

26 OCT 2023

Introduction to Probability Statistics & Machine Learning

Rishi Chandra

Chapter-1 Contents

1. Summarizing Data Sets : Histograms
2. Summarizing Data Sets : Mean
3. Summarizing Data Sets : Variance
4. Percentile
5. Quartiles and Box Plots
6. Normal Data Sets Histograms / Distributions & Outliers
7. Correlation = strength of an assumed linear relationship

References:

1. Pattern-Recognition-and-Machine-Learning - Christopher M. Bishop, available at <https://www.microsoft.com/en-us/research/uploads/prod/2006/01/Bishop-Pattern-Recognition-and-Machine-Learning-2006.pdf>
2. Introduction to Probability and statistics for engineers and scientists - Sheldon M. Ross, available at <https://minerva.it.manchester.ac.uk/~saralees/statbook3.pdf>

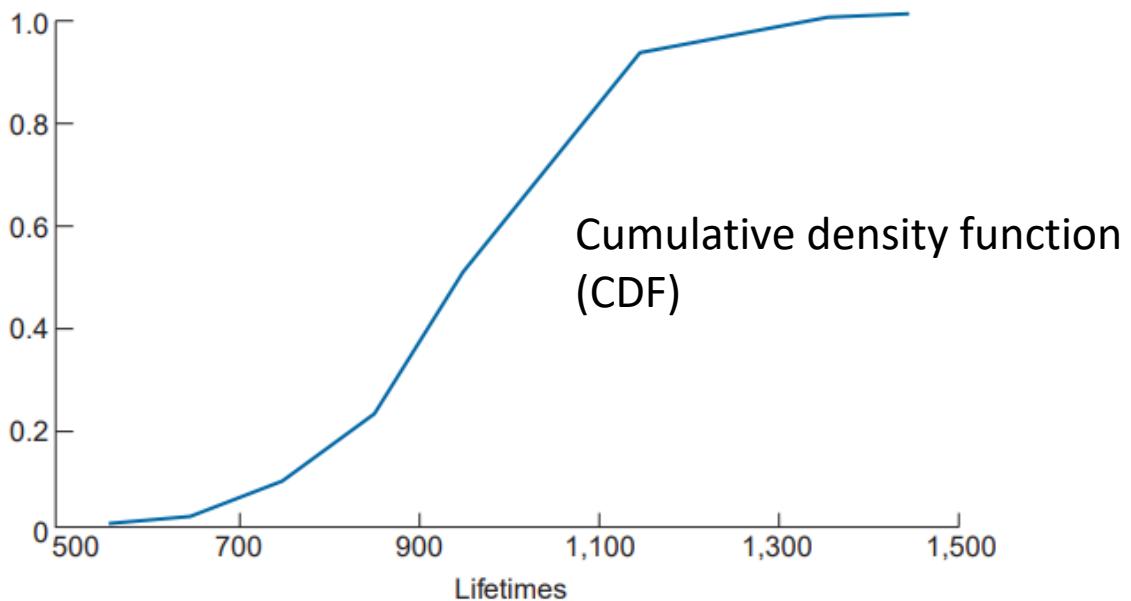
1.1 Summarizing Data Sets : Histograms

Relative frequency = Probability density (sums to 1)

For continuous values, bin into discrete buckets

Type of Cancer	Number of New Cases	Relative Frequency
Lung	42	.21
Breast	50	.25
Colon	32	.16
Prostate	55	.275
Melanoma	9	.045
Bladder	12	.06

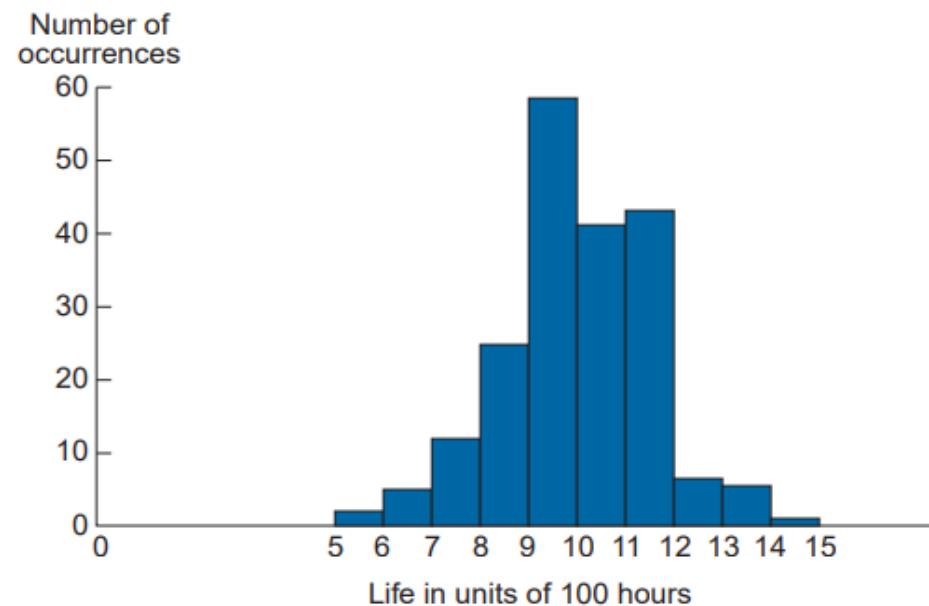
Probability
density function
(PDF)



A cumulative frequency plot.

TABLE 2.4 A Class Frequency Table

Class Interval	Frequency (Number of Data Values in the Interval)
500–600	2
600–700	5
700–800	12
800–900	25
900–1000	58
1000–1100	41
1100–1200	43
1200–1300	7
1300–1400	6
1400–1500	1



A frequency histogram.

1.2 Summarizing Data Sets : Mean

- Sample Mean, Sample Median, Sample Mode

$$\begin{aligned} \text{if } & Y = aX + b \\ \text{then } & \bar{Y} = a\bar{X} + b \end{aligned}$$

The *sample mean*, designated by \bar{x} , is defined by

Sample Mean $\bar{x} = \sum_{i=1}^n x_i/n$

$$\text{if } y_i = ax_i + b, \quad i = 1, \dots, n$$

then the sample mean of the data set y_1, \dots, y_n is

$$\bar{y} = \sum_{i=1}^n (ax_i + b)/n = \sum_{i=1}^n ax_i/n + \sum_{i=1}^n b/n = a\bar{x} + b$$

➤ Sample mean is NOT population mean !

e.g., Voter exit polls are only a sample mean intended to representation of the overall population mean, but is not the population mean

Median

8	15	50	200	1000
---	----	----	-----	------

Median

$$\text{Median} = (25+175)/2 = 100$$

1	3	25	175	880	9999
---	---	----	-----	-----	------

Order the values of a data set of size n from smallest to largest. If n is odd, the *sample median* is the value in position $(n+1)/2$; if n is even, it is the average of the values in positions $n/2$ and $n/2 + 1$.

Mode

- Most Frequent Value(s)

Another statistic that has been used to indicate the central tendency of a data set is the *sample mode*, defined to be the value that occurs with the greatest frequency. If no single value occurs most frequently, then all the values that occur at the highest frequency are called *modal values*.

1.3 Summarizing Data Sets : Variance

- Sample Variance and Standard Deviation, vs
- Population Variance and Std. Dev

The *sample variance*, call it s^2 , of the data set x_1, \dots, x_n is defined by

$$\text{Var} = (\text{Std. Dev})^2$$

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

- Like sample mean is not population mean,
Sample Variance != Population Variance
- Like sample mean is an estimate/approximation of the population mean, similarly sample variance is an estimate of the population variance.
- Sampling is done when it is not feasible to access/process the whole population.

An Algebraic Identity

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

Sample variance

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

S^2 =sample variance

x_i =value of i th element

\bar{x} = sample mean

n=sample size

Population variance

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

σ^2 =population variance

x_i =value of i th element

μ =population mean

N=population size

Variance of Translated variable

$$\text{if } y_i = a + bx_i, \quad i = 1, \dots, n$$

if s_y^2 and s_x^2 are the respective sample variances, then $s_y^2 = b^2 s_x^2$

If you score 99 percentile in an exam. It is not that you scored 99 % (percent).
 It means that 99% of people scored \leq you and 1% people score \geq you

The *sample $100p$ percentile* is that data value such that at least $100p$ percent of the data are less than or equal to it and at least $100(1 - p)$ percent are greater than or equal to it. If two data values satisfy this condition, then the sample $100p$ percentile is the arithmetic average of these two values.

To determine the sample $100p$ percentile of a data set of size n , we need to determine the data values such that

1. At least np of the values are less than or equal to it.
2. At least $n(1 - p)$ of the values are greater than or equal to it.

Deciles

Arranging a set of data points in deciles means:

- i. Sort the data points in increasing order,
- ii. Distribute them in 10 bins in increasing order such that each bucket contains equal number of data points (10% in each bin). These bins are the deciles, e.g:

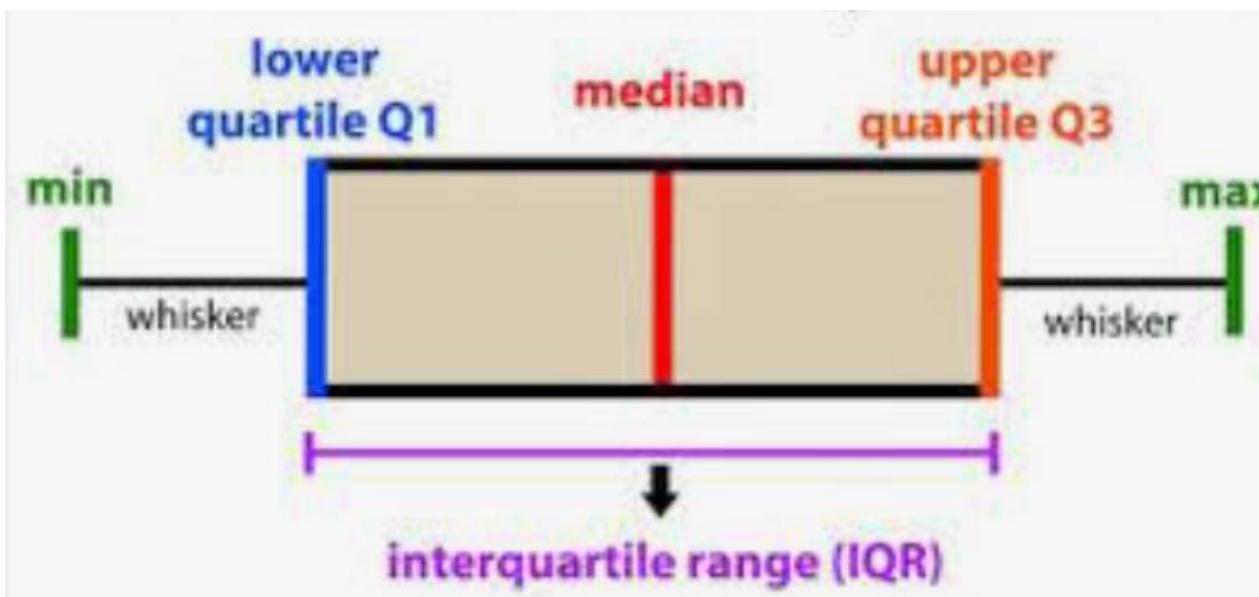
1	3	4	8	10	14	15	16	20	25	28	30	32	35	40	50	55	60	70	99
decile 1	decile 2	decile 3		decile 4		decile 5		decile 6		decile 7		decile 8		decile 9		decile 10			

Quartiles and Box Plots

- Box plot shows the min, 1st 2nd 3rd quartiles, and max values.
- 1st quartile = 25 percentile, 2nd quartile = 50 percentile = median, 3rd quartile = 75 percentile
- Box plot shows the skewness in data
- Also plot the mean in the box-plot. This gives a better view of the skew

The sample 25 percentile is called the *first quartile*; the sample 50 percentile is called the sample median or the *second quartile*; the sample 75 percentile is called the *third quartile*.

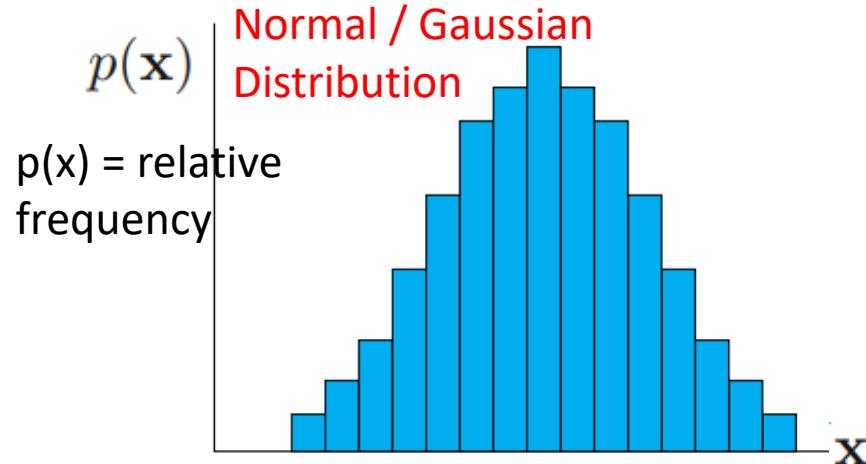
Box Plot



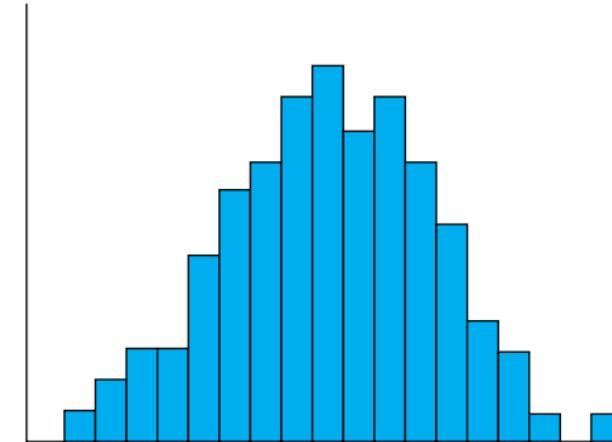
- The length of the line segment on the box plot, equal to the largest minus the smallest data value, is called the **range** of the data.
- the length of the box itself, equal to the third quartile minus the first quartile, is called the **interquartile range**.

Normal Data Sets Histograms / Distributions & Outliers

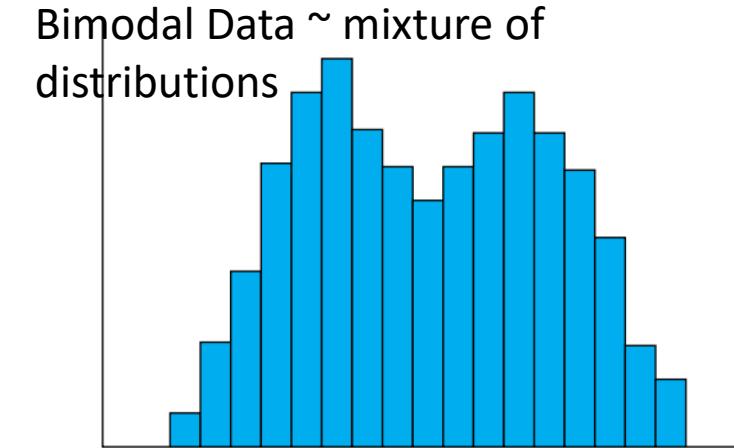
Many of the large data sets observed in practice have histograms that are similar in shape. These histograms often reach their peaks at the sample median and then decrease on both sides of this point in a bell-shaped symmetric fashion. Such data sets are said to be **Normal** and their histograms are called **Normal histograms**



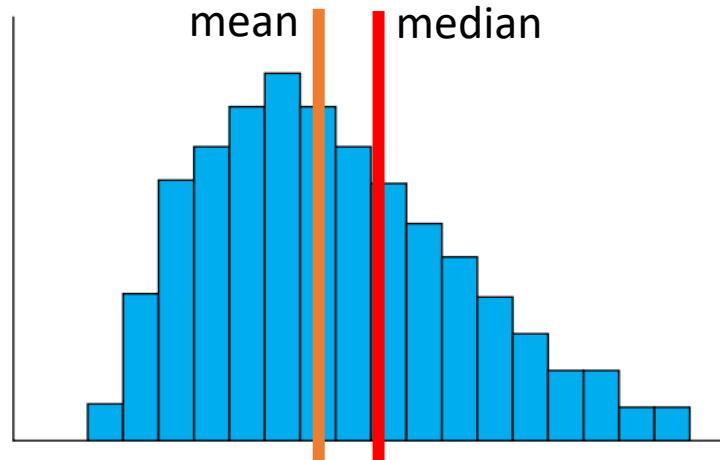
Histogram of a normal data set.



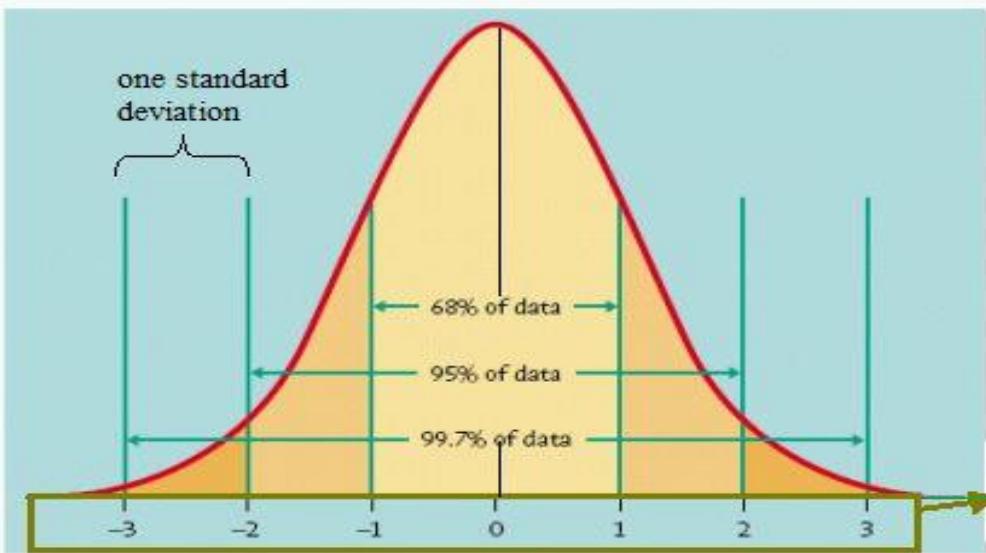
Histogram of an approximately normal data set.



Histogram of a bimodal data set.



Histogram of a data set skewed to the right.



Z – score = how many StDev away is a point from its mean

If $z > 3$, it maybe an **outlier**

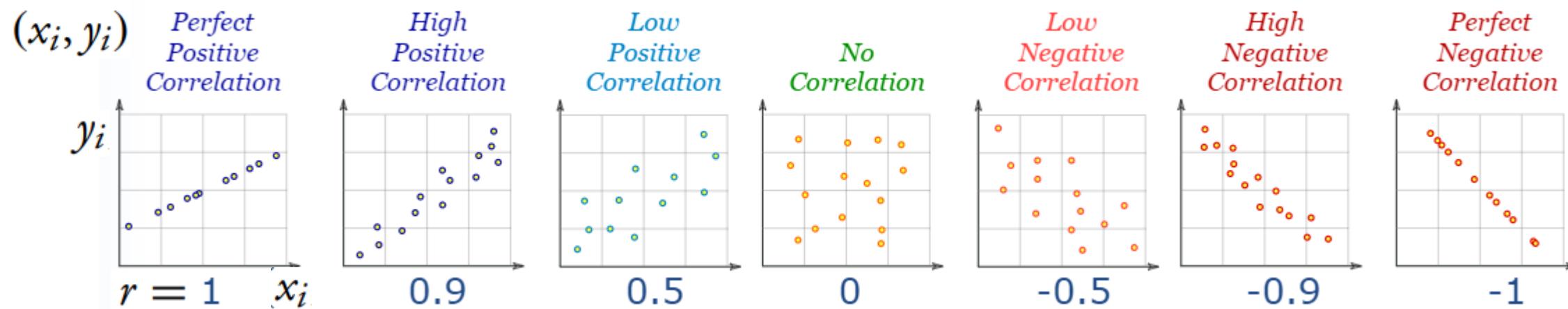
$$z = \frac{x - \mu}{\sigma}$$

$z > 3$: outlier

1.7

Correlation = strength of an assumed linear relationship

- Assumes and measures strength of linear relationship between two variables. Measures **Association, not Causation**
- An e.g., application: Features which are more correlated to the Target are more important. Fix all features at their mean except one, vary this by 1-sd and measure variation in target. The feature causing more variation in the Target is the more important one.



intuition: if $x_i - \bar{x}$ and $y_i - \bar{y}$ both have the same sign (either positive or negative), then their product $(x_i - \bar{x})(y_i - \bar{y})$ will be positive. Thus, it follows that when large x values tend to be associated with large y values and small x values are associated with small y values, then $\sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y})$ will tend to be a large positive number.

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

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

Properties of r

- $-1 \leq r \leq 1$
- If for constants a and b , with $b > 0$,
$$y_i = a + bx_i, \quad i = 1, \dots, n$$

then $r = 1$.

- If for constants a and b , with $b < 0$,
$$y_i = a + bx_i, \quad i = 1, \dots, n$$

then $r = -1$.

- If r is the sample correlation coefficient for the data pairs $x_i, y_i, i = 1, \dots, n$ then it is also the sample correlation coefficient for the data pairs

$$a + bx_i, \quad c + dy_i, \quad i = 1, \dots, n$$

provided that b and d are both positive or both negative.

