Module 5: Introduction to probability

Learning Goal for Module 5

Probability theory underlies statistics, econometrics, and simulation modelling. This Module introduces basic ideas of statistics, presents several probability distributions that find wide use in analytics, and develops the Excel needed to with probability problems.

By the end of this Module, you will:

- Understand and explain basic concepts of probability
- Appreciate the idea of a probability distribution and be able to work with the major discrete (binomial and Poisson) and continuous (normal and lognormal) distributions using Excel
- Understand the concepts of null and alternative hypothesis
- Apply the difference of means tests.

1. Introduction

We are all used to checking the weather each day. We rely on the forecasts to plan our day: what to wear; whether to plan a picnic; and, if you run a restaurant with a patio in the summer, whether to move your reservations indoors. One part of the forecast is the probability of precipitation (POP). A POP equal to 0 means there is no chance of any rain or snow, while a POP of 1 means it is currently raining or snowing.

Wet bias occurs in weather forecasting. Meteorologists tend to adjust POP upward under the assumption that it is "safer" to over-predict rain than to issue low probabilities.

If the probability is wrong, we may close the patio in anticipation of rain, and when it turns out to be sunny, our competitor across the street may get the business. Sometimes a wrong forecast is inconvenient, but other times it can be a matter of life and death, such as during the D-Day invasion of France by the Allies in the Second World War.

Probability is a fundamental idea in economics and business. While ancient philosophers had some ideas about probability, they really did little to develop the ideas. It was not until Italian mathematicians studied gambling games that formal probability ideas appeared.

1.1. Probability problems

Example - Outdoor restaurant

It is summer, and your outdoor patio is busy. However, you know that if it rains, fewer people will come to sit inside, and you will need fewer staff. You pay staff a show-up payment (two hours wages) but need to call them four hours ahead of their shift. It is noon; the weather forecast calls for 40% rain by 6 p.m. Do you call in the extra staff for the evening shift?

Example - Gambling problem

- Assume two gamblers are playing a best-of-five dice game and are interrupted after three games, with one gambler leading two games to one.
- What is the fairest way to split the pot if the game cannot resume?

The solution requires a probability model that predicts the winnings of each player on games 4 and 5 given the winnings to date. Do the winnings to date matter?

Example – Parking problem

Driving to the mall, you want to park as close to the entrance as possible since it is winter. You know as you get closer to the mall entrance, the number of free places will decline. Do you take the next spot or keep going, hoping to find a closer space? This is the "optimal stopping problem."

Example – Vaccine efficacy

Twenty-five thousand adults take part in a randomized, double-blind clinical trial for a new COVID vaccine. Researchers randomly distribute 12,500 participants to the treatment group and 12,500 participants to a control group. Neither the researchers not the participants are aware of the group assignment of any specific participant (double blindness). The treatment group receives the vaccine, and the control, a placebo (fake vaccine that looks exactly like the real vaccine). Monitoring of each group occurs for three months and, at the end, 567 members of the treatment group contract COVID while 1,201 members of the control group become infected. Is the vaccine effective?

1.2. Introductory ideas

Probability is all about the future and managing uncertainty. Any event that has occurred in the past is certain and has a probability of 1, which is not interesting. The probability of any future event, such as the amount of rain to occur tomorrow or whether a head or tail will come up on the next toss of a coin, must be between 0 and 1. The chance of a head from the toss of a fair coin is .5. We can express this as 50/50 or 1 in 2 odds.

Module 5: Introduction to probability

Probability can appear as a decimal, a percentage, or as odds. The odds of a head on a coin toss are 1 in 2. The probability of a "1" on the toss of a die is *1 in 6* or .16666666. The odds of any event is 1 divided by the total of all possible outcomes. The chance of a 1, 3, or 6 on any toss of a die is (1+1+1)/6, or 3 in 6, or 1 in 2. In other words, probability of a 1, 3, or 6 on a single toss of a fair die is .5 or 50%.

Table 1: Outcomes for a single toss of a die

Outcome (Result of the toss)	Probability (PDF)	Cumulative probability (CDF)
1	1/6	1/6
2	1/6	2/6
3	1/6	3/6
4	1/6	4/6
5	1/6	5/6
6	1/6	6/6 = 1
	Sum = 1	

PDF (probability density function) is the probability that x will take on exactly the value x_i or P(=xi).

CDF (cumulative density function) is the probability that x will assume at least a **specific** value x_i, ($P \leq x_i$).

Sample space

Listing all possible outcomes from a "trial" defines the *sample space.* A trial can be as simple as a single toss of a coin, throws of a single die, simultaneous tosses of several coins, or many sequential throws of a die. Complex experiments or simulations with many trials can create large and complex sample spaces.

A trial with two sequential tosses of a coin creates the following sample space – one version is a qualitative sample space and an identical one translated to a numerical sample space.

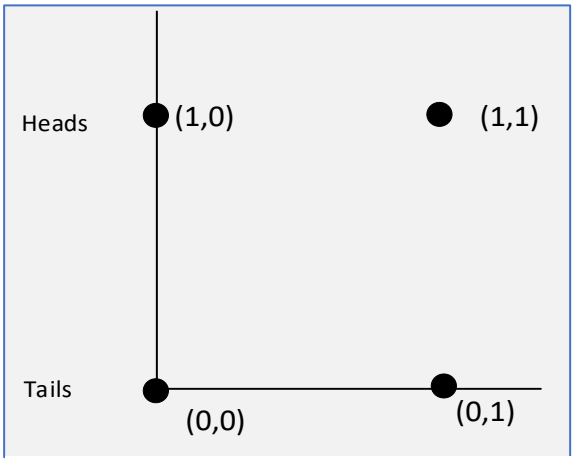


Figure 1: Sample space for sequential coin toss

Event

An event is a subset of a sample space. In a trial with two sequential tosses of a coin, an *event* might be "at least one tail occurs" or "no tails occur." The nature of the problem defines the sample space and any or all possible events. A trial is a specific experiment that produces one or more events. A simultaneous toss of two fair dice is a trial; the outcome "the sum of the two tosses of the dice is exactly 3" is the event, and the sample space appears below. The sample space has 36 possible outcomes, and the event "sum of the two tosses of the dice is exactly 3" has the probability of 2/36 or 1/18. Note that a trial with two sequential tosses of the same coin/dice or two simultaneous tosses of two coins/dice produces the same sample space.

Table 2: Sample space for tossing two dice

		Die 1					
Die 2		1	2	3	4	5	6
	1	1,1	2,1	3,1	4,1	5,1	6,1
	2	1,2	2,2	3,2	4,2	5,2	6,2
	3	1,3	2,3	3,3	4,3	5,3	6,3
	4	1,4	2,4	3,4	4,4	5,4	6,4
	5	1,5	2,5	3,5	4,5	5,5	6,5
	6	1,6	2,6	3,6	4,6	5,6	6,6

1.3. Probability

Three different views define probability.

- The classical or objective view is that one can calculate probabilities just by counting the events in the sample space and using arithmetic and algebra to figure out the probability of any event. For events arising from "two sequential tosses of a coin," the occurrence of "at least one tail" is 3/4. Enumerating all possible outcomes {HH, HT, TH, TT} = 4, of which 3 have at least one tail, supplies the answer. A key assumption in this simple example is that each single outcome is *equally likely*.

  Objective probabilities enumerate all possible outcomes, where natural laws, counting, and logic allow us to calculate the theoretical range of all possible outcomes, and then state that the probability of not getting a 7 or 11 on either of a single toss of a pair of fair dice is 7/9 and the probability of not getting a 7 or 11 on two tosses of the same pair of dice is 49/81. Work these examples out using pencil and paper.

- The frequentist view uses "real world" observations. If one flips a coin many times (thousands), then the probability of a head is just the number of occurrences of a head divided by the total number of tosses. Series of trials, each with 1,000 tosses, might produce results such as 489, 502, 515, 498… heads, and a frequentist would simply take

the average. A frequentist would also notice that results where there are 100, 25, or even 0 heads would be progressively less frequent.

By defining a trial as sequential tosses of two coins, we might expect, after 1,000 trials, that 3/4 (75%) would include a tail, but what if we cannot assume the chances of a head equals those of a tail? The process of creating a coin means that more material may be on one side or the other, biasing ever so slightly the changes of one result over the other. Also, just how large is "many times?" How many tosses do we need to ensure that the percentage of tails in any pairs has stabilized? Finally, imagine your thumb after even just 25 tosses – your ability to toss consistently wanes fast.

The frequentist approach is the only alternative when we have no logical basis for calculating the probability of events. For example, on average, out of 100 babies born, 51 will boys and 49 girls. This ratio is not stable. Variation exists country to country, but the constancy of results over many years and across many countries increases our confidence that this ratio is a fundamental relationship.

- The **subjective view** sees any statement of probability as an expression of opinion. For example, in December 2019, the statement "The next president of the United States will be a woman" might have been a reasonable opinion given that there were at least four women contending for the Democratic nomination. In early March 2020, when Joe Biden surged in the voting and Elizabeth Warren dropped out, the probability that the president of the United States in January 2021 would be a woman dropped to 0 as no women remained in contention.

