Basic Ideas in Probability and Statistics for Experimenters: Part I: Qualitative Discussion

He uses statistics as a drunken man uses lamp-posts – for support rather than for illumination … A. Lang

Shivkumar Kalyanaraman
Rensselaer Polytechnic Institute
shivkuma@ecse.rpi.edu
http://www.ecse.rpi.edu/Homepages/shivkuma
Overview

- Why Probability and Statistics: The Empirical Design Method...
- Qualitative understanding of essential probability and statistics
  - Especially the notion of inference and statistical significance
  - Key distributions & why we care about them...
- Reference: Chap 12, 13 (Jain) and
- [http://mathworld.wolfram.com/topics/ProbabilityandStatistics.html](http://mathworld.wolfram.com/topics/ProbabilityandStatistics.html)
Theory (Model) vs Experiment

- A model is only as good as the **NEW predictions** it can make (subject to “statistical confidence”)

- **Physics:**
  - Theory and experiment frequently changed roles as leaders and followers
  - Eg: Newton’s Gravitation theory, Quantum Mechanics, Einstein’s Relativity.
  - Einstein’s “thought experiments” vs real-world experiments that validate theory’s predictions

- **Medicine:**
  - FDA Clinical trials on New drugs: Do they work?
  - “Cure worse than the disease”?

- **Networking:**
  - How does OSPF or TCP or BGP behave/perform?
  - Operator or Designer: What will happen if I change .....?
Why Probability? BIG PICTURE

- Humans like *determinism*
- The real-world is unfortunately *random*!
- => CANNOT place ANY confidence on a single measurement or simulation result that contains potential randomness
- However,
  - We can make deterministic statements about *measures or functions of underlying randomness* …
    - … with bounds on degree of confidence
- Functions of Randomness:
  - Probability of a random event or variable
  - Average (mean, median, mode), Distribution functions (pdf, cdf), joint pdfs/cdfs, conditional probability, confidence intervals,
- Goal: Build “probabilistic” models of reality
  - Constraint: minimum # experiments
- Infer to get a model (i.e. maximum information)
  - Statistics: how to infer models about reality (“population”) given a SMALL set of expt results (“sample”)
Why Care About Statistics?

Measure, simulate, experiment

How to make this empirical design process EFFICIENT??

How to avoid pitfalls in inference!

Model, Hypothesis, Predictions

FIGURE 1.3. Data generation and data analysis in scientific investigation.
Why Care? Real-World Measurements: Internet Growth Estimates

- Growth of 80%/year
- Sustained for at least ten years ...
- ... before the Web even existed.

⇒ Internet is always changing. You do not have a lot of time to understand it.

Exponential Growth Model Fits
Probability

- Think of probability as **modeling an experiment**
- Eg: tossing a coin!
- The set of all possible **outcomes** is the **sample space**: $S$

- Classic “Experiment”: Tossing a die: $S = \{1,2,3,4,5,6\}$
- Any subset $A$ of $S$ is an **event**: $A = \{\text{the outcome is even}\} = \{2,4,6\}$
Probability of Events: Axioms

• $P$ is the Probability Mass function if it maps each event $A$, into a real number $P(A)$, and:
  
  i.) $P(A) \geq 0$ for every event $A \subseteq S$
  
  ii.) $P(S) = 1$

  iii.) If $A$ and $B$ are mutually exclusive events then,

  $$P(A \cup B) = P(A) + P(B)$$
Probability of Events

...In fact for any sequence of pair-wise-mutually-exclusive events, we have

\[ A_1, A_2, A_3, \ldots \quad (\text{i.e. } A_i A_j = 0 \text{ for any } i \neq j) \]

\[ P \left( \bigcup_{n=1}^{\infty} A_n \right) = \sum_{n=1}^{\infty} P(A_n) \]
Other Properties Can be Derived

- \( P(\overline{A}) = 1 - P(A) \)
- \( P(A) \leq 1 \)
- \( P(A \cup B) = P(A) + P(B) - P(AB) \)
- \( A \subseteq B \Rightarrow P(A) \leq P(B) \)

Derived by breaking up above sets into mutually exclusive pieces and comparing to fundamental axioms!!
Recall: Why care about Probability?

- … We can be deterministic about **measures or functions of underlying randomness** …
  - Functions of Randomness:
    - Probability or Odds of a random event or variable;
    - “Risk” associated with an event: Probability*RV
      - Or loosely “expectation” of the event
    - “Upside” risks vs “Downside” risks
    - Measures of dispersion, skew (a.k.a higher-order “moments”)
  
- Even though the experiment has a **RANDOM OUTCOME** (eg: 1, 2, 3, 4, 5, 6 or heads/tails) or **EVENTS** (subsets of all outcomes)
  - The probability function has a **DETERMINISTIC** value

- *If you forget everything in this class, do not forget this!*
Conditional Probability

- \( P(A | B) = (\text{conditional}) \) probability that the outcome is in \( A \) given that we know the outcome in \( B \)

\[
P(A | B) = \frac{P(AB)}{P(B)} \quad P(B) \neq 0
\]

- Example: Toss one die.

\[
P(i = 3 | i \text{ is odd}) = \quad \text{Note that: } P(AB) = P(B)P(A | B) = P(A)P(B | A)
\]

What is the value of knowledge that \( B \) occurred? How does it reduce uncertainty about \( A \)? How does it change \( P(A) \)?
Independence

- Events $A$ and $B$ are independent if $P(AB) = P(A)P(B)$.
- Also: $P(A | B) = P(A)$ and $P(B | A) = P(B)$
- Example: A card is selected at random from an ordinary deck of cards.
  - $A$=event that the card is an ace.
  - $B$=event that the card is a diamond.

\[
P(AB) = \\
\]
\[
P(A) = \\
\]
\[
P(A)P(B) = \\
\]
\[
P(B) = \\
\]
Random Variable as a Measurement

- We cannot give an exact description of a sample space in these cases, but we can *still describe specific measurements* on them.
  - The temperature change produced.
  - The number of photons emitted in one millisecond.
  - The time of arrival of the packet.
Random Variable as a Measurement

