# Statistical Methods

## Lecture 7 – Distributions of Sampling Statistics

Luke Dickens

Autumn 2020

**Introduction**

Sample Mean

Central Limit Theorem

Sampling proportions from a finite population

Distribution of sample variance of a normal population

# Conclusions and Assumptions

**The idea:** Suppose that a set of observed data (a sample) comes from a population, and each member of that population has a numerical value associated with it. The values in the sample can help us to draw conclusions about the values of the population.

**Assumptions:** Before we can draw these conclusions, we must first make assumptions about the population and its relationship with the sample. Two common assumptions to make are:

- There is an underlying probability distribution for the population values.
- Sample data represent independent values from this distribution.

# Samples

**Definition:** If $X_1, \ldots, X_n$ are independent RVs with a common probability distribution, we say they constitute a sample from that distribution.

In most cases, the probability distribution will not be (completely) known, and so the sample can be used to predict characteristics about the distribution.

Predicting characteristics about an unknown (or partly known) distribution is called **inference**.

Introduction

**Sample Mean**

Central Limit Theorem

Sampling proportions from a finite population

Distribution of sample variance of a normal population

# Properties of the sample mean

Assume a population of elements each with a value attached, and that these are the values of an RV with population mean $\mu$ and population variance $\sigma^2$.

We draw sample $X_1, \ldots, X_n$ which has sample mean:

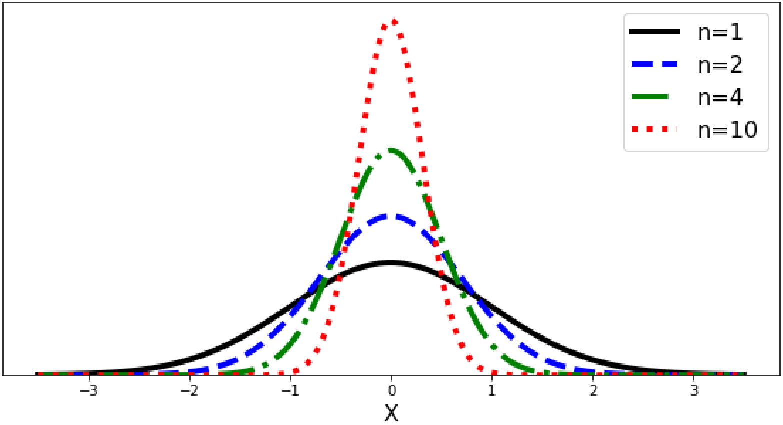
$$\overline{X} = \frac{X_1 + \ldots + X_n}{n}$$

As an arithmetic expression of a set of RVs, $\overline{X}$ is itself a RV with:

$$\mathbb{E}[\overline{X}] = \mu \quad , \quad \mathrm{var}(\overline{X}) = \frac{\sigma^2}{n} \quad \text{and} \quad \mathrm{SD}(\overline{X}) = \frac{\sigma}{\sqrt{n}}$$

# Sample mean illustration

Sample mean has the same expected value as a single value from population, and variance is reduced as sample size increases.



*Distribution of sample mean for standard normal with sample size n.*

Introduction

Sample Mean

Central Limit Theorem

Sampling proportions from a finite population

Distribution of sample variance of a normal population

# The Central Limit Theorem

**Theorem:** Let $X_1, X_2, \ldots, X_n$ be a sample from a (any) population with mean $\mu$ and SD $\sigma$. If $n$ is large then RV

$$X_1 + X_2 + \ldots + X_n$$

- will be approximately normally distributed
- have mean $n\mu$ and SD $\sqrt{n}\sigma$

**Corollary:** As a direct consequence, the mean of this sample, i.e.

$$\frac{X_1 + X_2 + \ldots + X_n}{n}$$

will also be approximately normally distributed.

