Stat 5421 Notes: To Accompany Agresti Ch 1

Charles J. Geyer

September 22, 2023

License

Statistical Terminology

Probability Models and Statistical Models

A *probability model* gives probabilities and expectations for some random process. They are also called probability distributions, probability laws, and probability measures.

In *classical probability theory* (undergraduate and master's level) they come in two kinds: discrete and continuous. In measure-theoretic probability theory (PhD level) here is only one kind (which includes the two kinds of classical theory and more).

This is a course in which all of the data are discrete, so all of our models for data will be discrete probability models. We will only encounter continuous distributions as approximations to discrete distributions. More on this throughout the course.

A *statistical model* is a family of probability models. The idea is that we have data produced by some random process. We assume some statistical model contains the probability model that describes the data. This is called the *true unknown distribution of the data* ("unknown" because we do not know which distribution in the statistical model is the truth). Then *statistical inference* is the process of saying something about which distribution is the true one.

In *frequentist inference* this process can take the form of point estimates, hypothesis tests, and confidence intervals, which should be familiar from other courses.

In *Bayesian inference* this process can take the form of posterior distributions, posterior expectations, or posterior probabilities. This should have been covered in a theory course, if you have had one, but we won't assume anyone has had a theory course. It should also have been covered in the statistical computing course, if you have had that, but that is also not a pre-requisite so we won't assume anyone has had that either. So we won't assume students have previous acquaintance with Bayesian inference (except, of course, the example in Chapter 0).

Statistical models also come in two kinds: parametric and nonparametric. In this course we only use parametric statistical models. Nonparametric models are the subject of a nonparametrics course (like Stat 5601 at the University of Minnesota).

The term "parameter" has two very closely related meanings in statistics, so closely related you cannot always tell which meaning is meant (but you usually can tell).

The first meaning is any quantity that is determined by the probability distributions in a statistical model. The mean (of the distributions) is a parameter. The median (of the distributions) is a parameter. The

standard deviation (of the distributions) is a parameter. In this meaning, even nonparametric families of distributions have parameters.

The second meaning is any quantity that determines probability distributions within a statistical model. It may take more than one variable to do this, in which case we say we have a vector parameter (collecting all of the parameter variables into one thing: a vector).

The mean (of the distributions) is the parameter of the Poisson family of distributions. The mean and variance (of the distributions) are the parameters of the normal family of distributions. And so forth. For each family of distributions (statistical model) we cover, we will explain the parameters.

A family of distributions can have more than one *parameterization*. The distributions can be determined in different ways. We start with one parameter and later may change to another. Much more on this theme throughout the course.

The set of allowed parameter values (those that correspond to distributions in the family) is called the *parameter space*.

We say a family of distributions is nonparametric (hence not the subject of this course) if no parameter vector of finite length parameterizes the family. Nonparametric families are too big to be parametric.

Probability Mass Functions and Probability Density Functions

A discrete probability model is specified by a *probability mass function* (PMF), a real-valued function that is nonnegative and sums to one. Its domain is called the *sample space*. Points in the sample space are called *outcomes*. Subsets of the sample space are called *events*.

If f is a PMF, then $f(x)$ is the probability of the outcome x. Probabilities are between zero and one, so $0 \le f(x) \le 1$ for all x.

A continuous probability model is specified by a *probability density function* (PDF), a real-valued function that is nonnegative and integrates to one. Its domain is called the *sample space*. Points in the sample space are called *outcomes*. Subsets of the sample space are called *events*.

If f is a PDF, then $f(x)$ is *not* the probability of the outcome x. It is probability per unit length (hence probability density) of intervals very near x. If dx is an infinitesimal length, then $f(x)\,dx$ is the probability of the interval from x to $x + dx$. This may not make any sense if you have not had calculus. We do always have $0 \le f(x)$ for all x, because probabilities are nonnegative. But we need not have $f(x) \le 1$ because $f(x)$ is not a probability; rather $f(x)\,dx$ is a probability.

