MTH499/599 Lecture Notes 02

Donghui Yan

Department of Math, Umass Dartmouth

December 23, 2015

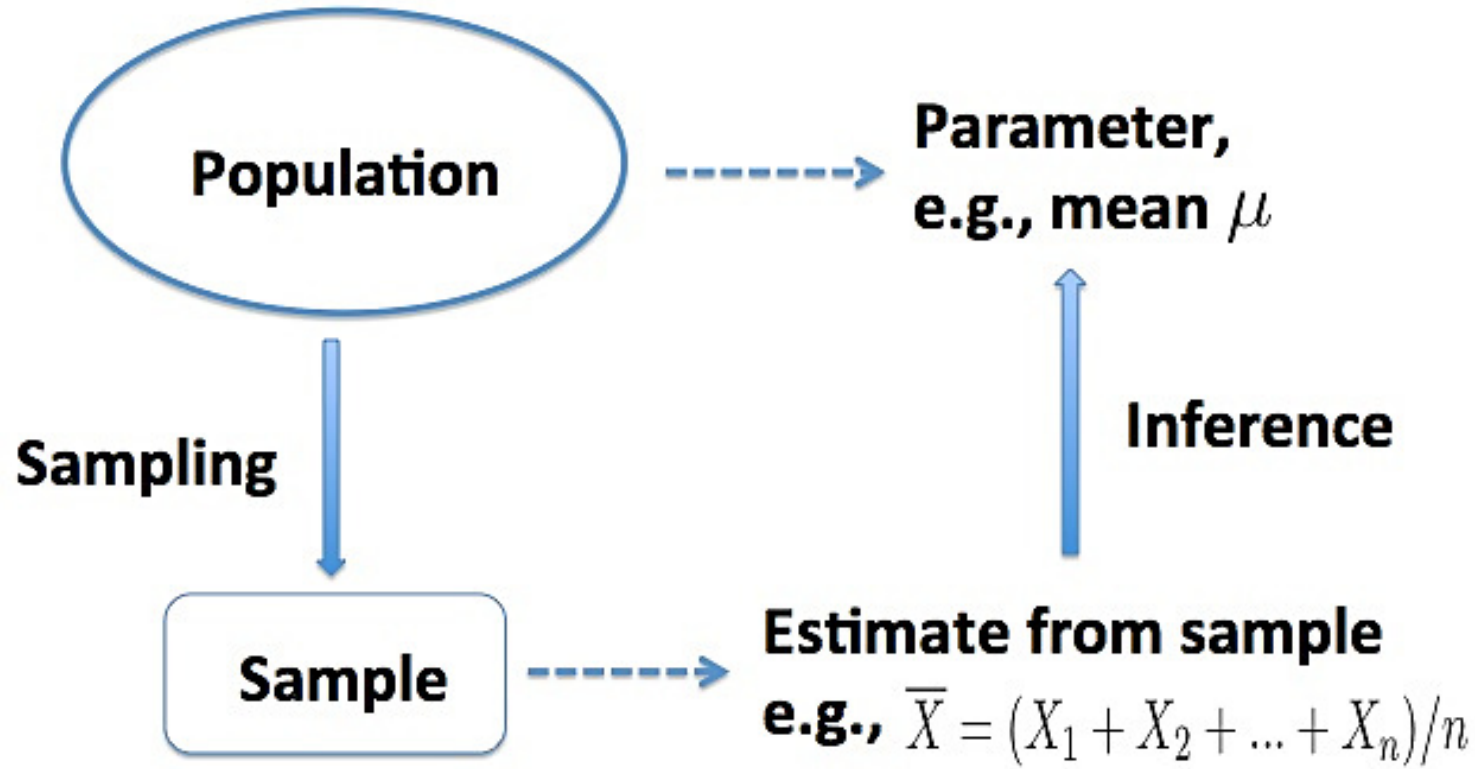
Outline

- Population and sampling
- The binomial, normal and t distribution
- Sampling distribution
- A little matrix algebra
- Basic optimization

Data collection

- Data is the essential part of data science
- To collect data, one needs to identify the population
  - The collection of all values, quantitative or qualitative, for a particular characteristic (or trait) of all units of interest
- In some cases, the entire population is used
  - Suitable when population size is small and easily accessible
- Other cases, difficult or not possible to get all data in population
  - The population is too big or even infinite, or keeps on changing
  - One way is sampling
    - To get a sub-collection of values from the population
    - *Sampling* as a stylish way of saying *data collection*
  - Other solutions?

