UQ, STAT2201, 2017,
Lecture 5
Unit 4 – Joint Distributions
and Unit 5 – Descriptive Statistics.

Unit 4 - Joint Probability Distributions

A **joint probability distribution** – two (or more) random variables in the experiment.

In case of two, referred to as **bivariate probability distribution**.

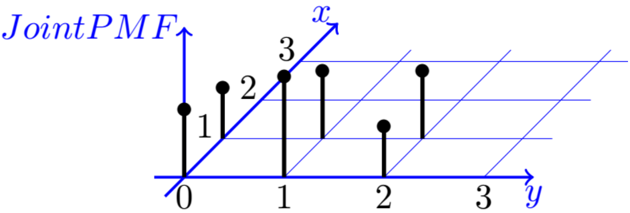
A **joint probability mass function** for discrete random variables X and Y, denoted as $p_{XY}(x, y)$, satisfies the following properties:

(1) $p_{XY}(x, y) \geq 0$ for all x, y.

(2) $p_{XY}(x, y) = 0$ for (x, y) not in the range.

(3) $\sum \sum p_{XY}(x, y) = 1$, where the summation is over all (x, y) in the range.

(4) $p_{XY}(x, y) = P(X = x,\ Y = y)$.


$JointPMF$
x
3
2
1
0
1
2
3
y

Example: Throw two independent dice and look at the,

$$X \equiv \text{Sum}, \qquad Y \equiv \text{Product}.$$

A **joint probability density function** for continuous random variables X and Y, denoted as $f_{XY}(x,y)$, satisfies the following properties:

(1) $f_{XY}(x,y) \geq 0$ for all x, y.

(2) $f_{XY}(x,y) = 0$ for (x,y) not in the range.

(3) $\int\limits_{-\infty}^{\infty} \int\limits_{-\infty}^{\infty} f_{XY}(x,y)\ dx\ dy = 1$.

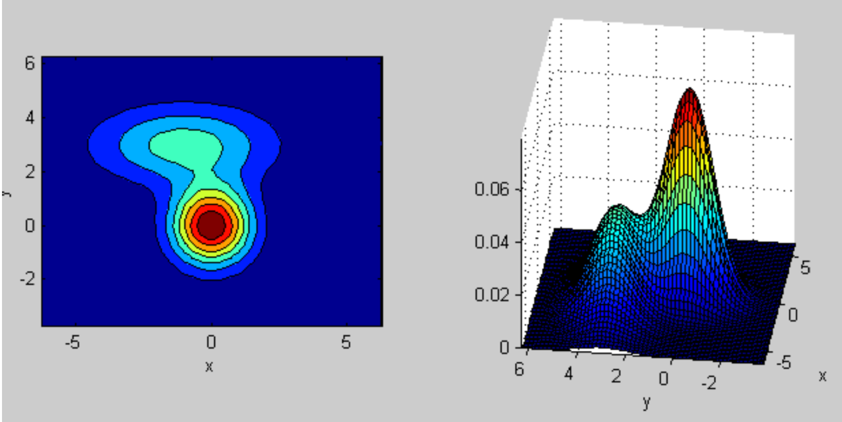
(4) For small Δx, Δy:

$$f_{XY}(x,y)\,\Delta x\,\Delta y \approx P\Big((X,Y) \in [x, x+\Delta\,x) \times [y, y+\Delta\,y)\Big).$$

(5) For any region R of two-dimensional space,

$$P\Big((X,Y) \in R\Big) = \iint\limits_{R} f_{XY}(x,y)\ dx\ dy.$$

e.g. Height and Weight.


6
4
2
0
-2
y
-5
0
5
x
0.06
0.04
0.02
0
6
4
2
0
-2
y
5
0
-5
x

A **joint probability density function** can also be defined for $n > 2$ random variables (as can be a **joint probability mass function**). The following needs to hold:

(1) $f_{X_1 X_2 \ldots X_n}(x_1, x_2, \ldots, x_n) \geq 0.$

(2) $\int\limits_{-\infty}^{\infty} \int\limits_{-\infty}^{\infty} \ldots \int\limits_{-\infty}^{\infty} f_{X_1 X_2 \ldots X_n}(x_1, x_2, \ldots, x_n) dx_1 \ dx_2 \ldots dx_n = 1.$

The **marginal distributions** of X and Y as well as **conditional distributions** of X given a specific value $Y = y$ and vice versa can be obtained from the joint distribution.

