Financial Econometrics – 2022 Lecture Notes

Rauli Susmel
Dept of Finance
Bauer College of Business
University of Houston

Lecture 1 – Review of Statistics and Linear Algebra (NOT Covered)

This lecture reviews basic probability concepts, from random variables to the Law of Large Numbers and the Central Limit Theory. In the Appendix, the lecture introduces Linear Algebra and its compact notation.

Random Variable

In probability and statistics, a *random variable* (RV), or *stochastic variable*, is described informally as a variable whose values depend on outcomes of an *experiment* (or phenomenon). An experiment is an act or a process with an unknown outcome. For example, the CEO of Apple announces a new product, the effect on the price of Microsoft is unknown, thus, the price of Microsoft is a RV.

Definitions & Notation:
Ω: The sample space –the set of possible outcomes from an experiment.
An event *A* is a set containing outcomes from the sample space.
Σ: The collection of all possible events involving outcomes chosen from Ω. (Formally: Σ is a σ - algebra of subsets of the sample space.)
P is a probability measure over Σ. P assigns a number between [0,1] to each event in Σ.

Remarks:
- A random variable is a convenient way to express the elements of Ω as numbers rather than abstract elements of sets.
- A random variable ***X*** is a function.
- It is a numerical quantity whose value is determined by a random experiment.
- It takes single elements in outcome set Ω, which can be abstract elements, and maps them to points in R.

Example: We compute the weekly sign of stock returns of two unrelated firms: Positive (U: up) or negative (D: down).

The sample space is Ω = {DD; DU; UD; UU}.

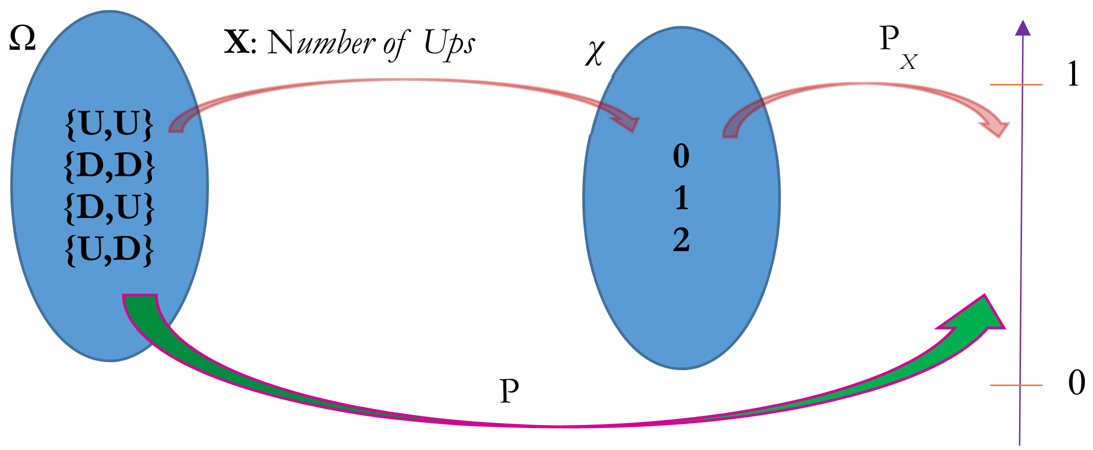
Possible events (*A*):
- Both firms have the same signed return: {U,U} & {D,D}.
- At least one firm has positive returns: {U,U}; {D,U} & {U,D}.
- The first firm is has positive returns: {U,U} & {U,D}

Collection of all possible events: Σ = [Φ, {U,U}, {U,D}, {D,U}, {D,D}, {UU, UD}, {UU, DU}, {UU, DD}, {DD, DU}, {DD, UD}, {DU, DD}, {UU, DU, UD}, {UD, DU, DD}, {UU, UD, DU, DD}]

Define RV: **X** = *"Number of Up cycles."* Recall, **X** takes Ω into χ, & induces P_X from P. Then,

χ = {0; 1; 2}

Σ_χ = {Φ; {0}; {1}; {2}; {0;1}; {0;2}; {1;2}; {0;1;2}}.



Assuming U and D have the same probability, P[U] = P[D] = ½, we define P_X:

Prob. of 0 Ups = P_X[0] = P[{DD}] = ¼
Prob. of 1 Ups = P_X[1] = P[{UD; DU}] = ½
Prob. of 2 Ups = P_X[2] = P[{UU}] = ¼
Prob. of 0 or 1 Ups = P_X[{0; 1}] = P[{DD; UD; DU}] = ¾
Prob. of 0 or 2 Ups = P_X[{0; 2}] = P[{DD; UU}] = ½
Prob. of 1 or 2 Ups = P_X[{1; 2}] = P[{DU; UD; DD}] = ¾
Prob. of 1, 2, or 3 Ups = P_X[{0; 1; 2}] = P[{DD; DU; UD; UU}] = 1
Prob. of "nothing" = P_X[Φ] = P[Φ] = 0

The empty set is simply needed to complete the σ-algebra (a technical point). Its interpretation is not important since P[Φ] = 0 for any reasonable P.

Technical detail: P is the probability measure over the sample space, Ω, and P_X is the probability measure over χ, the range of the random variable.

Example: IBM Returns

We buy a share IBM at USD 120 today and plan to sell the share next week. The return of IBM next week depends on how the market values IBM next week –this is the experiment.

The sample space is continuous, from -100% (worst case scenario) to potentially a huge undefined positive number. We set $\Omega = \{r_t : r \varepsilon [-1, K], K>0\}$.

Possible events:
- IBM returns are positive.
- IBM returns are higher than the 0.5%.
- IBM returns are lower than 10%.

The collection of all possible events, Σ, is very, very large. We use a probability distribution, for example, the normal distribution, to describe the likelihood of possible events.

Probability Function & CDF

Definition – The *probability function*, $p(x)$, of a RV, X.
For any random variable, X, and any real number, x, we define

$$p(x) = P[\, X = x\,] = P[\, \{X = x\}\,],$$

where $\{X = x\}$ = the set of all outcomes (event) with $X = x$.

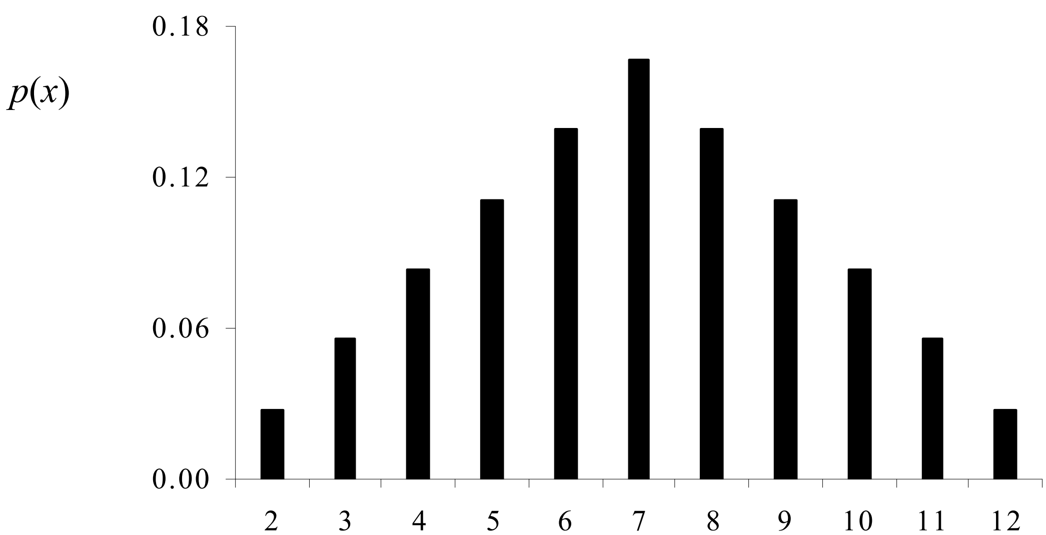
Definition – The *cumulative distribution function* (CDF), $F(x)$, of a RV, X.
For any random variable, X, and any real number, x, we define

$$F(x) = P[\, X \leq x\,] = P[\, \{X \leq x\}\,],$$

where $\{X \leq x\}$ = the set of all outcomes (event) with $X \leq x$.

Example: Two dice are rolled and X is the sum of the two upward faces. Sample space S = { 2:(1,1), 3:(1,2; 2,1), 4:(1,3; 3,1; 2,2), 5:(1,4; 2,3; 3,2; 4,1), 6, 7, 8, 9, 10, 11, 12}.

