# UCLA STAT 19

## Order & Organization in the Stochastic Universe

**A *Fiat Lux* Course**

•**Instructor**: **Ivo Dinov,**

**Asst. Prof. In Statistics and Neurology**

**University of California, Los Angeles, Fall 2004**

***http://www.stat.ucla.edu/~dinov/courses_students.html***



## Announcement

- **Marschak Colloquium Series**
  - http://www.anderson.ucla.edu/research/marschak



## Experiments, Observations & Distributions

- SOCR Demos (all available online, see class web-page)
  - C:\Ivo.dir\UCLA_Classes\Applets.dir\SOCR\Prototype1.1\classes\TestDistribution.html
  - C:\Ivo.dir\UCLA_Classes\Applets.dir\SOCR\Prototype1.1\classes\TestExperiment.html
- Describing processes using distributions, instead of using precise numerical quantitative descriptions:
- **Examples**: Outcome of a coin-toss experiment, number of arrivals for a fixed time interval, DNA mutation rates, particle velocities/positions, light intensities, exam/test scores, length/weight measurements, etc.



## Discrete & Continuous Patterns of Disorder

- **Examples of discrete stochastic processes:**
- **Examples of continuous processes:**



## Mathematical/Statistical Modeling

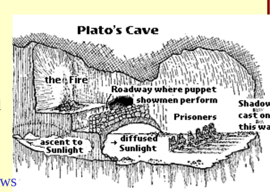
- **Modeling is an attempt to see the wood for the trees.**
- A model is a simplification or abstraction of reality separating the *important* from the *irrelevant*. Actually, modeling is a part of our existence.
- We could say that we do not perceive reality as it is. We only realize a model our mind has designed from sensory stimuli and their interpretation. It seems that certain animal species perceive different models of reality which, compared to ours, are based more on hearing and smell than on sight.



## Mathematical/Statistical Modeling

- Many philosophers have had deeper thoughts on this problem; following Plato's famous allegory of the cave. We may say that we only see the shadows of reality, or, following Kant, that we see the **phenomena** rather than the **noumena** (ground of the phenomena apprehended by the thought process).

Prisoners chained in a cave, unable to turn their heads. All they can see is the wall of the cave. Behind them burns a fire. Between the fire and the prisoners there is a parapet, along which puppeteers can walk. The puppeteers, who are behind the prisoners, hold up puppets that cast shadows on the wall of the cave. The prisoners are unable to see these puppets, the real objects, that pass behind them. What the prisoners see and hear are shadows and echoes cast by objects that they do not see.



## Mathematical/Statistical Modeling

- *We obtain our knowledge from models, and we make our predictions on the basis of models.*
- Since we are always modeling, modeling in the strict sense is the purposeful attempt to replace one model (the so-called "real world," which we typically accept without questioning) by another, deliberate, model which may give us more insight.
- There are two incentives for modeling:
  - either the **real-world** model is too complex to obtain the desired insight and so is replaced by a simpler or more abstract one.



## Mathematical/Statistical Modeling

- Or the **real-world** model does not allow certain experiments for ethical, practical, resource-limitations or other reasons. So, real-world model is replaced by a model in which all kinds of changes can be readily made and their consequences studied efficiently without causing harm.
- The word ***model*** traces back to the Latin word *modulus*, which means "little measure" (Merriam-Webster, 1994), alluding to a small-scale physical representation of a large object (e.g., a model airplane).



## Mathematical/Statistical Modeling

- **(Theoretical) Modeling** uses symbolic rather than physical representations, unleashing the power of mathematical analysis to increase scientific understanding. It can be divided into three stages (cf. Lin, Segel, 1974, 1988).

1. *Model formulation*: the translation of the scientific problem into mathematical terms.
2. *Model analysis*: the mathematical solution of the model thus created.
3. *Model interpretation and verification*: the interpretation of the solution and its empirical verification (validation) in terms of the original problem.



## Mathematical/Statistical Modeling

- The first step – **model formulation** – can lead to considerable insight. For building a math/stat model, one makes assumptions about the operating mechanisms, but often the real–model – the real-world – is far less understood than we expected.
- In many cases the modeling procedure – at least if one chooses parameters that are meaningful – already teaches what further knowledge is needed in order to apply the mathematical model successfully;
- The **model analysis** and its interpretation help to determine to what extent and precision new information and new data have to be collected.



## Mathematical/Statistical Modeling

- Analytic and numerical tools allow the extrapolation of present states of the mathematical model into the future and, sometimes, into the past.
- Assumptions, initial states, and parameters can easily be changed and the different outcomes compared. So, models can be used to identify trends or to estimate uncertainties in forecasts.
- While the model analysis may require sophisticated analytic or computational methods, mathematical modeling ideally leads to conceptual insight, which can be expressed without elaborate mathematics.



## Mathematical/Statistical Modeling

- A model is a simplification or abstraction; very often it is an **oversimplification** or **over-abstraction**. Insight obtained from a model should be checked against empirical evidence and common sense.
- It can also be checked against insight from other models: how much does the model's behavior depend on the degree of complexity, on the form of the model equations, on the choice of the parameters?
- Dealing with a concrete problem, a modeler should work with a whole scale of models starting from one which is as simple as possible.

