Goodness of Fit

MATH 2441, BCIT

Technical Mathematics for Food Technology

April 26, 2018

Kolmogorov-Smirnov Test

A while ago, we learned how to assess normality informally:

Histogram Construct a histogram. If the histogram departs dramatically from a bell shape, conclude that the data does not have a normal distribution.

Outliers Use a boxplot to identify outliers. If there is more than one outlier present, conclude that the data does not have a normal distribution.

NQP If the data passes the first two tests, use technology to generate a normal quantile plot. The population is probably not normally distributed if either one of the following two conditions apply:

1. The points do not lie reasonably close to a straight line.
2. The points show some systematic pattern that is not a straight-line pattern.

Kolmogorov-Smirnov Test

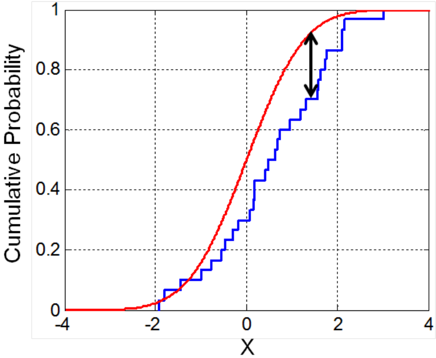
Now that we know how to do a hypothesis test, we can formalize these requirements. Consider a collection of data, for example the IQ test results on the following slide. Is the data normally distributed? Did another distribution generate the data? How good is the fit with a certain distribution? We will learn a goodness of fit procedure for categorical distributions in a moment. First, however, we will have a look at the Kolmogorov-Smirnov test, which tests the goodness of fit with respect to a continuous distribution.

Assessing Normality

104	127	91	91	87	113	112	100	97	87
78	99	87	88	103	95	102	104	114	87
63	93	106	110	82	97	102	100	107	91
105	109	103	115	95	95	128	92	120	107
121	81	101	102	105	119	82	105	73	115
116	95	85	119	113	108	160	128	101	69
87	104	78	85	93	95	128	85	83	82
124	67	106	126	106	103	105	98	76	128
104	122	105	90	110	86	82	87	100	108
83	118	102	109	80	78	112	89	113	92
107	79	111	111	102	84	101	82	93	87
108	113	131	108	87	89	83	92	117	95
85	104	114	113	78	120	102	114	74	97
103	88	90	124	92	120	81	91	139	115
142	99	119	87	109	73	94	95	91	101
101	122	89	107	118	108	97	109	123	125
107	89	117	105	122	92	91	44	106	74
133	125	95	111	128	74	97	112	79	107
127	104	98	109	99	101	104	121	99	119
94	119	84	87	94	92	93	86	104	113

Kolmogorov-Smirnov Test

The Kolmogorov-Smirnov test statistic is the maximum by which the categorical distribution deviates from the continuous distribution to which we compare it.



Kolmogorov-Smirnov Test

The Kolmogorov-Smirnov test statistic has a complicated distribution. Software usually helps us to evaluate whether the null hypothesis passes the test.

null hypothesis H_0 : the data set is generated by the continuous distribution under investigation, for example the standard normal distribution

alternative hypothesis H_a : the data set is not generated by the continuous distribution under investigation

If the IQ results are in the data set `iq`, then the command in R Statistics to perform the Kolmogorov-Smirnov test is

```
ks.test((iq-mean(iq))/sd(iq),pnorm)
```

Kolmogorov-Smirnov Test

Here is the output for the command on the last slide:

```
ks.test((iq-mean(iq))/sd(iq),pnorm)

One-sample Kolmogorov-Smirnov test

data:  (iq - mean(iq))/sd(iq)
D = 0.044518, p-value = 0.8229
alternative hypothesis: two-sided

Warning message:
In ks.test((iq - mean(iq))/sd(iq), pnorm) :
  ties should not be present for
  the Kolmogorov-Smirnov test
```

Kolmogorov-Smirnov Test

Exercise 1: Apply the Kolmogorov-Smirnov hypothesis test to find out if the following data sets are normally distributed. You can find the data on D2L in Lesson 7: Central Limit Theorem (when we learned how to assess normality using normal quantile plots).

