Lecture 5:

Statistical Methods for Computer Science

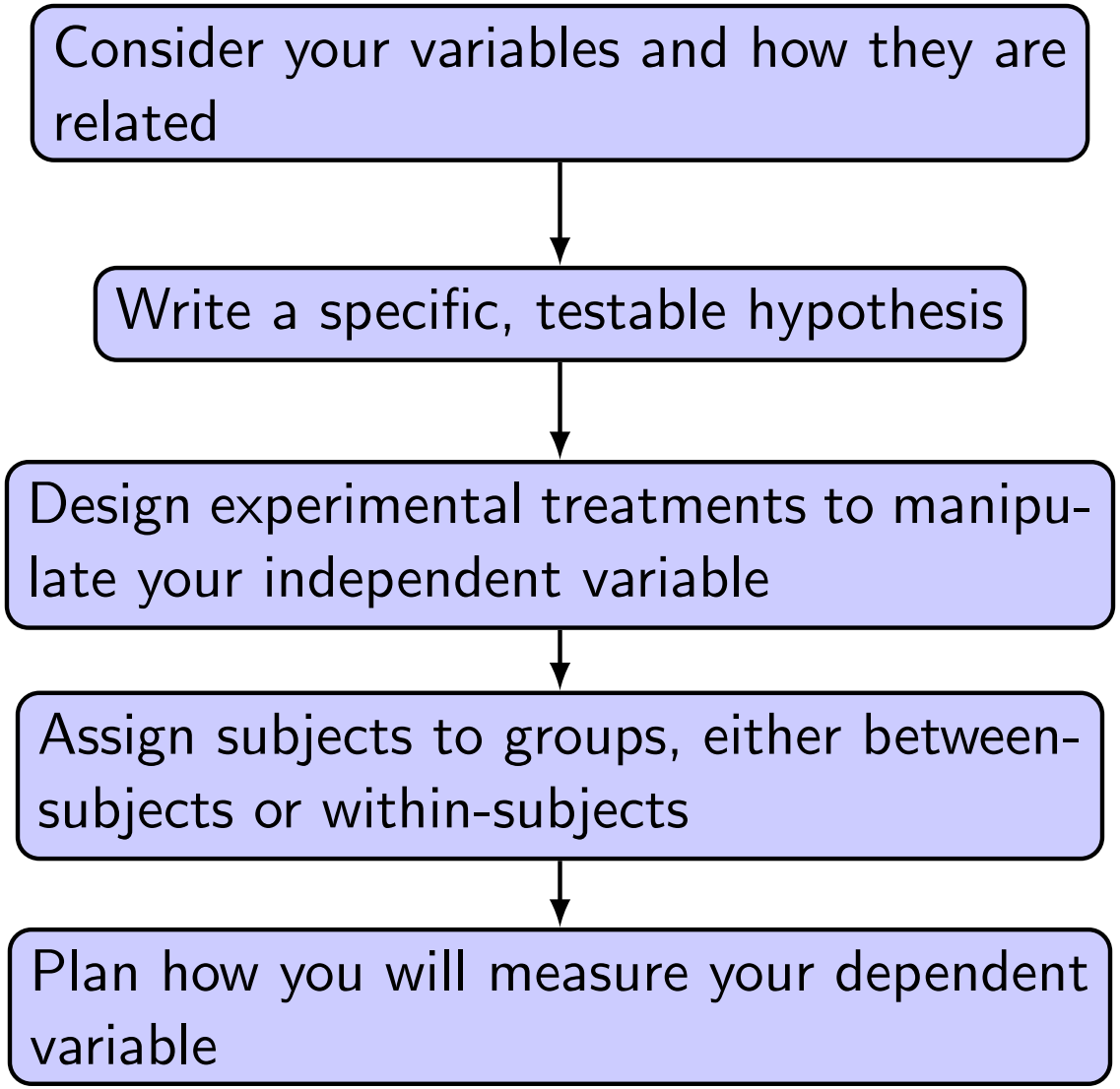
Faculty of Computing & Informatics
Multimedia University
Cyberjaya, Malaysia

Experimental/Empirical Research

- **Empirical Research** $\Rightarrow$ observation-based investigation seeking to discover and interpret facts, theories, or laws.
- Experiments are used to study causal relationships. You manipulate one or more independent variables and measure their effect on one or more dependent variables.
- variations:
  - Pre-experimental designs
  - Quasi-experimental designs

Key Steps

Experimental/Empirical Research



The simplest form of research design. In a pre-experiment either a single group or multiple groups are observed subsequent to some agent or treatment presumed to cause change.

- One-Shot Experimental Case Study
  Group 1: Tx $\rightarrow$ Obs
- One-Group Pretest-Posttest Design
  Group 1: Obs $\rightarrow$ Tx $\rightarrow$ Obs
- Static Group Comparison
  Group 1: Tx $\rightarrow$ Obs
  Group 2: $-\rightarrow$ Obs

True-Experimental Design

- Designs have a truly independent variable, manipulated by the researcher. Experimental designs offer control and internal validity
- Aims to establish a cause-and-effect relationship between an independent and dependent variable.
- Rely on random assignment of subjects into groups based on non-random criteria.

- *True* experimental designs have a truly independent variable, manipulated by the researcher. Experimental designs offer control and internal validity
- Like a true experiment, a quasi-experimental design aims to establish a cause-and-effect relationship between an independent and dependent variable.
- A quasi-experiment does not rely on random assignment. Instead, subjects are assigned to groups based on non-random criteria.
- Quasi-experimental design is a useful tool in situations where true experiments cannot be used for ethical or practical reasons.

