

Non-Parametric Statistics

Introduction

- **Parametric statistics** makes statements about population parameters, that come from well-specified distributions, such as μ, π or σ^2
- The estimators \bar{X}, P , or s^2 , their associated standard errors, sampling distributions (or asymptotic sampling distributions) are used to develop the critical regions, calculate confidence intervals, or p -values (significance or PROB-values).
- **Non-parametric** and/or **distribution free** statistical tests do not use as stringent assumptions about the underlying population distribution and/or ignore population parameters.
- When are non-parametric and/or distribution free methods used?
 - The **sample sizes** are fairly small.
 - If the **distribution of the population** is unknown and assumptions about it are not reasonable.

Note: There are tests to check if the underlying population follows a particular parametric distribution (e.g., goodness-of-fit test or the Kolmogorov-Smirnow test).

- If the random variables are **qualitative**, i.e., measured on the nominal or ordinal scale.
Exception: We know about the exact binomial tests for nominal scaled observations.
- For many parametric tests there are equivalent tests in the non-parametric domain.

Comparison of Parametric and Non-parametric Tests

- If the underlying parametric assumptions are satisfied and quality data are available parametric tests are superior.
- Non-parametric tests may be **conservative**, because they reject the H_0 at a nominal error level substantially less than their parametric counterparts.
- The likelihood of committing a β -error (not rejecting H_0 even though it is incorrect) is smaller for parametric tests if the parametric assumptions are satisfied.

Scope of Application

- Due to the robustness of non-parametric tests, they can be applied to a broader range of underlying populations and measurement scales.
- Remember we always can transform variables on a higher level measurement scale to variables on a lower order (e.g., interval to ordinal scale).

Sample Size

- For small sample sizes n we cannot apply the central limit theorem, when the underlying distribution assumptions are violated.
- For small sample sizes n , non-parametric test statistics have the advantage that their sampling distribution can be derived using combinatorial arguments.
- For larger sample sizes n , non-parametric tests frequently make use of [a] an assumed distribution of the observations, which is different from the normal distribution, and/or [b] a normal approximation by calculating the expected value $E(T|H_0)$ and standard error $\sqrt{Var(T|H_0)}$ of the test statistics T assuming that the null hypothesis is true

For instance,

- a. The counts in a contingency table are assumed to be Poisson distributed with $\lambda = E(X)$ and $\lambda = Var(X)$.

- b. Using the normal approximation of the test statistic: $\frac{T - E(T | H_0)}{\sqrt{Var(T | H_0)}} \sim N(0,1)$, or for a

squared z-transformed test statistic $\frac{(T - E(T|H_0))^2}{Var(T|H_0)} \sim \chi^2_{df=1}$.

- For non-parametric test statistics with an underlying discrete distribution we will ***not be able*** exhaust the nominal significance level α exactly (remember exact binomial test), and we will rather make use of the discrete PROB-values.

Goodness-of-Fit Tests

- Statistical tests often assume that the underlying population follows a particular distribution, like the normal distribution for most parametric test.
- Statistical Goodness-of-Fit tests checks whether the sample observations have been sampled from a hypothetical population.

The Chi-Square Test

- Depending on the measurement scale of the random variable a transformation needs to be applied:
 - If the underlying distribution is continuous, then it needs to re-scaled onto the ordinal scale by building a mutually exclusive and disjunctive partition of the underlying support of the random variable X .
 - If the underlying distribution is discrete one can work with the classified data immediately.

However, one may need to group low populated consecutive classes together to obtain a **critical count of 5 expected** observations.

- The underlying idea is to count the number of sample observations f_i (observed frequencies) of the random variable X within the classes $i \in \{1, \dots, k\}$.

Note: at least two classes are needed to perform the test.

- Assuming that the hypothetical population follows a particular distribution the expected counts within each class are calculated.

Both sets of counts are then compared within the classes.

