Probability Review

Lecture 2:
Probability Review and Hypothesis Testing

ECE/CS 498 DS

Professor Ravi K. Iyer

University of Illinois

Announcements

- **HW0 has been released and is due on 2/3 @ 11:59 PM**
  - Please upload your solved HW on Compass2G – **Let TAs know ASAP if you don't have access to Compass2G!**
  - You can either (i) handwrite and scan or (ii) type your answers
- **Submit your MP Groups by Wed 1/29 @ 11:59 PM**
  - Signup via Google Form, link shared on Piazza (https://forms.gle/nkjure9LXVjKKfEKA)
  - Look at "Search for Partners" page on Piazza for help finding groups
  - Guidelines:
    - Groups should consist of exactly 3 students
    - You will work with the same groups for all the MPs (and if applicable, final project)
    - 4-credit hour students should only work with other 4 credit-hour students, and 3 credit-hour students should only work with other 3-credit hour students

Course Timeline

- Last Week
  - **Wednesday 1/22**: Introduction to the course
- This Week
  - **Monday 1/27**: Probability review and p-value
    - HW 0 Released
  - **Wednesday 1/29**: Introduction to MP 1 and Jupyter Notebook Review
  - **Thursday 1/30:** MP 1 released
- Next Week
  - **Monday 2/3**: In Class Activity 1: Conditional probability and hypothesis testing
    - HW 0 Due at 11:59 PM
  - **Wednesday 2/5**: Naïve Bayes networks

PROBABILITY BASICS

Basic Terminology

- ***Random experiment*** is an experiment the outcome of which is not certain
- ***Sample Space (S)*** is the totality of the possible outcomes of a random experiment
- ***Discrete (countable) sample space*** is a sample space which is either
  - *Finite* - the set of all possible outcomes of the experiment is finite
  - *Countably Infinite* - the set of all outcomes can be put into a one-to-one correspondence with the natural numbers
- ***Continuous sample space*** is a sample space for which all elements constitute a continuum, such as all the points on a line, all the points in a plane
- An ***event*** is a collection of certain sample points, i.e., a subset of the sample space
  - ***Universal event*** is the entire sample space S
  - *The null set* $\varnothing$ is a **null or impossible event**

Algebra of Events

- **Algebra of Events**
  - The ***intersection*** of E_1 and E_2 is given by:
    - $\mathrm{E}_1 \cap \mathrm{E}_2 = \{s \in \mathrm{S} \mid s$ is an element of both E_1 and $\mathrm{E}_2\}$
  - The ***union*** E_1 and E_2 is given by:
    - $\mathrm{E}_1 \cup \mathrm{E}_1 = \{s \in \mathrm{S} \mid$ either $s \in \mathrm{E}_1$ or $s \in \mathrm{E}_2$ or both$\}$
  - In general: $|\mathrm{E}_1 \cup \mathrm{E}_2| \leq |\mathrm{E}_1| + |\mathrm{E}_2|$
    - where $|\mathrm{A}| =$ the number of elements in the set **(Cardinality)**
  - Definition of *union* and *intersection* extend to any finite number of sets:

$$\bigcup_{i=1}^{n} E_i = E_1 \cup E_2 \cup E_3 \cup \ldots \cup E_n$$

$$\bigcap_{i=1}^{n} E_i = E_1 \cap E_2 \cap E_3 \cap \ldots \cap E_n$$

Mutual Exclusive and Collectively Exhaustive

- Two events A and B are ***mutually exclusive/disjoint*** iff

$$A \cap B = \varnothing$$

- A list of events $A_1, A_2, \ldots, A_n$ is said to be
  - composed of ***mutually exclusive events*** iff:

$$A_i \cap A_j = \begin{cases} A_i, & \text{if } i = j \\ \emptyset, & \text{otherwise} \end{cases}$$

  - ***collectively exhaustive*** iff:

$$A_1 \cup A_2 \cup \cdots \cup A_n = S$$

- A list of events $A_1, A_2, \ldots, A_n$ is said to **partition** S if
  - $A_1, A_2, \ldots, A_n$ are pairwise disjoint
  - $A_1, A_2, \ldots, A_n$ are collectively exhaustive

- Note that "iff" is a condensed representation of "if and only if"

Probability Axioms

- **Probability Axioms**
  - Let S be a sample space of a random experiment and $P(A)$ be the probability of the event A
  - The probability function $P(.)$ must satisfy the three following axioms:
  - **(A1)** For any event A, $\mathrm{P(A)} \geq 0$

    (probabilities are nonnegative real numbers)
  - **(A2)** $\mathrm{P(S)} = 1$

    (probability of a certain event, an event that must happen is equal 1)
  - **(A3)** $\mathrm{P(A \cup B) = P(A) + P(B)}$, whenever A and B are mutually exclusive events, i.e., $\mathrm{A \cap B} = \emptyset$

    (probability function must be additive)
  - **(A3' - General)** For any countable sequence of events $\mathrm{A_1, A_2, \ldots, A_n} \ldots$, that are mutually exclusive (that is $\mathrm{A_j \cap A_k} = \emptyset$ whenever $\mathrm{j \neq k}$)

$$P\left(\bigcup_{n=1}^{\infty} A_n\right) = \sum_{n=1}^{\infty} P(A_n)$$

Derived Probability Relations

- **(Ra)** For any event A, $P(\bar{A}) = 1 - P(A)$
- **(Rb)** If $\varnothing$ is the impossible event, then $\mathrm{P}(\varnothing) = 0$
- **(Rc)** If A and B are any events, not necessarily mutually exclusive, then

$$\mathrm{P}(\mathrm{A} \cup \mathrm{B}) = \mathrm{P}(\mathrm{A}) + \mathrm{P}(\mathrm{B}) - \mathrm{P}(\mathrm{A} \cap \mathrm{B})$$

- **(Rc' - generalization of Rc**) If $\mathrm{A}_1, \mathrm{A}_2, \ldots, \mathrm{A}_\mathrm{n}$ are any events, then

$$P\left(\bigcup_{i=1}^{n} A_i\right) = \sum_{i} P(A_i) - \sum_{1 \le i < j \le n} P(A_i \cap A_j) + \sum_{1 \le i < j < k \le n} P(A_i \cap A_j \cap A_k) +$$

$$\ldots + (-1)^{n+1} P(A_1 \cap A_2 \cap \cdots \cap A_n)$$

where the successive sums are over all possible events, pairs of events, triples of events, and so on (Can prove this relation by induction…)

Discrete/Continuous Random Variables

- Random Variable $X: S \rightarrow \mathbb{R}$

- The ***discrete*** random variables are either a finite or a countable number of possible values.

- Random variables that take on a continuum of possible values are known as ***continuous*** random variables.
  - Example: A random variable denoting the time to disengagement of an AV, when the time is assumed to take on any value in some interval $(0, \infty)$ is *continuous*

- Common random variables (RVs)
  - Discrete: Bernoulli, Binomial, Geometric, Poisson
  - Continuous: Uniform, Gaussian, Exponential, Gamma, Beta, Weibull
  - Refer to end of slide deck for additional information about these distributions

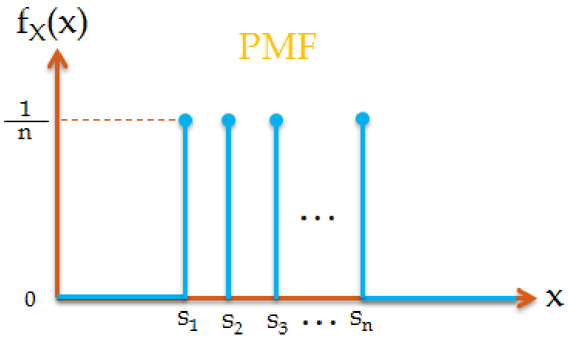
Discrete Random Variables

- **Probability Mass Function (PMF)**

$$p(a) = P\{X = a\}$$

$$p(x) = \begin{cases} > 0, & x = x_1, x_2, \dots \\ 0, & \text{for other values of } x \end{cases}$$

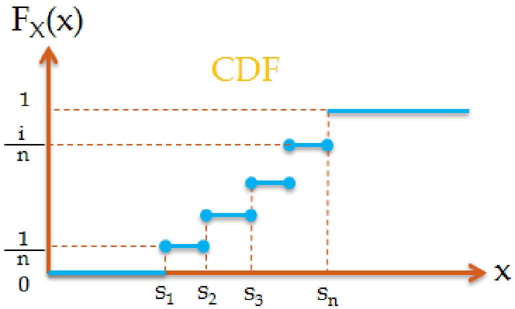
$$\sum_{i=1}^{\infty} p(x_i) = 1$$



- **Cumulative Mass Function (CMF)**

$$F(a) = \sum_{all\ x_i \leq a} p(x_i)$$

  - Staircase, right-continuous function



Discrete Random Variables: Probability Mass Function (pmf)

- A random variable that can take on at most countable number of possible values is said to be *discrete*.
- For a discrete random variable X, we define the ***probability mass function*** $p(a)$ of X by:

$$p(a) = P\{X = a\}$$

- $p(a)$ is positive for at most a countable number of values of a.

i.e., if X must assume one of the values $x_1, x_2, \ldots$ then

$$p(x) \begin{cases} > 0, & x = x_1, x_2, \ldots \\ = \ 0, & \text{for other values of } x \end{cases}$$

