Probability Refresher

MFT Topics discussed in these review slides are in bold:

1. **Measure of set operations**
2. **Conditional/joint probabilities**
3. **Counting rules**
4. Measures of central tendency and dispersion
5. **Distributions (including normal and binomial)**
6. Sampling and estimation
7. Hypothesis testing
8. Correlation and regression
9. Time-series forecasting
10. Statistical concepts in quality control

Sets

Sets are used to describe groups of things (called **elements**) – numbers, letter, cards, etc.

Examples: A = {1, 2, 3, 4, 5} and B = {2, 4, 6, 10}

Sets can be manipulated using set operations.

Two sets can be combined using the **union** operation (∪). Unions are associated with the word "and" – A ∪ B is read as "A and B":

Example: A ∪ B = {1, 2, 3, 4, 5, 6, 10}

Notice: 2 and 4, which appear in both A and B, are only written once in the union.

Sets

Recall: A = {1, 2, 3, 4, 5} and B = {2, 4, 6, 10}

The **intersection** operation (∩) results in the values shared by both sets. Intersections are associated with the word "or" – A ∪ B is read as "A or B":

Example: A ∩ B = {2, 4}

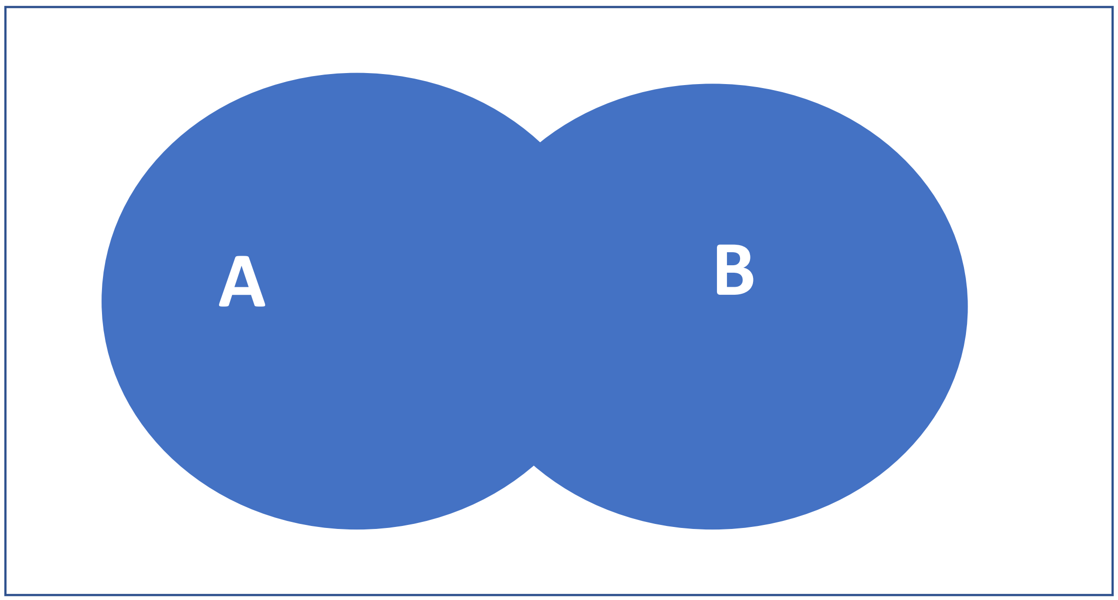
Sets are used to define **sample spaces** in probability applications.

A sample space is the set of all possible outcomes of a random event or random experiment.

Example: Flip a fair coin. The possible outcomes are heads and tails. The sample space can be written as S = {heads, tails}

