San José State University

Math 161A: Applied Probability & Statistics

Sampling distributions

Prof. Guangliang Chen

Chapter 1 Descriptive statistics

Section 5.3 Statistics and their distributions

Section 5.4 The distribution of the sample mean

Introduction

So far, we have covered the distribution of a single random variable (discrete or continuous) and the joint distribution of two discrete random variables.

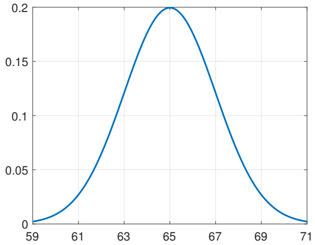
Sampling distributions concern the randomness associated to a **statistic** based on a **random sample** from a **population**.

It serves as the bridge between probability and statistics.

We present this important concept using a practical example - egg weight (see next slide).

Motivating example

Suppose that the weights (in grams) of brown eggs produced at a local farm have a normal distribution: $X \sim N(65, 2^2)$.

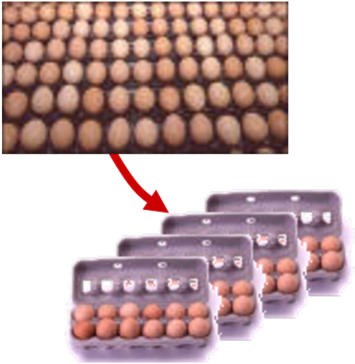


Those eggs are divided into cartons of size 12, to be sold on the market.

You randomly select a carton and measure the weights of all the 12 eggs in it.

Let $\bar{X}$ be their **average weight**.

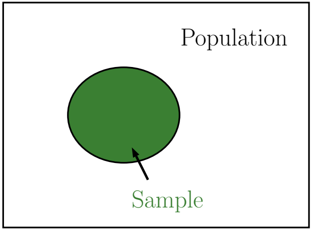
$\bar{X}$ clearly may vary from carton to carton, and thus is a (continuous) random variable.



Question: What is the distribution of $\bar{X}$?

The above question is about the **sampling distribution of a statistic**.

- **Population**: all brown eggs produced at the farm
- **Sample**: a carton of 12 eggs
- **Statistic**: $\bar{X}$ (average weight of the 12 eggs in the sample)



To study the distribution of $\bar{X}$, we denote individual weights of the 12 to-be-selected eggs as $X_1, \ldots, X_{12}$.

We then have

$$\bar{X} = \frac{X_1 + \cdots + X_{12}}{12}.$$

What we know about $X_1, \ldots, X_{12}$:
They are *identically and independently distributed (iid)*:

$$X_1, \ldots, X_{12} \overset{iid}{\sim} N(65, 2^2)$$

and are called a **random sample** (of size 12) from the distribution $N(65, 2^2)$.

Random sample

Def 0.1. More generally, a collection of n random variables $X_1, \ldots, X_n$ is called a random sample if they are

(1) identically distributed according to some pmf/pdf $f(x)$, and

(2) independent.

In short, we write $X_1, \ldots, X_n \overset{\text{iid}}{\sim} f(x)$.

Example 0.1. Suppose you toss a coin (with probability of heads p) repeatedly and independently for a total of n times, and let $X_1, \ldots, X_n$ denote the numerical outcomes of individual trials: 1 (heads) or 0 (tails). This constitutes a random sample from the Bernoulli(p) distribution because

$$X_1, \ldots, X_n \stackrel{\text{iid}}{\sim} \text{Bernoulli}(p).$$

Example 0.2. Let $X_1, \ldots, X_n$ represent n repeated and independent measurements of an object's length. They can be thought of as a random sample from a normal distribution

$$X_1, \ldots, X_n \overset{iid}{\sim} N(\mu, \sigma^2)$$

where

- μ: true length (if the measurement process is unbiased)
- σ^2: variance of the measurement error.