Graph: Probability function:



$$p(2) = P[X = 2] = P\left[\{(1,1)\}\right] = \frac{1}{36}$$

$$p(3) = P[X = 3] = P\left[\{(1,2),(2,1)\}\right] = \frac{2}{36}$$

$$p(4) = P[X = 4] = P[\{(1,3),(2,2),(3,1)\}] = \frac{3}{36}$$

$$p(5) = \frac{4}{36},\ p(6) = \frac{5}{36},\ p(7) = \frac{6}{36},\ p(8) = \frac{5}{36},\ p(9) = \frac{4}{36}$$

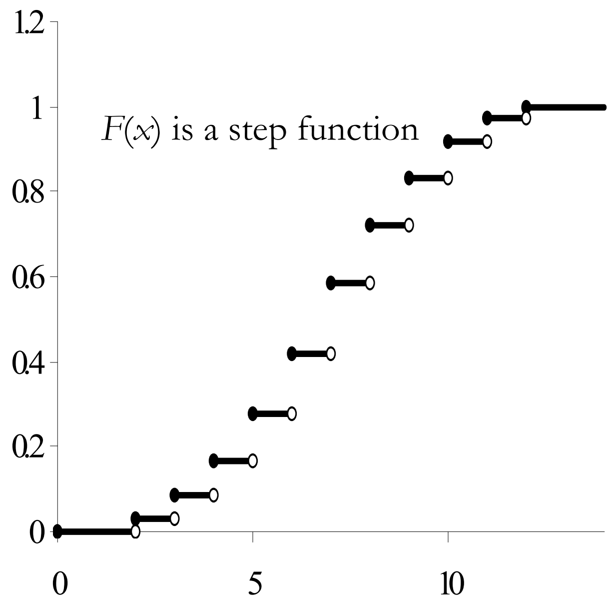
$$p(10) = \frac{3}{36},\ p(11) = \frac{2}{36},\ p(12) = \frac{1}{36}$$

and $p(x) = 0$ for all other x

Note: $\{X = x\} = \phi$ for all other x.

Graph: CDF

$$F(x) = \begin{cases} 0 & x < 2 \\ \frac{1}{36} & 2 \le x < 3 \\ \frac{3}{36} & 3 \le x < 4 \\ \frac{6}{36} & 4 \le x < 5 \\ \frac{10}{36} & 5 \le x < 6 \\ \frac{15}{36} & 6 \le x < 7 \\ \frac{21}{36} & 7 \le x < 8 \\ \frac{26}{36} & 8 \le x < 9 \\ \frac{30}{36} & 9 \le x < 10 \\ \frac{33}{36} & 10 \le x < 11 \\ \frac{35}{36} & 11 \le x < 12 \\ 1 & 12 \le x \end{cases}$$



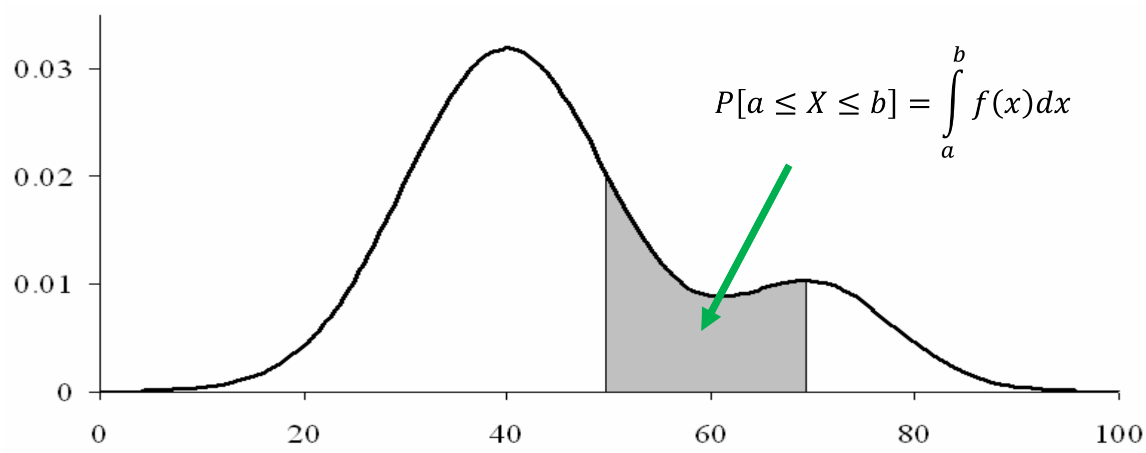
PDF for a Continuous RV

Definition: Suppose that X is a random variable. Let $f(x)$ denote a function defined for $-\infty < x < \infty$ with the following properties:

1. $f(x) \ge 0$
2. $\int_{-\infty}^{\infty} f(x)dx = 1.$
3. $P[a \le X \le b] = \int_a^b f(x)dx$

Then, $f(x)$ is called the *probability density function* (pdf) of X. The random variable X is called *continuous*.

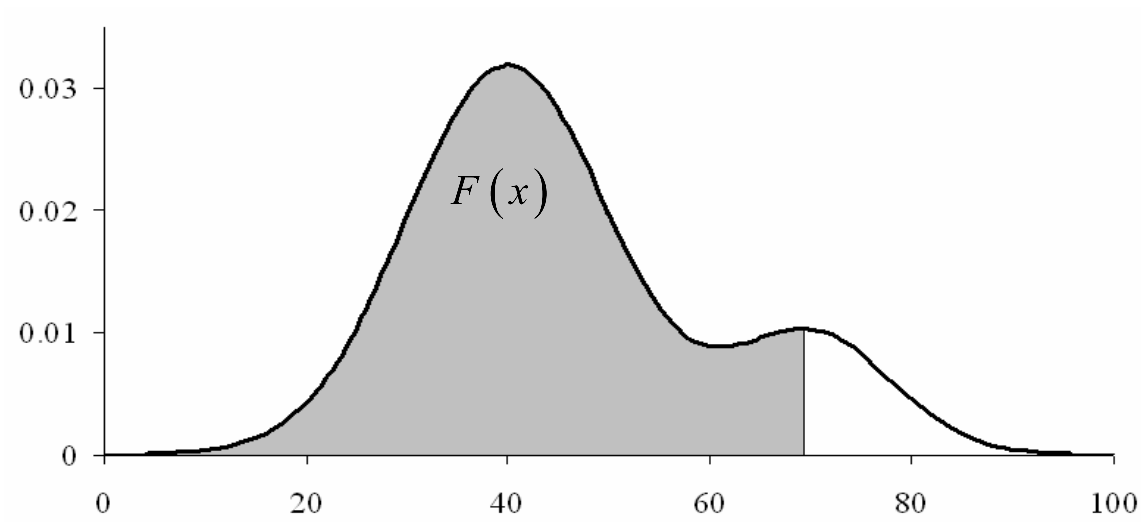
• PDF



• If X is a continuous random variable with probability density function, $f(x)$, the *cumulative distribution function* of X is given by:

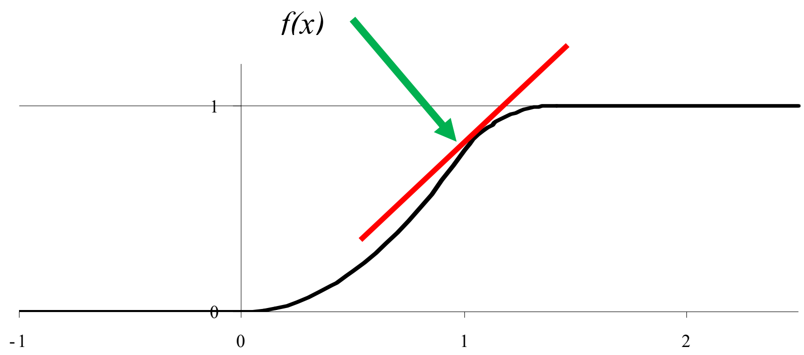
$$F(x) = P[X \leq x] = \int_{-\infty}^{x} f(t)dt$$

