Probability Distributions

Ananda V. Mysore

SJSU
Measurement Error and Statistical Analysis

- Measurement error can be categorized into two major types, systematic and random.

- For which of these types is statistical analysis more useful, and why?
Probability, Distributions, and Random Variables

- **Probability** $P$ is a quantitative expression for the likelihood of occurrence of some event.
  - The probability of a particular event is the number of times the event occurs divided by the total number of trials.

- A probability **distribution** is a mathematical model that expresses how the probability of an event varies according to the value of some variable.

- A **random** variable is one for which its numerical value follows a probability distribution, while still being subject to variability.
  - Random variables may be categorized as continuous (e.g. temperature) or discrete (e.g. outcome of rolling dice).
Common Probability Laws

- For any event $A$, $0 \leq P\{A\} \leq 1$
- If $\bar{A}$ is the (exclusive) complement of $A$, $P\{A\} + P\{\bar{A}\} = 1$
- If $A$ and $B$ are independent events, $P\{A \text{ and } B\} = P\{A\}P\{B\}$
- If $A$ and $B$ are mutually exclusive, $P\{A \text{ or } B\} = P\{A\} + P\{B\}$
- If $A$ and $B$ are not mutually exclusive it is necessary to subtract the joint probability, and $P\{A \text{ or } B\} = P\{A\} + P\{B\} - P\{A\}P\{B\}$
- If $B$ has already known to have occurred, the conditional probability that $A$ will occur is $P\{A \mid B\} = P\{A \text{ and } B\} / P\{B\}$
Measurements, Samples, and Populations

- A **measurement** for a random variable $x$ produces a numerical value for that variable.

- A **sample** is a subset of a population for which something is to be quantified (i.e. measured).
  - In statistical analysis, a sample usually implies more than a single measurement (e.g. $n = 5$ measurements taken from a population of $N = 1,000$).

- The **population** comprises the entire collection of possible measurements “whose properties are under consideration and about which some generalizations are to be made.”

Quoted text from *Introduction to Engineering Experimentation* by A. J. Wheeler and A. R. Ganji © 2004 Pearson Education, Inc., Upper Saddle River, NJ. All rights reserved.
Histogram and Cumulative Frequency Plot

- Individual values of a variable are sorted into intervals, then stacked to count the number of observations that fall into each interval.

- Intervals of constant span are almost always preferred.

- A good default for choosing the number of bins is the square root of the number of total observations \( n \).

- Histograms may be used for both continuous (e.g. layer thickness) and discrete variables (e.g. number of defects).

- The cumulative frequency plot is a related display that shows what fraction of the observations fall under a given value.
Statistical Inference

- Statistical inference draws conclusions about a population based on a sample selected from that population.

- A random sample of size $n$ is a subset of the population, which has size $N$.
  
  ▪ (Technically, random samples from finite populations must be drawn with replacement to ensure equal selection probability.)
Central Tendency & Variability in Samples

- **Sample Mean:**
  \[ \bar{x} = \frac{\sum_{i=1}^{n} x_i}{n} \]

- **Deviation (from mean):**
  \[ d_i = x_i - \bar{x} \]

- **Sample Range:**
  \[ R = x_{\text{max}} - x_{\text{min}} \]

- **Sample Variance:**
  \[ s^2 = \frac{\sum_{i=1}^{n} (x_i - \bar{x})^2}{n - 1} \]

- **Sample Standard Deviation:**
  \[ s = \sqrt{\frac{\sum_{i=1}^{n} (x_i - \bar{x})^2}{n - 1}} \]
Central Tendency & Variability in Populations

**Figure 2-11** The mean of a distribution.

**Figure 2-12** Two probability distributions with different means.

**Figure 2-13** Two probability distributions with the same mean but different standard deviations.
Discrete and Continuous Probability Distributions

- In a discrete distribution (left), the probability $P$ that a random variable $x$ has the specific value $x_i$ has a discrete value.