## Mathematical/Statistical Modeling

- The use of a range of models also educates the modeler on how critically qualitative and quantitative results depend on the assumptions one has made.
- When modeling concrete phenomena, there is typically a dilemma between incorporating enough complexity (or realism) on the one hand and keeping the model tractable on the other.
- Extremely complex mathematical models will be of limited value for quantitative and maybe even qualitative forecasts, but still have the other benefits of being realistic.



## Mathematical/Statistical Modeling

- Mathematical modeling has its place in all sciences.
- Deterministic models (as opposed to stochastic models), which neglect the influence of random events.
- To some degree one can dispute whether stochastic models are more realistic than deterministic models; there is still the possibility that everything is deterministic, but just incredibly complex. In this case, stochasticity would simply be a certain way to deal with the fact that there are many factors we do not know.



## Mathematical/Statistical Modeling

- While a typical tool of deterministic-model analysis consists of discussing large-time limits, stochastic models take account of the **truism** that *nothing lasts forever* and make it possible to analyze the expected time until extinction--a concept that has no counterpart in deterministic models.



## Mathematical/Statistical Modeling

- In many cases, deterministic models can theoretically be justified as approximations of stochastic models for large populations sizes; however, the population size needed to make the approximation good enough may be unrealistically large.
- Nevertheless, deterministic models have the values which we described above, as long as one keeps their limitations in mind. The latter particularly concerns predictions, which are of very limited use in this uncertain world if no confidence intervals for the predicted phenomena are provided.

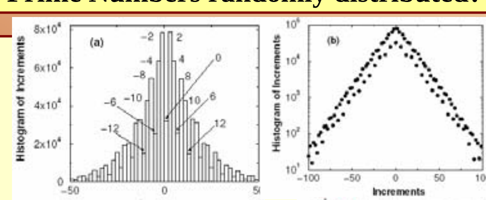


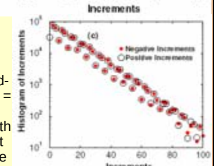
## Are Prime Numbers randomly distributed?

- The difference between two consecutive prime numbers is called the distance between the primes. This study of the statistical properties of the distances and their increments is for a sequence comprising the first $5\times10^7$ prime numbers. Results: the histogram of the increments follows an ***exponential distribution*** with superposed periodic behavior of **period three**, similar to previously-reported period six oscillations for the distances.
- ***Information Entropy and Correlations in Prime Numbers by*** P. Kumar, P.Ivanov, H. E. Stanley (2003)



## Are Prime Numbers randomly distributed?




Histogram of increments in the distances between consecutive prime numbers for the sequence of the first $10^6$ primes. The occurrence frequency of increments with given values exhibits a robust period-three oscillation. Increments with values ±(6k + 2) (k = 0, 1, 2, 3, ....) occur most often, increments with values ±(6k + 4) occur less often, and increments with values ±6k are rare. This regularity is always present regardless of the sequence length $N_p$. (b) Tent-shape of the histogram of increments on a **linear-log plot**.

## Are Prime Numbers randomly distributed?

- **Why Care?**
- The findings might have implications in the real world, as some systems in physics and biology - such as interacting prey and predator species with different life cycles - show patterns that depend on prime numbers.
- Coding Theory (e.g., Internet Security)
- ***Riemann hypothesis*** in number theory is intimately related to the distribution of primes. In 2001 the Clay Institute in the USA offered a prize of a million dollars for a proof of the this conjecture.
- Prime Number Th: number of primes ≤ x is: p(*x*) ~*x*/log *x*
- The ***Riemann hypothesis*** is that all nontrivial zeros of the zeta function are on the line Re(s)=½.

$$\zeta(s) = 1 + \frac{1}{2^s} + \frac{1}{3^s} + \frac{1}{4^s} + \frac{1}{5^s} + \cdots = \sum_{n=1}^{\infty} \frac{1}{n^s}$$



## Are Prime Numbers randomly distributed?

- RSA is an encryption method invented in 1978 by **R**ivest, **S**hamir, and **A**dleman at MIT in the USA, which is widely used nowadays in hardware and software to secure electronic data transport.
- This 'public key' method is based on the fact that, given the **product** of two carefully chosen **large prime numbers**, it is difficult to recover those numbers. The key question is: how large is sufficiently large to make this recovery virtually impossible? In the 1980s it was generally held that prime numbers of a **fifty odd digits** would suffice. However, developments went much faster than initially foreseen.

≤



## Are Prime Numbers randomly distributed?

- In 1977 Rivest challenged the world to factor RSA-129, a 129 digit number (from a special list), he estimated that on the basis of contemporary computational methods and computer systems this would take about $10^{16}$ years of computing time.
- Seventeen years later it **took only eight months** in a world-wide cooperative effort to do the job. Moreover, one should realize that it always remains possible that a new computational method is invented which makes factoring 'easy' (for example **quantum computing**, if an operative quantum computer will ever be realized).

≤



## Random Noise Generates 1-Way Spin

- A simple top converts foghorn noise to one-way spin. The device raises the hope that useful energy could be collected from ambient sounds. Normally, random vibrations, which physicists and engineers call noise, produce useless random motion. You can't move a cart from A to B by shoving it randomly in every direction.
- But in the new device, a flat plate encounters more friction when it spins in one direction than in the other, meaning it always rotates predictably.

