STAT 2011
Workshop on Data Exploration and Technical Writing

Section 1: Introduction to Excel VBA Programming

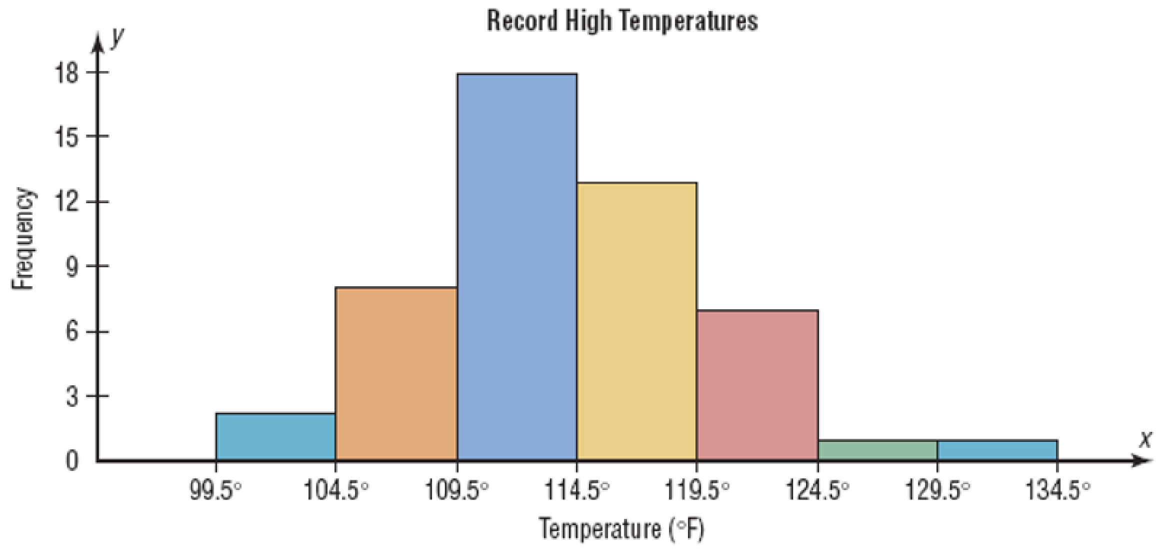
Dr. OUYANG Ming

2023/2024 Term 2

Lecture Notes 2
Frequency Distributions, descriptive statistics and Graphs

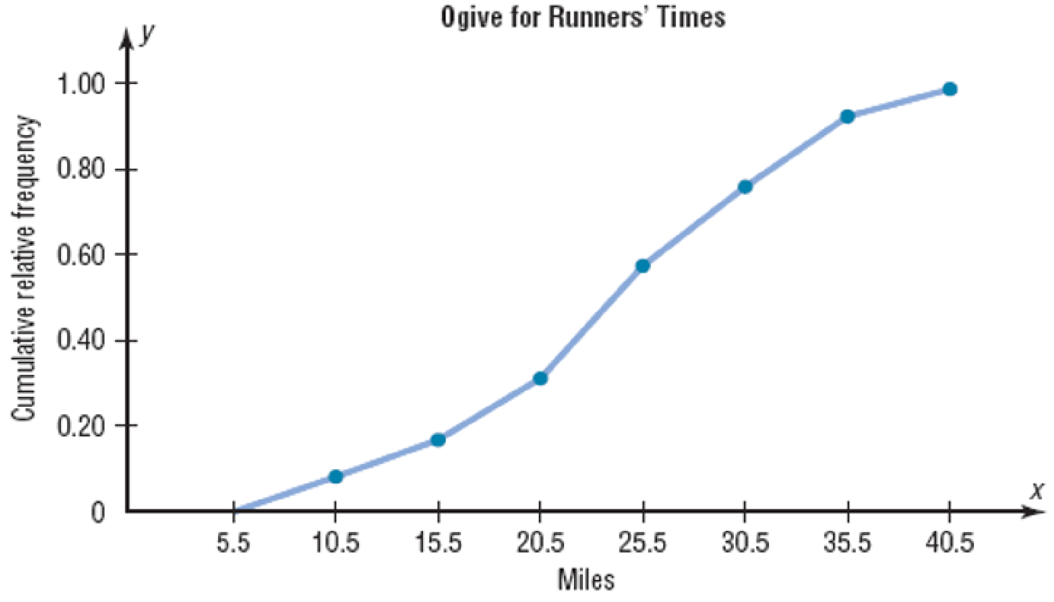
Histograms

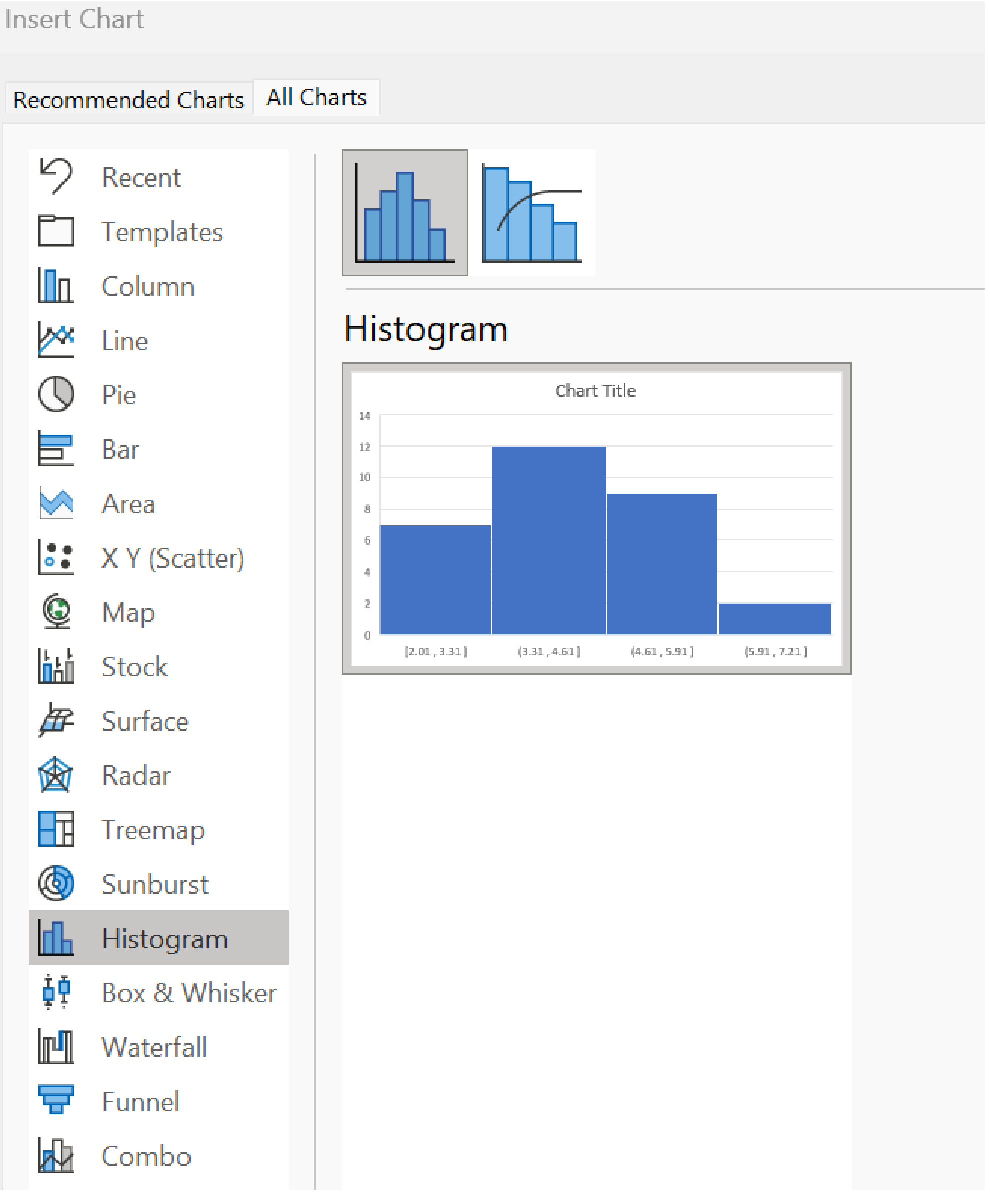
Histograms use class boundaries and frequencies of the classes.

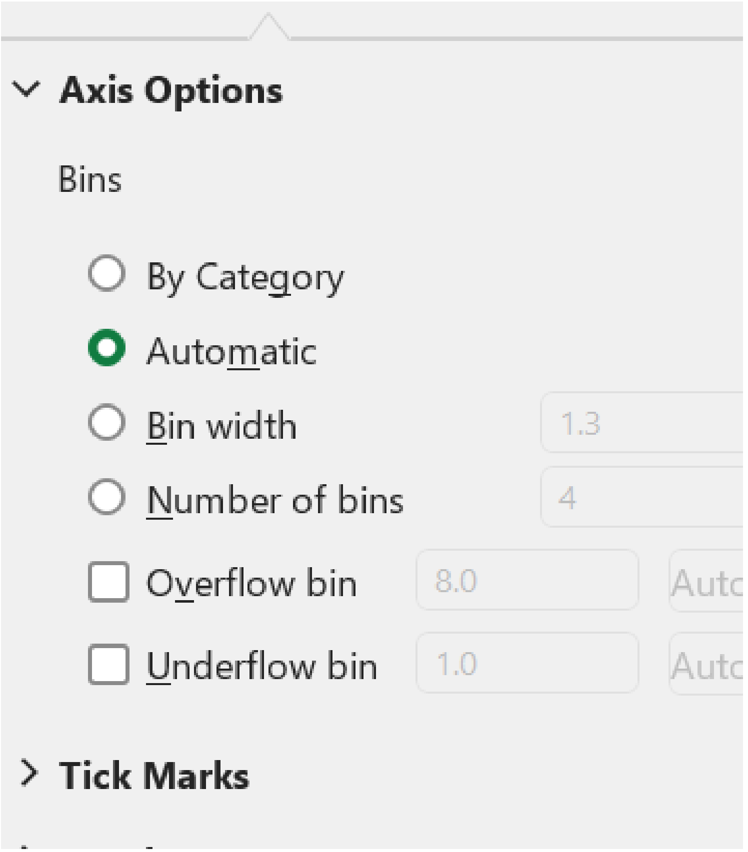


Ogives

Use the upper class boundaries and the cumulative relative frequencies.




Insert Chart
Recommended Charts
All Charts
Recent
Templates
Column
Line
Pie
Bar
Area
X Y (Scatter)
Map
Stock
Surface
Radar
Treemap
Sunburst
Histogram
Box & Whisker
Waterfall
Funnel
Combo
Histogram
Chart Title
14
12
10
8
6
4
2
0
[2.01, 3.31]
(3.31, 4.61]
(4.61, 5.91]
(5.91, 7.21]


Axis Options
Bins
By Category
Automatic
Bin width
1.3
Number of bins
4
Overflow bin
8.0
Underflow bin
1.0
Tick Marks

- **Frequency Table:**

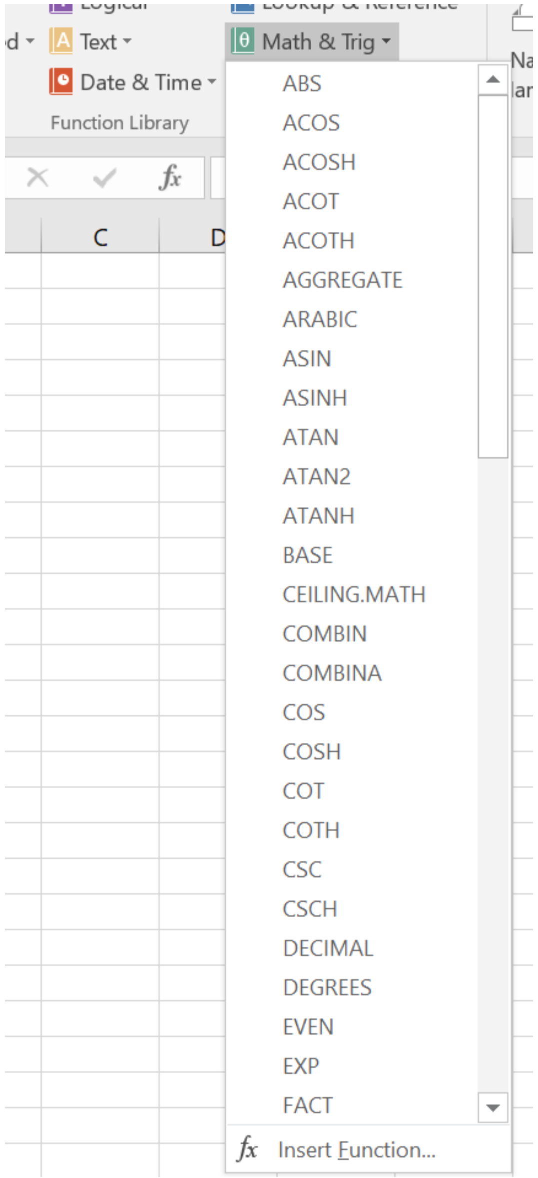
FREQUENCY

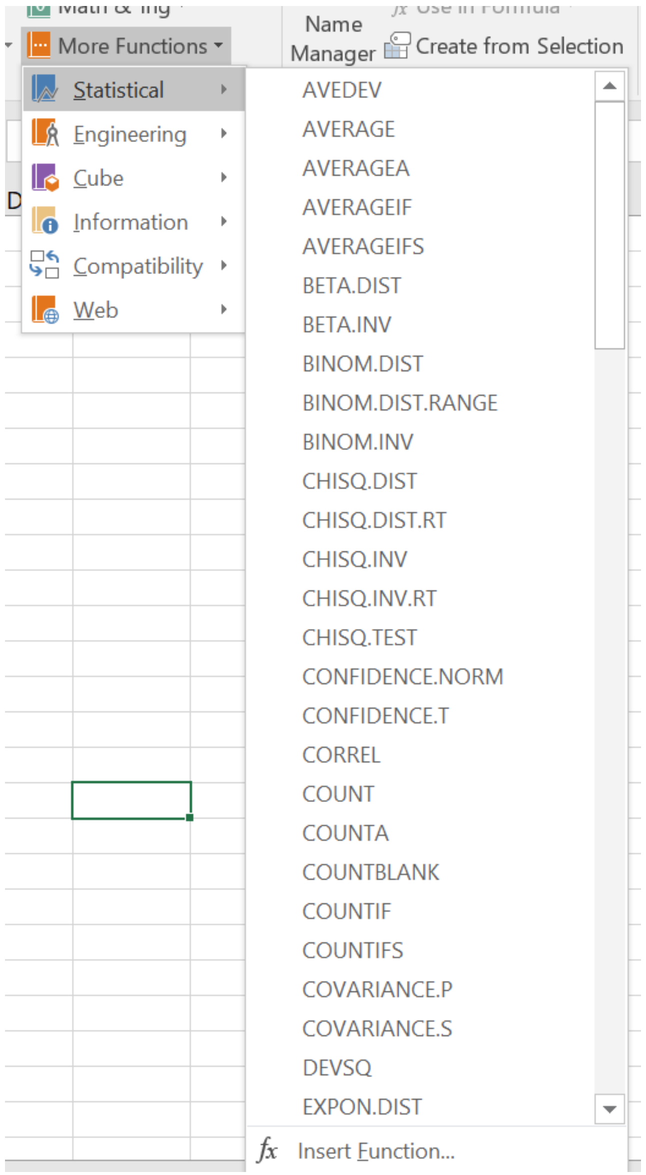
Students	Systolic blood pressure	ABO Blood Type
1	93.73	A
2	94.33	A
3	94.73	B
4	93.36	AB
5	92.63	B
6	90.87	O
7	94.35	O
8	94.15	A
9	92.72	AB
10	92.58	B
11	90.76	A
12	91.59	O
13	93.74	O
14	93.23	O
15	90.97	AB

Systolic Blood pressure	Frequency	Proportion
<92	4	26.67%
92 to 93	3	20%
93 to 94	4	26.67%
>94	4	26.66%
TOTAL	15	100%

Need bins to group ranges of data.

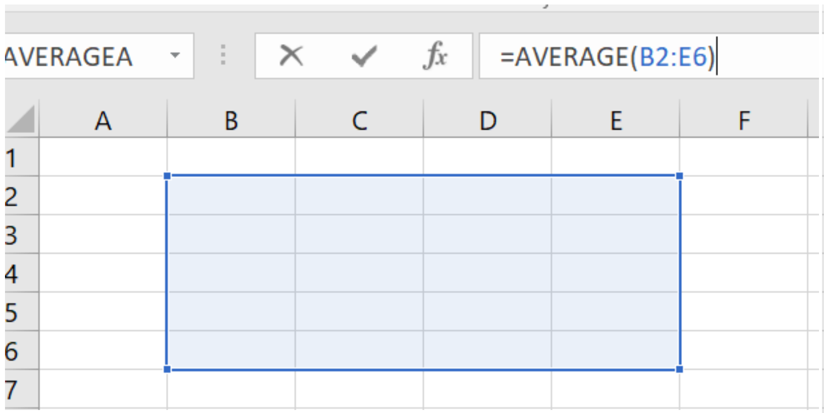
Statistical Formulas

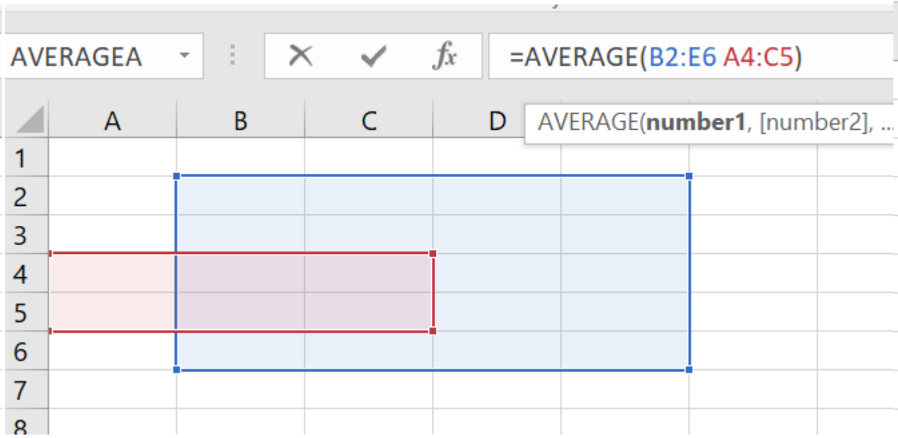


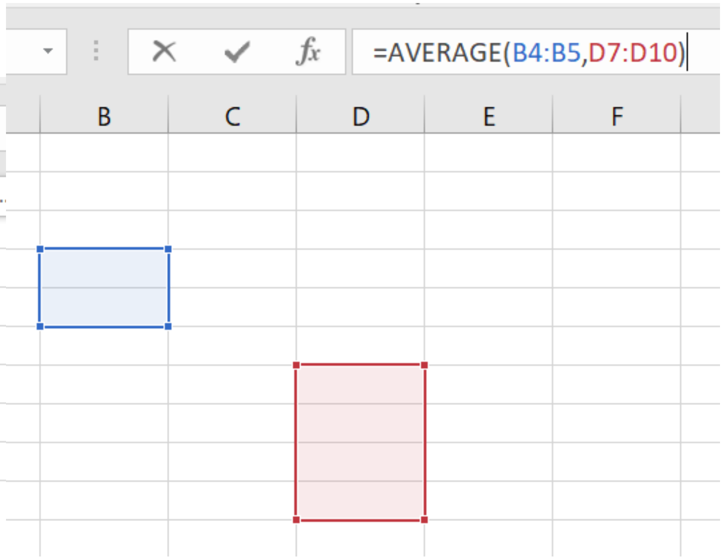


- **Mean (or Sample mean) :** Average value of the n observations:

$$\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i = \frac{1}{n}(x_1 + x_2 + \cdots + x_n)$$

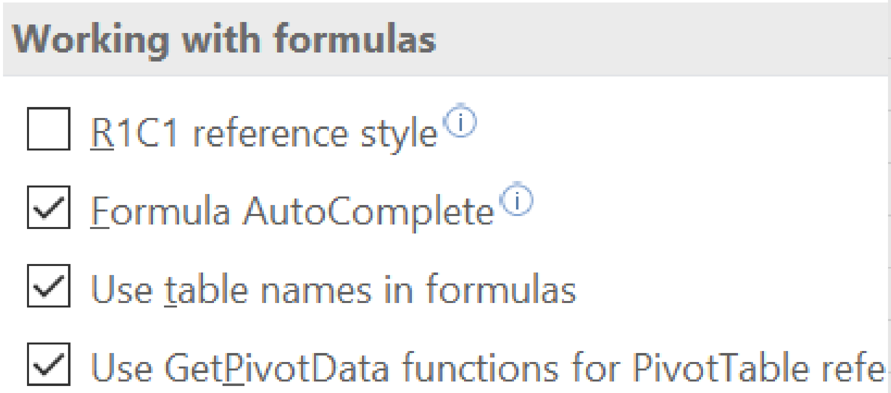


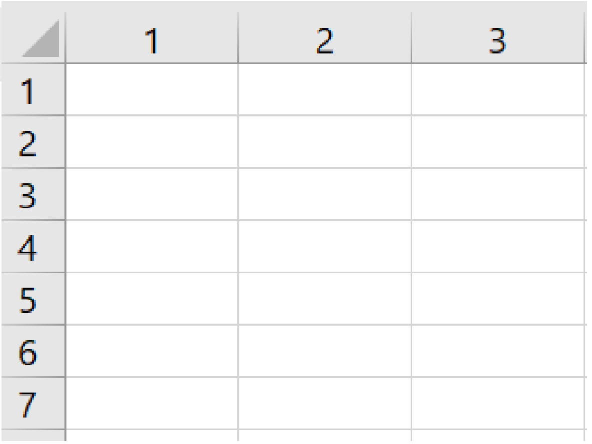


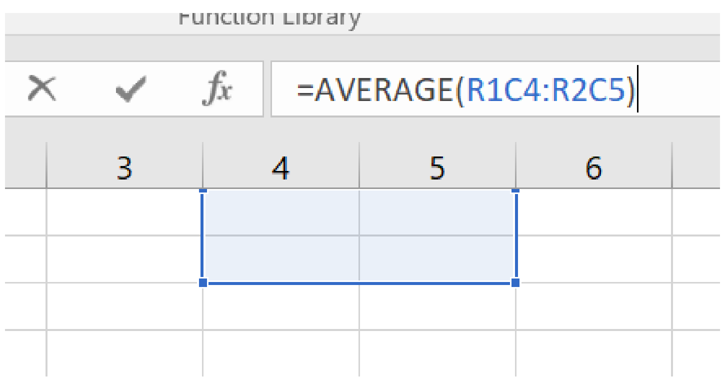


- **Mean (or Sample mean) :** Average value of the n observations:

$$\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i = \frac{1}{n}(x_1 + x_2 + \cdots + x_n)$$

