Big Data Fundamentals and Applications

Statistical Analysis (IV)

Reliability & Validity Analyses

Asst. Prof. Chan, Chun-Hsiang

Master program in Intelligent Computing and Big Data, Chung Yuan Christian University, Taoyuan, Taiwan
Undergraduate program in Intelligent Computing and Big Data, Chung Yuan Christian University, Taoyuan, Taiwan
Undergraduate program in Applied Artificial Intelligence, Chung Yuan Christian University, Taoyuan, Taiwan

Outlines

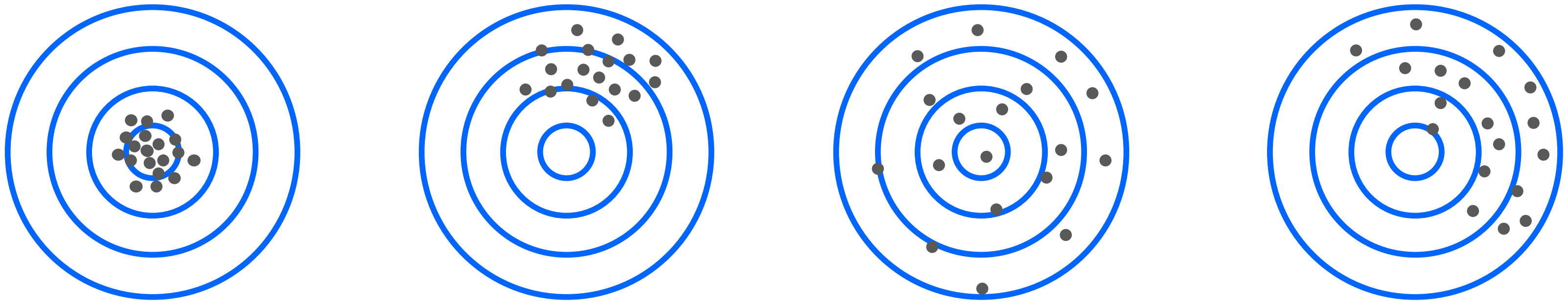
1. Introduction
2. Road Map of Statistical Analysis
3. Hypothesis Testing
4. Type I and Type II Errors
5. Reliability & Validity Analyses
6. Inferential Statistics
7. Test of Normality
8. Differences between Parametric and Nonparametric Statistics
9. Parametric Statistics
10. Nonparametric Statistics
11. Correlation Analysis
12. Question Time

Introduction
Road Map of Statistical Analysis
Hypothesis Testing
Type I and Type II Errors
Reliability & Validity Analyses
Inferential Statistics
Test of Normality
Differences between Parametric and Nonparametric Statistics
Parametric Statistics
Nonparametric Statistics
Correlation Analysis

Reliability & Validity Analyses

Reliability & Validity Analyses

Before we start analyzing our data, you need check the source of your dataset. If the dataset was collected from questionnaire or experiments, then you have to confirm the reliability and validity of each feature or column.



