Descriptive Statistics

- Descriptive statistics is the term given to the analysis of data that helps describe, show or summarize data in a meaningful way.

- Descriptive statistics are very important since raw data is hard to interpret, and visualization is not quantitively accurate.

- Descriptive statistics do not, however, allow us to make conclusions beyond the data we have analyzed or reach conclusions regarding any hypotheses we might have made.

Descriptive Statistics (cont.)

- The goal of descriptive statistics is to obtain some partial descriptions of the joint distribution of the data.

- Three aspects of the data are of importance:

1. *Central Tendency*. What is a typical value for each variable?
2. *Dispersion*. How far apart are the individual observations deviate from a central value for a given variable?
3. *Association*. When more than one variable are studied *together,* how does each variable relate to the remaining variables? How are the variables simultaneously related to one another? Are they positively or negatively related?

Population

- A **population** is the collection of all people, plants, animals, or objects of interest about which we wish to make statistical inferences (generalizations).

- The **population** may also be viewed as the collection of all possible random draws from a stochastic model; for example, independent draws from a normal distribution with a given population mean and population variance.

- A **population parameter** is a numerical characteristic of a population.

- In nearly all statistical problems we do not know the value of a parameter because we do not measure the entire population. We use sample data to make an inference about the value of a parameter.

Sample

- A sample is the subset of the population that we actually measure or observe.
- A sample statistic is a numerical characteristic of a sample. A sample statistic estimates the unknown value of a population parameter.
- Information collected from sample statistic is sometimes referred to as Descriptive Statistic.
- *Statistics, as a subject matter, is the science and art of using sample information to make generalizations about populations*

Example: USDA Women's Health Survey

Population

- Intake of the 5 nutrients for all women aged between 25 and 50 in United States.

Sample

- Intake of the 5 nutrients observed from 737 women aged between 25 and 50 in United States.

A Population Parameter

- Average intake of Calcium

A Sample Statistic

- Sample mean of Calcium intake

Notation

- $p = \text{number of variables}, n = \text{number of observations}.$
- $x_{ij} = i\text{-th observation of variable } j.$

- Vector of observations for the j-th variable

$$\boldsymbol{x}^j = \begin{pmatrix} x_{1j} \\ x_{2j} \\ \vdots \\ x_{nj} \end{pmatrix} = (x_{1j}, \quad x_{2j}, \quad \dots, \quad x_{nj})^{\mathrm{T}}$$

Notation (cont.)

- Data matrix (sometimes called design, regressor and model matrix) whose j-th column is the vector of observations for the j-th variable

$$\mathbf{X} = \left(\boldsymbol{x}^{\mathbf{1}}, \ldots, \boldsymbol{x}^{\boldsymbol{p}}\right) = \begin{pmatrix} x_{11} & \cdots & x_{1p} \\ \vdots & \ddots & \vdots \\ x_{n1} & \cdots & x_{np} \end{pmatrix} \in \mathbb{R}^{n \times p}$$

- Each column of $\mathbf{X}$ contains the observations of one variable.
- Each row of $\mathbf{X}$ contains all variables for one subject/observation.

Measures of Central Tendency

- Throughout this lecture, we use μ_j to represent the population mean of the j-th variable and the $\bar{x}_j$ to represent the sample mean based on the observed data for j-th variable.

- The population mean/expectation is the measure of central tendency for the population. The population mean/expectation for the j-th variable is

$$\mu_j = \mathbb{E}(x_{ij})$$

- The population mean can be estimated by the sample mean

$$\bar{x}_j = \frac{1}{n}\sum_{i=1}^{n} x_{ij}$$

Population Mean Vector

- A collection of population means of all variables forms the population mean vector

$$\boldsymbol{\mu} = \begin{pmatrix} \mu_1 \\ \mu_2 \\ \vdots \\ \mu_p \end{pmatrix} = \mathbb{E}\begin{pmatrix} x_{i1} \\ x_{i2} \\ \vdots \\ x_{ip} \end{pmatrix} = \mathbb{E}(\boldsymbol{x_i})$$

Sample Mean Vector

- We can estimate the population mean vector by sample mean vector

$$\bar{\boldsymbol{x}} = \begin{pmatrix} \bar{x}_1 \\ \bar{x}_2 \\ \vdots \\ \bar{x}_p \end{pmatrix} = \begin{pmatrix} \frac{1}{n}\sum_{i=1}^{n} x_{i1} \\ \frac{1}{n}\sum_{i=1}^{n} x_{i2} \\ \vdots \\ \frac{1}{n}\sum_{i=1}^{n} x_{ip} \end{pmatrix} = \frac{1}{n}\sum_{i=1}^{n} \boldsymbol{x_i}$$