•Norden, B., Zolotaryuk, Y., Christiansen, P.L. & Zolotaryuk, A.V. Ratchet device with broken friction symmetry. *Applied Physics Letters* **80,** 2601 - 2603 (2002).



## Noise breaks ice

- Natural randomness in the world's climate system may have caused the frequent, fast and fleeting returns to warm conditions during past ice ages.
- It's suggested that the events were caused by some kind of periodic influence on climate that repeated every 1,500 years. Perhaps a very weak periodic signal alters the ocean salt content every 1,500 years.
- There is evidence of a 1,500-year periodic forcing in many climate records. It is widely suspected to originate from repetitive changes in the activity of the Sun.

•Ganopolski, A. & Rahmstorf, S.
•Abrupt glacial climate changes due to noise-Infectious noise.
•*Physical Review Letters* **88,** 038501, (2002)



## Central Limit Theory and the Normal Distribution

- **SOCR Dice Demo** + CLT Applet Demo

  Applets.dir/SamplingDistributionApplet.html

## Duality on Completeness vs. Consistency

- In a famous lecture in 1900, David Hilbert listed 23 difficult problems he felt deserved the attention of mathematicians in the coming century.
- Some of these problems were solved quickly, others might never be completed, but all have influenced mathematics.
- Hilbert highlighted the need to **clarify the methods of mathematical reasoning**, using a formal system of explicit assumptions, or ***axioms***.

*Calude & Chaitin*, **Mathematics: Randomness everywhere**, *Nature* **400**, 319 - 320 (1999)



## Duality on Completeness vs. Consistency

- Hilbert stipulated that such a formal axiomatic system should be both **consistent** (free of contradictions) and **complete** (in that it represents all the truth).
- He also argued that any *well-posed* mathematical problem should be **decidable**, in the sense that there exists a mechanical procedure, a computer program, for deciding whether something is true or not.
- A problem is *ill-posed* if it may not have a solution, or the solution is not unique, or if small changes in initial conditions yield unpredictable/large changes in the final solution.



## Duality on Completeness vs. Consistency

- In 1931 Kurt Gödel showed that if you **assume a formal axiomatic system, containing elementary arithmetic, is consistent**, ➔ you can prove that it is **incomplete**. This was a huge surprise; everyone else thought Hilbert was right.
- The third condition (solvability of well-posed problems) was demolished by Alan Turing.
- Turing showed that no mechanical procedure, and therefore no formal axiomatic theory, can solve Turing's halting problem, the question of whether a given computer program will eventually halt.



## Duality on Completeness vs. Consistency

Hilbert's concern for consistency proofs led to Godel's Second Incompleteness Theorem.

Let *T* be a theory in the predicate calculus, satisfying certain mild conditions. Then:

1. *T* is incomplete.
2. The statement "*T* is consistent" is not a theorem of *T*. (Godel 1931)
3. The problem of deciding whether a given formula is a theorem of *T* is algorithmically unsolvable.



## Duality on Completeness vs. Consistency

- Turing's argument was based on computable real numbers. A real number, such as $\pi$, is a length measured with arbitrary precision, with an infinite number of digits.
- A **real number is computable** if there is a computer program or algorithm for calculating its digits one by one. There are programs for calculating $\pi$, but it is a surprising fact that **nearly all real numbers are not computable**.
- Turing showed that if you could find a mechanical procedure to decide if a computer program will ever halt, then you could compute a real number that is not actually computable, which is impossible.



## Randomness in Biology, Genetics, Engineering & Physics

- **Life is not ordered – life is organized.** Order is what a crystal (lattice) has.
- If you have 26 letters, as in English, you would expect a long sequence of (randomly chosen) characters to give each letter 1/26$^{th}$ of the time. That would be (uniformly) random.
- Random sequences have a high informational content, using information theory. A sequence can have lots of information regardless of whether it has any meaning.
- Now comes the problem that most anti-evolutionists don't quite grasp. Organized sequences are quite similar to random sequences.

## Randomness in Biology, Genetics, Engineering & Physics

- Organisms are often characterized as being **highly ordered** and in the same time as being highly organized. Clearly these terms have opposite meanings
- The message 01010101010101010101 is highly ordered and has a low entropy. A message highly organized is 01101100110111110001000.
- **Highly-organized** means that a long algorithm is needed to describe the sequence and therefore highly organized systems have a large entropy.
- **Highly-ordered** systems and **highly-organized** ones occupy opposite ends of the entropy scale.

Slide 31 — Stat 19, UCLA, Ivo Dinov

## Randomness in Biology, Genetics, Engineering & Physics

- **Highly-organized** systems are found embedded among random sequences, the latter occupying the high end of the entropy scale.
- Both, **random sequences** and **highly organized** sequences are complex (the shortest algorithm needed to compute a sequence is its **complexity**).
- **Information theory shows that it is fundamentally undecideable whether a given sequence has been generated by a stochastic process or by a highly organized process.**
- Algorithmic information theory shows that truth or validity may also be indeterminate or fundamentally undecidable.

