Lecture 5:
Fundamentals of Statistical Analysis
and
Distributions Derived from Normal Distributions

ELE 525: Random Processes in Information Systems

Hisashi Kobayashi

Department of Electrical Engineering
Princeton University
September 27, 2013

6 Fundamentals of Statistical Data Analysis

6.1 Sample mean and sample variance

The sample mean (or the empirical average) is defined as

\[ \bar{x} = \frac{1}{n} \sum_{i=1}^{n} x_i. \]  

(6.1)

Each sample \( x_i \) is an instance or realization of the associated RV \( X_i \). The sample mean of (6.1) is an instance of the sample mean variable defined by

\[ \overline{X} = \frac{1}{n} \sum_{i=1}^{n} X_i. \]  

(6.2)

- The expectation is

\[ E[\overline{X}] = \mu_X, \]  

(6.4)

- The variance is

\[ \text{Var} [\overline{X}] = \frac{\sigma^2_X}{n}. \]  

(6.8)
The sample variance is defined by

\[ s_x^2 \triangleq \frac{1}{n-1} \sum_{i=1}^{n} (x_i - \bar{x})^2. \]  

which can be viewed as an instance of the sample variance variable

\[ S_x^2 \triangleq \frac{1}{n-1} \sum_{i=1}^{n} (X_i - \bar{X})^2, \]  

which is also often called the sample variance. We can show

\[ E[S_x^2] = \frac{1}{n} \sum_{i=1}^{n} E[Y_i^2] = \sigma_x^2. \]  

- Equations (6.4) and (6.12) show that the sample mean variable of (6.2) and the sample variance variable (6.10) are unbiased estimates of the (population) mean \( \mu_X \) and the (population) variance \( \sigma_X^2 \), respectively.

- The square root of the sample variance (6.9), i.e., \( s_x \), is called the sample standard deviation.
6.2 Relative frequency and histograms

Consider observed data of sample size $n$, and they take on $k$ distinct discrete values.

Let

$$n_j = \text{number of times that the } j\text{th value is observed, } j=1, 2, ..., k.$$ 

Then

$$f_j = \frac{n_j}{n}, \quad j = 1, 2, ..., k, \tag{6.13}$$

is called the relative frequency of the $j$th value.

When the underlying RV $X$ is continuous, we group or classify the data.

Divide the range of observations into $k$ class intervals, at points $c_0, c_1, c_2, ..., c_k$.

$$\Delta_j \triangleq c_j - c_{j-1}, \quad j = 1, 2, ..., k$$

$$h(x) = \frac{f_j}{\Delta_j} = \frac{n_j}{n \Delta_j}, \quad \text{for } x \in (c_{j-1}, c_j], \quad j = 1, 2, ..., k. \tag{6.14}$$

is called a histogram, and is an estimate of the PDF of the population.
Cumulative relative frequency

Let \( \{x_k: 1 \leq k \leq n\} \) be \( n \) observations in the order observed, and \( \{x_{(i)}: 1 \leq i \leq n\} \) be the same observations in order of magnitude. \( H(x) \) be the frequency of observations that are smaller than or equal to \( x \):

\[
H(x) = \begin{cases} 
0, & \text{for } x < x_{(1)} \\
\frac{i}{n}, & \text{for } x_{(i)} \leq x < x_{(i+1)}, \quad i = 1, 2, \ldots, n - 1 \\
1, & \text{for } x \geq x_{(n)},
\end{cases}
\]  

which can be more concisely written as

\[
H(x) = \frac{1}{n} \sum_{i=1}^{n} u(x - x_{(i)}) = \frac{1}{n} \sum_{k=1}^{n} u(x - x_k), \quad -\infty < x < \infty. \tag{6.17}
\]

When grouped data are presented as a cumulative relative frequency distribution, it is called the cumulative histogram.

The cumulative histogram is far less sensitive to variation in class lengths than the histogram.
6.3 Graphical presentations
6.3.1 Histogram on probability paper
6.3.1.1 Testing the normal distribution hypothesis

For a given distribution function $F(x)$, let

$$P = F(x) \quad (6.18)$$

The inverse

$$x_P = F^{-1}(P) \quad (6.19)$$

is the value of $x$ that corresponds to the cumulative probability $P$.

- $x_P$ is called the **P-fractile** (or P-percentile or P-quantile).

