Statistics and Data Analysis

Distributions and Sampling

Ling-Chieh Kung

Department of Information Management
National Taiwan University

Introduction

- We have learned two separate topics.
  - Descriptive statistics: visualization and summarization of **existing data** to understand the data.
  - Probability: using **assumed probability distributions** (for, e.g., inventory management).
- Now it is time to connect them.
- This lecture:
  - We will study how to **estimate the distribution** of a random variable from existing data.
  - We will study how to **sample** from a population.
  - We will study **sampling distribution**: the distribution of a sample.

Road map

- **Estimating probability distributions**.
  - When the sample space is small.
  - When the sample space is large.
- Sampling techniques.
- Sample means.
- Distribution of sample means.

Estimating probability distributions

- Given a random variable, how to know its **probability distribution**?
  - Given a population of people, what will be the age of a randomly selected person?
  - Given a potential customer, will she/he buy my product?
  - Given a web page and a time horizon, how many visitors will we have?
  - Given a batch of products, how many will pass a given quality standard?
- We want more than one value; we want a **distribution**.
  - For each possible value, how likely it will be realized.
- To do the estimation, we **do experiments** or **collect past data**.

Estimating probability distributions

- Given a random variable, how to know its probability distribution?
  - Given a random variable X, how to get $F(x) = \Pr(X \leq x)$?
- Given a coin, how to know whether it is fair?
  - Let X be the outcome of tossing a coin.
  - Let $X = 1$ if the outcome is a head or 0 otherwise.
  - Let $\Pr(X = 1) = p = 1 - \Pr(X = 0)$.
  - Is $p = 0.5$?

Frequency and probability distributions

- The most straightforward way: Use a **frequency distribution** to be the **probability distribution**.
  - We may flip the coin for 100 times.
  - Suppose we see 46 heads and 54 tails.
  - We may "estimate" that $p = 0.46$.
- A frequency distribution and a probability distribution are different.
  - A frequency distribution is what we observe. It is an outcome of investigating a **sample**.
  - A probability distribution is what governs the random variable. It is a property of a **population**.
- The frequency distribution will be "approximately" the probability distribution if we have enough data.

Estimating a discrete distribution

- Consider a discrete random variable whose number of possible values are not too many.
- Let X be the random variable and S be the sample space.
  - We are saying that S does not contain too many values.
- We want to know $\Pr(X = x) = p_x$ for any $x \in S$.
- In this case, let $\{x_i\}_{i=1,...,n}$ be our observed sample data. Given a value $x \in S$, we may simply use the **proportion**

$$\frac{\text{number of } x_i\text{s that is } x}{\text{number of } x_i\text{s}}$$

  to be our estimated p_x.
  - Sometimes manual adjustments are helpful.

When the sample space is small: example

- A data set records the daily weather for the 731 days in two years.
  - 1 for sunny or partly cloudy, 2 for misty and cloudy, 3 for light snow or light rain, and 4 for heavy snow or thunderstorm.
- Let X be the daily weather for a future day. We have $S = \{1, 2, 3, 4\}$.
- By looking at the data set, we obtain

x	1	2	3	4
Frequency	463	247	21	0
Proportion	0.633	0.338	0.029	0

- Let $p_i = \Pr(X = i)$, we then estimate that $p_1 = 0.633$, $p_2 = 0.338$, $p_3 = 0.029$, and $p_4 = 0$.
  - This estimation is just based on a sample. It is never "right."
  - Manual adjustments based on experiences or knowledge are allowed.
  - E.g., we may adjust it to $p_1 = 0.65$, $p_2 = 0.3$, $p_3 = 0.03$, and $p_4 = 0.02$.

When the sample space is large

- When the sample space is large, this method is not very helpful.
  - E.g., a data set records the daily bike rentals in 731 days.
  - Let X be the daily bike rental.
  - X is discrete. Its sample space contains more than 8000 values.
  - The naive counting for frequencies does not help.
- In this case, we rely on **frequency distributions** to estimate the probability for the value to be **within a class**.

