Basics of Probability & Statistics

Chiranjit Mukhopadhyay
Indian Institute of Science

1 Introduction

Though our ultimate goal is to get into formal ANOVA and Regression model building as soon as possible, to get there, we first need to have a basic understanding of the underlying probability theory and the related probability distributions and basic elements of statistical inference, which encompasses estimation and hypothesis testing. Of these two elements of statistical inference, we only need to have good understanding of hypothesis testing, because selection between alternative models or testing for significance of a factor are all formulated as hypothesis testing problems. However we need to have a passing understanding of estimation theory as well, though the focus of this notes will rest upon the mechanics and the logic behind hypothesis testing.

We begin with some very basics of probability theory, but quickly move on to only those issues and models which are relevant to us for AutoDOE. We take up the issues pertaining to statistical inference next.

2 Probability Models

Statistical model building problems are typically concerned with developing relationships between a set of variables based on observations on these variables. Though these relationships are built empirically based on the observations, conceptually they attempt to capture the kind of relationships these variables might have in an abstract "population", which consists of the totality of all possible values the variables might assume, of which the observed values are just a part of the whole or a sample. Now the values (of the variables under consideration) occur in the population according to certain frequency, meaning some values occur more often than other. Or in other words when we sample the values through observations, some values would be more likely to occur than others. This frequency law of the values in the population is modeled in terms of a probability distribution.

That is a probability distribution is nothing but a model, which describes how the values of a variable are distributed in a population. More generally a model [1] specifies how different variables are (probabilistically) related to one another in the population. Since in the population, the totality of all possible values of a variable is handled by its probability distribution, a model in general specifies how the values of different variables under consideration are jointly distributed in the population. That is a (probability) model is equivalent

[1] We will not try to nitpick with the terms "probability model" or "statistical model" because they are used interchangeably. A model to us is always a probability model. But a probability model cannot be written down without any observations, it necessarily has to be empirically built from a sample of observations. This process of building a probability model from empirical statistical observations, called a sample, using statistical techniques is called statistical model building, and as a result the final model is also referred to as a statistical model.

to the specification of a (joint) probability distribution of the variables under consideration. A probability model is a model in the population, or in other words, it tries to describe what is happening *i.e.* how values are distributed and the relationship between values of variables, in the totality of all possible values in the population.

The way the probability distribution of a variable is specified, depends on the nature of the values the variable can take. They essentially fall in one of the two categories - discrete or continuous. Number of times a component fails within a specified time period, number of times a driver applies brakes during a 10 km journey, number of piston movements in a second etc. are examples of discrete variables; while mileage (number of km per litre), reaction time of a driver, average speed during a journey etc. are examples of continuous variables.

2.1 Discrete Probability Models

In general, for a discrete variable its probability distribution is specified using its probability mass function (p.m.f.) as follows. Let the discrete variable X takes values $\{x_0, x_1, x_2, \ldots\}$. Then its p.m.f. $p(x)$ gives $P(X = x)$. That is for a discrete variable its p.m.f. can be simply viewed as a sequence $\{p_0, p_1, p_2, \ldots\}$ such that $0 \leq p_i \leq 1 \ \forall i = 0, 1, 2 \ldots$ and $\sum_{i=0}^{\infty} p_i = 1$, with the interpretation that p_i gives $P(X = x_i)$. Alternatively many times, actually more often than not, the p.m.f. $p(x)$ is expressed as a formula with a few parameters (instead of a(n) (in)finite sequence), in which case they are referred to as discrete probability models. For example, if probability of Head in a single toss of a coin is p, the coin is tossed n times and X denotes the number of Heads in these n tosses then the p.m.f. of X can be expressed as $p(x) = \begin{pmatrix} n \\ x \end{pmatrix} p^x (1-p)^{n-x}$ for $x = 0, 1, 2, \ldots, n$ (Binomial model); if the coin is tossed till a Head appears and X denotes the number of Tails before the Head then the p.m.f. of X can be expressed as $p(x) = (1-p)^x p$ for $x = 0, 1, 2, \ldots$ (Geometric Model); if the coin is tossed till n Heads appear and X denotes the total number of Tails in this experiment, then the p.m.f. of X can be expressed as $p(x) = \begin{pmatrix} n+x-1 \\ x \end{pmatrix} p^n (1-p)^x$ for $x = 0, 1, 2, \ldots$ (Negative Binomial model); the number of times a driver applies brakes during a journey, say X, may be reasonably modeled using the p.m.f. $p(x) = e^{-\lambda}\lambda^x/x!$, $x = 0, 1, 2, \ldots$, for some $\lambda > 0$, called the parameter of the Poisson model.

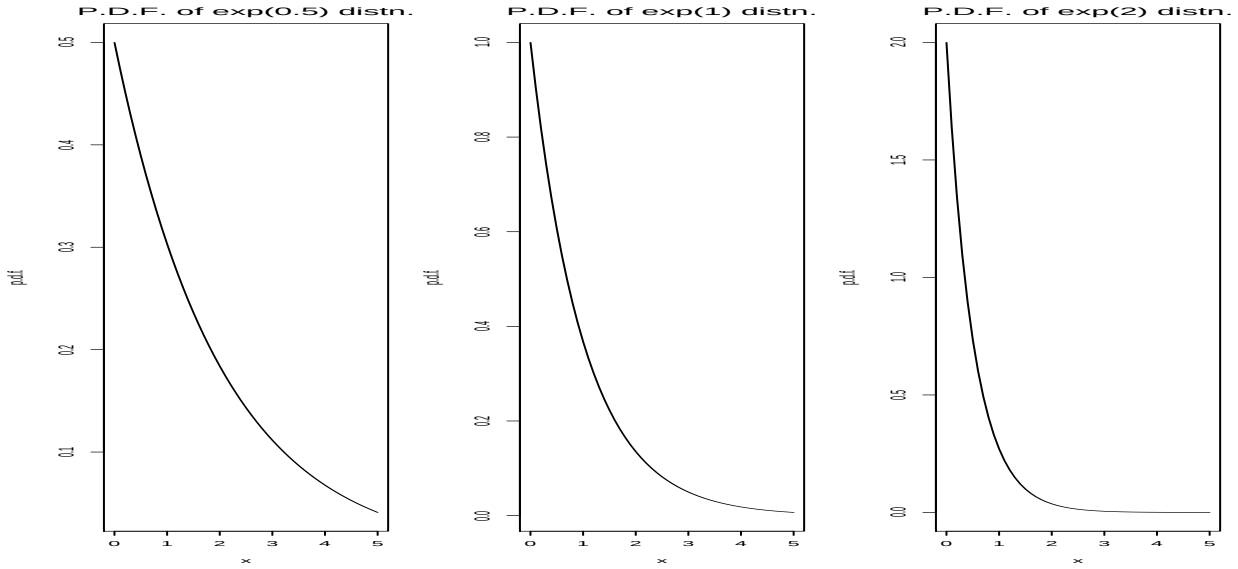
2.2 Continuous Probability Models

In the AutoDOE context, all the probability models or probability distributions we will be concerned with are continuous. Probability distributions of continuous variables are not specified in terms of its p.m.f. Because a variable is called continuous if $P(X = x) = 0$ $\forall x$ - that is its definition! Thus by definition, for a continuous variable its p.m.f. $p(x)$ is hopelessly identically equal to 0. A new calculus-based devise is utilized to define the probability distribution of a continuous variable, which is called probability density function or p.d.f..

The p.d.f. of a continuous variable X is defined as $f(x) = \lim_{dx \to 0} \frac{P(x < X < x+dx)}{dx}$. Or in other words $f(x)dx$ may be interpreted as the probability of X taking value very close

to x in a dx neighborhood. Probabilities of X taking values in some set A may then be computed as $\int_A f(x)dx$. In general any function $f(x)$ which is non-negative *i.e* $f(x) \geq 0\ \forall x$ with a total area underneath it as 1 *i.e.* $\int_{-\infty}^{+\infty} f(x)dx = 1$ is a bona fide p.d.f. Mean and variance of continuous variable X is computed using its p.d.f. as $\mu = E[X] = \int_{-\infty}^{+\infty} xf(x)dx$ and $\sigma^2 = V[X] = E[(X-\mu)^2] = \int_{-\infty}^{+\infty}(x-\mu)^2 f(x)dx = \int_{-\infty}^{+\infty} x^2 f(x)dx - \mu^2 = E[X^2] - \mu^2$ respectively.

Continuous probability models are typically specified in terms of its p.d.f. For example an exponential distribution with parameter λ, denoted by $\exp(\lambda)$, is defined as that distribution which has a p.d.f. $\lambda e^{-\lambda x}$ for $x > 0$. An $\exp(\lambda)$ distribution has mean $1/\lambda$ and variance $1/\lambda^2$. Some typical $\exp(\lambda)$ p.d.f. are depicted below:



The most important distribution in statistical applications is called Normal or Gaussian distribution. This is because most of the standard continuous variables (at least approximately) tend to follow this distribution. (And hence the name "Normal"). Furthermore whenever we measure something using a device typically it will be contaminated with some measurement errors, which arise from innumerable many sources. Theoretical considerations of the structure of such measurement errors also lead to Normally distributed measurements. Because of its importance and utility in the modeling exercise in AutoDOE, we devote a little bit of time in studying the Normal distribution.

