Inspire...Educate...Transform.






Inspire…Educate…Transform.

Foundations of Statistics and Probability for Data Science

Probability Distributions: Discrete and Continuous, Sampling Distribution of Means, CLT

Dr. Sridhar Pappu
Executive VP – Academics, INSOFE

June 16, 2018

Analyzing attributes

PROBABILITY DISTRIBUTIONS

Random Variable

- A variable that can take multiple values with different probabilities.
- The mathematical function describing these possible values along with their associated probabilities is called a probability distribution.

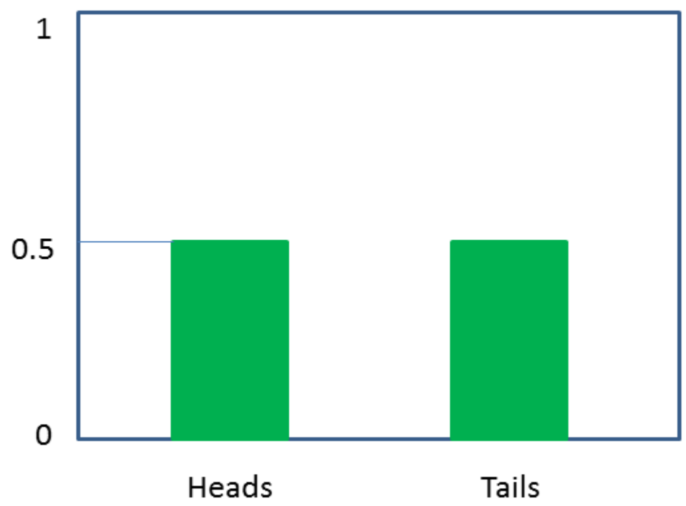
Points scored per game	**0**	**1**	**2**	**3**	**4**	**5**	**6**
Frequency, f	1	4	6	12	5	1	1

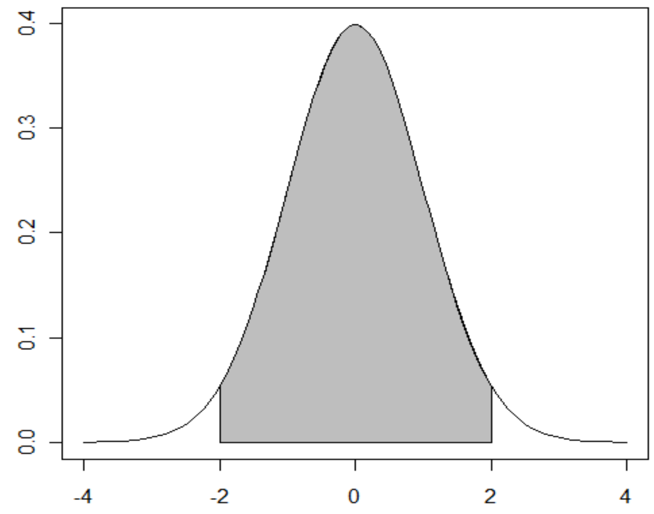
Points scored per game	**0**	**1**	**2**	**3**	**4**	**5**	**6**
Probability *Recall the Frequentist (empirical) approach of assigning probabilities*	$\frac{1}{30}$	$\frac{4}{30}$	$\frac{6}{30}$	$\frac{12}{30}$	$\frac{5}{30}$	$\frac{1}{30}$	$\frac{1}{30}$

Leads to Descriptive Stats

Leads to Inferential Stats

Discrete and Continuous





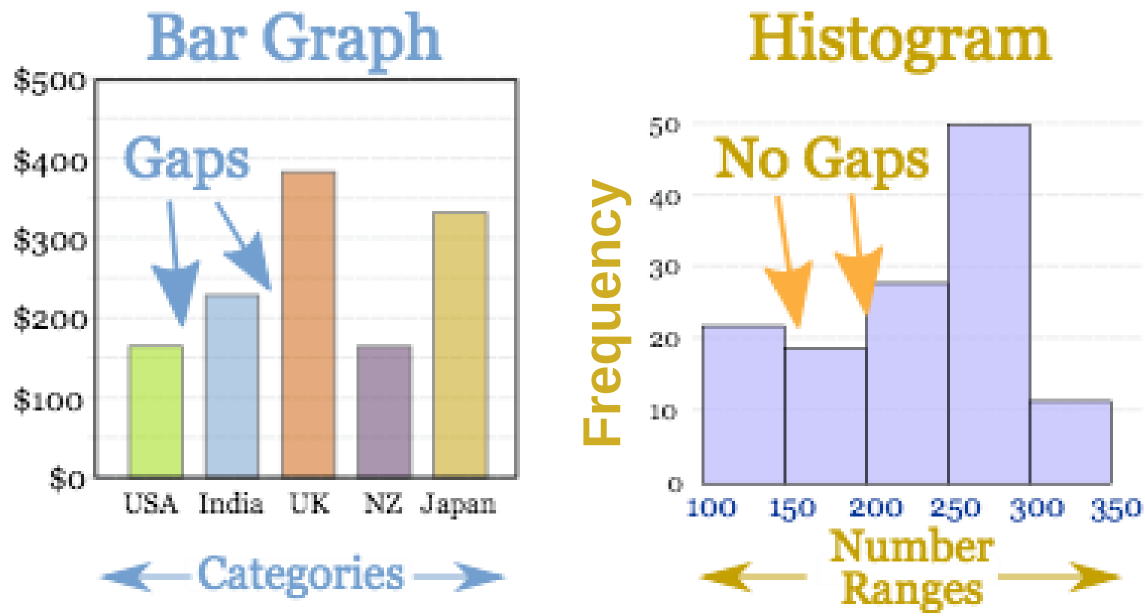
Countable

Measurable

Can any function be a probability distribution?

Discrete Distributions	**Continuous Distributions**
Probability that X can take a specific value x is $P(X = x) = p(x)$.	Probability that X is between two points a and b is $P(a \leq X \leq b) = \int_a^b f(x)dx$.
It is non-negative for all real x.	It is non-negative for all real x.
The sum of $p(x)$ over all possible values of x is 1, i.e., $\sum p(x) = 1$.	$\int_{-\infty}^{\infty} f(x)dx = 1$
Probability Mass Function	**Probability Density Function**

Histogram



A series of **contiguous rectangles** that represent the frequency of data in given class intervals.

How many class intervals?

- Rule of thumb: 5-15 (not too many and not too few)
- Freedman-Diaconis rule:

$$No.of\ bins = \frac{(max - min)}{2 * IQR * n^{\frac{-1}{3}}},$$

$$where\ the\ denominator\ is\ the\ bin - width$$

Histogram - Excel

Annual traffic data for 30 busiest airports in the world – 2013 and 2011

Source: http://www.aci.aero/Data-Centre/Annual-Traffic-Data/Passengers/2011-final and http://www.aci.aero/Data-Centre/Annual-Traffic-Data/Passengers/2013-final

Last accessed: February 04, 2016

Passenger Traffic 2011 FINAL (Annual)
Last Update: 8 July 2013
Passenger Traffic
Total passengers enplaned and deplaned, passengers in transit counted once

Rank	City (Airport)	Total Passengers	% Change
1	ATLANTA GA, US (ATL)	92389023	3.5
2	BEIJING, CN (PEK)	78675058	6.4
3	LONDON, GB (LHR)	69433565	5.4
4	CHICAGO IL, US (ORD)	66701241	-0.1
5	TOKYO, JP (HND)	62584826	-2.5
6	LOS ANGELES CA, US (LAX)	61862052	4.7
7	PARIS, FR (CDG)	60970551	4.8
8	DALLAS/FORT WORTH TX, US (DFW)	57832495	1.6
9	FRANKFURT, DE (FRA)	56436255	6.5
10	HONG KONG, HK (HKG)	53328613	5.9
11	DENVER CO, US (DEN)	52849132	1.7
12	JAKARTA, ID (CGK)	51533187	16.2
13	DUBAI, AE (DXB)	50977960	8
14	AMSTERDAM, NL (AMS)	49755252	10
15	MADRID, ES (MAD)	49653055	-0.4
16	BANGKOK, TH (BKK)	47910904	12
17	NEW YORK NY, US (JFK)	47644060	2.4
18	SINGAPORE, SG (SIN)	46543845	10.7
19	GUANGZHOU, CN (CAN)	45040340	9.9
20	SHANGHAI, CN (PVG)	41447730	2.1
21	SAN FRANCISCO CA, US (SFO)	40927786	4.3
22	PHOENIX AZ, US (PHX)	40591948	5.3
23	LAS VEGAS NV, US (LAS)	40560285	2
24	HOUSTON TX, US (IAH)	40128953	-0.9
25	CHARLOTTE NC, US (CLT)	39043708	2.1
26	MIAMI FL, US (MIA)	38314389	7.3
27	MUNICH, DE (MUC)	37763701	8.8
28	KUALA LUMPUR, MY (KUL)	37704510	10.6
29	ROME, IT (FCO)	37651222	3.9
30	ISTANBUL, TR (IST)	37406025	16.3

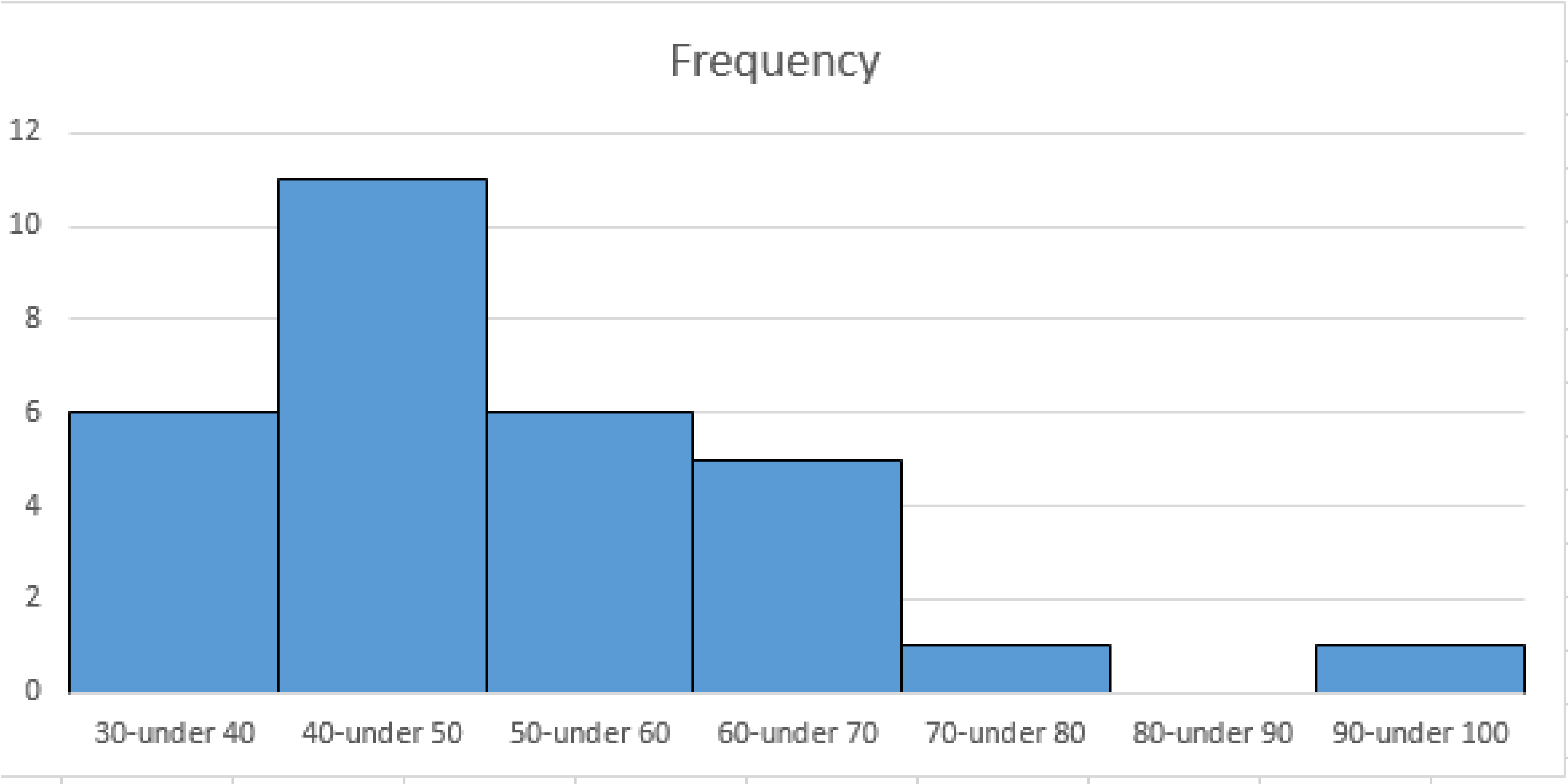
Passenger Traffic 2013 FINAL (Annual)
Last Update: 22 December 2014
Passenger Traffic
Total passengers enplaned and deplaned, passengers in transit counted once

Rank	City (Airport)	Passengers 2013	Passengers 2012	% Change
1	ATLANTA GA, US (ATL)	9,44,31,224	9,55,13,828	-1.1
2	BEIJING, CN (PEK)	8,37,12,355	8,19,29,359	2.2
3	LONDON, GB (LHR)	7,23,68,061	7,00,38,804	3.3
4	TOKYO, JP (HND)	6,89,06,509	6,67,95,178	3.2
5	CHICAGO IL, US (ORD)	6,67,77,161	6,66,29,600	0.2
6	LOS ANGELES CA, US (LAX)	6,66,67,619	6,36,88,121	4.7
7	DUBAI, AE (DXB)	6,64,31,533	5,76,84,550	15.2
8	PARIS, FR (CDG)	6,20,52,917	6,16,11,934	0.7
9	DALLAS/FORT WORTH TX, US (DFW)	6,04,70,507	5,86,20,160	3.2
10	JAKARTA, ID (CGK)	6,01,37,347	5,77,72,864	4.1
11	HONG KONG, HK (HKG)	5,95,88,081	5,60,61,595	6.3
12	FRANKFURT, DE (FRA)	5,80,36,948	5,75,20,001	0.9
13	SINGAPORE, SG (SIN)	5,37,26,087	5,11,81,804	5
14	AMSTERDAM, NL (AMS)	5,25,69,200	5,10,35,590	3
15	DENVER CO, US (DEN)	5,25,56,359	5,31,56,278	-1.1
16	GUANGZHOU, CN (CAN)	5,24,50,262	4,83,09,410	8.6
17	BANGKOK, TH (BKK)	5,13,63,451	5,30,02,328	-3.1
18	ISTANBUL, TR (IST)	5,13,04,654	4,51,23,758	13.7
19	NEW YORK NY, US (JFK)	5,04,23,765	4,92,91,765	2.3
20	KUALA LUMPUR, MY (KUL)	4,74,98,127	3,98,87,866	19.1
21	SHANGHAI, CN (PVG)	4,71,89,849	4,48,80,164	5.1
22	SAN FRANCISCO CA, US (SFO)	4,49,45,760	4,43,99,885	1.2
23	CHARLOTTE NC, US (CLT)	4,34,57,471	4,12,28,372	5.4
24	INCHEON, KR (ICN)	4,16,79,758	3,91,54,375	6.4
25	LAS VEGAS NV, US (LAS)	4,09,33,037	4,07,99,830	0.3
26	MIAMI FL, US (MIA)	4,05,62,948	3,94,67,444	2.8
27	PHOENIX AZ, US (PHX)	4,03,41,614	4,04,48,932	-0.3
28	HOUSTON TX, US (IAH)	3,97,99,414	3,98,91,444	-0.2
29	MADRID, ES (MAD)	3,97,17,850	4,51,76,978	-12.1
30	MUNICH, DE (MUC)	3,86,72,644	3,83,60,604	0.8

Histogram

Annual traffic data for 30 busiest airports in the world – 2011

Source: http://www.aci.aero/Data-Centre/Annual-Traffic-Data/Passengers/2011-final
Last accessed: November 22, 2014

