# **Principles of Statistics**

## Introduction

## Descriptive Statistics

Descriptive Statistics deals with understanding of the type of data currently available or the **sample data**.

### Frequency Distribution

### Summary Statistics for Univariate Distribution

#### Mean

#### Median:

Central value after ranking, or the value at 50th percentile.

#### Mode:

(hint: bi-modal, multi-modal): Value occurring at highest frequency

#### Extremes:

#### Percentile:

#### Variance:

For Linear statistics:

#### Standard Deviation:

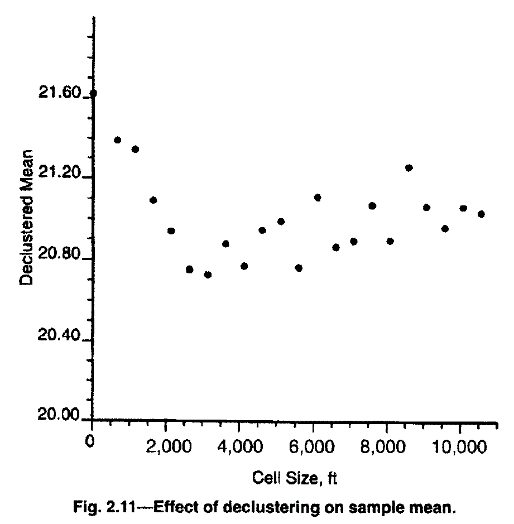
For Linear statistics:

#### Coefficient of variation:

#### Range

### Spatial Data Sets

#### Sample declustering

* + Step 1: Weights for each points are calculated:
  + Step 2: Arithmetic mean of samples after calculating weights:
  + Step 3: What size the subarea should be?
    - Option 1:
    - Option 2: Different subareas should be tried and declustered mean plotted against the subarea size. The size should be chosen where the declustered mean reaches maximum or minimum. E.g. while calculating the porosity usually wells are drilled in the sweet spots, thus there are more sample values of higher porosity. Hence, the subarea where the minimum average porosity is reached is chosen.
  + Since the declustered data is more uniform, it has more variance/standard deviation than the original clustered data.

#### Moving Window Statistics

* + A small window of desired size is chosen and all the samples within that window is used to calculate the local summary statistics e.g. mean, variance, s.d. If *local means and variances* are uniform it is called *homoscedastic* and if there is significant variation it is called *heteroscedastic*.
  + Choosing the right window size and shape: Recommended is a rectangular window and large enough with enough sample points inside, so that two adjacent windows can overlap and have common samples.
  + Several possibilities exist: Local mean and variance can both vary, local mean can stay fairly constant and local variance can vary and vice versa. In many earth science problems, local mean has been observed to be proportional to local standard deviation which is called *proportional effect*.

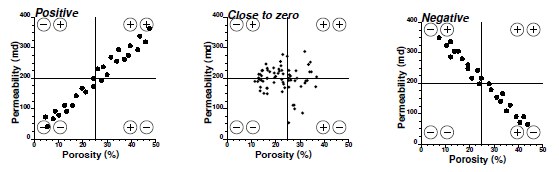
### Bivariate Statistics

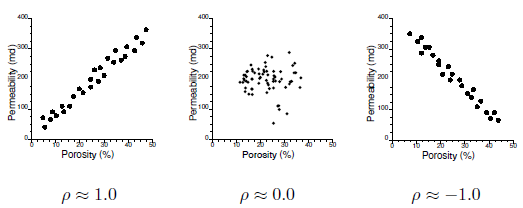
#### Conditional frequency distribution

#### Summary Statistics for Bivariate Distribution

##### Covariance:

For Linear statistics:

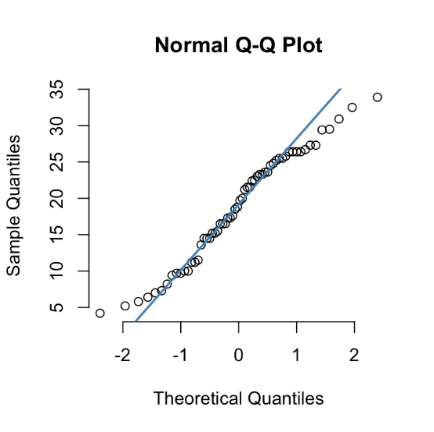




##### Correlation Coefficient

##### Rank correlation coefficient:

Steps: All data are sorted in ascending order and assigned a rank, depending on where it falls where the smallest number gets the lowest rank. With each value assigned a rank, the correlation is calculated for the ranks from the two sets of data, instead of the values themselves.

