Stat 213: Intro to Statistics 9
Central Limit Theorem

H. Kim

Fall 2007

unknown parameters

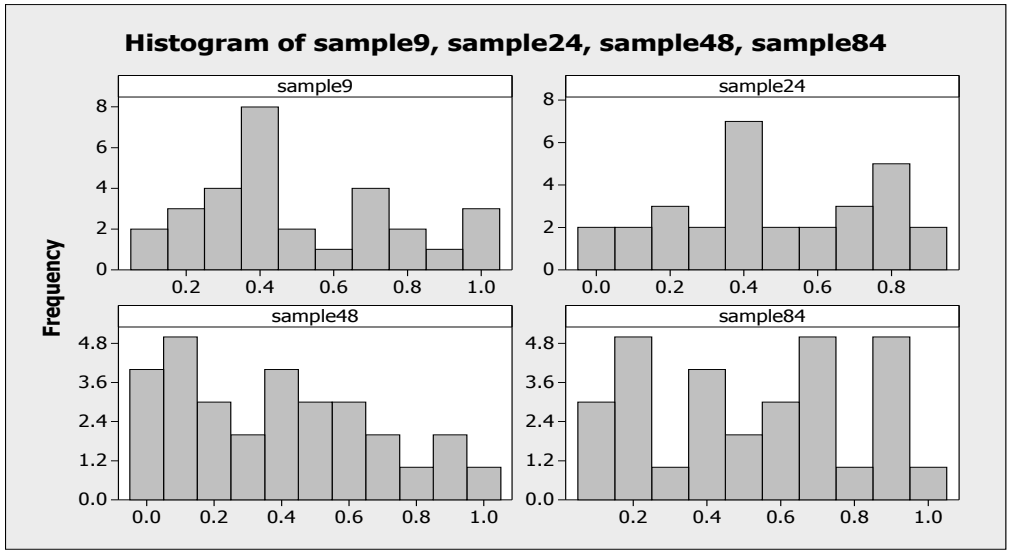
- Example: A pollster is sure that the responses to his "agree/disagree" questions will follow a binomial distribution, but p, the proportion of those who "agree" in the population, is unknown.
- In practice, the parameters of the distribution are unknown. Most rely on the `sample` to learn about the parameter.
- Want to the `sample` to provide reliable information about the population.

statistic

- A statistic is the numerical descriptive measures calculated from a sample: $\hat{p}$ and $\overline{X}$.
- A statistic is a random variable, their values vary from sample to sample $\Longrightarrow$ a statistic has a probability distribution.
- My sample represents the population?
  - the sampling distribution of a statistic is the probability distribution for all possible values of the statistic that results when random samples of size n are repeatedly drawn from the population
  - the expected value (mean) of sampling distribution is the true parameter, i.e. $E(\overline{X}) = \mu$ or $E(\hat{p}) = p$