Illustration of sampling



Population and sample

- Population in statistics has a special meaning
  - Unit (element) $\leftrightarrow$ Population
  - Variables (characteristic) $\leftrightarrow$ Statistical population
  - Examples.

Population	Unit (element)	Variable (characteristic)
All students@UMassD	student	GPA, # Credits, major
All campus restaurant	restaurant	# employees, seating capacity, menu
All books in a library	book	Cost, # pages, author
US population	individual in US	Age, hight, weight, gender

Binomial r.v.

- Repeat coin tossing for n times $\Longrightarrow$ Binomial r.v.
  - $X \triangleq$ # successes out of n trials
  - $\Omega = \{0, 1, 2, ..., n\}$
- Four characteristics
  - The number of trials is fixed, n
  - Two possible outcomes, 'H' or 'T' (call one of these success)
    - $\Omega = \{\text{Success}, \text{Failure}\}$
  - Probability of success remains the same, p, for each trial
  - Individual trials are independent
- Random variable X is called a *Binomial random variable*.

Probability distribution of a binomial r.v.

- We have the following structure
  - $P(\text{Success in a trial}) = p$
  - $\Omega = \{0, 1, 2, ..., n\}$
- The probability of k successes out of n trials, $k = 0, 1, ..., n$.

X	Probability
0	$\binom{n}{0}p^0(1-p)^n$
1	$\binom{n}{1}p^1(1-p)^{n-1}$
2	$\binom{n}{2}p^2(1-p)^{n-2}$
...	...
k	$\binom{n}{k}p^k(1-p)^{n-k}$
...	...
n	$\binom{n}{n}p^n(1-p)^0$

Normal distribution

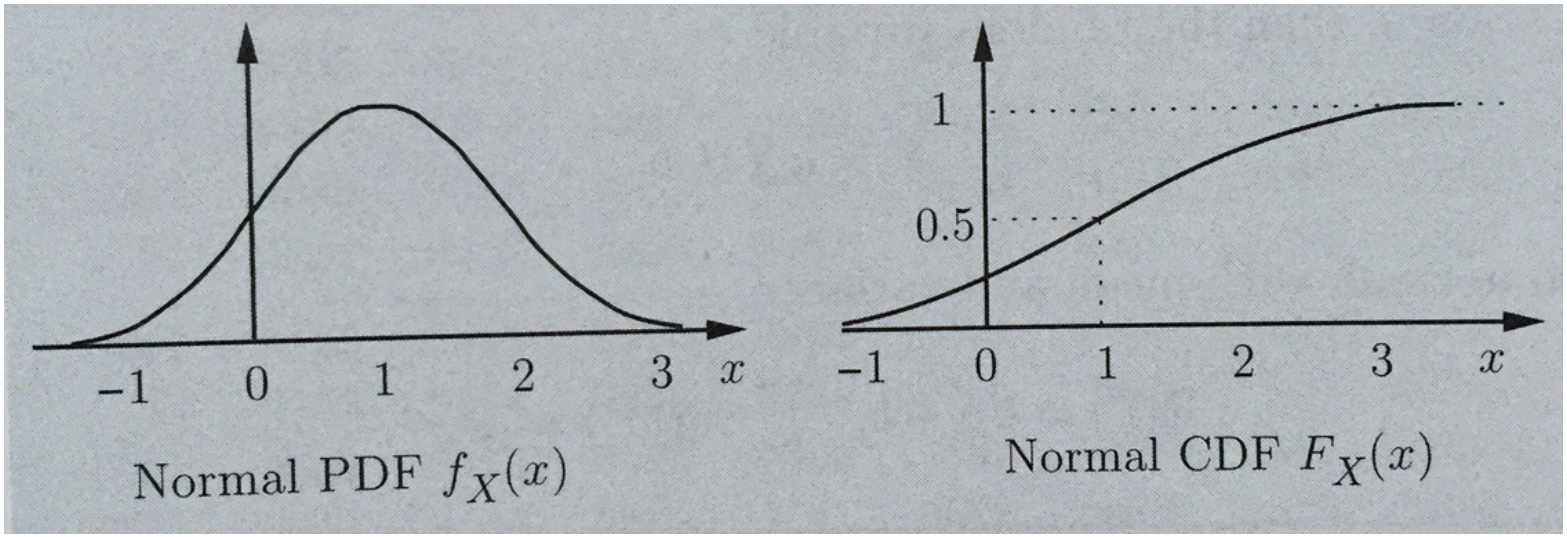
- Commonly attributed to C. F. Gauss
  - *Theoria motus corporum coelestium in sectionibus conicis solem ambientium* (1809)
  - Also called *Gaussian distribution*
- One of the most widely used probability distributions in practice
  - Due to *Central Limit Theorem* (Laplace, 1810).