• CDF



• Also because of the FTC (*fundamental theorem of calculus*):

$$F'(x) = \frac{dF(x)}{dx} = f(x)$$



PDF for a Discrete RV

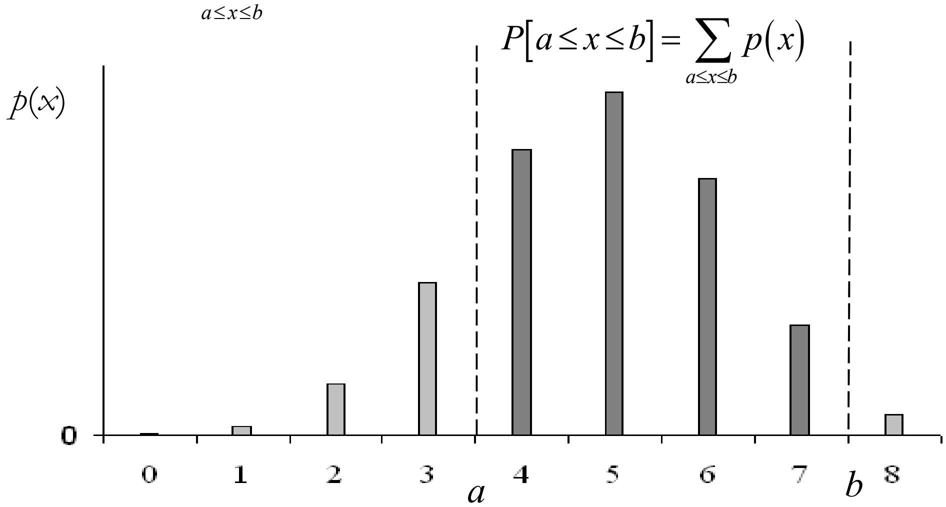
A random variable X is called *discrete* if

$$\sum_{x} p(x) = \sum_{i=1}^{\infty} p(x_i) = 1$$

All the probability is accounted for by values, x, such that $p(x) > 0$.

• For a discrete random variable X the probability distribution is described by the probability function $p(x)$, which has the following properties:

1. $0 \le p(x) \le 1$
2. $\sum_{x} p(x) = \sum_{i=1}^{\infty} p(x_i) = 1$
3. $P[a \le x \le b] = \sum_{a \le x \le b} p(x)$



Bernouille and Binomial Distributions

Suppose that we have a *Bernoulli trial* (an experiment) that has 2 results:

1. Success (S)
2. Failure (F)

Suppose that p is the probability of success (S) and $q = 1 - p$ is the probability of failure (F). Then, the probability distribution with probability function:

$$p(x) = P[X = x] = \begin{cases} q & x = 0 \\ p & x = 1 \end{cases}$$

is called the *Bernoulli distribution.*

• We observe an independent Bernoulli trial (**S, F)** n times. Let X be the number of successes in the n trials. Then, X has a *binomial distribution*:

$$p(x) = P[X = x] = \binom{n}{x} p^x q^{n-x} \quad x = 0, 1, 2, \ldots, n$$

where

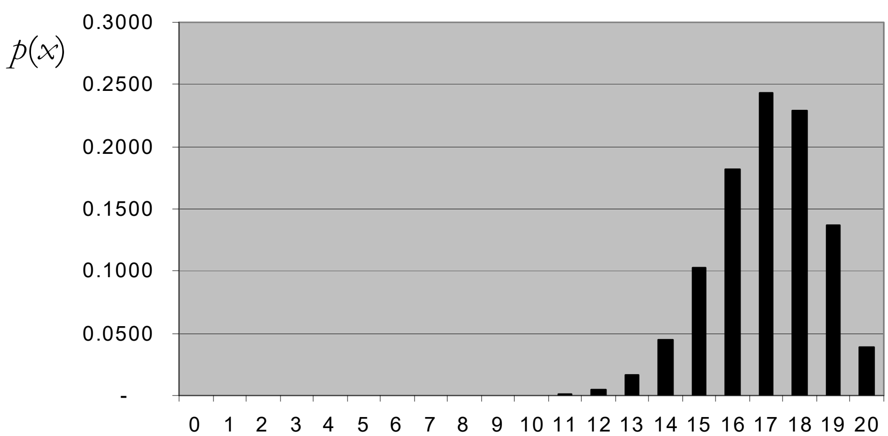
1. p = the probability of success (**S**), and
2. $q = 1 - p$ = the probability of failure (**F**)

Example: If a firm announces profits and they are "surprising," the chance of a stock price, P, increase is 85%. Assume there are n=20 (independent) announcements.
Let X be the number of increases in the stock price following *surprising announcements* in the n = 20 trials.

$$p(x) = P[X = x] = \binom{n}{x} p^x q^{n-x} \quad x = 0, 1, 2, \ldots, n$$

$$= \binom{20}{x} (.85)^x (.15)^{20-x} \quad x = 0, 1, 2, \ldots, 20$$

x	0	1	2	3	4	5
$p(x)$	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
x	6	7	8	9	10	11
$p(x)$	0.0000	0.0000	0.0000	0.0000	0.0002	0.0011
x	12	13	14	15	16	17
$p(x)$	0.0046	0.0160	0.0454	0.1028	0.1821	0.2428
x	18	19	20			
$p(x)$	0.2293	0.1368	0.0388			



The Poisson Distribution

Suppose events are occurring randomly and uniformly in time.

• The events occur with a known average.

• Let X be the number of events occurring (arrivals) in a fixed period of time (time-interval of given length).

• Typical example: X = Number of crime cases coming before a criminal court per year (original Poisson's application in 1838.)

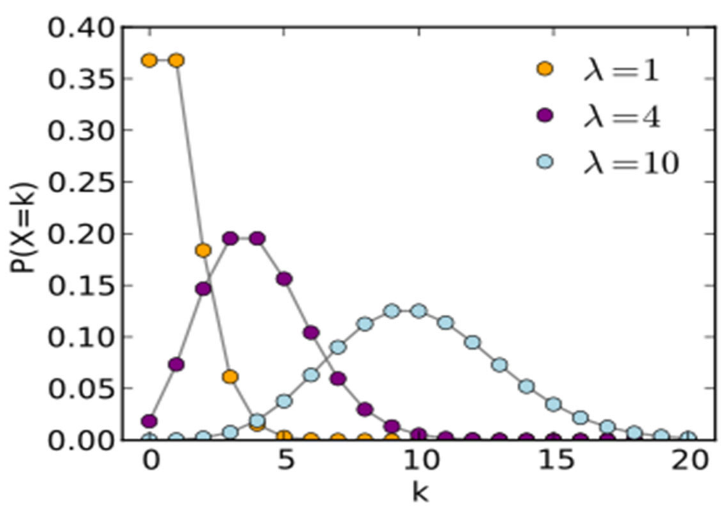
• Then, X will have a *Poisson distribution* with parameter λ:

$$p(x) = \frac{\lambda^x}{x!} e^{-\lambda} \qquad x = 0, 1, 2, 3, 4, \ldots$$

• The parameter λ represents the expected number of occurrences in a fixed period of time. The parameter λ is a positive real number.

Example: On average, a trade occurs every 15 seconds. Suppose trades are independent. We are interested in the probability of observing 10 trades in a minute (X=10). A Poisson distribution can be used with λ = 4 (4 trades per minute).

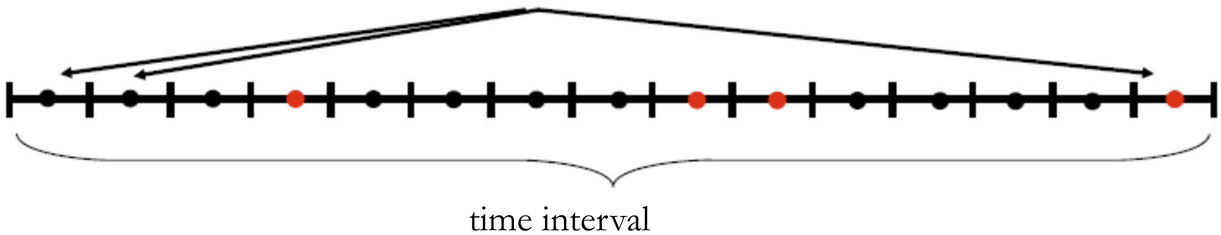
• Poisson probability function



Poisson Distribution: Illustration

Suppose a time interval is divided into n equal parts and that one event may or may not occur in each subinterval.

n subintervals



● - Event occurs
● - Event does not occur

X = # of events is $Bin(n,p)$
As $n \rightarrow \infty$, events can occur over the continuous time interval

X = # of events is $Poisson(\lambda)$

Poisson Distribution: Comments

• The Poisson distribution arises in connection with Poisson processes - a stochastic process in which events occur continuously and independently of one another.

