Data and Excel basic

Bob Xie

Sep 2020

What's Data

Data is a set of values of qualitative or quantitative variables. An example of qualitative data would be an anthropologist's handwritten notes about his or her interviews with people of an Indigenous tribe. Pieces of data are individual pieces of information. While the concept of data is commonly associated with scientific research, data is collected by a huge range of organizations and institutions, including businesses (e.g., sales data, revenue, profits, stock price), governments (e.g., crime rates, unemployment rates, literacy rates) and non-governmental organizations (e.g., censuses of the number of homeless people by non-profit organizations).

Data is measured, collected and reported, and analyzed, whereupon it can be visualized using graphs, images or other analysis tools. Data as a general concept refers to the fact that some existing information or knowledge is represented or coded in some form suitable for better usage or processing. Raw data ("unprocessed data") is a collection of numbers or characters before it has been "cleaned" and corrected by researchers. Raw data needs to be corrected to remove outliers or obvious instrument or data entry errors (e.g., a thermometer reading from an outdoor Arctic location recording a tropical temperature). Data processing commonly occurs by stages, and the "processed data" from one stage may be considered the "raw data" of the next stage. Field data is raw data that is collected in an uncontrolled "in situ" environment. Experimental data is data that is generated within the context of a scientific investigation by observation and recording. Data has been described as the new oil of the digital economy.

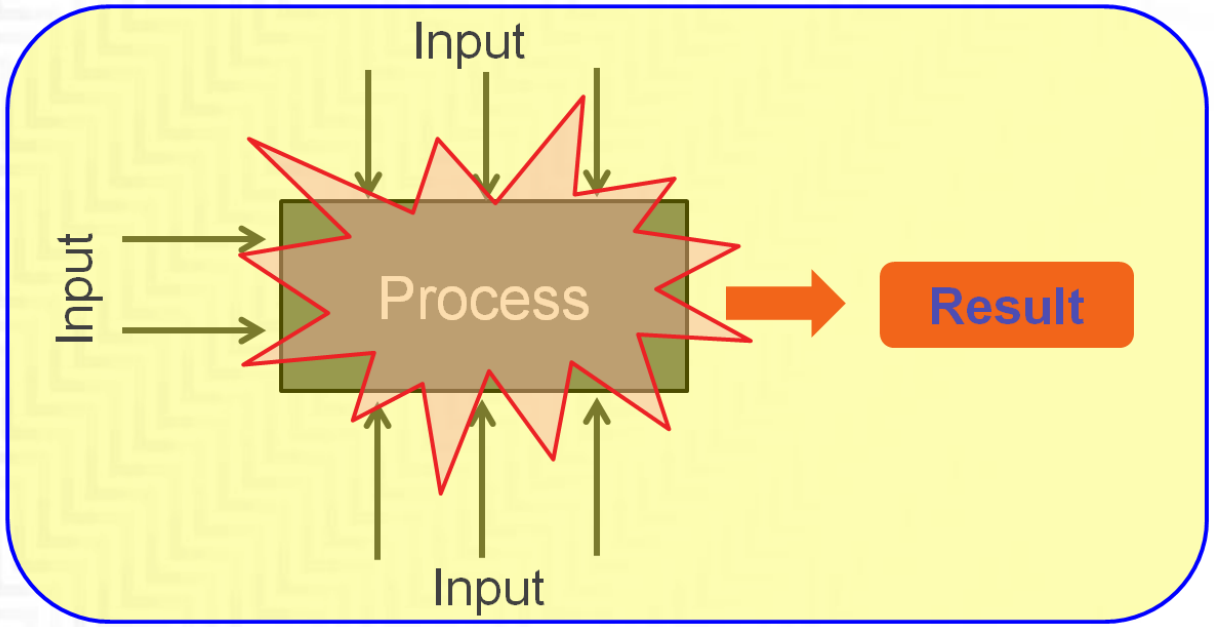
From Wikipedia

What's Data Analysis

Data analysis, also known as **analysis of data** or **data analytics** , is a process of inspecting, cleansing, transforming, and modeling data with the goal of discovering useful information, suggesting conclusions, and supporting decision-making. Data analysis has multiple facets and approaches, encompassing diverse techniques under a variety of names, in different business, science, and social science domains

