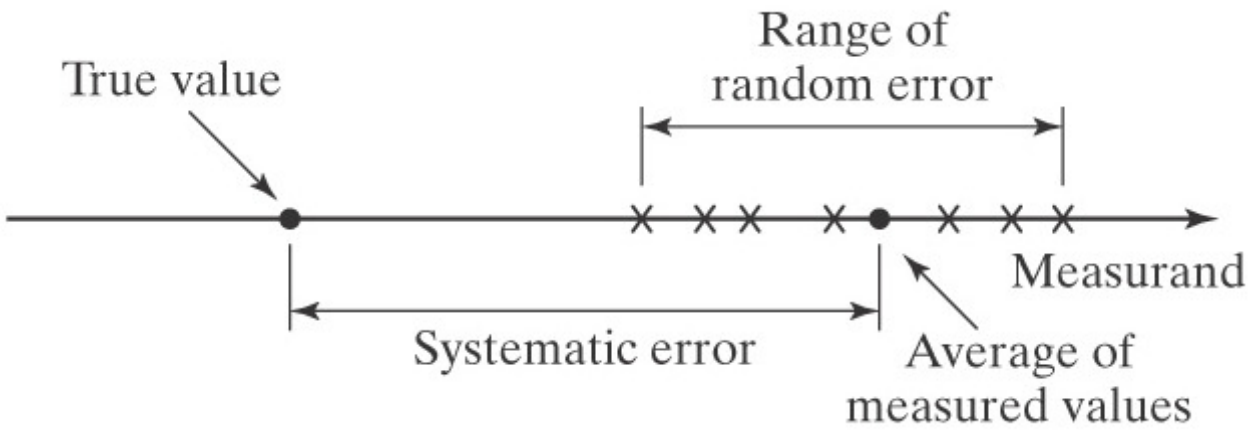
Probability Distributions

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 is 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

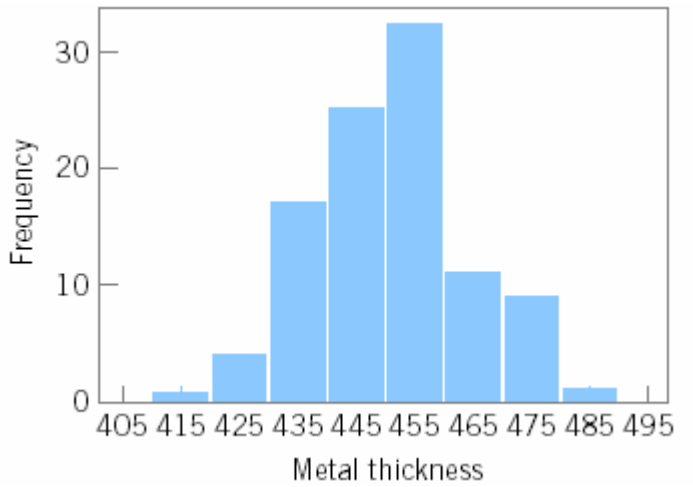
- For any event A, $0 \leq P\{A\} \leq 1$
- If $\overline{A}$ is the (exclusive) complement of A, $P\{A\} + P\{\overline{A}\} = 1$
- If A and B are independent events, $P\{A \text{ and } B\} = P\{A\}P\{B\}$
- If A and B are mutually exclusive, $P\{A \text{ or } B\} = P\{A\} + P\{B\}$
- If A and B are 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 $\mathrm{P}\{A \mid B\} = P\{A \text{ and } B\} / P\{B\}$

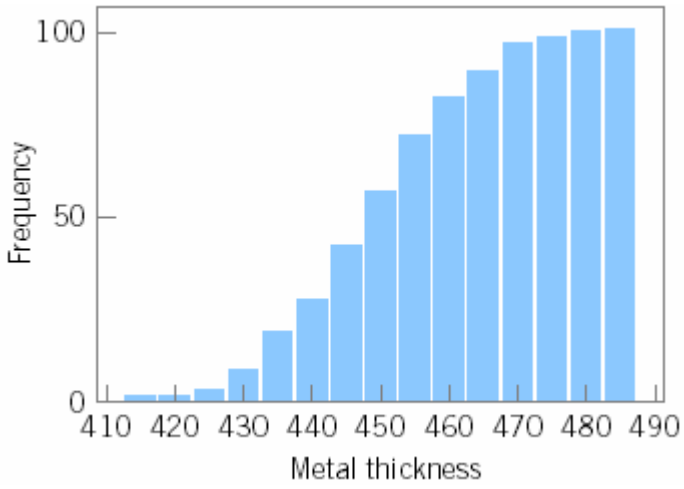
Measurements, Samples, and Populations

- 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."