TABLE 11-13
Structure of a χ^2 Goodness-of-Fit Test

Class	Observed frequency	Expected frequency with H_0	Relative squared difference
1	f_1	F_1	$(f_1 - F_1)^2/F_1$
2	f_2	F_2	$(f_2 - F_2)^2/F_2$
...
k	f_k	F_k	$(f_k - F_k)^2/F_k$
Total:	n	n	$X^2 = \sum_{i=1}^k (f_i - F_i)^2/F_i$

- The test statistic is $\chi^2 = \sum_{i=1}^k z_i^2$ with $z_i^2 = \frac{(f_i - F_i)^2}{F_i}$

- The individual counts follow under H_0 a Poisson distribution: $f_i \sim \text{Poisson}(F_i)$.

The Poisson distribution has the expectation $E(f_i) = F_i$ and the variance $\text{Var}(f_i) = F_i$.

Therefore, $z_i = \frac{f_i - E(f_i)}{\sqrt{\text{Var}(f_i)}}$ follows approximately a standard normal distribution.

- On average this approximation works well for a Poisson distribution if $E(f_i) \geq 5$.
- The square of a standard normal distribution follow a χ^2 -distribution with one degree of freedom, that is, $z_i^2 \sim \chi^2_{df=1}$.
- The sum of χ^2 -distributed random variables is again χ^2 -distributed. The sum's degrees of freedom become the sum of the individual degrees of freedom.
- The total degrees of freedom over all classes intervals are equal to

$$df = \underbrace{\# \text{ of classes}}_k - \underbrace{\# \text{ of estimated parameters}}_{\substack{2 \text{ for } \bar{X} \text{ and } s^2 \text{ of the normal distribution}}} - \underbrace{1}_{\substack{\text{for the constraint } n = \sum_{i=1}^k f_i = \sum_{i=1}^k F_i}}$$

Several degrees of freedom are lost due the used of estimated parameters based on the sample observations to calibrate the hypothetical distribution and the constraint at the overall sample and expected counts need to be equal.

TABLE 10-11
Chi-Square Goodness-of-Fit Test

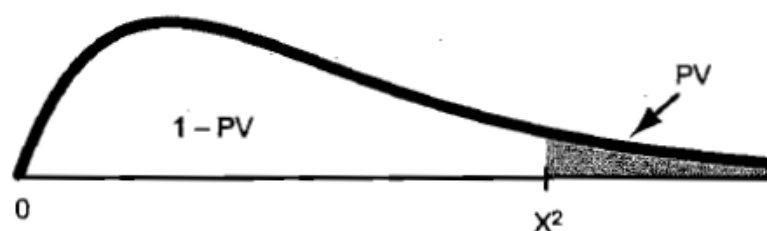
Background
<p>The χ^2 test is used to check if a sample distribution agrees with a theoretical probability distribution. The underlying random variable can be measured at the nominal level or higher. The sample consists of n observations distributed over k categories. Each category has O_j observations, thus $\sum O_j = n$. The theoretical distribution is used to generate expected frequencies for each category in the sample. For example, in testing a sample against the uniform distribution, each sample category would have the same expected frequency, equal to the sample size divided by the number of categories. The expected count for each category is denoted E_j. The test is based on the differences between O_j and E_j. For this test to be reliable, all E_j should exceed 2, and 80% of the E_j should exceed 5.</p>
Hypotheses
<p>Letting Y be the random variable, and $f(Y)$ be the theoretical probability distribution of Y, the null and alternative hypotheses are:</p> <p>H_0: The sample was drawn from a population $f(Y)$ H_A: The sample was drawn from some distribution other than $f(Y)$</p>
Test statistic
<p>Under H_0 the following has an approximate χ^2 distribution</p> $X^2 = \sum_{j=1}^k \frac{(O_j - E_j)^2}{E_j}$ <p>with degrees of freedom $df = k - m - 1$ where m is the number of parameters in $f(Y)$ estimated from the sample.</p>

 PROB-VALUE (PV) and decision rule


Large values of X^2 cast doubt on the truth of H_0 , because they arise rarely when H_0 is true. The PROB-VALUE is

$$PV = P(\chi^2 > X^2)$$

Reject H_0 if PROB-VALUE $< \alpha$



H_0 is rejected when $PV < \alpha$, and we conclude that the sample did not arise from the distribution $f(Y)$. Otherwise (if $PV \geq \alpha$), we fail to reject H_0 and draw no conclusion about the underlying population.

