Why Probability ?

- **Statistics uses probability to convey inferential conclusions from sample data about a larger population.**

- **Probability distributions (e.g., normal, binomial) describe how data is spread. Understanding these distributions allows statisticians to solve probability problems with appropriate statistical methods.**

Probability

THE

DOCTRINE

OF

CHANCES:

OR,

A Method of Calculating the Probability of Events in Play.

By *A. De Moivre.* F. R. S.

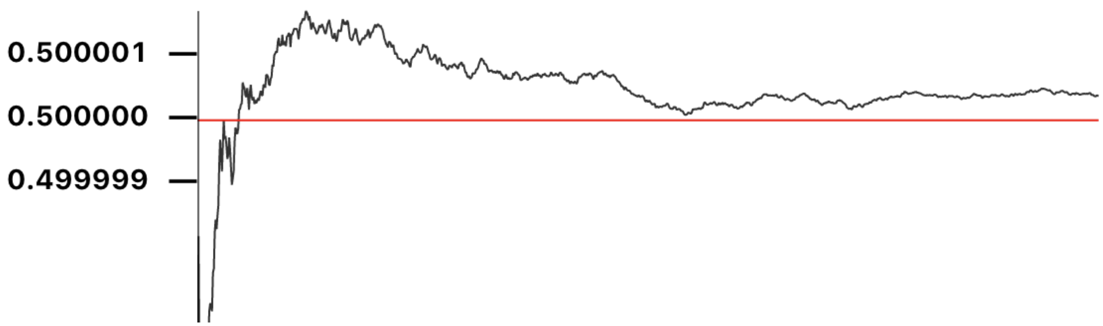
LONDON:

Printed by *W. Pearson,* for the Author. MDCCXVIII.

Probability :

The convergent ratio of "successes" over the number of trials where the trials are performed "ad infinitum".

Go



1000000

How many?

Probability :

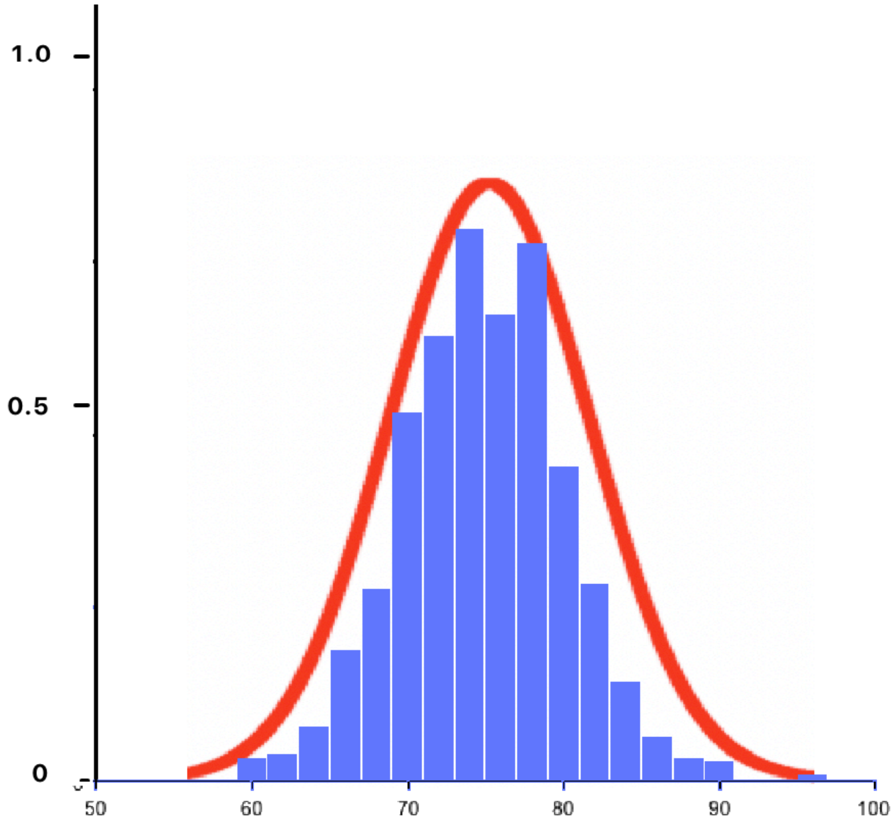
To calculate we must have either

- **a "history" of trials and successes**

or

- **a complete mathematical representation of the system in question**

Relative Frequency Histograms and Probability Distributions



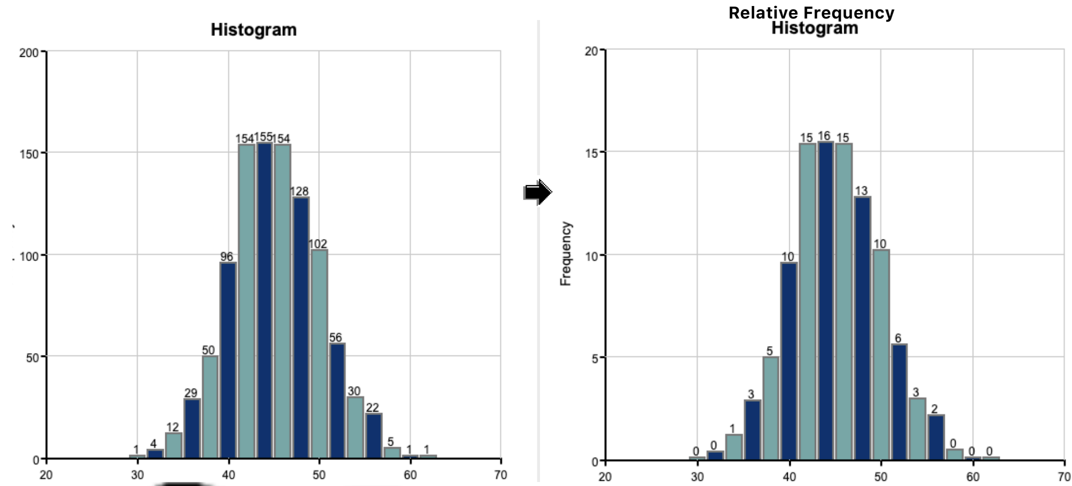
Many statistics texts do not discuss the concept of probability density in detail, but you should keep the following ideas in mind about the curve that describes a continuous distribution (like the normal distribution).

- **First, the area under the curve equals 1.**
- **Second, the probability of any exact value of *X* is 0.**
- **Finally, the area under the curve and bounded between two given points on the *X*-axis is the probability that a number chosen at random will fall between the two points.**

Also,

- $P(A) = 1 - P(\text{not } A)$
- $P(A \text{ and } B) = P(A) \times P(B)$
- $P(A \text{ or } B) = P(A) + P(B) - P(A) \times P(B)$

Data Distributions vs Probability Distributions



- Uniform
- Binomial
- Poisson
- Geometric
- Hypergeometric
- Multinomial

- Normal
- Standard Normal
- Exponential
- Gamma
- Beta
- Erlang

⋮

Sampling Distributions:

- means
- variances
- standard deviations
- medians
- ranges
- ratio of variances

⋮

Contrived Distributions:

- Student's t
- Chi-Square
- Snedecor - Fisher F

⋮

Probability :

- The length of human pregnancies from conception to birth approximates a normal distribution with a mean of 266 days and a standard deviation of 16 days.

  - What proporton of pregnancies last less than 240 days ?
  - What fraction of pregnancies last less than 240 days ?
  - What percent of pregnancies last less than 240 days ?
  - What is the probability that a pregnancy will last less than 240 days ?
  - What length of time marks the shortest 70% of all pregnancies?

$$\int_0^{240} \frac{1.0}{16\sqrt{2*\pi}} e^{-\frac{1}{2}\left(\frac{x-266}{16}\right)^2} \, dx$$

0.0520813

$$\text{Solve}\left[0.70 == \int_0^{x} \frac{1.0}{16\sqrt{2*\pi}} e^{-\frac{1}{2}\left(\frac{x-266}{16}\right)^2} \, dx,\ x\right]$$

$x \to 274.39$

Data Distributions vs Probability Distributions

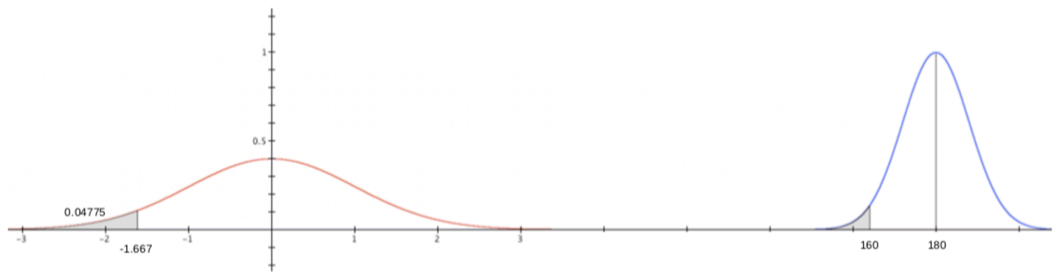
Three ways to approach our statistical work.

1.

$$\int_0^{160} \frac{1}{\sqrt{2.0\,\pi\,(12)^2}}\, e^{\frac{-(x-180)^2}{2\,(12)^2}}\, dx$$

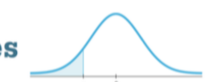
0.0477904

Using the Standard Normal Probability Distribution

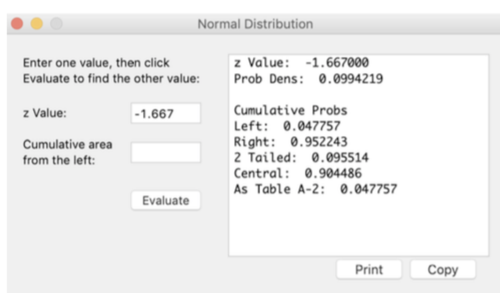


2.

