URP 6931. Introduction to Urban Analytics

Lecture 02: Review probability, vectors, and statistics

Instructor: Shenhao Wang
Assistant Professor, Director of Urban AI Lab
Department of Urban and Regional Planning
University of Florida

Outline

1

General diagram for data analytics – Connecting probability, vectors/numbers, and statistics

2

Review probability distributions as data generating process

3

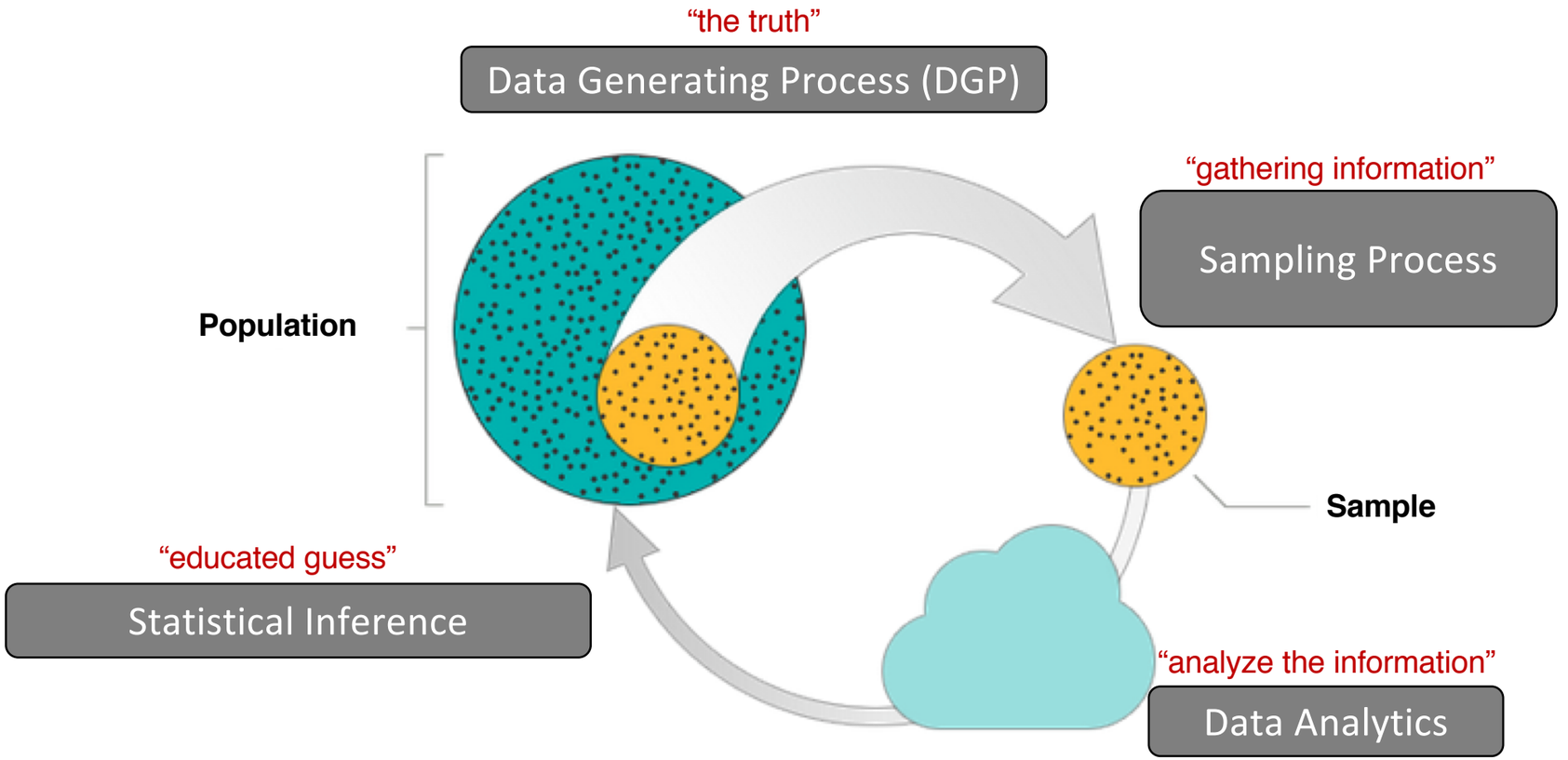
Review vectors/numbers

4

Review statistics

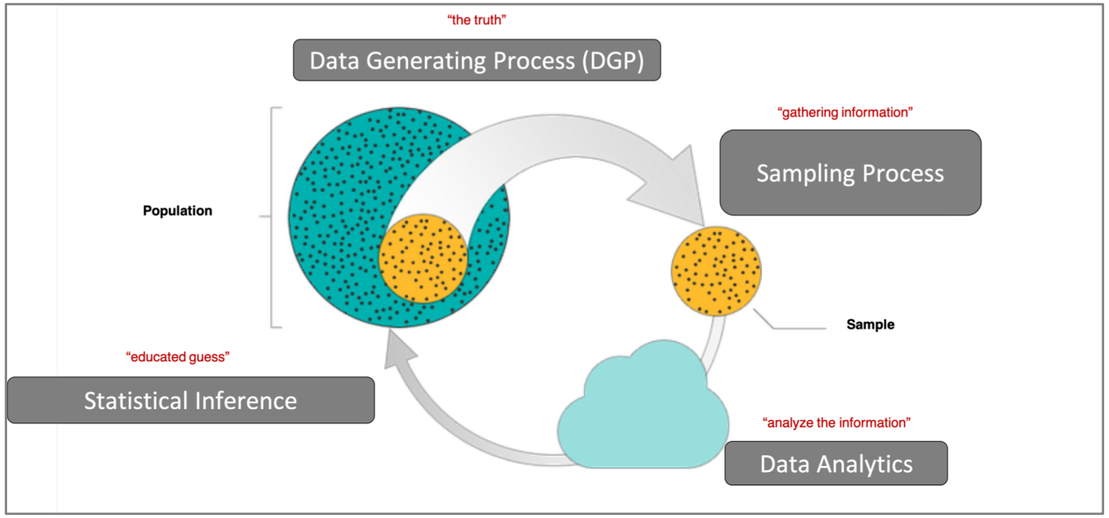
Part 1. General diagram for data analytical process

