PHIL6334/ECON6614 - Lecture Notes 1:
From Probability theory to Statistical Inference

Aris Spanos [Spring 2019]

1 Statistical modeling: an brief introduction

Empirical modeling, broadly speaking, refers to the process, methods and strategies grounded on statistical modeling and inference whose primary aim is to give rise to 'learning from data' about stochastic observable phenomena, using *statistical models.* Real world phenomena of interest are said to be 'stochastic', and thus amenable to statistical modeling, when the data they give rise to exhibit *chance regularity patterns*, irrespective of whether they arise from passive observation or active experimentation. In this sense, empirical modeling has three crucial features:
(a) it is based on observed data that exhibit chance regularities,
(b) its cornerstone is the concept of a *statistical model* that describes a probabilistic generating mechanism that could have given rise to the data in question,
(c) it provides the framework for combining the statistical and substantive information with a view to elucidate (understand, predict, explain) stochastic phenomena of interest.

In practice, statistical modeling begins with (i) a (substantive) question of interest pertaining to an observable stochastic phenomenon of interest, and (ii) the selection of the appropriate data $\mathbf{x}_0{:=}(x_1, x_2, \ldots, x_n)$ (a non-trivial problem) to answer that question. Both of these components are used to specify the premises of inductive inference in the form of a statistical model $\mathcal{M}_{\boldsymbol{\theta}}(\mathbf{x})$; the notation will be explained below.

Probability theory provides the framework for:

(a) Modeling: the mathematical concepts needed to account for (model) *chance regularity patterns* in the form of a **statistical model** $\mathcal{M}_{\boldsymbol{\theta}}(\mathbf{x})$: a set of probabilistic assumptions specifying a generating mechanism that could have given rise to the data $\mathbf{x}_0{:=}(x_1, x_2, \ldots, x_n)$, and

(b) Inductive Inference: the framework for model-based ($\mathcal{M}_{\boldsymbol{\theta}}(\mathbf{x})$) statistical inference - estimation (point and interval), testing, prediction and policy analysis.

2 Stochastic phenomena and chance regularities

Stochastic observable phenomena which exhibit *chance* (non-deterministic) regularities. Such phenomena vary from simple games of chance (tossing coins, casting dice, playing the roulette, etc.), to highly complicated experiments in physics and chemistry, as well as observable phenomena in economics and other social sciences, astronomy, geology, biology, epidemiology, etc.

Example 1. Tossing a coin and noting the outcome: Heads (H) or Tails (T).

Example 2. Sampling *with replacement* from an urn which contains red (R) and black (B) balls.

Example 3. Observing the gender (B or G) of newborns during a certain period (a month) in NY city.

Example 4. Casting two dice and adding the dots of the two sides facing up.

Example 5. Tossing a coin twice and noting the outcome.

Example 6. Tossing a coin until the first "Heads" occurs.

Example 7. Counting the number of emergency calls to a regional hospital during a certain period (a week).

Example 8. Sampling *without replacement* from an urn which contains red (R) and black (B) balls.

Example 9. Observing the daily changes of the Dow Jones (D-J) index during a certain period (a year).

Empirical example. To get some idea about 'chance regularity' patterns, consider the data given in table 1.

Table 1 - Observed data

3	10	11	5	6	7	10	8	5	11	2	9	9	6	8	4	7	6	5	12
7	8	5	4	6	11	7	10	5	8	7	5	9	8	10	2	7	3	8	10
11	8	9	5	7	3	4	9	10	4	7	4	6	9	7	6	12	8	11	9
10	3	6	9	7	5	8	6	2	9	6	4	7	8	10	5	8	7	9	6
5	7	7	6	12	9	10	4	8	6	5	4	7	8	6	7	11	7	8	3

A glance at this table suggests that the observed data constitute integers between 2 to 12, but no real patterns are apparent, at least at first sight. To bring out any chance regularity patterns we use a graph shown in fig. 1.1: **t-plot**: $\{(t, x_t),\ t{=}1, 2, ..., n\}$

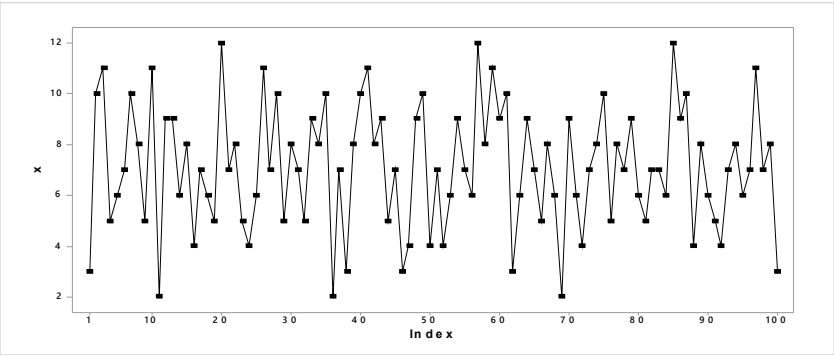


Fig. 1: t-plot of a sequence of 100 observations

The first distinction to be drawn is that between chance regularity patterns and deterministic regularities that are easy to detect.

Deterministic regularity. When a t-plot exhibits a clear pattern which would enable one to predict (guess) the value of the next observation *exactly,* the data are

said to exhibit *deterministic* regularity. The easiest way to think about deterministic regularity is to visualize the graphs of mathematical functions; see figure 2.

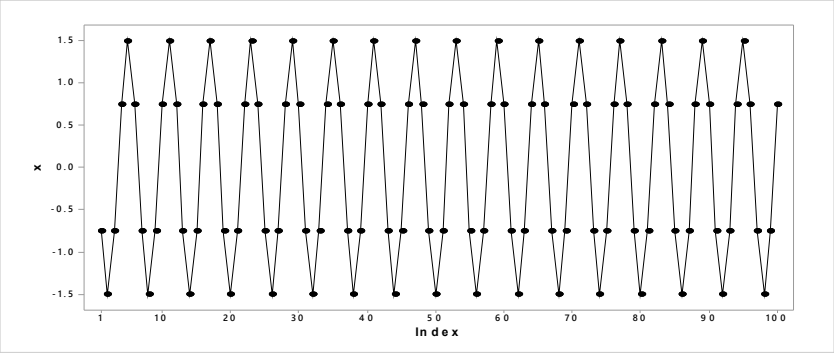


Fig. 2: The graph of $x{=}1.5\cos((\pi/3)t{+}(\pi/3))$

In contrast to deterministic regularities, to detect chance patterns one needs to perform a number of thought experiments.

Thought experiment 1. Associated each observation with identical squares and rotate the figure 1 anti-clockwise by 90^o letting the squares fall vertically to form a pile on the x-axis. The pile represents the well known histogram (see figure 3).

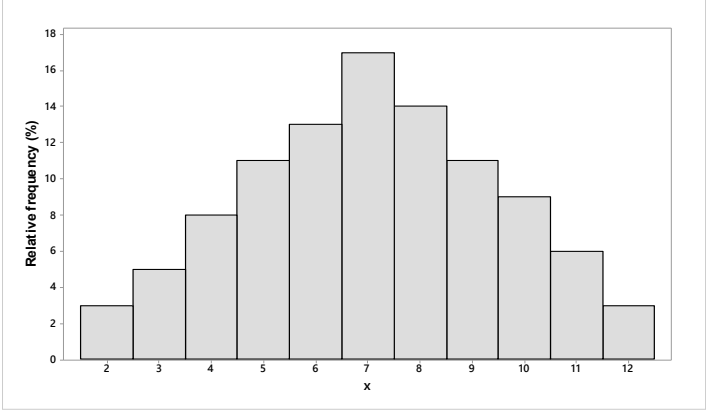
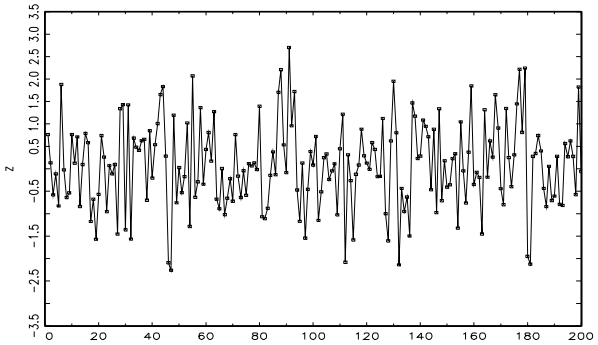


Fig. 3: Histogram of the data in fig. 1

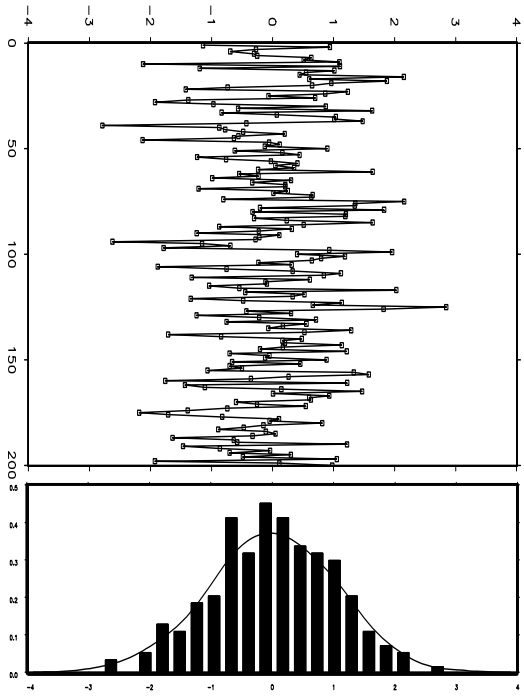
The histogram exhibits a clear triangular shape reflecting a form of regularity often associated with *stable (unchanging) relative frequencies (RF)* expressed as percentages (%). Each bar of the histogram represents the frequency of each of the integers 2-12. For example, since the value 3 occurs 5 times in this data set its relative frequency is: RF(3)=$\frac{5}{100}$=.05. The relative frequency of the value 7 is: RF(7)=$\frac{17}{100}$ = .17, which is the highest among the values 2-12. For reasons that will become apparent shortly we name this discernible regularity:

[1] **Distribution:** after a large enough number of trials, the relative frequency of the outcomes forms a seemingly stable distribution shape.

t-plots of data-examples.