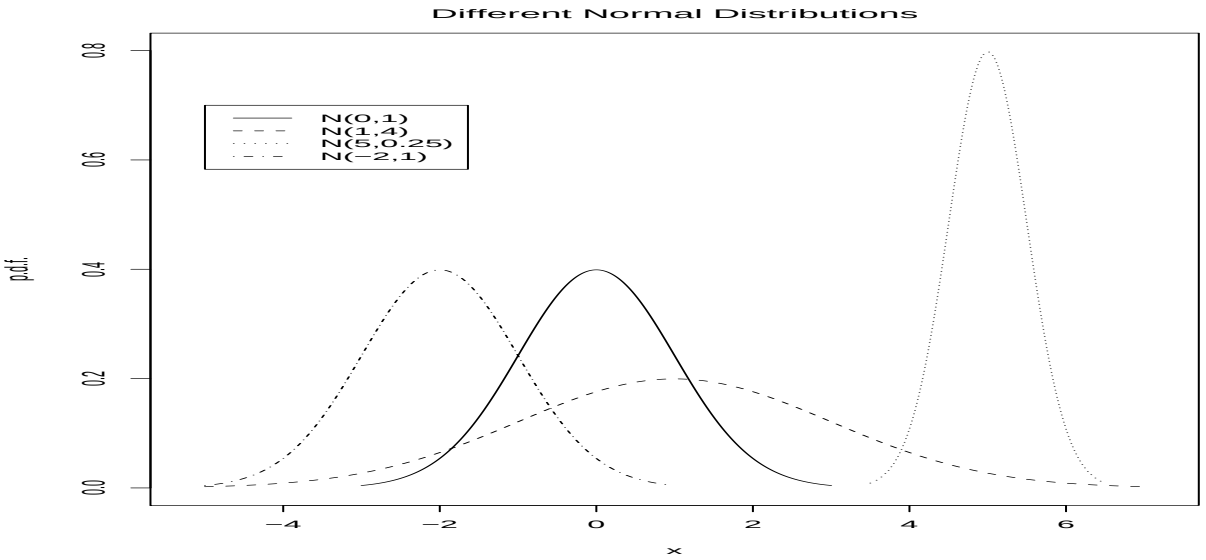
2.3 Normal or Gaussian Distribution

The Normal distribution is completely specified or characterized by two parameters (as exponential has one parameter λ) namely its mean μ and variance σ^2. Modeling with Normal distribution typically concentrates on the structure of μ and sometimes σ^2. For instance the different types of models that were discussed in the first session, basically are specifications of mean μ of a response which is assumed to have a Normal distribution.

The p.d.f. of a Normal distribution with mean μ and variance σ^2 is given by

$$f(x;\mu,\sigma^2) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{1}{2}\frac{(x-\mu)^2}{\sigma^2}} \qquad -\infty < x < +\infty \tag{1}$$

which has a symmetric bell shape, which is also popularly known as the so-called bell-curve. Such a Normal distribution with mean μ and variance $\sigma^2$is denoted by $X \sim N(\mu, \sigma^2)$. The mean parameter μ dictates where the main hump or the major probability mass of the distribution would be located, while the variance parameter σ^2 determines how this probability mass should be spread about or distributed along the x-axis. Larger the value of σ^2 more diffused the probability mass would be. The effects of changing the values of these two parameters on the resulting shapes of the p.d.f. given in equation (1) are depicted in the following figure:



We will need to compute probabilities and quantiles for a $N(\mu, \sigma^2)$ distribution while performing hypothesis testing. Thus it would be helpful to know that if $X \sim N(\mu, \sigma^2)$, then $Z = \frac{X-\mu}{\sigma} \sim N(0, 1)$, a Normal distribution with mean 0 and variance 1. The $N(0, 1)$ distribution is called the standard Normal distribution for which probability tables are available (this is because unlike the exponential distribution the indefinite integral of the Normal distribution cannot be obtained in a closed form). Any required probability or quantile computation of an arbitrary $N(\mu, \sigma^2)$ distribution is performed by first formulating the problem in terms of the standard Normal distribution and then solving it using the standard Normal probability tables.

3 Statistical Inference

As mentioned in §1, statistical inference deals with estimation and hypothesis testing about unknown population parameters, given a set of observations on the variable whose population behavior we want to study or model. As seen in §2 above, population parameters are nothing but some quantities which appear in the underlying (population) probability distribution of the variable under consideration as unknown constants, like the p of Binomial, Geometric or Negative Binomial distribution, or λ of the Poisson or exponential model or (μ, σ^2) of the Normal probability model of §2.3. Since we will almost exclusively deal with the Normal probability model, here we will only discuss the inferential techniques or the meth-

ods pertaining to the Normal distribution, instead of trying to invoke general mathematical statistics principles and results of statistical inference.

Thus suppose we have n observations $Y_1, Y_2, \ldots, Y_n$ on some response of interest Y. Life would be simple if we can somehow convince ourselves that the underlying population from which these observations are coming is Normal, because the inferential techniques for the Normal distribution is very standard and fairly easy to implement. Once the Normality is established, we would like to estimate and test hypothesis about the unknown parameters μ and σ^2. Thus we shall first address the issue of validating the assumption of the (underlying) Normal distribution. This leads to what is called Normal Probability Plots (NPP). Note that NPP has also been implemented in the final diagnostic module of AutoDOE and thus is a very important topic of discussion.

3.1 Normal Probability Plots

The problem here is to validate the assumption that the observations $Y_1, Y_2, \ldots, Y_n$ are coming from some $N(\mu, \sigma^2)$ population. One naive way to attempt to solve this problem would be to construct a histogram of of these observations and then visually check whether it roughly resembles a symmetric bell-shape or not. But this approach is fraught with danger for numerous reasons. First, histogram of majority of variables indeed exhibit a hump in the middle values with decaying frequency in the tails. Second, there are numerous other distributions (probability models) which have this bell-shape, as we shall see shortly. Third, we humans are not very good at judging the nature of curvature (quadratic versus cubic for instance). These considerations compel us to somehow "linearize" the problem, so that we can inspect whether something follows a straight-line or not, which we humans are very good at.

For any variable Y its cumulative distribution function or c.d.f.is defined as $F(y) = P(Y \leq y)$. [2] Note that if Y is discrete taking values $\{y_0, y_1, \ldots\}$ with p.m.f. $\{p_0, p_1, \ldots\}$, $F(y) = \sum_{i:y_i \leq y} p_i$; and if Y is continuous with p.d.f. $f(y)$, $F(y) = \int_{-\infty}^{y} f(t)dt$. Though in general the (population) c.d.f. of a variable Y remains unknown, given n observations $Y_1, Y_2, \ldots, Y_n$ on Y its c.d.f. may be estimated as $\hat{F}(y) = \#\{i : Y_i \leq y\}/n$ without making any assumption about the underlying probability model whatsoever. $\hat{F}(y)$ is called the empirical c.d.f. of Y.

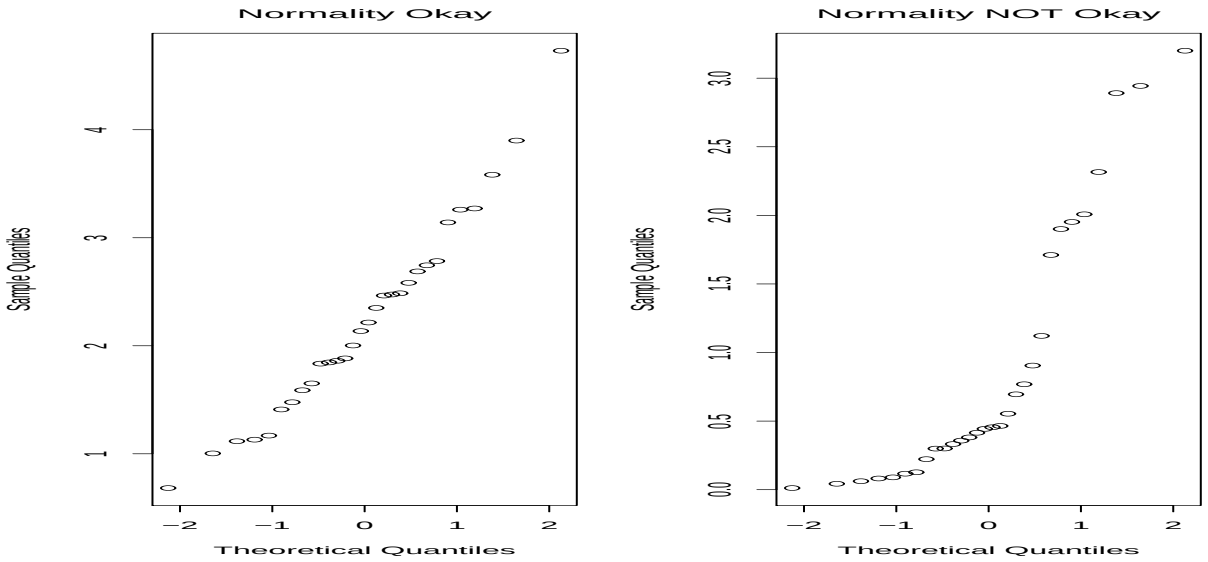
Now if Y has a $N(\mu, \sigma^2)$ distribution, its c.d.f. $F(y)$ is given by $F(y) = \Phi(\frac{y-\mu}{\sigma})$, where $\Phi(x) = \int_{-\infty}^{x} \phi(t)dt$ and $\phi(t) = \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}t^2}$, which is same as equation (1) with $\mu = 0$ and $\sigma^2 = 1$, the standard Normal p.d.f. Note that $\phi(\cdot)$ and thus $\Phi(\cdot)$ are known functions *i.e.* they do not depend on μ and σ and thus may be computed at least numerically.