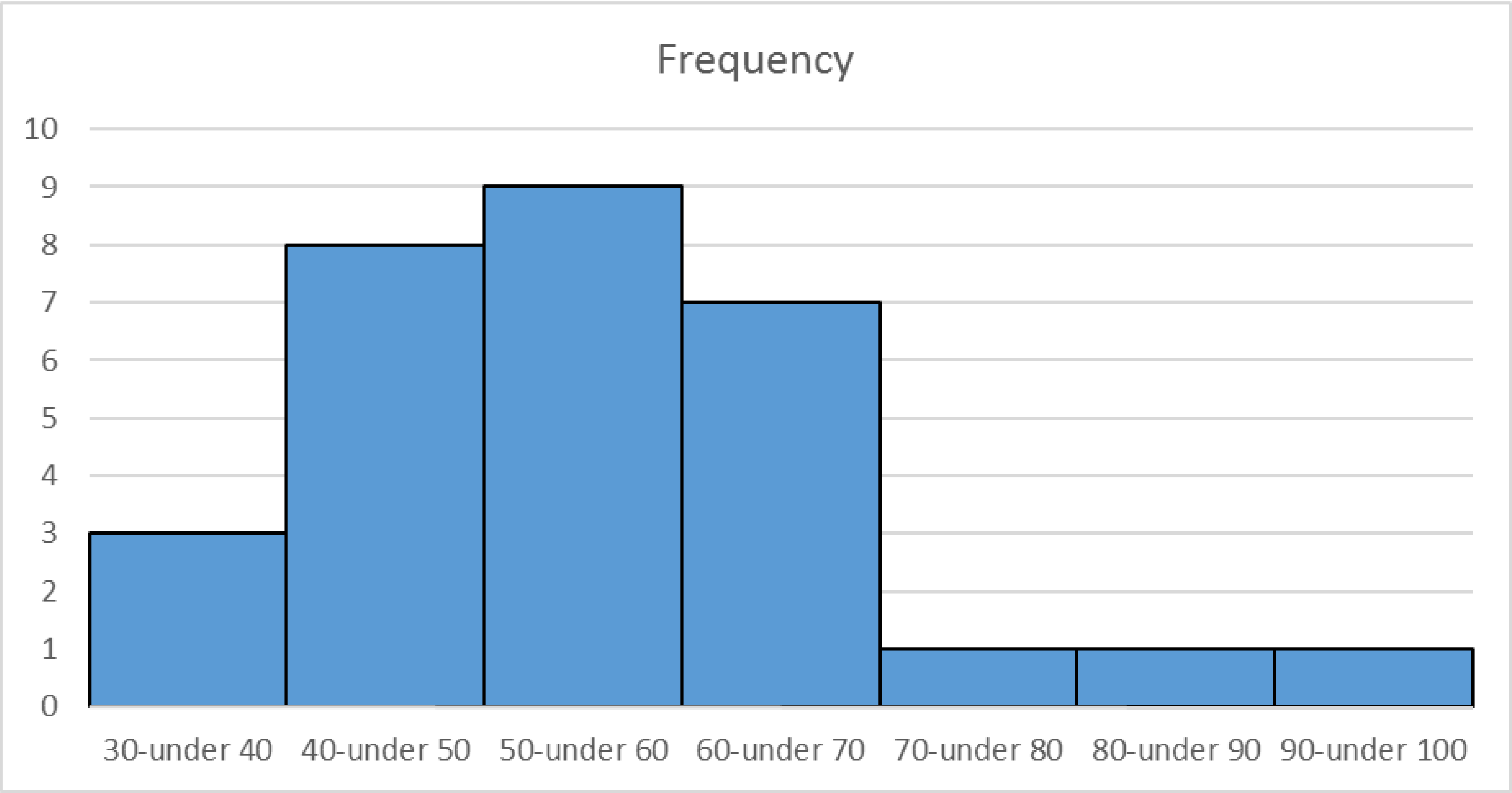


CSE 7315c

Histogram

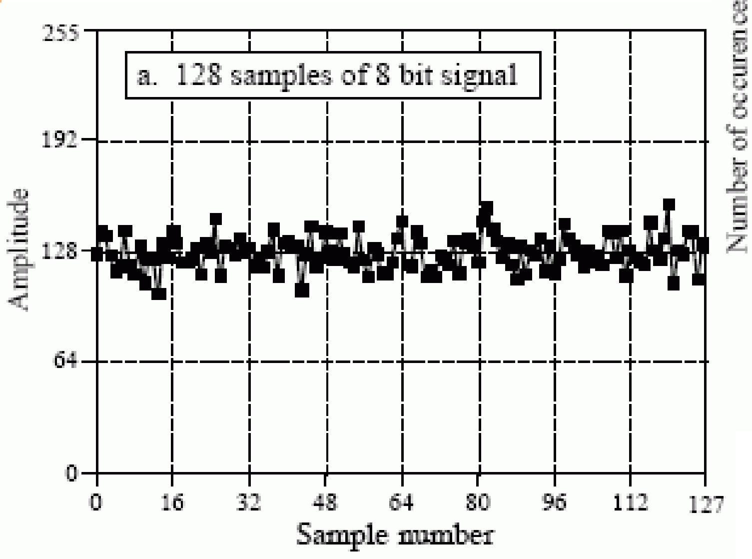
Annual traffic data for 30 busiest airports in the world – 2013

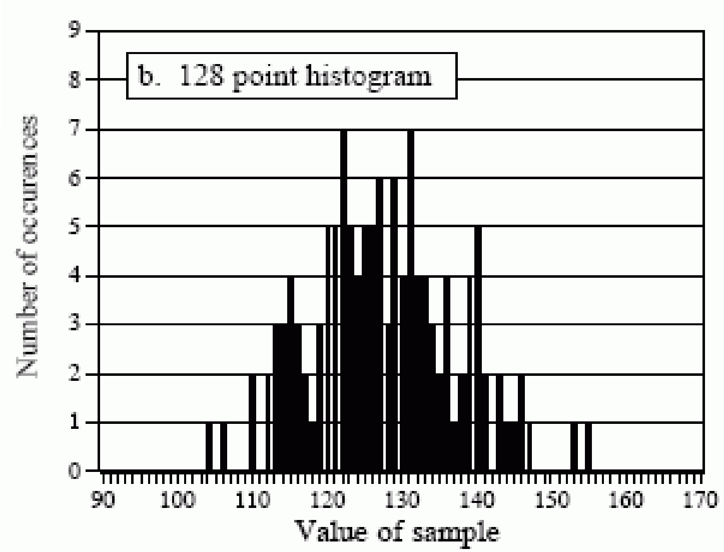
Source: http://www.aci.aero/Data-Centre/Annual-Traffic-Data/Passengers/2013-final
Last accessed: February 04, 2016

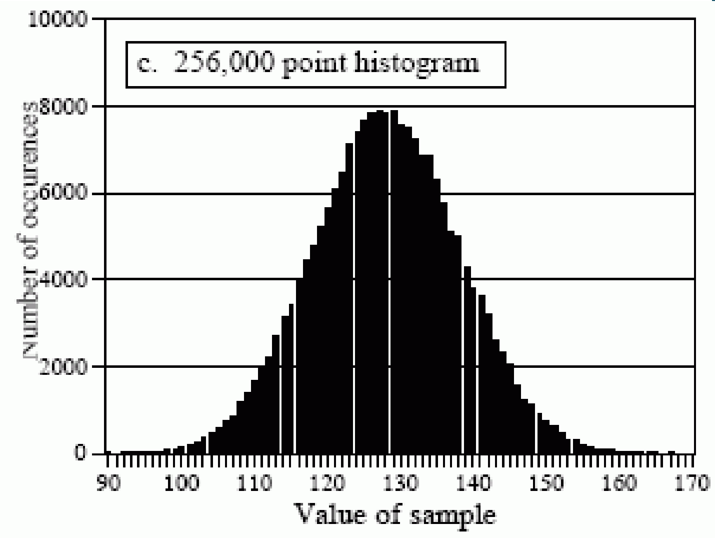


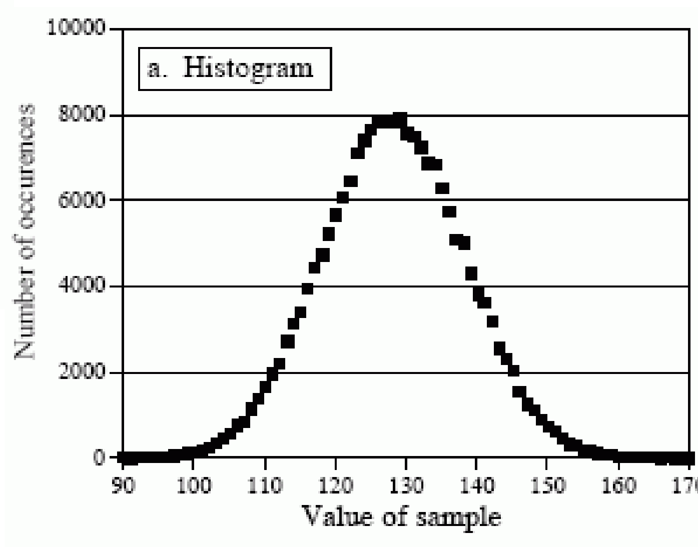
Histogram, PMF and PDF

Signal from an 8-bit analog-to-digital converter attached to a computer, e.g., 0-255 mV converted to digital numbers between 0 and 255.



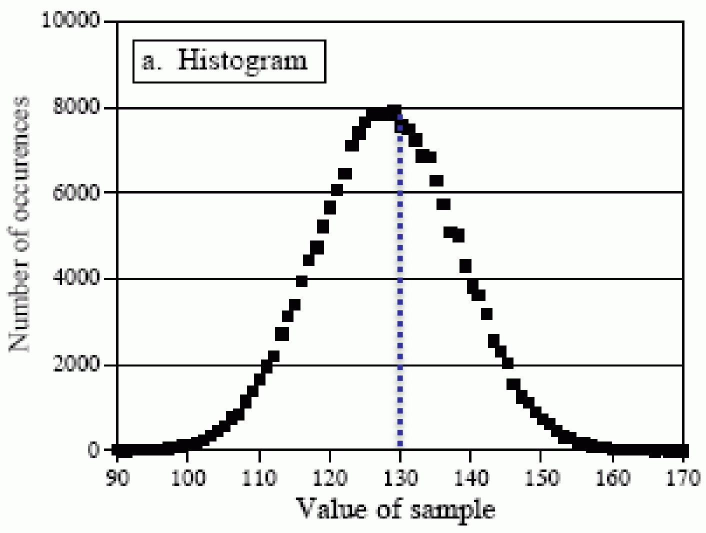


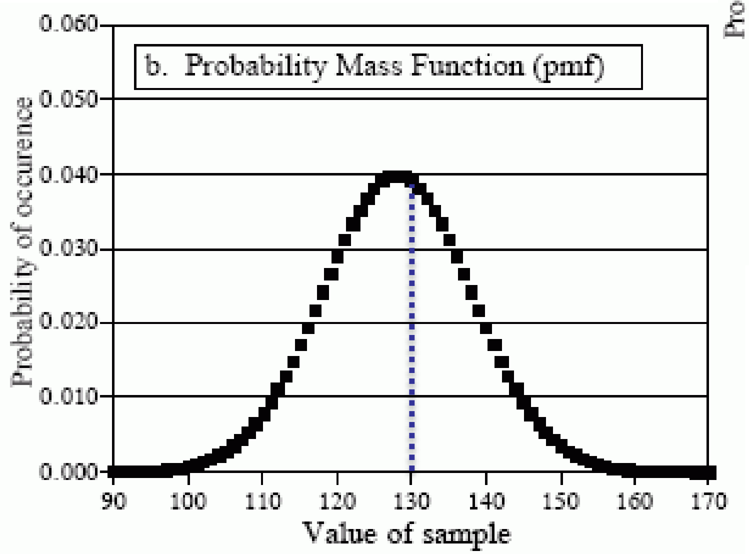


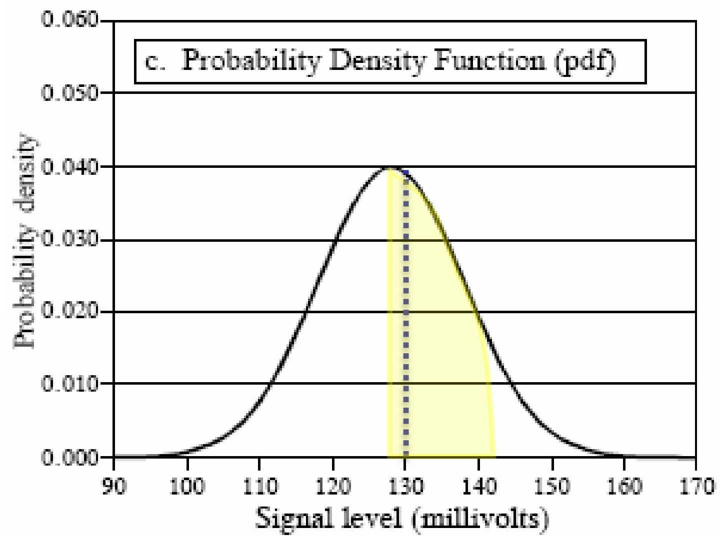


Histogram, PMF and PDF

Signal from an 8-bit analog-to-digital converter attached to a computer, e.g., 0-255 mV converted to digital numbers between 0 and 255.







Possible Outcome	$	Cherry	Lemon	Other
Probability of Outcome	0.1	0.2	0.2	0.5

Cost: $1 for each game

Winning combinations:

 = $20

 = $15 (any order)

 = $10

 = $5

Probability Distribution of Winnings

Combination	None	Lemons	Cherries	Dollars/Cherry	Dollars
Probability	0.977	0.008	0.008	0.006	0.001
Gain	-$1	$4	$9	$14	$19

Cost: $1 for each game

Winning combinations:

= $20

= $15 (any order)

= $10

= $5

Probability Distributions of Winnings and Income

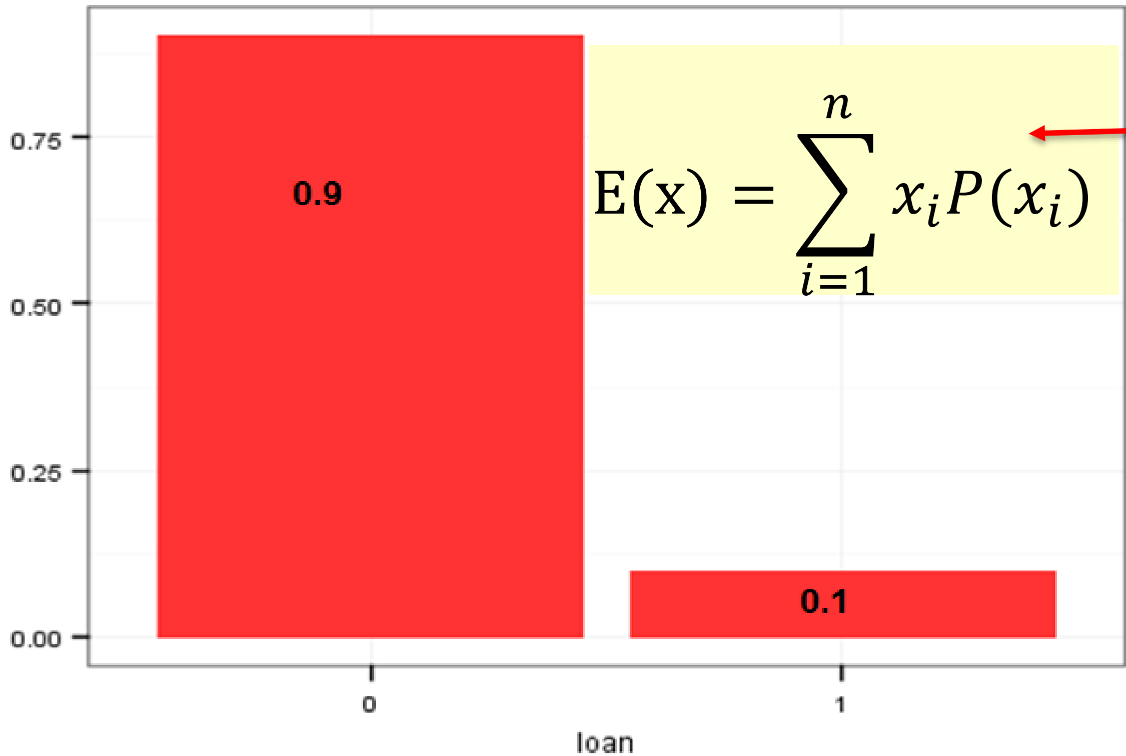
Combination	None	Lemons	Cherries	Dollars/Cherry	Dollars
Probability	0.977	0.008	0.008	0.006	0.001
Gain	-$1	$4	$9	$14	$19

Probability	0.43	0.04	0.43	0.09
Income (BHD)	100	345	1000	9833

Why do you need a probability distribution?

Once a distribution is calculated, it can be used to determine the EXPECTED outcome.

