Design & Analysis of Data Science Experiments

Hypothesis Testing & Tests

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

Senior Scientist, Computer Science & Mathematics Division
Oak Ridge National Laboratory

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

Statistical Distributions & Functions in R

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

Statistical Distributions: Mean and Variance

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

$$a = 2n_2^2(n_1 + n_2 - 2)$$
$$b = n_1(n_2 - 2)^2\,(n_2 - 4)$$

Reminder: Statistic & its Proxy

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

- <u>Step 1</u>: 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

- <u>Step 2 (optional)</u>: 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

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

- <u>Step 4</u>: 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 ≥ 0.05**)

Important Note

<u>DO NOT SAY</u>: We **ACCEPT** the Hypothesis

<u>INSTEAD</u>: We **FAIL TO REJECT** the Hypothesis

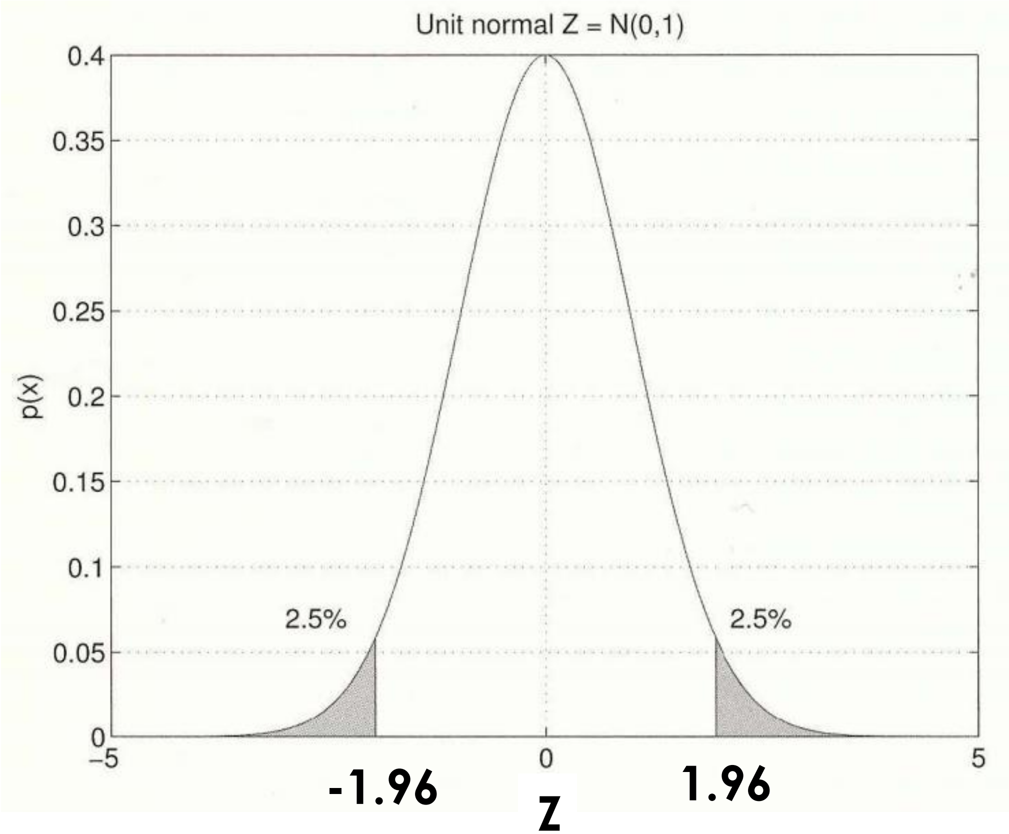
- Given the sample we had to calculate the statistic