If A is an event, we write $\Pr(A)$ for the probability of the event A. We have $0 \le \Pr(A) \le 1$ for all events A, regardless of whether the distribution is discrete or continuous. Probabilities are calculated by doing sums or integrals (depending on whether the distribution is discrete or continuous), as explained in theory courses, but these sums or integrals are often not ones explained in calculus classes. Often these sums or integrals cannot be done symbolically by any method known to mathematics, so we use computer programs to do them or approximate them by so-called asymptotic approximation. More on asymptotic approximation below and further below. More on calculation by computer simulation in these notes and these notes.

A real-valued function on the sample space is called a random variable. If X is a random variable, we write $E(X)$ for the expectation of the random variable X. Expectations are calculated by doing sums or integrals (depending on whether the distribution is discrete or continuous), as explained in theory courses, but, as with probabilities, these sums or integrals are often not ones explained in calculus classes and often cannot be done symbolically by any method known to mathematics, so we use computer programs to do them or approximate them by so-called asymptotic approximation.

Another word for expectation is *mean*. If $\mu = E(X)$, then we can say μ is the expectation of X or the mean of X. We can also say μ is the mean of the distribution of X. Another word for expectation is *expected value*.

Addition rule: the mean of a sum of random variables is the sum of the means of these random variables.

Independent and Identically Distributed

Random variables are *independent* if probabilities and expectations for them can be calculated by multiplication. Sometimes we say *stochastically independent* or *statistically independent* to single out this notion of independence, but in probability and statistics the words "independent" and "independence" refer to this concept and no other.

If X_1, X_2, ..., X_n are independent, then

$$\Pr(X_i \in A_i \text{ for all } i) = \prod_{i=1}^{n} \Pr(X_i \in A_i)$$

$$E\left(\prod_{i=1}^{n} g_i(X_i)\right) = \prod_{i=1}^{n} E(g_i(X_i))$$

In particular, for PMF or PDF we have

$$f(x_1, \ldots, x_n) = \prod_{i=1}^{n} f_i(x_i)$$

On the left-hand side we have the PMF or PDF of the so-called joint distribution (of all the random variables). On the right-hand side we have the PMF or PDF of the so-called marginal distributions (of each of the random variables separately).

How do we know when to apply this concept to data? When the random variables in question have nothing to do with each other. What happens with one has no influence on what happens with another.

When the random variables in question are *independent and identically distributed* every one has the same marginal distribution so

$$f(x_1, \ldots, x_n) = \prod_{i=1}^{n} f(x_i)$$

(all the PMF or PDF on the right-hand side are the same). (There is some "abuse of notation" here in using the same letter f for two different functions, but you can tell them apart because the f on the left-hand side has n arguments and the f on the right-hand side has one argument).

Multiplication rule: the mean of a product of independent random variables is the product of the means of these random variables. (This is false if the word "independent" is omitted.)

Variance and Standard Deviation

A special expectation that gets its own name is *variance,* which is expected squared deviation from the mean

$$\operatorname{var}(X) = E\{(X - \mu)^2\}$$

where

$$\mu = E(X)$$

When doing applied statistics, variance has a problem. If X has units (say feet), then $\operatorname{var}(X)$ has different units (feet squared or square feet). So $\operatorname{var}(X)$ is not comparable to X.

Thus *standard deviation*

$$\operatorname{sd}(X) = \sqrt{\operatorname{var}(X)}$$

was invented. Taking this square root gets back to the units of X.

As you were taught in intro stats, standard deviation is one measure of spread of a random variable (of how spread out its distribution is).

This measure (standard deviation) is primarily important because it is a parameter of the normal distribution. Hence standard deviation is important in any discussion of asymptotic normality.

But the standard deviation concept becomes completely inadequate as soon as there is more than one random variable under discussion. The generalization of the variance concept to random vectors has no analogous standard deviation concept.

Thus, unlike some intro stats courses, we cannot ignore variance and only use standard deviation.

Addition rule: the variance of a sum of independent random variables is the sum of the variances of these random variables. (This is false if the word "independent" is omitted.) This implies an addition rule for standard deviations, but that rule is not simple, involving squares and square roots.

