Statistics Bootcamp using R

DAY 2 DATA VISUALIZATION & UNDERSTANDING PATTERN

2.3 INTRODUCTION TO NORMAL DISTRIBUTION

GU Zhan (Sam)
Institute of Systems Science
National University of Singapore

issgz@nus.edu.sg

Agenda

Day 2 : Data Visualization & Understanding Pattern

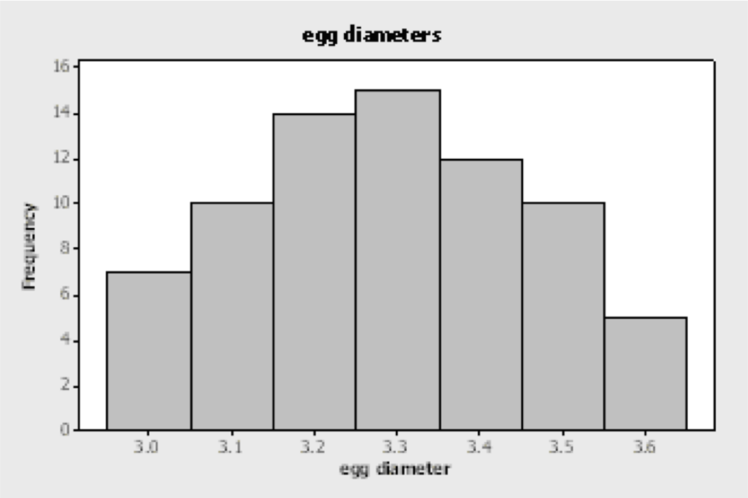
- Data Visualization
- Descriptive Statistics & Sampling
- **Introduction to Normal Distribution**

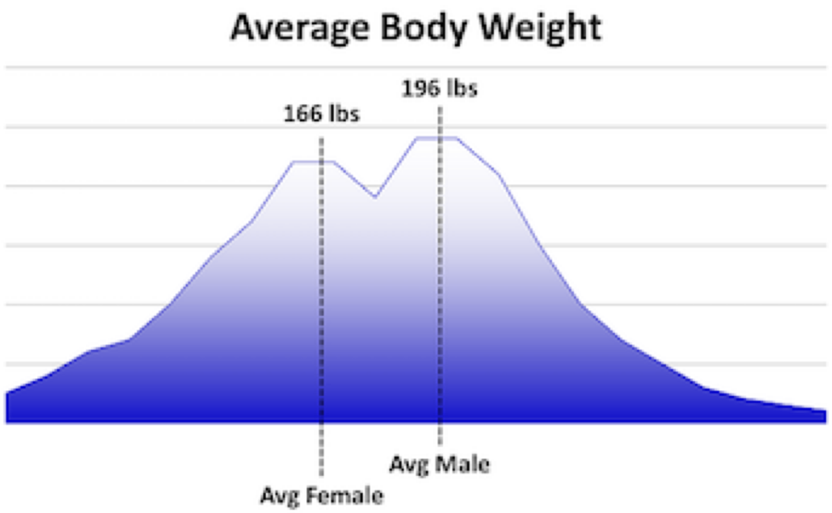
Learning objectives

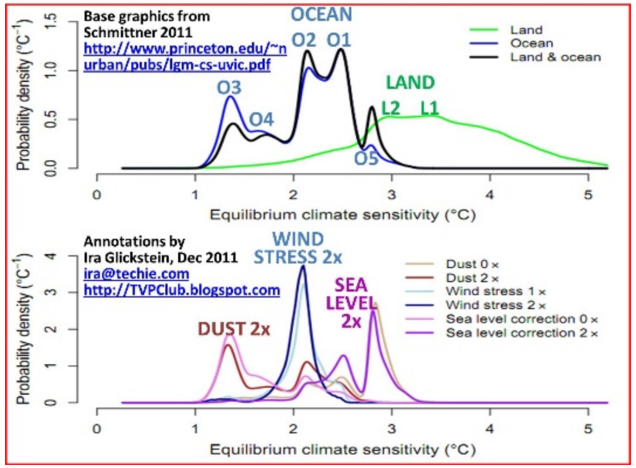
- Understand Normal Distribution & Central Limit Theorem (CLT)
- Understand distributions other than normal distribution
- Understand confidence interval (CI) and its applications for statistical inference
- Learn how to calculate CI & test normality using R

The Normal Distribution

- Usually when we take a sample of data from some experiment, we don't know how it is distributed
- It could be uni-modal Or bi-modal or multimodal ……….



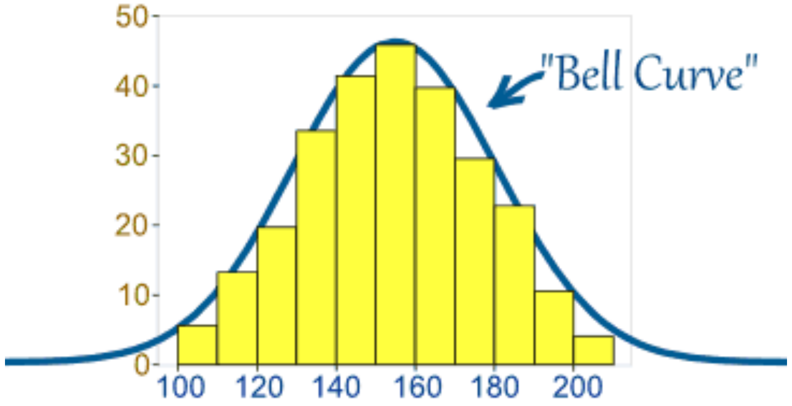




The Normal Distribution (continued)

The normal distribution (Gaussian Distribution) is the most common continuous distribution used in statistics. The main three reasons for this are as follows:

- Numerous continuous variables common in business have distributions that closely resembles normal distribution
- The normal distribution can be used to approximate various discrete probability distributions
- The normal distribution provides the basis for classical statistical inference because of its relationship to the **Central Limit Theorem**



Examples : IQ Score, Height, "Employees of a Company: No. of years worked" … all of these have approximately normal distribution

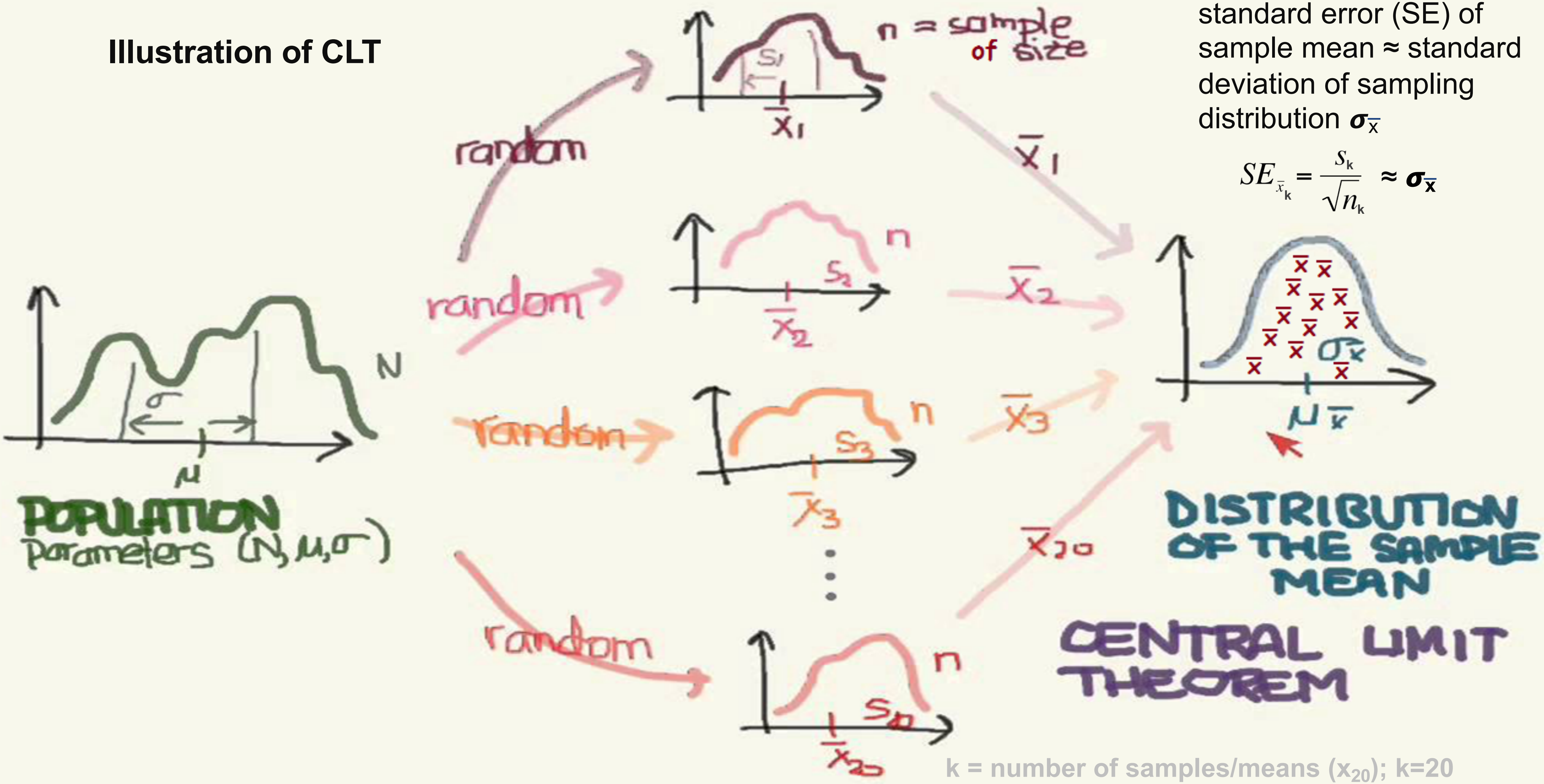
Central Limit Theorem (CLT)

