CHAPTER 3. UNCERTAINTY CHARACTERIZATION

This chapter discusses statistical analysis based on available sample data that characterizes uncertain data in a statistical form. Specifically, it introduces statistical procedures to determine an appropriate probability distribution for a random variable based on a limited set of sample data. There are two approaches in the statistical data analysis techniques: (a) conventional statistical methods (graphical methods and statistical hypothesis tests) and (b) Bayesian methods.

3.1 Conventional (or Frequentist) Statistical Methods

The conventional statistical methods impose models (both deterministic and probabilistic) on the data. Deterministic models include, for example, regression models and analysis of variance (ANOVA) models. The most common probabilistic models include the graphical methods and quantitative methods.

3.1.1 Graphical Methods

- Histogram (Fig. 3.1)
  The purpose of a histogram is to graphically summarize the distribution of a univariate data set. This histogram graphically shows the following:

  1. center (i.e., the location) of the data;
  2. spread (i.e., the variation) of the data;
  3. skewness of the data;
  4. presence of outliers; and
  5. presence of multiple modes in the data.

  These features provide strong indications of the proper distributional model for the data. The probability plot or a goodness-of-fit test can be used to verify the distributional model.

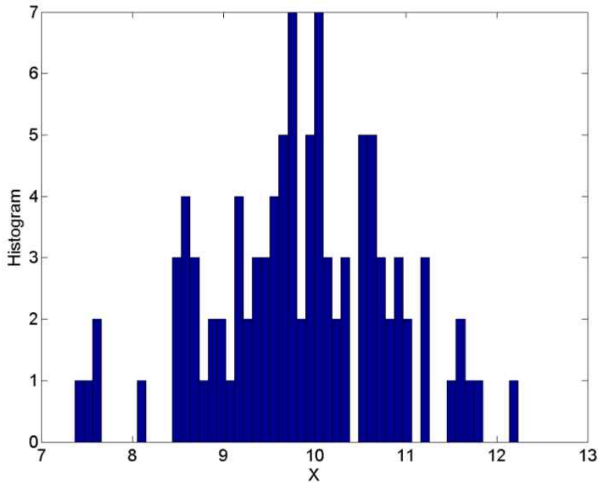
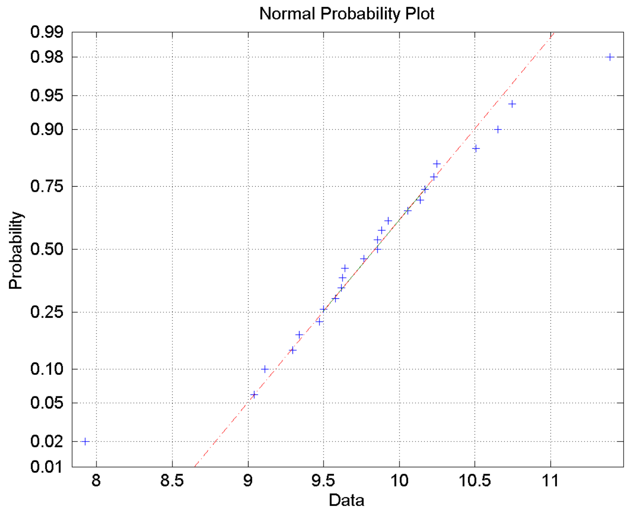


Figure 3.1: Histogram

- Normal probability plot

The normal probability plot is a graphical technique for assessing whether or not a data set can be approximated as a normal distribution. The data are plotted against a theoretical normal distribution in such a way that the points should form an approximate straight line. Departures from this straight line indicate departures from normality. The normal probability plot is a special case of the probability plot.

```
>> x = normrnd(10,1,25,1);
>> normplot(x);
```



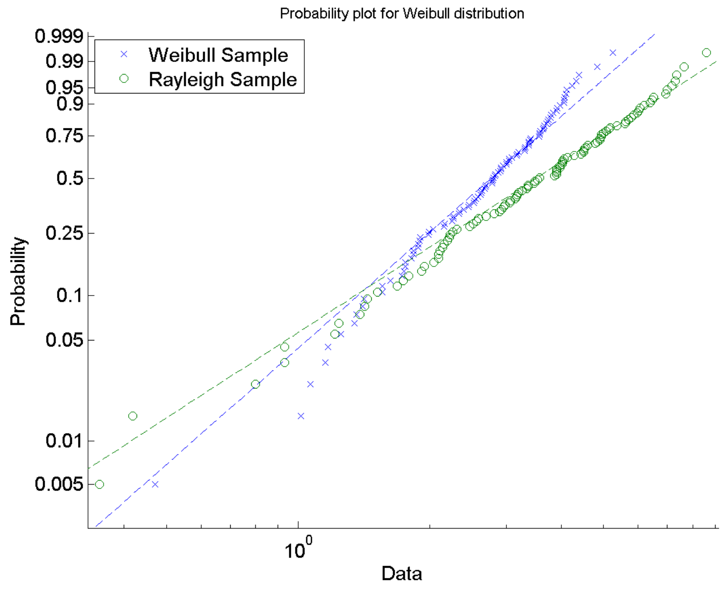
- Probability plot

The uniform distribution has a linear relationship between ordered physical data and probability. So any probability distribution can be used for approximating a given data set if a probability distribution is related to the uniform distribution. The relationship can be defined as

$$U(\mathrm{i}) = G(P(x(\mathrm{i})))$$

where $P(\mathrm{i})$ is the probability of the event $E = \{X \mid x(\mathrm{i}) \in \Omega\}$ and $U(\mathrm{i})$ follows a uniform distribution.

```
>> x1 = wblrnd(3,2,100,1);
>> x2 = raylrnd(3,100,1);
>> probplot('weibull',[x1 x2])
>> legend('Weibull Sample','Rayleigh Sample','Location','NW')
```



Rayleigh distribution is a special case of weibull distribution when a shape parameter is 2. Therefore both distributions follow the straight lines very closely.

Homework 8: *Graphical methods*
Use the data set for elastic modulus and yield strength in the excel file named 'tensile_test.xlex'. Build histograms and plot each data set on the normal probability plot to determine if they follow a normal distribution. Discuss your observation.

