Review of Statistical Theory

Properties of Linear Functions

A linear function is of the form, $y = ax + b$

- The **a** variable is the slope of the line and controls its 'steepness'. A positive value has the slope going up to the right. A negative slope goes down to the right.
- The **b** variable is the intercept - the point where the line crosses the y axis.

The probability framework for statistical inference

(a) Population, random variable, and distribution

(b) Moments of a distribution (mean, variance, standard deviation, covariance, correlation)

(c) Conditional distributions and conditional means

(d) Distribution of a sample of data drawn randomly from a population: $Y_1, \ldots, Y_n$

(a) Population, random variable, and distribution

Population

- The group or collection of all possible entities of interest (school districts)
- We do not observe the entire population of interest. We select a sample.

Random variable Y

- Numerical summary of a random outcome
- The value the random variable takes when the experiment is run, is called realization.
- Example of throwing a fair die: y is 1, 2, 3, 4, 5 or 6.

(b) Moments of a population distribution: mean, variance, standard deviation, covariance, correlation

mean $= E(Y) = \sum_i f(y_i)y_i = \int yf(y)dy$

$= \mu_Y$

= long-run average value of Y over repeated realizations of Y

variance $= E(Y-\mu_Y)^2 = \sum_i f(y_i)(y_i-\mu_Y)^2 = \int (y-\mu_Y)^2 f(y)dy$

$= \sigma_Y^2$

= measure of the squared spread of the distribution

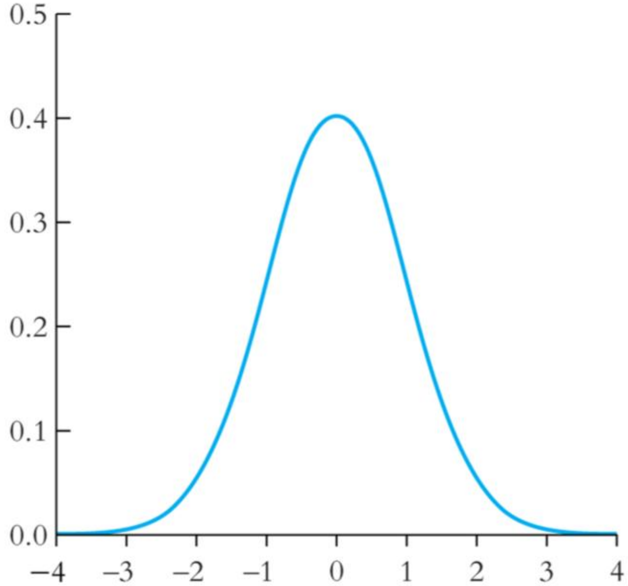
standard deviation $= \sqrt{\text{variance}} = \sigma_Y$

Moments, ctd.

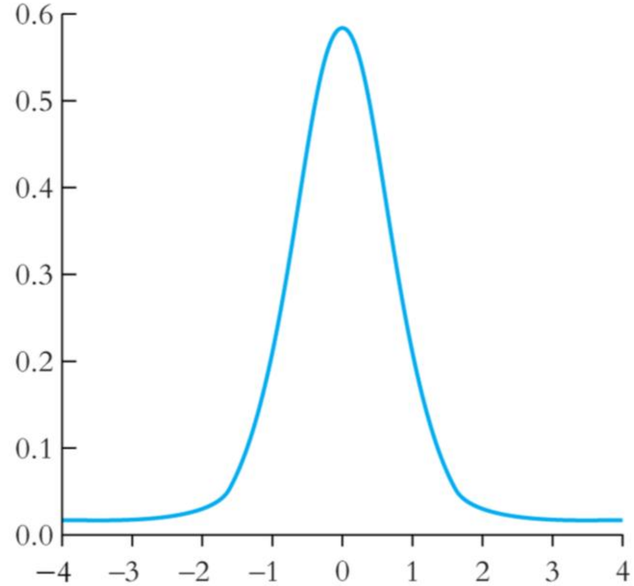
$$\textbf{\textit{skewness}} = \frac{E\left[\left(Y-\mu_Y\right)^3\right]}{\sigma_Y^3}$$

