Normal Distribution

(Gaussian Distribution)

<u>Lecture-04</u>

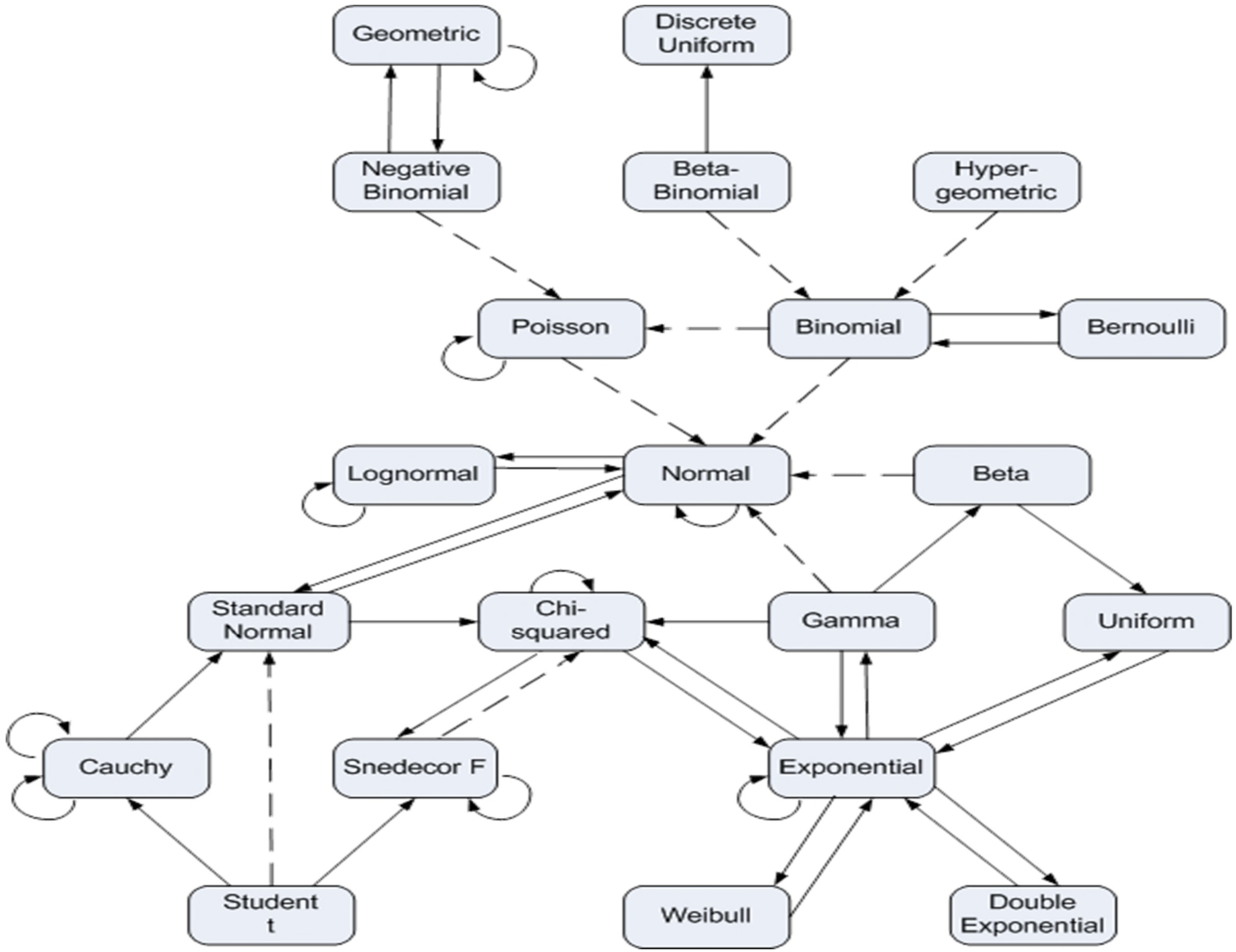
Dr. Tahseen A. Jilani
Postdoc (England-UK), PhD (UoK, HEC Indigenous program),
MSc(Statistics), MA(Economics), BSc(CS)

