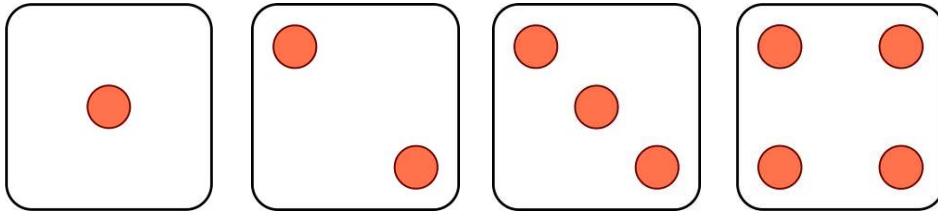
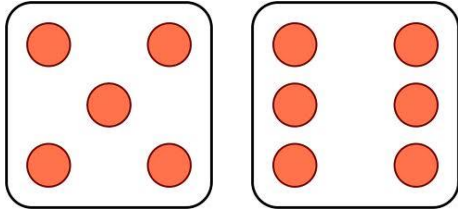


RANDOM VARIABLES

HOW TO MODEL A SIMPLE GAME



I get \$5 from you



You get \$10 from me

RANDOM VARIABLES

Definition

A **random variable** X on a sample space Ω is a real-valued function on Ω ; that is, $X : \Omega \rightarrow \mathcal{R}$. A **discrete random variable** is a random variable that takes on only a finite or countably infinite number of values.

Discrete random variable X and real value a : the event “ $X = a$ ” represents the set $\{s \in \Omega : X(s) = a\}$.

$$\Pr(X = a) = \sum_{s \in \Omega: X(s)=a} \Pr(s)$$

INDEPENDENCE

Definition

Two random variables X and Y are **independent** if and only if

$$\Pr((X = x) \cap (Y = y)) = \Pr(X = x) \cdot \Pr(Y = y)$$

for all values x and y . Similarly, random variables X_1, X_2, \dots, X_k are mutually independent if and only if for **any** subset $I \subseteq [1, k]$ and any values $x_i, i \in I$,

$$\Pr\left(\bigcap_{i \in I} X_i = x_i\right) = \prod_{i \in I} \Pr(X_i = x_i).$$

EXPECTATION

Definition

The **expectation** of a discrete random variable X , denoted by $\mathbf{E}[X]$, is given by

$$\mathbf{E}[X] = \sum_i i \Pr(X = i), \quad m$$

where the summation is over all values in the range of X . The expectation is finite if $\sum_i |i| \Pr(X = i)$ converges; otherwise, the expectation is unbounded.

The expectation (or mean or average) is a weighted sum over all possible values of the random variable.

LINEARITY OF EXPECTATION

Theorem

For any two random variables X and Y

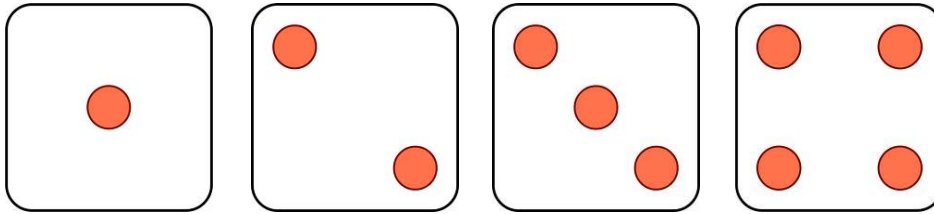
$$E[X + Y] = E[X] + E[Y].$$

Lemma

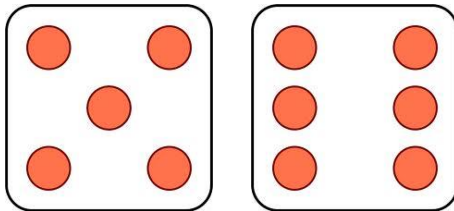
For any constant c and discrete random variable X ,

$$E[cX] = cE[X].$$

EXAMPLE: A SIMPLE GAME



I get \$5 from you



You get \$10 from me

$$P(X=-\$5) = 4/6 = 2/3$$

$$P(X=\$10) = 2/6 = 1/3$$

$$E(X) = -\$5 * 2/3 + \$10 * 1/3 = \$0$$

Would you play this game?

VARIANCE

Definition

The **variance** of a random variable X is

$$\text{Var}[X] = \mathbf{E}[(X - \mathbf{E}[X])^2] = \mathbf{E}[X^2] - (\mathbf{E}[X])^2.$$

Definition

The **standard deviation** of a random variable X is

$$\sigma(X) = \sqrt{\text{Var}[X]}.$$

VARIANCE

Theorem

If X and Y are independent random variables

$$\text{Var}[X + Y] = \text{Var}[X] + \text{Var}[Y].$$

Theorem

For any random variable X and constant a ,

$$\text{Var}[aX] = a^2 \text{Var}[X].$$

CLICKER QUESTION

You have two dice: **X and **Y****

What is the likelihood that the sum of both dice is **6**

a) $1/6$

b) $5/36$

c) $1/36$

EXAMPLE: ROLLING 2 DICE

You have two dice: **X** and **Y**

What is the likelihood that the sum of both dice is **6**?

| | | Die X | | | | | |
|-------|---|-------|---|---|----|----|----|
| | | 1 | 2 | 3 | 4 | 5 | 6 |
| Die Y | 1 | 2 | 3 | 4 | 5 | 6 | 7 |
| | 2 | 3 | 4 | 5 | 6 | 7 | 8 |
| | 3 | 4 | 5 | 6 | 7 | 8 | 9 |
| | 4 | 5 | 6 | 7 | 8 | 9 | 10 |
| | 5 | 6 | 7 | 8 | 9 | 10 | 11 |
| | 6 | 7 | 8 | 9 | 10 | 11 | 12 |

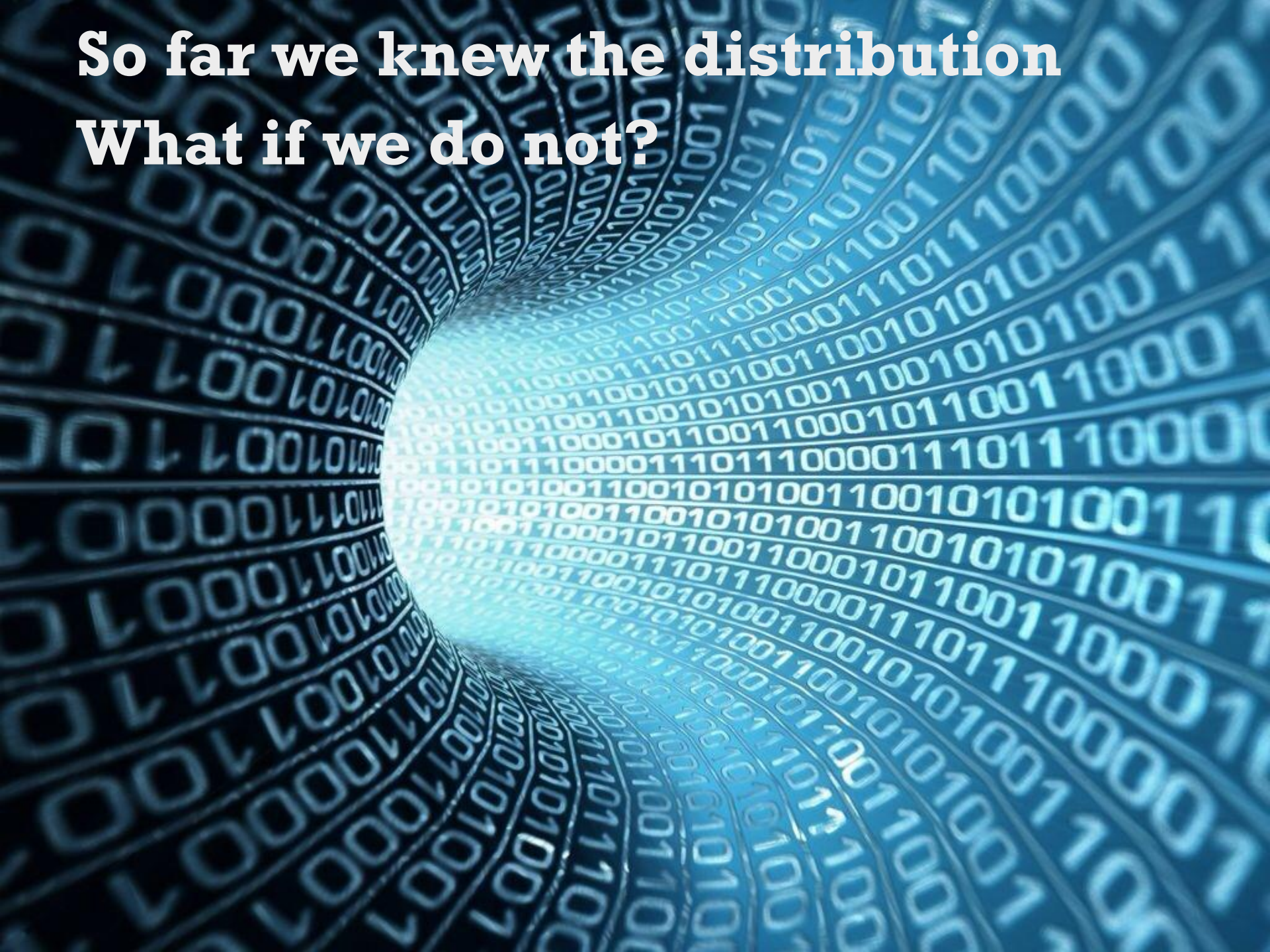
$$Z = X + Y$$

$$P(Z) = P(X + Y = n) = \sum_{m=-\infty}^{\infty} P(X = m) P(Y = n - m)$$

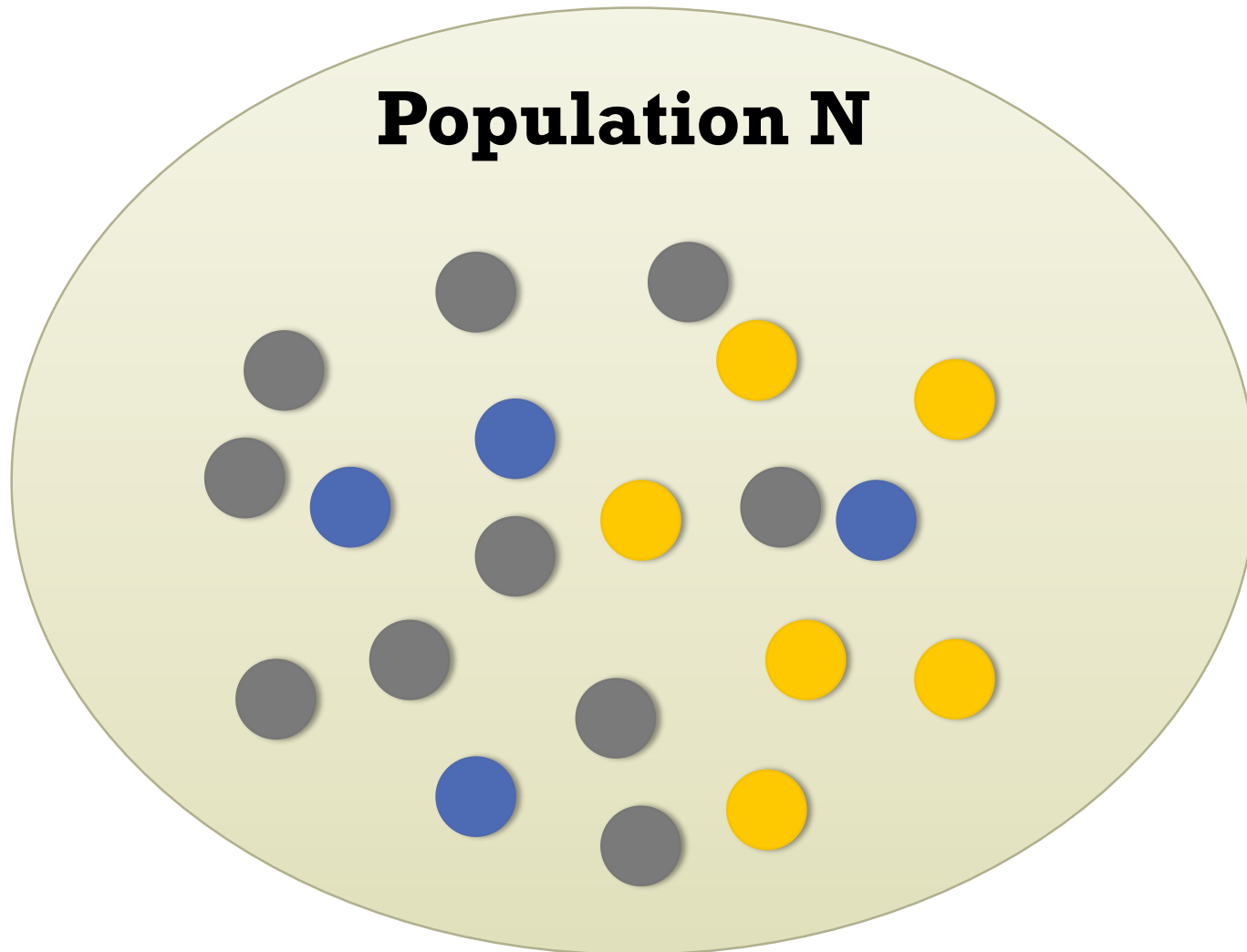
5/36

$$h(n) = f * g(n) = \sum_{m=-\infty}^{\infty} f(m) g(n - m)$$

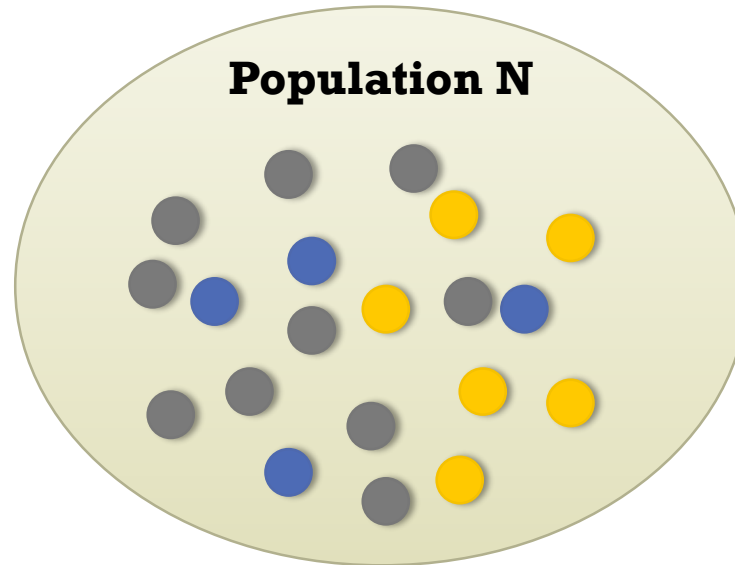
So far we knew the distribution
What if we do not?



