Categorical Data Analysis

The Universal Machinery of Chi-square

Radu T. Trîmbiţaş

May 19, 2016

1 Multinomial Experiments

Multinomial Experiments

A *multinomial experiment* has the following characteristics

1. The experiment consist of n identical trials.
2. The outcome of each trial falls into one of k categories or cells.
3. The probability that the outcome of a single trial will fall in a particular cell, say cell i, is p_i, where $i = \overline{1,k}$ and remains the same from trial to trial. Notice that
$$p_1 + p_2 + p_3 + \cdots + p_k = 1.$$
4. The trials are independent.
5. We are interested in $n_1, n_2, \ldots, n_k$, where n_i for $i = \overline{1,k}$ is equal to the number of trials in which the outcome falls into cell i. Notice that $n_1 + n_2 + \cdots + n_k = n$.

- Objective: inferences about the cell probabilities $p_1, p_2, \ldots, p_k$.
- Examples:
  - Employees can be classified into one of five income brackets.
  - Mice might react in one of three ways when subjected to a stimulus.
  - Motor vehicles might fall into one of four vehicle types.
  - Paintings could be classified into one of k categories according to style and period.

2 The Chi-Square Test

The Chi-Square Test

- The expected number of outcomes falling in the cell C_i may be calculated using the formula
$$E(n_i) = np_i, \quad i = \overline{1,k}.$$
- Now suppose that we hypothesize values for $p_1, p_2, \ldots, p_k$ and calculate the expected value for each cell. Certainly, if our hypothesis is true, the cell counts n_i should not deviate greatly from their expected values np_i, for $i = \overline{1,k}$.
- In 1900 Karl Pearson proposed the following test statistic
$$X^2 = \sum_{i=1}^{k} \frac{[N_i - E(n_i)]^2}{E(n_i)} = \sum_{i=1}^{n} \frac{(N_i - np_i)^2}{np_i}. \tag{1}$$
- Meaning:
$$X^2 = \sum_{i=1}^{k} \frac{(O_i - E_i)^2}{E_i}$$
where O_i - observed frequencies, E_i- expected frequencies.
- Q: Why this numerator and why this denominator? A: To distiguish 15-5 of 110-100
- This statistic is asymptotically standardized chi square with $k-1$ degrees of freedom distributed.

Distribution of the Chi-Square Statistics

Theorem 1. *The statistic*
$$X^2 = \sum_{i=1}^{k} \frac{(N_i - np_i)^2}{np_i}$$
has a standardized χ^2 distribution with $k-1$ degrees of freedom, when $n \to \infty$.

Proof

We start from Stirling formula $n! \sim n^n e^{-n}\sqrt{2\pi n}$. Since for a multinomial distribution it holds
$$P(N_1 = n_1, N_2 = n_2, \ldots, N_k = n_k) = \frac{n!}{n_1! n_2! \ldots n_k!} p_1^{n_1} p_2^{n_2} \ldots p_k^{n_k},$$
we have
$$P(N_1 = n_1, \ldots, N_k = n_k) \approx \frac{n^n e^{-n}\sqrt{2\pi n}}{\prod_{i=1}^{k}(\sqrt{2\pi n_i} n_i^{n_i} e^{-n_i})} p_1^{n_1} \ldots p_k^{n_k}$$

or

$$P(N_1 = n_1, \ldots, N_k = n_k) \approx K \prod_{i=1}^{k} \left(\frac{np_i}{n_i} \right)^{n_i + \frac{1}{2}}$$

where $K > 0$ is a constant.

