Part IV

OTHER DISTRIBUTIONS

In this part we consider parameter tests for two distributions different from the already covered Gaussian and binomial distributions, namely the Poisson and exponential distributions. Tests for these two distributions are presented in Chapters 5 and 6, respectively. For the Poisson distribution we also consider a test on the difference between two Poisson parameters.

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

5

Poisson distribution

The Poisson distribution is often called the distribution of rare events for count data. In health sciences it is used to model survival or mortality data. Other applications include the number of traffic accidents and the number of machine failures over a specific time frame. The Poisson distribution is closely related to the binomial distribution (if the sample size n is large and the success probability p is small) and to the Gaussian distribution which is deployed in the statistical tests presented next.

5.1 Tests on the Poisson parameter

In this section we deal with the question, if the parameter λ of a Poisson distribution differs from a value λ_0. The first test gives a solution to this hypothesis through a Gaussian approximation of the Poisson distribution, the second test shows how to perform an exact test. The third test deals with the question if the difference between two Poisson parameters differs from zero.

5.1.1 z-test on the Poisson parameter

Description: Tests if the parameter λ of a Poisson distribution differs from a specific value λ_0.

Assumptions:
- Data are measured as counts.
- The random variables $X_i, i = 1, \ldots, n$, are Poisson distributed with parameter λ.

Hypotheses:
(A) $H_0 : \lambda = \lambda_0$ vs $H_1 : \lambda \neq \lambda_0$
(B) $H_0 : \lambda \leq \lambda_0$ vs $H_1 : \lambda > \lambda_0$
(C) $H_0 : \lambda \geq \lambda_0$ vs $H_1 : \lambda < \lambda_0$

Test statistic: $$Z = \frac{\sum_{i=1}^{n} X_i - n\lambda_0}{\sqrt{n\lambda_0}}$$

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) $p = \Phi(z)$

Annotations:

- The test statistic Z follows a standard normal distribution, if $n\lambda$ is sufficiently large. If each X_i follows a $Poi(\lambda)$ distribution, then $\sum_{i=1}^{n} X_i$ follows a Poisson distribution with parameter $n\lambda$, which approximately follows a Gaussian distribution with mean and variance $n\lambda$. A continuity correction can improve this approximation.
- For an exact test see Test 5.1.2.

Example: Of interest are hospital infections on the islands of Laputa and Luggnagg with a conjecture of 4 expected infections per hospital in a year. The null hypothesis is $H_0 : \lambda = \lambda_0$, with $\lambda_0 = 4$. The dataset summarizes for how many of the 42 hospitals on the islands the number of infections ranges from zero to six (dataset in Table A.5).

SAS code

```
* Calculate the number of total infections;
data infections1;
 set c.infections;
 n_infections=infections*total;
run;

proc means data=infections1 sum;
 var n_infections total;
 output out=infections2 sum=x n_hospital;
run;

* Calculate test statistic and p-value;
data infections3;
 set infections2;
 format p_value pvalue.;

 lambda0=4; * Set lambda under the null hypothesis

 * Test statistic and p-values;
 z=(x-n_hospital*lambda0)/sqrt(n_hospital*lambda0);
 p_value_A=2*probnorm(-abs(z));
 p_value_B=1-probnorm(z);
 p_value_C=probnorm(z);
```

```
* Output results;
proc print;
 var z p_value_A p_value_B p_value_C;
run;
```

SAS output

```
    z      p_value_A  p_value_B   p_value_C
-1.69734   0.089633   0.95518     0.044816
```

Remarks:

- There is no basic SAS procedure to calculate the z-test on the Poisson parameter directly.

R code

```
# Number of observed total infections
x<-sum(infections$infections*infections$total)
n_hospital<-sum(infections$total)

# Set lambda under the null hypothesis
lambda0<-4

# Test statistic and p-value
z<-(x-n_hospital*lambda0)/sqrt(n_hospital*lambda0)
p_value_A=2*pnorm(-abs(z))
p_value_B=1-pnorm(z)
p_value_C=pnorm(z)

# Output results
z
p_value_A
p_value_B
p_value_C
```

R output

```
> z
[1] -1.697337
> p_value_A
[1] 0.08963299
> p_value_B
[1] 0.9551835
> p_value_C
[1] 0.0448165
```

Remarks:

- There is no basic R function to calculate the z-test on the Poisson parameter directly.

