GMBA 7098: Statistics and Data Analysis (Fall 2014)

Sampling and Sampling Distributions

Ling-Chieh Kung

Department of Information Management
National Taiwan University

October 27, 2014

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 sample is typically 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

- My mom asks me to produce bags of candies that weigh within 1.8 and 2.2 kg. She allows a 5% defective rate.
  - In a random sample of 1 bag of candies, suppose it weighs 2.1 kg. How likely that the true defective rate is less than 5%?
  - What if the average weight of 5 bags in a random sample is 2.1 kg?
  - What if the sample size is 10? 50? 100?
  - What if the mean is 2.18 kg?
- Recall the three pairs of concepts:
  - Populations vs. samples.
  - Parameters vs. statistics.
  - Census vs. **sampling**.
- To estimate or test parameters of interests, we rely on statistics obtained from our sample.
- We need to know the sampling distribution of those statistics.

Road map

- **Sampling techniques**.
- Sample means from a normal population.
- Sample means from a non-normal population.

Random vs. nonrandom sampling

- Sampling is the process of selecting a **subset** of entities from the whole population.
- Sampling can be **random** or **nonrandom**.
- If random, whether an entity is selected is **probabilistic**.
  - Randomly select 1000 phone numbers on the telephone book and then call them.
- If nonrandom, it is **deterministic**.
  - Ask all your classmates for their preferences on iOS/Android.
- Most statistical methods are **only** for random sampling.
- Some popular random sampling techniques:
  - Simple random sampling.
  - Stratified random sampling.
  - Cluster (or area) random sampling.

Simple random sampling

- In simple random sampling, each entity has **the same probability** of being selected.
- Each entity is assigned a label (from 1 to N). Then a sequence of n random numbers, each between 1 and N, are generated.
- One needs a **random number generator**.
  - E.g., `sample()` in R.

Simple random sampling

- Suppose we want to study all students graduated from NTU IM regarding the number of units they took before their graduation.
  - $N = 1000$.
  - For each student, whether she/he double majored, the year of graduation, and the number of units are recorded.

i	1	2	3	4	5	6	7	...	1000
Double major	Yes	No	No	No	Yes	No	No		Yes
Class	1997	1998	2002	1997	2006	2010	1997	...	2011
Unit	198	168	172	159	204	163	155		171

  - Suppose we want to sample $n = 200$ students.

Simple random sampling

- To run simple random sampling, we first generate a sequence of 200 random numbers:
  - Suppose they are 2, 198, 7, 268, 852, ..., 93, and 674.
  - Sampling with or without replacement?
- Then the corresponding 200 students will be sampled. Their information will then be collected.

i	1	**2**	3	4	5	6	**7**	...	1000
Double major	Yes	**No**	No	No	Yes	No	**No**		Yes
Class	1997	**1998**	2002	1997	2006	2010	**1997**	...	2011
Unit	198	**168**	172	159	204	163	**155**		171

- We may then calculate the sample mean, sample variance, etc.

Simple random sampling

- The good part of simple random sampling is **simple**.
- However, it may result in **nonrepresentative** samples.
- In simple random sampling, there are some possibilities that **too much** data we sample fall in **the same stratum**.
  - They have the same property.
  - For example, it is possible that all 200 students in our sample did not double major.
  - The sample is thus nonrepresentative.

Simple random sampling

- As another example, suppose we want to sample 1000 voters in Taiwan regarding their preferences on two candidates. If we use simple random sampling, what may happen?
  - It is possible that 65% of the 1000 voters are men while in Taiwan only around 51% voters are men.
  - It is possible that 40% of the 1000 voters are from Taipei while in Taiwan only around 28% voters live in Taipei.
- How to fix this problem?

Stratified random sampling

- We may apply **stratified random sampling**.
- We first split the whole population into several **strata**.
  - Data in **one** stratum should be (relatively) **homogeneous**.
  - Data in **different** strata should be (relatively) **heterogeneous**.
- We then use simple random sampling for each stratum.
- Suppose 100 students double majored, then we can split the whole population into two strata:

Stratum	Strata size
Double major	100
No double major	900

Stratified random sampling

- Now we want to sample 200 students.
- If we sample $200 \times \frac{100}{1000} = 20$ students from the double-major stratum and 180 ones from the other stratum, we have adopted **proportionate** stratified random sampling.