Associate Professor, Department of Computer Science, University of Karachi

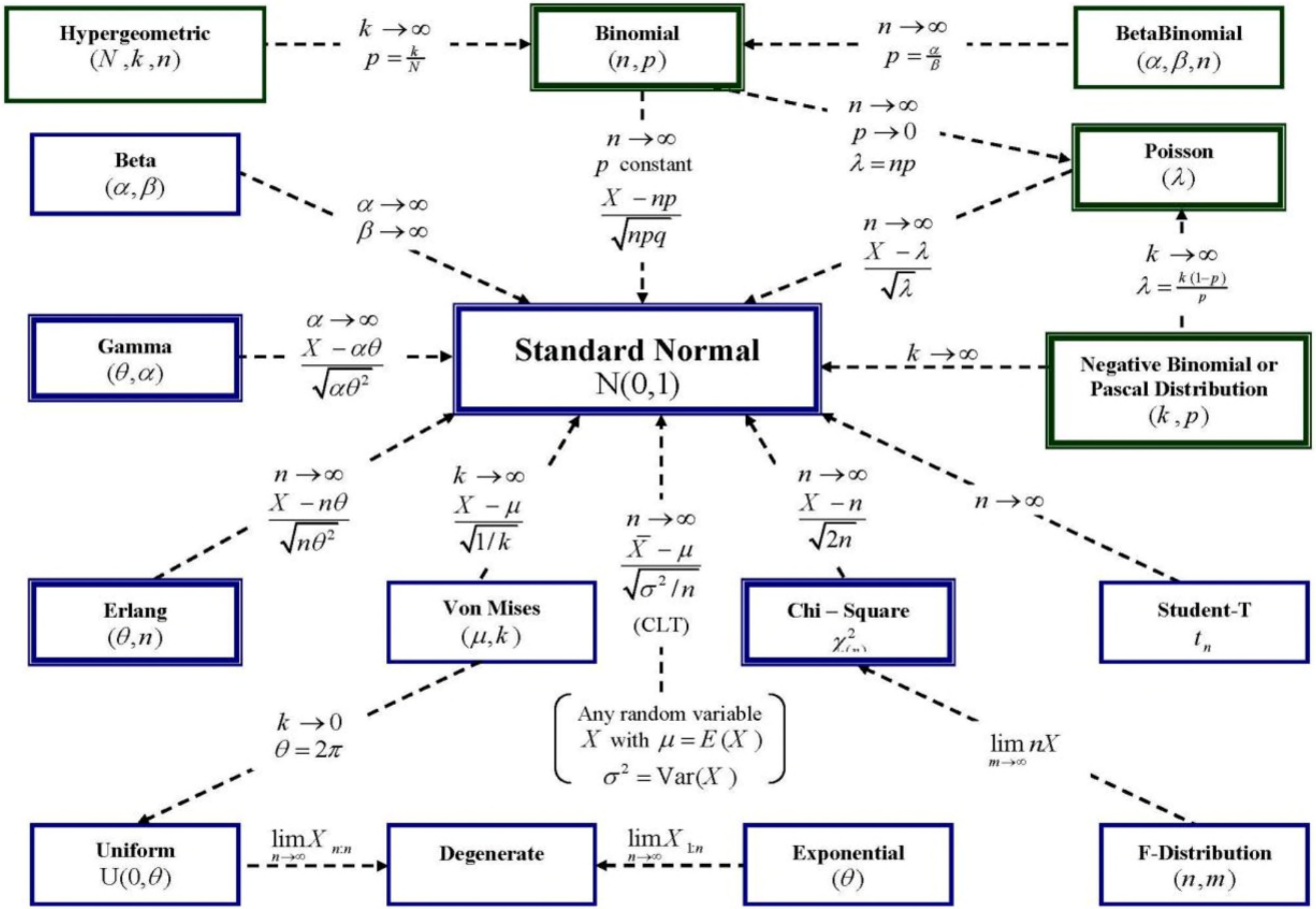
Some common probability distributions

- The most important distribution developed from Radial functions (in Mathematics).
- Various distributions are generated using Normal distribution such as Chi-square distribution. Chi-square further creates Erlang family, F and t distributions etc.
- For large sample size nearly every distribution converges to normal distribution.
- Heavily used in AI, Machine learning, business analytics, financial risk management, actuarial sciences, engineering systems, social sciences, environmental sciences etc.

Normal distribution is the Centre of all distributions



Asymptotic distributions ($n \to \infty$)



Some common probability distributions

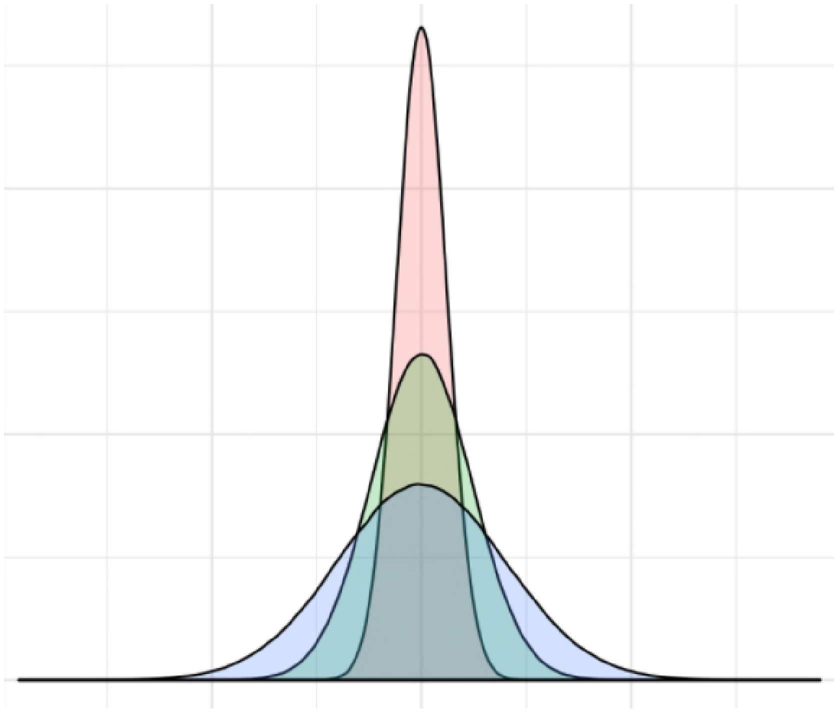
- Normal distribution x ~N(μ, σ^2) has two parameters, the mean μ and the variance σ^2.

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

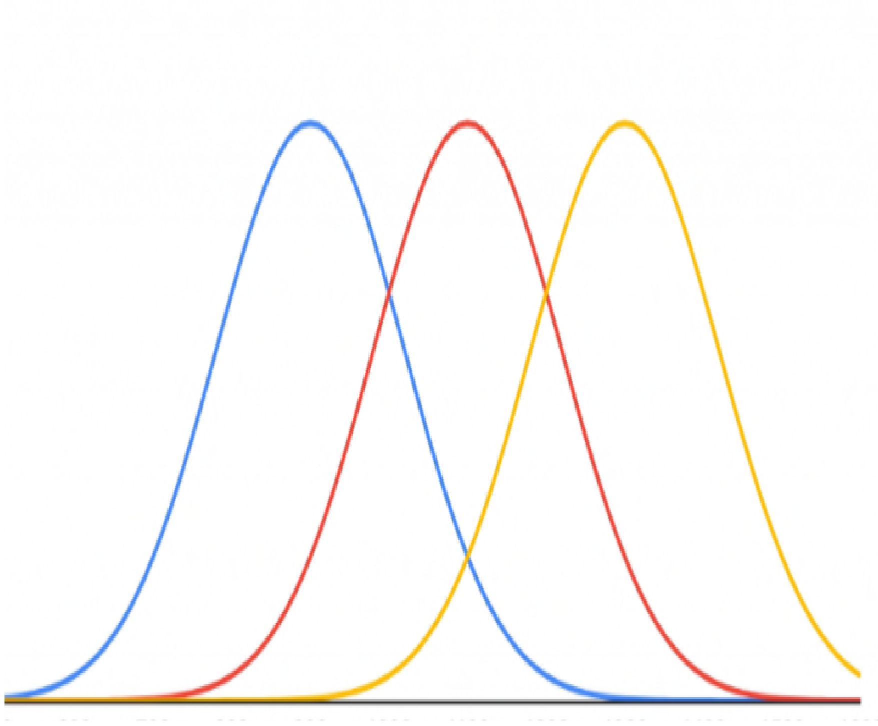
- Mean = median = mode
- Symmetric and Unimodal distribution.
- Sum of two normal random variables is again normal random variable.

Effect of Parameters (μ, σ) on Normal distribution

σ controls spread



μ controls location



(Left) Same mean with different standard deviations (spread).
(Right) Same standard deviation with different means (Locations).