Null Hypothesis vs. Alternative Hypothesis

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

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

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



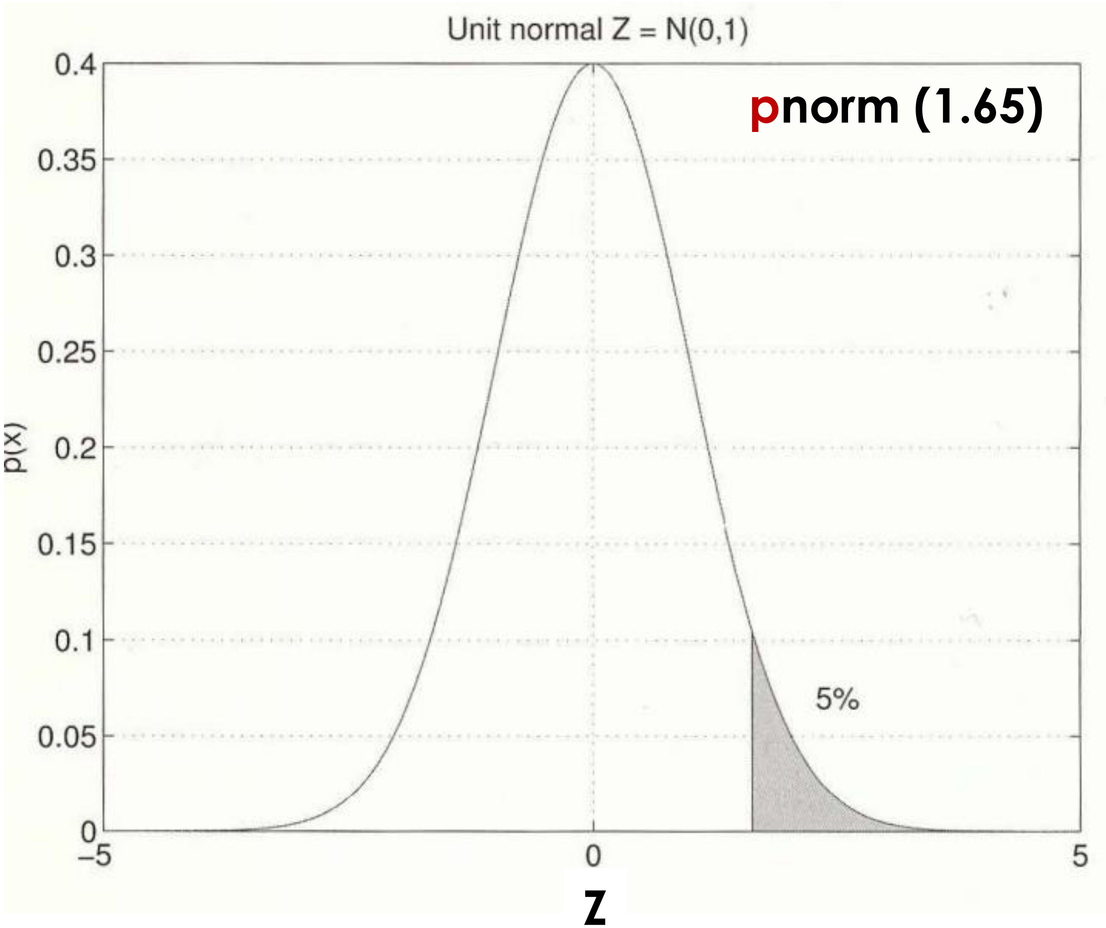
95% of the unit normal distribution lies between - 1.96 and 1.96

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

pnorm (1.96) – pnorm (-1.96)

What is (1 - pnorm(1.96))?

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



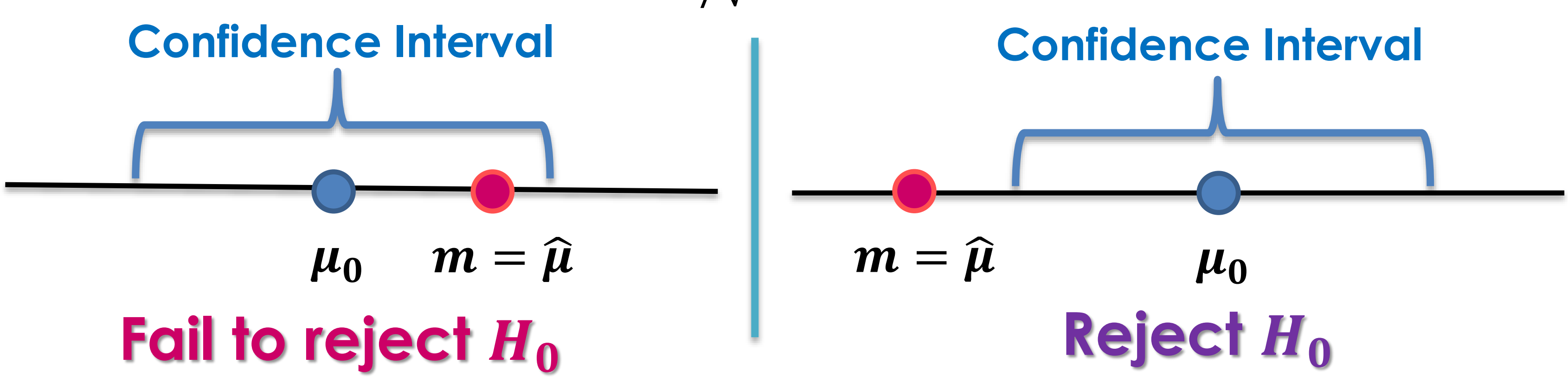
95% of the unit normal distribution lies below 1.64

