A short overview of variable types and how to choose them

Lovik, A.

BJPsych Advances (2026), vol. 32, 28–32 doi: 10.1192/bja.2025.10118

RESEARCH METHODS

A short overview of variable types and how to choose them

Anikó Lovik

Anikó Lovik, PhD, is an assistant professor in the Methodology and Statistics Unit at Leiden University in The Netherlands and is also affiliated to the Unit of Integrative Epidemiology of Karolinska Institute, Stockholm, Sweden. She obtained her PhD in biostatistics from KU Leuven, Belgium.

Correspondence Anikó Lovik. Email: a.lovik@fsw.leidenuniv.nl

First received 28 Mar 2025
Revised 18 May 2025
Accepted 20 May 2025

SUMMARY

This article provides a brief introduction (or recapitulation) of what variable types are and how the choice of the variable type may affect which research questions can be answered and the data analysis. The nine-item Patient Health Questionnaire and a simulated data-set are used as an illustration throughout.

KEYWORDS

Clinical outcome measures; depressive disorders; experiment design; statistical methodology; variable types.

An important, but sometimes overlooked or underestimated, aspect of research design is the choice of variable type or statistical data type. There are several taxonomies or ways to categorise variable types, but a common one divides variables into three distinct categories: (a) nominal or categorical (binary or dichotomous variables with only two categories being a special case), (b) ordinal and (c) numeric (more about each type a bit later). Because the variable type directly affects which statistical methods are applicable, it is important to select the type of each variable carefully when designing a study.

The aim of this short article is to (re-)introduce these variable types, describing their advantages and disadvantages to assist researchers in psychiatry make informed decisions when setting up a new project. To make this article more hands-on, all concepts will be illustrated using the nine-item Patient Health Questionnaire (PHQ-9), a primary care screening tool for depressive symptoms, which is commonly used in research as well as in primary care and clinical settings (Kroenke et al. 2001). Data with 5000 observations have been simulated based on the Omtanke2020 study (Lovik et al. 2023). Boxes 1 and 2 provide more information about the PHQ-9 questionnaire and the Omtanke 2020 study. In the following paragraphs, each variable type is discussed separately, and the information is summarised in Fig. 1.

BOX 1 The Patient Health Questionnaire-9 (PHQ-9)

The nine-item Patient Health Questionnaire (PHQ-9) is a validated screening tool for depressive symptoms and their severity (Kroenke 2001). Each item assesses the frequency of a specific symptom and is measured on a four-point Likert scale (a type of ordinal variable), where respondents indicate how frequently they experience a symptom. The response options are: 0 = not at all, 1 = some days, 2 = more than half of the days, 3 = nearly every day. The even-point scale means there is no neutral answer option (such as 'neither agree nor disagree' for scales ranging from 'completely disagree' to 'completely agree').

To obtain the total score (discrete numeric scale), the nine items are summed, resulting in a score between 0 and 27. These scores can be split into a binary variable using a cut-off at 10 (indicating clinically relevant symptoms or their absence) or an ordinal variable with suggested cut-off points for mild, moderate, moderately severe and severe depression at 5, 10, 15 and 20 points respectively.

Because the binary PHQ-9 variable is created from a total score, it is in fact also an ordinal variable with only two categories. However, from an analytical perspective there is no difference between a purely qualitative binary variable and a binary variable created through dichotomisation (the process of creating a binary variable from an ordinal or numeric variable by splitting the original variable into two categories).

The internal consistency of the PHQ-9 based on the Omtanke2020 study is high ($\alpha = 0.88$) (Lovik et al. 2023) and a PHQ-9 score ≥ 10 has a sensitivity of 88% and a specificity of 88% for major depression (Kroenke et al. 2001).

Nominal or categorical variables

Nominal or categorical variables are variables such as gender, blood type or diagnosis. These variables have two or more distinct categories and often there is a qualitative difference between the categories.

BOX 2 The Omtanke2020 Study

The Swedish Omtanke2020 Study is a cohort study established in June 2020 aiming to assess mental and physical health in Sweden during and after the COVID-19 pandemic. The study, involving more than 28 500 Swedish-speaking adult participants, employs online self-reported survey data (collected monthly during the pandemic and annually since January 2022) and linkage to Swedish National and Health Registers. The study is part of the international COVIDMENT consortium (Unnarsdóttir et al. 2021).