Though we do not know $F(y)$, if we replace it by its estimate the empirical c.d.f. $\hat{F}(y)$ in the above relationship $F(y) = \Phi(\frac{y-\mu}{\sigma})$, we get that $y = \mu + \sigma\Phi^{-1}\left(\hat{F}(y)\right)$ where $\Phi^{-1}(\cdot)$ is the inverse function of $\Phi(\cdot)$ (the inverse exists because $\Phi(\cdot)$ is strictly increasing and thus one-to-one). This gives a linear relationship of the Y_i's with the corresponding $\Phi^{-1}\left(\hat{F}(Y_i)\right)$'s.

[2] C.d.f's play a very important role in calculation of probabilities. For example $P(Y > a)$ can be calculated as $1 - F(a)$, $P(a < Y \leq b) = F(b) - F(a)$, $P(Y < y) = F(y-)$, the left hand limit of $F(\cdot)$ at y etc. As a matter of fact, probability tables of standard distributions like the standard Normal mentioned above, or Binomial or Poisson etc. all tabulates nothing but the c.d.f. of the corresponding variables.

That is since $F(y)$ has been estimated by $\hat{F}(y)$, without any assumption, and if $F(y)$ indeed has a Normal form, if we plot the $\Phi^{-1}\left(\hat{F}(Y_i)\right)$'s (called theoretical quantiles) in the x-axis (abscissa) and Y_i's (called sample quantiles) the y-axis (ordinate) we should get a straight line.

This plot of sample quantiles against the theoretical quantiles is called the Normal Probability Plot. If the points in the plot appear to be in a straight-line then with the above logic it may seem reasonable to assume that the observations $Y_1, Y_2, \ldots, Y_n$ are coming from some $N(\mu, \sigma^2)$ population, otherwise they are not. Two prototype NPP are provided in the following figure, where in the first case there is no reason to suspect the Normality assumption, while in the second case we should reject the Normality assumption.



3.2 Inference for a Single Normal Population

Suppose after the preliminary NPP analysis we are now willing to assume that $Y_1, Y_2, \ldots, Y_n$ are coming from some $N(\mu, \sigma^2)$ population. The next issue is to estimate the values of the unknown population parameters (μ, σ^2). The estimate for the population mean μ is obvious, which is given by the sample mean $\overline{Y} = \frac{1}{n}\sum_{i=1}^{n} Y_i$. It can also be shown that the sample mean $\overline{Y}$ is indeed the "best" estimate of μ in any sense you may put forward a criterion for being the "best". The estimate of the population variance σ^2 is however slightly less intuitive. Its estimate is given by the so-called sample variance $s_{n-1}^2 = \frac{1}{n-1}\sum_{i=1}^{n}(Y_i - \overline{Y})^2$. Note that the divisor is $n-1$ and not n. To understand the reasoning behind this estimate, we have to first understand the very important concept of **sampling distributions**, which will be used very critically from now on.

In classical statistics, as we are doing here, the optimality of any method is judged in terms of its repeated use over different samples from the same population. That is if we use the same method over and over again for all possible samples that we can draw from a population, the method would be called "good" if its average performance is satisfactory over this repeated sampling.

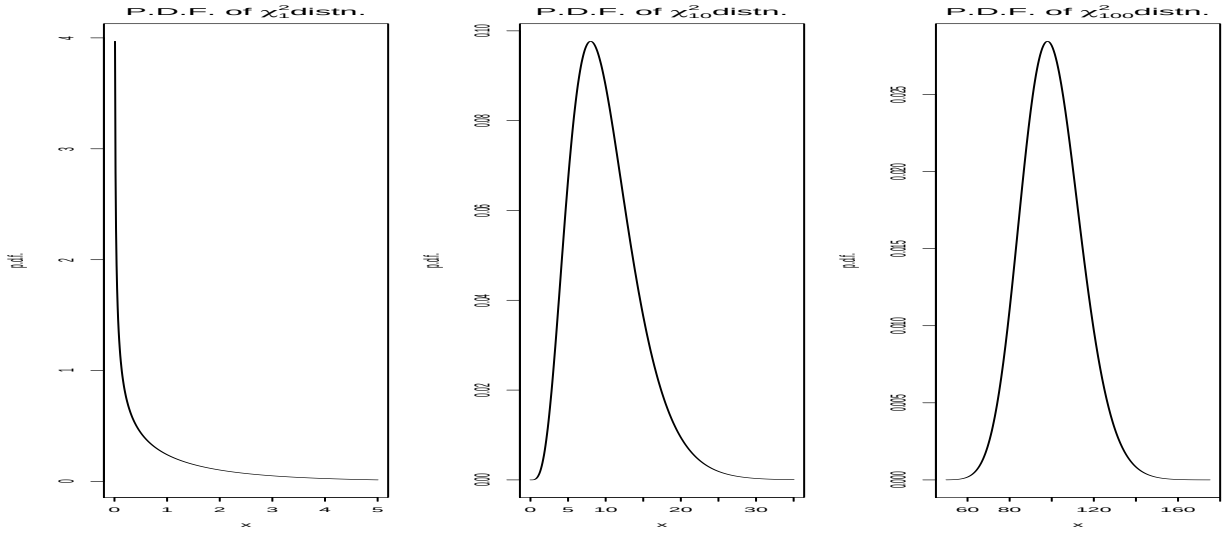
To put things in a little bit more concrete terms consider the sample mean $\overline{Y}$ as an estimate of the population mean μ. Though for a given population its mean μ is something fixed (but unknown), we cannot expect to get the same value of the sample mean $\overline{Y}$ for all different possible samples that we can draw from the population. However we can consider and theoretically derive how $\overline{Y}$ would behave over repeated sampling for all possible samples in terms of its probability distribution. This probability distribution of $\overline{Y}$ over all possible samples is called the sampling distribution of the sample mean $\overline{Y}$. If the sample is drawn from a $N(\mu, \sigma^2)$ population it may be shown that the sampling distribution of $\overline{Y}$ is $N(\mu, \frac{\sigma^2}{n})$.

Noticing a couple of features of this sampling distribution of $\overline{Y}$ should convince us as to why $\overline{Y}$ is a good estimate of μ. First note that the mean of the sampling distribution of $\overline{Y}$ is μ, which coincides with the parameter that $\overline{Y}$ is supposed to estimate. This property of an estimate *viz.* the mean of its sampling distribution coinciding with the parameter it is trying to estimate, is called *unbiasedness*. Loosely speaking, if an estimate is *unbiased*, on an average it hits the target. Thus $\overline{Y}$ is an unbiased estimate of μ. Second the variance of the sampling distribution of $\overline{Y}$ is σ^2/n, which means that as the sample size n increases the variance of $\overline{Y}$ decreases, and as the mean of the sampling distribution of $\overline{Y}$ is μ, this means that for large sample with very high probability the value of $\overline{Y}$ will be concentrated around μ.

Coming back to the estimation of population variance σ^2 from where we left it, it may be shown that the sampling distribution of $(n-1)s^2_{n-1}/\sigma^2$ has a so called χ^2 distribution with $(n-1)$ degrees of freedom (d.f.). [3] The mean and variance of a χ^2_ν variable are ν and 2ν respectively. From this result it may be seen that s^2_{n-1} is an unbiased estimated of σ^2 and thus is the preferred estimate over $s^2_n = \frac{1}{n}\sum_{i=1}^{n}(Y_i - \overline{Y})^2$, which slightly underestimates σ^2.

After settling the estimation issue we next turn our attention towards hypothesis testing. Methodologically the sampling distributions are again used for this purpose. However the logic and the related concepts of hypothesis testing is somewhat subtle which requires special

[3] **Definition:** If $Z_1, Z_2, \ldots, Z_\nu$ are independent and identically distributed $N(0,1)$ variables, $X = \sum_{i=1}^{\nu} Z_i^2$ is said to have a χ^2 distribution with ν d.f. and we write $X \sim \chi^2_\nu$.
Typical p.d.f.'s of some χ^2 distributions are plotted in the following figures:



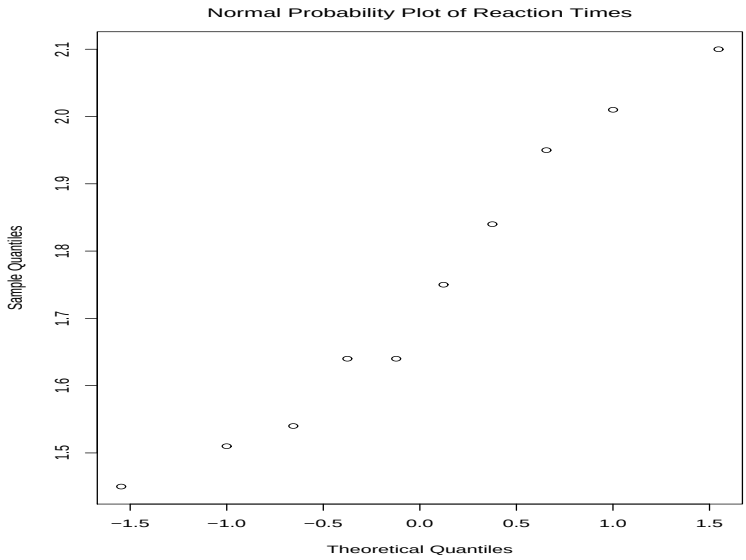
care and attention to follow. First consider hypothesis testing about the population mean μ. To see how the situation of hypothesis testing arises let us consider a concrete example as follows.