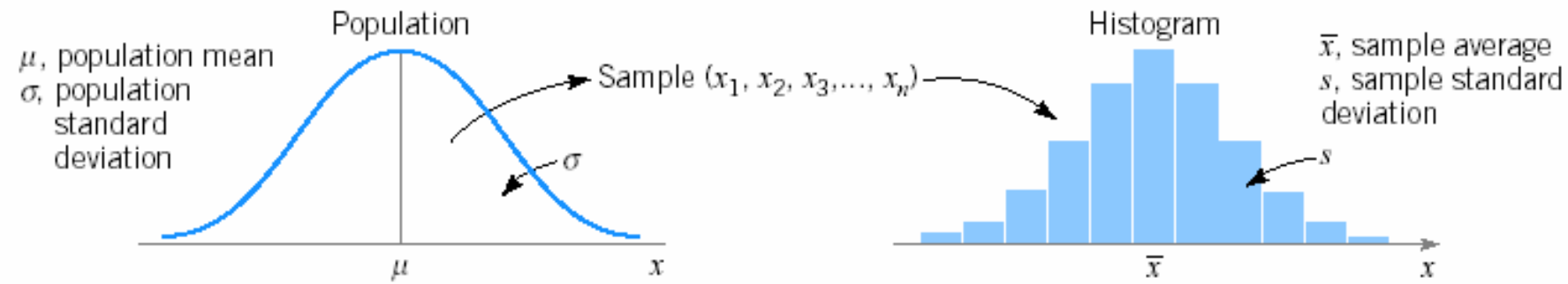
Histogram and Cumulative Frequency Plot

- 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:

$$\bar{x} = \frac{\sum_{i=1}^{n} x_i}{n}$$

- Sample Variance:

$$s^2 = \frac{\sum_{i=1}^{n} (x_i - \bar{x})^2}{n-1}$$

- Deviation (from mean):

$$d_i = x_i - \bar{x}$$

- Sample Standard Deviation:

$$s = \sqrt{\frac{\sum_{i=1}^{n} (x_i - \bar{x})^2}{n-1}}$$

- Sample Range:

$$R = x_{\max} - x_{\min}$$

Central Tendency & Variability in Populations

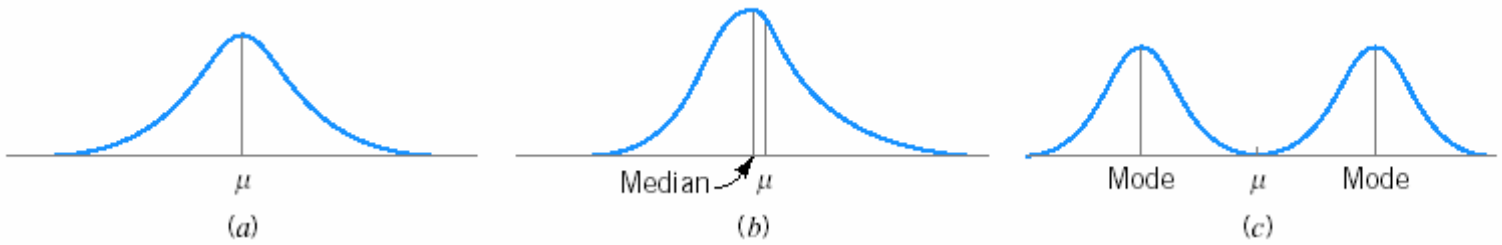


Figure 2-11 The mean of a distribution.

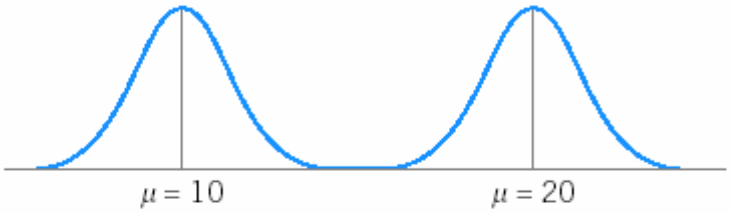


Figure 2-12 Two probability distributions with different means.

