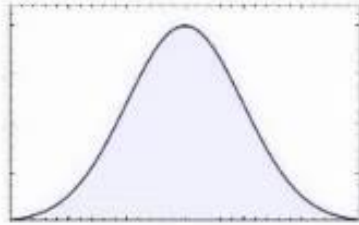


Normal Distribution?



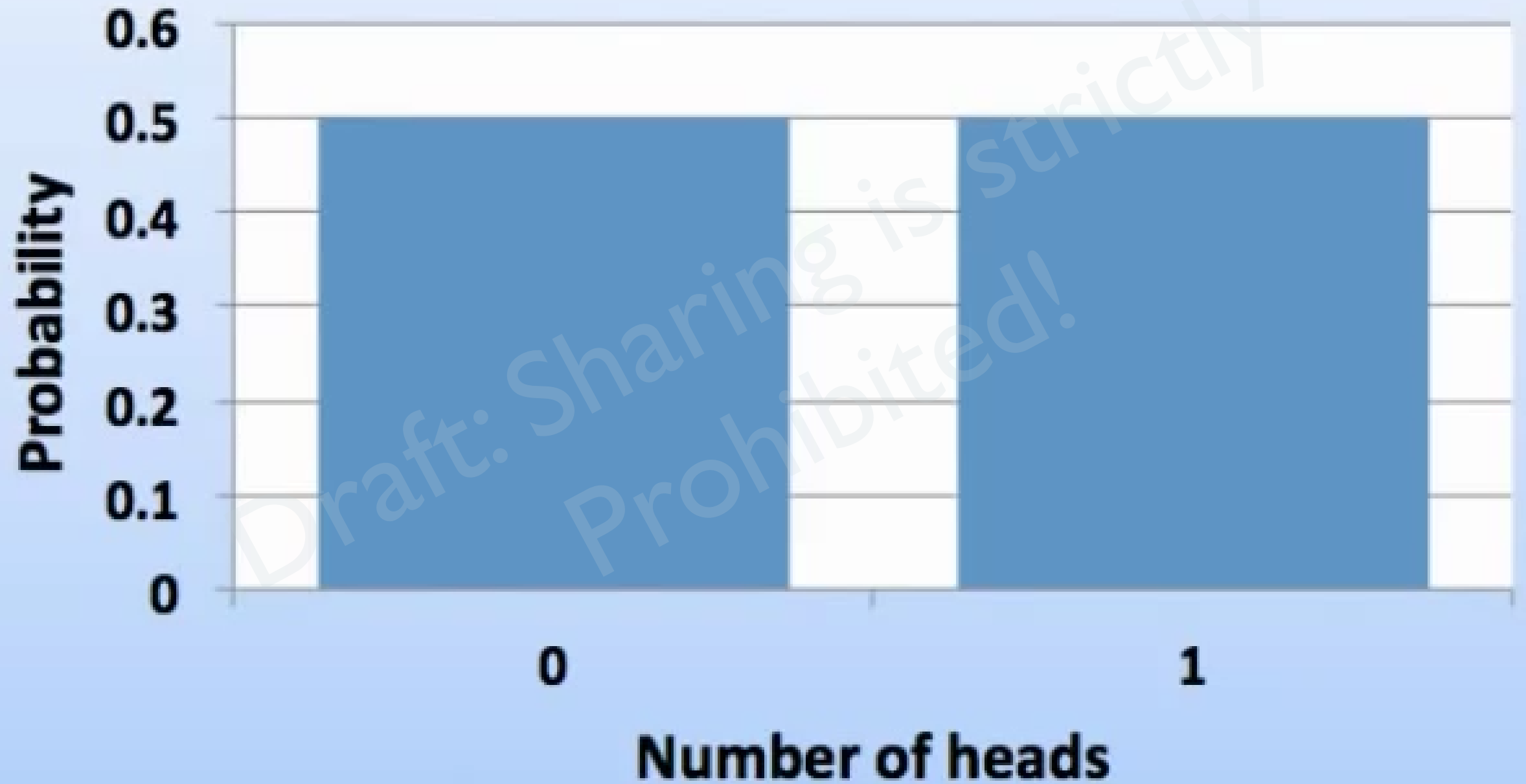
Central Limit Theorem



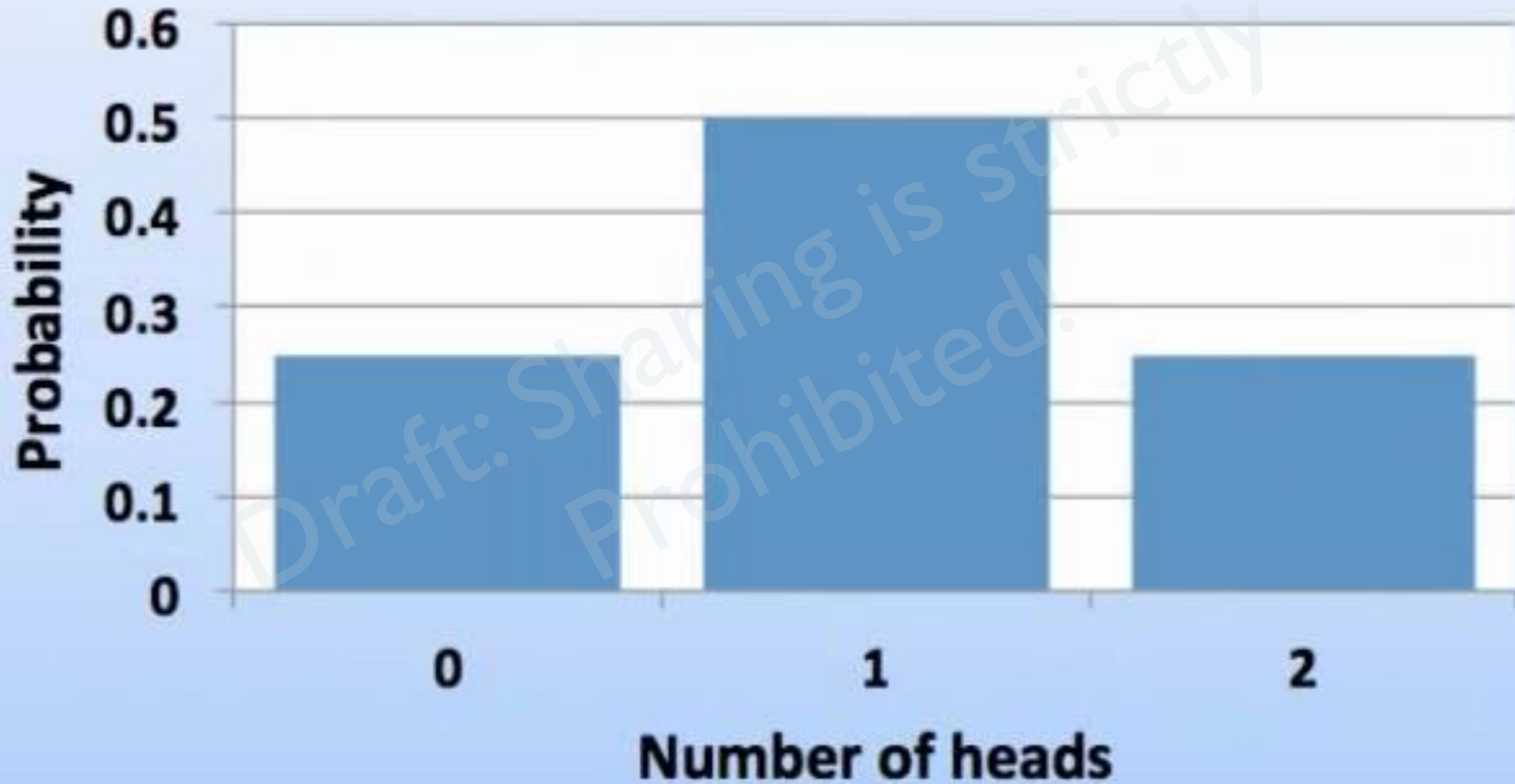
as n increases, the distribution of the sample mean or sum approaches a normal distribution



