Normal Approximation & Binomial Distribution

→ Empirical rule and normal approximation of data

→ Many data have histograms that look bell-shaped. If data has a normal bell, then it is a normal data / normal curve

↓

① has bell curve

② is symmetrical

eg:- height, weight, B.P

The Normal Approximation:-

If data has a normal curve, then

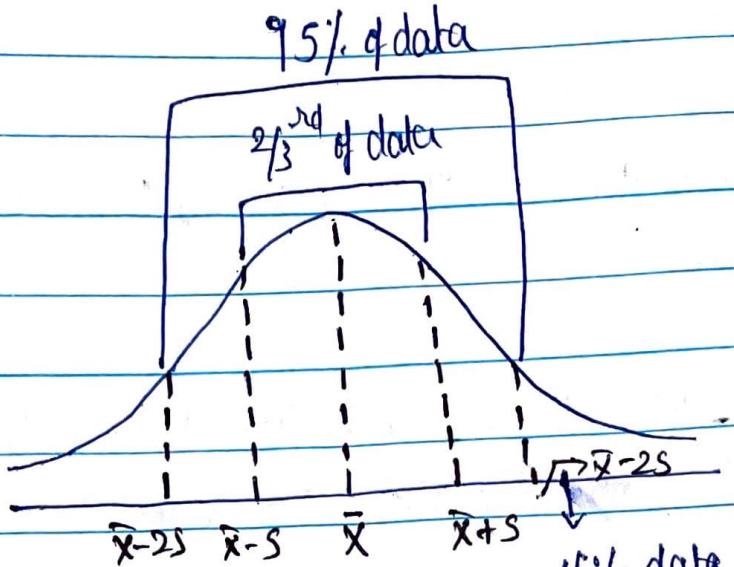
- about $2/3^{rd}$ of data fall within one standard deviation of mean
- about 95% fall within 2 standard deviation of mean.
- about 99.7% fall within 3 standard deviations.

eg:- if $S = 1.8$ & $\bar{x} = 68.3$, then

$\bar{x} - 2(1.8) = 64.7$

$\bar{x} - 2(S) = 71.9$

∴ 95% of data lies b/w 64.7 and 71.9.



16% data is bigger then 1 S.

Standardizing data & standard normal curve :-

Normal curve is completely determined by $\bar{x}$ and S. Once we know them, we know whole histogram & we can compute %. To do that, we need to standardize data.

Standardizing data that we take data by subtracting off $\bar{x}$ & dividing by S.

$$z = \frac{data - \bar{x}}{s}$$

z = z-score = standardized value.

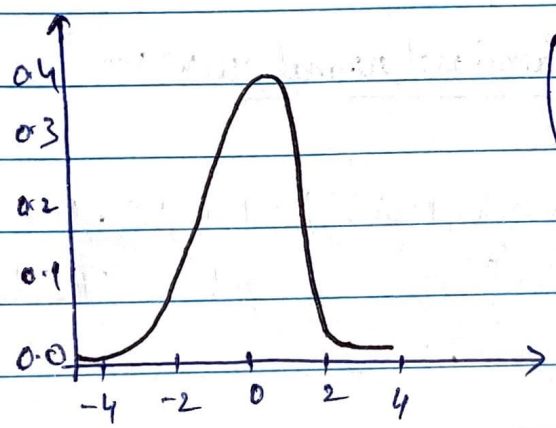
⇒ z-score tells how many standard deviations, the measurement is above or below average.

eg :- if z=2, i.e data is 2 standard deviations above average.
if z=-1.5, i.e data is 1.5 standard deviation below average.

Once we standardize data that have $\bar{x}=0$ & $S=1$, this is point of standardizing.

If $\bar{x}=68.3$ & $S=1.8$, after we standardize, the standardized values follow "standard normal curve" which ~~means~~ has $\bar{x}=0$ & $S=1$.

Standard normal curve graph, $y=\frac{1}{\sqrt{2\pi}}e^{-1/2 \cdot x^2}$



(This is a special case of empirical curve normal approximation)

Normal Approximation :-

Normal approximation means to use the area under the normal curve to figure percentages.