Area under the Normal Curve

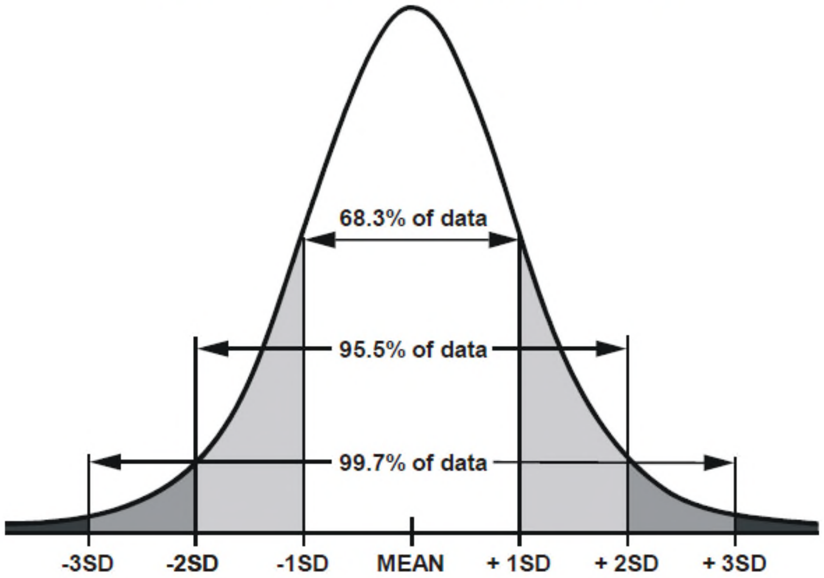
- Mean$\pm 1 S.D$ *contains* 68.26% *observations*.
- Mean$\pm 2 S.D$ *contains* 95.5% *observations*.
- Mean$\pm 3 S.D$ *contains* 99.72% *observations*.

$$\int_{\mu-\sigma}^{\mu+\sigma} \frac{1}{\sigma\sqrt{2\pi}} \bullet e^{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2} dx = .6826$$

$$\int_{\mu-2\sigma}^{\mu+2\sigma} \frac{1}{\sigma\sqrt{2\pi}} \bullet e^{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2} dx = .95$$

$$\int_{\mu-3\sigma}^{\mu+3\sigma} \frac{1}{\sigma\sqrt{2\pi}} \bullet e^{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2} dx = .9972$$

Areas under the normal curve that lie between 1, 2, and 3 standard deviations on each side of the mean



Standard Normal Distribution (sing Z-transformation)

- Cumulative density function for Normal distribution can not be obtained through integration.

$$F(x) = P(X \leq x) = \int_{-\infty}^{x} \frac{1}{\sqrt{2\pi\sigma^2}} e^{\frac{-1}{2}\left(\frac{x-\mu}{\sigma}\right)^2} dx$$

- Alternatively, one can use numerical (approximate) integration techniques to compute cumulative probabilities.
- One approach is to first standardise x ~ N(μ, σ^2) to z ~ N(0,1) using $z = \frac{x-\mu}{\sigma}$ (called z- transformation or z - scores). Therefore,

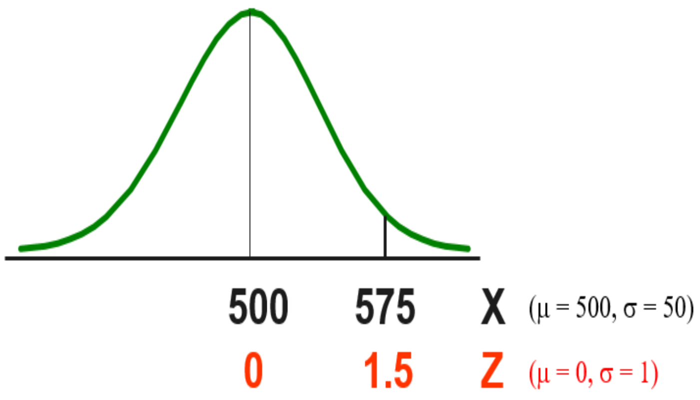
$$F(z) = P(Z \leq z) = \int_{-\infty}^{z} \frac{1}{\sqrt{2\pi}} e^{\frac{-1}{2}z^2} dz$$

Replaced using numerical integration as :

- $F(z) = P(Z \leq z) = \int_{-4.0}^{z} \frac{1}{\sqrt{2\pi}} e^{\frac{-1}{2}z^2} dz \quad -4.00 < z < 4.00$

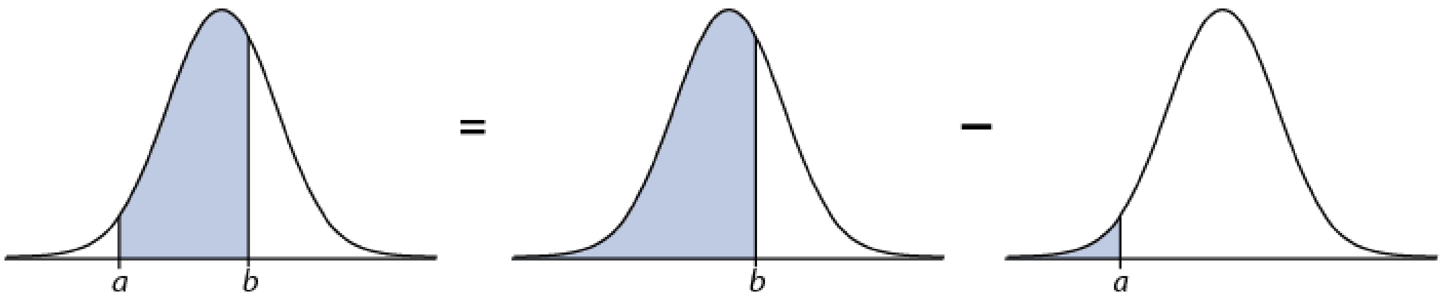
Standard Normal Distribution probability table

- Like Binomial and Poisson distributions, we have the standard Normal distribution table (z-table) to compute cumulative probabilities (area under the normal curve P(X$\leq x$).
- For $P(X \geq x) = 1 - P(X \leq x)$.



