Hypothesis Testing & Tests

Nagiza F. Samatova, samatova@csc.ncsu.edu
Professor, Department of Computer Science
North Carolina State University

NC STATE Executive Education

NC STATE UNIVERSITY
Department of Computer Science

Learning Objectives: Hypothesis Tests

- **Describe the difference between a hypothesis test and a confidence interval**
- **Correctly use vocabulary of hypothesis testing:**
  - **Null Hypothesis:** H_0
  - **Reject or Fail to Reject the Null Hypothesis**
  - **Alternate Hypothesis:** H_1
  - p**-value**
- **Test hypotheses for:**
  - **a single proportion**
  - **two proportions**
  - **two means**
- **Distinguish when to use one-way and two-way hypothesis test**

Descriptive vs. Inferential Statistics

- Descriptive or Summary Statistics
  - **Goal: to describe the features of a collection of data in a quantitative way**
  - **Measures of Central Tendency:**
    - mean, median, and mode
  - **Measures of Variability, Dispersion, or Spread:**
    - range, variance, standard deviation, quartiles
- Inferential or Inductive Statistics
  - **Goal: to summarize a sample of the data to infer or draw conclusions about the population from which the sample is drawn**
    - Hypothesis testing
      - A/B testing
      - p-value
      - t-tests, χ^2-tests, F-tests
    - Confidence intervals

Hypothesis Testing

Hypothesis testing tell us *how extreme* the observed result is compared to what random chance might produce, and it helps us make decisions on that basis.

Lecture Outline

- **Hypothesis Testing**
  - **Testing Procedure**
  - **Null Hypothesis & Alternative Hypothesis**
  - **p-value and Degrees of Freedom (df)**
  - **Type I and Type II Errors**
- **Examplar Tests: Parametric & Nonparametric**
  - **T-Tests and Wilcoxon Tests**
  - **Single Sample: T-Test**
  - **Two-Sample: Independent Groups**
  - **Paired Two-Sample: Dependent Groups**
  - **Multiple Samples: Independent Groups**

Python: Statistical Distributions & Functions

Distribution	Random Variable Sample	Density	Probability
Normal	scipy.stats.norm.rvs()	scipy.stats.norm.pdf()	scipy.stats.norm.cdf()
t	scipy.stats.t.rvs()	scipy.stats.t.pdf()	scipy.stats.t.cdf()
F	scipy.stats.f.rvs()	scipy.stats.f.pdf()	scipy.stats.f.cdf()
χ^2	scipy.stats.chi2.rvs()	scipy.stats.chi2.pdf()	scipy.stats.chi2.cdf()

distribution_abbreviation.{rvs/pdf/cdf} ()

- **rvs** = random variable (RV) sample generation
- **pdf** = probability density function of a given RV
- **cdf** = cumulative probability distribution function of a given RV

- scipy.stats.norm.**cdf(a)** $\equiv P(X \leq a)$: probability that a or smaller number occurs in the normal distribution
- scipy.stats.norm.**cdf(b)** – scipy.stats.norm.**cdf(a)** $\equiv P(a \leq X \leq b)$: probability that the variable falls between two values in the normal distribution

R: Statistical Distributions & Functions

Distribution	Random Number Generator	Density	Distribution	Quantile
Normal	rnorm	dnorm	pnorm	qnorm
t	rt	dt	pt	qt
F	rf	df	pf	qf
χ^2	rchisq	dchisq	pchisq	qchisq

{dpqr}*distribution_abbreviation*()

- **d** = density
- **p** = distribution function
- **q** = quantile function
- **r** = random generation

- **pnorm(a)** $\equiv \boldsymbol{P(X \leq a)}$: probability that a or smaller number occurs
- **pnorm(b) – pnorm(a)** $\equiv \boldsymbol{P(a \leq X \leq b)}$: probability that the variable falls between two points
- qnorm(): given the cumulative probability distribution, it returns the quantile

Statistical Distributions: Mean & Variance

Distribution	Degrees of freedom	Mean	Variance	Comments
Normal		μ	σ^2	
t	n	0	$n/(n-2)$	
F	n_1 and n_2	$n_2/(n_2-2)$	a/b	$a = 2n_2^2(n_1+n_2-2)$ $b = n_1(n_2-2)^2\,(n_2-4)$
χ^2	r	r	$2r$	

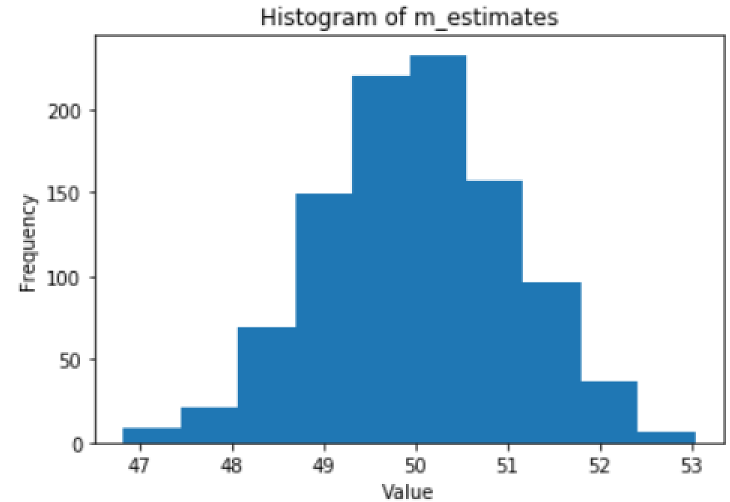
Sample Mean vs. Population Mean

Population Parameters: mu and sd

```
m_estimates = [stats.norm.rvs(loc=mean,
                              scale=sd,
                              size=sample_size).mean()
               for _ in range(n_samples)]

plt.hist(m_estimates)
plt.title("Histogram of m_estimates")
plt.xlabel("Value")
plt.ylabel("Frequency")
#plt.gcf()
plt.show()
```

data_prep_sampling.ipynb

Sample Statistic




```
print ("Mean of sample means: ", np.array(m_estimates).mean())
print ("Standard Error: ", np.array(m_estimates).var())
```

```
Mean of sample means:  50.0111166372
Standard Error:  0.956456485027
```

How sample statistic approximates population parameters for different sample sizes, n?

Ex: Sample from Unit Normal Distribution

$$N(\mu = 0, \sigma = 1)$$

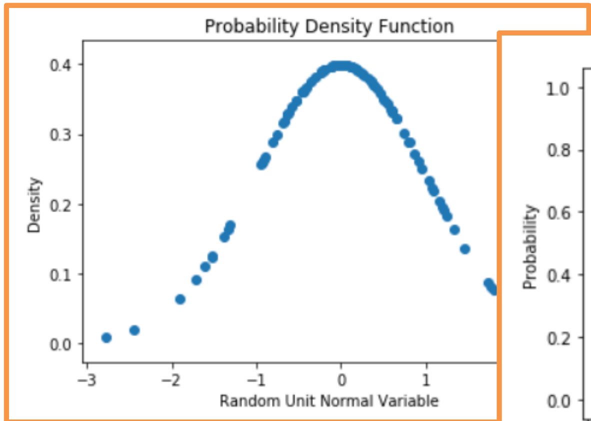
scipy.stats.norm.pdf()

