Fundamentals to Biostatistics

Prof. Chandan Chakraborty
Associate Professor
School of Medical Science & Technology
IIT Kharagpur
collection, analysis, interpretation of data

development of new statistical theory & inference

application of the methods derived from math. statistics to subject specific areas like psychology, economics and public health

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.

\[ 5.53 = 30^{th} \text{ percentile} \]
Measures of spread: IQR

- The 25th ($Q_1$), 50th (median), and 75th ($Q_3$) 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, Q₁, Median, Q₃, 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

- Median
- Q1
- Q3
- Suspected outliers
- Whiskers
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

- $r$ 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)

<table>
<thead>
<tr>
<th></th>
<th>Age</th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>≥30</td>
<td>≤29</td>
</tr>
<tr>
<td>BC</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>case</td>
<td></td>
<td>683 (21.2%)</td>
<td>2537</td>
</tr>
<tr>
<td>control</td>
<td></td>
<td>1498 (14.6%)</td>
<td>8747</td>
</tr>
<tr>
<td>Total</td>
<td></td>
<td>2181</td>
<td>11284</td>
</tr>
</tbody>
</table>

(MacMahon, B. et al., 1970, Rosner, B. 1985, p346)
How does taking Oral Contraceptive (OC) affect Blood Pressure (BP) in women

- paired samples

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

(Rosner, B. 1995, p253)
Birthweight Example

- Determine the effectiveness of drug A on preventing premature birth.
- Independent samples.
Histogram and Distribution (Blood pressure)
Histogram and Distribution (Birthweight)
Normal distribution

- $N(\mu, \sigma)$: mean $\mu$, standard deviation $\sigma$
- $N(0, 1)$: Standard normal
- Standardization: $X \rightarrow Z = (X - \mu) / \sigma$
Binomial distribution

- Two possible outcome for each trial:
  1. “success” and (0) “failure” \( \Pr(\text{success}) = p \)
  2. Example: each women with breast cancer had first birth either before age 30 or after.

- \( n \) independent trials, \( X = \# \text{ of successes} \sim B(n, p) \)
  1. Example: number of cases whose age at first birth \( \approx 30 \)

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

\[
\mu_X = np, \quad \sigma^2_X = np(1 - p)
\]
Characteristics of population distribution (parameters)

- **Center/location**
  - Population mean $\mu$, or “expected value”, “average”
  - Population median $M$, or “middle number”

- **Spread**
  - Range $R = \text{max} - \text{min}$
  - Variance $\sigma^2$, average squared distance of each value from the mean
  - Standard deviation $\sigma$

- **Coefficient of Variation (CV):** $\frac{\mu}{\sigma}$
  - 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 $R = \text{max} - \text{min}$
  - 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**

<table>
<thead>
<tr>
<th>Birthweight</th>
<th>n</th>
<th>Mean $\bar{x}$</th>
<th>$s$</th>
<th>Minimum</th>
<th>Median</th>
<th>Maximum</th>
</tr>
</thead>
<tbody>
<tr>
<td>Treatment</td>
<td>15</td>
<td>7.08</td>
<td>0.899</td>
<td>5.5</td>
<td>7.2</td>
<td>8.6</td>
</tr>
<tr>
<td>Control</td>
<td>15</td>
<td>6.26</td>
<td>0.961</td>
<td>4.2</td>
<td>6.4</td>
<td>8.2</td>
</tr>
</tbody>
</table>