If a random variable has a finite mean and variance, then the average of an independent sample of the random variable (draw many independent samples to calculate many average values) will approximately follow a normal distribution, regardless of the original distribution (population distribution) of the random variable itself, provided the sample size (**n**) is sufficiently large.

Illustration of CLT

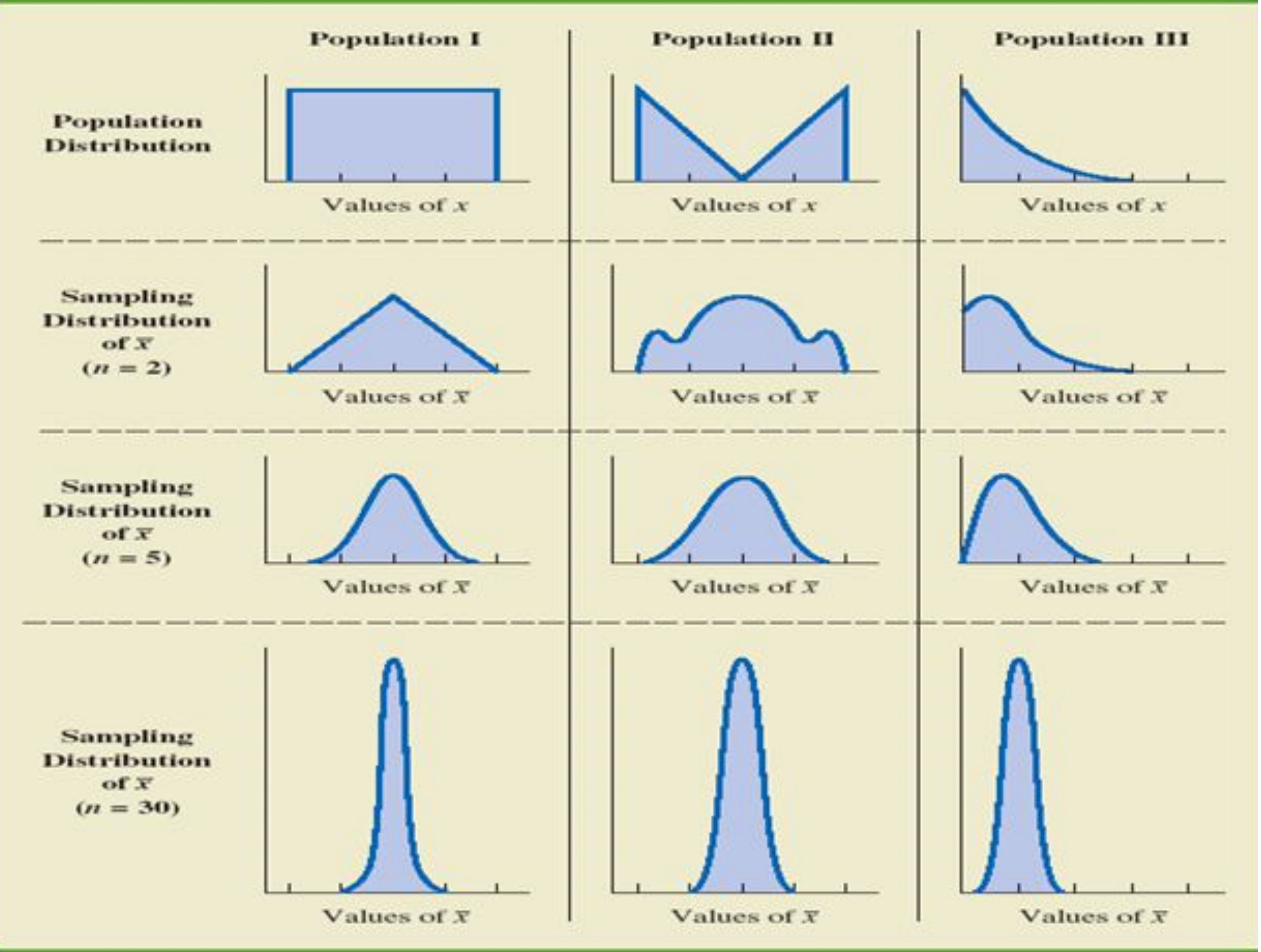
standard error (SE) of sample mean ≈ standard deviation of sampling distribution $\boldsymbol{\sigma}_{\overline{x}}$

$$SE_{\bar{x}_k} = \frac{s_k}{\sqrt{n_k}} \approx \boldsymbol{\sigma}_{\overline{x}}$$



k = number of samples/means (x_{20}); k=20

source: https://medium.com/@birajparikh/what-is-central-limit-theorem-clt-db3679433dcb



NUS National University of Singapore | ISS INSTITUTE OF SYSTEMS SCIENCE

Illustration of CLT

Continuous Probability Distributions

- A probability density function is a mathematical expression that defines the distribution of the values for a **continuous (numeric) variable**

- The most commonly used (other than normal) are Beta, Cauchy, Chi-square, Exponential, Gamma, **Logistic**, Pareto, Pearson, Uniform and Weibull Distribution

Normal Probability Density Function :

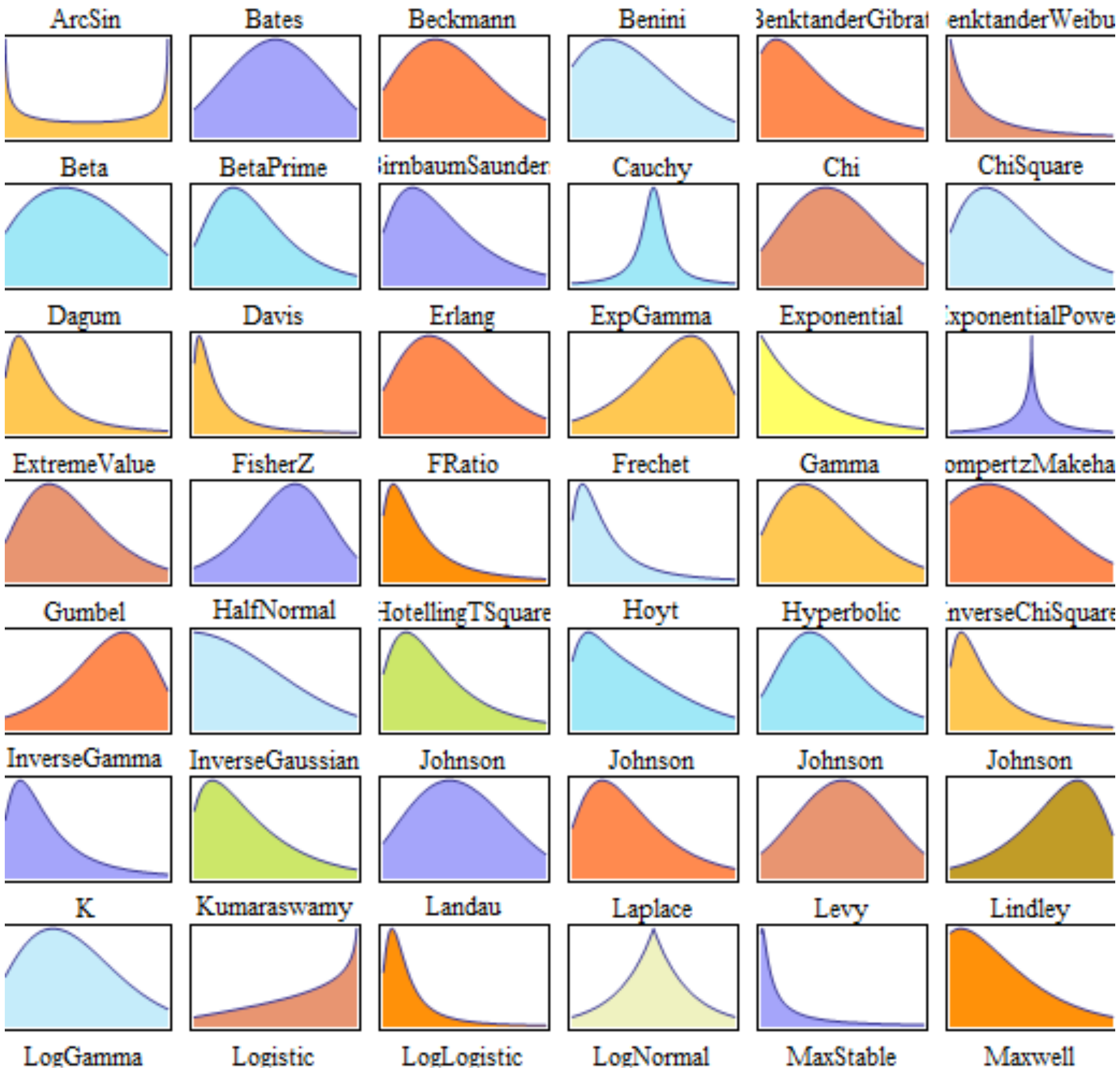
(Intuition: Distribution/Histogram normalized to 1 or 100%)