1. Ages of Oscar-Winning Actresses `OSCRF.TXT`
2. Body Temperatures `BTEMP.TXT`
3. White Blood Cell Counts for Males `MWHT.TXT`
4. Flight Departure Delays `DPDLY.TXT`

Goodness of Fit

A goodness of fit test is used to test the hypothesis that an observed frequency distribution fits (or conforms to) some claimed distribution. We now know how to apply the Kolmogorov-Smirnov goodness of fit test for a continuous distribution. We will now learn how to apply a test for a categorical distribution.

null hypothesis H_0 : The frequency counts agree with the claimed distribution.

alternative hypothesis H_a : The frequency counts do not agree with the claimed distribution.

Test Statistic for Goodness of Fit Tests

$$\chi^2 = \sum \frac{(O - E)^2}{E}$$

Benford's Law

Abbotsford	141485	Central Saanich	15895
Alert Bay	436	Chase	2365
Anmore	2322	Chetwynd	2877
Armstrong	4842	Chilliwack	90390
Ashcroft	1557	Clearwater	2368
Barriere	1751	Clinton	629
Belcarra	618	Coldstream	10938
Bowen Island	3580	Colwood	17583
Burnaby	238728	Comox	14400
Burns Lake	1803	Coquitlam	147619
Cache Creek	972	Courtenay	26056
Campbell River	33696	Cranbrook	20452
Canal Flats	744	Creston	4661
Castlegar	7934	Cumberland	3562

Benford's Law

Dawson Creek	12115	Grand Forks	4029
Delta	101997	Granisle	307
Duncan	4768	Greenwood	688
Elkford	2630	Harrison Hot Springs	1407
Enderby	2815	Hazelton	257
Esquimalt	16830	Highlands	2394
Fernie	4333	Hope	5796
Fort St. James	1755	Houston	3155
Fort St. John	22618	Hudson's Hope	1022
Fraser Lake	1178	Invermere	2941
Fruitvale	2098	Kamloops	91402
Gibsons	4550	Kaslo	1000
Gold River	1254	Kelowna	125737
Golden	3862	Kent	6220

Benford's Law

Keremeos	1348	Lytton	240
Kimberley	7050	Mackenzie	3492
Kitimat	7664	Maple Ridge	85653
Ladysmith	8342	Masset	859
Lake Country	14183	McBride	576
Lake Cowichan	3169	Merritt	7607
Langford	39936	Metchosin	4792
Langley (city)	27283	Midway	667
Langley (district)	122415	Mission	39873
Lantzville	3408	Montrose	1020
Lillooet	2403	Nakusp	1571
Lions Bay	1325	Nanaimo	93351
Logan Lake	2099	Nelson	11249
Lumby	1772	New Denver	519

New Hazelton	642	Penticton	33016
New Westminster	73771	Pitt Meadows	19090
North Cowichan	30229	Port Alberni	16236
North Saanich	11143	Port Alice	785
North Vancouver (city)	52794	Port Clements	366
North Vancouver (district)	86602	Port Coquitlam	61187
Northern Rockies	5384	Port Edward	474
Oak Bay	17368	Port Hardy	3731
Oliver	4568	Port McNeill	2500
One Hundred Mile House	1860	Port Moody	34193
Osoyoos	4800	Pouce Coupe	689
Parksville	12883	Powell River	13729
Peachland	4959	Prince George	70912
Pemberton	2511	Prince Rupert	11261