High= Random Orgz'd — Entropy — Low= Order

Slide 32 — Stat 19, UCLA, Ivo Dinov

## Randomness in Biology, Genetics, Engineering & Physics

- It is **impossible** (or at least not clear how to) **to tell an organized** (designed) **sequence from one which is merely random**.
- If you can't tell an organized sequence with high informational content from a random sequence, then you **can't tell if the sequence arose through random processes or through an intelligence** who designed it.
- Meaningful high informational content patterns are rare compared to meaningless sequences with a high info content. However, **randomness can give rise to meaningful patterns**.
- **Difference between life and matter is information.**

Slide 33 — Stat 19, UCLA, Ivo Dinov

## Randomness in Biology, Genetics, Engineering & Physics

- Although humans have 30 times the DNA of some insects, there are insects that have more than double the DNA in humans.
- The amount of DNA is not a reliable measure of complexity because not all the DNA may have to do with complexity; part of a genome may be just many repeats of the same section, or random sections or just meaningless patterns.
- There are bacteria that are resistant to very high dosages or radiation – their DNA is mainly devoted to real time identification and correction of DNA breakage/mutations.

Slide 34 — Stat 19, UCLA, Ivo Dinov

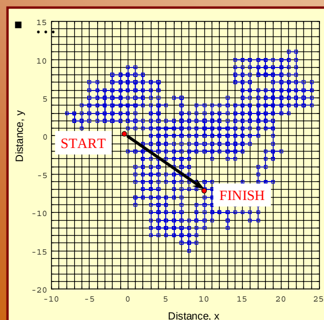
## Random Walks – The Gambler's Ruin

- Betting $1 on a game with a 50/50 chance: If you win, you get $2. If you lose, you get $0.
- A gambler bets $1 each round of a game random walk. However, he starts with **n** amount of money, whereas the bank/casino has unlimited funds.
- Let PN(n) denote the probability that, starting with **n** dollars, the gambler goes broke before winning N > **n** dollars.

  Previous Trial: Win Loose
- Solving PN(n) = (0.5)[ PN(n-1) + PN(n+1) ]
- Boundary conditions: PN(0) = 1, PN(N) = 0.
- Solution: PN(n) = 1 – n/N.

Slide 35 — Stat 19, UCLA, Ivo Dinov

## Random Walks




Each step has 4 possibilities:
- East
- West
- North
- South

Slide 36 — Stat 19, UCLA, Ivo Dinov

## Duality Principles: The Uncertainty Principle (momentum vs. position)

- Some physics experiments (such as blackbody radiation, the photoelectric effect, and Compton scattering) can be explained using the **photon picture of light** (discrete nature), and not with its wave properties.
- Other experiments, such as diffraction and interference, all need the **wave characteristics of light**. Considered as a photon (particle) the picture fails in these cases.
- We say that light exhibits a **wave-particle duality**: Light has a dual nature; in some cases it behaves as a wave, and in other cases it behaves as a photon.



## Duality Principles: The Uncertainty Principle (momentum vs. position)

- One important consequence of the wave-particle duality of nature was discovered by Heisenberg in 1926, and is called the (Heisenberg's) **uncertainty principle**.
- Imagine that we want to measure the ***position*** and the ***momentum*** of a particular particle. To do so we must see the particle, and so we shine some **light (as a wave)** of wavelength $\lambda$ on it. There is a limit to the resolving power of the light used to see the particle given by the wavelength of light used. This gives an *uncertainty* in the **particle's position**: $\Delta x \sim \lambda$.



## Duality Principles: The Uncertainty Principle (momentum vs. position)

- However, viewed as a photon, the light strikes the particle and gives up some or all of its momentum to the particle. Since we don't know how much it gave up, as we don't measure the photon's properties, there is an uncertainty in the **momentum** of the particle; $\Delta p \sim h/\lambda$, there h>0 is a constant.
- Hence, $\Delta x \times \Delta p \sim h$.
- A more refined treatment, developed by Heisenberg, results in the following relation:
  - $\Delta x \times \Delta p \geq h/4\pi$



## Duality Principles: The Uncertainty Principle (momentum vs. position)

  - $\Delta x \times \Delta p \geq h/4\pi$
- Note that this is **independent of the wavelength used**, and says there is a **limit as to how accurately one can simultaneously measure the position ($\Delta x$) and momentum of a particle ($\Delta p$)**.
- If one tries to measure the position more accurately, by using light of a shorter wavelength ($\lambda \to 0$), then the uncertainty in the momentum grows.
- Whereas if one uses light of a longer wavelength in order to reduce the uncertainty in momentum, then the uncertainty in position grows.



## Balancing Quality vs Volume of Information

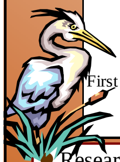
- **Quality – Quantity Duality**: You can't have both a large amount of information (data) with perfect quality. Increasing the volume of the data usually decreases its quality, conversely increasing the quality requires a decrease of the quantity.



## Statistical vs. Practical Significance

- **Is a second child gender influenced by the gender of the first child, in families with >1 kid?**
- When analyzing real data, investigators frequently employ statistical analytic techniques to detect real signal/effects in the data. Hence **statistically significant effects** are determined by a statistical analysis.
- How **practically meaningful**, however, are these statistically significant effects? Answer: Not clear, in general.

