

# جمع آوری یا تولید داده (اولین قدم در علم داده)

جواد وحدت

[vahdatjavad@gmail.com](mailto:vahdatjavad@gmail.com)

یکی از مهم ترین مراحل در روش تحقیق، جمع آوری داده ها یا Data Collecting است

۹۱ درصد زمان هر محقق صرف جمع آوری داده می شود!

داده ارزشمندترین ابزار در هر کسب و کاری هست.

# تفاوت داده با اطلاعات



# روش‌های گردآوری داده‌ها

- منابع دست اول و دست دومی

**منابع دست اول:** مستقیماً در ارتباط با حادثه یا پدیده بدست می‌آیند. ممکن است به شکل کتبی/شفاهی/تصویری/مادی و ... باشند.

**منابع دست دوم:** به طور غیر مستقیم در ارتباط با حادثه قرار دارند و به اتکای منابع دست اول تهیه شده اند. ممکن است به صورت نقل قول مستقیم یا غیر مستقیم/کتبی یا شفاهی و .. باشند.

- بررسی داده های موجود
- روش مشاهده
- روش مصاحبه
- روش پرسش نامه

- روش استفاده از داده های موجود:

داده هایی که قبلاً جمع آوری شده است استفاده می نماید. یکی از منابع چنین داده -  
هایی پرونده های بیمارستانی هستند. این روش در مطالعات گذشته نگر اهمیت زیادی  
دارد.

تمام فرم های اطلاعاتی باید یک فیلد به نام شماره شناسایی داشته باشند.

- روش مشاهده:

- - اولین و قدیمی ترین راه جمع آوری داده ها مشاهده است.
- - بیشتر در علوم انسانی کاربرد دارد.
- - مشاهده گر، مستقیماً، جامعه پژوهش را مورد مشاهده قرار می دهد.
- - بهتر است خودتون رو معرفی نکنین تا خطای کمتری داشته باشد (مشاهده غیرمستقیم)
- - این روش معمولاً بسیار هزینه بر هست.

## روش مصاحبه:

- - پژوهشگر می‌تواند پرسش‌های خود را بپرسد. این روش با نام مصاحبه شناخته شده است.
- - دلیل انعطاف در پرسش، از بهترین روش‌های جمع‌آوری اطلاعات می‌باشد.
- - مصاحبه ساختارمند (پرسشنامه) و مصاحبه غیرساختارمند.
- - مصاحبه غیرساختارمند در تحقیقات کمی علوم پزشکی کاربرد چندانی ندارد و به شدت تحت تأثیر مهارت مصاحبه‌گر است.
- - استفاده از ضبط صوت روش مناسب و شایعی است.
- - محدودیت‌هایی نیز دارد. برای مثال حضور، جنسیت، ظاهر و .. مصاحبه‌گر. برای مثال اگر مصاحبه‌کننده در رابطه با روش پیشگیری از بارداری افراد از آن‌ها سوال کند، جنسیت مصاحبه‌کننده بر صحت پاسخ‌ها تأثیر زیادی دارد. طولانی شدن زمان جمع‌آوری داده‌ها است. از طرفی تربیت مصاحبه‌گر حرفه‌ای، کاری دشوار است.

## روش پرسش نامه :

زمانی که حجم نمونه مطالعه بالا باشد و یا دسترسی به پاسخ دهندگان مشکل باشد.

ساختار پرسش نامه:

- بخش راهنما
- سوالات دموگرافیک
- سوالات اصلی



## انواع پرسش:

- پرسش های تشریحی (باز)
- پرسش های چند گزینه ای (بسته)

مراحل تهیه پرسش نامه:  
اگر پرسش نامه استاندارد موجود باشد بهتر است از آن استفاده شود.

- تعیین هدف
- رسم فلوچارت
- تهیه بانک سؤالات
- تعیین روایی و پایایی سؤالات
- انتخاب نهایی سؤالات

# روایی ابزار سنجش

• **روایی محتوا :** میزان تحت پوشش قرار دادن جنبه های مختلف متغیر مورد سنجش توسط سؤالات پرسش نامه است.

برای تعیین اعتبار محتوا معمولاً از **روش هامیلتون** استفاده می شود. (حداقل ۲ فرد خبره - به هر سوال به ۴ درجه (بی ارتباط - کم ارتباط - مرتبط - بسیار مرتبط) بیان کند)

• **سازگاری درونی:** اگر بخواهیم برای سنجش یک متغیر از چند سوال استفاده کنیم می بایست تمامی آن سوالات موید یکدیگر باشند.