From Wikipedia

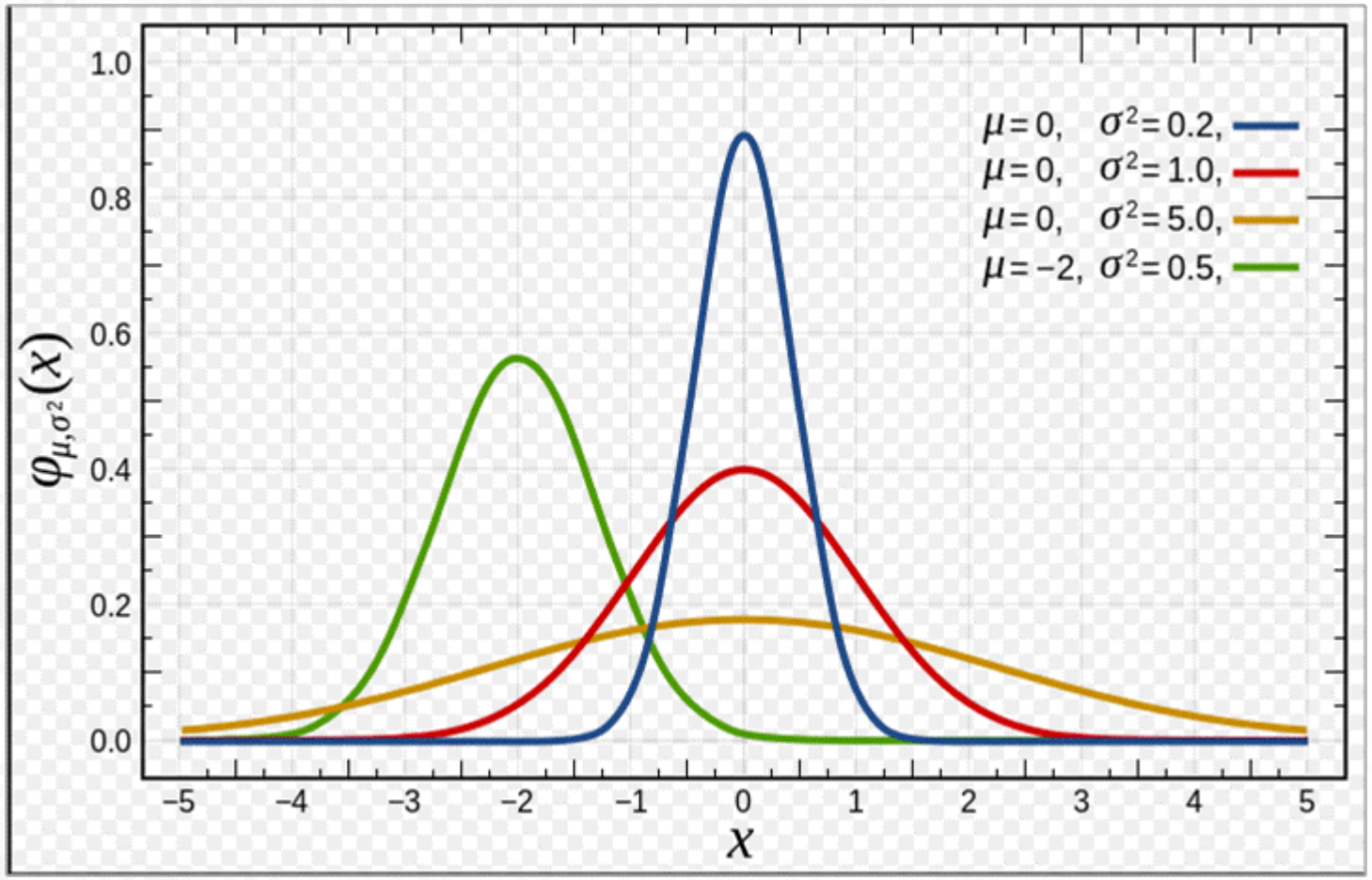
Why Data Analysis



$$Y = f(X) \begin{cases} f(\text{  }) = \text{"你好"} \\ f(\text{  }) = \text{"cat"} \\ f(\text{  }) = \text{"5-5"} \end{cases}$$