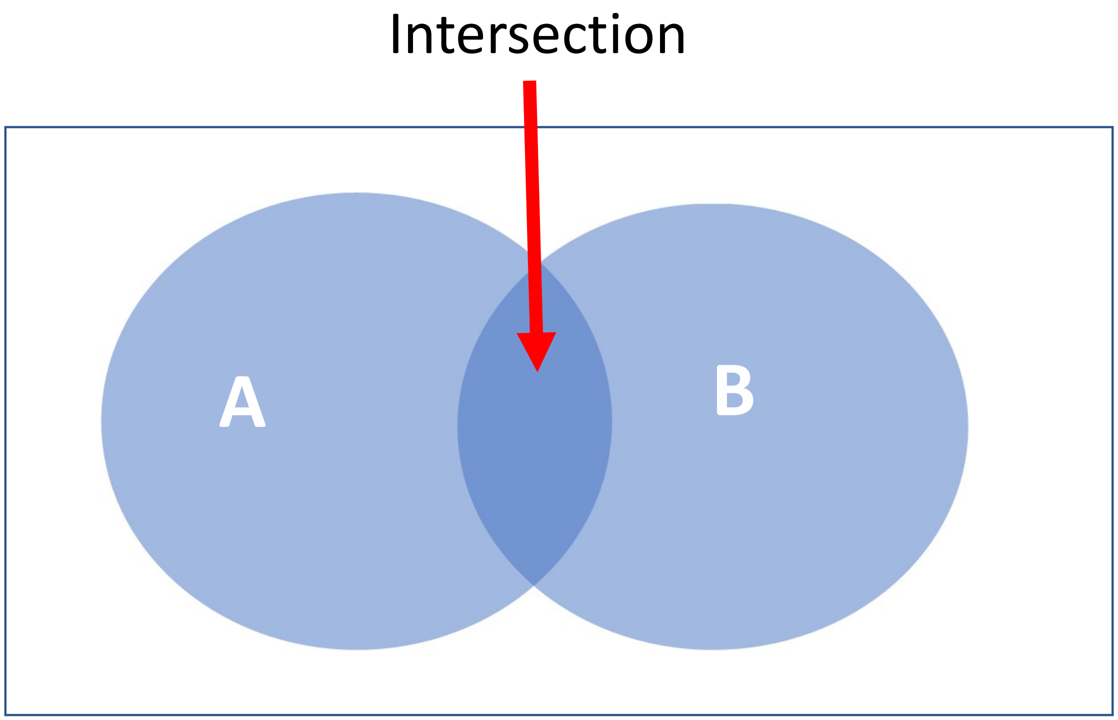
Sets and Venn diagrams

Venn diagrams are often used in conjunction with sets. This Venn diagram illustrates the union of two sets, A and B:



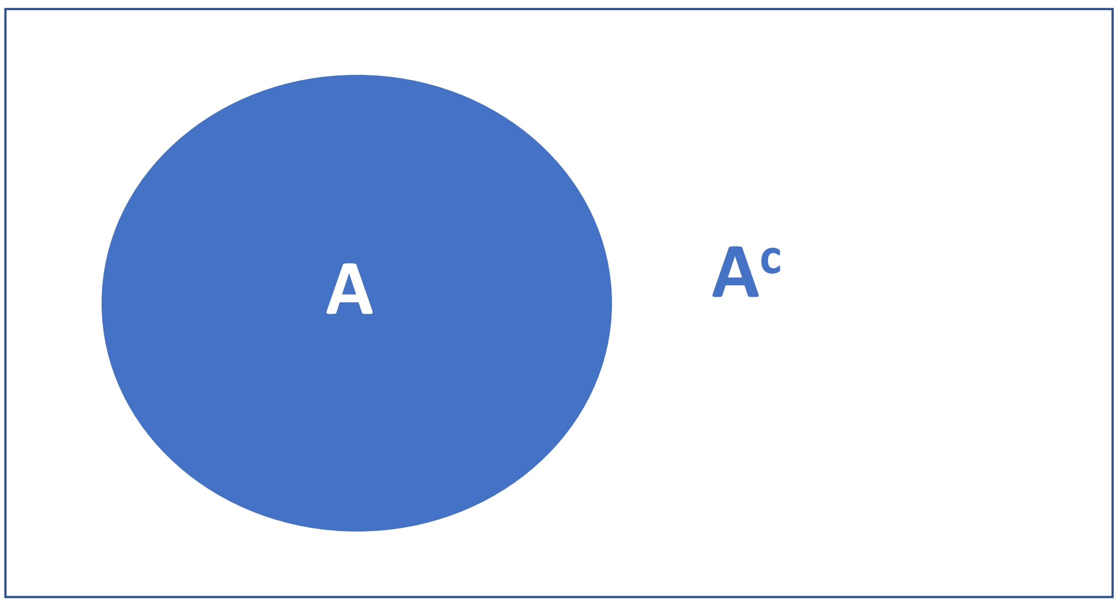
Sets and Venn diagrams

In the below, interception of the two sets, A and B, is shaded dark blue:



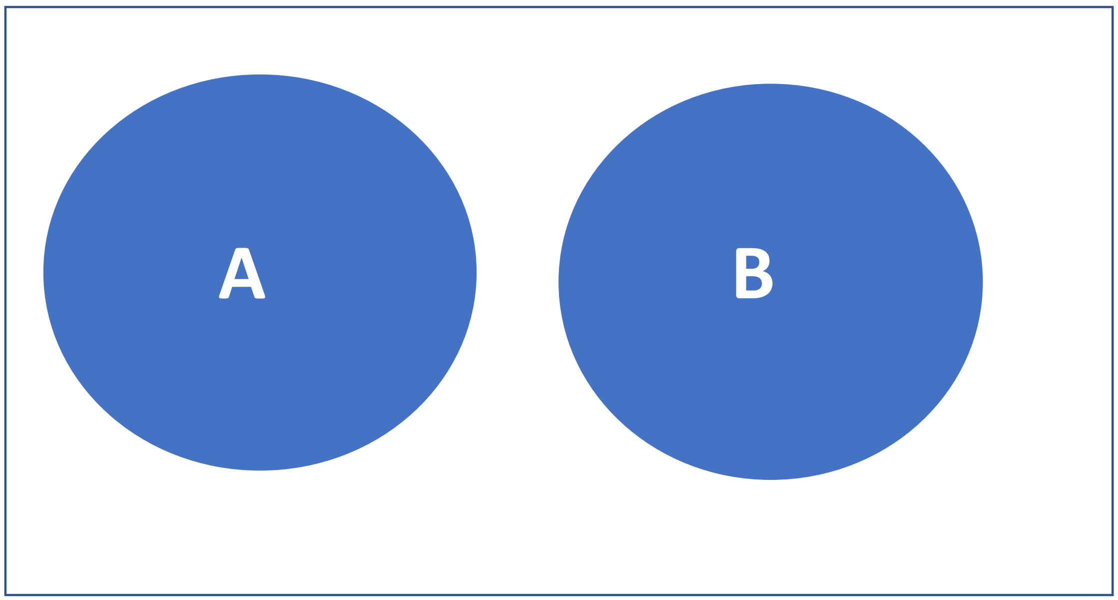
Sets and Venn diagrams

A complement is all the members of the sample space thatare not part of the set. In the diagram below, A^c is all of the unshaded space outside of the circle.



Sets and Venn diagrams

Sets are **mutually exclusive**, or **disjoint**, if they have no overlap. In the diagram below, sets A and B are mutually exclusive.



Probabilities

A probability is a number that describes the **likelihood** that an event will happen.

Probabilities are always between 0 and 1.0: $\mathbf{0 \leq P(A) \leq 1}$.

P(A) = probability that event A will occur

Probabilities are used to describe the future, not things that have already happened.

The **frequentist interpretation** of probability is widely accepted:

The probability of an outcome is the proportion of times the outcome would occur if we observed the random process an infinite number of times.

Sets and Probability

Set notation is used in probability formulas:

General addition rule: P(A or B) = P(A) + P(B) − P(A and B)

Multiplication rule, independent events: P(A and B) = P(A) × P(B)

Conditional Probability: P(A|B) = P(A and B)/P(B)

Multiplication rule, dependent events: P(A and B) = P(A) × P(B|A)

or P(A and B) = P(B) × P(A|B)

Complements: P(A) = 1 – P(A^c) and P(A^c) = 1 – P(A)

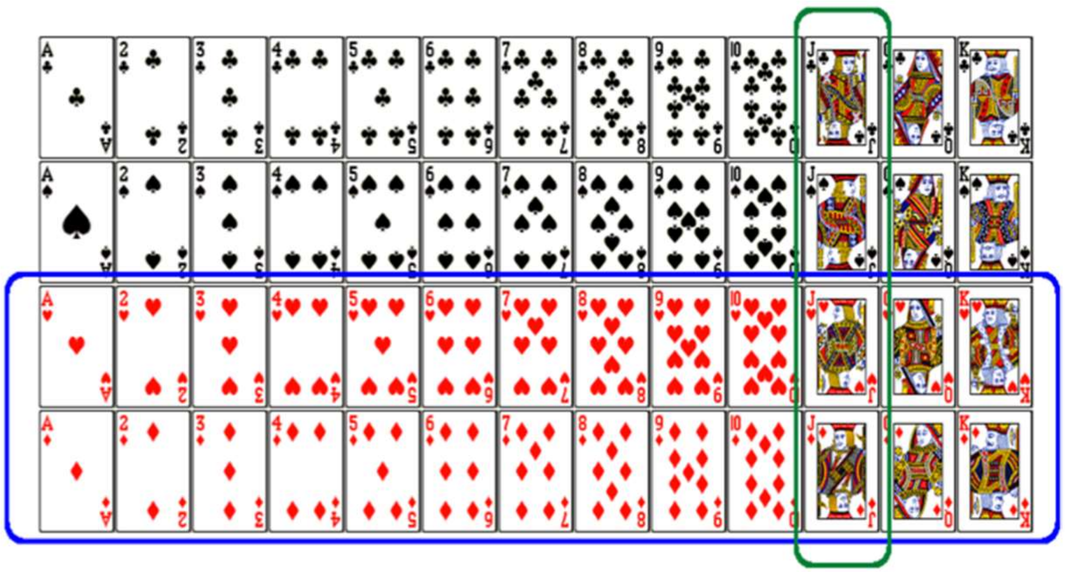
Note: P(A|B) means the **conditional probability** of A given B.