Normal distribution

- A normal r.v. $X \sim \mathcal{N}(\mu, \sigma^2)$ with mean μ and variance σ^2
- X has PDF

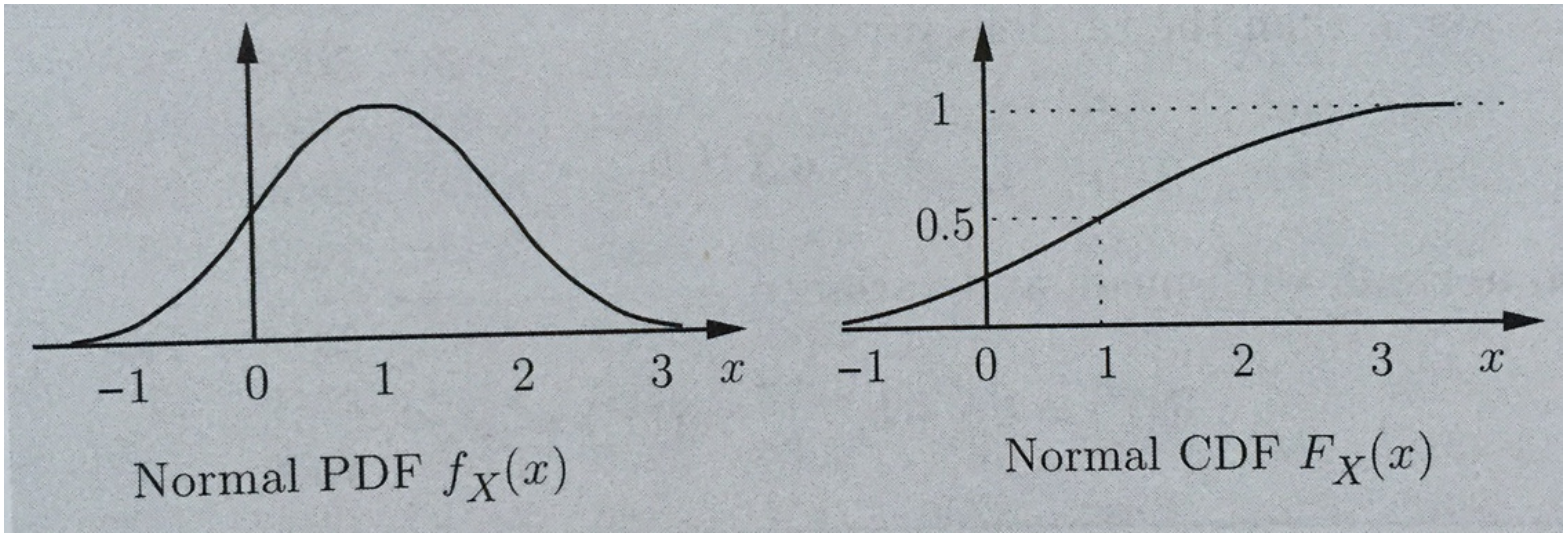
$$f_X(x) = \frac{1}{2\pi\sigma} e^{-(x-\mu)^2/(2\sigma^2)}$$

- $\mathbb{E}X = \mu$, $Var(X) = \sigma^2$.