- Since X takes values x_i:

$$\sum_{i=1}^{\infty} p(x_i) = 1$$

An Example of Discrete RV
- *Geometric Distribution*

- Consider a geometric random variable Z
  - By definition, the event $[Z = i]$ occurs iff we have a sequence of $(i - 1)$ failures followed by one success
  - In other words, a sequence of independent Bernoulli trials with the probability of success being p and of failure being $q = (1 - p)$
- Thus, we have

$$p_Z(i) = q^{i-1}p = p(1-p)^{i-1} \qquad \text{for i=1, 2, ...}$$

- Using the formula for the sum of a *geometric* series, we have:

$$\sum_{i=1}^{\infty} p_Z(i) = \sum_{i=1}^{\infty} pq^{i-1} = \frac{p}{1-q} = \frac{p}{p} = 1$$

- The corresponding CDF is:

$$F_Z(t) = \sum_{i=1}^{\lfloor t \rfloor} p(1-p)^{i-1} = 1 - (1-p)^{\lfloor t \rfloor} \text{ for } t \geq 0$$

Continuous Random Variables

- **Continuous Random Variables:**
  - **Probability density function (pdf)**:

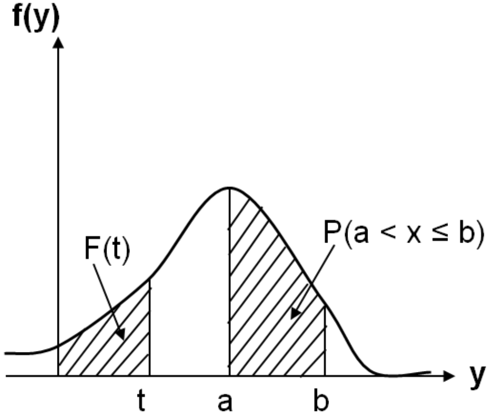
$$P\{X \in B\} = \int_B f(x)dx$$

    - Properties:

$$1 = P\{X \in (-\infty, \infty)\} = \int_{-\infty}^{\infty} f(x)dx$$

    - All probability statements about *X* can be answered by *f(x):*

$$P\{a \le X \le b\} = \int_a^b f(x)dx\,, \quad P\{X = a\} = \int_a^a f(x)dx = 0$$

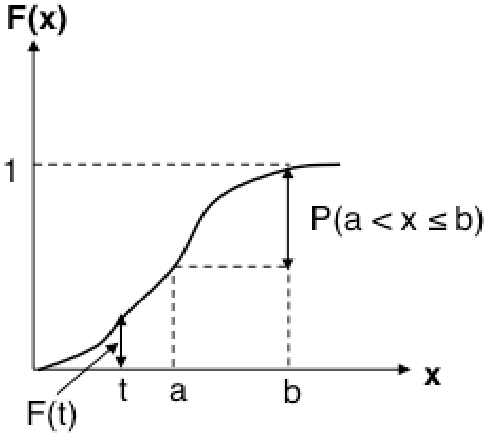


pdf

  - **Cumulative distribution function (CDF)**:

$$F_X(x) = P(X \le x) = \int_{-\infty}^{x} f_X(t)dt, \qquad -\infty < x < \infty$$

    - Properties: $\dfrac{d}{da}F(a) = f(a)$
    - **A right continuous function**



CDF

Continuous Random Variables

- Random variables whose set of possible values is uncountable

- *X* is a continuous random variable if there exists a nonnegative function *f(x)* defined for all real *x* Î (- ¥,¥), having the property that for any set *B* of real numbers,

$$P\{X \in B\} = \int_B f(x)dx$$

- *f(x)* is called the *probability density function (pdf)* of the random variable *X*

- The probability that *X* will be in *B* may be obtained by integrating the probability density function over the set *B*. Since *X* must assume some value, *f(x)* must satisfy

$$P\{X \in (-\infty,\infty)\} = \int_{-\infty}^{\infty} f(x)dx = 1$$

Continuous Random Variables Cont'd

- All probability statements about *X* can be answered in terms of *f(x)*
  e.g. letting *B*=[*a, b*], we obtain $P\{a \le X \le b\} = \int_a^b f(x)dx$
- If we let *a*=*b* in the preceding, then $P\{X = a\} = \int_a^a f(x)dx = 0$
- This equation states that the probability that a continuous random variable will assume any *particular* value is zero
- The relationship between the cumulative distribution *F(·)* and the probability density *f(·):*

$$F_X(a) = \; P(X \in (-\infty, a]) = \int_{-\infty}^{a} f_X(t)dt$$

- Differentiating both sides of the preceding yields

$$\frac{d}{da}F(a) = f(a)$$

Continuous Random Variables Cont'd

- That is, the density function is the derivative of the cumulative distribution function.

- A somewhat more intuitive interpretation of the density function

$$P\left\{a-\frac{\varepsilon}{2}\leq X\leq a+\frac{\varepsilon}{2}\right\}=\int_{a-\varepsilon/2}^{a+\varepsilon/2} f(x)dx \approx \varepsilon f(a)$$

  when ε is small

- The probability that X will be contained in an interval of length ε around the point a is approximately $\varepsilon f(a)$

An Example of Continuous RV - *Uniform Distribution*

- A continuous random variable X is said to have a uniform distribution over the interval (a, b) if its density is given by:

$$f(x) = \begin{cases} \dfrac{1}{b-a}, & a < x < b \\ 0, & \text{otherwise} \end{cases}$$

- And the distribution function is given by:

$$F(x) = \begin{cases} 0, & x < a \\ \dfrac{x-a}{b-a}, & \text{a} \leq x < b \\ 1, & x \geq b \end{cases}$$

Limit Theorems

- **Markov's Inequality:** If X is a random variable that takes only nonnegative values, then for any value $a>0$:

$$P\{X \geq a\} \leq \frac{E[X]}{a}$$

- **Chebyshev's Inequality:** If X is a random variable with mean μ and variance σ^2 then for any value $k>0$,

$$P\{|X-\mu| \geq k\} \leq \frac{\sigma^2}{k^2}$$

- **Strong law of large numbers:** Let $X_1, X_2, \ldots$ be a sequence of independent random variables having an identical distribution, and let $E[X_i]=\mu$. Then, almost surely,

$$\frac{X_1+X_2+\ldots+X_n}{n} \rightarrow \mu \text{ as } n \rightarrow \infty$$

Limit Theorems

- **Central Limit Theorem:** Let X_1 , X_2, ... be a sequence of independent, identically distributed random variables, each with mean μ and variance σ^2 then the distribution of

$$\frac{X_1+X_2+\cdots+X_n-n\mu}{\sigma\sqrt{n}} \rightarrow N(0,1) \text{ as } n \rightarrow \infty$$

- That is,

$$\mathrm{P}\left(\frac{X_1+X_2+\cdots+X_n-n\mu}{\sigma\sqrt{n}} \leq a\right) \rightarrow \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{a} e^{-x^2/2} dx \quad \text{as} \quad n \rightarrow \infty$$

- Note that like the other results, this theorem holds for any distribution of the X_i's ; herein lies its power.

Conditional Probability

- **Conditional Probability** *of A given B ($P(A/B)$)* defines the conditional probability of the event A given that the event B occurs and is given by:

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

if $\mathrm{P(B)} \neq 0$, and ***is undefined otherwise.***

- A rearrangement of the above definition gives the following ***multiplication rule (MR):***

$$P(A \cap B) = \begin{cases} P(B)P(A|B), & if\ P(B) \neq 0 \\ P(A)P(B|A), & if\ P(A) \neq 0 \\ 0, & otherwise \end{cases}$$

Theorem of Total Probability, Bayes Formula

- **Theorem of Total Probability**
  - Any event A can be partitioned into two disjoint subsets:

$$A = (A \cap B) \cup (A \cap \bar{B})$$

  - Then:

$$\begin{aligned} P(A) &= P(A \cap B) + P(A \cap \bar{B}) \\ &= P(A|B)P(B) + P(A|\bar{B})P(\bar{B}) \end{aligned}$$

  - In general, if $B_1, B_2, \dots, B_n$ partition S ,

$$\boxed{P(A) = \sum_{i=1}^{n} P(A \cap B_i) = \sum_{i=1}^{n} P(A|B_i)P(B_i)}$$

- **Bayes Formula:**

$$\boxed{P\left(B_j|A\right) = \frac{P(B_j \cap A)}{P(A)} = \frac{P\left(A\middle|B_j\right)P(B_j)}{\sum_i P(A|B_i)P(B_i)}}$$

Bayes Formula Example

- Suppose there is security software that monitors a system for suspicious activity
  - It raises a flag if, based on some recent activity, the probability that the user's account is compromised is greater than 50%
- Past data tells us that
  - P (normal user downloading executable file) = 0.2
  - P (hacker downloading executable files) = 0.8
  - P (account compromise) = 0.1
- **If a user recently downloaded an executable file onto the system, should the security software raise a flag?**
- A: event that an executable file is downloaded
  B: event that user's account is not compromised
- From past data, we know $P(A|B) = 0.2,\ \ P(A|\bar{B}) = 0.8,$
  $P(\bar{B}) = 0.1,\ \ P(B) = 1 - P(\bar{B}) = 0.9$
- Using Bayes formula, $P(\bar{B}|A) = \frac{0.8*0.1}{0.8*0.1+0.2*0.9} = \frac{0.08}{0.26} = 0.31 < 0.5$