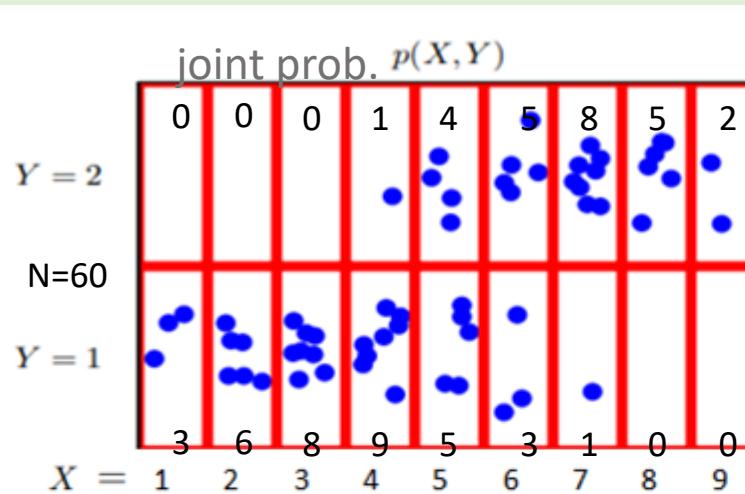
Chapter-2 Contents

1. Joint, marginal, conditional probability distribution
2. Three Laws of Probability
3. Expectation of $f(x)$ where x has PDF $p(x)$
4. Point Estimates: Max Likelihood (MLE) vs Max Posterior (MAPE)
5. Covariance between bi-variates (x, y) as matrix operation
6. Covariance among multi-variates
7. Summary: Probability Densities, expectation, variance, covariance, correlation

2.1

Joint, marginal, conditional probability distribution

- Multivariate data – joint probability distribution e.g., $p(x,y)$. (e.g., used for anomaly detection)
- Integrate the joint probability distribution to get **marginal probability distribution** e.g., $p(x)$

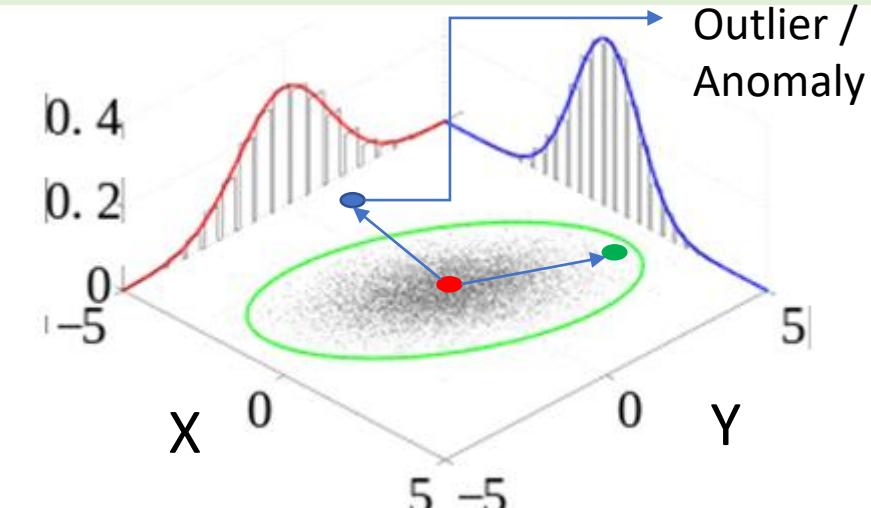
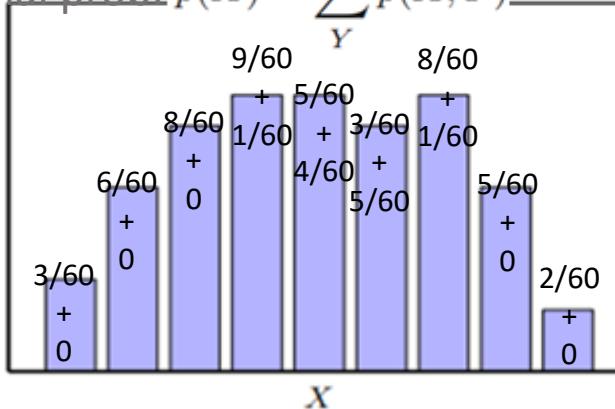
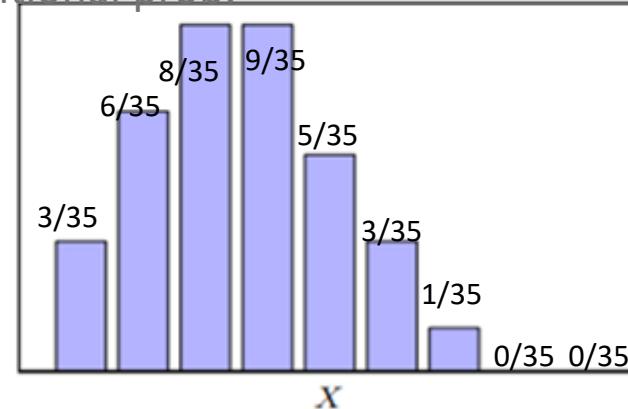
**Note:**

$$\begin{aligned} p(x=9) &> 0 \\ p(y=1) &> 0 \\ \text{but} \\ p(x=9, y=1) &= 0 ! \end{aligned}$$

marginal prob. $p(Y) = \sum_X p(X,Y)$

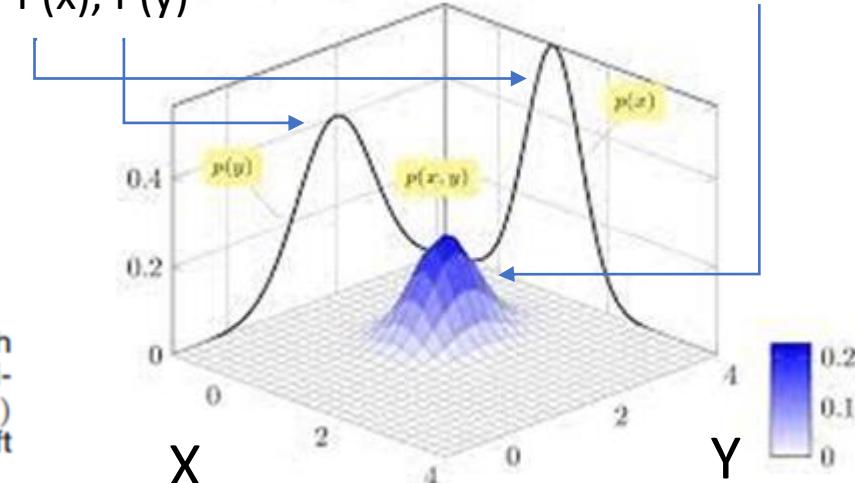
$$\begin{aligned} 0/60+0/60+0/60+1/60+4/60+5 \\ /60+8/60+5/60+2/60 = 25/60 \end{aligned}$$

$$\begin{aligned} 3/60+6/60+8/60+9/60+5/60+3 \\ /60+1/60+0/60+0/60 = 35/60 \end{aligned}$$

marginal prob. $p(X) = \sum_Y p(X,Y)$.conditional prob. $p(X|Y=1)$ 

Marginal
Prob. Distbn.
 $P(x), P(y)$

Joint Prob.
Distbn. $P(x,y)$



An illustration of a distribution over two variables, X , which takes 9 possible values, and Y , which takes two possible values. The top left figure shows a sample of 60 points drawn from a joint probability distribution over these variables. The remaining figures show histogram estimates of the marginal distributions $p(X)$ and $p(Y)$, as well as the conditional distribution $p(X|Y=1)$ corresponding to the bottom row in the top left figure.

2.2

Three Laws of Probability

- Sum rule and Product rule
- Baye's Theorem

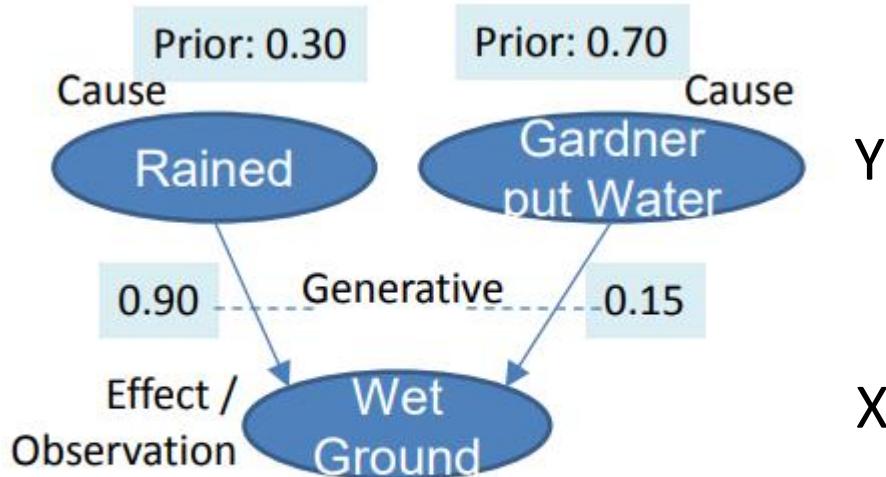
sum rule
product rule

$$p(X) = \sum_Y p(X, Y)$$

$$p(X, Y) = p(X|Y) p(Y)$$

$$p(Y|X) = \frac{p(X|Y)p(Y)}{p(X)}$$

$$p(X) = \sum_Y p(X|Y)p(Y)$$



'Wet Ground' is a weak entity. Its probability can be accounted only upon conditioning over the independent entities.

$$P(\text{rain}) = 0.3, P(\text{gardener}) = 0.7$$

$$P(\text{wet} | \text{rain}) = 0.9, P(\text{wet} | \text{gardener}) = 0.15$$

$$P(\text{rain} | \text{wet}) = ?$$

Baye's Theorem

Prior : $P(Y)$
Generative : $P(X|Y)$ – this is given.
Predictive : $P(Y|X)$ – need to estimate this

$$P(Y|X) = \frac{P(Y, X)}{P(X)} = \frac{P(Y) P(X|Y)}{P(X)} = \frac{P(Y) P(X|Y)}{\sum_Y P(X, Y)} = \frac{P(Y) P(X|Y)}{\sum_Y P(Y) P(X|Y)}$$

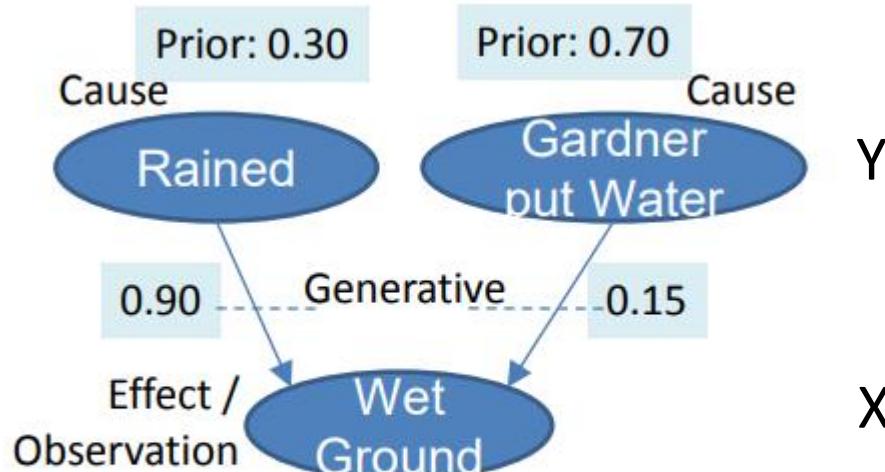
$$\text{Hence } P(Y|X) = \frac{P(Y) P(X|Y)}{\sum_Y P(Y) P(X|Y)}$$

Beauty of Baye's Theorem is that :
future (Prediction) is modeled through **past** (prior, generative)

2.2 Three Laws of Probability: question-1

- Sum rule and Product rule
- Baye's Theorem

sum rule	$p(X) = \sum_Y p(X, Y)$	$p(Y X) = \frac{p(X Y)p(Y)}{p(X)}$	$p(X) = \sum_Y p(X Y)p(Y)$
product rule	$p(X, Y) = p(X Y) p(Y)$		



'Wet Ground' is a weak entity. Its probability can be accounted only upon conditioning over the independent entities.

$$\begin{aligned} P(\text{rain}) &= 0.3, P(\text{gardener}) = 0.7 \\ P(\text{wet} | \text{rain}) &= 0.9, P(\text{wet} | \text{gardener}) = 0.15 \\ P(\text{rain} | \text{wet}) &= ? \end{aligned}$$

$$\begin{aligned} P(\text{rain} | \text{wet}) &= \frac{P(\text{rain}) P(\text{wet} | \text{rain})}{P(\text{rain}) P(\text{wet} | \text{rain}) + P(\text{gardener}) P(\text{wet} | \text{gardener})} \\ &= \frac{0.3 * 0.9}{0.3 * 0.9 + 0.7 * 0.15} = 0.72 \end{aligned}$$

Baye's Theorem

Prior : $P(Y)$ – this is given.

Generative : $P(X|Y)$ – this is given.

Predictive : $P(Y|X)$ – need to estimate this

$$P(Y|X) = \frac{P(Y, X)}{P(X)} = \frac{P(Y) P(X|Y)}{P(X)} = \frac{P(Y) P(X|Y)}{\sum_Y P(X, Y)} = \frac{P(Y) P(X|Y)}{\sum_Y P(Y) P(X|Y)}$$

$$\text{Hence } P(Y|X) = \frac{P(Y) P(X|Y)}{\sum_Y P(Y) P(X|Y)}$$

Beauty of Baye's Theorem is that :
future (Prediction) is modeled through past (prior, generative)

For weak/dependent entities we can only derive marginal probabilities (sum rule)

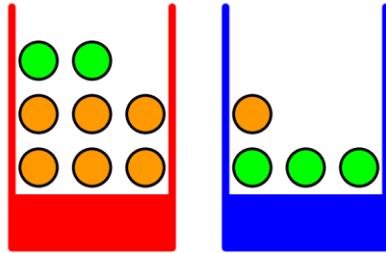
2.2 Three Laws of Probability : question-2

- Sum rule and Product rule
- Baye's Theorem

sum rule $p(X) = \sum_Y p(X, Y)$

product rule $p(X, Y) = p(X|Y) p(Y)$

$$p(Y|X) = \frac{p(X|Y)p(Y)}{p(X)} \quad p(X) = \sum_Y p(X|Y)p(Y)$$



Total #apple = 5

Total #Fruits = 12

Hence $p(\text{apple}) = 5/12$?

- No, this is Wrong

Given a red and blue box. Red box has 2 apples & 6 oranges. Blue box has 3 apples and 1 orange. Your eyes are closed and you have to pick any fruit. Blue box is bigger in size hence probability of selecting blue-box is

0.6. Problem:

- What is probability to pick an apple ?
- If an orange is picked, what is prob. if the box is red ?

Baye's Theorem

Prior : $P(Y)$ – this is given.

Generative : $P(X|Y)$ – this is given.

Predictive : $P(Y|X)$ – need to estimate this

$$P(Y|X) = \frac{P(Y, X)}{P(X)} = \frac{P(Y) P(X|Y)}{P(X)} = \frac{P(Y) P(X|Y)}{\sum_Y P(X, Y)} = \frac{P(Y) P(X|Y)}{\sum_Y P(Y) P(X|Y)}$$

$$\text{Hence } P(Y|X) = \frac{P(Y) P(X|Y)}{\sum_Y P(Y) P(X|Y)}$$

Beauty of Baye's Theorem is that : future (Prediction) is modeled through past (prior, generative)

2.2 Three Laws of Probability : question-2

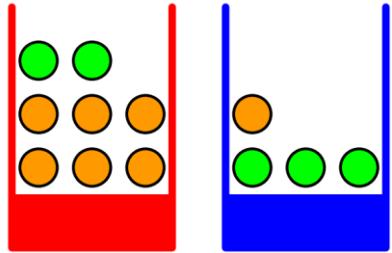
- Sum rule and Product rule
- Baye's Theorem

sum rule $p(X) = \sum_Y p(X, Y)$

product rule $p(X, Y) = p(X|Y) p(Y)$

$$p(Y|X) = \frac{p(X|Y)p(Y)}{p(X)}$$

$$p(X) = \sum_Y p(X|Y)p(Y)$$



Given a red and blue box. Red box has 2 apples & 6 oranges. Blue box has 3 apples and 1 orange. Your eyes are closed and you have to pick any fruit. Blue box is bigger in size hence probability of selecting blue-box is

0.6. Problem:

- What is probability to pick an apple ?
- If an orange is picked, what is prob. if the box is red ?

Total #apple = 5
Total #Fruits = 12
Hence $p(\text{apple}) = 5/12$?
– No, this is Wrong

Random Vars

B (Box):

r = red

b = blue

F (Fruit):

a = apple (green color)

o = orange (orange color)

Given

$$p(B = r) = 4/10$$

$$p(B = b) = 6/10$$

$$p(F = a|B = r) = 1/4$$

$$p(F = o|B = r) = 3/4$$

$$p(F = a|B = b) = 3/4$$

$$p(F = o|B = b) = 1/4.$$

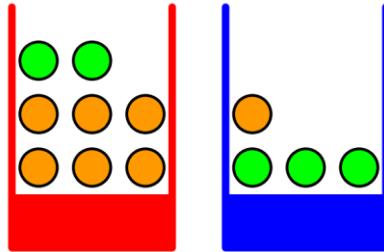
2.2 Three Laws of Probability : question-2

- Sum rule and Product rule
- Baye's Theorem

sum rule $p(X) = \sum_Y p(X, Y)$

product rule $p(X, Y) = p(X|Y) p(Y)$

$$p(Y|X) = \frac{p(X|Y)p(Y)}{p(X)} \quad p(X) = \sum_Y p(X|Y)p(Y)$$



Given a red and blue box. Red box has 2 apples & 6 oranges. Blue box has 3 apples and 1 orange. Your eyes are closed and you have to pick any fruit. Blue box is bigger in size hence probability of selecting blue-box is

0.6. Problem:

- What is probability to pick an apple ?
- If an orange is picked, what is prob. if the box is red ?

Total #apple = 5
Total #Fruits = 12
Hence $p(\text{apple}) = 5/12$?
– No, this is Wrong

Random Vars

B (Box):

r = red

b = blue

F (Fruit):

a = apple (green color)

o = orange (orange color)

Given

$$p(B = r) = 4/10$$

$$p(B = b) = 6/10$$

$$p(F = a|B = r) = 1/4$$

$$p(F = o|B = r) = 3/4$$

$$p(F = a|B = b) = 3/4$$

$$p(F = o|B = b) = 1/4.$$

$P(F = a) = ?$

$$p(F = a) = p(F = a, B = r) + p(F = a, B = b) \quad : \text{Sum Rule}$$

$$p(F = a) = \underbrace{p(F = a|B = r)p(B = r)}_{= \frac{1}{4} \times \frac{4}{10}} + \underbrace{p(F = a|B = b)p(B = b)}_{= \frac{3}{4} \times \frac{6}{10}} \quad : \text{Product Rule}$$

$$p(F = a) = \frac{1}{4} \times \frac{4}{10} + \frac{3}{4} \times \frac{6}{10} = \frac{11}{20} = 0.55$$

$$P(F = o) = 1 - P(F = a) = 9/20 = 0.45$$

$$p(x) = \int p(x, y) dy \quad \text{If continuous vars}$$

If we had been asked which box had been chosen before being told the identity of the selected item of fruit, then the most complete information we have available is provided by the probability $p(B)$. We call this the **prior probability** because it is the probability available *before* we observe the identity of the fruit. Once we are told that the fruit is an orange, we can then use Bayes' theorem to compute the probability $p(B|F)$, which we shall call the **posterior probability** because it is the probability obtained *after* we have observed F .

$P(B = r | F = o) = ?$

: Baye's Theorem

$$p(B = r | F = o) = \frac{p(F = o | B = r)p(B = r)}{p(F = o)}$$

$$= \frac{3}{4} \times \frac{4}{10} \times \frac{20}{9} = \frac{2}{3}.$$

Posterior $P(B=r | F=o)$ is higher even when prior $P(B=b)$ is higher!

Expectation of $f(x)$ where x has PDF $p(x)$

$$E[f(x)] = \sum_x f(x) p(x) \quad \text{or} \quad \int_x f(x) p(x) dx$$

$$\text{if } f(x) = x \rightarrow E[x] = \int_x x p(x) dx$$

$$\mathbb{E}[f] \simeq \frac{1}{N} \sum_{n=1}^N f(x_n). \quad x \text{ sampled from } p(x)$$

A person can walk max. up to 10 km. Probability that a person walks a distance of x km. is proportional to x . What is the expected distance a person walks ?

2.3

Expectation of $f(x)$ where x has PDF $p(x)$: question-1

$$E[f(x)] = \sum_x f(x) p(x) dx \text{ or } \int_x f(x) p(x) dx$$

$$\text{if } f(x) = x \rightarrow E[x] = \int_x x p(x) dx$$

$$\mathbb{E}[f] \simeq \frac{1}{N} \sum_{n=1}^N f(x_n). \quad x \text{ sampled from } p(x)$$

A person can walk max. up to 10 km. Probability that a person walks a distance of x km is proportional to x . What is the expected distance a person walks ?

