

MATLAB统计分析与应用

——描述性统计

主讲人：谢中华

ssstudy.com

科学软件学习网

主要内容

- 常用描述性统计量
- 统计图
- 频数和频率分布表

第一节 常用描述性统计量

常用描述性统计量函数列表

函数名	说 明	函数名	说 明
corr	线性（或秩）相关系数	moment	中心矩
corrcoef	线性相关系数	prctile	百分位数
cov	协方差矩阵	quantile	分位数
iqr	内4分位极差	range	极差
kurtosis	峰度	skewness	偏度
mean	均值	std	标准差
median	中位数	tabulate	频率分布表
mode	众数	var	方差

【例8.1-1】 现有某两个班的某门课程的考试成绩，如下表所示。试对成绩数据进行统计分析，计算描述性统计量。

	A	B	C	D	E	F	G	H
1	序号	班名	学号	姓名	平时成绩	期末成绩	总成绩	备注
2	1	60101	6010101	陈亮	0	63	63	
3	2	60101	6010102	李旭	0	73	73	
4	3	60101	6010103	刘鹏飞	0	0	0	缺考
5	4	60101	6010104	任时迁	0	82	82	
6	5	60101	6010105	苏宏宇	0	80	80	
7	6	60101	6010106	王海涛	0	70	70	
8	7	60101	6010107	王洋	0	88	88	
9	8	60101	6010108	徐靖磊	0	80	80	
10	9	60101	6010109	阎世杰	0	92	92	
11	10	60101	6010110	姚前树	0	84	84	
12	11	60101	6010111	张金铭	0	95	95	
13	12	60101	6010112	朱星宇	0	82	82	
14	13	60101	6010113	韩宏洁	0	75	75	
15	14	60101	6010114	刘菲	0	71	71	
16	15	60101	6010115	苗艳红	0	70	70	
17	16	60101	6010116	宋佳艺	0	80	80	
18	17	60101	6010117	王峰瑶	0	78	78	

【例8.1-1续】根据表中数据计算描述性统计量。

%% 求均值

```
score = xlsread('examp4_3_1.xls','Sheet1','G2:G52');
```

```
score = score(score > 0);
```

```
score_mean = mean(score)
```

%% 求方差和标准差

```
SS1 = var(score)
```

```
SS1 = var(score,0)
```

```
SS2 = var(score,1)
```

```
s1 = std(score)
```

```
s1 = std(score,0)
```

```
s2 = std(score,1)
```

%% 求最大值和最小值

```
score_max = max(score)
```

```
score_min = min(score)
```

%% 求极差

```
score_range = range(score)
```

%% 求中位数

```
score_median = median(score)
```

%% 求分位数

```
score_m1 = quantile(score,[0.25,0.5,0.75])
```

```
score_m2 = prctile(score,[25, 50, 75])
```

%% 求众数

```
score_mode = mode(score)
```

%% 求中心矩

B1 = moment(score,1)

B2 = moment(score,2)

%% 求偏度

score_skewness = skewness(score)

%% 求峰度

score_kurtosis = kurtosis(score)