Practice #1

a) What is the probability of drawing a jack or a red card from a well shuffled full deck of 52 regular playing cards?
b) You toss a coin twice, what is the probability of getting two tails in a row?
c) Are the events “major in accounting” and “major in finance” mutually exclusive?

Practice #1 solutions

a) What is the probability of drawing a jack or a red card from a well shuffled full deck of 52 regular playing cards?



$$P(\text{jack or red}) = P(\text{jack}) + P(\text{red}) - P(\text{jack and red})$$
$$= \frac{4}{52} + \frac{26}{52} - \frac{2}{52} = \frac{28}{52}$$

Figure from *http://www.milefoot.com/math/discrete/counting/cardfreq.htm*

Practice #1 solutions

b) You toss a coin twice, what is the probability of getting two tails in a row?

Assume coin tosses are indepdent. Use multiplication rule.

P(T on the first toss) x P(T on the second toss)

$$= (1/2) \times (1/2) = 1/4$$

c) Are the events “major in accounting” and “major in finance” mutually exclusive?

No – a student could double major in accounting and finance at the same time.

Probabilities from tables

Probabilities that are calculated from data are called **empirical probabilities**.

Categorical variables classify observations into one or more groups.

Data from two categorical variables is often summarized using a type of table called a **contingency table** or **crosstabulation**.

Example: cocaine addiction treatments

Researchers randomly assigned 72 chronic users of cocaine into three groups: desipramine (antidepressant), lithium (standard treatment for cocaine) and placebo. Results of the study are summarized in the contingency table displayed below.

	relapse	no relapse	total
desipramine	10	14	24
lithium	18	6	24
placebo	20	4	24
total	48	24	72

http://www.oswego.edu/~srp/stats/2_way_tbl_1.htm

Marginal probability

Refer to the cocaine addiction treatment study. What is the probability that a patient relapsed?

	relapse	no relapse	total
desipramine	10	14	24
lithium	18	6	24
placebo	20	4	24
total	48	24	72

P(relapsed) = 48 / 72 ~ 0.67

This is called a **marginal probability** because the number used in the calculation are from the margins of the table.

Joint probability

Refer to the cocaine addiction treatment study. What is the probability that a patient received the antidepressant (desipramine) <u>and</u> relapsed?

	relapse	no relapse	total
desipramine	10	14	24
lithium	18	6	24
placebo	20	4	24
total	48	24	72

P(relapsed <u>and</u> desipramine) = 10 / 72 ~ 0.14

This **joint probability** is calculated by dividing the count at the intersection of the row and column by the table total.

Marginal probability

Refer to the cocaine addiction treatment study. What is the probability that a patient relapsed?

	relapse	no relapse	total
desipramine	10	14	24
lithium	18	6	24
placebo	20	4	24
total	48	24	72

P(relapsed) = 48 / 72 ~ 0.67

This is called a **marginal probability** because the number used in the calculation are from the margins of the table.

Conditional probability

Refer to the cocaine addiction treatment study. If we know that a patient received the antidepressant (desipramine), what is the probability that they relapsed?

	relapse	no relapse	total
desipramine	10	14	24
lithium	18	6	24
placebo	20	4	24
total	48	24	72

P(relapse | desipramine) = 10 / 24 ~ 0.42

This **conditional probability** can be calculated using the relapse count and row total on the row that corresponds to the condition, desipramine.

Conditional probability

Alternatively, use the formula for conditional probability:

$$P(A|B) = \frac{P(A \text{ and } B)}{P(B)}$$

$$P(relapse|desipramine)$$
$$= \frac{P(relapse \text{ and } desipramine)}{P(desipramine)}$$
$$= \frac{10/72}{24/72}$$
$$= \frac{10}{24}$$
$$= 0.42$$

	relapse	no relapse	total
desipramine	10	14	24
lithium	18	6	24
placebo	20	4	24
total	48	24	72

Independence

If $P(A \mid B) = P(A)$ then the events A and B are said to be **independent**.

Conceptually: Giving B doesn't give useful information about A.

If $P(A \mid B) \neq P(A)$ then the events A and B are said to be **dependent**.

