# Lectures 13 and 14. Central Limit Theorem (CLT)

**YULIA R. GEL**

**CS/SE/STAT 3341 Probability and Statistics in Computer Science and Software Engineering**

February 28 and March 2, 2017

1. From Empirical Histograms to Probability Histograms: Path to CLT

2. Central Limit Theorem (CLT)

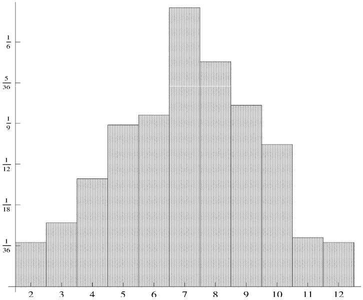
3. Applications

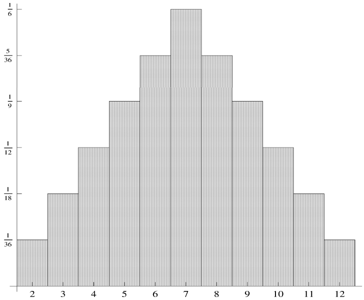
# Probability histograms

- The histograms that we usually deal with are empirical histograms, i.e.
  - based on data
  - area under the histogram represents the percentage of cases
- Now we look at a new type of histograms, i.e. **probability histograms**:
  - they are **not based on data but on theory**
  - area under the histogram represents chance

Empirical histogram converges to probability histogram when the number of trials $n \to \infty$, i.e. the empirical histogram looks more and more like the probability histogram.

Example 1. Let us roll two dice 300 times. Then we get a sample histogram (the left plot), the corresponding (exact) probability histogram is at the right plot. Notice that the probability histogram is a **limiting** case of a sample histogram when number of rolls → $\infty$.





# Central Limit Theorem (CLT)

The Central Limit Theorem (CLT) is one of the corner-stones of probability theory and statistics. CLT explains why many distributions tend to be close to the normal distribution and, hence, why we can use bell curve for approximation of many real world events.

CLT was originally proposed by Bernoulli, de Moivre, Laplace and Turing (!) under the assumption that all trials are identical and independent. In 19th and 20th centuries, CLT was extended to a way more general framework.

# Central Limit Theorem (CLT) – contd

Most people have a good intuitive understanding of the Law of Averages, but in many cases it is important to determine whether a particular size of deviation between the sample mean and the (usually unknown) expected value is probable or improbable.

In other words, what is the chance that the sample average is more than some value $d$ away from the **true** expected value $EX$?

Essentially, CLT allows one to describe how accurately the Law of Averages works.

# Central Limit Theorem

**Central Limit Theorem**. Formally, let $X_1, X_2, \ldots$ be independent random variables with the same expectation $\mu = E(X_i)$ and the same standard deviation $\sigma = Std(X_i)$, and let

$$S_n = X_1 + X_2 + \ldots + X_n.$$

Then,

$$P\left\{\frac{S_n - n\mu}{\sigma\sqrt{n}}\right\} \to \Phi(z),$$

where $\Phi(z)$ is a cfd for standard normal distribution.

# Central Limit Theorem for Averages

Alternatively we can say that if $\bar{X} = \frac{1}{n}\sum_{i=1}^{n} X_i$, then CLT implies that

$$\bar{X} \sim N(\mu, \frac{\sigma^2}{n}), \qquad \Leftrightarrow \qquad \frac{\bar{X} - \mu}{\sigma/\sqrt{n}} \sim N(0, 1).$$

Hence, a sample average is a random number that is approximately normal with mean zero and standard deviation one.

The approximation gets better as $n$ gets larger. A key point is that to use the Central Limit Theorem, we need to know mean $EX = \mu$ and standard deviation $Std(X) = \sigma$ for the observed data $X_1, X_2, \ldots$.

Modifications of this formula hold for many other situations, e.g., when there is some weak dependence in observed data etc.

