Normal/Gaussian distribution, properties

- The pdf of the normal distribution is:

$$f(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2}(\frac{x-\mu}{\sigma})^2}$$

  where μ is the mean and $\sigma^2 > 0$ is the variance. A random variable X with a normal distribution is written $X \sim N(\mu, \sigma^2)$.
- The standard normal distribution (SND) is: $N(0, 1)$ - that is with mean 0 and variance 1. If $X \sim N(\mu, \sigma^2)$ and $Z = \frac{X-\mu}{\sigma}$ then $Z \sim N(0, 1)$. So, all normal distributions are in some sense similar and we can work with just the standard normal distrubtion.
- Normal distributions are closed under linear transformations. If $X \sim N(\mu, \sigma^2)$ and $Y = aX + b$, $a, b \in \mathbb{R}$ then $Y \sim N(a\mu + b, a^2\sigma^2)$.
- If RVs X, Y are independent and $X \sim N(\mu_1, \sigma_1^2)$, $Y \sim N(\mu_2, \sigma_2^2)$ then $X + Y \sim N(\mu_1 + \mu_2, \sigma_1^2 + \sigma_2^2)$.
- The mode and median is μ, skewness is 0 and excess kurtosis is 0 (the actual kurtosis of SND is 3).

Skewness and Kurtosis

- **Skewness:** is a measure of asymmetry of the distibution of the sample/pdf.
- Defined as: for sample of size n,

$$g = \frac{\frac{\sum_{i=1}^{n}(X-\bar{X})^3}{n}}{s^3}$$

for a distribution:

$$g = \frac{E((X-\mu)^3)}{\sigma^3} = E((\frac{(X-\mu)}{\sigma})^3)$$

- Skewness can be +ve, -ve or zero.
- Asymmetry is not just visual symmetry but how the distribution is distributed around the mean. A distribution that appears visually symmetric will have a 0 or very low skewness. But converse is not true.
- The normal distribution has skewness of 0.

Kurtosis

- **Kurtosis:** is a measure of outliers.
- Defined as: for sample of size n,

$$\kappa = \frac{\frac{\sum_{i=1}^{n}(X-\bar{X})^4}{n}}{s^4}$$

for a distribution:

$$\kappa = \frac{E((X-\mu)^4)}{\sigma^4} = E((\frac{(X-\mu)}{\sigma})^4)$$

- The SND has a $\kappa = 3$. κ is often measured as excess kurtosis, that is the value beyond 3. Important to know which definition is being used

Parameters and statistics

- Parameters characterize a probability distribution of a population. Ex. μ and σ^2 for the Gaussian distribution; p and N for the Binomial distribution.
- A sample is a subset drawn from the population. A sample is **random** when every element of the population has the same probability of being part of the sample. Henceforth, all samples are assumed to be random.
  All the tests and results depend on the sample(s) being random.
- Any quantity or value (like mean, median, variance, proportion, quantile etc.) computed from the elements of the sample is called a **sample statistic** or just **statistic**.
- The main goal of statistical inference is to infer population parameters from sample statistics. More rarely one infers causality.
- In practice samples are rarely, if ever, random and we do not have a large number of samples to produce a sample distribution to estimate the parameters. In practice we have just one sample.

The simple random sample

- There are multiple ways to sample. The most common and assumed in the results that follow is the **simple random sample** - often just called random sample.
- Math defn:
  N picks corresponding to RVs $X_1, \ldots, X_N$ are a simple random sample if i) if the N RVs are independent and ii) each RV has the same distribution as the population. Often called idependent, identically distributed or *iid* in short.
- Informal defns:
  i) A sample of N elements from a population is a simple random sample if all possible samples of N elements from the population had the same probability of being selected.
  ii) A sample of N elements from a population is a simple random sample if all elements in the population had equal and independent probabilities of being selected.
  iii) A sample of N elements is a simple random sample if at every selection every unit in the population has the same probability of

Example

You want to draw random samples of size 3 from the numbers 1 to 9. Here are two ways you can do it.

1. On identical pieces of paper write all possible combinations of 3 digits, put them in a box, shake thoroughly and draw one piece of paper.
2. Write the numbers 1 to 9 on identical pieces of paper, put them in a box, shake thoroughly and pick one piece of paper at a time three times without replacement.

Do both ways of choosing the sample give us a random sample? Where random sample is interpreted as: all samples of size 3 have the same probability of being chosen. All selections are without replacement.

Example contd.

First case: the number of pieces of paper will be the number of ways one can choose 3 from a set of 9 or $\binom{9}{3} = \frac{9\times8\times7}{3\times2} = 84$. Probability of choosing any triple is: $\frac{1}{84}$. So, every triple has same probability and the sample is indeed a simple random sample.

Second case: Probability of choosing a specific triple, say (i, j, k) is: probability that one of i, j, k will be chosen in the first pick is:$P(L1) = \frac{3}{9}$, the probability that one of the remaining two will be picked in the second pick given one of i, j, k has already been picked is $P(L2) = \frac{2}{8}$, similarly the probability the remaining third will be picked in the third pick is $P(L3) = \frac{1}{7}$. So, $P(L1, L2, L3) = P(L1)P(L2|L1)P(L3|L1, L2) = \frac{3}{9} \times \frac{2}{8} \times \frac{1}{7} = \frac{1}{84}$.

We get exactly the same probability as the first case and so the sample is a random sample. It is much less obvious that this is way of selecting a sample also gives a simple random sample.

So, simple random samples can be taken one at a time from the population without replacement provided all remaining members of the population have the same probability of being chosen in every pick. Or the whole sample can be picked in one shot where all possible such samples in the population have the same probability of being selected.