$\rho(X,Y)$ close to 1 :

4. If the covariance and correlation are ≈ 0, then there is ____________________ between X and Y.

5. For any two random variables X and Y, $-1 \le \rho \le 1$. ✓

6. If a and c are either both positive or both negative, and any b & d :

corr(X,Y) ⇒ $\rho(aX+b, cY+d) = \rho(X,Y)$
(if $ac<0$ then $\rho(aX+b, cY+d) = -\rho(X,Y)$)

7. If X and Y are independent or uncorrelated, then $\rho = 0$, but $\rho = 0$ does not imply independence. It just means that there is no **linear** association between X and Y. But it can also mean that X and Y may have a non-linear association. → →

8. $\rho = 1$ or -1 if and only if $Y = aX + b$ for some numbers a and b with $a \neq 0$.

Lec23 ⚠ correlation $\not\Rightarrow$ causation

5.3 Statistics and Their Distributions

Definition 28. The random variables $X_1, X_2, \cdots, X_n$ are said to form a (simple) **random sample** of size n if

1. identical distribution
2. independent

Consider taking a random sample from a population, and compute the sample mean, $\overline{y}$, for the observations.

- Because the sample is random, the observations will also be random.

  Hence, $\overline{y}$ will also be random.
- Because $\overline{y}$ is random, it has a mass fn / pdf & so mean & variance / st dev associated with it. This distribution plays an important role in drawing conclusion about the population, This is what we called Inferential statistics.

Definition 29. A **statistic** is any quantity whose value can be calculated from the sample. Prior to obtaining data, there is uncertainty as to what value of any particular statistic will result. Therefore, a statistic is a random variable.

Examples: the sample mean $\overline{y}$, the sample median $\widetilde{y}$, the sample standard deviation s, etc. are all statistics.

Definition 30. The distribution of the statistics is called the sampling distribution of the statistics.

$\mathrm{Var}(\overline{X}) = \mathrm{Var}\left(\frac{X_1 + \cdots + X_n}{n}\right) = \frac{1}{n^2}\mathrm{Var}(X_1 + \cdots + X_n)$

$\overline{X} = \frac{X_1 + \cdots + X_n}{n}$

5.4 The Distribution of the Sample Mean

indep $\rightarrow = \frac{1}{n^2}(\mathrm{Var}(X_1) + \cdots + \mathrm{Var}(X_n))$

$E[\overline{X}] = \frac{1}{n}E[X_1 + \cdots + X_n]$

The importance of the sample mean $\overline{X}$ arises from its use in drawing conclusions about

$= \frac{1}{n^2}(\sigma^2 + \sigma^2 + \cdots + \sigma^2)$

$= \frac{1}{n}(E[X_1] + \cdots + E[X_n])$

$= \frac{1}{n}(\mu + \mu + \cdots + \mu) = \mu$

Let $X_1, X_2, \cdots, X_n$ be a random sample from a distribution with mean μ and standard deviation σ. Then

- $E(\overline{X}) = \mu$
- $\mathrm{Var}(\overline{X}) = \frac{\sigma^2}{n}$

$\mathrm{stdev}(\overline{X}) = \frac{\sigma}{\sqrt{n}}$

- The distribution of $\overline{X}$ becomes more concentrated about μ as sample size n increase, i.e. averaging moves probability in toward the middle.

NOTE: The standard deviation $\sigma_{\overline{X}}$ is often called the **standard error of the mean**; it describes the magnitude of a typical or representative deviation of the sample mean from the population mean.

NOTE: These formulas are true for any distribution.

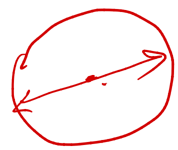
Example 74. The inside diameter of a randomly selected piston ring is a random variable with mean value 12cm and standard deviation 0.04cm.

a. If $\overline{X}$ is the sample mean diameter for a random sample of $n = 16$ rings, where is the sampling distribution of $\overline{X}$ centered, and what is the standard deviation of the $\overline{X}$ distribution?

b. Answer the questions posted in part (a) for a sample size of $n = 64$ rings.

c. For which of the two random samples, the one of part (a) or the one of part (b), is $\overline{X}$ more likely to be within 0.01cm of 12cm? Explain your reasoning.