True- vs. Quasi-Experimental Design

Experimental/Empirical Research

	True experimental design	Quasi-experimental design
Assignment to treatment	The researcher **randomly assigns** subjects to control and treatment groups.	Some other, **non-random** method is used to assign subjects to groups.
Control over treatment	The researcher usually **designs the treatment**.	The researcher often **does not have control over the treatment**, but instead studies pre-existing groups that received different treatments after the fact.
Use of control groups	Requires the use of **control and treatment groups**.	Control groups are not required (although they are commonly used).

Statistical Analysis of Data

- Given a set of measurements of a value, how certain can we be of the value?
- Given a set of measurements of two values, how certain can we be that the two values are different?
- Given a measured outcome, along with several condition or treatment values, how can we remove the effect of unwanted conditions or treatments on the outcome?

Central Limit Theorem

- The *Central Limit Theorem* says that the distribution of the sample means is normally distributed.
- If the original data is normally distributed with mean μ and standard deviation σ, then the sample means will be normally distributed with mean μ and standard deviation $\sigma' = \sigma/\sqrt{n}$ (but we don't know the original μ and σ...):

$$p(x) = \frac{1}{\sqrt{2\pi\sigma'}} e^{-\frac{1}{2}[\frac{x-\mu}{\sigma'}]^2}$$

- Note that it isn't important to remember this formula, since Matlab, R, etc. will do this for you. But it is very important to understand why you are computing it!

Central value

Give information concerning the average or typical score of a number of scores

- *mean*
- *median*
- *mode*

Median

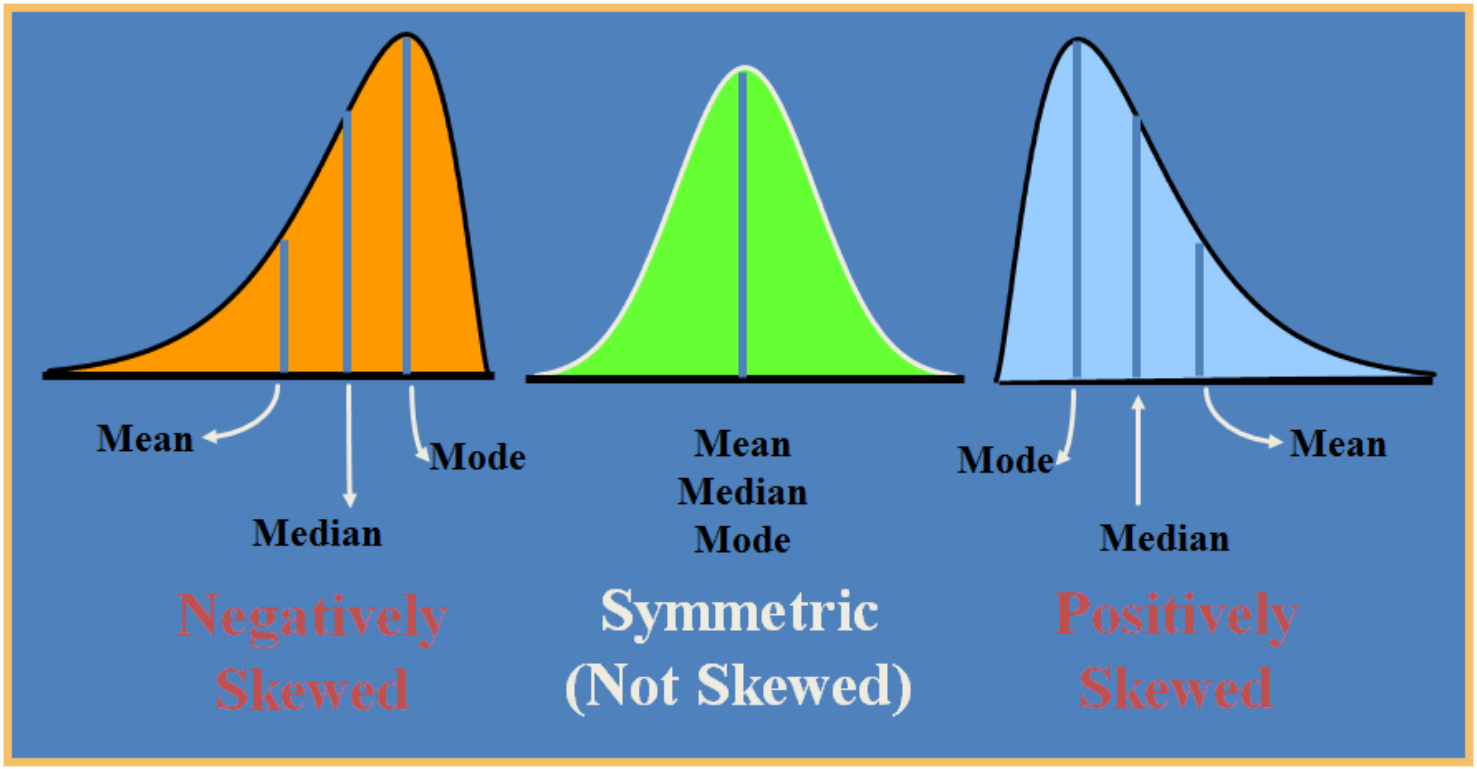
Central value

- Middlemost or most central item in the set of ordered numbers; it separates the distribution into two equal halves
- If $odd\ n$, middle value of sequence. if X = [1,2,4,6,$\underline{9}$,10,12,14,17], then 9 is the median
- If $even\ n$, average of 2 middle values if X = [1,2,4,6,$\underline{9,10}$,11,12,14,17] then 9.5 is the median; i.e., (9+10)/2 Median is not affected by extreme values

