Statistics

Week 2: Summarizing and Exploring Data (Chapter 4)

ESD, SUTD

Term 5, 2017

SINGAPORE UNIVERSITY OF TECHNOLOGY AND DESIGN

Established in collaboration with MIT

Outline

Histogram, continued

In an *Excel* histogram, you should:

- Properly label the bins (the values generated are only the upper boundaries of the bins).
- Change the Gap Width to 0%.

There is no universal formula for choosing the number of bins.

- *Excel* uses $[\sqrt{n}]$ bins.
- Another recommendations is to use $[\log_2 n] + 1$ bins.
- Yet another rule is to set the bin width to $2\,\mathrm{IQR}/n^{1/3}$.

In practice, aim for between 5 and 20 bins, and make the boundaries 'nice' numbers.

Other measures of spread

The *sample coefficient of variation* is defined as

$$\mathrm{CV} = \frac{s}{\bar{x}}.$$

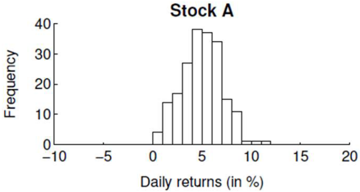
It is used in, for example, queueing theory (for an exponential distribution, CV should be 1).

The *z-score* or standard score calculates how many standard deviations a data value is above the sample mean:

$$z_i = \frac{x_i - \bar{x}}{s}.$$

You have seen this used in the normal distribution. It is useful for comparing different data sets.

Why study spread – example



- Are these two stocks similar for investors?
- Which one would you invest in?

Sample covariance

Bivariate data can be represented on a scatter plot.

Recall that the covariance of two random variables X and Y is given by $\mathsf{E}[(X - \mathsf{E}[X])(Y - \mathsf{E}[Y])]$.

Sample covariance and correlation

Given data values $(x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n)$, the sample covariance is defined as

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

The sample correlation coefficient is given by

$$r = \frac{s_{xy}}{s_x s_y}.$$

We will use these when studying linear regression.

Demo: look up Anscombe's quartet.

Tables

Bivariate data can also be represented in table form.

Before making any conclusions from tables, be careful of the way samples are drawn.

Example: a respiratory problem is studied by first finding 500 smokers and 500 non-smokers and then determining whether or not each individual has the problem. The results are shown below.

	Yes	No	Row total
Smokers	250	250	500
Non-smokers	50	450	500
Column total	300	700	1000

Exercise: are the following statements true?

About $5/6$ of all people with the respiratory problem are smokers.

About $1/2$ of all smokers have the respiratory problem.

Simpson's paradox

Real life example comparing two treatments for kidney stones:

	Treatment A	Treatment B
Small stones	93% (81/87)	87% (234/270)
Large stones	73% (192/263)	69% (55/80)
Both	78% (273/350)	83% (289/350)

This can occur when the group sizes are uneven, so watch out.

Q-Q plot

A **Q-Q plot** compares two probability distributions by plotting their quantiles against each other.

A point (x, y) on the plot corresponds to a quantile of the 2nd distribution plotted against the same quantile of the 1st one.

The *normal probability plot* is a special case of the Q-Q plot, when the 2nd distribution is the standard normal.

If a normal probability plot is close to a *straight line*, then the 1st distribution is approximately normal (since all normal distributions are related by linear transformations).

Normal probability plot

Consider some ordered data values $x_{(1)} \leq x_{(2)} \leq \cdots \leq x_{(n)}$.

Then $x_{(i)}$ is the $\frac{i}{n+1}$ quantile.

We plot $x_{(i)}$ against the $\frac{i}{n+1}$ quantile of the standard normal distribution, which is given by $\Phi^{-1}\left(\frac{i}{n+1}\right)$.

Intuition for using $n+1$ and not n: (1) imagine drawing $x_{(i)}$ from a distribution... (2) $\Phi^{-1}\left(\frac{n}{n}\right) = \infty$.

See *Excel* demo on speed of light data. For Φ^{-1}, use `norm.s.inv`.

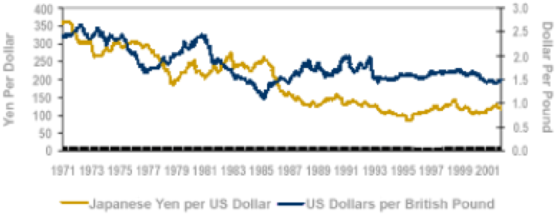
Outline

Time-series data

A **time series** is a sequence of data points $x_1, x_2, x_3, \ldots$, measured at successive points in time (typically spaced at uniform intervals). For examples, daily closing value of a stock, or annual rainfall.

Usually, a time-series has the following components: stable, trend (long-term pattern), seasonal (short-term, periodic fluctuation), random.

There are many examples from economics and finance:



Forecasting

We now describe some methods to

- Smooth out short-term fluctuations and highlight long-term trends in a time series, and/or
- Attempt to predict (forecast) the value of a time series at the next point in time.

A naïve way to forecast is to use the last data point:

$$F_{t+1} = x_t.$$

A more sophisticated approach is the **moving average**:

$$F_{t+1} = \frac{x_{t-w+1} + \cdots + x_{t-1} + x_t}{w}.$$

This also allows us to smooth out the time series, but can introduce a lag.

Exponentially weighted moving average

Weighted moving average: α_i are the weights; the idea is to give more importance to more recent data.

$$F_{t+1} = \frac{\alpha_{w-1} x_{t-w+1} + \cdots + \alpha_1 x_{t-1} + \alpha_0 x_t}{\alpha_{w-1} + \cdots + \alpha_1 + \alpha_0}.$$

We can choose α_i to be decreasing **exponentially**.

Let $\alpha \in (0,1)$, then define $F_{t+1} = \mathsf{EWMA}_t$, where

$$\mathsf{EWMA}_t = \alpha\, x_t + (1-\alpha)\,\mathsf{EWMA}_{t-1},$$

with $\mathsf{EWMA}_0 = x_1$.

If we apply this formula repeatedly, then after simplification,

$$\begin{aligned}\mathsf{EWMA}_t = \alpha\Big[x_t + (1-\alpha)x_{t-1} + (1-\alpha)^2 x_{t-2} + \cdots + (1-\alpha)^{t-1} x_1\Big] \\ + (1-\alpha)^t x_1.\end{aligned}$$

Forecasting error

How do we pick α?

The error of the forecast is $e_t = x_t - F_t$.

α can be chosen to minimize some total error.

One commonly used measure of total error is the *mean absolute percent error*, defined as

$$\text{MAPE} = \frac{1}{T-1} \sum_{t=2}^{T} \left| \frac{e_t}{x_t} \right| \times 100\%.$$

In *Excel*, we can use Solver to find the value of α that minimizes MAPE.