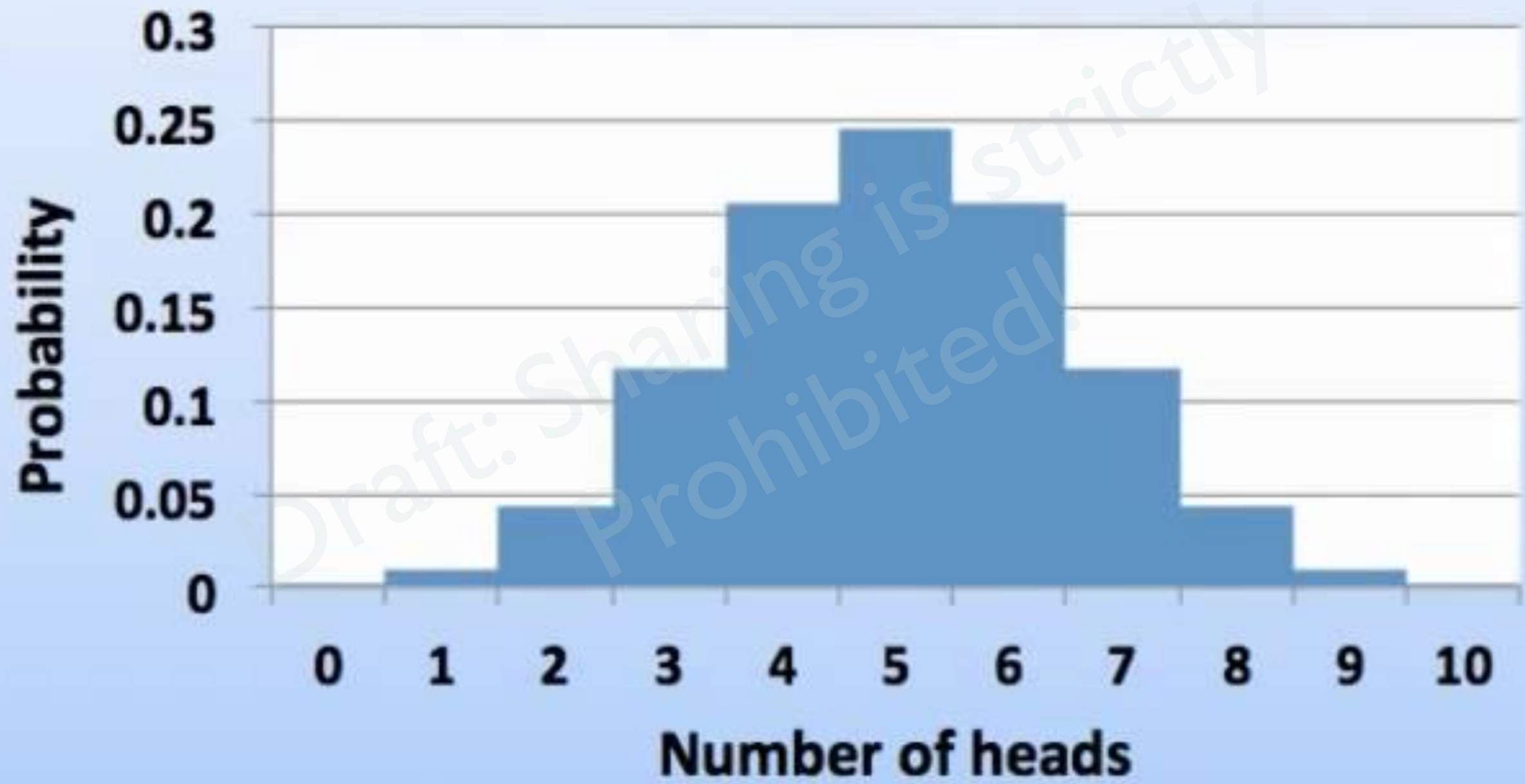
1 Coin toss



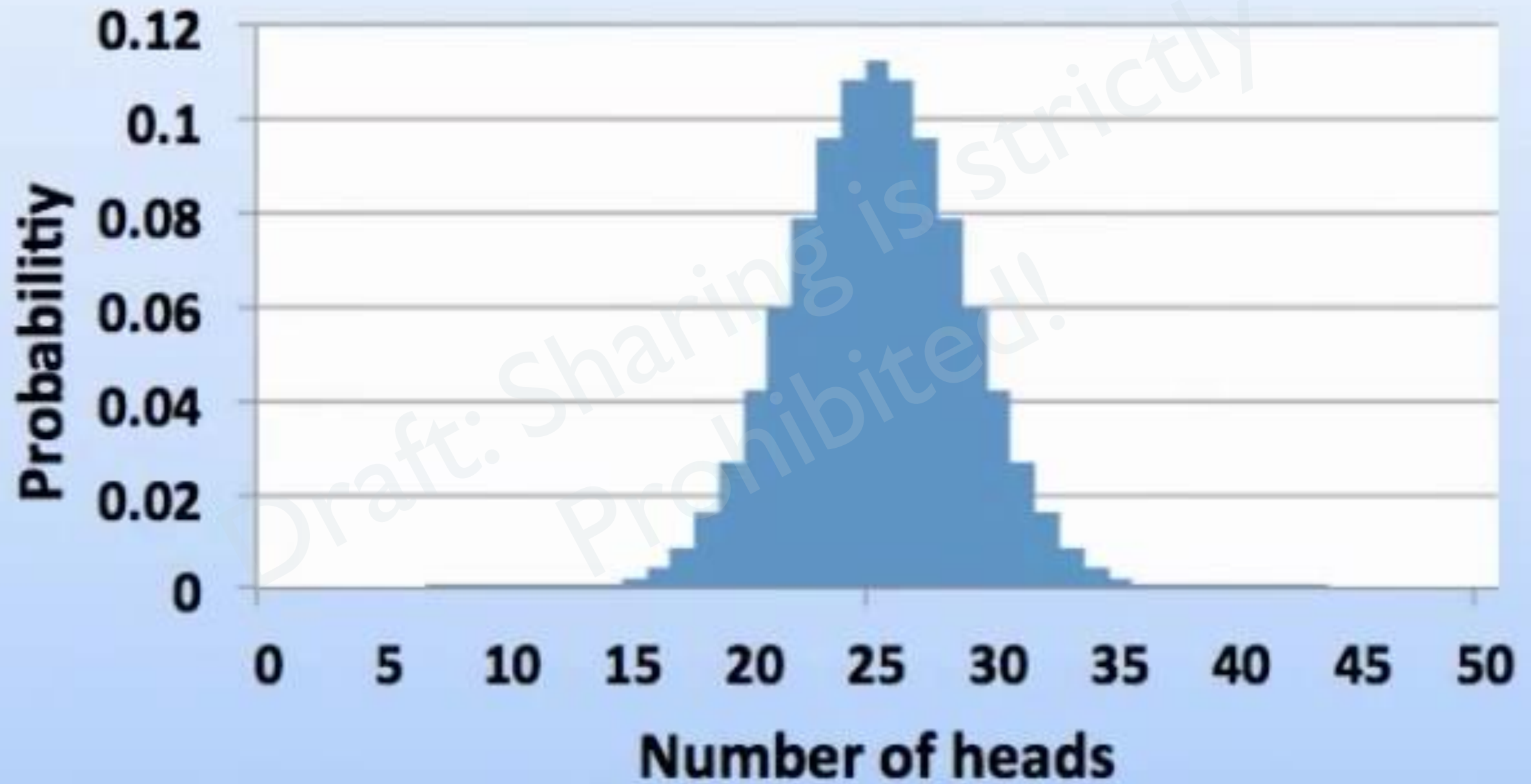
2 coin tosses



10 coin tosses



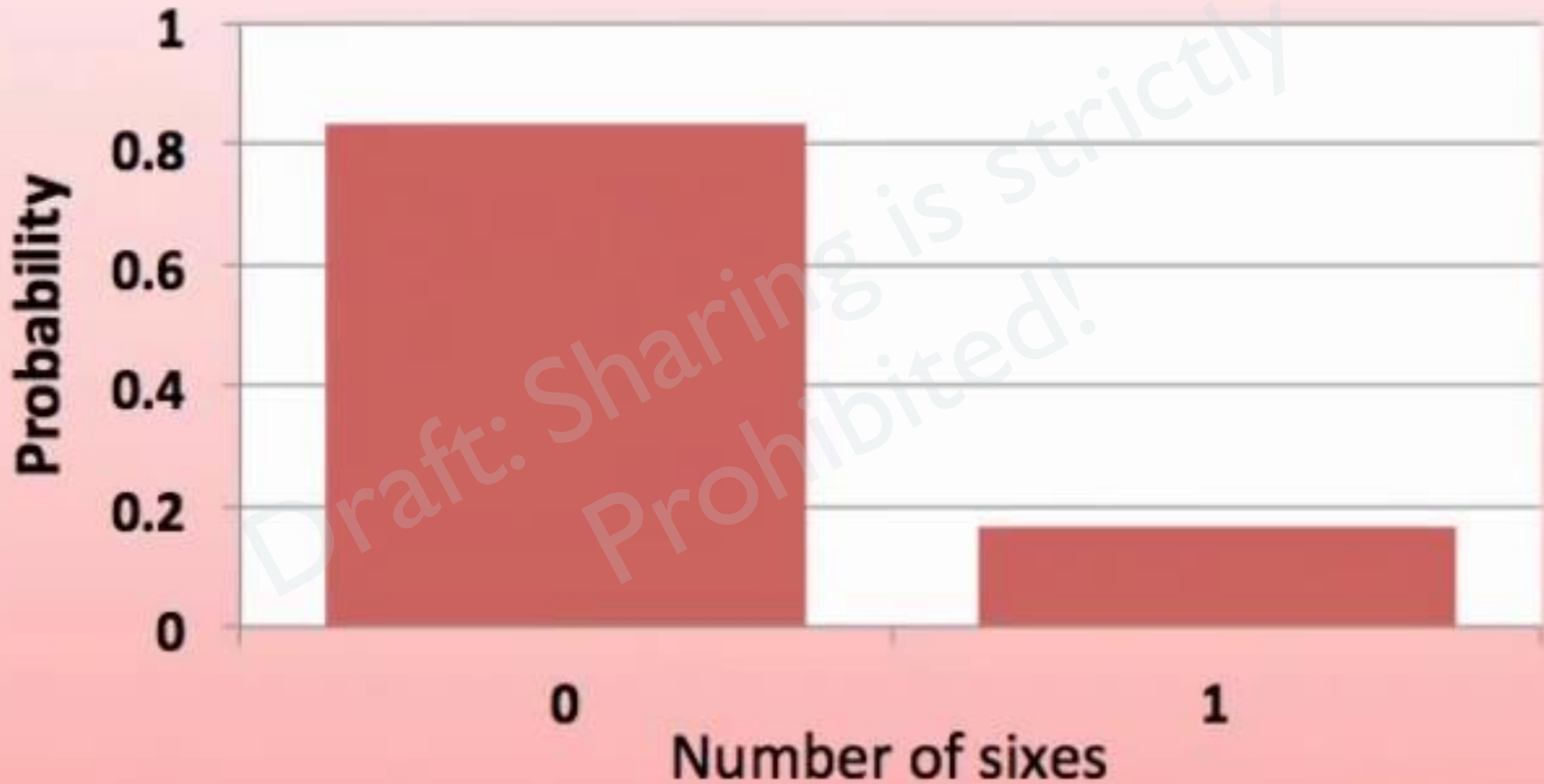
50 coin tosses



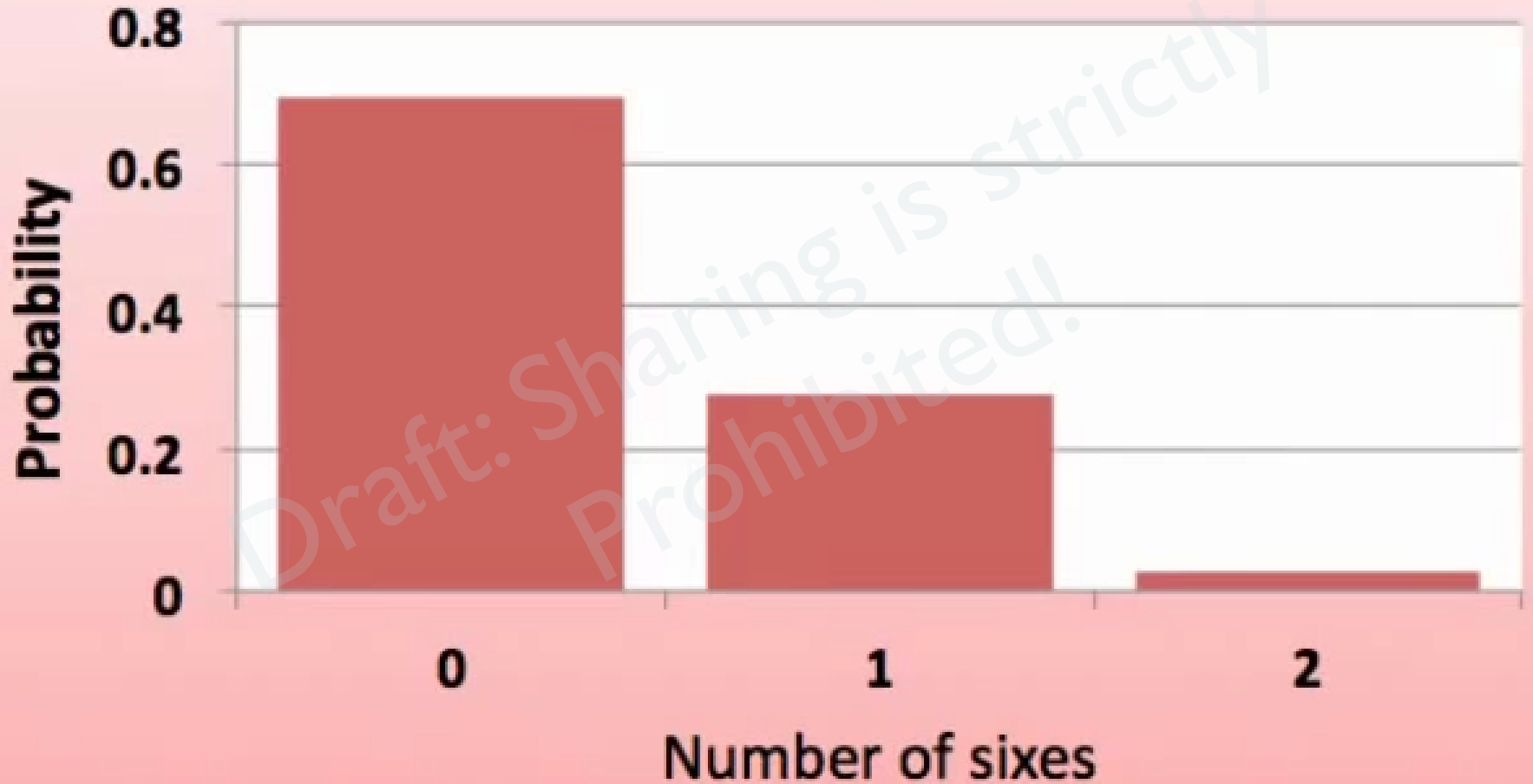


Draft: Sharing is strictly Prohibited!

