Central Limit Theorem

The Most Important Result in Statistics

Central limit theorem

One of the most frequently used statistics in engineering applications is the sample mean. We can relate the mean and variance of the probability distribution of the sample mean to those of a single observation when an iid model is appropriate.

In the case of the sample mean, if the sample size (n) is large enough, we can also approximate the *shape* of the *probability distribution function* of the sample mean!

Central limit theorem

If $X_1, \ldots, X_n$ are **independent** and **identically** distributed (iid) random variable (with mean μ and variance σ^2), then for large n, the variable $\overline{X}$ is approximately normally distributed. That is,

$$\overline{X} \dot{\sim} Normal\left(\mu, \frac{\sigma^2}{n}\right)$$

* This is one of the **most important** results in statistics.

Example: [Tool serial numbers]

Consider selecting the last digit of randomly selected serial numbers of pneumatic tools. Let

- $W_1 =$ the last digit of the serial number observed next Monday at 9am
- $W_2 =$ the last digit of the serial number observed the following Monday at 9am

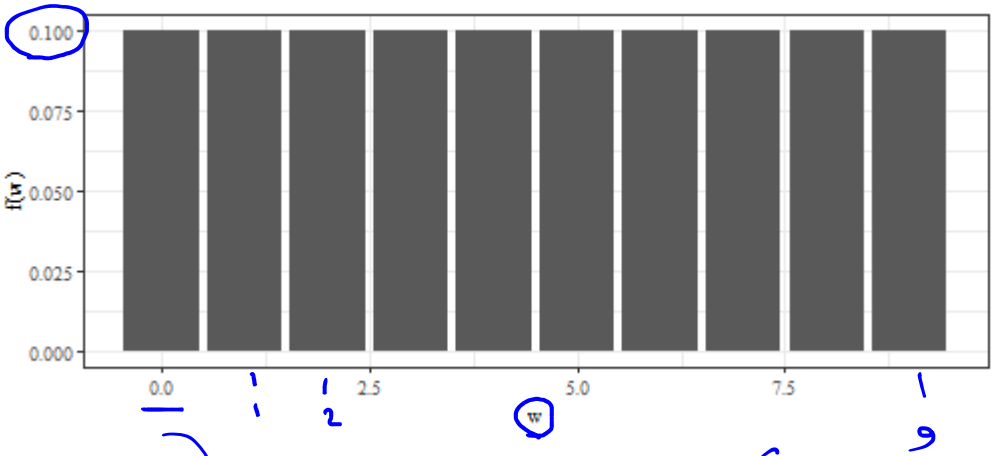
A plausible model for the pair of random variables W_1, W_2 is that they are independent, each with the marginal probability function

$$f(w) = \begin{cases} .1 & w = 0, 1, 2, \ldots, 9 \\ 0 & \text{otherwise} \end{cases}$$

W_i discrete r.v.

$i = 1, 2$

Example: [Tool serial numbers]

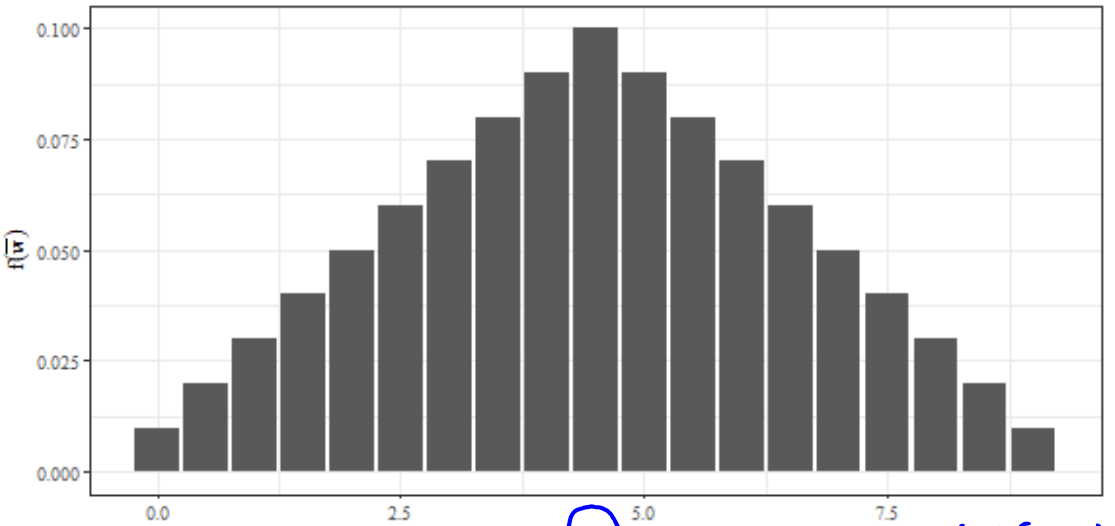


With $\mathrm{E}W = 4.5$ and $\mathrm{Var}W = 8.25$.