- Thus a random variable can be thought of as a measurement on an experiment
Histogram: Plotting Frequencies

<table>
<thead>
<tr>
<th>Class</th>
<th>Freq.</th>
</tr>
</thead>
<tbody>
<tr>
<td>15 but &lt; 25</td>
<td>3</td>
</tr>
<tr>
<td>25 but &lt; 35</td>
<td>5</td>
</tr>
<tr>
<td>35 but &lt; 45</td>
<td>2</td>
</tr>
</tbody>
</table>

Count

Frequency

Relative Frequency

Percent

Bars Touch

Lower Boundary
Probability Distribution Function (pdf)

**FIGURE 2.4.** Probability distribution for a conceptual population of yield values.

a.k.a. frequency histogram, p.m.f (for discrete r.v.)
The probability mass function (PMF) for a (discrete valued) random variable $X$ is:

$$P_X(x) = P(X = x) = P\{s \in S \mid X(s) = x\}$$

- Note that $P_X(x) \geq 0$ for $-\infty < x < \infty$

- Also for a (discrete valued) random variable $X$

$$\sum_{x=-\infty}^{\infty} P_X(x) = 1$$
PMF and CDF: Example
The cumulative distribution function (CDF) for a random variable \( X \) is

\[
F_X(x) = P(X \leq x) = P(\{s \in S \mid X(s) \leq x\})
\]

Note that \( F_X(x) \) is non-decreasing in \( x \), i.e.

\[
x_1 \leq x_2 \implies F_X(x_1) \leq F_X(x_2)
\]

Also

\[
\lim_{x \to -\infty} F_X(x) = 0 \quad \text{and} \quad \lim_{x \to \infty} F_X(x) = 1
\]
**Probability density functions (pdf)**

Emphasizes main body of distribution, frequencies, various modes (peaks), variability, skews
Cumulative Distribution Function (CDF)

- Lognormal(0,1)
- Gamma(.53,3)
- Exponential(1.6)
- Weibull(0.7,0.9)
- Pareto(1,1.5)

Emphasizes skews, easy identification of median/quartiles, converting uniform rvs to other distribution rvs.
**Complementary CDFs (CCDF)**

Useful for focusing on “tails” of distributions:
- Line in a log-log plot => “heavy” tail

Graph showing various distributions with log-log axes.
Real Example: Distribution of HTTP Connection Sizes is Heavy Tailed

CCDF plot on log-log scale
Recall: Why care about Probability?

- Humans like *determinism*
  - The real-world is unfortunately *random*!
  - **CANNOT place ANY confidence** on a **single** measurement

- We can be deterministic about *measures or functions of underlying randomness* ...

- Functions of Randomness:
  - Probability of a random event or variable
  - Average (mean, median, mode), Distribution functions (pdf, cdf), joint pdfs/cdfs, conditional probability, confidence intervals,
Numerical Data Properties

Central Tendency (Location)

Variation (Dispersion)

Shape
Numerical Data Properties & Measures

Numerical Data Properties

Central Tendency
- Mean
- Median
- Mode

Variation
- Range
- Interquartile Range
- Variance
- Standard Deviation

Shape
- Skew
Indices of Central Tendency (Location)

- Goal: Summarize (compress) entire distribution with one “representative” or “central” number
- Eg: Mean, Median, Mode
- Allows you to “discuss” the distribution with just one number
- Eg: What was the “average” TCP goodput in your simulation?
- Eg: What is the median HTTP transfer size on the web (or in a data set)?
Expectation (Mean) = Center of Gravity

**FIGURE 2.7.** The mean $\eta = E(y)$ as the center of gravity of a distribution.
The expectation (average) of a (discrete-valued) random variable $X$ is

$$
\bar{X} = E(X) = \sum_{x=-\infty}^{\infty} xP(X = x) = \sum_{x=-\infty}^{\infty} xP_X(x)
$$

Three coins example:

$$
E(X) = \sum_{x=0}^{3} xP_X(x) = 0 \times \frac{1}{8} + 1 \times \frac{3}{8} + 2 \times \frac{3}{8} + 3 \times \frac{1}{8} = 1.5
$$
Median, Mode

- **Median** = $F^{-1}(0.5)$, where $F = \text{CDF}$
  - Aka 50% percentile element
  - I.e. Order the values and pick the middle element
  - Used when distribution is skewed
  - Considered a “robust” measure

- **Mode**: Most frequent or highest probability value
  - Multiple modes are possible
  - Need not be the “central” element
  - Mode may not exist (e.g., uniform distribution)
  - Used with categorical variables
FIGURE 12.1 Five distributions showing relationships among mean, median, and mode.
# Summary of Central Tendency Measures

<table>
<thead>
<tr>
<th>Measure</th>
<th>Equation</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>Mean</td>
<td>$\frac{\sum X_i}{n}$</td>
<td>Balance Point</td>
</tr>
<tr>
<td>Median</td>
<td>$(n+1)$ Position [ \frac{2}{2} ]</td>
<td>Middle Value When Ordered</td>
</tr>
<tr>
<td>Mode</td>
<td>none</td>
<td>Most Frequent</td>
</tr>
</tbody>
</table>
Games with Statistics!

$400,000  

... employees cite low pay -- most workers earn only $20,000. [mode]

$70,000  

... President claims average pay is $70,000! [mean]

$50,000  

$30,000  

$20,000  

Who is correct? BOTH!

Issue: skewed income distribution
FIGURE 12.2  Selecting among the mean, median, and mode.
Indices/Measures of Spread/Dispersion: Why Care?

You can drown in a river of average depth 6 inches!

**Lesson:** The measure of uncertainty or dispersion may matter more than the index of central tendency.
Real Example: dispersion matters!