Thus, the security software should not raise a flag

Independence of Events

- **Independence of Events:**
  - Two events A and B are independent iff:

$$\mathrm{P(A|B) = P(A)}$$

  - Alternately, two events A and B are said to be independent iff:

$$\mathrm{P(A \cap B) = P(A)P(B)}$$

HYPOTHESIS TESTING AND P-VALUES

Hypothesis Testing

- Many practical problems require decision making from the limited information contained in a sample

- To make a decision, we begin by formulating a **statistical hypothesis**, which is an assertion (i.e. guess) about the situation that may or may not be valid

- We use procedures called **statistical tests** to decide whether or not we will reject or accept these hypotheses

- For example,
  - A system engineer/administrator may want to know if any one of the memory manufacturers is better than others
  - A Clinician may want to determine the efficacy of new drug vs the current drug in treating a given disease

Important Terms

- **Population =** all possible elements from a dataset (e.g. all students at UIUC)

- **Sample** = a selection from the population (e.g. 10 students who responded to an online questionnaire)

- **Statistical inference** = process of generalizing a result from a sample to a population with calculated degree of certainty

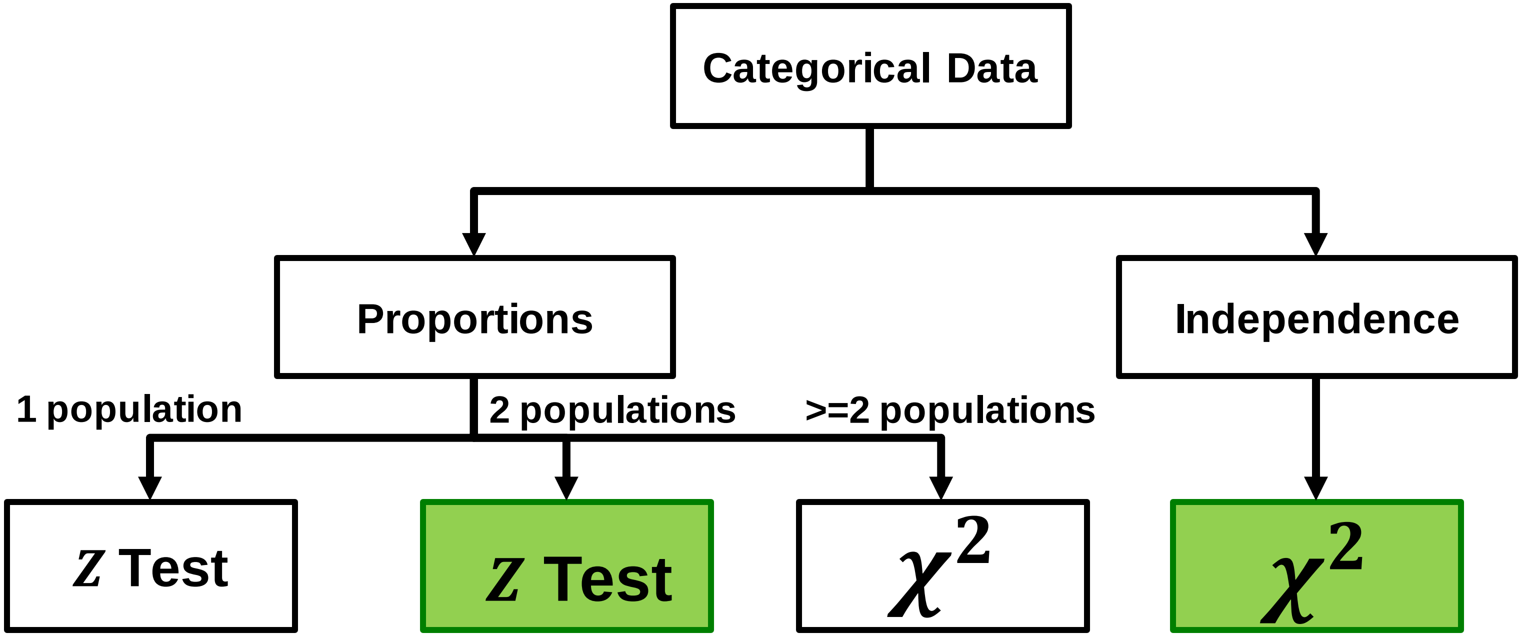
- Two forms of statistical inference
  - **Hypothesis testing**
  - **Estimation**

- **Parameter =** a characteristic of population (e.g. population mean μ)

- **Statistic =** calculated from data in the sample (e.g. sample mean $\bar{x}$)

Hypothesis Tests for Categorical Data

- We often perform hypothesis tests with categorical data
  - Quantitative random variables yield responses that can be put in categories
    - Example: Logic (True, False), Weather (Sunny, Rainy)
  - Data can be collected as continuous but recoded to categorical data
    - Example: electrical signal (On or Off)



Hypothesis Testing Steps

1. Develop null and alternative hypotheses
2. Calculate the test statistic
3. Calculate the p-value
4. Interpret p-value based on the significance level and make a decision

Step 1: Null and Alternative Hypotheses

- An important first step in hypothesis testing is to define a null and an alternative hypotheses to guide your analysis
- **Null Hypothesis ($\boldsymbol{H_0}$): S**tatement that some population parameter is equal to either (i) some fixed value or (ii) another population parameter
  - In the case of two populations, the claim is that there is no significant difference between two populations' parameters
  - Although usually a statement of equality, it can sometimes be expressed mathematically with $\leq$ or $\geq$ (**never with $\neq$**)
  - We test H_0 by first assuming that it's true and then reaching a conclusion that rejects it or fails to reject it
  - e.g. Consider two populations of students: Population A with students that get at least 8 hours of sleep each night, and Population B with those who don't. One plausible null hypothesis might be "There is no significant difference in academic performance between Populations A and B."

Null and Alternative Hypotheses

- **Alternate Hypothesis (H_a/H_1):** Statement that usually claims that the opposite of the null hypothesis is true
  - It can also serve as the claim that you wish to support with your research/study
  - Usually expressed mathematically with $\neq$, $<$, and $>$ (**never $=$**)
  - e.g. "There is a significant difference in academic performance between Populations A and B."
- Note that the only outcomes from analysis are **(i) rejecting H_0** or **(ii) failing to reject H_0**
  - Some textbooks may say "accept H_0", but in reality you just fail to reject H_0
  - By rejecting H_0, you can indirectly support the claim of H_a

Step 2: Z Test for Difference in Two Proportions

- A **population proportion** is a single numerical value that is used to represent the population.
  - It is often convenient to use proportions in statistical calculations
  - Common proportions include (i) the mean value in a dataset and (ii) the fraction of positive values over the size of the dataset
  - Which data points qualify as "positive" depends on the context of the analysis – for example, a single patient might be considered as a "positive" data point if they responded well to some treatment

- Z-Test motivation: We need a way to compare two different populations (using their proportions) that can yield statistically meaningful results
  - Null Hypothesis: There is no significant difference in the population proportions

- Z-Test assumptions:
  - Populations are independent
  - Populations follow binomial distribution
  - Normal approximation can be used for large samples (all expected counts > 5)

Z Test for Difference in Two Proportions

- Suppose that
  - Population 1 has
    - n_1 total data points
    - $X_1 < n_1$ positive data points
    - True proportion p_1
    - Estimated proportion $\hat{p}_1 = \frac{X_1}{n_1}$ (from the dataset)
  - Similarly, n_2, X_2, p_2, and $\hat{p}_2$ describe population 2
  - Let $\hat{p} = \frac{X_1+X_2}{n_1+n_2}$, $\hat{q} = 1 - \hat{p}$

- Then, the corresponding Z-statistic is

$$Z_{stat} = \frac{(\hat{p}_1 - \hat{p}_2) - (p_1 - p_2)}{\sqrt{\hat{p}.\hat{q}.\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}}$$

- Under the null hypothesis, $p_1 = p_2$ so

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

- Under the null hypothesis, the corresponding Z distribution is

$$Z_{dist} \cong N\left(0, \sqrt{\hat{p}.\widehat{q}.\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}\right)$$

Step 3: P-value

- The *P*-value answers the question: **What is the probability of the observed test statistic or a more extreme one <u>when H_0 is true</u>?**
- This corresponds to the AUC in the tail of the Standard Normal distribution beyond z_{stat}.