When the sample space is large: example

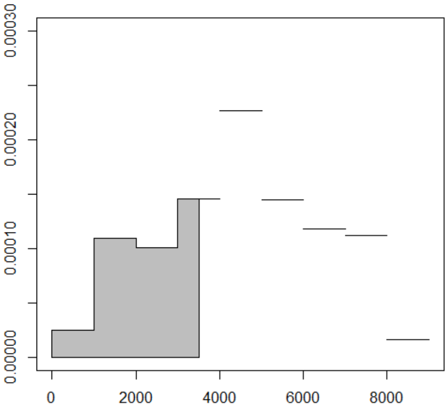
- Let X be the daily bike rental for a given day in the future.
- A data set contains the daily bike rentals in 731 days.
- We obtain the frequency distribution of daily bike rentals:

x	Frequency	Proportion
$[0, 1000)$	18	0.025
$[1000, 2000)$	80	0.109
$[2000, 3000)$	74	0.101
$[3000, 4000)$	107	0.146
$[4000, 5000)$	166	0.227
$[5000, 6000)$	106	0.145
$[6000, 7000)$	86	0.118
$[7000, 8000)$	82	0.112
$[8000, 9000)$	12	0.016

Generating uniform distributions for classes

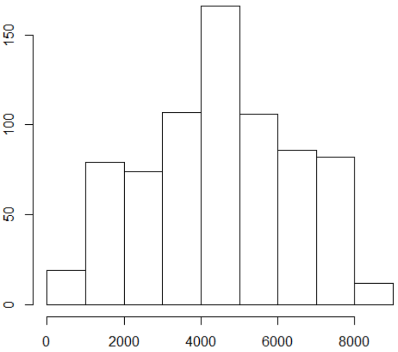
x	Proportion
$[0, 1000)$	0.025
$[1000, 2000)$	0.109
$[2000, 3000)$	0.101
$[3000, 4000)$	0.146
$[4000, 5000)$	0.227
$[5000, 6000)$	0.145
$[6000, 7000)$	0.118
$[7000, 8000)$	0.112
$[8000, 9000)$	0.016

- The cdf $F(x)$ can be constructed:



Distribution fitting

- There are two reasons not to use the 9-class distribution.
  - It is hard to use.
  - It is obtained from a sample.
- We typically want to **fit a theoretical distribution** to the observed distribution.
  - We "believe" that the population follows a certain distribution.
  - E.g., the histogram suggests us that the daily bike rental may actually be normal.
  - We do **distribution fitting**.

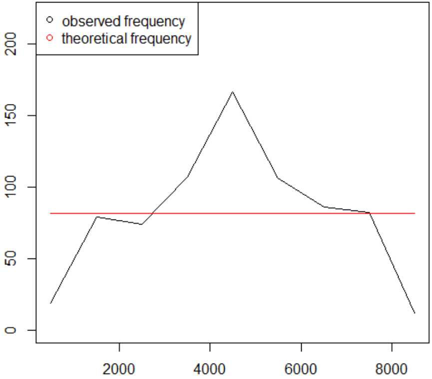


Distribution fitting

- We want to **fit** a distribution to a histogram.
- To do so, we select a distribution (by investigation and some experiences), find the theoretical frequency for each class following the distribution, and then plot the two sequences of frequencies together.
  - **Observed frequencies** are from the histogram.
  - **Theoretical frequencies** are from the assumed distribution.
  - If the two sequences are "close to each other," the fitting is appropriate.
- To visualize the fitting, we may depict the the assumed and observed distributions as two frequency polygons.
- We may try a few assumed distributions and select the best one.

Distribution fitting: uniform distribution

- Consider the daily bike rental example again.
- If we assume $X \sim \text{Uni}(0, 9000)$, the theoretical frequency of each class would be $\frac{731}{9} \approx 81.2$.
- We then compare those theoretical frequencies with the observed frequencies 18, 80, 74, 107, 166, etc.
- X does not seem to be $\text{Uni}(0, 9000)$.