  Therefore, under a subjective view, probability measures the "degree of belief" arising from judgment. That judgment may derive from logic or observation or a "feeling." Measurement of subjective probabilities appears often in sports betting, where bookies set odds based on history, recent race results, and how individual bettors place their money. A more complex way to measure subjective probability is to imagine a choice where you could win $1,000 tomorrow with a certain probability or could receive x% of that sum today with certainty. The value of "x" needed for you to take the money now reflects your degree of belief that you could win the prize the next day. If the chance of winning the money tomorrow is 90%, would you take $800 or wait till tomorrow?

Each view of probability has its purposes and limitations. The objective view requires a finite sample space that we can count. But no actual mechanical process, being it flipping a coin or tossing a die, can be "fair." Physically, the head and tail on any coin has different amounts of metal, and the depressions on dice are different for a 1 or a 6.

Frequentists rely on measuring repeatable instances under constant conditions. Many natural laws resulted from experiments with many trials. But recording outcomes over an extended

period requires that we know the error in the process that generates the outcome, and to calculate the error we need to know the probability … now the dog is chasing its tail.

Finally, a subjectivist faces the challenge of consistency. When asked the chances of 1 boy in a family of 6 children and the probability of 6 heads in 6 tosses of a coin, they need to respond 1/64. Since subjective probability appears in statements such as "Canada's GDP will increase by more than 3% has odds of 1 in 5," validation occurs when time reveals the truth. A subjectivist must then revise their views in the light of added information. Such revising of subjective probability in the face of new evidence reflects a Bayesian approach, favoured by statisticians.

> ***Example***: The Manitoba Public Insurance Rate Calculator lets you calculate the approximate costs of insurance based on the attributes of the driver. A massive database of accidents supports the probability of accidents and damages based on the attributes of a driver (age, gender, etc.), the attributes of the car (new luxury cars cost more to replace/repair than older, basic models), and other factors. This approach calculates probabilities using the frequentist approach.
>
> ***Example***: The POP (probability of precipitation) on weather websites uses complex meteorological models to develop the probability of rain/snow.

Definitions are important:

- Ex-ante (before the fact). When we calculate frequency distributions objectively or subjectively, we assess variability before events occur. Since the events have not occurred, the final outcomes may not align with our beliefs before the fact --- this is termed prediction risk.

- Ex-post (after the fact). We measure the variability after events have occurred, typically by recording events from large datasets. There is no risk associated with events that have occurred, but we use that information to project forward.

- Risk has two measures:

  1. Estimation of probabilities of events occurring based on formal (mathematical models) (ex-ante)

     Example: Tossing a coin has a 50% chance of a head and 50% chance of a tail.

  2. Estimation of probabilities of events based on past occurrences (ex-post)

     Example: Car insurance rates are based on driving records within a class (age, gender), driving history (accidents, speeding tickets increase rates), or age/cost of the car. This is the frequentist approach just discussed.

- Uncertainty. Strictly this reflects the absence of any basis for assigning a probability to outcomes. Uncertainty occurs when measuring the risks of outcomes is difficult.

Strategies to manage uncertainty include using a range of alternative values for risk and simulating outcomes over many trials.

1.4. Basic rules of probability

Probability follows basic rules. Any logical or mathematical expression that follows these rules can serve as a probability statement. Start by defining an event (such as the result of a coin toss or dice toss). The result from any toss is x_i and may be a head or a tail (or a 1,2,3… in the case of a dice toss). The rules are as follows:

1. The probability of any event must lie between 0 (no chance it can occur) and 1 (complete certainty it will occur). Formally, the rule is $0 \leq PX_i \leq 1$.
2. The sum of the probabilities of all outcomes is 1. The rule appears as $\sum_{i=1}^{n} PX_i = 1$.
3. The probability of a certain event is 1 and the probability of an impossible event is 0.

1.5. Pascal's triangle – An early probability model

Pascal's triangle shows the results of even odds games (tossing a coin). An illustration is the gender ratio, which we will assume to be 1:1, implying that the probability of a boy or a girl resulting from any birth is .5, as shown in Row 2 of Table 3.

Table 3: Pascal's triangle

Row	←Boys Girls→	Row sum	Number of children
1	1	1	0
2	1 1	2	1
3	1 2 1	4	2
4	1 3 3 1	8	3
5	1 4 6 4 1	16	4
6	1 5 10 10 5 1	32	5
7	1 6 15 20 15 6 1	64	6

With 6 children, the chance of 0 boys and 6 girls = 1/64

Chance of 3 boys and 3 girls = 20/64

Chance of 1 boy and 5 girls = 6/64

With two children (Row 3), the probability of all boys is 1/4 and all girls is also 1/4. The probability of a boy and a girl is 1/4 + 1/4 = 2/4, and this appears in Row 3. If a couple has 6 children, go to Row 7, representing 6 children, add the total (64). The chances of all boys or all girls is 1/64. The chance of 3 boys and 3 girls is 20/64.

1.6. Odds

Gambling, such as betting on horse races, uses the equivalent concept of "odds" as opposed to fractions of percentages to describe probabilities. The probability of getting exactly 3 on the toss of two fair dice is 1/18 or, equivalently, odds of 1 in 18, or 18 to 1 against getting the result "exactly 3." In Table 3, the odds of 2 boys (and 1 girl) in a family of three children is 3 in 7, while the odds of 3 boys (or 3 girls) is 1 in 7. Econometric models use odds or, more precisely, *log odds* to measure results.

1.7. Further concepts in probability

Other concepts and rules of probability support the analysis of the outcomes.

- Outcomes in probability are termed ***events.*** The event associated with a coin toss is either a head or a tail (but not both).

- ***Mutually exclusive*** events cannot occur at the same time. A head and a tail cannot occur on the same toss of a coin. Someone cannot be the sister and the daughter of the same person. The probability of two mutually exclusive events is the sum of their probabilities. If we toss a coin and a die at the same time, the probability of a head and a 1 is .5 *.166667 =.08. The outcome of the coin toss is independent of the dice throw. The value .08 states that, on 100 simultaneous tosses of a coin, the combination of a head and a 1 will occur about eight times on average. The results of the first 100 tosses may not be the same as the results of the second 100 tosses but repeat these sets of 100 tosses many times and, on average, the combination of a head and a 1 will occur 8 times out of a hundred.

- When event A changes the probability of event B (or vice versa), A and B are ***dependent*** events.

- When the occurrence of event A has no effect on the probability of event B, then A and B are ***statistically independent***.
  - ***Example***: The events of a head and a tail on the same toss are *mutually exclusive*.
  - ***Example***: The outcomes of two sequential tosses of a coin are *statistically independent. The result of the first toss does not affect the result on the second.*

  Statistically independent events imply that knowing the outcome of one event does not affect your estimate of the probability of the second. Tossing two coins in sequence (one after the other) is an example of statistical independence; the result on the first toss cannot affect your prediction of the outcome of the second toss. Mutually exclusive events are, by definition, statistically independent, but statistically independent events need not be mutually exclusive.

- The probability (occurrence), P, that an event can occur "r" ways out of "n" is r/n, and the probability of non-occurrence is 1-P.

  Example: The probability that the sum of the toss of two dice is less than or equal to 5 is 10/36, and the probability that the sum is more than 5 is 26/36.

  Example: (See Table 4) The probability of drawing an ace from a deck is P1 = 4/52, and (replacing the cards from the first draw) the probability of selecting a 6 is P2 = 4/54. The probability of drawing an ace ***and*** then a 6 is 4/52**+**4/52 = 2/13. These events are *mutually exclusive* (and therefore *independent events*), which allows us to simply add probabilities.

Table 4: Sample space for card draw from a complete deck

A♦	2♦	3♦	4♦	5♦	6♦	7♦	8♦	9♦	10♦	J♦	Q♦	K♦
A♥	2♥	3♥	4♥	5♥	6♥	7♥	8♥	9♥	10♥	J♥	Q♥	K♥
A♣	2♣	3♣	4♣	5♣	6♣	7♣	8♣	9♣	10♣	J♣	Q♣	K♣
A♠	2♠	3♠	4♠	5♠	6♠	7♠	8♠	9♠	10♠	J♠	Q♠	K♠

Example: The probability of drawing an ace on the first draw from a deck is E1 = 4/52, and the probability of drawing a spade is E2 = 13/52 on the second draw. The probability of drawing an ace on the first draw ***or*** a 6 on the second draw is 4/52**x**4/32 = 4/416. These events are statistically independent.

Example: The probability of drawing an ace ***and*** a spade is 1/52, while the probability of drawing an ace ***or*** a spade is 17/52 (Table 5).

Table 5: Sample space for card draw from a complete deck

A♦	2♦	3♦	4♦	5♦	6♦	7♦	8♦	9♦	10♦	J♦	Q♦	K♦
A♥	2♥	3♥	4♥	5♥	6♥	7♥	8♥	9♥	10♥	J♥	Q♥	K♥
A♣	2♣	3♣	4♣	5♣	6♣	7♣	8♣	9♣	10♣	J♣	Q♣	K♣
A♠	2♠	3♠	4♠	5♠	6♠	7♠	8♠	9♠	10♠	J♠	Q♠	K♠

Example: If the probability of rain tomorrow is .5 and the probability that it will rain the day after is .25, then the probability it rains on both days is .125. ***Caution!*** Assuming that the weather on successive days are independent events is a strong belief and wrong.