General Diagram for Data Analytical Process



Source: https://courses.lumenlearning.com/wm-concepts-statistics/chapter/wim-linking-probability-to-statistical-inference/

General Diagram for Data Analytical Process

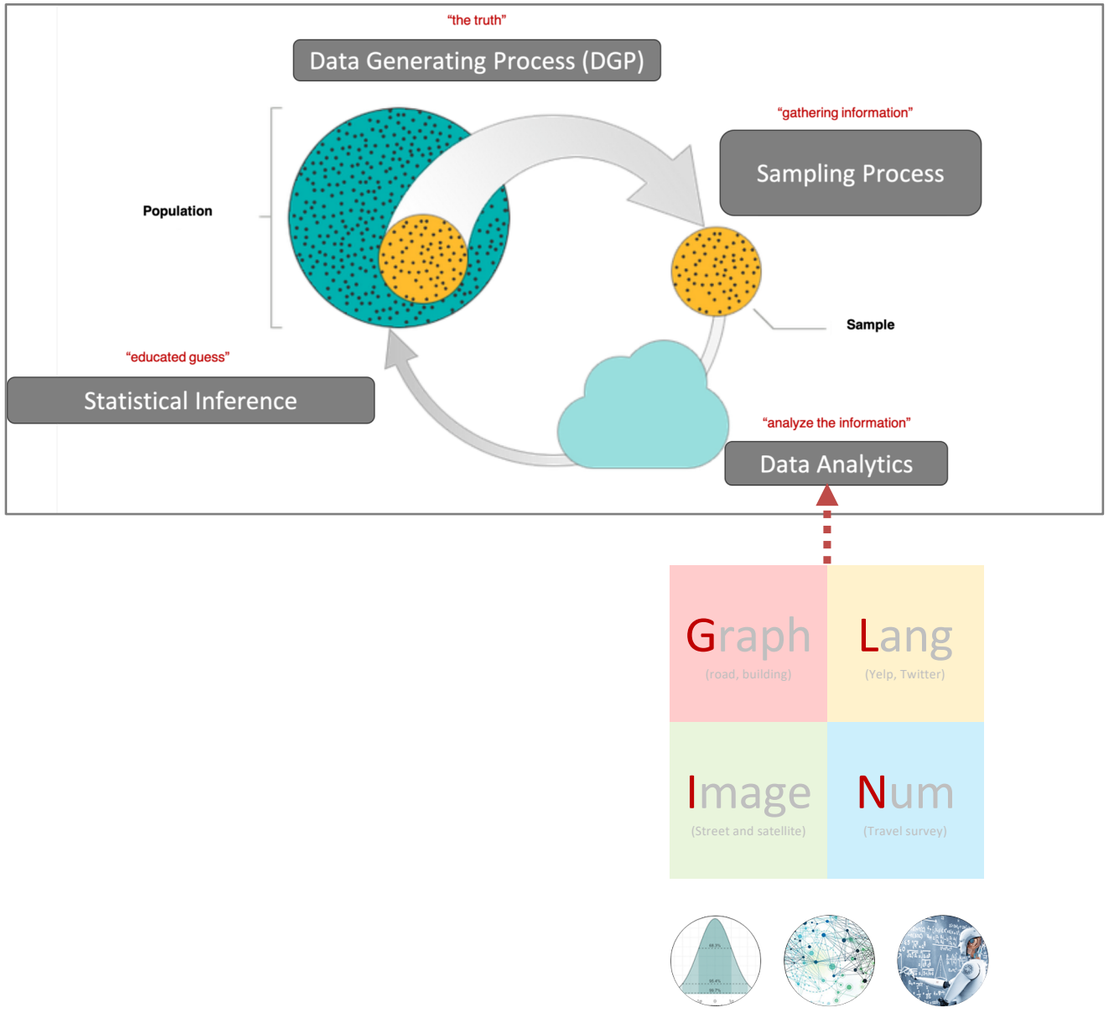


- **Data Generating Process**: a data generating process is a process in the real world that "generates" the data one is interested in.

- Usually, we **don't know** the DGP but we can **infer** it from the sample/data.

Source: https://courses.lumenlearning.com/wm-concepts-statistics/chapter/wim-linking-probability-to-statistical-inference/

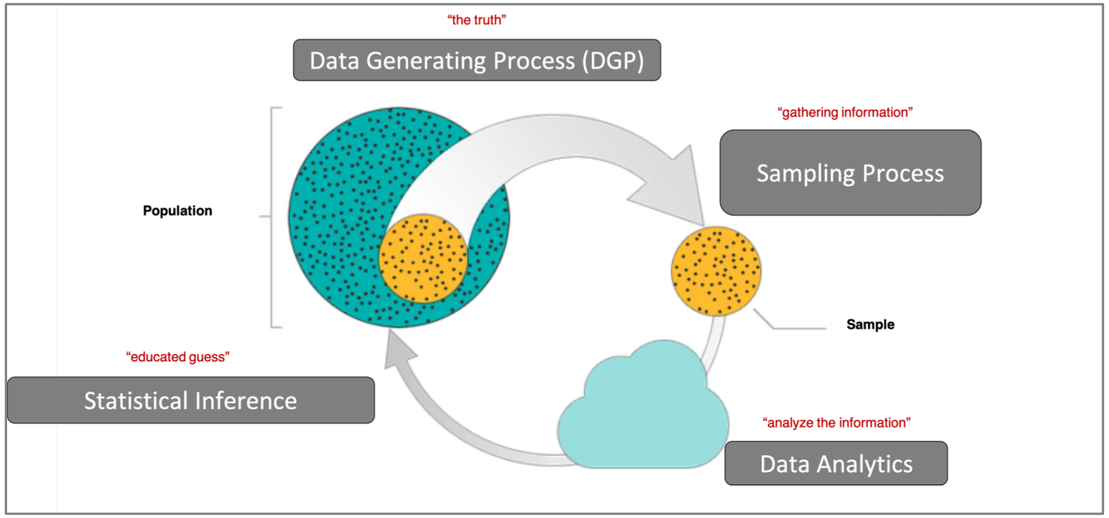
Examples in the General Diagram for Data Analytical Process