Stratum	Strata size	Number of samples
Double major	100	20
No double major	900	180

- If the opinions in some strata are more important, we may adopt **disproportionate** stratified random sampling.
  - E.g., opening a nuclear power station at a particular place.

Stratified random sampling

- We may further split the population into more strata.
  - Double major: Yes or no.
  - Class: 1994-1998, 1999-2003, 2004-2008, or 2009-2012.
  - This stratification makes sense **only if** students in different classes tend to take different numbers of units.
- Stratified random sampling is good in **reducing sample error**.
- But it can be hard to identify a reasonable stratification.
- It is also more **costly** and **time-consuming**.

Cluster (or area) random sampling

- Imagine that you are going to introduce a new product into all the retail stores in Taiwan.
- If the product is actually unpopular, an introduction with a large quantity will incur a huge lost.
- How to get an idea about the popularity?
- Typically we first try to introduce the product **in a small area**. We put the product on the shelves only in those stores in the specified area.
- This is the idea of **cluster (or area) random sampling**.
  - Those consumers in the area form a sample.

Cluster (or area) random sampling

- In stratified random sampling, we define strata.
- Similarly, in cluster random sampling, we define **clusters**.
- However, instead of doing simple random sampling in each strata, we will only choose **one or some clusters** and then collect **all** the data in these clusters.
  - If a cluster is too large, we may further split it into multiple **second-stage clusters**.
- Therefore, we want data in a cluster to be **heterogeneous**, and data across clusters somewhat **homogeneous**.

Cluster (or area) random sampling

- In practice, the main application of cluster random sampling is to understand the popularity of **new products**. Those chosen cities (counties, states, etc.) are called **test market cities** (counties, states, etc.).
- People use cluster random sampling in this case because of its feasibility and convenience.
- We should select test market cities whose population profiles are similar to that of the entire country.

Nonrandom sampling

- Sometimes we do **nonrandom sampling**.
- Convenience sampling.
  - The researcher sample data that are easy to sample.
- Judgment sampling.
  - The researcher decides who to ask or what data to collect.
- Quota sampling.
  - In each stratum, we use whatever method that is easy to fill the quota, a predetermined number of samples in the stratum.
- Snowball sampling.
  - Once we ask one person, we ask her/him to suggest others.
- Nonrandom sampling **cannot** be analyzed by the statistical methods we introduce in this course.

Road map

- Sampling techniques.
- **Sample means from a normal population**.
- Sample means from a non-normal population.

Sample means

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

Definition 1

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

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

is the sample mean.

- Let's assume that X_i and X_j are independent for all $i \neq j$.
  - We will discuss when is this assumption reasonable.

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 $\mathbb{E}[\bar{x}]$ and variance $\text{Var}(\bar{x})$.
  - These two numbers are also **fixed**.
  - They are sometimes denoted as $\mu_{\bar{x}}$ and $\sigma^2_{\bar{x}}$, respectively.
- We have a formula to calculate $\mu_{\bar{x}}$ and $\sigma^2_{\bar{x}}$. Before that, let's do an experiment first.

My bags of candies

- Suppose that I have produced 1000 bags of candies. Their weights follow a normal distribution with mean $\mu = 2$ and standard deviation $\sigma = 0.2$.
- Suppose my mom decides to sample 4 bags and calculate the sample mean $\bar{x}$. She will punish me if the sample mean is not in $[1.8, 2.2]$.
  - What is the mean of the sample mean $\mu_{\bar{x}}$?
  - What is the standard deviation of the sample mean $\sigma_{\bar{x}}$?
  - What is the distribution of the sample mean $\bar{x}$?
  - What is the probability that the sample mean is above 2?
  - What is the probability that I will be punished?

Experiments for estimating the probabilities

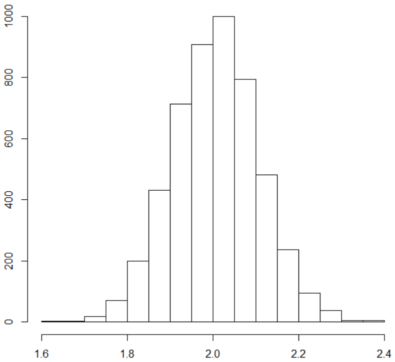
- Let's do an experiment.
  - First, use `x <- rnorm(1000, 2, 0.1)` to generate the weights of 1000 bags of candies.
  - Then use `mu.xbar <- mean(sample(x, 4))` to randomly sample 4 bags and calculate their sample mean.
  - Then repeat this for 5000 times!