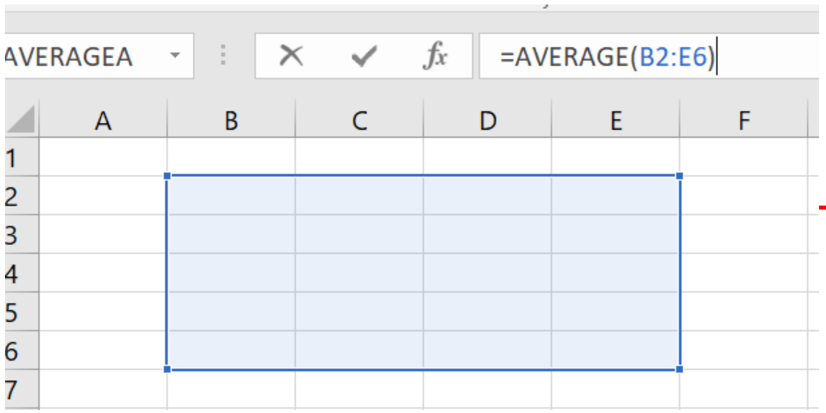




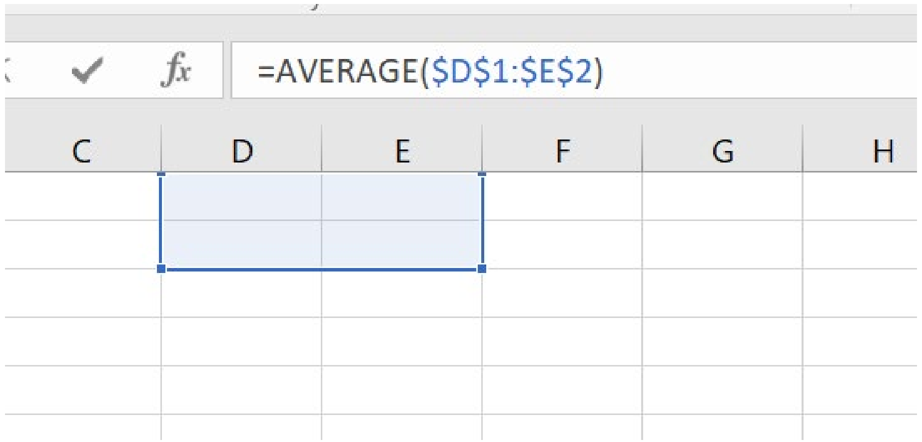


- **Mean (or Sample mean) :** Average value of the n observations:

$$\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i = \frac{1}{n}(x_1 + x_2 + \cdots + x_n)$$



F4



- **Mean (or Sample mean) :** Average value of the n observations:

TEXT

NAME	AMY	ID	A00202
SEX	MALE	Telephone	2851327

Number

Integer	1,2,3,...
Percentage	0.85%,1%,...

Logic value

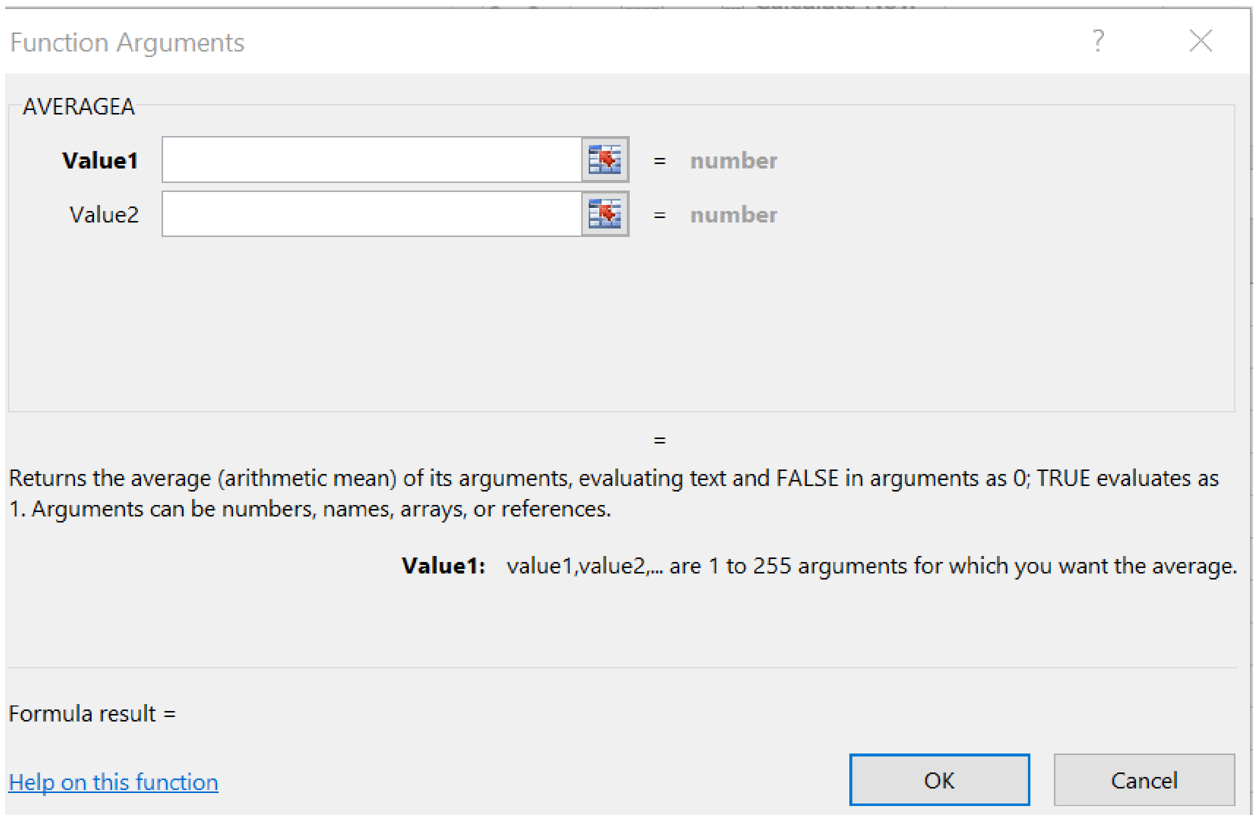
TRUE	0
FALSE	1

Date/Time

Birth date	2000/10
Time	8:00 AM

&

……



- **Arguments**

$$\cdots\cdots\ -2\ \ -1\ \ 0\ \ 1\ \ 2\ \cdots\cdots\ A\ \ B\ \cdots\cdots\ Z\cdots\ \ FALSE\ \ TRUE$$

1900/01/01

Dates are stored as the number of days that have elapsed since 1900/01/01

One hour: 1/24 = 0.04167
One minute: 1/(24*60) = 1/1440 = 0.00069

Operator	Example	Result	Example	Result
=	=5=3	FALSE	=8=2+6	TRUE
<>	=5<>3	TRUE	=5<>5	FALSE
>	=-5>2	FALSE	=8>1	TRUE
<	=8<2	FALSE	=0<7	TRUE
> =	=9>=7	TRUE	=5>=5	TRUE
< =	=5<=2	FALSE	=6<=6	TRUE

TEXT

NAME	AMY	ID	A00202
SEX	MALE	Telephone	2851327

Number

Integer	1,2,3,...
Percentage	0.85%,1%,...

Logic value

TRUE	0
FALSE	1

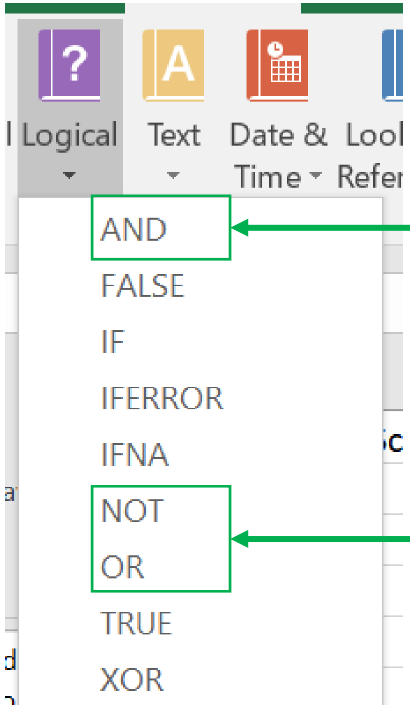
Date/Time

Birth date	2000/10
Time	8:00 AM

&

......

- **IF**



=IF(

IF(**logical_test**, [value_if_true], [value_if_false])

=IF(B2>=60,"PASS","FAIL")

$B2 \geq 60$

TRUE PASS

FALSE FAIL

	A	B	C
1	School	Score	
2	Band1	95	FAIL

=IF(60<=B2<=100,"PASS","FAIL")

$60 <= 95 <= 100 \Rightarrow TRUE <= 100 \Rightarrow FALSE$

=IF(FALSE,"PASS","FAIL")

=IF(AND(60<=B2,B2<=100),"PASS","FAIL")

- **Mean (or Sample mean) :** Average value of the n observations:

AVEDEV
AVERAGE
AVERAGEA
AVERAGEIF
AVERAGEIFS

=SUMIF(A2:A11,"????",B2:B11)

=COUNTIF(A1:A10,"*"&B1&"*")

=COUNTIF(A1:A10,"")

=COUNTIF(A1:A10,"<>")

=AVERAGEIF(

AVERAGEIF(**range**, criteria, [average_range])

?

=AVERAGEIF(A2:A5,"?????1",B2:B5)

	A	B	C
1	Year	Score	
2	Year1	11	12
3	Year2	1	
4	Year3	23	
5	Year11	12	

*

=AVERAGEIF(A2:A5,"*1",B2:B5)

	A	B	C
1	Year	Score	
2	Year1	11	11.5
3	Year2	1	
4	Year3	23	
5	Year11	12	

• Count

COUNT
COUNTA
COUNTBLANK
COUNTIF
COUNTIFS

=COUNTIFS(A2:A12,"=100",B2:B12,"=NUMBER")

	A	B	C	D
1	數據	數據説明		
2	10000	TEXT (number)	4	COUNT
3		TRUE EMPTY	8	COUNTA
4		FAKE EMPLTY(="")	5	COUNTBLANK
5	excel	TEXT	3	COUNTIF(A2:A12,"*")
6	100	NUMBER	8	COUNTIF(A2:A12,"<>")
7		TRUE EMPTY	1	
8		TRUE EMPTY		
9	0.5	NUMBER		
10		FAKE EMPLTY(="")		
11	20-Feb	DATE		
12	#DIV/0!	FALSE		

- **Mean (or Sample mean) :** Average value of the n observations:

AVEDEV
AVERAGE
AVERAGEA
AVERAGEIF
AVERAGEIFS

=AVERAGEIFS(
AVERAGEIFS(**average_range**, criteria_range1, criteria1, ...)

SUM | =AVERAGEIFS(D2:D11,$A$2:$A$11,$K$3,$C$2:$C$11,$L$3)

	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q
1	School	SID	Class	Chinese	Math	English	Physics	Total			Summary MEAN						
2	Band1	123	A	76	20	60	60	216			School	Class	Chinese	Math	English	Physics	Total
3	Band1	421	B	95	78	31	52	256			Band1	A	$L$3)	45	79.5	60.5	249
4	Band1	252	C	52	0	54	100	206									
5	Band1	213	B	77	68	75	66	286									
6	Band1	577	A	52	70	99	61	282									
7	Band2	568	A	65	44	16	86	211									
8	Band2	234	B	23	83	88	23	217									
9	Band2	164	B	49	46	50	33	178									
10	Band2	265	A	86	30	60	52	228									
11	Band2	123	C	97	71	41	35	244									
12																	

- **Mean (or Sample mean) :** Average value of the n observations:

AVEDEV
AVERAGE
AVERAGEA
AVERAGEIF
AVERAGEIFS

$$\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i = \frac{1}{n}(x_1 + x_2 + \cdots + x_n)$$

Returns the average of the absolute deviations of data points from their mean. Arguments can be numbers or names, arrays, or references that contain numbers.

$$d = \frac{\sum_{i=1}^{n} |x_i - \bar{x}|}{n}$$

Mean absolute deviation

- **Special MEAN**

GEOMEAN
HARMEAN
TRIMMEAN

Geometric Mean: Used to measure the rate of change of a variable over time

$$\bar{x}_G = (x_1 \times x_2 \times \cdots \times x_n)^{1/n}$$

Geometric Mean rate of return: Measures the status of an investment over time

Where r_i is the rate of return in time period i

$$\bar{r}_G = ((1 + r_1) \times (1 + r_2) \times \cdots \times (1 + r_n))^{1/n} - 1$$

Harmonic mean: it is appropriate for situations when the average rate is desired

$$\bar{x}_H = \frac{n}{\frac{1}{x_1} + \frac{1}{x_2} + \cdots + \frac{1}{x_n}} = \left(\frac{\sum_{i=1}^{n} x_i^{-1}}{n}\right)^{-1}$$

- **Mode**

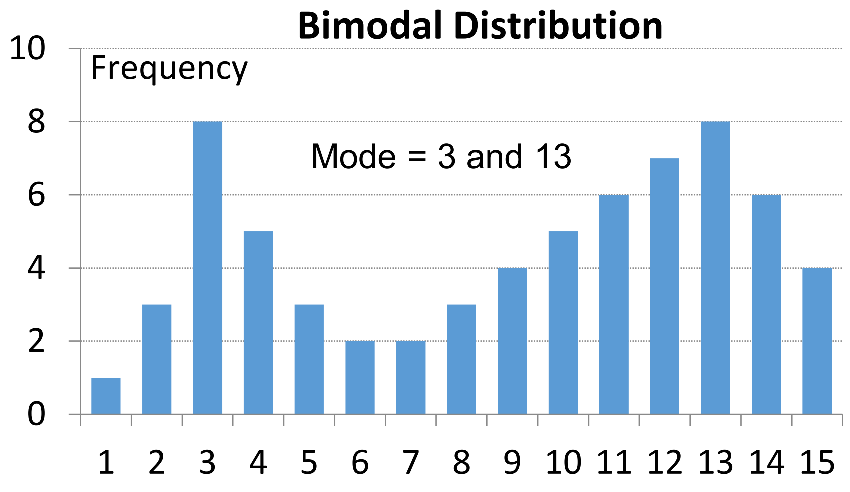
MODE.MULT

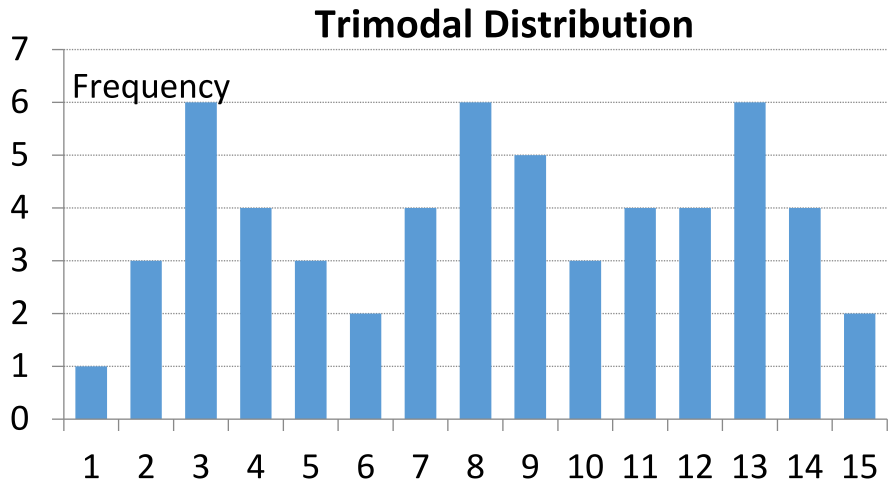
MODE.SNGL

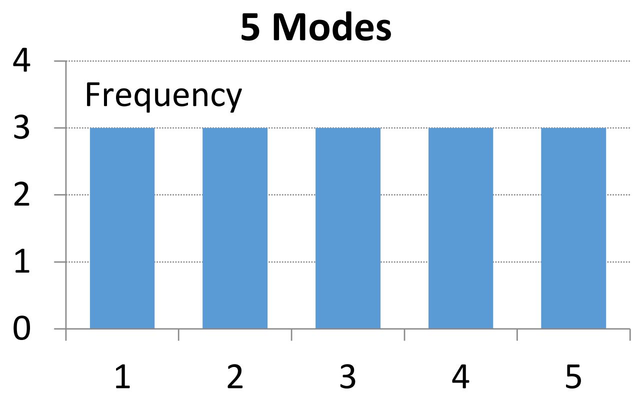
- **Advantage:** Easy to compute and interpret (most frequently)
- **Disadvantage:**

(1) Non-unique centering if more than one mode (Unimodal, Bimodal, Trimodal…)

(2) Sometimes unable to describe the majority of the data points







- **DEVSQ / VARIANCE / Standard Deviation**

DEVSQ

VAR.P	STDEV.P
VAR.S	STDEV.S
VARA	STDEVA
VARPA	STDEVPA

Arguments can be the following: numbers; names, arrays, or references that contain numbers; text representations of numbers; or logical values, such as TRUE and FALSE.

$$DEVSQ = \sum_{i=1}^{n} (x_i - \bar{x})^2$$

Sample variance

$$s^2 = \frac{1}{n-1} \sum_{i=1}^{n} (x_i - \bar{x})^2$$

Population variance

$$\sigma^2 = \frac{1}{n} \sum_{i=1}^{n} (x_i - \mu)^2$$

- **Percentile/Quartile**

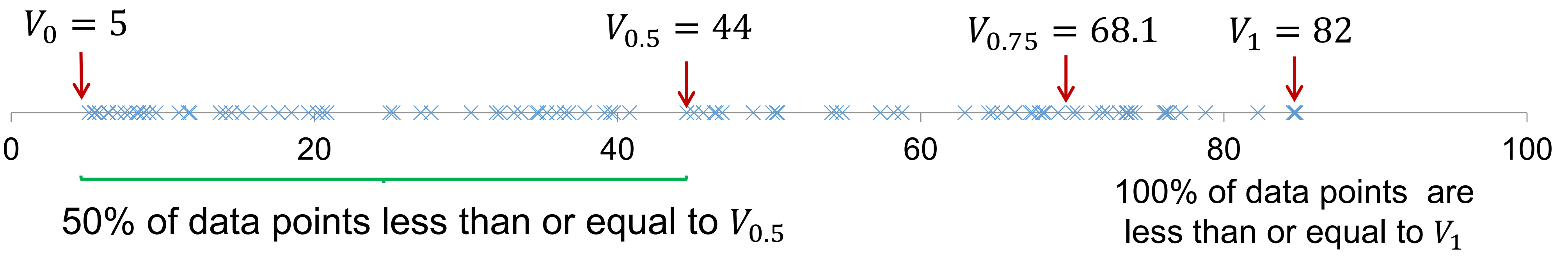
Returns the k^{th} smallest/largest value in a data set

Definition: The p^{th} percentile of a data set ($0 \leq p \leq 100$), denoted by $V_{p/100}$, is a value such that p percent of the data are less than or equal to $V_{p/100}$.

SMALL
LARGE
MAX
MAXA
MEDIAN
MIN
MINA
PERCENTILE.EXC
PERCENTILE.INC
QUARTILE.EXC
QUARTILE.INC

Percentiles for Median, Minimum and Maximum ($p = 50,100,0$):

- 50^{th}Percentile ($V_{0.5}$): Median, 50% of the data points are smaller than or equal to $V_{0.5}$.
- 100^{th} Percentile (V_1): Maximum = largest observation
- 0^{th}Percentile (V_0): Minimum = smallest observation



- **Percentile/Quartile**

Definition: The p^{th} percentile of a data set ($0 \le p \le 100$), denoted by $V_{p/100}$, is a value such that p percent of the data are less than or equal to $V_{p/100}$.

If more than one value has the same rank, the **average** rank is returned.

RANK.AVG

RANK.EQ

Returns the rank of a number in a list of numbers: its size relative to other values in the list.

If more than one value has the same rank, the **top** rank is returned.

PERCENTRANK.EXC

PERCENTRANK.INC

Returns the rank of a value in a data set as a percentage (0-1, *exclusive/inclusive*) of the data set.