Question 1

Reliability				
Validity				

Reliability & Validity Analyses

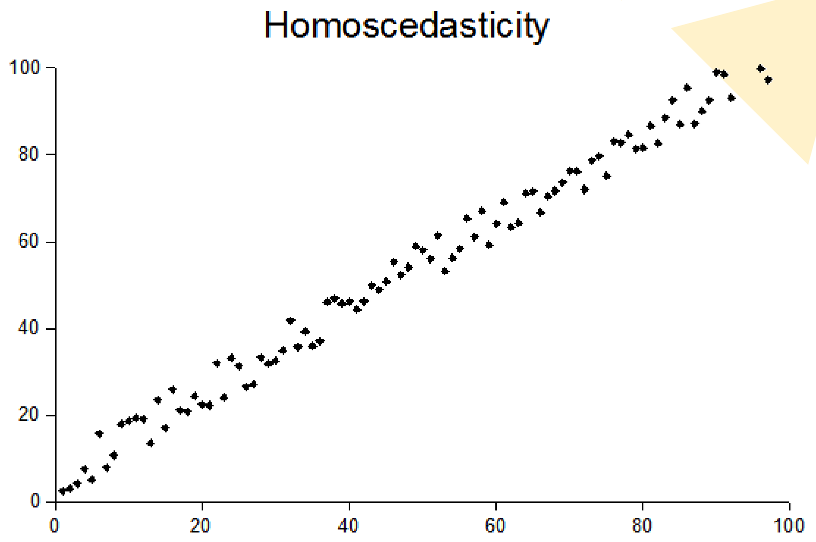
- **Question 2**

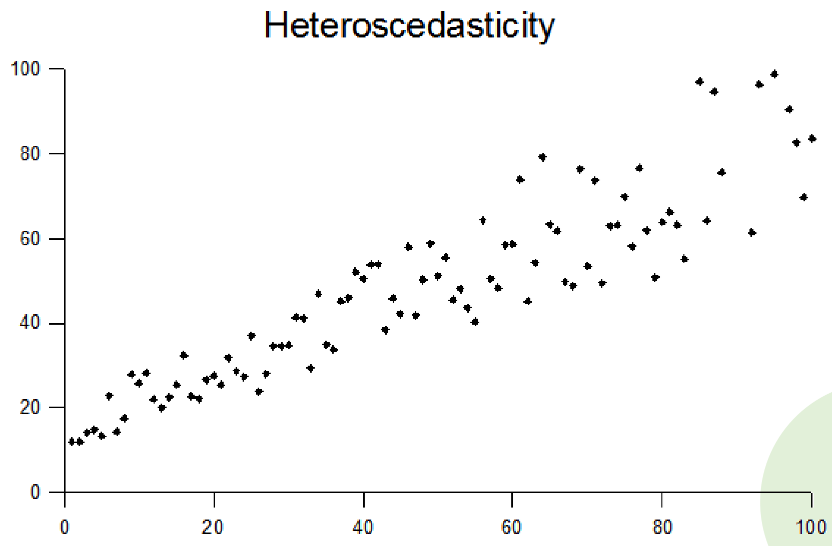
(1) What is the definition of random error and systematic error?

(2) If your data set has large random error and low systematic error, what would you expect to observe from a reliability and validity test?

Homoscedasticity and Heteroscedasticity

- In statistics, a sequence (or a vector) of random variables is **homoscedastic** if all its random variables have the same finite variance. This is also known as **homogeneity of variance**. The complementary notion is called **heteroscedasticity**.





Source: https://en.wikipedia.org/wiki/Homoscedasticity_and_heteroscedasticity

Homoscedasticity and Heteroscedasticity

- Assuming a variable is homoscedastic when in reality it is heteroscedastic results in unbiased but inefficient point estimates and in biased estimates of standard errors, and may result in overestimating the goodness of fit as measured by the Pearson coefficient.
- The existence of heteroscedasticity is a major concern in regression analysis and the analysis of variance, as it invalidates statistical tests of significance that assume that the modelling errors all have the same variance. While the ordinary least squares estimator is still unbiased in the presence of heteroscedasticity, it is inefficient and generalized least squares should be used instead.

Source: https://en.wikipedia.org/wiki/Homoscedasticity_and_heteroscedasticity

Homoscedasticity and Heteroscedasticity

- **Tests in regression**
  - Levene's test
  - Goldfeld–Quandt test
  - Park test
  - Glejser test
  - Brown–Forsythe test
  - Harrison–McCabe test
  - Breusch–Pagan test
  - White test
  - Cook–Weisberg test
- **Tests for grouped data**
  - F-test of equality of variances
  - Cochran's C test
  - Hartley's test
  - Bartlett's test

Source: https://en.wikipedia.org/wiki/Homoscedasticity_and_heteroscedasticity

Reliability

- **Reliability** is the overall consistency of a measure. A measure is said to have a high reliability if it produces similar results under consistent conditions.
- **Inter-rater reliability** assesses the degree of agreement between two or more raters in their appraisals. For example, a person gets a stomach ache and different doctors all give the same diagnosis.
- **Test-retest reliability** assesses the degree to which test scores are consistent from one test administration to the next. Measurements are gathered from a single rater who uses the same methods or instruments and the same testing conditions. This includes intra-rater reliability.

Source: https://en.wikipedia.org/wiki/Reliability_(statistics)

Reliability

- **Inter-method reliability** assesses the degree to which test scores are consistent when there is a variation in the methods or instruments used. This allows inter-rater reliability to be ruled out. When dealing with forms, it may be termed parallel-forms reliability.
- **Internal consistency reliability**, assesses the consistency of results across items within a test.