3.1.2 Quantitative Methods

- Statistical Moments:
  First-order moment (e.g., mean, location)

$$\text{mean}(\bar{X}) = \frac{\sum_{i=1}^{N} x_i}{N} \tag{19}$$

a. Confidence limits (or interval) for the mean (T-test)

$$\bar{X} \pm t_{(\alpha/2,N-1)}\, s/\sqrt{N} \tag{20}$$

where $\bar{X}$ and s are the sampled mean and standard deviation, N is the sample size, α is the desired significance level (or $1-\alpha$ = confidence level), and $t_{(\alpha/2,N-1)}$ is the critical value of the t-distribution with N-1.

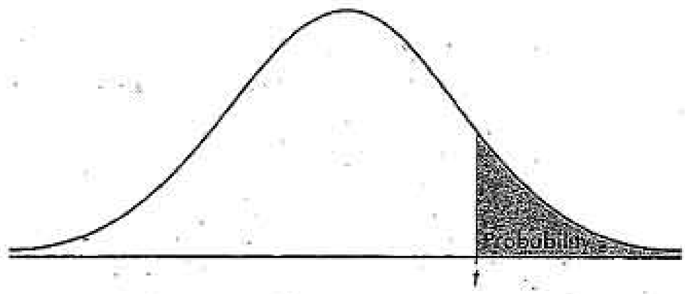


TABLE B: t-DISTRIBUTION CRITICAL VALUES

	Tail probability p											
df	.25	.20	.15	.10	.05	.025	.02	.01	.005	.0025	.001	.0005
1	1.000	1.376	1.963	3.078	6.314	12.71	15.89	31.82	63.66	127.3	318.3	636.6
2	.816	1.061	1.386	1.886	2.920	4.303	4.849	6.965	9.925	14.09	22.33	31.60
3	.765	.978	1.250	1.638	2.353	3.182	3.482	4.541	5.841	7.453	10.21	12.92
4	.741	.941	1.190	1.533	2.132	2.776	2.999	3.747	4.604	5.598	7.173	8.610
5	.727	.920	1.156	1.476	2.015	2.571	2.757	3.365	4.032	4.773	5.893	6.869
6	.718	.906	1.134	1.440	1.943	2.447	2.612	3.143	3.707	4.317	5.208	5.959
7	.711	.896	1.119	1.415	1.895	2.365	2.517	2.998	3.499	4.029	4.785	5.408
8	.706	.889	1.108	1.397	1.860	2.306	2.449	2.896	3.355	3.833	4.501	5.041
9	.703	.883	1.100	1.383	1.833	2.262	2.398	2.821	3.250	3.690	4.297	4.781
10	.700	.879	1.093	1.372	1.812	2.228	2.359	2.764	3.169	3.581	4.144	4.587
11	.697	.876	1.088	1.363	1.796	2.201	2.328	2.718	3.106	3.497	4.025	4.437
12	.695	.873	1.083	1.356	1.782	2.179	2.303	2.681	3.055	3.428	3.930	4.318
13	.694	.870	1.079	1.350	1.771	2.160	2.282	2.650	3.012	3.372	3.852	4.221
14	.692	.868	1.076	1.345	1.761	2.145	2.264	2.624	2.977	3.326	3.787	4.140
15	.691	.866	1.074	1.341	1.753	2.131	2.249	2.602	2.947	3.286	3.733	4.073
16	.690	.865	1.071	1.337	1.746	2.120	2.235	2.583	2.921	3.252	3.686	4.015
17	.689	.863	1.069	1.333	1.740	2.110	2.224	2.567	2.898	3.222	3.646	3.965
18	.688	.862	1.067	1.330	1.734	2.101	2.214	2.552	2.878	3.197	3.611	3.922
19	.688	.861	1.066	1.328	1.729	2.093	2.205	2.539	2.861	3.174	3.579	3.883
20	.687	.860	1.064	1.325	1.725	2.086	2.197	2.528	2.845	3.153	3.552	3.850
21	.686	.859	1.063	1.323	1.721	2.080	2.189	2.518	2.831	3.135	3.527	3.819
22	.686	.858	1.061	1.321	1.717	2.074	2.183	2.508	2.819	3.119	3.505	3.792
23	.685	.858	1.060	1.319	1.714	2.069	2.177	2.500	2.807	3.104	3.485	3.768
24	.685	.857	1.059	1.318	1.711	2.064	2.172	2.492	2.797	3.091	3.467	3.745
25	.684	.856	1.058	1.316	1.708	2.060	2.167	2.485	2.787	3.078	3.450	3.725
26	.684	.856	1.058	1.315	1.706	2.056	2.162	2.479	2.779	3.067	3.435	3.707
27	.684	.855	1.057	1.314	1.703	2.052	2.158	2.473	2.771	3.057	3.421	3.690
28	.683	.855	1.056	1.313	1.701	2.048	2.154	2.467	2.763	3.047	3.408	3.674
29	.683	.854	1.055	1.311	1.699	2.045	2.150	2.462	2.756	3.038	3.396	3.659
30	.683	.854	1.055	1.310	1.697	2.042	2.147	2.457	2.750	3.030	3.385	3.646
40	.681	.851	1.050	1.303	1.684	2.021	2.123	2.423	2.704	2.971	3.307	3.551
50	.679	.849	1.047	1.299	1.676	2.009	2.109	2.403	2.678	2.937	3.261	3.496
60	.679	.848	1.045	1.296	1.671	2.000	2.099	2.390	2.660	2.915	3.232	3.460
80	.678	.846	1.043	1.292	1.664	1.990	2.088	2.374	2.639	2.887	3.195	3.416
100	.677	.845	1.042	1.290	1.660	1.984	2.081	2.364	2.626	2.871	3.174	3.390
1000	.675	.842	1.037	1.282	1.646	1.962	2.056	2.330	2.581	2.813	3.098	3.300
∞	.674	.841	1.036	1.282	1.645	1.960	2.054	2.326	2.576	2.807	3.091	3.291
	50%	60%	70%	80%	90%	95%	96%	98%	99%	99.5%	99.8%	99.9%
	Confidence level C											

From the formula, it is clear that the width of the interval is controlled by two factors:

- ✓ As N increases, the interval gets narrower from the $\sqrt{N}$ term and $t_{(\alpha/2,N-1)}$. That is, one way to obtain more precise estimates for the mean is to increase the sample size.
- ✓ The larger the sample standard deviation, the larger the confidence interval. This simply means that noisy data, i.e., data with a large standard deviation, are going to generate wider intervals than data with a smaller standard deviation.