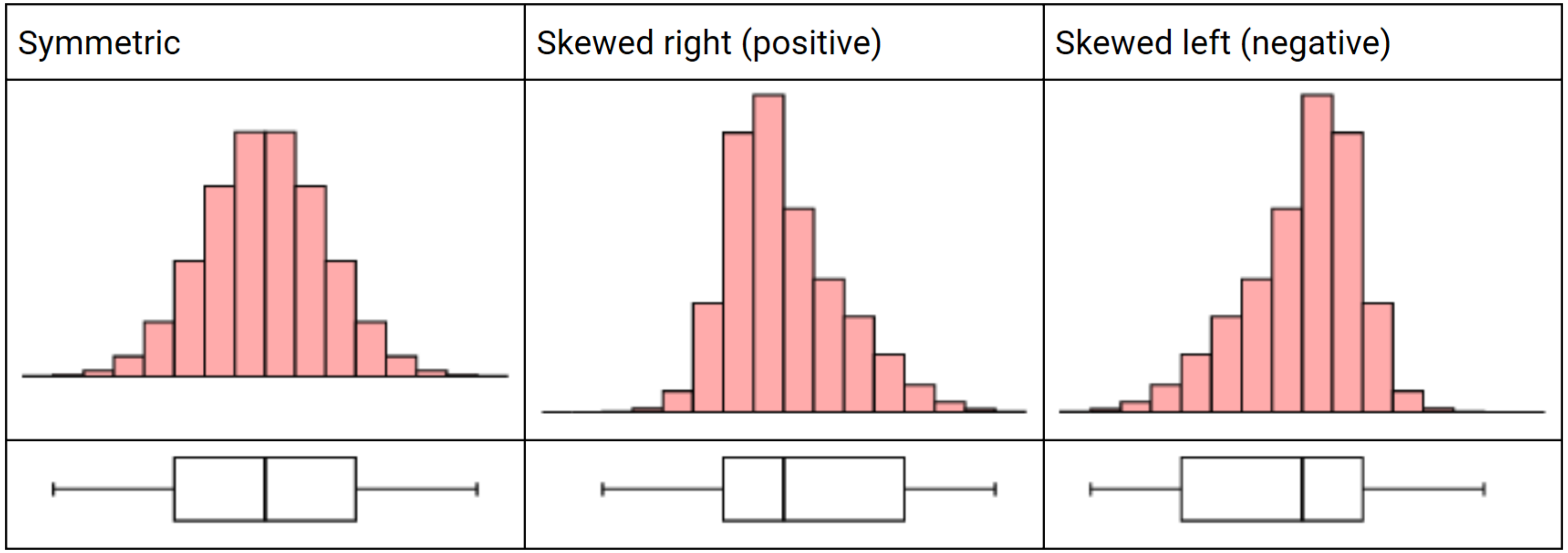
• Skew / Kurtosis

SKEW

SKEW.P

The skewness of a random variable X is the third standardized moment

$$\tilde{\mu}_3 = \mathrm{E}\left[\left(\frac{X-\mu}{\sigma}\right)^3\right] = \frac{\mu_3}{\sigma^3} = \frac{\mathrm{E}\left[(X-\mu)^3\right]}{\left(\mathrm{E}\left[(X-\mu)^2\right]\right)^{3/2}}$$



Standardized moment

$$\hat{\mu}_k = \frac{\mu_k}{\sigma^k} = \frac{\mathrm{E}\left[(X-\mu)^k\right]}{\left(\mathrm{E}\left[(X-\mu)^2\right]\right)^{k/2}}$$

• Skew / Kurtosis

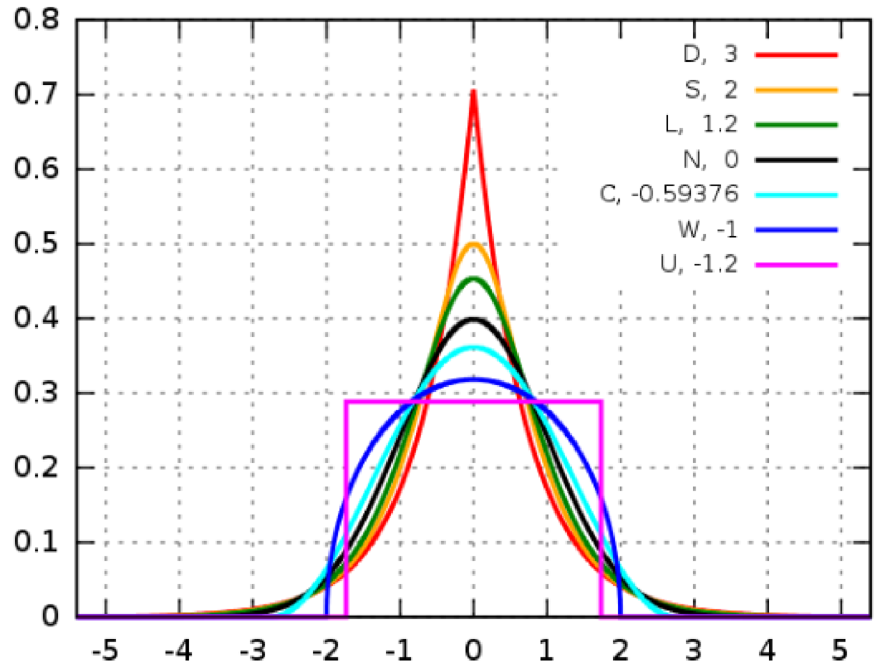
KURT

The kurtosis is the fourth standardized moment, defined as

$$\operatorname{Kurt}[X] = \operatorname{E}\left[\left(\frac{X-\mu}{\sigma}\right)^4\right] = \frac{\operatorname{E}[(X-\mu)^4]}{(\operatorname{E}[(X-\mu)^2])^2} = \frac{\mu_4}{\sigma^4}$$

Standardized moment

$$\hat{\mu}_k = \frac{\mu_k}{\sigma^k} = \frac{\operatorname{E}[(X-\mu)^k]}{(\operatorname{E}[(X-\mu)^2])^{k/2}}$$



Probability density functions for selected distributions with mean 0, variance 1 and different excess kurtosis

Lecture 3: Probability

Counting Rules

- **Factorial** is the product of all the positive numbers from 1 to a number.

$$n! = n(n-1)(n-2)\cdots 3\cdot 2\cdot 1$$

$$0! = 1$$

- **Permutation** is an arrangement of objects in a specific order. Order matters.

$$_{n}P_{r} = \frac{n!}{(n-r)!} = \underbrace{n(n-1)(n-2)\cdots(n-r+1)}_{r\ \text{items}}$$

- **Combination** is a grouping of objects. Order does not matter.

$$_{n}C_{r} = \frac{n!}{(n-r)!\,r!}$$

$$= \frac{_{n}P_{r}}{r!}$$

- **Factorial**
- **Permutations**
- **Combination**

FACT

PERMUT

PERMUTATIONA With replacement

COMBIN

COMBINA With replacement

$$P_k^n = \frac{n!}{(n-k)!}$$

$$n^k$$

$$C_k^n = \frac{n!}{(n-k)!\,k!}$$

$$C_k^{n+k-1}$$

Key Question: Select k objects be selected out of n objects ($0 \le k \le n$). How many ways can we select, if the selections are made **replacement**? Placing k markers on n numbers

n	$n-1$	$n-2$	……………	2	1

k identical balls into n distinct boxes→ Using $n-1$ walls to separate k balls

Lecture 4: Discrete Random Variables

- Probability Distributions, mean and variance
- Binomial Distribution
- Multinomial Distribution
- Poisson Distribution
- Hypergeometric Distribution

- **Probability Mass Function**

$\boxed{\text{PROB}}$

$$\Pr(a \le X \le b)$$

$$\begin{cases} \Pr(X=0) = \Pr(\{TT\}) = \dfrac{1}{4} \\ \Pr(X=1) = \Pr(\{TH, HT\}) = \dfrac{2}{4} = \dfrac{1}{2} \\ \Pr(X=2) = \Pr(\{HH\}) = \dfrac{1}{4} \end{cases}$$

- **Binomial distribution:** Probability distribution on the number of successes in n independent experiments, each experiment has a probability of success p

BINOM.DIST
BINOM.DIST.RANGE
BINOM.INV

$$X \sim Binomial(n, p)$$

$$\Pr(X = x) = \binom{n}{x} p^x (1-p)^{n-x},$$
$$for\ x = 0,1,2,\cdots,n$$

=BINOM.DIST(

BINOM.DIST(**number_s**, trials, probability_s, cumulative)

$\Pr(X = x)$ or $\Pr(X \leq x)$

=BINOM.DIST.RANGE(

BINOM.DIST.RANGE(**trials**, probability_s, number_s, [number_s2])

$\Pr(x_1 \leq X \leq x_2)$

=BINOM.INV(

BINOM.INV(trials, probability_s, alpha)

$\Pr(X \leq x) \geq \alpha$

Probability	Successes/Failures
$p^x(1-p)^{n-x}$	S S S … S S F F F … F
$p^x(1-p)^{n-x}$	S S S … S F S F F … F
………	………
$p^x(1-p)^{n-x}$	F F F F … F F F S … S

- **Poisson distribution**

POISSON.DIST

=POISSON.DIST(

POISSON.DIST(**x**, mean, cumulative)

Poisson Distribution is usually associated with rare events.

<u>Independent events occur over a period of time or space</u>

[The second most frequently used discrete distribution.]

The probability of k events occurring for a Poisson random variable with parameter μ is

$$\Pr(X = k) = \frac{e^{-\mu}\mu^{k}}{k!}, \mathrm{k} = 0,1,2,\cdots,$$

Where e is the Euler's constant, and μ is **<u>expected number of events to occur</u>**.

$$e = \lim_{n\to\infty}\left(1+\frac{1}{n}\right)^{n} \approx 2.718$$

$$\boxed{X \sim Poisson(\mu)}$$

- **Negative Binomial Distribution**

NEGBINOM.DIST

=NEGBINOM.DIST(

NEGBINOM.DIST(**number_f**, number_s, probability_s, cumulative)

Probability distribution on the number of times when the r success occurs[a fixed integer], each experiment has a probability of success p.

$X \sim Negative\ Binomial\ (r, p)$

Let X is the **number of failure times** , then

X follows Negative Binomial Distribution with parameters r and p.

The probability of k times when the r **success occurs** success occurs is

$$\Pr(X = k) = \binom{k + r - 1}{r - 1}(1 - p)^k p^r, \qquad k = 0,1,2\cdots,$$

Mean: $E(X) = \frac{r}{p}$ **Variance:** $\sigma^2 = \frac{r(1-p)}{p^2}$

Negative Binomial distribution reduces to Geometric distribution when $r = 1$.

- **Hypergeometric Distribution**

HYPGEOM.DIST

=HYPGEOM.DIST(

HYPGEOM.DIST(**sample_s**, number_sample, population_s, number_pop, cumulative)

Suppose we have a box containing **N_1 white balls and N_2 black balls**. We randomly select n balls ($n < N_1 + N_2 = N$) **without replacement**.

Let X be the **number of balls being white**

$$\boxed{X \sim Hypergeometric(N_1, N_2, n)}$$

The probability of x white balls in n balls is

$$\Pr(X = x) = \frac{\binom{N_1}{x}\binom{N_2}{n-x}}{\binom{N}{n}}$$

$k\ being\ nonnegative\ integer$: $x \le n, x \le N_1, n - x \le N_2$

Mean: $E(X) = n(\frac{N_1}{N})$ **Variance**: $\sigma^2 = n(\frac{N_1}{N})(\frac{N_2}{N})(\frac{N-n}{N-1})$

Normal distribution

Lecture 5 Normal Distributions

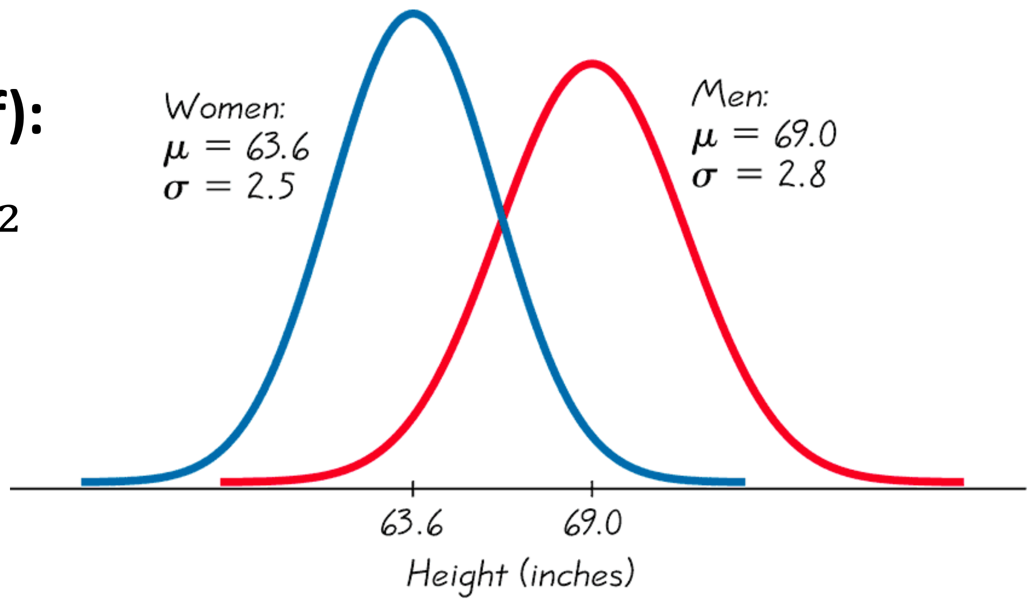
- NORM.DIST
- NORM.INV
- NORM.S.DIST
- NORM.S.INV
- LOGNORM.DIST
- LOGNORM.INV
- PHI
- STANDARDIZE
- GAUSS

$\boxed{X \sim Normal(\mu, \sigma^2)}$

μ (population mean)
σ^2 (population variance)

Probability Density Function (pdf):

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

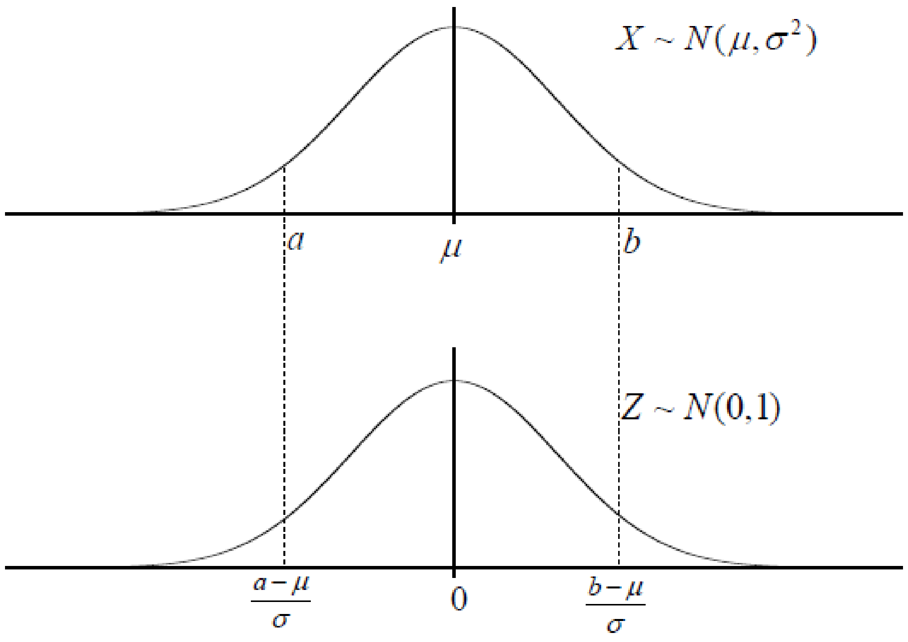


- **Normal distribution**

NORM.DIST
NORM.INV
NORM.S.DIST
NORM.S.INV
LOGNORM.DIST
LOGNORM.INV

PHI
STANDARDIZE
GAUSS

$$\Pr(a \le X \le b) = \Pr\left(\frac{a-\mu}{\sigma} \le Z \le \frac{b-\mu}{\sigma}\right)$$



$$Z = \frac{X-\mu}{\sigma} \quad Z \sim N(0,1)$$

Centralize

Shape of normal distribution preserves after translation (by $-\mu$) and rescaling (by $1/\sigma$)

- **Normal distribution**

NORM.DIST

NORM.INV

NORM.S.DIST

NORM.S.INV

LOGNORM.DIST

LOGNORM.INV

PHI

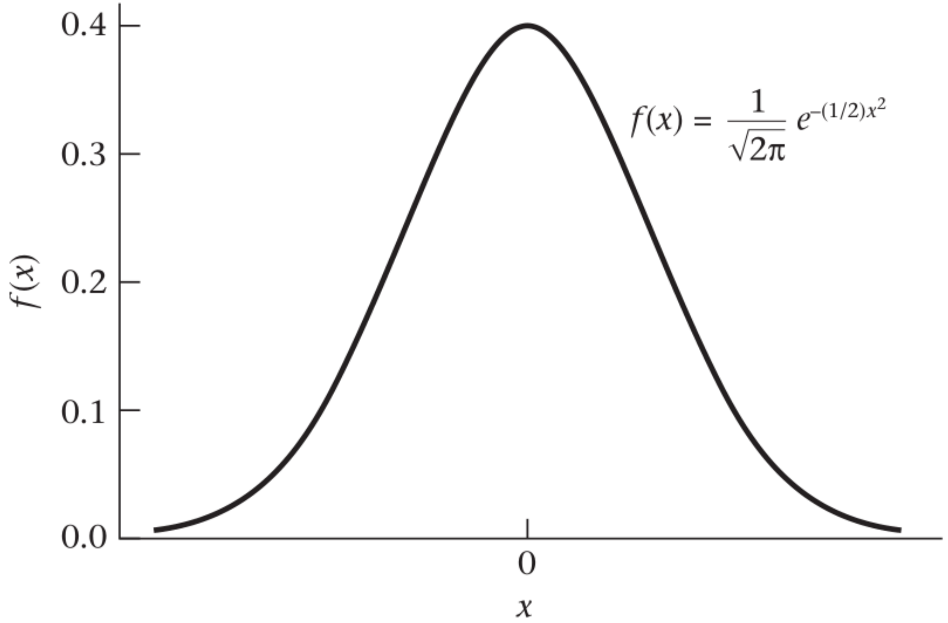
STANDARDIZE

GAUSS

The standard normal distribution $N(0,1)$ is a normal probability distribution with $\mu = 0$ and $\sigma^2 = 1$, **[Special Case]**

Probability density function (pdf) of $N(0,1)$: $-\infty < z < \infty$

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



$f(x) = f(-x)$: The distribution is symmetric about 0

Mean = Median = Mode = 0

- **Normal distribution**

NORM.DIST

NORM.INV

$\boxed{\text{NORM.S.DIST}}$

$\boxed{\text{NORM.S.INV}}$

LOGNORM.DIST

LOGNORM.INV

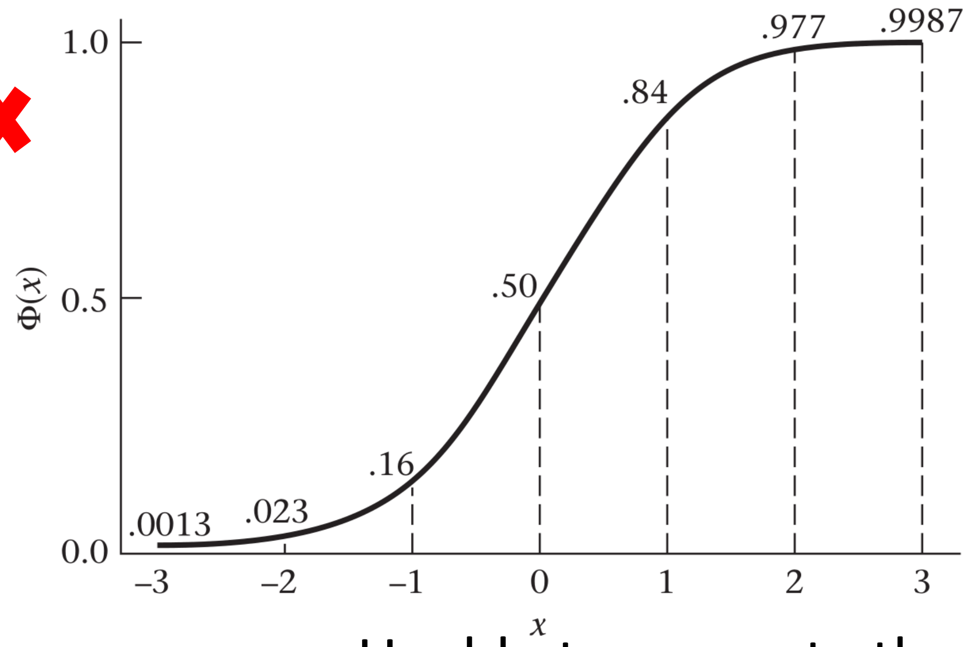
$\boxed{\text{PHI}}$

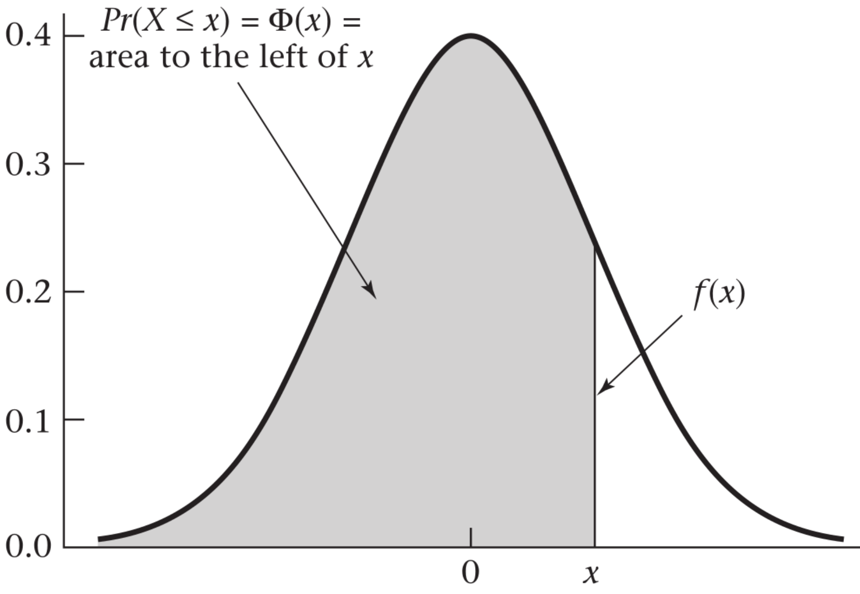
STANDARDIZE

$\boxed{\text{GAUSS}}$ -0.5

Notation for the cdf of $N(0,1)$:

$$\Phi(z) = F(z) = \Pr(Z \leq z)$$





Unable to compute the exact value of $\Phi(z)$ by hand

- **Normal distribution**

NORM.DIST

NORM.INV

NORM.S.DIST

NORM.S.INV

LOGNORM.DIST

LOGNORM.INV

PHI

STANDARDIZE

GAUSS

$$\boxed{X \sim Normal(\mu, \sigma^2)}$$

μ (population mean)

