8 STATISTICS:

(Reading Exercises: Yates and Goodman – Chapter 10)

So far, we have focused on theoretical models where knowledge of the relevant probability distribution functions is assumed to be known. Statistics deals with the real world where probability distributions are unknown; all we have are measured or collected data. Given a set of random data, we are interested in determining the sample mean, sample variance and the empirical distribution of the data.

Consider a set of n independent random data samples $\{x_1, x_2, \ldots, x_n\}$, with identical but unknown distributions. We say that the data set consists of n independent and identically distributed (IID) random variables.

Definition: A statistic is any quantity whose value can be calculated from a set of random data samples.

- A statistic is a function, $u(x_1, x_2, \ldots, x_n)$, of the collected random data
- A statistic is a random variable and is denoted by an uppercase letter; a lowercase letter is used to represent the value of the statistic.

Learning outcomes:

You will

- Know the formal definition for sample mean and sample variance; and how to find them from randomly collected data
- Know the formal definition for sample covariance and sample correlation coefficient; and how to find them from randomly collected data
- Empirical distribution functions from randomly collected data
- Understand and be able to create a quantile-quantile (q-q) plot.

8.1 Statistics and their Distributions:

Given a set of random data, we want to determine estimates of the first and second order statistics – mean and variance

Sample Mean – First order statistic

Consider a set of random samples, $X = \{x_1, x_2, \ldots, x_n\}$, consisting of n independent random variables that have identical mean $\mu_{X_i} = \mu_X$ and variance $\sigma^2_{X_i} = \sigma^2_X$. However, the true mean is unknown and needs to be estimated from these samples. A statistic that is often used to estimate the true mean is the sample mean, denoted $\overline{X}$, and defined as

$$\overline{\mathrm{X}} \triangleq \frac{1}{n}\sum_{i=1}^{n} X_i$$

The mean and variance of the sample mean are given, respectively, by

$$\mu_{\bar{X}} = E\left[\bar{X}\right] = \frac{1}{n}\sum_{i=1}^{n} E\left[X_i\right] = \mu_X$$

$$\sigma_{\bar{X}}^2 = E\left[\left(\bar{X} - \mu_{\bar{X}}\right)^2\right] = E\left[\left(\frac{1}{n}\sum_{i=1}^{n} X_i - \frac{1}{n}\sum_{i=1}^{n} E\left[X_i\right]\right)^2\right] = \frac{1}{n}\left(\frac{1}{n}\sum_{i=1}^{n} E\left[\left(X_i - \mu_X\right)^2\right]\right) = \frac{\sigma_X^2}{n}$$

- The mean of the sample mean is equal to the true mean of X.
- The variance of the sample mean reduces to zero as the sample size n, is increased.
- The accuracy of the estimated mean improves with increasing sample size.

Example: It is required to infer whether or not the population of resistors from a production line is within a tolerance range $1000\Omega \pm 10\%$. We pick a random sample of 10 resistors and measure their resistances in ohms, which are provided in the table below

X_1	X_2	X_3	X_4	X_5	X_6	X_7	X_8	X_9	X_{10}
900	1013	939	1062	1017	996	970	1079	1065	1049

Solution: Compute the sample mean

$$\overline{X} = \frac{1}{10}\sum_{j=1}^{10} X_j = 1009\ \Omega$$

The sample mean is $\bar{X} = 1009\ \Omega$ and is, therefore, within the tolerance range $1000\ \Omega \pm 10\%$

Sample Variance

The sample variance is an estimate of the true variance of X.

Definition: Consider a set of random data samples, $X = \{x_1,\ x_2,\ \ldots, x_n\}$, consisting of n IID random variables. The sample variance may be computed in one of two ways:

Mean is known:

$$S_X^2 = \frac{1}{n}\sum_{i=1}^{n}\left(x_i - \mu_X\right)^2$$

Mean is unkown:

$$(4)\ S_X^2 \triangleq \frac{1}{n}\sum_{j=1}^{n}\left(x_j - \bar{X}\right)^2$$

(5)

The sample variance can be expressed in a short-cut formula form as following:

$$S_X^2 \triangleq \frac{1}{n-1}\sum_{j=1}^{n}\left(x_j-\bar{X}\right)^2 = \frac{1}{n-1}\left(\sum_{j=1}^{n}x_j^2 - 2\bar{X}\sum_{j=1}^{n}x_j + \sum_{i=1}^{n}\bar{X}^2\right)$$

$$= \frac{1}{n-1}\left(\sum_{j=1}^{n}x_j^2 - \frac{2}{n}\left(\sum_{j=1}^{n}x_j\right)^2 + \frac{1}{n}\left(\sum_{j=1}^{n}x_j\right)^2\right) = \frac{n\sum_{j=1}^{n}x_j^2 - \left(\sum_{j=1}^{n}x_j\right)^2}{n(n-1)}$$

The sample standard deviation is defined as

$$S_X = \sqrt{\frac{n\sum_{j=1}^{n}X_j^2 - \left(\sum_{j=1}^{n}X_j\right)^2}{n(n-1)}}$$

For the previous example, the sample variance is

$$S_X^2 = \frac{10\sum_{j=1}^{10}X_j^2 - \left(\sum_{j=1}^{10}X_j\right)^2}{10(10-1)} = 58.7\Omega$$

Sample Covariance and Correlation Coefficient:

In order to infer whether two populations have any relationship, we collect random samples $X=\{X_i\}$ and $Y=\{Y_i\}$, $i=1,2,\ldots,n$, from each population and then estimate the covariance and the correlation coefficient. These two statistics will tell us whether the two populations are correlated or not.

