# PRINCIPLES OF DATA MANAGEMENT

UNIVERSITY of HOUSTON

DIVISION OF RESEARCH

# Hypothesis Testing

- where we use observed data in a sample to draw conclusions about unobserved data — often the population

- The general idea of hypothesis testing involves:

  1. Making an initial assumption
  2. Collecting evidence (data)
  3. Based on the available evidence (data), deciding whether to reject or not reject the initial assumption

- Every hypothesis test — regardless of the population parameter involved — requires the above three steps.

# Hypothesis testing

**Examples:**

- Is Normal Body Temperature Really 98.6 Degrees F?

- An online retailer tests whether customer engagement under website design A is different than under website design B

- A pharmaceutical company tests whether the efficacy of a new blood pressure medication differs from an existing competitor's product

- A polling organization tests whether the strength of support for one mayoral candidate differs from another

# Hypothesis testing

The null hypothesis, $H_0$, is usually the hypothesis that corresponds to the status quo, the standard, the desired level/amount, or it represents the statement of "no difference."

The alternative hypothesis, $H_1$, on the other hand, is the complement of $H_0$, and is typically the statement that the researcher would like to prove or verify.

Another word of caution: It is not proper for a researcher to set up the hypotheses after seeing the sample data; however, a data maybe used to generate a hypotheses, but to test these generated hypotheses you should gather a new set of sample data!

# Hypothesis testing

These hypotheses are usually set-up in such a way that deciding in favor of $H_1$ when in fact $H_0$ is the true (called a Type I error) statement is a very serious mistake.

# Hypothesis testing

Consider the population of adults.

**Hypothesis Test**: Average adult body temperature is lower than the often-advertised 98.6 degrees F
**Question:** Is the average adult body temperature 98.6 degrees? Or is it lower?
**Initial Assumption:**
**(Null Hypothesis):** Average adult body temperature was **98.6 degrees F**
**(Alternate Hypothesis):**Average adult body temperature is lower than **98.6 degrees F**

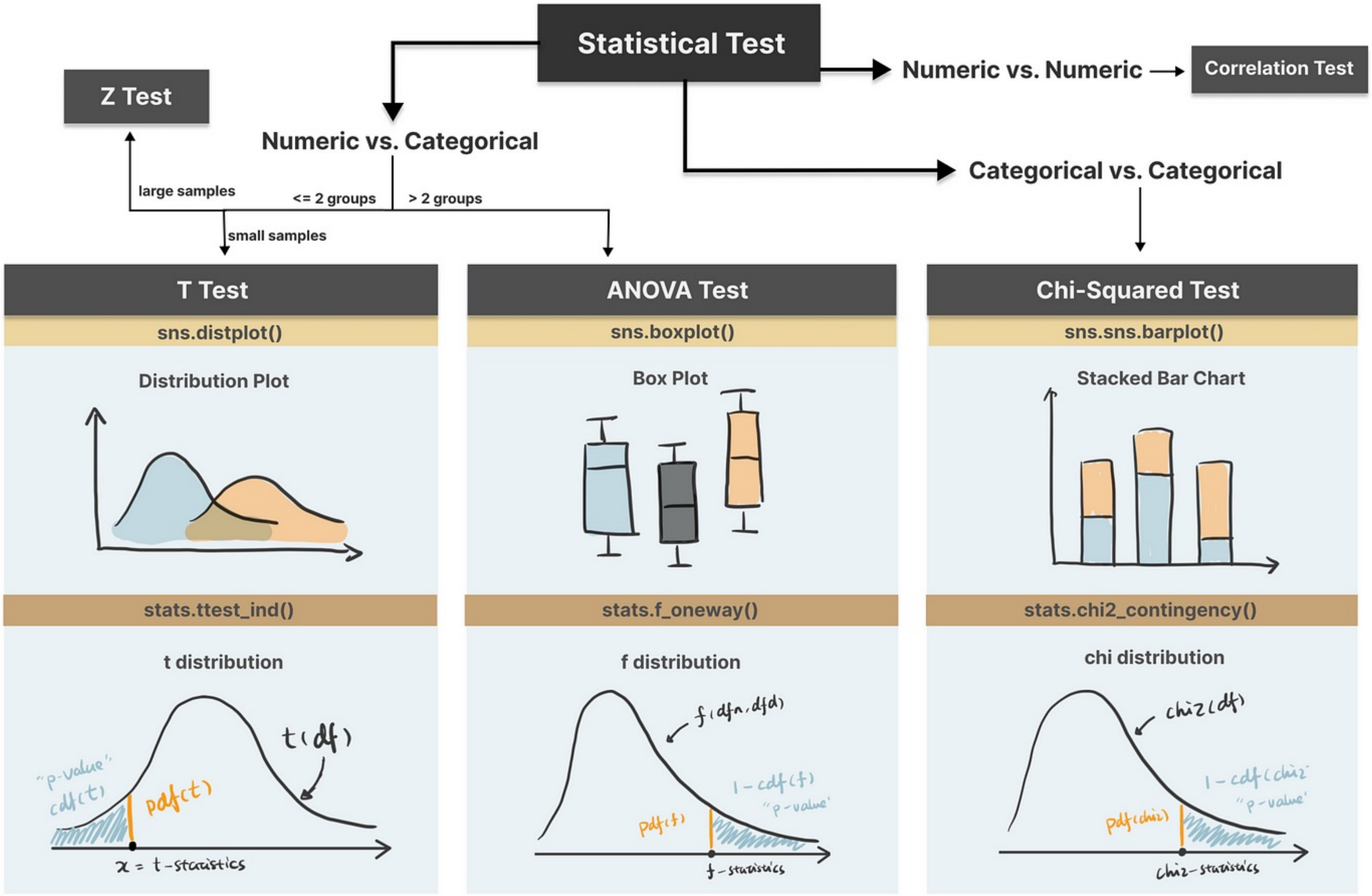
**Evidence(Data):** Select a **random sample** of **150** adults. The average body temperature of the 150 sampled adults **is 98.25 degrees**