Expectation: Discrete



$$E(x) = \sum_{i=1}^{n} x_i P(x_i)$$

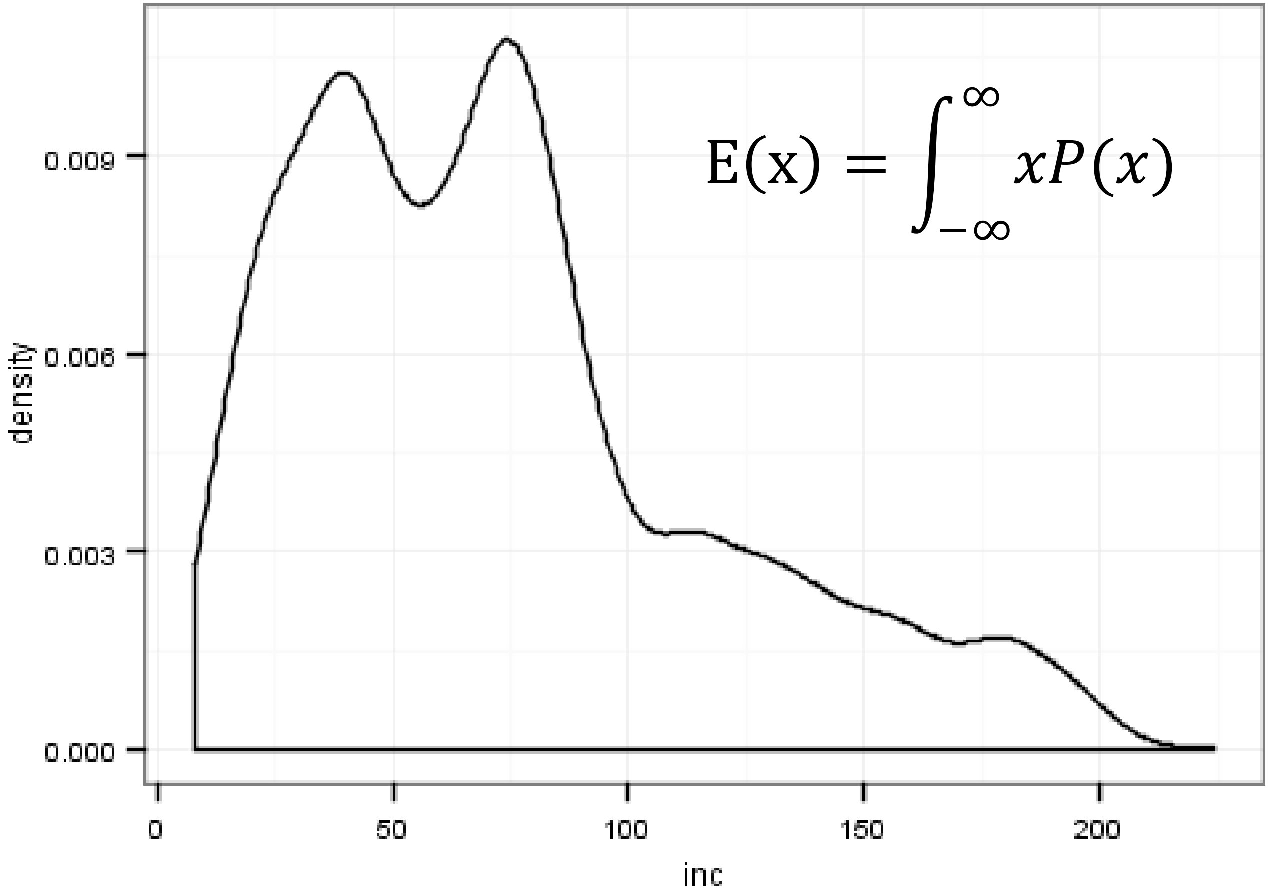
Recall anything like this?

Salary (BHD)	100	345	1000	9833
Frequency, f	10	1	10	2
Probability	0.43	0.04	0.43	0.09

Mean, $\mu = \frac{\Sigma x}{n} = \frac{\Sigma f x}{\Sigma f} = \frac{100X10+345X1+1000X10+9833X2}{10+1+10+2} = 1348$

Expectation, $E(X) = 100 * 0.43 + 345 * 0.04 + 1000 * 0.43 + 9833 * 0.09 = 1348$

Expectation: Continuous



$$E(x) = \int_{-\infty}^{\infty} xP(x)$$

Probability Distribution of Winnings

Combination	None	Lemons	Cherries	Dollars/Cherry	Dollars
P(X=x)	0.977	0.008	0.008	0.006	0.001
x	-$1	$4	$9	$14	$19

EXPECTATION, $\mathrm{E}(X) = \mu = \Sigma x P(X = x)$

E(X) = -0.77 (calculate and verify)

This is the amount of $ expected to be “gained” on each pull of the lever.

So, why play?

There is **VARIANCE**.

Probability Distribution of Winnings

Combination	None	Lemons	Cherries	Dollars/Cherry	Dollars
P(X=x)	0.977	0.008	0.008	0.006	0.001
x	-$1	$4	$9	$14	$19

VARIANCE, $Var(X) = E(X-\mu)^2 = \Sigma(x-\mu)^2 P(X=x)$

$$\sigma = \sqrt{Var(X)}$$

Simplifying the Formula

$$E[(X-\mu)^2]= E[X^2 - 2\mu X + \mu^2]$$

$$= E[X^2] - 2\mu E[X] + \mu^2$$ *(we get this as μ is just a number)*

$$= E[X^2] - 2\mu^2 + \mu^2$$

$$=E[X^2] - \mu^2 = E[X^2] - [E(X)]^2$$

Expectation Properties

$E(X+Y) = E(X) + E(Y)$ e.g., Playing a game each on 2 slot machines with different probabilities of winning. This is called **Independent Observation**.

$E(aX+b) = aE(X)+E(b) = aE(X) + b$ e.g., values x have been changed. This is called Linear Transformation.

If I have a portfolio of 30% Microsoft, 50% Bank of America and 20% Walmart stocks, the expected return of my portfolio is

$E(\text{Portfolio}) = 0.3\ E(\text{MS}) + 0.5\ E(\text{BofA}) + 0.2\ E(\text{Walmart})$

Variance Properties

- Var(X+a) = Var(X) (Variance does not change when a constant is added)
- Var(X+Y) = Var(X) + Var(Y) for Independent Observations
- Var(X-Y) = Var(X) + Var(Y)

Variance Properties

$Var(aX) = a^2 Var(X)$ for **Linear Transformation**

Say, $Y = aX$

$E(Y) = a E(X)$ (from the previous set of relations)
$Y - E(Y) = a(X - E(X))$

Squaring both sides and taking expectations
$E(Y-E(Y))^2 = a^2 E(X-E(X))^2$

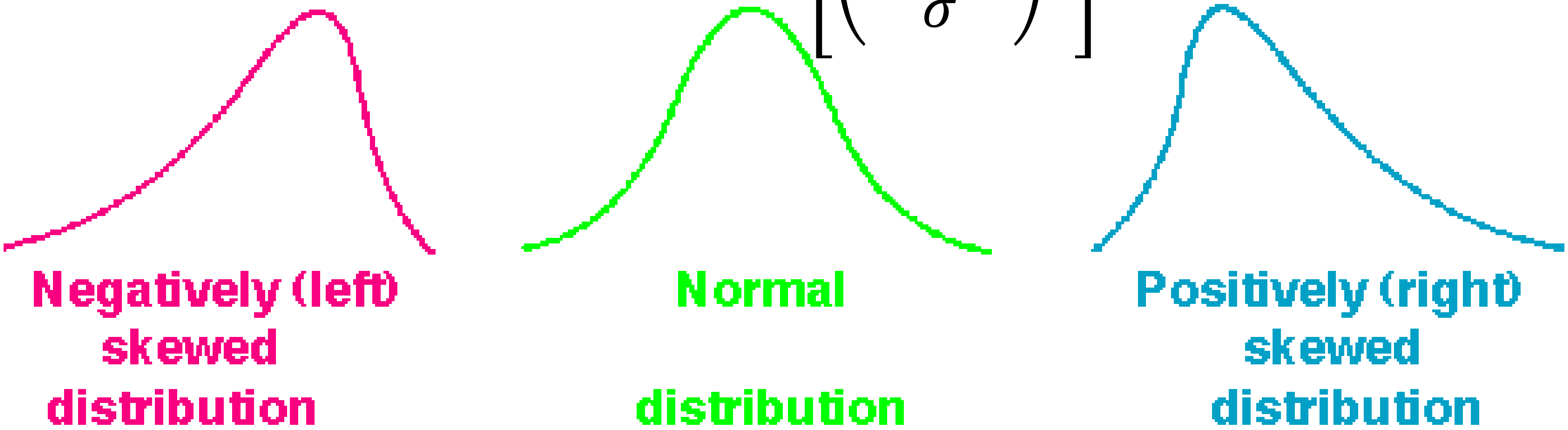
However, the left hand side is Variance of Y and RHS is Variance of X

$Var(Y) = a^2 Var(X)$ or $Var(aX) = a^2 Var(X)$

Understanding the Shape of a PDF - Skewness

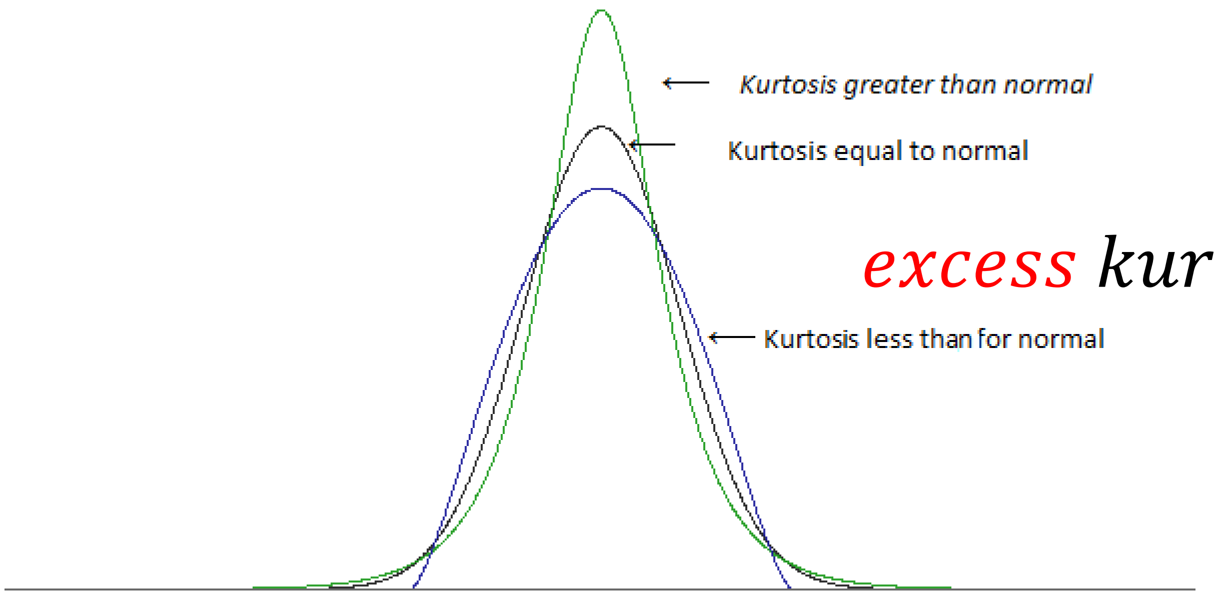
- A measure of symmetry. Negative skew indicates mean is less than median, and positive skew means median is less than mean.

$$skew(X) = E\left[\left(\frac{X-\mu}{\sigma}\right)^3\right]$$



Understanding the Shape of a PDF - Kurtosis

A measure of the 'tailed'ness of the data distribution as compared to a normal distribution. Negative kurtosis means a distribution with light tails (fewer extreme deviations from mean (or outliers) than in normal distribution). Positive kurtosis means a distribution with heavy tails (more outliers than in normal distribution).



$$\textcolor{red}{excess}\ kurt(X) = E\left[\left(\frac{X-\mu}{\sigma}\right)^4\right] - 3$$

Image Source: http://stats.stackexchange.com/questions/84158/how-is-the-kurtosis-of-a-distribution-related-to-the-geometry-of-the-density-fu
Last accessed: March 31, 2017

Rules of Thumb – Skewness and Kurtosis

Skewness

- Highly skewed: < -1 or > +1
- Moderately skewed: -1 to -0.5 or 0.5 to 1
- Symmetrical: -0.5 to 0.5

Excess Kurtosis

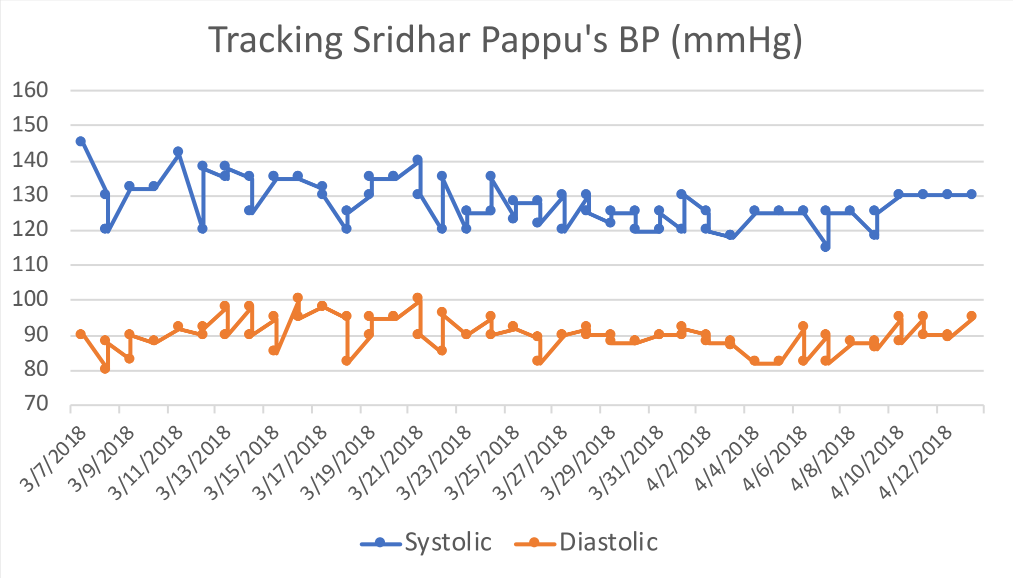
- High: < -1 or > +1
- Medium: -1 to -0.5 or 0.5 to 1
- Small: -0.5 to 0.5

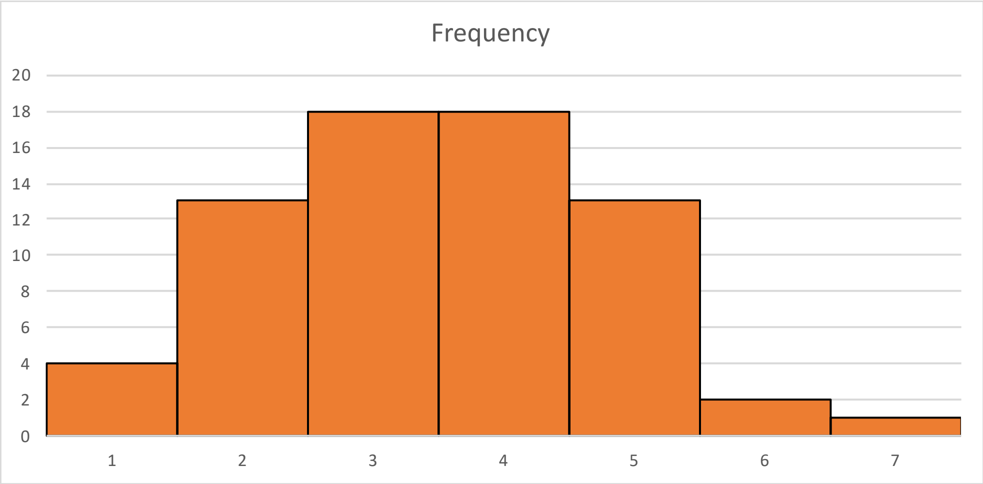
Describing a Distribution – Summary of Moments

