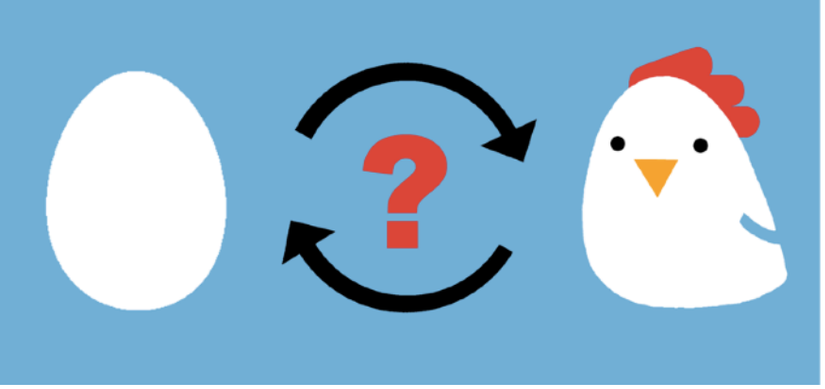


Causality and Machine Learning
(80-816/516)

Classes 3 (Jan 21, 2025)

From Probability Theory to Statistics

Instructor:

Kun Zhang (kunz1@cmu.edu)

Zoom link: https://cmu.zoom.us/j/8214572323)

Office Hours: W 3:00–4:00PM (on Zoom or in person); other times by appointment

Question to Answer

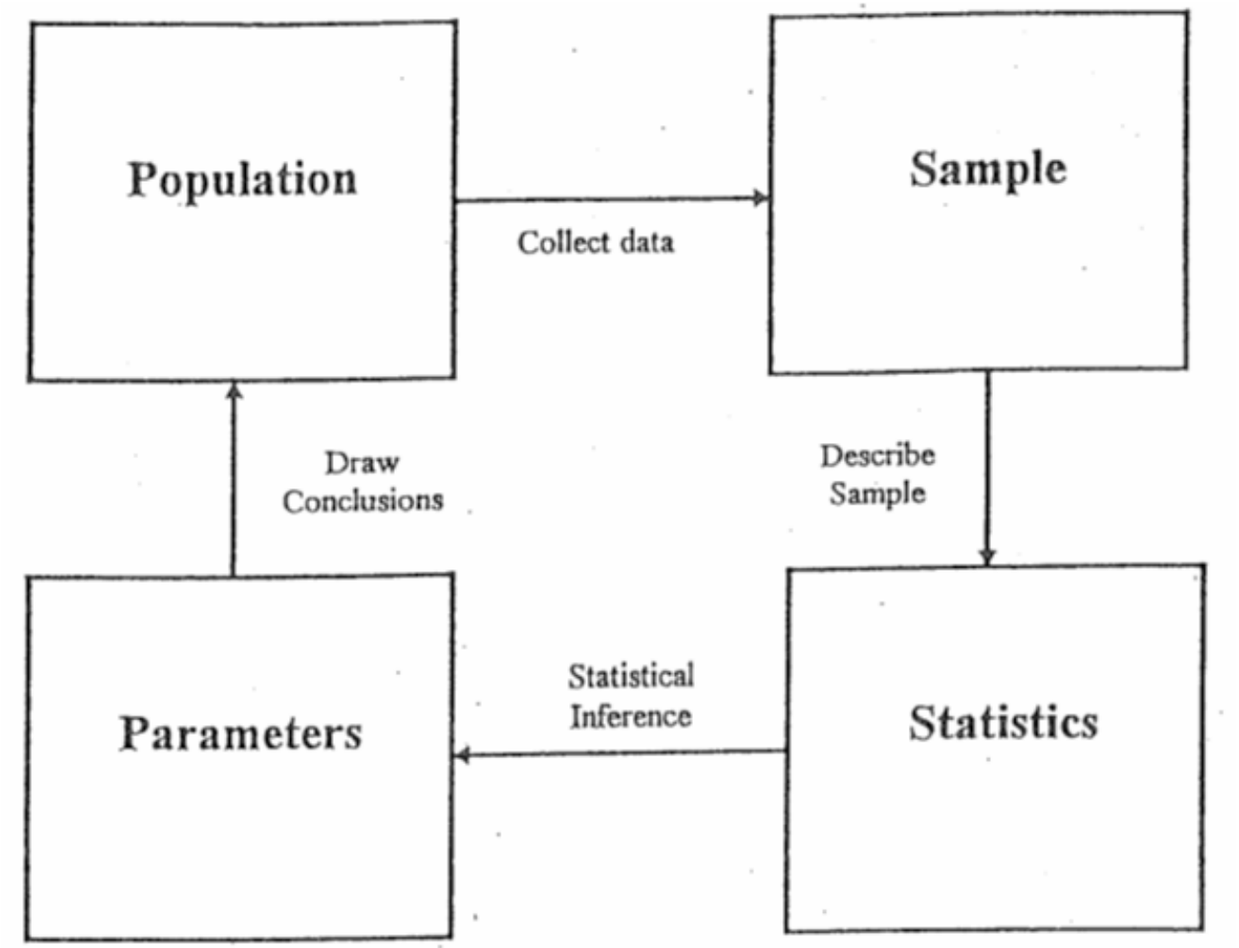
- Will the sun rise tomorrow?

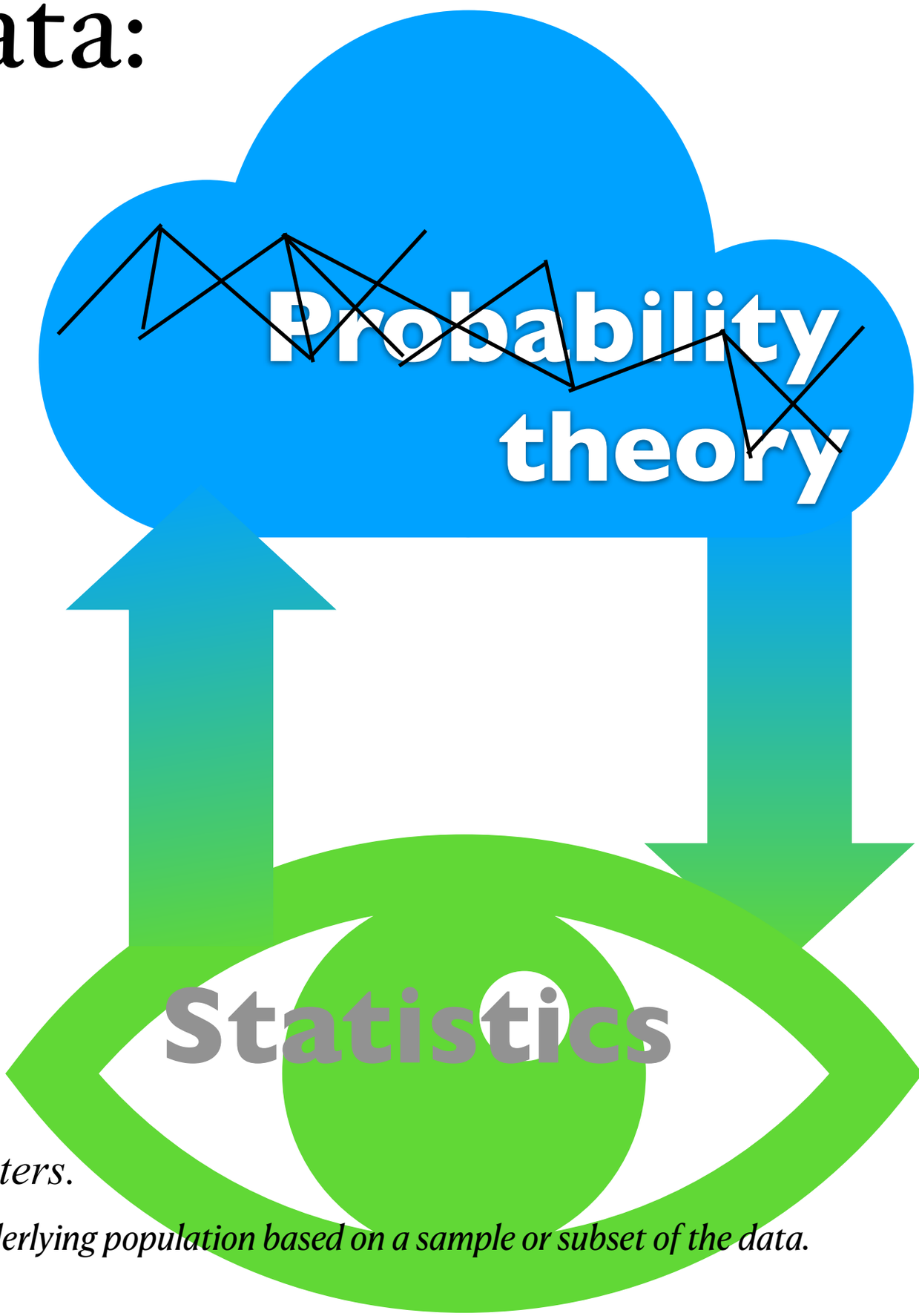
- Matters of fact, which are the second objects of human reason, are not ascertained in the same manner; nor is our evidence of their truth, however great, of a like nature with the foregoing. *The contrary of every matter of fact is still possible, because it can never imply a contradiction, and is conceived by the mind with the same facility and distinctness, as if ever so conformable to reality. That the sun will not rise tomorrow is no less intelligible a proposition, and implies no more contradiction, than the affirmation, that it will rise. We should in vain, therefore, attempt to demonstrate its falsehood.* Were it demonstratively false, it would imply a contradiction, and could never be distinctly conceived by the mind. —*An Enquiry Concerning Human Understanding* (1772). Hackett Publ Co. 1993; Chapter on Cause and Effect.

- Do you agree? What can we do?

Making Use of Data: Statistics…

- Connection between probability theory & statistics





Using sample statistics to estimate population parameters.

Statistical inference is the process of drawing conclusions about an underlying population based on a sample or subset of the data.

Why Probability?

- Probability is a tool to understand
  - what is randomness

- Random experiment: an experiment whose outcome is not known to us

To Define Probability

- Assume all possible outcomes of the random experiment are known
- Consider the following examples…

Why Probabilistic Thinking: A Story

The actual science of logic is conversant at present only with things either certain, impossible, or entirely doubtful, none of which (fortunately) we have to reason on. Therefore the true logic for this world is the calculus of Probabilities, which takes account of the magnitude of the probability which is, or ought to be, in a reasonable man's mind.

James Clerk Maxwell (1850)

Events

- Event: A set of outcomes of an experiment, as a subset of the **sample space** Ω

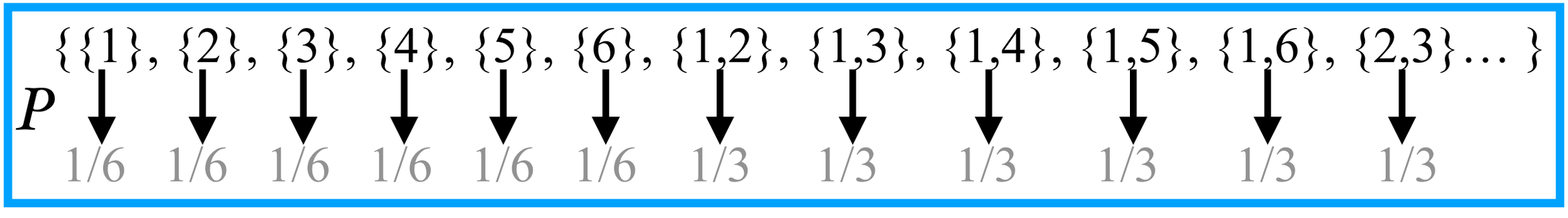
$\Omega = \{1, 2, 3, 4, 5, 6\}$

  - E.g., get 6 from rolling a die
  - Incompatible events, e.g., get 6 and get 2
  - Complementary events: Events A and [not A]
- Possible events form a field of sets (**event space**) F
  - {{1}, {2}, {3}, {4}, {5}, {6}, {1,2}, {1,3}, {1,4}, {1,5}, {1,6}, {2,3}… }
  - A field of sets over a nonempty set Ω is any collection of subsets of Ω that is closed under the intersection and union of pairs of sets and under complements of individual sets

Probability: Axioms

- Probability measure on F over Ω : a function P defined on all sets in F and assigning each set (event) a real number satisfying:

  - Nonnegativity: $P(A) \in \Re,\ P(A) \geq 0,\ \forall A \in F$
  - Normalization: $P(\Omega) = 1$
  - Additivity: $P(A \cup B) = P(A) + P(B)$ for incompatible A and B



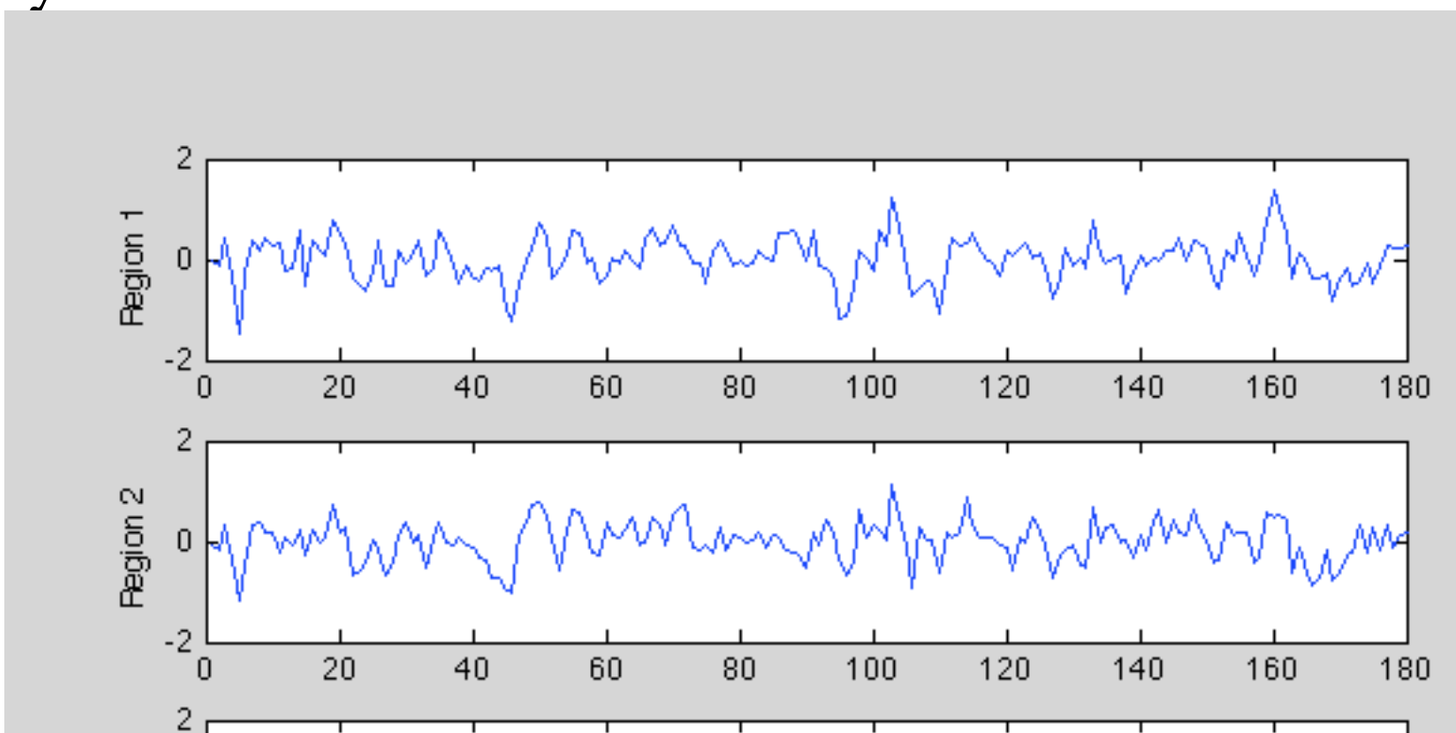
Conditional Probability

- P(A|B): probability of A given B has occurred
  - probability of getting 2 from the 2nd die given getting 6 from the 1st
- Suppose $P(B) \neq 0$; it is then defined to be $P(A \cap B) / P(B)$
- Fix B with $P(B) \neq 0$; then $P(\cdot|B)$ satisfies the probability axioms
- A and B are **independent** iff $P(A \cap B) = P(A)\ P(B)$
  - I.e., $P(A|B) = P(A)$
- **Product** and **sum rules** are fundamental: :-)
  - $P(A \cap B) = P(A|B)\ P(B)$
  - $P(A) = P(A \cap B) + P(A \cap \sim B)$

Roughly Speaking, Random Variables and Their Realizations...

- A random variable is a variable whose possible values are numerical outcomes of a random phenomenon.

- E.g., the sum of what I got on the two dice, the height of a MBZUAI student, the daily return of a stock...



