Summation Notation

ECON 340: Economic Research Methods — Instructor: Div Bhagia

The capital sigma (Σ) stands for summing everything on the right.

$$\sum_{i=1}^{N} X_i = X_1 + X_2 + ... + X_N$$

Things you CAN do to summations:

1. Pull constants out of them, or into them.

$$\sum_{i=1}^{N} bX_i = b \sum_{i=1}^{N} X_i$$

2. Split apart (or combine) sums (addition) or differences (subtraction)

$$\sum_{i=1}^{N} (bX_i + cY_i) = b \sum_{i=1}^{N} X_i + c \sum_{i=1}^{N} Y_i$$

3. Multiply through constants by the number of terms in the summation

$$\sum_{i=1}^{N} (a + bX_i) = aN + b \sum_{i=1}^{N} X_i$$

Things you CANNOT do to summations:

1. Split apart (or combine) products (multiplication) or quotients (division).

$$\sum_{i=1}^{N} X_i Y_i \neq \sum_{i=1}^{N} X_i \times \sum_{i=1}^{N} Y_i$$

2. Move the exponent out of or into the summation.

$$\sum_{i=1}^{N} X_i^a \neq \left(\sum_{i=1}^{N} X_i\right)^a$$

Exercise:

$$X = \{2, 9, 6, 8, 11, 14\} \qquad Y = \{7, 1, 3, 5, 0\}$$

1. $\sum_{i=1}^{4} X_i = X_1 + X_2 + X_3 + X_4 = 2 + 9 + 6 + 8 = 25$

2. $\sum_{i=1}^{4} 2X_i = 2\sum_{i=1}^{4} X_i = 2 \times 25 = 50$

3. $\sum_{i=1}^{4} (X_i + 4) = \sum_{i=1}^{4} X_i + 4 \cdot 4 = 25 + 16 = 41$

4. $\sum_{i=1}^{3} (X_i + Y_i) = (X_1 + Y_1) + (X_2 + Y_2) + (X_3 + Y_3) = (2 + 7) + (9 + 1) + (6 + 3) = 28$

5. $\sum_{i=1}^{2} X_i Y_i = X_1 Y_1 + X_2 Y_2 = 2 \cdot 7 + 9 \cdot 1 = 23$

6. $\sum_{i=1}^{2} X_i \times \sum_{i=1}^{2} Y_i = (X_1 + X_2) \times (Y_1 + Y_2) = (2 + 9) \times (7 + 1) = 88$

7. $\sum_{i=1}^{2} X_i^2 = X_1^2 + X_2^2 = 4 + 81 = 85$

Handout for Lecture 2

Empirical Distribution and Measures of Central Tendency

ECON 340: Economic Research Methods — Instructor: Div Bhagia

1. You rolled a six-sided die 100 times and noted down how many times each of the six outcomes were realized. Fill in the rest of the table below:

Outcome	Count (n_k)	Relative frequency (f_k)	Cumulative frequency (F_k)
1	18	0.18	0.18
2	18	0.18	0.36
3	12	0.12	0.48
4	16	0.16	0.64
5	21	0.21	0.85
6	15	0.15	1
Total	**100**	**1**	

Note that

$$f_k = \frac{n_k}{n} = \frac{\text{observations in category } k}{\text{total observations}}$$

(a) How many times did you get a die face with a value of at most 3? 48

(b) Are the proportions close to what you would have predicted?

Yes, we would have predicted each outcome's frequency to be close to $1/6 = 0.16$

2. Find the mean and median for: $3, 4, 1, 6, 8$

$$\text{Mean} = \frac{3+4+1+6+8}{5} = \frac{22}{5} = 4.4$$

Arrange in ascending order, $1, 3, \underline{4}, 6, 8$ → median

3. Amongst the mean and the median, which one is more affected by outliers? Mean is more affected by outliers as all values are used while calculating the mean.

4. We asked a sample of 10 individuals whether they like icecream or not. We then create a variable X that takes value 1 if the individual likes icecream, and 0 otherwise. Here is the data we collected:

$$1, 1, 0, 0, 0, 1, 0, 1, 1, 1$$

(a) How many individuals like icecream in our sample?