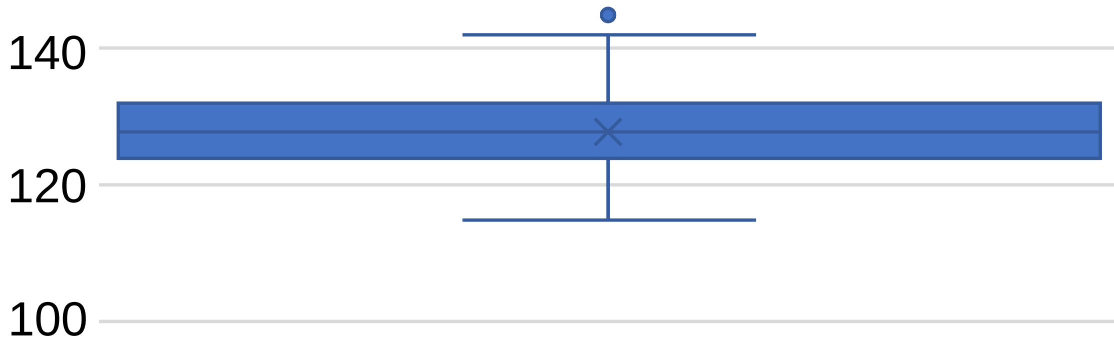
Measure	Formula	Description
Mean (μ)	$E(X)$	Measures the centre of the distribution of X
Variance (σ^2)	$E[(X-\mu)^2]$	Measures the spread of the distribution of X about the mean
Skewness	$E\left[\left(\frac{X-\mu}{\sigma}\right)^3\right]$	Measures asymmetry of the distribution of X
Kurtosis (excess)	$E\left[\left(\frac{X-\mu}{\sigma}\right)^4\right]-3$	Measures 'tailed'ness of the distribution of X and useful in outlier identification

Summary of Descriptive Statistics - Excel

- Central tendencies
- Measures of variability
- Box plot
- Histogram
- Scatterplot







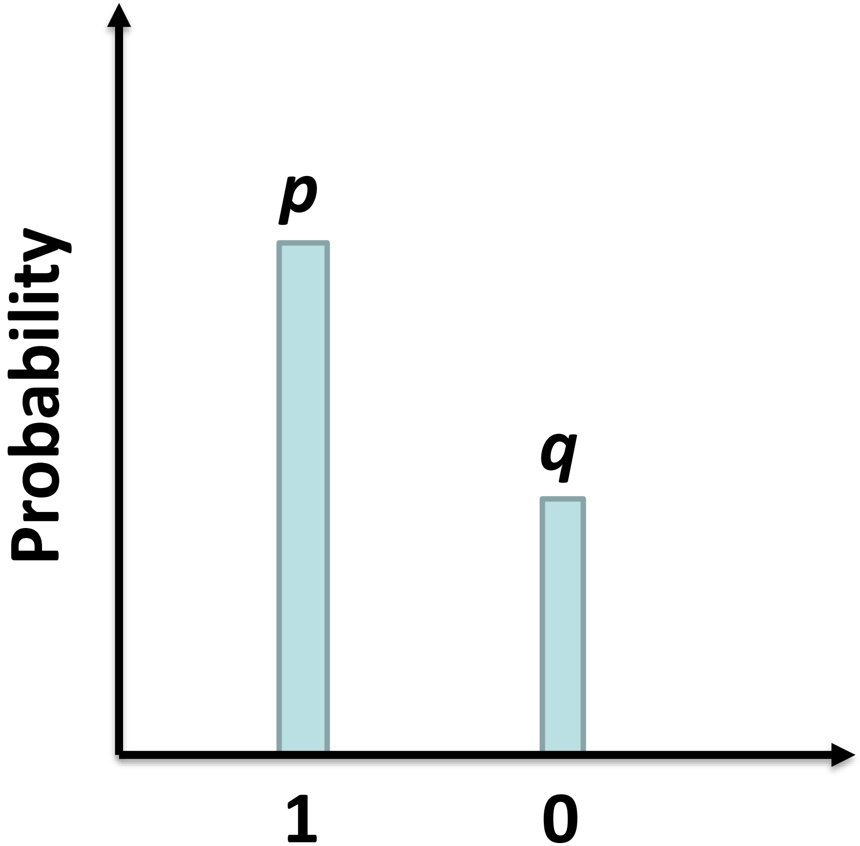
SOME COMMON DISTRIBUTIONS

Bernoulli

There are two possibilities (loan taker or non-taker) with probability p of success and $1-p$ of failure

- Expectation: p
- Variance: $p(1-p)$ or pq, where $q=1-p$

Bernoulli



$$Expectation, E(X) = \sum x_i P(x_i)$$

$$= 1 * p + 0 * q = p$$

$$Variance, Var = \sum (x_i - \mu)^2 P(x_i)$$

$$= (1-p)^2 * p + (0-p)^2 * (1-p)$$
$$= p(1-p)$$

Geometric Distribution

Number of independent and identical Bernoulli trials needed to get ONE success, e.g., number of people I need to call for the first person to accept the loan.

Geometric Distribution

PMF*, $P(X = r) = q^{r-1}p$ — *(r-1)* failures followed by ONE success.

$P(X > r) = q^r$ — Probability you will need more than *r* trials to get the first success.

CDF**, $P(X \leq r) = 1 - q^r$ — Probability you will need *r* trials or less to get your first success.

$$E(X) = \frac{1}{p} \qquad Var(X) = \frac{q}{p^2}$$

* Probability Mass Function ** Cumulative Distribution Function

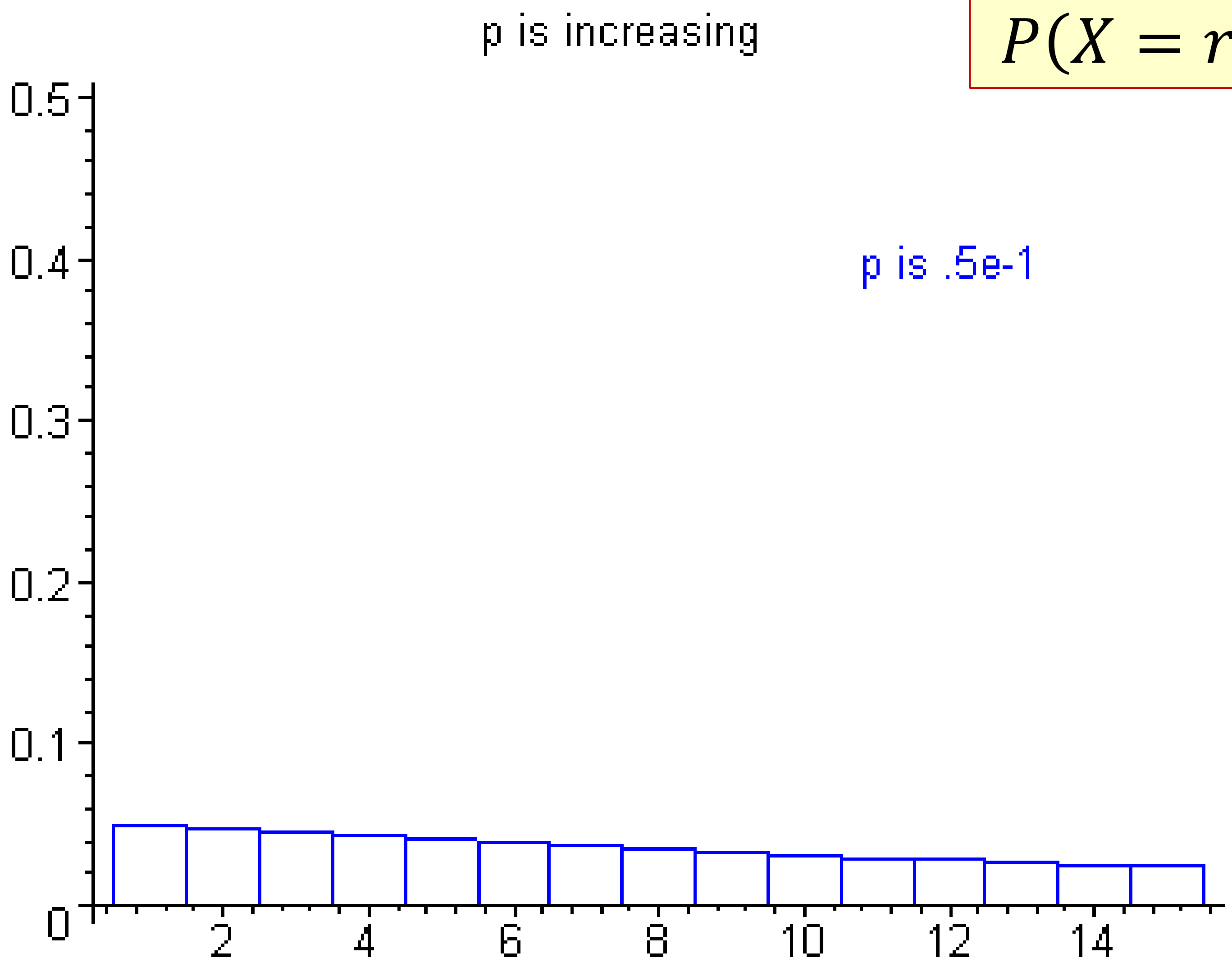
CSE 7315c

Geometric Distribution

- You run a series of independent trials.
- There can be either a success or a failure for each trial, and the probability of success is the same for each trial.
- The main thing you are interested in is how many trials are needed in order to get the first successful outcome.

X~Geo(p)

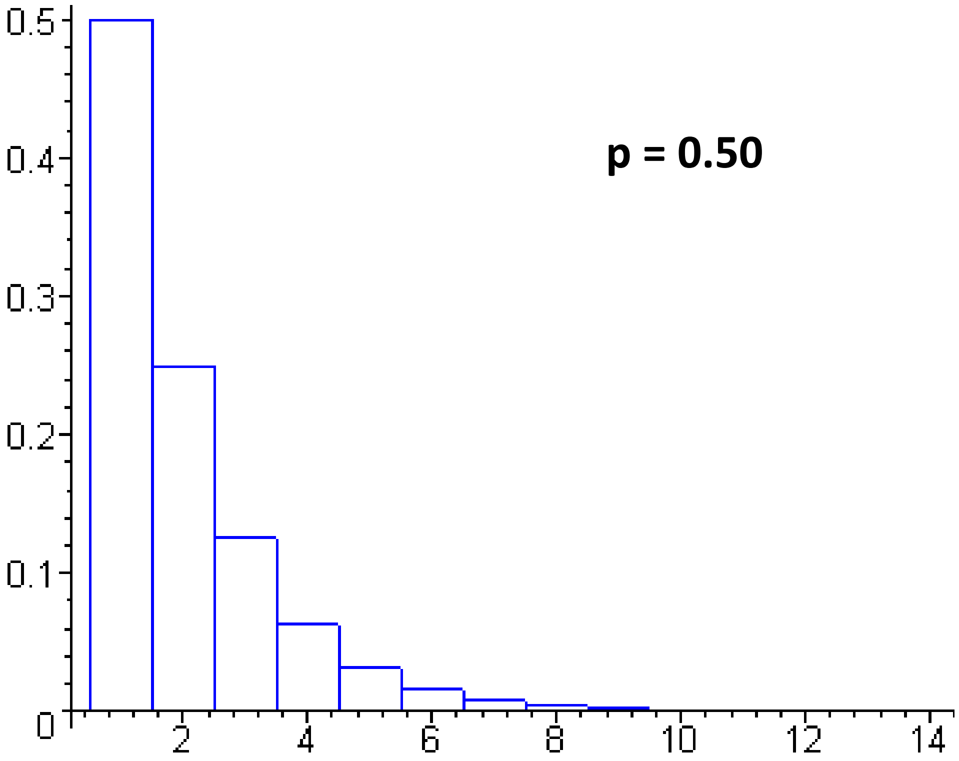
$$P(X = r) = q^{r-1}p$$

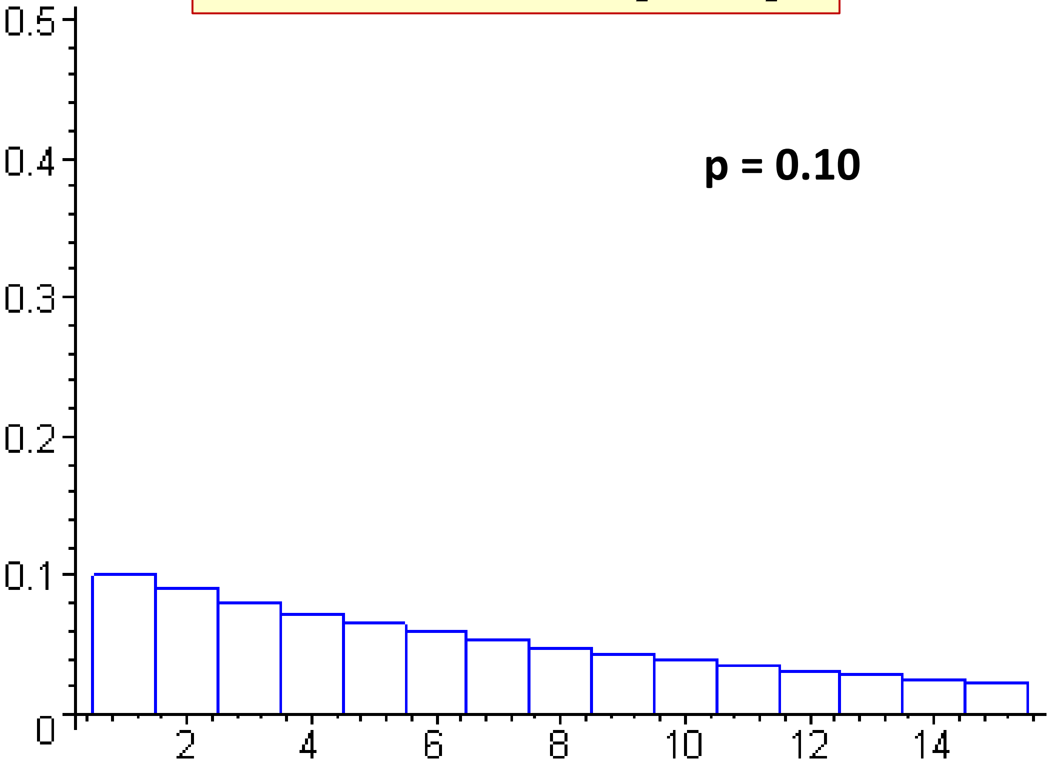


Ref: http://personal.kenyon.edu/hartlaub/MellonProject/Geometric2.html
Last accessed: June 12, 2015

X~Geo(p)

$$P(X = r) = q^{r-1}p$$



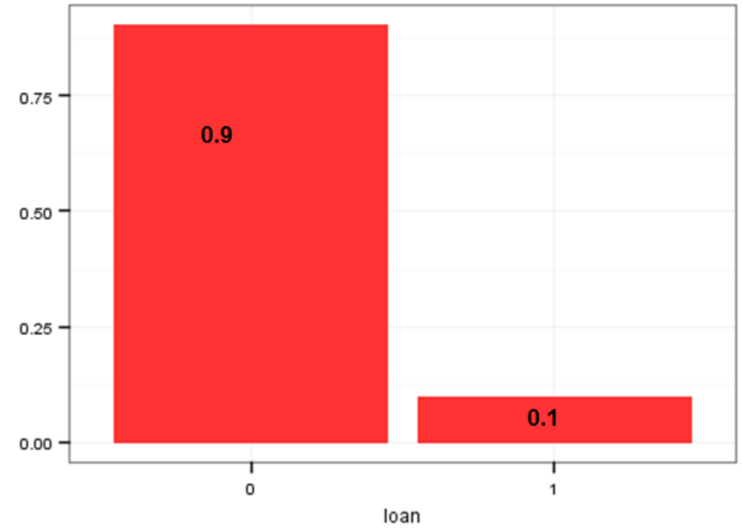


Ref: http://personal.kenyon.edu/hartlaub/MellonProject/Geometric2.html
Last accessed: December 09, 2017

Binomial Distribution

If I randomly pick 10 people, what is the probability that I will get exactly

- 0 loan takers = 0.9^{10}
- 1 loan taker = $10 * 0.1^{1} * 0.9^{9}$
- 2 loan takers = $C_{2}^{10} * 0.1^{2} * 0.9^{8}$



Binomial Distribution

If there are two possibilities with probability *p* for success and *q* for failure, and if we perform *n* trials, the probability that we see *r* successes is

