Model Based Statistics in Biology.

Part II. Quantifying Uncertainty.

Chapter 6.3 Fit of Observed to Model Distribution

ReCap. Part I (Chapters 1,2,3,4)
ReCap. Part II (Ch 5)
6.1 Frequency Distributions from Data
- Discrete Distributions
  - Example, Four Forms, Four Uses
- Continuous Distributions
  - Example, Four Forms, Four Uses
- Uses (Summary)

6.2 Frequency Distributions from a Model
- Notation
- Uses
- Computing Probabilities and Outcomes
  - Cell nuclei (binomial)
  - Lab3
- Model vs Observed Distributions

6.3 Fit of Observed to Model Distribution
- Grouped Data
- Case 1. Mining Disasters (poisson)
- Case 2. Students/row (poisson)
- Case 3. Ages of alumnae mothers (normal)
- Case 4. MUN student mother ages (normal)
- Case 5. Mortality (binomial)
  - Probability plots (Ungrouped Data)

On chalkboard

```
Red chalk for residuals
Yellow chalk for model
White chalk for data
```

```
Lab 3 uses statistical
packages to apply material
on fit of observed to
model distributions.
```

```
First exam just before undergrad drop
date.
                             Exam:open book.
Ability to use tools, not memorize.
Hence
 organize notes for ready access,
 review text for quick access,
 make sure you understand procedures
    and can make calculations.
 work through review material on
    Website (quizzes from past years)
```

ReCap Part I (Chapters 1,2,3,4)
Quantitative reasoning: Example of scallops, which combined stats and models
ReCap (Ch5)
Data equations summarize pattern in data.
Data equations apply to regression lines and to comparison of groups.
The sum of the squared residuals allows us to compare one model to another.
It allows us to quantify the improvement in fit, a key concept in statistics.
ReCap (Ch 6)
Frequency distributions are a key concept in statistics.
They are used to quantify uncertainty.
Observed frequency distributions are constructed from data
Frequency distributions from models are calculated from mathematical functions.

Today: Fit of Observed to Model Frequency Distributions.

Wrap-up. Graphical comparisons of observed to model frequency distributions allow judgement based on more information than a single measure of fit. We quantify the fit of an observed distribution to a probability model using data equations.

Fit of Observed Distribution to a Probability Model

Today we will look at graphical and formal comparison of observed distributions to probability models. We begin with the concept of data equations.

Data	=	Model	+ residual
Observed Distribution	=	Probability Model	+ residual
RF(*Q*=k)/n	=	Pr(*Q*=k))	+ residual

Case 1 (Poisson). Number of coal mining disasters, 1851 - 1866 (England).
Source: Andrews, D.F. A.M. Herzberg 1985. *Data. A Collection of Problems from Many Fields for the Student and Research Worker*. New York. Springer-Verlag. 442 pp

Ndisaster = [4 5 4 1 0 4 3 4 0 6 3 3 4 0 2 4]
sum(N) = 47
k = [0 1 2 3 4 5 6] = outcomes(N)
n = 16 observations

```
Sokal and Rohlf (1995)
Chapter 5, Section 3 uses
recursive formula to
calculate f-hat

f-hat  = Pr(N=k)
```

k	F(N=k)	F(N=k)/n
0	3	0.1875
1	1	0.0625
2	1	0.0625
3	3	0.1875
4	6	0.3750
5	1	0.0625
6	1	0.0625
n= 16		1.0000

k	F(N=k)	F(N=k)/n	Pr(N=k)	Obs-Exp
0	3	0.1875	0.0530	0.1345
1	1	0.0625	0.1557	-0.0932
2	1	0.0625	0.2287	-0.1662
3	3	0.1875	0.2239	-0.0364
4	6	0.3750	0.1644	0.2106
5	1	0.0625	0.0966	-0.0341
6	1	0.0625	0.0473	0.0152
n=	16	1.0000	0.9695	0.0305

From the observed frequency distribution F(N=k) we compute the relative frequency distribution F(N=k)/n, where n = 16.

We estimate the mean $\mu = 47/16$

We apply a formula to get the Poisson distribution for which the mean is
$\mu = 47/16 = 2.9375$

The formula is: $Pr(N=k) = e^{-\mu}\mu^k/k!$
(read in words)

To compare the observed to expected, we calculate observed - expected.

The largest deviation is 0.2106. This shows a slight tendency toward too many cases at a median value of 4 disasters. There is some evidence that data do not fit pattern of poisson model, of rare and random events.

Fit of Observed Distribution to Probability Model

An observed frequency distribution can be compared to a probability model.
This idea of Data = Model + Residual has appeared yet again.

Now we are comparing frequencies (observed versus model) instead of comparing a measured value to a value computed from a model.

Comparison can be made with any theoretical frequency distribution:
normal, t, F, chisquare, binomial, *etc* in addition to Poisson in this example.

