Model Based Statistics in Biology.

Part II. Quantifying Uncertainty.

Chapter 6.2 Frequency Distributions from Models

ReCap. Part I (Chapters 1,2,3,4)
ReCap Part II (Ch 5)
6.1 Frequency Distributions from Data
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

on chalk board

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

Lab 3a uses statistical package to apply material on theoretical frequency distributions.

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 quantifies goodness of fit and improvement in fit.
ReCap (Ch 6)
Frequency distributions are a key concept in statistics.
They are used to quantify uncertainty.
Frequency distributions can be either empirical (from data)
or theoretical (mathematical expression).

Today: Frequency Distributions from a Model

Wrap-up.
Empirical distributions are calculated from data.
Frequency distributions from a model are calculated from mathematical expressions.
They are used, just like distributions from data, to

1. Investigate underlying processes.
2. Compare distributions (now summarized as parameters)
3. Make statistical decisions. compute p-values.
4. Evaluate reliability of an estimate

They are compared to distributions from data to evaluate assumptions.

Frequency Distributions from Probability Models

When we infer from a sample to a population we do not know the true distribution of the population. The classical solution is to use a frequency distribution (probability model) to characterize the population. To do this we must assume that our observations are distributed in the same way as the probability model describing the population.

Here is a list of the most commonly used probability models.

Discrete distributions
- Binomial distribution
- Poisson distribution
- Negative binomial distribution

Continuous distributions.
- Normal distribution
- Sampling distributions generated by sampling from Normal distribution
  - Chi square distribution (from the greek letter "khai")
  - t-distribution (normal/chisquare)
  - F-distribution (chisquare/chisquare)

In 2002 these were presented as tour, first the discrete distributions, then the continuous distributions.
For each a brief summary of underlying process followed by discussion of shape, relation to other distributions, and typical application.
This was too abstract, there were not enough specific examples.
Try 4 cases under heading of Fit of theoretical to observed.
Case 1 Data = foraging success (students each guess and draw shape)
generating mechanism = trials k and success n, p fixed.
binomial pdf, fit.
Case 2. Data = counts per quadrat (students guess and draw)
generating mechanism = counts with low average per unit
(low success from unknown number of trials)
Poisson pdf, fit
Case 3. Data = random numbers (students guess and draw)
generating mechanism = same probability in each category
uniform pdf, fit
Case 4. Data = birth weights (students each guess and draw)
Then 3 cases under heading of calculating probabilities.
Case 1. Situation = trials and success. Binomial pdf
Case 2. Situation = many process. Normal pdf
Case 3. Situation = goodness of fit. X2 pdf

In 2015. Start with concept of Data Equation
Show notation for probability models
Give brief accounting of common probability models
Show Observed and Model notation in Data Eq. format
Link to next Chapter 6.3, several cases, with residuals.

Theoretical Frequency Distributions (Probability models)
We begin with a tour of probability models and how they are defined.
But first, a list of technical terms, for use during the tour, which begins on next page.

X is a random variable

P(X=x) is the relative frequency with which the values of X are equal to the outcome *x*

F(X=x) is the frequency of events in a sample of size n
F(X=x) = n· P(X=x)

P($X \leq x$) is the cumulative relative frequency (cumulative probability)

F($X \leq x$) is the cumulative frequency of events in a sample of size n
F($X \leq x$) = n· P($X \leq x$)

E(X) is the expected value of a random variable X
E(X) = μ_x the true value of the mean of the population of events X
$\hat{\mu}$ is the estimate of the true mean, from the sample. $\hat{\mu} = n^{-1} \sum Q_i$

V(X) is the expected value of the variance of the population of events X
V(X) = σ_x^2 the true value of the variance of the population of events X
$\hat{\sigma}^2$ estimate of variance, from a sample $\hat{\sigma}^2 = (n-1)^{-1} \sum (Q - \overline{Q})^2$, $\overline{Q} = n^{-1} \sum Q$

