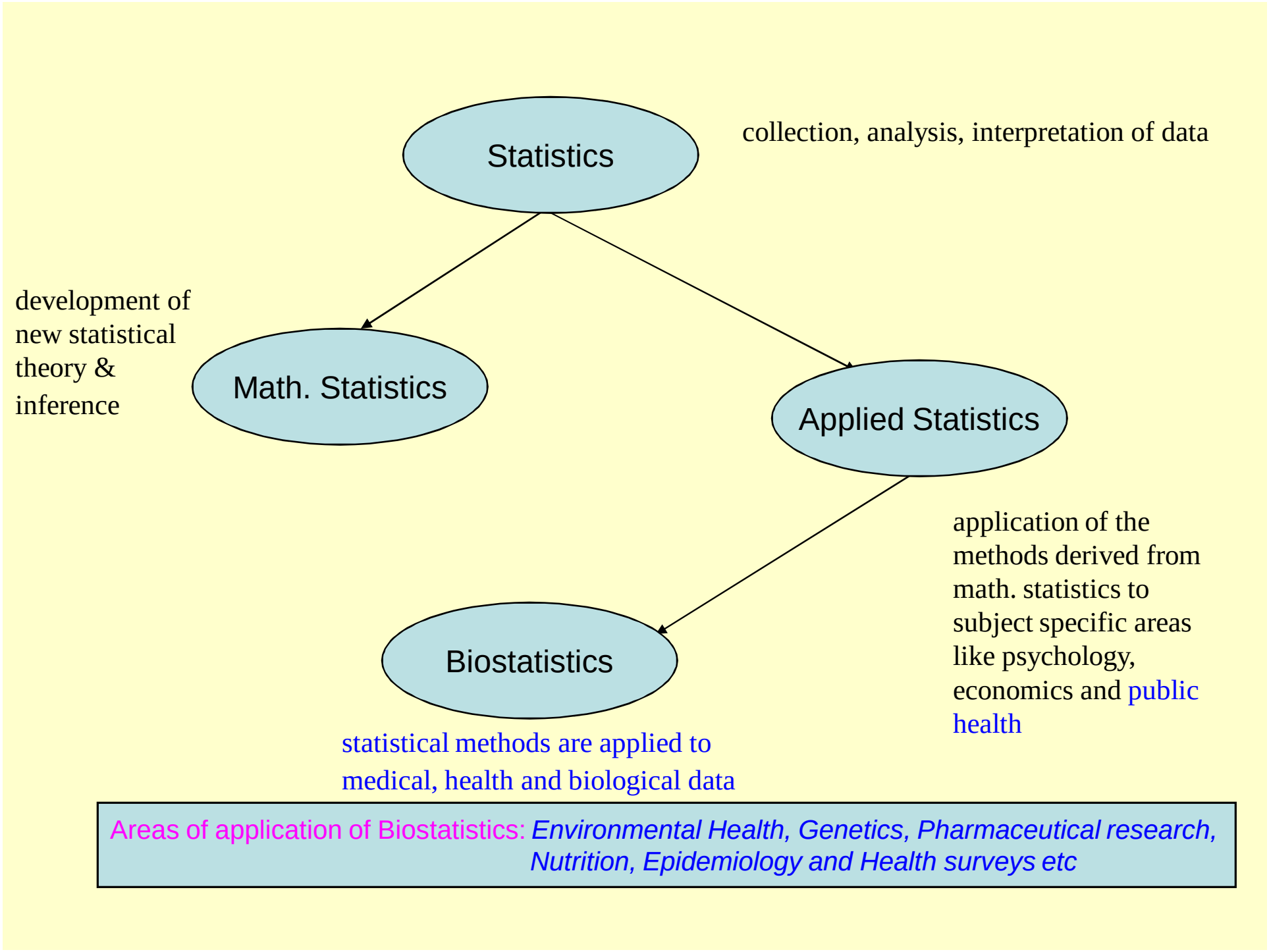
Fundamentals to Biostatistics

Prof. Chandan Chakraborty
Associate Professor
School of Medical Science & Technology
IIT Kharagpur


Statistics
collection, analysis, interpretation of data
development of new statistical theory & inference
Math. Statistics
Applied Statistics
application of the methods derived from math. statistics to subject specific areas like psychology, economics and public health
Biostatistics
statistical methods are applied to medical, health and biological data
Areas of application of Biostatistics: Environmental Health, Genetics, Pharmaceutical research, Nutrition, Epidemiology and Health surveys etc

Some Statistical Tools for Medical Data Analysis

- **Data collection and Variables under study**
- **Descriptive Statistics & Sampling Distribution**

  Statistical Inference – Estimation, Hypothesis Testing, Conf. Interval
- **Association**

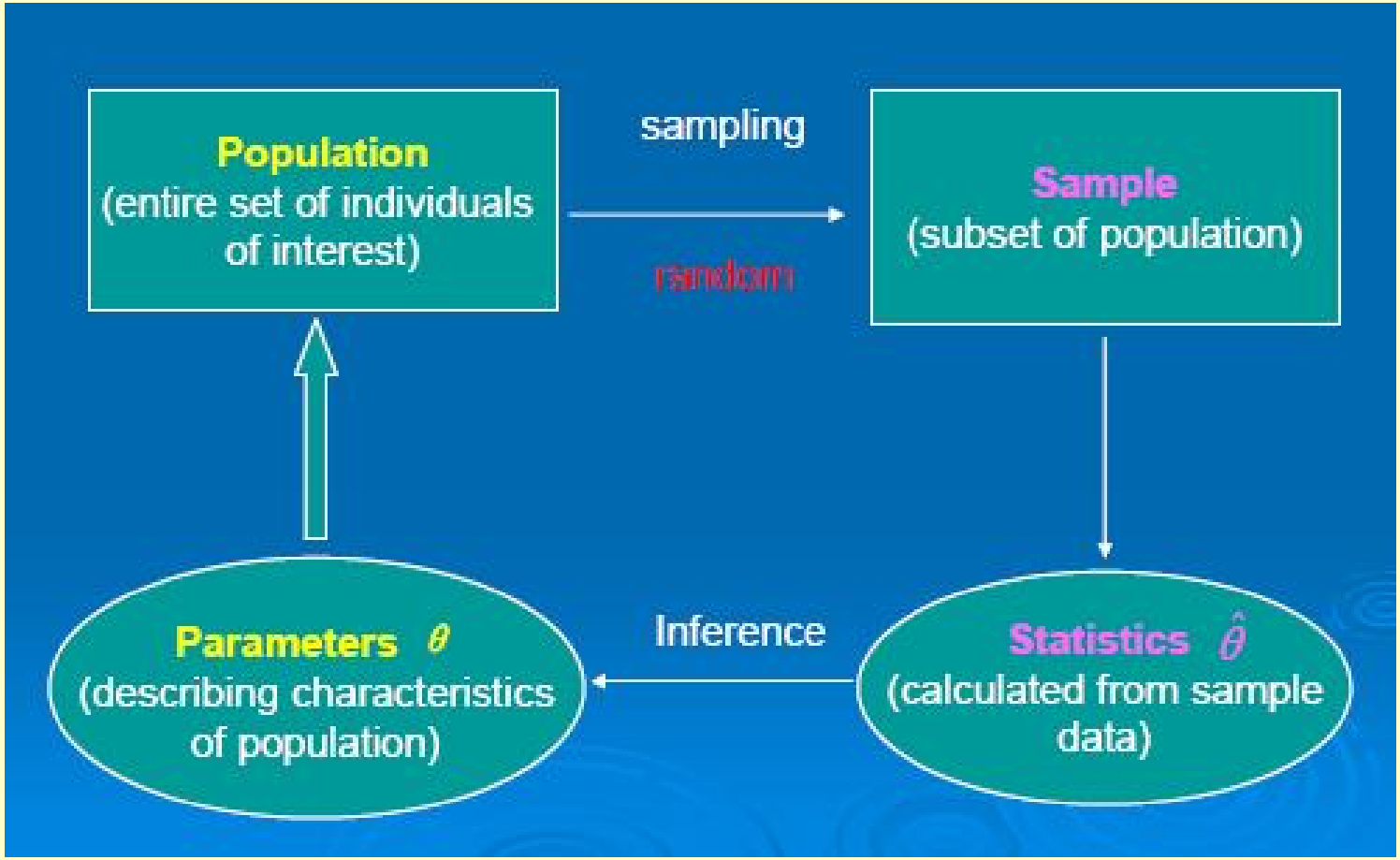
  Continuous: Correlation and Regression

  Categorical: Chi-square test
- **Multivariate Analysis**

  PCA, Clustering Techniques, Discriminantion & Classification
- **Time Series Analysis**

  AR, MA, ARMA, ARIMA

Population vs. Sample
Parameter vs. Statistics



Variable

- **Definition:** characteristic of interest in a study that has different values for different individuals.

- **Two types of variable**

□□□□ Continuous: values form continuum

□□□□ Discrete: assume discrete set of values

- **Examples**

Continuous: blood pressure

□□□□ Discrete: case/control, drug/placebo

Univariate Data

- Measurements on *a single* variable *X*
- Consider a *continuous (numerical)* variable
- Summarizing *X*
  - Numerically
    - Center
    - Spread
  - Graphically
    - Boxplot
    - Histogram

Measures of center: Mean

- The *mean* value of a variable is obtained by computing the total of the values divided by the number of values
- Appropriate for distributions that are fairly symmetrical
- It is sensitive to presence of outliers, since all values contribute equally

Measures of center: Median

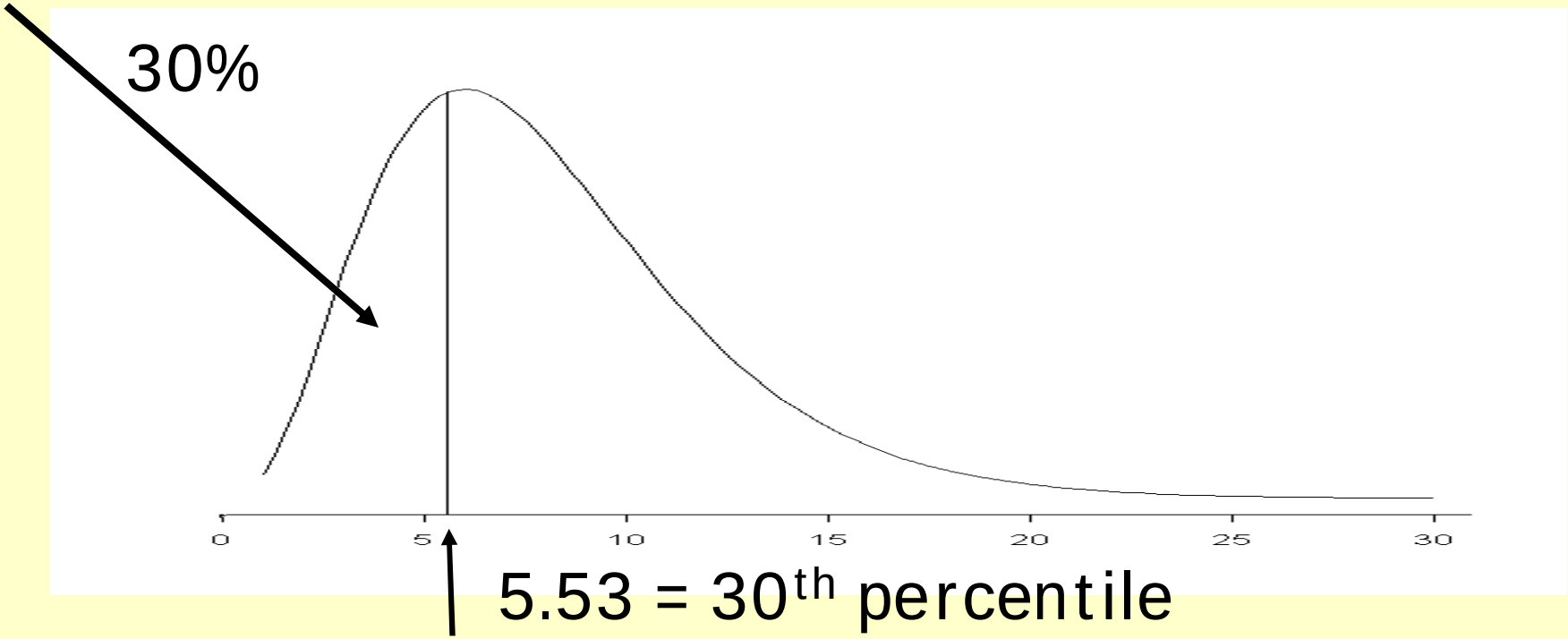
- The *median* value of a variable is the number having 50% (half) of the values smaller than it (and the other half bigger)
- It is NOT sensitive to presence of outliers, since it 'ignores' almost all of the data values
- The median is thus usually a more appropriate summary for skewed distributions

Measures of spread: SD

- The *standard deviation (SD)* of a variable is the square root of the average* of squared deviations from the mean (*for uninteresting technical reasons, instead of dividing by the number of values n, you usually divide by n-1)
- The *SD* is an appropriate measure of spread when center is measured with the *mean*

Quantiles

- The *p*th *quantile* is the number that has the proportion p of the data values smaller than it



Measures of spread: IQR

- The 25^{th} (Q_1), 50^{th} (median), and 75^{th} (Q_3) percentiles divide the data into 4 equal parts; these special percentiles are called *quartiles*
- The *interquartile range (IQR)* of a variable is the distance between Q_1 and Q_3:

$$IQR = Q_3 - Q_1$$

- The *IQR* is one way to measure spread when center is measured with the *median*

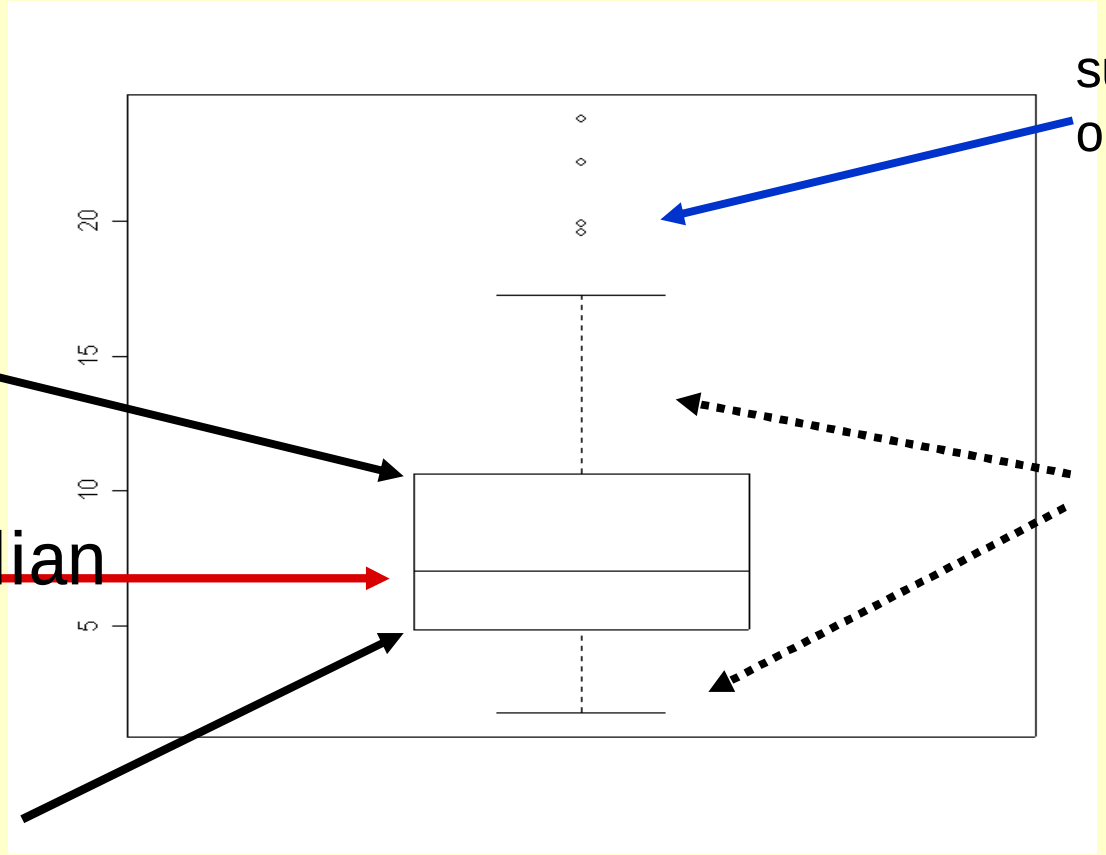
Five-number summary and boxplot

- An overall summary of the distribution of variable values is given by the five values:

  Min, Q_1, Median, Q_3, and Max

- A *boxplot* provides a visual summary of this five-number summary
- Display boxplots side-by-side to compare distributions of different data sets

Boxplot



Histogram

- A *histogram* is a special kind of bar plot
- It allows you to visualize the *distribution* of values for a numerical variable
- When drawn with a *density scale:*
  - the *AREA* (NOT height) of each bar is the proportion of observations in the interval
  - the *TOTAL AREA* is 100% (or 1)

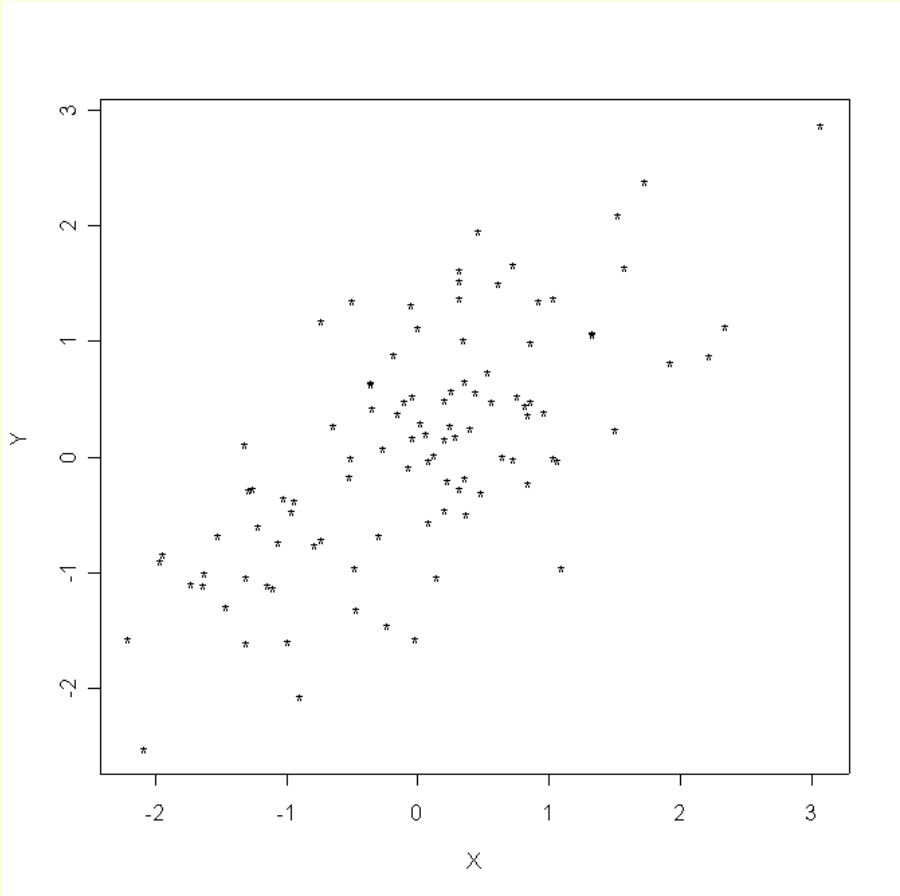
Bivariate Data

- *Bivariate data* are just what they sound like – data with measurements on *two* variables; let's call them *X* and *Y*
- Here, we are looking at two *continuous* variables
- Want to explore the *relationship* between the two variables

