Business Statistics

Agenda – Inferential Statistics I

1. Inferential Statistics
2. Some fundamental terms first
    a. Random Variables
    b. Distribution and its types
3. Binomial Distribution
4. Uniform Distribution
5. Normal Distribution
6. Sampling and Inference
    a. Simple Random Samples
    b. Sampling Distribution
    c. Central Limit Theorem
7. Estimation
    a. Point Estimation
    b. Interval Estimation

Inferential Statistics

Descriptive vs. Inferential Statistics

Summaries from data → **Is that enough?** → **Inferential Statistics**

Summaries give a sense of central tendency, variation, association

Tell a lot about 'what's happening'

Mean, standard deviation, correlation, etc.

Typically, we work with only 'samples' of data

It is not rough to make generalized statement about the population (what we are actually interested in)

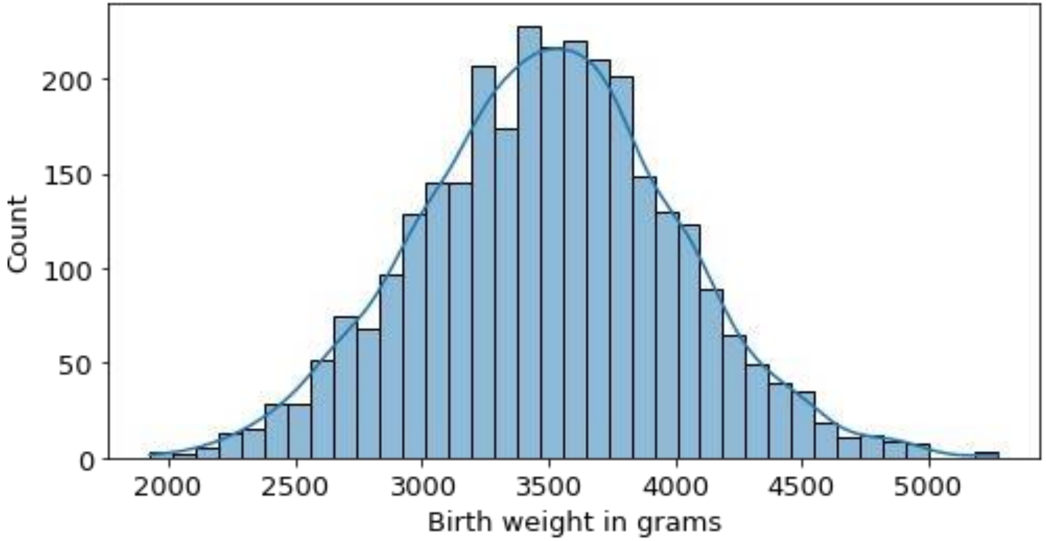
Challenge - how to learn about population from the sample?

Inferential statistics helps tackle that challenge

There are powerful methods to draw reasonable conclusions about the population from an observed sample

This becomes extremely critical in business decision making

Role of distributions in inferential statistics



What is the chance that birth weight is less than 3000 grams?

Descriptive answer: Count the number of births below 3000 grams from the histogram.

Inferential answer: Consider the underlying distribution and calculate the proportion (area) less than 3000 grams.

Business Problems

Quality Testing	Is the new manufacturing process better/more reliable than the old process?
Meteorology	How likely is it that temp will be more than 20 degree C on a specific day?
Human Resources	Does training the workforce improve sales?
Digital Marketing	What is the chance that the conversion rate on the website will be above x% next month?
...	...

Some fundamental terms

Random Variable

Suppose there are 1000 students in the university. What is the probability that 500 students will pass the upcoming exam?

There is a 50-50 chance that each student will pass or fail

The total number of students who pass can range from 0 to 1000

A random variable assigns a numerical value to each outcome of an experiment. It assumes different values with different probability.

Discrete Random Variable

You work for an Auto insurance company. Suppose the number of insurance claims filed by a driver in a month is a random variable (X) described by

$$X = \begin{cases} 0, \textit{ with prob } 0.95 \\ 1, \textit{ with prob } 0.04 \\ 2, \textit{ with prob } 0.008 \\ 3, \textit{ with prob } 0.002 \end{cases}$$

All probabilities must be non-negative and sum to 1.

When all possible values the random variable can take can be listed, we call it a **discrete random variable**

Continuous Random Variable

Suppose the volume of soda in a bottle is described by a random variable.

