ECO239 Statistics I

Week 5

Measures of Relationship between Variables

- Covariance
- Correlation Coefficient

[start here]Covariance

- A measure of the linear relationship between two variables
- Only concerned with the direction of the relationship.

Population Covariance

$$\mathrm{Cov}(X, Y) = \sigma_{xy} = \frac{\sum_{i=1}^{N}(X_i - \mu_x)(Y_i - \mu_Y)}{N}$$

Sample Covariance

$$COV(x, y) = S_{xy} = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{n - 1}$$

Covariance: Meaning

Cov(x,y) > 0 => X and Y tend to move in the same direction.

Cov(x,y) < 0 => X and Y tend to move in the opposite direction.

Cov(x,y)= 0 => X and Y are independent

Practice: Calculate COV(X,Y)

X (# Workers)	Y (# Cell phones)			
12	20			
30	60			
15	27			
24	50			
14	23			

X (# Workers)	Y (# Cell phones)	X-X_bar	Y- Y_bar	(X-X_bar)* (Y-Y_bar)
12	20	(12-19)=-7	(20-36)=-16	(-7)*(-16)=112
30	60	(30-19)=11	(60-36)=24	(11)*(24)=264
15	27	(15-19) = -4	(27-36)=-9	(-4)*(-9)=36
24	50	(24-19)=5	(50-36)=14	(5)*(14)=70
14	23	(14-19)=-5	(23-36)=-13	(-5)*(-13)=65
X_bar=19	Y_bar = 36			SUM = 547

COV(x,y)=547/(5-1)=136,75 (Positive Relationship)

Correlation Coefficient

- Measures the relative strength and direction of the linear relationship between two variables.

Population Correlation Coefficient

$$\rho = \frac{\mathrm{Cov}(X,Y)}{\sigma_X \sigma_Y}$$

Population Covariance

Standard Deviation of X, and Y.

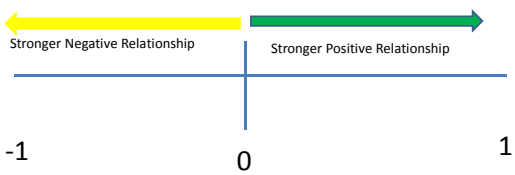
Sample Correlation Coefficient

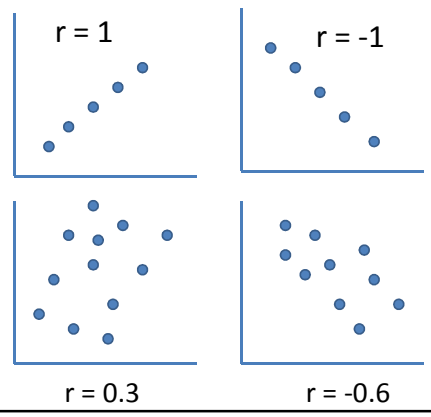
$$r = \frac{\mathrm{Cov}(x,y)}{s_X s_Y}$$

Sample Covariance

Standard Deviation of x and y.

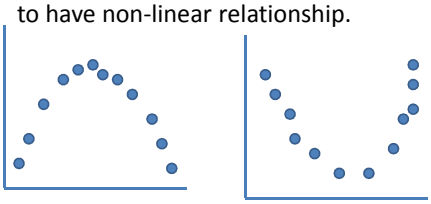
Correlation Coefficient: Meaning





Correlation Coefficient: Caution!

- This coefficient measures LINEAR relationship, not Non-Linear.
- Even if r = 0, it is possible for x and y variables to have non-linear relationship.



Calculate Correlation Coefficient

X (# Workers)	Y (# Cell phones)
12	20
30	60
15	27
24	50
14	23
X_bar=19	Y_bar = 36

- COV(x,y) = 136.75.

$$r = \frac{\mathrm{Cov}(x,y)}{s_X s_Y}$$

X	Y	X-X_bar	Y- Y_bar	(X-X_bar)^2	(Y-Y_bar)^2
12	20	(12-19)=-7	(20-36)=-16	(-7)^2 = 49	(-16)^2=256
30	60	(30-19)=11	(60-36)=24	(11)^2 = 121	(24)^2=576
15	27	(15-19) = -4	(27-36)=-9	(-4)^2=16	(-9)^2=81
24	50	(24-19)=5	(50-36)=14	(5)^2 =25	(14)^2=196
14	23	(14-19)=-5	(23-36)=-13	(-5)^2=25	(-13)^2=169
X_bar =19	Y_bar = 36			SUM=236	SUM=1278

r= 136.75/(sqrt(236/4)*sqrt(1278/4))
= 136.75/(7.68*17.87)
=0.996

- examscore

Coefficient of Variation

- A measure of relative variation
- Standard deviation as a percentage of the mean
- In %. => can compare multiple data measured in different units.

$$CV = \left(\frac{S}{\bar{X}}\right) * 100\% \quad \text{if } \bar{X} > 0$$

where s: standard deviation, $\bar{X}$: mean.

