Statistical Analysis

II. Probability and Decision-Making Approaches

What is Probability?

- **Probability** is a measure of how likely (probable) specific observations of a variable may or may not occur.

Example:

» P (lead concentration <= 150 ppm) = 10% or 0.10

» P (lead concentration > 233 ppm) = 82% or 0.82

The Probability

Classical Interpretation Approach

Suppose that an event **E** happens in **h** ways out of a total of **n** possible likely ways. Then,

$$p = \Pr(E) = P(E) = \frac{h}{n+1}$$

This is called **Probability of Occurrence** of the event (or **Success or exceedance**)

Therefore the **Probability of Nonoccurrence (Failure)** is

$$q = P(notE) = 1 - P(E)$$

Probability of an event is always between **0 and 1**

The Probability

Classical Interpretation Approach

Example:

Let E be an event of having the number 3 in a single toss of a die. What is the probability of having E=3?

Solution:

n=6 ➔faces of the die

h=1 ➔ face of the event E=3

P(3)= 1/7 = 0.14 ➔ means that this event might happen one time in every six tosses.

P(not 3) = 1-0.14= 0.86

The Probability

Classical Interpretation Approach

Consider the following **soil contamination example**:

What is the probability that a sample taken randomly from the soil at a refinery site has benzene concentration that exceeds a threshold value of 100,000 ug/kg? assume you have a total of 90 samples 11 out of which have concentration values above 100,000 ug/kg.

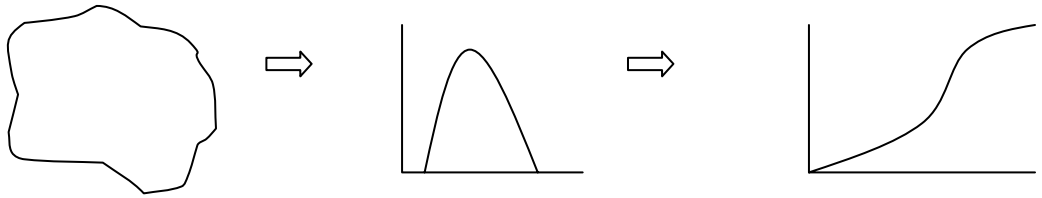
$$P(sample\ ) = \frac{h}{n+1}$$

Therefore, P(sample) = 11/91 = **0.121 or 12.1%**

The Probability

Cum. / Relative Frequency Interpretation Approach

Frequency distributions (e.g. relative frequency dist. And cumulative frequency dist.)