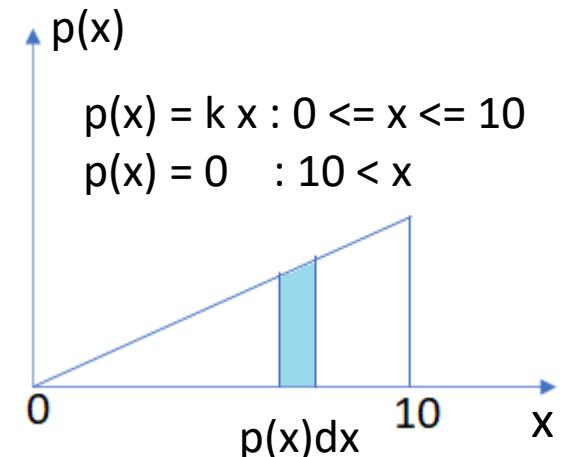
$$\begin{aligned} p(x) &= kx \\ \int_0^{10} p(x) dx &= 1 \end{aligned}$$

$$\text{i.e. } \int_0^{10} kx dx = 1$$

$$\text{i.e. } \left[\frac{kx^2}{2} \right]_0^{10} = 1$$

$$\text{i.e. } k[100 - 0] = 2 \\ \text{hence } k = 0.02$$

$$\begin{aligned} E(x) &= \int_0^{10} x p(x) dx \\ &= \int_0^{10} x(kx) dx \\ &= 0.02 \int_0^{10} x^2 dx \\ &= 0.02 \left[\frac{x^3}{3} \right]_0^{10} \\ &= \frac{0.02 * 1000}{3} \\ &= 6.67 \text{ km!} \end{aligned}$$



Point Estimates: Max Likelihood (MLE) vs Max Posterior (MAPE) Estimate

- MLE: An estimate which maximizes the likelihood of the observation.
- MAPE: Models the posterior probability distribution.

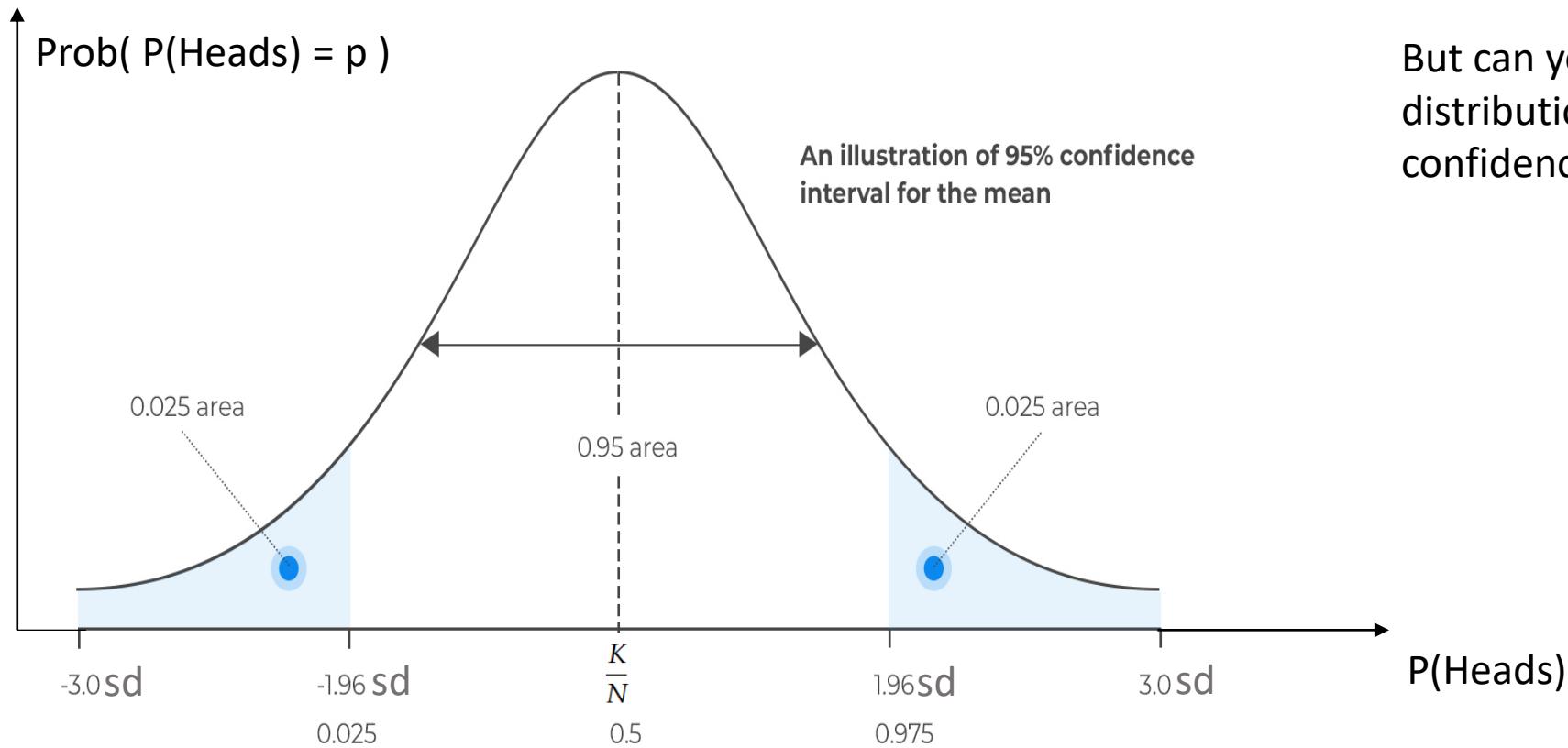
$$E[f(x)] = \sum_x f(x) p(x) dx \text{ or } \int_x f(x) p(x) dx$$

Prior info : A coin can show up Heads with probability anywhere between [0, 1].

Event E : out of N trials K heads occur.

Question : what is $P(\text{Head})$ now ?

Max. Likelihood: $P(\text{Head}) = \frac{K}{N}$



But can you predict a probability distribution for $P(\text{Heads})$ and suggest confidence intervals ?

2.4 Point Estimates: Max Likelihood (MLE) vs Max Posterior (MAPE) Estimate: question

- MLE: An estimate which maximizes the likelihood of the observation.
- MAPE: Models the posterior probability distribution.

$$E[f(x)] = \sum_x f(x) p(x) dx \text{ or } \int_x f(x) p(x) dx$$

Prior info: A coin can show up Heads with probability anywhere between [0, 1].

Event E: out of N trials K heads occur.

Question: what is $P(\text{Head})$ now ?

Max. Likelihood: $P(\text{Head}) = \frac{K}{N}$

Let \mathbf{p} be a random var. that is $P(\text{Head})$. Let $f()$ represent a prob. density func.

$$P(E|\mathbf{p} = p) = p^K (1-p)^{(n-K)}$$

$$P(E) = \int_0^1 P(E|\mathbf{p} = p) f(p) dp$$

$$P(E) = \int_0^1 p^K (1-p)^{(n-K)} dp = \frac{(n-k)! k!}{(n+1)!}$$

By Baye's Theorem :

$$f(p|E) = \frac{f(p) P(E|\mathbf{p} = p)}{P(E)}$$

MAPE

$$f(p|E) = \frac{(n+1)! p^K (1-p)^{(n-K)}}{(n-k)! k!}$$

- MAPE is considered more robust because it incorporates prior information
- Considering a probability distribution is more robust than a point estimate

MAPE: the point p at which this is maximized. In this case MAPE estimate is same as MLE: K/N

2.5

Covariance between bi-variates (x, y) as matrix operation

Covariance between variables (x, y):

$$COV(x, y) = \frac{[x - \bar{x}]^T[y - \bar{y}]}{n}$$

and variance :

$$VAR(x) = \frac{[x - \bar{x}]^T[x - \bar{x}]}{n}$$

X	Y
X1	Y1
X2	Y2
X3	Y3
...	...
Xn	Yn

$$COV(x, y) = \frac{[x - \bar{x}]^T[y - \bar{y}]}{n} = \frac{1}{n}[x_1 - \bar{x}, x_2 - \bar{x}, \dots, x_n - \bar{x}] \begin{bmatrix} y_1 - \bar{y} \\ y_2 - \bar{y} \\ \dots \\ y_n - \bar{y} \end{bmatrix}$$

$$= \frac{(x_1 - \bar{x})(y_1 - \bar{y}) + (x_2 - \bar{x})(y_2 - \bar{y}) + \dots + (x_n - \bar{x})(y_n - \bar{y})}{n}$$

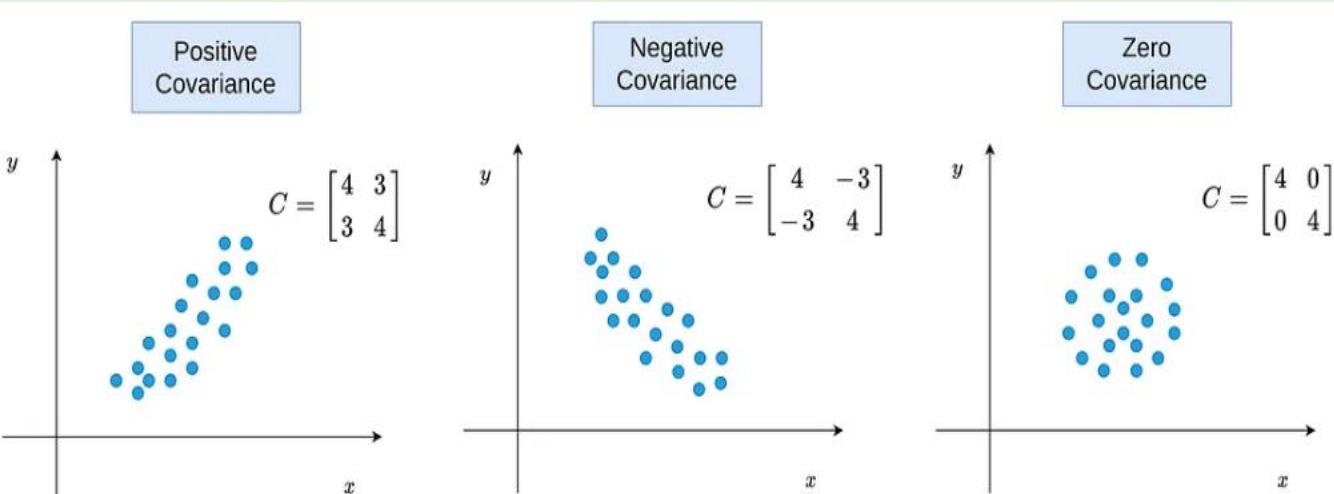
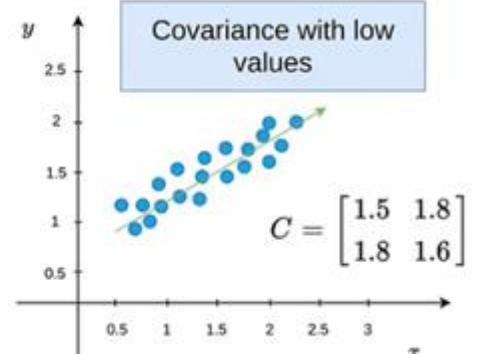
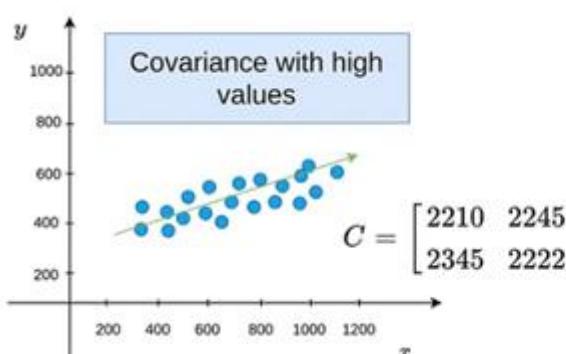
similarly

$$VAR(x) = \frac{[x - \bar{x}]^T[x - \bar{x}]}{n} = \frac{1}{n}[x_1 - \bar{x}, x_2 - \bar{x}, \dots, x_n - \bar{x}] \begin{bmatrix} x_1 - \bar{x} \\ x_2 - \bar{x} \\ \dots \\ x_n - \bar{x} \end{bmatrix}$$

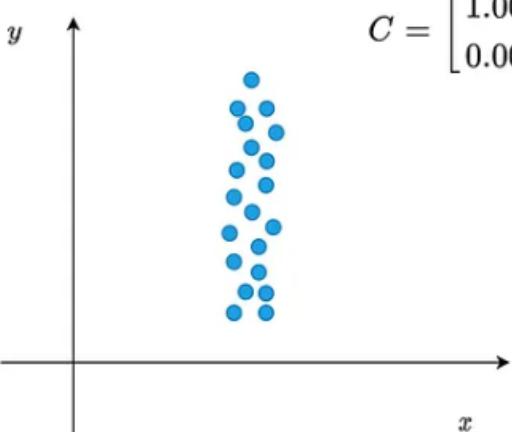
$$= \frac{(x_1 - \bar{x})(x_1 - \bar{x}) + (x_2 - \bar{x})(x_2 - \bar{x}) + \dots + (x_n - \bar{x})(x_n - \bar{x})}{n}$$

Covariance Matrix =

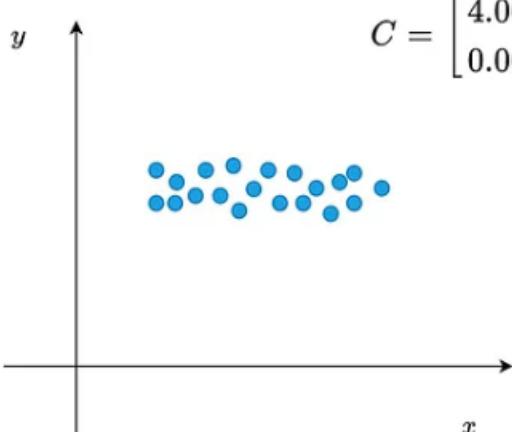
$$\begin{bmatrix} var(x) & cov(x, y) & cov(x, z) \\ cov(x, y) & var(y) & cov(y, z) \\ cov(x, z) & cov(y, z) & var(z) \end{bmatrix}$$



Low variance for x



Low variance for y



2.6

Covariance among multi-variates say $W = [x, y, z]$ is

$$\text{Cov}(W) = [W - \bar{W}]^T N^{-1} [W - \bar{W}]$$

$$W = \begin{bmatrix} x_1 & y_1 & z_1 \\ x_2 & y_1 & z_2 \\ .. & .. & .. \\ x_n & y_n & z_n \end{bmatrix} \quad W - \bar{W} = \begin{bmatrix} x_1 - \bar{x} & y_1 - \bar{y} & z_1 - \bar{z} \\ x_2 - \bar{x} & y_2 - \bar{y} & z_2 - \bar{z} \\ .. & .. & .. \\ x_n - \bar{x} & y_n - \bar{y} & z_n - \bar{z} \end{bmatrix} \quad \text{N}^{-1} = \begin{bmatrix} 1/n & 0 & 0.. & 0 \\ 0 & 1/n & 0.. & 0 \\ 0.. & 0.. & 1/n.. & 0.. \\ 0 & 0 & 0.. & 1/n \end{bmatrix}$$

$(n \times n)$ diagonal matrix

for population, but if for a sample then

$$\text{N}^{-1} = \begin{bmatrix} 1/(n-1) & 0 & 0.. & 0 \\ 0 & 1/(n-1) & 0.. & 0 \\ 0.. & 0.. & 1/(n-1).. & 0.. \\ 0 & 0 & 0.. & 1/(n-1) \end{bmatrix}$$

$(n \times n)$ diagonal matrix

$$\text{then } \text{Cov}(W) = [W - \bar{W}]^T N^{-1} [W - \bar{W}]$$

$$= \begin{bmatrix} x_1 - \bar{x} & x_2 - \bar{x} & .. & x_n - \bar{x} \\ y_1 - \bar{y} & y_2 - \bar{y} & .. & y_n - \bar{y} \\ z_1 - \bar{z} & z_2 - \bar{z} & .. & z_n - \bar{z} \end{bmatrix} \begin{bmatrix} 1/n & 0 & 0.. & 0 \\ 0 & 1/n & 0.. & 0 \\ 0.. & 0.. & 1/n.. & 0.. \\ 0 & 0 & 0.. & 1/n \end{bmatrix} \begin{bmatrix} x_1 - \bar{x} & y_1 - \bar{y} & z_1 - \bar{z} \\ x_2 - \bar{x} & y_2 - \bar{y} & z_2 - \bar{z} \\ .. & .. & .. \\ x_n - \bar{x} & y_n - \bar{y} & z_n - \bar{z} \end{bmatrix}$$

$$= \begin{bmatrix} \frac{(x_1 - \bar{x})}{n} & \frac{(x_2 - \bar{x})}{n} & .. & \frac{(x_n - \bar{x})}{n} \\ \frac{(y_1 - \bar{y})}{n} & \frac{(y_2 - \bar{y})}{n} & .. & \frac{(y_n - \bar{y})}{n} \\ \frac{(z_1 - \bar{z})}{n} & \frac{(z_2 - \bar{z})}{n} & .. & \frac{(z_n - \bar{z})}{n} \end{bmatrix} \begin{bmatrix} x_1 - \bar{x} & y_1 - \bar{y} & z_1 - \bar{z} \\ x_2 - \bar{x} & y_2 - \bar{y} & z_2 - \bar{z} \\ .. & .. & .. \\ x_n - \bar{x} & y_n - \bar{y} & z_n - \bar{z} \end{bmatrix}$$

$$= \begin{bmatrix} \frac{[x - \bar{x}]^T [x - \bar{x}]}{n} & \frac{[x - \bar{x}]^T [y - \bar{y}]}{n} & \frac{[x - \bar{x}]^T [z - \bar{z}]}{n} \\ \frac{[y - \bar{y}]^T [x - \bar{x}]}{n} & \frac{[y - \bar{y}]^T [y - \bar{y}]}{n} & \frac{[y - \bar{y}]^T [z - \bar{z}]}{n} \\ \frac{[z - \bar{z}]^T [x - \bar{x}]}{n} & \frac{[z - \bar{z}]^T [y - \bar{y}]}{n} & \frac{[z - \bar{z}]^T [z - \bar{z}]}{n} \end{bmatrix}$$

$$= \begin{bmatrix} \text{var}(x) & \text{cov}(x, y) & \text{cov}(x, z) \\ \text{cov}(y, x) & \text{var}(y) & \text{cov}(y, z) \\ \text{cov}(z, x) & \text{cov}(z, y) & \text{var}(z) \end{bmatrix} = \text{Cov}(W)$$

if $W = [x, y, z, ...]$ be a multivariate

$$\text{Let } W_{\text{centered}} = [(x - \bar{x}) \ (y - \bar{y}) \ (z - \bar{z}) \dots]$$

$$\text{then } \text{COV}(W) = W_{\text{centered}}^T N^{-1} W_{\text{centered}}$$

$$\text{Let } W_{\text{norm}} = \left[\frac{(x - \bar{x})}{S_x} \ \frac{(y - \bar{y})}{S_y} \ \frac{(z - \bar{z})}{S_z} \dots \right]$$

where $S_x \ S_y \ S_z \dots$ are the Std. dev of vars $x \ y \ z \dots$

$$\text{then } \text{CORR}(W) = W_{\text{norm}}^T N^{-1} W_{\text{norm}}$$

2.7 Summary: Probability Densities, expectation, variance, covariance, correlation

Important to get a hang of Expectation, Variance, Covariance, Correlation

Probability distribution/density properties

$$p(x \in (a, b)) = \int_a^b p(x) dx. \quad \text{or} \quad \sum_{a < x \leq b} p(x)$$

if x is continuous if x is discrete

$$\int_{-\infty}^{\infty} p(x) dx = 1. \quad \text{or} \quad \sum_x p(x) = 1$$

$$p(x) = \int p(x, y) dy \quad \text{or} \quad p(x) = \sum_y p(x, y)$$

$$p(x, y) = p(y|x)p(x).$$

Expected value of f(x)

$$E[f(x)] = \sum_x f(x) p(x) \quad \text{or} \quad \int_x f(x) p(x) dx$$

$$\mathbb{E}[f] \simeq \frac{1}{N} \sum_{n=1}^N f(x_n). \quad x \text{ sampled from } p(x)$$

$$E[X] = \sum_i x_i P\{X = x_i\}$$

Variance of f(x)

If x independent y: $E[XY] = E[X]E[Y]$.

$$\text{var}[f] = \mathbb{E}[(f(x) - \mathbb{E}[f(x)])^2] = \mathbb{E}[f(x)^2] - \mathbb{E}[f(x)]^2.$$

$$\text{var}[x] = E[(x - E[x])^2] = \mathbb{E}[x^2] - \mathbb{E}[x]^2.$$

$$E[aX + b] = aE[X] + b$$

$$E[X + Y] = E[X] + E[Y]$$

$$\text{Var}(aX + b) = a^2 \text{Var}(X)$$

$$\text{avg}(x - \bar{x}) = 0 \quad \text{or} \quad E(x - E(x)) = 0$$

$$\text{or} \quad \sum_x (x - E[x])p(x) = 0$$

$$\text{for independent } X_1, \dots, X_n, \quad \text{Var}\left(\sum_{i=1}^n X_i\right) = \sum_{i=1}^n \text{Var}(X_i)$$

$$\text{Corr}(X, Y) = \frac{\text{Cov}(X, Y)}{\sqrt{\text{Var}(X)\text{Var}(Y)}}$$

Covariance

$$\text{Cov}(X, Y) = E[(X - \mu_x)(Y - \mu_y)] = E[XY] - E[X]E[Y]$$

$$\text{cov}(aX, bY) = ab \text{ cov}(X, Y)$$

If X and Y are independent random variables, then

$$\text{Cov}(X, Y) = 0$$

Population Covariance Formula

$$\text{Cov}(x, y) = \frac{\sum (x_i - \bar{x})(y_i - \bar{y})}{N}$$

Sample Covariance

$$\text{Cov}(x, y) = \frac{\sum (x_i - \bar{x})(y_i - \bar{y})}{N - 1}$$

Correlation

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

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

Chapter-3 Contents

1. Special Random Variables: Uniform and Normal RVs
2. Uniform Random Variable: Max. Likelihood Estimate (MLE) example
3. Special Random Variables: Chi-Square, t-Distribution, F-Distribution
4. Distributions of Sampling Statistics

3.1

Special Random Variables: Uniform and Normal RVs

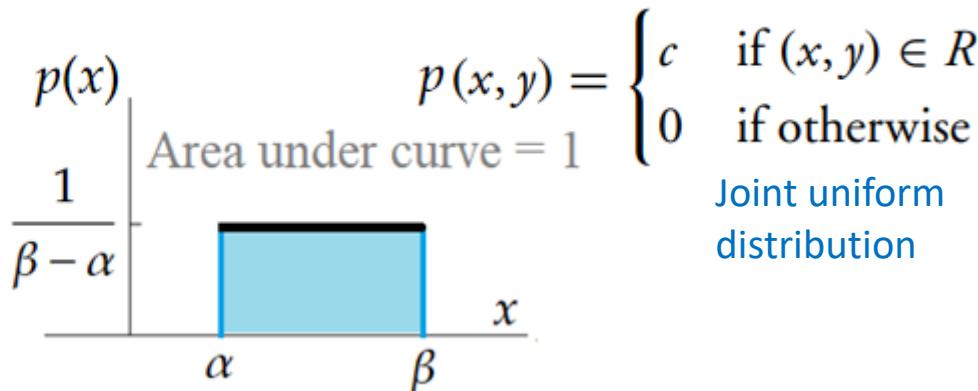
- Define Randomness: randomness is a probability distribution (think histogram) e.g.:
- Uniform randomness, Gaussian/Normal Randomness, etc..

