STATISTICS
A VIEW FROM CALCULUS PERSPECTIVE

Brigida A. Roscom

The dictum *Learn Calculus First* applies to all mathematical sciences including statistics. While no previous knowledge of statistics is assumed, a good background in calculus is a requisite for students to work their own way through the courses in mathematical statistics or the classical statistical theories. A sound background in calculus that is essential for learning statistics, includes a good working knowledge of multiple integration, partial differentiation and power series.

This paper is a tour of statistics via calculus with the aim of providing *a deeper look at statistics*. In the preparation of this tour, the author relied heavily on existing literature on the subject of mathematical statistics.

An extended view on mean, median, and mode. The three averages, the mean, median, and mode, are popularly seen as single numerical values which summarize the features of a data set, usually a sample or a finite population data. The computational concepts of these averages are straightforward. But it is obvious that the computations have limitations when the population data set is infinite. It is this limitation that compels us to extend our view on these measures of location. The background, with which this view is set up, is built upon the concepts of random variables, probability, distributions and expectations. To give you a feel of these concepts, consider the following definitions.

7.1 DEFINITION. If X is a discrete random variable with distinct values $x_1, x_2, \ldots, x_n, \ldots$ with $P[X = x]$ as the probability that the random variable X takes the value x, then the function $f(.)$ defined by

$$f(x) = \begin{cases} P[X = x], & \text{if } x = x_j, j = 1, 2, \ldots, n, \ldots \\ 0, & \text{if } x \neq x_j \end{cases}$$

is called the ***density function*** of X. It is also called the ***probability mass function*** or ***pmf*** of X.

7.2 DEFINITION. For a continuous random variable X with domain
$(-\infty,\infty)$, if there exists a function $f(.)$ such that

$$P(X \leq x) = \int_{-\infty}^{x} f(u)du,$$

then f is called the ***probability density function*** or ***pdf*** of X.

The density function $f(.)$ has the following properties :

(i) $f(x_j) > 0$ for $j = 1, 2, \ldots, r$ $f(x) = 0$ for $x \neq x_j$ and $\sum_j f(x_j) = 1$,
if X is discrete.

(ii) $f(x) \geq 0$ and $\int_{-\infty}^{\infty} f(x)dx = 1$ if X is continuous.

7.3 DEFINITION. The ***mean***, μ_X, or ***expected value***, $\xi[X]$ of a random variable X is defined by

$$\xi[X] = \begin{cases} \sum_j x_j f(x_j), & \text{if } X \text{ is discrete with mass points at } x_j, \text{ and} \\ \int_{-\infty}^{+\infty} x f(x)dx, & \text{if } X \text{ is continuous with pdf } f(x). \end{cases}$$

Note that the definition assumes that the series is absolutely convergent or the integral exists, respectively; otherwise, the mean does not exist.

7.4 DEFINITION. The ***median*** x_{med} of a random variable X is a number satisfying the following inequalities:

$$P[X \leq x_{\text{med}}] \geq \frac{1}{2} \text{ and } P[X \geq x_{\text{med}}] \geq \frac{1}{2}.$$

If X is a continuous random variable, then the median of X satisfies

$$\int_{-\infty}^{x_{med}} f(x)dx = \frac{1}{2} = \int_{x_{med}}^{\infty} f(x)dx .$$

7.5 DEFINITION. The ***mode***, denoted by x_{mo}, of a random variable X with density function $f(x)$ is a numerical value which satisfies

$$f(x_{mo}) = \max_{x}\{f(x)\} .$$

The mode, if it exists, is that value of X which maximizes $f(x)$.

