6
Statistical Experiments

Figure 6.1: Statue of the German polymath Gottfried Wilhelm Leibniz at the Göttingen auditorium (*courtesy: Wikimedia Commons*).

Statistical experiments enable us to make *inferences* from data about parameters that characterize a population. Generally speaking, inferences may be of two types, viz., *deductive* inference and *inductive* inference.[1] Deductive inference pertains to conclusions based on a set of premises (propositions) and their synthesis. Deductive reasoning has a definitive character. Eg., All men are mortal (first proposition); Socrates is a man (second proposition); hence, Socrates is mortal (deductive conclusion). On the other hand, inductive inference has a probabilistic character. One conducts an experiment and collects data. Based on this data, certain conclusions are drawn that may have a broader applicability beyond the contours of the particular experiment performed by the researcher. This generalization of the conclusions drawn from the particular experiment constitutes the framework of inductive reasoning. Eg., Measurement of heights of a small group of people belonging to a certain population is conducted. Based on the calculations of this smaller sample set; and upon finding that for this small group the average height of men is greater than the average height of women, it is inferred that men of this population are generally taller than the women.

The formal practice of inductive reasoning dates back to the thesis of Gottfried Wilhelm Leibniz. He was the first to propose that probability is a relation between hypothesis and evidence (data). His thesis was founded on three conceptual pillars: *chance* (probability), *possibilities* (realizable random events), and *ideas* (generalization of inferences by induction).[2] We have encountered the first two concepts in earlier chapters of this textbook. In this chapter, we will delve on the third theme whereby we will discuss methods to draw conclusions from data derived from statistical experiments based on the principles of inductive reasoning.

[1] *Introduction to the Theory of Statistics* by *Alexander Mood, Franklin lightgraybill, and Duane Boes*, McGraw Hill Education (third edition), 2017.

[2] *The Emergence of Probability* by *Ian Hacking*, Cambridge University Press (second edition), 2006.

6.1 Chapter objectives

The chapter objectives are listed as follows.

1. Students will learn the concept of random samples, population parameters, statistics, sampling distributions, and hypothesis.

2. Students will learn to deduce the probability distributions of test statistics and apply these sampling distributions to perform different tests of hypotheses.

3. Students will learn to analyze the asymptotic behavior of certain sampling distributions.

4. Students will learn to conduct analysis of variance (ANOVA).
5. Students will be trained to provide error estimates of inferences and confidence intervals of model parameters in statistical experiments.
6. Students will learn to organize data sourced from reconstruction projects following natural disasters and perform an analysis of variance experiment to prioritize post-disaster reconstruction efforts.

6.2 Chapter project: Prioritizing post-disaster reconstruction measures

6.2.1 Prologue: What factors severely impede the efficient implementation of post-disaster reconstruction efforts?

Reconstruction projects following catastrophic disasters like floods, hurricanes, earthquakes, etc. are often negatively impacted by a plethora of factors. These include availability of capital investment with the local government, availability of manpower resources for conducting rehabilitation and reconstruction efforts, existing laws relating to land acquisition for building temporary and permanent housing for displaced persons, etc.

The goal of this project is to highlight the issues and challenges in Post Disaster Reconstruction (PDR) efforts and to determine the significant differences between the issues and challenges in different locations where PDR projects are carried out. As a chief construction engineer of an international non-governmental organization, you are tasked with devising an emergency strategy for tackling the issues concerning the efficient implementation of PDR projects. Your decision making process relies on an extensive database across six international cities[3] where project engineers have rated the most pressing issues that are responsible for delay in PDR projects. Your first task (and the objective of this project) is to identify those issues that are common across geographical locations and address them as a priority. In order to accomplish this, you are provided with a database of responses by construction engineers from six different international cities. Construction engineers who have worked in these cities in the past have rated the significance of the respective issues on a scale of 1-10 with 1 being *strongly disagree* and 10 being *strongly agree*. The issues that will be investigated are:

- ***shortage of relief workers and technical staff,***
- ***land ownership and related laws,***
- ***funding and aid for PDR projects,* and**
- ***community participation in rebuilding efforts.***

In order to accomplish this task, you will use the *F-statistic* which is constructed by considering two different estimates for the sample variance. The ratings of the construction engineers from different locations (as mentioned above) constitute the sample data for this statistical experiment. Before analyzing this data, we must develop conceptual knowledge about sampling distributions (like the F-distribution which is used in performing an ANOVA experiment).

Figure 6.2: Photograph of the 1906 San Francisco earthquake showing the extent and magnitude of damage to property and mankind (*courtesy: Britannica*). The disaster brought about over 3000 fatalities and the monetary damage was estimated to over US $400 million in 1906 money. The relief and rehabilitation effort involved over 4000 US federal troops who built over 5600 makeshift houses to accommodate over 20000 displaced people in the immediate aftermath of the disaster. Millions of dollars in aid from around the world and from private enterprises within America poured into the city coffers for rehabilitation and reconstruction purposes. The committee entrusted with the rebuilding programs had to deal with costly land acquisition procedures by recommending the use of municipal bonds as financial guarantees.

[3] Port-au-Prince (Haiti), Tacloban City (Philippines), Latur (India), New Orleans (USA), Kathmandu (Nepal) and Bagh City (Pakistan).

6.3 Elements of statistical sampling

The quintessential idea is to conduct our statistical experiments on a smaller group instead of performing them on the entire population. The manner in which this smaller group is selected is crucial to the accuracies of the inferences drawn from the experiments. In this section, we will study some important sampling techniques and some important theoretical results that hold true in the asymptotic limit of large enough sample size.

Figure 6.3: The outcomes of the throws of multiple dice are independent and identically distributed. The outcome of each throw is independent of the other because the kinematic motion of the hands are not dependent on one another as they belong to different individuals. Yet, there is a certain sameness about the structure and motion of the hands (and the shape of the dice) due to which the outcome of each throw results in one of six numbers with the same probability $\frac{1}{6}$ (identically uniformly distributed).

6.3.1 Definition: Random sample

Consider a population $\mathfrak{P}$ from which we will pick our sample of size n randomly. The n random variables $X_1,\ X_2,\ ...,\ X_n$ constitute a *random sample* if they are *independent* and *identically distributed* (i.i.d.), i.e. (i) if the X_is are independent of each other, and (ii) if each of the X_is follow the same probability distribution (identically distributed).

If the sample is not chosen in a random manner from $\mathfrak{P}$, then the statistical techniques, that we will learn in this chapter, will not apply. Further, the inferences drawn from the experiments may be incorrect.

6.3.2 Definition: Statistic

A *statistic* is a function of the observations made from a random sample. It is a random variable because it depends on the randomly selected sample.

6.3.3 Example: Random sample and statistic

Let us consider that we want to know the proportion p of people in a certain organization who are interested to avail health coupons to obtain access to a nearby gym facility. There are over $10,000$ employees in the organization and soliciting an answer from every individual may not be logistically practical. So the organization randomly picks 100 people irrespective of gender, age, departmental affiliations, etc. and records the preferences of each member of this sample. It then computes and finds that 71 of these 100 people intend to avail the health coupons. Therefore, $\hat{p} = 0.71$ may be regarded as a reasonable estimate of p. It may be noted that had we chosen a different set of 100 people, the value of $\hat{p}$ might have been slightly different. So the statistic $\hat{p}$ is itself random as it is a function of the random sample.

6.3.4 Example: Sample statistics

Consider a random sample $X_1,\ X_2,\ ...,\ X_n$. Some frequently used sample statistics are

1. **sample mean**, $\overline{X} := \frac{X_1+X_2+\cdots+X_n}{n}$,
2. **sample variance**, $S^2 := \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \overline{X})^2$, and
3. **sample standard deviation**, $S = \sqrt{S^2}$.

It turns out that dividing by $(n-1)$ makes the sample variance an unbiased estimator of the population variance σ^2.[4] In fact, out of n degrees of freedom that is available from the n

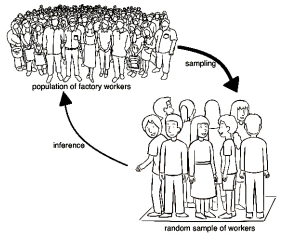

Figure 6.4: Random sampling of the whole population of factory workers enables us to make observations and estimates from a smaller representative group. We can then make inference about the whole population.

[4] S^2 is an unbiased estimator of σ^2, i.e. $E(S^2) = \sigma^2$.

observables $x_1,\ x_2,\ ...,\ x_n$; one unit of information is already used in the computation of $\overline{X}$. Hence we are left with $(n-1)$ degrees of freedom, thereby using $(n-1)$ as the normalization factor in the denominator makes sense.

6.3.5 *Sampling strategies*

Different sampling strategies may be employed depending on the experimental context. Broadly speaking, sampling strategies may be classified into two categories, viz., *non-probabilistic sampling* (eg. voluntary participation, convenience sampling based on availability of experimental participants, etc.), and *probabilistic sampling*. Some frequently used probabilistic sampling strategies are summarized below.[5]

1. **Random sampling**: Here, every member of the population has an equal probability of being included in the sample set whose size may be predefined.

2. **Stratified sampling**: Here, the entire population is classified into sub-populations based on certain characteristics. Sampling is performed randomly from within the sub-populations such that the groups are maintained in the same ratio in the sample set as they were in the original population. This type of sampling is most commonly used in pattern classification and machine learning applications.

3. **Clustered sampling**: In this strategy, the entire population is classified into sub-groups in a manner such that each of these sub-groups encompass all the features of the whole population. Thereafter, the clusters are randomly selected as a group, rather than individually, to form the sample set.

4. **Systematic sampling**: A pre-defined strategy is used to pick the members of the sample set from the whole population; eg. every 10^{th} (randomly decided) member of the population is sampled sequentially.

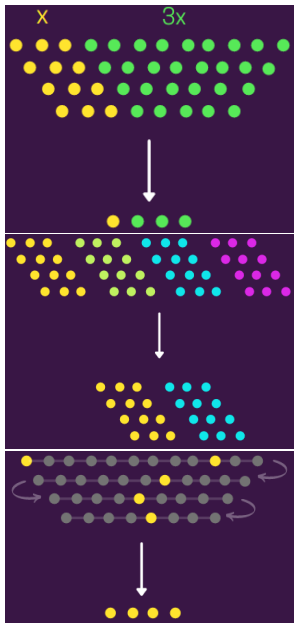

Figure 6.5: Probabilistic sampling strategies: stratified sampling (top), clustered sampling (middle), and systematic sampling (bottom).

[5] In many applications, re-sampling techniques like *jackknife* or *bootstrap* may be necessary.

6.3.6 *Definition: Sampling distribution*

The probability distribution of a statistic is called a *sampling distribution*.

The named probability distributions like Poisson, Bernoulli, exponential, etc., that were introduced in chapter three, are a consequence of theoretical considerations.[6] However, the sampling distributions are a consequence of real outcomes of a statistical experiment. We will study different types of sampling distributions and their applications in this chapter. We list here some of the most frequently used sampling distributions.

- Standard normal distribution.
- χ^2 distribution.
- t distribution.
- F distribution.

It is necessary to emphasize that the sampling distribution of a statistic depends on the following:

$\rightarrow$ distribution of the population,

[6] Eg., the probability density function of the exponential random variable can be derived theoretically from the Poisson distribution. We will consider the random variables: $X \sim \text{Poisson}(\lambda)$ and $T \sim exp(\lambda)$, where λ is the rate parameter. Since T is the inter-arrival time between two Poisson arrivals, $P(T > t)$ $\equiv P(X = 0 \text{ in } t \text{ time units})$ $= \underbrace{e^{-\lambda} \times e^{-\lambda} \times \cdots \times e^{-\lambda}}_{t \text{ times}}$ $= e^{-\lambda t}$. Therefore, $P(T \leq t)$ $= 1 - P(T > t) = 1 - e^{-\lambda t}$. Further, $f_T(t) = \frac{d}{dt} P(T \leq t)$ $= \lambda e^{-\lambda t}$. Thus, we see that the exponential probability distribution can be entirely constructed from the Poisson distribution based solely on theoretical arguments.

