Sampling Distributions I

Lecture 16

22/02/2013

Statistical Experiment

- Data is fixed
- Experiment is random
  - Many possible samples
- For example
  - Survey
    - ★ Every group of 1000 people is equally likely
  - Experiment
    - ★ Each assignment to control or treatment could happen

Example

In 82 basketball games, Kobe Bryant scored an average of 18.33 points.

- What is random?
- What is unknown?
- X = points scored on a single game

$$X \sim \mathcal{N}(\mu, \sigma)$$

  - What is μ?

Population vs. Sample

population_sample.jpg

Statistics vs. Parameters

- **Parameter**: a characteristic of the population. Typically unknown due to the large number of individuals in the population.
- **Statistic**: a quantity that is computed from the data that we collected from the population (sample).

Statistics vs. Parameters

- Statistics: can be calculated based on the sample
  - $\bar{x}$: sample mean
  - s: sample standard deviation
  - $\hat{p}$: sample proportion
- Parameters: unknown
  - μ: population mean
  - σ: population standard deviation
  - p: population proportion

Turn on your clickers!

The scores of a physic's exam are normally distributed with mean 80 and standard deviation 5. If we randomly select 20 students from the class, their average score is calculated to be 75.
Which one is a parameter and which one is a statistic?

(a) 80 is a parameter, 75 is a statistic

(b) 80 is a statistic, 75 is a parameter

(c) both 75 and 80 are statistics

(d) both 75 and 80 are parameters

Turn on your clickers!

The scores of a physic's exam are normally distributed with mean 80 and standard deviation 5. If we randomly select 20 students from the class, their average score is calculated to be 75.
Which one is a parameter and which one is a statistic?

(a) 80 is a parameter, 5 is a statistic

(b) 80 is a statistic, 5 is a parameter

(c) both 5 and 80 are statistics

(d) both 5 and 80 are parameters

Statistical Inference

- Use statistics to describe the parameter
- Parameter is the goal
- Statistics are the tools

Sampling Distribution

- Distribution of the statistics
- Depends on the parameter
- The statistics are random variables
  - Imagine the data are random variables
  - Statistic = function of data
  - Statistic is also random

Quantitative Measurements

Example: *Sample Mean*

- Statistic = $\bar{x}$
- Parameter = μ
- Sampling distribution

  Randomness in statistics

  Uncertainty in measurement

$\bar{x}$ as a random variable

- Data: $X_1, X_2, X_3, \ldots, X_n$
- Simple Random Sample

  Independent Identically Distributed observations
  - For example, every

$$X_i \sim \mathcal{N}(\mu, \sigma), \text{ for } i = 1, \ldots, n$$

    - ⋆ Unfortunately, μ and σ are unknown.

Expected Value of $\bar{x}$

$$\mathbb{E}(\bar{x}) = \frac{1}{n}\sum_{i=1}^{n}\mathbb{E}(X_i) = \mu$$

"The mean of the mean is the mean"

Why?

$$\mathbb{E}(\bar{x}) = \mathbb{E}\left(\frac{1}{n}\sum_{i=1}^{n}X_i\right) = \frac{1}{n}\sum_{i=1}^{n}\mathbb{E}(X_i)$$

$$= \frac{1}{n}\sum_{i=1}^{n}\mu = \frac{1}{n}\cdot n\cdot\mu = \mu$$

Standard Deviation of $\bar{x}$

$$\sigma_{\bar{x}} = \sqrt{Var(\bar{x})} = \frac{\sigma}{\sqrt{n}}$$

- Error in $\bar{x}$ shrinks with n
- Depends on the unknown σ

Standard Error of $\bar{x}$

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

What is the difference between $\sigma_{\bar{x}}$ and $s.e.(\bar{x})$

- σ is in general unknown
- We replace σ with the statistic s

5 Independent Normal Observations

- Dataset: $63, 65, 72, 74, 74$
- Sample Mean (statistic):

$$\bar{x} = 69.6$$

- Sample Standard Deviation (corresponds to 1 random variable):

$$s = \sqrt{\frac{1}{4}\left(\sum_{i=1}^{5} x_i^2 - 5(69.6)^2\right)} = 5.225$$

- Standard Error:

$$s.e.(\bar{x}) = \frac{s}{\sqrt{5}} = \frac{5.225}{\sqrt{5}} = 2.337$$

Distribution of $\bar{x}$

Rule:

When $X_1, X_2, \ldots, X_n$ are independent and normally distributed with mean μ and standard deviation σ (that is each $X_i \sim \mathcal{N}(\mu, \sigma)$), then

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

Standardization

- For one variable X:

$$Z = \frac{x - \mu}{\sigma}$$

- For the sample mean $\bar{x}$ of n variables $X_1, \ldots, X_n$:

$$Z = \frac{\bar{x} - \mu}{\sigma/\sqrt{n}}$$

OR

$$Z = \frac{\sqrt{n}\,(\bar{x} - \mu)}{\sigma}$$

Example

- You take an exam once

$$\mu_{score} = 78, \quad \sigma_{score} = 3.8$$

- You take the exam $n = 5$ times

  $\bar{x}$ = the average of your scores

$$\mu_{\bar{x}} = 78, \quad \sigma_{\bar{x}} = \frac{3.8}{\sqrt{5}} = 1.7$$

Q: $\mathbb{P}(\mathit{Score} > 80) = ?$ and $\mathbb{P}(\mathit{Average} > 80) = ?$

- You take an exam once ($\mu_{score} = 78, \quad \sigma_{score} = 3.8$)

$$\mathbb{P}(\mathit{Score} > 80) = \mathbb{P}\left(Z > \frac{80 - 78}{3.8}\right)$$
$$= \mathbb{P}(Z > 0.53) = 0.2981$$

- You take the exam $n = 5$ times ($\mu_{\bar{x}} = 78, \quad \sigma_{\bar{x}} = \frac{3.8}{\sqrt{5}} = 1.7$)

$$\mathbb{P}(\mathit{Average} > 80) = \mathbb{P}\left(Z > \frac{80 - 78}{3.8/\sqrt{5}}\right)$$
$$= \mathbb{P}(Z > 1.18) = 0.1190$$

Kobe Bryant Example

Suppose that each basketball game is normally distributed with $X_i \sim \mathcal{N}(28, 8.5)$ (there are 82 games in a season).

- $\mu = 28$ and $\sigma = 8.5$
- $\mathbb{E}(\bar{x}) = 28$ and $\sigma_{\bar{x}} = \frac{8.5}{\sqrt{82}} = 0.939$
- What is the probability that Kobe scores on average more than 28.33 points in a particular game?

$$\begin{aligned}\mathbb{P}(\bar{x} \geq 28.33) &= \mathbb{P}\left(Z \geq \frac{28.33 - 28}{8.5/\sqrt{82}}\right) \\ &= \mathbb{P}(Z \geq 0.352) = 0.3632\end{aligned}$$

What if σ is unknown??

Estimate σ with s

$$Z = \frac{\bar{x} - \mu}{s/\sqrt{n}}$$

Central Limit Theorem

- For large n
  - Rule of thumb: Large is when n>30
- When σ is known
  - Sums and averages look like normals
  - Binomials are approximately normal

Central Limit Theorem

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

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

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

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

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.

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?

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

Book: Sections 5.2, 5.3