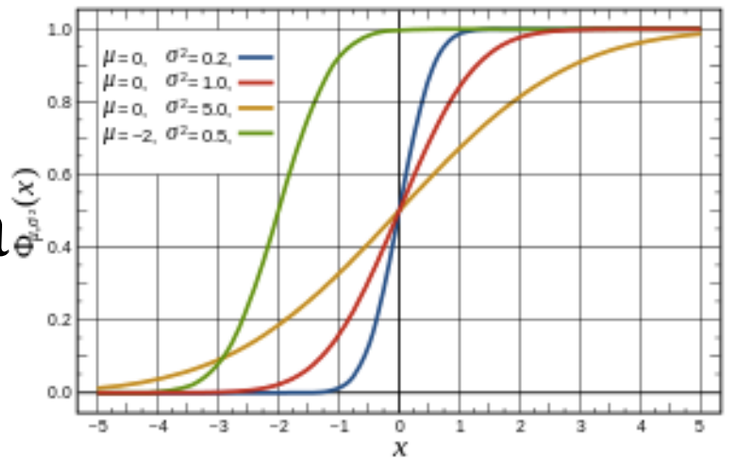
Discrete vs. Continuous Random Variables

- A random variable is **discrete** if its range (the set of values that it can take) is finite or at most countably infinite
  - E.g., the sum of what I got on the two dice
  - $P(X{=}k) = P(\{\omega: X(\omega) = k\})$; tabular representation for the probability mass function (PMF)
- A random variable is **continuous** (not discrete) if its range (the set of values that it can take) is uncountably infinite
  - E.g., the height of a CMU student
  - $P(a \leq X \leq b) = P(\{\omega: a \leq X(\omega) \leq b\})$

How to Specify Prob. Measures of Random Variables

- PMFs for *discrete* variables
- Cumulative distribution function (CDF): A function $F_X: \mathbb{R} \rightarrow [0,1]$ which specifies a probability measure as

$$F_X(x) \triangleq P(X \leq x)$$



- Probability density function (PDF): derivative of the CDF for *continuous* variables whose CDFs are differentiable everywhere

$$p_X(x) \triangleq \frac{dF_X(x)}{dx}$$

Probability Measure: Examples

- Discrete variables:
  - *Bernoulli*(*p*):

the discrete probability distribution of a random variable which takes the value 1 with probability *p* and the value 0 with probability *q=1-p*.

$$P(X=1) = p.$$

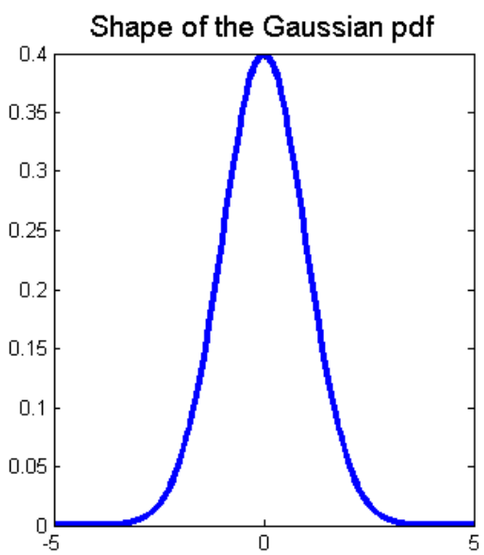
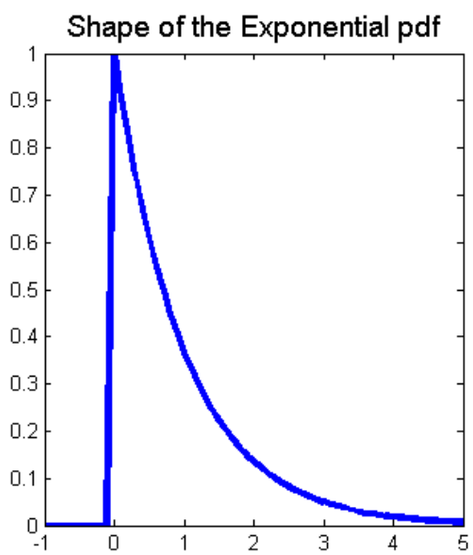
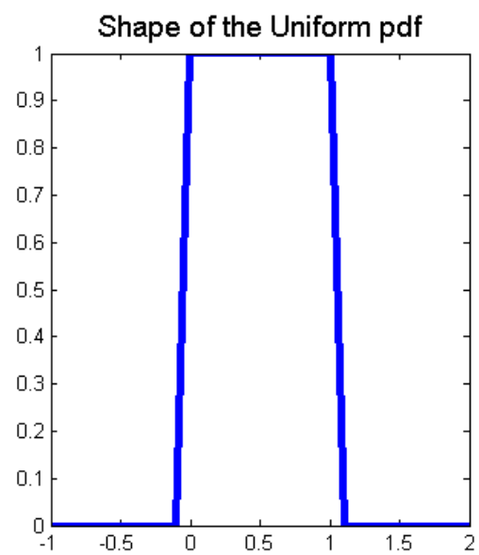
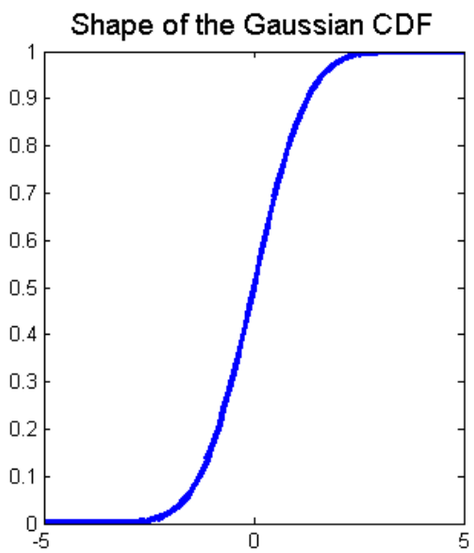
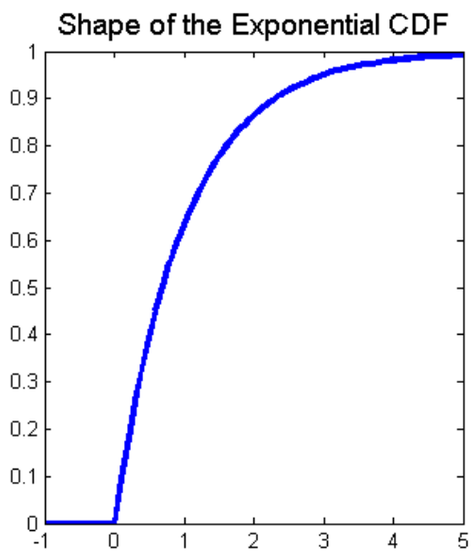
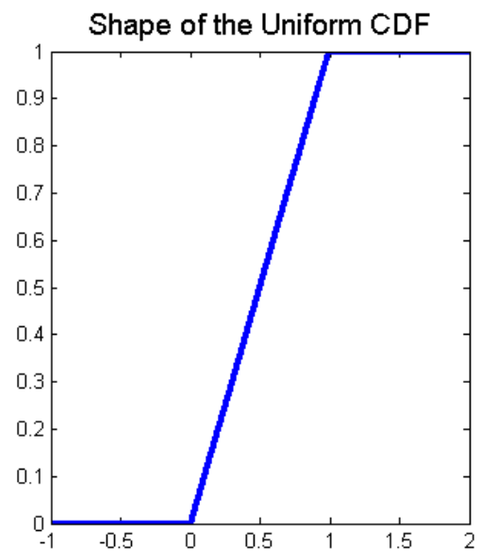
  - *Binomial*(*n*,*p*):

the discrete probability distribution of the number of successes in a sequence of *n independent* experiments, each with its own boolean-valued outcome: success (with probability *p*) or failure (with probability *q=1−p*).

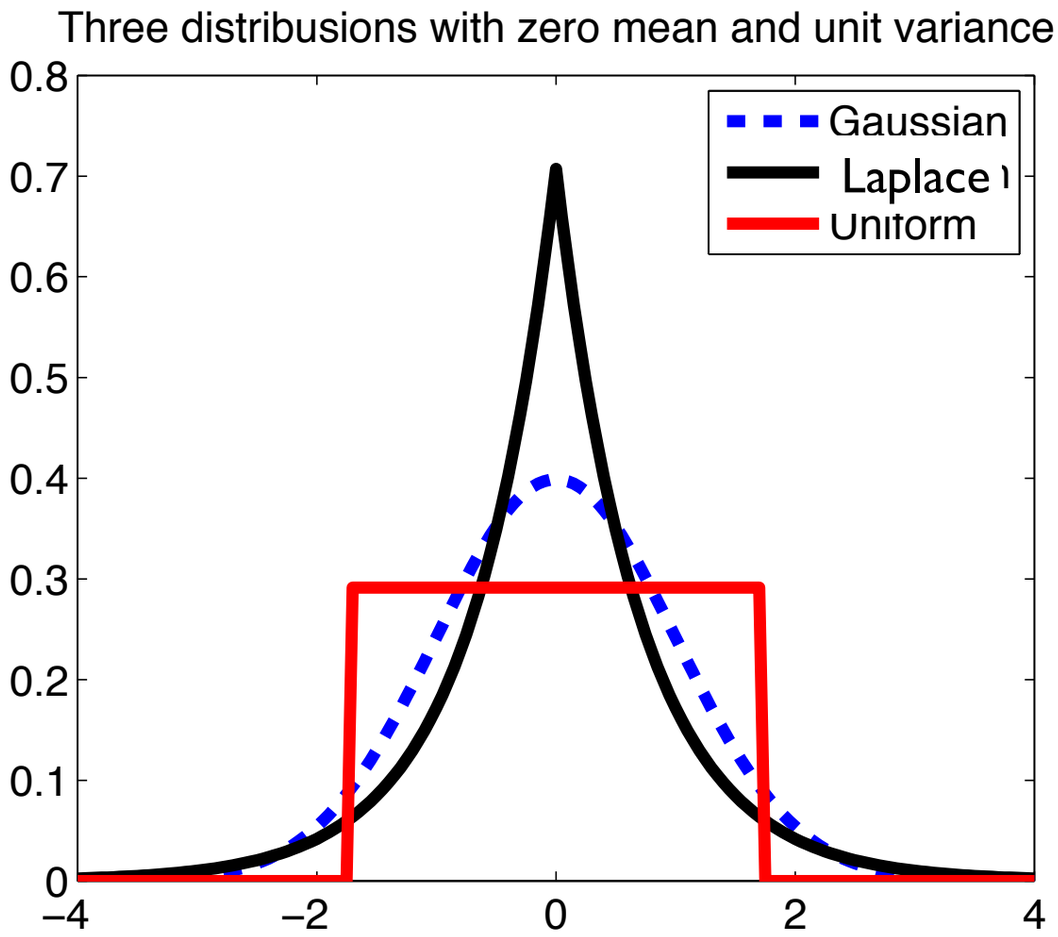
$$P(X=k) = \binom{n}{k} p^k (1-p)^{n-k}$$

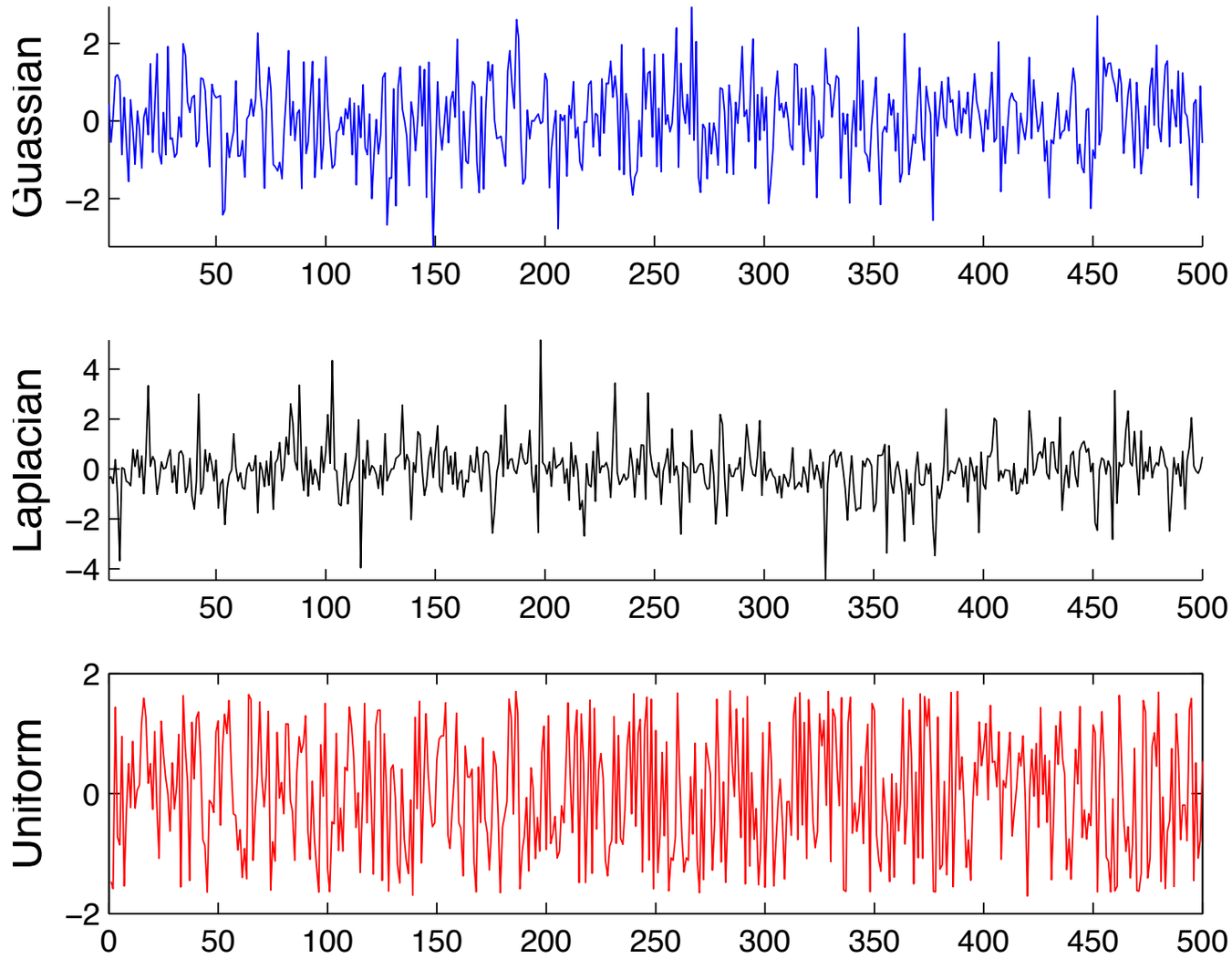
Probability Measure: Examples

- Continuous variables:

	Gaussian	Exponential	Uniform
PDF			
CDF			

Some Distributions





Conditional Distributions

- Joint/marginal PMFs, CDFs, and PDFs: *straightforward*
- What is the probability distribution over X, when we know Y must take a certain value y?
- Discrete case: Provided $P_Y(y) \neq 0$, conditional PMF of X given Y is

$$P_{X|Y} = \frac{P_{XY}(x,y)}{P_Y(y)}$$

- Continuous case: Provided $p_Y(y) \neq 0$, conditional PDF of X given Y is

$$p_{X|Y} = \frac{p_{XY}(x,y)}{p_Y(y)}$$

A Question...

- With 5 coins which are not necessarily fair, how many parameters to represent the joint probability distribution $P(O_1, O_2, \ldots, O_5)$?
- In practice we often need fewer parameters...
- Divide-and-conquer

Statistical Independence

- Two variables X and Y are independent if $F_{XY}(x,y) = F_X(x)\,F_Y(y)$ for all values of x and y. Equivalently,
  - For discrete variables, $P_{XY}(x,y) = P_X(x)P_Y(y)$, or $P_{X|Y}(x|y) = P_X(x)$ whenever $P_Y(y)\neq 0$
  - For continuous variables: p instead of P

Pairwise Independence vs. Mutual Independence

- Pairwise independent: every pair of random variables is independent
- **Mutually independent:** $F_{X1X2\ldots Xn}(x,y) = F_{X1}(x_1)\, F_{X2}(x_2)\ldots F_{Xn}(x_n)$
- Three-coin example: A $\perp\!\!\!\perp$ B; C is determined by A and B but C $\not\perp\!\!\!\perp$ B and C $\not\perp\!\!\!\perp$ A
  - Pairwise independence? Mutual independence?

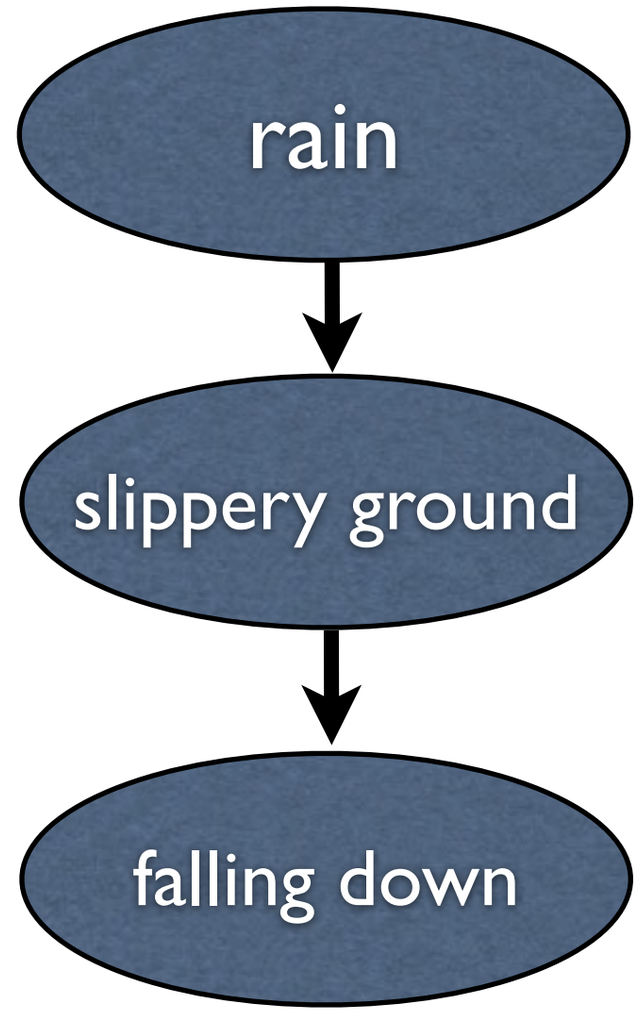
Ways to Produce Dependence

- Common cause underlying them
- causal relations between them
- Selection (conditioning on the effect)!

Another Example

- What if X_i's are not mutually independent but we know they were generated the following way?

$$X_1 \rightarrow X_2 \rightarrow ... \rightarrow X_n$$



Conditional Independence

- Two variables X and Y are conditionally independent given Z if $F_{XY|Z}(x,y|z) = F_{X|Z}(x|z)\, F_{Y|Z}(y|z)$ for all values of x, y and z. Equivalently,
  - For discrete variables, $P_{XY|Z}(x,y|z) = P_{X|Z}(x|z)P_{Y|Z}(y|z)$, or $P_{X|Y,Z}(x|y,z) = P_{X|Z}(x|z)$ whenever $P_{YZ}(y,z)\neq 0$
  - For continuous variables...
- $X \perp\!\!\!\perp Y \mid Z$: If Z is known, Y is not useful when modeling/predicting X

Some Properties of (Conditional) Independence

- Symmetry $X \perp\!\!\!\perp Y \quad \Rightarrow \quad Y \perp\!\!\!\perp X$
- Decomposition $X \perp\!\!\!\perp (A, B) \quad \Rightarrow \quad \text{and} \begin{cases} X \perp\!\!\!\perp A \\ X \perp\!\!\!\perp B \end{cases}$
- Weak union $X \perp\!\!\!\perp (A, B) \quad \Rightarrow \quad \text{and} \begin{cases} X \perp\!\!\!\perp A \mid B \\ X \perp\!\!\!\perp B \mid A \end{cases}$
- Contraction $\left.\begin{matrix} X \perp\!\!\!\perp A \mid B \\ X \perp\!\!\!\perp B \end{matrix}\right\} \text{and} \quad \Rightarrow \quad X \perp\!\!\!\perp (A, B)$

Relationship between independence & conditional independence?