- **Sample proportion** $\hat{p} = x / n$
Sampling distribution for statistics

- Sampling variability for statistics,
  - Standard error of sample mean: \( SE(\bar{x}) = \frac{s}{\sqrt{n}} \)
- If \( x \sim N(\mu, \sigma) \), then \( \bar{x} \sim N(\mu, \sigma/\sqrt{n}) \)
  \[
  \frac{\bar{x} - \mu}{\sigma/\sqrt{n}} \sim N(0,1) \quad \frac{\bar{x} - \mu}{s/\sqrt{n}} \sim t_{df=n-1}
  \]
- Central Limit Theorem: if \( n \) is large enough, \( \bar{x} \) is approximately normally distributed
  \[
  \frac{\bar{x} - \mu}{\sigma/\sqrt{n}} \sim N(0,1) \quad \frac{\bar{x} - \mu}{s/\sqrt{n}} \sim t_{df=n-1}
  \]
- 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
  \[ X_1 \sim N(7.08, 0.90), \quad n = 15 \]
- Sampling distribution
  \[ \bar{X}_1 \sim N(7.08, 0.90 / \sqrt{15}) = N(7.08, 0.23) \]
Statistical Inference

- **Point estimation**
  - Estimate parameter $\theta$ by $\hat{\theta}$, e.g., estimate $\mu$ by $\bar{x}$

- **Interval estimation: Confidence intervals (CI)**
  - An interval of plausible values of the parameter $\theta$ with a specific confidence level $1 - \alpha$; e.g., 95% CI
  - $\hat{\theta} \pm \text{margin of error} = \hat{\theta} \pm c_{1-\alpha} \text{SE}(\hat{\theta})$

- **Hypothesis testing (HT)**;
  - Testing hypotheses about parameter $\theta$
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 $\alpha$
  - $p$-value $< \alpha$, reject $H_0$
  - $p$-value $\geq \alpha$, do not reject $H_0$

- conclusion
**Statistic**

Mean: $\bar{X}$ estimates $\mu$

Standard deviation: $s$ estimates $\sigma$

Proportion: $p$ estimates $\pi$

**Parameter**

from sample

from entire population
Mean, $\mu$, is unknown

Population

Point estimate

$\bar{X} = 50$

Sample

Interval estimate

I am 95% confident that $\mu$ is between 40 & 60
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

\[ SE \ (\text{Mean}) = \frac{S}{\sqrt{n}} \]

Qualitative Variable

\[ SE \ (p) = \sqrt{\frac{p(1-p)}{n}} \]
Confidence Interval

\[ \bar{X} - 1.96 \times SE \quad \text{to} \quad \bar{X} + 1.96 \times SE \]

\[ (1 - \alpha) \]

\[ \alpha/2 \]

\[ \mu \]

\[ Z \text{-axis} \]

95% Samples
Confidence Interval

\[ p \pm 1.96 \times SE \]

\[ 1 - \alpha \]

\[ \alpha/2 \]

\[ \alpha/2 \]

95\% Samples

\[ p - 1.96 \times SE \]

\[ p + 1.96 \times SE \]
Interpretation of CI

Probabilistic

In repeated sampling 100(1-\(\alpha\))% of all intervals around sample means will in the long run include \(\mu\)

Practical

We are 100(1-\(\alpha\))% confident that the single computed CI contains \(\mu\)
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.

\[ \mu = \bar{X} \pm Z \times SE \]

\[ SE = \frac{10}{\sqrt{100}} = 1 \]

Then CI:

\[ \mu = 170 \pm 1.96 \times 1 \quad 168.04 \leq \mu \leq 171.96 \]
In a survey of 140 asthmatics, 35% had allergy to house dust. Construct the 95% CI for the population proportion.

\[ \pi = p \pm Z \sqrt{\frac{p(1-p)}{n}} \]

\[ SE = \sqrt{\frac{0.35(1-0.35)}{140}} = 0.04 \]

\[ 0.35 - 1.96 \times 0.04 \leq \pi \leq 0.35 + 1.96 \times 0.04 \]

\[ 0.27 \leq \pi \leq 0.43 \]

\[ 27\% \leq \pi \leq 43\% \]
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 $\neq 120$ $(H_1: \mu \neq 120)$. It may or may not be accepted and it is the hypothesis that is believed to be true by the researcher.
Level of Significance, $\alpha$

- 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, $\alpha$ and the Rejection Region
### Result Possibilities

