HCI: Empirical Research Methods

Learning Objective

- In the previous lecture module, we have already learnt several evaluation methods such as heuristic evaluation, cognitive walkthroughs or cognitive models to evaluate the interactive system designs at the early phases
- As we have already mentioned, interactive system design is not complete unless it is evaluated with real users at the end

Learning Objective

- In this lecture, we shall discuss user evaluation methods
- In particular, we shall discuss the following:
  - The Key Concerns in User Evaluation
  - Data Collection Procedure
  - Data Analysis Techniques

Empirical Research

- Empirical research is broadly defined as the "observation-based investigation" seeking to discover and interpret facts, theories, or laws
- Collection and Analysis of the end user data for determining the usability of an interactive system is an “observation-based investigation”, hence it is qualified as an empirical research study

Themes of Empirical Research

- Generally, Empirical Research is based on Three Themes
  - Answer and Raise Questions (Testable Research Questions) about a new or existing UI Design or Interaction Method
  - Observe and Measure (Ratio Measurement Scale is preferable)
  - User Studies

Research Question

- It is very important in an empirical research to formulate “appropriate and relevant” research questions
- For example, consider some questions about a system
  - Is it viable?
  - Is it as good as or better than current practice?
  - Which of several design alternatives is best?
  - What are its performance limits and capabilities?
  - What are its strengths and weaknesses?
  - How much practice is required to become proficient?

Testable Research Question

- Preceding questions, while unquestionably relevant, but are not *testable research questions*
- We have to come-up with testable research questions
- Let's illustrate the idea with the following example:

Suppose we have designed a new text entry technique for mobile phones. We think that the design is good. In fact, we feel that our method is better than the most widely used current technique, multi-tap. We decide to undertake some empirical research to evaluate our invention and to compare it with multi-tap? What are our testable research questions?

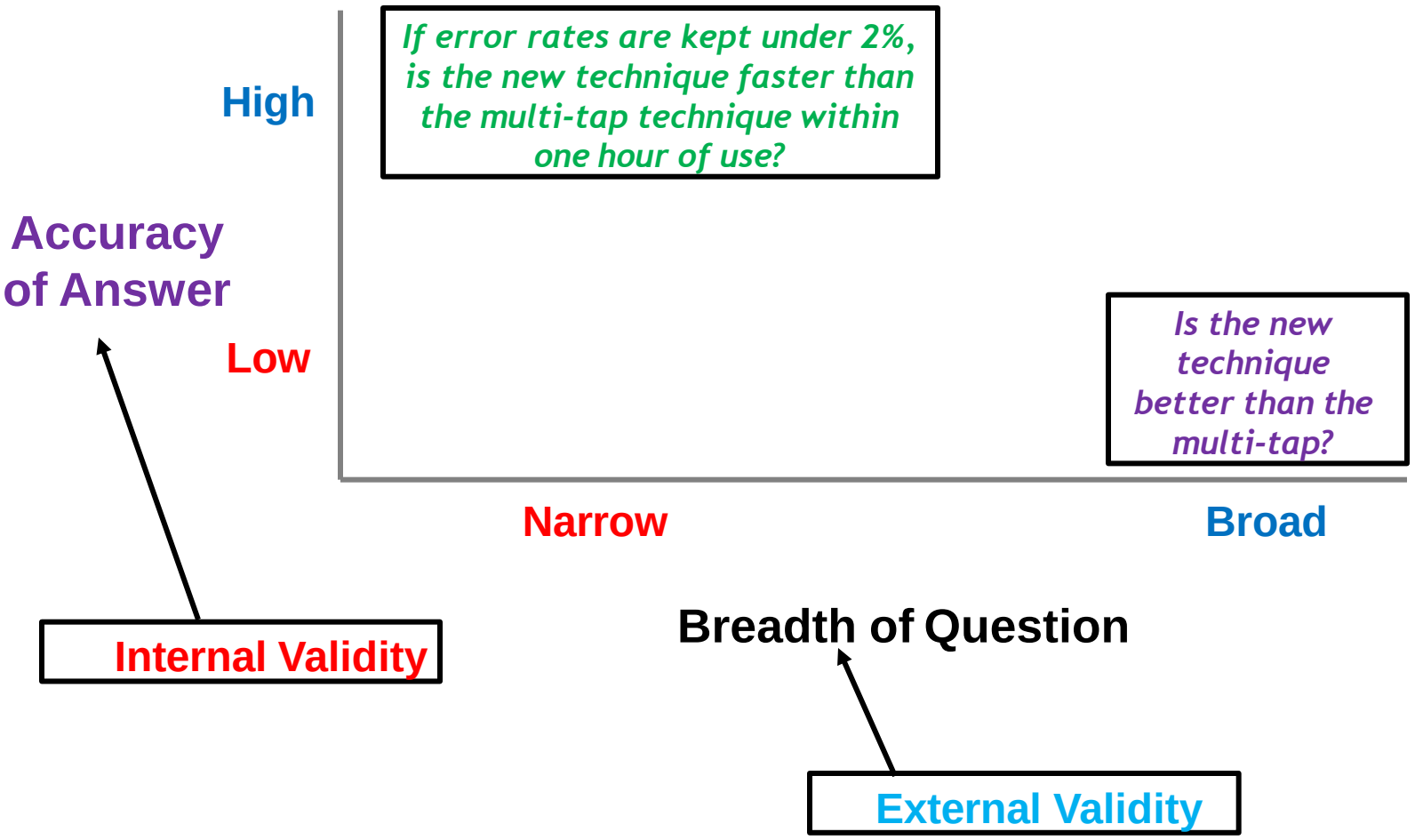
Testable Research Question

- Weak Question
  - Is the new technique better than multi-tap?
- Better
  - Is the new technique faster than multi-tap?
- Better still
  - Is the new technique faster than multi-tap within 1 hour of use?
- Even better
  - If error rates are kept under 2%, is the new technique faster than multi-tap technique within one hour of use?

Testable Research Question

- The questions are testable (we can actually conduct the experiments to test the answer to the testable questions)
- We can ask very specific questions (the last one) or relatively broader questions (the first one)
- For very specific questions, the accuracy of answers is high whereas for the broader questions, the breadth or the generalizability is very high

Testable Research Question



Internal and External Validity

- The extent to which the effects observed are due to the test conditions is called internal validity of the research question
- The extent to which the results are generalizable to other people and other situations is known as the external validity of the research question

More Examples on Validity

- Suppose we wish to compare two input devices for remote pointing (e.g., at a projection screen)
- External validity is improved if the test environment mimics expected usage
  - The test environment should use a projection screen, position participants at a significant distance from screen, have participants stand and include an audience

More Examples on Validity

- Note that creating the test environment mimicking the real usage scenario is not easy
- Instead we can go for controlled experiments where we can ask the user to sit in front of a computer in a laboratory and use the pointing devices to operate an application on the screen
  - The above setting can answer research questions with high internal validity but can not help in determining if the answers are applicable in real world scenario

More Examples on Validity

- Consider the scenario where we wish to compare two text entry techniques for the mobile devices
- To improve external validity, the test procedure should require participants to enter representative samples of text (e.g., phrases containing letters, numbers, punctuation, etc.) and correct mistakes
  - This may require compromising on internal validity

Trade-off

- There is tension between internal and external validities
  - The more the test environment and experimental procedures are "relaxed" (to mimic the real-world situations), the more the experiment is susceptible to uncontrolled sources of variation, such as pondering, distractions, or secondary tasks

Resolving the Trade-off

- Internal and external validities are increased by posing multiple narrow (testable) questions that cover the range of outcomes influencing the broader (un-testable) questions

Ex: A technique that is ***faster***, is ***more accurate***, takes ***fewer steps***, is ***easy to learn***, and is ***easy to remember***, is generally ***better***

Resolving the Trade-off

- The "good news" is that there is usually a positive correlation between the testable and un-testable questions
  - For example, participants generally find a UI ***better*** if it is ***faster***, ***more accurate***, ***takes fewer steps***, etc.
- The "good news", in fact, is not so good after all as it raises more confusions

Implication

- The “good news” actually implies we do not need empirical research!!
- We just do a user study and ask participants which technique they preferred
  - Because of the “positive correlation”, we need not take the pain in collecting and analyzing the data

Implication

- However, this is not true
- If participants are asked which technique they prefer (a broader question), they'll probably give an answer…even if they really have no particular preference!
  - There are many reasons, such as how recently they were tested on a technique, personal interaction with the experimenter, etc.

Implication

- Therefore, such preferences need not be indicative of the system performance
  - We need to scientifically ascertain the validity of the preferences expressed by the participants, which requires formulation of testable research questions
- Also, with broader questions, we may not get an idea about the feasibility or usefulness of the system
  - It is not enough to know if a system is better than another system only but we also need to know "how much better" (for example, it may not be feasible economically to develop a system that is only 5% better than the current system)

Implication

- Seeking the feedback from users on broader questions is not very helpful from another perspective
  - It does not help to identify the strengths, weaknesses, limits, capabilities of the design, thereby making it difficult to identify opportunities for improvements
- Such concerns can be addressed only with the raising of testable research questions
- A point to note is, to test the validity of empirical research questions through observations, we need **measurements**
  - This brings us towards the second theme of empirical research, i.e. to Observe and Measure.

Observe and Measure

- In empirical research, the observation is the most fundamental thing (activity) to do
- Observational (empirical) data can be gathered in two ways
  - Manual: In this case, a human observer manually records all the relevant observational data
  - Automatic: The observation can be recorded automatically, through the use of computers, software, sensor, camera etc.

Observe and Measure

- A measurement is, simply put, a recorded observation
- There are broadly four ***Scales of Measurements*** that are used (Nominal, Ordinal, Interval and Ratio)
- ***Nominal*:** Here, we assign some (arbitrary) codes to the attributes of the observational data (for example, male = 1, female = 2 etc.)

Scales of Measurements

- ***Ordinal*:** In this scale of measurement, the observations are ranked (for e.g., 1st, 2nd, 3rd etc.)
- ***Interval*:** In an interval measurement, we consider equally spaced units but no absolute starting point (for e.g., 20° C, 30° C, 40° C, …)
- ***Ratio*:** This scale of measurement has an absolute starting point (zero) and uses the ratios of two quantities (for e.g., 20 WPM, 30 CPS etc.)

Scales of Measurement

- Nominal
- Ordinal
- Interval
- **Ratio**

Crude

Sophisticated

Ratio measurements, being the most sophisticated scale of measurement, should be used as much as possible

Ratio Measurements

- As mentioned in the previous slide, ratio scales are the most preferred scale of measurement
  - This is because ratio scales make it convenient to compare or summarize observations
- If we are conducting an empirical research, then we should strive to report the “counts” as ratios wherever possible

Ratio Measurements

- For e.g., let us assume that we observed that " a 10-word phrase was entered by a participant in an empirical study in 30 seconds". What should we measure?
  - If we measure the "time to enter text" (e.g., t = 30 seconds) as an indicator of system performance, it is a bad measurement.
  - However, if we go for a ratio measurement (Entry Rate = 10/0.5 i.e. Entry Rate = 20 wpm), that is much better and gives a general indication of the performance.

Ratio Measurements

- Let us consider another example. Suppose in an empirical study, we observed that a participant has committed two errors while entering a 50 character phrase.
  - If we measure the "number of errors committed" (i.e., n = 2) as an indicator of system performance, it is a bad measurement.
  - However, if we go for a ratio measurement (Error Rate = 2/50, i.e. Error Rate = 0.04 = 4%), that is much better and is a more general performance indicator.

Summary

- We have discussed two of the three themes of empirical research, namely: (1) Answer and Raise (Ask) Questions about a new or existing UI Design or Interaction Method, (2) Observe and Measure
- We shall continue our discussion w.r.t. the third theme of empirical research (i.e. User Studies) in the next lecture

HCI: Empirical Research Methods

Learning Objective

- In previous lectures, we discussed the conceptual framework of empirical research (which is observation-based investigation)
  - In the context of HCI, it refers to an investigation about the usability of a system with the help of end users in a real-life scenario
- The investigation is based on testable research questions
  - We saw how to formulate the appropriate/relevant questions (the validity of questions, the trade-off and how to resolve the trade-off)
- We also discussed the process of observation and measurements
- In this lecture, we shall discuss other aspects of empirical research; In particular, we shall discuss the design of experiments.

Themes of Empirical Research

- To recap, the three themes of empirical research are
  - Answer and Raise Questions (Testable Research Questions)
  - Observe and Measure (Ratio Scale of Measurement is Preferable)
  - User Studies (Design of Experiments)

- In the previous lecture, we already discussed the first two HCI themes, namely: (1) Answer and Raise Questions, (2) Observe and Measure.

- In this lecture, let's discuss the third theme (User Studies) which is also equally important since we need to perform the observations in such user studies (design of Experiments).

User Study

- A user study, in the context of HCI, is a scientific way of collecting and analyzing observational data from the end users on an interactive system
- Collection of data involve the real users for conducting experiments and design of experiments

Experiment Design

- Experiment design is a general term refers to the organization of variables, procedures, etc., in an experiment
- The process of designing an experiment is the process of deciding on which variables to use, what procedure to use, how many participants (users) to use, how to solicit them etc.

Terminology

- Terms to know
  - Participant (User)
  - Independent Variable (Test Conditions)
  - Dependent Variable
  - Control Variable
  - Random Variable
  - Confounding Variable
  - Within Subjects vs. Between Subjects
  - Counterbalancing and Latin Square

Participant (User)

- The people participating in an experiment are referred to as participants (users)
  - When referring specifically to the experiment, use the term participants (e.g., “all participants exhibited a high error rate…”)
  - General comments on the problem or conclusions drawn from the results may use other terms (e.g., “these results suggest that users are less likely to…”

Independent Variable

- An independent variable is a variable that is selected or controlled through the design of the experiment
  - Examples including the device, feedback mode, the button layout, visual layout, gender, age, expertise, etc.
- The terms: independent variable and the factor are synonymous

Test Conditions

- The levels, values, or settings for an independent variable are the test conditions
- These conditions provide names for both an independent variable (factor) and the test conditions (levels) for the controlled variable (ex: table details)

Factor	Levels (Test Conditions)
Device	Mouse, Trackball, Joystick
Feedback Mode	Audio, Tactile, Visual
Task	Pointing, Dragging
Visualization	2D, 3D, Animated

Dependent Variable

- A variable representing the measurements or observations on a independent variable
- It is required to provide a name for both the dependent variable and its unit
  - Examples: Task completion time (ms), speed (WPM (word per minute)), selections per minute, etc.), error rate (%), throughput (bits/s (bps))

Control Variable

- Circumstances or factors that might influence a dependent variable, but are not under investigation need to be accommodated in some manner
- One way is to control them or to treat them as the control variables (e.g., room lighting, background noise, temperature)

Random Variable

- Instead of controlling all circumstances or factors, some might be allowed to vary randomly
- Such circumstances are random variables

Confounding Variable

- Any variable that varies systematically with an independent variable is a confounding variable
  - For example, if three devices are always administered in the same order, participant performance might improve due to practice; i.e., from the 1st to the 2nd to the 3rd condition; thus "practice" is a confounding variable (because it varies systematically with "device")

Within Subjects, Between Subjects

- The administering of levels of a factor is either within subjects or between subjects
  - If each participant is tested on each level, then the factor is said to be within subjects
  - If each participant is tested on only one level, then the factor is said to be between subjects. In this case a separate group of participants is used for each level
- The terms *repeated measures* and within subjects are synonymous
- A relevant question is, which of the two approaches (within subject and between subject) should be chosen in designing an experiment

Within Subjects, Between Subjects

- Answer: It depends!
  - Sometimes a factor must be between subjects (e.g., gender, age)
  - Sometimes a factor must be within subjects (e.g., session, block)
  - Sometimes there is a choice. In this case there is a trade-off
- The advantage of within subject design is, the variance due to participants' pre-dispositions should be the same across test conditions
- Between subjects design, on the other hand, has the advantage of avoiding interference effects (e.g., the practice effect while typing on two different layouts of keyboards)

Counterbalancing

- For repeated measures designs, participants' performance may tend to improve with practice as they progress from one level to the next level
  - Thus, users may perform better on the second level simply because they benefited from the practice on the first (this is undesirable)
- To compensate this, the order of presenting conditions must be counterbalanced

Latin Square

- Participants are divided into groups, and a different order of administration is used for each group
- The order is best governed by a Latin Square (the defining characteristic of a Latin Square is that each condition occurs only once in each row and column)

Latin Square

- Example: Suppose that we want to administer 4 levels (denoted by A, B, C and D) of a factor to the 4 participants (represented by P1, P2, P3 and P4)
  - We can construct a 4×4 Latin square arrangement to depict the order of administering the levels to each participant

P1	A	B	C	D
P2	B	C	D	A
P3	C	D	A	B
P4	D	A	B	C

Latin Square

- In a *balanced* Latin Square, each condition both precedes and follows each other condition an equal number of times
  - We can construct a balanced 4×4 Latin square arrangement for the previous example

P1	A	B	C	D
P2	B	D	A	C
P3	D	C	B	A
P4	C	A	D	B

Expressing Experiment Design

- Consider the statement “3 x 2 repeated-measures design”
  - It refers to an experiment with two factors, having three levels on the first, and two levels on the second. There are six test conditions in total. Both factors are repeated measures, meaning all participants were tested on all test conditions
- Any type of experiment is expressed similarly

Summary

- In this and the previous lectures, we discussed the fundamental ideas associated with empirical research (namely: question formulation, observation and measurement and experiment design)
- In the next lecture, we shall see another important aspect of empirical research, namely: the case for statistical analysis of empirical data

HCI: Analysis of Empirical Data

Learning Objective

- In the previous lectures, we have already discussed the basics of HCI - Empirical Research Methods
  - We discussed the three important themes, namely: Research Question Formulation, Observation and Measurement, and User Study (Experiment Design)
- In the present lecture, we shall mainly focus on the Analysis of Empirical Data

Learning Objective

- In particular, we shall learn the following:
  - The case for Statistical Analysis of Observed Data
  - Discussion on one of the commonly and widely used Statistical-based Empirical Analysis Techniques, namely: One-way ANalysis Of VAriance (ANOVA) Test

Answering Empirical Questions

- Suppose, we want to determine if the text entry speed of a proposed text input system is more than an existing system
- We know how to design an experiment and further we know how to observe and measure
- So, let us do the following:

  - We conduct a User Study and measure the performance on each of the given test conditions (our proposed system and the existing system) over a group of participants

  - For each test condition, let us compute the mean score (text entry speed) over the group of participants (users)

Answering Empirical Questions

- We now have the observed (empirical) data. What next?
- Now, we are faced with the following three questions:
  - Is there any difference?

    This is true as we are most likely to see some differences. However, can we conclude anything from this difference? This brings us to the second question.

  - Is the difference too large or too small?

    This is more difficult to answer. If we observe a difference of, say, 30%, we can definitely say this difference is too large. But, we can't say anything definite about, i.e., 5% difference. Clearly, the difference itself can't help us to draw any definite conclusion. This brings us to the third question.

Answering Empirical Questions

- We are faced with the following three questions (contd...):
  - Is the difference significant or is it due to chance?

    Even if the observed difference is "small", it can still lead us to conclude about our design if we can find the nature of difference. If the difference is found to be "significant" (not occurred due to any chance), then we can say something about our design.
- Pl. note that the term "significance" is a statistical term
- The test of (statistical) significance is an important aspect of empirical data analysis
- We can use statistical techniques for this purpose
  - The basic technique is ANOVA or ANalysis Of VAriance

Statistical Hypothesis Testing

In statistics, the Hypothesis is a claim or the statement about a property of a population.

Hypothesis Test (or Test of Significance) is considered as a standard procedure for testing a claim about a property of a population.

Basics of Hypothesis Testing

The following components of a statistical hypothesis test are widely used for carrying out the comprehensive procedures.

- **Null and Alternative Hypotheses**
- **Test Statistics**
- **Critical Region and Critical Values**
- **Significance Levels**
- ***P*-values**
- **Decision Criteria**
- **Type I and II Errors**
- **Power of a Hypothesis Test**

Learning Objectives

- Given a claim, How to identify both the null and the alternative hypotheses, and then express them in symbolic form.
- Given a claim and sample data, how to calculate the value of the test statistic.
- Given a significance level, How to identify the critical value(s).
- Given a value of the test statistic, How to identify the *P*-value.
- State the conclusion of a hypothesis test in simple, non-technical terms.

Example: Let's refer to the Gender Choice product that was once distributed by the ProCare Industries. ProCare claimed that couples using pink packages of Gender Choice would have girls at a rate greater than 50% or 0.5. Let's again consider an experiment whereby 100 couples use the Gender Choice in an attempt to have a baby girl; Let's assume that the 100 babies include exactly 52 girls, and let's formalize some of the analysis. Under normal circumstances the proportion of girls is 0.5, so a claim that Gender Choice is effective, can be expressed as $p > 0.5$. Using a normal distribution as an approximation to the binomial distribution, we find that P(52 girls or more in 100 births) = 0.3821. Figure 1 shows that with a probability of 0.5, the outcome of 52 girls in 100 births is not unusual.

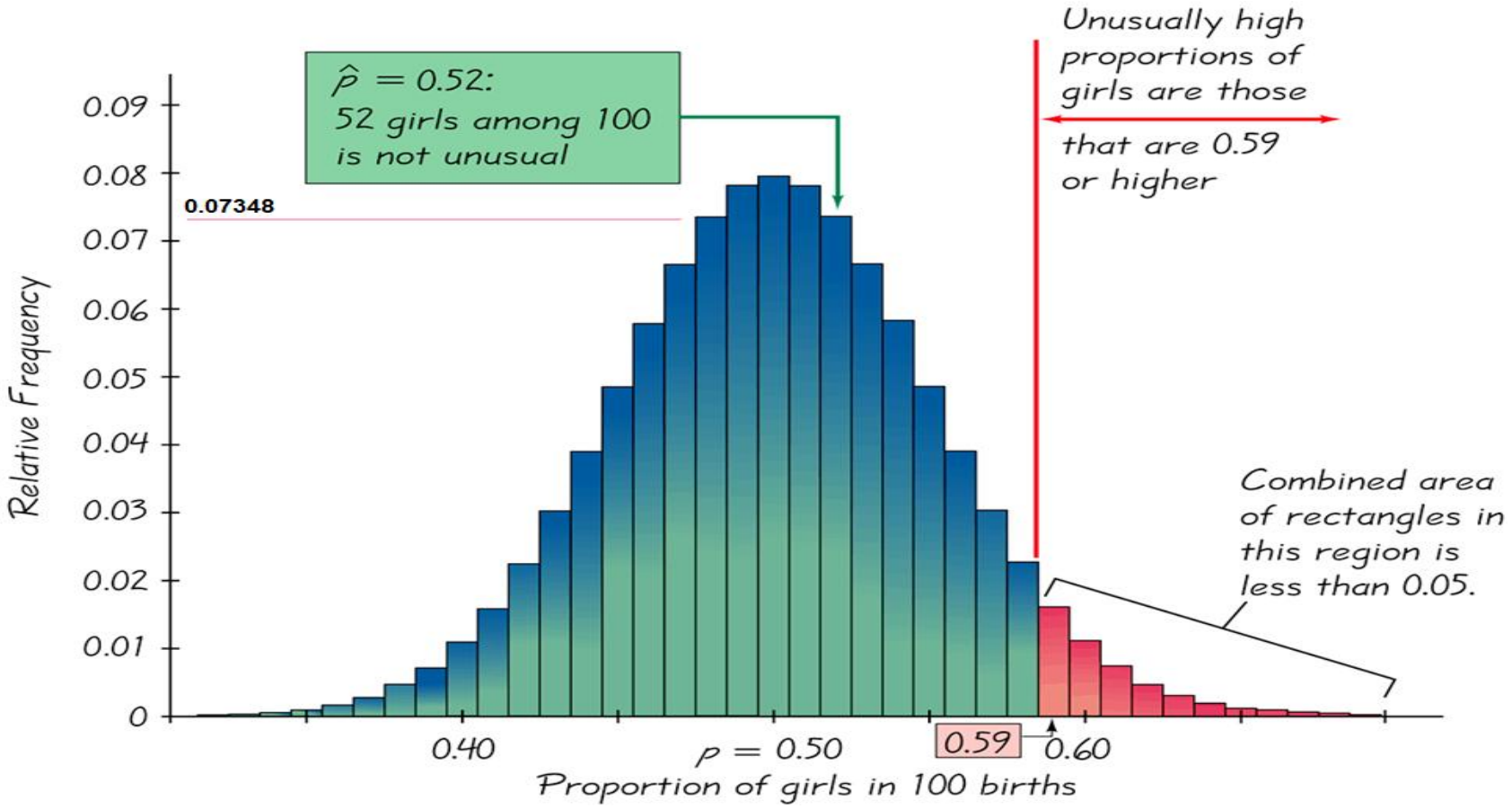


Figure 1

We do not reject random chance as a reasonable explanation. We conclude that the proportion of girls born to couples using Gender Choice is not significantly greater than the number that we would expect by random chance.

Observations

- **Claim: For couples using the Gender Choice product, the proportion of girls is $p > 0.5$.**
- **Working assumption: The proportion of girls is $p = 0.5$ (with no effect from the Gender Choice).**
- The sample resulted in 52 girls among 100 births, so the sample proportion is $\hat{p} = 52/100 = 0.52$.
- **Assuming that $p = 0.5$, we use a normal distribution as an approximation to the binomial distribution to find that P (at least 52 girls in 100 births) = 0.3821.**
- There are two possible explanations for the result of 52 girls out of 100 children births: Either a random chance event (with probability 0.3821) has occurred, or the proportion of girls born to couples using Gender Choice is greater than 0.5.
- There isn't sufficient evidence to support Gender Choice's claim.

Components of a Formal Statistical Hypothesis Test

Null Hypothesis: H_0

- **The Null Hypothesis (H_0) is a statement that the value of a population parameter (such as proportion, mean, or the standard deviation) is equal to some claimed value.**
- **We test the Null Hypothesis directly.**
- **Either reject H_0 or fail to reject (accept) H_0.**

Alternative Hypothesis: (H_1 or H_a or H_A)

- **The Alternative Hypothesis (H_1 or H_a or H_A) is the statement that the parameter has a value that somehow differs from The Null Hypothesis.**
- **The symbolic form of the alternative hypothesis must use one of these symbols: ≠, <, >.**

Note about Forming Our Own Claims (Hypotheses)

If we are conducting a study and want to use a hypothesis test to support our claim, then the claim must be worded so that it becomes the alternative hypothesis.

Note about Identifying H_0 and H_1

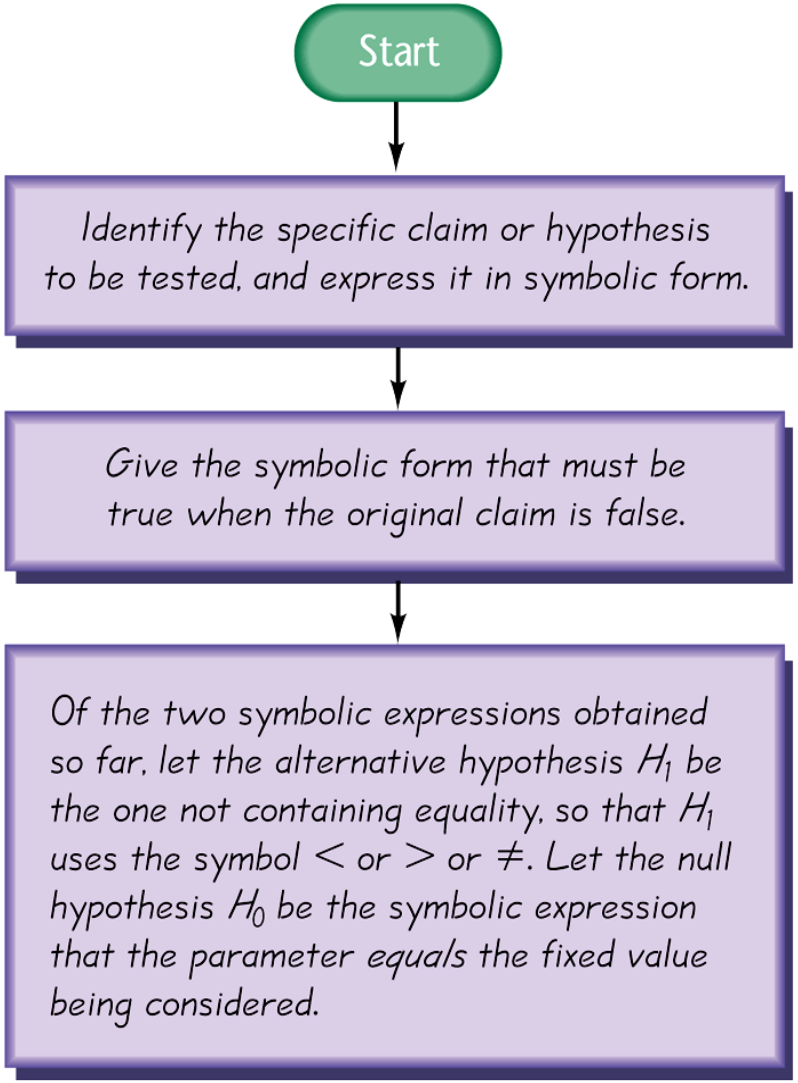


Figure 2

Example: Identify the Null and Alternative Hypotheses. Refer to Fig. 2 and use given claims to express corresponding null and alternative hypotheses in symbolic form.

a) The proportion of drivers who admit to running red lights is > 0.5.

b) The mean height of professional basketball players is at most 7 ft.

c) The standard deviation of IQ scores of actors is equal to 15.

Example: Identify the Null and Alternative Hypotheses. Refer to the Figure 2 and use the given claims to express the corresponding null and alternative hypotheses in symbolic form.

a) The proportion of drivers who admit to running red lights is greater than 0.5. In Step 1 of Figure 2, we express the given claim as $p > 0.5$. In Step 2, we see that if $p > 0.5$ is false, then $p \leq 0.5$ must be true. In Step 3, we see that the expression $p > 0.5$ does not contain equality, so we let the alternative hypothesis H_1 be $p > 0.5$, and we let H_0 be $p = 0.5$.

