Statistics and Data Analysis

Distributions and Sampling (2)

Ling-Chieh Kung

Department of Information Management
National Taiwan University

Introduction

- When we cannot examine the whole population, we study a **sample**.
  - One needs to choose among different **sampling techniques**.
  - What will be contained in a random sample is unpredictable.
  - We need to know the **probability distribution** of a sample so that we may connect the sample with the population.
- The probability distribution of a sample is a **sampling distribution**.

Introduction

- A factory produce bags of candies. Ideally, each bag should weigh 2 kg. As the production process cannot be perfect, a bag of candies should weigh between 1.8 and 2.2 kg.
- Let X be the weight of a bag of candies. Let μ and σ be its expected value and standard deviation.
  - Is $\mu = 2$?
  - Is $1.8 < \mu < 2.2$?
  - How large is σ?
- Let's sample:
  - In a random sample of 1 bag of candies, suppose it weighs 2.1 kg. May we conclude that $1.8 < \mu < 2.2$?
  - What if the average weight of 5 bags in a random sample is 2.1 kg?
  - What if the sample size is 10, 50, or 100?
  - What if the mean is 2.3 kg?
- We need to know the sampling distribution of those statistics (sample mean, sample standard deviation, etc.).

Road map

- **Sample means**.
- Distributions of sample means.
- Sample proportions.

Sample means

- The sample mean is one of the most important statistics.

Definition 1

Let $\{X_i\}_{i=1,\ldots,n}$ be a sample from a population, then

$$\bar{x} = \frac{\sum_{i=1}^{n} X_i}{n}$$

is the sample mean.

- Sometimes we write $\bar{x}_n$ to emphasize that the sample size is n.
- Let's assume that X_i and X_j are independent for all $i \neq j$.
  - This is fine if $n \ll N$, i.e., we sample a few items from a large population.
  - In practice, we require $n \leq 0.05N$.

Means and variances of sample means

- Suppose the population mean and variance are μ and σ^2, respectively.
  - These two numbers are fixed.
- A sample mean $\bar{x}$ is a **random variable**.
  - It has its expected value $\mathbb{E}[\bar{x}]$, variance $\text{Var}(\bar{x})$, and standard deviation $\sqrt{\text{Var}(\bar{x})}$. These numbers are all **fixed**
  - They are also denoted as $\mu_{\bar{x}}$, $\sigma^2_{\bar{x}}$, and $\sigma_{\bar{x}}$, respectively.
- For **any** population, we have the following theorem:

Proposition 1 (Mean and variance of a sample mean)

Let $\{X_i\}_{i=1,\ldots,n}$ be a size-n random sample from a population with mean μ and variance σ^2, then we have

$$\mu_{\bar{x}} = \mu, \quad \sigma^2_{\bar{x}} = \frac{\sigma^2}{n}, \textit{ and } \quad \sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}}.$$

Means and variances of sample means

- Do the terms confuse you?
  - The sample mean vs. the mean of the sample mean.
  - The sample variance vs. the variance of the sample mean.
- By definition, they are:
  - $\bar{x} = \frac{1}{n}\sum_{i=1}^{n} X_i$; a random variable.
  - $\mathbb{E}[\bar{x}]$; a constant.
  - $s^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \bar{x})^2$; a random variable.
  - $\text{Var}(\bar{x})$; a constant.
- The sample variance also has its mean and variance.

Example 1: Dice rolling

- Let X be the outcome of rolling a fair dice.
  - We have $\Pr(X = x) = \frac{1}{6}$ for all $x = 1, 2, ..., 6$.
  - We have

$$\mu = \sum_{x=1}^{6} x \Pr(X = x) = 3.5,$$

$$\sigma^2 = \sum_{x=1}^{6} (x - \mu)^2 \Pr(X = x) \approx 2.917, \text{ and}$$

$$\sigma = \sqrt{\sigma^2} \approx 1.708.$$

x	$\Pr(X = x)$	$(x - \mu)^2$
1	0.167	6.25
2	0.167	2.25
3	0.167	0.25
4	0.167	0.25
5	0.167	2.25
6	0.167	6.25
	$\mu = 3.5$	$\sigma^2 \approx 2.917$