Some Properties of (Conditional) Independence

$P(A,B|X) = P(A,B)$
$\Rightarrow P(A|X) = P(A)$ (by marginalizing B out)

- Symmetry $X \perp\!\!\!\perp Y \Rightarrow Y \perp\!\!\!\perp X$
- Decomposition $X \perp\!\!\!\perp (A, B) \Rightarrow$ and $\begin{cases} X \perp\!\!\!\perp A \\ X \perp\!\!\!\perp B \end{cases}$
- Weak union $X \perp\!\!\!\perp (A, B) \Rightarrow$ and $\begin{cases} X \perp\!\!\!\perp A \mid B \\ X \perp\!\!\!\perp B \mid A \end{cases}$
- Contraction $\left.\begin{matrix} X \perp\!\!\!\perp A \mid B \\ X \perp\!\!\!\perp B \end{matrix}\right\}$ and $\Rightarrow X \perp\!\!\!\perp (A, B)$

Relationship between independence & conditional independence?

Some Properties of (Conditional) Independence

$P(X|A,B) = P(X)$;
$P(X|A) = P(X)$.
$\Rightarrow P(X|A,B) = P(X|A)$, *i.e.*, $X \perp\!\!\!\perp B|A$

- Symmetry $X \perp\!\!\!\perp Y \Rightarrow Y \perp\!\!\!\perp X$
- Decomposition $X \perp\!\!\!\perp (A, B) \Rightarrow$ and $\begin{cases} X \perp\!\!\!\perp A \\ X \perp\!\!\!\perp B \end{cases}$
- Weak union $X \perp\!\!\!\perp (A, B) \Rightarrow$ and $\begin{cases} X \perp\!\!\!\perp A \mid B \\ X \perp\!\!\!\perp B \mid A \end{cases}$
- Contraction $\left.\begin{matrix} X \perp\!\!\!\perp A \mid B \\ X \perp\!\!\!\perp B \end{matrix}\right\}$ and $\Rightarrow X \perp\!\!\!\perp (A, B)$

Relationship between independence & conditional independence?

Relation between Independence and Conditional Independence

- If $X \perp\!\!\!\perp Y$, are they conditionally independent given Z ?

- If $X \perp\!\!\!\perp Y | Z$, are they independent ?

Expectation, Variance, and Standard Deviation

- Expectation:

$$E[g(X)] \triangleq \sum_x g(x) P_X(x) \quad \text{(for discrete varaibles) } or$$

$$E[g(X)] \triangleq \int_{-\infty}^{+\infty} g(x) p_X(x) dx \quad \text{(for continuous varaibles)}$$

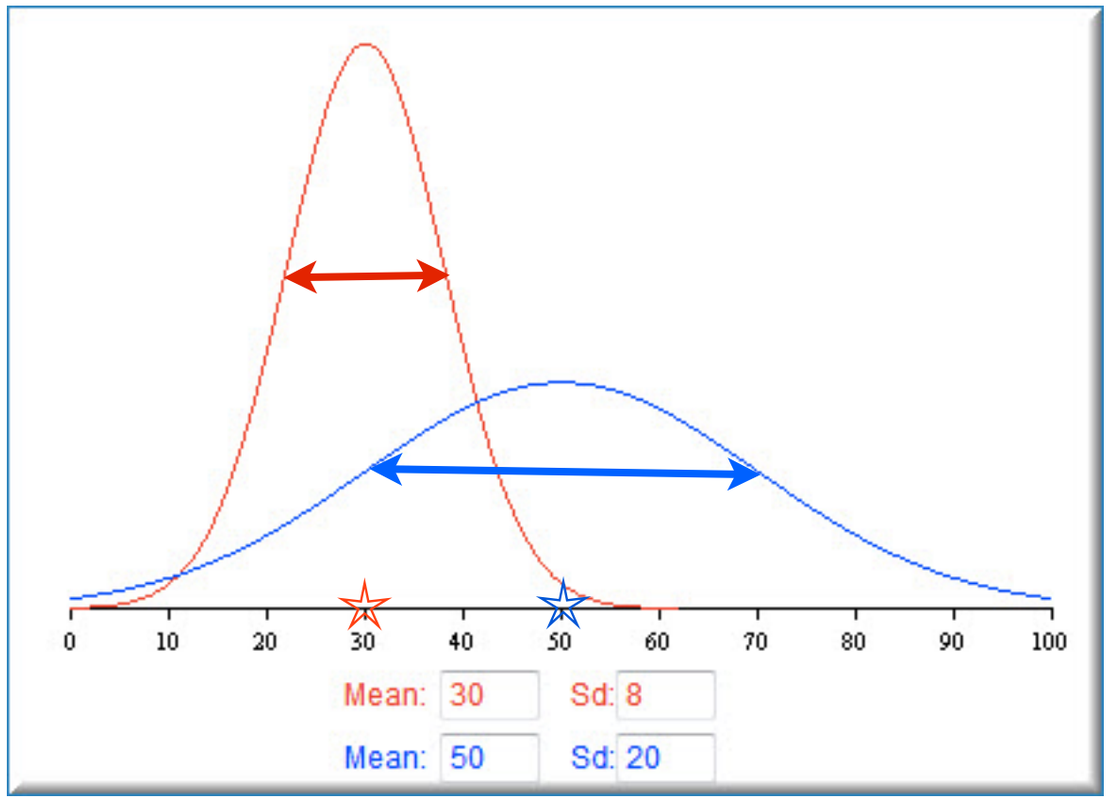
- Mean of X: E[X]

- Variance:

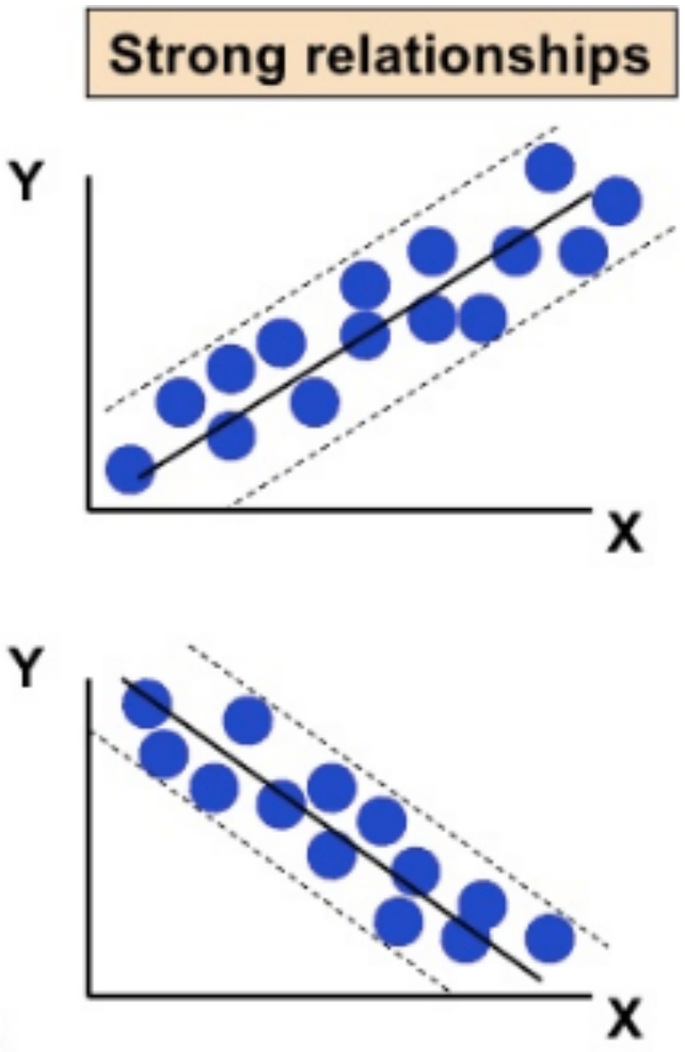
$$Var[X] \triangleq E\{[X - E(X)]^2\}$$

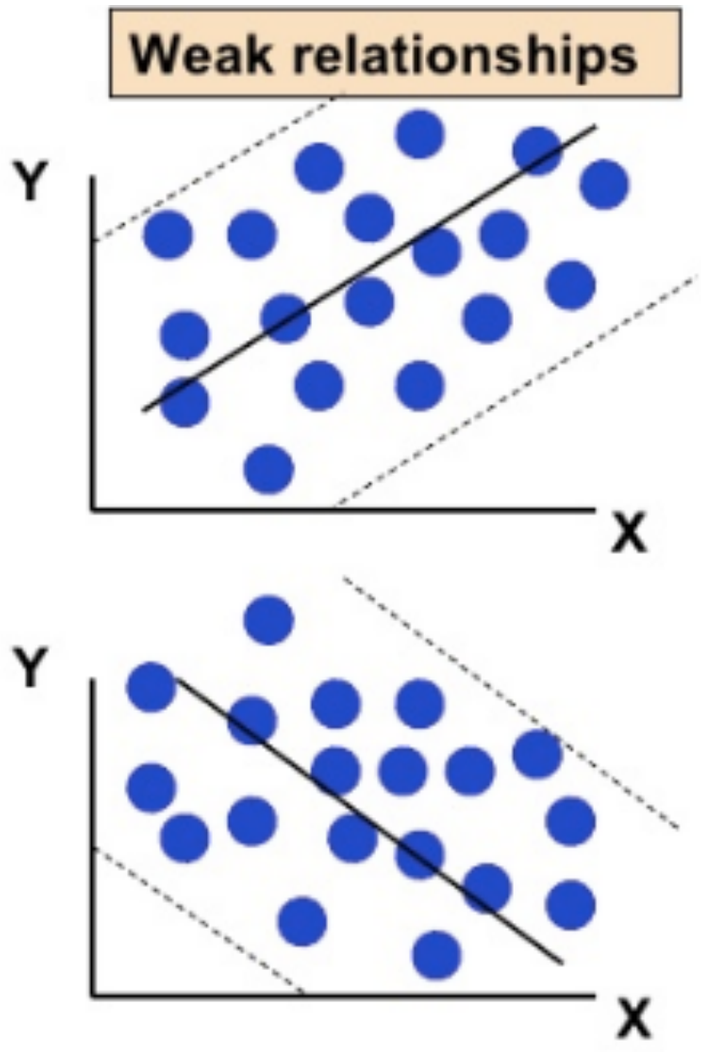
- Standard deviation:

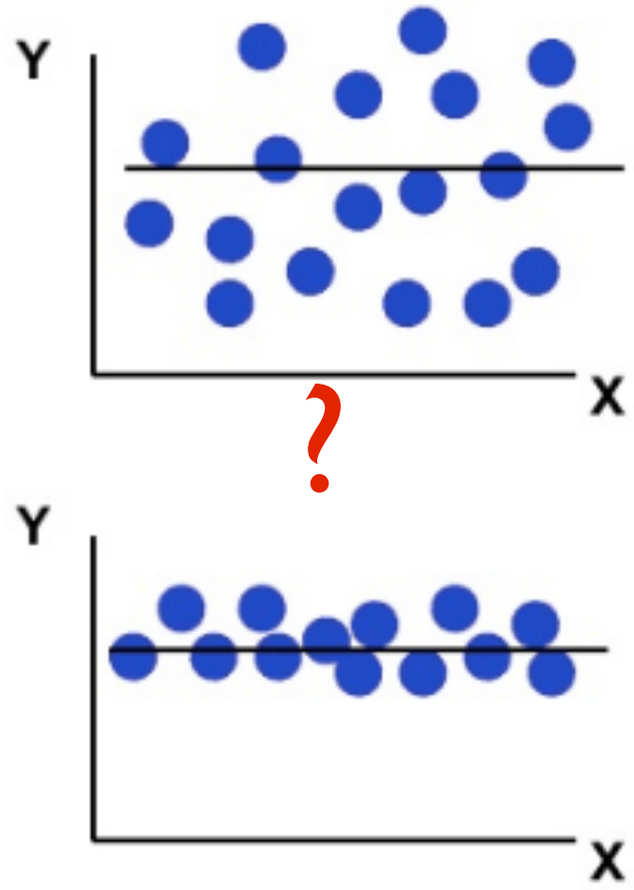
$$Std[X] \triangleq \sqrt{Var[X]}$$



Strong/Weak Relations?







Covariance and Correlation

- Covariance: $Cov[X,Y] \triangleq E[(X - E[X])(Y - E[Y])]$

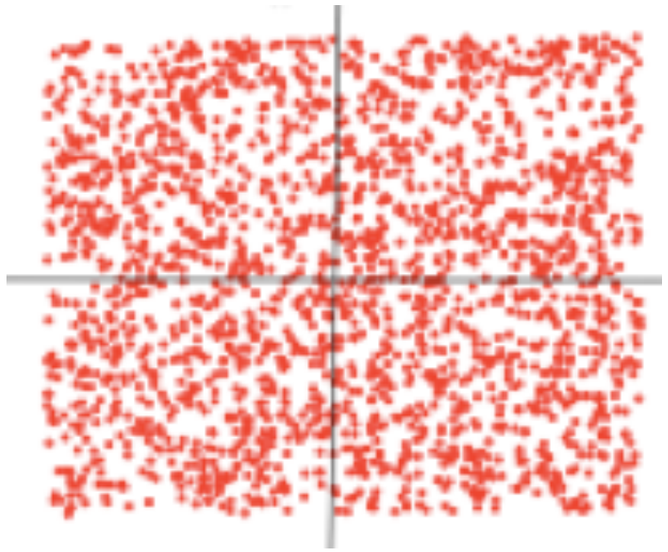
- Uncorrelated if $Cov[X,Y] = 0$

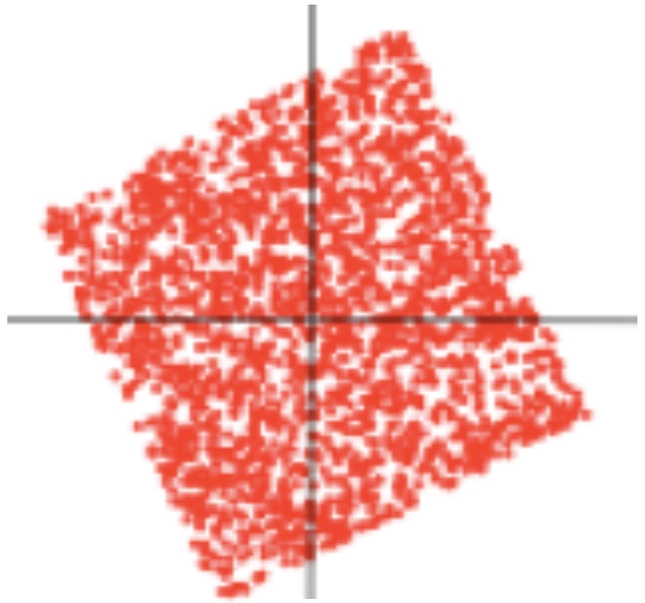
- Correlation: $Corr[X,Y] \triangleq \dfrac{Cov[X,Y]}{\sqrt{Var[X]Var[Y]}}$

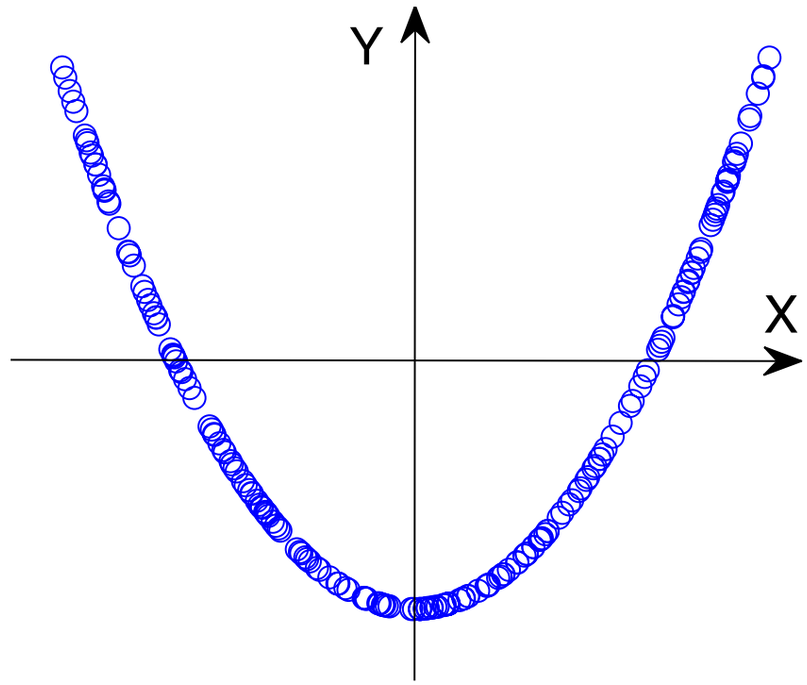
Some Properties of Expectation and Variance

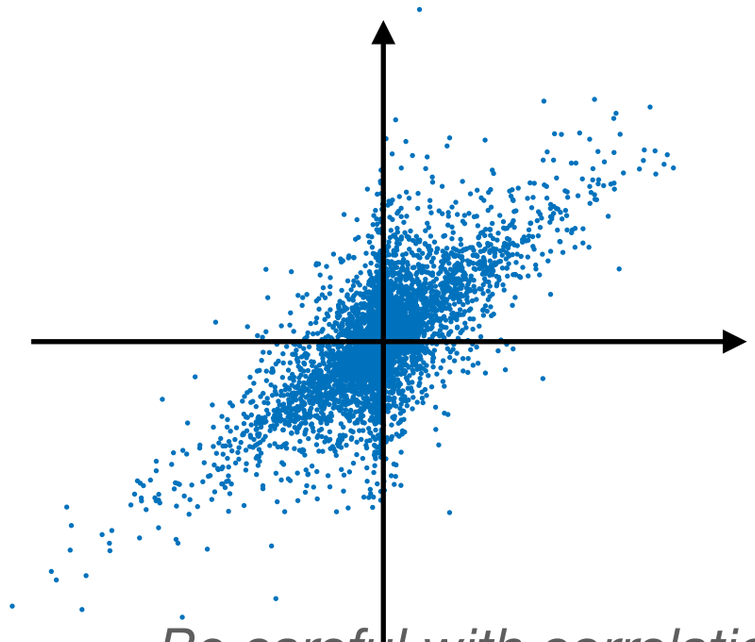
- $E\Big[\sum_{i=1}^{k} a_i X_i\Big] = \sum_{i=1}^{k} E[a_i X_i]$
- $E\Big[\prod_{i=1}^{k} X_i\Big] = \prod_{i=1}^{k} E[X_i]$ if all variables are independent!
- $\mathrm{Var}[aX + b] = a^2 \mathrm{Var}[X]$
- $\mathrm{Var}\Big[\sum_{i=1}^{k} a_i X_i\Big] = \sum_{i=1}^{k} a_i^2 \mathrm{Var}[X_i]$ if all variables are uncorrelated!

Are They Uncorrelated?