The idea of "on average" is important. If you toss a coin and it turns up heads, "on average" the next toss should return tails with 50% chance and heads with 50% chance. Tossing a coin 10 times should return heads on five tosses and tails on the other five, but a chance exists for all heads or all tails (.000976 to be precise). The actual pattern of outcomes on experiments with few trials will usually deviate from the pattern that appears from many trials.

This table summarizes the relationship between independent/dependent and exclusive/non-exclusive events.

Table 6: Events

	Independent	Dependent
Exclusive	No (not possible)	Yes (rolling a 2 on a die and a 3 on the same die, but separate toss)
Non-exclusive	Yes (rolling a 2 on a die and a 3 on another die)	Yes (rolling a 2 on a die and an even number on the same die)

- ***Conditional probability*** reflects how two events, E1 and E2, are linked. The probability of E1 given that E2 has occurred is P(E1|E2) or Probability of E1 given E2 has occurred. If P(E1|E2) = 0, then E1 and E2 are independent events.

- If E1E2 denotes E1 and E2 occurring at the same time, then P(E1E2) = P(E1) x P(E1|E2) for mutually exclusive events.

Example. Think about tossing a coin repeatedly. If the chance of heads is 1/2 on any toss, then the chance of a head on the 13^{th} and 17^{th} toss is 1/2 x 1/2 = 1/4 since the tosses are independent events.

1.8. Probability functions

It is possible to describe simple probability operations, such as tossing a coin or throwing a die. In other instances, important probability relationships require algebraic expressions termed distribution functions. In general, the term "probability distribution" covers the general idea of a probability operation; economists typically use discrete and continuous distributions:

- Examples of discrete distributions are the *binomial* (Bernoulli) and *Poisson*.
- Examples of continuous distributions are the *uniform*, *normal*, and *lognormal*. Other important distributions in regression modelling are the t and F.

The ***probability distribution function*** (PDF) shows the probability for an event as a single outcome, typically expressed as the probability of *exactly* a 2 on a single throw of the die. The ***cumulative distribution function*** (CDF) shows the probability for a range of single outcomes, typically expressed as the probability of *at least* a 4 on a single throw of a die. Interpreting the texts describing a problem is fundamental to deciding whether to choose the PDF or CDF. Interpreting the meaning of the text describing a probability problem is also important to deciding which probability distribution to use.

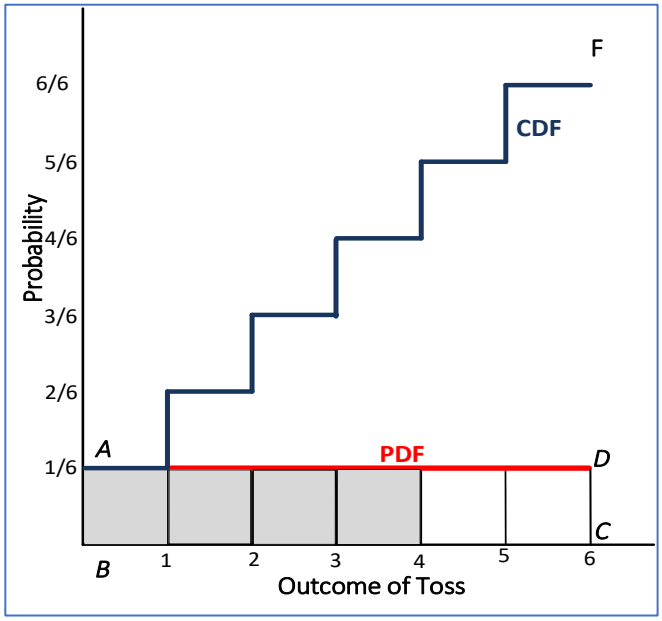


Figure 2: Die toss PDF and CDF

Consider the PDF and CDF for a single toss of a die (Figure *2*). Notice that the total area under the PDF (the rectangle ABCD) has the same value as the CDF at its maximum or the value 6. The probability of a "1" on the face of the die is 1/6. If we wish to calculate the probability of "at least a 4," then the shaded area of the PDF shows this, as does the value of the CDF at 4 (4/6). The area under the PDF for all possible events is 1, as is the value of the CDF when all events are counted .

The CDF always has a value equal to the sum of all possible values of the PDF thereafter. Both the PDF and the CDF uniquely describe the probability process in question. The language of the problem decides which to use.

If the problem is to find the probability when X is a specific value (income equals $90,000), use the PDF. If the problem is to estimate the probability over a range of X (e.g., income lies between $90,000 and $120,000), use the CDF.

Exa*mple*: Find the PDF f(x) and CDF F(x) for X=heads on both of two sequential tosses of a fair coin. There are four outcomes, so we phrase this as "What is the probability of getting at least 1 head in two tosses of a coin?"

Note that the CDF usually appears as F(X) and the PDF as f(x).

$$PDF = f(x) = \begin{cases} 0 & \infty < x \le x_1 \\ f(x_1) & x_1 \le x \le x_2 \\ f(x_1) + f(x_2) & x_2 \le x \le x_3 \\ : & \\ f(x_1) + f(x_2) + \ldots f(x_n) & x_n \le x < \infty \end{cases}$$

Here f(x1) – HH, f(x2) = HT….

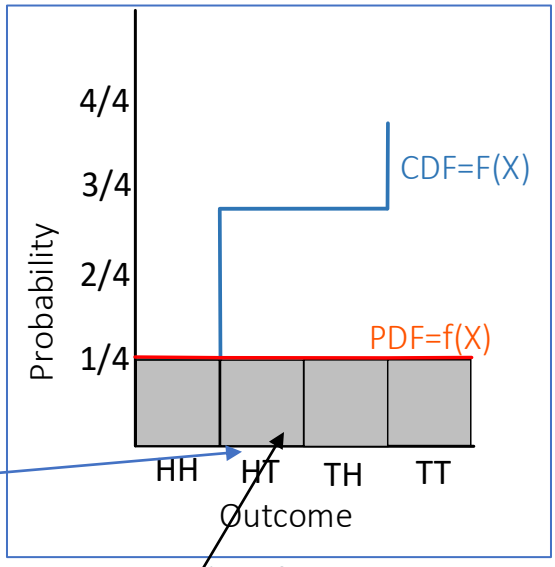


Figure 3: PDF and CDF for two, coin tosses

$$CDF = F(x) \begin{cases} 0 & -\infty < x \le 0 \\ 1/4 & 0 \le x \le 1 \\ 3/4 & 1 \le x \le 2 \\ 1 & 2 \le x \le \infty \end{cases}$$

P(HH) = 1/4, P(HT) =1/4; P(TH)=1/4 and P(TT)=1/4

An important idea is that probability may find expression at a point, f(X), but usually we are interested in the probability over a region, such as the probability of contracting COVID for those aged between 15 and 30.

1.9. Where do distributions come from?

Large datasets often reveal information through summary statistics. Measures of central tendency and variation are the most common summary statistics. Creating a "picture of data" can produce useful insights by classifying or grouping data. Ordinal classes rest on subjective

categories, such as good, better, or best. On surveys, such as course evaluations, the five-point Likert scale may be used – *strongly agree, agree, neither agree nor disagree, disagree, and strongly disagree*. Module 3 covered the mechanics of creating a frequency distribution using the function **=COUNTIF** and the **Data Analysis ToolPak**.

However, most empirical frequency distributions arising from research do not automatically define a probability distribution. A range of statistical tests support the inference that your data follow a specific distribution; often one cannot find a mathematical expression for the distribution.

In the abstract, the PDF shows the probability (a number between 0 and 1) that a specific value of "X_i" occurs. By adding up all the values of $P(X_i)$ we define the CDF. The exact (mathematical) relation between the PDF and the CDF will become plain as we move through four important probability distributions.

Note that the PDF (probability distribution y function) is called the probability mass function.

Video: Uniform Distribution

By choosing parameters carefully, many generic functions can function as probability functions/distributions. Here, by choosing a and b correctly, the shaded area will have the value of 1 and yield a CDF. Only quite special mathematical functions meet the condition to serve as probability distributions.

From **Error! Reference source not found.**, the a rea under the PDF (f(a →a*)) equals the height of the CDF at F(a*). The probability of events up to and including a* (but not higher) is the shaded area under the PDF and the value of the CDF at a* or F(a*).

Some probability specific functions appear from common gambling games and others from natural phenomena.

2. Four important probability distributions

Certain mathematical formulas define a probability distribution and cumulative distribution function. If these functions correspondent to
"natural" phenomena, then algebraic

manipulation of the function may return insights into those phenomena. The probability distributions of most interest include those used in modelling economic and business processes, namely the binomial, the normal, and the lognormal, as well as distributions used in the evaluation of statistical/econometric models, such as the t, F and Chi Square, which we do not study in detail. The simplest distribution is the uniform distribution, which we obtain when using the =RAND() formula. We will use this distribution extensively in modelling.