σ^2 (population variance)

Probability Density Function (pdf):

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

$$\boxed{Y = \exp(X) \sim Log - Normal(\mu, \sigma^2)}$$

A log-normal process is the statistical realization of the multiplicative product of many independent random variables, each of which is positive. This is justified by considering the central limit theorem in the log domain

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

only positive real values

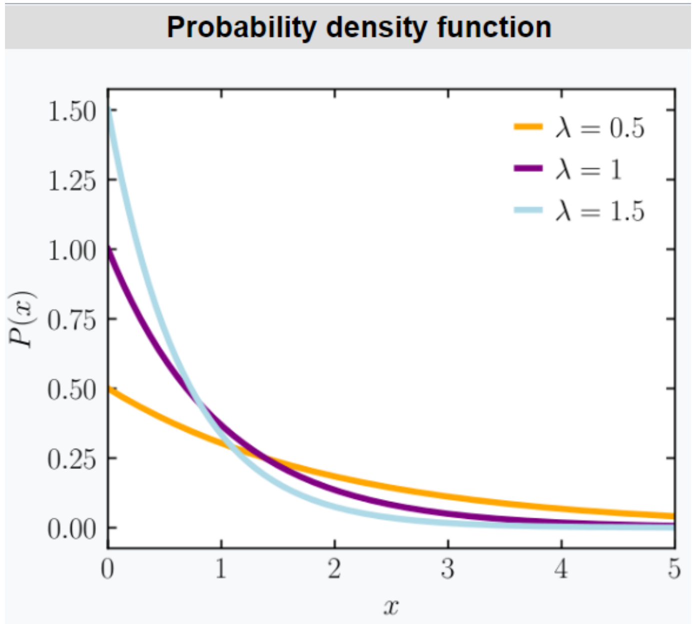
- **Exponential Distribution:** Used to model the length of time between two occurrences of an event

Let X is the length of time , then

X follows **Exponential Distribution** with parameters λ.

$$X \sim \text{exponential}(\lambda)$$

$$f(x) = \lambda e^{-\lambda x} \qquad for\ 0 < x < \infty$$



$$E(X) = \frac{1}{\lambda}$$

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

```
=EXPON.DIST(
EXPON.DIST(x, lambda, cumulative)
```

The probability that an arrival time is less than some specified time t

$$\Pr(arrival\ time < t) = \Pr(X < t) = \int_0^t \lambda e^{-\lambda x}\, dt = -e^{-\lambda x}\Big|_0^t = 1 - e^{-\lambda t}$$

- **Exponential Distribution:** Used to model the length of time between two occurrences of an event

Exponential-Poisson relationship

When events happen over time independently and at a constant rate:

1. The **number of events** in any fixed time period is **Poisson distribution**.
2. The **waiting time** between events is **Exponential distribution.**

Example: Suppose customers arrive independently at a constant mean rate of **40 Per Hour**. What's the probability that **at least one customer** arrives in the **next five Minutes**?

Let Y is the **number of events**

$$Y \sim Poisson(\frac{40}{60} * 5)$$

$$\Pr(Y \geq 1) = 1 - \Pr(Y = 0) = 1 - \frac{e^{-10/3}\mu^0}{0!}$$

Let X is the **arrival time for the first customers**

$$X \sim \text{exponential}(40)$$

$$\Pr\left(X < \frac{5}{60}\right) = 1 - e^{-40*5/60}$$

- **Gamma Distribution / Function**

GAMMA
GAMMA.DIST
GAMMA.INV
GAMMALN
GAMMALN.PRECISE

Gamma distribution is frequently used to model waiting times. For instance, in life testing, the waiting time until death is a random variable that is frequently modeled with a gamma distribution.

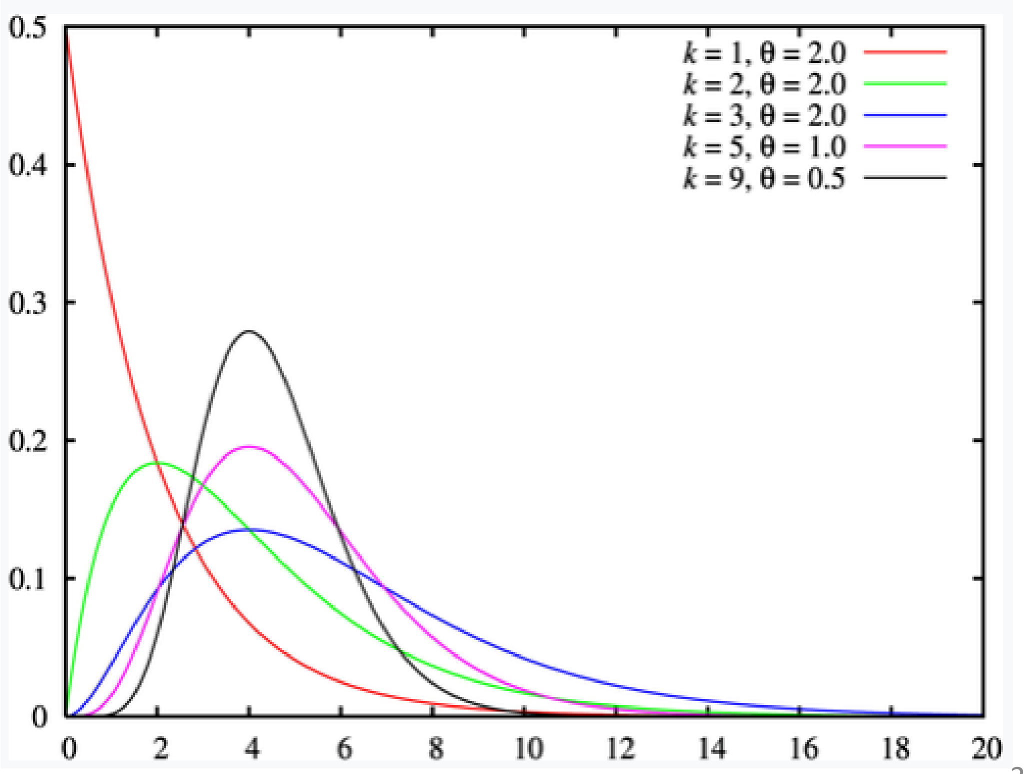
$\alpha \quad \beta$

$$X \sim \Gamma(k, \theta) \equiv \mathrm{Gamma}(k, \theta)$$

$$f(x; k, \theta) = \frac{x^{k-1} e^{-\frac{x}{\theta}}}{\theta^k \Gamma(k)} \quad \text{for } x > 0 \text{ and } k, \theta > 0.$$

$$E(X) = k\theta \qquad Var(X) = k\theta^2$$

$$\alpha = 1 \qquad \lambda = \frac{1}{\beta} = \frac{1}{\theta} \qquad \boxed{X \sim \mathrm{exponential}(\lambda)}$$



- **Gamma Distribution / Function**

GAMMA
GAMMA.DIST
GAMMA.INV
GAMMALN
GAMMALN.PRECISE

$Log(\Gamma(X))$

$$\Gamma(x) = \int_0^\infty t^x e^{-t} dt$$

$$\Gamma(n) = (n-1)!$$ n is positive integer

$$\Gamma(x+1) = x\Gamma(x)$$

$$\Gamma(0.5) = \sqrt{\pi}$$

$$\Gamma(-0.5) = -2\sqrt{\pi}$$

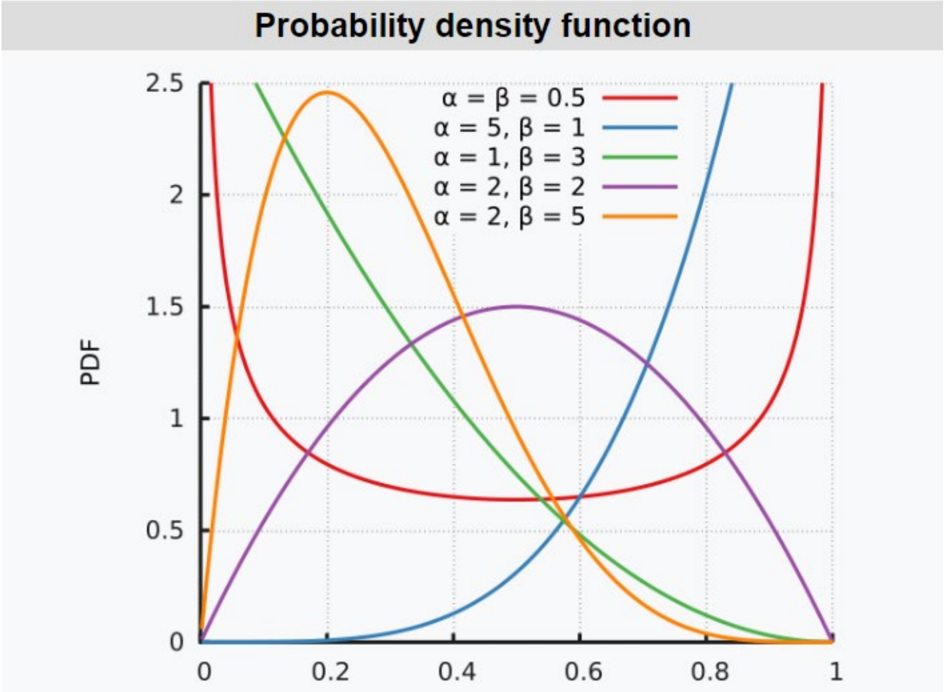
$$f(x;k,\theta) = \frac{x^{k-1}e^{-\frac{x}{\theta}}}{\theta^k \boxed{\Gamma(k)}} \quad \text{for } x > 0 \text{ and } k, \theta > 0$$

- **Beta distribution:** a family of continuous distributions defined on the interval $[0, 1]$ parameterized by two positive shape parameters by α and β

BETA.DIST

BETA.INV

$$\begin{aligned} f(x;\alpha,\beta) &= \mathrm{constant}\cdot x^{\alpha-1}(1-x)^{\beta-1} \\ &= \frac{x^{\alpha-1}(1-x)^{\beta-1}}{\int_0^1 u^{\alpha-1}(1-u)^{\beta-1}\,du} \\ &= \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)}\,x^{\alpha-1}(1-x)^{\beta-1} \\ &= \frac{1}{\mathrm{B}(\alpha,\beta)}x^{\alpha-1}(1-x)^{\beta-1} \end{aligned}$$



- **Weibull distribution**

If the quantity X is a "time-to-failure", the Weibull distribution gives a distribution for which the failure rate is proportional to a power of time.

WEIBULL.DIST

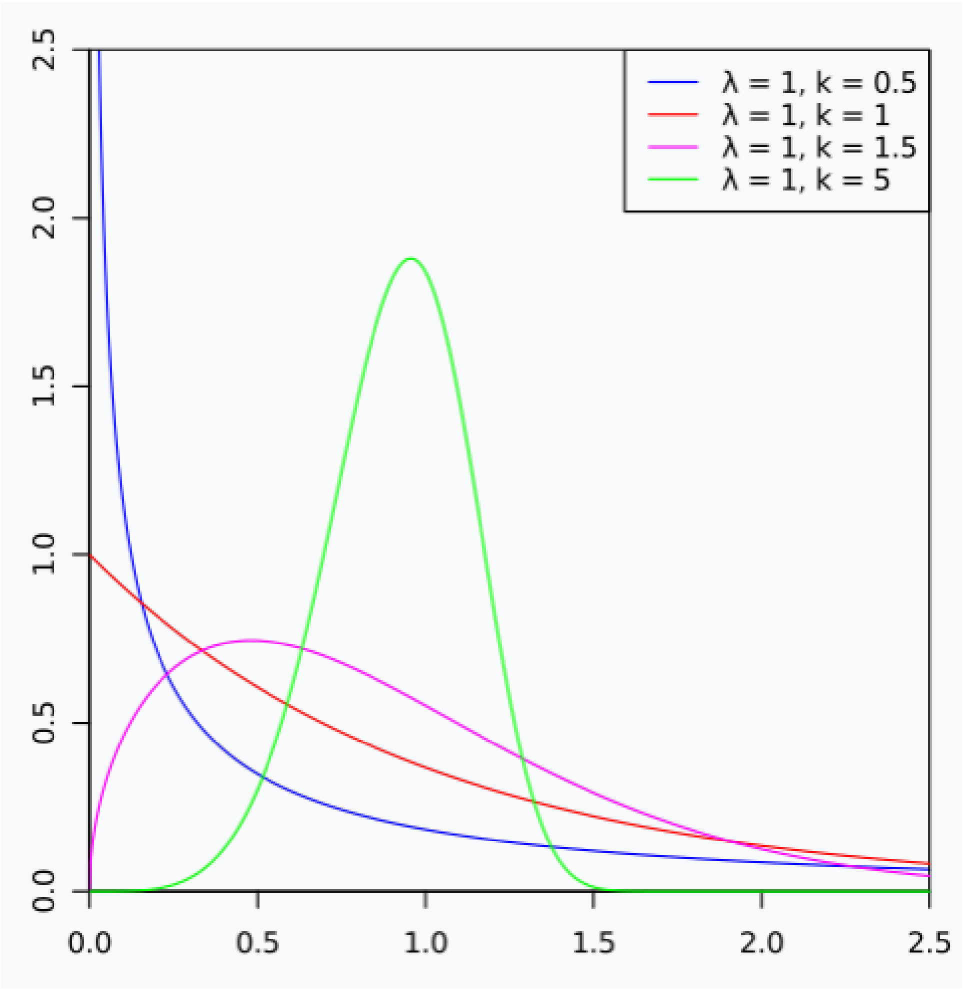
=WEIBULL.DIST(

WEIBULL.DIST(**x**, alpha, beta, cumulative)

$$f(x;\lambda,k)=\begin{cases}\frac{k}{\lambda}\left(\frac{x}{\lambda}\right)^{k-1}e^{-(x/\lambda)^{k}} & x\geq 0\\ 0 & x<0\end{cases}$$

$$E=\lambda\Gamma\left(1+\frac{1}{k}\right)$$

$$Var=\lambda^{2}\left[\Gamma\left(1+\frac{2}{k}\right)-\Gamma\left(1+\frac{1}{k}\right)^{2}\right]$$



- **Confidence Interval for Mean**

Lecture 6
Confidence Intervals and Sample Size

CONFIDENCE.NORM

CONFIDENCE.T

=CONFIDENCE.NORM(

CONFIDENCE.NORM(**alpha**, standard_dev, size)

=CONFIDENCE.T(

CONFIDENCE.T(**alpha**, standard_dev, size)

A **100(1-α)% Confidence Interval** for μ when **σ is known**

$$(\bar{x} - \boxed{\frac{z_{1-\alpha/2}\sigma}{\sqrt{n}}}, \bar{x} + \frac{z_{1-\alpha/2}\sigma}{\sqrt{n}})$$

A **100(1-α)% Confidence Interval** for μ when **σ is Unknown**

$$(\bar{x} - \boxed{\frac{t_{n-1,1-\alpha/2}S}{\sqrt{n}}}, \bar{x} + \frac{t_{n-1,1-\alpha/2}S}{\sqrt{n}})$$

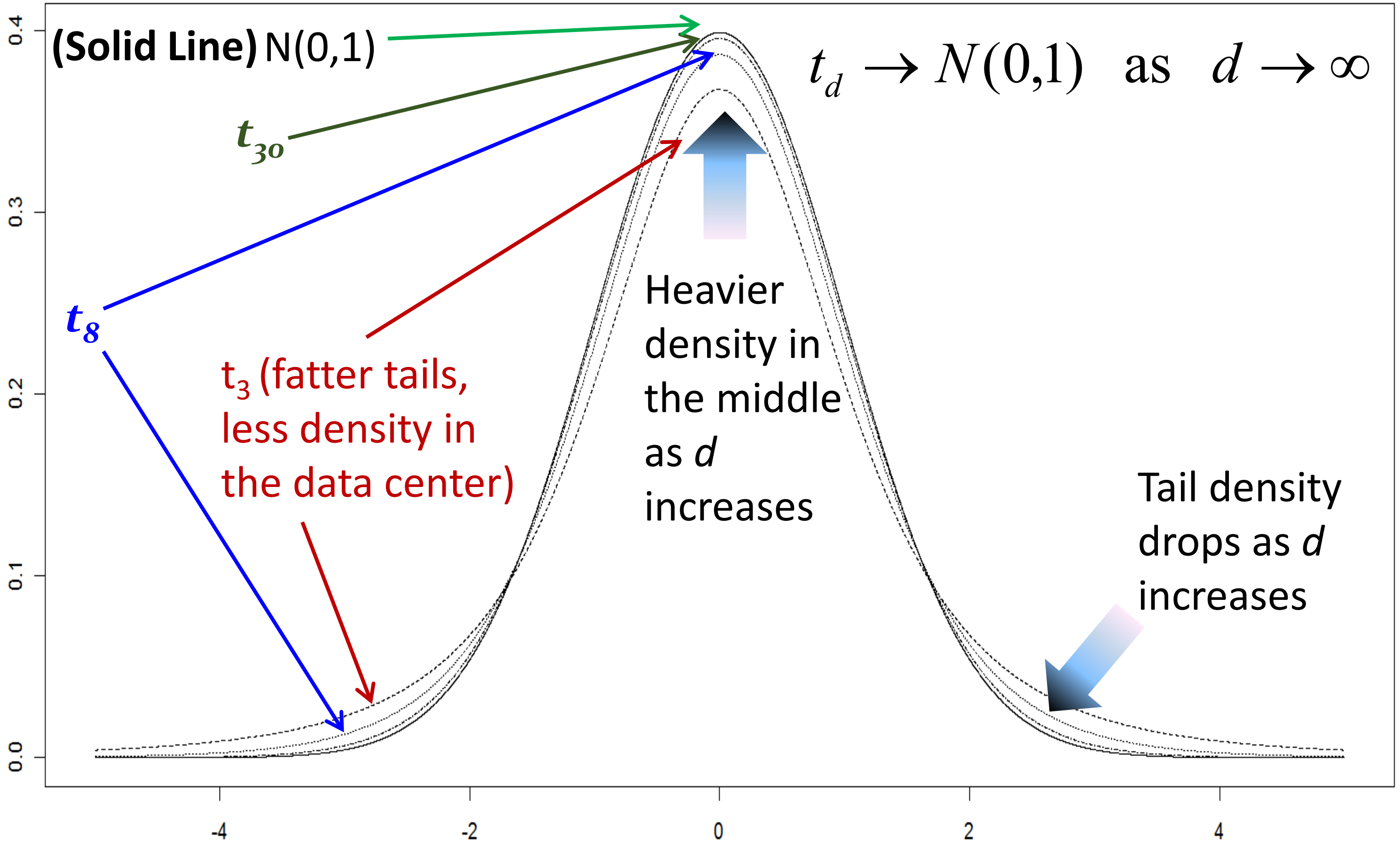
- **Student t-distribution**

If W is $Normal(0,1)$, V is $\chi^2(r)$ and W and V are independent

$$T = \frac{W}{\sqrt{V/r}}$$

$$T \sim t_r \quad t - distribution\ (r)$$

T.DIST
T.DIST.2T
T.DIST.RT
T.INV
T.INV.2T



Parameters: Degrees of freedom (df)

- **Bell-shaped symmetric pdf** (Similar to N(0,1)) but heavier densities at both tails
- As degree of freedom (d.f.) d increases, density of the (1) tails decrease, but (2) middle increases

- **Chi-square distribution**

Distribution can only be obtained when X_i follows Normal distribution

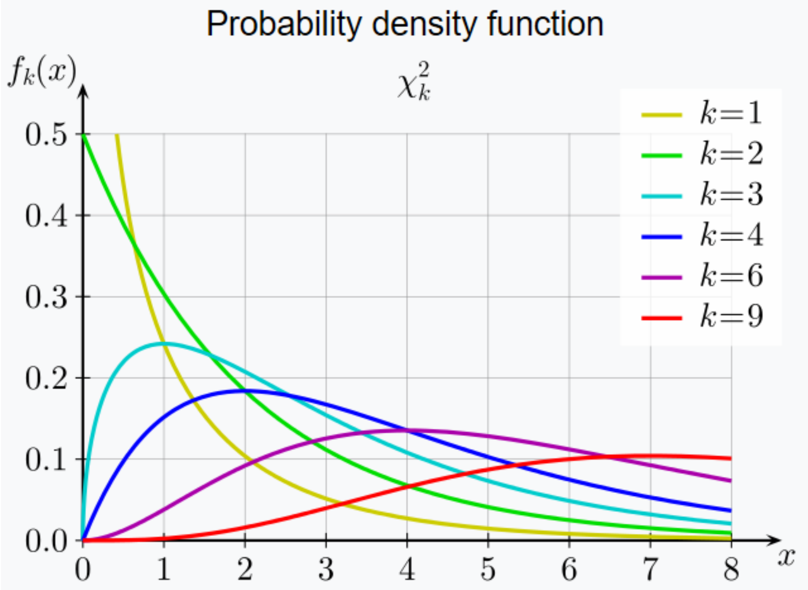
$X_i \sim \mathrm{N}(0,1)$ **&** $Y_i = X_i^2$ ⟶ $Y_i \sim \chi_1^2$ **[Chi-square distribution]**