- In a continuous distribution (right), the probability $P$ of occurrence for a random variable $x$ is expressed in terms of an interval.

\[
P\{x = x_i\} = p\{x_i\} \quad \quad P\{a \leq x \leq b\} = \int_a^b f\{x\}dx
\]
Mean, Variance, and Probability Distributions

- **Continuous Distribution**
  - Mean:
    \[ \mu = \int_{-\infty}^{+\infty} xf\{x\}dx \]
  - Variance:
    \[ \sigma^2 = \int_{-\infty}^{+\infty} (x - \mu)^2 f\{x\}dx \]

- **Discrete Distribution**
  - Mean:
    \[ \mu = \sum_{i=1}^{\infty} x_i p\{x_i\} \]
  - Variance:
    \[ \sigma^2 = \sum_{i=1}^{\infty} (x_i - \mu)^2 p\{x_i\} \]

Standard Deviation: \[ \sigma = \sqrt{\sigma^2} \]
Binomial Distribution

- Outcomes are discrete success or failure.
- Probability of success $p$.
- Number of successes $x$.
- Number of independent trials $n$.

- An example scenario would be to find probability of encountering $x$ number of non-conforming items in a random sample of $n$ items.

**Definition**

The binomial distribution with parameters $n \geq 0$ and $0 < p < 1$ is

$$p(x) = \binom{n}{x} p^x (1-p)^{n-x} \quad x = 0, 1, \ldots, n \quad (2-11)$$

The mean and variance of the binomial distribution are

$$\mu = np \quad (2-12)$$

and

$$\sigma^2 = np(1-p) \quad (2-13)$$

$$\binom{n}{x} = \frac{n!}{x!(n-x)!}$$

Images from Introduction to Statistical Quality Control, 5th Ed. by Douglas C. Montgomery, ©2005 John Wiley & Sons, Inc.
Poisson Distribution

- Number of defects per unit (or unit area, unit volume, etc.).
- Parameter $\lambda$ determines the shape of the distribution.
- Gives the probability that $x$ has a particular “defect” count.
- Useful in cases for example in which $\mu$ is known and probability that $x \leq b$ is of interest:

$$P\{x \leq b\} = \sum_{x=0}^{b} \frac{e^{-\mu} \mu^x}{x!}$$

### Definition

The Poisson distribution is

$$p(x) = \frac{e^{-\lambda} \lambda^x}{x!} \quad x = 0, 1, \ldots$$

(2-15)

where the parameter $\lambda > 0$. The mean and variance of the Poisson distribution are

$$\mu = \lambda$$

(2-16)

and

$$\sigma^2 = \lambda$$

(2-17)
Normal Distribution

Definition

The normal distribution is

\[ f(x) = \frac{1}{\sigma \sqrt{2\pi}} e^{-\frac{1}{2} \left( \frac{x-\mu}{\sigma} \right)^2} \quad -\infty < x < \infty \]  

(2-21)

The mean of the normal distribution is \( \mu \) \((-\infty < \mu < \infty)\) and the variance is \( \sigma^2 > 0 \).

- Relevant for “normal” random variables \( x \).
- Most common and arguably most important distribution in applied statistics.
- Abbreviated notation \( N(\mu, \sigma^2) \).
Standard Normal Distribution

- A standard normal distribution converts an \(N(\mu, \sigma^2)\) random variable to an \(N(0,1)\) random variable. Why is this useful?

\[
z = \frac{x - \mu}{\sigma}
\]

- The probability that the normal random variable \(x\) is less that or equal to a threshold \(a\) can be determined from the solution to the following integral expression.

\[
P\{x \leq a\} = \int_{-\infty}^{a} \frac{1}{\sigma \sqrt{2\pi}} e^{-\frac{1}{2} \left(\frac{x-\mu}{\sigma}\right)^2} \, dx = \int_{-\infty}^{a} \frac{1}{\sqrt{2\pi}} e^{-\frac{z^2}{2}} \, dz
\]

- Results are tabulated (thankfully) based on a single input, \(z\), in what is called a cumulative standard normal distribution table.

- Also: \(P\{x \geq a\} = 1 - P\{x \leq a\}\)
Central Limit Theorem

Definition: The Central Limit Theorem

If \( x_1, x_2, \ldots, x_n \) are independent random variables with mean \( \mu_i \) and variance \( \sigma_i^2 \), and if \( y = x_1 + x_2 + \ldots + x_n \), then the distribution of

\[
\frac{y - \sum_{i=1}^{n} \mu_i}{\sqrt{\sum_{i=1}^{n} \sigma_i^2}}
\]

approaches the \( N(0, 1) \) distribution as \( n \) approaches infinity.