CD(X) = V(X) / E(X) This ratio, the coefficient of dispersion, is often useful.
It can be estimated from data as $\hat{\sigma}^2 / \hat{\mu}$

E(X) and V(X) are calculated from functions obtained by integrating the function $f_x(x)$

$f_x(x)$ is a function that defines a random variable as normal, binomial, *etc.*
$f_x(x)$ assigns a relative frequency P$(X=x)$ to every possible outcome x
$f_x(x \ = P(X=x)$

$Q \sim fx(x)$ The quantity Q is assumed to be distributed as a random variable defined by a particular function (Normal, Binomial, *etc*).

Discrete distributions - Binomial.

The tour begins with the binomial distribution. This is a discrete distribution, for which outcomes k are whole numbers. Along with the normal distribution, the binomial distribution is one of the most useful distributions in the analysis of biological data. Like other discrete distributions, the binomial stems from Bernoulli trials, each with the same fixed success rate p.
For data based on Bernoulli trials, the Odds of success $p/(1-p)$ will often be of interest.

Binomial distribution. Each unit is scored as a success (1) or as a failure (0)
Examples: number live vs number dead. Number of purple vs white flowers.
Q = number of successes in n trials
Examples: Q = number of rats that develop tumors at a particular dose of a toxin
Q = number of seeds that germinate
We assume that the quantity Q is distributed according to a binomial distribution.
$Q \sim f_k(k;\, p,\, n) = \mathrm{P}(X=k)$
n identical and independent trials, each with success rate p
Under this assumption, we can use the model (the binomial distribution)
to compute the expected frequency distribution for each outcome k
$\mathrm{F}(X=k) = n!\ (k!\ (n-k)!)^{-1}\ \ p^{k}\ (1-p)^{\ n-1}$
k = 0 successes, 1 success, 2 successes, *etc*.
We can use the model to compute the expected (or ‘average’) value in the population.
$\mathrm{E}(X) = \mu = n \cdot p$ = expected number of successes
We can use the model to compute the variance,
a measure of dispersion around the expected value of the population.
$\mathrm{V}(X) = \sigma^2 = n \cdot p\ (1-p)$
$\mathrm{CD}(X) = (1-p)$, the chance of failure. As failure rate decreases,
reliability as measured by $\mathrm{CD}(X)$ increases.
$\mathrm{Odds}(X) = n^{-1}\ \mu^2 / \sigma^2$
To compute the expected value (mean) and dispersion (variance) of the population,
we need to know the value of the parameter p
This value can come from prior expectation such as the proportion of
children that inherit two recessive genes (25% in dihybrid cross)
This value can come from experimental design
(expected number of correct turns in a maze, before training, Lab 3)
This value can be estimated from a sample, $\hat{p} = Q / n$
if we want to compare the fit of the sample to the model (Ch6.3)

Discrete distributions - Poisson.

The next stop on the tour is the Poisson distribution. This is another discrete distribution, for which outcomes k are whole numbers. It is the distributional model that underpins analyses of counts where the number of trials is not known. It underpins the row by column contingency test, a staple in the test repertoire in textbooks. It underpins extension of the 2 x 2 contingency table to log-linear analysis (Bishop, Y.M.M., S.E. Feinberg, P.W. Holland. 1975. *Discrete Multivariate Analysis*. Cambridge, MA, MIT Press). The Poisson distributional model assumes that the variance/mean ratio is close to unity. Unfortunately, these assumptions usually go unmet for counts which are bounded at zero. Examples are physical phenomena (enumbers of hurricanes) and counts of biological phenomena (number of deaths, number patients infected, counts of a defined behaviour).

Poisson distribution. Counts are made within units: quadrats, periods of time, *etc.*

Q = number of counts in a defined unit (usually by space, time, or both).