Using such a distribution, it is possible to see that $\overline{W} = \frac{1}{2}(W_1 + W_2)$ has probability distribution

$\overline{w}$	$f(\overline{w})$	$\overline{w}$	$f(\overline{w})$	$\overline{w}$	$f(\overline{w})$	$\overline{w}$	$f(\overline{w})$	$\overline{w}$	$f(\overline{w})$
0.00	0.01	2.00	0.05	4.00	0.09	6.00	0.07	8	0.03
0.50	0.02	2.50	0.06	4.50	0.10	6.50	0.06	8.5	0.02
1.00	0.03	3.00	0.07	5.00	0.09	7.00	0.05	9	0.01
1.50	0.04	3.50	0.08	5.50	0.08	7.50	0.04		

Example: [Tool serial numbers]

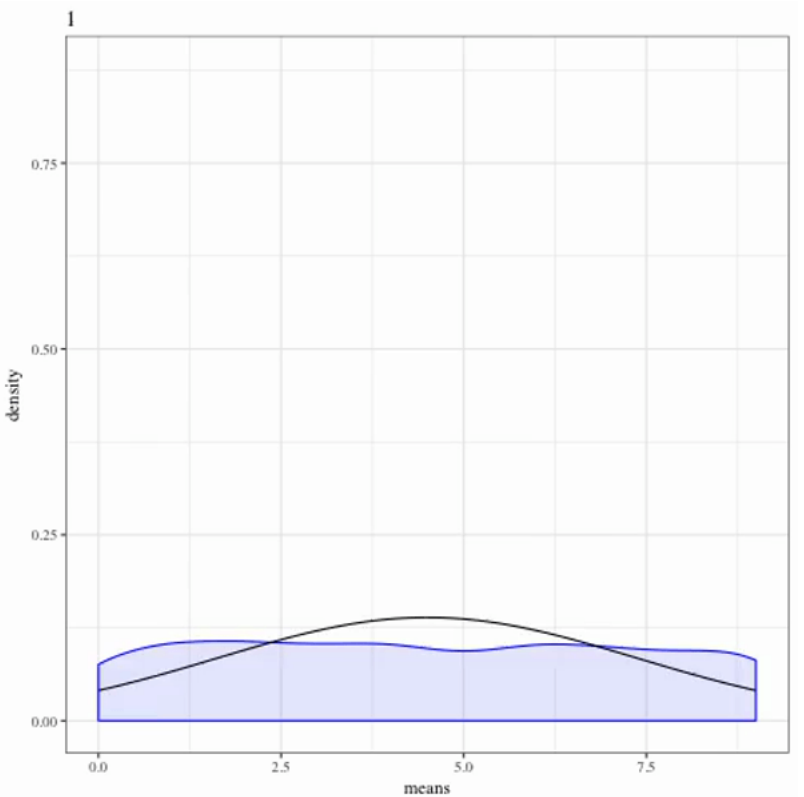


$E\bar{W} = EW_1 = 4.5$ $\quad Var(\bar{W}) = \frac{Var(W_1)}{2} = 4.125$

Comparing the two distributions, it is clear that even for a completely flat/uniform distribution of W and a small sample size of $n = 2$, the probability distribution of $\bar{W}$ looks more bell-shaped than the underlying distribution.

Now consider larger and larger sample sizes, $n = 1, \ldots, 40$:

Watch how CLT works here



$\bar{W}$ will always

ha $E\bar{W} = EW_1 = 4.5$,

but $\text{var}\,\bar{W} = \frac{\text{var}(W_1)}{n}$

decreases as $n \to \infty$

and $\bar{W}$ will approximately have Normal shape (distribution)

Example: [Stamp sale time]

Imagine you are a stamp salesperson (on eBay). Consider the time required to complete a stamp sale as S, and let

$\overline{S}$ = the sample mean time required to complete the next 100 sales

Each individual sale time should have an $Exp(\alpha = 16.5s)$ distribution. We want to consider approximating $P[\overline{S} > 17]$.

assume iid

$n \geq 25$

$S_i \overset{iid}{\sim} Exp(\alpha = 16.5) \longrightarrow ES_i = \alpha = 16.5$