7.6 EXAMPLE. A random variable X, which is ***normally distributed*** with mean μ and variance σ^2, where $\sigma > 0$, has pdf given by

$$f(x) = \frac{1}{\sigma\sqrt{2\pi}} \exp\left[-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2\right].$$

Let us show that the mean or expected value of X is μ. By Definition 7.3,

$$\xi[X] = \int_{-\infty}^{+\infty} x\left(\frac{1}{\sigma\sqrt{2\pi}}\right) \exp\left[-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2\right] dx .$$

Evaluating the integral by changing variables, i.e., by letting $z = \frac{x-\mu}{\sigma}$, we obtain

$$\xi[X] = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{+\infty} (\sigma z + \mu) e^{\frac{-z^2}{2}} dz ,$$

$$\xi[X] = \frac{\sigma}{\sqrt{2\pi}} \int_{-\infty}^{+\infty} z e^{\frac{-z^2}{2}} dz + \frac{\mu}{\sqrt{2\pi}} \int_{-\infty}^{+\infty} e^{\frac{-z^2}{2}} dz$$

$$= \frac{\sigma}{\sqrt{2\pi}} A + \frac{\mu}{\sqrt{2\pi}} B,$$

where

$$A = \int_{-\infty}^{+\infty} ze^{\frac{-z^2}{2}} dz \quad = -e^{\frac{-z^2}{2}} \Bigg|_{-\infty}^{+\infty}$$

$$= \lim_{z \to +\infty} (-e^{-z^2/2}) - \lim_{z \to -\infty} (-e^{-z^2/2}) = 0,$$

and

$$\frac{1}{\sqrt{2\pi}} B = \int_{-\infty}^{+\infty} \frac{1}{\sqrt{2\pi}} e^{z^2/2} dx = 1,$$

since the second integrand is the density function of that a normal random variable with mean 0 and variance 1.

Combining the results for A and B, the mean of a normally distributed random variable X is

$$\xi[X] = \frac{1}{\sqrt{2\pi}} \cdot 0 + \mu \cdot 1 = \mu.$$

7.7 EXAMPLE. Again, let us consider finding the mode of a normally distributed random variable X. Invoking Definition 7.5, we shall apply the methods of differentiation to obtain the maximum value of the pdf of X. Equating the first derivative of $f(x)$ of Example 7.6 to zero, we obtain

$$f'(x) = \frac{1}{\sqrt{2\pi\sigma}} e^{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2} \cdot -\left(\frac{x-\mu}{\sigma}\right) = 0, \quad \text{or}$$

$$e^{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2} \cdot -\left(\frac{x-\mu}{\sigma}\right) = 0.$$

The first factor cannot be zero, so $\left(\frac{x-\mu}{\sigma}\right) = 0,$ hence $x = \mu$.

The second derivative test would confirm if indeed the solution $x = \mu$ gives the maximum value of $f(x)$. Now since

$$f''(x) = \frac{-1}{\sqrt{2\pi}\sigma}\left\{\left(e^{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2}\right)\cdot\left(\frac{x-\mu}{\sigma}\right)^2 + \left(\frac{1}{\sigma}\right)\cdot\left(e^{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2}\right)\right\}$$

$$= \frac{-1}{\sqrt{2\pi}\sigma}\left(e^{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2}\right)\left[\left(\frac{x-\mu}{\sigma}\right)^2 + \left(\frac{1}{\sigma}\right)\right],$$

we immediately have

$$f''(\mu) = \left[-\frac{1}{\sqrt{2\pi}\sigma}\cdot 1\right]\left[\frac{1}{\sigma}\right] < 0,$$

since $\sigma > 0$. Thus the maximum of f is attained at $x = \mu$, which shows that the mode $x_{mo} = \mu$.

Behind that simple regression line. Consider a set of bivariate data $\{(x_i, y_i) : i = 1, 2, \ldots, n\}$. A simple linear regression model postulates a straight-line relationship between the variables X and Y. The model expressed as

$$y_i = A + Bx_i + \varepsilon_i$$

is intended to express an approximately linear relation between X and Y, where any deviation from a perfect straight line relationship is attributed to chance or random variation. In the model, x_i is some fixed value of X; the y_i's are independent normally distributed random variables with mean zero and the same unknown standard deviation σ. And for each x_i, a set of differing y_i values have mean $A + Bx_i$. The ε_i's, where

$$\varepsilon_i = y_i - (A + Bx_i),$$

are the deviations of the Y values from the regression line $Y = A + BX$.