第二节 统计图

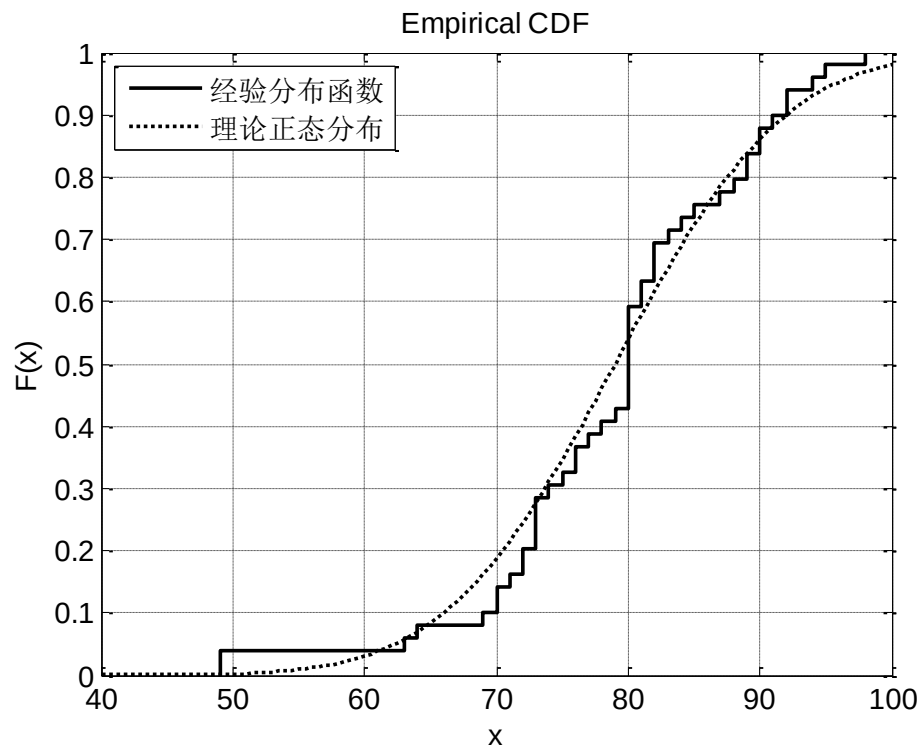
常用统计绘图函数列表

函数名	说 明	函数名	说 明
boxplot	箱线图	ksdensity	核密度估计图
cdfplot	分布函数图	lsline	添加最小二乘拟合线
ecdf	分布函数图	normplot	正态概率图
ecdfhist	频率直方图	parallelcoords	多元数据的平行坐标图
gline	添加参考线	probplot	概率图
glyphplot	星图或脸谱图	qqplot	Q-Q图（分位数图）
gplotmatrix	散点图矩阵	refcurve	添加参考多项式曲线
gscatter	分组散点图	refline	添加参考线
hist	二维直方图	scatterhist	二维散点图和边缘直方图
hist3	三维直方图	wblplot	威布尔分布概率图

【例8.2-1】根据例8.1-1中数据，绘制常用统计图。

➤ 经验分布函数图

```
score = xlsread('examp4_3_1.xls','Sheet1','G2:G52');  
score = score(score > 0);  
figure;  
[h,stats] = cdfplot(score)  
set(h,'color','k','LineWidth',2);  
x = 40:0.5:100;  
y = normcdf(x,stats.mean,stats.std);  
hold on  
plot(x,y,':k','LineWidth',2);  
legend('经验分布函数','理论正态分布','Location','NorthWest');
```



➤ 绘制频率直方图

%% 频数 (率) 直方图

figure;

[f, xc] = ecdf(score);

ecdfhist(f, xc, 7);

xlabel('考试成绩');

ylabel('f(x)');

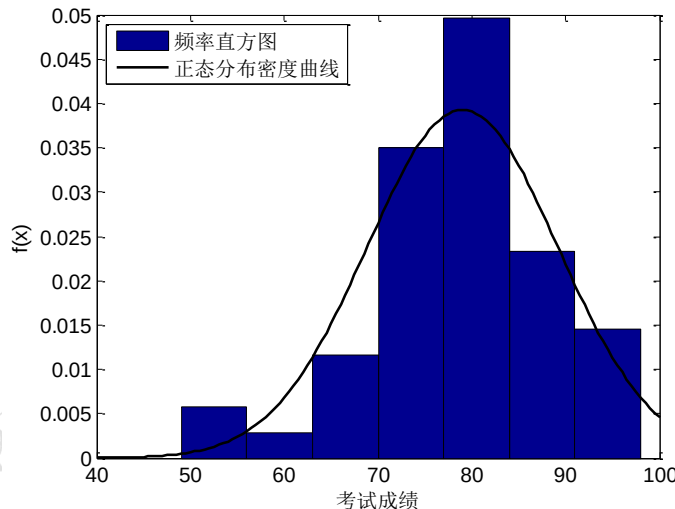
x = 40:0.5:100;

y = normpdf(x, mean(score), std(score));

hold on

plot(x, y, 'k', 'LineWidth', 2)

legend('频率直方图', '正态分布密度曲线', 'Location', 'NorthWest');