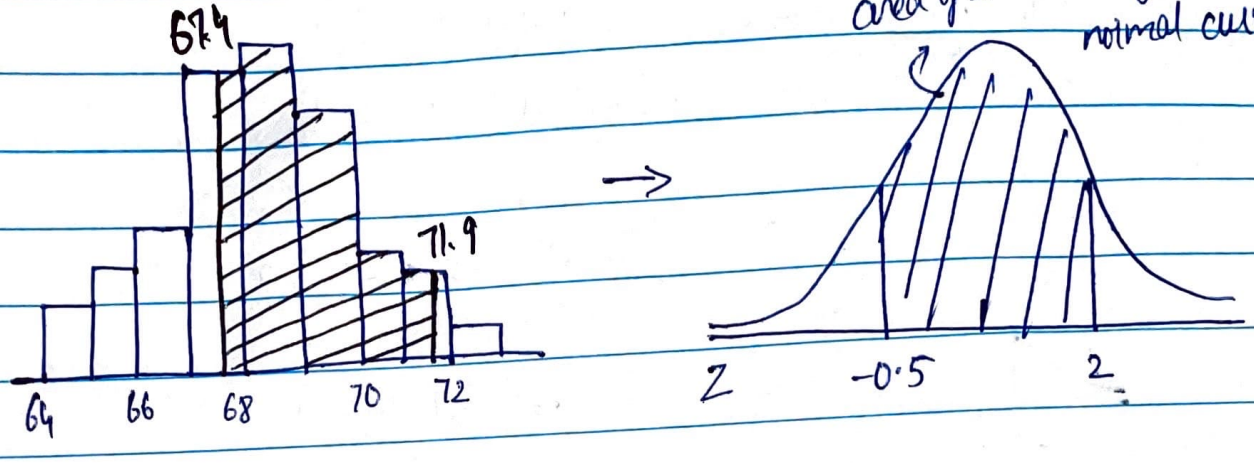
Q What % data is b/w 67.4 & 71.9, when $\bar{x}=68.3$ & $S=1.8$

A ① Standardize data

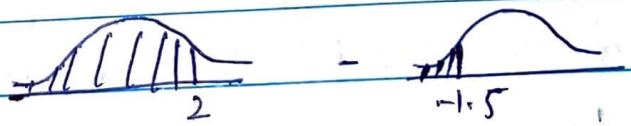
$$Z=\frac{67.4-68.3}{1.8}=-0.5$$

$$Z=\frac{71.9-68.3}{1.8}=2$$

② Mark the area under normal curve



③ Rewrite this graph as = (all of area to left of 2) − (all of area left of −1.5)



④ Find values on table or software

$\Rightarrow 97.7\% - 30.9\% = 66.8\%$

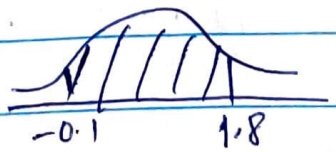
⑤ if $-1 \rightarrow 46\%$ & $1.8 \rightarrow 96.4\%$.

A) $\bar{x} = 68.3$

$s = 1.8$

$$Z = \frac{61.1 - 68.3}{1.8} \cong -0.1$$

$$Z = \frac{71.5 - 61.3}{1.8} \cong 1.8$$

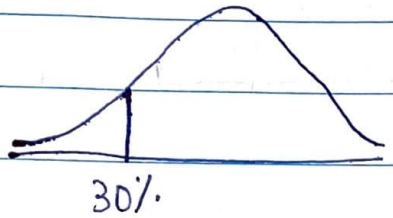


$\Rightarrow 96.4 - 46$

$\Rightarrow 50.4$

Computing percentiles with normal approximation :-

(V) What is the 30th percentile of given data?
↓ where 30% of data falls below



usually, $z = -0.52$

$$z = \frac{\text{data} - \bar{x}}{s}$$

$$\Rightarrow z s + \bar{x} = \text{data}$$

(V)

as z-score tells us how many standard deviations above or below average, so as $z = -0.52$, it is 0.52 standard deviation below average of $\bar{x}$, so

$$\text{data} = \bar{x} + zs$$
$$= 68.3 - (0.52 \times 1.8)$$
$$= 67.4$$

Binomial setting and Co-efficient :-

Revisiting, P(newborn is a girl) = 49%.
What are chances that if we look at 3 new borns, we have 2 girls?

