Sampling Distributions II

Lecture 17

02/25/2013

Parameter vs. Statistic

The mean GPA of all UCSB undergraduates is reported by the University to be 3.2 and the mean GPA of a sample of 500 UCSB undergraduates was found to be 2.96.

(a) 3.2 is a parameter and 2.96 is a statistic

(b) 3.2 is a statistic and 2.96 is a parameter

(c) both are statistics

(d) both are parameters

Example: *Average number of cars in US households*

- Record the number of cars in each household in the US.
- Population: All US Households

  Population size: $N = 312,615,000$

  Dataset: $x_1, \ldots, x_N$, where x_1 would be the number of cars in the 1st household in the population etc.
- Population Average:

$$\mu = \frac{x_1 + x_2 + \ldots + x_N}{N}$$

- Population Standard deviation:

$$\sigma = \sqrt{\frac{1}{N-1} \sum_{\text{for all } x} (x - \mu)^2}$$

Sample

Sample Dataset: $\{X_1, \ldots, X_n\}$

Sample size: n

Goal

Based on the sample values, $\{X_1, \ldots, X_n\}$, where n is the sample size, we try to reach some conclusions for the parameters of interest.

- The ideal sample:
  - representative
  - unbiased

- Choose the sample at *random*.

Simple random sample

Simple Random Sample

$\{X_1, \ldots, X_n\}$ is a Simple Random Sample if

i. if a certain member of the population is chosen, this does not affect the chances of another member to be chosen.

ii. every member of the population is equally likely to be chosen.

In other words...

i. $\{X_1, \ldots, X_n\}$ are independent

ii. $\{X_1, \ldots, X_n\}$ are identically distributed (i.e. have the same pmf or pdf).

Estimation of μ

- An *estimator* of a parameter is a *statistic* whose value in the sample is used to estimate this parameter.
  - An estimator for μ is the sample mean, $\bar{x}$.
  - An estimator for σ is the sample standard deviation, s.

Examples

- $X_1, \ldots, X_n$ is a sample of n American households.
- An estimate for μ will be

$$\bar{x} = \frac{x_1 + \ldots + x_n}{n}$$

- An estimator for σ will be $s = \sqrt{\frac{1}{n-1} \sum_{\text{for all } x} (x - \bar{x})^2}$.

Properties of $\bar{x}$

Question

Is $\bar{x}$ a “good" estimator for μ?

- This depends on <u>how</u> the sample $X_1, \ldots, X_n$ is chosen.
- If $X_1, \ldots, X_n$ are **iid**, i.e. if we have a *simple random sample (SRS)*, then the sample mean $\bar{X}$ has some has some very nice properties.

Remark

- $\bar{X}$ is a random variable, thus
  - It has a distribution, which we will call "sampling distribution".

Example

- Suppose that many different researchers collect SRS of households (all of them with the same sample size n).
- Then, each researcher will compute a different sample mean $\bar{x}$, i.e. we will have the following table:

Researcher 1	$\bar{x}_1$
Researcher 2	$\bar{x}_2$
...	...

Central Limit Theorem

Central Limit Theorem

$X_1, X_2, \ldots, X_n$ are independent and identically distributed and n is large (i.e. n>30), then

$$\mathbb{E}(\bar{x}) = \mu$$

$$s.d.(\bar{x}) = \frac{\sigma}{\sqrt{n}}$$

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

When do we use the CLT?

1. How large n should be?
2. The Central Limit Theorem does NOT say that every individual random variable is approximately normal.
3. The Central Limit Theory applies *independently of the distribution of* X. It suffices to know its expectation and its standard deviation.

Example: *Exam scores*

In a previous year of PSTAT 5A, the student scores on exams had mean 74 and standard deviation 14. The instructor gave a final exam in a class of 64 students.

- Approximate the probability that the average test scores in the class exceeds 80.

Remark!

In the problem, it is not mentioned that the exam scores follow a Normal distribution!

Example: *Exam scores*

- X_i= test score of the ith student in the class of 64 students, i=1,...,64.