a represents a lower boundary and *b* represents an upper boundary

$$\Pr(a \leq Z \leq b) \quad = \quad \Pr(Z \leq b) \quad - \quad \Pr(Z \leq a)$$



Standard Normal distribution table/ Z-table

z	0	0.01	0.02	0.03	0.04	0.05	0.06	0.07	0.08	0.09
-0	.50000	.49601	.49202	.48803	.48405	.48006	.47608	.47210	.46812	.46414
-0.1	.46017	.45620	.45224	.44828	.44433	.44034	.43640	.43251	.42858	.42465
-0.2	.42074	.41683	.41294	.40905	.40517	.40129	.39743	.39358	.38974	.38591
-0.3	.38209	.37828	.37448	.37070	.36693	.36317	.35942	.35569	.35197	.34827
-0.4	.34458	.34090	.33724	.33360	.32997	.32636	.32276	.31918	.31561	.31207
-0.5	.30854	.30503	.30153	.29806	.29460	.29116	.28774	.28434	.28096	.27760
-0.6	.27425	.27093	.26763	.26435	.26109	.25785	.25463	.25143	.24825	.24510
-0.7	.24196	.23885	.23576	.23270	.22965	.22663	.22363	.22065	.21770	.21476
-0.8	.21186	.20897	.20611	.20327	.20045	.19766	.19489	.19215	.18943	.18673
-0.9	.18406	.18141	.17879	.17619	.17361	.17106	.16853	.16602	.16354	.16109
-1	.15866	.15625	.15386	.15151	.14917	.14686	.14457	.14231	.14007	.13786
-1.1	.13567	.13350	.13136	.12924	.12714	.12507	.12302	.12100	.11900	.11702
-1.2	.11507	.11314	.11123	.10935	.10749	.10565	.10383	.10204	.10027	.09853
-1.3	.09680	.09510	.09342	.09176	.09012	.08851	.08692	.08534	.08379	.08226
-1.4	.08076	.07927	.07780	.07636	.07493	.07353	.07215	.07078	.06944	.06811
-1.5	.06681	.06552	.06426	.06301	.06178	.06057	.05938	.05821	.05705	.05592
-1.6	.05480	.05370	.05262	.05155	.05050	.04947	.04846	.04746	.04648	.04551
-1.7	.04457	.04363	.04272	.04182	.04093	.04006	.03920	.03836	.03754	.03673
-1.8	.03593	.03515	.03438	.03362	.03288	.03216	.03144	.03074	.03005	.02938
-1.9	.02872	.02807	.02743	.02680	.02619	.02559	.02500	.02442	.02385	.02330
-2	.02275	.02222	.02169	.02118	.02068	.02018	.01970	.01923	.01876	.01831
-2.1	.01786	.01743	.01700	.01659	.01618	.01578	.01539	.01500	.01463	.01426
-2.2	.01390	.01355	.01321	.01287	.01255	.01222	.01191	.01160	.01130	.01101
-2.3	.01072	.01044	.01017	.00990	.00964	.00939	.00914	.00889	.00866	.00842
-2.4	.00820	.00798	.00776	.00755	.00734	.00714	.00695	.00676	.00657	.00639
-2.5	.00621	.00604	.00587	.00570	.00554	.00539	.00523	.00508	.00494	.00480
-2.6	.00466	.00453	.00440	.00427	.00415	.00402	.00391	.00379	.00368	.00357
-2.7	.00347	.00336	.00326	.00317	.00307	.00298	.00289	.00280	.00272	.00264
-2.8	.00256	.00248	.00240	.00233	.00226	.00219	.00212	.00205	.00199	.00193
-2.9	.00187	.00181	.00175	.00169	.00164	.00159	.00154	.00149	.00144	.00139
-3	.00135	.00131	.00126	.00122	.00118	.00114	.00111	.00107	.00104	.00100
-3.1	.00097	.00094	.00090	.00087	.00084	.00082	.00079	.00076	.00074	.00071
-3.2	.00069	.00066	.00064	.00062	.00060	.00058	.00056	.00054	.00052	.00050
-3.3	.00048	.00047	.00045	.00043	.00042	.00040	.00039	.00038	.00036	.00035
-3.4	.00034	.00032	.00031	.00030	.00029	.00028	.00027	.00026	.00025	.00024
-3.5	.00023	.00022	.00022	.00021	.00020	.00019	.00019	.00018	.00017	.00017
-3.6	.00016	.00015	.00015	.00014	.00014	.00013	.00013	.00012	.00012	.00011
-3.7	.00011	.00010	.00010	.00010	.00009	.00009	.00008	.00008	.00008	.00008
-3.8	.00007	.00007	.00007	.00006	.00006	.00006	.00006	.00005	.00005	.00005
-3.9	.00005	.00005	.00004	.00004	.00004	.00004	.00004	.00004	.00003	.00003
-4	.00003	.00003	.00003	.00003	.00003	.00003	.00002	.00002	.00002	.00002

Question-01

- X is a normally distributed variable with mean μ = 30 and standard deviation σ = 4. Find (a). P(x<40), (b). P(x>21), (c). P(30<x<35).
  - For P(x < 40), first calculate Z-value

$$z = \frac{x - \mu}{\sigma} = \frac{40 - 30}{4} = 2.5$$

  - Therefore P(X<40) is transformed to P(Z<2.50). Therefore,
  - P(Z<2.50)=0.9938 = P(X< 40).

z	0.00	0.01	0.02	0.03	0.04	0.05	0.06	0.07	0.08	0.09
0.0	0.5000	0.5040	0.5080	0.5120	0.5159	0.5199	0.5239	0.5279	0.5319	0.5359
0.1	0.5398	0.5438	0.5478	0.5517	0.5557	0.5596	0.5636	0.5675	0.5714	0.5753
0.2	0.5793	0.5832	0.5871	0.5910	0.5948	0.5987	0.6026	0.6064	0.6103	0.6141
0.3	0.6179	0.6217	0.6255	0.6293	0.6331	0.6368	0.6406	0.6443	0.6480	0.6517
0.4	0.6554	0.6591	0.6628	0.6664	0.6700	0.6736	0.6772	0.6808	0.6844	0.6879
0.5	0.6915	0.6950	0.6985	0.7019	0.7054	0.7088	0.7123	0.7157	0.7190	0.7224
0.6	0.7257	0.7291	0.7324	0.7357	0.7389	0.7422	0.7454	0.7486	0.7517	0.7549
0.7	0.7580	0.7611	0.7642	0.7673	0.7704	0.7734	0.7764	0.7794	0.7823	0.7854
0.8	0.7881	0.7910	0.7939	0.7967	0.7995	0.8023	0.8051	0.8078	0.8106	0.8133
0.9	0.8159	0.8186	0.8212	0.8238	0.8264	0.8289	0.8315	0.8340	0.8365	0.8389
1.0	0.8413	0.8438	0.8461	0.8485	0.8508	0.8531	0.8554	0.8577	0.8599	0.8621
1.1	0.8643	0.8665	0.8686	0.8708	0.8729	0.8749	0.8770	0.8790	0.8804	0.8830
1.2	0.8849	0.8869	0.8888	0.8907	0.8925	0.8944	0.8962	0.8980	0.8997	0.9015
1.3	0.9032	0.9049	0.9066	0.9082	0.9099	0.9115	0.9131	0.9147	0.9162	0.9177
1.4	0.9192	0.9207	0.9222	0.9236	0.9251	0.9265	0.9279	0.9292	0.9306	0.9319
1.5	0.9332	0.9345	0.9357	0.9370	0.9382	0.9394	0.9406	0.9418	0.9429	0.9441
1.6	0.9452	0.9463	0.9474	0.9484	0.9495	0.9505	0.9515	0.9525	0.9535	0.9545
1.7	0.9554	0.9564	0.9573	0.9582	0.9591	0.9599	0.9608	0.9616	0.9625	0.9633
1.8	0.9641	0.9649	0.9656	0.9664	0.9671	0.9678	0.9686	0.9693	0.9699	0.9706
1.9	0.9713	0.9719	0.9726	0.9732	0.9738	0.9744	0.9750	0.9756	0.9761	0.9767
2.0	0.9773	0.9778	0.9783	0.9788	0.9793	0.9798	0.9803	0.9808	0.9812	0.9817
2.1	0.9821	0.9826	0.9830	0.9834	0.9838	0.9842	0.9846	0.9850	0.9854	0.9857
2.2	0.9861	0.9865	0.9868	0.9871	0.9874	0.9878	0.9881	0.9884	0.9887	0.9890
2.3	0.9893	0.9896	0.9898	0.9901	0.9904	0.9906	0.9909	0.9911	0.9913	0.9916
2.4	0.9918	0.9920	0.9922	0.9924	0.9927	0.9929	0.9931	0.9932	0.9934	0.9936
2.5	0.9938	0.9940	0.9941	0.9943	0.9945	0.9946	0.9948	0.9949	0.9951	0.9952
2.6	0.9953	0.9955	0.9956	0.9957	0.9959	0.9960	0.9961	0.9962	0.9963	0.9964
2.7	0.9965	0.9966	0.9967	0.9968	0.9969	0.9970	0.9971	0.9972	0.9973	0.9974
2.8	0.9974	0.9975	0.9976	0.9977	0.9977	0.9978	0.9979	0.9980	0.9980	0.9981
2.9	0.9981	0.9982	0.9982	0.9983	0.9984	0.9984	0.9985	0.9985	0.9986	0.9986