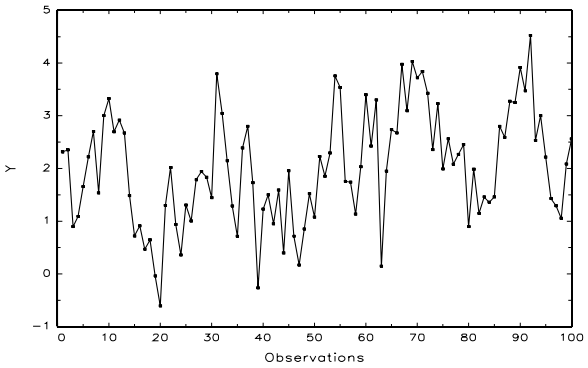
A typical realization of a NIID process



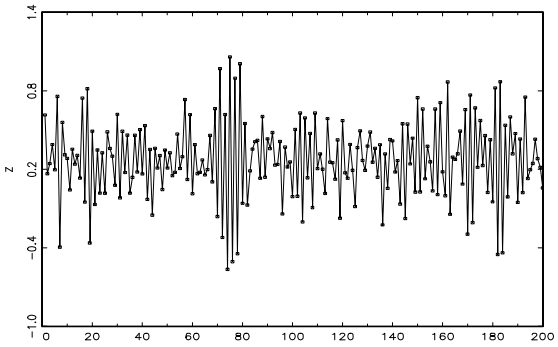
Thought experiment 2. In figure 1 one would hide the observations beyond a certain value of the index, say t=40, and try to guess the next outcome on the basis of the observations up to t=40. Repeat this along the x-axis for different index values and if it turns out that it is more or less impossible to use the previous observations to narrow down the potential outcomes, one would conclude that there is *no dependence* pattern that would enable the modeler to guess the next observation (within narrow bounds) with any certainty. In this experiment one needs to exclude the extreme values of 2 and 12 because following these values one is almost certain

to get a value greater and smaller, respectively. This type of predictability is related to the *distribution regularity* mentioned above. For reference purposes we name the chance regularity associated with the unpredictability of the next observation given the previous observations:

[2] Independence: in a sequence of trials the outcome of any one trial does not influence and is not influenced by the outcome of any other.



A typical realization of a *positively* dependent process



A typical realization of a *negatively* dependent process

Thought experiment 3. In figure 1 take a wide enough frame (to cover the spread of the fluctuations) that is also long enough (roughly less than half the length of the horizontal axis) and let it slide from left to right along the horizontal axis looking at the picture inside the frame as it slides along. In cases where the picture does not change significantly, the data exhibit the chance regularity we call *homogeneity*, otherwise *heterogeneity* is present. Another way to view this pattern is in terms of the arithmetic average and the *variation* around this average of the observations as we move from left to right. It appears as though this *sequential average* and its *variation* are relatively constant around 7. Moreover, the *variation* around this constant average value appears to be within fixed bands. This chance regularity can be intuitively described by the notion of:

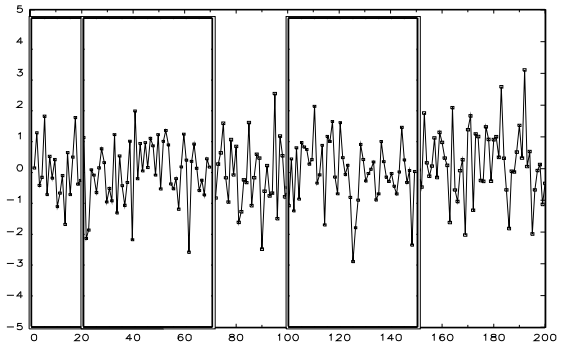
[3] Homogeneity: the probabilities associated with all possible outcomes remain the same for all trials.

In summary, the data in figure 1 exhibit the following chance regularity patterns:

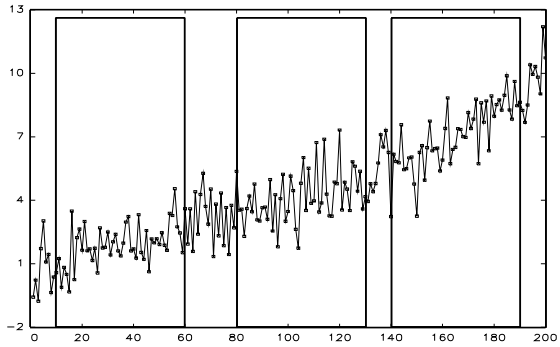
[1] A triangular distribution, [2] Independence, [3] Homogeneity (ID)

It is important to emphasize that these patterns have been discerned directly from the observed data without the use of any *substantive* subject matter information. Indeed, at this stage it is still unknown what these observations represent or measure, but that does not prevent one from discerning certain chance regularity patterns. The information conveyed by these patterns provides the raw material for constructing

statistical models aiming to adequately account for (or model) this (statistical) information. The way this is achieved is to develop probabilistic concepts which aim to formalize these patterns in a mathematical way and provide canonical elements for constructing statistical models.



A typical realization of a NIID process



A typical realization of a NI but non-ID process

3 From chance regularities to a statistical model

The formalization begins by representing the data as a set of n ordered numbers denoted generically by $\mathbf{x}_0{:=}(x_1, x_2, ..., x_n)$. These numbers are in turn interpreted as a *typical realization* of a finite initial segment $\mathbf{X}{:=}(X_1, X_2, ..., X_n)$ of a (possibly infinite) sequence of random variables $\{X_t,\ t{=}1, 2, ..., n, ...\}$, we call a *sample* $\mathbf{X}$; note that the random variables are denoted by capital letters and observations by small letters. The chance regularity patterns exhibited by the data are viewed as reflecting the probabilistic structure of $\{X_t,\ t{=}1, 2, ..., n, ...\}$. For the data in figure 1.1 the structure one can realistically ascribe to sample $\mathbf{X}$ is that they are Independent and Identically Distributed (IID) random variables, with a triangular (Δ) distribution. These probabilistic concepts give rise to a statistical model that takes the following simple form.

Table 2: Simple statistical model

[D] Distribution:	$X_t \backsim \Delta(\mu, \sigma^2),\ x_t{\in}\mathbb{N}_X{:=}(2, \ldots, 12)$, discrete triangular,
[M] Dependence:	$(X_1, X_2, ..., X_n)$ are Independent (I),
[H] Heterogeneity:	$(X_1, X_2, ..., X_n)$ are Identically Distributed (ID).

Note that $\mu{=}E(X_t)$ and $\sigma^2{=}E(X_t{-}\mu)^2$ denote the mean and variance of X_t, respectively.

It is worth emphasizing again that the choice of this statistical model, which aims to account for the regularities in figure 1, relied exclusively on the chance regularities, without invoking any substantive subject matter information relating to the actual

mechanism that gave rise to the particular data. Indeed, the generating mechanism was deliberately veiled in the discussion so far to make this point.

Where does probability come from?

4 A brief introduction to probability

Probability theory provides the mathematical framework for modeling stochastic phenomena of interest. That is, observable phenomena that give rise to data $\mathbf{x}_0{:=}(x_1, x_2, \ldots, x_n)$ that exhibit chance regularity patterns (statistical regularities).

4.1 Kolmogorov's Axiomatic Approach

The axiomatic approach to probability is specified by a probability space $(S, \Im, \mathbb{P}(.))$:

(a) S denotes the set of all possible distinct outcomes.

(b) $\Im$ denotes a set of subsets of S, called *events* of interest, endowed with the mathematical structure of a σ-field, that is, it satisfies the following conditions:

(i) $S{\in}\Im$, (ii) if $A{\in}\Im$, then $\overline{A}{\in}\Im$, (iii) if $A_i{\in}\Im$ for $i{=}1, 2, \ldots, n, \ldots$, then $\bigcup_{i=1}^{\infty} A_i{\in}\Im$.

(c) $\mathbb{P}(.)$**:** $\Im \to [0, 1]$ denotes a set function that satisfies axioms A1-A3 in table 3.

Table 3: Kolmogorov Axioms of Probability

[A1]	$\mathbb{P}(S){=}1$, for any outcomes set S,
[A2]	$\mathbb{P}(A) \geq 0$, for any event $A{\in}\Im$,
[A3]	*Countable Additivity.* For a countable sequence of mutually exclusive events, i.e., $A_i{\in}\Im$, $i{=}1, 2, ..., n, ..$ such that $A_i \cap A_j{=}\varnothing$, for all $i{\neq}j$, $i, j{=}1, 2, ..., n, ...$, then $\mathbb{P}(\bigcup_{i=1}^{\infty} A_i){=}\sum_{i=1}^{\infty} \mathbb{P}(A_i)$.

This formalization renders probability a sub-field of *measure theory* concerned with assigning size, length, content, area, volume, and etc. to sets. In this sense, the probability space $(S, \Im, \mathbb{P}(.))$ provides an idealized description of the stochastic mechanism that gives rise to the events of interest and related events $\Im$, with $\mathbb{P}(.)$ assigning probabilities to events in $\Im$. The mathematical structure of $\Im$, being a field or a σ-field is critically important for this axiomatization. Intuitively it amounts to ensuring that $\Im$ is a set of subsets of S which is closed under the set theoretic operations of union ($\cup$), intersection ($\cap$) and complimentation ( ). That is, if two events A and B belong to $\Im$, so would the events $(A \cup B)$, $(A \cap B)$ and $\overline{A}{=}S{-}A$ and $\overline{B}{=}S{-}B$. All these events will have a clearly assigned a probability by $\mathbb{P}(.)$.

