# 4. Survey Data, Scale Reliability & Common Method Bias

- In surveys, we can ask people how they feel, or about their own perceptions about behavior
- This is mostly done through *survey items* that the respondent is asked to evaluate, such as "I am very satisfied with my job"
- In commonly used *Likert-scales*, respondents are asked for their level of agreement on a number of given statements on a scale such as
    1. Strongly disagree
    2. Disagree
    3. Neither agree nor disagree
    4. Agree
    5. Strongly agree
- While practitioners are often tempted to use single items for a certain attitude or behavior, researchers stress the importance to use *multiple items* to assess a phenomenon

# 4.1 Psychological Constructs and Reliability

- Researchers typically use scales with *multiple items* that are supposed to measure certain psychological constructs
- A *psychological construct* is a label for a cluster of covarying behaviors or attitudes (such as job satisfaction, job engagement, but also of personality traits such as conscientiousness, extraversion, etc.)
- Typically
  - item responses are averaged to compute a score
  - the score then represents a person's position on the construct
- Important question: How *reliable* is a scale?
- That is, how consistently does a scale measure the same underlying construct?

**Python**

# Looking at Correlations

- A first step is often to assess the correlation between variables
- We can obtain a correlation matrix with `df.corr()`
  - Note: this can be a huge matrix as it shows the correlation coefficients between all variables in the DataFrame
  - Typically, it makes sense to only show it for a subset of the data
- To do so, we can filter the data frame (which gives us a smaller data frame selected by the filtering criteria)
  - Show correlation between two variables age and tenure:
    ```
    df.filter(items=['age', 'tenure']).corr()
    ```
  - Show correlation matrix for all variables starting with “Satis”:
    ```
    df.filter(regex='Satis*').corr()
    ```

## Your Task

## Analyze Survey Data

- The LPP includes a scale to measure employee engagement, a short version of the Utrecht Work Engagement Scale (Schaufeli et al. (2006)):
  - *At my work, I feel bursting with energy*
  - *At my job, I feel strong and vigorous*
  - *I am enthusiastic about my job*
  - *My job inspires me*
  - *When I get up in the morning, I feel like going to work*
  - *I feel happy when I am working intensely*
  - *I am proud of the work that I do*
  - *I am immersed in my work*
  - *I get carried away when I'm working*
- The response scale ranges from 1 "every day" to 5 "never"
- The respective 9 item variables in the data set start with menga
- Print the correlation matrix for these variables
- Save the notebook

# Assessing Reliability: Classical Test Theory in Psychology

- Suppose now we have $i = 1, \ldots, k$ items $X_i$ that measure a construct
- The observed score is $\bar{X} = \frac{1}{k}\sum_{i=1}^{k} X_i$
- Basic assumptions:
  - Response to a survey item = sum of "true score" $T$ and some noise
  $$X_i = T + \varepsilon_i$$
  - The noise terms $\varepsilon_i$ are independent and identically distributed
- The so-called reliability of the scale is often defined as
$$\frac{V[T]}{V[\bar{X}]}$$
- That is: A scale is reliable when much of the variance in the observed score $\bar{X}$ is due to variance in the true scores $T$

→ When a scale is more reliable we have a smaller *measurement error* and a weaker *attenuation bias* when using the construct's score in a regression

# The Length of a Scale

Note that we can write

$$\frac{V[T]}{V[\bar{X}]} = \frac{V[T]}{V[\frac{1}{k}\sum_{i=1}^{k}(T+\varepsilon_i)]}$$

$$= \frac{\sigma_T^2}{V[T+\frac{1}{k}\sum_{i=1}^{k}\varepsilon_i]} = \frac{\sigma_T^2}{V[T]+V[\frac{1}{k}\sum_{i=1}^{k}\varepsilon_i]}$$

$$= \frac{\sigma_T^2}{\sigma_T^2+\frac{1}{k^2}k\sigma_\varepsilon^2} = \frac{\sigma_T^2}{\sigma_T^2+\frac{1}{k}\sigma_\varepsilon^2}$$

Therefore:

- The reliability thus increases in the length of the scale $k$!
- Intuition: When responses to single items are noisy and this noise is independent, then the noise terms tend to “cancel each other out”
- The measurement error should be reduced when scales are longer

But how do we assess the reliability when we do not know $T$?