Be careful with correlation (covariance)…

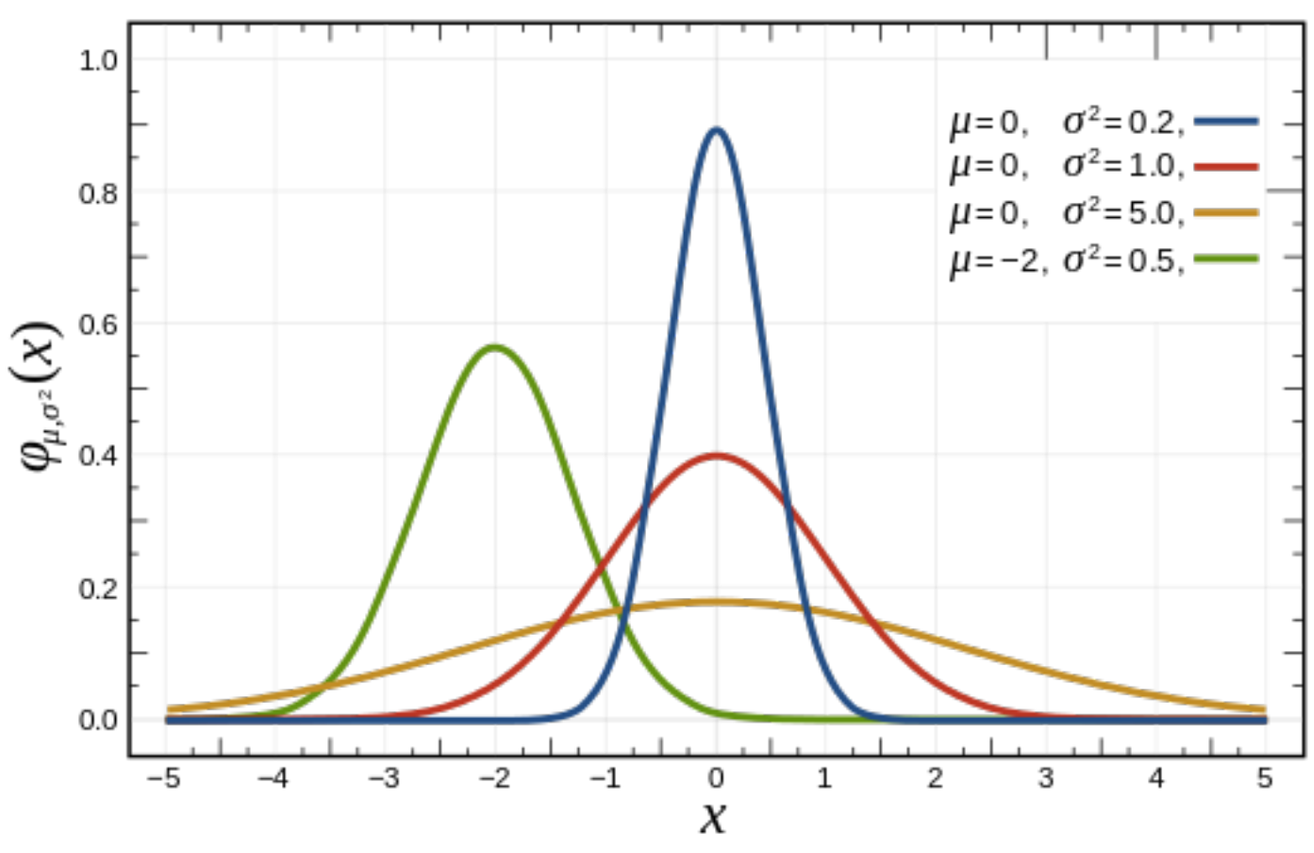
Independence and Uncorrelatedness

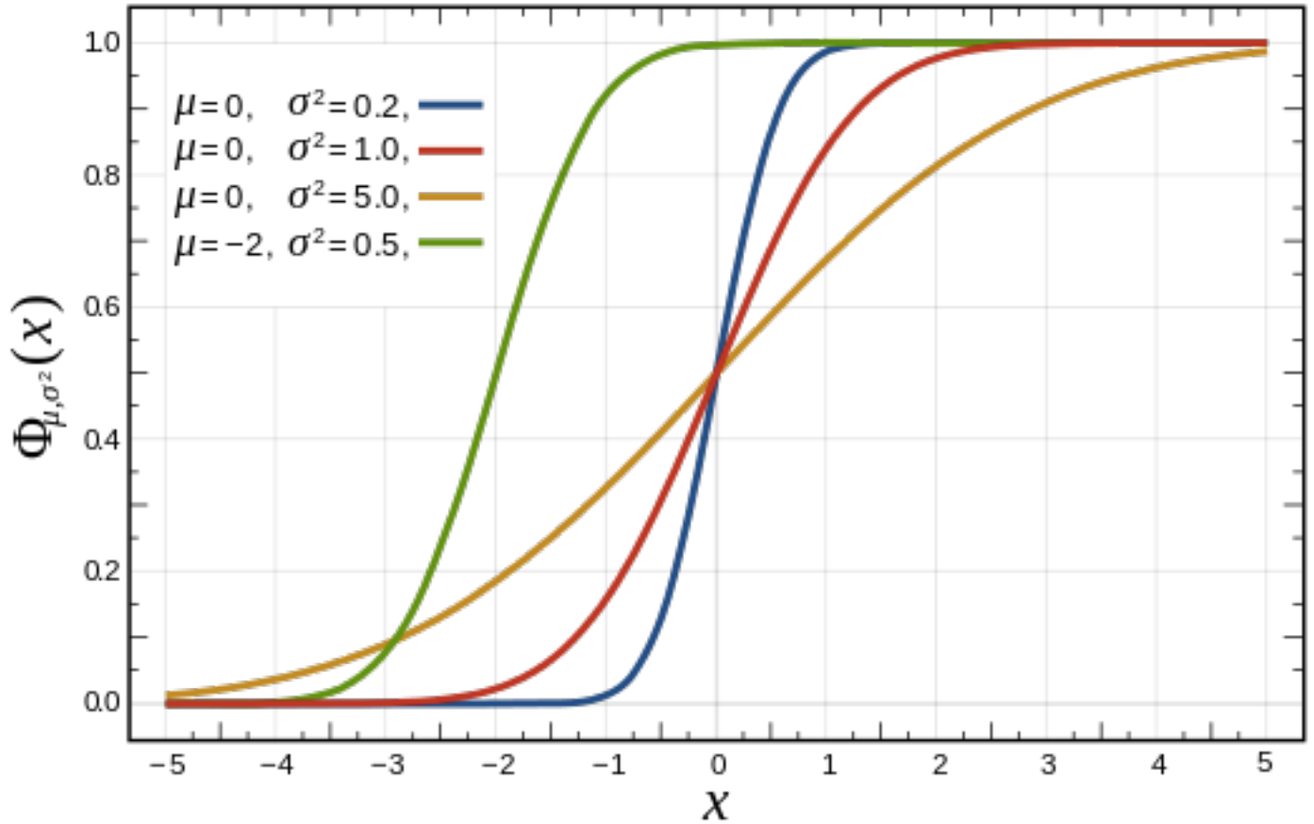
- Independence ⇒ uncorrelatedness
- How about the reverse direction?

Multivariate normal distribution !

Normal Distribution

$$p_X(x \mid \mu, \sigma) = \frac{1}{\sqrt{2\sigma^2\pi}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}$$



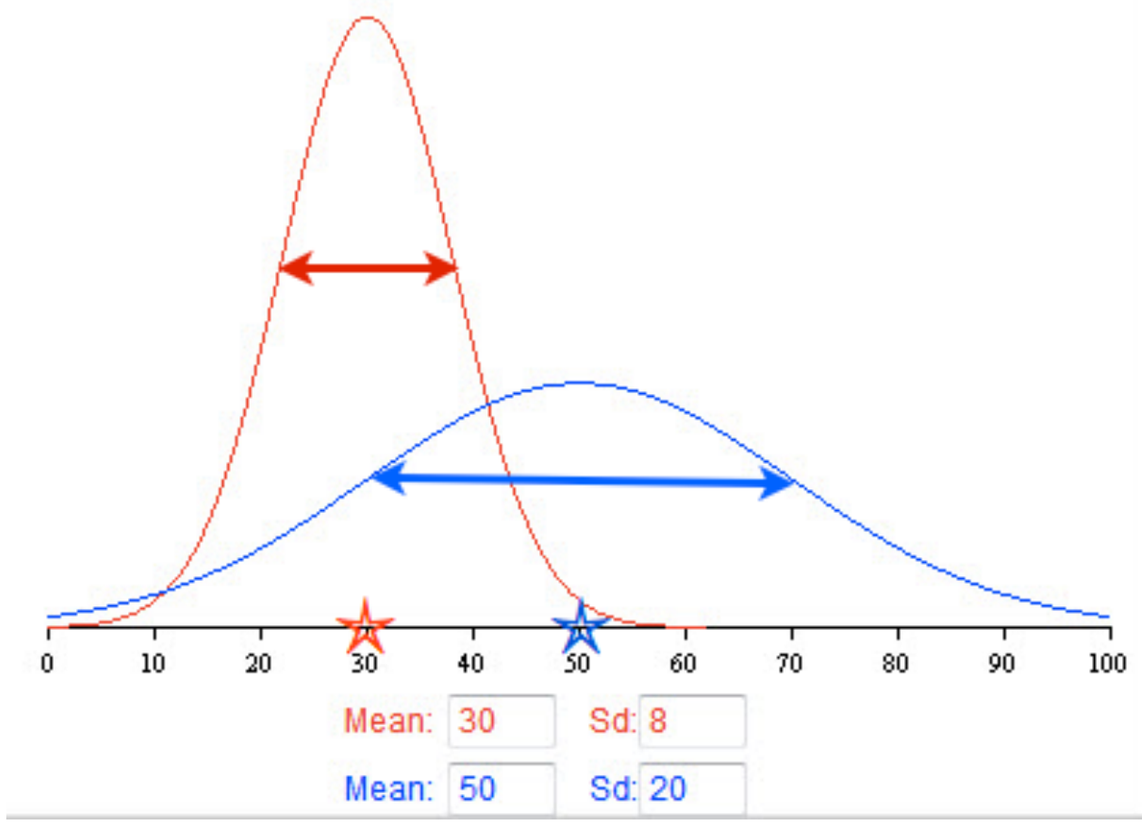


Normal Distribution

- Very common distribution (sometimes also informally known as bell curve)
- PDF specified by mean μ and standard deviation σ (or variance σ^2):

$$p_X(x \mid \mu, \sigma) = \frac{1}{\sqrt{2\sigma^2\pi}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}$$

Often denoted by $N(\mu, \sigma^2)$



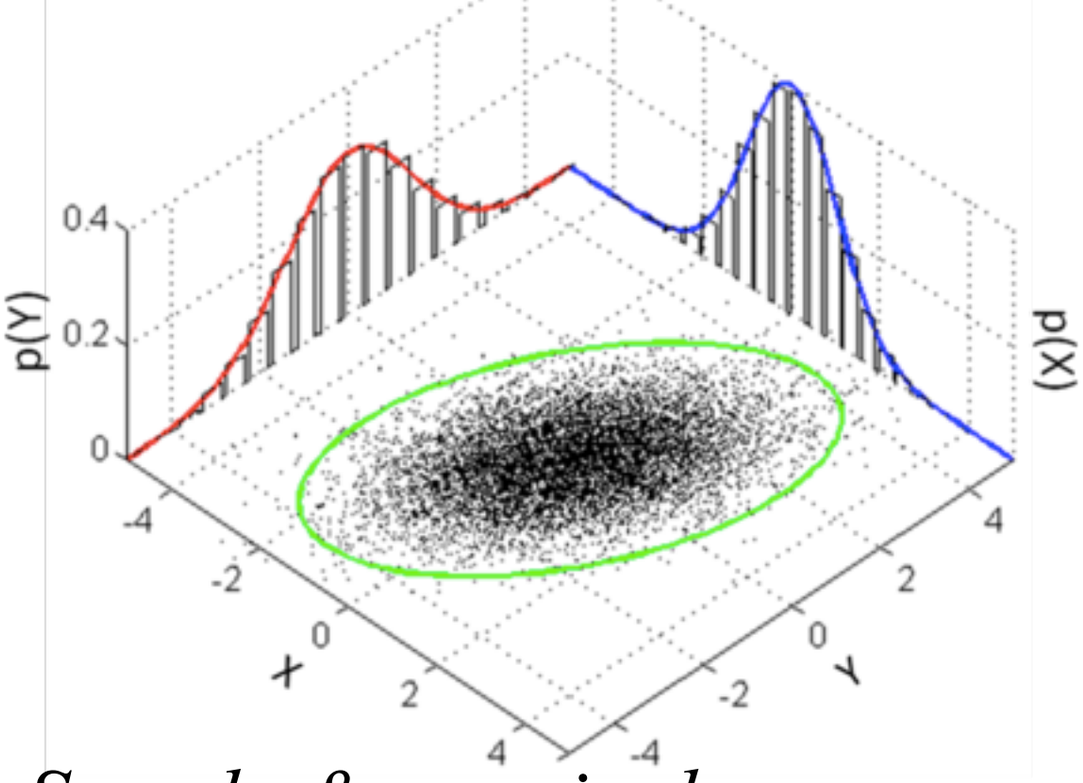
Multivariate Normal Distribution

Covariance matrix

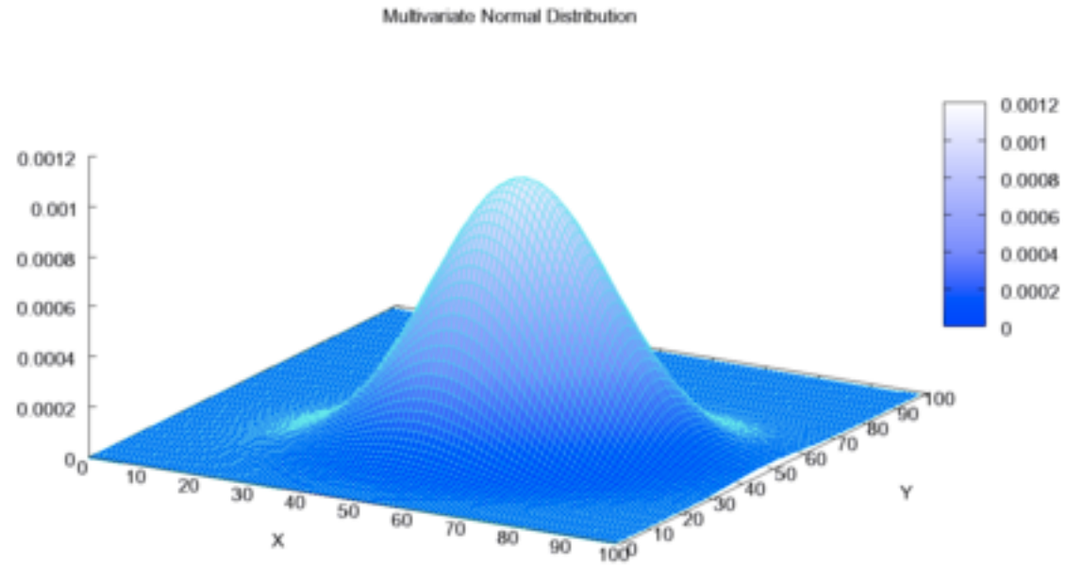
$$\Sigma = \begin{bmatrix} Var(X_1) & Cov(X_1, X_2) \\ Cov(X_2, X_1) & Var(X_2) \end{bmatrix}$$

- PDF for point $\mathbf{x} = (\mathbf{x}_1, \ldots, \mathbf{x}_k)$, specified by mean $\boldsymbol{\mu}$ and covariance matrix :

$$p_{\mathbf{X}}(\mathbf{x}) = \frac{1}{\sqrt{(2\pi)^k |\Sigma|}} \exp\left(-\frac{1}{2}(\mathbf{x} - \boldsymbol{\mu})^{\intercal} \Sigma^{-1} (\mathbf{x} - \boldsymbol{\mu}) \right)$$