Examples: Q = number of deaths by horsekick in the Prussian army, per year

Q = number of weevils in azuki beans

$Q \sim f_k(k;\,\lambda) = \mathrm{P}(X{=}k)$

Unknown number of independent trials,

each with fixed and small probability λ in a defined unit

$\mathrm{F}(X{=}k) = \mathrm{e}^{-\lambda}\ \ \lambda^{k}\,(k!)^{-1}$

$\mathrm{E}(X) = \mu = \lambda$, the expected (average) number of counts per unit

$\mathrm{V}(X) = \sigma^2 = \lambda$, the dispersion (variance) around the average count.

$\mathrm{CD}(X) = \mathrm{V}(X) / \mathrm{E}(X) = 1$

To compute the expected value (mean) and dispersion (variance),

we need to know the value of the parameter λ

Unlike the binomial, we rarely have theory or experimental guidance

to state the value of the parameter λ

Like binomial, or any other distribution, we can compare a sample

to the Poisson distribution by estimating the parameter λ from data (Ch6.3).

$\hat{\lambda} = n^{-1}\sum Q$ estimate of λ from a sample of size n

Discrete distributions – Negative Binomial.

The next stop on the tour is the negative binomial (Pascal) distribution, a discrete distribution for which outcomes k are whole numbers. For count data where the data exceeds the mean, it is the next probability model to consider after the Poisson.

Negative binomial distribution. Counts within units, just like Poisson.

Q = number of counts in a defined unit (usually by space, time, or both).

Examples: Number of plants per large quadrat

Counts from repeated trials where odds of failure are small/trial

Number of children in mother's family in B4605/B7220 (?)

$Q \sim f_k(k; p,r) = \mathrm{P}(X=k)$, where $k = r+1, r+2,$ *etc*.

Unknown number of independent trials,

each with success p that varies from trial to trial

r is number of failures until a success occurs.

X = number of successes until predefined number of failures r occur

$\mathrm{F}(X=k)$ = complicated!! Use a routine from a stat package, or from excel.

$\mathrm{E}(X) = \mu = r/p$ = expected number of counts per unit.

$\mathrm{V}(X) = \sigma^2 = r\,(1-p)/p^2$ expected dispersion (variance) of counts.

$\mathrm{CD}(X) = (1-p)/p$ hence not necessarily equal to 1, typically > 1

$\mathrm{CD}(X)$= 1/odds of success, hence low odds when $\mathrm{CD}(X) > 1$

Parameter p estimated from $(CD+1)^{-1}$ where CD is $\hat{\sigma}^2 / \hat{\mu}$

Parameter r estimated from $\hat{\mu}\,(CD+1)^{-1}$

Discrete distributions – Geometric distribution.

The next stop on the tour is the geometric distribution. It appears in engineering, in the context of quality control. It appears rarely in biology and the medical and health sciences. It is a special case of the negative binomial distribution.

Geometric distribution

Q = number of trials until first success.

Examples: Q = number of unsuccessful foraging attempts

$Q \sim f_k(k; p) = \mathrm{P}(X = k)$

n independent trials, with success rate p

$\mathrm{F}(X=k) = p\,(1-p)^{k}$

$\mathrm{E}(X) = \mu = 1/p$ = expected number of trials where p is known

$\mathrm{V}(X) = \sigma^2 = (1-p)/p^2$

$\mathrm{CD}(X) = (1-p)/p$ This is the inverse of the odds of success.

Discrete distributions – Uniform Distribution.

Another discrete distribution of interest in biology is the uniform distribution. This distribution is used to ensure random allocation of sampling units to experimental categories. It is also used in conducting scientifically valid surveys. Random allocation (experiments) and random selection of units from a defined frame (surveys) ensure that the sample is representative of the population (target of inference). For example in a rigorously conducted survey, random numbers from a discrete distribution are used to select the sampled units from a predefined frame (the population) such as list of residents in a defined area.