P(2 of 3 girls) = P(GGB or GBG or BGG)
= P(GGB) + P(GBG) or P(BGG)
= P(G) P(G) P(B) + ... + P(B) P(G) P(G)
= (3) × (0.49 × 0.49 × 0.51)

ways we can arrange 2G and 1B

same for all 3 terms. This is binomial setting.

binomial setting:

n = independent repeatitions

each of these have 2 outcomes (success/failure)

P(success) = same in each experiment.

⇒ Now what is P(2 out of 5 are girls)?

If we try to do it as before, enumerating possibilities, the number increases as n does. In this case, it is 10 ways.

To ease it we use

$$\frac{n!}{k!(n-k)!}$$

binomial coefficient

ways to arrange "k" successes out of "n" ~~possible~~ independent repeatitions.

∴ ways to arrange 2 girls out of 5 newborns,

$$\frac{5!}{2!\,3!} = \frac{5\times\cancel{4}^{2}\times\cancel{3}\times\cancel{2}\times\cancel{1}}{2\times\cancel{1}\times\cancel{3}\times\cancel{2}\times\cancel{1}} = \underline{10 \text{ possibilities}}$$

Binomial formula:- (Binomial probability)

Applying this coefficient in binomial setting gives

$$P(k \text{ success in } n \text{ experiments}) = \frac{n!}{k!(n-k)!}\, p^k (1-p)^{n-k}$$

$p^k(1-p)^{n-k}$: probability of having a particular pattern of k success & $(n-k)$ failures.

Q Play an online game 10 times. Each time there are 3 possible outcomes, P(win big) = 10%; P(win small) = 20%; P(win nothing) = 70%. P(2 small prizes)?

A P(win big) = 10%
P(win small) = 20%
P(win nothing) = 70%

Success = win small prize
failure = win big prize or win nothing

$$P(2 \text{ success in } 10 \text{ exp}) = \frac{10!}{2!\,8!} (0.2)^2 (0.1-0.2)^8$$

$$= \left(\frac{10^5 \times 9}{2 \times 1}\right) \times 0.04 \times (0.8)^8$$

$$= 0.3019$$

$$= \underline{30.19\%}$$

Random variables & Probability histograms:-

The outcome of 'n' experiments is due to chance, so the num of success we see 'k' is random.

X = 'num of successes' is a random variable.

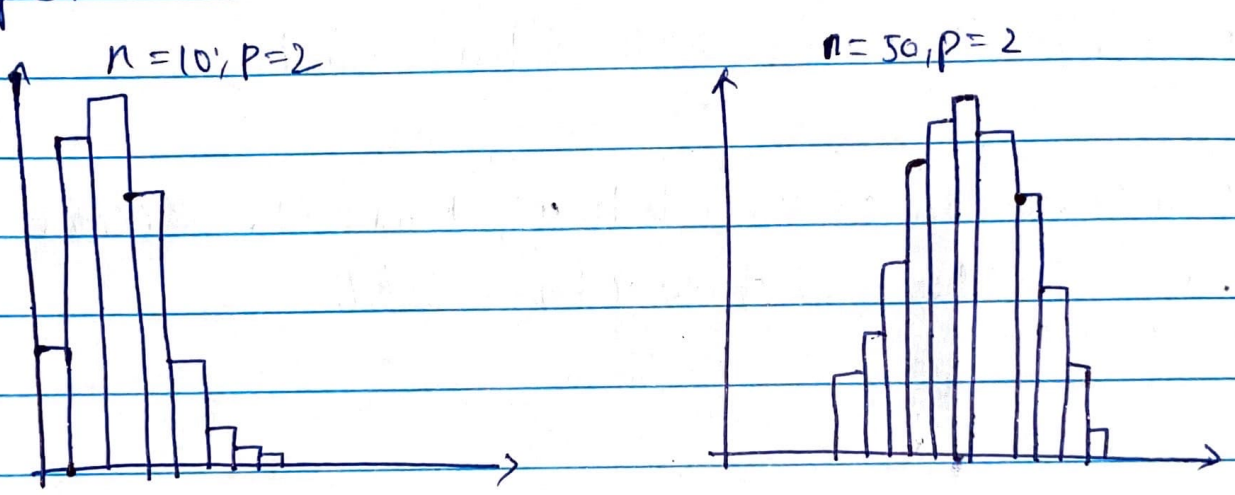
P(X=2) = 30.2% in binomial distribution.

We can visualize the probabilities of various outcomes of X with probability histogram.

