = 1, p-value = 0.1547
```

Remarks:

- The function *binom.test* calculates the exact test and the function *prop.test* the asymptotic test.
- The first parameter of both functions is for the number of successes, the second parameter for the number of trials and the third parameter for the proportion to test for.
- `alternative=`*"value"* is optional and indicates the type of alternative hypothesis: "two.sided"= two sided (A); "greater"=true proportion is greater (B); "less"=true proportion is lower (C). Default is "two.sided".
- The asymptotic test provides an additional parameter. With "corrected=TRUE" the test with continuity correction is applied. This yields a better approximation in some cases. A Yates' continuity correction is applied, but only if $0.5 \leq \left|\sum_{i=1}^{n} X_i - np_0\right|$. The default value is "correct=FALSE".
- Because the test statistics of the one-sample proportion test and the χ^2-test for one-way tables are equivalent, R uses the latter test.

4.2 Two-sample tests

In this section we deal with the question, if proportions of two independent populations differ from each other. We present two tests for this problem (Keller and Warrack 1997). In the first case the standard deviations of both distributions may differ from each other. In the second case equal but unknown standard deviations are assumed such that both samples can be pooled to obtain a better estimate of the standard deviation. Both presented tests are based on an asymptotic standard normal distribution.

4.2.1 z-test for the difference of two proportions (unpooled variances)

Description: Tests if two population proportions p_1 and p_2 differ by a specific value d_0.

Assumptions:

- Data are randomly sampled with two possible outcomes.
- Let $X = 1$ be denoted as "success" and $X = 0$ as "failure".
- The parameters p_1 and p_2 are the proportions of success in the two populations.
- Data are randomly sampled from two populations with sample sizes n_1 and n_2.
- The number of successes $\sum_{i=1}^{n_j} X_{ji}$ in the j^{th} sample follows a binomial distribution $B(n_j, p_j), j = 1, 2$.

Hypothesis:

(A) $H_0 : p_1 - p_2 = d_0$ vs $H_1 : p_1 - p_2 \neq d_0$
(B) $H_0 : p_1 - p_2 \leq d_0$ vs $H_1 : p_1 - p_2 > d_0$
(C) $H_0 : p_1 - p_2 \geq d_0$ vs $H_1 : p_1 - p_2 < d_0$

Test statistic:

$$Z = \left[(\hat{p}_1 - \hat{p}_2) - d_0\right] \Big/ \sqrt{\tfrac{\hat{p}_1(1-\hat{p}_1)}{n_1} + \tfrac{\hat{p}_2(1-\hat{p}_2)}{n_2}}$$

where $\hat{p}_1 = \frac{1}{n_1} \sum_{i=1}^{n_1} X_{1i}$ and $\hat{p}_2 = \frac{1}{n_2} \sum_{i=1}^{n_2} X_{2i}$

Test decision:

Reject H_0 if for the observed value z of Z
(A) $z < z_{\alpha/2}$ or $z > z_{1-\alpha/2}$
(B) $z > z_{1-\alpha}$
(C) $z < z_{\alpha}$

p-value:

(A) $p = 2\Phi(-|z|)$
(B) $p = 1 - \Phi(z)$
(C) $p = \Phi(z)$

Annotation:

- This is a large sample test. If the sample size is large enough the test statistic Z is a standard normal distribution. As a rule of thumb $n_1 p_1$, $n_1(1 - p_1)$, $n_2\, p_2$ and $n_2(1 - p_2)$ should all be ≥ 5.

Example: To test the hypothesis that the proportion of defective workpieces of company A and company B differ by 10%. The dataset contains $n_1 = 20$ observations from company A and $n_2 = 20$ observations from company B (dataset in Table A.4).

SAS code

```
* Determining sample sizes and number of successes;
proc means data=malfunction n sum;
 var malfunction;
 by company;
 output out=prop1 n=n sum=success;
run;
```

```
* Retrieve these results as two separate datasets;
data propA propB;
 set prop1;
 if company="A" then output propA;
 if company="B" then output propB;
run;