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

$-\infty < x < +\infty$

e = 2.71828

π = 3.14159



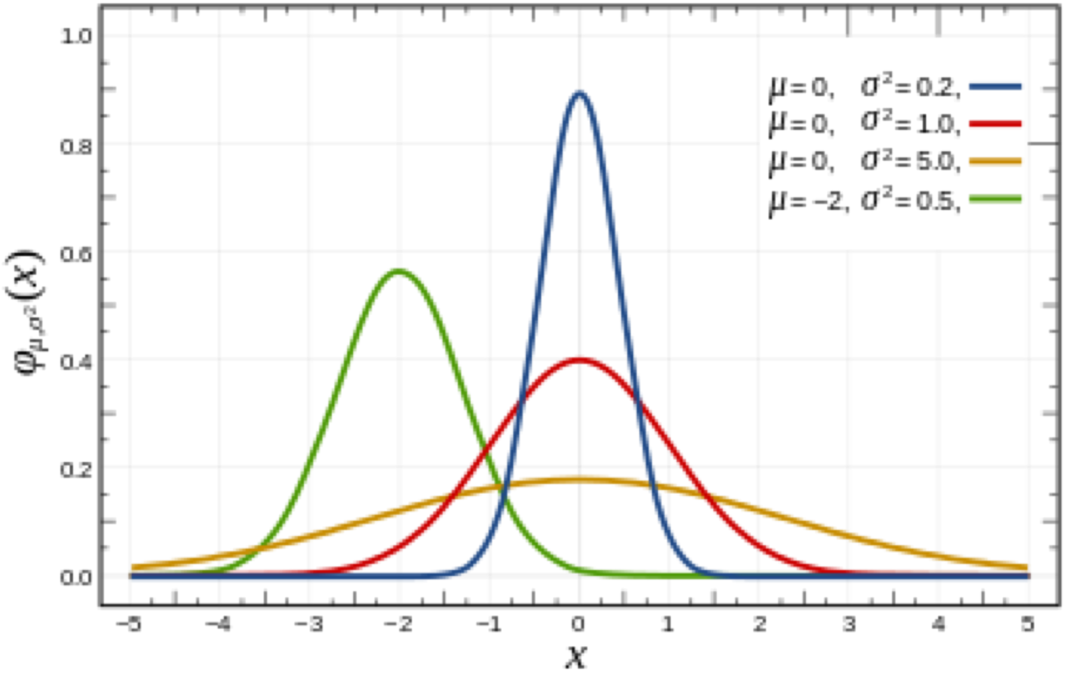
Characteristics of Normal Distribution

Standardization of a normally distributed variable X :

X is normally distributed with mean μ and standard deviation σ whereas Z is normally distributed with mean 0 and standard deviation 1

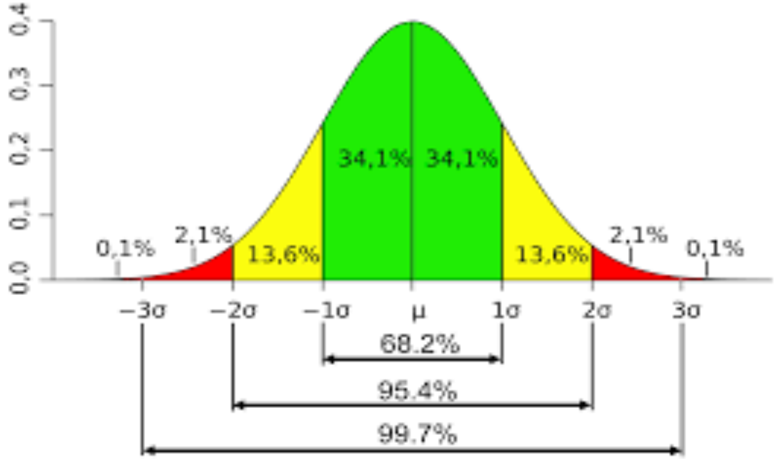
Z distribution is called standard normal distribution, with zero mean, unit variance.

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



Properties of Normal Distribution

- The mean, median and mode are equal
- The normal curve is bell shaped and symmetric about the mean
- The Total area under the curve is 1 (100% probability)
- The normal curve approaches but **never** touches the x-axis and it extends further and further away from the mean



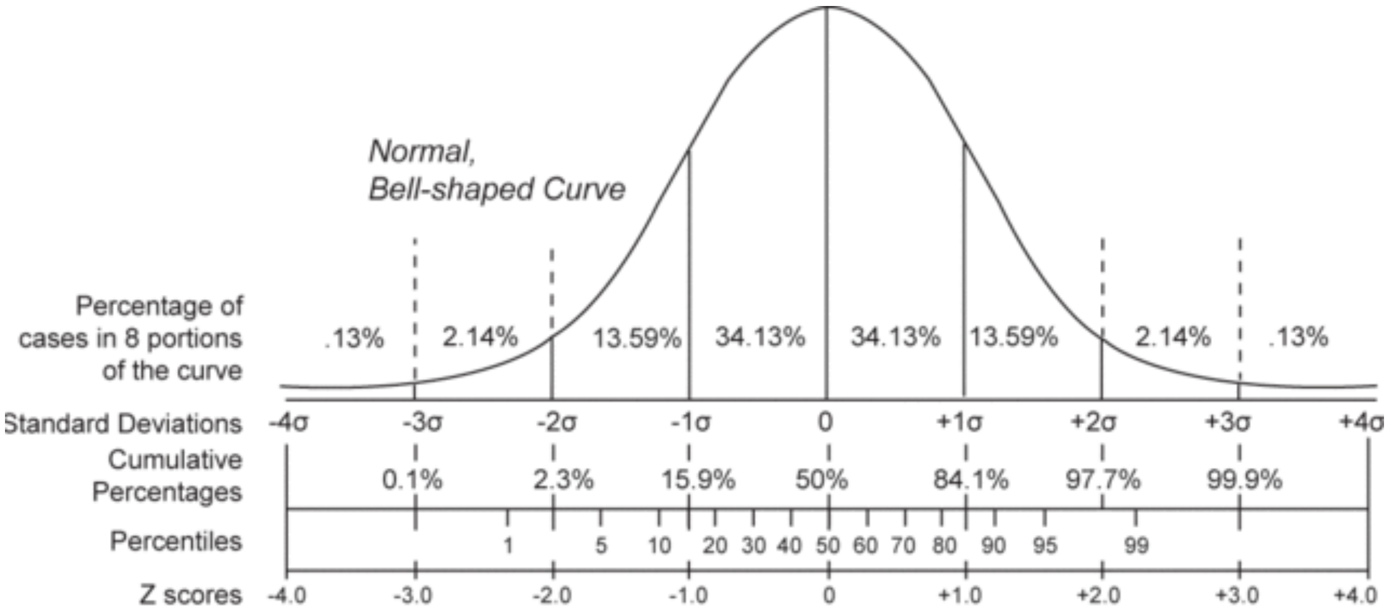
Normal Distribution To Calculate Probability