= measure of asymmetry of a distribution

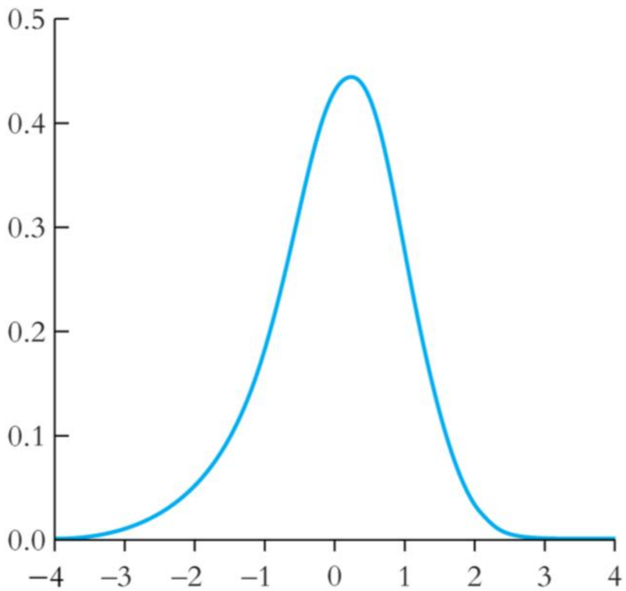
- *skewness* = 0: distribution is symmetric
- *skewness* > (<) 0: distribution has long right (left) tail



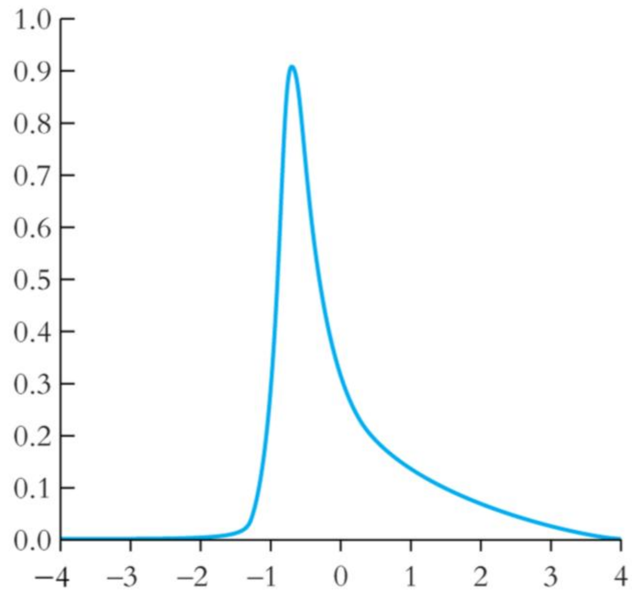
(a) Skewness = 0, kurtosis = 3



(b) Skewness = 0, kurtosis = 20



(c) Skewness = –0.1, kurtosis = 5



(d) Skewness = 0.6, kurtosis = 5

Some useful rules

I. Some properties of expected values:

1. If b is a nonstochastic (not random),

$$E(b) = b$$

2. If X and Y are two random variables,

$$E(X + Y) = E(X) + E(Y)$$

3. If a is nonstochastic,

$$E(aX) = aE(X)$$

$$E(aX + b) = aE(x) + b$$

II. Some properties of Variance:

- *Var(a)* = 0.

- If *a* and *b* are constants,

$$\begin{aligned} Var(aX+b) &= E(aX+b-(a\mu_X+b))^2 \\ &= E(aX-a\mu_X)^2 \\ &= a^2E(X-\mu_X)^2 \\ &= a^2Var(X). \end{aligned}$$

- It is easy to show that, for a constant a,

$$E(aX) = aE(X),\ Var(aX) = a^2 Var(X)$$

Interpretation of Variance

- The variance is a measure of the dispersion of the random variable around μ_X.
- The higher the variance, the less confident you are about whether the outcome will be near the mean (or expectation).

Standard deviation

- For comparison purposes it is usually more interesting to use the standard deviation, σ_X, instead. This is defined as.

$$\sigma_X = \sqrt{Var(X)}$$

- It shows how much variation or "dispersion" there is from the average (mean, or expected value).

- A low standard deviation indicates that the data points tend to be very close to the mean, whereas high standard deviation indicates that the data are spread out over a large range of values.

2 random variables: joint distributions and covariance

- When we have two random variables, the first question one may ask is whether they move together. That is when *X* is high, is *Y* high?

- The covariance is a measure of the linear association between X and Y:

$$Cov(X, Y) = E[(X - E(X))(Y - E(Y))].$$

Note that: $E(X) = \mu_X, E(Y) = \mu_Y, Var(X) = \sigma_X^2, Var(Y) = \sigma_Y^2.$

- cov(X,Y) > 0 means a positive relation between X and Y.
- The covariance between *X* and *Y* (often indicated as σ_{XY}).
- The formula $Cov(X,Y) = E(XY) - \mu_X\mu_Y$ can also be used.

Properties of covariance

$$Var(X+Y) = Var(X) + Var(Y) + 2Cov(X,Y)$$

$$Cov(X,X) = Var(X)$$

$$Cov(a_1X+b_1, a_2Y+b_2) = a_1a_2Cov(X,Y)$$

- This last property means that if we change the units of measurement of *X* and/or *Y*, the covariance changes.