From Result Focus to Process Focus

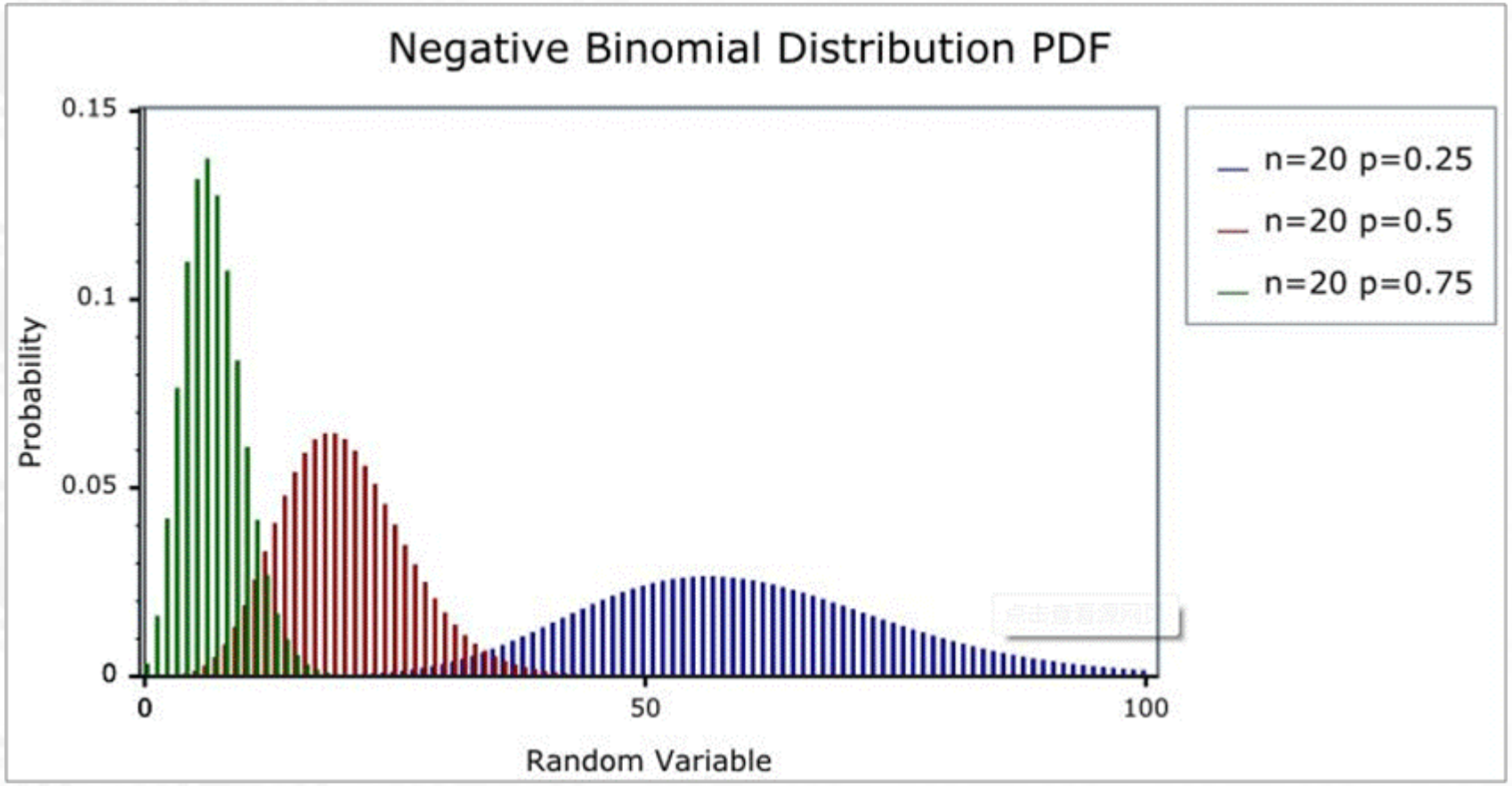
Data distribution _ Normal



Notation	$\mathcal{N}(\mu, \sigma^2)$
Parameters	$\mu \in \mathbf{R}$ — mean (location) $\sigma^2 > 0$ — variance (squared scale)
Support	$x \in \mathbf{R}$
PDF	$\frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}$
CDF	$\frac{1}{2}\left[1 + \operatorname{erf}\left(\frac{x-\mu}{\sigma\sqrt{2}}\right)\right]$
Quantile	$\mu + \sigma\sqrt{2}\operatorname{erf}^{-1}(2F-1)$
Mean	μ
Median	μ
Mode	μ
Variance	σ^2
Skewness	0
Ex. kurtosis	0
Entropy	$\frac{1}{2}\log(2\pi e\sigma^2)$