• It occurs most easily for time-events; such as the number of calls passing through a call center per minute, or the number of visitors passing through a turnstile per hour. However, it can apply to any process in which the mean can be shown to be constant.

• It is used in *finance* (number of jumps in an asset price in a given interval); *market microstructure* (number of trades per unit of time in a stock market); *sports economics* (number of goals in sports involving two competing teams); *insurance* (number of a given disaster - volcano eruptions/hurricanes/floods- per year); etc.

Example: The number of named storms over a period of a year in the Caribbean is known to have a Poisson distribution with λ = 13.1
Determine the probability function of X.
Compute the probability that X is at most 8.
Compute the probability that X is at least 10.
Given that at least 10 hurricanes occur, what is the probability that X is at most 15?
Solution:

$$p(x) = \frac{\lambda^x}{x!} e^{-\lambda} \qquad x = 0, 1, 2, 3, 4, \ldots$$

$$= \frac{13.1^x}{x!} e^{-13.1} \qquad x = 0, 1, 2, 3, 4, \ldots$$

x	$p(x)$	x	$p(x)$
0	0.000002	10	0.083887
1	0.000027	11	0.099901
2	0.000175	12	0.109059
3	0.000766	13	0.109898
4	0.002510	14	0.102833
5	0.006575	15	0.089807
6	0.014356	16	0.073530
7	0.026866	17	0.056661
8	0.043994	18	0.041237
9	0.064036	19	0.028432

$$P\left[\text{at most } 8\right] = P\left[X \le 8\right]$$
$$= p(0) + p(1) + \cdots + p(8) = .09527$$

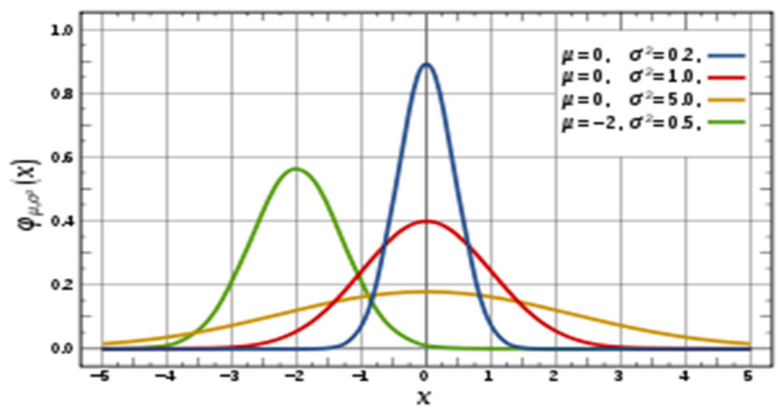
$$P\left[\text{at least } 10\right] = P\left[X \ge 10\right] = 1 - P\left[X \le 9\right]$$
$$= 1 - \left(p(0) + p(1) + \cdots + p(9)\right) = .8400$$

$$P\left[\text{at most } 15 \middle| \text{at least } 10\right] = P\left[X \le 15 \middle| X \ge 10\right]$$
$$= \frac{P\left[\{X \le 15\} \cap \{X \ge 10\}\right]}{P\left[X \ge 10\right]} = \frac{P\left[10 \le X \le 15\right]}{P\left[X \ge 10\right]}$$
$$= \frac{p(10) + p(11) + \cdots + p(15)}{.8400} = 0.708$$

The Normal distribution

A random variable, *X*, is said to have a *normal distribution* with mean *m* and standard deviation *s* if *X* is a continuous random variable with probability density function *f*(*x*):

$$f(x) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left[-\frac{(x-\mu)^2}{2\sigma^2}\right]$$



Normal distribution: Properties

1. Indexed by two parameters: μ *(central parameter)* & σ *(spread parameter).*

2. Symmetric around μ, which is the location of the maximum of *f*(*x*).
Check:

$$f(x) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left[-\frac{(x-\mu)^2}{2\sigma^2}\right]$$

$$f'(x) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left[-\frac{(x-\mu)^2}{2\sigma^2}\right] * \left[-\frac{(x-\mu)}{2\sigma^2}\right] = 0$$

The last equality holds when $\mu = x$. Thus, μ is an extremum point of *f*(*x*). Since *f*(*x*) is a pdf, it is the mode.

3. The inflection points of *f*(*x*) are $\mu - \sigma$, $\mu + \sigma$. (Check: set $f''(x) = 0$ and solve for *x*.)

Normal distribution: Comments

• The normal distribution is often used to describe or approximate any variable that tends to cluster around the mean. It is the most assumed distribution in economics and finance: rates of return, growth rates, IQ scores, observational errors, etc.

• The central limit theorem (CLT) provides a justification for the normality assumption when *n* is large.

Notation: PDF: $x \sim N(\mu,\sigma^2)$

CDF: $\Phi(x)$

The Expectation of X: E(X)

The expectation operator defines the mean (or population average) of a random variable or expression.

Definition

Let X denote a *discrete* RV with probability function $p(x)$ (probability density function $f(x)$ if X is *continuous*) then the expected value of X, $E(X)$ is defined to be:

$$E(X) = \sum_x xp(x) = \sum_i x_i p(x_i)$$

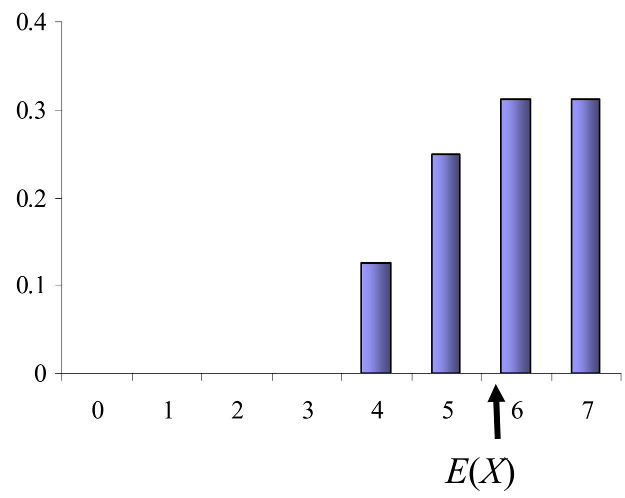
and if X is *continuous* with probability density function $f(x)$

$$E(X) = \int_{-\infty}^{\infty} xf(x)\,dx$$

Sometimes we use E[.] as E_X[.] to indicate that the expectation is being taken over $f_X(x)\,dx$

Interpretation of E(X)

1. The expected value of X, $E(X)$, is the center of gravity of the probability distribution of X.
2. The expected value of X, $E(X)$, is the *long-run average value* of X. (To be discussed later: *Law of Large Numbers*)



E[X]: The Normal Distribution

Suppose X has a Normal distribution with parameters m and s.
Then, E[X] = m.

Proof:

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

$$E(X) = \int_{-\infty}^{\infty} xf(x)\,dx = \int_{-\infty}^{\infty} x \frac{1}{\sqrt{2\pi}\,\sigma} e^{-\frac{(x-\mu)^2}{2\sigma^2}}\,dx$$

Making the substitution:

$$z = \frac{x-\mu}{\sigma} \quad \Rightarrow \quad dz = \frac{1}{\sigma}dx \quad \text{and } x = \mu + z\sigma$$

Then,

$$E(X) = \int_{-\infty}^{\infty} \frac{1}{\sqrt{2\pi}}(\mu + z\sigma)e^{-\frac{z^2}{2}}dz$$
$$= \mu \int_{-\infty}^{\infty} \frac{1}{\sqrt{2\pi}} e^{-\frac{z^2}{2}}dz + \frac{\sigma}{\sqrt{2\pi}} \int_{-\infty}^{\infty} ze^{-\frac{z^2}{2}}dz$$

Using the following results:

$$\int_{-\infty}^{\infty} \frac{1}{\sqrt{2\pi}} e^{-\frac{z^2}{2}}dz = 1 \quad \text{and} \quad \int_{-\infty}^{\infty} ze^{-\frac{z^2}{2}}dz = 0$$

Thus $E(X) = \mu$

Expectation of a function of a RV

Let X denote a *discrete RV* with probability function $p(x)$, then the expected value of $g(X)$, $E[g(X)]$, is defined to be:

$$E[g(X)] = \sum_x g(x)p(x) = \sum_i g(x_i)p(x_i)$$

and if X is *continuous* with probability density function $f(x)$

$$E[g(X)] = \int_{-\infty}^{\infty} g(x)f(x)\,dx$$

Examples: $g(x) = (x - \mu)^2 \Rightarrow E[g(x)] = E[(x - \mu)^2]$
$g(x) = (x - \mu)^k \Rightarrow E[g(x)] = E[(x - \mu)^k]$

Example: Suppose X has a uniform distribution from 0 to b. Then:

$$f(x) = \begin{cases} \frac{1}{b} & 0 \le x \le b \\ 0 & x < 0, x > b \end{cases}$$

Find the *expected value* of $A = X^2$.