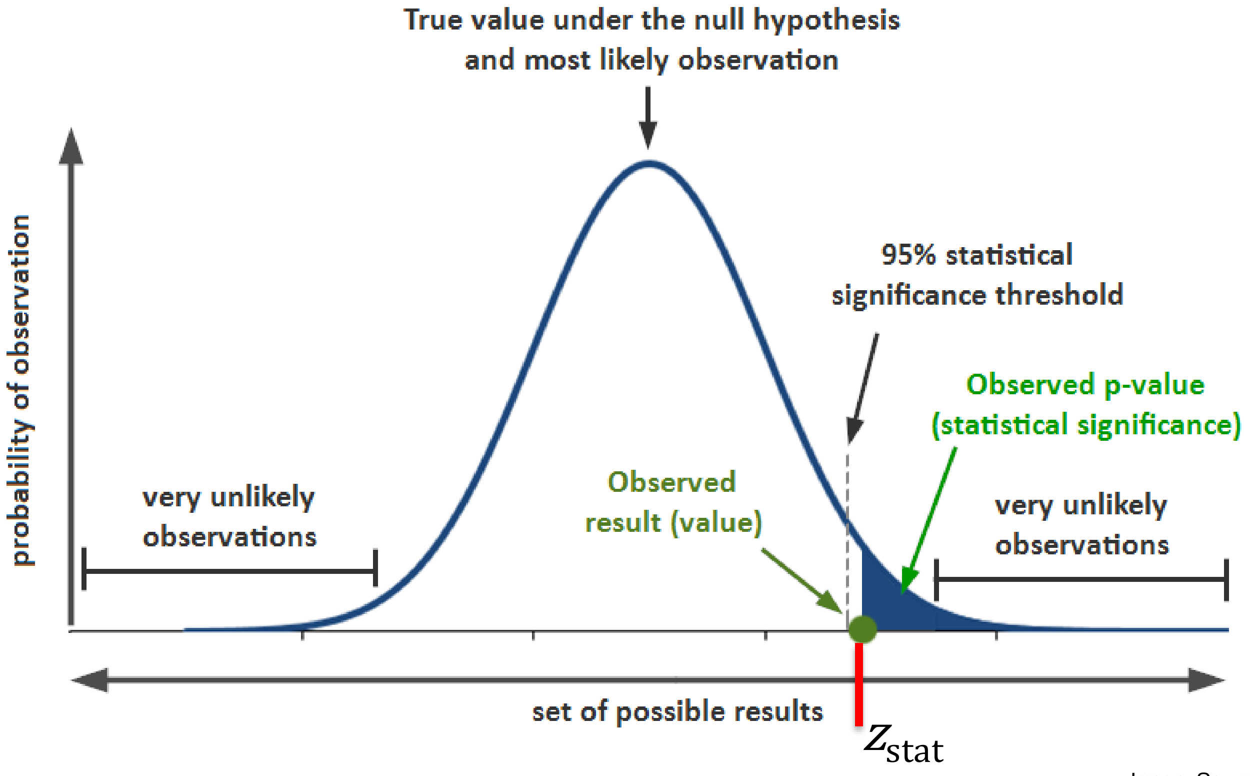


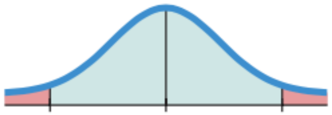
Image Source: https://www.simplypsychology.org/p-value.png.

P-value: Tails

- Convert z statistics to P-value :

For H_a: $\mathrm{p}_1 \neq \mathrm{p}_2$

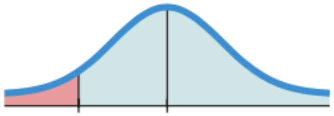
$$\Rightarrow P = 2 \times \text{one-tailed } P\text{-value}$$



Sign used in H_1: ≠
Two-tailed test

For H_a: $\mathrm{p}_1 < \mathrm{p}_2$

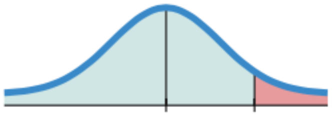
$$\Rightarrow P = \Pr(\mathrm{Z} < z_{\mathrm{stat}}) = \text{left tail beyond } \mathrm{z}_{\mathrm{stat}}$$



Sign used in H_1: <
Left-tailed test

For H_a: $\mathrm{p}_1 > \mathrm{p}_2$

$$\Rightarrow P = \Pr(\mathrm{Z} > z_{\mathrm{stat}}) = \text{right-tail beyond } \mathrm{z}_{\mathrm{stat}}$$



Sign used in H_1: >
Right-tailed test

Image Source: Triola – *Elementary Statistics* – 11th ed.

Step 4: α-Level (significance)

- Let α = probability of erroneously rejecting H_0
  - Set α threshold (e.g., let $\alpha = .10, \mathbf{.05}, \mathit{or\ other}$)
  - Ethical issues: don’t adjust α for the sake of getting better results
- Actions:
  - Reject H_0 when $P \leq \alpha$
  - Fail to reject H_0 when $P > \alpha$
- Example:
  - Set $\alpha = .10$. Find $P = 0.27 \Rightarrow$ fail to reject H_0
  - Set $\alpha = .01$. Find $P = .001 \Rightarrow$ reject H_0

Interpretation Language

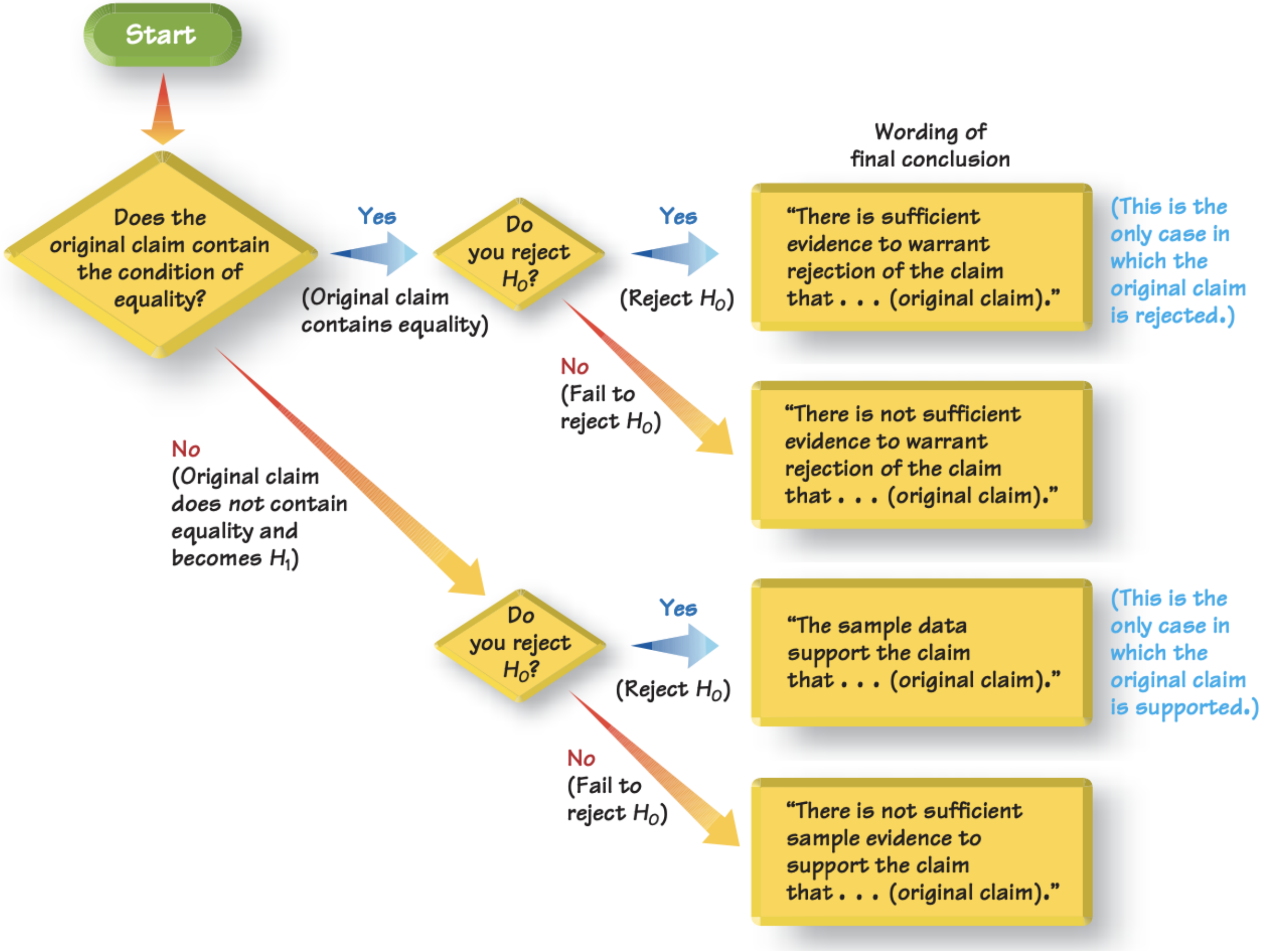

Start
Does the original claim contain the condition of equality?
Yes
(Original claim contains equality)
Do you reject H_0?
Yes
(Reject H_0)
Wording of final conclusion
"There is sufficient evidence to warrant rejection of the claim that . . . (original claim)."
(This is the only case in which the original claim is rejected.)
No
(Fail to reject H_0)
"There is not sufficient evidence to warrant rejection of the claim that . . . (original claim)."
No
(Original claim does not contain equality and becomes H_1)
Do you reject H_0?
Yes
(Reject H_0)
"The sample data support the claim that . . . (original claim)."
(This is the only case in which the original claim is supported.)
No
(Fail to reject H_0)
"There is not sufficient sample evidence to support the claim that . . . (original claim)."

Image Source: Triola – *Elementary Statistics* – 11th ed.

Z Test for Comparing Drug Efficacy

- We are studying the efficacy of two new drugs for treating depression.
- First drug was administered to 100 patients, 30 of which got cured.
- Second drug was administered to 200 patients, 85 of which got cured.
- **At .05 level, is the second drug better than the first drug?**

Let p_i be the efficacy of drug i.