To test whether the population mean has a specific value, μ_0, against the two-sided alternative that it does not have a value μ_0, the confidence interval is converted to hypothesis-test form. The test is a one-sample *t*-test, and it is defined as:

H_0:	$\bar{X} = \mu_0$
H_1:	$\bar{X} \neq \mu_0$
Tested statistics:	$T = (\bar{X} - \mu_0)/(s/\sqrt{N})$
Significance level:	α (=0.05 is most commonly used.)
Critical region:	Reject the null hypothesis that the mean is a specified value, μ_0, if $T < -t_{(\alpha/2,N-1)}$ or $T > t_{(\alpha/2,N-1)}$

Let's say the null hypothesis is rejected. The p-value indicates the probability that the rejection of the null hypothesis is wrong.

```
>> x1 = normrnd(0.1,1,1,100);
>> [h,p,ci] = ttest(x1,0)
h =
     0
p =
    0.8323
ci =
   -0.1650    0.2045
```

The test fails to reject the null hypothesis at the default α. The 95% confidence interval on the mean contains 0.

```
>> x2 = normrnd(0.1,1,1,1000);
>> [h,p,ci] = ttest(x2,0)
h =
     1
p =
    0.0160
ci =
    0.0142    0.1379
```

The test rejects the null hypothesis at the default α. The p-value has fallen below $\alpha = 0.05$ and 95% confidence interval on the mean does not contain 0.

b. 1-factor ANOVA (Analysis of Variance)
http://www.itl.nist.gov/div898/handbook/eda/section3/eda354.htm

Second-order moment (e.g., variation)

$$\text{variation}(s^2) = \frac{\sum_{i=1}^{N}\left(x_i - \bar{X}\right)^2}{(N-1)} \tag{21}$$

a. Bartlett's test: http://www.itl.nist.gov/div898/handbook/eda/section3/eda357.htm
b. Chi-Square test: http://www.itl.nist.gov/div898/handbook/eda/section3/eda358.htm
c. F-test: http://www.itl.nist.gov/div898/handbook/eda/section3/eda359.htm
d. Levene test: http://www.itl.nist.gov/div898/handbook/eda/section3/eda35a.htm

The formula for computing the covariance of the variables X and Y is

$$\text{COV} = \frac{\sum_{i=1}^{N}\left(x_i - \bar{X}\right)\left(y_i - \bar{Y}\right)}{N-1} \tag{22}$$

- Maximum Likelihood Estimation (MLE):
  The principle behind the MLE method is that for a random variable X, if $x_1, x_2, \ldots, x_n$ are the N observations or sample values, then the estimated value of the parameter is the value most likely to produce these observed values. Consider the density function of X to be $f_X(x,\theta)$, where θ is the unknown parameter(s). In random sampling, the x_i's are assumed to be independent. If the likelihood of observing x_i's is proportional to their corresponding density functions, the likelihood function can be defined as

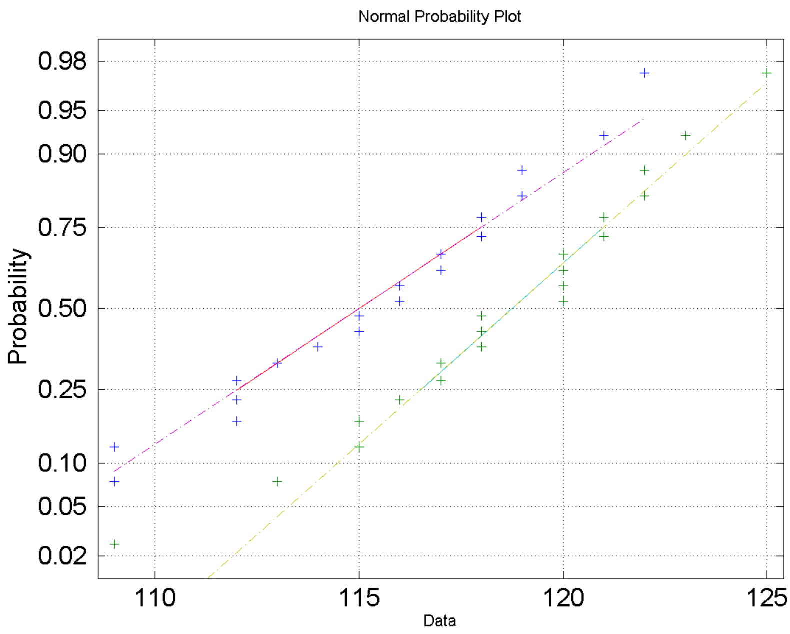
$$L(x_1, x_2, \cdots, x_n \mid \theta) = \prod_{i=1}^{N} f_X(x_i \mid \theta) \\ = f_X(x_1 \mid \theta) f_X(x_2 \mid \theta) \cdots f_X(x_n \mid \theta) \quad (23)$$

The MLE can be formulated as

$$\text{To determine } \theta, \text{maximize } \; \mathrm{L}(x_1, x_2, \ldots x_n)$$

Homework 9: *Quantitative methods*
Use the data set for elastic modulus and yield strength in the excel file named 'tensile_test.xlex'. Test whether or not the population mean has a specific value, μ_0=200 GPa, for a quality control. Let's assume the elastic modulus follow a normal distribution. Determine the optimal mean and standard deviation using the maximum likelihood method.

```
>> load gas

>> prices = [price1 price2];

>> normplot(prices)
```




```
>> sample_means = mean(prices)

sample_means =

  115.1500  118.5000
```

```
>> [h,pvalue,ci] = ttest(price2/100,1.1515)

h =

    1

pvalue =

  4.9517e-004

ci =

    1.1675    1.2025
```

- Distributional Measures:
  Chi-squared Goodness-of-Fit (GOF) Tests:
  The chi-square test is used to test if sampled data come from a population with a specific distribution. An attractive feature of the chi-square GOF test is that it can be applied to both continuous and discrete distributions. The chi-square GOF test is applied to binned data (i.e., data put into classes). So the values of the chi-square test statistic are dependent on how the data is binned. Another disadvantage of the chi-square test is that it requires a sufficient sample size in order for the chi-square approximation to be valid.