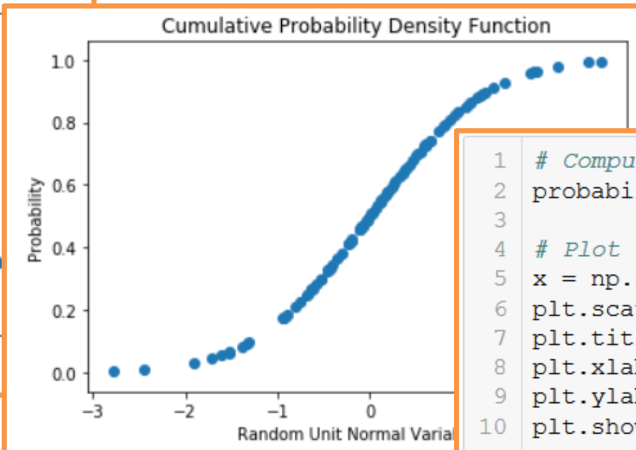
```
# Calculate and plot their probability density functions
densityRandUnitNormal = stats.norm.pdf(randUnitNormal)

x = np.linspace(norm.ppf(0.01), stats.norm.ppf(0.99), 100)
plt.scatter(randUnitNormal, densityRandUnitNormal)
plt.title("Probability Density Function")
plt.xlabel("Random Unit Normal Variable")
plt.ylabel("Density")
plt.show()
```

scipy.stats.norm.rvs()

```
# Generate 1000 points drawn
# from the unit normal distribution: N( 1.0, 0.0)
mean = 0.0
sd = 1.0
randUnitNormal = scipy.stats.norm.rvs(loc=mean,
                                      scale=sd,
                                      size=100)
randUnitNormal[0:3]
```

```
array([-0.96907238, -0.24168709,  2.19288252])
```





scipy.stats.norm.cdf()

```
# Compute and plot cumulative probability distribution
probabilityRandUnitNormal = stats.norm.cdf(randUnitNormal)

# Plot the distribution
x = np.linspace(norm.ppf(0.01), stats.norm.ppf(0.99), 1000)
plt.scatter(randUnitNormal, probabilityRandUnitNormal)
plt.title("Cumulative Probability Density Function")
plt.xlabel("Random Unit Normal Variable")
plt.ylabel("Probability")
plt.show()
```

Statistic & its Proxy: Hypothesis Testing

Aim	Model Statistic	Sample Statistic	Proxy Statistic	Formula for Proxy
Estimate the mean μ of a normal distribution with known variance σ^2	μ	m	Z-statistic	$Z \sim \frac{m-\mu}{\sigma/\sqrt{n}}$
Estimate the variance σ^2 of a normal distribution with known mean μ	σ^2	S^2	χ^2-statistic	$\chi^2_{n-1} \sim (n-1)\frac{S^2}{\sigma^2}$
Estimate the mean μ of a normal distribution with un-known variance σ^2	μ	m	t-statistic	$T_{n-1} \sim \frac{m-\mu}{S/\sqrt{n}}$

Ex.	Proxy Statistic	Distribution	Degrees of Freedom (df)
1	Z-statistic	$N(0,1)$	
2	χ^2-statistic	$\chi^2(n-1)$	$n-1$
3	t-statistic	T_{n-1}	$n-1$

Hypothesis Testing: Procedure

- Step 1: Define **a statistic** that obeys a certain **distribution** if the hypothesis is correct:
  - Ex-1: The mean μ from a normal distribution with known variance σ^2
  - Ex-2: The variance σ^2 from a normal distribution with known mean μ
  - Ex-3: The mean μ from a normal distribution with unknown variance σ^2

- Step 2 (optional): Transform the statistic to a **proxy statistic** with the **proxy distribution** of better understood properties/characteristics:
  - Ex-1: Z-statistic from a uniform normal distribution, $N(0,1)$
  - Ex-2: χ_{n-1}-statistic from a χ^2 distribution with n df
  - Ex-3: T_{n-1}-statistic from a t-distribution with $n-1$ df

- Step 3: Calculate the statistic (original/proxy) from the **sample**

- Step 4: Compute the **probability** (the $\boldsymbol{p}$**-value**) of this sample with this statistic to be drawn from this distribution (original/proxy)
  - **Reject the hypothesis** if probability is **low** (e.g., $\boldsymbol{p}$**-value < 0.05**)
  - **Fail to reject the hypothesis** otherwise (e.g., $\boldsymbol{p}$**-value** $\geq$ **0.05**)

Important Note

DO NOT SAY: We ACCEPT the Hypothesis

INSTEAD: We FAIL TO REJECT the Hypothesis

- **Given the sample we had to calculate the statistic**

Null Hypothesis vs. Alternative Hypothesis

- **Null Hypothesis** (H_0): **what is considered to be true**:
  - **Example:** $H_0 : \mu = \mu_0$: We want to test a hypothesis that the unknown mean μ for a sample from a normal distribution with known variance σ^2 is equal to a specific constant μ_0

- **Alternative Hypothesis** (H_1): **If the null hypothesis is rejected:**
  - **Example:** $H_1 : \mu \neq \mu_0$

Examples: Null and Alternative Hypotheses

- **Null = "no difference between the means of group A and group B"**
  - **Alternative = "A is different from B" (could be bigger or smaller)**
- **Null = "A $\leq$ B"**
  - **Alternative = "B > A"**
- **Null = "B is not $x\%$ greater than A"**
  - **Alternative = "B is $x\%$ greater than A"**

- The Null and Alternative Hypotheses must account for all possibilities.
- The nature of the null hypothesis determines the structure of the hypothesis test.

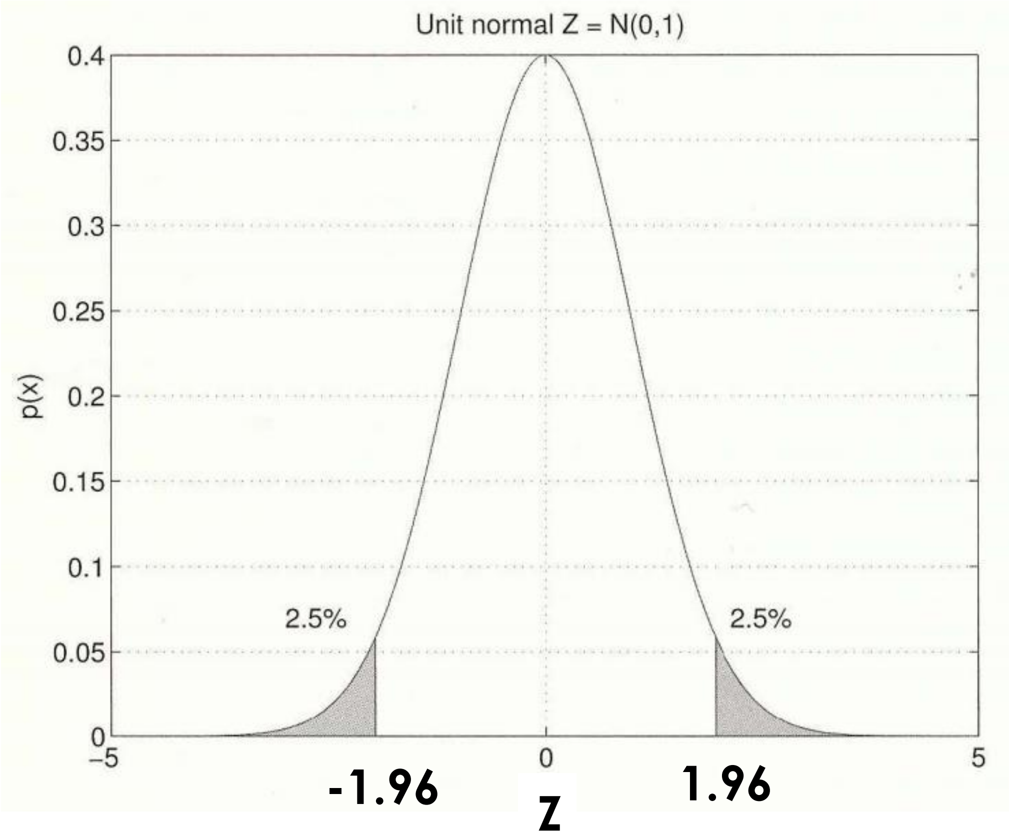
H1 Hypothesis: One-Way or Two-Way Tests?

