# Lecture 8: Principles of Data Reduction

02 Feb 2016

The previous lecture dealt with some properties of random samples. In the event when the samples were picked IID from a normal distribution, the sample mean and the sample variance were shown to be independent random variables. Also, a general expression for the cumulative distribution function (CDF) of the $j^{\text{th}}$ order statistic was derived in terms of the CDF governing the population. We begin this lecture by deriving the CDF of the $j^{\text{th}}$ order statistic for a uniformly distributed population. Next, we discuss the Student's $t$ distribution. We then study the principle of sufficiency, and derive sufficient statistics for a parameterized population governed by (a) binomial/Bernoulli distribution, and (b) normal distribution.

**Example 0.1 (Order Statistics of Uniform Distribution).** Let $X_1, \ldots, X_n \overset{IID}{\sim} \text{unif}\,[0,1]$. Then, for all $i = 1, \ldots, n$, $F_{X_i}(x) = x$, $x \in [0,1]$, where $F(.)$ denotes the CDF. We now derive the CDF of the $j^{th}$ order statistic. From Lecture 7, we know that for any $x \in [0,1]$, we have $F_{X_{(j)}}(x) = \sum_{k=j}^{n} \binom{n}{k} x^k (1-x)^{n-k}$. So,

$$
\begin{aligned}
f_{X_{(j)}}(x) &= \frac{d}{dx} F_{X_{(j)}}(x), \\
&= \sum_{k=j}^{n} \binom{n}{k} \{k x^{k-1}(1-x)^{n-k} - x^k (1-x)^{n-k-1}(n-k)\}, \\
&= \sum_{k=j+1}^{n} \binom{n}{k} k x^{k-1}(1-x)^{n-k} - \sum_{k=j}^{n-1} \binom{n}{k} x^k (1-x)^{n-k-1}(n-k) \\
&\qquad\qquad + \binom{n}{j} j x^{j-1}(1-x)^{n-j}, \\
&= \frac{n!}{(j-1)!\,(n-j)!} x^{j-1}(1-x)^{n-j}, \\
&= \frac{\Gamma(n+1)}{\Gamma(j)\Gamma(n-j+1)} x^{j-1}(1-x)^{(n-j+1)-1}, \qquad (1)
\end{aligned}
$$

where the first equality follows from the relation between the CDF and probability density function (pdf), the second equality follows from the chain rule of the differentiation, and the fourth equality follows by using the following expression:

$$\binom{n}{k+1}(k+1) = \frac{n!}{(k)!\,(n-k-1)!} = \binom{n}{k}(n-k).$$

**Definition 0.2 (Beta Distribution).** For $x \in [0,1]$, and shape parameters $\alpha, \beta > 0$, the Beta distribution is a power function of the variable $x$ and of its reflection $(1-x)$. It's pdf is defined as follows:

$$\begin{aligned} f_{\text{Beta}(\alpha,\beta)}(x) &= \frac{x^{\alpha-1}(1-x)^{\beta-1}}{\int_0^1 u^{\alpha-1}(1-u)^{\beta-1}\,du} \\ &= \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)}\,x^{\alpha-1}(1-x)^{\beta-1} \\ &= \frac{1}{B(\alpha,\beta)}x^{\alpha-1}(1-x)^{\beta-1}, \end{aligned}$$

where $\Gamma(.)$ denotes the gamma function, and the beta function $B(\alpha,\beta)$ is a normalization factor to ensure that the pdf integrates to 1.

Thus, from (1), and the definition of the Beta distribution, the $j^{\text{th}}$ order statistic of a unif $[0,1]$ random sample has a Beta$(j, n-j+1)$ distribution. The mean and variance of the $j^{\text{th}}$ order statistic are as follows:

$$\mathbb{E}\left[X_{(j)}\right] = \frac{j}{n+1} \text{ and } \text{Var}\left[X_{(j)}\right] = \frac{j(n-j+1)}{(n+2)(n+1)^2}.$$