Example 1: Dice rolling

- Suppose now we roll the dice **twice** and get X_1 and X_2 as the outcomes.
- Let $\bar{x}_2 = \frac{X_1+X_2}{2}$ be the sample mean.
- The theorem says that $\mu_{\bar{x}_2} = \mu = 3.5$ and $\sigma_{\bar{x}_2} = \frac{\sigma}{\sqrt{n}} \approx \frac{1.708}{1.414} = 1.208$.
- $\mu_{\bar{x}_2} = \mu$: We expect $\bar{x}$ to be **around** 3.5, just like X.
  - The expected value of each outcome is 3.5. So the average is still 3.5.
- $\sigma_{\bar{x}_2} = \frac{\sigma}{\sqrt{2}} < \sigma$: The variability of $\bar{x}_2$ is smaller than that of X.
  - For X, $\Pr(X \geq 5) = \frac{1}{3}$.
  - For $\bar{x}_2$,

$$\begin{aligned}\Pr(\bar{x}_2 \geq 5) &= \Pr\left((X_1, X_2) \in \left\{(4,6),(5,5),(6,4),(5,6),(6,5),(6,6)\right\}\right) \\ &= \frac{1}{6}.\end{aligned}$$

  - To have a large value of $\bar{x}_2$, we need **both** values to be large.

Example 1: Dice rolling

- Let $\bar{x}_4 = \frac{\sum_{i=1}^4 X_i}{4}$ be the sample mean of rolling the dice **four times**.
- The theorem says that $\mu_{\bar{x}_4} = \mu = 3.5$ and $\sigma_{\bar{x}_4} = \frac{\sigma}{\sqrt{n}} \approx \frac{1.708}{2} = 0.854$.
- We have
$$\sigma_{\bar{x}_4} = \frac{\sigma}{\sqrt{4}} < \sigma_{\bar{x}_2} = \frac{\sigma}{\sqrt{2}} < \sigma.$$
The variability of $\bar{x}_4$ is **even smaller** than that of $\bar{x}_2$.
  - To have a large $\bar{x}_4$, we need most of the four values to be large.

Proposition 2

For two random samples of size n and m from the same population, let $\bar{x}_n$ and $\bar{x}_m$ be their sample means. Then we have

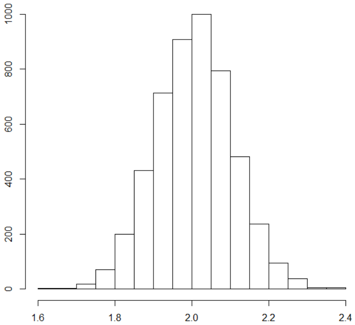
$$\sigma_{\bar{x}_n} < \sigma_{\bar{x}_m} \quad \textit{if} \quad n > m.$$

Example 2: Quality inspection

- The weight of a bag of candies follow a normal distribution with mean $\mu = 2$ and standard deviation $\sigma = 0.2$.
- Suppose the quality control officer decides to sample 4 bags and calculate the sample mean $\bar{x}$. She will punish me if $\bar{x} \notin [1.8, 2.2]$.
  - Note that my production process is actually "good:" $\mu = 2$.
  - Unfortunately, it is not perfect: $\sigma > 0$.
  - We may still be punished (if we are unlucky) even though $\mu = 2$.
- What is the probability that I will be punished?
  - We want to calculate $1 - \Pr(1.8 < \bar{x} < 2.2)$.
  - We know that $\mu_{\bar{x}} = \mu = 2$ and $\sigma_{\bar{x}} = \frac{\sigma}{\sqrt{4}} = 0.1$.
  - But we do not know the **probability distribution** of $\bar{x}$!

Experiments for estimating the probabilities

- Let's do an experiment.
  - Generate the weights of 4 bags of candies following $\text{ND}(2, 0.2)$.
  - Calculate $\bar{x}$.
  - Repeat this for 5000 times.
  - Draw a histogram for these 5000 $\bar{x}$s.
- The result of my experiment:
  - The mean of the 5000 $\bar{x}$ is 1.993741.
  - The standard deviation of the 5000 $\bar{x}$ is 0.1002187.
  - It looks like a normal distribution.
  - The proportion of $\bar{x}$s above 2.2 or below 1.8 is 4.68%.
- Is $\bar{x} \sim \text{ND}(2, 0.1)$?