**H₀: Innocent**

<table>
<thead>
<tr>
<th>Jury Trial</th>
<th>Hypothesis Test</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Actual Situation</strong></td>
<td><strong>Actual Situation</strong></td>
</tr>
<tr>
<td>Verdict</td>
<td>Decision</td>
</tr>
<tr>
<td>Innocent</td>
<td>Accept H₀</td>
</tr>
<tr>
<td>Guilty</td>
<td>Reject H₀</td>
</tr>
</tbody>
</table>

- **Verdict**
  - Innocent
    - Correct
    - Error
  - Guilty
    - Error
    - Correct

- **Decision**
  - H₀ True
  - H₀ False

- **Type I Error (α)**: Reject H₀, False Negative
- **Type II Error (β)**: Accept H₀, False Positive

- **Power (1 - β)**: H₀ False

- **1 - α**: H₀ True
Factors Increasing Type II Error

- True Value of Population Parameter
  - Increases When Difference Between Hypothesized Parameter & True Value Decreases
- Significance Level $\alpha$
  - Increases When $\alpha$ Decreases
- Population Standard Deviation $\sigma$
  - Increases When $\sigma$ 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 $\alpha$ $\alpha = 0.05$

Choose $n$ $n = 100$

Choose Test: $Z, t, \chi^2$ 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°C (on average)?

State the null and alternative hypotheses

\[ H_0: \mu = 37.6^\circ C \]
\[ H_a: \mu \neq 37.6^\circ C \]
Example Normal Body Temp (cont)

**Data:** random sample of \( n = 18 \) normal body temps

<table>
<thead>
<tr>
<th>Variable</th>
<th>n</th>
<th>Mean</th>
<th>SD</th>
<th>SE</th>
<th>t</th>
<th>P</th>
</tr>
</thead>
<tbody>
<tr>
<td>Temperature</td>
<td>18</td>
<td>37.22</td>
<td>0.68</td>
<td>0.161</td>
<td>2.38</td>
<td>0.029</td>
</tr>
</tbody>
</table>

Summarize data with a test statistic

\[
t = \frac{\text{sample mean} - \text{null value}}{\text{standard error}} = \frac{\bar{x} - \mu_0}{s/\sqrt{n}}
\]
<table>
<thead>
<tr>
<th>Degrees of freedom</th>
<th>Probability (p value)</th>
<th>0.10</th>
<th>0.05</th>
<th>0.01</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td></td>
<td>6.314</td>
<td>12.706</td>
<td>63.657</td>
</tr>
<tr>
<td>5</td>
<td></td>
<td>2.015</td>
<td>2.571</td>
<td>4.032</td>
</tr>
<tr>
<td>10</td>
<td></td>
<td>1.813</td>
<td>2.228</td>
<td>3.169</td>
</tr>
<tr>
<td>17</td>
<td></td>
<td>1.740</td>
<td>2.110</td>
<td>2.898</td>
</tr>
<tr>
<td>20</td>
<td></td>
<td>1.725</td>
<td>2.086</td>
<td>2.845</td>
</tr>
<tr>
<td>24</td>
<td></td>
<td>1.711</td>
<td>2.064</td>
<td>2.797</td>
</tr>
<tr>
<td>25</td>
<td></td>
<td>1.708</td>
<td>2.060</td>
<td>2.787</td>
</tr>
<tr>
<td>∞</td>
<td></td>
<td>1.645</td>
<td>1.960</td>
<td>2.576</td>
</tr>
</tbody>
</table>
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 ($p$)  
  - Test is called Z test where:
  - $Z$ is computed value
  - $\pi$ is proportion in population (null hypothesis value)

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

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

Critical Values: 1.96 at $\alpha=0.05$
2.58 at $\alpha=0.01$
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 $\alpha = 0.05$ significance level.
Solution

**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

- $\mu_1$: mean BP at baseline (no OC)
- $\mu_2$: 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,
  \[ T = \frac{\bar{d} - 0}{SE(\bar{d})} = \frac{\bar{d}}{s_d / \sqrt{n}} = \frac{\bar{d}}{4.566 / \sqrt{10}} = t_{df=9} \]