PMF, $\mathrm{P(X = r)} = C_r^n p^r q^{n-r}$

CDF, $\mathrm{P(X \leq r)} = \sum_{i=0}^{r} C_i^n p^i q^{n-i}$

Binomial Distribution

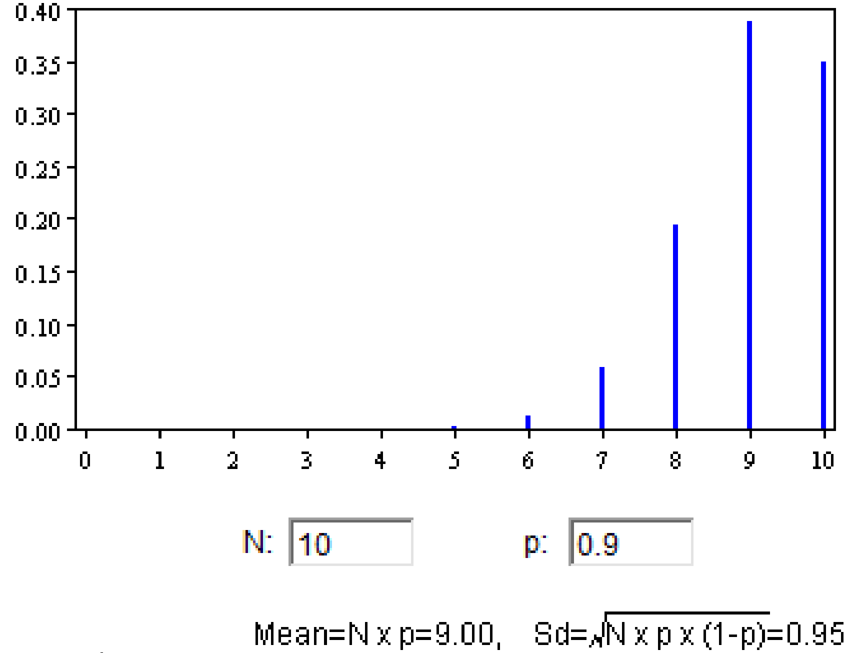
$$E(X) = np \qquad\qquad Var(X) = npq$$

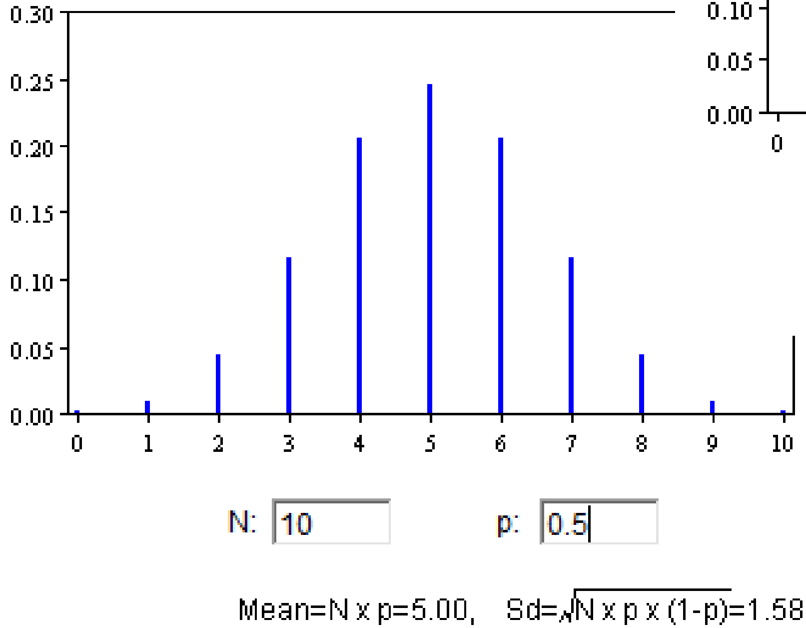
When to use?

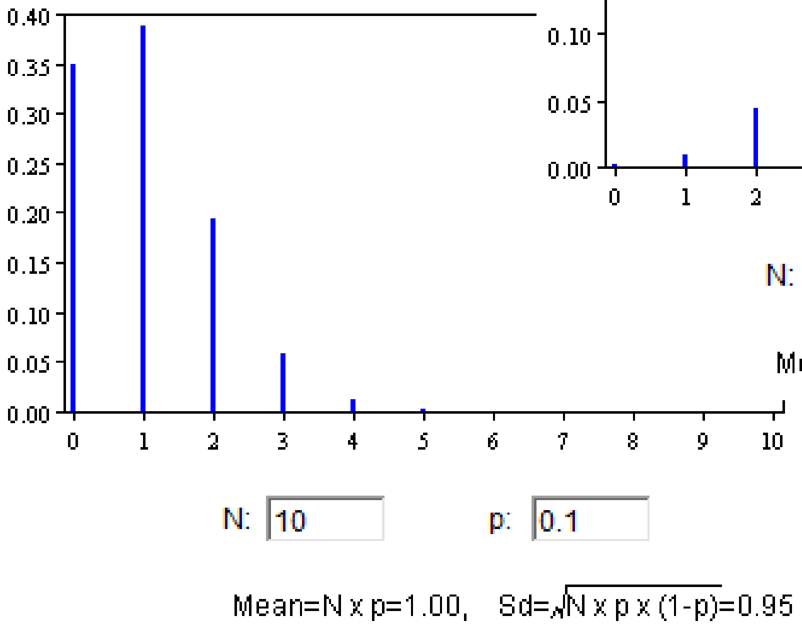
- You run a series of independent trials.
- There can be either a success or a failure for each trial, and the probability of success is the same for each trial.
- There are a finite number of trials, and you are interested in the number of successes or failures.

X~B(n,p)

$$P(X = r) = C_r^n p^r q^{n-r}$$







Ref: http://onlinestatbook.com/2/probability/binomial_demonstration.html
Last accessed: December 09, 2017 on Safari

Poisson Distribution

French pronunciation: [pwasɔ̃]; in English often rendered /ˈpwɑːsɒn/ - Wikipedia

Binomial: We are interested in number of successes/events (discrete) occurring randomly in fixed *number of trials* (discrete).

Poisson: We are interested in number of successes/events (discrete) occurring randomly in fixed *duration or space* (continuous).

Poisson Distribution

- No. of deaths by horse and mule kicking between 1875-1894 in the Prussian army (http://blog.minitab.com/blog/quality-data-analysis-and-statistics/no-horsing-around-with-the-poisson-distribution-troops)
- No. of birth defects
- No. of defects in a batch of semiconductor wafers
- No. of typing errors per page
- No. of insurance claims (or policies sold) per week
- No. of vehicles passing through a busy traffic junction per minute
- No. of car accidents per hour

Poisson Distribution

Probability of getting 15 customers requesting for loans in a given day, given on average we see 10 customers

$\lambda = 10 \; and \; r = 15$

PMF, $\mathrm{P(X = r)} = \dfrac{e^{-\lambda}\lambda^r}{r!}$

CDF, $\mathrm{P(X \leq r)} = e^{-\lambda} \sum_{i=0}^{r} \dfrac{\lambda^i}{i!}$

Poisson Distribution

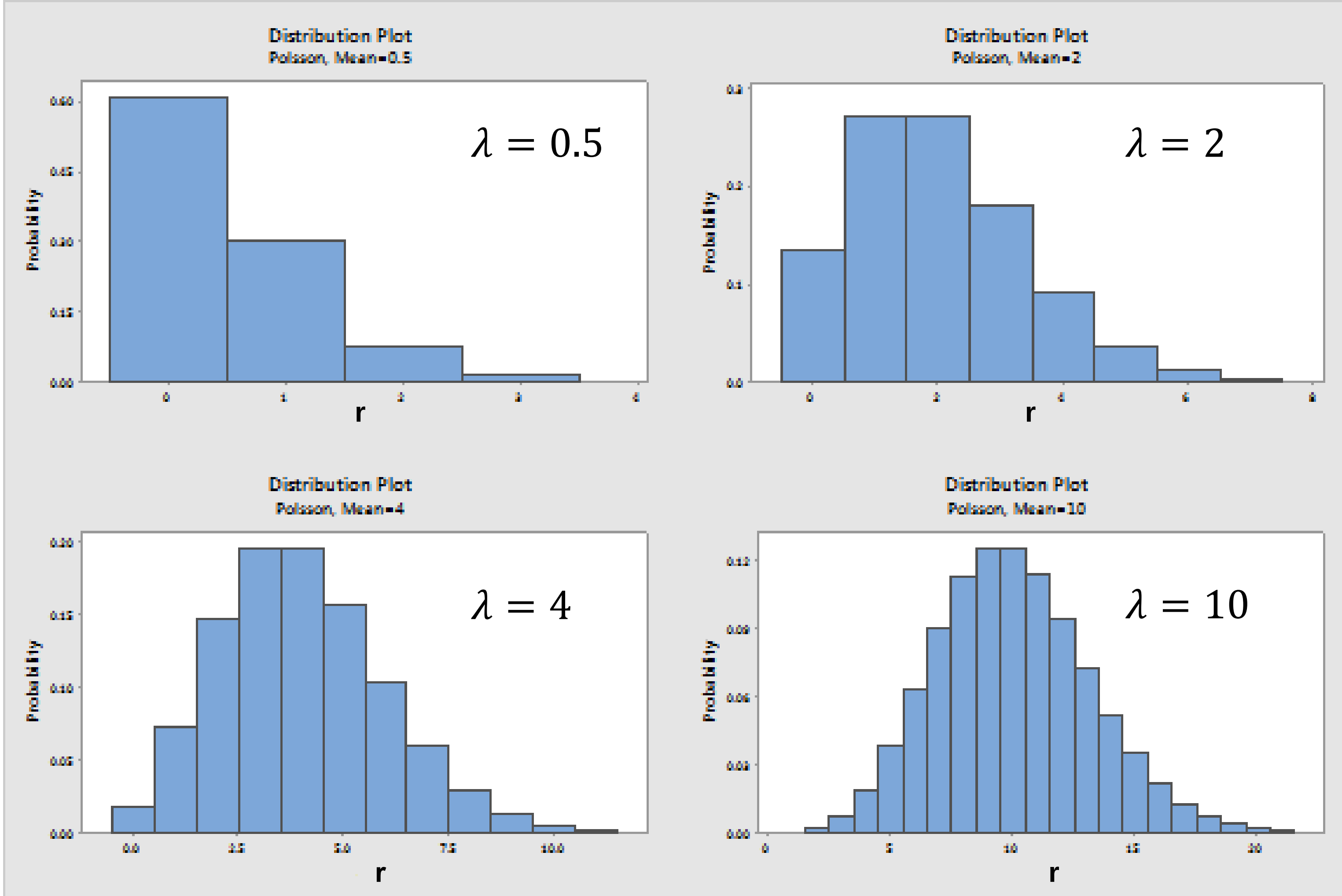
$$E(X) = \lambda$$
$$Var(X) = \lambda$$

When to use?

- Individual events occur at random and independently in a given interval (time or space).
- You know the mean number of occurrences, λ, in the interval or the rate of occurrences, and it is finite.

X~Po(λ)

$$P(X = r) = \frac{e^{-\lambda}\lambda^{r}}{r!}$$



Ref: http://blog.minitab.com/blog/quality-data-analysis-and-statistics/no-horsing-around-with-the-poisson-distribution-troops
Last accessed: March 02, 2018

Poisson Distribution

- Limiting case of Binomial distribution when $n \to \infty$ (infinite trials) and $p \to 0$ (infinitesimally small probability, i.e., "rare" events).
- As a rule of thumb, if $n > 50$ and $p < 0.1$, Binomial can be approximated by Poisson, i.e., $np \to \lambda$.
- That is, Poisson distribution is used to model occurrences of events that ***could*** happen a very large number of times (large n), but ***actually*** happen very rarely (small p).

Poisson Distribution

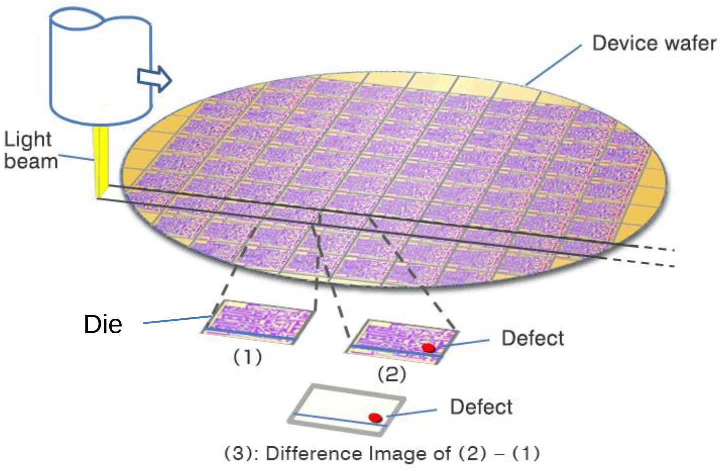
Example

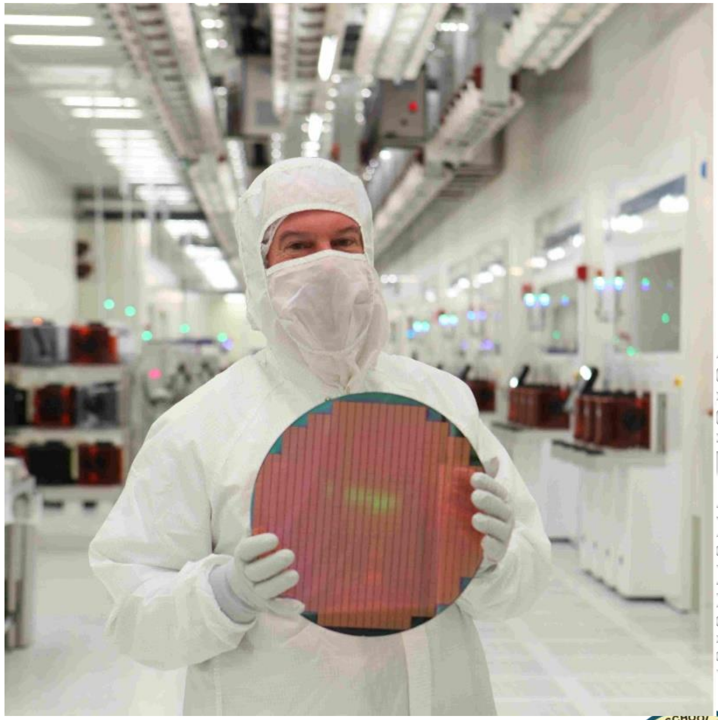
In a tie-breaking T20 Super Over, there are fixed number of opportunities to hit a six, and the probability of hitting a six is very high. So, the number of sixes in a T20 Super Over is **Binomial**.

On the other hand, in a cricket Test Match, a six can be hit almost every few minutes, but a six is probably hit once in a few hours. So, the number of sixes in a Test Match is **Poisson**.

A company makes semiconductor wafers. The probability of a defective die on the wafer is 0.001. What is the probability that a random sample of 500 dies will contain exactly 5 defective dies?

What distribution is this?





Poisson Distribution

Approach 1: Binomial

n = 500, p = 0.001, r = 5

$^{500}C_5 * (0.001)^5 * (1-0.001)^{495} = 0.00156$

Approach 2: Poisson

$$\lambda = np = 0.5, r = 5$$

$$\frac{2.718^{-0.5} 0.5^5}{5!} = 0.00158$$

Note: e = 2.718

Poisson Distribution

The probability that no customer will visit the store in one day

$$P(X=0) = \frac{e^{-\lambda}\lambda^{0}}{0!} = e^{-\lambda}$$

Probability that no customer will visit in *n* days

$$e^{-n\lambda}$$

Exponential Distribution

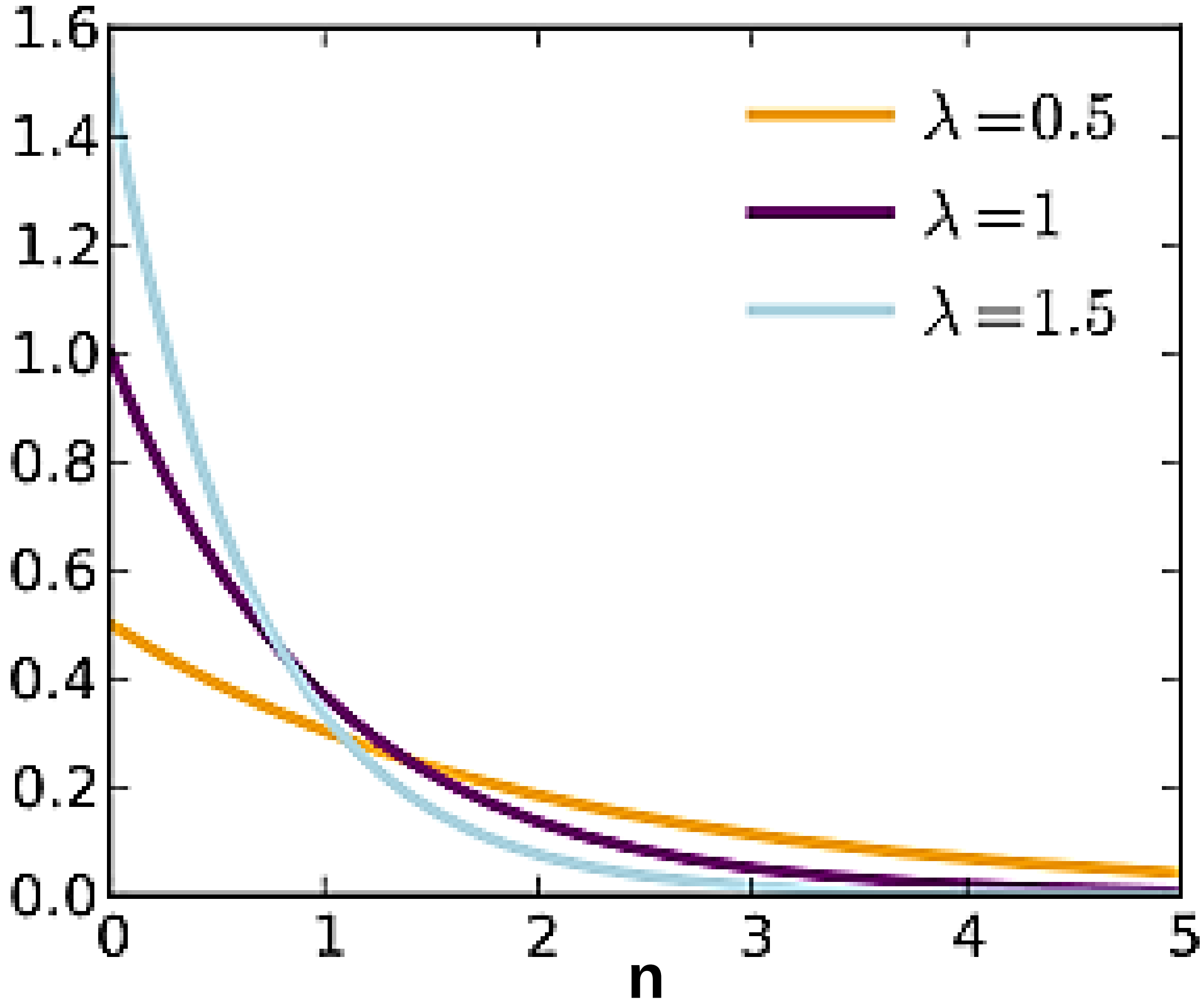
Probability that a customer will visit in *n* days:

$$1 - e^{-n\lambda}$$

$$CDF = 1 - e^{-n\lambda}, n \geq 0$$
$$PDF = \lambda e^{-n\lambda}, n \geq 0$$

X~Exp(λ)

$$PDF = \lambda e^{-n\lambda}, n \geq 0$$



Ref: http://en.wikipedia.org/wiki/Exponential_distribution
Last accessed: June 12, 2015