Which stock is the most risky one?

	Stock A	Stock B	Stock C
Average price last year	50 TL	100 USD	100 Euro
St.dev.	5 TL	20 USD	5 Euro
CV			

Which stock is the most risky?

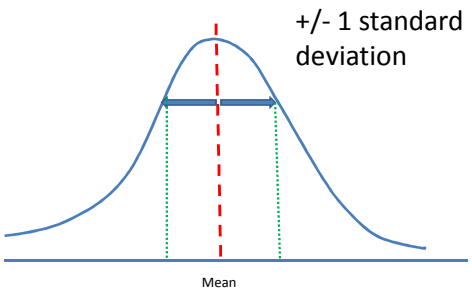
	Stock A	Stock B	Stock C
Average price	50 TL	100 USD	100 Euro
St.dev.	5 TL	20 USD	5 Euro
CV	(5/50) *100% =10%	(20/100) *100% =20%	(5/100)* 100% =5%

Chebyshev's Theorem

- Answers the question "How much percentage of observations can be found in the interval $\mu \pm k\sigma$?"

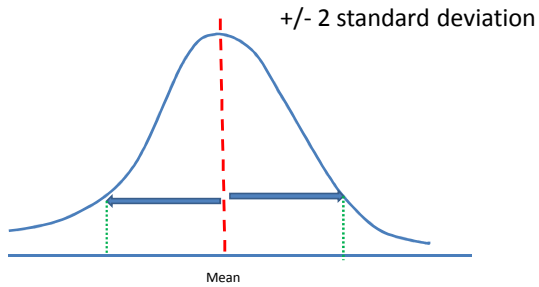
Example
examscore

More generally…

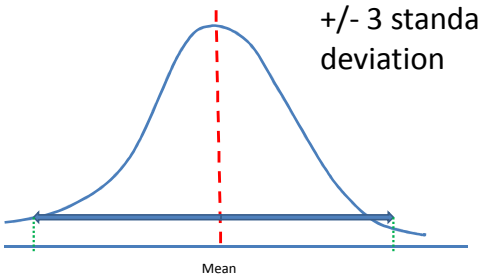


Chevyshev's theorem does not work for this case, k = 1.

More generally…

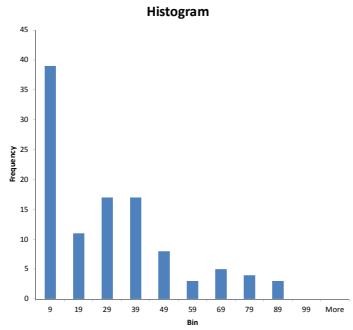


More generally…



How much % of students are included within +/- 1, 2, or 3 standard deviation?

mean	24.44
stdev	23.27
mean + 1 stdev	47.71
mean - 1 stdev	1.17
mean + 2 stdev	70.9728
mean + 3 stdev	94.2396



Chebyshev's Theorem

- For any mean and standard deviation, and **k >1**, the % of observation that fall within the interval $\mu \pm k\sigma$ is at least

$$100\left[1-\left(\frac{1}{k^2}\right)\right]\%$$

- Regardless of how the data are distributed.
- Does not work for k = 1.

Within	At least
K=2 (mean +/- 2 stdev)	(1-(1/(2^2))*100% = 75%
K=3 (mean +/- 3 stdev)	(1-(1/(3^2))*100% = 89%

- Does not work for k = 1.
- K does not have to be integers.

Chebyshev's Theorem

- **Advantage**: Applicable to any population & distributional shapes.
- **Disadvantage**: In reality, distributions are relatively close to symmetric, and % of observations in a specific range is much higher.

Practice

- A large class with 280 students.
- Midterm exam result: mean = 74, stdev=6.
- At least how many students scored between 50 and 98 according to Chevyshev's Theorem?

- (74+k*6)=98
- K=(98-74)/6 = 4
- (1-(1/4^2))*100% = 0.9375*100% = 93.75%
- 280*0.9375=262.5 or at least 263 students.

If stdev = 8, instead of 6,

- At least how much % of students are included in the same range (50 & 98) ?
- Do you think it's more /less than the previous question? And WHY?

- (74+k*8)=98
- K=(98-74)/8 = 3
- (1-(1/3^2))*100% = 88.9%
- Less # of students are included in the same range.

