Week 4 Statistical theory
5. Central Limit Theorem

Stat 140 - 04

Mount Holyoke College

Dr. Shan Shan

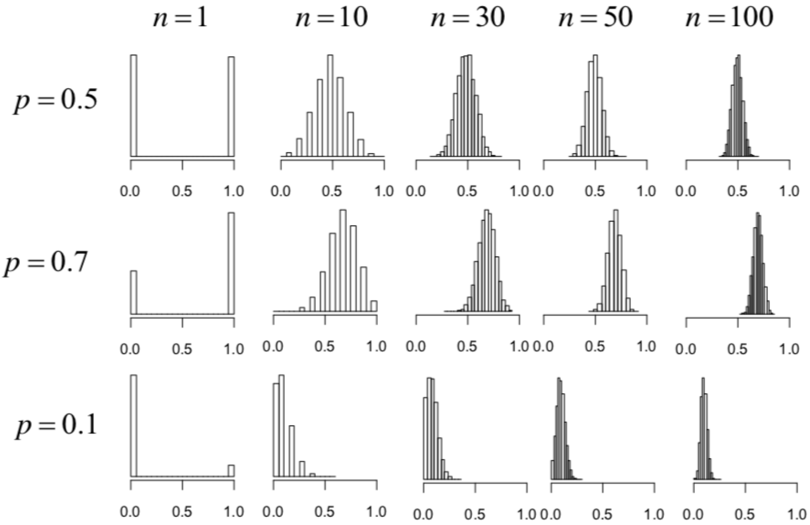
Slides posted at *http://sshanshans.github.io/stat140*

For a sufficiently large sample size, the distribution of sample proportion or sample mean is normal.

For a sufficiently large sample size, the distribution of sample proportion or sample mean is normal.

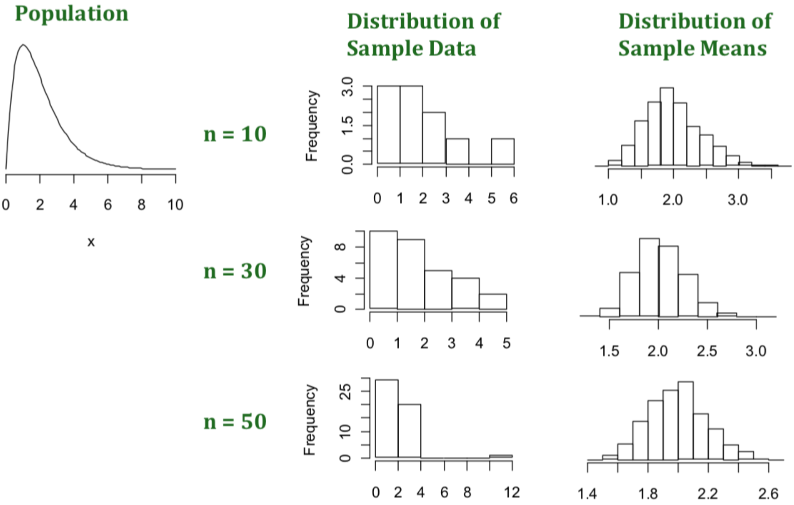
Also true for

- difference in sample mean
- difference in sample proportion
- ... (unbiased estimator)



$n=1$
$n=10$
$n=30$
$n=50$
$n=100$
$p=0.5$
$p=0.7$
$p=0.1$
0.0
0.5
1.0

CLT for mean


Population
x
0
2
4
6
8
10
Distribution of Sample Data
Distribution of Sample Means
n = 10
n = 30
n = 50
Frequency
0.0
1.5
3.0
0
1
2
3
4
5
6
1.0
2.0
3.0
0
4
8
1.5
2.5
10
25
12
1.4
1.8
2.2
2.6

- The central limit theorem holds for ANY original distribution, although "sufficiently large sample size" varies
- The more skewed the original distribution is (the farther from normal), the larger the sample size has to be for the CLT to work
- For small samples, it is more important that the data itself is approximately normal

"The theory of probabilities is at bottom nothing but common sense reduced to calculus."
– Laplace, in *Théorie analytique des probabilités*, 1812

We need to check the following two conditions

- Independence: Sampled observations must be independent. This is difficult to verify, but is more likely if
  - random sampling/assignment is used, and,
  - if sampling without replacement, $n < 10\%$ of the population.
- Sample size/skew: Either
  - the population distribution is normal or
  - $n > 30$ and the population dist. is not extremely skewed, or
  - n is much larger than 30 (approx. gets better as n increases).

- For distributions of a quantitative variable that are not very skewed and without large outliers, $n \geq 30$ is usually sufficient to use the CLT
- For distributions of a categorical variable, counts of at least 10 within each category is usually sufficient to use the CLT

The central limit theorem says ...

- For a sample proportion

$$\hat{p} \sim \mathcal{N}\left(p, \sqrt{\frac{p(1-p)}{n}}\right)$$

  where p is the population proportion, and n is the sample size

- For a sample mean

$$\bar{x} \sim \mathcal{N}\left(\mu, \frac{\sigma}{\sqrt{n}}\right)$$

  where μ is the population mean, σ is the population standard deviation and n is the sample size

In March 2011, a random sample of 1000 US adults were asked

"Do you think exercise is important"

753 adults responded they think exercise is important

- Use $753/1000 = 0.753$ to approximate p
- Sample size $n = 1000$

Plugging in $\hat{p} \sim \mathcal{N}\left(p, \sqrt{\frac{p(1-p)}{n}}\right)$, we get

$$\hat{p} \sim \mathcal{N}\left(0.75, 0.01\right)$$

The mean price of a house Topanga, CA was roughly \$1.3 million with a standard deviation of \$300,000. There were no houses listed below \$600,000 but a few houses above \$3 million.

What is the probability that the mean of 60 randomly chosen houses in Topanga is more than \$1.4 million?

According to CLT with $\mu = 1.3, \sigma = 0.3, n = 60$

$$\bar{x} \sim \mathcal{N}\left(1.3, \frac{0.3}{\sqrt{60}}\right)$$

$$P(\bar{x} > 1.4) = P(z > \frac{1.4 - 1.3}{0.0387}) = 0.0049$$