H_0:	The data follow a specified distribution.
H_1:	The data do not follow the specified distribution.
Significance level:	α (=0.05 is most commonly used.)
Test statistics:	For the chi-square goodness-of-fit computation, the data are divided into k bins and the test statistics is defined as

$$\chi^2 = \sum_{i=1}^{k} \left(O_i - E_i\right)^2 / E_i \tag{24}$$

where O_i is the observed frequency for bin i and E_i is the expected frequency for bin i. The expected frequency is calculated by

$$E_i = N \cdot \left(F(X_u) - F(X_l)\right)_i = N \cdot \bar{f}_i \tag{25}$$

where F is the cumulative distribution function (CDF) for the distribution being tested, X_u is the upper limit for bin i, X_l is the lower limit for a bin i, and N is the sample size.

Critical region: The hypothesis that the data are from a population with the specified distribution is rejected if

$$\chi^2 > \chi^2_{(\alpha,k-c)}$$

where $\chi^2_{(\alpha,k-c)}$ is the chi-square percent point function with k–c degrees of freedom and a significant level of α. k is the number of non-empty cells and c = the number of estimated parameters (including location and scale parameters and shape parameters) for the distribution.

Anderson-Darling (A-D) Goodness-of-Fit Test:
http://www.itl.nist.gov/div898/handbook/eda/section3/eda35e.htm

Kolmogorov-Smirnov (K-S) Goodness-of-Fit Test:

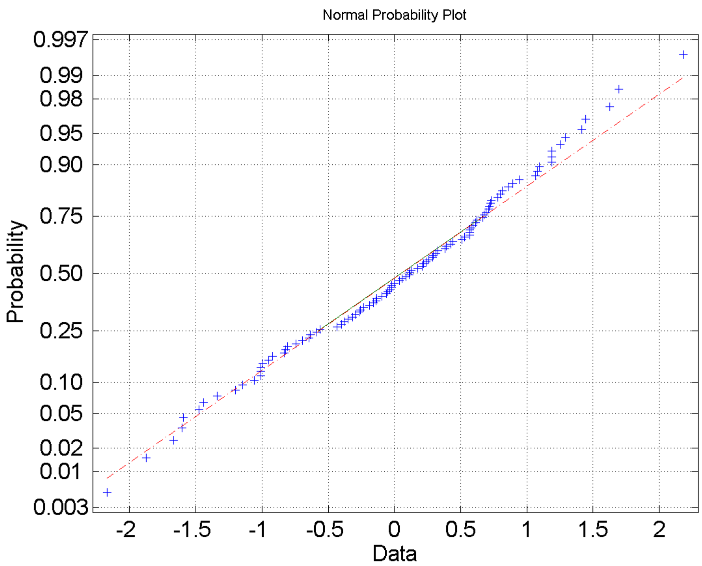
http://www.itl.nist.gov/div898/handbook/eda/section3/eda35g.htm

```
>> price2=normrnd(118,3.8,100,1);
>> [h,p] = chi2gof(price2,'cdf',{@normcdf,mean(price2),std(price2)})
>> [h,p] = chi2gof(price2,'cdf',{@normcdf,119,3.5})
```

```
>> x = randn(100,1);
>> [h,p,st] = chi2gof(x,'cdf',@normcdf)

h =

    0

p =

   0.370

st =

   chi2stat: 7.5909
         df: 7
      edges: [-2.1707 -1.2999 -0.8645 -0.4291 0.0063 0.4416 0.8770 1.3124 2.1832]
          O: [8 9 10 19 18 21 10 5]
          E: [9.6817 9.6835 14.0262 16.8581 16.8130 13.9138 9.5546 9.4690]
```

Degrees of Freedom	Probability										
	0.95	0.90	0.80	0.70	0.50	0.30	0.20	0.10	0.05	0.01	0.001
1	0.004	0.02	0.06	0.15	0.46	1.07	1.64	2.71	3.84	6.64	10.83
2	0.10	0.21	0.45	0.71	1.39	2.41	3.22	4.60	5.99	9.21	13.82
3	0.35	0.58	1.01	1.42	2.37	3.66	4.64	6.25	7.82	11.34	16.27
4	0.71	1.06	1.65	2.20	3.36	4.88	5.99	7.78	9.49	13.28	18.47
5	1.14	1.61	2.34	3.00	4.35	6.06	7.29	9.24	11.07	15.09	20.52
6	1.63	2.20	3.07	3.83	5.35	7.23	8.56	10.64	12.59	16.81	22.46
7	2.17	2.83	3.82	4.67	6.35	8.38	9.80	12.02	14.07	18.48	24.32
8	2.73	3.49	4.59	5.53	7.34	9.52	11.03	13.36	15.51	20.09	26.12
9	3.32	4.17	5.38	6.39	8.34	10.66	12.24	14.68	16.92	21.67	27.88
10	3.94	4.86	6.18	7.27	9.34	11.78	13.44	15.99	18.31	23.21	29.59

```
>> normplot(x)
```



Homework 10: *Probability Distribution & Statistical Moments*

Let us recall the example of fatigue tests. The sample data can be obtained about the physical quantities in the damage model below.

$$\frac{\Delta\varepsilon}{2} = \frac{\sigma_f'}{E}\left(2N_f\right)^b + \varepsilon_f'\left(2N_f\right)^c$$

Let us consider a 30 data set (Table 3.1) for the fatigue ductility coefficient (ε_f') and exponent (c) used in the strain-life formula shown above. Answer the following questions and ***provide a matlab code***:

(1) Construct the covariance matrix and find out the coefficient of correlation using the data set given in Table 3.1.

(2) Use normal, weibull, and lognormal distributions. Determine the most suitable parameters of three distributions for the fatigue ductility coefficient (ε_f') and exponent (c) using the MLE method.

(3) Find out the most suitable distributions for the data set (ε_f', c) using a GOF test.

(4) Verify the results with the graphical methods (histogram and probability plots).

