Model Based Statistics in Biology.

Part II. Quantifying Uncertainty and Evidence.

Chapter 6.1 Frequency Distributions from Data

ReCap. Part I (Chapters 1,2,3,4)
ReCap Part II (Ch 5)
6.1 Frequency Distributions from Data
Discrete Distribution
Example, Four Forms, Four Uses
Continuous Distribution
Example, Four Forms, Four Uses
Uses (Summary)
6.2 Frequency Distributions from a Model
6.3 Fit of Observed to Model Distribution

on chalk board

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

ReCap Part I (Chapters 1,2,3,4)
Quantitative reasoning: Example of scallops, which combined stats and models
Quantities: Five part definition
Equations express an idea or concept about the relation of one quantity to another

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.

Today: Frequency Distributions.
Frequency distributions are a key concept in statistics.
For a variable quantity, these distributions summarize information.
They will be used throughout the course, for a variety of purposes.
Frequency distributions can be either empirical (from data)
or theoretical (mathematical expression).

Wrap-up.

We constructed a frequency distribution for discrete data (people per family)
We constructed a frequency distribution for continuous data that had to be grouped (age of mothers)
We obtained alternative forms of the distribution. 4 such
Frequency distributions are a fundamental concept in statistics.
They have multiple uses:

1. Shape is a clue to underlying process
2. Comparison based on all of the information in the data.
3. Statistical inference
4. Reliability of an estimate

Frequency Distributions from Data. Discrete Distributions

Frequency distributions summarize data.
They have a number of uses.

We'll begin by constructing a distribution, then look at the uses to which frequency distributions can be applied.

We'll construct a discrete distribution describing the family size of your mother.

```
The concepts in this lecture
    (outcomes---> probability)
    (probability---> outcome)
are sophisticated and non-intuitive.
Thus it is important to cover and briefly
review the background material that
students have already had (even before
University).  To review, construct
empirical frequency distributions.
```

We begin with the quantity.

Q = [ ] = number per family

```
Record family size from
several students
```

From this, we construct the list of outcomes k = 1,2,3,*etc*

```
Arrange outcomes on
number line, then use
bidding to find largest
family size.
```

k = outcomes(Q) = [ ]

```
Fill in the outcomes = k
Eg.  k = 0 1 2 3.. per family
```

For each outcome we tabulate the frequency of that outcome:

k = outcomes(Q) = [____ ____ ____ ____]

$F(Q = k)$ = [____ ____ ____ ____]

```
Use show of hands to
obtain frequencies
```

This symbol $F(Q = k)$ is read "frequency with which Q equals outcome k"

One frequency for each outcome. This is a compact summary of the distribution of family sizes. Here is the data recorded in 2010.

Family size of student mothers (number of siblings, including your mother)

k = outcomes(Q)	0	1	2	3	4	5	6	7	8	9	10	11	12	13	Sum
F(Q = k)	-	1	8	11	6	0	1	1	3	2	0	0	0	1	34

Frequency Distributions from Data – Discrete Distributions (continued)

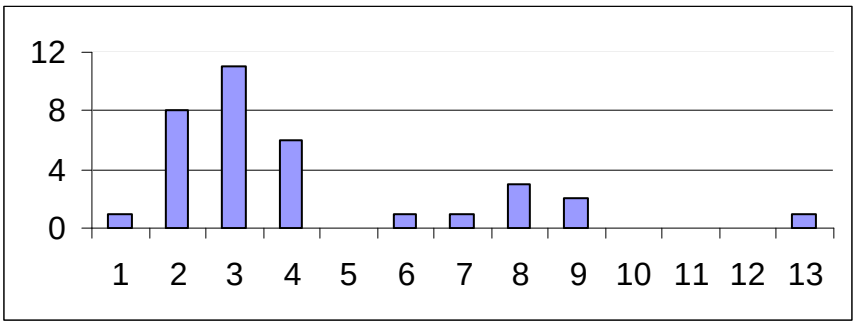
A frequency distribution $F(Q = k)$ can be expressed in three other forms, depending on how we wish to use the distribution.

k = outcomes(Q)	0	1	2	3	4	5	6	7	8	9	10	11	12	13	Sum
F(Q = k)	-	1	8	11	6	0	1	1	3	2	0	0	0	1	34
F(Q = k) / n		0.03	0.24	0.32	0.18	0.00	0.03	0.03	0.09	0.06	0.00	0.00	0.00	0.03	1.00
F($Q \leq$ k)		1	9	20	26	26	27	28	31	33	33	33	33	34	
F($Q \leq$ k) / n		0.03	0.26	0.59	0.76	0.76	0.79	0.82	0.91	0.97	0.97	0.97	0.97	1.00	