2.3. Binomial distribution

If p is the probability that an event occurs on any specific trial (probability of success), and the only other possibility is as failure q, defined as q=1-p (probability of a failure), then the probability of X successes in N trials is

$$p(X) = \binom{N}{X} p^X q^{N-X} = \frac{N!}{X!(N-X)!} p^X Q^{N-X}$$

Example: The probability of getting exactly five heads in six tosses of a fair coin is

$$p(X) = \binom{6}{5} \frac{1}{2}^5 \frac{1}{2}^1 = \frac{6*5*4*3*2*1}{5*4*3*2*1} * \frac{1}{32} * \frac{1}{2} = \frac{6}{64}$$

We assume a fair coin in which p(Heads) =.5 = p(Tails) = 1 - p

Aside from pass/fail courses, the binomial distribution occurs anytime events take on stop/go, on/off, life/death situations. Only two outcomes exist

Video: Binomial Distribution

Binomial.xlsx

Properties of the Binomial Distribution	
Mean	μ=Np
Variance	σ^2= Npq
Standard Deviation	σ =$(Npq)^{1/2}$

Example: In 1,000 tosses of an unfair coin (p(Heads) =.4), the mean number of heads is 400, which is the number of heads on average or the expected number of heads. Then the standard deviation = $(1000*.4*.6)^{1/2}$ = (1000*.4*.6)^.5=15.49.

2.4. The normal distribution

Certainly, the workhorse of statistics is normal distribution, developed by Carl Gauss in the early 19th century to explain measurement errors in astronomy. It has the characteristic "bell" shape and found application in modelling human attributes (height, intelligence…). See [].

Standardizing data: We standardize a set of numbers by subtracting the mean from each number.

Standardize.xlsx

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

With standardized data z-(X-μ)/σ, the equation becomes $f(x)=\frac{1}{\sqrt{2\pi}}e^{-1/2x^2}$.

See Normal.xlsx

Video: Normal Distribution

Video: Plotting the Normal Distribution

Example: Assume that elite athletes have careers characterized by the normal – slow start, then acceleration, followed by decline and retirement. If the career of a National Hockey League player is 825 games on average, with a standard deviation of 250: a) What is the probability a newly drafted player (with no games played) will play exactly 700 games; b: What is the probability the player will play fewer than 700 games; c) More than 1,000 games? Use Excel.

Answer: a. Using the =NORMDIST(700,825,250.FALSE) =.001408. b. =NORMDIST(700,825,250,TRUE)=.308538, c) 1 – NORMDIST(1000,825,250,TRUE)=1 - .758036. Note: In part c, you first need to calculate the probability of playing at least 1,000 games. Then we subtract this from 1.

2.5. The lognormal distribution

By simply changing the "x" values in the lognormal distribution, one obtains the lognormal distribution shown below. Note that the normal distribution runs from -∞ to +∞, while the lognormal exists only for positive values of x.

$$f(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(\ln x - \mu)^2}{2\sigma^2}}, x > 0$$

Video: Lognormal Distribution

In economics, the lognormal appears in studies of inequality, the distribution of resources, and the size distribution of firms. The lognormal produces a range of forms depending on the value of the mean and standard deviation (be sure to experiment).

Lognormal.xlsx

2.6. The Poisson distribution

Superficially, the Poisson distribution appears close to the binomial distribution. It too reflects a situation where two states exist – success/failure, pass/fail, etc. A key difference is that the Poisson distribution is open ended, while the binomial specifies the number of successes (or failures) within N trials. The formula for the Poisson distribution appears as

$$f(x,\mu) = \frac{e^{-u}u^{x}}{x!}$$, where u is the average number of successes and x is the actual number derived from an experience or experiment.

Example: The average number of 911 calls on any day is 345. What is the probability that tomorrow will experience 200 calls?

$P(x,u) = (e^{-u})(u^{x})/x! = P(200,345) = (2.71828^{-345})(345^{200}/200!)$, which is a tedious calculation.

Here, note the average number of calls in a day, which sets the value of μ, and plugging in the actual value of calls produces the specific value needed. Fortunately, Excel makes this easy see the following examples.

Poisson.xlsx

Poisson 2.xlsx

Poisson 3.xlsx

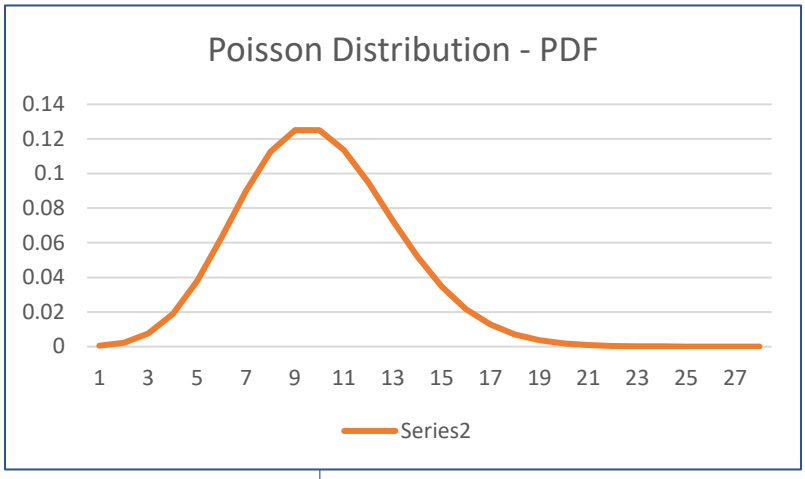


Figure 4: Poisson PDF

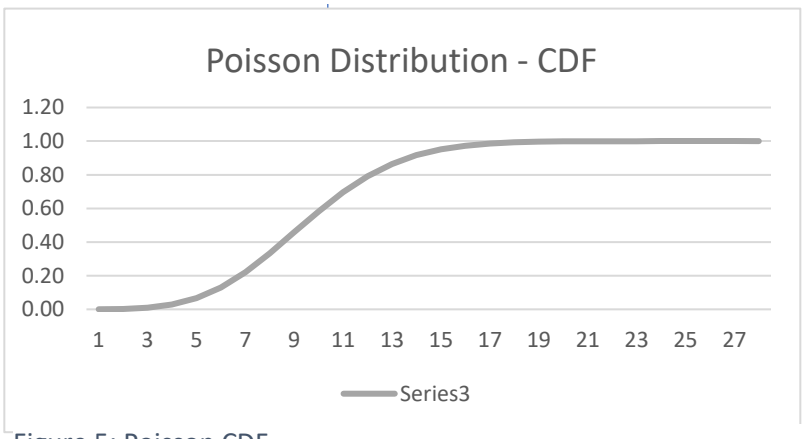


Figure 5: Poisson CDF

The mean of the distribution is 10 (see the spreadsheet), but because the Poisson is not symmetric (unlike the normal), the probability at the mean (see PDF table in spreadsheet) is .12511 (the same as 9), and the value of the CDF is .58304, more than half the area under the PDF.

Note the flat top of the PDF, a reflection of the fact that the Poisson is a discrete distribution, as is the binomial.

e: If a vaccine does not protect 5% of recipients on average, then out of 5,000 doses, what is the probability that exactly 300 will not be protected? What is the probability that fewer than 300 will not have protection? What is the probability that more than 300 will not have protection?

Answer: For the first, we need the value of the PDF when x = 400 and u = 250 (5% of 5,000). In Excel, this is =POISSON(300,250,False) or .00021. For the second, we need the CDF or =POISSON(300,250,True)=.999, and the probability that more than 300 will not have protection is 1 -.999 =.001.

See Section 5 for a summary of the Excel expressions for the probability distributions. Note that all Excel probability functions have a last argument that is either TRUE for the CDF or FALSE for the PDF.

2.7. Detecting the correct distribution

It can be puzzling to decide which distribution one needs for any specific problem. Here are some hints, given that you only have four distributions from which to choose. Note that hundreds of mathematical functions exist to support analytics.

- If you have no information to suggest it is binary (pass/fail, on/off…) or a similar problem, then use the normal. It is the general "go to" probability function.
- With a problem that has just two outcomes, it is either a binomial or a Poisson. A binomial problem will speak of success or failure in a set number of trials (four passes out of seven tests attempted), while a Poisson will define a range, often time (number of successes in the next four days, rejected job applications in the next week…). Note that the Poisson also has application with areas, such as the pattern of rocket strikes around an intended target.
- Certain phenomena tend to skew right, suggesting a lognormal would be right. Natural phenomena and competitive processes in both nature and society often produce "lopsided" results the lognormal can capture.
- Tests for normality and other types of distributions are beyond the scope of this text.

Many probability distributions exist, and economic analytics, especially Monte Carlo methods, use a range of these functions. (See Module 13).

3. Sampling theory

The concept of a sample is familiar, through the many political and attitudinal polls presented in the media. Less commonly understood is that official statistics, such as consumer prices and unemployment rates, depend on sample surveys. In business, surveys of consumers produce strategic information on preferences, price elasticities, and response to marketing.

