Basic Statistical Notation

Notation	Meaning	Remark
y	random variable	it takes different values with probabilities
$\mu = E(y) = \Sigma_j y_j f(y_j)$	population mean	the center of a distribution weighted average of possible values (y_j) weight is probability ($f(y_j)$)
$\sigma^2 = \texttt{var}(y) = E(y-\mu)^2$	population variance	the dispersion of the distribution distribution is wide if $\texttt{var}(y)$ is big
$\sigma = \sqrt{\texttt{var}(y)}$	standard deviation (sd)	another measure of dispersion
$\{y_1, y_2, \ldots, y_n\}$	sample	we get a random or i.i.d sample if $E(y_i) = \mu, \texttt{var}(y_i) = \sigma^2, \texttt{cov}(y_i, y_j) = 0, \forall i, j$
$\bar{y} = \frac{\Sigma_{i=1}^n y_i}{n}$	sample mean	estimate for population mean it is a random variable $E(\bar{y}) = \mu$ if we use random sample $\texttt{var}(\bar{y}) = \frac{\sigma^2}{n}$ if we use random sample $\frac{\sigma}{\sqrt{n}}$ is standard error (se) of $\bar{y}$
$\bar{y} \sim N\left(\mu, \frac{\sigma^2}{n}\right)$	central limit theorem	it holds when n is big
$t = \frac{\bar{y}-\mu}{\frac{\sigma}{\sqrt{n}}}$	t value	standardized $\bar{y}$ $t \sim N(0,1)$ when n is big big t value rejects a hypothesis
$2Pr(T > \|t\|)$	p-value for two-tailed test	small p value rejects a hypothesis
$s^2 = \frac{\Sigma(y_i-\bar{y})^2}{n-1}$	sample variance	estimate for population variance
$s = \sqrt{s^2}$	sample standard deviation	
$\texttt{cov}(x, y)$ $= E((x-\mu_x)(y-\mu_y))$	population covariance	measure of association x and y are positively correlated if $\texttt{cov} > 0$ x and y are negatively correlated if $\texttt{cov} < 0$
$\rho = \frac{\texttt{cov}(x,y)}{\sqrt{\texttt{var}(x)\texttt{var}(y)}}$	correlation (coefficient)	$-1 \le \rho \le 1$
$y = \beta_0 + \beta_1 x + u$	simple regression	y is dependent variable x is independent variable (regressor) u is error term (other factors)
$E(y\|x) = \beta_0 + \beta_1 x$ $\beta_0 = E(y\|x=0)$ $\beta_1 = \frac{dE(y\|x)}{dx}$	PRF intercept (constant) term slope	we assume $E(u\|x) = 0$ which implies $\texttt{cov}(x, u) = 0$ (exogeneity, ceteris paribus) $\Delta E(y\|x) = \beta_1$ when $\Delta x = 1$

Some Useful Intuitions

Let c denotes a constant number, and x and y denote two random variables

1. The expectation or mean value (E or μ) means "average". It measures the central tendency of the distribution of a random variable.

2. By definition, variance is the average squared deviation:

$$\texttt{var}(x) = E[(x - \mu_x)^2].$$

   Variance is big when x varies a lot. Variance cannot be negative.

3. Since a constant has zero variation, we have $\texttt{var}(c) = 0$

4. Since variance is average squared something, we have $\texttt{var}(cx) = c^2\texttt{var}(x)$

5. We have $(a + b)^2 = a^2 + b^2 + 2ab$. Similarly we can show

$$\texttt{var}(x + y) = \texttt{var}(x) + \texttt{var}(y) + 2\texttt{cov}(x, y).$$

   Do not forget the cross product or covariance.

6. Covariance measures (linear) co-movement. Since c stays constant no matter how x moves, we have $\texttt{cov}(x, c) = 0$.

7. Formally, covariance is the average product of deviation of x from its mean and deviation of y from its mean:

$$\texttt{cov}(x, y) = E[(x - \mu_x)(y - \mu_y)].$$

   The covariance is positive if both x and y move up beyond their mean values, or both move below their mean values. The covariance is negative if one moves up, while the other moves down. In short, covariance is positive if two variables move in the same direction, while negative when they move in opposite direction.

8. Two variables are uncorrelated if covariance is zero.

9. For example, from eco201, we know price and quantity demanded are negatively correlated, while price and quantity supplied are positively correlated.

10. The link between variance and covariance is that $\texttt{cov}(x, x) = \texttt{var}(x)$

11. Variance and covariance are <u>not</u> unit-free, i.e., they can be manipulated by changing the units. For example, we have $\texttt{var}(cx) = c^2\texttt{var}(x)$ and $\texttt{cov}(cx, y) = c\texttt{cov}(x, y)$

12. By contrast, the <u>correlation coefficient</u> (ρ or $\texttt{corr}$) cannot be manipulated since it stays the same after we multiply x by c :

$$\rho_{cx,y} = \frac{\texttt{cov}(cx, y)}{\sqrt{\texttt{var}(cx)}\sqrt{\texttt{var}(y)}} = \frac{c\texttt{cov}(x, y)}{\sqrt{c^2\texttt{var}(x)}\sqrt{\texttt{var}(y)}} = \rho_{x,y}$$

13. In a similar fashion we can show the OLS estimator $\hat{\beta} = \frac{S_{xy}}{S_x^2}$ is not unit-free, so can be manipulated, while the t-value is unit-free and cannot be manipulated. That is why we want to pay more attention to the correlation coefficient and t-value.

14. We have $-\sqrt{a^2}\sqrt{b^2} \leq ab \leq \sqrt{a^2}\sqrt{b^2}$, Similarly we can show or $-\sqrt{\texttt{var}(x)}\sqrt{\texttt{var}(y)} \leq \texttt{cov}(x, y) \leq \sqrt{\texttt{var}(x)}\sqrt{\texttt{var}(y)}$, or by using the absolute value $|\texttt{cov}(x, y)| \leq \sqrt{\texttt{var}(x)}\sqrt{\texttt{var}(y)}$. This implies that

$$-1 \leq \rho_{x,y} \leq 1$$

So the correlation coefficient is unit-free, moreover, it is also bounded between minus one and one.

15. The equality holds ($\rho = 1$ or -1) only when x and y have <u>perfect linear</u> relationship $y = a + bx$. In general, the relationship is not perfectly linear so we need to add the error term: $y = a + bx + u$, then we have $-1 < \rho_{x,y} < 1$. In short, the correlation coefficient measures the degree to which two variables are linearly related.

16. The sample mean is in the middle of sample in the sense that positive deviation cancels out negative deviation. As a result,

$$\sum(x_i - \bar{x}) = 0$$

17. Treat $\bar{x}$ as constant (since it has no subscript i) when it appears in the sigma notation (summation). For instance,

$$\sum \bar{x} = n\bar{x}; \quad \sum \bar{x}x_i = \bar{x}\sum x_i$$