- Average test score: $\bar{x} = \frac{X_1 + \ldots + X_{64}}{64}$
- CLT assumptions? (iid, n>30)

$$P(\bar{x} > 80) = P\left(Z > \frac{80 - 74}{14/\sqrt{64}}\right) = P(Z > 3.429)$$
$$= 1 - P(Z \leq 3.429) = 1 - 0.9997 = 0.0003$$

Turn on your clickers!

We have a sample of 100 uniform random variables $X_1,\ X_2, \ldots, X_{100} \sim U[0, 10]$ with mean 5 and standard deviation 8.3. The distribution of $\bar{x} = \frac{X_1+X_2+\ldots+X_{100}}{100}$ can be approximated by a

(a) Normal(5, 8.3/10)

(b) Uniform[0, 10]

(c) Uniform[5, 8.3]

(d) Normal(5, 8.3/100)

Turn on your clickers!

We have a sample of 100 uniform random variables
$X_1,\ X_2, \ldots, X_{100} \sim U[0, 10]$, with mean 5 and standard deviation 8.3.
The distribution of $\bar{x} = \frac{X_1+X_2+...+X_{100}}{100}$ can be approximated by a

(a) Normal(5, 8.3/10)

(b) Uniform[0, 10]

(c) Uniform[5, 8.3]

(d) Normal(500, 8.3/100)

Population & Sample Proportion

Survey

- Simple Random Sample
- 55% are in favor to a statement
- Repeat survey
  - 56% are in favor
  - 52% are in favor
  - ...
- Sample is random

Example: *Elections*

- 55% support the Democrats
- 56% support the Democrats
- 52% support the Democrats
- ...

Sampling Proportion

- $X \sim \text{Binomial}(n, p)$
  - p is a population parameter (typically unknown)
- If we have X but not p,
  - Sample proportion
    $$\hat{p} = \frac{X}{n}$$
  - Sample proportion $\approx$ Population proportion
  - $\hat{p}$ is an estimator for p.

Sampling Proportion

Example: *Elections*

- X = number of respondents who support the Democrats
- n = total number of respondents in the survey
- $\hat{p} = \frac{X}{n}$ = proportion of respondents who support the Democrats
- We estimate the proportion of respondents in the population who support the democrats based on the sample proportion.