These comparisons are usually made for relative frequency distributions, using pdf and cdf functions. But they can be computed for absolute as well as relative frequencies.

The mining disaster example compared the relative frequencies (used the pdf)

$RF(Q{=}k)/n \;=\; Pr(Q{=}k)) \;+\; \text{Residual}$ (uses pdf)

We could have compared the absolute frequencies

$F(Q{=}k) \;=\; n \cdot Pr(Q{=}k)) \;+\; \text{Residual}$

We could also have compared the cumulative distributions

$F(Q \leq k) \;=\; n \cdot Pr(Q \leq k) \;+\; \text{Residual}$

$F(Q \leq k) \;=\; Pr(Q \leq k) \;+\; \text{Residual}$ (uses cdf)

Graphical comparisons of observed to theoretical frequency distributions are useful in that they allow judgement based on all the information about the observed distribution. Thus, the graphical comparison (or the tabular one above, for the mining disasters) will be more informative than computing a single measure of goodness of fit of the observed distribution to the theoretical distribution.

```
Add Graph to compare
observed to theoretical
```

Fit of Observed Distribution to a Probability Model
Case 2 (Poisson). People per row.
Another example, with different dynamics, that of taking seats in a room.
Quickly tally the frequency distribution of number of people per row (or table)
This is the observed frequency distribution F(N=k)
Now calculate the expected number in each row (or table) based on the Poisson distribution.
Choose Poisson as expected distribution if people sit down at random.
$n \cdot \Pr(N{=}k)] =$ total students/number of rows $= e^{-\mu} \cdot \mu^k \cdot (k!)^{-1}$
k = 0 through 12 people per row in 2000
= 48 people / 6 rows = 8 people per row (= estimate of true value of μ)
$\Pr(N{=}k) = e^{-\mu}\mu^k/k!$ (read in words) estimate of μ is 48/6 = 8

		white	yellow	red	<--- chalk
		Data =	Model	+ Residual	
		Obs =	Expected	+ Residual	
k	F(N=k)	F(N=k)/n	Pr(N=k)	Obs-Exp	
0	0	0.000	0.000	0.000	
1	0	0.000	0.003	-0.003	
2	0	0.000	0.011	-0.011	
3	0	0.000	0.029	-0.029	
4	0	0.000	0.057	-0.057	
5	1	0.167	0.092	0.075	
6	1	0.167	0.122	0.045	
7	0	0.000	0.140	-0.140	
8	2	0.333	0.140	0.194	
9	0	0.000	0.124	-0.124	
10	1	0.167	0.099	0.067	
11	1	0.167	0.072	0.094	
12	0	0.000	0.048	-0.048	
n=	6	1.000	0.936	0.064	

We apply the concept of Data = Model + Residual
Does data fit model?
In 2000 there were 6 rows of 12 seats in the room.
The room was half full (48 of 72 seats).
Poisson model does not fit - There tend to be too few cases of 0,1,2,3, or 4 people/row because the room is nearly full, so we cannot have 4/row or less.
Poisson model (rare and random taking of seats) does not fit because it is not appropriate to the dynamics of taking seats in this room. It would be appropriate if there had been few enough students per row that seat choice was not affected by the number of students already in a row.

Examination of the pattern of residuals will tell us more than a single measure of fit.

Fit of Observed Distribution to a Probability Model.
Case 3 (normal distribution). Age of alumni mothers.

In a class of 1500 entering students at Duke University in 2000, there were 63 students with alumni mothers for which year of graduation was reported. Age of mother in year of birth of student was calculated from date of graduation by assuming age 22 at graduation, which is known to be a reasonable assumption.

Age of mother = 2000 – year of parent's graduation + 22

Mothers Age	Obs Freq	Sum(Age)	Sum(Age*Age)	Expected Freq	Deviation Obs-Exp	Obs is:
23	1	23	529	0.49	0.51	
24	0	0	0	1.08	-1.08	low
25	3	75	1875	2.09	0.91	
26	4	104	2704	3.57	0.43	
27	4	108	2916	5.40	-1.40	
28	5	140	3920	7.20	-2.20	low
29	13	377	10933	8.49	4.51	high
30	11	330	9900	8.84	2.16	high
31	8	248	7688	8.13	-0.13	
32	2	64	2048	6.60	-4.60	low
33	5	165	5445	4.74	0.26	high
34	3	102	3468	3.00	0.00	
35	2	70	2450	1.68	0.32	
36	1	36	1296	0.83	0.17	
37	1	37	1369	0.36	0.64	
Sum	63	1879	56541	62.50	0.50	
mean(Age)		29.8254				
var(Age)			8.0497			
stdev(Age)			2.8372			

```
Draw observed normal on board (sideways, yellow chalk)
Add theoretical normal   (sideways, white chalk)
Colour in the difference between observed and theoretical
     (red chalk)
Draw Rootogram (red chalk sideways from vertical white line)
Add confidence limits
```

Fit of Observed Distribution to a Probability Model.
Case 3 (normal distribution). Age of alumnae mothers. (continued)

Draw observed in white chalk, normal in yellow
Draw difference in each class, in red.