- **p-value**: the chance of observing something as extreme or more extreme than what we observed,
  \[ t = \frac{4.8}{1.444} = 3.32. \]

  **p-value**
  \[ = 2P(T < -3.32) = 2 \times 0.0045 = 0.009 < \alpha = 0.05 \]

  Significant!
  Reject $H_0$. 
CI: blood pressure

- Approximately,
  \[
  \frac{\bar{d} - \mu_d}{s_d / \sqrt{n}} \sim t_{\text{DF}}, \text{ where } \mu_d = \mu_1 - \mu_2
  \]

- 95% confidence interval for \( \mu_d \), \( t_{\text{DF}=9,0.975} = 2.262 \)
  \[
  \bar{d} \pm t_{\text{DF}=9,1-\alpha/2} \frac{s_d}{\sqrt{n}} = 4.8 \pm 2.262(1.444)
  \]
  \[
  = 4.8 \pm 3.27 = (1.53, 8.07)
  \]
Two-sample t-test: birthweight

- $\mu_1$: mean baby weight of treatment group
- $\mu_2$: 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 $s_2$

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

- $t=2.41$, $df=28$,
- $p\text{-value}=0.011 < 0.05$,
- Reject $H_0$ at 5% level
Cl: birthweight

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

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

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

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

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

$$= 0.82 \pm 2.04841 \times 0.3397766 = (0.0124, 0.516)$$
Two-sample proportion Z-test

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

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

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

<table>
<thead>
<tr>
<th>BC</th>
<th>Age</th>
<th>Yes (case)</th>
<th>No (control)</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>≥30</td>
<td>683 (p^1 = 21.2%)</td>
<td>1498 (p^2 = 14.6%)</td>
<td>2181</td>
</tr>
<tr>
<td></td>
<td>≤29</td>
<td>2537</td>
<td>8747</td>
<td>11284</td>
</tr>
<tr>
<td></td>
<td>Total</td>
<td>n_1 = 3220</td>
<td>n_2 = 10245</td>
<td>13465</td>
</tr>
</tbody>
</table>

- \( z = 8.9 \)
- \( p\text{-value}=2P(Z>8.9)=2*[1-\Phi(8.9)] \approx 0 < 0.01 \)
- 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 ($\text{FEV}_1/\text{FVC} \leq 74\%$)
  - Group B ($75\% \leq \text{FEV}_1/\text{FVC} \leq 84\%$)
  - Group C ($\text{FEV}_1/\text{FVC} \geq 85\%$)
- Question: is there significant difference in mean bronchial reactivity among the 3 groups
Plot of data

Individual Value Plot of Bronchial Reactivity vs Group
One-way ANOVA

- $y = \mu + \alpha_i + \epsilon_{ij}$
  - $\mu$: overall effect, underlying mean of all groups taken together
  - $\alpha_i$: $i=1,2,3$, group effect, difference between the mean of $i$th group and the overall mean
  - $\epsilon_{ij}$: 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

<table>
<thead>
<tr>
<th>Source</th>
<th>DF</th>
<th>SS</th>
<th>MS</th>
<th>F</th>
<th>P</th>
</tr>
</thead>
<tbody>
<tr>
<td>Group</td>
<td>2</td>
<td>503.5</td>
<td>251.8</td>
<td>4.99</td>
<td>0.018</td>
</tr>
<tr>
<td>Error</td>
<td>19</td>
<td>958.8</td>
<td>50.5</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Total</td>
<td>21</td>
<td>1462.4</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

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

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

- Strong Positive
- Strong Negative
- Weak Positive
- Moderate Negative
- None
- Weak Negative
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**

<table>
<thead>
<tr>
<th>Predictor</th>
<th>Coef</th>
<th>SE Coef</th>
<th>T</th>
<th>P</th>
</tr>
</thead>
<tbody>
<tr>
<td>Constant</td>
<td>21.523</td>
<td>2.620</td>
<td>8.21</td>
<td>0.000</td>
</tr>
<tr>
<td>Estriol</td>
<td>0.6082</td>
<td>0.1468</td>
<td>4.14</td>
<td>0.000</td>
</tr>
</tbody>
</table>