- Consider the **standard normal distribution**

$$\Phi(u) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{u} \exp\left(-\frac{t^2}{2}\right) dt. \quad (6.20)$$

The fractile $u_P$ of the distribution $N(0,1)$ is

$$u_P = \Phi^{-1}(P). \quad (6.21)$$
For a given cumulative relative frequency $H(x)$, we wish to test whether

$$H(x) \approx \Phi \left( \frac{x - \mu}{\sigma} \right)$$

(6.22)

holds for some $\mu$ and $\sigma$. Testing the above is equivalent to testing the relation

$$u_H(x) \approx \frac{x - \mu}{\sigma}$$

(6.23)

The plot of $u_H(x)$ versus $x$ forms a step (or staircase) curve:

$$u_H(x) = \begin{cases} -\infty, & \text{for } x < x(1) \\ \frac{i}{n}, & \text{for } x(i) \leq x < x(i+1), \quad i = 1, 2, \ldots, n-1, \\ \infty, & \text{for } x \geq x(n). \end{cases}$$

(6.24)

The plot in the $(x, u)$-coordinates of the staircase function

$$u = u_H(x)$$

(6.25)

is called the fractile diagram, and provides an estimate of the straight line

$$u = \frac{x - \mu}{\sigma}$$

(6.26)
The **probability paper**

On the ordinate axis, the values $P = \Phi(u)$ are marked, rather than the $u$ values.
The **dot diagram**: Instead of the step curve, we plot $n$ points $(x_{(i)}, (i-\frac{1}{2})/n)$, which are situated at the middle points.

![The fractile diagram of normal variates: (b) dot diagram. (n=50)](image)
6.3.1.2 Testing the log-normal distribution hypothesis

The log-normal paper: Modify the probability paper by changing the horizontal axis from the linear scale to the logarithmic scale, i.e., $\log_{10} x$.

$n=50, x_i= \exp y_i$ where $y_i$ is drawn from $N(2,4)$.
6.3.2 Log-survivor function curve

- The **survivor function** or the **survival function**: 
  \[ S_X(t) \triangleq P[X > t] = 1 - F_X(t) \]  
  \(^{(6.27)}\)

- The **log-survivor function** or the **log survival function**: 
  \[ \log S_X(t) = \log (1 - F_X(t)). \]  
  \(^{(6.28)}\)

- The **sample log-survivor function** or **empirical log-survivor function**: 
  \[ \log [1 - H(t)], \]  
  \(^{(6.31)}\)

where \(H(x)\) is the **cumulative relative frequency** (for ungrouped data) or the **cumulative histogram** (for grouped data).

✓ For the ungrouped case: Plot 
  \[ \log \left(1 - \frac{i}{n}\right), \quad 1 \leq i \leq n \]  
  \(^{(6.32)}\)

against \(x_{(i)}\)
In order to avoid difficulty at $i=n$, we may modify (6.32) into

$$
\log \left( 1 - \frac{i}{n+1} \right), \quad 1 \leq i \leq n. \tag{6.33}
$$

**Example:** A mixed exponential (or hyperexponential) distribution:

$$
F_X(x) = \pi_1 (1 - e^{-\alpha_1 x}) + \pi_2 (1 - e^{-\alpha_2 x}), \quad \alpha_1 > \alpha_2, \quad \pi_1 + \pi_2 = 1, \tag{6.29}
$$

$$
\log S_X(t) = \log \left[ \pi_1 e^{-\alpha_1 t} + \pi_2 e^{-\alpha_2 t} \right]
\approx \begin{cases} 
-\alpha_1 t + \log \pi_1, & \text{for small } t \\
-\alpha_2 t + \log \pi_2, & \text{for large } t. 
\end{cases} \tag{6.30}
$$

**Numerical example:**

$\pi_2 = 0.0526, \quad \pi_1 = 1 - \pi_2, \quad \alpha_2 = 0.1 \text{ and } \alpha_1 = 2.0$

**Note:** To be consistent with the assumption in (6.29), we should exchange the subscripts 1 and 2.
Correction to the figure caption: Exchange the subscripts 1 and 2 of $\pi$ and $\alpha$ to be consistent with (6.29) and (6.30)

Figure 6.3. The log-survivor function of a mixed-exponential (or hyper-exponential) distribution with $\pi_1 = 0.0526$, $\pi_2 = 1 - \pi_1$, $\alpha_1 = 0.1$, and $\alpha_2 = 2.0$. 
6.3.3 Hazard function and mean residual life curves

- The hazard function or the failure rate:

\[
h_X(t) = \frac{f_X(t)}{S_X(t)} = \frac{f_X(t)}{1 - F_X(t)},
\]

which is called the completion rate function, when \( X \) represents a service time variable.