绘制箱线图

%% 箱线图

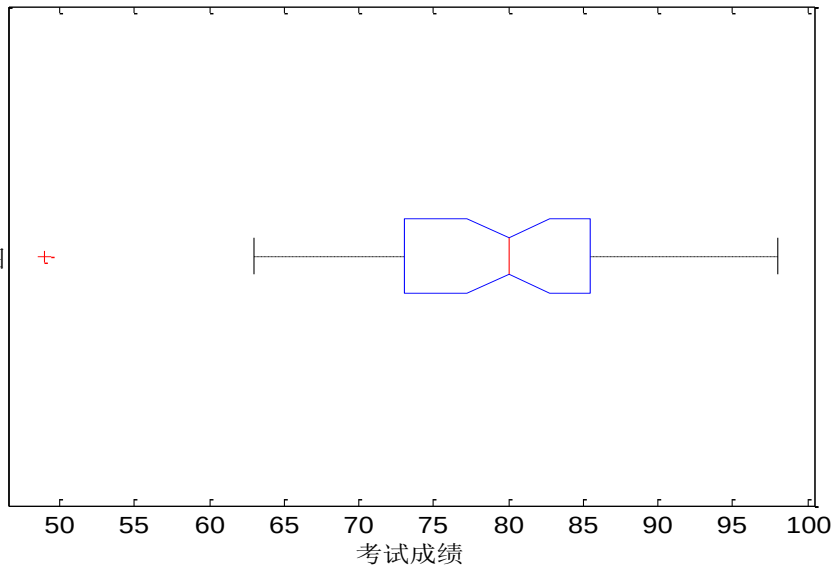
figure;

boxlabel = {'考试成绩箱线图'};

boxplot(score,boxlabel,'notch','on','orientation','horizontal')

xlabel('考试成绩');

考试成绩箱线图

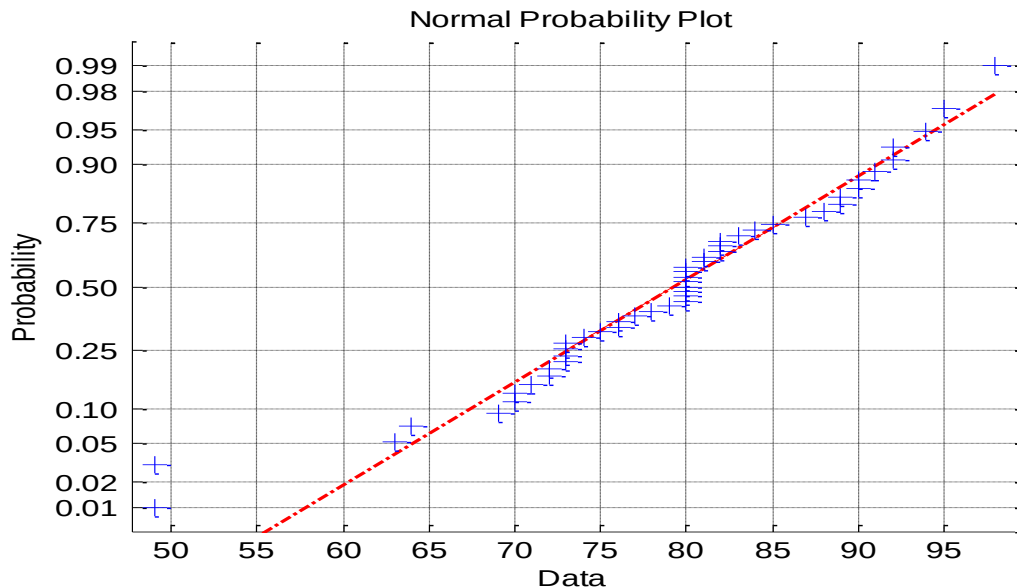


绘制正态概率图

%% 正态概率图

figure;

normplot(score);