5.1.2 Exact test on the Poisson parameter

Description: Tests if the parameter λ of a Poisson distribution differs from a pre-specified value λ_0.

Assumptions:
- Data are measured as counts.
- The random variables $X_i, i = 1, \dots, n$, are Poisson distributed with parameter λ.

Hypotheses:
(A) $H_0 : \lambda = \lambda_0$ vs $H_1 : \lambda \neq \lambda_0$
(B) $H_0 : \lambda \leq \lambda_0$ vs $H_1 : \lambda > \lambda_0$
(C) $H_0 : \lambda \geq \lambda_0$ vs $H_1 : \lambda < \lambda_0$

Test statistic: $S = \sum_{i=1}^{n} X_i$

Test decision: Reject H_0 if for the observed value s of S
(A) $s < c_{n\lambda_0;\alpha/2}$ or $s > c_{n\lambda_0;1-\alpha/2}$
(B) $s > c_{n\lambda_0;1-\alpha}$
(C) $s < c_{n\lambda_0;\alpha}$
where the critical values $c_{\lambda,\alpha}$ are coming from the cumulative distribution function of a Poisson distribution with parameter λ and quantile α.

p-value:
(A) If $s < n\lambda_0 : p = \min\left\{2 \times \sum_{k=0}^{S} \frac{n\lambda_0^k}{k!} e^{-n\lambda_0}, 1\right\}$

If $s > n\lambda_0 : p = \min\left\{2 \times \left(1 - \sum_{k=0}^{S-1} \frac{n\lambda_0^k}{k!} e^{-n\lambda_0}\right), 1\right\}$

(B) $p = 1 - \sum_{k=0}^{S-1} \frac{n\lambda_0^k}{k!} e^{-n\lambda_0}$

(C) $p = \sum_{k=0}^{S} \frac{n\lambda_0^k}{k!} e^{-n\lambda_0}$

Annotations:
- This test is the exact version of Test 5.1.1.
- The p-values are calculated using the sum of the observed counts s and the expected counts λ_0 using the cumulative distribution function of the Poisson distribution [see Rosner 2011 for details].

Example: Of interest are hospital infections on the islands of Laputa and Luggnagg with a conjecture of 4 expected infections per hospital in a year. The null hypothesis is $H_0 : \lambda = \lambda_0$, with $\lambda_0 = 4$. The dataset summarizes for how many of the 42 hospitals on the islands the number of infections ranges from zero to six (dataset in Table A.5).

SAS code

```
* Calculate the number of total infections;
data infections1;
 set c.infections;
 n_infections=infections*total;
run;

proc means data=infections1 sum;
 var n_infections total;
 output out=infections2 sum=x n_hospital;
run;

* Calculate p-values;
data infections3;
 set infections2;
 format p_value_A p_value_B p_value_C pvalue.;

 lambda0=4; * Set lambda under the null hypothesis

 * Test on hypothesis A;
 if x<n_hospital*lambda0  then
     p_value_A=min(2*CDF('Poisson',x,n_hospital*lambda0),1);
 if x>=n_hospital*lambda0 then
     p_value_A=
         min(2*(1-CDF('Poisson',x-1,n_hospital*lambda0)),1);

 * Test on hypothesis B;
 p_value_B=1-CDF('Poisson',x-1,n_hospital*lambda0);

 * Test on hypothesis C;
 p_value_C=CDF('Poisson',x,n_hospital*lambda0);
run;

* Output results;
proc print;
 var p_value_A p_value_B p_value_C;
run;
```

SAS output

```
p_value_A    p_value_B   p_value_C
 0.0924       0.9611      0.0462
```

Remarks:

- There is no basic SAS procedure to calculate the exact test directly.
- The SAS function `CDF()` is used to calculate the p-values. The first parameter sets the distribution function to a Poisson distribution, the second parameter is the number of observed cases and the third parameter the mean of the Poisson distribution.