If the random variables X and Y are independent, then $f_{XY}(x, y) = f_X(x)\, f_Y(y)$ and similarly in the discrete case.

Generalized Moments

The **expected value of a function of two random variables** is:

$$E\Big[h(X,Y)\Big] = \iint h(x,y) f_{XY}(x,y)\,dx\,dy \qquad \text{for } X, Y \text{ continuous.}$$

The **covariance** is a common measure of the relationship between two random variables (say X and Y). It is denoted as $\text{cov}(X, Y)$ or σ_{XY}, and is given by:

$$\sigma_{XY} = E\Big[(X - \mu_X)(Y - \mu_Y)\Big] = E(XY) - \mu_X \, \mu_Y .$$

The covariance of a random variable with itself is its variance.

The **correlation** between the random variables X and Y, denoted as ρ_{XY}, is

$$\rho_{XY} = \frac{\text{cov}(X, Y)}{\sqrt{V(X)V(Y)}} = \frac{\sigma_{XY}}{\sigma_X \sigma_Y}.$$

For any two random variables X and Y, $-1 \leq \rho_{XY} \leq 1$.

If X and Y are independent random variables then $\sigma_{XY} = 0$ and $\rho_{XY} = 0$.

The opposite case does not always hold: In general $\rho_{XY} = 0$ does not imply independence.

For jointly Normal random variables it does.

In any case, if $\rho_{XY} = 0$ then the random variables are called **uncorrelated**.

When considering several random variables, it is common to consider the (symmetric) **Covariance Matrix**, Σ with $\Sigma_{i,j} = \text{cov}(X_i, X_j)$.

Bivariate Normal

The **probability density function** of a **bivariate normal distribution** is

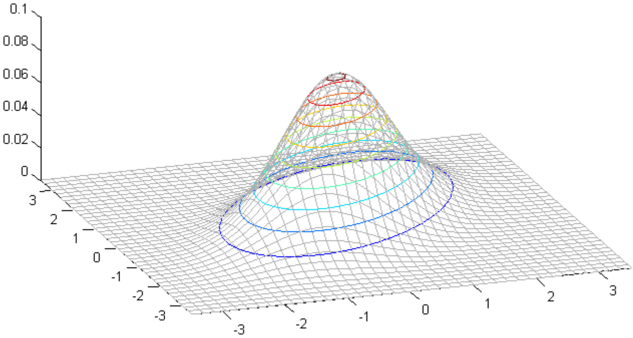
$$f_{XY}(x, y; \sigma_X, \sigma_Y, \mu_X, \mu_Y, \rho) = \frac{1}{2\pi\sigma_X\sigma_Y\sqrt{1-\rho^2}}$$

$$\times \exp\left\{ \frac{-1}{2(1-\rho^2)} \left[\frac{(x-\mu_X)^2}{\sigma_X^2} - \frac{2\rho(x-\mu_X)(y-\mu_Y)}{\sigma_X\sigma_Y} + \frac{(y-\mu_Y)^2}{\sigma_Y^2} \right] \right\}$$

for $-\infty < x < \infty$ and $-\infty < y < \infty$.

The parameters are

$$\sigma_X > 0,\ \sigma_Y > 0,$$
$$-\infty < \mu_X < \infty,\ -\infty < \mu_Y < \infty,$$
$$-1 < \rho < 1.$$


0.1
0.08
0.06
0.04
0.02
0
3
2
1
0
-1
-2
-3
-3
-2
-1
0
1
2
3

Linear Combinations of Random Variables

Given random variables $X_1, X_2, \ldots, X_n$ and constants $c_1, c_2, \ldots, c_n$, the (scalar) **linear combination**

$$Y = c_1 X_1 + c_2 X_2 + \cdots + c_n X_n$$

is often a random variable of interest.

The mean of the linear combination is the linear combination of the means,

$$E(Y) = c_1 E(X_1) + c_2 E(X_2) + \cdots + c_n E(X_n).$$

This holds even if the random variables are not independent.

The variance of the linear combination is as follows:

$$V(Y) = c_1^2 V(X_1) + c_2^2 V(X_2) + \cdots + c_n^2 V(X_n) + 2 \sum_{i<j} \sum c_i c_j \text{cov}(X_i, X_j)$$

If $X_1, X_2, \ldots, X_n$ are **independent** (or even if they are just uncorrelated).

$$V(Y) = c_1^2 V(X_1) + c_2^2 V(X_2) + \cdots + c_n^2 V(X_n).$$

Example: Derive Mean and variance of the Binomial Distribution.