Example: Length, weight, Voltage, Current .etc

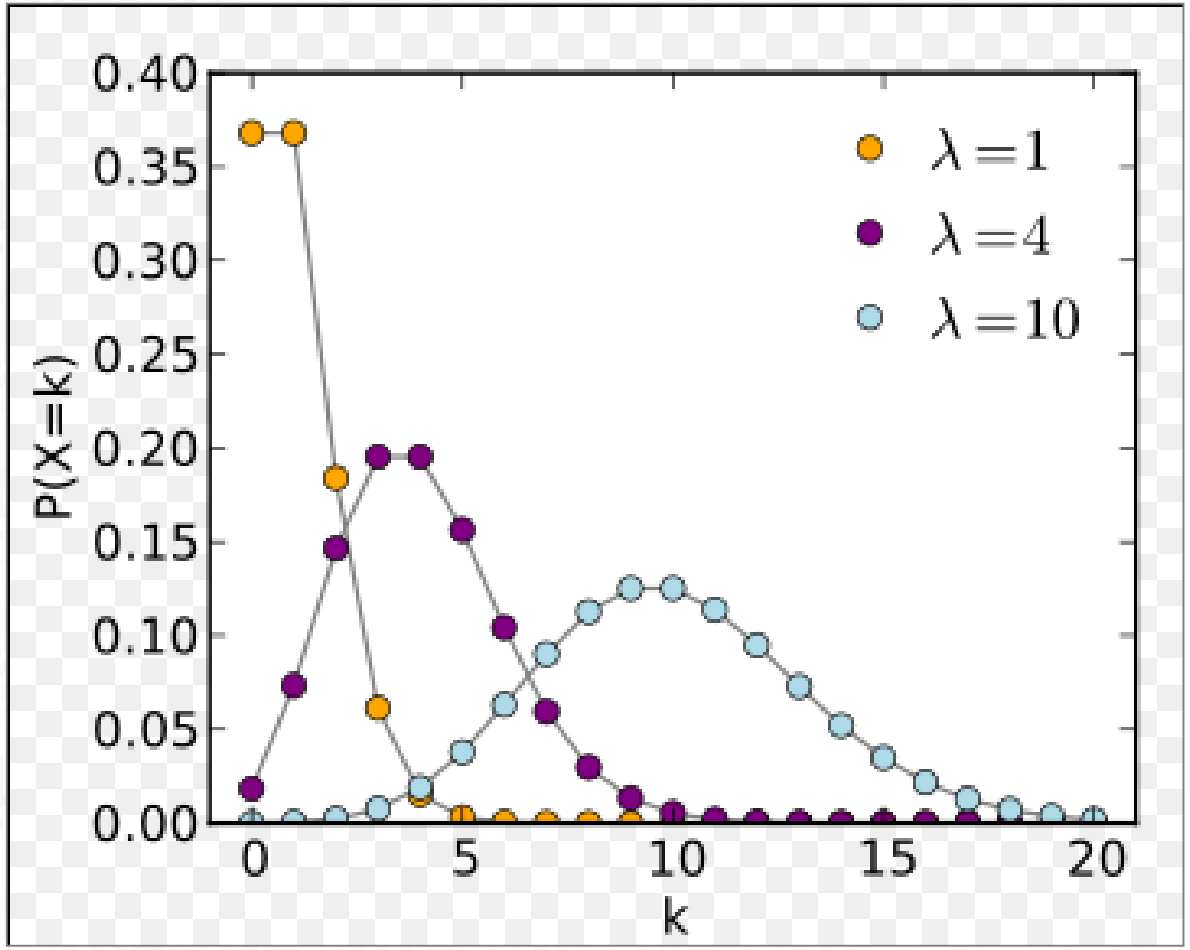
Data distribution _ Binomial



Notation	$\mathcal{B}(n, p)$
Parameters	$n \in \mathbb{N}_0$ — number of trials $p \in [0,1]$ — success probability in each trial
Support	$k \in \{0, \ldots, n\}$ — number of successes
pmf	$\binom{n}{k} p^k (1-p)^{n-k}$
CDF	$I_{1-p}(n-k, 1+k)$
Mean	np
Median	$\lfloor np \rfloor$ or $\lceil np \rceil$
Mode	$\lfloor (n+1)p \rfloor$ or $\lceil (n+1)p \rceil - 1$
Variance	$np(1-p)$
Skewness	$\frac{1-2p}{\sqrt{np(1-p)}}$
Ex. kurtosis	$\frac{1-6p(1-p)}{np(1-p)}$
Entropy	$\frac{1}{2}\log_2\left(2\pi e\, np(1-p)\right) + O\left(\frac{1}{n}\right)$ in shannons. For nats, use the natural log in the log.