Try these

- P(x>21) = 1 – P(X<21)
- P(30<x<35)= P(X<35)-P(x<30)

Question-02

- A radar unit is used to measure speeds of cars on a motorway. The speeds are normally distributed with a mean of 90 km/hr and a standard deviation of 10 km/hr. What is the probability that a car picked at random is travelling at more than 100 km/hr?

- Here $\mu = 90\,\frac{km}{hr}\, and\ \sigma = 10\,\frac{km}{hr}$.
- P(X>100) = 1- P(X<100)

$$= 1 - \ P\left(\frac{x-\mu}{\sigma} < \frac{100-90}{10}\right)$$
$$= 1 = P(Z < 1.00)$$
$$= 1 - 0.8413 = 0.1587$$

- You can use online calculator for this calculation during classes.

https://onlinestatbook.com/2/calculators/normal_dist.html

Problem: (Test your understanding)

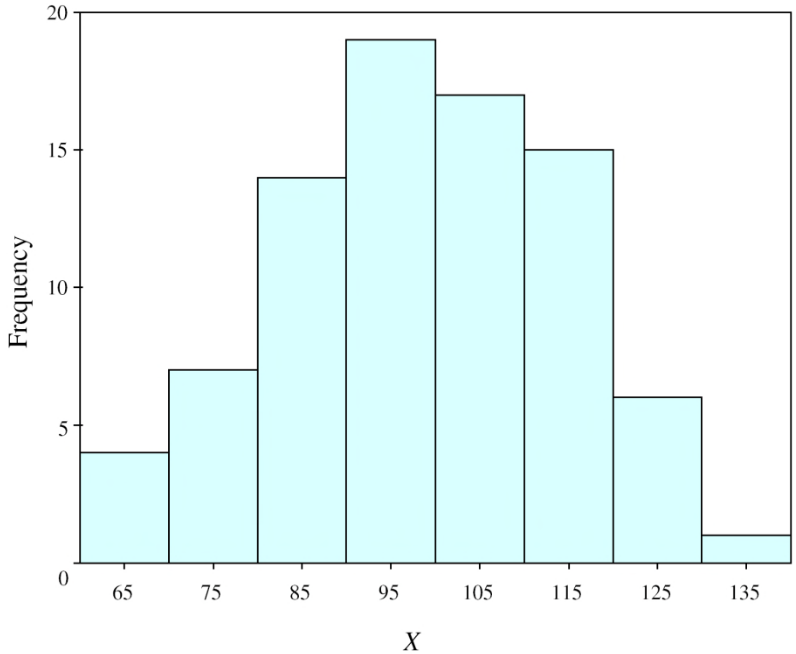
- For a certain type of computers, the length of time between charges of the battery is normally distributed with a mean of 50 hours and a standard deviation of 15 hours. John owns one of these computers and wants to know the probability that the length of time will be between 50 and 70 hours.

Applied Data Science and normal distribution

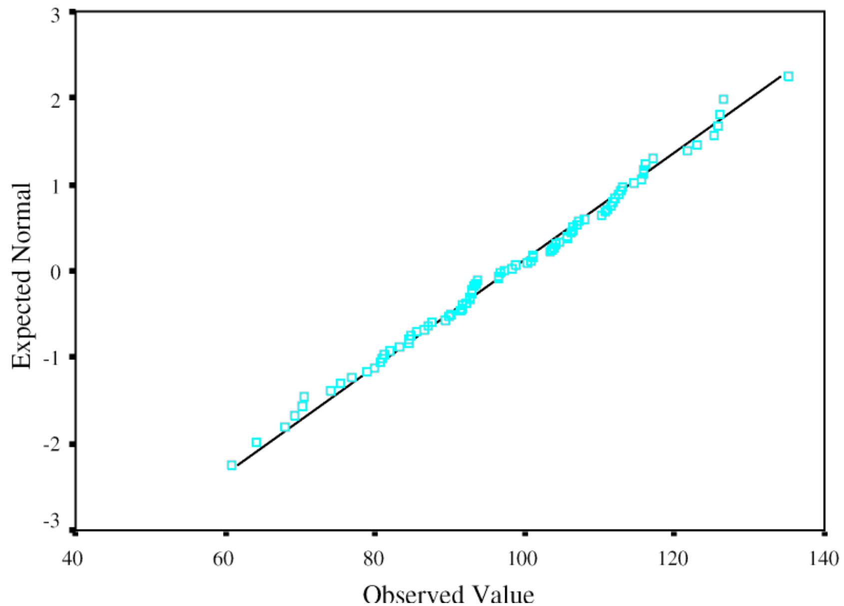
1. Look at the histogram! Does it appear bell shaped?
2. Compute descriptive summary measures—are mean, median, and mode similar?
3. Do 2/3 of observations lie within 1 SD dev of the mean? Do 95% of observations lie within 2 SD dev of the mean?
4. Look at a normal probability plot—is it approximately linear?
5. Run tests of normality (such as Kolmogorov-Smirnov). But, be cautious, highly influenced by sample size!
6. If the data is not following normal distribution, try different transformations.