The relative frequency distribution is $F(Q = k)/n$

The symbol $F(Q = k)/n$ was devised to be easily read:
"relative frequency with which Q equals outcome k"

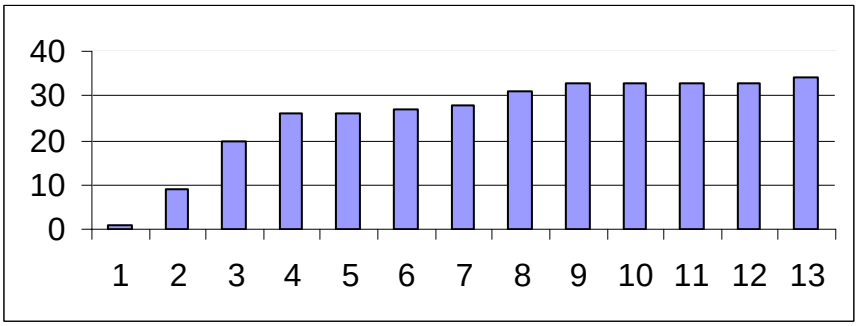
This distribution will have the same shape as the frequency distribution $F(Q=k)$



$F(Q=k)$

The cumulative frequency distribution $F(Q \leq k)$ is taken by summing the frequencies $F(Q=k)$ from the lowest to the highest values of k.

The symbol $F(Q \leq k)$ is read
"frequency that Q is less than or equal to k"
or "cumulative frequency of Q"



The graph of this distribution always increases from left to right.

$F(Q \leq k)$

Finally, we can construct the relative cumulative frequency distribution $F(Q \leq k)/n$

The symbol $F(Q \leq k)/n$ is read "relative cumulative frequency of Q"

This distribution has the same shape as the cumulative distribution.
It ranges upward to 100%

Frequency Distributions from Data. Uses.

Frequency distributions summarize data.
They have a number of uses. As an example of the uses, we use the distribution of number of children per family, for students taking this course in 2010.

1. Use: Summarization is a clue to process.
   Distribution of number of children per family was highly skewed, clustering around 3 per family in parent's generation, but with a few cases of very large families. What factors do you think led to this distribution?

2. Use: Summarization is useful in making comparisons.
   How does the family size of your mother (brothers and sisters) compare to mothers of students from previous year?

Family size of student mothers (number of siblings, including your mother)

	k	0	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	Sum
2010	F(*Q* = k)	-	1	8	11	6	0	1	1	3	2	0	0	0	1	0	0	0	0	0	34
2011	F(*Q* = k)	-	2	8	6	9	3	1	1	0	1	0	0	0	0	0	0	1	0	0	32
2013	F(*Q* = k)	-	1	0	9	5	4	1	0	1	0	1	1	1	0	0	1	0	0	1	24
2014	F(*Q* = k)		2	6	7	6	5	1	1	0	1	0	0	0	0	0	0	0	0	0	21
2015	F(*Q* = k)		1	9	13	7	4	6	2	3	2	1	1	0	0	0	1	0	0	0	40

   We see that the distribution from 2010 was unusual in having two modes, one at 3 children and a smaller mode at 8 children. The other years have a single mode at 3 children (or 4) and long tail

3. Use: Frequency distributions are used in evaluating reliability.
   Application: Reliability of a statistic, the average family size

   In 2010: *mean*(*Q*) = 142 sibs / 34 mothers = 4.2 children/family= sum(*Q*)/n

   We can see that for this data, the average is not a representative value for many people in the class.
   It is true for 6 people (family size of 4)
   But it is not true for the other 34 - 6 = 28 students, especially where the family was large.

4. Use: Frequency distributions are used to calculate probabilities.
   Bayesian Application.
   We begin with a prior probability distribution that considers each child an independent event, with the count is the number of trials to obtain a specified number of children. If we take three as the specified number, we then calculate the posterior probabilities, given the data. The posterior probability is a measure of certainty, given the prior probability model.

Evidentialist application.
Does family size differ among years?
We calculate the goodness of fit to a negative binomial distribution, using maximum likelihood to estimate parameters of the negative binomial parameters from the data. We use a likelihood ratio to measure the strength of evidence for a fit to the negative binomial distribution.