- The test can be set up in  by using the probabilities of each class using the reference distribution.

For instance if we have 4 classes and assume a uniform distribution then the theoretical probabilities become $\pi_1 = 0.25, \pi_2 = 0.25, \pi_3 = 0.25$ and $\pi_4 = 0.25$.

Example:

```
> chi.results <- chisq.test(c(20,29,35,16), p=c(0.25,0.25,0.25,0.25))
> chi.results
```



```
Chi-squared test for given probabilities
data:  c(20, 29, 35, 16)
X-squared = 8.88, df = 3, p-value = 0.03093
> chi.results$expected
[1] 25 25 25 25
```

- Notes:
 - The expected frequencies need to be checked for the low count condition.
 - In this case the degrees of freedom are correct but in other situations they need to be reduced by the number of estimated parameters, which are used to obtain the theoretical distribution.
 - The theoretical probabilities need to sum to one.
 - Explore the online help for the function `chisq.test()`.

Contingency Tables:

- Contingency tables are a cross-tabulation of two (or more) nominal or ordinal scaled variables.
- Row and column sums represent the univariate sample frequency distributions of the variables X_1 and X_2 ,
whereas the cells X_{ij} denote the sample frequency of the joint occurrences of the events

$$X_1 = i \cap X_2 = j$$

		Car ownership				
		0	1	2	3	Total
Household size	2	10	8	3	2	23
	3	7	10	6	3	26
	4	4	5	12	6	27
	5	1	2	6	15	24
Total		22	25	27	26	100

- Contingency tables allow testing the **null hypothesis** whether two nominal or ordinal scaled variables are **jointly statistically independent** of each other against the **alternative hypothesis** that they are statistically related with each other.
- Recall the concept of statistical independence: $\Pr(X_1 = i \cap X_2 = j) = \Pr(X_1 = i) \cdot \Pr(X_2 = j)$
- Therefore, the set hypotheses become: $H_0 : \pi_{ij} = \pi_i^r \cdot \pi_j^c$ against $H_A : \pi_{ij} \neq \pi_i^r \cdot \pi_j^c$ for at least one pair (i, j)
- Construction of the test statistics.
 - At first glance the underlying test statistic could center around the absolute difference between the observed proportion and the expected probabilities: $|p_{ij} - \pi_{ij}|$.

Just small difference – due to sampling variations – would support the null hypothesis.

However, two issues need to be considered:

- The expected probabilities need to be calculated from **estimates** of the marginal relative row and column frequencies:

$$\hat{\pi}_{ij} = \hat{\pi}_i^r \cdot \hat{\pi}_j^c \quad \text{with} \quad \hat{\pi}_i^r = p_i^r = x_{i+}/n \quad \text{and} \quad \hat{\pi}_j^c = p_j^c = x_{+j}/n.$$
 - Just using the expected and observed probabilities ignores that the sampling variability which depends on the **sample size** n .
 - The distribution of $|p_{ij} - \pi_{ij}|$ is unknown and cannot be approximated by the standard normal distribution.
- For these reasons the test statistic is expressed in terms of **observed** and **expected counts** (i.e., frequencies) rather than probabilities:

$$\begin{aligned} E(X_{ij} | H_0) &= n \cdot \hat{\pi}_{ij} = n \cdot p_i^r \cdot p_j^c \\ &= n \cdot \frac{\sum_{j=1}^J x_{ij}}{n} \cdot \frac{\sum_{i=1}^I x_{ij}}{n} = \frac{\sum_{j=1}^J x_{ij} \cdot \sum_{i=1}^I x_{ij}}{n} \\ &= \frac{x_{i+} \cdot x_{+j}}{n} \end{aligned}$$