Below is a table of pairs of height x and weight y from a sample of 15 female juniors. The pairs (x,y) are plotted in Figure 1.

x	150	158	155	155	156	157	157	158	160	162	162	163	165	166	166
y	104	100	112	98	105	95	114	108	99	114	95	110	100	118	106

Table 1. Heights and weights of 15 female students

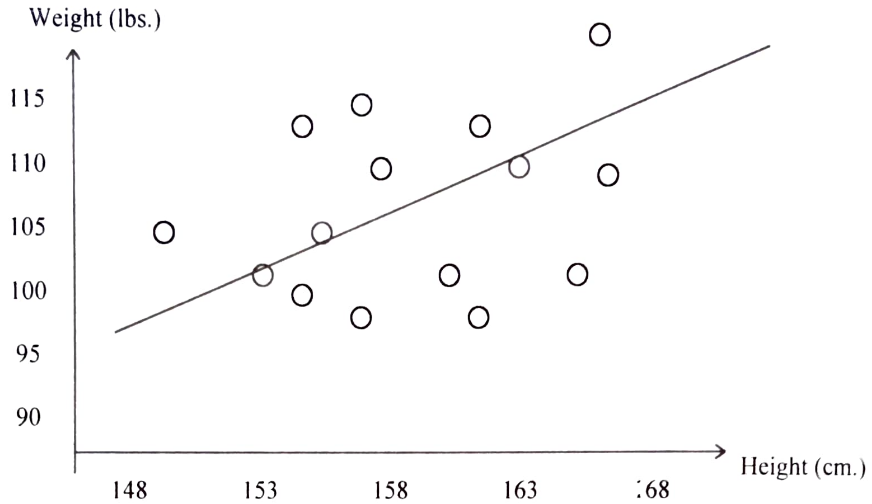

Figure 1. Scatter Plot of Weight vs. Height

Figure 2 below shows a picture that may help in understanding the model.

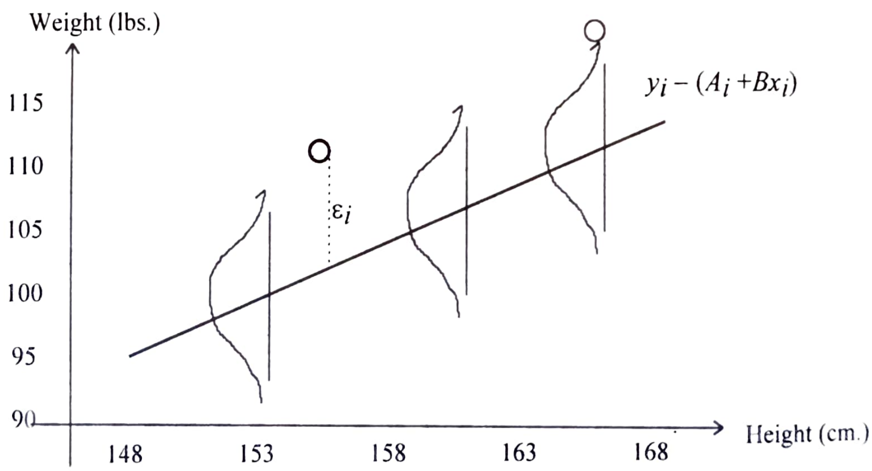

Figure 2. Assumptions in a Regression Model

The parameters A and B are estimated using sample data. Behind the estimation of these parameters is the *principle of least squares*. The principle declares that when choosing among the possible lines to represent a bivariate set of data, the line that best fits is that line which makes as small as possible the sum S of the squared vertical distances from the points to the line. The mathematical procedure of finding this best line uses partial differentiation. What follows is the solution towards the least squares estimates for the regression line coefficients A and B. Let S be the sum of the squared deviations, then

$$S = \sum_{i=1}^{n} \varepsilon_i = y_i - \left(A + Bx_i\right).$$

Employing partial differentiation technique to minimize S, we proceed with

$$\frac{\partial S}{\partial A} = -2\sum_{i=1}^{n}(Y_i - A - BX_i) = 0, \text{ and}$$