Independent **ADDITIVITY** **Parameters: Degrees of freedom (df)**

$X_1, \cdots, X_n \sim \mathrm{iid}\ \mathrm{N}(0,1)$ $G = X_1^2 + X_2^2 + \cdots + X_n^2 = Y_1 + \cdots + Y_n$ $G \sim \chi_n^2$

CHISQ.DIST	$\Pr(X < x)$ or $f(x)$
CHISQ.DIST.RT	$\Pr(X > x)$
CHISQ.INV	$\Pr(X < x) = \alpha$
CHISQ.INV.RT	$\Pr(X > x) = \alpha$

$$f_k(x) = \frac{1}{2^{\frac{k}{2}}\Gamma(\frac{k}{2})} x^{\frac{k}{2}-1} e^{\frac{-x}{2}}$$



- **Hypothesis testing for Means**

$\boxed{\text{Z.TEST}}$ **[One sample]** $\boxed{\text{T.TEST}}$ **[Two sample]**

Objective: Compare population mean μ with some value μ_0

Assumption: A random sample $X_1, X_2, \ldots X_n$ with **unknown μ but known σ**, with either of the following condition:

1. X_i are normally distributed, or
2. n is large $(n > 30)$

$$\boxed{\bar{X} \sim Normal(\mu, \frac{\sigma^2}{n})}$$

Possible Hypotheses:

$$\begin{cases} H_0: \mu = \mu_0 \\ H_1: \mu \neq \mu_0 \end{cases} \quad \text{or} \quad \begin{cases} H_0: \mu = \mu_0 \\ H_1: \mu > \mu_0 \end{cases} \quad \text{or} \quad \begin{cases} H_0: \mu = \mu_0 \\ H_1: \mu < \mu_0 \end{cases}$$

Test Statistic and its Distribution:

Under H_0 (i.e. Assume that H_0 is true), $z_0 = \frac{\bar{x} - \mu_0}{\sigma/\sqrt{n}}$

- **Hypothesis testing for Means**

Z.TEST **[One sample]**

Objective: Compare population mean μ with some value μ_0

Assumption: A random sample $X_1, X_2, \ldots X_n$ with **unknown μ but known σ,** with either of the following condition: **Unknown**

1. X_i are normally distributed, or
2. n is large $(n > 30)$

$$\boxed{\bar{X} \sim Normal(\mu, \frac{\sigma^2}{n})}$$

Possible Hypotheses:

$$\begin{cases} H_0: \mu = \mu_0 \\ H_1: \mu > \mu_0 \end{cases}$$

$$\mathrm{p-value} = \Pr(Z > z_0)$$

Test Statistic and its Distribution:

Under H_0 (i.e. Assume that H_0 is true), $z_0 = \frac{\bar{x} - \mu_0}{\sigma/\sqrt{n}}$

- **Hypothesis testing for Means**

Z.TEST **[One sample]**

Objective: Compare population mean μ with some value μ_0

Assumption: A random sample $X_1, X_2, \ldots X_n$ with **unknown μ but known σ, Unknown**
with either of the following condition:

1. X_i are normally distributed, or
2. n is large $(n > 30)$

$$\boxed{\bar{X} \sim Normal(\mu, \frac{\sigma^2}{n})}$$

Possible Hypotheses:

$$\begin{cases} H_0: \mu = \mu_0 \\ H_1: \mu \neq \mu_0 \end{cases}$$

$\mathrm{p-value} = 2 * \min(\Pr(Z > z_0), \Pr(Z < z_0))$

$$\begin{cases} H_0: \mu = \mu_0 \\ H_1: \mu < \mu_0 \end{cases}$$

More extreme

$\mathrm{p-value} = \Pr(Z < z_0)$

Test Statistic and its Distribution:

Under H_0 (i.e. Assume that H_0 is true), $z_0 = \frac{\bar{x} - \mu_0}{\sigma/\sqrt{n}}$

- **Hypothesis testing for Means**

Z.TEST **[One sample]** T.TEST **[Two sample]**

A more frequently encountered situation is **the two-sample hypothesis testing problem**. Compare the underlying parameters of two different populations, neither of whose values is assumed known, are compared.

$$H_0: \mu_1 = \mu_2$$

Compare two sample (Unknown value)

Population 1 μ_1, σ_1^2 → **Random Sample** $X_1, X_2, \ldots, X_n$

Compare

Population 2 μ_2, σ_2^2 → **Random Sample** $Y_1, Y_2, \ldots, Y_m$

• Hypothesis testing for Means

T.TEST **[Two sample]**

Longitudinal Study (縱向研究) vs Cross-sectional study (橫斷面研究)

Same group of subjects are followed over time/condition (Repeated observations)

Interested in the **difference** in measurements **over time/condition.**

Two observations (different time) from a single subject

[Two observations are Dependent]

$X_1 \sim Normal(\mu_1, \sigma^2)$ $X_2 \sim Normal(\mu_2, \sigma^2)$ ……

Dependent

Independent (different distribution) ……

$Y_1 \sim Normal(\mu_1 + \Delta, \sigma^2)$ $Y_2 \sim Normal(\mu_2 + \Delta, \sigma^2)$

Multiple groups of subjects are seen at one time point.

Interested in the **difference** in measurements **over groups.**

One observation from a single subject in different groups

[Groups are Independent]

$X_1, X_2, \cdots, X_n \sim Normal(\mu_1', \sigma^2)$

Independent

$Y_1, Y_2, \cdots, Y_m \sim Normal(\mu_2', \sigma^2)$

- **Hypothesis testing for Means**

T.TEST **[Two sample]** **TYPE I**

Objective: Compare the population means for paired data

Example: $X_1 \sim Normal(\mu_1, \sigma_1^2)$ $X_2 \sim Normal(\mu_2, \sigma_1^2)$ ……

$Y_1 \sim Normal(\mu_1 + \Delta, \sigma_2^2)$ $Y_2 \sim Normal(\mu_2 + \Delta, \sigma_2^2)$ ……

Minus

$D_1 \sim Normal(\Delta, \sigma_3^2)$ $D_2 \sim Normal(\Delta, \sigma_3^2)$ ……

Reduce to one sample case

$$D_1, D_2, \cdots, D_n \sim \boxed{Normal}(\Delta, \sigma_3^2)$$

The difference in the two measurements for the i^{th} subject

Possible Hypotheses:

$$\begin{cases} H_0\colon \Delta = 0 \\ H_1\colon \Delta \neq 0 \end{cases} \quad \begin{cases} H_0\colon \Delta = 0 \\ H_1\colon \Delta > 0 \end{cases} \quad \begin{cases} H_0\colon \Delta = 0 \\ H_1\colon \Delta < 0 \end{cases}$$

Two-tailed **one-tailed** **P-value**

$$t_0 = \frac{\bar{d} - 0}{s/\sqrt{n}}$$ Test statistic

$$s^2 = \frac{1}{n-1}\sum_{i=1}^{n}(d_i - \bar{d})^2$$

- **Hypothesis testing for Means**

T.TEST **[Two sample]** **TYPE II**

Objective: Compare the population means for two independent samples [Cross Sectional]

$$\frac{\bar{X}-\bar{Y}-(\mu_1-\mu_2)}{\sqrt{(\sigma_1^2)/n+(\sigma_2^2)/m}} \sim Normal(0,1)$$

Estimate population variance by sample variance

$$\frac{\bar{X}-\bar{Y}-(\mu_1-\mu_2)}{\sqrt{(S_1^2)/n+(S_2^2)/m}}$$

Assumptions:

- When n and m are small $n, m \leq 30$, assumption on Normal distribution is required (Linear combination):
  $$X_1, \cdots, X_n \sim Normal(\mu_1, \sigma_1^2)$$
  $$Y_1, \cdots, Y_m \sim Normal(\mu_2, \sigma_2^2)$$
- If σ_1^2, σ_2^2 are Unknown but **equal to each other:** $\sigma_1^2 = \sigma_2^2 = \sigma^2$

Define the pooled estimate of variance as:

$$S_p^2 = \frac{(n-1)S_1^2 + (m-1)S_2^2}{n+m-2}$$

which is a weighted average of two sample variances.

$$E\left(S_p^2\right) = \sigma^2 = \sigma_1^2 = \sigma_2^2$$

• Hypothesis testing for Means

T.TEST [Two sample] TYPE II

Objective: Compare the population means for two independent samples [Cross Sectional]

$$\frac{\bar{X}-\bar{Y}-(\mu_1-\mu_2)}{\sqrt{(\sigma_1^2)/n+(\sigma_2^2)/m}} \sim Normal(0,1)$$

Estimate population variance by sample variance

$$\frac{\bar{X}-\bar{Y}-(\mu_1-\mu_2)}{\sqrt{(S_1^2)/n+(S_2^2)/m}}$$

Assumptions:

- When n and m are small $n, m \leq 30$, assumption on Normal distribution is required (Linear combination):

$$X_1, \cdots, X_n \sim Normal(\mu_1, \sigma_1^2)$$
$$Y_1, \cdots, Y_m \sim Normal(\mu_2, \sigma_2^2)$$

- If σ_1^2, σ_2^2 are Unknown but **equal to each other:** $\sigma_1^2 = \sigma_2^2 = \sigma^2$

F-test

$$\frac{\bar{X}-\bar{Y}-(\mu_1-\mu_2)}{S_p\sqrt{1/n+1/m}} \quad t-distribution\ (n+m-2)$$

$$= \frac{\bar{X}-\bar{Y}-(\mu_1-\mu_2)}{\sigma\sqrt{1/n+1/m}} \div \sqrt{\left[\frac{(n-1)S_1^2}{\sigma^2}+\frac{(m-1)S_2^2}{\sigma^2}\right]/(n+m-2)}$$

- **Hypothesis testing for Means**

T.TEST [Two sample] TYPE II

Objective: Compare the population means for two independent samples [Cross Sectional]

$$\frac{\bar{X}-\bar{Y}-(\mu_1-\mu_2)}{\sqrt{(\sigma_1^2)/n+(\sigma_2^2)/m}} \sim Normal(0,1) \xrightarrow{\text{Estimate population variance by sample variance}} \frac{\bar{X}-\bar{Y}-(\mu_1-\mu_2)}{\sqrt{(S_1^2)/n+(S_2^2)/m}}$$

$$\begin{cases} H_0: \mu_1-\mu_2=0 \\ H_1: \mu_1-\mu_2 \neq 0 \end{cases}$$

$$\begin{cases} H_0: \mu_1-\mu_2=0 \\ H_1: \mu_1-\mu_2 > 0 \end{cases}$$

$$\begin{cases} H_0: \mu_1-\mu_2=0 \\ H_1: \mu_1-\mu_2 < 0 \end{cases}$$

$$\frac{\bar{X}-\bar{Y}-(\mu_1-\mu_2)}{S_p\sqrt{1/n+1/m}} \quad t-distribution\ (n+m-2)$$

$$=\frac{\bar{X}-\bar{Y}-(\mu_1-\mu_2)}{\sigma\sqrt{1/n+1/m}} \div \sqrt{\left[\frac{(n-1)S_1^2}{\sigma^2}+\frac{(m-1)S_2^2}{\sigma^2}\right]/(n+m-2)}$$

$$t_0=\frac{\bar{x}-\bar{y}}{s_p\sqrt{1/n+1/m}}$$

Test statistic

- **Hypothesis testing for Means**

$\boxed{\text{T.TEST}}$ **[Two sample]** **TYPE III**

Objective: Compare the population means for two independent samples [Cross Sectional]

$$\frac{\bar{X}-\bar{Y}-(\mu_1-\mu_2)}{\sqrt{(\sigma_1^2)/n+(\sigma_2^2)/m}} \sim Normal(0,1)$$

Estimate population variance by sample variance

$$\frac{\bar{X}-\bar{Y}-(\mu_1-\mu_2)}{\sqrt{(S_1^2)/n+(S_2^2)/m}}$$

Assumptions:

- When n and m are small $n, m \le 30$, assumption on Normal distribution is required (Linear combination):

$$X_1, \cdots, X_n \sim Normal(\mu_1, \sigma_1^2)$$
$$Y_1, \cdots, Y_m \sim Normal(\mu_2, \sigma_2^2)$$

- ~~If σ_1^2, σ_2^2 are Unknown but **equal to each other**: $\sigma_1^2 = \sigma_2^2 = \sigma^2$~~

NO RELATIONSHIP ASSUMPTION

Difficult to find the appropriate t-distribution

Satterthwaite's Method

Approximately to t-distribution:

$$d' = \frac{(s_1^2/n + s_2^2/m)^2}{(s_1^2/n)^2/(n-1) + (s_2^2/m)^2/(m-1)}$$

- **F - Distribution:**

F.DIST

F.DIST.RT

F.INV

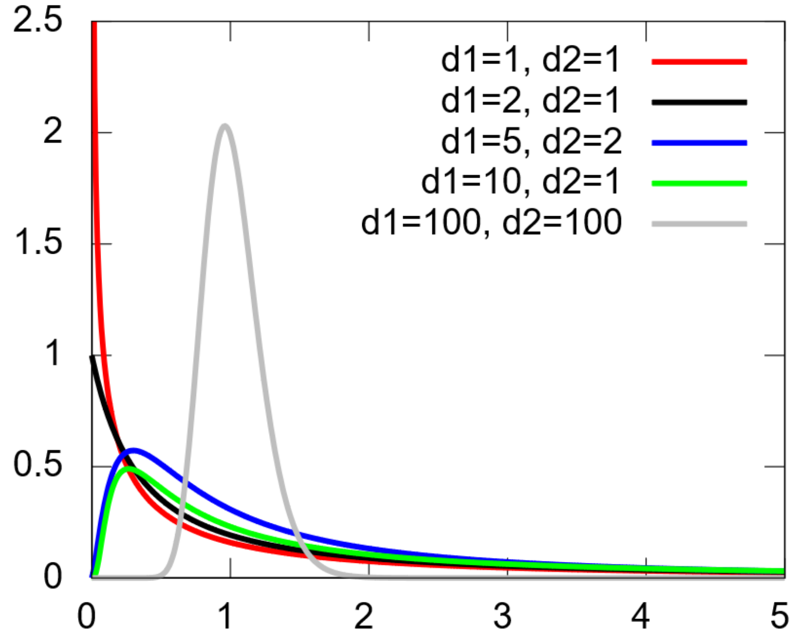
F.INV.RT

$$f(x; d_1, d_2) = \frac{\sqrt{\frac{(d_1 x)^{d_1}\, d_2^{d_2}}{(d_1 x + d_2)^{d_1+d_2}}}}{x\,\mathrm{B}\left(\frac{d_1}{2}, \frac{d_2}{2}\right)}$$

$$= \frac{1}{\mathrm{B}\left(\frac{d_1}{2}, \frac{d_2}{2}\right)} \left(\frac{d_1}{d_2}\right)^{\frac{d_1}{2}} x^{\frac{d_1}{2}-1} \left(1 + \frac{d_1}{d_2}\,x\right)^{-\frac{d_1+d_2}{2}}$$

If $X \sim \chi^2_{d_1}$ and $Y \sim \chi^2_{d_2}$ are independent, then:

$$\frac{X/d_1}{Y/d_2} \sim F(d_1, d_2)$$



df for denominator, d_2	p	*df* for numerator, d_1: 1	2	3	4	5	6	7	8	12	24	∞
1	.90	39.86	49.50	53.59	55.83	57.24	58.20	58.91	59.44	60.71	62.00	63.33
	.95	161.4	199.5	215.7	224.6	230.2	234.0	236.8	238.9	243.9	249.1	254.3
	.975	647.8	799.5	864.2	899.6	921.8	937.1	948.2	956.7	976.7	997.2	1018.
	.99	4052.	5000.	5403.	5625.	5764.	5859.	5928.	5981.	6106.	6235.	6366.
	.995	16211.	20000.	21615.	22500.	23056.	23437.	23715.	23925.	24426.	24940.	25464.
	.999	405280.	500000.	540380.	562500.	576400.	585940.	592870.	598140.	610670.	623500.	636620.
2	.90	8.53	9.00	9.16	9.24	9.29	9.33	9.35	9.37	9.41	9.45	9.49
	.95	18.51	19.00	19.16	19.25	19.30	19.33	19.35	19.37	19.41	19.45	19.50
	.975	38.51	39.00	39.17	39.25	39.30	39.33	39.36	39.37	39.42	39.46	39.50
	.99	98.50	99.00	99.17	99.25	99.30	99.33	99.36	99.37	99.42	99.46	99.50
	.995	198.5	199.0	199.2	199.2	199.3	199.3	199.4	199.4	199.4	199.5	199.5
	.999	998.5	999.0	999.2	999.2	999.3	999.3	999.4	999.4	999.4	999.5	999.5
3	.90	5.54	5.46	5.39	5.34	5.31	5.28	5.27	5.25	5.22	5.18	5.13
	.95	10.13	9.55	9.28	9.12	9.01	8.94	8.89	8.85	8.74	8.64	8.53
	.975	17.44	16.04	15.44	15.10	14.88	14.74	14.62	14.54	14.34	14.12	13.90
	.99	34.12	30.82	29.46	28.71	28.24	27.91	27.67	27.49	27.05	26.60	26.13
	.995	55.55	49.80	47.47	46.20	45.39	44.84	44.43	44.13	43.39	42.62	41.83
	.999	167.00	148.5	141.1	137.1	134.6	132.8	131.6	130.6	128.3	125.9	123.5
4	.90	4.54	4.32	4.19	4.11	4.05	4.01	3.98	3.95	3.90	3.83	3.76
	.95	7.71	6.94	6.59	6.39	6.26	6.16	6.09	6.04	5.91	5.77	5.63
	.975	12.22	10.65	9.98	9.60	9.36	9.20	9.07	8.98	8.75	8.51	8.26
	.99	21.20	18.00	16.69	15.98	15.52	15.21	14.98	14.80	14.37	13.93	13.46
	.995	31.33	26.28	24.26	23.16	22.46	21.98	21.62	21.35	20.70	20.03	19.32
	.999	74.14	61.25	56.18	53.44	51.71	50.53	49.66	49.00	47.41	45.77	44.05
5	.90	4.06	3.78	3.62	3.52	3.45	3.40	3.37	3.34	3.27	3.19	3.10
	.95	6.61	5.79	5.41	5.19	5.05	4.95	4.88	4.82	4.68	4.53	4.36
	.975	10.01	8.43	7.76	7.39	7.15	6.98	6.85	6.76	6.52	6.28	6.02
	.99	16.26	13.27	12.06	11.39	10.97	10.67	10.46	10.29	9.89	9.47	9.02
	.995	22.78	18.31	16.53	15.56	14.94	14.51	14.20	13.96	13.38	12.78	12.14
	.999	47.18	37.12	33.20	31.09	29.75	28.83	28.16	27.65	26.42	25.13	23.79
6	.90	3.78	3.46	3.29	3.18	3.11	3.05	3.01	2.98	2.90	2.82	2.72
	.95	5.99	5.14	4.76	4.53	4.39	4.28	4.21	4.15	4.00	3.84	3.67
	.975	8.81	7.26	6.60	6.23	5.99	5.82	5.70	5.60	5.37	5.12	4.85
	.99	13.75	10.92	9.78	9.15	8.75	8.47	8.26	8.10	7.72	7.31	6.88
	.995	18.64	14.54	12.92	12.03	11.46	11.07	10.79	10.57	10.03	9.47	8.88
	.999	35.51	27.00	23.70	21.92	20.80	20.03	19.46	19.03	17.99	16.90	15.75
7	.90	3.59	3.26	3.07	2.96	2.88	2.83	2.78	2.75	2.67	2.58	2.47
	.95	5.59	4.74	4.35	4.12	3.97	3.87	3.79	3.73	3.57	3.41	3.23
	.975	8.07	6.54	5.89	5.52	5.29	5.12	4.99	4.90	4.67	4.42	4.14
	.99	12.25	9.55	8.45	7.85	7.46	7.19	6.99	6.84	6.47	6.07	5.65
	.995	16.24	12.40	10.88	10.05	9.52	9.16	8.89	8.68	8.18	7.65	7.08
	.999	29.25	21.69	18.77	17.20	16.21	15.52	15.02	14.63	13.71	12.73	11.70

- **F - TEST:** **Objective:** Compare the population variance for two **Normal distribution** independent samples

Assumptions

F.TEST

$X_1, \cdots, X_n \sim Normal(\mu_1, \sigma_1^2)$ $\quad Y_1, \cdots, Y_m \sim Normal(\mu_2, \sigma_2^2)$

$$\begin{cases} H_0: \sigma_1^2 = \sigma_2^2 \\ H_1: \sigma_1^2 \neq \sigma_2^2 \end{cases}$$

Reject H_0 if $f_0 > F_{n-1,m-1,1-\alpha/2}$ or $f_0 < F_{n-1,m-1,\alpha/2}$ then H_1 is accept