# 1 The Student's $t$ Distribution

Consider a random sample $X_1, \ldots, X_n$, with $X_i \sim \mathcal{N}(\mu, \sigma^2)$, where $\sigma^2$ is known but $\mu$ is unknown. Consider the problem of finding $\mu$. With the knowledge of sample values, the statistic

$$T_1 = \frac{\bar{X} - \mu}{\sigma/\sqrt{n}} \tag{2}$$

has only $\mu$ as the unknown quantity. Here, $\bar{X} = \frac{1}{n}\sum_{i=1}^{n} X_i$ denotes the sample mean. Since $T_1 \sim \mathcal{N}(0,1)$, an estimate of $\mu$ can be obtained as follows: let $\alpha \in (0,1)$ be fixed, and let $c > 0$ be a number such that

$$\mathbb{P}\left(-c \le T_1 \le c\right) = \alpha. \tag{3}$$

Equivalently, we have

$$\mathbb{P}\left(-c \leq \frac{\bar{X}-\mu}{\sigma/\sqrt{n}} \leq c\right) = \alpha, \tag{4}$$

which may be rearranged to obtain a probabilistic estimate for $\mu$ as follows:

$$\mathbb{P}\left(\mu \in \left[\bar{X} - \frac{c\sigma}{\sqrt{n}}, \bar{X} + \frac{c\sigma}{\sqrt{n}}\right]\right) = \alpha. \tag{5}$$

For example, if $\alpha = 0.99$, (5) provides us with a range of values in which $\mu$ lies with probability 0.99.

If both $\sigma^2$ and $\mu$ are unknown, the approach outlined in the preceding paragraph cannot be used. Instead, however, the sample variance $S^2 = \frac{1}{n-1}\sum_{i=1}^{n}\left(X_i - \bar{X}\right)^2$ can be evaluated with the knowledge of sample values $X_1, \ldots, X_n$, and the quantity $S = \sqrt{S^2}$ can be used in place of $\sigma$ in (4). But, in order to do so, the distribution of the statistic $T_2 = \frac{\bar{X}-\mu}{\sqrt{S^2}/\sqrt{n}}$ has to necessarily be known.

If $X_1, \ldots, X_n \overset{IID}{\sim} \mathcal{N}(\mu, \sigma^2)$, then we know that $\frac{\bar{X}-\mu}{\sqrt{\sigma}/\sqrt{n}} \sim \mathcal{N}(0,1)$ and $\frac{(n-1)S^2}{\sigma^2} \sim \chi^2_{n-1}$ (see Lecture 7), where $\chi^2_{n-1}$ denotes the ch-squared distribution with $n-1$ degrees of freedom. Thus, the expression for $T_2$ can be modified as follows:

$$\begin{aligned} T_2 &= \frac{\bar{X}-\mu}{\sqrt{S^2}/\sqrt{n}} \\ &= \frac{\left(\bar{X}-\mu\right)/\left(\sigma/\sqrt{n}\right)}{\sqrt{S^2/\sigma^2}}. \end{aligned} \tag{6}$$

As pointed out before, the numerator of (6) is an $\mathcal{N}(0,1)$ random variable and the denominator is a $\sqrt{\frac{1}{n-1}\chi^2_{n-1}}$ random variable, independent of the denominator (since $\bar{X}$ and $S$ are independent random variables). Thus, the distribution of $T_2$ can be found by solving the simplified problem of finding the distribution of $U/\sqrt{V/p}$, where $U$ is $\mathcal{N}(0,1)$ and $V$ is $\chi^2_p$, $p = n-1$, and $U$ and $V$ are independent. This gives us Student's $t$ distribution.

**Definition 1.1.** Let $X_1, \ldots, X_n$ be a random sample from $\mathcal{N}(\mu, \sigma^2)$ distribution. The quantity $\frac{\bar{X}-\mu}{\sqrt{S^2}/\sqrt{n}}$ has Student's $t$ distribution with $n-1$ degrees of freedom. Equivalently, a random variable $T$ has Student's $t$ distribution with $p$ degrees of freedom, and we write $T \sim t_p$, if it has the following pdf:

$$f_T(t) = \frac{\Gamma\left(\frac{p+1}{2}\right)}{\Gamma\left(\frac{p}{2}\right)} \frac{1}{\sqrt{p\pi}} \frac{1}{\left(1+\frac{t^2}{p}\right)^{\frac{p+1}{2}}}, \ -\infty < t < \infty. \tag{7}$$