Source: https://en.wikipedia.org/wiki/Reliability_(statistics)

Reliability – Test-retest Method

- **Test-retest method (再測信度)**
- It directly assesses the degree to which test scores are consistent from one test administration to the next.
  - Administering a test to a group of individuals
  - Re-administering the same test to the same group at some later time
  - Correlating the first set of scores with the second
- The correlation between scores on the first test and the scores on the retest is used to estimate the reliability of the test using the Pearson product-moment correlation coefficient.

Source: https://en.wikipedia.org/wiki/Reliability_(statistics)

Reliability – Parallel-forms Method

- The key to this **parallel-forms method (複本信度)** is the development of alternate test forms that are equivalent in terms of content, response processes and statistical characteristics. For example, alternate forms exist for several tests of general intelligence, and these tests are generally seen equivalent.
- With the parallel test model it is possible to develop two forms of a test that are equivalent in the sense that a person's true score on form A would be identical to their true score on form B. If both forms of the test were administered to a number of people, differences between scores on form A and form B may be due to errors in measurement only.

Source: https://en.wikipedia.org/wiki/Reliability_(statistics)

Reliability – Parallel-forms Method

- It involves:
  - Administering one form of the test to a group of individuals
  - At some later time, administering an alternate form of the same test to the same group of people
  - Correlating scores on form A with scores on form B
- The correlation between scores on the two alternate forms is used to estimate the reliability of the test.
- Problems
  - It may be very difficult to create several alternate forms of a test
  - It may also be difficult if not impossible to guarantee that two alternate forms of a test are parallel measures

Source: https://en.wikipedia.org/wiki/Reliability_(statistics)

Reliability – Split-half Method

- **Split-half method (折半信度)** treats the two halves of a measure as alternate forms. It provides a simple solution to the problem that the **parallel-forms method** faces: the difficulty in developing alternate forms.
- It involves:
  - Administering a test to a group of individuals
  - Splitting the test in half
  - Correlating scores on one half of the test with scores on the other half of the test

Reliability – Split-half Method

- The correlation between these two split halves is used in estimating the reliability of the test. This halves reliability estimate is then stepped up to the full test length using the Spearman–Brown prediction formula.
- Spearman–Brown prediction formula:

$$\rho^*_{xx'} = \frac{n\rho_{xx'}}{1 + (n-1)\rho_{xx'}}, n = 2\ for\ split - half\ method$$

where n is the number of "tests" and $\rho_{xx'}$ is the reliability of the current "test". The formula predicts the reliability of a new test composed by replicating the current test n times (or, equivalently, creating a test with n parallel forms of the current exam).

Reliability – Kuder – Richardson

- **Kuder-Richardson (庫李信度)** is used to measure the reliability for "yes/no" questions.

$$r_{KR_{20}} = \frac{n}{n-1}\left(1 - \frac{\sum_{i=1}^{n} p_i q_i}{S_X^2}\right), q_i = 1 - p_i$$

where S_X^2 is the variance of the total score, p_i is the proportion of people with the correct answer to total people, n is the number of questions.

Reliability – Internal Consistency Cronbach's α

- Cronbach's α is the most common reliability method for **internal consistency (內部一致性信度)**. α value ranges from 0 to 1. The higher α value indicates the higher reliability, as well as, internal consistency. Cronbach's α is suitable for a multi-choice question, not a "yes/no" question.

$$\alpha = \frac{n}{n-1}\left(1 - \frac{\sum_{i=1}^{n} S_i^2}{S_X^2}\right)$$

Where n is the number of questions, S_i^2 is variance associated with each, and S_X^2 is variance associated of the total scores.

Reliability – Internal Consistency Cronbach's α

Cronbach's α	Internal Consistency
$0.9 \leq \alpha$	Excellent
$0.8 \leq \alpha \leq 0.9$	Good
$0.7 \leq \alpha \leq 0.8$	Acceptable
$0.6 \leq \alpha \leq 0.7$	Questionable
$0.5 \leq \alpha \leq 0.6$	Poor
$\alpha < 0.5$	Unacceptable

Validity

- Validity is defined by a ratio of reference variance to total variance (from observation) in a series of measurements.

$$\because obs = ref + e = ref + [e_s + e_r]$$

e, e_s, e_r are total error, systematic error, and random error, respectively.

$$r_{obs,ref}^2 = \frac{S_{ref}^2}{S_{obs}^2} = \frac{S_{ref}^2}{S_{obs}^2 + S_e^2}$$

- $r_{obs,ref}^2$ is the coefficient of validity, S_{ref}^2 is the reference variance, and S_{obs}^2 is the total variance.