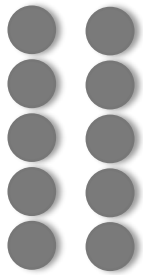
EXAMPLE: GRAY/BLUE/YELLOW LOTTERY



EMPIRICAL PROBABILITY



$$f_i = \frac{n_i}{N} = \frac{n_i}{\sum_i n_i}$$



$$f_{gray} = \frac{10}{20}$$

$$f_{blue} = \frac{4}{20}$$

$$f_{yellow} = \frac{6}{20}$$

POPULATION

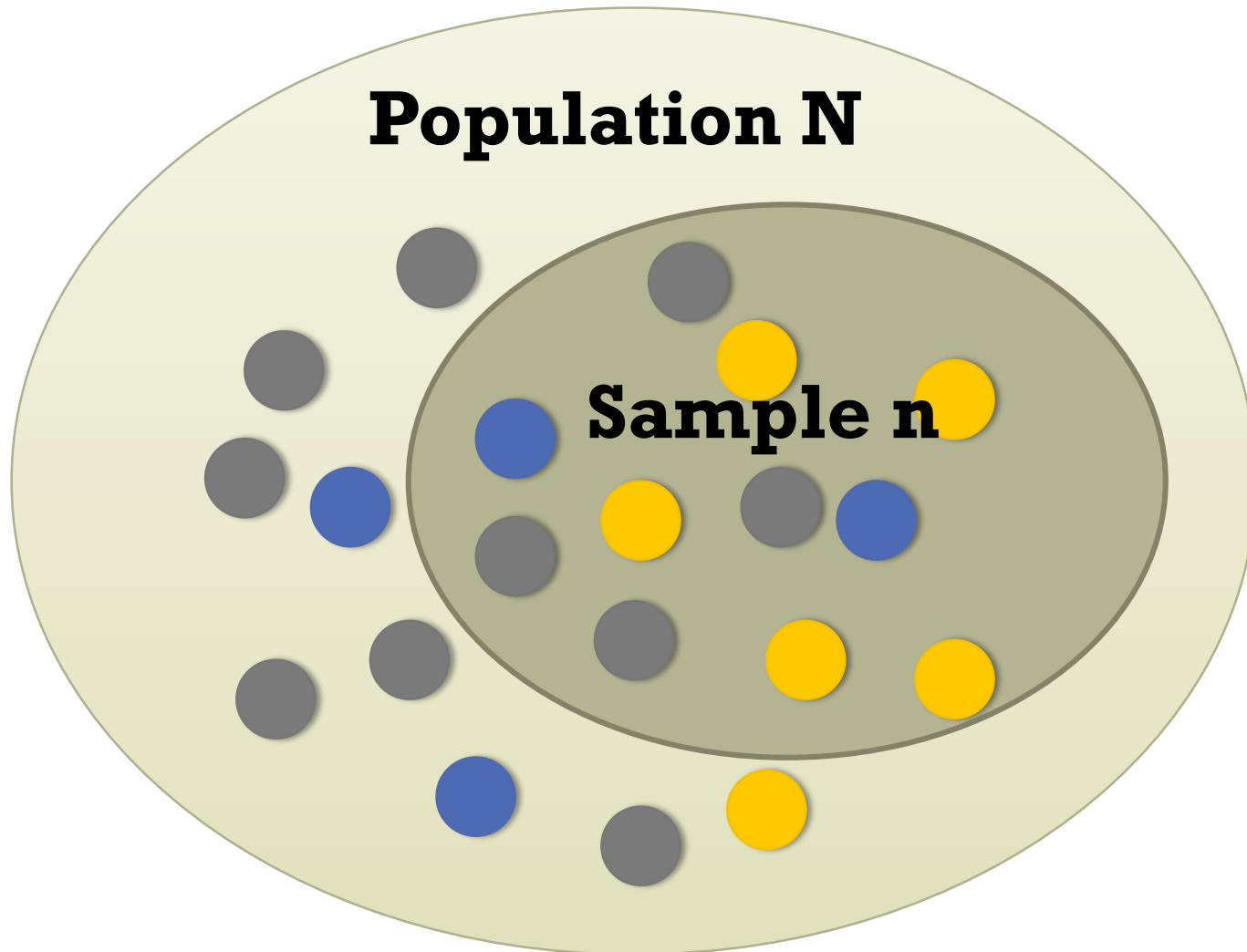
Mean

$$m = \frac{\sum x_i}{N}$$

Variance

$$s^2 = \frac{\sum (x_i - m)^2}{N}$$

EXAMPLE: GRAY/BLUE/YELLOW LOTTERY



POPULATION VS. SAMPLE

Sample (Statistic)
→ Estimates

Population (parameter)

Mean

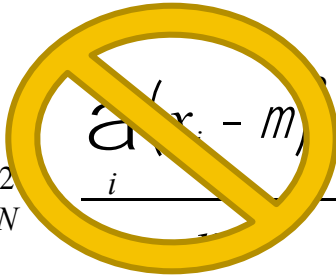
$$m = \frac{\sum x_i}{N}$$

$$\bar{x} = \frac{\sum x_i}{n}$$

Variance

$$s^2 = \frac{\sum (x_i - m)^2}{N}$$

Biased
Estimate

$$s_N^2 = \frac{\sum (x_i - m)^2}{n}$$


Un-Biased
Estimate

$$s_{N-1}^2 = \frac{\sum (x_i - m)^2}{n - 1}$$

BIG DATA

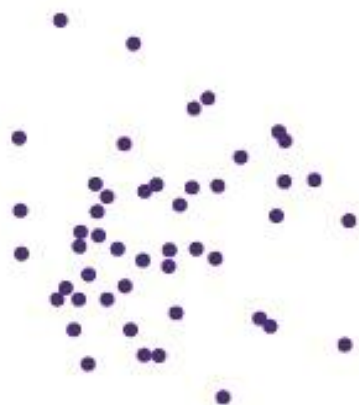
How to calculate the Variance in 1-Pass over large data:

$$S_{N-1}^2 = \frac{\sum_i (x_i - m)^2}{n - 1}$$

PEARSON CORRELATION

$$r = \frac{\sum (X - \bar{X})(Y - \bar{Y})}{\sqrt{\sum (X - \bar{X})^2} \sqrt{\sum (Y - \bar{Y})^2}}$$

$$r_{xy} = \frac{n \sum x_i y_i - \sum x_i \sum y_i}{\sqrt{n \sum x_i^2 - (\sum x_i)^2} \sqrt{n \sum y_i^2 - (\sum y_i)^2}}$$



Correlation $r = 0$



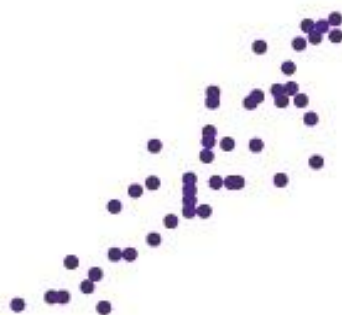
Correlation $r = -0.3$



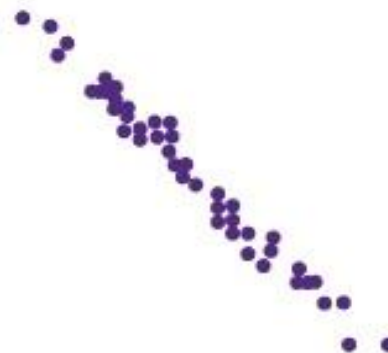
Correlation $r = 0.5$



Correlation $r = -0.7$



Correlation $r = 0.9$



Correlation $r = -0.99$

LAW OF LARGE NUMBERS

LAW OF LARGE NUMBERS

- Draw independent observations at random from any population with finite mean μ .
- As the number of observations increases, the sample mean approaches mean μ of the population.
- The more variation in the outcomes, the more trials are needed to ensure that is close to μ .

WEAK LAW OF LARGE NUMBERS

Theorem

Let X_1, \dots, X_n be independent, identically distributed, random variables. Let $\bar{X}_n = \frac{1}{n} \sum_{i=1}^n X_i$. For any constant $\varepsilon > 0$,

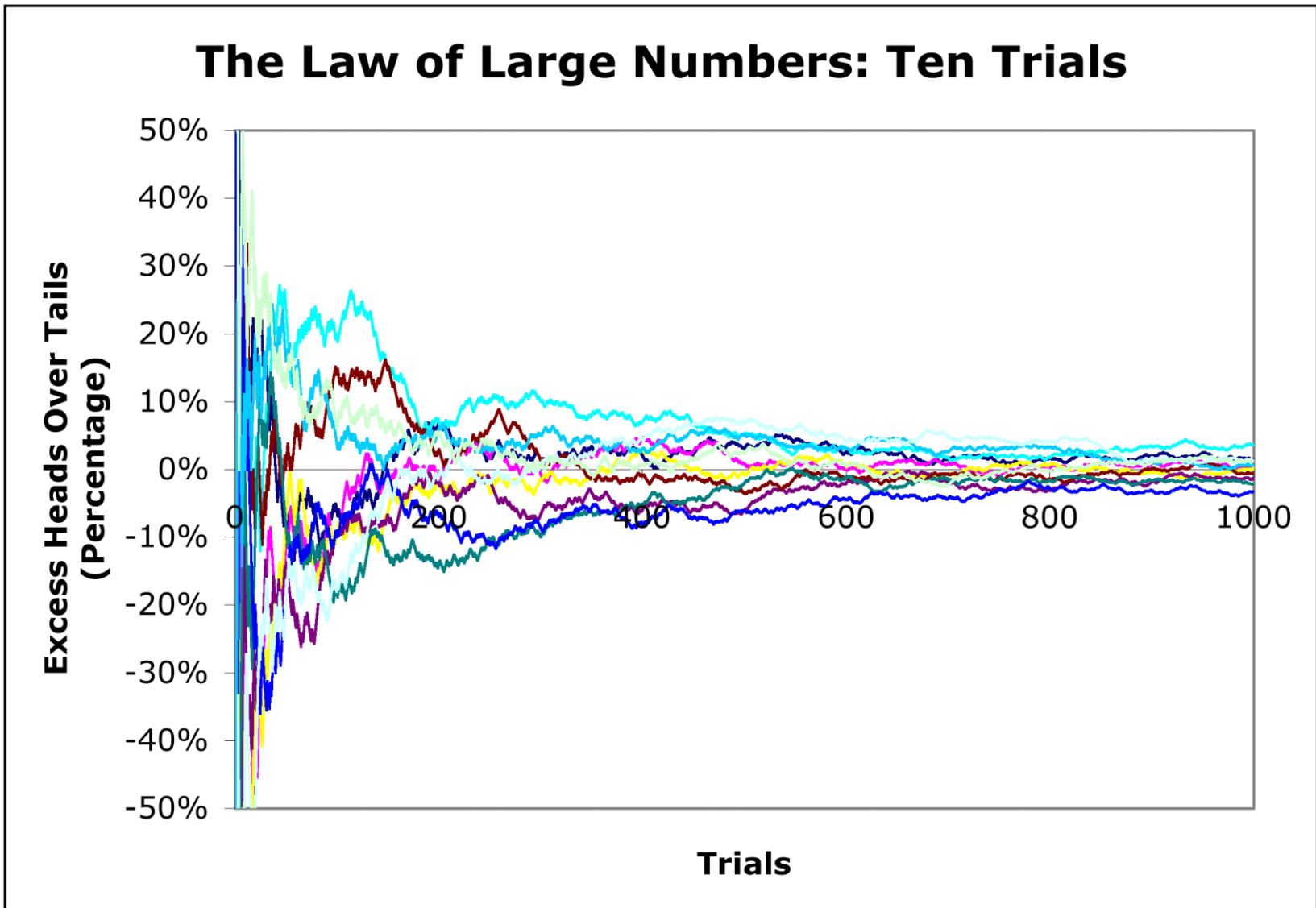
$$\lim_{n \rightarrow \infty} \Pr(|\bar{X}_n - \mathbf{E}[X]| \leq \varepsilon) = 1.$$

Strong law of large numbers

$$\bar{X} \rightarrow m$$

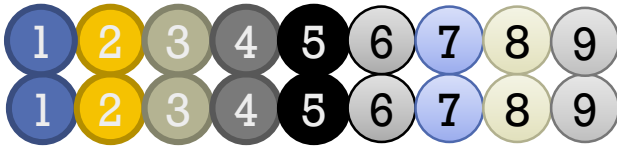
$$\Pr\left(\lim_{n \rightarrow \infty} \bar{X}_n = m\right) = 1$$

EXAMPLE: REPEATED COIN FLIPS



CENTRAL LIMIT THEOREM

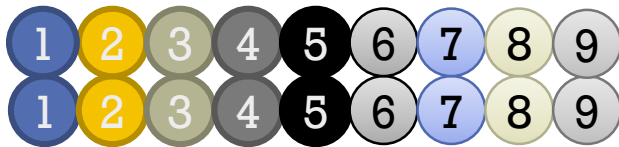
WHAT DOES IT MEAN?



$$m = 5$$

$$S^2 \gg 6.66$$

WHAT DOES IT MEAN?



$$m = 5$$

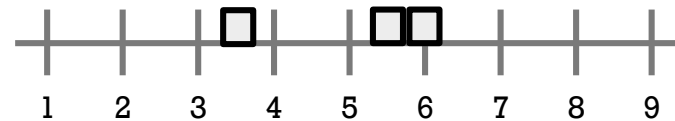
$$S^2 \gg 6.66$$

$$S_2 = \frac{X_1 + X_2}{n}$$

$$(\text{4}, \text{8}) = 6$$

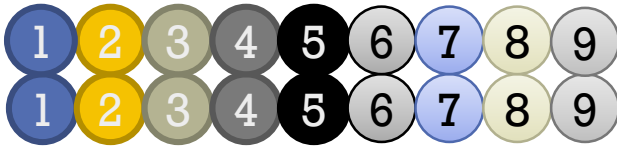
$$(\text{2}, \text{9}) = 5.5$$

$$(\text{3}, \text{4}) = 3.5$$



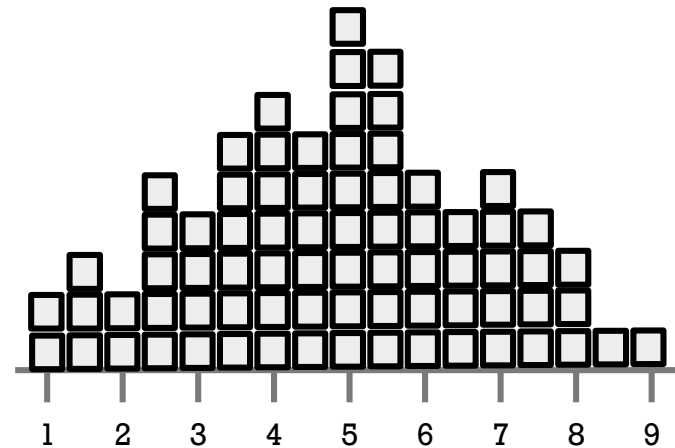
WHAT DOES IT MEAN?

