STA13: Elementary Statistics

Lecture 12
Book Sections 4.8 - 4.9

Dmitriy Izyumin

February 9 2018

Population and Sample

The population is the set of ALL subjects of interest.

- All undergrads at UCD
- All copies of an electronic component
- All US citizens

A sample is a subgroup of the population.

- 100 undergrads at UCD
- 1000 copies of an electronic component
- A group of US citizens

Parameters and Statistics

A parameter is a value corresponding to the population.

- Maximum height of all undergrads at UCD
- Average time until failure for an electronic component
- Average lifespan of all US citizens

A statistic is a value corresponding to the sample.

- Maximum height of 100 undergrads at UCD
- Average time until failure among 1000 identical electronic components
- Average lifespan of a sample of US citizens

Some Statistics

Some statistics we'll focus on:

- $\bar{x}$, the sample mean
- s, the sample standard deviation
- s^2, the sample variance
- $\hat{p}$, the sample proportion (next week)

Statistical Inference

We want to reach conclusions about the whole population, but we can only analyze a sample.

We want to know the parameters, but we only know the statistics.

We take what we find in the sample and generalize it to the population.

This is called statistical inference.

Random Variable

An experiment produces random outcomes.

A random variable X takes on numerical values based on the outcomes.

The distribution of X tells us the possible values and their probabilities.

Motivation

We want to estimate the population mean μ.

We take a sample, and obtain the sample mean $\bar{x}$.

Since $\bar{x}$ is a random variable, it has different possible values.

We only see the value corresponding to our sample.

Thus our estimate has a degree of uncertainty.

It would be useful to know how $\bar{x}$ varies from sample to sample so we can quantify the uncertainty of the estimate.

Statistics as Random Variables

A sample is chosen randomly.

The value of a statistic depends on the sample.

IMPORTANT!
The statistic is itself a random variable.

- Different samples may lead to different statistics.
- Some values may be more likely than others.
- We observe just one value.

Sampling Distribution

The distribution of a statistic describes how the values of a statistic vary across different samples of the same size, and taken from the same population.

It is called the sampling distribution.

It depends on:

- The sample size n
- The population distribution and parameters.

The s.d. of a statistic is called the standard error (SE).

Example 1

A family has three children with ages 8, 10, and 12.

Two kids are chosen at random *with replacement*, and their sample mean $\bar{x}$ of their ages is calculated.

What is the sampling distribution of $\bar{x}$?

Example 1

The possibilites for different samples are:

Sample	$\bar{x}$	probability
(8,8)	8	1/9
(8,10)	9	1/9
(8,12)	10	1/9
(10,8)	9	1/9
(10,10)	10	1/9
(10,12)	11	1/9
(12,8)	10	1/9
(12,10)	11	1/9
(12,12)	12	1/9

The sampling distribution of $\bar{x}$ is:

k	$p(k)$
8	1/9
9	2/9
10	3/9
11	2/9
12	1/9

Why the Normal Distribution is Important

Often we look at sums or means of random variables.

- Average height/weight/lifespan of a group of people.
- Total number of car accidents for 10 days.
- Average temperature change over a year.

These sums and means are themselves random variables.

It is often useful to know their sampling distributions.

Central Limit Theorem

Suppose we have *i.i.d.* random variables $X_1, X_2, \cdots, X_n$.

The CLT says that if n is "large," then

- $(X_1 + \cdots + X_n)$ is approximately Normal$(n\mu, \sqrt{n}\sigma)$.
- $\bar{x} = \frac{1}{n}(X_1 + \cdots + X_n)$ is approximately Normal$(\mu, \frac{\sigma}{\sqrt{n}})$.

Notice that $X_1, X_2, \cdots, X_n$ do not have to be normal.

(i.i.d. means "independent and identically distributed").

How large?

How large n has to be for the CLT to take effect depends on the problem.

Rule of Thumb:

Population distribution	Sampling distribution of $\bar{x}$
Normal(μ, σ)	Normal$(\mu, \frac{\sigma}{\sqrt{n}})$ regardless of n
symmetric, mean μ, s.d. σ	approximately Normal$(\mu, \frac{\sigma}{\sqrt{n}})$ even for small n
skewed, mean μ, s.d. σ	approximately Normal$(\mu, \frac{\sigma}{\sqrt{n}})$ only for $n \geq 30$