Example 1: Suppose a safety feature has been implemented in a vehicle design which will work if the reaction time of the driver is less than 2 seconds. Otherwise the feature would be useless. Thus experiments were conducted to measure the reaction times of 10 drivers which are as follows:

$$2.01,\ 1.54,\ 1.95,\ 2.10,\ 1.64,\ 1.75,\ 1.51,\ 1.84,\ 1.64,\ 1.45$$

Furthermore from similar experiments it may be assumed that the variance of the reaction times $\sigma^2 = 0.04$. The question we need to answer is, is it worth implementing the safety feature?

This is a typical hypothesis testing problem. Here we formulate the problem as follows. The first question to answer is can we assume that the reaction times of the drivers has a Normal distribution? For this we prepare the NPP as follows:



Since the plot looks reasonably linear, we next proceed to model the reaction times with a $N(\mu, 0.02^2)$ distribution. Now with this formulation, we may say that if $\mu < 2$ the safety feature will be useful for more than 50% of the drivers in the population and thus would be worth it. Thus based on the sampled data we are to decide whether it is reasonable to assume that $\mu < 2$ or not.

The point we want prove is $\mu < 2$ and this becomes what is called our *alternative hypothesis* or H_a. Unless the data says otherwise we will go with the *status quo* of the situation which will say that the average reaction times of drivers is at least 2 minutes. This gives us our *null hypothesis* H_0. That is for this problem the next stage of formulation is to decide upon the pair of hypotheses

$$\begin{aligned} H_0 &: \quad \mu \geq 2 \\ H_a &: \quad \mu < 2 \end{aligned}$$

To decide upon one of the competing hypotheses regarding population mean, it is only but natural to look at the sample mean. The basic intuitive idea behind deciding on one

hypothesis over the other is to look at the value of the sample mean $\overline{Y}$ and see to which region does it fall - H_0 or H_a? That is loosely speaking one should reject H_0 if $\overline{Y}$ is small compared to 2 and accept it otherwise. However care must be taken while trying to do this. The regions specified in the hypothesis pertain to that of a population parameter while the value of $\overline{Y}$ we get, which is 1.743 in this example, is specific to this sample at hand. For the same problem if we had another sample of 10 drivers we would have possibly observed a different sample mean. To resolve this issue of all possible values of the sample means we must look at its sampling distribution. But its sampling distribution, which we know is $N(\mu, 0.2^2/10)$ depends on the unknown value of μ, about which we are trying to test the hypotheses. Thus even if H_0 is true there is a possibility for $\overline{Y}$ to be less than 2 and similarly there is a possibility for $\overline{Y}$ to be more than 2 even when H_a is true.

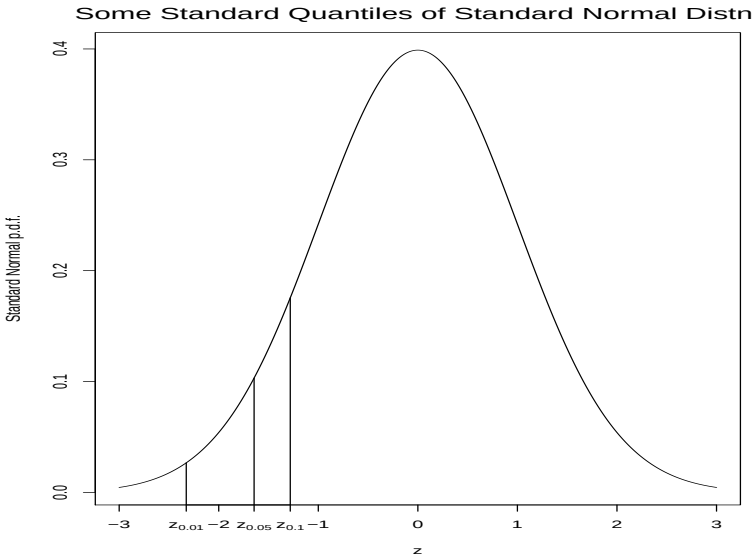
Thus though we decide to reject H_0 for $\overline{Y}$ small compared to 2, there is a possibility of rejecting the right hypothesis. Similarly there is the possibility of wrongly accepting H_0. This dilemma is overcome by considering the probabilities of committing these two types of errors. The former *i.e.* Rejecting H_0 when it is true is called Type-I error, and the later *i.e.* Accepting H_0 when it is false is called Type-II error. In classical hypothesis testing, as mentioned above, since the null hypothesis is always taken to be the *status quo* of a situation *i.e.* assuming it to be true by default unless proven otherwise, and the alternative is a point one wants to prove, Type-I error is deemed to be more expensive than the Type-II error. Thus the optimal decision is taken by fixing a low value for the probability of committing a Type-I error, called α. [4]

That is the optimum decision rule would be: Reject H_0 if $\overline{Y} < \overline{Y}_c$, where $\overline{Y}_c$ is found by fixing an α. With this decision rule $P(\text{Type-I Error}) = P(\overline{Y} < \overline{Y}_c | \mu \geq 2)$ and $P(\text{Type-II Error}) = P(\overline{Y} \geq \overline{Y}_c | \mu < 2)$. Note that the "probability" is coming in Type-I/II error because the decision rule is stochastic or the occurrence of the event $\overline{Y} < \overline{Y}_c$ (or its compliment) depends on the sampling distribution of $\overline{Y}$.

Thus now we are finally ready to carry out the hypothesis test. To do this, only thing remaining is to find $\overline{Y}_c$, the critical value of $\overline{Y}$, which is found as a solution to the equation $P(\overline{Y} < \overline{Y}_c | \mu \geq 2) = \alpha$. However note that though there is no unique solution to this problem (because the left hand side assumes different values for different values of $\mu \geq 2$) the left hand side of the equation attains a maximum when $\mu = 2$. Thus if we fix some small value of α and solve this equation for $\mu = 2$, which incidentally has a unique solution, the $P(\text{Type-I Error})$ for other values μ in H_0 can only be less than α. Thus the desired $\overline{Y}_c$ is determined by solving $P(\overline{Y} < \overline{Y}_c | \mu = 2) = \alpha$ for a given value of α. Or in other words in the present problem, $\overline{Y}_c$ is nothing but the α-th quantile of a $N(\mu, 0.2^2/10)$ distribution.

As mentioned in the last paragraph of §2.3, α-th quantile of an arbitrary $N(\mu, \sigma^2)$ distribution can be found as $\mu + \sigma z_\alpha$, where z_α is the α-th quantile of a standard Normal distribution. Some standard quantiles of the standard Normal distribution is plotted in the following figure:

[4] So what happens to the probability of committing a Type-II error? Among all possible decision rules or tests, which has the same level of α, that test is said to be *most powerful* which has the minimum probability of Type-II error. For example, for the problem of Normal mean μ, it may also be interpreted as the population median and a test may be formulated in terms of the population median. However it can be shown that tests based on the mean has the least probability of Type-II error or most powerful and thus the preferred way of testing.



Thus for $\alpha = 0.01$, 0.05 and 0.1, $\overline{Y}_c$ may be computed as 1.853, 1.896 and 1.919 respectively. That is for the given problem we will Reject H_0 even for $\alpha = 0.01$. Instead of stating the rule of rejecting H_0 directly in terms $\overline{Y}_c$ it is customary to state it in the following standardized form: Reject H_0 if $Z = \frac{\overline{Y}-\mu_0}{\sigma/\sqrt{n}} < z_\alpha$, where μ_0 is the boundary value specified in the null hypothesis and σ is the (supposedly known) population variance. For the numerical example at hand, the observed value of Z called $Z_{observed}$ equals $\frac{1.743-2}{0.2/\sqrt{n}} = -4.063$ and since it is less than -2.326 ($=Z_{0.01}$) our decision would be to reject H_0.