Properties of normal distribution

- Symmetric w.r.t. its mean
- Same mean, mode, median
- Normality preserved by linear transformation, $aX + b$ is normal.

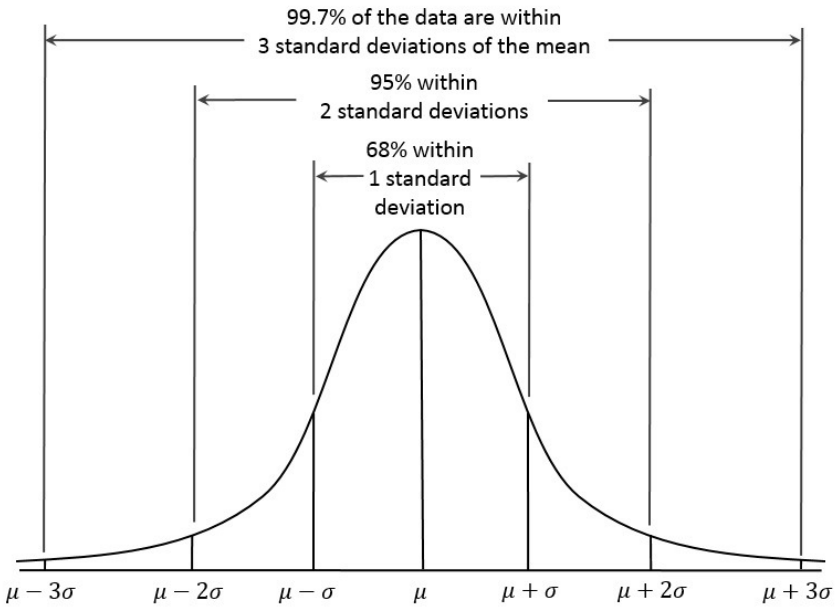


Standard normal

- $X \sim \mathcal{N}(\mu, \sigma^2)$ transformed by

$$X \to Z = \frac{X - \mu}{\sigma} \text{ (Standardization)}$$

- $P(X < x) = P(Z < (x - \mu)/\sigma)$.



Using the normal table

- What is $P(X \geq 80)$ if $X \sim \mathcal{N}(60, 20^2)$?
- $P(X > 80) = P(Z > (80 - 60)/20) = P(Z > 1)$.