→ sample size, and

→ sampling strategy.

6.4 *Sampling distributions*

We will begin with the sampling distributions of the sample mean $\overline{X}$ and the sample variance S^2. If the random sample is drawn from the normal distribution, i.e. $X_1, X_2, ..., X_n \sim N(\mu, \sigma^2)$, then clearly by the linearity principle, we have $\overline{X} \sim N(\mu, \frac{\sigma^2}{n})$.[7] Notably, $\overline{X}$ and S^2 are independent random variables.[8] It turns out that this independence of $\overline{X}$ and S^2 is a unique property of the normal distribution. We will return to the distribution of the sample variance later in this section. Now, we will state our first result on the asymptotic distribution of $\overline{X}$ without assuming the distribution of the samples $X_1, X_2, ..., X_n$. This is one of the most fundamental result in the theory of statistics.

[7] This result can be easily deduced by using the method of moment generating functions that we studied earlier in chapter three.

[8] This result is a consequence of *Basu's theorem* because $\overline{X}$ is a complete sufficient statistic for estimating the model parameter μ and S^2 is an ancillary statistic whose distribution does not depend on μ. The independence of $\overline{X}$ and S^2 can also be proven by using *Cochran's theorem*. A detailed study of these techniques and concepts can be found in more advanced textbooks on statistics (**cf.** pg. 289 and pg. 572 in *Statistical Inference* by *George Casella* and *Roger L. Berger*, Duxbury-Thomson Learning (second edition), 2002.)

6.4.1 *Sampling distribution of $\overline{X}$:* ***Central Limit Theorem*** *(CLT)*

Consider a random sample of size n denoted by $X_1, X_2, ..., X_n$ which is drawn from a population of mean μ (i.e. $E(X_i) = \mu,\ i = 1, 2, ..., n$) and variance σ^2 (i.e. $Var(X_i) = \sigma^2$). Then the asymptotic distribution of $Z = \frac{\overline{X}-\mu}{\sigma/\sqrt{n}}$ as $n \to \infty$ is the standard normal distribution $N(0, 1)$.

Most importantly, $X_1, X_2, ..., X_n$ need not be normally distributed. Many random variables encountered in several science and engineering applications are normally distributed due to the effect manifested by the CLT.

6.4.2 *How large a sample suffices for the CLT result to take effect?*

This is a rather subjective question. If the underlying population distribution is much alike the standard normal distribution, then only a few (say 3 to 5) samples are required to get a decent normal approximation of the sample mean. However, if the population distribution is very different from the standard normal distribution, then about 30 samples will suffice to obtain a good normal approximation for the sample mean.

6.4.3 *Computer illustration of CLT*

In order to fully comprehend the meaning of the above result, we provide a computer simulated random sampling of exponentially distributed random variables with rate parameter $\mu = 3$. We will notice that with an increasing number of samples/realizations of the simulated experiment, the distribution of the sample means (mean subtracted and appropriately normalized)[9] resembles the standard normal distribution.

[9] The computer simulation shows that $\frac{\overline{X}-\mu}{\sigma/\sqrt{n}} \sim N(0, 1)$ as $n \to \infty$. Here $X_i \sim exp(\mu)$, for all $i = 1, 2, ..., n$.

The Matlab routine for conducting the aforementioned computer experiment is presented below.

```
mu = 3; var = mu^2; k=1; numsamples = [2 5 25 2000];
for N =  numsamples
    ni = [];
    for i=1:N
        ni = [ni; exprnd(mu,1,10000)];
```

```
    end
    n = (mean(ni)-mu)/(sqrt(var)/sqrt(N)); x = -max(n):0.1:max(n);
    subplot(length(numsamples),1,k);
    h=histogram(n,'Normalization','probability'); hold on;
    h.BinWidth = 1.0;
    [yn,xn]=ksdensity(n);
    plot(xn,yn,'k','LineWidth',2); hold on;
    pd = makedist('Normal'); pdf_normal = pdf(pd,x);
    plot(x,pdf_normal,'rx-','LineWidth',2);
    xlabel(,'FontSize',16); ylabel('probability distribution','FontSize',16);
    legend('histogram of samples from exponential distribution',...
        'pdf of distribution with N samples','standard normal distribution','FontSize',16);
    k=k+1;
end
```

It is necessary to comment on some of the Matlab commands used in the above routine.

→ `numsamples = [2 5 25 2000]`: Here, `numsamples` is a variable array which is assigned the values of the size of the experiments (number of realizations) that will be considered.

→ `exprnd(mu,1,10000)`: It returns an array of random numbers of dimensions 1×10000 chosen from the exponential distribution with mean parameter `mu`.

→ `ni = [ni; exprnd(mu,1,10000)]`: `ni` is a two-dimensional array of dimensions $N \times 10000$ that stores the N realizations of the exponentially distributed random variables. Each realization (sample) is an array of 10000 observables.

→ `histogram(n,'Normalization','probability')`: This function returns a histogram of the elements in the array `n`, the resulting histogram is normalized in such a way that the cumulative area covered by the histogram is normalized to unity.

→ `h.BinWidth = 1.0`: In conjunction with the normalization mode mentioned in the above paragraph, the width of each bin in the histogram must be set to unity to ensure compliance with the second axiom of probability (axiom of unitarity).

→ `ksdensity(n)`: It computes a probability density estimate of the sample stored in the array `n`.

→ `pdf(makedist('Normal'),x)`: This function generates the probability density function of a standard normal distribution for the observable variables in the range specified by `x`.

It is essential to emphasize that irrespective of the type of the probability distribution of the population, the sampling distribution of $Z := \frac{\overline{X}-\mu}{\sigma/\sqrt{n}}$ tends to the standard normal distribution. This can be easily verified by replacing the Matlab function `exprnd` with any other probability distribution, eg., `Poisson`, `Geometric`, etc. For more details, readers may check out the Matlab documentation page for `random`[10] for generating random variables (random samples) drawn from different probability distributions. The results of the simulation experiment are shown below in Figure 6.6.

[10] Type `>>doc random` in the Matlab command window and hit return.

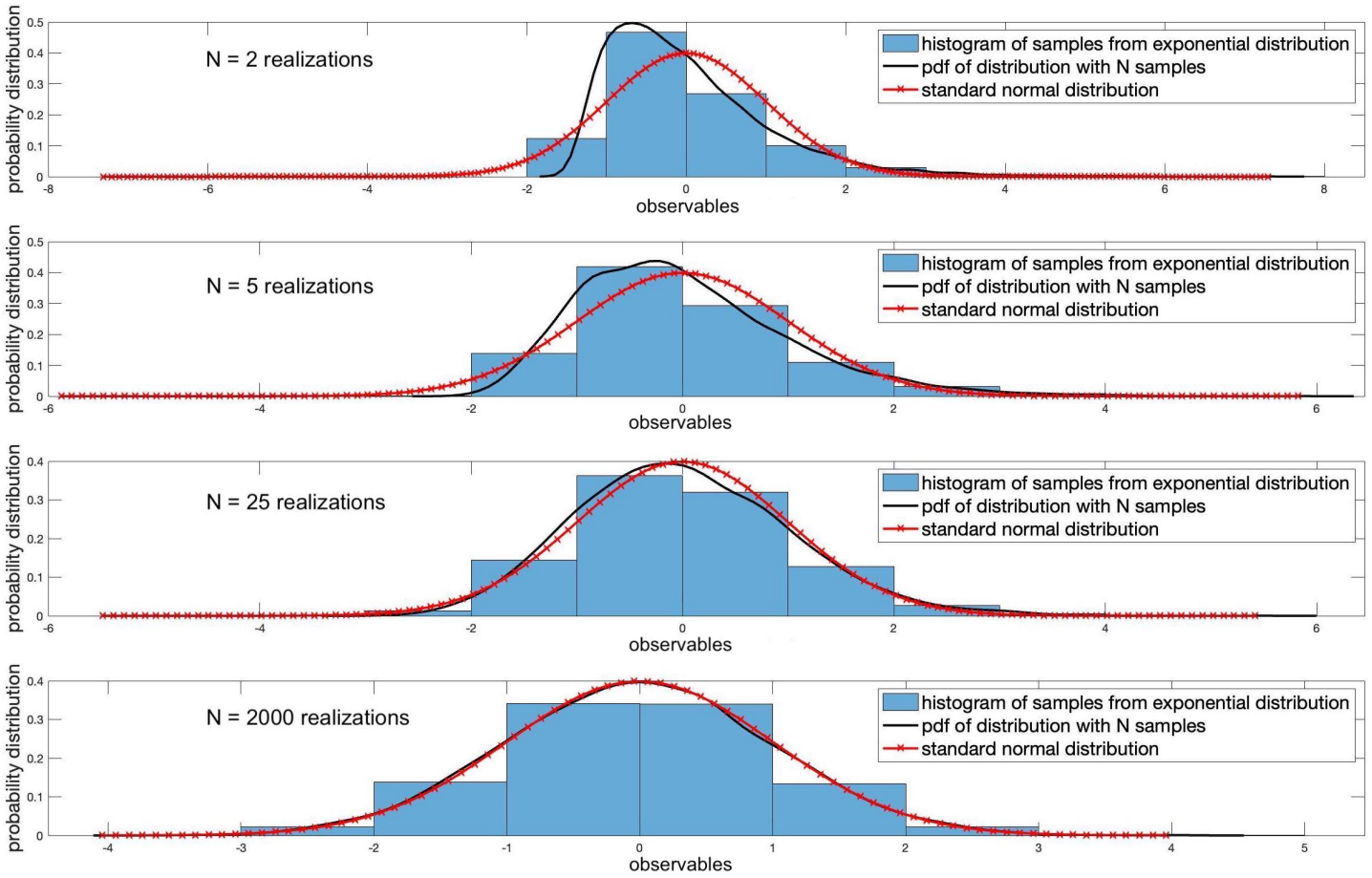

Figure 6.6: The Central Limit Theorem (CLT) demonstrates that as the size of the samples is increased from $N = 2$ to $N = 2000$, the probability distribution of the mean subtracted and appropriately normalized sample mean (shown by black solid lines) resembles the standard normal distribution (shown by cross-cut red lines).

6.4.4 *Example: Rating of frontal cameras of an autonomous vehicle (application of CLT)*

Safety features of a self-driving car are of paramount importance to manufacturers. This is the reason why modern autonomous cars have anywhere between 10 to 30 cameras on board. A certain self-driving car has three frontal cameras which are activated when the car gets close to a vehicle or object in front of the car. This is especially useful in a traffic jam and/or on the highway when a certain distance must be maintained between two successive cars. During foggy winter conditions, measurement of the exact distance of the car (d meters (m)) from the one in front, by a single input from a camera, may become less reliable. Therefore, the system must have built-in redundancies whereby multiple frames per second (fps) must be obtained successively by the three frontal cameras. Each of these measurements (all made every second) may be regarded to be an independent random variable with mean d m and standard deviation 2 m based on multiple statistical tests performed by the manufacturer. Then the average of all these measurements must be taken and processed

per second by the electronic computer of the car to make the autonomous technology more accurate. What must be the fps rating of the frontal camera unit of the car so that the manufacturer is at least 95% certain that the estimated information is accurate to within ± 0.5 m?

Solution: Consider that the requisite fps rating of the frontal camera system is n. Then, if $\overline{X}$ is the mean measurement made every second, then it is reasonable to assume that $\overline{X} \sim N(d, \frac{4}{n})$ based on the CLT (**cf.** note below). Therefore,

$$\begin{aligned} P(-0.5 < \overline{X} - d < 0.5) &= P\left(\frac{-0.5}{2/\sqrt{n}} < \frac{\overline{X} - d}{2/\sqrt{n}} < \frac{0.5}{2/\sqrt{n}}\right) \\ &\approx P\left(-\sqrt{n}/4 < Z < \sqrt{n}/4\right) \\ &\quad \nearrow \text{direct consequence of CLT} \\ &= 2P(Z < \sqrt{n}/4) - 1, \qquad (6.1) \\ &\quad \nearrow \text{see caption of Figure 6.7 below.} \end{aligned}$$

where Z is the standard normal random variable. This means we must have $2P(Z < \sqrt{n}/4) - 1 \geq 0.95$ or equivalently $P(Z < \sqrt{n}/4) \geq 0.975$. Following the standard normal distribution table shown in Figure 6.8, since we have $P(Z < 1.96) = 0.975$; n must be chosen such that $\sqrt{n}/4 \geq 1.96$ or $n \geq 61.46$. This means that the fps rating of the frontal camera system of the autonomous car must be at least 62.