NEGATIVE z Scores



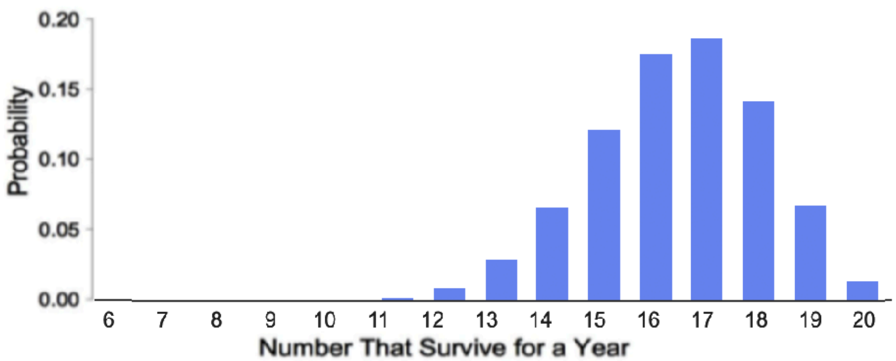
−2.0	.0228	.0222	.0217	.0212	.0207	.0202	.0197	.0192
−1.9	.0287	.0281	.0274	.0268	.0262	.0256	.0250	.0244
−1.8	.0359	.0351	.0344	.0336	.0329	.0322	.0314	.0307
−1.7	.0446	.0436	.0427	.0418	.0409	.0401	.0392	.0384
−1.6	.0548	.0537	.0526	.0516	.0505	.0495	.0485	.0475



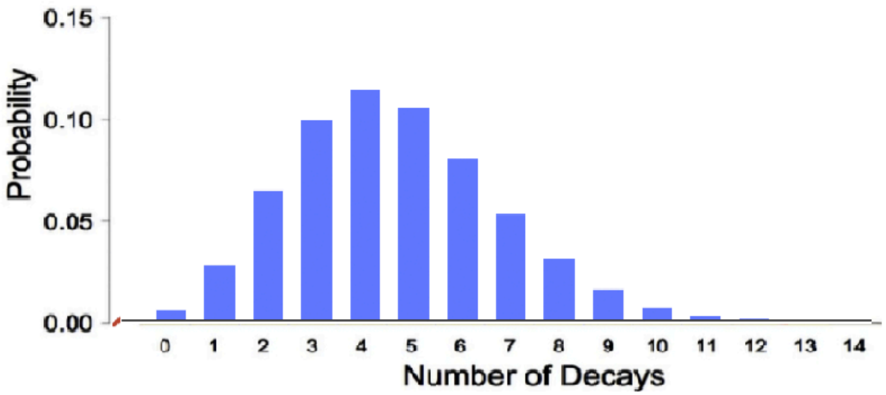
3.

Discrete Distributions

Binomial Distribution

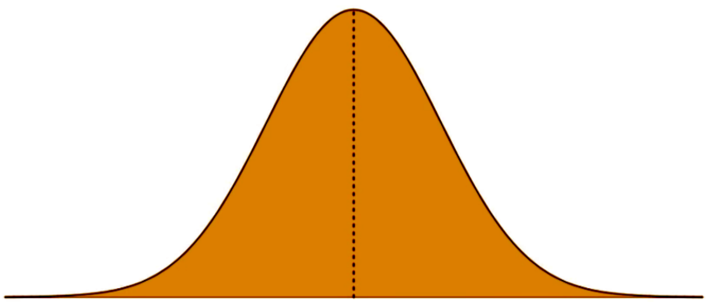


Poisson Distribution



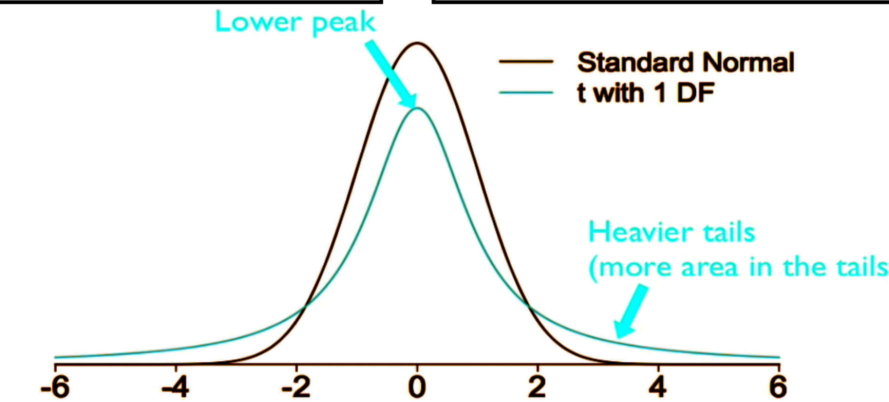
Continuous Distributions

Normal Distribution



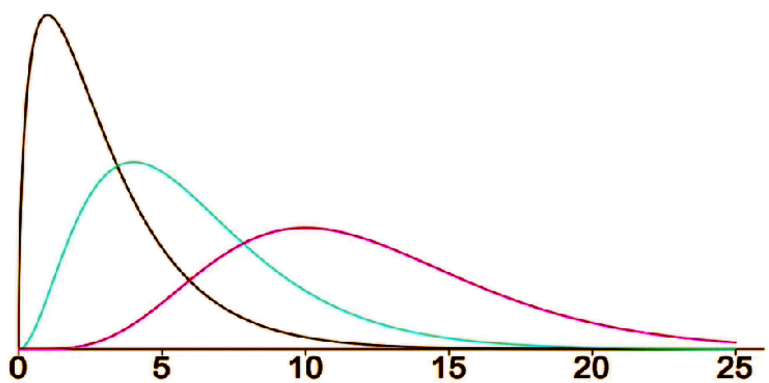
Standard Normal Distribution

Student's *t*-Distribution

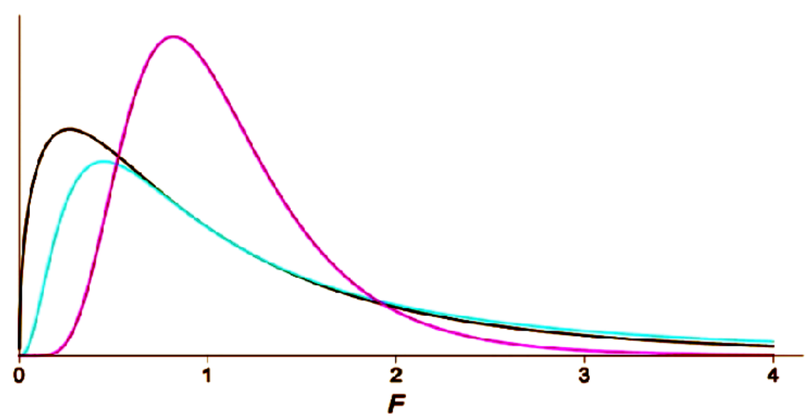


As the degrees of freedom increase, the t distribution tends toward the standard normal distribution

Chi Square Distribution



Fischer F Distribution



IQ Scores

Rainfall Amounts

Product Lifetimes

Wind Speeds

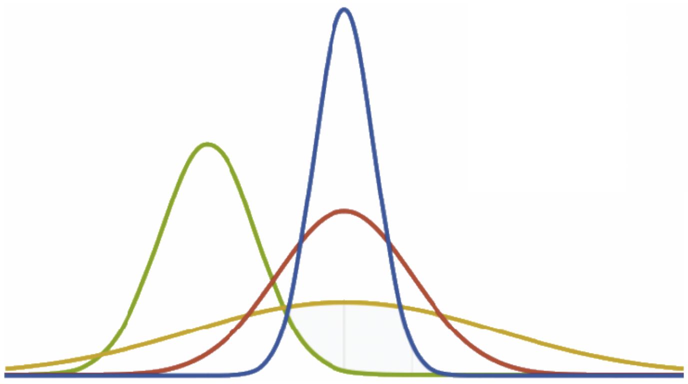
Cholesterol Levels

Sensor Measurements

Network Latency

Human Reaction Times

Normal Dustribution



The normal probability distribution was first discovered by **Abraham De Moivre** in 1733 as an approximation to the binomial distribution. While often called the Gaussian distribution and attributed to Carl Friedrich Gauss (who published his work on it in 1809), Gauss's work built upon and refined the existing principles of the distribution.