$$S_2 = \frac{X_1 + X_2}{n}$$



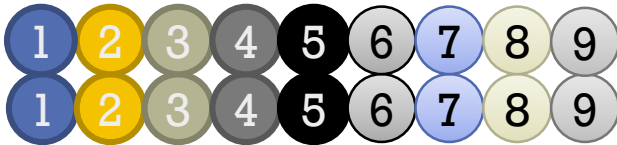
$$m = 5$$

$$S^2 \gg 6.66$$



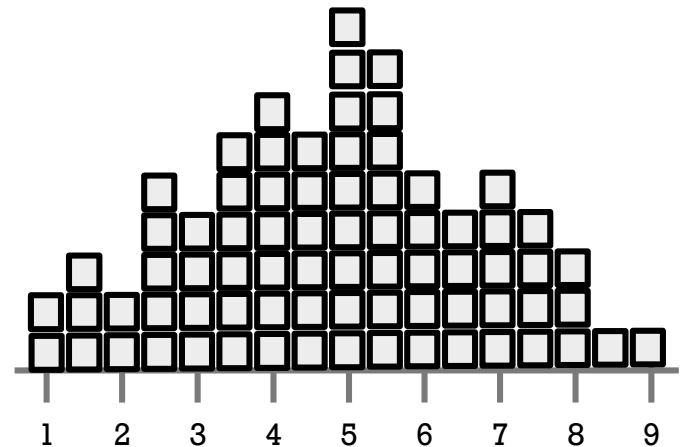
WHAT DOES IT MEAN?

$$S_2 = \frac{X_1 + X_2}{n}$$



$$m = 5$$

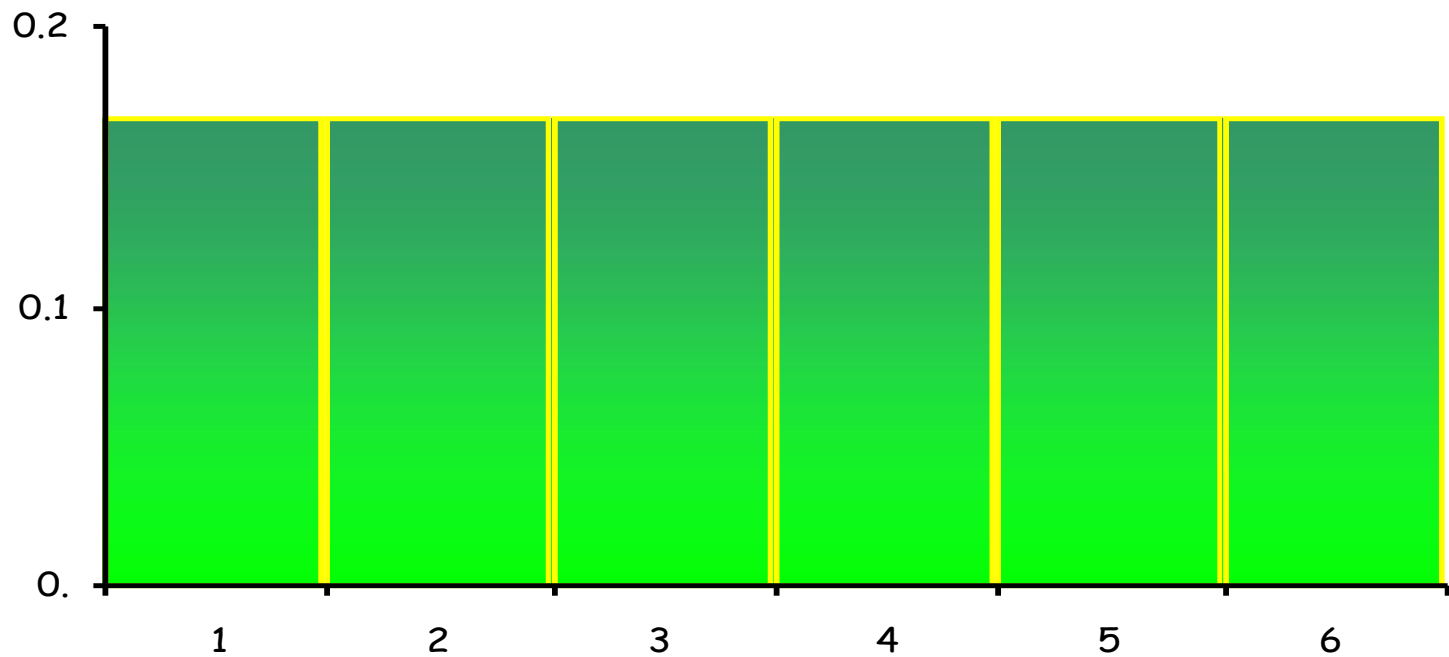
$$S^2 \gg 6.66$$



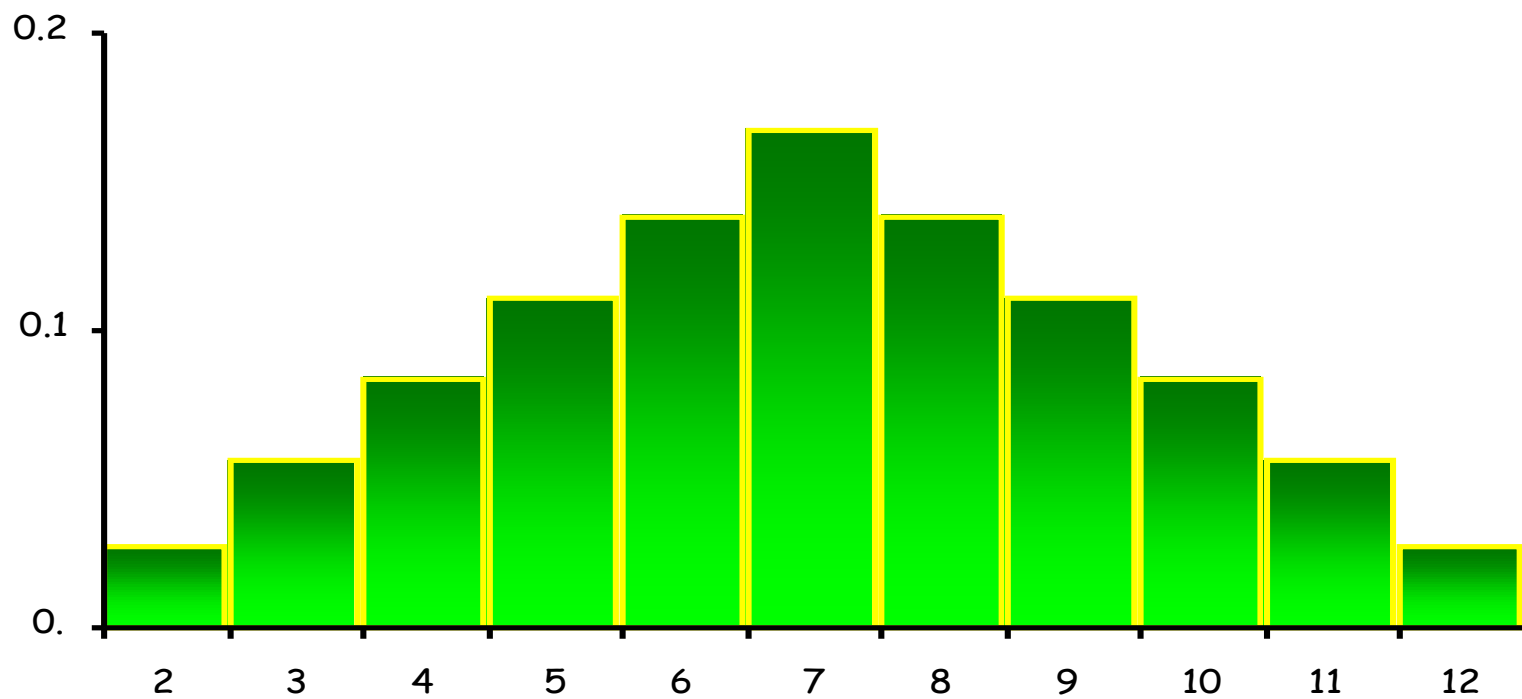
For large n :