## Is a second child gender influenced by the gender of the first child, in families with >1 kid?



**First and Second Births by Sex**

| | | Second Child | | |
|---|---|---|---|---|
| | | **Male** | **Female** | **Total** |
| First Child | Male | 3,202 | 2,776 | 5,978 |
| | Female | 2,620 | 2,792 | 5,412 |
| | Total | 5,822 | 5,568 | 11,390 |

- Research hypothesis needs to be formulated first before collecting/looking/interpreting the data that will be used to address it. Mothers whose 1st child is a girl are more likely to have a girl, as a second child, compared to mothers with boys as 1st children.
- Data: 20 yrs of birth records of 1 Hospital in Auckland, NZ.



## Analysis of the birth-gender data – data summary

| | Girl as a Second Child | |
|---|---|---|
| **Group** | **Number of births** | **Number of girls** |
| 1 (Previous child was girl) | 5412 | 2792 (approx. 51.6%) |
| 2 (Previous child was boy) | 5978 | 2776 (approx. 46.4%) |

- Let $p_1$=true proportion of girls in mothers with girl as first child, $p_2$=true proportion of girls in mothers with boy as first child. Parameter of interest is $p_1 - p_2$.
- $H_0$: $p_1 - p_2 = 0$ (skeptical reaction). $H_a$: $p_1 - p_2 > 0$ (research hypothesis)



## Hypothesis testing as decision making

**Decision Making**

| | Actual situation | |
|---|---|---|
| Decision made | $H_0$ is true | $H_0$ is false |
| Accept $H_0$ as true | OK | Type II error |
| Reject $H_0$ as false | Type I error | OK |

- Sample sizes: $n_1$=5412, $n_2$=5978, Sample proportions (estimates) $\hat{p}_1 = 2792/5412 \approx 0.5159$, $\hat{p}_2 = 2776/5978 \approx 0.4644$,
- $H_0$: $p_1 - p_2 = 0$ (skeptical reaction). $H_a$: $p_1 - p_2 > 0$ (research hypothesis)



## Analysis of the birth-gender data

- Samples are large enough to use Normal-approx. Since the two proportions come from totally diff. mothers they are independent → use formula 8.5.5.a

$$t_0 = \frac{\text{Estimate - HypothesizedValue}}{SE} = 5.49986 =$$

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

$$P - value = \Pr(T \geq t_0) = 1.9 \times 10^{-8}$$



## Analysis of the birth-gender data

- We have strong evidence to reject the $H_0$, and hence conclude mothers with first child a girl a more likely to have a girl as a second child.
- How much more likely? A 95% CI:

CI ($p_1 - p_2$) =[0.033; 0.070]. And computed by:

$$\text{estimate} \pm z \times \text{SE} = \hat{p}_1 - \hat{p}_2 \pm 1.96 \times SE\left(\hat{p}_1 - \hat{p}_2\right) =$$

$$\hat{p}_1 - \hat{p}_2 \pm 1.96 \times \sqrt{\frac{\hat{p}_1(1-\hat{p}_1)}{n_1} + \frac{\hat{p}_2(1-\hat{p}_2)}{n_2}} =$$

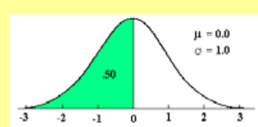
$$0.0515 \pm 1.96 \times 0.0093677 = [3\%;7\%]$$



## Standard Normal Curve (cf. Modeling)

- The standard normal curve is described by the equation:

$$y = \frac{e^{-\frac{x^2}{2}}}{\sqrt{2\pi}}$$

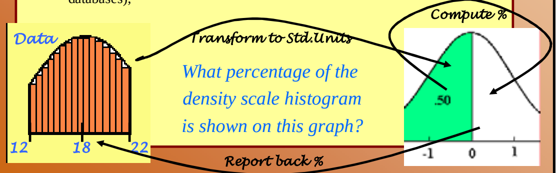



Where remember, the natural number $e \sim 2.7182...$

We say: $X \sim Normal(\mu, \sigma)$, or simply $X \sim N(\mu, \sigma)$

## Standard Normal Approximation

- The standard normal curve can be used to estimate the percentage of entries in an interval for any process. Here is the protocol for this approximation:
  - Convert the interval (we need the assess the percentage of entries in) to standard units. We saw the algorithm already.
  - Find the corresponding area under the normal curve (from tables or online databases);



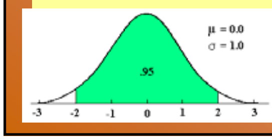


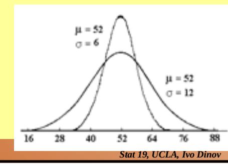

## General Normal Curve

- The general normal curve is defined by:
  - Where $\mu$ is the average of (the symmetric) normal curve, and $\sigma$ is the standard deviation (spread of the distribution).

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

  - Why worry about a standard and general normal curves?
  - How to convert between the two curves?