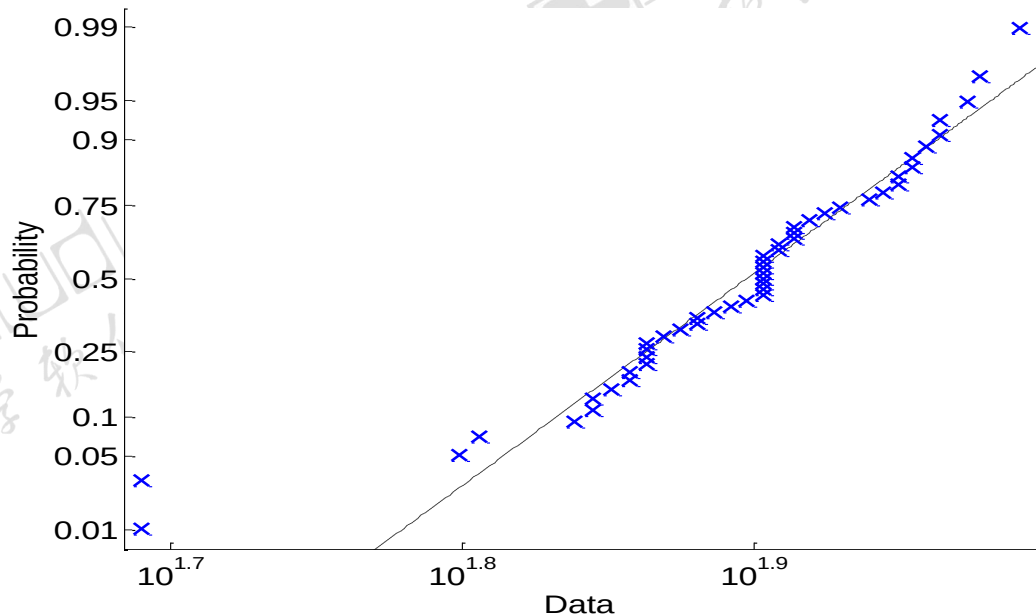
绘制对数正态概率图

%% 对数正态概率图

figure;

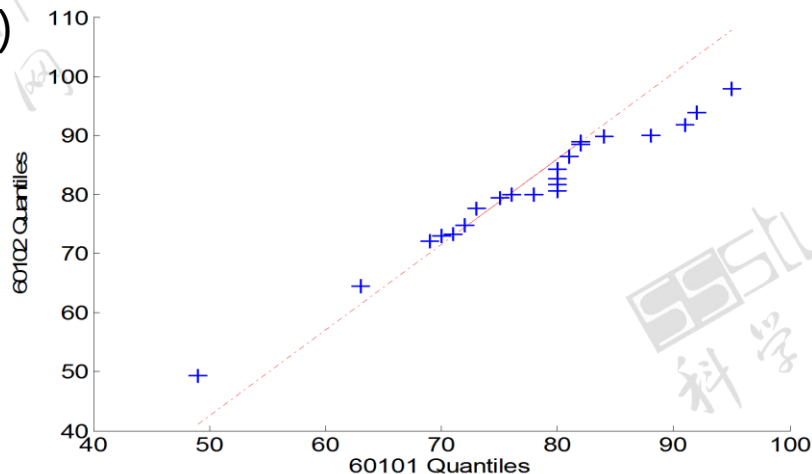
probplot('lognormal',score);

Probability plot for Lognormal distribution



➤ 绘制q-q图

```
banji = xlsread('examp4_3_1.xls','Sheet1','B2:B52');  
score = xlsread('examp4_3_1.xls','Sheet1','G2:G52');  
banji = banji(score > 0);  
score = score(score > 0);  
score1 = score(banji == 60101);  
score2 = score(banji == 60102);  
qqplot(score1,score2)
```



第三节 频数和频率分布表

一、调用tabulate函数作频数和频率分布表

➤ tabulate函数

功能：用来作频数和频率分布表

调用方式：

TABLE = tabulate(x)

【例8.3-1】统计数值型数组中各元素出现的频数、频率。

```
>> x = [2 2 6 5 2 3 2 4 3 4 3 4 4 4 4 2 2  
6 0 4 7 2 5 8 3 1 3 2 5 3 6 2 3 5  
4 3 1 4 2 2 2 3 1 5 2 6 3 4 1 2 5];  
>> tabulate(x(:))
```

Value	Count	Percent
0	1	1.96%
1	4	7.84%
2	14	27.45%
3	10	19.61%
4	10	19.61%
5	6	11.76%
6	4	7.84%
7	1	1.96%
8	1	1.96%

【例8.3-2】统计字符串中各字符出现的频数、频率。

```
>> x = ['If x is a numeric array, TABLE is a numeric  
matrix.'];
```

```
>> tabulate(x)
```

Value	Count	Percent
-------	-------	---------

I	1	2.44%
---	---	-------

f	1	2.44%
---	---	-------

x	2	4.88%
---	---	-------

i	5	12.20%
---	---	--------

s	2	4.88%
---	---	-------

a	5	12.20%
---	---	--------

n	2	4.88%
---	---	-------

u	2	4.88%
---	---	-------

...