A slightly different but equivalent (and the preferred) way of expressing the same result is in terms of what is called the "observed significance level" or p-value. In general p-value gives the probability of observing a data as "extreme" as the one at hand, under H_0, where what is "extreme" is determined by the alternative hypothesis H_a. The logic behind the p-value is as follows. As seen before, since H_0 is the *status quo* of a situation, the benefit of doubt would go to H_0. Or in other words, unless proven otherwise the verdict would go in favour of the null hypothesis H_0. (This is exactly analogous to the situation of an accused being considered to be not guilty in jurisprudence, which is the null hypothesis. Strong evidence must be provided for rejecting this null hypothesis, and thus accepting the alternative of the accused being guilty.) Thus we start with assuming the null hypothesis to be true. Then decide what kind of phenomenon constitutes an evidence against H_0. For the example at hand, a small value of $\overline{Y}$ (compared to 2) would constitute such an evidence. Then we see what has the data at hand said about this phenomenon. For the given example, the observed data yields an *observed* value of 1.743 for $\overline{Y}$, or $\overline{Y}_{observed} = 1.743$. Now (by assuming H_0 to be true) we compute what is the probability of observing such a phenomenon as the observed data has depicted. For the example, this amounts to calculating $P(\overline{Y} < \overline{Y}_{observed} = 1.743|\mu = 2)$ which is same $P(Z < Z_{observed} = -4.063) = 0.00002$. This is the p-value of this test. Now let us try to interpret this p-value. The p-value is saying that if indeed H_0 were true, *i.e.* if indeed the population mean were 2 or more, the chance that we would observe a sample mean of 1.743 or less is about 2 in a lakh. This is a very small chance of occurrence. But however such a phenomenon has happened in terms of the observed data. So what is going wrong? In the calculation of the probability the assumption that H_0 is true. Thus the p-value gives a very compelling evidence against H_0, which should lead to its rejection. Note that since the

p-value $= P(Z < Z_{observed})$ and previously we were rejecting H_0 if $Z_{observed} < z_\alpha$, these two methods are identical with the rejection rule: Reject H_0 if p-value$< \alpha$. Though these two approaches are thus equivalent, it is always better to provide the p-value or the "observed significance level" (instead just accepting/rejecting H_0 for a given fixed value of α or the so-called fixed significance level testing), because in a way p-value provides the amount of evidence the observed data is carrying against H_0 in 0-1 scale, with the credibility of H_0 being proportional to the numerical value of the p-value.

The above example illustrates the way we test the hypotheses $\begin{matrix} H_0: & \mu \geq \mu_0 \\ H_{a1}: & \mu < \mu_0 \end{matrix}$ for a Normal mean with a given known variance σ^2 for a given hypothesized value μ_0 of μ. The mechanics of the test may be stated as follows. Either, fix a significance level or maximum value of probability of committing a Type-I error α, and then reject H_0 if $Z_{observed} = \frac{\overline{Y}_{observed} - \mu_0}{\sigma/\sqrt{n}} < z_\alpha$. Or, better provide the p-value$=P(Z < Z_{observed})$, where $Z_{observed}$ is as defined in the previous sentence, to the user who can then compare it with his/her personal α to arrive at a decision.

There are two other common types of alternative hypotheses which are also tested for Normal mean (with a given known variance σ^2). These hypotheses and the mechanics of testing them are presented in the following table:

Hypotheses	Fixed Significance Level Testing	p-value
$\begin{matrix} H_0: & \mu \leq \mu_0 \\ H_{a2}: & \mu > \mu_0 \end{matrix}$	Reject H_0 if $Z_{observed} = \frac{\overline{Y}_{observed} - \mu_0}{\sigma/\sqrt{n}} > z_{1-\alpha}$	$P(Z > Z_{observed})$
$\begin{matrix} H_0: & \mu = \mu_0 \\ H_{a3}: & \mu \neq \mu_0 \end{matrix}$	Reject H_0 if $\lvert Z_{observed}\rvert = \left\lvert \frac{\overline{Y}_{observed} - \mu_0}{\sigma/\sqrt{n}} \right\rvert > z_{1-\alpha/2}$	$2P(Z > \lvert Z_{observed}\rvert)$

While the logic of testing for H_{a2} is exactly analogous to that of H_{a1}, just note that in this case one should reject H_0 if $\overline{Y}_{observed}$ is large compared to μ_0, and how large is large is determined by α. Thus for a fixed α in this case $\overline{Y}_c$ is such that $P(\overline{Y} > \overline{Y}_c | \mu = \mu_0) = \alpha$. Transforming this to Z-computation yields $\overline{Y}_c = \mu_0 + (\sigma/\sqrt{n}) z_{1-\alpha}$ and the above rejection rule. For H_{a3} the test is called a two-tailed test. In this case H_0 should be rejected if $Z_{observed}$ is far away from 0 in either direction. Since the distribution of Z is symmetric about 0 this leads to a symmetric rejection region. That is in this case Z_c is such that $P(|Z| > Z_c) = \alpha$. Now using the symmetry of the standard Normal distribution about 0, it may be seen that $Z_c = z_{1-\alpha/2}$, which yields the given rejection rule as listed in the above table for H_{a3}.

Hypothesis testing about the Normal population variance σ^2 like wise is carried out using the sample variance s^2_{n-1}. The logic is exactly same as before *i.e.* for some given σ^2_0, we reject H_0 for small s^2_{n-1} value compared to σ^2_0 while testing for $H_{a1}: \sigma^2 < \sigma^2_0$; we reject H_0 for large s^2_{n-1} value compared to σ^2_0 while testing for $H_{a2}: \sigma^2 > \sigma^2_0$; and we reject H_0 for either small or large s^2_{n-1} value compared to σ^2_0 while testing for $H_{a3}: \sigma^2 \neq \sigma^2_0$. The "large" or "small" values are determined by a pre-fixed value of α and sampling distribution of $(n-1)s^2_{n-1}/\sigma^2_0$, which has a χ^2_{n-1} distribution under H_0. The rejection rules and the corresponding p-value expressions for testing for Normal variance is summarized in the following table:

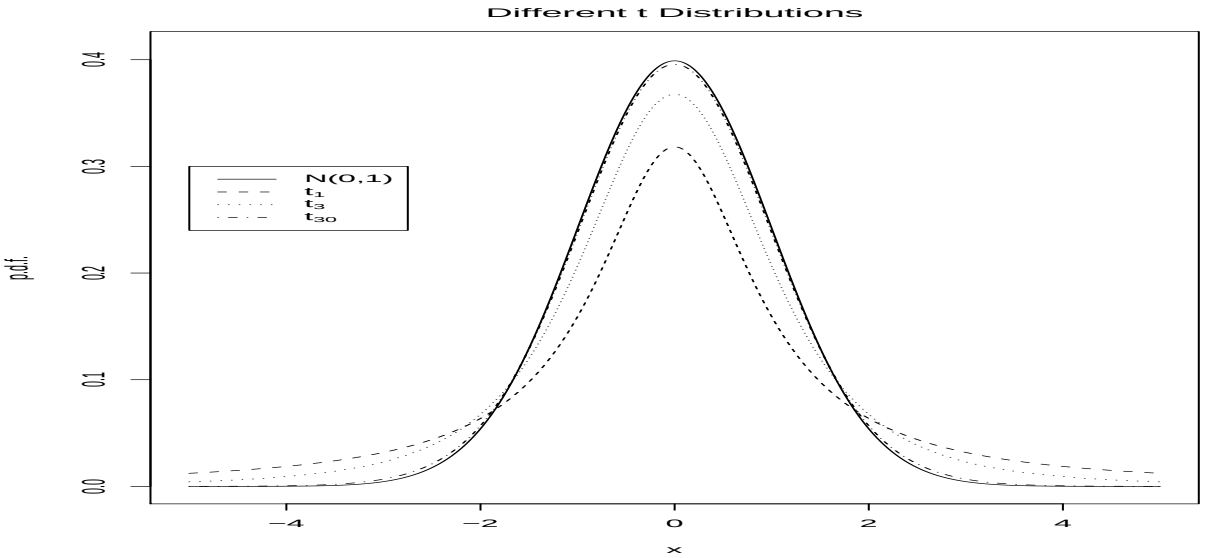
Hypotheses	Fixed Significance Level Testing	p-value
$H_0: \sigma^2 \geq \sigma_0^2$ $H_{a1}: \sigma^2 < \sigma_0^2$	Reject H_0 if $(n-1)\frac{s_{n-1}^2}{\sigma_0^2} < \chi^2_{n-1,\alpha}$	$P\left(\chi^2_{n-1} < (n-1)\frac{s_{n-1}^2}{\sigma_0^2}\right)$
$H_0: \sigma^2 \leq \sigma_0^2$ $H_{a2}: \sigma^2 > \sigma_0^2$	Reject H_0 if $(n-1)\frac{s_{n-1}^2}{\sigma_0^2} > \chi^2_{n-1,1-\alpha}$	$P\left(\chi^2_{n-1} > (n-1)\frac{s_{n-1}^2}{\sigma_0^2}\right)$
$H_0: \sigma^2 = \sigma_0^2$ $H_{a3}: \sigma^2 \neq \sigma_0^2$	Reject H_0 if $(n-1)\frac{s_{n-1}^2}{\sigma_0^2} < \chi^2_{n-1,\alpha/2}$ or $(n-1)\frac{s_{n-1}^2}{\sigma_0^2} > \chi^2_{n-1,1-\alpha/2}$	$2p$ of H_{a1} if $s_{n-1}^2 < \sigma_0^2$ $2p$ of H_{a2} if $s_{n-1}^2 > \sigma_0^2$

Unknown Population Variance:

Now let us turn our attention to the case of testing hypothesis about a Normal population mean μ when the the population variance σ^2 is unknown. In this case the logic remains the same, with only difference lying in the use of sampling distribution of $\overline{Y}$. Though the sampling distribution of $\overline{Y}$ is still $N(\mu, \sigma^2/n)$, it can no longer be directly used because σ^2 is unknown. A natural solution to this problem would be to replace σ^2 by its unbiased estimate s^2_{n-1}. However this changes the resulting distribution to a so-called t-distribution.

Definition: If $Z \sim N(0,1)$ and $X \sim \chi^2_\nu$ and Z and X are independent, then the variable $T = \frac{Z}{\sqrt{X/\nu}}$ is said to have a t distribution with ν degrees of freedom (d.f.) and is written as $T \sim t_\nu$.