- **Examples in DGP**: random and non-random.
- **Examples in sampling process**: simple random, stratified, cluster or other sampling.
- **Examples in data**: GLIN
- **Examples in analytics**: statistics, network, and ML/DL
- **Statistical inference**: interpretation, prediction, etc.

Source: https://courses.lumenlearning.com/wm-concepts-statistics/chapter/wim-linking-probability-to-statistical-inference/

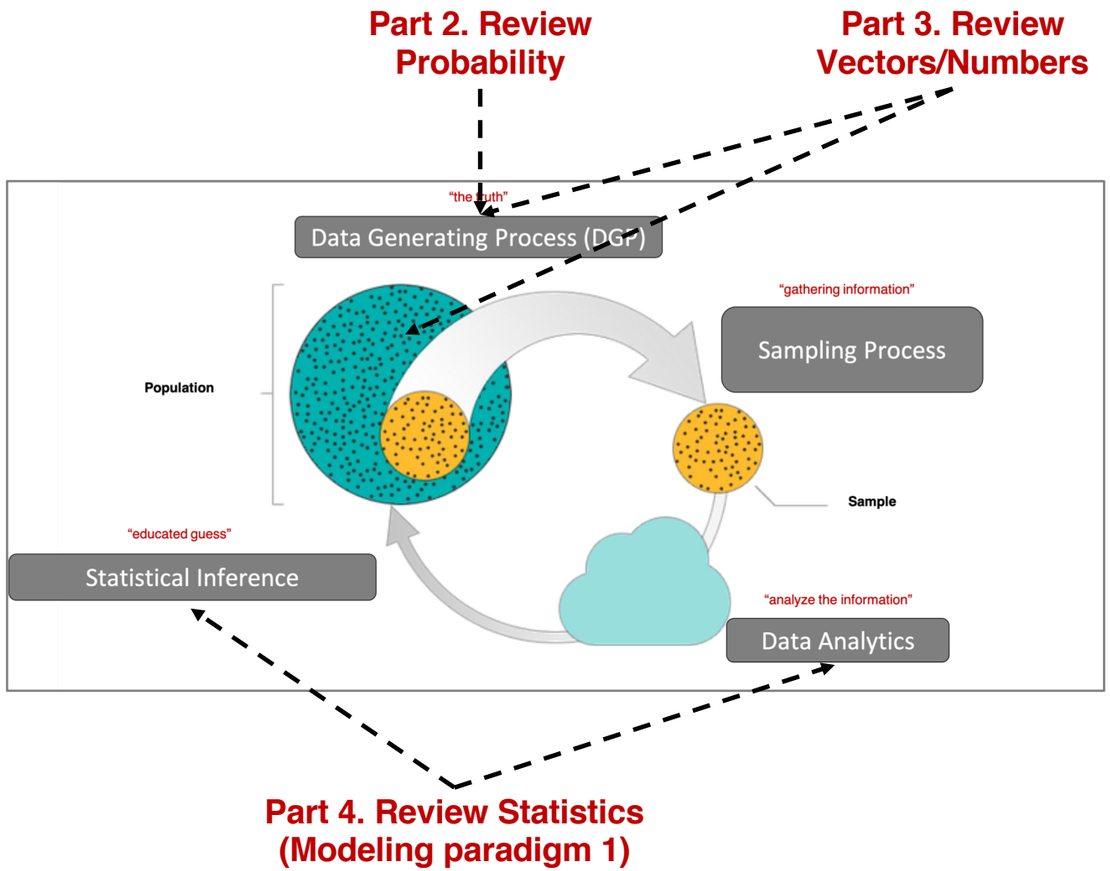
How to criticize any data analysis?



- **Using DGP**: random and non-random?
- **Using sampling process**: simple random, stratified, cluster or other sampling?
- **Using data**: GLIN?
- **Using analytics**: statistics, network, or ML/DL?
- **Statistical inference**: interpretation or prediction?

Source: https://courses.lumenlearning.com/wm-concepts-statistics/chapter/wim-linking-probability-to-statistical-inference/

Connecting probability, vectors/numbers, and statistics



Source: https://courses.lumenlearning.com/wm-concepts-statistics/chapter/wim-linking-probability-to-statistical-inference/

Part 2. Probability distributions as data generating process

Disclaimer: probability distributions are much broader – but this title serves the purpose for today's class

Random variables and key concepts

- The behavior of a random variable is determined by its probability distribution function.
- The probability distribution function has its support and probability density/mass function.
- The shape of the probability distribution of a random variable is often characterized by parameters.
    - The mean value: E[X]
    - The variance: V[X]
- Other important concepts:
    - Independence of X and Y
    - Conditioning P(Y|X)
    - Joint distributions P(X, Y)
    - Probability density function P(X)
    - etc.
- **Common probability distributions**

Two basic types of random variables

Discrete support

Take on a countable number of distinct values. e.g. {0, 1}

Examples:

- Household size (0, 1, …)
- Number of vehicles (0, 1, …)

Distribution function:

- Probability Mass Function (PMF)