Experiments for estimating the probabilities

```
trial <- 5000
x <- rnorm(1000, 2, 0.2)
mu.xbar <- rep(0, trial)
for(i in 1:trial)
{
  mu.xbar[i] <- mean(sample(x, 4))
}
mean(mu.xbar) # mean of sample mean
sd(mu.xbar) # standard deviation of sample mean
hist(mu.xbar) # distribution of sample mean
length(which(mu.xbar > 2)) / trial # Pr(xbar > 2)
f1 <- length(which(mu.xbar < 1.8))
f2 <- length(which(mu.xbar > 2.2))
(f1 + f2) / trial # Pr(xbar < 1.8 or xbar > 2.2)
```

Experiments for estimating the probabilities

- The result of my experiment:
  - The mean of sample means is 1.993741.
  - The standard deviation of sample mean is 0.1002187.
  - The distribution looks like a normal distribution.
  - The probability for the sample mean to be above 2 is 0.473.
  - The probability for me to be punished is 0.0468.
- Is $\bar{x} \sim \text{ND}(2, 0.1)$?



Experiments for estimating the probabilities

- If we do multiple rounds of this experiment:

Round	Mean	Standard deviation	$\Pr(\bar{x} > 2)$	$\Pr(\bar{x} < 1.8) + \Pr(\bar{x} > 2.2)$
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?

Sampling from a normal population

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

Proposition 1

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

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

Sampling from a normal population

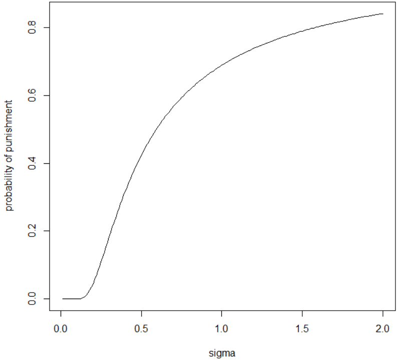
- What is the mean of the sample mean $\mu_{\bar{x}}$?
  - $\mu_{\bar{x}} = \mu = 80$.
- What is the standard deviation of the sample mean $\sigma_{\bar{x}}$?
  - $\text{Var}(\bar{x}) = \frac{\sigma^2}{n} = \frac{0.04}{4} = 0.01$. The standard deviation is $\sqrt{0.01} = 0.1$.
- What is the distribution of the sample mean $\bar{x}$?
  - $\text{ND}(2, 0.1)$.
- What is the probability that the sample mean is above 2?
  - $\Pr(\bar{x} > 2) = 0.5$.
- What is the probability that I will be punished?
  - $\Pr(\bar{x} < 1.8) + \Pr(\bar{x} > 2.2) \approx 0.045$.
  - Use `pnorm(1.8, 2, 0.1)` and `1 - pnorm(2.2, 2, 0.1)`.
- Summary: Because the population is normal with $\mu = 2$, $\sigma = 0.2$, and $n = 4$, indeed we have $\bar{x} \sim \text{ND}(0.2, 0.1)$.

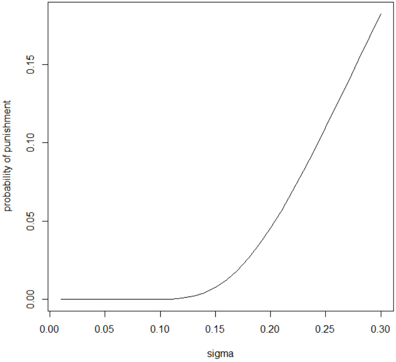
Means and variances of sample means

- 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.
  - $\mathrm{Var}(\bar{x})$; a constant.
- The sample variance also has its mean and variance.

Adjusting the standard deviation

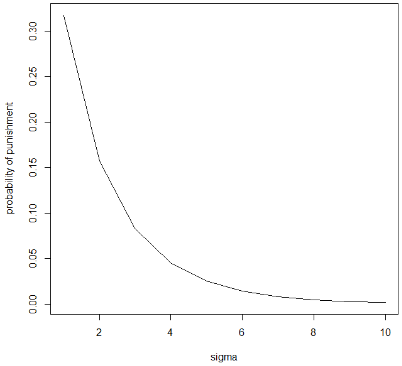
- When the population is ND($\mu = 2, \sigma = 0.2$) and the sample size is $n = 4$, the probability of punishment is 0.045.
- If I adjust my standard deviation σ (by paying more or less attention to my production process), the probability will change.