**When we have not determined what significance level is and the p-value:**

**The outcome will be as follows**
It is either *likely(fail to reject null H)* or *unlikely(reject null H)*

- If it is *likely*, then the researcher **does not reject** his initial assumption that the average adult body temperature is 98.6 degrees. There is not enough evidence to do otherwise.
- If it is *unlikely*, then:
  - either the initial assumption is correct, and there is an unusual event
  - or the initial assumption is incorrect

# How to decide the outcomes?



## Statistical Test

Numeric vs. Numeric → Correlation Test

Numeric vs. Categorical

- <= 2 groups
  - large samples → Z Test
  - small samples → T Test
- > 2 groups → ANOVA Test

Categorical vs. Categorical → Chi-Squared Test

### T Test

sns.distplot()

Distribution Plot

stats.ttest_ind()

t distribution

### ANOVA Test

sns.boxplot()

Box Plot

stats.f_oneway()

f distribution

### Chi-Squared Test

sns.sns.barplot()

Stacked Bar Chart

stats.chi2_contingency()

chi distribution

visit www.visual-design.net for step by step guide

# Statistical test types

- **Parametric Test:**

  Parametric statistics are based on assumptions about the **distribution of population** from which the sample was taken

  such as Normally distributed population

- **Non-Parametric Test:**

  Nonparametric statistics are not based on assumptions, that is, the data can be collected from a sample that **does not follow a specific distribution**

**Using the test above we can decide if the result is statistically significant or not**

# Statistical Test

- **T-test:** compare two groups/categories of numeric variables with **small sample size**
- **Z-test:** compare two groups/categories of numeric variables with **large sample size**
- **ANOVA test:** compare the difference between two or more groups/categories of numeric variables
- **Chi-Squared test:** examine the relationship between two categorical variables
- **Correlation test:** examine the relationship between two numeric variables

# Assumptions (T-test/Z-test)

- The observations in the data should be randomly selected
- The data should follow a continuous or ordinal scale
- It is useful for
  - Small sample size of data-set: t-test
  - Large sample size of data-set(>30): z-test
- The data should be normally distributed. What if my data isn't nearly normally distributed?
  - If you cannot safely assume normality, you can perform a **nonparametric test** that doesn't assume normality.
- Variances among the groups should be equal (f*or independent two-sample t-test*)

# Z-test &T-test types

- **One-Sample t-test**: is used to determine whether an unknown population mean is different from a specific value

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

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

- **Two Sample t-test**: is used to compare the means of two different samples

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

$$S^2 = \frac{\sum (x - m_A)^2 + \sum (x - m_B)^2}{n_A + n_B - 2}$$

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

**Degrees of Freedom:** is used to measure the amount of variability in the data. It can be the number of observations in a sample that are free to vary once conditions are applied