- The sum \( y \) of \( n \) independent random variables \( x \) has a distribution that is approximately normal, regardless of the distribution of each individual random variable \( x_i \) in the sum.
- The approximation improves as \( n \) increases.
- In many circumstances this theorem is often used to justify the assumption of a normal distribution regardless of underlying distribution.
# Tabulated Standard Normal Distribution Values

<table>
<thead>
<tr>
<th>z</th>
<th>0.00</th>
<th>0.01</th>
<th>0.02</th>
<th>0.03</th>
<th>0.04</th>
<th>0.05</th>
<th>0.06</th>
<th>0.07</th>
<th>0.08</th>
<th>0.09</th>
<th>z</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.0</td>
<td>0.50000</td>
<td>0.50399</td>
<td>0.50798</td>
<td>0.51197</td>
<td>0.51595</td>
<td>0.51994</td>
<td>0.52392</td>
<td>0.52790</td>
<td>0.53188</td>
<td>0.53586</td>
<td>0.0</td>
</tr>
<tr>
<td>0.1</td>
<td>0.53983</td>
<td>0.54379</td>
<td>0.54776</td>
<td>0.55172</td>
<td>0.55567</td>
<td>0.55962</td>
<td>0.56356</td>
<td>0.56749</td>
<td>0.57142</td>
<td>0.57534</td>
<td>0.1</td>
</tr>
<tr>
<td>0.2</td>
<td>0.57926</td>
<td>0.58317</td>
<td>0.58706</td>
<td>0.59095</td>
<td>0.59483</td>
<td>0.59871</td>
<td>0.60257</td>
<td>0.60642</td>
<td>0.61026</td>
<td>0.61409</td>
<td>0.2</td>
</tr>
<tr>
<td>0.3</td>
<td>0.61791</td>
<td>0.62172</td>
<td>0.62551</td>
<td>0.62930</td>
<td>0.63307</td>
<td>0.63683</td>
<td>0.64058</td>
<td>0.64431</td>
<td>0.64803</td>
<td>0.65173</td>
<td>0.3</td>
</tr>
<tr>
<td>0.4</td>
<td>0.65542</td>
<td>0.65910</td>
<td>0.66276</td>
<td>0.66640</td>
<td>0.67003</td>
<td>0.67364</td>
<td>0.67724</td>
<td>0.68082</td>
<td>0.68438</td>
<td>0.68793</td>
<td>0.4</td>
</tr>
<tr>
<td>0.5</td>
<td>0.69146</td>
<td>0.69497</td>
<td>0.69847</td>
<td>0.70194</td>
<td>0.70540</td>
<td>0.70884</td>
<td>0.71226</td>
<td>0.71566</td>
<td>0.71904</td>
<td>0.72240</td>
<td>0.5</td>
</tr>
<tr>
<td>0.6</td>
<td>0.72575</td>
<td>0.72907</td>
<td>0.73237</td>
<td>0.73565</td>
<td>0.73891</td>
<td>0.74215</td>
<td>0.74537</td>
<td>0.74857</td>
<td>0.75175</td>
<td>0.75490</td>
<td>0.6</td>
</tr>
<tr>
<td>0.7</td>
<td>0.75030</td>
<td>0.76115</td>
<td>0.76424</td>
<td>0.76730</td>
<td>0.77035</td>
<td>0.77337</td>
<td>0.77637</td>
<td>0.77935</td>
<td>0.78230</td>
<td>0.78523</td>
<td>0.7</td>
</tr>
<tr>
<td>0.8</td>
<td>0.78814</td>
<td>0.79103</td>
<td>0.79389</td>
<td>0.79673</td>
<td>0.79954</td>
<td>0.80234</td>
<td>0.80510</td>
<td>0.80785</td>
<td>0.81057</td>
<td>0.81327</td>
<td>0.8</td>
</tr>
<tr>
<td>0.9</td>
<td>0.81594</td>
<td>0.81859</td>
<td>0.82121</td>
<td>0.82381</td>
<td>0.82639</td>
<td>0.82894</td>
<td>0.83147</td>
<td>0.83397</td>
<td>0.83646</td>
<td>0.83891</td>
<td>0.9</td>
</tr>
<tr>
<td>1.0</td>