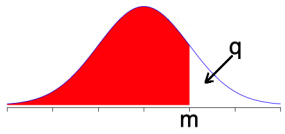

Figure 6.7: The shaded portion of the distribution $N(0,1)$ shows $P(Z < m)$, $m > 0$. $P(-m < Z < m) = (1 - 2q) = 1 - 2\{1 - P(Z < m)\} = 2P(Z < m) - 1$.

Note: In addition to the statement of the CLT, we can verify that

$$E(\overline{X}) = \frac{E(X_1) + E(X_2) + \cdots + E(X_n)}{n} = \mu, \text{ and } Var(\overline{X}) = \frac{\sum_{i=1}^{n} Var(X_i)}{n^2} = \frac{n\sigma^2}{n^2} = \sigma^2/n$$

.

	0.00	0.01	0.02	0.03	0.04	0.05	0.06	0.07	0.08	0.09
0.0	0.5000	0.5040	0.5080	0.5120	0.5160	0.5199	0.5239	0.5279	0.5319	0.5359
0.1	0.5398	0.5438	0.5478	0.5517	0.5557	0.5596	0.5636	0.5675	0.5714	0.5753
0.2	0.5793	0.5832	0.5871	0.5910	0.5948	0.5987	0.6026	0.6064	0.6103	0.6141
0.3	0.6179	0.6217	0.6255	0.6293	0.6331	0.6368	0.6406	0.6443	0.6480	0.6517
⋮										
1.6	0.9452	0.9463	0.9474	0.9484	0.9495	0.9505	0.9515	0.9525	0.9535	0.9545
1.7	0.9554	0.9564	0.9573	0.9582	0.9591	0.9599	0.9608	0.9616	0.9625	0.9633
1.8	0.9641	0.9649	0.9656	0.9664	0.9671	0.9678	0.9686	0.9693	0.9699	0.9706
1.9	0.9713	0.9719	0.9726	0.9732	0.9738	0.9744	0.9750	0.9756	0.9761	0.9767
2.0	0.9772	0.9778	0.9783	0.9788	0.9793	0.9798	0.9803	0.9808	0.9812	0.9817
2.1	0.9821	0.9826	0.9830	0.9834	0.9838	0.9842	0.9846	0.9850	0.9854	0.9857
2.2	0.9861	0.9864	0.9868	0.9871	0.9875	0.9878	0.9881	0.9884	0.9887	0.9890

Figure 6.8: An excerpt from the standard normal distribution table compatible with the shaded portion of the probability distribution shown in Figure 6.7. In the example discussed in section 6.4.4, $m = 1.96$. $P(Z < 1.96)$ can be computed from the table by reading off the entry corresponding to $(1.9, 0.06) = 0.9750$.

6.4.5 *Sampling distributions derived from the standard normal distribution*

In this section, we will study a few sampling distributions which either arise from the standard normal distribution or converge to the standard normal distribution asymptotically. These sampling distributions will be extensively used for performing statistical experiments and testing hypotheses.

(i) **The chi-square distribution:** The chi-square distribution, denoted by $\chi^2(\nu)$, has ν degrees of freedom. It is the distribution of sum of squares of normally distributed random variables.

$$X_\nu = Z_1^2 + Z_2^2 + \cdots + Z_\nu^2; \text{ where } Z_i \sim N(0,1);\ i = 1, 2, \cdots, \nu. \tag{6.2}$$

Then $X_\nu \sim \chi^2(\nu)$. The Matlab routine for simulating the χ^2 distribution based on the above definition is given below along with the results in Figure 6.9.

```
x = [0:0.1:20]; num_max = 5000;
nu = 5; mu = 0; sigma = 1;

for k=1:num_max
    X = 0;
    for i=1:nu
        Z_gauss = normrnd(mu,sigma);
        X = X + Z_gauss^2;
    end
    X_chi(k) = X;
end

h = histogram(X_chi,'Normalization','probability'); hold on;
h.BinWidth = 1.0;
[yn,xn]=ksdensity(X_chi);
plot(xn,yn,'k','LineWidth',2); hold on;

fx_chi = chi2pdf(x,nu);
plot(x,fx_chi,'m--','LineWidth',4);
legend(['histogram of samples from X_{\nu} = ' ...
    'Z_1^2 + Z_2^2 + \cdot \cdot + Z_{\nu}^2; where Z_i \sim N(0,1), i=1,2,...,\nu'],...
    'pdf of X_{\nu}','pdf of \chi^2(\nu)','FontSize',16);
xlabel('x','FontSize',16);
ylabel('f_{X_\nu}(x)','FontSize',16);
title('Number of degrees of freedom, \nu = 5','FontSize',16);
xlim([0 25]);
```

In Figure 6.9, the probability density function (pdf) of X_ν based on five thousand realizations of the ν-tuple random sample $\{Z_1, Z_2, \cdots, Z_\nu\}$ (solid lines) is compared against the χ^2 pdf generated by the Matlab function `chi2pdf` (broken lines). It may be easily verified that the accuracy of the pdf of X_ν generated from multiple realizations of $\{Z_i\}_{i=1,2,\ldots,\nu}$ increases by considering a larger number of realizations (i.e. by increasing the value of `num_max` in the above Matlab routine).

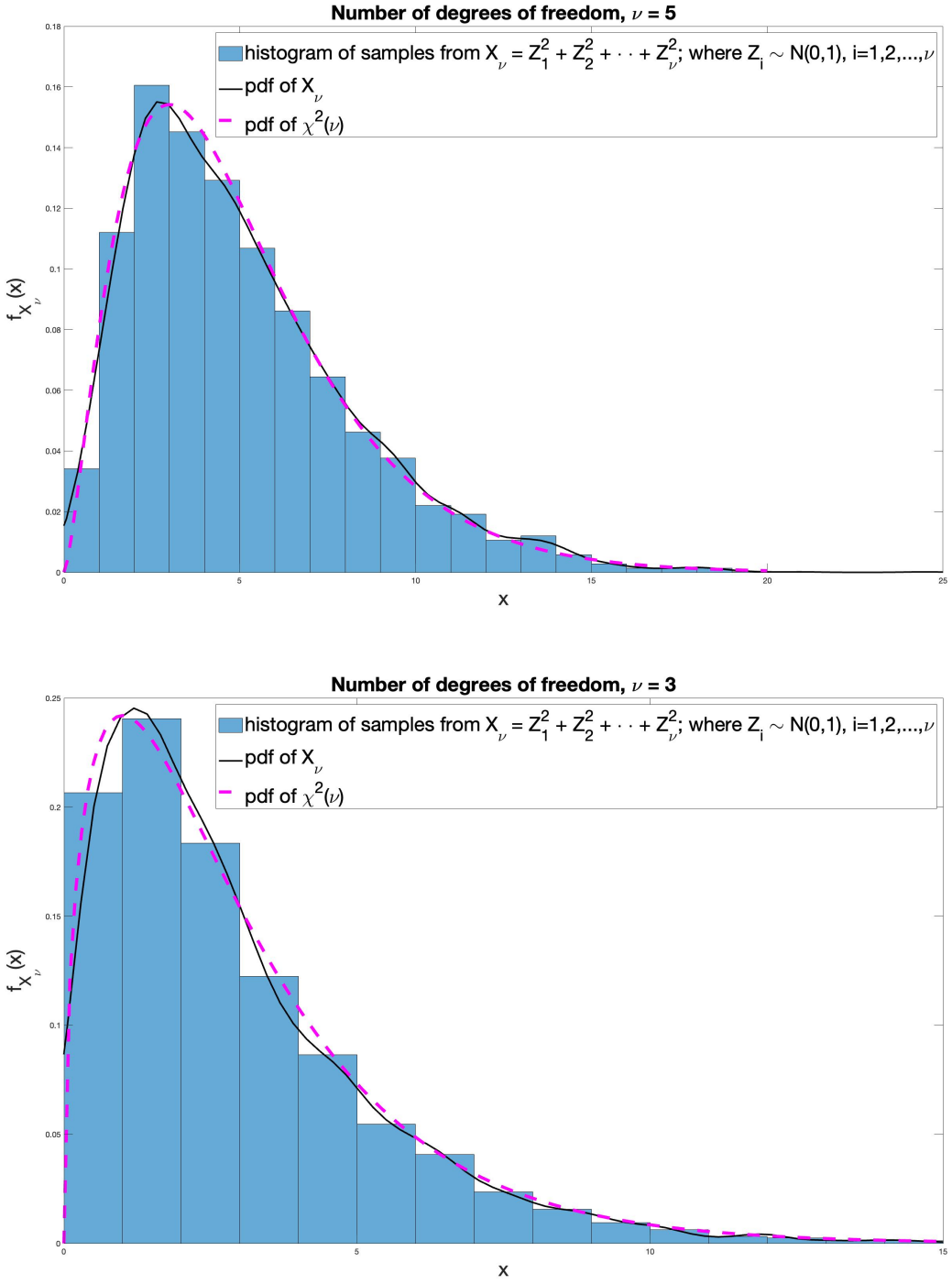

Figure 6.9: Computer simulation of the χ^2 distribution with ν degrees of freedom.

(i.i) **Properties of chi-square random variables**

1. χ^2 values are always positive (for $\nu > 1$) or non-negative (for $\nu = 1$).
2. The shape of the pdf of χ^2 random variables differs with ν.
3. $E(X_\nu) = \nu$ and $Var(X_\nu) = 2\nu$.

4. $\chi^2(\nu) \to N(\nu, 2\nu)$ as $\nu \to \infty$. The approximation is very good for $\nu \geq 30$. This entails that for large values of ν, $Z := \frac{X_\nu - \nu}{\sqrt{2\nu}} \sim N(0,1)$ where $X_\nu \equiv \chi^2_\nu \sim \chi^2(\nu)$. An illustration of this property is demonstrated below in Figure 6.10 with $\nu = 100$ degrees of freedom. The pdf is generated based on one hundred thousand realizations.

5. $\chi^2(n) \equiv \Gamma(\alpha, \lambda)$ when the shape parameter $\alpha = n/2$ and the rate parameter $\lambda = 1/2$.[11] Further, recall that $\Gamma(1, \lambda) \equiv exp(\lambda)$. So, when $n = 2$, the χ^2 distribution collapses to the exponential distribution with rate $\lambda = 1/2$, i.e. $\chi^2(2) \equiv exp(1/2)$. Here, $\Gamma(\alpha, \gamma)$ refers to the gamma probability distribution, whereas $\Gamma(\alpha)$ refers to the gamma function.

[11] The pdf of $\chi^2(n)$ distribution is $f_X(x) = \frac{(\frac{1}{2})^{\frac{n}{2}} x^{\frac{n}{2}-1} e^{-\frac{x}{2}}}{\Gamma(\frac{n}{2})}$, $x > 0$.