# Differences

## Z-test

- Z tests require you to know the population standard deviation
- It evaluates Z-scores in the context of the standard normal distribution

## t-test

- T tests use a sample estimate of the standard deviation
- The standard normal distribution doesn’t change shape as the sample size changes. Consequently, the critical values don’t change with the sample size

# Z-test

- One Sample z-test:

$$Z = \frac{\overline{x} - \mu_0}{\frac{\sigma}{\sqrt{n}}}$$

The numerator is the difference between your sample mean and a hypothesized value for the population mean ($\mu_0$). This value is often a strawman argument that you hope to disprove

The denominator is the standard error of the mean. It represents the uncertainty in how well the sample mean estimates the population mean

- Two Sample z-test:

$$Z = \frac{\overline{x_1} - \overline{x_2}}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}}$$

The numerator is the difference between your two sample means
The denominator calculates the pooled standard error of the mean by combining both samples. In this Z test formula, enter the population variances (σ2) for each sample

# Analysis of Variance (ANOVA)

is a statistical method that separates observed variance data into different components to use for additional tests

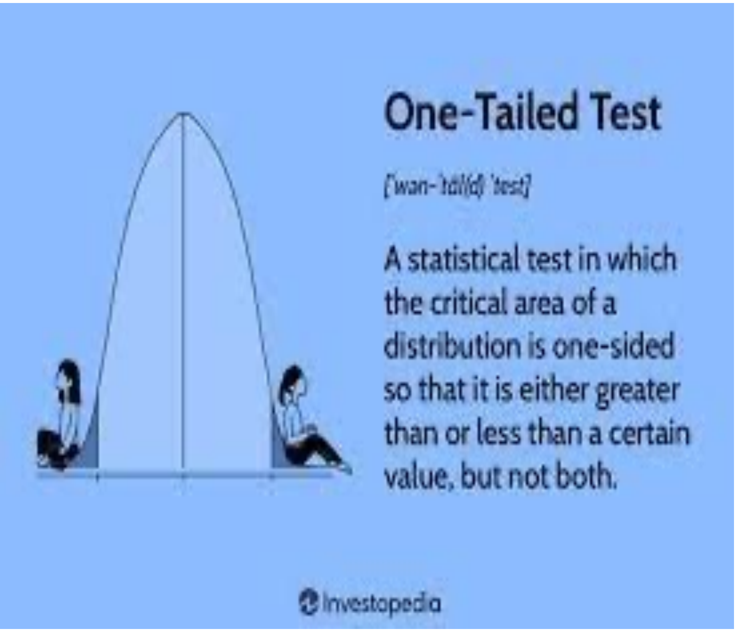
If no real difference exists between the tested groups, which is called the null hypothesis, the result of the ANOVA's F-ratio statistic will be close to 1.

F=MSE/MST

**where:**
F=ANOVA coefficient
MST=Mean sum of squares due to treatment
MSE=Mean sum of squares due to error

The ANOVA test allows a comparison of more than two groups at the same time to determine whether a relationship exists between them
The result of the ANOVA formula, the F statistic (also called the F-ratio), allows for the analysis of multiple groups of data to determine the variability between samples and within samples.