- A directional alternative hypothesis:
  - B is better than A
  - Use a one-way or one-tail hypothesis test
    - An extreme chance results in only one direction count toward the p-value
  - A one-tail hypothesis test often fits the nature of A/B decision making
    - decision is required and one option is assigned “default” unless the other proves better
- A bi-directional alternative hypothesis
  - A is different from B (could be bigger or smaller)
  - Use a two-way or two-tail hypothesis
    - Extreme chance results in either direction count toward the the p-value
  - R software typically provides a two-tail test in default output

Summary: Key Ideas

- A null hypothesis embodies the notion that nothing special has happened
  - any effect you observe is due to random chance
- The hypothesis test assumes the null hypothesis is true, creates a "null model" (a probability model), and tests whether the effect you observe is a reasonable outcome of that model

Two-sided Confidence Interval for $Z \sim N(0, 1)$



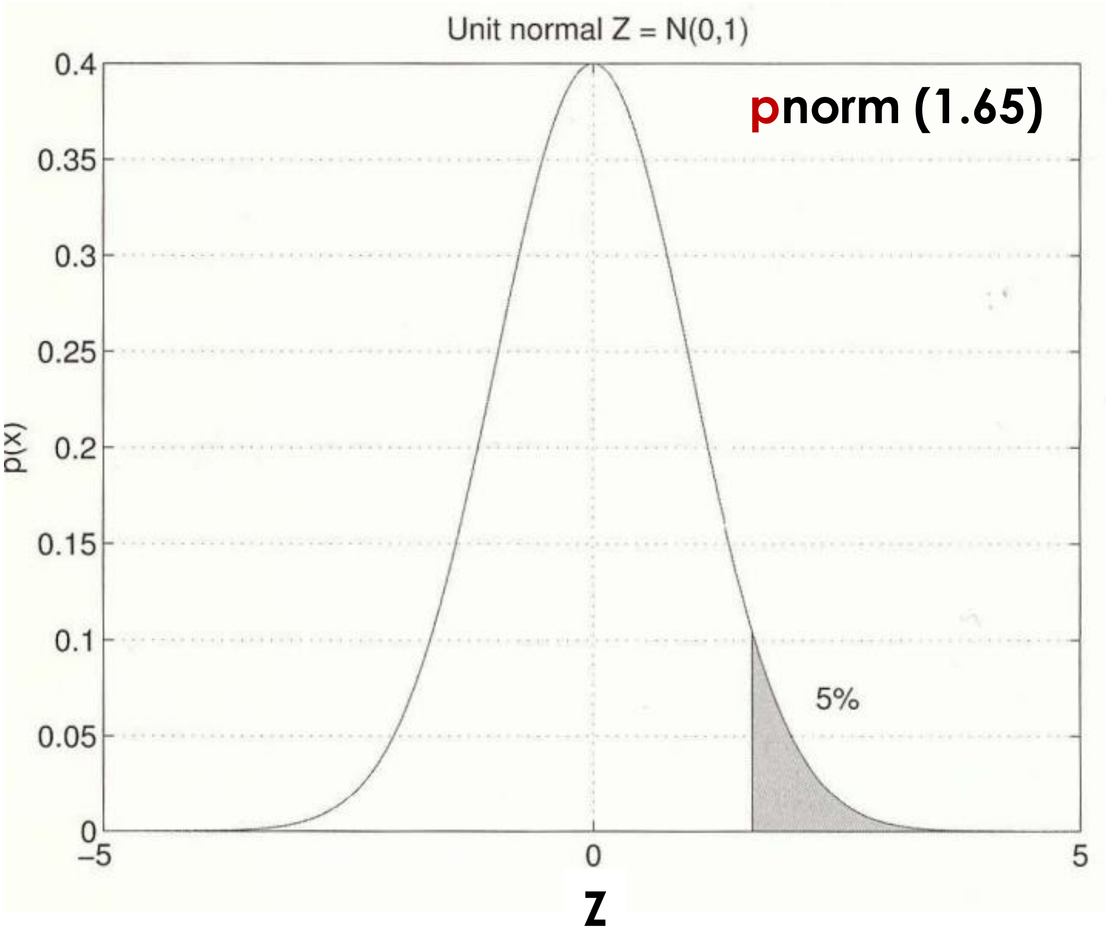
95% of the unit normal distribution lies between - 1.96 and 1.96

$$P\{\,|Z - 0| < 1.96\,\} = 0.95$$

pnorm (1.96) – pnorm (-1.96)

What is (1 - pnorm(1.96))?

One-sided Confidence Interval for $Z \sim N(0,1)$



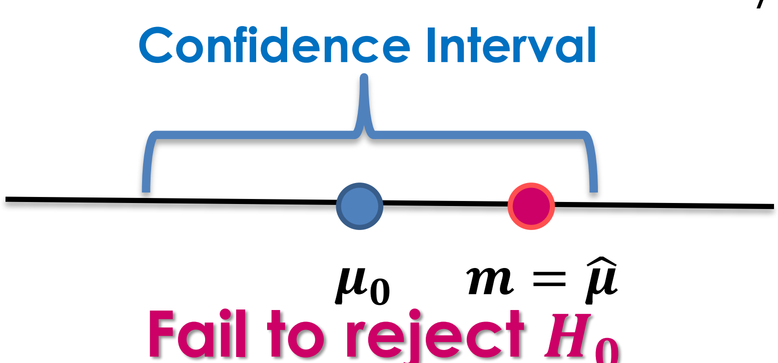
95% of the unit normal distribution lies below 1.64

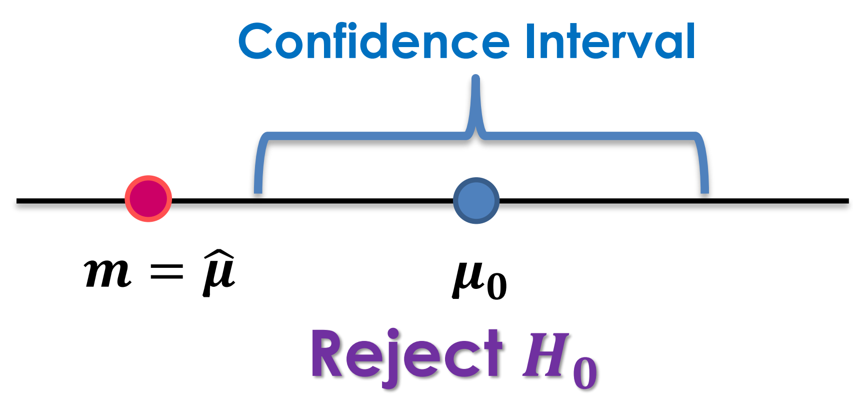
$$P\{\, Z < 1.64 \,\} = 0.95$$

Significance Level: Two-sided Test

- Null Hypothesis: $H_0: \ \mu = \mu_0$
- Alternative Hypothesis: $H_1: \ \mu \neq \mu_0$

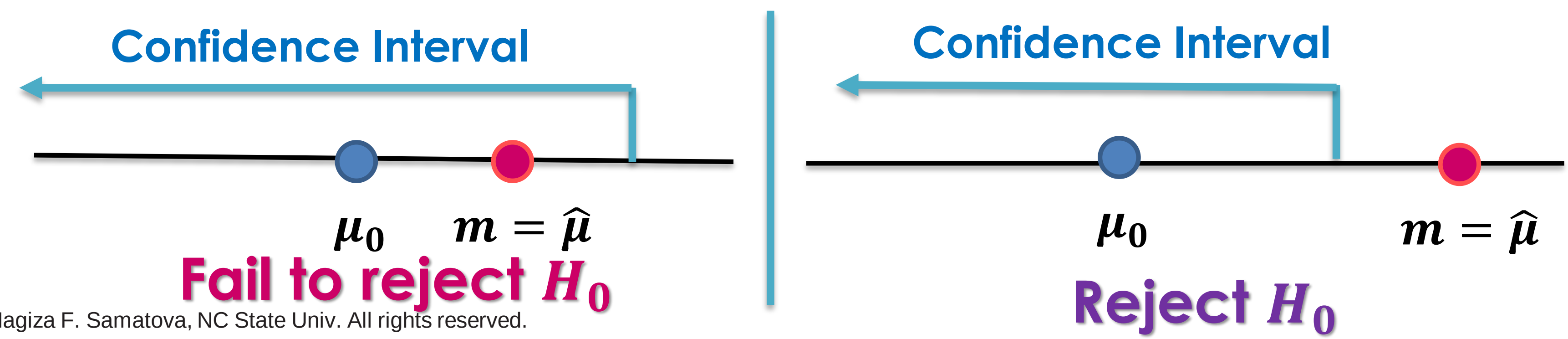
- **Significance Level** (α): We fail to reject the null hypothesis with *level of significance* α if the estimate of the sample statistic lies within the $100(1-\alpha)$ *percent two-sided confidence interval* (CI) for the hypothesized value of the statistic:
  - m is the point estimate of μ :
    - We **fail to reject H_0 if m is close to μ_0**, i.e., within the confidence interval, namely, if $Z \sim \frac{m-\mu_0}{\sigma/\sqrt{n}} \in (-z_{\alpha/2}, z_{\alpha/2})$
    - We **reject H_0 if m is too far from μ_0**, i.e., outside the confidence interval, namely, if $Z \sim \frac{m-\mu_0}{\sigma/\sqrt{n}} \notin (-z_{\alpha/2}, z_{\alpha/2})$