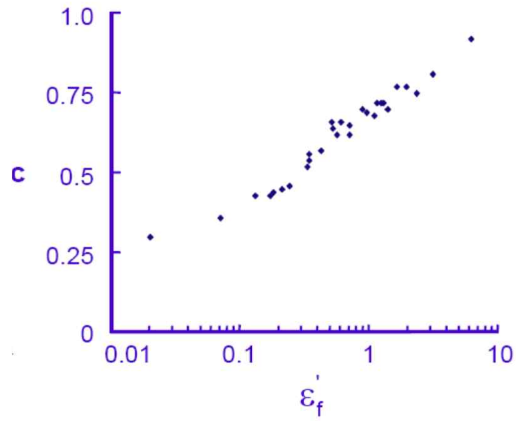


Figure 3.2: Statistical Correlation

Table 3.1: Data for the fatigue ductility coefficient and exponent

(ε_f', c)		(ε_f', c)		(ε_f', c)		(ε_f', c)		(ε_f', c)	
0.022	0.289	0.253	0.466	0.539	0.630	0.989	0.694	1.611	0.702
0.071	0.370	0.342	0.531	0.590	0.621	1.201	0.690	1.845	0.760
0.146	0.450	0.353	0.553	0.622	0.653	1.304	0.715	1.995	0.759
0.185	0.448	0.354	0.580	0.727	0.635	1.388	0.717	2.342	0.748
0.196	0.452	0.431	0.587	0.729	0.645	1.392	0.716	3.288	0.821
0.215	0.460	0.519	0.655	0.906	0.703	1.426	0.703	6.241	0.894

3.2 Bayesian

We have discussed methods of statistical inference which view the probability as relative frequency and exclusively rely on the sample data to estimate the underlying probability distribution of the population. In addition to these frequentist statistical methods, the Bayesian approach utilizes some prior information in conjunction with the sample information. The Bayesian inference is capable of continuously updating the prior information with evolving sample data to obtain the posterior information.

3.2.1 Bayes' Theorem

Bayes' theorem (also known as Bayes' rule or Bayes' law) is developed based on conditional probabilities. If A and B denote two events, $P(A|B)$ denotes the conditional probability of A occurring, given that B occurs. An important application of Bayes' theorem is that it gives a rule how to update or revise a prior belief to *a posterior* belief. Bayes' theorem relates the conditional and marginal probabilities of stochastic events A and B:

$$P(A|B)=\frac{P(B|A)\cdot P(A)}{P(B)} \tag{26}$$

Each term in Bayes' theorem has a conventional name:

- $P(A)$ is the prior probability or marginal probability of A. The prior probability can be treated as the subjective probability which expresses our belief prior to the occurrence of A. It is "prior" in the sense that it does not take into account any information about B.
- $P(B)$ is the prior or marginal probability of B, and acts as a normalizing constant.
- $P(A|B)$ is the conditional probability of A, given B. It is also called the posterior probability of A, given B because it depends upon the specified value of B.
- $P(B|A)$ is the conditional probability of B given prior information of A.

An important application of Bayes' theorem is that it gives a rule how to update or revise a *prior* belief to *a posterior* belief. Let us take a look at an interesting example to get a better understanding.

Example 3.1

There are three doors and behind two of the doors are goats and behind the third door is a new car with each door equally likely to provide the car. Thus the probability of selecting the car for each door at the beginning of the game is simply 1/3. After you have picked a door, say A, before showing you what is behind that door, Monty opens another door, say B, revealing a goat. At this point, Monty gives you the opportunity to switch doors from A to C if you want to. What should you do? (Given that Monty is trying to let you get a goat.)

Solution

The question is whether the probability is 0.5 to get the car since only two doors left, or mathematically, $P(\mathrm{A}|\mathrm{B}_{\mathrm{Monty}}) = P(\mathrm{C}|\mathrm{B}_{\mathrm{Monty}}) = 0.5$. Basically we need to determine the probabilities of two event E_1 = {A|B_{Monty}}, E_2 = {C|B_{Monty}}. We elaborate the computation in the following steps:

1. The prior probabilities read $P(A) = P(B) = P(C) = 1/3$.

2. We also have some useful conditional probabilities $P(B_{Monty}|A) = ½$, $P(B_{Monty}|B) = 0$, and $P(B_{Monty}|C) = 1$.

3. We can compute the probabilities of joint events as $P(B_{Monty}, A) = ½\times1/3 = 1/6$, $P(B_{Monty}, B) = 0$, and $P(B_{Monty}, C) = 1\times1/3 = 1/3$.

4. Finally, with the denominator computed as $P(B_{Monty}) = 1/6 + 0 + 1/3 = ½$, we then get $P(A|B_{Monty}) = 1/3$, $P(C|B_{Monty}) = 2/3$. Thus, it is better to switch to C.

3.2.2 Bayesian Inference

Let X and Θ be random variables with a joint probability density function $f(x, \theta)$, $\theta \in \Omega$. When the amount of data for X is small or X is rapidly evolving, its statistical parameter θ (e.g., μ, σ) is considered to be random. From the Bayesian point of view, θ is interpreted as a realization of a random variable Θ with a probability density $f_\Theta(\theta)$. Based on the Bayes' theorem, the posterior distribution of Θ given a new observation X can be expressed as

$$f_{\Theta|X}(\theta\,|\,x) = \frac{f_{X,\Theta}(x,\theta)}{f_X(x)} = \frac{f_{X|\Theta}(x\,|\,\theta)\cdot f_\Theta(\theta)}{f_X(x)} \quad (27)$$

It can be seen that the Bayesian inference employs both the prior distribution of θ, $f(\theta)$, and the conditional probability distribution of the sample (evidence or likelihood), $f_{X|\Theta}(x|\theta)$, to find a posterior distribution of θ, $f_{\Theta|X}(\theta|x)$. Let us consider a normal inference model as one example to illustrate the Bayesian inference process.

Example 3.2: Suppose that we have a set of random samples $\mathbf{x} = \{x_1, x_2,\ldots, x_M\}$ from a normal PDF $f_X(x; \mu, \sigma)$ of a random variable X, where μ is unknown and σ is known. Assume that the prior distribution of μ, $f_M(\mu)$, is a normal distribution with its mean, u, and variance, τ^2. Determine the posterior distribution of μ, $f_{M|X}(\mu|\mathbf{x})$.