Practice

- In company A, the average salary is 6000 TL with standard deviation of 1200 TL.
- According to Chevyshev's theorem, what is the interval in which at least 80% of the salaries lie?

- (1-(1/k^2))=0.8 => (1/k^2)=0.2 => k^2=5, k=sqrt(5).
- 6000+sqrt(5)*1200 = 8683.
- 6000-sqrt(5)*1200 = 3317.
- 80% receives between 3317 and 8683 TL.

Empirical Rule

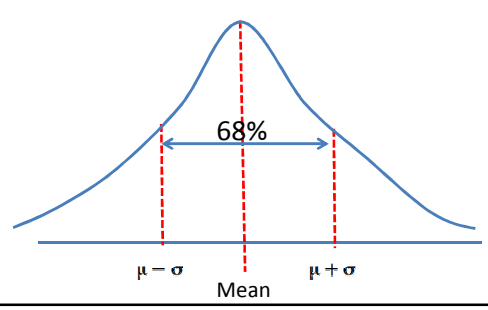
- If the data distribution is bell-shaped,

$\mu \pm \sigma$ contains about 68% of observations

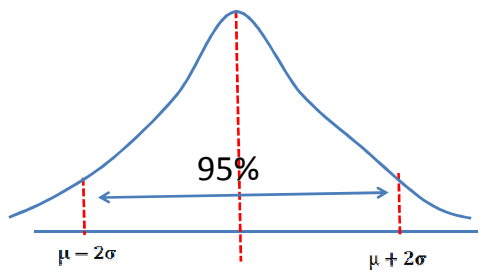
$\mu \pm 2\sigma$ contains about 95% of observations

$\mu \pm 3\sigma$ contains about 99.7% of observations

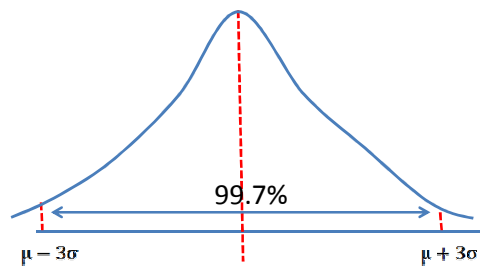
$\mu \pm \sigma$ contains about 68% of observations



$\mu \pm 2\sigma$ contains about 95% of observations



$\mu \pm 3\sigma$ contains 99.7% of observations



Practice

- n = 280 students
- μ = 74, σ = 6.

Q1: How many students scored between 62 and 86 according to Empirical Rule?

Q2: If you score 92, you are in top _____ %.

Practice

- n = 280 students
- μ = 74, σ = 6.

Q1: How many students scored between 62 and 86 according to Empirical Rule?

74 – k*6 = 62 => k = (74-62)/6 = 2. 2 st.dev => 95% => 280*0.95= about 266 students.

Q2: If you score 92, you are in top _____ %.

92 = 74+k*6 => k = (92-74)/6 = 3. 3 st.dev. => 99.7%. By assuming perfect symmetry, 0.3%/2 or 0.15%.

Empirical vs. Chevyshev

- Compare this result with Chevyshev's theorem.
- n = 280 students; mean= 74, stdev= 6.

Q1: How many students scored between 62 and 86 according to Empirical Rule?

74 – k*6 = 62 => k = (74-62)/6 = 2. 2 st.dev => 95% => 280*0.95= about 266 students.

Q1': How many students scored between 62 and 86 according to Chevyshev's theorem?

K = 2 => (1-(1/2^2))*100% = (3/4)*100% = 75%. => 280*0.75 = 210 students.

Practice

- Average bill in Quick China, mean = 55 TL,st.dev. = 8.3 TL.

Q: 99.7% of the time you expect your bill to be between [ ] and [ ] TL according to Empirical Rule

99.7% => k = 3 => 55 +/- 3*8.3 = [30.1 79.9]

Q:95% of the time you expect your bill to be between [ ] and [ ] TL according to Empirical Rule.

95% =>k = 2 => 55+/- 2*8.3 = [38.4 71.6].

Q: According to Chevyshev's Theorem, 95% will be included within [,] range.

(1-(1/k^2))=0.95 => K^2 = 1/0.05 => k= $\sqrt{20}$ = about 4.47.
55-4.47*8.3 = 92.1. 55+4.47*8.3 = 17.88.

Quiz 3 (Nov.1.2016)

Sample Data: {0, 4, 14, 25, 32}

Q1. Calculate sample mean. = 15 (0.25 point)

Q2. Calculate sample variance. = 184 (0.5 point)

Q3. Calculate sample standard deviation. = 13.56 (0.25 point)

Sqrt not calculated = -0.05.