Blood Glucose Levels

Memory Recall Time

Heart Rate

Material Strength

Delivery Times

Defect Rates

Minor Earthquake Magnitudes

Birth Weights

Website Traffic

Temperature Variations

Production Costs

Battery Life

Body Weight

Insurance Claim Amounts

Reading Speeds

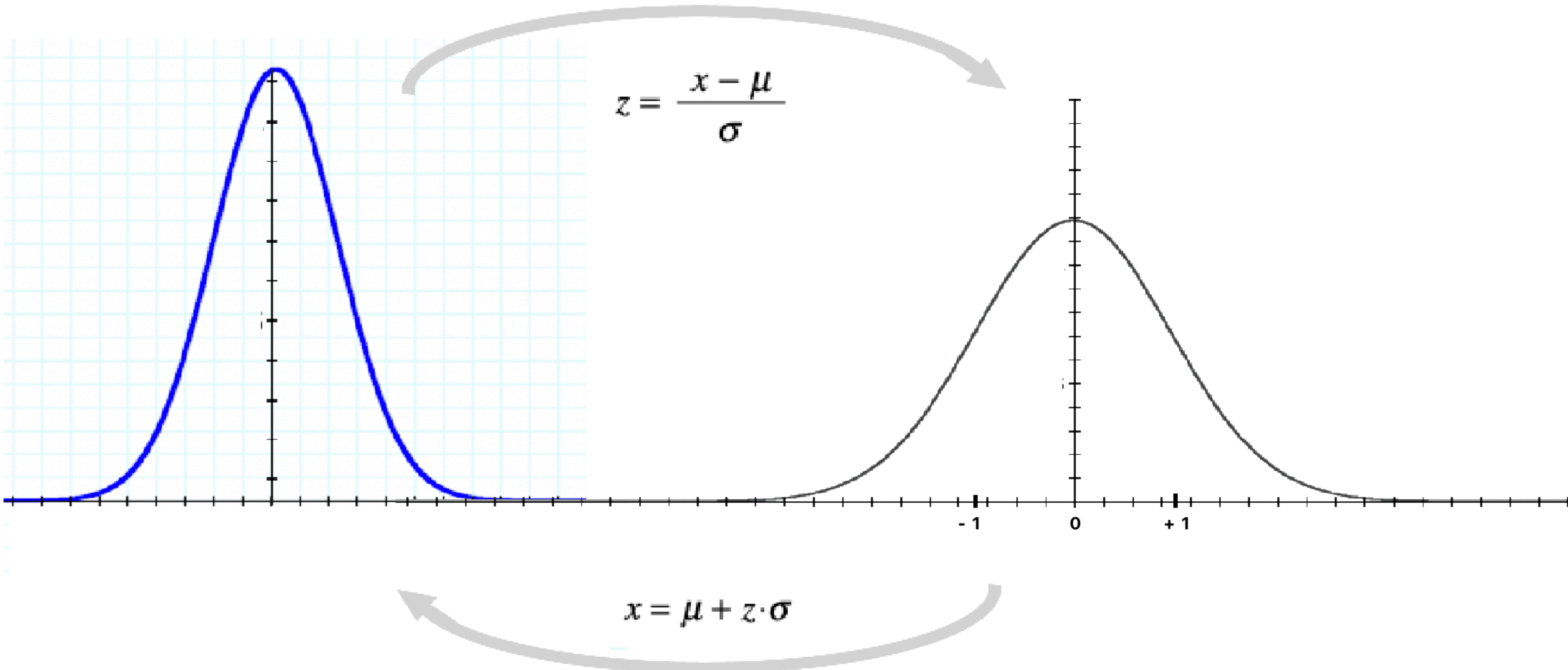
Body Temperature

River Flow Rates

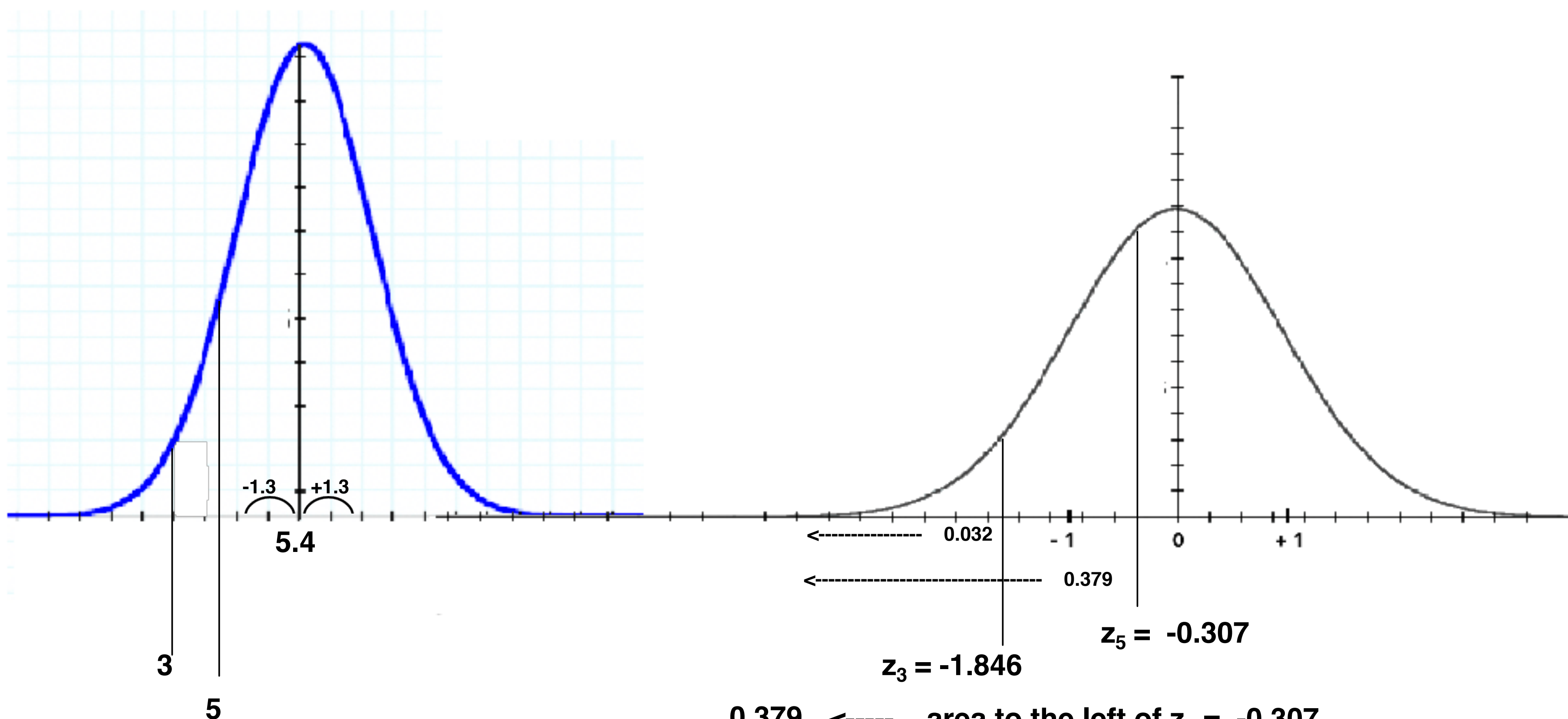
Machine Tolerances

Time to Complete Tasks

Standardize any Normal Distribution



Standardize any Normal Distribution



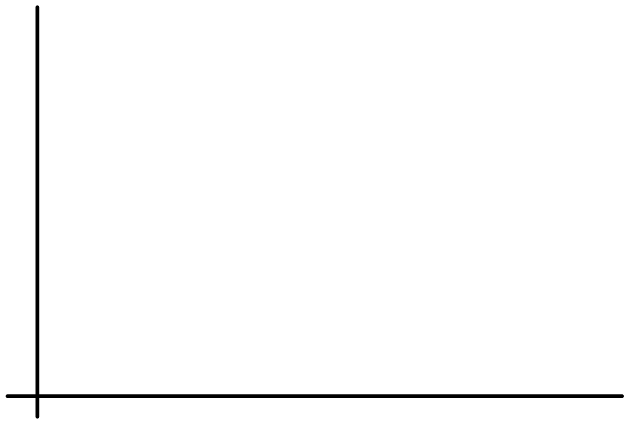
0.379	<-----	area to the left of z_5 = -0.307
- 0.032	<-----	area to the left of z_3 = -1.846
0.347		

Histogram

Population Size: 1000

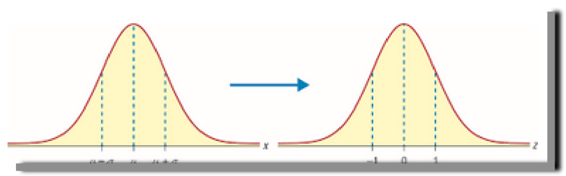
Number of intervals:

- ○ 10
- ◉ 20

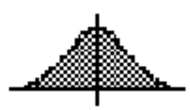


54.18192	1.851768
35.89285	-1.88463
43.48103	-0.33439
52.01798	1.409681
39.49108	-1.14953
44.5143	-0.12330
46.19136	0.219318
42.44947	-0.54514
42.29704	-0.57628
42.7461	-0.48453
42.59781	-0.51483
40.40228	-0.96337
43.07388	-0.41757
49.85313	0.967408
53.36640	1.68516
47.43559	0.473511
52.86845	1.58343
49.45608	0.886292
46.12574	0.205912
46.63426	0.309802
43.16321	-0.39932
44.30296	-0.16647
46.20177	0.221446

Standard Normal Distribution



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



Normal Distribution

MU :	45.117836	MU :	0
SIGMA :	4.894827	SIGMA :	1.0005
VAR :	23.959327	VAR :	1.001001

μ -> 45 σ -> 5

Areas and Probabilities (Normal)