* Relative frequencies of successes for company A;
data propA;
 set propA;
 keep n success p1;
 rename n=n1
        success=success1;
 p1=success/n;
run;

* Relative frequencies of successes for company B;
data propB;
 set propB;
 keep n success p2;
 rename n=n2
        success=success2;
 p2=success/n;
run;

* Merge datasets of company A and B;
data prop2;
 merge propA propB;
run;

* Calculate test statistic and p-value;
data prop3;
 set prop2;
 format p_value pvalue.;

 p_diff=p1-p2; *Difference of proportions;
 d0=0.10;      *Difference to be tested;

 * Test statistic and p-values;
 z=(p_diff-d0)/sqrt((p1*(1-p1))/n1 + (p2*(1-p2))/n2);
 p_value=2*probnorm(-abs(z));
run;

proc print;
 var z p_value;
run;
```

SAS output

```
   z        p_value
1.75142     0.0799
```

Remarks:

- There is no SAS procedure to calculate this test directly.
- The data do not fulfill the criteria to ensure that the test statistic Z is a Gaussian distribution, because $n_2 * p_2 = 4 \ngeq 5$, therefore the p-value is questionable.

R code

```
# Number of observations for company A
n1<-length(malfunction$malfunction
                             [malfunction$company=='A'])
# Number of successes for company A
s1<-length(malfunction$malfunction[malfunction$company=='A'
                              & malfunction$malfunction==1])
# Number of observations for company B
n2<-length(malfunction$malfunction
                             [malfunction$company=='B'])
# Number of successes for company B
s2<-length(malfunction$malfunction[malfunction$company=='B'
                              & malfunction$malfunction==1])

# Proportions
p1=s1/n1
p2=s2/n2

# Difference of proportions
p_diff=p1-p2
# Difference to test
d0=0.10

# Test statistic and p-values
z=(p_diff-d0)/sqrt((p1*(1-p1))/n1 + (p2*(1-p2))/n2)
p_value=2*pnorm(-abs(z))

# Output results
z
p_value
```

R output

```
> z
[1] 1.751424
> p_value
[1] 0.07987297
```

Remarks:

- There is no R function to calculate this test directly.
- The data do not fulfill the criteria to ensure that the test statistic Z is a Gaussian distribution, because $n_2 * p_2 = 4 \ngeq 5$, therefore the p-value is questionable.

4.2.2 z-test for the equality between two proportions (pooled variances)

Description: Tests if two population proportions p_1 and p_2 differ from each other.

Assumptions:

- Data are randomly sampled with two possible outcomes.
- Let $X = 1$ be denoted as "success" and $X = 0$ as "failure".
- The parameters p_1 and p_2 are the proportions of success in the two populations.
- Data are randomly sampled from two populations with sample sizes n_1 and n_2.
- The number of successes $\sum_{i=1}^{n_j} X_{ji}$ in the j^{th} sample follow a binomial distribution $B(n_j, p_j), j = 1, 2$.

Hypothesis:
(A) $H_0 : p_1 - p_2 = 0$ vs $H_1 : p_1 - p_2 \neq 0$
(B) $H_0 : p_1 - p_2 \leq 0$ vs $H_1 : p_1 - p_2 > 0$
(C) $H_0 : p_1 - p_2 \geq 0$ vs $H_1 : p_1 - p_2 < 0$

Test statistic:

$$Z = \frac{\hat{p}_1 - \hat{p}_2}{\sqrt{\hat{p}(1-\hat{p})\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}} \quad \text{with} \quad \hat{p} = \frac{\hat{p}_1 n_1 + \hat{p}_2 n_2}{n_1 + n_2}.$$

where $\hat{p}_1 = \frac{1}{n_1} \sum_{i=1}^{n_1} X_{1i}$ and $\hat{p}_2 = \frac{1}{n_2} \sum_{i=1}^{n_2} X_{2i}$

Test decision: Reject H_0 if for the observed value z of Z
(A) $z < z_{\alpha/2}$ or $z > z_{1-\alpha/2}$
(B) $z > z_{1-\alpha}$
(C) $z < z_{\alpha}$

p-value:
(A) $p = 2\Phi(-|z|)$
(B) $p = 1 - \Phi(z)$
(C) $p = \Phi(z)$

Annotation:

- This is a large sample test. If the sample size is large enough the test statistic Z is a standard normal distribution. As a rule of thumb following $n_1 p_1$, $n_1(1 - p_1)$, $n_2\, p_2$ and $n_2(1 - p_2)$ should all be ≥ 5.
- This test is equivalent to the χ^2-test of a 2×2 table, that is, $Z^2 = \chi^2 \sim \chi_1^2$. The advantage of the χ^2-test is that there exists an exact test for small samples, which calculates the p-values from the exact distribution. This test is the famous Fisher's exact test. More information is given in Chapter 14.

Example: To test the hypothesis that the proportion of defective workpieces of company A and company B are equal. The dataset contains $n_1 = 20$ observations from company A and $n_2 = 20$ observations from company B (dataset in Table A.4).

SAS code