$$S_n = \frac{X_1 + \dots + X_n}{n} \approx N\left(m, \frac{S^2}{n}\right)$$

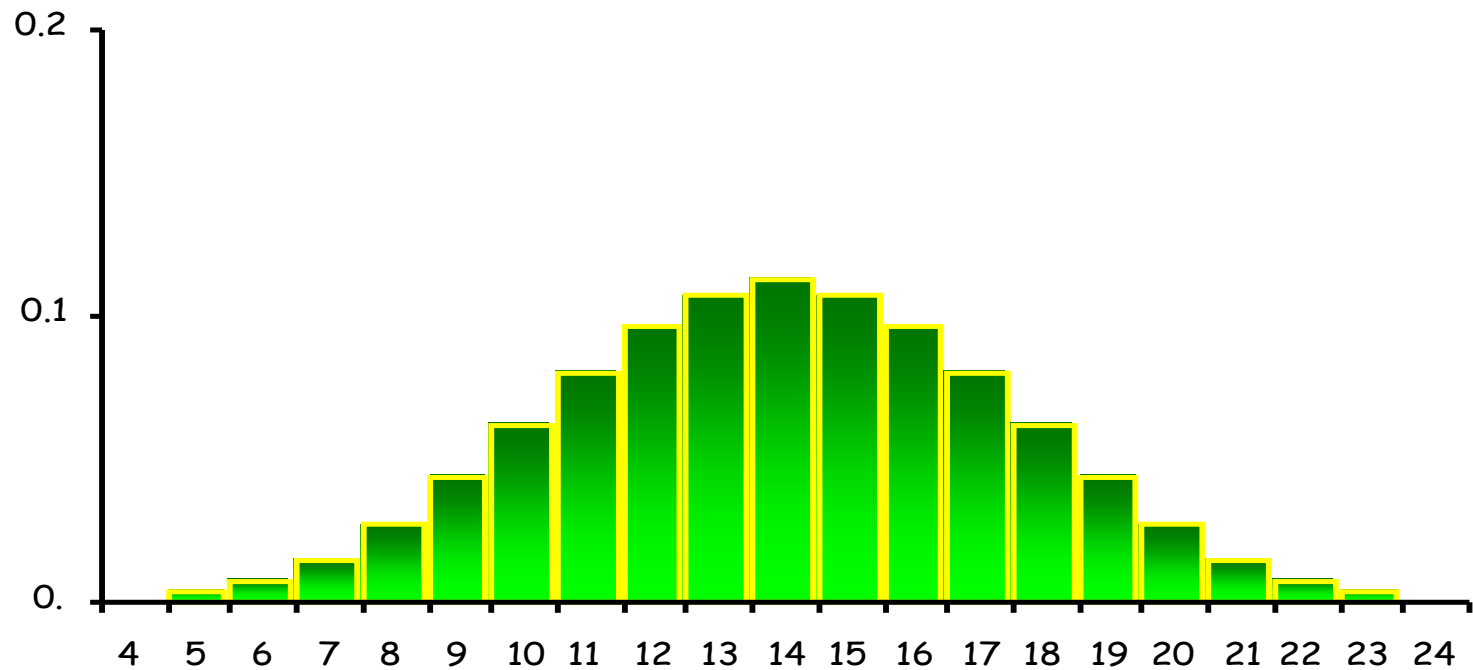
DISTRIBUTION OF X_1 : DIE 1 OR DIE 2



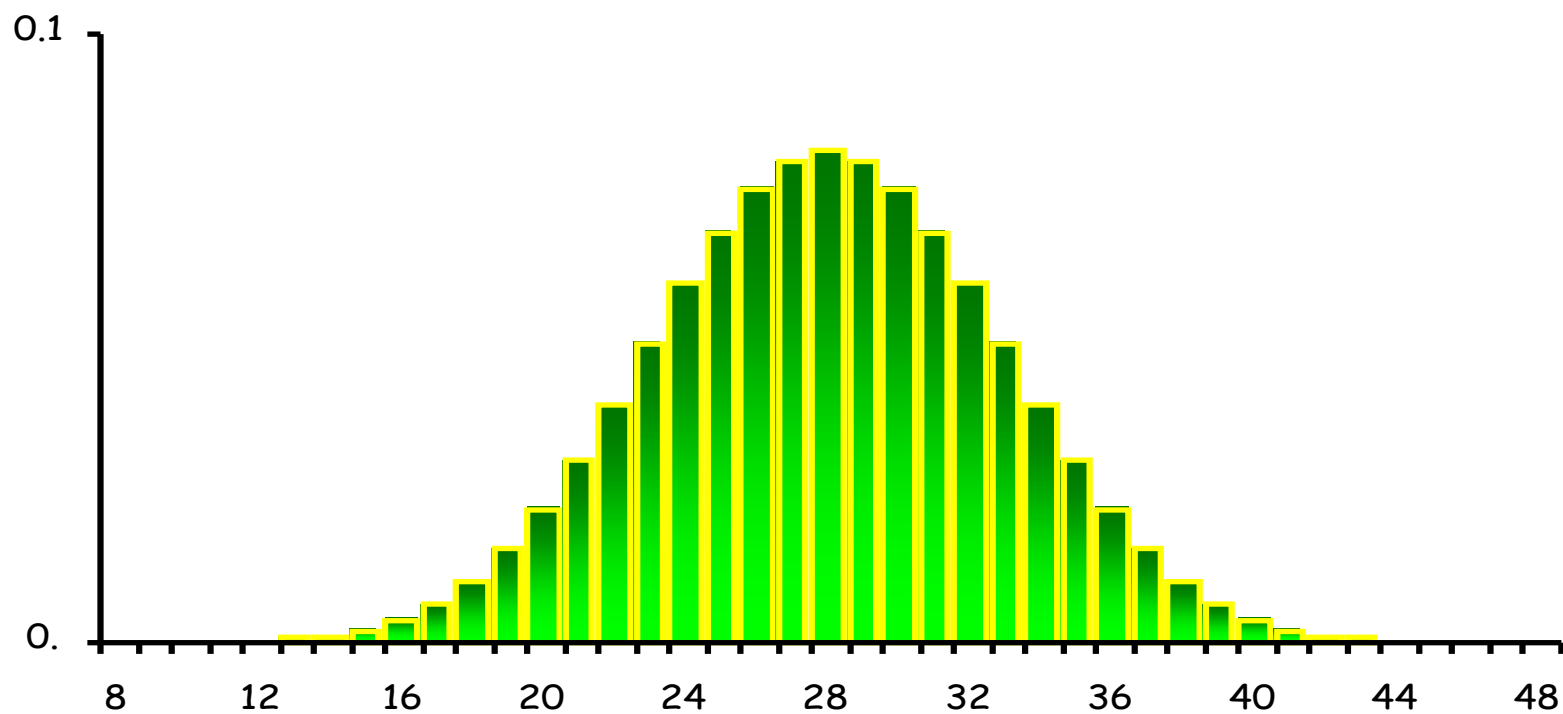
DISTRIBUTION OF S_1 : 2 DICE



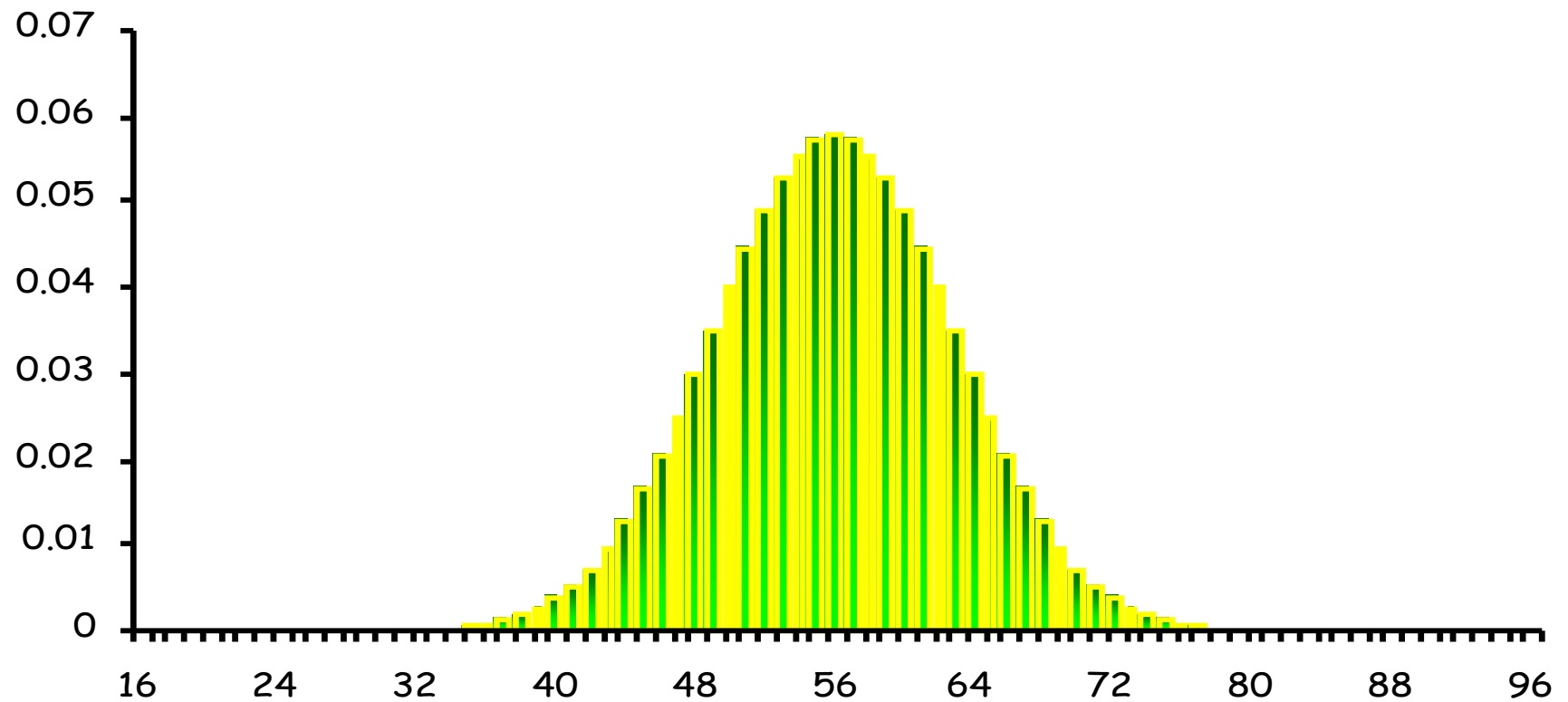
DISTRIBUTION OF S_4



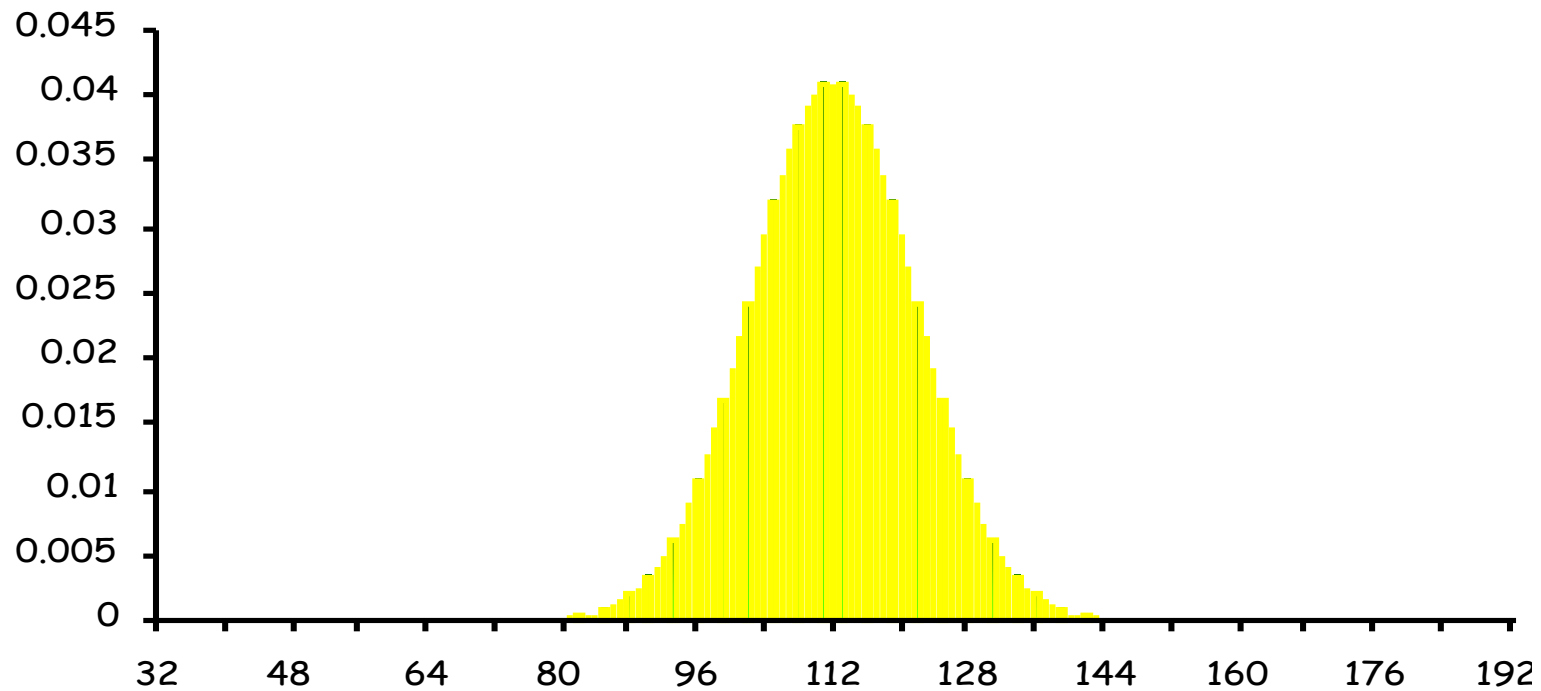
DISTRIBUTION OF S_8



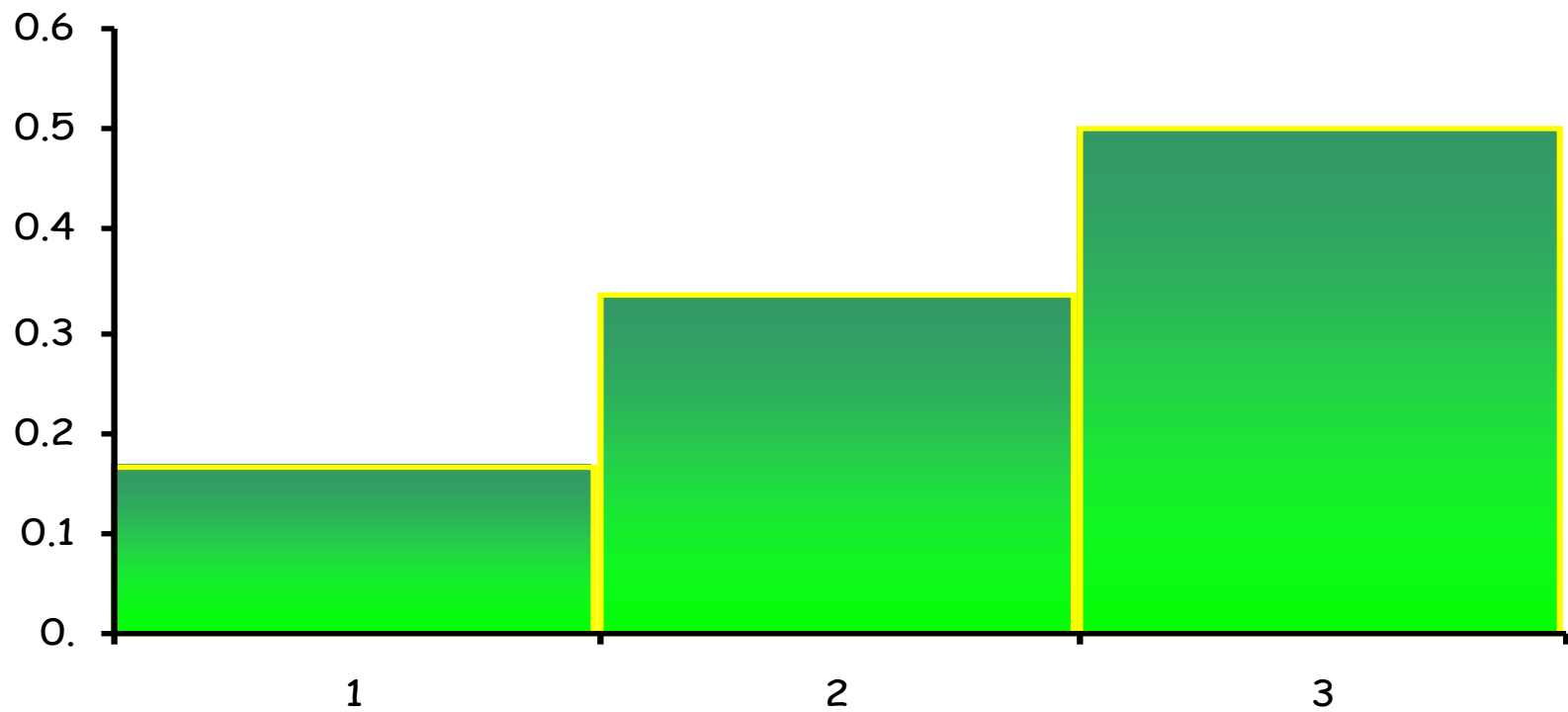
DISTRIBUTION OF S_{16}



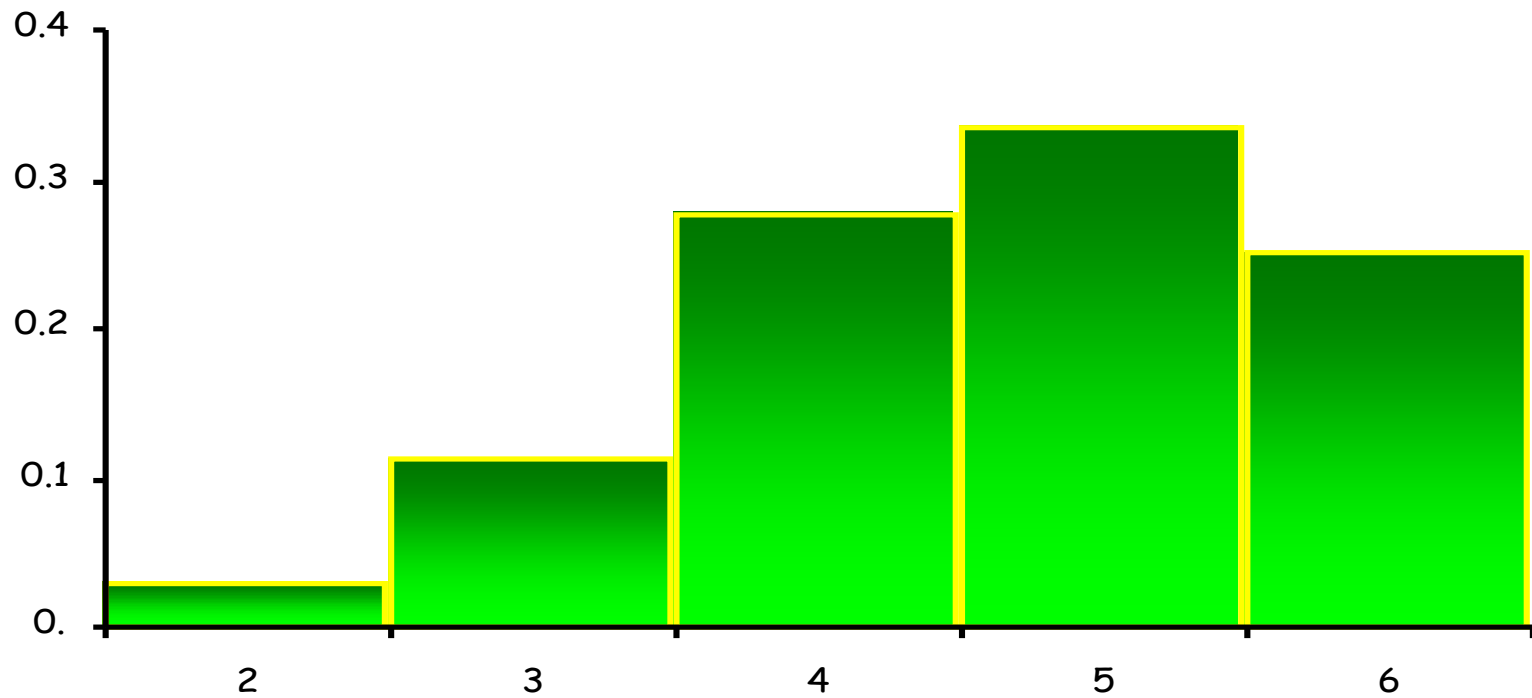
DISTRIBUTION OF S_{32}



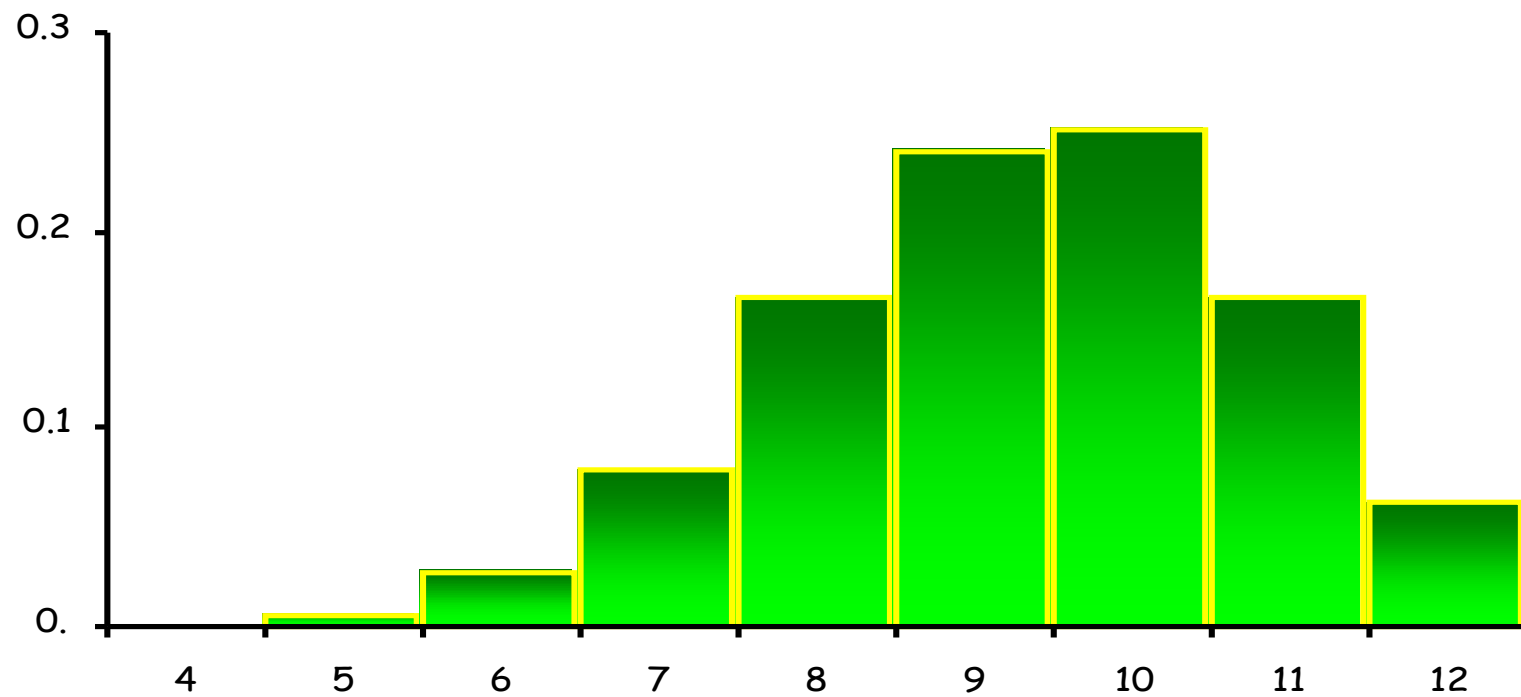
DISTRIBUTION OF X_1



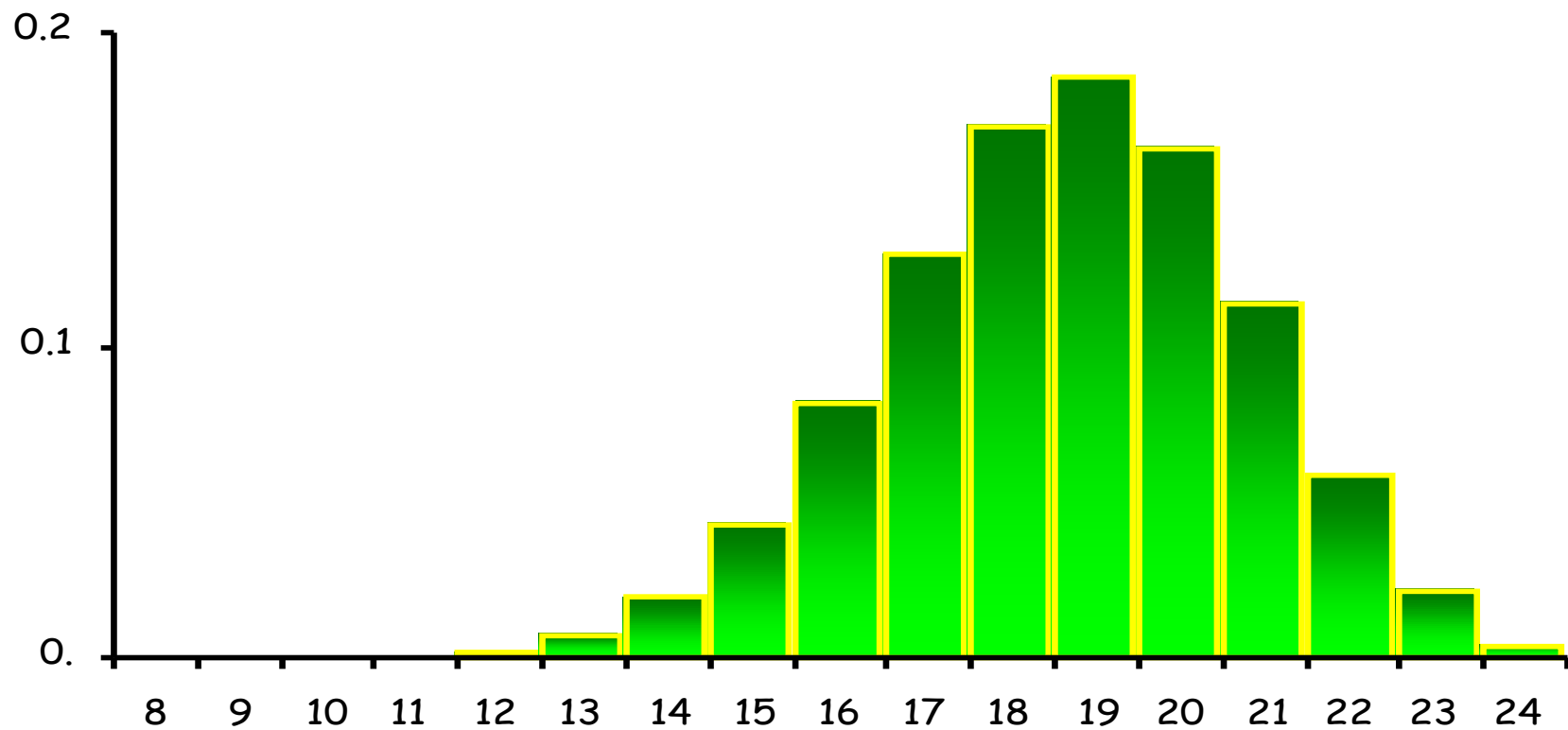
DISTRIBUTION OF S_2



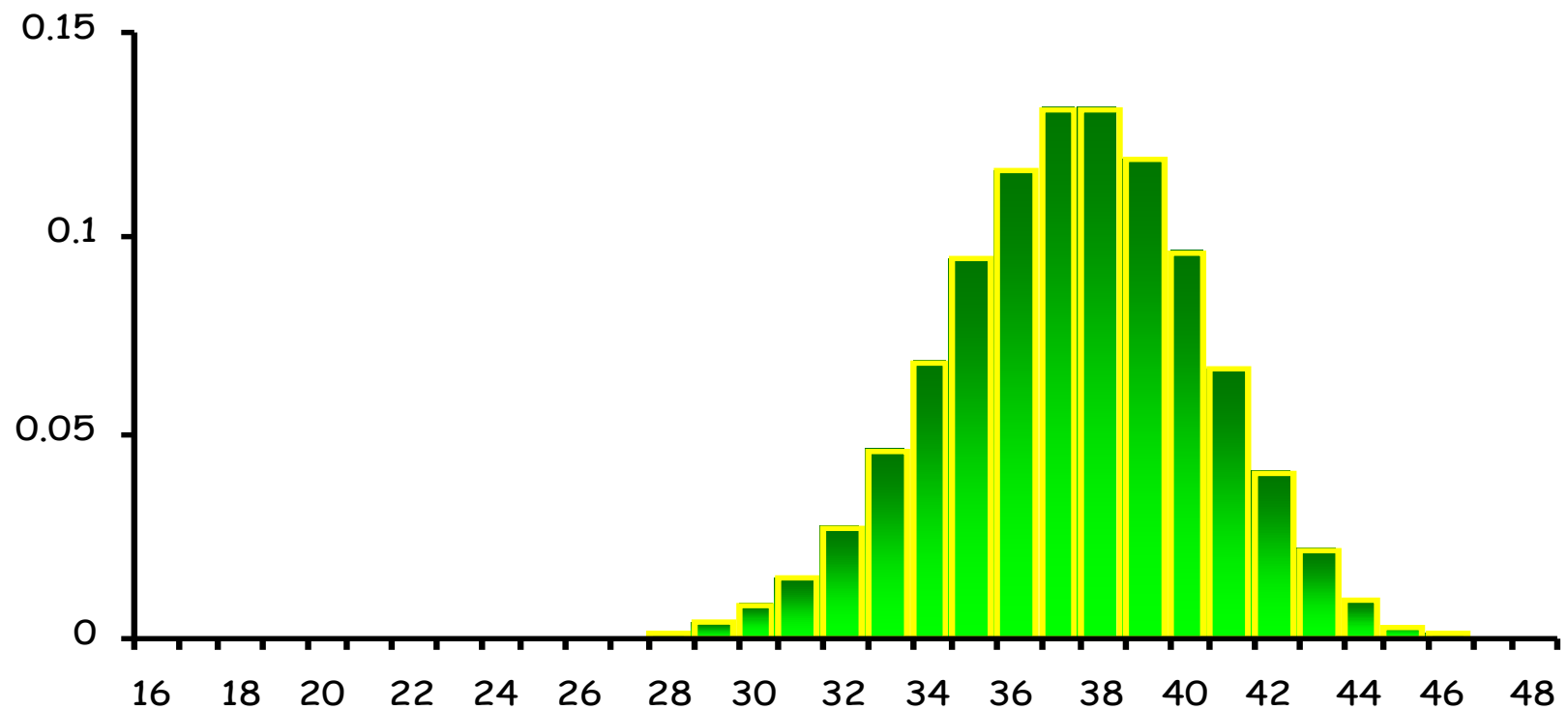
DISTRIBUTION OF S_4



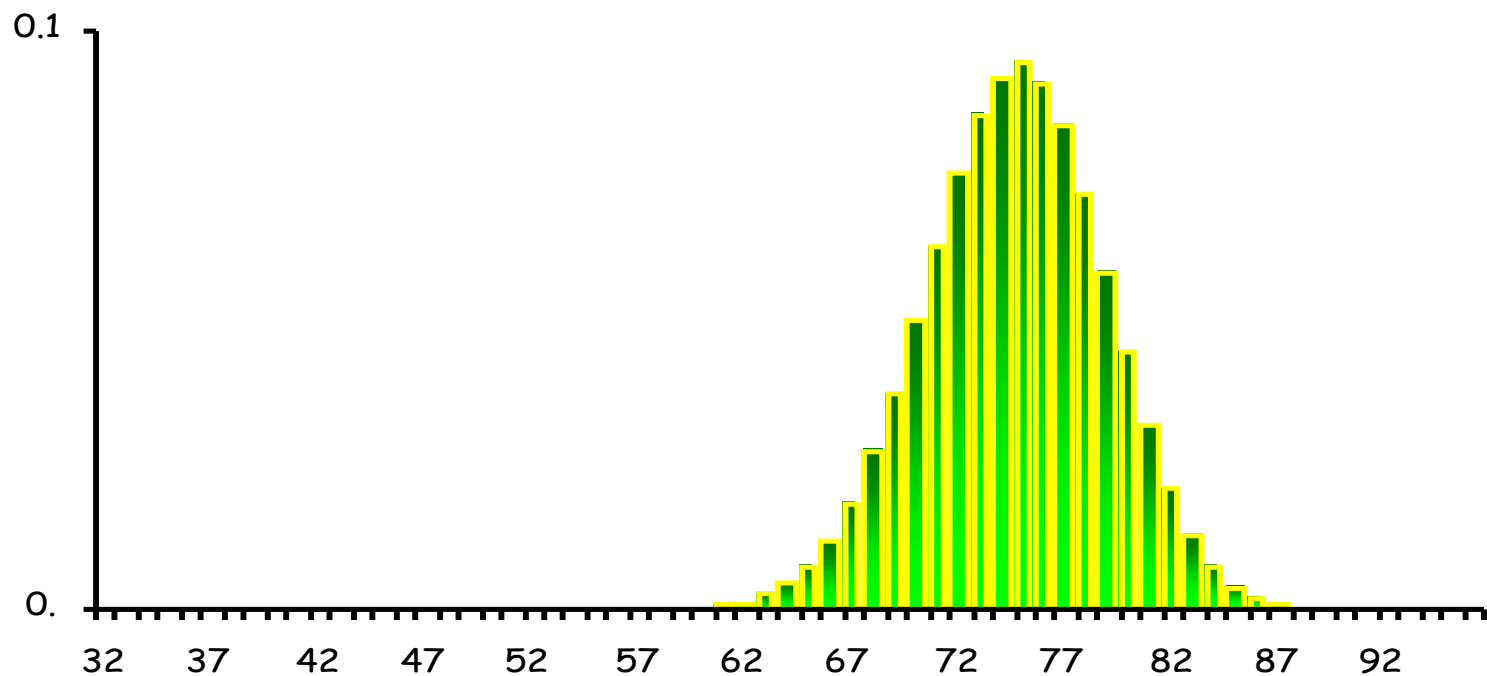
DISTRIBUTION OF S_8



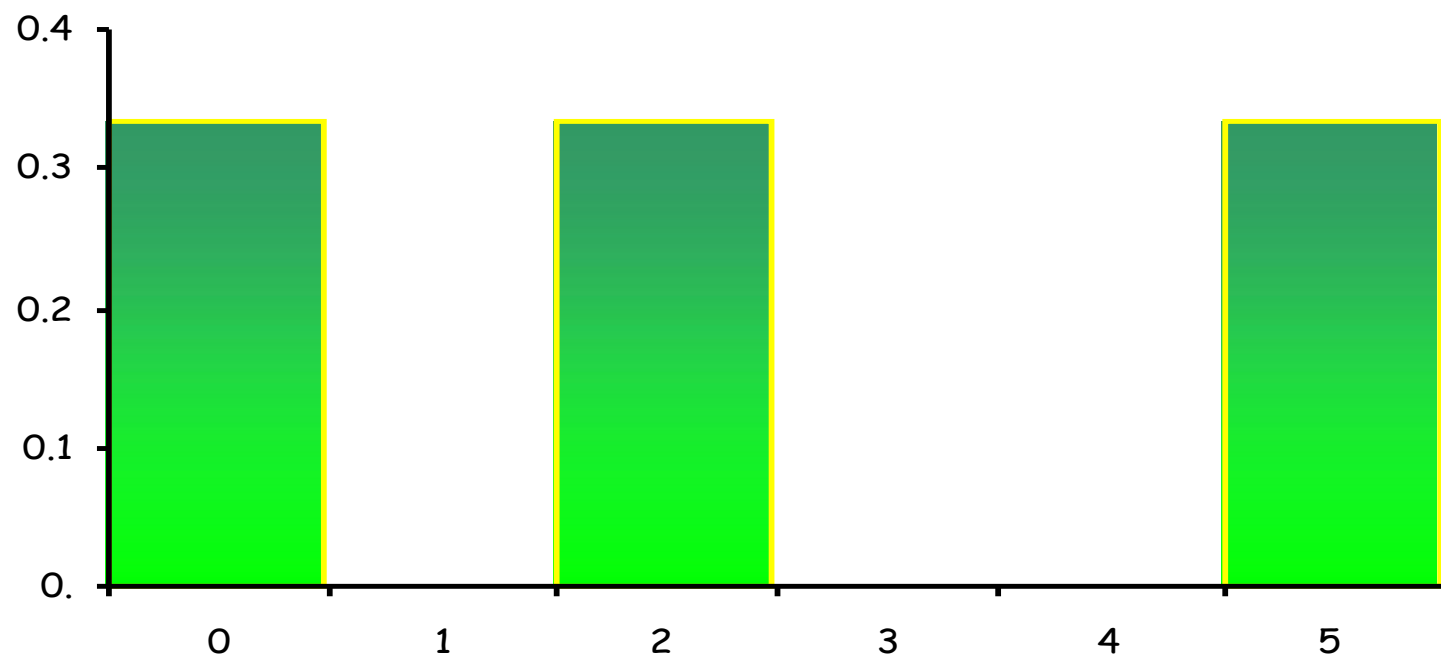
DISTRIBUTION OF S_{16}



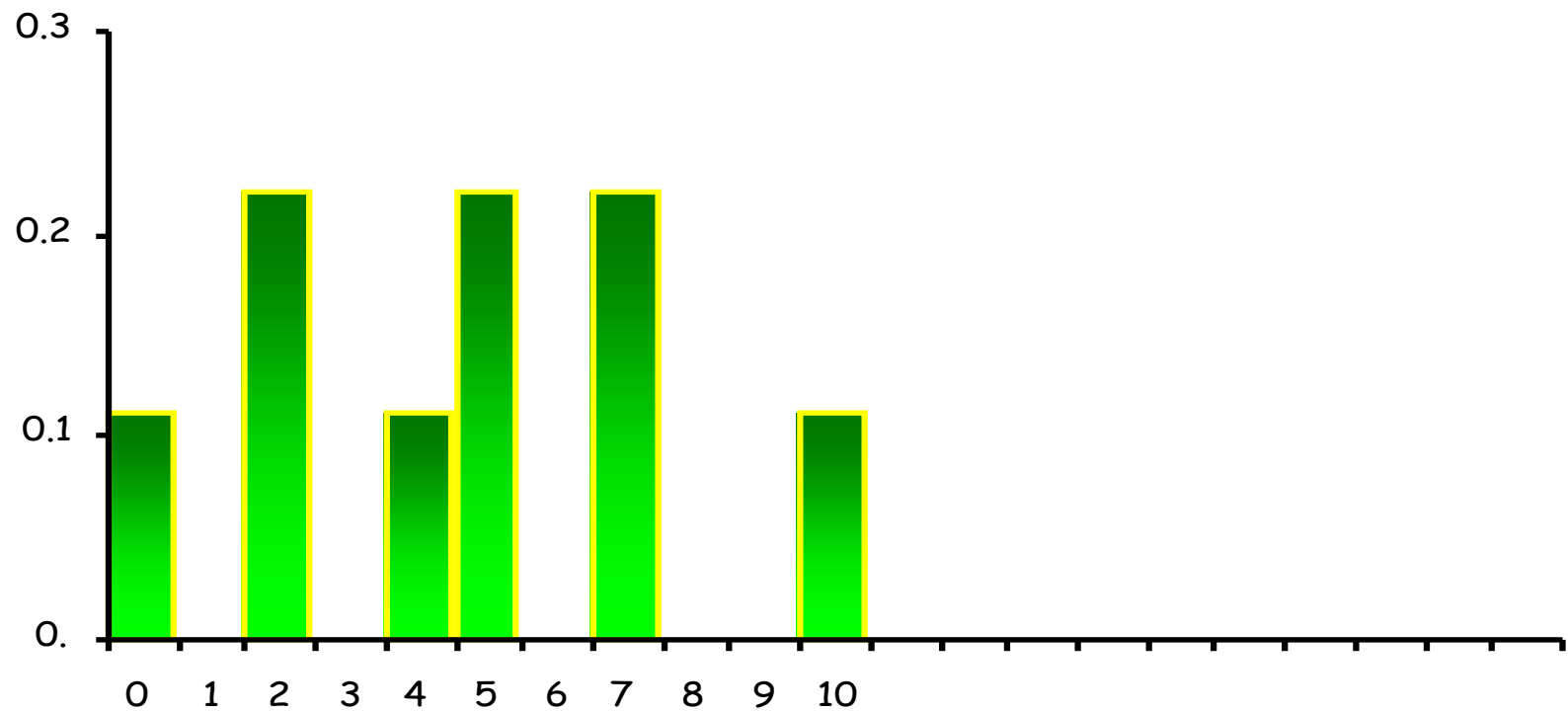
DISTRIBUTION OF S_{32}



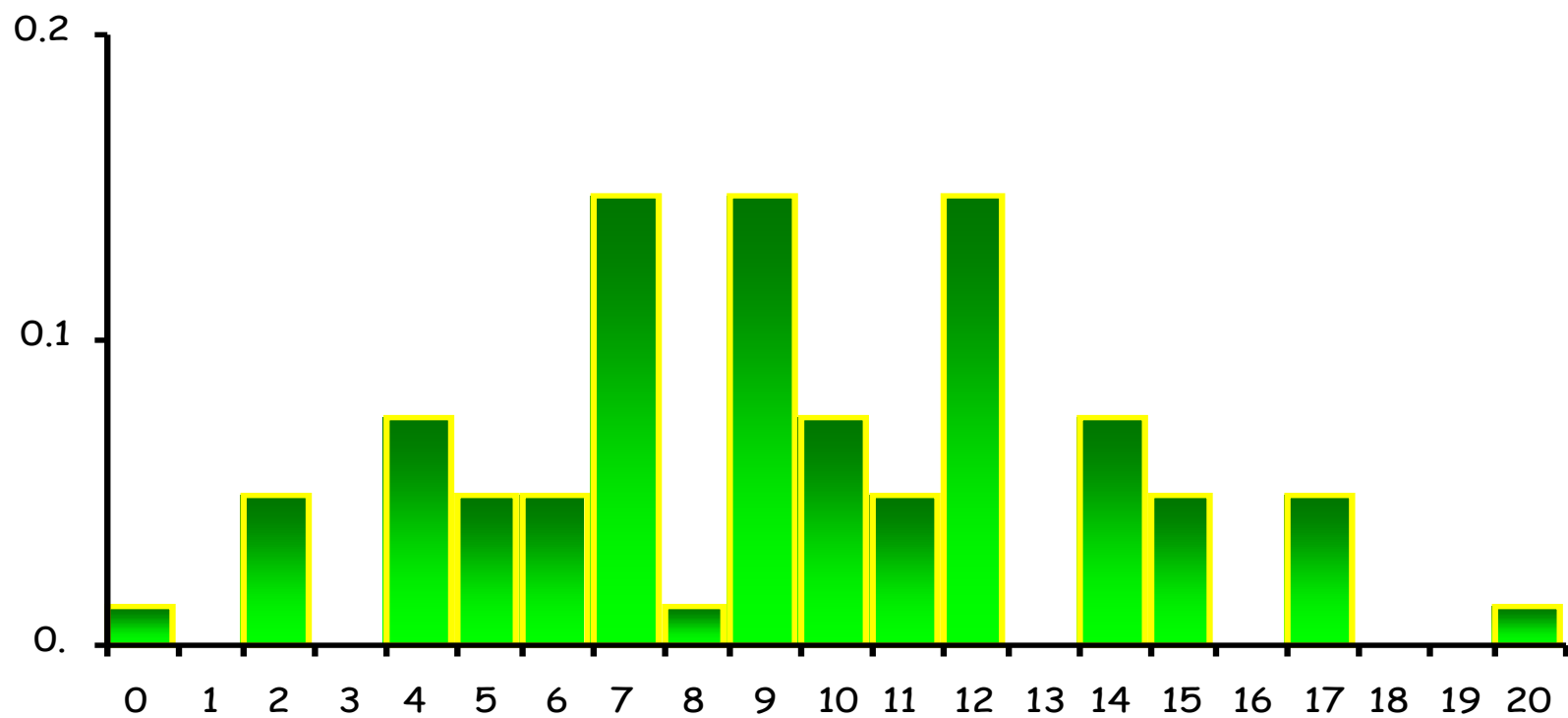
DISTRIBUTION OF X_1



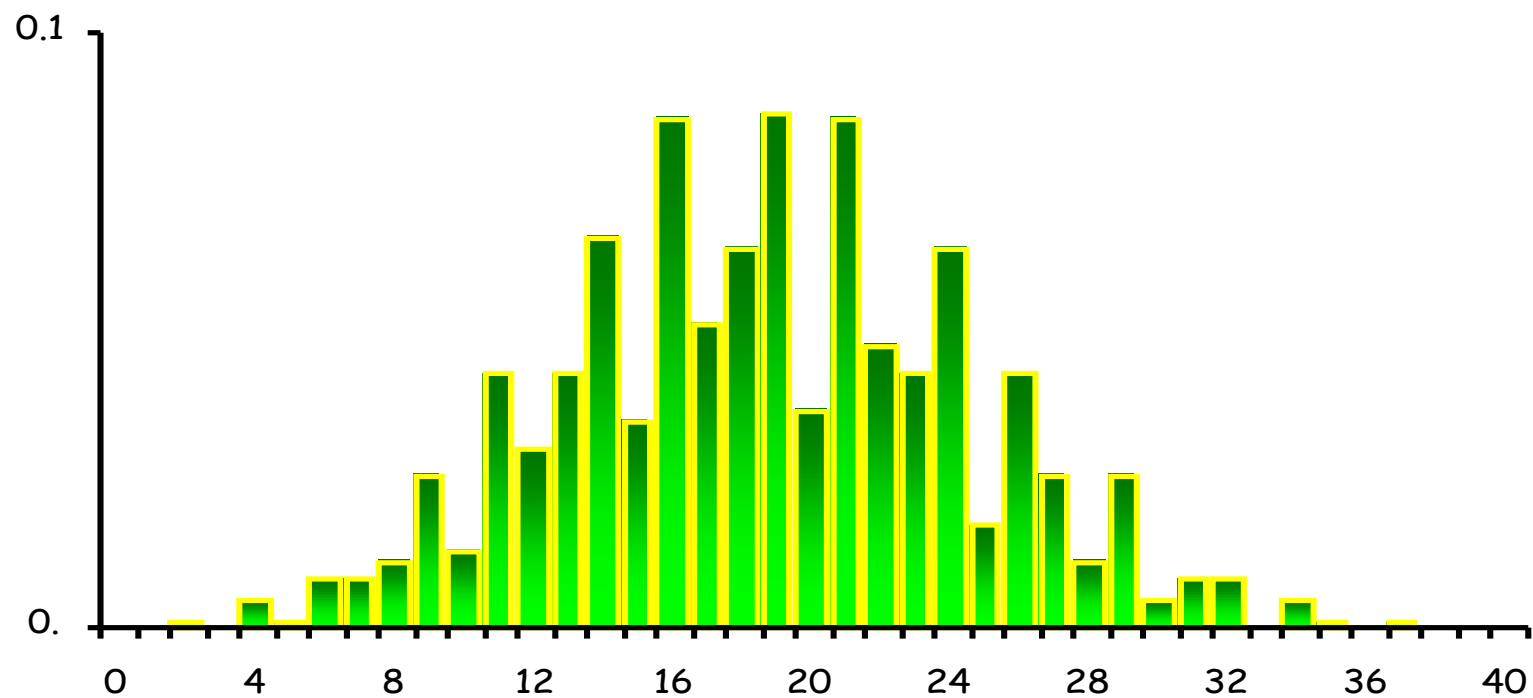
DISTRIBUTION OF S_2



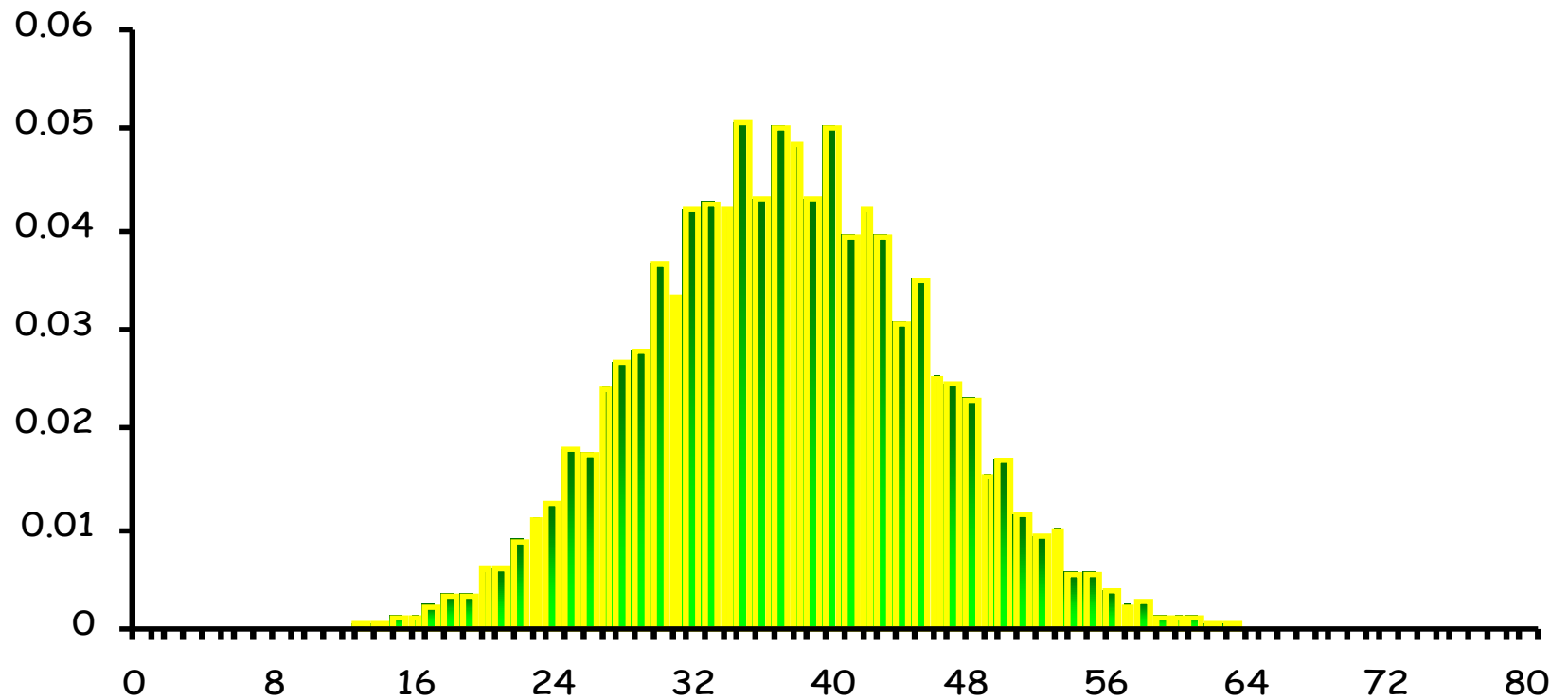
DISTRIBUTION OF S_4



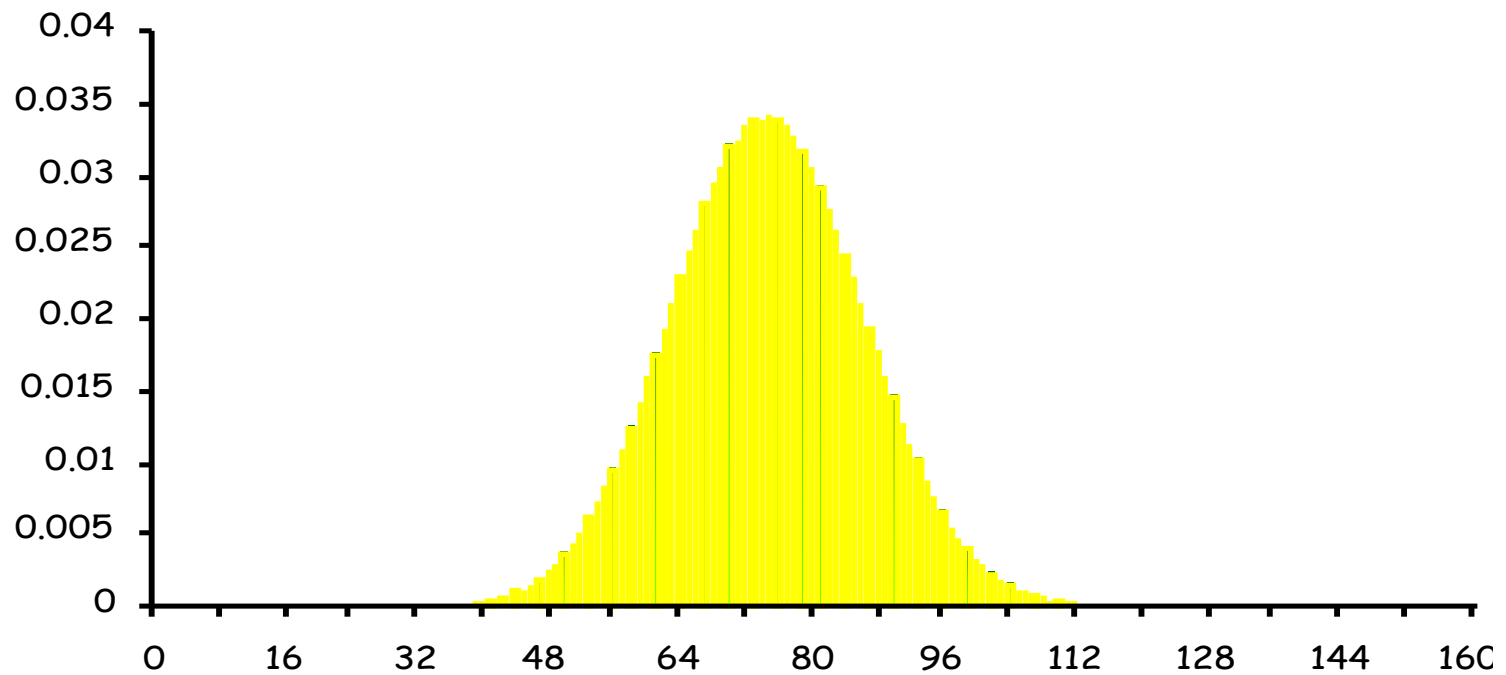
DISTRIBUTION OF S_8



DISTRIBUTION OF S_{16}



DISTRIBUTION OF S_{32}

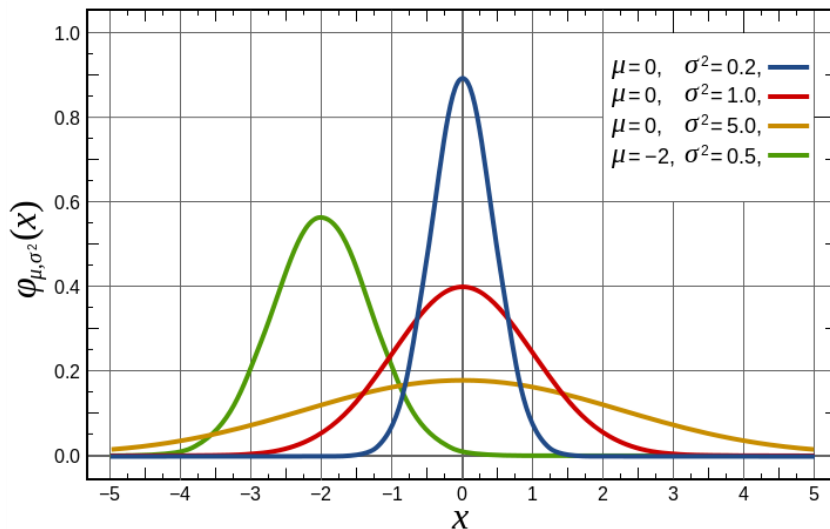


NORMAL DISTRIBUTION

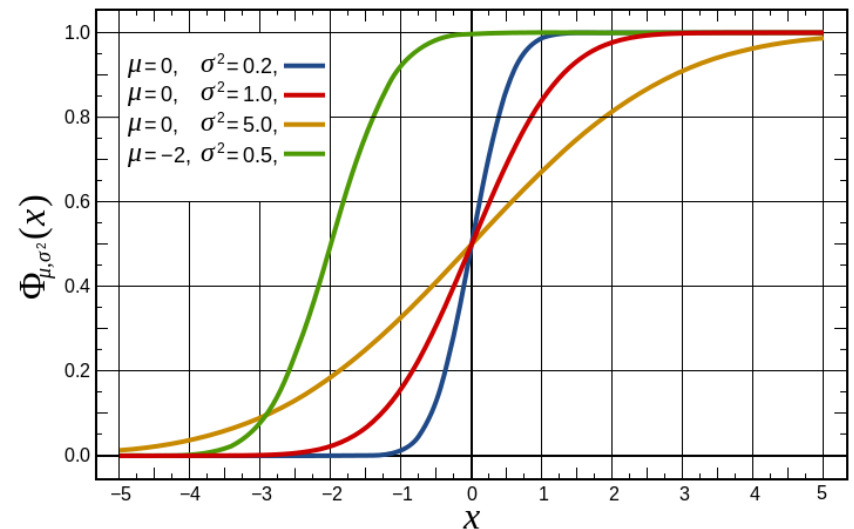
$$N(m, \sigma^2)$$

Probability Density Function

$$f(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-(x-\mu)^2 / (2\sigma^2)}$$



