

*COMP 1433: Introduction to Data Analytics &
COMP 1003: Statistical Tools and Applications*

Lecture 3 – Statistics Basics for Data Analytics

Dr. Jing Li

Department of Computing

The Hong Kong Polytechnic University

25&27 Jan 2022

Discussion: Tears of Sally

- *Sally Clark* was the victim of a miscarriage of justice in 1999.
- She was found guilty of the murder of two infant sons, both died within the first few weeks of their births.
- Here are the arguments from the defense:
 - Only the mother is present when the babies die.
 - 1 in 8,500 babies died from sudden infant death syndrome (SIDS)
 - The probability that two babies both died from SIDS is 1 in 73 million (the square of 1/8500)
 - The probability that the mother kills the sons is $1 - \frac{1}{73M} \approx 1$
 - The mother is the murderer!
- *QUESTION: How can you fight back?*



Roadmap

- Expectation and Variance
- Sample Statistics
- Hypothesis Testing
- Statistics vs. Data Analytics
 - Naïve Bayes (cont.)

Roadmap

- Expectation and Variance
- Sample Statistics
- Hypothesis Testing
- Statistics vs. Data Analytics
 - Naïve Bayes (cont.)

Expectation of Random Variables

- The *expected value* for discrete random variable X
 - $E[X] \equiv \sum_k x_k P(X = x_k) = \sum_k x_k p_X(x_k)$
 - It represents the *weighted average sum* of X
 - Also called the *mean* of X
- **Example.** The *expected value* of *rolling a die* is:
 - $E(X) = 1 \cdot \frac{1}{6} + 2 \cdot \frac{1}{6} + \dots + 6 \cdot \frac{1}{6} = \frac{1}{6} \cdot \sum_{k=1}^6 k = \frac{7}{2}$
- **PROPERTIES**
 - $E[aX] = aE[X]$
 - $E[aX + b] = aE[X] + b$



Probability mass function $p_{X,Y}(x,y)$

	$y = 6$	$y = 8$	$y = 10$	$p_X(x)$
$x = 1$	$\frac{1}{5}$	0	$\frac{1}{5}$	$\frac{2}{5}$
$x = 2$	0	$\frac{1}{5}$	0	$\frac{1}{5}$
$x = 3$	$\frac{1}{5}$	0	$\frac{1}{5}$	$\frac{2}{5}$
$p_Y(y)$	$\frac{2}{5}$	$\frac{1}{5}$	$\frac{2}{5}$	1

What is $E[X]$, $E[Y]$, and $E[XY]$?

Are X and Y independent?

Variance and Standard Deviation

- Let the mean of X be $\mu = E[X]$
- Then the *variance* of X is:
 - $Var(X) \equiv E[(X - \mu)^2] = \sum_k (x_k - \mu)^2 p_X(x_k)$
 - The weighted *square distance* from the *mean*.
- We have another form of *variance* as:
 - $Var(X) = E[X^2] - \mu^2$
- The *standard deviation* is $\sigma(X) = \sqrt{Var(x)}$
 - The weighted *distance* from the *mean*.
- The *variance* of *rolling the die* is:
 - $Var(X) = \sum_{k=1}^6 [k^2 \cdot \frac{1}{6}] - \mu^2 = \frac{1}{6} \cdot \frac{6(6+1)(2 \cdot 6 + 1)}{6} - \left(\frac{7}{2}\right)^2 = \frac{35}{12}$
 - The *standard deviation* $\sigma = \sqrt{\frac{35}{12}} \cong 1.70$

Roadmap

- Expectation and Variance
- Sample Statistics
- Hypothesis Testing
- Statistics vs. Data Analytics
 - Naïve Bayes (cont.)

What is *Sampling*?

- *Sampling* can consist:
 - *Gathering random data* from a *large* population, e.g.,
 - Measuring the *height of randomly selected adults*.
 - Measuring the *starting salary of random CS students*.
 - *Recording the results of experiments*, e.g.,
 - Measuring the *breaking strengths* of randomly selected bolts
 - Measuring the *lifetime* of randomly selected bulbs.
- **Assumptions:**
 - The population is *infinite* (or very *large*).
 - The observations are *independent*.
 - The experiment outcome does not affect other experiments.

What is *Sample Statistics*?

- A random sample from a population consists of:
 - *Independent, identically distributed* random variables,
 X_1, X_2, \dots, X_n
- The values of X_i are called the *outcomes* of the experiment.
- A *statistic* is a *function* of X_1, X_2, \dots, X_n .
- Thus a *statistic* itself is a *random variable*.

Important Statistics

- The *sample mean*:

- $\bar{X} \equiv \frac{1}{n}(X_1 + X_2 + \dots + X_n)$

- The sample variance and standard deviation:

- $S^2 \equiv \frac{1}{n-1} \sum_{k=1}^n (X_k - \bar{X})^2$ and $S = \sqrt{\frac{1}{n-1} \sum_{k=1}^n (X_k - \bar{X})^2}$

- The *order statistic* in which the observations are *ordered in size*, e.g., *increasing order*

- The *sample median* is:

- The *midvalue* of the *order statistic* (if n is an odd)
- The *average* of the *two middle values* (if n is an even)

- The *sample range* is the *difference between the largest and smallest observations*.