# More on the Central Limit Theorem

The result holds even when RVs $X_i$ do not come from the same distribution (so long as their variances are of the same scale):

- helps to explain why so many naturally occuring populations of values exhibit a normal distribution
- experimental measurements can be affected by lots of small errors – repeated measures of the same thing are therefore normally distributed.
- partially explains why biological characteristics are normally distributed – lots of small environmental influences
- big influencers, e.g. gender for height, do not qualify

# Reasoning using the standard normal

Again, let $X_1, X_2, \ldots, X_n$ be a sample from a population with mean $\mu$ and SD $\sigma$. If we define sample mean

$$\overline{X} = \frac{X_1 + X_2 + \ldots + X_n}{n}$$

then if $n$ is sufficiently large, we can reason about the distribution of $\overline{X}$ by converting to the standard normal with:

$$\begin{aligned} \Pr\left(\overline{X} \leq a\right) &= \Pr\left(\frac{\overline{X} - \mu}{\sigma/\sqrt{n}} \leq \frac{a - \mu}{\sigma/\sqrt{n}}\right) \\ &\approx \Pr\left(Z \leq \frac{a - \mu}{\sigma/\sqrt{n}}\right) \end{aligned}$$

Since $\overline{X}$ has expectation $\mu$, SD $\sigma/\sqrt{n}$ and is approximately normally distributed.

# Example

A museum has a donation point where every family that enters either donates \$0, \$5, \$10 and \$20 with respective probabilities 0.8, 0.16, 0.03 and 0.01. Assuming $1000 = 10^3$ families visit in a day:

a) What is the expected total donated that day?

b) What is the standard deviation of this total?

c) With what probability will this exceed \$1500?

# Example

A museum has a donation point where every family that enters either donates \$0, \$5, \$10 and \$20 with respective probabilities 0.8, 0.16, 0.03 and 0.01. Assuming $1000 = 10^3$ families visit in a day:

a) What is the expected total donated that day?

$$E[X_i] = \mu = 0.8 \cdot 0 + 0.16 \cdot 5 + 0.03 \cdot 10 + 0.01 \cdot 20 = \$1.30$$

$$E[\sum_i X_i] = 1000\mu = \$1300$$

b) What is the standard deviation of this total?

$$var(X_i) = \sigma^2 = E[X_i^2] - \mu^2 = 9.31$$

$$\mathrm{SD}(\sum_i X_i) = \sqrt{n}\sigma = \$96.5$$

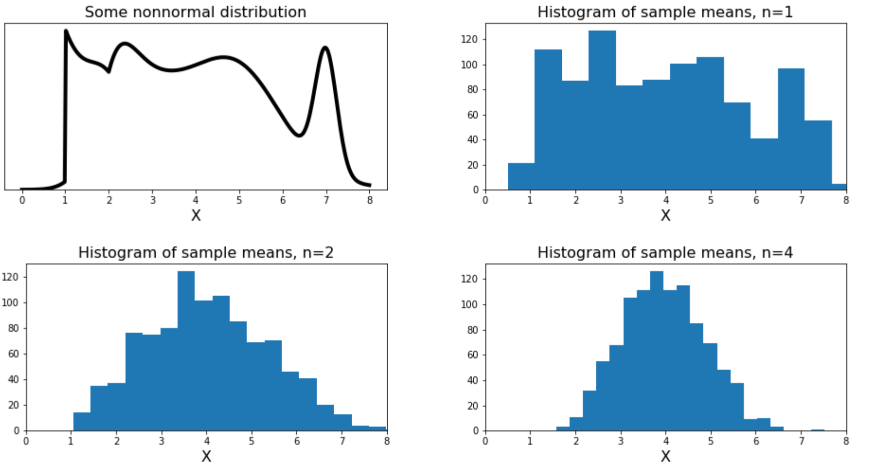
c) With what probability will this exceed \$1500?

$$\Pr\left(\sum_i X_i > \$1500\right) = \Pr\left(Z > \frac{1500 - 1300}{96.5}\right) = 0.192$$

# How big a sample is needed. . .

. . . before you can consider the mean to be normally distributed?

[Ros17, p. 309], says $n = 30$: “no matter how nonnormal the underlying population is”, and $n$ can be much smaller.



Introduction

Sample Mean

Central Limit Theorem

Sampling proportions from a finite population

Distribution of sample variance of a normal population

# Samples from finite populations

**Definition:** A sample of size $n$ selected from a population of $N$ elements is **a random sample** if it is selected so that it is equally likely to be any subset of size $n$.

Suppose such a random sample is taken, a characteristic of interest is given, and for $i = 1, \ldots, n$

$$X_i = \begin{cases} 1 & \text{if the } i\text{th member of the sample has the characteristic} \\ 0 & \text{otherwise} \end{cases}$$

Consider the sum

$$X = X_1 + X_2 + \ldots X_n$$

This counts the number in the sample with the characteristic.

# Approximate independence

As we have a finite population, $N$ the number of elements with our characteristic, say $K$, means that a proportion $p = \frac{K}{N}$ have that characteristic.