- The survivor function and the hazard function are related by

\[
S_X(x) = e^{-\int_0^x h_X(t) \, dt}, \quad x \geq 0,
\]

and

\[
h_X(t) = -\frac{d \log S_X(t)}{dt}, \quad t \geq 0.
\]
Given that the service time variable $X$ is greater than $t$,

$$R = X - t$$

is the residual life conditioned on $X > t$.

The mean residual life function

$$R_X(t) = E[R|X > t] = \frac{\int_t^\infty S_X(u) \, du}{S_X(t)}.$$  \hspace{1cm} (6.41)

$$R_X(0) = \int_0^\infty S_X(u) \, du = E[X],$$  \hspace{1cm} (6.42)
Figure 6.5 The mean residual life curves of a Pareto distribution with $\alpha = 3.0$ and $\beta = 1.0$, and a Weibull distribution with $\alpha = 1.5$ and $\beta = 1.0$. 
6.3.4 Dot diagram and correlation coefficient

Dot or scatter diagram:

\[(x_i, y_i); 1 \leq i \leq n\]  

*Figure 6.6* Scatter diagram of delay measurements from Internet host at Stanford University to 79 other hosts across the U.S. [363]
Correlation coefficient

\[ \rho_{XY} = \frac{\sigma_{XY}}{\sigma_X \sigma_Y}. \]  

(6.46)

where

\[ \sigma_{XY} \triangleq \text{Cov}[X, Y] = E[(X - \mu_X)(Y - \mu_Y)] = E[XY] - \mu_X \mu_Y. \]  

(6.45)

X and Y are said to be **properly linearly correlated** if

\[ P[aX - bY = c] = 1. \]  

(6.48)

\[ \text{Var } [aX - bY - c] = 0, \]  

(6.49)

\[ \rho_{XY} = +1 \text{ or } -1 \]  

(6.50)

depending on whether \( ab \) is positive or negative.
Conversely, if \( \rho = \pm 1 \), then (Problem 6.17)

\[
P \left[ \pm \frac{(X - \mu_X)}{\sigma_X} + \frac{Y - \mu_Y}{\sigma_Y} = 0 \right] = 1. \quad (6.51)
\]

- The **sample variance** based on observations \( \{(x_i, y_i) : 1 \leq i \leq n\} \)

\[
s_{xy} = \frac{1}{n-1} \sum_{i=1}^{n} (x_i - \overline{x})(y_i - \overline{y})
\]

\[
= \frac{1}{n-1} \sum_{i=1}^{n} x_i y_i - \frac{n \overline{xy}}{n-1}, \quad (6.52)
\]

- The sample correlation coefficient

\[
R_{xy} = \frac{s_{xy}}{s_x s_y}, \quad (6.53)
\]
7 Distributions Derived from the Normal Distribution

7.1 Chi-Squared Distribution

Let $U_i, 1 \leq i \leq n$ be $n$ i.i.d RVs with the standard normal distribution $N(0,1)$. Define

$$\chi^2_n = \sum_{i=1}^{n} U_i^2.$$  \hfill (7.1)

The PDF of this RV (Problem 7.2)

$$f_{\chi^2_n}(x) = \frac{x^{(n/2)-1}e^{-x/2}}{2^{n/2}\Gamma(n/2)}dx, \quad 0 \leq x < \infty,$$  \hfill (7.2)

$$\Gamma(z) = \int_{0}^{\infty} t^{z-1}e^{-t}dt.$$  \hfill (7.3)

$$\Gamma(1) = 1, \quad \Gamma\left(\frac{1}{2}\right) = \sqrt{\pi}$$  \hfill (7.4)
\[ \Gamma \left( \frac{n}{2} \right) = \begin{cases} \left( \frac{n}{2} - 1 \right)!, & \text{for } n \text{ even} \\ \left( \frac{n}{2} - 1 \right) \left( \frac{n}{2} - 2 \right) \cdots \frac{3}{2} \cdot \frac{1}{2} \sqrt{\pi}, & \text{for } n \text{ odd.} \end{cases} \] (7.5)

Figure 7.1 The \( \chi^2_n \) distribution with degree of freedom \( n \).
\[ f_{\chi_1^2}(x) = \frac{x^{-1/2}e^{-x/2}}{\sqrt{2\pi}}, \quad x > 0. \]  \hspace{1cm} (7.6)

\[ f_{\chi_2^2}(x) = \frac{e^{-x/2}}{2}, \quad x \geq 0 \]  \hspace{1cm} (7.7)

\[ f_{\chi_3^2}(x) = \frac{x^{1/2}e^{-x/2}}{\sqrt{2\pi}}, \quad x \geq 0. \]  \hspace{1cm} (7.8)

\[ E[\chi_n^2] = n, \]  \hspace{1cm} (7.9)

\[ \text{Var}[\chi_n^2] = 2n. \]  \hspace{1cm} (7.10)