Specific realizations of a random sample

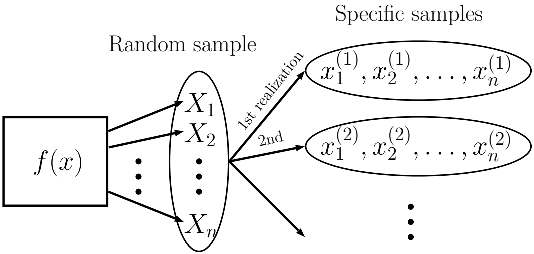
Example 0.3. Suppose you *actually* buy a carton of $n = 12$ eggs from the farm and measure their weights individually. Then you may obtain a data set like the following (called a **specific sample**):

$$x_1 = 65.4,\ x_2 = 65.0,\ x_3 = 64.8,\ x_4 = 65.1,\ x_5 = 64.8,\ x_6 = 64.4,$$
$$x_7 = 65.0,\ x_8 = 65.1,\ x_9 = 65.5,\ x_{10} = 64.8,\ x_{11} = 64.8,\ x_{12} = 65.2$$

Notation. We use lowercase letters such as $x_1, x_2, \ldots$ to represent specific values of the random variables $(X_1, X_2, \ldots)$ in a random sample.

Remark. If we realize the sampling process again, then we may obtain a different set of weights. For example,

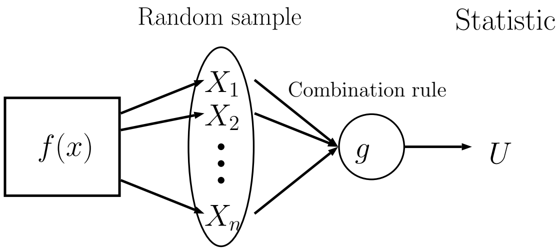
$$x_1 = 65.6,\ x_2 = 64.3,\ x_3 = 64.2,\ x_4 = 65.4,\ x_5 = 64.9,\ x_6 = 64.4,$$
$$x_7 = 65.2,\ x_8 = 65.2,\ x_9 = 65.0,\ x_{10} = 64.7,\ x_{11} = 64.5,\ x_{12} = 65.1$$



Statistic

Def 0.2. Mathematically, a statistic is just a summary of a random sample by certain combination rule g:

$$U = g(X_1, X_2, \ldots, X_n)$$



Remark. Depending on purpose, different statistics may be defined on the same random sample. Two common ones are

- **Sample mean**

$$\bar{X} = \frac{1}{n}\sum_{i=1}^{n} X_i \quad \longleftarrow \text{a measure of center, or location}$$

- **Sample variance**

$$\begin{aligned} S^2 &= \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \bar{X})^2 \quad \longleftarrow \text{a measure of variability} \\ &= \frac{1}{n-1}\left[\sum_{i=1}^{n} X_i^2 - n\cdot \bar{X}^2\right] \end{aligned}$$

Other examples of statistics include

- sample median (also a measure of center)
- sample minimum or maximum
- sample range (i.e., sample maximum - sample minimum)
- trimmed mean

See Chapter 1 for details.

Statistics are random variables

Clearly, for different realizations of the sampling process, the values of the statistic may vary. For the egg weight example (and the statistic $\bar{X}$),

(1) One realization ($\bar{x} = 64.992$):

$$x_1 = 65.4,\ x_2 = 65.0,\ x_3 = 64.8,\ x_4 = 65.1,\ x_5 = 64.8,\ x_6 = 64.4,$$
$$x_7 = 65.0,\ x_8 = 65.1,\ x_9 = 65.5,\ x_{10} = 64.8,\ x_{11} = 64.8,\ x_{12} = 65.2$$

(2) Another realization ($\bar{x} = 64.875$) :

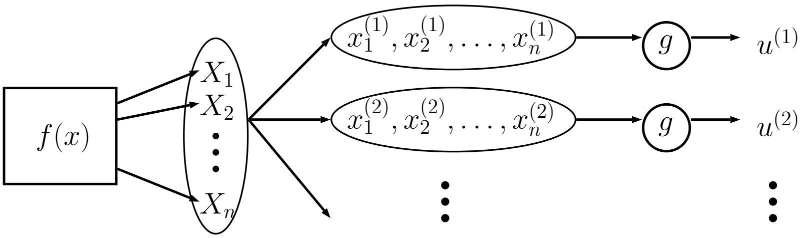
$$x_1 = 65.6,\ x_2 = 64.3,\ x_3 = 64.2,\ x_4 = 65.4,\ x_5 = 64.9,\ x_6 = 64.4,$$
$$x_7 = 65.2,\ x_8 = 65.2,\ x_9 = 65.0,\ x_{10} = 64.7,\ x_{11} = 64.5,\ x_{12} = 65.1$$