THE UNIFORM RANDOM VARIABLE

A random variable X is said to be uniformly distributed over the interval $[\alpha, \beta]$ if its probability density function is given by

$$p(x) = \begin{cases} \frac{1}{\beta - \alpha} & \text{if } \alpha \leq x \leq \beta \\ 0 & \text{otherwise} \end{cases}$$

Note $\sum_{\alpha}^{\beta} p(x) \text{ or } \int_{\alpha}^{\beta} p(x) dx = \frac{1}{\beta - \alpha} \int_{\alpha}^{\beta} dx = 1$



NORMAL RANDOM VARIABLES

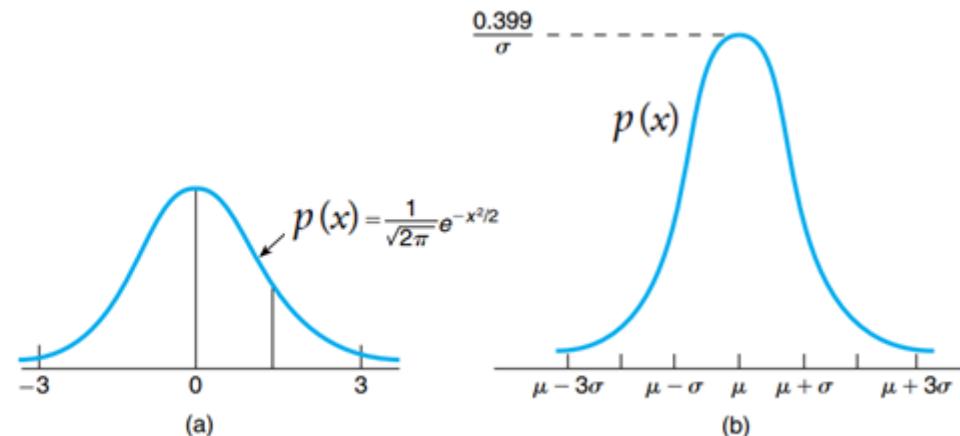
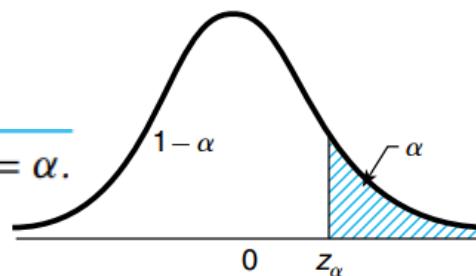
A random variable is said to be normally distributed with parameters μ and σ^2 , and we write $X \sim \mathcal{N}(\mu, \sigma^2)$, if its density is

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

The normal density $p(x)$ is a bell-shaped curve that is symmetric about μ and that attains its maximum value of $\frac{1}{\sqrt{2\pi}\sigma} \approx 0.399/\sigma$ at $x = \mu$

if $X \sim N(\mu, \sigma^2)$ then $Z = \frac{x-\mu}{\sigma} \sim N(0, 1)$ is Std. Normal

$$\underline{P\{Z > z_{\alpha}\} = \alpha.}$$



The normal density function (a) with $\mu = 0, \sigma = 1$ and (b) with arbitrary μ and σ^2 .

Uniform Random Variable: Max. Likelihood Estimate (MLE)

What is some param (e.g. mean) θ of a given population distribution when some samples from that population are observed ?

Maximum Likelihood Estimate of θ = that value which maximizes the likelihood of the observations.

Given : Samples X_1, X_2, \dots, X_n from a uniformly distributed population over (0, θ)

Question : What is the population mean ?

Uniform Random Variable: Max. Likelihood Estimate (MLE): question-1

What is some param (e.g. mean) θ of a given population distribution when some samples from that population are observed ?

Maximum Likelihood Estimate of θ = that value which maximizes the likelihood of the observations.

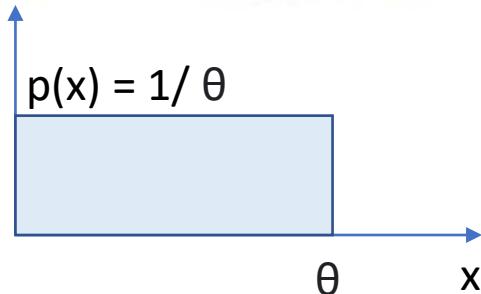
Given : Samples X_1, X_2, \dots, X_n from a uniformly distributed population over $(0, \theta)$

Question : What is the population mean ?

Estimating the Mean of a Uniform Distribution

Suppose X_1, \dots, X_n constitute a sample from a uniform distribution on $(0, \theta)$, where θ is unknown. Their joint density is thus

$$f(x_1, x_2, \dots, x_n | \theta) = \begin{cases} \frac{1}{\theta^n} & 0 < x_i < \theta, \quad i = 1, \dots, n \\ 0 & \text{otherwise} \end{cases}$$



This density is maximized by choosing θ as small as possible. Since θ must be at least as large as all of the observed values x_i , it follows that the smallest possible choice of θ is equal to $\max(x_1, x_2, \dots, x_n)$. Hence, the maximum likelihood estimator of θ is

$$\hat{\theta} = \max(X_1, X_2, \dots, X_n)$$

$$\text{mean} = \frac{\hat{\theta}}{2} = \frac{\max(X_1, X_2, \dots, X_n)}{2}$$

3.3 Special Random Variables: Chi-Square, t-Distribution, F-Distribution

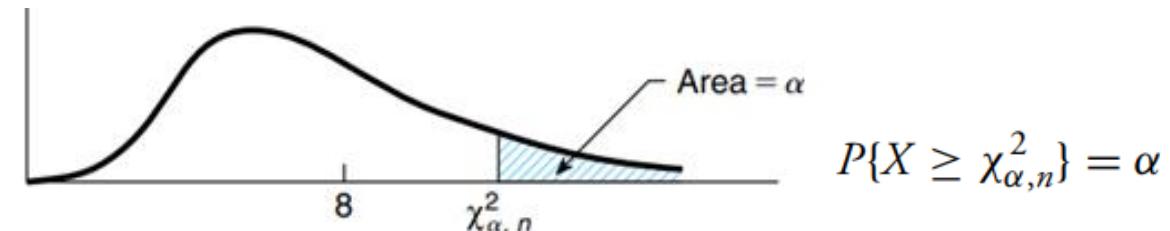
Some distributions are 1-sided ($x \geq 0$), some are 2-sided

THE CHI-SQUARE DISTRIBUTION

If Z_1, Z_2, \dots, Z_n are independent standard normal random variables,

$$\text{then } X, \text{ defined by } X = Z_1^2 + Z_2^2 + \dots + Z_n^2 \quad X \sim \chi_n^2$$

is said to have a *chi-square distribution with n degrees of freedom*.



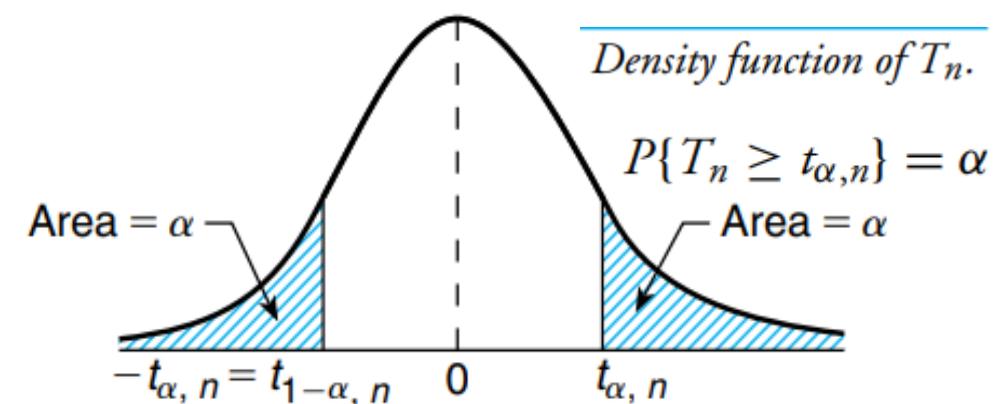
The chi-square density function with 8 degrees of freedom.

THE t-DISTRIBUTION

If Z and χ_n^2 are independent random variables, with Z having a standard normal distribution and χ_n^2 having a chi-square distribution with n degrees of freedom, then the random variable T_n defined by

$$T_n = \frac{Z}{\sqrt{\chi_n^2/n}} \quad \text{where} \quad \frac{\chi_n^2}{n} = \frac{Z_1^2 + \dots + Z_n^2}{n} \quad E[T_n] = 0, \quad n > 1$$

is said to have a *t-distribution with n degrees of freedom*. $\text{Var}(T_n) = \frac{n}{n-2}, \quad n > 2$

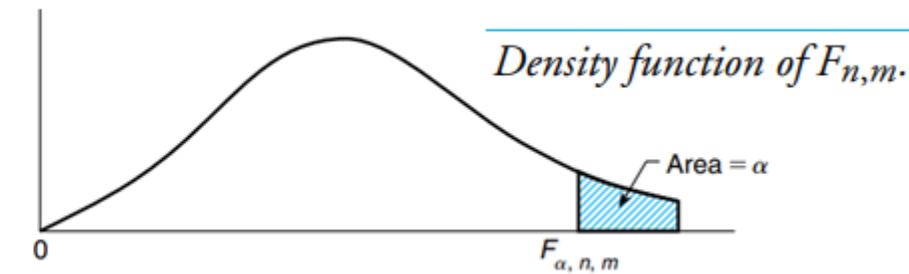


THE F-DISTRIBUTION

If χ_n^2 and χ_m^2 are independent chi-square random variables with n and m degrees of freedom, respectively, then the random variable $F_{n,m}$ defined by

$$F_{n,m} = \frac{\chi_n^2/n}{\chi_m^2/m} \quad \text{is said to have an } F\text{-distribution with } n \text{ and } m \text{ degrees of freedom.}$$

For any $\alpha \in (0, 1)$, let $F_{\alpha, n, m}$ be such that $P\{F_{n,m} > F_{\alpha, n, m}\} = \alpha$



3.4 Distributions of Sampling Statistics

Sample mean is a point estimate of the population mean. Sample mean is a random variable

Properties of Mean

The quantities μ and σ^2 are called the *population mean* and the *population variance*, respectively. Let X_1, X_2, \dots, X_n be a sample of values from this population. The sample mean is defined by

$$\bar{X} = \frac{X_1 + \dots + X_n}{n} \quad E(\bar{X}) = \mu \quad \text{Var}(\bar{X}) = \frac{\sigma^2}{n}$$

The Central Limit Theorem

Let X_1, X_2, \dots, X_n be a sequence of independent and identically distributed random variables each having mean μ and variance σ^2 . Then for n large, the distribution of $X_1 + \dots + X_n$ is approximately normal with mean $n\mu$ and variance $n\sigma^2$.

Properties of Variance

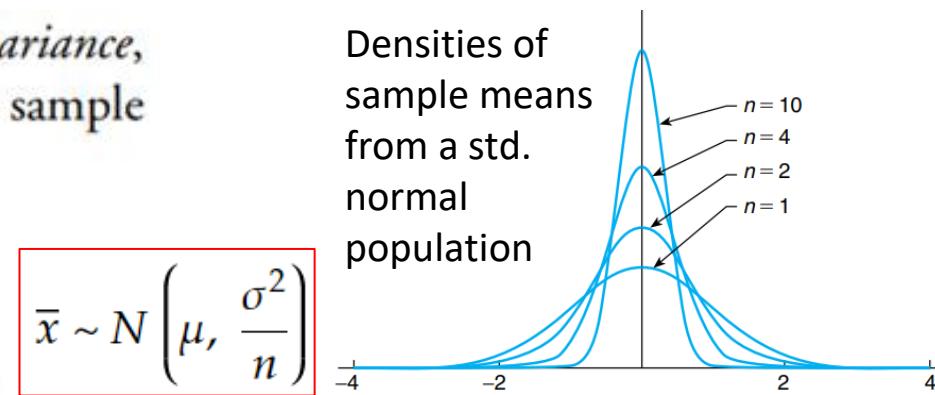
If X_1, \dots, X_n is a sample from a normal population having mean μ and variance σ^2 , then \bar{X} and S^2 are independent random variables, with \bar{X} being normal with mean μ and variance σ^2/n and $(n-1)S^2/\sigma^2$ being chi-square with $n-1$ degrees of freedom.

Binomial Distribution

The probability mass function of a binomial random variable with parameters n and p

$$P\{X = i\} = \binom{n}{i} p^i (1-p)^{n-i}, \quad i = 0, 1, \dots, n \quad \text{where } \binom{n}{i} = n!/[i!(n-i)!]$$

$$E[X] = np \quad \text{and} \quad SD(X) = \sqrt{np(1-p)}$$



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

A general rule of thumb is that one can be confident of the normal approximation whenever the sample size n is at least **30**.

$$(n-1) \frac{S^2}{\sigma^2} \sim \chi^2_{n-1}$$

$$\sqrt{n} \frac{(\bar{X} - \mu)}{S} \sim t_{n-1}$$

$$E(\bar{X}) = np$$

$$\text{Var}(\bar{X}) = \text{Var}(X)/n^2 = p(1-p)/n$$

$$\bar{X} \sim N\left(np, \frac{p(1-p)}{n}\right)$$

Chapter-4 Contents

1. Interval Estimate (Confidence Interval) for population mean
2. Point Estimate (MLE) vs Interval Estimate (Confidence Interval)
3. Two-sided and one-sided Confidence Interval
4. Distributions of Sampling Statistics
5. Parameter Estimation
6. Hypothesis Testing

4.1

Interval Estimate (Confidence Interval) for population mean

If we do sampling multiple times, each time the sample mean maybe different. This tells that the sample mean is a random var and 95% times actual population mean will lie within $1.96 * \text{Stdev}$ of sample mean.

$$x_1 | x_2 | x_2 | \dots | x_n | \dots$$

$$x_1 | x_2 | x_2 | \dots | x_n | \dots$$

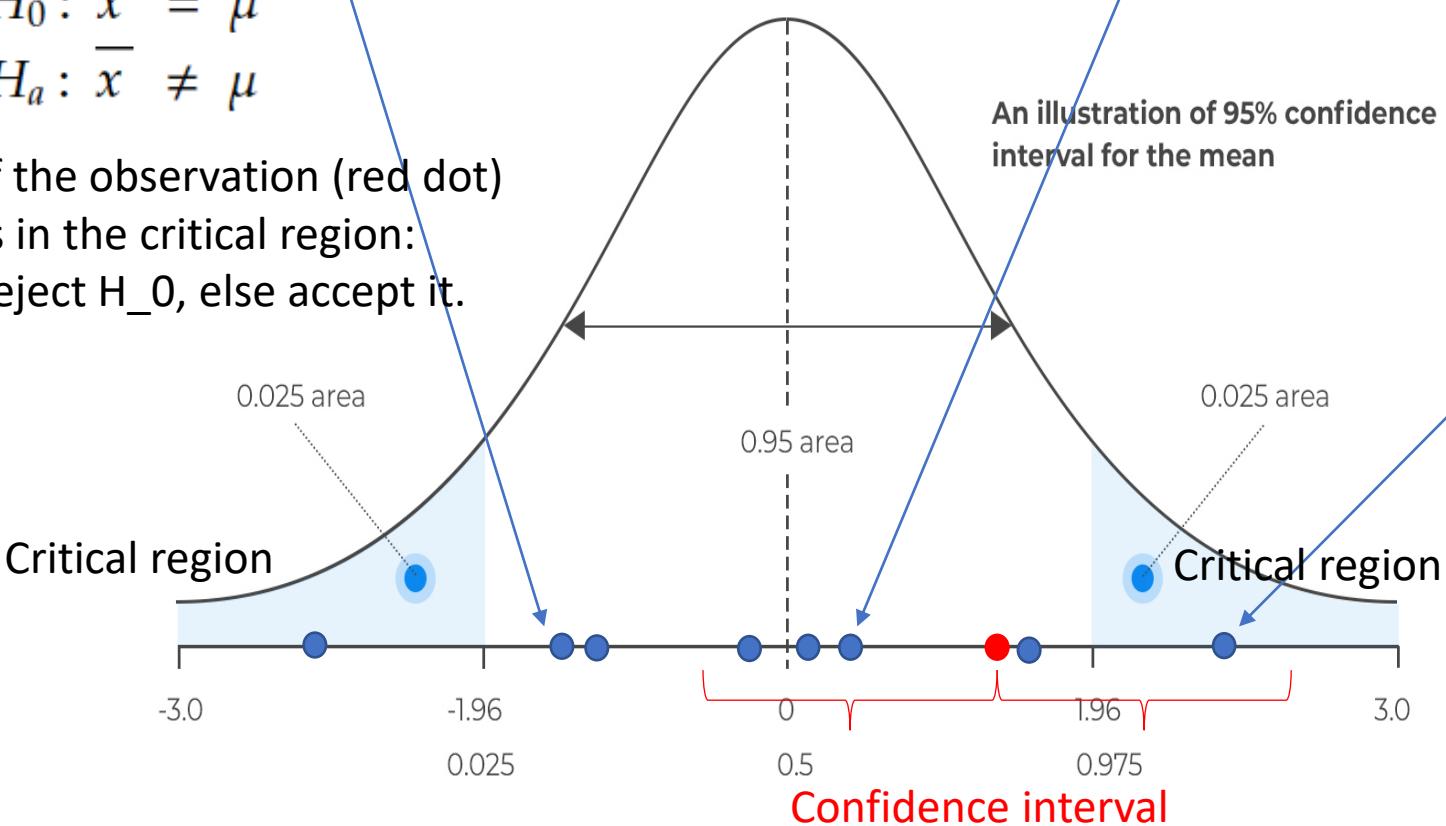
$$x_1 | x_2 | x_2 | \dots | x_n$$

$$\frac{\bar{x} - \mu}{\sigma / \sqrt{n}} \sim \text{Normal} (\text{mean} = 0, \text{stdev} = 1)$$

$$H_0: \bar{x} = \mu$$

$$H_a: \bar{x} \neq \mu$$

If the observation (red dot) is in the critical region: reject H_0 , else accept it.



$$\frac{\bar{X} - \mu}{\sigma / \sqrt{n}}$$

$$\frac{\bar{X} - \mu}{\sigma / \sqrt{n}}$$

Z-score

- Suppose the red dot is the only observed value.
- Then the red range is the 95% confidence interval estimate:

$$-1.96 < \frac{\bar{x} - \mu}{\sigma / \sqrt{n}} < 1.96$$

or, equivalently,

$$P \left\{ \bar{X} - 1.96 \frac{\sigma}{\sqrt{n}} < \mu < \bar{X} + 1.96 \frac{\sigma}{\sqrt{n}} \right\} = .95$$

Confidence interval

Two-sided confidence interval assuming stdev is known.

Suppose that X_1, \dots, X_n is a sample from a normal population having unknown mean μ and known variance σ^2 . It has been shown that $\bar{X} = \sum_{i=1}^n X_i/n$ is the maximum likelihood estimator for μ . However, we don't expect that the sample mean \bar{X} will exactly equal μ , but rather that it will "be close." Hence, rather than a point estimate, it is sometimes more valuable to be able to specify an interval for which we have a certain degree of confidence that μ lies within.