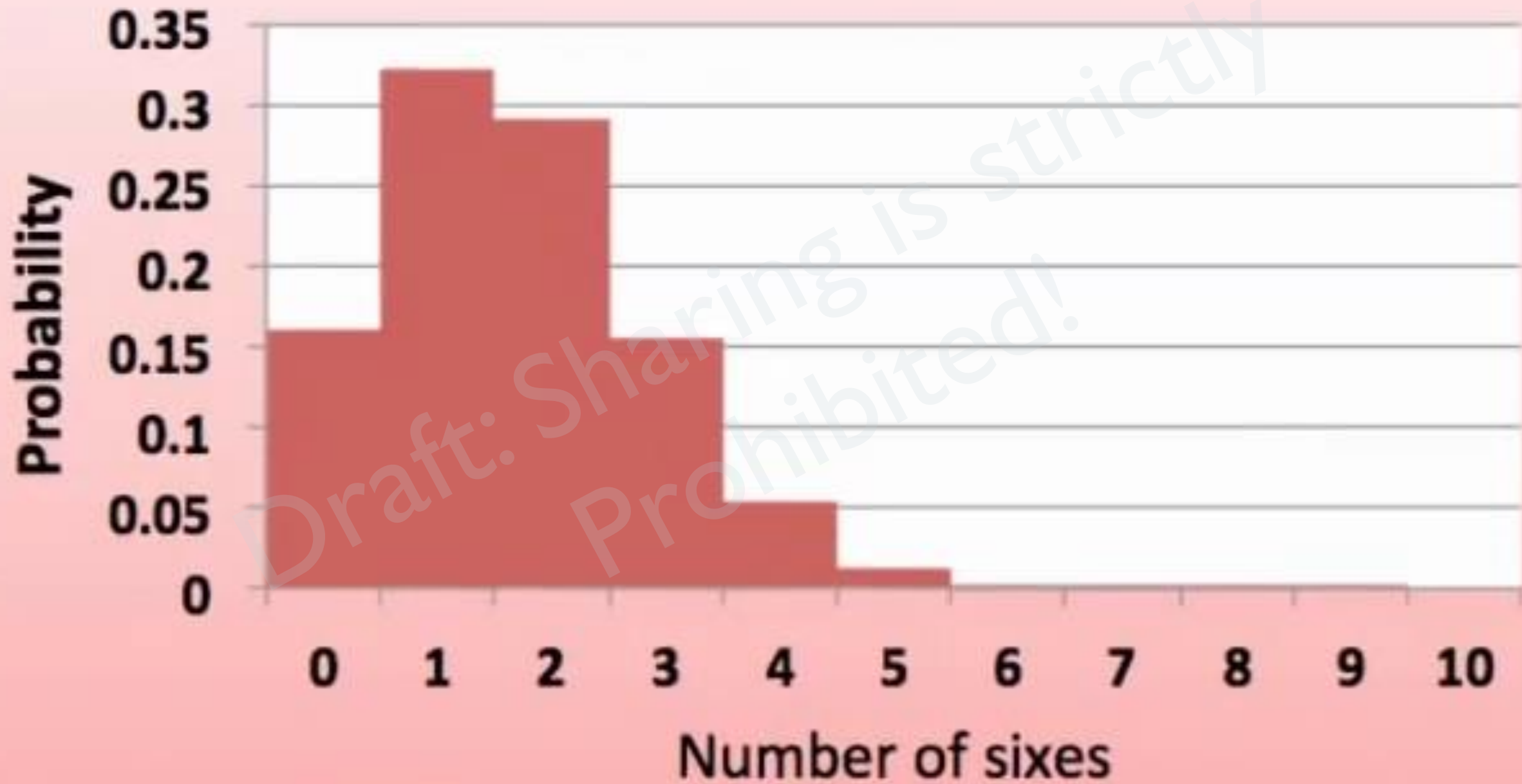
1 dice roll



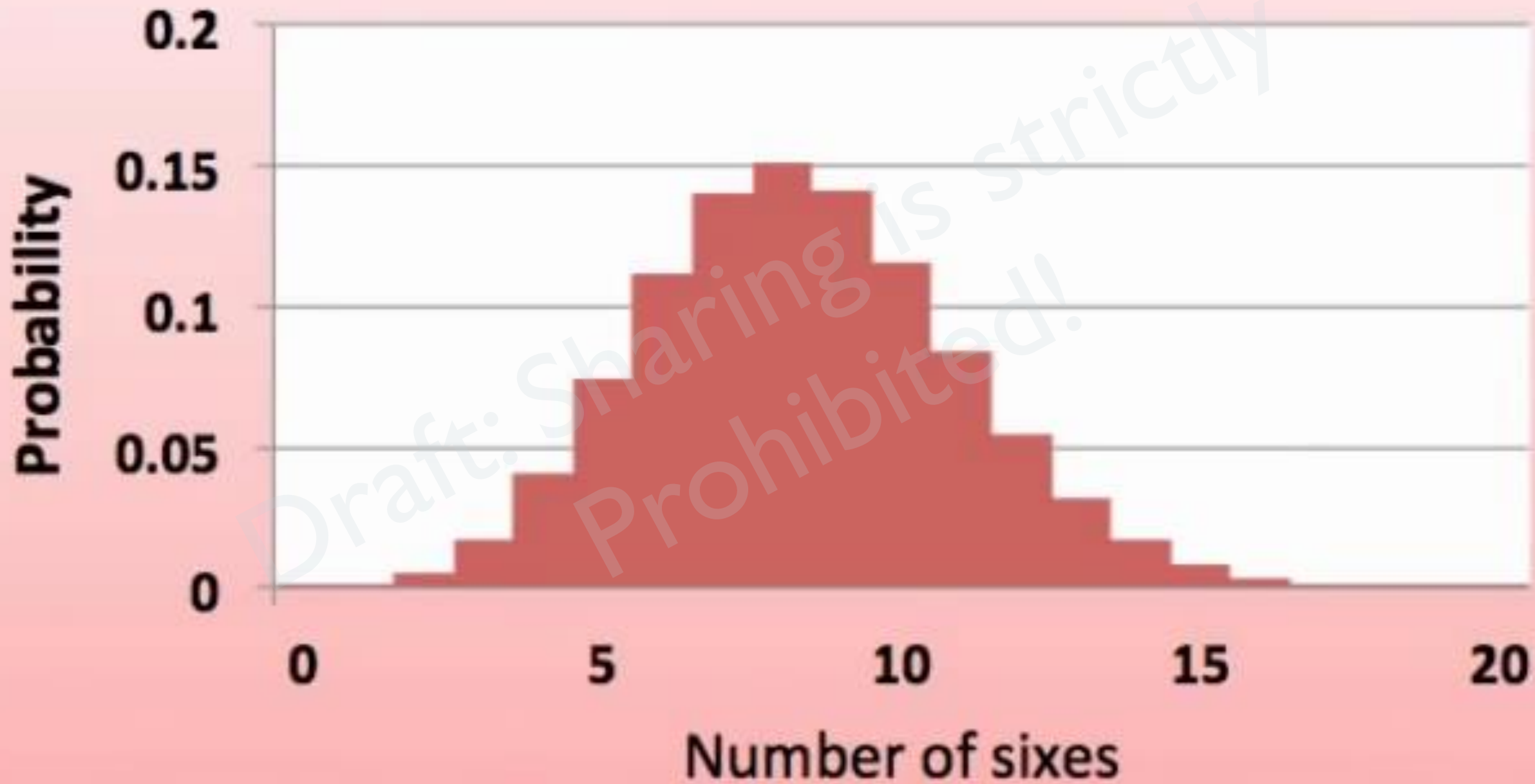
2 dice rolls

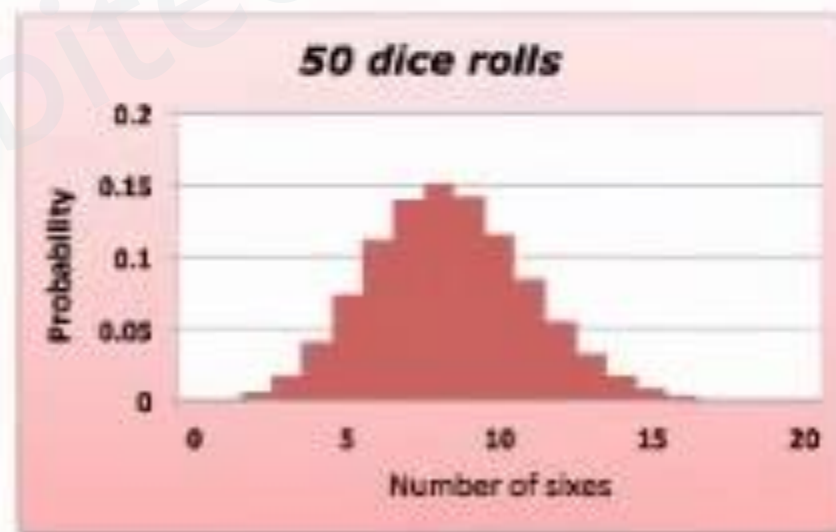
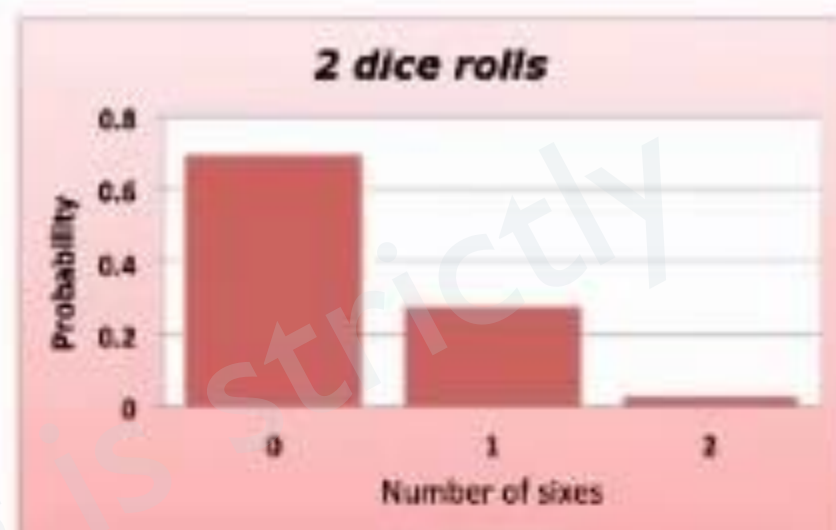
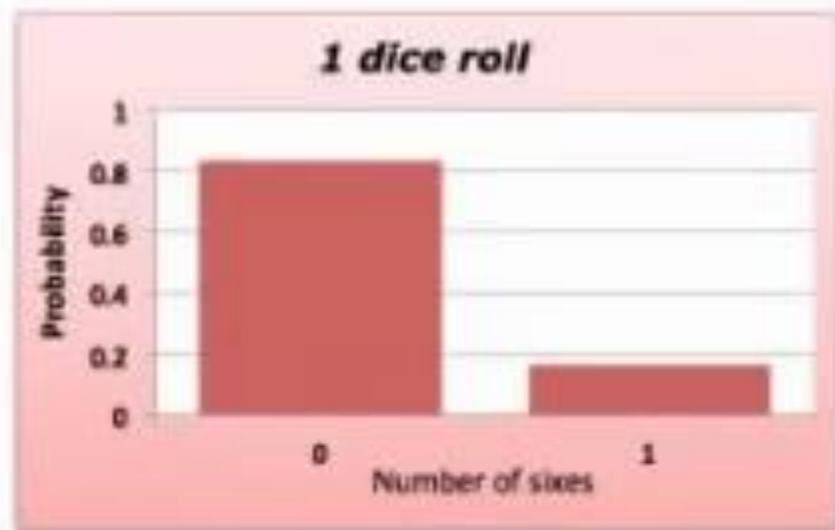


10 dice rolls



50 dice rolls

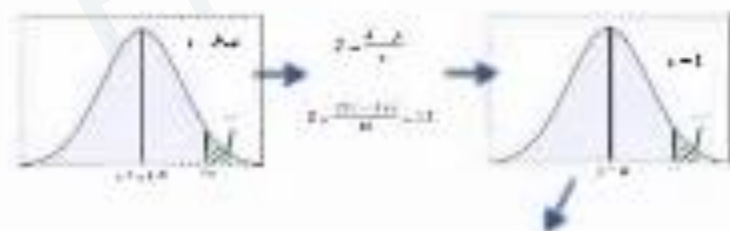




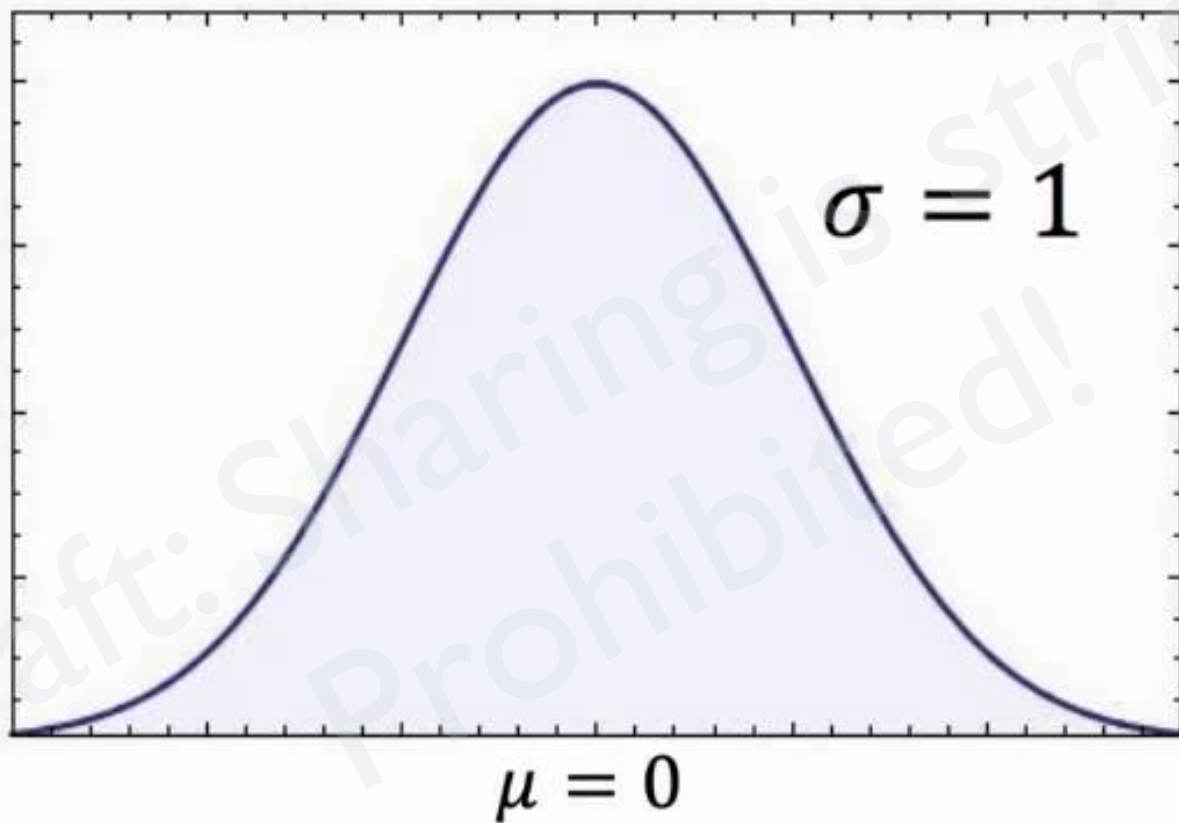
as n increases, the distribution of the sample mean or sum approaches a normal distribution.

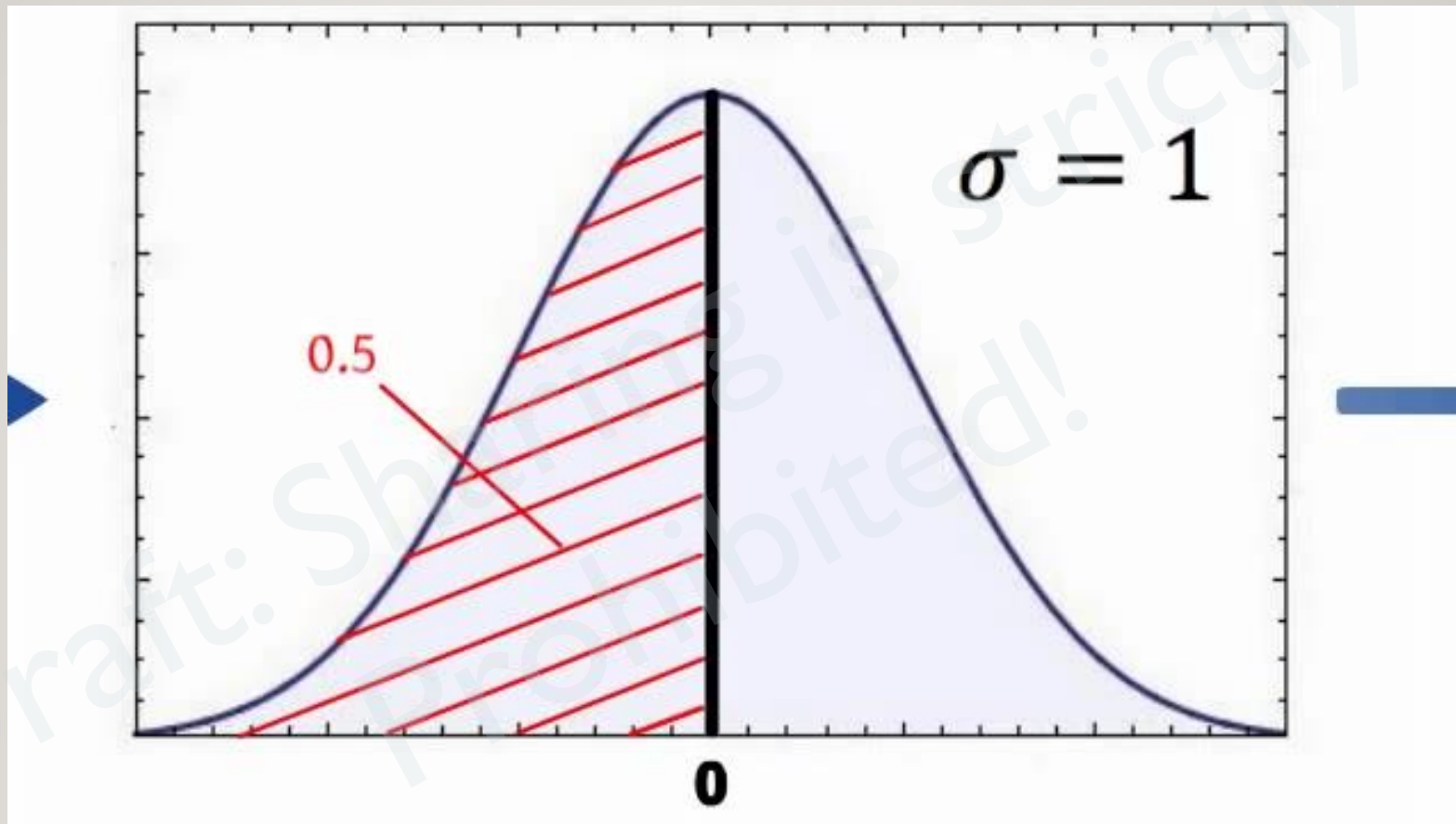
=!?