Experiments for estimating the probabilities

- If $\bar{x} \sim \text{ND}(2, 0.1)$:
  - $\Pr(\bar{x} > 2) = 0.5$.
  - $\Pr(\bar{x} < 1.8) + \Pr(\bar{x} > 2.2) \approx 0.0455$.
- Our experiments only give us sample outcomes. However, our outcomes should be close to the theoretical outcomes.
- If we do multiple rounds of this experiment:

Round	Mean	Standard deviation	Proportion of $\bar{x} > 2$	Proportion of $\bar{x} < 1.8$ and $\bar{x} > 2.2$
1	1.994	0.100	0.473	0.047
2	2.006	0.100	0.530	0.047
3	2.003	0.104	0.513	0.058
4	1.996	0.104	0.486	0.054

- It seems that $\bar{x} \sim \text{ND}(2, 0.1)$ is true. Is it?

Road map

- Sample means.
- **Distributions of sample means**.
- Sample proportions.

Sampling from a normal population

- If the population is normal, the sample mean is also **normal**!

Proposition 3

Let $\{X_i\}_{i=1,\ldots,n}$ *be a size-*n *random sample from a normal population with mean* μ *and standard deviation* σ*. Then*

$$\bar{x} \sim \text{ND}\left(\mu, \frac{\sigma}{\sqrt{n}}\right).$$

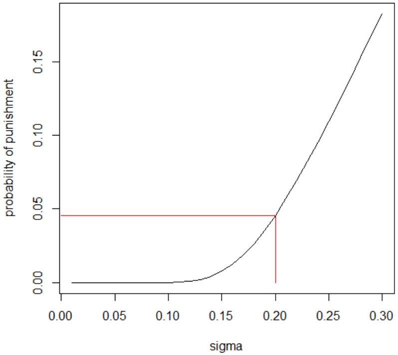
- We already know that $\mu_{\bar{x}} = \mu$ and $\sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}}$. This is true regardless of the population distribution.
- When the population is normal, the sample mean will also be normal.

Example 2 revisited: Quality inspection

- The weight of a bag of candies follow a normal distribution with mean $\mu = 2$ and standard deviation $\sigma = 0.2$.
- Suppose the quality control officer decides to sample 4 bags and calculate the sample mean $\bar{x}$. She will punish me if $\bar{x} \notin [1.8, 2.2]$.
- What is the probability that I will be punished?
  - the distribution of the sample mean $\bar{x}$ is $\mathrm{ND}(2, 0.1)$.
  - $\Pr(\bar{x} < 1.8) + \Pr(\bar{x} > 2.2) \approx 0.045$.

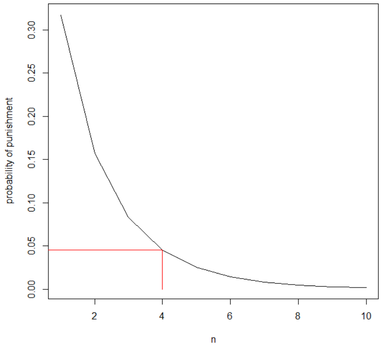
Adjusting the standard deviation

- When the population is ND($\mu = 2, \sigma = 0.2$) and the sample size is $n = 4$, the probability of punishment is 0.045.
- If we adjust our standard deviation σ (by paying more or less attention to the production process), the probability will change.
- Reducing σ reduces the probability of being punished. With the sampling distribution of $\bar{x}$, we may **optimize** σ.
  - An improvement from 0.2 to 0.15 is helpful; from 0.15 to 0.1 is not.



Adjusting the sample size

- When the population is $\text{ND}(2, 0.2)$ and the sample size is $n = 4$, the probability of punishment is 0.045.
- If the quality control officer increases the sample size n, the probability will decrease.
- $\mu = 2$ is actually ideal. A larger sample size makes the officer less likely to make a mistake.