Continuous support

Take infinite number of possible values. e.g., $[-\infty, \infty]$

Examples:

- Income (dollars, cents)
- Property values (dollars, cents)

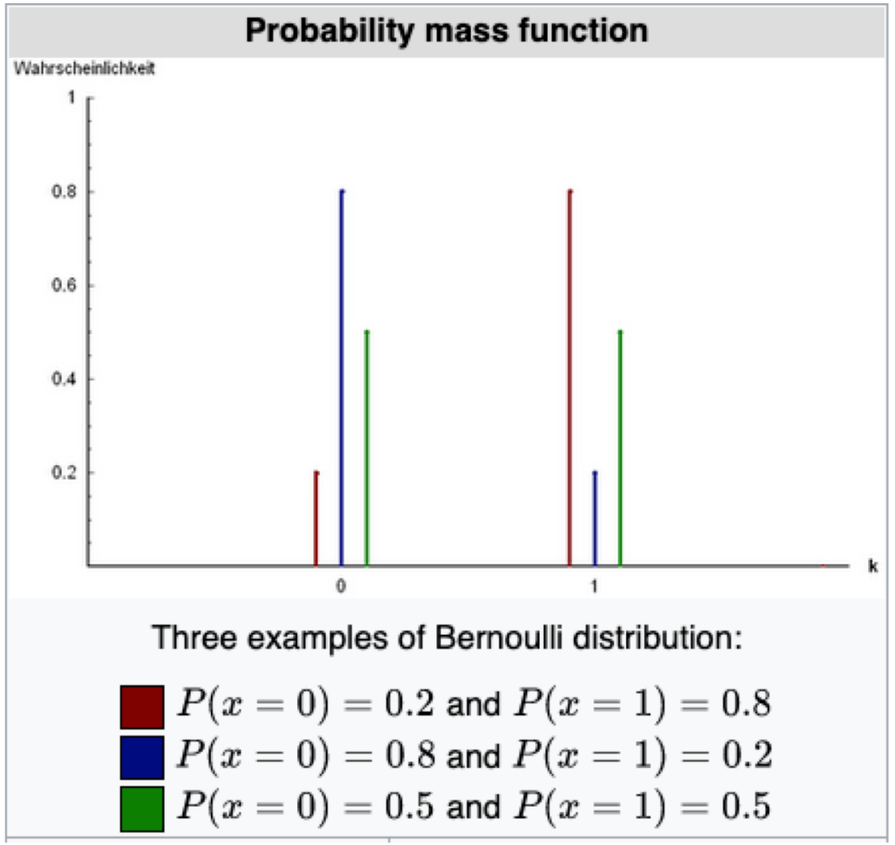
Distribution function:

- Probability Density Function (PDF)

Bernoulli distribution (discrete support)

Bernoulli random variable x ~ B(p)

- x takes two values: 0 and 1
- Pr(x = 0) = p and Pr(x = 1) = 1- p



Example (DGP)

- Parameter p = 0.6
- Sample 10 numbers
- [1, 1, 0, 1, 1, 0, 0, 1, 0, 1]

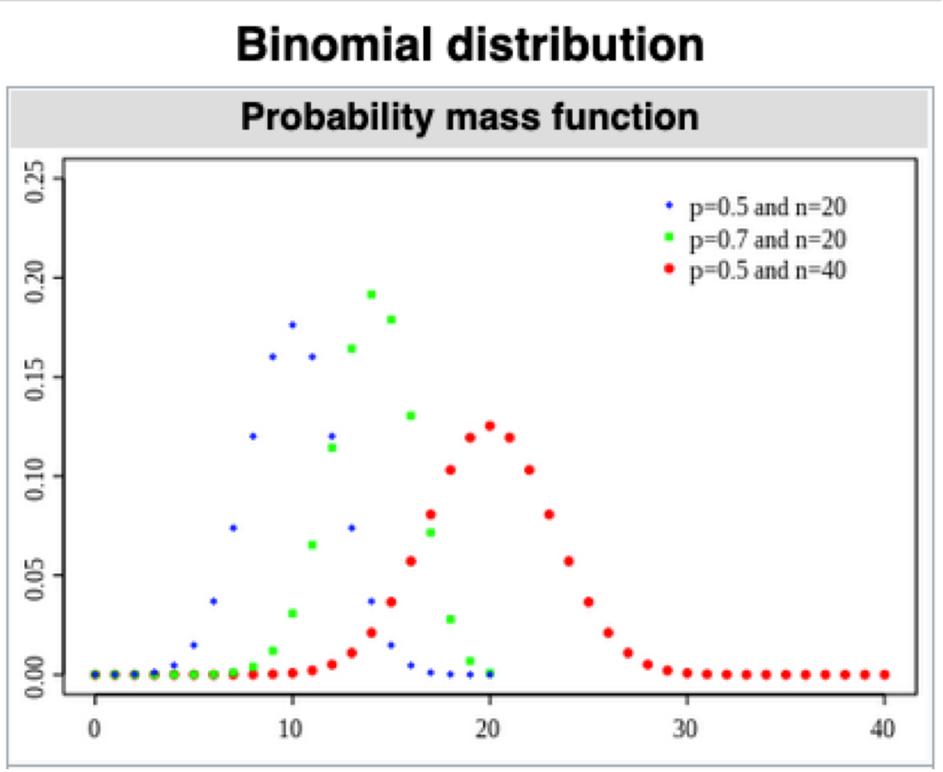
Urban applications

- Coin flip; high/low income; young/old generations; racial majority/minority
- Very basic distribution but with broad usage

Binomial distribution (discrete support)

Random variable x ~ B(n, p)

- p: probability of heads up in a coin flip
- n: repeated coin flips



Example (DGP)

- Parameter n = 5, p = 0.6 (DGP)
- Sample 10 numbers
- [3, 3, 3, 2, 4, 2, 2, 3, 5, 3]