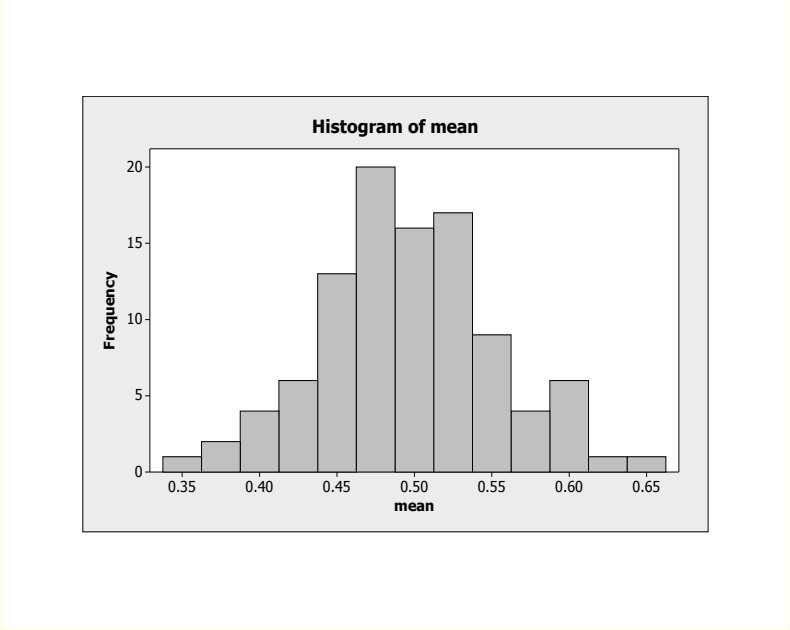
simulation 1

- If we draw 100 repeated random samples of the same size 30 from uniform population with mean $\mu = 0.5$ and standard deviation $\sigma = \frac{1}{12}$,



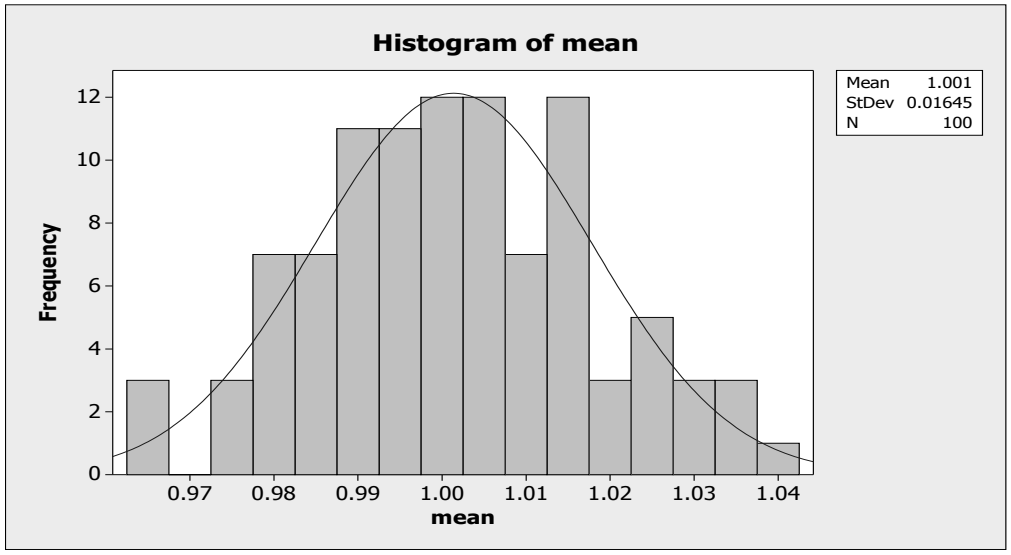
simulation 1

measure the means $(\overline{X})$ for each sample, and draw histogram:



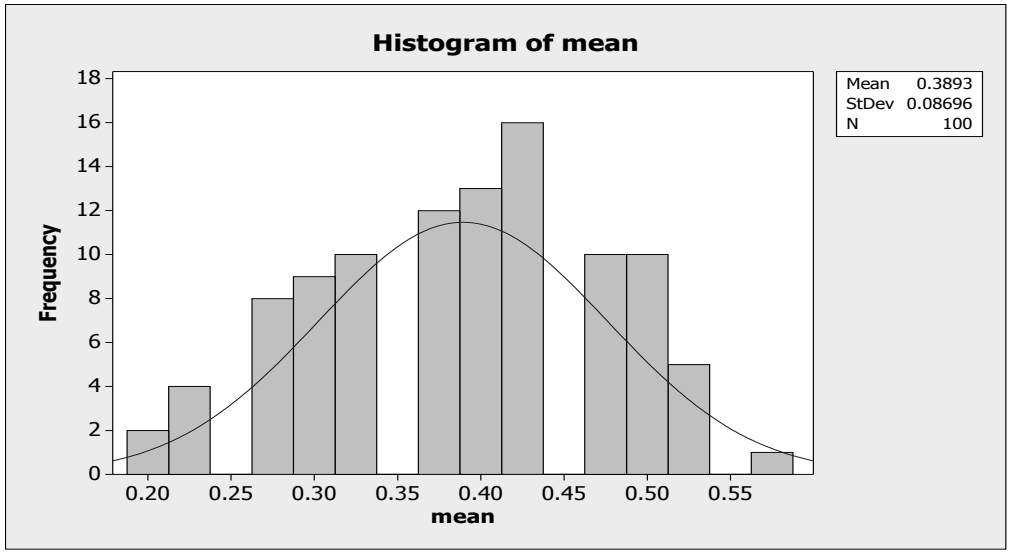
simulation 2

- If we draw 100 repeated random samples of the same size 30 from normal population with mean $\mu = 1$ and standard deviation $\sigma = 0.1$, and measure the means $(\overline{X})$ for each sample, and draw histogram:



simulation 3

- If we draw 100 repeated random samples of the same size 30 from Bernolli population with $p = 0.4$ and measure the means $(\overline{X})$ for each sample, and draw histogram:



mean and variance for sample mean, $\overline{X}$

- Random variables $X_1, X_2, \cdots, X_n$ are independent with mean $E(X_i) = \mu$ and variance $V(X_i) = \sigma^2$, $i = 1, 2, \cdots, n$:

$$\overline{X} = \frac{1}{n} \sum_{i=1}^{n} X_i$$

- $E(\overline{X})$ and $V(\overline{X})$

- Sampling distribution of the random variable $\overline{X}$?

mean and variance for sample proportion, $\hat{p}$

- If $X_1, \cdots, X_n$ are independent Bernoulli random variables with mean $E(X_i) = p$ and variance $V(X_i) = p(1-p)$, $i = 1, 2, \cdots, n$:
  - $X_i = \begin{cases} 1 & \text{if success} \\ 0 & \text{if failure} \end{cases}$
  - $Y = \sum_{i=1}^{n} X_i \sim Binomial(n, p)$

  the sample mean, $\overline{X} = \frac{1}{n}\sum X_i = \frac{Y}{n} = \hat{p}$: proportion
- $E(\overline{X})$ and $V(\overline{X})$

- Sampling distribution of the random variable $\hat{p}$?

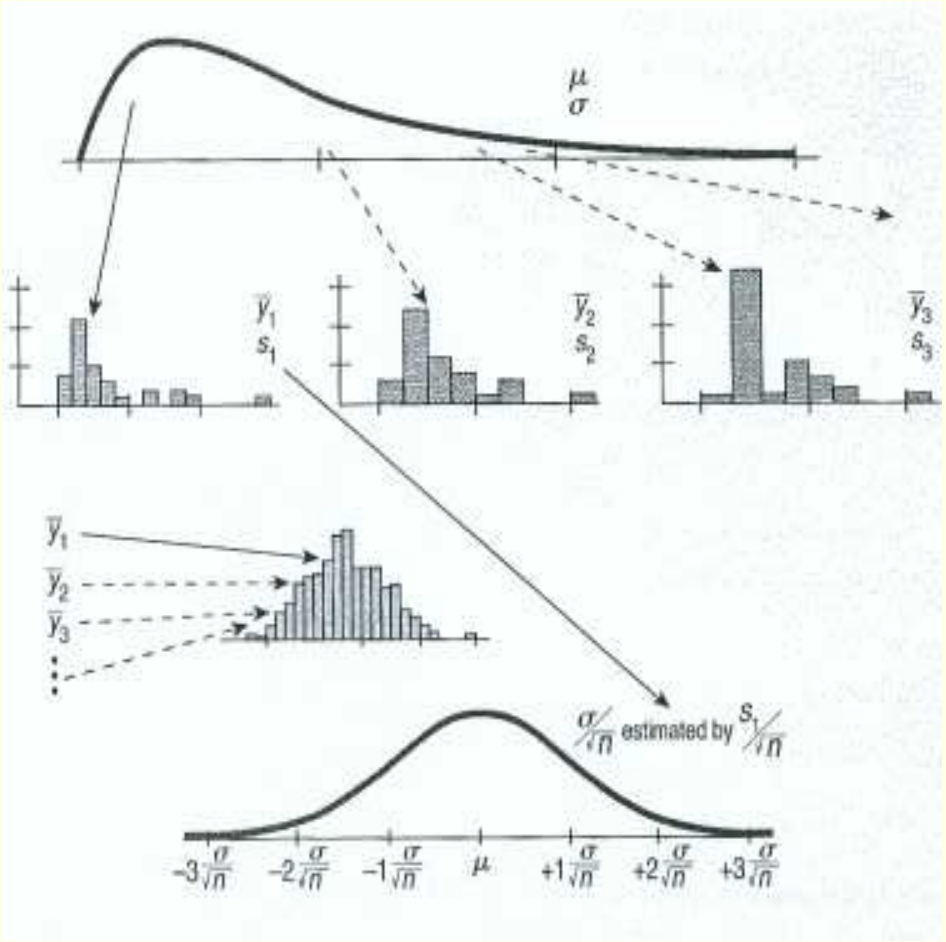
sampling distributions of $\overline{X}$ and $\hat{p} \Longrightarrow$ Normal ?