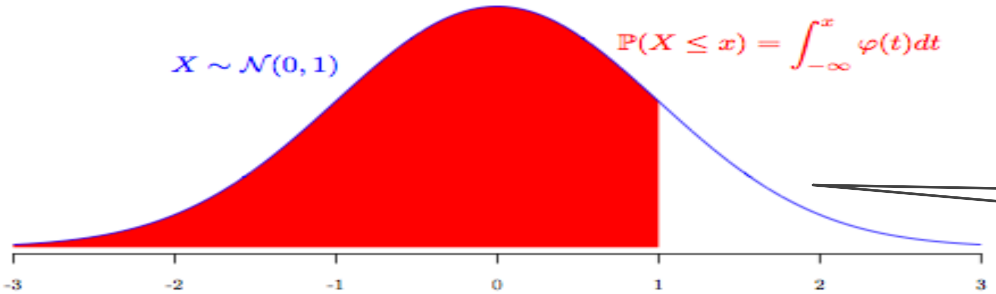
Example: for a normal distribution with parameters μ and σ.

So we know most of the features of the normal distribution, including the probabilities we will get values outside a certain range

- We can calculate the probability of getting a value outside the range
  - (μ - 1.64*σ, μ + 1.64*σ) = 10% → 90% confidence interval
- We can calculate the probability of getting a value outside the range
  - (μ - 1.96*σ, μ + 1.96*σ) = 05% → 95% confidence interval

So we can estimates a statistic with probabilistic confidence represented as confidence intervals.





z score (tenths)

Question: How much is the total area under bell curve?

Z score (hundredths)

	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	0.5160	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	0.7852
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	0.8810	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	0.9772	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	0.9864	0.9868	0.9871	0.9875	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	0.9925	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	0.9979	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
3.0	0.9987	0.9987	0.9987	0.9988	0.9988	0.9989	0.9989	0.9989	0.9990	0.9990

Normal Distribution to Calculate Probability

(manual table lookup without using statistical software like R)

Source https://www.statisticshowto.datasciencecentral.com/tables/z-table/

Source https://home.ubalt.edu/ntsbarsh/Business-stat/StatistialTables.pdf

Confidence Intervals - Formula

You estimate population parameters by using either **point estimate** or **interval estimate**

A **point estimate** is the value of a **single sample statistic** e.g. sample mean (average) or proportion (rate)

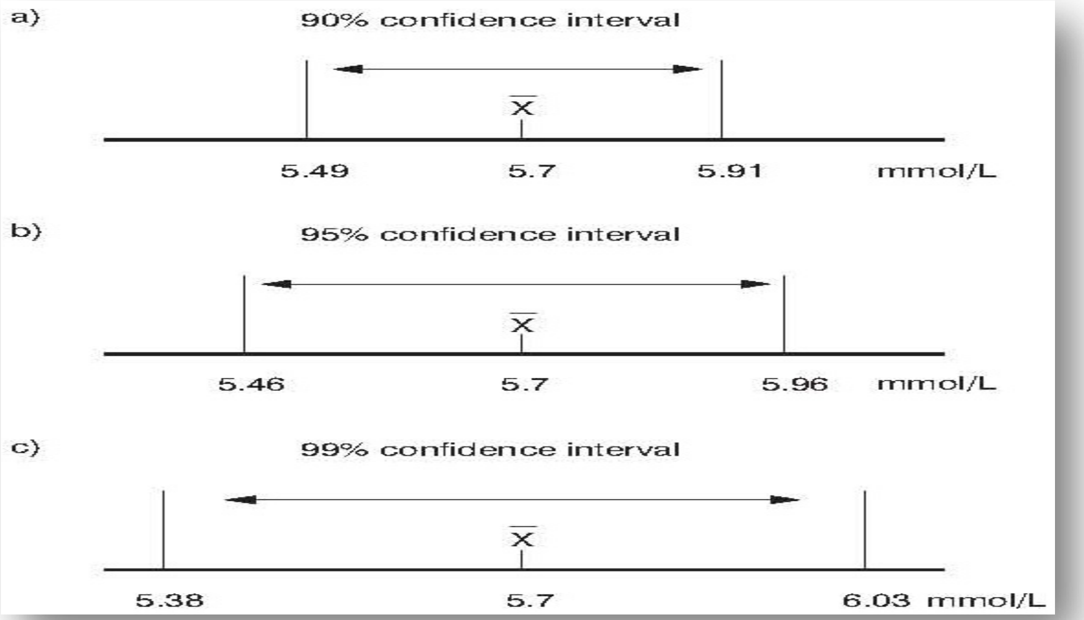
A confidence **interval estimate** is a range of numbers, called an **interval** constructed ***around*** the **point estimate**

For Mean		
with known variance & n>= 30	Use Z Distn Confidence Interval	$C.I. = \bar{X} \pm Z_{\frac{\alpha}{2}} \frac{\sigma}{\sqrt{n}}$
with unknown variance or n<30	Use t Dstn Confidence Interval	$C.I. = \bar{X} \pm t_{\frac{\alpha}{2}} \frac{S}{\sqrt{n}}$
For Variance	Use the χ^2 Distn Confidence Interval	$C.I. = \frac{(n-1)\,S^2}{\chi^2_{\alpha/2,n-1}} \leq \sigma^2 \leq \frac{(n-1)\,S^2}{\chi^2_{1-\alpha/2,n-1}}$
For Proportion	Use Z Distn Confidence Interval if $n \times \hat{p}$ and $n \times \hat{q}$ are each >= 5	$C.I. = \hat{p} \pm Z_{\alpha/2} \sqrt{\frac{\hat{p}\hat{q}}{n}}$ q=1-p