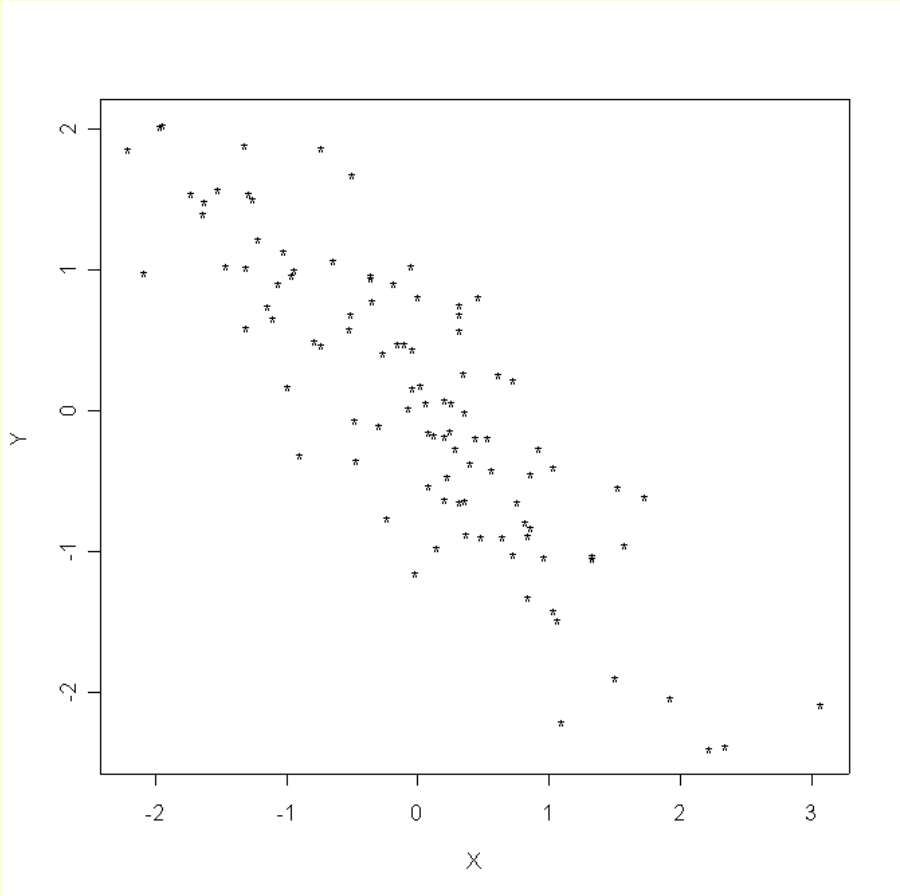
Scatter plot

- We can graphically summarize a bivariate data set with a *scatter plot* (also sometimes called a *scatter diagram*)
- Plots values of one variable on the horizontal axis and values of the other on the vertical axis
- Can be used to see how values of 2 variables tend to move with each other (*i.e*. how the variables are *associated*)

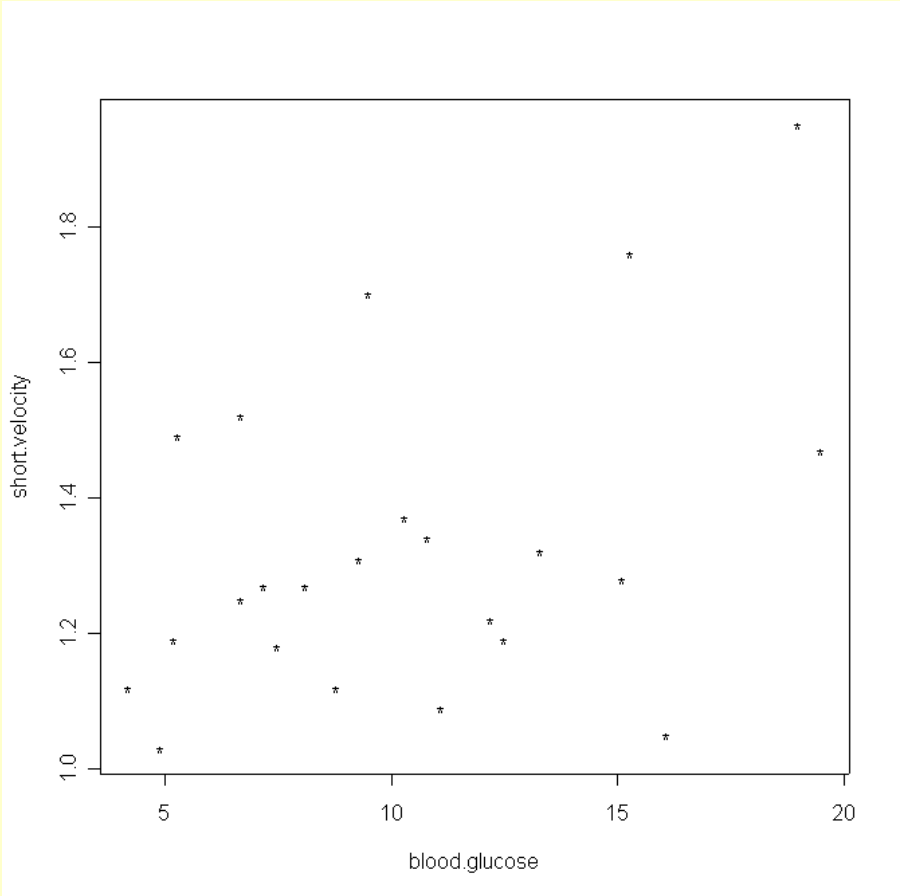
Scatter plot: positive association



Scatter plot: negative association



Scatter plot: real data example



Correlation Coefficient

$$r = \frac{\sum_{i=1}^{n}\left((x_i - \bar{x})(y_i - \bar{y})\right)}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2 \sum_{i=1}^{n}(y_i - \bar{y})^2}}$$

- *r* is a *unitless quantity*
- $-1 \leq r \leq 1$
- *r* is a measure of *LINEAR ASSOCIATION*
- When *r* = 0, the points are *not LINEARLY ASSOCIATED* – this does NOT mean there is *NO ASSOCIATION*

Breast cancer example

- Study on whether age at first child birth is an important risk factor for breast cancer (BC)

BC \ Age	⩾30	⩽29	Total
case	683 (21.2%)	2537	3220
control	1498 (14.6%)	8747	10245
Total	2181	11284	13465

(MacMahon, B. et al., 1970, Rosner, B. 1995, p348)

Blood Pressure Example

- How does taking Oral Conceptive (OC) affect Blood Pressure (BP) in women
- paired samples

	Systolic blood pressure		
Subject i	Baseline (Not using OC)	1-year (Using OC)	Difference
1	115	128	13
2	112	115	3
3	107	106	-1
4	119	128	9
5	115	122	7
6	138	145	7
7	126	132	6
8	105	109	4
9	104	102	-2
10	115	117	2

(Rosner, B. 1995, p253)

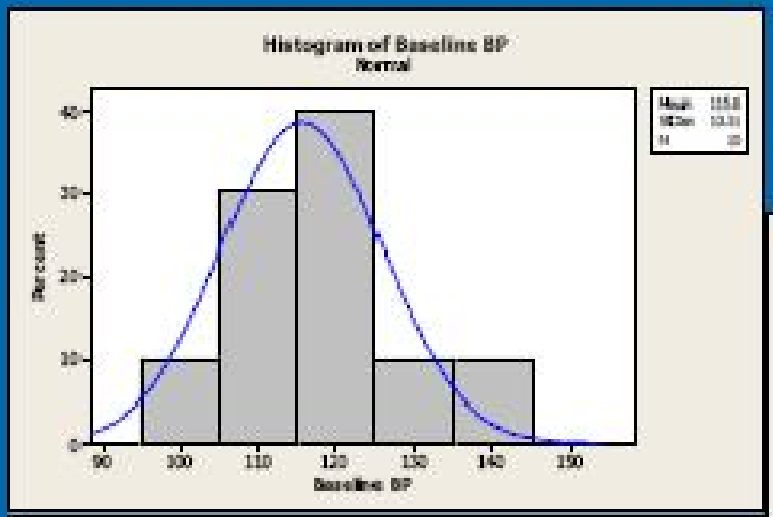
Birthweight Example

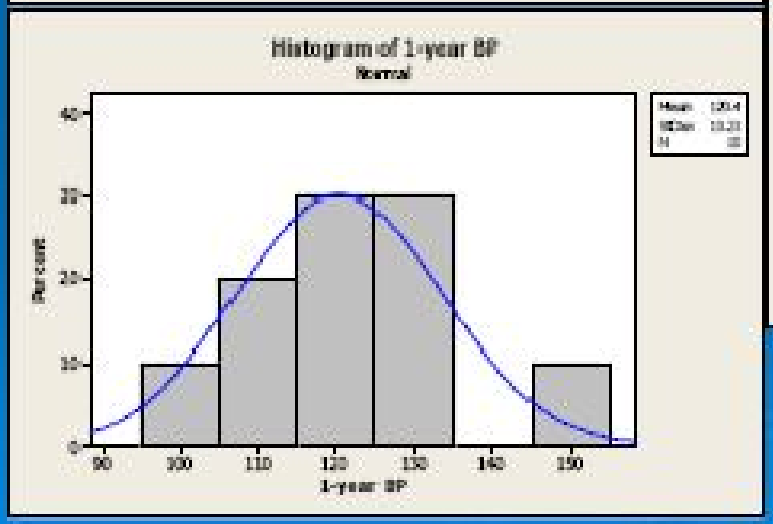
- Determine the effectiveness of drug A on preventing premature birth.
- Independent samples.

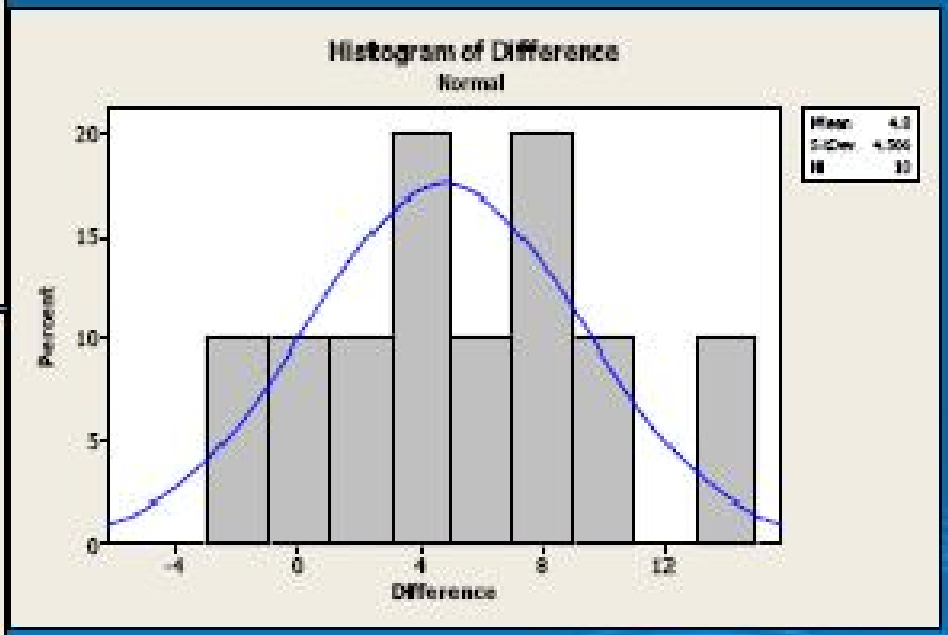
	Birthweight	
Patient #	Treatment	Control
1	6.9	6.4
2	7.6	6.7
3	7.3	5.4
4	7.6	8.2
5	6.8	5.3
6	7.2	6.6
7	8	5.8
8	5.5	5.7
9	5.8	6.2
10	7.3	7.1
11	8.2	7
12	6.9	6.9
13	6.8	5.6
14	5.7	4.2
15	8.6	6.8

(Rosner, B. 1995, p290)

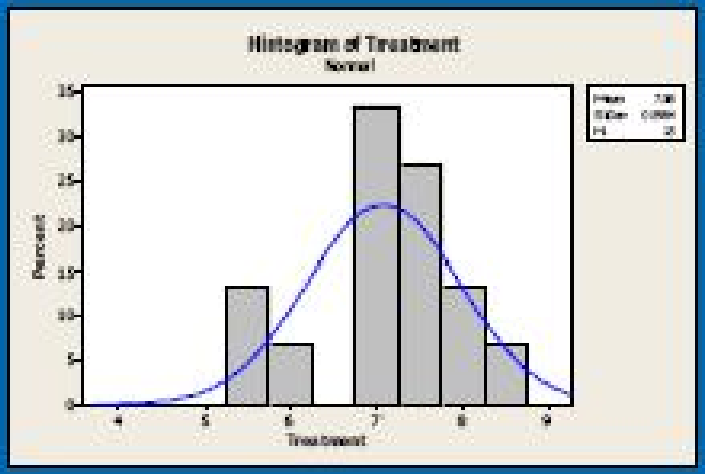
Histogram and Distribution (Blood pressure)

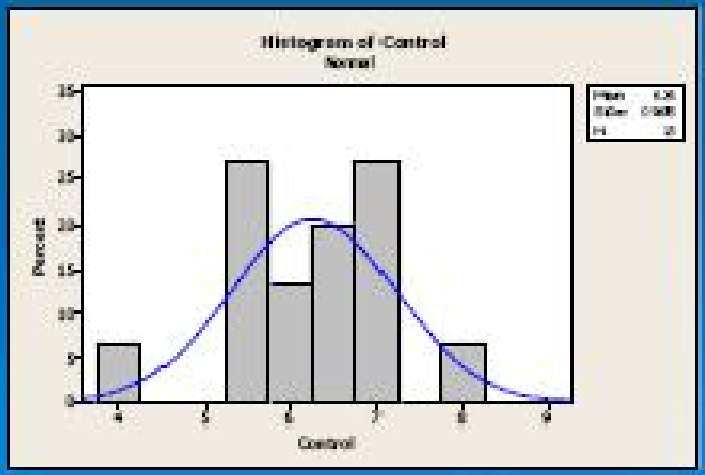


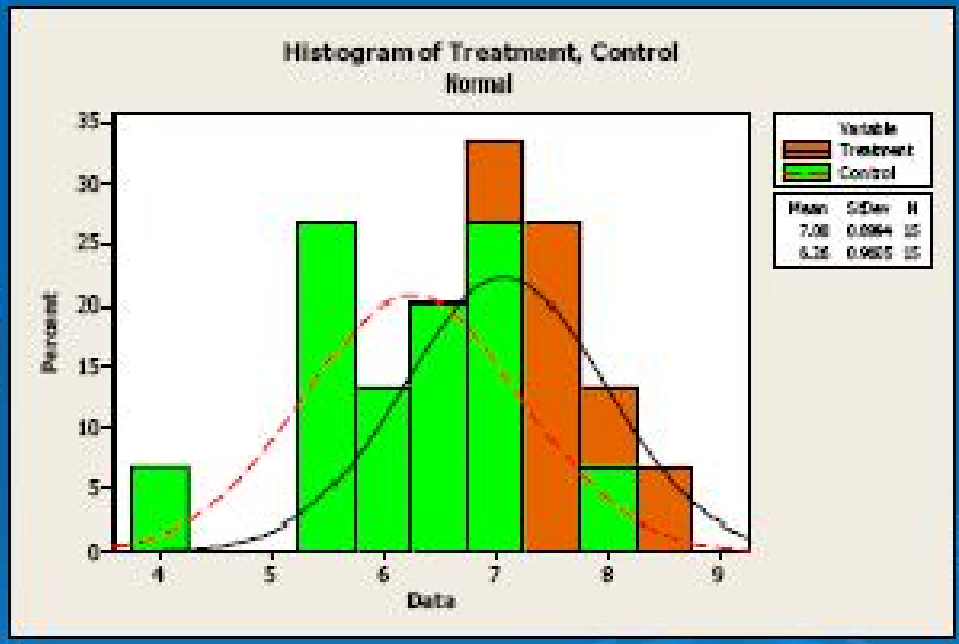




Histogram and Distribution (Birthweight)

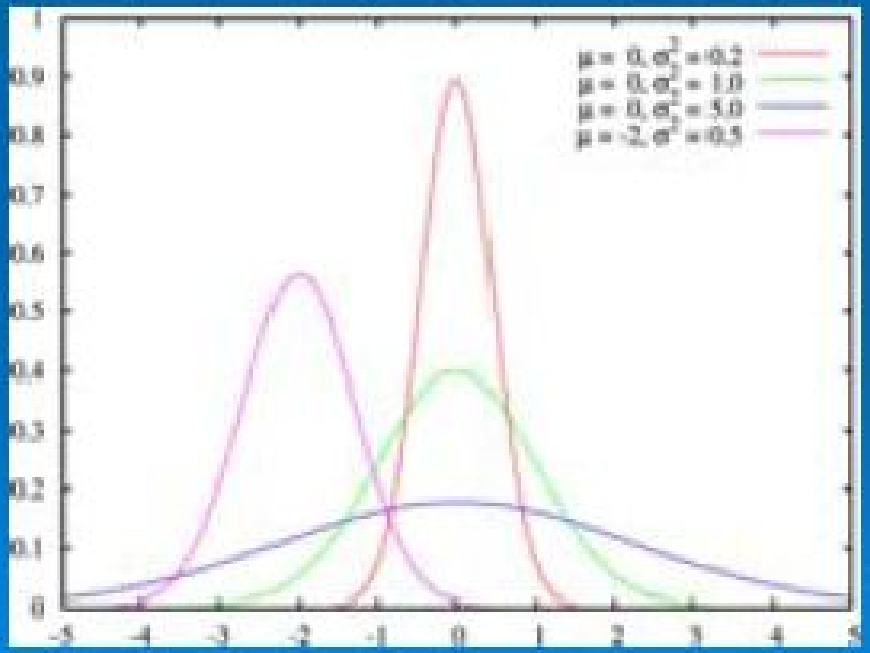


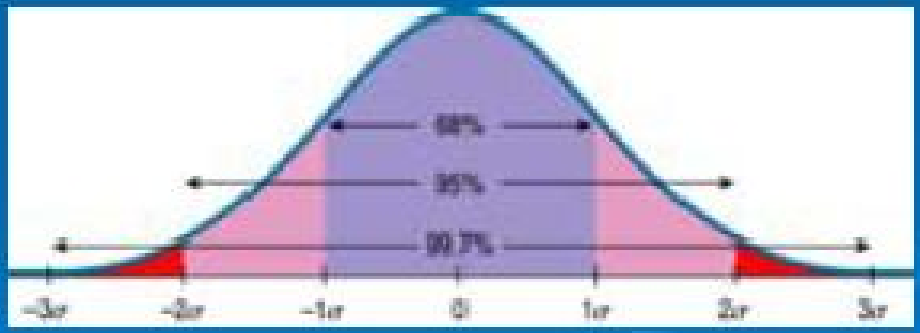




Normal distribution

- $N(\mu, \sigma)$: mean μ, standard deviation σ
- $N(0, 1)$: Standard normal
- Standardization: X → $Z=(X-\mu)/\sigma$





Binomial distribution

- Two possible outcome for each trial:

  (1) "success" and (0) "failure" Pr(success) = *p*

  - Example: *each women with breast cancer had first birth either before age 30 or after.*
- *n independent* trials, X = # of successes ~ B(*n*, *p*)
  - Example: *number of cases whose age at first birth ¤ 30*

$$P(X=k)=\binom{n}{k}p^k(1-p)^{n-k}$$

$$\mu_x = np,\quad \sigma_x^2 = np(1-p)$$

Characteristics of population distribution (parameters)

- Center/location
  - Population mean μ, or "expected value", "average"
  - Population median M, or "middle number"
- Spread
  - Range $R = max - min$
  - Variance σ^2, average squared distance of each value from the mean
  - Standard deviation σ
- Coefficient of Variation (CV): μ/σ
  - Useful for comparing variability of several different samples with different means
- Proportion p for Binomial distribution

Sample statistics

- Center/location
  - Sample mean $\bar{x}$
  - Sample median *M*
- Spread
  - Sample range $R = max - min$
  - Sample standard deviation $s = \sqrt{s^2} = \sqrt{\frac{\sum_{i=1}^{n}(x_i - \bar{x})^2}{n-1}}$
  - Sample variance s^2
- Birthweight example