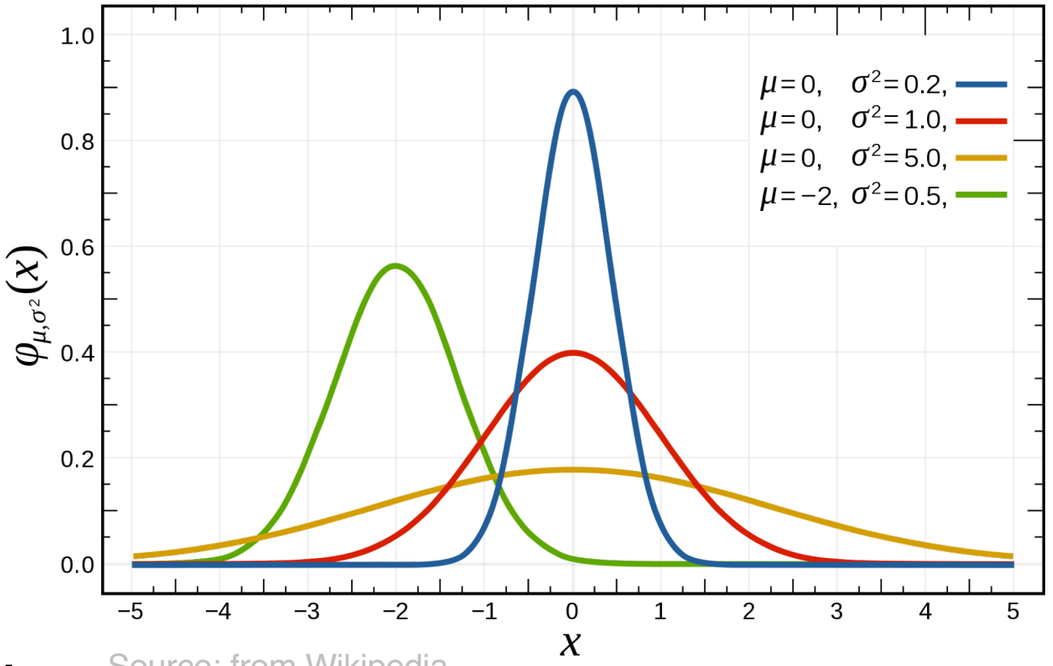
Urban applications

- Household size, number of vehicles per household, etc.

Normal/Gaussian distribution (continuous support)

Random variable $x \sim N(\mu, \sigma^2)$

- μ: mean (location)
- σ^2 : variance (squared scale)



Source: from Wikipedia

Example (DGP)

- Parameter $\mu = 0$, $\sigma^2 = 1$
- Sample 10 numbers
- [0.399, 1.302, -0.557, 0.116, -1.237, -0.572, 0.080, -0.041, -1.191, 0.796]

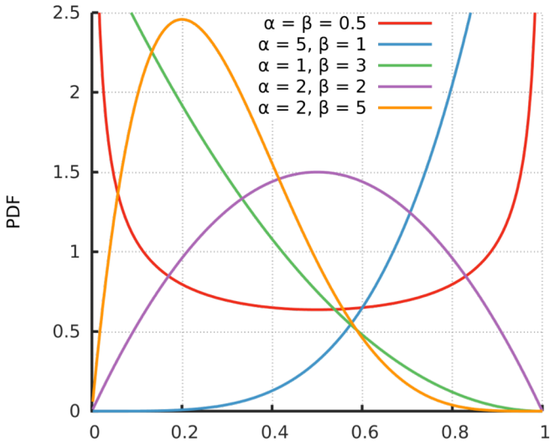
Urban applications

- Gaussian distribution is commonly used as theoretical benchmarks.
- But the urban data don't have a lot of negative values

Other distributions

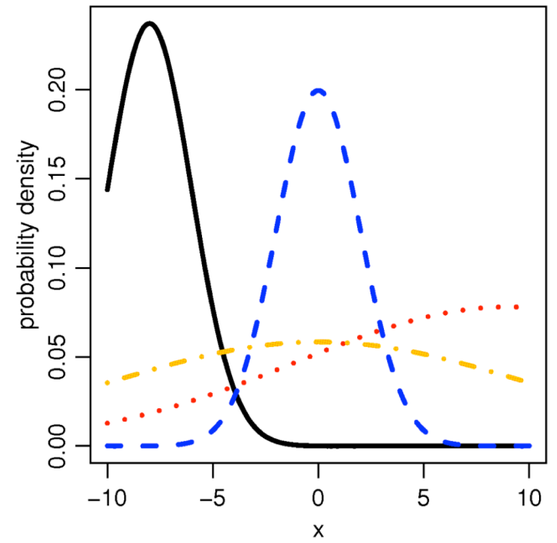
Support at [0, 1]

- e.g. Beta distributions
- Urban applications: ratio of anything (aggregate)



Support at $[0, \infty]$

- e.g. truncated normal distribution
- Urban applications : income, property values, etc.



For other distributions, please check probability textbooks

Q: How to connect the probability distributions to the spread sheets we collect?