Validity – Content Validity

- **Content validity (內容效度)** is the degree to which a test or assessment instrument evaluates all aspects of the topic, construct, or behavior that it is designed to measure.
- Content validity reflects the relevance and representativeness of questionnaire content based on the principle of question distribution:
  - The items measured fall into the measurement field.
  - The items measured cover all perspectives of the measurement field.
  - The proportion of items measured is appropriate.
- Therefore, content validity is also called logical validity (邏輯效度), internal validity (內在效度) and circular validity (循環效度).

Source: https://statisticsbyjim.com/basics/content-validity/

Validity – Content Validity

Table 1 The table added to the cover letter to guide experts for scoring method

Relevancy		**Clarity**	
1	Not relevant	1	Not clear
2	Item need some revision	2	Item need some revision
3	Relevant but need minor revision	3	Clear but need minor revision
4	Very relevant	4	Very clear

Zamanzadeh, V., Ghahramanian, A., Rassouli, M., Abbaszadeh, A., Alavi-Majd, H., & Nikanfar, A. R. (2015). Design and implementation content validity study: development of an instrument for measuring patient-centered communication. Journal of caring sciences, 4(2), 165-178.

[1] Abdollahpour E, Nejat S, Nourozian M, Majdzadeh R. The process of content validity in instrument development. *Iranian Epidemiology.* 2010;6(4):66–74.

Validity – Content Validity

- The item-level content validity index (I-CVI)
  - Based on Abdollahpour et al. (2010), …
  - I-CVI > 0.79 → appropriate
  - 0.70 < I-CVI < 0.79 → revision
  - I-CVI < 0.70 → eliminated
- The scale-level content validity index (S-CVI)
  - **U**niversal **a**greement among experts (S-CVI/**UA**)
  - A less conservative method – **ave**rages the item-level CVIs (S-CVI/Ave)

[1] Abdollahpour E, Nejat S, Nourozian M, Majdzadeh R. The process of content validity in instrument development. *Iranian Epidemiology.* 2010;6(4):66–74.

Validity – Content Validity

- I-CVI & S-CVI

Table 2 The number of experts and its implication on the acceptable cut-off score of CVI

Number of experts	**Acceptable CVI values**	**Source of recommendation**
Two experts	At least 0.80	Davis (1992)
Three to five experts	Should be 1	Polit & Beck (2006); Polit et al. (2007)
At least six experts	At least 0.83	Polit & Beck (2006); Polit et al. (2007)
Six to eight experts	At least 0.83	Lynn (1986)
At least nine experts	At least 0.78	Lynn (1986)

1. **CVI for item (I-CVI) =** item level content validity index the proportion of content experts giving item a relevance of 3 or 4
2. **CVI for scale (S-CVI) =** scale level content validity index

Validity – Content Validity

	Expert									
Questions	#1	#2	#3	#4	#5	#6	#7	#8	#9	#10
Q1	3	4	4	4	3	4	4	4	4	4
Q2	4	4	4	3	4	4	3	4	3	3
Q3	4	2	3	2	3	4	4	3	3	4
Q4	0	0	0	0	0	0	0	0	0	0
Q5	3	4	4	4	4	3	3	4	4	3
Q6	4	4	4	3	3	3	3	3	4	4
Q7	3	4	3	4	4	3	4	3	4	3
Q8	4	4	4	3	4	4	3	3	4	4
Q9	4	3	4	4	4	4	4	4	3	3
Q10	3	4	4	4	4	4	3	3	4	4

1. **Statistics is interesting**
   Relevance Level
   O 1 O 2 O 3 V O 4

Scoring ≥3: agree
Scoring <3: disagree

Validity – Content Validity

	Expert										Expert in Agreement
Questions	#1	#2	#3	#4	#5	#6	#7	#8	#9	#10	
Q1	1	1	1	1	1	1	1	1	1	1	10
Q2	1	1	1	1	1	1	1	1	1	1	10
Q3	1	0	1	0	1	1	1	1	1	1	8
Q4	0	0	0	0	0	0	0	0	0	0	0
Q5	1	1	1	1	1	1	1	1	1	1	10
Q6	1	1	1	1	1	1	1	1	1	1	10
Q7	1	1	1	1	1	1	1	1	1	1	10
Q8	1	1	1	1	1	1	1	1	1	1	10
Q9	1	1	1	1	1	1	1	1	1	1	10
Q10	1	1	1	1	1	1	1	1	1	1	10
Proportion of Relevance	0.9	0.8	0.9	0.8	0.9	0.9	0.9	0.9	0.9	0.9	