Conceptually: Giving B is useful information about A because it changes the probability.

Some notes about counting

When calculating probabilities, we often have to count how many different ways our event of interest could occur.

Example: In a group of five students, an instructor might wish to select three of them to make presentations.

If order does not matter, this is called a **combination**.

If order does matter, it is called a **permutation**.

Combinations

The number of combinations for selecting *n* objects from a set of *N* is given by the formula:

$$C(n,\ N) = \binom{N}{n} = \frac{N!}{n!(N-n)!}$$

Example: In a group of five students, an instructor might wish to select three of them to make presentations. *Order does not matter.*

$$C(3,\ 5) = \binom{5}{3} = \frac{5!}{3!(5-3)!} = \frac{(5)(4)(3)(2)(1)}{3(2)(1)\times(2)(1)} = 10$$

There are 10 ways of selecting three students from a group of five.

Permutations

The number of permutations of n objects from a set of N is given by the formula:

$$P(n, N) = n!\binom{N}{n} = \frac{N!}{(N-n)!}$$

Example: In a group of five students, an instructor might wish to select three of them the order is important (for instance, knowing which student presents first, second, and third).

$$P(3, 5) = 3!\binom{5}{3} = \frac{5!}{(5-3)!} = \frac{(5)(4)(3)(2)(1)}{(2)(1)} = 60$$

There are 60 potential ordered sets of 3 students from a group of 3.

Tips for solving probability problems

Visualizations often help.

- Sketch the sample space.
- Draw a Venn diagram.
- Arrange the data into a table format.

Practice #2

a) Consider the contingency table. What is the probability that a randomly sampled student thinks marijuana should be legalized <u>or</u> they agree with their parents' political views?

(a) (40 + 36 - 78) / 165
(b) (114 + 118 - 78) / 165
(c) 78 / 165
(d) 78 / 188
(e) 11 / 47

	Share Parents' Politics		
Legalize MJ	No	Yes	Total
No	11	40	51
Yes	36	78	114
Total	47	118	165

b) Roll two fair six-sided dice (like the kind from Monopoly or Yahtzee). What is the probability that they sum to 9 or more?

Practice #2

c) In a large firm, the pool of applicants for the positions of sales manager and sales associate consists of three men and three women. If each candidate has an equal chance of being selected, what is the probability that both positions will be filled by men?

(A) 0.20

(B) 0.25

(C) 0.33

(D) 0.50

Practice #2 solutions

a) Consider the contingency table. What is the probability that a randomly sampled student thinks marijuana should be legalized <u>or</u> they agree with their parents' political views?

(b) (114 + 118 - 78) / 165

	Share Parents' Politics		
Legalize MJ	No	Yes	Total
No	11	40	51
Yes	36	78	114
Total	47	118	165

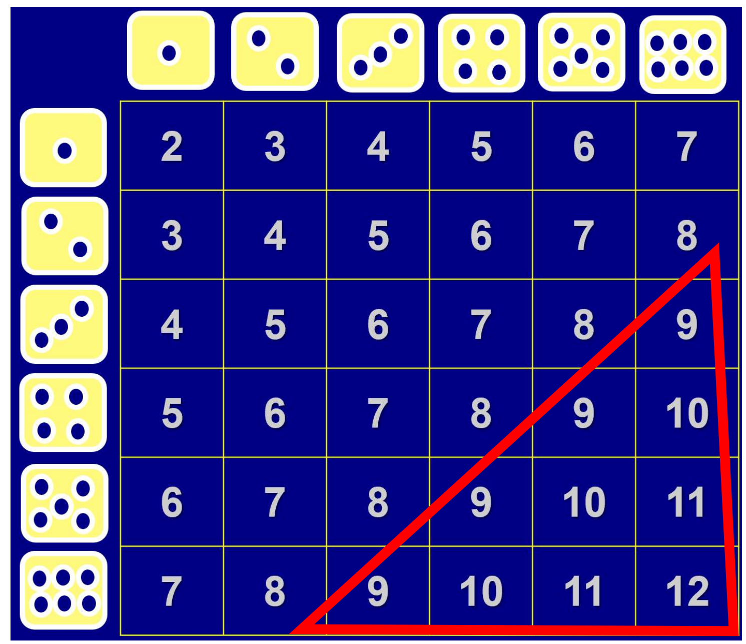
(Don't count the 78 at the intersection twice!)

Practice #2 solutions

b) Roll two fair six-sided dice (like the kind from Monopoly or Yahtzee). What is the probability that they sum to 9 or more?