Example: Important Statistics

- For the 8 observations:
 - $-0.737, 0.511, -0.083, 0.066, -0.562, -0.906, 0.358, 0.359$
- What is the *sample mean*:
 - $\bar{X} = \frac{1}{8} (-0.737 + 0.511 - 0.083 + 0.066 - 0.562 - 0.906 + 0.358 + 0.359) = -0.124$
- What is the *sample variance*:
 - $S^2 = \frac{1}{8-1} [(-0.737 + 0.124)^2 + (0.511 + 0.124)^2 + (-0.083 + 0.124)^2 + (0.066 + 0.124)^2 + (-0.562 + 0.124)^2 + (-0.906 + 0.124)^2 + (0.358 + 0.124)^2 + (0.359 + 0.124)^2] = 0.297$
- What is the *sample standard deviation*:
 - $S = \sqrt{0.297} = 0.545$

Example: Important Statistics

- For the 8 observations:
 - $-0.737, 0.511, -0.083, 0.066, -0.562, -0.906, 0.358, 0.359$
- What is the *order statistics*?
 - $-0.906, -0.737, -0.562, -0.083, 0.066, 0.358, 0.359, 0.511.$
- What is the *sample median*?
 - $\frac{-0.083+0.066}{2} = -0.0085$
- What is the *sample range*?
 - $0.511 - (-0.906) = 1.417$

Discussion:

The Trap of Mean



- **NEWS TITLE:** The *average income* of Company T's employees is *100K HKD/month*
 - Is the news title *misleading*?
 - Considering the following example:
 - *10% of the employees earn 910K*
 - *90% of the employees earn 10K*
 - What is the *mean* of their monthly income?
 - Is it able to reflect the overall situation?
 - What is a better alternative?

Sample Mean vs. Population Mean

- Suppose the population mean and standard deviation are μ and σ .
- The sample mean is $\bar{X} = \frac{1}{n}(X_1 + X_2 + \dots + X_n)$
 - It is also a *random variable*:
 - Expected value
 - $\mu_{\bar{X}} \equiv E[\bar{X}] = E\left[\frac{1}{n}(X_1 + X_2 + \dots + X_n)\right] = \mu$
 - Variance
 - $\sigma_{\bar{X}}^2 \equiv Var(\bar{X}) = \frac{\sigma^2}{n} \rightarrow n \rightarrow +\infty$ and $\sigma_{\bar{X}}^2 \rightarrow 0$
- So, the *expected value* of *sample mean* is the *population mean* μ .

Markov Inequality and Chebyshev's Inequality

- **Markov Inequality.** For discrete random variable $X \geq 0$ and $\epsilon > 0$:
 - $P(X \geq \epsilon) \leq \frac{E[X]}{\epsilon}$
 - $E[X] = \sum_{x \geq 0} xp(x) \geq \sum_{x \geq \epsilon} xp(x) \geq \epsilon \sum_{x \geq \epsilon} p(x) = \epsilon P(X \geq \epsilon)$
 - It also holds for continuous random variables!
- **Chebyshev's Inequality.** For any random variables, we have:
 - $P(|X - \mu| \geq \epsilon) \leq \frac{\sigma^2}{\epsilon^2}$
 - where $\mu = E[X]$ and $\sigma = \sqrt{Var(X)}$
 - $P(|X - \mu| \geq \epsilon) = P((X - \mu)^2 \geq \epsilon^2) \leq \frac{\sigma^2}{\epsilon^2}$

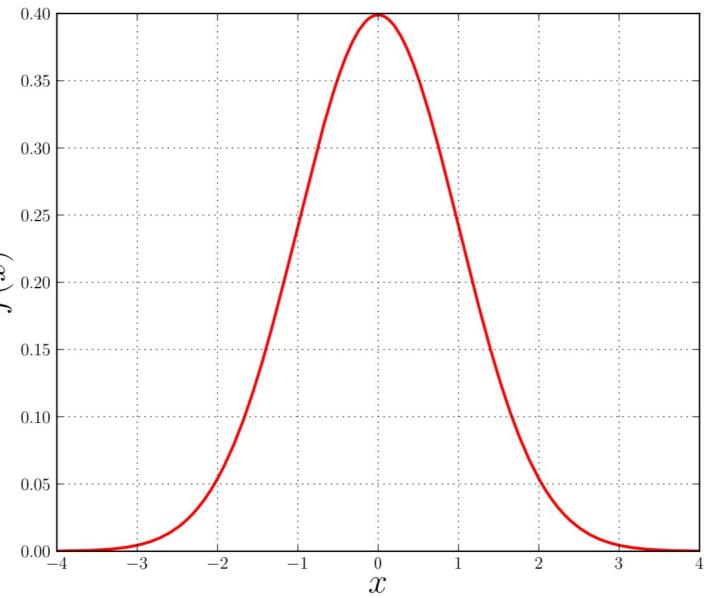
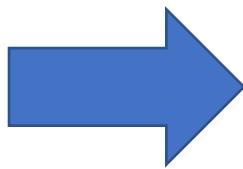
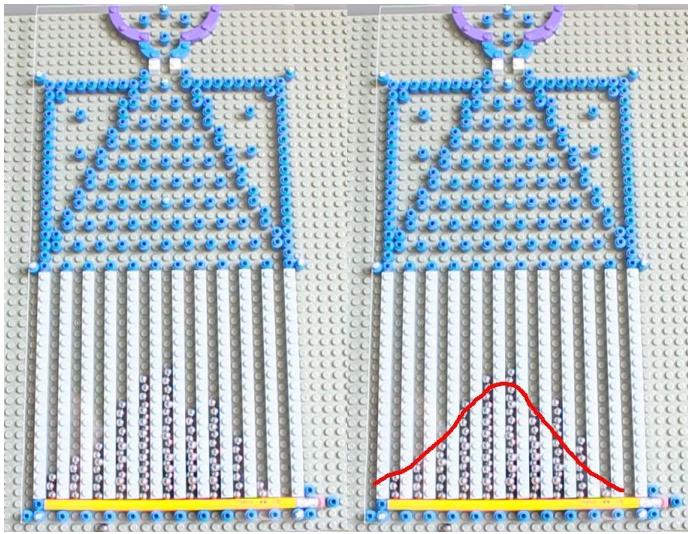
Law of Large Number

- *Chebyshev's Inequality*. For any random variables:
 - $P(|X - \mu| \geq \epsilon) \leq \frac{\sigma^2}{\epsilon^2}$ ($\mu = E[X]$ and $\sigma = \sqrt{Var(X)}$)
- Then $P(|\bar{X} - \mu| \leq \epsilon) \geq 1 - \frac{\sigma_{\bar{X}}^2}{\epsilon^2} = 1 - \frac{\sigma^2}{n\epsilon^2}$
- For $n \rightarrow +\infty$, $P(|\bar{X} - \mu| \leq \epsilon) = 1$ for any $\epsilon > 0$
- The *sample mean* approximates the *population mean* μ for very large n .