The Mean of a Random Vector

The mean of a random vector Y is the vector whose components are the means of the components of Y, that is, if

$$\mu_i = E(Y_i), \qquad \text{for all } i$$

and $Y = (Y_1, Y_2, \ldots, Y_k)$ and $\mu = (\mu_1, \mu_2, \ldots, \mu_k)$, then

$$\mu = E(Y)$$

Covariance and the Variance of a Random Vector

If X and Y are random variables having means μ_X and μ_Y, respectively, then

$$\operatorname{cov}(X, Y) = E\{(X - \mu_X)(Y - \mu_Y)\}$$

is called the *covariance* of X and Y.

Knowing only the variances of the components of a random vector tells you little about the distribution of that random vector. (This will become evident throughout the course.) You have to also know the covariances of pairs of components. Thus the following.

The *variance* of a random vector Y is the (nonrandom) matrix whose i, j component is $\operatorname{cov}(Y_i, Y_j)$. Note that the covariance of a random variable with itself is the variance

$$\operatorname{cov}(X, X) = \operatorname{var}(X)$$

so

$$\operatorname{var}(Y) = \begin{pmatrix} \operatorname{var}(Y_1) & \operatorname{cov}(Y_1, Y_2) & \cdots & \operatorname{cov}(Y_1, Y_k) \\ \operatorname{cov}(Y_2, Y_1) & \operatorname{var}(Y_2) & \cdots & \operatorname{cov}(Y_2, Y_k) \\ \vdots & \vdots & \ddots & \vdots \\ \operatorname{cov}(Y_k, Y_1) & \operatorname{cov}(Y_k, Y_2) & \cdots & \operatorname{var}(Y_k) \end{pmatrix}$$

where k is the dimension of Y.

Your humble author prefers the name *variance matrix* for this concept, the reason being that it does play the same role that the variance (scalar) of a random variable plays in univariate theory. (This will become evident throughout the course.)

But many people disagree. At least three other terms are widely used. One such term is *covariance matrix*. The logic behind this name is that all of the components of the matrix are covariances. (Even the terms that are variances, the ones on the diagonal, are also covariances of components of Y with themselves.) But this is a bad term because it uses up the name that should be properly applied to the covariance of two random vectors.

Hence many people call this the *variance-covariance matrix*.

Others, disgusted with the confusion of terminology call it the *dispersion matrix*.

These notes will always say "variance matrix". You can call it what you like.

Note that there is no analog of standard deviation of a random vector. The components of a random vector, which are random scalars, have standard deviations. The random vector does not. There is no way to take the square root of a matrix. (At least no unique way, and no way that has a simple interpretation.)

Correlation

Correlation is covariance divided by standard deviations

$$\operatorname{cor}(X, Y) = \frac{\operatorname{cov}(X, Y)}{\operatorname{sd}(X) \operatorname{sd}(Y)}$$

If $\operatorname{sd}(X) = 0$ or $\operatorname{sd}(Y) = 0$, then $\operatorname{cov}(X, Y) = 0$, so the correlation is $0/0$ which is undefined.

There are two important properties of correlation. The *correlation inequality* says

$$-1 \le \operatorname{cor}(X, Y) \le 1$$

(when the correlation is defined). Hence one can tell what is large correlation (near one or minus one).

If $|\operatorname{cor}(X, Y)| = 1$, then Y is linearly related to X, that is, $Y = a + bX$ for some constants a and b, and vice versa (of course, not the same constants for the inverse relation). Thus Y can be perfectly predicted from X despite the randomness, and vice versa.

The slope of the equation relating Y and X (the b in $a + bX$) has the same sign as the correlation.

The correlation matrix of a random vector with components Y_i has i, j component $\operatorname{cor}(Y_i, Y_j)$.

Some Statistical Models

Here we are syncing up with Section 1.2.1 in Agresti.

The Bernoulli Distribution

The section title is a misnomer, since "Bernoulli" denotes a parametric statistical model (a family of distributions) so there is not just one Bernoulli distribution so the "the" in "the Bernoulli distribution" is wrong. But everybody talks this way, and you can't change how people talk, so we will use the same sloppy way of talking about every statistical model (saying, for example, the normal distribution instead of the normal family of distributions or the normal statistical model). To be pedantically correct, we should say the Bernoulli family of distributions or the Bernoulli statistical model.

A random variable is *Bernoulli* if it has only two possible values: zero or one. Bernoulli means the same thing as zero-or-one-valued.