4.2 Conditional probabilities

For any two events A and B in $\Im$, the *conditional probability* formula is:

$$P(A|B){=}\frac{P(A \cap B)}{P(B)}, \text{ for } P(B) > 0, \tag{1}$$

initially proposed by De Moivre (1718). This formula treats the events A and B symmetrically, and thus:

$$P(B|A)=\frac{P(A\cap B)}{P(A)},\ \text{for } P(A)>0. \tag{2}$$

Solving (1) and (2) for $P(A\cap B)$ yields the *multiplication formula*:

$$P(A\cap B)=P(B|A)\cdot P(A)=P(A|B)\cdot P(B). \tag{3}$$

Substituting (3) into (1) yields an alternative but equivalent formula for conditional probability:

$$P(A|B)=\frac{P(B|A)\cdot P(A)}{P(B)},\ \text{for } P(B)>0. \tag{4}$$

When one uses the partition of S stemming from event A and its complement A^c to define the *total probability formula*:

$$P(B)=P(B|A)\cdot P(A)+P(B|A^c)\cdot P(A^c),\ P(A)>0,\ P(A^c)>0. \tag{5}$$

Substituting (5) into (4) yields:

$$\boxed{P(A|B)=\frac{P(B|A)\cdot P(A)}{P(B|A)\cdot P(A)+P(B|A^c)\cdot P(A^c)},\ \text{for } P(B)>0.} \tag{6}$$

More generally, the total probability formula holds for any set of events $(A_1, A_2, ..., A_m)$ that constitute a partition of S in the sense that:

$$A_1\cup A_2\cup ...\cup A_m=S,\ A_i\cap A_j=\varnothing,\ \text{for any } i\neq j,\ i,j=1,2,...,m,$$

$$\begin{aligned} P(B)&=\textstyle\sum_{i=1}^{m} P(A_i)\cdot P(B|A_i),\\ P(A_i|B)&=\frac{P(B|A_i)\cdot P(A_i)}{\sum_{i=1}^{m} P(A_i)\cdot P(B|A_i)},\ \text{for } P(B)>0. \end{aligned} \tag{7}$$

Example 1. Consider the random experiment of tossing a coin twice:

$$S=\{(HH),(HT),(TH),(TT)\}$$

Let the events of interest be:

$$A_1=\{(HH),(HT)\},\ A_2=(TH)\},\ A_3=\{(TH)\},\ B=\{(HT),(TH),(TT)\}.$$

Assuming that the coin is fair, one can assign probabilities to all events in $\Im$:

$$P(A_1)=.5,\ P(A_2)=.25,\ P(A_3)=.25,\ P(B)=.75.$$

It is important to note that:

$$\begin{aligned} &P(A_1\cap B)=.25,\ P(A_2\cap B)=.25,\ P(A_3\cap B)=.25,\\ &P(B)=\textstyle\sum_{i=1}^{3} P(A_i\cap B)=.75. \end{aligned}$$

Hence, the conditional probability formula in (7) yields:

$$P(A_i|B)=\frac{P(A_i\cap B)}{P(B)}=\frac{.25}{.75}=\frac{1}{3},\ i=1,2,3.$$

4.3 Bayes' rule in terms of 'events'

The conditional probability formula in (4) is transformed into an *updating rule* by interpreting the two events A and B as a *hypothesis* H (e.g. $\theta = .5$) and *evidence* E, (e.g. data $\mathbf{x}_0$), respectively, to yield **Bayes' rule**:

$$P(H|E)=\frac{P(E|H)\cdot P(H)}{P(E)},\ P(E)>0, \tag{8}$$

attributed to Bayes' (1763). Its components are interpreted as follows:

(i) $P(H|E)$ is the *posterior probability* of H given E,

(ii) $P(E|H)$ is the *likelihood* of E given H,

(iii) $P(H)$ is the *prior probability* of H, and

(iv) $P(E)$ is the initial *probability of evidence* E.

It is claimed that Bayes' rule in (8) is self-evident since it constitutes an instance of (4) or (6). This claim is misleading because the nature of the events involved and the assignment of probabilities raise legitimate questions of both substance and interpretation. Contrary to the conventional wisdom, (8) is not an instantiation of (4) or (6), because it raises several crucial issues of meaningfulness and applicability; see Appendix B.

4.4 Bayesian confirmation in terms of events

The Bayesian confirmation theory relies on comparing the prior with the posterior probability of a particular hypothesis H:

$$\text{[i] Confirmation:} \qquad P(H|E) > P(H)$$

$$\text{[ii] Disconfirmation:} \quad P(H|E) < P(H)$$

The *degree of confirmation* is evaluated using some measure $\mathfrak{c}(H,E)$ of the 'degree to which E raises the probability of H'. Examples of such Bayesian measures are (Fitelson, 1999):

$$d(H,E)=P(H|E)-P(H),$$

$$m(H,E)=P(E|H)-P(E),$$

$$r(H,E)=\frac{P(H|E)}{P(H)}.$$

Using $\mathfrak{c}(H,E)$ one can define *relative evidence* as:

$\mathfrak{c}(H,E)$ indicates that evidence E favors hypothesis H_1 over H_0, iff:

$$\mathfrak{c}(H_1,E) > \mathfrak{c}(H_0,E).$$

For instance using the measure $r(H,E)$ in the case of two competing hypotheses H_0 and H_1 :

$$\frac{P(H_1|E)}{P(H_1)} > \frac{P(H_0|E)}{P(H_0)} \overset{\text{Bayes}}{\Leftrightarrow} \frac{P(E|H_1)}{P(E)} > \frac{P(E|H_0)}{P(E)} \Leftrightarrow \frac{P(E|H_1)}{P(E|H_0)} > 1$$

where $\frac{P(E|H_1)}{P(E|H_0)}$ is the (Bayesian) likelihood ratio. For comparison purposes let us contrast this to the *ratio of posteriors*:

$$\frac{P(H_1|E)}{P(H_0|E)}=\frac{\frac{P(E|H_1)\cdot P(H_1)}{P(E)}}{\frac{P(E|H_0)\cdot P(H_0)}{P(E)}}=\left(\frac{P(E|H_1)}{P(E|H_0)}\right)\left(\frac{P(H_1)}{P(H_0)}\right)>1, \tag{9}$$

which is the product of the 'likelihood ratio' $\frac{P(E|H_1)}{P(E|H_0)}$ *and* the ratio of the priors $\frac{P(H_1)}{P(H_0)}$.

In light of the fact that the above Bayesian confirmation theory inherits all the weaknesses (a)-(d) raised in Appendix B relating to the potential arbitrariness of the various probabilistic assignments pertaining to unobservable events, the whole confirmation endeavor seems like playing war games on an imaginary map without any actual connection to the reality we would like to understand.

4.5 Random Variables and Statistical Models

For empirical modeling purposes the probability space $(S, \Im, \mathbb{P}(.))$ is much too abstract because real data come in the form of numbers that live on the real line $\mathbb{R}=(-\infty, \infty)$. This calls for mapping $(S, \Im, \mathbb{P}(.))$ onto $\mathbb{R}$ using the concept of a *random variable:* a real-valued function:

$$X(.)\textbf{:}\ S \to \mathbb{R}, \text{ such that } \{X \leq x\} \in \Im \text{ for all } (\forall)\ x \in \mathbb{R}.$$

That is, $X(.)$ assigns numbers to the elementary events in S in such a way so as to preserve the original event structure of interest $(\Im)$. The key role of the random variable $X(.)$ is to transform the original $(S, \Im, \mathbb{P}(.))^n$ into a statistical model $\mathcal{M}_{\boldsymbol{\theta}}(\mathbf{x})$ defined on the real line:

$$(S, \Im, \mathbb{P}(.))^n \ \stackrel{X(.)}{\longrightarrow}\ \mathcal{M}_{\boldsymbol{\theta}}(\mathbf{x})=\{f(\mathbf{x};\boldsymbol{\theta}),\ \boldsymbol{\theta} \in \Theta \subset \mathbb{R}^m\},\ \mathbf{x} \in \mathbb{R}^n_X,\ m<n, \tag{10}$$

where $f(\mathbf{x};\boldsymbol{\theta}),\ \mathbf{x} \in \mathbb{R}^n_X$ denotes the **joint distribution of the sample** $\mathbf{X}:=(X_1, \ldots, X_n)$, $\mathbb{R}^n_X$ the sample space (the range of value of the sample $\mathbf{X}$), and Θ the parameter space (the range of values of θ). The statistical model $\mathcal{M}_{\boldsymbol{\theta}}(\mathbf{x})$ can be viewed as a parameterization of the stochastic process $\{X_k, k \in \mathbb{N}\}$ whose probabilistic structure is chosen so as to render data $\mathbf{x_0}\textbf{:}=(x_1, \ldots, x_n)$ a *typical realization* thereof.

Why we need the **joint** distribution of the sample to specify a statistical model in (10). Because when the data exhibit any form of **dependence or/and heterogeneity**, one needs the joint $f(\mathbf{x};\boldsymbol{\theta}),\ \mathbf{x} \in \mathbb{R}^n_X$, distribution because the marginal distributions cannot account for them.

Example. Using the random variable $X(H)=1,\ X(T)=0$, one can define the relevant density function via:

$$\mathbb{P}(H)=\theta=f(x=1),\ \mathbb{P}(T)=1-\theta=f(x=0),\ 0 \leq \theta \leq 1,$$

giving rise to the well-known **Bernoulli density**:

$$f(x;\theta)=\theta^x(1-\theta)^{1-x}, \quad x=0, 1,\ \theta \in [0, 1].$$

Alternatively, this formula can be expressed as a table given below.

x	$f(x;\theta)$
0	$(1-\theta)$
1	θ

Parameters and moments of distributions

For modeling and inference purposes we need to focus on both the marginal $f(x_k;\boldsymbol{\theta})$, and joint distributions $f(\mathbf{x};\boldsymbol{\theta})$, $\mathbf{x}\in\mathbb{R}_X^n$, to learn how to handle them. In particular, in practice we often relate the unknown parameters to the moments of the underlying density function because it makes the interpretation, the modeling and the inference easier to handle.

Raw moments: $E(X^k)=\sum_{x\in\mathbb{R}_X} x^k f(x;\theta)$ (discrete),
or $E(X^k)=\int_{x\in\mathbb{R}_X} x^k f(x;\theta)dx$ (continuous), $k=1,2,...$
The most widely used raw moment is the **mean**:

$$\text{Discrete: } E(X)=\sum_{x\in\mathbb{R}_X} xf(x;\theta), \quad \text{Continuous: } E(X)=\int_{x\in\mathbb{R}_X} xf(x;\theta)dx$$

Example. In the case of the simple Bernoulli distribution:

$$E(X)=\sum_{x\in\mathbb{R}_X} xf(x;\theta)=(1)\theta+(0)(1-\theta)=\theta.$$

Central moments: $E([X-E(X)]^k)=\sum_{x\in\mathbb{R}_X}[x-E(X)]^k f(x;\theta)$ (discrete),
or $E(X^k)=\int_{x\in\mathbb{R}_X}[x-E(X)]^k f(x;\theta)dx$ (continuous), $k=2,...$
The most widely used raw moment is the **variance**:

$$Var(X)=E([X-E(X)]^2)=\sum_{x\in\mathbb{R}_X}[x-E(X)]^2 f(x;\theta).$$

Example. In the case of the simple Bernoulli distribution:

$$Var(X)=\sum_{x\in\mathbb{R}_X}(x-\theta)^2 f(x;\theta)=(1-\theta)^2\theta+(0-\theta)^2(1-\theta)=\theta(1-\theta).$$

Example. Consider the discrete random variable X with a density function:

x	0	1	2
$f(x)$	.3	.3	.4

(11)

$$E(X)=0(.3)+1(.3)+2(.4)=1.1, \quad Var(X)=(0-1.1)^2(.3)+(1-1.1)^2(.3)+(2-1.1)^2(.4)=.69$$

4.6 Random (IID) Sample

The concept of a *random sample* can be defined in terms of the *joint distribution of the sample* $\mathbf{X}:=(X_1,X_2,...,X_n)$, say $f(x_1,x_2,...,x_n;\boldsymbol{\theta})$, for all $(x_1,x_2,...,x_n)\in\mathbb{R}_X^n$.