Central Limit Theorem

- Suppose the population mean and standard deviation are μ and σ .
- The sample mean is $\bar{X} = \frac{1}{n}(X_1 + X_2 + \dots + X_n)$
 - $\mu_{\bar{X}} \equiv E[\bar{X}] = E\left[\frac{1}{n}(X_1 + X_2 + \dots + X_n)\right] = \mu$
 - $\sigma_{\bar{X}}^2 \equiv Var(\bar{X}) = \frac{\sigma^2}{n}$
- **NOTE.** \bar{X} is approximately *general normal* (or satisfies *normal distribution*) for very large n .
 - The proof requires advanced knowledge in Calculus.
 - This explains why normal distribution is so important!

General Normal



Galton Knocked Boards

$$\frac{P(x \leq X \leq x + \Delta x)}{\Delta x}$$

where Δx is very small

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

General and Standard Normal

- For *continuous random variable* X , if for all $x \in R$

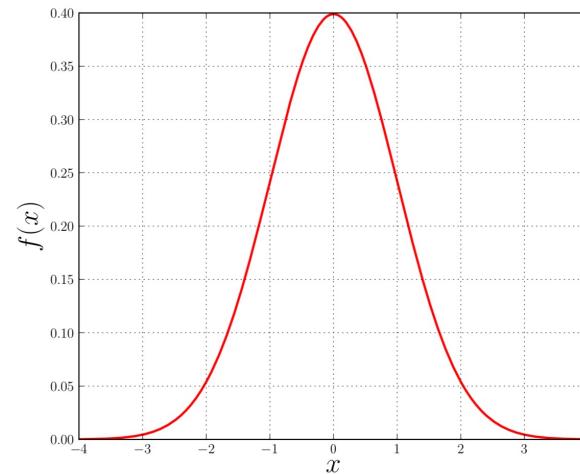
- We have $\frac{P(x \leq X \leq x + \Delta x)}{\Delta x} \equiv f(x) = \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{1}{2\sigma^2}(x-\mu)^2}$
- where Δx is a very small value

Probability
Density
Function

- We say X is *general normal* or $X \sim N(\mu, \sigma^2)$

PROPERTIES.

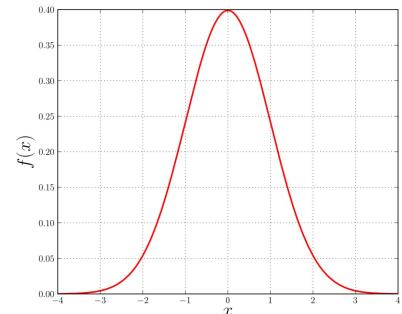
- $E[X] = \mu$ and $Var(X) = \sigma^2$.
- When $\mu = 0$ and $\sigma = 1$
 - $f(x) = \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}x^2}$
 - X is *standard normal* or $X \sim N(0, 1)$



General and Standard Normal

- For *continuous random variable* X , if for all $x \in R$
 - We have $\frac{P(x \leq X \leq x + \Delta x)}{\Delta x} = f(x) = \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{1}{2\sigma^2}(x-\mu)^2}$
 - where Δx is a very small value
- We say X is *general normal* or $X \sim N(\mu, \sigma^2)$
- **PROPERTIES.**
 - $\frac{X-\mu}{\sigma} \sim N(0,1)$
 - **BECAUSE:**

$$\bullet \quad f\left(\frac{x-\mu}{\sigma}\right) = \frac{P\left(x \leq \frac{X-\mu}{\sigma} \leq x + \Delta x\right)}{\Delta x} = \frac{P(x\sigma + \mu \leq X \leq (x + \Delta x)\sigma + \mu)}{\Delta x \sigma} \sigma = \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}x^2}$$



Central Limit Theorem

- Suppose the population mean and standard deviation are μ and σ .
- The sample mean is $\bar{X} = \frac{1}{n}(X_1 + X_2 + \dots + X_n)$
 - $\mu_{\bar{X}} \equiv E[\bar{X}] = E\left[\frac{1}{n}(X_1 + X_2 + \dots + X_n)\right] = \mu$
 - $\sigma_{\bar{X}}^2 \equiv Var(\bar{X}) = \frac{\sigma^2}{n}$
- **NOTE.** \bar{X} is approximately *general normal* (or satisfies *normal distribution*) for very large n .
 - $\frac{\bar{X}-\mu_{\bar{X}}}{\sigma_{\bar{X}}} = \frac{\bar{X}-\mu}{\frac{\sigma}{\sqrt{n}}}$ is approximately *standard normal* for very large n .