Sampling Distribution of $\hat{p}$

- Expectation

$$\mathbb{E}(\hat{p}) = \frac{n\,p}{n} = p$$

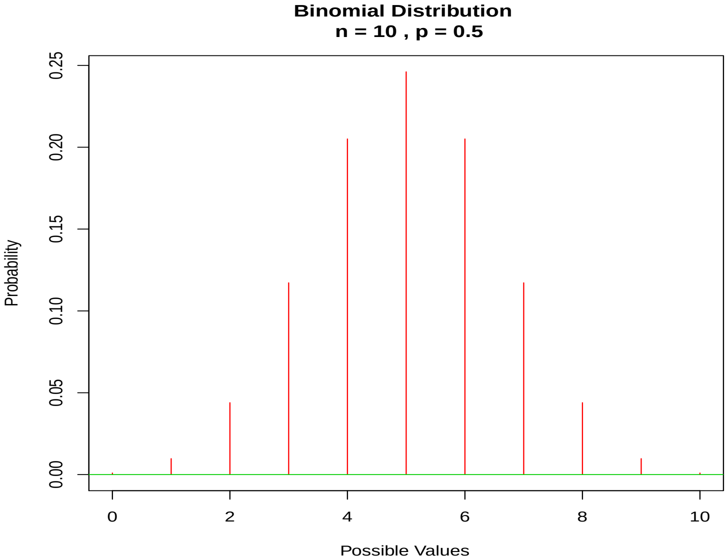
- Standard deviation

$$s_{\hat{p}} = \sqrt{\frac{p\,(1-p)}{n}}$$

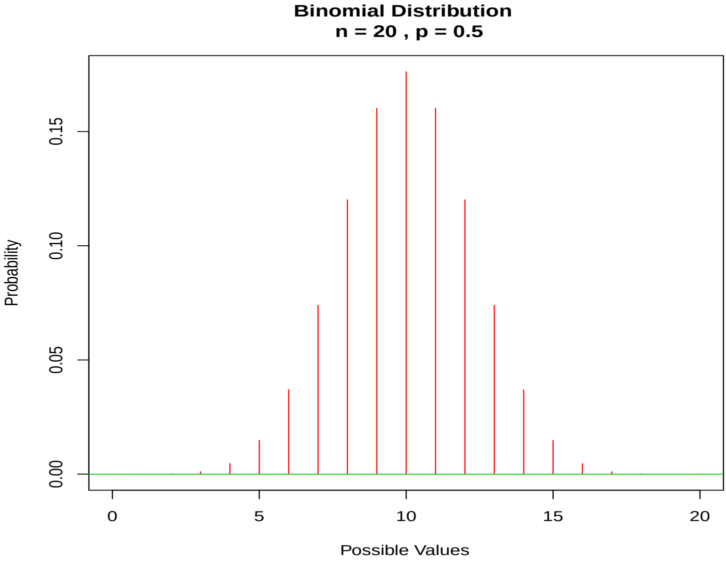
- Standard error

$$s.e.(\hat{p}) = \sqrt{\frac{\hat{p}(1-\hat{p})}{n}}$$

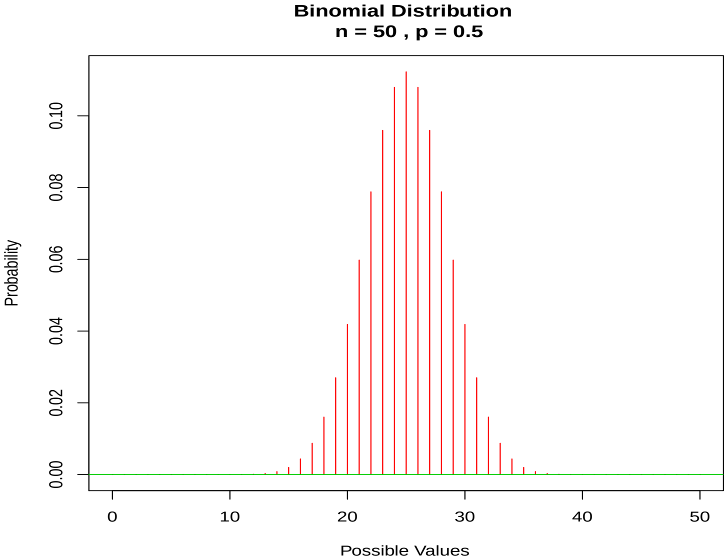
Let n get big: $n = 10, \quad p = 0.5$



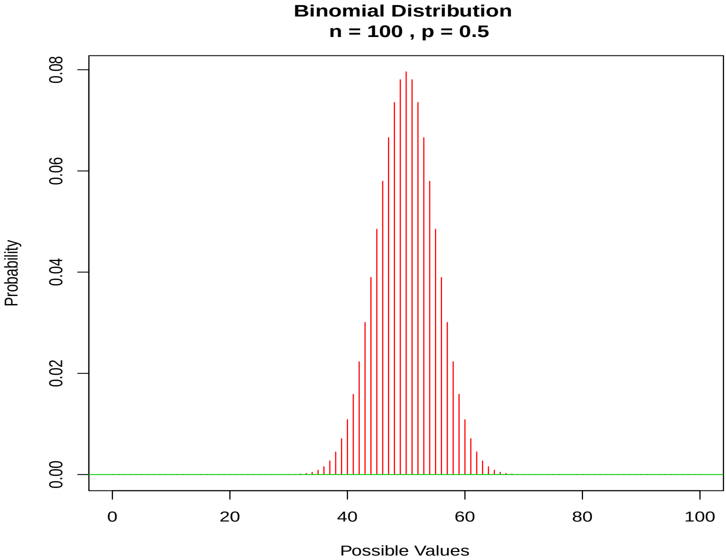
Let n get big: $n = 20, \quad p = 0.5$



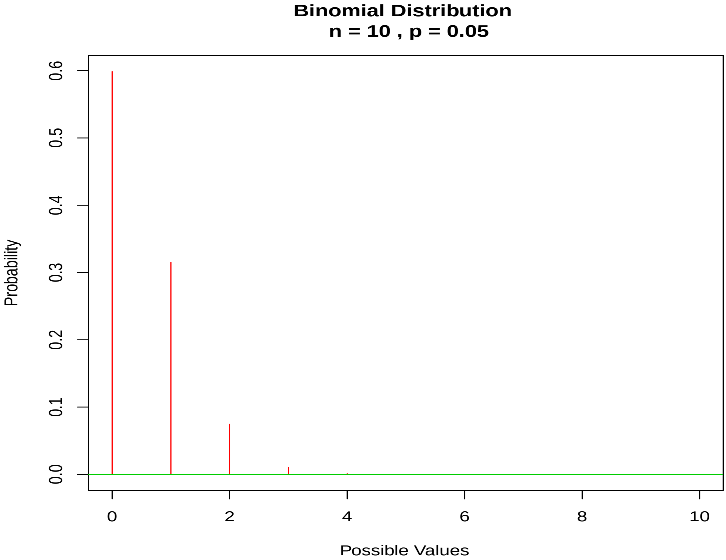
Let n get big: $n = 50, \quad p = 0.5$



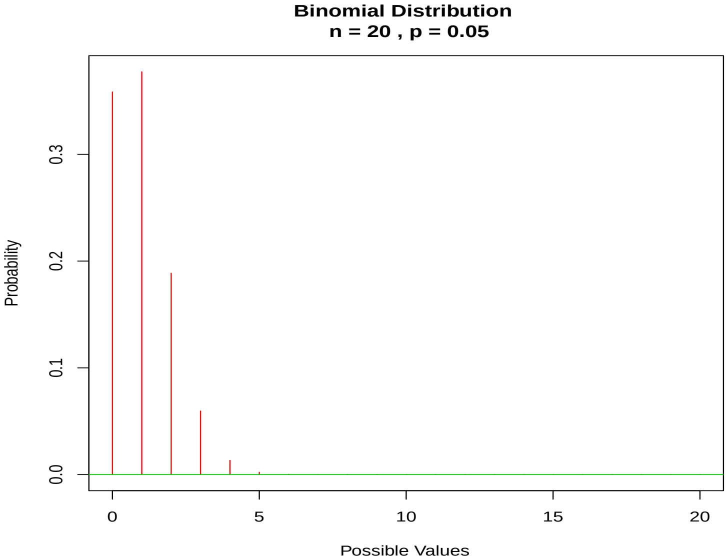
Let n get big: $n = 100, \quad p = 0.5$



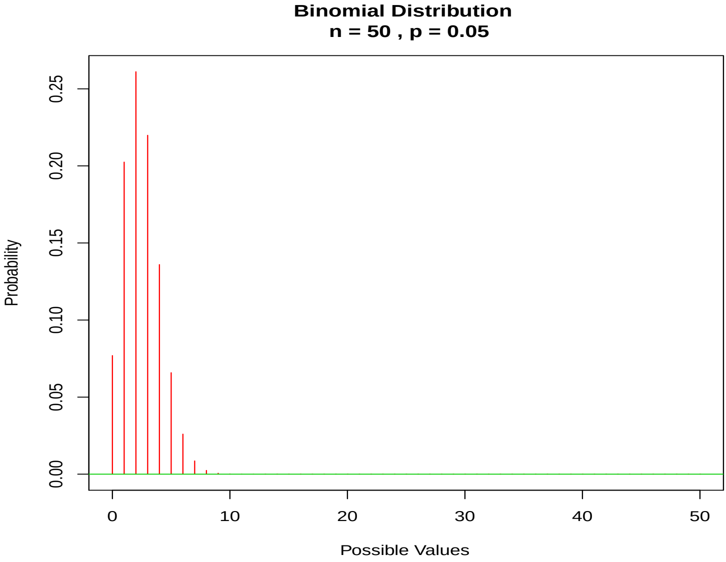
Let n get big: $n = 10, \quad p = 0.05$



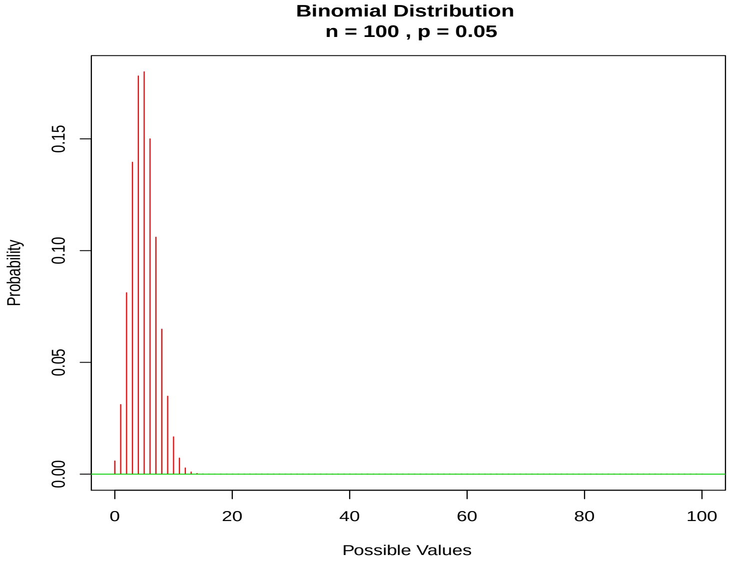
Let n get big: $n = 20, \quad p = 0.05$



Let n get big: $n = 50, \quad p = 0.05$



Let n get big: $n = 100, \quad p = 0.05$



Law of Large Numbers