• روش دونیم کردن

• روش آلفا کرونباخ

• روش کودر ریچاردسون

## پایایی ابزار سنجش

- درجه یکسانی و ثبات نتایج در تکرار یک سنجش (آزمون – دوباره آزمون)

### سنجش نگرش

- مقیاس لیکرت : کاملاً موافق، موافق، بی نظر، مخالف، کاملاً مخالف
- مقیاس گوتمن: موافقت یا مخالفت (سپس به هر فرد به تعداد عبارت هایی که با آنها موافق بوده نمره داده می شود.)

## Government and global data

- [www.data.gov](http://www.data.gov) : From science and research to manufacturing and climate
- [www.census.gov/data.html](http://www.census.gov/data.html) : For demographical data on U.S. inhabitants
- [www.data.gov.uk](http://www.data.gov.uk) : Similar to the data.gov's source for U.S. data
- [www.ukdataservice.ac.uk](http://www.ukdataservice.ac.uk) : A perfect complement to data.gov.uk is the UK Data Service
- <http://data.europa.eu/euodp/en/data/> : 14,000 datasets available, EUROPA
- [www.opendatanetwork.com/](http://www.opendatanetwork.com/) : look for data using a robust search engine
- <https://data.unicef.org/> : These valuable open datasets monitor and report on the situations of children and women everywhere

## Financial and economic data

- <https://data.worldbank.org/> : world bank open data, GDP rates, logistics,
- <https://markets.ft.com/data/> : the most robust data sources for global markets
- <https://www.globalfinancialdata.com/> : access GFD's complete datasets
- <https://comtrade.un.org/> : All data is accessible through API.
- <https://www.imf.org/en/Data> : International Monetary Fund
- <http://www.bea.gov/> : Bureau of Economic Analysis
- <https://www.sec.gov/dera/data/financial-statement-data-sets.html> : U.S. Securities and Exchange Commission
- <https://www.nber.org/> : National Bureau of Economic Research
- <https://fred.stlouisfed.org/> : Federal Reserve Economic Database

## Crime and drug data

- <https://www.fbi.gov/services/cjis/ucr> : Uniform Crime Reporting Program
- <https://www.bjs.gov/index.cfm?ty=dca> : Bureau of Justice Statistics
- <https://www.icpsr.umich.edu/icpsrweb/content/NACJD/index.html> : National Archive of Criminal Justice Data
- <https://www.drugabuse.gov/related-topics/trends-statistics> : National Institute on Drug Abuse
- <https://www.unodc.org/unodc/en/data-and-analysis/> : United Nations Office on Drugs and Crime

## Health and scientific data

- <https://www.who.int/gho/database/en/> : World Health Organization
- <https://www.fda.gov/Food/default.htm> : Food and Drug Administration
- <https://healthdata.gov/> : HealthData.gov
- <http://portals.broadinstitute.org/cgi-bin/cancer/datasets.cgi> : Broad Institute
- <https://seer.cancer.gov/faststats/selections.php?series=cancer> : National Cancer Institute
- <https://www.cdc.gov/datastatistics/> : Center for Disease Control
- <https://digital.nhs.uk/> : NHS Digital
- <https://www.opensciencedatacloud.org/> : Open Science Data Cloud
- <https://pds.nasa.gov/> : NASA Planetary Data System
- <https://earthdata.nasa.gov/> : NASA Earth Data



## Academic data

- <https://scholar.google.com/> : Google Scholar
- 
- <http://www.pewinternet.org/datasets/> : Pew is one of the largest open data sources in the U.S.
- <https://nces.ed.gov/> : National Center for Education Statistics

## Environmental data

- <https://www.ncdc.noaa.gov/cdo-web/datasets> : Climate Data Online
- <https://www.cdc.gov/nceh/data.htm> : National Center for Environmental Health
- <http://energyatlas.iea.org/#!/tellmap/-1118783123> : IEA Atlas of Energy

## Business directory data

- <https://www.glassdoor.com/research/> : Glassdoor
- <https://www.yelp.com/dataset> : Yelp
- <https://opencorporates.com/> : Open Corporates

## Media and journalism data

- <https://data.fivethirtyeight.com/> : FiveThirtyEight
- <https://developer.nytimes.com/> : The New York Times Developer Network
- <https://developer.ap.org/index.html> : Associated Press Developer

## Marketing and social media data

- <https://developers.facebook.com/docs/graph-api> : Graph API
- <http://www.socialmention.com/> : Social Mention
- <https://trends.google.com/trends/?geo=US> : Google Trends

## Miscellaneous data

- <https://www.kaggle.com/datasets?sortBy=relevance&group=featured&search=tag%3A%27internet%27> : Kaggle
- <https://www.reddit.com/r/datasets/> : Datasets Subreddit
- <https://wiki.dbpedia.org/> : DBpedia
- <https://www.google.com/publicdata/directory> : Google Public Data Explorer