Taking the logarithm we get

$$\ln P(N_1 = n_1, \ldots, N_k = n_k) \approx \ln K + \sum_{i=1}^{k} \left(n_i + \frac{1}{2} \right) \ln \frac{np_i}{n_i},$$

and setting

$$x_i = \frac{n_i - np_i}{\sqrt{np_i}}, \text{ i.e. } \frac{n_i}{np_i} = 1 + \frac{x_i}{\sqrt{np_i}},$$

it follows that

$$\ln P(N_1 = n_1, \ldots, N_k = n_k) \approx \ln K - \sum_{i=1}^{k} \left(n_i + \frac{1}{2} \right) \ln \left(1 + \frac{x_i}{\sqrt{np_i}} \right).$$

Now, using a Taylor expansion of natural logarithm truncated to two terms

$$\ln \left(1 + \frac{x_i}{\sqrt{np_i}} \right) \approx \frac{x_i}{\sqrt{np_i}} - \frac{x_i^2}{2np_i}$$

and taking into account that

$$\sum_{i=1}^{k} x_i \sqrt{np_i} = \sum_{i=1}^{k} (n_i - np_i) = n - n = 0$$

we get successively

$$\begin{aligned} \ln P(N_1 = n_1, N_2 = n_2, \ldots, N_k = n_k) & \\ \approx \ln K - \sum_{i=1}^{k} \left(n_i + \frac{1}{2} \right) \left(\frac{x_i}{\sqrt{np_i}} - \frac{x_i^2}{2np_i} \right) & \\ = \ln K - \sum_{i=1}^{k} \left(np_i + x_i \sqrt{np_i} + \frac{1}{2} \right) \left(\frac{x_i}{\sqrt{np_i}} - \frac{x_i^2}{2np_i} \right) & \\ \approx \ln K - \sum_{i=1}^{k} \left(x_i \sqrt{np_i} + \frac{x_i^2}{2} \right) = \ln K - \frac{1}{2} \sum_{i=1}^{k} x_i^2 & \end{aligned}$$

or

$$P(N_1 = n_1, \ldots, N_k = n_k) \approx K e^{-\frac{1}{2} \sum_{i=1}^{k} x_i^2}.$$

Now putting $X_i = \frac{N_i - np_i}{\sqrt{np_i}}$, one gets

$$P(X_1 = x_1, \ldots, X_k = x_k) \approx K e^{-\frac{1}{2} \sum_{i=1}^{k} x_i^2},$$

that is, the random vector

$$(X_i)_{i=\overline{1,k}} = \left(\frac{N_i - np_i}{\sqrt{np_i}}\right)_{i=\overline{1,k}}$$

has, for $n \to \infty$ a degenerated k-dimensional normal distribution, since each X_i is a linear combination of the others. Since a sum of squares of normally distributed random variable has a chi-square distribution the proof is complete.

Remarks

- The approximation stated in Theorem 1 is good if all theoretical frequencies $E_i = np_i \geq 5$ and $k \geq 5$. For $k < 4$, $E_i \gg 5$.
- *The appropriate number of degrees of freedom will equal the number of cells k less 1 degree of freedom for each independent linear restriction placed upon the observed cell counts*. For example, one linear restriction is present because the sum of the cell counts must equal n; that is

$$n_1 + n_2 + \cdots + n_k = n.$$

- Other restrictions will be introduced for some applications because of the necessity for estimating unknown parameters required in the calculation of the expected all frequencies or because of the method by which the sample is collected.
- When unknown parameters must be estimated in order to compute X^2, a maximum likelihood estimator should be employed. The degrees of freedom for the approximating chi-square distribution will be reduced by 1 for each parameter estimated. These cases will arise as we consider various practical examples.

3 A Test of a Hypothesis Concerning Specified Cell Probabilities

A Test of a Hypothesis Concerning Specified Cell Probabilities

- The simplest hypothesis concerning the cell probabilities: $H_0 : p_1 = p_1^{(0)}, \ldots, p_k = p_k^{(0)}$, where $p_i^{(0)}$ denotes a specified value for p_i.
- The alternative is the general one that states that at least one of the equalities does not hold: $H_1 : \exists j \in \{1, \ldots, k\}$ such that $p_j \neq p_j^{(0)}$.
- Because the only restriction on the observations is that $\sum_{i=1}^{k} n_i = n$, the X^2 test statistic will have approximately a χ^2 distribution with $k-1$ degrees of freedom.