The PHQ-9 simulation in this article uses data from the baseline survey (collected between June 2020 and June 2021), which is described in detail in Lovik et al (2023). A research paper from the Omtanke2020 study in which the PHQ-9 is used in various variable types in distinct analyses can be found in González-Hijón et al (2023), where the standardised total score (treated as a discrete numeric variable) is used to depict seasonal changes over time, and, for descriptive network analysis, a binary variable is employed to assess relative risk of experiencing clinically relevant symptoms. In some supplementary analyses, the ordinal version is utilised for predicting sleep quality outcomes. Generally, the binary version is used to estimate prevalence in epidemiological and public mental health studies, whereas psychological studies more often use the total score.

When entered in a database, the assigned codes in any statistical software are arbitrary. Therefore, to describe these variables we commonly use the frequency (number of observations) of each category and the mode (the most common value). The number of categories is an important choice when creating nominal variables: typically, more categories result in more finely tuned information but also require a larger sample size and may be computationally intensive. Moreover, it is also important to have enough observations in small categories when the categories are unbalanced. For example, if we have a non-binary category for gender, either a large enough sample needs to be collected so enough non-binary individuals are included or we need stratified sampling where we include more samples of minority classes (in the statistical sense).

The left-hand column of Fig. 1 shows the typical questions and analyses for the binary PHQ-9. Typically, we can answer questions about the incidence or prevalence of depressive symptoms with this binary variable as we compare the proportion of individuals with and without depressive symptoms for a specific demographic group or treatment or other exposure of interest. In this case, we obtain the categorical variable by first computing the total score for the numeric PHQ-9 scale and then creating two groups at the validated cut-off of 10. Note that this works in only one direction: we can create a categorical variable from a numeric one but never create a numeric variable from a categorical one. An advantage is that both groups are reasonably large despite the imbalance (about 80% have no depressive symptoms).

Ordinal variables

Ordinal variables also consist of distinct categories, but these categories have a strict order although not necessarily equal distance between the categories. For example, educational level could be defined in a study as 'no completed compulsory education', 'completed compulsory education', 'high school diploma' and 'university/doctoral degree'. Clearly, someone with a high school diploma has a higher educational level than someone who did not complete compulsory education, but we cannot tell how much higher. Because there is no clear distance between the levels (in fact, we do not assume at all that they should be the same), the assigned values are still arbitrary (e.g. one could code the four levels in the educational level variable with 0, 1, 2, 3 but also with 1, 2, 4, 8). For this reason, the mean and standard deviation make no sense here, and ordinal variables are best described with the median (the middle value after ranking from smallest to highest) and interquartile range (range of the middle 50% of the observations), although the mode, minimum, maximum and percentiles in general are also valid and useful. This means that many of the basic methods (such as the *t*-test, analysis of variance (ANOVA), Pearson's correlation and linear regression) should not be applied since the calculations are based on means and standard deviations.

The middle column of Fig. 1 shows the ordinal PHQ-9: with five ordered categories it is more fine-grained than the binary version, distinguishing minimal, mild, moderate, moderately severe and severe depressive symptoms, but the number of categories is limited enough that pairwise comparisons do not result in large tables or too many analyses requiring correction for multiple testing. In fact, many analyses compare categories with a selected baseline category only. Research questions about the severity of depressive symptoms could be answered with this ordinal version, for example by employing multinomial logistic or ordinal regression where we could compare each severity level with those who have minimal symptoms. The size of the categories here ranges from 140 to 2645, but for