Independence (I): the sample $\mathbf{X}:=(X_1,X_2,\ldots,X_n)$ is said to be *Independent* if the (n-dimensional) joint distribution $f(\mathbf{x};\boldsymbol{\theta})$, $\mathbf{x}\in\mathbb{R}_X^n$, reduces into a product of (1-dimensional) marginal distributions (for all $(x_1,x_2,...,x_n)\in\mathbb{R}_X^n$):

$$f(x_1,x_2,...,x_n;\boldsymbol{\theta})=f_1(x_1;\boldsymbol{\theta}_1)\cdot f_2(x_2;\boldsymbol{\theta}_2)\cdot\cdots\cdot f_n(x_n;\boldsymbol{\theta}_n)=\prod_{k=1}^n f_k(x_k;\boldsymbol{\theta}_k), \quad (12)$$

Identically Distributed (ID): the sample $\mathbf{X}:=(X_1,X_2,\ldots,X_n)$ is said to be *Identically Distributed* if the marginal distributions are identical:

$$f_k(x_k;\boldsymbol{\theta}_k)=f(x_k;\boldsymbol{\theta}), \quad \text{for all } k=1,2,...,n. \quad (13)$$

Note that this means both, the density functions have the same formula and the unknown parameters are common to all of them.

Examples. (i) Tossing a coin once with outcomes: Heads (H), Tails (T).
(ii) Sampling *with replacement* from an urn containing red (R) and black (B) balls.
(iii) Observing the gender (B or G) of newborns during a year in NY city.

Example 4.33. Consider the bivariate density given below.

$x \diagdown y$	0	1	$f_x(x)$
0	$f(0,0)$=.3	$f(0,1)$=.2	0.5
2	$f(2,0)$=.3	$f(0,0)$=.2	0.5
$f_y(y)$	0.6	0.4	1

(14)

To check whether X and Y are independent, we need to verify that the equality in (12) holds, for *all* values of X and Y:

(X,Y)=$(0,0)$: $f(0,0)$=$f_x(0)\cdot f_y(0)$=(.3)=(.5)(.6),

(X,Y)=$(1,0)$: $f(1,0)$=$f_x(1)\cdot f_y(0)$=(.3)=(.5)(.6),

(X,Y)=$(0,2)$: $f(0,2)$=$f_x(0)\cdot f_y(2)$=(.2)=(.5)(.4),

(X,Y)=$(1,2)$: $f(1,2)$=$f_x(1)\cdot f_y(2)$=(.2)=(.5)(.4).

These results show that (12) holds, and thus X and Y are independent. They are not Identically distributed, however, because the two marginal distributions are different:

x	$f_x(x)$
0	0.5
2	0.5
	1

y	$f_y(y)$
0	0.6
1	0.4
	1

Their differences are (i) X and Y take different values $\mathbb{R}_X$:=$\{0,2\} \neq \mathbb{R}_Y$:=$\{0,1\}$, and (ii) their probabilities are not equal.

Example. Consider the three bivariate distributions (a)-(c) given below.

$x \diagdown y$	0	2	$f_x(x)$
1	0.18	0.12	0.3
2	0.42	0.28	0.7
$f_y(y)$	0.6	0.4	1

(a)

$x \diagdown y$	0	1	$f_x(x)$
0	0.18	0.12	0.3
1	0.42	0.28	0.7
$f_y(y)$	0.6	0.4	1

(b)

$x \diagdown y$	0	1	$f_x(x)$
0	0.36	0.24	0.6
1	0.24	0.16	0.4
$f_y(y)$	0.6	0.4	1

(c)

The random variables (X,Y) are independent in all three cases (verify!). The random variables in (a) are not Identically Distributed because $\mathbb{R}_X \neq \mathbb{R}_Y$, and $f_x(x) \neq f_y(y)$ for some $(x,y) \in \mathbb{R}_X \times \mathbb{R}_Y$.

The random variables in (b) are not Identically Distributed because even though $\mathbb{R}_X$=$\mathbb{R}_Y$, $f_x(x) \neq f_y(y)$ for some (x,y)∈$\mathbb{R}_X \times \mathbb{R}_Y$.

Finally, the random variables in (c) are Identically Distributed because $\mathbb{R}_X$=$\mathbb{R}_Y$, and $f_x(x)$=$f_y(y)$ for all (x,y)∈$\mathbb{R}_X \times \mathbb{R}_Y$.

Example. Consider the following bivariate distribution.

$y \diagdown x$	0	1	$f_Y(y)$
0	.0	.25	.25
1	.25	.50	.75
$f_X(x)$	.25	.75	1

(15)

$E(X)$=$0(.25)+1(.75)$=.75, $E(Y)$=$0(.25)+1(.75)$=.75,

$Var(X)$=$(0-.75)^2(.25)+(1-.75)^2(.75)$=.1875,

$Var(Y)$=$(0-.75)^2(.25)+(1-.75)^2(.75)$=.1875.

Are these two r.v's X and Y *Independent*? The answer is no because for (x,y)=$(0,0)$:

$$f(0,0)=0 \neq f_X(0) \cdot f_Y(0)=(.25)(.25).$$

The correlation coefficient $Corr(X,Y)$ is defined by:

$$Corr(X,Y)=\frac{Cov(X,Y)}{\sqrt{Var(X)\cdot Var(Y)}},$$

where $Cov(X,Y)$=$E[(X-E(X))(Y-E(Y))]$=$\sum_{x\in\mathbb{R}_X}\sum_{x\in\mathbb{R}_X}(x-E(X))(y-E(Y))f(x,y)$. In the case of (15):

$$\begin{aligned} Cov(X,Y) \quad &=(0-.75)(0-.75)(0)+(0-.75)(1-.75)(.25)+ \\ &+(1-.75)(0-.75)(.25)+(1-.75)(1-.75)(.5)=-.0625 \end{aligned}$$

Hence, $Corr(X,Y)$=$\frac{Cov(X,Y)}{\sqrt{Var(X)\cdot Var(Y)}}$=$\frac{-.0625}{\sqrt{(.1875)^2}}$=$-0.333$.

4.7 Simple statistical models

The **simple Bernoulli model** is specified by:

$$\mathcal{M}_\theta(\mathbf{x})\text{: } X_k \backsim \mathsf{BerIID}(\theta,\theta(1-\theta)),\ x_k=0,1,\ \theta\in[0,1],\ k=1,2,\dots n,...$$

where '$\mathsf{BerIID}(\theta,\theta(1-\theta))$' stands for Bernoulli, Independent and Identically Distributed (IID), with mean θ and variance $\theta(1-\theta)$, k is an index that denotes the order of the sample.

Example. In the case of the simple Bernoulli model, the derivation of the distribution of the sample, $f(\mathbf{x};\boldsymbol{\theta})$, $\mathbf{x}\in\mathbb{R}_X^n$, by imposing the IID and Bernoulli assumptions sequentially, takes the following form:

$$\begin{aligned} f(x_1,x_2,...,x_n;\boldsymbol{\theta}) \quad &\overset{\text{I}}{=}f_1(x_1;\theta_1)\cdot f_2(x_2;\theta_2)\cdot\dots\cdot f_n(x_n;\theta_n)=\textstyle\prod_{k=1}^n f_k(x_k;\theta_k)\overset{\text{IID}}{=}\prod_{k=1}^n f(x_k;\theta)= \\ &\overset{\text{BerIID}}{=}\textstyle\prod_{k=1}^n \theta^{x_k}(1-\theta)^{1-x_k}=\theta^{\sum x_k}(1-\theta)^{\sum(1-x_k)}=\theta^y(1-\theta)^{n-y},\ \mathbf{x}\in\{0,1\}^n, \end{aligned} \tag{16}$$

where $Y=\sum X_k$. Do we need the above complicated notation? Yes, because without it one would never understand the concept of the distribution of the sample $f(\mathbf{x};\boldsymbol{\theta}),\ \mathbf{x}\in\mathbb{R}^n_X$, or the likelihood function $L(\boldsymbol{\theta};\mathbf{x}_0),\ \boldsymbol{\theta}\in\Theta$; see Appendices A and C.

The Binomial distribution

In the case of an IID sample from a simple Bernoulli distribution, one can show that the summation of $(X_1, X_2, ..., X_n)$:

$$Y=\sum_{i=1}^{n} X_i,\ \ y=0,1,...,n,$$

is Binomially distributed and Y is the number of 1's in n trials; the values X_i=0 contribute nothing to the summation.

The distribution of Y takes the form:

$$f(y;\theta,n)=\binom{n}{y}\theta^y(1-\theta)^{n-y},\ \ y=0,1,2,...,n,$$

denoted by $Y\backsim \mathsf{Bin}(n\theta,\ n\theta(1-\theta))$. Note: $\binom{n}{y}=\frac{n!}{(n-y)!y!}$, where $n!=n(n-1)(n-2)\cdots(2)(1)$, is called n factorial. The factor $\binom{n}{y}$ represents the different number of different ways y 1's can appear in a sequence of length n.