Effect of n

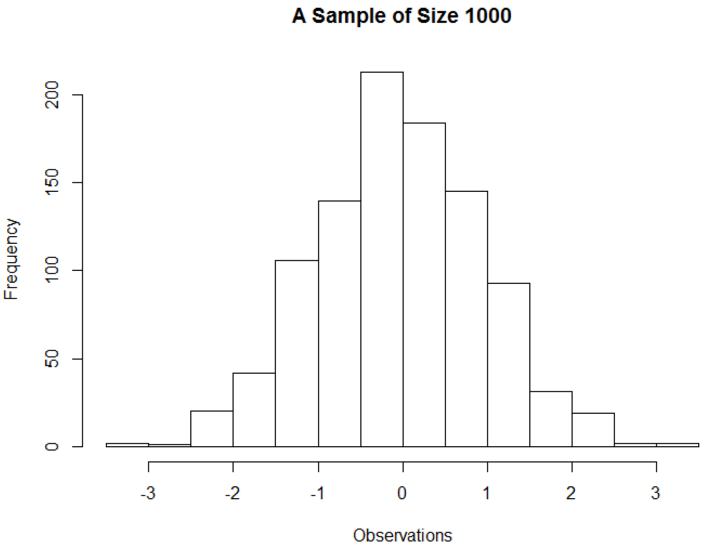
The standard error of $\bar{x}$ is $\frac{\sigma}{\sqrt{n}}$.

As the sample size n increases,

- the standard error decreases
- the pdf curve of $\bar{x}$ becomes more narrow
- the values of $\bar{x}$ become more concentrated around μ
- $\bar{x}$ becomes a more accurate estimate of μ

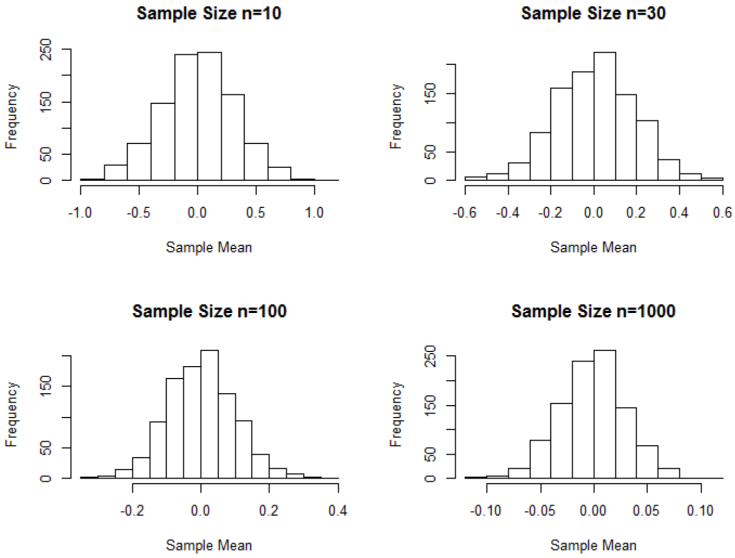
CLT - An Illustration

Population Distribution - Normal(0,1)



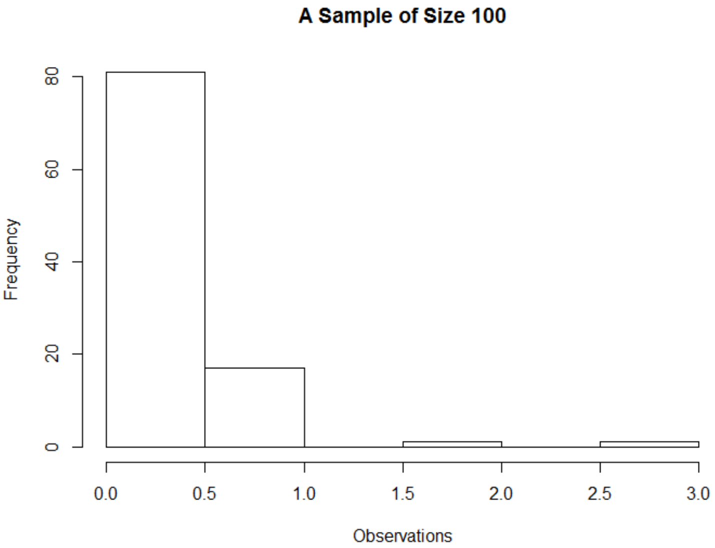
CLT - An Illustration

Sample means of 1000 samples of each size



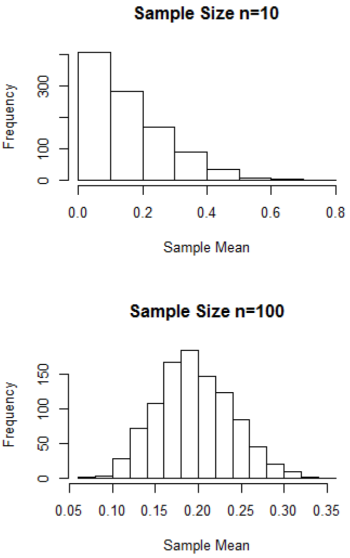
CLT - An Illustration

Population Distribution - Poisson(0.2)



CLT - An Illustration

Sample means of 1000 samples of each size



CLT and Sampling Distributions

The CLT is a result about sums of i.i.d. random variables.