6

(b) What proportion of individuals like icecream in our sample?

$6/10 = 0.6$

(c) Use the frequency distribution table and the following formula to calculate the mean of X.

$$\bar{X} = \frac{\sum_{k=1}^{K} n_k X_k}{n} = \sum_{k=1}^{K} f_k X_k \quad = 0.6$$

X_k	n_k	f_k	$X_k f_k$
1	6	0.6	0.6
0	4	0.4	0
			0.6

5. We have the following data on shoe sizes (X_i) for four individuals.

$$X = \{8, 6, 6, 8\}$$

(a) Calculate the mean:

$$\mu = \frac{\sum_{i=1}^{N} X_i}{N} = \frac{8+6+6+8}{4} = \frac{28}{4} = 7$$

(b) Calculate the weighted mean with weights $w = \{1, 1, 1, 1\}$.

$$\mu_{Weighted} = \frac{\sum_{i=1}^{N} w_i X_i}{\sum_{i=1}^{N} w_i} = \frac{1.8+1.6+1.6+1.8}{1+1+1+1} = 7$$

(c) Calculate the weighted mean with weights $w = \{1, 2, 2, 1\}$.

$$\mu_{Weighted} = \frac{\sum_{i=1}^{N} w_i X_i}{\sum_{i=1}^{N} w_i} = \frac{1.8+2.6+2.6+1.8}{6} = 6.66$$

(d) Calculate the weighted mean with weights $w = \{0.5, 0, 0, 0.5\}$.

$$\mu_{Weighted} = \frac{\sum_{i=1}^{N} w_i X_i}{\sum_{i=1}^{N} w_i} = \frac{0.5 \times 8 + 0.5 \times 8}{1} = 8$$

Handout for Lecture 3

Variance, Standard Deviation, Z-Score

ECON 340: Economic Research Methods — Instructor: Div Bhagia

1. We surveyed a group of 100 individuals to determine their preference for ice cream. Among the respondents, 70 individuals expressed a liking for ice cream, while the remaining 30 individuals reported not liking it. To represent this data, we introduced a variable denoted as X, assigned a value of 1 to individuals who enjoy ice cream, and a value of 0 to those who do not.

   Use the frequency distribution table to calculate the mean and variance of X. Here are the formulas you will need.

$$\bar{X} = \sum_{k=1}^{K} f_k X_k \qquad S_X^2 = \frac{n}{n-1} \sum_{k=1}^{K} f_k (X_k - \bar{X})^2$$

X_k	f_k	$f_k X_k$	$(X_k - \bar{X})^2$	$f_k (X_k - \bar{X})^2$
1	0.7	0.7	0.09	0.063
0	0.3	0	0.49	0.147
		0.7		0.21

   Answer:

$$\bar{X} = 0.7, \quad S_X^2 = \frac{100}{99} \cdot 0.21 = 0.21$$

2. You're in a statistics class with 30 students. Everyone takes the final exam, and the grades are all over the place. The average score for the class turns out to be 70, and the standard deviation is 10. You scored an 85. How did you fare relative to the class?

   Remember the Z-score formula:

   $$Z = \frac{X - \mu}{\sigma}$$

   How does your answer change if the standard deviation is 20? Why should the standard deviation affect your relative standing in the class?

   When the standard deviation is 10:

   $$Z = \frac{85 - 70}{10} = 1.5$$

   You scored 1.5 standard deviations above the class average.

   When the standard deviation is 20:

   $$Z = \frac{85 - 70}{20} = 0.75$$

   You are 0.75 standard deviations above the class average.

   The standard deviation measures how spread out the grades are around the average. A smaller standard deviation (10 in this case) implies that most students scored close to the average. Your high score then stands out more, which is reflected in the higher Z-score (1.5).

Handout for Lecture 4

Covariance and Correlation

ECON 340: Economic Research Methods — Instructor: Div Bhagia

1. Let X be the average hours of sleep per day you got last week, and let Y be the average hours you exercised per day last week. You want to look at the relationship between these two variables over the last three weeks. Given values of X and Y fill in the following table.