【例8.3-3】统计字符型数组中各行元素出现的频数、频率.

```
>> x = ['崔家峰','孙乃喆','安立群','王洪武','王玉杰','高纯静','崔家峰';  
        '叶 鹏','关泽满','谢中华','王宏志','孙乃喆','崔家峰','谢中华'];  
>> tabulate(x)
```

Value	Count	Percent
崔家峰	3	21.43%
孙乃喆	2	14.29%
安立群	1	7.14%
王洪武	1	7.14%
王玉杰	1	7.14%
高纯静	1	7.14%
叶 鹏	1	7.14%
关泽满	1	7.14%
谢中华	2	14.29%
王宏志	1	7.14%

【例8.3-4】统计字符串元胞数组中各字符串出现的频率。

```
>> x = {'崔家峰';'孙乃喆';'安立群';'王洪武';'王玉杰';'高纯静';'崔家峰';  
'叶 鹏';'关泽满';'谢中华';'王宏志';'孙乃喆';'崔家峰';'谢中华'};
```

```
>> tabulate(x)
```

Value	Count	Percent
-------	-------	---------

崔家峰	3	21.43%
-----	---	--------

孙乃喆	2	14.29%
-----	---	--------

安立群	1	7.14%
-----	---	-------

王洪武	1	7.14%
-----	---	-------

王玉杰	1	7.14%
-----	---	-------

高纯静	1	7.14%
-----	---	-------

叶 鹏	1	7.14%
-----	---	-------

关泽满	1	7.14%
-----	---	-------

谢中华	2	14.29%
-----	---	--------

王宏志	1	7.14%
-----	---	-------

【例8.3-5】统计名义尺度（如性别，职业，产品型号等）数组中各元素出现的频数、频率。

% 载入MATLAB自带的鸢尾花数据

```
>> load fisheriris
```

% 将字符串元胞数组species转为名义尺度数组

```
>> species = nominal(species);
```

```
>> tabulate(species)
```

Value	Count	Percent
setosa	50	33.33%
versicolor	50	33.33%
virginica	50	33.33%

二、应用扩展 —— 宋词密码

水调歌头

——苏轼

明月几时有？

把酒问青天。

不知天上宫阙、今夕是何年？

我欲乘风归去，惟恐琼楼玉宇，高处不胜寒。

起舞弄清影，何似在人间？

转朱阁，低绮户，照无眠。

不应有恨、何事长向别时圆？

人有悲欢离合，月有阴晴圆缺，此事古难全。

但愿人长久，千里共婵娟。

宋词中那些词出现
频率最高呢？

1. 宋词中的高频词汇

➤ 99个两字高频词

东风	梅花	黄昏	风吹	时节	杨柳	桃李	而今	时候	悠悠
何处	千里	当年	风月	平生	西湖	人生	鸳鸯	肠断	几度
人间	回首	天涯	多情	凄凉	桃花	十分	为谁	富贵	青山
风流	明月	相逢	故人	春色	扁舟	心事	十年	蓬莱	何时
归去	多少	芳草	当时	匆匆	消息	黄花	去年	昨夜	天气
春风	如今	尊前	无人	功名	憔悴	一声	少年	行人	惟有
西风	阑干	一枝	斜阳	一点	何事	佳人	海棠	今夜	一曲
归来	年年	风雨	不知	无限	芙蓉	长安	寂寞	谁知	月明
江南	万里	流水	不见	今日	神仙	东君	无情	不似	往事
相思	一笑	依旧	深处	天上	一片	断肠	不是	江上	

1. 宋词中的高频词汇（续）

➤ 218个三字高频词

到如今	又还是	谩赢得	记年时	情脉脉	倚东风	终不似
君知否	人不见	那堪更	人间事	斜阳外	又何须	肠断处
谁知道	当此际	君不见	又只恐	月明中	多少事	西源好
功名事	记当年	家山好	回首处	谁念我	安阳好	阑干外
须信道	东风里	归来也	夜沈沈	还知么	对东风	与谁同
最好是	怎奈向	思往事	断人肠	问何如	对西风	从此去
人间世	春去也	还知否	早归来	东风恶	归去也	倚西风
从今去	须知道	追往事	有多少	人何处	归来晚	分明是
谁信道	争知道	最苦是	空凝伫	人正在	愿年年	南徐好
倚阑干	更那堪	空怅望	向尊前	今老矣	江南岸	思晴好
到而今	留不住	空惆怅	微雨过	从别后	空回首	想当年
.....						

2. 让我们也成为词人

- 为高频词编号
- 通过程序随机抽取高频词

3. 小试牛刀

《最苦是》

——谢中华【天津科技大学】

最苦是

鸳鸯归来晚

西湖风吹记当年

明月时节空肠断

3. 小试牛刀（续）

《相思春色佳丽地》

谢中华【天津科技大学】

相思春色佳丽地

富贵凄凉有谁知

相逢当时人不见

今夜风流吾老矣

Thank You