Solution

Firstly, we compute the conditional probability of obtaining $\mathbf{x}$ given μ as

$$\begin{aligned} f_{X|M}(\mathbf{x}\,|\,\mu) &= \prod_{i=1}^{M}\frac{1}{\sqrt{2\pi}\sigma}\exp\left[-\frac{1}{2}\left(\frac{x_i-\mu}{\sigma}\right)^2\right] \\ &= \left(2\pi\sigma^2\right)^{-M/2}\exp\left[-\frac{1}{2\sigma^2}\sum_{i=1}^{M}\left(x_i-\mu\right)^2\right] \end{aligned} \quad (28)$$

Next, we compute the joint probability of $\mathbf{x}$ and μ as

$$f_{X,M}(\mathbf{x},\mu) = f_{X|M}(\mathbf{x}\,|\,\mu) f_M(\mu)$$
$$= \left(2\pi\sigma^2\right)^{-M/2}\left(2\pi\tau^2\right)^{-1/2}\exp\left[-\frac{1}{2\sigma^2}\sum_{i=1}^{M}(x_i-\mu)^2-\frac{1}{2\tau^2}(\mu-u)^2\right]$$
$$= K_1(x_1,\ldots,x_M,\sigma,u,\tau)\exp\left[-\left(\frac{M}{2\sigma^2}+\frac{1}{2\tau^2}\right)\mu^2+\left(\frac{M\bar{x}}{\sigma^2}+\frac{u}{\tau^2}\right)\mu\right]$$

We then set up a square with μ in the exponent as

$$f_{X,M}(\mathbf{x},\mu) = K_2(x_1,\ldots,x_M,\sigma,u,\tau)\exp\left[-\frac{1}{2}\left(\frac{M}{\sigma^2}+\frac{1}{\tau^2}\right)\left(\mu-\frac{\frac{M\bar{x}}{\sigma^2}+\frac{u}{\tau^2}}{\frac{M}{\sigma^2}+\frac{1}{\tau^2}}\right)^2\right]$$
$$= K_2(x_1,\ldots,x_M,\sigma,u,\tau)\exp\left[-\frac{1}{2}\left(\frac{M}{\sigma^2}+\frac{1}{\tau^2}\right)\left(\mu-\frac{M\tau^2\bar{x}+\sigma^2u}{M\tau^2+\sigma^2}\right)^2\right]$$

Since the denominator $f_X(x_1, x_2,\ldots, x_M)$ does not depend on μ, we then derive the posterior distribution of μ as

$$f_{M|X}(\mu\,|\,\mathbf{x}) = K_3(x_1,\ldots,x_M,\sigma,u,\tau)\exp\left[-\frac{1}{2}\left(\frac{M}{\sigma^2}+\frac{1}{\tau^2}\right)\left(\mu-\frac{M\tau^2\bar{x}+\sigma^2u}{M\tau^2+\sigma^2}\right)^2\right]$$

Clearly, this is a normal distribution with the mean and variance as

$$\hat{u} = \frac{M\tau^2\bar{x}+\sigma^2u}{M\tau^2+\sigma^2},\quad \hat{\tau} = \left(\frac{M}{\sigma^2}+\frac{1}{\tau^2}\right)^{-1} = \frac{\sigma^2\tau^2}{M\tau^2+\sigma^2} \tag{29}$$

Therefore, the Bayes estimate of μ is essentially a weighted-sum of the sample mean and the prior mean. In contrast, the maximum likelihood estimator is only the sample mean. As the number of samples M approaches the infinity, the Bayes estimate becomes equal to the maximum likelihood estimator since the sample data tend to have a predominant influence over the prior information. However, for the case of a small sample size, the prior information often plays an important role, especially when the prior variance τ^2 is small (or we have very specific prior information).

3.2.3 Conjugate Bayes Models

As can be seen in the Example 3.2, the Bayes inference and the maximum likelihood estimation essentially provide the same estimator if we have a very large sample size. In engineering practice, however, we often have very limited sample data possibly due to the high expense to obtain the data. In such cases, the maximum likelihood estimation may not give an accurate or even reasonable estimator. In contrast, the Bayesian inference would give much better estimator if we assume a reasonable prior assumption. By "reasonable", we mean that the prior assumption is at least consistent with the underlying distribution of the population. If there is no such consistency, the Bayesian inference may give an erroneous estimator due to the misleading prior information.

Another important observation we can make from Example 3.2 is that the posterior distribution shares a similar form (i.e., normal distribution) with the prior. In this case, we say that the prior is *conjugate* to the likelihood. If we have a conjugate prior, the posterior distribution can be obtained in an explicit form. Looking back to Example 3.2, we note that the normal or Gaussian family is conjugate to itself (or self-conjugate): if the likelihood function is normal, choosing a normal prior will ensure that the posterior distribution is also normal. Other conjugate Bayes inference models include the binomial inference, exponential inference, and Poisson inference. Among these inferences, the binomial inference is the most widely used. Consider a Bernoulli sequence of n experimental trials with x occurrences of an outcome whose probability of occurrence p_0 is unknown. We assume a beta prior $B(a,b)$ for the unknown binomial probability p_0, expressed as

$$f_{P_0}(p_0) = \frac{\Gamma(a,b)}{\Gamma(a)\Gamma(b)} p_0^{a-1}(1-p_0)^{b-1}$$

The likelihood function can be expressed according to a binomial distribution as

$$L(x;n,p_0) = C(n,x)\, p_0^x (1-p_0)^{n-x} \propto p_0^x (1-p_0)^{n-x}$$

We can easily obtain the posterior distribution of p_0 as a beta distribution, expressed as

$$f_{P_0|X}(p_0 \mid x) = \frac{\Gamma(x+a,n+b-x)}{\Gamma(x+a)\Gamma(n+b-x)} p_0^{x+a-1}(1-p_0)^{n+b-x-1}$$

The posterior distribution has the same form (beta distribution) as the prior distribution, leading to the conjugacy condition. Let us take a look at the use of this inference with a simple reliability analysis problem.

Example 3.3

Suppose that we intend to quantify the reliability of a product by conducting a sequence of 10 repeated tests. The product passes 8 of these tests and fails at the other two. We assume a beta prior $B(4, 4)$ for the probability of success (or reliability) p_0 in each test. Compute the posterior distribution of p_0 with the reliability test data.

Solution

Clearly, the parameters in this example take the following values: $a = 4$, $b = 4$, $x = 8$, $n = 10$. Then the posterior distribution can be obtained as $B(x+a, n+b-x)$, or $B(12, 6)$. The prior and posterior distributions of p_0 are plotted in Figure 3.3, where we can see the posterior distribution combines the prior information and the testing information (evidence) and achieves a compromise between the prior distribution and the maximum likelihood estimator.