Week	X_i	Y_i	$(X_i - \mu_X)^2$	$(Y_i - \mu_Y)^2$	$(X_i - \mu_X)(Y_i - \mu_Y)$
1	6	0.5	4	0.01	0.2
2	9	0.3	1	0.09	-0.3
3	9	1	1	0.16	0.4
Total	24	1.8	6	0.26	0.3

Note that $\mu_X = 24/3 = 8$ and $\mu_Y = 1.8/3 = 0.6$.

(a) Calculate the variance of X and Y.

$$\sigma_X^2 = \frac{1}{N}\sum_{i=1}^{N}(X_i - \mu_X)^2 = \frac{6}{3} = 2$$

$$\sigma_Y^2 = \frac{1}{N}\sum_{i=1}^{N}(Y_i - \mu_Y)^2 = \frac{0.26}{3} = 0.087$$

(b) Calculate the covariance and correlation between X and Y.

Covariance:

$$\sigma_{XY} = \frac{1}{N}\sum_{i=1}^{N}(X_i - \mu_X)(Y_i - \mu_Y) = \frac{0.3}{3} = 0.1$$

Correlation:

$$\rho_{XY} = \frac{\sigma_{XY}}{\sigma_X \sigma_Y} = \frac{0.1}{\sqrt{2}\sqrt{0.087}} = 0.24$$

(c) In class, we learned that covariance is positive when two variables move together, meaning that they increase or decrease together. Can you explain how the formula you used in (c) ensures that this is the case? Explain it to your peer.

While calculating the covariance, we add a positive number to the numerator in the formula whenever both variables are either above or below their respective averages. Conversely, we add a negative number to the numerator when one variable is above its average while the other is below. Therefore, a positive covariance indicates that, on average, the variables move together—either both being above or both being below their averages. While a negative covariance indicates that, on average, the variables move in opposite directions—one being above its average while the other is below.

Now say instead of recording the exercise in hours, you had recorded it in minutes. Then your data would look as below, where Z is the average minutes of exercise per day.

Week	X_i	Z_i	$(X_i - \mu_X)^2$	$(Z_i - \mu_Z)^2$	$(X_i - \mu_X)(Z_i - \mu_Z)$
1	6	30	4	36	12
2	9	18	1	324	-18
3	9	60	1	576	24
Total	26	108	6	936	18

(d) Do you think the covariance between sleep and exercise is going to be larger, smaller or the same now that exercise is measured in minutes instead of hours?

The covariance between sleep and exercise will be larger when exercise is measured in minutes instead of hours. This is because covariance is sensitive to the scale of the variables.

(e) Fill in the table above and calculate the covariance and correlation between X and Z.

Note that $\mu_Z = 108/3 = 36$ and $\sigma_Z^2 = 936/3 = 312$.

$$\text{Covariance: } \sigma_{XZ} = \frac{18}{3} = 6 \qquad \text{Correlation: } \rho_{XZ} = \frac{6}{\sqrt{312}\sqrt{2}} = 0.24$$

2. If a study finds a strong positive correlation between the number of houses and house prices across US cities, can we conclude that more housing supply leads to higher house prices? Why or why not? Discuss.

If we observe a strong positive correlation between the number of houses and house prices across US cities, it does not necessarily imply that an increase in housing supply causes higher house prices. In fact, it is possible that higher house prices lead to more housing supply, as builders try to maximize their profits by building in more profitable locations, resulting in a positive correlation between the stock of housing and prices. This is known as reverse causality.

It is also possible that external confounding factors are responsible for the observed positive correlation between housing stock and housing prices. For instance, if certain cities have more desirable amenities, more people may want to live there, leading to higher demand for housing. This, in turn, can result in higher prices and increased housing stock as builders construct more houses to meet the growing demand.

Handout for Lecture 9

Distribution, Expectation, Variance

ECON 340: Economic Research Methods — Instructor: Div Bhagia

X is a random variable.