Height of bar = P(Corresponding outcome)

Normal approximation to the binomial:-

As number of experiment, n, gets larger, probability histogram of binomial distribution looks more similar to normal curve.



This suggest we can use normal approximation to compute binomial probability, subtract off $\boxed{np}$

$$\boxed{\sqrt{np(1-p)}}$$

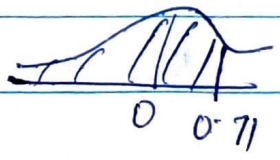
① Same question, n = 50, P(at most 2 small)

p = 0.2

Standardize, $\frac{12 - np}{\sqrt{np(1-p)}} = \frac{12 - (10)}{2.83} = \underline{0.71}$

So, we have to find area under curve to left of 0.71

which is roughly 76%.



⇒ In the setting of survey, poll is a simple random sample, that is sampling without replacement. So this is not binomial setting, because 'p' changes after a subject has been removed. But if population is much larger then it has no effect

⇒ A simple random sample selects subjects without replacement & each subject has equal chance of being selected.

Q Coin is tossed 400 times. Chances of getting more than 210 tails?

A $n = 400$

$k = 210$

$(n-k) = 190$

$p = 0.2$

$$\frac{400!}{210!\,190!}(0.2)^{210}(0.8)^{190}$$

$$Z = \frac{210 - np}{\sqrt{np(1-p)}} = \frac{210 - (400 \times 0.2)}{\sqrt{(400 \times 0.2)(1-0.2)}}$$

$$= \frac{65}{4} = 16.25$$

Q Coin is tossed 6 times. Chances of getting 2 tails in each of first 3 and last 3?

A $n=6$ TTH TTH
$p=0.2$ THT THT
$k=2$ HTT HTT
15
$\frac{6!}{2!4!}$ = 15 possibilities

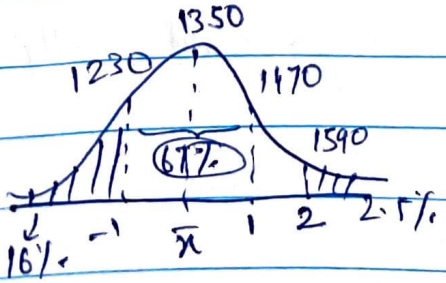
P(2 tails 1-3 and 3 last) = P(2 tails in first 3) and P(2 tails in last 3)

$= \left(\frac{3!}{2!1!}(0.2)^2(0.8)^1\right)\left(\frac{3!}{2!1!}\right)$

Q MCQ with 5 ques, 4 possible answer, 1 correct, chance of getting 2 correct?

A $p(\text{success}) = 1/4 = 0.25$ ways $= \frac{5!}{2!3!}$
$\Rightarrow p(\text{wrong}) = 0.75$
$P(2 \text{ correct}) = \left(\frac{1}{4}\times\frac{1}{4}\right) = \left(\frac{1}{4}\right)^2$
$P(3 \text{ wrong}) = \left(\frac{3}{4}\right)^3$

Q $\bar{x} = 1350$, $S = 120$, what % score below 1230 & what score is needed to be in top 2.5%.



$z = \frac{1230-1350}{120} = -1$

top 2.5% $\Rightarrow$ 2nd std deviation

data $= zS + \bar{x} = 2(120) + 1350$

$= 1590$

Parameter & Statistic :-

→ That in every statistic analysis, there are several histograms floating around. The probability histogram which generates the data. The histogram of observed data and the probability histogram of an estimator.

→ We can estimate average height of population with a relatively small sample.

μ = average of population

σ = standard deviation of population

Parameter → quantity of interest measured ~~in sample~~ about population

Statistic → quantity of interest measured in sample.

Expected value & Standard error :-

If we draw a random sample from population, then we expect their size to be around the average population μ give or take about 1 σ

⇒ The expected value of one random draw is population average μ.

⇒ Average of n draws, $\bar{x}_n$ is $E(\bar{x}_n) = \mu$

⇒ Since n is random, $\bar{x}_n$ is also random, so not exactly is $\bar{x}_n = \mu$.

How far off from μ is $\bar{x}_n$.

The standard error of a statistic tells roughly how far off the statistic will be from its expected value.

SE (of a statistic) plays same role as σ for random observation drawn at random.

Square root law,

$$SE(\bar{x}_n) = \frac{\sigma}{\sqrt{n}}$$

① It shows that SE becomes smaller if we use large samples we can use this formula for calculating SE.