Intuitively, a sample is a small subset of the entire population. Collecting information from the sample is much less costly than collecting data from every member of the population; when we can get statistics from the sample to infer (stand for) the same attributes in the population, resource savings are considerable. What makes a sample, usually a tiny fraction of the population, a proxy (replacement) for a population?

In a word, random samples of "sufficient" size will allow inference from the attributes of the sample to the same attributes of the population. The term *attribute* stands for the measures of central tendency, variation, and any other measures used to describe a collection of numbers.

When are samples not useful? If we know the probability distribution underlying a phenomenon, taking a sample has no point. For example, the expected pattern of heads and tails for 10 coin tosses is the same as the pattern for 100 tosses or 100 million. However, the political preferences of residents in Hong Kong are unknown except very generally, and a random sample supporting a survey is the best way to infer preference of the population.

In Module 4 we used histograms to create frequency distributions. The key idea is that a population can generate countless different samples. A large university may have 30,000 or more students, which can produce many samples of 1,000 students. Imagine we select 1,000 students for sample 1, then another 1,000 students (without replacement) for sample 2, and so on. If we are interested in grade point averages (GPAs), it is not hard to see that sample 1 and sample 2 will have different average GPAs. But let us take the average GPAs for these two samples. And let us keep taking averages of the ever-expanding numbers of samples. What happens is that, initially, when we have few samples, the average of the average GPAs varies, but as we build the number of samples, each new sample contributes a smaller weight to the growing pool of samples. As the number of samples continues to build, the *average of the sample averages* converges (becomes ever closer) to the average GPA of the population.

Further, if one were to create a histogram of the sample averages, what results is the normal distribution. This powerful idea is the *Central Limit Theorem*.

> Central Limit Theorem: The mean of samples of sufficient size (n>30) from a population, will be close equal to the population mean, and the distribution of the sample means will be approximately normal. As the number of samples increases, the closer the mean of the samples will lie to the population mean.

3.3. Sampling distributions

One of the important (and more difficult ideas to grasp) is the idea of a sampling distribution. Figure 4 below is a visual presentation of sampling from a population with replacement.

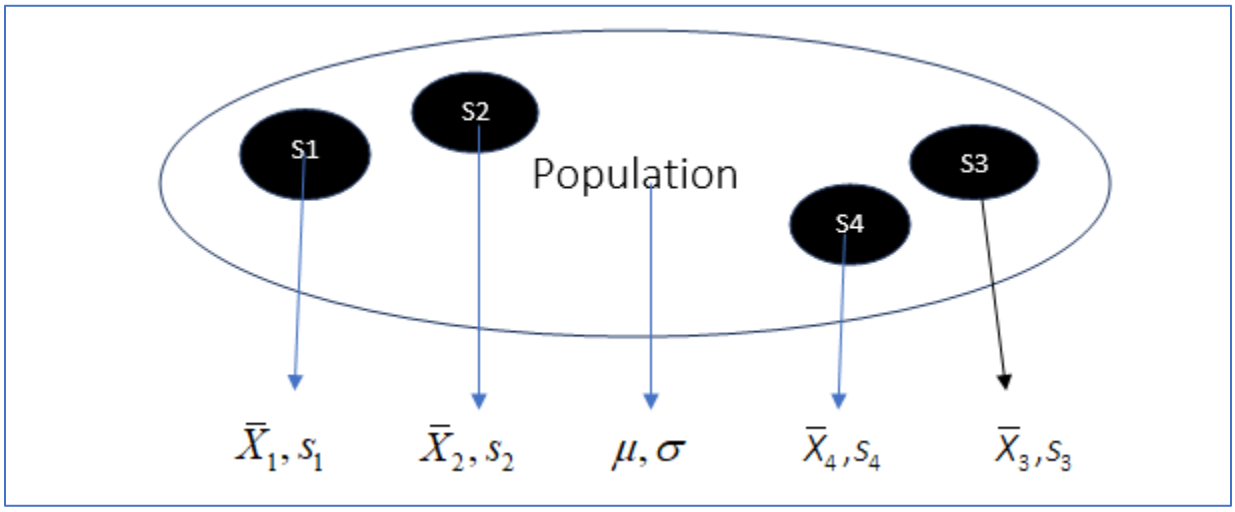


Figure 6: Population and sample means, variances

Each sample $S_1 - S_4$ has a mean $\bar{X}$ and sample standard deviation s that will typically not equal the population mean μ or σ, except by chance. Here is the key idea. If one takes all possible samples from the population, the average of the means of all possible samples will equal the population mean μ. The distribution of all values of $\bar{X}$ is normal. See the example for a demonstration of important ideas behind samples from a population.

Video: Sampling Distribution

Sampling Distribution.xlsx

3.4. Central Limit Theorem

At the heart of classical decision-making, there is the Central Limit Theorem (CLT), which the earlier section and associated example and video strongly suggested. The CLT has four elements:

- Provided the sample size is large enough, the sampling distribution of the sample means is approximately normal (with the approximation getting better as the sample size increases).

- The mean of the sampling distribution of the means equals the population mean, which makes the mean of a sample, selected at random from the population, an *unbiased estimator* of the population mean. This idea is useful because, if we wish to measure the average incomes of the residents of, say, Beijing, we do not need to ask everyone, just a sample of 1,000. Of course, the sample must be random, and everyone included in the survey must respond to the questions.

Video: Central Limit Theorem

- The standard deviation of the sampling distributions of means (AKA the *standard error*) has a special relationship with the standard deviation of the population, namely $s_{\bar{X}}^2 = \frac{N}{n-1}\sigma^2$. In fact, the standard deviation of a sample is a *biased estimator* of the population standard deviation, which is why we need the correction $\frac{N}{N-1}$.

- The population that generates the samples need not be normal. See [].

Central Limit Theorem.xlsx

3.5. Unusual observations

We all experience unusual or out-of-the-ordinary events or people. We develop "rules of thumb" that allow us to take shortcuts in our decision-making. If rules of thumb derive from observation and experience, their use can be efficient. Real world data analytics deals with administrative data that often has mistakes and missing values. Data cleaning may seem like "rigging information" to produce a specific outcome but resolving errors in data will form a critical element of your work as applied economists.

3.5.1. Deciding what is unusual

Any time an event deviates too much from what we consider the norm or average, we wonder whether it is just a single aberration or signals a new trend.

> ***Example***: If we experience a poor meal or bad service at our favourite restaurant, we can wonder whether the regular cook is off for the night or staff did not show up for work. But we usually discount it as a one-off experience. However, repeated bad experiences may lead us to revise our opinions, and it ceases to remain our favourite restaurant.

Outliers vs unusual observations

- The difference often comes down to judgment based on experience and knowledge.
- Our experience informs us as to what might be an unusual but **credible (possible)** result.
- In other instances, we might have repeated experimental or survey data that shows values for the mean, median, or variance for each survey. This encourages us to develop ideas of what the usual heights, incomes, or level sales are for specific groups.

> *Example*: Someone who is 190 cm tall is not unusual in our experience of the 21st century, but three hundred years ago, they would be. Someone who is 210 cm tall would be unusual. However, at 230 cm, they would be well outside the normal range and may be a mistake.

When do we judge something to be out of the norm? When do we decide that an observation is so far out of the norm that we reject the idea that it is a member of the group?

Graphing the data may suggest that certain observations do not belong to the group. This occurs for two reasons:

- Someone made a mistake in entering the data.
- The outlier comes from another population and your data has hidden dimensions.

Which is it? Including or excluding an outlier can affect statistics and other analytics, so deciding whether to include or exclude the outlier is important.

Imagine we recorded sales made by each salesperson within a car dealership. The sales manager may believe that experience matters and salespeople with more experience tend to generate higher revenues. Figure 8 shows the relation between sales and years of experience. A new employee arrives, with modest experience, but who posts impressive sales numbers (Figure 7)? Is this employee someone who is radically different from the existing sales force? Or are their numbers within the norm of experience?

Figure 8: Sales and experience

Figure 7: New employee posts big numbers

How unusual does an observation need to be? As the new employee's "numbers" lie further from the historic averages, the more inclined we are to accept that this individual is unique in terms of natural ability, training, or another attribute. Formally, we say this observation comes from a different population. It may be that the new employee belongs to a particular ethnic group and attracts new clients to the business.

The reason here is not as important as finding a rule or procedures for assessing when we can reject the idea that the new employee is not a member of the population that generated the historic norm. The "distance" that observation lies from the historic norm is an important measure.