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

Significance Level: Two-sided Test

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

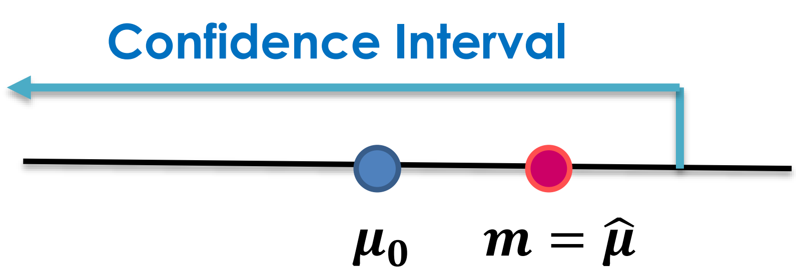
- **Significance Level** (α): We fail to reject the null hypothesis with *level of significance* α if the estimate of the sample statistic lies within the $100(1-\alpha)$ *percent two-sided confidence interval* (CI) for the hypothesized value of the statistic:
  - m is the point estimate of μ :
    -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



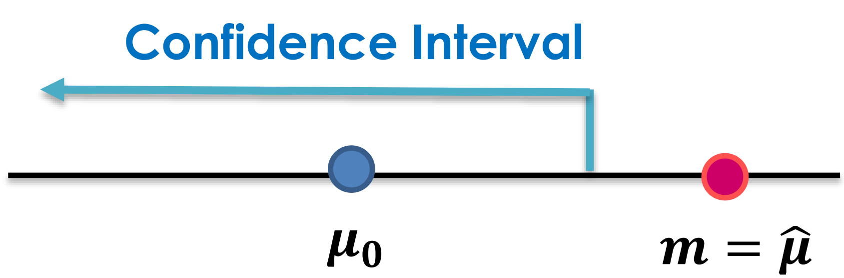
Significance Level: One-sided Test

- 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 $Z \sim \frac{m-\mu_0}{\sigma/\sqrt{n}} \notin (-\infty, z_\alpha)$



Fail to reject H_0



Reject H_0

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**:
  - $\boldsymbol{H_0 : \ \mu = \mu_0}$ **:** 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 $\boldsymbol{t}$**-statistic**: $\boldsymbol{T_{n-1} \sim \frac{m - \mu}{S / \sqrt{n}}}$

Two-sided Test:

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

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

Example: T-Test Hypothesis Testing in R

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
```

One-Sample T-Test (cont.)

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

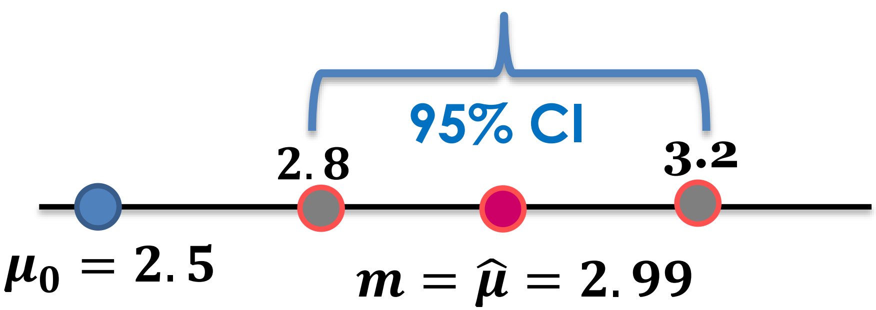
```
> t.test(tips$tip, alternative="two.sided", mu=2.5)