Can we list all possible values?

498 mL, 499 mL, 500 mL, … … …   What about 499.2129415 mL?

Sometimes it is just not possible to list all values a random variable can take

If the random variable can take any value in a given range, we call it a **continuous random variable**

Probability Distribution

Probability Distribution

Describes the values that a random variable can take, along with the probabilities of those values

Discrete Probability Distribution

Arises from discrete random variables

Has an associated **probability mass function**, which gives the probability with which the random variable takes a particular value

Continuous Probability Distribution

Arises from continuous random variables

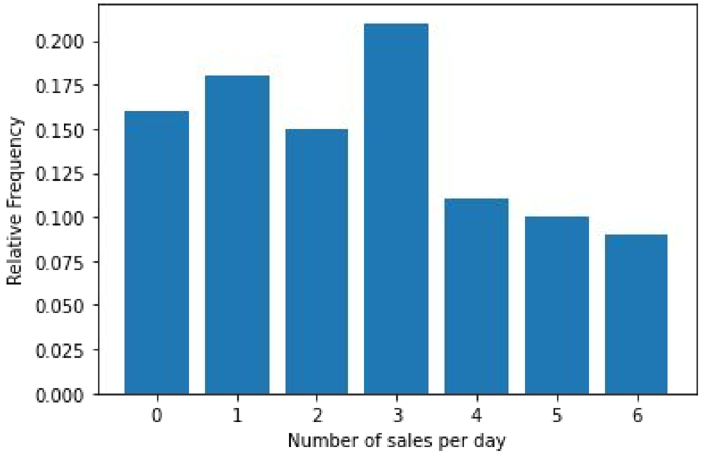
Has an associated **probability density function**, which helps determine the probability with which the random variable lies between two given numbers

Probability Distribution: Example

A company tracks the number of sales new employees make each day during a 100-day probationary period. The results for one new employee are shown. Construct and plot a probability distribution.

Sales	#Days
0	16
1	18
2	15
3	21
4	11
5	10
6	9

Sales	#Days	Relative Frequency
0	16	0.16
1	18	0.18
2	15	0.15
3	21	0.21
4	11	0.11
5	10	0.10
6	9	0.09



Distributions around us (commonly occuring)

Bernoulli	Company has introduced a new drug to cure a disease, it either cures the disease (it's successful) or it doesn't (it's a failure)
Binomial	The number of defective products in a batch production run
Uniform	The number of of microwave oven sold daily at a busy consumer goods store
Normal	Income distribution of a country on a logarithmic scale

Basic distributions - Binomial

Bernoulli Distribution

Success and failure are non-judgemental. Any one outcome may be termed as success

It has only **two possible outcomes**, namely 1 (success) and 0 (failure), of **one single trial**.

$$X = \begin{cases} \textbf{1, with prob } \boldsymbol{p} \\ \textbf{0, with prob } \boldsymbol{1-p} \end{cases}$$

Very useful in many scenarios:

Manufacturing defective parts	Outcome of medical test

Binomial Distribution

Suppose we ask any adult who uses the app TikTok if he/she has ever posted a video on the app

The answer can be Yes or No (success or failure)

We can use the Bernoulli distribution to model this scenario

Now let us extend this into a survey of 25 adults chosen at random

We can define **a random variable *X* which counts the number of successes** (say, the number of adults who responded Yes)

Binomial Distribution

In many situations an experiment may have only two outcomes - success and failure

These experiments can be modelled using the Binomial probability distribution.

Bernoulli Distribution is a special case of Binomial Distribution with a single trial.

Probability Mass Function

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

Binomial Distribution : Assumptions

Number of trials (n) is fixed.

Each trial is independent of the other trials.

There are only two possible outcomes (success or failure) for each trial.

The probability of a success (p) is the same for each trial.

What happens if these assumptions are violated?

In a month of 30 days, what is the probability that it will rain on more than 10 days, if on average the chance of rain on a given day is 20%?

If we assume that:

1. The event of rain on a particular day is independent of it raining on the previous day.
2. The chance of rain does not increase or decrease over the duration of the month.