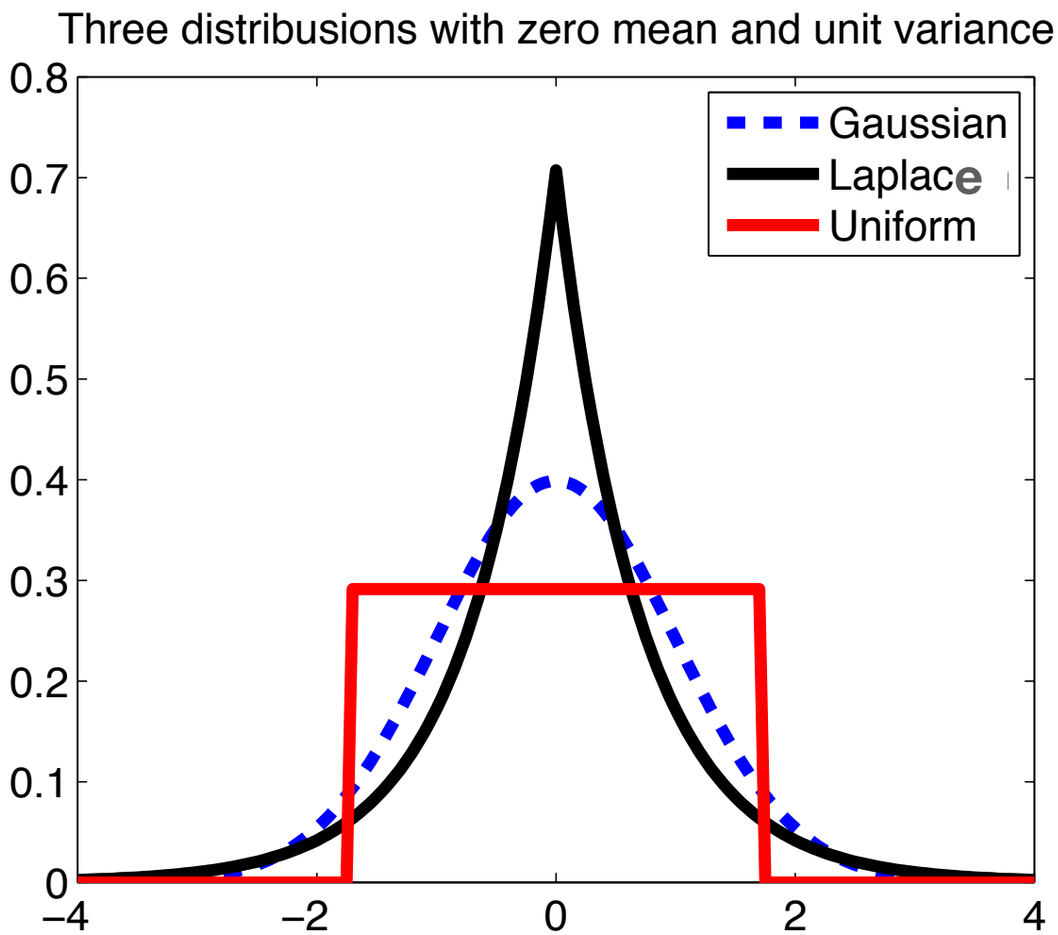


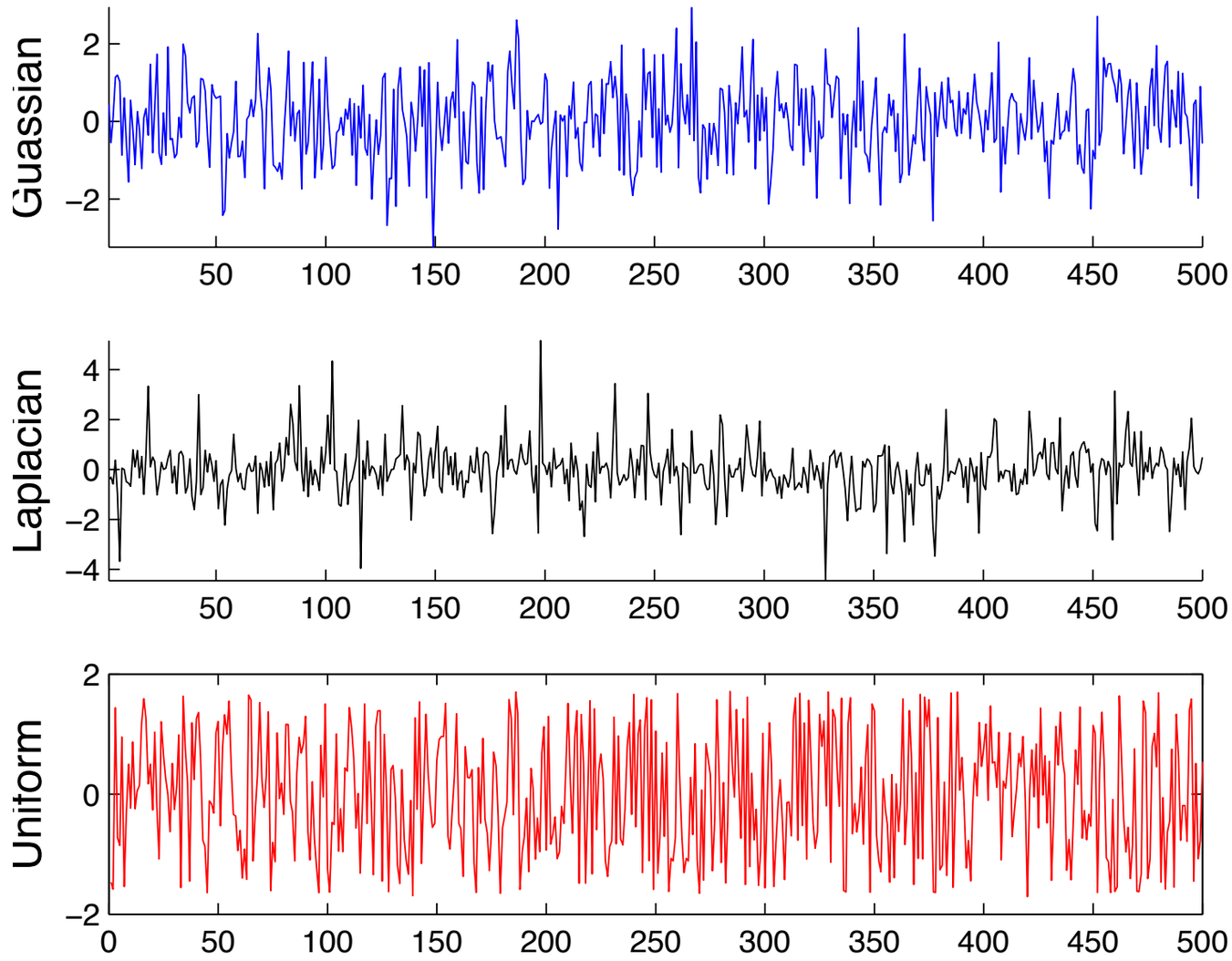
Sample & marginal



pdf

Some Distributions





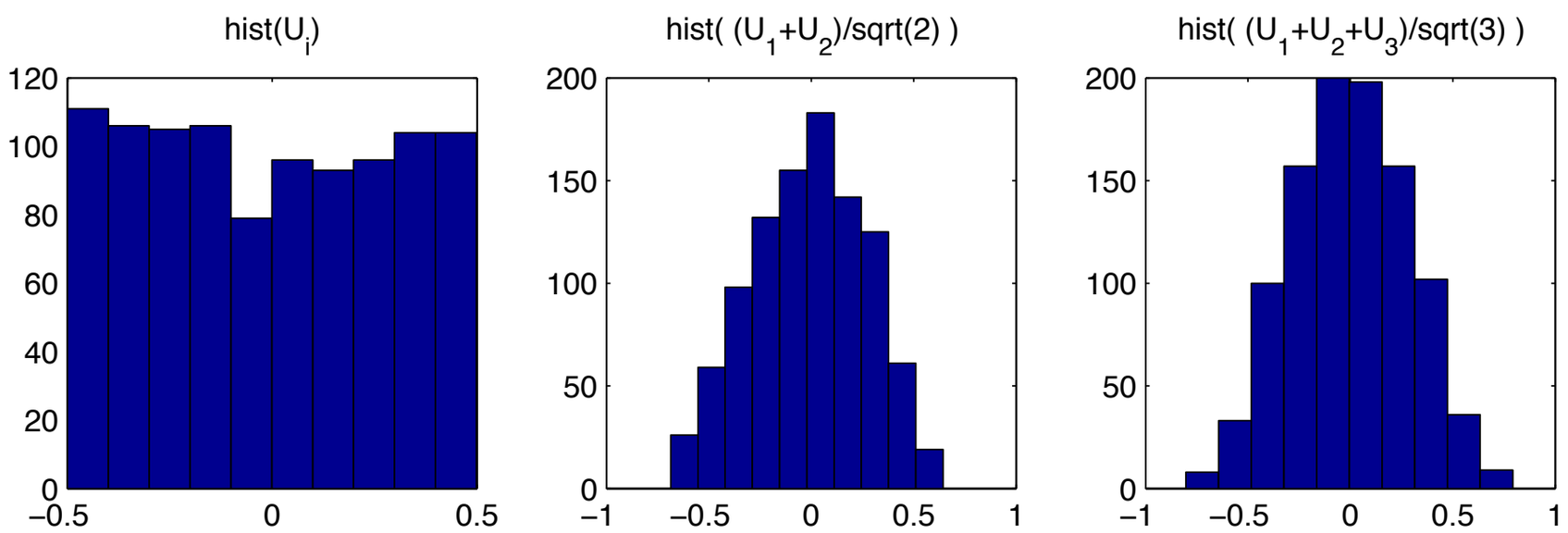
Some Properties of Normal Distributions

- "Simplicity" of the form; completely characterized by mean and covariance; marginal and conditionals are also Gaussian
- Uncorrelatedness implies independence
- Has maximum entropy, given values of the mean and the covariance matrix
- Approximately holds in many cases because of *central limit theorem* (CLT; see demonstrations)

Interested students may refer to Chapter 7 of "Probability theory: The logic of science"

Central Limit Theorem: An Illustration

- CLT: Under some conditions, $S = \frac{1}{n}\sum_{i=1}^{n} X_i$ converges to a normal distribution for independent X_i with finite mean and variance



- Are they really normal? Cramer's decomposition theorem!

E. T. Jaynes. Probability Theory: The Logic of Science. 1994. Chapter 7.

CHAPTER 7

THE CENTRAL GAUSSIAN, OR NORMAL, DISTRIBUTION

> *"My own impression ··· is that the mathematical results have outrun their interpretation and that some simple explanation of the force and meaning of the celebrated integral ··· will one day be found ··· which will at once render useless all the works hitherto written."* - - - Augustus de Morgan (1838)

Here, de Morgan was expressing his bewilderment at the "curiously ubiquitous" success of methods of inference based on the gaussian, or normal, "error law" (sampling distribution), even in cases where the law is not at all plausible as a statement of the actual frequencies of the errors. But the explanation was not forthcoming as quickly as he expected.

In the middle 1950's the writer heard an after–dinner speech by Professor Willy Feller, in which he roundly denounced the practice of using gaussian *probability* distributions for errors, on the grounds that the *frequency* distributions of real errors are almost never gaussian. Yet in spite of Feller's disapproval, we continued to use them, and their ubiquitous success in parameter estimation continued. So 145 years after de Morgan's remark the situation was still unchanged, and the same surprise was expressed by George Barnard (1983): "*Why have we for so long managed with normality assumptions?*"

Today we believe that we can, at last, explain (1) the inevitably ubiquitous use, and (2) the ubiquitous success, of the gaussian error law. Once seen, the explanation is indeed [illegible] yet to the best of our knowledge it is not recognized in any of the previous liter[illegible] because of the universal tendency to think of probability distributions in terms [illegible] cannot understand what is happening until we learn to to think of probabilit[illegible] terms of their demonstrable *information content* instead of their imagined (an[illegible] irrelevant) frequency connections.

Interested students may refer to Chapter 7 of "Probability theory: The logic of science"

A simple explanation of these properties – stripped of past irrelevancies – has been achieved only very recently, and this development changed our plans for the present work. We decided that it is so important that it should be inserted at this somewhat early point in the narrative, even though we must then appeal to some results that are established only later. In the present Chapter, then, we survey the historical basis of gaussian distributions and get a quick preliminary understanding of their functional role in inference. This understanding will then guide us directly – without the usual false starts and blind alleys – to the computational procedures which yield the great majority of the useful applications of probability theory.

Three Ways to Derive Gaussian PDFs

- Found by de Moivre (1733), without realizing its importance

- Independence + isotropy (Herschel 1785)

- Maximum likelihood estimate = arithmetic mean (Gauss, 1809)

- Stability in its form under small perturbation (Landon, 1941)

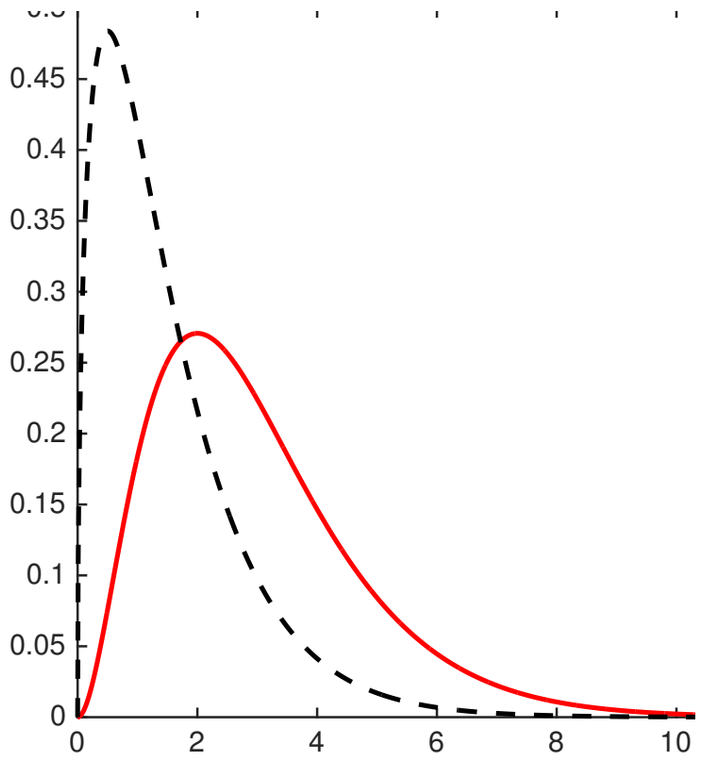
Interested students may refer to Chapter 7 of "Probability theory: The logic of science"

Distance Between Distributions: Are Two Distributions the Same?

- Kullback-Leibler divergence:

$$D_{\mathrm{KL}}(P\|Q) = \sum_i P(i)\,\log\frac{P(i)}{Q(i)}.$$

$$D_{\mathrm{KL}}(p(x)\|q(x)) = \int_{-\infty}^{\infty} p(x)\,\log\frac{p(x)}{q(x)}\,\mathrm{d}x.$$



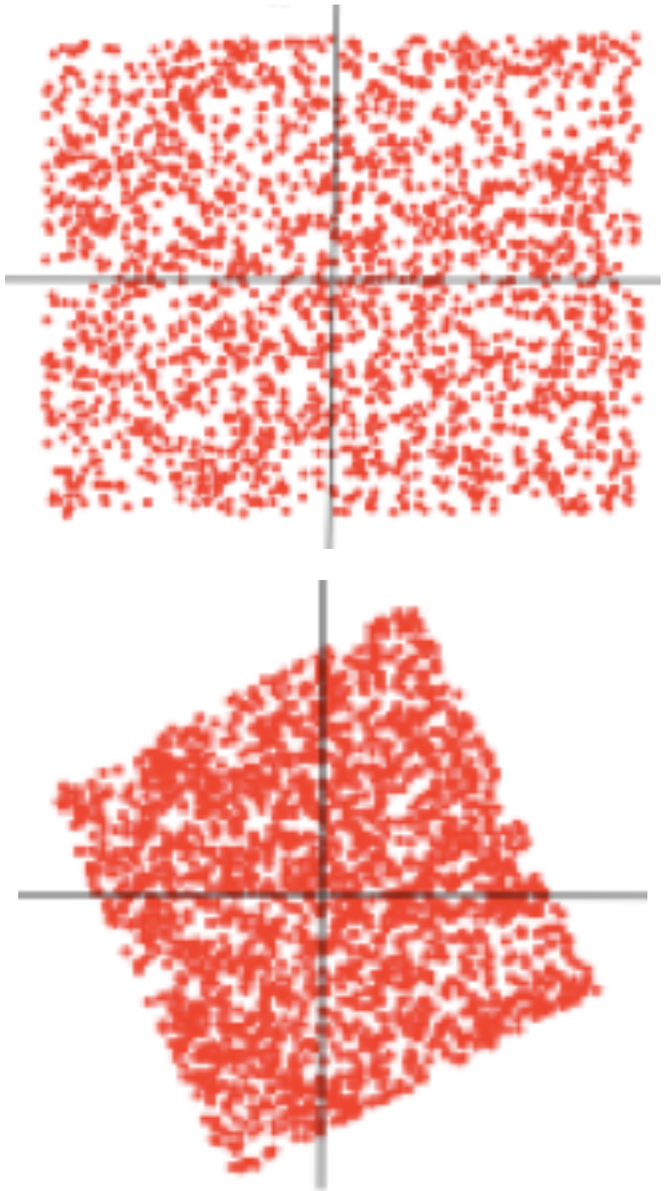
- Non-negative; asymmetric; zero iff identical

Are Two Variables Independent?

- Natural measure of statistical dependence: mutual information

$$I(X;Y) = \sum_y \sum_x P(x,y) \log \left(\frac{P(x,y)}{P(x)\,P(y)} \right),$$

$$I(X;Y) = \int \int p(x,y) \log \left(\frac{p(x,y)}{p(x)\,p(y)} \right) \, dx\, dy,$$



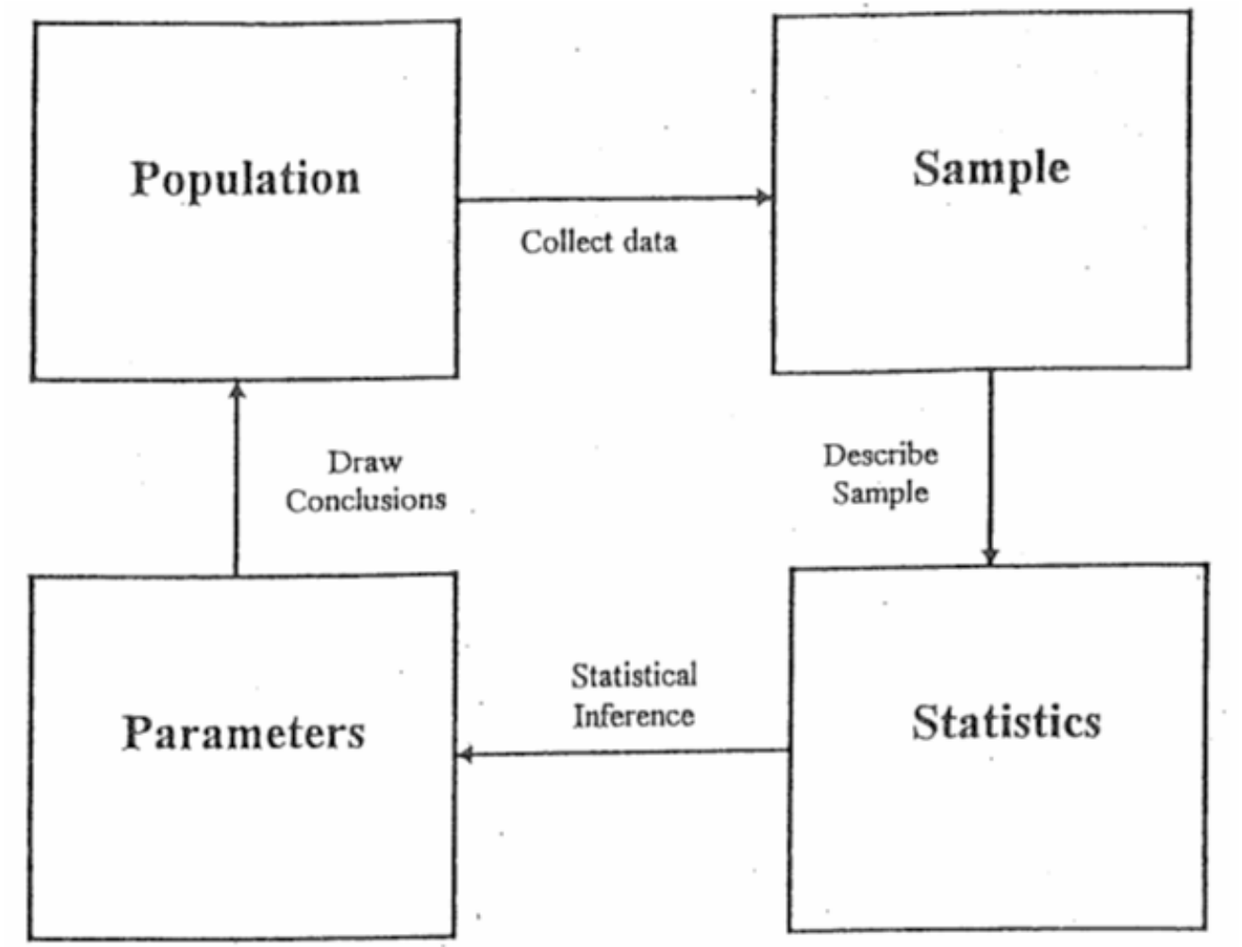
- Non-negative; is zero iff X and Y are independent