Example: Identify the Null and Alternative Hypotheses. Refer to the Figure 2 and use the given claims to express the corresponding null and alternative hypotheses in symbolic form.

b) The mean height of professional basketball players is at most 7 ft. In Step 1 of Figure 2, we can express that "The mean of at most 7 ft" it symbols as $\mu \leq 7$. In Step 2, we see that if $\mu \leq 7$ is false, then $\mu > 7$ must be true. In Step 3, we see that the expression $\mu > 7$ does not contain equality, so we let the alternative hypothesis H_1 be $\mu > 7$, and we let H_0 be $\mu = 7$.

Example: Identify the Null and Alternative Hypotheses. Refer to Figure 2 and use the given claims to express the corresponding null and alternative hypotheses in symbolic form.

c) The standard deviation of IQ scores of actors is equal to 15. In Step 1 of Figure 2, we express the given claim as $\sigma = 15$. In Step 2, we see that if $\sigma = 15$ is false, then $\sigma \neq 15$ must be true. In Step 3, we let the alternative hypothesis H_1 be $\sigma \neq 15$, and we let H_0 be $\sigma = 15$.

Test Statistic

The Test Statistic is a value used in making a decision about the null hypothesis, and is found by converting the sample statistic to a score with the assumption that the null hypothesis is true.

Test Statistic - Formulas

$$z = \frac{\hat{p} - p}{\sqrt{\frac{pq}{n}}}$$

Test Statistic for Proportions

$$z = \frac{\bar{x} - \mu_{\bar{x}}}{\frac{\sigma}{\sqrt{n}}}$$

Test Statistic for Mean

$$\chi^2 = \frac{(n-1)s^2}{\sigma^2}$$

Test Statistic for Standard Deviation

Example: A survey of $n = 880$ randomly selected adult drivers showed that 56% (or $p = 0.56$) of those respondents admitted to running the red lights. Find the value of test statistic for the claim that the majority of adult drivers admit to running red lights. (Assume that the required assumptions are satisfied and the focus must be on finding the indicated test statistic).

Solution: The example of the previous slide showed that the given claim results in the following null and alternative hypotheses: H_0: $p = 0.5$ and H_1: $p > 0.5$. Because we work under the assumption that the null hypothesis is true for a value of $p = 0.5$, we get the following test statistic:

$$z = \frac{\hat{p} - p}{\sqrt{\frac{pq}{n}}} = \frac{0.56 - 0.5}{\sqrt{\frac{(0.5)(0.5)}{880}}} = 3.56$$

Critical Region

The Critical Region (or Rejection Region) is the set of all values of the test statistic that cause us to reject the null hypothesis. For example, see the red-shaded region as shown in Figure 3.

Significance Level

The Significance Level (α) is the probability that the test statistic will fall in the critical region when the null hypothesis is actually true. Common choices for α are 0.05, 0.01, and 0.10.

Critical Value

A Critical Value is any value that separates the critical region (where we reject null hypothesis) from the values of the test statistic that do not lead to rejection of the null hypothesis. The critical values depend on the nature of the null hypothesis, sampling distribution that applies, and the significance level (α). Pl. See Figure 3 where a critical value of $z = 1.645$ corresponds to a significance level of $\alpha = 0.05$.

Test Statistic

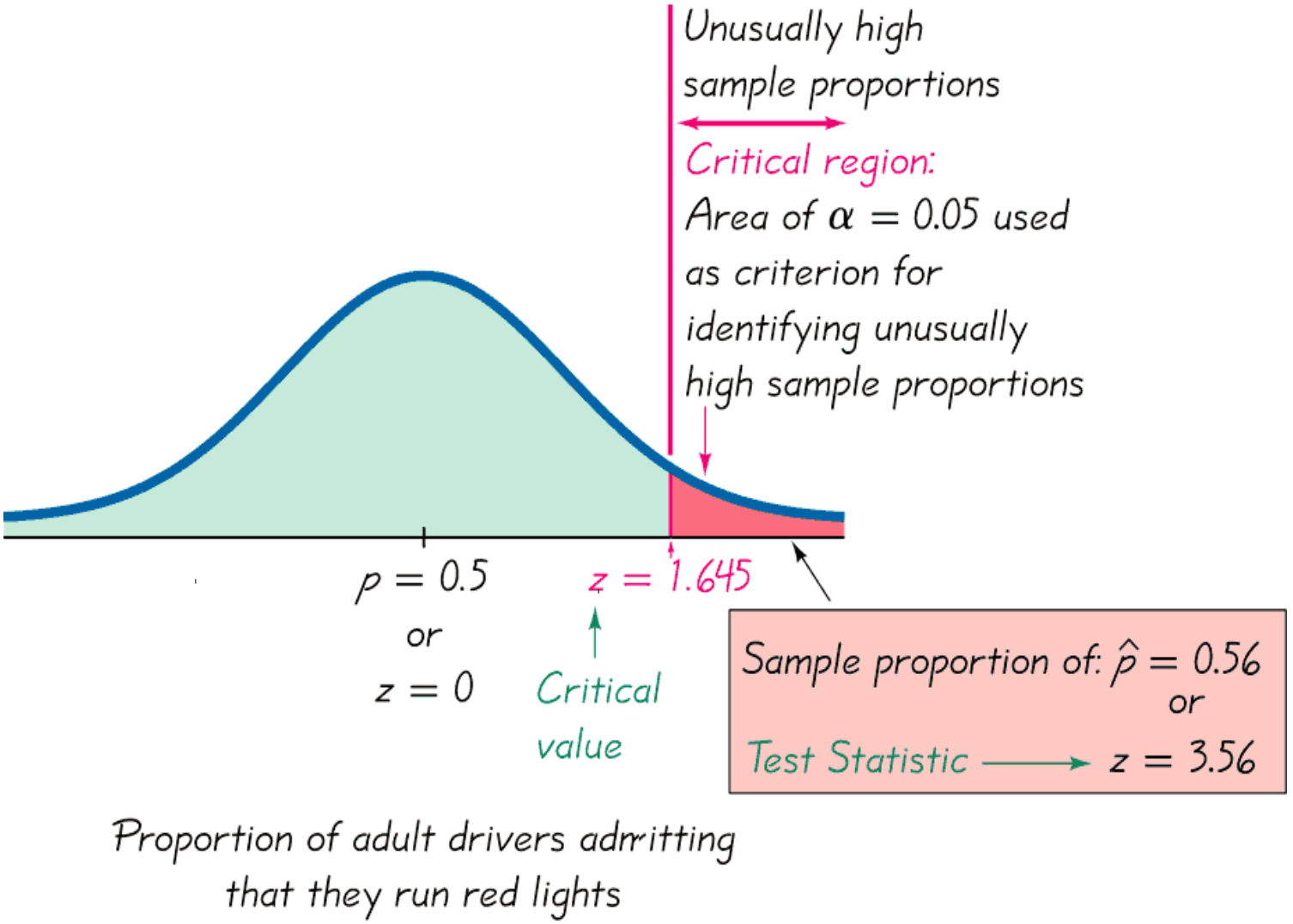


Figure 3

Two-tailed Test

H_0: =

H_1: ≠

The Tails in a distribution are the extreme regions bounded by critical values.

Significance level (α) is divided equally between the two tails of the critical region

Here ≠ means Less than or Greater than

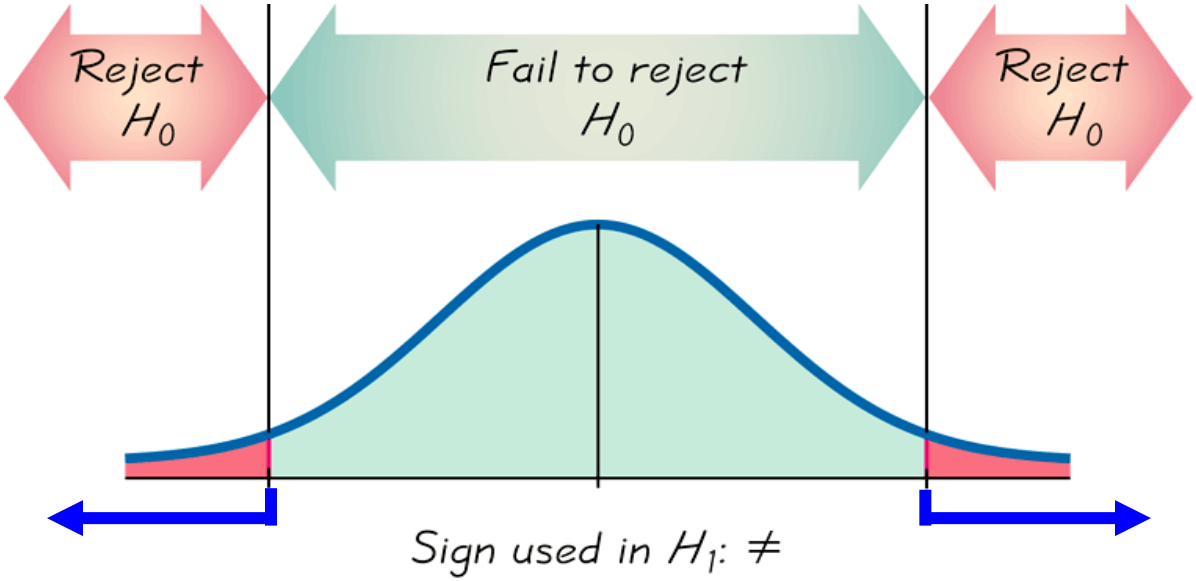


Figure 4

Right-tailed Test

H_0: =

H_1: >

Points Right

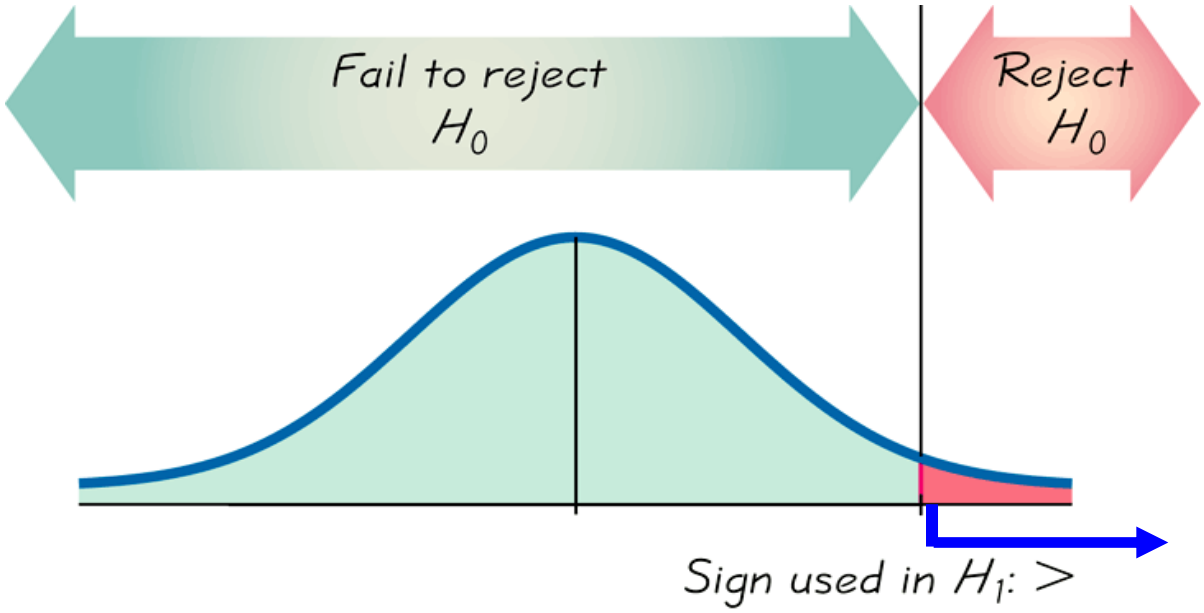


Figure 5

Left-tailed Test

H_0: =

H_1: <

Points Left

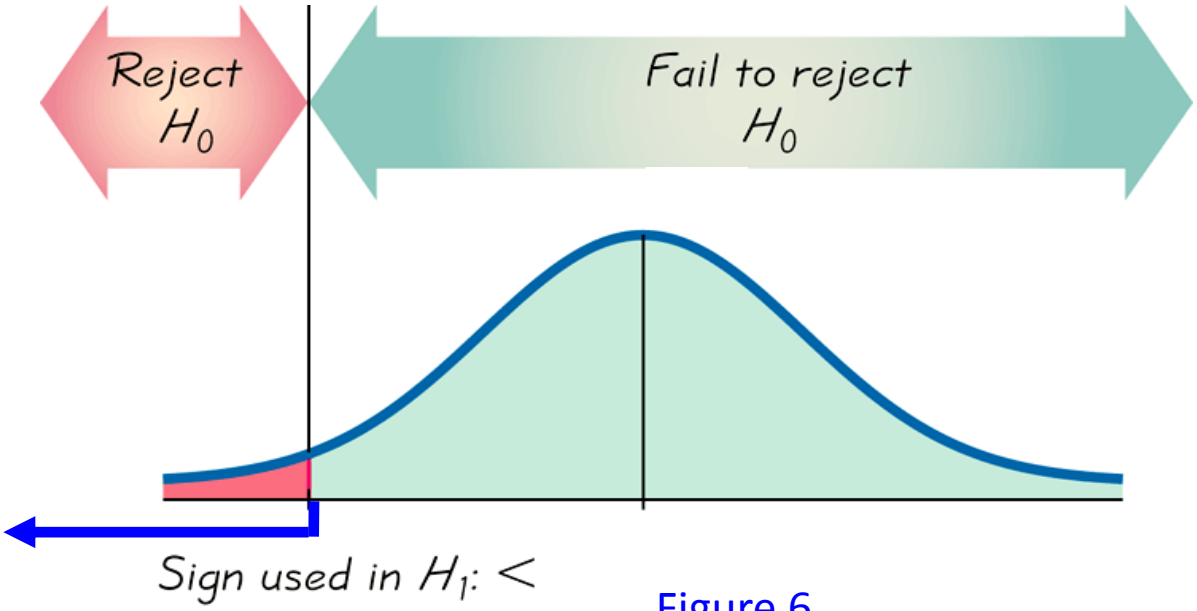


Figure 6

P-Value

The *P*-value (or) *p*-value (or) probability value is the probability of getting a value of the test statistic that is at least as extreme as the one representing the sample data, assuming that null hypothesis is true, and it is rejected if the *P*-value is very small, such as 0.05 or less.

Initial Conclusions in Hypothesis Testing

We always test the null hypothesis. The inference will be one of the following:

1. Reject the Null Hypothesis.

2. Fail to Reject the Null Hypothesis.

Decision Criterion

Traditional Method:

Reject the *Null Hypothesis* (H_0) if the test statistic falls within the critical region.

Fail to Reject the H_0 if the test statistic does not fall within the critical region.

Decision Criterion (Contd…)

P-value Method:

Reject H_0 **if the *P*-value** $\leq \alpha$ (where α is the significance level, such as 0.05).

Fail to Reject H_0 if the *P*-value > α.

Decision Criterion (Contd…)

Another Option:

Instead of using a significance level (α) such as 0.05, simply identify the *P*-value and leave the decision to the reader.

Decision Criterion (Contd…)

Confidence Intervals:

Because a confidence interval estimate of a population parameter contains likely values of that parameter, reject a claim that the population parameter has a value that is not included in the confidence interval.

Procedure for Finding *P*-Values

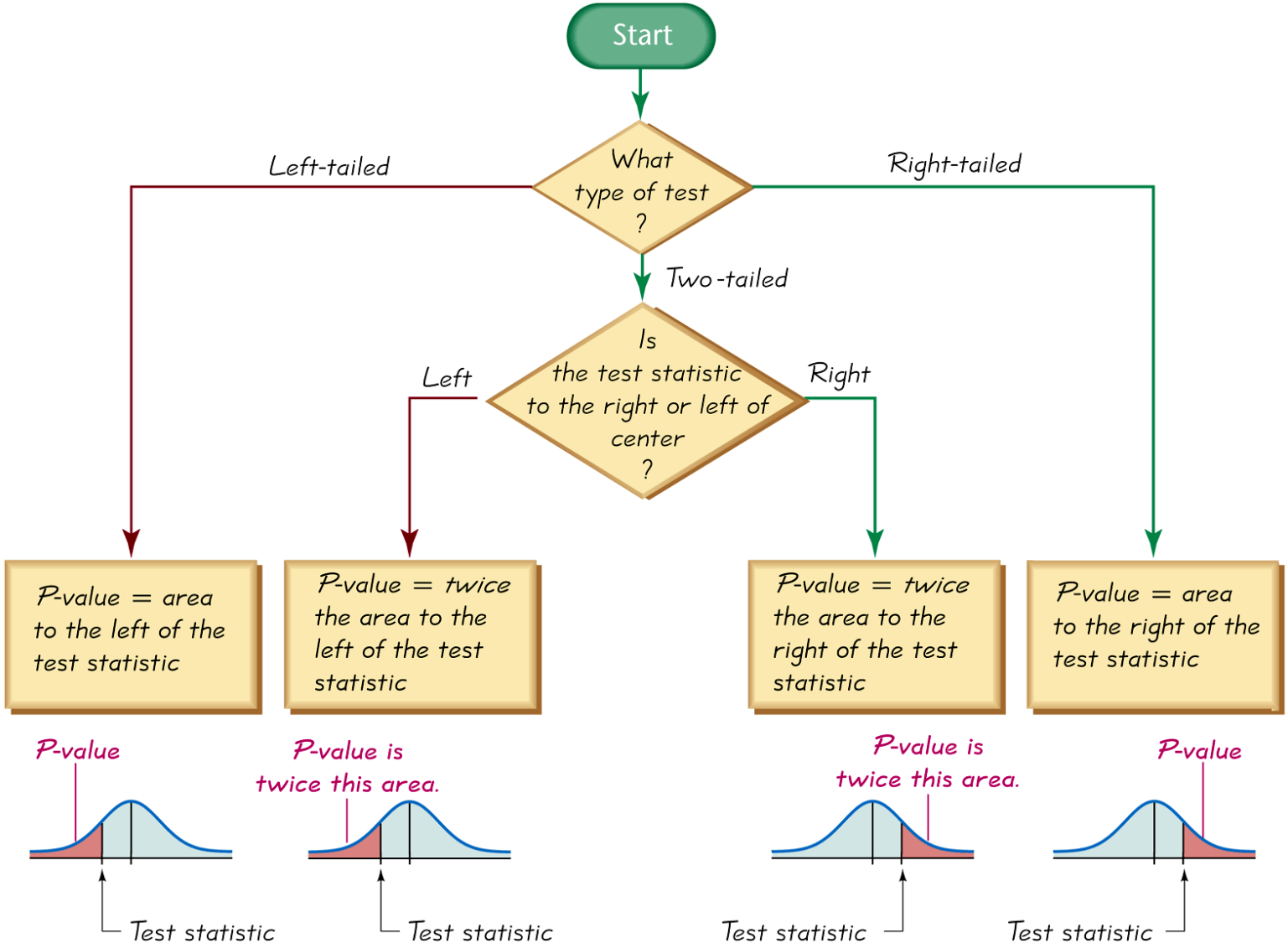


Figure 7

Wording of Final Conclusion

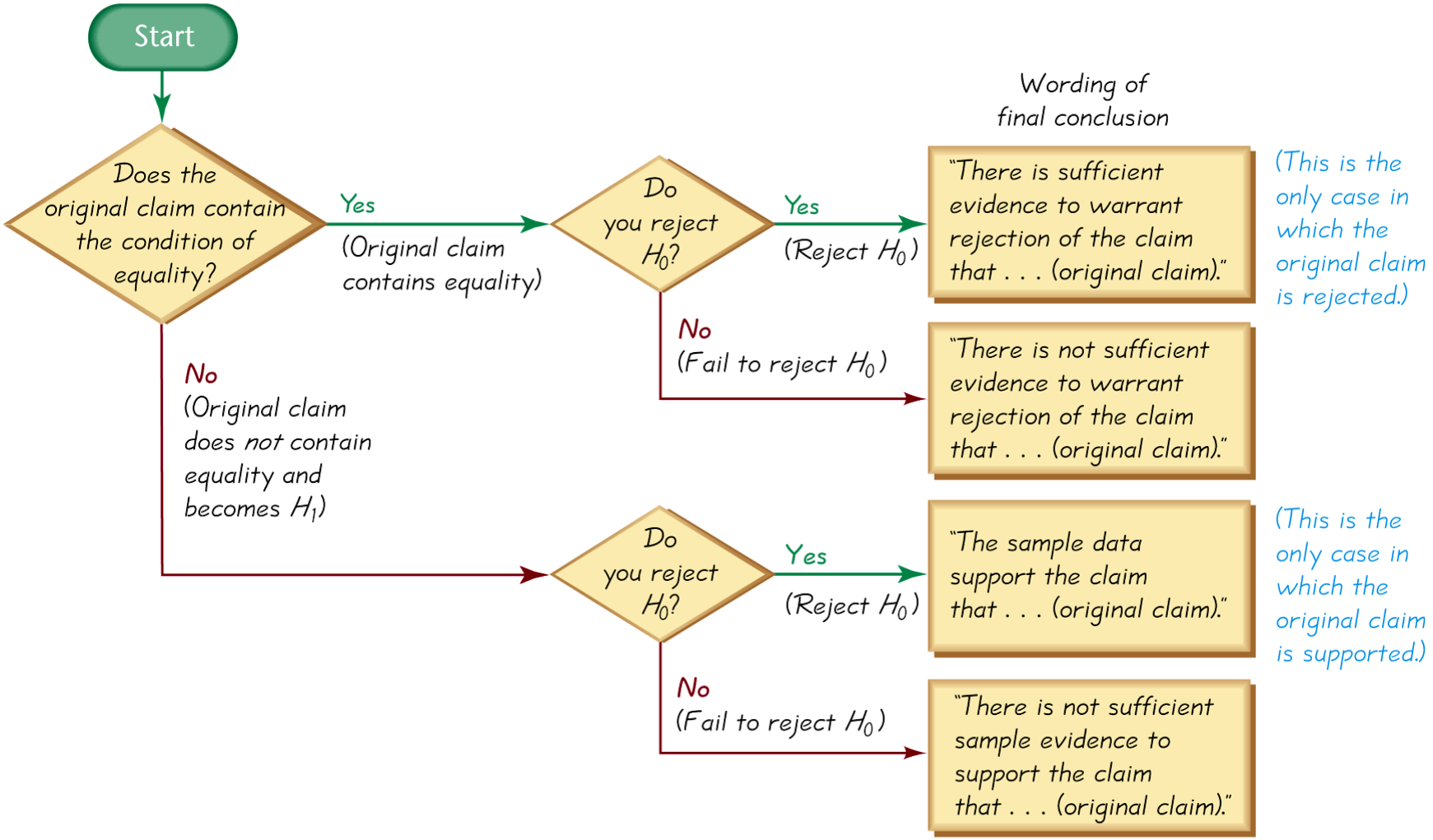


Figure 8

Accept Versus Fail to Reject

- **Some texts use "accept the null hypothesis".**
- **We are not proving the null hypothesis.**
- **The sample evidence is not strong enough to warrant the rejection (such that there is not enough evidence to convict a suspect).**

Type I Error

- **A Type I Error is the mistake of rejecting the null hypothesis when it is true.**

- **The symbol α (alpha) is used to represent the probability of a Type I error.**

Type II Error

- **A Type II Error is the mistake of failing to reject the null hypothesis when it is false.**

- **The symbol β (beta) is used to represent the probability of a Type II error.**

Type I and Type II Errors

Table: Type I and Type II Errors

		True State of Nature	
		The null hypothesis is true	The null hypothesis is false
Decision	We decide to reject the null hypothesis	**Type I error** (rejecting a true null hypothesis) α	Correct decision
	We fall to reject the null hypothesis	Correct decision	**Type II error** (failing to reject a false null hypothesis) β

Controlling Type I and Type II Errors

- **For any fixed value of α, an increase in the sample size *n* will cause a decrease in β.**

- **For any fixed sample size *n*, a decrease in α will cause an increase in β. Conversely, an increase in α will cause a decrease in β.**

- **To decrease both α and β, we need to increase the sample size.**

Power of a Hypothesis Test

The Power of a Hypothesis Test is the probability (1 - β) of rejecting a false null hypothesis, which is computed by using a particular significance level α and a particular value of population parameter that is an alternative to the value which is assumed to be true in the null hypothesis. That is, the power of the hypothesis test is known as the probability of supporting an alternative hypothesis that is true.

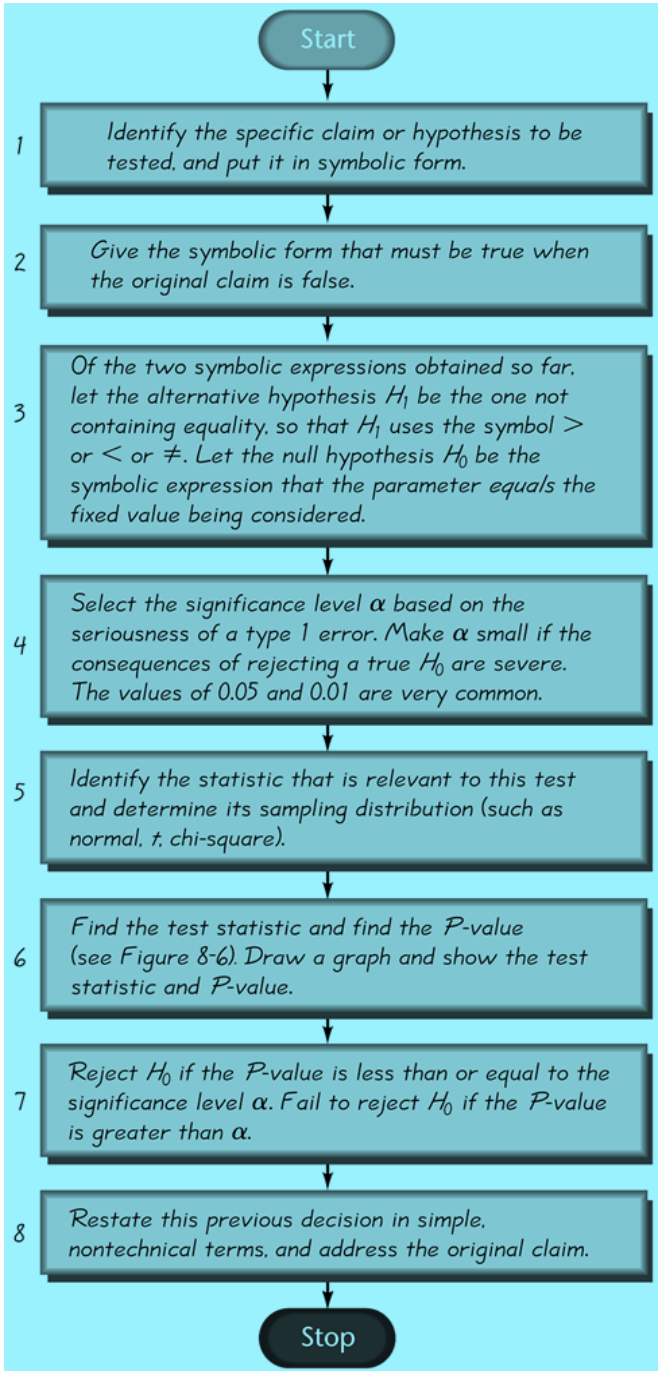


Figure 9

Comprehensive Hypothesis Test: *P*-Value Method

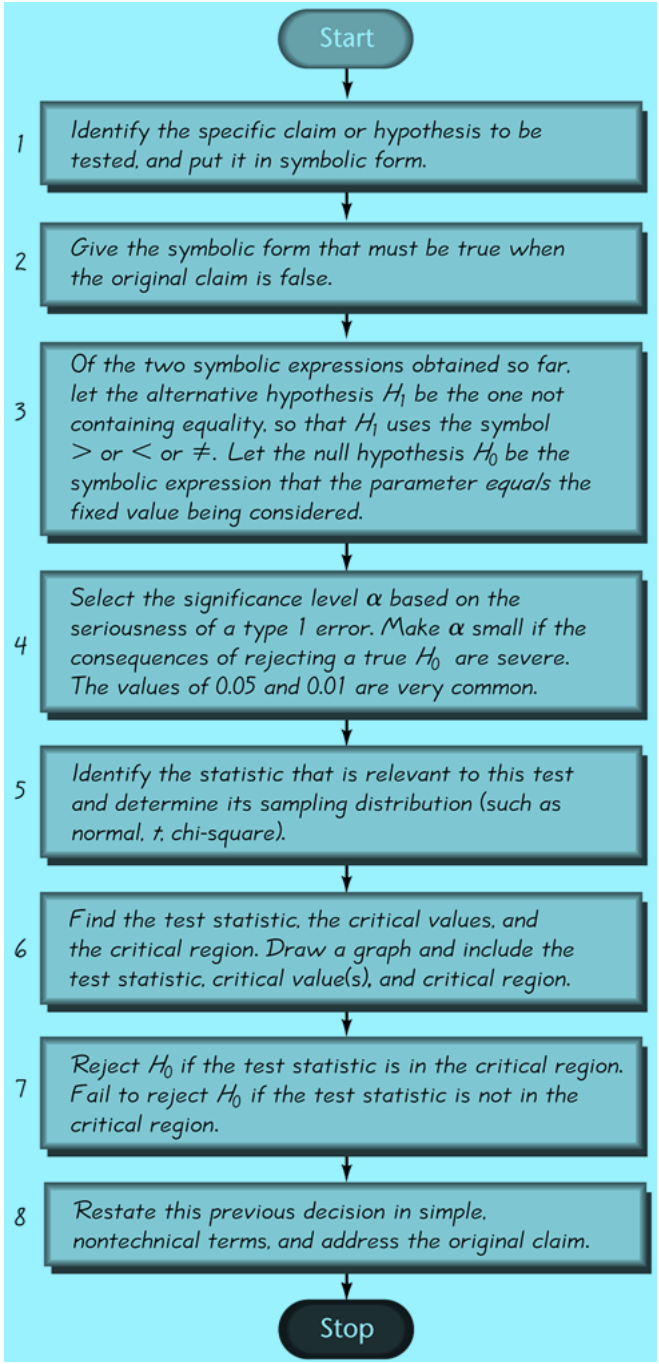


Figure 10

Comprehensive Hypothesis Test: Traditional Method

Comprehensive Hypothesis Test

A confidence interval estimate of a population parameter contains the values of that parameter. We should therefore reject a claim that the population parameter has a value that is not included in the confidence interval.