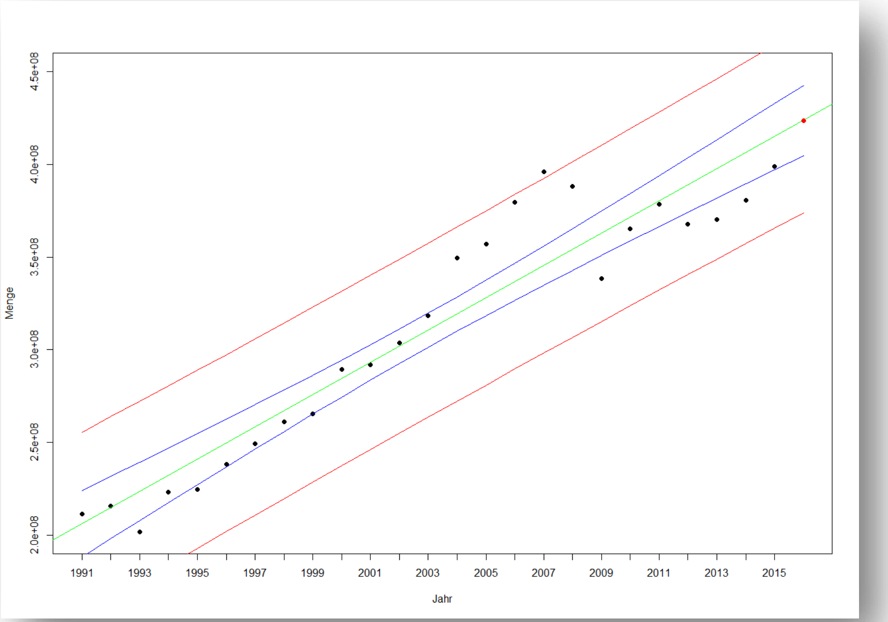
At (1-α) X 100% confidence level

Commonly used α value is 0.10 and 0.05 i.e. 90% and 95% level of confidence

Confidence intervals – significance level

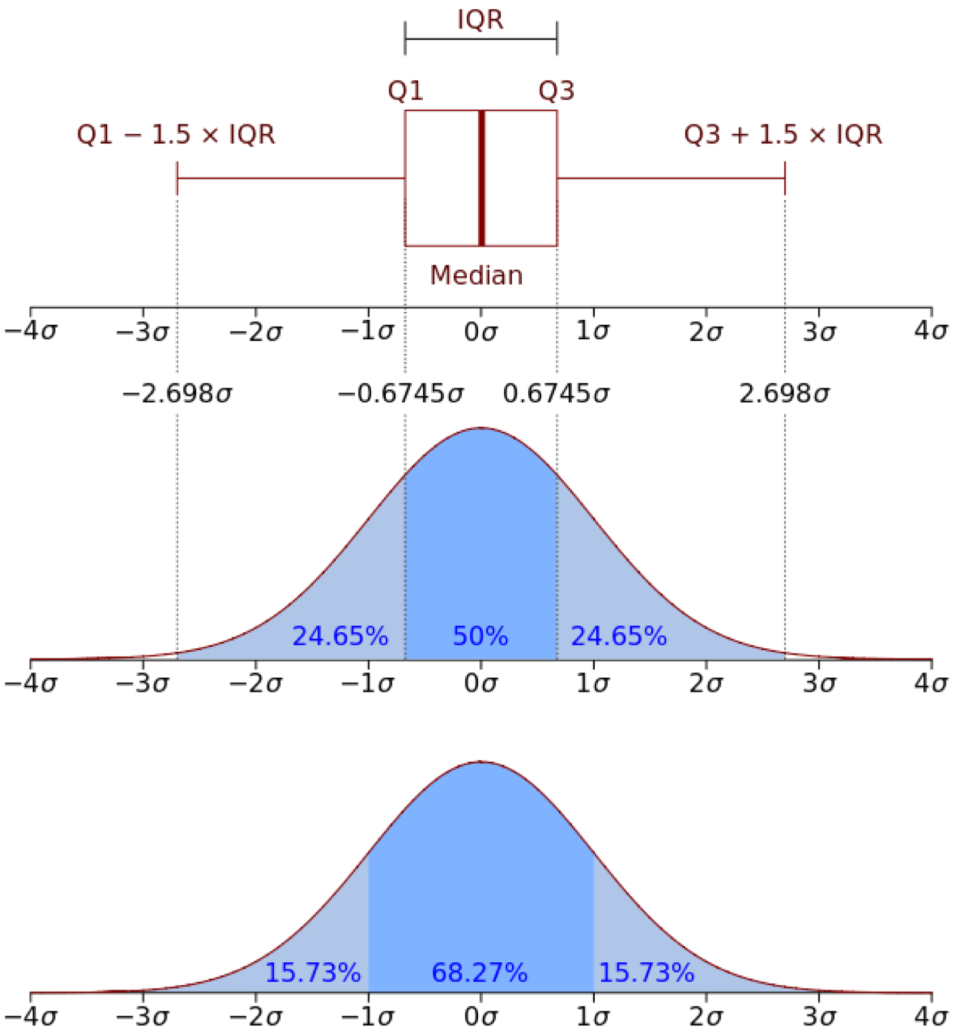


Estimation context



Prediction context

Box plot & Distribution



Source: https://en.wikipedia.org/wiki/File:Boxplot_vs_PDF.svg

Confidence interval

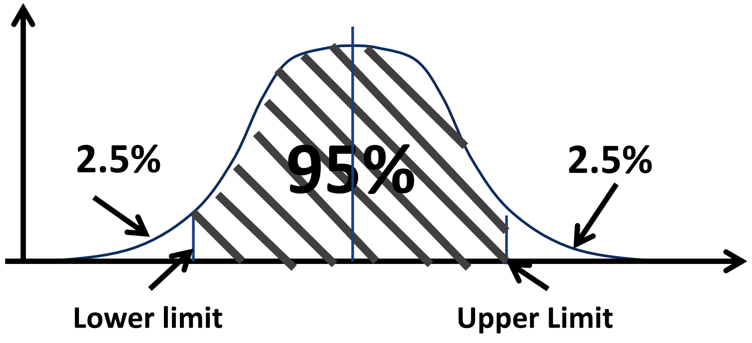
- The possible range/values for the sample mean
- 95% (1 - a=0.05) confidence interval: An interval we are 95% confident that the **actual population mean** will fall within (hopefully with this xx% chance)
- But, not to forget: 5% chance fall outside the range
- Confidence Interval:

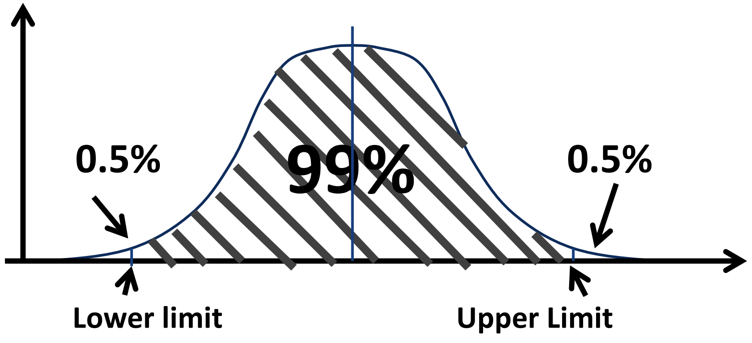
1 to 1 mapping

standard error (of sample mean): $SE_{\bar{x}} = \frac{s}{\sqrt{n}}$

Lower Limit: Mean – quantile * standard error

Upper Limit: Mean + quantile * standard error





Obtain Confidence Interval using R

Refer to 20 person’s age example in topic:

2.2 descriptive statistics & sampling

Confidence interval

1st sample of four numbers (person ages)

```
# create the vector
> s4 = c(21, 26, 28, 20)
> t.test(s4)


One Sample t-test

data:  s4
t = 12.299, df = 3, p-value = 0.001158
alternative hypothesis: true mean is not equal to 0
95 percent confidence interval:
 17.60436 29.89564
sample estimates:
mean of x
    23.75
```

Confidence interval

Point estimate

Confidence interval

2nd sample of eight numbers (person ages)

```
# create the vector
> s8 = c(23, 25, 29, 26, 25, 28, 33, 20)
> t.test(s8)


One Sample t-test

data:  s8
t = 18.736, df = 7, p-value = 3.064e-07
alternative hypothesis: true mean is not equal to 0
95 percent confidence interval:
 22.8279 29.4221
sample estimates:
mean of x
   26.125
```