PHQ-9 variable type										
Binary variable				Ordinal variable				Discrete numeric		
Yes/No variable	PHQ-9 values	*n* (%)	Descriptive statistics	Categories	PHQ-9 values	*n* (%)	Descriptive statistics	PHQ-9 total score	*n* (%)	Descriptive statistics
No clinically relevant symptoms (coded 0)	0–9	3973 (79.5)	Minimum*: 0 ('No clin. rel. symptoms') Maximum*: 1 ('Clin. rel. symptoms') Mode: 0 ('No symptoms')	Minimal (coded 0)	0–4	2645 (52.9)	Minimum: 0 ('Minimal') Maximum: 4 ('Severe') Mode: 0	0	667 (13.3)	Minimum: 0 Maximum: 27 Mode: 0
								1	526 (10.5)	
								2	557 (11.1)	
								3	490 (9.8)	
								4	405 (8.1)	
				Mild (coded 1)	5–9	1328 (26.6)		5	359 (7.2)	
								6	306 (6.1)	
								7	247 (4.9)	
								8	219 (4.4)	
								9	197 (3.9)	
Clinically relevant symptoms (coded 1)	10–27	1027 (20.5)	1st quartile: ND Median: ND 3rd quartile: ND IQR: ND	Moderate (coded 2)	10–14	613 (12.3)	1st quartile: 0 ('Minimal') Median: 0 ('Minimal') 3rd quartile: 1 ("Mild") IQR: 1	10	164 (3.3)	1st quartile: 2 Median: 4 3rd quartile: 8 IQR: 6
								11	132 (2.6)	
								12	118 (2.4)	
								13	102 (2.0)	
								14	97 (1.9)	
				Moderately severe (coded 3)	15–19	274 (5.5)		15	77 (1.5)	
								16	71 (1.4)	
								17	52 (1.0)	
								18	35 (0.7)	
								19	39 (0.8)	
				Severe (coded 4)	20–27	140 (2.8)		20	33 (0.7)	
								21	33 (0.7)	
			Mean: ND Standard deviation: ND				Mean: ND Standard deviation: ND	22	27 (0.5)	Mean: 5.692 Standard deviation: 5.382
								23	16 (0.3)	
								24	11 0.2)	
								25	5 (0.1)	
								26	3 (0.1)	
								27	12 0.2)	
A few typical research questions that could be answered with each variable type										
'What is the prevalence of depressive symptoms?' *'Can we predict the presence of clinically relevant depressive symptoms?'*				*'Can we predict the severity of depressive symptoms based on a set of predictors?'* *'How is the level of severity of the depressive symptoms associated with which treatment the patient received?'*				*'Is there a difference in the mean depressive symptom scores between treatment and control group?'* *'Is there an association between depressive symptoms and BMI?'*		
A few typical statistical methods for each variable type										
Simple: contingency tables, χ^2-test, McNemar's test, (multinomial) logistic regression *Multiple independent variables: (multinomial) logistic regression*				*Simple: χ^2-test and multinomial logistic regression (ignoring the existing order in the five levels), Mann-Whitney and Wilcoxon tests, Spearman correlation, ordinal regression* *Multiple independent variables: multinomial logistic regression, ordinal regression*				*Simple: t-test, ANOVA, Pearson or Spearman correlation, linear regression* *Multiple independent variables: multiple linear regression, ANCOVA, mediation and moderation analysis*		
*Minimum and maximum are rarely defined for binary variables (usually only when an ordinal or numeric variable has been dichotomised).										

FIG 1 Simulated data example of the nine-item Patient Health Questionnaire (PHQ-9) total score presented as a binary, ordinal or discrete numeric variable. Clin. rel., clinically relevant; IQR, interquartile range; ND, not determined; BMI, body mass index; ANOVA, analysis of variance; ANCOVA, analysis of covariance.

BJPsych Advances (2026), vol. 32, 28–32 doi: 10.1192/bja.2025.10118

most analyses these groups are still large enough.

Numeric variables

Numeric(al) (sometimes called quantitative) variables take numerical values. They come in two categories: continuous (where all values are possible within a certain range, e.g. the exact weight or height of a person) and discrete (where many but not infinitely many values are possible, e.g. the weight in kg or height in cm of a person). Sometimes ordinal variables with five or more categories are also treated as discrete numeric. Depending on the distribution and the statistical method, this may cause only limited bias in the results, especially if the number of categories is at least 11 (Wu and Leung 2017). The PHQ-9 total score is a good example: although actually ordinal and (in the general population) highly skewed towards the left (towards low values, with a tail to the right), in many studies it is treated as a numeric variable.