10/36 ~ .28

Visualize the sample space, which has 36 possible outcomes. There are 10 ways to get a sum of 9 or more.



Practice #2 solutions

c) In a large firm, the pool of applicants for the positions of sales manager and sales associate consists of three men and three women. If each candidate has an equal chance of being selected, what is the probability that both positions will be filled by men?
**** (A) 0.20**

Why? P(first is man) = 3/6; P(second is man|first is man) = 2/5; The probability of this sequence of events is (3/6)(2/5) = 6/30 = 1/5 = .2

Random variables

A **random variable** is a numeric quantity whose value depends on the outcome of a random event.

We often use a capital letter, like X or Y, to name a random variable. The values of a random variable are denoted with a lowercase letter, such as x. $P(X = x)$, for example.

Discrete random variables often take only integer values and often correspond to counts. Examples: number of hits on a website in a day, number of credit hours.

A **continuous** random variable has outcomes over one or more continuous intervals of real numbers. Examples: daily temperature, time to run a 5K race.

Probability distributions

A **probability distribution** is a characterization of the possible values that a random variable may assume along with the probability of assuming these values.

Probability distributions can be described using graphs, tables, or formulas.

We can calculate the **relative frequencies** from a sample of empirical data to develop a probability distribution. Because this is based on sample data, we usually call this an **empirical probability distribution**.

Distribution characteristics

Important characteristics of probability distributions are the expected value and variance.

The **expected value**, or **mean**, of a probability distribution is interpreted as the long-run average if you observed the outcome of the random variable many, many times and averaged the results.

The **variance** (and related standard deviation) describe the variability in the values of the random variable. The variance is based on the idea of **deviations from the mean** value. The standard deviation is the square root of the variance.

Calculations for discrete RVs

For discrete random variables, the expected value (μ) and variance (σ^2) can be calculated using the following formulas:

$$\mu = E(X) = \sum_{i=1}^{k} x_i P(X = x_i)$$

$$\sigma^2 = Var(X) = \sum_{i=1}^{k} (x_i - E(X))^2 P(X = x_i)$$

$$\sigma = SD(X) = \sqrt{Var(X)}$$

A simple game

Consider a simple gambling game using a regular deck of 52 playing cards that has been randomly shuffled. Wager $5 to draw a card at random. If the card is a face card (J, Q, K, A), win $10 (and get your wager back). If the card is not a face card, you lose your wager.

Let X = the result of one wager.

Write out the probability distribution of X and calculate E(X).

Example calculations

Since there are four face cards for each of four suits, there are 16 winning cards. The probability of winning the wager is 16/52.

The probability distribution can be written as a table:

X	**P(x)**
-5	36/52
10	16/52

$E(X) = (-5)(36/52) + (10)(16/52) = -0.385$

Since the game has a negative expected value, the players are expected to lose money in the long run and the house is expected to make money in the long run.