Significance Level: One-sided Test

- **Null Hypothesis:** $H_0 : \mu \le \mu_0$
- **Alternative Hypothesis:** $H_1 : \mu > \mu_0$
- **Significance Level** (α)**: We fail to reject the null hypothesis with *level of significance* α if the estimate of the sample statistic lies within the $100(1-\alpha)$ *percent one-sided confidence interval* (CI) for the hypothesized value of the statistic:**
  - m is the point estimate of μ :
    - We **fail to reject H_0 if m is close to μ_0**, i.e., within the confidence interval, namely, if $Z \sim \frac{m-\mu_0}{\sigma/\sqrt{n}} \in (-\infty, z_\alpha)$
    - We **reject H_0 if m is too far from μ_0**, i.e., outside the confidence interval, namely, if $Z \sim \frac{m-\mu_0}{\sigma/\sqrt{n}} \notin (-\infty, z_\alpha)$



Exercise: Test the null hypothesis

- **Null Hypothesis** $(\boldsymbol{H_0})$**:** **what is considered to be true**:
  - $\boldsymbol{H_0}$ **:** $\boldsymbol{\mu = \mu_0}$ **:** We want to test a hypothesis that the **unknown** mean μ for a sample from a normal distribution with **unknown** variance σ^2 is equal to a specific constant μ_0
- Hint: Use $\boldsymbol{t}$**-statistic** rather than $\boldsymbol{Z}$**-statistic** from the previous examples

Solution: Test the null hypothesis

- **Null Hypothesis** (H_0): **what is considered to be true**:
  - $H_0 : \ \mu = \mu_0$: We want to test a hypothesis that the **unknown** mean μ for a sample from a normal distribution with **unknown** variance σ^2 is equal to a specific constant μ_0

Use t**-statistic**: $T_{n-1} \sim \dfrac{m-\mu}{S/\sqrt{n}}$

Two-sided Test:

- We **fail to reject H_0 at significance level α** if
$$T_{n-1} \sim \frac{m-\mu_0}{S/\sqrt{n}} \in (-t_{\alpha/2,n-1}, t_{\alpha/2,n-1})$$

- We **reject H_0 at significance level α** if
$$T_{n-1} \sim \frac{m-\mu_0}{S/\sqrt{n}} \notin (-t_{\alpha/2,n-1}, t_{\alpha/2,n-1})$$

R Example: T-Test Hypothesis Testing

Null Hypothesis (H_0): The average tip is equal to $2.50

```
> data(tips, package = "reshape2")
> head (tips)
  total_bill  tip    sex smoker day   time size
1      16.99 1.01 Female     No Sun Dinner    2
2      10.34 1.66   Male     No Sun Dinner    3
3      21.01 3.50   Male     No Sun Dinner    3
4      23.68 3.31   Male     No Sun Dinner    2
5      24.59 3.61 Female     No Sun Dinner    4
6      25.29 4.71   Male     No Sun Dinner    4
> unique (tips$sex)
[1] Female Male
Levels: Female Male
> unique (tips$day)
[1] Sun  Sat  Thur Fri
Levels: Fri Sat Sun Thur
```

- **Assumption: Variance is unknown**
- **Use t-statistic**

File: HypothesisTesting.R

Python Example: T-Test Hypothesis Testing

Null Hypothesis (H_0): The average tip is equal to $2.50

```
tips = pd.read_csv("../data_raw/ht_tips.csv")
```

```
tips.head()
```

	total_bill	tip	sex	smoker	day	time	size
0	16.99	1.01	Female	No	Sun	Dinner	2
1	10.34	1.66	Male	No	Sun	Dinner	3
2	21.01	3.50	Male	No	Sun	Dinner	3
3	23.68	3.31	Male	No	Sun	Dinner	2
4	24.59	3.61	Female	No	Sun	Dinner	4

```
tips.mean (axis=0, numeric_only=True)
```

```
total_bill    19.785943
tip            2.998279
size           2.569672
dtype: float64
```

```
tips["day"].describe()
```

```
count      244
unique       4
top        Sat
freq        87
Name: day, dtype: object
```

```
tips["sex"].describe()
```

```
count       244
unique        2
top        Male
freq        157
```

File: ht_hypothesis_testing.ipynb

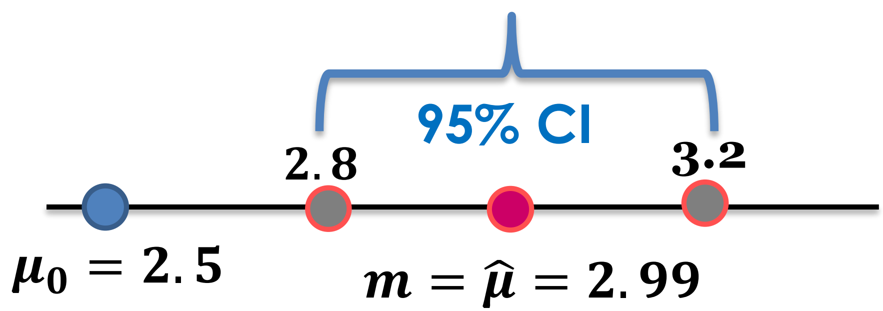
R: One-Sample T-Test (cont.)

Null Hypothesis (H_0): The average tip is equal to $2.50

```
> t.test(tips$tip, alternative="two.sided", mu=2.5)

 One Sample t-test

data:  tips$tip
t = 5.6253, df = 243, p-value = 5.08e-08
alternative hypothesis: true mean is not equal to 2.5
95 percent confidence interval:
 2.823799 3.172758
sample estimates:
mean of x
 2.998279
```

Reject Null Hypothesis

- **The p-value (less than 0.05) indicates the null hypothesis should be rejected**



Reject H_0

Conclusion: The mean is not equal to $2.50

Python: One-Sample T-Test (cont.)

Null Hypothesis (H_0): The average tip is equal to $2.50

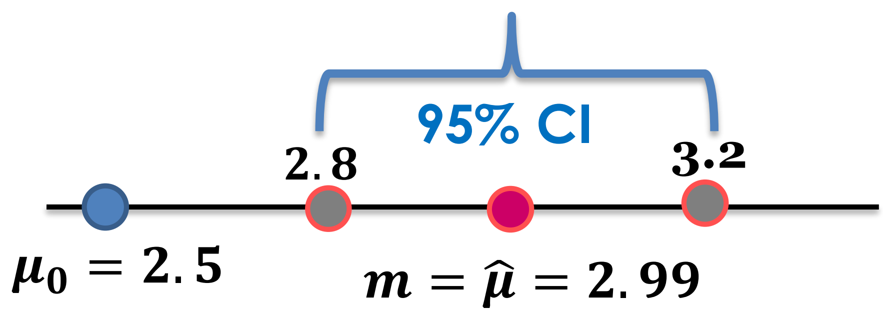
```python
from scipy.stats import stats
```

```python
# two-sided test for the null hypothesis that the expected value
# (mean) of a sample of independent observations `a` is equal to the given
# population mean, `popmean`
stats.ttest_1samp (tips["tip"], popmean = 2.5)
```

```
Ttest_1sampResult(statistic=5.625287009994555, pvalue=5.07998845968649e-08)
```

Reject Null Hypothesis

- **The p-value (less than 0.05) indicates the null hypothesis should be rejected**