- Expectation of X, $\mu_X = E(X) = \sum_x xf(x)$
- Variance of X, $\sigma_X^2 = Var(X) = E[(X - \mu_X)^2] = \sum_x (x - \mu_X)^2 f(x)$
- Standard deviation of X, $\sigma_X = \sqrt{\sigma_X^2}$

If X is a random variable and $Y = a + bX$, then Y is also a random variable with

- $E(Y) = a + bE(X)$
- $Var(Y) = b^2 Var(X)$

You are at a fair and considering playing the following game — flip a coin, if you get heads, you gain \$10, else you lose \$10. Denote X as your winnings/loss from the game.

1. Find the expected value, variance, and standard deviation of X.

x	$f(x)$	$xf(x)$	$(x - \mu_X)^2$	$f(x)(x - \mu_X)^2$
10	0.5	5	10^2	50
-10	0.5	-5	$(-10)^2$	50
		0		100

Answer:

$$\mu_X = 0, \quad \sigma_X^2 = 100, \quad \sigma_X = 10$$

2. You look up and realize that you have to pay \$5 in order to play the game. So your actual winnings/loss from the game will be $Y = X - 5$. Find the expected value, variance, and standard deviation of Y.

y	$f(y)$	$yf(y)$	$(y-\mu_Y)^2$	$f(y)(y-\mu_Y)^2$
5	0.5	2.5	$(5-(-5))^2$	50
-15	0.5	-7.5	$(-15-(-5))^2$	50
		-5		100

Answer:

$$\mu_Y = -5, \quad \sigma_Y^2 = 100, \quad \sigma_Y = 10$$

Note that, $E(Y) = E(X) - 5$ and $Var(Y) = Var(X)$.

3. You see another stall offering a lower stakes game — flip a coin, if you get heads, you gain \$5, else you lose \$5. Your winnings/loss from this game will be $Z = 0.5X$. Find the expected value, variance, and standard deviation of Z.

z	$f(z)$	$zf(z)$	$(z-\mu_Z)^2$	$f(z)(z-\mu_Z)^2$
5	0.5	2.5	5^2	12.5
-5	0.5	-2.5	$(-5)^2$	12.5
		0		25

Answer:

$$\mu_Z = 0, \quad \sigma_Z^2 = 25, \quad \sigma_Z = 5$$

Note that, $E(Z) = E(X)$, $Var(Z) = (0.5)^2 Var(X)$, and $\sigma_Z = 0.5\sigma_X$.

Handout for Lecture 10

Normal Distribution and Z-Score

ECON 340: Economic Research Methods — Instructor: Div Bhagia

If $X \sim N(\mu, \sigma^2)$, then the standardized random variable,

$$Z = \frac{X - \mu}{\sigma} \sim N(0, 1)$$

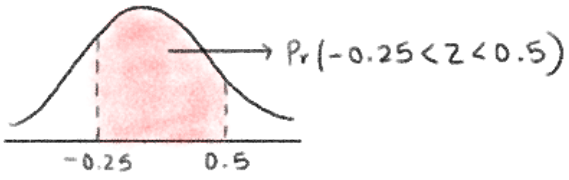
Given $X \sim N(\mu, \sigma^2)$, to find $Pr(x_0 < X < x_1)$:

- Find $z_0 = (x_0 - \mu)/\sigma$ and $z_1 = (x_1 - \mu)/\sigma$
- Use standard normal table to find $Pr(z_0 < Z < z_1)$

Exercises: Refer to the standard normal table to answer the following.

1. Given $X \sim N(3, 16)$, find $Pr(2 < X < 5)$.

$Pr(2<X<5) = Pr\left(\frac{2-3}{4} < Z < \frac{5-3}{4}\right) = Pr(-0.25<Z<0.5)$



From the standard Normal table:

