Surveys

HCI/PSYCH 522
Iowa State University

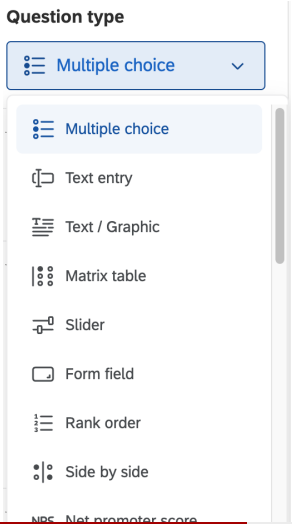
March 29, 2022

Overview

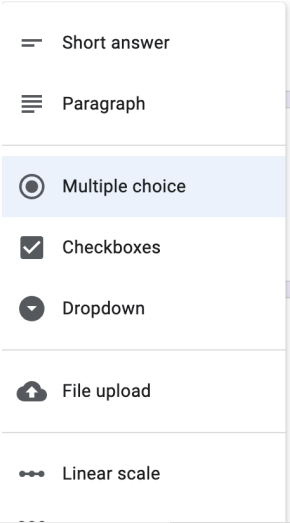
- Question types
  - Multiple choice
  - Multiple answer
  - Likert
  - Numerical
  - Text
- Statistical analysis
  - Binary response
  - Numerical (continuous) response
- Missing data
  - Missing completely at random
  - Missing at random
  - Not missing at random
- Additional considerations

Question types

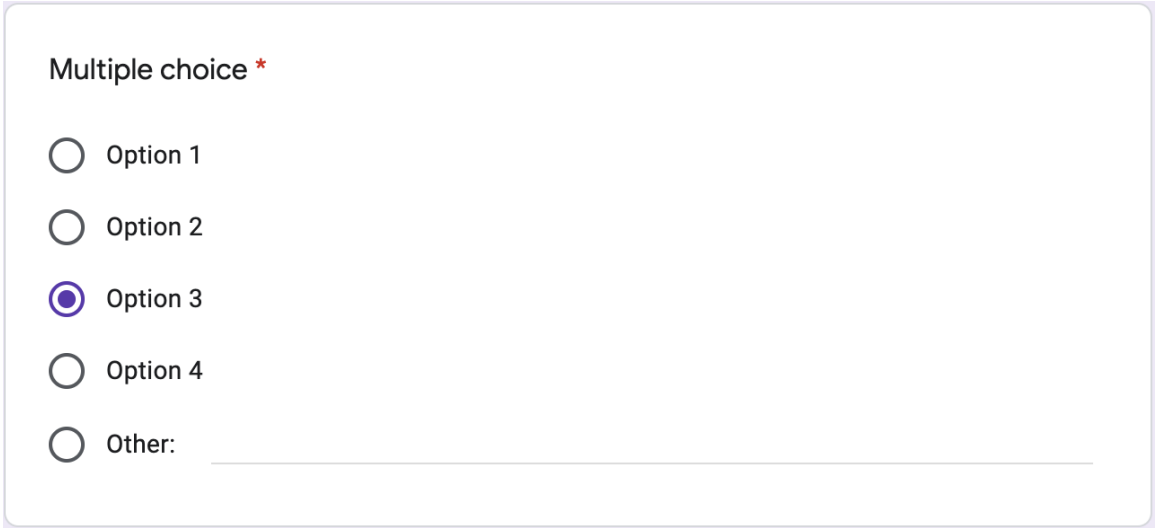
Qualtrics:



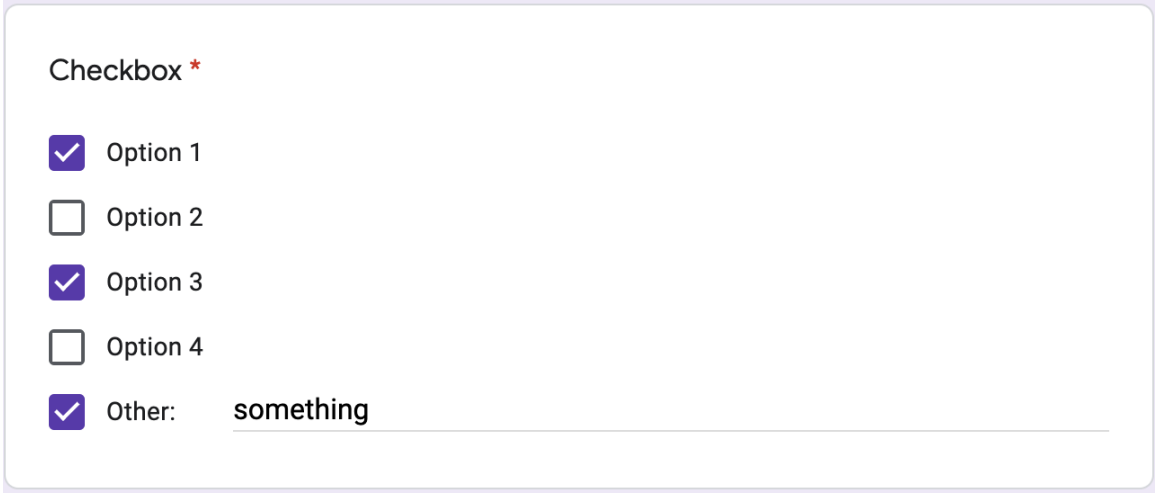
Google forms:



Multiple choice



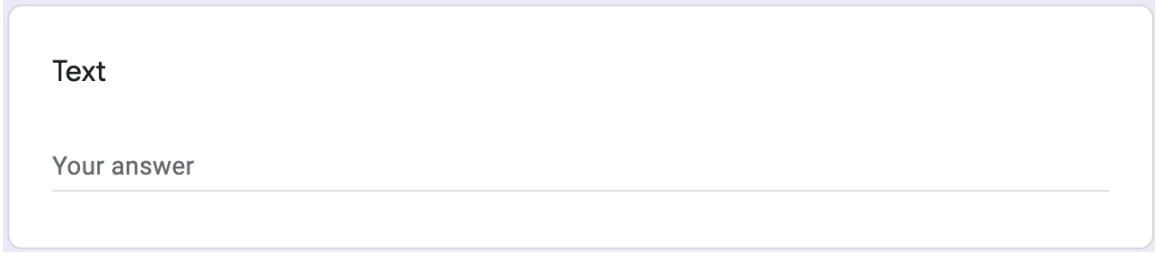
Multiple answer



Likert scale

Numerical

Text



Statistical analyses

- Binary dependent variable
  - Binomial analysis
  - Logistic regression
- Numerical (continuous) dependent variable
  - Normal analysis
  - (Linear) regression
- Numerical (count) dependent variable
  - Poisson analysis
  - Poisson regression

Binary

Construct binary variables from survey responses: Did respondent ...

- choose "option 1" on multiple choice question 1?
- choose "option 1" or "option 2" on multiple choice question 1?
- choose "option 1" on multiple answer question 2?
- choose "option 1" and/or "option 2" on multiple answer question 2?
- choose "Strongly agree" on Likert question 3?
- choose "Strongly agree" or "Agree" on Likert question 3?
- indicate a value greater than X on numerical question 4?
- indicate a value between X and Y on numerical question 4?
- mention Z on text question 5?

Binomial model

What proportion of respondents Agreed or Strongly Agreed with the statement in question 4? Let Y_i be an indicator that respondent i Agreed or Strongly Agreed with the statement in question 4.

Assume

$$Y_i \overset{ind}{\sim} Bin(n, \theta)$$

and construct a credible interval for θ.

Logistic regression model

How does the proportion of people who indicated they Agreed or Strongly Agreed with the statement in question 4 relate to their Salary on question 5?
Let Y_i be an indicator that respondent i Agreed or Strongly Agreed with the statement in question 4. Let X_i be the reported Salary on question 5. Assume a binomial model

$$Y_i \overset{ind}{\sim} Bin(n, \theta_i) \quad \text{and} \quad \log\left(\frac{\theta_i}{1-\theta_i}\right) = \beta_0 + \beta_1 X_i$$

and construct a credible interval for β_1.