R code

```
# Number of observed total infections
x<-sum(infections$infections*infections$total)

# Number of hospitals
n_hospital<-sum(infections$total)

# Set lambda under the null hypothesis
lambda0<-4

# Test of Hypothesis A
poisson.test(x,n_hospital*lambda0,alternative="two.sided")

# Test of Hypothesis B
poisson.test(x,n_hospital*lambda0,alternative="greater")

# Test of Hypothesis C
poisson.test(x,n_hospital*lambda0,alternative="less")
```

R output

```
# Test of Hypothesis A
p-value=0.09692

# Test of Hypothesis B
p-value=0.9611

# Test of Hypothesis C
p-value=0.04619
```

Remarks:

- This test can be conducted using the `poisson.test()` function of R.
- `alternative=`“*value*” is optional and defines the type of alternative hypothesis: “two.sided”= hypothesis (A); “greater”= hypothesis (B); “lower”= hypothesis (C). Default is “two.sided”.
- The R function `poisson.test()` uses a slightly different algorithm than SAS to calculate the two-sided p-value, which is reflected in the output.

5.1.3 z-test on the difference between two Poisson parameters

Description: Tests if the difference between two Poisson parameters λ_1 and λ_2 is zero.

Assumptions:
- Data are measured as counts.
- Data are randomly sampled from two independent Poisson distributions.
- The random variables $X_{11}, \dots, X_{1n_1}$ are coming from a Poisson distribution with parameter λ_1 and $X_{21}, \dots, X_{2n_2}$ from a Poisson distribution with parameter λ_2.

Hypotheses:
(A) $H_0 : \lambda_1 - \lambda_2 = 0$ vs $H_1 : \lambda_1 - \lambda_2 \neq 0$
(B) $H_0 : \lambda_1 - \lambda_2 \leq 0$ vs $H_1 : \lambda_1 - \lambda_2 > 0$
(C) $H_0 : \lambda_1 - \lambda_2 \geq 0$ vs $H_1 : \lambda_1 - \lambda_2 < 0$

Test statistic:

$$Z = \frac{X_1/n_1 - X_2/n_2}{\sqrt{X_1/n_1^2 + X_2/n_2^2}},$$

with $X_1 = \sum_{i=1}^{n_1} X_{1i}$ and $X_2 = \sum_{i=1}^{n_2} X_{2i}$

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) $p = \Phi(z)$

Annotations:

- The test statistic Z follows a standard normal distribution, if $n_1\lambda_1$ and $n_2\lambda_2$ are large enough to fulfill the approximation to the Gaussian distribution.
- A continuity correction does not improve the test according to Detre and White (1970).
- For details on this test see Thode (1997).

Example: To test, if the difference in free kicks between two soccer teams is zero. Team A had 88 free kicks in 15 soccer games and team B had 76 free kicks in 10 games.

SAS code

```
data kick;
 format p_value pvalue.;

 * Define free kicks of team A and B;
 fk_A=88;
 fk_B=76;

 * Number of soccer games of team A and team B;
 n_A=15;
 n_B=10;

 * Test statistic and p-values;
 z=(fk_A/n_A-fk_B/n_B)/sqrt(fk_A/(n_A)**2+fk_B/(n_B)**2);
 p_value_A=2*probnorm(-abs(z));
 p_value_B=1-probnorm(z);
 p_value_C=probnorm(z);
```

```
* Output results;
proc print;
 var z p_value_A p_value_B p_value_C;
run;
```

SAS output

```
   z        p_value_A   p_value_B   p_value_C
-1.61556   0.10619      0.94691     0.053095
```

Remarks:

- There is no basic SAS procedure to calculate the z-test on Poisson parameters directly.

R code

```
# Define free kicks of team A and B
fk_A<-88
fk_B<-76

# Number of soccer games of team A and team B;
n_A=15;
n_B=10;

# Test statistic and p-values
z<-(fk_A/n_A-fk_B/n_B)/sqrt(fk_A/n_A^2+fk_B/n_B^2)
p_value_A=2*pnorm(-abs(z))
p_value_B=1-pnorm(z)
p_value_C=pnorm(z)

# Output results
z
p_value_A
p_value_B
p_value_C
```

R output

```
> z
[1] -1.615561
> p_value_A
[1] 0.1061892
> p_value_B
[1] 0.9469054
> p_value_C
[1] 0.05309459
```

Remarks:

- There is no basic R function to calculate the z-test on Poisson parameters directly.