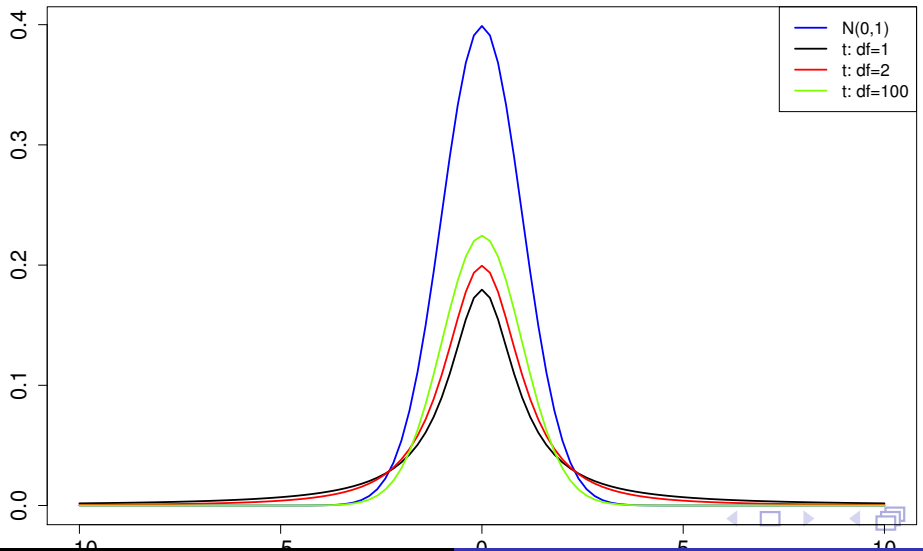
	.00	.01	.02	.03	.04	.05	.06	.07	.08	.09
0.0	.5000	.5040	.5080	.5120	.5160	.5199	.5239	.5279	.5319	.5359
0.1	.5398	.5438	.5478	.5517	.5557	.5596	.5636	.5675	.5714	.5753
0.2	.5793	.5832	.5871	.5910	.5948	.5987	.6026	.6064	.6103	.6141
0.3	.6179	.6217	.6255	.6293	.6331	.6368	.6406	.6443	.6480	.6517
0.4	.6554	.6591	.6628	.6664	.6700	.6736	.6772	.6808	.6844	.6879
0.5	.6915	.6950	.6985	.7019	.7054	.7088	.7123	.7157	.7190	.7224
0.6	.7257	.7291	.7324	.7357	.7389	.7422	.7454	.7486	.7517	.7549
0.7	.7580	.7611	.7642	.7673	.7704	.7734	.7764	.7794	.7823	.7852
0.8	.7881	.7910	.7939	.7967	.7995	.8023	.8051	.8078	.8106	.8133
0.9	.8159	.8186	.8212	.8238	.8264	.8289	.8315	.8340	.8365	.8389
1.0	.8413	.8438	.8461	.8485	.8508	.8531	.8554	.8577	.8599	.8621
1.1	.8643	.8665	.8686	.8708	.8729	.8749	.8770	.8790	.8810	.8830
1.2	.8849	.8869	.8888	.8907	.8925	.8944	.8962	.8980	.8997	.9015
1.3	.9032	.9049	.9066	.9082	.9099	.9115	.9131	.9147	.9162	.9177
1.4	.9192	.9207	.9222	.9236	.9251	.9265	.9279	.9292	.9306	.9319
1.5	.9332	.9345	.9357	.9370	.9382	.9394	.9406	.9418	.9429	.9441
1.6	.9452	.9463	.9474	.9484	.9495	.9505	.9515	.9525	.9535	.9545

t-distribution

- Density specified as

$$\frac{\Gamma((\nu+1)/2)}{\sqrt{\nu\pi}\Gamma(\nu/2)}(1+x^2/\nu)^{-(\nu+1)/2}$$

  - ν degrees of freedom
  - Mean, median, and mode all 0; tends to $\mathcal{N}(0,1)$ as $\nu \to \infty$.



t-table

t-table

Table entry for p and C is the critical value t^* with probability p lying to its right and probability C lying between $-t^*$ and t^*.