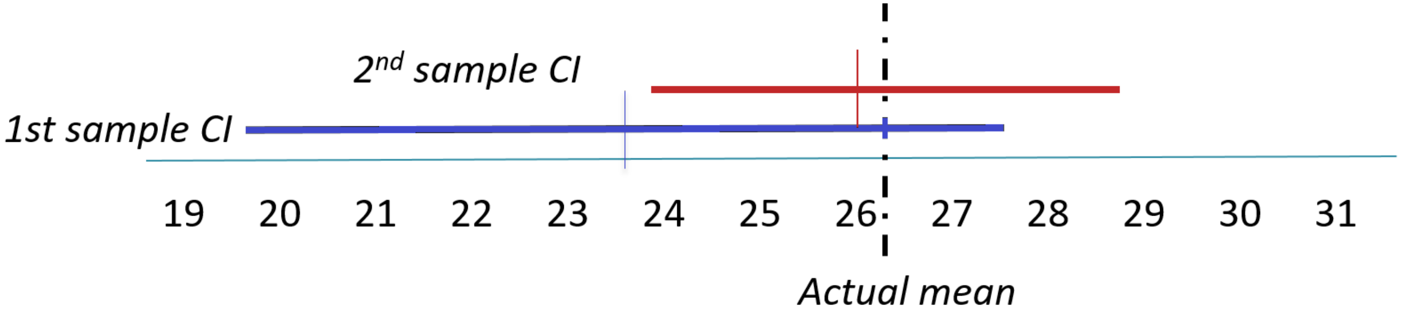
Confidence interval

Point estimate

Calculate confidence interval

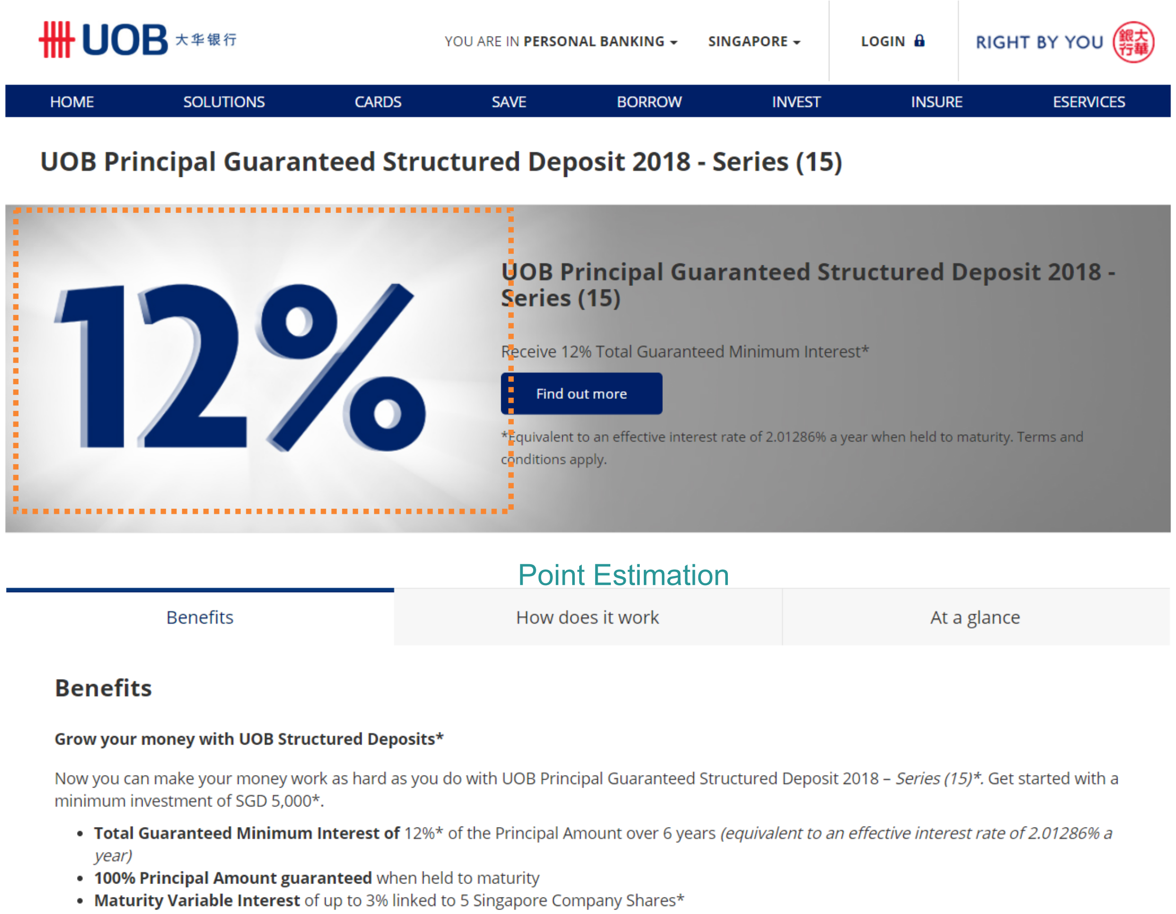
Continue previous examples

- For the **1st sample** of four person's age: 21,26,28,20
- For the **2nd sample** of eight person's age : 23,25,29,26,25,28,33,20
- The sample mean is **23.75**
- The sample mean is **26.125**
- The 95% confidence interval are **(17.6, 29.9)**
- The 95% confidence interval are **(22.8, 29.4)**



(average age is 26.35 based on 20 person population)

Point estimation isn’t neat and nice?
Why bother Confidence Interval?



Source: https://www.uob.com.sg/personal/invest/structured-deposit/current-plans-sgd.page

Benefits | How does it work | At a glance

At a glance

Investment Currency	Singapore Dollars (SGD)
Tenor	6 Years
Offer Period	3 October 2018 to 23 October 2018^
Guaranteed Fixed Interest	Payable at the end of: **Year** **% of the Principal Amount** 1 2.30% 2 2.30% 3 2.30% 4 2.30% 5 2.30% CI [0.5%, 3%] Total Guaranteed Fixed Interest = 11.50%
Maturity Variable Interest	Minimum at 0.5% of the Principal Amount and maximum at 3% of the Principal Amount*
Total Guaranteed Minimum Interest	12%* of the Principal Amount (Being Total Guaranteed Fixed Interest of 11.50% plus minimum Maturity Variable Interest of 0.5%)
Minimum Effective Interest Rate	2.01286% per annum (Based on total guaranteed minimum interest of 12%* of the Principal Amount.)
Shares in Underlying Basket	• CapitaLand Limited ("CAPL") • DBS Group Holdings Limited ("DBS") • Keppel Corporation Limited ("KEP") • Oversea-Chinese Banking Corporation Limited ("OCBC") • Singapore Telecommunications Limited ("ST")

^ Subject to changes, please refer to Indicative Term Sheet for full details.

* Note: Product terms and conditions apply. Please refer to the Indicative Term Sheet for more details. A copy of the Indicative Term Sheet is available at the point of sale or at any United Overseas Bank Limited ("UOB") branch.

All 'secrets' here

+ Only for existing UOB Personal Internet Banking (PIB) customers with a fixed deposit account. Online purchase is subject to passing the Customer Knowledge Assessment (CKA) on UOB PIB.

Calculating Confidence Limits for a Population Mean (σ known)

The manager of a super market wants to estimate the amount of rice contained in 5-kg(μ) bags imported from a regionally known rice exporter. The exporter's specifications state that the standard deviation (σ known) of the amount of rice is equal to 0.03 kg. For a random sample of 40 bags, the sample mean amount of rice per 5 kg bag is 4.995kg.

- On the basis of these results, do you think the manager has a right to complain to the exporter? Why?

We can calculate a confidence interval assuming the sample mean follows the **normal distribution**

Specifically , if we want a probability of 95% of getting a value inside the range, then the confidence interval will be $(4.995 - 1.96 * 0.03/\sqrt{40}\ ,\ 4.995 + 1.96 * 0.03/\sqrt{40}) = (4.623, 5.367)$

i.e. *Probability* $(\boldsymbol{\mu} - 1.96 * \sigma / \sqrt{n} \le \textbf{Value} \le \boldsymbol{\mu} + 1.96 * \sigma / \sqrt{n}) = 95\%$

For a 90% confidence limit the confidence interval replace 1.96 with 1.64.

Calculating Confidence Limits for a Population Mean (σ unknown) (n < 30)

Example: For the class of 20 people with ages : (assume we don't know σ)

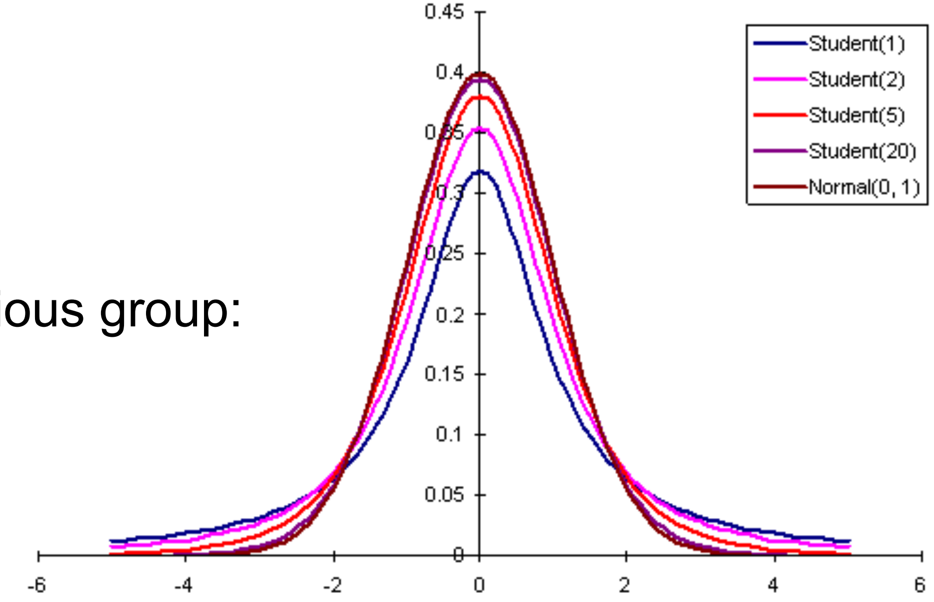
- 21, 23, 24, 19, 25. 35, 29, 31, 34, 26, 23, 25, 29, 27, 28, 23, 33, 31, 21, 20

We took a sample of four numbers from the above group: 21, 26, 28, 20

- The mean of these was 23.75
- The estimate of the standard error is 1.93 ($SE_{\bar{x}} = \frac{s}{\sqrt{n}}$)
- The 95% t-dist confidence limits are (17.60, 29.90) $\bar{x} \pm t_{\alpha/2,n-1} \frac{s}{\sqrt{n}}$
- $t_{\alpha/2,n-1}$ -> Student's t with n-1 degrees of freedom

Similarly we took a sample of eight numbers from the previous group: 23, 25, 29, 26, 25, 28, 33, 20

- The mean of these was 26.13
- The estimate of the standard error is 1.39.
- The 95% t-dist confidence limits are (22.83, 29.42)
- If n>30, use normal distribution Z.