Frequentist application.
Does family size of mothers fit a negative binomial distribution?
We use a chisquare distribution to calculate the uncertainty on our measure of evidence.

Frequency Distributions from Data. Continuous Distribution.

Anothere Example. Age of your mother, when you were born.

The previous example was fairly easy to construct because the quantity of interest, number of people per family had only discrete outcomes: it has to be an integer number, it can't be some fraction, as we defined it.
The next example is for a quantity that is continuous, and for which there are fractional outcomes (fractions of years, days, or even hours).

Q = Age of mother at birth. This is an important quantity in demography.

In the absence of substantial mortality due to diseases (plague, cholera) we can calculate rates of population change and hence demand on services (hospitals, schools) from age of mothers and number of children. Age of mother is at least as important as number of children in its effect of population growth.

T_g = generation time (years). T_g^{-1} = % year^{-1}
This is an approximate estimate of the rate of population increase.
The full estimate is $\log_e$(Family size/(2 parents) T_g^{-1}
The generation time T_g has a large influence on population increase
T_g = 23 years, Then T_g^{-1} = 4.3 % / year (2 children per family)
T_g = 42 years, Then T_g^{-1} = 2.4 % / year (2 children per family)

Let's construct a frequency distribution to see how this important quantity is distributed.

Ask class to work out age of mother when they were born. Then use this to construct frequency distribution, which will require grouping of outcomes into classes.

Frequency Distributions from Data. Continuous Distribution.

Define Quantity T_g = generation time

Define outcomes(T_g) in convenient classes.

Find lower value. Bid up from 15 years ? 16 years ? etc

Find upper value. Bid up from 35 years ? 36 years ? etc.

Tabulate frequencies in each year (show of hands).

Here is the 1997 data, in detail.

19 19
21 21 21 23 23 24 24 24 24 25
26 26 26 27 27 27 28 28 29 29 30
31 31 32 32 33 33 35 35
36
41

Construct classes by dividing range (age 19 to 41) into several classes.

1997: (41 - 19) years / (1 yr/class) = 22 classes

1997: (41 - 19) years / (5 yr/class) = 4.4 classes

5 year groupings showed same pattern as 1 year groupings,

so we will use 5 year groupings to display results.

We assign <u>all</u> of the observations from 16-20 to a single class, age 18

Age 18 is the class mark for ages 16-20.

We assign <u>all</u> of the cases from 21-25 to a single number, k = age 23.

We continue in this fashion for all of the classes.

outcomes(T_g)	= [	____	____	____	____	____	_____] <---class marks in here
Class	=	16-20	21-25	26-30	31-35	36-40	41-45
k	=	18	23	28	33	38	43 <—class marks

Here is the result, showing class marks for each group.

All of the ages have been assigned to the class mark (the midpoint of the age class).

Age Range	Class Mark x	Obs Freq F(Age=x)	n = 33 Relative Freq F(Age=x)/n	Cumulative Freq F(Age≤ x)	Relative Cumulative Freq F(Age≤x)/n
16-20	18	2	0.061	2	0.061
21-25	23	10	0.303	12	0.364
26-30	28	11	0.333	23	0.697
31-35	33	8	0.242	31	0.939
36-40	38	1	0.030	32	0.970
41-45	43	1	0.030	33	1.000
	Sum	33	1		

Now, see if you can do the calculations for 1998 data

Age Range	Class Mark x	Obs Freq F(Age=x)	n = 55 Relative Freq F(Age=x)/n	Cumulative Freq F(Age≤ x)	Relative Cumulative Freq F(Age≤x)/n
16-20	18	11			
21-25	23	19			
26-30	28	18			
31-35	33	7			
36-40	38	0			
41-45	43	0			
	Sum	55	0		

Frequency Distributions from Data -- Uses

1. Use: Summarization as a clue to process.

Application: What factors tend to produce this distribution of ages?

Application: How would this distribution compare to students from a place with different demographics? What factors do you think alter distribution of age of mothers? What factors do you think alter distribution in number of children in a population?

2. Use: Summarization is useful in making comparisons.

Application: Compare to data from previous years. What changes do you see? Do ages of mothers in this course differ between graduate and undergrad students?