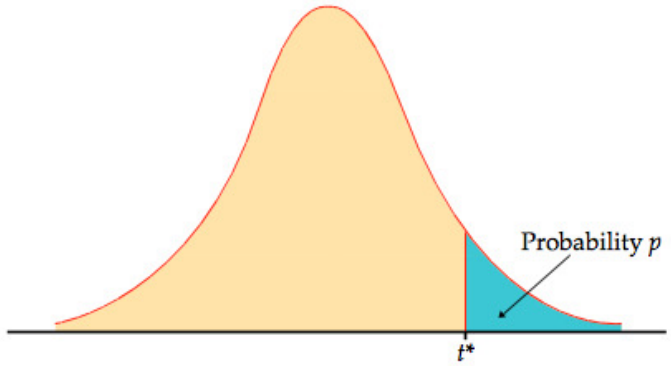


TABLE D

t distribution critical values

	Upper-tail probability p											
df	.25	.20	.15	.10	.05	.025	.02	.01	.005	.0025	.001	.0005
1	1.000	1.376	1.963	3.078	6.314	12.71	15.89	31.82	63.66	127.3	318.3	636.6
2	0.816	1.061	1.386	1.886	2.920	4.303	4.849	6.965	9.925	14.09	22.33	31.60
3	0.765	0.978	1.250	1.638	2.353	3.182	3.482	4.541	5.841	7.453	10.21	12.92
4	0.741	0.941	1.190	1.533	2.132	2.776	2.999	3.747	4.604	5.598	7.173	8.610
5	0.727	0.920	1.156	1.476	2.015	2.571	2.757	3.365	4.032	4.773	5.893	6.869
6	0.718	0.906	1.134	1.440	1.943	2.447	2.612	3.143	3.707	4.317	5.208	5.959
7	0.711	0.896	1.119	1.415	1.895	2.365	2.517	2.998	3.499	4.029	4.785	5.408
8	0.706	0.889	1.108	1.397	1.860	2.306	2.449	2.896	3.355	3.833	4.501	5.041
9	0.703	0.883	1.100	1.383	1.833	2.262	2.398	2.821	3.250	3.690	4.297	4.781
10	0.700	0.879	1.093	1.372	1.812	2.228	2.359	2.764	3.169	3.581	4.144	4.587
11	0.697	0.876	1.088	1.363	1.796	2.201	2.328	2.718	3.106	3.497	4.025	4.437
12	0.695	0.873	1.083	1.356	1.782	2.179	2.303	2.681	3.055	3.428	3.930	4.318
13	0.694	0.870	1.079	1.350	1.771	2.160	2.282	2.650	3.012	3.372	3.852	4.221

Central Limit Theorem

- The distribution of sum of large number of *independent* r.v.'s tends to a normal distribution
  - Let $X_1, X_2, ..., X_n$ be independent, then

$$X_1 + X_2 + ... + X_n \Rightarrow \mathcal{N}(0, 1) \text{ as } n \to \infty$$

- Key words
  - Independent (or some notions of independence)
  - Large number
- Partially explains why so many normal distributions in practice.

Sampling distribution of the sample mean

- Given a sample $X_1, ..., X_n$, the sample mean is defined as

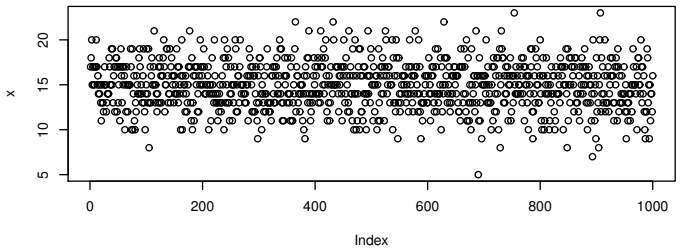
$$\overline{X} = (X_1 + ... + X_n)/n$$

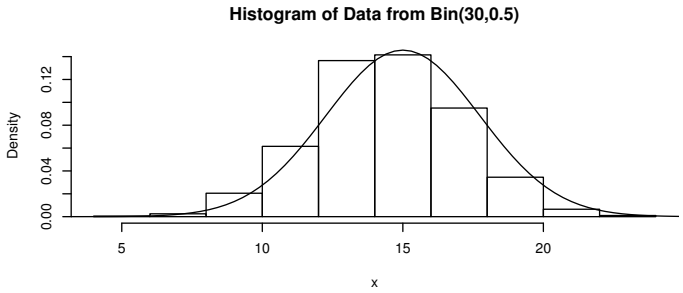
  - Value of $\overline{X}$ fluctuates as different samples are used in calculation
- By CLT, $\overline{X}$ also tends to a normal distribution with
  - Mean $\mathbb{E}(\overline{X}) = \mu$
  - Variance

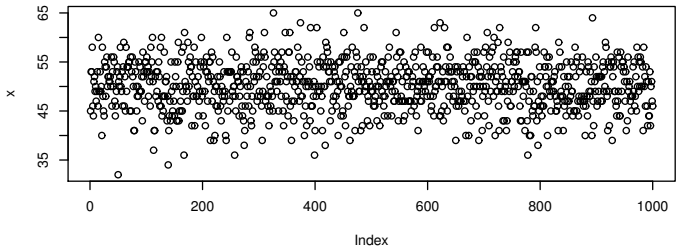
$$\begin{aligned} Var(\overline{X}) &= (Var(X_1) + ... + Var(X_n))/n^2 \\ &= \sigma^2/n. \end{aligned}$$