Birthweight	n	Mean $\bar{x}$	s	Minimum	Median	Maximum
Treatment	15	7.08	0.899	5.5	7.2	8.6
Control	15	6.26	0.961	4.2	6.4	8.2

- Sample proportion $\hat{p} = x / n$

Sampling distribution for statistics

- Sampling variability for statistics,
  - Standard error of sample mean $SE(\bar{x}) = s/\sqrt{n}$
- If x ~ $N(\mu, \sigma)$, then $\bar{x}$ ~ $N(\mu, \sigma/\sqrt{n})$

$$\frac{\bar{x}-\mu}{\sigma/\sqrt{n}} \sim N(0,1) \qquad \frac{\bar{x}-\mu}{s/\sqrt{n}} \sim t_{df=n-1}$$

- Central Limit Theorem: if n is large enough, $\bar{x}$ is approximately normally distributed

$$\frac{\bar{x}-\mu}{\sigma/\sqrt{n}} \sim N(0,1) \qquad \frac{\bar{x}-\mu}{s/\sqrt{n}} \sim t_{df=n-1}$$

- Binomial, $\hat{p}$ is approximately normally distributed

$$\hat{p} \sim N(p, \sqrt{\frac{p(1-p)}{n}})$$

Birthweight example

- X_1 =birthweight for treatment group
- Population distribution

$$X_1 \sim N(7.08, 0.90), n = 15$$

- sampling distribution

$$\overline{X}_1 \sim N(7.08, 0.90/\sqrt{15}) = N(7.08, 0.23)$$

Statistical Inference

- **Point estimation**
  - Estimate parameter θ by $\hat{\theta}$, e.g., estimate ❍ by $\bar{x}$
- Interval estimation: **Confidence intervals** (CI)
  - An interval of plausible values of the parameter θ with a specific confidence level 1- α;, e.g., 95% CI
  - $$\hat{\theta} \pm \text{margin of error} = \hat{\theta} \pm c_{1-\alpha} SE(\hat{\theta})$$
- **Hypothesis testing** (HT);
  - Testing hypotheses about parameter θ

Hypothesis Testing

- null (H_0) and alternative hypothesis (H_a)
  - Two-sided: $H_0: \theta = \theta_0$ versus $H_a: \theta \neq \theta_0$
  - One-sided: $H_0: \theta = \theta_0$ versus $H_a: \theta < \theta_0$
  - One-sided: $H_0: \theta = \theta_0$ versus $H_a: \theta > \theta_0$
- Test statistics
- p-value
- Compare p-value with significance level α
  - p-value < α, reject H_0
  - p-value $\geq$ α, do not reject H_0
- conclusion

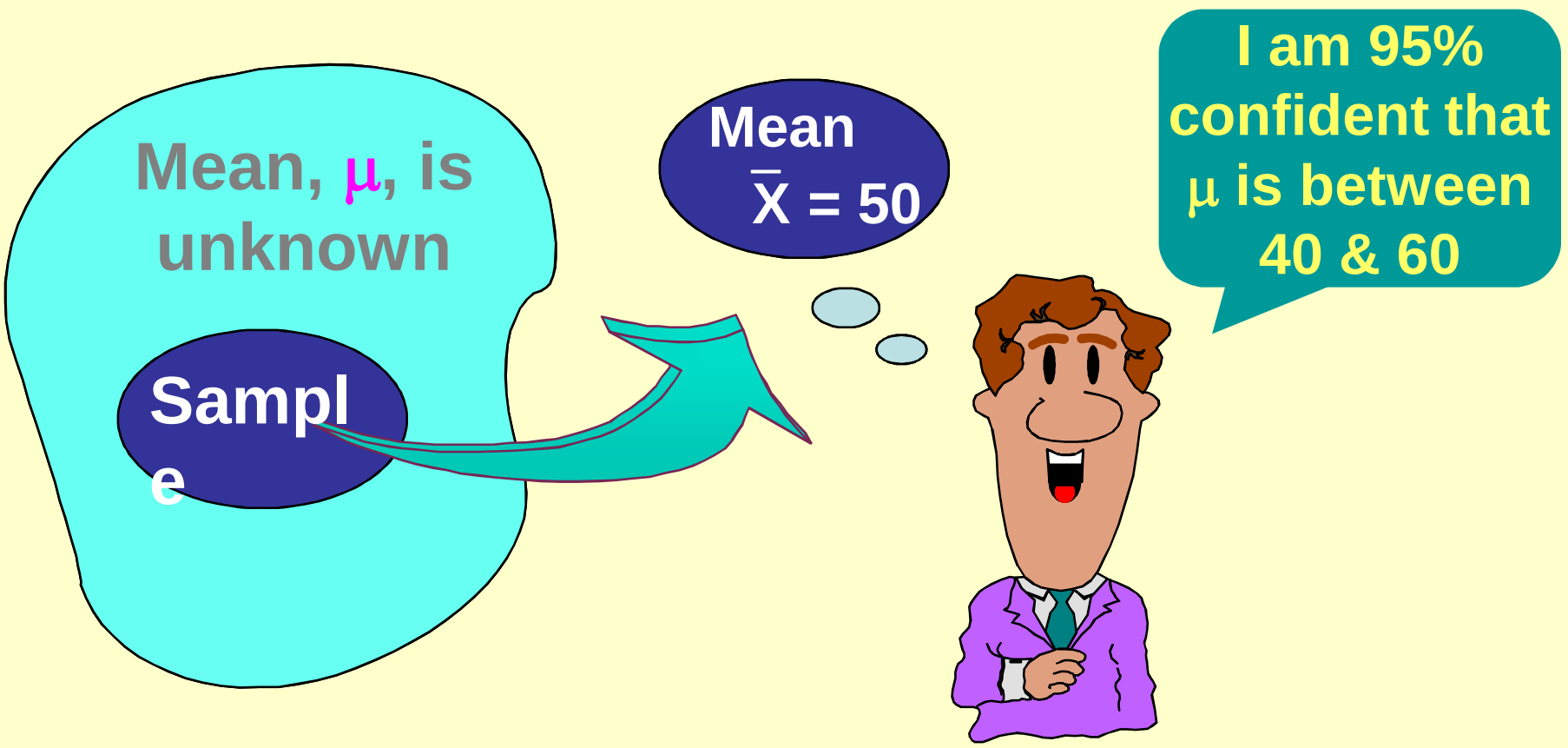
	Statistic		Parameter
Mean:	$\bar{x}$	estimates	μ
Standard deviation:	s	estimates	σ
Proportion:	p	estimates	π

from sample

from entire population

Estimation of parameters

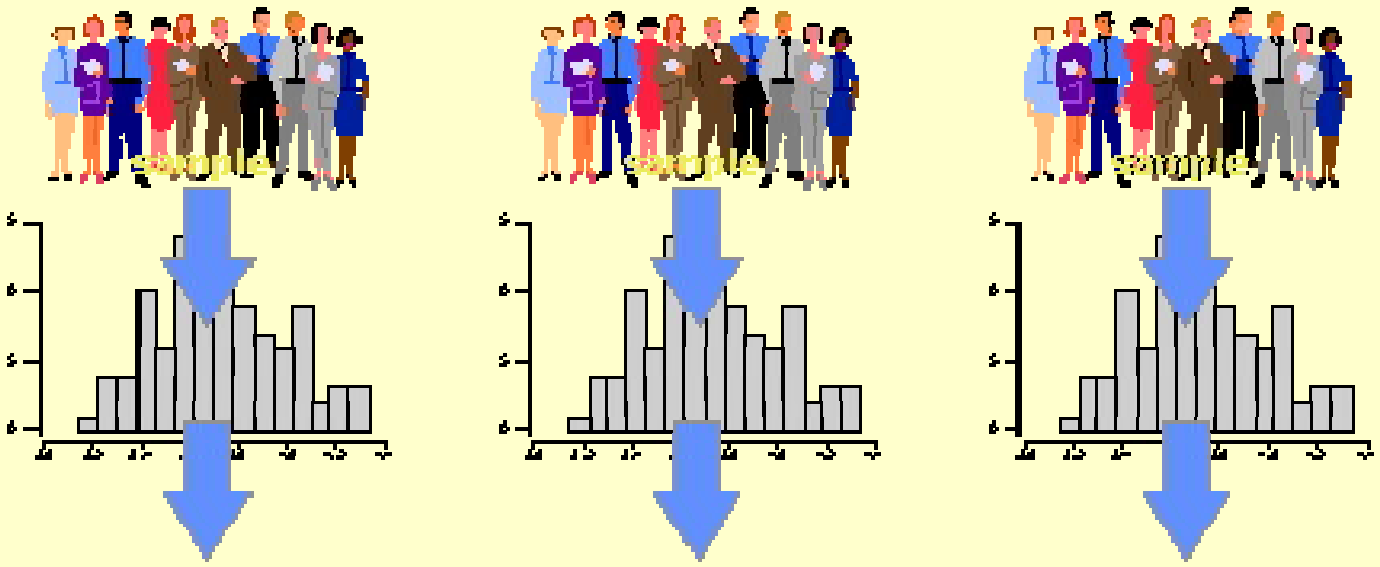
Population **Point estimate** **Interval estimate**



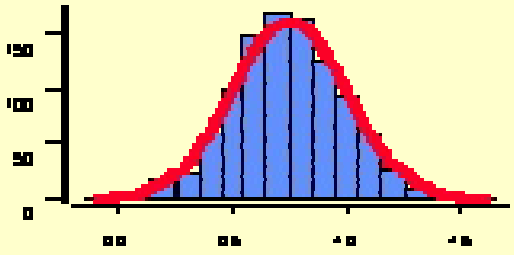
Parameter

= Statistic ± Its Error

Sampling Distribution



The Sampling Distribution...



...is the distribution of a statistic across an infinite number of samples

Standard Error

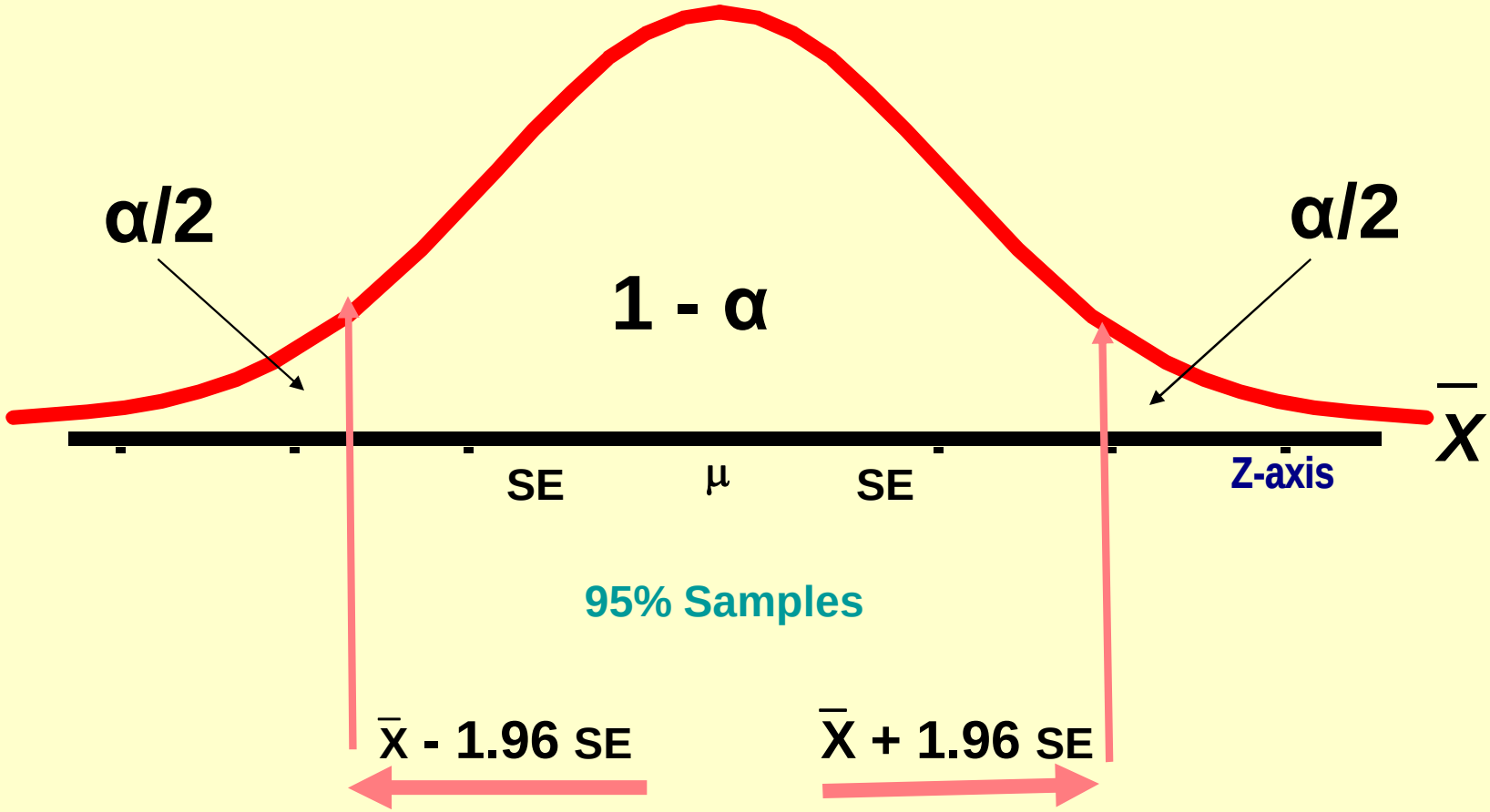
Quantitative Variable

$$SE\ (Mean) = \frac{S}{\sqrt{n}}$$

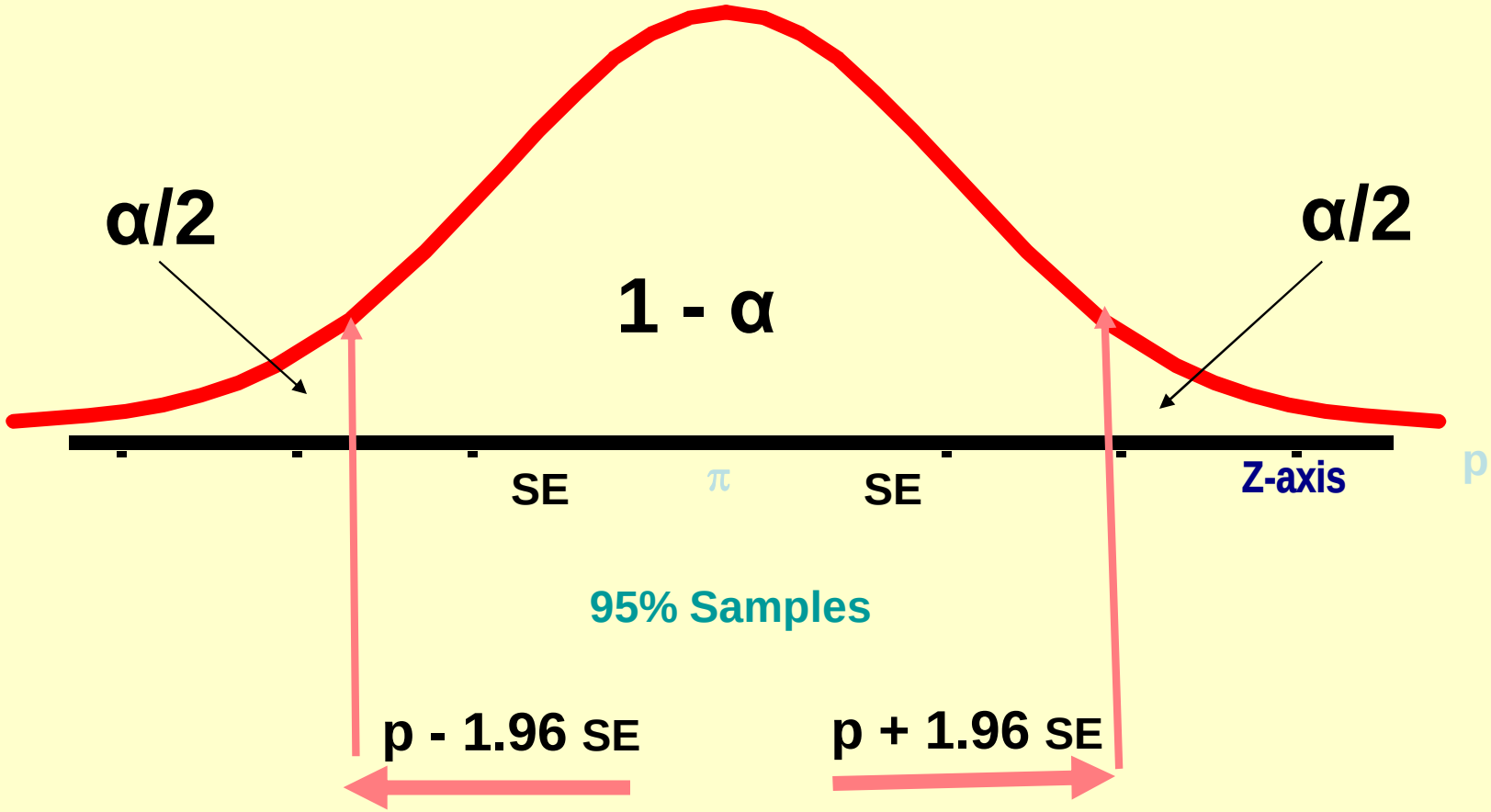
Qualitative Variable

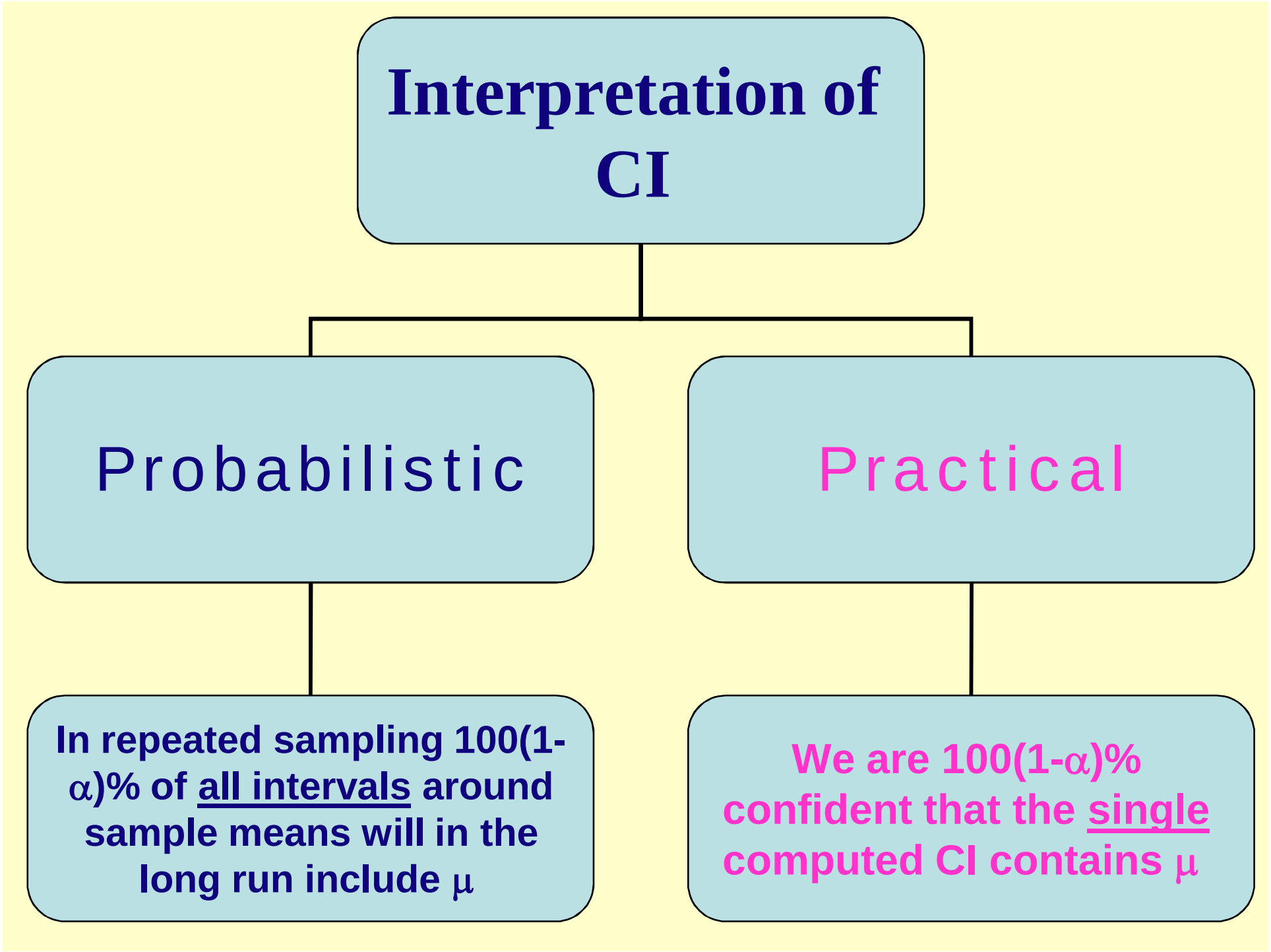
$$SE\ (p) = \sqrt{\frac{p(1-p)}{n}}$$

Confidence Interval



Confidence Interval




Interpretation of CI
Probabilistic
Practical
In repeated sampling 100(1-α)% of all intervals around sample means will in the long run include μ
We are 100(1-α)% confident that the single computed CI contains μ

Example (Sample size≥30)

An epidemiologist studied the blood glucose level of a random sample of 100 patients. The mean was 170, with a SD of 10.

$$\mu = \bar{X} \pm Z \times SE$$

SE = 10/10 = 1

95%

Then CI:

$\mu = 170 \pm 1.96 \times 1$ $\quad 168.04 \leq \mu \leq 171.96$

Example (Proportion)

In a survey of 140 asthmatics, 35% had allergy to house dust. Construct the 95% CI for the population proportion.

$$\pi = p \pm Z\sqrt{\frac{P(1-p)}{n}} \qquad SE = \sqrt{\frac{0.35(1-0.35)}{140}} = 0.04$$

$$0.35 - 1.96 \times 0.04 \leq \pi \geq 0.35 + 1.96 \times 0.04$$

$$0.27 \leq \pi \geq 0.43$$

$$27\% \leq \pi \geq 43\%$$

Hypothesis testing

A statistical method that uses sample data to evaluate a hypothesis about a population parameter. It is intended to help researchers differentiate between real and random patterns in the data.