- *Note that these are estimates of the confidence limits, as σ is estimated by **s**.*

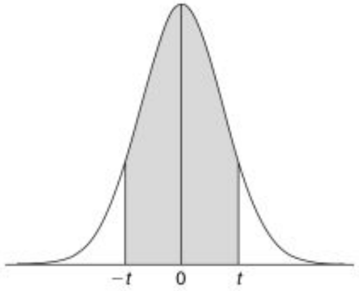
Calculating Confidence Limits for a Population Mean (σ unknown) (n < 30)

Student's *t* Table

Gives the value of *t* for a given ν and P % confidence .

TWO-SIDED STUDENT'S *t* VARIABLE VALUES

ν	$t_{\nu,P=50\%}$	$t_{\nu,P=90\%}$	$t_{\nu,P=95\%}$	$t_{\nu,P=99\%}$
1	1.000	6.341	12.706	63.657
2	0.816	2.920	4.303	9.925
3	0.765	2.353	3.192	5.841
4	0.741	2.132	2.770	4.604
5	0.727	2.015	2.571	4.032
6	0.718	1.943	2.447	3.707
7	0.711	1.895	2.365	3.499
8	0.706	1.860	2.306	3.355
9	0.703	1.833	2.262	3.250
10	0.700	1.812	2.228	3.169
11	0.697	1.796	2.201	3.106
12	0.695	1.782	2.179	3.055
13	0.694	1.771	2.160	3.012
14	0.692	1.761	2.145	2.977
15	0.691	1.753	2.131	2.947
16	0.690	1.746	2.120	2.921
17	0.689	1.740	2.110	2.898
18	0.688	1.734	2.101	2.878
19	0.688	1.729	2.093	2.861
20	0.687	1.725	2.086	2.845
21	0.686	1.721	2.080	2.831
30	0.683	1.697	2.042	2.750
40	0.681	1.684	2.021	2.704
50	0.680	1.679	2.010	2.679
60	0.679	1.671	2.000	2.660
120	0.677	1.658	1.980	2.617
∞	0.674	1.645	1.960	2.576



What is *t* for n = 12 ?

$\boldsymbol{\nu}$ = n - 1

Degree of freedom

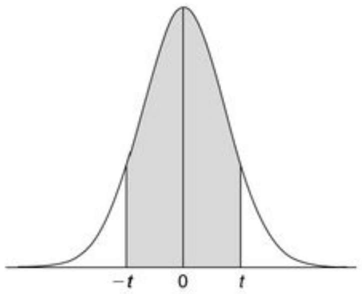
Calculating Confidence Limits for a Population Mean (σ unknown) (n < 30)

Student's t Table

Gives the value of t for a given ν and P % confidence .

TWO-SIDED STUDENT'S t VARIABLE VALUES

ν	$t_{\nu,P=50\%}$	$t_{\nu,P=90\%}$	$t_{\nu,P=95\%}$	$t_{\nu,P=99\%}$
1	1.000	6.341	12.706	63.657
2	0.816	2.920	4.303	9.925
3	0.765	2.353	3.192	5.841
4	0.741	2.132	2.770	4.604
5	0.727	2.015	2.571	4.032
6	0.718	1.943	2.447	3.707
7	0.711	1.895	2.365	3.499
8	0.706	1.860	2.306	3.355
9	0.703	1.833	2.262	3.250
10	0.700	1.812	2.228	3.169
11	0.697	1.796	2.201	3.106
12	0.695	1.782	2.179	3.055
13	0.694	1.771	2.160	3.012
14	0.692	1.761	2.145	2.977
15	0.691	1.753	2.131	2.947
16	0.690	1.746	2.120	2.921
17	0.689	1.740	2.110	2.898
18	0.688	1.734	2.101	2.878
19	0.688	1.729	2.093	2.861
20	0.687	1.725	2.086	2.845
21	0.686	1.721	2.080	2.831
30	0.683	1.697	2.042	2.750
40	0.681	1.684	2.021	2.704
50	0.680	1.679	2.010	2.679
60	0.679	1.671	2.000	2.660
120	0.677	1.658	1.980	2.617
∞	0.674	1.645	1.960	2.576



Examine P = 95 %

What is t for n = 12 ?

N = 12

>> ν = N – 1 = 11

$t_{11,95} = 2.201$

Degrees of Freedom

While mentioning t distribution we mention degrees of freedom

In statistics, the number of **degrees of freedom** is the number of values in the final calculation of a statistic that are free to vary.

Estimates of statistical parameters can be based upon different amounts of information or data. The number of independent pieces of information that go into the estimate of a parameter are called the degrees of freedom.

Examples : Suppose you have n observations y_i . And you are calculating standard deviation s which is given by following formula:

$$\boldsymbol{Average} = \bar{y}$$

$$\boldsymbol{Sample\ Variance} = s^2 = \frac{\sum_{i=1}^{n}(y_i - \bar{y})2}{n-1}$$

Sample Standard deviation = *s* = $\sqrt[2]{variance}$

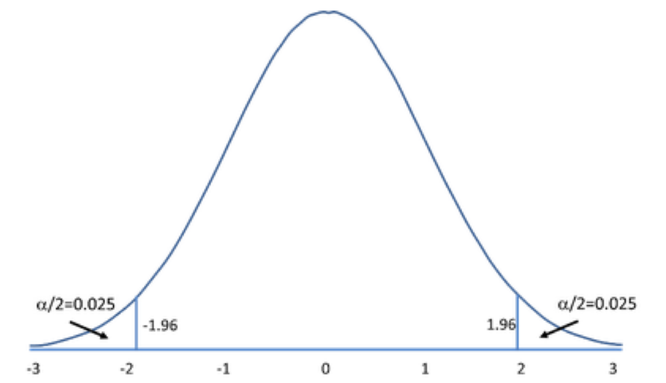
In the above calculation if one knows the average and y_1 , y_2, ….y_{n-1} then y_n is fixed $\{\bar{y}$ - (y_1 + y_2+ ….+y_{n-1})} i.e. we lose 1 degree of freedom . But y_1 , y_2, ….y_{n-1} can take any value i.e. in standard deviation calculation we have n – 1 degrees of freedom.

Determining Sample Size for the Mean

To compute the sample size n you must know three quantities :

- **the desired confidence level, which determines the value of $Z_{\alpha/2}$, the critical value from the standardized normal distribution**
- **The acceptable sampling error, E**
- **The standard deviation σ**

$$n = \left(\frac{(Z_{\alpha/2})(\sigma)}{E}\right)^2$$



Rejection Region for Two-Tailed Z Test (H_1: $\mu \neq \mu_0$) with $\alpha = 0.05$

The decision rule is: Reject H_0 if $Z \leq -1.960$ or if $Z \geq 1.960$.

Two-Tailed Test	
α	Z
0.20	1.282
0.10	1.645
0.05	1.960
0.010	2.576
0.001	3.291
0.0001	3.819

A consumer group wants to estimate mean electric bill for the month of July for single-family homes in a large city

Based on the studies conducted in other cities, the standard deviation is assumed to be $25. The group wants to estimate, with 95% confidence, the mean bill for July to be within ± $5

- What Sample size is needed?
- If 99% is desired, how many homes need to be selected?