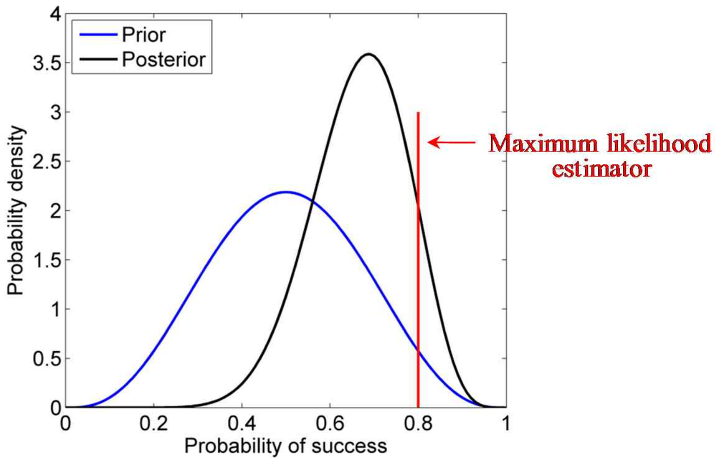


Figure 3.3 Prior and posterior distributions

Homework 11. Matlab coding for Bayesian statistics
Build your own Matlab coding for accomplishing the Example 3.3 (results and figure) above.

In many engineering problems, the conjugacy condition does not hold and explicit solutions cannot be readily obtained with simple mathematical manipulations. In such cases, we can build the posterior distributions by random sampling. A commonly used simulation method for drawing samples from the posterior distribution is referred to as *Markov chain Monte Carlo* (MCMC) in which the two most common techniques, the Metropolis–Hastings algorithm and Gibbs sampling, are used. Others include particle filtering, (extended) Karman filtering, etc. An in-depth theoretical discussion of these techniques is beyond the scope of this book. Readers are recommended to refer to some Bayesian statistics books for detailed information.

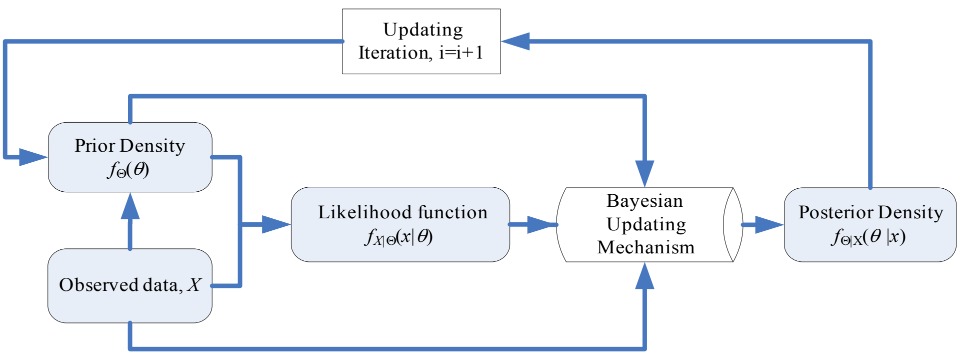


Figure 3.4: Process of Bayesian Updating

The Bayesian approach is used for updating information about the parameter θ. First, a prior distribution of Θ must be assigned before any future observation of X is taken. Then, the prior distribution of Θ is updated to the posterior distribution as the new data for X is employed. The posterior distribution is set to a new prior distribution and this process can be repeated with an evolution of data sets. This updating process can be briefly illustrated in Fig. 3.4.

Markov model is widely used in various fields such as word recognition, voice recognition and gesture recognition in which sequence of the data is very meaningful. Markov chain which consists of Markov model defines probability of posterior event given the prior events. For example, 1st Markov chain considers just last event and 2nd Markov chain take last two events into consideration to calculate probability of the current event, expressed as

$$1^{st}\ Markov\,chain$$
$$: P\left(X_t = x \middle| X_{t-1} = x_{t-1}, X_{t-2} = x_{t-2}, \cdots, X_1 = x_1\right) = P\left(X_t = x \middle| X_{t-1} = x_{t-1}\right)$$

$$2^{nd}\ Markov\,chain$$
$$: P\left(X_t = x \middle| X_{t-1} = x_{t-1}, X_{t-2} = x_{t-2}, \cdots, X_1 = x_1\right) = P\left(X_t = x \middle| X_{t-1} = x_{t-1}, X_{t-2} = x_{t-2}\right)$$

A state diagram for a simple example of the 1st Markov chain is shown in the **Figure 3.5.**

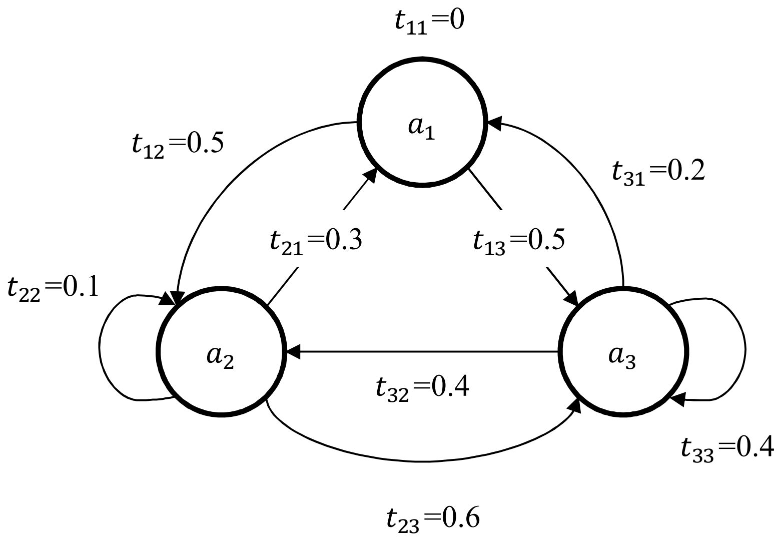


Figure 3.5: State diagram of a simple Markov chain

'a' represents the observations which can be obtained from the model, and 't_{ij}' is probability that a_j occurs when a_i is given. For example, probability that the posterior event X_t becomes a_2 can be defined based on prior events as follows

$$P(X_t = a_2) = P(X_t = a_2 | X_{t-1} = a_1) \times P(X_t = a_2 | X_{t-1} = a_2) \times P(X_t = a_2 | X_{t-1} = a_3)$$
$$= t_{12} \times t_{22} \times t_{32} = 0.5 \times 0.1 \times 0.4 = 0.02$$

For more convenient interpretation of the model, transition matrix can be defined as

$$T = \begin{bmatrix} 0 & 0.5 & 0.5 \\ 0.3 & 0.1 & 0.6 \\ 0.2 & 0.4 & 0.4 \end{bmatrix}$$

It can be noticed that sum of the probability of all posterior events given one prior event is 1.