Discrete Uniform distribution

Each integer from *Qmax* + *Qmin* has same probability

Q = random integer from *Qmax* to *Qmin*

$Q \sim f_k(k; p,r\ ) = \mathrm{P}(X=k)$

Each integer from *Qmax* to *Qmin* has equal probability of occurrence

$\mathrm{F}(X=k) = 1/\mathrm{n}$

$\mathrm{E}(X) = (Xmax + Xmin)/2$

$\mathrm{V}(X) = \sigma^2 = ((Xmax - Xmin + 1)^2 - 1)/12$

Example: integer from a random number table

Continuous distributions.

Normal distribution

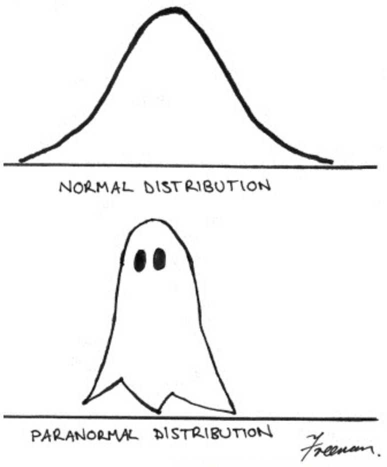
Outcome due to larger number of independent factors

Examples: birth weights, adult heights, measurement errors

Lognormal distribution

Outcomes due to a small number of factors, or interacting factors

Examples:



Copyright Matthew Freeman

Gamma distribution

Continuous analog of negative binomial distribution.

Used as probability model for data with residuals showing asymmetrical skew.

Examples:

Beta distribution.

Analog of binomial distribution, in that it is bounded between zero and one.

Examples: habitat percent cover.

Beta-binomial distribution.

Outcomes due to a process that determines presence absence and a second process that governs number if present.

Continuous distributions used to calculate Type I error (p-values).

Chi-square distribution

Sum of squared normal variables

Examples:

ratio of variance of a sample to variance of the population

sum of squared differences between observed and expected numbers in a genetics experiment

t-distribution

Ratio of random normal variate to chi-square variate

F-distribution

Ratio of one chisquare variate to another

F, t, and Chi-square are related to each other.
The all do the same job – produce a p-value from a likelihood ratio.

Oh God - I look just like my mother!
Uncle Zed
The T-distribution Family at home

Probability Models – Symbols and notation.

There are 4 forms of any distribution. The following notation distinguishes probability models from empirical distributions.

n = sample size

N = finite population

Work through handout (below), distinguishing F(*Q* = k) from Pr(*X*=*x*)

Empirical	Theoretical	Theoretical
	(k discrete)	(x continuous)
F(*Q*=k)	N ·Pr(*X*=k)	N ·Pr(*X*=*x*)]
F(*Q*=k)/n	Pr(*X*=k) = pmf	Pr(*X*=*x*) = pdf
F(*Q*≤k)	N ·Pr(*X*≤k)	N ·Pr(*X*≤*x*)
F(*Q*≤k)/n	Pr(*X*≤k)	Pr(*X*≤*x*) = cdf
white	yellow chalk	yellow

All 4 forms of a theoretical distribution can obtained one from another. This is done either by integrating, which is like going from a frequency distribution F(*Q*=k) to a cumulative distribution F(*Q*≤k). Or it is done by differentiation, which is like going from F(*Q*≤k) to F(*Q*=k).

The distributions encountered in most statistical texts are theoretical distributions, in one of these 4 forms. For example the normal distribution is usually pictured in the form of the Pr(*X*=*x*), while the p-values are calculated from the cumulative normal distribution Pr(*X*≤*x*). Some of the theoretical forms are used so often that they have abbreviated names: the pmf is the probability mass function, the pdf is the probability density function, the cdf is the cumulative distribution function.

Table 6.1 Notation for Frequency Distributions and Probability Models.

Notation for frequency distributions and probability functions vary from text to text. Here are some notational conventions that tend to be widely used. Equivalent notation is also shown.