Example: IFIR, RA, LRR .etc

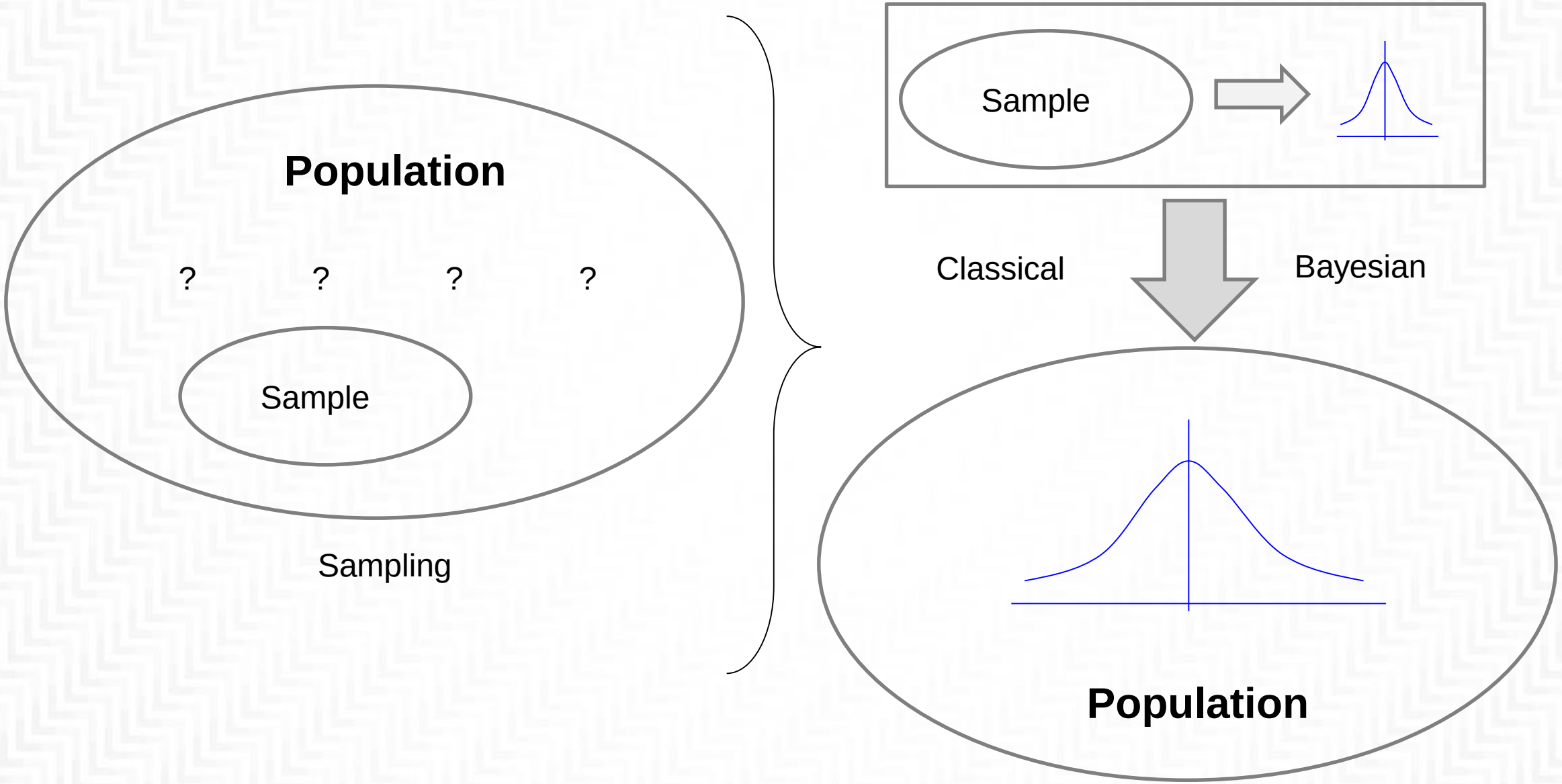
Data distribution _ Poisson



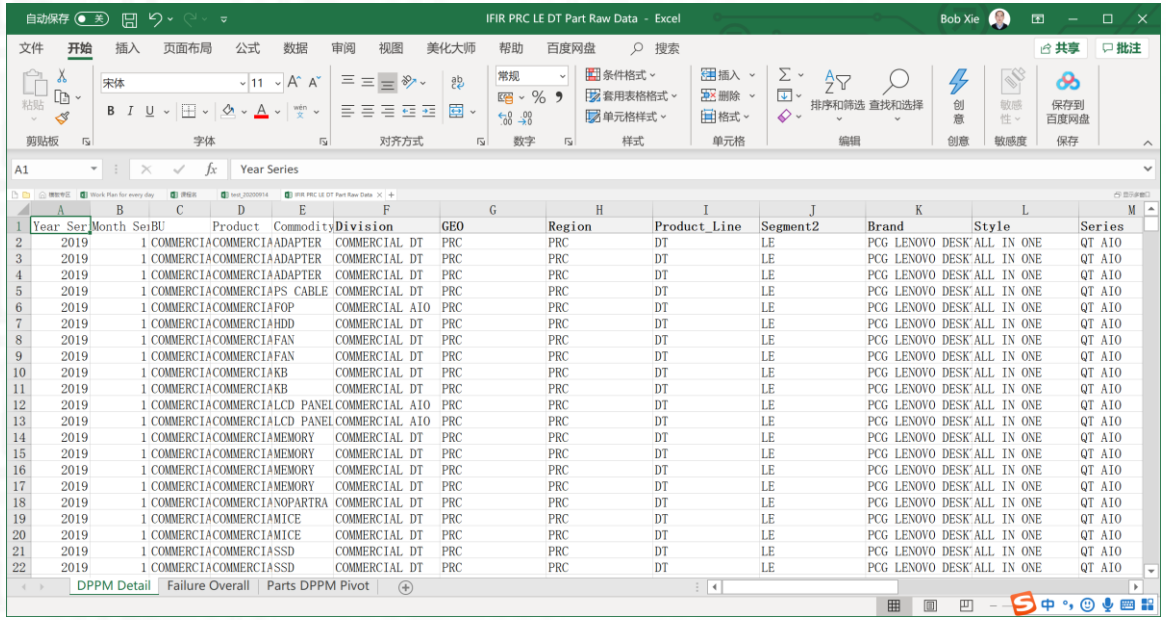
Parameters	$\lambda > 0$ (real) — rate
Support	$k \in \mathbb{N} \cup \{0\}$;
pmf	$\frac{\lambda^k e^{-\lambda}}{k!}$
CDF	$\frac{\Gamma(\lfloor k+1 \rfloor, \lambda)}{\lfloor k \rfloor !}$, or $e^{-\lambda} \sum_{i=0}^{\lfloor k \rfloor} \frac{\lambda^i}{i!}$, or $Q(\lfloor k+1 \rfloor, \lambda)$ (for $k \geq 0$, where $\Gamma(x, y)$ is the upper incomplete gamma function, $\lfloor k \rfloor$ is the floor function, and Q is the regularized gamma function)
Mean	λ
Median	$\approx \lfloor \lambda + 1/3 - 0.02/\lambda \rfloor$
Mode	$\lceil \lambda \rceil - 1, \lfloor \lambda \rfloor$
Variance	λ
Skewness	$\lambda^{-1/2}$
Ex. kurtosis	λ^{-1}
Entropy	$\lambda[1 - \log(\lambda)] + e^{-\lambda} \sum_{k=0}^{\infty} \frac{\lambda^k \log(k!)}{k!}$ (for large λ) $\frac{1}{2} \log(2\pi e\lambda) - \frac{1}{12\lambda} - \frac{1}{24\lambda^2} - \frac{19}{360\lambda^3} + O\left(\frac{1}{\lambda^4}\right)$