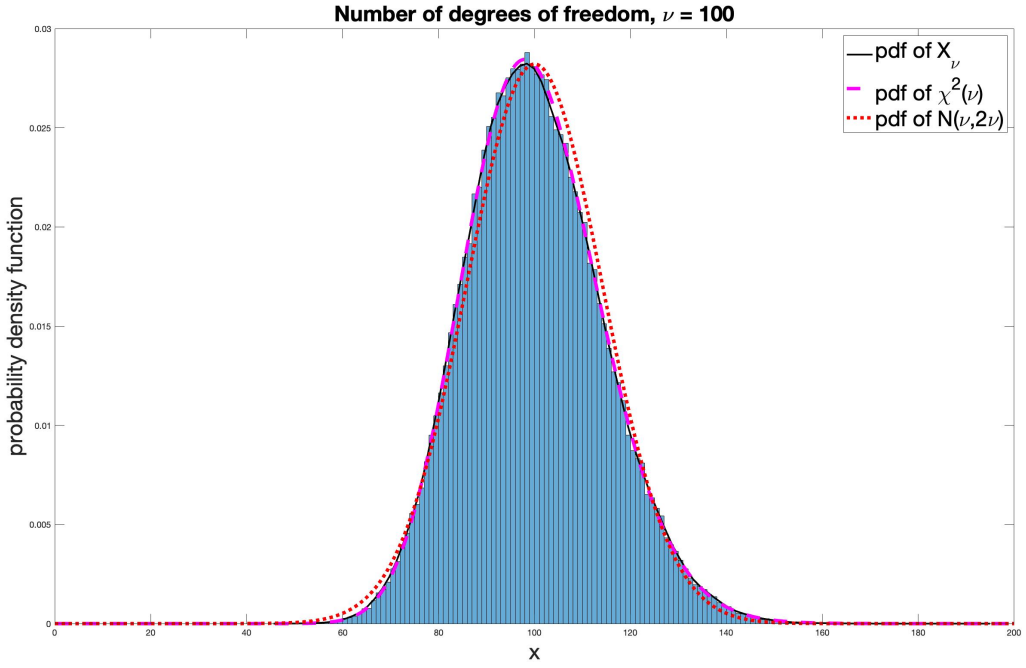

Figure 6.10: For large values of ν (say 100), the $\chi^2(\nu)$ distribution converges to a normal distribution with mean ν and variance 2ν.

(i.ii) **Application of χ^2 distribution: sampling distribution of $\frac{(n-1)S^2}{\sigma^2}$**

Let $X_i \sim N(\mu, \sigma^2)$; $i = 1, 2, ..., n$; then $\frac{(n-1)S^2}{\sigma^2} \sim \chi^2(n-1)$ where $S^2 := \frac{\sum_{i=1}^n (X_i - \overline{X})^2}{(n-1)}$ is the *unbiased* sample variance. This result is elucidated here by simulating a large number of normally distributed random variables and subsequently, computing the sample variance as shown in Figure 6.11. The Matlab code for this simple simulation is furbished below.

```
mu = 10; var = 2;
s = sqrt(var);
n=5; max_realizations = 100000;
for realizations = 1:max_realizations
    y = s.*randn(n,1) + mu;
    sample_var(realizations) = (sum((y - mean(y)).^2))/(length(y)-1);
end
weighted_sample_var = (n-1)*sample_var/s^2;
h=histogram(weighted_sample_var,'Normalization','probability'); hold on;
h.BinWidth = 1.0;
```

```
[yn,xn]=ksdensity(weighted_sample_var,'Support','positive');
plot(xn,yn,'k','LineWidth',2); hold on;
x=[0:30];
fx_chi = chi2pdf(x,n-1);
plot(x,fx_chi,'m--','LineWidth',4);
legend('histogram of $\frac{(n-1)S^2}{\sigma^2}$ based on 100000 realizations',...
    'pdf of $\frac{(n-1)S^2}{\sigma^2}$','pdf of $\chi^2(n-1)$',...
    'Interpreter','latex','FontSize',25);
xlim([min(xn) max(x)]);
xlabel('x','FontSize',25);
ylabel('probability density function','FontSize',25);
title('Demonstrating $\frac{(n-1)S^2}{\sigma^2} \sim \chi^2(n-1)$ with n = 5',...
    'FontSize',25,'Interpreter','latex');
```

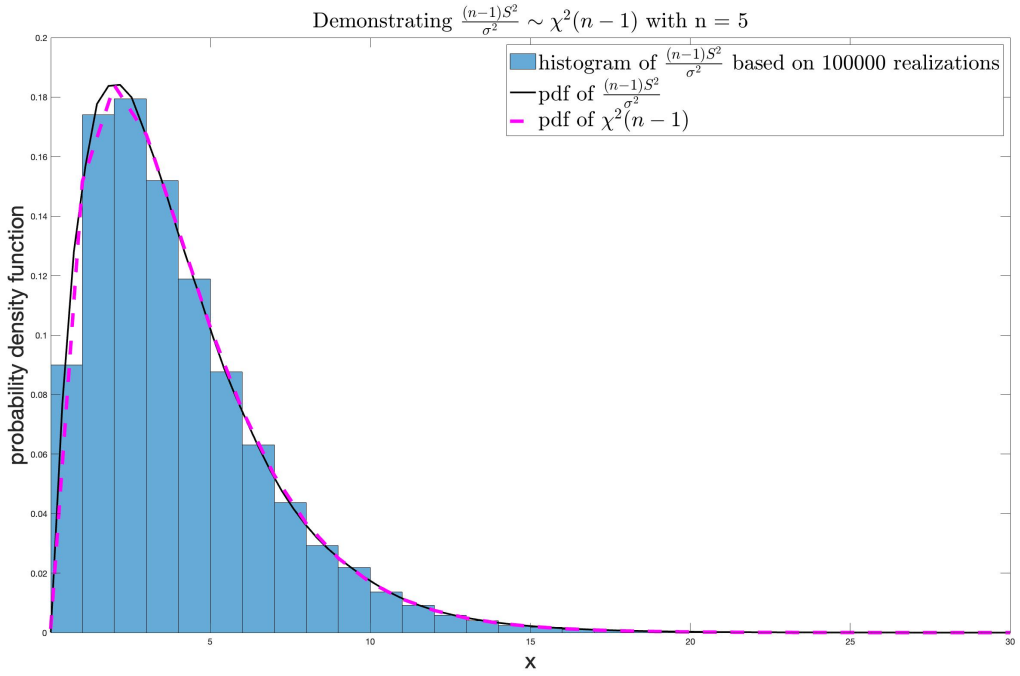

Figure 6.11: Sampling distribution of the sample variance.

The above result can be proved more rigorously by using the method of moment generating functions. Interested readers may refer to the bibliography provided at the end of this chapter. We will omit the presentation of the proof here.

(i.iii) **Example: success rate of a sniper**

A sniper can locate a target at a distance of 2 kms. In desert conditions and/or foggy weather, the optical scope on the marksman's rifle has an imprecision in each of the horizontal and vertical coordinates that is normally distributed with mean zero and variance of four sq. meters. What is his success rate to hit a target within a radius of 0.1 m?

Solution: Let $R_{err}^2 = X^2 + Y^2$ denote the square of the error to hit the target, $X, Y \sim N(0,4)$. Consider $Z_1 = X/2$ and $Z_2 = Y/2$ whence we have $Z_i \sim N(0,1),\ i = 1,2$. Now, following equation 6.2, we have,

$$P(R_{err}^2 < 0.01) = P(Z_1^2 + Z_2^2 < 0.01/4 = 0.0025) = P(\chi_2^2 < 0.0025) = 1 - e^{-\frac{0.0025}{2}} = 0.0012.$$

The penultimate equality above arises from the fact that $\chi_2^2 \sim exp(\lambda = 1/2)$ as stated above in subsection (i.i.5) of section 6.4.5. Alternatively, the equality may also be arrived at by using the Matlab command: >> `chi2cdf(0.0025,2)`. Therefore, the success rate of the sniper under the given adverse atmospheric condition drops to 0.12%.

(ii) **The t distribution:**[12] Let us recall from the Central Limit Theorem that if we have random samples Y_i of size n from any population distribution with mean μ and variance σ^2, then $\overline{Y} \sim N(\mu, \frac{\sigma^2}{n})$ as $n \to \infty$. This implies $Z := \frac{\overline{Y}-\mu}{\sigma/\sqrt{n}} \sim N(0,1)$.

However, in most practical cases, the population variance σ^2 is unknown. This limits the utility of Z as a suitable statistic for conducting experiments. This entails that we consider an appropriate test statistic where the population variance σ^2 is replaced by the sample variance S^2 which is a knowable (computable) quantity. From our discussion on χ^2 distribution, since we know that $\frac{(n-1)S^2}{\sigma^2} \sim \chi^2(n-1)$, let us consider a test statistic defined as follows:

$$T \equiv T_{n-1} := \frac{Z}{\sqrt{\frac{\chi^2_{(n-1)}}{n-1}}}. \tag{6.3}$$

Here Z and $\chi^2_{(n-1)}$ are independent random variables. The above test statistic T simplifies to $T = \frac{\overline{Y}-\mu}{S/\sqrt{n}} \sim t(n-1)$.[13] This defines the t distribution with $(n-1)$ degrees of freedom.[14]

(ii.i) **Properties of t distribution**

1. $t(n) \xrightarrow{n\to\infty} N(0,1)$.
2. $E(T) = 0$, for $n > 1$ (otherwise, undefined), and $Var(T) = \frac{n}{n-2}$, for $n > 2$ (∞ for $1 < n \leq 2$).
3. For some $\alpha \in [0,1]$, let us consider the observable quantity $t_{\alpha,n}$ such that $P(T_n > t_{\alpha,n}) = \alpha$. Now, given that the bell-shaped t distribution is a symmetric curve about the mean zero (**cf.** Figures 6.12 and 6.13), we can easily establish the following symmetry relation.

$$-t_{\alpha,n} = t_{1-\alpha,n}. \tag{6.4}$$

The first property stated above is illustrated through a Matlab routine given below.

```
x = [-6:0.1:6];
for nu=1:1000
    plot(x,tpdf(x,nu)); hold on;
end
pd = makedist('Normal'); pdf_normal = pdf(pd,x);
```

[12] It is also known as the *Student's* distribution.

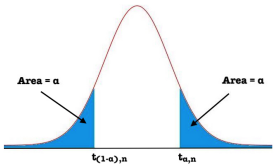

Figure 6.12: The pdf of the t distribution is symmetric about the origin, i.e. $-t_{\alpha,n} = t_{1-\alpha,n} = t_{\alpha,n}$.

[13] For clarity, here $\chi^2(\nu)$ refers to the chi-square distribution with ν degrees of freedom and $\chi^2_\nu \equiv X_\nu$ refers to the random variable that is sampled from the $\chi^2(\nu)$ distribution. The random variable T follows the t distribution with $(n-1)$ degrees of freedom.

[14] We will not present here the actual expression of the pdf of the t-distribution because it is complicated. Interested readers may refer to the texts mentioned in the bibliography of this chapter.

```
plot(x,pdf_normal,'rd:','LineWidth',4);
xlabel('observables','FontSize',24); ylabel('probability density function','FontSize',24);
set(gca,'FontSize',24)
```

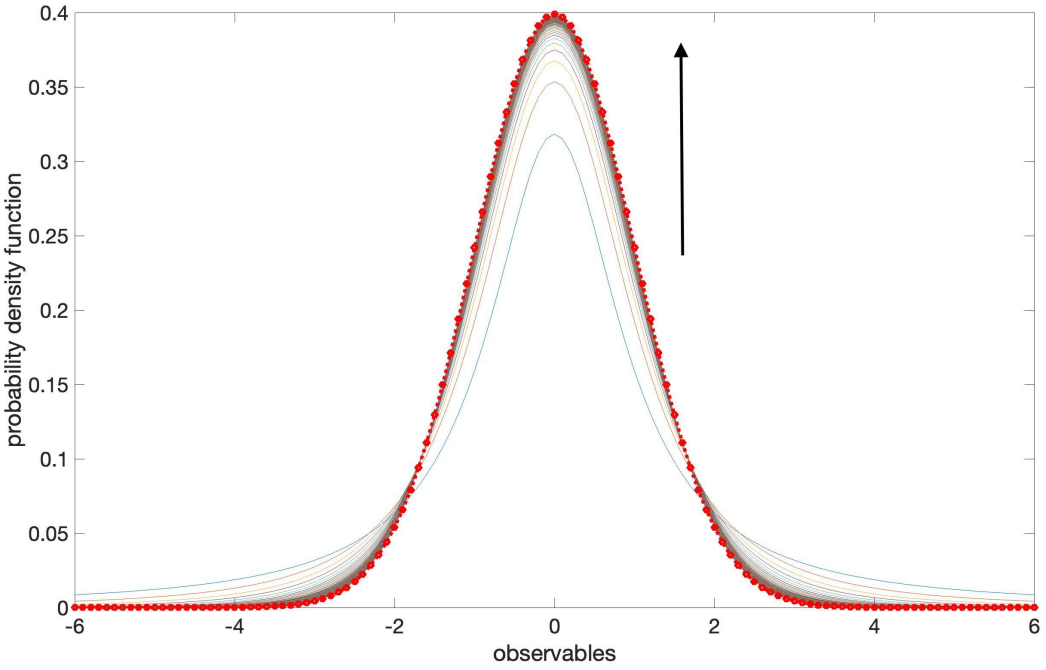

Figure 6.13: The thin solid curves represent the pdf of the t distribution with varying degrees of freedom. The solid vertical arrow points in the direction of increasing degrees of freedom. The checkered-dotted curve represents the pdf of the standard normal distribution. $t(n) \xrightarrow{n\to\infty} N(0,1)$.