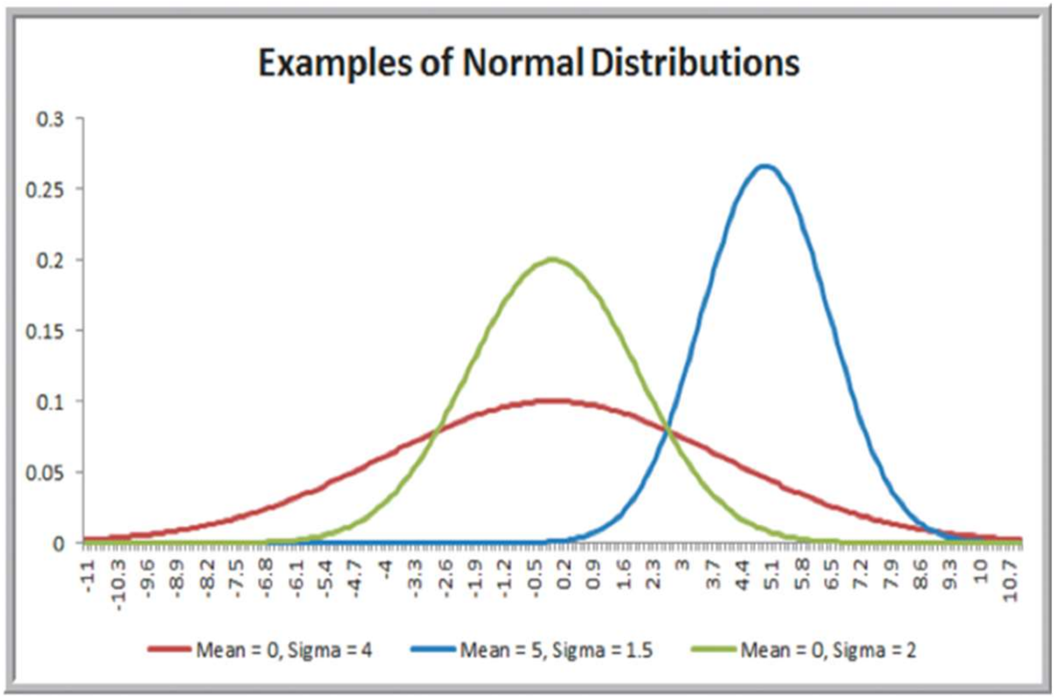
Common probability distributions

There are some probability distributions that are particularly useful for describing or modeling real life phenomena.

Two particularly common probability distributions are the normal distribution and the binomial distribution.

Normal distributions

- A **continuous** distribution
- Unimodal and symmetric, bell shaped curve
- Many variables are nearly normal, but none are exactly normal
- Denoted as N(μ, σ) → Normal with mean μ and standard deviation σ



Normal distributions

Calculations are typically performed using software. For example, the Excel function **NORM.DIST** can calculate ≤ probabilities.

Example:

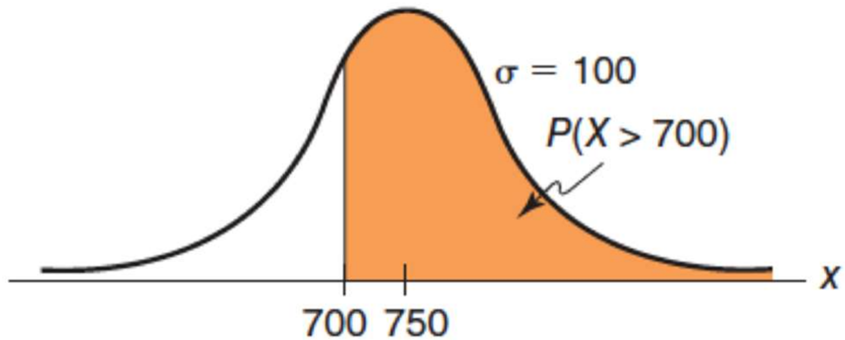
Suppose X ~ Normal(μ = 100, σ = 15). Find P(X ≤ 90).

In Excel, P(X ≤ 90) =NORM.DIST(90, 100, 15, true).

To get ≥ probabilities, use the complement (i.e. subtract from 1). To get the probability between two values, find the difference in the ≤ probabilities.

Normal calculation example

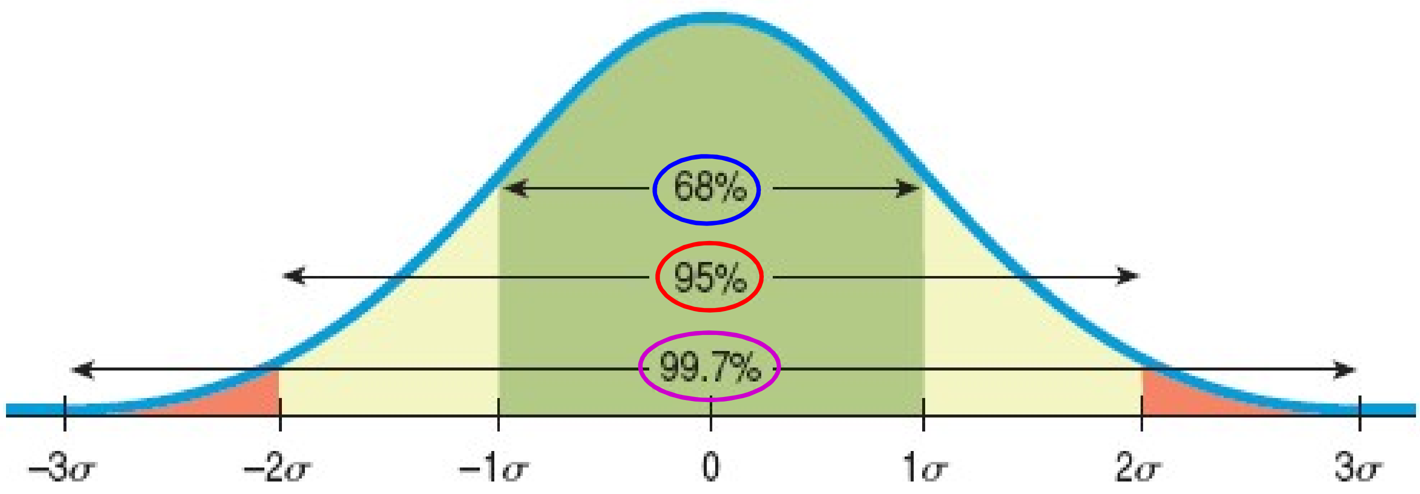
The distribution for customer demand (units per month) averages 750 with a standard deviation of 100. If the distribution is accurately described by a normal distribution, what is the probability that demand will exceed 700 units?