- What is the fairness between TCP goodputs when we use different queuing policies?
- What is the confidence interval around your estimates of mean file size?
Standard Deviation, Coeff. Of Variation, SIQR

- **Variance**: second moment around the mean:
  
  \[ \sigma^2 = E((X-\mu)^2) \]

- **Standard deviation** = \( \sigma \)

\[
\text{stdv}(x) = \sigma = \sqrt{\langle x^2 \rangle - \langle x \rangle^2} = \sqrt{\mu'_2 - \mu^2},
\]

- **Coefficient of Variation (C.o.V.)** = \( \sigma/\mu \)

- **SIQR** = Semi-Inter-Quartile Range (used with median = 50\(^{th}\) percentile)
  
  \( (75^{th} \text{ percentile} - 25^{th} \text{ percentile})/2 \)
Real Example: CoV

- Goal: TCP rate control vs RED in terms of fairness (measured w/ CoV)

- Lower CoV is better

<table>
<thead>
<tr>
<th>Provider Metrics</th>
<th>User Metrics</th>
</tr>
</thead>
<tbody>
<tr>
<td>Speed Mbps</td>
<td>Utilization Percent</td>
</tr>
<tr>
<td>0.056</td>
<td>98.79</td>
</tr>
<tr>
<td>0.384</td>
<td>99.68</td>
</tr>
<tr>
<td>1.5</td>
<td>99.85</td>
</tr>
<tr>
<td>10</td>
<td>99.54</td>
</tr>
<tr>
<td>45</td>
<td>94.17</td>
</tr>
</tbody>
</table>

Table 1: RED, 100 sources, Heterogeneous RTTs. Result: Optimized provider metrics. Tradeoff: High CoV, low avg. goodput.

Figure 2: Plot for coefficient of variation of per-flow goodput - Heterogeneous RTTs, long transfers
Covariance and Correlation: Measures of Dependence

- **Covariance:**
  \[
  \langle (x_i - \mu_i)(x_j - \mu_j) \rangle = \langle x_i x_j \rangle - \langle x_i \rangle \langle x_j \rangle,
  \]

  - For \(i = j\), covariance = variance!
  - Independence => covariance = 0 (not vice-versa!)

- **Correlation (coefficient)** is a normalized (or scaleless) form of covariance:
  \[
  \text{cor}(x_i, x_j) \equiv \frac{\text{cov}(x_i, x_j)}{\sigma_i \sigma_j},
  \]

  - Between \(-1\) and \(+1\).
    - Zero => no correlation (uncorrelated).
    - Note: uncorrelated **DOES NOT** mean independent!
Real Example: Randomized TCP vs TCP Reno

- **Goal:** to reduce synchronization between flows (that kills performance), we inject randomness in TCP
- **Metric:** Cov-Coefficient between flows (close to 0 is good)

### Table IV
Comparison of Covariance Coefficient of Congestion Window for Two Flows for TCP Reno, Paced and Randomized. (Value around 0 is Good.)

<table>
<thead>
<tr>
<th>Flow Pair</th>
<th>Reno</th>
<th>Paced</th>
<th>Randomized</th>
</tr>
</thead>
<tbody>
<tr>
<td>(1,2)</td>
<td>0.5183</td>
<td>-0.1454</td>
<td>0.2525</td>
</tr>
<tr>
<td>(1,3)</td>
<td>0.5416</td>
<td>-0.1537</td>
<td>0.1422</td>
</tr>
<tr>
<td>(1,4)</td>
<td>0.3492</td>
<td>-0.1833</td>
<td>0.1535</td>
</tr>
</tbody>
</table>

### Table V
Comparison of Covariance Coefficient of Congestion Windows for 3 Flows for TCP Reno, Paced and Randomized. (Value around 0 is Good.)
Recall: Why care about Probability?

- Humans like *determinism*
  - The real-world is unfortunately *random*!
  - **CANNOT** place ANY confidence on a single measurement

- We can be deterministic about *measures or functions of underlying randomness* …

- Functions of Randomness:
  - Probability of a random event or variable
  - Average (mean, median, mode), Distribution functions (pdf, cdf), joint pdfs/cdfs, conditional probability, confidence intervals,
Continuous-valued Random Variables

- So far we have focused on discrete(-valued) random variables, e.g. $X(s)$ must be an integer.
- Examples of discrete random variables: number of arrivals in one second, number of attempts until success.
- A continuous-valued random variable takes on a range of real values, e.g. $X(s)$ ranges from 0 to as $s$ varies.
- Examples of continuous(-valued) random variables: time when a particular arrival occurs, time between consecutive arrivals.
Continuous-valued Random Variables

- Thus, for a continuous random variable $X$, we can define its probability density function (pdf)

$$f_X(x) = F'_X(x) = \frac{dF_X(x)}{dx}$$

- Note that since $F_X(x)$ is non-decreasing in $x$ we have

$$f_X(x) \geq 0 \quad \text{for all } x.$$
Properties of Continuous Random Variables

- From the Fundamental Theorem of Calculus, we have
  \[ F_X(x) = \int_{-\infty}^{x} f_X(t) \, dt \]

- In particular, \[ \int_{-\infty}^{\infty} f_X(x) \, dx = F_X(\infty) = 1 \]

- More generally, \[ \int_{a}^{b} f_X(x) \, dx = F_X(b) - F_X(a) = P(a < X \leq b) \]
The expectation (average) of a continuous random variable $X$ is given by

$$E(X) = \int_{-\infty}^{\infty} xf_X(x)dx$$

Note that this is just the continuous equivalent of the discrete expectation

$$E(X) = \sum_{x=-\infty}^{\infty} xP_X(x)$$
Important (Discrete) Random Variable: Bernoulli

- The simplest possible measurement on an experiment:
  - **Success** \( (X = 1) \) or **failure** \( (X = 0) \).

- Usual notation:

\[
P_X(1) = P(X = 1) = p \\
P_X(0) = P(X = 0) = 1 - p
\]

- \( E(X) = \)
Important (discrete) Random Variables: Binomial

- Let $X$ = the number of success in $n$ independent Bernoulli experiments (or trials).
  