- The mode is the most frequently occurring number in a distribution. if X = [1,2,4,7,7,7,8,10,12,14,17], then 7 is the mode.
- Easy to see in a simple frequency distribution
- Possible to have no modes or more than one mode
  – bimodal & multimodal
- Don't have to be exactly equal frequency
  – major mode, minor mode
- Mode is not affected by extreme values

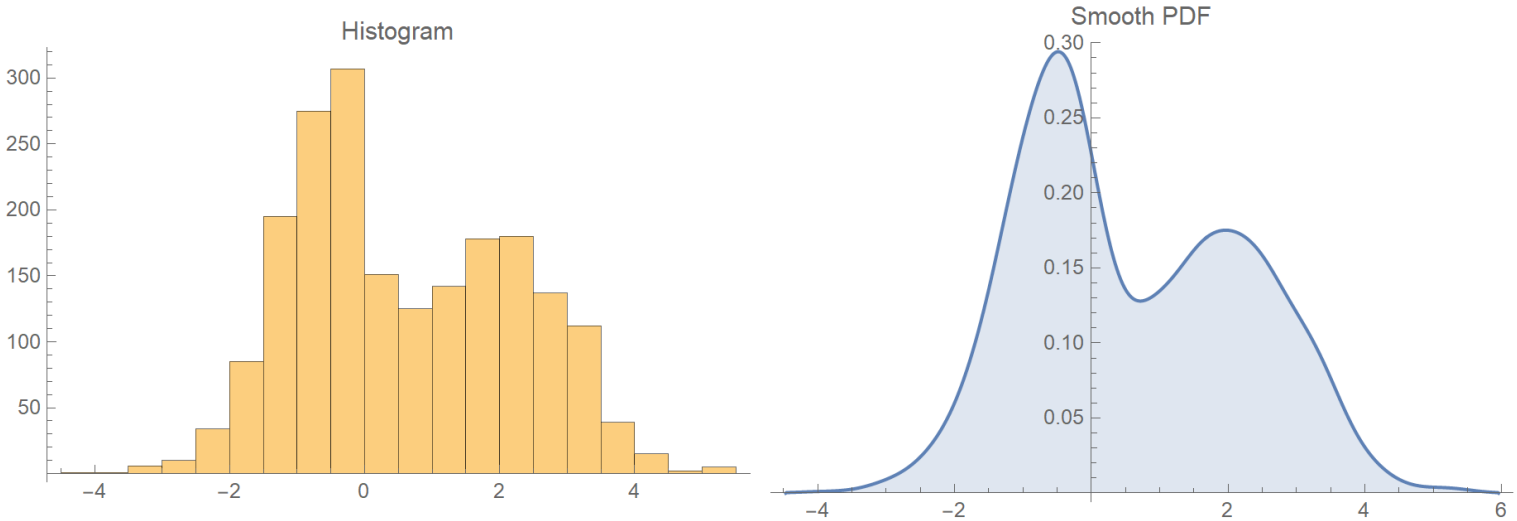
Mean, Median, Mode

Central Value



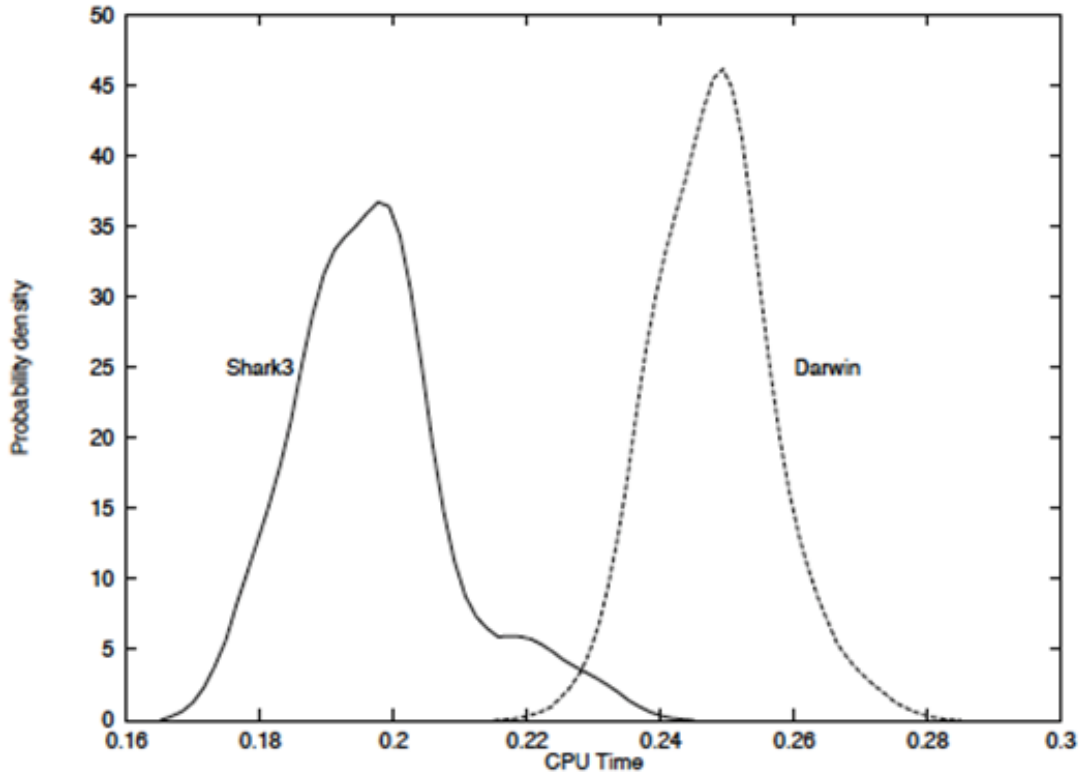
Kernel Density Estimation

kernel density estimation (KDE) is a non-parametric way to estimate the probability density function of a random variable. Kernel density estimation is a fundamental data smoothing problem where inferences about the population are made, based on a finite data sample.