Sampling distribution of a statistic

Def 0.3. The probabilistic distribution of a statistic (as a random variable)

$$U = g(X_1, X_2, \ldots, X_n)$$

is called the *sampling distribution* of the statistic.

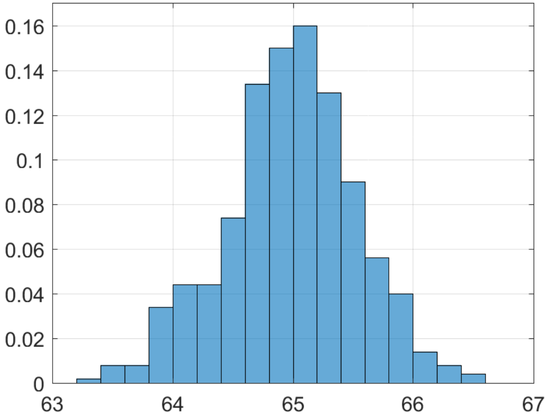


Simulation

We "selected" 500 cartons of eggs randomly from the farm (through computer simulation) and computed their average weights. Below shows 50 observations of $\bar{X}$:

65.0506	64.7592	65.0571	64.9674	65.4973	64.7503	65.0393	64.6714
65.3764	65.2525	65.2012	64.4910	65.6002	65.1868	65.0916	63.8280
65.2636	64.9638	65.2998	65.5587	63.9801	65.3903	64.9052	65.7352
64.6329	64.5109	65.7044	64.3291	65.1044	64.8036	66.0407	65.3560
65.3534	65.4668	64.7394	65.1690	64.5668	64.8478	64.0334	65.7562
64.8553	64.9939	65.6044	64.5237	64.2092	64.5860	65.2096	65.5114
64.6195	65.0312						

We can display all 500 values through a histogram shown below



The sample mean

We focus on the sample mean statistic

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

where

$$X_1, \ldots, X_n \overset{\text{iid}}{\sim} f(x)$$

and

$$\mathrm{E}(X_i) = \mu, \quad \mathrm{Var}(X_i) = \sigma^2, \text{ for all } i.$$

We present three different results for the statistic $\bar{X}$:

1. **Expectation and variance of $\bar{X}$** (for any distribution $f(x)$)
2. **Exact distribution of $\bar{X}$** when $f(x)$ is a normal distribution
3. **Approximate distribution of $\bar{X}$** for nonnomral distributions in the setting of a large sample

General distributions: Expectation and variance of $\bar{X}$

Theorem 0.1. Suppose $X_1, \ldots, X_n \overset{\text{iid}}{\sim} f(x)$, with $\mathrm{E}(X_i) = \mu$ (population mean) and $\mathrm{Var}(X_i) = \sigma^2$ (population variance). Then

$$\mathrm{E}(\bar{X}) = \mu, \quad \mathrm{Var}(\bar{X}) = \frac{\sigma^2}{n}, \quad \mathrm{Std}(\bar{X}) = \frac{\sigma}{\sqrt{n}}$$

Remark. This result does NOT concern the specific distribution of $\bar{X}$!