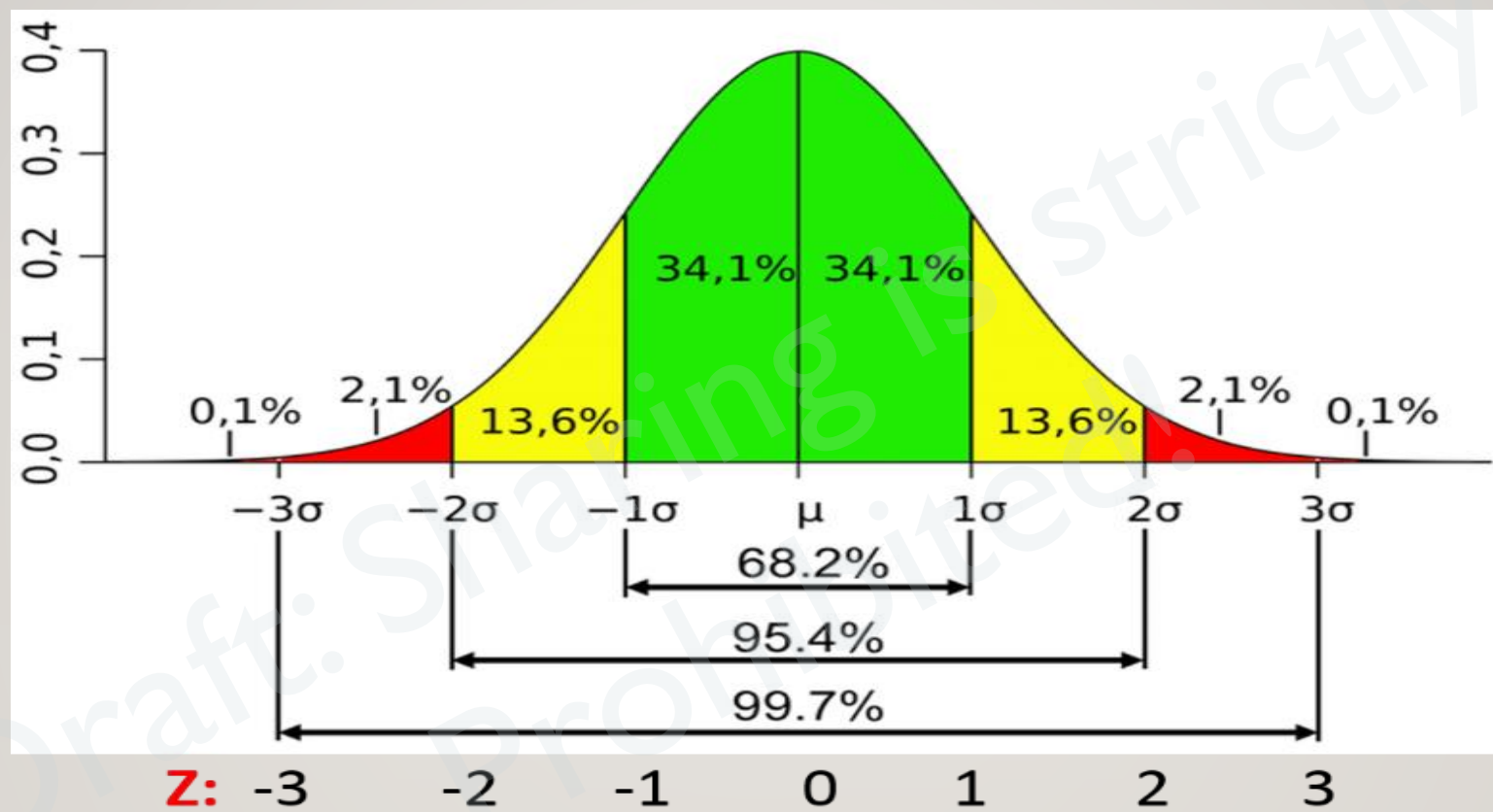
Standard normal

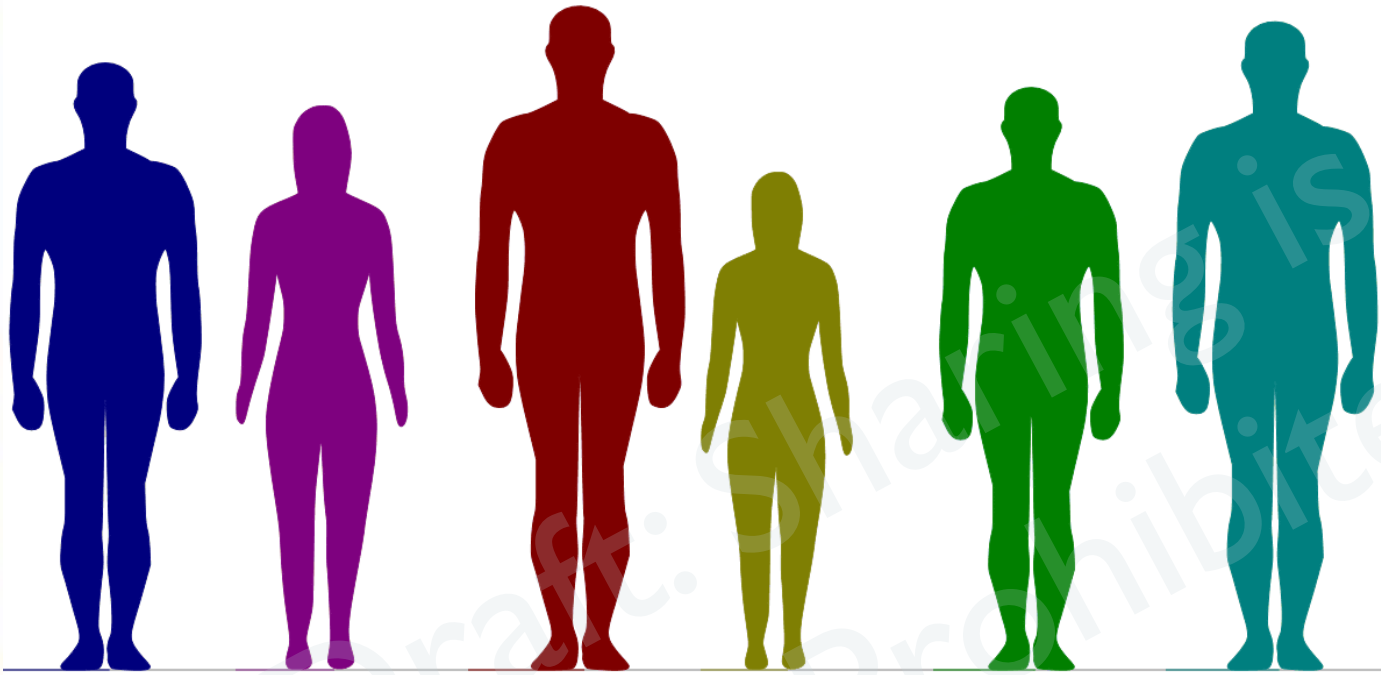


$P(\text{Height} > 190) = 0.057 \leftarrow P(Z > 1.5) = 0.067$



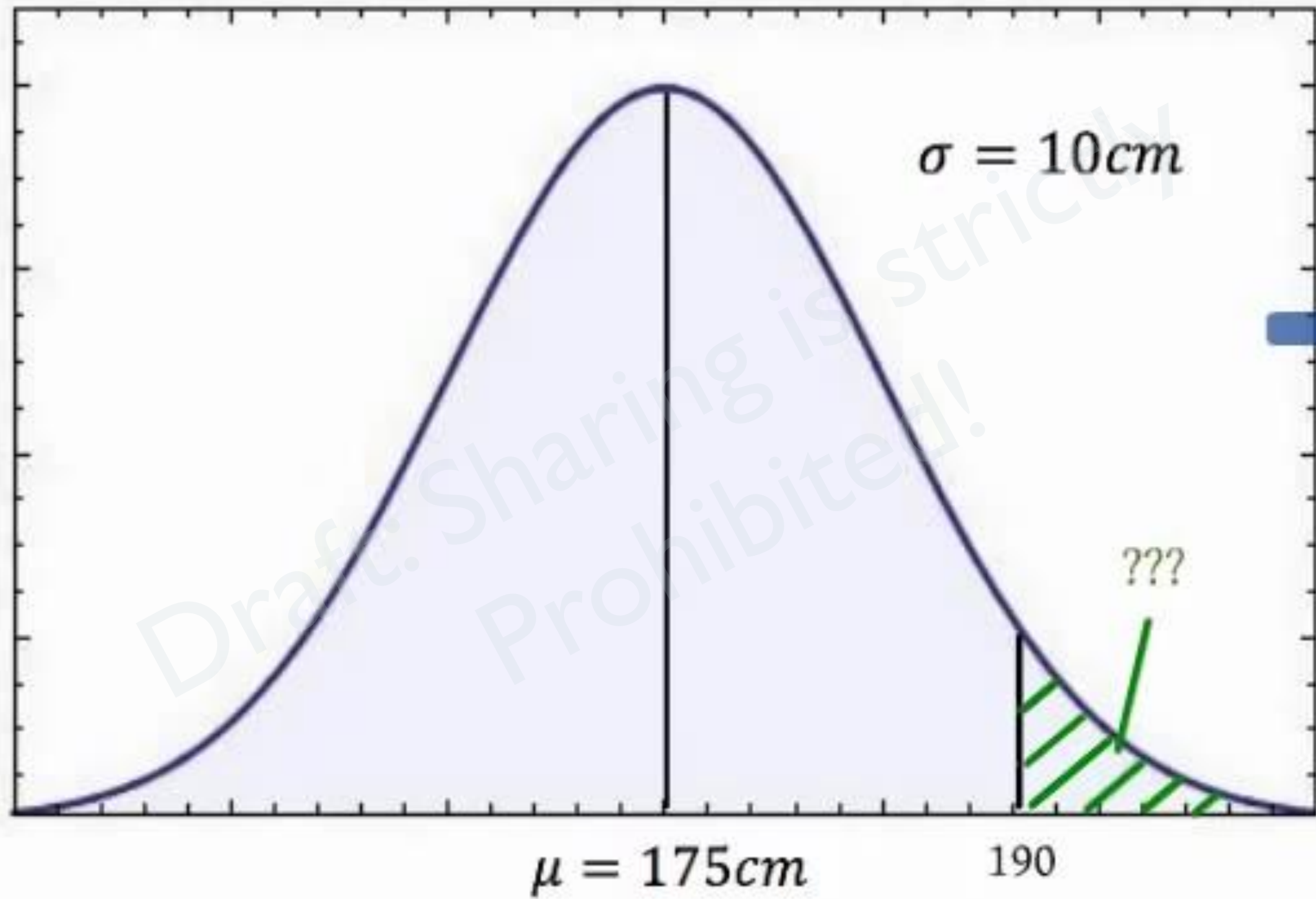


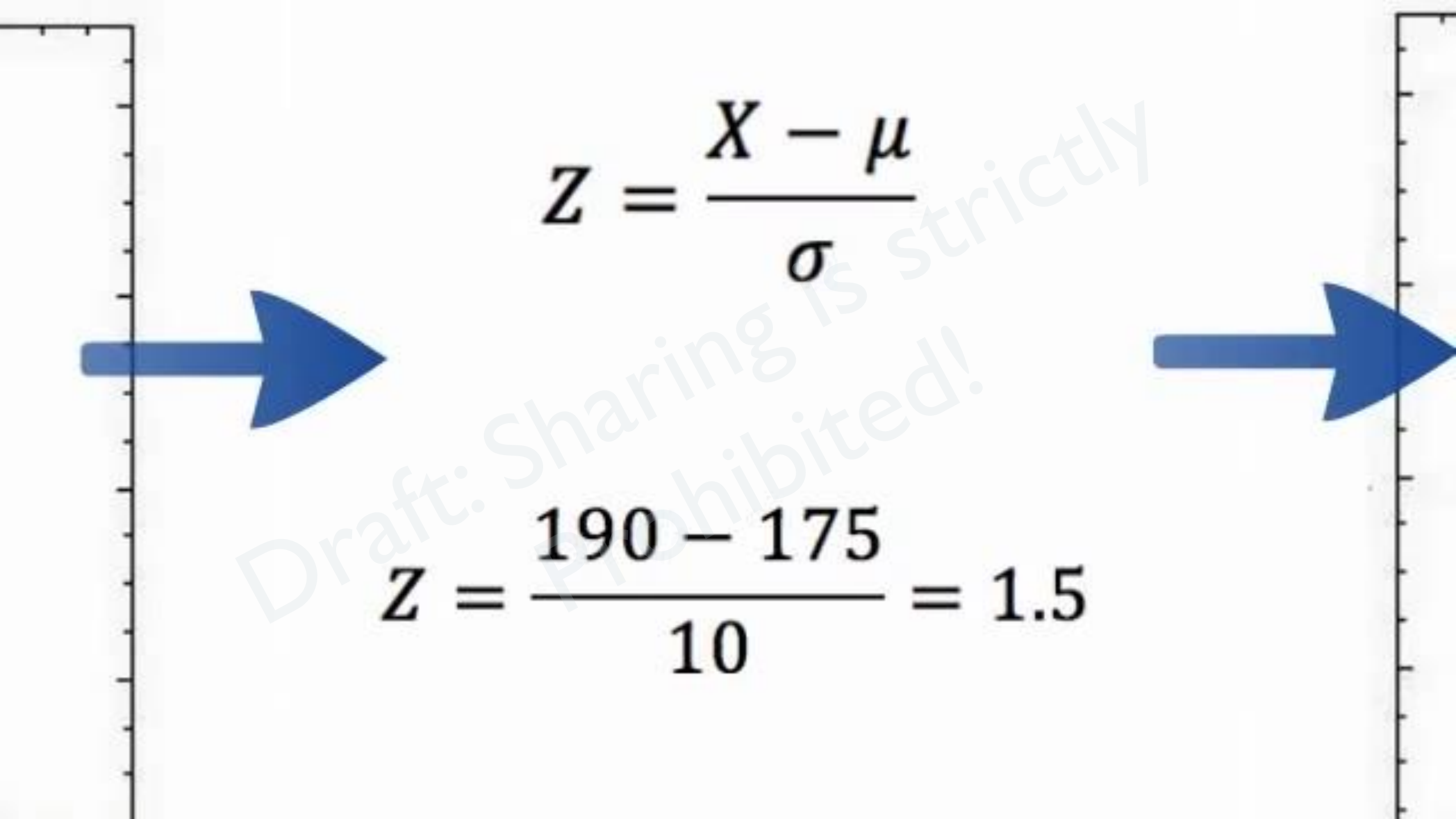




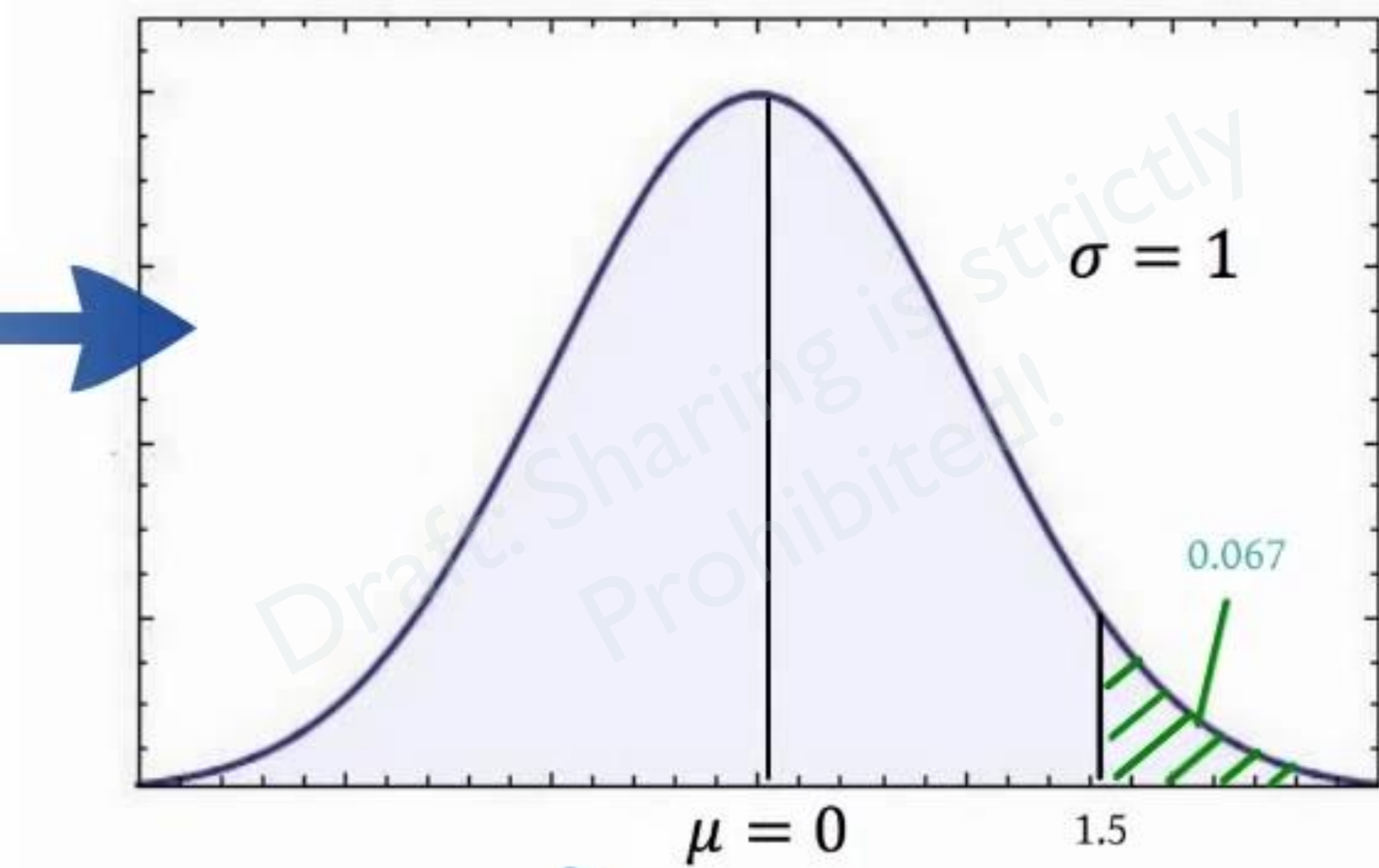
$$\mu = 175cm$$

$$\sigma = 10cm$$



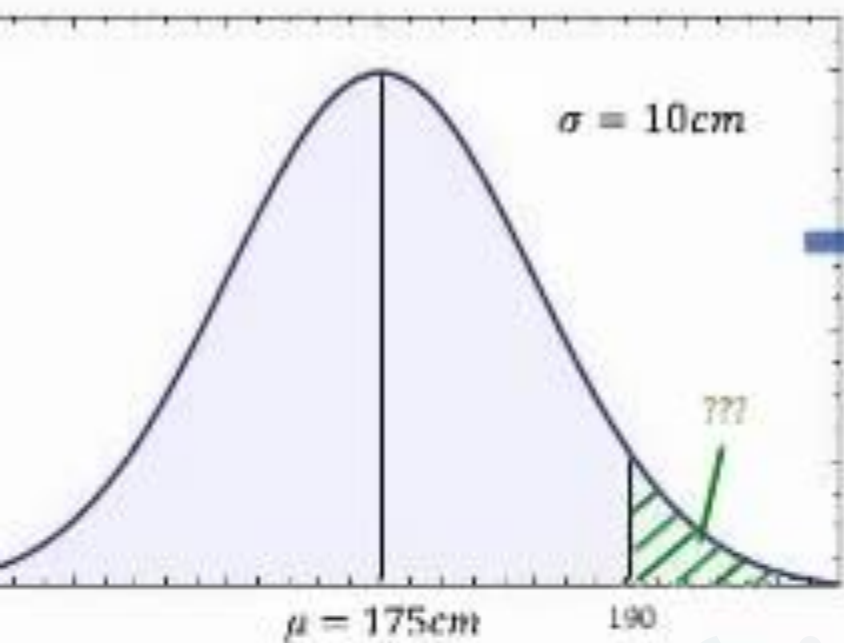

$$Z = \frac{X - \mu}{\sigma}$$

$$Z = \frac{190 - 175}{10} = 1.5$$



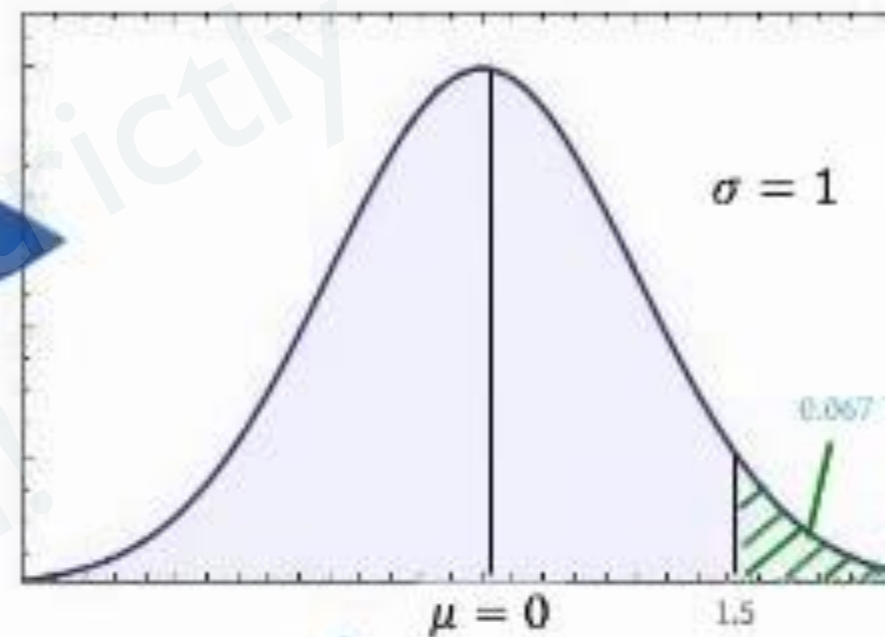
$$P(Z > 1.5) = 0.0667$$

$$P(\text{Height} > 190) = 0.067$$



$$Z = \frac{X - \mu}{\sigma}$$

$$Z = \frac{190 - 175}{10} = 1.5$$

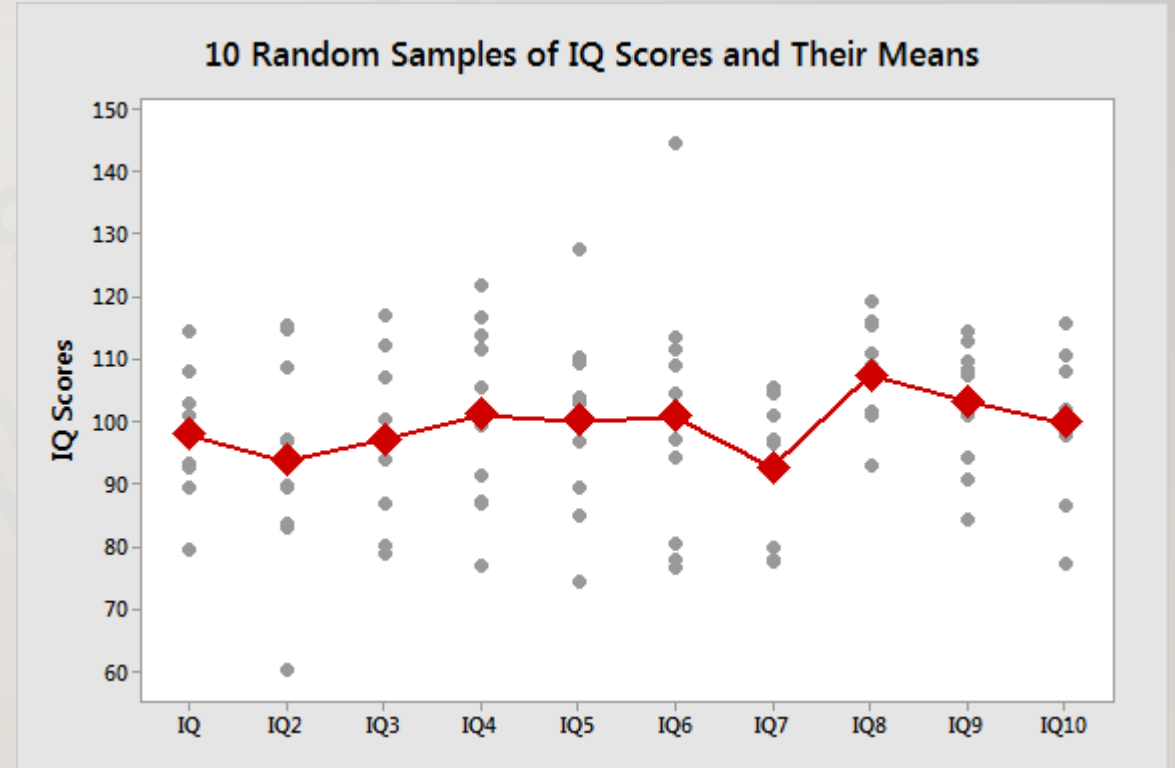
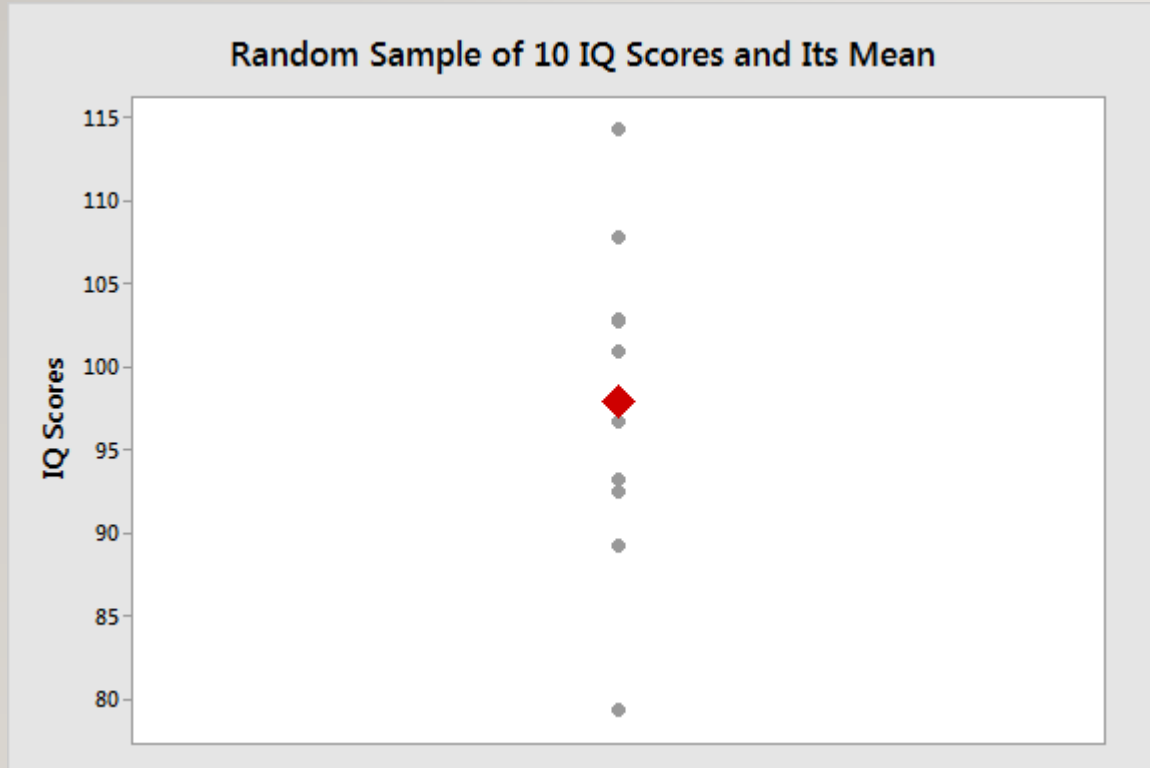


$$P(\text{Height} > 190) = 0.067 \longleftarrow P(Z > 1.5) = 0.067$$

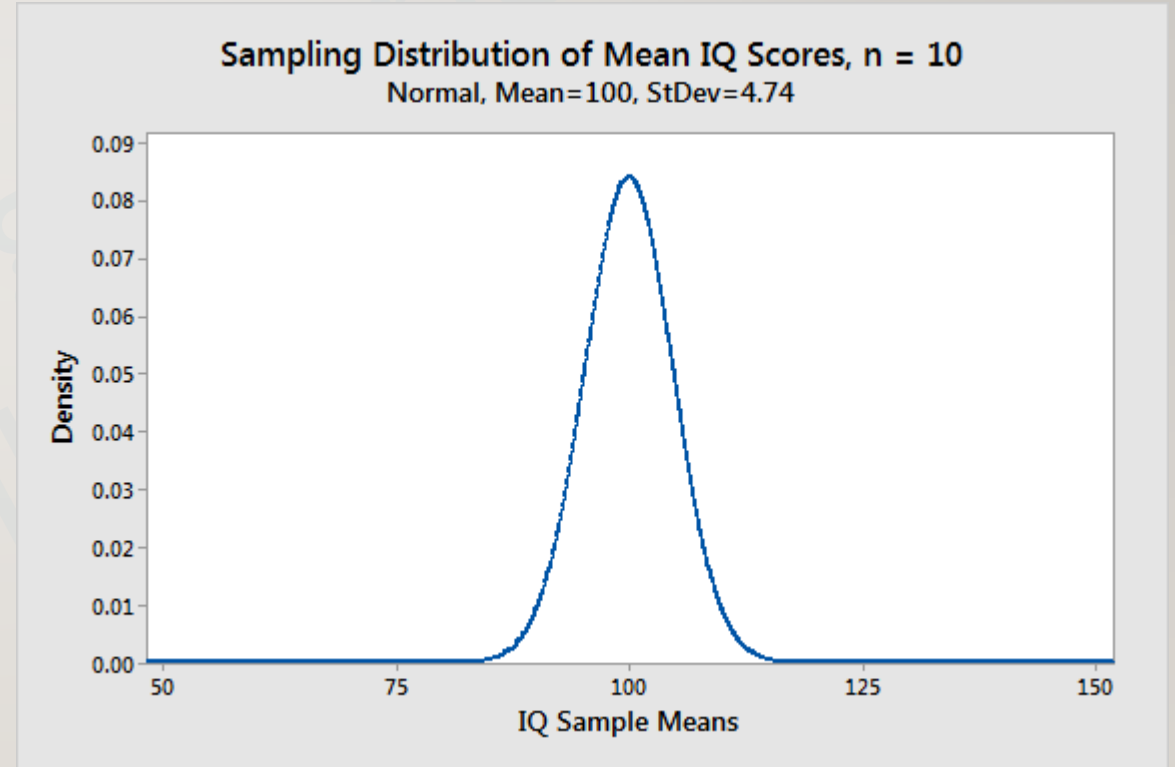
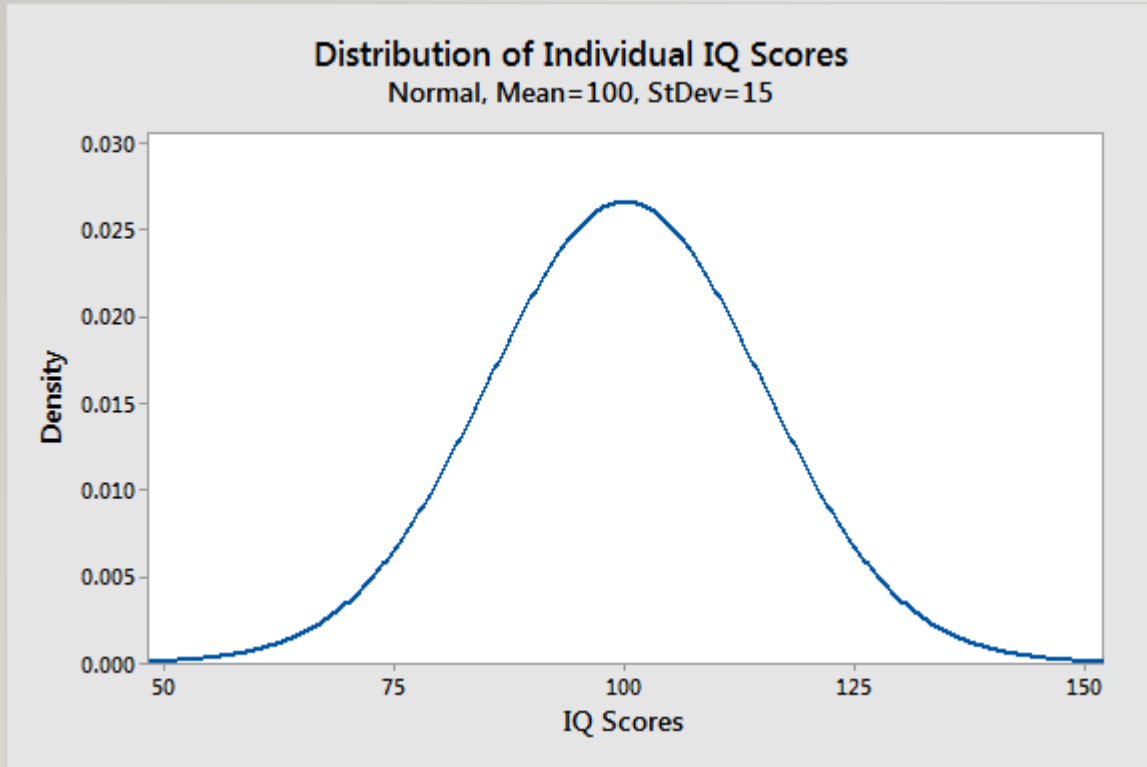
Standard Deviation (SD) vs. Standard Error (SE)

- **Standard deviation:** Quantifies the variability of values in a dataset. It assesses how far a data point likely falls from the mean.
- **Standard error:** Quantifies the variability between samples drawn from the same population. It assesses how far a sample statistic (i.e., Sample Mean) likely falls from a population parameter (i.e., Population Mean).

Standard Deviation (SD) vs. Standard Error (SE)



Standard Deviation (SD) vs. Standard Error (SE)



Standard Deviation (SD) vs. Standard Error (SE)

To calculate standard deviation for population data, the formula is:

$$\sigma = \sqrt{\frac{1}{N} \sum_{i=1}^N (x_i - \mu)^2}$$

To calculate standard deviation for sample data, you can use the following formula:

$$s = \sqrt{\frac{1}{N-1} \sum_{i=1}^N (x_i - \bar{x})^2}$$

Standard Deviation (SD) vs. Standard Error (SE)

To calculate the standard error, the formula is:

$$SE = \frac{\sigma}{\sqrt{n}}$$

If the population standard deviation is not known, use this formula:

$$SE = \frac{s}{\sqrt{n}}$$

Suppose a large number of students from multiple schools participated in a design competition. From the whole population of students, evaluators chose a sample of 300 students for a second round. The mean of their competition scores is 650, while the sample standard deviation of scores is 220. Now let's calculate the standard error.

Draft: Sharing is strictly Prohibited!

Suppose a large number of students from multiple schools participated in a design competition. From the whole population of students, evaluators chose a sample of 300 students for a second round. The mean of their competition scores is 650, while the sample standard deviation of scores is 220. Now let's calculate the standard error.

1. Find the square root of the sample size. In our example, $n = 300$, and you can calculate the square root in Excel or Google Sheets using the following formula: $=300^{0.5}$. So $n = 17.32$

2. Find the standard deviation for your data sample. You can do this following the steps laid out in section three, but for now we'll take it as known that the sample standard deviation $S = 220$.