When $p = 1$, (7) becomes the pdf of Cauchy distribution. In order to show that the statistic $T_2$ has this distribution, let $U$ and $V$ be random variables as defined in the previous paragraph, and let $p = n - 1$. Then, the joint pdf of $U$ and $V$ is given by:

$$\begin{aligned} f_{U,V}(u,v) &= f_U(u) f_V(v) \text{ (since } U \text{ and } V \text{ are independent)}, \\ &= \left(\frac{1}{\sqrt{2\pi}} e^{-\frac{u^2}{2}}\right)\left(\frac{1}{\Gamma(p)}\frac{1}{2^{\frac{p}{2}}} v^{\frac{p}{2}-1} e^{-\frac{v}{2}}\right), \ -\infty < u < \infty, \ 0 < v < \infty. \end{aligned} \quad (8)$$

We now apply the transformation $a = \frac{u}{\sqrt{v/p}}$ and $b = v$. Then, the Jacobian matrix of the transformation is given by:

$$\begin{aligned} J(u,v) &= \begin{bmatrix} \frac{\partial a}{\partial u} & \frac{\partial a}{\partial v} \\ \\ \frac{\partial b}{\partial u} & \frac{\partial b}{\partial v} \end{bmatrix}, \\ &= \begin{bmatrix} \sqrt{\frac{p}{v}} & -\frac{u}{2}\sqrt{\frac{p}{v}} \\ \\ 0 & 1 \end{bmatrix}. \end{aligned} \quad (9)$$

Thus, the joint distribution of $A = \frac{U}{\sqrt{V/p}}$ and $B = V$ is given by:

$$\begin{aligned} f_{A,B}(a,b) &= \frac{f_{U,V}(u,v)}{|\det(J(u,v))|}, \\ &= f_{U,V}(u,v)\sqrt{\frac{v}{p}}, \end{aligned} \quad (10)$$

where $\det(J(u,v))$ denotes the determinant of the Jacobian matrix, which evaluates to $\sqrt{\frac{p}{v}}$. We note that the random variable $A$ corresponds to the statistic $T_2$ whose pdf is to be computed. Thus, we have

$$\begin{aligned} f_{T_2}(t) = f_A(t) &= \int_0^\infty f_{A,B}(t,b)\ db, \\ &= \int_0^\infty f_{U,V}\left(t\sqrt{\frac{v}{p}}, v\right)\sqrt{\frac{v}{p}}\ dv, \end{aligned} \quad (11)$$

which upon simplification yields the expression on the RHS of (7), with $p = n - 1$.

**An Application of Student's $t$ Distribution:** We now return to the problem of finding the mean $\mu$ of a population governed by $\mathcal{N}(\mu, \sigma^2)$ distribution, given

the values of the random sample $X_1, \ldots, X_n$, when both $\mu$ and $\sigma^2$ are unknown. Let $\alpha \in (0,1)$ be fixed, and let $c > 0$ be a number such that

$$\mathbb{P}\left(-c \leq T_2 \leq c\right) = \alpha. \tag{12}$$

Equivalently, we have

$$\mathbb{P}\left(-c \leq \frac{\bar{X} - \mu}{\sqrt{S^2}/\sqrt{n}} \leq c\right) = \alpha. \tag{13}$$

Since $\frac{\bar{X}-\mu}{\sqrt{S^2}/\sqrt{n}}$ follows Student's $t$ distribution with $n-1$ degrees of freedom, (13) can be solved to obtain the value of $c$, and the following equation then provides the range of values in which $\mu$ lies with probability $\alpha$:

$$\mathbb{P}\left(\mu \in \left[\bar{X} - \frac{c\sqrt{S^2}}{\sqrt{n}}, \bar{X} + \frac{c\sqrt{S^2}}{\sqrt{n}}\right]\right) = \alpha. \tag{14}$$

# 2 Data Reduction / Principle of Sufficiency.

(Reference: Chapter 6 of *Statistical Inference* by George Casella and Roger L. Berger).

A statistician uses the information in a sample $x_1, \ldots, x_n$ provided by a data collector to draw inferences about an unknown parameter $\theta$ of the distribution governing the population. If the sample size $n$ is large, then the observed sample may be cumbersome to deal with.

So, from the statistician's point of view, it is desirable to summarize the information in a sample by determining its key features. This is usually done by computing statistics, which are functions of the sample. Let $\underline{X} \triangleq \{X_1, \ldots, X_n\}$ denote a random sample, and $\underline{x} \triangleq \{x_1, \ldots, x_n\}$ denote the sample values (or simply, sample). We wish to construct functions (or statistics) $T(\underline{X})$ that are "sufficient" for the purpose of determining the parameter $\theta$. Such functions are called sufficient statistics.