Validity – Content Validity

Proportion of Relevance	0.9	0.8	0.9	0.8	0.9	0.9	0.9	0.9	0.9	0.9
Average across 10 experts	0.88									

Expert in Agreement	I-CVI	Universal Agreement	
10	1	1	
10	1	1	
8	0.8	0	
0	0	0	
10	1	1	
10	1	1	
10	1	1	
10	1	1	
10	1	1	
10	1	1	
Sum of I-CVI	8.8	Sum of UA	8
S-CVI Average (Sum of I-CVI/ No. of Questions)	0.88	S-CVI Relevance (Sum of UA/ No. of Questions)	0.8

→ Should be eliminated (<0.7)

→ Appropriate (>0.78)

Validity – Content Validity

(from Abdollahpour et al., 2010)

Although content validity index is extensively used to estimate content validity by researchers, **this index does not consider the possibility of inflated values because of the chance agreement**. Therefore, Wynd *et al*., propose both content validity index and multi-rater kappa statistic in content validity study because, unlike the CVI, it adjusts for chance agreement. Chance agreement is an issue of concern while studying agreement indices among assessors, especially when we place four-point scoring within two relevant and not relevant classes.[7] In other words, **kappa statistic is a consensus index of inter-rater agreement that adjusts for chance agreement**[10] and is an important supplement to CVI **because Kappa provides information about degree of agreement beyond chance.**[7] Nevertheless, content validity index is mostly used by researchers because it is simple for calculation, easy to understand and provide information about each item, which can be used for modification or deletion of instrument items.[6,10]

[1] Abdollahpour E, Nejat S, Nourozian M, Majdzadeh R. The process of content validity in instrument development. *Iranian Epidemiology.* 2010;6(4):66–74.

Validity – Content Validity

- To calculate modified kappa statistic, the probability of chance agreement was first calculated for each item by following formula:

$$P_C = \binom{N}{A} \times 0.5^N = \frac{N!}{(N-A)!\,A!} \times 0.5^N$$

where N is the number of experts in a panel, A is the number of panelists who agree that the item is relevant.

[1] Abdollahpour E, Nejat S, Nourozian M, Majdzadeh R. The process of content validity in instrument development. *Iranian Epidemiology.* 2010;6(4):66–74.

Validity – Content Validity

- Kappa was computed by entering the numerical values of probability of chance agreement (PC) and content validity index of each item (I-CVI) in following formula:

$$\kappa = \frac{I\text{-}CVI - P_C}{1 - P_C}$$

- Evaluation criteria for kappa is the values above 0.74, between 0.60 and 0.74, and the ones between 0.40 and 0.59 are considered as excellent, good, and fair, respectively.

[1] Abdollahpour E, Nejat S, Nourozian M, Majdzadeh R. The process of content validity in instrument development. *Iranian Epidemiology.* 2010;6(4):66–74.
[2] Cicchetti DV, Sparrow SA. Developing criteria for establishing interrater reliability of specific items: applications to assessment of adaptive behavior. *Am J Ment Defic.* 1981;86(2):127–37.

Validity – Criterion-related Validity

- **Criterion validity** (or **criterion-related validity, 校標關聯效度**) evaluates how accurately a test measures the outcome it was designed to measure.
- An **outcome** can be a disease, behavior, or performance.
- Concurrent validity measures tests and criterion variables in the present, while predictive validity measures those in the future.
- To establish criterion validity, you need to compare your test results to **criterion variables**. Criterion variables are often referred to as a “gold standard” measurement. They comprise other tests that are widely accepted as valid measures of a **construct**.

Validity – Criterion-related Validity

- **Criterion validity** (or **criterion-related validity, 校標關聯效度**)
- Usually, criterion validity uses Pearson correlation coefficient to verify the level of validity. The other method is t test to understand the difference between the observations and reference.