Examples

Example 2. We want to check if a die is fair. This means that $p = P$ (any one number) $= \frac{1}{6}$. Suppose we decide to roll the die 60 times. If the die is fair, we expect that each number $1, 2, \ldots, 6$ should appear approximately $\frac{1}{6}$ of the time (that is, 10 times). It is roled from a cup onto a smooth flat surface 60 times and the frequency recorded in the table:

Number	1	2	3	4	5	6
Occurences	7	12	10	12	8	11

Solution. $H_0 : p_1 = p_2 = \cdots = p_6 = \frac{1}{6}$; $H_1 : \exists\, i_0 \in \{1, \ldots, 6\}$ s.t. $p_{i_0} \neq \frac{1}{6}$.

Rejection region $RR = (\chi^2_{5,0.95}, \infty) = (11.0705, \infty)$

The value of test statistic is

$$\chi^{2*} = \frac{(7-10)^2}{10} + \frac{(12-10)^2}{10} + \frac{(10-10)^2}{10}$$

$$+ \frac{(12-10)^2}{10} + \frac{(8-10)^2}{10} + \frac{(11-10)^2}{10} = 2.2$$

Decision: Fail to reject H_0 (χ^*_2 is not in RR). See `catego/dice.pdf` □

Example 3. The Mendelian theory of inheritance claims that the frequencies of round and yellow, wrinkled and yellow, round and green, and wrinkled and green will occur in the ratio $9 : 3 : 3 : 1$ when two specific varieties of peas are crossed. In testing this theory, Mendel obtained frequencies of 315, 101, 108 and 32 respectively. Do these sample data provide sufficient evidence to reject this theory, at the 0.05 level of significance?

Solution. H_0: the ratio of inheritance is 9 : 3 : 3 : 1 or

$$H_0 : p_1 = \frac{9}{16}, \quad p_2 = \frac{3}{16}, \quad p_3 = \frac{3}{16}, \quad p_4 = \frac{1}{16}$$

$$\alpha = 0.05, \quad k = 4, \quad df = 3, n = 556$$

Expected frequencies:

$$E = (np_i) = [312.75, 104.25, 104.25, 34.75]$$
$$RR = (7.81, \infty)$$

Statistic:

$$\chi^{2*} = \sum \frac{(n_i - E_i)^2}{E_i} = 0.47$$

Decision: Fail to reject H_0. Conclusion: There is not sufficient evidence to reject Mendel's theory. See `catego/Mendel.pdf` □

4 Contingency Tables

4.1 Testing independence

Testing Independence

- We wish to investigate a *dependency* (or contingency) between two classification criteria.
- Examples
  - we might classify a sample of people by gender and by opinion on a political issue in order to test the hypothesis that opinions on this issue are independent of gender.
  - we might classify patients suffering from a certain disease according to the type of medication and the rate of recovery in order to see if recovery rate depends upon the type of medication
- The input data (counts) are presented in a *contingency table*

$$\begin{array}{cccc|c}
n_{11} & n_{12} & \cdots & n_{1c} & n_{1\cdot} \\
n_{21} & n_{22} & \cdots & n_{2c} & n_{2\cdot} \\
\vdots & \vdots & \ddots & \vdots & \vdots \\
n_{r1} & n_{r2} & \cdots & n_{2c} & n_{r\cdot} \\
\hline
n_{\cdot 1} & n_{\cdot 2} & \cdots & n_{\cdot c} & n_{\cdot\cdot}
\end{array}$$

- Let n_{ij} denote the observed frequency in row i and column of the contingency table and let p_{ij} denote the probability of an observation falling this cell.
- The null hypothesis: the two classification factors are independent

$$H_0 : p_{ij} = p_{i\cdot} p_{\cdot j}, \qquad i = 1, \ldots, r, j = 1, \ldots, c.$$

- *If observations are independently selected*, then the cell frequencies have a multinomial distribution, and the maximum-likelihood estimator for p_{ij} is

$$\widehat{p}_{ij} = \frac{n_{ij}}{n}, \quad i = \overline{1,r},\ j = \overline{1,c}.$$

- Viewing row i as a single cell, the probability for row i is given by p_i and hence

$$\widehat{p}_{i\cdot} = \frac{n_{i\cdot}}{n}$$

- Analogously,

$$\widehat{p}_{\cdot j} = \frac{n_{\cdot j}}{n}$$

- Under the null hypothesis, the maximum-likelihood estimator to the expected value of n_{ij} is

$$E\left(\widehat{n}_{ij}\right) = n(\widehat{p}_{i\cdot}\widehat{p}_{\cdot j}) = n\frac{n_{i\cdot}}{n}\frac{n_{\cdot j}}{n} = \frac{n_{i\cdot}n_{\cdot j}}{n}.$$

  This can be interpreted as distributing each row total according to the proportions in each column (or vice versa) or as distributing the grand total according to the products of the row and column proportions.