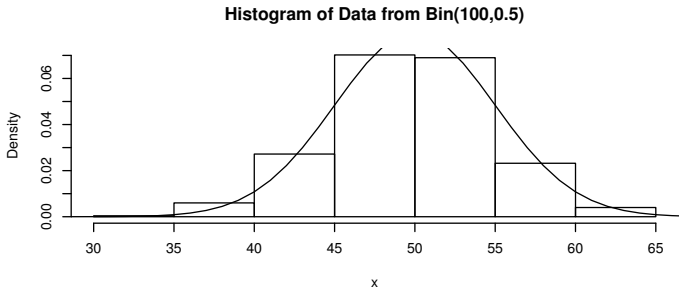
Sampling distribution

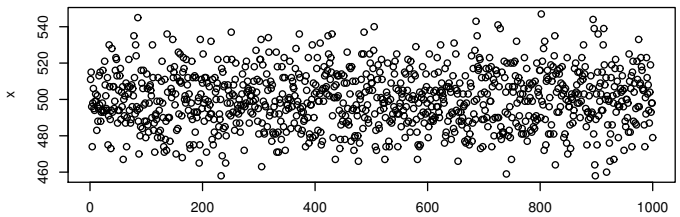
- Sampling distribution of Bin(n, 0.5) with n=30,100,1000