- For any two constant a, b, Cov(aX,bY) = abCov(X,Y).

Correlation of random variables

- The correlation is a measure of the association between two random variables that is not affected by the unit of measurement:

$$\rho(X, Y) = \frac{Cov(X, Y)}{\sqrt{Var(X)}\sqrt{Var(Y)}}.$$

- $\rho(X, Y)$ is between -1 and 1.
- When X and Y are independent, so there is ‘no relation’ between X and Y , $\rho = 0$
- If $\rho > 0$ then X and Y go up and down together.
- If $\rho < 0$ then when X goes up, Y tends to go down and vice versa.

Examples of Correlation

Positively correlated random variables:

- Years of school, earnings
- Stock price, profit of firm

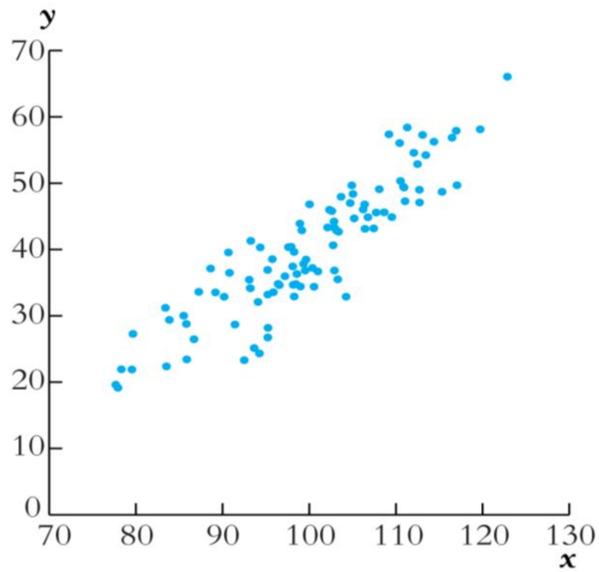
Negatively correlated random variables:

- GDP growth, unemployment
- Husband’s income, wife’s hours worked

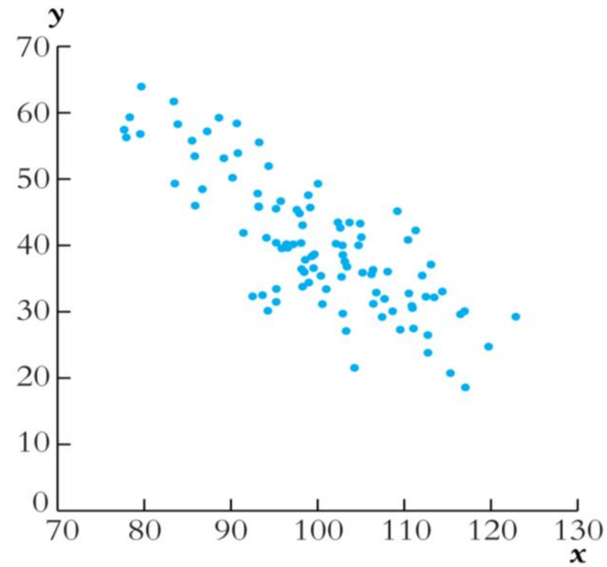
Random variables with zero or near zero correlation

- Today's weather, Number of questions asked in this class today

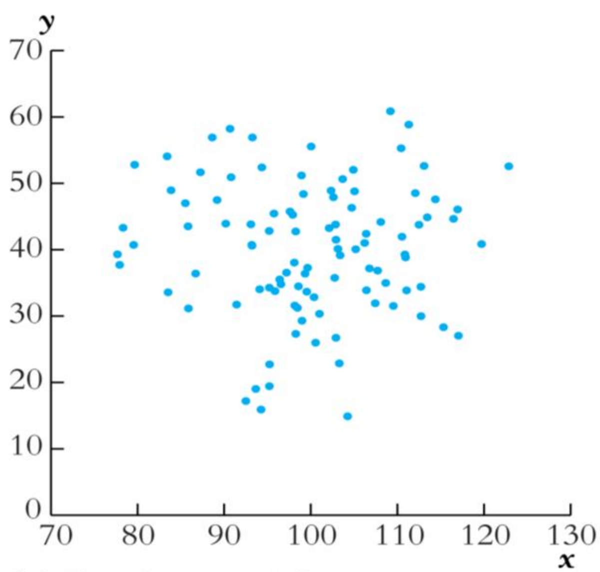
The correlation coefficient measures linear association