Exponential Distribution

- Poisson process
- Continuous analog of Geometric distribution

$$E(X) = \frac{1}{\lambda} \qquad Var(X) = \frac{1}{\lambda^2}$$

Probability Distributions

Geometric:	For estimating number of attempts before first success
Binomial:	For estimating number of successes in *n* attempts
Poisson:	For estimating *n* number of events in a given time period when on average we see *m* events
Exponential:	Time between events

Probability Distributions - Scenarios

Identify the distribution and calculate expectation, variance and the required probabilities.

Q1. A man is bowling. The probability of him knocking all the pins over is 0.3. If he has 10 shots, what is the probability he will knock all the pins over less than 3 times?

Probability Distributions - Scenarios

X ~ B(10,0.3); n=10, p=0.3, q=1-0.3=0.7, r=0, 1, 2 (< 3)

E(X) = np = 3

Var(X) = npq = 2.1

$$P(X = r) = \ {}^{n}C_{r} p^{r} q^{n-r}$$

P(X=0) = 0.028; P(X=1) = 0.121; P(X=2) = 0.233

∴ P(X<3) = 0.028 + 0.121 + 0.233 = 0.382

Probability Distributions - Scenarios

Identify the distribution and calculate expectation, variance and the required probabilities.

Q2. On average, 1 bus stops at a certain point every 15 minutes. What is the probability that no buses will turn up in a single 15 minute interval?

Probability Distributions - Scenarios

X ~ Po(1); λ=1, r=0

E(X) = λ = 1

Var(X) = λ = 1

$$P(X = r) = \frac{e^{-\lambda}\lambda^{r}}{r!}$$

P(X=0) = 0.368

Probability Distributions - Scenarios

Identify the distribution and calculate expectation, variance and the required probabilities.

Q3. 20% of cereal packets contain a free toy. What is the probability you will need to open fewer than 4 cereal packets before finding your first toy?

Probability Distributions - Scenarios

X ~ Geo(0.2); p=0.2, q=1-0.2=0.8, r<4 or ≤3

E(X) = $\frac{1}{p}$ = 5

Var(X) = $\frac{q}{p^2}$ = 20

$$P(X \leq r) = 1 - q^r$$

P(X ≤ 3) = 0.488

Poisson Distribution Formula Differences?

$$P(X = r) = \frac{e^{-\lambda}\lambda^r}{r!} \ or \ \frac{e^{-\lambda t}(\lambda t)^r}{r!} \ ?$$

Suppose births in a hospital occur randomly at an average rate of 1.8 births per hour. What is the probability of 5 births in a given 2 hour interval?

What is λ?

$$P(X = 5) = \frac{e^{-3.6}3.6^5}{5!} \ or \ \frac{e^{-1.8*2}(1.8 * 2)^5}{5!} \ ?$$

If you use 1.8, use t=2 in the second formula. Alternatively, you could say that since the average is 1.8 per hour, it is 3.6 per 2 hours (the interval of interest).

Poisson Distribution Formula Differences?

$$P(X = r) = \frac{e^{-\lambda}\lambda^r}{r!} \ or \ \frac{e^{-\lambda t}(\lambda t)^r}{r!} \ ?$$

Now suppose head injury patients (due to not wearing helmets) arrive in Hospital A randomly at an average rate of 0.25 patients per hour, and in Hospital B randomly at an average rate of 0.75 per hour. What is the probability of more than 3 such patients arriving in a given 2 hour interval in both hospitals together?

What is the probability distribution?

$X \sim Po(\lambda_1) \ and \ Y \sim Po(\lambda_2)$
$X + Y \sim Po(\lambda_1 + \lambda_2)$

What are λ_1 and λ_2 if we use first formula?

$\lambda_1 = 0.5 \text{ and } \lambda_2 = 1.5$
$\lambda_1 + \lambda_2 = 2$

$$P(X+Y > 3) = P(X+Y=4) + P(X+Y=5) + P(X+Y=6) + \cdots$$
$$= 1 - P(X+Y \leq 3) = 1 - (P(X+Y=0) + P(X+Y=1) + P(X+Y=2) + P(X+Y=3))$$

Poisson or Exponential?

Given a Poisson process:

- The *number* of events in a given time period — Poisson
- The *time* until the first event
- The *time* from now until the next occurrence of the event
- The *time interval* between two successive events

Exponential

Poisson or Exponential?

The tech support centre of a computer retailer receives 5 calls per hour on an average. What is the probability that the centre will receive 8 calls in the next hour? What is the probability that more than 30 minutes will elapse between calls?

$$P(X = 8) = \frac{e^{-5}5^8}{8!} = 0.065$$

$$P(Time\ between\ calls > 0.5) = \int_{0.5}^{\infty} \lambda e^{-\lambda T} dT = -e^{-\lambda T}]_{0.5}^{\infty}$$

$$= e^{-5*0.5} = 0.082$$

Probability Distributions

Babyboom Data - Excel

Forty-four babies -- a new record -- were born in one 24-hour period at the Mater Mothers' Hospital in Brisbane, Queensland, Australia, on December 18, 1997. For each of the 44 babies, *The Sunday Mail* recorded the time of birth, the sex of the child, and the birth weight in grams.

Probability Distributions

Determine the distributions for the following scenarios for this dataset:

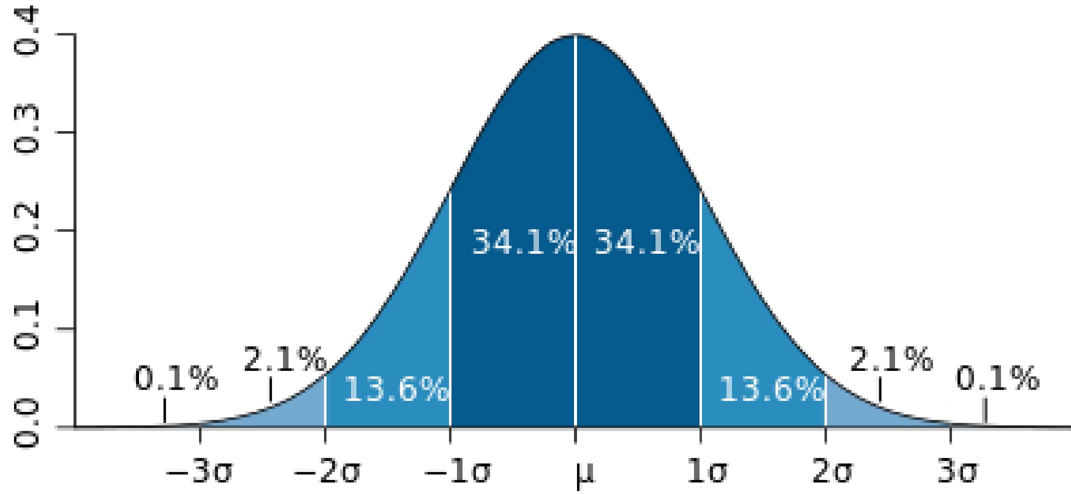
1. Probability of observing at least 26 boys in 44 births assuming equal probability of a boy or a girl being born.
2. Probability that 3 births occur before the birth of a girl.
3. Probability of 4 births per hour given 44/24 = 1.83 births per hour on average.
4. Probability that more than 60 minutes will elapse between births.

1. Binomial; 2. Geometric; 3. Poisson; 4. Exponential

NORMAL DISTRIBUTION

Normal (Gaussian) Distribution

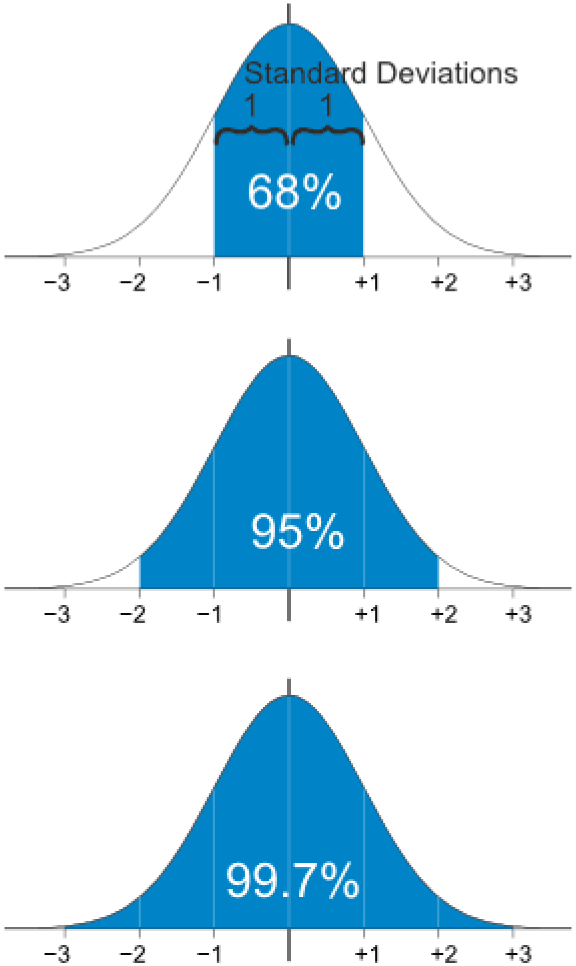
- Mean = Median = Mode
- 68-95-99.7 empirical rule
- Zero Skew and Kurtosis
- X ~ N(μ, σ^2)
- Shaded area gives the probability that X is between the corresponding values



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

Measures of Spread (Dispersion)

You know the 68-95-99.7 rule.



A company produces a valve that is specified to weigh 1500g, but there are imperfections in the process. While the mean weight is 1500g, the standard deviation is 300g.

Q1. What is the range of weights within which 95% of the valves will fall?

Q2. Approximately 16% of the weights will be more than what value?

Q3. Approximately 0.15% of the weights will be less than what value?

Image source: http://www.mathsisfun.com/data/standard-normal-distribution.html
Last accessed: December 15, 2017

Sample Software Output

SUMMARY OUTPUT

Regression Statistics	
Multiple R	0.717055011
R Square	0.514167888
Adjusted R Square	0.494734604
Standard Error	4.21319131
Observations	27

ANOVA

	df	SS	MS	F	Significance F
Regression	1	469.6573265	469.6573265	26.4581054	2.57053E-05
Residual	25	443.7745253	17.75098101		
Total	26	913.4318519			

	Coefficients	Standard Error	t Stat	P-value	Lower 95%	Upper 95%	Lower 99.0%	Upper 99.0%
Intercept	-4.154014573	2.447784673	-1.697050651	0.102104456	-9.195321476	0.88729233	-10.97705723	2.669028089
Big Mac Price ($)	3.547427488	0.689658599	5.143744297	2.57053E-05	2.127049014	4.967805962	1.625048409	5.469806567

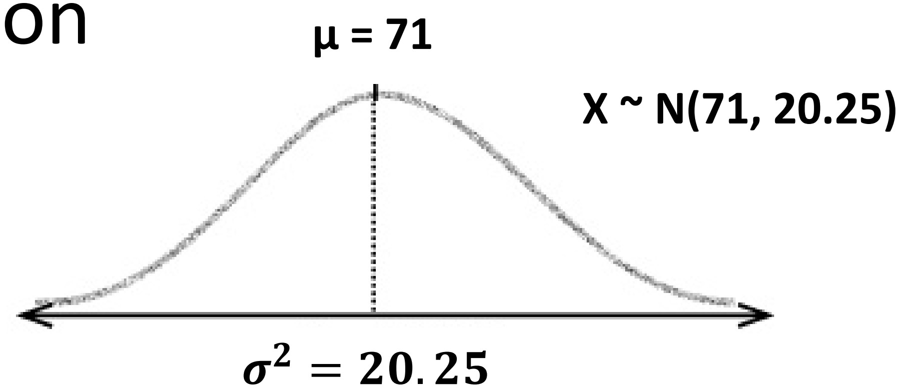
Sample Software Output

```
Call:
glm(formula = Response ~ Age, family = "binomial", data = flierresponse)

Deviance Residuals:
     Min        1Q    Median        3Q       Max
-1.95015  -0.32016  -0.05335   0.26538   1.72940

Coefficients:
             Estimate Std. Error z value Pr(>|z|)
(Intercept) -20.40782    4.52332  -4.512 6.43e-06 ***
Age           0.42592    0.09482   4.492 7.05e-06 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 123.156  on 91  degrees of freedom
Residual deviance:  49.937  on 90  degrees of freedom
AIC: 53.937

Number of Fisher Scoring iterations: 7
```

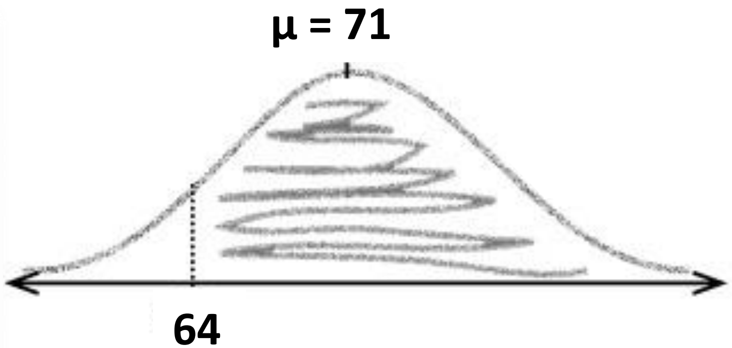
Calculating Normal Probabilities

Step 1: Determine the distribution

Julie wants to marry a person taller than her and is going on blind dates. The mean height of the 'available' guys is 71" and the variance is 20.25 inch^2 (yuck!).



Oh! By the way, Julie is 64" tall.