Princeton	2782	Sechelt (Sunshine Coast)	21
Qualicum Beach	8687	Sicamous	2468
Queen Charlotte	943	Sidney	11129
Quesnel	9026	Silverton	199
Radium Hot Springs	764	Slocan	309
Revelstoke	7316	Smithers	5462
Richmond	213392	Sooke	11868
Rossland	3639	Spallumcheen	5222
Saanich	110889	Sparwood	4078
Salmo	1165	Squamish	19067
Salmon Arm	18128	Stewart	423
Sayward	311	Summerland	11375
Sechelt Municipality	9490	Sun Peaks Mountain	457
Sechelt (Powell River)	831	Surrey	543940

Benford's Law

Tahsis	295	Warfield	1669
Taylor	1544	Wells	231
Telkwa	1328	West Kelowna	34930
Terrace	10659	West Vancouver	40923
Tofino	2190	Whistler	10627
Trail	7376	White Rock	19288
Tumbler Ridge	2853	Williams Lake	11028
Ucluelet	1634	Zeballos	99
Valemount	947	Wells	231
Vancouver	653046	West Kelowna	34930
Vanderhoof	4526	West Vancouver	40923
Vernon	41671	Whistler	10627
Victoria	85192	White Rock	19288
View Royal	10137	Williams Lake	11028
Warfield	1669	Zeballos	99

Benford's Law

The distribution of first digits is as follows:

1	2	3	4	5	6	7	8	9
52	28	20	18	8	9	11	7	9

Now let's see if this is consistent with a uniform distribution.

	1	2	3	4	5	6	7	8	9
O	52	28	20	18	8	9	11	7	9
E	18	18	18	18	18	18	18	18	18
$O-E$	34	10	2	0	-10	-9	-7	-11	-9
$(O-E)^2$	1156	100	4	0	100	81	49	121	81
$(O-E)^2/E$	64.22	5.56	0.22	0.00	5.56	4.50	2.72	6.72	4.50

The sum of the last row is 94. The degree of freedom is 8 (the degree of freedom is $k-1$, where k is the number of categories, *not* the sample size n). Consulting the Chi-Square χ^2 Distribution table, we record the critical value as 15.507, using $\alpha = 0.05$. We reject the null hypothesis.

Benford's Law

A set of numbers is said to satisfy Benford's law if the leading digit d ($d \in \{1, \ldots, 9\}$) occurs with probability

$$P(d) = \log_{10}(d+1) - \log_{10}(d) = \log_{10}\left(1 + \frac{1}{d}\right) \qquad (1)$$

The distribution of first digits is as follows:

1	2	3	4	5	6	7	8	9
52	28	20	18	8	9	11	7	9

Now let's see if this is consistent with Benford's Law.

	1	2	3	4	5	6	7	8	9
O	52	28	20	18	8	9	11	7	9
E	48.8	28.5	20.2	15.7	12.8	10.8	9.4	8.3	7.4
$O - E$	3.23	-0.53	-0.24	2.30	-4.83	-1.85	1.61	-1.29	1.59
$(O - E)^2$	10.45	0.28	0.06	5.29	23.30	3.41	2.58	1.66	2.52
$(O - E)^2/E$	0.58	0.015	0.0032	0.29	1.29	0.19	0.14	0.09	0.14

The sum of the last row is 3.508727, which is also the test statistic. The critical value is still $\chi^2 = 15.507$. We fail to reject the null hypothesis.

R Commands for Goodness of Fit Test

Multinomial goodness of fit tests (goodness of fit tests for a categorical distribution) are usually right-tailed using the chi-squared distribution. To find the critical value on the last slide, use the R command `qchisq(0.95,8)` or the Microsoft Excel command `CHISQ.INV(0.95,8)`. To perform the goodness of fit test in R Statistics, use the following commands.

```
o<-c(52,28,20,18,8,9,11,7,9)
s<-seq(1,9,1)
bf<-log(1+(1/s))/log(10)
e<-bf*162
chisq.test(o,p=bf)
```

Exercise 2: Mario Triola purchased a slot machine (Bally Model 809) and tested it by playing it 1197 times. There are 10 different categories of outcomes, including no win, win jackpot, win with three bells, and so on. When testing the claim that the observed outcomes agree with the expected frequencies, the author obtained a test statistic of $\chi^2 = 8.185$. Use a 0.05 significance level to test the claim that the actual outcomes agree with the expected frequencies. Does the slot machine appear to be functioning as expected?