$$\frac{\partial S}{\partial B} = -2\sum_{i=1}^{n} X_i(Y_i - A - BX_i) = 0.$$

Thus the estimates for A and B that minimizes S, denoted by a and b respectively, are the solutions to the equations below:

$$\sum_{i=1}^{n}(Y_i - a - bX_i) = 0, \text{ and}$$

$$\sum_{i=1}^{n} X_i(Y_i - a - bX\ ) \ = 0.$$

Summing up each of the equations above, we have

$$\sum_{i=1}^{n} Y_i - na - b\sum_{i=1}^{n} X_i = 0$$

$$\sum_{i=1}^{n} X_iY_i - a\sum_{i=1}^{n} X_i - b\sum_{i=1}^{n} X_i^2 = 0\text{, or}$$

$$an + b\sum_{i=1}^{n} X_i = \sum_{i=1}^{n} Y_i\text{, and}$$

$$a\sum_{i=1}^{n} X_i + b\sum_{i=1}^{n} X_i^2 = \sum_{i=1}^{n} X_iY_i\,.$$

The last pair of equations above, are known as the normal equations. They have solutions for b and a as follows:

$$b = \frac{\sum_{i=1}^{n} X_iY_i - \frac{1}{n}\left[(\sum_{i=1}^{n} X_i)(\sum_{i=1}^{n} Y_i)\right]}{\sum_{i=1}^{n} X_i^2 - \frac{1}{n}(\sum_{i=1}^{n} X_i)^2} = \frac{\sum_{i=1}^{n}(X_i - \overline{X})(Y_i - \overline{Y})}{\sum_{i=1}^{n}(X_i - \overline{X})^2} \quad \text{and}$$

$$a = \overline{Y} - b\overline{X}\,.$$

The Poisson distribution as a limiting distribution for the binomial distribution. Two of the commonly used discrete distribution models in the life and social sciences are the *Poisson* and *Binomial distributions*. The Poisson distribution provides a probabilistic model for a wide class of phenomena. Examples are the number of telephone calls during a given period of time, the number of particles emitted from a radioactive source, and the number of cars passing by an intersection point. The binomial distribution is by far the most important discrete distribution. An experiment or an activity follows a ***binomial model*** if it has n independent trials with two possible outcomes per trial: either a specific event occurs or does not occur. The probability p of the occurrence of the said event remains the same from trial to trial. Typical examples are flipping of a coin, getting a defective or nondefective product, and having a boy or a girl for a child.

7.8 THEOREM. *Let $B(x;\, n, p)$ and $P(x;\, \lambda)$ be the binomial and Poisson probability mass functions, respectively. For each fixed x, as $n \to \infty$ and $p \to 0$ such that $\lambda = np$ remains constant,*

$$\lim_{n\to\infty} B(x; n, p) \;=\; P(x; np).$$

Proof. $B(x;\, n, p)$ denotes the probability that an event E has x number of occurrences in n trials. The event E has a probability p of occurrence in each trial. The function $B(x;\, n, p)$ is defined by

$$B(x;\, n, p) \;=\; \begin{bmatrix} n \\ p \end{bmatrix} p^x\, q^{n-x}$$

$$= \frac{n\,!}{x\,!\,(n-x)\,!}\, p^x\,(1-p)^{n-x}$$

$$= \frac{n\,(n-1)\,\cdots\,(n-x+1)}{x\,!}\, p^x\,(1-p)^{n-x} \qquad (7.8.1)$$

Now, allow p to vary with n, so take $p = \frac{\lambda}{n}$, $n \geq 1$, $\lambda > 0$. Thus, substituting $p = \frac{\lambda}{n}$ in (7.8.1), we have,

$$B(x;\, n, p) \;=\; \frac{n\,(n-1)\,\cdots\,(n-x+1)}{x\,!}\left[\frac{\lambda}{n}\right]^x \left[1-\frac{\lambda}{n}\right]^{n-x}$$

$$= 1\left[1-\frac{1}{n}\right]\cdots\left[1-\frac{x-1}{n}\right]\frac{\lambda^x}{x\,!}\left[1-\frac{\lambda}{n}\right]^n\left[1-\frac{\lambda}{n}\right]^{-x}$$