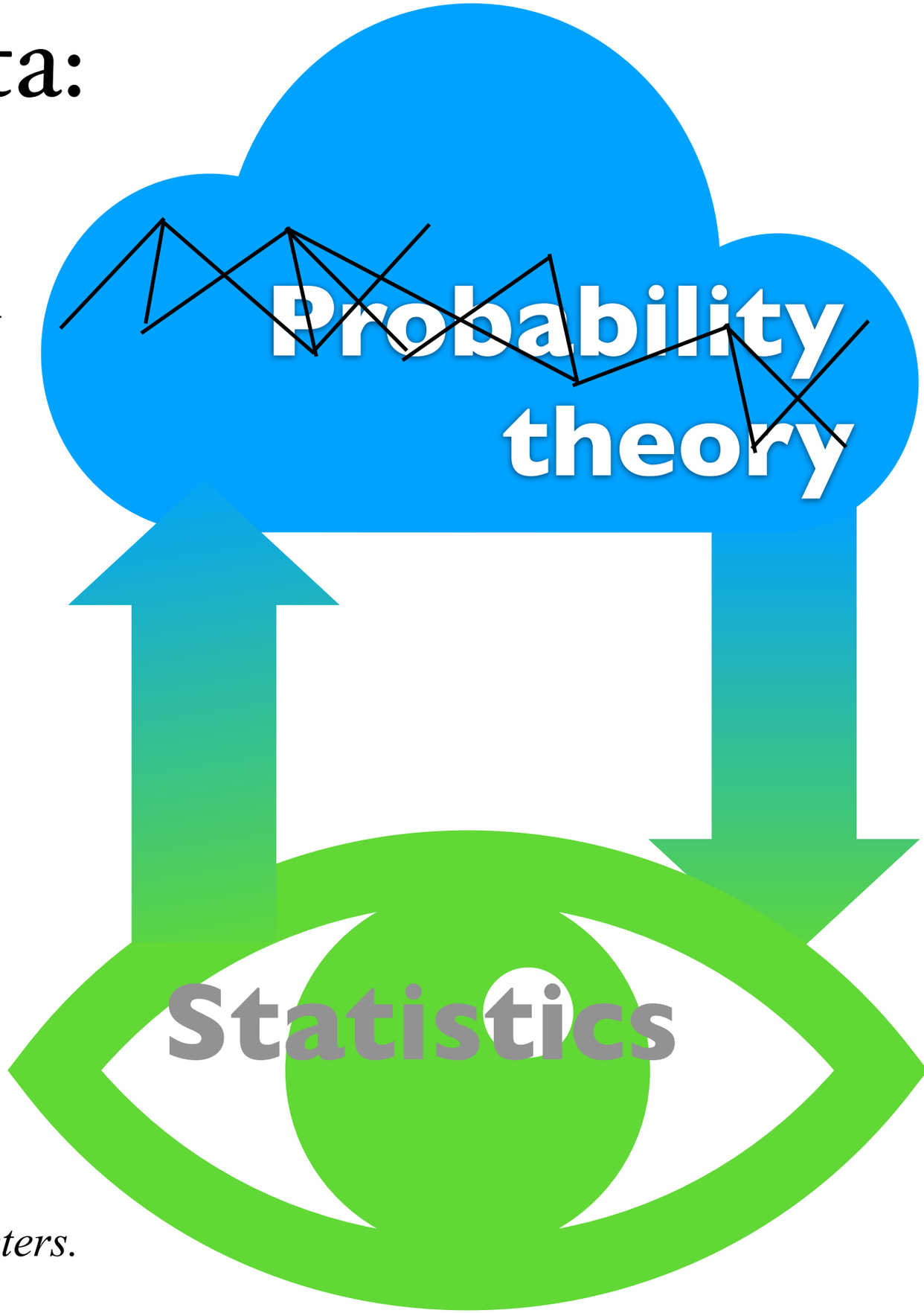
Summary: Probability Theory

- How to understand probability?
- Typical distributions
- **Independence & conditional independence**
- Basic statistics: expectation, variance…
- Independence vs. zero correlation
- Gaussian distribution
- Distance “between” two distributions & measure of dependence

Making Use of Data: Statistics…

- Relationship between probability theory & statistics



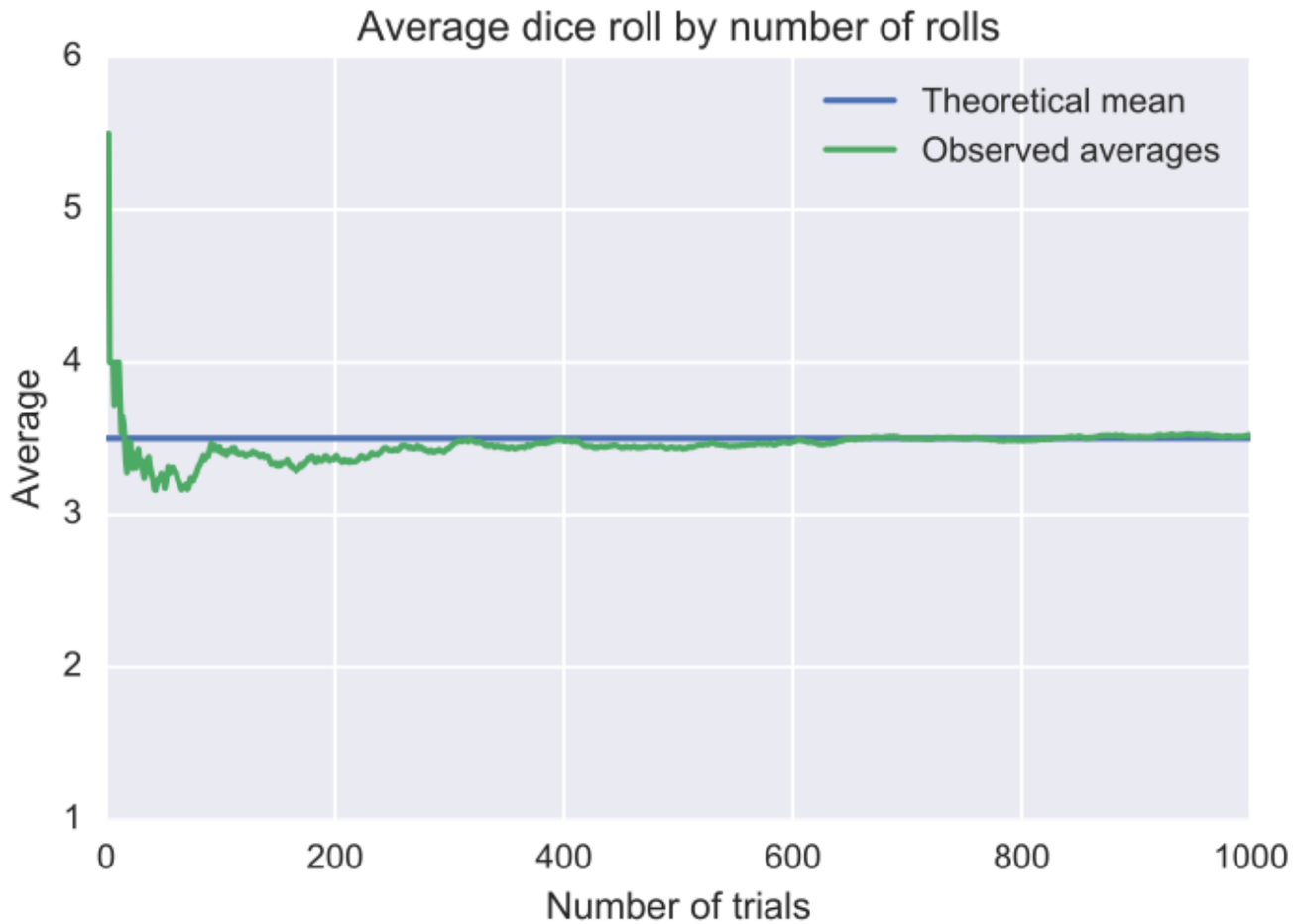


Using sample statistics to estimate population parameters.

Terms

- Population
- Random variable
- Parameter (A parameter is **a number describing a whole population** (e.g., population mean), while a statistic is a number describing a sample)
- Sample
- Statistic
- Likelihood function
- Null hypothesis, null distribution, *p* value

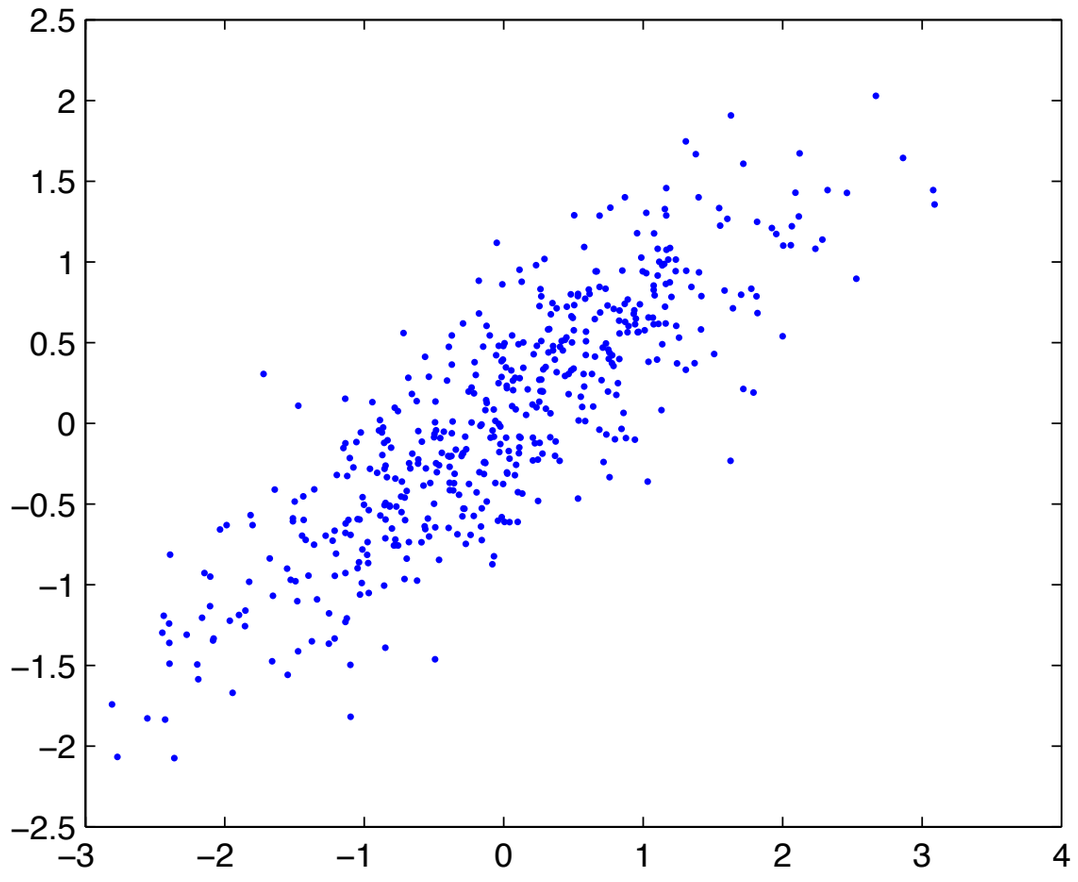
Law of Large Numbers

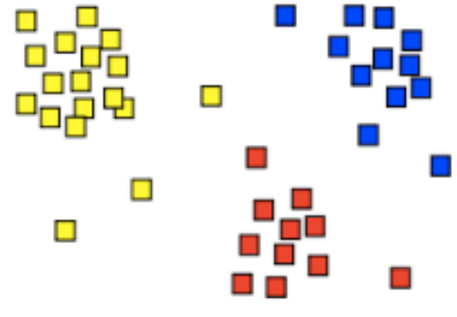


- Law of large numbers (LLN) is a theorem that describes the result of performing the same experiment a large number of times: the average of the results obtained from a large number of trials should be close to the expected value, and will tend to become closer as more trials are performed.

Let's Come Closer to Reality...

- Find knowledge from data, which has randomness. E.g.,
- Bayesian inference
- Parameter estimation and hypothesis test
- Learning
  - Supervised learning
  - Unsupervised learning...
  - Causal discovery





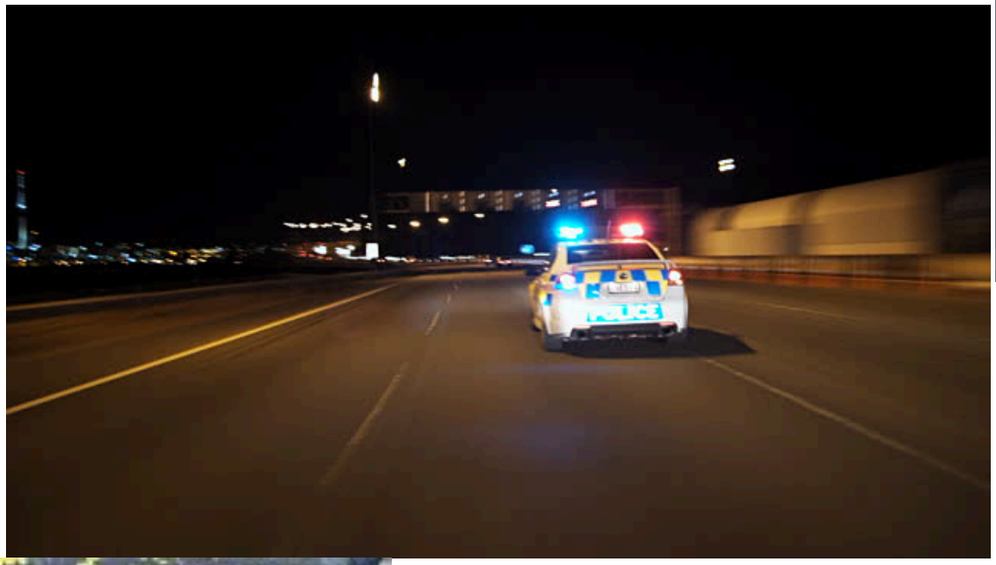
* We'll See More Detail: Bayesian vs. Frequentist Inference

	Frequentist/ Likelihood	Bayesian
Setup	Parameter θ is unknown but fixed	Parameter θ is a random variable
What do we want?	$\hat{\theta}$ – best estimate of the unknown (but fixed) θ based on the given data $\boldsymbol{x}$	$p(\theta\|\boldsymbol{x})$ – posterior distribution of θ that is informed by the given data $\boldsymbol{x}$
What do we need?	Statistical model $\{f_\theta : \theta \in \Omega\}$, data $\boldsymbol{x}$	Statistical model $\{f_\theta : \theta \in \Omega\}$, data $\boldsymbol{x}$, and **prior** $\pi(\theta)$

Related Question: What is Probability?

- Frequentist view: treats "probability" in equivalent terms to "frequency"
- Frank Ramsey: Probability is a rational degree of belief
  - The measure of degrees of belief must satisfy the axioms for probability measures.
  - As new evidence is acquired, the measure of degrees of belief in a system of events must change to their conditional probabilities on the evidence.

How to Update Our Belief?



Bayes' Rule

$$P(A \mid B) = \frac{P(B \mid A)P(A)}{P(B)}$$

- Use evidence (B) to update probabilities (info about A)
  - How to find $P(B)$?
- $P(A_i \mid B) = \frac{P(B|A_i)P(A_i)}{P(B)} = \frac{P(B|A_i)P(A_i)}{\sum_{k=1}^{m} P(B|A_k)P(A_k)}$

Bayes' Rule: Example

$$P(A_i \mid B) = \frac{P(B|A_i)P(A_i)}{P(B)} = \frac{P(B|A_i)P(A_i)}{\sum_{k=1}^{m} P(B|A_k)P(A_k)}$$

- Suppose a drug test is 99% sensitive and 99% specific. That is, the test will produce 99% true positive results for drug users and 99% true negative results for non-drug users.
- Suppose that 0.1% of people are users of the drug.
- If a randomly selected individual tests positive, what is the probability he or she is a user?
- P(User | +) ≈ ? *A. 0.1,* *B. 0.4,* *C. 0.9*

$$\frac{P(User|+)}{P(\sim User|+)} = \frac{P(+|User)P(User)/P(+)}{P(+|\sim User)P(\sim User)/P(+)} = \frac{P(+|User)P(User)}{P(+|\sim User)P(\sim User)}$$
$$= \frac{0.99 \cdot 0.001}{(1-0.99)\cdot(1-0.001)} = \frac{99}{999}$$

* Bayesian Inference: An Example

An Example:

A. You measured my height *4* times, with *n* observations *1.78m, 1.80m, 1.79m, 1.81m. They are assumed to be generated from* $N(\theta, 0.03^2)$. σ^2 x_i

B. The prior distribution of my height is $\theta \sim N(1.75, 0.1^2)$.

C. The posterior distribution of my height θ ? $\alpha=\{\mu_0, \sigma_0^2\}$

* Bayesian Inference: An Example

- x, a data point in general.
- θ, the parameter of the data point's distribution, i.e., $X \sim p_X(x|\theta)$.
- α, the hyperparameter of the parameter distribution, i.e., $\theta \sim p(\theta|\alpha)$
- $\mathbf{X}$ is the sample, a set of n observed data points, i.e., $x_1, \ldots, x_n$.
- $\tilde{x}$, a new data point whose distribution is to be predicted.

An Example:

A. You measured my height *4* times, with *n* observations *1.78m, 1.80m, 1.79m, 1.81m. They are assumed to be generated from* $N(\theta, 0.03^2)$. σ^2 x_i

B. The prior distribution of my height is $\theta \sim N(1.75, 0.1^2)$.

C. The posterior distribution of my height θ ? $\alpha=\{\mu_0, \sigma_0^2\}$

* Bayesian Inference: An Example

- x, a data point in general.
- θ, the parameter of the data point's distribution, i.e., $X \sim p_X(x|\theta)$.
- α, the hyperparameter of the parameter distribution, i.e., $\theta \sim p(\theta|\alpha)$
- $\mathbf{X}$ is the sample, a set of n observed data points, i.e., $x_1, \ldots, x_n$.
- $\tilde{x}$, a new data point whose distribution is to be predicted.