An empirical distribution constructed from a sample of size n can be expressed in any of four different ways:

$F(Q = k)$	histogram of values	frequencies
$F(Q = k)/n$	histogram of proportions	relative frequencies
$F(Q \leq k)$	histogram of cumulative values	cumulative frequencies
$F(Q \leq k)/n$	histogram of proportions	cumulative relative frequencies

Theoretical distributions can be either discrete (binomial, Poisson) or continuous (normal, chisquare, F, t). These are functional expressions. The probability density function pdf is a function for the probability, or relative frequency. The cumulative density function cdf is for the cumulative probability, or cumulative frequency. These function can thus be considered models for the frequency distribution obtained from data.

	Observed	Expected	k is discrete	Q is measured
	n = sample	N = finite population	x is continuous	X is continuous
Frequency	$F(Q = k)$	Frequency of Q in the sample of size n		(the histogram)
	$n \cdot Pr(X \leq k)$	Expected frequency that X in sample, limited to k values		
	$n \cdot Pr(X \leq x)$	Expected frequency X in sample, X continuous		
	$N \cdot Pr(Q \leq k)$	Expected frequency that Q in population, k values only		
	$N \cdot Pr(X \leq x)$	Expected frequency X in population, X continuous		
Relative Frequency	$F(Q = k)/n$	Proportion of Q in the sample of size n		
	$Pr(Q = k)$	Probability that $Q = k$	probability mass function, pmf	
	$Pr(X=x)$	Probability that $X = x$	probability density function, pdf	
Cumulative Frequency	$F(Q \leq k)$	Cumulative frequency of Q		
	$n \cdot Pr(Q \leq k)$	Expected frequency that $Q \leq k$ in sample, limited to k values		
	$n \cdot Pr(X \leq x)$	Expected frequency $X \leq x$ in sample, X continuous		
	$N \cdot Pr(Q \leq k)$	Expected frequency that $Q \leq k$ in population, k values only		
	$N \cdot Pr(X \leq x)$	Expected frequency $X \leq x$ in population, X continuous		
Cum. Relative Frequency	$F(Q \leq k)/n$	Proportion of $Q \leq k$ in the sample of size n		
	$Pr(Q \leq k)$	Probability that $Q \leq k$	cumulative mass function, cmf	
	$Pr(X \leq x)$	Probability that $X \leq x$	cumulative density function, cdf	

Equivalent notation	$Pr(Q = k)$	$f(x)$	pmf	$P(Q = k)$	for discrete variables
	$Pr(X = x)$	$f(x)$	pdf	$P(X = x)$	for continuous
	$Pr(Q \leq k)$	$F(x)$	cmf	$P(Q \leq k)$	for discrete variables
	$Pr(X \leq x)$	$F(x)$	cdf	$P(X \leq x)$	for continuous

Frequency Distributions from Probability Models -- Notation and Concordance

The notation adopted here distinguishes theoretical and empirical distributions. It distinguishes the four forms of any distribution. The symbols are consistent with prevalent usage in the literature.

However this is not the only notation. In text it is sometimes unclear whether a theoretical or an empirical distribution is being used. In texts it is sometimes difficult to determine which of the four different forms is under discussion. This usually takes some scrutiny to determine. It helps to keep in mind that there are four different forms, then work out which distribution is being used.

Because notation varies among texts, the recourse is to set up a concordance. That is, list one set of symbols, then list the corresponding symbols from the second set immediately beneath the first set.

Application: Table 6.1 in Sokal and Rohlf (1995).
Discuss and translate each column.
In Box 6.1 Q = birth weights

This goes well if most people have books.
Dropped in 2003 because few people had books.