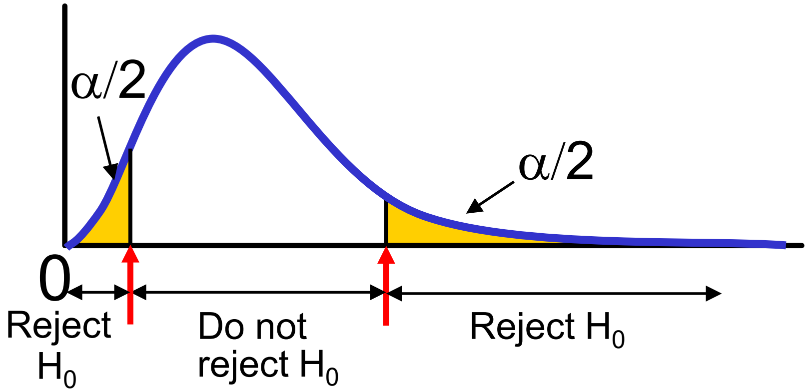
$$\frac{(n-1)S_1^2}{\sigma_1^2} \sim \chi^2_{n-1}$$

$$\frac{(m-1)S_2^2}{\sigma_2^2} \sim \chi^2_{m-1}$$

$$\Rightarrow \quad \frac{S_1^2/\sigma_1^2}{S_2^2/\sigma_2^2} \sim F(n-1, m-1)$$

Assuming the null hypothesis is true, we have $\sigma_1^2 = \sigma_2^2$

Test statistics: $\quad f_0 = \frac{s_1^2}{s_2^2}$



P-VALUE

- **F - TEST:** **Objective:** Compare the population variance for two **Normal distribution** independent samples

F.TEST

Assumptions

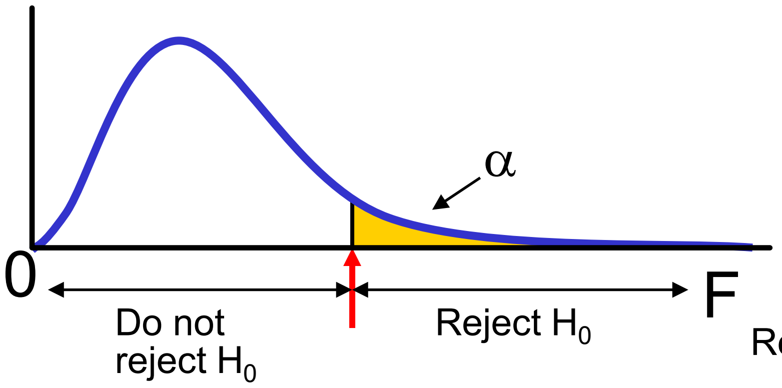
$X_1, \cdots, X_n \sim Normal(\mu_1, \sigma_1^2)$ $Y_1, \cdots, Y_m \sim Normal(\mu_2, \sigma_2^2)$

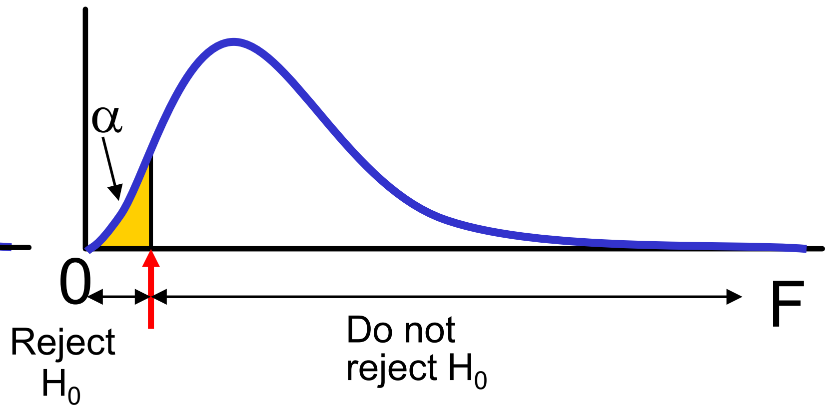
$\begin{cases} H_0: \sigma_1^2 = \sigma_2^2 \\ H_1: \sigma_1^2 > \sigma_2^2 \end{cases}$ Reject H_0 if $f_0 > F_{n-1,m-1,1-\alpha}$, then H_1 is accept

$\begin{cases} H_0: \sigma_1^2 = \sigma_2^2 \\ H_1: \sigma_1^2 < \sigma_2^2 \end{cases}$ Reject H_0 if $\boxed{f_0 < F_{n-1,m-1,\alpha},}$ then H_1 is accept

Assuming the null hypothesis is true, we have $\sigma_1^2 = \sigma_2^2$

Test statistics: $f_0 = \dfrac{s_1^2}{s_2^2}$





- **Chi-square test**

Contingency Tables

=CHISQ.TEST(

CHISQ.TEST(**actual_range**, expected_range)

- Useful in situations involving multiple population proportions
- Used to classify sample observations according to two or more characteristics

Sample results organized in a contingency table:

Hand Preference	Gender: Female	Gender: Male	
Left	12	24	36
Right	108	156	264
	120	180	300

120 Females, 12 were left handed

180 Males, 24 were left handed

Hand Preference vs. **Gender**

Dominant Hand: Left vs. Right

Gender: Male vs. Female

2 categories for each variable, so the table is called a 2 x 2 table

300 college students, sample size $= n = 300$

- **Chi-square test**

=CHISQ.TEST(

CHISQ.TEST(**actual_range**, expected_range)

$$H_0: p_{ij} = p_{i\cdot}p_{\cdot j}$$
$$for\ all\ i,j$$
$$H_1: Exist\ p_{ij} \neq p_{i\cdot}p_{\cdot j}$$

Under H_0: Hand preference is independent of gender/ Gender will not affect the Hand Preference (Not relevant)

Under H_1: Hand preference and gender are dependent

Hand Preference	Gender		
	Female	Male	
Left	12, $\boxed{p_{11}}$	24, $\boxed{p_{12}}$	36, $\boxed{p_{1\cdot}}$
Right	108, p_{21}	156, p_{22}	264, $p_{2\cdot}$
	120, $\boxed{p_{\cdot 1}}$	180, $\boxed{p_{\cdot 2}}$	300

Pr(Left preference Femal in Human)

Pr(Left preference Male in Human)

Pr(Left preference in Human)

$$p_{1\cdot} = \sum_{j=1}^{2} p_{1j} = p_{11} + p_{12}$$

Pr(Femal in Human) Pr(Male in Human)

Under H_0

Pr(Left preference in Female) $\longleftrightarrow$ $p_{11}/p_{\cdot 1}$

= Pr(Left preference in Male) $\longleftrightarrow$ $= p_{12}/p_{\cdot 2}$

= Pr(Left preference in Human) $\longleftrightarrow$ $= p_{1\cdot}$

- **Chi-square test**

=CHISQ.TEST(

CHISQ.TEST(**actual_range**, expected_range)

$$H_0: p_{ij} = p_{i\cdot}p_{\cdot j}$$
$$for\ all\ i,j$$
$$H_1: Exist\ p_{ij} \neq p_{i\cdot}p_{\cdot j}$$

Under H_0: Hand preference is independent of gender/ Gender will not affect the Hand Preference (Not relevant)

Under H_1: Hand preference and gender are dependent

Hand Preference	Gender		
	Female	Male	
Left	12, p_{11}	24, p_{12}	36, $p_{1\cdot}$
Right	108, p_{21}	156, p_{22}	264, $p_{2\cdot}$
	120, $p_{\cdot 1}$	180, $p_{\cdot 2}$	300

The expect proportion Under H_0

$$\widehat{p_{1\cdot}} = \frac{36}{300} \quad \widehat{p_{2\cdot}} = \frac{264}{300} \quad \widehat{p_{\cdot 1}} = \frac{120}{300} \quad \widehat{p_{\cdot 2}} = \frac{180}{300}$$

$$\widehat{p_{11}} = \widehat{p_{1\cdot}}\widehat{p_{\cdot 1}} \qquad \widehat{p_{12}} = \widehat{p_{1\cdot}}\widehat{p_{\cdot 2}}$$

$$\widehat{p_{21}} = \widehat{p_{2\cdot}}\widehat{p_{\cdot 1}} \qquad \widehat{p_{22}} = \widehat{p_{2\cdot}}\widehat{p_{\cdot 2}}$$

- **Chi-square test**

$$H_0: p_{ij} = p_{i\cdot}p_{\cdot j} \quad for\ all\ i,j$$

$$H_1: Exist\ p_{ij} \neq p_{i\cdot}p_{\cdot j}$$

Under H_0: Hand preference is independent of gender/ Gender will not affect the Hand Preference (Not relevant)

Under H_1: Hand preference and gender are dependent

Hand Preference	Gender: Female	Gender: Male	
Left	12,p_{11}	24,p_{12}	36,$p_{1\cdot}$
Right	108,p_{21}	156,p_{22}	264,$p_{2\cdot}$
	120,$p_{\cdot 1}$	180,$p_{\cdot 2}$	300

Under H_0 → **The expect frequency**

Hand Preference	Gender: Female	Gender: Male	
Left	$12\ vs\ 14.4$	$24\ vs\ 21.6$	36,$\widehat{p_{1\cdot}}$
Right	$108\ vs\ 105.6$	$156\ vs\ 158.4$	264,$\widehat{p_{2\cdot}}$
	120,$\widehat{p_{\cdot 1}}$	180,$\widehat{p_{\cdot 2}}$	300

- **Chi-square test**

=CHISQ.TEST(

CHISQ.TEST(**actual_range**, expected_range)

$$H_0: p_{ij} = p_{i\cdot}p_{\cdot j}$$
$$for\ all\ i,j$$
$$H_1: Exist\ p_{ij} \neq p_{i\cdot}p_{\cdot j}$$

Under H_0: Hand preference is independent of gender/ Gender will not affect the Hand Preference (Not relevant)

Under H_1: Hand preference and gender are dependent

Test statistics:

$$q = \sum_{j=1}^{2}\sum_{i=1}^{2}\frac{\left(x_{ji} - n\widehat{p_{i\cdot}}\widehat{p_{\cdot j}}\right)^2}{n\widehat{p_{i\cdot}}\widehat{p_{\cdot j}}}$$

The expect frequency

Hand Preference	Gender		
	Female	Male	
Left	$12\ vs\ 14.4$	$24\ vs\ 21.6$	$36, \widehat{p_{1\cdot}}$
Right	$108\ vs\ 105.6$	$156\ vs\ 158.4$	$264, \widehat{p_{2\cdot}}$
	$120, \widehat{p_{\cdot 1}}$	$180, \widehat{p_{\cdot 2}}$	300

(q is large if there is a big different between observed data and the expected frequency)

[H_0 is wrong]

- **Chi-square test**

=CHISQ.TEST(

CHISQ.TEST(**actual_range**, expected_range)

$H_0: p_{ij} = p_{i\cdot}p_{\cdot j}$ $for\ all\ i,j$

$H_1: Exist\ p_{ij} \neq p_{i\cdot}p_{\cdot j}$

Under H_0: Hand preference is independent of gender/ Gender will not affect the Hand Preference (Not relevant)

Under H_1: Hand preference and gender are dependent

$$Q = \sum_{j=1}^{2}\sum_{i=1}^{2}\frac{(X_{ij} - n\widehat{P_{i\cdot}}\widehat{P_{\cdot j}})^2}{n\widehat{P_{i\cdot}}\widehat{P_{\cdot j}}} \sim \boxed{\chi_U^2}$$

$$\widehat{P_{1\cdot}} = \frac{X_{11} + X_{12}}{n} \qquad \widehat{P_{2\cdot}} = \frac{X_{21} + X_{22}}{n} = 1 - \widehat{P_{1\cdot}}$$

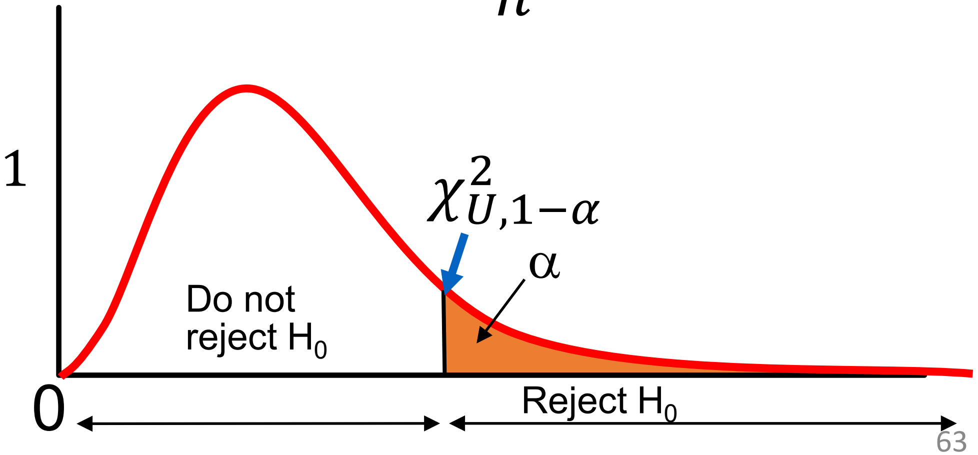
$$\widehat{P_{\cdot 1}} = \frac{X_{11} + X_{21}}{n} \qquad \widehat{P_{\cdot 2}} = \frac{X_{12} + X_{22}}{n} = 1 - \widehat{P_{\cdot 1}}$$

2 Estimated parameter: minus 2 degree of freedom

Degree of freedom $U = (2 * 2 - 1) - \boxed{1} - \boxed{1} = (2 - 1)(2 - 1) = 1$

Three parameters need to be estimated p_{11}, p_{12}, p_{21}

Reject H_0 if test statistic $q > \chi^2_{1,1-\alpha}$, then H_1 is accept



Lecture 9: Regression

- **Correlation** is a statistical method used to determine whether a linear relationship between variables exists.
- **Regression** is a statistical method used to describe the nature of the relationship between variables—that is, positive or negative, linear or nonlinear.

- **Covariance**

COVARIANCE.P

COVARIANCE.S

$$Cov(x,y) = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})$$

Sample Covariance

$$Cov(x,y) = \frac{1}{N}\sum_{i=1}^{n}(x_i - \mu_x)(y_i - \mu_y)$$

Population Covariance

$Cov(x,y) > 0$ X and Y tend to move in the **same** direction

$Cov(x,y) < 0$ X and Y tend to move in the **opposite** direction

$Cov(x,y) = 0$ X and Y have **NO** linear relationship

• Coefficient of Correlation

=CORREL(
CORREL(**array1**, array2)

=PEARSON(
PEARSON(**array1**, array2)

To measures the <u>relative</u> strength of the linear relationship **between two variables**

$$r = \frac{Cov(x,y)}{s_x s_y} = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2}\sqrt{\sum_{i=1}^{n}(y_i - \bar{y})^2}}$$

r^2 RSQ

- ✓ Unit free
- ✓ Ranges between <u>–1 and 1</u>
- ✓ The closer to **<u>–1</u>**, the stronger the ***<u>negative linear relationship</u>***
- ✓ The closer to **<u>1</u>**, the stronger the ***<u>positive linear relationship</u>***
- ✓ The closer to **0**, **the weaker any positive linear relationship**

- **Regression**

$$Y = a + bX$$

INTERCEPT $\rightarrow$ $a = \bar{Y} - b\bar{X}$

SLOPE $\rightarrow$ $b = \dfrac{\sum(X_i - \bar{X})(Y_i - \bar{Y})}{\sum(X_i - \bar{X})^2}$

STEYX

Residual standard error:
Standard error of the predicted y-value for each x in the regression

$$\frac{\sum_{i=1}^{N}\left(\widehat{Y}_i - Y_i\right)^2}{N-2}$$

Simple Regression

LINEST

$$Y = b + m_1X_1 + m_2X_2 + \cdots + m_nX_n$$

	A	B	C	D	E	F
1	m_n	m_{n-1}	. . .	m_2	m_1	b
2	se_n	se_{n-1}	. . .	se_2	se_1	se_b
3	r_2	se_y				
4	F	d_f				
5	ss_{reg}	ss_{resid}				

• Regression

An F-test in regression compares the fits of different linear models.

$$e_i = Y_i - \hat{Y}_i = Y_i - (b_0 + b_1 X_{i1} + \cdots + b_{p-1} X_{i,p-1}),\ i = 1,\ldots,n,$$

$$s^2 = \frac{1}{n-p}\sum_{i=1}^{n} e_i^2 = \frac{1}{n-p}\sum_{i=1}^{n}(Y_i - \hat{Y}_i)^2 = \frac{SSE}{n-p} = MSE.$$

$$H_0:\ \beta_1 = \beta_2 = \cdots = \beta_{p\text{-}1} = 0, \text{ versus}$$

$$H_A:\ \text{At least one } \beta_j \neq 0, \text{ for } j = 1,\ldots,p-1.$$

Source	Degrees of Freedom	Sum of Squares	Mean Squares	*F*-statistic
Regression	$p-1$	SSR	MSR	MSR/MSE
Error	$n-p$	SSE	MSE	
Total	$n-1$	$SSTO$		

$\boxed{\text{LINEST}}$

$$Y = b + m_1 X_1 + m_2 X_2 + \cdots + m_n X_n$$

	A	B	C	D	E	F
1	m_n	m_{n-1}	...	m_2	m_1	b
2	se_n	se_{n-1}	...	se_2	se_1	se_b
3	r_2	se_y				
4	F	d_f				
5	ss_{reg}	ss_{resid}				

- **Exponential Regression/Linear exponential forecasting**

GROWTH

FORECAST.LINEAR TREND $\widehat{Y}_i$

$$Y = b * m^X \qquad \log(Y) = \log(b) + \log(m)\, X$$

$$Y' = b' + m'X$$

LOGEST

	A	B	C	D	E	F
1	m_n	m_{n-1}	...	m_2	m_1	b
2	se_n	se_{n-1}	...	se_2	se_1	se_b
3	r_2	se_y				
4	F	d_f				
5	ss_{reg}	ss_{resid}				

- **Fisher transformation:** Used to model the length of time between two occurrences of an event

FISHER

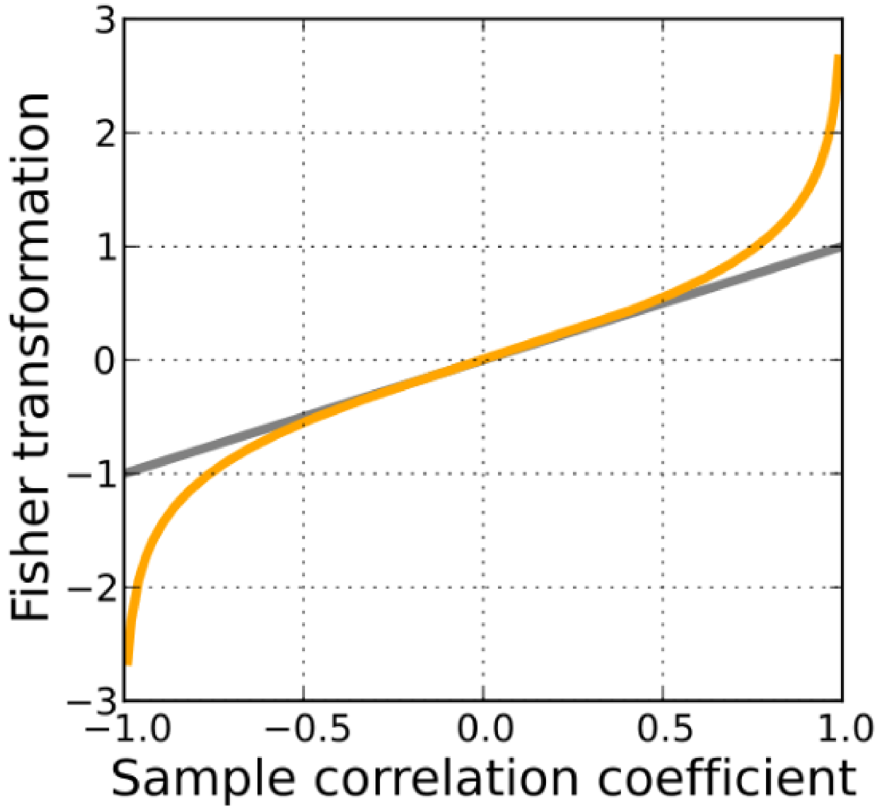
FISHERINV

$$z = \frac{1}{2}\ln\left(\frac{1+r}{1-r}\right) \qquad r = \frac{\exp(2z)-1}{\exp(2z)+1} = \tanh(z)$$

If (X, Y) has a bivariate normal distribution with correlation ρ and the pairs (X_i, Y_i) are independent and identically distributed.

$$z = \frac{1}{2}\ln\left(\frac{1+r}{1-r}\right) \text{ follows } N\left(\frac{1}{2}\ln\left(\frac{1+\rho}{1-\rho}\right) \quad \frac{1}{\sqrt{N-3}} \right)$$

It can be used to construct a large-sample confidence interval for r using standard normal theory and derivations.