Distribution fitting: normal distribution

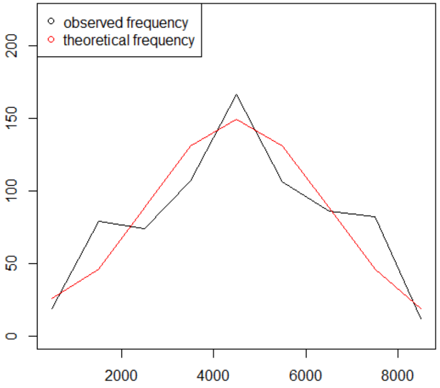
- Let's try to fit a normal distribution to the histogram.
- We need to choose a mean and a standard deviation to construct the normal curve.
  - A typical way: Use the sample mean and sample standard deviation.
  - For the 731 values, we have $\bar{x} \approx 4504$ and $s \approx 1937$.
- If $X \sim \text{ND}(4504, 1937)$, we have:[1]

$[l, u)$	Theoretical proportion $\Pr(l \leq X < u)$	Theoretical frequency $731 \times \Pr(l \leq X < u)$
$[0, 1000)$	0.035	25.75
$[1000, 2000)$	0.063	45.92
	$\vdots$	
$[8000, 9000)$	0.025	18.59

[1]In MS Excel, use `NORM.DIST` to find $\Pr(l \leq X < u)$.

Distribution fitting: normal distribution

- If we assume $X \sim \text{ND}(4504, 1937)$:
- $\text{ND}(4504, 1937)$ seems to fit the observed data better.
- Further trials and adjustments are always possible.



Summary

- We want to estimate the probability distribution of a random variable.
- When the sample space is small:
  - Use the relative frequency of each possible value to be its probability.
- When the sample space is large:
  - Construct a frequency distribution.
  - Use the relative frequency of each class to be its probability.
  - Look at a histogram and guess which probability distribution fits it.
  - Find the theoretical frequency for each class.
  - Compare the observed and theoretical frequencies.
  - Stop when the overall difference is "small."[2]
- Human judgments may be needed.

[2] For example, one may try a few theoretical distributions and select the one with the minimum error.

Road map

- Estimating probability distributions.
- **Sampling techniques**.
- Sample means.
- Distribution of sample means.

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.

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., `RAND()` and `RANDBETWEEN()` in MS Excel.
- Sampling with or without replacement:
  - **With replacement**: One may be selected for many times.
  - **Without replacement**: One may be selected for at most once.

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 not representative.
- 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 stratified random sampling.[3]

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

[3] More precisely, we say this is proportionate stratified random sampling. If the proportions of entities sampled from the strata are not identical, that is disproportionate stratified random sampling.

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**.

Road map

- Estimating probability distributions.
- Sampling techniques.
- **Sample means**.
- Distribution of sample means.

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 $1.8 < \mu < 2.2$?
- 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 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.).
  - The probability distribution of a sample is a **sampling distribution**.

Sample means

- We will focus on the **sample mean**, one of the most important statistics, to illustrate the concept.

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 $\mu_{\bar{x}}$, $\sigma_{\bar{x}}^2$, 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

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

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 $\bar{x}_2$,

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

  - 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 $\bar{x}$!
  - Is it normal? Is it uniform? Is it something else?

Road map

- Estimating probability distributions.
- Sampling techniques.
- Sample means.
- **Distribution of sample means**.

Sampling from a normal population

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

Proposition 3

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

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

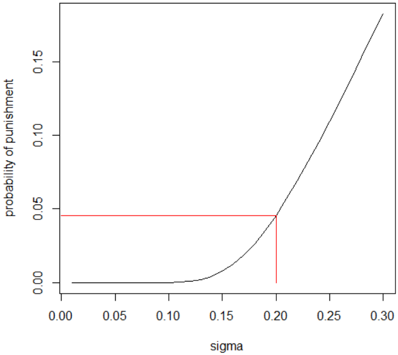
- We already know that $\mu_{\bar{x}} = \mu$ and $\sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}}$. This is true regardless of the population distribution.
- When the population is normal, the sample mean will also be normal.

