Lecture 24

The Sample Variance S^2

The squared variation

Suppose we have n numbers $x_1, x_2, \ldots, x_n$. Then their squared variation $sv = sv(x_1, x_2, \ldots x_n)$

$$sv(x_1, x_2, \ldots, x_n) = \sum_{i=1}^{n} (x_i - \bar{x})^2$$

Their mean (average) squared variation msv or σ_n^2 (denoted σ^2 and called the "population variance on page 33 of our text) is given by

$$msv = \sigma_n^2 = \frac{1}{n} sv = \frac{1}{n} \sum_{i=1}^{n} (x_i - \bar{x})^2$$

Here $\bar{x}$ is the average $\frac{1}{n} \sum_{i=1}^{n} x_i$.

The msv measure how much the numbers $x_1, x_2, \ldots, x_n$ vary (precisely how much they vary from their average $\bar{x}$). For example if they are all equal then they will be all equal to their average $\bar{x}$

so

$$sv = 0 \quad \text{and} \quad msv = 0$$

We also define the sample variance s^2 by

$$s^2 = \frac{1}{n-1} sv = \frac{n}{n-1} msv$$

$$s^2 = \frac{1}{n-1} \sum_{i=1}^{n} (x_i - \bar{x})^2$$

Amazingly, s^2 is more important than msv in statistics.

The Shortcut Formula for the Squared Variation

Theorem

$$SV(x_1, x_2, \ldots, x_n) = \sum_{i=1}^{n} x_i^2 - \frac{1}{n}\left(\sum_{i=1}^{n} x_i\right)^2 \quad (*)$$

Proof

Note since $\overline{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$

we have $\sum_{i=1}^{n} x_i = n\overline{x}$

Now

$$\sum_{i=1}^{n} (x_i - \overline{x})^2 = \sum_{i=1}^{n} (x_i^2 - 2x_i\overline{x} + \overline{x})^2$$

$$= \sum_{i=1}^{n} x_i^2 - \sum_{i=1}^{n} 2x_i\overline{x} + \sum_{i=1}^{n} \overline{x}^2$$

$$= \sum_{i=1}^{n} x_i^2 - 2\overline{x}\sum_{i=1}^{n} x_i + \overline{x}^2\sum_{i=1}^{n} 1$$

$$= \sum_{i=1}^{n} x_i^2 - 2\bar{x}(n\bar{x}) + n\bar{x}^2$$

$$= \sum_{i=1}^{n} x_i^2 - 2n\,\bar{x}^2 + n\bar{x}^2$$

$$= \sum_{i=1}^{n} x_i^2 - n\bar{x}^2$$

$$= \sum_{i=1}^{n} x_i^2 - n\left(\frac{\sum_{i=1}^{n} x_i}{n}\right)^2$$

$$= \sum_{i=1}^{n} x_i^2 - \cancel{n}\frac{\left(\sum_{i=1}^{n} x_i\right)^2}{\cancel{n}^2}$$

$$= \sum_{i=1}^{n} x_i^2 - \frac{1}{n}\left(\sum_{i=1}^{n} x_i\right)^2$$

$\square$

Corollary 1

Divide both sides of (*) by n to get

$$msv = \frac{1}{n}\sum_{i=1}^{n} x_i^2 - \frac{1}{n^2}\left(\sum_{i=1}^{n} x_i\right)^2$$

Corollary 2 (Shortcut formula for s^2)

Divide both sides of (*) by $n-1$ to get

$$s^2 = \frac{1}{n-1}\sum_{i=1}^{n} x_i^2 - \frac{1}{n(n-1)}\left(\sum_{i=1}^{n} x_i\right)^2$$

It is this last formula that we will need.

Let me give a conceptual proof of the theorem — the way a professional mathematician would prove the theorem.

Definition

A polynomial $p(x_1, x_2, \dots, x_n)$ is symmetric if it is unchanged by permuting the variables.

Examples

$p(x,y,z) = x^2+y^2+z^2$ is symmetric

$p(x,y,z) = xy+z^2$ is not symmetric

Theorem

Any symmetric polynomial p in $x_1, x_2, \dots, x_n$ can be rewritten as a polynomial in the power sums $\sum_{i=1}^{n} x_i^k$ that is

$$p(x_1, \dots, x_n) = q\left(\sum x_i, \sum x_i^2, \dots, \sum x_i^l\right)$$

if $\deg p = l$.

Bottom Line

$SV = \sum_{i=1}^{n} (x_i - \bar{x})^2$ is a symmetric polynomial in $x_1, x_2, .., x_n$ so there exist a and b with

$$SV(x_1, x_2, .., x_n) = a \sum_{i=1}^{n} x_i^2 + b\left(\sum_{i=1}^{n} x_i\right)^2 \quad (**)$$

This is true for all $x_1, .., x_n$ (an "identity") so we just choose $x_1, .., x_n$ cleverly to get a and b.