Add in:	outcomes(*Q*)=k	Column (1)	
	Pr(X=x)	Column (5)	= probability density function pdf
	Pr(*Q*=k)	Column (6)	
	N·Pr(*Q*=k)	Column (7)	
	Residuals	Column (8)	
	z-scores	Column (3)	= (class mark – mean)/st.dev

Another application from text. Eq 6.1 page 101
The normal distribution. What is Z ?
Which of the four forms are we looking at ?
Looking at the graph, it seems that Pr(Y=y) is meant,
where Y = birth weight and y is outcomes(Y).
it is a theoretical distribution (i.e. an equation)
it is on relative basis (sum = 1) based on Fig 6.2 and 6.3
it is not cumulative
by elimination it is Pr(Y=y)= Z
but note that this Z has nothing to do with z-score
z scores are single data values standardized relative to mean and to sd

Another application. Eq 6.2 in Sokal and Rohlf. what is meaning of z here ?
It is now for a sample, not the population.
It is E[F(*Q*=k)] not E[F(*Q*=k)/n]

Another application. p 103 in Sokal and Rohlf cdf and pdf.

page 103	pdf = probability density function	Pr(X=x)
	cdf = cumulative distribution function	Pr(X≤x)

Frequency Distributions from Probability Models -- Uses

1. Clue to underlying process.

If an empirical distribution fits one of the following, then this suggests the kind of mechanism that generated the data.

Uniform distribution
e.g. number of people per table in crowded room
generating mechanism is that all outcomes have equal probability
Normal distribution
e.g. oxygen intake per day
generating mechanism is usually several independent factors
t-distribution = small sample from normal distribution
Chisquare = distribution of a variance e.g Fig 7.5 Sokal & Rohlf95
F-distribution = ratio of variances
Poisson distribution
e.g. counts of rare plant per quadrat
gen. mechanism is rare and random event
Binomial distribution
e.g. number of heads on several tosses of coins
e.g. number of successful captures in several tries by predator
gen. mechanism is yes/no outcome on repeated trials
Beta binomial
e.g. captures by several predators, each with diff. probability of success
gen. mechanism is collection of binomial processes, each with different probablity of success
Negative binomial
e.g. number of plants in quadrat, if not rare
gen. mechanism is sequence of prior events, such as several critical events leading up to successful colonization of quadrat

2. Used to summarize data. For example, number of births per unit area summarized as Poisson parameter. Estimate of parameter of appropriate distribution is a powerful summary of set of observations.
Summarization in this form lends itself to comparisons. For example: birth rate in one habitat as a parameter) versus that in another.
The information in a distribution is reduced to a few parameters, which we use to compare the information.

Theoretical Frequency Distributions -- Uses (continued)

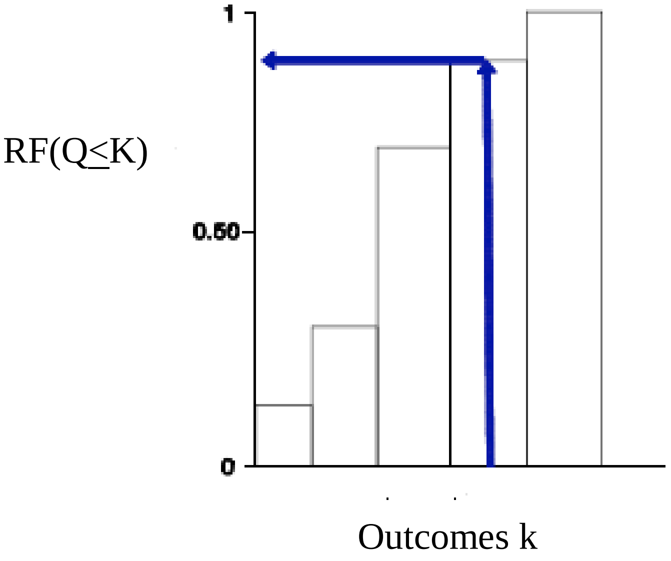
3. Decision making. Can use theoretical distribution to calculate p-value, instead of undertaking the effort of tabulating a frequency distribution.