Validity – Construct Validity

- **Construct validity (建構效度)** concerns how well a set of indicators represent or reflect a concept that is not directly measurable.
- Construct validity is the appropriateness of inferences made on the basis of observations or measurements, specifically whether a test can reasonably be considered to reflect the intended construct. Constructs are abstractions that are deliberately created by researchers in order to conceptualize the latent variable, which is correlated with scores on a given measure.
- Construct validity is essential to the perceived overall validity of the test.

Source: https://en.wikipedia.org/wiki/Construct_validity

Validity – Construct Validity

- There are six aspects of construct validity in Messick's unified theory of construct validity:
  - **Consequential** – What are the potential risks if the scores are invalid or inappropriately interpreted? Is the test still worthwhile given the risks?
  - **Content** – Do test items appear to be measuring the construct of interest?
  - **Substantive** – Is the theoretical foundation underlying the construct of interest sound?
  - **Structural** – Do the interrelationships of dimensions measured by the test correlate with the construct of interest and test scores?
  - **External** – Does the test have convergent, discriminant, and predictive qualities?
  - **Generalizability** – Does the test generalize across different groups, settings and tasks?

Source: https://en.wikipedia.org/wiki/Construct_validity

Validity – Nomological Validity

- **Nomological validity (學說效度)** refers to the degree to which predictions in a formal theoretical network containing a construct of interest are confirmed. In one sense, the difference between predictive and nomological validity is one of degree and not kind.

Bagozzi R. P. (2010). Structural equation models are modelling tools with many ambiguities: comments acknowledging the need for caution and humility in their use. *J. Consum. Psychol.* 20 208–214. 10.1016/j.jcps.2010.03.001

Box's M Test

- **Box's M test** (also called Box's Test for **Equivalence of Covariance Matrices**) is a parametric test used to compare variation in multivariate samples. More specifically, it tests if two or more covariance matrices are equal (homogeneous). The test is commonly used to test the assumption of homogeneity of variances and covariances in *MANOVA* and *linear discriminant analysis (LDA)*.
- **The null hypothesis for this test is that the observed covariance matrices for the dependent variables are equal across m groups (H_0: $\Sigma_1 = \Sigma_2 = \cdots = \Sigma_m$).**

Sources: https://en.wikipedia.org/wiki/Box%27s_M_test ; https://www.statisticshowto.com/boxs-m-test/

Box’s M Test

- **Box’s M Test is extremely sensitive to departures from normality**; the fundamental test assumption is that your data is multivariate normally distributed.
- **Box’s M has very little power (Cohen, 2008) for small sample sizes**; if your small-sample result is not significant, it doesn’t necessarily indicate that the covariance matrices are equal. The test has also been criticized for being overly sensitive for large sample sizes. To address this particular issue, a smaller alpha level (e.g., 0.001) is recommended (Hahs-Vaughn, 2016).

Sources: https://en.wikipedia.org/wiki/Box%27s_M_test ; https://www.statisticshowto.com/boxs-m-test/

Box's M Test

- Now suppose that $S_1, \ldots, S_m$ are sample covariance matrices from the m populations where each S_j is based on n_j independent observations each consisting of $k \times 1$ column vector (or alternatively a $1 \times k$ row vector).
- Now define S as the pooled covariance matrix

$$S = \frac{1}{n-m}\sum_{j=1}^{m}(n_j - 1)S_j\,, where\ n = \sum_{j=1}^{m} n_j\ define\ the\ following:$$

$$M = (n - G)\ln(|S|) - \sum_{j=1}^{m}(n_j - 1)ln(|S_j|)$$

$$c = \frac{6k^2 + 3k - 1}{6(k+1)(m-1)}\left(\sum_{j=1}^{m}\frac{1}{n_j - 1} - \frac{1}{n-m}\right)$$

Box’s M Test

- **Box’s M Test** uses a **chi-squared** approximation.

Then

$$\Lambda = M \times (1 - C) \sim \chi^2_{k(k+1)(m-1)/2}$$

Where

$$df = \frac{k(k+1)(m-1)}{2}$$

The null hypothesis (of equal covariance matrices) is rejected when $M \times (1 - c) > \chi^2{}_{crit}$ $(or\ P\ value\ < \alpha)$.