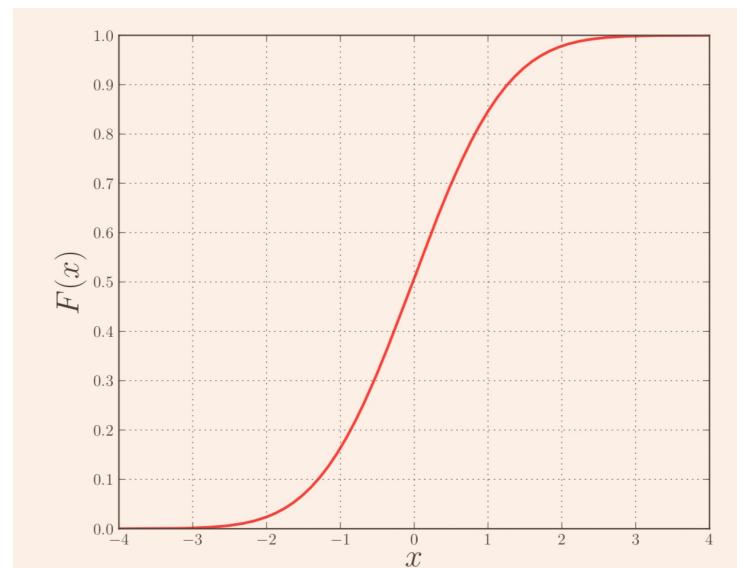
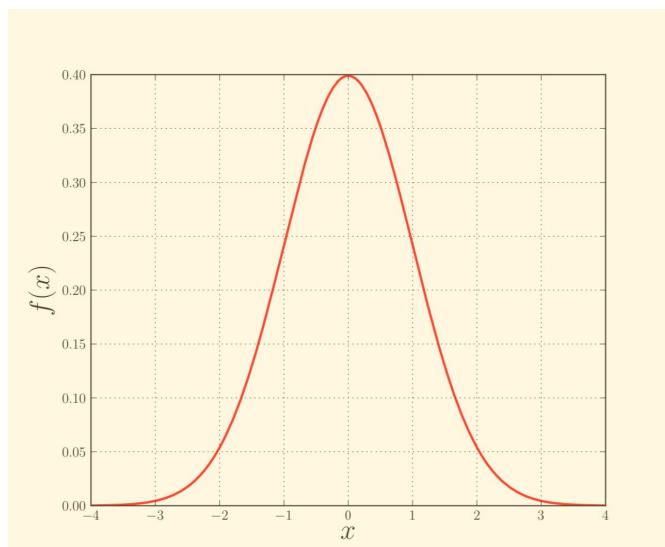
Standard Normal

- For *continuous random variable* X , if for all $x \in R$

- $f(x) = \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}x^2}$, say $X \sim N(0,1)$

- The *cumulative distribution function*

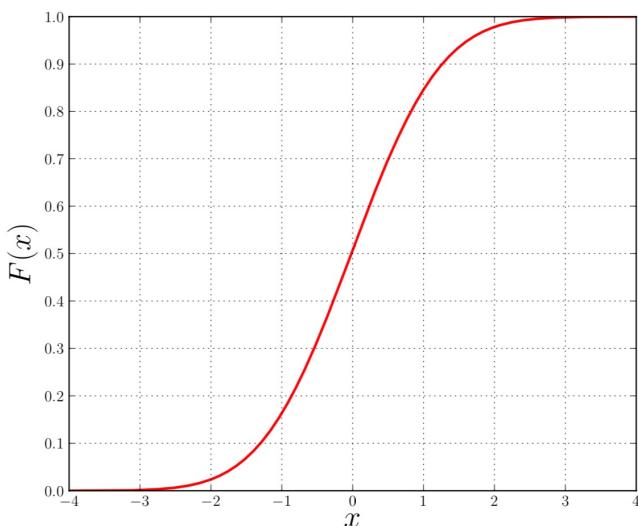
- $F(x) \equiv P(X \leq x) \equiv \Phi(x)$



Standard Normal

- For *continuous random variable* X , if for all $x \in R$

- $f(x) = \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}x^2}$, say $X \sim N(0,1)$
- The *cumulative distribution function*
 - $F(x) \equiv P(X \leq x) \equiv \Phi(x)$
 - $P(-x \leq X \leq x) = 1 - 2\Phi(-x)$
 $(x > 0)$



z	$\Phi(z)$	z	$\Phi(z)$
0.0	.5000	-1.2	.1151
-0.1	.4602	-1.4	.0808
-0.2	.4207	-1.6	.0548
-0.3	.3821	-1.8	.0359
-0.4	.3446	-2.0	.0228
-0.5	.3085	-2.2	.0139
-0.6	.2743	-2.4	.0082
-0.7	.2420	-2.6	.0047
-0.8	.2119	-2.8	.0026
-0.9	.1841	-3.0	.0013
-1.0	.1587	-3.2	.0007

QUESTION. How about positive z ?

Roadmap

- Expectation and Variance
- Sample Statistics
- Hypothesis Testing
- Statistics vs. Data Analytics
 - Naïve Bayes (cont.)

Confidence Interval Estimate

- $\frac{\bar{X} - \mu}{\frac{\sigma}{\sqrt{n}}}$ is approximately *standard normal* for very large n .
- $$P\left(\left|\frac{\bar{X} - \mu}{\frac{\sigma}{\sqrt{n}}}\right| \leq z\right) = P\left(|\bar{X} - \mu| \leq \frac{\sigma z}{\sqrt{n}}\right)$$
$$= P\left(\mu \in \left[\bar{X} - \frac{\sigma z}{\sqrt{n}}, \bar{X} + \frac{\sigma z}{\sqrt{n}}\right]\right)$$
$$\cong 1 - 2\Phi(-z)$$
- This gives the *confidence interval estimate* of μ .