Many statistical texts have tabled valued of theoretical distributions for this purpose. In the days before computers thse tables were the best way to use theoretical distributions. With computers these tables are obsolete and imprecise. It takes no longer to compute an exact probability with a computer (e.g. with Minitab) than to look up a critical value in a table. The computer value is better, because the precise probability is returned, not an outcome corresponding to one of a small number of tabled probability. The use of computers to calculate exact probabilities is consistent with modern practice in statistics, which is to report exact p-values corresponding to an outcome. This is much better then simply declaring "the result was significant at 5%"

Many examples of this use of theoretical distributions in statistics.

Will use Minitab to calculate p-values from observed outcome (cdf command)

This can be visualized as same maneuver used with empirical distribution. Start with outcome, move up to curve (the theoretical distribution), then across to the probability.

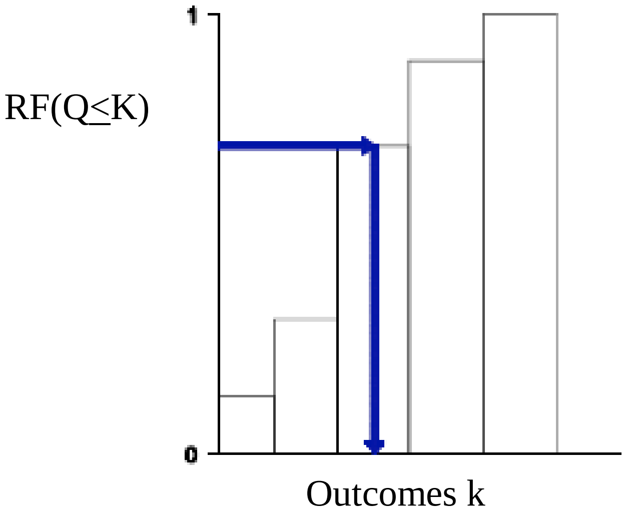


Theoretical Frequency Distributions -- Uses (continued)

4. Reliability. Put probability range around an outcome. Read from probability space to sample space.

This use of theoretical frequency distributions also made easier with computers. In Minitab this is done with invcdf command.

Can be visualized as starting with range on the y-axis (probability range) then proceeding across to the curve, then down to pair of outcomes that correspond to the probability range.



Theoretical Frequency Distributions -- Computing probabilities

Cell nuclei.

The most common use of theoretical distributions in statistics is calculating p-values and setting confidence limits. Theoretical distributions have additional uses. These will be covered here briefly.

One use of theoretical distributions is to calculate expected outcomes, before data are collected.

Here is an example, using the Binomial distribution.

At any one time, 2% of all cell nuclei are undergoing mitosis.
Question: How many nuclei must be examined in order to have a 95% chance of finding at least one nucleus dividing ?

Two percentges here, makes it confusing.
2% chance of each nuclei dividing
100% – 2 % = 98% of each not dividing

95% chance of finding at least one nucleus dividing
100% – 95% = 5% chance of finding none dividing.

Similar questions that are easier :
How many examined to have 2% success ? 1
How many to have 98% failure ? 1
How many to have 5% failure ? more than 1
How many to have 95% success more than 1, same as above

Looking for the frequency of an outcome (at least one event)
Same as the frequency of another outcome 1 – RF(no events)

Can calculate either Pr(at least 1) or Pr(none)
Easier to calculate Pr(none)

Computing Probabilities (Binomial example continued)

With problem in mind, search for right formula for Pr(X=x) or Pr(*Q*=k)
Want the pdf Pr(X=x) or the probability mass function pmf = Pr(*Q*=k)
What kind of distribution ?
Series of success-failures, each independent, each with same probability (2%)
The appropriate distribution is the binomial distribution.
The problem is like coin flipping, except that the "coin" lands on heads 2% tails 98%
This is a discrete distribution so we want the pmf
The formula for the binomial distribution pmf is:

$$\Pr(Y=k)) \;=\; C(k,Y) \cdot p^{k-Y}\, q^{Y}$$

trials k = cell nuclei examined for mitosis
outcomes Y = mitosis happening (number of successes)
rate p = proportion <u>not</u> undergoing mitosis at any one time
q = proportion undergoing mitosis = 1–p
C(k,Y) = combinations of k things taken Y at a time.