A: Probability distributions as DGP

	pop_total	age_median	households	property_value_median	inc_median_household	race_white_ratio	race_black_ratio	race_native_ratio	race_asian_ratio
0	2812	39.4	931	240400	53533	0.741820768	0.183854908	0	0.043385491
1	4709	34.2	1668	179900	33958	0.505839881	0.414737736	0	0.018687619
2	5005	34.1	1379	254900	40250	0.466333666	0.440759241	0.044755245	0.031768232
3	6754	31.3	2238	147800	39962	0.599940776	0.247408943	0.048267693	0.052413385
4	3021	44.1	1364	205900	63889	0.947037405	0.040052963	0	0
5	2599	55.8	1278	433600	73125	0.939976914	0.029242016	0	0.010773374
6	2079	64.6	1017	808500	114653	0.927849928	0	0	0.037518038
7	4215	38.8	1032	232200	68214	0.789323843	0.152313167	0.02514828	0.004270463
8	7032	34.5	2191	150400	42583	0.655147895	0.268486917	0	0.00867463
9	4019	34.9	1235	213500	46174	0.782532968	0.075889525	0.012192088	0.032843991
10	5477	34.7	1705	80500	45228	0.723023553	0.1327369	0	0.04710608
11	3916	42.2	1277	238900	85024	0.877425945	0.027579162	0	0.026302349
12	4673	33.4	1558	140700	31190	0.702118553	0.213995292	0.01455168	0.001283972
13	8212	36.7	2397	344400	88707	0.813078422	0.082075012	0	0.04846566
14	3385	75	1769	258200	55402	0.971048744	0.006794682	0	0
15	3496	28.6	1031	170100	52862	0.796052632	0.110983982	0	0.026887872
16	7173	33.3	1948	159400	44216	0.815140109	0.073330545	0	0.033877039
17	6009	32.5	1548	282800	57111	0.508404061	0.419703778	0	0.054418372
18	6089	43.1	2043	183300	50938	0.639349647	0.259320085	0.029068813	0
19	7535	45.2	2741	260200	91279	0.840079628	0.082149967	0	0.025348374
20	3926	40.3	1590	454900	91929	0.693326541	0.255476312	0.004075395	0.003311258

Probability distributions as DGP

Row: observation i (census tracts); **Column**: features x_i or y_i

Q: How to choose a probability distribution to generate the age column?

	pop_total	age_median	households	property_value_median	inc_median_household	race_white_ratio	race_black_ratio	race_native_ratio	race_asian_ratio
0	2812	39.4	931	240400	53533	0.741820768	0.183854908	0	0.043385491
1	4709	34.2	1668	179900	33958	0.505839881	0.414737736	0	0.018687619
2	5005	34.1	1379	254900	40250	0.466333666	0.440759241	0.044755245	0.031768232
3	6754	31.3	2238	147800	39962	0.599940776	0.247408943	0.048267693	0.052413385
4	3021	44.1	1364	205900	63889	0.947037405	0.040052963	0	0
5	2599	55.8	1278	433600	73125	0.939976914	0.029242016	0	0.010773374
6	2079	64.6	1017	808500	114653	0.927849928	0	0	0.037518038
7	4215	38.8	1032	232200	68214	0.789323843	0.152313167	0.02514828	0.004270463
8	7032	34.5	2191	150400	42583	0.655147895	0.268486917	0	0.00867463
9	4019	34.9	1235	213500	46174	0.782532968	0.075889525	0.012192088	0.032843991
10	5477	34.7	1705	80500	45228	0.723023553	0.1327369	0	0.04710608
11	3916	42.2	1277	238900	85024	0.877425945	0.027579162	0	0.026302349
12	4673	33.4	1558	140700	31190	0.702118553	0.213995292	0.01455168	0.001283972
13	8212	36.7	2397	344400	88707	0.813078422	0.082075012	0	0.04846566
14	3385	75	1769	258200	55402	0.971048744	0.006794682	0	0
15	3496	28.6	1031	170100	52862	0.796052632	0.110983982	0	0.026887872
16	7173	33.3	1948	159400	44216	0.815140109	0.073330545	0	0.033877039
17	6009	32.5	1548	282800	57111	0.508404061	0.419703778	0	0.054418372
18	6089	43.1	2043	183300	50938	0.639349647	0.259320085	0.029068813	0
19	7535	45.2	2741	260200	91279	0.840079628	0.082149967	0	0.025348374
20	3926	40.3	1590	454900	91929	0.693326541	0.255476312	0.004075395	0.003311258

What probability distribution can be used to generate the age column?

Let's try $N(\mu, \sigma^2)$ with μ = 44 and σ = 10. Sample size = 20

Generated age column: [66.3, 48.6, 35.1, 52.6, 35.9, 33.8, 49.7, 28.5, 55.3, 67.4, 50.7, 46.2, 52.8, 42.5, 36.9, 42.7, 51.2, 48.3, 43.4, 26.7]

Notes

- It is a naïve practice to connect probability, vectors/matrix, and statistics.
- We can formalize this intuition.

	pop_total	age_median	households
0	2812	39.4	931
1	4709	34.2	1668
2	5005	34.1	1379
3	6754	31.3	2238
4	3021	44.1	1364
5	2599	55.8	1278
6	2079	64.6	1017
7	4215	38.8	1032
8	7032	34.5	2191
9	4019	34.9	1235
10	5477	34.7	1705
11	3916	42.2	1277
12	4673	33.4	1558
13	8212	36.7	2397
14	3385	75	1769
15	3496	28.6	1031
16	7173	33.3	1948
17	6009	32.5	1548
18	6089	43.1	2043
19	7535	45.2	2741
20	3926	40.3	1590

Part 2. Reviewing vectors and numbers

- Discussing the vectors and numbers in the **population** (not strictly).
- Extension is matrix algebra (not reviewed today).
- Covering three types of numbers: nominal, ordinal, and cardinal numbers (different from the discrete vs. continuous numbers).
- Summarizing central tendency (mean, mode, and median) and variability (range, quartiles, and variance).

1. Nominal Numbers

Nominal Numbers. For identification only; They cannot be used for ranking or counting.

Example 1: city names

City List: Boston, Lima, Los Angeles, Minneapolis, Osaka, Cairo

Number representation: 0, 1, 2, 3, 4, 5

Example 2: travel mode choice

List: automobile, other modes

Number representation: 0, 1

How to summarize the central tendency of nominal numbers?

Mode: the most frequent result.