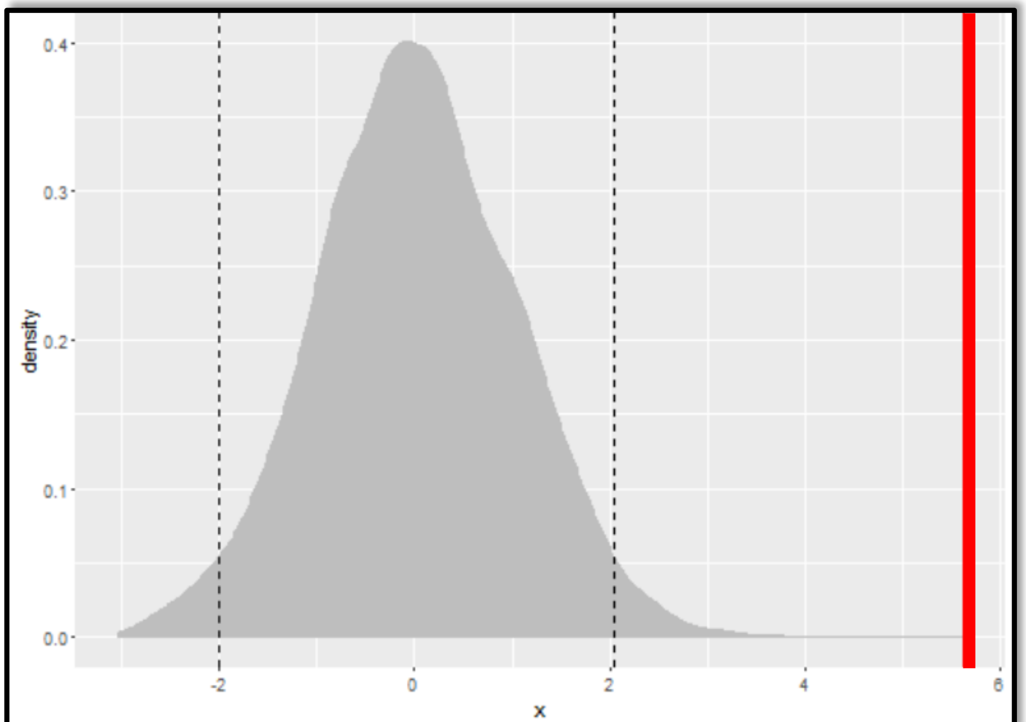


Reject H_0

Conclusion: The mean is not equal to $2.50

R: Examine *t*-statistic & its probability

```
randT <- rt(3000, df=NROW(tips)-1)
tipTTest <- t.test(tips$tip,
                   alternative="two.sided",
                   mu=2.5)
require (ggplot2)
ggplot(data.frame(x=randT)) +
  geom_density(aes(x=x), fill="grey", color="grey") +
  geom_vline(xintercept=tipTTest$statistic, color="red") +
  geom_vline(xintercept=mean(randT) +
             c(-2,2)*sd(randT), linetype=2)
```



Probability of t-statistic

p-value = 5.08e-08

t-statistic = 5.62

t-distribution and t-statistic for tip data:

- *dashed lines are two sd's from the mean in either direction*
- *thick red line (t-statistic) is far outside the distribution → reject null hypothesis → true mean is not equal to $2.50*

Python: Visualizing *t*-statistic & its probability

```python
tipTTest = stats.ttest_1samp (tips["tip"], popmean = 2.5)
tipTTest.statistic
```

5.625287009994555

```python
# draw a sample of size=3000 from t-distribution
# with df degrees of freedom
r = t.rvs(df=df, size = 3000)
```

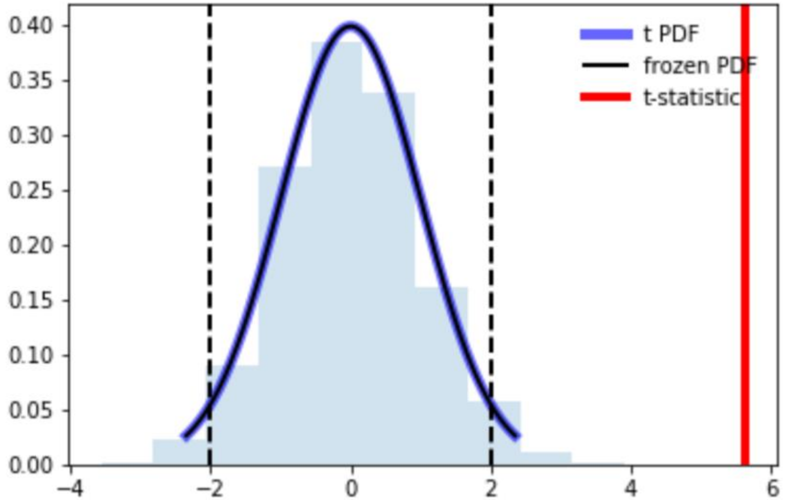
```python
# df: degrees of freedom
df = tips.shape[0]-1
df
```

243

t-statistic = 5.62 **p-value = 5.08e-08**

***t*-distribution and *t*-statistic for tip data**:

- *dashed lines are two sd's from the mean in either direction*
- *thick red line (t-statistic) is far outside the distribution → reject null hypothesis → true mean is not equal to $2.50*




```python
fig, ax = plt.subplots(1, 1)
x = np.linspace(t.ppf(0.01, df), t.ppf(0.99, df), 100)
ax.plot(x, t.pdf(x, df),'b-', lw=5, alpha=0.6, label='t PDF')
rv = t(df)
ax.plot(x, rv.pdf(x), 'k-', lw=2, label='frozen PDF')
ax.hist(r, normed=True, histtype='stepfilled', alpha=0.2)
plt.axvline(x=tipTTest.statistic, color='r',
            linestyle='solid', lw=4, label='t-statistic')
plt.axvline(x=-2*r_sd, color='k', linestyle='--', lw=2)
plt.axvline(x=2*r_sd, color='k', linestyle='--', lw=2)
ax.legend(loc='best', frameon=False)
plt.show()
```

R: What about one-sided T-Test?

Null Hypothesis (H_0): The average tip is less than $2.50

```
> t.test(tips$tip, alternative="greater", mu=2.5)

 One Sample t-test

data:  tips$tip
t = 5.6253, df = 243, p-value = 2.54e-08
alternative hypothesis: true mean is greater than 2.5
95 percent confidence interval:
 2.852023      Inf
sample estimates:
mean of x
 2.998279
```

Reject Null Hypothesis

- **The p-value (less than 0.05) indicates the null hypothesis should be rejected**

Conclusion: The mean is greater than $2.50

Python: What about one-sided T-Test?

Null Hypothesis (H_0): The average tip is less than $2.50

```
# alpha = 0.5: 95% confidence interval
def one_tailed_t_test(data, mu, alternative="less than", alpha=0.05):
    # one-tailed, less than 130

    results = stats.ttest_1samp(data, mu)
    print ("t = {0:5.4f}  df = {1}  p-value = {2:10.9f}".
           format(results[0], str(data.shape[0]-1), results[1]))

    if re.search("less than", alternative):
        if (results[0] < 0) & (results[1]/2 < alpha):
            print ("reject null hypothesis, mean is less than {}".format(mu))
        else:
            print ("fail to reject null hypothesis, mean is greater than {}".format(mu))
            print ("sample estimate: \n\t mean of data: {}".format(data.mean(axis=0)))
    elif re.search ("greater", alternative):
        if (results[0] > 0) & (results[1]/2 < alpha):
            print ("reject null hypothesis, mean is greater than {}".format(mu) )
        else:
            print("fail to reject null hypothesis, mean is less than {}".format(mu))
            print ("sample estimate: \n\t mean of data: {}".format(data.mean(axis=0)))
    else:
        print ("invalid argument for alternative: {}".format(alternative))
```

- The p-value (less than 0.05) indicates the null hypothesis should be rejected

```
one_tailed_t_test(data=tips["tip"], mu=2.5,
                  alternative="less than")
```

```
t = 5.6253  df = 243  p-value = 0.000000051
fail to reject null hypothesis, mean is greater than 2.5
sample estimate:
         mean of data: 2.9982786885245902
```