Probability Density Function (PDF)



Cumulative Distribution Function (CDF)

Clicker:

You know from a previous study, that the customers have a normally distributed income with mean $\mu = \$50k$ and $\sigma = \$10k$

If you randomly select 100 customers, how many do you expect to have an income of over \$70k?

- a) ≤ 0.02
- b) ≈ 1.00
- c) ≈ 2.28
- d) ≈ 15.87

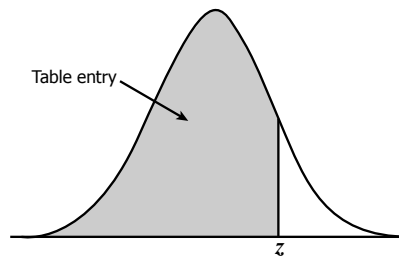


Table entry for z is the area under the standard normal curve to the left of z .

| z | .00 | .01 | .02 | .03 | .04 | .05 | .06 | .07 | .08 | .09 |
|-----|-------|-------|-------|-------|-------|-------|-------|-------|-------|-------|
| 0.0 | .5000 | .5040 | .5080 | .5120 | .5160 | .5199 | .5239 | .5279 | .5319 | .5359 |
| 0.1 | .5398 | .5438 | .5478 | .5517 | .5557 | .5596 | .5636 | .5675 | .5714 | .5753 |
| 0.2 | .5793 | .5832 | .5871 | .5910 | .5948 | .5987 | .6026 | .6064 | .6103 | .6141 |
| 0.3 | .6179 | .6217 | .6255 | .6293 | .6331 | .6368 | .6406 | .6443 | .6480 | .6517 |
| 0.4 | .6554 | .6591 | .6628 | .6664 | .6700 | .6736 | .6772 | .6808 | .6844 | .6879 |
| 0.5 | .6915 | .6950 | .6985 | .7019 | .7054 | .7088 | .7123 | .7157 | .7190 | .7224 |
| 0.6 | .7257 | .7291 | .7324 | .7357 | .7389 | .7422 | .7454 | .7486 | .7517 | .7549 |
| 0.7 | .7580 | .7611 | .7642 | .7673 | .7704 | .7734 | .7764 | .7794 | .7823 | .7852 |
| 0.8 | .7881 | .7910 | .7939 | .7967 | .7995 | .8023 | .8051 | .8078 | .8106 | .8133 |