As $n \to \infty$, while x and λ remain constant, we obtain the following:

$$\lim_{n\to\infty} 1\left[1-\frac{1}{n}\right]\cdots\left[1-\frac{x-1}{n}\right] \;=\; 1 \quad \text{and}$$

$$\lim_{n\to\infty}\left[1-\frac{\lambda}{n}\right]^{-x} = 1.$$

From the definition of the number e, we also have

$$\lim_{n\to\infty}\left[1-\frac{\lambda}{n}\right]^{n} = e^{-\lambda}. \qquad (7.8.2)$$

Therefore, under the given limiting conditions,

$$\lim_{n\to\infty} B(x;n,p) = P(x;\lambda) = \frac{e^{-\lambda}\lambda^{x}}{x!}.$$

The amazing central limit theorem. When a finite set of independent random variables $X_1, X_2, \ldots, X_n$ has a common distribution, the probability distribution for their mean $\overline{X}$ is approximately normal for large n. The precise statement of this observation is one of the most celebrated theorems of mathematics, the so-called *Central Limit Theorem* or CLT.

The Central Limit Theorem implies that if the sample n is large and yet a small fraction of the population size N so that independence of $X_1, X_2, \ldots, X_n$ is reasonable, we can approximate the probabilities of the sample mean $\overline{X}$ using the table of areas under the normal curve. A special case of this theorem is stated below.

7.9 THEOREM (CENTRAL LIMIT THEOREM). *Let $f(.)$ be a density function with mean μ and variance σ^2. Let $\overline{X_n}$ be the sample mean of a random sample of size n from $f(.)$. Let the random variable Z_n be defined by*

$$Z_n = \frac{\overline{X_n} - \xi\left(\overline{X_n}\right)}{\sqrt{var\left(\overline{X_n}\right)}} = \frac{\overline{X_n} - \mu}{\sigma/\sqrt{n}}.$$

Then, the distribution of Z_n approaches the standard normal distribution as n tends to infinity.

Discussion. The amazing part of the Central Limit Theorem is the fact that nothing is assumed about the form of the original density function. The importance of this theorem as far as applications are concerned, is the fact that the mean of a random sample of size n from any distribution with finite variance σ^2 and mean μ, is approximately distributed as a normal random variable with mean μ and variance σ^2/n. The proof of this theorem will make use of the concepts of moments, and moment generating function of a standard normal random variable.

The ***moment generating function*** or ***mgf*** of a random variable X is defined as

$$m(t) = \xi\left[e^{tX}\right] = \int_{-\infty}^{+\infty} e^{tX} f(x)dx.$$

If the mgf of X exists, then $m(t)$ is continuously differentiable at some neighborhood of the origin. Differentiating $m(t)$, r times with respect to t, we obtain

$$\frac{d^r}{dt^r} m(t) = \int_{-\infty}^{+\infty} x^r e^{tx} f(x)dx.$$

Letting $t \to 0$, we have

$$\frac{d^r}{dt^r} m(0) = \int_{-\infty}^{+\infty} x^r f(x)dx = \mu_r.$$

The number μ_r is called the ***rth moment*** *of f*(x). Replacing e^{tX} by its series expansion in $\xi\left[e^{tX}\right]$, we obtain

$$m(t) = \xi\left[1 + Xt + \frac{1}{2!}(Xt)^2 + \frac{1}{3!}(Xt)^3 + \dots\right]$$

$$= 1 + \mu_1 + \frac{1}{2!}t^2\mu_2 + \frac{1}{3!}t^3\mu_3 + \dots$$

$$= \sum_{i=0}^{\infty} \frac{\mu_r}{i!} t^r .$$

Hence, the mgf of a standard normal random variable is $m(t) = e^{\frac{1}{2}t^2}$.

If the moment generating function of Z_n, $m_{Z_n}(t)$ can be shown to approach the standard normal random variable, $m(t)$ as n becomes large then the Theorem 7.9 is proved.