Example: Confidence Interval Estimate

- The population *mean* is *unknown*.
- The population standard deviation is $\sigma = 3$.
- The sample size is $n = 25$.
- The sample mean is $\bar{X} = 4.5$.
- Then taking $z = 2$ we have

$$\begin{aligned} \bullet \quad P\left(\mu \in \left[\bar{X} - \frac{3 \cdot 2}{\sqrt{25}}, \bar{X} + \frac{3 \cdot 2}{\sqrt{25}}\right]\right) &= P(\mu \in [3.3, 5.7]) \\ &\cong 1 - 2\Phi(-2) \cong 95\% \end{aligned}$$

Go back to the table of $\Phi(\cdot)$ on page 24 to check!

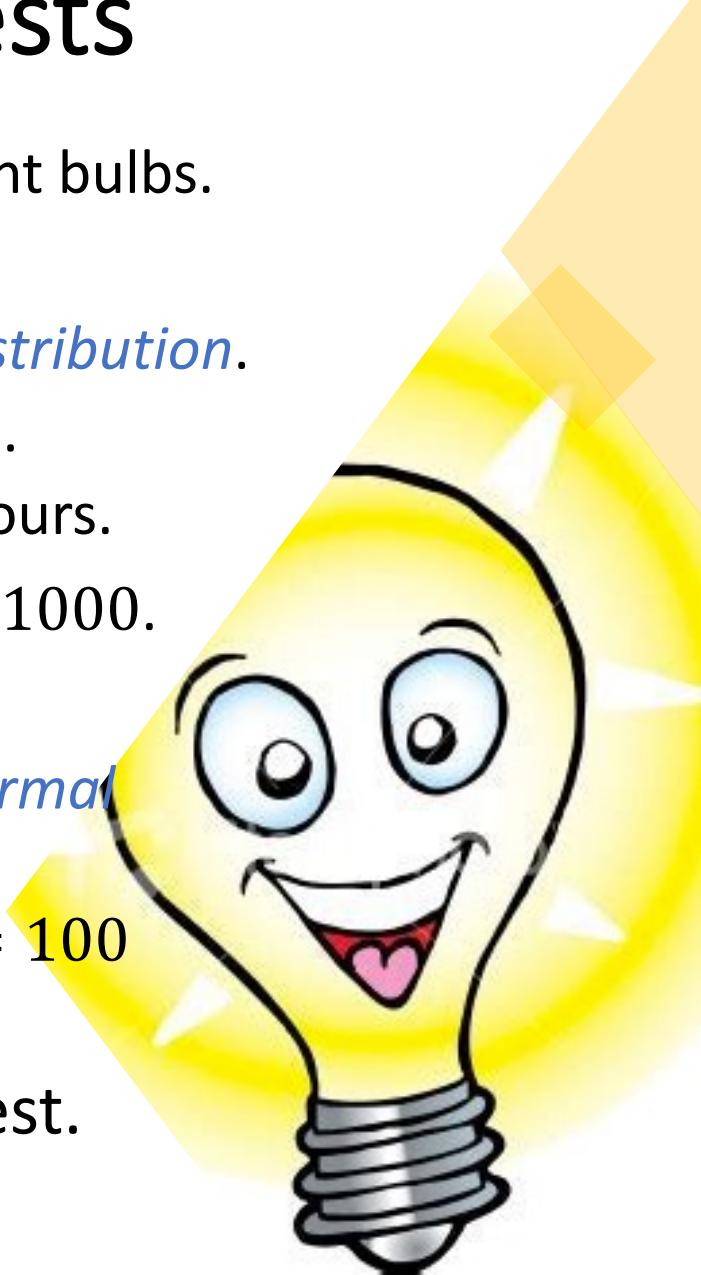
- We say $[3.3, 5.7]$ is the 95% *confidence interval estimate* of μ .

Hypothesis Testing

- How to *test* whether a *hypothesis* is *true or false*?
- First, we gather *data* (i.e., *samples*).
- Second, we hypothesize that a random variable *has a given mean*.
- Third, we decide to *accept* or *reject* the hypothesis based on the data we collect.

Example: Hypothesis Tests

- We want to order a large shipment of light bulbs.
- The manufacturer *claims* that:
 - The lifetime of bulbs has a *normal distribution*.
 - The *mean* lifetime is $\mu = 1000$ hours.
 - The *standard deviation* is $\sigma = 100$ hours.
- We want to *test the hypothesis* that $\mu = 1000$.
- We assume that:
 - The lifetime of bulbs has indeed a *normal distribution*.
 - The *standard deviation* is indeed $\sigma = 100$ hours.
- We sample 25 light bulbs for the test.