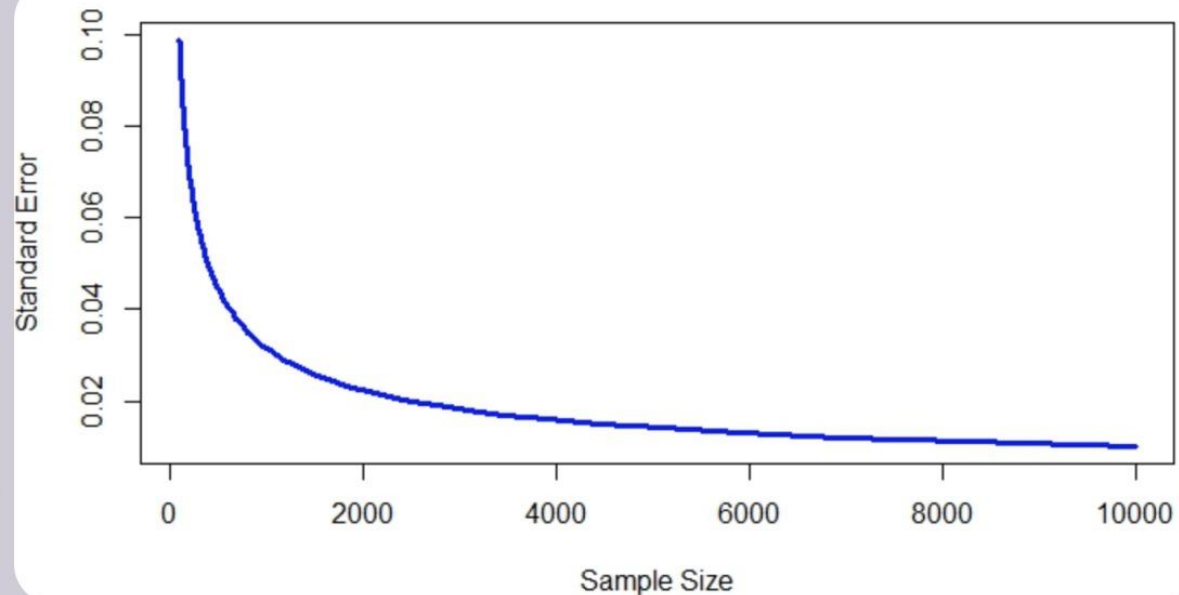
3. Divide the sample standard deviation by the square root of the sample size. So, in our example, $220 / 17.32 = 12.7$. So, the standard error is 12.7.

When reporting the standard error, you would write (for our example): The mean test score is 650 ± 12.7 (SE).

Standard Deviation (SD) vs. Standard Error (SE)

What is the relationship between standard error (SE) and the sample size?

Sample size is inversely proportional to standard error, and so the standard error can be minimized by using a large sample size. As you can see from this graph, the larger the sample size, the lower the standard error.



Standard Deviation (SD) vs. Standard Error (SE)

Standard error vs standard deviation: What's the difference?

- Standard deviation describes variability within a single sample, while standard error describes variability across multiple samples of a population.
- Standard deviation is a descriptive statistic that can be calculated from sample data, while standard error is an inferential statistic that can only be estimated.
- Standard deviation measures how much observations vary from one another, while standard error looks at how accurate the mean of a sample of data is compared to the true population mean.

Student's t Distribution

When to use *t-distribution* instead of *normal distribution*?

- when estimating the mean of a normally distributed population in situations where the sample size is small and the population standard deviation is unknown.
- An unknown population standard deviation implies that it would have to be estimated from the samples itself which is inaccurate with small sample sizes.
- A sample size ≥ 30 implies the use of a normal distribution, a sample size < 30 implies the use of the t-distribution.

1. Overview

+

t Distribution

* t-test developed in 1908 by William Sealy Gosset



Draft: Sharing is strictly Prohibited!

t Distribution

- * t-test developed in 1908 by William Sealy Gosset
- * Solved the problem of "small sample statistics"



Draft: Sharing is strictly Prohibited!

t Distribution



- * t-test developed in 1908 by William Sealy Gosset
- * Solved the problem of "small sample statistics"

Underlying distribution is
NORMAL

t Distribution



- * t-test developed in 1908 by William Sealy Gosset
- * Solved the problem of "small sample statistics"

Underlying distribution is
NORMAL

Population standard
deviation unknown

t Distribution



- * t-test developed in 1908 by William Sealy Gosset
- * Solved the problem of "small sample statistics"

Underlying distribution is
NORMAL

Population standard
deviation unknown

Sample size is too small for
C.L.T to apply

t Distribution



- * t-test developed in 1908 by William Sealy Gosset
- * Solved the problem of "small sample statistics"

Underlying distribution is
NORMAL

Population standard
deviation unknown

Sample size is too small for
C.L.T to apply

- * The following measures would be t-distributed

$$\frac{\bar{x} - \mu}{s / \sqrt{n}}$$

$$\frac{b - \beta}{SE(b)}$$

$$\frac{(\bar{x}_1 - \bar{x}_2) - (\mu_1 - \mu_2)}{\sqrt{\left(\frac{s_1^2}{\sqrt{n_1}}\right) + \left(\frac{s_2^2}{\sqrt{n_2}}\right)}}$$

2. Sampling Recap

+

t Distribution

SAMPLING RECAP!

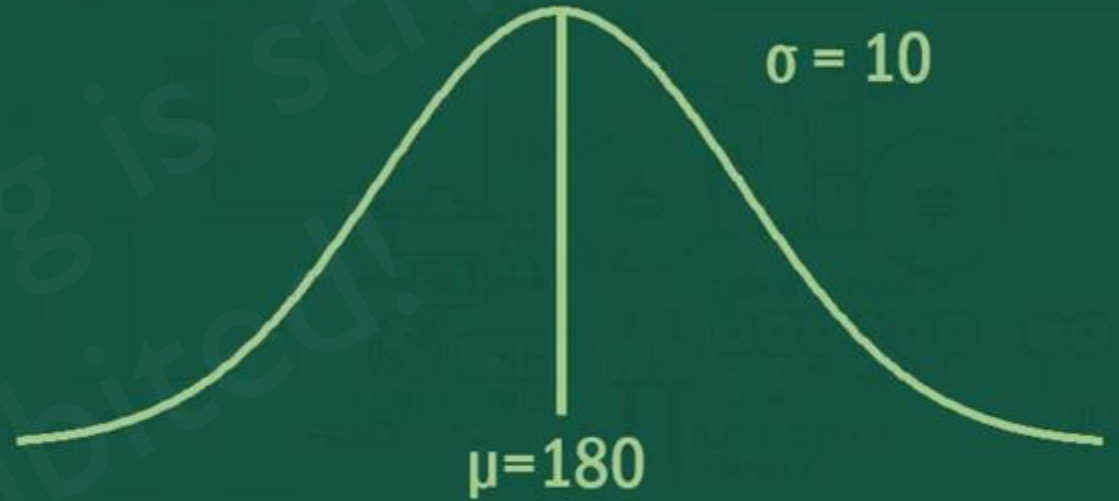
- * Take a sample of five observations
+ from a normally distributed population

t Distribution

SAMPLING RECAP!

- * Take a sample of five observations from a normally distributed population

Heights of female basketballers (cm)



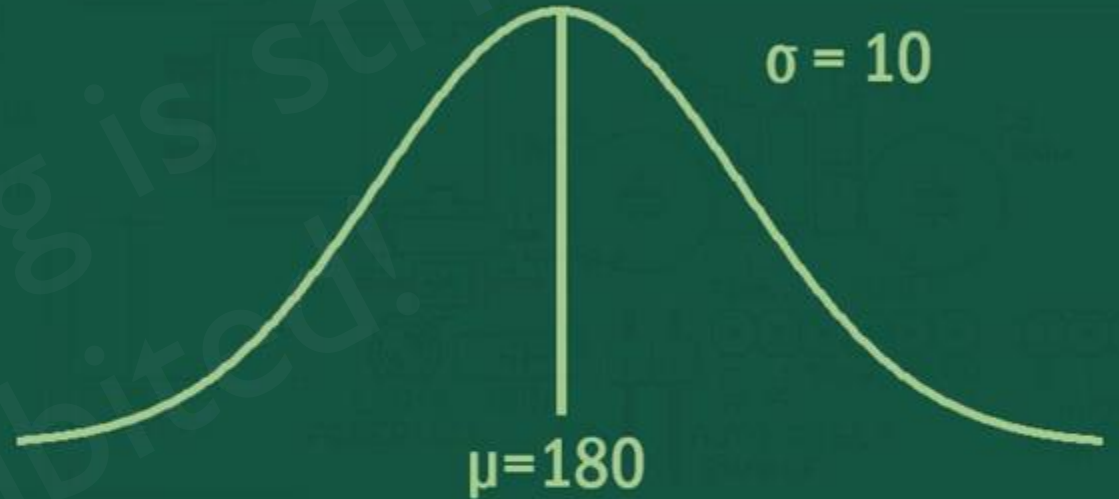
t Distribution

SAMPLING RECAP!

- * Take a sample of five observations from a normally distributed population

[183 , 170 , 189 , 191 , 203]

Heights of female basketballers (cm)



t Distribution

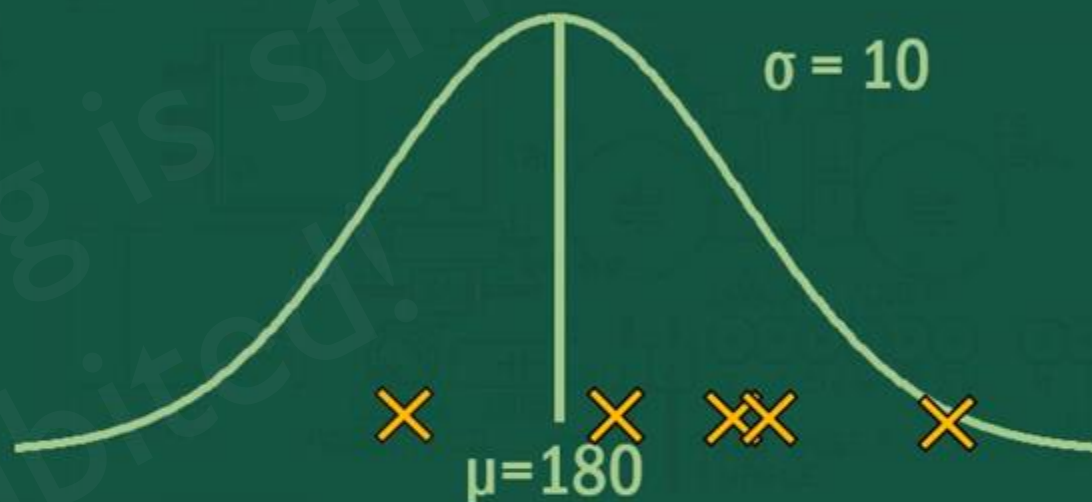
SAMPLING RECAP!

- * Take a sample of five observations from a normally distributed population

[183 , 170 , 189 , 191 , 203]

- * Find the average of that sample

Heights of female basketballers (cm)



+

t Distribution

SAMPLING RECAP!

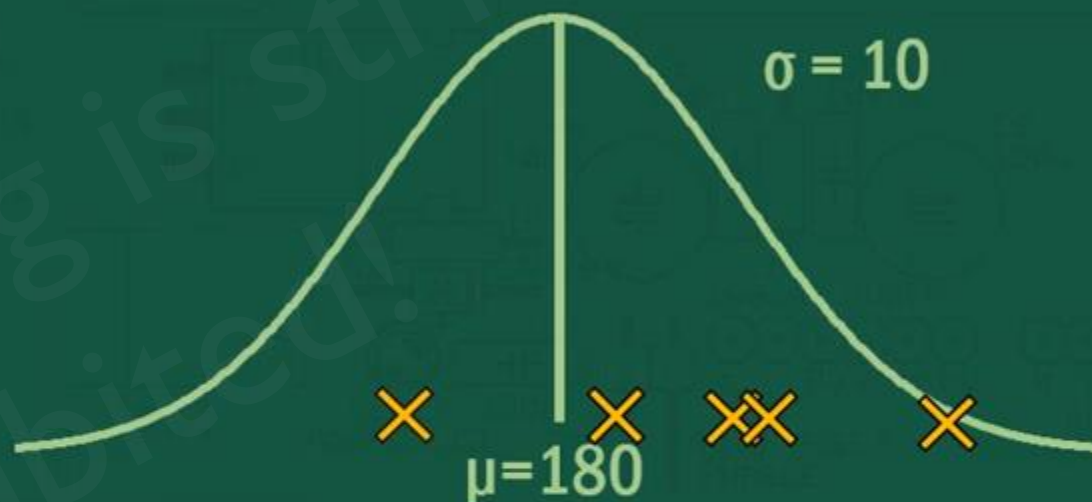
- * Take a sample of five observations from a normally distributed population

[183 , 170 , 189 , 191 , 203]

- * Find the average of that sample

$$\bar{X} = 187.2 \text{ cm}$$

Heights of female basketballers (cm)



t Distribution

SAMPLING RECAP!

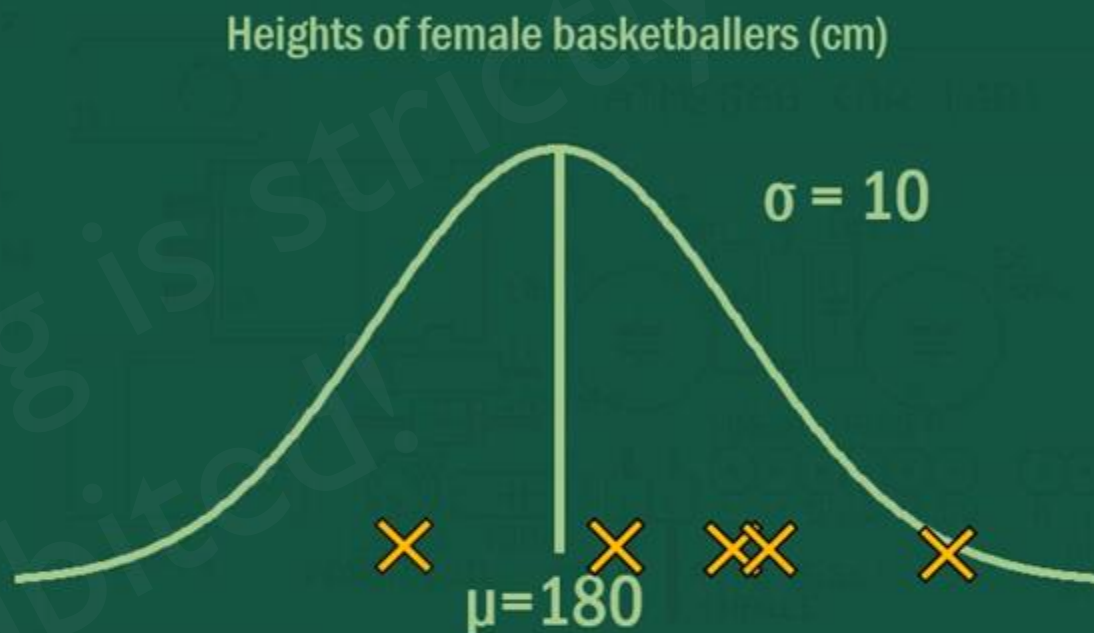
- * Take a sample of five observations from a normally distributed population

[183 , 170 , 189 , 191 , 203]

- * Find the average of that sample

$$\bar{X} = 187.2 \text{ cm}$$

- * How would such a sample mean (of size 5) be distributed?



t Distribution

SAMPLING RECAP!