```
* Determining sample sizes and number of successes;
proc means data=malfunction n sum;
 var malfunction;
 by company;
 output out=prop1 n=n sum=success;
run;

* Retrieve these results in two separate datasets;
data propA propB;
 set prop1;
 if company="A" then output propA;
 if company="B" then output propB;
run;

* Relative frequencies of successes for company A;
data propA;
 set propA;
 keep n success p1;
 rename n=n1
        success=success1;
 p1=success/n;
run;

* Relative frequencies of successes for company B;
data propB;
 set propB;
 keep n success p2;
 rename n=n2
        success=success2;
 p2=success/n;
run;

* Merge datasets of company A and B;
data prop2;
 merge propA propB;
run;

* Calculate test statistic and p-value;
data prop3;
 set prop2;
 format p_value pvalue.;

 * Test statistic and p-values;
 p=(p1*n1+p2*n2)/(n1+n2);
 z=(p1-p2)/sqrt((p*(1-p))*(1/n1+1/n2));
 p_value=2*probnorm(-abs(z));
run;

proc print;
 var z p_value;
run;
```

SAS output

```
   z       p_value
2.28619    0.0222
```

Remarks:

- There is no SAS procedure to calculate this test directly.
- The data do not fulfill the criteria to ensure that the test statistic Z is a Gaussian distribution, because $n_2 * p_2 = 4 \ngeq 5$. In this case it is better to use Fisher’s exact test, see Chapter 14.

R code

```
# Number of observations for company A
n1<-length(malfunction$malfunction
                       [malfunction$company==’A’])

# Number of successes for company A
s1<-length(malfunction$malfunction[malfunction$company==’A’
                        & malfunction$malfunction==1])

# Number of observations for company B
n2<-length(malfunction$malfunction
                       [malfunction$company==’B’])

# Number of successes for company A
s2<-length(malfunction$malfunction[malfunction$company==’B’
                       & malfunction$malfunction==1])