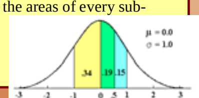





## Areas under Standard Normal Curve – Normal Approximation

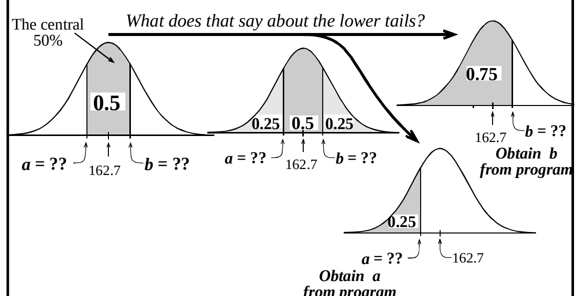
- Protocol:
  - Convert the interval (we need to assess the percentage of entries in) to Standard units. Actually convert the end points in Standard units.
    - ❑ In general, the transformation X → (X-μ)/σ, standardizes the observed value X, where μ and σ are the average and the standard deviation of the distribution X is drawn from.
  - Find the corresponding area under the normal curve (from tables or online databases);
    - ❑ Sketch the normal curve and shade the area of interest
    - ❑ Separate your area into individually computable sections
    - ❑ Check the Normal Table and extract the areas of every sub-section
    - ❑ Add/compute the areas of all sub-sections to get the total area.






## Obtaining central range for symmetric distributions

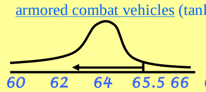
What values contain the **central** 50%?





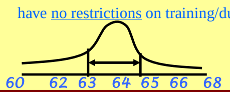

## Areas under Standard Normal Curve – Normal Approximation, Scottish Army Recruits

- The mean height is 64 *in* and the standard deviation is 2 *in*.
  - Only recruits shorter than 65.5 in will be trained for tank operation. What percentage of the incoming recruits will be trained to operate armored combat vehicles (tanks)?




X → (X-64)/2
65.5 → (65.5-64)/2 = ¾
Percentage is 77.34%

  - Recruits within ½ standard deviations of the mean will have no restrictions on duties. About what percentage of the recruits will have no restrictions on training/duties?




X → (X-64)/2
65 → (65-64)/2 = ½
63 → (63-64)/2 = -½
Percentage is 38.30%



## Statistics of Extremes

- SOCR Demo:
  - C:\Ivo.dir\UCLA_Classes\Applets.dir\SOCR\Prototype1.1\classes\TestDistribution.html

## Interpolation vs. Extrapolation

- **Interpolation** is the process of estimating a value for a point that lies on a curve between known data points
  - Linear interpolation assumes a straight line between the known data points
- One Method:
  - Select the two points with known coordinates
  - Determine the equation of the line that passes through the two points (Find m and b of the line)
  - Insert the X value of the desired point in the equation and calculate the Y value (knowing ***y = mx + b***)



## Interpolation vs. Extrapolation

Linear Interpolation:

$$\text{Fractional Difference} = \left(\frac{x - x_1}{x_2 - x_1}\right) = \left(\frac{y - y_1}{y_2 - y_1}\right)$$

$$y = y_1 + \left(\frac{y_2 - y_1}{x_2 - x_1}\right)(x - x_1)$$

- Given the following set of points, find the dependent variable y2 using linear interpolation.

  (x1, y1) = (1.0, 18)

  (x, y) = (2.4, **y** )

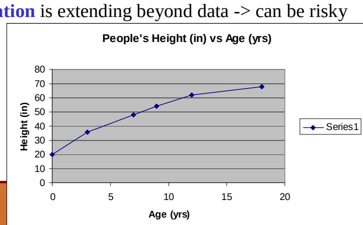
  (x2, y2) = (4.0, 35)



## Interpolation vs. Extrapolation

- **Interpolation** is extending between data points

  -> usually safe
- **Extrapolation** is extending beyond data -> can be risky






## Bayesian Theory – Statistical Inference

- Statistical inference is based on probability
  - To be useful probability must be interpreted.
    - Relative Frequency (Venn, Fisher, Neyman, etc.)
    - Degree of Belief (Bayes, Laplace, Gauss, Jeffreys, etc.)
    - Propensity (Popper, etc.)
  - The validity of these interpretations cannot be decided by an appeal to Nature.
  - Statistical inference is based on principles that can always be challenged by anyone who doesn't find all of them compelling. Again, Nature cannot help.
  - Statistical inference cannot be fully objective.



## Frequentist Inference

- The Good
  - No **arbitrary priors**: Absence of prior anxiety!
  - Coverage property is powerful (some say beautiful)
  - There is a **badness-of-fit test**
  - One can play delightful MC games on a computer
- The Bad
  - No systematic method to incorporate prior information
  - "Grosse Fuge" reasoning is difficult and unnatural
- The Ugly
  - Difficult to teach
  - Doesn't do what we want: Probability(Theory|Data)



## Bayesian Inference

- The Good
  - Natural **model of inferential reasoning**
  - General theory for handling uncertainty in all its forms
  - Results depend only on data observed
  - Does what we want: Probability(Theory|Data)
  - Easy to teach and understand
- The Bad
  - Can be computationally demanding
  - Until recently, no goodness-of-fit test
- The Ugly
  - Choosing prior probabilities can be, well, a "Grosse Fuge"!

## Bayesian Theory – Statistical Inference

- "A Frequentist uses impeccable logic to answer the wrong question, while a Bayesian answers the right question by making assumptions that nobody can fully believe in."