# The Reliability of Scales: Cronbach's Alpha

- Consider again the reliability coefficient $\frac{V[T]}{V[\bar{X}]}$
- For any two items, we have that

$$Cov[X_1, X_2] = Cov[T + \varepsilon_1, T + \varepsilon_2] = V[T]$$

- Now estimate $V[T]$ by the mean of all covariances between any two items:

$$\overline{\sigma_{ij}} = \frac{1}{k(k-1)} \sum_{i=1}^{k} \sum_{j \neq i}^{k} Cov[X_i, X_j]$$

- The ratio $\alpha = \frac{\overline{\sigma_{ij}}}{V[\bar{X}]}$ is called **Cronbach's alpha**

**Note:**

- $\alpha$ is a very frequently applied measure for the *internal consistency* of a scale
- Scale is considered to have a good internal consistency if $\alpha > 0.8$

Python

## Computing Item Averages and Standardizing

- To obtain the score for the scale we typically compute the average across all items of the scale
- In Python we can do that for instance (say we have four items measuring satisfaction called satis1, … , satis4)
  - by "manually" summing up the items and averaging:
    `df['satis'] = (df.satis1+df.satis2+… ) / 4`
  - or averaging across all columns of a filtered DataFrame:
    `df['satis'] = df.filter(regex='satis*').mean(axis=1)`
  - Note: the method mean returns the mean of the values either over rows/observations (axis=0) or columns/variables (axis=1)
- Frequently, scores are *standardized* $X_{STD} = \frac{X - m_X}{\sigma_X}$ where $m_X$ is the mean and $\sigma_X$ the standard deviation of $X$
- We can do that for instance by
  `df['sat_std'] = (df.satis-df.satis.mean())/df.satis.std()`

Python

## Computing Cronbach's alpha

- We can use method cronbach_alpha **from package** pingouin
  - **To so we must first install** pingouin **with** !pip install pingouin –q
  - **Then we can** import pingouin as pg
  - **You call the function with** pg.cronbach_alpha(data=df)
- Or we can define our own function:

```
def cronbach(data):
  k = data.shape[1]
  varX = data.sum(axis=1).var()
  sumVar = data.var(axis=0).sum()
  return k / (k-1) * (1 - sumVar/varX)
```

- Note: the DataFrame you pass to either function must only consist of the variables of the specific scale, you can generate such a DataFrame with df.filter(regex='menga*')

Your Task

## Estimate the Reliability of a Scale

- Please open again the notebook LPPanalysis.ipynb used to look at the engagement data
- Generate a new variable enga for the mean engagement score
- Note: As the variable is coded, low values indicate high engagement. To avoid later confusion, it makes sense to reverse the scale (you can do that by simply stating df['enga'] = 6 - df['enga']
- Also generate a standardized version of the variable (call it enga_std)
- What is the value of Cronbach’s alpha? To what extent is the engagement scale internally consistent?

# 4.2 Common Method Bias

- The noise terms of different items in a survey will often likely be correlated due to factors beyond a common true score of a construct

- This can for instance be due to: respondents‘ different …
  - … tendencies to reply in a socially acceptable manner (social desirability)
  - … tendencies or more or less consistently (consistency motif)
  - … personality traits or cognitive abilities
  - … levels of fatigue or mood when filling out the survey…

- We must thus be aware that the $\varepsilon_i$ will be correlated beyond their connection through some underlying “true score“ $T$

- Importantly:
  This will also affect *separate* constructs elicited in the same survey

- Particularly problematic *when dependent and independent variable of a regression are collected in same survey*
  - Then you tend to overestimate the association between the constructs
  - This phenomenon is often called *common method bias*

- Hence: Be careful when you use different constructs from the same survey in a regression!
  - When regressing one psychological construct on another one measured in the same survey (for instance regress job satisfaction on engagement) you likely overestimate their true association
  - Less problematic for factual survey items where respondents‘ assessments are less subjective