- **Exponential Triple Smoothing (ETS)**

Forecasting Section

Time series

FORECAST.ETS

FORECAST.ETS.CONFINT

FORECAST.ETS.SEASONALITY

FORECAST.ETS.STAT

Exponential Smoothing DOUBLE

$$s_t = \alpha y_t + (1-\alpha)(s_{t-1} + b_{t-1})$$

$$b_t = \beta(s_t - s_{t-1}) + (1-\beta)b_{t-1}$$

trend smoothing factor

$$\widehat{y_{t+1}} = s_t + b_t$$

Moving Average

$$\widehat{y_{t+1}} = \frac{1}{k}(y_t + y_{t-1} + \cdots + y_{t-k-1})$$

$$\widehat{y_{t+1}} = \frac{1}{t}(y_t + y_{t-1} + \cdots + y_1)$$

Exponential Smoothing

data smoothing factor

$$\widehat{y_{t+1}} = \alpha y_t + (1-\alpha)\hat{y}_t$$

$$= \alpha y_t + \alpha(1-\alpha)y_{t-1} + \cdots$$

$$= \alpha[y_t + (1-\alpha)y_{t-1} + (1-\alpha)^2 y_{t-2} + \cdots$$

Exponential Smoothing TRIPLE

$$s_t = \alpha(y_t - c_{t-L}) + (1-\alpha)(s_{t-1} + b_{t-1})$$

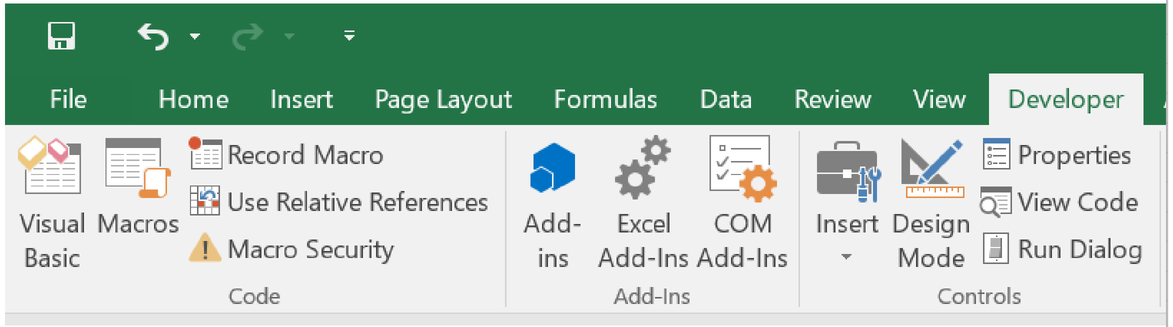
a cycle of seasonal change of length

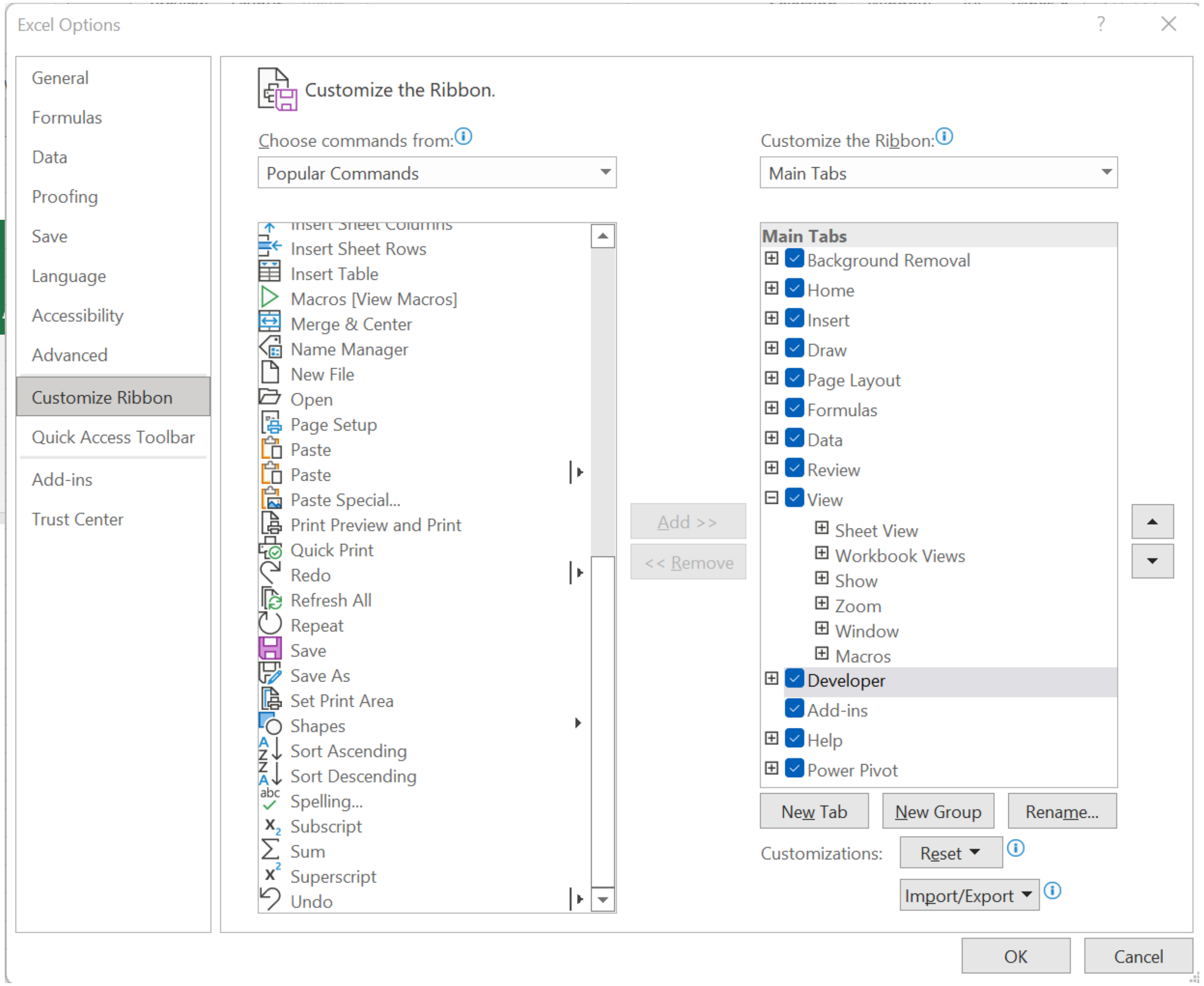
$$c_t = \gamma(y_t - s_{t-1} - b_{t-1}) + (1-\gamma)c_{t-L}$$

seasonal change smoothing factor

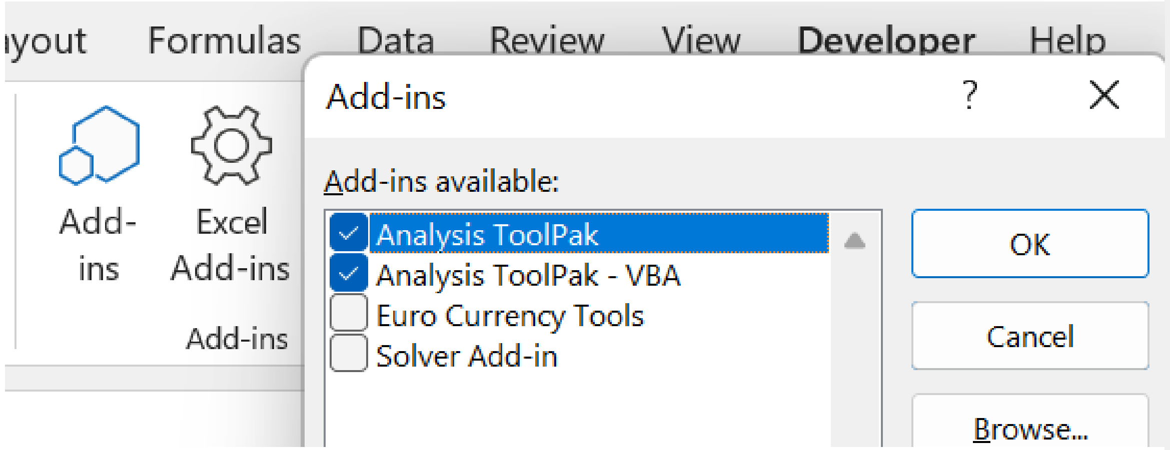
$$\widehat{y_{t+1}} = s_t + b_t + c_{t-L+1}$$

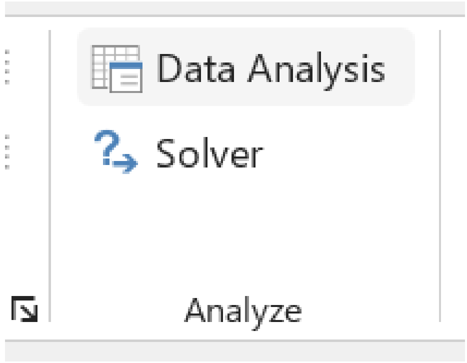
Visual Basic for Applications [VBA]

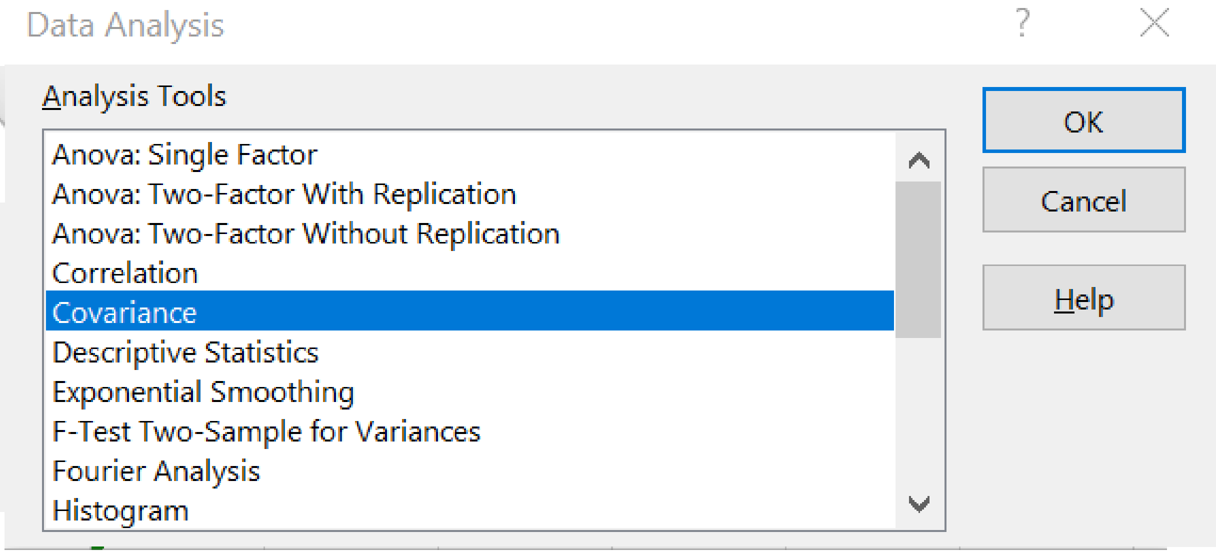


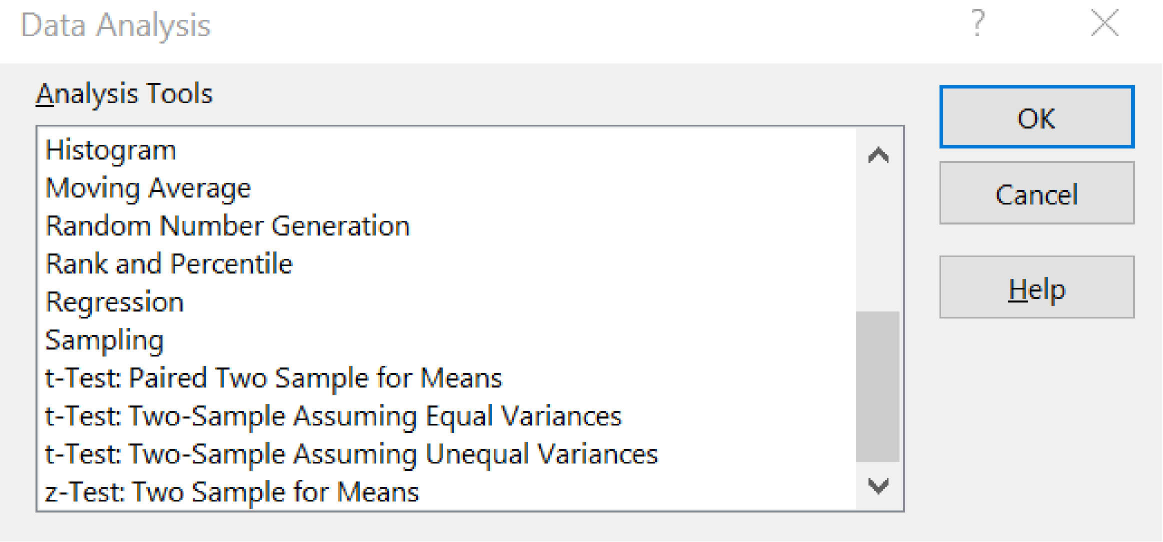


Add-in Macros



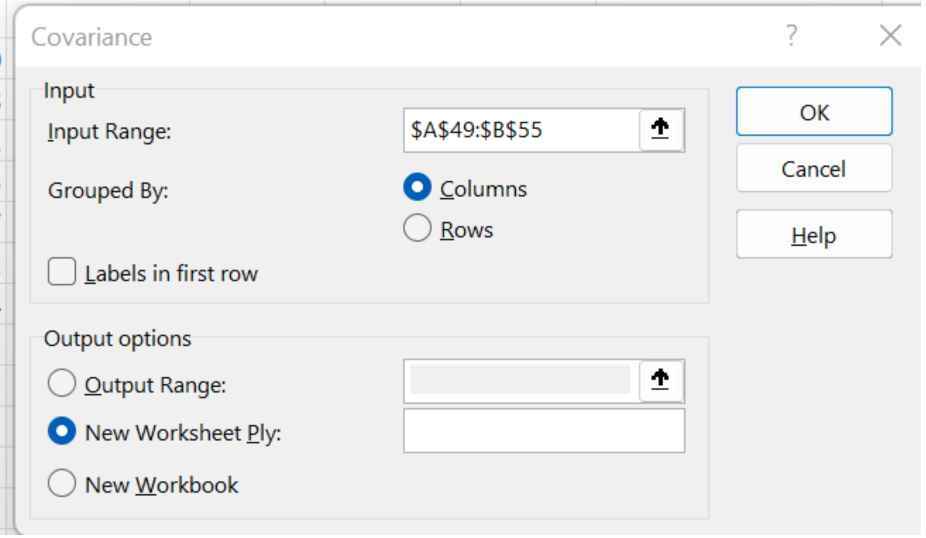






Add-in Macros

48		
49	134	70
50	146	118
51	104	101
52	119	85
53	124	107
54	161	132
55	107	94
56	83	
57	113	
58	129	
59	97	
60	123	

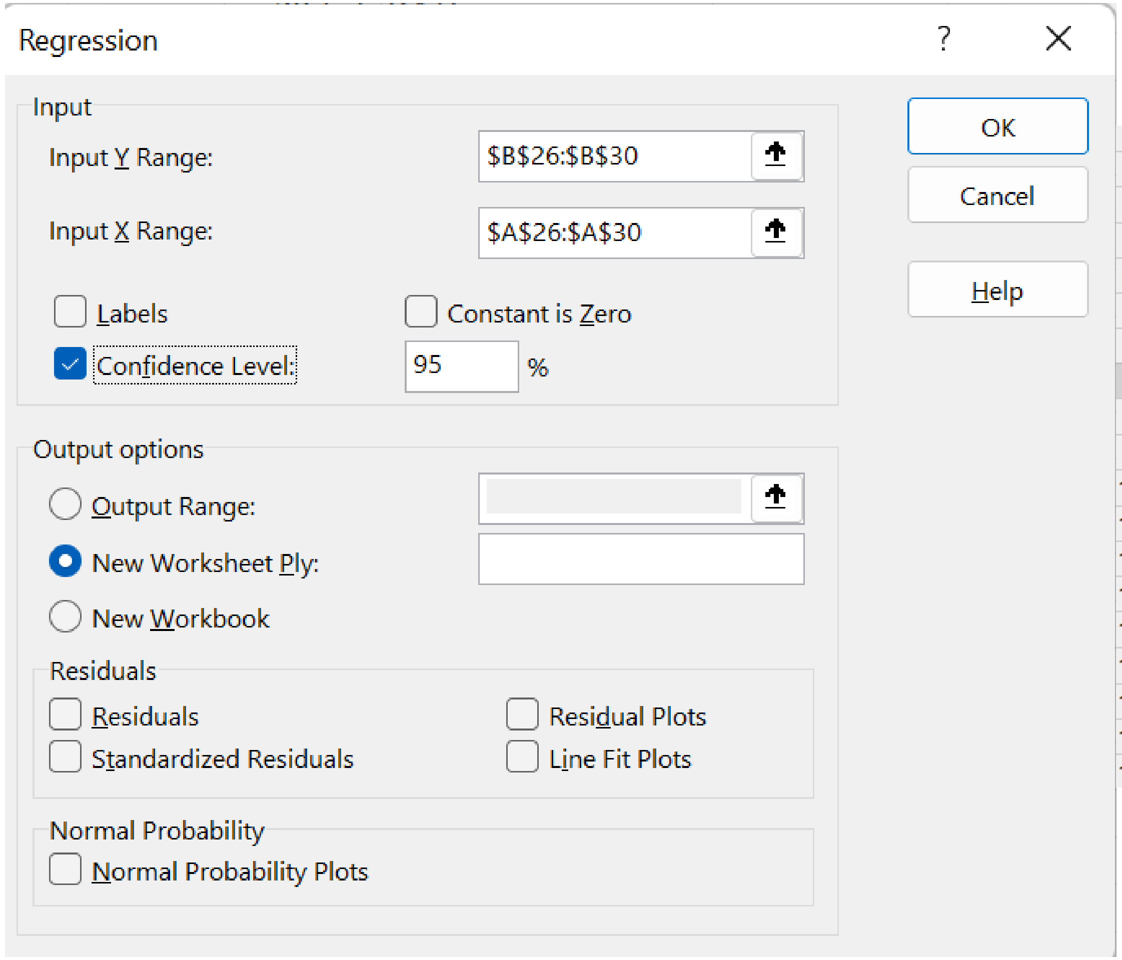


	A	B	C
1		*Column 1*	*Column 2*
2	Column 1	366.1224	
3	Column 2	201.4286	364.5714

C49 =COVARIANCE.P($A$49:$A$55,$B$49:$B$55)

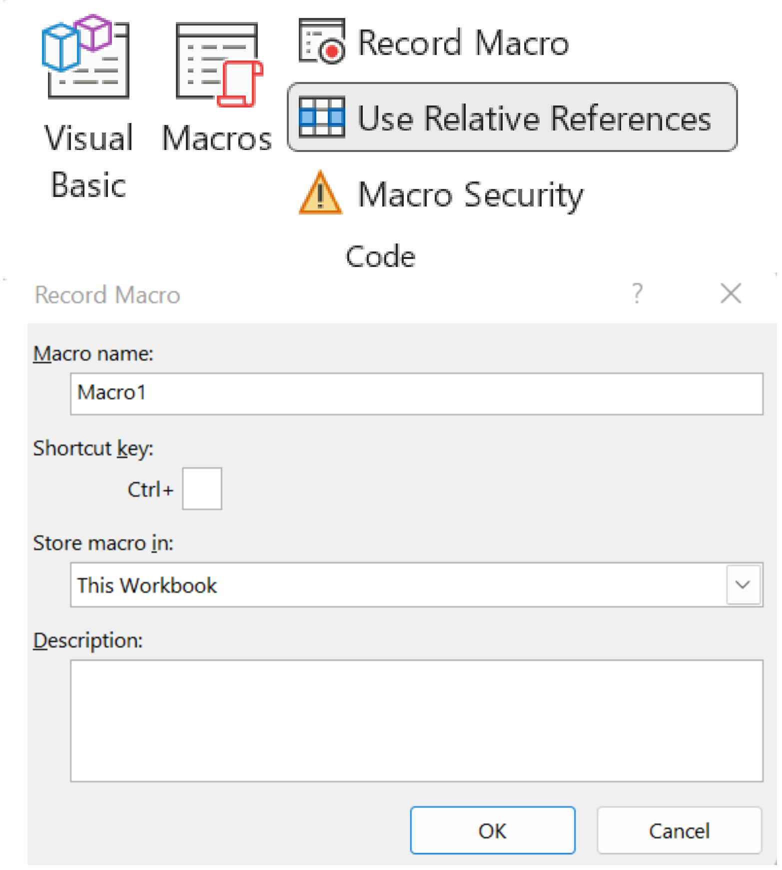
	A	B	C	D	E	F
49	134	70	201.4285714			
50	146	118				
51	104	101				
52	119	85				
53	124	107				
54	161	132				
55	107	94				

Add-in Macros



	A	B	C	D	E	F	G	H	I
1	SUMMARY OUTPUT								
2									
3	*Regression Statistics*								
4	Multiple R	0.904808094							
5	R Square	0.818677686							
6	Adjusted F	0.758236915							
7	Standard I	3.95806113							
8	Observatio	5							
9									
10	ANOVA								
11		*df*	*SS*	*MS*	*F*	*gnificance F*			
12	Regressior	1	212.2012563	212.2013	13.54512	0.034748			
13	Residual	3	46.99874372	15.66625					
14	Total	4	259.2						
15									
16		*Coefficients*	*Standard Error*	*t Stat*	*P-value*	*Lower 95%*	*Upper 95%*	*ower 95.0%*	*Upper 95.0%*
17	Intercept	18.00753769	4.046575303	4.450069	0.021114	5.129529	30.88555	5.129529	30.88554631
18	X Variable	1.154522613	0.313697434	3.68037	0.034748	0.156197	2.152848	0.156197	2.152847851