Kernel Density Visualization

Visually, the second machine (Shark3) is much faster than the first (Darwin):



Normal Distribution

Normal distribution, also known as the Gaussian distribution, is a probability distribution that is symmetric about the mean, showing that data near the mean are more frequent in occurrence than data far from the mean.

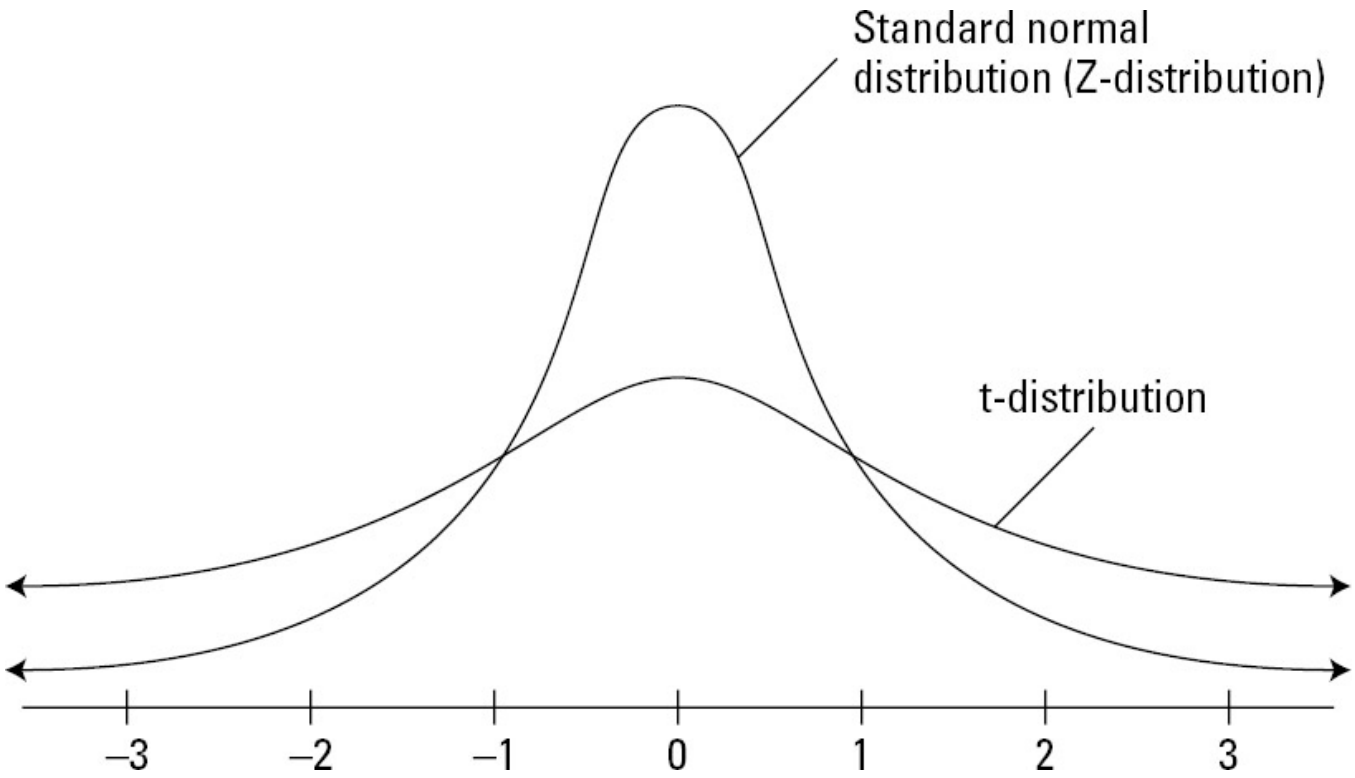
1. A normal distribution is the proper term for a probability bell curve.
2. In a normal distribution the mean is zero and the standard deviation is 1. It has zero skew and a kurtosis of 3.
3. The 68-95-99.7 rule states that the percentage of values that lie within a band around the mean in a normal distribution with a width of two, four and six standard deviations, comprise 68%, 95% and 99.7% of all the values.

t-Distribution

- Instead of assuming a normal distribution, we can use a t distribution (sometimes called a "Student's t distribution"), which has three parameters: μ, σ, and the degrees of freedom (d.f. $= n$-1)
- The probability distribution function looks somewhat like a normal distribution, but gives a tighter peak (with longer tails) as n increases
- This distribution yields just slightly tighter confidence limits, using the central limit theorem:

t-Distribution

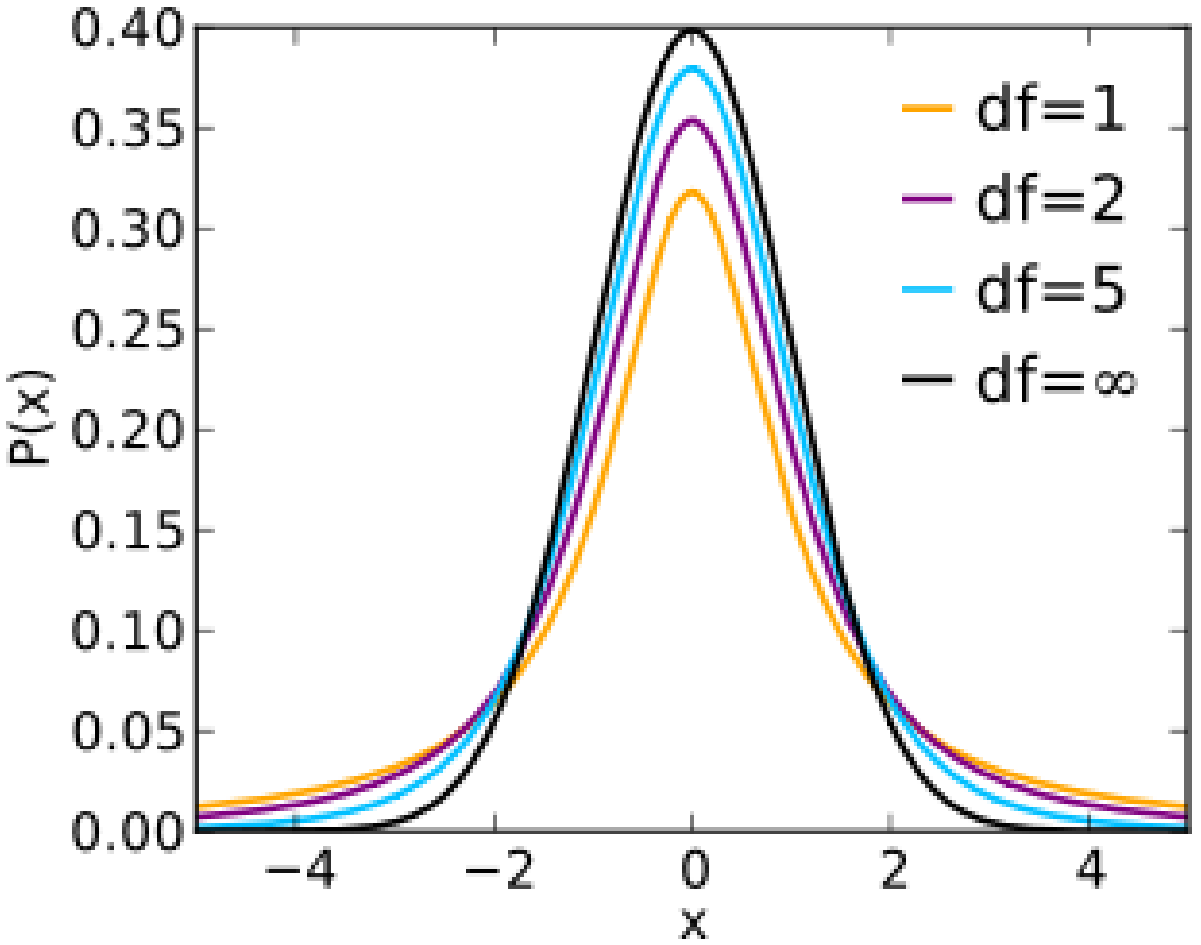
Comparing with Normal Distribution



When sample size increases, t-distribution $=$ z-distribution

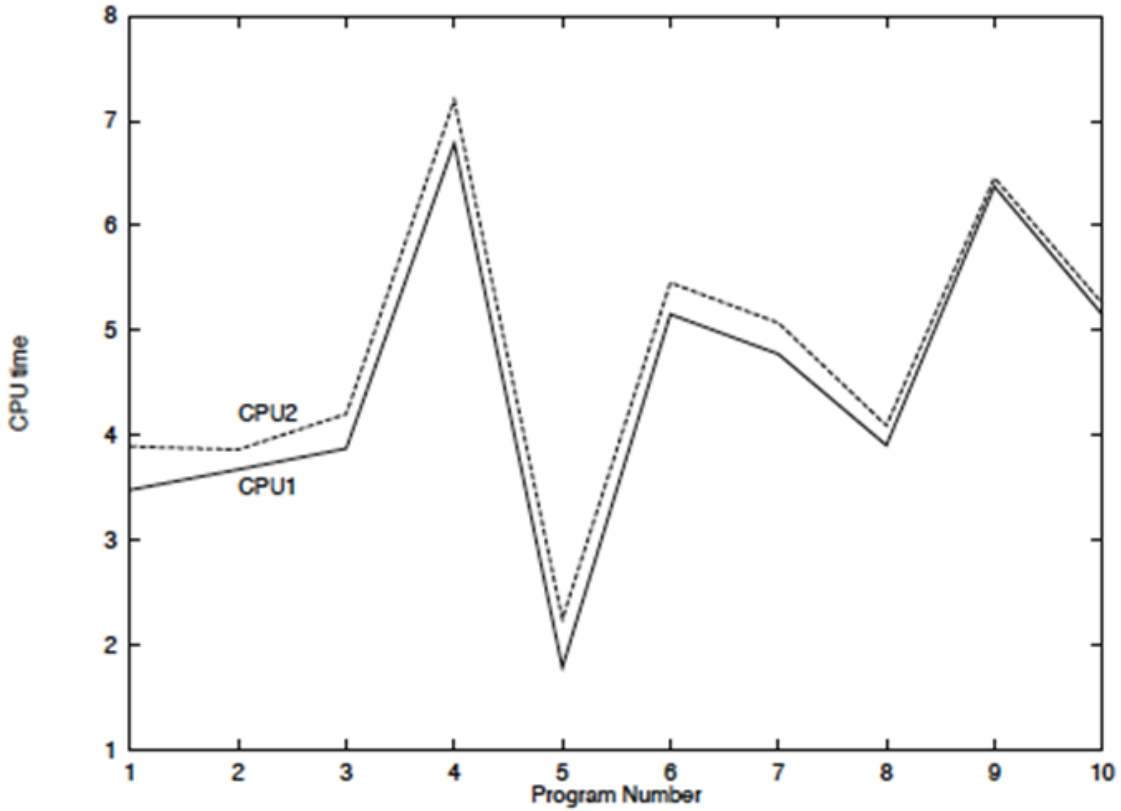
t-Distribution

with different df



Sequential Visualization

- The co-correlation of program "difficulty" (and faster CPU speed of CPU1) is even more obvious in this ordered (by program number) line plot:



Hypothesis Testing

- Is the true difference zero, or more than zero?
- Use classical statistical rejection testing
  1. Null hypothesis: The two machines have the same speed (i.e., μ, the difference in sample rate, is equal to zero)
  2. Can we reject this hypothesis, based on the observed data?
  3. If the null hypothesis were true, what is the probability we would have observed this data?
  4. We can measure this probability using the t distribution
  5. In this case, the computed t value $= (\mu 1 - \mu 2) \ / \ \sigma' = 21.69$
  6. The probability of seeing this t value, if μ was actually zero, is nearly nonexistent: The 99.999% confidence interval (for the null hypothesis) is [-4.59, 4.59], so the probability of this t value is (much) less than 0.00001

Intro to t-test

Introduction to t-test

One of the most popular ways to test a hypothesis is a concept called the t-test. There are different types of t-tests.

Scenario:
Consider a telecom company that has two service centers in the city. The company wants to find whether the average time required to service a customer is the same in both stores. The company measures the average time taken by 50 random customers in each store. Store A takes 22 minutes while Store B averages 25 minutes. Can we say that Store A is more efficient than Store B in terms of customer service?

Can we merely based on the average sample time to make conclusion?

Assumptions for Performing a t-test

1. The data should follow a continuous or ordinal scale
2. The observations in the data should be randomly selected
3. The data should resemble a bell-shaped curve when we plot it, i.e., it should be normally distributed.
4. Large sample size should be taken for the data to approach a normal distribution
5. Variances among the groups should be equal (for independent two-sample t-test)

There are three types of t-tests we can perform based on the data at hand:

1. One sample t-test
2. Independent two-sample t-test
3. Paired sample t-test

In a one-sample t-test, we compare the average (or mean) of **one group** against the **set average** (or mean). This set average can be any theoretical value (or it can be the population mean).

Consider the following example – A research scholar wants to determine if the average eating time for a (standard size) burger differs from a set value. Let's say this value is 10 minutes. How do you think the research scholar can go about determining this? He/she can broadly follow the below steps:

1. Select a group of people
2. Record the individual eating time of a standard size burger
3. Calculate the average eating time for the group
4. Finally, compare that average value with the set value of 10

we can perform a one-sample t-test. Here's the formula to calculate this:

$t = \frac{m-\mu}{\frac{s}{\sqrt{n}}}$

where,

t = t-statistic

m = mean of the group

μ = theoretical value or population mean

s = standard deviation of the group

n = group size or sample size

Note: Once we have calculated the t-statistic value, the next task is to compare it with the critical value of the t-test.

Suppose a botanist wants to know if the mean height of a certain species of plant is equal to 15 inches. She collects a random sample of 12 plants and records each of their heights in inches.

```
import scipy.stats as stats
data = [14, 14, 16, 13, 12, 17, 15, 14, 15, 13, 15, 14]
stats.ttest_1samp(a=data, popmean=15)
```

The t-test statistic is -**1.6848** and the corresponding two-sided p-value is **0.1201**.

The two hypotheses for this particular one sample t-test are as follows:

$H_0 : \mu = 15$ (the mean height for this species of plant is 15 inches)
$H_1 : \mu \neq 15$ (the mean height is not 15 inches)

Because the p-value of our test (0.1201) is greater than $\alpha = 0.05$, we fail to reject the null hypothesis of the test. We **do not** have sufficient evidence to say that the mean height for this particular species of plant is different from 15 inches.

The two-sample t-test is used to compare the means of two different samples.

Let's say we want to compare the average height of the male employees to the average height of the females. Of course, the number of males and females should be equal for this comparison. This is where a two-sample t-test is used.

The formula to calculate the t-statistic for a two-sample t-test:

$$t = \frac{m_A - m_B}{\sqrt{\frac{S^2}{n_A} + \frac{S^2}{n_B}}}$$

where,
m_A and m_B are the means of two different samples
n_A and n_B are the sample sizes
S_2 is an estimator of the common variance of two samples
Here, the degree of freedom is $n_A + n_B - 2$.

```
import scipy.stats as stats
data_group1 = np.array([14, 15, 15, 16, 13, 8, 14,
                        17, 16, 14, 19, 20, 21, 15,
                        15, 16, 16, 13, 14, 12])
data_group2 = np.array([15, 17, 14, 17, 14, 8, 12,
                        19, 19, 14, 17, 22, 24, 16,
                        13, 16, 13, 18, 15, 13])
print(np.var(data_group1), np.var(data_group2))
```

make sure you check the variance of the 2 groups

Here, the ratio is **12.260 / 7.7275** which is less than **4:1**.