$Pr(Z<-0.25) = 0.4013$

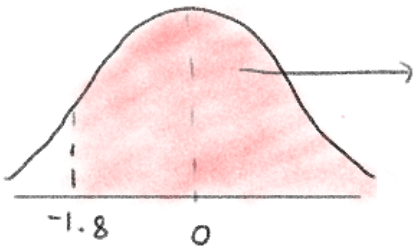
$Pr(Z<-0.5) = Pr(Z>0.5) = 0.3085$

So, $Pr(-0.25<Z<0.5) = 1 - 0.4013 - 0.3085$

$= \boxed{0.2902}$

2. Given $X \sim N(15, 100)$, find $Pr(X > -3)$.

$Pr(X>-3) = Pr\left(\frac{X-15}{10} > \frac{-3-15}{10}\right) = Pr(Z>-1.8)$



$Pr(Z>-1.8) = 1 - Pr(Z<-1.8)$

$= 1 - 0.0359$

$= \boxed{0.9641}$

Got this from the standard normal table

Alternatively, $Pr(Z>-1.8) = Pr(Z<1.8)$

$= 0.9641$

Given $X \sim N(\mu, \sigma^2)$ and $Pr(X < x) = p$, to find x :

- Use standard normal table to find z where $Pr(Z < z) = p$
- Find $x = \mu + z \cdot \sigma$

Follows analogously for when we are given $Pr(X > x) = p$.

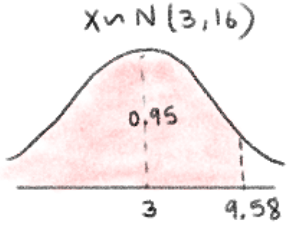
Exercises: Refer to the standard normal table to answer the following.

1. Given $Pr(Z > z) = 0.95$. Find z.

   From the standard normal table, $Pr(z < 1.645) = 0.95$
   Since the normal distribution is symmetric, $Pr(z > -1.645) = 0.95$
   So, $z = -1.645$

2. Given $X \sim N(3, 16)$ and $Pr(X < x) = 0.95$. Find x.

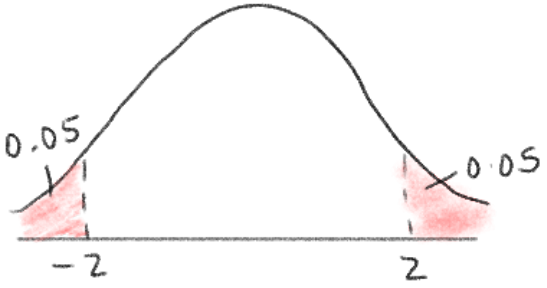
   From the standard normal table,
   $Pr(z < 1.645) = 0.95$
   Since, $z = \frac{x - \mu}{\sigma} \rightarrow x = \mu + z\sigma$
   $x = 3 + 1.645 \times 4 = 9.58$

3. Given $Pr(|Z| > z) = \sout{0.90}$ 0.10 (typo). Find z.

   Note: Since the normal distribution is symmetric $Pr(Z > z) = Pr(Z < -z)$, so we have that: $Pr(|Z| > z) = 2Pr(Z > z) = 2Pr(Z < -z)$.

   $Pr(|z| > z) = Pr(z < -z) + Pr(z > z) = 2\,Pr(z < -z)$
   Since $Pr(|z| > z) = 0.1$, trying to find z such that
   $Pr(z < -z) = 0.05$
   From standard normal table,
   $z = 1.645$

Handout for Lecture 11
Independence and Correlation

ECON 340: Economic Research Methods — Instructor: Div Bhagia

Consider two random variables X and Y.

- The *joint probability* $f(x,y) = Pr(X = x, Y = y)$ represents the likelihood that X equals x and Y equals y.
- The marginal probability of $Y = y$, denoted $f(y)$, is obtained by summing the joint probability $f(x,y) = Pr(X = x, Y = y)$ over all possible values of x.
- The *conditional probability* $f(y|x) = Pr(Y = y|X = x)$ represents the likelihood that Y is equal to y, given that X is equal to x.

$$f(y|x) = Pr(Y = y|X = x) = \frac{Pr(X = x, Y = y)}{Pr(X = x)} = \frac{f(x,y)}{f(x)}$$

- The *conditional expectation* $E(Y|x)$ is the expected value of Y given that $X = x$.

$$E(Y|X = x) = \sum_y yPr(Y = y|X = x) = \sum_y yf(y|x)$$

Uncorrelated vs Independent

- Covariance and correlation:

$$\sigma_{XY} = Cov(X,Y) = E[(X - \mu_X)(Y - \mu_Y)]$$

$$\rho_{XY} = corr(X,Y) = \frac{\sigma_{XY}}{\sigma_X \sigma_Y} \quad \text{where } -1 \leq \rho \leq 1$$

- Two random variables are *uncorrelated* if $corr(X,Y) = 0$.
- Two random variables are *independent* if $f(y|x) = f(y)$ for all x and y or equivalently $E(Y|X) = E(Y)$.
- If X and Y are independent, then they are also uncorrelated. (The converse is not necessarily true.)