Example. Consider the case of n=20 and θ=.5. The table with the Binomial probabilities is given below:

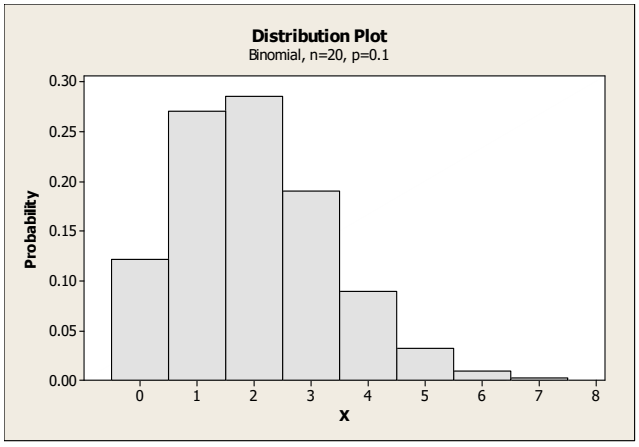
x	3	4	5	6	7	8	9	10	11	12	13	14	15	16
$f(x;.5)$	.001	.005	.015	.037	.074	.12	.16	.176	.16	.12	.074	.037	.015	.005

where $f(0;.5)$, $f(1;.5)$, $f(2;.5)$, $f(17;.5)$, $f(18;.5)$, $f(19;.5)$, $f(20;.5)$ are *not* shown because they are very close to zero. To illustrate evaluation of the Binomial probabilities take the case where Y=12. The probability associated with this value is given by:

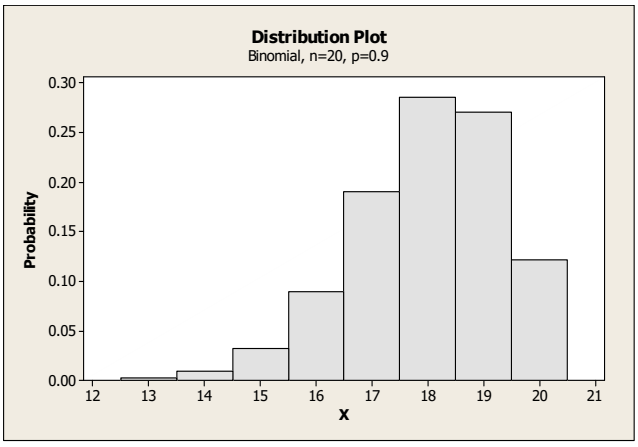
$$f(y=12;.5,n=20)=\binom{20}{12}(.5)^{12}(1\text{-}.5)^8=(125970)(.00244)(.00391)=.120. \qquad (17)$$

In practice we might also be interested in tail areas of the form:

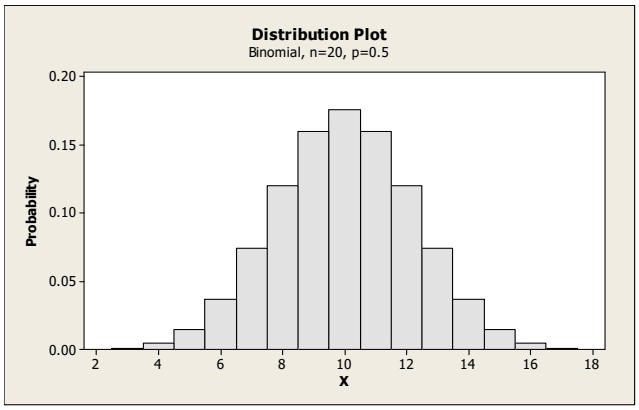
$$\mathbb{P}(Y\geq 12)=\sum_{y=12}^{20} f(y;.5,20)=.12+.074+.037+.015+.005+.001=.252. \qquad (18)$$



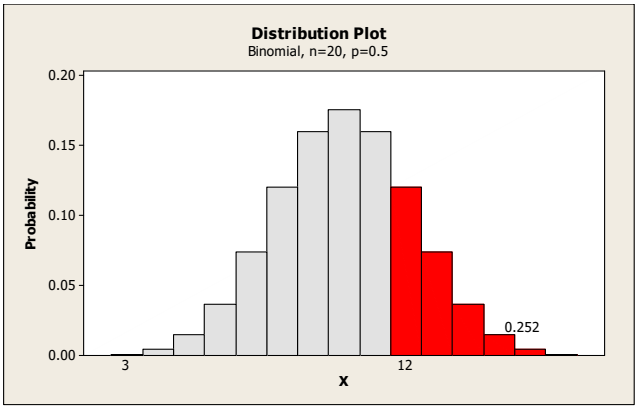
Binomial: $f(y;\theta=.1,n=20)$



Binomial: $f(y;\theta=.9,n=20)$



Binomial: $f(y;\theta{=}.5, n{=}20)$



Binomial: $\mathbb{P}(Y \geq 12){=}\sum_{y=12}^{20} f(y;.5,20)$

The Normal distribution

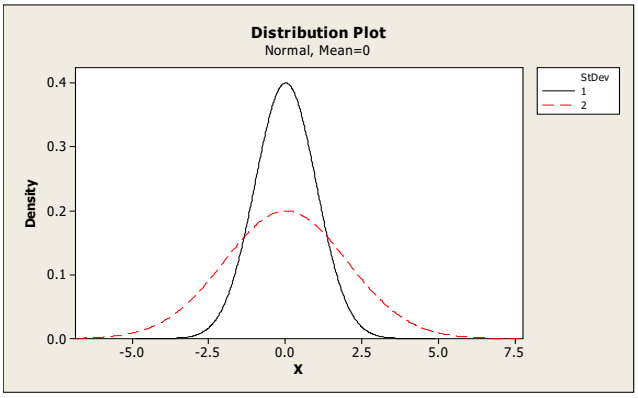
Another distribution of even greater interest in statistical inference is the Normal distribution whose density function takes the form:

$$f(x;\boldsymbol{\theta}){=}\tfrac{1}{\sqrt{2\pi\sigma^2}}\exp\left\{-\tfrac{1}{2\sigma^2}(x-\mu)^2\right\}, \quad \boldsymbol{\theta}{:=}(\mu,\sigma^2){\in}\mathbb{R}{\times}\mathbb{R}_+, \;\; x{\in}\mathbb{R}.$$

where exp is the base of the natural logarithm. Note that the Normal distribution has two unknown parameters (μ,σ^2) where $\mu{\in}\mathbb{R}{:=}(-\infty,\infty)$ and $\sigma^2{\in}\mathbb{R}_+{:=}(0,\infty)$. In practice we denote this using the notation:

$$X \backsim \mathsf{N}(\mu,\sigma^2), \;\; x{\in}\mathbb{R},$$

where $E(X){=}\mu$, $Var(X){=}\sigma^2$. The graph below shows the densities of $\mathsf{N}(0,1)$ [in solid line] and $\mathsf{N}(0,4)$ [dotted line], shows the effect of increasing σ^2.

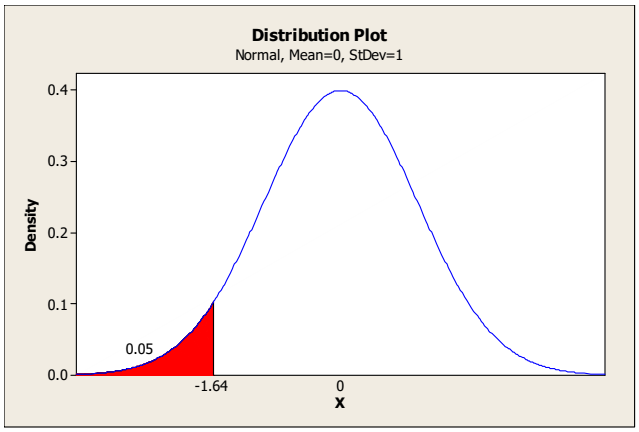


Normal density with $\sigma = 1$ and $\sigma = 2$

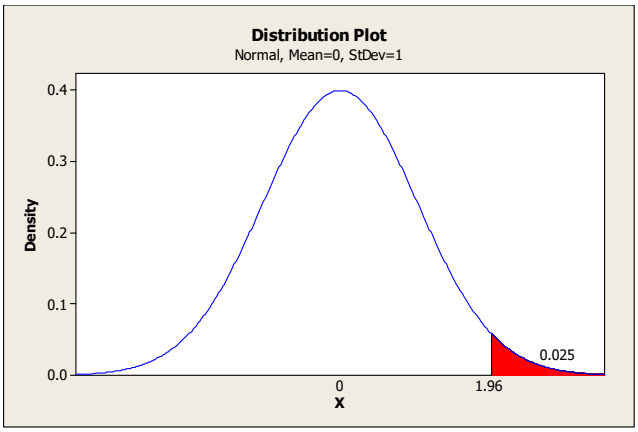
The *standard Normal density*, denoted by $X \backsim \mathsf{N}(0,1)$, $x{\in}\mathbb{R}$, is widely used in statistical inference and the graphs below give certain features of this density relating

to probabilities of interest in such a context. Due to the symmetry of the Normal distribution one can use the tail areas shown below to derive the following interesting features concerning the area around the mean in terms of standard deviation units:

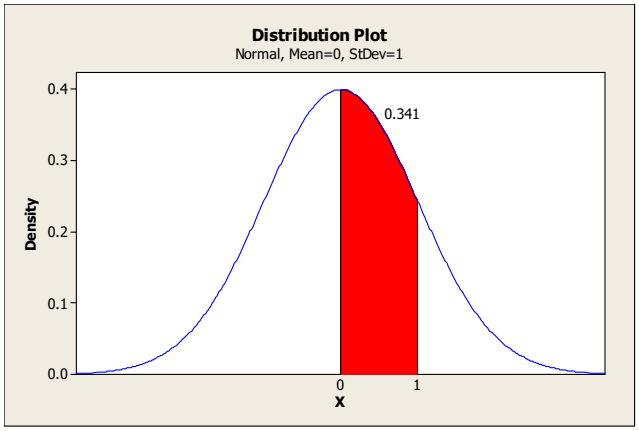
$$\mathbb{P}(-1 < X < 1){=}.682, \quad \mathbb{P}(-2 < X < 2){=}.954, \quad \mathbb{P}(-3 < X < 3){=}.997.$$



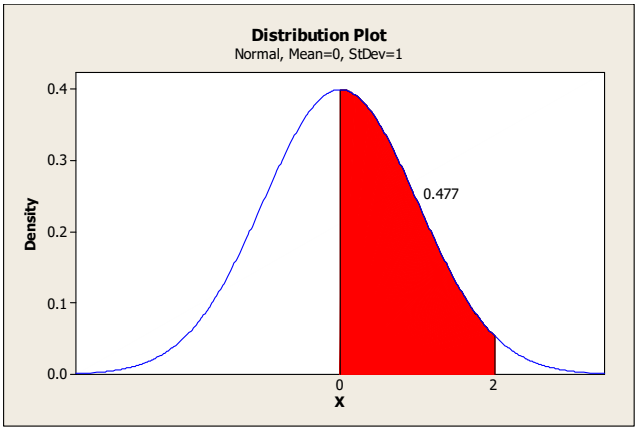
Normal: $\mathbb{P}(X < -1.64) = .05$



Normal: $\mathbb{P}(X > 1.96) = .025$



Normal: $\mathbb{P}(0 < X < 1) = .34$



Normal: $\mathbb{P}(0 < X < 2) = .477$

It is trivial to show that when $X \backsim \mathsf{N}(\mu, \sigma^2)$ one can transform it into a $\mathsf{N}(0,1)$ by subtracting the mean (μ) and dividing by the standard deviation ($\sigma{=}\sqrt{Var(X)}{=}\sqrt{\sigma^2}$):

$$Z{=}\left(\tfrac{X-\mu}{\sigma}\right) \backsim \mathsf{N}(0,1), \quad z{\in}\mathbb{R}.$$