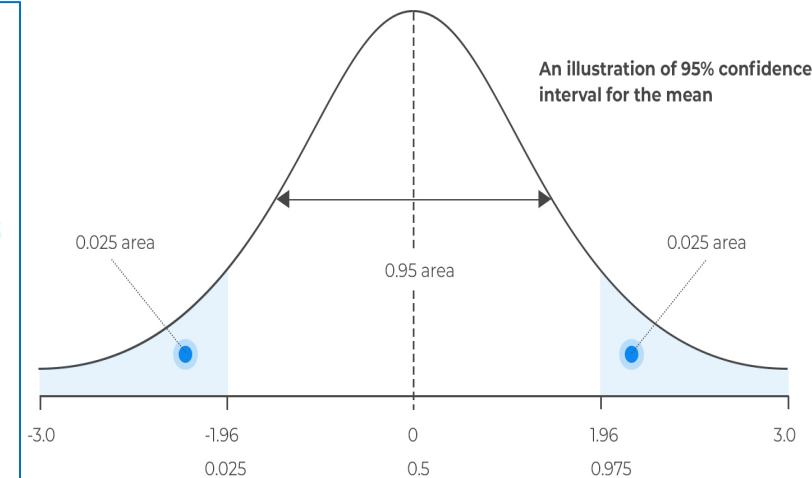
$$\frac{\bar{X} - \mu}{\sigma/\sqrt{n}} = \sqrt{n} \frac{(\bar{X} - \mu)}{\sigma} \text{ has a standard normal distribution. Therefore,}$$

$$P \left\{ -1.96 < \sqrt{n} \frac{(\bar{X} - \mu)}{\sigma} < 1.96 \right\} = .95$$

or, equivalently,

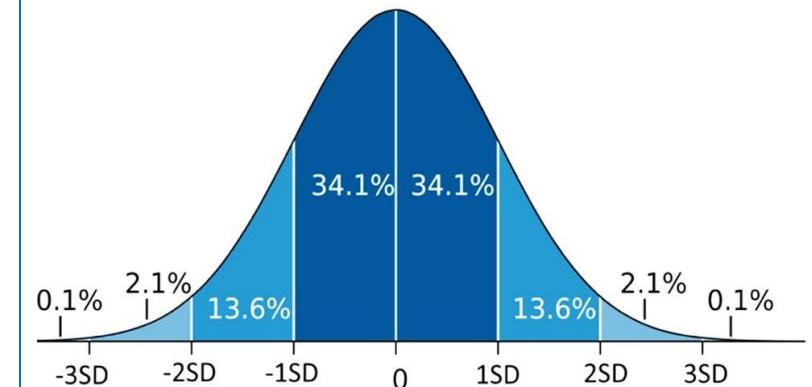
$$P \left\{ \bar{X} - 1.96 \frac{\sigma}{\sqrt{n}} < \mu < \bar{X} + 1.96 \frac{\sigma}{\sqrt{n}} \right\} = .95$$

That is, 95 percent of the time the value of the sample average \bar{X} will be such that the distance between it and the mean μ will be less than $1.96 \sigma/\sqrt{n}$. If we now observe the sample and it turns out that $\bar{X} = \bar{x}$, then we say that "with 95 percent confidence"



That is, "with 95 percent confidence" we assert that the true mean lies within $1.96 \sigma/\sqrt{n}$ of the observed sample mean. The interval

$$\left(\bar{x} - 1.96 \frac{\sigma}{\sqrt{n}}, \bar{x} + 1.96 \frac{\sigma}{\sqrt{n}} \right) \text{ is called a } 95 \text{ percent confidence interval estimate of } \mu.$$



Two-sided and one-sided Confidence Interval

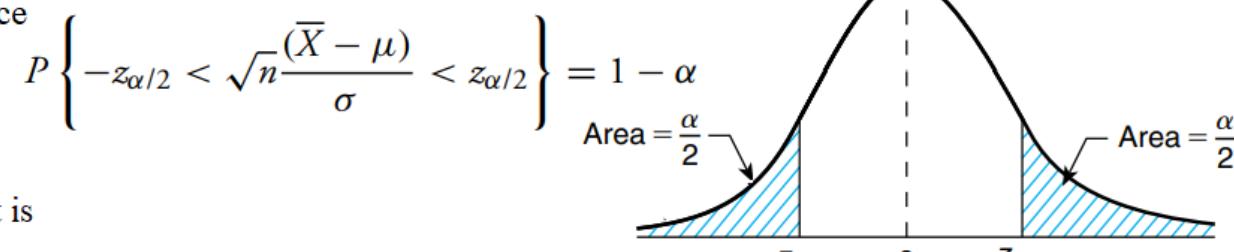
Two-sided vs. one-sided confidence interval assuming stdev is known.

We can also obtain confidence intervals of any specified level of confidence. To do so, recall that z_α is such that

$$P\{Z > z_\alpha\} = \alpha$$

when Z is a standard normal random variable. But this implies (see Figure 7.1) that for any α

$$\text{hence } P\{-z_{\alpha/2} < Z < z_{\alpha/2}\} = 1 - \alpha$$

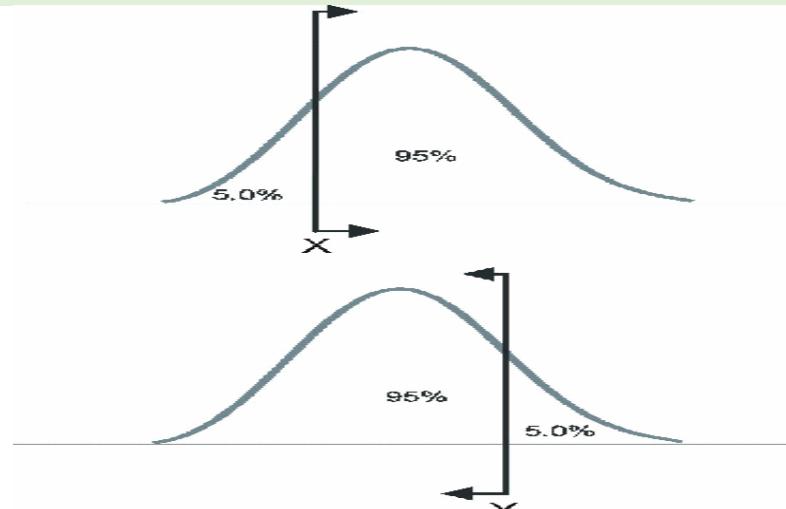


that is

$$P\left\{\bar{X} - z_{\alpha/2} \frac{\sigma}{\sqrt{n}} < \mu < \bar{X} + z_{\alpha/2} \frac{\sigma}{\sqrt{n}}\right\} = 1 - \alpha \quad \underline{P\{-z_{\alpha/2} < Z < z_{\alpha/2}\} = 1 - \alpha.}$$

Hence, a $100(1 - \alpha)$ percent two-sided confidence interval for μ is

$$\left(\bar{x} - z_{\alpha/2} \frac{\sigma}{\sqrt{n}}, \quad \bar{x} + z_{\alpha/2} \frac{\sigma}{\sqrt{n}}\right)$$



Similarly, knowing that $Z = \sqrt{n}(\bar{X} - \mu) / \sigma$ is a standard normal random variable, along with the identities

$$P\{Z > z_\alpha\} = \alpha$$

and

$$P\{Z < -z_\alpha\} = \alpha$$

results in one-sided confidence intervals of any desired level of confidence. Specifically, we obtain that

$$\left(\bar{x} - z_\alpha \frac{\sigma}{\sqrt{n}}, \infty\right)$$

and

$$\left(-\infty, \bar{x} + z_\alpha \frac{\sigma}{\sqrt{n}}\right)$$

are, respectively, $100(1 - \alpha)$ percent one-sided upper and $100(1 - \alpha)$ percent one-sided lower confidence intervals for μ .

Chapter-5 Contents

1. Hypothesis Testing
2. Known variance: Normal Distbn., Unknown variance: T-Distribution
3. Testing equality of means of two Normal populations
4. Tests comparing the Variance of populations

5.1

Hypothesis Testing

When null hypothesis H_0 is True then observations X_i s are very unlikely to be in the critical region C_0

Hence, if X_i s in C_0 then H_0 is very likely False ! Usually C_0 is the region of least 0.05 or 5 % probability

null hypothesis

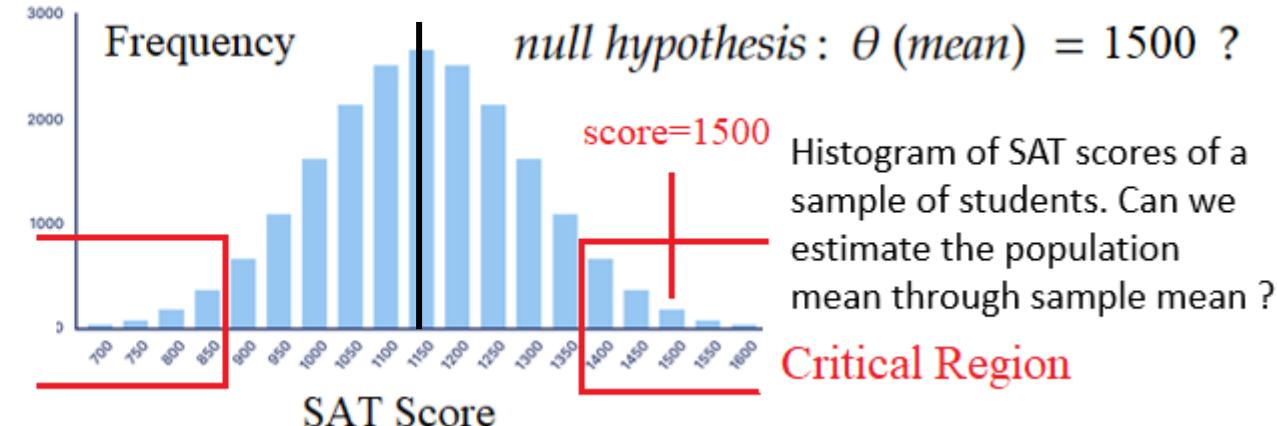
$$(a) H_0 : \theta = 1$$

Simple hypothesis: completely specifies the population distribution

$$(b) H_0 : \theta \leq 1$$

Composite hypothesis: incomplete specification of the distribution

in the figure, 1500 is too far away from the sample mean, and lie in the Critical region. Hence we reject the null hypothesis !



Histogram of SAT scores of a sample of students. Can we estimate the population mean through sample mean ?
Critical Region

Normal distribution with known variance

null hypothesis $H_0 : \mu = \mu_0$ against alternative hypothesis $H_1 : \mu \neq \mu_0$

$$P_{\mu_0} \{ |\bar{X} - \mu_0| > c \} = \alpha$$

$$\text{or } P_{\mu_0} \left\{ \frac{|\bar{X} - \mu_0|}{\sigma/\sqrt{n}} > \frac{c}{\sigma/\sqrt{n}} \right\} = \alpha$$

$$\text{or } P_{\mu_0} \left\{ |Z| > \frac{c}{\sigma/\sqrt{n}} \right\} = \alpha$$

$$\text{or } 2 P_{\mu_0} \left\{ Z > \frac{c}{\sigma/\sqrt{n}} \right\} = \alpha$$

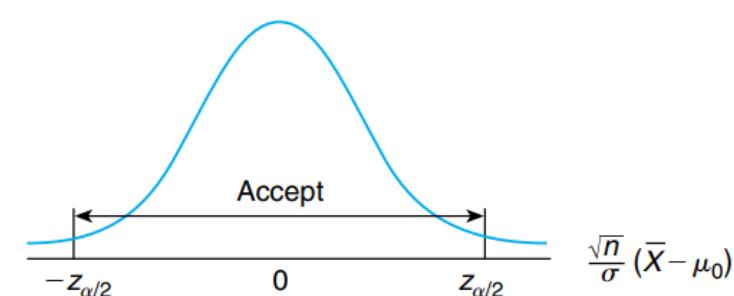
$$\text{or } P_{\mu_0} \left\{ Z > \frac{c}{\sigma/\sqrt{n}} \right\} = \alpha/2$$

$$\text{and } P_{\mu_0} \{ Z > z_{\alpha/2} \} = \alpha/2$$

$$\Rightarrow \frac{c}{\sigma/\sqrt{n}} = z_{\alpha/2} \text{ or } c = \frac{z_{\alpha/2} \sigma}{\sqrt{n}}$$

reject H_0 if $\frac{\sqrt{n}}{\sigma} |\bar{X} - \mu_0| > z_{\alpha/2}$

accept H_0 if $\frac{\sqrt{n}}{\sigma} |\bar{X} - \mu_0| \leq z_{\alpha/2}$



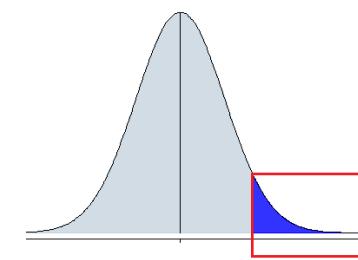
One sided test

$H_0 : \mu = \mu_0$ versus $H_1 : \mu > \mu_0$

$$P_{\mu_0} \{ \bar{X} - \mu_0 > c \} = \alpha$$

accept H_0 if $\frac{\sqrt{n}}{\sigma} (\bar{X} - \mu_0) \leq z_{\alpha}$

reject H_0 if $\frac{\sqrt{n}}{\sigma} (\bar{X} - \mu_0) > z_{\alpha}$



5.2

Known variance: Normal Distbn., Unknown variance: T-Distribution

- Reject null hypothesis if p-value < α (level of significance)
- Smaller the p-value worse the fit

 X_1, \dots, X_n Is a Sample from a $\mathcal{N}(\mu, \sigma^2)$ Population

$$\sigma^2 \text{ Is Known, } \bar{X} = \sum_{i=1}^n X_i/n$$

if σ is unknown, estimate it by S

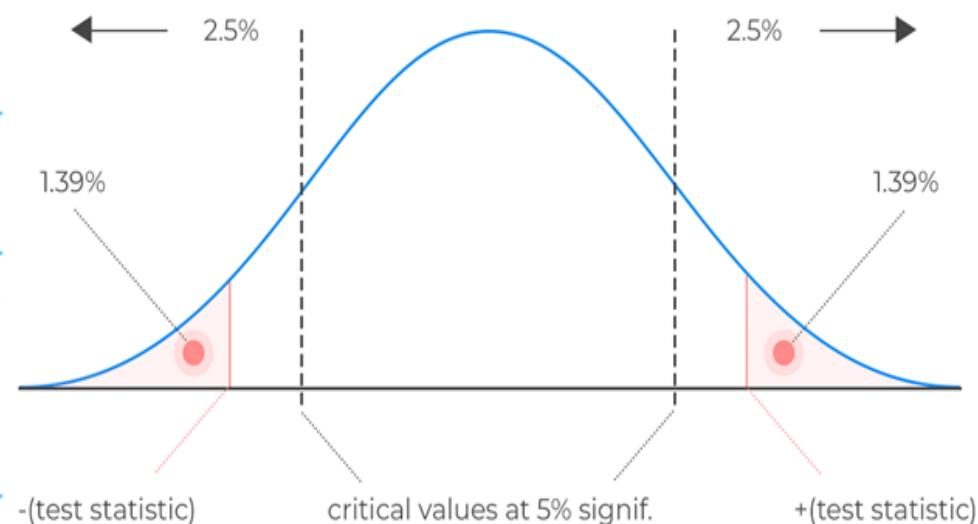
$$S^2 = \frac{\sum_{i=1}^n (X_i - \bar{X})^2}{n - 1} \quad \text{hence}$$

H_0	H_1	Test Statistic TS	Significance Level α	Test	p-Value if $TS = t$
$\mu = \mu_0$	$\mu \neq \mu_0$	$\sqrt{n}(\bar{X} - \mu_0)/\sigma$		Reject if $ TS > z_{\alpha/2}$	$2P\{Z \geq t \}$
$\mu \leq \mu_0$	$\mu > \mu_0$	$\sqrt{n}(\bar{X} - \mu_0)/\sigma$		Reject if $TS > z_\alpha$	$P\{Z \geq t\}$
$\mu \geq \mu_0$	$\mu < \mu_0$	$\sqrt{n}(\bar{X} - \mu_0)/\sigma$		Reject if $TS < -z_\alpha$	$P\{Z \leq t\}$

 Z is a standard normal random variable. X_1, \dots, X_n Is a Sample from a $\mathcal{N}(\mu, \sigma^2)$ Population

$$\sigma^2 \text{ Is Unknown, } \bar{X} = \sum_{i=1}^n X_i/n \quad S^2 = \sum_{i=1}^n (X_i - \bar{X})^2/(n - 1)$$

H_0	H_1	Test Statistic TS	Significance Level α	Test	p-Value if $TS = t$
$\mu = \mu_0$	$\mu \neq \mu_0$	$\sqrt{n}(\bar{X} - \mu_0)/S$		Reject if $ TS > t_{\alpha/2, n-1}$	$2P\{T_{n-1} \geq t \}$
$\mu \leq \mu_0$	$\mu > \mu_0$	$\sqrt{n}(\bar{X} - \mu_0)/S$		Reject if $TS > t_{\alpha, n-1}$	$P\{T_{n-1} \geq t\}$
$\mu \geq \mu_0$	$\mu < \mu_0$	$\sqrt{n}(\bar{X} - \mu_0)/S$		Reject if $TS < -t_{\alpha, n-1}$	$P\{T_{n-1} \leq t\}$

 T_{n-1} is a t-random variable with $n - 1$ degrees of freedom: $P\{T_{n-1} > t_{\alpha, n-1}\} = \alpha$.

Testing equality of means of two Normal populations

Paired t-test can work even if the two populations are not independent and have different variances.
 Separate tests for i) known variance, ii) unknown but equal variance, iii) unknown, unequal variance.

THE PAIRED t-TEST

Before	After	$A - B$
30.5	23	-7.5
18.5	21	2.5
24.5	22	-2.5
32	28.5	-3.5
16	14.5	-1.5
15	15.5	.5
$W_i = X_i - Y_i, i = 1, \dots, n.$		
$H_0 : \mu_w = 0$ versus $H_1 : \mu_w \neq 0$		
normal population having unknown mean and unknown variance		

X_1, \dots, X_n Is a Sample from a $\mathcal{N}(\mu_1, \sigma_1^2)$ Population; Y_1, \dots, Y_m Is a Sample from a $\mathcal{N}(\mu_2, \sigma_2^2)$ Population

The Two Population Samples Are Independent to Test

$$H_0 : \mu_1 = \mu_2 \text{ versus } H_0 : \mu_1 \neq \mu_2$$

Assumption	Test Statistic TS	Significance Level α	Test	p-Value if $TS = t$
σ_1, σ_2 known	$\frac{\bar{X} - \bar{Y}}{\sqrt{\sigma_1^2/n + \sigma_2^2/m}}$		Reject if $ TS > z_{\alpha/2}$	$2P\{Z \geq t \}$
$\sigma_1 = \sigma_2$	$\frac{\bar{X} - \bar{Y}}{\sqrt{\frac{(n-1)s_1^2 + (m-1)s_2^2}{n+m-2} \sqrt{1/n + 1/m}}}$		Reject if $ TS > t_{\alpha/2, n+m-2}$	$2P\{T_{n+m-2} \geq t \}$
n, m large	$\frac{\bar{X} - \bar{Y}}{\sqrt{s_1^2/n + s_2^2/m}}$		Reject if $ TS > z_{\alpha/2}$	$2P\{Z \geq t \}$

accepting H_0 if $-\bar{t}_{\alpha/2, n-1} < \sqrt{n} \frac{\bar{W}}{S_w} < \bar{t}_{\alpha/2, n-1}$
 rejecting H_0 otherwise

Tests comparing the Variance of populations

$$H_0 : \sigma^2 = \sigma_0^2 \quad \text{vs.} \quad H_1 : \sigma^2 \neq \sigma_0^2$$

$$\frac{(n-1)S^2}{\sigma_0^2} \sim \chi^2_{n-1}$$

$$P_{H_0} \left\{ \chi^2_{1-\alpha/2, n-1} \leq \frac{(n-1)S^2}{\sigma_0^2} \leq \chi^2_{\alpha/2, n-1} \right\} = 1 - \alpha$$

Therefore, a significance level α test is to

$$\text{accept } H_0 \quad \text{if} \quad \chi^2_{1-\alpha/2, n-1} \leq \frac{(n-1)S^2}{\sigma_0^2} \leq \chi^2_{\alpha/2, n-1}$$

reject H_0 otherwise

$$H_0 : \sigma_x^2 = \sigma_y^2 \quad \text{versus} \quad H_1 : \sigma_x^2 \neq \sigma_y^2$$

$$S_x^2 = \frac{\sum_{i=1}^n (X_i - \bar{X})^2}{n-1} \quad S_y^2 = \frac{\sum_{i=1}^m (Y_i - \bar{Y})^2}{m-1}$$

denote the sample variances, then as shown in Section 6.5, $(n-1)S_x^2/\sigma_x^2$ and $(m-1)S_y^2/\sigma_y^2$ are independent chi-square random variables with $n-1$ and $m-1$ degrees of freedom, respectively. Therefore, $(S_x^2/\sigma_x^2)/(S_y^2/\sigma_y^2)$ has an F -distribution with parameters $n-1$ and $m-1$. Hence, when H_0 is true

$$S_x^2/S_y^2 \sim F_{n-1, m-1}$$

$$S_x^2/S_y^2 \sim F_{n-1, m-1} \quad \text{hence}$$

$$P_{H_0} \{ F_{1-\alpha/2, n-1, m-1} \leq S_x^2/S_y^2 \leq F_{\alpha/2, n-1, m-1} \} = 1 - \alpha$$

hence

$$\text{accept } H_0 \quad \text{if} \quad F_{1-\alpha/2, n-1, m-1} < S_x^2/S_y^2 < F_{\alpha/2, n-1, m-1}$$

reject H_0 otherwise

Chapter-6 Contents

1. One way - Analysis of Variance (ANOVA)
2. Chi-squared test – test independence of two categorical columns

Get two estimates of variance: i. When assuming all group means are same, ii. irrespective of the group means being same or not. If they differ much then the group means are not same.