| 0.9 | .8159 | .8186 | .8212 | .8238 | .8264 | .8289 | .8315 | .8340 | .8365 | .8389 |
| 1.0 | .8413 | .8438 | .8461 | .8485 | .8508 | .8531 | .8554 | .8577 | .8599 | .8621 |
| 1.1 | .8643 | .8665 | .8686 | .8708 | .8729 | .8749 | .8770 | .8790 | .8810 | .8830 |
| 1.2 | .8849 | .8869 | .8888 | .8907 | .8925 | .8944 | .8962 | .8980 | .8997 | .9015 |
| 1.3 | .9032 | .9049 | .9066 | .9082 | .9099 | .9115 | .9131 | .9147 | .9162 | .9177 |
| 1.4 | .9192 | .9207 | .9222 | .9236 | .9251 | .9265 | .9279 | .9292 | .9306 | .9319 |
| 1.5 | .9332 | .9345 | .9357 | .9370 | .9382 | .9394 | .9406 | .9418 | .9429 | .9441 |
| 1.6 | .9452 | .9463 | .9474 | .9484 | .9495 | .9505 | .9515 | .9525 | .9535 | .9545 |
| 1.7 | .9554 | .9564 | .9573 | .9582 | .9591 | .9599 | .9608 | .9616 | .9625 | .9633 |
| 1.8 | .9641 | .9649 | .9656 | .9664 | .9671 | .9678 | .9686 | .9693 | .9699 | .9706 |
| 1.9 | .9713 | .9719 | .9726 | .9732 | .9738 | .9744 | .9750 | .9756 | .9761 | .9767 |
| 2.0 | .9772 | .9778 | .9783 | .9788 | .9793 | .9798 | .9803 | .9808 | .9812 | .9817 |
| 2.1 | .9821 | .9826 | .9830 | .9834 | .9838 | .9842 | .9846 | .9850 | .9854 | .9857 |
| 2.2 | .9861 | .9864 | .9868 | .9871 | .9875 | .9878 | .9881 | .9884 | .9887 | .9890 |
| 2.3 | .9893 | .9896 | .9898 | .9901 | .9904 | .9906 | .9909 | .9911 | .9913 | .9916 |
| 2.4 | .9918 | .9920 | .9922 | .9925 | .9927 | .9929 | .9931 | .9932 | .9934 | .9936 |