**Q-Q Plots**: Graphical method of comparing two probability plots by plotting their quantiles against each other.

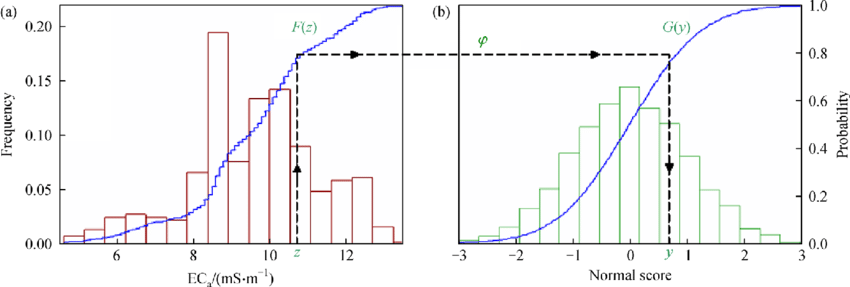
If points lie on x=y line, the two distributions are similar.

If points lie on a line, the two distributions are linearly related.

Often used to compare a dataset against a pre-existing model.

**Normal score transformation**: A given data point is assigned a value which is either exactly, or an approximation, to the expectation of the order statistic of the same rank in a sample of standard normal random variables of the same size as the observed data set. (It preserves the expected value of each data point.)

In the below diagram, point z is transformed to point y. (so for n ‘z’ points, we generate n ‘y’ points)



##### Linear Regression

Establishing relationship between two variables, so as to predict the value of one variable when the value of the other variable is known.

##### Bivariate relationship for spatial data:

See the Chapter 2 on the relationships of spatial data.

## Inferential Statistics

It is a logical extension of the descriptive statistics, where it deals mostly with sample data sets. However, from the characteristics of the samples conclusions can be drawn about the population (from which the sample was taken). Inferential statistics is known to handle this kind of problems.

How ‘samples statistics’ are different from the ‘inferential or population statistics’? Jerry Jensen UPDC notes

What is Random Experiment? (Kelkar Perez)

Conditional probability (Kelkar Perez, Jerry Jensen UPDC notes)

Bayes Theorem (Kelkar Perez, Jerry Jensen UPDC notes)

### Random Experiment

A random function model replaces reality by a set of ‘alternate possible realities’ & all these ‘alternate possible realities’ are realizations of a random function. All these ‘alternate possible realities’ share a few properties i.e. mean, variance and covariance. At any location, there are a series of possible values determined by the random function model and any of these realizations can be the real one with certain probability factor. (So there is always an uncertainty associated with any given number.)

Small case alphabet **z**, denotes simply a ‘deterministic’ variable

Uppercase alphabet **Z**, denotes a Random Variable

A random function is a rule that assigns a realization to the outcome of an experiment.

Each random variable is characterized by a probability distribution function (pdf)

E.g., a Gaussian dist.

[Jamie UPDC notes]

If we are able to develop a completely deterministic model based on evolution of reservoir

Input from Kelkar Perez.

### Sample Space and Events

Sample space is the set of all possible outcomes of a random experiment.

### Probability

#### Laws of probability

#### Conditional Probability

###### Bayes Theorem

### Random Variables

A random variable is a variable whose values are generated by a random experiment on the basis of some probabilistic function.

#### Probability Function

#### Cumulative Distribution Function

#### Bivariate Functions

### Mathematical Expectation

#### Expected Value

##### Characteristics of Expected Value

#### Important parameters of Univariate Distribution

##### Arithmetic mean

##### Variance

For discrete random variable:

Sample standard deviation is denoted by **s** and s.d. of random variable is denoted by **σ**

For continuous random variable:

###### Characteristics of variance:

##### Moment

The n-th moment of a real-valued continuous function f(x) of a real variable about a value c is:

* The first **raw moment** is the mean or [When c=0, is raw moment].
* The 2nd central moment is the variance.

##### Central moment

In probability theory and statistics, a central moment is a moment of a probability distribution of a random variable about the random variable's mean **µ**; that is, it is the expected value of a specified integer power of the deviation of the random variable from the mean.

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Significance of moments (raw, central, normalised) and cumulants (raw, normalised), in connection with named properties of distributions | | | | | |
| **Moment  ordinal** | **Moment** | | | **Cumulant** | |
| **Raw** | **Central** | **Standardized**  where | **Raw** | **Normalized** |
| 1 | Mean | 0 | 0 | Mean | N/A |
| 2 | – | Variance | 1 | Variance | 1 |
| 3 | – | – | Skewness | – | Skewness |
| 4 | – | – | (Non-excess or historical) kurtosis | – | Excess kurtosis |
| 5 | – | – | Hyperskewness | – | – |
| 6 | – | – | Hypertailedness | – | – |
| 7+ | – | – | – | – | – |

##### Standardized moment

The standardized moment of degree k is . () The standardized moment is scale invariant.