Table Confidence Level for Confidence Interval

		Two-Tailed Test	One-Tailed Test
Significance Level for Hypothesis Test	0.01	99%	98%
	0.05	95%	90%
	0.10	90%	80%

Summary of Hypothesis Testing

An **objective** method of making decisions or **inferences** from sample data (evidence)

Sample data used to choose between two choices i.e. **Hypotheses** or Statements about a population

We do this by comparing what we have observed to what we expected if one of the statements (**Null Hypothesis**) was true

The cat is faced with a decision to go out into the snow or not. It has two hypotheses; Null Hypothesis: Cat won't get its paws wet and be cold, Alternative Hypothesis: It will get its paws wet and will be cold. Cat could base its decision on data that is collected in its mind in the past; when it went out in the snow in the past its paws did get cold and wet and so the cat may make the decision not to go out based on that evidence.

Summary of Hypothesis Testing

Always we have Two Hypotheses:

H_1: Alternative (Research) Hypothesis

What we aim to gather evidence of.

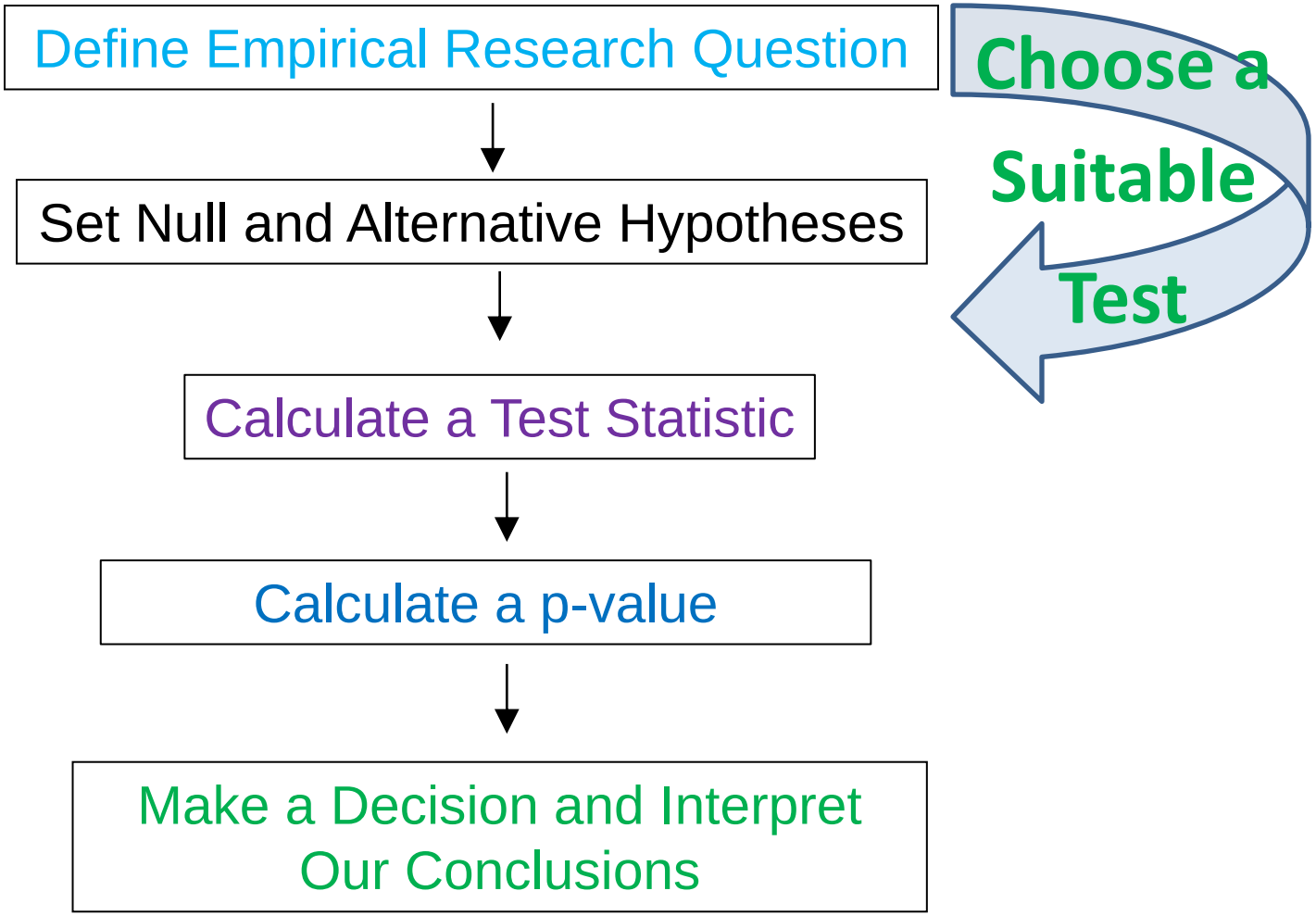
Typically that there **is** a difference/effect/relationship etc.

H_0: Null Hypothesis

What we assume is true to begin with.

Typically that there is **no** difference/effect/relationship etc.

Summary of Hypothesis Testing



Summary of Hypothesis Testing

We can use statistical software to undertake a hypothesis test
e.g. SPSS (Statistical Package for Social Sciences)
One part of the output is the p-value (P)

If P < **0.05 reject** H_0 (Reject Null Hypothesis) => **Evidence** of H_1 being true (i.e. **IS** association (Alternative Hypothesis))

If P > **0.05 do not** reject H_0 (Accept the Null Hypothesis)
(i.e. **NO** association (Null Hypothesis))

What if P = 0.049 or 0.051? Discuss the fact that hypothesis testing involves the weight of evidence and "shades of grey" rather than being a clear cut decision making process.

Summary of Hypothesis Testing (Choosing the Right Test)

1) A clearly defined research question
2) What is the dependent variable and what type of variable is it?
3) How many independent variables are there and what data types are they?
4) Are you interested in comparing means or investigating relationships?
5) Do you have repeated measurements of the same variable for each subject?

Summary of Hypothesis Testing (Choosing the Right Test)

- Clarity of empirical research questions with measurable quantities
- Which variables will help answer these empirical research questions
- Think about what test is needed before carrying out a study so that the right type of variables are collected

Summary of Hypothesis Testing (Choosing the Right Test)

- How many variable are involved?
- Two – interested in the relationship
- One dependent and one independent
- One dependent and several independent variables: some may be controls
- Relationships between more than two: multivariate techniques (not covered here)

Summary of Hypothesis Testing (Choosing the Right Test)

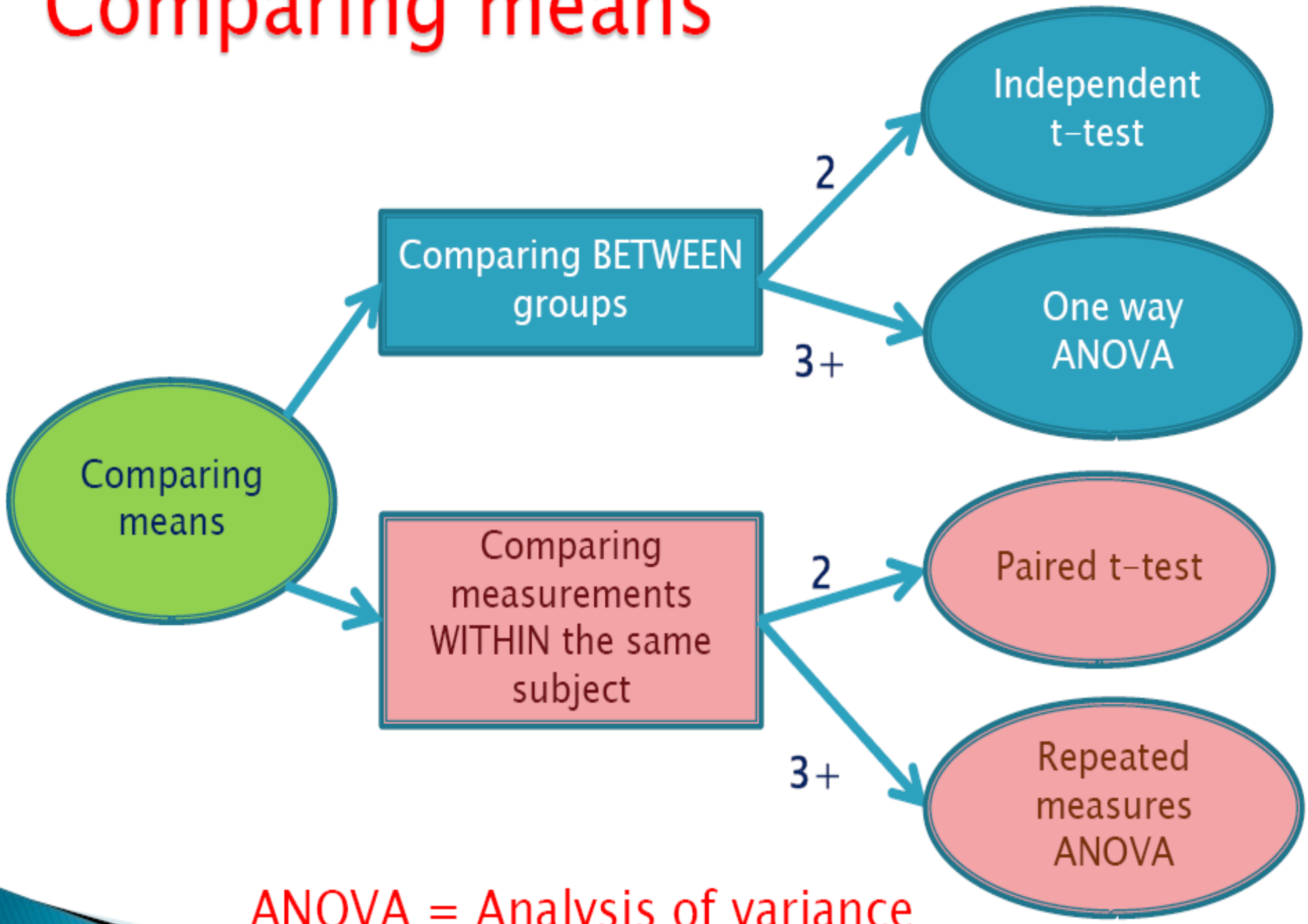
Comparing the Means

- Dependent = Scale
- Independent = Categorical

- How many means are we comparing?

- Do we have independent groups or repeated measurements on each person?

Comparing means



ANOVA = Analysis of variance

Exercise – Comparing the Means

Research Question	Dependent variable	Independent variable	Test
Do women do more housework than men?	Housework (hrs per week) (Scale)	Gender (Nominal)	Independent t-test
Does Margarine X reduce cholesterol? Everyone has cholesterol measured on 3 occasions	Cholesterol (Scale)	Occasion (Nominal)	Repeated measures ANOVA
Which of 3 diets is best for losing weight?	Weight lost on diet (Scale)	Diet (Nominal)	One-way ANOVA

Parametric or non-parametric?

Statistical tests fall into two types:

Parametric tests

Assume data follows a particular distribution e.g. normal

Non-parametric

Nonparametric techniques are usually based on ranks/ signs rather than actual data

Non-parametric Tests

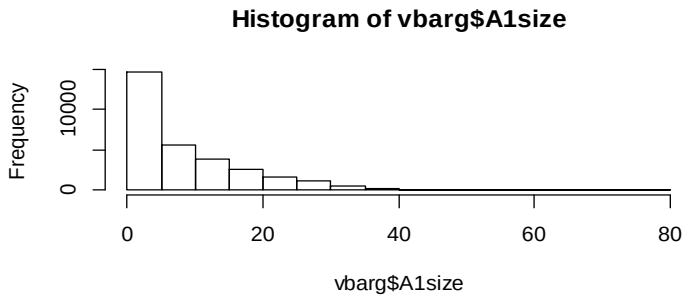
- Non-parametric methods are used when:
  - Data is ordinal
  - Data does not seem to follow any particular shape or distribution (e.g. Normal)
  - Assumptions underlying parametric test not met
  - A plot of the data appears to be very skewed
  - There are potential influential outliers in the dataset
  - Sample size is small

Note: Parametric tests are fairly robust to non-normality.
Data has to be very skewed to be a problem

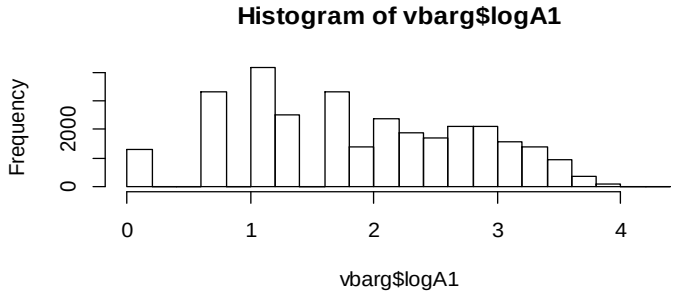
What can be done about non-normality?

If the data are not normally distributed, there are two options:

1. **Use a non-parametric test**
2. **Transform the dependent variable**



For positively skewed data, taking the log of the dependent variable often produces normally distributed values

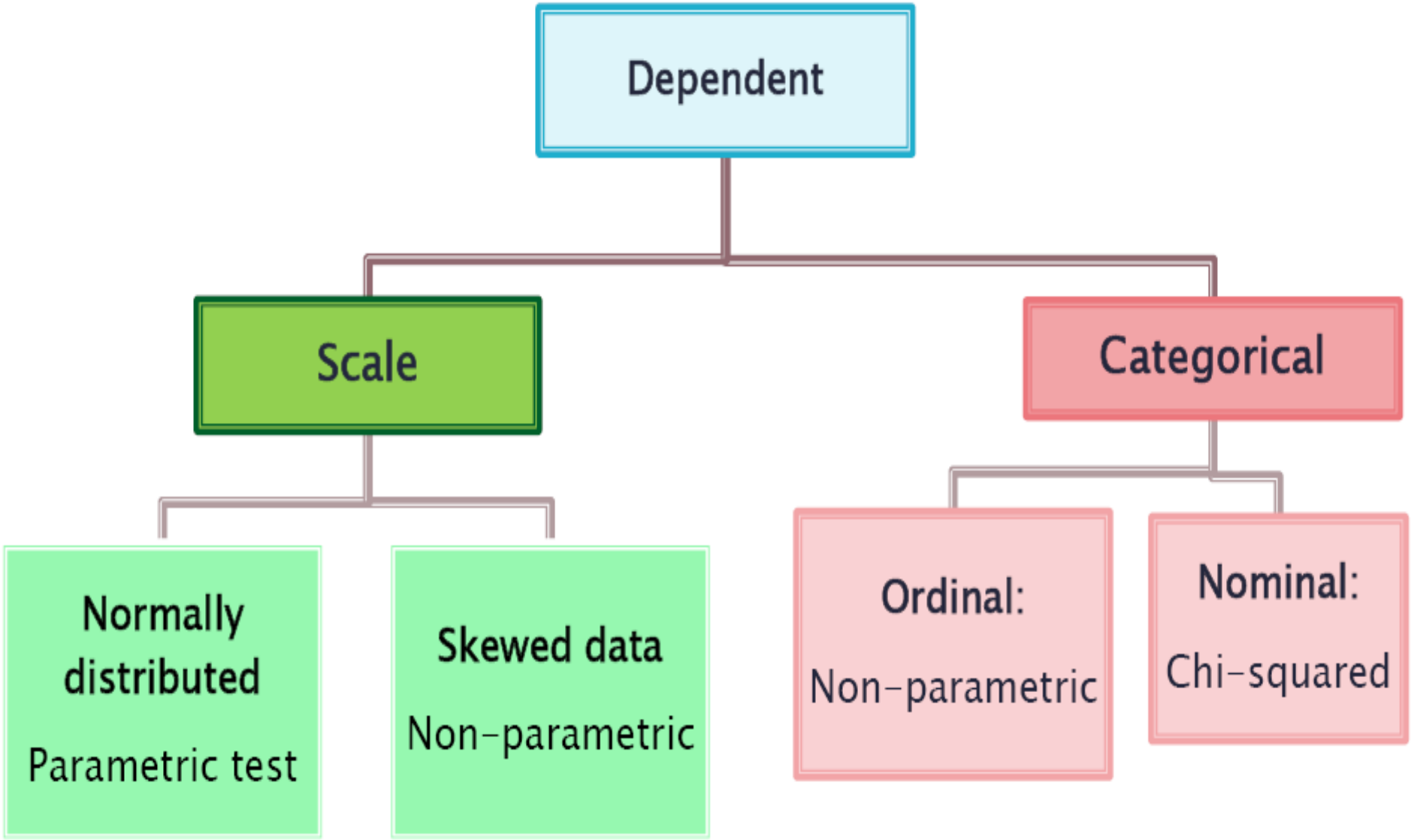


Non-parametric tests

Parametric test	What to check for normality	Non-parametric test
Independent t-test	Dependent variable by group	Mann-Whitney test
Paired t-test	Paired differences	Wilcoxon signed rank test
One-way ANOVA	Residuals/Dependent	Kruskal-Wallis test
Repeated measures ANOVA	Residuals	Friedman test
Pearson's Correlation Co-efficient	At least one of the variables should be normal	Spearman's Correlation Co-efficient
Linear Regression	Residuals	None - transform the data

Notes: The residuals are the differences between the observed and expected values.

Summary


Dependent
Scale
Categorical
Normally distributed
Parametric test
Skewed data
Non-parametric
Ordinal:
Non-parametric
Nominal:
Chi-squared

Statistical Hypothesis Testing: T-tests

(Paired or Independent (Unpaired) Data?)

We are often interested in comparing two sets of data, prior to analysis we must determine whether this data is paired or not.

T-tests are used to compare two population means

- **Paired Data:** Same individuals studied at two different times or under two conditions PAIRED T-TEST
- **Independent (Unpaired) Data:** Data is collected from two separate groups INDEPENDENT SAMPLES T-TEST (UNPAIRED T-TEST)

What is the t-distribution?

- The t-distribution is similar to the standard normal distribution but has an additional parameter called degrees of freedom (*df*). The df is calculated as the number of observations – 1.

  - For a paired t-test, *df* (or ν) = number of pairs – 1

  - For an independent t-test, $\nu = n_{group1} + n_{group2} - 2$

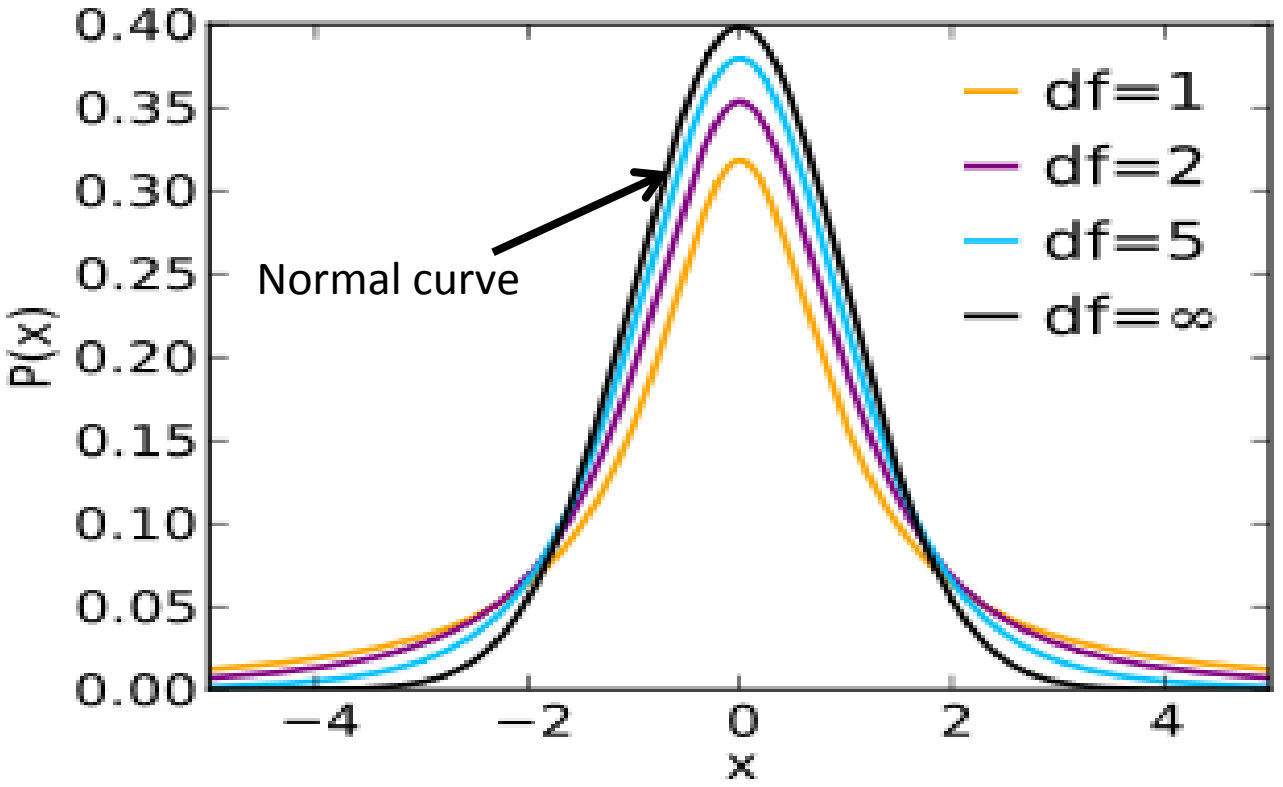
- **Used for small samples and when the population standard deviation is unknown**

- **Small sample sizes have heavier tails**

When we have small sample sizes (n < 30), we replace the normal distribution with ***Student's t distribution***, which has slightly less probability of being close to the mean and a somewhat larger probability of being in the tails. As the sample size increases, the critical values tend towards those of the normal distribution e.g. for a two tailed test with 5% significance, the critical value gets closer to 1.96, as the sample size increases.

Relationship to Normal Distrbution

As the sample size increases then the df increases, thus the t-distribution becomes more like the Normal Distribution.



Assumptions in t-Tests

Normality: Plot Histograms

One plot of the paired differences for any paired data.

Two (One for each group) for independent samples.

Don't have to be perfect, just roughly symmetric.

Equal Population Variances: Compare sample standard deviations

As a rough estimate, one should be no more than twice the other.

Do an F-test (Levene's in SPSS) to formally test for differences.

However the *t*-test is very robust to violations of the assumptions of Normality and equal variances, particularly for moderate (i.e. >30) and larger sample sizes.

Suggests ways of checking the assumptions in a t-test. For paired data it is the single column of differences that are assumed to be normal not each set of data from the two time points. The equal variances assumption does not apply to the paired t-test since there is only one sample (and population!)

What if the assumptions are not met?

There are alternative tests which do not have these assumptions

Test	Check	Equivalent non-parametric test
Independent t-test	Histograms of data by group	Mann-Whitney
Paired t-test	Histogram of paired differences	Wilcoxon signed rank

ANOVA Test

- Let us go through the procedure for one-way ANOVA
  - That means, one independent variable
- Multi-way ANOVA computations are cumbersome and very time consuming to do manually
  - So it is better to do computations using statistical packages

ANOVA Test

Compares the means of several groups.

- Which diet is the best?
  Dependent: Weight lost (Scale)
  Independent: Diet 1, 2 or 3 (Nominal)
- **Null Hypothesis: The mean weight lost on diets 1, 2 and 3 is the same.**
- **Alternative Hypothesis: The mean weight lost on diets 1, 2 and 3 are not the same.**

Summary Statistics

	Overall	Diet 1	Diet 2	Diet 3
Mean	3.85	3.3	3.03	5.15
Standard deviation	2.55	2.24	2.52	2.4
Number in group	78	24	27	27

- **Which diet was the best?**
- **Are the standard deviations similar?**

ANOVA Test

ANOVA = ANalysis Of VAriance

We compare variation **between** groups relative to variation **within** groups

Population variance estimated in two ways:

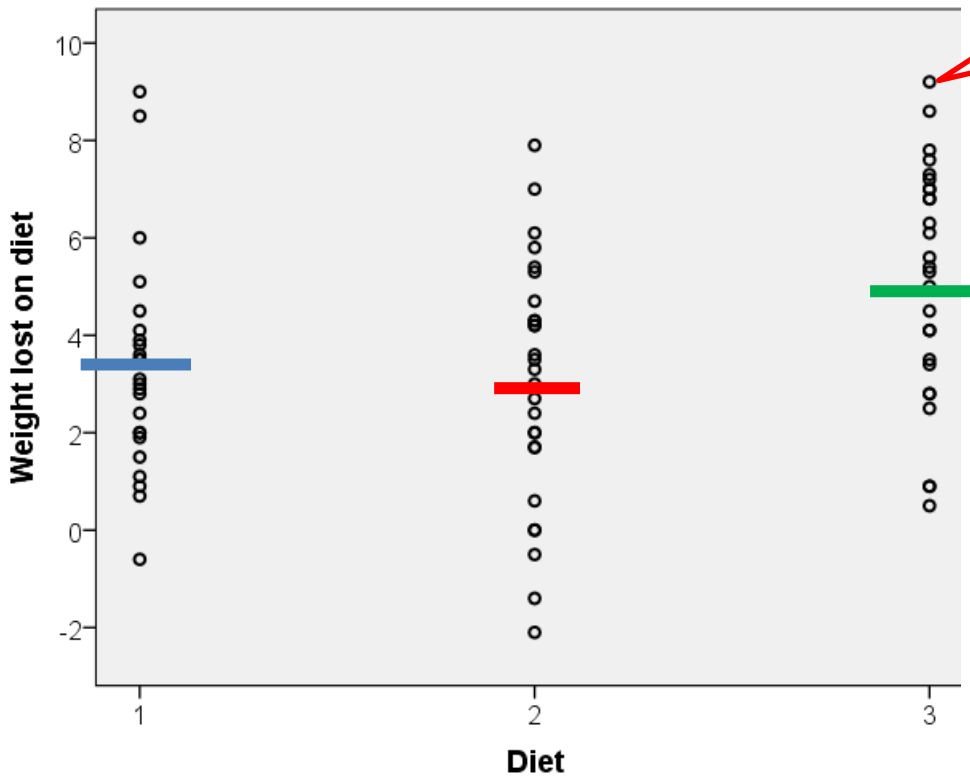
One based on variation **between** groups we call the **M**ean **S**quare due to **T**reatments/ **MST/** $\mathbf{MS_{between}}$

Other based on variation **within** groups we call the **M**ean **S**quare due to **E**rror/ **MSE/** $\mathbf{MS_{within}}$

Within the Group Variation

Residual =difference between an individual and the group mean

SS_{within}=sum of squared residuals

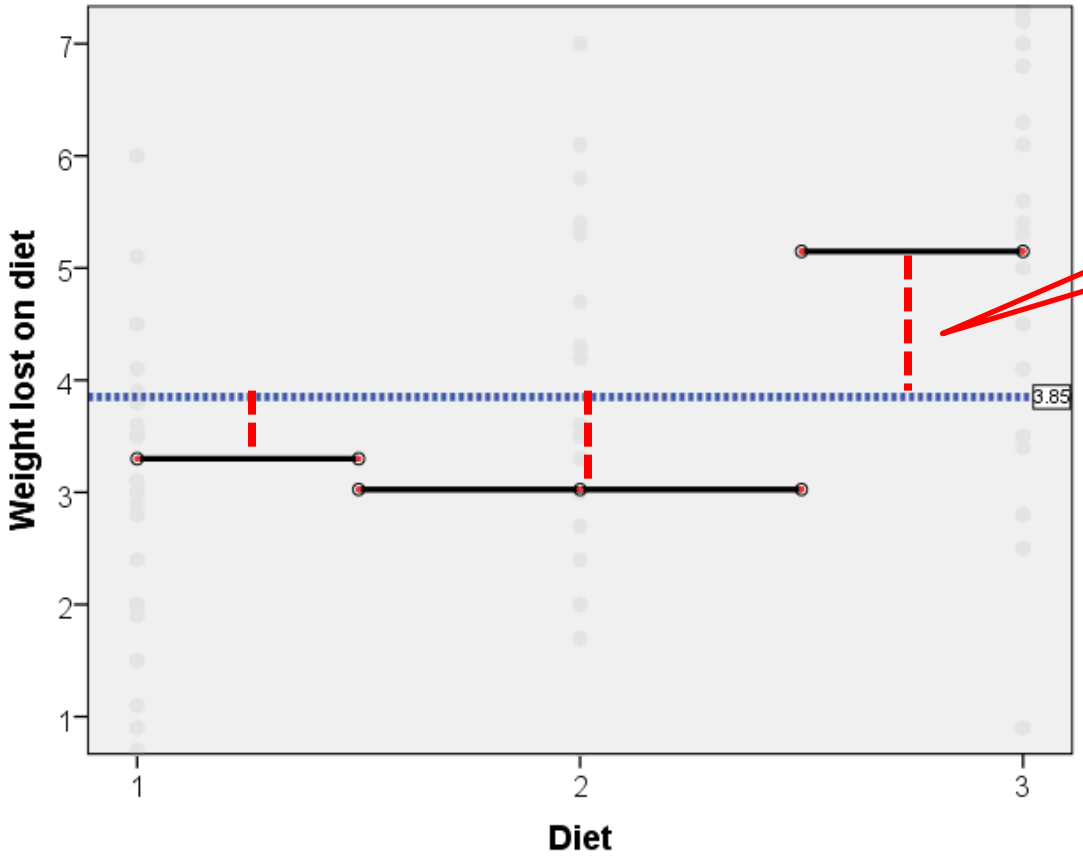


Person lost 9.2kg kg, hence residual =9.2 – 5.15 = 4.05

Mean weight lost on diet 3 = 5.15kg

Between the Group Variation

Differences between each group mean and the overall mean



Diet 3 difference
=5.15 – 3.85 = 1.3

Overall
mean = 3.85

Sum of Squares Calculations

K = Number of Groups

$$SS_{within} = \sum_{j=1}^{k}\sum_{i=1}^{n}\left(x_{ij} - \overline{x}_j\right)^2$$

$$= \sum_{i=1}^{24}\left(x_i - 3.3\right)^2 + \sum_{i=1}^{27}\left(x_i - 3.03\right)^2 + \sum_{i=1}^{27}\left(x_i - 5.15\right)^2 = 430.179$$

$$SS_{Between} = \sum_{j=1}^{k} n_j\left(\overline{x}_j - \overline{x}_T\right)^2$$

$$= 24\left(3.3 - 3.85\right)^2 + 27\left(3.03 - 3.85\right)^2 + 27\left(5.15 - 3.85\right)^2 = 71.094$$