For our sample, $\Pr(X_i = 1) = \frac{K}{N} = p$ for any $i$.

However, if we know one binary value then this changes the condition probability of another, e.g.

$$\Pr(X_2 = 1 | X_1 = 1) = \frac{K-1}{N-1} \neq p$$

For large enough $N$ (and small enough $n$) we can ignore these differences, and simply say, e.g.

$$\Pr(X_j | X_{i_1} = 1, X_{i_2} = 1, X_{i_3} = 0, \ldots) \approx p = \Pr(X_j)$$

# Binomial Approximation

For large enough $N$, the sum

$$X = X_1 + X_2 + \ldots X_n$$

is a RV approximately distributed as Binomial$(n, p)$ with

$$\mathbb{E}[X] = np \qquad \text{and} \qquad \text{SD}[X] = \sqrt{np(1-p)}$$

The mean $\overline{X} = X/\,n$ is the proportion of those in the sample with the characteristic and

$$\mathbb{E}[\overline{X}] = p \qquad \text{and} \qquad \text{SD}[\overline{X}] = \sqrt{p(1-p)/\,n}$$

# Sample proportion as a sample mean

If proportion, $p$, of a population has a given characteristic then:

- Let $\overline{X}$ denote the proportion of a sample of size $n$ with the characteristic
- Make use of the fact that $X = n\overline{X}$ is approximately binomial with parameters $n$ and $p$ (successes in $n$ independent trials)
- Binomial RV, $X$ can be thought of as $X_1 + X_2 + \ldots X_n$

$$X_i = \begin{cases} 1 & \text{if trial } i \text{ is a success} \\ 0 & \text{if trial } i \text{ is a failure} \end{cases}$$

Thus $\overline{X} = X/n$ can be regarded as the sample mean of a sample of size $n$ from population having mean $p$ and SD $\sqrt{p(1-p)}$.

# Probabilities associated with sample proportions

From the Central Limit theorem, for large enough $n$, $X$ and thus $\overline{X}$ are approximately normal and

$$\frac{X/n - p}{\sqrt{p(1-p)/n}} = \frac{X - np}{\sqrt{np(1-p)}}$$

has an approximately standard normal distribution, and so we can compute (to high accuracy) the probability that $\overline{X}$ will lie in a given range using [Ros17, Tbl. 6.1] (also [Ros17, Tbl. D.1])

From[Ros17, p. 316]: "the normal approximation to the binomial is quite good provided $n$ is large enough that the quantities $np$ and $n(1-p)$ are both greater than 5."

Introduction

Sample Mean

Central Limit Theorem

Sampling proportions from a finite population

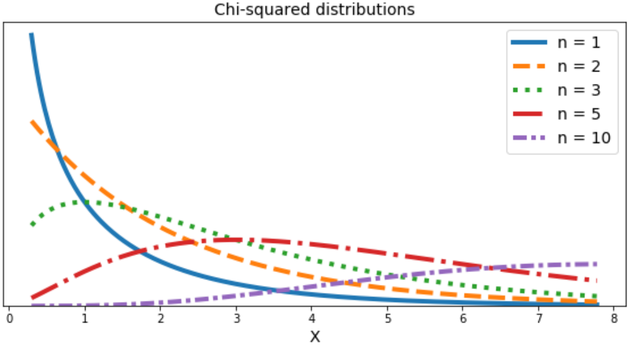
**Distribution of sample variance of a normal population**

# Chi-squared random variables

**Definition:** If $Z_1, \ldots, Z_n$ are independent samples from a standard normal distribution then

$$\sum_{i=1}^{n} Z_i^2$$

is a chi-squared RV with $n$ degrees of freedom.



# Distribution of normal sample variance

Suppose $X_1, \ldots, X_n$ is a sample from a normal population with mean $\mu$ and SD $\sigma$, then the sample variance:

$$S^2 = \frac{\sum_{i=1}^{n}(X_i - \overline{X})^2}{n-1}$$

is a RV. Moreover

$$\frac{(n-1)S^2}{\sigma^2} = \frac{\sum_{i=1}^{n}(X_i - \overline{X})^2}{\sigma^2}$$

is a chi-squared RV with $n-1$ degrees of freedom.

If you are interested in the proof see **here**.

# References I

[Ros17] Sheldon M. Ross, *Introductory Statistics*, 4 ed., Academic Press, 2017.