$Var S_i = \alpha^2 = 16.5^2 = 272.25$

one of them

Now: $E\bar{S} = ES_1 = 16.5$

$Var(\bar{S}) = \frac{Var(S_1)}{n} = \frac{272.25}{100} = 2.72225$

Since $n = 100 \geq 25$,

using CLT: $\bar{S} \sim N(\mu = 16.5, \sigma^2 = 2.72225)$ (1.65^2)

$$\rightarrow P(\bar{S} > 17) = P\left(\frac{\bar{S} - 16.5}{\sqrt{2.7225}} > \frac{17 - 16.5}{\sqrt{2.7225}}\right)$$

$$= P\left(Z > \frac{17 - 16.5}{1.65}\right) \quad \frac{0.5}{1.65} \approx 0.303$$

$$= 1 - P(Z \leq 0.303)$$

$$= 1 - \Phi(0.303)$$

table $\approx 1 - 0.6217 = 0.3783$

Example: [Cars]

Suppose a bunch of cars pass through certain stretch of road. Whenever a car comes, you look at your watch and record the time. Let X_i be the time (in minutes) between when the i^{th} car comes and the $(i+1)^{th}$ car comes for $i = 1, \ldots, 44$. Suppose you know the average time between cars is 1 minute.

Find the probability that the average time gap between cars for the next 44 cars exceeds 1.05 minutes.

X_i : time (minute) between i^{th} car & $(i+1)^{th}$ car

$X_i \overset{iid}{\sim} Exp(\alpha = 1)$

$\bar{X} = \frac{1}{44}\sum_{i=1}^{44} X_i$: the average time gap between cars for 44 cars.

$P(\bar{X} > 1.05) = ?$

we have iid sample + $n = 44 \geq 25$

use CLT.

$$\bar{X} \overset{\cdot}{\sim} N\left(\underbrace{EX_1 = \alpha = 1}_{\text{Exp}(1)}, \frac{\text{Var}(X_1)}{n} = \frac{\alpha^2}{n} = \frac{1}{44}\right)$$

$$P(\bar{X} > 1.05) = P\left(\frac{\bar{X} - 1}{\sqrt{1/44}} > \frac{1.05 - 1}{\sqrt{1/44}}\right)$$

$$= P(Z > 0.332) \quad , Z \sim N(0,1)$$

$$= 1 - P(Z \leq 0.332)$$

$$= 1 - \Phi(0.332)$$

$$= 1 - 0.6331 = 0.3669.$$

with 36.69% probability, the average time gap between 44

cars is > 1.05 minute.

Example: [Baby food jars, cont'd]

The process of filling food containers appears to have an inherent standard deviation of measured fill weights on the order of $1.6g$. Suppose we want to calibrate the filling machine by setting an adjustment knob and filling a run of n jars. Their sample mean net contents will serve as an indication of the process mean fill level corresponding to that knob setting.

$\sigma = 1.6$

we're looking for n.

You want to choose a sample size, n, large enough that there is an 80\% chance the sample mean is within .3g of the actual process mean.

$\bar{x}$

μ

(assume iid)

* $P(\mu - 0.3 < \bar{x} < \mu + 0.3) = 0.8$

Note that, μ & σ are not given!

we're looking for n

X_i : the weight of one jar.

$\bar{X}$: The sample mean weight of n jars.