# CLT for Sums

In view of the **Central Limit Theorem for sums**, we can also say that:

$$S_n \sim N(n\mu, n\sigma^2), \qquad \Leftrightarrow \qquad \frac{S_n - n\mu}{\sqrt{n}\sigma} \sim N(0,1).$$

This formula is useful when calculating the chance of winning a given amount of money when gambling, or getting more than a specific score on a test.

With these two CLT formulas, we can answer all sorts of practical questions.

# Example: Average Income

Example 1. Suppose we need to estimate the average income of people in Richardson, TX. You draw a random sample of 100 households and find the mean income to be $42,000 with standard deviation of $10,000.

What is the probability that your sample estimate is higher than the true value by $500 or more?

# Example: Average Income

Let us assume that the mean and standard deviation of your sample equals to the mean and standard deviation of the whole population in Richardson, i.e. that you know mean $\mu$ and variance $\sigma$ exactly.

Clearly, this is an approximation, and later we shall discuss how to improve this procedure.

# Example: Average Income

Solution.

$$
\begin{aligned}
P(\bar{X} - EX > 500) &= P\big(\frac{\bar{X} - \mu}{std.dev.(X)} > \frac{500}{std.dev.(X)}\big) \\
&= P\big(\frac{\bar{X} - \mu}{\sigma/\sqrt{n}} > \frac{500}{\sigma/\sqrt{n}}\big) \\
&= P\big(z > \frac{500}{10000/\sqrt{100}}\big) = P(z > 0.5).
\end{aligned}
$$

From the standard normal table, we know this has chance 30.85%.

So the probability of the sample average income to be overestimated by \$500 or more is 0.3085.

How can we improve our estimate?

# Example: Roulette

Example 2. You are playing Red and Black in roulette. A roulette wheel has 38 pockets: 18 are red, 18 are black, and 2 are green - the house takes all the money on green.

You pick either red or black; if the ball lands in the color you pick, you win a dollar. Otherwise, you lose a dollar.

## Example: Roulette

- What is the expected dollar amount you get after each trial?
- What is the standard deviation?
- Suppose you make 100 plays. What is the chance that you lose $10 or more?

# Example: Roulette

Solution. There are 38 tickets: 18 are labeled 1 and the 20 are labeled -1. Let $X$ be a random variable describing your gain/loss in a single trial.

So the expected value is

$$\begin{aligned} EV &= \frac{1}{38}\Big\{1+1+1+1+1+1+1+1+1+1+1+1+1+1+1+1+1+1 \\ &+ (-1)+(-1)+(-1)+(-1)+(-1)+(-1)+(-1)+(-1)+(-1)+(-1)+(-1) \\ &+ (-1)+(-1)+(-1)+(-1)+(-1)+(-1)+(-1)+(-1)+(-1)\Big\} = \frac{-2}{38} = -\frac{1}{19}. \end{aligned}$$

The standard deviation for gain/loss on each trial is

$$\begin{aligned} \sigma &= \sqrt{\frac{1}{38}\sum_{i=1}^{38} X_i^2 - (EX)^2} \\ &= \sqrt{\left(1 - \left[-\frac{1}{19}\right]^2\right)} = 0.998614\$. \end{aligned}$$

## Example: Roulette

Hence, given that all trial are independent and taking into account that expected value on a single trial is $-1/19$ and standard deviation is 0.998614, the probability of losing more than \$10 or more in 100 plays is given by

$$\begin{aligned}
P(\mathrm{sum} < -10) &= P(\mathrm{sum} - \mathrm{nEX} < -10 - \mathrm{nEX}) \\
&= P\Big(\frac{\mathrm{sum} - \mathrm{nEX}}{\sqrt{n}\sigma} < \frac{-10 - nEV}{\sqrt{n}\sigma}\Big) \\
&= P\Big(z < \frac{-10 - nEX}{\sqrt{n}\sigma}\Big) \\
&= P\Big(z < \frac{-10 - 100 \times \frac{-1}{19}}{10 \times 0.998614}\Big) \\
&= P(z < -0.4734).
\end{aligned}$$

## Example: Roulette

From the standard normal, the chance of $z$ falling below -0.4734 is 31.76%.

It's pretty high!