- Collection of the mean values will pile up around the underlying (μ) in such way that a histogram of the sample means ($\overline{X}$) can be modeled well by a Normal model: sampling distribution of the mean

$$\overline{X} \sim \mathcal{N}\left(\mu, \frac{\sigma^2}{n}\right)$$

$$\hat{p} \sim \mathcal{N}\left(p, \frac{p(1-p)}{n}\right), \quad np > 5, n(1-p) > 5$$

Central Limit Theorem



Central Limit Theorem

- When a random sample is drawn from any population with mean μ and standard deviation σ, its sample mean, $\overline{X}$, has a sampling distribution with the mean μ and standard deviation $\frac{\sigma}{\sqrt{n}}$ and the shape of the sampling distribution is approximately Normal as long as the sample size is large enough (at least 30).
- sampling distribution models tame the variation in statistics ($\overline{X}$) enough to know us to measure how close our computed statistic values are likely to be to the unknown underlying parameters (μ)
- standard error (se): estimated standard deviation $\left(\frac{\hat{\sigma}}{\sqrt{n}}\right)$ of the sampling distribution

the real world and the model world

- we never actually get to see the sampling distribution; we imagine repeated samples to develop the theory and own intuition about sampling distribution models
- sampling distributions act as a bridge from real world to imaginary model of the statistic and enable to say something about the population when all we have is data from the real world

example 1

- The length of stay of patients in a chronic health facility is normally distributed with a mean of 40 days and a standard deviation of 12 days. Suppose that a sample of $n = 16$ patients is randomly selected. Of interest is the mean length of the sample of $n = 16$ patients.

  a. Specify the distribution for the mean length of stay of the sample of 16 patients is less than 34 days ?

  b. What is the probability that the mean length of stay for the 16 patients is less than 34 days ?

example 1

c. What is the probability that the mean length of stay for the 16 patients is between 34 and 46 days ?

d. What is the probability that the length of stay of one of the 16 patients is less than 34 days?

example 2

- The population of healthy females in Canada has a mean potassium concentration of 4.36 mEq/l and a standard deviation of $0.12 mEq/l$. Suppose that a sample of 50 females is selected.

  a. Specify the distribution for the mean potassium concentration of the sample of 50 females. What is the standard error of this sample mean ?

  b. What is the probability that the mean potassium concentration for 50 females is below $4.4 mEq/l$?

example 3

- The duration of Alzheimer's disease from the onset of symptoms until death ranges from 3 to 20 years: the average is 8 years with a standard deviation of 4 years. The administrator of a large medical center randomly selects the medical records of 36 deceased Alzheimer's patients from the medical center's database and records the average duration. Find the approximate probability for these events:

  a. the average duration is less than 7 years

example 3

b. the average duration lies within 1 year of the population mean, $\mu = 8$.

example 4

- Statistics Canada reported that 33.1% of all 1997 family incomes in New Brunswick were below $30,000$. Suppose a random sample of 80, 1997 family incomes from New Brunswick is selected. What is the probability that the percentage of incomes in the sample that are under $30,000$ is over 30 percent ?