Range		1997	1998	2000	2001	2002	2003	2004	2005	2007	2008 under	2008 grad	2009	2010	2011	2013
16-20	18	2	11	2	3	3	2	4	3	2	1	1	2	1	1	1
21-25	23	10	19	17	12	4	7	15	7	12	2	6	10	9	9	6
26-30	28	11	18	11	17	13	12	21	17	12	17	7	16	11	10	12
31-35	33	8	7	10	4	4	4	12	6	6	10	6	14	13	11	8
36-40	38	1	0	2	1	2	2	3	3	2	1	1	4	1	3	2
41-45	43	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0
n		33	55	42	37	26	27	55	36	34	31	21	46	35	34	29
Mean(Age)		27.8	24.9	27.2	26.4	27.6	27.4	27.5	27.9	27.1	29.3	28.0	28.9	28.6	28.9	28.7
stdev(Age)		5.5	4.8	5.05	4.4	5.3	5.06	5.0	5.1	5.0	3.9	5.00	5.1	4.7	5.1	4.8

2. Use: Summarization is useful in making comparisons.

Here is another comparison: How does age of mothers in this course compare to 4th year students at another North America university?

Here is comparison of age of mothers taking this course at MUN in 2003, to age of mothers of 63 students entering Duke University in 2000 (4th year students in fall term of 2003).

Age	Duke		MUN	
16-20	0	0.0%	2	7.4%
21-25	4	6.3%	7	25.9%
26-30	37	58.7%	12	44.4%
31-35	20	31.7%	4	14.8%
36-40	2	3.2%	2	7.4%
41-45	0	0.0%	0	0.0%

Students from same generation, born in early 1980s

Do you think the distributions differ?

Each frequency distribution fully characterizes the information we have for both groups. We can compare any aspect of the distribution: its location at around 28 years of age (mean or median), its dispersion (range, standard deviation), its shape (symmetrical, skewed).

Frequency Distributions from Data -- Uses

3. Use: Frequency distributions are used to calculate strength of evidence.
   How good is the fit to a normal distribution?
   Use: Frequency distribution are used to calculate uncertainty on a measure of strength of evidence.

4. Reliability of estimates

Average T_g for this data is
Mean(T_g) = 910/33 = 27.6 years in 1997
Mean(T_g) = 1359/55= 24.7 years in 1998
Mean(T_g) = 1141/42= 27.2 years in 2000
Similar values, in years after 2000.

How reliable ? How wide a range to encompass most of the values ?
In 1997, most of the values (29 out of 33 = 88%) ranged from 21 to 35 years.
In 1998, most of the values (44 out of 55 = 80%) ranged from 21 to 35 years.
Similarly for subsequent years.
So the mean is not that representative, because of the wide range around the mean, within the range that we see in this variable.

Frequency Distributions from Data. More about uses.

1. Sumarization a clue to process.
It is a good idea, in working with any data, to have the computer make a simple plot of the frequency distribution F(*Q*=k), to see what it looks like.

"This data is tightly clustered around a single value"
"This data is clustered around several values"
"This data is strongly skewed toward mostly zeros"

The shape of the distribution can be a clue to the process that generated the data.
Examples:

If distribution fits Poisson, then a good guess is that it was generated by rare and random events
If distribution looks Normal (clustered symmetrically around a central value), then a good guess is generation by additive effects of a number of independent processes.
If distribution has central peak, but a long tail on the right, then a good guess is that it was generated by interacting processes, such that every once in a while several factors interact to produce large value. This leads to an occasional large value, and a heavy tail on right side of distribution. (This would be guess for age of mothers).

Exploratory analysis with frequency distributions.

Change width of grouping interval (ie distance between class marks) to get another picture.
Construct cumulative distribution to get another view. This often useful in comparing several distributions. Especially useful if only the frequencies are available, not the data, because cumulative distributions allow comparison of data with different class marks.
Stem and leaf diagrams show more detail than F(*Q*=k)
(All statistical packages will construct these)

2. Empirical frequency distributions are useful in making comparisons.
Example: seed production per unit area in two different habitats.
Frequency distribution in each habitat compared.
If shape differs substantially, this suggests that habitat affects seed production.