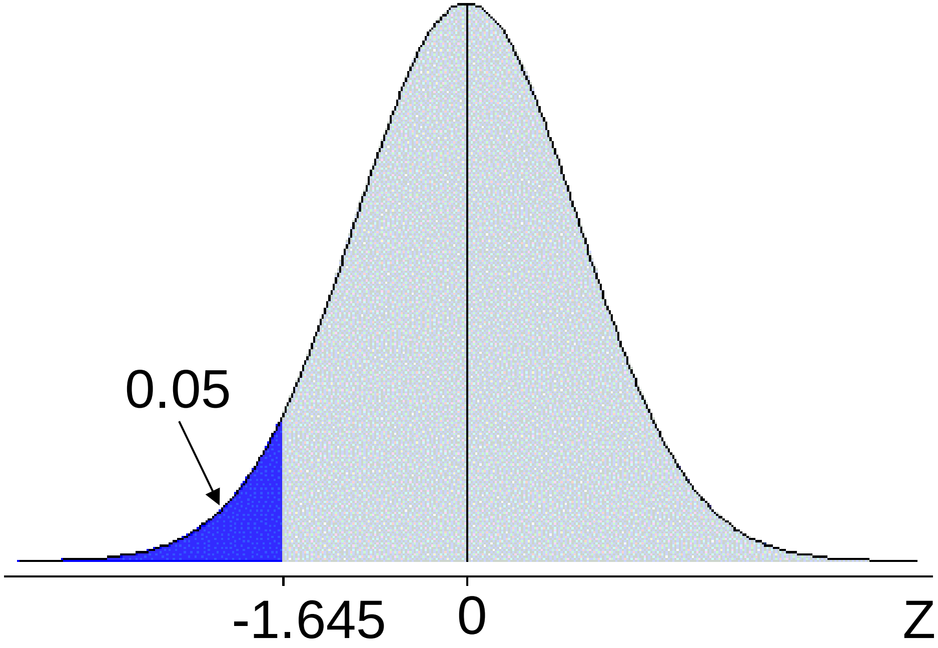
H_0: $p_1 - p_2 \geq 0$

H_a: $p_1 - p_2 < 0$

$\alpha = 0.05$

$n_1 = 100\,, n_2 = 200$

We want probability in the tail to be less than 0.05, so Z < -1.645



Z-stat calculation

$$\hat{p}_1 = \frac{30}{100} = 0.3$$

$$\hat{p}_2 = \frac{85}{200} = 0.425$$

$$\hat{p} = \frac{30 + 85}{100 + 200} = \frac{115}{300} = 0.383$$

$$Z = \frac{(0.3\ - 0.425) - 0}{\sqrt{(0.383)(1 - 0.383)\left(\frac{1}{100} + \frac{1}{200}\right)}} = \frac{-0.125}{0.0595} = -2.1$$

Z-test for two proportions solution

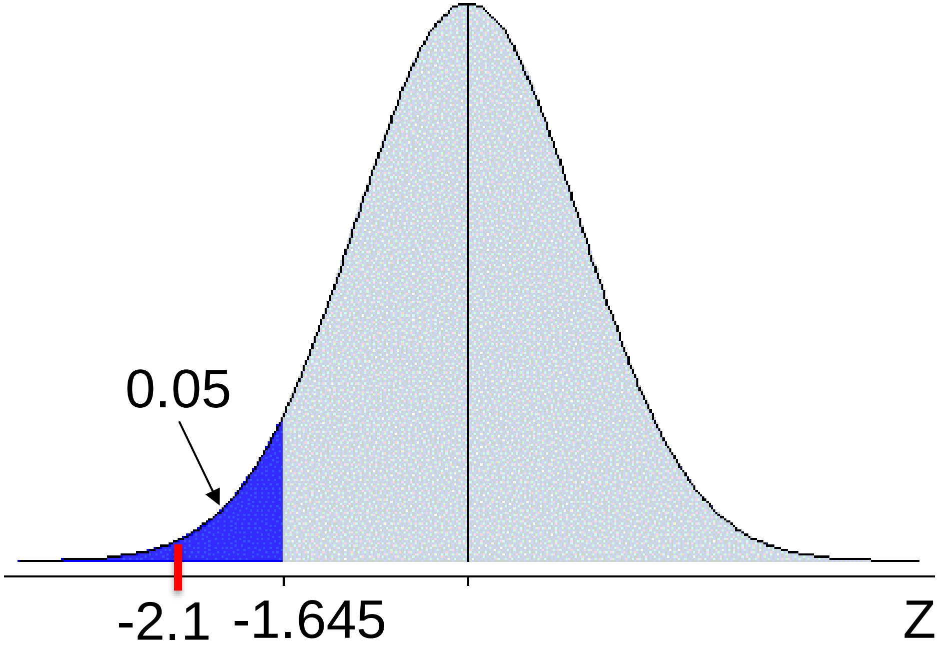
Let p_i be the efficacy of drug i.

H_0: $p_1 - p_2 \geq 0$; Ha: $p_1 - p_2 < 0$
$\alpha = 0.05$
$n_1 = 100$, $n_2 = 200$

Test statistic: $Z =$ -2.1

Decision:
Z = -2.1 < -1.645
Reject at $\alpha = 0.05$

Conclusion:
Drug 2 is better than Drug 1



χ^2 Test for Independence

- Test the independence of two categorical variables

- Assumptions
  - The sampling method is simple random sampling.
  - The variables under study are each categorical.
  - If sample data are displayed in a contingency table, the expected frequency count for each cell of the table is at least 5

- Hypothesis
  - H_0: In the population, the two categorical variables are independent.
    H_a: In the population, two categorical variables are dependent

- Gather data and summarize in the two-way contingency table
  - Represents the observed counts and is called the **Observed Counts Table**

Steps in χ^2 Test for Independence

1. Develop null and alternative hypotheses
2. Calculate the test statistic
3. Analyze the sample data
4. Calculate the P-value
5. Interpret P-value based on significance level and make a decision

χ^2 Test for Comparing AV performance by weather

- An AV manufacturer wants to know whether there is a statistically significant dependency between AV disengagement (first categorical variable) and the weather conditions while driving (second variable).
- For that, they count the number of miles driven with and without disengagement.
- A mile is counted as disengaged if there was at least one disengagement within it.

	Sunny	**Rainy**	**Snow**	total
Disengagement	28	59	34	121
No Disengagement	821	465	283	1569
total	849	524	317	1690

Hypothesis Formulation

- **State the hypotheses.** The first step is to state the null hypothesis and an alternative hypothesis.
  - H_0: The categorical variables, disengagement and weather, are independent
  - H_a: Disengagement and weather are not independent

- Test Statistic: chi-squared

- Significance level: $\alpha = 0.05$

Analyze Sample Data

"Contingency Table"

	Sunny	**Rainy**	**Snow**	total
Disengagement	28	59	34	121
No Disengagement	821	465	283	1569
total	849	524	317	1690

Degrees of freedom (DF):

$$\text{DF} = (r - 1)*(c - 1)$$

where r and c are levels of categorical variables

Expected frequencies:

$$E_{rc} = \frac{n_r \times n_c}{n}$$

Test Statistic:

$$\chi^2 = \sum \frac{(O_{rc} - E_{rc})^2}{E_{rc}}$$

Analyze Sample Data

DF = (r - 1) * (c - 1) = (2 - 1) * (3 - 1) = 2

$E_{r,c} = (n_r * n_c)/n$

"Expectation Table"

	Sunny	Rainy	Snow
Disengagement	60.8	37.5	22.7
No Disengagement	788.2	486.5	294.3

$$\chi^2 = \sum \frac{(O_{rc} - E_{rc})^2}{E_{rc}}$$

$$= \frac{(28-60.8)^2}{60.8} + \frac{(59-37.5)^2}{37.5} + \frac{(34-22.7)^2}{22.7} + \frac{(821-788.2)^2}{788.2} + \frac{(524-486.5)^2}{486.5} + \frac{(283-294.3)^2}{294.3}$$

$$= 38.36$$

P-Value

- Using a lookup table, we see that the p-value (the probability that a chi-square statistic having 2 degrees of freedom is more extreme than 38.86) is

$$P(\chi^2 \ > 38.86) = 4.7 \times 10^{-9}$$

- This p-value (4.7×10^{-9}) is less than the significance level (α =0.05)
  - Reject the null hypothesis
  - **Conclude that there is a relationship between AV disengagement and weather condition**

Kolmogorov Smirnov Test

- Till now we have seen statistical tests that compare the means and/or proportions of distributions

- Not all distributions are completely summarized by their mean; therefore we need tests to compare distributions

- Suppose we have observations $X_1, X_2 \ldots X_n$ which we think come from a distribution P.

- The Kolmogorov-Smirnov (KS) Test is used for the following hypothesis test:

$$\mathrm{H}_0\text{: the samples come from } P$$
$$\mathrm{H}_\mathrm{a}\text{: the samples do not come from } P$$

KS Test: Test statistic

- The CDF uniquely characterizes a probability distribution.
- Suppose $F(x)$ is the CDF if the samples are generated from the probability distribution P. Then

$$F(x) = P(X < x),$$

- Now, let $F_n(x)$ be the empirical CDF that is calculated from the samples using the indicator function

$$F_n(x) = \frac{1}{n}\sum_{i=1}^{n} \mathbb{I}[X_i \leq x]$$

$$\mathbb{I}[X_i \leq x] = \begin{cases} 1, & X_i \leq x \\ 0, & X_i > x \end{cases}$$

- The test statistic is:

$$D_n = \max_x |F_n(x) - F(x)|$$

  - We are just looking at the max “distance” between the distributions!