It is easier to work with failures where C(k,Y) = 1
There is only <u>1</u> way to get all failures in k trials
C(Y=k) = 1

```
This material requires
extensive preparation,
to go well.
```

If we work with failures (p = 98%), then the equation reduces to:

$$\Pr(Y=k)) \;=\; p^{k-Y}\, q^{Y}$$

if p = 98% failure at any one time, how many nuclei have to be examined in order to have 5% failures ?
This is the same as asking the chance (95%) of finding more than 0
k is unknown
p = 98% failure
Y = 0
Pr(Y=k) = 1 – 95% = 5%
$\Pr(Y=k) = \; p^{k-Y}\, q^{Y} = \; 0.98^{k-0}\, 0.02^{0}$
$0.05 = .98^{k-0}\, .02^{0}$
$\ln(0.05) = k(\ln(0.98) + 0 \cdot \ln(0.02)$
$k = \ln(0.05)/\ln(0.98)$
$k = 148$

It is sometimes hard to tell if formula has been applied correctly.
To check, ask whether the answer look reasonable.
do other answers (similar values) look reasonable ?
e.g. if p = 80% rather than 98% should k go up or down ?
then calculalate k for 80% of time not dividing.

Computing probabilities from observed versus theoretical distributions.

Examples

- randomization test in S&R section 18.3
- Fisher's exact test.
- Non-parametric tests based on ranks
  - Tables are based on tally of outcomes, not on equation for th. distribution

Examples of probabilities computed from theoretical distributions (equation for each)

- t, F, Chi-square, Anova

What are the advantages and disadvantages of these ?

Observed distributions

- Advantages
  - no assumptions
  - easiest to defend because no assumptions
- Disadvantages
  - lots of computation
  - not always easy to carry out.
  - no ready made commands in stats packages

Theoretical distributions

- Advantages
  - Easy to do in statistical packages,
    - which are set up to use theoretical distributions
  - Familiar
  - Good recipes, known performance
- Disadvantages
  - Assumptions may not apply, so p-value may be wrong
  - Checking assumptions can be laborious
    - Almost as easy to do the observed distribution as
      - to do a thorough check on assumptions

Probabilities from Theoretical versus Observed Frequency Distributions

A note on "non-parametric" tests: It is a confusing term.

Does this refer to statistical test with no parameters ?

Does this refer to test with no parametric distribution of outcomes ?

I.e. does it apply to use of empirical rather than theoretical frequency distributions ?

For example, some texts refer to a chisquare test as a "non-parametric" test. In fact such a test has parameters (expected row and column proportions). It typically employs a theoretical frequency distribution (Chi-square, with parameter = df). A chisquare test is most certainly a parametric test.

In practice non-parametric has come to mean a special type of test where data are reduced to ranks

These are becoming historical artifacts (from the days of no computers)

Reasoning was that linear models (t-tests, regressions, etc) based on theoretical distributions could not be trusted because residuals were not normally distributed.

So the solution, before computers, was to reduce data to rank scale so that all outcomes could be enumerated for a statistic based on ranks.

Texts for natural scientists have whole chapter sdevoted to these kinds of tests.

Kruskal Wallis etc.

These are randomization tests that use the strategy of reducing data to ranks.

If you have a computer or access to the web this is no longer necessary

It is not a very good solution because information is lost.

Randomization tests based on data are better.

As we have seen, a t-test can now be done with non-normal data.

Just do a randomization test, using the t-statistic (no assumptions about distribution needed).

Practice that developed from 20th century texts was once necessary:.

Start with set of classic methods (regression, ANOVA, etc)

If normal error assumptions violated, use another set of methods

These are based on defensive position of a 'non-parametric test'

The cost is stupefying the data (to ranks, loss of information).

The 21st century solution: present randomization tests early in the course, rather than bringing in 'non-parametric' tests at the end as curatives whenever first set of methods cannot be used.