  $P(X=0) =$
  
  $P(X=1) =$
  
  $P(X=2) =$
  
  In general, $P(X = x) =$

Binomial Variables are useful for *proportions* (of successes. Failures) for a small number of repeated experiments. For larger number (n), under certain conditions (p is small), Poisson distribution is used.
Binomial Distribution Characteristics

**Mean**

\[ \mu = E(x) = np \]

**Standard Deviation**

\[ \sigma = \sqrt{np(1 - p)} \]
Binomial can be skewed or normal

Depends upon $p$ and $n$!
Important Random Variable: Poisson

- A Poisson random variable $X$ is defined by its PMF:

$$P(X = x) = \frac{\lambda^x e^{-\lambda}}{x!} \quad x = 0, 1, 2, ...$$

Where $\lambda > 0$ is a constant.

- Exercise: Show that

$$\sum_{x=0}^{\infty} P_X(x) = 1 \quad \text{and} \quad E(X) = \lambda$$

- Poisson random variables are good for counting frequency of occurrence: like the number of customers that arrive to a bank in one hour, or the number of packets that arrive to a router in one second.
Continuous Probability Density Function

- 1. Mathematical Formula
- 2. Shows All Values, $x$, & Frequencies, $f(x)$
  - $f(X)$ is **Not** Probability
- 3. Properties

\[ \int f(x) \, dx = 1 \]

All $X$ (Area Under Curve)

$f(x) \geq 0$, $a \leq x \leq b$
**Important Continuous Random Variable: Exponential**

- Used to represent time, e.g. until the next arrival
- Has PDF

\[
f_X(x) = \begin{cases} \lambda e^{-\lambda x} & \text{for } x \geq 0 \\ 0 & \text{for } x < 0 \end{cases}
\]

for some \( \lambda > 0 \)

- Properties:

\[
\int_{0}^{\infty} f_X(x) \, dx = 1 \quad \text{and} \quad E(X) = \frac{1}{\lambda}
\]

- Need to use integration by Parts!
Memoryless Property of the Exponential

- An exponential random variable $X$ has the property that “the future is independent of the past”, i.e. the fact that it hasn’t happened yet, tells us nothing about how much longer it will take.

- In math terms:

$$P(X > s + t \mid X > t) = P(X > s) \quad \text{for } s, t > 0$$

$$P(X > s) = e^{-\lambda s}$$
Recall: Why care about Probability?

- Humans like *determinism*
  - The real-world is unfortunately *random*!
  - **CANNOT** place ANY confidence on a **single** measurement

- We can be deterministic about *measures or functions of underlying randomness* …

- Functions of Randomness:
  - Probability of a random event or variable
  - Average (mean, median, mode), Distribution functions (pdf, cdf), joint pdfs/cdfs, conditional probability, confidence intervals,
Important Random Variables: **Normal**

![Graph of the normal distribution]

\[ p(y) = \text{constant} \frac{1}{\sigma} e^{-\frac{(y-\eta)^2}{2\sigma^2}} \]

**FIGURE 2.9.** The normal (Gaussian) distribution.
Normal Distribution: PDF & CDF

- **PDF:**

  
  \[
  P(x) \, dx = \frac{1}{\sigma \sqrt{2\pi}} e^{-\frac{(x-\mu)^2}{2\sigma^2}} \, dx,
  \]

- With the transformation: (a.k.a. unit normal deviate)

  \[
  z \equiv \frac{x - \mu}{\sigma},
  \]

- **z-normal-PDF:**

  
  \[
  P(x) \, dx = \frac{1}{\sqrt{2\pi}} e^{-\frac{z^2}{2}} \, dz.
  \]

Double exponential & symmetric

Can “standardize” to have \( \mu = 0, \sigma = 1 \).

Simplifies!
Nice things about Normal Distribution

1. ‘Bell-Shaped’ & Symmetrical
2. Mean, Median, Mode Are Equal
3. ‘Middle Spread’ Is 1.33 $\sigma$
4. Random Variable is continuous and has Infinite Range
Rapidly Dropping Tail Probability!

Why? Doubly exponential PDF (\(e^{-z^2}\) term…)
A.k.a: “Light tailed” (not heavy-tailed).
No skew or tail => don’t have two worry about > 2^{nd} order parameters (mean, variance)
Height & Spread of Gaussian Can Vary!

FIGURE 2.11. Normal distributions with different means and variances.
But this can be specified with just 2 parameters ($\mu$ & $\sigma$): $N(\mu, \sigma)$

Parsimonious! Only 2 parameters to estimate!
Normal Distribution
Probability Tables

Probability is area under curve!

\[ P(c \leq x \leq d) = \int_c^d f(x) \, dx \]
Why Standardize? Easy Table Lookup

Normal distributions differ by mean & standard deviation.
Each distribution would require its own table.

That’s an *infinite* number!

Shivkumar Kalyanaraman
Standardize the Normal Distribution

\[ Z = \frac{X - \mu}{\sigma} \]

Normal Distribution

Standardized Normal Distribution

One table!
## Obtaining the Probability

### Standardized Normal Probability Table (Portion)

<table>
<thead>
<tr>
<th>Z</th>
<th>.00</th>
<th>.01</th>
<th>.02</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.0</td>
<td>.0000</td>
<td>.0040</td>
<td>.0086</td>
</tr>
<tr>
<td>0.1</td>
<td>.0398</td>
<td>.0438</td>
<td>.0478</td>
</tr>
<tr>
<td>0.2</td>
<td>.0793</td>
<td>.0832</td>
<td>.0871</td>
</tr>
<tr>
<td>0.3</td>
<td>.1179</td>
<td>.1217</td>
<td>.1255</td>
</tr>
</tbody>
</table>

- **Shaded area exaggerated**

- **Probabilities**

- **Shivkumar Kalyanaraman**

- **Rensselaer Polytechnic Institute**
Example

$P(3.8 \leq X \leq 5)$

$$Z = \frac{X - \mu}{\sigma} = \frac{3.8 - 5}{10} = -.12$$
$P(2.9 \leq X \leq 7.1)$

$Z = \frac{X - \mu}{\sigma} = \frac{2.9 - 5}{10} = -0.21$

$Z = \frac{X - \mu}{\sigma} = \frac{7.1 - 5}{10} = 0.21$

Standardized Normal Distribution

Normal Distribution

Shaded area exaggerated
Example

\[ P(X \geq 8) \]

\[
Z = \frac{X - \mu}{\sigma} = \frac{8 - 5}{10} = .30
\]

Normal Distribution

Shaded area exaggerated

Standardized Normal Distribution

\[
Z = \frac{X - \mu}{\sigma} = \frac{8 - 5}{10} = .30
\]
Finding Z Values for Known Probabilities

What is Z given $P(Z) = .1217$?