# Hypothesis Testing

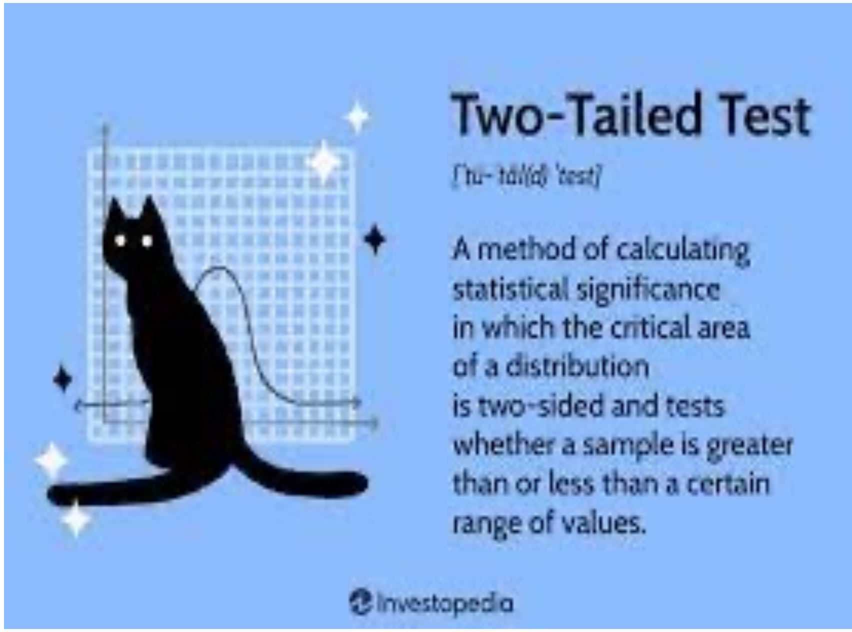


## One-Tailed Test

[ˈwən-ˈtāl(d) ˈtest]

A statistical test in which the critical area of a distribution is one-sided so that it is either greater than or less than a certain value, but not both.

Investopedia



## Two-Tailed Test

[ˈtü-ˈtāl(d) ˈtest]

A method of calculating statistical significance in which the critical area of a distribution is two-sided and tests whether a sample is greater than or less than a certain range of values.

Investopedia

*Image courtesy: Investopedia*

# Hypothesis Testing

- **Significance Level(α):** is a threshold used in hypothesis testing.

It is a pre-determined value at which you reject or retain your null hypothesis(H0)

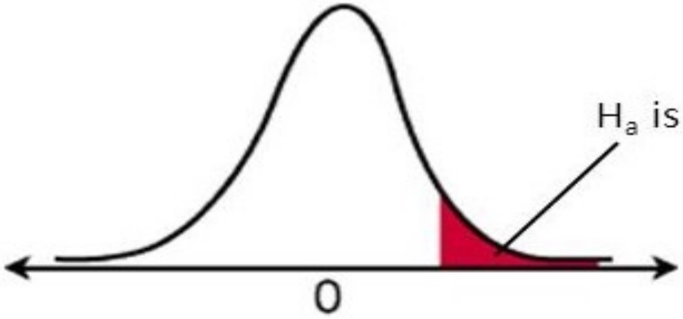
Common **α** values are:

- 0.05(5%)
- 0.01(1%)

- **p-value**: measures the probability of obtaining the observed results, assuming that the null hypothesis is true. It helps in validating a HYPOTHESIS against the observed data

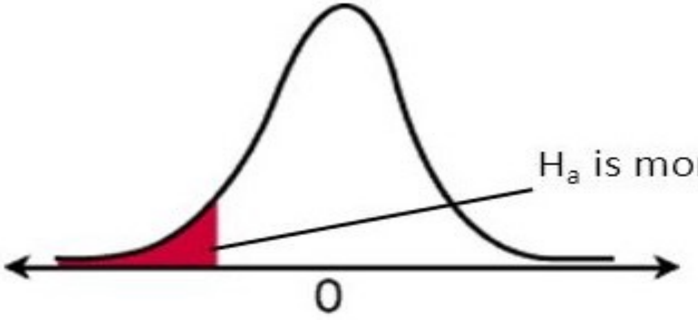
If p value is **larger than** the threshold, it means that the value is likely to occur in the distribution when the null hypothesis is true

if **lower than** significance level, it means it is very unlikely to occur in the null hypothesis distribution — so **reject the null hypothesis**



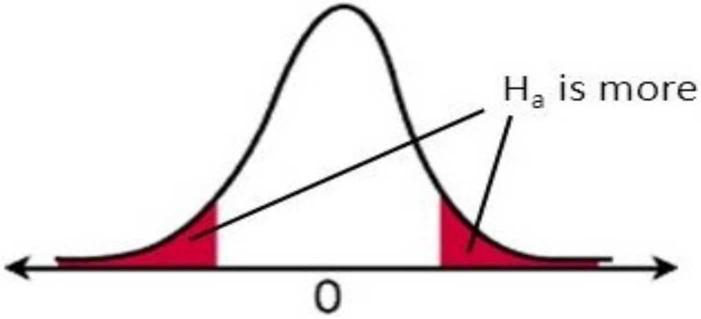
**Right-tail test**

$H_a$: $\mu >$ value



**Left-tail test**

$H_a$: $\mu <$ value



**Two-tail test**

$H_a$: $\mu \neq$ value

*Image courtesy: www.fromthegenesis.com*