Besides the asymptotic convergence of the t-distribution to the standard normal distribution, the above result also demonstrates that the t-distribution is mostly suitable for data with bell-shaped distribution.

(ii.ii) **Applications of t distribution**

1. Let us consider two different samples of sizes n_1 and n_2 with means $\overline{x}_1$ and $\overline{x}_2$, respectively. We may want to know if the two means are sufficiently alike to warrant an inference that both the samples are drawn from the same population. In such a case, $T := \frac{(\overline{x}_1-\overline{x}_2)\sqrt{n_1+n_2-2}}{\sqrt{S_1+S_2}}\sqrt{\frac{n_1 n_2}{n_1+n_2}}$ is a suitable test statistic that follows the t distribution. Here S_1 and S_2 are the respective sample variances.

2. The t distribution also finds applications in statistical experiments to find the significance of differences between regression coefficients obtained from different samples.

We will study the first application in greater detail in this chapter. We will re-visit regression analysis in a subsequent chapter of this book. Interested readers may refer to the seminal work by R. A. Fisher to know more about the applications of the t distribution.[15]

(ii.iii) **Example: Estimating the spread of viral infection**

1. Consider that the daily number of reported influenza cases in a village is denoted

[15] *Applications of Student's Distribution by* R. A. Fisher, Metron, **5**, pp. 90-104, (1925).

by X. $X \sim N(70, 9)$. What is the probability that on a given day the total number of reported cases exceeds seventy five?

2. Now, consider that the actual daily mean number of influenza cases in the village is seventy, i.e. $\mu_Y = 70$. It is not known if Y follows a normal distribution. Specifically, over a period of fifteen days, the sample mean of number of infections $\overline{Y}$ is computed. Additionally, it is observed that the sample standard deviation is four reported cases. What is the probability that $\overline{Y}$ is greater than seventy four?

Solution:

1. X is sampled from a population which is normally distributed with mean 70 and variance 9. So $Z := \frac{X-\mu}{\sigma/\sqrt{n}} = \frac{75-70}{3} \approx 1.67$, $Z \sim N(0,1)$. In this case, $n = 1$. Therefore, $P(Z > 1.67) = 1 - P(Z \leq 1.67) = 0.0475$. This result may be computed by using the standard normal distribution look-up table or by using the following Matlab command.

   `>> 1-cdf('Normal',1.67,0,1)`

2. In this case, since the population standard deviation is unknown, an appropriate test statistic is $T := \frac{\overline{Y}-\mu_Y}{S/\sqrt{n}} \sim t(n-1)$ which depends on the sample standard deviation S.

$$P(\overline{Y} > 74) = P\left(\frac{\overline{Y}-\mu}{S/\sqrt{n}} > \frac{74-70}{4/\sqrt{15}}\right) = P(T > 3.8730) = 0.00084461.$$

   The above may be computed by using the Matlab command:

   `>>1-tcdf(3.8730,14)`

   It may be carefully noted that we have used $(n-1) = (15-1) = 14$ degrees of freedom to compute the probability because $T \sim t(n-1)$.

(iii) **The F distribution:**[16] If $X_n = \chi^2_n$ and $X_m = \chi^2_m$ are two independent chi-square random variables with n and m degrees of freedom respectively. Then the ratio

$$F := \frac{X_n/n}{X_m/m} \tag{6.5}$$

is a random variable which follows the F distribution. Further, if we have two independent samples of sizes n_1 and n_2 from two independent normal populations with variances σ_1^2 and σ_2^2 respectively, then the statistic

$$F := \frac{S_1^2/\sigma_1^2}{S_2^2/\sigma_2^2} \tag{6.6}$$

follows the $F(n_1 - 1, n_2 - 1)$ distribution.

[16] It is also known as the *Fisher–Snedecor distribution*.

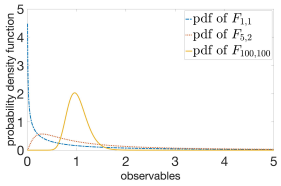

Figure 6.14: The pdfs of the F distribution with different pairs of n and m are generated using the Matlab command `>>plot(x,fpdf(x,n,m));` where x takes the values of the observables.

(iii.i) **Properties of F distribution**

1. The F distribution is defined for non-negative values.
2. The pdf of the F random variables is not symmetric in shape over the range of the observables (**cf.** Figure 6.14).

3. $F(1,m) \equiv t^2(m)$ (**cf.** Figure 6.15).
The Matlab routine to demonstrate this result computationally is given below.

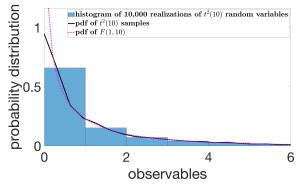

Figure 6.15: Computer simulation demonstrating $F(1,m) \equiv t^2(m)$.

```
x = [0:0.1:6];
n=1; m=10;
nu = m;
max_realizations = 10000;
for i=1:max_realizations
    t_rv = trnd(nu);
    tsq_rv(i) = (t_rv)^2;
end
h = histogram(tsq_rv,'Normalization','probability'); hold on;
h.BinWidth = 1.0;
[yn,xn]=ksdensity(tsq_rv);
plot(xn,yn,'k','LineWidth',4); hold on;
plot(x,fpdf(x,n,m),'m:','LineWidth',4);
xlim([0 max(x)]);
xlabel('observables');
ylabel('probability distribution');
set(gca,'FontSize',60);
legend('\bf histogram of 10,000 realizations of $t^2(10)$ random variables',...
    '\bf pdf of $t^2(10)$ samples',...
    '\bf pdf of $F(1,10)$','Interpreter','latex','FontSize',33);
```

4. $F(n,m) \overset{m\to\infty}{\equiv} \frac{\chi^2(n)}{n}$ (**cf.** Figure 6.16).
The Matlab routine to demonstrate this result computationally is given below.

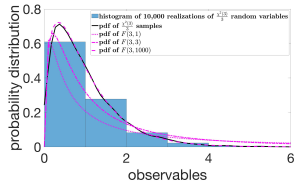

Figure 6.16: Computer simulation demonstrating $F(n,m) \overset{m\to\infty}{\equiv} \frac{\chi^2(n)}{n}$.

```
x = [0:0.1:6];
n=3; m=1000; m1=1; m2=3;
nu = n;
max_realizations = 10000;
for i=1:max_realizations
    X_chisq(i) = chi2rnd(nu)/nu;
end
h = histogram(X_chisq,'Normalization','probability'); hold on;
h.BinWidth = 1.0;
[yn,xn]=ksdensity(X_chisq);
plot(xn,yn,'k','LineWidth',4); hold on;
plot(x,fpdf(x,n,m1),'m:',x,fpdf(x,n,m2),'m-.','LineWidth',4); hold on;
plot(x,fpdf(x,n,m),'m--','LineWidth',4);
xlim([0 max(x)]);
xlabel('observables');
ylabel('probability distribution');
set(gca,'FontSize',60);
legend('\bf histogram of 10,000 realizations of $\frac{\chi^2(3)}{3}$ random variables',...
```

```
'\bf pdf of $\frac{\chi^2(3)}{3}$ samples','\bf pdf of $F(3,1)$',...
'\bf pdf of $F(3,3)$','\bf pdf of $F(3,1000)$','Interpreter','latex','FontSize',30);
```

5. $E(F) = \frac{m}{m-2}$ and $Var(F) = \frac{2m^2(n+m-2)}{n(m-2)^2(m-4)}$ where n is the degrees of freedom corresponding to the numerator, and m is the degrees of freedom corresponding to the denominator.

6. Consider $P(F > f_{\alpha,n,m}) = \int_{f_{\alpha,n,m}}^{\infty} f_F(x)dx = \alpha$ and $P(F < f_{1-\alpha,n,m}) = \int_0^{f_{1-\alpha,n,m}} f_F(x)dx = \alpha$. Here $f_{\alpha,n,m}$ is the *upper-tailed* α*-percentage point* and $f_{1-\alpha,n,m}$ is the *lower-tailed* $(1-\alpha)$*-percentage point*. The lower-tailed percentage point can be found in terms of the upper-tailed percentage point as follows: $f_{1-\alpha,n,m} = \frac{1}{f_{\alpha,m,n}}$. Here α is known as the *level of significance*.[17] $f_{\alpha,n,m}$ is the *critical value*. We will re-visit α in a latter section of this chapter when we discuss methods to test hypotheses.

[17] α corresponds to the *rejection region*. The area under the curve $f_F(x)$ to the right of $f_{\alpha,n,m}$ is equal to α; the area under the curve $f_F(x)$ to the left of the curve $f_{\alpha,n,m}$ is equal to $1-\alpha$ and corresponds to the *acceptance region*.

(iii.ii) **Applications of F distribution**

1. The test statistic used for performing an *analysis of variance* (ANOVA) experiment to test the difference between means of different populations is an F random variable that follows the F distribution.

2. F distribution is also used to test the existence of any significant difference between the variances of two different groups of population. Eg., (i) a university academic policy may prefer that two instructors, co-teaching a course, grade exams in such a way so as to have the same variation in their grading; (ii) to ensure a tight fit, a manufacturing unit may require the variation in the lid and the container to be similar (**cf.** definition of F in equation (6.6) and Figure 6.17).

Figure 6.17: The variation in the pattern of grooves in the inner lining of the lid and the outer lining of the container must be similar. A manufacturing unit making covered containers must perform a statistical experiment to ensure quality check of its products. Such an experiment would rely on the application of the F distribution.

(iii.iii) **Example: Inverse symmetry of the upper and lower tailed percentage points of $F(n,m)$**

Given that the upper 5-percentage point of $F(5,10)$ can be found by using the Matlab command >> `finv(1-0.05,5,10)`.[18] Compute the lower 95-percentage point of $F(5,10)$ both analytically as well as using an appropriate Matlab command.

Solution: Since the Matlab command >> `finv(1-0.05,5,10)` gives `ans = 3.3258`, and the command >> `finv(1-0.05,10,5)` gives `ans = 4.7351`. We have $f_{0.05,5,10} = 3.3258$ and $f_{0.05,10,5} = 4.7351$ whence we must have $f_{0.95,5,10} = \frac{1}{f_{0.05,10,5}} = \frac{1}{4.7351} = 0.2112$. Indeed using the Matlab command >>`finv(1-0.95,5,10)`, we get `ans = 0.2112` which means that we would observe the value greater than 0.2112 about 95% of the time by chance (when the numerator degrees of freedom is 5 and the denominator degrees of freedom is 10).

[18] The convention used in the Matlab command >> `finv(1-`α`,n,m)` may seem antithetical to the corresponding notation $f_{\alpha,n,m}$; so we warn the readers here to be mindful of this.

6.5 Chapter project: Prioritizing post-disaster reconstruction measures

6.5.1 Interlude: Ratings of the field managers on bottlenecks in project implementation

The ratings of the field managers on the issues mentioned in the project prologue section are printed below.