Sample Mean is Unbiased

- Sample mean (vector) is an unbiased descriptive statistic of the population mean (vector)

$$\mathbb{E}(\bar{x}_j) = \mathbb{E}\left(\frac{1}{n}\sum_{i=1}^{n} x_{ij}\right) = \frac{1}{n}\sum_{i=1}^{n} \mathbb{E}(x_{ij}) = \frac{1}{n}\sum_{i=1}^{n} \mu_j = \mu_j,$$

$$\text{and } \mathbb{E}(\bar{\boldsymbol{x}}) = \begin{pmatrix} \mathbb{E}(\bar{x}_1) \\ \mathbb{E}(\bar{x}_2) \\ \vdots \\ \mathbb{E}(\bar{x}_3) \end{pmatrix} = \begin{pmatrix} \mu_1 \\ \mu_2 \\ \vdots \\ \mu_p \end{pmatrix} = \boldsymbol{\mu}.$$

Why We Care Bias?

- Statistical bias is defined as the difference between population parameter and the expectation of the estimator
- For example

$$\mu_j - \mathbb{E}(\bar{x}_j).$$

- The expectation of an estimator is a quantity (non-random) that your estimator converges to when sample size is large enough. This is the "best" you can expect from your estimator.

- If the bias is non-zero, there will be a non-vanishing estimation error even if you increase your sample size.

Measures of Dispersion

- A variance measures the degree of dispersion (spread) in a variable's values.

- The population variance of the j-th variable is

$$\sigma_j^2 = \mathrm{Var}(x_{ij}) = \mathbb{E}(x_{ij} - \mu_j)^2 = \mathbb{E}(x_{ij}^2) - \mu_j^2.$$

The population standard deviation of the j-th variable is

$$\sigma_j = \sqrt{\mathrm{Var}(x_{ij})}.$$

Sample Variance

- The population variance σ_j^2 can be estimated by the sample variance

$$s_j^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_{ij}-\bar{x}_j)^2.$$

- The sample standard deviation for the j-th variable is simply the square root of the sample variance, that is, s_j.

- Question: why dividing $n-1$ instead of n?

Why dividing $n-1$?

- $\sum_{i=1}^{n}(x_{ij}-\bar{x}_j)=0$, thus, if we know $n-1$ of the deviations, we can compute the last one.

- This means that there are only $n-1$ freely varying deviations, i.e. $n-1$ degrees of freedom.

- Dividing $n-1$ makes sample variance an unbiased descriptive statistic for population variance

$$\mathbb{E}(s_j^2)=\sigma_j^2.$$

Dividing n

Pros

- From a purely descriptive viewpoint, to divide by n in the definition of the sample variance makes more sense.

Cons

- Biased sample variance

Dividing $n-1$

Pros

- Unbiased sample variance.

Cons

- Not intuitive.

When n is large $n \approx n-1$ and the difference is negligible.

Relation between center and dispersion

- Measures of center and measures of dispersion are best thought of together, in the context of an error function.
- The error function measures how well a single number a represents the entire data set.
- The values of a (if they exist) that minimize the error functions are our measures of center.
- The minimum value of the error function is the corresponding measure of spread.

Mean Squared Error Function

- The mean squared error (MSE) function is defined by

$$\mathrm{MSE}(a) = \frac{1}{n-1}\sum_{i=1}^{n}(x_{ij}-a)^2.$$

- Minimizing MSE with respect to a is equivalent to solving

$$\frac{d}{da}\mathrm{MSE}(a) = \frac{2}{n-1}\sum_{i=1}^{n}(x_{ij}-a) = 0.$$

- MSE is minimized at $a = \bar{x}_j$, the sample mean.
- The minimum value of MSE is s^2, the sample variance.

Measures of Association: Covariance

- The population covariance is a measure of the association between pairs of variables. The population covariance between variables j and k is

$$\sigma_{jk} = \mathbb{E}\{(x_{ij} - \mu_j)(x_{ik} - \mu_k)\}.$$

- The production $(x_{ij} - \mu_j)(x_{ik} - \mu_k)$ is a function of random variables x_{ij} and x_{ik}. Therefore, it is also a random variable and has a population mean.

- Positive population covariance means that the two variables are positively associated (similar to negative).

Population Covariance Matrix

- The population variances and covariances can be collected into the population variance-covariance matrix. This is also known by the name population dispersion matrix.

$$\boldsymbol{\Sigma} = \begin{pmatrix} \sigma_1^2 & \sigma_{12} & \cdots & \sigma_{1p} \\ \sigma_{21} & \sigma_2^2 & \cdots & \sigma_{2p} \\ \vdots & \vdots & \ddots & \vdots \\ \sigma_{p1} & \sigma_{p2} & \cdots & \sigma_p^2 \end{pmatrix} \in \mathbb{R}^{p \times p}.$$

- The population variance-covariance matrix is a symmetric and positive semi-definite (PSD) matrix.