<td>0.84134</td>
<td>0.84375</td>
<td>0.84613</td>
<td>0.84849</td>
<td>0.85083</td>
<td>0.85314</td>
<td>0.85543</td>
<td>0.85769</td>
<td>0.85993</td>
<td>0.86214</td>
<td>1.0</td>
</tr>
<tr>
<td>1.1</td>
<td>0.86433</td>
<td>0.86650</td>
<td>0.86864</td>
<td>0.87076</td>
<td>0.87285</td>
<td>0.87493</td>
<td>0.87697</td>
<td>0.87900</td>
<td>0.88100</td>
<td>0.88297</td>
<td>1.1</td>
</tr>
<tr>
<td>1.2</td>
<td>0.88493</td>
<td>0.88686</td>
<td>0.88877</td>
<td>0.89065</td>
<td>0.89251</td>
<td>0.89435</td>
<td>0.89616</td>
<td>0.89796</td>
<td>0.89973</td>
<td>0.90147</td>
<td>1.2</td>
</tr>
<tr>
<td>1.3</td>
<td>0.90320</td>
<td>0.90490</td>
<td>0.90658</td>
<td>0.90824</td>
<td>0.90988</td>
<td>0.91149</td>
<td>0.91308</td>
<td>0.91465</td>
<td>0.91621</td>
<td>0.91773</td>
<td>1.3</td>
</tr>
<tr>
<td>1.4</td>
<td>0.91924</td>
<td>0.92073</td>
<td>0.92219</td>
<td>0.92364</td>
<td>0.92506</td>
<td>0.92647</td>
<td>0.92785</td>
<td>0.92922</td>
<td>0.93056</td>
<td>0.93189</td>
<td>1.4</td>
</tr>
<tr>
<td>1.5</td>
<td>0.93319</td>
<td>0.93448</td>
<td>0.93574</td>
<td>0.93699</td>
<td>0.93822</td>
<td>0.93943</td>
<td>0.94062</td>
<td>0.94179</td>
<td>0.94295</td>
<td>0.94408</td>
<td>1.5</td>
</tr>
<tr>
<td>1.6</td>
<td>0.94520</td>
<td>0.94630</td>
<td>0.94738</td>
<td>0.94845</td>
<td>0.94950</td>
<td>0.95053</td>
<td>0.95154</td>
<td>0.95254</td>
<td>0.95352</td>
<td>0.95448</td>
<td>1.6</td>
</tr>
<tr>
<td>1.7</td>
<td>0.95543</td>
<td>0.95637</td>
<td>0.95728</td>
<td>0.95818</td>
<td>0.95907</td>
<td>0.95994</td>
<td>0.96080</td>
<td>0.96164</td>
<td>0.96246</td>
<td>0.96327</td>
<td>1.7</td>
</tr>
<tr>
<td>1.8</td>
<td>0.96407</td>
<td>0.96485</td>
<td>0.96562</td>
<td>0.96637</td>
<td>0.96711</td>
<td>0.96784</td>
<td>0.96856</td>
<td>0.96926</td>
<td>0.96995</td>
<td>0.97062</td>
<td>1.8</td>
</tr>
<tr>
<td>1.9</td>
<td>0.97128</td>
<td>0.97193</td>
<td>0.97257</td>
<td>0.97320</td>
<td>0.97381</td>
<td>0.97441</td>
<td>0.97500</td>
<td>0.97558</td>
<td>0.97615</td>
<td>0.97670</td>
<td>1.9</td>
</tr>
<tr>
<td>2.0</td>
<td>0.97725</td>
<td>0.97778</td>
<td>0.97831</td>
<td>0.97882</td>
<td>0.97932</td>
<td>0.97982</td>
<td>0.98030</td>
<td>0.98077</td>
<td>0.98124</td>
<td>0.98169</td>
<td>2.0</td>
</tr>
</tbody>
</table>

\[ \Phi(z) = \int_{-\infty}^{z} \frac{1}{\sqrt{2\pi}} e^{-u^2/2} du \]
What’s the Probability of...
Discrete Random Variables w/ $N$ “Equal Chances”

- Real measurements are generally discrete measurements.

- For discrete random variables with $N$ equally-likely values, the following is true:

$$p\{x_i\} = \frac{1}{N}$$

$$\mu = \frac{\sum_{i=1}^{N} x_i}{N} \quad \sigma^2 = \frac{\sum_{i=1}^{N} (x_i - \mu)^2}{N} \quad \sigma = \sqrt{\frac{\sum_{i=1}^{N} (x_i - \mu)^2}{N}}$$