Adjusting the sample size

- When the population is $\mathrm{ND}(2, 0.2)$ and the sample size is $n = 4$, the probability of punishment is 0.045.
- If my mom adjust the sample size n, the probability will also change.
  - Why is it decreasing in n?
  - What is the implication?



Road map

- Sampling techniques.
- Sample means from a normal population.
- **Sample means from a non-normal population**.

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 2 (Central limit theorem)

Let $\{X_i\}_{i=1,...,n}$ be a size-n random sample from a population with mean μ and standard deviation σ, i.e., $\mathbb{E}[X_i] = \mu$ and $\mathrm{Var}(X_i) = \sigma^2$. Let $\bar{x}$ be the sample mean. If $\sigma < \infty$, then

$$Z_n \equiv \frac{\bar{x} - \mu}{\sigma / \sqrt{n}}$$

converges to $Z \sim ND(0,1)$ 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:[1]

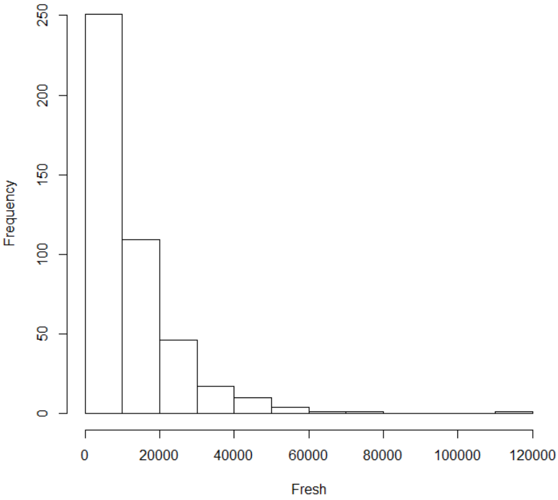
	Channel	Region	Fresh	Milk	Grocery	Frozen	D_Paper	Delicassen
1	1	1	30624	7209	4897	18711	763	2876
2	1	1	11686	2154	6824	3527	592	697
3	1	1	9670	2280	2112	520	402	347
4	1	1	25203	11487	9490	5065	284	6854
5	1	1	583	685	2216	469	954	18
6	1	1	1956	891	5226	1383	5	1328

- Let's ignore `Channel` and `Region` and consider the whole `Fresh` column as our population.

[1] To load this data set into your R program, first set the work directory and then use the function `read.table(file, header = TRUE)`.

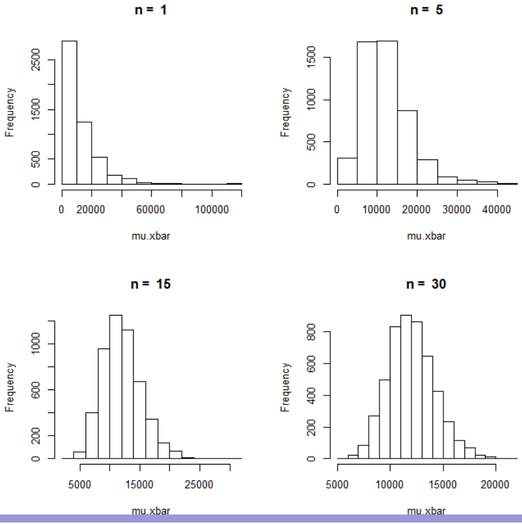
Experiments on the central limit theorem

- This population is definitely not normal.
- It is highly skewed to the right (positively skewed).



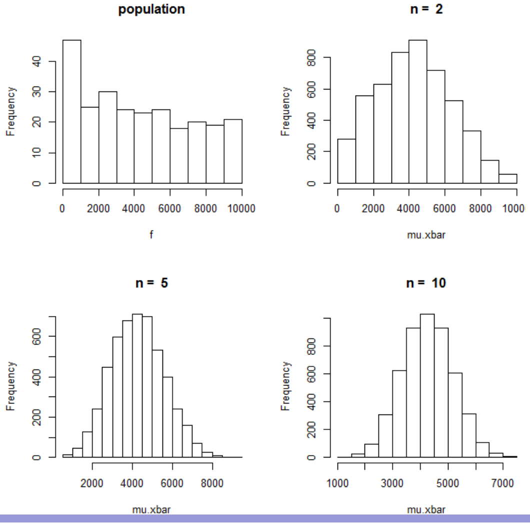
Experiments on the central limit theorem

- 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.