The first four standardized moments can be written as:

|  |  |  |
| --- | --- | --- |
| Degree *k* |  | Comment |
| 1 | 0μ ~ 1 = μ 1 σ 1 = E ⁡ [ ( X − μ ) 1 ] ( E ⁡ [ ( X − μ ) 2 ] ) 1 / 2 = μ − μ E ⁡ [ ( X − μ ) 2 ] = 0 {\displaystyle {\tilde {\mu }}\_{1}={\frac {\mu \_{1}}{\sigma ^{1}}}={\frac {\operatorname {E} \left[(X-\mu )^{1}\right]}{(\operatorname {E} \left[(X-\mu )^{2}\right])^{1/2}}}={\frac {\mu -\mu }{\sqrt {\operatorname {E} \left[(X-\mu )^{2}\right]}}}=0} | The first standardized moment is zero, because the first moment about the mean is always zero. |
| 2 | μ ~ 2 = μ 2 σ 2 = E ⁡ [ ( X − μ ) 2 ] ( E ⁡ [ ( X − μ ) 2 ] ) 2 / 2 = 1 {\displaystyle {\tilde {\mu }}\_{2}={\frac {\mu \_{2}}{\sigma ^{2}}}={\frac {\operatorname {E} \left[(X-\mu )^{2}\right]}{(\operatorname {E} \left[(X-\mu )^{2}\right])^{2/2}}}=1} 1 | The second standardized moment is one, because the second moment about the mean is equal to the variance σ2. |
| 3 | μ ~ 3 = μ 3 σ 3 = E ⁡ [ ( X − μ ) 3 ] ( E ⁡ [ ( X − μ ) 2 ] ) 3 / 2 {\displaystyle {\tilde {\mu }}\_{3}={\frac {\mu \_{3}}{\sigma ^{3}}}={\frac {\operatorname {E} \left[(X-\mu )^{3}\right]}{(\operatorname {E} \left[(X-\mu )^{2}\right])^{3/2}}}} | The third standardized moment is a measure of skewness. |
| 4 | μ ~ 4 = μ 4 σ 4 = E ⁡ [ ( X − μ ) 4 ] ( E ⁡ [ ( X − μ ) 2 ] ) 4 / 2 {\displaystyle {\tilde {\mu }}\_{4}={\frac {\mu \_{4}}{\sigma ^{4}}}={\frac {\operatorname {E} \left[(X-\mu )^{4}\right]}{(\operatorname {E} \left[(X-\mu )^{2}\right])^{4/2}}}} | The fourth standardized moment refers to the kurtosis. |

But **coefficient of variation** is (measure of the spread of the distribution) , the reciprocal of the first standardized moment.

##### Moment Generating Function

The Moment generating function of a random variable **X** (of pdf **f(x)**) is

always exists and is equal to 1. However, a key problem with moment-generating functions is that moments and the moment-generating function may not exist, as the integrals need not converge absolutely. By contrast, the characteristic function or Fourier transform always exists (because it is the integral of a bounded function on a space of finite measure), and for some purposes may be used instead.

The nth moment is:

***Relation of MGF to other functions:***

1. Characteristic function is the moment generating function of iX or moment generating function of X evaluated on imaginary axis iX.
2. Cumulant generating function
3. Probability generating function:

##### Cumulant

In probability theory and statistics, the cumulants of a probability distribution are a set of quantities that provide an alternative to the moments of the distribution. The moments determine the cumulants in the sense that *any two probability distributions whose moments are identical will have identical cumulants as well, and similarly the cumulants determine the moments*.

can be obtained by differentiating n times and evaluating the result to 0.

If the moment-generating function *does not exist*, the cumulants can be defined in terms of the *relationship between cumulants and moments*.

***Alternative definition of cumulant generating function:***

Some writers prefer to definite the cumulant generating function as the natural logarithm of the **characteristic function** . Hence the is sometimes called the *second characteristic function*.

* Adv: is well defined for all real values of t even when is not well defined for all real values of t.
* Use in statistics:

For statistically independent random variables X and Y.

***Properties of cumulants:***

* Invariance and equivariance:

* Homogeniety
* Additivity

* A negative result

Normal distribution doesn’t have cumulants for m>3 with non-zero lower order cumulants.

#### Important properties of Bivariate Distribution

##### Expectation

##### Covariance

For C[X,Y]=0 means the X & Y variables are independent.

For X=Y, C[X,Y]=V[X]

##### Some properties of variance, covariance

**Variance** of single variable composed of multiple variables:



**Covariance** of multiple variables



##### Correlation coefficient

* Population Corr. Coeff.:
* Sample Corr. Coeff.: **r**

### Important Distribution Functions

#### Uniform Distribution

#### Normal (Gaussian) Distribution

#### Log-Normal Distribution

### Inference of Parameters (MVUE)

In terms of variables x:

For random variables, true value:

For random variables, estimated value:

Two requirements are to be applied to the error : unbiasedness & variance minimization

#### Unbiasedness

For unbiasedness condition, the expected value of the error should be 0.

#### Minimum variance

The variance of the error has to be kept minimum.

To minimize the error variance, differentiating w.r.t. and :

Solving for :

Or

Solving for :

#### Generalized linear regression

A generalized expression for estimating Y from random variables.

Applying unbiased condition:

Applying the minimum variance condition:

To minimize variance, we need to equate partial derivative of wrt (i=0-n) to 0.

In matrix form all the n equations can be written as:

Value of will be replaced in the above **equation 1** to get the value of