Linear Combinations of Normal Random Variables

Linear combinations of Normal random variables remain Normally distributed:

If $X_1, \ldots, X_n$ are jointly Normal then,

$$Y \;\sim\; \text{Normal}\Big(E(Y), V(Y)\Big).$$

i.i.d. Random Samples

A collection of random variables, $X_1, \ldots, X_n$ is said to be **i.i.d.**, or **independent and identically distributed** if they are mutually independent and identically distributed.

The (n - dimensional) joint probability density is a product of the individual densities.

In the context of statistics, a **random sample** is often modelled as an i.i.d. vector of random variables. $X_1, \ldots, X_n$.

An important linear combination associated with a random sample is the **sample mean**:

$$\overline{X} = \frac{\sum_{i=1}^{n} X_i}{n} = \frac{1}{n}X_1 + \frac{1}{n}X_2 + \ldots + \frac{1}{n}X_n.$$

If X_i has mean μ and variance σ^2 then sample mean (of an i.i.d. sample) has,

$$E(\overline{X}) = \mu, \qquad V(\overline{X}) = \frac{\sigma^2}{n}.$$

Unit 5 – Descriptive Statistics

Descriptive statistics deals with summarizing **data** using numbers, qualitative summaries, tables and graphs.

There are many possible data configurations...

Single sample: $x_1, x_2, \ldots, x_n$.

Single sample over time (time series): $x_{t_1}, x_{t_2}, \ldots, x_{t_n}$ with $t_1 < t_2 < \ldots < t_n$.

Two samples: $x_1, \ldots, x_n$ and $y_1, \ldots, y_m$.

Generalizations from two samples to k samples (each of potentially different sample size, $n_1, \ldots, n_k$).

Observations in tuples: $(x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n)$.

Generalizations from tuples to vector observations (each vector of length ℓ),

$$(x_1^1, \ldots, x_1^\ell), \ldots, (x_n^1, \ldots, x_n^\ell).$$

Individual **variables** may be **categorical** or **numerical**.

Categorical variables may be **ordinal** meaning that they be sorted (e.g. "a", "b", "c", "d"), or not ordinal (e.g. "cat", "dog", "fish").

A Statistic

A **statistic** is a quantity computed from a sample (assume here a single sample $x_1, \ldots, x_n$).

The **sample mean**: $\bar{x} = \dfrac{x_1 + \cdots + x_n}{n} = \dfrac{\sum\limits_{i=1}^{n} x_i}{n}$.

The **sample variance**: $s^2 = \frac{\sum_{i=1}^{n}(x_i - \bar{x})^2}{n-1} = \frac{\sum_{i=1}^{n} x_i^2 - n\overline{x}^2}{n-1}$.

The **sample standard deviation**: $s = \sqrt{s^2}$.

Order Statistics

Order statistics: Sort the sample to obtain the sequence of sorted observations, denoted $x_{(1)}, \ldots, x_{(n)}$ where, $x_{(1)} \leq x_{(2)} \leq \ldots \leq x_{(n)}$.

Some common order statistics:

The **minimum** $\min(x_1, \ldots, x_n) = x_{(1)}$.

The **maximum** $\max(x_1, \ldots, x_n) = x_{(n)}$.

The **median**

$$\text{median} = \begin{cases} x_{(\frac{n+1}{2})} & \text{if } n \text{ is odd,} \\ \frac{1}{2}\left(x_{(\frac{n}{2})} + x_{(\frac{n}{2}+1)}\right) & \text{if } n \text{ is even.} \end{cases}$$

The median is the 50'th percentile and the 2nd quartile (see below).

The qth **quantile** ($q \in [0, 1]$) or alternatively the $p = 100q$ **percentile** (measured in percents instead of a decimal), is the observation such that p percent of the observations are less than it and $(1 - p)$ percent of the observations are greater than it.

The first **quartile**, denoted $Q1$ is the 25th percentile. The second quartile ($Q2$) is the median. The **third quartile**, denoted $Q3$ is the 75th percentile. Thus half of the observations lie between $Q1$ and $Q3$. In other words, the quartiles break the sample into 4 quarters. The difference $Q3 - Q1$ is the **interquartile range**.

The **sample range** is $x_{(n)} - x_{(1)}$.

Interlude: The quantile of a probability distribution?