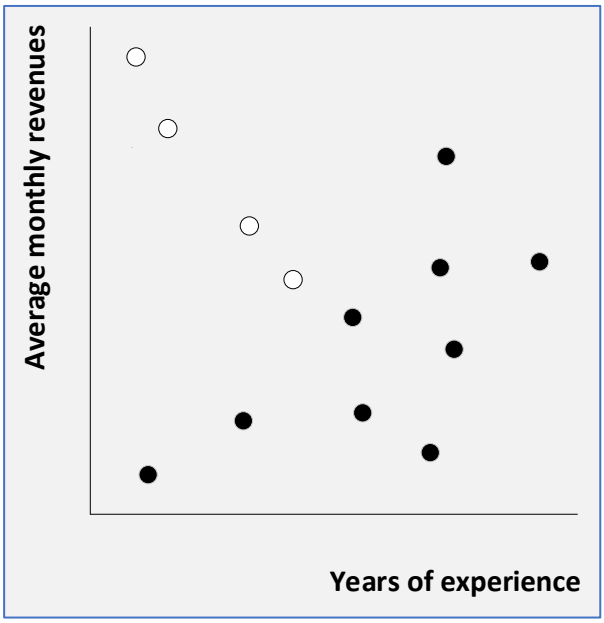


Figure 9: Which are the unusual salespersons?

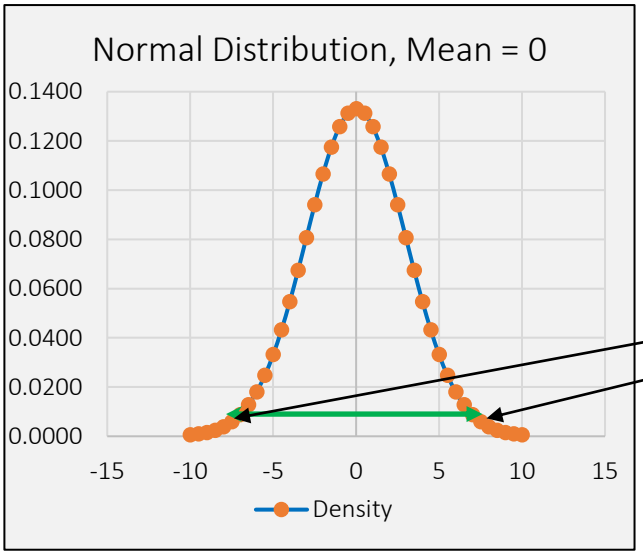
In other words, the farther away the new observation lies from the cluster of historic data, the more *confident* we are in rejecting the idea that this new employee's numbers are different from the norm. In Figure 9, which numbers lie outside the norm?

3.5.2. Two standard deviations rule

Fortunately, probability theory offers a way to detect an outlier, with two assumptions of course. First, we accept that a normal distribution is the rule producing our data; for natural, social, and economic phenomena, this often works. Second, referring to **[Area under the Normal.xlsx]**, if we replace the fixed x values with a standard deviation scale, we can understand how deviations from the mean mark areas under the probability distribution. Table 7 is an extract from the spreadsheet, showing the added scale for standard deviations from the mean. The mean (10) lies 0 standard deviations away from itself, and the fixed value of 7 is -1 standard deviations away, while the fixed value of 13 lies +1 standard deviations from the mean.

Table 7: Standard deviations and values of X			
7	0.08066	0.158655	-1.00
7.5	0.09397	0.202328	-0.83
8	0.10648	0.252493	-0.67
8.5	0.11736	0.308538	-0.50
9	0.12579	0.369441	-0.33
9.5	0.13115	0.433816	-0.17
10	0.13298	0.5	0.00
10.5	0.13115	0.566184	0.17
11	0.12579	0.630559	0.33
11.5	0.11736	0.691462	0.50
12	0.10648	0.747507	0.67
12.5	0.09397	0.797672	0.83
13	0.08066	0.841345	1.00

Now we know the CDF traces out the area under the probability distribution, so to calculate the area associated with a ± 1 std. deviation we only need to subtract the area of the upper SD from the area of the lower SD, which here is .841345 -.158655 =.68269 = 68.27% of the area under the PDF. A ± 2 SD band includes about 95.45% of the area under the PDF. Make sure to watch the video to understand this derivation.



The area shown by ± 1 standard deviation encloses 68% of the area under a normal curve. The area shown by ± 2 standard deviations make up about 95% of the area under a normal curve.

A + 2 SDs "rule" leaves 5% in the tails together (2.5% in each tail).

A common decision is to reject observations t

hat lie more than two standard deviations from the

Figure 10: Normal distribution

Example: If the marks in a class have a mean of 70, with a standard deviation of 10, that means that 2.5% of the class will have a mark of 90 and over, and 2.5% of the class will have a mark of 50. These are the assumptions that govern the usual numerical mark to letter grade conversion. Courses that are mathematical and technical have marks with bimodal distributions (two peaks) and do not conform to the normal. Others are skew left and skew right. Such double peaks and asymmetry create problems in setting simple rules for classifying outliers and aligning numerical and letter grades.

3.6. Classical statistical decision theory

Classical statistical decision theory rests on the idea that one knows the underlying probability distribution governing the data (call this the "source" or "reference" distribution). Then, we can assess whether a new observation "belongs" to that reference distribution. The question is simple: Is the new observation "part of the family?"

We never *prove* that the new observation is from the source distribution; rather, we calculate the probability of being wrong if we reject that claim. We create a statement, the *null hypothesis*, termed H_0. After gathering evidence, we "judge" whether can accept the null; if not, we turn to an alternative hypothesis termed H_a.

Here are examples of null hypotheses and their alternatives:

- H_0: Men and women have equal wages in welding occupations. H_a: Men and women have unequal wages in welding occupations.
- H_0: A coin that returns 70% heads on 20 tosses of a coin is unfair. H_a: The coin is fair.
- H_0: Before receiving an interview, a Black graduate in economics sends out the same number of resumés as a white graduate. H_a: Black and white graduates send out different numbers of resumés before receiving an interview. Or H_a'': Black graduates send out more resumés than white graduates before receiving an interview.

Notice the neutrality of the null in the first and third example. This is intentional. Often researchers have beliefs, such as wage discrimination exists based on sex, race, or age. By creating a null that asserts no discrimination exists, in this case, the test is the equality of wages, we create a rigorous decision model, where the evidence must be sufficiently strong to reject the null. Statistical decision theory uses the rejection of wrong ideas and not the acceptance of correct ideas. This ideal of "falsification" is a tenet of modern science. As we assemble evidence, such as male and female wages, we measure the difference between average wages and when that difference becomes large, accepting the null becomes untenable. The host of statistical tests and experimental designs create the structure for making that judgment about when to reject the null.

As evidence mounts that the gap between Black and white peoples' wages are not zero, we take less risk in rejecting the null. The wider the gap, the lower the probability of being wrong when we reject the null, and we reject the null in favour of an alternative hypothesis.

A null hypothesis for a statistical test and a research hypothesis framing a research study or program related: a research hypothesis may result in a series of null hypotheses. A null hypothesis poses a specific statement, capable of falsification by assembling evidence. The null is not a proposition you prove directly, but a statement you can reject based on the evidence, *with a specific chance of being wrong in that rejection*. Recall the discussion about unusual observations. The more unusual the observation, the less risk we take in rejecting the idea that it is associated within the pool of accumulated observation.

It is also possible to frame null hypotheses as two-sided or one-sided by how one creates the alternative. The first two hypotheses above have two-sided alternatives – the wages of male and female welders are different. A one-sided alternative might be men's wages are higher (or lower) than women's wages. The third example presents a two-sided alternative (H_a') and a one-sided alternative H_a''. In this course, we will only consider the simplest two-sided decision problem.

In simple decision scenarios with two options, we can make two types of errors. A Type 1 error occurs when we reject a null hypothesis when we should accept it; a Type 2 error occurs when we accept a null hypothesis when we should reject it. In a court of law, we assume an accused is innocent until proven guilty (null hypothesis is that someone is innocent), then a Type 1 error is convicting a guilty person. Alternatively, a Type 2 error is accepting the null when we should reject it (releasing a guilty party).

This section focuses on a common statistical test, namely, evaluating the difference of means involving normal distributions. Other tests exist involving differences of means with other distributions, differences of proportions, differences in variances, etc.

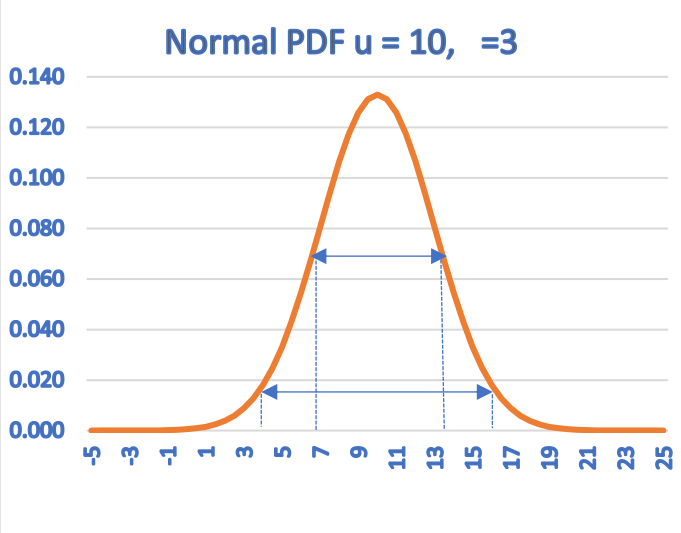


Figure 11: Normal showing standard deviations

Figure 11 is the PDF for a normal distribution with μ =10 and σ=3. The range of ± 3 around the mean (from 7 to 13) is one standard deviation and defines 68% of the area under the curve; the range of ±6 (4 to 16) defines 95% of the area. By implication, 2.5% is in each tail, and so 5% combined. When the risk of being wrong (in rejecting the null) falls to 5%, most accept the

evidence and switch to the alternative hypothesis. Setting this risk tolerance of being wrong depends on context. Plane manufacturers (and passengers) would not accept a 5% rule for rejecting a null hypothesis.