Skewness is an important property of the distribution of a numeric variable and in cases of positive skewness like that of the PHQ-9, a logarithmic or square root transformation could make the distribution more symmetrical and allow it to meet assumptions. Conversely, if the data are negatively skewed an exponential or square transformation may be useful.

Next to skewness, kurtosis is also an important property of the distribution of numeric variables. Kurtosis describes the 'tailedness' of a distribution: a normal distribution has a kurtosis of zero, whereas negative kurtosis typically means less than expected values towards the tails (and fewer outliers) and positive kurtosis implies more values towards the tails.

An advantage of numeric variables is the wide range of statistical models available to answer nuanced research questions. Many of these models make use of means and standard deviations and make additional assumptions about the distribution of the error terms or linearity (assuming, for example, a linear relationship between outcome and predictor). Having to deal with such assumptions may seem a high price but, on the other hand, these models also have higher power and thus need smaller sample sizes. Depending on the model and the goal of the analysis, violation of the assumptions may or may not be too problematic. For example, if the aim is to obtain accurate predictions of the depression score, the assumptions are less relevant than when the interest is in obtaining precise estimates (James et al. 2021). Having said this, generally it is better to err on the side of caution.

The right-hand column of Fig. 1 shows the numeric PHQ-9, which is clearly positively skewed, with fewer than 10 observations (out of 5000) for scores 25 and 26. It would be very hard to estimate anything for categories with so few values, which is one of the reasons such scores are often treated as numeric. However, treating the PHQ-9 total score as discrete numeric means that we assume that the distance between scores 6 and 7 is the same as the distance between scores 26 and 27 and, more importantly, in many (basic) analyses, we will assume linearity. Yet a numeric variable allows us to compare group means, for example between the treatment and the control group, or to assess the mean change before and after treatment. Additionally, we can employ a variety of methods taught in basic statistics courses to obtain these comparisons, making the analyses easily accessible. An important point here is that we need to have a clear idea about what a meaningful (i.e. clinically relevant) difference is to interpret the results.

In summary, looking at Fig. 1 globally, as we progress from binary (categorical) to numeric variables, we have more finely grained information and more descriptives at our disposal. However, each variable type is best suited to answer slightly different research questions and employs different statistical methods. Therefore, there is no overall 'best' variable type, and the variable type should be selected with care when designing a study.

This brief article gives only an overview of variable types and how the choice of the variable type (in particular, the choice of the outcome variable) affects the formulation of the research question and the statistical analyses. For interested readers, I recommend the short article by Ranganathan & Gogtay (2019) addressing this topic from a different and broader perspective and Nikolaou's guide in this journal for basic statistical analysis (Nikolaou 2016).

Funding

This research received no specific grant from any funding agency, commercial or not-for-profit sectors.

Declaration of interest

None.

References

González-Hijón J, Kähler AK, Frans EM, et al (2023) Unravelling the link between sleep and mental health during the COVID-19 pandemic. *Stress Health*, **39**: 828–40.

James G, Witten D, Hastie T, et al (2021) *An Introduction to Statistical Learning*. Springer.

Kroenke K, Spitzer RL, Williams JB (2001) The PHQ-9: validity of a brief depression severity measure. *Journal of General Internal Medicine*, **16**: 606–13.

Lovik A, González-Hijón J, Kähler AK, et al (2023) Mental health indicators in Sweden over a 12-month period during the COVID-19 pandemic. *Journal of Affective Disorders*, **322**: 108–17.

Nikolaou V (2016) Statistical analysis: a practical guide for psychiatrists. *BJPsych Advances*, **22**: 251–9.

Ranganathan P, Gogtay NJ (2019) An introduction to statistics – data types, distributions and summarizing data. *Indian Journal of Critical Care Medicine*, **23**(suppl 2): s169–70.

Unnarsdóttir AB, Lovik A, Fawns-Ritchie C, et al (2021) COVIDMENT: COVID-19 cohorts on mental health across six nations. *International Journal of Epidemiology*, **51**: e108–22.

Wu H, Leung SO (2017) Can Likert scales be treated as interval scales? – A simulation study. *Journal of Social Service Research*, **43**: 527–32.