- X_{ij} are counts and thus again can be assumed to follow a Poisson distribution

$$X_{ij} \sim \text{Poisson}(n \cdot \pi_{ij})$$

with $n \cdot \pi_{ij} = E(X_{ij} | H_0) = \text{Var}(X_{ij} | H_0)$ (because the expectation and variance of a Poisson distribution are identical)

- Summing the squared and standardized differences between the observed and expected frequencies (i.e., squared z-transformed variables) leads to the χ^2 -statistic:

$$\chi^2 = \sum_{i=1}^I \sum_{j=1}^J \frac{[X_{ij} - E(X_{ij} | H_0)]^2}{\text{Var}(X_{ij} | H_0)} = \sum_{i=1}^I \sum_{j=1}^J \frac{[X_{ij} - E(X_{ij} | H_0)]^2}{E(X_{ij} | H_0)}$$

- Degrees of freedom: In total there are $I \times J$ cells in a contingency table. However, since the expected cell frequency are estimated from the row and column sums minus the double counting of the sample size n in total $(I + J - 1)$ degrees for freedom are lost.

$$df = I \times J - (I + J - 1) = (I - 1) \times (J - 1)$$

Once we know the $(I - 1)$ row sums and $(J - 1)$ column sums as well as the total sum n only $(I - 1) \times (J - 1)$ expected cell counts can vary freely.



- Minimum expected cell counts: Approximating summands $\frac{[X_{ij} - E(X_{ij} | H_0)]^2}{\text{Var}(X_{ij} | H_0)}$ by the χ^2 -

distribution is only feasible if $E(X_{ij} | H_0) \geq 5$ for almost all cells.

Just a few cells are allowed to have an expectation $E(X_{ij} | H_0) \geq 2$.

Other books quote different minimum expected counts.

Should the minimum count rules **not** been satisfied then:

- One could switch to *exact* tests
 - Aggregate classes along rows or columns if this is meaningful from a contextual perspective. This will increase the expected count in the aggregated classes.
 - Drop either a class that has a low row or column sums (thus leading to low expected counts)
- Power concerns: As the sample size increase – while keeping all table rates constant – it will become more likely to reject the null hypothesis because the χ^2 statistic will increase. For instance, doubling the cell counts (i.e., $2 \cdot n$) doubles the χ^2 statistic, while leaving its degrees of freedom unchanged. \Rightarrow It becomes more significant.
 -  exercise:
 - Load the **TITANIC** dataset into .
 - Reformat the dataset to individual records rather than aggregate frequencies with
`TitanicIndividual <- as.data.frame(lapply(Titanic, function(x)
rep(x, Titanic$Freq)))`
 - Test whether the survival status is independent of age, sex or class.
 - Redo tests after deviding **TITANIC\$FREQ** by four. Compare to previous test results.

- The χ^2 -test proceeds as follows:

TABLE 11-22

Chi-Square Test for Independence in Contingency Tables

Specification

A total of n observations of two nominal/ordinal variables are cross-classified in a contingency table of dimensions $r \times c$. The observed frequency in the ij th cell is denoted f_{ij} . Expected frequencies for these cells are calculated from

$$F_{ij} = \frac{R_i C_j}{n}$$

Here R_i is the observed frequency count of the i th row, and C_j is the observed frequency count of the j th column.

Hypotheses

$$H_0: \pi_{ij} = \pi_i^r \pi_j^c \quad i = 1, 2, \dots, r; j = 1, 2, \dots, c$$

The variables are statistically independent.

$$H_A: \pi_{ij} \neq \pi_i^r \pi_j^c \quad \text{for at least one } ij \text{ pair}$$

The variables are not statistically independent.

Test statistic

$$X^2 = \sum_{i=1}^r \sum_{j=1}^c \frac{(f_{ij} - F_{ij})^2}{F_{ij}} \quad \text{with } (r-1)(c-1) \text{ degrees of freedom}$$

Decision rule

Reject H_0 if $X^2 > A$, where $A = \chi^2_{1-\alpha, (r-1)(c-1)}$.