Solution.



mean 12cm

stdev 0.04

$\mu = 12$

$\sigma = 0.04$

a) $E(\overline{X}_{16}) = 12$

$\text{stdev}(\overline{X}_{16}) = \frac{0.04}{\sqrt{16}} = 0.01$

b) $E(\overline{X}_{64}) = 12$

$\text{stdev}(\overline{X}_{64}) = \frac{0.04}{\sqrt{64}} = \frac{0.04}{8}$

c) $\overline{X}_{64}$ has smaller stdev. than $\overline{X}_{16}$

So it is closer (on average) to 12.

5.4.1 Samples from Normal Distribution

> Let $X_1, X_2, \cdots, X_n$ be a random sample from a normal distribution with mean μ and standard deviation σ. Then for any n, $\overline{X}$ is a normal
> w/ mean μ and stdev $\frac{\sigma}{\sqrt{n}}$

Example 75. (Example 74 continued) For the random sample in part (a), suppose X_i's are normally distributed, what is the probability that $\overline{X}$ is within one standard error of the mean?

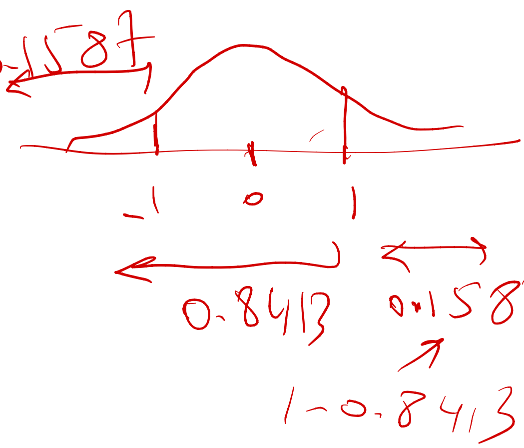
Solution. $N(\mu, \frac{\sigma^2}{n})$

$$P\left(|\overline{X} - \mu| \le \frac{\sigma}{\sqrt{n}}\right) \approx 0.68$$

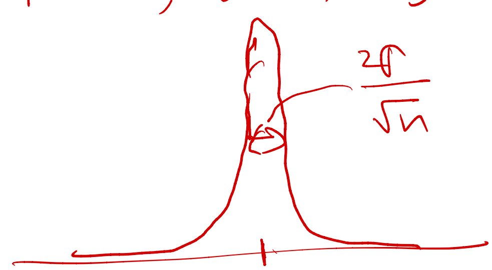
$$Z = \frac{\overline{X} - \mu}{\sigma/\sqrt{n}}$$

$$P(|Z| \le 1) \approx 0.6826$$

N(0,1)



pdf of $X_1, X_2, X_3 \cdots$



pdf of $\overline{X}$ μ

5.4.2 The Central Limit Theorem

When the X_i's are normally distributed, so is $\overline{X}$ for any sample size n. Even when the population distribution is highly non-normal, if n is large, a normal curve will approximate the actual distribution of $\overline{X}$.

The Central Limit Theorem (CLT)
Let $X_1, X_2, \cdots, X_n$ be a random sample from **any** distribution with mean μ and variance σ^2. Then if n is sufficiently large,

- $\overline{X}$ is approximately $N(\mu, \frac{\sigma^2}{n})$
- The larger the n is, the better the approximation
- $n \geq 30$

In summary,

	small sample size: $n \leq 30$	larger sample size $n > 30$
$X_1 \cdots X_n \sim N(\mu, \sigma^2)$	$\overline{X} \sim N(\mu, \frac{\sigma^2}{n})$	$\overline{X} \sim N(\mu, \frac{\sigma^2}{n})$
$X_1 \cdots X_n$ not $N(\mu, \sigma^2)$ but have mean μ var σ^2	$\overline{X}$ has mean μ stdev. $\frac{\sigma}{\sqrt{n}}$	$\overline{X} \sim$ (approximately) $N(\mu, \frac{\sigma^2}{n})$