KS Test: Critical value and Conclusion

- At the 95% level, the critical value is approximately given by

$$D_{crit,0.05} = \frac{1.36}{\sqrt{n}}$$

- Fail to reject the null hypothesis if,

$$D_n < D_{crit,0.05}$$

- Intuition: If the samples are drawn from the probability distribution P, then the CDF and the empirical CDF would be “close” to each other.

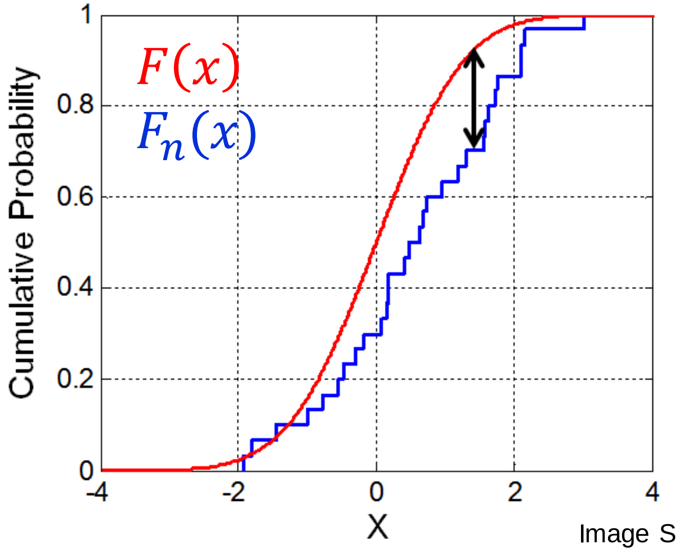


Image Source: https://commons.wikimedia.org/w/index.php?curid=25222928

ADDITIONAL SLIDES

COMMON PROBABILITY DISTRIBUTIONS

Binomial Theorem

- **Combinatorial Problems**
  - Permutations with replacement:
    - Ordered samples of size k, with replacement *P(n, k)*
  - Permutations without replacement
    - Ordered Samples of size k, without replacement

$$n(n-1)\ldots(n-k+1)=\frac{n!}{(n-k)!}\qquad k=1,2,\ldots,n$$

  - Combinations
    - Unordered sample of size k, without replacement

- **Binomial Coefficient** $\binom{n}{k}=\frac{n!}{k!\,(n-k)!}$

- **Binomial Theorem** $(x+y)^n=\sum_{k=0}^{n}\binom{n}{k}x^k y^{n-k}$

Bernoulli and Binomial Distributions

- *X* is said to be a *Bernoulli* random variable if its probability mass function is given by the equation below for some $p \in [0,1]$, where is the probability that the trial is a success

$$p(0) = P\{X = 0\} = 1 - p$$

$$p(1) = P\{X = 1\} = p$$

- In *n* independent trials, each of which results in a "success" with *p* and in a "failure" with probability *1-p,*
- If *X* represents the number of successes that occur in the *n* trials, *X* is said to be a *binomial* random variable with parameters (*n*, *p*) and its PMF is given by:

$$p(i) = \binom{n}{i} p^i (1-p)^{n-i}, \qquad \text{where} \binom{n}{i} = \frac{n!}{i!\,(n-i)!}$$

Binomial Random Variable Example 1

- In Alzheimer's disease (AD), the ability to retrieve memories gets compromised due to deterioration in brain health. The process of retrieving memories can be thought of as being probabilistic, and the probability reduces in AD.

- Researchers found that in AD, the probability of successfully retrieving a memory of a previously viewed picture is 0.1. Assume that the retrieval of a memory is independent of retrieval of any other memory.

- What is the probability that out of three pictures stored in memory, at most one will be retrieved?

- If X is the number of pictures that are retrieved, then X is a binomial random variable with parameters (3, 0.1). Hence, the desired probability is given by:

$$P\{X=0\}+P\{X=1\}=\binom{3}{0}(0.1)^0(0.9)^3+\binom{3}{1}(0.1)^1(0.9)^2=0.972$$

Poisson Distribution

- A random variable X, taking on one of the values 0,1,2,…, is said to be a *Poisson* random variable with parameter λ, if for some $\lambda > 0$,

$$p(i) = P\{X = i\} = e^{-\lambda}\frac{\lambda^i}{i!}, \qquad i = 0,1,\ldots$$

defines a probability mass function since

- **Relationship to Exponential Distribution**

If the number of arrivals in an interval *t* is Poisson distributed with parameter λ, the inter-arrival times will be Exponentially distributed with parameter λ.

- **Intuition**: Within certain time, how many events have happened
- **Example**: Number of failures occurred in a certain component up to time *t*

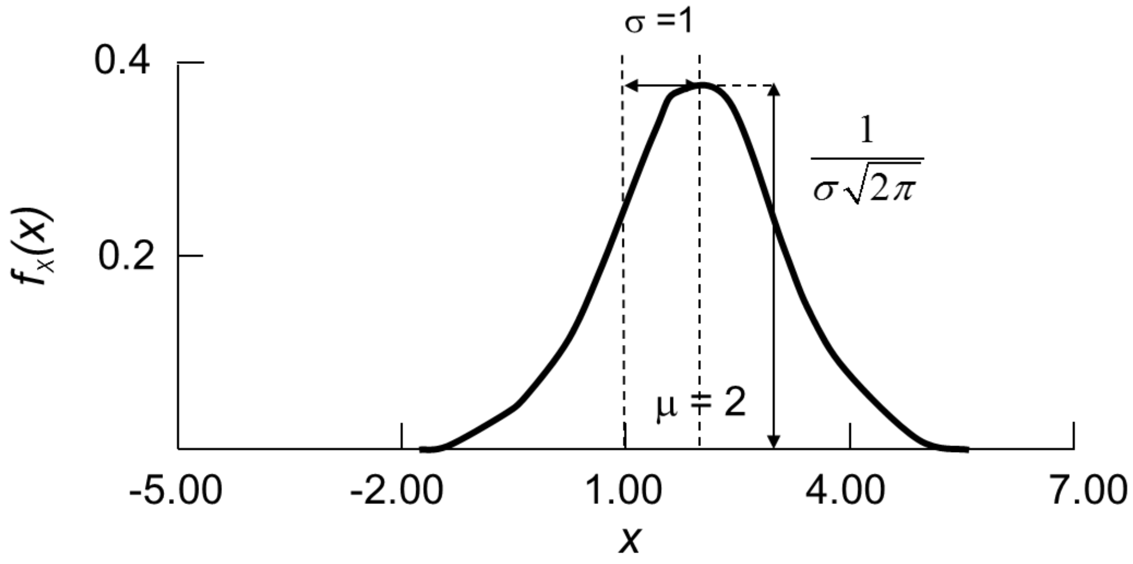
Normal/Gaussian Distribution

- The normal density is given by:

$$f(x) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left[-\frac{(x-\mu)^2}{2\sigma^2}\right], \qquad -\infty < x < \infty$$

where μ and σ are two parameters of the distribution.

- Normal density with parameters $\mu = 2$ and $\sigma = 1$



Standard Normal Distribution

- The distribution function *F(x)* has no closed form, so between every pair of limits *a* and *b*, probabilities relating to normal distributions are usually obtained numerically and recorded in special tables.

- These tables apply to the **standard normal distribution** $Z \sim N(0,1)$
  -- a normal distribution with parameters μ = 0 , σ = 1
  -- Their entries are the values of:

$$F_Z(z) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{z} e^{-t^2/2} dt$$

The Exponential Distribution

- The exponential distribution occurs in applications such as reliability theory and queuing theory. Reasons for its use include:
  - Its memoryless (Markov) property
  - Its relation to the (discrete) Poisson distribution

- The following random variables will often be modeled as exponential:
  - Time between two successive job arrivals to a computing center
  - Service time at a server in a queuing network
  - Time to failure (lifetime) of a component
  - Time required to repair a component that has malfunctioned