- * Take a sample of five observations from a normally distributed population

[183 , 170 , 189 , 191 , 203]

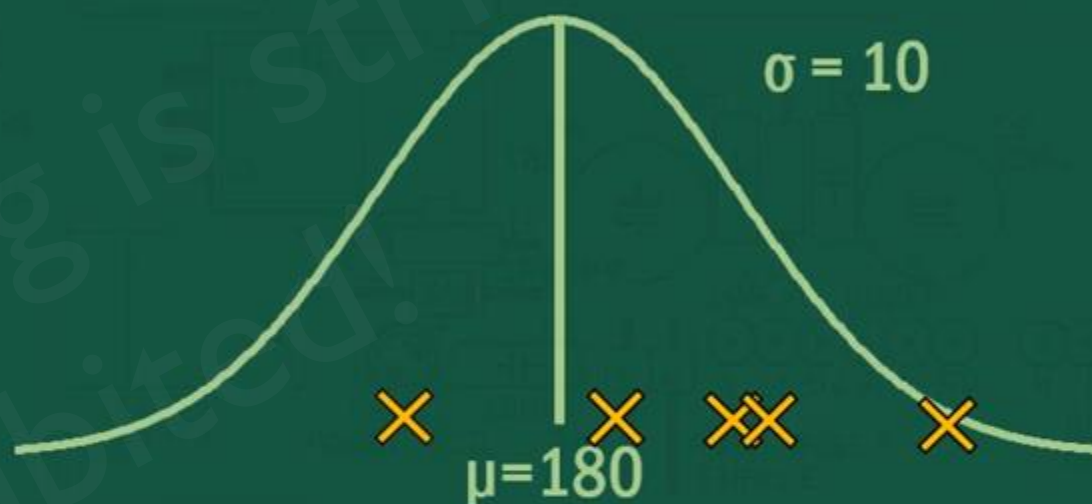
- * Find the average of that sample

$$\bar{X} = 187.2 \text{ cm}$$

- * How would such a sample mean (of size 5) be distributed?

$$\bar{x} \sim N \left(\mu, \left(\frac{\sigma}{\sqrt{n}} \right)^2 \right) \quad \bar{x} \sim N \left(180, \left(\frac{10}{\sqrt{5}} \right)^2 \right)$$

Heights of female basketballers (cm)



t Distribution

SAMPLING RECAP!

- * Imagine we are **TESTING** the population mean value of 180cm by using our sample
- * $H_0: \mu = 180$ $H_1: \mu \neq 180$

t Distribution

SAMPLING RECAP!

* Imagine we are **TESTING** the population mean value of 180cm by using our sample

[183 , 170 , 189 , 191 , 203]

$$\bar{X} = 187.2 \text{ cm}$$

* $H_0: \mu = 180$ $H_1: \mu \neq 180$

t Distribution

SAMPLING RECAP!

* Imagine we are **TESTING** the population mean value of 180cm by using our sample

[183 , 170 , 189 , 191 , 203]

$$\bar{X} = 187.2 \text{ cm}$$

* $H_0: \mu = 180$ $H_1: \mu \neq 180$

$$\frac{\bar{x} - \mu}{\sigma / \sqrt{n}} \sim Z$$

t Distribution

SAMPLING RECAP!

* Imagine we are **TESTING** the population mean value of 180cm by using our sample

[183 , 170 , 189 , 191 , 203]

$$\bar{X} = 187.2 \text{ cm}$$

* $H_0: \mu = 180$ $H_1: \mu \neq 180$

$$\frac{\bar{x} - \mu}{\sigma / \sqrt{n}} \sim Z \longrightarrow \frac{\bar{x} - \mu}{s / \sqrt{n}}$$

t Distribution

SAMPLING RECAP!

* Imagine we are **TESTING** the population mean value of 180cm by using our sample

[183 , 170 , 189 , 191 , 203]

$$\bar{X} = 187.2 \text{ cm}$$

$$s = 12.05 \text{ cm}$$

* $H_0: \mu = 180$ $H_1: \mu \neq 180$

$$\frac{\bar{x} - \mu}{\sigma / \sqrt{n}} \sim Z \longrightarrow \frac{\bar{x} - \mu}{s / \sqrt{n}} +$$

t Distribution

SAMPLING RECAP!

* Imagine we are **TESTING** the population mean value of 180cm by using our sample

[183 , 170 , 189 , 191 , 203]

$$\bar{X} = 187.2 \text{ cm}$$

$$s = 12.05 \text{ cm}$$

* $H_0: \mu = 180$ $H_1: \mu \neq 180$

$$\frac{\bar{x} - \mu}{\sigma / \sqrt{n}} \sim z \longrightarrow \frac{\bar{x} - \mu}{s / \sqrt{n}} \sim t$$

t Distribution

SAMPLING RECAP!

* Imagine we are **TESTING** the population mean value of 180cm by using our sample

[183 , 170 , 189 , 191 , 203]

$$\bar{X} = 187.2 \text{ cm}$$

$$s = 12.05 \text{ cm}$$

* $H_0: \mu = 180$ $H_1: \mu \neq 180$

$$\frac{\bar{x} - \mu}{\sigma / \sqrt{n}} \sim z \longrightarrow \frac{\bar{x} - \mu}{s / \sqrt{n}} \sim t_{n-1}$$

t Distribution

SAMPLING RECAP!

- * Imagine we are **TESTING** the population mean value of 180cm by using our sample

[183 , 170 , 189 , 191 , 203]

$$\bar{X} = 187.2 \text{ cm}$$

$$s = 12.05 \text{ cm}$$

- * $H_0: \mu = 180$ $H_1: \mu \neq 180$

$$\frac{\bar{x} - \mu}{\sigma / \sqrt{n}} \sim Z \longrightarrow \frac{\bar{x} - \mu}{s / \sqrt{n}} \sim t_{n-1}$$

- * We need to adjust for the additional uncertainty around **s**.

t Distribution

SAMPLING RECAP!

- * Imagine we are **TESTING** the population mean value of 180cm by using our sample

[183 , 170 , 189 , 191 , 203]

$$\bar{X} = 187.2 \text{ cm}$$

$$s = 12.05 \text{ cm}$$

- * $H_0: \mu = 180$ $H_1: \mu \neq 180$

$$\frac{\bar{x} - \mu}{\sigma / \sqrt{n}} \sim Z \longrightarrow \frac{\bar{x} - \mu}{s / \sqrt{n}} \sim t_{n-1}$$

- * We need to adjust for the additional uncertainty around **s**.
- * The smaller the sample size, the more uncertain we are.

t Distribution

SAMPLING RECAP!

- * Imagine we are **TESTING** the population mean value of 180cm by using our sample

[183 , 170 , 189 , 191 , 203]

$$\bar{X} = 187.2 \text{ cm}$$

$$s = 12.05 \text{ cm}$$

- * $H_0: \mu = 180$ $H_1: \mu \neq 180$

$$\frac{\bar{x} - \mu}{\sigma / \sqrt{n}} \sim Z \longrightarrow \frac{\bar{x} - \mu}{s / \sqrt{n}} \sim t_{n-1}$$

$$t = \frac{187.2 - 180}{12.05 / \sqrt{5}} \sim t_4$$

- * We need to adjust for the additional uncertainty around s .
- * The smaller the sample size, the more uncertain we are.

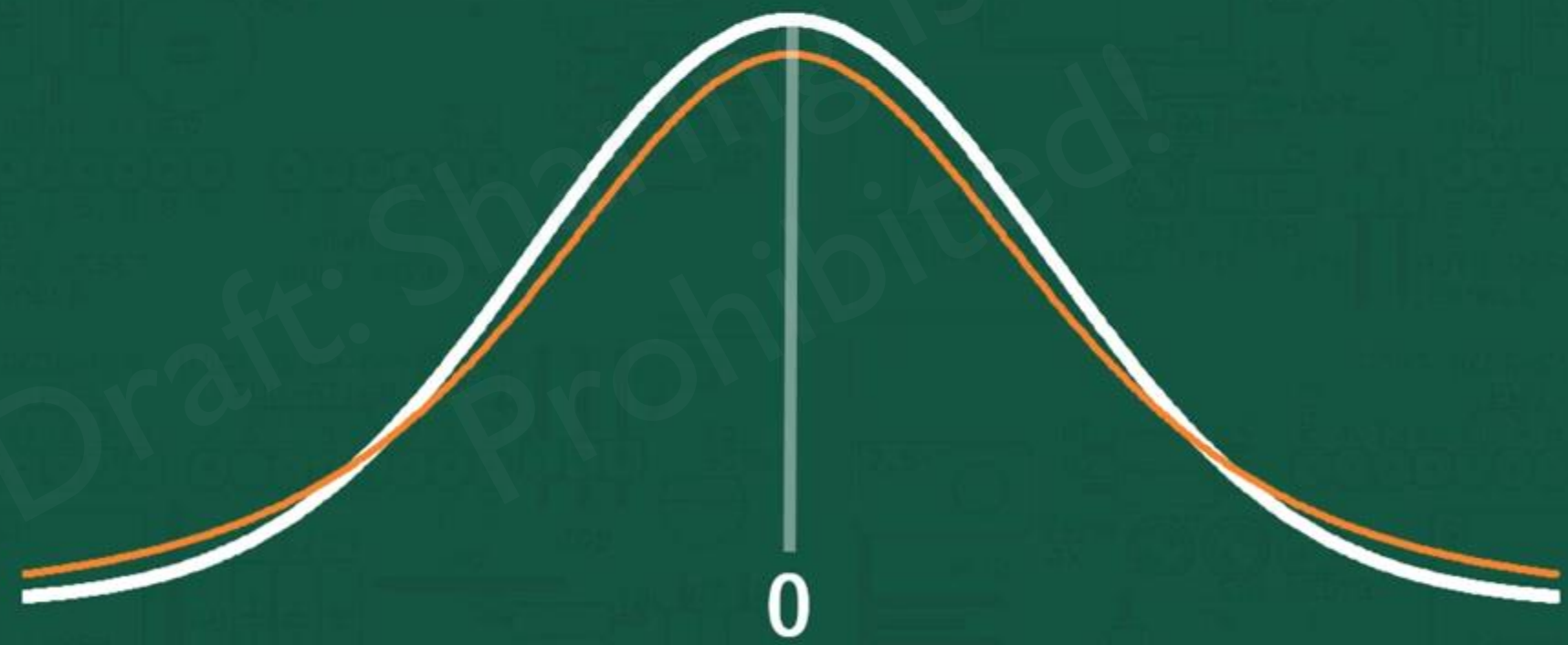
3. Visualisation

+

t Distribution

Probability Distribution Function

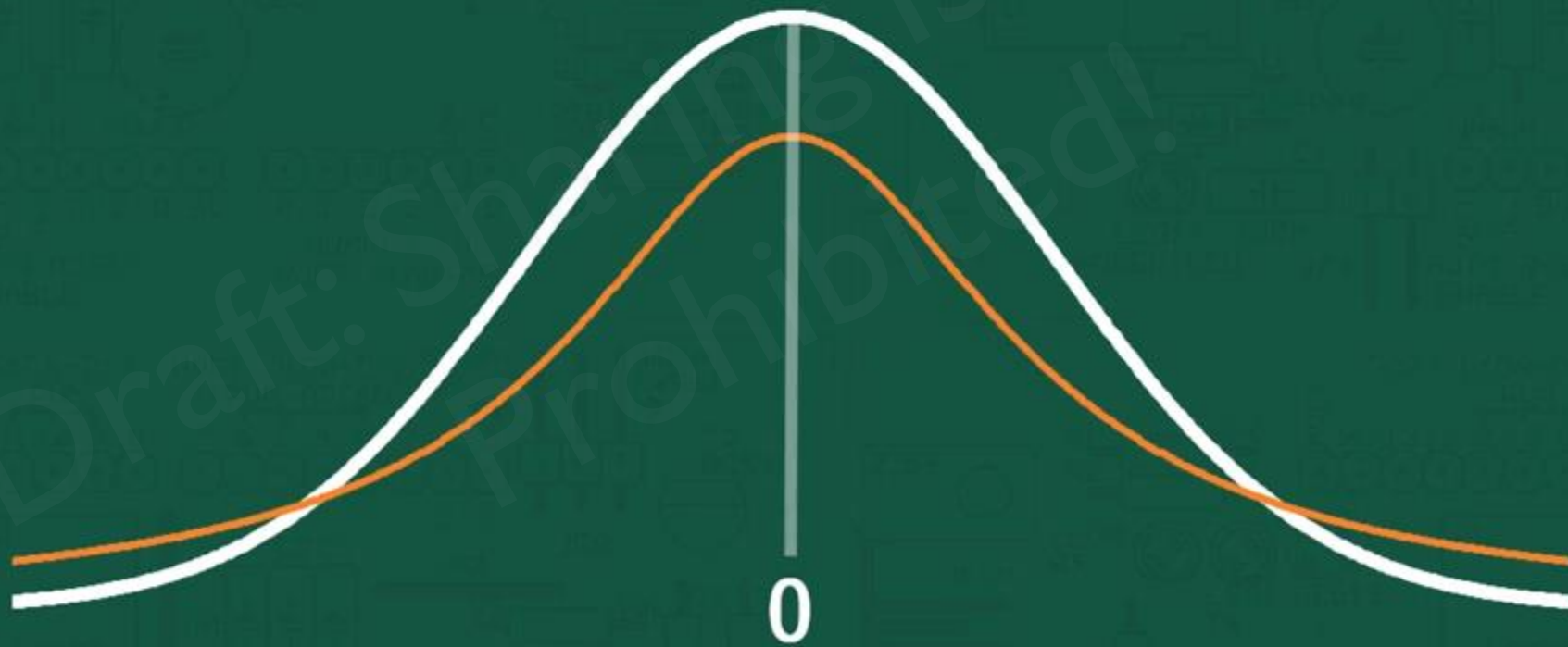
—Z —t (df=4)



t Distribution

Probability Distribution Function

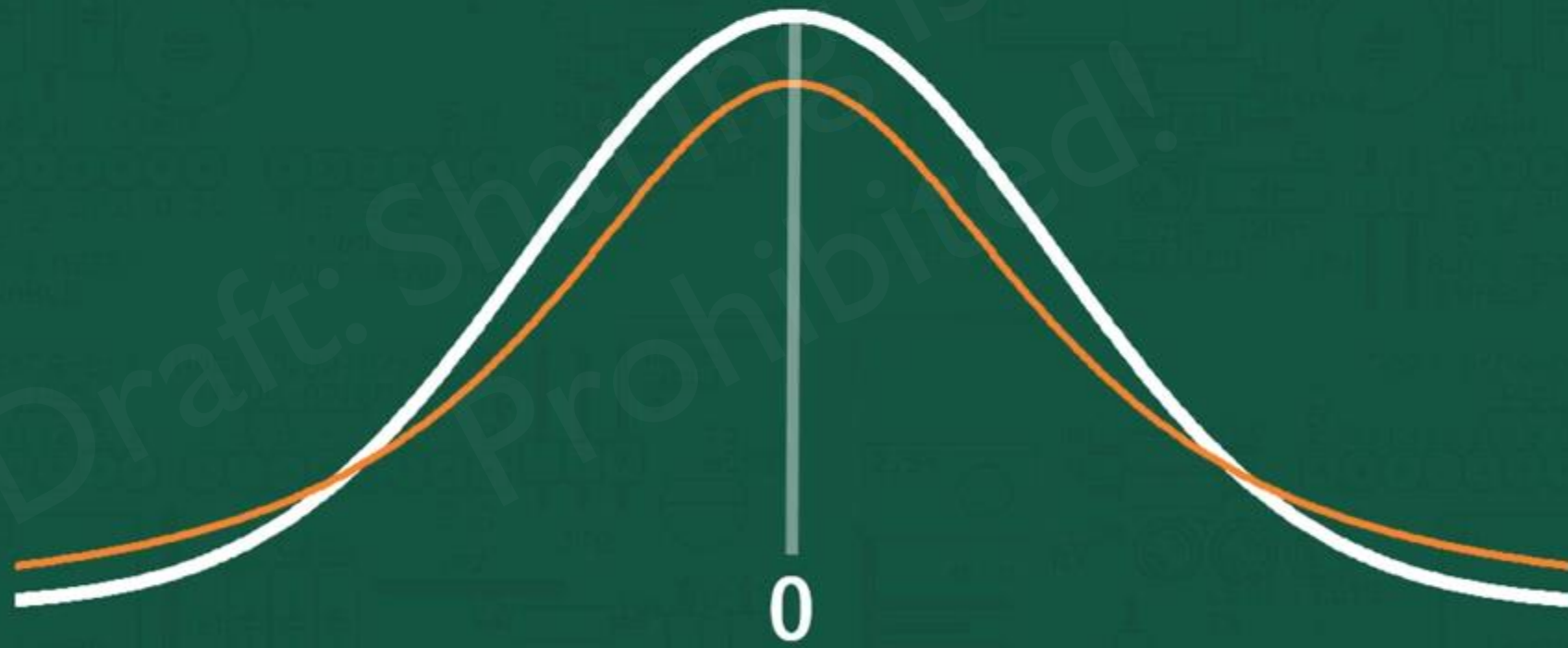
—Z —t (df=1)