Frequency Distributions from Data More about uses (continued)

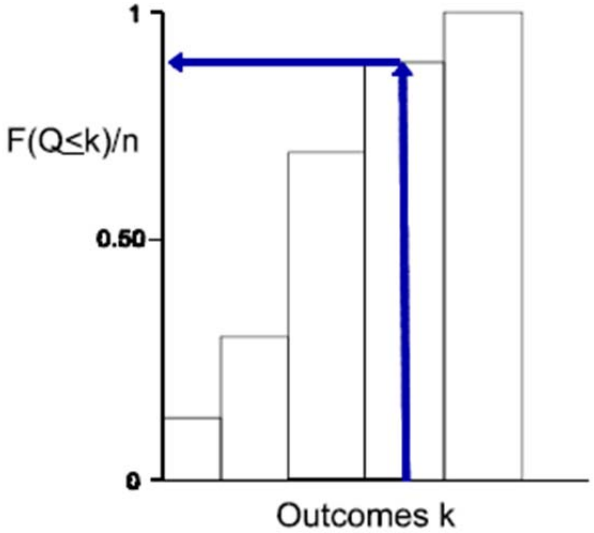
3. Empirical frequency distributions are used in calculate likelihood and probabilities.
   Does the distribution of data compel us to change our minds? (Bayesian)
   Can we reject 'chance' as an explanation? (Decision-theoretic frequentist)

The Bayesian approach begins with a statement about what we believe to be true. We use the distribution to put probabilities on all the outcomes. Then we look at whether the probabilities lead us to change our belief.

The Fisherian frequentist approach rests on long-run probability—the law of large numbers. What is the strength of evidence and the degree of uncertainty given the strength of evidence?

The decision-theoretic approach sets a limit on Type I error (false positives). Is the probability of the observed outcome so small (less the 5%) that we can reject chance as an explanation of pattern in the data?

As an example of the frequentist approach, what it the chance that the mother of the first person you asked, in class today, had more than 8 siblings?



For the class in 2010 it was
p = (2+1)/34 = 0.088 - somewhat improbable.
F($Q \leq k$)/n sums from the left so to obtain p:
p = 1– F($Q \leq k$)/n
p = 1– (31/34) = 1–0.912 = 0.088

For the class in 2011 it was again improbable
p = 1– (30/32) = 0.062

If we assume that this data is a sample from a population that includes the class today, then the chance is p = 1– (61/66) = 0.076

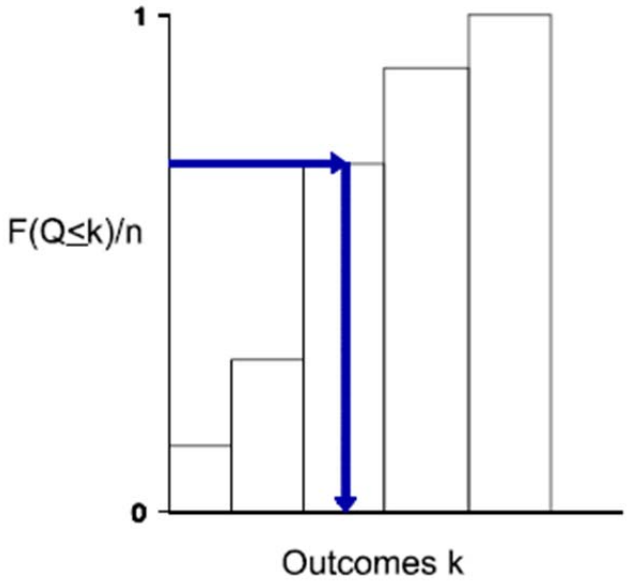
```
Draw arrow from particular outcome
     up to the curve,
     then over to the y-axis
     labelled F(Q ≤ k)/n

This emphasizes that estimate is made by
proceeding from the observed outcome, via a
frequency distribution, to a probability.
```

Frequency Distributions from Data -- Uses (continued)

4. Frequency distributions used in evaluating reliability.

How reliable is an estimate ? This is evaluated by constructing confidence limits.

A frequency distribution is again used. Reliability is evaluated by <u>starting</u> with a probability. Then go across to frequency distribution, then <u>down</u> to outcome corresponding to the probability.



These arrows should be placed on a new drawing of the distribution, rather than using the one already on the board.

This maneuver will be used to construct confidence limits.

Frequency Distributions from Data. Population or sample?

Empirical distributions can be considered as populations or as samples from populations. In the example of ages of mothers, the empirical distribution fully described the population in attendance on the day that students in this course were censused.
The distribution could also be considered:
- a sample of the students registered in the course (because a few students were absent when the census was taken). The sampling fraction relative to registered students in 1997 was about 33/37, or about 90%
- a sample of a much larger population, all students at Memorial. The sampling fraction is on the order of 33/15,000 = 0.02%
- a sample from a still larger population, all graduate and undergraduate students at Canadian and American universities. The comparison of numbers from MUN and Duke suggests that we can combine the MUN and Duke numbers as a sample from this population.