Definition: Consider two collected random data sets $X=\{X_i\}$ and $Y=\{Y_i\}$, $i=1,2,\ldots,n$ from two populations. The sample covariance is defined as

$$C_{XY} = \frac{1}{n-1}\sum_{i=1}^{n}\left(x_i-\bar{X}\right)\left(y_i-\bar{Y}\right) = \frac{n\sum_{i=1}^{n}x_iy_i - \left(\sum_{i=1}^{n}x_i\right)\left(\sum_{i=1}^{n}y_i\right)}{n(n-1)}$$

$$C_{XY}=\frac{1}{n-1}\sum_{i=1}^{n}\left(x_i-\bar{X}\right)\left(y_i-\bar{Y}\right)=\frac{\frac{1}{n}\sum_{i=1}^{n}x_iy_i-\sum_{1}^{n}x_i\sum_{1}^{n}y_i}{n\left(n-1\right)}$$

The sample correlation coefficient is defined as

$$\rho_{XY}=\frac{C_{XY}}{S_XS_Y}$$

- $-1\le\rho_{XY}\le1$
- If $\rho_{XY}=0$, then X and Y are uncorrelated or independent.
- If $\rho_{XY}=\pm1$, then X and Y are correlated or linearly related.

Example: Find the correlation coefficient for the data pairs in the table below.

X_i	0.68	0.72	1.27	2.01	2.63	3.06	3.15	4.00	4.03	4.50
Y_i	12.45	9.93	6.64	10.14	8.93	13.34	11.56	16.72	19.62	15.03

Solution:

$$\bar{X}=\frac{1}{10}\sum_{i=1}^{10}x_i=2.6;\quad \bar{Y}=\frac{1}{10}\sum_{i=1}^{10}y_i=12.44$$

$$S_X^2=\frac{10\sum_{i=1}^{n}x_i^2-\left(\sum_{i=1}^{n}x_i\right)^2}{10\left(10-1\right)}=1.9475\Rightarrow S_X=1.3955$$

$$S_Y^2=\frac{10\sum_{i=1}^{n}y_i^2-\left(\sum_{i=1}^{n}y_i\right)^2}{10\left(10-1\right)}=3.88^2\Rightarrow S_Y=3.88$$

$$C_{XY}=\frac{10\sum_{i=1}^{10}x_iy_i-\sum_{i=1}^{10}x_i\sum_{i=1}^{10}y_i}{10\left(10-1\right)}=3.864$$

$$\rho_{XY}=\frac{C_{XY}}{S_XS_Y}=\frac{3.864}{1.395\times3.88}=0.714$$

- There is a strong correlation between the two populations

8.2 Empirical Distribution Functions:

Suppose a random variable, X, has a CDF $F_X(x)$, which we do not know. However, we have a number of IID experimental data samples $X = \{x_1, x_2, \ldots, x_n\}$.

Definition: The empirical CDF is defined as

$$\tilde{F}_X(x) = \frac{\text{Number of samples } x_1, x_2, \ldots, x_n \text{ no greater than } x}{n}$$

Example: Obtain the empirical distribution of the resistance samples in the table below.

X_1	X_2	X_3	X_4	X_5	X_6	X_7	X_8	X_9	X_{10}
900	1013	939	1062	1017	996	970	1079	1065	1049

Solution: List the samples in ascending order, then compute $\tilde{F}_X(x)$

y_1	y_2	y_3	y_4	y_5	y_6	y_7	y_8	y_9	y_{10}
900	939	949	970	1013	1017	1049	1065	1070	1079
$\tilde{F}(x) = 0.1$	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9	1.0

8.3 Quantile-Quantile (Q-Q) Plot:

When we collect data samples, we often want to know the probability distribution. A quantile-quantile (Q-Q) plot is used to determine whether a set of experimental data fits some specified theoretical distribution. Quantile-Quantile plots are an excellent graphical tool for comparing a sample data set to a theoretical distribution, most often, the standard Gaussian model.

Procedure for Computing Quantiles:

Consider the case where we collect a set of n experimental data samples, $X_1, X_2, \ldots, X_n$, and we want to see how well this data set fits a standard Gaussian CDF function $F_X(x)$. The following are the steps for computing the quantiles:

- **Data or Sample Quantile:**
  The sample quantile is obtained by sorting the values in the data sample set into ascending order and denoting them as $X_{\min} = X_{(1)}, X_{(2)}, \ldots, X_{(n)} = X_{\max}$. The ordered data set is known as the sample quantile.

  Example: The table below shows a data set and its sample quantile.