oneStep

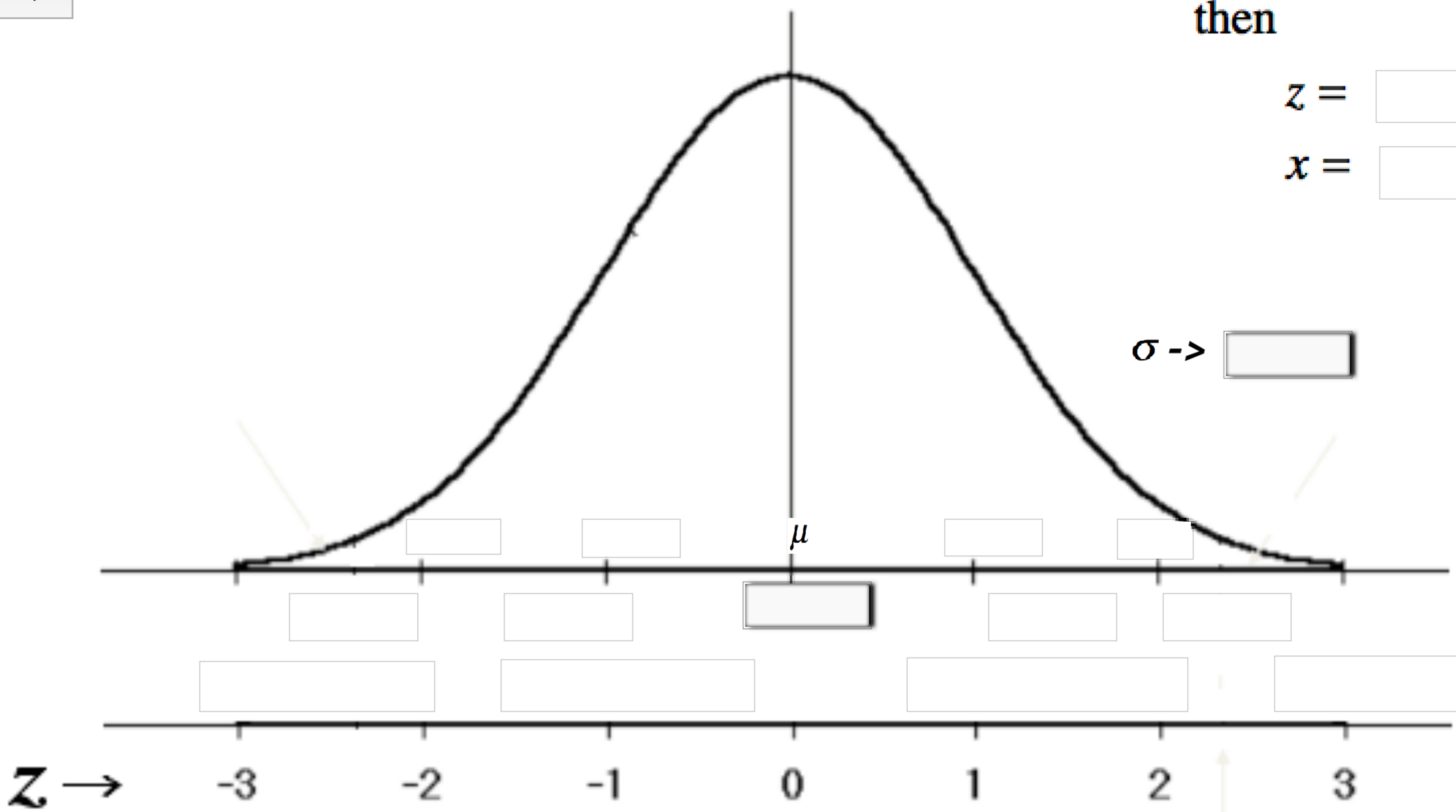
if $p =$

then

$z =$

$x =$

σ ->

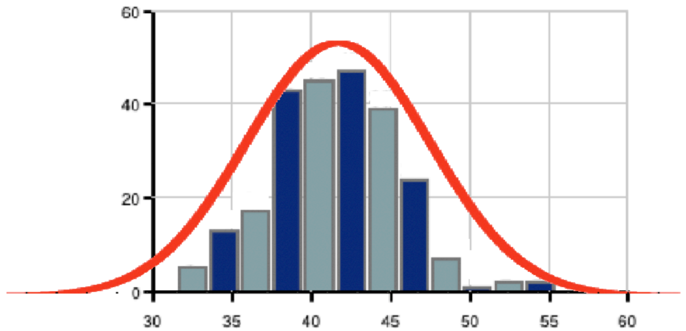


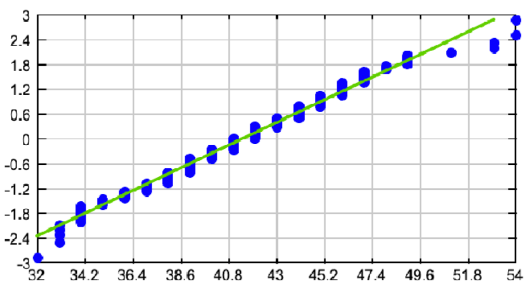
$p(z \leq \quad) =$

$p(\quad \leq z \leq \quad) =$

$p(z \geq \quad) =$

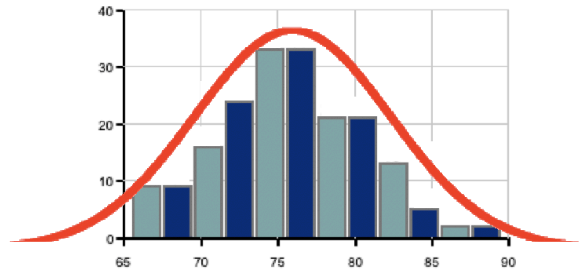
RESET

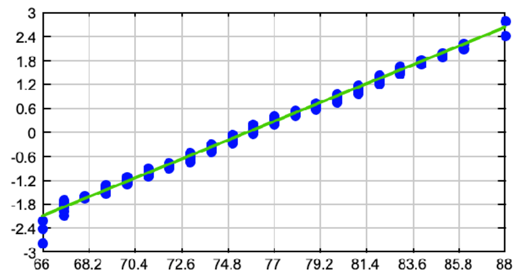




Ryan-Joiner Test
Test statistic, Rp: 0.9924
Critical value for 0.05 significance level: 0.9939
Critical value for 0.01 significance level: 0.9913

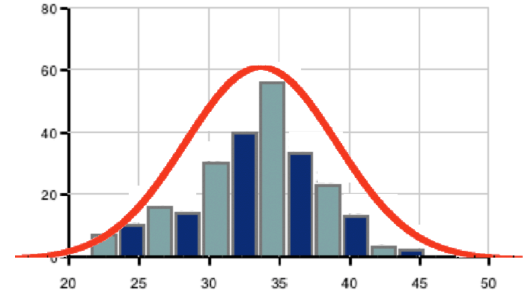
Reject normality with a 0.05 significance level.
Fail to reject normality with a 0.01 significance level.

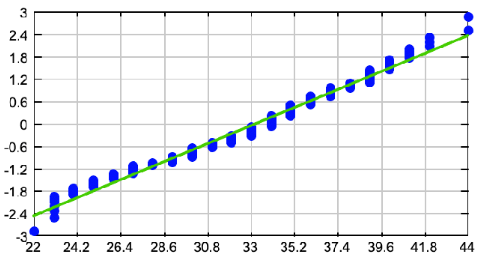




Ryan-Joiner Test
Test statistic, Rp: 0.9959
Critical value for 0.05 significance level: 0.9923
Critical value for 0.01 significance level: 0.989

Fail to reject normality with a 0.05 significance level.
Fail to reject normality with a 0.01 significance level.





Ryan-Joiner Test
Test statistic, Rp: 0.9911
Critical value for 0.05 significance level: 0.9939
Critical value for 0.01 significance level: 0.9914

Reject normality with a 0.05 significance level.
Reject normality with a 0.01 significance level.

BINOMIAL

Number of Side Effects from Medications
Number of Fraudulent Transactions
Number of Spam Emails per Day
Number of River Overflows
Shopping Returns per Week
.
.
.

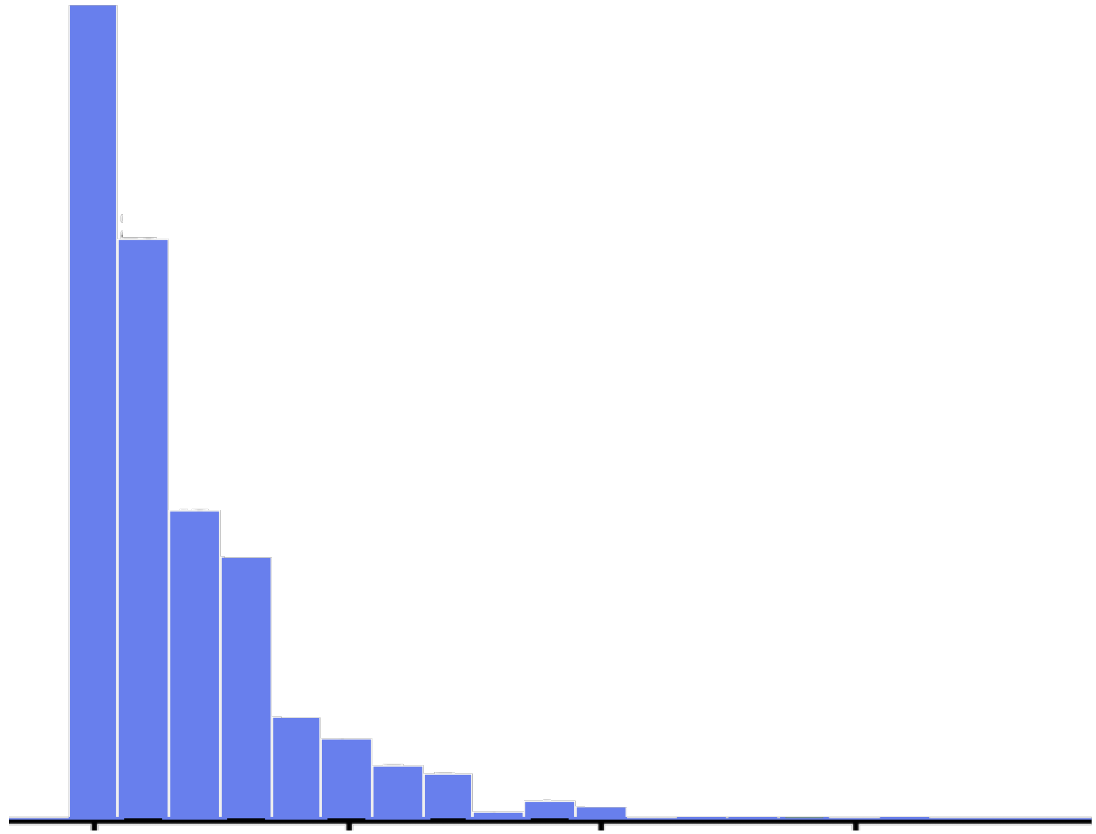
POISSON

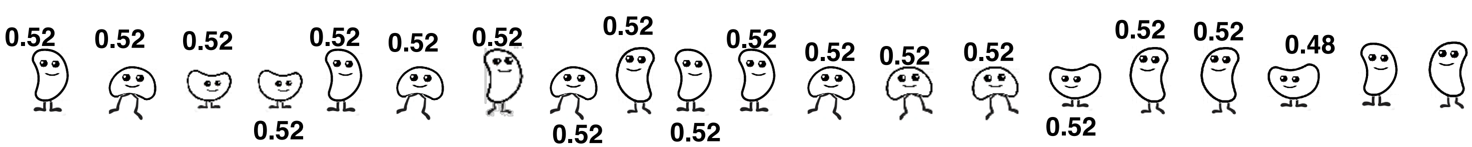
Calls per Hour at a Call Center
Number of Arrivals at a Restaurant
Number of Website Visitors per Hour
Number of Bankruptcies Filed per Month
Number of Network Failures per Week
.
.
.

Areas and Probabilities (Geometric)

- success at location -> 18
- *p* -> 0.48

0.000007





Clinical Side Effects

Sports Betting

Telemarketing Success

Employee Retention

Customer Satisfaction Surveys

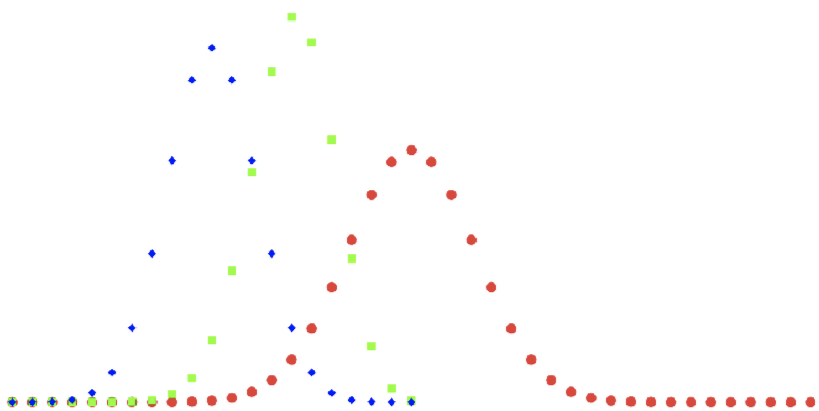
Credit Risk Assessment

Vaccine Efficacy

Binomial Distribution

Fraud Detection

Genetic Trait Inheritance



The binomial probability distribution was discovered by the Swiss mathematician **Jal Bernoulli** (also known as James Bernoulli) his posthumously published work Ars Conjectandi in 1713.

