Supplementary material to 'Class notes 1'

Brief introduction to some known probability models

Probability

To understand probability, it is useful to introduce the concepts *event*, *random experiment*, and *relative frequency*. To make things simple, we define an event as the outcome of a random experiment. The relative frequency is the number of times an event occurs divided by the number of times the random experiment is repeated.

Let A be an event and let $P(A)$ denote the probability of A occurring. Then, $P(A)$ can be approximated by the relative frequency

$$P(A) \approx \frac{\#\, times\, A \text{ occurs}}{n}$$

where n is the number of times the random experiment is repeated. According to the *law of large numbers*, as n tend to infinity (as n becomes a very large number), the relative frequency tend to the exact probability of *A*.

Random variables

Broadly speaking, a random (or stochastic) variable is a variable whose outcome is subject to random (or stochastic) change.

Characteristics of random variables

Random variables have certain characteristics that is of interest to us. The two most important characteristics of a random variable is its *expectation* and *variance*. The expectation of X (or the mean of X) is written as

$$E(X)$$

Simplifying the understanding somewhat, the expectation can be thought of as the average outcome of the variable X as the random experiment is repeated infinitely many times.

Another interesting characteristic is the variance. The variance is defined as

$$Var(X) = E\big(X - E(X)\big)^2 \geq 0$$

The variance of a random variable is a measure of dispersion (how spread the distribution is). Sometimes it is beneficial to work with the standard deviation of X rather than the variance. The standard deviation is simply the square root of the variance

$$Std(X) = \sqrt{Var(X)}$$

The normal distribution

In statistics, the normal distribution plays a crucial role. There are several reasons for the popularity of the normal (or Gaussian) distribution. First, many real world phenomena may be well approximated by the normal distribution. Second, the normal distribution plays an important role in the *Central limit theorem* (CLT).

Let X be a normal random variable with expectation μ and variance σ^2. We typically use the notation: $X \sim N(\mu, \sigma^2)$. The *probability density function* (or just the pdf) is given by the mathematical expression

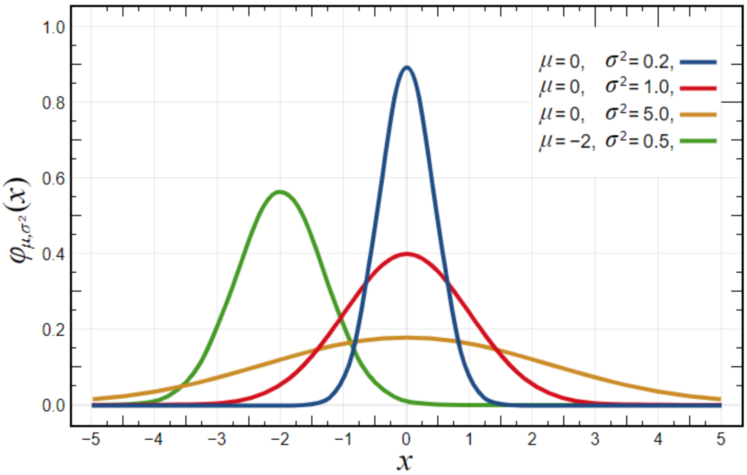
$$f(x) = \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{1}{2\sigma^2}(x-\mu)^2}$$

where $-\infty < x < \infty$. In this expression, the parameter μ determines the location (or the central tendency) and σ^2 determines the spread (or the dispersion) of the distribution. One can show that

$$E(X) = \mu$$

$$Var(X) = \sigma^2$$

A graphical representation of the normal pdf is referred to as the normal curve. The following figure (taken from Wikipedia) shows three different normal density curves:



The density function $f(x)$ has the following properties

1. The area below the density curve is exactly 1.
2. $f(x) \geq 0$. That is, the curve is never below zero.

Note that these properties hold for any pdf (including the normal distribution). A special, and important, case of the normal distribution is the *standard* normal distribution. A normal variable is said to be standard normal if the expected value is 0 and the variance is 1. Typically, the letter Z is used to denote a standard normal random variable. Formally, we write $Z \sim N(0,1)$.

Obtaining probabilities of a normal random variable

For simplicity, we start with the case of a standard normal variable. That is, $Z \sim N(0,1)$. We are interested in finding the probability that $Z \leq z$, where (lower case) z is some value. This is also written as

$$P(Z \leq z)$$

Some examples involving numbers are:

$$P(Z \leq 2)$$

$$P(Z \leq -1 \text{ } or \text{ } Z \geq 1)$$

$$P(-1.96 \leq Z \leq 1.96)$$

There is no closed form formula for computing these probabilities. Thus, we need to obtain these values, either by the normal table or by computers.

Standardizing a normal variable

How can we obtain probabilities in case $X \sim N(\mu, \sigma^2)$, where μ , σ^2 can be any values? Note that the normal table only provide probabilities for the standard normal. The problem is solved by standardization.