Note: probably want to consider using $\log(X_i)$ rather than X_i to reduce the impact of those with large reported Salaries.

Logistic regression model

How does the proportion of people who indicated they Agreed or Strongly Agreed with the statement in question 4 relate to their Gender on question 5?
Let Y_i be an indicator that respondent i Agreed or Strongly Agreed with the statement in question 4. Let X_i be an indicator that Gender was reported as Female. Assume a binomial model

$$Y_i \overset{ind}{\sim} Bin(n, \theta_i) \quad \text{and} \quad \log\left(\frac{\theta_i}{1-\theta_i}\right) = \beta_0 + \beta_1 X_i$$

and construct a credible interval for β_1.

Numerical (continuous) dependent variable

Construct numerical (continuous) variables from survey responses:

- Use responses from numerical answers.
  - Consider taking a logarithm for strictly positive values.
- Convert Likert scale to integers.
  - There are better methods, but this can often be reasonable.

Normal model

What is the average Salary of respondents?
Let Y_i be the Salary response for individual i.

Assume

$$Y_i \overset{ind}{\sim} N(\mu, \sigma^2)$$

and construct a credible interval for μ.

(Linear) regression model

How does Gender affect Salary?
Let Y_i be the Salary response for individual i. Let X_i be an indicator that Gender was reported as Female. Assume a (linear) regression model

$$Y_i \overset{ind}{\sim} N(\beta_0 + \beta_1 X_i, \sigma^2)$$

and construct a credible interval for β_1.

Note: probably want to consider using $\log(Y_i)$ rather than Y_i and interpret $100(e^{\beta_1} - 1)$ as the percent change in salary for women vs men.

(Linear) regression model

How does IQ affect Salary?
Let Y_i be the reported Salary response for individual i. Let X_i be the reported IQ. Assume a (linear) regression model

$$Y_i \overset{ind}{\sim} N(\beta_0 + \beta_1 X_i, \sigma^2)$$

and construct a credible interval for β_1.

Note: probably want to consider using $\log(Y_i)$ rather than Y_i and interpret $100(e^{\beta_1} - 1)$ as the percent change per point increase in IQ in salary.

Missing data

https://stefvanbuuren.name/fimd/sec-MCAR.html

- Missing completely at random (MCAR)
- Missing at random (MAR)
- Not missing at random (NMAR)

Types of missing

- Missing values
- Missing respondents

```
##   SDMVPSU SDMVSTRA WTMEC2YR HI_CHOL race   agecat RIAGENDR
## 1       2       77 37760.57      NA    2   (0,19]        2
## 2       1       89 19399.67      NA    3  (19,39]        1
## 3       2       83 33405.02      NA    3  (19,39]        1
## 4       2       76 22538.87      NA    2 (59,Inf]        1
## 5       1       81 31182.47      NA    1  (19,39]        2
## 6       2       83 48409.15      NA    2   (0,19]        2
```

Missing completely at random

Definition

If the probability of being missing is the same for all cases, then the data are said to be missing completely at random (MCAR).

For example,

- Respondents sit at home at flip a coin to determine if they will respond
- Scale runs out of batteries

If the missing data are MCAR, then you can ignore the missingness and proceed with your analysis. Often called complete-case analysis.

Missing at random (MAR)

Definition

If the probability of being missing is the same only within groups defined by the observed data, then the data are missing at random (MAR).

For example,

- probability of responding depends on county or
- probability of responding to agree/disagree question depends on gender.

You will likely need to model the missingness and incorporate it into your analysis.

Missing not at random (MNAR)

Definition

If neither MCAR nor MAR holds, then we speak of missing not at random (MNAR) or NMAR (not missing at random). MNAR means that the probability of being missing varies for reasons that are unknown to us.

For example,

- probability of responding depends politcal affiliation or
- probability of responding about salary depends on salary.

You will need to make assumptions about the probability of missing and your analysis will depend on those assumptions. These assumptions can only be checked by gather additional data.

Additional considerations

- 10 tips for building effective surveys
- Writing survey questions
- 7 proven practical tips for creating your next questionnaire
- Good practice in the conduct and reporting of survey research