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

Analysis of Variance

<table>
<thead>
<tr>
<th>Source</th>
<th>DF</th>
<th>SS</th>
<th>MS</th>
<th>F</th>
<th>P</th>
</tr>
</thead>
<tbody>
<tr>
<td>Regression</td>
<td>1</td>
<td>250.57</td>
<td>250.57</td>
<td>17.16</td>
<td>0.000</td>
</tr>
<tr>
<td>Residual Error</td>
<td>29</td>
<td>423.43</td>
<td>14.60</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Total</td>
<td>30</td>
<td>674.00</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Breast cancer example

<table>
<thead>
<tr>
<th></th>
<th>age $\geq 30$</th>
<th>age $\leq 29$</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>case</td>
<td>$n_{11}=683$</td>
<td>$n_{12}=2537$</td>
<td>$n_{1}=3220$</td>
</tr>
<tr>
<td>control</td>
<td>$n_{21}=1498$</td>
<td>$n_{22}=8747$</td>
<td>$n_{2}=10245$</td>
</tr>
<tr>
<td>Total</td>
<td>$n_{1}=2181$</td>
<td>$n_{2}=11284$</td>
<td>$n_{..}=13465$</td>
</tr>
</tbody>
</table>

- Odds ratio = \( \frac{n_{11}}{n_{12}} \div \frac{n_{21}}{n_{22}} = \frac{n_{11}n_{22}}{n_{12}n_{21}} = 1.57 \)

- 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 
- $df = (nrow-1) \times (ncol-1) = 1$ 
- $p$-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 $\Omega$. An *event* $A$ is a (set of) possible outcomes of the experiment, and corresponds to a subset of $\Omega$.

- A probability law/measure is a function $P(A)$ 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 $P(A)$ must satisfy the following:
  
  1. $0 \leq P(A_i) \leq 1$
  2. $P(\Omega) = \sum P(A_i) = 1$
  3. If $A_i \cap A_j = \phi$, then $P(A_i \cup A_j) = P(A_i) + P(A_j)$
     
     otherwise $P(A_i \cup A_j) = P(A_i) + P(A_j) - P(A_i \cap A_j)$
Important Theorems of Probability

- \( P(A^C) = 1 - P(A) \)
- \( A_i \subset A_j \Rightarrow P(A_i) < P(A_j) \)
- \( P(\emptyset) = 0 \)
- \( P(A) \leq 1 \)

- If \( \{A_i \cap A_j = \emptyset, \forall i, j\} \Rightarrow P\left( \bigcup_{k=1}^{N} A_k \right) = \sum_{k=1}^{N} P(A_k) \)
A random variable $X$ is a function that maps every possible event in the space $\Omega$ of a random experiment to a real number.

For the fish example, if $X=\textit{the next fish selected from the pile}$, we have $X(\text{Salmon})=1$, $X(\text{Trout})=2$, $X(\text{Sword})=3$, $X(\text{Catfish})=4$ and $X(\text{Seabass})=5$.

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 be 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.
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

\[
P(X = x) > 0, \quad \sum_{x \in X} P(x) = 1
\]

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

\[
F(x) = P(X \leq x) = \sum_{u \leq x} P(u)
\]
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 $f(x)$.