Assessing Departures from Normality

Approximately Normal histogram

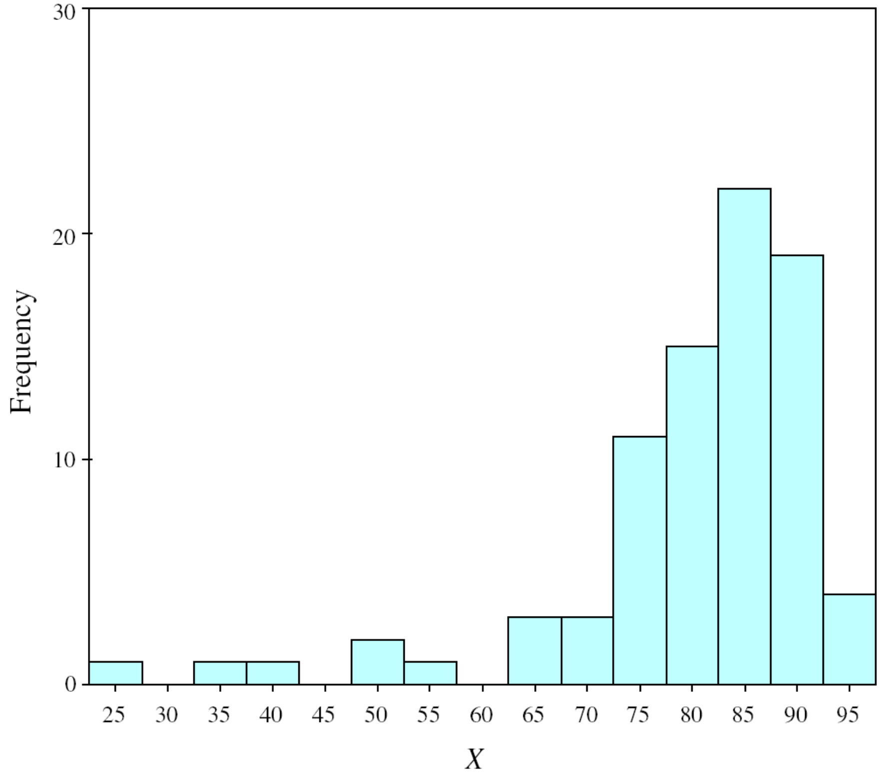


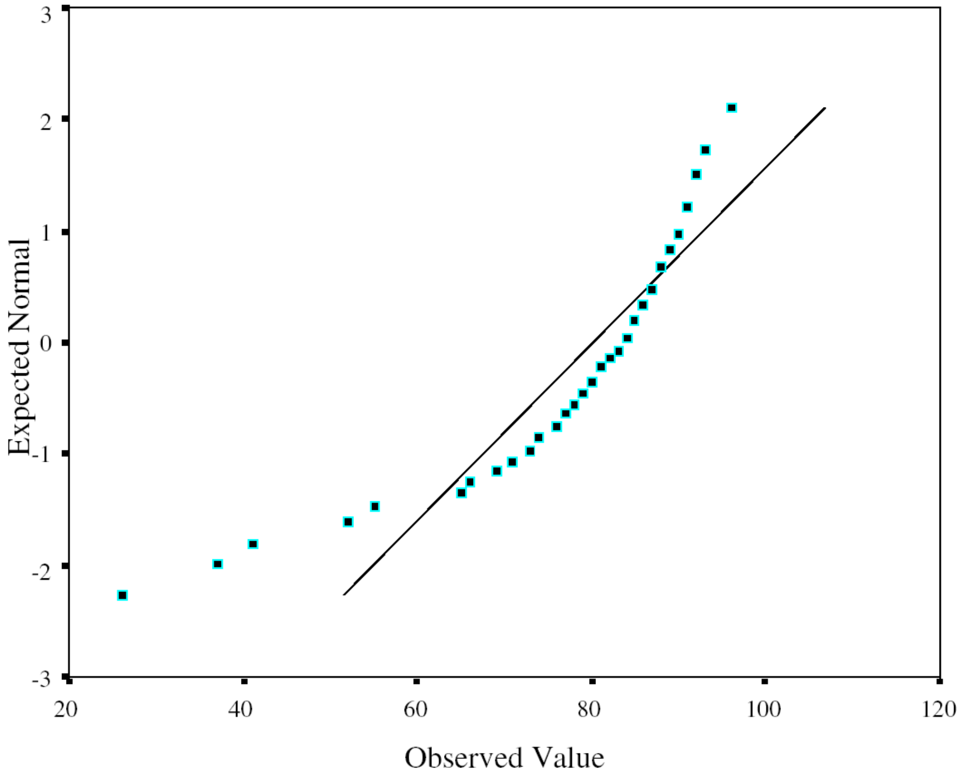
Same distribution on Normal "Q-Q" Plot



Normal distributions adhere to diagonal line on Q-Q plot

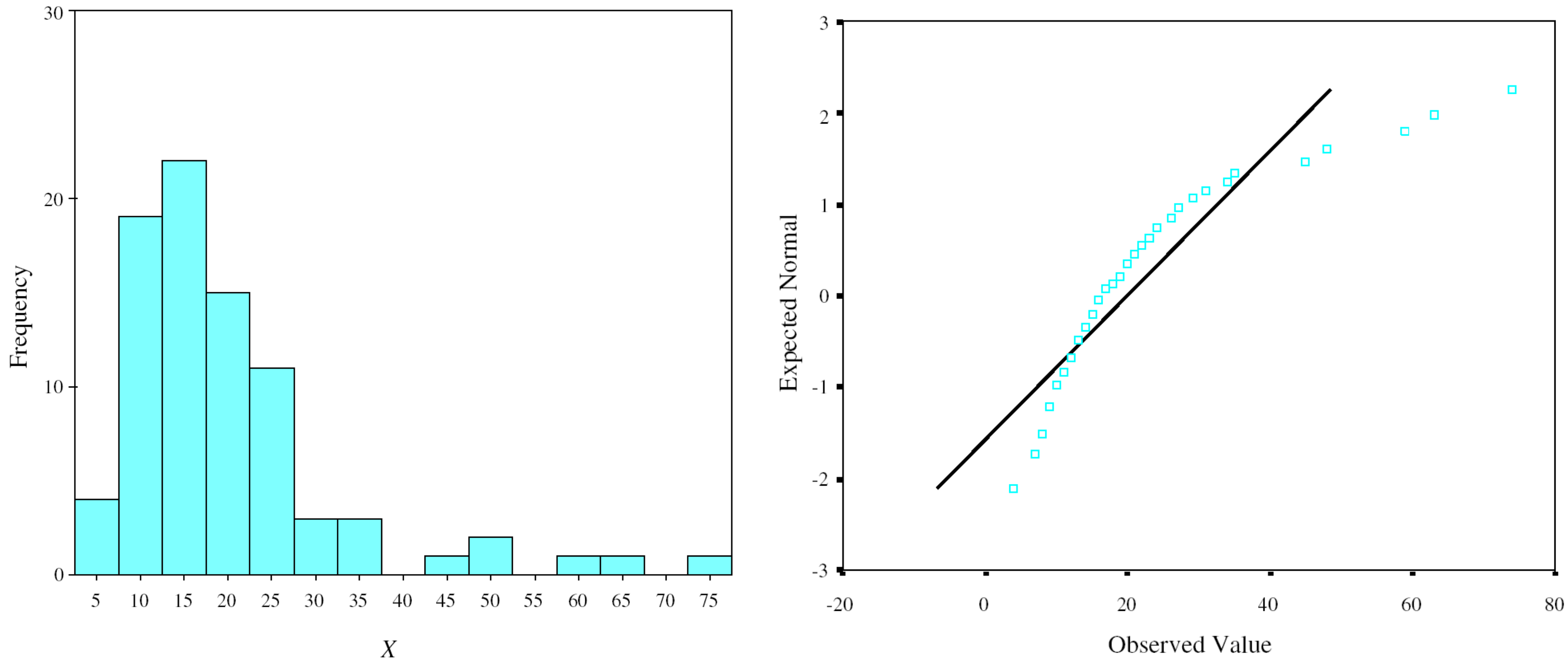