Call center success rates

Agree/Disagree polling

Online Course Completion

Disease testing prevalence

Online Ad Click-Through Rates

Customer Churn

Weather Event Prediction

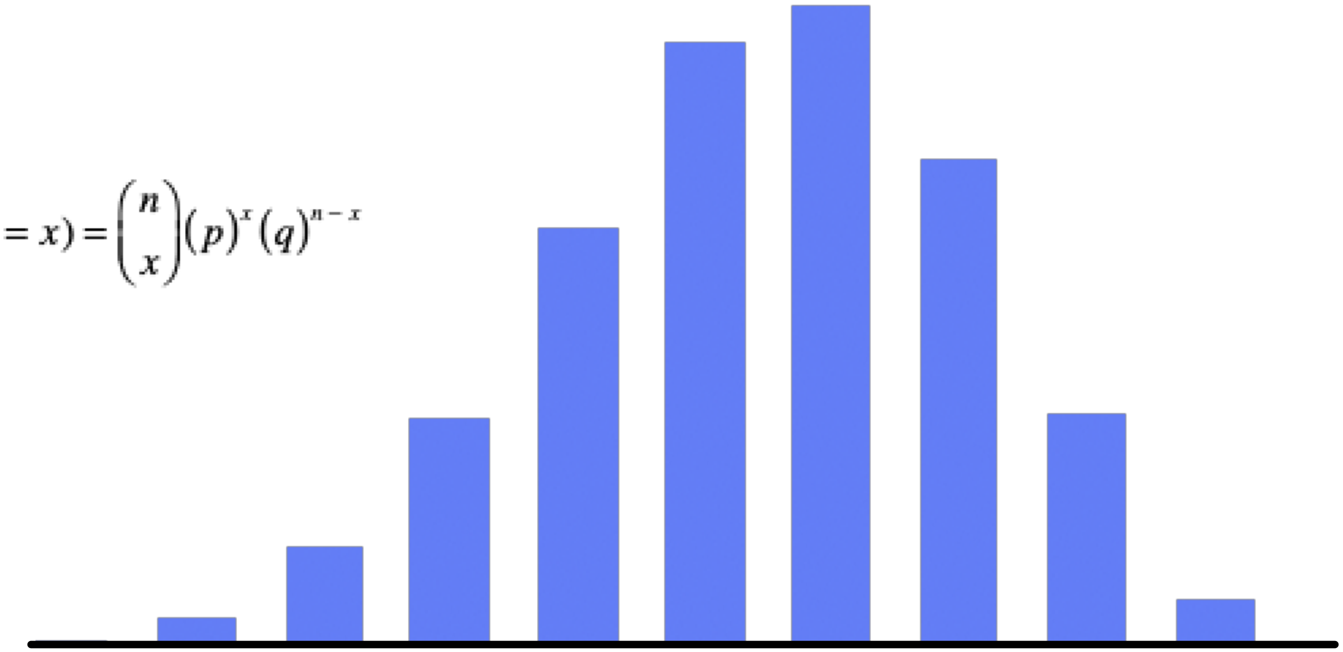
Customer Feedback

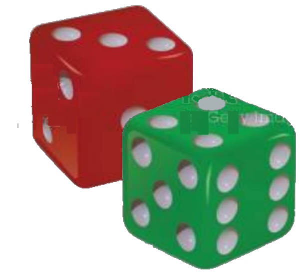
Supply Chain Delays

Product Sampling

Areas and Probabilities (Binomial)

$$p(X = x) = \binom{n}{x}(p)^x(q)^{n-x}$$

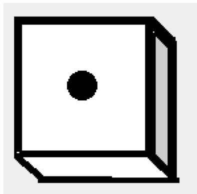




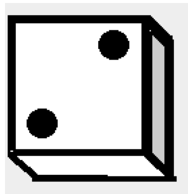
Roll	Occurrences	Probabilit	
2	1	1/36	0.0
3	2	2/36	0.0
4	3	3/36	0.0
5	4	4/36	0.1
6	5	5/36	0.1
7	6	6/36	0.1
8	5	5/36	0.1
9	4	4/36	0.1
10	3	3/36	0.0
11	2	2/36	0.0
12	1	1/36	0.0

- **Find the probability of "winning" 4 times in 10 rolls.**
- **Find the probability of "winning" at most 3 times out of 10 rolls.**
- **Find the probability of "winning" more than half the time out of 10 rolls.**

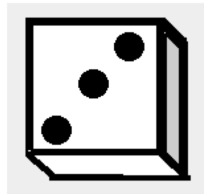
Dice Simulator



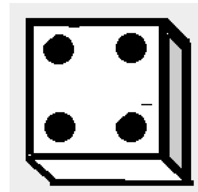
0.166667



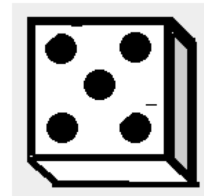
0.166667



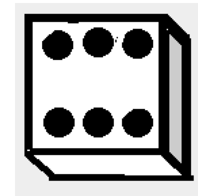
0.166667



0.166667



0.166667



0.166667

Options

Number of Dice : 2

Number of Rolls : 100

GO

1	1 is not possible.
2	2
3	3
4	10
5	9
6	15
7	23
8	13
9	7
10	10
11	3
12	5
13	13 is not possible.
14	
15	
16	
17	
18	
19	
:	

○ **Unbiased**

○ **Biased**

☐ **Show Rolls**

1-2-3-4-5-6

Coin Simulator

How many occurences of 2 heads?

375

Heads: 2020
Tails: 1980

Unbiased

Biased

Options

Number of Coins : 4

Number of Trials : 1000

GO

H,H,H,H	4H, 0T
T,H,T,T	1H, 3T
T,H,T,T	1H, 3T
H,T,T,T	1H, 3T
T,T,T,H	1H, 3T
H,H,T,T	2H, 2T
T,H,T,T	1H, 3T
T,T,H,T	1H, 3T
H,T,T,H	2H, 2T
H,H,T,H	3H, 1T
T,T,T,H	1H, 3T
H,H,T,T	2H, 2T
H,H,T,H	3H, 1T
H,T,T,H	2H, 2T
H,T,T,H	2H, 2T
T,T,H,T	1H, 3T
T,H,H,T	2H, 2T
H,H,T,H	3H, 1T
T,T,T,H	1H, 3T
T,H,H,T	2H, 2T