Combining the Normal distribution with an IID sample yields the **simple Normal model**:

$$\mathcal{M}_{\boldsymbol{\theta}}(\mathbf{x})\text{: } X_k \backsim \mathsf{NIID}(\mu, \sigma^2),\ x_k{\in}\mathbb{R},\ \mu{\in}\mathbb{R},\ \sigma^2 > 0,\ k{\in}\mathbb{N}.$$

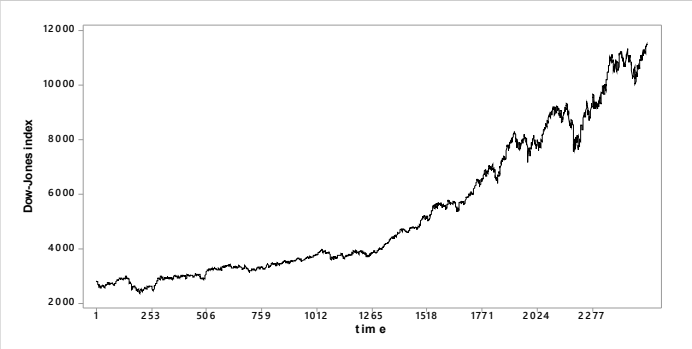
More realistic statistical models

The notion of a *simple statistical model* (IID sample) is likely to be inappropriate for the following stochastic phenomena.

Examples. (i) Sampling *without replacement* from an urn which contains red (R) and black (B) balls.

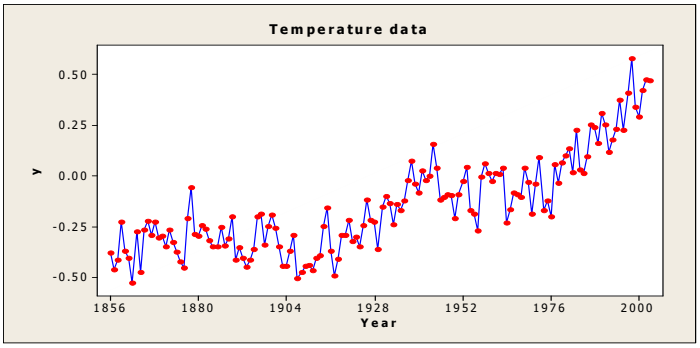
It turns out that for this experiment the underlying distribution is Bernoulli but the sample $\mathbf{X}:=(X_1, X_2, \ldots, X_n)$ is *no* longer random; $\mathbf{X}$ is *dependent*.

(ii) Observing the daily changes of the Dow Jones (D-J) index during a certain period (a year).



Daily D-J index 1990-1999

(iii) Observing the annual average global surface air temperature since 1856.



Surface air temperature, annual series

Both of the above data series exhibit strong mean heterogeneity (trend) as well as temporal dependence (irregular cycles) that will require statistical models that account for such systematic statistical information; a random (IID) sample will not do.

5 Appendix A: The Likelihood Function

The **primary aim** of the frequentist approach is to *learn from data* about the "true" statistical data GM:

$$\mathcal{M}_{\boldsymbol{\theta}^*}(\mathbf{x})=\{f(\mathbf{x};\boldsymbol{\theta}^*)\},\ \forall \mathbf{x}\in\mathbb{R}^n_X.$$

The expression "$\boldsymbol{\theta}^*$ denotes the true value of $\boldsymbol{\theta}$" is a shorthand for saying that "data $\mathbf{x}_0$ constitute a typical realization of the sample $\mathbf{X}$ with distribution $f(\mathbf{x};\boldsymbol{\theta}^*)$".

Likelihood function. A crucial role in 'learning from data' is played by the *likelihood function* defined by:

$$\boxed{L(\boldsymbol{\theta};\mathbf{x}_0)\propto f(\mathbf{x}_0;\boldsymbol{\theta}),\ \forall\boldsymbol{\theta}\in\Theta,}$$

where $\propto$ reads 'proportional to' the distribution of the sample $f(\mathbf{x}_0;\boldsymbol{\theta})$ evaluated at $\mathbf{x}_0$. In light of the fact that we view the statistical model as the stochastic mechanism that could have generated $\mathbf{x}_0{:=}(x_1, x_2, ..., x_n)$, it seems intuitively obvious to evaluate $f(\mathbf{x};\boldsymbol{\theta}),\ \mathbf{x}{\in}\mathbb{R}^n_X$, at $\mathbf{X}{=}\mathbf{x}_0$, and pose the reverse question:

▶ how likely does $f(\mathbf{x}_0;\boldsymbol{\theta})$ render the different values of $\boldsymbol{\theta}$ in Θ to have been the 'true' value $\boldsymbol{\theta}^*$?

Note that '$\boldsymbol{\theta}^*$ denotes the true value of $\boldsymbol{\theta}$' is a shorthand for saying that 'data $\mathbf{x}_0$ constitute a typical realization of the sample $\mathbf{X}$ with distribution $f(\mathbf{x};\boldsymbol{\theta}^*),\ \mathbf{x}{\in}\mathbb{R}^n_X$'. Hence, the likelihood function yields the *likelihood* (proportional to the probability) of getting $\mathbf{x}_0$ under different values of $\boldsymbol{\theta}$.

Example. Consider the *simple Bernoulli model* specified above. The distribution of the sample (derived above) is:

$$f(\mathbf{x};\boldsymbol{\theta}){=}\theta^Y(1{-}\theta)^{n-Y},\ \mathbf{x}{\in}\{0,1\}^n,\ \text{where } Y{=}\textstyle\sum_{k=1}^n X_k.$$

Hence, the Likelihood Function (LF) is:

$$L(\theta;\mathbf{x}_0) \propto \theta^y(1{-}\theta)^{(n-y)},\ \forall\theta{\in}[0,1]. \tag{19}$$

Note that $f(\mathbf{x};\theta), \mathbf{x}{\in}\{0,1\}^n$ is a discrete density function of Y, but the LF, $L(\theta;\mathbf{x}_0),\ \theta{\in}[0,1]$, is a continuous function of $\theta{\in}[0,1]$. In general a crucial distinction is:

$$f(\mathbf{x};\boldsymbol{\theta}),\ \mathbf{x}{\in}\mathbb{R}^n_X \text{ vs. } L(\boldsymbol{\theta};\mathbf{x}_0),\ \boldsymbol{\theta}{\in}\Theta.$$

In the simple Bernoulli model, Y is Binomially distributed:

$$Y{=}\textstyle\sum_{k=1}^n X_k \backsim \mathsf{Bin}\left(n\theta,\ n\theta(1{-}\theta); n\right), \tag{20}$$

Example. The distribution $f(y;\theta),\ y{=}1, 2, .., n$, is shown in figure 4 for $n{=}100,\ \theta{=}.56$, is a one-dimensional representation of $f(\mathbf{x};\theta), \mathbf{x}{\in}\{0,1\}^n$ using $f(\mathbf{x};\theta){=}\theta^Y(1{-}\theta)^{n-Y},\ y{=}0, 1, 2, ..., n$. This discrete distribution in fig. 4 should be contrasted with the Likelihood Function (LF) $L(\theta;\mathbf{x}_0){=}\theta^y(1{-}\theta)^{n-y}$, $\theta{\in}[0,1]$, (figure 5) which is a continuous and differentiable function of θ.

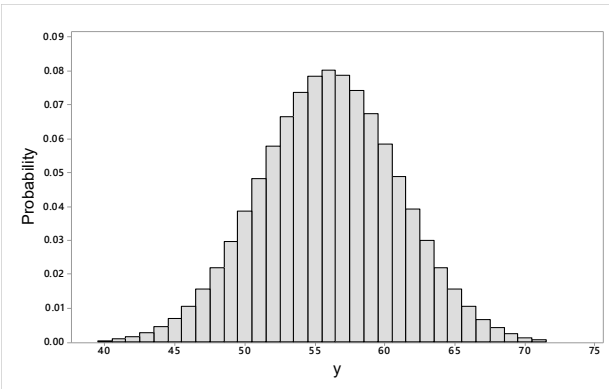


Fig. 4: $Y \backsim \mathsf{Bin}(n\theta,\ n\theta(1{-}\theta))$, $n{=}100,\ \theta{=}.56$

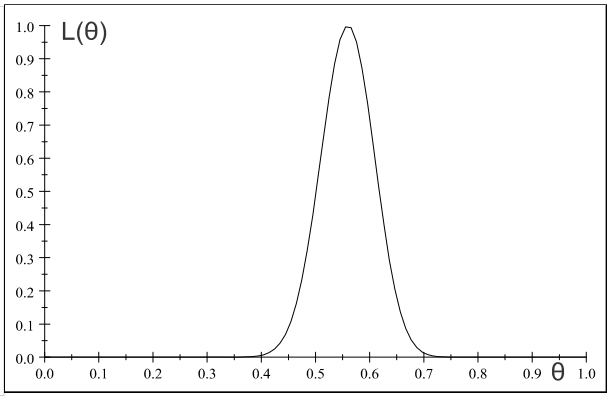


Fig. 5: $L(\theta;\mathbf{x}_0),\ \theta{\in}[0,1],\ Y{=}.56$

This brings out an important feature of the likelihood function that pertains to the scaling on the vertical axis. This scaling is arbitrary since one can define the Likelihood Function (LF), equivalently as:

$$L(\boldsymbol{\theta};\mathbf{x}_0)=c(\mathbf{x}_0)\cdot f(\mathbf{x}_0;\boldsymbol{\theta}),\ \ \boldsymbol{\theta}\in\Theta, \tag{21}$$

where $c(\mathbf{x}_0)$ depends only on the data $\mathbf{x}_0$ and *not* on $\boldsymbol{\theta}$. Indeed, the likelihood function graph in figure 5 has been scaled using $c(\mathbf{x}_0)=[1/L(\widehat{\theta};\mathbf{x}_0)]$, where $L(\widehat{\theta};\mathbf{x}_0)$ denotes the estimated likelihood with $\widehat{\theta}=.56$, being the Maximum Likelihood (ML) estimate of θ; see Lecture Notes 1. This renders the graph of the likelihood function easier to read as well as compare the likelihood values for different θ's. To get some idea of comparing likelihood values for different values of θ consider the following example.