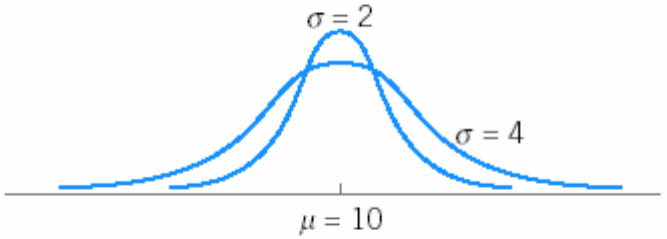
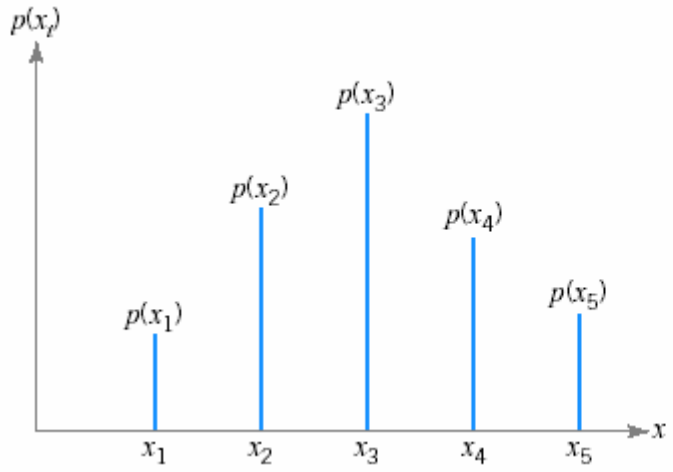


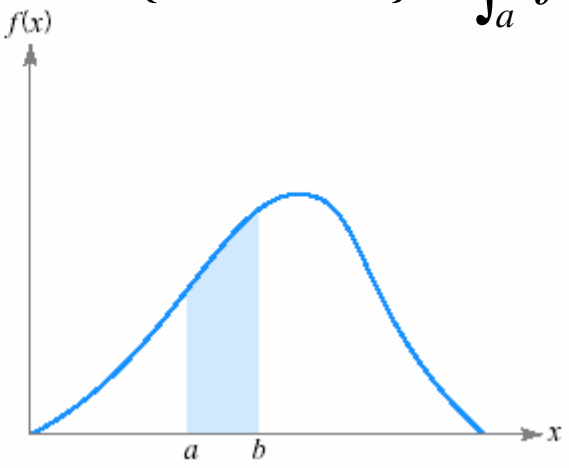
Figure 2-13 Two probability distributions with the same mean but different standard deviations.

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.

$$P\{x = x_i\} = p\{x_i\} \qquad P\{a \le x \le b\} = \int_a^b f\{x\}dx$$





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

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

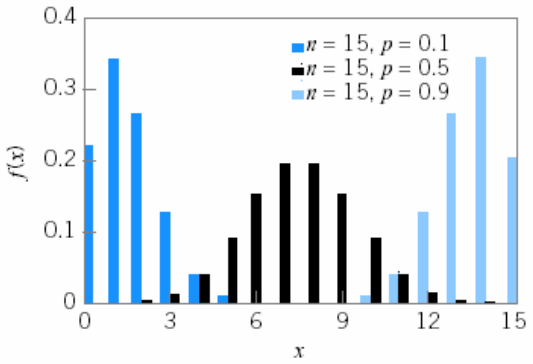
The mean and variance of the binomial distribution are

$$\mu = np \qquad (2\text{-}12)$$

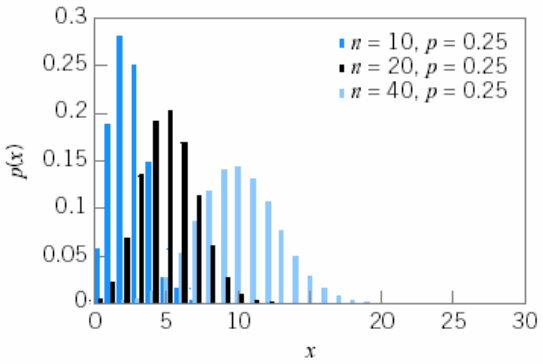
and

$$\sigma^2 = np(1-p) \qquad (2\text{-}13)$$

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



(*a*) Binomial distributions for different values of p with $n = 15$.