Mode: \( n-2 \)
The relations to other distributions:

- Let

\[ Y_n = \frac{\chi_n^2}{2} \]  

\[ f_{Y_n}(y) = \frac{y^{(n/2)-1}e^{-y}}{\Gamma \left( \frac{n}{2} \right)}, \]  

which is a special case \( \lambda = 1, \beta = n/2 \) in the **gamma distribution** (4.30)

\[ f(y) = \begin{cases} \frac{e^{-\lambda y}}{\Gamma(\beta)} \lambda(y)^{\beta-1}, & y \geq 0 \\ 0, & y < 0\end{cases}, \]  

- The case where \( n \) is an even integer:

\[ n = 2k, \]  

\[ f_{Y_{2k}}(y) = \frac{y^{k-1}e^{-y}}{(k-1)!}, \]  

which is the \( k \)-stage **Erlang distribution** with mean \( k \).
The relation to the \textbf{Poisson distribution}

\[
P[\chi^2_{2k} > 2\lambda] = \int_{\lambda}^{\infty} \frac{y^{k-1}e^{-y}}{(k-1)!}dy
\]
\[= \int_{\lambda}^{\infty} P(k-1; y)dy = Q(k-1; \lambda), \quad (7.17)
\]

\textbf{Example 7.1:} Independent observations from \(N(\mu, \sigma^2)\)

\textbullet\ Case 1: An estimate of \(\sigma^2\), when the population mean \(\mu\) is known

\[
\bar{s}^2 = \frac{1}{n} \sum_{i=1}^{n} (X_i - \mu)^2. \quad (7.18)
\]
\[
\sum_{i=1}^{n} U_i^2, \quad (7.19)
\]

where

\[
U_i = \frac{X_i - \mu}{\sigma}, \quad 1 \leq i \leq n, \quad (7.20)
\]

Thus, we can write

\[
\bar{s}^2 = \frac{\sigma^2}{n} \chi^2_n. \quad (7.21)
\]
An estimate of $\sigma^2$ when $\mu$ is unknown.

$$s^2 = \frac{1}{n-1} \sum_{i=1}^{n} (X_i - \bar{X})^2.$$  \hfill (7.22)

We can show (Problem 7.1)

$$s^2 = \frac{\sigma^2}{n-1} \chi_{n-1}^2.$$  \hfill (7.26)

Karl Pearson (1857-1936) was a British statistician who applied statistics to biological problems of heredity and evolution.
7.2 Student’s $t$-Distribution

The sample mean $\bar{X}$ of $n$ independent observations $\{X_1, X_2, \ldots, X_n\}$ from $N(\mu, \sigma^2)$ is normally distributed according to $N(\mu, \sigma^2/n)$.

Thus,

$$U = \frac{(\bar{X} - \mu)\sqrt{n}}{\sigma}$$  \hspace{1cm} (7.30)

is a standard normal variable.

We wish to estimate the population mean $\mu$.

- If $\sigma$ is known, we can use the table of the standard normal distribution to test whether $U$ is significantly different from 0.

- If $\sigma$ is unknown, we use

$$t_{n-1} = \frac{(\bar{X} - \mu)\sqrt{n}}{s}.$$  \hspace{1cm} (7.31)

Using (7.26)

$$t_{n-1} = \frac{(\bar{X} - \mu)\sqrt{n}/\sigma}{s/\sigma} = \frac{U}{\sqrt{\chi^2_{n-1}/(n-1)}}.$$  \hspace{1cm} (7.32)
The distribution of the variable $t_k$ is called the (Student’s) $t$-distribution with $k$ degrees of freedom (d.f.).

Its PDF is given by (Problem 7.6)

$$f_{t_k}(t) = \frac{\Gamma\left(\frac{k+1}{2}\right)}{\Gamma\left(\frac{k}{2}\right)\sqrt{\pi k}} \left(1 + \frac{t^2}{k}\right)^{-\frac{k+1}{2}}, \quad -\infty < t < \infty. \quad (7.33)$$

$k=1$,

$$f_{t_1}(t) = \frac{1}{\pi(1+t^2)}, \quad (7.34)$$

which is called the Cauchy’s distribution.