P(X ≥ 700) = 1 – NORM.DIST(700, 750, 100, true) = .6915

Empirical rule (or 68-95-99.7 rule)

In a unimodal, symmetric distribution, about 68% of the values fall within one standard deviation of the mean, about 95% of the values fall within two standard deviations of the mean, and about 99.7% of the values fall within three standard deviations of the mean.



Binomial distributions

The binomial distribution describes the probability of having exactly k successes in n independent binary outcome (success-failure) trials with probability of success p.

"Success" is arbitrarily defined as the outcome being counted. The result is a count and therefore binomial distributions are **discrete**.

Example: Suppose you are late to class and missed a pop quiz. Your instructor gives you time to write down answers, but refuses to show the questions on the screen again. Suppose each of the ten questions had five answers choices, a-e. Let X = the count of the number you guessed correctly. Then X ~ binomial (n = 10, p = 1/5 = .2).

Binomial calculations

If *p* represents probability of success, (1-*p*) represents probability of failure, *n* represents number of independent trials, and *k* represents number of observed successes, then calculate probabilities using the following formula:

$$P(k \text{ successes in } n \text{ trials}) = \binom{n}{k} p^k \, (1-p)^{(n-k)}$$

We often use software, such as Excel's **BINOM.DIST** function to automate the calculations.

Binomial calculation example

Example: Suppose you are late to class and missed a pop quiz. Your instructor gives you time to write down answers, but refuses to show the questions on the screen again. Suppose each of the ten questions had five answers choices, a-e. Let X = the count of the number you guessed correctly. Then X ~ binomial (*n* = 10, *p* = 1/5 = .2).

Find the probability that you score 0 on the quiz.

$$P(X = 0) = \binom{10}{0} .2^0\, .8^{10} = .1073$$

Alternatively, P(X = 0) = BINOM.DIST(0, 10, .2, false) = .1073

Practice #3

a) Suppose you are told that the average return on investment for a particular class of investments was 7.8% with a standard deviation of 2.3. Furthermore, the histogram of the distribution of returns is approximately bell-shaped. We would expect that 95 percent of all of these investments had a return between what two values?

A) 5.5% and 10.1%

B) 0% and 15%

C) 3.2% and 12.4%

D) 0.9% and 14.7%

Practice #3

b) A 2012 Gallup survey suggests that 26.2% of Americans are obese. Among a random sample of 10 Americans, what is the probability that exactly 8 are obese?

- ❑ pretty high
- ❑ pretty low

Practice #3

c) A 2012 Gallup survey suggests that 26.2% of Americans are obese. Among a random sample of 10 Americans, what is the probability that exactly 8 are obese?

(a) $0.262^8 \times 0.738^2$

(b) $\binom{8}{10} \times 0.262^8 \times 0.738^2$

(c) $\binom{10}{8} \times 0.262^8 \times 0.738^2$

(d) $\binom{10}{8} \times 0.262^2 \times 0.738^8$

Practice #3 solutions

a) Suppose you are told that the average return on investment for a particular class of investments was 7.8% with a standard deviation of 2.3. Furthermore, the histogram of the distribution of returns is approximately bell-shaped. We would expect that 95 percent of all of these investments had a return between what two values?

****C) 3.2% and 12.4%**

Use the empirical rule (or 68-95-99.7 rule): 95% is mean ± 2 standard deviations → 7.8 ± (2)(2.3) → 7.8 ± 4.6 → 3.2 to 12.4

Practice #3 solutions

b) A 2012 Gallup survey suggests that 26.2% of Americans are obese. Among a random sample of 10 Americans, what is the probability that exactly 8 are obese?

- **pretty low**

Practice #3 solutions

c) A 2012 Gallup survey suggests that 26.2% of Americans are obese. Among a random sample of 10 Americans, what is the probability that exactly 8 are obese?

(a) $0.262^8 \times 0.738^2$

(b) $\binom{8}{10} \times 0.262^8 \times 0.738^2$

(c) $\binom{10}{8} \times 0.262^8 \times 0.738^2 = 45 \times 0.262^8 \times 0.738^2 = 0.0005$

(d) $\binom{10}{8} \times 0.262^2 \times 0.738^8$