Exponential Distribution

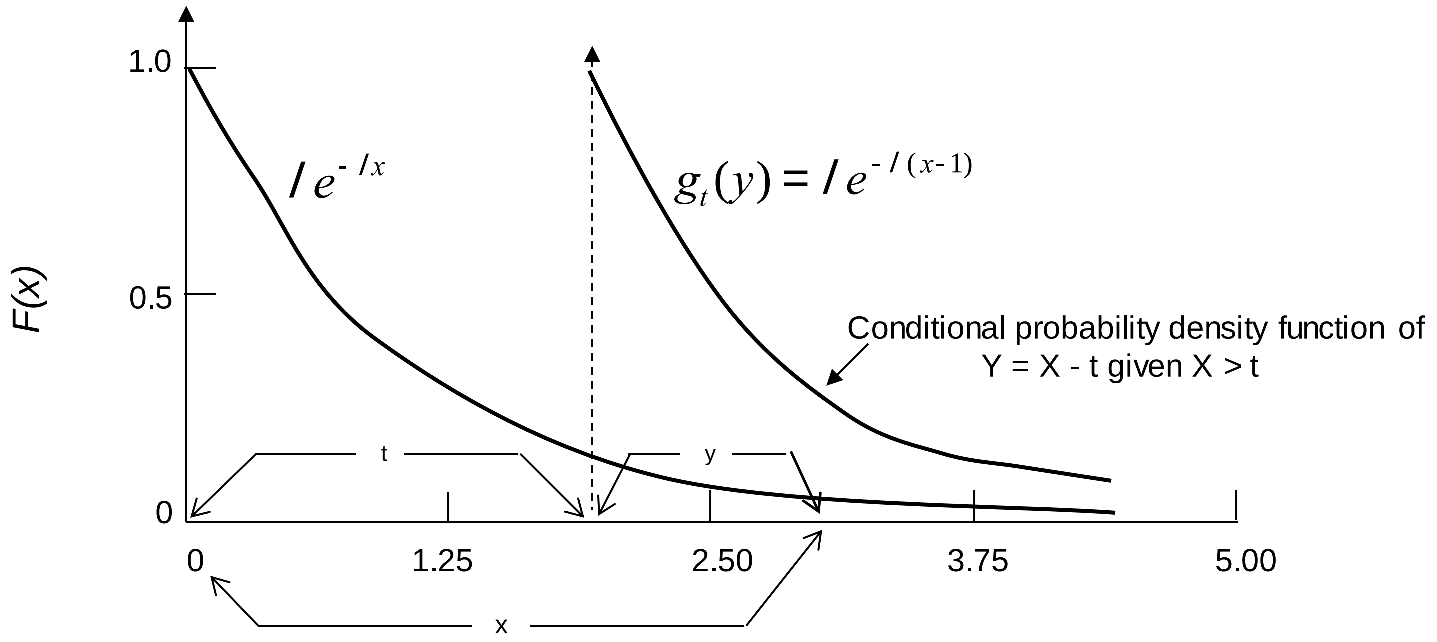
- The CDF of the exponential distribution is given by:

$$F(x) = \begin{cases} 1 - e^{-\lambda x}, & \text{if } 0 \leq x \\ 0, & \text{otherwise} \end{cases}$$

- If the CDF of a random variable X is given by the above equation, we use the notation $X \sim EXP(\lambda)$, for brevity. The pdf of X is given by:

$$f(x) = \begin{cases} \lambda e^{-\lambda x}, & if\ x > 0 \\ 0, & \text{otherwise} \end{cases}$$

Memory-less Property of Exponential



Conditional probability density function of Y = X - t given X > t

Gamma Distribution

- Two-parameter family of continuous probability distributions
  - Exponential, Erlang, and chi-squared distributions are special cases of Gamma distribution
- Shape α and rate β

$$X \sim \Gamma(\alpha, \beta) \equiv Gamma(\alpha, \beta)$$

PDF: $f(x; \alpha, \beta) = \frac{\beta x^{\alpha-1} e^{-\beta x}}{\Gamma(\alpha)}$, where, $\Gamma(\alpha)$ is the gamma function

$$\Gamma(z) = \begin{cases} (z-1)!, & \text{positive integer} \\ \int_0^\infty x^{z-1} e^{-x}\, dx\,, & \text{otherwise} \end{cases}$$

- Intuition: how much time it takes for α events to happen
- Example: the number of requests on web servers

Beta Distribution

- Two-parameter continuous probability distributions defined on [0,1]
  - A special case of the Dirichlet distribution

- Shape parameters α and β

$$PDF\text{:}\ f(x;\alpha,\beta) = \frac{1}{B(\alpha,\beta)}\, x^{\alpha-1}(1-x)^{\beta-1}$$

where $\mathrm{B}(\alpha,\beta)$ is the beta function: $B(\alpha,\beta) = \dfrac{\Gamma(\alpha)\Gamma(\beta)}{\Gamma(\alpha+\beta)}$

($\Gamma(\cdot)$ is the gamma function shown in last slide)

- Intuition: the likelihood of simulated Bernoulli experiments (with a finite sequence of probabilities) that agrees with the observation
- Example: Density of product rating

Weibull Distribution

- Two-parameter continuous probability distributions
  - λ defines scale; k defines shape
- Shape parameters α and β

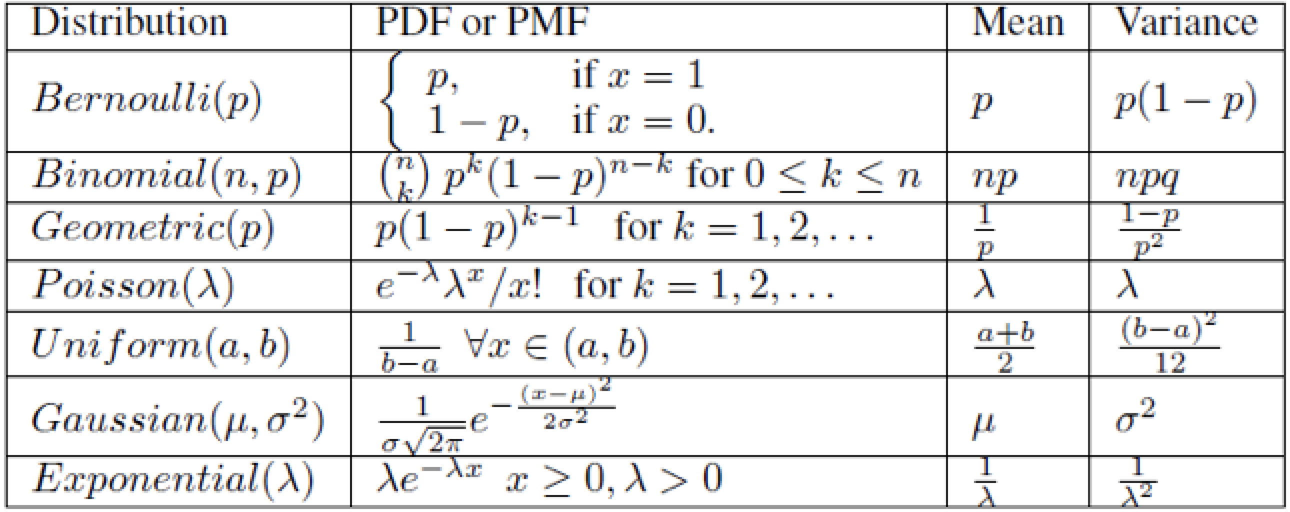
$$pdf: f(x;\lambda,k) = \begin{cases} \frac{k}{\lambda}\left(\frac{x}{\lambda}\right)^{k-1} e^{-\left(\frac{x}{\lambda}\right)^{k}} & (x \geq 0) \\ 0 & (x < 0) \end{cases}$$

  - Shape parameter k:
    - $k < 1$: failure rate decreases over time
    - $k = 1$: failure rate is constant over time
    - $k > 1$: failure rate increases over time
- Example: time to failure; reaction time for an AV disengagement

https://en.wikipedia.org/wiki/Weibull_distribution

General Analysis

- **Summary of important distributions:**



Distribution	PDF or PMF	Mean	Variance
$Bernoulli(p)$	$\begin{cases} p, & \text{if } x = 1 \\ 1-p, & \text{if } x = 0. \end{cases}$	p	$p(1-p)$
$Binomial(n,p)$	$\binom{n}{k}\ p^k(1-p)^{n-k}$ for $0 \leq k \leq n$	np	npq
$Geometric(p)$	$p(1-p)^{k-1}$ for $k = 1, 2, \ldots$	$\frac{1}{p}$	$\frac{1-p}{p^2}$
$Poisson(\lambda)$	$e^{-\lambda}\lambda^x/x!$ for $k = 1, 2, \ldots$	λ	λ
$Uniform(a,b)$	$\frac{1}{b-a}\ \ \forall x \in (a,b)$	$\frac{a+b}{2}$	$\frac{(b-a)^2}{12}$
$Gaussian(\mu,\sigma^2)$	$\frac{1}{\sigma\sqrt{2\pi}}e^{-\frac{(x-\mu)^2}{2\sigma^2}}$	μ	σ^2
$Exponential(\lambda)$	$\lambda e^{-\lambda x}\ \ x \geq 0, \lambda > 0$	$\frac{1}{\lambda}$	$\frac{1}{\lambda^2}$

JOINTLY DISTRIBUTED RANDOM VARIABLES

Joint Distribution Functions

- **Joint distribution functions:**
  - For any two random variables X and Y, the ***joint cumulative probability distribution function*** of X and Y is:

$$F(a, b) = P\{X \leq a, Y \leq b\}, \qquad -\infty < a, b, < \infty$$

  - **Discrete:**
    - The ***joint probability mass function*** of X and Y

$$p(x, y) = P\{X = x, Y = y\}$$

    - Marginal PMFs of X and Y:

$$p_X(x) = \sum_{y:\, p(x,y)>0} p(x, y)$$

$$p_Y(y) = \sum_{x:\, p(x,y)>0} p(x, y)$$

Joint Distribution Functions

- **Joint distribution functions:**
  - **Continuous:**
    - The ***joint probability density function*** of X and Y:

$$P\{X \in A, Y \in B\} = \int_B \int_A f(x,y)dx\, dy$$

    - Marginal PDFs of X and Y:

$$f_X(x) = \int_{-\infty}^{\infty} f(x,y)dy$$

$$f_Y(y) = \int_{-\infty}^{\infty} f(x,y)dx$$

    - Relation between joint CDF and PDF

$$F(a,b) = P(X \le a, Y \le b) = \int_{-\infty}^{a}\int_{-\infty}^{b} f(x,y)dy\, dx$$