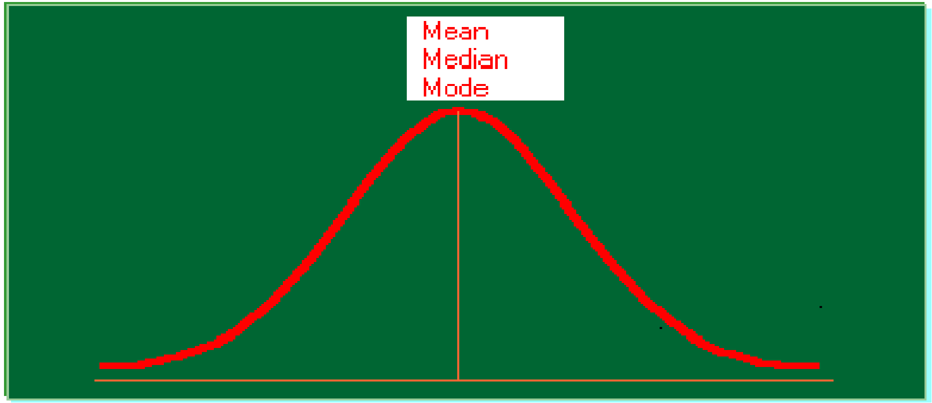
Then we can use the binomial distribution with n=30 and p=0.2 to calculate the probability.

Assumptions 1 and 2 in the example are not strictly valid, but they allow for a direct calculation that may be good enough for practical purposes.

What is Probability Distribution?

- In precise terms, a probability distribution is a total listing of the various values the random variable can take along with the corresponding probability of each value. A real life example could be the pattern of distribution of the machine breakdowns in a manufacturing unit.
- The random variable in this example would be the various values the machine breakdowns could assume.
- The probability corresponding to each value of the breakdown is the relative frequency of occurrence of the breakdown.
- The probability distribution for this example is constructed by the actual breakdown pattern observed over a period of time. Statisticians use the term “observed distribution” of breakdowns.

What is Normal Distribution?



- The Normal Distribution is the most widely used continuous distribution. It occupies a unique place in the field of statistics. In fact, the entire quality control function that employs the statistical techniques heavily will come to a grinding halt without the use of the normal distribution. The control charts for reducing and stabilizing variation relies on the normal distribution. Process capability studies to meet the customer specifications needs the normal distribution. The whole subject matter inferential statistics is based on the normal distribution. In all management functions including the human side, the observed frequency distributions encountered are all fairly close to the normal distribution when the sample size is reasonably large.

Properties of Normal Distribution

- The normal distribution is a continuous distribution looking like a bell. Statisticians use the expression “Bell Shaped Distribution”.
- It is a beautiful distribution in which the mean, the median, and the mode are all equal to one another.
- It is symmetrical about its mean.
- If the tails of the normal distribution are extended, they will run parallel to the horizontal axis without actually touching it. (asymptotic to the x-axis)
- The normal distribution has two parameters namely the mean μ and the standard deviation σ

Normal Probability Density Function

In the usual notation, the probability density function of the normal distribution is given below:

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

x is a continuous normal random variable with -∞ < x < ∞.

Standard Normal Distribution

- The Standard Normal Variable is defined as follows:

$$z = \frac{x - \mu}{\sigma}$$

- Please note that Z is a pure number independent of the unit of measurement. The random variable Z follows a normal distribution with mean=0 and standard deviation =1.

$$f(Z) = \frac{1}{\sqrt{2\pi}} e^{-\left[\frac{Z^2}{2}\right]}$$

Example Problem

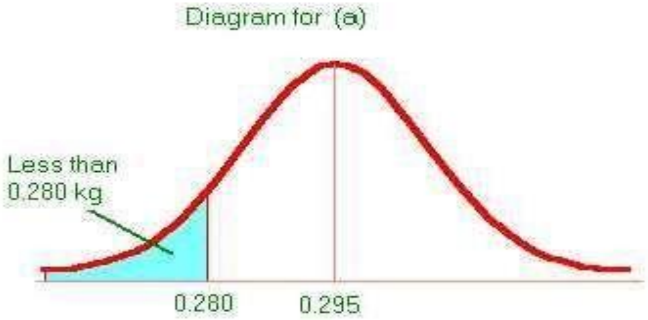
The mean weight of a morning breakfast cereal pack is 0.295 kg with a standard deviation of 0.025 kg. The random variable weight of the pack follows a normal distribution.

a) What is the probability that the pack weighs less than 0.280 kg?

b) What is the probability that the pack weighs more than 0.350 kg?