Standardized Normal Probability Table (Portion)

<table>
<thead>
<tr>
<th>Z</th>
<th>.00</th>
<th>.01</th>
<th>.2</th>
</tr>
</thead>
<tbody>
<tr>
<td>.0</td>
<td>.000</td>
<td>.004</td>
<td>.008</td>
</tr>
<tr>
<td>.1</td>
<td>.0398</td>
<td>.0438</td>
<td>.0478</td>
</tr>
<tr>
<td>.2</td>
<td>.0793</td>
<td>.0832</td>
<td>.0871</td>
</tr>
<tr>
<td>.3</td>
<td>.1179</td>
<td>.1217</td>
<td>.1255</td>
</tr>
</tbody>
</table>
Finding X Values for Known Probabilities

\[ X = \mu + Z \cdot \sigma = 5 + (0.31)(10) = 8.1 \]
Why Care? Ans: Statistics

- Sampling: Take a sample
  - How to sample?
  - Eg: Infer country's GDP by sampling a small fraction of data. Important to sample right.
  - Randomness gives you power in sampling

- Inferring properties (eg: means, variance) of the population (a.k.a “parameters”)
  - … from the sample properties (a.k.a. “statistics”)
  - Depend upon theorems like Central limit theorem and properties of normal distribution

- The distribution of sample mean (or sample stdev etc) are called "sampling distributions"

- Powerful idea: sampling distribution of the sample mean is a normal distribution (under mild restrictions): **Central Limit Theorem (CLT)**!
Estimation Process

Population

Mean, $\mu$, is unknown

Random Sample

Mean $\overline{X} = 50$

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

1. Involves
   - Estimation
   - Hypothesis Testing

2. Purpose
   - Make Decisions About Population Characteristics
Inference Process

Estimates & tests

Sample statistic ($X$)

Population

Sample

Shivkumar Kalyanaraman
Key Terms

- 1. Population (Universe)
  - All Items of Interest

- 2. Sample
  - Portion of Population

- 3. Parameter
  - Summary Measure about Population

- 4. Statistic
  - Summary Measure about Sample
Normal Distribution: Why?

Uniform distribution looks nothing like bell shaped (gaussian)!
Large spread (σ)!

CENTRAL LIMIT TENDENCY!

Sum of r.v.s from a uniform distribution after very few samples looks remarkably normal
BONUS: it has decreasing σ!
Gaussian/Normal Distribution: Why?

Goal: estimate mean of Uniform distribution (distributed between 0 & 6)

Sample mean = (sum of samples)/N
Since sum is distributed normally, so is the sample mean (a.k.a sampling distribution is NORMAL), w/ decreasing dispersion!
Numerical Example: Population Characteristics

Summary Measures

\[ \mu = \frac{\sum_{i=1}^{N} X_i}{N} = 2.5 \]

\[ \sigma = \sqrt{\frac{\sum_{i=1}^{N} (X_i - \mu)^2}{N}} = 1.12 \]

Population Distribution

1 2 3 4
Summary Measures of All Sample Means

\[ \mu_{\bar{x}} = \frac{1}{N} \sum_{i=1}^{N} \bar{X}_i = \frac{1.0 + 1.5 + \cdots + 4.0}{16} = 2.5 \]

\[ \sigma_{\bar{x}} = \sqrt{\frac{1}{N} \sum_{i=1}^{N} (\bar{X}_i - \mu_{\bar{x}})^2} \]

\[ = \sqrt{\frac{(1.0-2.5)^2 + (1.5-2.5)^2 + \cdots + (4.0-2.5)^2}{16}} = 0.79 \]
Comparison

Population

<table>
<thead>
<tr>
<th>X</th>
<th>P(X)</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>0.3</td>
</tr>
<tr>
<td>2</td>
<td>0.2</td>
</tr>
<tr>
<td>3</td>
<td>0.1</td>
</tr>
<tr>
<td>4</td>
<td>0.1</td>
</tr>
</tbody>
</table>

\[ \mu = 2.5 \]
\[ \sigma = 1.12 \]

Sampling Distribution

<table>
<thead>
<tr>
<th>X</th>
<th>P(\bar{X})</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>0.0</td>
</tr>
<tr>
<td>1.5</td>
<td>0.1</td>
</tr>
<tr>
<td>2</td>
<td>0.2</td>
</tr>
<tr>
<td>2.5</td>
<td>0.3</td>
</tr>
<tr>
<td>3</td>
<td>0.2</td>
</tr>
<tr>
<td>3.5</td>
<td>0.1</td>
</tr>
<tr>
<td>4</td>
<td>0.0</td>
</tr>
</tbody>
</table>

\[ \mu_{\bar{X}} = 2.5 \]
\[ \sigma_{\bar{X}} = 0.79 \]
Recall: Rapidly Dropping Tail Probability!

Sample mean is a gaussian r.v., with $x = \mu$ & $s = \sigma/(n)^{0.5}$
⇒ With larger number of samples, avg of sample means is an excellent estimate of true mean.
⇒ If (original) $\sigma$ is known, invalid mean estimates can be rejected with HIGH confidence!