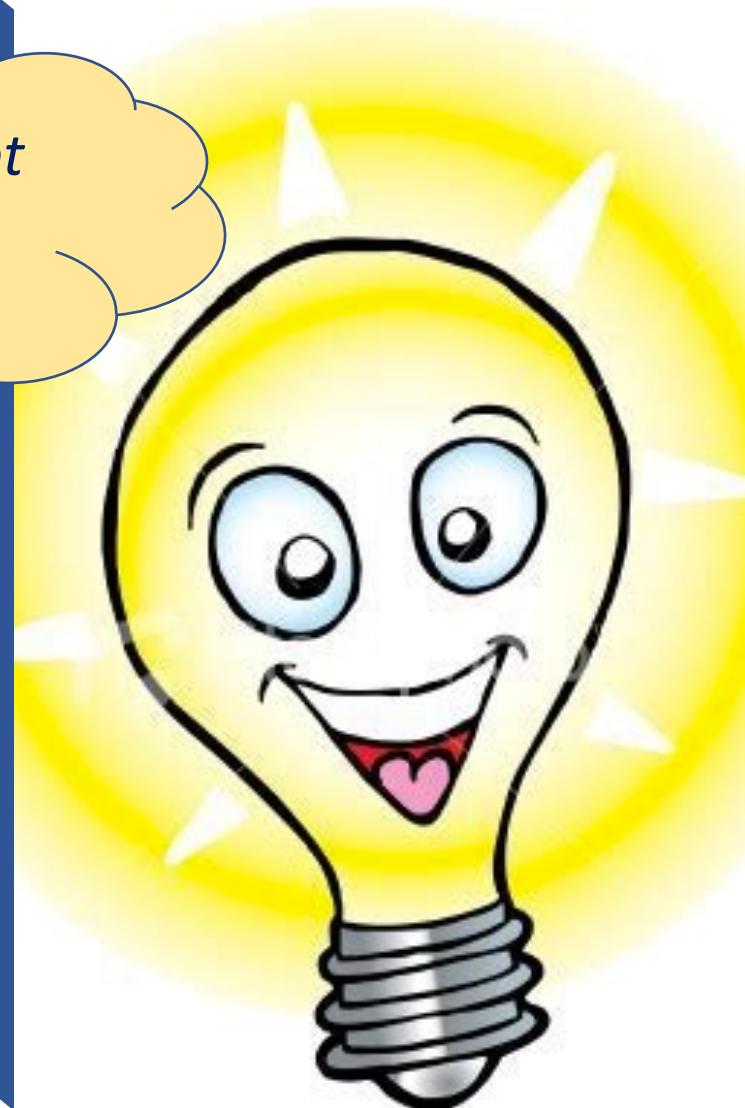


Example: Hypothesis Tests

- The manufacturer *claims* that:
 - $\mu = 1000$ hours.
 - $\sigma = 100$ hours.
- We assume that:
 - Indeed, $\sigma = 100$ hours.
- We sample 25 light bulbs for the test.
 - The *sample* average lifetime is $\bar{X} = 960$
 - Do we accept the hypothesis that $\mu = 1000$?
- We know $P(\bar{X} \leq 960) = \Phi\left(\frac{960-1000}{\frac{100}{\sqrt{25}}}\right) = \Phi(-2.0) = 2.28\%$

One-sided probability

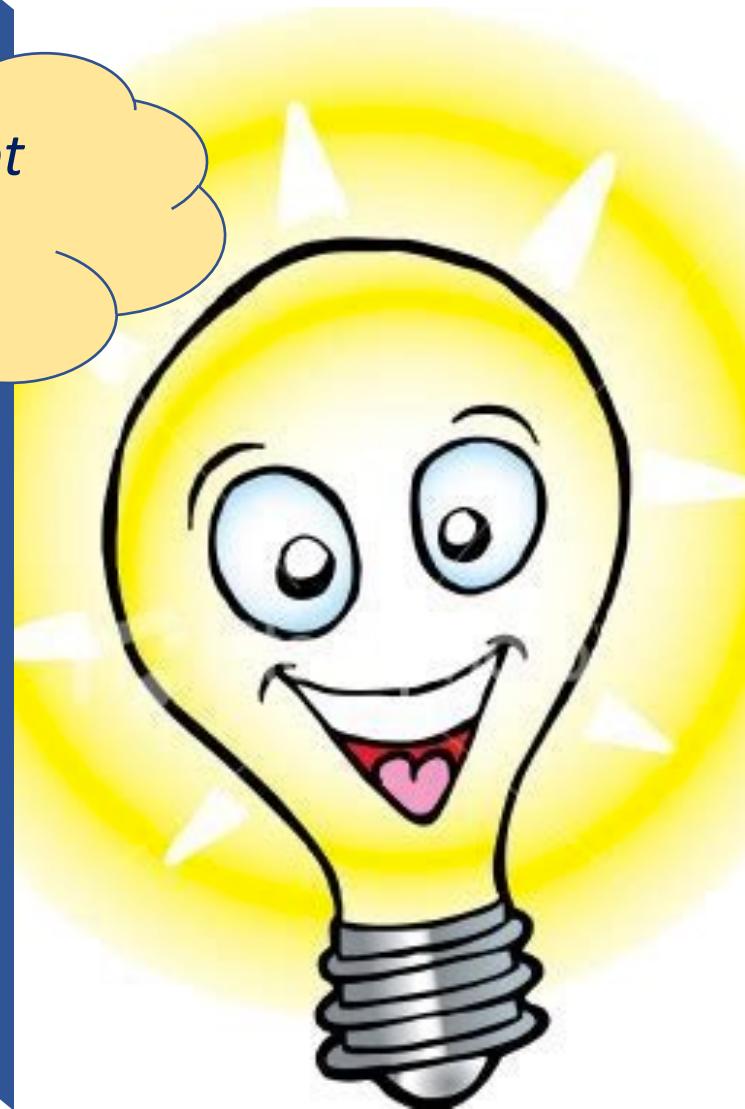
*Would you accept
and pay for the
shipment!*



Example: Hypothesis Tests

- The manufacturer *claims* that:
 - $\mu = 1000$ hours.
 - $\sigma = 100$ hours.
- We assume that:
 - Indeed, $\sigma = 100$ hours.
- We sample 25 light bulbs for the test.
 - The *sample* average lifetime is $\bar{X} = 1040$
 - Do we accept the hypothesis that $\mu = 1000$?
- We know $P(\bar{X} \geq 1040) = \Phi\left(\frac{1040 - 1000}{\frac{100}{\sqrt{25}}}\right)$
One-sided probability $= \Phi(-2.0) = 2.28\%$