Null & Alternative Hypotheses

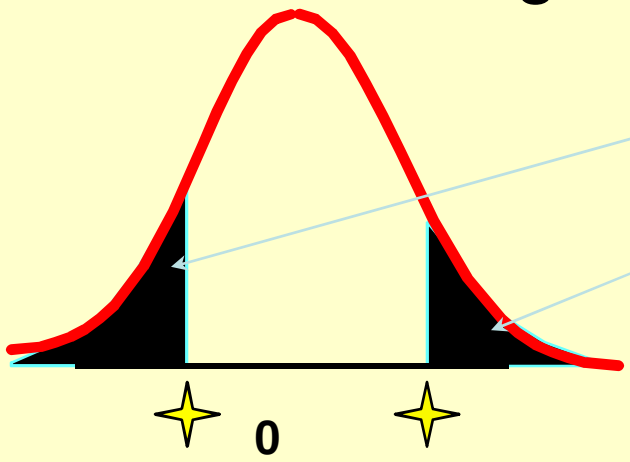
- **H_0 Null Hypothesis states the Assumption to be tested e.g. SBP of participants = 120 (H_0: $\mu = 120$).**

- **H_1 Alternative Hypothesis is the opposite of the null hypothesis (SBP of participants ≠ 120 (H_1: $\mu \neq 120$). It may or may not be accepted and it is the hypothesis that is believed to be true by the researcher**

Level of Significance, α

- **Defines unlikely values of sample statistic if null hypothesis is true. Called rejection region of sampling distribution**
- **Typical values are 0.01, 0.05**
- **Selected by the Researcher at the Start**
- **Provides the Critical Value(s) of the Test**

Level of Significance, α and the Rejection Region



Result Possibilities

H_0: Innocent

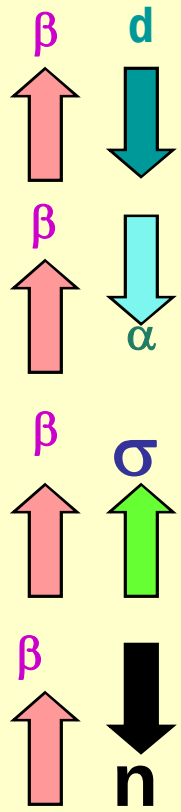
Jury Trial			Hypothesis Test		
	Actual Situation			Actual Situation	
Verdict	Innocent	Guilty	Decision	H_0 True	H_0 False
Innocent	Correct	Error	Accept H_0	$1 - \alpha$	Type II Error (β)
Guilty	Error	Correct	Reject H_0	Type I Error (α)	Power ($1 - \beta$)

False Positive

False Negative

Factors Increasing Type II Error

- True Value of Population Parameter
  - Increases When Difference Between Hypothesized Parameter & True Value Decreases
- Significance Level α
  - Increases When α Decreases
- Population Standard Deviation σ
  - Increases When σ Increases
- Sample Size *n*
  - Increases When *n* Decreases



p Value Test

- **Probability of Obtaining a Test Statistic More Extreme ($\leq$ or $\geq$) than Actual Sample Value Given H_0 Is True**
- **Called Observed Level of Significance**
- **Used to Make Rejection Decision**
  - **If *p* value $\geq \alpha$, Do Not Reject H_0**
  - **If *p* value $< \alpha$, Reject H_0**

Hypothesis Testing: Steps

Test the Assumption that the true mean SBP of participants is 120 mmHg.

State H_0	$H_0: \mu = 120$
State H_1	$H_1: \mu \neq 120$
Choose α	$\alpha = 0.05$
Choose *n*	n = 100
Choose Test:	***Z, t, X^2 Test (or p Value)***

Hypothesis Testing: Steps

Compute Test Statistic *(or compute P value)*

Search for Critical Value

Make Statistical Decision rule

Express Decision

One sample-mean Test

- **Assumptions**
  - **Population is normally distributed**

- **t test statistic**

$$t = \frac{\text{sample mean} - \text{null value}}{\text{standard error}} = \frac{\bar{x} - \mu_0}{s/\sqrt{n}}$$

Example Normal Body Temperature

What is **normal body temperature**? Is it actually 37.6^{o}C (on average)?

State the null and alternative hypotheses

H_0: $\mu = 37.6^{o}C$

H_a: $\mu \neq 37.6^{o}C$

Example Normal Body Temp (cont)

Data: random sample of n = 18 normal body temps

37.2	**36.8**	**38.0**	**37.6**	**37.2**	**36.8**	**37.4**	**38.7**	**37.2**
36.4	**36.6**	**37.4**	**37.0**	**38.2**	**37.6**	**36.1**	**36.2**	**37.5**

Summarize data with a test statistic

Variable	n	Mean	SD	SE	t	P
Temperature	18	37.22	0.68	0.161	2.38	0.029

$$t = \frac{\text{sample mean} - \text{null value}}{\text{standard error}} = \frac{\bar{x} - \mu_0}{s/\sqrt{n}}$$

STUDENT'S t DISTRIBUTION TABLE

Degrees of freedom	Probability (p value)		
	0.10	**0.05**	**0.01**
1	**6.314**	**12.706**	**63.657**
5	**2.015**	**2.571**	**4.032**
10	**1.813**	**2.228**	**3.169**
17	**1.740**	**2.110**	**2.898**
20	**1.725**	**2.086**	**2.845**
24	**1.711**	**2.064**	**2.797**
25	**1.708**	**2.060**	**2.787**
∞	**1.645**	**1.960**	**2.576**

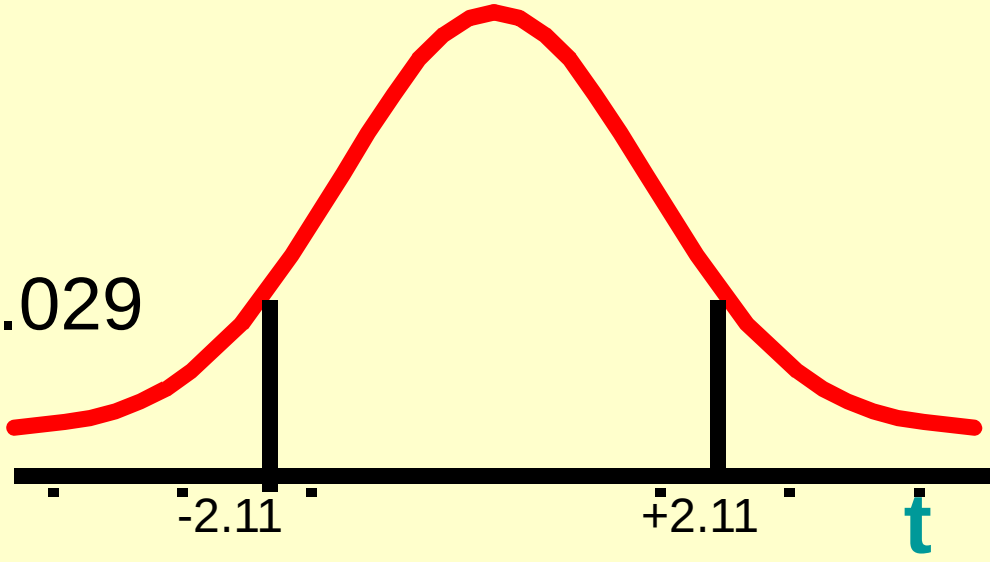
Example Normal Body Temp (cont)

Find the *p*-value

Df = n – 1 = 18 – 1 = 17

From SPSS: *p*-value = 0.029

From t Table: *p*-value is between 0.05 and 0.01.



Area to left of *t* = -2.11 equals area to right of *t* = +2.11.

The value *t* = 2.38 is between column headings 2.110& 2.898 in table, and for df =17, the *p*-values are 0.05 and 0.01.

Example Normal Body Temp (cont)

Decide whether or not the result is statistically significant based on the *p*-value

Using $\alpha = 0.05$ as the level of significance criterion, the results are **statistically significant** because 0.029 is less than 0.05. In other words, we can reject the null hypothesis.

Report the Conclusion

We can conclude, based on these data, that the mean temperature in the human population does not equal 37.6.

One-sample test for proportion

- Involves categorical variables
- Fraction or % of population in a category
- Sample proportion (*p*)

$$p = \frac{X}{n} = \frac{number\ of\ successes}{sample\ size}$$

- Test is called Z test where:
- Z is computed value
- π is proportion in population (null hypothesis value)

$$Z = \frac{p - \pi}{\sqrt{\frac{\pi\ (1 - \pi\ )}{n}}}$$

Critical Values: 1.96 at α=0.05

2.58 at α=0.01

Example

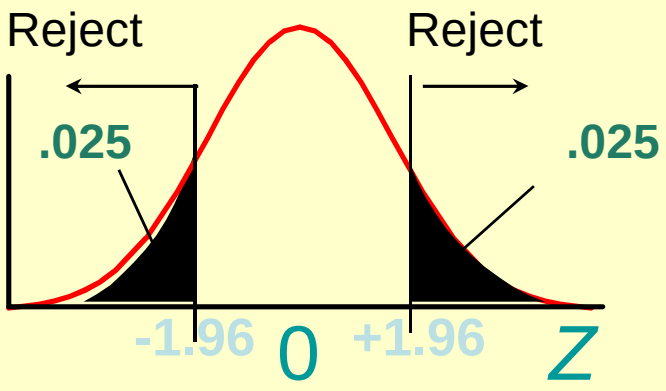
- **In a survey of diabetics in a large city, it was found that 100 out of 400 have diabetic foot. Can we conclude that 20 percent of diabetics in the sampled population have diabetic foot.**
- **Test at the α =0.05 significance level.**

Solution

H_o: π = 0.20

H_1: π ≠ 0.20

$$Z = \frac{0.25 - 0.20}{\sqrt{\frac{0.20\,(1 - 0.20)}{400}}} = 2.50$$



Critical Value: 1.96

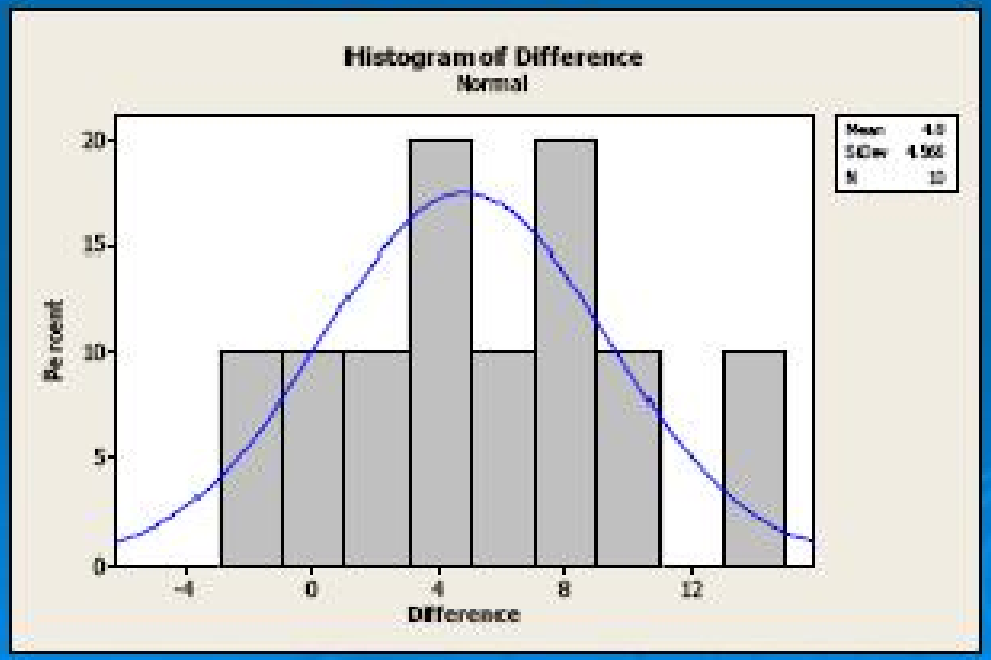
Decision:

We have sufficient evidence to reject the Ho value of 20%

We conclude that in the population of diabetic the proportion who have diabetic foot does not equal 0.20

Paired t-test: blood pressure

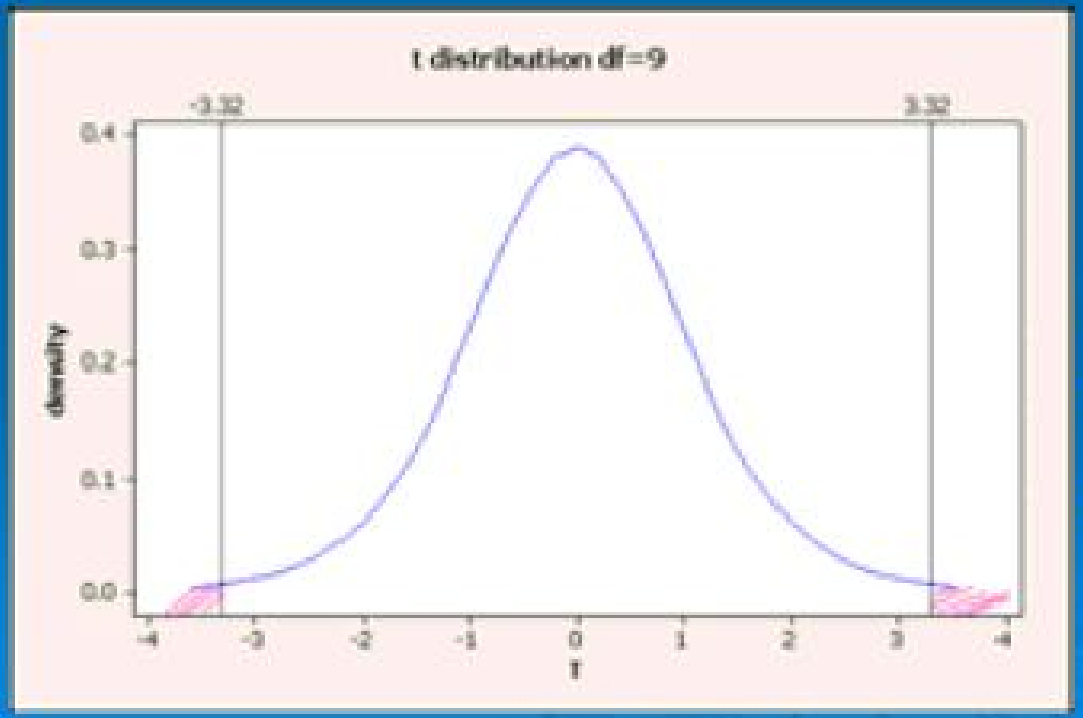
- μ_1: mean BP at baseline (no OC)
- μ_2: mean BP at 1-year (with OC)
- H_0: $\mu_1 - \mu_2 = 0$ vs. H_a: $\mu_1 - \mu_2 \neq 0$



Paired t-test: blood pressure

- Assume H_0 is true, $T = \frac{\bar{d} - 0}{SE(\bar{d})} = \frac{\bar{d}}{s_d / \sqrt{n}} = \frac{\bar{d}}{4.566 / \sqrt{10}} \sim t_{df=9}$
- p-value: the chance of observing something as extreme or more extreme than what we observed,

  t=4.8/1.444=3.32.



p-value

=2P(T<-3.32)

=2 * 0.0045

=0.009 < α

=0.05

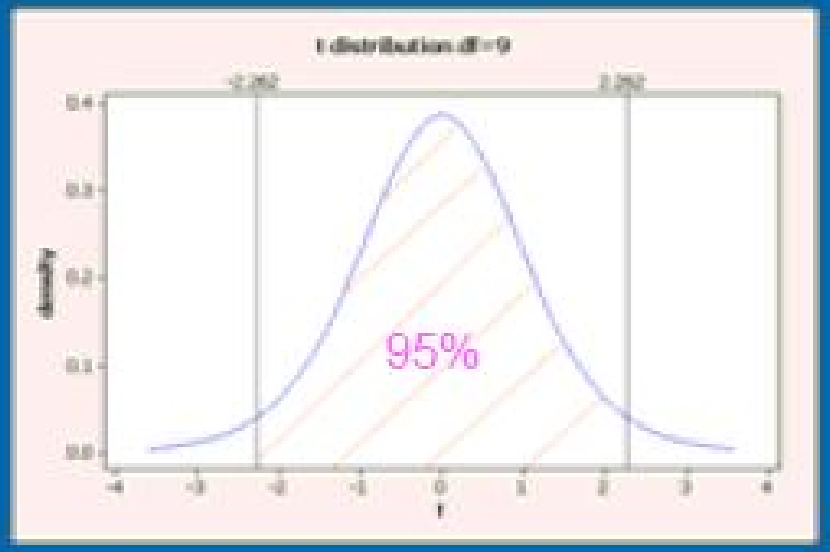
Significant!

Reject H_0.