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 $\text{ND}(2, 0.1)$.
  - $\Pr(\bar{x} < 1.8) + \Pr(\bar{x} > 2.2) \approx 0.045$.

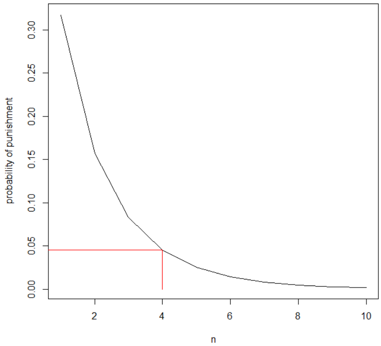
Adjusting the standard deviation

- 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.



Central limit theorem

- When the population is normal, the sample mean is also normal.
- What if the population is **non-normal**?
- The **central limit theorem** says that, for any population, a sample mean is **approximately normal** if the sample size is **large enough**.

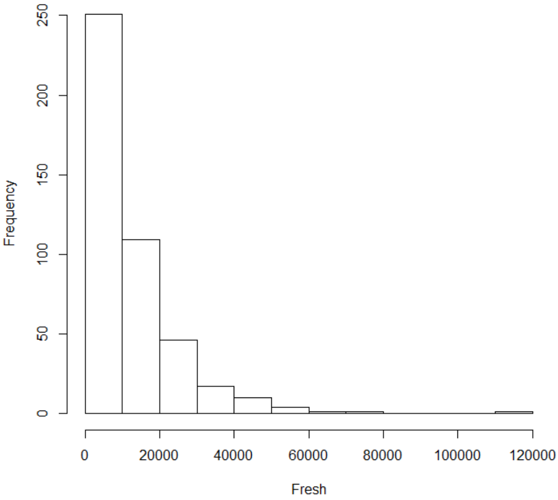
Proposition 4 (Central limit theorem)

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.

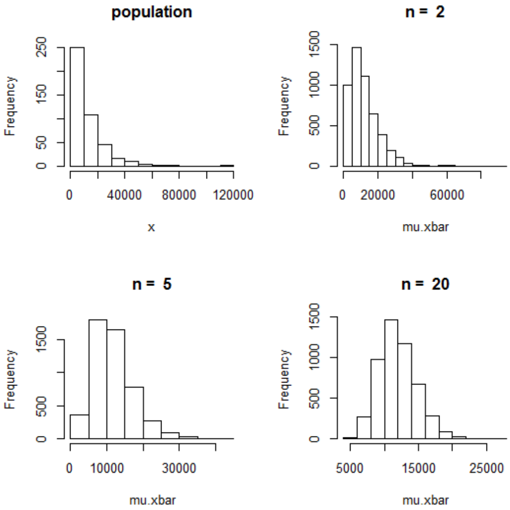
Experiments on the central limit theorem

- 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).



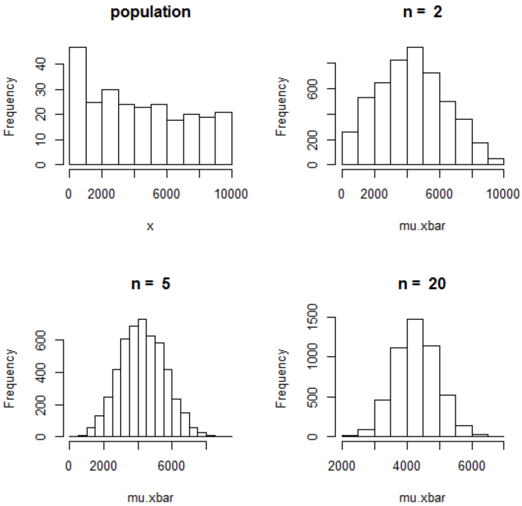
Experiments on the central limit theorem

- 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.