*Would you accept
and pay for the
shipment!*

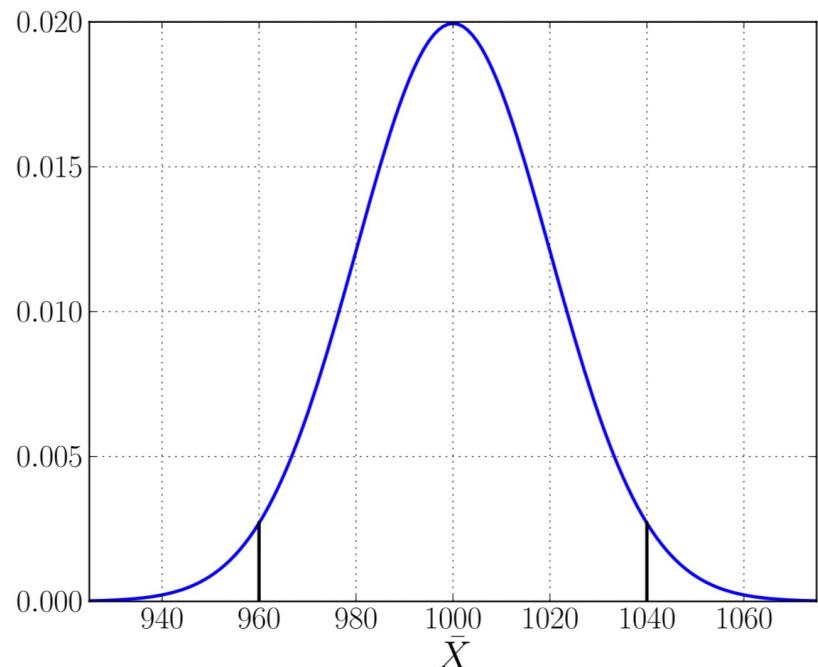


- The manufacturer *claims* that:
 - $\mu = 1000$ hours.
 - $\sigma = 100$ hours.
- We assume that:
 - Indeed, $\sigma = 100$ hours.
 - We *accept* that $\mu = 1000$ if

$$960 \leq \bar{X} \leq 1040$$

- We sample 25 light bulbs for the test
- The probability that we *accept*:
 - $P(|\bar{X} - 1000| \leq 40) = 1 - 2\Phi\left(\frac{960-1000}{100/\sqrt{25}}\right)$
 $= 1 - 2\Phi(-2.0) \cong 95\%$
- And we *reject* with 5% probability

*Example
(cont.)*



Another Example

- Given a sample size 9 from a normal population with $\sigma = 0.2$, has sample mean $\bar{X} = 4.88$
- Claim: The population mean is $\mu = 5.00$ (**Null Hypothesis** H_0)
- Since $Z \equiv \frac{\bar{X} - \mu}{\sigma/\sqrt{n}}$ is standard normal, the **p-value** is
 - $P(|\bar{X} - \mu| \geq 0.12) = P\left(|Z| \geq \frac{0.12}{\frac{0.2}{\sqrt{9}}}\right) = 2\Phi(-1.8) \cong 7.18\%$
 - We *reject* the hypothesis if $P(|\bar{X} - \mu| \geq 0.12)$ is *rather small*, say if $P(|\bar{X} - \mu| \geq 0.12) < 10\%$ (*level of significance* 10%).
 - We *reject* the hypothesis if $P(|\bar{X} - \mu| \geq 0.12) < 5\%$ (*level of significance* 5%) --- *We are more tolerant* .

Another Example

- Given a sample size 64 from a normal population with $\sigma = 0.234$, has sample mean $\bar{X} = 4.847$
- We will test the *Null Hypothesis* $H_0: \mu \leq 4.8$
- We will *reject* H_0 if $P(\bar{X} \geq 4.847)$ is small, e.g., $P(\bar{X} \geq 4.847) < 5\%$
- Then the p-value:
 - $P(\bar{X} \geq 4.847) = P\left(\frac{\bar{X}-\mu}{\sigma} \geq \frac{4.847-4.8}{\frac{0.234}{\sqrt{64}}}\right) = P\left(\frac{\bar{X}-\mu}{\sigma} \geq 1.6\right) = \Phi(-1.6) = 5.48\%$
- We (barely) accept H_0 at the level of significance 5%
- We reject H_0 at the level of significance 10%

Roadmap

- Expectation and Variance
- Sample Statistics
- Hypothesis Testing
- Statistics vs. Data Analytics
 - Naïve Bayes (cont.)

Descriptive vs. Inferential Statistics

- **Descriptive:** e.g., Median; describes data you have but can't be generalized beyond that
- **Inferential:** e.g., t-test, that enable inferences about the population beyond our data
 - These are the techniques we'll leverage for Machine Learning and Prediction

Examples of Business Questions

- **Simple (descriptive) Stats**
 - “Who are the most profitable customers?”
- **Hypothesis Testing**
 - “Is there a difference in value to the company of these customers?”
- **Segmentation/Classification**
 - What are the common characteristics of these customers?
- **Prediction**
 - Will this new customer become a profitable customer? If so, how profitable?

Statistics vs. Data Analytics

- Most business questions are causal: **what would happen if?** (e.g., I show this ad)
- But it's easier to ask **correlational** questions, (what happened in this past when I showed this ad).
- **Supervised Learning:**
 - Classification and Regression
- **Unsupervised Learning:**
 - Clustering and Dimension reduction
- Note: Unsupervised Learning is often used inside a larger Supervised learning problem.
 - E.g., latent topic discovery before topic categorization

Statistics vs. Data Analytics

- **Supervised Learning (Classification):**

- Naïve Bayes
- Logistic Regression
- Support Vector Machines
- Random Forests
- Neural Networks

A Data Sample (e.g., an image, a document)

↓
Classifier

A Class label (e.g., positive or negative sentiment)

- **Unsupervised Learning:**

- Clustering
- Factor analysis
- Latent Topic Modeling
- Auto-encoders

A Collection of Data Samples

↓
Clustering

Groups of data where data samples inside a group are more similar to each other.

Roadmap

- Expectation and Variance
- Sample Statistics
- Hypothesis Testing
- Statistics vs. Data Analytics
 - Naïve Bayes (cont.)