② It does not depend on population size but on sample size.

$\Rightarrow$ Sum of n draws $= S_n = n\bar{x}_n$

$\Rightarrow$ Expected value of sum $= E(S_n) = n\mu$

Standard error of sum $= \sqrt{n}\sigma$

Let revisit what % of population approve way of P.M?

But the approval % is an average.
% of likely voters

→ Population falls into either of two categories. Put '1' on one who approves & '0' doesn't

Let there be 5 votes, 10010
Sum of labels = 2 which is no. of people who approved.

∴ % of likely votes who approve is % of 1s among the labels.

So, in a sample of n,
the number of votes in sample who approve is S_n

% approval $\Rightarrow \frac{S_n}{n} \times 100\% = \bar{x}_n \times 100\%$

Expected value of 1s $= \mu \times 100\%$

SE of 1s $= \frac{\sigma}{\sqrt{n}} \times 100\%$

All these formulas are true even when data is simulated, generated according to histogram.

If a random variable X that is simulated has K possible outcomes x_1 to x_K, then

$$\mu = \sum_{i=1}^{K} x_i P(X = x_i) \qquad \sigma^2 = \sum_{i=1}^{K} (x_i - \mu) P(X = x_i)$$

The square root law :-

(E) Toss a coin 100 times. How many tails we can see? give a false.

(A) 1 = tails, 0 = heads } $P(0) = P(1) = 1/2$

no. of tails = sum of 100 draws

$\Rightarrow E(Sn) = 100 \times \mu$ $\qquad \mu = SnP = 0(P(0)) + 1(P(1)) = \frac{1}{2}$

$= 50$

give a false,

$SE(Sum) = \sqrt{100}\,\sigma = 10 \times \frac{1}{2} = 5$

$\sigma^2 = (0 - \frac{1}{2})^2 \frac{1}{2} + (1 - \frac{1}{2})^2 \frac{1}{2} = \frac{1}{4}$

50 tails, give or take 5. | $SE(\text{percentage}) = \frac{\frac{1}{2}}{\sqrt{100}} = 5\%$

The sampling distribution :-

Toss a coin 100 times. Possible tails range from 0 to 100
How likely is each outcome?

This is binomial distribution, $n=100$ & $p=1/2$ because if we call a coin landing success then no. of tails = no. of success in 100 exp.

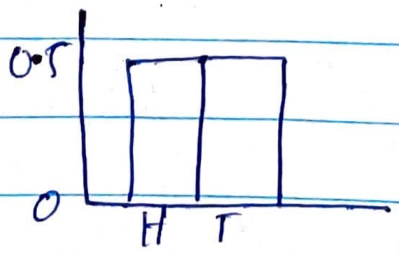
So, if statistic interest of S_n = no. of tails, then S_n is a random variable whose probability histogram is given by binomial distribution. This is called sampling distribution of statistic S_n.

Provides more details about chance properties

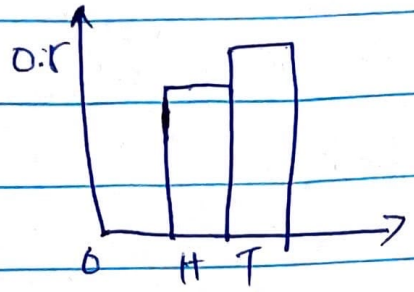
3 histograms :-

The chance process of tossing a coin 100 times with 3 different histograms

① Probability histogram → theoretical



② Actual experimented histogram, let's say 53T & 47H



③ Sampling distribution: if we are interested in a statistic such as the no. of tails in 100 tosses, then we get this. which gives Sampling distribution of that statistic

Law of large numbers :-

The square root law says that SE of sample mean goes to zero as sample size increases. So, the sample mean will likely to be close to its expected value μ if the sample size is large. This is law of large numbers

→ Applies only for averages & %
→ Sampling with replacement.

Central limit Theorem:-

Let go back to the online game · $n=10$, $p=0.2$ (small win) · We gamble n times & found X has binomial dist.

When we sample with replacement, and n get large, then the sampling distribution of sample average/sum/% approximately follows normal curve.

To standardize, subtract expected value of statistic, then divide by SE.