If X is the length of a side of a square (chosen at random from 0 to b) then A is the area of the square

$$E(X^2) = \int_{-\infty}^{\infty} x^2 f(x)\,dx = \int_a^b x^2 \tfrac{1}{b-a}\,dx = \left[\tfrac{1}{b}\frac{x^3}{3}\right]_0^b = \frac{b^3 - 0^3}{3(b)} = \frac{b^2}{3}$$

= 1/3, the maximum area of the square

Median: Another central measure

A median is the numeric value separating the higher half of a sample, a population, or a probability distribution, from the lower half.

Definition: Median

The *median* of a random variable ***X*** is the unique number *m* that satisfies the following inequalities:

$P(\boldsymbol{X} \leq m) \geq ½$ and $P(\boldsymbol{X} \geq m) \geq ½$.

For a continuous distribution, we have that *m* solves:

$$\int_{-\infty}^{m} f_X(x)dx = \int_{m}^{\infty} f_X(x)dx = 1/2$$

<u>Note</u>: If the **mean** > **median** > **mode** (= most popular observation), the distribution will be skewed to the right. If the **mean** < **median** < **mode**, the distribution will be skewed to the left.

• Calculation of medians is a popular technique in summary statistics and summarizing statistical data, since it is simple to understand and easy to calculate, while also giving a measure that is more robust in the presence of outlier values than is the mean.

An optimality property

A median is also a central point which minimizes the average of the absolute deviations. That is, a value of *c* that minimizes

$$E(|\mathbf{X} - c|)$$

is the median of the probability distribution of the random variable ***X***.

Example: Let *X* have an exponential distribution with parameter *λ*. The probability density function of *X* is:

$$f(x) = \begin{cases} \lambda e^{-\lambda x} & x \geq 0 \\ 0 & x < 0 \end{cases}$$

The median *m* solves the following integral of *X*:

$$\int_{m}^{\infty} f_X(x)dx = 1/2$$

$$\int_{m}^{\infty} \lambda e^{-\lambda x}dx = \lambda \int_{m}^{\infty} e^{-\lambda x}dx = -e^{-\lambda x}\Big|_{m}^{\infty} = e^{-\lambda m} = 1/2$$

That is, $m = \ln(2)/\lambda$.

Moments of Random Variables

The moments of a random variable *X* are used to describe the behavior of the RV (discrete or continuous).

Definition: K^{th} Moment
Let X be a RV (discrete or continuous), then the k^{th} *moment of X* is:

$$\mu_k = E(X^k) = \begin{cases} \sum_x x^k p(x) & if X \text{ is discrete} \\ \int_{-\infty}^{\infty} x^k f(x)dx & if X \text{ is continuous} \end{cases}$$

Definition: Central Moments
Let X be a RV (discrete or continuous). Then, the k^{th} *central moment* of X is defined to be:

$$\mu_k^0 = E[(X-\mu)^k] = \begin{cases} \sum_x (x-\mu)^k p(x) & \text{if } X \text{ is discrete} \\ \int_{-\infty}^{\infty} (x-\mu)^k f(x)dx & \text{if } X \text{ is continuous} \end{cases}$$

where $m = m_1 = E(X)$ = the first moment of X.

• The central moments describe how the probability distribution is distributed about the center of gravity, m.

• The first central moments is given by:

$$\mu_1^0 = E[X-\mu]$$

• The second central moment depends on the *spread* of the probability distribution of X about m. It is called the variance of X and is denoted by the symbol σ^2 = var(X):

$$\mu_2^0 = E[(X-\mu)^2] = \text{var(X)} = \sigma^2$$

The square root of var(X) is called the *standard deviation* of X and is denoted by the symbol s = SD(X). We also refer to it as *volatility*:

$$\sqrt{\mu_2^0} = \sqrt{E[(X-\mu)^2]} = \sigma$$

Moments of a RV: Skewness

The third central moment:

$$\mu_3^0 = E[(X-\mu)^3]$$

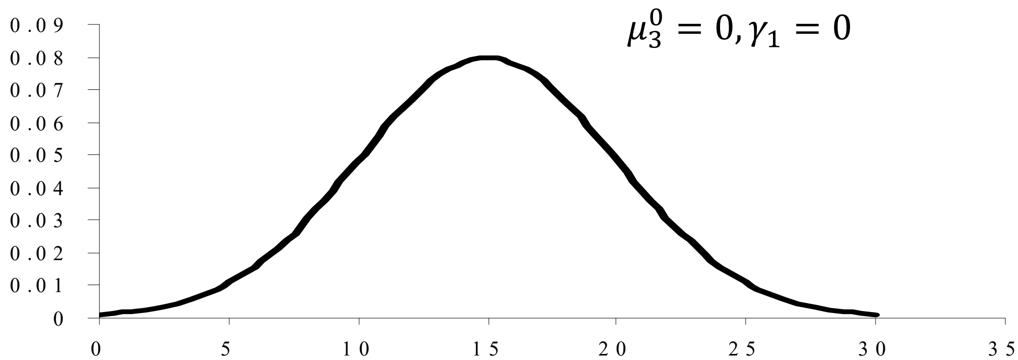
μ_3^0 contains information about the *skewness* of a distribution.

• A popular measure of skewness:

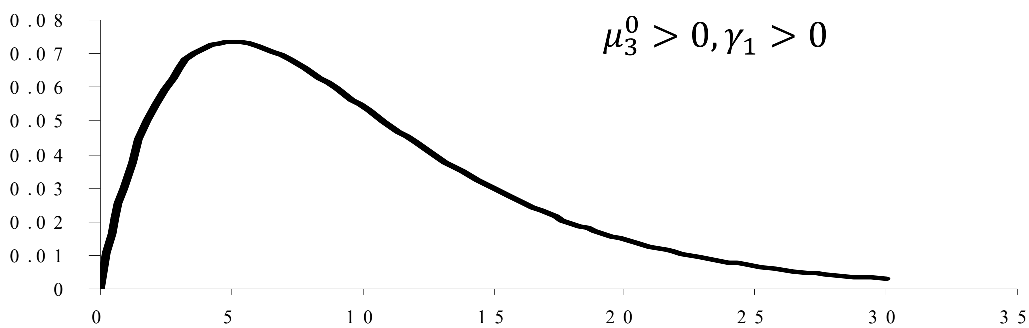
$$\gamma_1 = \frac{\mu_3^0}{\sigma^3} = \frac{\mu_3^0}{(\mu_2^0)^{\frac{3}{2}}}$$

• Distribution according to skewness:

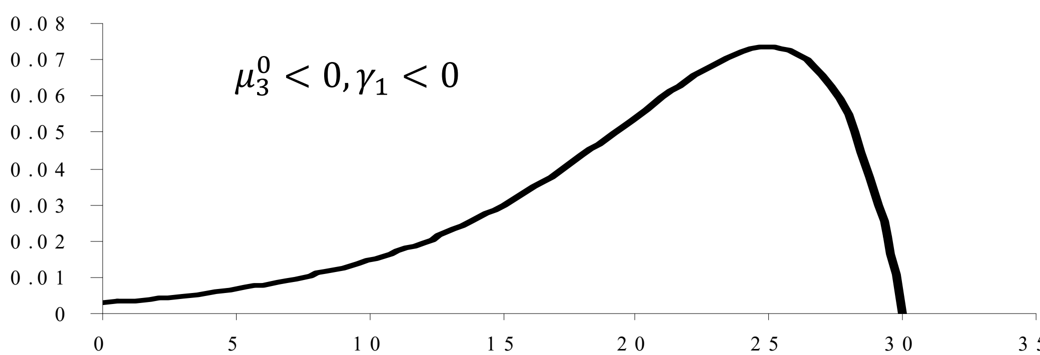
1) Symmetric distribution



2) Positively (right-) skewed distribution (with mode < median < mean)



3) Negatively (left-) skewed distribution (with mode > median > mean)



• Skewness and Economics
- Zero skew means symmetrical gains and losses.
- Positive skew suggests many small losses and few rich returns.
- Negative skew indicates lots of minor wins offset by rare major losses.