| 2.5 | .9938 | .9940 | .9941 | .9943 | .9945 | .9946 | .9948 | .9949 | .9951 | .9952 |
| 2.6 | .9953 | .9955 | .9956 | .9957 | .9959 | .9960 | .9961 | .9962 | .9963 | .9964 |
| 2.7 | .9965 | .9966 | .9967 | .9968 | .9969 | .9970 | .9971 | .9972 | .9973 | .9974 |
| 2.8 | .9974 | .9975 | .9976 | .9977 | .9977 | .9978 | .9979 | .9979 | .9980 | .9981 |
| 2.9 | .9981 | .9982 | .9982 | .9983 | .9984 | .9984 | .9985 | .9985 | .9986 | .9986 |
| 3.0 | .9987 | .9987 | .9987 | .9988 | .9988 | .9989 | .9989 | .9989 | .9990 | .9990 |
| 3.1 | .9990 | .9991 | .9991 | .9991 | .9992 | .9992 | .9992 | .9992 | .9993 | .9993 |
| 3.2 | .9993 | .9993 | .9994 | .9994 | .9994 | .9994 | .9994 | .9995 | .9995 | .9995 |
| 3.3 | .9995 | .9995 | .9995 | .9996 | .9996 | .9996 | .9996 | .9996 | .9996 | .9997 |
| 3.4 | .9997 | .9997 | .9997 | .9997 | .9997 | .9997 | .9997 | .9997 | .9997 | .9998 |

Ignore that there are no negative incomes

You know from a previous study, that the customers have a normally distributed income with mean $\mu = \$50k$ and $\sigma = \$10k$

If you randomly select 100 customers, how many do you expect to have have an income of over \$70k?

$$Z = \frac{X - m}{S} = \frac{70k - 50k}{10k} = 2$$

$$P(X \geq 70k) \approx 100 \\ = 100 \cdot (1 - 0.9772) = 2.28$$

Ignore that there are no negative incomes

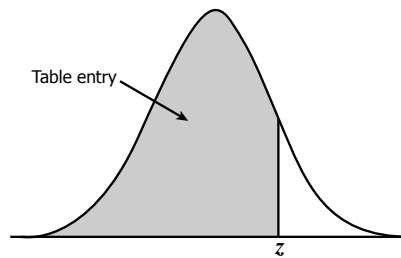


Table entry for z is the area under the standard normal curve to the left of z .