Given $\alpha \in [0,1]$: What is x such that $P(X \leq x) = \alpha$,

$$F(x) = \alpha.$$

Or,

$$\int_{-\infty}^{x} u\, du = \alpha.$$

To find the quantile, solve the equation for x.

Visualization

Histogram (with Equal Bin Widths):

(1) Label the bin (class interval) boundaries on a horizontal scale.

(2) Mark and label the vertical scale with **frequencies** or **counts**.

(3) Above each bin, draw a rectangle where height is equal to the frequency (or count).

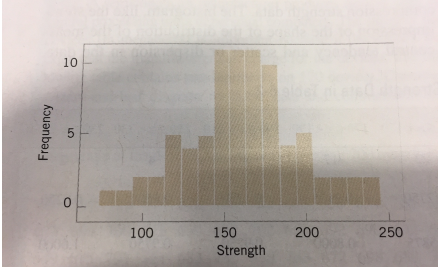
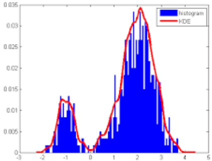
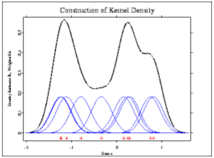


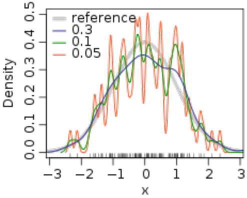
FIGURE 6-8 A histogram of the compressive strength data with 17 bins.

A **Kernel Density Estimate** (KDE) is a way to construct a **Smoothed Histogram**.

While construction is not as straightforward as steps (1)–(3) above, automated tools can be used.







Both the histogram and the KDE are not unique in the way they summarize data.

With these methods, different settings (e.g. number of bins in histograms or **bandwidth** in a KDE) may yield different representations of the same data set.

Nevertheless, they are both very common, sensible and useful visualisations of data.

The **box plot** is a graphical display that simultaneously describes several important features of a data set:

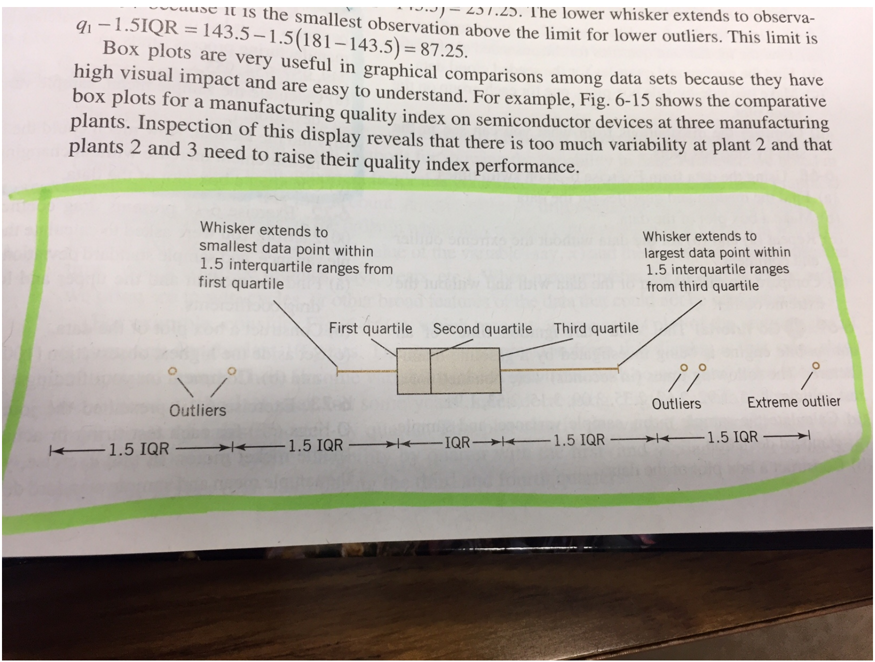
Centre.

Spread.

Departure from symmetry.

Identification of unusual observations or **outliers**.

It is often common to plot several box plots next to each other for comparison.