- The test statistic is

$$X^2 = \sum_{i=1}^{r}\sum_{j=1}^{c}\frac{[n_{ij} - E\left(\widehat{n}_{ij}\right)]^2}{E\left(\widehat{n}_{ij}\right)} = \sum_{i=1}^{r}\sum_{j=1}^{c}\frac{\left(n_{ij} - \frac{n_{i\cdot}n_{\cdot j}}{n}\right)^2}{\frac{n_{i\cdot}n_{\cdot j}}{n}}.$$

- The degrees of freedom associated with a contingency table possessing r rows and c columns is given by

$$df = \begin{cases} r-1, & \text{if} \quad c = 1 \\ c-1, & \text{if} \quad r = 1 \\ (r-1)(c-1), & \text{if} \quad r \neq 1 \wedge c \neq 1 \end{cases}$$

  You will recall that the number of degrees of freedom associated with the χ^2 statistic will equal the number of cells (in this case, $k = rc$) less 1 degree of freedom for each independent linear restriction placed upon the observed cell frequencies; 1 for $\sum_i \sum_j n_{ij} = n$ and $c-1$ and $r-1$ for columns and rows probabilities.

Example

Example 4. Suppose that we wish to classify defects found on furniture produced in a manufacturing plant according to (1) the type of defect and (2) production shift. A total of $n = 309$ furniture defects was recorded and the defects were classified as one of the four types A, B, C or D. At the same time each piece of furniture was identified according to the production shift in which it was manufactured. These counts are presented in Table 1 (Number in parantheses are the estimated expected cell frequencies). Our objective is to test the null hypothesis that type of defect is independent of shift against the alternative that the two categorization schemes are dependent.

Solution

- The estimated expected cell frequencies for our example are shown in parantheses in Table 1. For example

$$E(\widehat{n}_{11}) = \frac{n_{1\cdot}n_{\cdot 1}}{n} = \frac{94 \cdot 74}{309} = 22.51.$$

	Type of Defect				
Shift	*A*	*B*	*C*	*D*	Total
1	15(22.51)	21(20.99)	45(38.94)	13(11.56)	94
2	26(22.99)	31(21.44)	34(39.77)	5(11.81)	96
3	33(28.50)	17(26.57)	49(49.29)	20(14.63)	119
Total	74	69	128	38	309

Table 1: A contingency table

- The value of the test statistic is

$$
\begin{aligned}
X^2 &= \sum_{i=1}^{3}\sum_{j=1}^{4} \frac{\left(n_{ij} - \frac{n_{i\cdot}i_{\cdot j}}{n}\right)^2}{\frac{n_{i\cdot}n_{\cdot j}}{n}} \\
&= \frac{(15-22.51)^2}{22.51} + \cdots + \frac{(20-14.63)^2}{14.63} = 19.17.
\end{aligned}
$$

- In our case $df = (4-1)(3-1) = 6$.

- Since $P = 1 - F_6(19.17) < 0.05$, exist a dependence between defect type and manufacturing shift. See *catego/furniture. pdf*

4.2 Tables with Fixed Row or Column Total

Tables with Fixed Row or Column Total

- There exists methods of collecting data that may not meet the requirement of a multinomial experiment. For example, due to chance one category could be completely missing.

- We might decide beforehand to interview a specified number of people in each column or row category, thereby fixing the column or row total in advance. (We actually are testing the equivalence of several binomial distributions).