Conditional and Marginal Distribution Functions

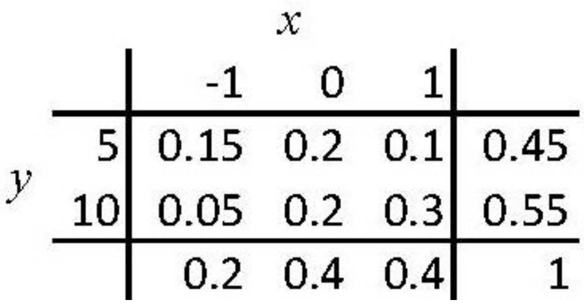
- Joint distribution: $f(x, y) = P[X = x, Y = y]$
- Marginal distributions:

$$f(x) = \sum_{all\ y} f(x, y) \qquad\qquad f(y) = \sum_{all\ x} f(x, y)$$

- Conditional distributions:

$$f(y|x) = \frac{f(x,y)}{f(x)} \qquad\qquad f(x|y) = \frac{f(x,y)}{f(y)} = \frac{f(y|x)f(x)}{y(y)}$$

- Example



		x			
		-1	0	1	
y	5	0.15	0.2	0.1	0.45
	10	0.05	0.2	0.3	0.55
		0.2	0.4	0.4	1

Independent Random Variables

- **Independent Random Variables:** Two random variables X and Y are said to be independent if:

$$F(x, y) = F_X(x)F_Y(y), -\infty < x < \infty, -\infty < y < \infty$$

- If X and Y are continuous:

$$f(x, y) = f_X(x)f_Y(y), -\infty < x < \infty, -\infty < y < \infty$$

- If X is discrete and Y is continuous:

$$P(X = x, Y \pounds y) = p_X(x)f_Y(y), \text{ all } x \text{ and } y$$

Moments

- Recall the definition of moments:

$$E[X^n] = \begin{cases} \sum x^n p(x), & if\ X\ is\ discrete \\ \int_{-\infty}^{\infty} x^n f(x) df\,, & if\ X\ is\ continuous \end{cases}$$

- Moments convey important information about the distribution of the variable
  - First moment = mean/expected value
  - First and second moments are used to calculate variance, which measures the expected square of the deviation of a random variable X from its mean

$$Var(X) = E[X^2] - (E[X])^2$$

  - In general, moments are especially important when considering joint distributions, which can be much more complex than distributions of individual random variables

Covariance

- The ***covariance*** of any two random variables, X and Y, denoted by $Cov(X,Y)$, is defined by

$$\begin{aligned} Cov(X,Y) &= E[(X-E[X])(Y-E[Y])] \\ &= E\big[XY - YE[X] - XE[Y] + E[X]E[Y]\big] \\ &= E[XY] - E[Y]E[X] - E[X]E[Y] + E[X]E[Y] \\ &= E[XY] - E[X]E[Y] \end{aligned}$$

- Covariance generalizes variance, in the sense that ***$Var(X) = Cov(X,X)$.***
- If either X or Y has mean zero, then E[XY] = Cov(X,Y).

- If X and Y are ***independent*** then it follows that ***$Cov(X,Y) = 0$***
- **But the converse is not true:**
  - **Cov(X,Y) = 0** means X and Y are **uncorrelated, but it doesn't imply** that X and Y are **independent.**

Properties of Covariance

- For any random variable X, Y, Z, and constant c, we have:
  1. Cov(X,X) = Var(X),
  2. Cov(X,Y) = Cov(Y,X),
  3. Cov(cX,Y) = cCov(X,Y),
  4. Cov(X,Y+Z) = Cov(X,Y) + Cov(X,Z).
  - Whereas the first three properties are immediate, the final one is proven as follows:

$$\begin{aligned} Cov(X,Y+Z) &= E[X(Y+Z)] - E[X]E[Y+Z] \\ &= E[XY] - E[X]E[Y] + E[XZ] - E[X]E[Z] \\ &= Cov(X,Y) + Cov(X,Z) \end{aligned}$$

- The last property generalizes to give the following result:

$$Cov\left(\sum_{i=1}^{n} X_i, \sum_{j=1}^{m} Y_j\right) = \sum_{i=1}^{n}\sum_{j=1}^{m} Cov(X_i, Y_j)$$

Correlation Coefficient

- The correlation between two random variable X and Y is measured using the ***correlation coefficient***:

$$\rho_{X,Y} = \frac{Cov(X,Y)}{\sqrt{Var(X)Var(Y)}} = \frac{Cov(X,Y)}{\sigma_X \sigma_Y}$$

($\rho_{X,Y}$ is well defined if $\mathrm{Var(X)} > 0$ and $\mathrm{Var(Y)} > 0$)

- If ***Cov(X,Y) = 0,*** X and Y are ***uncorrelated,***

  ⇒ ***E[XY] = E[X] E[Y].***

- If ***Cov(X,Y) > 0***, X and Y are ***positively correlated,***

  ⇒ ***Y*** **tends to increase as** ***X*** **increases.**

- If ***Cov(X,Y) < 0***, X and Y are ***negatively correlated,***

  ⇒ ***Y*** **tends to decrease as** ***X*** **increases.**

BINARY HYPOTHESIS TESTING

Binary Hypothesis Testing

- **“**Binary” hypothesis testing involves consideration of two hypotheses
  - **Null Hypothesis (H_0):** Assumes that any observations/patterns in the data are due to random chance
    - e.g. “There is no significant difference in academic performance when a student gets 8 hours of sleep at night.
  - **Alternate Hypothesis (H_1):** An alternate hypothesis to the null hypothesis
    - e .g. “There is a significant difference in GPA when a student gets 8 hours of sleep at night.”

- Two commonly reported probabilities used as error metrics are
  - $\boldsymbol{p_{false\ alarm}}$: Comes from rejecting a true null hypothesis
    - e.g. A doctor incorrectly reports a fever, when the patient in fact has a healthy temperature
  - $\boldsymbol{p_{miss}}$: Comes from accepting a false null hypothesis
    - e.g. A doctor incorrectly reports no fever, when the patient in fact does have one

Binary Hypothesis Testing (ML)

- **Maximum Likelihood (ML) Decision Rule**
  - Declares the hypothesis which maximizes the probability (or likelihood) of the probability
    - Likelihood Ratio Test (LRT)

$$\Lambda(k) = \frac{p_1(k)}{p_0(k)} \qquad \Lambda(k) = \begin{cases} > 1 \text{ declare } \mathrm{H}_1 \text{ is true} \\ \leq 1 \text{ declare } \mathrm{H}_0 \text{ is true} \end{cases}$$

    - More generally,

$$\Lambda(k) = \begin{cases} > \tau \text{ declare } \mathrm{H}_1 \text{ is true} \\ \leq \tau \text{ declare } \mathrm{H}_0 \text{ is true} \end{cases}$$

  - As τ increases, fewer observations lead to decide H_1 to be true, thus $p_{false\,alarm}$ decreases and p_{miss} increases.
  - Therefore, τ can be applied to select an operating point on the tradeoff between the two error probabilities

Binary Hypothesis Testing Example

- To assist doctors in the diagnosis of breast cancer, researchers have created a classification model that uses the image of breast tissue from biopsies.
- **Model must decide whether the image is of normal or cancerous tissue**
- Model assigns the following probabilities after receiving tissue biopsy image from a new patient:
  - P (image | normal tissue) = 0.3
  - P (image | cancerous tissue) = 0.5
- If the decision threshold $\tau = 1$ for this model, then using the ML rule, will it report that the patient has cancer?

H_0: The patient does not have cancerous tissue, $p_0(k) = 0.3$

H_1: The patient has cancerous tissue, $p_1(k) = 0.5$

$$\Lambda(k) = \frac{p_1(k)}{p_0(k)} = \frac{0.5}{0.3} = 1.67 > 1 = \tau$$

Therefore, the patient has cancer.

Binary Hypothesis Testing (MAP)

- **Maximum a Posteriori (MAP) Decision Rule**
  - Prior, observation $\xRightarrow{Bayes\ formula}$ Posterior
  - By Bayes Formula,

$$P(H_1|X=k) = \frac{P(H_1, X=k)}{P(X=k)} = \frac{P(H_1, X=k)}{P(H_1, X=k) + P(H_0, X=k)}$$

$$= \frac{\overset{\textbf{observation}}{\boxed{P(X=k|H_1)}} \cdot \overset{\textbf{prior}}{\boxed{P(H_1)}}}{P(H_1, X=k) + P(H_0, X=k)}$$

  - Priors: $\pi_0 = P(H_0), \pi_1 = P(H_1)$
  - The MAP rule declares hypothesis H_1 true if $\pi_1 p_1(k) > \pi_0 p_0(k)$, or if $\Lambda(k) > \frac{\pi_0}{\pi_1}$
  - Therefore, the MAP rule is equivalent to the ML rule with threshold $\tau = \frac{\pi_0}{\pi_1}$
  - The MAP rule minimizes $p_e = \pi_0 p_{false\ alarm} + \pi_1 p_{miss}$