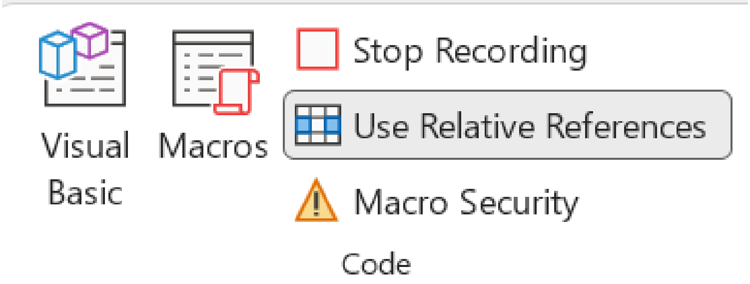
Record Macros

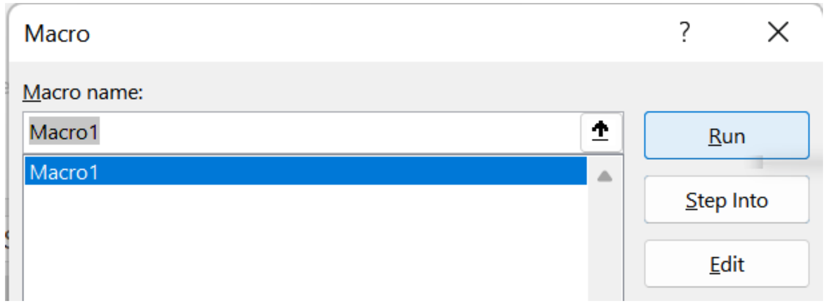


	A	B	C	D
1	Students ID	Course	Code	Final Grade
2	1155159540	STAT	5101	A
3	1155159540	STAT	5101	A
4	1155160008	STAT	5101	A
5	1155157370	STAT	5101	A
6	1155158100	STAT	5101	A
7	1155159534	STAT	5101	A
8	1155157369	STAT	5101	A-
9	1155159305	STAT	5102	A-
10	1155159305	STAT	5101	A-
11	1155159305	STAT	5101	A-
12	1155158619	STAT	5101	A-
13	1155159994	STAT	5101	A-
14	1155159290	STAT	5101	A-

	A	B	C	D
1	Students ID	Course	Code	Final Grade
2	1155159540	STAT	5101	A
3				
4	Students ID	Course	Code	Final Grade
5	1155159540	STAT	5101	A
6				
7	Students ID	Course	Code	Final Grade
8	1155160008	STAT	5101	A
9				
10	Students ID	Course	Code	Final Grade
11	1155157370	STAT	5101	A
12				
13	Students ID	Course	Code	Final Grade
14	1155158100	STAT	5101	A

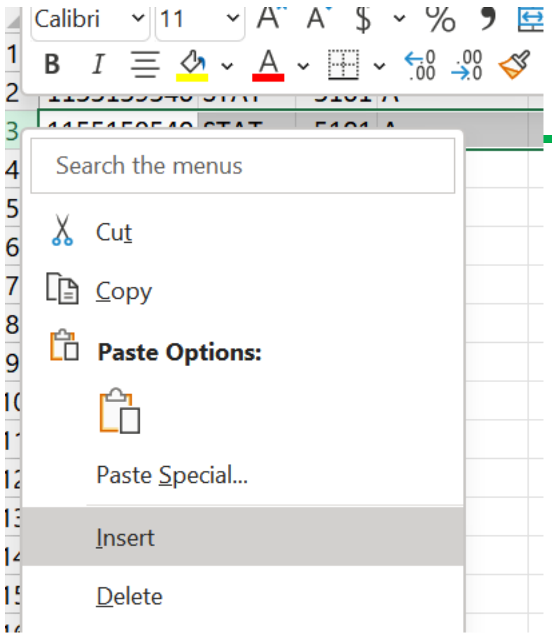
Record Macros



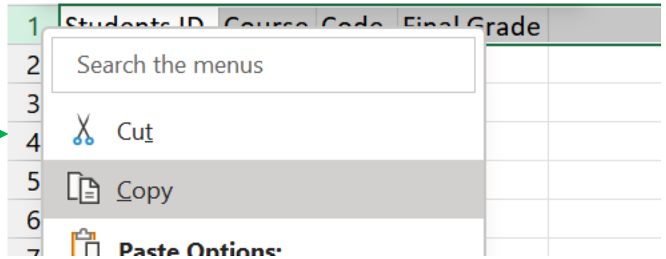


	A	B	C	D
1	Students ID	Course	Code	Final Grade
2	1155159540	STAT	5101	A
3	1155159540	STAT	5101	A
4	1155160008	STAT	5101	A
5	1155157370	STAT	5101	A
6	1155158100	STAT	5101	A

Select Line 1



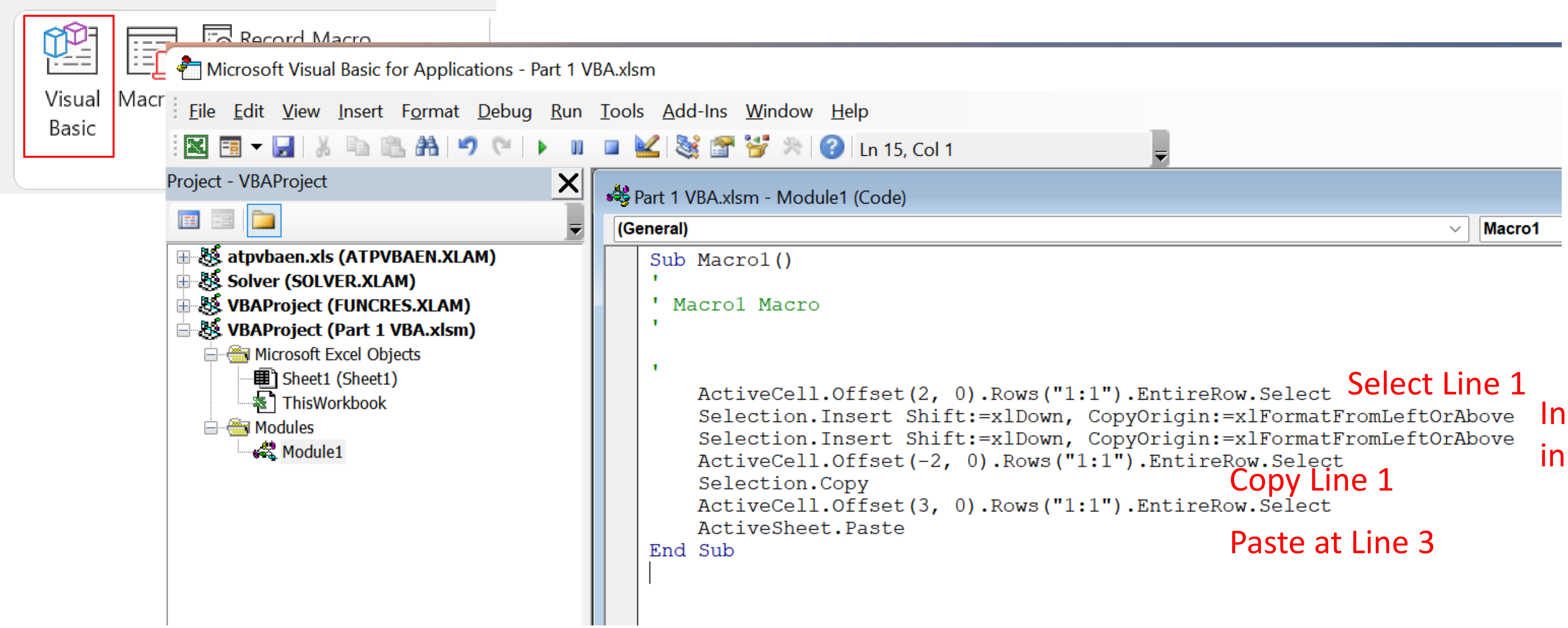
Insert Twice in Line 3



Copy Line 1 and paste at Line 3

	A	B	C	D
1	Students ID	Course	Code	Final Grade
2	1155159540	STAT	5101	A
3				
4	Students ID	Course	Code	Final Grade
5	1155159540	STAT	5101	A

Record Macros



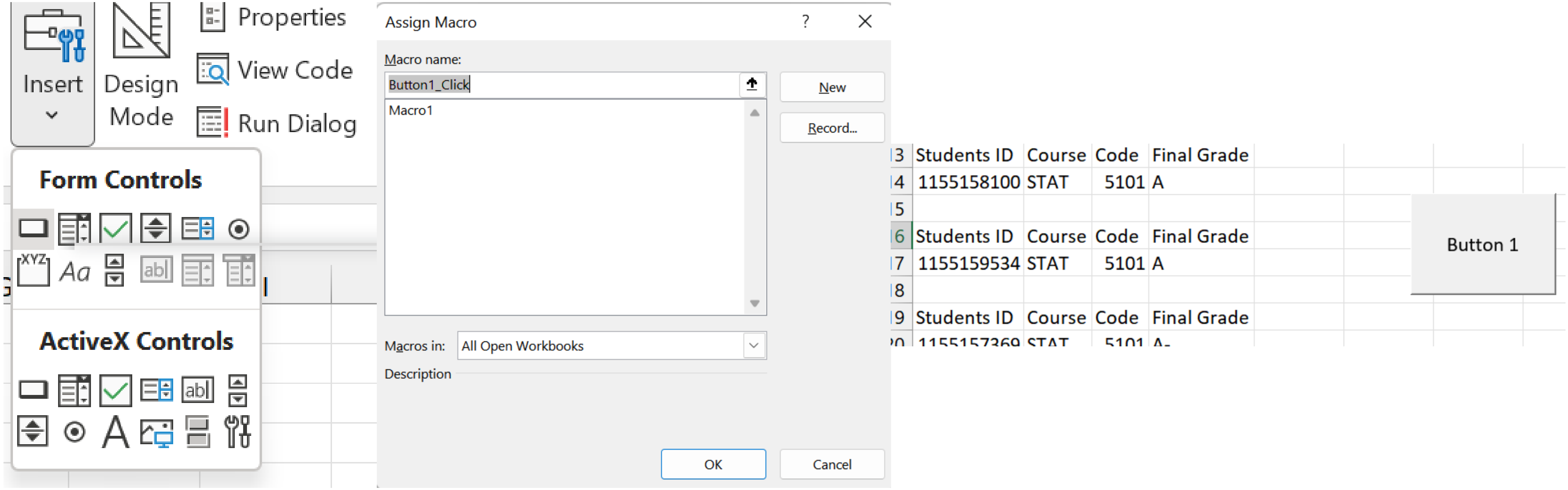
Select Line 1

Insert Twice in Line 3

Copy Line 1

Paste at Line 3

Record Macros



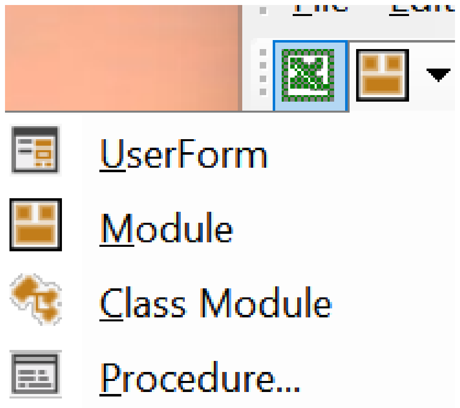
Record Macros

Part 1 VBA.xlsm - Module1 (Code)

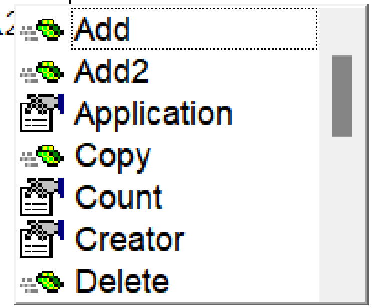
(General) Macro1

```
Sub Macro1()
'
' Macro1 Macro
'

'
    Dim i As Long
    For i = 2 To 13
    ActiveCell.Offset(2, 0).Rows("1:1").EntireRow.Select
    Selection.Insert Shift:=xlDown, CopyOrigin:=xlFormatFromLeftOrAbove
    Selection.Insert Shift:=xlDown, CopyOrigin:=xlFormatFromLeftOrAbove
    ActiveCell.Offset(-2, 0).Rows("1:1").EntireRow.Select
    Selection.Copy
    ActiveCell.Offset(3, 0).Rows("1:1").EntireRow.Select
    ActiveSheet.Paste
    Next i

End Sub
```

Macros




```
Sub Macro2()

Application.ActiveWorkbook.Worksheets("Sheet4").Rows("3:3").Select
Range("A1").Value = 100
Worksheets.
Range("A2

End Sub
```



Object model

```
Sub Macro2()

Application.ActiveWorkbook.Worksheets("Sheet4").Rows("3:3").Select
Range("A1").Value = 100
Range("A2").

End Sub
```

properties

methods

https://docs.microsoft.com/en-us/office/vba/api/overview/excel/object-model

Macros—control Workbook and worksheet

Part 1 VBA.xlsm - Module2 (Code)

(General)

```
Sub Macro2()

Application.ActiveWorkbook.Worksheets("Sheet4").Rows("3:3").Select
Worksheets("Sheet4").Range("A1").Value = 100
Worksheets.Add(after:=Worksheets(4)).Name = "1"
Range("A2").Value = 200
ActiveSheet.Name = "2"
Worksheets("2").Delete
Worksheets("Sheet4").Copy before:=Worksheets("Sheet1")
Worksheets(1).Move after:=Worksheets("Sheet4")
Worksheets("Sheet4").Activate

MsgBox "Hello"
MsgBox "Worksheet number " & Worksheets.Count

Workbooks.Open ("D:\Book1.xlsx")
'Workbooks.Open Filename:="D:\Book1.xlsx"
Workbooks("Book1").Close SaveChanges:=False

End Sub
```

Macros—control range and cells in worksheet

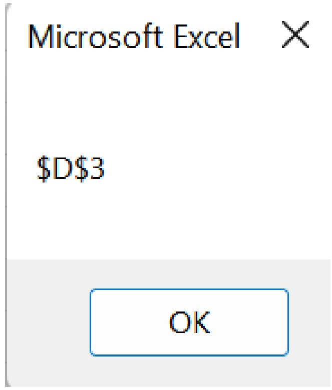
```
Sub ran()

Worksheets("Sheet4").Activate
Range("A2:A10").Value = 200
Dim n As String
n = "B1:B10"
Range(n).Value = 20

Range("C1:C4 C3:D5").Value = 11
Range("C1:C4,C3:D5").Value = 12
Range("C1:C4", "C3:D5").Value = 15
Range("D7:F8").Offset(-5, -2).Value = 1

Cells(2, 5).Value = 20
Cells(2, "F").Value = 25
Cells(2, 3).Select

Rows("3:5").Select
Rows("3").Select
Columns("D").Select
Columns(6).Select
```

Different meaning

2

5

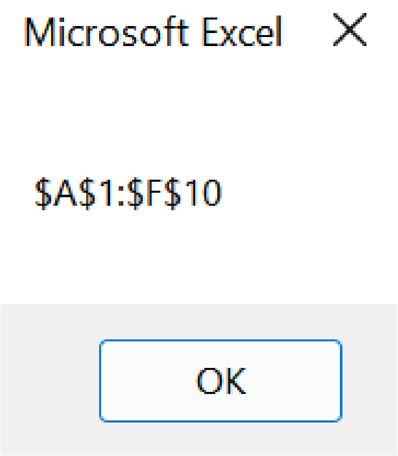
Macros—control range and cells in worksheet




```
Rows("3").End(xlToRight).Select
MsgBox ActiveCell.Address
Columns(1).End(xlDown).Value = 5
ActiveSheet.UsedRange.Select
MsgBox ActiveSheet.UsedRange.Address
Range("A2").CurrentRegion.Select
```

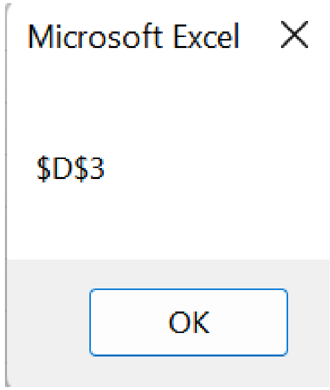
Rows three End + →

Last one of the Colum



Crtl + A

Macros—control range and cells in worksheet



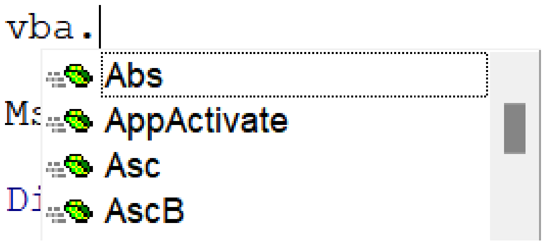
Different Properties

Rows three

9:36:00 AM	0.4 -> time	`Range("H5").Value = "=0.1+0.3" 'TIME`
9:36:00 am	Only text	`Range("H6").Value = Range("H5").Text`
0.4	Formula	`Range("H7").Value = Range("H5").Formula`
0.4	Value	`Range("H8").Value = Range("H5").Value`

Macros—copy and paste

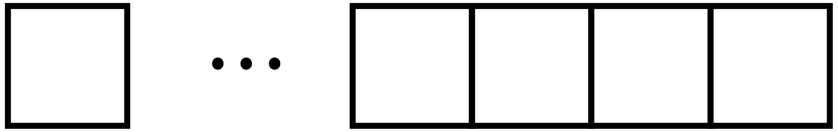
```
Range("A1").Copy Destination:=Range("B11")
Range("A1:C11").Copy Range("A12")
Range("C1:D5").Copy
Range("I1").PasteSpecial Paste:=xlPasteValues
Range("K1").PasteSpecial
Range("A10").Cut Range("B10")
Range("A12").Delete
Range("A11").Value = Application.WorksheetFunction.Sum(Range("A1:A10"))
```



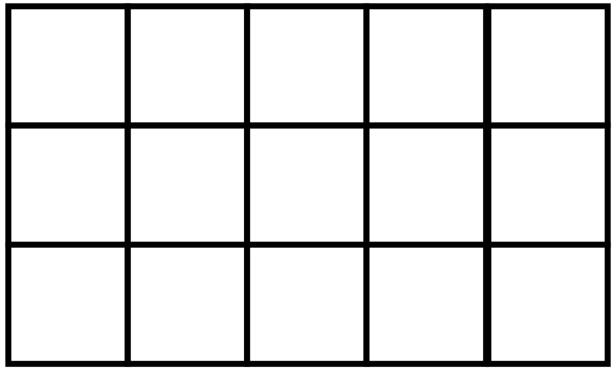
Applied Build in function in Macros

[Visual Basic for Applications functions | Microsoft Docs]

Macros



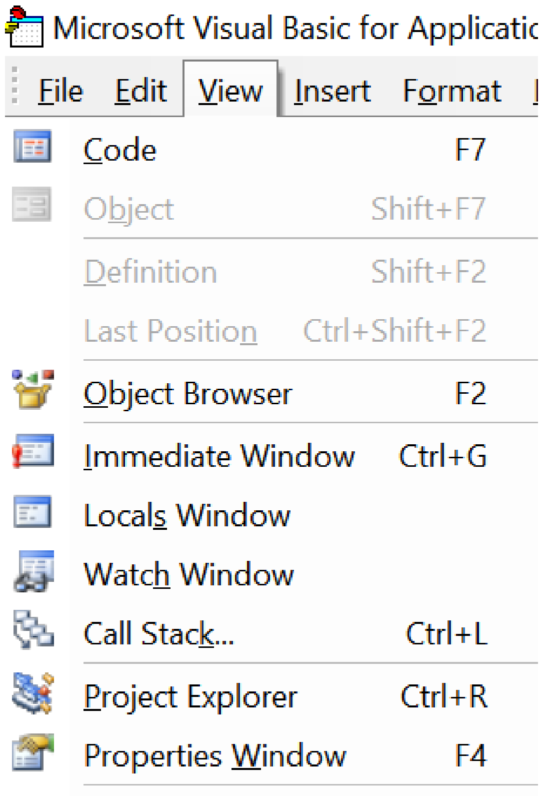
```
Dim arry(1 To 100) As Double
arry(1) = 0.01
MsgBox arry(1)
```



```
Dim arrd(1 To 3, 1 To 5) As Integer
arrd(1, 1) = Range("A2").Value
Range("A3").Value = arrd(1, 1)
```

```
Dim AR() As Integer
Dim num As Integer
Dim i As Integer
num = Range("B1").CurrentRegion.Rows.Count
'Dim AR(1 To num) As Integer 'Constant Required
ReDim AR(1 To num)
For i = 1 To num
AR(i) = Cells(i, "B").Value
Next i
ReDim Preserve AR(1 To num + 1)
AR(num + 1) = 588
Dim stn As String
stn = "M1:M" & (num + 1)
Range(stn).Value = WorksheetFunction.Transpose(AR)
```

Macros



Byte	Integer: 0-255
Boolean	True or False
Integer	Integer: -2768~32767
Long	Integer: -+2147483648
Single	Single-precision floating-point format
Double	Double-precision floating-point format
String	
Date	
Variant	

Watches

Expression	Value	Type
AR		Integer(1 to 23)
ak	986	Variant/Integer
num	22	Integer
stn	"M1:M23"	String

Macros——Iteration

```
For <variable> = <initial> To <End> [STEP]
    <Loop body>
     [Exit For]
Next [variable]
```

```
Do [While <logic>]
   <Loop body>
Loop
```

```
Do
   <Loop body>
Loop [While <logic>]
```

```
Do [Until <logic>]
   <Loop body>
Loop
```

```
Do
   <Loop body>
Loop [Until <logic>]
```

Macros——Judgment

```
If <logic> Then
        <body>
ElseIf <logic> Then
        <body>
Else
        <body>
End If
```

```
Select Case <Variable>
Case Is <logic>
        <body>
Case Is <logic>
        <body>
Case Else
        <body>
End Select
```