- In such a case the null hypothesis is, say, for fixed column total

$$H_0 : p_1 = p_2 = \cdots = p_c.$$

- It can be shown that the resulting X^2 statistic will possess a probability distribution in repeated sampling that is approximated by a χ^2 distribution with $(r-1)(c-1)$ dfs.

	Ward				
Opinion	1	2	3	4	Total
Favor A	76(59)	53(59)	59(59)	48(59)	236
Do not favor A	124(141)	147(141)	141(141)	152(141)	564
Total	200	200	200	200	800

Table 2: Data tabulation for example 5

Example

Example 5. A survey of voter sentiment was conducted in four mid city political wards to compare the fraction of voters favoring candidate A. Random sample of 200 voters were polled in each of the four wards, with results as shown in Table 2. Do the data present sufficient evidence to indicate that the fraction of voters favoring candidate A differ in the four wards?

Solution

$H_0 : p_1 = p_2 = p_3 = p_4$ that is the fraction p of voters favorizing A is the same for all four wards.

The maximum likelihood estimate (combining the results from all four samples) for the common value of p is $\widehat{p} = 236/800 = r_{1\cdot}/n$.

The expected number of individuals who favor candidate A in Ward 1 is $E(n_{11}) = 200p$ which is estimated by

$$\widehat{E(n_{11})} = 200\widehat{p} = 200 \cdot 236/800 = n_{\cdot 1}n_{1\cdot}/n.$$

The estimated call frequencies are given in parantheses in table 2.

We see that

$$X^2 = \sum_{i=1}^{2}\sum_{j=1}^{4} \frac{[n_{ij} - \widehat{E(n_{ij})}]}{\widehat{E(n_{ij})}} = 10.72.$$

The critical value χ^2 for $\alpha = 0.05$ and $(r-1)(c-1) = 3$ degrees of freedom is 7.81. Because X^{2*} is in the rejection region we conclude that the fraction of voters favoring candidate A is not the same for all four wards. The associated p-value is $p = P(X^2 > 10.72) = 0.013$.

See *catego/exforbin.pdf*

5 Goodness of Fit Tests

5.1 The Chi-square test

The Chi-square goodness of fit test

- Let X be a characteristic having an unknown cdf F. We want to test the null hypothesis $H_0 : F = F_0$ w.r.t the alternative $H_a : F \neq F_0$. For the parametrical variant of this test F_0 depends on unknown parameters.

- If X range is, say, (a, b) and the classes are determined by points $a = a_0 < a_1 < \cdots < a_k = b$, we introduce notations

$$p_i := P(a_{i-1} < X \leq a_i) = F(a_i) - F(a_{i-1})$$

- Let E_i be the event that a randomly chosen individual from our population be in $[a_{i-1}, a_i)$. The null hypothesis, considered above becomes $H_0 : p_i = p_i^{(0)}$, $i = \overline{1,k}$ and the alternative is rewritten as: there exists i_0 such that $p_{i0} \neq p_{i0}^{(0)}$, where

$$p_i^{(0)} = P(a_{i-1} < x \leq a_i | H_0) = F_0(a_i) - F_0(a_{i-1}).$$

- Thus we reduced this test to a chi-square test for proportions.
- If F_0 depends on s unknown parameters, $\theta_1, \theta_2, \ldots, \theta_s$ i.e. $F_0 = F_0(X; \theta_1, \theta_2, \ldots, \theta_s)$ we replace these parameters by their MLE, say $\widehat{\theta_1}, \widehat{\theta_2}, \ldots, \widehat{\theta_s}$. The differences are that

$$\begin{aligned} p_i^{(0)} &= P(a_{i-1} < x \leq a_i | H_0) \\ &= F_0(a_i; \widehat{\theta_1}, \widehat{\theta_2}, \ldots, \widehat{\theta_s}) - F_0(a_{i-1}, \widehat{\theta_1}, \widehat{\theta_2}, \ldots, \widehat{\theta_s}) \end{aligned}$$

for $i = \overline{1,k}$ and chi-square distribution has $k - s - 1$ degrees of freedom.