ANOVA Test Statistics

Summary ANOVA

Source	Sum of Squares	Degrees of Freedom	Variance Estimate (Mean Square)	*F* Ratio
Between	SS_B	$K-1$	$MS_B = \frac{SS_B}{K-1}$	$\frac{MS_B}{MS_W}$
Within	SS_W	$N-K$	$MS_W = \frac{SS_W}{N-K}$	
Total	$SS_T = SS_B + SS_W$	$N-1$		

Test Statistic (usually reported)

N = Total observations in all groups,

K = Number of groups

Test Statistic (by hand)

Filling in the boxes

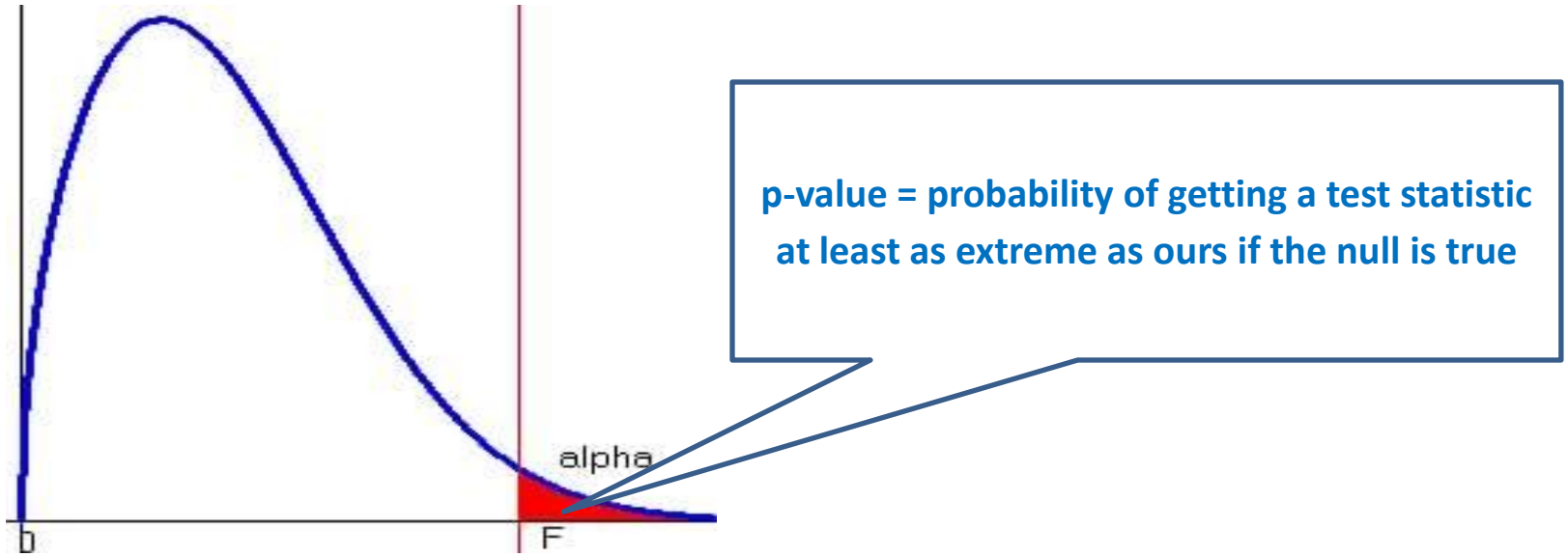
	Sum of Squares	Degrees of Freedom	Mean Square	F-ratio (Test Statistic)
$SS_{between}$	71.045	2	35.522	**6.193**
SS_{within}	430.180	75	5.736	
SS_{total}	501.275	77		

F-ratio = Mean between group sum of squared differences / Mean within group sum of squared differences

If F-ratio > 1, then there is a bigger difference between the groups than within the groups

P-value

- The p-value for ANOVA is calculated using the F-distribution
- If we repeated the experiment several times, then we would get a variety of test statistics



One Way ANOVA

$$\text{Test Statistic} = \frac{\text{between group variation}}{\text{within group variation}} = \frac{MS_{Diet}}{MS_{Error}} = 6.197$$

Tests of Between-Subjects Effects

Dependent Variable: Weight lost on diet (kg)

Source	Type III Sum of Squares	df	Mean Square	F	Sig.
Corrected Model	71.094[a]	2	35.547	6.197	.003
Intercept	1137.494	1	1137.494	198.317	.000
Diet	71.094	2	35.547	6.197	.003
Error	430.179	75	5.736		
Total	1654.350	78			
Corrected Total	501.273	77			

a. R Squared = .142 (Adjusted R Squared = .119)

$MS_{between}$

MS_{within}

There was a significant difference in weight lost between the diets (p=0.003)

Post-hoc Tests

If there is a significant ANOVA result, then Pairwise comparisons are made

These are t-tests with adjustments to keep the type 1 error to a minimum

- Tukey's and Scheffe's tests are the most commonly used post-hoc tests.
- Hochberg's GT2 is better where the sample sizes for the groups are very different.

Post-hoc Tests

Which diets are significantly different?

Multiple Comparisons

Dependent Variable: Weight lost on diet (kg)

	(I) Diet	(J) Diet	Mean Difference (I-J)	Std. Error	Sig.	95% Confidence Interval	
						Lower Bound	Upper Bound
Tukey HSD	1	2	.2741	.67188	.912	-1.3325	1.8806
		3	-1.8481*	.67188	.020	-3.4547	-.2416
	2	1	-.2741	.67188	.912	-1.8806	1.3325
		3	-2.1222*	.65182	.005	-3.6808	-.5636
	3	1	1.8481*	.67188	.020	.2416	3.4547
		2	2.1222*	.65182	.005	.5636	3.6808

Write up the results and conclude with which diet is the best.

Pairwise Comparisons

Test	p-value
Diet 1 vs Diet 2	P = 0.912
Diet 1 vs Diet 3	P = 0.02
Diet 2 vs Diet 3	P = 0.005

There is no significant difference between Diets 1 and 2 but there is between diet 3 and diet 1 ($p = 0.02$) and diet 2 and diet 3 ($p = 0.005$).

The mean weight lost on Diets 1 (3.3kg) and 2 (3kg) are less than the mean weight lost on diet 3 (5.15kg).

Assumptions for ANOVA Test

Assumption	How to check	What to do if assumption not met
Normality: The residuals (difference between observed and expected values) should be normally distributed	Histograms/ QQ plots/ normality tests of residuals	Do a Kruskall-Wallis test which is non-parametric (does not assume the normality)
Homogeneity of variance (each group should have a similar standard deviation)	Levene's test	Welch test instead of ANOVA and Games-Howell for post-hoc or Kruskall-Wallis

ANOVA Illustrated

- Let's illustrate the idea with the following example:

  Suppose we have designed a new text entry technique for mobile phones. We think the design is good. In fact, we feel that our method is *better* than the most widely used state-of-the-art techniques: multi-tap and T9. We decide to undertake some empirical research to evaluate our design invention and to compare it with these current techniques?

  Suppose "better" is defined in terms of error rate

Empirical Data

- In order to ascertain the validity of our claim, we conducted the experiments and thus collected the following empirical data (error rate of participants under different test conditions)

Participants	Our Method	Multi-tap	T9
1	3	5	7
2	2	2	4
3	1	4	5
4	1	2	3
5	4	3	6

ANOVA Steps - 1

- Now Let's compute means, standard deviations (SD) and variances for each test condition (over all participants)

	Our Method	Multi-tap	T9
Mean	**2.20**	**3.20**	**5.00**
SD	**1.30**	**1.30**	**1.58**
Variance	**1.70**	**1.70**	**2.50**

ANOVA Steps - 1

- **Also calculate "grands" – values involving all irrespective of groups**
  - **Grand Mean (mean of means) = 3.467**
  - **Grand SD (w.r.t. grand mean) = 1.767**
  - **Grand Variance (w.r.t. grand mean) = 3.124**

ANOVA Steps - 2

- **Calculate "total sum of squares (SS_T)"**

$$SS\_T = \sum (x\_i - mean\_grand)^2$$
$$= 43.74$$

Where, *x_i* is the error rate value of the i-th participant (among all)

ANOVA Steps - 2

- **An associated concept is the degrees of freedom (DoF (df)), which is the number of observations that are free to vary**
- **DoF (df) can be calculated simply as the (number of things used to calculate – 1)**
  - **For SS_T calculation, DoF (df) = N-1**

ANOVA Steps - 3

- **Next calculate the "model sum of square (SS_M)"**
  - **Calculate (mean_group_*i*-mean_grand) for the i-th group**
  - **Square the above**
  - **Multiply by *n_i*, the number of participants in the *i*-th group**
  - **Sum for all groups**

ANOVA Steps - 3

- **In the example,**

$$SS\_M = 5(2.200-3.467)^2 + 5(3.200-3.467)^2 + 5(5.000-3.467)^2$$

$$= 20.135$$

- **DoF (df) = Number of group means – 1**

= 3 - 1 = 2 (in our example)

ANOVA Steps - 4

- **Calculate the "residual sum of square (SS_R)" and the corresponding DoF**

  SS_R = SS_T – SS_M

  DoF (SS_R) = DoF (SS_T) – DoF (SS_M)

- **Thus, in this example,**

  SS_R = 43.74 – 20.14 = 23.60

  DoF (SS_R) = 14 – 2 = 12

ANOVA Steps - 5

- **Calculate two "average sum of squares" or "mean squares (MS)"**
- **Model MS (MS_M) = SS_M/DoF(SS_M)**

  = 20.135/2 = 10.067 (for our example)
- **Residue MS (MS_R) = SS_R/DOF(SS_R)**

  = 23.60/12 = 1.967 (for our example)

ANOVA Steps - 6

- Calculate the "F-ratio" (simply divide MS_M by MS_R)
  - F-ratio (F) = 10.067/1.967 = 5.12 (for our example)
- DoFs associated with the F-ratio are the DoFs used to calculate the DoF(SS_M) and DoF(SS_R)
  - In our case, these are 2, 12 respectively
- Hence, in our case, the F-ratio would be written as F(2, 12) i.e. F-ratio (F) = 5.12

ANOVA Steps - 6

- Look up the critical value of F-ratio (F)
  - The critical values for different "significance levels" / thresholds (α) are available in a tabular form
  - The critical values signifies the value of F that we would expect to get by chance for α% of tests

ANOVA Steps - 6

- Example (for more details, please see the next slide)
  - To find the critical value of F(2, 12) from the Table for α=.05, look at 2nd column, 12th row for .05
  - Which is 3.89
  - That means, 3.89 is the F-value we would expect to get by chance for 5% of the tests.

df (Denominator)	p	df (Numerator)																
		1	2	3	4	5	6	7	8	9	10	15	20	25	30	40	50	1000
1	.05	161.45	199.50	215.71	224.58	230.16	233.99	236.77	238.88	240.54	241.88	245.95	248.01	249.26	250.10	251.14	251.77	254.19
	.01	4052.18	4999.50	5403.35	5624.58	5763.65	5858.99	5928.36	5981.07	6022.47	6055.85	6157.31	6208.74	6239.83	6260.65	6286.79	6302.52	6362.70
2	.05	18.51	19.00	19.16	19.25	19.30	19.33	19.35	19.37	19.38	19.40	19.43	19.45	19.46	19.46	19.47	19.48	19.49
	.01	98.50	99.00	99.17	99.25	99.30	99.33	99.36	99.37	99.39	99.40	99.43	99.45	99.46	99.47	99.47	99.48	99.50
3	.05	10.13	9.55	9.28	9.12	9.01	8.94	8.89	8.85	8.81	8.79	8.70	8.66	8.63	8.62	8.59	8.58	8.53
	.01	34.12	30.82	29.46	28.71	28.24	27.91	27.67	27.49	27.35	27.23	26.87	26.69	26.58	26.50	26.41	26.35	26.14
4	.05	7.71	6.94	6.59	6.39	6.26	6.16	6.09	6.04	6.00	5.96	5.86	5.80	5.77	5.75	5.72	5.70	5.63
	.01	21.20	18.00	16.69	15.98	15.52	15.21	14.98	14.80	14.66	14.55	14.20	14.02	13.91	13.84	13.75	13.69	13.47
5	.05	6.61	5.79	5.41	5.19	5.05	4.95	4.88	4.82	4.77	4.74	4.62	4.56	4.52	4.50	4.46	4.44	4.37
	.01	16.26	13.27	12.06	11.39	10.97	10.67	10.46	10.29	10.16	10.05	9.72	9.55	9.45	9.38	9.29	9.24	9.03
6	.05	5.99	5.14	4.76	4.53	4.39	4.28	4.21	4.15	4.10	4.06	3.94	3.87	3.83	3.81	3.77	3.75	3.67
	.01	13.75	10.92	9.78	9.15	8.75	8.47	8.26	8.10	7.98	7.87	7.56	7.40	7.30	7.23	7.14	7.09	6.89
7	.05	5.59	4.74	4.35	4.12	3.97	3.87	3.79	3.73	3.68	3.64	3.51	3.44	3.40	3.38	3.34	3.32	3.23
	.01	12.25	9.55	8.45	7.85	7.46	7.19	6.99	6.84	6.72	6.62	6.31	6.16	6.06	5.99	5.91	5.86	5.66
8	.05	5.32	4.46	4.07	3.84	3.69	3.58	3.50	3.44	3.39	3.35	3.22	3.15	3.11	3.08	3.04	3.02	2.93
	.01	11.26	8.65	7.59	7.01	6.63	6.37	6.18	6.03	5.91	5.81	5.52	5.36	5.26	5.20	5.12	5.07	4.87
9	.05	5.12	4.26	3.86	3.63	3.48	3.37	3.29	3.23	3.18	3.14	3.01	2.94	2.89	2.86	2.83	2.80	2.71
	.01	10.56	8.02	6.99	6.42	6.06	5.80	5.61	5.47	5.35	5.26	4.96	4.81	4.71	4.65	4.57	4.52	4.32
10	.05	4.96	4.10	3.71	3.48	3.33	3.22	3.14	3.07	3.02	2.98	2.85	2.77	2.73	2.70	2.66	2.64	2.54
	.01	10.04	7.56	6.55	5.99	5.64	5.39	5.20	5.06	4.94	4.85	4.56	4.41	4.31	4.25	4.17	4.12	3.92
11	.05	4.84	3.98	3.59	3.36	3.20	3.09	3.01	2.95	2.90	2.85	2.72	2.65	2.60	2.57	2.53	2.51	2.41
	.01	9.65	[illegible]	6.22	5.67	5.32	5.07	4.89	4.74	4.63	4.54	4.25	4.10	4.01	3.94	3.86	3.81	3.61
12	.05	4.75	3.89	3.49	3.26	3.11	3.00	2.91	2.85	2.80	2.75	2.62	2.54	2.50	2.47	2.43	2.40	2.30
	.01	9.33	6.93	5.95	5.41	5.06	4.82	4.64	4.50	4.39	4.30	4.01	3.86	3.76	3.70	3.62	3.57	3.37
13	.05	4.67	3.81	3.41	3.18	3.03	2.92	2.83	2.77	2.71	2.67	2.53	2.46	2.41	2.38	2.34	2.31	2.21
	.01	9.07	6.70	5.74	5.21	4.86	4.62	4.44	4.30	4.19	4.10	3.82	3.66	3.57	3.51	3.43	3.38	3.18
14	.05	4.60	3.74	3.34	3.11	2.96	2.85	2.76	2.70	2.65	2.60	2.46	2.39	2.34	2.31	2.27	2.24	2.14
	.01	8.86	6.51	5.56	5.04	4.69	4.46	4.28	4.14	4.03	3.94	3.66	3.51	3.41	3.35	3.27	3.22	3.02
15	.05	4.54	3.68	3.29	3.06	2.90	2.79	2.71	2.64	2.59	2.54	2.40	2.33	2.28	2.25	2.20	2.18	2.07
	.01	8.68	6.36	5.42	4.89	4.56	4.32	4.14	4.00	3.89	3.80	3.52	3.37	3.28	3.21	3.13	3.08	2.88
16	.05	4.49	3.63	3.24	3.01	2.85	2.74	2.66	2.59	2.54	2.49	2.35	2.28	2.23	2.19	2.15	2.12	2.02
	.01	8.53	6.23	5.29	4.77	4.44	4.20	4.03	3.89	3.78	3.69	3.41	3.26	3.16	3.10	3.02	2.97	2.76
17	.05	4.45	3.59	3.20	2.96	2.81	2.70	2.61	2.55	2.49	2.45	2.31	2.23	2.18	2.15	2.10	2.08	1.97
	.01	8.40	6.11	5.18	4.67	4.34	4.10	3.93	3.79	3.68	3.59	3.31	3.16	3.07	3.00	2.92	2.87	2.66
18	.05	4.41	3.55	3.16	2.93	2.77	2.66	2.58	2.51	2.46	2.41	2.27	2.19	2.14	2.11	2.06	2.04	1.92
	.01	8.29	6.01	5.09	4.58	4.25	4.01	3.84	3.71	3.60	3.51	3.23	3.08	2.98	2.92	2.84	2.78	2.58
19	.05	4.38	3.52	3.13	2.90	2.74	2.63	2.54	2.48	2.42	2.38	2.23	2.16	2.11	2.07	2.03	2.00	1.88
	.01	8.18	5.93	5.01	4.50	4.17	3.94	3.77	3.63	3.52	3.43	3.15	3.00	2.91	2.84	2.76	2.71	2.50
20	.05	4.35	3.49	3.10	2.87	2.71	2.60	2.51	2.45	2.39	2.35	2.20	2.12	2.07	2.04	1.99	1.97	1.85
	.01	8.10	5.85	4.94	4.43	4.10	3.87	3.70	3.56	3.46	3.37	3.09	2.94	2.84	2.78	2.69	2.64	2.43
22	.05	4.30	3.44	3.05	2.82	2.66	2.55	2.46	2.40	2.34	2.30	2.15	2.07	2.02	1.98	1.94	1.91	1.79
	.01	7.95	5.72	4.82	4.31	3.99	3.76	3.59	3.45	3.35	3.26	2.98	2.83	2.73	2.67	2.58	2.53	2.32

Implication

- Thus, we get the critical value = 3.89 for F(2,12), $\alpha = 0.05$
- Note that F(2, 12)=5.12 > the critical value
  - Implies that the effect of test conditions has a significant effect on the outcome w.r.t. α=.05

Reporting F-Statistic

- We can report the result as “our proposed method has a significant effect on reducing user errors [$F(2,12)=5.12$, $p < 0.05$] as compared to the other methods”.
- If it is found that the effect is not significant, it is reported as “our method has no significant effect on reducing user errors [$F(1,9)=0.634$, ns] as compared to the other methods”.

A Note of Caution

- **ANOVA requires that**
  - Empirical Data should have normally distributed sampling distribution and from a normally distributed population
  - Variances in each experimental condition are fairly similar
  - Observations should be independent
  - Dependent variables are measured on at least an interval scale
- The first two may be ignored if group sizes are equal
  - Otherwise, ALL conditions MUST have to be met

HCI: Empirical Research Case Study

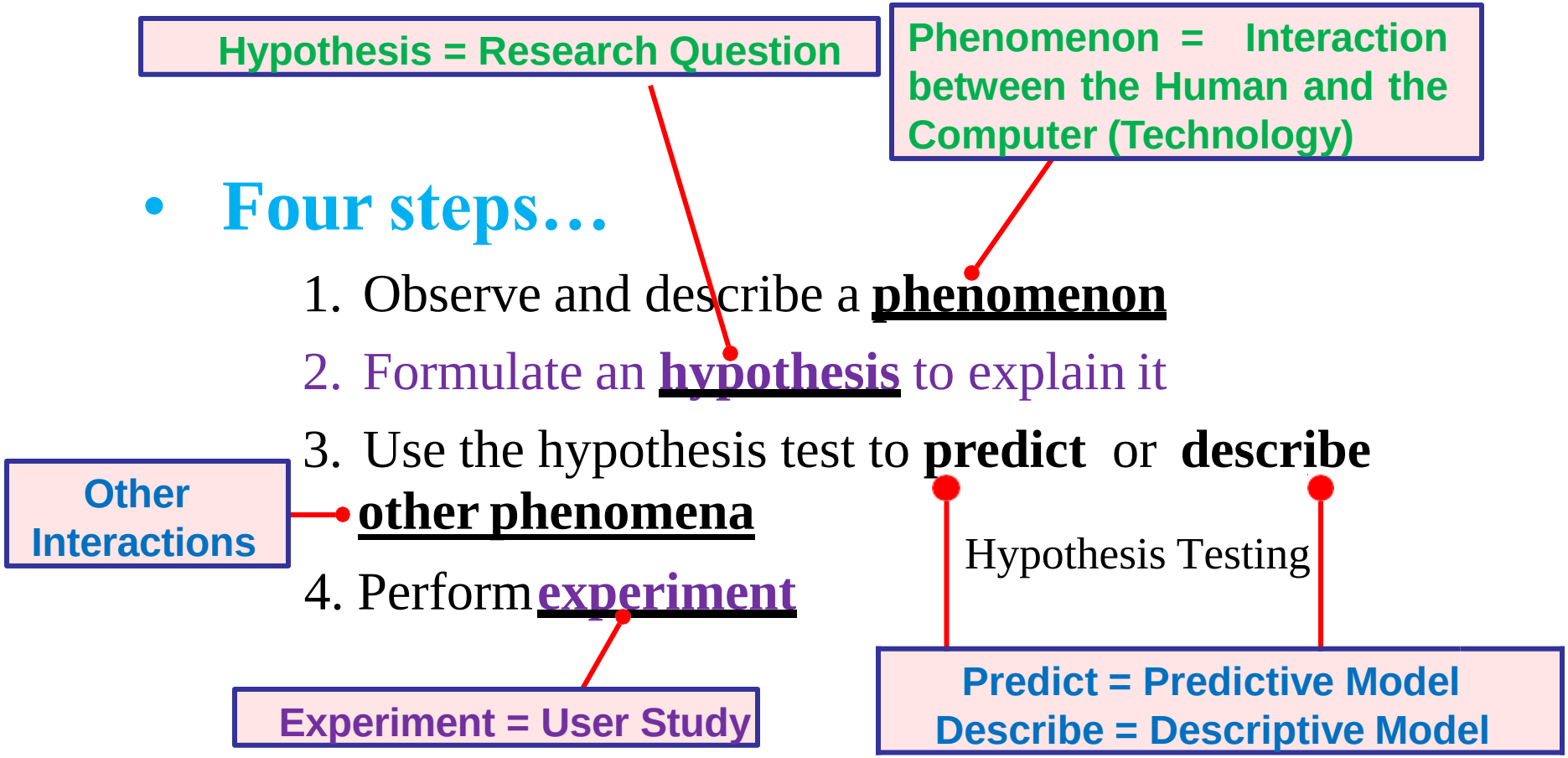
Learning Objective

- In the previous lectures, we discussed different empirical research methods involved in HCI
- We introduced several concepts such as testable empirical question formulation, experiment design, data collection and statistical analysis of data
- In this lecture, we shall consider a case study

Case Study

- Suppose, we want to study the application of eye tracking technology for the text entry task (i.e., typing through eye gaze). Let us initiate an empirical inquiry to explore performance limits and the capabilities of various feedback modalities for keys in on-screen keyboards used with eye gaze based typing.
  - Suppose four feedback modalities are considered by us, namely: Audio only [A], Click+Visual [C], Speech+Visual [S], Visual only [V].

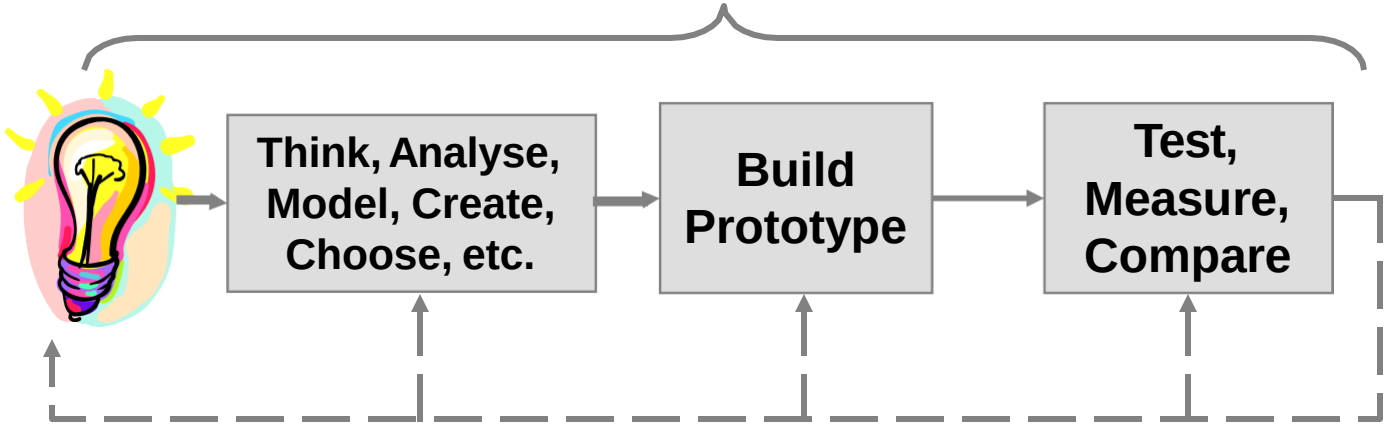
Steps in Empirical Research (Classical View)



Steps in Empirical Research (Practical View)

Phase I – The Prototype

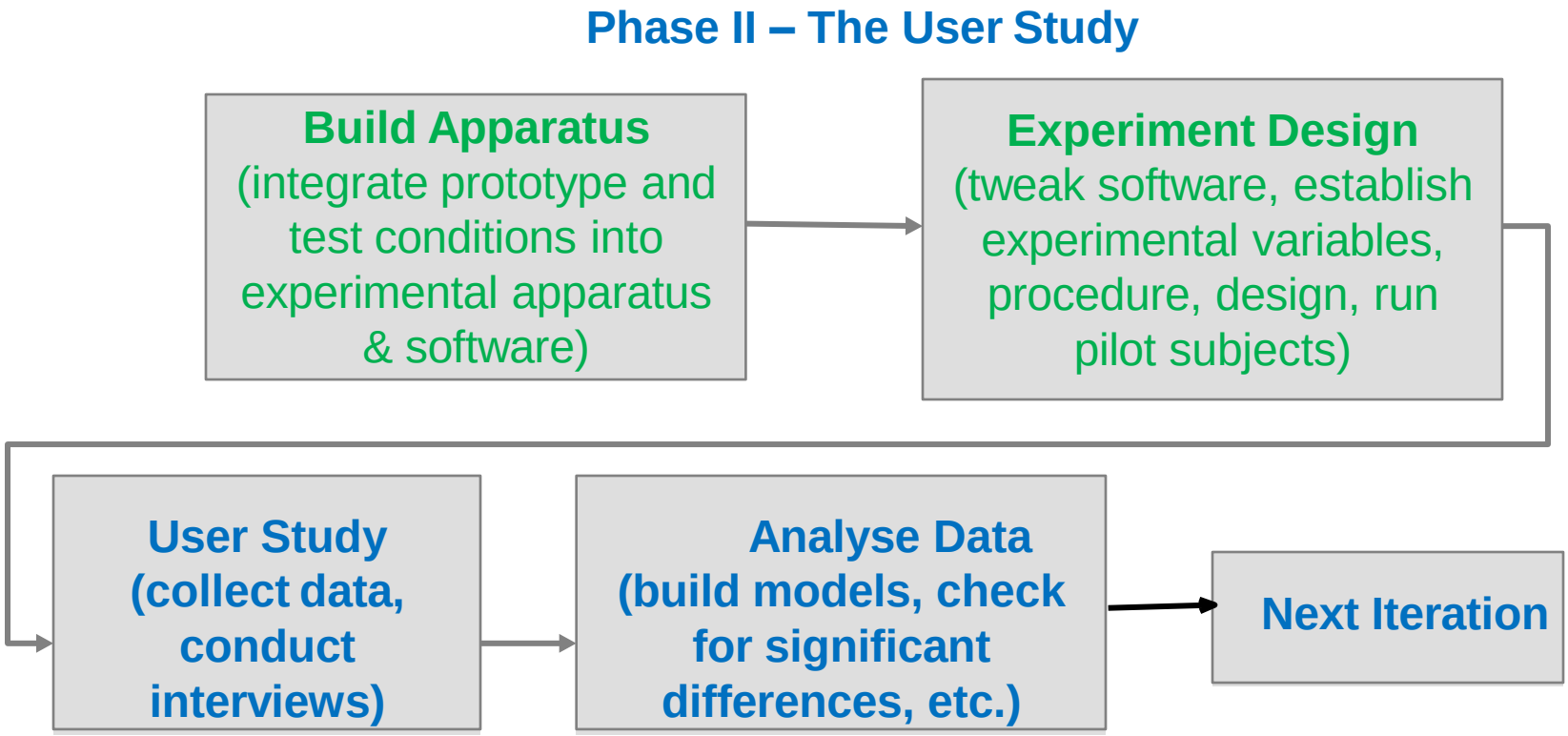
Steps 1-3 (previous slide)



Iterations are frequent, unstructured, intuitive, informed, …

Empirical Research Questions "take shape" (i.e., certain measurable aspects of the interaction suggest "test conditions", and "tasks" for empirical inquiry

Steps in Empirical Research (Practical View)



The User Study

- Describe the participants employed for our study
  - Thirteen, volunteers, recruited from university campus, age, gender, computer experience, eye tracking/typing experience
- Apparatus
  - Describe hardware and software, etc.