1. **Community Participation:**

Rating (scale: 1-10, 1: strongly disagree, 10: strongly agree)						
Cities	**Mgr1**	**Mgr2**	**Mgr3**	**Mgr4**	**Mgr5**	**Mgr6**
Port-au-Prince (Haiti)	$y_{11}=3$	$y_{12}=2$	$y_{13}=9$	$y_{14}=8$	$y_{15}=9$	$y_{16}=9$
Tacloban City	$y_{21}=5$	$y_{22}=9$	$y_{23}=10$	$y_{24}=5$	$y_{25}=8$	$y_{26}=9$
Latur	$y_{31}=6$	$y_{32}=7$	$y_{33}=10$	$y_{34}=5$	$y_{35}=7$	$y_{36}=8$
New Orleans	$y_{41}=8$	$y_{42}=9$	$y_{43}=9$	$y_{44}=8$	$y_{45}=2$	$y_{46}=8$
Kathmandu	$y_{51}=3$	$y_{52}=8$	$y_{53}=7$	$y_{54}=10$	$y_{55}=10$	$y_{56}=4$
Bagh City	$y_{61}=2$	$y_{62}=7$	$y_{63}=9$	$y_{64}=10$	$y_{65}=6$	$y_{66}=7$

2. **Funding:**

Rating (scale: 1-10, 1: strongly disagree, 10: strongly agree)						
Cities	**Mgr1**	**Mgr2**	**Mgr3**	**Mgr4**	**Mgr5**	**Mgr6**
Port-au-Prince (Haiti)	$y_{11}=3$	$y_{12}=2$	$y_{13}=9$	$y_{14}=8$	$y_{15}=9$	$y_{16}=7$
Tacloban City	$y_{21}=5$	$y_{22}=4$	$y_{23}=4$	$y_{24}=5$	$y_{25}=3$	$y_{26}=2$
Latur	$y_{31}=5$	$y_{32}=2$	$y_{33}=4$	$y_{34}=5$	$y_{35}=1$	$y_{36}=2$
New Orleans	$y_{41}=3$	$y_{42}=1$	$y_{43}=1$	$y_{44}=2$	$y_{45}=6$	$y_{46}=2$
Kathmandu	$y_{51}=3$	$y_{52}=8$	$y_{53}=7$	$y_{54}=10$	$y_{55}=10$	$y_{56}=4$
Bagh City	$y_{61}=3$	$y_{62}=1$	$y_{63}=9$	$y_{64}=8$	$y_{65}=6$	$y_{66}=7$

3. **Land Ownership:**

Rating (scale: 1-10, 1: strongly disagree, 10: strongly agree)						
Cities	**Mgr1**	**Mgr2**	**Mgr3**	**Mgr4**	**Mgr5**	**Mgr6**
Port-au-Prince (Haiti)	$y_{11}=9$	$y_{12}=9$	$y_{13}=10$	$y_{14}=8$	$y_{15}=7$	$y_{16}=8$
Tacloban City	$y_{21}=5$	$y_{22}=4$	$y_{23}=4$	$y_{24}=5$	$y_{25}=3$	$y_{26}=2$
Latur	$y_{31}=4$	$y_{32}=6$	$y_{33}=7$	$y_{34}=2$	$y_{35}=8$	$y_{36}=9$
New Orleans	$y_{41}=3$	$y_{42}=1$	$y_{43}=5$	$y_{44}=2$	$y_{45}=6$	$y_{46}=2$
Kathmandu	$y_{51}=7$	$y_{52}=4$	$y_{53}=5$	$y_{54}=1$	$y_{55}=2$	$y_{56}=3$
Bagh City	$y_{61}=3$	$y_{62}=2$	$y_{63}=9$	$y_{64}=8$	$y_{65}=6$	$y_{66}=7$

4. **Shortage of technical staff:**

Rating (scale: 1-10, 1: strongly disagree, 10: strongly agree)						
Cities	**Mgr1**	**Mgr2**	**Mgr3**	**Mgr4**	**Mgr5**	**Mgr6**
Port-au-Prince (Haiti)	$y_{11}=6$	$y_{12}=9$	$y_{13}=5$	$y_{14}=5$	$y_{15}=7$	$y_{16}=6$
Tacloban City	$y_{21}=6$	$y_{22}=4$	$y_{23}=6$	$y_{24}=5$	$y_{25}=7$	$y_{26}=8$
Latur	$y_{31}=4$	$y_{32}=6$	$y_{33}=7$	$y_{34}=2$	$y_{35}=8$	$y_{36}=9$
New Orleans	$y_{41}=4$	$y_{42}=6$	$y_{43}=6$	$y_{44}=1$	$y_{45}=8$	$y_{46}=9$
Kathmandu	$y_{51}=10$	$y_{52}=7$	$y_{53}=8$	$y_{54}=1$	$y_{55}=5$	$y_{56}=6$
Bagh City	$y_{61}=3$	$y_{62}=2$	$y_{63}=9$	$y_{64}=8$	$y_{65}=6$	$y_{66}=7$

The data set above comprises of ratings of construction managers on issues plaguing reconstruction projects post disaster (PDR). In each group (city), there are six different observations from six construction engineers who have been involved in PDR projects over the last many years. For each of the above mentioned issues, you have to perform a one-way ANOVA calculation and test if the data provided in the tables presents a statistically significant difference in the mean rating of the construction engineers between the six different cities. This will reveal if the issues plaguing the implementation of PDR projects is affected in an identical manner across the six different cities around the world. Once the issues that are universally relevant have been identified, then the total mean rating score across all cities for a given issue should be computed and a decision on necessary corrective measure should be taken if this grand mean is greater than 5 (on a scale of 10).

The one-way ANOVA analysis is a specific statistical experiment where we test whether there is a significant difference in means (μ) of different populations. In the next few sections of this chapter, we will study the fundamental principles of performing the test of hypothesis that will aid us to complete the chapter project.

6.6 *Test of hypothesis and statistical inference*

The primary objective of performing a test of hypothesis is to use data from a sample (random) to make inferences about the population. Such statistical experiments rely on the use of test statistics and sampling distributions like the standard normal distribution, χ^2 distribution, t distribution, F distribution, etc.

Such a statistical experiment involves (i) a statement (*hypothesis*) about the parameter (characterizing the concerned population), and (ii) a measure of reliability of that statement in terms of probability.

6.6.1 *What is a hypothesis?*

It is a statement about a parameter(s) characterizing a population. A hypothesis usually results from speculation concerning an observed behavior, a natural phenomenon, or an established theory. If a hypothesis is stated in terms of population parameters such as the mean and variance, then it is called a *statistical hypothesis*. Data from the sample is used to test the validity of the hypothesis.

6.6.2 *Components of an experiment to test hypothesis*

The key components are itemized below.

- Construct the statement to be tested: *null* (H_0) vs *alternate* (H_1 or H_a) hypothesis.
- Identify the *rejection* (or *critical*) region to enable a decision about the hypothesis (eg. the evidence based on the test statistic may prompt us to either reject or fail to reject the null hypothesis).
- Quantify the likely error in the aforementioned decision in terms of a probability measure. There are generally two types of errors, viz., *type-1 error* (with a probability of occurrence denoted by α) and *type-2 error* (with a probability of occurrence denoted by β). The

objective is to reduce these errors while making a decision. However, in many circumstances, reducing one type of error can lead to an increase in the other type of error.

6.6.3 *Steps involved in performing a test of hypothesis*

The steps involved in performing a test of hypothesis are listed below.

1. Step 1: Specify H_0 and H_a and an acceptable level of significance α.
2. Step 2: Define a sample based test statistic (eg. $\overline{X}, S^2$, etc.) and a rejection (or critical) region for H_0 that is most suitable for the experiment.
3. Step 3: Collect the sample data and calculate the test statistic.
4. Step 4: Make a decision to either reject or fail to reject H_0.
5. Step 5: Interpret the result in the language of the problem at hand (eg., provide confidence intervals, etc.) and provide an estimate of the error in the decision.

6.6.4 *Errors in inference*

The type and measure of the error is captured succinctly in the following table.

Decision \Truth condition	H_0 is true	H_0 is not true
H_0 is not rejected	Decision is correct (with probability $1-\alpha$)	type-2 error (with probability β)
H_0 is rejected	type-1 error (with probability α)	Decision is correct (with probability $1-\beta$)

Figure 6.18: A sample of sixteen peanut jars from a packaging unit is subjected to statistical tests to check for discrepancies in weights.

6.6.5 *What might determine our choice of α?*

Deciding on the level of significance of a test is as much an art as it is guided by the context of the problem. This may be best illustrated through an example which we study in this section.

Example: Quality control in a packaging industry

A company that packages salted peanuts in 8 kg jars is interested in maintaining control on the amount of peanuts put in the jars by one of the machines in its packaging units. *Control* is defined as averaging 8 kg per jar and not consistently over or under filling the jars. To monitor this control, a sample of 16 jars is taken from the packaging line at random time intervals and their contents weighed. The mean weight of peanuts in these 16 jars will be used to test the null hypothesis that the machine is indeed working properly. If it is found not to be doing so, an expensive adjustment will be required. What may be a suitable level of significance for this test? For convenience, let us suppose the population standard deviation $\sigma = 0.2$ kg of the weight of the jars is known to us.[19]

Solution:

1. Step 1: The hypothesis is stated as follows.

$$H_0 : \quad \mu = 8 \tag{6.7}$$
$$H_a : \quad \mu \neq 8 \tag{6.8}$$

   In many problems, we may want to decide on α at this point. But what if we are not sure about how to choose α? This example will attempt to address this issue.

2. Steps 2 & 3: The most appropriate test statistic for the case in hand is the sample mean, $\overline{X} = \frac{\sum_{i=1}^{16} X_i}{16}$.

3. Step 4: A suitable rejection criteria may be selected as $\overline{X} < 7.9$ or $\overline{X} > 8.1$.

4. Step 5: The identification of the rejection criteria must aid us in assigning the level of significance of the test α.[20] This may be estimated as follows.

$$\alpha = \text{Prob}\Big(\overline{X} < 7.9 \text{ or } \overline{X} > 8.1 \text{ when } \mu = 8\Big). \tag{6.9}$$

$$\begin{aligned} P(\overline{X} < 7.9) &= P(\frac{\overline{X} - \mu}{\sigma/\sqrt{n}} < \frac{7.9 - 8}{0.2/\sqrt{16}}) && (6.10)\\ &= P(Z < -2.0) && (6.11)\\ &= 0.0228. && (6.12) \end{aligned}$$

   Likewise, $P(\overline{X} > 8.1) = P(Z > 2.0) = 0.0228$. Since the events $\overline{X} < 7.9$ and $\overline{X} > 8.1$ are *disjoint*; we have

$$\alpha = \text{Prob(type-1 error)} = P(\overline{X} < 7.9) + P(\overline{X} > 8.1) = 0.0228 + 0.0228 = 0.0456.$$

[19] In most practical cases, σ will not be known and we may have to base our analysis on the t distribution as opposed to the standard normal distribution that is used in this problem.
[20] We may also interpret α as the maximum allowable type-1 error.

6.6.6 *Additional comments on the level of significance α*

1. Notwithstanding the insight that may be gleaned from the previous example, we may not have a clear idea of what an appropriate maximum allowable type-1 error should be for a typical statistical experiment. There is no general rule of thumb to choose α.

2. α may also be sensitive to minor changes in the sample statistic and may conflate matters in appropriately testing the veracity of the hypothesis.

3. There is always a trade-off between α and β[21] because any efforts to reduce one may likely increase the other.

[21] β is the probability of making type-2 error. We will not discuss much about β in this introductory text. Interested readers may refer to more advanced texts listed in the bibliography.

6.6.7 *Two sample test for means*

In this section, we will discuss another test of hypothesis which uses the t distribution.