5.2 The Kolmogorov's Test

The Kolmogorov's Test

- Let X be a continuous characteristic and F its theoretical cdf. We wish to test the null hypothesis $H_0 : F = F_0$ versus one of the alternative
  1. $H_a : F \neq F_0$ (two-tailed test)
  2. $H_a : F > F_0$ (upper-tailed test)
  3. $H_a : F < F_0$ (lower-tailed test)
- The *empirical cdf*

$$\overline{F}_n(x) = \frac{\text{card}\{X \leq n\}}{n}.$$

- Test statistics

$$\begin{aligned} D_n &= \sup_{x \in \mathbb{R}} \{ |\overline{F}_n(x) - F_0(x)| \} \\ D_n^+ &= \sup_{x \in \mathbb{R}} \{ \overline{F}_n(x) - F_0(x) \} \\ D_n^- &= \sup_{x \in \mathbb{R}} \{ F_0(x) - \overline{F}_n(x) \} \end{aligned}$$

Theorem 6 (Valery Ivanovich Glivenko and Francesco Paolo Cantelli)**.**

$$P\left(\lim_{n\to\infty}\sup_{x\in\mathbb{R}}\left|\overline{F}_n(x)-F(x)\right|=0\right)=1.$$

Theorem 7 (Kolmogorov)**.** *If F is continuous the*

$$\lim_{n\to\infty}P\left(\sqrt{n}D_n\le x\right)=\begin{cases}K(x), & \text{if } x\ge 0,\\ 0, & \text{if } x<0,\end{cases}$$

where

$$K(x)=\sum_{k=-\infty}^{+\infty}(-1)^k e^{-2k^2x^2}.$$

- Also,

  $$\begin{aligned}\lim_{n\to\infty}P(\sqrt{n}D_n^+\le x)&=\lim_{n\to\infty}(\sqrt{n}D_n^-\le x)\\ &=K^{\pm}(x)=1-e^{-2x^2},\quad x>0.\end{aligned}$$

  $K^{\pm}$ is called χ-law (not χ^2) with 2 degrees of freedom.

- So, for $\alpha\in(0,1)$ fixed we compute the quantiles $k_{1-\alpha}$ and $k^{\pm}_{1-\alpha}$ of K and $K^{\pm}$, respectively,such that

  $$P(\sqrt{n}D_n\le k_{1-\alpha})=1-\alpha,\text{ i.e. } K(k_{1-\alpha})=1-\alpha,$$

  for a two-tailed test, and

  $$P(\sqrt{n}D_n^+\le k^{\pm}_{1-\alpha})=1-\alpha \text{ and } P(\sqrt{n}D_n^-\le k^{\pm}_{1-\alpha})=1-\alpha$$

  for a one-tailed test.

- As a conclusion, H_0 should be rejected when

  $$\begin{aligned}\sqrt{n}d_n&\ge k_{1-\alpha}\quad\text{for}\quad H_a:F=F_0\\ \sqrt{n}d_n^+&\ge k^{\pm}_{1-\alpha}\quad\text{for}\quad H_a:F>F_0\\ \sqrt{n}d_n^-&\ge k^{\pm}_{1-\alpha}\quad\text{for}\quad H_a:F<F_0\end{aligned}$$

- For a practical implementation we follow a probability based approach.

- It is a good practice to sort the sample values in ascending order: $x_1<x_2<\cdots<x_n$. In this case, for the value of test statistics one gets

  $$\begin{aligned}d_n^+&=\max_{k=\overline{1,n}}\{\overline{F}_n(x_k)-F_0(x_k)\}=\max_{k=\overline{1,n}}\left\{\frac{k}{n}-F_0(x_k)\right\}\\ d_n^-&=\max_{k=\overline{1,n}}\{F_0(x_k)-\overline{F}_n(x_k-0)\}=\max_{k=\overline{1,n}}\left\{F_0(x_k)-\frac{k-1}{n}\right\}\\ d_n&=\max_{k=\overline{1,n}}\{|\overline{F}_n(x_k)-F_0(x_k)|\}=\max\{d_n^+,d_n^-\}\end{aligned}$$

  For grouped data we can employ the test using class limits.