Before conducting the two-sample T-Test we need to find if the given data groups have the same variance. If the ratio of the larger data groups to the small data group is less than 4:1 then we can consider that the given data groups have **equal variance**.

```
stats.ttest_ind(a=data_group1, b=data_group2,
    equal_var=True)
```

Hypothesis:

$H_0 : \mu_1 = \mu_1$ (population mean of dataset1 is equal to dataset2)
$H_1 : \mu_1 \neq \mu_1$ (population mean of dataset1 is different from dataset2)

since the p-value (0.53004) is greater than alpha $= 0.05$ so we **cannot reject the null hypothesis**. We do not have sufficient evidence to say that the mean height of students between the two data groups is different.

This test is also known as the *dependent sample t-test*. It is a statistical concept and is used to check whether the mean difference between the two sets of observation is equal to zero. Each entity is measured is two times in this test that results in the pairs of observations.

The formula to calculate the t-statistic for a paired t-test is:

$$t = \frac{m}{\frac{s}{\sqrt{n}}}$$

where,
t = t-statistic
m = mean of the group
μ = theoretical value or population mean
s = standard deviation of the group
n = group size or sample size

We can take the degree of freedom in this test as $n - 1$ since only one group is involved.

Suppose below is the pre-test and post-test result:

```
# Pre-test result
pre = [88, 82, 84, 93, 75, 78, 84, 87,
95, 91, 83, 89, 77, 68, 91]

# Post-test result
post = [91, 84, 88, 90, 79, 80, 88, 90,
90, 96, 88, 89, 81, 74, 92]

# Performing the paired sample t-test
stats.ttest_rel(pre, post)
```

The test statistic comes out to be equal to **2.584** and the corresponding two-sided p-value is **0.029**.

The paired samples t-test follows the null and alternative hypotheses:

H_0: It signifies that the mean pre-test and post-test scores are equal
H_1: It signifies that the mean pre-test and post-test scores are not equal

As the p-value comes out to be equal to **0.029** which is less than 0.05 hence **we reject the null hypothesis**. So, we have *enough proof* to claim that the true mean test score is different for cars before and after applying the different engine oil.

Intro to Annova

Introduction to ANOVA

An **ANOVA** test is a way to find out if experiment results are significant. you're testing groups to see if there's a difference between them. Examples of when you might want to test different groups:

1. A group of psychiatric patients are trying three different therapies: counseling, medication and biofeedback. You want to see if one therapy is better than the others.
2. A manufacturer has two different processes to make light bulbs. They want to know if one process is better than the other.
3. Students from different colleges take the same exam. You want to see if one college outperforms the other.

Types of ANOVA test

1. One-way ANOVA between groups: used when you want to test two groups to see if there's a difference between them.
2. Two way ANOVA without replication: used when you have one group and you're double-testing that same group. For example, you're testing one set of individuals before and after they take a medication to see if it works or not.
3. Two way ANOVA with replication: Two groups, and the members of those groups are doing more than one thing. For example, two groups of patients from different hospitals trying two different therapies.

Scenario:

Researchers took **20** cars of the same to take part in a study. These cars are randomly doped with one of the four-engine oils and allowed to run freely for 100 kilometers each. At the end of the journey, the performance of each of the cars is noted.

One-Way ANOVA using Python

```
# Importing library
from scipy.stats import f_oneway

# Performance when each of the engine oil is applied
performance1 = [89, 89, 88, 78, 79]
performance2 = [93, 92, 94, 89, 88]
performance3 = [89, 88, 89, 93, 90]
performance4 = [81, 78, 81, 92, 82]

# Conduct the one-way ANOVA
f_oneway(performance1, performance2,
    performance3, performance4)
```

One-Way ANOVA using Python

A one-way ANOVA uses the following null and alternative hypotheses:

H_0: $\mu_1 = \mu_2 = \mu_3 = \mu_4$ (all the population means are equal)
H_1: at least one population mean is different from the rest

The F test statistic is 4.625 and the corresponding p-value is 0.01633. Since the p-value is less than .05, we **can reject** the null hypothesis. Therefore, indicating that we do have sufficient evidence to say that there is a difference in engine oil.

Two-Way ANOVA

A two-way ANOVA is used to determine whether or not there is a statistically significant difference between the means of three or more independent groups that have been split on two factors.

The purpose of a two-way ANOVA is to determine how two factors impact a response variable, and to determine whether or not there is an interaction between the two factors on the response variable.

Scenario:

A botanist wants to know whether or not plant growth is influenced by sunlight exposure and watering frequency. She plants 30 seeds and lets them grow for two months under different conditions for sunlight exposure and watering frequency. After two months, she records the height of each plant, in inches.

Use the following steps to perform a two-way ANOVA to determine if watering frequency and sunlight exposure have a significant effect on plant growth, and to determine if there is any interaction effect between watering frequency and sunlight exposure.

Two-Way ANOVA using Python

```
import numpy as np
import pandas as pd

#create data
df = pd.DataFrame(
 {'water': np.repeat(['daily', 'weekly'], 15),
  'sun': np.tile(np.repeat(['low', 'med', 'high'], 5), 2),
  'height': [6, 6, 6, 5, 6, 5, 5, 6, 4, 5,
             6, 6, 7, 8, 7, 3, 4, 4, 4, 5,
             4, 4, 4, 4, 4, 5, 6, 6, 7, 8]})

#view first ten rows of data
df[:10]
```

Two-Way ANOVA using Python

```
import statsmodels.api as sm
from statsmodels.formula.api import ols

#perform two-way ANOVA
model = ols('height ~ C(water) +
                      C(sun) +
                      C(water):C(sun)',
                      data=df).fit()
sm.stats.anova_lm(model, typ=2)
```

Two-Way ANOVA using Python

We can see the following p-values for each of the factors in the table:

water: p-value = **.000527**
sun: p-value = **.0000002**
water*sun: p-value = **.120667**

Since the p-values for water and sun are both less than .05, this means that both factors have a statistically significant effect on plant height.