Exercise 3: For a recent year, the following are the numbers of homicides that occurred each month in New York City: 38, 30, 46, 40, 46, 49, 47, 50, 50, 42, 37, 37. Use a 0.05 significance level to test the claim that homicides in New York City are equally likely for each of the 12 months. Is there sufficient evidence to support the police commissioner's claim that homicides occur more often in the summer when the weather is better?

Exercise 4: In his book *Outliers*, author Malcolm Gladwell argues that more baseball players have birthdates in the months immediately following July 31, because that was the cutoff date for nonschool baseball leagues. Here is a sample of frequency counts of months of birthdates of American-born major league baseball players starting with January: 387, 329, 366, 344, 336, 313, 313, 503, 421,434, 398, 371. Using a 0.05 significance level, is there sufficient evidence to warrant rejection of the claim that American-born major league baseball players are born in different months with the same frequency? Do the sample values appear to support Gladwell's claim?

Exercise 5: Mario Triola drilled a hole in a die and filled it with a lead weight, then proceeded to roll it 200 times. Here are the observed frequencies for the outcomes of 1, 2, 3, 4, 5, and 6, respectively: 27, 31, 42, 40, 28, 32. Use a 0.05 significance level to test the claim that the outcomes are not equally likely. Does it appear that the loaded die behaves differently than a fair die?

Exercise 6: Records of randomly selected births were obtained and categorized according to the day of the week that they occurred (based on data from the National Center for Health Statistics). Because babies are unfamiliar with our schedule of weekdays, a reasonable claim is that births occur on the different days with equal frequency. See the table that follows. Use a 0.01 significance level to test that claim. Can you provide an explanation for the result?

Day	Sun	Mon	Tue	Wed	Thu	Fri	Sat
Number	77	110	124	122	120	123	97

Exercise 7: Do World War II Bomb Hits Fit a Poisson Distribution? In analyzing hits by V-1 buzz bombs in World War II, South London was subdivided into regions, each with an area of 0.25 km^2. Shown below is a table of actual frequencies of hits and the frequencies expected with the Poisson distribution (first row: Number of Bomb Hits; second row: Actual Number of Regions; third row: Expected Number of Regions from Poisson Distribution). Use the values listed and a 0.05 significance level to test the claim that the actual frequencies fit a Poisson distribution.

0	**1**	**2**	**3**	**4**
229	211	93	35	8
227.5	211.4	97.9	30.5	8.7

Contingency Tables

A contingency table is a table consisting of frequency counts of categorical data corresponding to two different variables.

In a test of independence, we test the null hypothesis that in a contingency table, the row and column variables are independent. (That is, there is no dependency between the row variable and the column variable.)

Example 1: Smoking Cessation. The accompanying table summarizes successes and failures when subjects used different methods when trying to stop smoking. The determination of smoking or not smoking was made five months after the treatment was begun, and the data are based on results from the Centers for Disease Control and Prevention. Test the claim that success is independent of the method used at a significance level of 0.05.

Nicotine	Gum	Patch	Inhaler
Smoking	191	263	95
Not Smoking	59	57	27

Test of Independence

Let O be the observed frequencies; E the expected frequencies; r represents the number of rows, c the number of columns. The requirements are as follows:

1. The sample data are randomly selected.
2. The sample data are represented as frequency counts in a two-way table.
3. For every cell in the contingency table, the expected frequency E is at least 5.

The null hypothesis is that the row and column variables are independent. The test statistic is

$$\chi^2 = \sum \frac{(O - E)^2}{E} \tag{2}$$

The expected values are

$$E = \frac{(\text{row total})(\text{column total})}{(\text{grand total})} \tag{3}$$

Use degree of freedom $(r - 1) \cdot (c - 1)$.