[illegible]

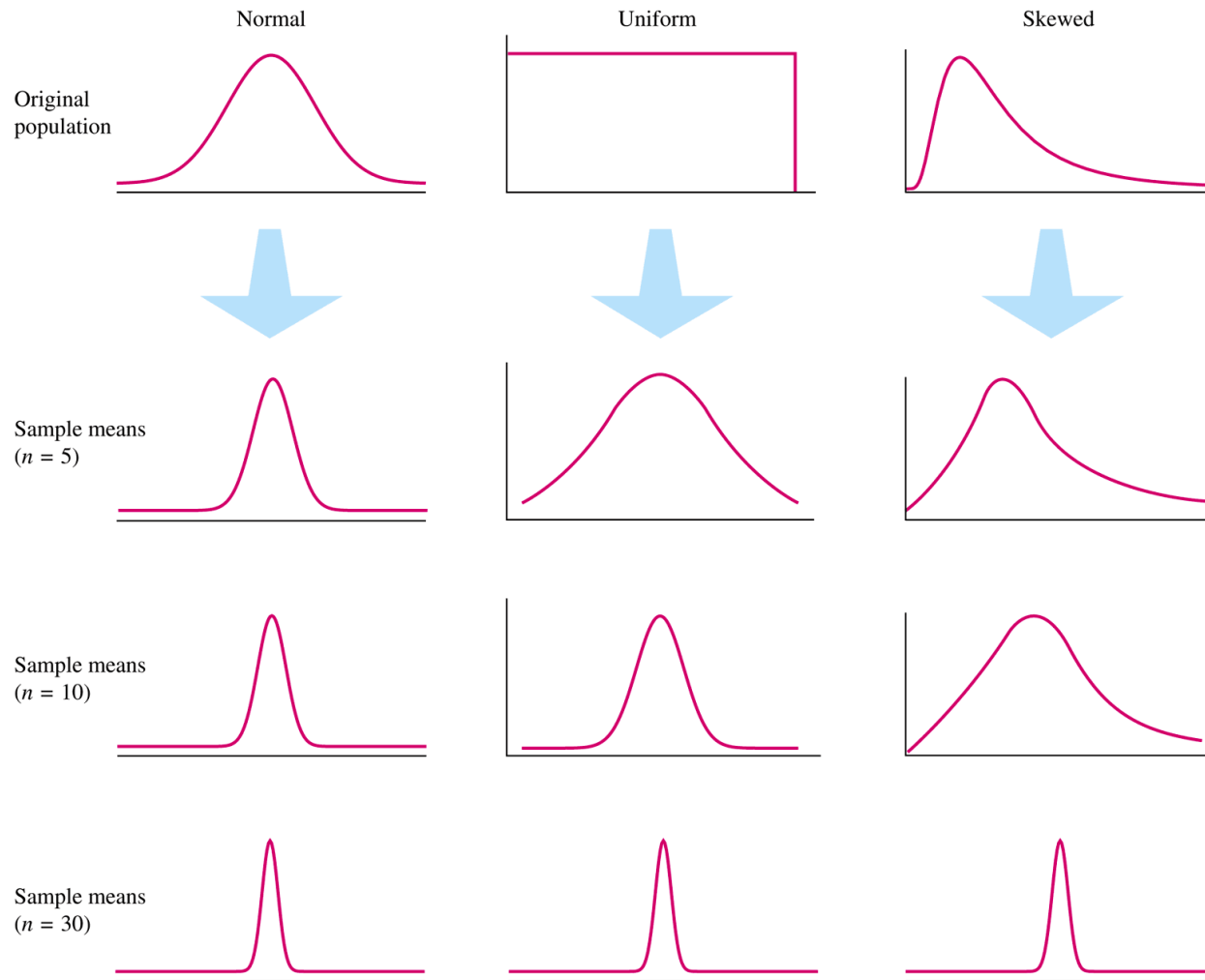
THE CENTRAL LIMIT THEOREM

1. The distribution of means will be approximately a normal distribution for larger sample sizes
2. The mean of the distribution of means approaches the population mean, μ , for large sample sizes
3. The **standard deviation of the distribution of means** approaches σ/\sqrt{n} for large sample sizes, where σ is the standard deviation of the population and n is the sample size

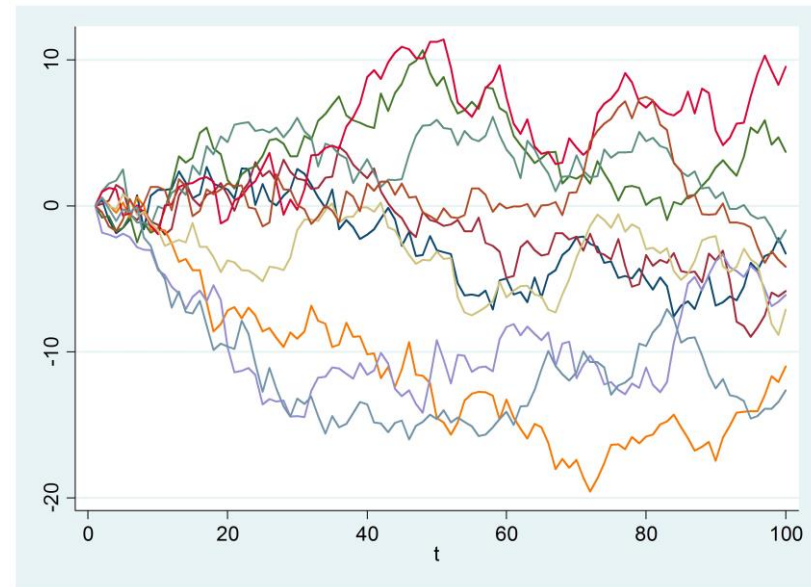
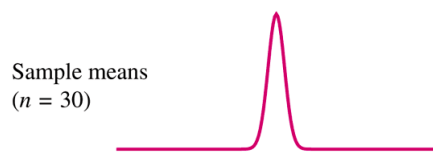
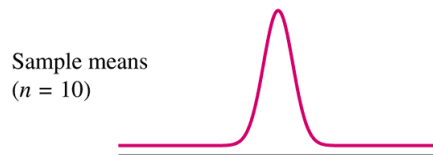
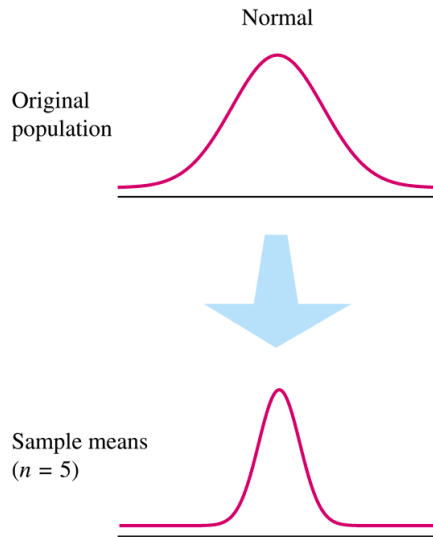
THE CENTRAL LIMIT THEOREM: NOTES

1. For practical purposes, the distribution of means will be nearly normal if the sample size is larger than 30
2. If the original population is normally distributed, then the sample means will remain normally distributed for any sample size n , and it will become narrower
3. The original variable can have any distribution, it does not have to be a normal distribution

SHAPES OF DISTRIBUTIONS AS SAMPLE SIZE INCREASES



THIS THEOREM IS EVERYWHERE!



| | A | B | C | D | E | F | G | H | I | J | K |
|----|---------------------|-----------|-----------------|----------------|-----------|-----------|-----------|------------------|-----------|-----------|-----------|
| 44 | Portfolio Scenarios | | | Risk Minimized | | | | Return Maximized | | | |
| 45 | | Equal- | | Min | Long | 0% to 25% | 5% to 15% | | Long | 0% to 25% | 5% to 15% |
| 46 | | Weighted | | Var | Only | Weights | Weights | | Only | Weights | Weights |
| 47 | | Portfolio | Stock | Portfolio | Portfolio | Portfolio | Portfolio | Stock | Portfolio | Portfolio | Portfolio |
| 48 | Stock | | | | | | | | | | |
| 49 | ARO | 10.00% | ARO | (43.08%) | 0.00% | 0.00% | 5.00% | ARO | 0.00% | 25.00% | 15.00% |
| 50 | ARW | 10.00% | ARW | 87.99% | 16.98% | 25.00% | 15.00% | ARW | 0.00% | 0.00% | 15.00% |
| 51 | ASI | 10.00% | ASI | 7.72% | 0.00% | 0.00% | 15.00% | ASI | 0.00% | 0.00% | 5.00% |
| 52 | GLW | 10.00% | GLW | (14.02%) | 0.00% | 0.00% | 5.00% | GLW | 0.00% | 25.00% | 15.00% |
| 53 | GTIV | 10.00% | GTIV | 70.77% | 59.67% | 25.00% | 15.00% | GTIV | 0.00% | 0.00% | 5.00% |
| 54 | KLIC | 10.00% | KLIC | 18.61% | 0.00% | 0.00% | 5.00% | KLIC | 0.00% | 25.00% | 15.00% |
| 55 | MKSI | 10.00% | MKSI | (28.90%) | 0.00% | 25.00% | 15.00% | MKSI | 0.00% | 0.00% | 5.00% |
| 56 | OME | 10.00% | OME | 1.61% | 0.00% | 0.00% | 5.00% | OME | 0.00% | 0.00% | 5.00% |
| 57 | PL | 10.00% | PL | 5.13% | 23.35% | 25.00% | 15.00% | PL | 0.00% | 0.00% | 5.00% |
| 58 | SNDK | 10.00% | SNDK | (5.83%) | 0.00% | 0.00% | 5.00% | SNDK | 100.00% | 25.00% | 15.00% |
| 59 | Sum of Weights | 100.00% | Sum of Weights | 100.00% | 100.00% | 100.00% | 100.00% | Sum of Weights | 100.00% | 100.00% | 100.00% |
| 60 | Expected Return | 27.75% | Expected Return | 7.54% | 15.93% | 14.35% | 21.55% | Expected Return | 60.40% | 45.26% | 34.00% |
| 61 | Std Dev | 44.51% | Std Dev | 0.00% | 14.12% | 21.45% | 33.28% | Std Dev | 104.46% | 78.43% | 56.74% |

メトロノーム同期 (32個)

Synchronization of thirty two metronomes

2012年09月14日, 池口研究室前廊下にて撮影

Filmed at Ikeguchi Laboratory, on September 14, 2012.

Clicker:

For a normal distributed random variable X with

$$m = 5$$

$$S = 10$$

What is the likelihood to see a value less than 10.5

$$P(X < 10.5)$$

- a) 0.5
b) 0.6915
c) 0.7088
d) 0.9987

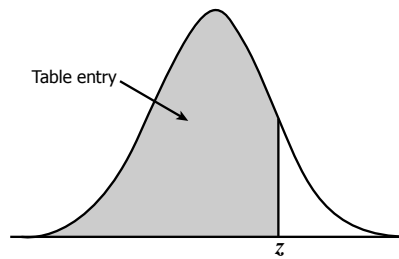


Table entry for z is the area under the standard normal curve to the left of z .

[illegible]

Clicker:

For a normal distributed
random variable X with

$$m = 5$$

$$S = 10$$

What is the likelihood that to see a value above 10.5

$$Z = \frac{X - m}{S}$$
$$= \frac{10.5 - 5}{10} = 0.55$$

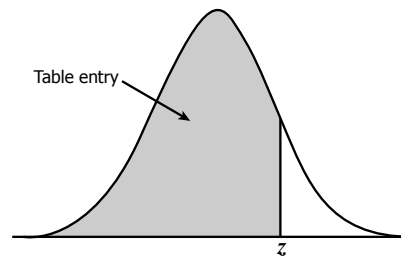


Table entry for z is the area under the standard normal curve to the left of z .

[illegible]

Clicker:

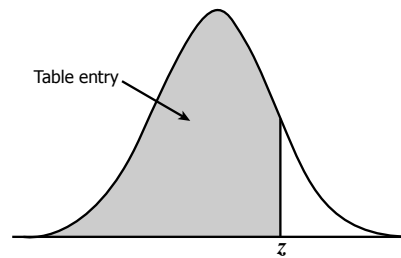


Table entry for z is the area under the standard normal curve to the left of z .

For a normal distributed random variable X with

$$m = 5$$

$$S = 10$$

What is the value of d so that

$$P(Z < d) \leq 0.05$$

- a) -21.5
b) -11.5
c) 5
d) 21.5

[illegible]

Clicker:

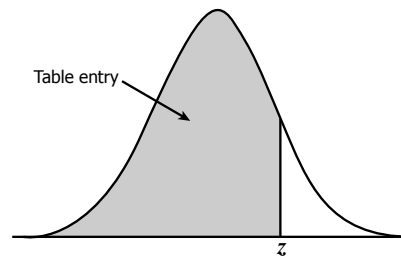


Table entry for z is the area under the standard normal curve to the left of z .

For a normal distributed random variable X with

$$m = 5$$

$$S = 10$$

What is the value of d so that

$$P(Z < d) \leq 0.05$$

$$-1.65 = \frac{x - m}{s} = \frac{d - 5}{10}$$

$$d = -1.65 \times 10 + 5 = -11.5$$

[illegible]