(*b*) Binomial distributions for different values of n with $p = 0.25$.

Poisson Distribution

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

$$P\{x \leq b\} = \sum_{x=0}^{b} \frac{e^{-4}\mu^{x}}{x!}$$

Definition

The **Poisson distribution** is

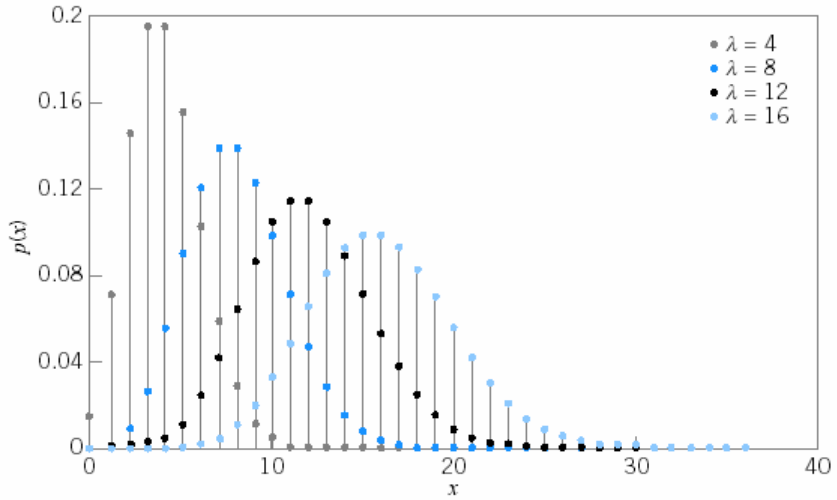
$$p(x) = \frac{e^{-\lambda}\lambda^{x}}{x!} \qquad x = 0,1,\ldots \qquad (2\text{-}15)$$

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

$$\mu = \lambda \qquad (2\text{-}16)$$

and

$$\sigma^2 = \lambda \qquad (2\text{-}17)$$



Normal Distribution

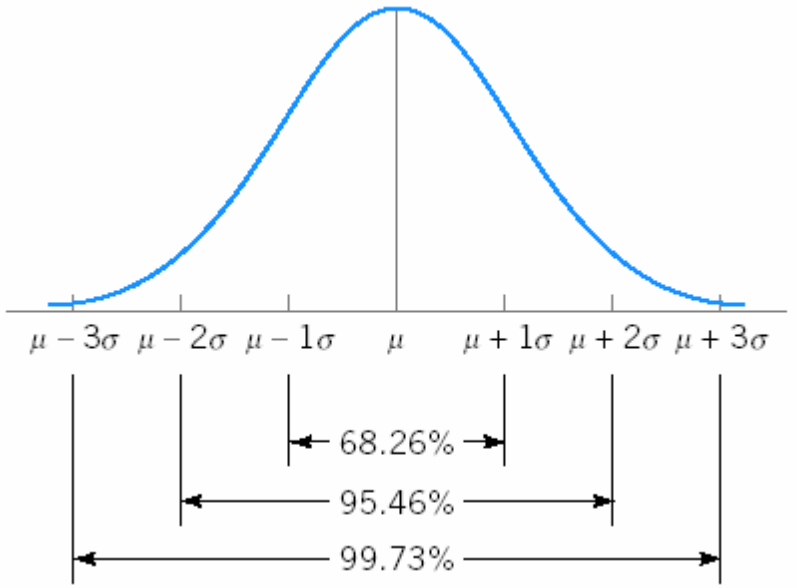
Definition

The **normal distribution** is

$$f(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2} \qquad -\infty < x < \infty \qquad (2\text{-}21)$$

The mean of the normal distribution is μ ($-\infty < \mu < \infty$) 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.