The t-distribution in a way can be thought as a generalization of the standard Normal distribution, except with fatter tails. As a matter of fact a t distribution with ∞ degrees of freedom is same as the standard Normal distribution. P.d.f.'s of different t distributions are plotted in the following figure:



Based on the t distribution the testing for Normal mean in case of unknown population variance can be summarized as follows:

Hypotheses	Fixed Significance Level Testing	p-value
$H_0: \mu \geq \mu_0$ $H_{a1}: \mu < \mu_0$	Reject H_0 if $t_{observed} = \frac{\overline{Y}_{observed} - \mu_0}{s_{n-1}/\sqrt{n}} < t_{n-1,\alpha}$	$P(t_{n-1} < t_{observed})$
$H_0: \mu \leq \mu_0$ $H_{a2}: \mu > \mu_0$	Reject H_0 if $t_{observed} = \frac{\overline{Y}_{observed} - \mu_0}{s_{n-1}/\sqrt{n}} > t_{n-1,1-\alpha}$	$P(t_{n-1} > t_{observed})$
$H_0: \mu = \mu_0$ $H_{a3}: \mu \neq \mu_0$	Reject H_0 if $\lvert t_{observed}\rvert = \left\lvert \frac{\overline{Y}_{observed} - \mu_0}{s_{n-1}/\sqrt{n}} \right\rvert > t_{n-1,1-\alpha/2}$	$2P(t_{n-1} > \lvert t_{observed}\rvert)$

To see why $\frac{\overline{Y}-\mu}{s_{n-1}/\sqrt{n}}$ has a t-distribution with $n-1$ d.f., observe that $\overline{Y} \sim N(\mu, \sigma^2/n)$ and $(n-1)\frac{s^2_{n-1}}{\sigma^2} \sim \chi^2_{n-1}$ (and they can be shown to be independent). Thus $\frac{\overline{Y}-\mu}{s_{n-1}/\sqrt{n}} = \frac{(\overline{Y}-\mu)/(\sigma/\sqrt{n})}{\sqrt{\{(n-1)s^2_{n-1}/\sigma^2\}/(n-1)}} = \frac{Z}{\sqrt{\chi^2_{n-1}/(n-1)}}$ which by definition has a t distribution with $n-1$ d.f.

Example 1 (Revisited): If we did not have any information about population variance σ^2 we would have done a t-test in this case. Here $\overline{Y}_{observed} = 1.743$ and $s_{n-1} = 0.2245$. This yields a $t_{observed} = \frac{1.743-2}{0.2245/\sqrt{10}} = -3.62$. Since here it is a left-tailed test of H_{a1}, comparing the observed t-value of -3.62 with $t_{9,0.01} = -2.821$, $t_{9,0.05} = -1.833$ and $t_{9,0.1} = -1.383$ we can again safely reject H_0. More precisely the p-value $P(t_9 < -3.62)$ is given by 0.0028, indicating that if indeed the null hypothesis were true there is only a chance of 28 in 10,000 of observing a data set as the one we have got, leading to a possible rejection of H_0.

3.3 Inference for Two Normal Populations

Now suppose we have two variables Y_1 and Y_2 both of which are Normally distributed with $Y_1 \sim N(\mu_1, \sigma^2_1)$ and $Y_2 \sim N(\mu_2, \sigma^2_2)$. Also suppose we have two samples of sizes n_1 and n_2 respectively from these two population of values of Y_1 and Y_2 with $Y_{11}, Y_{12}, \ldots, Y_{1n_1}$ and $Y_{21}, Y_{22}, \ldots, Y_{2n_2}$ as the respective samples. The main problem of interest is to compare the two population means μ_1 and μ_2 in terms of their difference $\mu_1 - \mu_2$, with a passing interest on the comparison of the two population variances σ^2_1 and σ^2_2 in terms of their ratio σ^2_1/σ^2_2. Let $\overline{Y}_1 = \frac{1}{n_1}\sum_{j=1}^{n_1} Y_{1j}$ and $\overline{Y}_2 = \frac{1}{n_2}\sum_{j=2}^{n_2} Y_{2j}$ denote the two sample means and $s^2_1 = \frac{1}{n_1-1}\sum_{j=1}^{n_1}(Y_{1j} - \overline{Y}_1)^2$ and $s^2_2 = \frac{1}{n_2-1}\sum_{j=1}^{n_2}(Y_{2j} - \overline{Y}_2)^2$ denote the two sample variances respectively.

The methods of inference for $\mu_1 - \mu_2$ now depend on the kind of assumptions we are willing to make about the population variances σ^2_1 and σ^2_2. This is because the sampling distribution of $\overline{Y}_i$ is $N(\mu_i, \sigma^2_i/n_i)$ for $i = 1, 2$ and thus the sampling distribution of $\overline{Y}_1 - \overline{Y}_2$, the point estimate of the parameter of interest $\mu_1 - \mu_2$, is $N(\mu_1 - \mu_2, \sigma^2_1/n_1 + \sigma^2_2/n_2)$, which critically depends on the two population variances σ^2_1 and σ^2_2. This leads to considerations of various cases which are as follows.

Case I: Known σ^2_1 and σ^2_2

Here for drawing inference on $\mu_1 - \mu_2$ we look at the quantity $\frac{(\overline{Y}_1 - \overline{Y}_2) - (\mu_1 - \mu_2)}{\sqrt{\frac{\sigma^2_1}{n_1} + \frac{\sigma^2_2}{n_2}}}$ which has a standard Normal $N(0,1)$ distribution. This is utilized for both testing of hypothesis and

interval estimation [5] of the parameter $\mu_1 - \mu_2$. The test procedure for the three different scenarios of alternative hypotheses are summarized in the following table:

Hypotheses	Fixed Significance Level Testing	p-value
$H_0: \quad \mu_1 - \mu_2 \geq \mu_0$ $H_{a1}: \quad \mu_1 - \mu_2 < \mu_0$	Reject H_0 if $Z_{obs} = \frac{(\overline{Y}_1 - \overline{Y}_2) - \mu_0}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}} < z_\alpha$	$P(Z < Z_{obs})$
$H_0: \quad \mu_1 - \mu_2 \leq \mu_0$ $H_{a2}: \quad \mu_1 - \mu_2 > \mu_0$	Reject H_0 if $Z_{obs} = \frac{(\overline{Y}_1 - \overline{Y}_2) - \mu_0}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}} > z_{1-\alpha}$	$P(Z > Z_{obs})$
$H_0: \quad \mu_1 - \mu_2 = \mu_0$ $H_{a3}: \quad \mu_1 - \mu_2 \neq \mu_0$	Reject H_0 if $Z_{obs} = \left\lvert \frac{(\overline{Y}_1 - \overline{Y}_2) - \mu_0}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}} \right\rvert > z_{1-\alpha/2}$	$2P(Z > Z_{obs})$

In most practical applications the μ_0 value we are typically interested in 0. A $100(1-\alpha)\%$ confidence interval for $\mu_1 - \mu_2$ is given by $(\overline{Y}_1 - \overline{Y}_2) \pm z_{1-\alpha/2}\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}$.

Case II: σ_1^2 and σ_2^2 Unknown but Large Samples, $n_1, n_2 \geq 30$

In this case the test procedure and the confidence interval formula are exactly as that of Case I, except here σ_1^2 and σ_2^2 are replaced by their unbiased sample estimates s_1^2 and s_2^2 respectively.

Case III: σ_1^2 and σ_2^2 Unknown but Equal, $\sigma_1^2 = \sigma_2^2$

To begin with, a few remarks are in order about this assumption of equality of population variances. First, the technical term for this assumption is called *homoscedasticity*. Second, while comparing the means of two Normal Populations, this is the most interesting or interpretable case. To see why, refer back to the figure in page 4. There suppose one of the populations is the standard Normal (solid line), and for the second population consider two cases - $N(-2,1)$ (dot-dashed line) and $N(1,2^2)$ (dashed line). When one compares $N(0,1)$ to $N(-2,1)$, the homoscedastic case, there is a clear-cut difference, namely it can be said that the values coming from the $N(0,1)$ population are in general larger than the values coming from the $N(-2,1)$ population. However when one compares $N(0,1)$ to $N(1,2^2)$, the