Example 3.4 (Gambler's ruin)
Suppose that a gambler having \$20 is going to gamble at roulette in a Casino. The gambler bets \$10 on odd number, and makes \$10 when it occurs. If even number occurs, he loses the money betting the roulette.
He has to leave the Casino if he loses his entire money or make \$20 to have \$40 in his pocket. What is likelihood that the gambler lose his entire money from ten times of the roulette game given that probability of winning at each game is 50%?

Solution
First, we have to develop Markov chain to solve the example. 1st Markov chain is used in this example. Graphical model can be illustrated as

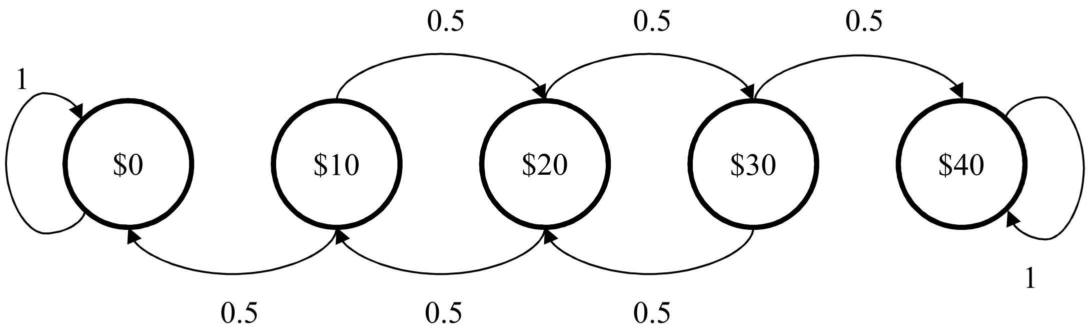


Figure 3.6 Markov chain for the Example 3.4

And the corresponding transition matrix is

$$T = \begin{bmatrix} 1 & 0 & 0 & 0 & 0 \\ 0.5 & 0 & 0.5 & 0 & 0 \\ 0 & 0.5 & 0 & 0.5 & 0 \\ 0 & 0 & 0.5 & 0 & 0.5 \\ 0 & 0 & 0 & 0 & 1 \end{bmatrix}$$

After ten times of roulette games, multiplication of the transition matrix gives

$$T^{10} \approx \begin{bmatrix} 1 & 0 & 0 & 0 & 0 \\ 0.734 & 0.016 & 0 & 0.016 & 0.234 \\ 0.484 & 0 & 0.031 & 0 & 0.484 \\ 0.234 & 0.016 & 0 & 0.016 & 0.734 \\ 0 & 0 & 0 & 0 & 1 \end{bmatrix}$$

What this result is saying is that probability of losing all initial pocket money ($20) as a result of 10 times of roulette games is about 48.4% under the given condition.

The idea of Markov Chain Monte Carlo (MCMC) is basically the same as the Markov model in that it defines posterior position of the sampling point based on the prior information of the sampled points. Two most important techniques can be employed in MCMC, the Metropolis-Hastings algorithm and Gibbs sampling.

Metropolis algorithm, which is the most simplified MCMC method can be performed by the following steps

Step 1. Set a sample index i to 0 and initial sampling point x_0
Step 2. Pick a random value $u \sim U(0,1)$, where u follows the uniform distribution
Step 3. Define a candidate of the next sampling point $x^* \sim P(x^*|x_i)$, where P is
'proposal distribution' → generate 'random walk' using a proposal density
Step 4. If $u < \min\left\{1, \frac{p(x^*)}{p(x^i)}\right\}$

$$x^{i+1} = x^* \quad \rightarrow \text{accept a proposal}$$

else

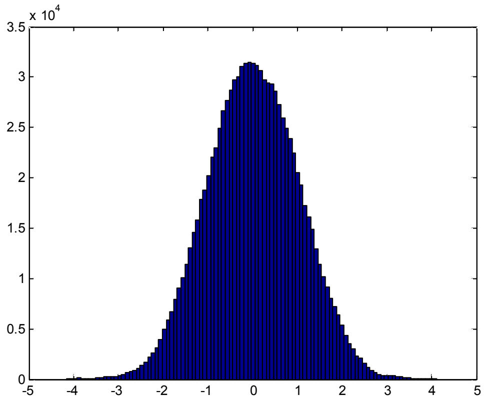
$$x^{i+1} = x^i \quad \rightarrow \text{reject a proposal}$$

end

In step 4, decision criterion is defined based on the ratio of probability of the candidate position and probability of the prior sampling point. Thus, the next position of the sampling point is defined in most likely direction.

For example, it is possible to design the sampling position for the Gaussian distribution with mean of zero and standard deviation of one using the Metropolis algorithm, where 'proposal function' P follows Gaussian distribution (norm(x^i, 0.05)).

```
>> n=1000000;
>> x=zeros(n,1);
>> x0=0.5;                                          % Step 1
>> x(1)=x0;
>> for i=1:n-1
>>  x_star=normrnd(x(i),0.05);                      % Step 2
>>    u=rand;                                       % Step 3
>>    if u<min(1,normpdf(x_star)/normpdf(x(i)))     % Step 4
>>       x(i+1)=x_star;
>>    else
>>       x(i+1)=x(i);
>>    end
>> end
>>
>> figure;
>> hist(x,100);
```



Others include particle filtering, (extended) Karman filtering, etc. An in-depth theoretical discussion of these techniques is beyond the scope of this book. Readers are recommended to refer to some Bayesian statistics books for detailed information.

3.2.4 How to Model Prior Distribution?

- Informative Prior Distribution
  Generally we have two ways to handle known information (**x**):
  1. Histogram
  2. Select a prior density function with unknown parameters firstly, and then estimate the unknown parameters for the data.

- Non-informative Prior Distribution
  Non-informative prior distribution means determining the prior distribution when no other information about the parameter θ is available except its feasible field Θ.

References for Bayesian statistics:

1. http://en.wikipedia.org/wiki/Bayesian
2. Singpurwalla, N.D., 2006, Reliability and Risk: A Bayesian Perspective, Wiley.
2. Andrew Gelman, John B. Carlin, Hal S. Stern and Donald B. Rubin, 2004, Bayesian Data Analysis, Second Edition, Chapman & Hall/CRC.
3. Bernardo, J.M., and Smith A.F.M., 1994, *Bayesian Theory*, John Wiley & Son Ltd..