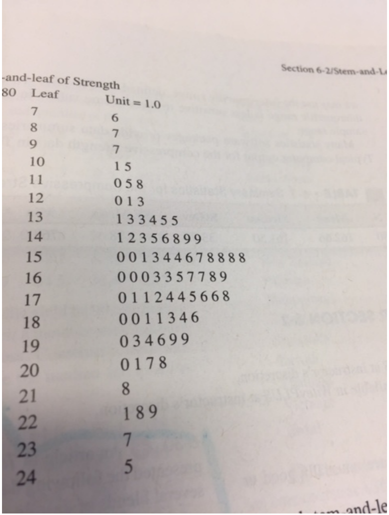


5) = 237.25. The lower whisker extends to observa- cause it is the smallest observation above the limit for lower outliers. This limit is $q_1 - 1.5\text{IQR} = 143.5 - 1.5(181 - 143.5) = 87.25$.

Box plots are very useful in graphical comparisons among data sets because they have high visual impact and are easy to understand. For example, Fig. 6-15 shows the comparative box plots for a manufacturing quality index on semiconductor devices at three manufacturing plants. Inspection of this display reveals that there is too much variability at plant 2 and that plants 2 and 3 need to raise their quality index performance.

Whisker extends to smallest data point within 1.5 interquartile ranges from first quartile
Whisker extends to largest data point within 1.5 interquartile ranges from third quartile
First quartile
Second quartile
Third quartile
Outliers
Outliers
Extreme outlier
1.5 IQR
1.5 IQR
IQR
1.5 IQR
1.5 IQR

An anachronistic, but useful way for summarising small data-sets is the **stem and leaf diagram**.



Section 6-2/Stem-and-L

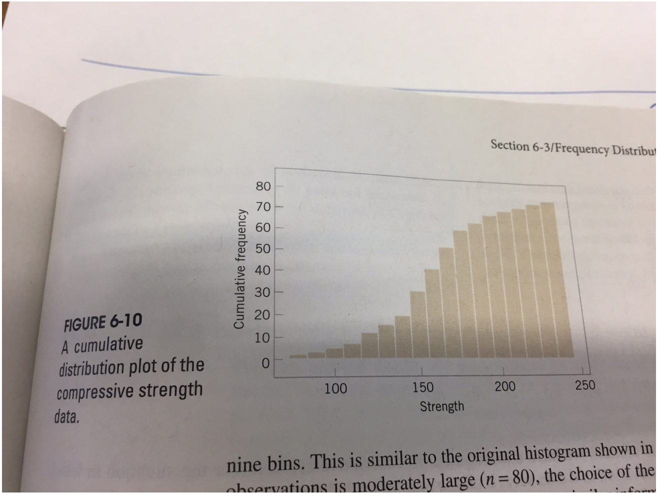
-and-leaf of Strength

80 Leaf Unit = 1.0

7	6
8	7
9	7
10	15
11	058
12	013
13	133455
14	12356899
15	001344678888
16	0003357789
17	0112445668
18	0011346
19	034699
20	0178
21	8
22	189
23	7
24	5

em-and-le

In a **cumulative frequency plot** the height of each bar is the total number of observations that are less than or equal to the upper limit of the bin.



Section 6-3/Frequency Distribut

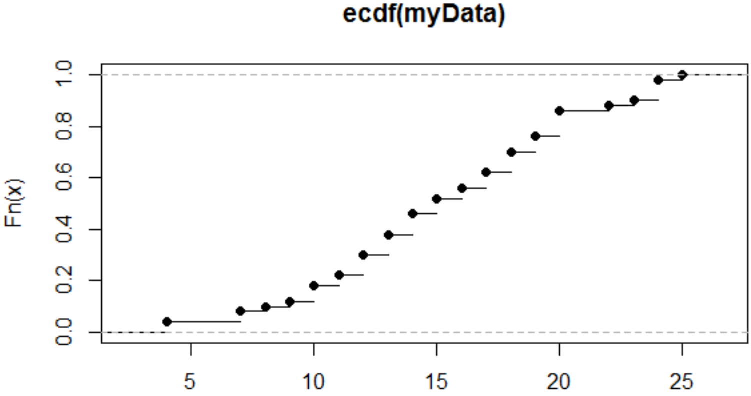
FIGURE 6-10 A cumulative distribution plot of the compressive strength data.

nine bins. This is similar to the original histogram shown in
observations is moderately large ($n = 80$), the choice of the

The **Empirical Cumulative Distribution Function** (ECDF) is,

$$\hat{F}(x) = \frac{1}{n}\sum_{i=1}^{n} \mathbf{1}\{x_i \le x\}.$$

Here $\mathbf{1}\{\cdot\}$ is the **indicator function**. The ECDF is a function of the data, defined for all x.