Probability

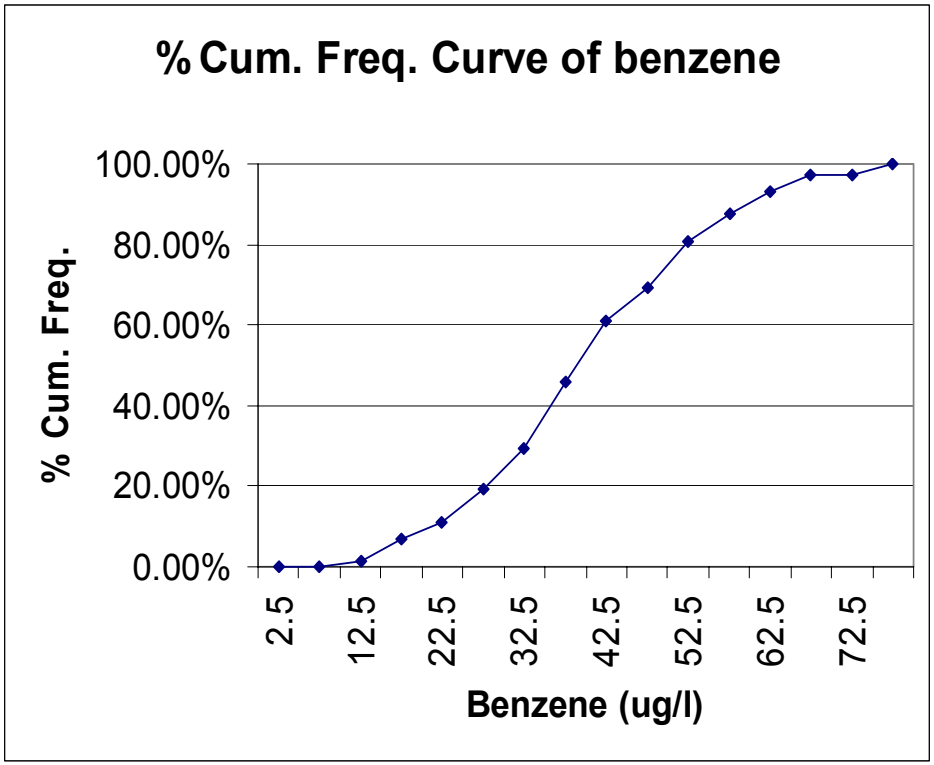
The Probability

Cum. / Relative Frequency Interpretation Approach

Now, consider the following **example**:

What is the probability that benzene concentration is between 32.5 ug/l and 42.5 ug/l at a contaminated shallow aquifer?

Construct a cum. / relative frequency histogram and interpret the probability value from it (30% **how??**).



Decision-Making Practices

1. Standardized variable
2. Point and interval estimates
3. Estimation of the required number of samples
4. Comparison of means
5. Comparison of variances
6. Monte Carlo Simulation

Decision-Making Practices

1) Standardized Variable

Suppose you need to find **probability** of occurrence of a specific contaminant, below a threshold, what to do?

1. Construct a CDF ➔ large sample
2. Go to Z-Tables ➔ large or small sample

What is the Z variable?

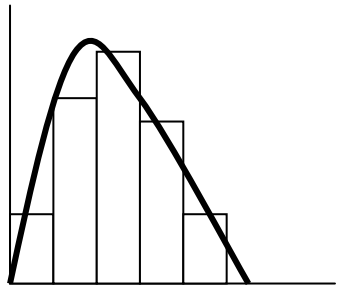
It is a **standardized normally distributed** variable with m = 0 and V = 1 (i.e. **N(0,1)**)and it has no units.

$$Z = \frac{x - m}{S}$$

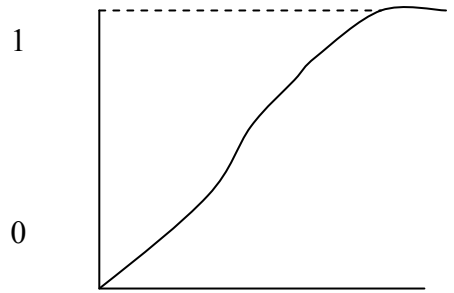
Why is it used?

We standardize values of variable x so that they come from the standard normal distribution because it is not possible to produce different probability tables for different variables and values.

What is the area under the PDF curve?



PDF

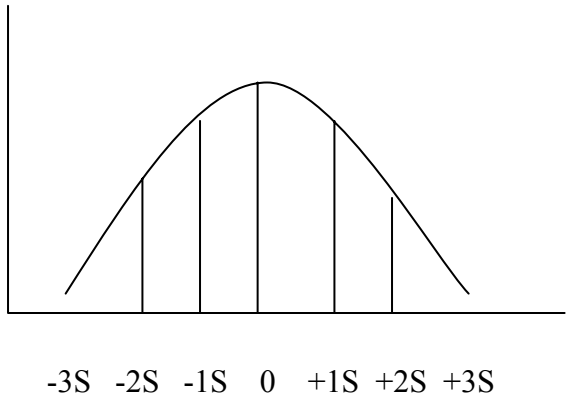


CDF

$$CDF = \int PDF$$

The area under the curve is always equal to **ONE**.