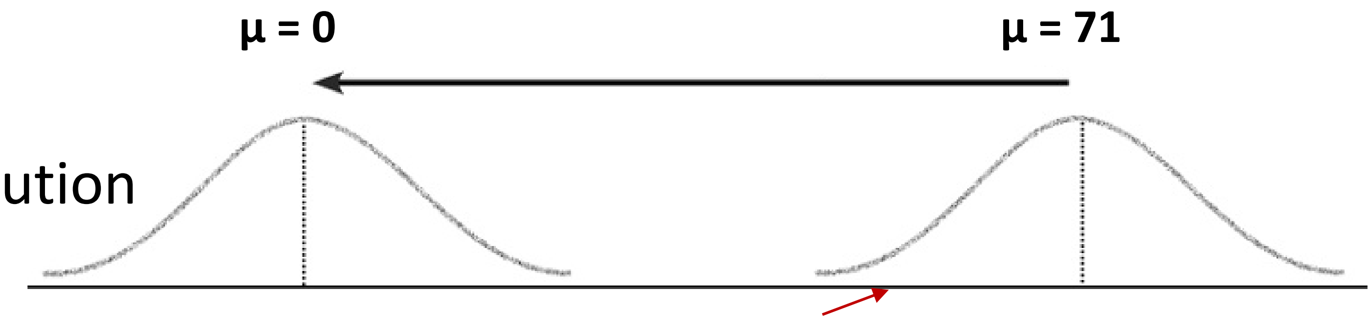
Calculating Normal Probabilities

Step 2: Standardize to Z ~ N(0,1)

1. Move the mean

   This gives a new distribution

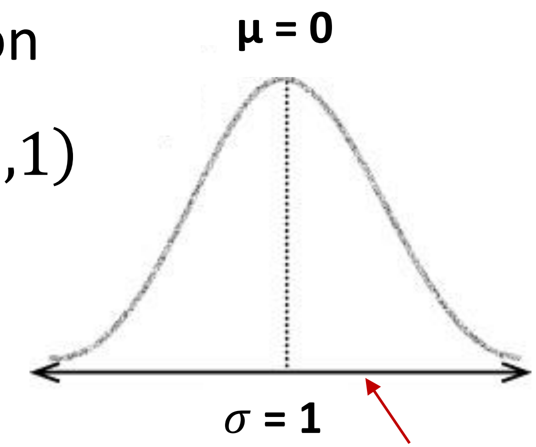
   X-71 ~ N(0,20.25)



Random variable is **x**, the ***actual*** heights of available guys

2. Squash the width by dividing by the standard deviation

   This gives us $\frac{X-71}{4.5} \sim N(0,1)$

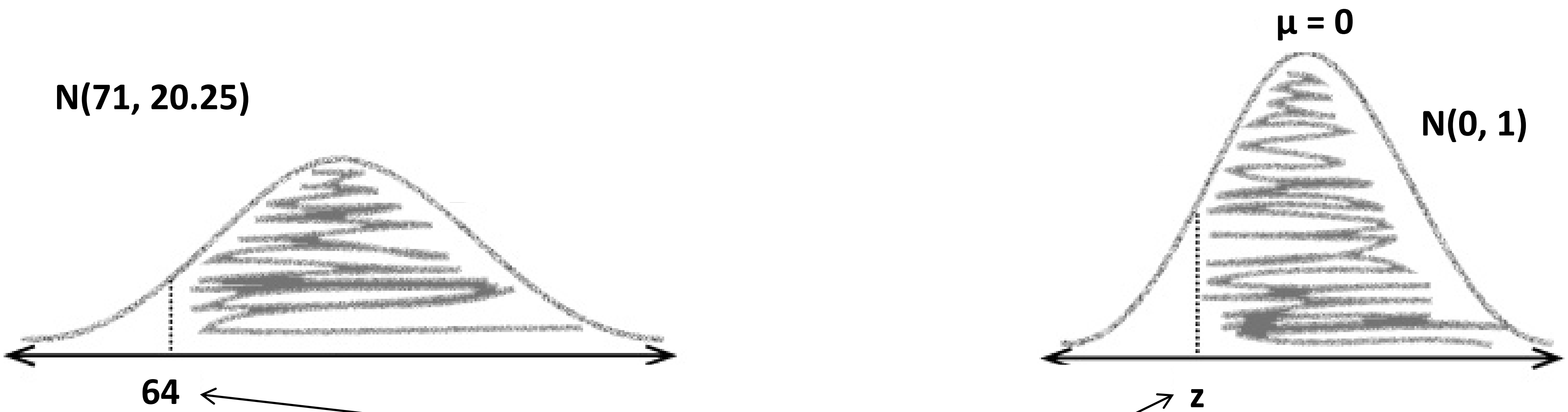


$Z = \frac{X-\mu}{\sigma}$ is called the Standard Score or the z-score.

Random variable is **z**, the the ***standardized*** heights of available guys

Calculating Normal Probabilities

Step 2: Standardize to Z ~ N(0,1)



$$z = \frac{64 - 71}{4.5} = -1.56$$

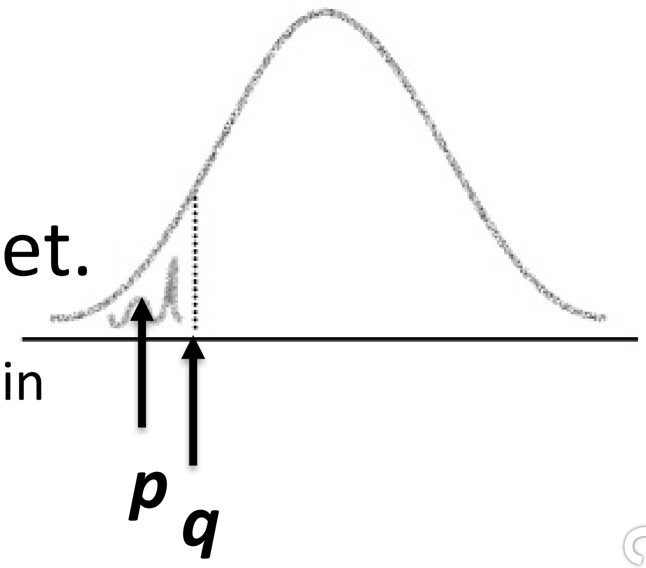
Julie is 64” tall, i.e., she is 1.56 standard deviations shorter than the average height of the available guys.

Calculating Normal Probabilities

Step 3: Look up the probability in the tables

Note the tables give P(Z<z).

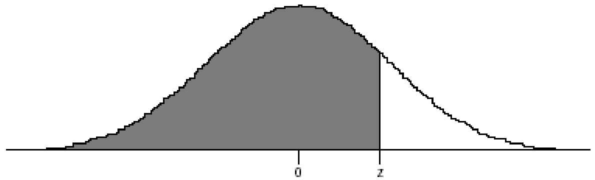
In R functions, the distribution is abbreviated and prefixed with an alphabet.



pnorm: Probability (Cumulative Distribution Function, CDF) in a *Normal Distribution*

qnorm: Quantile (Inverse CDF) in a *Normal Distribution* – The value corresponding to the desired probability.

Calculating Normal Probabilities



Step 3: Look up the probability in the tables

Note the tables give P(Z<z).

$$z = \frac{64-71}{4.5} = -1.56$$ in the case of our problem.

P(Z>-1.56) = 1- P(Z<-1.56) = 1- 0.0594 = 0.9406

Normal Deviate z	.00	.01	.02	.03	.04	.05	.06	.07	.08	.09
-4.0	.0000	.0000	.0000	.0000	.0000	.0000	.0000	.0000	.0000	.0000
-3.9	.0000	.0000	.0000	.0000	.0000	.0000	.0000	.0000	.0000	.0000
-3.8	.0000	.0000	.0000	.0000	.0000	.0000	.0000	.0000	.0000	.0000
-3.7	.0001	.0001	.0000	.0000	.0000	.0000	.0000	.0000	.0000	.0000
-3.6	.0002	.0002	.0001	.0001	.0001	.0001	.0001	.0001	.0001	.0001
-3.5	.0002	.0002	.0002	.0002	.0002	.0002	.0002	.0002	.0002	.0002
-3.4	.0003	.0003	.0003	.0003	.0003	.0003	.0003	.0003	.0003	.0002
-3.3	.0005	.0005	.0005	.0004	.0004	.0004	.0004	.0004	.0004	.0003
-3.2	.0007	.0007	.0006	.0006	.0006	.0006	.0006	.0005	.0005	.0005
-3.1	.0010	.0009	.0009	.0009	.0008	.0008	.0008	.0008	.0007	.0007
-3.0	.0013	.0013	.0013	.0012	.0012	.0011	.0011	.0011	.0010	.0010
-2.9	.0019	.0018	.0018	.0017	.0016	.0016	.0015	.0015	.0014	.0014
-2.8	.0026	.0025	.0024	.0023	.0023	.0022	.0021	.0021	.0020	.0019
-2.7	.0035	.0034	.0033	.0032	.0031	.0030	.0029	.0028	.0027	.0026
-2.6	.0047	.0045	.0044	.0043	.0041	.0040	.0039	.0038	.0037	.0036
-2.5	.0062	.0060	.0059	.0057	.0055	.0054	.0052	.0051	.0049	.0048
-2.4	.0082	.0080	.0078	.0075	.0073	.0071	.0069	.0068	.0066	.0064
-2.3	.0107	.0104	.0102	.0099	.0096	.0094	.0091	.0089	.0087	.0084
-2.2	.0139	.0136	.0132	.0129	.0125	.0122	.0119	.0116	.0113	.0110
-2.1	.0179	.0174	.0170	.0166	.0162	.0158	.0154	.0150	.0146	.0143
-2.0	.0228	.0222	.0217	.0212	.0207	.0202	.0197	.0192	.0188	.0183
-1.9	.0287	.0281	.0274	.0268	.0262	.0256	.0250	.0244	.0239	.0233
-1.8	.0359	.0351	.0344	.0336	.0329	.0322	.0314	.0307	.0301	.0294
-1.7	.0446	.0436	.0427	.0418	.0409	.0401	.0392	.0384	.0375	.0367
-1.6	.0548	.0537	.0526	.0516	.0505	.0495	.0485	.0475	.0465	.0455
-1.5	.0668	.0655	.0643	.0630	.0618	.0606	.0594	.0582	.0571	.0559
-1.4	.0808	.0793	.0778	.0764	.0749	.0735	.0721	.0708	.0694	.0681
-1.3	.0968	.0951	.0934	.0918	.0901	.0885	.0869	.0853	.0838	.0823
-1.2	.1151	.1131	.1112	.1093	.1075	.1056	.1038	.1020	.1003	.0985
-1.1	.1357	.1335	.1314	.1292	.1271	.1251	.1230	.1210	.1190	.1170
-1.0	.1587	.1562	.1539	.1515	.1492	.1469	.1446	.1423	.1401	.1379

Calculating Normal Probabilities

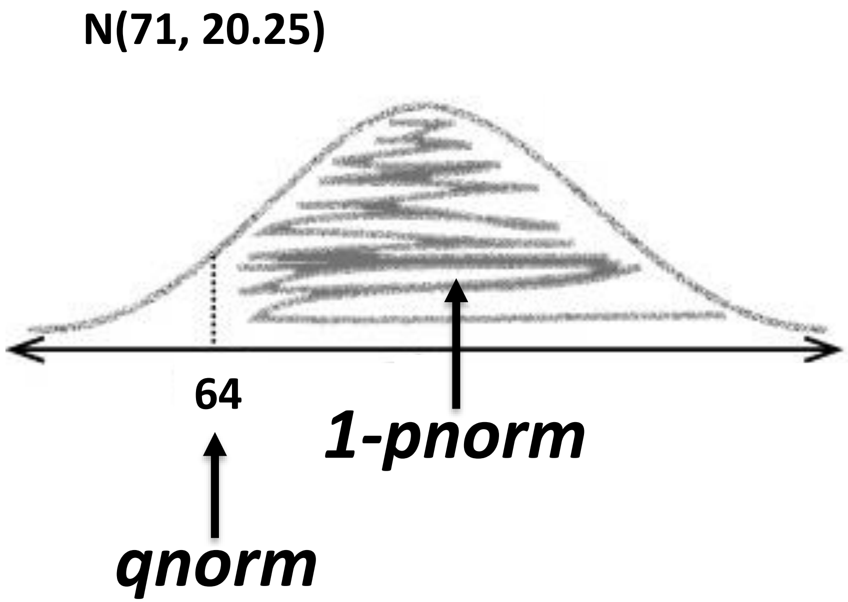
Step 3: Get the probability from R

1-pnorm(64, mean=71, sd=sqrt(20.25))
or
1-pnorm(64, 71, 4.5)

Answer: 1-0.0599 = 94.01%

qnorm(0.0599, 71, 4.5)

Answer: 64



Attention Check

Q. What is the standard score for N(10,4), value 6?

A. $z = \frac{6-10}{2} = -2$

Q. The standard score of value 20 is 2. If the variance is 16, what is the mean?

A. $2 = \frac{20-\mu}{4}$. $\therefore\ \mu = 20 - 8 = 12$

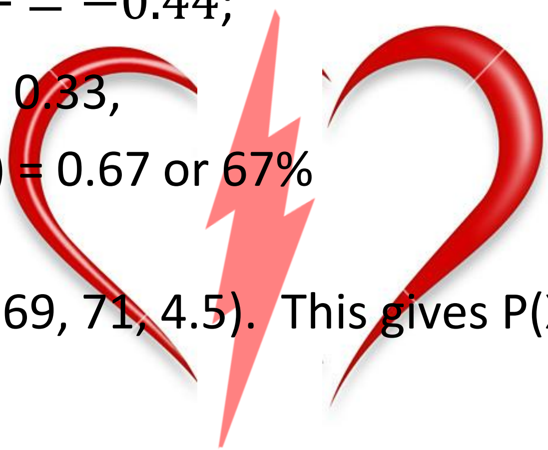
Attention Check

Q. Julie just realized that she wants her date to be taller when she is wearing her heels, which are 5" high. Find the new probability that her date will be taller.

A. $z = \frac{69-71}{4.5} = -0.44;$

P(Z<-0.44) = 0.33,

∴ P(Z>-0.44) = 0.67 or 67%



z	0.00	0.01	0.02	0.03	0.04	0.05	0.06	0.07	0.08	0.09
-3.4	0.0003	0.0003	0.0003	0.0003	0.0003	0.0003	0.0003	0.0003	0.0003	0.0002
-3.3	0.0005	0.0005	0.0005	0.0004	0.0004	0.0004	0.0004	0.0004	0.0004	0.0003
-3.2	0.0007	0.0007	0.0006	0.0006	0.0006	0.0006	0.0006	0.0005	0.0005	0.0005
-3.1	0.0010	0.0009	0.0009	0.0009	0.0008	0.0008	0.0008	0.0008	0.0007	0.0007
-3.0	0.0013	0.0013	0.0013	0.0012	0.0012	0.0011	0.0011	0.0011	0.0010	0.0010
-2.9	0.0019	0.0018	0.0018	0.0017	0.0016	0.0016	0.0015	0.0015	0.0014	0.0014
-2.8	0.0026	0.0025	0.0024	0.0023	0.0023	0.0022	0.0021	0.0021	0.0020	0.0019
-2.7	0.0035	0.0034	0.0033	0.0032	0.0031	0.0030	0.0029	0.0028	0.0027	0.0026
-2.6	0.0047	0.0045	0.0044	0.0043	0.0041	0.0040	0.0039	0.0038	0.0037	0.0036
-2.5	0.0062	0.0060	0.0059	0.0057	0.0055	0.0054	0.0052	0.0051	0.0049	0.0048
-2.4	0.0082	0.0080	0.0078	0.0075	0.0073	0.0071	0.0069	0.0068	0.0066	0.0064
-2.3	0.0107	0.0104	0.0102	0.0099	0.0096	0.0094	0.0091	0.0089	0.0087	0.0084
-2.2	0.0139	0.0136	0.0132	0.0129	0.0125	0.0122	0.0119	0.0116	0.0113	0.0110
-2.1	0.0179	0.0174	0.0170	0.0166	0.0162	0.0158	0.0154	0.0150	0.0146	0.0143
-2.0	0.0228	0.0222	0.0217	0.0212	0.0207	0.0202	0.0197	0.0192	0.0188	0.0183
-1.9	0.0287	0.0281	0.0274	0.0268	0.0262	0.0256	0.0250	0.0244	0.0239	0.0233
-1.8	0.0359	0.0351	0.0344	0.0336	0.0329	0.0322	0.0314	0.0307	0.0301	0.0294
-1.7	0.0446	0.0436	0.0427	0.0418	0.0409	0.0401	0.0392	0.0384	0.0375	0.0367
-1.6	0.0548	0.0537	0.0526	0.0516	0.0505	0.0495	0.0485	0.0475	0.0465	0.0455
-1.5	0.0668	0.0655	0.0643	0.0630	0.0618	0.0606	0.0594	0.0582	0.0571	0.0559
-1.4	0.0808	0.0793	0.0778	0.0764	0.0749	0.0735	0.0721	0.0708	0.0694	0.0681
-1.3	0.0968	0.0951	0.0934	0.0918	0.0901	0.0885	0.0869	0.0853	0.0838	0.0823
-1.2	0.1151	0.1131	0.1112	0.1093	0.1075	0.1056	0.1038	0.1020	0.1003	0.0985
-1.1	0.1357	0.1335	0.1314	0.1292	0.1271	0.1251	0.1230	0.1210	0.1190	0.1170
-1.0	0.1587	0.1562	0.1539	0.1515	0.1492	0.1469	0.1446	0.1423	0.1401	0.1379
-0.9	0.1841	0.1814	0.1788	0.1762	0.1736	0.1711	0.1685	0.1660	0.1635	0.1611
-0.8	0.2119	0.2090	0.2061	0.2033	0.2005	0.1977	0.1949	0.1922	0.1894	0.1867
-0.7	0.2420	0.2389	0.2358	0.2327	0.2296	0.2266	0.2236	0.2206	0.2177	0.2148
-0.6	0.2743	0.2709	0.2676	0.2643	0.2611	0.2578	0.2546	0.2514	0.2483	0.2451
-0.5	0.3085	0.3050	0.3015	0.2981	0.2946	0.2912	0.2877	0.2843	0.2810	0.2776
-0.4	0.3446	0.3409	0.3372	0.3336	0.3300	0.3264	0.3228	0.3192	0.3156	0.3121
-0.3	0.3821	0.3783	0.3745	0.3707	0.3669	0.3632	0.3594	0.3557	0.3520	0.3483
-0.2	0.4207	0.4168	0.4129	0.4090	0.4052	0.4013	0.3974	0.3936	0.3897	0.3859
-0.1	0.4602	0.4562	0.4522	0.4483	0.4443	0.4404	0.4364	0.4325	0.4286	0.4247
-0.0	0.5000	0.4960	0.4920	0.4880	0.4840	0.4801	0.4761	0.4721	0.4681	0.4641

A. 1-pnorm(69, 71, 4.5). This gives P(X>69) = 67%

Attention Check

Q. Julie wants to have at least 80% probability of finding the right guy. What is the maximum size of heels she can wear?