$$E\bar{X} = EX_1 = EX_i = \mu$$

$$Var(\bar{X}) = \frac{Var(X_1)}{n} = \frac{1.6^2}{n}$$

For n large enough, by CLT : $\bar{X} \sim N(\mu, \frac{\sigma^2}{n})$

$$\Rightarrow \bar{X} \sim N(\mu, \frac{1.6^2}{n})$$

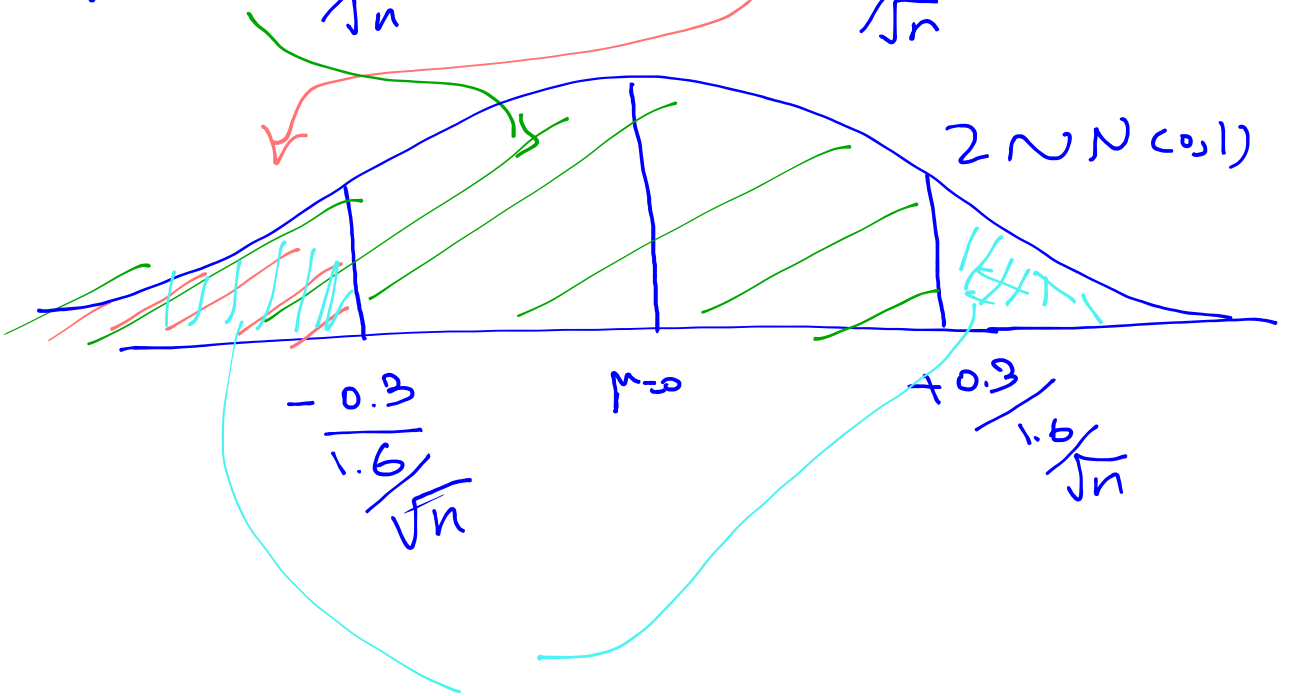
$$P(\mu - 0.3 < \bar{X} < \mu + 0.3) = 0.8$$

$$= P\left(\frac{\mu - 0.3 - \mu}{1.6/\sqrt{n}} < \frac{\bar{X} - \mu}{1.6/\sqrt{n}} < \frac{\mu + 0.3 - \mu}{1.6/\sqrt{n}} \right) = 0.8$$

$$Z \sim N(0,1)$$

$$= P\left(\frac{-0.3}{1.6/\sqrt{n}} < Z < \frac{+0.3}{1.6/\sqrt{n}} \right) = 0.8$$

$$= \Phi\left(\frac{0.3}{1.6/\sqrt{n}} \right) - \Phi\left(\frac{-0.3}{1.6/\sqrt{n}} \right) = 0.8$$



$$= \Phi\left(\frac{0.3}{1.6/\sqrt{n}}\right) - \left(1 - \Phi\left(\frac{0.3}{1.6/\sqrt{n}}\right)\right) = 0.8$$

$$= 2\,\Phi\left(\frac{0.3}{1.6/\sqrt{n}}\right) - 1 = 0.8$$

$$\Rightarrow \Phi\left(\frac{0.3}{1.6/\sqrt{n}}\right) = \frac{1.8}{2} = 0.9$$

by the table $\Rightarrow \dfrac{0.3}{1.6/\sqrt{n}} = 1.29 \Rightarrow \sqrt{n} = \dfrac{1.29 \times 1.6}{0.3}$

$$\Rightarrow n = \left(\frac{1.29 \times 1.6}{0.3}\right)^2 = 47.3344$$

Choose : $n = 48$

Example: [Printing mistakes]

Suppose the number of printing mistakes on a page follows some unknown distribution with a mean of 4 and a variance of 9. Assume that number of printing mistakes on a printed page are iid.

$\sigma^2 = 9$

- What is the approximate probability distribution of the average number of printing mistakes on 50 pages?

$\bar{X}$

$n \geq 25$

$\bar{X} \sim N(4, \frac{9}{50})$ by CLT.

- Can you find the probability that the number of printing mistakes on a single page is less than 3.8?

X

No, because the probability dist. of # of printing mistakes is Unkown on a single page.

Example: [Printing mistakes]

- Can you find the probability that the average number of printing mistakes on 10 pages is less than 3.8?

No, because $n = 10 < 25$ & we cannot use CLT: Thus, the dist. of $\bar{x}$ is unknown.

- Can you find the probability that the average number of printing mistakes on 50 pages is less than 3.8?

Yes. because $n = 50 > 25$ & x_i are i.i.d.

So, by CLT $\bar{x} \sim N(\mu = 4, \frac{9}{50})$