

Statistical computations in R

- In this lecture, you will learn basic statistical computing in R:
- 1-sample t-test/confidence interval for one normal mean
- 2-sample t-tests/confidence interval for $\mu_1 - \mu_2$ (independent samples, paired samples)
- Test/ t-test/confidence interval for 1 population proportion
- Test/ t-test/confidence interval for difference in 2 population proportions

Example 1: A random sample of 50 golf balls of Brand X were hit by a Robot-Driver. Can we conclude that the mean distance obtained exceeds 300 yards?

Yards

303	317	284	305	300
282	297	290	277	293
289	308	304	278	300
298	317	290	301	276
283	293	311	304	318
303	295	299	296	315
309	294	303	285	292
293	291	299	288	303
316	297	282	279	301
302	300	318	310	292

$H_0 : \mu \leq 300$, vs. $H_1 : \mu > 300$

is tested by the 1-sample t-test. In R:

```
yy <- read.csv("G:/DataMining/Data/GolfTest.csv", header=TRUE)
attach(yy)
yyt <- t.test(Yards, mu = 300, alt = "greater")
```

Result of the 1-sample t-test (R OUTPUT) is shown on the next slide.

```
> yyt
```

One Sample t-test

data: Yards

t = -1.5022, df = 49, p-value = 0.9303

alternative hypothesis: true mean is greater than
300

95 percent confidence interval:

294.9214 Inf

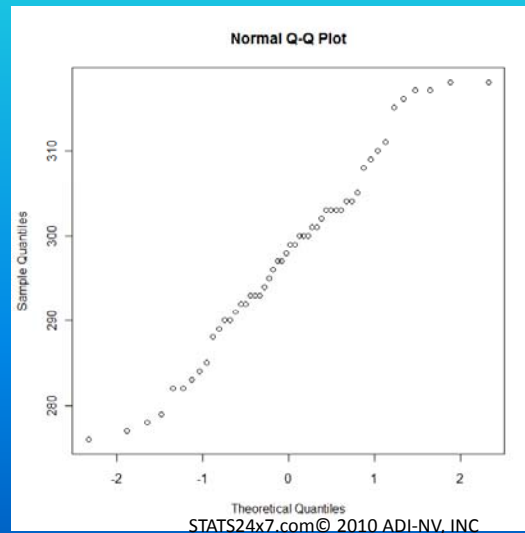
sample estimates:

mean of x

297.6

Verifying that the sample of 50 Yards values are normally distributed (assumption of the t-test) .

```
qqnorm(Yards) # qq plot for testing normality
```



Example 2: A random sample of 35 advertisements in the used car section of a local newspaper resulted in the following prices of late model cars. Test the hypothesis that that the average price of a late model car this year is less than that from last year (\$10,000 from an extensive study done the previous year). Use data file UsedCars.csv)

9100	9520	9600	11390
10170	11660	9910	6740
10550	10070	9650	10840
9670	9120	10520	10440
10590	9160	10300	10170
9560	9200	9640	
6720	9620	9160	
9540	10410	12750	
10190	8520	9100	
10090	9920	10610	

Example 2: (solution)

$$H_0 : \mu \geq 10000$$

$$H_1 : \mu < 10000$$

In R:

```
yy <- read.csv("G:/DataMining/Data/UsedCars.csv", header=TRUE)
> attach(yy)
> yyt2 <- t.test(Yards, mu = 300, alt = "less")
> yyt2
```

```
>yyt2 # this will show the results of t-test (R OUTPUT)
```

One Sample t-test

data: Yards

t = -1.5022, df = 49, p-value = 0.06974

alternative hypothesis: true mean is less than 300

95 percent confidence interval:

-Inf 300.2786

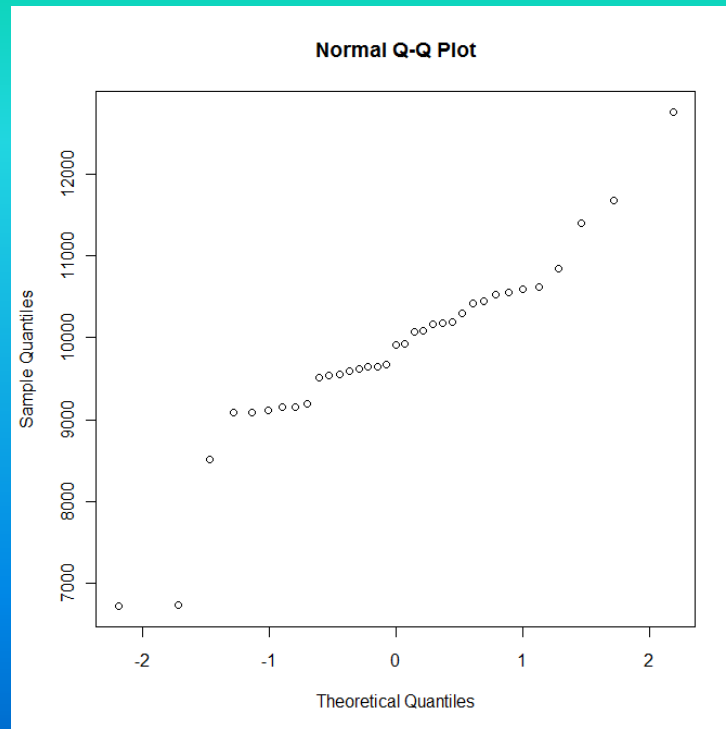
sample estimates:

mean of x

297.6

P-value > .05, do not reject null, conclude mean price is higher than last year's

Verification of normality for data of Example 2: except for a couple of samples at the low end, data seems normally distributed as the points seem to fall along a line.



STATS24x7.com© 2010 ADI-NV, INC

9

Price of 1 gal UL gas on a certain day in Henderson

2.46	2.61	2.68	2.81
3.04	2.78	2.96	2.52
2.84	2.7	2.53	2.47
2.96	2.93	3.08	2.66
2.34	2.67	2.51	
2.69	2.61	3.18	

Is the mean price per gal = \$2.60?

```
> xy <- read.csv("G:/DataMining/Data/UL.csv", header=TRUE)
```

```
> attach(xy)
```

```
# run 1-sample t-test for testing  $\mu = 2.6$  vs.  $\mu \neq 2.6$ 
```

```
xyt3 <- t.test(UL, mu = 2.6, alt = "two.sided")
```

STATS24x7.com© 2010 ADI-NV, INC

10

OUTPUT FROM R

```
> xyt3
```

One Sample t-test

data: UL

t = 2.6916, df = 21, p-value = 0.01366 (*reject null, since $p < .05$*)

alternative hypothesis: true mean is not equal to 2.6

95 percent confidence interval:

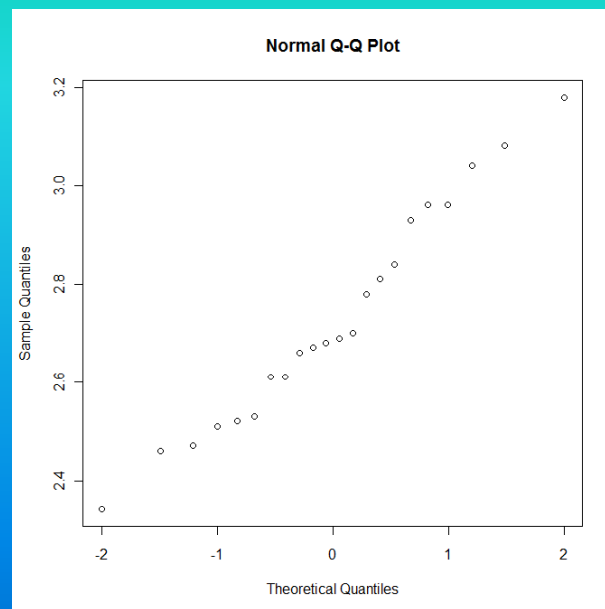
2.629247 2.828026

sample estimates:

mean of x

2.728636

Normal QQ Plot for Price of UL Gas (Example 3)



Example 3:

CaseNo	Type_Fish	Price_1970	Price_1980
1	COD	13.1	27.3
2	FLOUNDER	15.3	42.4
3	HADDOCK	25.8	38.7
4	MENHADEN	1.8	4.5
5	OCEAN PERCH	4.9	23.0
6	SALMON, CHINOOK	55.4	166.3
7	SALMON, COHO	39.3	109.7
8	TUNA, ALBACORE	26.7	80.1
9	CLAMS, SOFT-SHELLED	47.5	150.7
10	CLAMS, HARD-SHELLED	6.6	20.3
11	LOBSTERS, AMERICAN	94.7	189.7
12	OYSTERS, EASTERN	61.1	131.3
13	SEA SCALLOPS	135.6	404.2
14	SHRIMP	47.6	149.0

- Example 3 (Continued): Test if average fish price has gone up.

$$H_0 : \mu_{1970} = \mu_{1980} , H_1 : \mu_{1970} < \mu_{1980}$$

```
x <- read.csv("M:/DataMining/Data/fish_prices.csv")  
attach(x)
```

```
t.test(Price_1970, Price_1980, paired = TRUE, alt = "less")
```

Paired t-test

data: Price_1970 and Price_1980

t = -3.7017, df = 13, p-value = 0.001331

alternative hypothesis: true difference in means is less than 0

95 percent confidence interval:

-Inf -35.83333

sample estimates:

mean of the differences

-68.7

Example 4: The following table shows effectiveness of two brands of car wax. Compare the two brands in terms of mean wax effectiveness at test size 5%. (use data file wax_effectiveness.csv)

$$H_0 : \mu_{\text{Sureglow}} = \mu_{\text{Mirrorsheen}} , H_1 : \mu_{\text{Sureglow}} \neq \mu_{\text{Mirrorsheen}}$$

The 2-sample t-test for independent samples is used, since the samples are independent.

Type	Effectiveness	Type	Effectiveness
Sureglow	93	Mirrorsheen	90
Sureglow	96	Mirrorsheen	97
Sureglow	87	Mirrorsheen	91
Sureglow	91	Mirrorsheen	94
Sureglow	88	Mirrorsheen	100
Sureglow	85	Mirrorsheen	95
Sureglow	88	Mirrorsheen	88
Sureglow	91	Mirrorsheen	92
Sureglow	82	Mirrorsheen	94
Sureglow	91	Mirrorsheen	89
Sureglow	86	Mirrorsheen	96
Sureglow	93	Mirrorsheen	91
Sureglow	91	Mirrorsheen	97
Sureglow	87	Mirrorsheen	92
Sureglow	88	Mirrorsheen	92

```
xx <- read.csv("M:/DataMining/Data/wax_effectiveness.csv")
x1 <- xx$Effectiveness[xx$Type=="Sureglow"]
x2 <- xx$Effectiveness[xx$Type=="Mirrorsheen"]

t.test(x1, x2, var.equal = TRUE, alt = "two.sided")
```

Two Sample t-test

data: x1 and x2

t = -3.2048, df = 28, p-value = 0.003364

alternative hypothesis: true difference in means is not equal to 0

95 percent confidence interval:

-6.665932 -1.467401

sample estimates:

mean of x mean of y

89.13333 93.20000

Confidence Intervals and Test for Proportions in R

Example 5 (single proportion) : A large casino has collected data on a sample of 12844 loyal customers, and classified each as a “high roller” or not. The data file hi_rollers.txt has this data.

- (a) Calculate a 90% confidence interval for the proportion of ‘high rollers’ in the loyal customers database of this casino.
- (b) Test if the true proportion of ‘high rollers’ > 0.15 .

```

x <- read.table("G:/DataMining/Data/Hi_rollers.txt")
# R needs x = number of successes , n = total number of trials
y <- table(x)
> y
x
  0    1
11434 1410

# use prop.test(x, n, p = .15, alt = "greater") to test H0:p=.15 vs >.15

```

```

# run approximate z-test for binomial proportion

```

```

prop.test(11434,11434+1410, p = .15, alt = "greater", conf.level=.9)