Posterior distribution of the parameter:

$$p(\theta \mid \mathbf{X}, \alpha) = \frac{p(\theta, \mathbf{X}, \alpha)}{p(\mathbf{X}, \alpha)} = \frac{p(\mathbf{X} \mid \theta, \alpha)p(\theta, \alpha)}{p(\mathbf{X} \mid \alpha)p(\alpha)} = \frac{p(\mathbf{X} \mid \theta, \alpha)p(\theta \mid \alpha)}{p(\mathbf{X} \mid \alpha)} \propto p(\mathbf{X} \mid \theta, \alpha)p(\theta \mid \alpha)$$

Bayesian prediction (posterior predictive distribution):

$$p(\tilde{x} \mid \mathbf{X}, \alpha) = \int p(\tilde{x} \mid \theta)p(\theta \mid \mathbf{X}, \alpha)\, \mathrm{d}\theta$$

An Example:

A. You measured my height *4* times, with *n* observations *1.78m, 1.80m, 1.79m, 1.81m. They are assumed to be generated from* $N(\theta, 0.03^2)$. σ^2 x_i

B. The prior distribution of my height is $\theta \sim N(1.75, 0.1^2)$.

C. The posterior distribution of my height θ ? $\alpha=\{\mu_0, \sigma_0^2\}$

* Bayesian Inference: An Example

- x, a data point in general.
- θ, the parameter of the data point's distribution, i.e., $X \sim p_X(x|\theta)$.
- α, the hyperparameter of the params

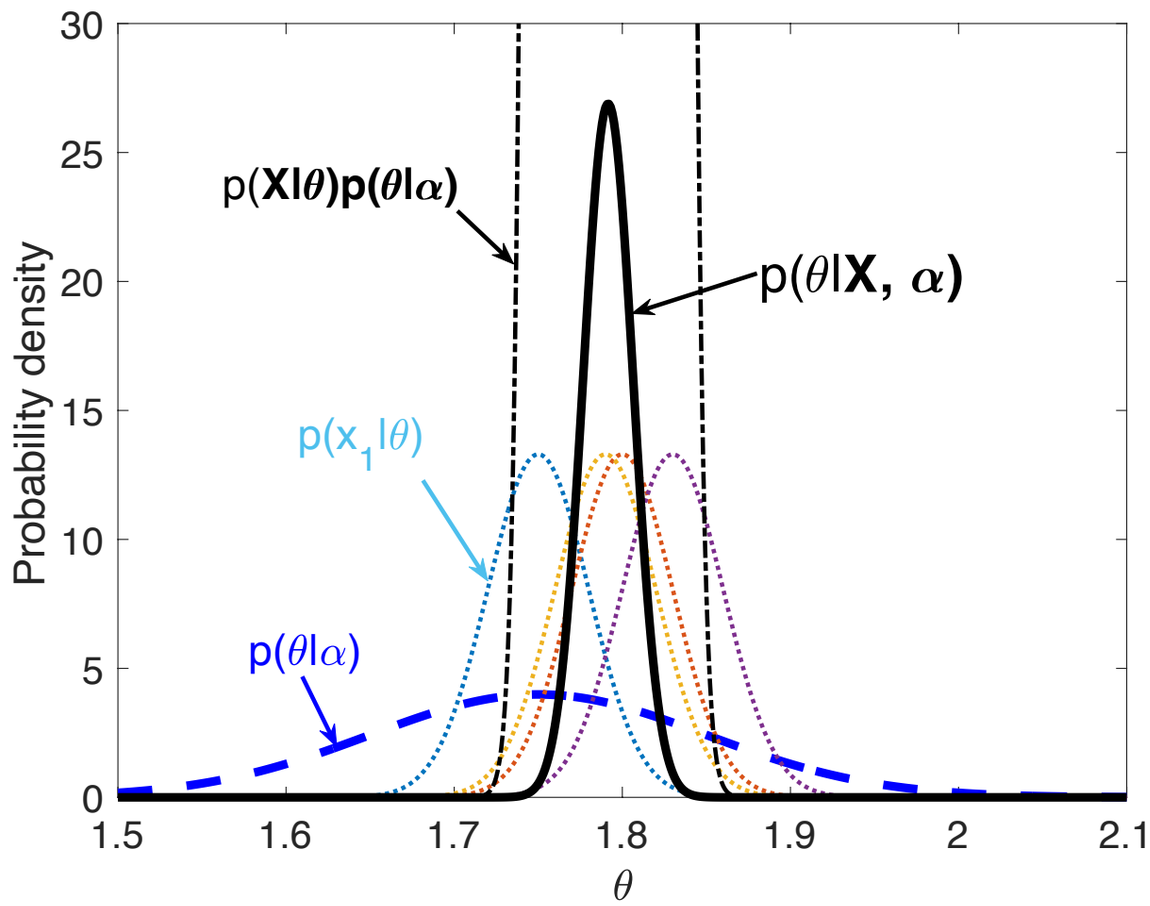
Prior distribution: $\theta \sim N(\mu_0, \sigma_0^2)$;

likelihood: $p(\mathbf{X}|\theta, \sigma^2) = \prod_{i=1}^{n} p(x_i|\theta, \sigma^2)$;

$p(\theta|\mathbf{X}, \alpha) = N(\mu_1, \sigma_2^2)$,

where $\sigma_1^2 = \left(\frac{1}{\sigma_0^2} + \frac{n}{\sigma^2}\right)^{-1}$, and

$\mu_1 = \sigma_1^2\left(\frac{\mu_0}{\sigma_0^2} + \frac{n\bar{x}}{\sigma^2}\right) = \frac{\frac{\mu_0}{\sigma_0^2} + \frac{n\bar{x}}{\sigma^2}}{\frac{1}{\sigma_0^2} + \frac{n}{\sigma^2}}$.

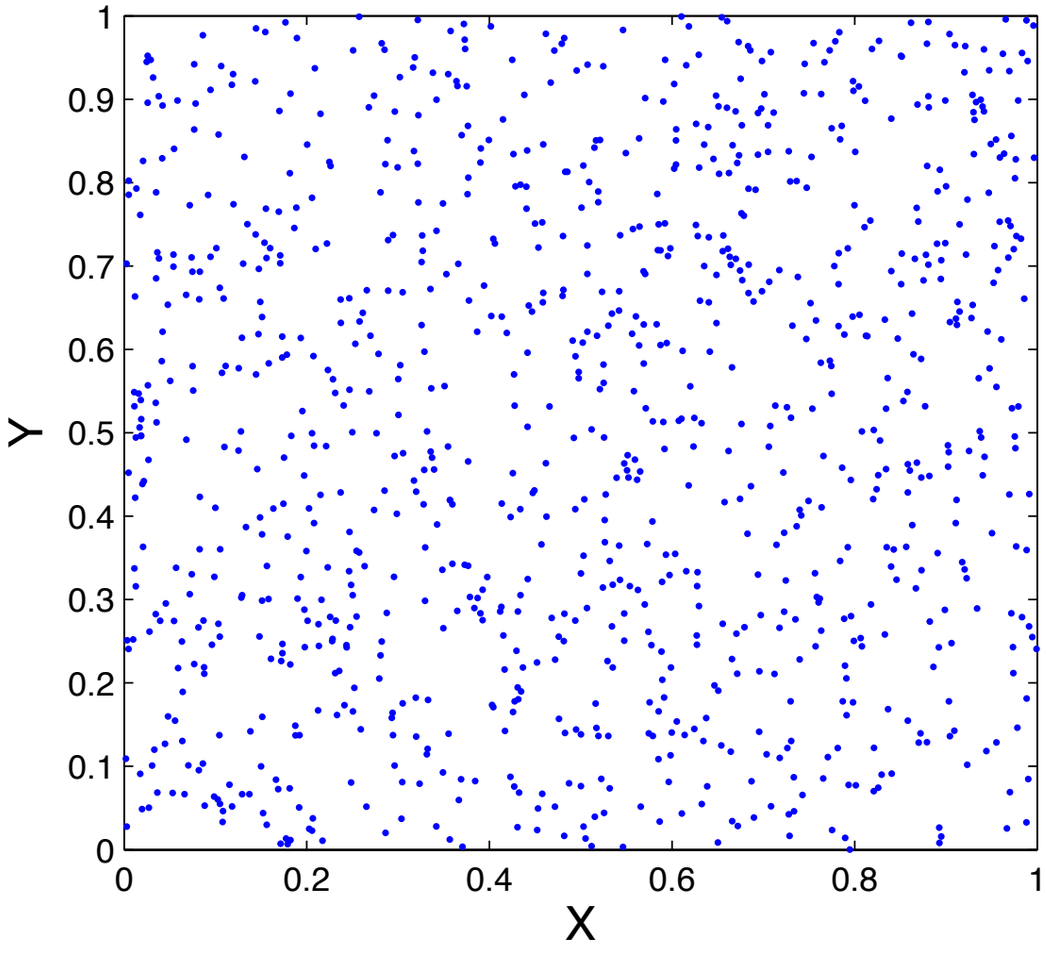


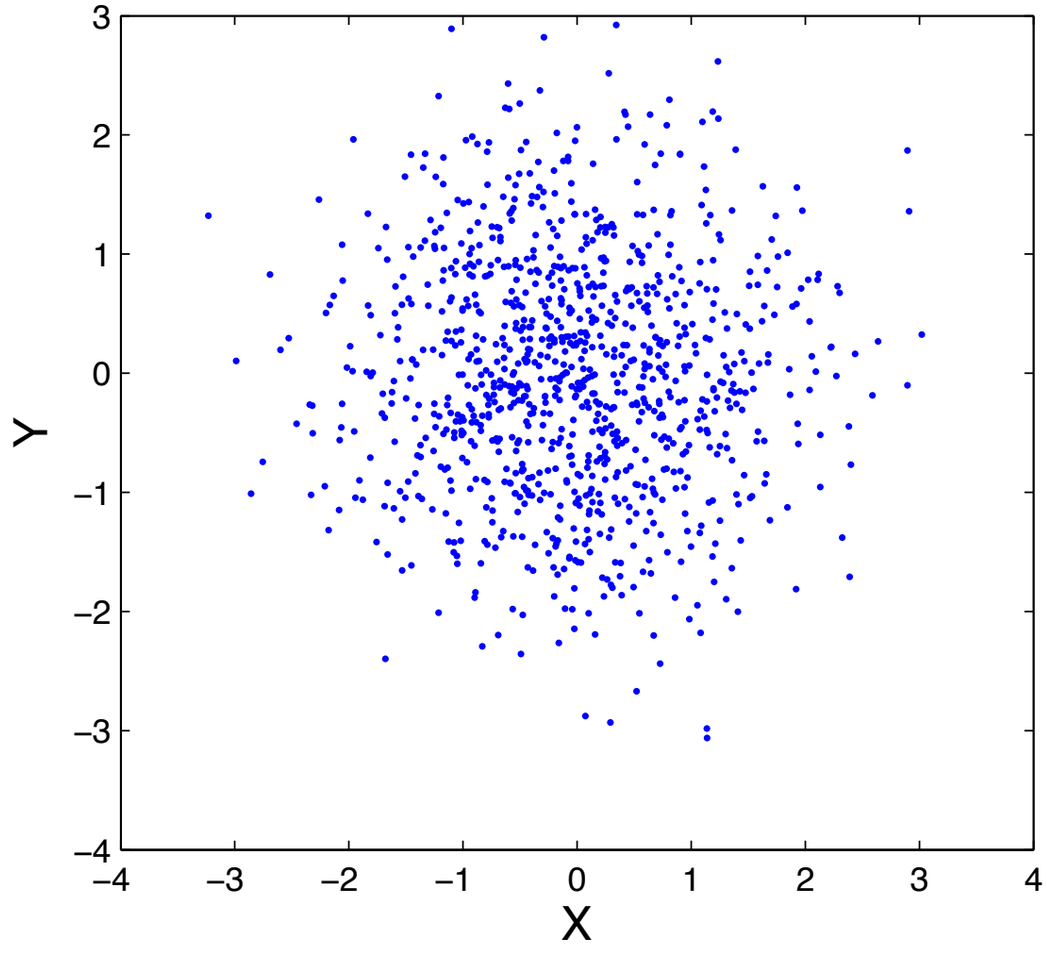
A. You measured my height *n* times, with *n* observations *1.78m, 1.80m, 1.79m, 1.81m. They are assumed to be generated from* $N(\theta, 0.03^2)$. σ^2 x_i

B. The prior distribution of my height is $\theta \sim N(1.75, 0.1^2)$. $\bar{x}$ *is the average of* x_i.

C. The posterior distribution of my height θ ? $\alpha=\{\mu_0, \sigma_0^2\}$

Can You See Whether They Are Independent?





- $p_{XY}(x,y)$ has the same shape for different values of y...
- Further consider two examples with different sample sizes...

A Simple Testing Problem

The one-sample t-test is used to determine whether a sample of observations could have been generated by a process with a specific mean.

- I claim I am *1.80m* tall. You measured my height *n* times, with *n* observations *1.78m, 1.79m,..., 1.81m*
- Null hypothesis H_0: $\mu = 1.80m$; alternative hypothesis H_1: $\mu \neq 1.80m$
- Let's use the one-sample t-test…
  - Calculate the sample mean: $\bar{x} = \frac{x_1 + x_2 + ... + x_n}{n}$
  - Calculate the sample standard deviation: $\hat{\sigma} = \sqrt{\frac{(x_1 - \bar{x})^2 + (x_2 - \bar{x})^2 + ... + (x_n - \bar{x})^2}{n}}$
  - Calculate the **test statistic** $t = \frac{\bar{x} - \mu}{\hat{\sigma}/\sqrt{n}}$
  - Find the *p* value by comparing *t* to a ***t*-distribution** with $(n - 1)$ degrees of freedom

$p = 2 \cdot \mathrm{P}(T > |t|)$ (two-tailed)

  - Draw conclusion by comparing the *p* value with α

Independence Test: Discrete Case

r \ c	Have you taken an online course?		
	Yes	**No**	
Men	43	63	106
Women	95	113	208
	138	176	314

- Set hypotheses:

H_0: Variables are independent.
H_a: Variables are not independent.

- Formulate a plan:
  - Significance level: 0.01, 0.05…
  - Test method: Here we use chi-square test for ⫫
- Analyze the sample
  - Statistic: $Q = \sum_{i=1}^{r}\sum_{j=1}^{c}[(O_{ij} - E_{ij})^2/E_{ij}]$ (E_{ij}: expected freq.)
  - Null dstr of Q; degrees of freedom: $DF = (r\text{-}1)*(c\text{-}1)$
  - p-value: probability of observing a sample statistic as extreme as the test statistic

A Simple Testing Problem

The one-sample t-test is used to determine whether a sample of observations could have been generated by a process with a specific mean.

- I claim I am
observations
- Null hypoth
- Let's use the
  - Calculate
  - Calculate
  - Calculate
  - Find the
degrees
- Draw conclusion by comparing the p value with α

Table of error types		Null hypothesis (H_0) is	
		True	**False**
Decision about null hypothesis (H_0)	**Don't reject**	Correct inference (true negative) (probability = $1-\alpha$)	Type II error (false negative) (probability = β)
	Reject	Type I error (false positive) (probability = α)	Correct inference (true positive) (probability = $1-\beta$) *power*

$$p = 2 \cdot \mathrm{P}(T > |t|)$$
(two-tailed)

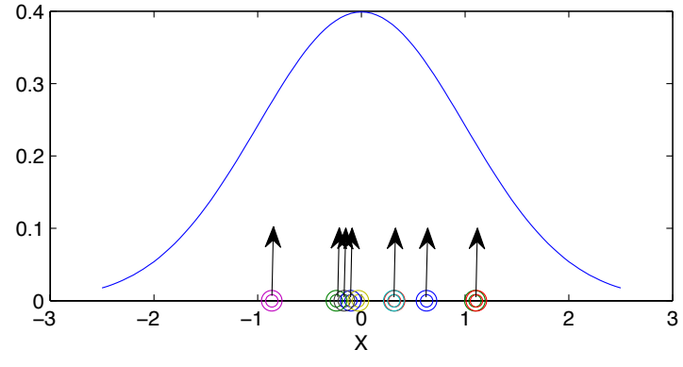
Remember the Example?

What is the best estimate of θ?

An Example:

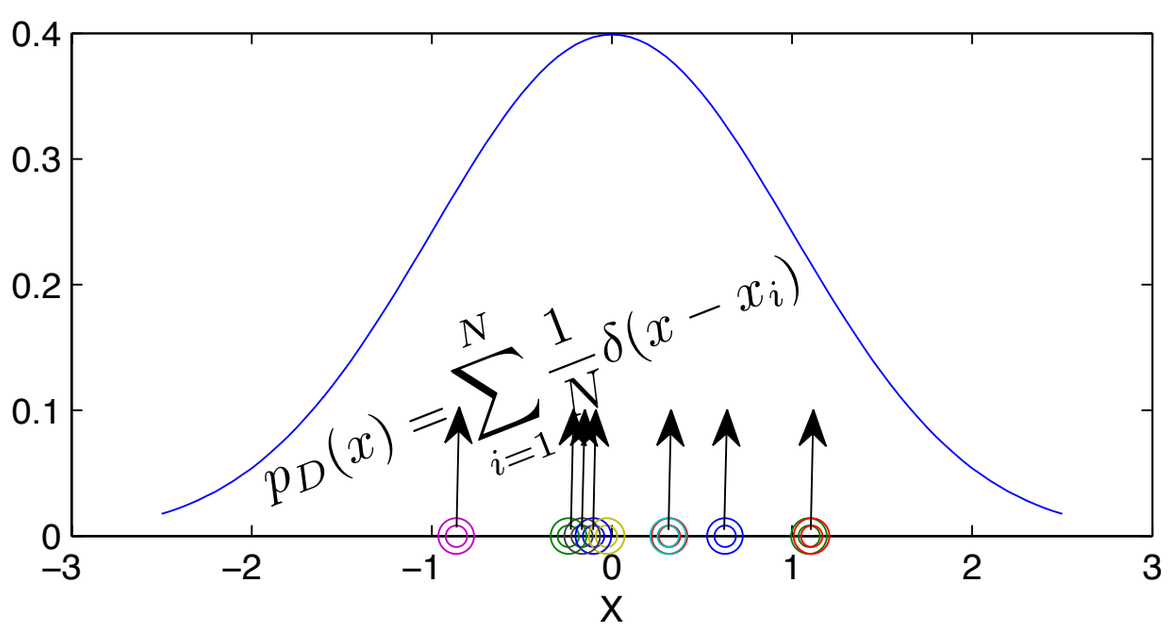
A. You measured my height *4* times, with *n* observations *1.78m, 1.80m, 1.79m, 1.81m. They are assumed to be generated from* $N(\theta, 0.03^2)$.

Maximum Likelihood Estimation



- Estimate characteristics of the model distribution from the sample
  - so that the distribution underlying the sample is close to the model distribution
- Suppose we have functional form of the pdf/pmf $f(x;\theta)$ with unknown parameters $\theta \in \Theta$
- Aim to find a point estimator of θ, i.e., a member of $\{f(x;\theta) \mid \theta \in \Theta\}$ as the most likely pmf/pdf

*How to Find the Best Parameter



- $f(x;\theta)$ should be as close as possible to $p_D(x) = \sum_{i=1}^{N} \frac{1}{N} \delta(x - x_i)$

Definitions [edit]

The Dirac delta can be loosely thought of as a function on the real line which is zero everywhere except at the origin, where it is infinite,

$$\delta(x) = \begin{cases} +\infty, & x = 0 \\ 0, & x \neq 0 \end{cases}$$

and which is also constrained to satisfy the identity

$$\int_{-\infty}^{\infty} \delta(x)\, dx = 1.$$[18]

This is merely a heuristic characterization. The Dirac delta is not a function in the traditional sense as no function defined on the real numbers has these properties.[17] The Dirac delta function can be rigorously defined either as a distribution or as a measure.