1. The following table gives the joint-probability distribution of rain (X) *and* commute time in minutes (Y).

	Rain $(X = 1)$	No Rain $(X = 0)$	Total
60-min commute $(Y = 60)$	0.3	0.2	**0.5**
30-min commute $(Y = 30)$	0.1	0.4	**0.5**
Total	**0.4**	**0.6**	**1**

(a) Fill the marginal probabilities in the column and row labeled *Total*. For example, the first row in column *Total* should contain the marginal probability of having a 60-minute commute $(Pr(Y = 60))$.

(b) Find the following conditional probabilities:

- Probability of having a 60-min commute conditional on *raining*

$$Pr(Y = 60|X = 1) = \frac{Pr(Y = 60, X = 1)}{Pr(X = 1)} = \frac{0.3}{0.4} = \frac{3}{4}$$

- Probability of having a 60-min commute conditional on *not raining*

$$Pr(Y = 60|X = 0) = \frac{Pr(Y = 60, X = 0)}{Pr(X = 0)} = \frac{0.2}{0.6} = \frac{1}{3}$$

(c) Calculate $E(Y|X = 1)$ and $E(Y|X = 0)$.

$$\begin{aligned} E(Y|X = 1) &= \sum_y yPr(Y = y|X = 1) \\ &= 60 \cdot Pr(Y = 60|X = 1) + 30 \cdot Pr(Y = 30|X = 1) \\ &= 60 \cdot \frac{3}{4} + 30 \cdot \frac{1}{4} = 52.5 \end{aligned}$$

$$\begin{aligned} E(Y|X=0) &= \sum_y yPr(Y=y|X=0) \\ &= 60 \cdot Pr(Y=60|X=0) + 30 \cdot Pr(Y=30|X=0) \\ &= 60 \cdot \frac{1}{3} + 30 \cdot \frac{2}{3} = 40 \end{aligned}$$

(d) How does rain impact the expected commute time in this example?

Rain increases the expected commute time by 12.5 minutes.

2. You flipped a coin six times and got tails each time. The likelihood of getting a head in your seventh flip is

- ☒ 1/2
- ☐ More than 1/2
- ☐ Less than 1/2

3. Note that for sums of two random variables X and Y:

$$E(aX + bY) = aE(X) + bE(Y)$$

$$Var(aX + bY) = a^2Var(X) + b^2Var(Y) + 2abCov(X, Y)$$

In the above expressions, a and b are constants.

You are contemplating investing in two stocks with the same average return and spread.

$$\mu_X = \mu_Y = \mu \qquad \sigma_X^2 = \sigma_Y^2 = \sigma^2$$

Should you pick any one stock at random or invest equally in both?

Consider a new random variable W that represents the return from investing equally in both stocks:

$$W = 0.5X + 0.5Y.$$

Using the formula above

$$E(W) = 0.5E(X) + 0.5E(Y) = 0.5\mu + 0.5\mu = \mu.$$

The expected return from investing equally in both stocks is the same as the return from investing in individual stocks.

Using the formula for the variance of the sum of two random variables, we have

$$\begin{aligned}\text{Var}(W) &= 0.5^2\text{Var}(X) + 0.5^2\text{Var}(Y) + 2 \times 0.5 \times 0.5 \times \text{Cov}(X, Y) \\ &= 0.25\sigma^2 + 0.25\sigma^2 + 0.5\text{Cov}(X, Y) \\ &= 0.5\sigma^2 + 0.5\text{Cov}(X, Y).\end{aligned}$$

If X and Y are uncorrelated, then $\text{Cov}(X, Y) = 0$, and $\text{Var}(W) = 0.5\sigma^2$, which is lower than the variance of individual stocks. In this case, it would be better to invest equally in both stocks to minimize risk.