• In financial markets, stock returns at the firm level show positive skewness, but at the aggregate (index) level show negative skewness.

• From horse race betting and from U.S. state lotteries there is evidence supporting the contention that gamblers are not necessarily risk-lovers but skewness-lovers: Long shots are overbet (positive skewness loved!).

Moments of a RV: Kurtosis

The fourth central moment:

$$\mu_4^0 = E[(X-\mu)^4]$$

It contains information about the *shape* of a distribution. The property of shape that is measured by this moment is called *kurtosis,* usually estimated by

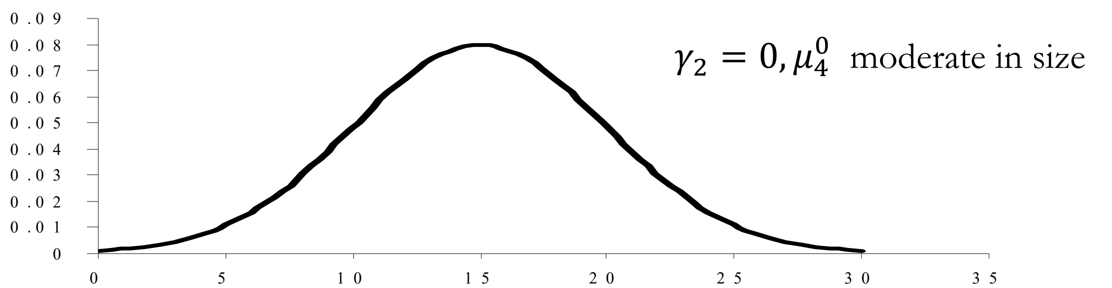
$$\gamma_2 = \frac{\mu_4^0}{\sigma^4}.$$

• The *measure of (excess) kurtosis*:

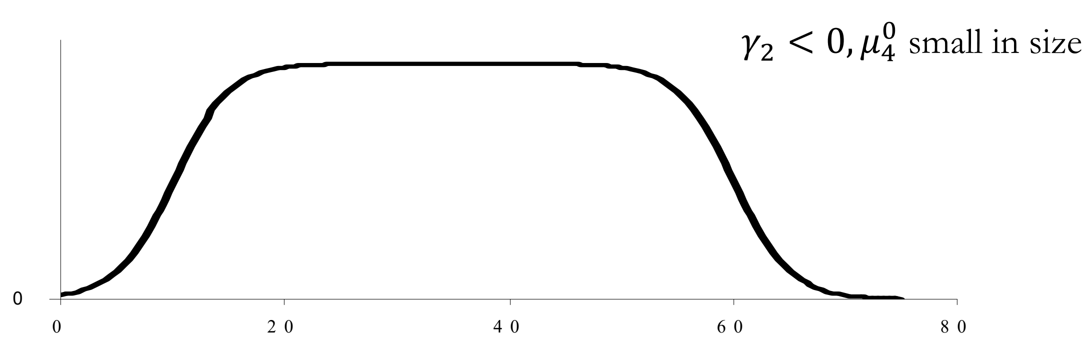
$$\gamma_2 = \frac{\mu_4^0}{\sigma^4} - 3 = \frac{\mu_4^0}{(\mu_2^0)^2} - 3$$

• Distributions:

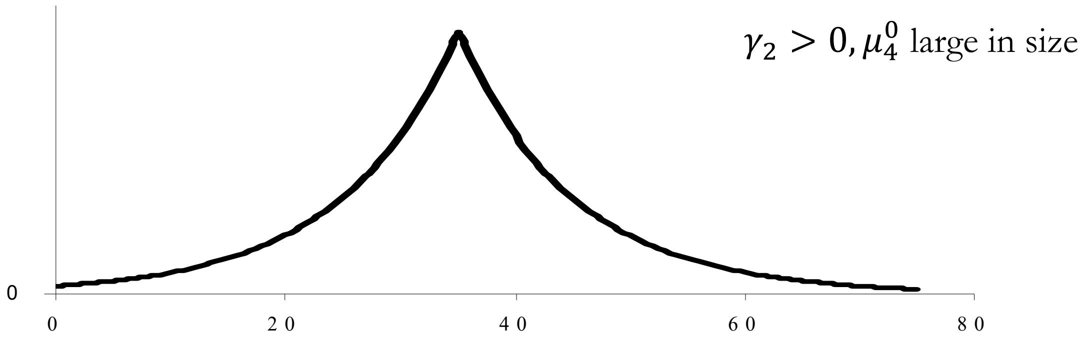
1) Mesokurtic distribution



2) Platykurtic distribution



3) Leptokurtic distribution (usual shape for asset returns)



Moments and Expected Values

Note that moments are defined by expected values. We define the expected value of a function of a continuous RV X, *g*(*X*), as

$$E[g(X)] = \int_{-\infty}^{\infty} g(x)f(x)dx$$

• If *X* is *discrete* with probability function *p*(*x*)

$$E[g(X)] = \sum_x g(x)p(x) = \sum_i g(x_i)p(x_i)$$

Examples: $g(x) = (x - \mu)^2 \Rightarrow E[g(x)] = E[(x - \mu)^2]$

$g(x) = (x - \mu)^k \Rightarrow E[g(x)] = E[(x - \mu)^k]$

• We estimate expected values with sample averages. The Law of Large Numbers (LLN) tells us they are *consistent* estimators of expected values.

Estimating Moments

We estimate expected values with sample averages. For example, the first moment, the mean, and the second central moment, the variance, are estimated by:

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

$$s^2 = \frac{\sum_{i=1}^{N}(X_i - \bar{X})^2}{N-1}$$ (*N*-1 adjustment needed for $E[s^2] = \sigma^2$)

• Besides consistent, they are both are *unbiased* estimators of their respective population moments (unbiased = "on average, I get the population parameter"). That is,

$E[\bar{X}] = \mu$ "population parameter"

$E[s^2] = \sigma^2$

The Law of Large Numbers (LLN)

Long history: Gerolamo Cardano (1501-1576) stated it without proof. Jacob Bernoulli published a rigorous proof in 1713.

Theorem (Weak LLN)

Let $X_1, \ldots, X_N$ be n mutually independent random variables each having mean μ and a finite variance σ^2 -i.e, the sequence $\{x_N\}$ is *i.i.d.*

Let $\bar{X} = \frac{\sum_{i=1}^{N} X_i}{N}$.

Then, for any $\delta > 0$ (no matter how small)

$$P[|\bar{X} - \mu| < \delta] = P[\mu - \delta < \bar{X} < \mu + \delta] \rightarrow 1, \qquad \text{as } N \rightarrow \infty$$

• There are many variations of the LLN. It is a general result: A sample average as the sample size goes to infinite tends to its expected value. Also written as:

$$\bar{X}_N \xrightarrow{p} \mu. \qquad \text{(convergence in probability)}$$

The Central Limit Theorem (CLT)

The Central Limit Theorem (CLT) states conditions for the sequence of RV $\{x_N\}$ under which the mean or a sum of a sufficiently large number of x_i*'s* will be approximately normally distributed.

Let $X_1, X_2, \ldots, X_N$ be a sequence of *i.i.d.* RVs with finite mean μ, and finite variance σ^2. Then, as N increases, $\bar{X}_N$, the sample mean, approaches the normal distribution with mean μ and variance σ^2/N.

This theorem is sometimes stated as:

$$\frac{\sqrt{N}(\bar{X} - \mu)}{\sigma} \xrightarrow{d} N(0,1)$$

where $\xrightarrow{d}$ means "the limiting distribution (asymptotic distribution) is" (or *convergence in distribution*).

• Many version of the CLT. Two versions are commonly used in economics and finance:

- The one above is the *Lindeberg-Lévy CLT*, with $\{x_N\}$ are *i.i.d.*, with finite μ and finite σ^2.
- The other one is the *Lindeberg-Feller CLT*. It requires $\{x_N\}$ are independent, with finite μ_i, $\sigma_i^2 < \infty$, $S_n = \Sigma_i x_i$, $s_n^2 = \Sigma_i \sigma_i^2$ and for ε>0,

$$\lim_{n \to \infty} \frac{1}{s_n^2} \sum_{i=1}^{n} \int_{|x_i - \mu_i| > \varepsilon S_n} (x_i - \mu_i)^2 f(x_i) dx = 0$$

Note:

Lindeberg-Levy assumes random sampling –observations are *i.i.d.*, with the same mean and same variance.