# Proportions
p1=s1/n1
p2=s2/n2

# Test statistic and p-value
p=(p1*n1+p2*n2)/(n1+n2)
z=(p1-p2)/sqrt((p*(1-p))*(1/n1+1/n2))
p_value=2*pnorm(-abs(z))

# Output results
z
p_value
```

R output

```
> z
[1] 2.286190
> p_value
[1] 0.02224312
```

Remarks:

- There is no R function to calculate this test directly.
- The data do not fulfill the criteria to ensure that the test statistic Z is a Gaussian distribution, because $n_2 * p_2 = 4 \ngeq 5$. In this case it is better to use Fisher's exact test, see Chapter 14.

4.3 *K*-sample tests

Next we present the population proportion equality test for K samples [see Bain and Engelhardt (1991) for further details]. If we have K independent binomial samples we can arrange them in a $K \times 2$table and take advantage of results on contingency tables. We concentrate on the χ^2-test based on asymptotic results, although Fisher's exact test can be used as well. More details are given in Chapter 14.

4.3.1 *K*-sample binomial test

Description: Tests if k population proportions, p_i, $i = 1, \ldots, K$, differ from each other.

Assumptions:

- Data are randomly sampled with two possible outcomes.
- Let $X = 1$ be denoted as "success" and $X = 0$ as "failure".
- The parameters p_k are the proportions of success in the k^{th} populations, $k = 1, \ldots, K$.
- Data are randomly sampled from the K populations with sample sizes n_k, $k = 1, \ldots, K$.
- The number of successes $\sum_{i=1}^{n_k} X_{ki}$ in the k^{th} sample follow a binomial distribution $B(n_k, p_k)$, $k = 1, \ldots, K$.

Hypothesis: $H_0 : p_1 = \ldots = p_K$ vs $H_1 : p_k \neq p_{k'}$ for at least one $k \neq k'$

Test statistic:

$$\chi^2 = \sum_{k=1}^{K} \sum_{j=0}^{1} \frac{(O_{kj} - E_{kj})^2}{E_{kj}}$$

where $O_{k1} = \sum_{i=1}^{n_k} X_{ki}$, $O_{k0} = n_k - O_{k1}$,

$$\hat{p} = \frac{1}{n} \sum_{k=1}^{K} O_{k1},\ n = \sum_{k=1}^{K} n_k,\ E_{k1} = n_k \hat{p},\ E_{k0} = n_k(1 - \hat{p}).$$

Test decision: Reject H_0 if for the observed value χ_0 of χ^2
$\chi_0 > \chi_{1-\alpha;K-1}$

p-value: $p = 1 - P(\chi^2 \leq \chi_0)$

Annotation:

- The test statistic χ^2 is χ^2_{K-1}-distributed.
- $\chi_{1-\alpha;K-1}$ is the (1-α)-quantile of the χ^2-distribution with $K-1$ degrees of freedom.
- If not all expected absolute frequencies E_{kj} are larger or equal to 5, use Fisher's exact test (see Test 14.1.1).

Example: The proportions of male carp in three ponds are tested for equality. The observed relative frequency of male carp in pond one is 10/19, in pond two 12/20, and in pond three 14/21.

SAS code

```
data counts;
input r c counts;
datalines;
1 1 10
1 0  9
2 1 12
2 0  8
3 1 14
3 0  7
;
run;

proc freq;
 tables r*c /chisq;
 weight counts;
run;
```

SAS output

```
Statistic      DF     Value     Prob
Chi-Square     2      0.8187    0.6641
```

Remarks:

- The data step constructs a 3 × 2 contingency, with r for rows (ponds 1 to 3) and c for columns (1 for male and 0 for female carp). The variable counts includes the counts for each combination between ponds and sex. In `proc freq` these counts can be passed by using the `weight` statement.
- With `proc freq` it is also possible to use raw data instead of a predefined contingency table to perform these tests. In this case there must be one variable for the ponds and one for the sex and one row for each carp. Use the same SAS statement but omit the `weight` command.
- Because the null hypothesis is rejected if $\chi^2 \geq \chi^2_{1-\alpha;2}$, the p-value must be calculated as `1-pchisq(0.8187,2)`.

R code

```
x1 <- matrix(c(10, 12, 14, 9, 8, 7), ncol = 2)
chisq.test(x1)
```

R output

```
X-squared = 0.8187, df = 2, p-value = 0.6641
```

Remarks:

- The matrix command constructs a matrix X with the ponds in the columns and the male carp population in the first row and the female carp population the second row. This matrix can then be passed on to the `chisq.test` function.
- Because the null hypothesis is rejected if $\chi^2 \geq \chi^2_{1-\alpha;2}$, the p-value must be calculated as `1-pchisq(0.8187,2).`

References

Bain L.J. and Engelhardt M. 1991 *Introduction to Probability and Mathematical Statistics*, 1st edn. Duxbury Press.

Keller G. and Warrack B. 1997 *Statistics for Management and Economics*, 4th edn. Duxbury Press.