Further reading: https://www.youtube.com/watch?v=64CnIq13hBA&ab_channel=Sangwoo.StatisticsKim

Box's M Test

- This estimate works pretty well provided $n_j > 20, m \le 5\ and\ k \le 5$.
- A better estimate can be obtained using the *F* distribution by defining the following:

$$c_2 = \frac{(k-1)(k+2)}{6(m-1)}\left(\sum_{j=1}^{k}\frac{1}{\left(n_j-1\right)^2} - \frac{1}{(n-m)^2}\right)$$

$$df_2 = \frac{df+2}{|c_2 - c^2|}$$

$$a^+ = \frac{df}{1-c-\frac{df}{df_2}}, F = \frac{M}{a^+}; a^- = \frac{df_2}{1-c+\frac{2}{df_2}}, F^- = \frac{df_2 M}{df(a^- - M)},$$

$If\ c_2 > c^2\ \ define\ F = F^+, while\ if\ c_2 < c^2\ \ define\ F = F^-. Then\ F \sim F(df, df_2). The\ null\ hypothesis\ is\ rejected\ if\ F > F_{crit}$

Further reading: https://www.youtube.com/watch?v=64CnIq13hBA&ab_channel=Sangwoo.StatisticsKim

Levene’s Test

- In statistics, **Levene's test** is an inferential statistic used to assess the equality of variances for a variable calculated for two or more groups.
- Some common statistical procedures assume that variances of the populations from which different samples are drawn are equal. Levene's test assesses this assumption.
- It tests **the null hypothesis that the population variances are equal (called homogeneity of variance or homoscedasticity).** Some of the procedures typically assuming homoscedasticity, for which one can use Levene's tests, include analysis of variance and t-tests.

Source: https://en.wikipedia.org/wiki/Levene%27s_test

Levene’s Test

- Levene's test is equivalent to a 1-way between-groups analysis of variance (ANOVA) with the dependent variable being the absolute value of the difference between a score and the mean of the group to which the score belongs (shown below as $Z_{ij} = |Y_{ij} - \overline{Y}_i|$).
- The test statistic, W, is equivalent to the $F_{k-1,N-k}$ statistic that would be produced by such an ANOVA, and is defined as follows:

$$W = \frac{(N-k)}{(k-1)} \cdot \frac{\sum_{i=1}^{k} N_i\,(Z_{i.} - Z_{..})^2}{\sum_{i=1}^{k}\sum_{j=1}^{N_i}\left(Z_{ij} - Z_{i.}\right)^2}$$

k is the number of different groups to which the sampled cases belong.
N_i is the number of cases in the ith group.
N is the total number of cases in all groups.
Y_{ij} is the value of the measured variable for the jth case from the ith group.

$$Z_{ij} = \begin{cases} |Y_{ij} - \overline{Y}_{i.}| & \overline{Y}_i \text{ is a mean of the ith group} \\ |Y_{ij} - \widetilde{Y}_{i.}| & \widetilde{Y}_i \text{ is a medjan of the ith group} \end{cases}$$

$Z_{i.} = \frac{1}{N_i}\sum_{j=1}^{N_i} Z_{ij}$ is the mean of the Z_{ij} for group i.

$Z_{..} = \frac{1}{N}\sum_{i=1}^{k}\sum_{j=1}^{N_i} Z_{ij}$ is the mean of all Z_{ij}.

Source: https://en.wikipedia.org/wiki/Levene%27s_test

Kaiser-Meyer-Olkin Test (KMO)

- **The Kaiser–Meyer–Olkin (KMO) test is a statistical measure to determine how suited data is for factor analysis.**
- The test measures sampling adequacy for each variable in the model and the complete model. The statistic is a measure of the proportion of variance among variables that might be common variance. The higher the proportion, the higher the KMO-value, the more suited the data is to factor analysis. Its value ranges from 0 to 1.

$$KMO = \frac{\sum\sum_{j\neq k} r_{jk}^2}{\sum\sum_{j\neq k} r_{jk}^2 + \sum\sum_{j\neq k} p_{jk}^2},$$

where r_{jk} is the correlation between the variance in question and another, and p_{jk} is the partial correlation.

Partial Correlation p_{jk}

- When the relationship between two variables may be caused by the third variable, we want to exclude (control) the influence of the third variable, and then observe the relationship between the two variables.
- Null hypothesis (H_0): $r = 0$ two variables have no partial correlation
- Alternative hypothesis (H_1): $r \neq 0$ two variables have partial correlation

$$p_{12.3} = \frac{r_{12} - r_{13} \times r_{23}}{\sqrt{1 - r_{13}^2}\sqrt{1 - r_{23}^2}}$$

r_{12}, r_{13}, r_{23} are the Pearson correlation between two variables.