Imagine that the difference in average male and female wages for two random samples of welders were less than one standard deviation, falling in the area noted as 68.28%. If we reject the null hypothesis, we take a risk in that rejection. To be specific, if the difference were -.33 standard deviations (male wages minus female wages), while this may conform to our expectations, the difference is small. Now imagine the difference in the wages were -2.00 standard deviations. If we reject the null, we assume an approximately 2.5% chance (2.27%) of being wrong in that rejection.

> ***Example***: Imagine that male welder wages are $27.50 per hour on average with a standard deviation of $5.75. If female wages are $23.25, is the difference between male and female welders' wages statistically significant?

Assuming that the variance of male and female wages is the same (a strong assumption, relaxed below), the Excel function =NORM.DIST finds the probability for a value of X with a given mean and standard deviation, which is .0000078, meaning that we face a small risk in rejecting the hypothesis that this wage of $23.25 comes from the population of male wages. If the standard deviation is $3.1, then this probability falls to .013 if the mean of male wages is $24.15. See .

Male and Female Welders.xlsx

As an aside, this example presented a *one-tailed test* for reasons explained in the next section.

Video: Male and Female Welders

3.7. The Z score and critical values

The process of calculating the probabilities associated with a test value (female wages) with a reference distribution defined by mean and standard deviation (male wages) is quite cumbersome. It would be handy to have a single score and a table that would tell us whether we were taking a 10%, 5%, or 1% chance of making a Type 1 error when we reject the null hypothesis. The Z score does that and has the formula:

$$Z = (X_0 - \overline{X}) / \sigma$$

In the spreadsheet [**Male and Female Welders Standard.xlsx]** the wage scale uses the formula above, and the Excel function =STANDARDISE transforms wages to a Z score. This creates a standard normal distribution with a mean of 0. The same probabilities are available after this

transformation, but the standard normal distribution supports more direct estimation of values associated with a 10%, 5% and 1% chance of making a Type 1 error in rejecting the null.

3.8. One-tail and two-tail tests

Consider two null hypotheses:

- H_0: Male and female welder wages are equal.
- H_0: Male welder wages are more than female welder wages.

In the first instance H_0: $\mu_m=\mu_f$, and in the second H_0: $\mu_m>\mu_f$. The first case uses a two-tail test, while the second a one-tail test. Continuing with the simple (simplistic) assumption where the information on the distribution of male wages follows the normal with a known mean (μ_m) and standard deviation σ_m, with a single observation for female wages X_f, the example **[Male and Female Welders Standard.xlsx]** shows the probability that a specific female wage came from the proposed male wage.

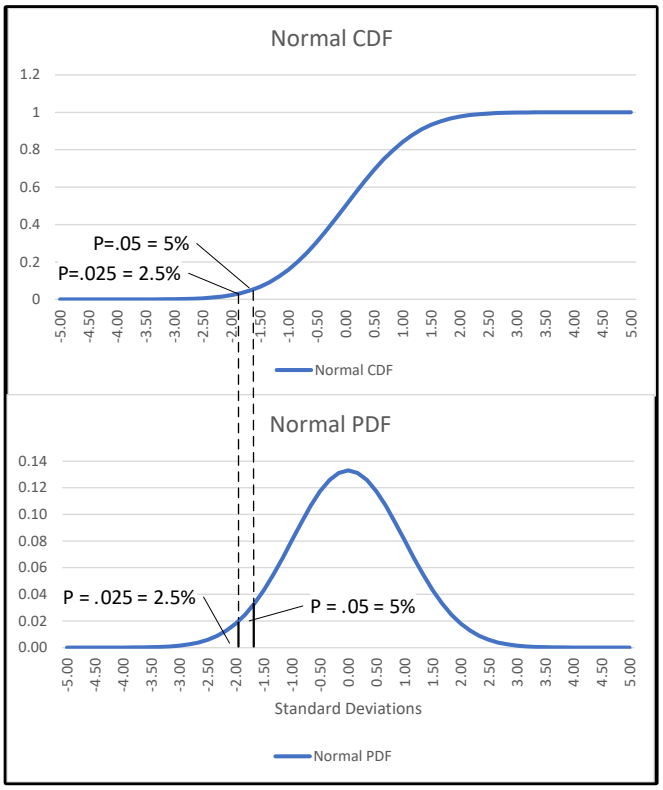


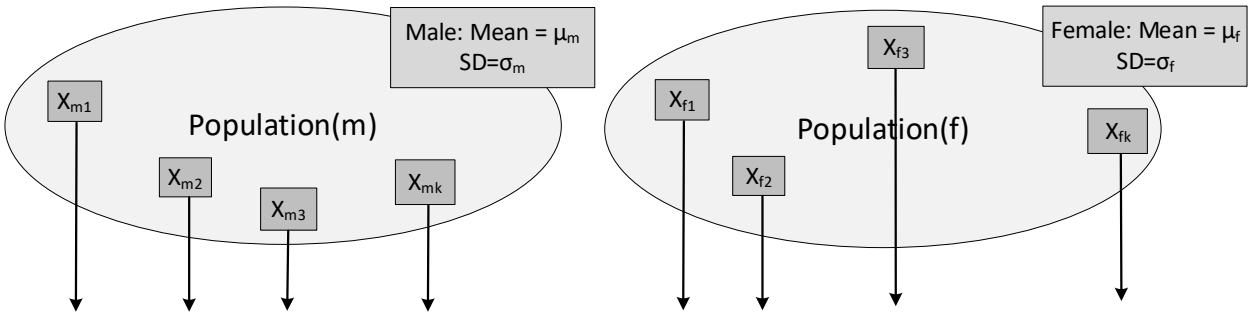
Figure 12: Critical values

Figure 12 shows the critical values associated with a 5% chance of making a Type 1 error in rejecting the null. If the null poses equality, then a + 2.5% critical value reflects that a female wage above or below the male mean is possible. With a null posing inequality, specifically that we hypothesize that male welder wages are greater than female welder wages, the one-tail test is right.

For a one-tail test (the null states that male wages exceed female wages), the critical values associated with the 10%, 5%, and 1% Type 1 error rates are $\pm$1.28, $\pm$1.65, and $\pm$2.33; the lower the error one is willing to accept, the larger the difference between the mean of the reference distribution (male wages) and the sole test observation (the single observation on female wages). The critical values may be positive or negative, depending on whether the test observation lies above or below the reference mean (μ_m).

Now consider a more sophisticated (and scientifically better) example of two samples of wages, one for male welders, the other for female welders. We now have two means, μ_m and μ_f, as well as two standard deviations σ_m and σ_f. The null hypotheses are the same but slightly rewritten

- H_0: $\mu_m - \mu_f = 0$; $H_a = \mu_m - \mu_f \geq 0$ (or $H_a = \mu_m - \mu_f \leq 0$) (one-tailed test)
- H_0: $\mu_m - \mu_f = 0$; $H_a = \mu_m - \mu_f \neq 0$ (two-tailed test)



$\bar{X}_{m1}$ $\bar{X}_{m2}$ $\bar{X}_{m3}$ $\bar{X}_{mk}$ $\bar{X}_{f1}$ $\bar{X}_{f2}$ $\bar{X}_{f3}$ $\bar{X}_{fk}$

$(\bar{X}_{m1} - \bar{X}_{f1})_1$ $(\bar{X}_{m2} - \bar{X}_{f2})_2$ $(\bar{X}_{m3} - \bar{X}_{f3})_3$ $(\bar{X}_{mk} - \bar{X}_{fk})_k$

Above, two populations of the wages of *all* male welders (m) and *all* female welders (f) have means of μ_m and μ_f and standard deviations of σ_m and σ_f. X_{m1} is a *sample* (not a single observation) of male wages drawn from Population(m), X_{f1} is a sample of female wages from Population(f). We have many such samples: $\bar{X}_{m1}$ is the mean of sample 1 from the male population and $\bar{X}_{f1}$ the mean of sample 1 from the female population. This leads to $\bar{X}_{m1} - \bar{X}_{f1}$ as the first element of the *sampling distribution of the difference of means.* To make sense of the sampling distribution of the means, several pairs of samples from Population(m) and Population(f), at least 10 and preferably 30 or more, will support reliable and valid estimates of *the mean of the differences in means*. Most research studies will involve a single pair. This needs a concrete example, so see….