Sample Covariance

- The population covariance between variables *j* and *k* can be estimated by the sample covariance

$$s_{jk} = \frac{1}{n-1}\sum_{i=1}^{n}\left(x_{ij} - \bar{x}_j\right)\left(x_{ik} - \bar{x}_k\right).$$

- $s_{jk} = 0$: **suggests** two variables are uncorrelated (not independence!);
- $s_{jk} > 0$: **suggests** two variables are positively correlated ($j \uparrow$when $k \uparrow$);
- $s_{jk} < 0$: **suggests** two variables are negatively correlated ($j \downarrow$when $k \uparrow$).

Unbiasedness: $\mathbb{E}\left(s_{jk}\right) = \sigma_{jk}$

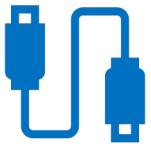
Sample Covariance Matrix

- The population variance-covariance matrix can be estimated by the sample variance-covariance matrix

$$\mathbf{S} = \begin{pmatrix} s_1^2 & s_{12} & \cdots & s_{1p} \\ s_{21} & s_2^2 & \cdots & s_{2p} \\ \vdots & \vdots & \ddots & \vdots \\ s_{p1} & s_{p2} & \cdots & s_p^2 \end{pmatrix} \in \mathbb{R}^{p \times p}.$$

- The sample variance-covariance matrix is also a symmetric matrix.

Unbiasedness: $\mathbb{E}(\mathbf{S}) = \boldsymbol{\Sigma}$ *(implied by element-wise result)*



Measures of Association: Correlation

- The sign of a covariance value is useful to suggest positive, negative correlations or uncorrelation.
- The magnitude of a covariance value is not particularly helpful as it depends on the magnitudes (scales) of the two variables. It does not tell us the strength of the associations.
- To assess the strength of an association, we use correlation values. The population correlation between variables j and k is

$$\rho_{jk} = \frac{\sigma_{jk}}{\sigma_j \sigma_k}.$$

Correlation and Data Transformation

- Correlation of raw data is equivalent to the covariance of Z-score standardized data.
- After Z-score standardization, $\sigma_j = \sigma_k = 1$.
- Correlation is a "standardized" version of covariance.
- Correlation is "scale invariant" as its value does not change if we apply a linear transformation (except multiply by 0) to the variable.

Population Correlation

- The population correlation ρ_{jk} has the same sign with σ_{jk}.
- The population correlation ρ_{jk} lies between -1 and 1

$$-1 \leq \rho_{jk} \leq 1.$$

- $\rho_{jk} = 0$: two variables are uncorrelated;
- ρ_{jk} close to 1: strong positive dependence;
- ρ_{jk} close to -1: strong negative dependence.

Sample Correlation

- The population correlation can be estimated by substituting into the formula the sample covariances and sample standard deviations.
- The sample correlation between variables j and k is

$$r_{jk} = \frac{s_{jk}}{s_j s_k}.$$

- $r_{jk} = 0$: **suggests** two variables are uncorrelated;
- r_{jk} close to 1: **suggests** strong positive dependence;
- r_{jk} close to -1: **suggests** strong negative dependence.

Biased: $\mathbb{E}(r_{jk}) \neq \rho_{jk}$

Asymptotic unbiasedness: $\mathbb{E}(r_{jk}) \to \rho_{jk}$ as $n \to \infty$

Correlation Matrix

- The population correlation matrix and sample correlation matrix are

$$\mathbf{P} = \begin{pmatrix} \rho_1^2 & \rho_{12} & \cdots & \rho_{1p} \\ \rho_{21} & \rho_2^2 & \cdots & \rho_{2p} \\ \vdots & \vdots & \ddots & \vdots \\ \rho_{p1} & \rho_{p2} & \cdots & \rho_p^2 \end{pmatrix} \text{ and } \mathbf{R} = \begin{pmatrix} r_1^2 & r_{12} & \cdots & r_{1p} \\ r_{21} & r_2^2 & \cdots & r_{2p} \\ \vdots & \vdots & \ddots & \vdots \\ r_{p1} & r_{p2} & \cdots & r_p^2 \end{pmatrix}.$$

- The above two matrices are also symmetric & positive semi-definite.

Example 1: USDA Women's Health Survey

- In 1985, the USDA commissioned a study of women's nutrition. Nutrient intake was measured for a random sample of 737 women aged between 25 and 50.

- The following variables were measured:
  1. Calcium (mg), 2. Iron (mg), 3. Protein (g)
  4. Vitamin A (µg), 5. Vitamin C (mg)

Q: Find the descriptive statistics of this dataset.

Sample Mean and Sample Standard Deviation

- Here we calculate sample mean and sample standard deviation for each variable in the dataset.