Example 1: city names

City List: Boston, Lima, Los Angeles, Lima, Minneapolis, Osaka, Cairo

Result: Lima.

2. Ordinal Numbers

Ordinal Numbers. For identification and ranking. But cannot be used for algebraic operations (e.g. addition/subtraction)

Example 1: Housing quality rating

3, 2, 3, 4, 2, 3, 1

Ranking:

1, 2, 2, 3, 3, 3, 4

Addition:

1 + 2 = 3(?)

How to summarize the central tendency of ordinal numbers?

Median: the middle observation when everything is in order.

Example 1: Housing quality rating

3, 2, 3, 4, 2, 3, 1

Ranking: 1, 2, 2, 3, 3, 3, 4

The median is 3.

3. Cardinal Numbers

Cardinal Numbers. For identification, ranking, and algebraic operations (e.g. addition/subtraction)

Example 1: Life expectancy at birth

59, 59, 61, 62, 71, 71, 73

How to summarize the central tendency of cardinal numbers?

Mean: what we normally call “the average” – it is the sum of the scores divided by the number of observations (N).

$$\mu = \frac{\sum_{i=1}^{N} X_i}{N}$$

Example 1: Life expectancy at birth

59, 59, 61, 62, 71, 71, 73

The mean is: $\frac{59+62+71+59+73+71+61}{7} = 65.14$

Can we use *the mean* to summarize ordinal numbers?

Example 1: Housing quality rating

3, 2, 3, 4, 2, 3, 1

Re-order: 1, 2, 2, 3, 3, 3, 4

$$\mu = \frac{\sum_{i=1}^{N} X_i}{N}$$

The mean is: $\frac{3+2+3+4+2+3+1}{7}$ = 2.57

What is wrong with this?

Can we use *the mean* to summarize nominal numbers?

Example 2: travel mode choice

List: automobile, other modes

Number representation: 0, 1

$$\mu = \frac{\sum_{i=1}^{N} X_i}{N}$$

The mean is: $\frac{0+1}{2}$ = 0.50

What is wrong with this?

Means vs. Medians: sensitivity to extreme cases

Net Worth (I made up the numbers)

Shenhao Wang: $10

Emre Tepe: $97

Ruth Steiner: $109

Zhong-Ren Peng: $121

Chimay Anumba: $200

Median: $109; Mean: $107

Means vs. Medians: sensitivity to extreme cases

Net Worth (I made up the numbers)

Shenhao Wang: $10

Emre Tepe: $97

Ruth Steiner: $109

Zhong-Ren Peng: $121

Chimay Anumba: $200

Warren Buffett: $87,000,000,000

Median: $115; Mean: $14.5BN

Which one to use? Depending on the distribution and the issue at hand, you may care more about medians or means. Remember: for ordinal data, the median is really all that is meaningful.

Measures of variability: Range

The Range is simply the difference between the highest and the lowest value in the distribution.

Examples

- 67, 67, 97, 98, 99, 100, 101
- The range is: 101 - 67 = 34.
- 2.4 3.5 3.5 6.7 7.0 7.0 9.1 9.9 11.2
- The range is: 11.2 – 2.4 = 8.8

Measures of variability: Quartiles

Quartiles divide up the range into four segments with an equal number of scores in each segment.

Quartiles are the most common, but some people use quintiles, deciles, or percentiles. These are all called “quantiles.”

Measures of variability: Variance of a population

Variance: the average squared deviation of the scores from the mean.

$$\sigma^2 = \frac{\sum_{i=1}^{N}(X_i - \mu)^2}{N}$$

Note: this is the variance for a population, not a sample. The symbol is σ^2 (not s^2), the mean is μ (not $\bar{X}$). The denominator is N (not n-1).

Measures of variability: Standard deviation of a population

Standard deviation: simple the square root of the variance; more useful for most statistics.

$$\sigma = \sqrt{\frac{\sum_{i=1}^{N}(X_i - \mu)^2}{N}}$$

Note: this is the standard deviation for a population, not a sample. The symbol is σ (not s), the mean is μ (not $\bar{X}$). The denominator is N (not n-1).

Information Richness

Nominal numbers	<	Ordinal numbers	<	Cardinal numbers
Mode		Mode & Median		Mode & Median & Mean
Range/Support		Range and quantiles		Range, quantiles, variance, and standard deviation

Information Richness Indicates the Direction of Data Processing

Example: Income

[High, Low] ← [3, 2, 1] ← [100K, 50K, 30K]

Part 3. Statistics as “educated guesses”

Statistical Inference

Unknown Population Distribution:

$$Y \sim ?\,(\mu, \sigma^2)$$

Parameters: μ, σ^2

Sample: $(Y_1, Y_2, Y_3, \ldots, Y_N)$ – Note each data point is random.

Let's infer: what is the mean of the population?

1. Creating an estimator/statistic: $\hat{g}(Y_1, Y_2, Y_3, \ldots, Y_N)$.

Estimators are function of sample data which we use to learn about the parameters, often denoted with a hat. $\hat{\mu}$, $\hat{\sigma}$, $\hat{\theta}$, $\hat{\beta}$, etc.

2. Computing an estimate: $\hat{g}(Y_1 = y_1, Y_2 = y_2, Y_3 = y_3, \ldots, Y_N = y_N)$.

Estimates are particular values of estimators that are realized in a given sample.

Let's estimate the population mean μ

If we think of the data as randomly sampled from the population distribution, then the sample $(Y_1, Y_2, Y_3, \ldots, Y)$ are independently and identically distributed (IID) random variables with $E[Y_i] = \mu$, and $V[Y_i] = \sigma^2$ for all $i \in \{1, \ldots N\}$.