The User Study

- Experiment Design
  - We decided to have a 4 x 4 repeated measures design
  - There are two independent variables (factors) with four levels each
    - Feedback modality (with the levels A, C, S, V)
    - Users were asked to enter blocks of text at a time and four such blocks were there for each user. So, “block” is a factor with four levels 1, 2, 3, 4

Note: Audio only [A], Click+Visual [C], Speech+Visual [S], Visual only [V].

The User Study

- Experiment Design
  - We have identified dependent variables (measures)
    - Speed of text entry (in "words per minutes")
    - Accuracy of text entry (in "percentage of characters in error")
    - Key selection activity (in "keystrokes per character")
    - Also… responses to "broader" questions
  - Order of Conditions
    - Feedback modality order differed for each user (Latin Square)

The User Study

- Procedure for Data Collection
  - We first explained to the participants the general objectives of the experiment
  - Then the eye tracking apparatus was calibrated
  - The participants were put through some practice trials for familiarization
  - Afterwards, let us begin data collection

The User Study

- Procedure for Data Collection
  - Phrases of text presented to the participants by experimental software
  - Participants instructed to enter phrases “as quickly and accurately as possible”
  - Five phrases were entered by the participants per block
  - Total number of phrases entered in this experiment is found to be 13 x 4 x 4 x 5 = 1040

Experiment Replication

- The description of the experimental methodology (i.e., participants, participant selection, apparatus, design, procedure) must be sufficient to allow the experiment to be replicated by other researchers
  - This is necessary to allow the possibility for the results to be verified or refuted as part of performance evaluation
  - **An experiment that cannot be replicated is useless**

Data Tables

- Next slide contains example data on text entry speed, recorded in this empirical study (user study)
  - Create a Table to arrange data i.e. Data Table
  - From the Data Table, let us calculate other quantities such as the grand mean = 6.96 WPM
  - The Data Table also allows us to make salient observations (for example, 4th block speed for best condition was…)

Data Tables

Note: Audio only [A], Click+Visual [C], Speech+Visual [S], Visual only [V].

Factors and Levels

Speed																
	A	A	A	A	C	C	C	C	S	S	S	S	V	V	V	V	
Participant	1	2	3	4	1	2	3	4	1	2	3	4	1	2	3	4	Mean
1	6.17	7.19	7.04	7.09	6.76	7.40	7.54	7.94	6.44	6.17	7.84	6.81	5.20	6.29	7.39	7.63	6.93
2	6.71	7.25	7.05	7.15	7.73	7.57	8.04	7.26	7.00	6.75	7.68	7.46	7.50	7.07	7.32	7.06	7.29
3	6.80	6.65	7.62	7.98	6.61	7.18	7.34	8.19	6.65	7.53	7.09	7.90	5.73	7.24	6.94	7.13	7.16
5	6.30	6.31	7.59	7.38	6.85	7.64	7.58	7.88	7.07	6.43	7.26	7.65	6.75	6.59	6.97	7.72	7.12
7	6.68	6.89	7.32	7.51	7.00	7.81	7.64	7.2						72	7.57	7.20	7.11
8	6.08	6.55	6.83	5.92	7.44	6.93	7.56	6.4						3	7.45	7.16	6.98
9	7.62	7.01	6.60	7.07	6.91	6.81	6.91	7.73	6.50	7.57	7.59	7.80	6.62	7.06	7.16	7.41	7.15
10	5.88	5.71	7.33	7.11	6.66	7.97	7.64	8.15	6.35	7.21	6.56	7.33	5.00	6.97	6.54	6.36	6.80
12	6.89	7.61	7.42	7.88	7.79	8.28	8.20	8.39	6.62	6.87	7.99	8.23	**9.57**	8.17	7.91	7.09	7.81
13	6.85	6.57	8.14	6.00	5.92	7.89	7.49	6.98	6.05	7.45	5.34	7.46	7.21	6.81	6.80	8.24	6.95
14	5.37	5.56	**6.04**	6.86	6.20	6.82	7.71	7.76	5.85	6.37	6.74	6.69	5.98	6.43	6.38	5.87	6.41
15	5.51	6.12	6.32	7.00	6.16	6.49	7.21	7.19	5.65	6.52	6.49	7.10	5.31	6.88	6.36	6.93	6.45
16	5.88	7.18	5.95	6.00	4.85	6.98	7.37	6.98	6.88	6.21	4.96	5.34	6.72	7.14	4.96	6.80	6.26
																	6.96

Outlier

Each cell is the mean for five phrases of input

Statistical Analysis of Data

- The data recorded in the Data Table is analyzed statistically to identify the significant effects
- For example, we may have the following findings:
  - Significant effect for Feedback mode [$F(3,36)=8.77$, $p<.0005$]
  - Insignificant Effect for Feedback mode by the block interaction [$F(9,108)=0.767$, ns]

Data Tables

- Apart from the main tables, other tables are also created, which helps in making more useful observations
- The next slide shows an example of a summary table created from the data on text entry speed

Data Tables

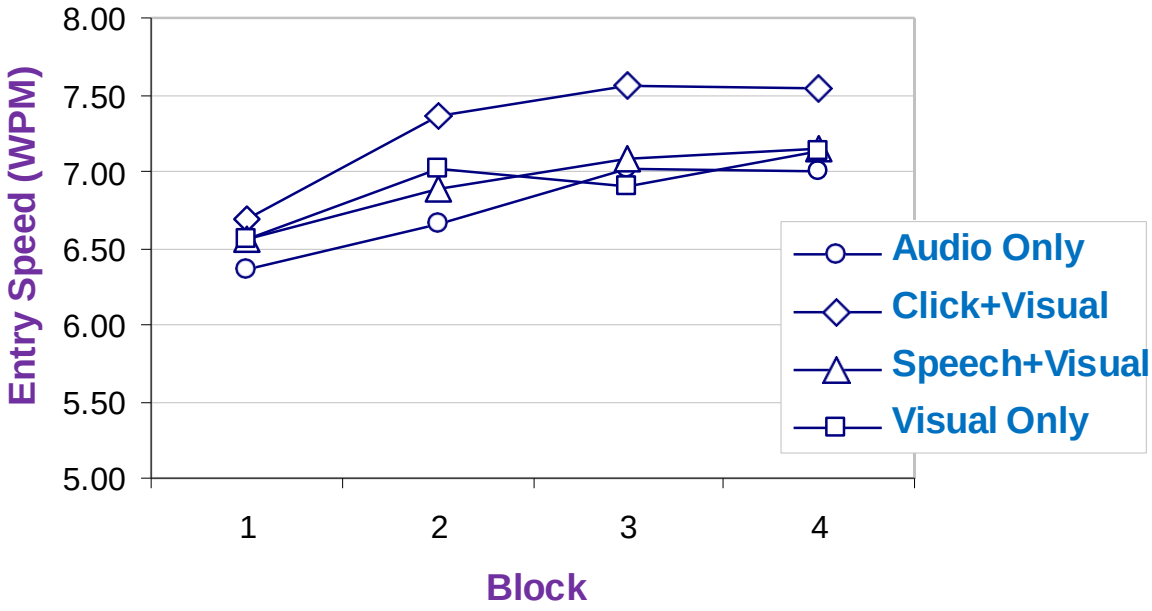
Speed (WPM)					
	Feedback Mode				
Block	Audio Only	Click+Visual	Speech+Visual	Visual Only	Mean
1	6.36	6.68	6.56	6.55	**6.54**
2	6.66	7.37	6.88	7.02	**6.98**
3	7.02	7.56	7.09	6.90	**7.14**
4	**7.00**	**7.55**	**7.14**	**7.12**	**7.20**
Mean	**6.76**	**7.29**	**6.92**	**6.90**	**6.97**

5.7% faster on 4th block

Each cell is the mean for 13 participants

Charts/Graphs

- Also let us create graphs/charts to visualize findings



The Broader Questions

- Along with the data analysis, an empirical study typically collects the direct feedback from all the participants on “broader” questions
  - For example, all participants can be asked about their preferences, satisfaction levels or even their suggestions for further improvements

The Broader Questions

- In the study, we asked the participants to rank (between 1 to 4) the feedback mode based on their personal preferences
- We obtained the following results:
  - Six of thirteen participants gave a 1st place ranking to the fastest feedback modality

The Broader Questions

- The results obtained are not strong enough to come to any conclusions
  - A reason may be that the differences just weren't large enough for the participants to really tell the difference in overall performance
- However, we have also made another observation, namely ten of the thirteen participants gave a 1st or 2nd place ranking to the fastest feedback modality
  - This can be treated as a strong indication that the better performance yields a better preference rating

What's Missing?

- **The case study just described show that the user study involves collection and analysis of usage data as well as participants' feedback**
- However, that's not all (it misses an important aspect of empirical research)
  - There is no theoretical account of the phenomena

What's Missing?

- **There is no delineation, description, categorization of known and observed behaviors (…that can form such a theoretical account)**
- **It is not sufficient to simply observe and conclude, it is also necessary to theorize about the observations (e.g., why the text entry speed is the least in a particular feedback mode)**

Empirical Research in HCI

- **The direct conclusions from observations help us to decide an interaction method; a theory about observed behavior can help us do much more**
  - **Such theories can eliminate the need for further investigations as well as can suggest the suitable ways for further improvement**
- **Such theories, if found, are another motivation for conducting empirical research in HCI (in fact, many models in HCI have been derived empirically)**

Case Study: The Case for a Model

- Is there a "model of interaction" suggested by the observations in this case study?
- Perhaps. Here's one possibility
  - All gaze point changes were logged as "events".
    What was the total number of such events?
    Are there categories of such events?
- The identification, labelling, and tabulation of such could form the basis of a model of interaction for eye typing

Analysis of Variance (ANOVA)

Analysis of variance (ANOVA)

- is a statistical technique that is used to check if the means of two or more groups are significantly different from each other.
- ANOVA checks the impact of one or more factors by comparing the means of different samples.
- One-way or two-way refers to the number of independent variables in the analysis of variance test.

Steps:

1. Define Hypothesis (Null and alternative)
2. Calculate the sum of squares
3. Determine degrees of freedom

4. Find the mean for each of the groups.
5. Find the **overall mean** (the mean of the groups combined).
6. Find the **Within Group Variation**; the total deviation of each member's score from the Group Mean.
7. Find the **Between Group Variation**: the deviation of each Group Mean from the Overall Mean.
8. Find the F statistic: the ratio of Between Group Variation to Within Group Variation.

Formulas Used

One-Way ANOVA Table					
Source	**Degrees of Freedom** **DF**	**Sum of Squares** **SS**	**Mean Square** **MS**	**F-Stat**	**P-Value**
Between Groups	$k - 1$	SS_B	$MS_B = SS_B / (k - 1)$	$F = MS_B / MS_W$	Right tail of $F(k-1,N-k)$
Within Groups	$N - k$	SS_W	$MS_W = SS_W / (N - k)$		
Total:	$N - 1$	$SS_T = SS_B + SS_W$			

Between Groups Degrees of Freedom: $\mathbf{DF = k - 1}$, where $\mathbf{k}$ is the number of groups

Within Groups Degrees of Freedom: $\mathbf{DF = N - k}$, where $\mathbf{N}$ is the total number of subjects

Total Degrees of Freedom: $\mathbf{DF = N - 1}$

Sum of Squares Between Groups: $\mathbf{SS_B = \sum_{i=1}^{k} n_i (\bar{x}_i - \bar{x})^2}$, where $\mathbf{n_i}$ is the number of subjects in the i-th group

Sum of Squares Within Groups: $\mathbf{SS_W = \sum_{i=1}^{k} (n_i - 1) S_i^2}$, where $\mathbf{S_i}$ is the standard deviation of the i-th group

$$SS_{within} = \Sigma (X_i - \overline{X}_j)^2$$

Total Sum of Squares: $\mathbf{SS_T = SS_B + SS_W}$

Mean Square Between Groups: $\mathbf{MS_B = SS_B / (k - 1)}$

Mean Square Within Groups: $\mathbf{MS_W = SS_W / (N - k)}$

F-Statistic (or F-ratio): $\mathbf{F = MS_B / MS_W}$

where

- $\overline{X}_j$ denotes a group mean;
- X_i denotes an individual observation (“data point”).

URLs for ANOVA test

- https://goodcalculators.com/one-way-anova-calculator/
- ANOVA Calculator | AAT Bioquest (https://www.aatbio.com/tools/anova-analysis-of-variance-one-two-way-calculator)

Example 1:

- Suppose we want to know whether or not three different exam prep programs lead to different mean scores on a certain exam. To test this, we recruit 30 students to participate in a study and split them into three groups. The students in each group are randomly assigned to use one of the three exam prep programs for the next three weeks to prepare for an exam. At the end of the three weeks, all of the students take the same exam. The exam scores for each group are shown below:

Group 1	Group 2	Group 3
85	91	79
86	92	78
88	93	88
75	85	94
78	87	92
94	84	85
98	82	83
79	88	85
71	95	82
80	96	81

Group	Degrees of Freedom (DF)	Sum of Squares (SS)	Mean Square (MS)	F-Statistic	P-Value
Between Groups	2	192.2	96.1	2.3575	0.1138
Within Groups	27	1100.6	40.763		
Total	29	1292.8			

- α (significance level) = 0.05
- DF1 (numerator degrees of freedom) = df treatment = 2
- DF2 (denominator degrees of freedom) = df error = 27
- We find that the F critical value is **3.3541**.
- F test statistic in the ANOVA table is less than the F critical value in the F distribution table, we fail to reject the null hypothesis. This means we don’t have sufficient evidence to say that there is a statistically significant difference between the mean exam scores of the three groups.

- To find P-value please refer

https://www.statology.org/here-is-how-to-find-the-p-value-from-the-f-distribution-table/

https://www.statology.org/f-distribution-calculator/

Example2:

Participant	Test Condition			
	A	B	C	D
1	11	11	21	16
2	18	11	22	15
3	17	10	18	13
4	19	15	21	20
5	13	17	23	10
6	10	15	15	20
7	14	14	15	13
8	13	14	19	18
9	19	18	16	12
10	10	17	21	18
11	10	19	22	13
12	16	14	18	20
13	10	20	17	19
14	10	13	21	18
15	20	17	14	18
16	18	17	17	14
Mean	14.25	15.13	18.75	16.06
SD	3.84	2.94	2.89	3.23

Key Parameters

Source	F-statistic	P-value
Between Groups	5.7587	0.0016

ANOVA Summary

Group	Degrees of Freedom (DF)	Sum of Squares (SS)	Mean Square (MS)	F-Statistic	P-Value
Between Groups	3	182.1719	60.724	5.7587	0.0016
Within Groups	60	632.6875	10.5448		
Total	63	814.8594			

HCI: Dialog Design

Learning Objective

- One key aspect of HCI is the dialog (communication), which is the interaction that takes place between a human user and the computer (machine)
- In this lecture, we shall learn about the representation, modelling and analysis of dialogs

Dialog

- A dialog refers to the *structure* of the interaction
- Dialog in HCI can be analyzed at three levels:
  - Lexical - At this level, the details such as the shape of icons, actual keys pressed etc. are dealt with
  - Syntactic - The order of inputs and outputs in an interaction are described at this level
  - Semantic - The effect a dialog has on the internal application/data is the subject matter at this level

Dialog Representation

- We need (formal) techniques to represent dialogs, which serves two purposes
  - It helps to understand the proposed design better
  - Formal representation makes it possible to analyze dialogs to identify usability problems (e.g., we can answer questions such as “does the design *actually* support *undo*”?)

Dialog Representation

- There are several formalisms that we can use to represent dialogs
- We shall discuss three of these formalisms in this lecture:
  - The state transition networks (STN)
  - The state charts (Finite State Machine (FSM))
  - The (classical) Petri-Nets

State Transition Network (STN)

- STNs are the most intuitive among all formalisms
- It assumes that a dialog essentially refers to a progression from one state of the system to the next in the system state space (in fact this assumption holds for all formalisms that we shall discuss)

State Transition Network (STN)

- The syntax of an STN is simple and consists of the following two entities
  - Circles: a circle in a STN refers to a state of the system, which is labeled (by giving a name to the state)
  - Arcs: the circles are connected with arcs, each of which refers to the action/event (represented by arc labels) that results in the system making a transition from the state where the arc originates to the state where it terminates

State Transition Network (STN)

Let’s illustrate the idea with an example. Suppose, we are using a drawing interface that allows us to draw lines and circles, by choosing appropriate menu item. To draw a circle, we need to select a center and circumference. A line can be drawn by selecting points on the line. How can we model this dialog using an STN?

State Transition Network (STN)

- So, what are states and transitions here?
  - We shall have a “start” state
  - From this “start” state, we shall go to a “menu” state, where we are shown the menu options. If we select the circle option, we go to a “circle” state. Otherwise, we select the “line” option and go to the “line” state

State Transition Network (STN)

- So, what are states and transitions here?
  - While at the “circle” state, we select a point as the circle center (through mouse click, say), which takes us to the “center” state
  - In the “center” state, we select the circle periphery (through mouse movement, say) and double click to indicate the end of input (the “finish” state). At this stage, the circle is displayed

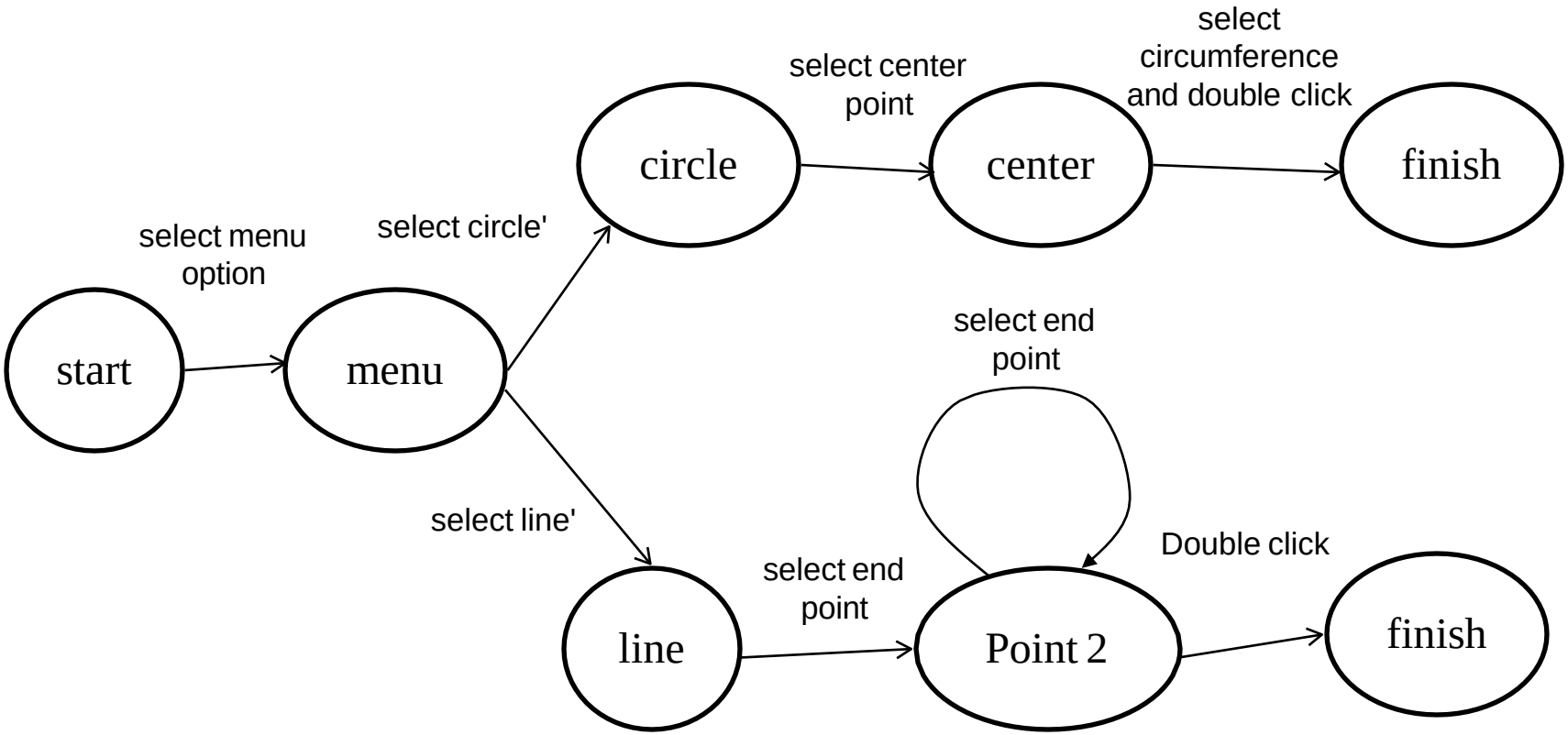
State Transition Network (STN)

- So, what are states and transitions here?
  - While at the “line” state, we select a point as the beginning of the line (through mouse click, say)
  - Then, we select another point to denote the last point on the line and transit to “point 2”. At this stage, a line is displayed between the two points

State Transition Network (STN)

- So, what are states and transitions here?
  - We can select another point, while at “point 2” to draw another line segment between this point and the point last selected. We can actually repeat this as many times as we want, to draw line of arbitrary shape and size
  - When we perform a double click, it indicates the end of input and the dialog comes to the “finish” stage

State Transition Network (STN)



STN – Pros and Cons

- Pros
  - Intuitive
  - Easy to understand
- Cons
  - Good for simple systems
  - Quickly becomes messy as the number of states/arcs grow

How to Model Complex Dialogs

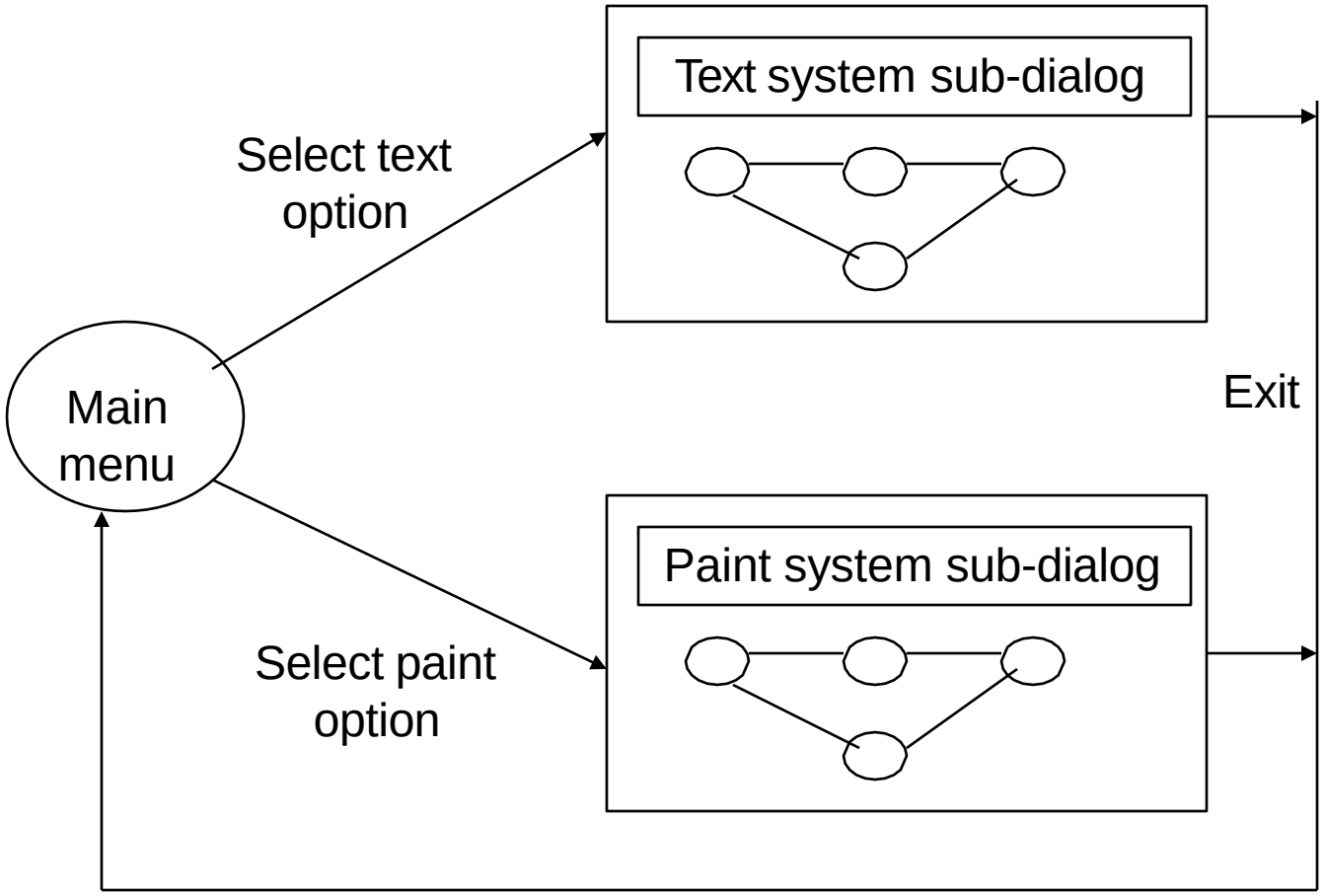
- Hierarchical STNs provide a way to manage complex dialogs
- Here, we divide the dialog into sub-dialogs
- Each sub-dialog is modeled with STNs
- Upper level STNs are designed to connect sub-dialogs

Hierarchical STN - Example

Suppose we want to model the dialog for a menu-based system. There are two menu items, one for a text system and the other for a paint-like system.

Each of these systems has its own dialog. For example, the paint system may have the dialog shown in the previous example. We can model the overall dialog as a hierarchical STN, as shown in the next slide

Hierarchical STN - Example

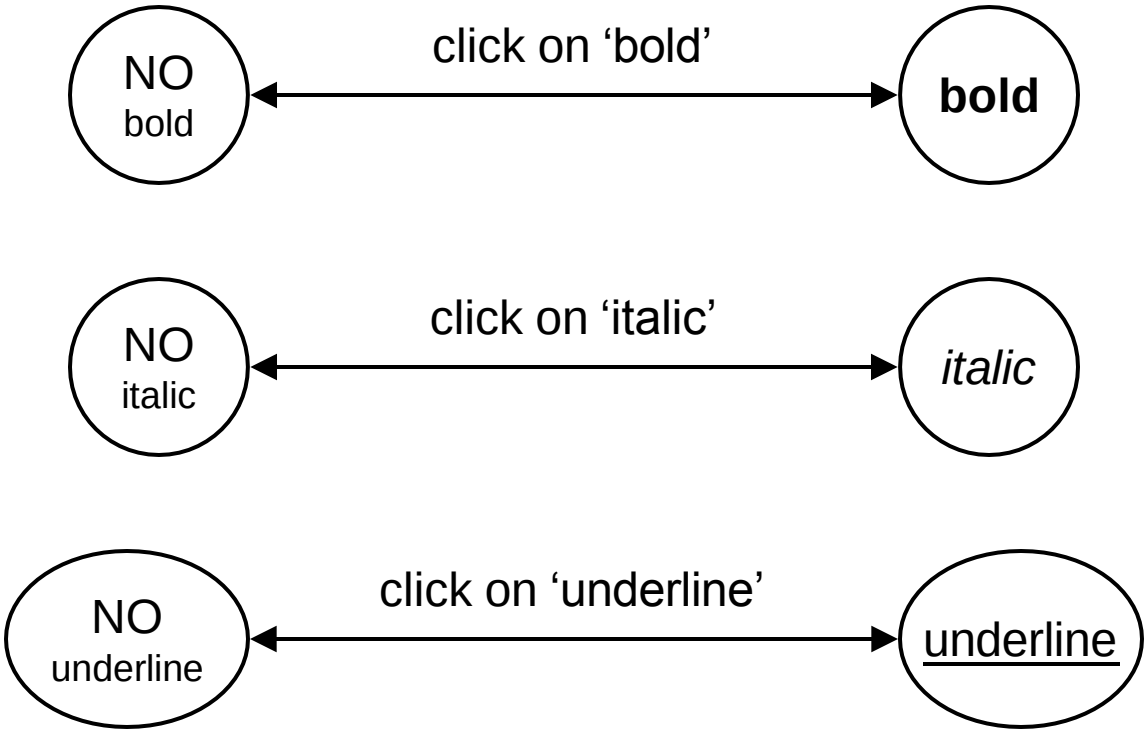


How to Model Complex Dialogs

- However, even hierarchical STNs are inadequate to model many “common” dialogs

  For example, consider a text editor that supports three operations: underline, **bold** and *italic*. Let us try to model this dialog with an STN, assuming first that we can perform (only) one operation on a piece of text

Modelling Complex Dialogs

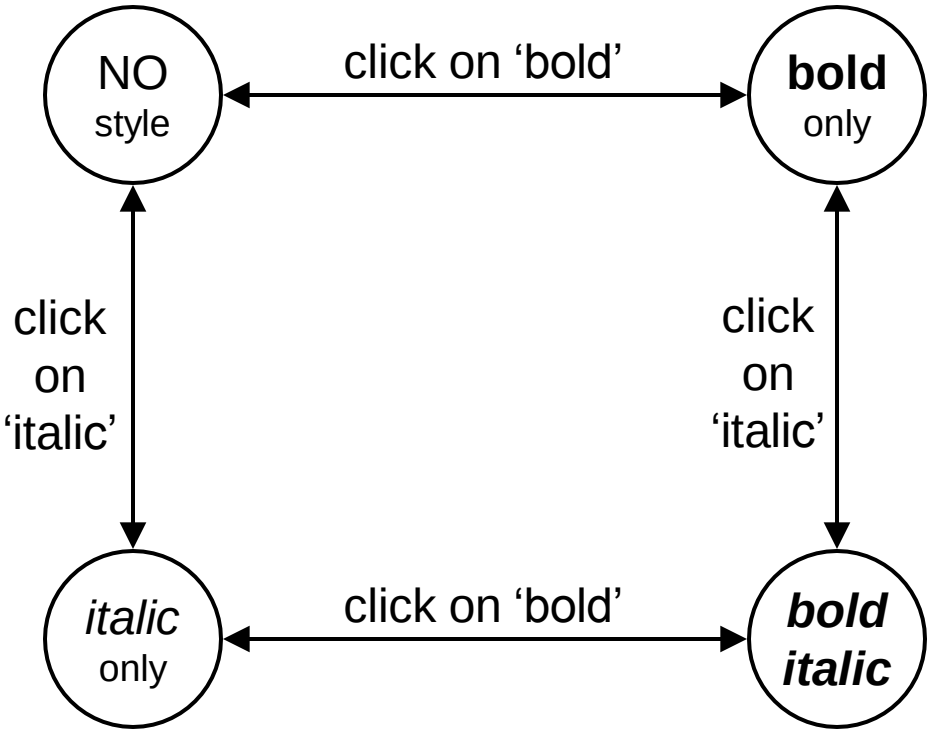


How to Model Complex Dialogs

Now suppose we relax the condition “we can perform (only) one operation on a piece of text”. Now we can perform two operations together on the same piece of text (e.g., **bold** followed by *italic*).
How the STN for this new system looks?