Shivkumar Kalyanaraman
Many Experiments: Coin Tosses

Sample Mean = \frac{Total\ Heads}{Number\ of\ Tosses}

Variance reduces over trials by factor $\sqrt{n}$

Corollary: more experiments ($n$) is good, but not great (why? $\sqrt{n}$ doesn’t grow fast)
Standard Error of Mean

1. Standard Deviation of All Possible Sample Means, $\bar{X}$
   - Measures Scatter in All Sample Means, $\bar{X}$
2. Less Than Pop. Standard Deviation
3. Formula (Sampling With Replacement)

$$\sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}}$$
Properties of Sampling Distribution of Mean

1. **Unbiasedness**
   - Mean of Sampling Distribution **Equals** Population Mean

2. **Efficiency**
   - Sample Mean **Comes Closer to** Population Mean Than Any Other Unbiased Estimator

3. **Consistency**
   - As Sample Size Increases, **Variation** of Sample Mean from Population Mean **Decreases**
Unbiasedness

Unbiased

Biased

P(\bar{X})

\mu

Unbiased

Biased

A

C

Shivkumar Kalyanaraman

Rensselaer Polytechnic Institute
Efficiency: Sample mean vs Sample median

- $P(\bar{X})$
- Sampling distribution of mean
- Sampling distribution of median
Consistency: Sample Size

\[ P(\bar{X}) \]

- **Larger sample size**
- **Smaller sample size**

Diagram illustrating the probability distribution of sample means \( \bar{X} \) with two curves:
- Curve A: Smaller sample size
- Curve B: Larger sample size
Sampling from Non-Normal Populations

- **Central Tendency**
  \[ \mu_{\bar{X}} = \mu \]

- **Dispersion**
  \[ \sigma_{\bar{X}} = \frac{\sigma}{\sqrt{n}} \]

- **Sampling with replacement**

**Population Distribution**

\[ \sigma = 10 \]

\[ \mu = 50 \]

**Sampling Distribution**

\[ n = 4 \]
\[ \sigma_{\bar{X}} = 5 \]

\[ n = 30 \]
\[ \sigma_{\bar{X}} = 1.8 \]
As sample size gets large enough \((n \geq 30)\) ...
Central Limit Theorem (CLT)

As sample size gets large enough (n ≥ 30) ...

\[ \sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}} \]

The sampling distribution becomes almost normal.

\[ \mu_{\bar{x}} = \mu \]
Aside: Caveat about CLT

- Central limit theorem works if original distribution are not heavy tailed
- Moments converge to limits
- Trouble with aggregates of “heavy tailed” distribution samples

- Non-classical version of CLT for such cases...
  - Sum converges to stable Levy-noise (heavy tailed and long-range dependent auto-correlations)
Other interesting points reg. Gaussian

- Uncorrelated r.vs. + gaussian => INDEPENDENT!
  - Important in random processes (i.e. sequences of random variables)

- Random variables that are independent, and have exactly the same distribution are called IID (independent & identically distributed)

- IID and normal with zero mean and variance $\sigma^2$ => IIDN($0, \sigma^2$)
Recall: Why care about Probability?

- Humans like determinism
  - The real-world is unfortunately random!
  - CANNOT place ANY confidence on a single measurement

- We can be deterministic about measures or functions of underlying randomness …

- Functions of Randomness:
  - Probability of a random event or variable
  - Average (mean, median, mode), Distribution functions (pdf, cdf), joint pdfs/cdfs, conditional probability, confidence intervals,

- Goal: Build “probabilistic” models of reality
  - Constraint: minimum # experiments
  - Infer to get a model (i.e. maximum information)
  - Statistics: how to infer models about reality (“population”) given a SMALL set of expt results (“sample”)
Point Estimation … vs..

1. Provides Single Value
   - Based on Observations from 1 Sample
2. Gives No Information about How Close Value Is to the Unknown Population Parameter
3. Example: Sample Mean $\bar{X} = 3$ Is Point Estimate of Unknown Population Mean
... vs ... Interval Estimation

1. Provides **Range** of Values
   - Based on Observations from 1 Sample

2. Gives Information about Closeness to Unknown Population Parameter
   - Stated in terms of Probability
     - Knowing Exact Closeness Requires Knowing Unknown Population Parameter

3. Example: Unknown Population Mean Lies Between 50 & 70 with 95% Confidence
Key Elements of Interval Estimation

A probability that the population parameter falls somewhere within the interval.
Confidence Interval

- Probability that a measurement will fall within a closed interval \([a,b]\): (mathworld definition…)

\[
CI(a, b) \equiv \int_b^a P(x) \, dx, = (1 - \alpha)
\]

- Jain: the interval \([a,b]\) = “confidence interval”;
  - the probability level, \(100(1-\alpha)\) = “confidence level”;
  - \(\alpha\) = “significance level”

- Sampling distribution for means leads to high confidence levels, i.e. small confidence intervals
Confidence Limits for Population Mean

Parameter = Statistic ± Error

(1) \( \mu = \bar{X} \pm Error \)

(2) \( Error = \bar{X} - \mu \) or \( \bar{X} + \mu \)

(3) \( Z = \frac{\bar{X} - \mu}{\sigma_{\bar{X}}} = \frac{Error}{\sigma_{\bar{X}}} \)

(4) \( Error = Z \sigma_{\bar{X}} \)

(5) \( \mu = \bar{X} \pm Z \sigma_{\bar{X}} \)
Many Samples Have Same Interval

\[ \bar{X} = \mu \pm Z\sigma_{\bar{X}} \]

\[ \mu - 1.65\sigma_{\bar{X}} \quad \mu \quad \mu + 1.65\sigma_{\bar{X}} \]

\[ \mu - 1.96\sigma_{\bar{X}} \quad \mu \quad \mu + 1.96\sigma_{\bar{X}} \]

90% Samples

95% Samples

Shivkumar Kalyanaraman
Many Samples Have Same Interval

\[ \overline{X} = \mu \pm Z \sigma_{\overline{X}} \]

- 90% Samples: \( \mu - 1.65 \sigma_{\overline{X}}, \mu + 1.65 \sigma_{\overline{X}} \)
- 95% Samples: \( \mu - 1.96 \sigma_{\overline{X}}, \mu + 1.96 \sigma_{\overline{X}} \)
- 99% Samples: \( \mu - 2.58 \sigma_{\overline{X}}, \mu + 2.58 \sigma_{\overline{X}} \)
Confidence Level

1. Probability that the Unknown Population Parameter Falls Within Interval

2. Denoted \((1 - \alpha)\) %
   - \(\alpha\) is probability that parameter is Not within interval

3. Typical Values Are 99%, 95%, 90%
Intervals & Confidence Level

Sampling Distribution of Mean

\( \bar{X} - Z\sigma_{\bar{X}} \) to \( \bar{X} + Z\sigma_{\bar{X}} \)

Intervals extend from \( \bar{X} - Z\sigma_{\bar{X}} \) to \( \bar{X} + Z\sigma_{\bar{X}} \)

(1 - \( \alpha \)) % of intervals contain \( \mu \).
\( \alpha \) % do not.