Reject Null Hypothesis

Conclusion: The mean is greater than $2.50

Comments on p-value & degrees of freedom

- **p-value**: The probability, if the null hypothesis were correct, of getting as extreme, or more extreme, a result for the tested statistic (e.g., the estimated mean):
  - It is a measure of how extreme the statistic is
  - If the statistic is too extreme, we conclude that H_0 should be rejected
  - Typical p-value to reject H_0: 0.10, 0.05 or 0.01 to be too extreme

- **Degrees of freedom (df)**: Represents the effective number of observations:
  - Usually, df is the number of observations minus the number of parameters being estimated

P-values: Six principles

- *P*-values can indicate how incompatible the data are with a specified statistical model
- *P*-values do *not* measure the probability that the studied hypothesis is true, or probability that the data were produced by random chance alone
- Scientific conclusions and business or policy decisions should not be based only on whether *p*-value passes a specific threshold
- Proper inference requires full reporting and transparency
- *P*-value, or statistical significance, does *not* measure the size of an effect or the importance of a result
- By itself, a *p*-value does *not* provide a good measure of evidence regarding a model of hypothesis

Guidelines by the American Statistical Association (ASA)

Type I and Type II Errors, Power Function

	Decision	
Truth	Fail to reject H_0	Reject H_0
True	Correct	Type I Error
False	Type II Error	Correct (Power)

- **Type I Error: Reject the null hypothesis H_0, when H_0 is correct**
  - The significance level α set before the test defines how much Type I Error we can tolerate
    - Typical values for $\alpha = 0.1, 0.05, 0.01$
- **Type II Error: Fail to reject the null hypothesis H_0, when H_0 is false**
  - Fail to reject the null hypothesis when the true mean μ is unequal to μ_0.
  - The probability that H_0 is not rejected when the true mean is μ is a function of μ: $$\beta(\mu) = P_\mu\{-z_{\alpha/2} \leq \frac{m - \mu_0}{\sigma/\sqrt{n}} \leq z_{\alpha/2}\}$$
- **Power function of the test ($1 - \beta(\mu)$)**: The probability of rejection when μ is the true value
  - Type II error probability increases as μ and μ_0 get closer

Comparing Two Groups of Observations

- **Parametric vs. Nonparametric**
  - Parametric tests are more powerful if the underlying assumptions hold true → Always try parametric tests first
  - Nonparametric tests are more appropriate when the assumptions are grossly unreasonable (e.g., rank ordered data)

- **Dependent vs. Independent** Groups
  - **Paired Tests** (**paired = TRUE**) for ***dependent*** groups

Examples: Hypothesis Tests

Sample	Paired	Null Hypothesis	Assumptions	R Test
One Sample		$H_0 : \mu = \mu_0$	i.i.d. $N(\mu, \sigma^2)$	t.test()
Two Samples	No	$H_0 : \sigma_1^2 = \sigma_2^2$	Normally distributed	F-test: var.test() Bartlett: bartlett.test()
Two Samples	No	$H_0 : \sigma_1^2 = \sigma_2^2$	Non-parametric	Ansari-Bradley test: ansari.test()
Two Samples	No	$H_0 : \mu_1 = \mu_2$	$\sigma_1^2 = \sigma_2^2$	t.test(**var.equal=TRUE**)
Two Samples	No	$H_0 : \mu_1 = \mu_2$	$\sigma_1^2 \neq \sigma_2^2$	Welch t-test t.test(var.equal=FALSE)
Two samples	No	$p_1(x) = p_2(x)$ p: probab. distr	Non-parametric	Wilcoxon rank sum wilcox.test ()
Two Samples	Yes	$H_0 : \mu_1 = \mu_2$	$\sigma_1^2 \neq \sigma_2^2$	t.test(**paired=TRUE**)
Two samples	Yes	$p_1(x) = p_2(x)$ p: probab. distr	Non-parametric	wilcox.test (**paired=TRUE**)

Non-parametric Test of Equal Variance

$$\boxed{H_0 : \ \sigma_1^2 = \sigma_2^2}$$

- Input: Two ***independent*** samples (i.e., two groups of observations)
- Null Hypothesis: The variances of two populations are equal

- Assumption: The data does not appear to be normally distributed
  - Hence, parametric tests can not be applied:
    - Neither F-test (var.test) nor Bartlett test can be applied

- **Ansari-Bradley Test**: **ansari.test**()
  - Non-parametric (no assumptions about population distribution)
  - Fail to reject the null hypothesis if the p-value is large, i.e.,
    - in this case, we conclude that the test indicates that the variances are equal

Ex: Ansari-Bradley Test: Equality of Variances

H_0: The variances in tips between female and male groups are equal

R code

Quick look into variances

```
> aggregate (tip ~ sex, data = tips, var)
     sex      tip
1 Female 1.344428
2   Male 2.217424
```

Python code

```
female_tips = tips[tips["sex"]=="Female"]
male_tips = tips[tips["sex"]=="Male"]
print ("Female Tip Variance: {0:4.3}".format(female_tips["tip"].var(axis=0)))
print ("Male Tip Variance: {0:4.3f}".format(male_tips["tip"].var(axis=0)))
```

```
Female Tip Variance: 1.34
Male Tip Variance: 2.217
```

R: Ansari-Bradley Test: Equality of Variances

H_0: The variances in tips between female and male groups are equal

```
> shapiro.test(tips$tip[tips$sex == "Female"])

 Shapiro-Wilk normality test

data:  tips$tip[tips$sex == "Female"]
W = 0.9568, p-value = 0.005448

> shapiro.test(tips$tip[tips$sex == "Male"])

 Shapiro-Wilk normality test

data:  tips$tip[tips$sex == "Male"]
W = 0.8759, p-value = 3.708e-10
```

Check the assumptions:
test for normality of tip distributions

- p-value < 0.05: **the null hypothesis should be rejected**
- **Conclusion: groups are not normally distributed**

```
> ansari.test(tip ~ sex, tips)

 Ansari-Bradley test

data:  tip by sex
AB = 5582.5, p-value = 0.376
alternative hypothesis: true ratio of scales
is not equal to 1
```

Assumption appears to be correct:
apply a non-parametric test

- p-value > 0.05: **fail to reject the null hypothesis**
- **According to this test, the results were not significant;**
- **Conclusion: the variances are equal**

Python: Ansari-Bradley Test: Equality of Variances

H_0: The variances in tips between female and male groups are equal

```
from scipy.stats import shapiro
```

```
shapiro(female_tips["tip"])
```

```
(0.9567776918411255, 0.005448382347822189)
```

```
shapiro(male_tips["tip"])
```

```
(0.875868916515356, 3.708431339788376e-10)
```

Check the assumptions: test for normality of tip distributions

- **p-value < 0.05: the null hypothesis should be rejected**
- **Conclusion: groups are not normally distributed; hence we need to apply a non-parametric test**

```
from scipy.stats import ansari
```

```
ansari(female_tips["tip"], male_tips["tip"])
```

```
AnsariResult(statistic=5582.5, pvalue=0.3760472514100246)
```

Assumption appears to be correct: apply a non-parametric test

- **p-value > 0.05: fail to reject the null hypothesis**
- **According to this test, the results were not significant;**
- **Conclusion: the variances are equal**

R: Two-Sample T-Test: Equality of Means

H_0: Female and male groups are, on average, tipped equally

- Based on the Ansari-Bradley test, the variances in tips between two groups are equal
- Hence, a standard two sample t-test can be used rather than the Welch test for unequal variances

Check the assumptions: test for equal variances

Assumption appears to be correct: apply a standard two sample t-test

```
> t.test (tip ~ sex, data = tips, var.equal=TRUE)

 Two Sample t-test

data:  tip by sex
t = -1.3879, df = 242, p-value = 0.1665
alternative hypothesis: true difference in means
is not equal to 0
95 percent confidence interval:
 -0.6197558  0.1074167
sample estimates:
mean in group Female   mean in group Male
            2.833448             3.089618
```

- **p-value > 0.05: fail to reject the null hypothesis**
- **According to this test, the results were not significant;**
- **Conclusion: female and male workers are tipped roughly equally**