**Definition 2.1 (Sufficient Statistics (Informal)).** A statistic $T(\underline{X})$ is called a sufficient statistic for a population $F$ with parameter $\theta$ if $T(\underline{X})$ captures all the information about $\theta$.

**Definition 2.2 (Sufficient Statistics (Formal)).** A statistic $T(\underline{X})$ is said to be sufficient for a parameterized population $F_\theta$ if the conditional distribution of the random sample $\underline{X}$ given $T(\underline{X})$ does not depend on $\theta$.

**Example 2.3 (The Trivial Sufficient Statistic).** We now show that $T(\underline{X}) = \underline{X}$ is trivially a sufficient statistic for the parameter $\theta$. We have,

$$\mathbb{P}\left(\underline{X} = \underline{x} | \; T\left(\underline{X}\right) = (y_1, \ldots, y_n)\right) = \mathbb{1}_{(x_i = y_i) \; \forall i}. \tag{15}$$

Since (15) is independent of $\theta$, we conclude that $T\left(\underline{X}\right) = \underline{X}$ is a sufficient statistic for the parameter $\theta$.

**Example 2.4 (Bernoulli/Binomial sufficient statistic).** Consider $n$ IID random variables $X_1, \ldots, X_n$ distributed according to a Bernoulli distribution with bias parameter $\theta \in [0, 1]$, i.e., $X_i \in \{0, 1\} \forall i = 1, \ldots, n$. We now wish to see if $T\left(\underline{X}\right) = \sum_{i=1}^{n} X_i$ a sufficient statistic? For any $k \in \mathbb{Z}^+$, where $\mathbb{Z}^+$ denotes the set of all positive integers, we have

$$\mathbb{P}\left(\underline{X} = \underline{x} | \; T\left(\underline{X}\right) = k\right) = \frac{1}{\binom{n}{k}} \cdot \mathbb{1}_{\left(\sum_{i=1}^{n} x_i = k\right)}. \tag{16}$$

Since (16) is independent of $\theta$, we conclude that $T\left(\underline{X}\right) = \sum_{i=1}^{n} X_i$ is a sufficient statistic for the parameter $\theta$.

**Example 2.5 (Normal Sufficient Statistic (when $\sigma^2$ is known)).** Without loss of generality, let $\sigma^2 = 1$. Let $X_1, \ldots, X_n \overset{iid}{\sim} \mathcal{N}(\mu, 1)$. Here, $\mu$ is unknown, and the objective is to construct at a sufficient statistic for $\mu$. Let

$$T(\underline{X}) = \frac{1}{n} \sum_{i=1}^{n} X_i = \bar{X}, \tag{17}$$

where $\bar{X}$ is the sample mean. We wish to see if $\bar{X}$ is a sufficient statistic for $\mu$. We have

$$\begin{aligned}
f_{\underline{X}|T(\underline{X})}(\underline{x}|t) &= \frac{f_{\underline{X}}(\underline{x})}{f_{T(\underline{X})}(t)}, \\
&= \frac{\left(\frac{1}{\sqrt{2\pi}}\right)^n \exp\left[-\frac{1}{2} \sum_{i=1}^{n} (x_i - \mu)^2\right]}{\frac{1}{\sqrt{2\pi}} \frac{1}{\sqrt{n}} \exp\left[-\frac{1}{2\left(\frac{1}{n}\right)} (t - \mu)^2\right]}, \\
&= \frac{\left(\frac{1}{\sqrt{2\pi}}\right)^n \exp\left[-\frac{1}{2}\left(\sum_{i=1}^{n} (x_i - \bar{x})^2 + n(\bar{x} - \mu)^2\right)\right]}{\frac{1}{\sqrt{2\pi}} \frac{1}{\sqrt{n}} \exp\left[-\frac{1}{2\left(\frac{1}{n}\right)} (t - \mu)^2\right]}, \\
&= \sqrt{n} \left(\frac{1}{\sqrt{2\pi}}\right)^{n-1} \exp\left[-\frac{1}{2} \sum_{i=1}^{n} (x_i - \bar{x})^2\right],
\end{aligned} \tag{18}$$

where the last line follows from the fact that $\bar{x} = t$. Since the RHS of (18) is independent of $\mu$, we conclude that the sample mean $\bar{X}$ is a sufficient statistic for the mean of a population following normal distribution when the variance is known. However, this is not true in general for other distributions.