Kaiser-Meyer-Olkin Test (KMO)

KMO Value (0 ~ 1)	Implication for Factor Analysis
0.9 ≤ KMO	Marvelous
0.8 ≤ KMO ≤ 0.9	Meritorious
0.7 ≤ KMO ≤ 0.8	Middling
0.6 ≤ KMO ≤ 0.7	Mediocre
0.5 ≤ KMO ≤ 0.6	Miserable
0.5 ≤ KMO	Unacceptable

Bartlett’s Test

- **Bartlett’s test** is used to test **homoscedasticity**, referring that if multiple samples are from populations with equal variances.
- **Bartlett's test is sensitive to departures from normality**. That is, if the samples come from non-normal distributions, then **Bartlett's test may simply be testing for non-normality**.
- Levene's test and the Brown–Forsythe test are alternatives to the Bartlett test that are less sensitive to departures from normality.
- Bartlett's test is used to test the null hypothesis, $\boldsymbol{H_0}$ **that all** $\boldsymbol{k}$ **population variances are equal against the alternative that at least two are different**.

Source: https://en.wikipedia.org/wiki/Bartlett%27s_test

Bartlett's Test

- If there are k samples with sizes n_i and samples variances S_i^2 then Bartlett's statistic is

$$\chi^2 = \frac{(N-k)\ln\left(S_p^2\right) - \sum_{i=1}^{k}(n_i - 1)\ln(S_i^2)}{1 + \frac{1}{3(k+1)}\left(\sum_{i=1}^{k}\left(\frac{1}{n_i - 1}\right) - \frac{1}{N-k}\right)},$$

$where\ N = \sum_{i=1}^{k} n_i\ \ and\ S_p^2 = \frac{1}{N-k}\sum_i (n_i - 1)S_i^2\ \ is\ the\ pooled\ estimate\ for\ the\ variance.$

- The test statistics has approximately a χ_{k-1}^2 distribution. Thus, the null hypothesis is rejected if $\chi^2 > \chi_{k-1,\alpha}^2$ (where $\chi_{k-1,\alpha}^2$ is the upper tail critical value for the χ_{k-1}^2 distribution).

Source: https://en.wikipedia.org/wiki/Bartlett%27s_test

Validity Analysis Procedure

- You may follow this analysis procedure…
  1) Descriptive statistics (scale if item deleted)
  2) Correlation matrix
  3) KMO (if KMO > 0.5) and Bartlett’s test of sphericity (if *p value* < 0.05)
  4) Factor Analysis (Principal Component Analysis)

Please see the PCA slides for further details.

Reading

1. Abdollahpour E, Nejat S, Nourozian M, Majdzadeh R. The process of content validity in instrument development. *Iranian Epidemiology.* 2010;6(4):66–74.
2. Cicchetti DV, Sparrow SA. Developing criteria for establishing interrater reliability of specific items: applications to assessment of adaptive behavior. Am J Ment Defic. 1981;86(2):127–37.
3. Polit DF, Beck CT. The content validity index: are you sure you know what's being reported? Critique and recommendations. *Res Nurs Health.* 2006;29(5):489–97. doi: 10.1002/nur.20147.
4. Lynn MR. Determination and quantification of content validity. *Nurs Res.* 1986;35(6):382–5.
5. Davis LL. Instrument review: Getting the most from a panel of experts. *Applied Nursing Research.* 1992;5(4):194–7. doi: 10.1016/S0897-1897(05)80008-4.
6. Polit DF, Beck CT, Owen SV. Is the CVI an acceptable indicator of content validity? Appraisal and recommendations. *Res Nurs Health.* 2007;30(4):459–67. doi: 10.1002/nur.20199.
7. Waltz C, Bausell BR. Nursing research: design statistics and computer analysis. Philadelphia: Davis FA; 1981.
8. Wynd CA, Schmidt B, Schaefer MA. Two quantitative approaches for estimating content validity. *West J Nurs Res.* 2003;25(5):508–18. doi: 10.1177/0193945903252998.

Question Time

If you have any questions, please do not hesitate to ask me.

The End

Thank you for your attention))