Large number of intervals

\( \mu_{\bar{X}} = \mu \)
Factors Affecting Interval Width

1. Data Dispersion
   - Measured by $\sigma$

2. Sample Size
   - $\sigma_{\bar{X}} = \sigma / \sqrt{n}$

3. Level of Confidence
   - (1 - $\alpha$)
   - Affects $Z$

Intervals Extend from $\bar{X} - Z\sigma_{\bar{X}}$ to $\bar{X} + Z\sigma_{\bar{X}}$
Meaning of Confidence Interval

![Diagram showing the concept of confidence interval]

**FIGURE 13.1 Meaning of a confidence interval.**

- Total 'Yes' ≥ 100(1-α)
- Total 'No' ≤ 100α
Intervals & Confidence Level

Sampling Distribution of Mean

Intervals extend from $\bar{X} - Z\sigma_{\bar{X}}$ to $\bar{X} + Z\sigma_{\bar{X}}$

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

Large number of intervals

$(1 - \alpha)\%$ of intervals contain $\mu$.

$\alpha\%$ do not.
Statistical Inference: Is $\mu_A = \mu_B$?

- Note: sample mean $y_A$ is not $\mu_A$, but its estimate!
- Is this difference statistically significant?
- Is the null hypothesis $y_A = y_B$ false?

**TABLE 2.1. Yield data from an industrial experiment (plant trial)**

<table>
<thead>
<tr>
<th>time order</th>
<th>method</th>
<th>yield</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>A</td>
<td>89.7</td>
</tr>
<tr>
<td>2</td>
<td>A</td>
<td>81.4</td>
</tr>
<tr>
<td>3</td>
<td>A</td>
<td>84.5</td>
</tr>
<tr>
<td>4</td>
<td>A</td>
<td>84.8</td>
</tr>
<tr>
<td>5</td>
<td>A</td>
<td>87.3</td>
</tr>
<tr>
<td>6</td>
<td>A</td>
<td>79.7</td>
</tr>
<tr>
<td>7</td>
<td>A</td>
<td>85.1</td>
</tr>
<tr>
<td>8</td>
<td>A</td>
<td>81.7</td>
</tr>
<tr>
<td>9</td>
<td>A</td>
<td>83.7</td>
</tr>
<tr>
<td>10</td>
<td>A</td>
<td>84.5</td>
</tr>
<tr>
<td>11</td>
<td>B</td>
<td>84.7</td>
</tr>
<tr>
<td>12</td>
<td>B</td>
<td>86.1</td>
</tr>
<tr>
<td>13</td>
<td>B</td>
<td>83.2</td>
</tr>
<tr>
<td>14</td>
<td>B</td>
<td>91.9</td>
</tr>
<tr>
<td>15</td>
<td>B</td>
<td>86.3</td>
</tr>
<tr>
<td>16</td>
<td>B</td>
<td>79.3</td>
</tr>
<tr>
<td>17</td>
<td>B</td>
<td>82.6</td>
</tr>
<tr>
<td>18</td>
<td>B</td>
<td>89.1</td>
</tr>
<tr>
<td>19</td>
<td>B</td>
<td>83.7</td>
</tr>
<tr>
<td>20</td>
<td>B</td>
<td>88.5</td>
</tr>
</tbody>
</table>

$\bar{y}_A = 84.24$, $\bar{y}_B = 85.54$

$\bar{y}_B - \bar{y}_A = 1.30$
Step 1: Plot the samples

**FIGURE 2.1.** Yield values plotted in time order for comparative experiment.
Compare to (external) reference distribution (if available)

Since 1.30 is at the tail of the reference distribution, the difference between means is **NOT statistically significant**!

**FIGURE 2.6.** Reference distribution of 191 differences between averages of adjacent sets of 10 observations.
Under random sampling assumption, and the null hypothesis of $y_A = y_B$, we can view the 20 samples from a common population & construct a reference distributions from the samples itself!
$t$-distribution: Create a Reference Distribution from the Samples Itself!

**FIGURE 2.13.** The $t$ distribution for $\nu = 1, 9,$ and $\infty$. 
**t-distribution**

**FIGURE 13.2** The ratio \((\bar{x} - \mu) / (s/\sqrt{n})\) for samples from normal populations follows a \(t(n - 1)\) distribution.
Student’s t Distribution

- Bell-Shaped
- Symmetric
- ‘Fatter’ Tails

\( t (df = 5) \): wider

\( t (df = 13) \)

Standard Normal
### Student’s t Table

<table>
<thead>
<tr>
<th>v</th>
<th>$t_{.10}$</th>
<th>$t_{.05}$</th>
<th>$t_{.025}$</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>3.078</td>
<td>6.314</td>
<td>12.706</td>
</tr>
<tr>
<td>2</td>
<td>1.886</td>
<td>2.920</td>
<td>4.303</td>
</tr>
<tr>
<td>3</td>
<td>1.638</td>
<td>2.353</td>
<td>3.182</td>
</tr>
</tbody>
</table>

Assume:
- $n = 3$
- $df = n - 1 = 2$
- $\alpha = .10$
- $\alpha/2 = .05$
Assume:
- \( n = 3 \)
- \( df = n - 1 = 2 \)
- \( \alpha = .10 \)
- \( \alpha/2 = .05 \)