Consider m independent samples, each of size n , where the members of the i th sample — $X_{i1}, X_{i2}, \dots, X_{in}$ — are normal random variables with unknown mean μ_i and unknown variance σ^2 . That is,

$$X_{ij} \sim N(\mu_i, \sigma^2), \quad i = 1, \dots, m, \quad j = 1, \dots, n$$

We will be interested in testing $H_0 : \mu_1 = \mu_2 = \dots = \mu_m$ Let $\mu_i = \mu$ for all i .
 versus $H_1 : \text{not all the means are equal}$

expected value of a chi-square random variable is equal to its number of degrees of freedom

$$E(X_n^2) = n$$

$$\sum_{i=1}^m \sum_{j=1}^n (X_{ij} - E[X_{ij}])^2 / \sigma^2 = \sum_{i=1}^m \sum_{j=1}^n (X_{ij} - \mu_i)^2 / \sigma^2 \sim \chi_{nm}^2$$

Let $X_{i\cdot} = \sum_{j=1}^n X_{ij} / n$ then $\sum_{i=1}^m \sum_{j=1}^n (X_{ij} - X_{i\cdot})^2 / \sigma^2 \sim \chi_{nm-m}^2$

Sum of squares of $n*m$ std norm vars is chi^2 with $n*m$ degrees of freedom

1 degree of freedom is lost for each parameter that is estimated

$$\text{Let } SS_W = \sum_{i=1}^m \sum_{j=1}^n (X_{ij} - X_{i\cdot})^2 \text{ then } E[SS_W] / \sigma^2 = nm - m \quad \text{or} \quad E[SS_W / (nm - m)] = \sigma^2$$

$SS_W / (nm - m)$ is an estimator of σ^2 .
 irrespective of H_0 being true or not

$$\text{if } H_0 \text{ is true then } \frac{\sqrt{n}(X_{i\cdot} - \mu)}{\sigma} \text{ is Std. Normal. Hence, sum of squares of } n \sum_{i=1}^m (X_{i\cdot} - \mu)^2 / \sigma^2 \sim \chi_m^2 \text{ the estimator of } \mu \text{ is } X_{..} = \frac{\sum_{i=1}^m \sum_{j=1}^n X_{ij}}{nm} = \frac{\sum_{i=1}^m X_{i\cdot}}{m}$$

$$\text{hence } n \sum_{i=1}^m (X_{i\cdot} - X_{..})^2 / \sigma^2 \sim \chi_{m-1}^2 \quad \text{Let } SS_b = n \sum_{i=1}^m (X_{i\cdot} - X_{..})^2 \quad \text{when } H_0 \text{ is true, } E[SS_b] / \sigma^2 = m - 1 \quad \text{or} \quad E[SS_b / (m - 1)] = \sigma^2$$

$n \sum_{i=1}^m (X_{i\cdot} - X_{..})^2 / \sigma^2 \sim \chi_{m-1}^2$ is called the between samples sum of squares.

6.1

One way - Analysis of Variance (ANOVA)

H0: all group means are same.

- Calculate the test statistic and p-value.
- Reject H0 if p-value < 0.05 (95% confidence interval)

Thus we have shown that

$SS_W/(nm - m)$ always estimates σ^2

$SS_b/(m - 1)$ estimates σ^2 when H_0 is true

Because it can be shown that $SS_b/(m - 1)$ will tend to exceed σ^2 when H_0 is not true, let the test statistic

be given by

$$TS = \frac{SS_b/(m - 1)}{SS_W/(nm - m)} \sim F_{m-1, nm-m}$$

reject H_0 when TS is sufficiently large.

The significance level α test of H_0 is as follows:

reject H_0 if $\frac{SS_b/(m - 1)}{SS_W/(nm - m)} > F_{m-1, nm-m, \alpha}$

do not reject H_0 otherwise

If the value of the test statistic is $TS = v$,
then the p-value will be given by $p\text{-value} = P\{F_{m-1, nm-m} \geq v\}$

Chi-squared test – test independence of two categorical columns

H_0 : X and Y are independent. If H_0 is rejected, we conclude that the two columns are dependent/correlated. Hence columns with minimum p-values are the most important ones!

X	Y
1	9
1	8
1	9
2	6
3	3
3	6
4	8
...	...

Suppose that there are r possible values for the X – characteristic, and s for the Y – characteristic, and let $P_{ij} = P\{X = i, Y = j\}$ for $i = 1, \dots, r$, $j = 1, \dots, s$.

Let $p_i = P\{X = i\} = \sum_j P_{ij}$ and $q_j = P\{Y = j\} = \sum_i P_{ij}$

We are interested in testing:

$H_0: P_{ij} = p_i q_j$ for all i, j vs $H_a: P_{ij} \neq p_i q_j$ for some i, j

Solution: Let $N_i = \sum_j N_{ij}$ then $\hat{p}_i = \frac{N_i}{n}$ $i = 1, \dots, r$

Let $M_j = \sum_i N_{ij}$ then $\hat{q}_j = \frac{M_j}{n}$ $j = 1, \dots, s$

$$T = \sum_{j=1}^s \sum_{i=1}^r \frac{(N_{ij} - n\hat{p}_i \hat{q}_j)^2}{n\hat{p}_i \hat{q}_j} = \sum_{j=1}^s \sum_{i=1}^r \frac{N_{ij}^2}{n\hat{p}_i \hat{q}_j} - n$$

$$T = \sum_i \sum_j \frac{(Observed\ frequency - Expected\ frequency)^2}{Expected\ frequency}$$

reject H_0 if $T \geq \chi_{\alpha, (r-1)(s-1)}^2$

not reject H_0 otherwise

Let $T = t$, then

$$p-value = P\{\chi_{\alpha, (r-1)(s-1)}^2 \geq t\}$$

at 95% confidence interval i.e

at 5% level of significance

$$p-value = P\{\chi_{0.05, (r-1)(s-1)}^2 \geq t\}$$

Question: is gender independent of political affiliation ?

EXAMPLE 11.4a A sample of 300 people was randomly chosen, and the sampled individuals were classified as to their gender and political affiliation, Democrat, Republican, or Independent. The following table, called a *contingency table*, displays the resulting data.

<i>i</i>	<i>j</i>			Total
	Democrat	Republican	Independent	
Women	68	56	32	156
Men	52	72	20	144
Total	120	128	52	300

Thus, for instance, the contingency table indicates that the sample of size 300 contained 68 women who classified themselves as Democrats, 56 women who classified themselves as Republicans, and 32 women who classified themselves as Independents; that is, $N_{11} = 68$, $N_{12} = 56$, and $N_{13} = 32$. Similarly, $N_{21} = 52$, $N_{22} = 72$, and $N_{23} = 20$.

Use these data to test the hypothesis that a randomly chosen individual's gender and political affiliation are independent.

6.2 Chi-squared test – test independence of two categorical columns : question-1

Question: is gender independent of political affiliation ?

Answer: No, political affiliation is dependent on gender.

$n\hat{p}_i\hat{q}_j = N_iM_j/n$ are as follows:

$$\frac{N_1M_1}{n} = \frac{156 \times 120}{300} = 62.40$$

$$\frac{N_1M_2}{n} = \frac{156 \times 128}{300} = 66.56$$

$$\frac{N_1M_3}{n} = \frac{156 \times 52}{300} = 27.04$$

$$\frac{N_2M_1}{n} = \frac{144 \times 120}{300} = 57.60$$

$$\frac{N_2M_2}{n} = \frac{144 \times 128}{300} = 61.44$$

$$\frac{N_2M_3}{n} = \frac{144 \times 52}{300} = 24.96$$

The value of the test statistic is thus

$$TS = \frac{(68 - 62.40)^2}{62.40} + \frac{(56 - 66.56)^2}{66.56} + \frac{(32 - 27.04)^2}{27.04} + \frac{(52 - 57.60)^2}{57.60} \\ + \frac{(72 - 61.44)^2}{61.44} + \frac{(20 - 24.96)^2}{24.96} \\ = 6.433$$

Since $(r - 1)(s - 1) = 2$, we must compare the value of TS with the critical value $\chi^2_{0.05,2}$.
From Table A2

$$\chi^2_{0.05,2} = 5.991$$

Since $TS \geq 5.991$, the null hypothesis is rejected at the 5 percent level of significance.
That is, the hypothesis that gender and political affiliation of members of the population
are independent is rejected at the 5 percent level of significance. ■

Chapter-7 Contents

1. Linear regression (linear in weights)
2. MLE solution: Moore Penrose pseudo inverse – Least square approximation
3. Evaluation: R^2 - higher the better
4. Gradient Descent : 1st order approximation of Taylor's theorem
5. Sequential Gradient Descent Optimization
6. Regularized Least Squares Error (weight decay)
7. Underfitting and Overfitting
8. Bootstrap Aggregation (Bagging) – reduces both Bias and Variance
9. Bagging demo
10. Linear Regression: Loss = Modeling loss + Noise
11. Bias – Variance tradeoff !
12. Regularization – balance between Bias and Variance
13. Kernel Regression

7.1 Linear regression (linear in weights)

Target (t) = Linear ($w \cdot x$) + GaussianNoise

$E(t|x) = y = w \cdot x$ (Linear in w , not in x)

Assumption: $t = y(w, x) + \epsilon$ where $\epsilon \sim N(0, \sigma^2)$
 or $t = N(y(w, x), \sigma^2)$ hence $E[t] = y(w, x)$

$$y(\mathbf{x}, \mathbf{w}) = \sum_{j=0}^{M-1} w_j \phi_j(\mathbf{x}) = \mathbf{w}^T \boldsymbol{\phi}(\mathbf{x})$$

$$y(\mathbf{x}, \mathbf{w}) = w_0 + \sum_{j=1}^{M-1} w_j \phi_j(\mathbf{x})$$

$\phi_j(\mathbf{x})$ are known as *basis functions*.

define an additional dummy ‘basis function’ $\phi_0(\mathbf{x}) = 1$

$$y(\mathbf{x}, \mathbf{w}) = \sum_{j=0}^{M-1} w_j \phi_j(\mathbf{x}) = \mathbf{w}^T \boldsymbol{\phi}(\mathbf{x})$$

common basis functions $\sim x^n, e^x, e^{x^2}, \log(x), \tanh(x)$ e.g. if $\mathbf{x} = \begin{bmatrix} x_1 \\ x_2 \end{bmatrix}$

$$\text{let } y(\mathbf{w}, \mathbf{x}) = w_0 + w_1 x_1 + w_2 x_2 + w_3 x_1^2 + w_4 x_2^2 + w_5 e^{x_1} + w_6 e^{x_2} = \begin{bmatrix} w_0 \\ w_1 \\ \dots \\ w_6 \end{bmatrix}^T \begin{bmatrix} 1 \\ x_1 \\ \dots \\ e^{x_2} \end{bmatrix} = \mathbf{w}^T \boldsymbol{\phi}(\mathbf{x})$$

Assumption: $t = y(\mathbf{w}, \mathbf{x}) + \epsilon$ where $\epsilon \sim N(0, \sigma^2)$

or $t = N(y(\mathbf{w}, \mathbf{x}), \sigma^2)$ hence $E[t|x] = y(\mathbf{w}, \mathbf{x})$

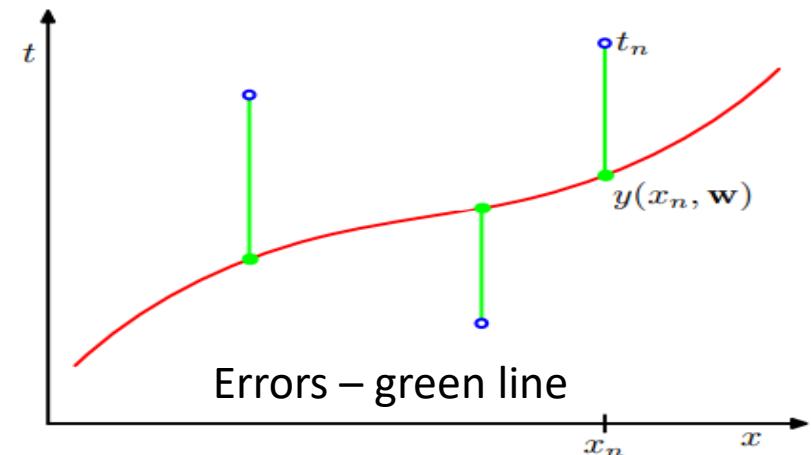
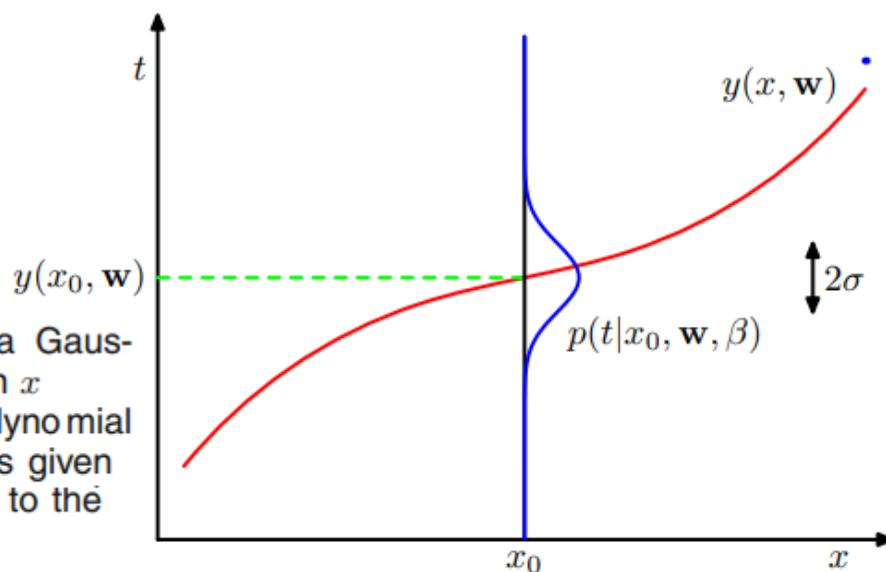


Figure Schematic illustration of a Gaussian conditional distribution for t given x in which the mean is given by the polynomial function $y(x, \mathbf{w})$, and the precision is given by the parameter β , which is related to the variance by $\beta^{-1} = \sigma^2$.



- Moore Penrose pseudo inverse can go out of memory. - It represents a least square projection of targets T (n -dim) onto the vector-subspace spawned by columns of X (k -dim): $k < n$

$$\text{Let } X = \begin{bmatrix} x_{10} & x_{11} & x_{12} & .. & x_{2k} \\ x_{20} & x_{21} & x_{22} & .. & x_{2k} \\ : & : & : & :: & : \\ x_{n0} & x_{n1} & x_{n2} & .. & x_{nk} \end{bmatrix} T = \begin{bmatrix} t_1 \\ t_2 \\ : \\ t_n \end{bmatrix}$$

sum-of-squares error function is defined by

$$E_D(\mathbf{w}) = \frac{1}{2} \sum_{n=1}^N \{t_n - \mathbf{w}^T \phi(\mathbf{x}_n)\}^2.$$

*This is also the
Max.
Likelihood
Estimate (MLE)
of W .*

then MLE estimate for W s.t. $T = W^T X + \epsilon$

is $W_{MLE} = (X^T X)^{-1} X^T T$ where

$(X^T X)^{-1} X^T$ = Moore Penrose pseudo inverse

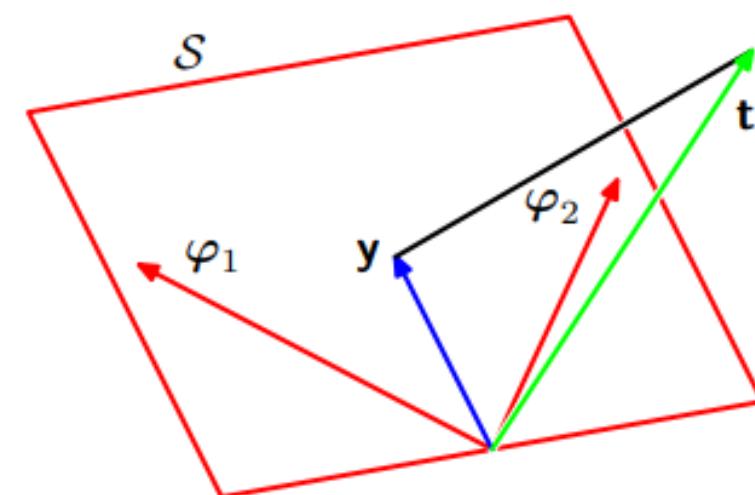
X : large matrix can go out of memory.
Hence, use sequential optimization
approaches as Seq. Gradient Descent.

Online Learning: Sequential learning is
also appropriate for real-time
applications in which the data
observations are arriving in a continuous
stream, and predictions must be made
before all of the data points are seen.

Geometrical interpretation of the least-squares solution, in an N -dimensional space whose axes are the values of t_1, \dots, t_N . The least-squares regression function is obtained by finding the orthogonal projection of the data vector \mathbf{t} onto the subspace spanned by the basis functions $\phi_j(\mathbf{x})$ in which each basis function is viewed as a vector φ_j of length N with elements $\phi_j(\mathbf{x}_n)$.

if features = $\Phi = \begin{pmatrix} \phi_0(\mathbf{x}_1) & \phi_1(\mathbf{x}_1) & \cdots & \phi_{M-1}(\mathbf{x}_1) \\ \phi_0(\mathbf{x}_2) & \phi_1(\mathbf{x}_2) & \cdots & \phi_{M-1}(\mathbf{x}_2) \\ \vdots & \vdots & \ddots & \vdots \\ \phi_0(\mathbf{x}_N) & \phi_1(\mathbf{x}_N) & \cdots & \phi_{M-1}(\mathbf{x}_N) \end{pmatrix}$

Then:



7.3 Evaluation: R^2 - higher the better

R^2 = Proportion of variance explained by the linear model. It is a statistical measure used to evaluate the goodness of fit of a regression model. max R^2 = 1 is the best it can get.

LINEAR REGRESSION IN POPULATION: ANALYSIS OF VARIANCE

USING THE DECOMPOSITION OF Y

$$Y = \beta' X + \epsilon$$

$$\text{NORMAL EQUATIONS } EX\epsilon = 0$$

$$EY^2 = E(\beta' X)^2 + E\epsilon^2$$

DEFINE THE POPULATION MEAN SQUARED PREDICTION ERROR

$$MSE_{pop} = E\epsilon^2$$

$$R^2_{pop} := \frac{E(\beta' X)^2}{EY^2} = 1 - \frac{E\epsilon^2}{EY^2} \in [0, 1]$$

PROPORTION OF VARIATION OF Y EXPLAINED BY THE BLP

TRAINING **V**ALIDATION
 n OBSERVATIONS m OBSERVATIONS
 V INDEXES OBSERVATIONS IN THE TEST/VALIDATION SAMPLE

$$MSE_{test} = \frac{1}{m} \sum_{k \in V} (Y_k - \hat{\beta}' X_k)^2$$

$$R^2_{test} = 1 - \frac{MSE_{test}}{\frac{1}{m} \sum_{k \in V} Y_k^2}$$

Gradient Descent : 1st order approximation of Taylor's theorem

1st order approx. : $f(x) = f(a) + f'(a)(x - a)$ where x is close to a

$x = x - \mu f'(x)$ decreases $f(x)$: Gradient Descent !

If a real-valued function $f(x)$ is differentiable at the point $x = a$, then it has a linear approximation near this point. This means that there exists a function $h_1(x)$ such that

$$f(x) = f(a) + f'(a)(x - a) + h_1(x)(x - a), \quad \lim_{x \rightarrow a} h_1(x) = 0.$$

Here Limiting Taylor's theorem to 1st derivative gives the 1st order approximation

$$P_1(x) = f(a) + f'(a)(x - a) \quad \approx f(x) \text{ if } x \text{ is near } a$$

is the linear approximation of $f(x)$ for x near the point a , whose graph $y = P_1(x)$ is the tangent line to the graph $y = f(x)$ at $x = a$. The error in the approximation is:

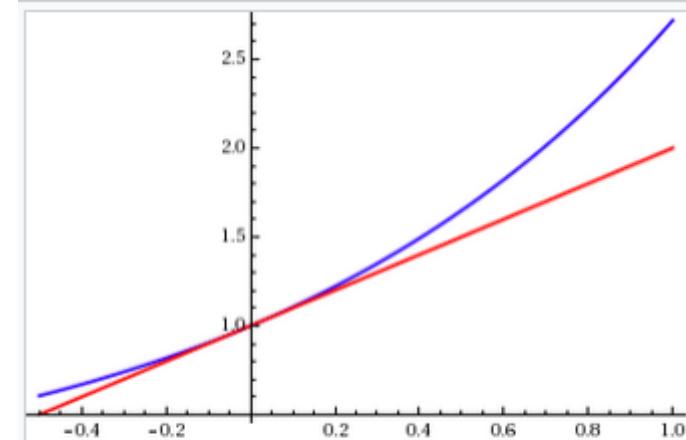
$$R_1(x) = f(x) - P_1(x) = h_1(x)(x - a).$$

As x tends to a , this error goes to zero much faster than $f'(a)(x-a)$, making $f(x) \approx P_1(x)$ a useful approximation.