Let's compare:

1. $\hat{\mu}_1 = Y_1$ (the first data observation)
2. $\hat{\mu}_2 = 2$ (because 2 is Shenhao's lucky number)
3. $\hat{\mu}_3 = \bar{Y}_N \equiv \frac{1}{N}(Y_1 + \ldots + Y_N)$ (the sample average)
4. $\hat{\mu}_4 = \bar{Y}_N \equiv \frac{1}{N+5}(Y_1 + \ldots + Y_N)$ (an "adjusted" sample average)

What is your intuition?

Let's estimate the population mean μ

Formalize our intuition

Finite-sample properties (apply for any sample size):

- Unbiasedness: is the sampling distribution of our estimator centered at the true parameter value? $E[\hat{\mu}] = \mu$
- Efficiency: Is the variance of the sampling distribution of our estimator reasonably small? $V[\hat{\mu}]$ is small.

Asymptotic properties (kick in when N is large):

- Consistency: As our sample size grows to infinity, does the sampling distribution of our estimator converge to the true parameter value?
- Asymptotic Normality: As our sample size grows large, does the sampling distribution of our estimator approach a normal distribution?

Definition (Bias)

Bias is the expected difference between an estimator $\hat{\theta}$ and a parameter θ

$$Bias(\hat{\theta}) = E[\hat{\theta} - \theta]$$

An estimator is unbiased iff: $Bias(\hat{\theta}) = 0$

Let's compare:

$\hat{\mu}_1 = Y_1$; $\hat{\mu}_2 = 2$; $\hat{\mu}_3 = \frac{1}{N}(Y_1 + \ldots + Y_N)$; $\hat{\mu}_4 = \frac{1}{N+5}(Y_1 + \ldots + Y_N)$

- $E[\hat{\mu}_1] = \mu$; (Unbiased)
- $E[\hat{\mu}_2] = 2$; (Biased)
- $E[\hat{\mu}_3] = \mu$; (Unbiased)
- $E[\hat{\mu}_4] = \frac{N}{N+5}\mu$; (Biased)

Definition (Efficiency)

If $\hat{\theta}_1$ and $\hat{\theta}_2$ are two estimators of θ, then $\hat{\theta}_1$ is more efficient relative to $\hat{\theta}_2$ iff:

$$V[\hat{\theta}_1] < V[\hat{\theta}_2]$$

Let's compare:

$\hat{\mu}_1 = Y_1; \hat{\mu}_2 = 2; \hat{\mu}_3 = \frac{1}{N}(Y_1 + \ldots + Y_N); \hat{\mu}_4 = \frac{1}{N+5}(Y_1 + \ldots + Y_N)$

- $V[\hat{\mu}_1] = \sigma^2;$
- $V[\hat{\mu}_2] = 0;$
- $V[\hat{\mu}_3] = \frac{1}{N}\sigma^2;$
- $V[\hat{\mu}_4] = \frac{N}{(N+5)^2}\sigma^2;$

- It is very hard to evaluate estimators using ONLY efficiency.

Definition (Consistency)

An estimator $\hat{\theta}_N$ is consistent if the sequence $\hat{\theta}_1, \ldots, \hat{\theta}_N$ converges in probability to the true parameter θ as sample size N grows to infinity:

$$\hat{\theta}_N \xrightarrow{p} \theta$$

Let's compare:

$\hat{\mu}_1 = Y_1; \hat{\mu}_2 = 2; \hat{\mu}_3 = \frac{1}{N}(Y_1 + \ldots + Y_N); \hat{\mu}_4 = \frac{1}{N+5}(Y_1 + \ldots + Y_N)$

- $E[\hat{\mu}_1] = \mu$ and $V[\hat{\mu}_1] = \sigma^2$ (inconsistent)
- $E[\hat{\mu}_2] = 2$ and $V[\hat{\mu}_2] = 0$ (inconsistent)
- $E[\hat{\mu}_3] = \mu$ and $V[\hat{\mu}_3] = \frac{1}{N}\sigma^2$; (consistent)
- $E[\hat{\mu}_4] = \frac{N}{N+5}\mu$ and $V[\hat{\mu}_4] = \frac{N}{(N+5)^2}\sigma^2$; (consistent)

About sample mean - Law of Large Numbers

Definition (Law of Large Numbers)

Let $Y_1, Y_2, \ldots, Y_N$ be a sequence of i.i.d. random variables, each with finite mean μ. Then for all $\epsilon > 0$,

$$\bar{X}_N \xrightarrow{p} \mu \text{ as } N \to \infty$$

or equivalently,

$$\lim_{n \to \infty} \Pr(|\bar{Y}_N - \mu|) \geq \epsilon) = 0$$

where $\bar{Y}_N$ is the sample mean.

Notes

- About sample mean.
- It says that sample mean is a consistent estimator for ANY population distribution as long as the mean exists.

About sample mean - Central Limit Theorem

Definition (Central Limit Theorem)

Let $Y_1, Y_2, \ldots, Y_N$ be a sequence of i.i.d. random variables, each with finite mean μ and variance $\sigma^2 < \infty$. Then for any population distribution of Y,

$$\sqrt{N}(\bar{Y}_N - \mu) \xrightarrow{d} N(0, \sigma^2).$$

CLT also implies that the standardized sample mean converges to a standard normal distribution:

$$\frac{\bar{Y}_N - \mu}{\sigma/\sqrt{N}} \xrightarrow{d} N(0, 1).$$

Notes

- It says that standardized sample mean follows a normal distribution for ANY population distribution.
- Both LLN and CLT describe the asymptotic properties (large sample size).

Recap

1. Diagram

- General Diagram for Data Analytics.
- Connecting probability, vectors/numbers, and statistics

2. Probability

- Probability distributions as data generating process
- Basic probability distributions (discrete vs. continuous support)
- From random variables to numbers/vectors

3. Vectors/Numbers

- Three types of numbers: nominal, ordinal, and cardinal
- Centrality and variability

4. Statistics

- Three criteria
- LLN and CLT