Because probabilities sum to one, $f(0) + f(1) = 1$, where f is the PMF. If

$$f(1) = \pi,$$

then

$$f(0) = 1 - \pi.$$

We see that π determines the probabilities, hence it parameterizes the statistical model. We can write

$$f_\pi(x) = \begin{cases} \pi, & x = 1 \\ 1 - \pi, & x = 0 \end{cases}$$

The parameter space of the Bernoulli statistical model is the interval $[0, 1]$ (the square brackets mean the endpoints are included in the interval). In other words, $0 \le \pi \le 1$.

An abbreviation for the Bernoulli distribution with parameter π is $\mathrm{Ber}(\pi)$.

If X is a random variable having the $\mathrm{Ber}(\pi)$ distribution, then

$$\begin{aligned} E(X) &= \pi \\ \mathrm{var}(X) &= \pi(1 - \pi) \\ \mathrm{sd}(X) &= \sqrt{\pi(1 - \pi)} \end{aligned}$$

Addition rule: see under the binomial distribution

The Binomial Distribution

As noted above, the section title is a misnomer, since "binomial" denotes a parametric statistical model (a family of distributions) so there is not just one binomial distribution so the "the" in "the binomial distribution" is wrong. To be pedantically correct, we should say the binomial family of distributions or the binomial statistical model.

Suppose we have an IID sequence of Bernoulli random variables X_1, X_2, ..., X_n. Then

$$Y = \sum_{i=1}^{n} X_i$$

has the *binomial distribution*.

If the X_i have the $\mathrm{Ber}(\pi)$ distribution, then the distribution of Y is binomial with *sample size* n and *parameter* π. We abbreviate this distribution as $\mathrm{Bin}(n, \pi)$.

The *rationale* of the binomial distribution is that it is a distribution of counts when we are looking at any dichotomous (two-valued) variable. We have a random sample of individuals and we ask them all one question and we are interested in one particular answer, which we code as one and we code all the other answers as zero. Then the sum counts the individuals who gave the answer we are coding as one. It might be the number of individuals who answered "yes" or "vanilla" or "purple" depending on what the question was. In short, the rationale is counts for a dichotomous classification.

When we use a binomial statistical model, n is known and π is the unknown parameter. So if we are really being fussy "the" binomial statistical model is wrong. There is a different binomial statistical model for each different n.

The PMF of the $\mathrm{Bin}(n, p)$ distribution is

$$f_\pi(x) = \binom{n}{x} \pi^x (1 - \pi)^n, \qquad x = 0, 1, \ldots, n,$$

where the symbol called a *binomial coefficient* is defined as

$$\binom{n}{x} = \frac{n!}{x!(n - x)!}$$

where $n!$ (n factorial) is the product of the numbers from 1 to n and by definition $0! = 1$. The sample space for this distribution is the set of integers from zero to n. The parameter space is $[0, 1]$ as for the Bernoulli distribution.

The notation becomes ambiguous when $\pi = 0$ or $\pi = 1$ (because 0^0 is undefined), but we can work out what it should be using the rationale. If $\pi = 0$, then all of the X_i are equal to zero (with probability one), so Y is zero with probability one. And similarly for $\pi = 1$.

$$f_0(x) = \begin{cases} 1, & x = 0 \\ 0, & x \neq 0 \end{cases}$$

$$f_1(x) = \begin{cases} 1, & x = n \\ 0, & x \neq n \end{cases}$$

We say these two cases are *degenerate*. The random variable X is not really random but rather constant (it has only one value with certainty, that is, with probability one). But mathematically, certainty is just a special case of randomness (probability one), so we call these *constant random variables*. (If we adopt the convention $0^0 = 1$, then the general formula works for these special cases.)