First choose $x_1 = 1, x_2 = -1, x_3 = .. = x_n = 0$

so $\sum_{i=1}^{n} x_i = 0$ and $\sum_{i=1}^{n} x_i^2 = 2$

Since $\bar{x} = 0$

and $SV(1, -1, 0, .. 0) = \sum_{i=1}^{n} (x_i - \bar{x})^2 = \sum_{i=1}^{n} x_i^2$

$(**)$ becomes

$$2 = a2 + b(0) \quad \text{so } a = 1$$

To find b take all the x_i's to be 1. so $\bar{x} = 1$ and $sv(1, 1, 1) = 0$ (there is no variation in the x_i's)

$\sum_{i=1}^n x_i^2 = n$, $\sum_{i=1}^n x_i = n$ so

$$sv(x_1, \dots, x_n) = \sum_{i=1}^n x_i^2 + b\left(\sum x_i\right)^2$$

gives us

$$0 = n + bn^2 \quad \text{so} \quad b = -\frac{1}{n}$$

and

$$sv(x_1, x_2, \dots x_n) = \sum_{i=1}^n x_i^2 - \frac{1}{n}\left(\sum x_i\right)^2$$

as before.

Remark Any symmetric quadratic function $q(x_1, x_2, \dots, x_n)$ is a linear combination of $\sum_{i=1}^n x_i^2$ and $\left(\sum_{i=1}^n x_i\right)^2$ that is

$$q(x_1, \dots, x_n) = a \sum_{i=1}^n x_i^2 + b\left(\sum_{i=1}^n x_i\right)^2$$

In Which We Return to Statistics

Estimating the Population Variance

We have seen that $\overline{X}$ is a good (the best) estimator of the population mean μ, in particular it was an unbiased estimator

$$E(\overline{X}) = \mu$$

sample mean random variable

population mean

How do we estimate the population variance?

$$\boxed{\begin{matrix} X \\ V(X)=\sigma^2 \end{matrix}} \longrightarrow x_1, x_2, \ldots, x_n \longrightarrow s^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i-\bar{x})^2$$

Answer – use the sample variance s^2 to estimate the population variance σ^2

The reason is that if we take the associated sample variance random variable

$$S^2 = \frac{1}{n-1}\sum_{i=1}^{n-1}(X_i-\bar{X})^2$$

then we have

Amazing Theorem

$$E(S^2) = \sigma^2$$

sample variance

population variance

Why do you need $\frac{1}{n-1}$? We will see.

Before starting the proof we first note the Corollary 2, page 2 implies

<u>Proposition</u> (Shortcut formula for the sample variance random variable S^2)

$$S^2 = \frac{1}{n-1} \sum_{i=1}^{n} X_i^2 - \frac{1}{n(n-1)} \left(\sum_{i=1}^{n} X_i \right)^2 \qquad (b)$$

Why does this follow from the formula for s^2?

We will also need the following

<u>Proposition</u>

Suppose Y is a random variable then

$$E(Y^2) = E(Y)^2 + V(Y) \qquad (\#)$$

<u>Proof</u> $V(Y) = E(Y^2) - (E(Y))^2$
(shortcut formula for $V(Y)$)

$\square$

Corollary

Suppose $X_1, X_2, \ldots, X_n$ is a random sample from a population of mean μ and variance σ^2. Then

(i) $E(X_i^2) = \mu^2 + \sigma^2$

(ii) $E(T_0) = n^2\mu^2 + n\sigma^2$

Proof

(i) $E(X_i) = \mu$ and $V(Y) = \sigma^2$

so plug into (#)

(ii) $E(T_0) = n\mu$ and $V(T_0) = n\sigma^2$

so plug into (#)

We can now prove (b)

$$E(S^2) = E\left(\frac{1}{n-1}\sum_{i=1}^{n} X_i^2 - \frac{1}{n(n-1)}\underbrace{\left(\sum X_i\right)^2}_{T_0}\right)$$

since E is linear

$$= \frac{1}{n-1}\sum_{i=1}^{n} E(X_i^2) - \frac{1}{n(n-1)}E(T_0^2)$$

by (i) and (ii)

$$= \frac{1}{n-1}\sum_{i=1}^{n}(\mu^2+\sigma^2) - \frac{1}{n-1}\frac{1}{n}(n^2\mu^2 + n\sigma^2)$$

$$= \frac{1}{n-1}\left[n\mu^2 + n\sigma^2 - \frac{1}{n}(n^2\mu^2 + n\sigma^2)\right]$$

$$= \frac{1}{n-1}\left[\cancel{n\mu^2} + n\sigma^2 - \cancel{n\mu^2} - \sigma^2\right]$$

$$= \frac{1}{n-1}\left[(n-1)\sigma^2\right]$$

$$= \sigma^2$$

$\square$

Amazing – you need $\frac{1}{n-1}$ not $\frac{1}{n}$.