If X and Y are negatively correlated, i.e., $\text{Cov}(X, Y) < 0$, then $\text{Var}(W)$ would be even lower than $0.5\sigma^2$, making the portfolio considerably less risky. Therefore, if the stocks are negatively correlated, diversifying by investing equally in both would be an even more advantageous strategy to minimize risk.

The only case where investing equally in both stocks would not offer a risk-reduction advantage is when the stocks are perfectly positively correlated. Specifically, if $\text{Corr}(X, Y) = 1$, then $\text{Cov}(X, Y) = \sigma^2$ and the variance of W would equal σ^2, the same as the variance for individual stocks. In such a scenario, diversifying by investing in both stocks would offer no risk-minimizing benefits, making us indifferent between the two investment options.

Handout for Lecture 12

Good Estimators, Sample Mean Distribution, and Confidence Intervals

ECON 340: Economic Research Methods — Instructor: Div Bhagia

Good Estimators

Denote $\hat{\theta}$ as an estimator for the population parameter θ. Some desirable properties for an estimator

- *Unbiasedness*: $E(\hat{\theta}) = \theta$
- *Efficiency*: lower variance is better
- *Consistency*: as the sample size becomes infinitely large, $\hat{\theta} \to \theta$

Question 1: If some sample estimator $\hat{\theta}$ is an unbiased estimator for the true population parameter θ i.e. $E(\hat{\theta}) = \theta$. This implies that:

- ☐ $\hat{\theta} = \theta$ in all samples.
- ☒ If we take repeated samples, average of $\hat{\theta}$ is equal to θ

Question 2: We are choosing between two estimators $\hat{\theta}_1$ and $\hat{\theta}_2$, both of which are unbiased i.e. $E(\hat{\theta}_1) = \mu$ and $E(\hat{\theta}_2) = \mu$. But the variance of $\hat{\theta}_1$ is lower than that of $\hat{\theta}_2$ i.e. $Var(\hat{\theta}_1) < Var(\hat{\theta}_2)$. Which of the following is true?

- ☐ We are indifferent between the two estimators.
- ☒ We prefer $\hat{\theta}_1$ over $\hat{\theta}_2$.
- ☐ We prefer $\hat{\theta}_2$ over $\hat{\theta}_1$.
- ☐ We need more information to reach any conclusion.

Sample Mean Distribution

Let $X_1, X_2, ..., X_n$ denote independent random draws (random sample) from a population with mean μ and variance σ^2. Then the sample mean $\bar{X} = \frac{1}{n}\sum_{i=1}^{n} X_i$ is a random variable with:

$$E(\bar{X}) = \mu \qquad Var(\bar{X}) = \frac{\sigma^2}{n}$$

In addition, the distribution of the sample mean is *normal* if *either* of the following is true:

- The underlying population is normal
- The sample size is large, say $n \geq 100$

Given the variance of the sample mean as $\sigma^2_{\bar{X}} = \frac{\sigma^2}{n}$, its standard deviation, commonly referred to as the *standard error*, is $\sigma_{\bar{X}} = \frac{\sigma}{\sqrt{n}}$.

Question 3: If the average of hourly wages in the population is $\mu = \$30$ and the variance of hourly wages is $\sigma^2 = 16$. Then what is the expected value, variance, standard error, and distribution of the sample mean estimator for a sample size of 400?

$$E(\bar{X}) = 30, \quad \sigma^2_{\bar{X}} = \frac{16}{400} = 0.04, \quad \sigma_{\bar{X}} = \sqrt{0.04} = 0.2$$

Since $n \geq 100$, by Central Limit Theorem $\bar{X}$ is normally distributed.