  P.G. Hamer

- Do you see another version of the Duality-Principle in action? (Freq/Bayes)

Slide 63 Stat 19, UCLA, Ivo Dinov

## Bayesian Decision Theory

- Bayesian decision theory is a fundamental statistical approach to the problem of pattern classification.
  - Decision making when all the probabilistic information is known.
  - For given probabilities the decision is optimal.
  - When new information is added, it is assimilated in optimal fashion for improvement of decisions.

Slide 64 Stat 19, UCLA, Ivo Dinov

## Bayesian Decision Theory cont.

- Fish Example: Each fish is in one of 2 states: **sea bass** or **salmon**
- Let ω denote the ***state of nature***
  - $\omega = \omega_1$ for sea bass
  - $\omega = \omega_2$ for salmon

Slide 65 Stat 19, UCLA, Ivo Dinov

## Bayesian Decision Theory cont.

- The State of nature is unpredictable ➔ ω is a variable that must be described probabilistically.
- If the catch produced as much *salmon* as *sea-bass* the next fish is equally likely to be sea bass or salmon.
- Define
  - $P(\omega_1)$ : a priori probability that the next fish is sea bass
  - $P(\omega_2)$: a priori probability that the next fish is salmon.

Slide 66 Stat 19, UCLA, Ivo Dinov

## Bayesian Decision Theory cont.

- If other types of fish are irrelevant:

$$P(\omega_1) + P(\omega_2) = 1.$$

- *Prior probabilities reflect our prior knowledge (e.g. time of year, fishing area, …)*
- *Simple decision Rule:*
  - *Make a decision without seeing the fish.*
  - *Decide $w_1$ if $P(\omega_1) \geq P(\omega_2)$; $\omega_2$ otherwise.*
  - *OK if deciding for one fish*
  - *If several fish, all assigned to same class.*

Slide 67 Stat 19, UCLA, Ivo Dinov

## Bayesian Decision Theory cont.

- In general, we will have some features and more information.
- Feature: lightness measurement = $x$
  - *Different fish yield different lightness readings (x is a random variable)*

Slide 68 Stat 19, UCLA, Ivo Dinov

## Bayesian Decision Theory cont.

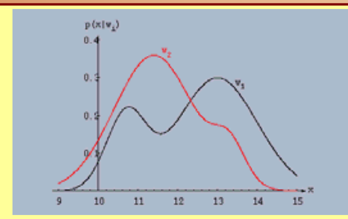
- Define

  **p(x | $\omega_1$)** = Class Conditional Probability Density Probability density function for x given that the state of nature is $\omega_1$

- The difference between **p(x | $\omega_1$)** and **p(x | $\omega_2$)** describes the difference in lightness between sea bass and salmon.



## Bayesian Decision Theory cont.




Hypothetical class-conditional probability Density functions are normalized (area under each curve is 1.0)



## Bayesian Decision Theory cont.

- Suppose that we know

  The prior probabilities P($\omega_1$) and P($\omega_2$),

  The conditional densities $p(x|\omega_1)$ $p(x|\omega_2)$

  Measure lightness of a fish = x.

- What is the category (class) of the fish given the evidence (light) $p(\omega_j | x)$ ?



## Bayesian Formula

$$P(\omega_j | x) = P(x | \omega_j)\, P(\omega_j) / P(x),$$

where

$$P(x) = \sum_{j=1}^{2} p(x|\omega_j) P(\omega_j)$$

$$\text{Posterior} = \frac{\text{Likelihood} * \text{Prior}}{\text{Evidence}}$$

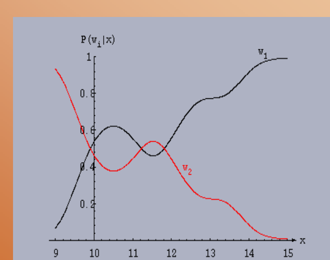


## Bayesian Formula

- $p(x|\omega_j)$ is called the ***likelihood*** of $\omega_j$ with respect to ***x***. (the $\omega_j$ category for which $p(x|\omega_j)$ is large is more "likely" to be the true category)
- p(x) is the ***evidence*** how frequently we will measure a pattern with feature value x.
- Scale factor that guarantees that the posterior probabilities sum to 1.



**Bayes' formula cont.**




Posterior probabilities for the particular priors P($\omega_1$)=2/3 and P($\omega_2$)=1/3. At every x the posteriors sum to 1.

## Error

$$P(error \mid x) = \begin{cases} \text{If we decide } \omega_2 \Rightarrow P(\omega_1 \mid x) \\ \text{If we decide } \omega_1 \Rightarrow P(\omega_2 \mid x) \end{cases}$$

For a given $x$, we can minimize the probability of error by deciding $\omega_1$ *if* $P(\omega_1|x) > P(\omega_2|x)$ *and* $\omega_2$ *otherwise.*



## Bayes' Decision Rule
(Minimizes the probability of error)

$\omega_1$ : *if* $P(\omega_1|x) > P(\omega_2|x)$

$\omega_2$ : *otherwise*

or

$\omega_1$ : *if* $P(x|\omega_1)P(\omega_1) > P(x|\omega_2)P(\omega_2)$

$\omega_2$ : *otherwise*

and

$P(Error|x) = \min[P(\omega_1|x), P(\omega_2|x)]$



## Bayesian Decision Theory: Continuous Features: General Case

Formalize the ideas just considered in 4 ways:

- *Allow more than one feature*

  Replace the scalar x by the feature vector $\mathbf{x} \in R^d$. An d-dimensional Euclidean space $R^d$ is called the feature space.
- *Allow more than 2 states of nature*

  Generalize to several classes
- *Allow actions other than merely deciding the state of nature*

Possibility of rejection, i.e., of refusing to make a decision in close cases.