$k=2$,

$$f_{t_2}(t) = (2 + t^2)^{-3/2}, \quad (7.35)$$

which has zero mean but infinite variance.
Figure 7.2 The student's $t$-distribution with $k$ degrees of freedom ($k = 1, 2, 5, \infty$).
William S. Gosset (1876-1937) was a statistician of the Guinness brewing company.

\[
E[t_{k}^{2r}] = E[(x_{1}^{2})^{r}]E \left[ \left( \frac{x_{k}^{2}}{k} \right)^{-r} \right],
\]

\[
E[t_{k}^{2r}] = \frac{k^{r} \Gamma \left( \frac{1}{2} + r \right) \Gamma \left( \frac{k}{2} - r \right)}{\Gamma \left( \frac{1}{2} \right) \Gamma \left( \frac{k}{2} \right)},
\]

\[
E[t_{k}] = 0, \quad \text{Var} \ [t_{k}] = \frac{k}{k - 2}.
\]
7.3 Fisher’s \( F \)-distribution

RVs \( V_1 \) and \( V_2 \) are independent and are \( \chi^2 \) distributed with \( n_1 \) and \( n_2 \) degrees of freedom (d.f.), respectively. Then the variable \( F \) defined by

\[
F = \frac{V_1/n_1}{V_2/n_2}.
\]

has the following PDF:

\[
f_F(x) = \frac{\Gamma\left(\frac{n_1+n_2}{2}\right)\left(\frac{n_1}{n_2}\right)^{n_1/2}}{\Gamma\left(\frac{n_1}{2}\right)\Gamma\left(\frac{n_2}{2}\right)} x^{(n_1/2)-1}\left(1 + \frac{n_1x}{n_2}\right)^{-(n_1+n_2)/2}
\]

which is called the \textbf{F-distribution with} \( (n_1, n_2) \), also called the \textbf{Snedecor distribution}.

\[
E[F^r] = \left(\frac{n_2}{n_1}\right)^r \frac{\Gamma\left(\frac{n_1}{2} + r\right)\Gamma\left(\frac{n_2}{2} - r\right)}{\Gamma\left(\frac{n_1}{2}\right)\Gamma\left(\frac{n_2}{2}\right)}
\]

which exists for \( -n_1 < 2r < n_2 \).
Figure 7.3 The $F$-distributions for various degrees of freedom $(n_1, n_2)$.

$$E[F] = \frac{n_2}{n_2 - 2} \text{ for } n_2 > 2$$ \hspace{1cm} (7.42)

$$\text{Var} [F] = \frac{2n_2^3(n_1 + n_2 - 2)}{n_1(n_2 - 2)^2(n_2 - 4)} \text{ for } n_2 > 4.$$ \hspace{1cm} (7.43)

$$\text{mode } F = \frac{n_2(n_1 - 2)}{n_1(n_2 + 1)}. \hspace{1cm} (7.44)$$
7.4 Log-normal distribution

A positive RV $X$ is said to have the log-normal distribution if

$$Y = \ln X$$

Is normally distributed, i.e.,

$$f_Y(y) = \frac{1}{\sqrt{2\pi}\sigma_Y} \exp \left\{ -\frac{(y - \mu_Y)^2}{2\sigma_Y^2} \right\}, \quad -\infty < y < \infty. \quad (7.46)$$

Then, by using $dy = \frac{dx}{x}$, and $f_Y(y) \, dy = f_X(x) \, dx$, we readily find

$$f_X(x) = \frac{1}{\sqrt{2\pi}\sigma_Y x} \exp \left\{ -\frac{(\ln x - \mu_Y)^2}{2\sigma_Y^2} \right\}, \quad x > 0. \quad (7.47)$$

In order to find the expectation and variance, we use the moment generating function (MGF) (to be studied in Section 8.1)

$$M_Y(t) = E[e^{tY}] = \exp \left\{ \mu_Y t + \frac{\sigma_Y^2 t^2}{2} \right\}. \quad (7.48)$$

$$\mu_X = E[X] = E[e^Y] = M_Y(1) = \exp \left\{ \mu_Y + \frac{\sigma_Y^2}{2} \right\}. \quad (7.49)$$

$$E[X^2] = E[e^{2Y}] = M_Y(2) = \exp \left\{ 2\mu_Y + 2\sigma_Y^2 \right\} = \mu_X^2 e^{\sigma_Y^2}. \quad (7.50)$$
Then

\[
\sigma_X^2 = \mu_X^2 \left( \exp \{ \sigma_Y^2 \} - 1 \right). \quad (7.51)
\]

From (7.49) and (7.51) we find

\[
\mu_Y = \ln \mu_X - \frac{1}{2} \ln \left( 1 + \frac{\sigma_X^2}{\mu_X^2} \right), \quad (7.52)
\]

\[
\sigma_Y^2 = \ln \left( 1 + \frac{\sigma_X^2}{\mu_X^2} \right). \quad (7.53)
\]