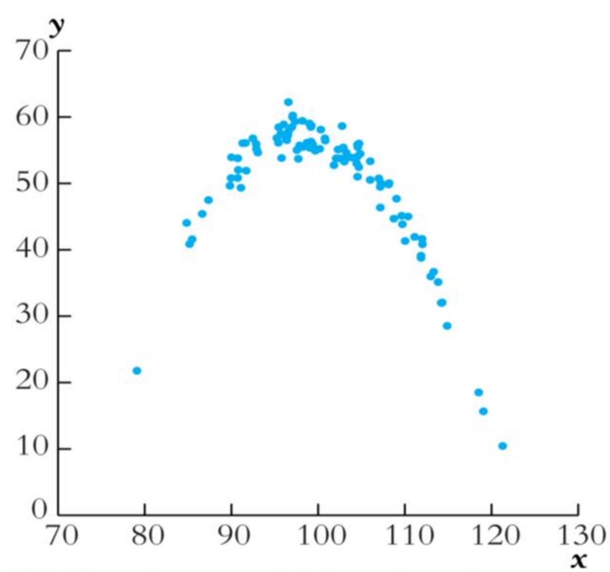
(a) Correlation = +0.9



(b) Correlation = –0.8



(c) Correlation = 0.0



(d) Correlation = 0.0 (quadratic)

Correlation versus Causality

- The difference between causation and correlation is a key concept in econometrics.

- We would like to identify causal effects and estimate their magnitude. It means we want to give a causal interpretation to the results derived from an appropriate statistical procedure.

(c) Conditional distributions and conditional means

Conditional distributions

- The distribution of X, given fixed value(s) of other random variable(s) Z.
- We obtain conditional distributions from the joint distribution of (X,Z) and the ***marginal distribution*** of Z:

$$\Pr(X = x \mid Z = z)\Pr(Z = z) = \Pr(X = x, Z = z)$$

$$\Pr(X = x \mid Z = z) = \frac{\Pr(X = x, Z = z)}{\Pr(Z = z)}$$

- The marginal distribution of Z is obtained by summing (or integrating) over all possible values of X:

$$\Pr(Z = z) = \sum_{i=1} \Pr(X = x_i, Z = z)$$

Conditional Expectation

- We define conditional expectation

$$E(Y|X)$$

<u>to mean:</u> if I ‘condition’ *X* to be some value, what is the expected value of *Y* ?

- In almost all interesting cases
  - *Y* is a random variable so after choosing *X* we don’t know exactly what *Y* will be.
  - *E(Y | X)* depends on *X*, so changing *X* will change the expected value of *Y*.

<u>Example :</u> Wages and Gender. Let *X* be the gender of an individual, we may be very interested in how wages vary with gender.

$$E(\text{wage}|\text{male})$$
$$E(\text{wage}|\text{female})$$

(d) Distribution of a sample of data drawn randomly from a population: $Y_1,\ldots, Y_n$

We will assume simple random sampling

- Choose and individual (district, entity) at random from the population

Randomness and data

- Prior to sample selection, the value of Y is random because the individual selected is random
- Once the individual is selected and the value of Y is observed, then Y is just a number – not random
- The data set is $(Y_1, Y_2,\ldots, Y_n)$, where Y_i = value of Y for the i^{th} individual (district, entity) sampled

Distribution of $Y_1,\ldots, Y_n$ under simple random sampling

- Because individuals #1 and #2 are selected at random, the value of Y_1 has no information content for Y_2. Thus:
  - Y_1 and Y_2 are ***independently distributed***
  - Y_1 and Y_2 come from the same distribution, that is, Y_1, Y_2 are ***identically distributed***
  - That is, under simple random sampling, Y_1 and Y_2 are independently and identically distributed (***i.i.d.***).
  - More generally, under simple random sampling, $\{Y_i\}$, $i = 1,\ldots, n$, are i.i.d.

Estimation

How we can estimate $\bar{Y}$ the natural estimator of the mean?

The starting point is the sampling distribution of $\bar{Y}$ …

Probability Density Functions (pdf)

- Suppose that X is a random variable that takes on J possible values x_1, ...x_J. The probability density function (pdf), f (.) of X is defined as:

$$f(x_j) = \Pr(X = x_j).$$

CDFs

- Another representation of the distribution of a random variable is the cumulative distribution function (CDF), usually denoted F (x) and defined to be the probability that X falls at or below the value x:

$$F(x) = Pr(X \leq x).$$

$P(X \leq x)$: the probability associated with the event $\{X \leq x\}$

Normal Random Variables

- Normal random variables play a big role in econometrics

- The distribution of a normal random variable depends only on its mean and variance.

- The sum of normal random variables is normal.

- We will use the notation $X \sim N(\mu, \sigma^2)$ to mean 'Random variable X is normally distributed with expected value μ and variance σ^2.

- If a random variable Z has a Normal(0,1) distribution, then we say it has a standard normal distribution.