Lindeberg-Feller allows for heterogeneity in the drawing of the observations --through different variances. The cost of this more general case: More assumptions about how the $\{x_N\}$ vary.

• The CLT gives only an asymptotic distribution. We usually take it as an approximation for a finite number of observations. In these cases, the notation goes from $\xrightarrow{d}$ to $\xrightarrow{a}$.

Technical Note: The *Berry–Esseen theorem* (*Berry–Esseen inequality*) attempts to quantify the rate at which the convergence to normality takes place.

$$| F_n(x) - \Phi(x) | \leq \frac{C\rho}{\sigma^3 n^{1/2}}$$

where $\rho = E(|X|) < \infty$ and C is a constant (best current C=0.7056).

Asymptotic Distribution

An asymptotic distribution is a hypothetical distribution that is the *limiting* distribution of a sequence of distributions.

We will use the asymptotic distribution as a finite sample *approximation* to the true distribution of a RV when *N* -i.e., the sample size- is *large*.

Practical question: When is *N* large?

Sampling Distributions

All statistics, T(*X*), are functions of RVs and, thus, they have a distribution. Depending on the sample, we can observe different values for T*(X),* thus, the finite sample distribution of T(*X*) is called the *sampling distribution*.

For the sample mean, $\bar{X}$, if the X_i's are normally distributed, then the sampling distribution is normal with mean μ and variance σ^2/N. Or

$$\bar{X} \sim N(\mu, \sigma^2/N).$$

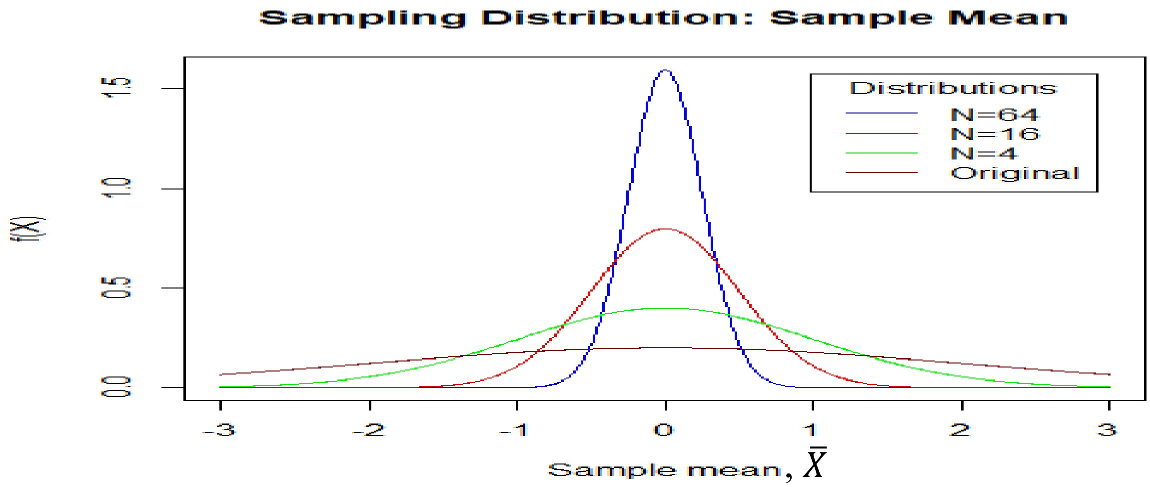
Then, $E[\bar{X}] = \mu$

$Var[\bar{X}] = \sigma^2/N$ ⇒ variance of sample mean decreases as N increases!

The SD of the sampling distribution is called the *standard error* (SE). Then, $SE(\bar{X}) = \sigma/\text{sqrt}(N)$.

We usually associate the standard error with the precision of the estimate. That is, the precision of the estimation of the mean increases as *N* increases.

• Below, we show the sampling distribution for the sample mean of a normal population for different sample sizes (*N)*.



Note: As $N \to \infty$, $\bar{X} \to \mu$ –i.e., the distribution becomes a spike at μ!

Note: If the data is not normal, the CLT is used to approximate the sampling distribution by the asymptotic one, usually, after some manipulations. Again, in those cases, the notation goes from $\xrightarrow{d}$ to $\xrightarrow{a}$.

• For the sample variance σ^2, if the X_i's are normally distributed, then the sampling distribution is derived from this result:

$$(N-1)\, s^2/\sigma^2 \sim \chi^2_{N-1}.$$

It can be shown that a random variable that follows a χ^2_v distribution has a variance equal to 2 times the degrees of freedom (=2*v). Then,

$$\text{Var}[(N\text{-}1)\, s^2/\sigma^2\,] = 2 * (N-1) \qquad \Rightarrow \text{Var}[s^2] = 2 * \sigma^4 /(N-1)$$

Then, $\text{SE}(s^2) = \text{SD}(s^2) = \sigma^2 * \sqrt{2/(N-1)}]$.

Note: If the data is not normal (& N is large), the CLT can be used to approximate the sampling distribution by the asymptotic one:

$$s^2 \xrightarrow{a} \text{N}(\sigma^2, \sigma^4 * (\kappa - 1)/ N)$$

where $\kappa = \frac{\mu_4^0}{\sigma^4}$ (recall when data is normal, $\kappa = 3$).

Remark: The precision of the estimation increases as N increases.

This remark is especially relevant in Finance, where we derive relations between expected returns and risk factors, like market risk or volatility. As we gather more data, expected returns and the volatility of returns will be more precisely estimated.

Hypothesis Testing

A *statistical hypothesis test* is a method of making decisions using experimental data. A result is called *statistically significant* if it is unlikely to have occurred by chance.

• These decisions are made using (null) hypothesis tests. A hypothesis can specify a particular value for a population parameter, say $q=q_0$. Then, the test can be used to answer a question like:

Assuming q_0 is true, what is the probability of observing a value for the (test) statistic used that is at least as big as the value that was actually observed?

• Uses of hypothesis testing:
- Check the validity of theories or models.
- Check if new data can cast doubt on established facts.

• In general, there are two kinds of hypotheses:

(1) About the form of the probability distribution

Example: Is the random variable normally distributed?

(2) About the parameters of a distribution function

Example: Is the mean of a distribution equal to 0?

• The second class is the traditional material of econometrics. We may test whether the effect of income on consumption is greater than one, or whether the size coefficient on a CAPM regression is equal to zero.

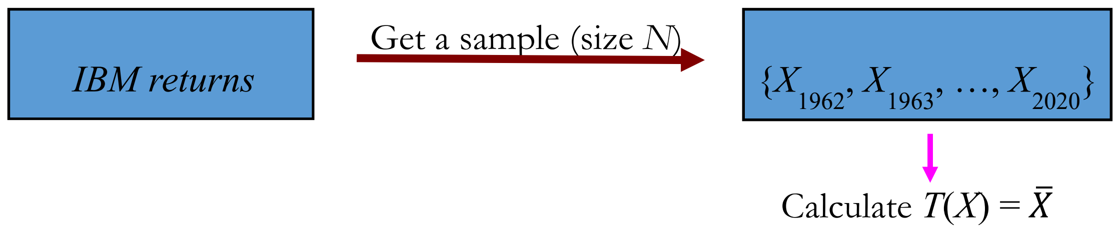
• Hypothesis testing involves the comparison between two competing hypothesis (sometimes, they represent partitions of the world).
- The null hypothesis, denoted H_0, is sometimes referred to as the maintained hypothesis.
- The alternative hypothesis, denoted H_1, is the hypothesis that will be considered if the null hypothesis is "rejected."

Idea: We collect a sample of data $X = \{X_1, \ldots, X_N\}$. We construct a statistic $T(X) = f(X)$, called the *test statistic*. Now we have a decision rule:
- If $T(X)$ is contained in space R, we reject H_0 (& we learn).
- If $T(X)$ is in the complement of R (R^C), we fail to reject H_0.

Note: $T(X)$, like any other statistic, is a RV. It has a distribution.

Example: Suppose we want to test if the mean of IBM annual returns, μ_{IBM}, is 10%. That is, H_0: $\mu_{IBM} = 10\%$.