```

1-sample proportions test with continuity correction

```

data: 11434 out of 11434 + 1410, null probability 0.15
X-squared = 55190.89, df = 1, p-value < 2.2e-16 (reject null)
alternative hypothesis: true p is greater than 0.15
90 percent confidence interval:
 0.8865966 1.0000000
sample estimates:
      p
0.8902211

```

alternatively, run exact binomial test

```
binom.test(y, p=.15, conf.level=.9, alt="greater")
```

Exact binomial test

data: y

number of successes = 11434, number of trials = 12844, p-value

<

2.2e-16

alternative hypothesis: true probability of success is greater than 0.15

90 percent confidence interval:

0.8866032 1.0000000

sample estimates:

probability of success

0.8902211

Example 6 (comparison of two proportions)

A blackjack player has collected the following data on two blackjack dealers in one casino :

n = # hands dealt, x = # of blackjacks dealt to player

	n	x
Dealer 1	1000	59
Dealer 2	1200	61

The player wants to know if the proportion of blackjacks dealt by the two dealers are equal.

$H_0 : p_1 = p_2$, $H_1 : p_1 \neq p_2$

```
prop.test(c(59, 61), c(1000, 1200), alt="two.sided")
```

2-sample test for equality of proportions with continuity correction

data: c(59, 61) out of c(1000, 1200)

X-squared = 0.5559, df = 1, p-value = 0.4559 (Null NOT rejected)

alternative hypothesis: two.sided

95 percent confidence interval:

-0.01192630 0.02825963

sample estimates:

prop 1 prop 2

0.05900000 0.05083333

Ex 7. In planning her campaign strategy to determine how to approach a particular issue, a congressional candidate wants to know if there are any differences in the proportion of voters who favor the issue among her rural, suburban, and urban constituents. She has collected sample opinions and has obtained the following results:

	Rural	Suburban	Urban
In favor	65	63	52
Not in favor	35	37	48

Null is: P(In Favor) is same for Rural, Suburban, and Urban constituents.

Alternative : null is false.

```
> x <- read.csv("K:/DataMining/Data/campaign.csv",header=FALSE)
```

```
> x
```

```
  V1 V2 V3
```

```
1 65 63 52
```

```
2 35 37 48
```

```
> chix <- chisq.test(x)
```

```
> chix
```

Pearson's Chi-squared test

```
data: x
```

```
X-squared = 4.0833, df = 2, p-value = 0.1298 (Null NOT rejected,  
probability of (In Favor) is same for all constituents.)
```

Example 8: For the following data, test if Age and GAME PREFERENCE are independent.

GAME	21-25	26-50	Over 50
Multi-Line Slots	15	37	16
Video Poker	25	25	17
Wheel of Fortune	14	40	27
Sports Book	11	4	1
Blackjack	9	23	14
Megabucks	3	8	1

```

> xx <- read.csv("K:/DataMining/Data/game_preference.csv",header=FALSE)
> xx
  V1 V2 V3
1 15 37 16
2 25 25 17
3 14 40 27
4 11  4  1
5  9 23 14
6  3  8  1
> chixx <- chisq.test(xx)
Warning message:
In chisq.test(xx) : Chi-squared approximation may be incorrect
> chixx

```

Pearson's Chi-squared test

```

data: xx
X-squared = 28.5528, df = 10, p-value = 0.001471 (NULL is rejected – AGE and GAME
PREFERENCE are associated.)

```

Example 9: You are given the results of 360 rolls of a pair of fair dice (data file Pair-of_Dice.csv). Test the hypothesis that the two dice are fair.

```

y <- read.csv("K:/TDataMining/Data/Pair_of_Dice.csv",header=TRUE)
ty <- table(y)
> > ty
y
 2 3 4 5 6 7 8 9 10 11 12
14 22 23 39 52 75 48 33 31 14 9

tt <- as.matrix(ty)
chi <- chisq.test(tt, p=c(1/36, 2/36, 3/36, 4/36, 5/36, 6/36, 5/36, 4/36, 3/36, 2/36,
1/36))
> Chi
  Chi-squared test for given probabilities
data: tt
X-squared = 10.5267, df = 10, p-value = 0.3956 (conclude dice are fair)

```