- As $n \to \infty$
- Sample Proportion $\to$ Population proportion
- Standard Deviation $\to 0$

Central Limit Theorem

If n is large enough, the sample proportion $\hat{p}$ behaves approximately as a normal random variable with

- Mean: $\mu = p$
- Standard deviation: $\sigma = \sqrt{\frac{p\,(1-p)}{n}}$.

Standardization for Binomial

In other words, if n is *large enough*

$$Z \approx \frac{\hat{p} - p}{\sqrt{\frac{p\,(1-p)}{n}}} \approx \frac{\hat{p} - p}{\sqrt{\frac{\hat{p}\,(1-\hat{p})}{n}}}$$

$$Z \sim \mathcal{N}(0, 1)$$

When can we use this approximation?

Rule

1. A population with a fixed proportion
2. Random Sample

    Independent

    Equally likely (equal chance)

3. Sample size is large

    $np > 10$

    $n(1-p) > 10$

i.e. this is a binomial experiment with normal approximation.

Example: *Tossing a coin n times*

Suppose we flip a coin 50 times with $\mathbb{P}(\textit{Heads}) = 0.25$.

(a) What is the distribution of the sample proportion?

(b) What is the probability to have more than 50% Heads in the 50 tosses?

Example: *Tossing a coin n times*

X = # of Heads in the 50 tosses

1. We have a fixed number of tosses ($n = 50$).
2. Each toss is independent of the others.
3. There are two possible outcomes (Heads, Tails)
4. $\mathbb{P}(success) = \mathbb{P}(Heads) = 0.25$

This is a Binomial experiment!

Check also

$$n\,p = 50 \cdot 0.25 = 12.5 > 10$$
$$n\,(1-p) = 50 \cdot 0.75 = 37.5 > 10$$

(This guarantees that the sample size is large enough.)

Example: *Tossing a coin n times*

(a) $\mu = p = 0.25$

$$\sigma = \sqrt{\frac{p\,(1-p)}{n}} = \sqrt{\frac{0.25\,(1-0.25)}{50}} = 0.06$$

$$\hat{p} \sim \mathcal{N}(0.25,\ 0.06)$$

(b)

$$\begin{aligned} \mathbb{P}(\hat{p} \geq 0.5) &= \mathbb{P}\left(Z \geq \frac{0.5 - 0.25}{0.06}\right) \\ &= \mathbb{P}(Z \geq 4.17) \approx 0 \end{aligned}$$