Test of Independence

=(B2-B6)^2/B6

A	B	C	D	E
	Nicotine Gum	Nicotine Patch	Nicotine Inhaler	Row Totals
Smoking	191	263	95	549
Not Smoking	59	57	27	143
Column Totals	250	320	122	**692**
	Nicotine Gum	Nicotine Patch	Nicotine Inhaler	Row Totals
Smoking	198.34	253.87	96.79	549
Not Smoking	51.66	66.13	25.21	143
Column Totals	250	320	122	**692**
	Nicotine Gum	Nicotine Patch	Nicotine Inhaler	Row Totals
Smoking	0.27	0.33	0.03	
Not Smoking	1.04	1.26	0.13	
Column Totals				**3.0618**

Test of Independence

Now let's do this in R Statistics. Here is the data:

```
+------------+------+-------+---------+
|Nicotine    |Gum   |Patch  |Inhaler  |
+------------+------+-------+---------+
|Smoking     |191   |263    |95       |
+------------+------+-------+---------+
|Not Smoking |59    |57     |27       |
+------------+------+-------+---------+
```

Consider the following commands and output:

```
> nicotine<-matrix(c(191,59,263,57,95,27),ncol=3)
> table<-as.table(nicotine)
> chisq.test(table)
Pearson's Chi-squared test
data:  table
X-squared = 3.0618, df = 2, p-value = 0.2163
```

Exercise 8: The table below includes results from polygraph (lie detector) experiments conducted by researchers Charles R. Honts (Boise State University) and Gordon H. Barland (Department of Defense Polygraph Institute). In each ease, it was known if the subject lied or did not lie, so the table indicates when the polygraph test was correct. Use a 0.05 significance level to test the claim that whether a subject lies is independent of the polygraph test indication. Do the results suggest that polygraphs are effective in distinguishing between truths and lies?

	subject did not lie	subject lied
polygraph indicated lie	15	42
polygraph indicated no lie	32	9

Exercise 9: Alert nurses at the Veteran's Affairs Medical Centre in Northampton, Massachusetts, noticed an unusually high number of deaths at times when another nurse, Kristen Gilbert, was working. Those same nurses later noticed missing supplies of the drug epinephrine, which is a synthetic adrenaline that stimulates the heart. Kristen Gilbert was arrested and charged with four counts of murder and two counts of attempted murder. When seeking a grand jury indictment, prosecutors provided a key piece of evidence consisting of the table below. Use a 0.01 significance levd to test the defence claim that deaths on shifts are independent of whether Gilbert was working. What does the result suggest about the guilt or innocence of Gilbert?

	shifts with death	shifts w/o death
Gilbert working	40	217
Gilbert not working	34	1350

Test of Independence Exercise

Exercise 10: Each one of a large population of objects has the following two properties: shape (circular, square, hexagonal, cross) and colour (green, yellow, red, blue). Do shape and colour depend on each other? Use the following sample data of 591 objects and a 10% significance level. The data is provided as a comma-separated values (CSV) table for easy import into Microsoft Excel.

```
,circular,square,hexagonal,cross,
green,57,44,7,23,131
yellow,25,23,18,83,149
red,68,48,32,4,152
blue,47,7,51,54,159
,197,122,108,164,591
```

Exercise 11: In soccer, serious fouls result in a penalty kick with one kicker and one defending goalkeeper. The table below summarizes results from 286 kicks during games among top teams. In the table, jump direction indicates which way the goalkeeper jumped, where the kick direction is from the perspective of the goalkeeper. Use a 0.05 significance level to test the claim that the direction of the kick is independent of the direction of the goalkeeper jump. Do the results support the theory that because the kicks are so fast, goalkeepers have no time to react, so the directions of their jumps are independent of the directions of the kicks?

	GK left	GK centre	GK right
kick left	54	1	37
kick right	46	7	59

End of Lesson

Next Lesson: ANOVA