Distribution of the sample mean

- So now we have one general conclusion: When we sample from a normal population, the sample mean is also normal.
  - And its mean and standard deviation are μ and $\frac{\sigma}{\sqrt{n}}$, respectively.
- What if the population is **non-normal**?
- Fortunately, we have a very powerful theorem, the **central limit theorem**, which applies to **any** population.

Central limit theorem

- The theorem says that a sample mean is **approximately normal** when the sample size is **large enough**.

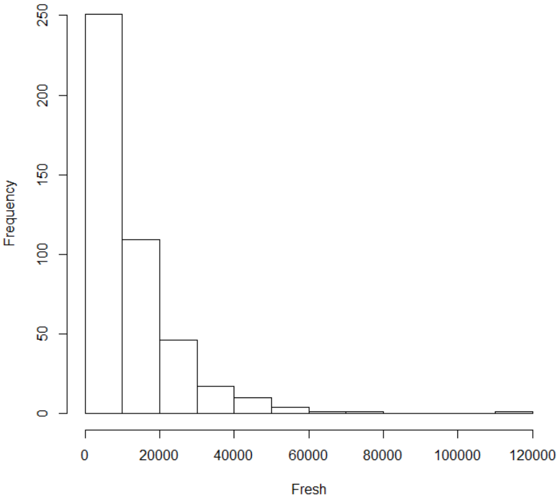
Proposition 4 (Central limit theorem)

Let $\{X_i\}_{i=1,\ldots,n}$ be a size-n random sample from a population with mean μ and standard deviation σ.Let $\bar{x}_n$ be the sample mean. If $\sigma < \infty$, then $\bar{x}_n$ converges to $\mathrm{ND}(\mu, \frac{\sigma}{\sqrt{n}})$ as $n \to \infty$.

- Obviously, we will not try to prove it.
- Let's get the idea with experiments.

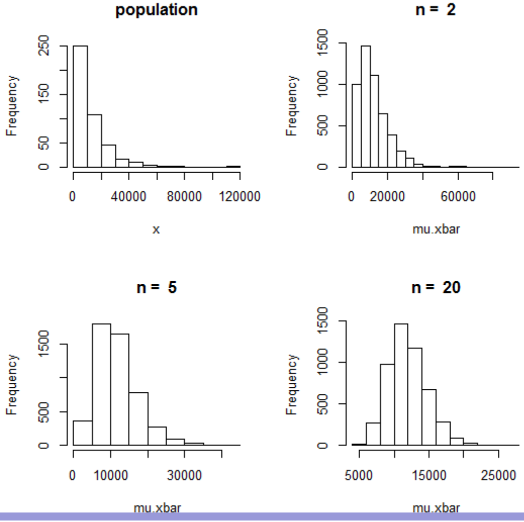
Experiments on the central limit theorem

- Consider our wholesale data again. Let the "Fresh" variable to be our population.
- This population is definitely not normal.
- It is highly skewed to the right (positively skewed).



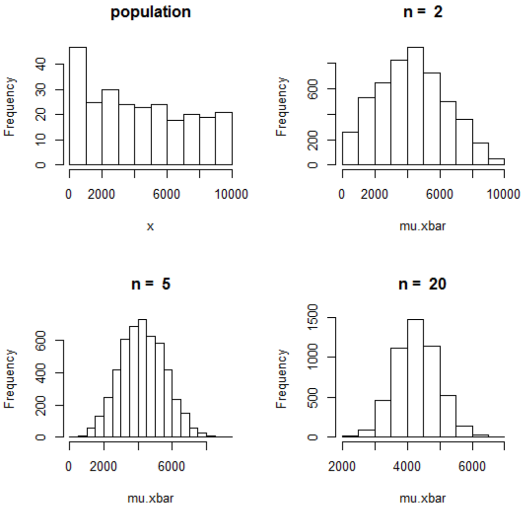
Experiments on the central limit theorem

- When the sample size n is small, the sample mean does not look like normal.
- When the sample size n is **large enough**, the sample mean is **approximately normal**.