Question 4: You are interested in the average starting salary of CSUF graduates and are considering taking a random sample of 120 students. I advise you to take as large of a sample as feasible. This is sound advice because taking an even larger sample would ensure that

- ☐ The sample average $\bar{x} = \mu$
- ☐ The sample average $\bar{x}$ is drawn from a normal distribution
- ☒ The sample average $\bar{x}$ is drawn from a distribution with lower variance

Note: I am using $\bar{x}$ to denote a realization of $\bar{X}$.

Question 5: Can you explain intuitively why the variance of the sample mean increases with σ^2 and decreases with n?

Why does the variance of the sample mean decrease with n?

Suppose you aim to determine the average test score for all students at a university. When you use smaller samples, you could encounter significant sample-to-sample variability. For example, one sample might consist of students who performed exceptionally well, while another might include students who scored poorly. In contrast, a larger sample is more likely to accurately represent the overall student population, thereby reducing the variability between different samples.

Why does the variance of the sample mean increase with σ^2?

If the range of scores is quite wide, one sample could consist of students who performed exceptionally well, leading to a high sample mean, while another might include students who scored poorly, resulting in a low sample mean. However, if there is little to no variation in the student test scores across the university, you are likely to obtain similar sample means across different samples.

Confidence Intervals

Let $z_{\alpha/2}$ be the z-value that leaves area $\alpha/2$ in the upper tail of the normal distribution. Then $1-\alpha$ confidence interval is given by

$$\bar{x} \pm \underbrace{z_{\alpha/2}\frac{\sigma}{\sqrt{n}}}_{\text{Margin of Error}}$$

Question 6: Continuing with Question 3, say you took a sample of 400 individuals and found the average hourly wages in your sample of $\bar{x} = 26$. Create a 95% confidence interval for the true population mean.

Note that here $1-\alpha = 0.95$, so $\alpha/2 = 0.025$. From the Standard Normal Table, $z_{0.025} = 1.96$. In which case, the 95% confidence interval is given by:

$$26 \pm 1.96 \times 0.2 = [25.6, 26.4]$$

Handout for Lecture 13

Confidence Intervals

ECON 340: Economic Research Methods — Instructor: Div Bhagia

How to construct a confidence interval?

Known population variance: $1-\alpha$ confidence interval for the population mean μ:

$$\bar{x} \pm z_{\alpha/2} \cdot \frac{\sigma}{\sqrt{n}}$$

where $z_{\alpha/2}$ is the z-value that leaves area $\alpha/2$ in the upper tail of the standard normal distribution.

Unknown population variance: $1-\alpha$ confidence interval for the population mean μ:

$$\bar{x} \pm t_{n-1,\alpha/2} \frac{S}{\sqrt{n}}$$

where $t_{n-1,\alpha/2}$ is the t-value that leaves area $\alpha/2$ in the upper tail of the t-distribution. $n-1$ is the degrees of freedom.

Note: Since the t distribution looks just like the standard normal for large n, for $n \geq 100$ you can continue using the standard normal table.

Question: A car manufacturer wants to estimate the mean CO2 emissions of a new model of car. A sample of 196 cars is randomly selected and their CO2 emissions are measured. The sample mean and standard deviation are 120 g/km and 20 g/km, respectively. Construct a 95% confidence interval for the true mean CO2 emissions of this car model. (Note: $Pr(Z > 1.96) = 0.025$.)

Answer: We are given:

$$\bar{x} = 120 \text{ g/km} \quad \text{(sample mean)}$$
$$S = 20 \text{ g/km} \quad \text{(sample standard deviation)}$$
$$n = 196 \quad \text{(sample size)}$$

We can use the following formula to create a confidence interval:

$$\bar{x} \pm t_{n-1,\alpha/2} \frac{S}{\sqrt{n}}$$

Since we want to create a 95% confidence interval, here $1 - \alpha = 0.95$. In which case, the T statistic we want is $t_{195,0.025}$. However, since the degrees of freedom are large enough, $t_{195,0.025} \approx z_{0.025} = 1.96$. So the 95% confidence interval is given by:

$$120 \pm 1.96 \left(\frac{20}{\sqrt{196}} \right) = [117.2, 122.8]$$

Therefore, we are 95% confident that the true mean CO2 emission of this car model is between 117.2 g/km and 122.8 g/km.