(Q) POP
$N = 10{,}000$; $n = 100$
$1 = 6000$ % 1's will be
$0 = 4000$

(A) $n = 100$ $\sigma_1 = \sigma_0 = 0.5$
$P(1) = \frac{6000}{10000} = 0.6$
$P(0) = 0.4$

$E(S) = 100 \times \mu = 100 \times (0 \times P(0) + 1 P(1)) = 100 \times 0.6 = 60$

$SE\% = \frac{0.6}{\sqrt{100}} = \frac{6}{10 \times 10} = 6\%$

$SE = \sqrt{100}\,\sigma = 10 \times 0.5 = \underline{5}$

Q) 100 pledges, each pledge is equally likely to be $10, $50 & $100.
$\sigma_{10} = \sigma_{50} = \sigma_{100} = \$37$. Expected value of sum of 100 pledges?

A) expected value of sum $= n\mu$
$= 100 \times \mu$

$$\mu = \sum_{i=1}^{k} x_i P(X = x_i) = \frac{\$10 + \$50 + \$100}{3} = \$53.33$$

$i = 10,$

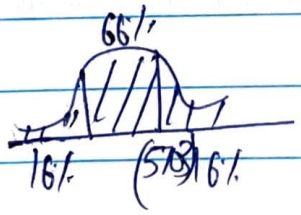
$\therefore E(S_n) = \$5333$

Q) Same question. Chance that 100 pledges more than $5700

A) $SE = \sqrt{n}\,\sigma = \sqrt{100} \times \sigma = \$370$

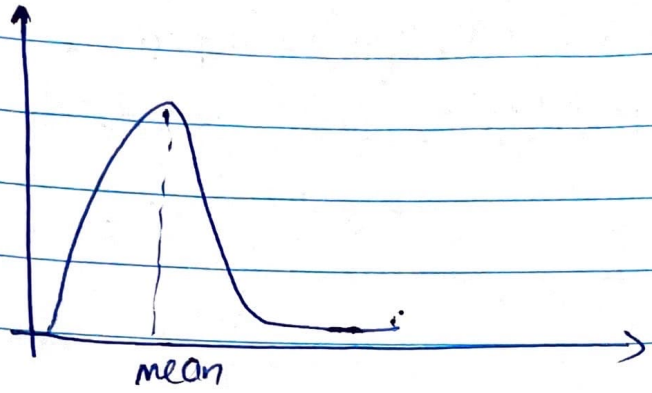
5333
370
$5703

$5703 ≈ $5700 is one standard deviation away,
so it's 16%.



(Contd CLT):-
It shows the statistic has normal distribution no matter what population histogram is.

let $\sigma = \$38,000$ in an income graph skewed right.
$\mu = \$67,000$



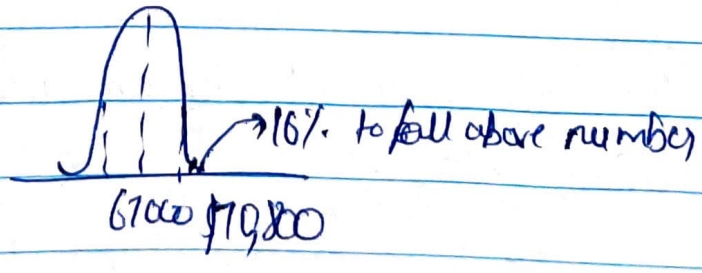
So, if we same "n" incoms on random, then

$\bar{X}_n$ = sample average follows normal curve. To do normal approximation, we do subtract off the expected value of statistic which in this case is average of all incomes $E(\bar{X}_n) = \mu = \$67000$ and then divide by $SE(\bar{X}_n) = \frac{\sigma}{\sqrt{n}} = \frac{\$38000}{\sqrt{n}}$

(1) Let take 100 incomes, if $n = 100 \Rightarrow \sqrt{n} = 10$

$\therefore SE(\bar{X}_{10}) = \$3800$

The CLT says, the sample mean follows normal curve centered at \$67000 with SE = \$3800



(Q) we did 'n' gambles and we had X = no of small prizes. Since we use labels, 1 = everything of small prize, everything else = 0

$\Rightarrow X$ = no. of small prizes = sum of labels

Since we are looking at sum, we use CLT.

$\Rightarrow \mu = p$ & $\sigma = \sqrt{p(1-p)}$ for 1 gamble

$\Rightarrow \mu = np$ & $\sigma = \sqrt{np(1-p)}$