<table>
<thead>
<tr>
<th>( v )</th>
<th>( t_{.10} )</th>
<th>( t_{.05} )</th>
<th>( t_{.025} )</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>3.078</td>
<td>6.314</td>
<td>12.706</td>
</tr>
<tr>
<td>2</td>
<td>1.886</td>
<td>2.920</td>
<td>4.303</td>
</tr>
<tr>
<td>3</td>
<td>1.638</td>
<td>2.353</td>
<td>3.182</td>
</tr>
</tbody>
</table>
### Student’s t Table

<table>
<thead>
<tr>
<th>ν</th>
<th>( t_{.10} )</th>
<th>( t_{.05} )</th>
<th>( t_{.025} )</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>3.078</td>
<td>6.314</td>
<td>12.706</td>
</tr>
<tr>
<td>2</td>
<td>1.886</td>
<td><strong>2.920</strong></td>
<td>4.303</td>
</tr>
<tr>
<td>3</td>
<td>1.638</td>
<td>2.353</td>
<td><strong>3.182</strong></td>
</tr>
</tbody>
</table>

**Assume:**
- \( n = 3 \)
- \( df = n - 1 = 2 \)
- \( \alpha = .10 \)
- \( \alpha/2 = .05 \)

![t Value Diagram](image)
Degrees of Freedom (df)

1. Number of Observations that Are Free to Vary After Sample Statistic Has Been Calculated

2. Example
   - Sum of 3 Numbers Is 6
     - $X_1 = 1$ (or Any Number)
     - $X_2 = 2$ (or Any Number)
     - $X_3 = 3$ (Cannot Vary)
   - Sum = 6

   degrees of freedom = $n - 1$
   = 3 - 1
   = 2
Statistical Significance with Various Inference Techniques

- Normal population assumption not required
- Random sampling assumption required
- Std.dev. estimated from samples itself!

*t-distribution an approx. for gaussian!*

---

Shivkumar Kalyanaraman
Rensselaer Polytechnic Institute
Normal, $\chi^2$ & t-distributions: Useful for Statistical Inference

Note: not all sampling distributions are normal

FIGURE 3.4. Random sampling from a normal distribution to produce sampling distributions of $\bar{y}$, $s^2$, and $t$. 
Relationship between Confidence Intervals and Comparisons of Means

**FIGURE 13.3** Testing for a zero mean.

- (a) CI includes zero ⇒ mean is zero
- (b) CI does not include zero ⇒ mean is nonzero

**FIGURE 13.4** Comparing two alternatives.

- (a) CIs do not overlap ⇒ A is higher than B
- (b) CIs overlap and mean of one is in the CI of the other ⇒ alternatives are not different
- (c) CIs overlap but mean of any one is not in the CI of the other ⇒ need to do the t-test
Application:
Internet Measurement and Modeling: 
A snapshot
Internet Measurement/Modeling: Why is it Hard?
There is No Such Thing as “Typical”

- Heterogeneity in:
  - Traffic mix
  - Range of network capabilities
    - Bottleneck bandwidth (orders of magnitude)
    - Round-trip time (orders of magnitude)
  - Dynamic range of network conditions
    - Congestion / degree of multiplexing / available bandwidth
    - Proportion of traffic that is adaptive/rigid/attack
  - Immense size & growth
  - Rare events will occur
  - New applications explode on the scene
There is No Such Thing as “Typical”, con’t

- New applications **explode** on the scene
  - Not just the Web, but: Mbone, Napster, KaZaA etc., IM
- Event **robust statistics** fail.
  - E.g., **median** size of FTP data transfer at LBL
    - Oct. 1992: 4.5 KB (60,000 samples)
    - Mar. 1993: 2.1 KB
    - Mar. 1998: 10.9 KB
    - Dec. 1998: 5.6 KB
    - Dec. 1999: 10.9 KB
    - Jun. 2000: 62 KB
    - Nov. 2000: 10 KB
- **Danger**: if you misassume that something is “typical”, **nothing** tells you that you are wrong!
The Search for Invariants

- In the face of such diversity, identifying things that don’t change (aka “invariants”) has immense utility.

- Some Internet traffic invariants:
  - Daily and weekly patterns
  - Self-similarity on time scales of 100s of msec and above
  - Heavy tails
    - both in activity periods and elsewhere, e.g., topology
  - Poisson user session arrivals
  - Log-normal sizes (excluding tails)
  - Keystrokes have a Pareto distribution
Web traffic …
X = Htailed HTTP Requests/responses

… is streamed onto the Internet …
(TCP-type transport)

… creating “bursty-looking” link traffic
Y = “colored” noise {self-similar}
So, why does this burstiness matter?  
**Case: Failure of Poisson Modeling**

- Long-established framework: *Poisson modeling*
- Central idea: network events (packet arrivals, connection arrivals) are well-modeled as *independent*
- In simplest form, there’s just a rate parameter, $\lambda$
- It then follows that the time between “calls” (events) is exponentially distributed, # of calls $\sim$ Poisson

- Implications or Properties (if assumptions correct):
  - Aggregated traffic will smooth out quickly
  - Correlations are fleeting, bursts are limited
Burstiness: Theory vs. Measurement

- For Internet traffic, Poisson models have fundamental problem: they greatly underestimate burstiness.

- Consider an arrival process: $X_k$ gives the number of packets arriving during the $k^{th}$ interval of length $T$.
  - Take a 1-hour trace of Internet traffic (1995).
  - Generate (batch) Poisson arrivals with the same mean and variance.
  - I.e. “fit” a Poisson (batch) process to the trace...
Poisson Packet Arrivals

Seconds

Packets

0 1 2 3 4 5 6
Poisson Packet Arrivals

Seconds

Packets

Previous Region

×10
Poisson Packet Arrivals

Seconds

Packets

×600

0 1000 2000 3000

40000

60000

80000

2000

3000

Rensselaer Polytechnic Institute

Shivkumar Kalyanaraman
Amen!