Python: Two-Sample T-Test: Equality of Means

H_0: **Female and male groups are, on average, tipped equally**

- Based on the Ansari-Bradley test, **the variances in tips between two groups are equal**
- Hence, **a standard two sample t-test can be used rather than the Welch test for unequal variances**

Check the assumptions: test for equal variances

Assumption appears to be correct: apply a standard two sample t-test

```
from scipy.stats import ttest_ind
```

```
t, p = ttest_ind(female_tips["tip"], male_tips["tip"],
                 equal_var=True)
print("ttest_ind: t = %g p = %g" % (t, p))
```

```
ttest_ind: t = -1.38786 p = 0.166456
```

- **p-value > 0.05: fail to reject the null hypothesis**
- **According to this test, the results were not significant;**
- **Conclusion: female and male workers are tipped roughly equally**

R: Paired Two-Sample T-Test: Dependent Groups

H_0: **Fathers and sons have equal heights, on average**

Check the assumptions:
- *test for normal distribution*
- *test for equal variances*

```
install.packages("UsingR")
require(UsingR)
head(father.son)
```

```
> t.test(father.son$fheight, father.son$sheight, paired=TRUE)

 Paired t-test

data:  father.son$fheight and father.son$sheight
t = -11.7885, df = 1077, p-value < 2.2e-16
alternative hypothesis: true difference in means is not equal
to 0
95 percent confidence interval:
 -1.1629160 -0.8310296
sample estimates:
mean of the differences
             -0.9969728
```

Reject H_0

- **p-value < 0.05: the null hypothesis should be rejected**
- **Conclusion: fathers and sons (at least for this data set) have different heights**

Python: Paired Two-Sample T-Test: Dependent Groups

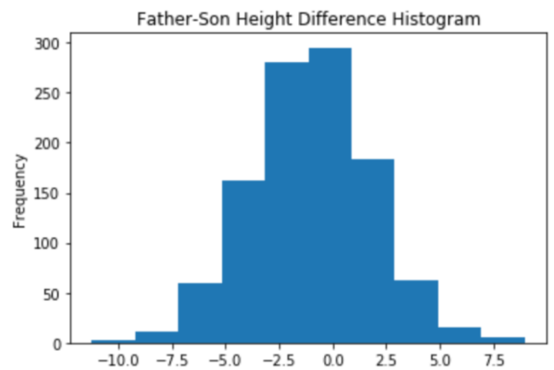
H_0: Fathers and sons have equal heights, on average

```
father_son = pd.read_csv("../data_raw/ht_father_son.csv")
```

```
father_son.head()
```

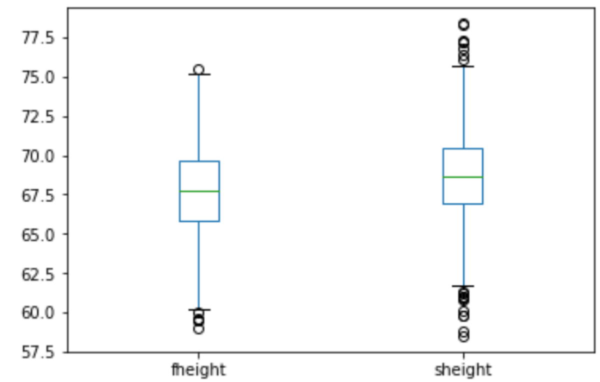
	fheight	sheight
1	65.04851	59.77827
2	63.25094	63.21404
3	64.95532	63.34242
4	65.75250	62.79238
5	61.13723	64.28113

```
# Test the difference for normality
father_son['difference'] = father_son['fheight'] - father_son['sheight']
father_son['difference'].plot(kind='hist',
          title= 'Father-Son Height Difference Histogram')
shapiro(father_son['difference'])
plt.show()
```



Check the assumptions:
- *test for normal distribution*
- *test for equal variances*

```
import matplotlib.pyplot as plt
father_son.plot(kind='box')
plt.show()
```




```
from scipy.stats import ttest_rel
```

```
t, p = ttest_rel(father_son["fheight"], father_son["sheight"])
print("ttest_ind: t = %g p = %g" % (t, p))
```

```
ttest_ind: t = -11.7885 p = 2.95723e-30
```

Reject H_0

- **p-value < 0.05: the null hypothesis should be rejected**
- **Conclusion: fathers and sons (at least for this data set) have different heights**

Wilcoxon Rank Sum Test

Non-parametric comparison of two (in)dependent groups

H_0: **Both groups are sampled from the same probability distribution:**

$$p_1(x) = p_2(x)$$

- **Assumptions** for using Wilcoxon Rank Sum Test: wilcox.test():
  - Two groups are ***independent***
  - If two groups are ***dependent*** then use parameter **paired = TRUE**
  - Unable to meet the parametric assumptions of a t-test or ANOVA
  - Outcome variables are severely ***skewed*** or
  - Outcome variables are ***ordinal*** in nature (**rank ordered data**):
    - Probability of obtaining higher scores is greater in one population than the other

R: Wilcoxon Rank Sum Test

Non-parametric comparison of two independent groups

H_0: Incarceration rates are the same in Southern & non-Southern states

```
library(MASS)
head(UScrime)
# So: Southern vs non-Southern state
# Prob: Probability of incareceration
# (i.e., being imprisoned if committed a crime)
with (UScrime, by(Prob, So, median))

wilcox.test (Prob ~ So, data = UScrime)
```

```
 Wilcoxon rank sum test

data:  Prob by So
W = 81, p-value = 8.488e-05
alternative hypothesis: true location shift
is not equal to 0
```

Reject H_0

- p-value < 0.05: the null hypothesis should be rejected
- Conclusion: incarceration rates are not the same

Python: Wilcoxon Rank Sum Test

Non-parametric comparison of two independent groups

H_0: Blood pressure is the same between Before & After Intervention

```
bp = pd.read_csv("../data_raw/ht_synthetic_blood_pressure.csv")
bp.head(3)
```

```
from scipy.stats import wilcoxon
```

	patient	sex	agegrp	bp_before	bp_after
0	1	Male	30-45	143	153
1	2	Male	30-45	163	170
2	3	Male	30-45	153	168

```
# Calculate the differences between the two conditions
bp['difference'] = bp['bp_after'] - bp['bp_before']
```

```
# Method-1: Using the difference
wilcoxon (bp['difference'])
```

```
WilcoxonResult(statistic=2234.5, pvalue=0.0014107333565442858)
```

Reject H_0

```
# Method-2: Using both variables
wilcoxon (bp['bp_after'], bp['bp_before'])
```

```
WilcoxonResult(statistic=2234.5, pvalue=0.0014107333565442858)
```

- p-value < 0.05: the null hypothesis should be rejected
- Conclusion: The blood pressure before the intervention was higher (M= 156.45 ± 11.39 units) compared to the blood pressure post intervention (M= 151.36 ± 14.18 units); there was a statistically significant decrease in blood pressure (t=2,234.5, p= 0.0014).