Proof. By linearity and independence,

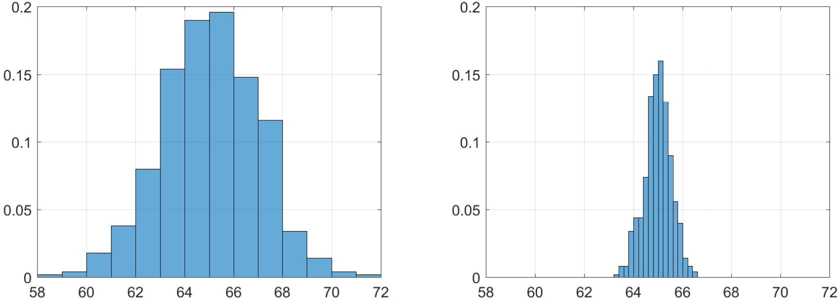
$$\mathrm{E}(\bar{X}) = \frac{1}{n}\left(\mathrm{E}(X_1) + \cdots + \mathrm{E}(X_n)\right) = \frac{1}{n}(\mu + \cdots + \mu) = \mu$$

$$\mathrm{Var}(\bar{X}) = \frac{1}{n^2}(\mathrm{Var}(X_1) + \cdots + \mathrm{Var}(X_n)) = \frac{1}{n^2}(\sigma^2 + \cdots + \sigma^2) = \frac{\sigma^2}{n}.$$

Remark. The theorem indicates that

- expectation of $\bar{X}$ is μ (population mean), and
- variance of $\bar{X}$ is only $1/n$ of the population variance (for single X_i)

Example 0.4. Weights of 500 single eggs (left) and average weights of 500 cartons (right), all selected at random.



Normal populations: Exact distribution of $\bar{X}$

Assume a random sample

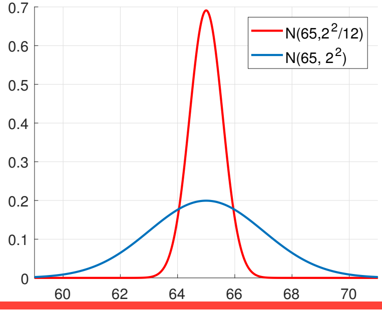
$$X_1, \ldots, X_n \overset{\text{iid}}{\sim} N(\mu, \sigma^2).$$

Theorem 0.2. We have

$$\bar{X} \sim N(\mu, \frac{\sigma^2}{n}).$$

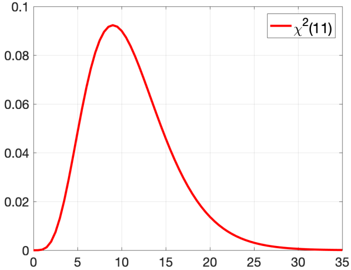
This also implies that

$$\frac{\bar{X} - \mu}{\sigma/\sqrt{n}} \sim N(0, 1).$$



Remark. In this setting of a normal population, the sample variance statistic S^2, after being properly scaled, can be shown to follow a chi-square distribution:

$$\frac{(n-1)S^2}{\sigma^2} \sim \chi^2(n-1) \quad \longleftarrow \quad \text{Gamma}(\alpha = \frac{n-1}{2}, \beta = 2).$$



Example 0.5. In the brown egg example, suppose the population distribution is $N(65, 2^2)$. For a random sample of size 12, what is the probability that the sample mean $\bar{X}$ is within 65 ± 1? What about an individual egg? (Answers: $.9164, .3829$)

Example 0.6. In the library elevator of a large university, there is a sign indicating a 16-person limit as well as a weight limit of 2500 lbs. When the elevator is full, we can think of the 16 people in the elevator as a random sample of people on campus. Suppose that the weight of students, faculty, and staff is normally distributed with a mean weight of 150 lbs and a standard deviation of 27 lbs. What is the probability that the total weight of a random sample of 16 people in the elevator will exceed the weight limit? (*Answer*: $.1762$)

Solution:

Nonnormal populations: Approximate distribution of $\bar{X}$

Assume a random sample

$$X_1, \ldots, X_n \overset{\text{iid}}{\sim} f(x) \quad \longleftarrow \text{any distribution}$$

and that the population has finite mean μ and variance σ^2.