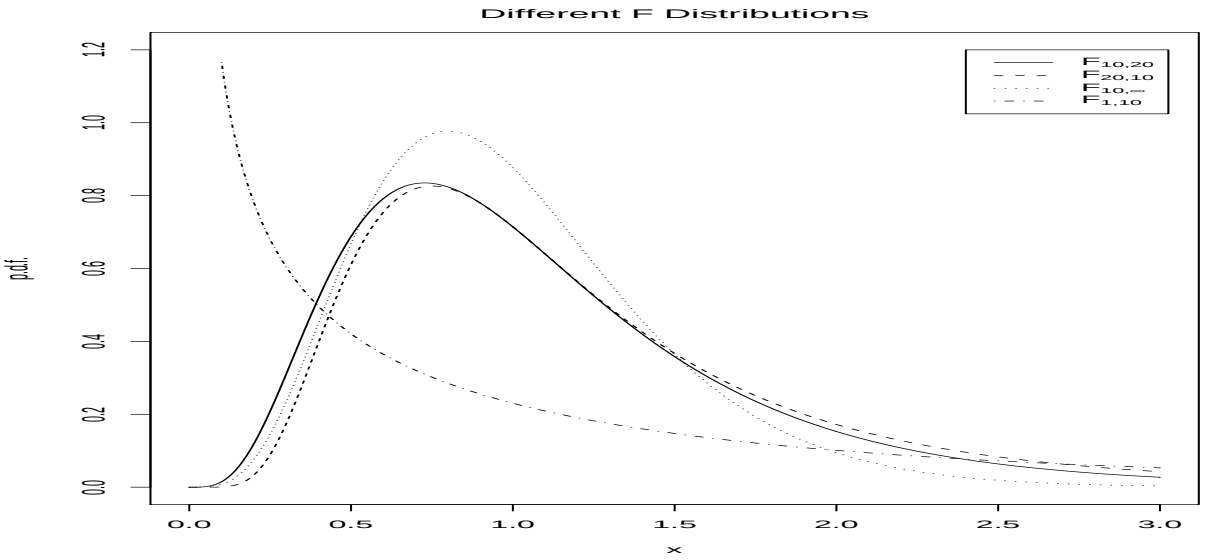
[5] We have not talked about interval estimation in the earlier case of a single Normal distribution in §3.2. There, for known σ^2, $\frac{\overline{Y}-\mu}{\sigma/\sqrt{n}}$ has a standard Normal distribution. Thus for a desired confidence level of $1-\alpha$, one could write that $P(z_{\alpha/2} < \frac{\overline{Y}-\mu}{\sigma/\sqrt{n}} < z_{1-\alpha/2}) = 1-\alpha$, and then after rearranging the terms and exploiting the fact that $z_{\alpha/2} = -z_{1-\alpha/2}$ one could rewrite the above probability statement as $P(\overline{Y} - z_{1-\alpha/2}\frac{\sigma}{\sqrt{n}} < \mu < \overline{Y} + z_{1-\alpha/2}\frac{\sigma}{\sqrt{n}}) = 1-\alpha$. This gives a $100(1-\alpha)\%$ confidence interval for μ. However caution is required for interpreting the above probability statement or confidence intervals in general. In the above probability statement, what is random is $\overline{Y}$ and not μ. Thus the probability statement says that if one keeps on using the random interval $\overline{Y} \pm z_{1-\alpha/2}\frac{\sigma}{\sqrt{n}}$ repeatedly for all possible samples of size n that can be drawn from the population of Y values, then $100(1-\alpha)\%$ of the time this random interval will successfully be able to capture the unknown value of μ. Similarly a $100(1-\alpha)\%$ confidence interval for σ^2 may be obtained as $\left[\frac{(n-1)s_{n-1}^2}{\chi^2_{n-1,1-\alpha/2}}, \frac{(n-1)s_{n-1}^2}{\chi^2_{n-1,\alpha/2}}\right]$, by utilizing the sampling distribution of $\frac{(n-1)s_{n-1}^2}{\sigma^2} \sim \chi^2_{n-1}$. For unknown σ^2, $\frac{\overline{Y}-\mu}{s_{n-1}/\sqrt{n}} \sim t_{n-1}$. Thus a $100(1-\alpha)\%$ confidence interval for μ in this case may be obtained as $\overline{Y} \pm t_{n-1,1-\alpha/2}\frac{s_{n-1}}{\sqrt{n}}$.

heteroscedastic case, there is no clear way of saying population of which values are in general larger or smaller than the other. Thus for the ease of interpretation, whenever possible, we prefer to compare two population means under the umbrella of homoscedasticity.

This leads to the responsibility of first checking this assumption of homoscedasticity. That is before applying the methods applicable for this case, we must first ensure that indeed the two population variances σ_1^2 and σ_2^2 are equal. But note that this brings us to the problem of comparing two population variances. For this comparison the natural statistic to look at is the ratio of the two sample variances s_1^2/s_2^2. If this ratio is "close" to 1 then we may safely assume homoscedasticity, otherwise we will be forced to reject the (null) hypothesis of equality of two population variances. That immediately raises the question of how "close" is "close" and for the answer we already know what to do. We need to figure out the kind of behavior one may expect the ratio of sample variances to exhibit under the assumption of homoscedasticity, or in other words we need to know the sampling distribution of s_1^2/s_2^2. This now leads to the final sampling distribution that we need to study for the Normal models, called the F distribution defined below.

Definition: If $U \sim \chi^2_{\nu_1}$ and $V \sim \chi^2_{\nu_2}$ and U and V are independent, then the variable $X = \frac{U/\nu_1}{V/\nu_2}$ is said to have an F distribution with numerator degrees of freedom ν_1 and denominator degrees of freedom ν_2 and is written as $X \sim F_{\nu_1,\nu_2}$.

The p.d.f's of some typical F distributions are plotted in the following figure:



As the $N(0,1)$ or t distributions are studied with symmetry around 0 in mind, the F distribution is studied around 1. Here also there is a symmetry around 1, but it is in a reciprocal scale. A few other properties of the F distribution would be illuminating in its application.

Note that by definition, for the numerator degrees of freedom equal to 1, $F_{1,\nu} = \left(\frac{Z}{\chi^2_\nu/\nu}\right)^2 = t^2_\nu$. Or in other words an F distribution with 1 numerator degrees of freedom is same as the square of a t distribution with its d.f. same as the denominator d.f. of the F. Next note that $1/F_{\nu_1,\nu_2}$ is same as F_{ν_2,ν_1}. Also note that $F_{\nu,\infty}$ is same as χ^2_ν/ν and thus $F_{\infty,\nu}$ is same

as the reciprocal of χ^2_ν/ν. Finally note that $F_{\infty,\infty}$ is a degenerate variable concentrating all its probability mass at 1.

Now coming back to the problem of comparing two population variances, note that $\frac{(n_1-1)s_1^2}{\sigma_1^2} \sim \chi^2_{n_1-1}$ and $\frac{(n_2-1)s_2^2}{\sigma_2^2} \sim \chi^2_{n_2-1}$ and they are independent by sampling design. Thus $\frac{\sigma_2^2}{\sigma_1^2}\frac{s_1^2}{s_2^2}$ has an F_{n_1-1,n_2-1} distribution. Utilizing this, the test of hypotheses comparing two population variances are summarized in the following table:

Hypotheses	Fixed Significance Level Testing	p-value
$H_0: \frac{\sigma_1^2}{\sigma_2^2} \geq \sigma_0^2$ $H_{a1}: \frac{\sigma_1^2}{\sigma_2^2} < \sigma_0^2$	Reject H_0 if $\frac{s_1^2}{\sigma_0^2 s_2^2} < F_{n_1-1,n_2-1,\alpha}$	$P\left(F_{n_1-1,n_2-1} < \frac{s_1^2}{\sigma_0^2 s_2^2}\right)$
$H_0: \frac{\sigma_1^2}{\sigma_2^2} \leq \sigma_0^2$ $H_{a2}: \frac{\sigma_1^2}{\sigma_2^2} > \sigma_0^2$	Reject H_0 if $\frac{s_1^2}{\sigma_0^2 s_2^2} > F_{n_1-1,n_2-1,1-\alpha}$	$P\left(F_{n_1-1,n_2-1} > \frac{s_1^2}{\sigma_0^2 s_2^2}\right)$
$H_0: \frac{\sigma_1^2}{\sigma_2^2} = \sigma_0^2$ $H_{a3}: \frac{\sigma_1^2}{\sigma_2^2} \neq \sigma_0^2$	Reject H_0 if $\frac{s_1^2}{\sigma_0^2 s_2^2} < F_{n_1-1,n_2-1,\alpha/2}$ or $\frac{s_1^2}{\sigma_0^2 s_2^2} > F_{n_1-1,n_2-1,1-\alpha/2}$	$2p$ of H_{a1} if $\frac{s_1^2}{s_2^2} < \sigma_0^2$ $2p$ of H_{a2} if $\frac{s_1^2}{s_2^2} > \sigma_0^2$

For testing for homoscedasticity we shall test H_0 against the alternative H_{a3} with $\sigma_0^2 = 1$. A $100(1-\alpha)\%$ confidence interval for σ_1^2/σ_2 is given $\left[(s_1^2/s_2^2)F_{n_2-1,n_1-1,\alpha/2}, (s_1^2/s_2^2)F_{n_2-1,n_1-1,1-\alpha/2}\right]$. Note that for any ν_1 and ν_2 $F_{\nu_2,\nu_1,\alpha} = 1/F_{\nu_1,\nu_2,1-\alpha}$.