Example: Call in # /hr, Cosmetic defects .etc

Sampling VS Population

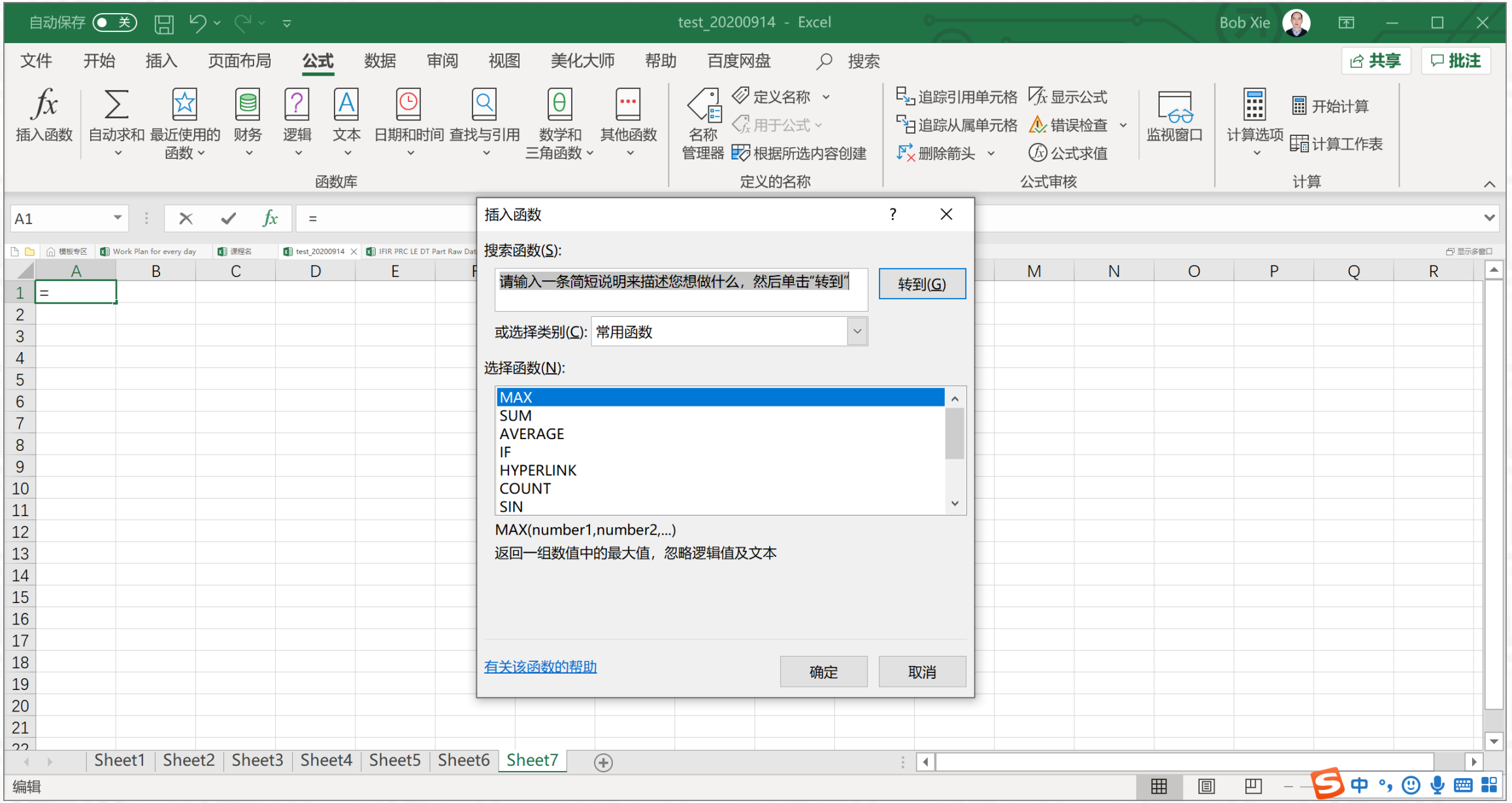


Excel Basic _ Pivot

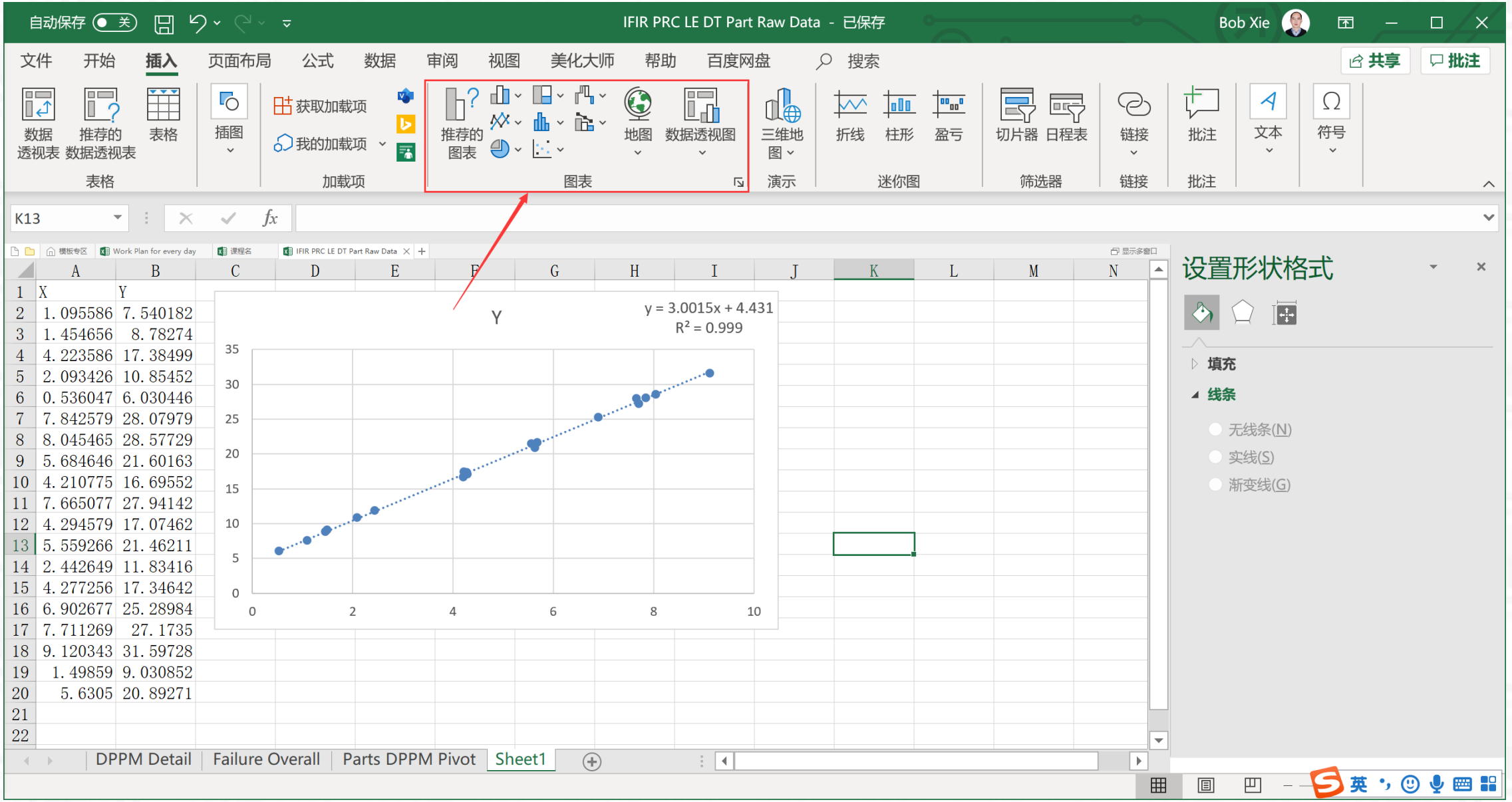


	A	B	C	D	E
1					
2	Model	(全部)			
3	Division	(全部)			
4	Commodity	(全部)			
5	PN	SP50H29558			
6					
7		值			
8	Year Month	Summary:Failed Qty	Summary:Volume	求和项:DPPM	
9	Dec-19	7	29131	240	
10	Jan-20	2	4919	407	
11	Feb-20	3	4866	617	
12	Mar-20	15	24271	618	
13	Apr-20	9	17856	504	
14					
15					
16					
17					

Excel Basic _ Function



Excel Basic _ Charts



	X	Y
2	1.095586	7.540182
3	1.454656	8.78274
4	4.223586	17.38499
5	2.093426	10.85452
6	0.536047	6.030446
7	7.842579	28.07979
8	8.045465	28.57729
9	5.684646	21.60163
10	4.210775	16.69552
11	7.665077	27.94142
12	4.294579	17.07462
13	5.559266	21.46211
14	2.442649	11.83416
15	4.277256	17.34642
16	6.902677	25.28984
17	7.711269	27.1735
18	9.120343	31.59728
19	1.49859	9.030852
20	5.6305	20.89271

Home work

- Use Excel to draw curve of F（X）= X^2+2*X+1;
- Figure out how to set up the Anaconda for Python;

Author information

- Bob Xie: Lenovo Senior SQE

If you interesting about Data analysis, can follow my column and WeChat

WeChat

column

THANK YOU

DAKUJEM DANK BEDANKT MERCI TAKK 谢谢
ありがとう СПАСИБО GRACIAS DZIĘKUJĘ DANKE
OBRIGADO БЛАГОДАРЯ GRAZIE תודה GRACIAS

Lenovo™