For a better approximation to $f(x)$, we can fit a quadratic polynomial instead of a linear function:

Limiting Taylor's theorem to 2nd derivative gives the 2nd order approximation

$$P_2(x) = f(a) + f'(a)(x - a) + \frac{f''(a)}{2}(x - a)^2. \quad \approx f(x) \text{ if } x \text{ is near } a$$



Graph of $f(x) = e^x$ (blue) with its linear approximation $P_1(x) = 1 + x$ (red) at $a = 0$.

1st order approx. : $f(x) = f(a) + f'(a)(x - a)$ when x is close to a

Let $x_{\text{new}} = a - \mu f'(a)$ here $\mu \ll 1$ s.t. $\mu f'(a) \equiv 0.01$, then :

$$f(a - \mu f'(a)) = f(a) + f'(a)(a - \mu f'(a) - a)$$

$$\text{or } f(a - \mu f'(a)) = f(a) + f'(a)(-\mu f'(a))$$

$$\text{or } f(a - \mu f'(a)) = f(a) - \mu f'^2(a)$$

now $f'^2(a) \geq 0$ because it is squared, hence $f(a - \mu f'(a)) \leq f(a)$

thus $x = x - \mu f'(x)$ decreases $f(x)$: Gradient Descent !

Gradient Descent : gradually reduce the learning rate

1st order approx. : $f(x) = f(a) + f'(a)(x - a)$ where x is close to a

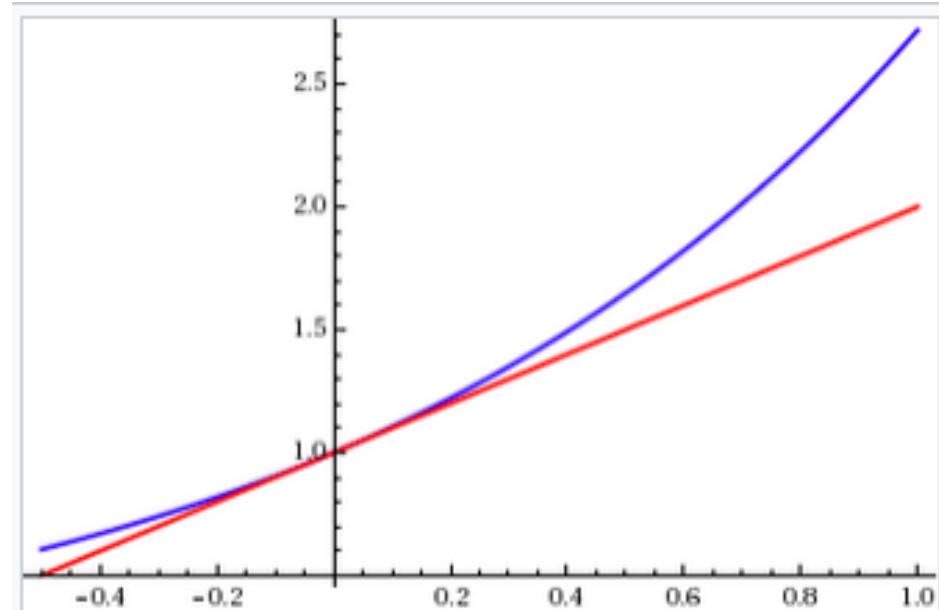
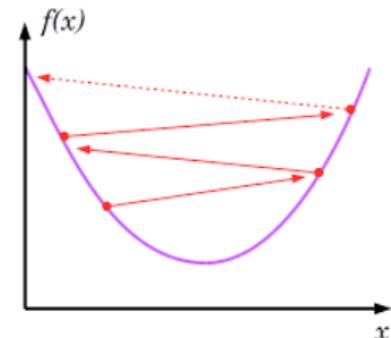
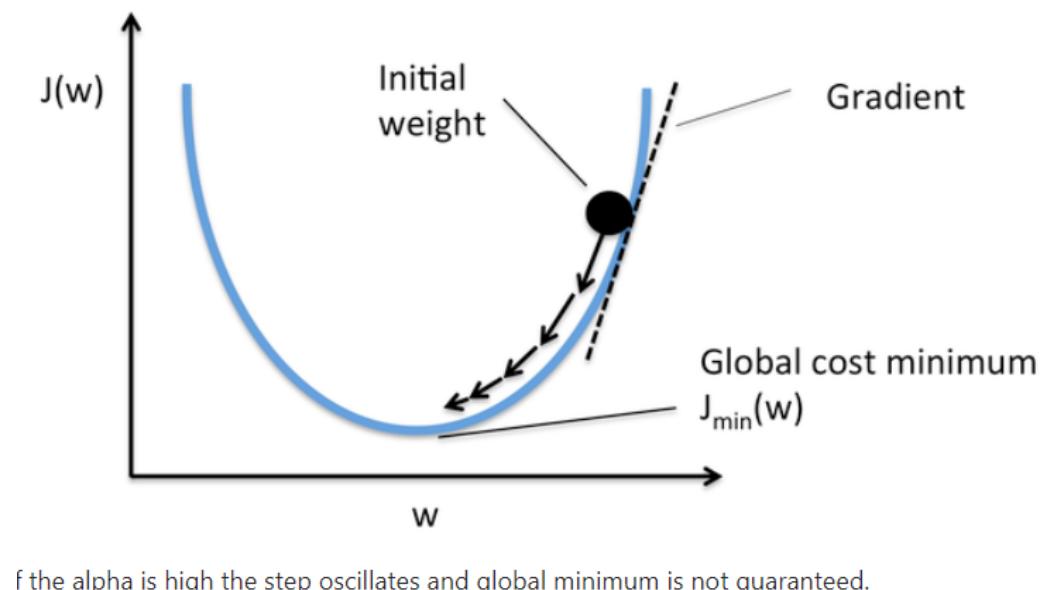
$x = x - \mu f'(x)$ decreases $f(x)$: Gradient Descent !

Limiting Taylor's theorem to 1st derivative gives the 1st order approximation

Limiting Taylor's theorem to 2nd derivative gives the 2nd order approximation

If learning rate (LR)
is high, GD may
never converge.

Hence, gradually
reduce the LR.



Graph of $f(x) = e^x$ (blue) with its linear approximation $P_1(x) = 1 + x$ (red) at $a = 0$.

There are libraries such as JAX which use N-th order approximation for GD optimization.
<https://jax.readthedocs.io/en/latest/index.html>

Sequential Gradient Descent Optimization

Sequential mini-batch gradient descent
can help in online-learning also

$$\mathbf{w}_{T+1} = \mathbf{w}_T + \eta \sum_{n=1}^B \{t_n - \mathbf{w}^T \phi_n\} \phi_n$$

B = no. of samples in a mini batch

error function comprises a sum over data points $E = \sum_n E_n$

Instead of the total E at once, apply E_n sequentially for all x_n :

$$E'_D(\mathbf{w}) = - \sum_{n=1}^N \{t_n - \mathbf{w}^T \phi(x_n)\} \phi(x_n)$$

Gradient Descent: $\mathbf{w}_{n+1} = \mathbf{w}_n + \eta \sum_{n=1}^N \{t_n - \mathbf{w}^T \phi_n\} \phi_n$
 η = learning rate parameter e.g. $\mu = 0.001$ $\phi_n = \phi(x_n)$.

The value of \mathbf{w} is initialized to some starting vector $\mathbf{w}^{(0)}$.

for epoch = 1 to K:

 for n = 1 to N:

$$\mathbf{w}^{(\tau+1)} = \mathbf{w}^{(\tau)} - \eta \nabla E_n \quad \text{or} \quad \mathbf{w}^{(\tau+1)} = \mathbf{w}^{(\tau)} + \eta(t_n - \mathbf{w}^{(\tau)T} \phi_n) \phi_n$$

where τ denotes the iteration number, and η is a learning rate parameter. $\phi_n = \phi(x_n)$.

Sequential Gradient Descent

- Faster, less expensive
- Less overfitting

The value of \mathbf{w} is initialized to some starting vector $\mathbf{w}^{(0)}$.

for epoch = 1 to k:

 for batch = 1 to M:

$$\mathbf{w}^{(\tau+1)} = \mathbf{w}^{(\tau)} - \eta \nabla E_n \quad \text{or} \quad \mathbf{w}_{T+1} = \mathbf{w}_T + \eta \sum_{n=1}^B \{t_n - \mathbf{w}^T \phi_n\} \phi_n$$

B = no. of samples in a mini batch

$B \times M = N$: Total no of training samples

Stochastic (mini batch) Gradient Descent (fastest)

- In between Sequential and full Grad. descent

7.6

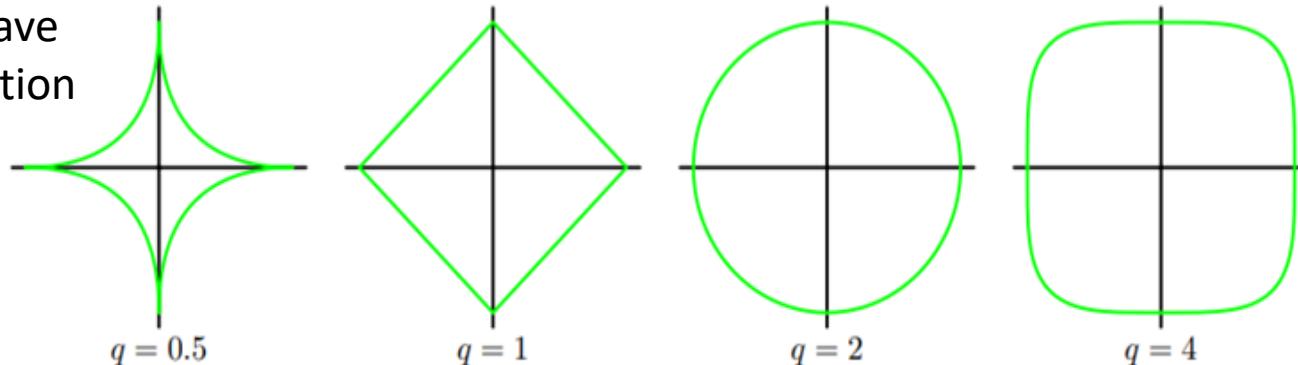
Regularized Least Squares Error (weight decay)

Regularization constrains \mathbf{W} from growing as it penalizes \mathbf{W} .
Thus, it prevents overfitting.

$$\frac{1}{2} \sum_{n=1}^N \{t_n - \mathbf{w}^T \phi(\mathbf{x}_n)\}^2 + \frac{\lambda}{2} \sum_{j=1}^M |w_j|^q$$

$$\text{Err} = \frac{1}{2} \sum_{n=1}^N \{t_n - \mathbf{w}^T \phi(\mathbf{x}_n)\}^2 + \frac{\lambda}{2} \sum_{j=1}^M |w_j|^q$$

If $q = 1$ we have
L1 regularization
(Lasso)



If $q = 2$ we have L2 regularization (Ridge regression):

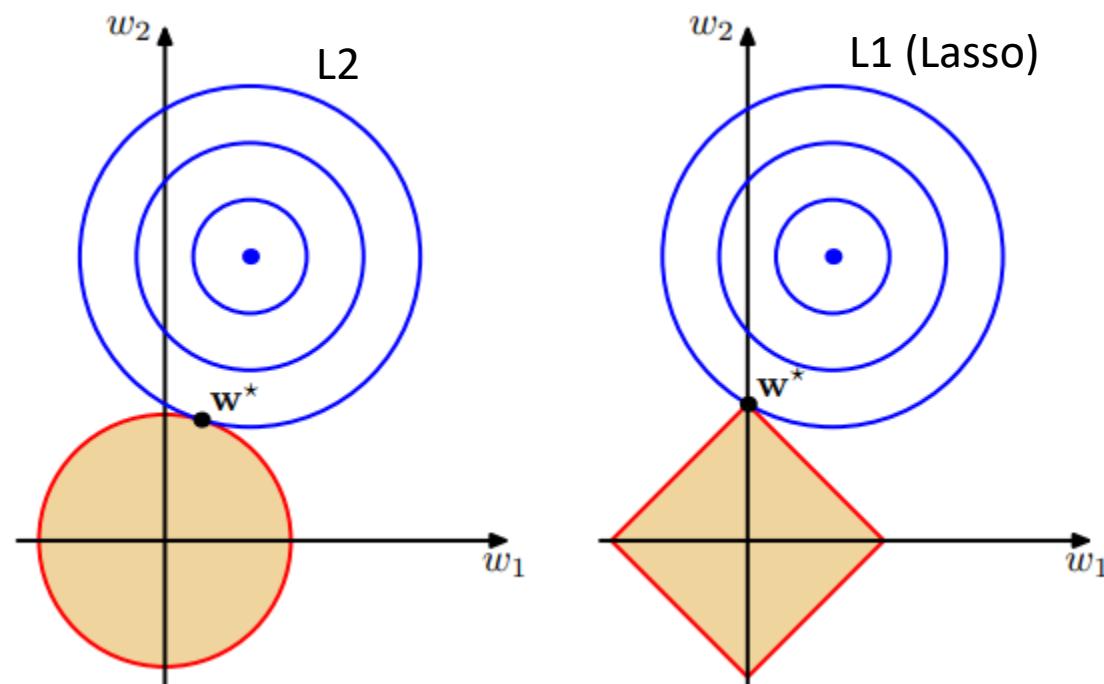
$$\frac{1}{2} \sum_{n=1}^N \{t_n - \mathbf{w}^T \phi(\mathbf{x}_n)\}^2 + \frac{\lambda}{2} \mathbf{w}^T \mathbf{w}.$$

Now $\mathbf{w} = (\lambda \mathbf{I} + \Phi^T \Phi)^{-1} \Phi^T \mathbf{t}$.

L1 regularizer (Lasso) yields sparse \mathbf{W} as it sets most of the not so important dimensions to 0 or very small values close to 0.

(i.e. features which are not so important their weights will be near 0.)

Figure Plot of the contours of the unregularized error function (blue) along with the constraint region (3.30) for the quadratic regularizer $q = 2$ on the left and the lasso regularizer $q = 1$ on the right, in which the optimum value for the parameter vector \mathbf{w} is denoted by \mathbf{w}^* . The lasso gives a sparse solution in which $w_1^* = 0$.

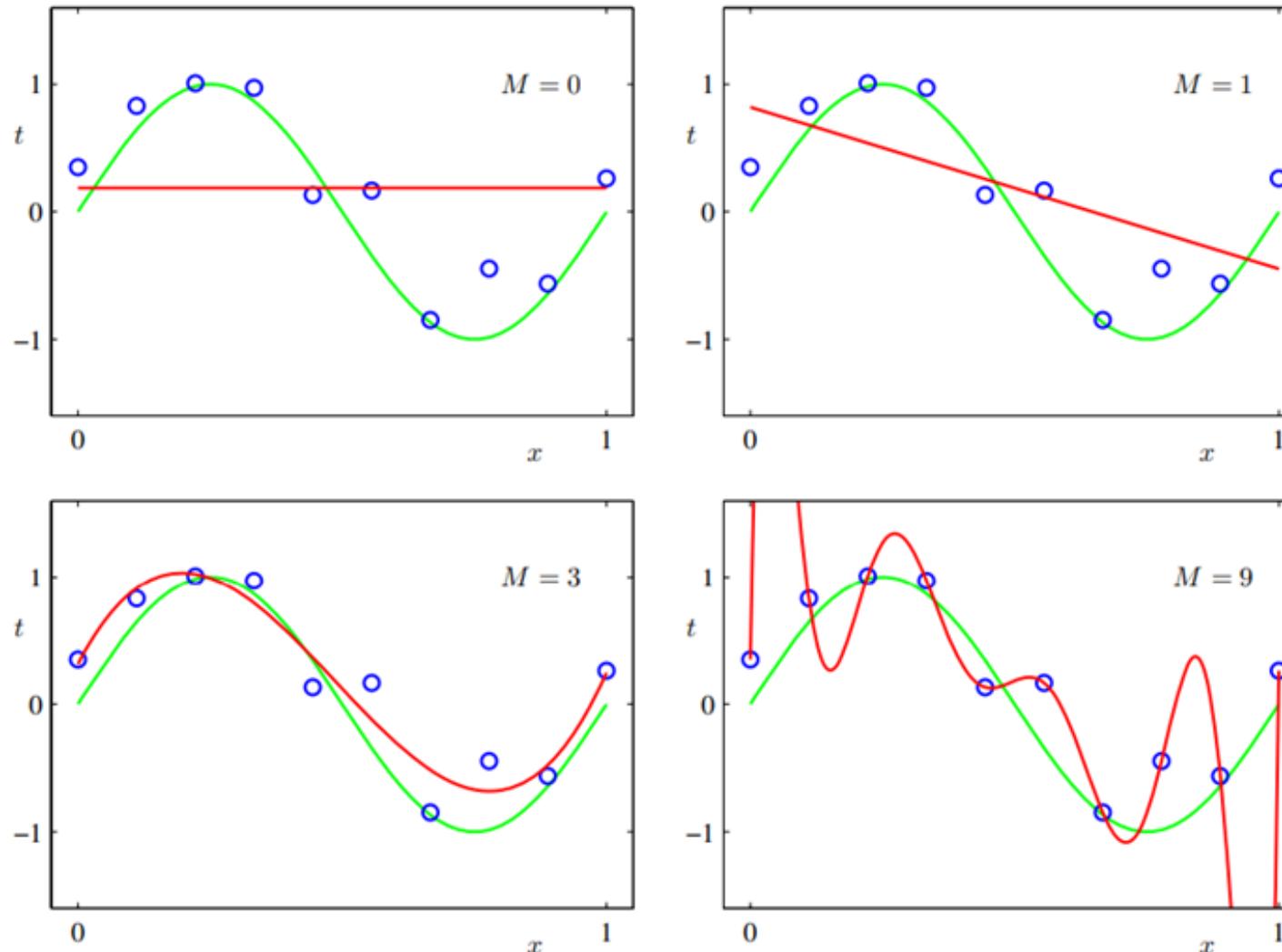


7.7

Underfitting and Overfitting

Underfitting \sim high Bias : Scope to add more complex features.

Overfitting \sim high Variance : Starts memorizing training data and give near 100% accuracy on it.



Graphs of the root-mean-square error, defined by (1.3), evaluated on the training set and on an independent test set for various values of M .

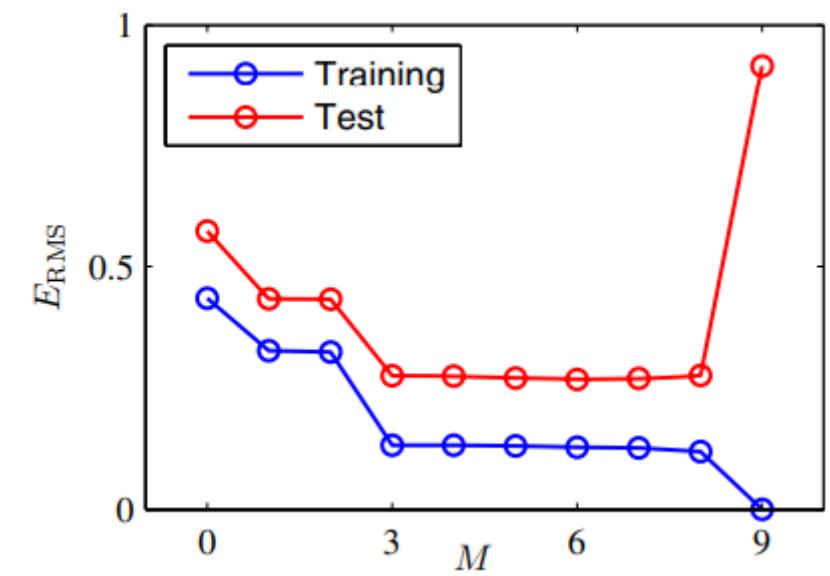
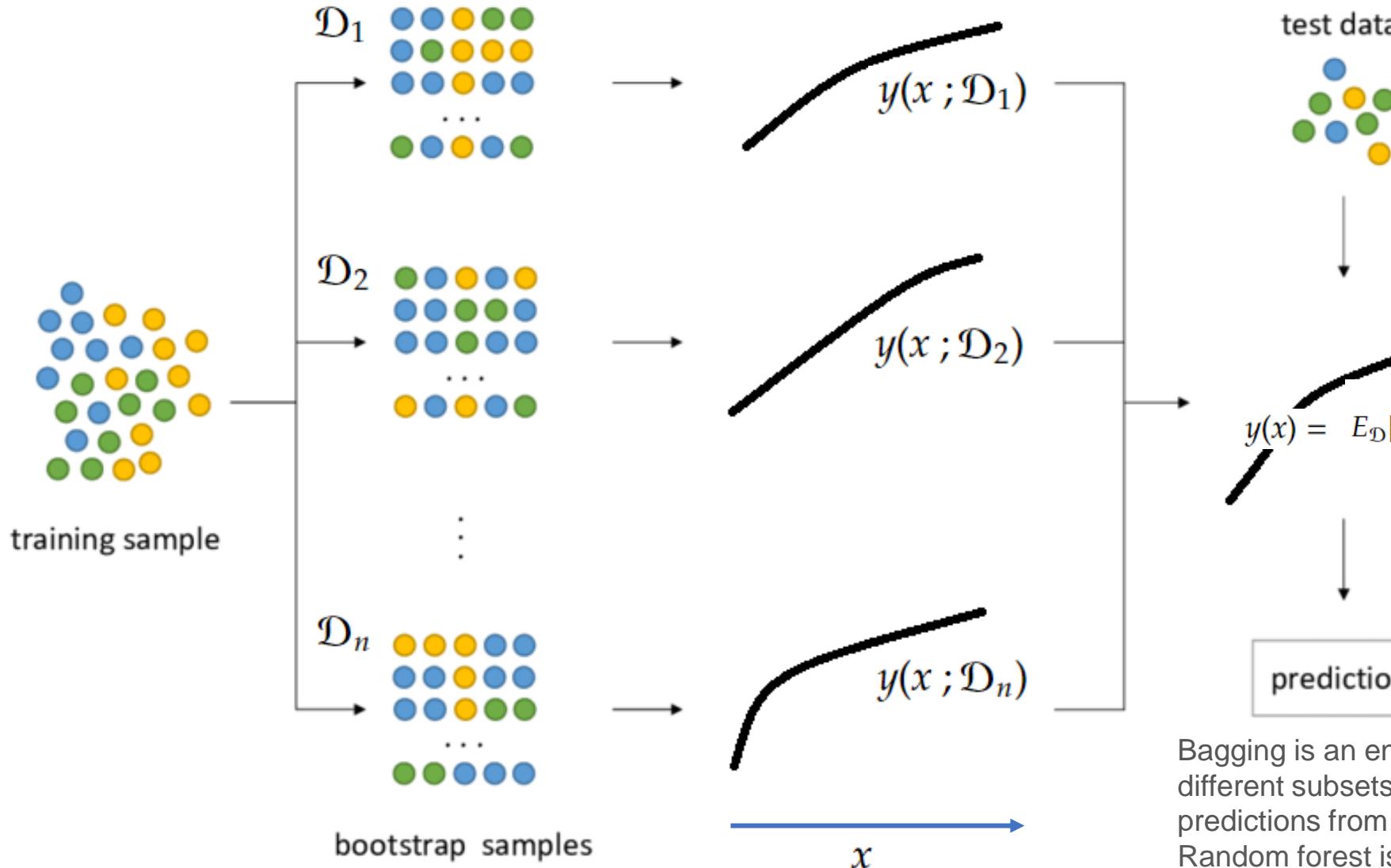