Example. For the *simple Bernoulli model*, with $n=100$, $\theta=.56$, let us compare the likelihood of two values of $\theta=\mathbb{P}(X=1)$ within the interval $[0,1]$, $\theta_1=.45$ and $\theta_2=.62$; see fig. 5. The values of the likelihood function are:

$$L(.45;\mathbf{x}_0)=(.45)^{56}(1-.45)^{44}=1.431\,7\times 10^{-31},$$
$$L(.62;\mathbf{x}_0)=(.62)^{56}(1-.62)^{44}=7.663\,2\times 10^{-31},$$

which are tiny, and thus highly vulnerable to approximation errors. Having said that, due to the presence of the arbitrary constant $c(\mathbf{x}_0)$ in (21), the LF can be scaled to avoid such problems. An obvious way to scale the LF is to divie by the estimated LF:

$$L(\widehat{\theta};\mathbf{x}_0)=(.56)^{56}(1-.56)^{44}=1.623\,5\times 10^{-30},$$

which is also a tiny number. The scaled likelihood function $\frac{L(\theta;\mathbf{x}_0)}{L(\widehat{\theta};\mathbf{x}_0)}$, however, takes values between zero and one:

$$\frac{L(.45;\mathbf{x}_0)}{L(\widehat{\theta};\mathbf{x}_0)}=\frac{(.45)^{56}(1-.45)^{44}}{(.56)^{56}(1-.56)^{44}}=.0882,\ \ \frac{L(.62;\mathbf{x}_0)}{L(\widehat{\theta};\mathbf{x}_0)}=\frac{(.62)^{56}(1-.62)^{44}}{(.56)^{56}(1-.56)^{44}}=.472,$$

which renders the comparison of the two easier. CAUTION, however, is advised to avoid misconstruing the scaled likelihood function as assigning probabilities to $\theta\in[0,1]$, just because of the particular scaling used.

In light of the arbitrariness of the scaling factor $c(\mathbf{x}_0)$, the only meaningful measure of relative likelihood for two values of θ comes in the form of the ratio:

$$\frac{L(.62;\mathbf{x}_0)}{L(.45;\mathbf{x}_0)}=\frac{c(\mathbf{x}_0)(.62)^{56}(1-.62)^{44}}{c(\mathbf{x}_0)(.45)^{56}(1-.45)^{44}}=\frac{(.62)^{56}(1-.62)^{44}}{(.45)^{56}(1-.45)^{44}}=5.353,$$

since the scaling factor *cancels out*, being the same for all values $\theta\in[0,1]$. This renders the value $\theta=.62$ more than 5 times likelier than $\theta=.45$. Does that mean that $\mathbf{x}_0$ this provides evidence that $\theta=.62$ is close to the θ^*, the true θ?

Not necessarily! This is because, by definition, the values of the likelihood function $L(\theta;\mathbf{x}_0)$ are dominated by the Maximum Likelihood (ML) estimate $\widehat{\theta}=.56$. Moreover, in point estimation there is no warranted inferential claim that $\widehat{\theta}=.56$ is approximately equal to θ^* due to the sampling variability associated with the ML estimator:

$$\widehat{\theta}(\mathbf{X})=\overline{X}=\tfrac{1}{n}\textstyle\sum_{i=1}^{n}X_i\backsim \mathsf{sBin}(\theta,\frac{\theta(1-\theta)}{n}),$$

where $\mathsf{sBin}(\theta, \frac{\theta(1-\theta)}{n})$ denotes a 'scaled' Binomial distribution with mean θ and variance $\frac{\theta(1-\theta)}{n}$; see (20). This suggests that for a particular sample realization $\mathbf{x}_0$ there is no reason to presume that $\widehat{\theta}(\mathbf{X}) \simeq \theta^*$, since for an unbiased estimator $\widehat{\theta}(\mathbf{X})$ of θ^* is only its mean that has such property: $E(\widehat{\theta}(\mathbf{X}))=\theta^*$. That is, if one were to use the long-run metaphor to visualize the sampling distribution of $\widehat{\theta}(\mathbf{X})$, one would have to draw N (say N=10000) sample realizations $\mathbf{x}_i$, i=1, 2, ...N and construct the empirical sampling distribution of $\widehat{\theta}(\mathbf{X})$ and evaluate its mean to be able to claim that $\widehat{E}(\widehat{\theta}(\mathbf{X}))\simeq\theta^*$.

In contrast to a point estimator, both confidence intervals and hypothesis testing account for this sampling variability by using statistics of the form:

$$\widehat{\theta}(\mathbf{X}) \pm c_{\frac{\alpha}{2}} \frac{\sqrt{\widehat{\theta}(\mathbf{X})(1-\widehat{\theta}(\mathbf{X}))}}{\sqrt{n}}, \ \ \frac{\sqrt{n}(\widehat{\theta}(\mathbf{X})-\theta_0)}{\sqrt{\theta_0(1-\theta_0)}}.$$

More details will be given later during this course.

6 Appendix B: A closer look at Bayes' formula

Let us look more closely at *Bayes' rule*:

$$P(H|E)=\frac{P(E|H)\cdot P(H)}{P(E)}, \ \ P(E)>0, \tag{22}$$

where H denotes a *hypothesis* and E *evidence relevant to* H, and:

(i) $P(H|E)$ is the *posterior probability* of H given E,

(ii) $P(E|H)$ is the *likelihood* of E given H,

(iii) $P(H)$ is the *prior probability* of H, and

(iv) $P(E)$ is the initial *probability of evidence* E.

(1) For Bayes' rule in (8) to represent an instantiation of the conditional probability formula (4), the events H and E are required to be:

(a) defined on the same event space $\Im$,

(b) potentially *observable,* and

(c) related in the sense that the events $H \cap E$, $H \cup E$, H^c, E^c belong to $\Im$, where H^c denotes the complement of H with respect to S).

Conditions (a)-(c) are potentially problematic for Bayes' rule since E is, in principle, observable and lies in the real world, but H is usually *unobservable* and belongs to the world of mathematics. Ignoring the gap between these two worlds by assuming they are interrelated via $\Im$ in the above simplistic way raises fundational issues in empirical modeling.

(2) It is not obvious how the likelihood function $P(E|H)$ assigns a probability to E by conditioning on an *unobservable* event H. How does one 'condition on the occurrence of an unobservable event H' without running into an oxymoron? A generous possible interpretation might be that the conditioning is only *notional* in the sense that the hypothesis H relates to a particular instance of the mechanism that gave

rise to E. A generous interpretation might be that $P(E|H)$ refers to the 'objective probability of the occurrence of E presuming that H is true'. In practice, however, 'presuming that H is true' could not represent the occurrence of an unobservable event as such. It could, however, be interpreted as an instantiation of a chance set-up pertaining to the mechanism giving rise to events like E.

(3) Despite the bold move of merging hypotheses (mathematical world) and evidence (real world) into the same S, when it comes to acknowledging this overlap $(E\cap H)$, Bayesians sidestep the issue by using the identity $P(E\cap H)=P(E|H)\cdot P(H)$ in:

$$P(H|E) = \tfrac{P(E\cap H)}{P(E)},\ P(E) > 0. \tag{23}$$

It is presumed that the assignments $P(E|H)$ and $P(H)$ are easier to justify!

(4) The most problematic of the probabilistic assignments (i)-(iv) is $P(E)$ because it's not obvious where the probability could come from; see Earman (1992), p. 172. The Bayesians attempt to address this conundrum by defining (iv) in terms of (ii)-(iii) which they deem less questionable. In particular, they use H and $\overline{H}$ denoted by ($\overline{H}$, the complement of H with respect to S), to define a *partition* of S: $S=H\cup\overline{H}$, and then use:

$$P(H\cup\overline{H})=P(H)+P(\overline{H})=P(S)=1,$$

to deduce the *total probability formula*:

$$P(E)=P(H)\cdot P(E|H) + P(\overline{H})\cdot P(E|\overline{H}). \tag{24}$$

The formula in (24) is used to rewrite *Bayes' rule* as:

$$P(H|E)=\frac{P(E|H)\cdot P(H)}{P(H)\cdot P(E|H)+P(\overline{H})\cdot P(E|\overline{H})},\ P(E) > 0. \tag{25}$$

The so-called *Bayesian catchall factor* $P(E|\overline{H})$ in (25) has been severely criticized by Mayo (1996), pp. 116-118, as highly misleading in practice.

6.1 Appendix C: Too many gratuitous symbols?

A common complaint by students who are learning probability theory and statistics is that the Spanos perspective uses too many concepts accompanied by gratuitous symbols without any apparent reason; no other textbooks seem to need so many. For instance, why define the concept of a generic statistical model using a joint distribution $f(\mathbf{x};\boldsymbol{\theta}),\ \mathbf{x}\in\mathbb{R}^n_X$, by:

$$\mathcal{M}_{\boldsymbol{\theta}}(\mathbf{x})=\{f(\mathbf{x};\boldsymbol{\theta}),\ \boldsymbol{\theta}\in\Theta\},\ \mathbf{x}\in\mathbb{R}^n_X,\ \Theta\subset\mathbb{R}^m,\ m<n.$$

Elementary and intermediate books ignore $\mathcal{M}_{\boldsymbol{\theta}}(\mathbf{x})$ altogether, and some advanced textbooks use $\{P_{\boldsymbol{\theta}},\ \boldsymbol{\theta}\in\Theta\}$, without the additional details.

▶ The reason is that to be able to talk about *model validation* (testing the validity of the model assumptions) one needs a complete list of testable probabilistic assumptions comprising the statistical model in question; $f(\mathbf{x};\boldsymbol{\theta}),\ \mathbf{x}\in\mathbb{R}^n_X$, encapsulates all

that! Using marginal distributions will not do because they cannot account for any dependence or/and heterogeneity; endemic in economic data. Worse, the overwhelming majority of the concepts, notation and terminology used in statistics textbooks makes sense only for simple statistical models based on random (IID) samples. Think of the notion of a sample of representative data from a target population! This is a highly misleading and misplaced metaphor for empirical modeling in general. What renders data $\mathbf{x}_0{:=}(x_1,x_2,...,x_n)$ amenable to statistical modeling and inference is not whether they can be viewed as a random sample from a target population, but their exhibiting chance regularity patterns. Hence, the need for viewing a statistical model as a stochastic generating mechanism that could have given rise to data $\mathbf{x}_0$, and define the alternative mechanisms one is comparing $\mathcal{M}_{\boldsymbol{\theta}}(\mathbf{x})$ with.

