

# 1 Random Sampling Environment

$$\{X_1, \dots, X_T\}$$

are independent and identically distributed (iid) random variables with unknown pdf  $p(x)$ .

Observed Sample:

$$\{X_1 = x_1, \dots, X_T = x_T\}$$

are observations generated by the random sample

Descriptive Statistics

Data summaries (statistics) to describe certain features of the data and to learn about the unknown pdf,  $p(x)$ .

# Histograms

Goal: Describe the shape of the distribution of the data

Histogram Construction:

1. Order data from smallest to largest values

$\text{min} = \text{smallest value}$

$\text{max} = \text{largest value}$

$\text{range} = \text{max} - \text{min}$

2. Divide range into  $N$  equally spaced bins

$[- | - | - | \cdots | - | - | -]$

3. Count number of observations in each bin
4. Create bar chart (optionally normalize area to equal 1)

## R Functions

Function	Description
<code>sort()</code>	sort elements of data vector
<code>min()</code>	compute minimum value of data vector
<code>max()</code>	compute maximum value of data vector
<code>range()</code>	compute min and max of a data vector
<code>hist()</code>	compute histogram
<code>density()</code>	compute smoothed histogram

## Empirical Quantiles

Percentiles:

For  $\alpha \in [0, 1]$ , the  $\alpha^{th}$  percentile of a sample of data is the data value  $\hat{q}_\alpha$  such that  $\alpha \cdot 100\%$  of the data are less than  $\hat{q}_\alpha$ .

Quartiles

$\hat{q}_{.25}$  = first quartile

$\hat{q}_{.50}$  = second quartile (median)

$\hat{q}_{.75}$  = third quartile

$\hat{q}_{.75} - \hat{q}_{.25}$  = interquartile range (IQR)

## R functions

`quantile()`

`median()`

`summary()`

## Sample Statistics

Plug-In Principle: Estimate population quantities using sample statistics

Sample Average (Mean)

$$\frac{1}{T} \sum_{t=1}^T x_t = \bar{x} = \hat{\mu}_x$$

Sample Variance

$$\frac{1}{T-1} \sum_{t=1}^T (x_t - \bar{x})^2 = s_x^2 = \hat{\sigma}_x^2$$

Sample Standard Deviation

$$\sqrt{s_x^2} = s_x = \hat{\sigma}_x$$

Sample Skewness

$$\frac{1}{T-1} \sum_{t=1}^T (x_t - \bar{x})^3 / s_x^3 = \widehat{skew}$$

Sample Kurtosis

$$\frac{1}{T-1} \sum_{t=1}^T (x_t - \bar{x})^4 / s_x^4 = \widehat{kurt}$$

Sample Excess Kurtosis

$$\widehat{kurt} - 3$$

## R Functions

Function	Package	Description
<code>mean()</code>	base	compute mean of data vector
<code>colMeans()</code>	base	compute column means of matrix
<code>var()</code>	base	compute sample variance of data
<code>summary()</code>		compute column variances of matrix
<code>sd()</code>		compute sample standard deviation
<code>skewness()</code>	TSA	compute sample skewness of data
<code>kurtosis()</code>	TSA	compute sample excess kurtosis of data
<code>summary()</code>		compute quantiles, mean, min, max

## Empirical Cumulative Distribution Function

Recall, the CDF of a random variable  $X$  is

$$F_X(x) = \Pr(X \leq x)$$

The empirical CDF of a random sample is

$$\begin{aligned}\hat{F}_X(x) &= \frac{1}{n}(\#x_i \leq x) \\ &= \frac{\text{number of } x_i \text{ values } \leq x}{\text{sample size}}\end{aligned}$$

How to compute and plot  $\hat{F}_X(x)$  for a sample  $\{x_1, \dots, x_n\}$

- Sort data from smallest to largest values:  $\{x_{(1)}, \dots, x_{(n)}\}$
- Plot  $\hat{F}_X(x)$  against sorted data  $\{x_{(1)}, \dots, x_{(n)}\}$