Negative Skew





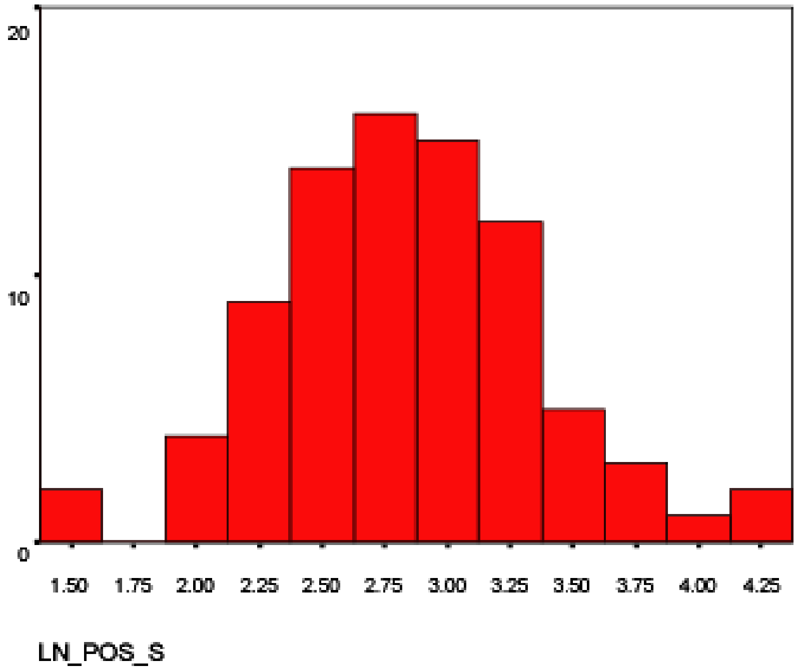
Negative skew shows upward curve on Q-Q plot

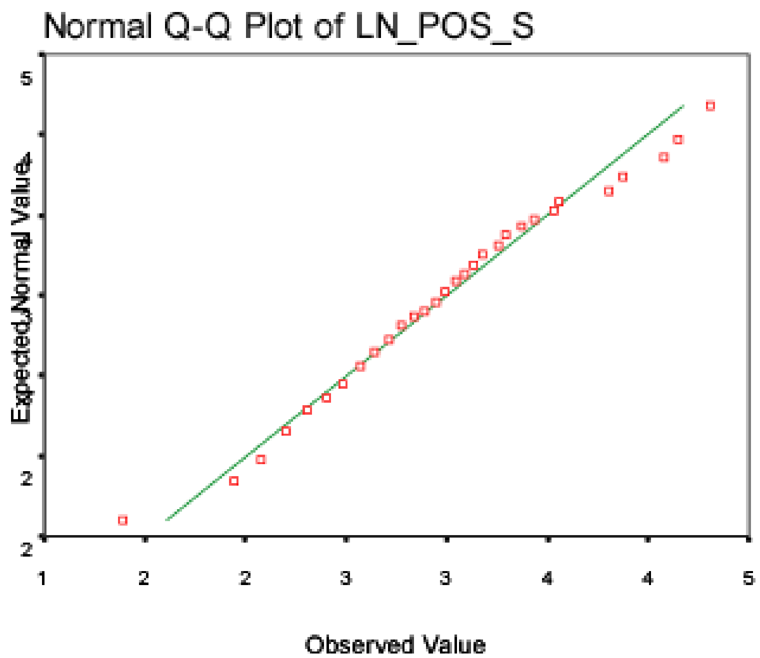
Positive Skew



Positive skew shows downward curve on Q-Q plot

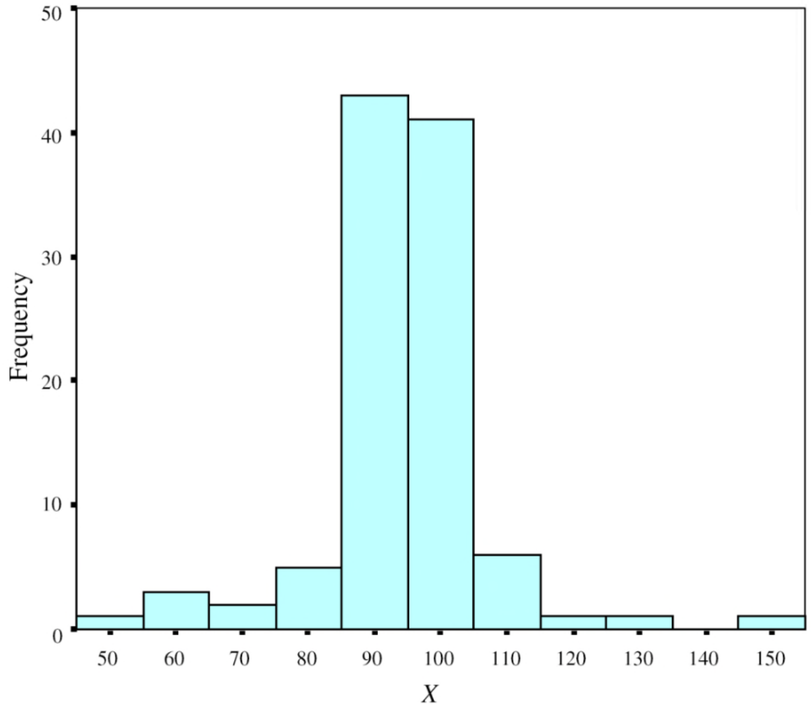
Same data as prior slide with log-transformation