(let us construct the STN for only the dialog involving **bold** and *italic*. STN for other pairs will be similar)

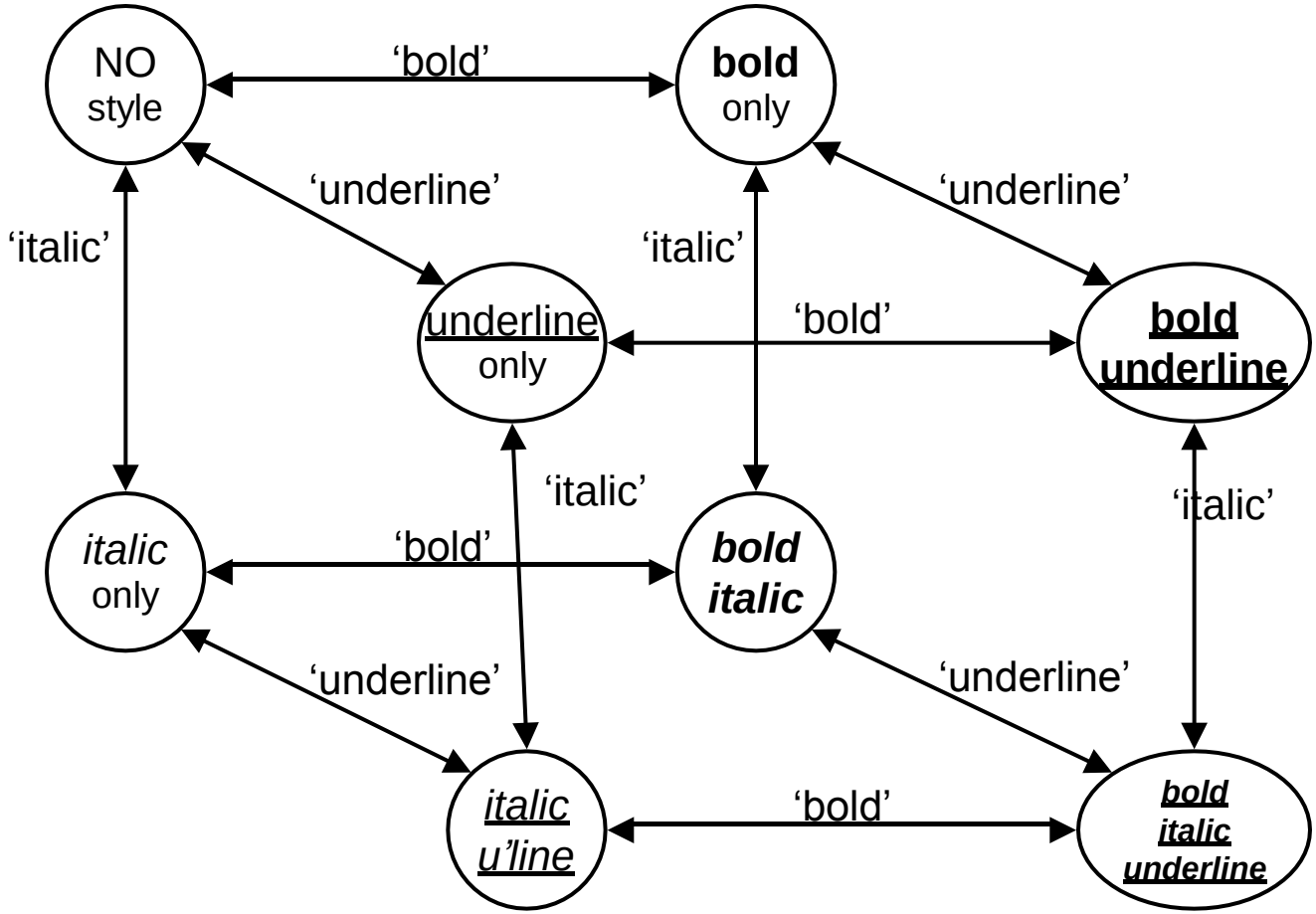
Modelling Complex Dialogs



How to Model Complex Dialogs

Now suppose we relax the condition further. Now we can perform all the three operations together on the same piece of text. This is a fairly common scenario and supported by all text editors. Let us see how the STN for this new system looks.

Modelling Complex Dialogs



How to Model Complex Dialogs

- As we can see, the STN has become very complex with too many states and arcs
- This is because we are trying to model activities that occur on the same object or at the same time. Such behaviors are known as “concurrent behaviors”
- STNs are not very good at modeling concurrent behaviors, which are fairly common in dialogs that we encounter in HCI

Summary

- To better model “concurrent” dialogs, other formalisms are used
- We shall discuss two of those formalisms, namely the State-Charts and the Petri Nets, in the following lectures

HCI: Dialog Design (State-Charts)

Learning Objective

- In the previous lecture, we introduced the need for dialog design
- We also learned about the advantages about formal modeling of dialogs
- We discussed how to use STNs for the purpose

Learning Objective

- As we mentioned, STNs are good for modeling simple systems; for complex systems as well as systems having concurrency, STNs fail
- In this lecture, we shall learn about the State-Chart formalism that can overcome the problems with STNs

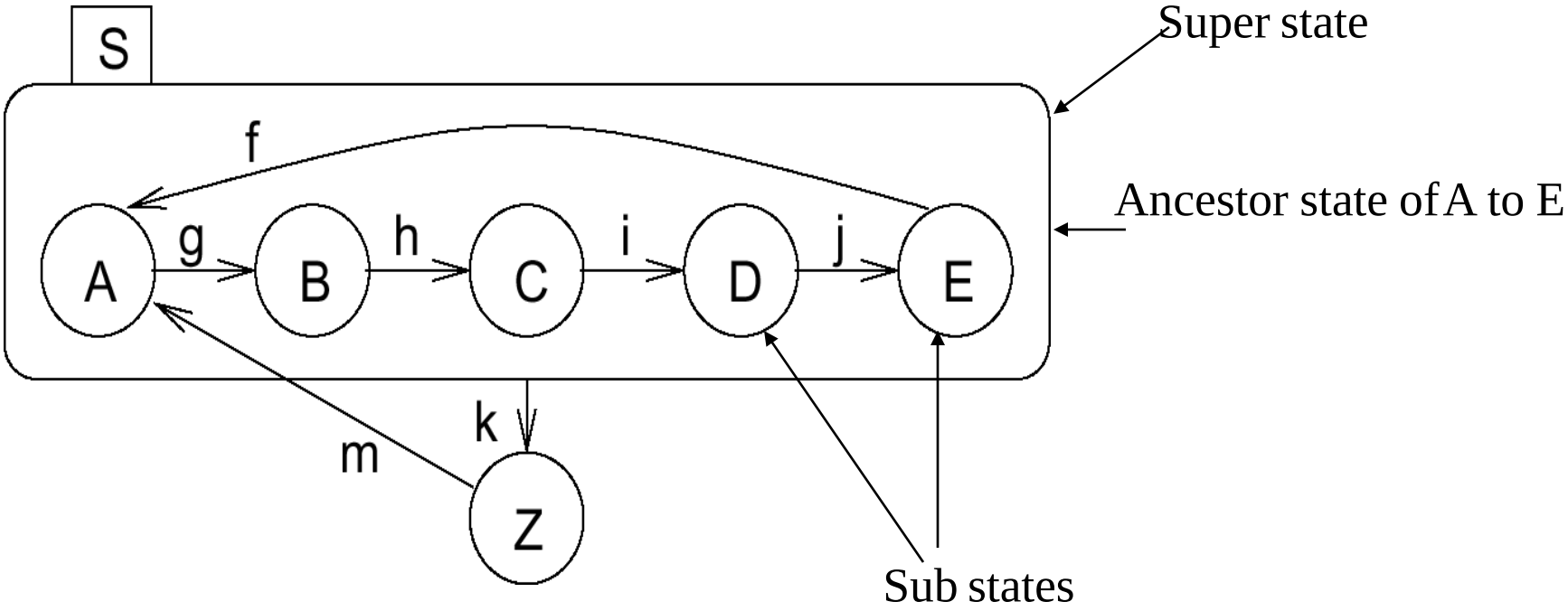
State-Charts

- Proposed by David Harel (1987) to represent complex *reactive* systems
- Extends finite state machines (FSM)
  - Better handle concurrency
  - Adds memory, conditional statements to FSM
- Simplifies complex system representation (states and arcs) to a great extent

Definitions

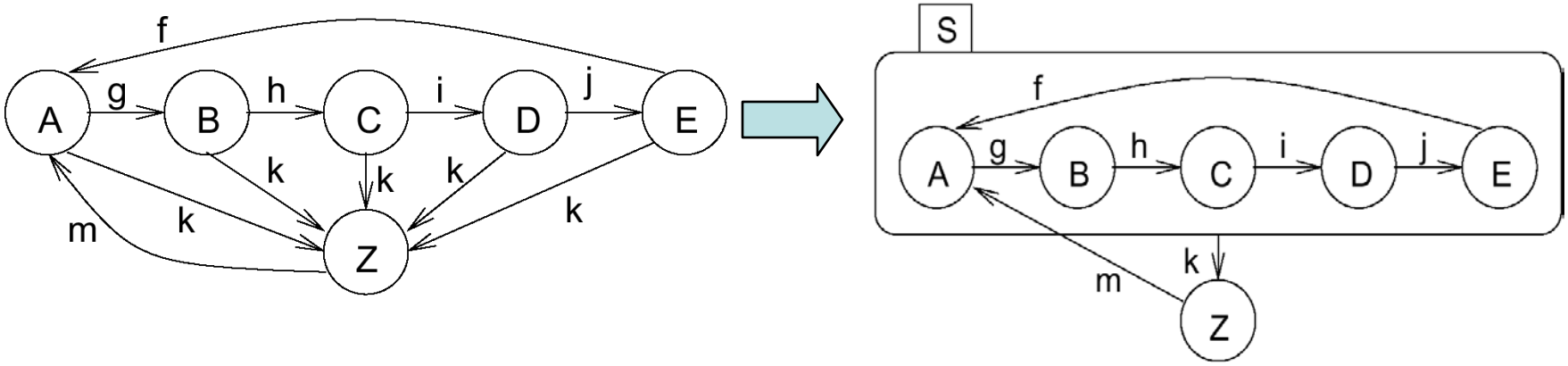
- **Active state**: the current state of the underlying FSM
- **Basic states**: states that are not composed of other states
- **Super states**: states that are composed of other states
  - For each basic state *b*, the super state containing *b* is called the *ancestor* state
  - A super state is called OR super state if exactly one of its sub states is active, whenever it is active

Definitions



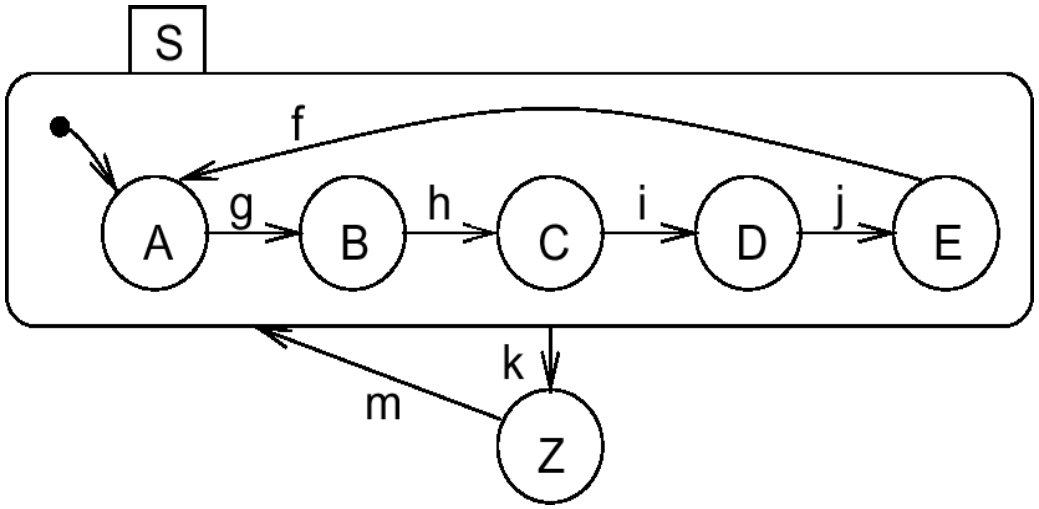
Super State Advantage

- It allows us to represent complex FSM in a nice way, by clustering states



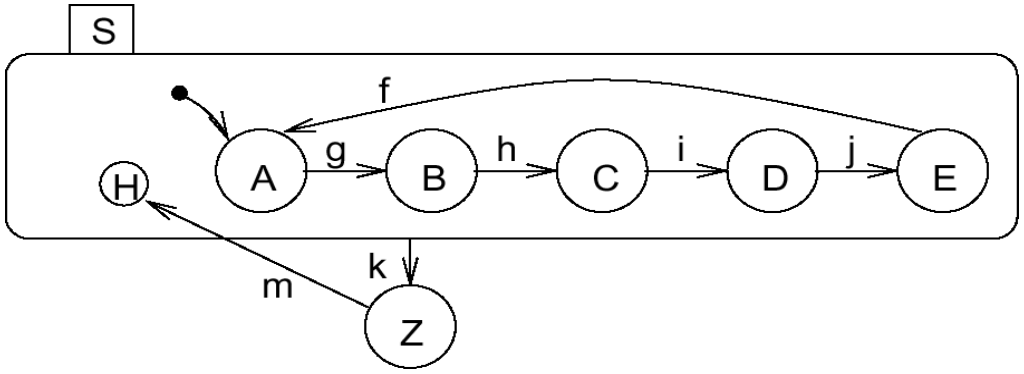
Default State Mechanism

- Indicates the sub state entered whenever super state is entered – represented using a filled circle
  - Not a state by itself

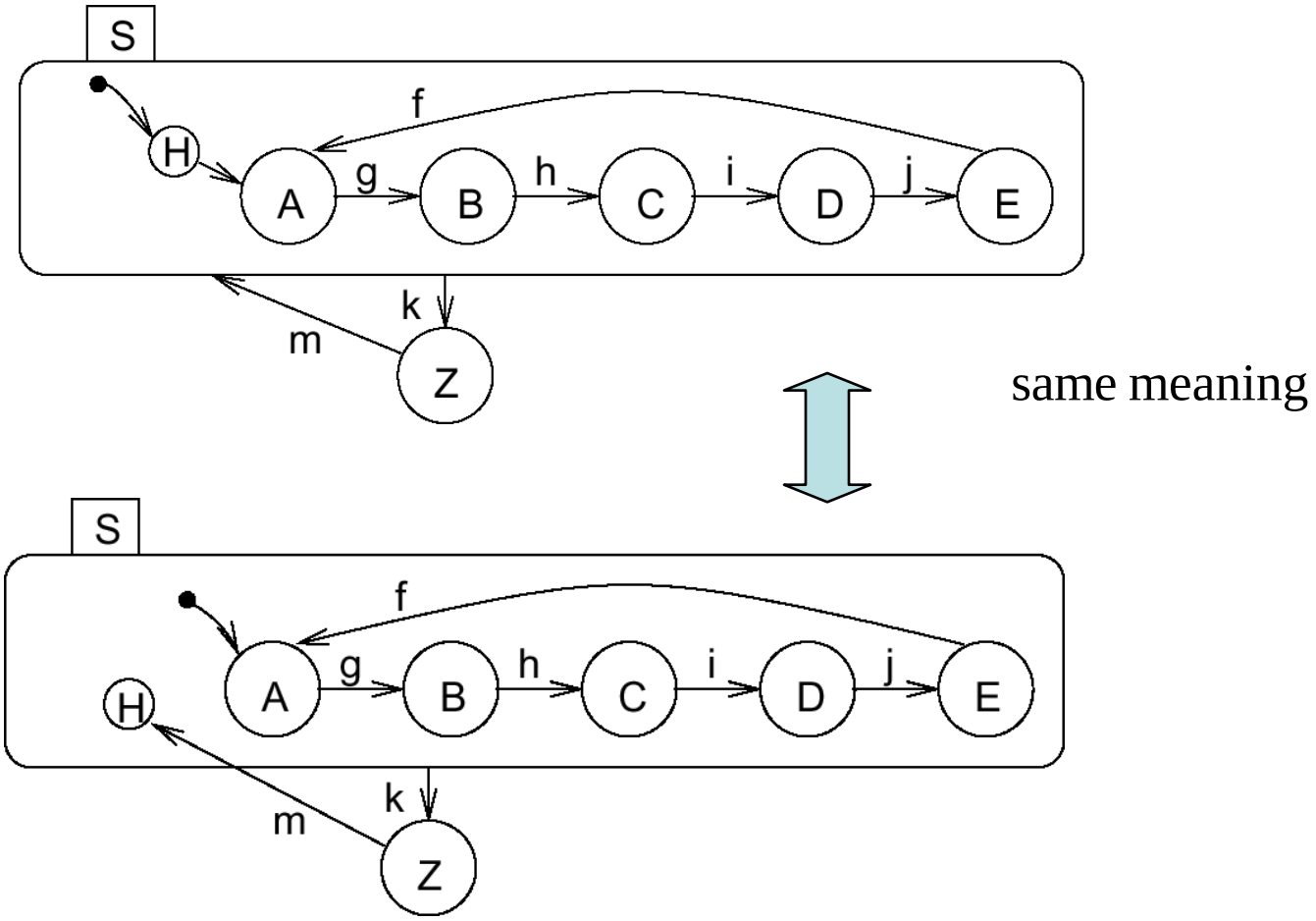


History Mechanism

- For input m, S enters the state it was in before S was left
  - If S is entered for the very first time, the default mechanism applies
  - History and default mechanisms can be used hierarchically

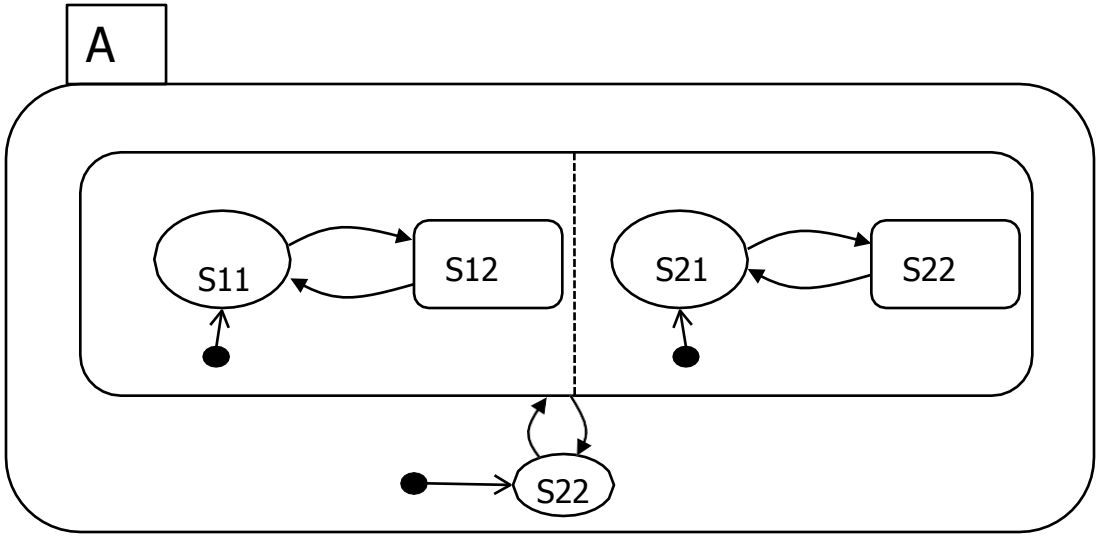


Combining History and Default State Mechanism



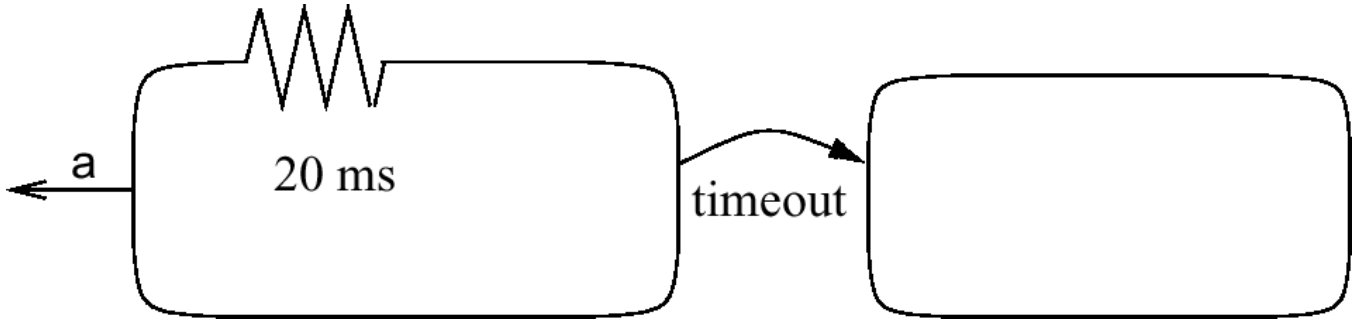
Concurrency

- State-Charts supports concurrency using the notion of the AND super states
  - In AND super states, the FSM is active in all (immediate) sub states simultaneously



Timing Constraints

- State-Chart supports delay/timeout modeling
  - using special edges
    - Do we need it??



If event **a** does not happen while the system is in the left state for 20 ms, a timeout will take place.

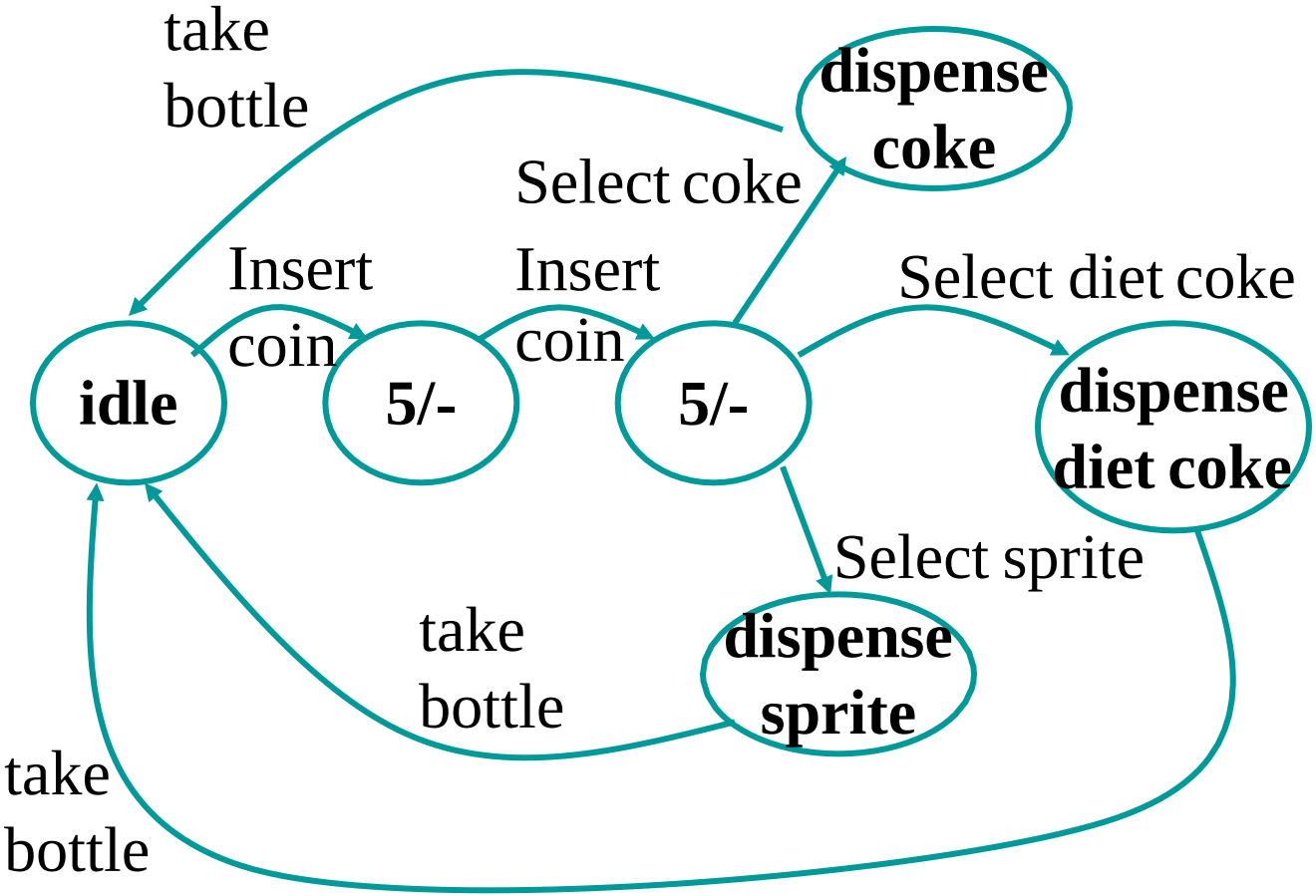
Example: Vending Machine v1.0

- Suppose we have a juice/beverages vending machine:
  - When turned on, this vending machine waits for money
  - When Rs. 5/- coin is deposited, the machine waits for another Rs. 5/- coin
  - When the second coin is deposited, the machine waits for a selection
  - When the user presses “COKE,” a coke is dispensed

Example: Vending Machine v1.0

- Suppose we have a juice/beverages vending machine:
  - When the user takes the bottle, the machine waits again
  - When the user presses either “SPRITE” or “DIET COKE,” a Sprite or a Diet Coke is dispensed
  - When the user takes the bottle, the machine waits again
- Let us represent this behavior using FSM

Vending Machine v1.0



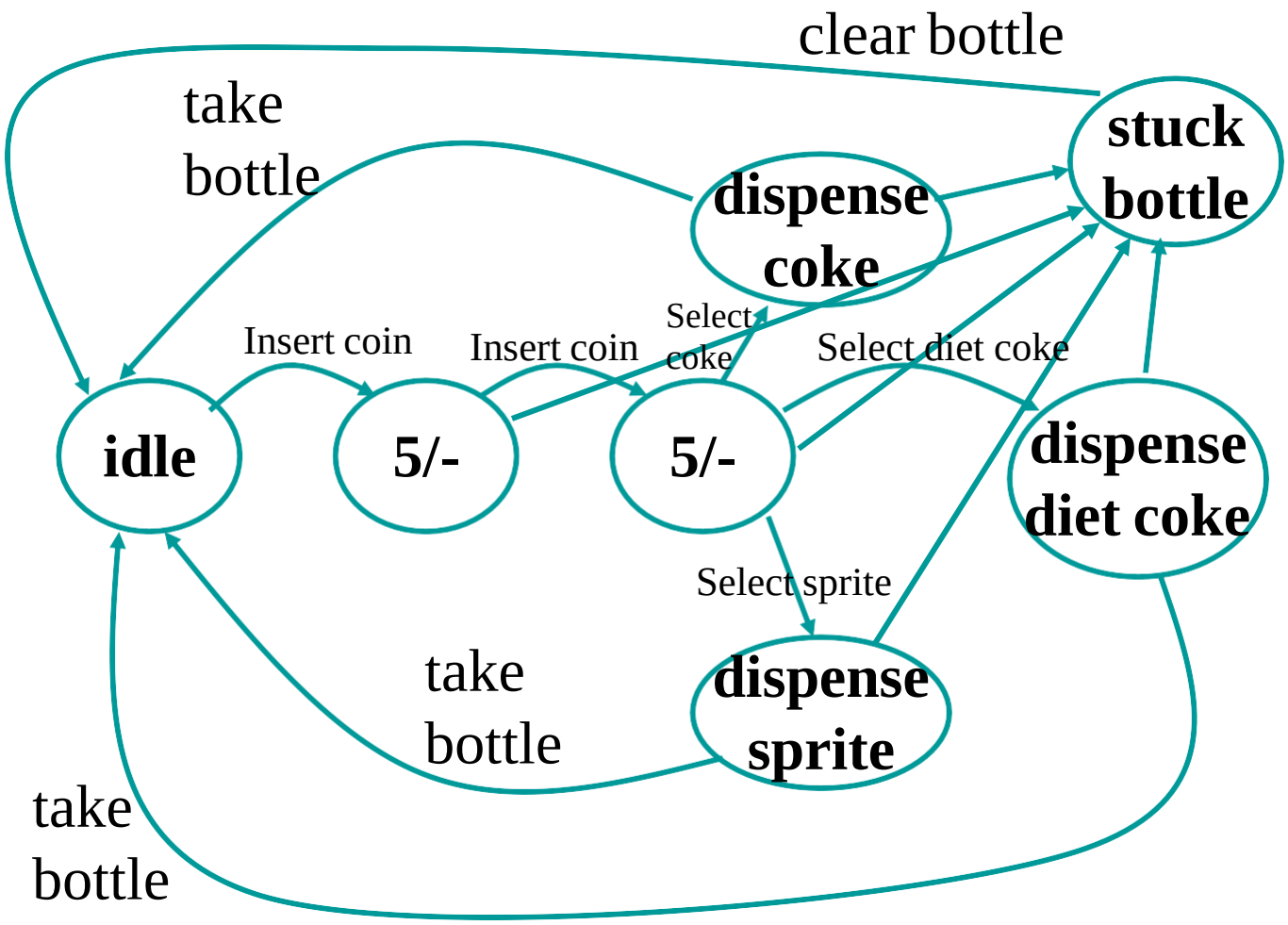
Vending Machine v2.0

- Let us include some more features in the vending machine
- The Bottles can get stuck in the machine
  - An automatic indicator will notify the system when a bottle is stuck
  - When this occurs, the machine will not accept any money or issue any bottles until the bottle is cleared
  - When the bottle is cleared, the machine will wait for money again

Vending Machine v2.0

- State machine changes
  - How many new states are required?
  - How many new transitions?