References

Rosner B. 2011 *Fundamentals of Biostatistics*, 7th edn. Brooks/Cole Cengage Learning.

Thode H.C. 1977 Power and sample size requirements for tests of differences between two Poisson rates. *The Statistician* **46**, 227–230.

Detre K. and White C. 1970 The comparison of two Poisson distributed observations. *Biometrics* **26**, 851–854.

6

Exponential distribution

The exponential distribution is typically used to model waiting times, for example, to model survival times of cancer patients or life times of machine components. Often these are waiting times between the events of a Poisson distribution. The exponential distribution is memoryless in the sense that the remaining waiting time at a specific time point is independent of the time which has already elapsed since the last event. The parameter λ can be seen as a constant failure rate (failures over a specific unit such as hours).

6.1 Test on the parameter of an exponential distribution

6.1.1 z-test on the parameter of an exponential distribution

Description: Tests if the parameter λ of an exponential distribution differs from a specific value λ_0.

Assumptions:
- Data are measured on an interval or ratio scale.
- Data are randomly sampled from an exponential distribution.
- The random variables $X_1, \dots, X_n, i = 1, \dots, n$, are exponentially distributed with parameter λ.
- Let p be the probability of failure during a time period T. Then $p = 1 - e^{-\lambda T}$.

Hypotheses: (A) $H_0 : \lambda = \lambda_0$ vs $H_1 : \lambda \neq \lambda_0$
(B) $H_0 : \lambda \leq \lambda_0$ vs $H_1 : \lambda > \lambda_0$
(C) $H_0 : \lambda \geq \lambda_0$ vs $H_1 : \lambda < \lambda_0$

Test statistic:

$$Z = \frac{M - np}{\sqrt{np(1-p)}} \quad \text{with} \quad p = 1 - e^{-\lambda_0 T}$$

$$\text{and} \quad M = \sum_{1}^{n} 1\!\!1_{[0,T]}\left\{X_i\right\} \text{ number of failures}$$

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) $p = \Phi(z)$

Annotations:

- The variable M is binomially distributed with parameters n and p (Bain and Engelhardt 1991, p. 555).
- The test statistic Z follows a standard normal distribution if n is large. This condition usually holds if $np(1-p) \geq 9$.
- z_{α} is the α-quantile of the standard normal distribution.

Example: To test, if the failure rate of pocket lamps is equal to $\lambda_0 = 0.2$ within a year. For 100 lamps we observe 25 failures during a time period $T = 1$ year.

SAS code

```
data expo;
 format p_value pvalue.;

 n=100;       * Number of observations;
 T=1;         * Time interval;
 M=25;        * Number of failures;
 lambda0=0.2; * Failure rate under the null hypothesis;
 p=1-exp(-lambda0*T); *Probability of failure

* Test statistic and p-value;
z=(25-n*p)/sqrt(n*p*(1-p));
p_value=2*probnorm(-abs(z));

* Output results;
proc print;
 var z p_value;
run;
```

SAS output

```
z         p_value
1.78410   0.0744
```

Remarks:

- There is no basic SAS procedure to calculate the test directly.
- The one-sided p-value for hypothesis (B) can be calculated with `p_value_B=1-probnorm(z)` and the p-value for hypothesis (C) with `p_value_C=probnorm(z)`.

R code

```
n<-100        # Number of observations
T<-1          # Time interval
M<-25         # Number of failures
lambda0<-0.2 # Failure rate under the null hypothesis
p=1-exp(-lambda0*T) #Probability of failure

# Test statistic and p-value
z=(25-n*p)/sqrt(n*p*(1-p))
p_value=2*pnorm(-abs(z))

# Output results
z
p_value
```

R output

```
> z
[1] 1.784097
> p_value
[1] 0.07440789
```

Remarks:

- There is no basic R function to calculate the test directly.
- The one-sided p-value for hypothesis (B) can be calculated with `p_value_B=1-pnorm(z)` and the p-value for hypothesis (C) with `p_value_C=pnorm(z)`.

Reference

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