Figure Plots of polynomials having various orders M , shown as red curves

Bootstrap Aggregation (Bagging) – reduces both Bias and Variance

create models $y(x ; \mathcal{D}_1)$ $y(x ; \mathcal{D}_2)$... $y(x ; \mathcal{D}_n)$ on samples \mathcal{D}_1 \mathcal{D}_2 ... \mathcal{D}_n

$$\text{Final model } y(x) = \text{Avg. or Expectation over } \mathcal{D} \text{ of } y(x ; \mathcal{D}) = E_{\mathcal{D}}[y(x ; \mathcal{D})] = \frac{y(x ; \mathcal{D}_1) + y(x ; \mathcal{D}_2) + \dots + y(x ; \mathcal{D}_n)}{n}$$



Let variable $\mathcal{D} = \{\mathcal{D}_1 \ \mathcal{D}_2 \ \dots \ \mathcal{D}_n\}$

Final model $y(x) = \text{Avg. or Expectation over } \mathcal{D} \text{ of } y(x ; \mathcal{D}) = E_{\mathcal{D}}[y(x ; \mathcal{D})]$

$$y(x) = E_{\mathcal{D}}[y(x ; \mathcal{D})] = \frac{y(x ; \mathcal{D}_1) + y(x ; \mathcal{D}_2) + \dots + y(x ; \mathcal{D}_n)}{n}$$

- Averaging minimizes variance in predictions.
- Averaging is costly

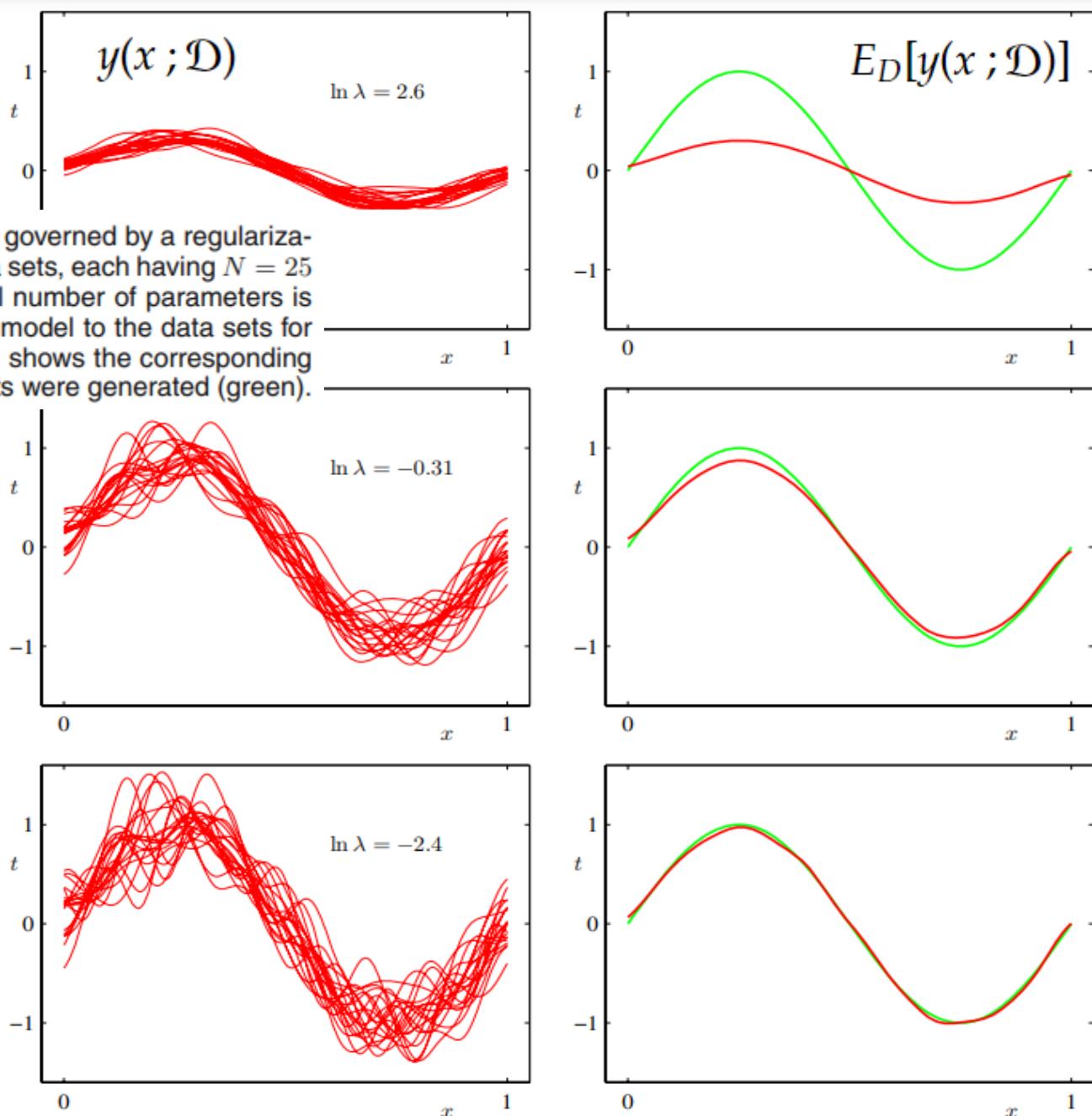
Bagging is an ensemble algorithm that fits multiple models on different subsets of a training dataset, then combines the predictions from all models. Random forest is an extension of bagging that also randomly selects subsets of features used in each data sample.

Ensembel tends not to overfit (less variance)

Ensemble tends to be more accurate (less bias)

Figure Illustration of the dependence of bias and variance on model complexity, governed by a regularization parameter λ , using the sinusoidal data set from Chapter 1. There are $L = 100$ data sets, each having $N = 25$ data points, and there are 24 Gaussian basis functions in the model so that the total number of parameters is $M = 25$ including the bias parameter. The left column shows the result of fitting the model to the data sets for various values of $\ln \lambda$ (for clarity, only 20 of the 100 fits are shown). The right column shows the corresponding average of the 100 fits (red) along with the sinusoidal function from which the data sets were generated (green).

$$E_D[y(x ; \mathcal{D})] = \frac{y(x ; \mathcal{D}_1) + y(x ; \mathcal{D}_2) + \dots + y(x ; \mathcal{D}_n)}{n}$$



7.10 Linear Regression: Loss = Modeling loss + Noise

Least square loss decomposed into :

i) Modeling loss, and ii) Noise. Independent noise signals cannot be predicted.

$$\text{Let } E[t|x] = \int t p(t|x) dt = h(x)$$

$$\begin{aligned} \text{Loss} &= E_{xt}[(y(x) - t)^2] = E_{xt}[(y(x) - E[t|x] + E[t|x] - t)^2] \\ &= E_x[(y(x) - E[t|x])^2] + E_{xt}[(E[t|x] - t)^2] + 2E_{xt}[(y(x) - E[t|x])(E[t|x] - t)] \end{aligned}$$

$$\text{now } 2E_{xt}[(y(x) - E[t|x])(E[t|x] - t)] = 0 \text{ hence :}$$

$$\text{Loss} = E_x[(y(x) - E[t|x])^2] + E_{xt}[(E[t|x] - t)^2]$$

$$\text{Loss} = \underbrace{E_x[(y(x) - h(x))^2]}_{\text{modeling loss}} + \underbrace{E_{xt}[(h(x) - t)^2]}_{\text{Noise}}$$

$$\text{Noise} = E_{xt}[(h(x) - t)^2] = \int_x \int_t (h(x) - t)^2 p(t|x) p(x) dt dx$$

Noise cannot be predicted, hence this is the minimum expected loss !

The loss component which can be minimized is the :

$$\text{modeling loss} = E_x[(y(x) - h(x))^2] = \int_x (y(x) - h(x))^2 p(x) dx$$

min. modeling loss = 0 when prediction $y(x) == h(x) = E[t|x]$

7.11 Bias – Variance tradeoff !

Modeling loss = bias² + variance

If bias decreases, variance can increase and vice-versa.

Hence

$$\text{Loss} = \text{bias}^2 + \text{variance} + \text{noise}$$

If we have multiple trained models across different datasets $\mathcal{D}_1 \mathcal{D}_2.. \mathcal{D}_n$
and instead of avg. the predictions, we avg. the loss, then :

modeling loss = Avg. loss across datasets $\mathcal{D}_1 \mathcal{D}_2.. \mathcal{D}_n = E_{xD}[(y(x; \mathcal{D}) - h(x))^2]$

remember $E_D[y(x; \mathcal{D})] = \frac{y(x; \mathcal{D}_1) + y(x; \mathcal{D}_2) + \dots + y(x; \mathcal{D}_n)}{n}$

$$\begin{aligned} \text{then modeling loss} &= E_{xD}[(y(x; \mathcal{D}) - E_D[y(x; \mathcal{D})])^2] + E_D[y(x; \mathcal{D})] - h(x)]^2 \\ &= E_{xD}[(y(x; \mathcal{D}) - E_D[y(x; \mathcal{D})])^2] + E_x[(E_D[y(x; \mathcal{D})] - h(x))^2] \\ &\quad + 2E_{xD}[(y(x; \mathcal{D}) - E_D[y(x; \mathcal{D})])(E_D[y(x; \mathcal{D})] - h(x))] \end{aligned}$$

the last term = 0 hence

$$\begin{aligned} \text{modeling loss} &= \underbrace{E_{xD}[(y(x; \mathcal{D}) - E_D[y(x; \mathcal{D})])^2]}_{\text{variance}} + \underbrace{E_x[(E_D[y(x; \mathcal{D})] - h(x))^2]}_{\text{bias}^2} \\ &= \underbrace{\int_x E_D[(y(x; \mathcal{D}) - E_D[y(x; \mathcal{D})])^2] p(x) dx}_{\text{variance}} + \underbrace{\int_x (E_D[y(x; \mathcal{D})] - h(x))^2 p(x) dx}_{\text{bias}^2} \end{aligned}$$

Hence

$$\begin{aligned} \text{Loss} &= \text{modeling loss} + \text{noise} \\ &= \text{bias}^2 + \text{variance} + \text{noise} \end{aligned}$$

There is a trade-off between bias and variance, with very flexible models having low bias and high variance, and relatively rigid models having high bias and low variance. The model with the optimal predictive capability is the one that leads to the best balance between bias and variance.

7.12 Regularization – balance between Bias and Variance

Bias: deviation between $h(x)$ and averaged curve

Variance: deviation between $y(x)$ and averaged curve

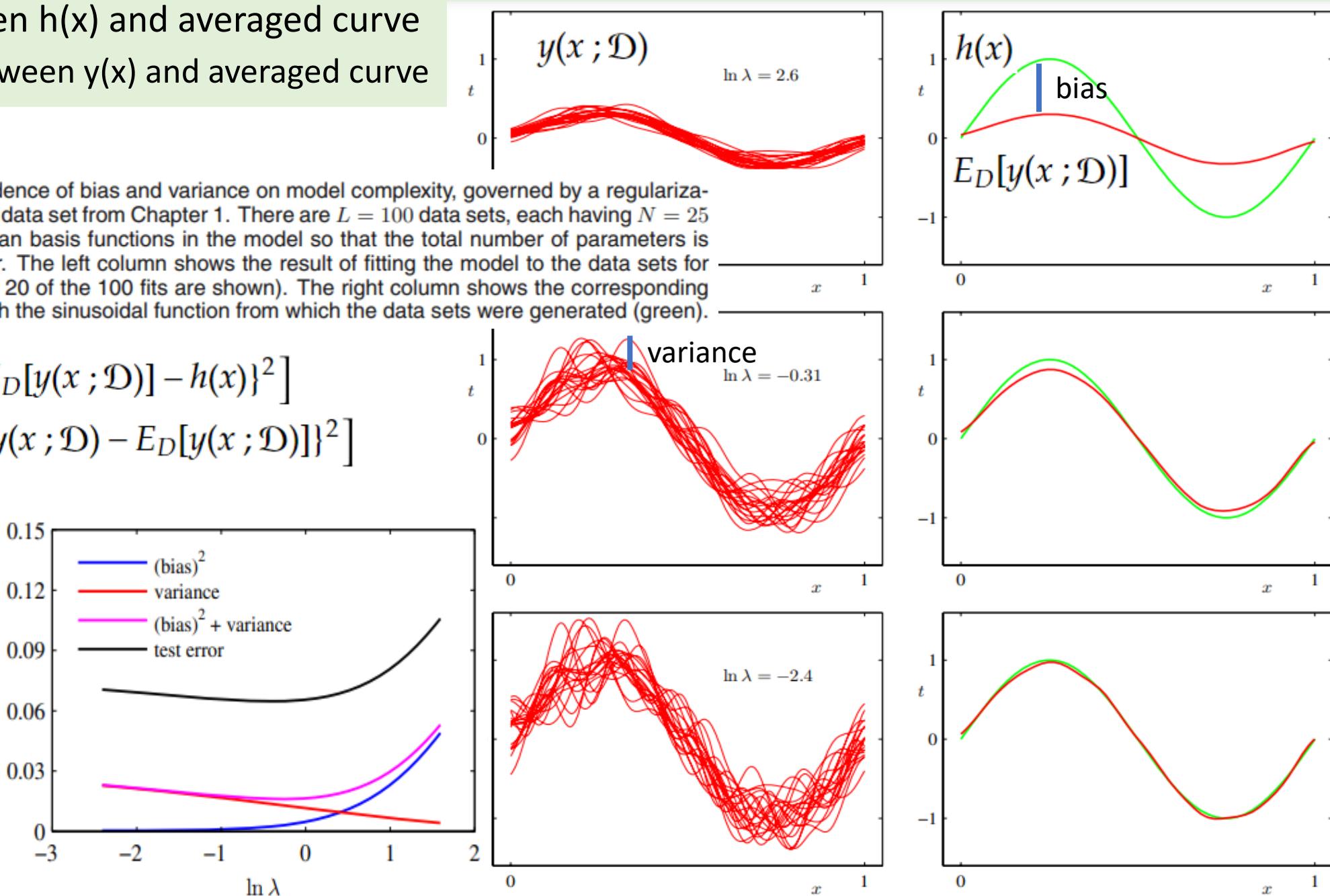
Figure Illustration of the dependence of bias and variance on model complexity, governed by a regularization parameter λ , using the sinusoidal data set from Chapter 1. There are $L = 100$ data sets, each having $N = 25$ data points, and there are 24 Gaussian basis functions in the model so that the total number of parameters is $M = 25$ including the bias parameter. The left column shows the result of fitting the model to the data sets for various values of $\ln \lambda$ (for clarity, only 20 of the 100 fits are shown). The right column shows the corresponding average of the 100 fits (red) along with the sinusoidal function from which the data sets were generated (green).

$$\text{bias}^2 = E_x [\{ E_D[y(x ; \mathcal{D})] - h(x) \}^2]$$

$$\text{variance} = E_{xD} [\{ y(x ; \mathcal{D}) - E_D[y(x ; \mathcal{D})] \}^2]$$

Plot of squared bias and variance, together with their sum, corresponding to the results shown in Figure 3.5. Also shown is the average test set error for a test data set size of 1000 points. The minimum value of $(\text{bias})^2 + \text{variance}$ occurs around $\ln \lambda = -0.31$, which is close to the value that gives the minimum error on the test data.

Stronger regularization
yields higher bias & lower
variance



7.13 Kernel Regression

Purely data driven: interpolate smoothly between the observations
Not assuming any kind of linear relationships

Gaussian Process Regression

https://scikit-learn.org/stable/modules/gaussian_process.html#

<https://github.com/jwangjie/Gaussian-Processes-Regression-Tutorial>

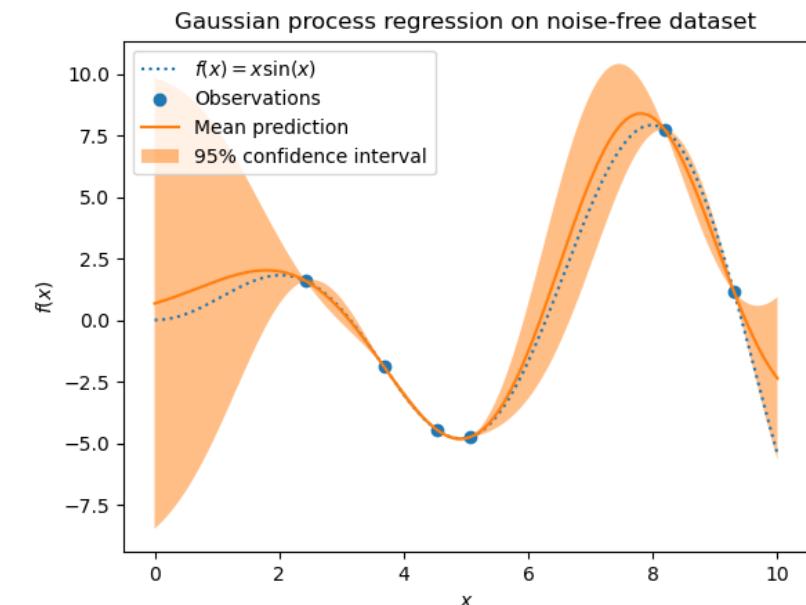
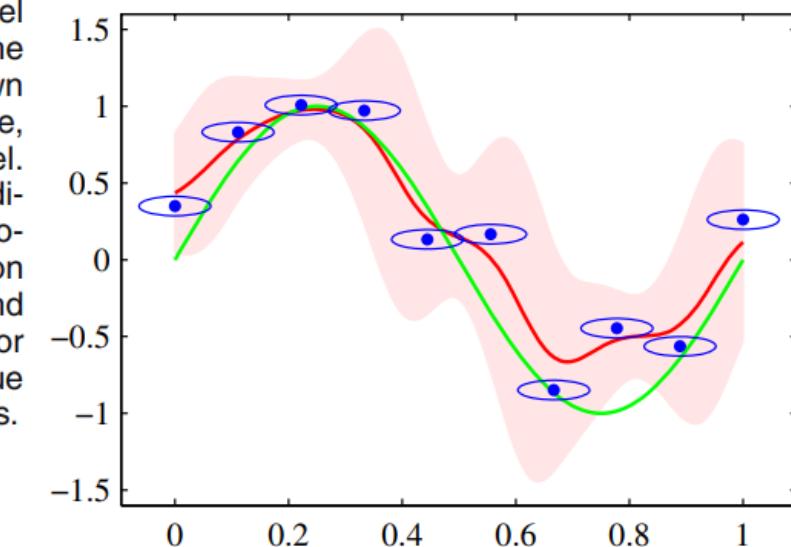


Figure Illustration of the Nadaraya-Watson kernel regression model using isotropic Gaussian kernels, for the sinusoidal data set. The original sine function is shown by the green curve, the data points are shown in blue, and each is the centre of an isotropic Gaussian kernel. The resulting regression function, given by the conditional mean, is shown by the red line, along with the two-standard-deviation region for the conditional distribution $p(t|x)$ shown by the red shading. The blue ellipse around each data point shows one standard deviation contour for the corresponding kernel. These appear noncircular due to the different scales on the horizontal and vertical axes.



My two favorite books from where I re-used most of the material

References

- Pattern-Recognition-and-Machine-Learning - Christopher M. Bishop, available at <https://www.microsoft.com/en-us/research/uploads/prod/2006/01/Bishop-Pattern-Recognition-and-Machine-Learning-2006.pdf>
- Introduction to Probability and statistics for engineers and scientists - Sheldon M. Ross, available at <https://minerva.it.manchester.ac.uk/~saralees/statbook3.pdf>