Experiments on the central limit theorem

- When the population is **uniform**, the sample mean still becomes normal when n is large enough.
  - Those values in "Fresh" that are less than 10000.
- We only need a small n for the sample mean to be normal.



Timing for central limit theorem

- In short, the central limit theorem says that, for any population, the sample mean will be approximately normally distributed as long as the sample size is large enough.
  - With the distribution of the sample mean, we may then calculate all the probabilities of interests.
- How large is "large enough"?
- In practice, typically $n \geq 30$ is believed to be large enough.

Road map

- Sample means.
- Distributions of sample means.
- **Sample proportions**.

Means vs. proportions

- For interval or ratio data, we have defined sample means.
  - We have studied the distributions of sample means.
- For **ordinal or nominal** data, there is no sample mean.
  - Instead, there are sample **proportions**.

Population proportions

- How to know the **proportions** of girls and boys in NTU?
- We first **label** girls as 0 and boys as 1.
- Let $X_i \in \{0, 1\}$ be the sex of student i, $i = 1, ..., N$.
- Then the **population proportion** of boys is defined as

$$p = \frac{1}{N} \sum_{i=1}^{N} X_i$$

- The population proportion of girls is $1 - p$.

Sample proportions

- Let $\{X_i\}_{i=1,...,N}$ be the population.
- With a sample size n, let $\{X_i\}_{i=1,...,n}$ be a sample. Suppose X_i and X_j are independent for all $i \neq j$.
  - E.g., 100 randomly selected students.
- Then the **sample proportion** is defined as

$$\hat{p} = \frac{1}{n} \sum_{i=1}^{n} X_i.$$

- The population proportion p is deterministic (though unknown) while the sample proportion $\hat{p}$ is **random**.
- We are interested in the distribution of $\hat{p}$.

Bernoulli random variables

- A random variable X whose sample space is $\{0, 1\}$ is a **binary** variable.
- Let $p = \Pr(X = 1)$ be the **success probability**.
- We say X follows a **Bernoulli distribution** with probability p.
  - Denoted as $X \sim \text{Ber}(p)$.
- We may calculate its expected value:

$$\mu_X = p \times 1 + (1 - p) \times 0 = p.$$

- We may calculate its standard deviation:

$$\sigma_X^2 = p(1 - p)^2 + (1 - p)(0 - p)^2 = p(1 - p), \text{ and}$$
$$\sigma_X = \sqrt{p(1 - p)}.$$

Distributions of sample proportions

- What is the distribution of the sample proportion

$$\hat{p} = \frac{1}{n}\sum_{i=1}^{n} X_i?$$

- Note that the sample proportion is a special type of **sample mean**!
  - It is special as $X_i \in \{0, 1\}$.
  - However, it is still a sample mean. The arithmetic average does have a physical meaning: the proportion.
- We may apply the **central limit theorem**:
  - If $n \geq 30$, the sample proportion is approximately normally distributed.
  - Its mean and standard deviations are

$$\mu_{\hat{p}} = \mu = p \quad \text{and} \quad \sigma_{\hat{p}} = \frac{\sigma}{\sqrt{n}} = \sqrt{\frac{p(1-p)}{n}}.$$

Sample proportions: An example

- In 2011, there are 19756 boys and 13324 girls in NTU.
- The population proportion of boys is

$$p = \frac{19756}{33080} \approx 0.597.$$

- Let's sample 100 students and find the sample proportion $\hat{p}$.
  - What is the distribution of $\hat{p}$?
  - What is the probability that to see fewer boys than girls?

Sample proportions: An example

- What is the distribution of $\hat{p}$?
  - As $n \geq 30$, it follows a normal distribution.
  - Its mean is $p \approx 0.597$.
  - Its standard deviation is $\sqrt{\frac{p(1-p)}{n}} \approx 0.049$.
- The probability that $\hat{p} < 0.5$ is

$$\Pr(\hat{p} < 0.5) \approx 0.024.$$

- Summary:
  - A sample proportion "is" a sample mean of qualitative data.
  - It is normal when the sample size is large enough.
  - Thanks to the central limit theorem.