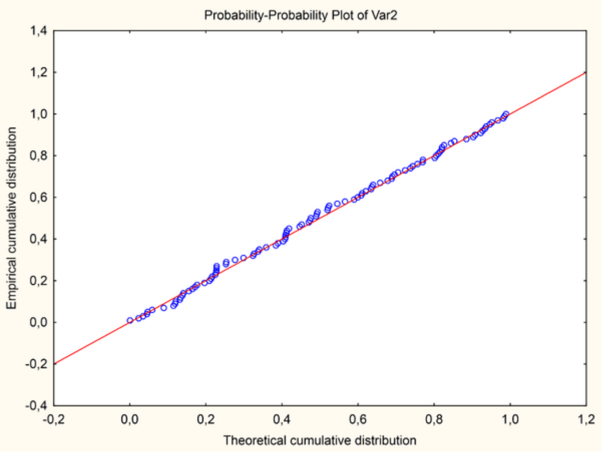
Given a **candidate distribution** with CDF $F(x)$, a **probability plot** is a plot of the ECDF (or sometimes just it's jump points) with the y-axis stretched by the inverse of the CDF $F^{-1}(\cdot)$.

The monotonic transformation of the y-axis is such that if the data comes from the candidate $F(x)$, the points would appear to lie on a straight line.

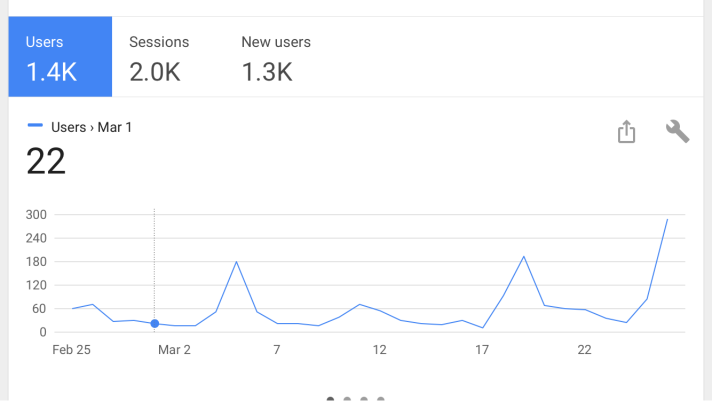
Names of variations of probability plots are the **P-P plot** and **Q-Q plot** (these plots are similar to the probability plot).

A very common probability plot is the **Normal probability plot** where the candidate distribution is taken to be Normal$(\overline{x}, s^2)$.

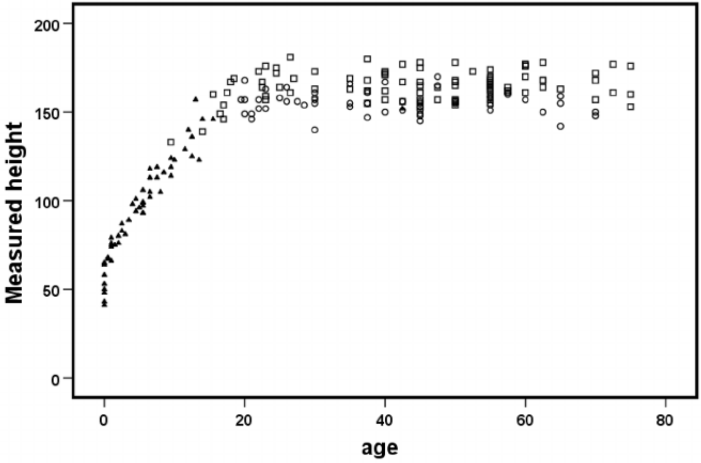
The Normal probability plot can be useful in identifying distributions that are symmetric but that have tails that are "heavier" or "lighter" than the Normal.



A **time series plot** is a graph in which the vertical axis denotes the observed value of the variable and the horizontal axis denotes time.



A **scatter diagram** is constructed by plotting each pair of observations with one measurement in the pair on the vertical axis of the graph and the other measurement in the pair on the horizontal axis.



The **sample correlation coefficient** r_{xy} is an estimate for the correlation coefficient, ρ, presented in the previous unit:

$$r_{xy} = \frac{\sum_{i=1}^{n} (y_i - \bar{y})(x_i - \bar{x})}{\sqrt{\sum_{i=1}^{n} (y_i - \bar{y})^2 \ \sum_{i=1}^{n} (x_i - \bar{x})^2}}.$$