We use it determine the sampling distributions of statistics based on sums of i.i.d. random variables:

- Sums / totals
- Scaled sums
- Means
- etc.

Sampling Distribution of the Sample Mean

How to determine the sampling distribution of $\bar{x}$:

1. In any case $\bar{x}$ has mean μ and s.d. $\frac{\sigma}{\sqrt{n}}$.
2. Is the population Normal(μ, σ)?
   - YES! $\rightarrow$ $\bar{x}$ is Normal$(\mu, \frac{\sigma}{\sqrt{n}})$
   - No... Keep going
3. Is the sample large $(n \geq 30)$?
   - YES! $\rightarrow$ CLT says $\bar{x}$ is approx. Normal$(\mu, \frac{\sigma}{\sqrt{n}})$
   - No... The normal approximation may not be reliable :(

Working with $\bar{x}$

1. Focus on a population with mean μ and s.d. σ
2. Take a sample of size n: $x_1, x_2, \cdots, x_n$
3. Obtain the sample mean $\bar{x} = \frac{1}{n}(x_1 + x_2 + \cdots + x_n)$
4. Find the distribution of $\bar{x}$:
   - $\bar{x}$ has mean μ and s.d. $\frac{\sigma}{\sqrt{n}}$.
   - Population is Normal$(\mu, \sigma) \rightarrow \bar{x}$ is Normal$(\mu, \frac{\sigma}{\sqrt{n}})$
   - Sample is large $(n \geq 30) \rightarrow \bar{x}$ is approx. Normal$(\mu, \frac{\sigma}{\sqrt{n}})$
5. Make inferences about μ (will cover later)

Example 2

A candy factory uses a machine that packages candy into bags with mean weight 7oz and a standard deviation of 0.2oz.

Take a random sample of 100 bags of candy from the factory.

Let $\bar{x}$ denote the mean weight of the sample.

(a) What is the probability that $\bar{x}$ exceeds 7.5oz?

(b) It turns out $\bar{x}$ is 7.5oz. What might you conclude?

Example 2

A candy factory uses a machine that packages candy into bags with mean weight 7oz and a standard deviation of 0.2oz. Take a random sample of 100 bags of candy from the factory. Let $\bar{x}$ denote the mean weight of the sample.

(a) What is the probability that $\bar{x}$ exceeds 7.2oz?

$n \geq 30$
by the CLT $\bar{x}$ is approximately normal
with mean 7 and s.d $\frac{0.2}{\sqrt{100}} = 0.02$

$P(\bar{x} > 7.2) = P(Z > \frac{7.2-7}{0.02}) = P(Z > 10) \approx 0$

Example 2

A candy factory uses a machine that packages candy into bags with mean weight 7oz and a standard deviation of 0.2oz. Take a random sample of 100 bags of candy from the factory. Let $\bar{x}$ denote the mean weight of the sample.

(b) It turns out $\bar{x}$ is 7.5oz. What might you conclude?

From (a) we know it's very unlikely that $\bar{x}$ exceeds 7.2oz.

It's even more unlikely that $\bar{x}$ exceeds 7.5oz.

If that's the case, I might conclude that the information about the machine is incorrect.

Sampling Distribution of $\sum x_i$

The sampling distribution of the sample sum $\sum_{i=1}^{n} x_i$:

1. In any case $\sum_{i=1}^{n} x_i$ has mean $n\mu$ and s.d. $\sigma\sqrt{n}$.
2. Is the population Normal(μ, σ)?
   - YES! $\rightarrow$ $\sum_{i=1}^{n} x_i$ is Normal$(n\mu,\ \sigma\sqrt{n})$
   - No... Keep going
3. Is the sample large ($n \geq 30$)?
   - YES! $\rightarrow$ CLT says $\sum_{i=1}^{n} x_i$ is approx. Normal$(n\mu,\ \sigma\sqrt{n})$
   - No... The normal approximation may not be reliable :(

Example 3

A candy factory uses a machine that packages candy into bags with mean weight 7oz and a standard deviation of 0.2oz. Take a random sample of 100 bags of candy from the factory.

Let T denote the total weight of the sample ($T = \sum_{i=1}^{n} x_i$).

What is the probability that the sample weights between than 720 and 740oz?

Example 3

A candy factory uses a machine that packages candy into bags with mean weight 7oz and a standard deviation of 0.2oz. Take a random sample of 100 bags of candy from the factory. What is the probability that the sample weights between than 700 and 705oz?

Since $n \geq 30$, the CLT applies, and T is approximately normal with mean 700 and s.d $0.2\sqrt{100} = 2$

$$
\begin{aligned}
P(700 < T < 705) &= P\left(\frac{700-700}{2} < Z < \frac{705-700}{2}\right) \\
&= P(0 < Z < 2.5) \\
&= P(Z < 2.5) - P(Z < 0) \\
&= 0.99379 - 0.5 = 0.49379
\end{aligned}
$$