Consider a case where we have two different populations that are normally distributed with the same variance. A random variable sampled from each population is denoted by

$X_1 \sim N(\mu_1, \sigma^2)$ and $X_2 \sim N(\mu_2, \sigma^2)$. Further, let there be n_1 samples taken from the first population: $X_{1i} \sim N(\mu_1, \sigma^2)$; $\quad i = 1, 2, 3, ..., n_1$ and let there be n_2 samples taken from the second population: $X_{2j} \sim N(\mu_2, \sigma^2)$; $\quad j = 1, 2, 3, ..., n_2$.

1. Step 1: Construct the hypothesis.

$$\begin{aligned} H_0: &\quad \mu_1 = \mu_2 \\ H_1: &\quad \mu_1 \neq \mu_2 \qquad \text{(double sided test)} \\ &\quad \text{(alternatively, } \mu_1 > \mu_2 \text{ or } \mu_1 < \mu_2) \qquad \text{(single sided test)} \end{aligned}$$

   Further, choose and set α.

2. Steps 2 & 3: The test statistic is $t := \frac{(\overline{X_1} - \overline{X_2}) - (\mu_1 - \mu_2)}{S\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}}$ where $S^2 = \frac{(n_1 - 1)S_1^2 + (n_2 - 1)S_2^2}{n_1 + n_2 - 2}$ and S_j^2 is the sample variance corresponding to the samples taken from the j^{th} population set. Here $j = 1, 2$.

3. Step 4: Identify a rejection criteria.[22] There may be three distinct cases depending on the type of test (double or single sided alternate hypotheses).

   - Reject H_0 in favor of H_1 $(\mu_1 \neq \mu_2)$ if

$$|t| \geq t(\alpha/2, n_1 + n_2 - 2).$$

   - Reject H_0 in favor of H_1 $(\mu_1 > \mu_2)$ if

$$t \geq t(\alpha, n_1 + n_2 - 2).$$

   - Reject H_0 in favor of H_1 $(\mu_1 < \mu_2)$ if

$$t \leq -t(\alpha, n_1 + n_2 - 2).$$

   The right hand side term refers to the t observable value from the t distribution with a given significance level ($\alpha/2$ or α as stated above) and $(n_1 + n_2 - 2)$ degrees of freedom.

[22] In the parlance of statistics, one never accepts a null hypothesis. One either rejects or fails to reject the null hypothesis against an alternate hypothesis.

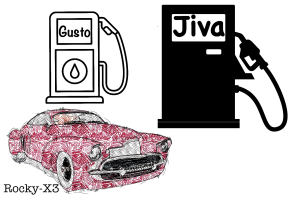

Figure 6.19: My vintage car `Rocky-X3` is likely to derive better mileage when powered by gasoline brand Gusto.

6.6.8 *Example: Choosing between two gasoline brands for optimal mileage and performance*

While comparing two different gasoline brands, a consumer survey reveals the following:

- a full tank of brand Gusto requires 4 cans and covers 546 km with a standard deviation of 31 km,
- a full tank of brand Jiva requires 4 cans and covers 492 km with a standard deviation of 26 km.

Assume that the performance parameters (mentioned above) of both brands are sampled from Normal distributions with equal variances; test if there is a significantly

better value in terms of mileage offered by Gusto over Jiva or if the mileage of both brands are statistically similar. Choose $\alpha = 0.05$.

Solution: Let us define the hypothesis at the outset as follows. We will use the subscript 1 for the brand Gusto and the subscript 2 for the brand Jiva.

$$H_0: \quad \mu_1 = \mu_2$$
$$H_1: \quad \mu_1 > \mu_2$$

For samples from brand Gusto, we have: $\overline{X}_1 = 546,\ S_1 = 31,\ n_1 = 4$.
For samples from brand Jiva, we have: $\overline{X}_2 = 492,\ S_2 = 26,\ n_2 = 4$.

The sample variance $S^2 = \frac{(4-1)31^2+(4-1)26^2}{4+4-2} \implies S = 28.609$. Therefore,

$$t \text{ (under } H_0) = \frac{546 - 492}{28.609\sqrt{1/4 + 1/4}} = 2.67.$$

The observable value $t(0.05, 6) = 1.9432$ may be computed either from the t distribution look-up table or using Matlab `>>tinv(1-0.05,6)` where $(n_1 + n_2 - 2) = (4 + 4 - 2) = 6$. Since this is a single tailed test

$$H_1: \ \mu_1 > \mu_2,$$

and because

$$t_{\text{calculated}} = 2.67 > t(0.05, 6) = 1.9432;$$

we reject H_0 in favor of H_1. In other words, brand Gusto will likely give us better mileage than brand Jiva at the level of significance $\alpha = 0.05$ (or with 95% confidence level).

6.6.9 *Analysis of variance (ANOVA)*

It must be emphasized that the t statistic based test for two means cannot be generalized for more than two different population sets. For multi-population tests, we may have to resort to *analysis of variance* (ANOVA). In order to appreciate the machinery of ANOVA, it is crucial to understand some notations and conventions. We will start here with a note on the data representation.

One-way ANOVA

Data collated from survey samples is denoted by y_{ij} where the first subscript represents the i^{th} population groups ($i = 1, 2, ..., t$) and the second subscript represents the j^{th} observation (data point; $j = 1, 2, ..., n$). We will consider $n_1,\ n_2,\ ...,\ n_t$ observations for the t population groups. If $n_1 = n_2 = ... = n_t = n$, then we have a *balanced* data set. The total number of observations is $\sum_{i=1}^{t} n_i$ ($= nt$ in the case of balanced data).

Null hypothesis:

$$\begin{aligned} H_0 &: \quad \mu_1 = \mu_2 = \cdots = \mu_t \\ H_1 &: \quad \text{one of the above equality is not satisfied.} \end{aligned} \tag{6.13}$$

Assumption: Data from each population group is normally distributed $N(\mu_i, \sigma^2)$ where σ^2 is the same across population groups.

Organization of data:

Population group	observations/data	$\sum_j y_{ij}$ (totals)	$\frac{Y_{i.}}{n_i}$ (means)	sum of squares
1	$y_{11}\ y_{12}\ \cdots\ y_{1n_1}$	$Y_{1.}$	$\overline{y}_{1.}$	SS_1
2	$y_{21}\ y_{22}\ \cdots\ y_{2n_2}$	$Y_{2.}$	$\overline{y}_{2.}$	SS_2
3	$y_{31}\ y_{32}\ \cdots\ y_{3n_3}$	$Y_{3.}$	$\overline{y}_{3.}$	SS_3
.	.	.	.	.
.	.	.	.	.
.	.	.	.	.
t	$y_{t1}\ y_{t2}\ \cdots\ y_{tn_t}$	$Y_{t.}$	$\overline{y}_{t.}$	SS_t
	Overall	$Y_{..}$	$\overline{y}_{..}$	SS_p

We have used the convention whereby the position of the . (dot) in the subscript represents which of the two indices (in the subscript) are being summed. Eg., for some variable α_{ij}, we will use the following convention for summation: $\sum_i = \alpha_{ij} = \alpha_{.j}$ where the summation is performed over the first index i.

Sum of squares: $SS_i = \sum_j (y_{ij} - \overline{y}_{i.})^2 \equiv \sum_j y_{ij}^2 - \frac{Y_{i.}^2}{n_i}$.
Pooled sum of squares: $SS_p = \sum_{i=1}^{t} SS_i$.
Pooled degrees of freedom: $\sum_{i=1}^{t} n_i - t = t(n-1)$.[23]

[23] The last equality is true for balanced data.

The pooled variance s_p^2 can now be defined as $s_p^2 = \frac{SS_p}{\sum_{i=1}^{t} n_i - t}$. Now another estimate of the sample variance is possible by considering the mean data across the population groups (factor levels). This sample variance estimate is formulated as follows: $s_{means}^2 = \frac{\sum_i (\overline{y}_{i.} - \overline{y}_{..})^2}{t-1}$. Under the null hypothesis and based on the discussions under the first paragraph of the section on sampling distributions, we may deduce that the factor level means have a distribution with mean μ and variance $\frac{\sigma^2}{n}$. Thus we have an estimate for the population variance $\sigma^2 = n s_{means}^2$ with $(t-1)$ degrees of freedom. Of course, an alternate estimate of the population variance is s_p^2 with $t(n-1)$ degrees of freedom. We know from the definition of the F statistic that the F value represents the ratio of two independent estimates of a common variance. Therefore,

$$F_{cal} = \frac{n s_{means}^2}{s_p^2}. \tag{6.14}$$

If $F_{cal} > F_\alpha(t-1, (n-1)t)$, then we reject H_0.

Alternate formulation of one-way ANOVA

This alternate formulation leads to the same inference. In this formulation, we have the following definitions.

$$SSB \text{ (sum of sqs. between groups)} = \sum_i \frac{(Y_{i.})^2}{n_i} - \frac{Y_{..}^2}{\sum_i n_i} \text{ with } (t-1) \text{ degrees of freedom.} \tag{6.15}$$

$$SSW \text{ (sum of sqs. within groups)} = \sum_j \sum_i y_{ij}^2 - \sum_i \frac{Y_{i.}^2}{n_i} \text{ with } (\sum_i n_i - t) \text{ degrees of freedom.} \tag{6.16}$$

Consequently, the total sum of squares is $TSS = SSB + SSW$. The one-way ANOVA table can now be re-formulated thusly.

Source	d.o.f.†	SS†	MS† = $\frac{SS}{d.o.f.}$	F_{cal}
between groups	$t-1$	SSB	MSB	$\frac{MSB}{MSW}$
within groups	$(\sum_i n_i - t)$	SSW	MSW	
total	$\sum_i n_i - 1$	TSS		

† d.o.f. means degrees of freedom, SS means sum of squares, MS means mean sum of squares.

Figure 6.20: Rice plantation in multiple plots of land to test the productivity across different rice varieties.

Inference: If $F_{cal} > F_\alpha(t-1, (n-1)t)$, then we reject H_0 in favor of H_1.

6.6.10 Example: Rice yield across varieties

An experiment to compare the yield of four varieties of rice is conducted. Each of the plots on a test farm where soil fertility is fairly homogeneous is treated alike relative to water and fertilizer. Four plots are randomly assigned each of the four varieties of rice. The yield in kg/acre is recorded for each plot for this randomized experiment. Does the data presented in the following table indicate a difference in the mean yield between the four varieties? Choose $\alpha = 0.01$.

variety	yield			
1	934	1041	1028	935
2	880	963	924	946
3	987	951	976	840
4	992	1143	1140	1191

Solution:

The hypothesis is stated as follows.

$$H_0: \quad \mu_1 = \mu_2 = \mu_3 = \mu_4$$
$$H_1: \quad \text{not all varieties have the same mean.}$$

Here μ_i denotes the mean yield of the i^{th} variety. $n = 4$, $t = 4$. The one-way ANOVA table is printed below.

Rice variety	yield data	$Y_{i.}$ (totals)	$\overline{y_{i.}}$ (means)	SS_i
1	934 1041 1028 935	3938	984.50	10085
2	880 963 924 946	3713	928.25	3868.75
3	987 951 976 840	3754	938.50	13617
4	992 1143 1140 1191	4466	1116.5	22305
	overall	15871	991.94	49875.75

$ns^2_{means} = n\frac{\sum_i(\overline{y}_{i.}-\overline{y}_{..})^2}{t-1} = 29977.06$. Further, $s^2_p = \frac{\sum_i SS_i}{t(n-1)} = 4156.31$.

$$F_{cal} = \frac{29977.06}{4156.31} = 7.21.$$

Using the F distribution table or using Matlab, we can find $F_{0.01}(3,12) = 5.95$. Since $F_{cal} > F_{0.01}(3,12)$, we reject H_0 and infer that there is significant difference in yield between the different rice yield.

We would arrive at the same conclusion if we had used the alternate approach to perform the calculation. In the alternate approach, the one-way ANOVA table is as follows.