t Distribution

Probability Distribution Function

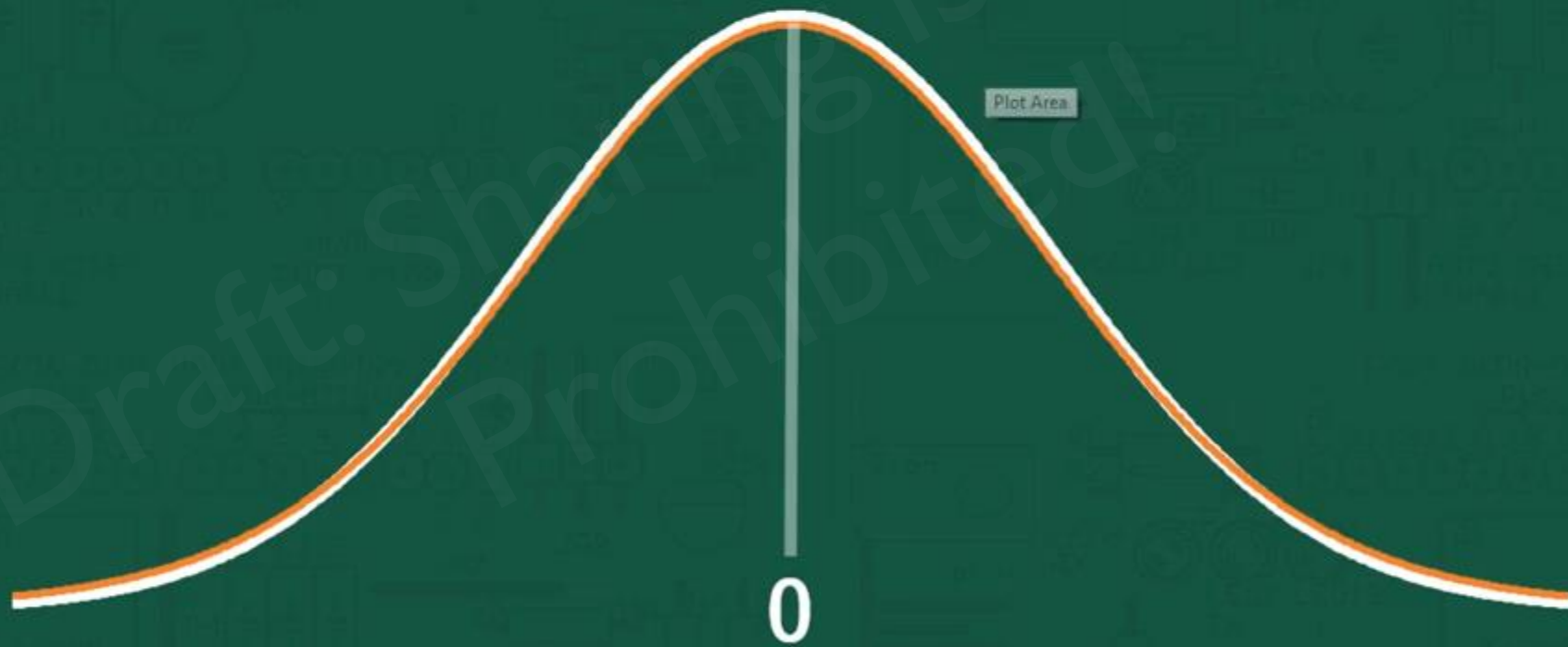
—Z —t (df=2)



t Distribution

Probability Distribution Function

—Z —t (df=20)



t Distribution

Cumulative Distribution Function

What proportion of the t-distribution (with 4 df) exists above $t=1.61$?

Draft: Sharing is strictly prohibited!

t Distribution

Cumulative Distribution Function

What proportion of the t-distribution (with 4 df) exists above $t=1.61$?

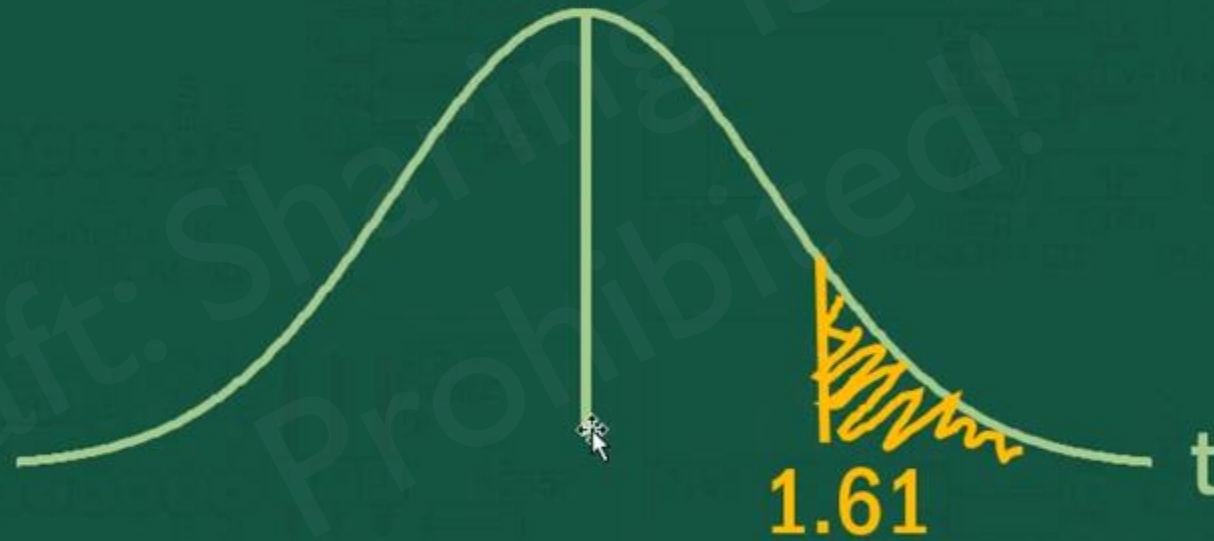


+

t Distribution

Cumulative Distribution Function

What proportion of the t-distribution (with 4 df) exists above $t=1.61$?



$$=1-T.DIST(1.61,4,TRUE)$$

t Distribution

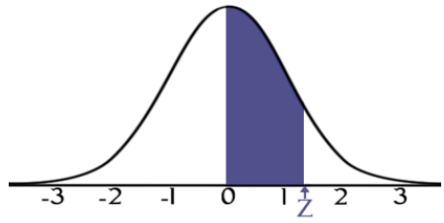
Cumulative Distribution Function

What proportion of the t-distribution (with 4 df) exists above $t=1.61$?



$$=1-T.DIST(1.61,4,TRUE)$$

$$=0.091$$



STANDARD NORMAL TABLE (Z)

Entries in the table give the area under the curve between the mean and z standard deviations above the mean. For example, for $z = 1.25$ the area under the curve between the mean (0) and z is 0.3944.

Z	0.00	0.01	0.02	0.03	0.04	0.05	0.06	0.07	0.08	0.09
0.0	0.0000	0.0040	0.0080	0.0120	0.0160	0.0190	0.0239	0.0279	0.0319	0.0359
0.1	0.0398	0.0438	0.0478	0.0517	0.0557	0.0596	0.0636	0.0675	0.0714	0.0753
0.2	0.0793	0.0832	0.0871	0.0910	0.0948	0.0987	0.1026	0.1064	0.1103	0.1141
0.3	0.1179	0.1217	0.1255	0.1293	0.1331	0.1368	0.1406	0.1443	0.1480	0.1517
0.4	0.1554	0.1591	0.1628	0.1664	0.1700	0.1736	0.1772	0.1808	0.1844	0.1879
0.5	0.1915	0.1950	0.1985	0.2019	0.2054	0.2088	0.2123	0.2157	0.2190	0.2224
0.6	0.2257	0.2291	0.2324	0.2357	0.2389	0.2422	0.2454	0.2486	0.2517	0.2549
0.7	0.2580	0.2611	0.2642	0.2673	0.2704	0.2734	0.2764	0.2794	0.2823	0.2852
0.8	0.2881	0.2910	0.2939	0.2969	0.2995	0.3023	0.3051	0.3078	0.3106	0.3133
0.9	0.3159	0.3186	0.3212	0.3238	0.3264	0.3289	0.3315	0.3340	0.3365	0.3389
1.0	0.3413	0.3438	0.3461	0.3485	0.3508	0.3513	0.3554	0.3577	0.3529	0.3621
1.1	0.3643	0.3665	0.3686	0.3708	0.3729	0.3749	0.3770	0.3790	0.3810	0.3830
1.2	0.3849	0.3869	0.3888	0.3907	0.3925	0.3944	0.3962	0.3980	0.3997	0.4015
1.3	0.4032	0.4049	0.4066	0.4082	0.4099	0.4115	0.4131	0.4147	0.4162	0.4177
1.4	0.4192	0.4207	0.4222	0.4236	0.4251	0.4265	0.4279	0.4292	0.4306	0.4319
1.5	0.4332	0.4345	0.4357	0.4370	0.4382	0.4394	0.4406	0.4418	0.4429	0.4441
1.6	0.4452	0.4463	0.4474	0.4484	0.4495	0.4505	0.4515	0.4525	0.4535	0.4545
1.7	0.4554	0.4564	0.4573	0.4582	0.4591	0.4599	0.4608	0.4616	0.4625	0.4633
1.8	0.4641	0.4649	0.4656	0.4664	0.4671	0.4678	0.4686	0.4693	0.4699	0.4706
1.9	0.4713	0.4719	0.4726	0.4732	0.4738	0.4744	0.4750	0.4756	0.4761	0.4767
2.0	0.4772	0.4778	0.4783	0.4788	0.4793	0.4798	0.4803	0.4808	0.4812	0.4817
2.1	0.4821	0.4826	0.4830	0.4834	0.4838	0.4842	0.4846	0.4850	0.4854	0.4857
2.2	0.4861	0.4864	0.4868	0.4871	0.4875	0.4878	0.4881	0.4884	0.4887	0.4890
2.3	0.4893	0.4896	0.4898	0.4901	0.4904	0.4906	0.4909	0.4911	0.4913	0.4916
2.4	0.4918	0.4920	0.4922	0.4925	0.4927	0.4929	0.4931	0.4932	0.4934	0.4936
2.5	0.4938	0.4940	0.4941	0.4943	0.4945	0.4946	0.4948	0.4949	0.4951	0.4952
2.6	0.4953	0.4955	0.4956	0.4957	0.4959	0.4960	0.4961	0.4962	0.4963	0.4964
2.7	0.4965	0.4966	0.4967	0.4968	0.4969	0.4970	0.4971	0.4972	0.4973	0.4974
2.8	0.4974	0.4975	0.4976	0.4977	0.4977	0.4978	0.4979	0.4979	0.4980	0.4981
2.9	0.4981	0.4982	0.4982	0.4983	0.4984	0.4984	0.4985	0.4985	0.4986	0.4986
3.0	0.4987	0.4987	0.4987	0.4988	0.4988	0.4989	0.4989	0.4989	0.4990	0.4990
3.1	0.4990	0.4991	0.4991	0.4991	0.4992	0.4992	0.4992	0.4992	0.4993	0.4993
3.2	0.4993	0.4993	0.4994	0.4994	0.4994	0.4994	0.4994	0.4995	0.4995	0.4995
3.3	0.4995	0.4995	0.4995	0.4996	0.4996	0.4996	0.4996	0.4996	0.4996	0.4997
3.4	0.4997	0.4997	0.4997	0.4997	0.4997	0.4997	0.4997	0.4997	0.4997	0.4998

Numbers in each row of the table are values on a t -distribution with (df) degrees of freedom for selected right-tail (greater-than) probabilities (p).



df/p	0.40	0.25	0.10	0.05	0.025	0.01	0.005	0.0005
1	0.324920	1.000000	3.077684	6.313752	12.70620	31.82052	63.65674	636.6192
2	0.288675	0.816497	1.885618	2.919986	4.30265	6.96456	9.92484	31.5991
3	0.276671	0.764892	1.637744	2.353363	3.18245	4.54070	5.84091	12.9240
4	0.270722	0.740697	1.533206	2.131847	2.77645	3.74695	4.60409	8.6103
5	0.267181	0.726687	1.475884	2.015048	2.57058	3.36493	4.03214	6.8688
6	0.264835	0.717558	1.439756	1.943180	2.44691	3.14267	3.70743	5.9588
7	0.263167	0.711142	1.414924	1.894579	2.36462	2.99795	3.49948	5.4079
8	0.261921	0.706387	1.396815	1.859548	2.30600	2.89646	3.35539	5.0413
9	0.260955	0.702722	1.383029	1.833113	2.26216	2.82144	3.24984	4.7809
10	0.260185	0.699812	1.372184	1.812461	2.22814	2.76377	3.16927	4.5869
11	0.259556	0.697445	1.363430	1.795885	2.20099	2.71808	3.10581	4.4370
12	0.259033	0.695483	1.356217	1.782288	2.17881	2.68100	3.05454	4.3178
13	0.258591	0.693829	1.350171	1.770933	2.16037	2.65031	3.01228	4.2208
14	0.258213	0.692417	1.345030	1.761310	2.14479	2.62449	2.97684	4.1405
15	0.257885	0.691197	1.340606	1.753050	2.13145	2.60248	2.94671	4.0728
16	0.257599	0.690132	1.336757	1.745884	2.11991	2.58349	2.92078	4.0150
17	0.257347	0.689195	1.333379	1.739607	2.10982	2.56693	2.89823	3.9651
18	0.257123	0.688364	1.330391	1.734064	2.10092	2.55238	2.87844	3.9216
19	0.256923	0.687621	1.327728	1.729133	2.09302	2.53948	2.86093	3.8834
20	0.256743	0.686954	1.325341	1.724718	2.08596	2.52798	2.84534	3.8495
21	0.256580	0.686352	1.323188	1.720743	2.07961	2.51765	2.83136	3.8193
22	0.256432	0.685805	1.321237	1.717144	2.07387	2.50832	2.81876	3.7921
23	0.256297	0.685306	1.319460	1.713872	2.06866	2.49987	2.80734	3.7676
24	0.256173	0.684850	1.317836	1.710882	2.06390	2.49216	2.79694	3.7454
25	0.256060	0.684430	1.316345	1.708141	2.05954	2.48511	2.78744	3.7251
26	0.255955	0.684043	1.314972	1.705618	2.05553	2.47863	2.77871	3.7066
27	0.255858	0.683685	1.313703	1.703288	2.05183	2.47266	2.77068	3.6896
28	0.255768	0.683353	1.312527	1.701131	2.04841	2.46714	2.76326	3.6739
29	0.255684	0.683044	1.311434	1.699127	2.04523	2.46202	2.75639	3.6594
30	0.255605	0.682756	1.310415	1.697261	2.04227	2.45726	2.75000	3.6460
z	0.253347	0.674490	1.281552	1.644854	1.95996	2.32635	2.57583	3.2905
CI	——	——	80%	90%	95%	98%	99%	99.9%

THANK YOU!