Vending Machine v2.0



Vending Machine v3.0

- Let us add some more features in the vending machine
- Bottles sometimes shake loose
  - An additional, automatic indicator will indicate that the bottle is cleared
  - When the bottles are cleared, the machine will return to the same state it was in before the bottle got stuck

Vending Machine v3.0

- State machine changes
  - How many new states are required?
  - How many new transitions?

Vending Machine v4.0

- We can add even more features
- Automatic bottle filler
  - If a button is pressed, the machine will toggle between bottle filling and dispensing modes
  - When in bottle filling mode
    - Bottles may be inserted if the Coke/Diet Coke/Sprite machine is ready
    - When a bottle is inserted, the machine will NOT be ready to accept another bottle and will check the bottle

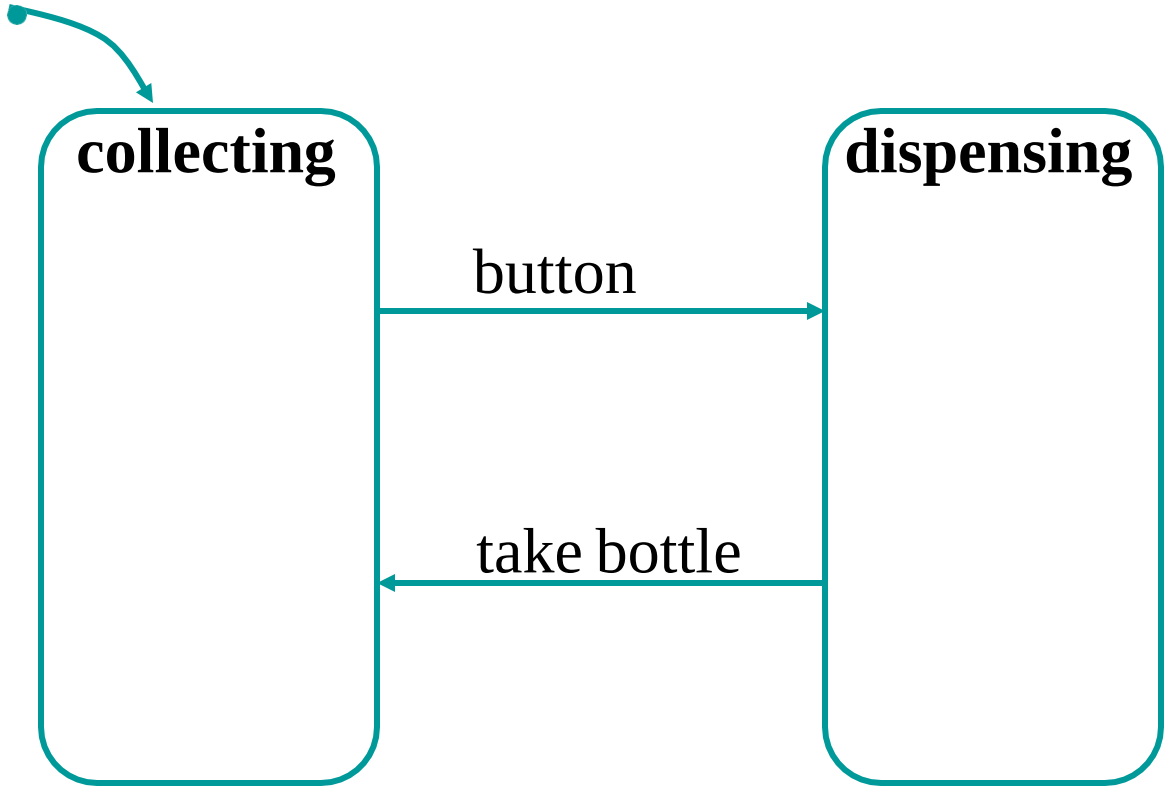
Vending Machine v4.0

- We can add even more features
- Automatic bottle filler
  - If a button is pressed, the machine will toggle between bottle filling and dispensing modes
  - When in bottle filling mode
    - If the bottle check finds a Coke was inserted, it will signal Coke_OK and return to ready
    - If the bottle check finds a Diet Coke was inserted, the coke machine will signal Diet_OK and return to ready
    - Otherwise, the bottle will be immediately dispensed

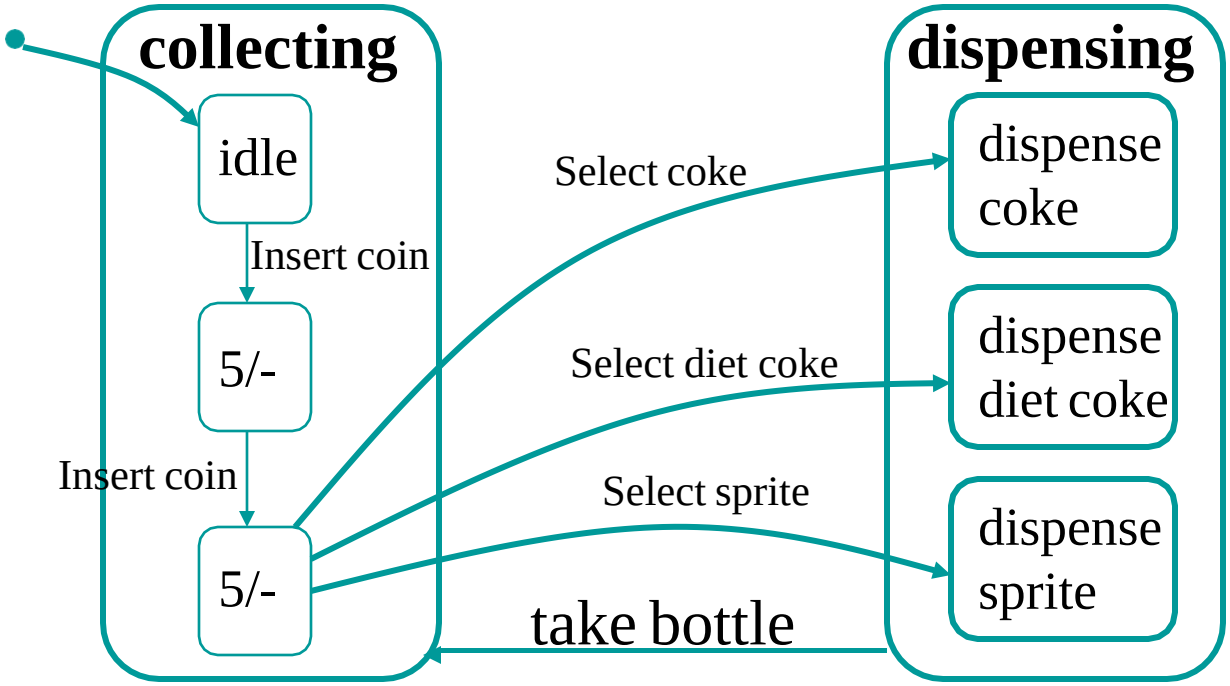
Vending Machine v4.0

- State machine changes
  - How many new states are required?
  - How many new transitions?

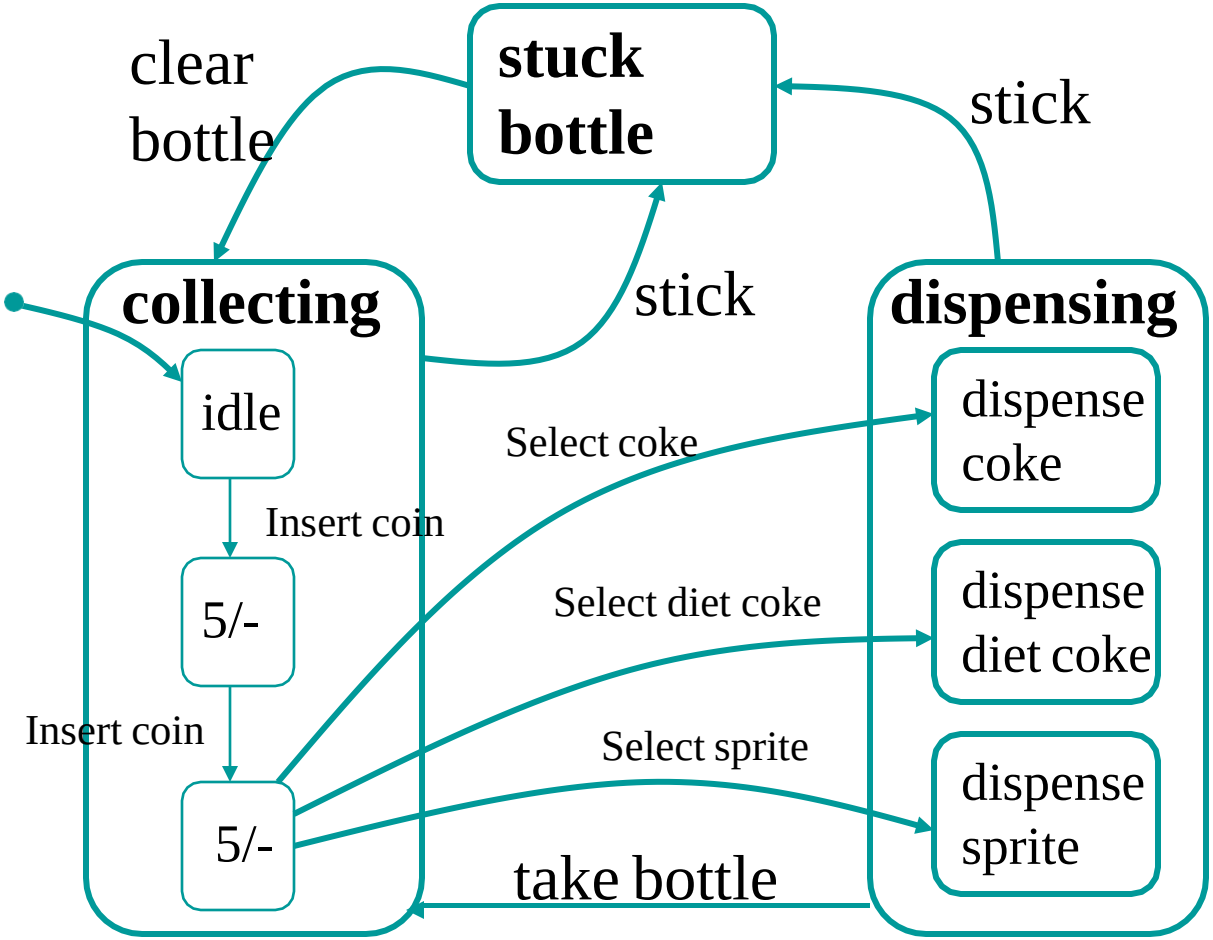
State-Chart Construction: Bottle Dispenser



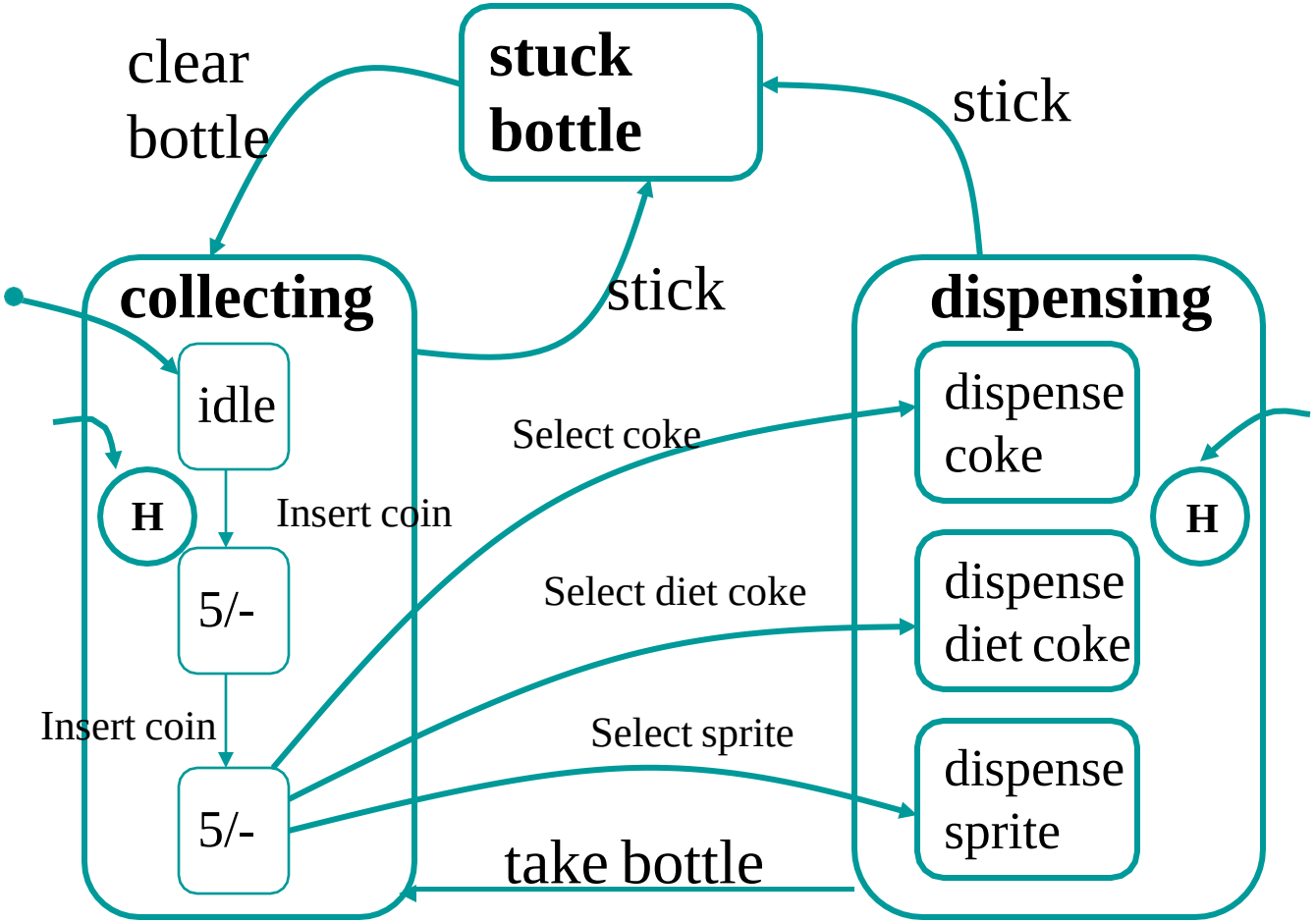
State-Chart Construction: Bottle Dispenser



State-Chart Construction: Bottle Dispenser



State-Chart Construction: Adding History



State-Chart Pros

- Large number of commercial simulation tools available (StateMate, StateFlow, BetterState, ...)
- Available “back-ends” translate State-Charts into C or VerilogHDL or VHDL, thus enabling software or hardware implementations

HCI: Dialog Design Using Petri Nets

Learning Objective

- In the previous lecture, we discussed about State-Charts, a formalism suitable for dialog design, which is potentially more expressive than State Transition Networks (STNs)
- In this lecture, we shall discuss a powerful formalism for dialog design, namely the (classical) Petri Nets

(Classical) Petri Net (PN)

- The formalism was first proposed by Carl Adam Petri (1962, PhD thesis)
- It is a simple model of dynamic behavior
  - Just four elements are used to represent behavior: **places**, **transitions, arcs and tokens**
  - Graphical and mathematical description for easy understanding
  - Formal semantics allow for analysis of the behavior

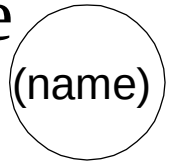
Elements of PN

- Place: used to represent passive elements of the reactive system
- Transition: used to represent active elements of the reactive system
- Arc: used to represent causal relations
- Token: elements subject to change

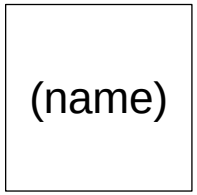
The state (space) of a process/system is modeled by places and tokens and state transitions are modeled by transitions

Elements of PN: Notation

- A place is represented by a circle
- Transitions are represented by squares/rectangles
- Arcs are represented by arrows
- Tokens are represented by small filled circles



place



transition

arc (directed connection)

token

Role of a Token

- Tokens can play the following roles
  - A **physical object**, for example a product, a part, a drug, a person
  - An **information object**, for example a message, a signal, a report
  - A **collection of objects**, for example a truck with products, a warehouse with parts, or an address file
  - An **indicator of a state**, for example the indicator of the state in which a process is, or the state of an object
  - An **indicator of a condition**: the presence of a token indicates whether a certain condition is fulfilled

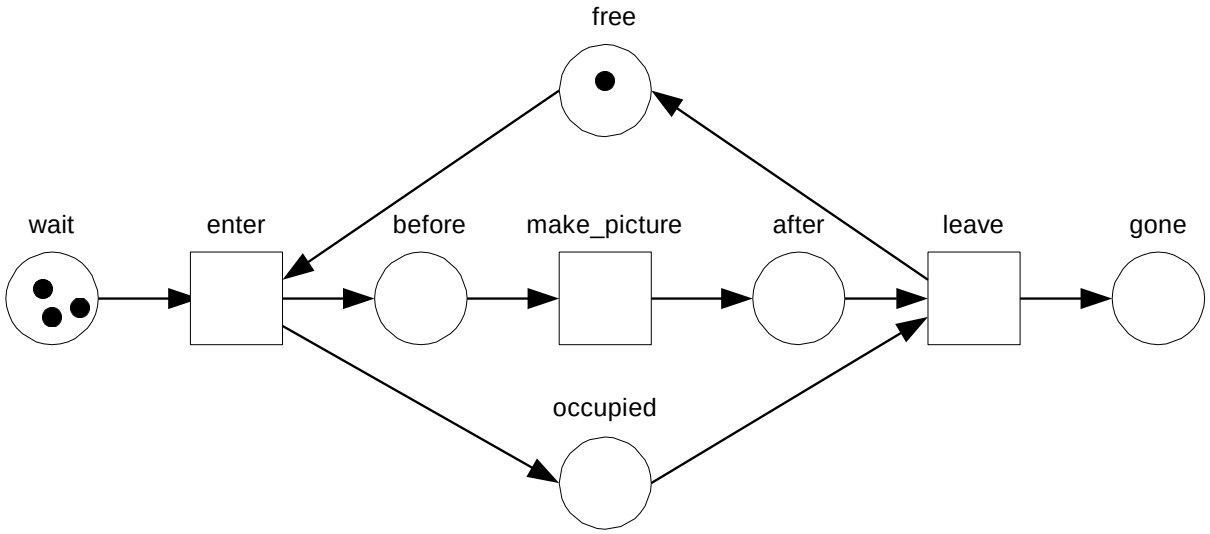
Role of a Place

- A place in a PN can represent the following
  - A type of **communication medium**, like a telephone line, a middleman, or a communication network
  - A **buffer**: for example, a depot, a queue or a post bin
  - A **geographical location**, like a place in a warehouse, office or hospital
  - A possible **state or state condition**: for example, the floor where an elevator is, or the condition that a specialist is available

Role of a Transition

- A transition can be used to represent things such as
  - An **event** (e.g., starting an operation, the switching of a traffic light from red to green)
  - A **transformation of an object**, like adapting a product, updating a database, or updating a document
  - A **transport of an object**: for example, transporting goods, or sending a file

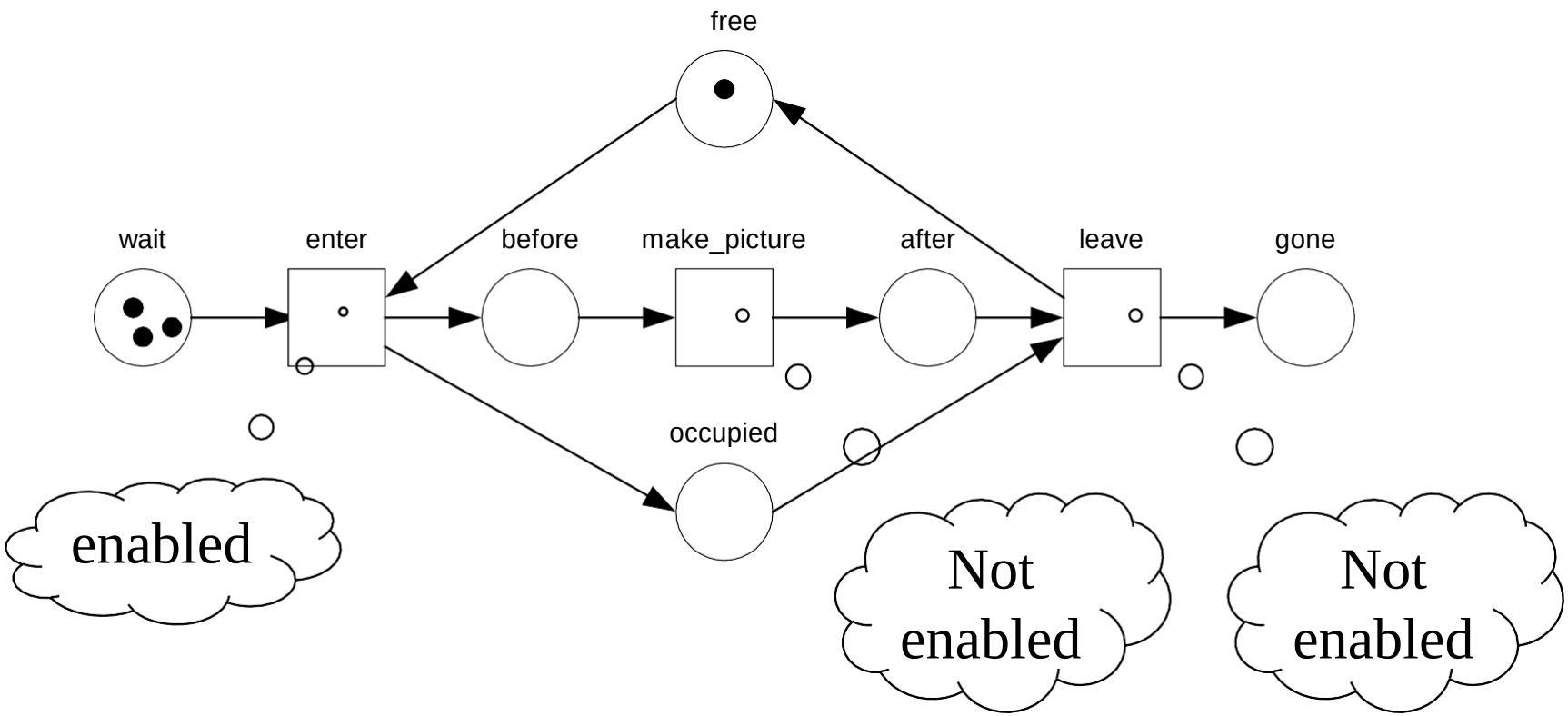
PN Construction Rules



- Connections are directed
- No connections between two places or two transitions is allowed
- Places may hold zero or more tokens

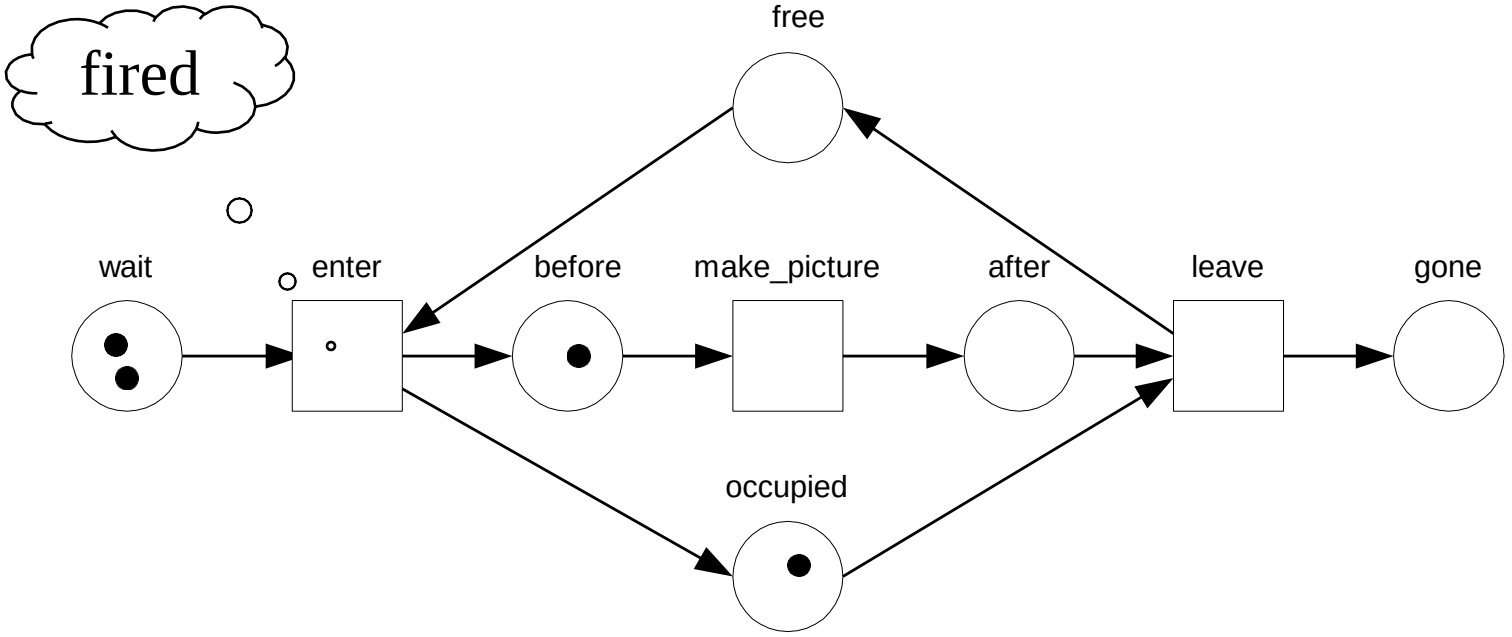
Enabled

- A transition is **enabled** if each of its input places contains at least one token



Firing

- An **enabled** transition can **fire** (i.e., it occurs)
- When it **fires** it **consumes** a token from each input place and **produces** a token for each output place

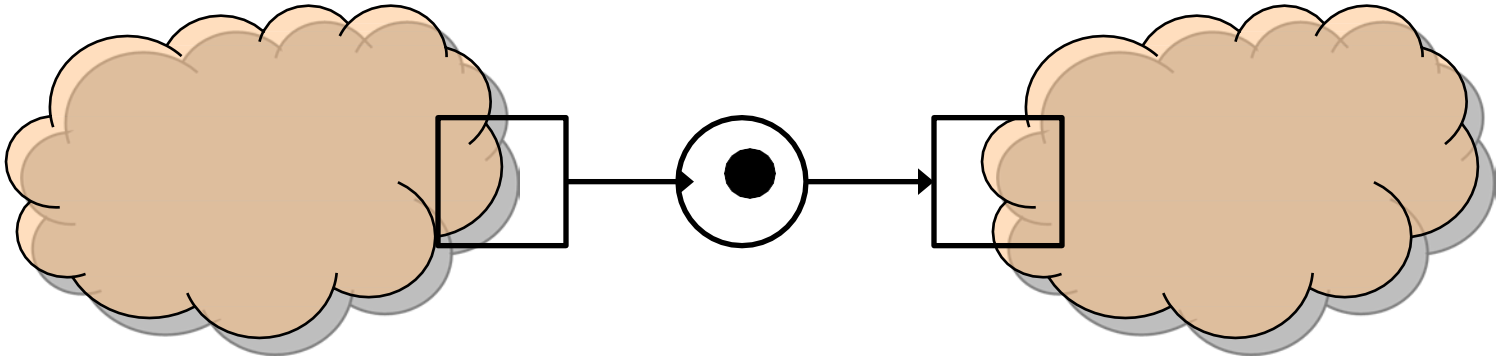


Remarks

- Firing is **atomic** (i.e., it always completes after start)
- Non-determinism: multiple transitions may be enabled, but only one fires at a time
- The **state** of the reactive system is represented by the distribution of tokens over places (also referred to as **marking**)