Source	d.o.f.	SS	MS = $\frac{SS}{d.o.f.}$	F_{cal}
between varieties	3	89931.19	29977.06	$\frac{MSB}{MSW} = 7.21$
within varieties	12	49875.75	4156.31	
total	15	139806.94		

6.6.11 *Advanced ANOVA techniques*

An extension of the one way ANOVA is a two way ANOVA technique where we may consider the influence of two independent factors (eg., we may want to test the influence of rice variety and temperature of the environment in our aforementioned experimental example) on the dependent observable or outcome (eg., yield of the rice). In such a scenario, certain combinations of the two factors may *interact* differently from a simple additive approach. This interaction phenomena must necessarily be captured in the analysis. There may be problems where more than two factors may be involved, and we may have to consider what is known as a 2^k *factorial experimental design*. We will not elaborate on these multi-factor ANOVA designs here. However, an illustrative exercise problem is included at the end of the chapter for the interested audience to learn this technique.

6.7 *Chapter project: Prioritizing post-disaster reconstruction measures*

6.7.1 *Epilogue: Identifying issues that are universally plaguing reconstruction efforts by performing ANOVA on the survey ratings of the managers*

Now that we know the basic principles involved in performing ANOVA, let us use the survey ratings of the managers to populate the following ANOVA table.

source	degree of freedom	sum of sqs.	mean of sqs.	F_{cal}
between groups	$dfB = t - 1$	**SSB**	**MSB**	$F_{cal} = \frac{MSB}{MSW}$
within groups	$dfW = \sum_i n_i - t$	**SSW**	**MSW**	
total	$\sum_i n_i - 1 = n - 1$	**TSS = SSB+SSW**		

Here number of groups = $t = 6$ **and number of observations in group** i **=** $n_i = 6$ **and** $n = 36$.

Software Implementation

Use matlab to construct the one-way ANOVA table and implement the following algorithm.

⇒ **Compute the entries of the ANOVA table. One-way ANOVA computation involves the following calculations:**

1. **Compute** $Y_{i.} = \sum_{j=1}^{n_i} y_{ij}, Y_{..} = \sum_{i,j} y_{ij}$.
2. **Compute** $SSB = \sum_i \frac{Y_{i.}^2}{n_i} - \frac{Y_{..}^2}{\sum_i n_i}$, $SSW = \sum_{i,j} y_{ij}^2 - \sum_i \frac{Y_{i.}^2}{n_i}$, $TSS = SSB + SSW$.
3. **Set** $MSB = \frac{SSB}{t-1}$ **and** $MSW = \frac{SSW}{\sum_i n_1 - t}$.
4. **Compute** $F_{cal} = \frac{MSB}{MSW}$.

⇒ **Next, compare** F_{cal} **and** $F_{tab} = F_\alpha(dfB, dfW)$
(F_{tab} **is found from** $F-$**distribution table).**

⇒ **If** $F_{tab} > F_{cal}$**, then fail to reject** H_0 **(i.e. possibly all means (mean rating values) are *statistically* equal); else if** $F_{tab} < F_{cal}$**, then reject** H_0 **(and abandon dwelling on the issue for the time being).**

⇒ **If** $F_{tab} > F_{cal}$ **and if** $\mu = \frac{\sum_{i,j} y_{ij}}{n} > 5.0$**, then the issue is universally relevant across cities and demands corrective measures to successfully implement PDR projects.**

Questions

1. **State the assumptions of one-way ANOVA. Comment if these assumptions seem reasonable in the context of the PDR data provided here.**
2. **Implement the algorithm prescribed above in Matlab. Specifically, write a matlab script to construct and display the ANOVA table for each of the dataset provided in the interlude section 6.5.1 of the project.**
3. **Compute** F_{tab} **for the given problem from the** $F-$ **distribution table corresponding to a level of significance of test** $\alpha = 0.01$.
4. **Based on the strategy prescribed in the algorithm, select and mention the issues that are universally relevant across different cities and that need immediate redressal.**

5. Mention at least two drawbacks of one-way ANOVA test.

6.8 Selected bibliography

1. *Probability and Statistical Inference* by *R. V. Hogg, E. A. Tanis, D. L. Zimmerman*, Pearson (ninth edition), 2020.

2. *Applied Statistics and Probability for Engineers* by *D. C. Montgomery, G. C. Runger*, Wiley (sixth edition), 2019.

3. *Fundamentals of Mathematical Statistics* by *S. C. Gupta, V. K. Kapoor*, Sultan Chand & Sons (eleventh edition), 2019.

4. *Probability and Measure* by *P. Billingsley*, John Wiley and Sons (second edition), 1990.

5. *Introduction to Probability and Statistics for Engineers and Scientists* by *S. M. Ross*, Academic Press, Elsevier (sixth edition), 2021.

6.9 Exercise problems

Figure 6.21: Its a long queue!

1. (***Service time of a cashier in a shopping mall: application of CLT***) A cashier in a shopping mall serves customers standing in the queue one by one. Suppose that the service time X_i for the i^{th} customer has a mean $E(X_i) = 2$ min and $Var(X_i) = 1$ min. $\{X_i\}$ for all i customers are independent random variables. Let W be the total time spent by the cashier to serve fifty customers. Find $P(80 < W < 120)$.

2. (***Testing efficacy of student's health and welfare programs in schools***) A middle school conducted a survey of its students to collect health information for future planning purposes. The survey revealed that a substantial proportion of students were not engaging in regular exercise, many did not have access to optimal nutrition and a substantial number were exposed to environmental hazards like pollution. In response to a question on regular exercise, 60% of all parents of the students reported that their children were not regularly exercising, 25% reported their wards were exercising sporadically and 15% reported that their children were exercising regularly. The next year the school launched a health promotion campaign on campus in an attempt to increase health behaviors among students. The program included modules on exercise, nutrition, and environmental awareness. To evaluate the impact of the program, the school again surveyed students and their parents. The survey was completed by 470 students and the following data were collected on the exercise question:

	no exercise	sporadic exercise	regular exercise	total
number of students	255	125	90	470

Based on the data, is there evidence of a shift in the distribution of responses to the exercise question following the implementation of the health promotion campaign on campus? In order to answer this question, design a suitable statistical experiment and frame a hypothesis. Consider a 5% level of significance for testing your hypothesis.

3. (***Testing level of bacterial contamination in shipped packages***) Matlab provides an inbuilt database called *hogg* that is sourced from the work of Hogg and Ledolter.[24] This dataset has a record of total bacteria count in randomly selected cartons of milk packed in different shipping boxes. The relevant data may be obtained by typing the following commands in the Matlab command window.

[24] *Engineering Statistics* by R. V. Hogg, and J. Ledolter, Macmillan USA, (1987).

```
>> load hogg
>> hogg
hogg =

    24    14    11     7    19
    15     7     9     7    24
    21    12     7     4    19
    27    17    13     7    15
    33    14    12    12    10
    23    16    18    18    20
```

Here the columns represent the randomly selected shipping boxes and the rows constitute the bacterial count in randomly picked milk cartons from the respective shipping box. Perform a one-way ANOVA to test if the bacterial contamination is the same, on an average, across the five different shipping boxes. Use the level of significance of test $\alpha = 0.01$. Compare your analysis with the Matlab inbuilt `anova1` function that outputs the one-way ANOVA table and the p-value of the test.[25]

[25] You may use the following Matlab commands: `>>[p,tbl,stats] = anova1(hogg);` and `>>doc anova1` to learn about testing the validity of the null hypothesis from the p-value of the ANOVA experiment.

4. (***Consistency of caffeine content in coke across beverage counters***) Coke is available as a fountain soft drink from different vendors and beverage counters across the world. In order to attain, consistency in quality, it is desired that the caffeine content in mg per 12 oz does not exceed 34 mg. The data collected from fifty randomly selected beverage counters from across the world is available in a consolidated manner through this link here: coke.csv. Perform a suitable test of hypothesis to validate the null hypothesis (mean caffeine content across counters is 34 mg per 12 oz) against the alternate hypothesis that the mean caffeine content is greater than 34 mg per 12 oz. In this context, answer the following questions.

   (a) Clearly state the null and the alternate hypothesis in mathematical terms.

   (b) Identify the appropriate statistical test for this experiment.

   (c) Verify the assumptions for using this aforementioned test is met.[26]

   (d) Calculate the sampling distribution of mean under the null hypothesis.

   (e) Conduct the test of hypothesis and report your inference at $\alpha = 0.05$ significance level.

Figure 6.22: Watch out for the extra dose of caffeine in your drink!

5. (***Prognosis of pulmonary infection***) The change in the amount of carbon monoxide transfer, that is an indicator of improved pulmonary function in smokers with chickenpox, over a one week time frame, is recorded as follows: 33, 2, 24, 17, 4, 1, -6 (units are in ml). Is there an evidence of significant improvement in pulmonary function at a 95% confidence level,

   (a) if the data are normally distributed with variance $\sigma^2 = 100$,

[26] You may want to use the Matlab inbuilt function `kstest` to check one of these assumptions.

(b) if the data are normally distributed with unknown variance σ^2?

You may use $\alpha = 0.05$.

6. (***Efficacy of calcium channel blockers among hypertensive patients of age group 45-59***) The efficacy of a treatment for hypertension is to be studied using a randomized clinical trial. Thirty-eight hypertensive patients in the age group 45-59 were randomly allocated to either a placebo group (who were administered over the counter potassium tablets) or an intervention group (who were administered a calcium channel blocker) and a two-month follow-up study was performed at the clinic. At the end of the trial phase, the difference in systolic blood pressure was measured for patients in each group and recorded. A summary of the results is given below.

group	number of patient participants	mean difference in systolic blood pressure	sample variance
placebo	21	-0.108	2.101^2
intervention	17	3.753	4.630^2

Is there any evidence of significant improvement in the treatment group? Use $\alpha = 0.01$.

7. (***How long will you keep your first car?***) An economist wishes to investigate whether people are keeping cars longer now than in the past. He knows that five years ago, 38% of all passenger vehicles in operation were at least ten years old. He commissions a study in which 325 automobiles are randomly sampled. Of them, 132 are ten years old or older.

   (a) Find the sample proportion.

   (b) Find the probability that, when a sample of size 325 is drawn from a population in which the true proportion is 0.38, the sample proportion will be as large as the value you computed in part (a). You may assume that the normal distribution applies.

   (c) Give an interpretation of the result in part (b). Is there strong evidence that people are keeping their cars longer than was the case five years ago?

8. (***Cholesterol content in eggs***) Suppose the mean amount of cholesterol in eggs labeled "large" is 186 milligrams, with standard deviation 7 milligrams. Find the probability that the mean amount of cholesterol in a sample of 144 eggs will be within 2 milligrams of the population mean.

Figure 6.23: Estimating cholesterol in eggs using a sampling distribution.

9. (***Simulating Cochran's theorem***) Write a Matlab code to simulate and establish the veracity of the following theorem, *if* $Z_1, ..., Z_k$ *are independent and identically distributed (i.i.d.) standard normal random variables, then* $\sum_{i=1}^{k}(Z_i - \overline{Z})^2 \sim \chi^2_{k-1}$; *where* $\overline{Z}$ *is the sample mean.*

10. (***Two-way ANOVA***) Aircraft primer paints are applied to aluminium surfaces by two methods: *dipping* and *spraying*. The purpose of using the primer is to improve paint adhesion, and some parts can be primed using either method. The process engineering group responsible for this operation is interested whether three different primers differ in their adhesion properties. A factorial experiment was performed to investigate the effect of paint primer type and application method on paint adhesion. For each combination of primer type and application method, three specimens were painted, then a finish paint was applied, and the adhesion force was measured. The data from the experiment are shown in the table below. Perform a two-way ANOVA and identify the most effective primer type and the better of the two paint application method.

primer type	dipping	$y_{ij.}$	spraying	$y_{ij.}$	$y_{i..}$
1	4.0, 4.5, 4.3	12.8	5.4, 4.9, 5.6	15.9	28.7
2	5.6, 4.9, 5.4	15.9	5.8, 6.1, 6.3	18.2	34.1
3	3.8, 3.7, 4.0	11.5	5.5, 5.0, 5.0	15.5	27.0
$y_{.j.}$	40.2		49.6		$y_{...} = 89.8$

□