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

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

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

CDF

Weight of NJ residents

PDF
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, \( P(\text{weight} = 175.24654163546546874876454687987652354354385424148454257654) = 0 \)

The area under the pdf curve between any two abscissas gives the probability that the value lie between those two points:
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

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

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

The standard deviation is simply the positive square root of the variance. The $k^{th}$ moment in general is obtained as

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

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.
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

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

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

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

$$\sigma_y^2 = E[(y - \mu_y)^2] = \sum_x \sum_y (y - \mu_y)^2 \cdot P(x, y)$$
A cross-moment can also be defined as the covariance

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

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 \( \Sigma \)

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

If \( X \) and \( Y \) are statistically independent \( \Rightarrow \sigma_{xy} = 0 \)

If \( \sigma_{xy} = 0 \) \( \Rightarrow \) then the variables are said to be **uncorrelated**

Note that statistical independence is a stronger property than correlation:

statistical independence \( \nRightarrow \) uncorrelated
If \( \rho = 1 \), then the variables are identical, they move together,

If \( \rho = -1 \), then the variables are negatively correlated, one decreases as the other increases at the same rate,

If \( \rho = 0 \) \( \Rightarrow \) 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

\[
\Sigma_{xy} = \begin{bmatrix}
1.073 & -0.026 \\
-0.0264 & 0.9673
\end{bmatrix} \quad \Sigma_{xz} = \begin{bmatrix}
1.073 & 2.1476 \\
2.1476 & 4.2951
\end{bmatrix}
\]

\[
\rho_{xy} = -0.0259
\]
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

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

$$
P(x, y) = P(x | y)P(y)
$$

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 $X \cap Y$, that is $P(y)$ renormalizes the probability of events that occur jointly with $Y$
Let $B_1, \ldots, B_N$ be $N$ mutually exclusive events, whose union gives the sample space $\Omega$. Hence the events $B$ constitute a partition of $\Omega$.

Now consider an event $A$, a subset of $\Omega$. 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

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

$$P(A) = P(A | B_1)P(B_1) + \ldots + P(A | B_N)P(B_N) = \sum_{k=1}^{N} P(A | B_k)P(B_k)$$
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) \( P(A'|B') = P(B'|A')P(A')/P(B') = P(B'|A')P(A')/(1-P(B)) = 0.94*0.99/(1-.0691)=0.9997 \)
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

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

- 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:

\[ r = \frac{x-\mu_x}{\sigma_x} \]
In $d$-dimensional space, the Gaussian pdf is:

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

$$p(x) \sim N(\mu, \Sigma)$$

Gaussian distributions are very popular since:

- The parameters $(\mu, \Sigma)$ are **sufficient** to uniquely characterize the normal distribution.
- If the $x_i$'s are mutually **uncorrelated** ($\sigma_{ik}=0$), 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 $N \times N$ invertible matrix, then $Y=AX$ is also jointly Gaussian:
    $$p_Y(y) = p_X(A^{-1}y) / |A|$$
Multivariate Normal Density

• x is a vector of d Gaussian variables

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

\[ \mu = E[x] = \int x p(x) dx \]

\[ \Sigma = E[(x-\mu)(x-\mu)^T] = \int (x-\mu)(x-\mu)^T p(x) dx \]

• Mahalanobis Distance

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

• All conditionals and marginals are also Gaussian
Multivariate 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 \textit{a priori} chance of seeing a class

\[ P(\text{Class2}) \quad \& \quad P(\text{Class1}) \]
But How Do We Classify?

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

  \[ P(\text{Class1}), P(\text{Class2}) \]

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

  \[ P(\text{X}|\text{Class1}), P(\text{X}|\text{Class2}) \]

*But we want the probability of the class given a feature*  
*How do we get* \( P(\text{Class1}|X) \)?
Bayes Rule

\[
P(\text{Class} \mid \text{Feature}) = \frac{P(\text{Feature} \mid \text{Class})P(\text{Class})}{P(\text{Feature})}
\]

Belief before evidence

Evaluate evidence

Likelihood

Prior

Belief after evidence

Evidence

Bayes Decision Rule

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

The **Bayes Decision Rule** is simply choose Class 1 if:

$$P(\text{Class 1} \mid X) > P(\text{Class 2} \mid X)$$

$$\frac{P(X \mid \text{Class 1})P(L_1)}{P(X)} > \frac{P(X \mid \text{Class 2})P(L_2)}{P(X)}$$

This is the same number on both sides!
Discriminant Function

We can create a convenient representation of the Bayes Decision Rule

\[ P(X \mid \text{Class1})P(\text{Class1}) > P(X \mid \text{Class2})P(\text{Class2}) \]

\[ \frac{P(X \mid \text{Class1})P(\text{Class1})}{P(X \mid \text{Class2})P(\text{Class2})} > 1 \]

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

*If* \( G(X) > 0 \), *we classify as* \( \text{Class 1} \)
Stepping back

What do we have so far?