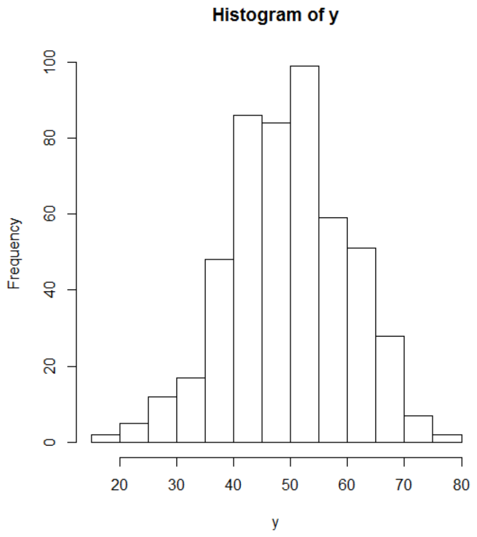
Evaluating normality

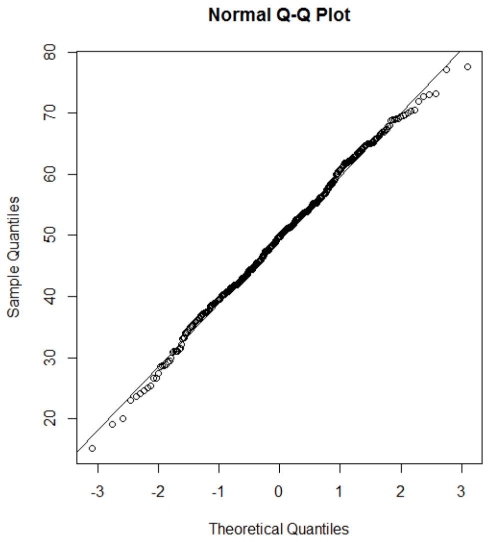
- Normality can be checked using a histogram or quantile-quantile plot (QQ plot)

```
# Generate 500 numbers from a normal distribution
with a mean of 50 and a standard deviation of 10
> y = rnorm(500, 50, 10)

# Create QQ plot
> qqnorm(y)
> qqline(y)

# histogram
> hist(y)
```

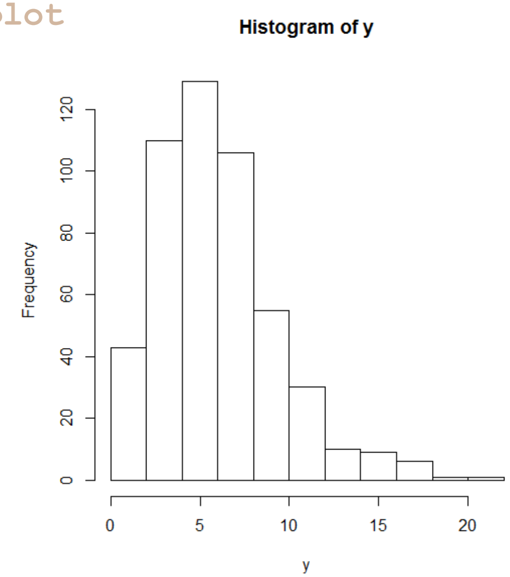


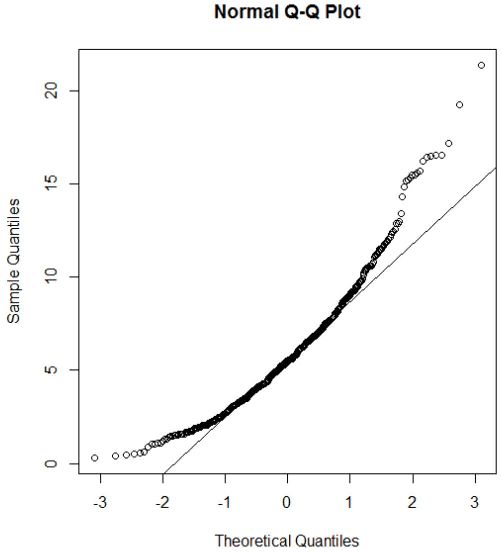


Evaluating normality

- Check the normality of another distribution

```
# Generate 500 numbers from not normal distribution
distribution
> z = rchisq(500, df=6)
# histogram
> hist(z)
# Create QQ plot
> qqnorm(z)
> qqline(z)
```





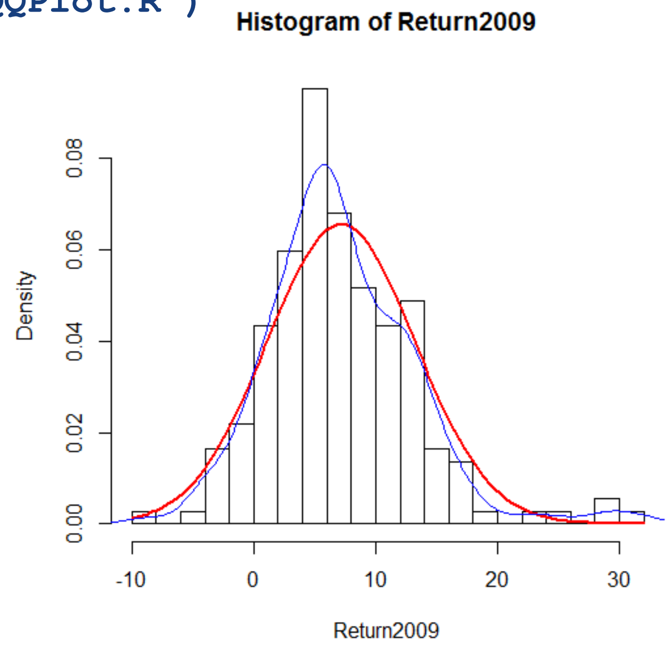
Evaluating normality

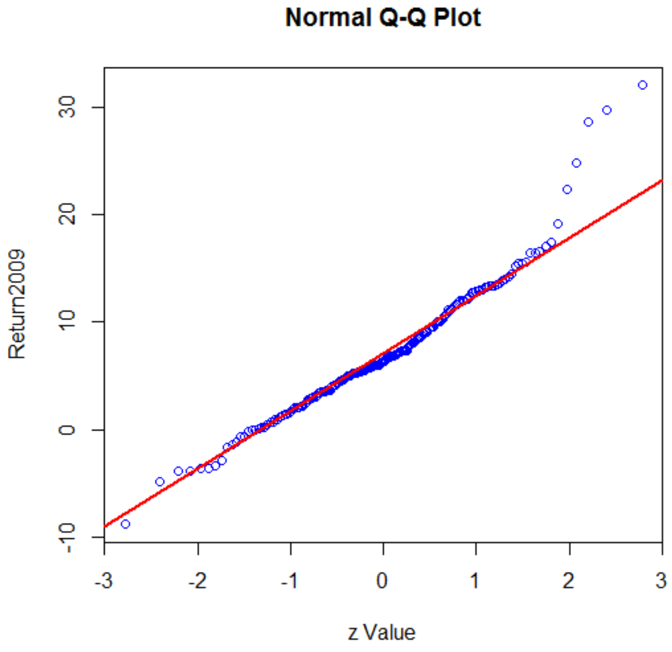
Example

- Check if the returns of 184 mutual funds in year 2009 followed a normal distribution

```
> source("BondFunds_Hist.R")
```

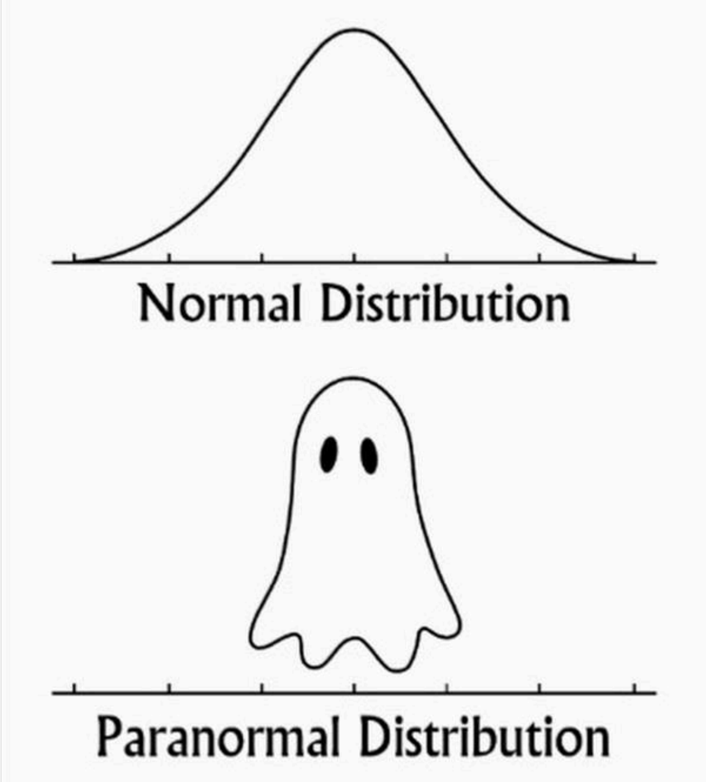
```
> source("BondFunds_QQPlot.R")
```





Summary of useful R functions

Function	Description
`sample`	Takes a sample of the specified size from the elements of x
`qqnorm()`	Produce a normal QQ plot
`qqline()`	Add a straight line to the QQ plot
`hist()`	Produce a histogram



Source: https://alvinalexander.com/photos/paranormal-distribution-stats-geeks-funny

End of Lecture Notes

Appendix

Examples IS Group Projects for 1st foundational grad-cert: Intelligent Reasoning Systems
One project for each modular course:

https://github.com/IRS-RS
https://github.com/IRS-MR
https://github.com/IRS-CGS