From the population, we get a sample: $\{X_{1962}, X_{1963}, \ldots, X_{N=2020}\}$, with N=59. We use $T(X) = \bar{X}$, which is unbiased, consistent, and, assuming X is normally distributed, we know its distribution, $\bar{X} \sim N(\mu, \sigma^2/N)$.

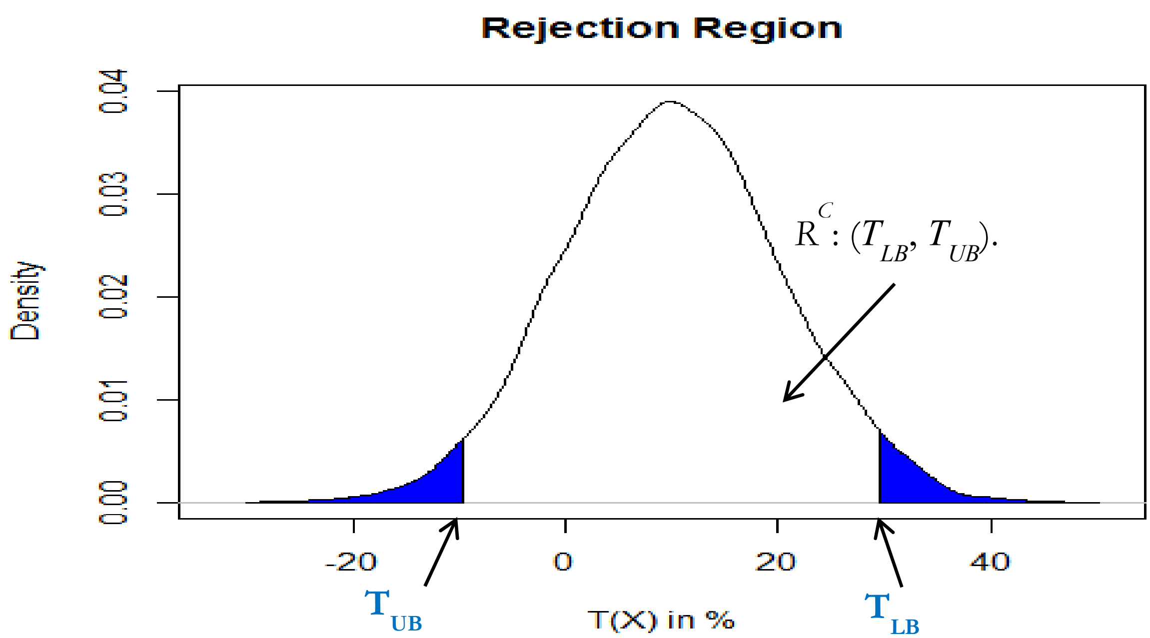


Now, we need to determine the rejection region, *R,* such that if

$$T(X) = \bar{X} \notin [T_{LB}, T_{UB}] \qquad \Rightarrow \text{Reject } H_0\text{: } \mu_{IBM} = 10\%.$$

That is,

$$R = [\bar{X} < T_{LB}, T_{UB} > \bar{X}]$$



Hypothesis Testing: Steps

We present the *classical approach*, a synthesized approach, known as *significance testing*. It relies on Fisher's *p-value*: the probability, of observing a result at least as extreme as the test statistic, under H_0.

We follow these steps:

Step 1. Identify H_0 & decide on a *significance level* (α%) to compare your test results.

Step 2. Determine the appropriate test statistic T(X) and its distribution under the assumption that H_0 is true.

Step 3. Calculate T(X) from the data.

Step 4. Decision Rule:

Reject H_0 if the *p-value* is sufficiently small, that is, we consider $T(X)$ in R (we learn). Otherwise, we reach no conclusion (no learning).

• Q: What *p-value* is "sufficiently small" as to warrant rejection of H_0?

Rule: If *p-value* $< \alpha$ (say, 5%) $\Longrightarrow$ test result is *significant:* Reject H_0.
If the results are "*not significant*," no conclusions are reached (no learning here). Go back gather more data or modify model.

• The father of this approach, Ronald Fisher, favored 5% or 1%.

Example: From the U.S. Jury System
H_0: The defendant is not guilty
H_1: The defendant is guilty

In statistics we learn when we reject. In this case, we learn a defendant is guilty when the jury finds the defendant guilty, by rejecting H_0.

Example: From the U.S. Jury System

Step 1. Identify H_0 & decide on a *significance level* (α%)
H_0: The defendant is not guilty
H_1: The defendant is guilty

Significance level α = "*beyond reasonable doubt*," presumably small level.

Step 2. After judge instructions, each juror forms an "innocent index" $T(X)_i$.

Step 3. Through deliberations, jury reaches a conclusion $T(X) = \sum_{i=1}^{12} T(X)_i$.

Step 4. Decision Rule:
If *p-value* of $T(X) < \alpha \Rightarrow$ Reject H_0. That is, guilty!
If *p-value* of $T(X) > \alpha \Rightarrow$ Fail to reject H_0. That is, non-guilty.
Alternatively, we build a rejection region around H_0.

Note: Mistakes are made. We want to quantify these mistakes.

• Failure to reject H_0 does not necessarily mean that the defendant is not guilty, or rejecting H_0 does not mean necessarily the defendant is guilty. *Type I error* and *Type II error* give us an idea of both mistakes.

Definition: Type I and Type II errors
A *Type I error* is the error of rejecting H_0 when it is true. A *Type II error* is the error of "accepting" H_0 when it is false (that is, when H_1 is true).

Notation: Probability of Type I error: $\alpha = P[X \in R \mid H_0 \text{ is true}]$
Probability of Type II error: $\beta = P[X \in R^C \mid H_1 \text{ is true}]$

	State of World	
Decision	H_0 true	H_1 true (H_0 false)
Cannot reject ("accept") H_0	Correct decision	*Type II error*
Reject H_0	*Type I error*	Correct decision

Need to control both types of error:

α = P[rejecting H_0 | H_0 is true]

β = P[not rejecting H_0 | H_1 is true]

Example: From the U.S. Jury System

Type I error is the error of finding an innocent defendant guilty.

Type II error is the error of finding a guilty defendant not guilty.

• In general, we think *Type I error* is the worst of the two errors, we try to minimize the error of sending to jail an innocent person.

Actually, we would like *Type I error* to be zero. However, the only way to do this (100% of innocent defendants are found not guilty) is to never reject H_0. Then, we maximize *Type II error*.

• There is a clear trade-off between both errors. Traditional view: Set *Type I error* equal to a small number (defined in the U.S. court system as "*beyond reasonable doubt*") and design a test that minimizes *Type II error*.

The usual tests (*t-tests*, *F-tests*, Likelihood Ratio tests) incorporate this traditional view.

Example: We want to test if the mean is equal to μ_0. Then,

1. H_0: $\mu = \mu_0$.

   H_1: $\mu \neq \mu_0$.

2. Appropriate T(X): *t-test* (based on σ unknown and estimated by s).

   Determine distribution of T(X) under H_0. Sampling distribution of $\bar{X}$, under H_0:

$$\bar{X} \sim N(\mu_0, \sigma^2/N).$$

Then, distribution of T(X) under H_0:

$$t = \frac{\bar{X} - \mu_0}{s/\sqrt{N}} \sim t_{N-1} \qquad \text{– when } N > 30,\ t \sim N(0, 1).$$

3. Compute t, $\hat{t}$, using $\bar{X}$, μ_0, s, and N. Get *p-value*($\hat{t}$).

4. <u>Rule</u>: Set an α level. If *p-value*($\hat{t}$) < α $\Rightarrow$ Reject H_0: $\mu = \mu_0$.

   Alternatively, if $|\hat{t}| > t_{N-1,\alpha/2}$ (=**1.96**, if α=.05) $\Rightarrow$ Reject H_0: $\mu = \mu_0$.

Technical Note 1: In step 2, the distribution of the t-test, t, is exact only if {X} follows a normal distribution, otherwise, the distribution is asymptotic (for this we need a large N); that is

$$t = \frac{\bar{X} - \mu_0}{s/\sqrt{N}} \xrightarrow{d} N(0, 1).$$

Technical Note 2: In step 2, we determine the distribution of t, by using the sampling distribution of $\bar{X}$ under H_0. If H_0 is not true, suppose $\mu = \mu_1$, then

$$\bar{X} \sim N(\mu_1, \sigma^2/N),$$

and, thus, t is distributed N(0, 1) only under H_0, since only under H_0 the $E[\bar{X} - \mu_0] = 0$.