If H_0 cannot be rejected against H_{a3} with $\sigma_0^2 = 1$, then we know that we are in Case III. In this situation the next thing one does is estimate this common variance $\sigma^2 = \sigma_1^2 = \sigma_2^2$. A Uniformly Minimum Variance Unbiased Estimate (called UMVUE [6]) for this common variance σ^2 is given by $s_p^2 = \frac{(n_1-1)s_1^2+(n_2-1)s_2^2}{n_1+n_2-2}$, s_p^2 indicates a *pooled*-estimate. It can be shown that $(n_1+n_2-2)\frac{s_p^2}{\sigma^2} \sim \chi^2_{n_1+n_2-2}$ and is independent of $\overline{Y}_1$ and $\overline{Y}_2$. Thus if we replace σ_1^2 and σ_2^2 by their common estimate s_p^2 in the Z-formula of Case I, we get $t = \frac{(\overline{Y}_1-\overline{Y}_2)-(\mu_1-\mu_2)}{s_p\sqrt{\frac{1}{n_1}+\frac{1}{n_2}}} \sim t_{n_1+n_2-2}$. These t-tests (called pooled t-tests) for the hypotheses involving $\mu_1-\mu_2$ are summarized in the following table:

Hypotheses	Fixed Significance Level Testing	p-value
$H_0: \mu_1-\mu_2 \geq \mu_0$ $H_{a1}: \mu_1-\mu_2 < \mu_0$	Reject H_0 if $t_{obs} = \frac{(\overline{Y}_1-\overline{Y}_2)-\mu_0}{s_p\sqrt{\frac{1}{n_1}+\frac{1}{n_2}}} < t_{n_1+n_2-2,\alpha}$	$P(t_{n_1+n_2-2} < t_{obs})$
$H_0: \mu_1-\mu_2 \leq \mu_0$ $H_{a2}: \mu_1-\mu_2 > \mu_0$	Reject H_0 if $t_{obs} = \frac{(\overline{Y}_1-\overline{Y}_2)-\mu_0}{s_p\sqrt{\frac{1}{n_1}+\frac{1}{n_2}}} > t_{n_1+n_2-2,1-\alpha}$	$P(t_{n_1+n_2-2} > t_{obs})$
$H_0: \mu_1-\mu_2 = \mu_0$ $H_{a3}: \mu_1-\mu_2 \neq \mu_0$	Reject H_0 if $t_{obs} = \left\lvert\frac{(\overline{Y}_1-\overline{Y}_2)-\mu_0}{s_p\sqrt{\frac{1}{n_1}+\frac{1}{n_2}}}\right\rvert > t_{n_1+n_2-2,1-\alpha/2}$	$2P(t_{n_1+n_2-2} > t_{obs})$

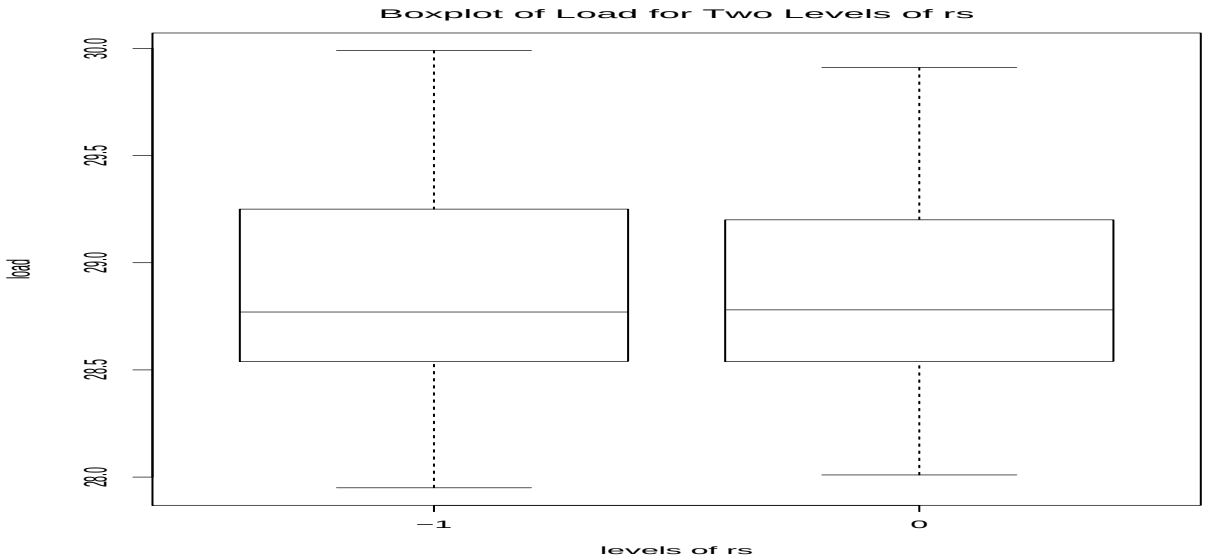
A $100(1-\alpha)\%$ confidence interval for $\mu_1-\mu_2$ is given by $(\overline{Y}_1-\overline{Y}_2) \pm t_{n_1+n_2-2,1-\alpha/2}s_p\sqrt{\frac{1}{n_1}+\frac{1}{n_2}}$.

[6] All the estimates we have considered so far, like $\overline{Y}$ for μ and s^2_{n-1} for σ^2 in a single Normal population are UMVUE, which is an estimate with minimum variance among the class of all unbiased estimates. For any population parameter, a UMVUE is the most desirable estimate one can think of.

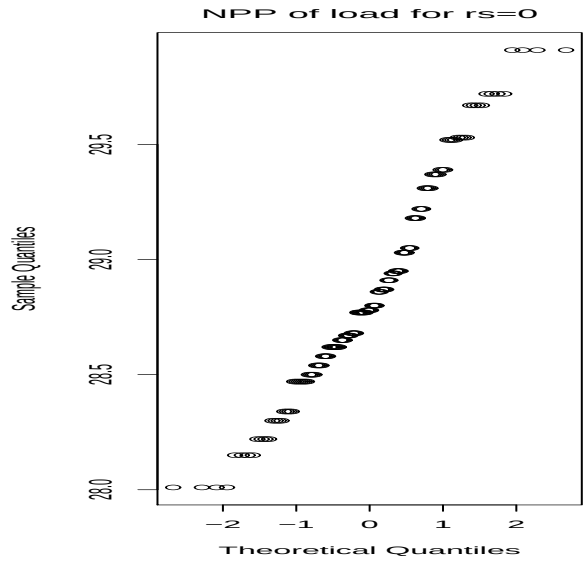
Case IV: σ_1^2 and σ_2^2 Unknown and Unequal, $\sigma_1^2 \neq \sigma_2^2$

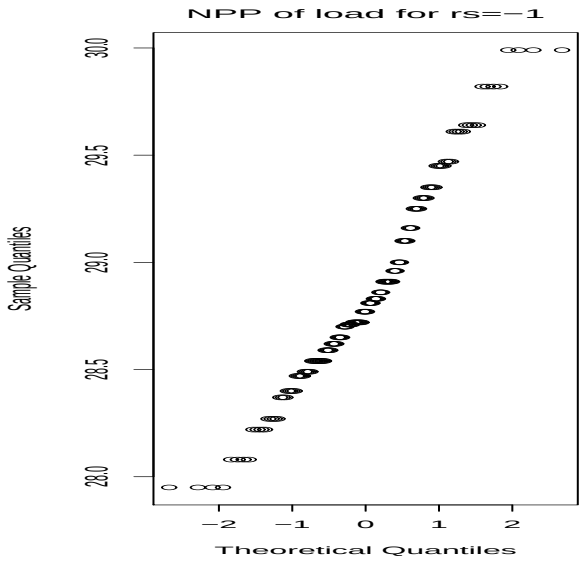
If $H_0 : \sigma_1^2 = \sigma_2^2$ is rejected against the alternative $H_a : \sigma_1^2 \neq \sigma_2^2$ then we cannot use the pooled t-test as discussed above. In this case the resulting t-test is called Welch's t-test. The formula of the t-statistic in this case is identical to Z-statistic that was used in Case II, namely $t = \frac{(\overline{Y}_1 - \overline{Y}_2) - \mu_0}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}}$ (which was the Z-statistic of Case II). But here the sampling distribution of this statistic is *approximated* by a t distribution with ν degrees of freedom, where a conservative value of ν is given by Minimum$\{n_1 - 1, n_2 - 1\}$ while a more accurate value of ν may be computed as $\frac{\left(\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}\right)^2}{\frac{1}{n_1 - 1}\left(\frac{s_1^2}{n_1}\right)^2 + \frac{1}{n_2 - 1}\left(\frac{s_2^2}{n_2}\right)^2}$.

Example 2: Here in an experiment we want to investigate whether the lower column revolution stiffness (say rs) has any effect on the response load on column joint of steering (say ls). For any statistical analysis one must first visually try to assess what is going on before launching any formal analysis. Towards this end we first make some plots:



By looking at this plot it appears that as such rs may not have any effect on load or the distribution of load for the two levels are not very different. This intuitive feeling will be formally tested next. But for this we must test for the assumption of Normality first using NPP. These plots are provided in the next page. From these NPP the load do not look very Normal. So the best possible power transformation was tried, but even that failed to provide a decent NPP. Thus for the sake of illustration we proceed with the formal analysis by first testing for homoscedasticity and then the pooled t-test.






```
> var.test(load~rs)

        F test to compare two variances

data:  load by rs
F = 1.0974, num df = 134, denom df = 134, p-value = 0.5914
alternative hypothesis: true ratio of variances is not equal to 1
95 percent confidence interval:
 0.7810349 1.5418990
sample estimates:
ratio of variances
          1.097395

> t.test(load~rs,var.equal=T)

        Two Sample t-test

data:  load by rs
t = -0.0609, df = 268, p-value = 0.9515
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 -0.1185705  0.1114594
sample estimates:
mean in group -1  mean in group 0
        28.86496         28.86852

> t.test(load~rs)

        Welch Two Sample t-test
```

```
data:  load by rs
t = -0.0609, df = 267.423, p-value = 0.9515
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 -0.1185716  0.1114605
sample estimates:
mean in group -1  mean in group 0
        28.86496         28.86852
```

Thus we may conclude that as such indeed lower column revolution stiffness does not affect on the response load on column joint of steering.