Geometric Interpretation

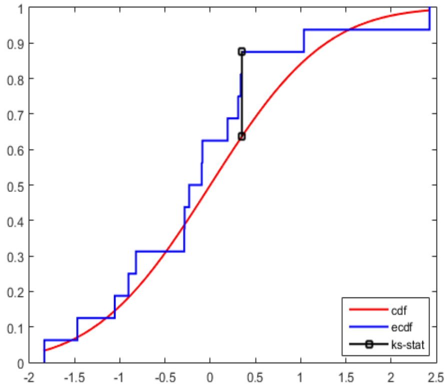


Figure 1: llustration of the Kolmogorov–Smirnov statistic. Red line is CDF, blue line is an ECDF, and the black arrow is the K–S statistic.

5.3 The Kolmogorov-Smirnov Test for Two Samples

The Kolmogorov-Smirnov Test for Two Samples

- X, Y continuous, independent with cdfs F_X, F_Y
- Null hypothesis

$$H_0 : F_X = F_Y$$

- Alternative hypotheses

$$\begin{aligned} H_a : & F_X \neq F_Y \\ & F_X > F_Y \\ & F_X < F_Y \end{aligned}$$

- Test statistics $\sqrt{\frac{n_1 n_2}{n_1+n_2}} D_{n_1,n_2}$, $\sqrt{\frac{n_1 n_2}{n_1+n_2}} D^+_{n_1,n_2}$, $\sqrt{\frac{n_1 n_2}{n_1+n_2}} D^-_{n_1,n_2}$where

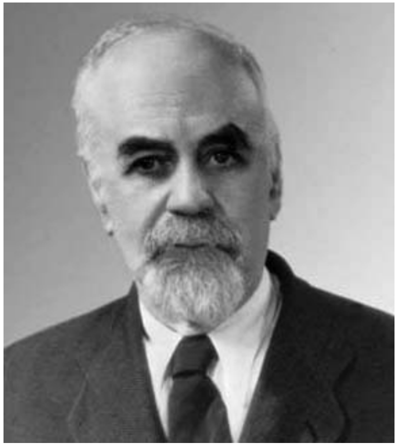
$$\begin{aligned} D_{n_1,n_2} &= \sup_{x\in\mathbb{R}}\{|\overline{F}_X(x) - \overline{F}_Y(x)|\} \\ D^+_{n_1,n_2} &= \sup_{x\in\mathbb{R}}\{\overline{F}_X(x) - \overline{F}_Y(x)\} \\ D^-_{n_1,n_2} &= \sup_{x\in\mathbb{R}}\{\overline{F}_Y(x) - \overline{F}_X(x)\} \end{aligned}$$

- Asymptotic behavior: Kolmogorov's distribution

$$
\begin{aligned}
\lim_{n_1,n_2\to\infty} P\left(\sqrt{\frac{n_1 n_2}{n_1+n_2}} D_{n_1,n_2} \leq x\right) &= \sum_{k=-\infty}^{\infty} (-1)^k e^{-2k^2x^2} \\
\lim_{n_1,n_2\to\infty} P\left(\sqrt{\frac{n_1 n_2}{n_1+n_2}} D^+_{n_1,n_2} \leq x\right) &= 1 - e^{-2x^2} \\
\lim_{n_1,n_2\to\infty} P\left(\sqrt{\frac{n_1 n_2}{n_1+n_2}} D^+_{n_1,n_2} \leq x\right) &= 1 - e^{-2x^2}
\end{aligned}
$$

for $x > 0$.

Andrey Nikolaevich Kolmogorov (1903-1987)



Vladimir Ivanovich Smirnov (1887-1974)

6 References

References

References

[1] Agresti, Alan. 2002. *Categorical Data Analysis*, 2d ed. New York: Wiley-Interscience.

[2] Agresti, Alan. 2007. *An Introduction to Categorical Data Analysis*, 2d ed. New York: Wiley-Interscience.

[3] Dennis D. Wackerly, William Mendenhall III, Richard L. Scheaffer. 2007. *Mathematical Statistics with Applications*, 7th ed., Thompson, Brooks/Cole.