Put them in in the log space

Instead of this:

This:

$$c_{NB} = \operatorname{argmax}_{c_j \in C} P(c_j) \prod_{i \in \text{positions}} P(x_i | c_j)$$

Parameters to
be estimated
from the data

$$c_{NB} = \operatorname{argmax}_{c_j \in C} \left[\log P(c_j) + \sum_{i \in \text{positions}} \log P(x_i | c_j) \right]$$

This is ok since log doesn't change the ranking of the classes
(class with highest prob still has highest log prob)

Model is now just max of sum of weights: a *linear* function of
the inputs

So naive bayes is a *linear classifier*

Learning the Multinomial Naive Bayes Model

- First attempt: maximum likelihood estimates
 - simply use the frequencies in the data

$$\hat{P}(c_j) = \frac{doccount(C = c_j)}{N_{doc}}$$

$$\hat{P}(w_i | c_j) = \frac{count(w_i, c_j)}{\sum_{w \in V} count(w, c_j)}$$

V is the vocabulary maintaining all the words used for classification.

Parameter estimation

$$\hat{P}(w_i | c_j) = \frac{\text{count}(w_i, c_j)}{\sum_{w \in V} \text{count}(w, c_j)}$$

fraction of times word w_i appears
among all words in documents
with class c_j

- Create mega-document for class j by concatenating all docs with the class label
 - Use frequency of w in mega-document

Problem with Maximum Likelihood

- What if we have seen no training documents with the word ***fantastic*** and classified in the topic **positive (*thumbs-up*)?**

$$\hat{P}(\text{"fantastic"} \mid \text{positive}) = \frac{\text{count}(\text{"fantastic"}, \text{positive})}{\sum_{w \in V} \text{count}(w, \text{positive})} = 0$$

- Zero probabilities cannot be conditioned away, no matter the other evidence!

$$c_{MAP} = \operatorname{argmax}_c \hat{P}(c) \prod_i \hat{P}(x_i \mid c)$$

Laplace (add-1) Smoothing for Naive Bayes (empirically, it works!)

$$\begin{aligned}\hat{P}(w_i | c) &= \frac{\text{count}(w_i, c) + 1}{\sum_{w \in V} (\text{count}(w, c) + 1)} \\ &= \frac{\text{count}(w_i, c) + 1}{\left(\sum_{w \in V} \text{count}(w, c) \right) + |V|}\end{aligned}$$

Unknown words

- What about unknown words
 - that appear in our test data
 - but not in our training data or vocab
- We *ignore* them
 - Remove them from the test document!
 - Pretend they weren't there!
 - Don't include any probability for them at all.
- Why don't we build an unknown word model?
 - It doesn't help: knowing which class has more unknown words is not generally a useful thing to know!

Stop words

- Some systems ignore another class of words:
- **Stop words:** very frequent words like *the* and *a*.
 - Sort the whole vocabulary by frequency in the training, call the top 10 or 50 words the *stopword list*.
 - Now we remove all stop words from the training and test sets as if they were never there.
- But in some specific text classification applications, removing stop words don't help, so it's more common to **NOT** use stopword lists and use all the words in Naive Bayes (case-by-case).

Multinomial Naïve Bayes: Learning

- From training corpus, extract *Vocabulary*
- Calculate $P(c_j)$ terms
 - For each c_j in C do
$$docs_j \leftarrow \text{all docs with class } = c_j$$
$$P(c_j) \leftarrow \frac{|docs_j|}{\text{total # documents}}$$
 - Calculate $P(w_k | c_j)$ terms
 - $Text_j \leftarrow \text{single doc containing all } docs_j$
 - For each word w_k in *Vocabulary*
$$n_k \leftarrow \# \text{ of occurrences of } w_k \text{ in } Text_j$$
$$P(w_k | c_j) \leftarrow \frac{n_k + \alpha}{n + \alpha |Vocabulary|}$$

Let's do a worked sentiment example!

	Cat	Documents
Training	- just plain boring - entirely predictable and lacks energy - no surprises and very few laughs + very powerful + the most fun film of the summer	
Test	?	predictable with no fun

A worked sentiment example

	Cat	Documents
Training	-	just plain boring entirely predictable and lacks energy no surprises and very few laughs
	+	very powerful the most fun film of the summer

Test	?	predictable with no fun
------	---	------------------------------------

Prior from training:

$$P(-) = 3/5$$

$$P(+) = 2/5$$

Drop "with"

Likelihoods from training:

$$P(\text{"predictable"}|-) = \frac{1+1}{14+20}$$

$$P(\text{"predictable"}|+) = \frac{0+1}{9+20}$$

$$P(\text{"no"}|-) = \frac{1+1}{14+20}$$

$$P(\text{"no"}|+) = \frac{0+1}{9+20}$$

$$P(\text{"fun"}|-) = \frac{0+1}{14+20}$$

$$P(\text{"fun"}|+) = \frac{1+1}{9+20}$$

Scoring the test set:

$$P(-)P(S|-) = \frac{3}{5} \times \frac{2 \times 2 \times 1}{34^3} = 6.1 \times 10^{-5}$$

$$P(+)P(S|+) = \frac{2}{5} \times \frac{1 \times 1 \times 2}{29^3} = 3.2 \times 10^{-5}$$

A slide to take away

- What is ***expectation*** and ***variance***?
- What are the ***statistics*** to represent the data ?
- How to connect expectation with the sample mean with the ***law of large number***?
- How to describe the distribution of sample means with the ***central limit theorem***?
- What is the ***p-value*** and when to ***accept/reject a hypothesis***?
- How to estimate the likelihoods and priors to ***train a Naïve Bayes classifier***?