If X has the $\mathrm{Bin}(n, \pi)$ distribution, then

$$\begin{aligned} E(X) &= n\pi \\ \mathrm{var}(X) &= n\pi(1 - \pi) \\ \mathrm{sd}(X) &= \sqrt{n\pi(1 - \pi)} \end{aligned}$$

Addition rule: the sum of independent binomial random variables with the same parameter value (but possibly different sample sizes) is again binomial with the same parameter value and sample size that is the sum of the separate sample sizes: if $X_i \sim \mathrm{Bin}(n_i, \pi)$ for $i = 1, 2, \ldots, k$, then $\sum_i X_i \sim \mathrm{Bin}(n_1 + \cdots + n_k, \pi)$. The special case where $n_1 = \cdots = n_k = 1$ is the addition rule for Bernoulli random variables: if $X_i \sim \mathrm{Ber}(\pi)$ for $i = 1, 2, \ldots, k$, then $\sum_i X_i \sim \mathrm{Bin}(n, \pi)$.

The Poisson Distribution

As noted above, the section title is a misnomer, since "Poisson" denotes a parametric statistical model (a family of distributions) so there is not just one Poisson distribution so the "the" in "the Poisson distribution" is wrong. To be pedantically correct, we should say the Poisson family of distributions or the Poisson statistical model.

The Poisson family of distributions has one parameter μ and its parameter space is the closed interval $[0, \infty)$. Parameter values range from zero (which is a possible parameter value) upwards with no upper bound. The Poisson distribution having parameter μ is abbreviated $\mathrm{Poi}(\mu)$.

The PMF of the $\mathrm{Poi}(\mu)$ distribution is

$$f_\mu(x) = \frac{\mu^x}{x!} e^{-\mu}, \qquad x = 0, 1, 2, \ldots,$$

where the symbol $x!$ (x factorial) was defined in the preceding section. The sample space for this distribution is the set of nonnegative integers.

If X has the $\mathrm{Poi}(\mu)$ distribution, then

$$\begin{aligned} E(X) &= \mu \\ \mathrm{var}(X) &= \mu \\ \mathrm{sd}(X) &= \sqrt{\mu} \end{aligned}$$

The parameter of the distribution is both the mean and the variance.

The notation becomes ambiguous when $\mu = 0$ (because 0^0 is undefined), but what the distribution should be can be worked out by taking limits (which involves calculus) or by using the rationale (which we haven't covered yet). The $\mathrm{Poi}(0)$ distribution is the distribution of the constant random variable that is always equal to zero (the same as the binomial distribution for $\pi = 0$). (If we adopt the convention $0^0 = 1$, then the general formula works for this special case.)

The Poisson distribution has two very closely related rationales.

- If n is very large and π is very small and $n\pi$ is moderate sized, then the $\mathrm{Bin}(n, \pi)$ distribution is very close to the $\mathrm{Poi}(n\pi)$ distribution in the sense both distributions give almost the same probability for each event. Since the two distributions have different sample spaces, this means the Poisson approximation must give nearly zero probability to all outcomes greater than n.

  As an example (that won't appear in the rest of the course), what is the probability of the number of winners of a lottery? This depends on the probability π of an individual ticket winning and on the

number n of tickets sold. For all lotteries π is very small (millions or billions to one) and n is very large (many millions of tickets sold) and $n\pi$ is moderate sized. In most drawings there are zero, one, or two winners. More than two winners is rare. So the Poisson distribution with parameter $\mu = n\pi$ is a very good approximation here.

Suppose the expected number of winners of a lottery is 0.5 (it often happens that when the jackpot is low that not many tickets are sold relative to the probability of winning). Then the probabilities of each number of winners are

```
x <- 0:6
p <- dpois(x, 0.5)
p <- round(p, 5)
names(p) <- x
p
```

```
##       0       1       2       3       4       5       6
## 0.60653 0.30327 0.07582 0.01264 0.00158 0.00016 0.00001
```

- A *spatial point process* is a random pattern of points, both the number of points and the locations of points being random. A *Poisson* point process has the following assumptions:
  - it is impossible to have more than one point at the same location,
  - no location has a nonzero probability of having a point,
  - counts of points in non-overlapping regions are independent random variables, and
  - the total number of points is finite with probability one.

  The dimension of the space in which the process lives is irrelevant. It can be one-dimensional, two-dimensional, three-dimensional, whatever.

  The second assumption is not problematic. Any continuous random variable has the same property: the probability of any particular value is zero. This seems strange because we don't usually take real numbers seriously, something taking an infinite number of digits to specify. When we say that the probability of observing X and it agreeing with a prespecified x to an infinite number of decimal places, it no longer seems so weird to say that has probability zero.