- *Introducing general loss function*



## The Loss Function

- ***Loss ( or cost ) function*** states exactly how costly each action is, and is used to convert a probability determination into a decision. Loss functions let us treat situations in which some kinds of classification mistakes are more costly than others.



## Problem Formulation

- Let $\{\omega_1, \dots, \omega_c\}$ be the finite set of $c$ states of nature ("categories").
- Let $\{\alpha_1, \dots, \alpha_a\}$ be the finite set of a possible actions.
- The loss function $\lambda(\alpha_i \mid \omega_j)$ = loss incurred for taking action $\alpha_i$ when the state of nature is $\omega_j$.
- $\mathbf{x}$ = d-dimensional feature vector (random variable)
- $P(\mathbf{x}|\omega_j)$ = the state conditional probability density function for $\mathbf{x}$ (The probability density function for $\mathbf{x}$ conditioned on $\omega_j$ being the true state of nature)
- $P(\omega_j)$ = prior probability that nature is in state $\omega_j$.



## Expected Loss

- Suppose that we observe a particular $\mathbf{x}$ and that we contemplate taking action $\alpha_i$.
- If the true state of nature is $\omega_j$ then loss is

$$\lambda(\alpha_i \mid \omega_j)$$

- Before we have done an observation the expected loss is

$$R(\alpha_i) = \sum_{j=1}^{C} \lambda(\alpha_i \mid \omega_j) P(\omega_j)$$

## Conditional Risk

- After the observation the expected risk which is called now "conditional risk" is given by

$$R(\alpha_i \mid \mathbf{x}) = \sum_{j=1}^{C} \lambda(\alpha_i \mid \omega_j) P(\omega_j \mid \mathbf{x})$$



## Total Risk

- Objective: Select the action that minimizes the conditional risk
- A general decision rule is a function $\alpha_i(\mathbf{x})$
- For every $\mathbf{x}$, the decision function $\alpha_i(\mathbf{x})$ assumes one of the a values $\alpha_1, \ldots, \alpha_a$
- The "total risk" is

$$\int R(\alpha(\mathbf{x}) \mid \mathbf{x})\, p(\mathbf{x})\, d\mathbf{x}$$



## Bayesian Decision Rule

- Compute the conditional risk

$$R(\alpha_i \mid \mathbf{x}) = \sum_{j=1}^{C} \lambda(\alpha_i \mid \omega_j) P(\omega_j \mid \mathbf{x})$$

for i =1, ... , a.

- Select the action $\alpha_i$ for which $R(\alpha_i \mid \mathbf{x})$ is minimum.
- The resulting minimum total risk is called the Bayes Risk, denoted R*, and is ***the best performance*** that can be achieved.



## Two-Category Classification

- Action $\alpha_1 =$ deciding that the true state is $\omega_1$
- Action $\alpha_2 =$ deciding that the true state is $\omega_2$ .
- Let $\lambda_{ij} = \lambda(\alpha_i \mid \omega_j)$ be the loss incurred for deciding $\omega_i$ when the true state is $\omega_j$.

$$R(\alpha_1 \mid \mathbf{x}) = \lambda_{11} P(\omega_1 \mid \mathbf{x}) + \lambda_{12} P(\omega_2 \mid \mathbf{x})$$

$$R(\alpha_2 \mid \mathbf{x}) = \lambda_{21} P(\omega_1 \mid \mathbf{x}) + \lambda_{22} P(\omega_2 \mid \mathbf{x})$$

- Decide $\omega_1$ if $R(\alpha_1 \mid \mathbf{x}) < R(\alpha_2 \mid \mathbf{x})$

  or if $(\lambda_{21} - \lambda_{11}) P(\omega_1 \mid \mathbf{x}) > (\lambda_{12} - \lambda_{22}) P(\omega_2 \mid \mathbf{x})$

  or if $(\lambda_{21} - \lambda_{11}) p(\mathbf{x} \mid \omega_1) P(\omega_1) > (\lambda_{12} - \lambda_{22}) p(\mathbf{x} \mid \omega_2) P(\omega_2)$

  and $\omega_2$ *otherwise*



## Two-Category Likelihood Ratio Test

- Under reasonable assumption that $\lambda_{21} > \lambda_{11}$ (why?)

  decide $\omega_1$ if

$$\frac{p(\mathbf{x} \mid \omega_1)}{p(\mathbf{x} \mid \omega_2)} > \frac{(\lambda_{12} - \lambda_{22})}{(\lambda_{21} - \lambda_{11})} \frac{P(\omega_2)}{P(\omega_1)} = T$$

and $\omega_2$ *otherwise.*

The ratio $\frac{p(\mathbf{x} \mid \omega_1)}{p(\mathbf{x} \mid \omega_2)}$ is called the *likelihood ratio*. We can decide $\omega_1$ if the likelihood ratio exceeds a threshold $T$ value that is independent of the observation x.