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

- Results are tabulated (thankfully) based on a single input, z, in what is called a cumulative standard normal distribution table.
- Also: $P\{x \ge a\} = 1 - P\{x \le a\}$

Central Limit Theorem

Definition: The Central Limit Theorem

If $x_1, x_2, \ldots, x_n$ are independent random variables with mean μ_i and variance σ_i^2, and if $y = x_1 + x_2 + \ldots + x_n$, then the distribution of

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

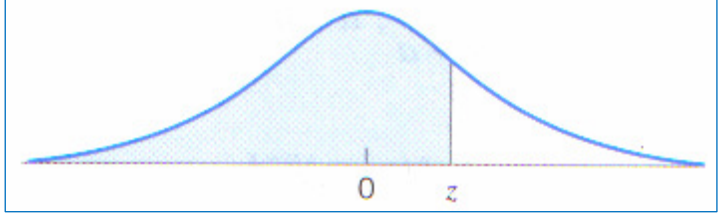
approaches the $N(0, 1)$ distribution as n approaches infinity.

- 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

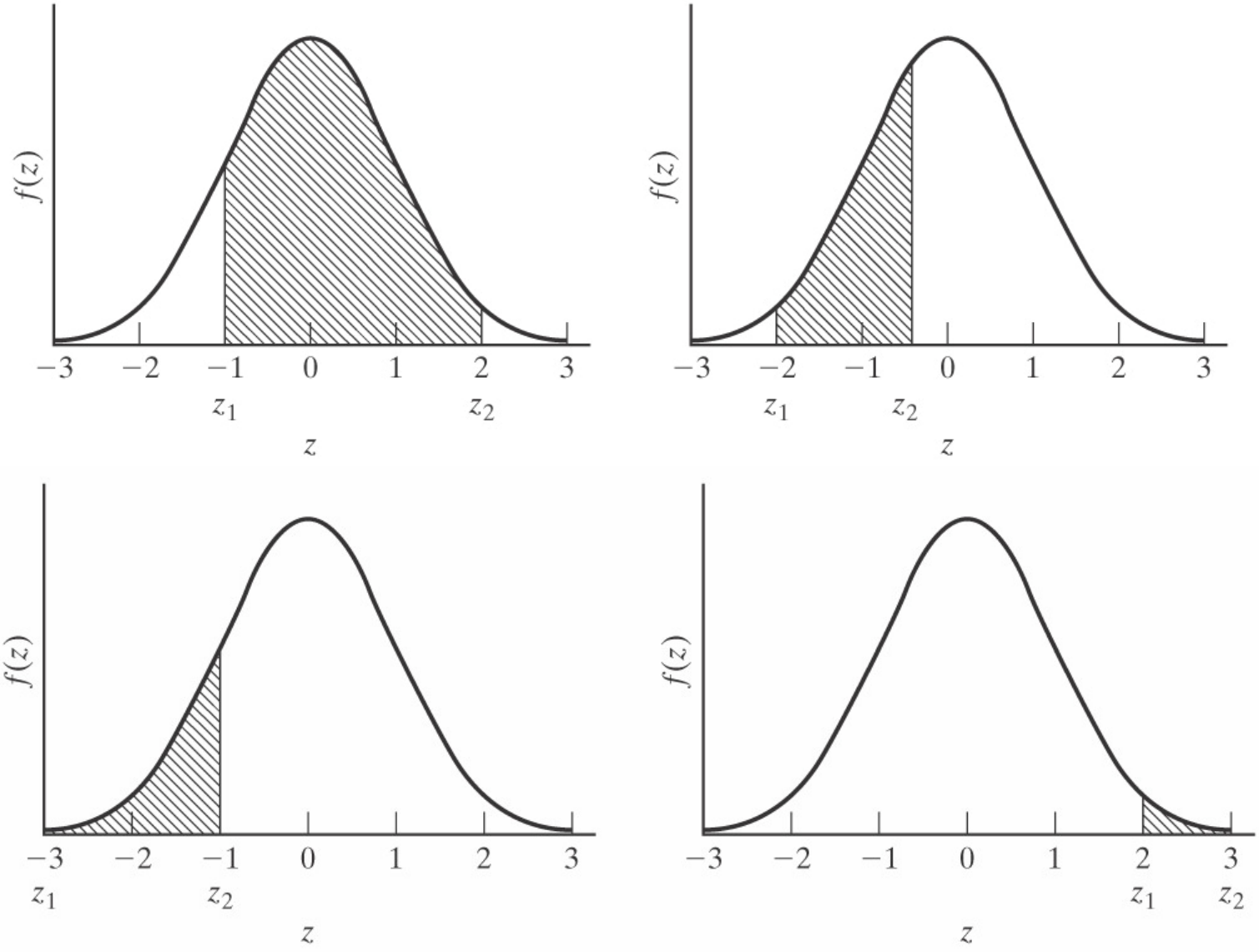
Tabulated Standard Normal Distribution Values