## Comparing Empirical CDF to Normal Distribution

**Question:** Does observed data come from a normal distribution?

- Standardize data to have zero mean and variance one

$$z_i = \frac{x_i - \bar{x}}{s_x}$$

- Sort standardized data from smallest to largest values:  $\{z_{(1)}, \dots, z_{(n)}\}$
- Compute standard normal CDF at each sorted value:  $\Phi(z_{(i)})$
- Plot  $\hat{F}_X(x)$  and  $\Phi(z_{(i)})$  against sorted data

## Quantile-Quantile (QQ) Plots

A QQ plot is useful for comparing your data with the quantiles of a distribution (usually the normal distribution) that you think is appropriate for your data. You interpret the QQ plot in the following way:

- If the points fall close to a straight line, your conjectured distribution is appropriate
- If the points do not fall close to a straight line, your conjectured distribution is not appropriate and you should consider a different distribution

## R functions

Function	Description
<code>qqnorm()</code>	QQ-plot against normal distribution
<code>qqline()</code>	draw straight line on QQ-plot

## Outliers

- Extremely large or small values are called “outliers”
- Outliers can greatly influence the values of common descriptive statistics. In particular, the sample mean, variance, standard deviation, skewness and kurtosis
- Percentile measures are more robust to outliers: outliers do not greatly influence these measures

Moderate Outlier

$$\hat{q}_{.75} + 1.5 \cdot IQR < x < \hat{q}_{.75} + 3 \cdot IQR$$
$$\hat{q}_{.25} - 3 \cdot IQR < x < \hat{q}_{.25} - 1.5 \cdot IQR$$

Extreme Outlier

$$x > \hat{q}_{.75} + 3 \cdot IQR$$
$$x < \hat{q}_{.25} - 3 \cdot IQR$$

## Boxplots

A box plot displays the locations of the basic features of the distribution of one-dimensional data—the median, the upper and lower quartiles, outer fences that indicate the extent of your data beyond the quartiles, and outliers, if any.

R function

```
boxplot()
```

## Bivariate Descriptive Statistics

$$\{(X_1, Y_1), (X_2, Y_2), \dots, (X_T, Y_T)\}$$

random sample of size  $T$  with realized values

$$\{(x_1, y_1), (x_2, y_2), \dots, (x_T, y_T)\}$$

Scatterplot

XY plot of bivariate data

R functions: `plot()`, `pairs()`

Sample Covariance

$$\frac{1}{T-1} \sum_{t=1}^T (x_t - \bar{x})(y_t - \bar{y}) = s_{xy} = \hat{\sigma}_{xy}$$

Sample Correlation

$$\frac{s_{xy}}{s_x s_y} = r_{xy} = \hat{\rho}_{xy}$$

## R functions

Function	Description
<code>var</code>	compute sample variance matrix
<code>cor</code>	compute sample correlation matrix

## Time Series Descriptive Statistics

$$\{X_1, \dots, X_T\}$$

is a covariance stationary time series with unknown pdf  $p(x)$  with realized values

$$\{x_1, \dots, x_T\}$$

Recall,

$$\begin{aligned} E[X_t] &= \mu \text{ indep of } t \\ \text{var}(X_t) &= \sigma^2 \text{ indep of } t \\ \text{cov}(X_t, X_{t-j}) &= \gamma_j \text{ indep of } t \\ \text{cor}(X_t, X_{t-j}) &= \rho_j \text{ indep of } t \end{aligned}$$

Sample Autocovariance

$$\frac{1}{T-1} \sum_{t=j+1}^T (x_t - \bar{x})(x_{t-j} - \bar{x}) = \hat{\gamma}_j, \quad j = 1, 2, \dots$$

Sample Autocorrelation

$$\hat{\rho}_j = \frac{\hat{\gamma}_j}{\hat{\sigma}^2}, \quad j = 1, 2, \dots$$

Sample Autocorrelation Function (SACF)

Plot  $\hat{\rho}_j$  against  $j$

## R functions

Function	Description
<code>acf()</code>	compute and plot sample autocorrelations
<code>acf.plot()</code>	plot sample autocorrelations