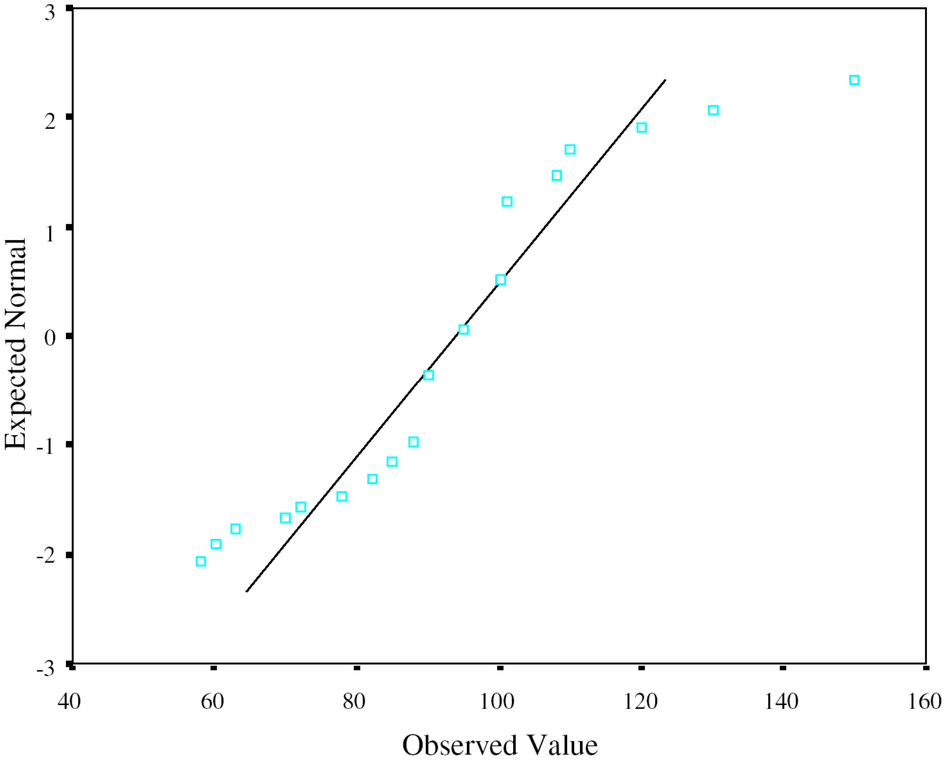




The log transform Normalize the skew

Leptokurtotic





Leptokurtotic distribution show S-shape on Q-Q plot

Normal Approximation to Binomial and Poisson distributions

Normal approximation to Binomial distribution

- If the random variable X has the binomial distribution with parameters n and p then for large n, and moderate p, the distribution of

$$z = \frac{x - np}{\sqrt{npq}}$$

- Has standard normal distribution with mean=0 and st.dev.=1 where np=mean and npq=variance.
- We use the normal approximation to evaluate binomial probabilities whenever p is not close to 0 or 1.
- The approximation of normal using binomial is excellent when n is large and fairly good for small values of n if p is reasonably close to 0.5.
- If both np and nq are greater than or equal to 5, then the approximation will be good.

Continuity Correction

- As we are approximating a discrete distribution using a continuous distribution, therefore a correction is made which is called continuity correction.

$$P(x=a) \Rightarrow P(a-0.5 < X < a+0.5)$$

$$P(x \le a) \Rightarrow P(X < a+0.5)$$

$$P(x < a) \Rightarrow P(X < a-0.5)$$

$$P(a \le x \le b) \Rightarrow P(a-0.5 \le X \le b+0.5)$$

$$P(a < x \le b) \Rightarrow P(a+0.5 \le X \le b+0.5)$$

$$P(a \le x < b) \Rightarrow P(a-0.5 \le X \le b-0.5)$$

and if the measurements are recorded upto one decimal place, then the values are considered as mod point of the interval $a \pm 0.5$

Problem (Normal approximation to Binomial distribution)

- It is known that in a sack of mixed grass seeds 35% are ryegrass. Use the normal approximation to the binomial distribution to find the probability that in a sample of 400 seeds there are
  - Less than 120 ryegrass seeds.
  - Between 120 and 150 ryegrass seeds (inclusive).
  - More than 160 ryegrass seeds.

Let X be the r.v. 'the number of ryegrass seeds'. Here n=400 and p=0.35. As n is large and p (or q) is not close to 1, therefore we can apply normal approximation with mean =np=(400)(0.35)=140 and variance= npq =(400)(0.35)(0.65)=91. SO X~N(140,91).

- $P(X<120)$ becomes $P(X<119.5)$ and use Normal distribution.
- $P(120<=X<=150)$ becomes $P(119.5 <=X<=150.5)$.
- $P(X>160)$ becomes $P(X>=160.5)$.

The normal approximation to the Poisson distribution

- If $X \sim Poisson(\lambda)$ then $E(X)=Var(X)= \lambda$. Now for large λ, we have $X \sim N(\lambda, \lambda)$ approximately. Generally, we require $\lambda > 20$ for a good approximation.
  - A radioactive disintegration gives counts that follows a Poisson distribution with mean count per second of 25. Find the probability that in 1 second the count is between 23 and 27 inclusive.
  - $P(23<=X<=27) = P(22.5 <X<27.5)= P(-0.5<z<0.5) = 0.383$
  - So the probability that the count is between 23 and 27 inclusive is 0.383.
- NOTE: If we solve this problem using Poisson distribution, we will get a nearly close answer. But conceptually not correct.

When to Use the Different Approximations

	Distribution of X	Restrictions on parameters	Approximation
(1)	$X \sim Bin(n,p)$	n large (say $n>20$) and p small (say $p<0.05$)	$X \sim Po(np)$
(2)	$X \sim Bin(n,p)$	$n>20$, p close to 0.5	$X \sim N(np,npq)$
(3)	$X \sim Po(\lambda)$	$\lambda>20$ (say)	$X \sim N(\lambda, \lambda)$

End of lesson-04

Thank you