Proof of Theorem 7.9. Using the independence of $X_1, X_2, \ldots, X_n$, we obtain

$$m_{Z_n}(\mathrm{t}) = \xi\left[e^{tZ_n}\right] = \xi\left[\exp\left(t\frac{\overline{X} - \mu}{\sigma / \sqrt{n}}\right)\right] = \xi\left[\exp\left(\frac{t}{n}\sum_{i=1}^{n}\frac{X_i - \mu}{\sigma / \sqrt{n}}\right)\right]$$

$$= \xi\left[\prod_{i=1}^{n}\exp\left(\frac{t}{n}\cdot\frac{X_i - \mu}{\sigma / \sqrt{n}}\right)\right] = \prod_{i=1}^{n}\xi\left[\exp\left(\frac{t}{\sqrt{n}}\cdot\frac{X_i - \mu}{\sigma}\right)\right].$$

Now if we let $Y_i = \left(X_i - \mu\right)/\sigma$, then $m_{Y_i}(t)$, the moment generating function of Y_i, is independent of i since all Y_i's have the same distribution. Let $m_Y(t)$ denote each $m_{Y_i}(t)$. Then

$$\prod_{i=1}^{n}\xi\left[\exp\left(\frac{t}{\sqrt{n}}\cdot\frac{X_i - \mu}{\sigma}\right)\right] = \prod_{i=1}^{n}\xi\left[\exp\left(\frac{\mathrm{t}}{\sqrt{n}}\cdot Y_i\right)\right]$$

$$= \prod_{i=1}^{n} m_{Y_i}\left(\frac{t}{\sqrt{n}}\right) = \left[m_Y\left(\frac{t}{\sqrt{n}}\right)\right]^n .$$

Hence,

$$m_{Z_n}(t) = \left[m_Y\left(\frac{t}{\sqrt{n}}\right)\right]^n .$$

The rth derivative of $m_Y(t/\sqrt{n})$ evaluated at $t = 0$ gives us the rth moment about the mean of the density $f(.)$ divided by $(\sigma\sqrt{n})^r$, so we may write its Taylor expansion

$$m_Y\left(\frac{t}{\sqrt{n}}\right) = 1 + \frac{\mu_1}{\sigma}\frac{t}{\sqrt{n}} + \frac{1}{2!}\frac{\mu_2}{\sigma^2}\left(\frac{t}{\sqrt{n}}\right)^2 + \frac{1}{3!}\frac{\mu_3}{\sigma^3}\left(\frac{t}{\sqrt{n}}\right)^3 + \cdots .$$

Since $\mu_1 = 0$ and $\mu_2 = \sigma^2$, this may be written

$$m_Y\left(\frac{t}{\sqrt{n}}\right) = 1 + \frac{1}{n}\left(\frac{1}{2}t^2 + \frac{1}{3!\sqrt{n}}\frac{\mu_3}{\sigma^3}t^3 + \frac{1}{4!n}\frac{\mu_4}{\sigma^4}t^4 + \cdots\right). \qquad (7.9.1)$$

Now recall that $\lim_{n\to\infty}\left(1+\frac{u}{n}\right)^n = e^u$, if u is constant. Thus, if u represents the expression within the parenthesis in (7.9.1), then it follows that $\lim_{n\to\infty}\left(1+\frac{u}{n}\right)^n = e^{\frac{1}{2}t^2}$, because $\lim_{n\to\infty} u = \frac{1}{2}t^2$. Therefore, we have

$$\lim_{n\to\infty} m_{Z_n}(t) = \lim_{n\to\infty}[m_Y(\tfrac{t}{\sqrt{n}})]^n = \lim_{n\to\infty}(1+\tfrac{\mu}{n})^n = e^{\frac{1}{2}t^2},$$

so that, at the limit, Z_n has the same moment generating function as that of the standard normal variable.

REFERENCES

[1] Mood, A. M. and Graybill, *Introduction to the Theory of Statistics*, McGraw-Hill, 1963

[2] Wilks, S.S. *Mathematical Statistics*, Wiley, 1962