Does one need the sample ($\mathbb{R}^n_X$) and parameter (Θ) space notation?

► Yes! $\mathbb{R}^n_X$ and Θ occupy center stage in the context of statistical inference, and their mathematical structure plays a crucial role in learning from data.

Focusing on a particular example, the **simple Normal**:

$$\mathcal{M}_{\boldsymbol{\theta}}(\mathbf{x})\text{: } X_k\backsim\mathsf{NIID}(\mu,\sigma^2),\ x_k{\in}\mathbb{R},\ \mu{\in}\mathbb{R},\ \sigma^2{>}0,\ k{\in}\mathbb{N}{:=}(1,2,...,n,...),$$

why would anybody need the notation $k{\in}\mathbb{N}{:=}(1,2,...,n,...)$, and not just write:

$$k{=}1,2,...,n,...?$$

► The reason is that the latter is too narrow and misleading for many statistical models in practice because the mathematical structure of the index set $\mathbb{N}$ (e.g. scale of measurement: ratio, interval, ordinal, nominal) plays a crucial role in statistical modeling and inference.

Why would anybody need the following derivation?

$$f(x_1,x_2,...,x_n;\boldsymbol{\theta})\ \overset{\text{I}}{=}f_1(x_1;\theta_1){\cdot}f_2(x_2;\theta_2){\cdot}\cdots{\cdot}f_n(x_n;\theta_n){=}\textstyle\prod_{k=1}^n f_k(x_k;\theta_k)\overset{\text{IID}}{=}\prod_{k=1}^n f(x_k;\theta){=}$$
$$\overset{\text{NIID}}{=}\textstyle\prod_{k=1}^n\frac{1}{\sqrt{2\pi\sigma^2}}\exp\{-\frac{(x_k-\mu)^2}{2\sigma^2}\}{=}(\frac{1}{\sqrt{2\pi\sigma^2}})^n\exp\left\{\frac{1}{2\sigma^2}\sum_{k=1}^n(x_k{-}\mu)^2\right\},\ \mathbf{x}{\in}\mathbb{R}^n.$$

► The reason is to bring out the role of the probabilistic assumptions giving rise to the distribution of the sample $f(\mathbf{x};\boldsymbol{\theta})$, and emphasize that it is a *function* of $\mathbf{x}{\in}\mathbb{R}^n_X$. This, in turn, helps one understand the likelihood *function*:

$$L(\boldsymbol{\theta};\mathbf{x}_0)\propto f(\mathbf{x}_0;\boldsymbol{\theta}),\ \forall\boldsymbol{\theta}{\in}\Theta.$$

Defining a function. It is extremely important to emphasize that a numerical function $y{=}h(x)$ is *not* just a bunch of numbers. Indeed, one cannot define a function without specifying the range of values of x and y, $\mathbb{R}_X$ and $\mathbb{R}_Y$. In addition, it matters whether the values for these variables are discrete or continuous. Hence, for clarity and completeness a **function** should be properly written as:

$$\boxed{y{=}h(x),\ \forall x{\in}\mathbb{R}_X,\forall y{\in}\mathbb{R}_Y.}$$

What can go wrong if the range of values of x and y, $\mathbb{R}_X$ and $\mathbb{R}_Y$, are not explicitly specified? Let us see!

Example. *Bayes' rule* in terms of *events* is written as:

$$P(H|E)=\frac{P(E|H)\cdot P(H)}{P(E)},\ P(E)>0, \tag{26}$$

where H denotes a *hypothesis* and E the relevant *evidence*.

When the relevant evidence comes in the form of numerical data $\mathbf{x}_0$, one needs to transform (26) in terms of density functions relating to random variables X and unknown parameters θ; we use a single variable instead of the sample $\mathbf{X}$ to simplify the noation. The traditional Bayesian *recasting* of (26) in terms of random variables X and θ is:

$$f(\theta|x)=\frac{f(x|\theta)\cdot f(\theta)}{f(x)},\ f(x)>0, \tag{27}$$

(see Howson and Urbach, 2006, p. 38, Lindley (1965), p. 118, O'Hagan (1994), p. 4, and Robert (2004), pp. 8-9), where $f(x|\theta)$ is interpreted as the conditional density of X given θ. This is conceptually and formally incorrect for several reasons.

(a) One cannot simply replace H and E with X and θ because a random variable is *not an event* in itself. One can define many different events using a random variable depending on its range of values.

(b) A density function is defined at particular values of each random variable, and thus for the formulae (27) to make probabilistic sense one needs to add the *missing quantifiers* for the relevant values of X and θ.

(c) Although the conditioning in (26) in terms of events is symmetric, the conditional density is *non-symmetric* with respect to the random variables X and θ.

▶ The proper recasting of (26) would be:

$$f(\theta|X=x)=\frac{f(X=x|\theta)\cdot f(\theta)}{f(X=x)},\ \text{for } f(x)>0,\ \forall\theta\in\Theta. \tag{28}$$

This, however, raises several additional problems for Bayesians claiming that: "$f(\theta)$ is the prior density function and $f(x|\theta)$ is the density of X, interpreted as the conditional density of X given θ. The numerator is the joint density of θ and X and the denominator is the marginal density of X." (Ghosh et al. (2006), p. 31)

▶ The conditional density of X given θ is properly defined by:

$$f(x|\theta=\theta_1),\ \forall x\in\mathbb{R}_X, \tag{29}$$

where θ_1 is a particular value of θ in Θ. This is very different from $f(X=x|\theta)$, $\forall\theta\in\Theta$, indicating that (29) cannot be accommodated within (28) because the latter is defined for a particular value of X and all values of θ. Having said that, leaving out the quantifier $\forall\theta\in\not\leq$ hides away the problem!

What is the source of the problem? In the case of two events A and B it is true that:

$$P(B|A)\cdot P(A)=P(A|B)\cdot P(B),$$

but for random variables and their density functions, no such symmetry exists.

What would render the claim " $f(x|\theta)$ is the density of X, interpreted as the conditional density of X given θ" valid? When a second quantifier is added, $\forall x \in \mathbb{R}_X$, that would transform Bayes' formula into:

$$f(\theta|X=x)=\frac{f(X=x|\theta)\cdot f(\theta)}{f(X=x)},\ \text{for}\ f(x)>0,\ \forall x\in\mathbb{R}_X,\ \forall\theta\in\Theta. \tag{30}$$

This, however, would be equivalent to reparameterizing the joint distribution since:

$$f(x,\theta)=f(x|\theta)\cdot f(\theta)=f(\theta|x)\cdot f(x),\ \forall x\in\mathbb{R}_X,\ \forall\theta\in\Theta.$$

Which raises a thorny construction problem for Bayesians because the formula in (30), when expressed in terms of the sample $\mathbf{X}$ and the data $\mathbf{x}_0$, takes the form:

$$\pi(\boldsymbol{\theta}|\mathbf{x})=\frac{f(\mathbf{x}|\boldsymbol{\theta})\cdot\pi(\boldsymbol{\theta})}{\int_{\boldsymbol{\theta}\in\Theta} f(\mathbf{x}|\boldsymbol{\theta})\cdot\pi(\boldsymbol{\theta})\mathbf{d}\boldsymbol{\theta}},\ \forall\boldsymbol{\theta}\in\Theta,\ \forall\mathbf{x}\in\mathbb{R}^n_X. \tag{31}$$

When the additional quantifier $\forall\mathbf{x}\in\mathbb{R}^n_X$ is included, as in (31), all the interpretations in the above quotation are valid since $f(\mathbf{x}|\boldsymbol{\theta})$, $\pi(\boldsymbol{\theta})$ and $f(\mathbf{x})=\int_{\boldsymbol{\theta}\in\Theta} f(\mathbf{x}|\boldsymbol{\theta})\cdot\pi(\boldsymbol{\theta})\mathbf{d}\boldsymbol{\theta}$ are proper densities. However, the presence of the quantifier $\forall\mathbf{x}\in\mathbb{R}^n_X$ in (31) belies the **Likelihood Principle**, which asserts that, for Bayesian inference purposes, the only relevant value of $\mathbf{X}$ is $\mathbf{x}_0$; see Mayo (2018). If we impose that (31) becomes:

$$\pi(\boldsymbol{\theta}|\mathbf{x}_0)=\frac{f(\mathbf{x}_0|\boldsymbol{\theta})\cdot\pi(\boldsymbol{\theta})}{\int_{\boldsymbol{\theta}\in\Theta} f(\mathbf{x}_0|\boldsymbol{\theta})\cdot\pi(\boldsymbol{\theta})\mathbf{d}\boldsymbol{\theta}},\ \forall\boldsymbol{\theta}\in\Theta, \tag{32}$$

and $f(\mathbf{x}_0|\boldsymbol{\theta})$, $\forall\boldsymbol{\theta}\in\Theta$, is no longer "the density of $\mathbf{X}$, interpreted as the conditional density of $\mathbf{X}$ given $\boldsymbol{\theta}$".

What is a statistical model? The truth is that no current or past textbook in statistics explains the concept of a statistical model (the cornerstone of statistical modeling and inference) adequately enough. Worse, you will be hard pressed to find any textbook that provides a complete list of the probabilistic assumptions comprising the different statistical models discussed. The reason is that the first published paper attempting to attempt to give an answer to the question 'what is a statistical model?' is rather recent. McCullagh (2002), in his path-breaking paper criticizes the statistical literature for largely ignoring this key question as well as the related question "What is a parameter?" He questions the widespread practice of specifying statistical models in ad hoc and idiosyncratic ways and ignoring crucial issues such as "statistical meaningfulness" and "parameterizations that have a well-defined meaning". He goes on to propose answers to these fundamental questions and issues using abstract 'category theory'; much too abstract for most statisticians as well as practitioners.

Technical background vs. discerning notation. Statistics require a moderate mathematical background on behalf of the practitioner, but the underlying reasoning is subtle and at times sophisticated. Without a meticulously precise notation the practitioner cannot appreciate the numerous subtle distinctions that permeate statistical modeling and inference, including:

(i) Statistical modeling vs. inference.

(ii) Estimation vs. testing reasoning.

(iii) Statistical vs. substantive information and models.

(iv) Statistical modeling vs. curve-fitting.

(v) Statistical adequacy vs. goodness-of-fit.

(vi) Pre-data vs. post-data error probabilities.

(vii) Testing within vs. testing outside the boundary of a statistical model.

(vii) The distribution of the sample vs. the likelihood function.

(ix) A statistical model specified in terms of a complete, internally consistent and testable probabilistic assumptions.