And since the p-value for the interaction effect (.120667) is not less than .05, this tells us that there is **no significant** interaction effect between sunlight exposure and watering frequency. Meaning the interaction between sunlight exposure and watering frequency does not significantly impact the height.

Pearson's Chi-Square Test

Pearson's Chi-Square Test

The Pearson's Chi-Square statistical hypothesis is a test for independence between categorical variables.

A Contingency table (also called crosstab) is used in statistics to summarise the relationship between several categorical variables.

Pearson's Chi-Square Test

Below is a table that shows the number of men and women buying different types of pets.

	dog	cat	bird	total
men	207	282	241	730
women	234	242	232	708
total	441	524	473	1438

The aim of the test is to conclude whether the two variables(gender and choice of pet) are related to each other.

Pearson's Chi-Square Test

Hypothesis:

H_0: there is no relation between the variables

H_1: there is a significant relation between the two variables.

Expected Values Table :

Next, we prepare a similar table of calculated(or expected) values. To do this we need to calculate each item in the new table as :

$$\frac{row\ total \times column\ total}{grand\ total}$$

The expected values table :

	dog	cat	bird	total
men	223.87343533	266.00834492	240.11821975	730
women	217.12656467	257.99165508	232.88178025	708
total	441	524	473	1438

Chi-Square Table :

We prepare this table by calculating for each item the following:

$$\frac{(Observed\ value - Calculated\ value)^2}{Calculated\ value}$$

	observed (o)	calculated (c)	(o-c)^2 / c
	207	223.87343533	1.2717579435607573
	282	266.00834492	0.9613722161954465
	241	240.11821975	0.003238139990850831
	234	217.12656467	1.3112758457617977
	242	257.99165508	0.991245364156322
	232	232.88178025	0.0033387601600580606
Total			**4.542228269825232**

From this table, we obtain the total of the last column, which gives us the calculated value of chi-square. Hence the calculated value of chi-square is **4.542228269825232**.

Now, we need to find the critical value of chi-square. We can obtain this from a table. To use this table, we need to know the degrees of freedom for the dataset. The degrees of freedom is defined as : (no. of rows – 1) * (no. of columns – 1). Hence, the degrees of freedom is **(2-1) * (3-1) = 2**

Pearson's Chi-Square Test

Now, let us look at the table and find the value corresponding to 2 degrees of freedom and 0.05 significance factor :

Critical values of the Chi-square distribution with d degrees of freedom

	Probability of exceeding the critical value						
d	0.05	0.01	0.001	d	0.05	0.01	0.001
1	3.841	6.635	10.828	11	19.675	24.725	31.264
2	5.991	9.210	13.816	12	21.026	26.217	32.910
3	7.815	11.345	16.266	13	22.362	27.688	34.528
4	9.488	13.277	18.467	14	23.685	29.141	36.123
5	11.070	15.086	20.515	15	24.996	30.578	37.697
6	12.592	16.812	22.458	16	26.296	32.000	39.252
7	14.067	18.475	24.322	17	27.587	33.409	40.790
8	15.507	20.090	26.125	18	28.869	34.805	42.312
9	16.919	21.666	27.877	19	30.144	36.191	43.820
10	18.307	23.209	29.588	20	31.410	37.566	45.315

The tabular or critical value of chi-square here is **5.991**.

Pearson's Chi-Square Test

Critical value of $\chi^2 \geqq$ calculated value of χ^2.

Therefore, H_0 **is accepted**, that is, the variables do not have a significant relation.

Pearson's Chi-Square Test using Python

```
from scipy.stats import chi2_contingency

# defining the table
data = [[207, 282, 241], [234, 242, 232]]
stat, p, dof, expected = chi2_contingency(data)

# interpret p-value
alpha = 0.05
print("p value is " + str(p))
if p <= alpha:
    print('Dependent (reject H0)')
else:
    print('Independent (H0 holds true)')
```

Another Example using Python

Pearson's Chi-Square Test using Python

```
import pandas as pd
import numpy as np
from scipy.stats import chi2_contingency
import seaborn as sns
import matplotlib.pyplot as plt
%matplotlib inline
df = pd.DataFrame(
  {'Gender' : ['M', 'M', 'M', 'F', 'F'] * 10,
    'isSmoker' : ['Smoker', 'Smoker',
    'Non-Smpoker', 'Non-Smpoker', 'Smoker'] * 10
   })
df.head()
```

Pearson's Chi-Square Test using Python

```
   Gender isSmoker
0  M      Smoker
1  M      Smoker
2  M      Non-Smpoker
3  F      Non-Smpoker
4  F      Smoker
```

Pearson's Chi-Square Test using Python

```
contigency= pd.crosstab(df['Gender'], df['isSmoker'])
contigency
```

isSmoker	Non-Smpoker	Smoker
Gender		
F	10	10
M	10	20

Pearson's Chi-Square Test using Python

```
# Chi-square test of independence.
c, p, dof, expected = chi2_contingency(contigency)
# Print the p-value
print(p)
```

The p-value is **0.3767** which means that we do not reject the null hypothesis at 95% level of confidence. The null hypothesis was that Smokers and Gender are independent.

Declaration & Acknowledgment

The contents presented in this slide are meant for teaching purpose only. No commercialized component exist. The texts are partially copied from the textbooks and images are downloaded from the internet.