Data Set	$X_i = 50, 76, 92, 83, 105, 102, 109, 106, 91, 110, 89$
Sample quantile	$X_{(i)} = 50, 76, 83, 89, 91, 92, 102, 105, 106, 109, 110$

- **Theoretical Quantile:**
  - **Definition:** The p^{th} quantile is a number q_p, such that

$$P\left[X \le q_p\right] = p$$

- We divide the interval $[0,n]$ into n equal bins and assume that the ordered data is in the middle (half way between) the bins. In that case, the n order statitics (the k^{th} percentile will be located between adjacent points $\frac{k}{n}$ and $\frac{k+1}{n}$, $k = 0,1,\cdots,n$. We expect the ordered values to be in the intervals $\left(\frac{0}{n},\frac{1}{n}\right)$, $\left(\frac{1}{n},\frac{2}{n}\right)$, $\cdots$, $\left(\frac{n-1}{n},1\right)$. Therefore, for $i = 1,2,\ \ldots,\ n$, we compute the percentiles $\frac{i-0.5}{n}$.
- Define the i^{th} quantile z_i, such that
-

$$F_Z(z_i) = P\left[Z \le z_i\right] = \frac{i-0.5}{n},\quad i = 1,2,\cdots,n$$

- If we are testing for a standard Gaussian distribution, we would have

$$F_Z(z_i) = \frac{1}{\sqrt{2\pi}}\int_{-\infty}^{z_i}\exp\left(-\frac{x^2}{2}\right)dx = 1 - Q(z_i) = \frac{i-0.5}{n} \Rightarrow Q(z_i) = 1 - \frac{i-0.5}{n}$$

- We then solve for z_i by checking, from the Q-function table, for the argument that satisfies the above equation. The theoretical quantile is the set of values of z_i, (also known as $z\text{-}scores$).

For the data collected above, the sample and theoretical quantiles are summarized below.

i	$X_{(i)}$	$(i-0.5/n)$	$1-\left(\frac{i-0.5}{n}\right)$	z_i
1	50	0.05	0.95	–1.65
2	76	0.15	0.85	–1.05
3	83	0.25	0.75	–0.675
4	89	0.35	0.65	–0.375
5	91	0.45	0.55	–0.125
6	92	0.55	0.45	0.125
7	102	0.65	0.35	0.375
8	105	0.75	0.25	0.675
9	106	0.85	0.15	1.05

10	109	0.95	0.05	1.65

- If we are testing for a regular Gaussian distribution, we would have to compute the sample mean and sample standard deviation. We find z_i from the CDF
- $$F_Z[z_i] = 1 - Q\left(z_i = \frac{X_{(i)} - \bar{X}}{S_X}\right) = \frac{i - 0.5}{n} \Rightarrow Q(z_i) = 1 - \frac{i - 0.5}{n}$$

  Finally, we express z_i as

$$X_{(i)} = S_X z_i + \bar{X}$$

Definition: A Q-Q plot is a plot of the sample quantile $X_{(i)}$, versus the theoretical quantile z_i, or vice versa.

- If the Q-Q plot is approximately linear, then our data are approximately Gaussian distributed.

Exercise: (a) Sketch the Q-Q-plot and conclude whether or not our data is coming from a Guassian population.

Example: The table below gives the sample quantiles and theoretical quantiles for a standard Gaussian distribution.

Sample Quantile $X_{(i)}$	i	$(i-0.5)/n$	Theoretical Quantile $z_i = F^{-1}((i-0.5)/n)$	$X_{(i)} = S_X z_i + \bar{X}$ (Regular Gaussian)
-1.96	1	0.1	-1.28	
-0.78	2	0.3	-0.52	
0.31	3	0.5	0.00	
1.15	4	0.7	0.52	
1.62	5	0.9	1.28	

Example: We are given the following random data set:

$$\mathbf{X} = \{25.0,\ 25.0,\ 27.7,\ 25.9,\ 25.9,\ 21.7,\ 22.8,\ 28.9,\ 26.4,\ 22.4\}$$

Compute $\bar{X}$ and S_X, fill in the corresponding values in the last column of the table below.

i	$X_{(i)}$	$p_i = (i - 0.5/n)$	z_i	$X_{(i)} = S_X z_i + \bar{X}$
1	21.7	0.05	–1.65	
2	22.4	0.15	–1.05	
3	22.8	0.25	–0.675	
4	25.0	0.35	–0.375	
5	25.0	0.45	–0.125	
6	25.9	0.55	0.125	
7	25.9	0.65	0.375	
8	26.4	0.75	0.675	
9	27.7	0.85	1.05	
10	28.9	0.95	1.65	

Plot $X_{(i)}$ versus z_i and $X_{(i)} = S_X z_i + \bar{X}$ on the same graph. How do they relate?

Monte Carlo Simulations:

A Monte Carlo simulation is a method for carrying out complex computer simulations of systems undergoing random perturbations. Suppose we have a system described in terms of input-output equations. The input or system parameters have random aspects (random noise, unknown component values, etc.). We can obtain a probabilistic description of the system by modeling the random quantities with random number generators on a computer, repeatedly solving the system equations as we draw different realizations for the random quantities and computing some statistics of the desired responses or outputs.