Likelihood function

Log likelihood

- (
- (

Equivalent to minimizing the Kullback-Leiber divergence $KL(p_D(x) \,||\, f(x;\Theta))$

Maximum Likelihood Estimation: Example

- Let $X_1, X_2, \ldots, X_n$ be a random sample from $N(\theta_1, \theta_2)$. Find the maximum likelihood estimate of θ_1 and θ_2

$$\hat{\theta}_1 = \frac{1}{n}\sum_{i=1}^{n} x_i; \quad \hat{\theta}_2 = \frac{1}{n}\sum_{i=1}^{n}(x_i - \hat{\theta}_1)^2$$

- Sample mean, sample variance, sample covariance

$$\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i; \qquad s^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2$$

A. You measured my height *n* times, with *n* observations *1.78m, 1.80m, 1.79m, 181cm. They are assumed to be generated from* $N(\theta, 0.03^2)$.

B. What is the point estimate of θ?

$\bar{x}$ *is the average of* x_i.

Identifiability of Parameters in Statistical Models

- Identifiability, in simple words, means that different values of a parameter must produce different probability distributions.
- Mathematically, a parameter θ is said to be identifiable if and only
  $\theta \neq \theta' \Rightarrow P_\theta \neq P_{\theta'}$, or equivalently $P_\theta = P_{\theta'} \Rightarrow \theta = \theta'$
- Is the mean of a Gaussian distribution identifiable?

Unbiased and Consistent Estimator

- If $E[\hat{\theta}] = \theta$, then $\hat{\theta}$ is called an unbiased estimate of θ
- If the estimate $\hat{\theta}_n$ converges in prob. to the true value of the parameter, then it is a consistent estimate

Definition 5.5.1 A sequence of random variables, $X_1, X_2, \ldots$, *converges in probability* to a random variable X if, for every $\epsilon > 0$,

$$\lim_{n\to\infty} P(|X_n - X| \geq \epsilon) = 0 \quad \text{or, equivalently,} \quad \lim_{n\to\infty} P(|X_n - X| < \epsilon) = 1.$$

- Are the MLEs of θ_1 and θ_2 we just derived unbiased? consistent?

Interested students may refer to https://stats.stackexchange.com/questions/31036/what-is-the-difference-between-a-consistent-estimator-and-an-unbiased-estimator .

Let's Check Their Properties...

- Sample mean $\bar{X} = \frac{1}{n}\sum_{i=1}^{n} X_i$. $\mathbb{E}[\bar{X}] = ?$ $\mathbb{V}\text{ar}[\bar{X}] = ?$
- Sample variance $S^2 = \frac{1}{n-1}\sum_{i=1}^{n} (X_i - \bar{X})^2$. $\mathbb{E}[S^2] = ?$

[See page 214 of "Statistical Inference"]

Definition 5.2.2 The *sample mean* is the arithmetic average of the values in a random sample. It is usually denoted by

$$\bar{X} = \frac{X_1 + \cdots + X_n}{n} = \frac{1}{n}\sum_{i=1}^{n} X_i.$$

Definition 5.2.3 The *sample variance* is the statistic defined by

$$S^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \bar{X})^2.$$

The *sample standard deviation* is the statistic defined by $S = \sqrt{S^2}$.

As is commonly done, we have suppressed the functional notation in the above definitions of these statistics. That is, we have written S rather than $S(X_1, \ldots, X_n)$. The dependence of the statistic on the sample is understood. As before, we will denote observed values of statistics with lowercase letters. So $\bar{x}$, s^2, and s denote observed values of $\bar{X}$, S^2, and S.

The sample mean is certainly familiar to all. The sample variance and standard deviation are measures of variability in the sample that are related to the population variance and standard deviation in ways that we shall see below. We begin by deriving some properties of the sample mean and variance. In particular, the relationship for the sample variance given in Theorem 5.2.4 is related to (2.3.1), a similar relationship for the population variance.

Theorem 5.2.4 *Let $x_1, \ldots, x_n$ be any numbers and $\bar{x} = (x_1 + \cdots + x_n)/n$. Then*
a. $\min_a \sum_{i=1}^n (x_i - a)^2 = \sum_{i=1}^n (x_i - \bar{x})^2$,
b. $(n-1)s^2 = \sum_{i=1}^n (x_i - \bar{x})^2 = \sum_{i=1}^n x_i^2 - n\bar{x}^2$.

Proof: To prove part (a), add and subtract $\bar{x}$ to get

$$\begin{aligned}\sum_{i=1}^{n}(x_i - a)^2 &= \sum_{i=1}^{n}(x_i - \bar{x} + \bar{x} - a)^2 \\ &= \sum_{i=1}^{n}(x_i - \bar{x})^2 + 2\sum_{i=1}^{n}(x_i - \bar{x})(\bar{x} - a) + \sum_{i=1}^{n}(\bar{x} - a)^2 \\ &= \sum_{i=1}^{n}(x_i - \bar{x})^2 + \sum_{i=1}^{n}(\bar{x} - a)^2. \qquad \text{(cross term is 0)}\end{aligned}$$

It is now clear that the right-hand side is minimized at $a = \bar{x}$. (Notice the similarity to Example 2.2.6 and Exercise 4.13.)

To prove part (b), take $a = 0$ in the above. □

The expression in Theorem 5.2.4(b) is useful both computationally and theoretically because it allows us to express s^2 in terms of sums that are easy to handle.

We will begin our study of sampling distributions by considering the expected values of some statistics. The following result is quite useful.

a. $\mathrm{E}\bar{X} = \mu$,
b. $\mathrm{Var}\,\bar{X} = \dfrac{\sigma^2}{n}$,
c. $\mathrm{E}S^2 = \sigma^2$.

Proof: To prove (a), let $g(X_i) = X_i/n$, so $\mathrm{E}g(X_i) = \mu/n$. Then, by Lemma 5.2.5,

$$\mathrm{E}\bar{X} = \mathrm{E}\left(\frac{1}{n}\sum_{i=1}^{n} X_i\right) = \frac{1}{n}\mathrm{E}\left(\sum_{i=1}^{n} X_i\right) = \frac{1}{n}n\mathrm{E}X_1 = \mu.$$

Similarly for (b), we have

$$\mathrm{Var}\,\bar{X} = \mathrm{Var}\left(\frac{1}{n}\sum_{i=1}^{n} X_i\right) = \frac{1}{n^2}\mathrm{Var}\left(\sum_{i=1}^{n} X_i\right) = \frac{1}{n^2}n\,\mathrm{Var}\,X_1 = \frac{\sigma^2}{n}.$$

For the sample variance, using Theorem 5.2.4, we have

$$\begin{aligned}\mathrm{E}S^2 &= \mathrm{E}\left(\frac{1}{n-1}\left[\sum_{i=1}^{n} X_i^2 - n\bar{X}^2\right]\right) \\ &= \frac{1}{n-1}\left(n\mathrm{E}X_1^2 - n\mathrm{E}\bar{X}^2\right) \\ &= \frac{1}{n-1}\left(n(\sigma^2 + \mu^2) - n\left(\frac{\sigma^2}{n} + \mu^2\right)\right) = \sigma^2,\end{aligned}$$

establishing part (c) and proving the theorem. □

The relationships (a) and (c) in Theorem 5.2.6, relationships between a statistic and a population parameter, are examples of *unbiased* statistics. These are discussed in Chapter 7. The statistic $\bar{X}$ is an *unbiased estimator* of μ, and S^2 is an *unbiased estimator* of σ^2. The use of $n-1$ in the definition of S^2 may have seemed unintuitive. Now we see that, with this definition, $\mathrm{E}S^2 = \sigma^2$. If S^2 were defined as the usual average of the squared deviations with n rather than $n-1$ in the denominator, then $\mathrm{E}S^2$ would be $\frac{n-1}{n}\sigma^2$ and S^2 would not be an unbiased estimator of σ^2.

We now discuss in more detail the sampling distribution of $\bar{X}$. The methods from Sections 4.3 and 4.6 can be used to derive this sampling distribution from the population distribution. But because of the special probabilistic structure of a random sample (iid random variables), the resulting sampling distribution of $\bar{X}$ is simply expressed.

First we note some simple relationships. Since $\bar{X} = \frac{1}{n}(X_1 + \cdots + X_n)$, if $f(y)$ is the pdf of $Y = (X_1 + \cdots + X_n)$, then $f_{\bar{X}}(x) = nf(nx)$ is the pdf of $\bar{X}$ (see Exercise 5.5). Thus, a result about the pdf of Y is easily transformed into a result about the pdf of $\bar{X}$. A similar relationship holds for mgfs:

$$M_{\bar{X}}(t) = \mathrm{E}e^{t\bar{X}} = \mathrm{E}e^{t(X_1 + \cdots + X_n)/n} = \mathrm{E}e^{(t/n)Y} = M_Y(t/n).$$

Since $X_1, \ldots, X_n$ are identically distributed, $M_{X_i}(t)$ is the same function for each i. Thus, by Theorem 4.6.7, we have the following.

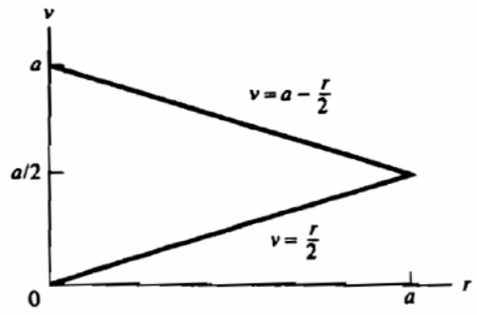


Figure 5.4.1. *Region on which $f_{R,V}(r,v) > 0$ for Example 5.4.7*

5.5 Convergence Concepts

This section treats the somewhat fanciful idea of allowing the sample size to approach infinity and investigates the behavior of certain sample quantities as this happens. Although the notion of an infinite sample size is a theoretical artifact, it can often provide us with some useful approximations for the finite-sample case, since it usually happens that expressions become simplified in the limit.

We are mainly concerned with three types of convergence, and we treat them in varying amounts of detail. (A full treatment of convergence is given in Billingsley 1995 or Resnick 1999, for example.) In particular, we want to look at the behavior of $\bar{X}_n$, the mean of n observations, as $n \to \infty$.

5.5.1 Convergence in Probability

This type of convergence is one of the weaker types and, hence, is usually quite easy to verify.

Definition 5.5.1 A sequence of random variables, $X_1, X_2, \ldots$, *converges in probability* to a random variable X if, for every $\epsilon > 0$,

$$\lim_{n\to\infty} P(|X_n - X| \geq \epsilon) = 0 \quad \text{or, equivalently,} \quad \lim_{n\to\infty} P(|X_n - X| < \epsilon) = 1.$$

The $X_1, X_2, \ldots$ in Definition 5.5.1 (and the other definitions in this section) are typically not independent and identically distributed random variables, as in a random sample. The distribution of X_n changes as the subscript changes, and the convergence concepts discussed in this section describe different ways in which the distribution of X_n converges to some limiting distribution as the subscript becomes large.

Frequently, statisticians are concerned with situations in which the limiting random variable is a constant and the random variables in the sequence are sample means (of some sort). The most famous result of this type is the following.

Theorem 5.5.2 (Weak Law of Large Numbers) *Let $X_1, X_2, \ldots$ be iid random variables with $\mathrm{E}X_i = \mu$ and $\mathrm{Var}\, X_i = \sigma^2 < \infty$. Define $\bar{X}_n = (1/n)\sum_{i=1}^n X_i$. Then,*

Recall that Theorem 10.1.6 stated that, under general conditions, MLEs are consistent. Under somewhat stronger regularity conditions, the same type of theorem holds with respect to asymptotic efficiency so, in general, we can consider MLEs to be consistent and asymptotically efficient. Again, details on the regularity conditions are in Miscellanea 10.6.2.

Theorem 10.1.12 (Asymptotic efficiency of MLEs) *Let $X_1, X_2, \ldots$, be iid $f(x|\theta)$, let $\hat{\theta}$ denote the MLE of θ, and let $\tau(\theta)$ be a continuous function of θ. Under the regularity conditions in Miscellanea 10.6.2 on $f(x|\theta)$ and, hence, $L(\theta|\mathbf{x})$,*

$$\sqrt{n}[\tau(\hat{\theta}) - \tau(\theta)] \to \mathrm{n}[0, v(\theta)],$$

where $v(\theta)$ is the Cramér–Rao Lower Bound. That is, $\tau(\hat{\theta})$ is a consistent and asymptotically efficient estimator of $\tau(\theta)$.

Proof: The proof of this theorem is interesting for its use of Taylor series and its exploiting of the fact that the MLE is defined as the zero of the likelihood function. We will outline the proof showing that $\hat{\theta}$ is asymptotically efficient; the extension to $\tau(\hat{\theta})$ is left to Exercise 10.7.

Recall that $l(\theta|\mathbf{x}) = \sum \log f(x_i|\theta)$ is the log likelihood function. Denote derivatives (with respect to θ) by $l', l'', \ldots$. Now expand the first derivative of the log likelihood around the true value θ_0,

$$l'(\theta|\mathbf{x}) = l'(\theta_0|\mathbf{x}) + (\theta - \theta_0)l''(\theta_0|\mathbf{x}) + \cdots, \tag{10.1.4}$$

where we are going to ignore the higher-order terms (a justifiable maneuver under the regularity conditions).

Now substitute the MLE $\hat{\theta}$ for θ, and realize that the left-hand side of (10.1.4) is 0. Rearranging and multiplying through by $\sqrt{n}$ gives us

$$\sqrt{n}(\hat{\theta} - \theta_0) = \sqrt{n}\frac{-l'(\theta_0|\mathbf{x})}{l''(\theta_0|\mathbf{x})} = \frac{-\frac{1}{\sqrt{n}}l'(\theta_0|\mathbf{x})}{\frac{1}{n}l''(\theta_0|\mathbf{x})}. \tag{10.1.5}$$

If we let $I(\theta_0) = \mathrm{E}[l'(\theta_0|X)]^2 = 1/v(\theta)$ denote the information number for one observation, application of the Central Limit Theorem and the Weak Law of Large Numbers will show (see Exercise 10.8 for details)

$$\begin{aligned} -\frac{1}{\sqrt{n}}l'(\theta_0|\mathbf{X}) &\to \mathrm{n}[0, I(\theta_0)], && \text{(in distribution)} \\ \frac{1}{n}l''(\theta_0|\mathbf{X}) &\to I(\theta_0). && \text{(in probability)} \end{aligned} \tag{10.1.6}$$

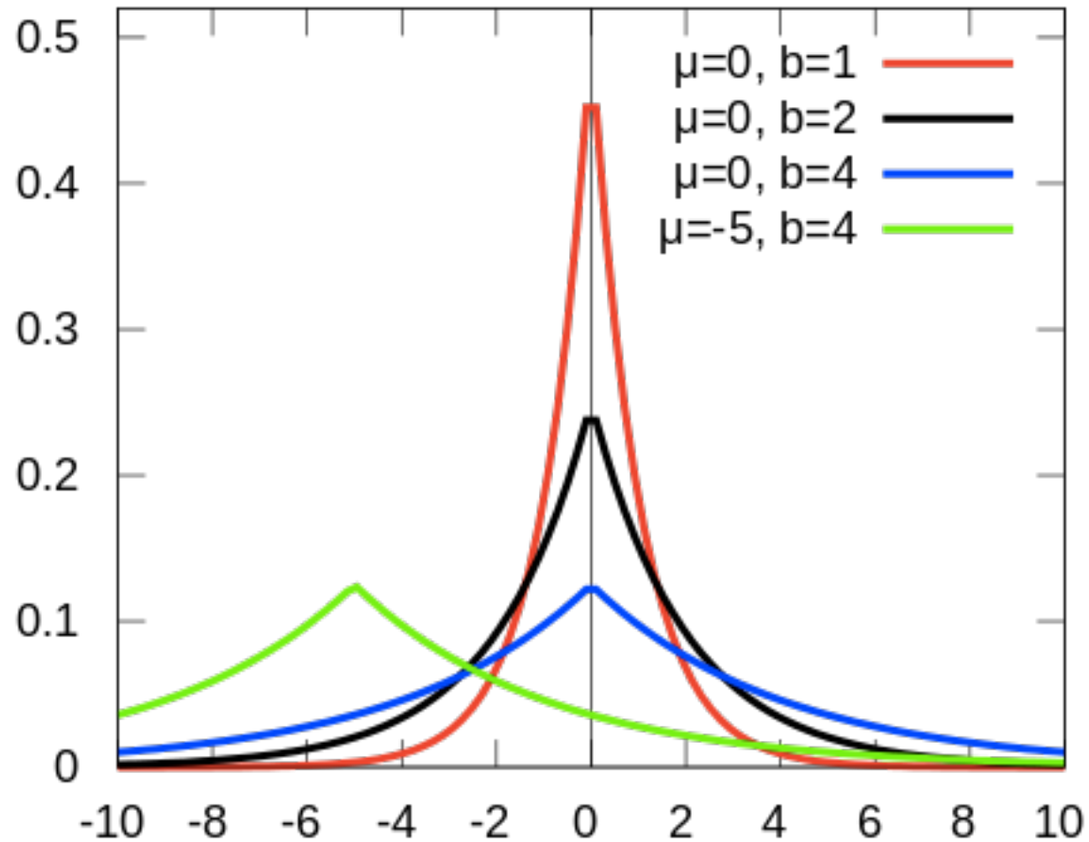
Thus, if we let $W \sim \mathrm{n}[0, I(\theta_0)]$, then $\sqrt{n}(\hat{\theta}-\theta_0)$ converges in distribution to $W/I(\theta_0) \sim \mathrm{n}[0, 1/I(\theta_0)]$, proving the theorem. □

Example 10.1.13 (Asymptotic normality and consistency) The above theorem shows that it is typically the case that MLEs are efficient and consistent. We

* MLE: What If We Use a Different Distribution?

- Let $X_1, X_2, \ldots, X_n$ be a random sample from *the following Laplace distribution*. Find the maximum likelihood estimate of μ

$$p_X(x\,;\,\mu, b) = \frac{1}{2b}\exp\left(-\frac{|x-\mu|}{b}\right)$$



- Median: separating the higher half from the lower

$$P(X \leq m) \geq 1/2, \text{ and } P(X \geq m) \geq 1/2$$

Mean vs. median

Summary: From Probability Theory to Statistics

- From probability theory to statistics
- Ways of making use of data
  - Bayesian inference, parameter estimation, and hypothesis test
  - Intuition behind maximum likelihood estimation
- From linear regression to ML