In a **normal distribution**:



68.5% of data lie within + - 1S
95.5% of data lie within + - 2S
99.7% of data lie within + - 3S

How to calculate probability using Z-variable?

Example: Suppose the mean concentration of lead (Pb) = 367 ppm and its V = 32400 ppm^2. Calculate the probability of having Pb <= 450 ppm

Solution: m = 367 ppm, S = 180 ppm, x = 450 ppm
Therefore,

$$Z = \frac{450 - 367}{180} = 0.46$$

P (x <= 450 ppm) = P (Z <= 0.46) = 0.68 = 68%

P (x > 450 ppm) = 1 - P (Z <= 0.46) = 1 – 0.68 = 0.32 = 32 %

Decision-Making Practices

2) Point and Interval Estimates

An **estimate** of a population is expressed by a **single value (i.e. point value)** like mean of sulfur content in groundwater … etc.

However a **single value** is statistically **meaningless (WHY?)**.

To be more meaningful, an interval with upper and lower limits should be associated with it. This interval is called **Confidence Interval (C.I.)** and its limits are called **Confidence Limits (C.L.).**

To calculate **C.I.,** either **Z-Distribution** or **t-Distribution** should be used depending on the available number of samples **(n)**

Student's t-Distribution

t-Distribution was discovered by W. S. Gossett who published his work under the name "**student**".

It is used to define **Confidence Intervals (CI)** to the estimated mean of a population.

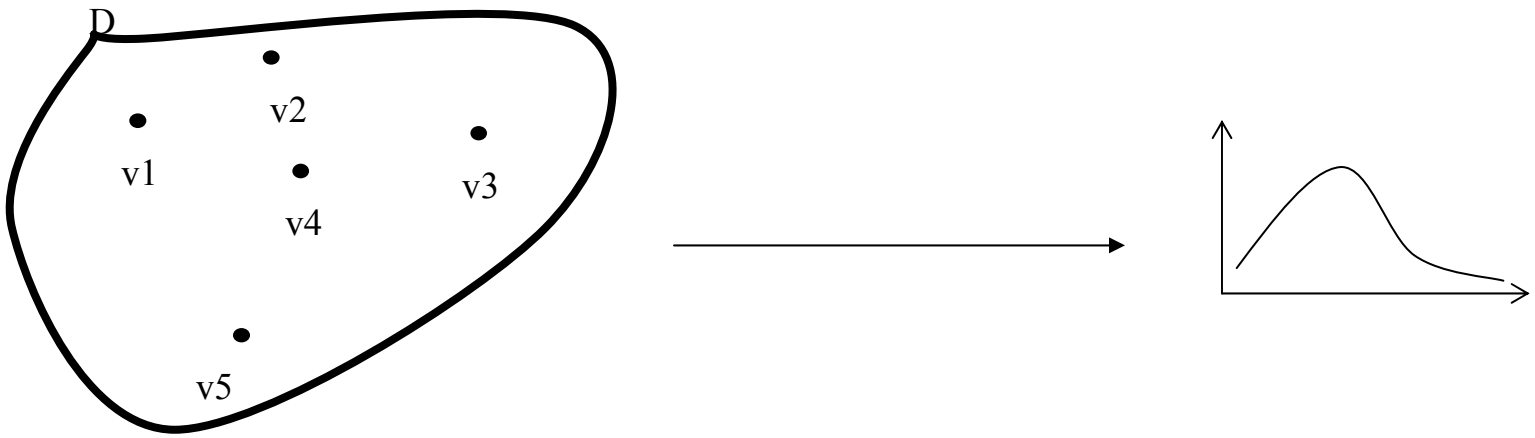
t-Distribution is used if the number of available data samples is less than 30 (i.e. **n<30**). However **Z-distribution** is used to assign CI to the estimated mean if **n>=30**.

Both distributions follow the normal distribution assumption.