CI: blood pressure

- Approximately,

$$\frac{\bar{d} - \mu_d}{s_d / \sqrt{n}} \sim t_9, \text{ where } \mu_d = \mu_1 - \mu_2$$



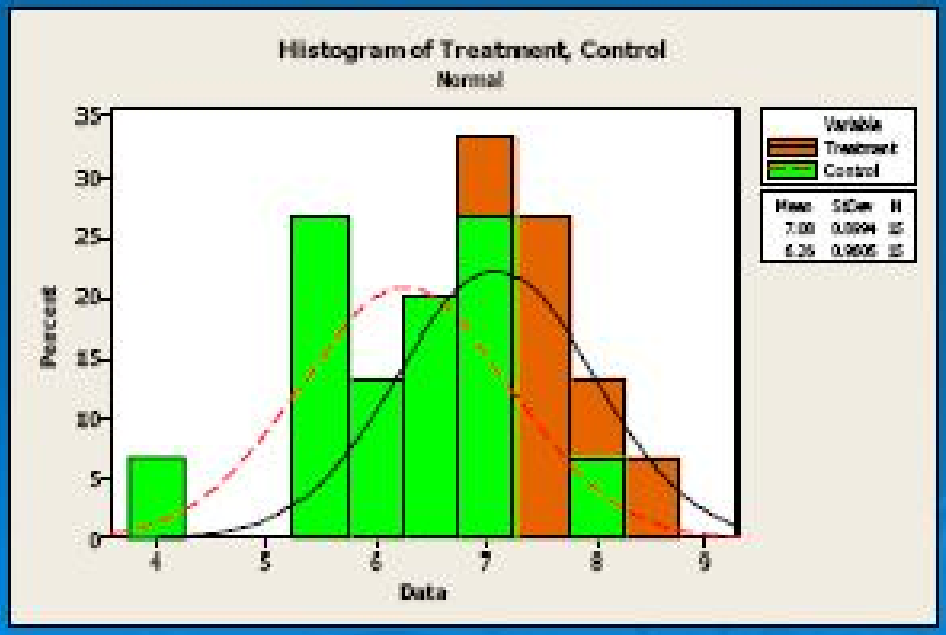
- 95% confidence interval for μ_d, ($t_{df=9,0.975}$=2.262)
  -

$$\bar{d} \pm t_{df=9,1-\alpha/2} \frac{s_d}{\sqrt{n}} = 4.8 \pm 2.262(1.444)$$

$$= 4.8 \pm 3.27 = (1.53, 8.07)$$

Two-sample t-test: birthweight

- μ_1: mean baby weight of treatment group
- μ_2: mean baby weight of control group
- H_0: $\mu_1 - \mu_2 = 0$ vs. H_a: $\mu_1 - \mu_2 > 0$

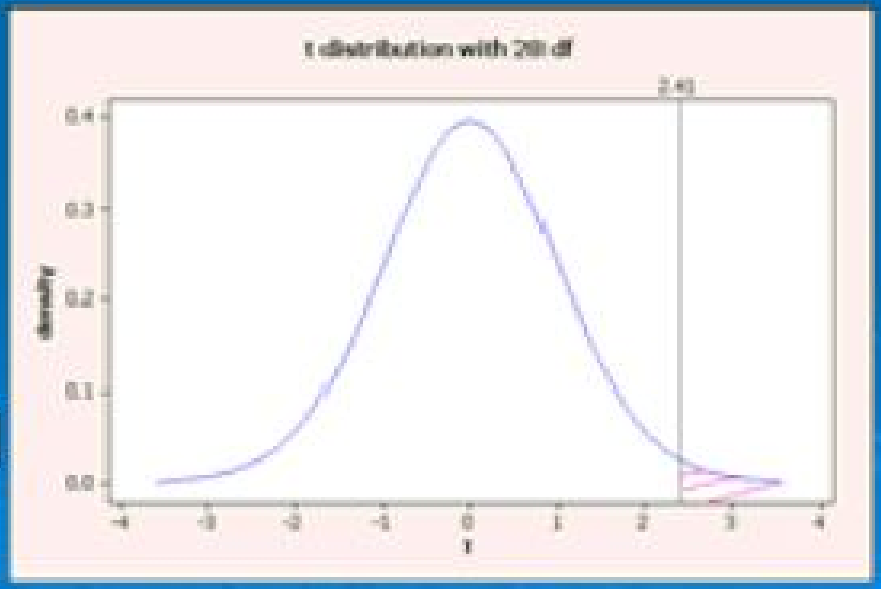


Two-sample t-test: birthweight

- Equal variance for the two groups
- Pooled standard deviation s_{pooled} from s_1 and s_2

$$T = \frac{\bar{x}_1 - \bar{x}_2 - \mu_1 - \mu_2}{s_{pooled}\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}} \overset{\text{Under } H_0}{=} \frac{\bar{x}_1 - \bar{x}_2}{s_{pooled}\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}} \sim t_{n_1+n_2-2}$$

- t=2.41, df=28,
- p-value=0.011 < 0.05
- Reject H_0 at 5% level



CI: birthweight

- 95% CI for $\mu_1 - \mu_2$
- Estimated difference $\pm$ Margin of Error

$$\bar{x}_1 = 7.08, \bar{x}_2 = 6.26, s_1 = 0.899, s_2 = 0.961$$

$$s_{pooled} = \sqrt{[(n_1 - 1)s_1^2 + (n_2 - 1)s_2^2]/(n_1 + n_2 - 2)} = 0.93$$

$$SE(\bar{x}_1 - \bar{x}_2) = s_{pooled}\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}$$

$$\bar{x}_1 - \bar{x}_2 \pm t_{df=28,0.975}\, s_{pooled}\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}$$

$$= 0.82 \pm 2.04841 * 0.3397766 = (0.0124, 0.516)$$

Two-sample proportion Z-test

- p_1: proportion of case women whose age at first birth $\geqslant$30
- p_2: proportion of case women whose age at first birth $\leqslant$29
- H_0: p_1-p_2 = 0 vs. H_a: p_1-$p_2 \neq 0$

$$z = \frac{\hat{p}_1 - \hat{p}_2}{\sqrt{\hat{p}(1-\hat{p})(1/n_1 + 1/n_2)}} \sim N(0,1)$$

$$\hat{p} = \frac{n_1\hat{p}_1 + n_2\hat{p}_2}{n_1 + n_2}, \text{ need } n_1\hat{p}(1-\hat{p}) \geq 5 \text{ and } n_2\hat{p}(1-\hat{p}) \geq 5$$

Breast cancer example

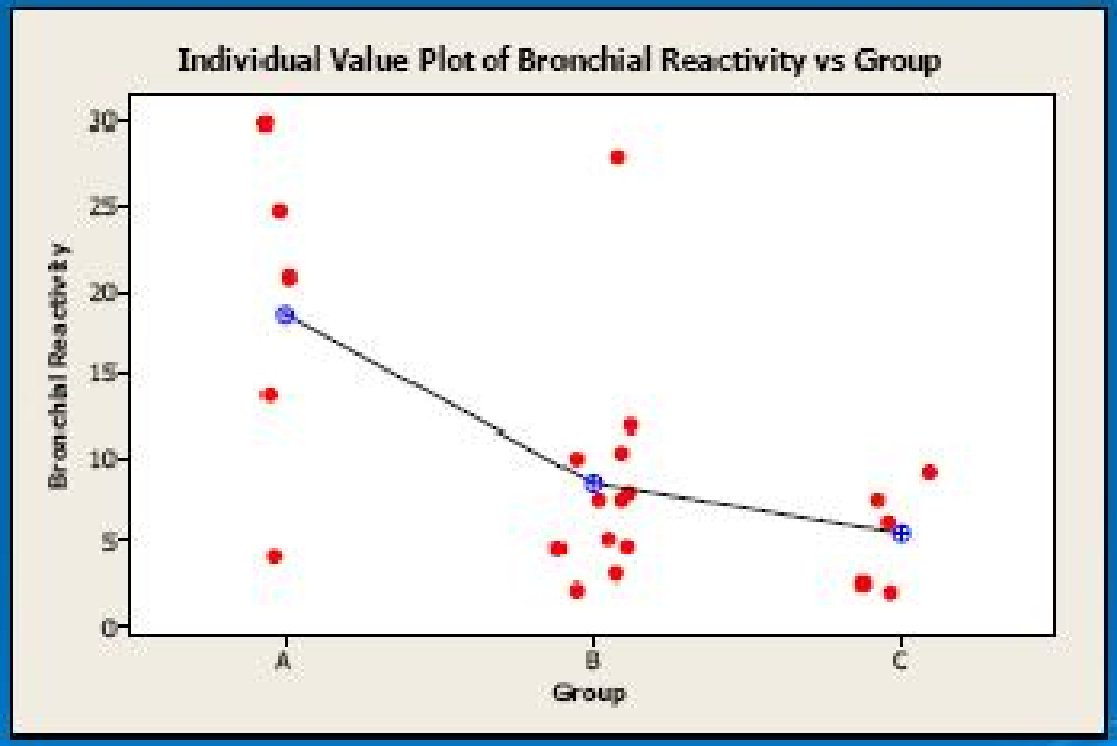
BC \ Age	$\geqslant 30$	$\leqslant 29$	Total
Yes (case)	683 ($\hat{p}_1$=21.2%)	2537	n_1= 3220
No (control)	1498 ($\hat{p}_2$=14.6%)	8747	n_2= 10245
Total	2181	11284	13465

- z = 8.9
- p-value=2P(Z>8.9)=2*[1-;Φ(8.9)] $\approx$0 < 0.01
- Reject H_0 at 1% significance level

ANOVA

- Analysis of Variance: comparing the mean of more than two distributions/groups
- Example: relationship between bronchial reactivity to SO_2 grouped by lung function among 22 asthmatic patients
  - Group A ($FEV_1/FVC \leqslant 74\%$)
  - Group B ($75\% \leqslant FEV_1/FVC \leqslant 84\%$)
  - Group C ($FEV_1/FVC \geqslant 85\%$)
- Question: is there significant difference in mean bronchial reactivity among the 3 groups

Plot of data


Individual Value Plot of Bronchial Reactivity vs Group
Bronchial Reactivity
30
25
20
15
10
5
0
A
B
C
Group

One-way ANOVA

- $y = \mu + \alpha_i + \varepsilon_{ij}$
  - μ: overall effect, underlying mean of all groups taken together
  - α_i : i=1,2,3, group effect, difference between the mean of ith group and the overall mean
  - ε_{ij} : random error about the mean $\mu + \alpha_i$ for subject j in group I
- ANOVA table
- F-test for overall comparison of group means
  - H_0: $\alpha_1 = \alpha_2 = \alpha_3$ vs. H_a: not all equal

ANOVA table

One-way ANOVA: Bronchial Reactivity versus Group

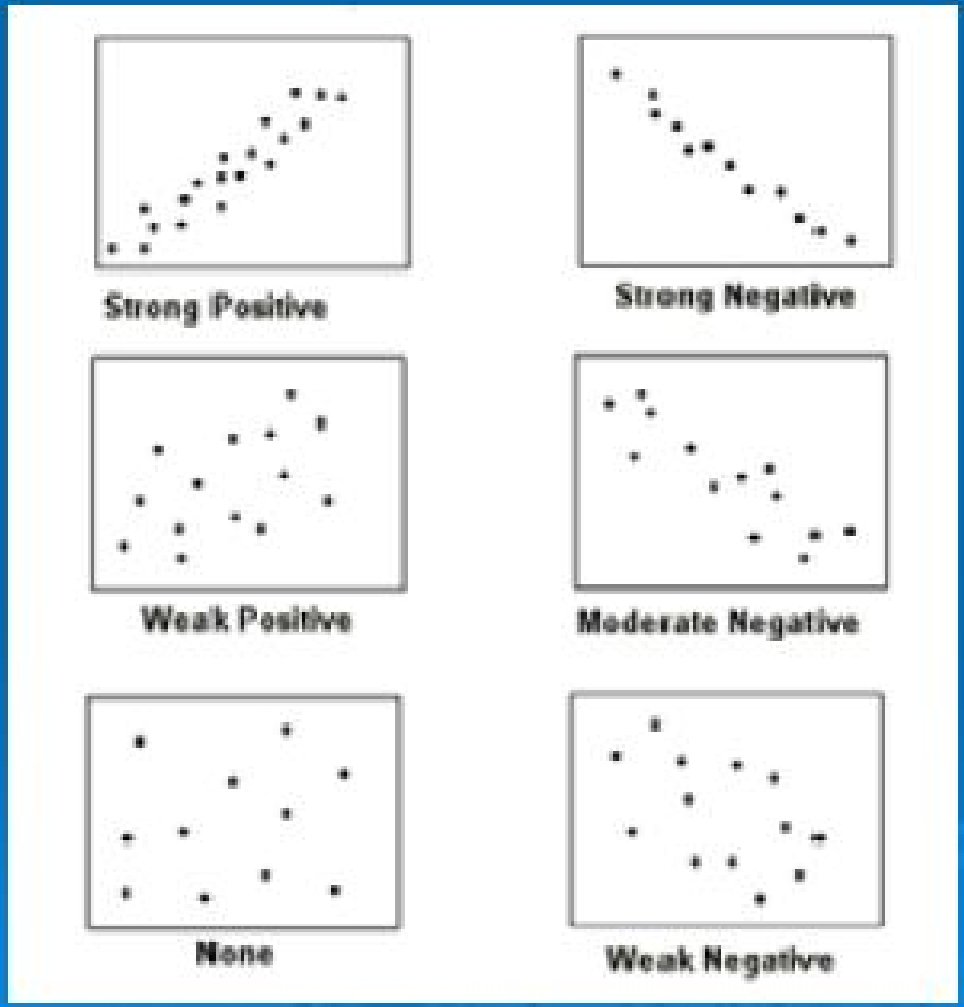
Source	DF	SS	MS	F	P
Group	2	503.5	251.8	4.99	0.018
Error	19	958.8	50.5		
Total	21	1462.4			

S = 7.104 R-Sq = 34.43% R-Sq(adj) = 27.53%

Association

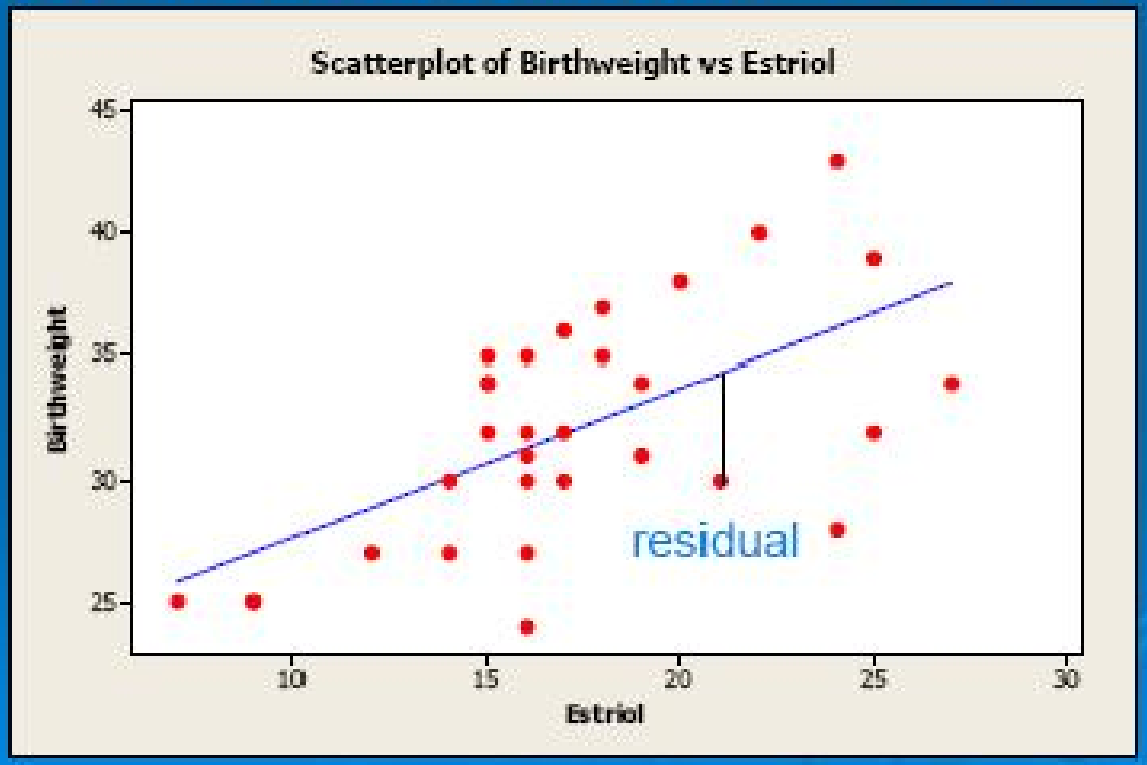
- Two quantitative variables
  - Scatter plot
  - Measure of linear association
    - Correlation r $(-1 \leqslant r \leqslant 1)$
  - regression
- Two categorical variable
  - Contingency table
  - Measure of association
    - odds ratio, relative risk, absolute risk
  - Chi-squared test for association

Degree of correlation



Obstetrics example

- X=Estriol level (mg/24hr) and Y=Birthweight (g/100)
- *r*=0.610, moderate positive, linear association



Regression: Obstetrics example

Regression Analysis: Birthweight versus Estriol

The regression equation is

Birthweight = 21.5 + 0.608 Estriol

Predictor	Coef	SE Coef	T	P
Constant	21.523	2.620	8.21	0.000
Estriol	0.6082	0.1468	4.14	0.000

S = 3.82111 R-Sq = 37.2% R-Sq(adj) = 35.0%

Analysis of Variance

Source	DF	SS	MS	F	P
Regression	1	250.57	250.57	17.16	0.000
Residual Error	29	423.43	14.60		
Total	30	674.00			

Breast cancer example

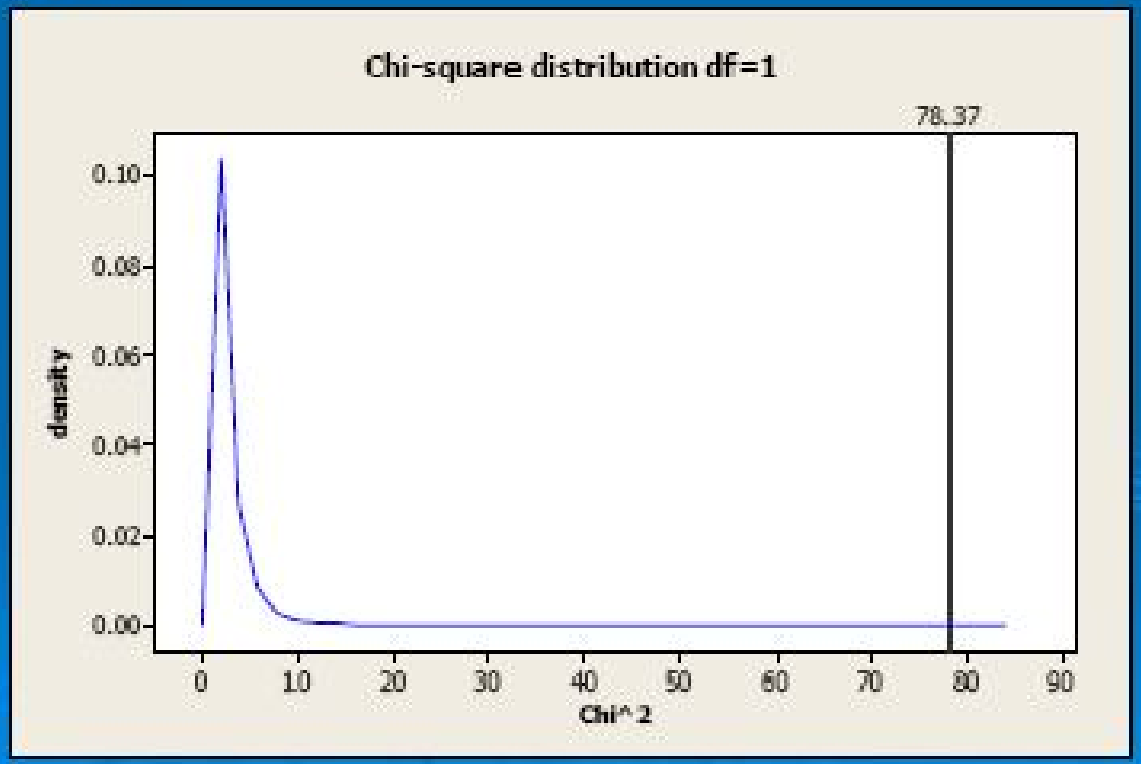
	age $\geq$ 30	age $\leq$ 29	Total
case	n_{11}=683	n_{12}=2537	$n_{1.}$ = 3220
control	n_{21}=1498	n_{22}=8747	$n_{2.}$ = 10245
Total	$n_{.1}$= 2181	$n_{.2}$= 11284	$n_{..}$ =13465