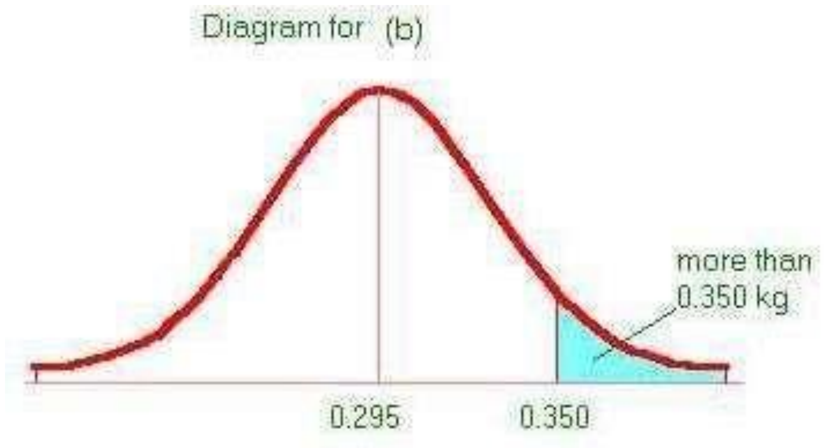
c) What is the probability that the pack weighs between 0.260 kg to 0.340 kg?

Solution – a)



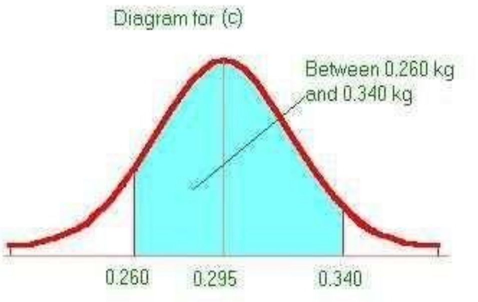
z = $\dfrac{x - \mu}{\sigma}$ = (0.280-0.295) / 0.025 = -0.6. Click "Paste Function" of Microsoft Excel, then click the "statistical" option, then click the standard normal distribution option and enter the z value. You get the answer. Excel accepts directly both the negative and positive values of z. Excel always returns the cumulative probability value. When z is negative, the answer is direct. When z is positive, the answer is =1- the probability value returned by Excel. The answer for part a) of the question = 0.2743(Direct from Excel since z is negative). This says that 27.43 % of the packs weigh less than 0.280 kg.

Solution – b)



z = $\dfrac{x-\mu}{\sigma}$ = (0.350-0.295) / 0.025 =2.2. Excel returns a value of 0.9861. Since z is positive, the required probability is = 1-0.9861 =0.0139. This means that 1.39% of the packs weigh more than 0.350 kg.

Solution – c)



For this part, you have to first get the cumulative probability up to 0.340 kg and then subtract the cumulative probability up to 0.260. $z = \frac{x - \mu}{\sigma}$ = (0.340-0.295)/0.025 =1.8(up to 0.340)

$z = \frac{x - \mu}{\sigma}$ = (0.260-0.295) / 0.025 = -1.4 (up to 0.260). These two probabilities from Excel are 0.9641 and 0.0808 respectively. The answer is = 0.9641 - 0.0808 = 0.8833. This means that 88.33% of the packs weigh between 0.260 kg and 0.340 kg.

Poisson Distribution

- Poisson Distribution is another discrete distribution which also plays a major role in **quality control** in the context of reducing the number of defects per standard unit.

- Examples include number of defects per item, number of defects per transformer produced, number of defects per 100 m2 of cloth, etc.

- Other examples would include 1) The number of cars arriving at a highway check post per hour; 2) The number of customers visiting a bank per hour during peak business period; 3) The number of pixels in the image that are corrupted.

Poisson Probability Function

Poisson Distribution Formula

$$P(x) = \frac{e^{-\lambda}\lambda^{x}}{x!}$$

where

$P(x)$ = Probability of x successes given an idea of λ

λ = Average number of successes

e = 2.71828(based on natural logarithm)

x = successes per unit which can take values 0, 1, 2, 3,............∞

λ is the Parameter of the Poisson Distribution.

Mean of the Poisson Distribution is = λ

Standard Deviation of the Poisson Distribution is = $\sqrt{\lambda}$

Example – Poisson Distribution

If on an average, 6 customers arrive every two minutes at a bank during the busy hours of working, a) what is the probability that exactly four customers arrive in a given minute? b) What is the probability that more than three customers will arrive in a given minute?

Questions if any…