Both distribution, in case of defining the C.I. satisfy the requirements of the **Central Limit Theory**

Central Limit Theory

If a random sample size n is taken from a **normal** distribution with mean **m** and standard deviation **s**, the distribution of sample means will also be **normal** with mean **m** and standard error **s/sqrt (n).**



If n >= 30 samples

$$m = \left(\bar{x}\right) \pm Z_{\frac{\alpha}{2}} \frac{s}{\sqrt{n}}$$

If n < 30 samples

$$m = \left(\bar{x}\right) \pm t_{(\frac{\alpha}{2},\nu)} \frac{s}{\sqrt{n}}$$

α = probabilit y of committing error (level of significan ce)

Decision-Making Practices

3) Estimation of the required number of samples

In order to **increase precision**, the size of confidence interval **(C.I.)** around a statistical parameter (say the mean) should be **decreased**.

To apply this, a **pilot sample n_1** of a pre-investigation project must be used to conduct a preliminary statistical analysis.

With the help of the Student's **t-distribution table**, the **required number of samples** can be estimated **by minimizing the half of the width (d) of C.I.**

To minimize **d**, **n** must be increased.

If the pilot number of samples is n_1, the **approximate more number of samples required**, **n**, is derived as follows:

$$d = t_{(\frac{\alpha}{2},\nu_1)} \frac{s_1}{\sqrt{n}}$$

$$n = \frac{\left\{ t^2_{(\frac{\alpha}{2},\nu_1)} s_1^2 \right\}}{d^2}$$

Decision-Making Practices

4) Comparison of two variances

It is revealed by testing hypothesis using the **F-distribution** or Fisher's distribution (After R. A. Fisher who discovered it).

It is a useful test to understand if a **significant difference** occurs between **the variances of two populations** (i.e. data from group of upgradient and downgradient wells).

The test is performed by:

1) Computing estimated F-statistics (F_{est})

$$F_{est} = \frac{S_1^2}{S_2^2} \qquad S_1 > S_2$$

2) Predicting the critical F-value (F_c)from the F-table

$$F_c = F(\alpha, \upsilon_1, \upsilon_2)$$

a) 3) Testing the hypothesis:
a- If $F_{est} < F_c$ ➔ Accept Null (No Difference) Hypothesis H_o (i.e. the two variances are equal)

b- If $F_{est} > F_c$ ➔ Reject Null Hypothesis H_o and accept the Alternative Hypothesis H_a (i.e. the two variances are not equal)

Decision-Making Practices

5) Comparison of two means

It is revealed by testing hypothesis using the **t-distribution**

It is a useful test to understand if a **significant difference** occurs between **the means of two data sets** (i.e. comparing pollutant concentrations between upgradient and downgradient wells intercepting a contaminated aquifer).

The test is performed by:

1) Computing **estimated t-statistics (t_{est})**

$$t_{est} = \frac{\bar{x}_1 - \bar{x}_2}{S\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}}$$

$$S = \sqrt{\frac{n_1 S_1^2 + n_2 S_2^2}{n_1 + n_2 - 2}}$$

2) Predicting the **critical t-value** (t_c)from the **Student's t-table**

$$t_c = t_{(\frac{\alpha}{2}, \nu)}$$

a) 3) Testing the hypothesis:

a) If $t_{est} < t_c$ ➔ Accept **Null (No Difference) Hypothesis H_o** (i.e. the two means are equal)
b) If $t_{est} > t_c$ ➔ Reject **Null Hypothesis H_o** and accept the **Alternative Hypothesis H_a** (i.e. the two means are not equal)

Decision-Making Practices

6) Monte Carlo Simulation

It is a statistical method for obtaining the probability distribution of output ***O*** given the probability distribution of input ***I*** . ***The output is analyzed statistically to reveal a decision.***

Examples:

- **Probability (Risk)** that a person will have a disease based on the probability that a pollutant intersects a water producing well
- **Uncertainty analysis** in travel time and spread of contaminants in porous media as a function of uncertainty in the transport model input parameters

Monte Carlo Simulation (Constructing the Model!!)