A normal variable is standardized by the use of the following expression

$$Z = \frac{X - \mu}{\sigma}$$

It is then the case that $Z \sim N(0,1)$. Thus, it is possible to obtain probabilities from the normal table even if the distribution is non-standard. Again, let $X \sim N(\mu, \sigma^2)$. Now use

$$\begin{aligned} P(X \leq x) &= P\left(\frac{X - \mu}{\sigma} \leq \frac{x - \mu}{\sigma}\right) \\ &= P(Z \leq z) \end{aligned}$$

which is easy to find using the normal table.

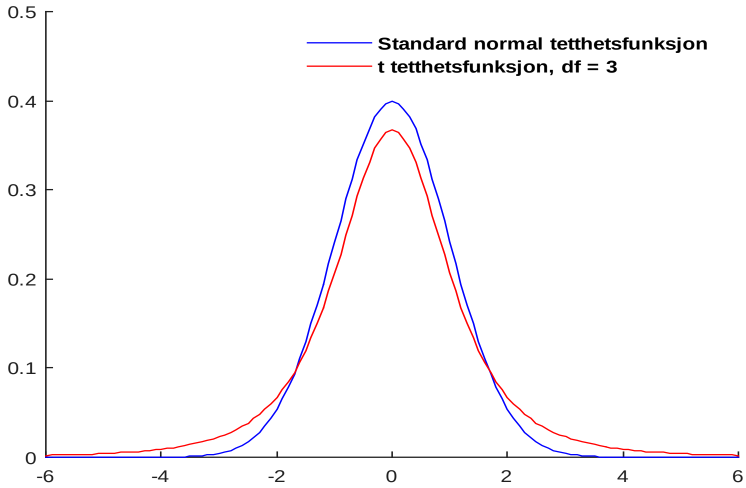
As an example, let $X \sim N(108, 15^2)$. Find $P(X > 130)$. By standardization, we have

$$\begin{aligned} P(X > 130) &= P\left(\frac{X - 108}{15} > \frac{130 - 108}{15}\right) \\ &= P(Z > 1.4667) \\ &\approx 7.1\% \end{aligned}$$

The t-distribution

Another continuous distribution is the t-distribution. As we shall see later, the t-distribution plays an important role in inferential statistics. The density function of a t-distributed random variable is a rather involved mathematical expression. Thus, the mathematical details will be skipped in these notes. Nevertheless, we briefly mention that the sole parameter determining the exact shape of the density function is called *degrees of freedom*, or simply *df*. The parameter *df* is a positive integer value (a number in the sequence: 1, 2, 3, …).

The following figure illustrates the resemblance between the t-distribution with $df = 3$ and the standard normal distribution



Note that as *df* $\rightarrow \infty$ (as *df* becomes a large number), the *t*-distribution converge to a standard normal distribution. For *df* $=$ 100, there is little difference between the two distributions. Thus, for large values of *df*, we have

$$P(T \leq t) = P(Z \leq z) \quad \text{for} \;\; t = z$$

The Central Limit Theorem

One of the most important results in statistics is the Central Limit Theorem (CLT). There are many versions of the CLT, all derived for various purposes. The CLT presented here is sometimes referred to as the classical CLT. Before presenting the CLT, it is necessary to introduce the concept of a *random sample*, and to outline a few simple results.

Random sample

A random sample is a collection of random variables, typically denoted $X_1, X_2, \dots, X_n$, which are *independently* and *identically distributed*. Thus, the Xs that make up the sample must be mutually unrelated, and they must come from the same distribution. The latter part of this statement implies that the Xs all have the same mean and variance.

The mean and variance of $\bar{X}$

Let $X_1, X_2, \dots, X_n$ be a *random sample* from a population with mean μ and variance σ^2. Moreover, let the sample mean be given by the expression

$$\bar{X} = \frac{1}{n}\sum_{i=1}^{n} X_i = \frac{X_1 + X_2 + \cdots + X_n}{n}$$

Under the stated conditions, we have

$$E(\bar{X}) = \mu$$

$$Var(\bar{X}) = \frac{\sigma^2}{n}$$

The Theorem

As before, let $X_1, X_2, \dots, X_n$ be a random sample from a population with mean μ and variance σ^2, and let the sample mean be

$$\bar{X} = \frac{1}{n}\sum_{i=1}^{n} X_i = \frac{X_1 + X_2 + \cdots + X_n}{n}$$

The CLT states that for a reasonably large n

$$\bar{X} \sim N\left(\mu, \frac{\sigma^2}{n}\right)$$

Note that this relation holds approximately. In less technical terms, the CLT tells us that for a reasonably large n (for instance, $n > 50$), the sample mean, given by $\bar{X}$, is approximately normally distributed with expectation μ and variance σ^2/n.

Also note the following:

- The CLT applies for any distribution of the Xs, given that they are independent and the distribution is the same for all.
- A special case is when the Xs all come from the same normal distribution. In that case, for any sample size n, $\bar{X}$ will be normally distributed with mean μ and variance σ^2/n.
- The square root of the variance $\sqrt{\frac{\sigma^2}{n}} = \frac{\sigma}{\sqrt{n}}$ is called the *standard error* of $\bar{X}$.