Variable	Sample mean	Sample SD
Calcium	624.0 mg	397.3 mg
Iron	11.1 mg	6.0 mg
Protein	65.8 mg	30.6 mg
Vitamin A	839.6 µg	1634.0 µg
Vitamin C	78.9 mg	73.6 mg

How to interpret?

Variable	Sample mean	Sample SD
Calcium	624.0 mg	397.3 mg
Iron	11.1 mg	6.0 mg
Protein	65.8 mg	30.6 mg
Vitamin A	839.6 µg	1634.0 µg
Vitamin C	78.9 mg	73.6 mg

However, whether the standard deviations are relatively large or not, will depend on the context of application. Skill in interpreting the statistical analysis depends very much on the researcher's subject matter knowledge.

- Sample mean estimates the central tendency (average amount of nutrient intake).
- Sample std estimates dispersion.
- Note that for Vitamin A & C, the std's are large relative to their respective means.
- This indicates a high variability among women in nutrient intake.

Sample Covariance Matrix

- The sample variance-covariance matrix is copied into the matrix below.

	Calcium	Iron	Protein	Vitamin A	Vitamin C
Calcium	157829.4	940.1	6075.8	102411.1	6701.6
Iron	940.1	35.8	114.1	2383.2	137.7
Protein	6075.8	114.1	934.9	7330.1	477.2
Vitamin A	102411.1	2383.2	7330.1	2668452.4	22063.3
Vitamin C	6701.6	137.7	477.2	22063.3	5416.3

How to interpret?

	Calcium	Iron	Protein	Vitamin A	Vitamin C
Calcium	157829.4	940.1	6075.8	102411.1	6701.6
Iron	940.1	35.8	114.1	2383.2	137.7
Protein	6075.8	114.1	934.9	7330.1	477.2
Vitamin A	102411.1	2383.2	7330.1	2668452.4	22063.3
Vitamin C	6701.6	137.7	477.2	22063.3	5416.3

However, the magnitude of the covariance value can NOT be directly interpreted as the strength of association because it depends on the scales of variables.

- Sample covariance estimates the association between variables.
- All off-diagonal elements in this table are positive, which indicates positive dependency.
- A woman with above-average Calcium intake may also have above-average intake of other nutrients.
- A woman with below-average Iron intake may also have below-average intake of other nutrients.

Sample Correlation Matrix

- The sample correlation matrix is copied into the matrix below.

	Calcium	Iron	Protein	Vitamin A	Vitamin C
Calcium	1.000	0.395	0.500	0.158	0.229
Iron	0.395	1.000	0.623	0.244	0.313
Protein	0.500	0.623	1.000	0.147	0.212
Vitamin A	0.158	0.244	0.147	1.000	0.184
Vitamin C	0.229	0.313	0.212	0.184	1.000

How to interpret?

	Calcium	Iron	Protein	Vitamin A	Vitamin C
Calcium	1.000	0.395	0.500	0.158	0.229
Iron	0.395	1.000	0.623	0.244	0.313
Protein	0.500	0.623	1.000	0.147	0.212
Vitamin A	0.158	0.244	0.147	1.000	0.184
Vitamin C	0.229	0.313	0.212	0.184	1.000

Why these three nutrients are highly correlated?
This can be a good research problem!

- Sample correlation estimates the association between standardized variables.
- All off-diagonal elements in this table are positive, which indicates positive dependency.
- Magnitude indicates strength of dependency.
- High correlation pairs: Calcium – Iron, Calcium – Protein, Iron-Protein

Overall Measures of Dispersion

- Sometimes it is also useful to have an overall measure of dispersion in the data. In this measure, it would be good to include all the variables simultaneously, rather than one at a time.

- The following two quantities are used to measure the dispersion of all variables together

1. Total variance
2. Generalized variance

Total Variance

- Population total variance is defined as the trace of the population variance-covariance matrix

$$\text{trace}(\boldsymbol{\Sigma}) = \sigma_1^2 + \sigma_2^2 + \cdots + \sigma_p^2.$$

- Total variance is the sum of variances for all variables in the dataset.
- Population total variance can be estimated by the trace of $\mathbf{S}$

$$\text{trace}(\mathbf{S}) = s_1^2 + s_2^2 + \cdots + s_p^2.$$

Generalized Variance

- Population generalized variance is defined as the determinant of the population variance-covariance matrix

$$\det(\boldsymbol{\Sigma}) \text{ or } |\boldsymbol{\Sigma}|$$

- Generalized variance also accounts for off-diagonal elements in $\boldsymbol{\Sigma}$ (e.g. covariance effects).
- Population generalized variance can be estimated by the determinant of $\mathbf{S}$

$$\det(\mathbf{S}) \text{ or } |\mathbf{S}|$$