A. qnorm(0.20, 71, 4.5). This gives a value of 67.2”. As Julie is 64” tall, the maximum heel size she should wear is about 3”.

Attention Check

Q. Julie is convinced of the dangers of high heels and decides to stick with only 1” heels. What is the probability of finding the right guy now?

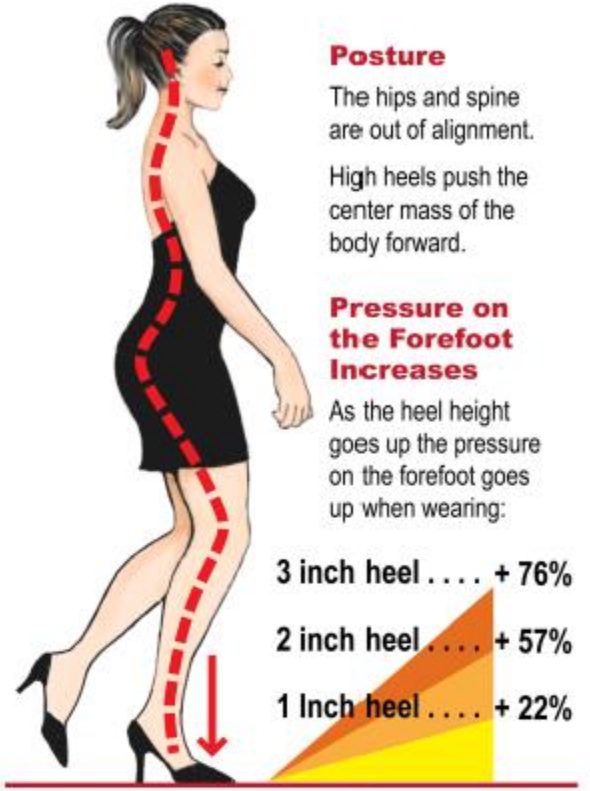


Figure 2

A. 1-pnorm(65, 71, 4.5). This gives a P(X>65)=90.9%.

PRIYANKA PRAVEEN

DECCAN CHRONICLE

FLAT REFUSAL

It's not just high heels that can be a pain, flat footwear is equally damaging

Almost everyone's favourite pair of 'killer' high heels have been notorious for bad posture and foot aches amongst other issue. Now reports say that its simple cousin — the flats — aren't really goody two shoes either.

Even celebrities like Victoria Beckham, who swear by their stilettos, have on quite a few occasions traded them for a pair of flats, but doctors feel that this really might not be the best thing for our feet. From agonising pain, spinal damage and even disorders — flats, are responsible for a host of problems.

Flats can cause spinal problems and inflammation of the thick band of tissues that connects the heel and the toes

"Our foot consists of the toes, the arch and the heel, this mechanism works so well that when we walk our entire weight is distributed equally," explains Dr Mithin Aachi, Senior Orthopedician. "The arch is what helps with the equal distribution of weight and so when we wear flat footwear unequal distribution of weight takes place and undue stress is put on the heel. This leads to several problems including plantar fasciitis and an inflammation of the thick band of tissues that connects the heel and the toes," he adds. In such cases, the pain is, several times, unbearable. Dr Praveen Rao, Orthopedic Surgeon, says, "When this happens, people find it difficult to walk after sitting for a long time."

Apart from pain, the lack of a cushioning and an arch in these footwear can eventually lead to spine troubles. "Since the pressure is on the heel, the gait of the person changes over the years and that leads to spinal problems and causes severe pain," explains Dr Rao.

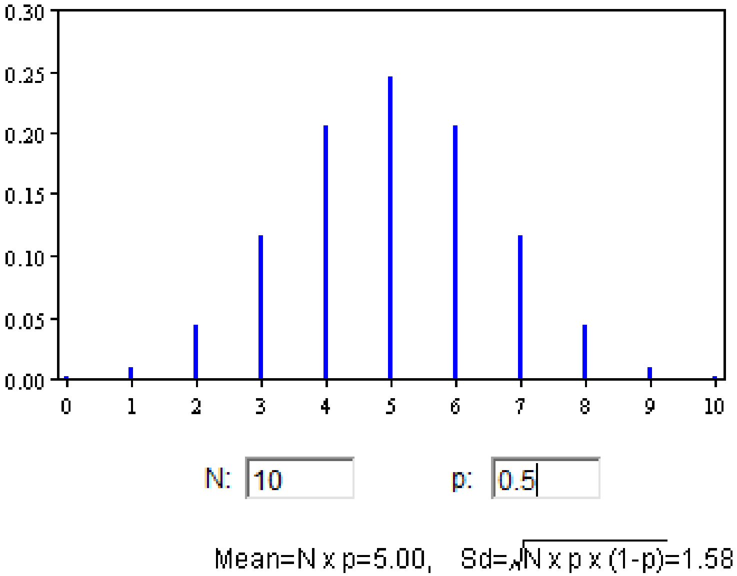
Doctors believe that we need to find a middle ground. "It's okay to wear high heels once in a while and since flats are more convenient, you can wear them occasionally, but you will need to find a balance. It helps to take a 'foot holiday' once a week by giving flats and heels a break and opting for an arched and cushioned footwear," explains Dr Aachi.

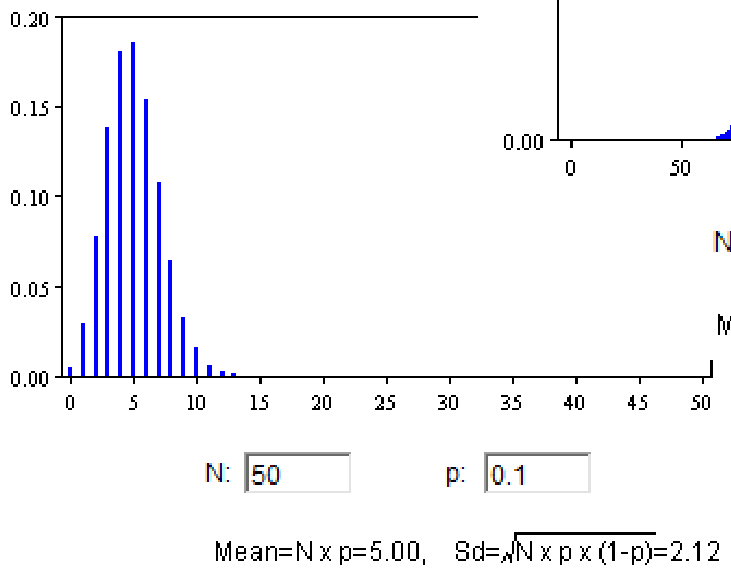
So, is there an ideal heel height that one needs to follow? "There isn't a number as such, but heels above one inch should be avoided regularly. Also wearing cushioned footwear with a small block-heel sometimes is fine," adds Dr. Rao.

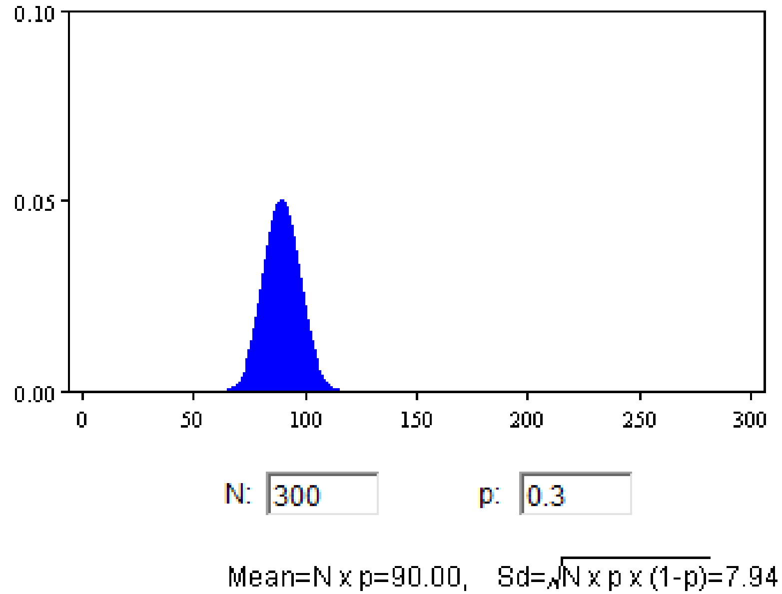
ALL TOO FLAT: Wearing flats regularly can be bad for your feet

Normal Distribution

Binomial distribution can be approximated to a Normal distribution if np>5 and nq>5.

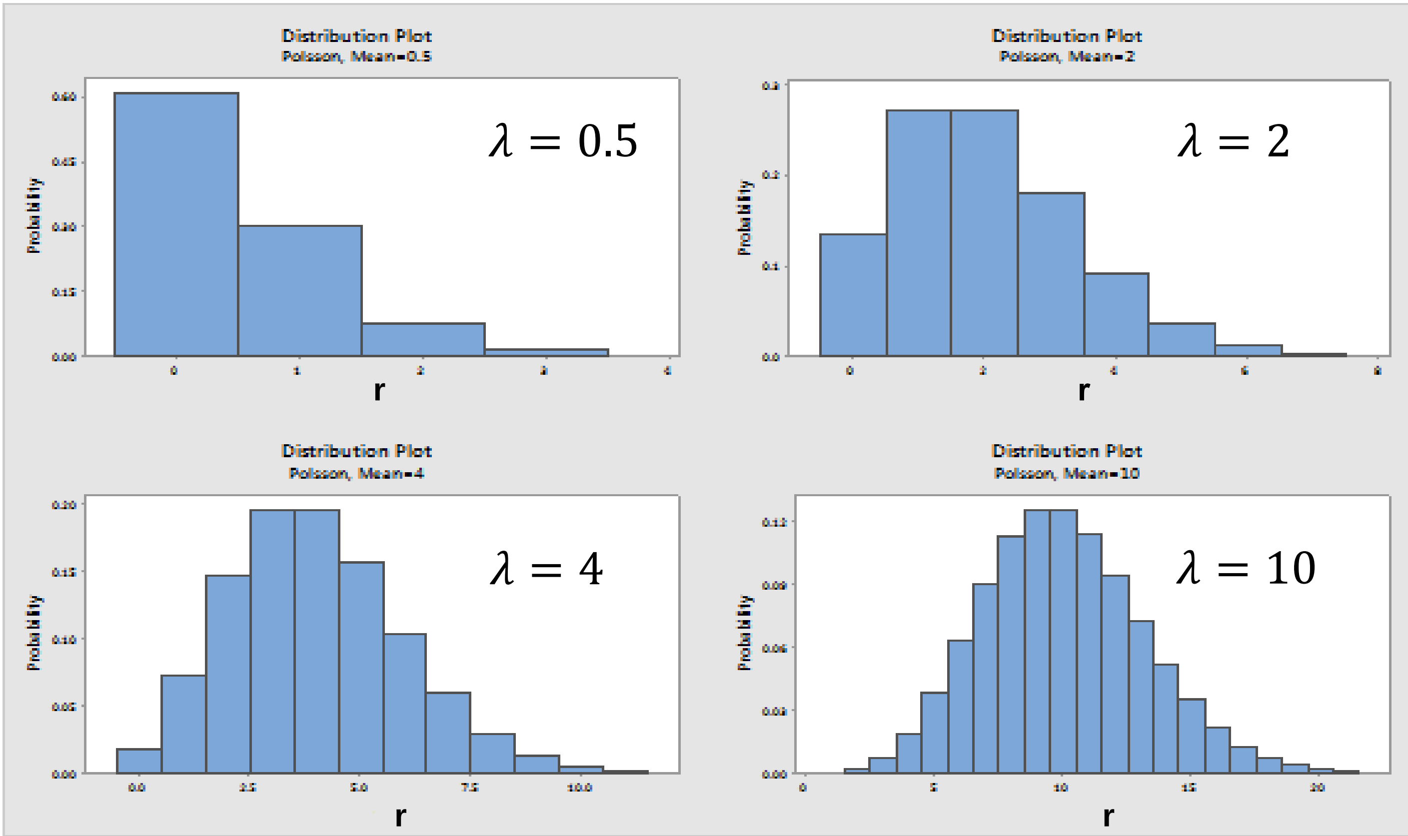


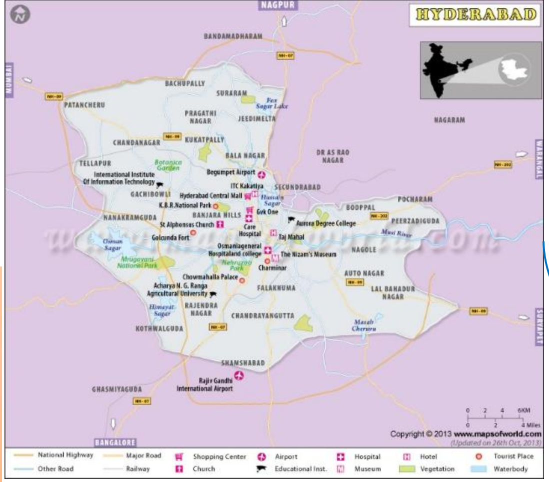




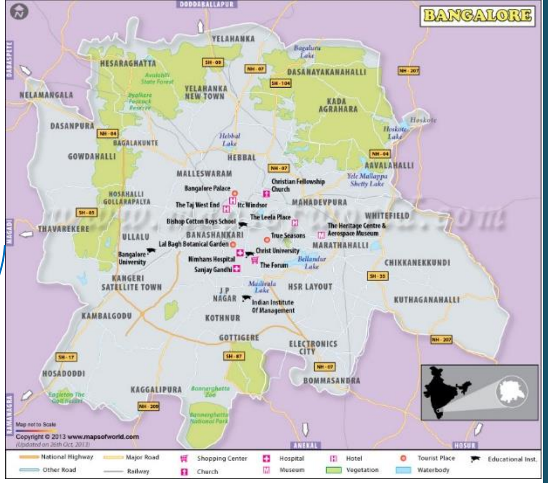
Normal Distribution

Poisson distribution can be approximated to a Normal distribution when $\lambda > 15$.





Inspire...Educate...Transform.



HYDERABAD
2nd Floor, Jyothi Imperial, Vamsiram Builders, Old Mumbai Highway, Gachibowli, Hyderabad - 500 032
+91-9701685511 (Individuals)
+91-9618483483 (Corporates)

BENGALURU
Floors 1-3, L77, 15th Cross Road, 3A Main Road, Sector 6, HSR Layout, Bengaluru – 560 102
+91-9502334561 (Individuals)
+91-9502799088 (Corporates)

Social Media

Web:	http://www.insofe.edu.in
Facebook:	https://www.facebook.com/insofe
Twitter:	https://twitter.com/Insofeedu
YouTube:	http://www.youtube.com/InsofeVideos
SlideShare:	http://www.slideshare.net/INSOFE
LinkedIn:	http://www.linkedin.com/company/international-school-of-engineering

This presentation may contain references to findings of various reports available in the public domain. INSOFE makes no representation as to their accuracy or that the organization subscribes to those findings.