R: Paired T-Test

Parametric comparison of two dependent groups

H_0: Unemployment rates are the same for younger and older males in Alabama

```
library(MASS)
head(UScrime)
sapply(UScrime[c("U1", "U2")],
       function(x) c(mean=mean(x), sd=sd(x)))
with (UScrime, t.test(U1, U2, paired = TRUE))
```

```
 Paired t-test

data:  U1 and U2
t = 32.4066, df = 46, p-value < 2.2e-16
alternative hypothesis: true difference in
means is not equal to 0
95 percent confidence interval:
 57.67003 65.30870
sample estimates:
mean of the differences
               61.48936
```

← Reject H_0

- the mean difference (61.5) is large to warrant rejection of H_0 that the mean unemployment rate for older and younger males is the same
- younger males have a higher rate
- probability of obtaining a sample difference that large if population means are equal is $2.2e^{-16}$

Python: Paired T-Test

Parametric comparison of two dependent groups

H_0: Unemployment rates are the same for younger and older males in Alabama

```
UScrime = pd.read_csv("../data_raw/ht_US_crime.csv")
UScrime.head(3)
```

	M	So	Ed	Po1	Po2	LF	M.F	Pop	NW	U1	U2	GDP	Ineq	Prob	Time	y
1	151	1	91	58	56	510	950	33	301	108	41	394	261	0.084602	26.2011	791
2	143	0	113	103	95	583	1012	13	102	96	36	557	194	0.029599	25.2999	1635
3	142	1	89	45	44	533	969	18	219	94	33	318	250	0.083401	24.3006	578

```
UScrime[['U1', 'U2']].describe()
```

	U1	U2
count	47.000000	47.000000
mean	95.468085	33.978723
std	18.028783	8.445450
min	70.000000	20.000000
25%	80.500000	27.500000
50%	92.000000	34.000000
75%	104.000000	38.500000
max	142.000000	58.000000

```
t, p = ttest_rel(UScrime["U1"], UScrime["U2"])
print("ttest_ind: t = %g p = %g" % (t, p))
```

```
ttest_ind: t = 32.4066 p = 2.53039e-33
```

Reject H_0

- the mean difference (61.5) is large to warrant rejection of H_0 that the mean unemployment rate for older and younger males is the same
- younger males have a higher rate
- probability of obtaining a sample difference that large if population means are equal is $2.2e^{-16}$

Comparing More than Two Groups

- **Parametric vs. Nonparametric**
  - Parametric tests: **ANOVA** ← **later as part of Experiment Design**
  - Nonparametric tests: **Kruskal Wallis or Friedman**

- **Dependent vs. Independent** Groups : **Nonparametric Tests**
  - **Independent Groups: Kruskal Wallis Test: kruskal.test()**
  - **Dependent Groups: Friedman Test: friedman.test()**

Hypothesis Testing

STATISTICAL SIMULATION

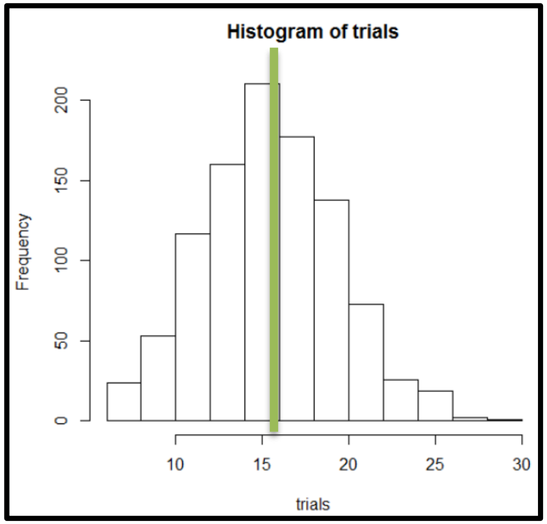
Null Hypothesis via Simulation

Null Hypothesis, Ho	Hat Model for Null Hypothesis	Quantitative Translation
Advertisements A and B are equally good (Goal: Reject Ho)	A single hat with 0s (no clicks) and 1s (clicks) where the 1s are the total clicks (A+B) and 0s are the total page-views	Ads A and B have the same click-through rate
A new targeted chemotherapy for advanced breast cancer is no more effective that the standard tamoxifen (Goal: Reject Ho)	A single hat for all patients (regardless of the group), with the number of days the person survived as values	Median survival time in a clinical trial is the same for both groups
A new & costly manufacturing process will not increase the chip-processing speed worth the investment (Goal: Reject Ho)	A single hat with all chips tested (both new and existing) in the sample, with the processing speed as values	Mean processing speed in a pilot new-process sample falls short of a 25% improvement

R Example: Hypothesis Testing via Simulation

Null Hypothesis: Reduction in the return rate to 8% have occurred by chance

- An online merchant has historically experienced 10% return rate in the “kitchen gadget” category.
- To increase returns, the merchant does a pilot in which it adds additional explanatory information and pictures about products to its website
- Out of the next 200 purchases, 16 (8%) are returned.
- Is the pilot effective?



- 16 or fewer returns are not unusual
- p-value = 0.564 → Fail to Reject Null Hypothesis

Python Example: Hypothesis Testing via Simulation

Null Hypothesis: Reduction in the return rate to 8% have occurred by chance

- An online merchant has historically experienced 10% return rate in the "kitchen gadget" category.
- To increase returns, the merchant does a pilot in which it adds additional explanatory information and pictures about products to its website
- Out of the next 200 purchases, 16 (8%) are returned.
- Is the pilot effective?

```python
hat = np.append(np.repeat(1,1),np.repeat(0,9))
print(hat)
```

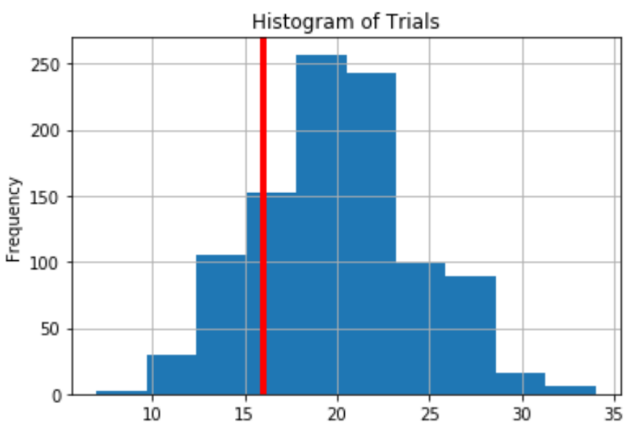
```
[1 0 0 0 0 0 0 0 0 0]
```

```python
def resample_statistic (data, size=200, replace = True):
    data_df = pd.DataFrame(data, columns=['trial'])
    trial_df = data_df.sample(200,replace = True).reset_index(drop=True)
    ret = trial_df['trial'].sum()
    return ret
```

```python
n = 1000
trials = []
for i in range(n):
    trials.append(resample_statistic(hat))
```

```python
trials_df = pd.DataFrame(trials, columns=['trial'])
pval =(trials_df['trial']<=16).sum()/float(n)
print("Estimated p-value: {0:.5f}".format(pval))
trials_df.hist()
plt.title("Histogram of Trials")
plt.ylabel("Frequency")
plt.axvline(x=16, color='r', linestyle='solid', lw=4)
plt.show()
```

```
Estimated p-value: 0.19800
```



- 16 or fewer returns are not unusual
- p-value > 0.05 → Fail to Reject Null Hypothesis

Basic Two-Sample Hypothesis Test: Concept

1. Establish a null model, or the null hypothesis
   - This represents a world in which nothing unusual is happening except by chance
   - Often, this null model is that the two samples come from the same population
2. Examine pairs of resamples drawn repeatedly from the null model to see how much they differ from one another
   - Alternatively, use formulas to learn about distribution of sample differences
   - If the observed difference is rarely encountered in this chance model, then we Reject the Null Hypothesis: random chance is not responsible

Basic Two-Sample Hypothesis Test: Details

1. Make sure you clearly understand:
   - the sizes of the two original samples
   - the statistic used to measure the difference between sample A and sample B:
     - the difference in means, proportions, ratio of proportions
   - the value of that statistic for the original two samples
2. Create a hat that represent the null model
   - e.g.: a hat with all the observed body weights in sample A and sample B
3. Draw two resamples of the same size as the original samples from the hat
   - with or without replacement: similar results unless small samples (<10)
4. Record the value of the statistic of interest
5. Repeat steps 3 and 4 many times for 1,000 trials
   - more trial can produce greater accuracy
6. Note a proportion of trials that yields a value for the statistic as large as that observed (known as p-value)

Alternative H_1: Hypothesis Tests

The alternative hypothesis is the theory you would like to accept, assuming that your results disprove the null hypothesis.

Null Hypothesis, Ho:
- No-fault reporting in hospitals is no more effective than the regular systems

Quantitatively:
- The no-fault system and the regular one both reduce errors to the same degree

Hat Model for Null Ho:
- A single hat with the total number of errors for both groups

Alternative Hypothesis, H_1:
- No-fault reporting in hospitals is BETTER the regular system; it reduces errors more
- One-way: Hypothesis is the question of whether a treatment is *better* than the control

Average error reduction	
Control	1.88
No-fault	2.80
Difference	0.92