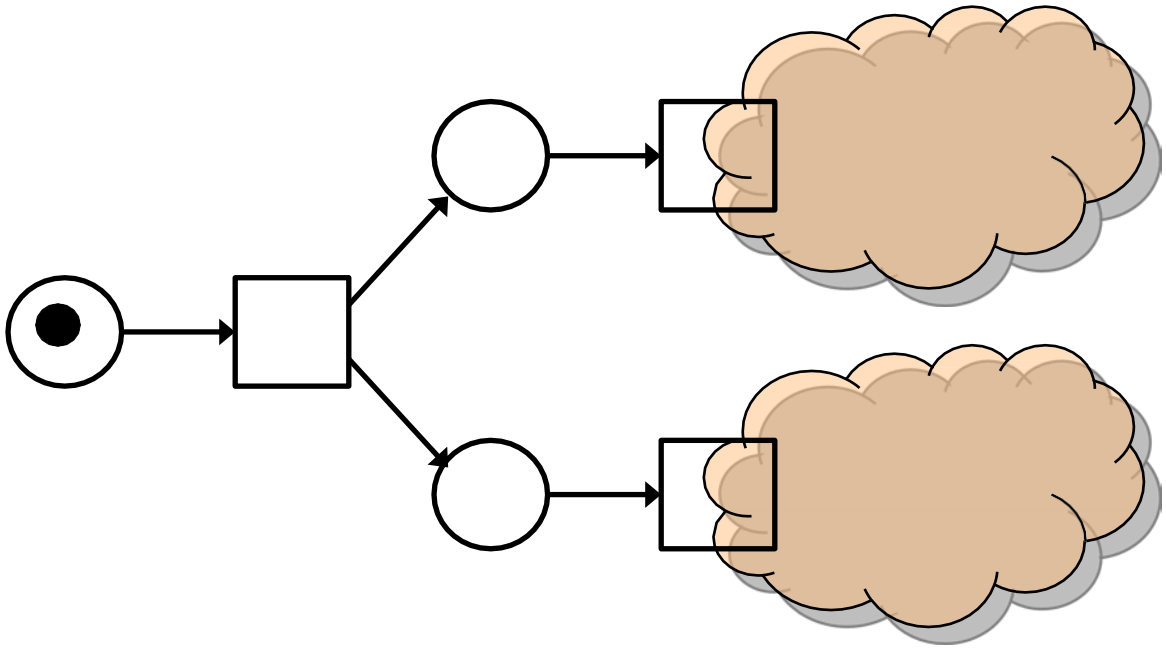
Typical PN Structures

- **Causality**, i.e., one part of the PN is caused by the other part



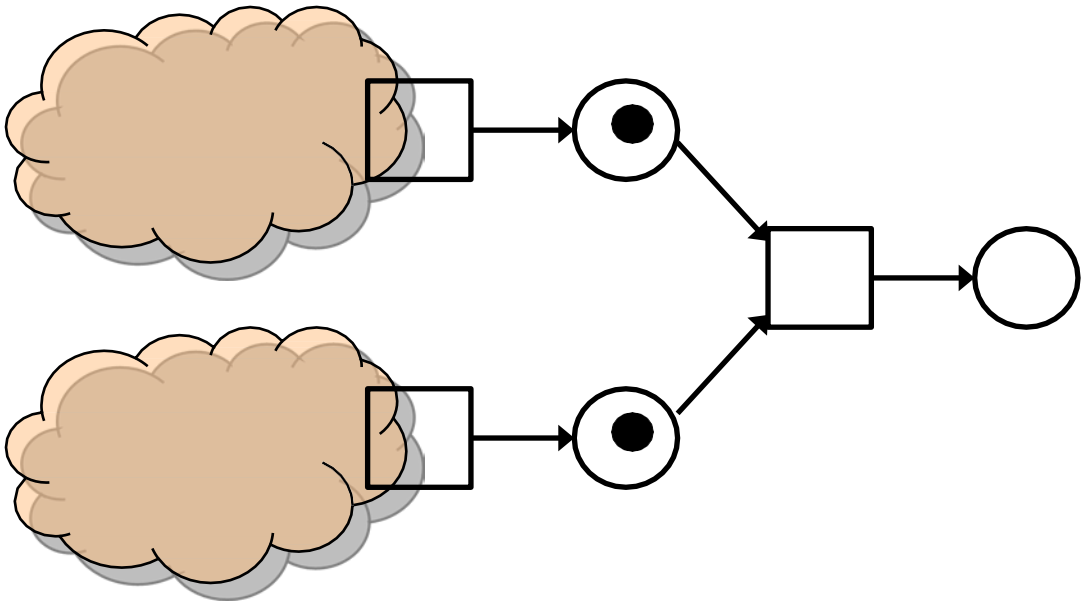
Typical PN Structures

- **Parallelism (AND-split)**, i.e., two parts of the PN can be activated at the same time



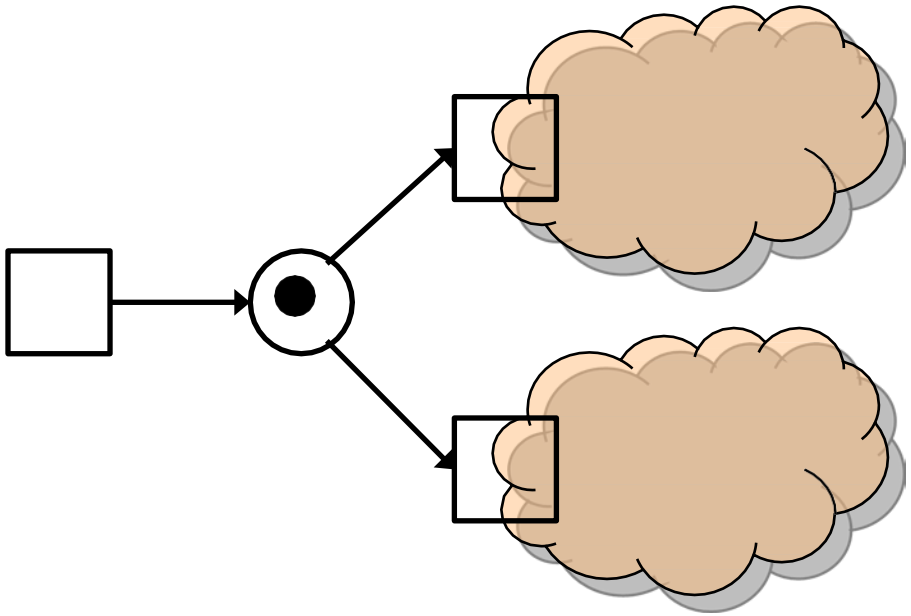
Typical PN Structures

- **Parallelism (AND-join)**, i.e., two parts of the PN must be active at the same time, or enable further firings



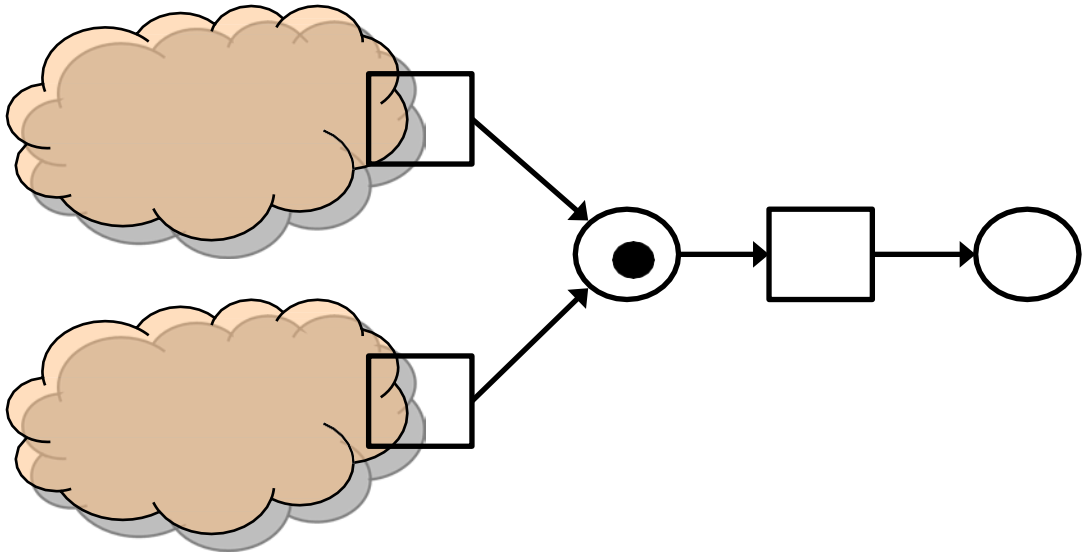
Typical PN Structures

- **Choice (XOR-split)**, i.e., either of the two sub nets of a PN can be activated



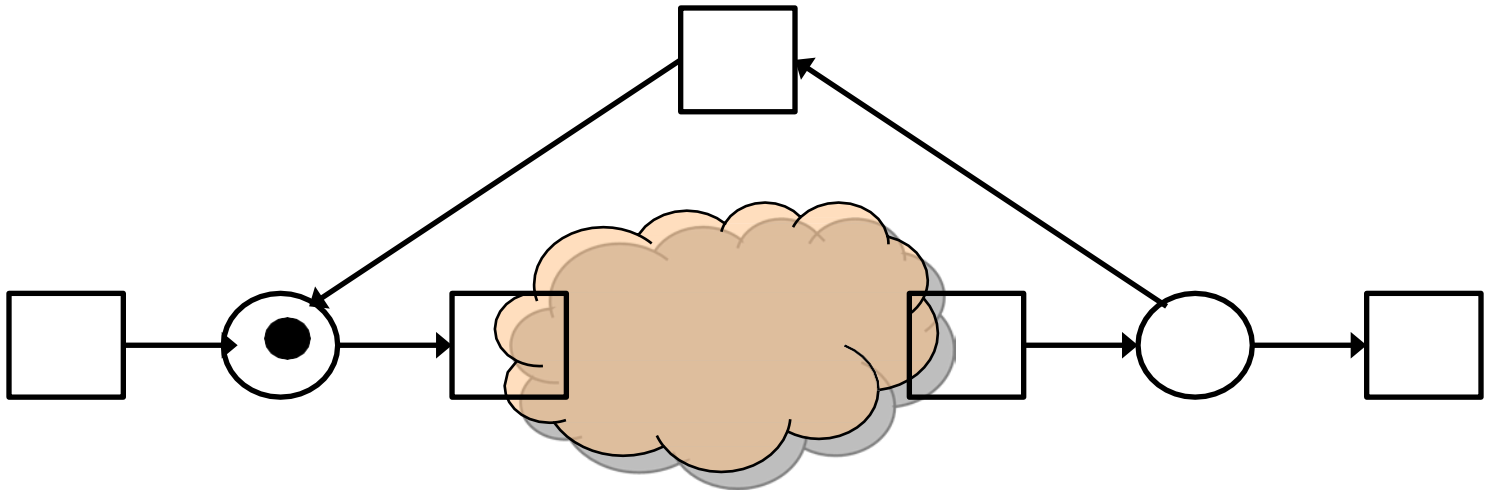
Typical PN Structures

- **Choice (XOR-join)**, i.e., either of the two sub nets of a PN is an enabler



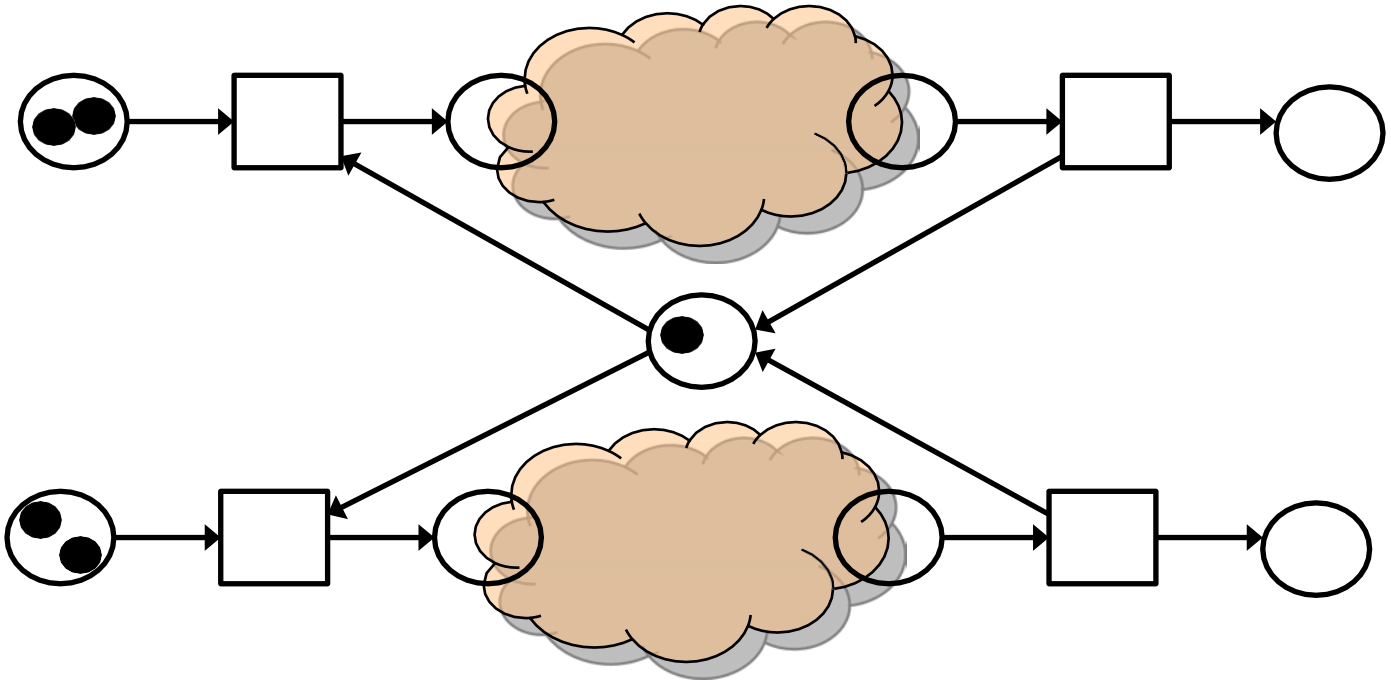
Typical PN Structures

- **Iteration (1 or more times)**, i.e., the firing iterates at least once



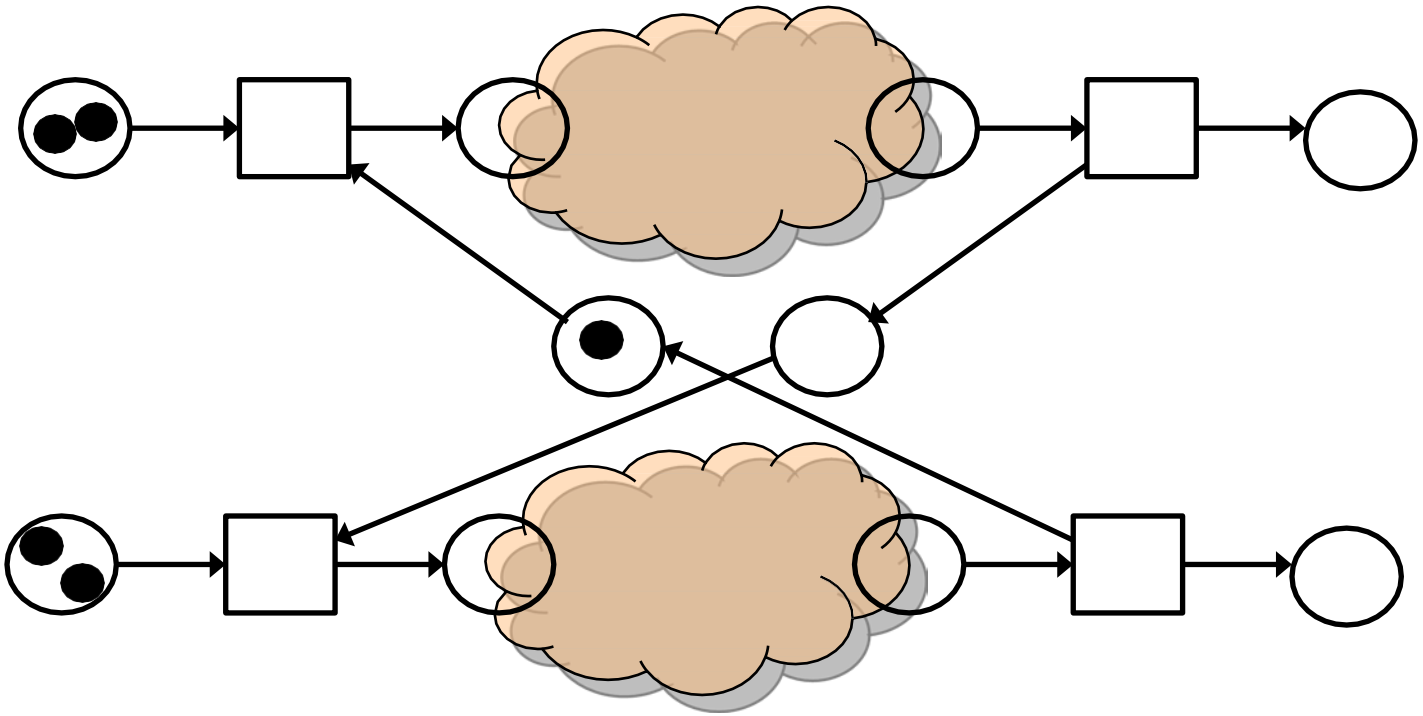
Typical PN Structures

- **Mutual exclusion**, i.e., only one of the sub nets should be active at a time



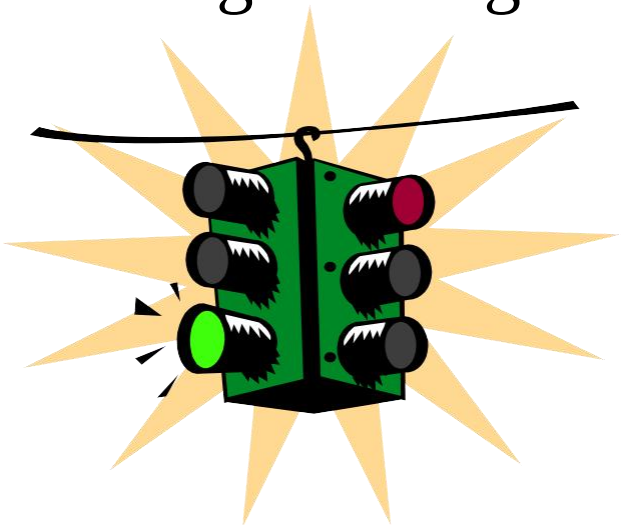
Typical PN Structures

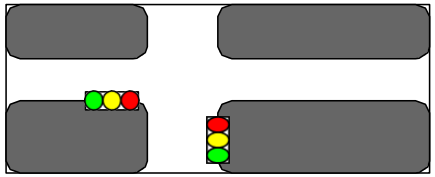
- **Alternating**, i.e., the sub nets of a PN should be alternatively activated



Example: Two Traffic Lights

- Let us illustrate the idea with an example. Suppose there are two traffic lights at a road junction. How we can model the behavior of these two lights using PN?

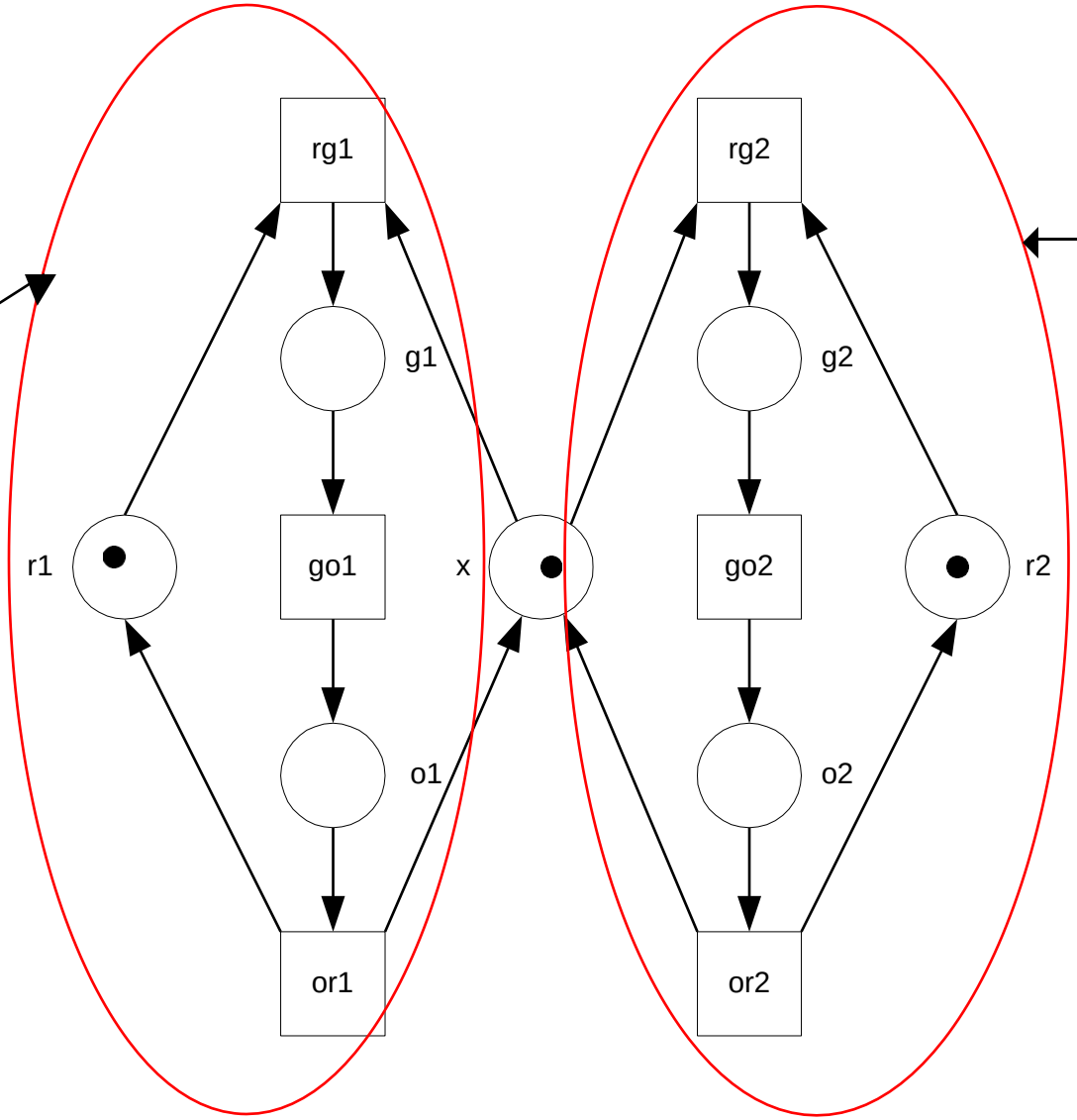




Example: Two Traffic Lights

- The characteristics of the combined system (of two lights)
  - They are mutually exclusive
  - They should alternate
- We can use the typical structures to model the behavior

Example: Two Traffic Lights



One light

(r=red,
g=green,
o=orange)

The other light

(r=red,
g=green,
o=orange)

Summary

- Look closely in the example how the elements of a PN are used to model the behavior of the system
- In the next lecture, we shall discuss with an example the usefulness of formal dialog representation
- Also we shall discuss about the properties we check with the formalisms and how?

HCI: Dialog Design – Use of Formalism

Learning Objective

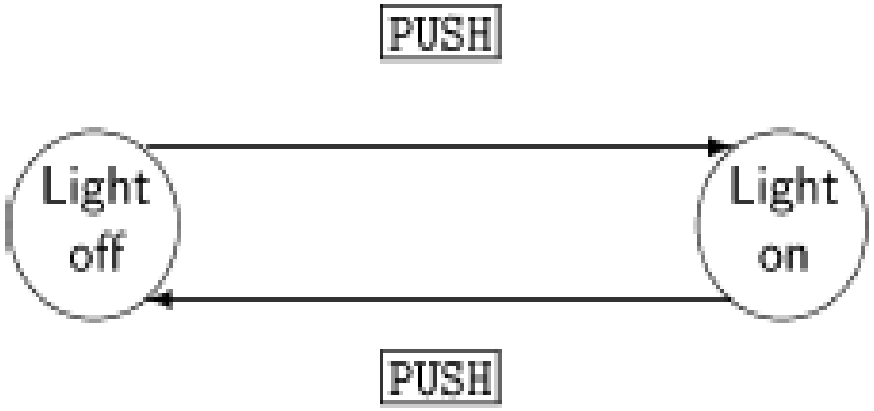
- In the previous lectures, we have learned different formalisms such as State Transition Network (STN), State-Charts (Finite State Machines (FSM)), and the Petri Nets (PN) to represent the dialogs
- Some are simple (STN) but not powerful enough to capture the typical interactive behavior (complexity and concurrency), others are more suitable (FSM, PN)

Learning Objective

- It’s clear from the discussions that representing interactive systems formally is no easy task and requires expertise
- This brings us to the question, why should we spend time and effort mastering formal representation techniques?
- In this lecture, we shall discuss how to answer this question

Example Use of Formalism

- Let us try to understand the use of formalism in dialog design with a simple example. Consider a simple two state system – a light bulb controlled by a push button switch, which alternately turns the light on or off.
- The corresponding STN is



a. Light with push-on/push-off action.

Example Use of Formalism

- The example system belongs to the general class of push button devices
- They belong to an important class of interactive system (Desktop GUI, touch-screen devices, WWW)
  - Ubiquitous (mobile phones, vending machines, aircraft flight deck, medical unit, cars, nuclear power stations)
- Thus, modelling such devices can actually help to model a large number of interactive systems
- Hence, Interaction with such devices can be modelled as the Matrix Algebra (MA)

Push Button Device: FSM to MA

- We shall use the following notations
  - N = number of states
  - States are numbered from 0 to N
  - A transition is represented by a matrix (N×N)
  - A state is represented by an unit vector of size N, with N-1 0s and one 1 at the position corresponding to the state number. e.g., states ON = (1 0) and OFF = (0 1)

Push Button Device: FSM to MA

- We shall use the following notations
  - New state = old state (vector) × transition(s) [a matrix multiplication]

    e.g., one push button action push

$$\boxed{\texttt{PUSH}} = \begin{pmatrix} 0 & 1 \\ 1 & 0 \end{pmatrix}$$

- We can check several properties of our simple push button device through matrix multiplication

Property Checking: Example

- When light is off, pushing the button puts the light on

$$\begin{aligned} \mathbf{off}\ \boxed{\texttt{PUSH}} &= (0\ 1)\begin{pmatrix} 0 & 1 \\ 1 & 0 \end{pmatrix} \\ &= (1\ 0) \\ &= \mathbf{on} \end{aligned}$$

- Similarly, we can show pushing the button when the light is on puts it off

Push Button Device: FSM to MA

- The case for *undo*

$$\boxed{\text{PUSH}}\,\boxed{\text{PUSH}} = \begin{pmatrix} 0 & 1 \\ 1 & 0 \end{pmatrix}\begin{pmatrix} 0 & 1 \\ 1 & 0 \end{pmatrix}$$

$$= \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}$$

$$= I$$

- Pressing the button twice return the system to the original state (s $\boxed{\text{PUSH}}\,\boxed{\text{PUSH}}$ = s)
  - What about pushing the button thrice, four times,…? Do the calculations and check for yourself

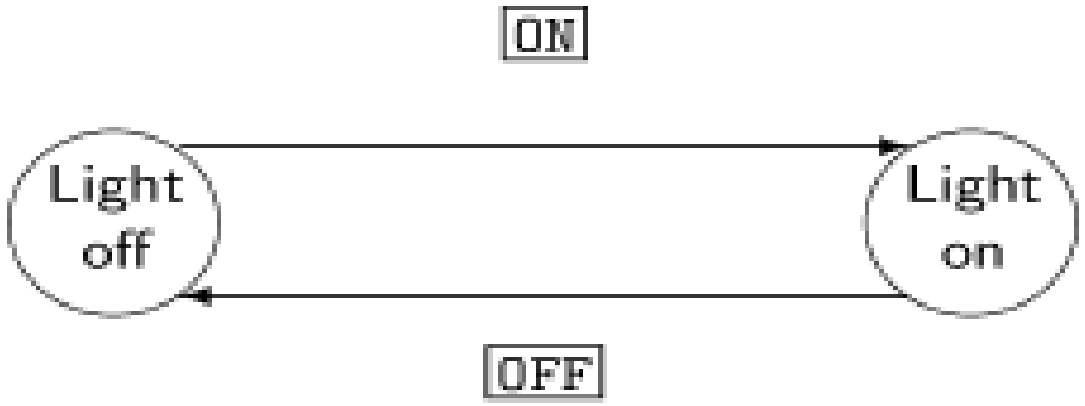
Push Button Device: FSM to MA

- A problem: the lamp has failed and we want to replace it
  - Need to know if it is safe (i.e. the system is in the off state)
  - How many times the user has to press the button so that he is sure that he is in a safe state (remember, we are not sure of the current state)

$$\boxed{\text{PUSH}}'' = \begin{pmatrix} 0 & 1 \\ 0 & 1 \end{pmatrix}$$

Push Button Device: FSM to MA

- Mathematically, we can show that NO value of n exists for this system that satisfies the equation
  - We need a two-position switch



b. Light with separate on/off actions.

Push Button Device: FSM to MA

- A two-position switch gives two options: ON and OFF

$$[ON] = \begin{pmatrix} 1 & 0 \\ 1 & 0 \end{pmatrix}$$

$$[OFF] = \begin{pmatrix} 0 & 1 \\ 0 & 1 \end{pmatrix}$$

- We can check, through matrix algebra, that OFF switch works as intended

$$\text{On}[OFF] = \begin{pmatrix} 1 & 0 \end{pmatrix} \begin{pmatrix} 0 & 1 \\ 0 & 1 \end{pmatrix} = \begin{pmatrix} 0 & 1 \end{pmatrix} = [OFF]$$

i.e., pressing **Off** after **On** keeps the system in [OFF] state

$$\text{Off}[OFF] = \begin{pmatrix} 0 & 1 \end{pmatrix} \begin{pmatrix} 0 & 1 \\ 0 & 1 \end{pmatrix} = \begin{pmatrix} 0 & 1 \end{pmatrix} = [OFF]$$

i.e., pressing **Off** after **Off** keeps the system in [OFF] state

Motivation to Use Formalism

- From the previous example, we can see that the formal representation (in this case, the matrices) allow us to check for the system properties through formal analysis (in this case, matrix algebra)
- From this analysis, we can decide if there are any usability problems with the system (e.g., single button is not good in case of failure, as it is dangerous to repair)
- That is the main motivation for having formalism in dialog representation (it allows us to check for properties that should be satisfied for having a usable system)

System Properties

- The properties that are checked for a usable system are of two types
  - Action properties
  - State properties

Action Properties

- There are mainly three action properties that a usable system should satisfy
  - Completeness: if all the transition leads to acceptable/final states or there are some *missed arcs*, i.e., some transition sequence don't lead to final states
  - Determinism: if there are several arcs (transitions) for the same action
  - Consistency: whether the same action results in the same effect (state transition) always

How Action Properties Help

- Completeness ensures that a user never gets struck at some state of the dialog, which s/he doesn't want, and can't come out from there (leads to frustration and less satisfaction)
- Lack of determinism introduces confusion and affects learnability and memorability
- An inconsistent interface reduces memorability and learnability, thus reducing the overall usability

State Properties

- There are mainly three state properties related to system usability
  - Reachability: can we get to any state from any other state?
  - Reversibility: can we return to the previous state from the current state?
  - Dangerous states: Are there any undesirable states that leads to deadlock (i.e. no further transitions are possible)?

How State Properties Help

- Reachability ensures that all system's features can be used
- Reversibility ensures that the user can recover from mistakes, thus increasing confidence and satisfaction
- Detection of dangerous states ensures that the user never goes to one, thus avoiding potential usability problems