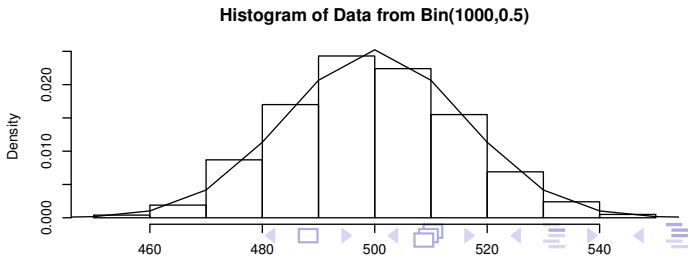






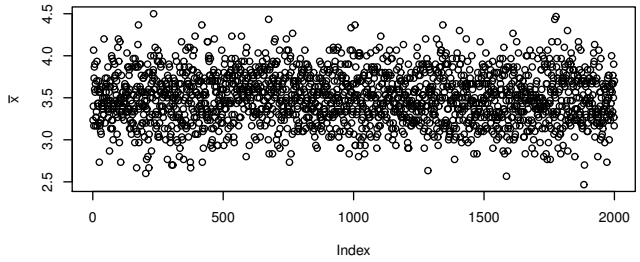


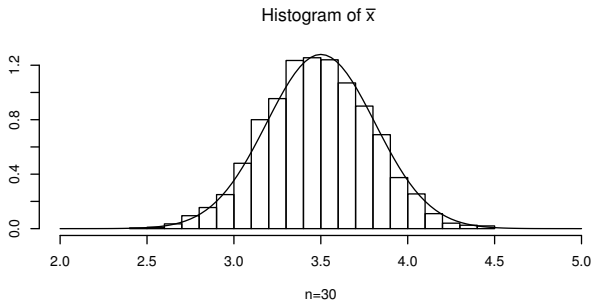


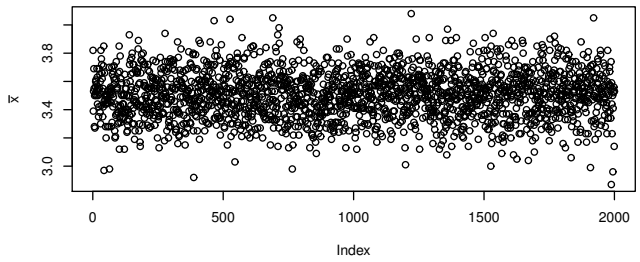


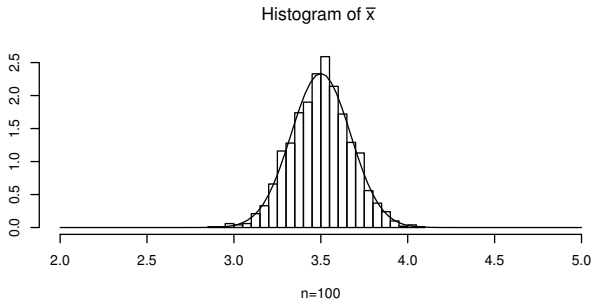
Sampling distribution

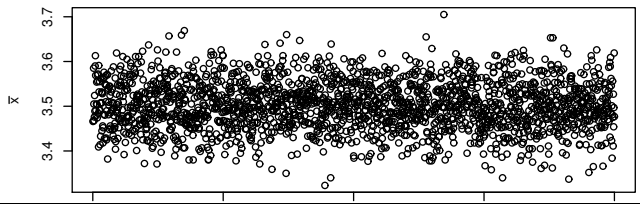
- Sampling distribution of average # dots when rolling a die

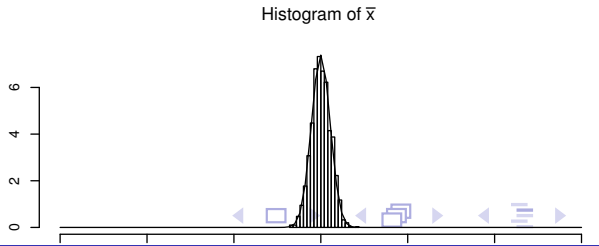












A little matrix algebra

- Let $\boldsymbol{A}_{n\times n}, B$ be matrices. Then

$$Trace(\boldsymbol{AB}) = Trace(\boldsymbol{BA}).$$

- Let $\mathbf{A}$ be a symmetric matrix. Then $\mathbf{A}$ is positive semidefinite iff

$$\boldsymbol{x}^T \boldsymbol{Ay} \geq 0 \text{ for all vecrors } \boldsymbol{x}, \boldsymbol{y} \neq 0.$$

- Let matrix $\boldsymbol{A}$ be symmetric. Then

$$\arg\max_{w} \frac{\boldsymbol{w}^T A \boldsymbol{w}}{\boldsymbol{w}^T \boldsymbol{w}} = \lambda_1$$

where λ_1 is the largest eigenvalue of $\boldsymbol{A}$.

A little matrix algebra

- Let $\mathbf{X}$ be a vector. Then

$$Cov(\mathbf{X}) = \mathbb{E}[(\mathbf{X} - \mathbb{E}(X))(\mathbf{X} - \mathbb{E}(X))^T]$$

- $\mathbf{A}$ and $\mathbf{B}$ constant matrices, $\mathbf{c}$ and $\mathbf{d}$ constant vectors. Then

$$\begin{aligned} Cov(\mathbf{Ax_1} + \mathbf{c}, \mathbf{Bx_2} + \mathbf{d}) &= \mathbf{A}Cov(\mathbf{x_1}, \mathbf{x_2})\mathbf{B}^T \\ &\triangleq \mathbf{A} < \mathbf{x_1}, \mathbf{x_2} > \mathbf{B}^T \end{aligned}$$

- Let matrix $\boldsymbol{W}$ be symmetric. Then

$$\frac{\partial}{\partial \boldsymbol{s}}(Y - \boldsymbol{As})^T \boldsymbol{W}(Y - \boldsymbol{As}) = -2\boldsymbol{A}^T \boldsymbol{W}(Y - \boldsymbol{As}).$$

Basic optimization

- Smooth function $f(x)$ achieves maximum at $x = x_0$ if

$$\frac{\partial f(x)}{\partial x} \mid_{x=x_0} = 0, f''(x_0) \leq 0.$$

- Solve constrained optimization problem

$$\begin{aligned} \max_{x,y} \quad & f(x,y) \\ \text{s.t.} \quad & g(x,y) = 0 \end{aligned}$$

by *Lagrange multiplier* $\mathcal{L}(x,y,\lambda) = f(x,y) + \lambda g(x,y)$ with

$$\nabla_{x,y,\lambda} \mathcal{L}(x,y,\lambda) = 0.$$