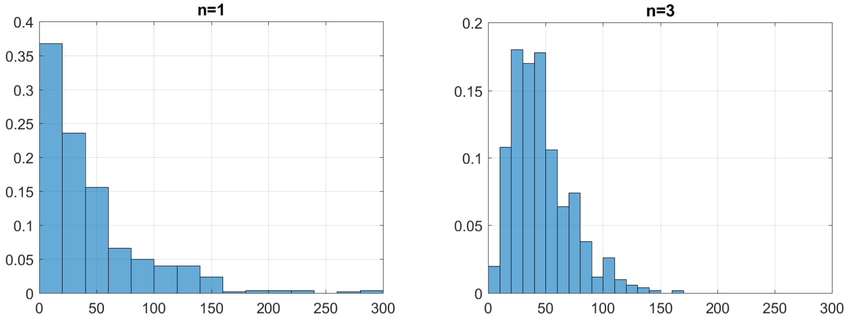
Theorem 0.3. If n is large (30 or greater), then

$$\bar{X} \overset{\text{approx.}}{\sim} N\left(\mu, \frac{\sigma^2}{n}\right), \quad \text{and} \quad \frac{\bar{X} - \mu}{\sigma/\sqrt{n}} \overset{\text{approx.}}{\sim} N(0, 1).$$

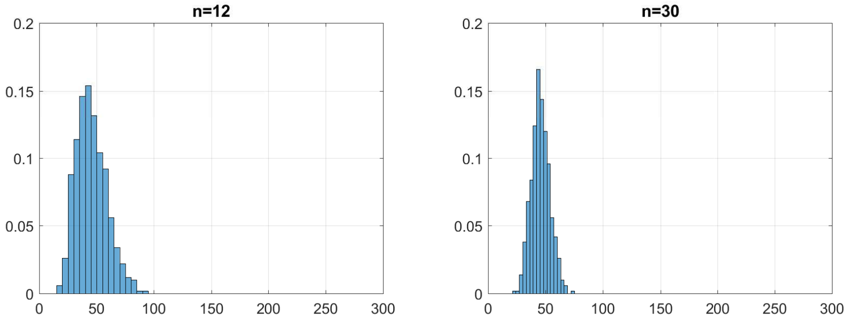
Remark. This is called the **Central Limit Theorem (CLT)**, one of the most important results in probability and statistics.

Example 0.7. Suppose salaries of all SJSU employees follow an exponential distribution with average salary $=$ 45K (which means that $\lambda = \frac{1}{45}$). We draw a random sample of size n from the population, and compute the sample mean $\bar{X}$.

We display the histograms of the simulated values of $\bar{X}$ through 500 repetitions for each of $n = 1, 3, 12, 30$.


n=1
0.4
0.35
0.3
0.25
0.2
0.15
0.1
0.05
0
0
50
100
150
200
250
300
n=3
0.2
0.15
0.1
0.05
0
0
50
100
150
200
250
300

Sampling distributions



Example 0.8 (Employee salary distribution, cont'd)**.** Suppose we draw a random sample of size 30 from the population. Find $P(\bar{X} > 55)$. *Answer: 0.1118 (CLT), 0.1157 (exact)*

The normal approximation to Binomial is a direct consequence of the CLT.

Corollary 0.4. Let $X \sim B(n,p)$. If n is large (i.e., $np, n(1-p) \geq 10$), then

$$\frac{X - np}{\sqrt{np(1-p)}} \overset{\text{approx.}}{\sim} N(0,1)$$

Proof. Consider the experiment of tossing a coin independently for a total of n times, and denote the results by $X_1, \ldots, X_n$. Then

$$X_1, \ldots, X_n \overset{\text{iid}}{\sim} \text{Bernoulli}(p), \quad \text{and} \quad X = \sum_{i=1}^n X_i \sim B(n,p).$$

According to the CLT,

$$\frac{\bar{X} - \mu}{\sigma/\sqrt{n}} = \frac{\bar{X} - p}{\sqrt{p(1-p)}/\sqrt{n}} = \frac{X - np}{\sqrt{np(1-p)}} \overset{\text{approx.}}{\sim} N(0,1).$$

Remark. In the setting of a random sample from a Bernoulli distribution,

$$X_1, \ldots, X_n \overset{\text{iid}}{\sim} \text{Bernoulli}(p)$$

the sample mean

$$\bar{X} = \frac{1}{n} \sum_{i=1}^{n} X_i \qquad \longleftarrow \text{sample proportion } \hat{p}$$

represents the proportion of successes in the sample.

