

ANALYZING DATA

Suppose we want to know the average weight in a large population. So we choose a random sample of n individuals, weigh them, and average the numbers. What do we now know about the population? The point of statistics is to draw reasonable conclusions from incomplete information.

Assume that X_1, \dots, X_n are independent identically distributed random variables, with an unknown mean μ and an unknown variance σ^2 . We can think of X_i as the i th measurement out of the total of n measurements; for example, X_i might be the weight in kilograms of the i th person chosen.

The **sample mean** is the average

$$\bar{X} = \frac{X_1 + \dots + X_n}{n}$$

which we hope is close to the unknown μ . We already know that $E(\bar{X}) = \mu$, $\text{var}(\bar{X}) = \sigma^2/n$, the standard deviation of \bar{X} is σ/\sqrt{n} , and by the law of large numbers, \bar{X} converges in probability to μ as $n \rightarrow \infty$.

The **sample variance** S^2 is defined by the two equations:

$$\begin{aligned} S^2 &= \frac{(X_1 - \bar{X})^2 + \dots + (X_n - \bar{X})^2}{n - 1} \\ &= \frac{X_1^2 + \dots + X_n^2}{n - 1} - \frac{n}{n - 1} \bar{X}^2 \end{aligned}$$

(The first equation is reminiscent of the “ $E((X - \mu)^2)$ ” version of the variance, and the second is reminiscent of the “ $E(X^2) - \mu^2$ ” version.) We hope that S^2 is close to σ^2 ; it is at least true that $E(S^2) = \sigma^2$ (see page 902). Consequently by the law of large numbers, S^2 converges in probability to σ^2 as $n \rightarrow \infty$.

The **sample standard deviation** S is the square root of S^2 ; we hope that S is close to the unknown σ .

The **standard error** S.E. is defined to be $\frac{S}{\sqrt{n}}$; we hope it is close to the standard deviation of \bar{X} .

By the central limit theorem, we know that

$$P(|\bar{X} - \mu| \leq z \frac{\sigma}{\sqrt{n}}) \approx P(|Z| \leq z)$$

where Z has the standard normal distribution. This equation can be rewritten:

$$P(\mu \in [\bar{X} - z \frac{\sigma}{\sqrt{n}}, \bar{X} + z \frac{\sigma}{\sqrt{n}}]) \approx P(|Z| \leq z)$$

That is, $P(|Z| \leq z)$ measures our degree of *confidence* that μ is somewhere in the interval $[\bar{X} - z\frac{\sigma}{\sqrt{n}}, \bar{X} + z\frac{\sigma}{\sqrt{n}}]$. This cannot be used directly if we don't know σ . But our data points do give us the standard error S.E. For large n , the standard error will be close to $\frac{\sigma}{\sqrt{n}}$ and hence

$$P(|\bar{X} - \mu| \leq z(\text{S.E.})) \approx P(|Z| \leq z)$$

or in other words:

$$P(\mu \in [\bar{X} - z(\text{S.E.}), \bar{X} + z(\text{S.E.})]) \approx P(|Z| \leq z)$$

The intervals

$$[\bar{X} - z\frac{\sigma}{\sqrt{n}}, \bar{X} + z\frac{\sigma}{\sqrt{n}}] \quad \text{and} \quad [\bar{X} - z(\text{S.E.}), \bar{X} + z(\text{S.E.})]$$

are called **confidence intervals**. For example, if $P(|Z| \leq z) = 0.95$ then they are called 95% confidence intervals (this happens when $z = 1.96$).

Binary variables. The foregoing is applicable no matter how the individual measurements X_i are distributed. But a special case of interest is where the range of X_i is $\{0, 1\}$:

$$X_i = \begin{cases} 1 & \text{with probability } p \\ 0 & \text{with probability } 1 - p \end{cases}$$

for some (unknown) number p . For example, this is the situation in the pill problem. In this case, we have n Bernoulli trials (where “success” is when $X_i = 1$). And $\mu = p$ and $\sigma^2 = p(1 - p)$. The sum $X_1 + \cdots + X_n$ is the number k of successes, and $\bar{X} = \frac{k}{n}$ is the average number of successes; call it \hat{p} :

$$\hat{p} = \bar{X} = \frac{k}{n}$$

Using this notation, we obtain

$$S^2 = \frac{k}{n-1} - \frac{n}{n-1}\hat{p}^2 = \frac{n\hat{p} - n\hat{p}^2}{n-1} = \frac{n}{n-1}(\hat{p} - \hat{p}^2)$$

and therefore

$$\text{S.E.} = \sqrt{\frac{\hat{p}(1 - \hat{p})}{n-1}} \approx \sqrt{\frac{\hat{p}(1 - \hat{p})}{n}}.$$

Compare this with what we did in §12.6: There we used the “worst-case” value of $1/4$ for σ^2 ; this value holds when $p = 1/2$. But if we have data showing that p is not close to $1/2$, then it makes sense to use S.E. instead of $1/2\sqrt{n}$. (The worst-case estimate is applicable *only* to binary variables.)

—H. B. Enderton