Show Flips

Heads

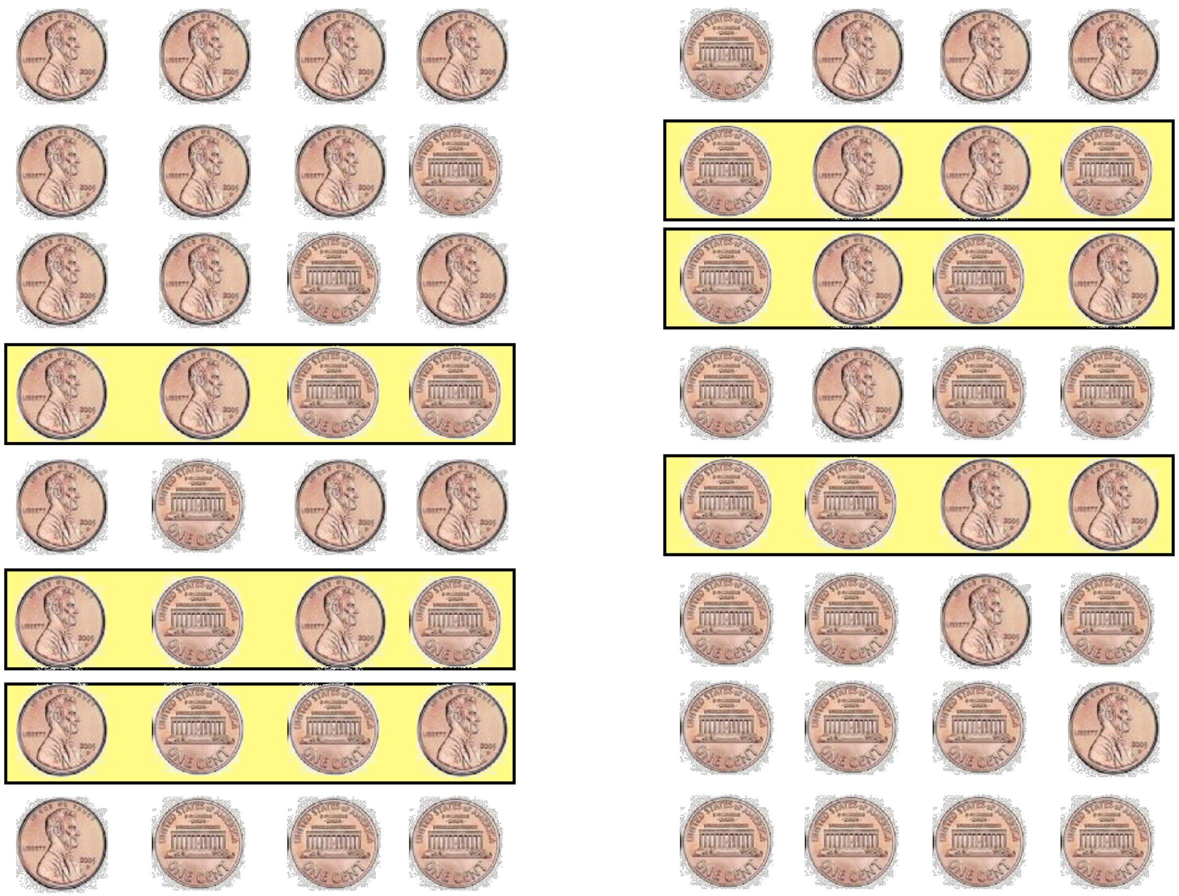
0.5

Tails

0.5

$$p(X = x) = \binom{n}{x}(p)^x(q)^{n-x}$$

Areas and Probabilities



$$\binom{n}{x} = \frac{n!}{x!(n - x!)}$$

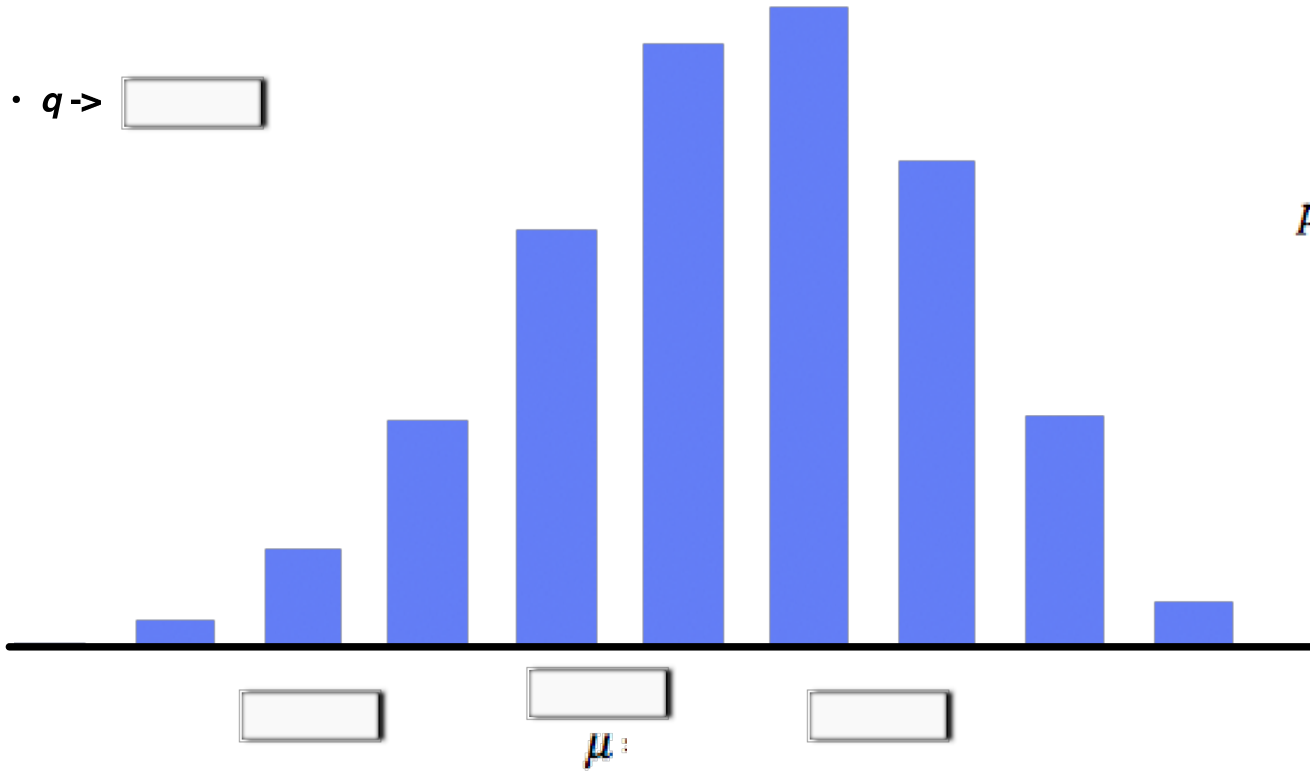
$$\binom{4}{2} = \frac{4!}{2!(4-2)!} = \frac{\overset{2}{4} \cdot 3 \cdot 2 \cdot 1}{2 \cdot 1(2 \cdot 1)}$$

$$p(X = x) = \binom{n}{x}(p)^{x}(q)^{n-x}$$

$$p(X = 2) = \binom{4}{2}\left(\frac{1}{2}\right)^{2}\left(\frac{1}{2}\right)^{2} = 6 \cdot \frac{1}{4} \cdot \frac{1}{4} = \frac{6}{16} = 0.375$$

Areas and Probabilities (Binomial)

- n -> 100
- p -> 0.6
- q ->



$$pmf = \binom{n}{x} p^x q^{n-x}$$

$$\mu = np$$

$$\sigma = \sqrt{npq}$$

$p(X = 50) = 0.010338$

$p(X \leq 50) = 0.027099$

$p(X > 50) = 0.972901$

$p(X = 65) = 0.049133$

$p(X \leq 65) = 0.869663$

$p(X > 65) = 0.130337$

$p(50 < X < 65) = 0.842564$

Mutations in a DNA sequence

Flood occurrences

Lightning strikes in an area

Meteor showers

Pedestrian crossings at intersections

Patient arrivals at an emergency

Fire outbreaks in a district

Disease outbreaks in a region

Animal sightings in a national park

Customer complaints

Earthquakes in a region

Poisson Distribution

Hospital readmissions

Spam email arrivals

The Poisson probability distribution was discovered by the French mathematician and physicist **Siméon Denis Poisson** in 1837. He introduced the distribution to model the probability of rare events occurring over a fixed interval, such as the number of wrongful convictions in a given population

Network packet arrivals

Online order placements

Medical equipment failures

ext message arrivals

Bus arrivals at a stop

Power outages in a grid

Traffic light failures

Prescription errors in a pharmacy

Elevator breakdowns in a building

Cosmic ray hits on a detector

Parking violations in a city zone

Phone calls to a call center

Tree falls in a forest

Police emergency calls

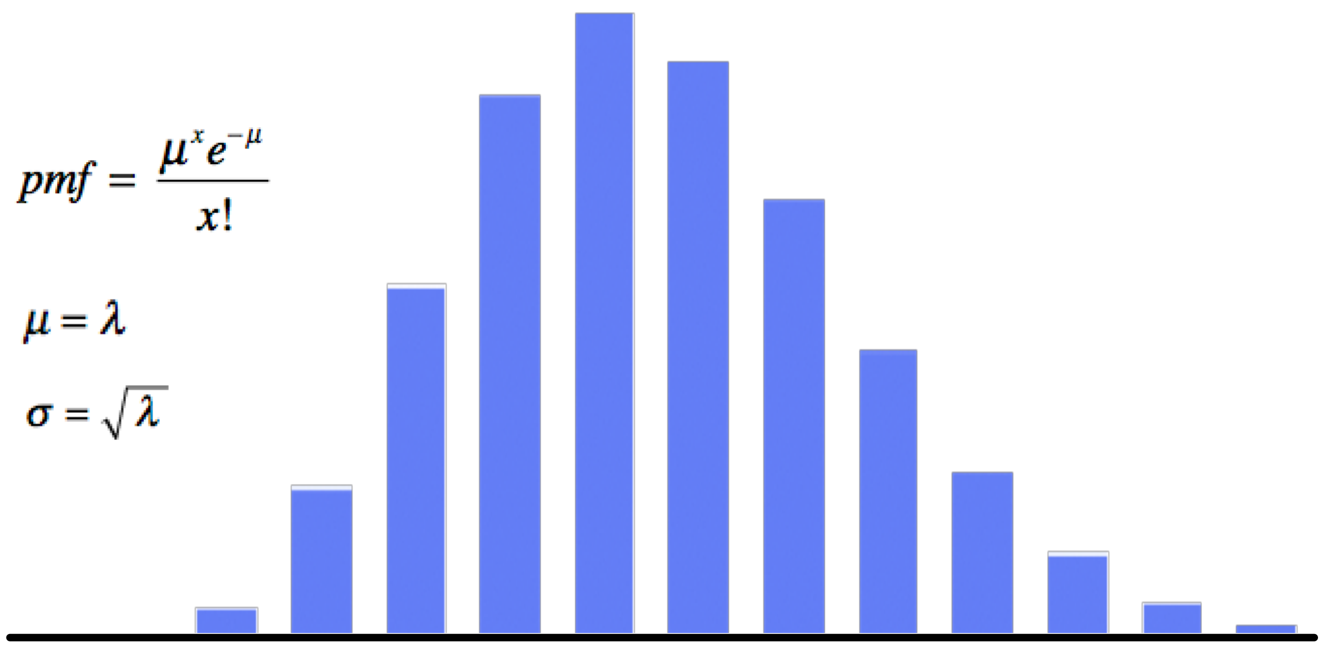
Areas and Probabilities (Poisson)

- λ -> 10
- iterations -> 20

$$pmf = \frac{\mu^x e^{-\mu}}{x!}$$

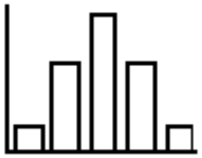
$$\mu = \lambda$$

$$\sigma = \sqrt{\lambda}$$



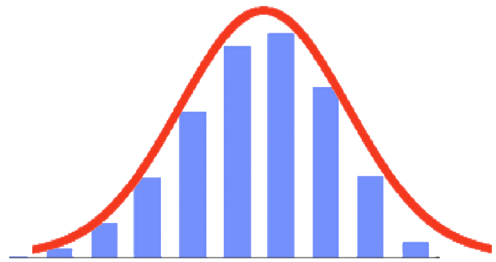
x	$P(X) = x$	$P(X) \le x$	$P(X) >$
0	0.000045	0.000045	0.99995
1	0.000454	0.000499	0.99950
2	0.00227	0.002769	0.99723
3	0.007567	0.010336	0.98966
4	0.018917	0.029253	0.97074
5	0.037833	0.067086	0.93291
6	0.063055	0.130141	0.86985
7	0.090079	0.220221	0.77977
8	0.112599	0.33282	0.66718
9	0.12511	0.45793	0.54207
10	0.12511	0.58304	0.41696
11	0.113736	0.696776	0.30322
12	0.09478	0.791556	0.20844
13	0.072908	0.864464	0.13553
14	0.052077	0.916542	0.08345
15	0.034718	0.95126	0.04874
16	0.021699	0.972958	0.02704
17	0.012764	0.985722	0.01427
18	0.007091	0.992813	0.00718
19	0.003732	0.996546	0.00345
20	0.001866	0.998412	0.00158
21			
22			
23			
24			
25			

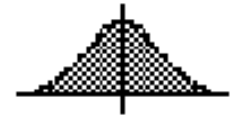
"Continuity" correction:



Binomial Distribution

when $npq \geq 5$





Normal Distribution

$$pmf = \binom{n}{x} p^x q^{n-x}$$

$$\mu = np$$

$$\sigma = \sqrt{npq}$$

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

Compute the probability that between 50 and 75 of 100 white blood cells will be neutrophils (white blood cells which form an essential part of the innate immune system). The probability that one cell is a neutrophil is 0.6

$$p(50 \leq x \leq 75)$$

$$\sum_{k=50}^{75} \text{Binomial}[100, k] * (0.6^k) * (0.4)^{100-k}$$

0.982677

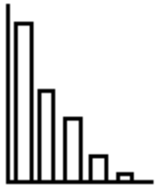
$$N\left[\left(\frac{1}{\sqrt{2\,(24)\,\pi}}\right)\int_{50}^{75} e^{\frac{-(x-60)^2}{2\,(24)}}\,dx\right]$$

0.978287

$$N\left[\left(\frac{1}{\sqrt{2\,(24)\,\pi}}\right)\int_{49.5}^{75.5} e^{\frac{-(x-60)^2}{2\,(24)}}\,dx\right]$$

0.983177

"Continuity" correction:



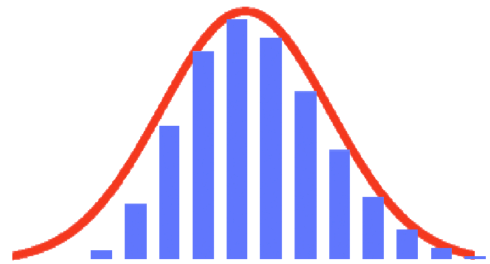
Poisson Distribution

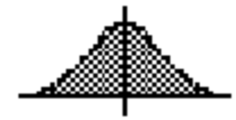
$$pmf = \frac{\mu^x e^{-\mu}}{x!}$$

$\mu = \lambda$

$\sigma = \sqrt{\lambda}$

when $\mu \geq 10$





Normal Distribution

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

Consider the distribution of the number of bacteria in a Petri plate whose area is 100 cm². Assume the probability of observing x bacteria is given by a Poisson distribution with $\lambda = 0.16(100)$. Suppose 20 bacteria are observed in this area. How unusual is this?

$$p(x \geq 19)$$

$$1 - N\left[\sum_{k=0}^{19} \frac{16^k e^{-16}}{k!}\right]$$

0.187751

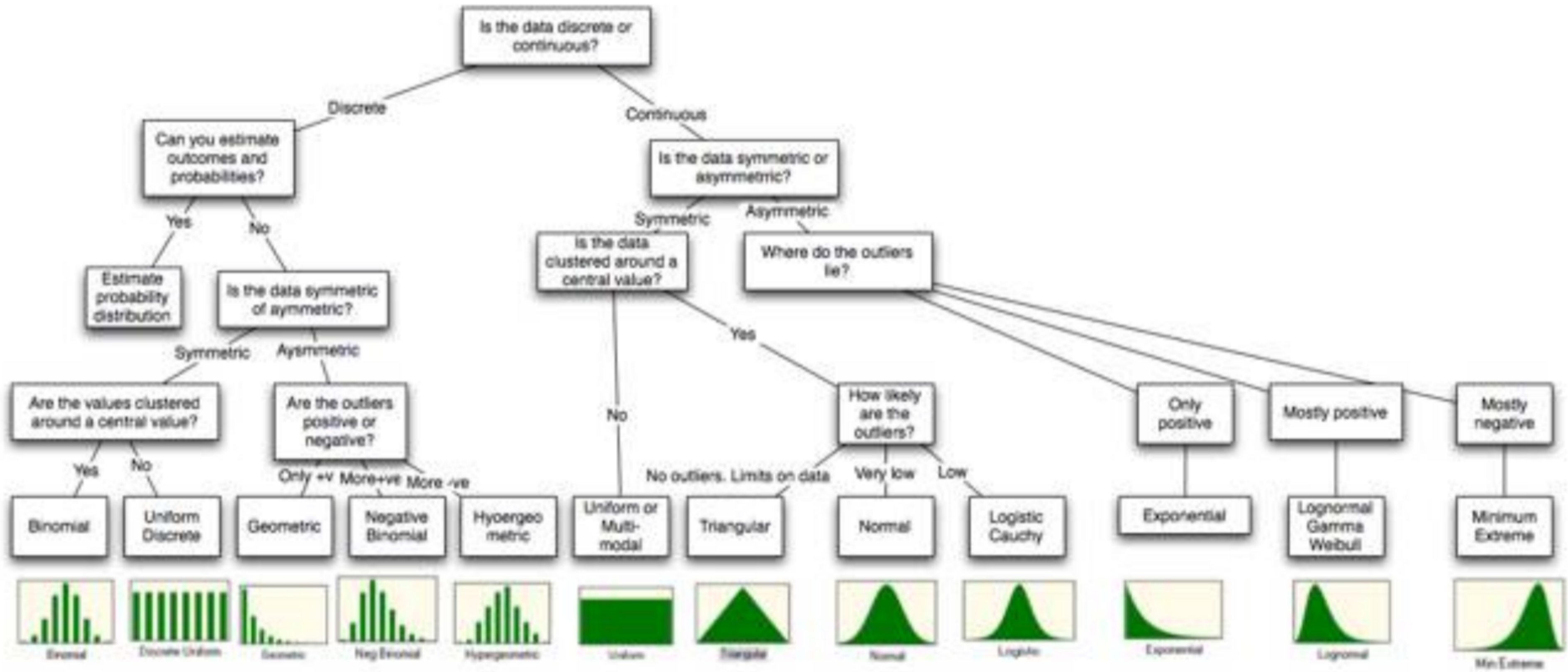
$$1 - N\left[\int_0^{19} \frac{1.0}{\sqrt{16}\ \sqrt{2 * \pi}} e^{\frac{-1}{2}\left(\frac{x-16}{\sqrt{16}}\right)^2} dx\right]$$

0.226659

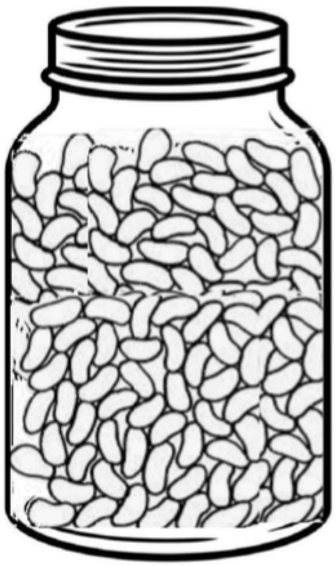
$$1 - N\left[\int_0^{19.5} \frac{1.0}{\sqrt{16}\ \sqrt{2 * \pi}} e^{\frac{-1}{2}\left(\frac{x-16}{\sqrt{16}}\right)^2} dx\right]$$