We have showed that if n is large, then

$$\frac{\bar{X} - \mu}{\sigma/\sqrt{n}} = \frac{\hat{p} - p}{\sqrt{p(1-p)/n}} \overset{\text{approx.}}{\sim} N(0, 1).$$

A “large-sample” joke

One day there was a fire in a wastebasket in the Dean’s office and in rushed a physicist, a chemist, and a statistician.

The physicist immediately starts to work on how much energy would have to be removed from the fire to stop the combustion. The chemist works on which reagent would have to be added to the fire to prevent oxidation.

While they are doing this, the statistician is setting fires to all the other wastebaskets in the office.

“What are you doing?” they demanded. “Well to solve the problem, obviously you need a large sample size” the statistician replies.

The distribution of a linear combination

Def 0.4. Given random variables $X_1, \ldots, X_n$ and constants $a_1, \ldots, a_n$,

$$Y = a_1 X_1 + \cdots + a_n X_n = \sum_{i=1}^{n} a_i X_i$$

is called a linear combination of the X_i's.

Example 0.9. For three variables X_1, X_2, X_3, the following are all linear combinations of them: $X_1 + 2X_2 - 3X_3, \frac{1}{3}(X_1 + X_2 + X_3), X_1 - X_2$

Remark. The sample mean is a special linear combination of a random sample $X_1, \ldots, X_n \overset{iid}{\sim} f(x)$ with equal weights: $a_1 = \cdots = a_n = 1/n$.

We have the following general result.

Theorem 0.5. Any linear combination of independent normal random variables is still normal. That is, if

$$X_1 \sim N(\mu_1, \sigma_1^2),\ \ldots,\ X_n \sim N(\mu_n, \sigma_n^2)$$

are independent random variables, then for any constants $a_1, \cdots, a_n$,

$$Y = \sum_{i=1}^n a_i X_i \sim N\left(\sum_{i=1}^n a_i \mu_i, \sum_{i=1}^n a_i^2 \sigma_i^2\right).$$

Remark. This reduces to $\bar{X} \sim N(\mu, \sigma^2/n)$ when $a_1 = \cdots = a_n = 1/n$, $\mu_1 = \cdots = \mu_n = \mu$ and $\sigma_1^2 = \cdots = \sigma_n^2 = \sigma^2$.

Summary

This presentation covers the following:

- **Basic concepts**
  - *Population*: set of all individuals (whose certain characteristic is of interest)
  - *Sample*: a subset of the population (to be measured)
  - *Random sample*: a collection of random variables $X_1, \ldots, X_n \overset{\text{iid}}{\sim} f(x)$, where $f(x)$ represents the pmf/pdf of the population
  - *Statistic*: a numerical summary of the sample, such as $\bar{X}, S^2$

- *Sampling distribution of a statistic*: probabilistic distribution of the statistic as a random variable

- **The sample mean statistic**: For any random sample $X_1, \ldots, X_n \stackrel{\text{iid}}{\sim} f(x)$, define

$$\bar{X} = \frac{1}{n} \sum X_i$$

If the population distribution $f(x)$ has mean μ and variance σ^2, then

$$\mathrm{E}(\bar{X}) = \mu, \quad \mathrm{Var}(\bar{X}) = \frac{\sigma^2}{n}, \quad \mathrm{Std}(\bar{X}) = \frac{\sigma}{\sqrt{n}}$$

- **Sampling distributions of $\bar{X}$**
  - If the population is normal ($N(\mu, \sigma^2)$), then the sample mean has the following sampling distribution:

$$\bar{X} \sim N\left(\mu, \frac{\sigma^2}{n}\right)$$

  - For non-normal populations, if the sample size is large (i.e., $n \geq 30$), then

$$\bar{X} \overset{\text{approx}}{\sim} N\left(\mu, \frac{\sigma^2}{n}\right)$$

This is called the **central limit theorem (CLT)**.