Carry pattern in red over to this graph

Interpretation:

no class marks outside the confidence limits. This indicates good fit)
no pattern in deviations

To do this in minitab

Data in c1

MTB> Rootogram c1 (also in Sokal and Rohlf 1995, p 123)

Extra: Modify example to demonstrate pattern in deviations.

```
Modify observed distribution to be platykurtotic
Colour in the deviations
   (positive on left, negative in centre, pos on right)
Draw red deviations as suspended rootogram
```

Case 4 (normal distribution). Ages of mothers of students at MUN.
55 students in quantitative biology course in 1997.

Age Range	Age x	Obs Freq F(Age=x)	Sum(Age)	Sum(Age*Age)	Expected Freq 55*Pr(Age=x)*5	Obs-Exp	Cumulative Frequency F(Age≤x)
16-20	18	4	72	1296	3.59	0.41	4
21-25	23	15	345	7935	14.50	0.50	19
26-30	28	21	588	16464	21.74	-0.74	40
31-35	33	12	396	13068	12.11	-0.11	52
36-40	38	3	114	4332	2.51	0.49	55
41-45	43	0	0	0	0.19	-0.19	55
	Sum	55	1515	43095	54.65	0.35	
	mean(Age)		27.55				
	var(Age)			25.25			
	stdev(Age)			5.03			

Note that to obtain the expected frequency, the probability in each class (computed from normal distribution) is multiplied by number of students (55) and by number of years in each age class (5 years). The formula for the expected distribution is:

$$\Pr(Age = x) = \frac{1}{\sigma 2\pi} e^{\frac{-1}{2}\left(\frac{\mathbf{x} - \mu}{\sigma}\right)^{\mathbf{2}}}$$

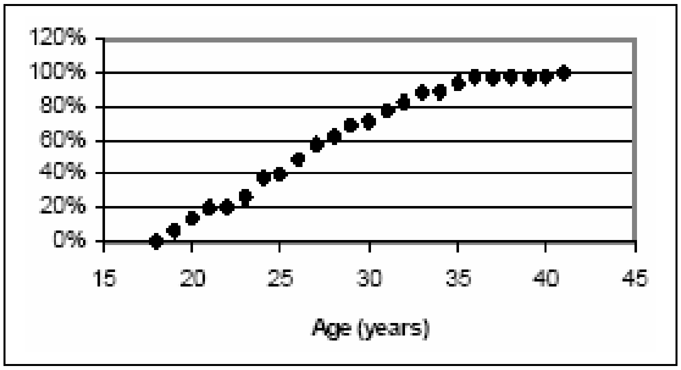
Fit of Observed Distribution to a Probability Model.
Case 5 (Ungrouped data) Ages of mothers of MUN students. Probability plot.

Another way to assess normality is to transform the data to obtain a probability plot.

Example.
Age of mothers of students in quantitative biology courses in 1997.
The mean age was 27.85 years.
Are these ages normally distributed?

The age of each person's mother was recorded in 1997.

We begin by plotting the observed cumulative distribution.



It is hard to judge whether this distribution fits the sigmoid shape of cumulative normal distribution.

k = Age (years)	F(T=k)	F(T<k)	F(T<k)/35	Pr(T<k)
18	0	0	0.00	0.051
19	2	2	0.06	0.074
20	3	5	0.14	0.105
21	2	7	0.20	0.144
22	0	7	0.20	0.191
23	2	9	0.26	0.247
24	4	13	0.37	0.311
25	1	14	0.40	0.382
26	3	17	0.49	0.456
27	3	20	0.57	0.532
28	2	22	0.63	0.607
29	2	24	0.69	0.678
30	1	25	0.71	0.743
31	2	27	0.77	0.801
32	2	29	0.83	0.850
33	2	31	0.89	0.890
34	0	31	0.89	0.922
35	2	33	0.94	0.946
36	1	34	0.97	0.964
37	0	34	0.97	0.977
38	0	34	0.97	0.985
39	0	34	0.97	0.991
40	0	34	0.97	0.995
41	1	35	1.00	0.997

Fit of Observed Distribution to a Probability Model.
Case 5 (Ungrouped data) Ages of mothers of MUN students.
Probability plot.

To aid judgement, we plot the normal distribution with the same mean and variance.

The data (points) match a normal distribution (line).

Most statistical packages produce similar plots, comparing each data point to the line expected for a normal distribution having the same mean and variance.

Most statistical packages will also produce a plot for which normally distributed data fall along a straight line rising from left to right.
Deviations from straight line indicate deviations from normality.

There are several ways of constructing straight line plots, including quantile-quantile plots and normal score plots.