 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**

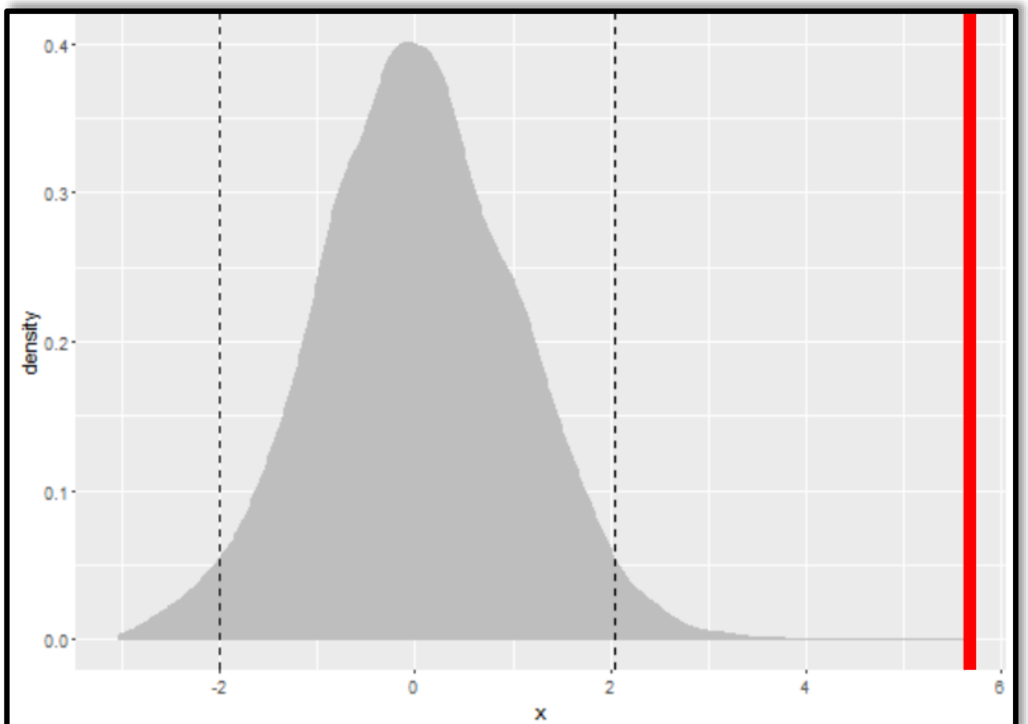


Reject H_0

<u>Conclusion:</u> The mean is not equal to \$2.50

Examine *t*-statistic & its probability visually

```r
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*

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**

<u>Conclusion:</u> 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

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: $$\boldsymbol{\beta(\mu) = P_\mu\{-z_{\alpha/2} \le \frac{m-\mu_0}{\sigma/\sqrt{n}} \le z_{\alpha/2}\}}$$
- **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	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

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

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

- <u>Assumption</u>: 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

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

Quick look into variances

Ex: 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**

Ex: 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**

Paired Two-Sample T-Test: Dependent Groups

H_0: Fathers and sons have equal heights, on average

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

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

```
> 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**

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

Example: Wilcoxon Rank Sum Test

Non-parametric comparison of two independent groups

H_0: Incareceration 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: incareceration rates are not the same

Example: Paired Wilcoxon Signed Rank Test

Non-parametric comparison of two dependent groups

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

```r
library(MASS)
head(UScrime)
sapply(UScrime[c("U1", "U2")], median)
with (UScrime, wilcox.test(U1, U2, paired = TRUE))
```

Check the assumptions: null hypothesis for normality is rejected

```
 Wilcoxon signed rank test with continuity
correction

data:  U1 and U2
V = 1128, p-value = 2.464e-09
alternative hypothesis: true location shift
is not equal to 0
```

Reject H_0

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

Example: 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 $2.2e^{-16}$

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()**