z	0.00	0.01	0.02	0.03	0.04	0.05	0.06	0.07	0.08	0.09	z
0.0	0.50000	0.50399	0.50798	0.51197	0.51595	0.51994	0.52392	0.52790	0.53188	0.53586	0.0
0.1	0.53983	0.54379	0.54776	0.55172	0.55567	0.55962	0.56356	0.56749	0.57142	0.57534	0.1
0.2	0.57926	0.58317	0.58706	0.59095	0.59483	0.59871	0.60257	0.60642	0.61026	0.61409	0.2
0.3	0.61791	0.62172	0.62551	0.62930	0.63307	0.63683	0.64058	0.64431	0.64803	0.65173	0.3
0.4	0.65542	0.65910	0.62276	0.66640	0.67003	0.67364	0.67724	0.68082	0.68438	0.68793	0.4
0.5	0.69146	0.69497	0.69847	0.70194	0.70540	0.70884	0.71226	0.71566	0.71904	0.72240	0.5
0.6	0.72575	0.72907	0.73237	0.73565	0.73891	0.74215	0.74537	0.74857	0.75175	0.75490	0.6
0.7	0.75803	0.76115	0.76424	0.76730	0.77035	0.77337	0.77637	0.77935	0.78230	0.78523	0.7
0.8	0.78814	0.79103	0.79389	0.79673	0.79954	0.80234	0.80510	0.80785	0.81057	0.81327	0.8
0.9	0.81594	0.81859	0.82121	0.82381	0.82639	0.82894	0.83147	0.83397	0.83646	0.83891	0.9
1.0	0.84134	0.84375	0.84613	0.84849	0.85083	0.85314	0.85543	0.85769	0.85993	0.86214	1.0
1.1	0.86433	0.86650	0.86864	0.87076	0.87285	0.87493	0.87697	0.87900	0.88100	0.88297	1.1
1.2	0.88493	0.88686	0.88877	0.89065	0.89251	0.89435	0.89616	0.89796	0.89973	0.90147	1.2
1.3	0.90320	0.90490	0.90658	0.90824	0.90988	0.91149	0.91308	0.91465	0.91621	0.91773	1.3
1.4	0.91924	0.92073	0.92219	0.92364	0.92506	0.92647	0.92785	0.92922	0.93056	0.93189	1.4
1.5	0.93319	0.93448	0.93574	0.93699	0.93822	0.93943	0.94062	0.94179	0.94295	0.94408	1.5
1.6	0.94520	0.94630	0.94738	0.94845	0.94950	0.95053	0.95154	0.95254	0.95352	0.95448	1.6
1.7	0.95543	0.95637	0.95728	0.95818	0.95907	0.95994	0.96080	0.96164	0.96246	0.96327	1.7
1.8	0.96407	0.96485	0.96562	0.96637	0.96711	0.96784	0.96856	0.96926	0.96995	0.97062	1.8
1.9	0.97128	0.97193	0.97257	0.97320	0.97381	0.97441	0.97500	0.97558	0.97615	0.97670	1.9
2.0	0.97725	0.97778	0.97831	0.97882	0.97932	0.97982	0.98030	0.98077	0.98124	0.98169	2.0
2.1	0.98214	0.98257	0.98300	0.98341	0.98382	0.98422					
2.2	0.98610	0.98645	0.98679	0.98713	0.98745	0.98778					
2.3	0.98928	0.98956	0.98983	0.99010	0.99036	0.99061					
2.4	0.99180	0.99202	0.99224	0.99245	0.99266	0.99286					
2.5	0.99379	0.99396	0.99413	0.99430	0.99446	0.99461					
2.6	0.99534	0.99547	0.99560	0.99573	0.99585	0.99598					
2.7	0.99653	0.99664	0.99674	0.99683	0.99693	0.99702					
2.8	0.99744	0.99752	0.99760	0.99767	0.99774	0.99781					
2.9	0.99813	0.99819	0.99825	0.99831	0.99836	0.99841					
3.0	0.99865	0.99869	0.99874	0.99878	0.99882	0.99886					

$$\Phi(z) = \int_{-\infty}^{z} \frac{1}{\sqrt{2\pi}} e^{-u^2/2} du$$



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:

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

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