- Odds ratio = $(n_{11}/ n_{12})/ (n_{21}/ n_{22}) = n_{11} n_{22} / n_{12} n_{21}$
  = 1.57
- Chi-square test, measure discrepancy between expected cell counts under independence and observed cell counts

$$\chi^2 = \frac{(O_{11}-E_{11})^2}{E_{11}} + \frac{(O_{12}-E_{12})^2}{E_{12}} + \frac{(O_{21}-E_{21})^2}{E_{21}} + \frac{(O_{22}-E_{22})^2}{E_{22}} = 78.37$$

Chi-square test

- H_0: independent vs.
  H_a: associated
- df=(nrow-1)*(ncol-1)=1
- p-value$\approx$0
- Reject H_0

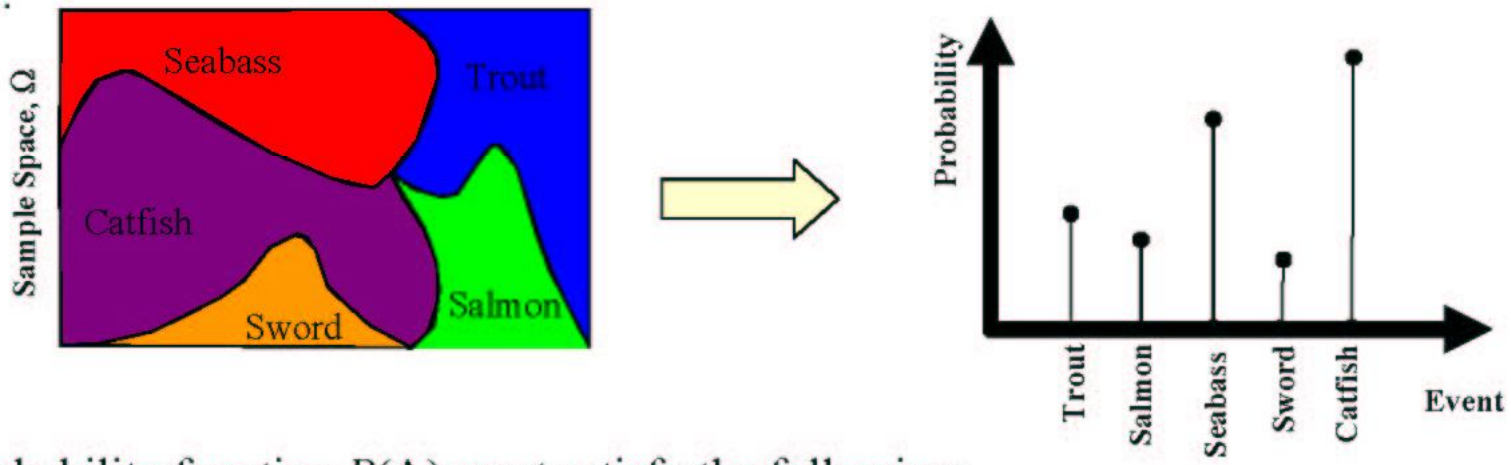


PROBABILITY THEORY

- Axioms of Probability Theory, Conditional Probability
- Discrete and continuous random variables
  - Probability mass / density / distribution function
  - Cumulative mass / density / distribution function
  - Expected value (average)
  - Variance and standard deviation
- Pairs of random variables
  - Joint probability, Joint distribution
  - Statistical independence
  - Expectation for two variables
  - Covariance / covariance matrix
  - Correlation / correlation coefficient

AXIOMS OF PROBABILITY THEORY

- The set of all possible outcomes of an experiment is the ***sample space***, denoted Ω. An ***event*** **A** is a (set of) possible outcomes of the experiment, and corresponds to a subset of Ω.
- A probability law / measure is a function $P(\mathbf{A})$ with the argument **A**, that assigns a value to **A** based on the expected proportion of number of times that event **A** is actually likely to happen.



- The probability function $P(\mathrm{A})$ must satisfy the following:
  - $0 \le P(\mathbf{A}_i) \le 1$
  - $P(\Omega) = \sum P(\mathbf{A}_i) = 1$
  - *if* $\mathbf{A}_i \bigcap \mathbf{A}_j = \phi$, *then* $P(\mathbf{A}_i \bigcup \mathbf{A}_j) = P(\mathbf{A}_i) + P(\mathbf{A}_j)$

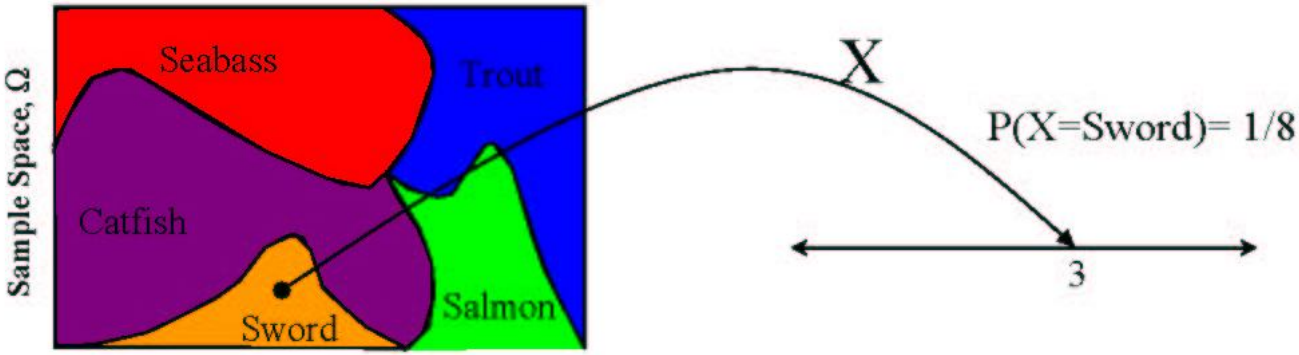
    otherwise $P(\mathbf{A}_i \bigcup \mathbf{A}_j) = P(\mathbf{A}_i) + P(\mathbf{A}_j) - P(\mathbf{A}_i \bigcap \mathbf{A}_j)$

Important Theorems of Probability

- $P(A^C) = 1 - P(A)$
- $A_i \subset A_j \Rightarrow P(A_i) < P(A_j)$
- $P(\phi) = 0$
- $P(A) \leq 1$
- $If \left\{ A_i \cap A_j = \phi, \forall i, j \right\} \Rightarrow P\left(\bigcup_{k=1}^{N} A_k \right) = \sum_{k=1}^{N} P(A_k)$

RANDOM VARIABLES

- A random variable X is a function that maps every possible event in the space Ω of a random experiment to a real number.
  - For the fish example, if **X= *the next fish selected from the pile***, we have X(Salmon) = 1, X(Trout)=2, X(Sword) = 3, X(Catfish)=4 and X(Seabass) = 5.
  - We can also assign probabilities to these events P(X=1)=1/8, P(X=2)=1/8, P(X=3)=1/8, P(X=4)=3/8, and P(X=5)=2/8.

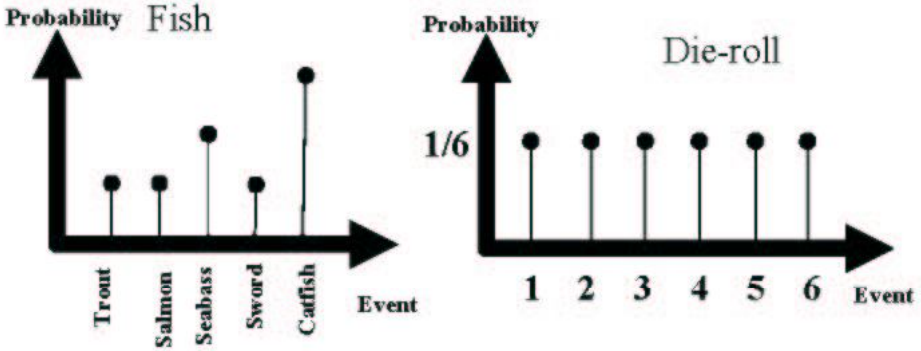


  - Random variables can discrete, e.g., the number of heads in three consecutive coin tosses, or continuous, the weight of a class member.
  - Note that, random variable is just like an ordinary variable, whose value may change based on its argument, except, now, this value is random, not deterministic.

Probability and Cumulative Mass Functions

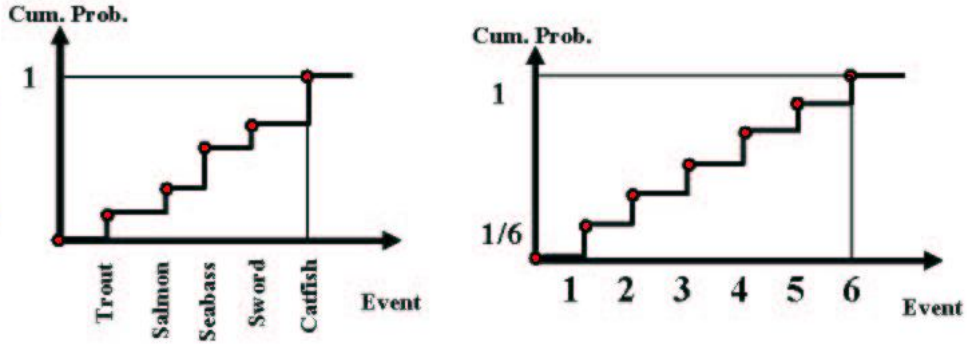
- A probability mass (distribution) function is a function that tells us the probability of *x*, an observation of X, assuming a specific value. This function also satisfies the axioms of the probability

$$P(X = x) > 0, \qquad \sum_{x \in X} P(x) = 1$$



- The cumulative mass (distribution) function indicates the probability of X assuming a value less then or equal to *x:*

$$F(x) = P(X \le x) = \sum_{u \le x} P(u)$$

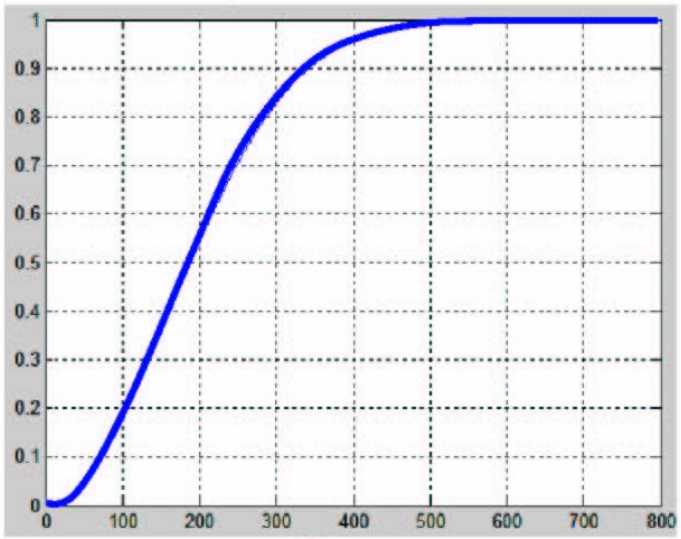


Probability and Cumulative Density Functions

- For continuous random variables, the *probability (pdf)* and *cumulative density function (cdf)* replace the mass functions
  - In continuous domain, pdf $p(x)$– if exists – is the derivative of the cdf $f(x)$.

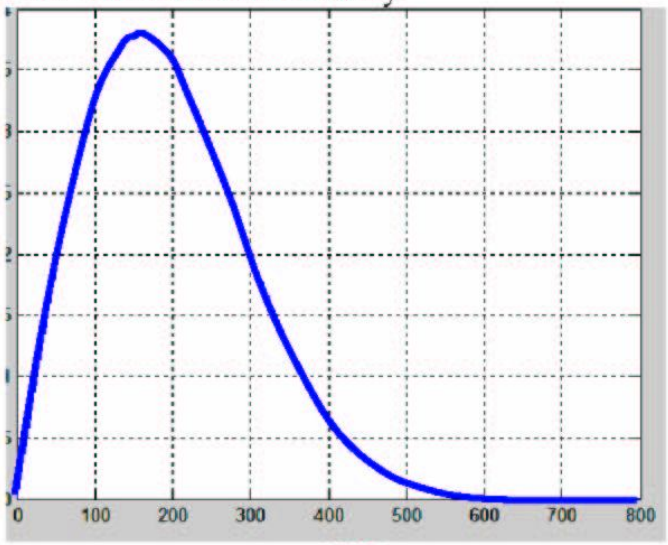
$$p(x) = \frac{df(x)}{d(x)}$$

  - In fact, even in discrete domain the pmf can also be defined similarly



cdf

$$P(x) = \frac{\Delta F(x)}{\Delta x}$$



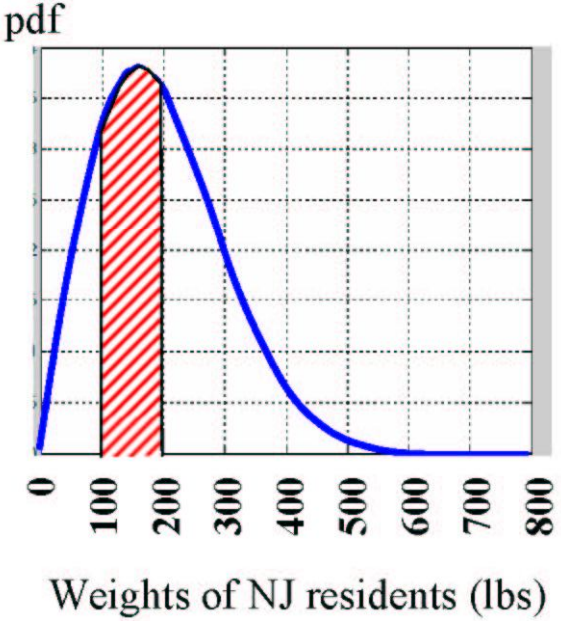
pdf

Weight of NJ residents

PDF & CDF

- Unlike the pmf in discrete case, in continuous domain, the points on the pdf do not represent the probabilities, as the probability of any single value on the continuous axis is zero, rather, the pdf represent *densities*.
- For example, P(weight = 175.24654163546546874876454687987652354354385424148454257654)=0
- The area under the pdf curve between any two abscissas gives the probability that the value lie between those two points:

$$P(a < x < b) = \int_{x=a}^{b} p(x)dx$$



Expected Value (Mean)

- The expected value, or average, of a random variable X, whose possible values are $\{x_1, \ldots, x_m\}$ with respective probabilities $p_1, \ldots, p_m$, is given as

$$E(x) = \mu = \sum_{x \in X} x \cdot P(x) = \sum_{i=1}^{m} x_i \cdot p_i$$

- In general, the expected value of a stochastic (random) function *f(x)* is given as

$$E(f(x)) = \sum_{x \in X} f(x) \cdot P(x)$$

- It is a linear function, and it can be used to compute *moments* of a random variable

$$E(x^k) = \sum_{x \in X} x^k \cdot P(x) = \sum_{i=1}^{m} (x_i)^k \cdot p_i$$

- A special, mean subtracted, form of the second moment is the *variance*, the average dispersion of the data from the mean

$$Var[x] = \sigma^2 = E\left((x-\mu)^2\right) = \sum_{x \in X} (x-\mu)^2 \cdot P(x)$$

Mean & Variance in Continuous Domain

- The expected value, or average, of a continuous random variable X, whose pdf is given by $p(x)$ is computed as

$$E(x) = \mu = \int_{-\infty}^{\infty} x \cdot p(x) \cdot dx$$

- The variance is then computed as the mean-removed 2nd moment:

$$Var[x] = \sigma^2 = E\left((x-\mu)^2\right) = \int_{x=-\infty}^{\infty} (x-\mu)^2 \cdot p(x) \cdot dx$$

- The standard deviation is simply the positive square root of the variance. The kth moment in general is obtained as

$$E(x^k) = \int_{x=-\infty}^{\infty} x^k \cdot p(x) \cdot dx$$

- As usual, the mean represents the center of mass of the density, and the variance represent the average dispersion of the density around the mean.

PAIRS OF RANDOM VARIABLES

- If we have two random variables, X and Y, assuming values from the sets $\{x_1, \ldots, x_m\}$ and $\{y_1, \ldots, y_n\}$, we can define a joint probability for each pair of values:

$$p_{ij} = P(x = x_i; y = y_j)$$

- Joint probability also need to satisfy the axioms of the probability theory

$$P(x, y) > 0 \quad \text{and} \quad \sum_{x \in X} \sum_{y \in Y} P(x, y) = 1$$

- Everything that relate to X, or Y – individually or together – can be obtained from the P(x,y). In particular, the individual pmfs, called the *marginal distribution functions* can be obtained as

$$P_x(x) = \sum_{y \in Y} P(x, y) \qquad P_y(y) = \sum_{x \in X} P(x, y)$$

STATISTICAL INDEPENDENCE

- Random variables X and Y are said to be *statistically independent*, if and only if

$$P(x,y) = P_x(x) \cdot P_y(y)$$

  - That is, if the outcome of one event does not effect the outcome of the other, they are statistically independent. For example, the outcome of two individual dice are independent, as one does not affect the other.

- Expected values, moments and variances of joint distributions can be computed similar to single variable cases:

$$\mu_x = \sum_x \sum_y x \cdot P(x,y) \qquad \mu_y = \sum_x \sum_y y \cdot P(x,y)$$

$$\sigma_x^2 = E\left[(x-\mu_x)^2\right] = \sum_x \sum_y (x-\mu_x)^2 \cdot P(x,y)$$

$$\sigma_y^2 = E\left[(y-\mu_y)^2\right] = \sum_x \sum_y (y-\mu_y)^2 \cdot P(x,y)$$

CO-VARIANCE

- A *cross-moment* can also be defined as the *covariance*

$$\sigma_{xy}^2 = E\left[(x-\mu_x)(y-\mu_y)\right] = \sum_x \sum_y (x-\mu_x)\cdot(x-\mu_x)\cdot P(x,y)$$

  - Covariance defines how the variables vary together as a pair – are they both increasing together, does one increase when the other decrease, etc.
  - Note that we can also define $\sigma_{xx} = \sigma_x^2$, $\sigma_{yy} = \sigma_y^2$, and $\sigma_{xy} = \sigma_{yx}$, all which can be represented with a single matrix, ***the covariance matrix***, denoted by $\sum$

$$\Sigma = \begin{bmatrix} \sigma_{xx} & \sigma_{xy} \\ \sigma_{yx} & \sigma_{yy} \end{bmatrix}$$

  - If X and Y are statistically independent ➔ $\sigma_{xy}=0$
  - If $\sigma_{xy}=0$ ➔ then the variables are said to be ***uncorrelated***
  - Note that statistical independence is a stronger property then correlation:

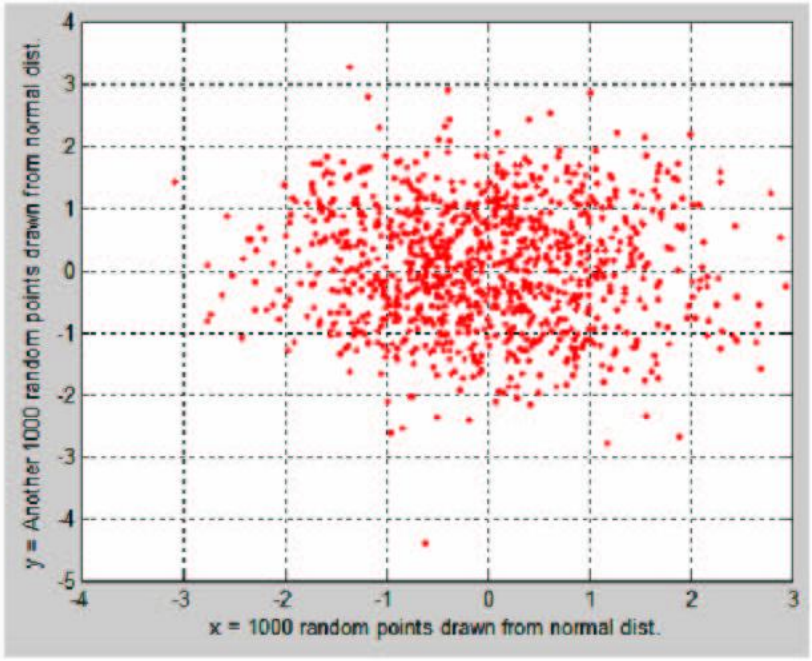
statistical independence  uncorrelated

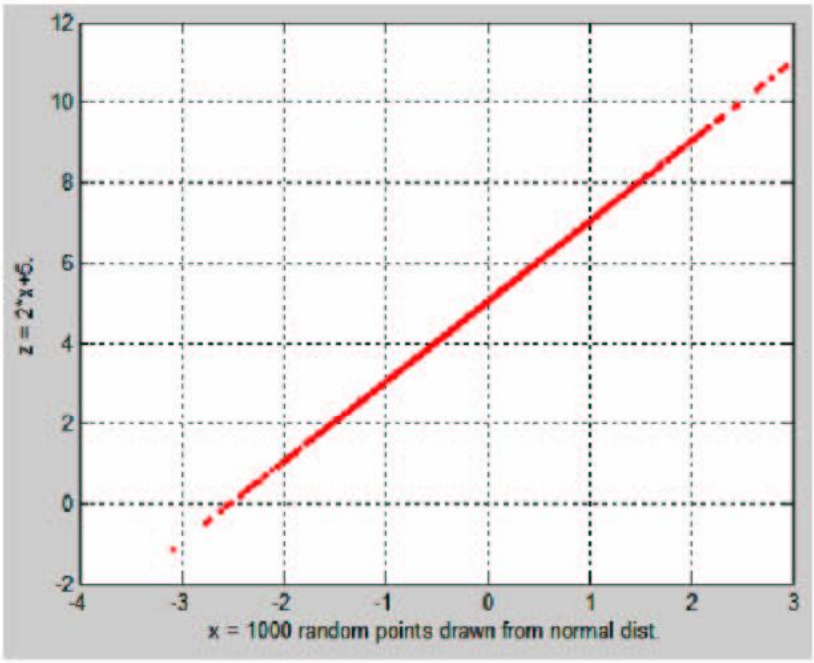
CORRELATION COEFFICIENT

$$\rho = \frac{\sigma_{xy}}{\sigma_x \sigma_y} \qquad -1 \le \rho \le 1$$

- If ρ=1, then the variables are identical, they move together,
- If ρ=-1, then the variables are negatively correlated, one decreases as the other increases at the same rate
- If ρ=0 ➔ the variables are uncorrelated. The variation of one, has no effect on the other.
- For all practical purposes, if $|\rho| < 0.05$, the variables are considered to be uncorrelated.