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

\[ P(X|\text{Class1}), P(X|\text{Class2}) \quad \text{and} \quad P(\text{Class1}), P(\text{Class2}) \]

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

\[ G(X) = \log \frac{P(X|\text{Class1}) P(\text{Class1})}{P(X|\text{Class2}) P(\text{Class2})} > 0 \]

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

…and if \( G(X) > 0 \) we classify as Class1
Getting $P(X|\text{Class})$ from Training Set

**One Simple Approach**

Divide $X$ values into bins

And then we simply count frequencies

There are 13 data points
Class conditional from Univariate Normal Distribution

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

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

\[ P(X|\text{Class1}), P(X|\text{Class2}) \quad P(\text{Class1}), P(\text{Class2}) \]

And we are ready to plug these into our discriminant function

\[ G(X) = \log \frac{P(X \mid \text{Class1})}{P(X \mid \text{Class2})} \cdot \frac{P(\text{Class1})}{P(\text{Class2})} > 0 \]

But there is one more little complication…..
Multidimensional feature space?

So \( P(X|\text{Class}) \) become \( P(X_1,X_2,X_3,\ldots,X_8|\text{Class}) \)
and our discriminant function becomes

\[
G(X) = \log \frac{P(X_1, X_2, \ldots, X_7 | \text{Class1})}{P(X_1, X_2, \ldots, X_7 | \text{Class2})} \frac{P(\text{Class1})}{P(\text{Class2})} > 0
\]

In \( d \)-dimensional space, the Gaussian pdf is:

\[
p(x) = \frac{1}{(2\pi)^{d/2} |\Sigma|^{1/2}} e^{-\frac{1}{2}[(x-\mu)^T \Sigma^{-1} (x-\mu)]} \\
p(x) \sim N(\mu, \Sigma)
\]
Naïve Bayes Classifier

We are going to make the following assumption:

*All features are independent given the class*

\[
P(X_1, X_2, ..., X_n \mid \text{Class}) = P(X_1 \mid \text{Class})P(X_2 \mid \text{Class})...P(X_n \mid \text{Class})
\]

\[
= \prod_{i=1}^{n} P(X_i \mid \text{Class})
\]

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:

\[ G(X_1, \ldots, X_7) = \log \frac{P(X_1, X_2, \ldots, X_7 \mid \text{Class1})}{P(X_1, X_2, \ldots, X_7 \mid \text{Class2})} \frac{P(\text{Class1})}{P(\text{Class2})} > 0 \]

As this:

\[ G(X_1, \ldots, X_7) = \log \prod_{i} \frac{P(X_i \mid \text{Class1})}{P(X_i \mid \text{Class2})} \frac{P(\text{Class1})}{P(\text{Class2})} > 0 \]
Classifying Parasitic RBC

Plug these and priors into the discriminant function

\[ G(X_1, ..., X_7) = \log \frac{\prod P(X_i | \text{Mito})}{P(\text{Mito})} \frac{P(\text{Mito})}{\prod P(X_i | \sim \text{Mito})} > 0 \]

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

<table>
<thead>
<tr>
<th></th>
<th>True (Mito)</th>
<th>False (~Mito)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Predicted True</td>
<td>TP</td>
<td>FP</td>
</tr>
<tr>
<td>Predicted False</td>
<td>FN</td>
<td>TN</td>
</tr>
</tbody>
</table>

Sensitivity = \( \frac{TP}{TP+FN} \)  
Specificity = \( \frac{TN}{TN+FP} \)

- **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