Video: Difference in Means

Mean male wage = 27.5, SD= 6

Mean female wage = 21.25, SD = 3

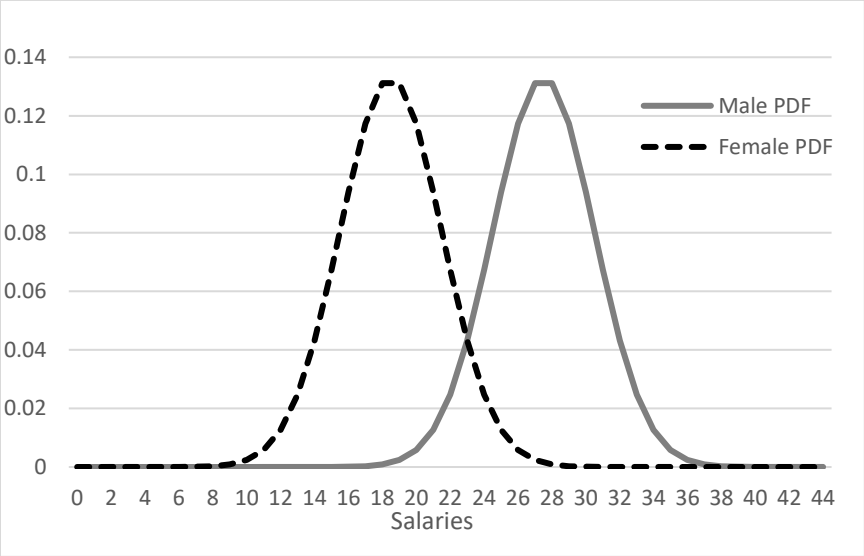
The important idea is that, compared to the distribution of female wages, the distribution of male wages has a higher mean (higher average wages) and a higher standard deviation (more variability, resulting in a "flatter" shape.

Figure 13: Example of wage distributions

Difference in Means.xlsx

Mean male wage = 27.5, SD= 3

Mean female wage = 18.5, SD = 3



In contrast to the example in Figure 13, the two distributions have different means, but the same standard deviation. Both the difference in means and the difference in standard deviations affect the decision on rejection of accepting the null hypothesis .

Figure 14: Example of wage distributions

As the distributions of male and female welder wages "separate," the risk falls in rejecting the null hypothesis that male and female wages are equal. It is possible to create a Z score for differences of means, but first note that

$$\mu_{\bar{X}_{m1}-\bar{X}_{f1}} = \mu\bar{X}_{m1} - \mu\bar{X}_{f1} \text{ and } \sigma_{\bar{X}_{m1}-\bar{X}_{f1}} = (\sigma^2_{m1}/N_{m1} + \sigma^2_{f1}/N_{f1})^{1/2}$$

The example shows the calculation in detail, and the video describes the process of creating this simulation.

Video: Difference in Means Z Score

Diff of Means Z score.xlsx

Excel has a function =Z.TEST, which calculates the Z score for a single observation against a set of values. (View in the help menu of Excel).

The three levels of 10%, 5%, and 1% are standard in economic research; 10% when for social or political reasons one wishes to (needs to) reject the null hypothesis, which is the common standard, and 1%, when the risks of a Type 1 error are high. Selecting the right Type 1 error level

requires judgment – when the costs of rejecting a true null hypothesis is high, one will typically want the Z score to be higher (p value lower). This table presents the critical values of Z for different Type 1 errors.

Table 8: Critical values of the Z score (standard normal distribution)

Level of Significance	10%	5%	1%	.5%
Value of Z for one-tailed tests	±1.28	±1.65	±2.33	±2.58
Value of Z for two-tailed tests	±1.65	**±1.96**	±2.58	±2.81

Just to recall, a one-tailed test is associated with an inequality as the null (H0: Female wages exceed male wages), while the two-tailed test is associated with an equality as the null (H0: Female and male wages are equal). The two-tailed test at the 5% level has a Z score of +1.96, and we will see how this becomes a rule of thumb in testing regression models later in the course.

4. A very brief note on Bayesian analysis

Bayesian analysis reflects the work of Thomas Bayes. Very briefly, this style of statistical decision-making starts our analysis with a belief or prior. Informally, we usually have a belief about the state of the world before we begin our research. For example, we might believe men and women receive the same pay as welders. This is the null and termed the *prior probability of distribution*. Then, as evidence mounts in the form of observed wages, we may choose to revise that expectation. The degree to which we revise our beliefs depends on the degree of discrepancy between the new observation on female wages and the number of these "discordant" observations.

> ***Example***: If I have 100 observations on male wages and calculate a mean of $25 and SD = 20, a single observation for a female welder at $23 would not cause any reason to expect the distribution of female wages was different than that of men. In other words, no reasons exist to revise the prior. But with more observations on female wages lying below $25, the analyst will revise the prior. Observations lying much below the male mean should cause a faster revision than a larger number just below the male mean.

5. Summary

This Module has presented a basic introduction to classical statistical decision-making. Advancements have occurred in decision theory and in the analysis of relationships.

- The normal distribution, and specifically the standard normal in evaluating the difference of means, is but one in common use. In the next two modules, the t and the F distributions will figure prominently in testing regression models.
- This chapter has not examined the risks associated with accepting a false hypothesis or accepting a null when we should reject it.
- One concept not covered in this Module is bias in a statistic. The experiment shown in [] shows why the mean of a sample is an unbiased estimator of the population mean, but the sample variance tends to overstate the population variance.

Bias in a Statistic.xlsx

Imagine our survey of male and female wages showed a statistically significant difference (we can reject the null of equality with less than a 1% chance of making a Type 1 error). Does it require government regulation? Not without more analysis. For example, women have only just recently entered the trades in significant numbers, so the sample of female welders may be much younger, on average, than the sample of male welders. Male welders may have seniority, worked in a range of environments, taken upgrading courses, and not experienced work interruption (due to childbirth). All these factors (lurking variables) and others that we lack the imagination to consider may affect the distribution of welding wages. The next two modules present techniques for introducing added information into our analysis of wages, which serves other purposes, such as measuring the interconnections among social and economic factors and forecasting.

Annex: Excel statistical/probability functions and formulas

Excel statistical formulas		
Formula	Explanation	Example
=AVERAGE(range)	Returns the mean of row, column, or array (set of contiguous cells).	=AVERAGE(A1:A30) returns the average of contents in cells A1 to A30. The cell location of this formula cannot be in cells A1 to A30 (circular reference), and it will ignore cells with blanks or text.
=STDEV.P(range) =STDEV.S(range) =VAR.P(range) =VAR.S(range)	Returns the standard deviation or variance of a column, row, or array. Use .P for the population value and .S for a sample. When unsure,	=STDEX.P(A1:F30) returns the standard deviation of an array of numbers (ignoring blanks and text).

Excel statistical formulas		
Formula	Explanation	Example
	use the .P version.	
=MIN(range) =MAX(range)	Returns the minimum/maximum value in the row, column, or array.	=MAX(A1:A30)-MIN(F2:F10) =MAX(A1:A30)-MIN(A1:A30) returns the statistic "range." These functions are useful in flagging outliers.
=MED(range)	Returns the median of a row, column, or array.	=MED(A1:A30) Comparing the results of the simple mean (=AVERAGE) and the median can give clues on whether the distribution is skewed. (Mean > Median implies skew right and Mean < Median suggest skew left).
=GEOMEAN(range)	Returns the geometric mean of a row, column, or array.	=GEOMEAN(A1:F20)
Notes: • The reference cell must not be in the specified range, otherwise you will get a "circular reference" error. • While these formulas will ignore blank cells and text, others will return errors if the rows are all blanks. • You can write the formulas in lower or uppercase (or a combination); Excel converts them all to uppercase.		

Excel probability formulas		
Formula	Explanation	Example
=BINOM.DIST (number_s,trials, probability_s, cumulative)	This provides the CDF (cumulative = True) or PDF (cumulative = False) for the binomial with "s" successes out of n trials, with a probability of success.	=BINOM.DIST(5,23,.03,T) yields the probability of, at most, five successes out of 23 trials when the probability of success is .03. = BINOM.DIST(5,23,.03,F) yields the probability of exactly five successes out of 23 trials when the probability of success is .03. (Note that the definition of T and F for cumulative is the same for all probability functions).
=NORM.DIST(x, mean, standard_dev, cumulative)	This provides the CDF/PDF(cumulative =	=NORM.DIST(5,15, 10,True/False) is the probability

Excel probability formulas		
Formula	Explanation	Example
	True/False) for the value of x, for normal distribution with the stated mean and standard deviation.	of seeing at least/exactly the value of 5 with a normal distribution with a mean of 15 and a standard deviation of 10.
=LOGNORM.DIST((x, mean, standard_dev, cumulative)	This provides the CDF/PDF(cumulative = True/False) for the value of x, for lognormal distribution with stated mean and standard deviation.	=LOGNORM.DIST(5,15, 10,True/False) is the probability of seeing at least/exactly the value of 5 with a lognormal distribution with a mean of 15 and a standard deviation of 10.
=POISSON.DIST(x, mean, cumulative)	This provides the CDF/PDF(cumulative = True/False) for the value of x, for Poisson with the stated mean.	=POISSON.DIST(5,10,True/False) is the probability of seeing at least/exactly 5 if the distribution is Poisson with a mean of 10.