CORRELATION





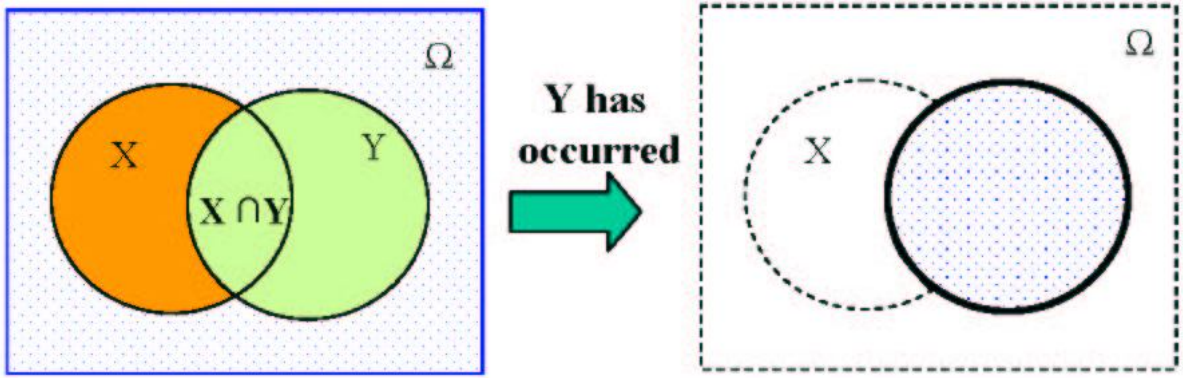
$$\Sigma_{xy} = \begin{bmatrix} 1.073 & -0.026 \\ -0.0264 & 0.9673 \end{bmatrix} \qquad \Sigma_{xz} = \begin{bmatrix} 1.073 & 2.1476 \\ 2.1476 & 4.2951 \end{bmatrix}$$

$$\rho_{xy} = -0.0259$$

CONDITIONAL PROBABILITY

- If two variables are statistically dependent, knowing the value of one may allow us to better estimate the other:
- The conditional probability of $X=x$ given the $Y=y$ has been observed is given as

$$P(X=x \mid Y=y) = \frac{P(X=x \cap Y=y)}{P(Y=y)} = \frac{P(x,y)}{P(y)} \Rightarrow P(x,y) = P(x \mid y)P(y)$$

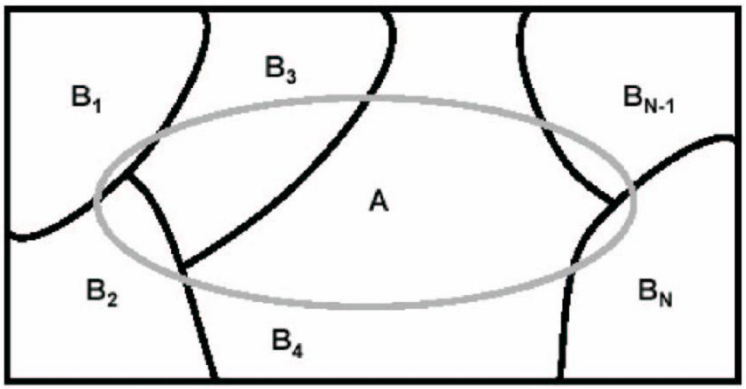


- The fact that Y=y has been observed has two main consequences:
  - The sample space effectively becomes the space of Y
  - The event X=x, effectively becomes X∩Y, that is P(y) renormalizes the probability of events that occur jointly with Y

Law of Total Probability

- Let $B_{1, \ldots} B_N$ be N mutually exclusive events, whose union gives the sample space Ω. Hence the events B constitute a partition of Ω
- Now consider an event A, a subset of Ω. This event can be represented as

$$A = A \cap \Omega = A \cap (B_1 \cup B_2 \cup \ldots \cup B_N) = (A \cap B_1) \cup (A \cap B_2) \cup \ldots (A \cap B_N)$$



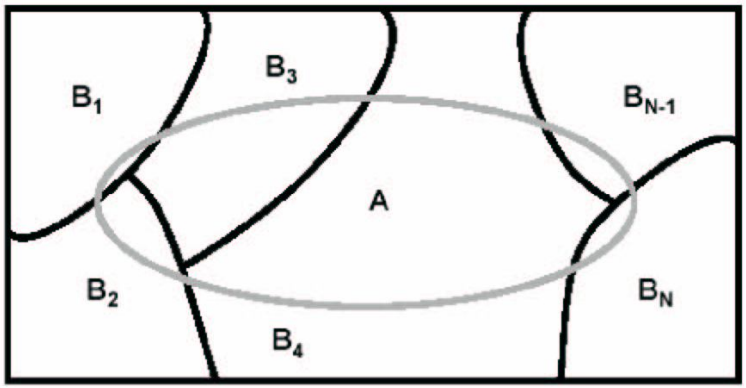
- Since the B_i are mutually exclusive

$$P(A) = P(A \cap B_1) + P(A \cap B_2) + \ldots + P(A \cap B_N)$$

$$P(A) = P(A \mid B_1)P(B_1) + \ldots + P(A \mid B_N)P(B_N) = \sum_{k=1}^{N} P(A \mid B_k)P(B_k)$$

Bayes Rule

- We now pose the following question: Given that the event A has occurred. What is the probability that any single one of the event B's occur?



$$P(B_j \mid A) = \frac{P(A \cap B_j)}{P(A)} = \frac{P(A \mid B_j) \cdot P(B_j)}{\sum_{k=1}^{N} P(A \mid B_k) \cdot P(B_k)}$$

This is known as the Bayes rule

Rev. Thomas Bayes,
(1702-1761)

Example

3. It is known that 1% of population suffers from a particular disease. A blood test has a 97% chance to identify the disease for a diseased individual, by also has a 6% chance of falsely indicating that a healthy person has a disease.

a. What is the probability that a random person has a positive blood test.

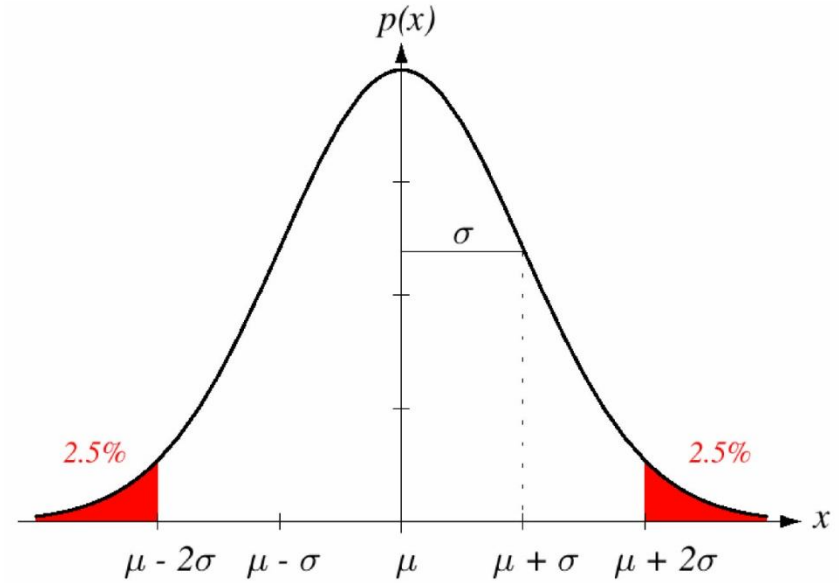
b. If a blood test is positive, what's the probability that the person has the disease?

c. If a blood test is negative, what's the probability that the person does not have the disease?

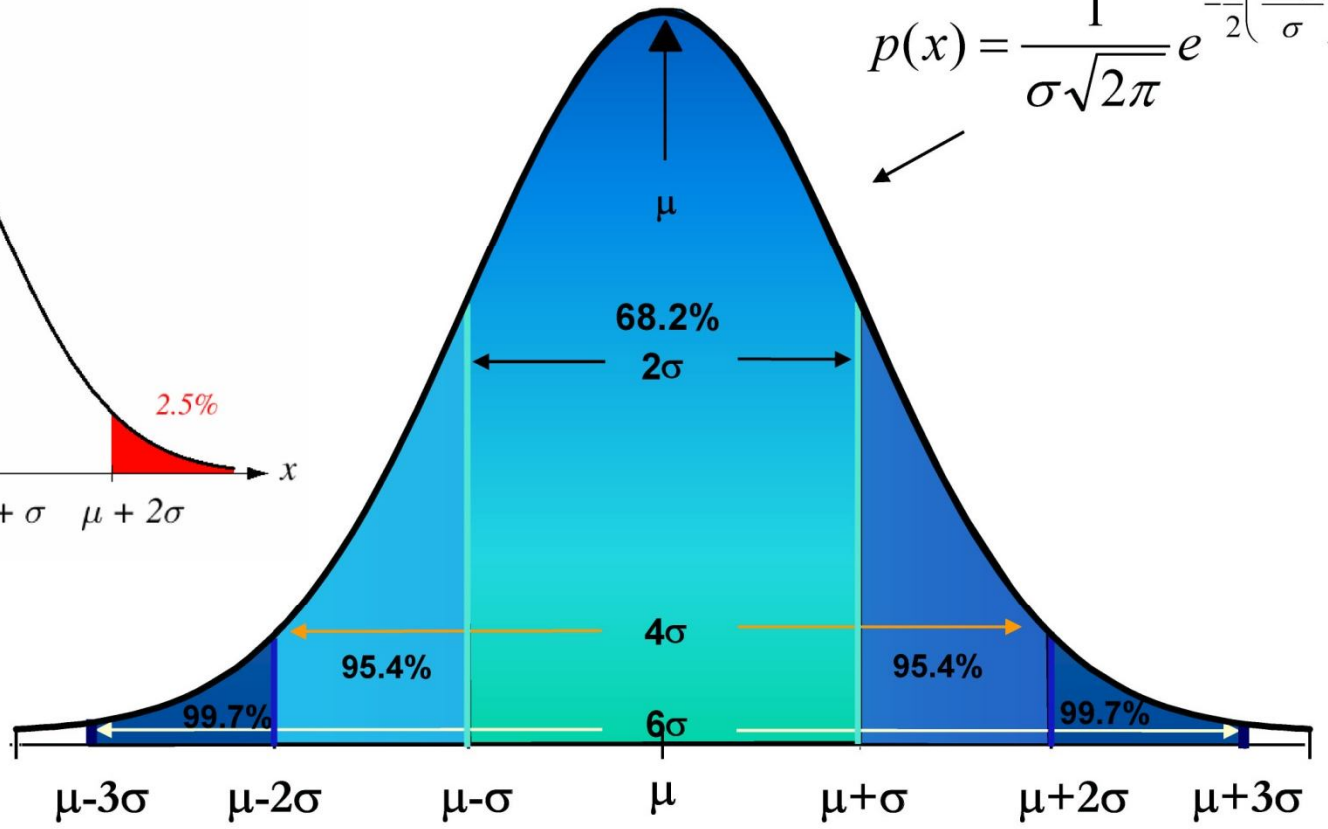
- A is the event that a person has a disease. P(A) = 0.01; P(A') = 0.99.
- B is the event that the test result is positive.
  - P(B|A) = 0.97; P(B'|A) = 0.03;
  - P(B|A') = 0.06; P(B'|A') = 0.94;
- (a) P(B) = P(A) P(B|A) + P(A')P(B|A') = 0.01*0.97 +0.99 * 0.06 = 0.0691
- (b) P(A|B)=P(B|A)*P(A)/P(B) = 0.97* 0.01/0.0691 = 0.1403
- (c) P(A'|B') = P(B'|A')P(A')/P(B')= P(B'|A')P(A')/(1-P(B))= 0.94*0.99/(1-.0691)=0.9997

GAUSSIAN DISTRIBUTION

- By far the most important and most commonly observed (cont.) probability distribution



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



Normal Distributions

- Gaussian distribution

$$p(x) = N(\mu_x, \sigma_x) = \frac{1}{\sqrt{2\pi}\sigma_x} e^{-(x-\mu_x)^2 / 2\sigma_x^2}$$

- Mean $E(x) = \mu_x$

- Variance $E[(x-\mu_x)^2] = \sigma_x^2$

- Central Limit Theorem says sums of random variables tend toward a Normal distribution.

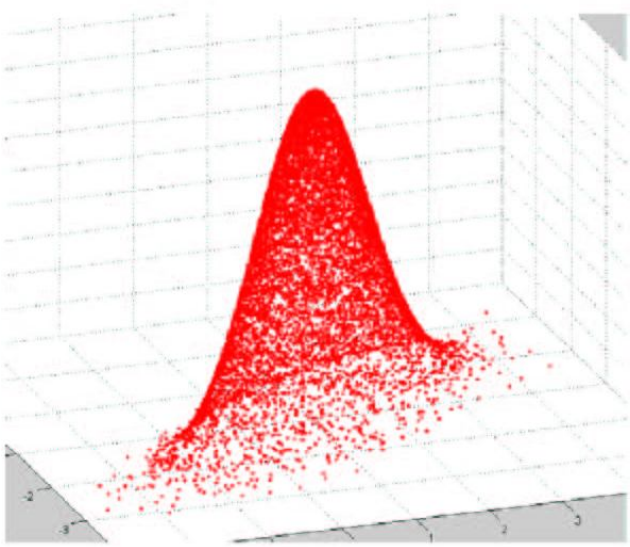
- Mahalanobis Distance: $r = \frac{x-\mu_x}{\sigma_x}$

Multivariate Gaussian Distribution

In *d-dimensional* space, the Gaussian pdf is:

$$p(\mathbf{x}) = \frac{1}{(2\pi)^{d/2}|\boldsymbol{\Sigma}|^{1/2}} e^{-\frac{1}{2}\left[(\mathbf{x}-\boldsymbol{\mu})^T \boldsymbol{\Sigma}^{-1}(\mathbf{x}-\boldsymbol{\mu})\right]}$$

$$p(\mathbf{x}) \sim N(\boldsymbol{\mu}, \boldsymbol{\Sigma})$$



- **Gaussian distributions are very popular since**
  - The parameters (μ,Σ) are **sufficient** to uniquely characterize the normal distribution
  - If the $\mathbf{x}_i$'s are mutually **uncorrelated** (σ_{ik}=0), then they are also **independent**
    - The covariance matrix becomes a diagonal matrix, with the individual variances in the main diagonal
  - **Central Limit Theorem**
  - The **marginal and conditional densities** are also Gaussian
  - Any **linear transformation** of any N jointly Gaussian rv's results in N rv's that are also Gaussian
    - For $X=[X_1\ X_2\ \ldots\ X_N]^T$ jointly Gaussian, and A an N×N invertible matrix, then Y=AX is also jointly Gaussian

$$p_Y(y) = \frac{p_X(A^{-1}y)}{|A|}$$

Multivariate Normal Density

- x is a vector of d Gaussian variables

$$p(x)=N(\mu,\Sigma)=\frac{1}{2\pi^{d/2}|\Sigma|^{1/2}}e^{-\frac{1}{2}(x-\mu)^T\Sigma^{-1}(x-\mu)}$$

$$\mu=E[x]=\int_{-\infty}^{\infty}xp(x)dx$$

$$\Sigma=E[(x-\mu)(x-\mu)^T]=\int_{-\infty}^{\infty}(x-\mu)(x-\mu)^T p(x)dx$$

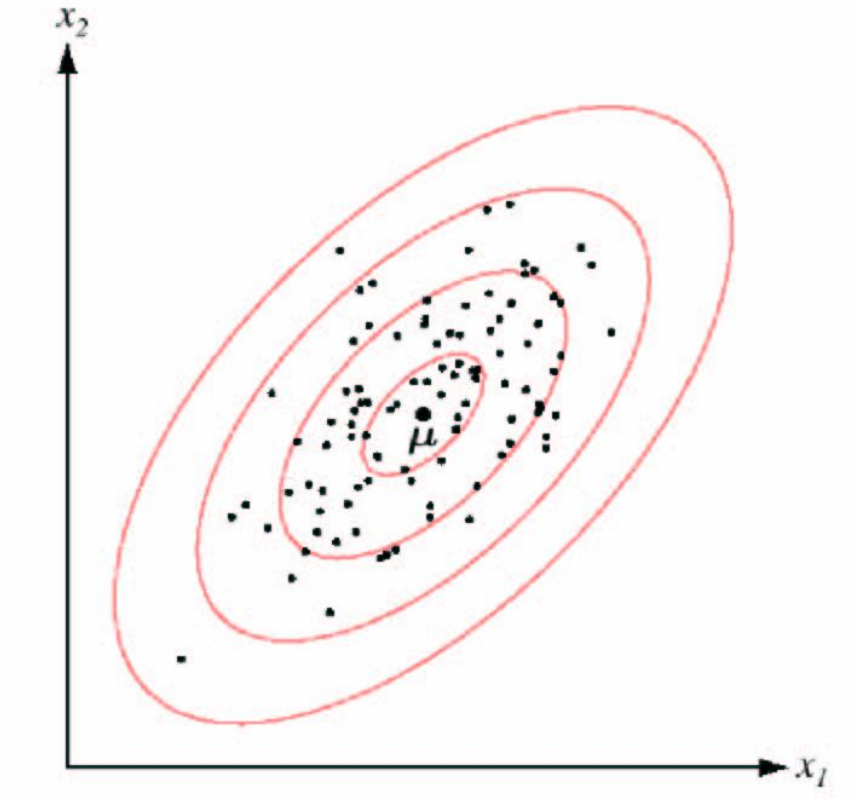
- Mahalanobis Distance

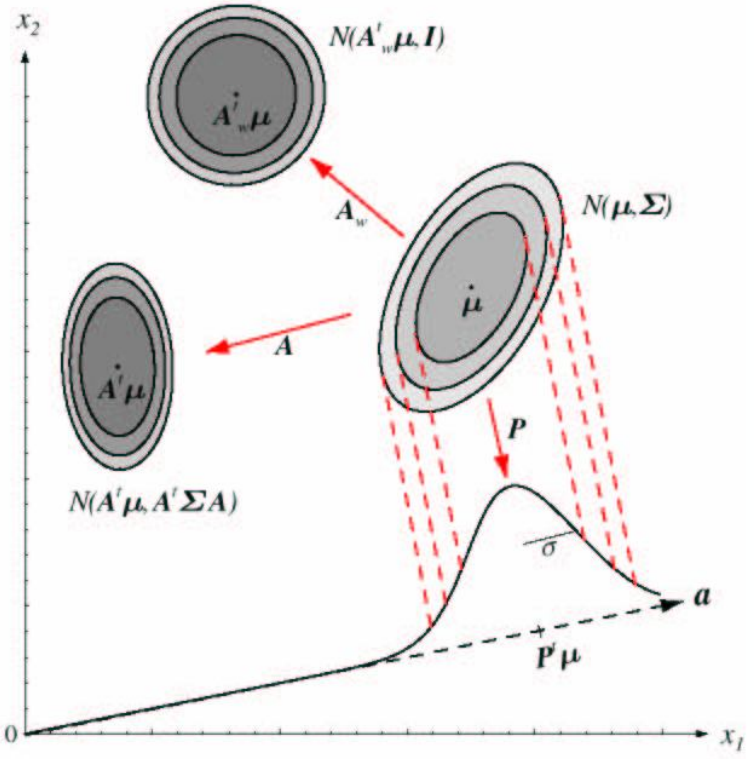
$$r^2=(x-\mu)^T\Sigma^{-1}(x-\mu)$$

- All conditionals and marginals are also Gaussian

MULTIVARIATE GAUSSIAN DISTRIBUTION

- Multivariate normal density function
- Mahalanobis distance
- Whitening Transform