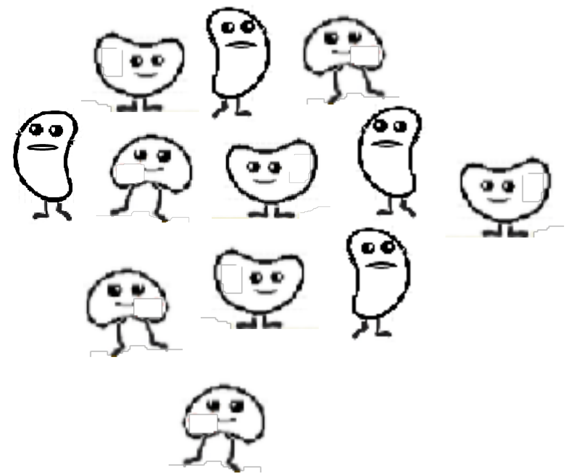
0.190819

Data Distributions vs Probability Distributions



Sampling



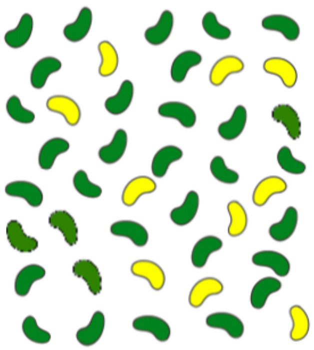


Previous

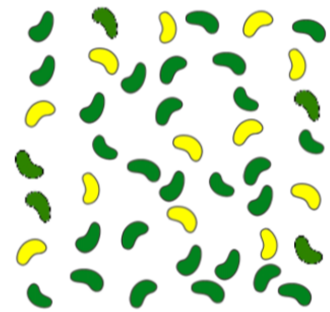
Next

Sampling Strategies

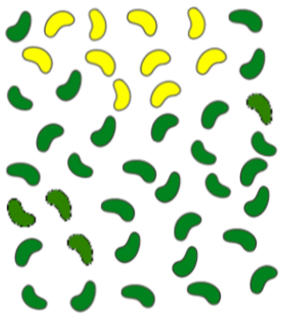
Simple Random

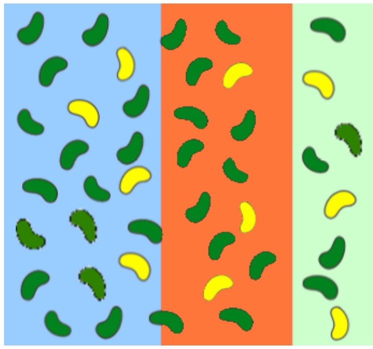


Systematic



Convenience





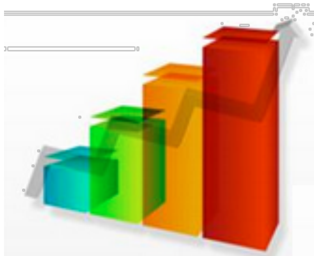
Cluster

Stratified

Previous

Next

Sampling Bias



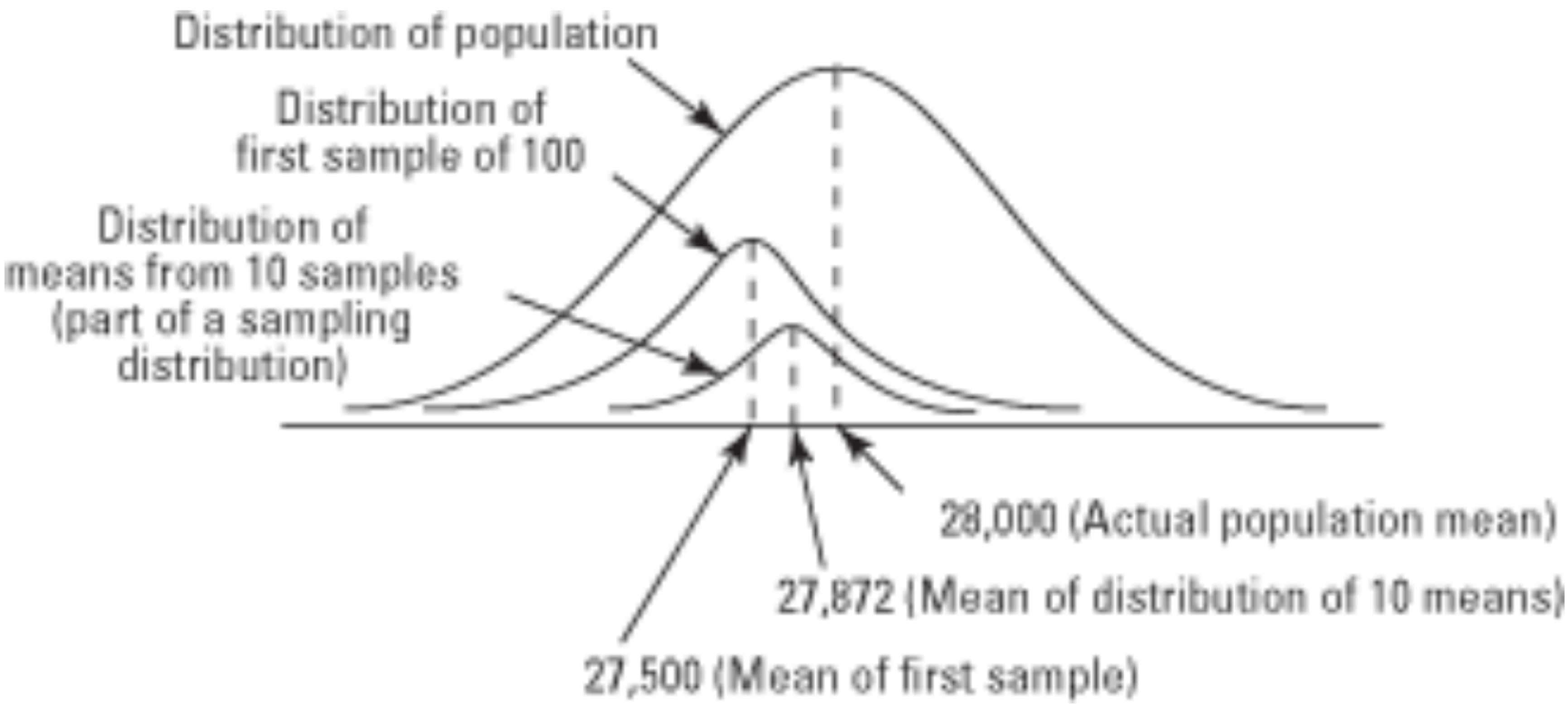
Types of Bias

1. Undercoverage
   - when you inadequately represent some members of your population in the
2. Voluntary Response Bias
   - only viewers who have strong opinions on who should win will participate
3. Convenience Sample Bias
   - the sample is taken from a group of people easy to contact or to reach
4. Nonresponse Bias
   - snail mail survey for young adults or a smartphone survey for older adults
5. Response Bias
   - they may feel pressure to give answers that are socially acceptable
6. Social-desirability Bias
   - "Trump deniabilty syndrome"

Previous

Next

Taking Samples



The Central Limit Theorem

The sample mean will be approximately normally distributed for large sample sizes, *regardless of the distribution from which we are sampling.*

$$\overline{X} \sim N\left(\mu_{\overline{X}}, \sigma_{\overline{X}}^2\right) \sim N\left(\mu, \frac{\sigma^2}{n}\right)$$

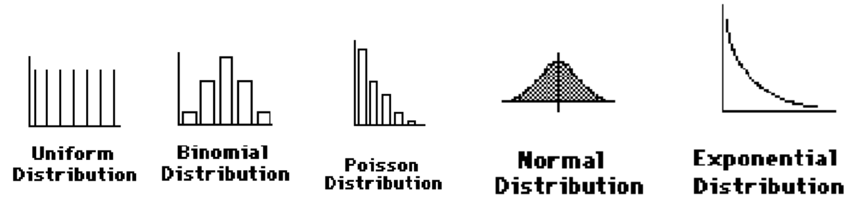
Sampling Distribution (Means)

Population Size: 1000

Histogram

Number of intervals:

- (•) 10
- ( ) 20



MU :	47.833
SIGMA :	28.451909
VAR :	809.51111

μ ->	23
σ ->	12
n ->	20
p ->	0.7
$\lambda \rightarrow$	4

9
12
5
86
40
59
56
41
67
90
28
25
41
87
45
57
52
66
6
53
75
9
50
59
18
54
23
41
23

Sample. Size:	40
Number of Samples:	200
Mean of Sampling Distribution:	47.235
Standard Deviation of Sampling Distribution:	4.73284
Standard Error: (Calculated)	4.498642

$$\sigma^* = \frac{\sigma}{\sqrt{n}}$$

SAMPLE

The Sample Means ->

41
60
46
49
40
47
48

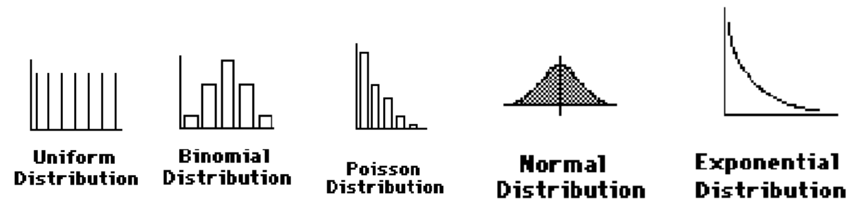
Sampling Distribution (Variances)

Histogram

Population Size: 250

Number of intervals:

- (•) 10
- ( ) 20



MU :	46.000152
SIGMA :	4.948392
VAR :	24.486585

μ ->	45
σ ->	5
n ->	23
p ->	.3
$\lambda \rightarrow$	4

Sample. Size:	20
Number of Samples:	1000
Mean of Sampling Distribution:	27.77631
Standard Deviation of Sampling Distribution:	8.002262
Standard Error: (Calculated)	1.106494

38.7
43.8
47.2
46.3
42.9
42.4
39.8
48.1
42.5
41.4
44.8
43.0
46.3
41.7
43.0
50.3
48.2
50.4
40.8
40.7
47.5
40.7
44.4
42.3
51.8
52.5
47.9
34.6
39.7

SAMPLE

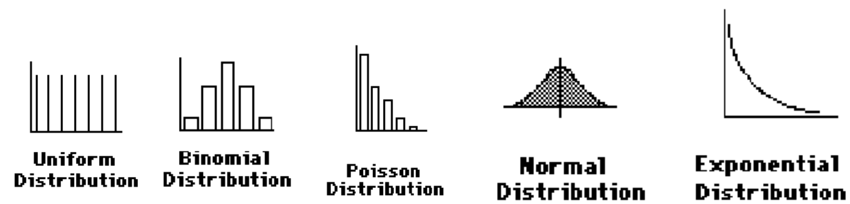
Sampling Distribution (Standard Deviations)

Population Size: 1000

Histogram

Number of intervals:

- ○ 10
- ◉ 20



MU :	51.807
SIGMA :	28.967115
VAR :	839.09375

μ ->	45
σ ->	5
n ->	23
p ->	.3
$\lambda \rightarrow$	4

42
15
9
21
75
39
6
40
14
92
15
81
93
7
16
33
96
70
91
89
8
90
25
38
78
72
43
90
73

Sample. Size:	10
Number of Samples:	400
Mean of Sampling Distribution:	31.442935
Standard Deviation of Sampling Distribution:	4.394522
Standard Error: (Calculated)	9.160206

SAMPLE

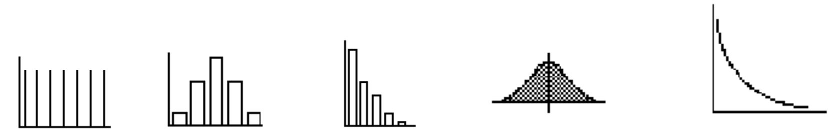
Sampling Distribution (Ranges)

Population Size: 200

Histogram

Number of intervals:

- 10
- 20



Uniform Distribution | Binomial Distribution | Poisson Distribution | Normal Distribution | Exponential Distribution

MU :	0.249429
SIGMA :	0.227932
VAR :	0.051953
RANGE :	1.316969

μ ->	13
σ ->	12
n ->	25
p ->	.7
$\lambda \rightarrow$	4

Sample. Size:	5
Number of Samples:	1000
Mean of Sampling Distribution:	0.427265
Standard Deviation of Sampling Distribution:	0.216379
Standard Error: (Calculated)	0.101934

SAMPLE

0.19
0.79
0.37
0.20
0.50
0.27
0.46
0.09
0.45
0.06
0.04
0.12
0.76
0.34
0.25
0.22
0.01
0.11
0.05
0.18
0.39
0.08
0.01
0.03
0.47
0.04
0.00
0.09
0.07

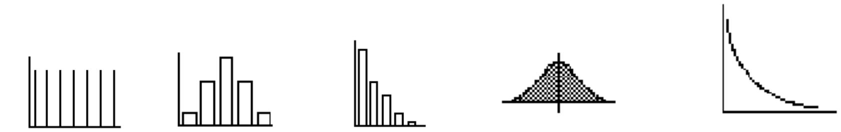
Sampling Distribution (Medians)

Histogram

Population Size: 250

Number of intervals:

- ◯ 10
- ◯ 20



Uniform Distribution | Binomial Distribution | Poisson Distribution | Normal Distribution | Exponential Distribution

MU :	23.32
SIGMA :	5.26665
VAR :	27.7376
MEDIAN:	27.7376

μ ->	23
σ ->	12
n ->	23
p ->	.3
$\lambda \rightarrow$	4

Sample. Size:	80
Number of Samples:	100
Mean of Sampling Distribution:	22
Standard Deviation of Sampling Distribution:	0.648074
Standard Error: (Calculated)	0.588829

SAMPLE

24
30
25
29
28
29
19
24
20
20
30
17
22
22
31
28
24
22
18
18
32
18
24
12
24
21
32
24
22