Bayesian Decision Making

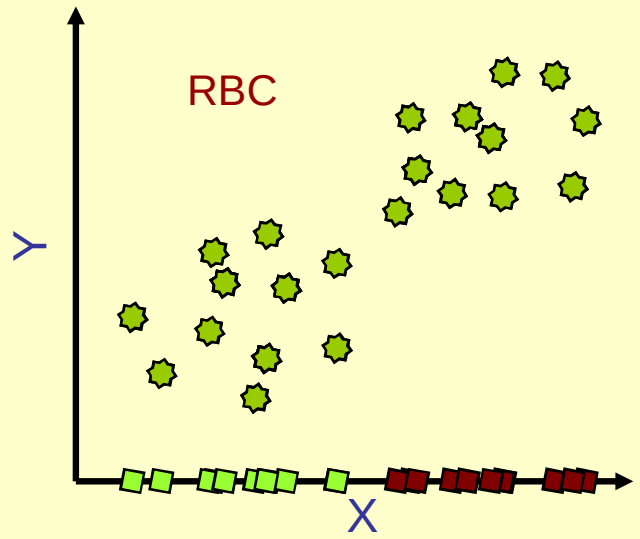
Classification problem in probabilistic terms

Create models for how features are distributed for objects of different classes

We will use probability calculus to make classification decisions

Lets Look at Just One Feature

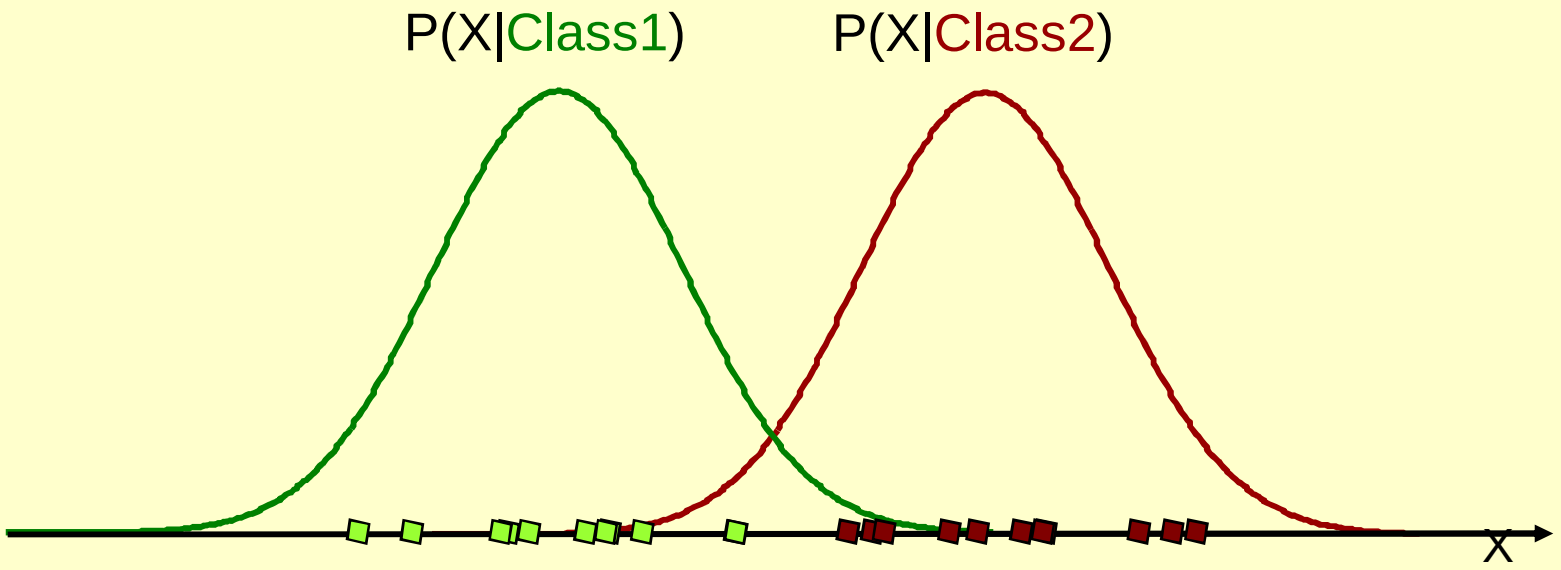
- Each object can be associated with multiple features
- We will look at the case of just one feature for now



We are going to define two key concepts….

The First Key Concept

Features for each class drawn from class-conditional probability distributions (CCPD)



Our first goal will be to *model* these distributions

The Second Key Concept

We model prior probabilities to quantify the expected *a priori* chance of seeing a class

P(Class2) & P(Class1)

But How Do We Classify?

- **So we have priors defining the *a priori* probability of a class**

P(Class1), P(Class2)

- **We also have models for the probability of a feature given each class**

P(X|Class1), P(X|Class2)

But we want the probability of the class given a feature
How do we get P(Class1|X)?

Bayes Rule

Belief before evidence

Evaluate evidence

Likelihood

Prior

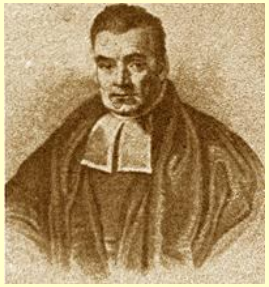
$$P(Class \mid Feature) = \frac{P(Feature \mid Class)P(Class)}{P(Feature)}$$

Belief after evidence

Posterior

Evidence

Bayes, Thomas (1763) An essay towards solving a problem in the doctrine of chances. Philosophical Transactions of the Royal Society of London, 53:370-418



Bayes Decision Rule

If we observe an object with feature X, how do decide if the object is from Class 1?

The Bayes Decision Rule is simply choose Class1 if:

$$P(Class1 \mid X) > P(Class2 \mid X)$$

$$\frac{P(X \mid Class1)P(L1)}{P(X)} > \frac{P(X \mid Class2)P(L2)}{P(X)}$$

This is the same number on both sides!

Discriminant Function

We can create a convenient representation of the Bayes Decision Rule

$$P(X \mid Class1)P(Class1) > P(X \mid Class2)P(Class2)$$

$$\frac{P(X \mid Class1)P(Class1)}{P(X \mid Class2)P(Class2)} > 1$$

$$G(X) = \log \frac{P(X \mid Class1)}{P(X \mid Class2)} \frac{P(Class1)}{P(Class2)} > 0$$

If G(X) > 0, we classify as Class 1

Stepping back

What do we have so far?

We have defined the two components, class-conditional distributions and priors

P(X|Class1), P(X|Class2) P(Class1), P(Class2)

We have used Bayes Rule to create a discriminant function for classification from these components

$$G(X) = \log \frac{P(X \mid Class1)}{P(X \mid Class2)} \frac{P(Class1)}{P(Class2)} > 0$$

Given a new feature, X, we plug it into this equation…

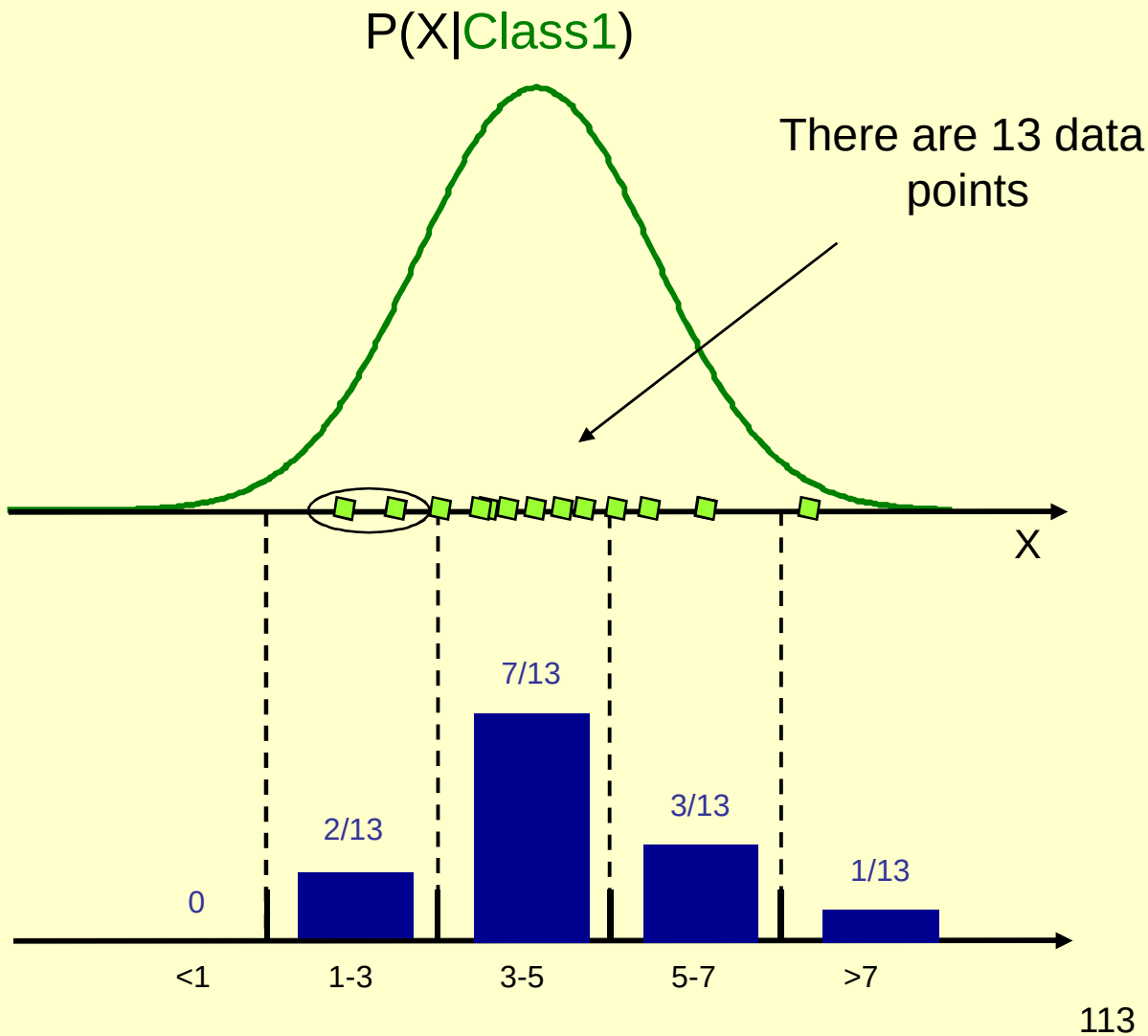
…and if G(X)> 0 we classify as Class1

Getting P(X|Class) from Training Set

One Simple Approach

Divide X values into bins

And then we simply count frequencies



Class conditional from Univariate Normal Distribution

$$p(x) = N(\mu_x, \sigma_x) = \frac{1}{\sqrt{2\pi}\sigma_x} e^{-(x-\mu_x)^2 / 2\sigma_x^2}$$

Mean :

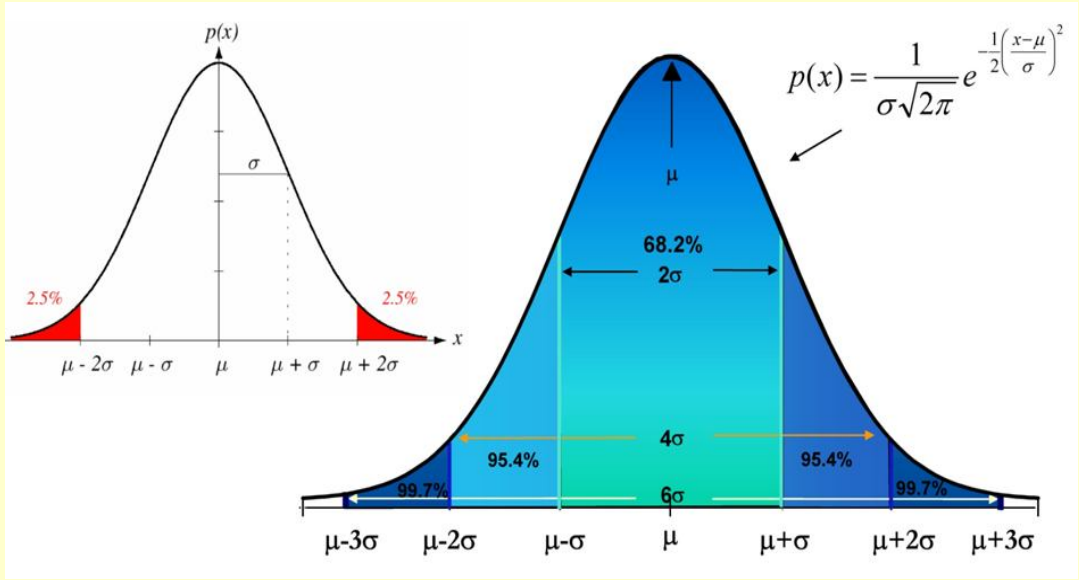
$$E(x) = \mu_x$$

Variance :

$$E[(x-\mu_x)^2] = \sigma_x^2$$

Mahalanobis Distance :

$$r = \frac{x-\mu_x}{\sigma_x}$$



We Are Just About There….

We have created the class-conditional distributions and priors

P(X|Class1), P(X|Class2) P(Class1), P(Class2)

And we are ready to plug these into our discriminant function

$$G(X) = \log \frac{P(X \mid Class1)}{P(X \mid Class2)} \frac{P(Class1)}{P(Class2)} > 0$$

But there is one more little complication…..

Multidimensional feature space ?

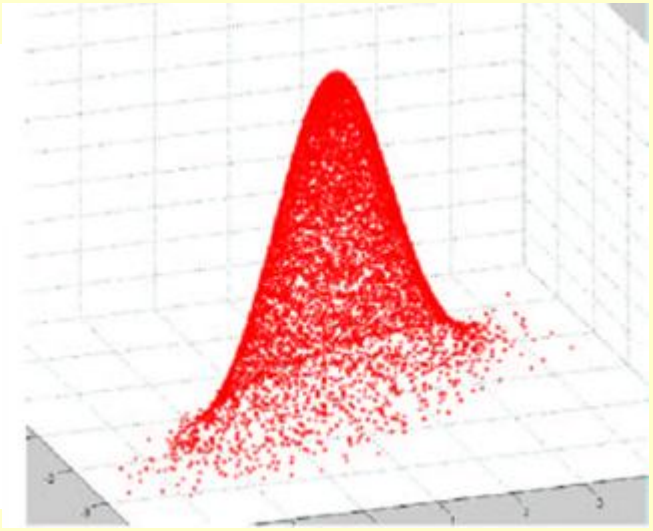
So P(X|Class) become P(X1,X2,X3,…,X8|Class)
and our discriminant function becomes

$$G(X) = \log \frac{P(X_1, X_2, ..., X_7 \mid Class1)}{P(X_1, X_2, ..., X_7 \mid Class2)} \frac{P(Class1)}{P(Class2)} > 0$$

In *d-dimensional* space, the Gaussian pdf is:

$$p(\mathbf{x}) = \frac{1}{(2\pi)^{d/2}|\boldsymbol{\Sigma}|^{1/2}} e^{-\frac{1}{2}\left[(\mathbf{x}-\boldsymbol{\mu})^T \boldsymbol{\Sigma}^{-1}(\mathbf{x}-\boldsymbol{\mu})\right]}$$

$$p(\mathbf{x}) \sim N(\boldsymbol{\mu}, \boldsymbol{\Sigma})$$



Naïve Bayes Classifier

We are going to make the following assumption:

All features are independent given the class

$$P(X_1, X_2, ..., X_n \mid Class) = P(X_1 \mid Class)P(X_2 \mid Class)...P(X_n \mid Class)$$

$$= \prod_{i=1}^{n} P(X_i \mid Class)$$

We can thus estimate individual distributions for each feature and just multiply them together!

Naïve Bayes Discriminant Function

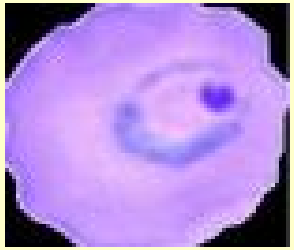
Thus, with the Naïve Bayes assumption, we can now rewrite, this:

$$G(X_1,...,X_7) = \log \frac{P(X_1, X_2,..., X_7 \mid Class1)}{P(X_1, X_2,..., X_7 \mid Class2)} \frac{P(Class1)}{P(Class2)} > 0$$

As this:

$$G(X_1,...,X_7) = \log \frac{\prod P(X_i \mid Class1)}{\prod P(X_i \mid Class2)} \frac{P(Class1)}{P(Class2)} > 0$$

Classifying Parasitic RBC



- Intensity
- Hu's moments
- Fractal dimension
- Entropy
- Homogeneity
- Correlation
- Chromatin dots

X_i

$P(X_i|Malaria)$

$P(X_i|\sim Malaria)$

Plug these and priors into the discriminant function

$$G(X_1,...,X_7) = \log \frac{\prod P(X_i \mid Mito)}{\prod P(X_i \mid \sim Mito)} \frac{P(Mito)}{P(\sim Mito)} > 0$$

IF G>0, we predict that the parasite is from class Malaria

How Good is the Classifier?

The Rule

We *must* test our classifier on a different set from the training set: the labeled test set

The Task

We will classify each object in the test set and count the number of each type of error

Binary Classification Errors

	True (Mito)	False (~Mito)
Predicted True	TP	FP
Predicted False	FN	TN

Sensitivity = TP/(TP+FN) Specificity = TN/(TN+FP)

- Sensitivity
  - Fraction of all Class1 (True) that we correctly predicted at Class 1
  - *How good are we at finding what we are looking for*

- Specificity
  - Fraction of all Class 2 (False) called Class 2
  - *How many of the Class 2 do we filter out of our Class 1 predictions*

In both cases, the higher the better

Thank you