  The reason why we call this a Poisson process, even though there is nothing in the assumptions about the Poisson distribution, is that it can be shown that every count of points in any region has a Poisson distribution.

  If the expected number of points in a region is proportional to the length of the region in one dimension, area in two dimensions, volume in three dimensions, hypervolume in four dimensions, and so forth, then we say we have a *homogeneous* Poisson process. Otherwise inhomogeneous.

Between these concepts we get the rationale for the Poisson distribution: either approximation to the binomial distribution or counts in a Poisson process. In either case, Poissonness ultimately arises from independence. The independence in the binomial distribution comes from its rationale: sum of IID Bernoulli random variables. The independence in the Poisson process is in the assumption that counts of points in non-overlapping regions are independent.

Somewhat sloppily, we can say that whenever we have a count random variable and when each thing counted has nothing whatsoever to do with any other thing counted, then the count has a Poisson distribution. If we need to be more careful, we can refer details of Poisson approximation and Poisson processes.

The rationale of the Poisson distribution is rather complicated, something that no human ever thought of before 1837. But it can be dumbed down to be the distribution of any count variable when the things being counted are statistically independent. What is the distribution of number of clicks of a Geiger counter in a specified time interval? What is the distribution of the number of anthills in your back yard? What is the distribution of the number of raisins in a box of raisin bran? What is the number of stars in a specified

region of the sky? The answer to all of these questions is Poisson with some mean (a different mean for each question, of course). Poisson will be our go to distribution in this course. It is the default distribution for count variables. Every other distribution we use will be related to the Poisson distribution (including the binomial distribution).

The assumption the data have a Poisson distribution can fail if the independence assumption fails. What is the distribution of the number of people walking by a certain location on Northrop mall during a certain time period? The parameter of the Poisson distribution (mean number of people) will vary with both the location and the time period, but this is not a problem. A Poisson distribution can have any mean. But people are gregarious. They tend to walk in groups. The count of groups may well be Poisson, but the count of individuals in those groups may well not be. The groups are a failure of stochastic independence. Agresti mentions failure of independence in Section 1.2.4, but we will ignore this until we get to the notes accompanying Section 4.7 in Agresti.

Addition rule: the sum of independent Poisson random variables is again Poisson. (This is false if the word "independent" is omitted.) The mean of the sum is the sum of the means.

The Multinomial Distribution

As noted above, the section title is a misnomer, since "multinomial" denotes a parametric statistical model (a family of distributions) so there is not just one multinomial distribution so the "the" in "the multinomial distribution" is wrong. It is also wrong for another reason: like the binomial distribution the multinomial distribution has a sample size n, which is known, so there is a different multinomial statistical model for each different n.

Suppose we have n IID individuals, and we classify them into k categories (every individual goes into exactly one of the k categories). Define Y_1, Y_2, ..., Y_k be the numbers of individuals in these categories. Then the random vector $Y = (Y_1, Y_2, \ldots, Y_k)$ has the multinomial distribution with PMF

$$f_\pi(y) = \binom{n}{y_1, y_2, \ldots, y_k} \pi_1^{y_1} \pi_2^{y_2} \cdots \pi_k^{y_k}, \qquad y \in S$$

where the symbol called a *multinomial coefficient* is defined as

$$\binom{n}{y_1, y_2, \ldots, y_k} = \frac{n!}{y_1! \, y_2! \cdots y_k!}$$

where the symbol $x!$ (x factorial) was defined in the section about the binomial distribution, and where the sample space S is the of all vectors y having nonnegative-integer-valued components that sum to n. ($Y_1 + Y_2 + \cdots + Y_k = n$ with probability one because every individual is classified in exactly one category.)

The number of categories k is the *dimension* of the multinomial random vector. The number of individuals classified n is the *sample size* of the multinomial distribution.

We abbreviate this distribution Multi(n, π).

The parameter space of this distribution is the set of all probability vectors, the set of all vectors π whose components are nonnegative and sum to one. Thus the parameters satisfy the equality constraint $\pi_1 + \pi_2 + \cdots + \pi_k = 1$.

The notation becomes ambiguous when any component of the parameter vector is zero (because 0^0 is undefined), but we can work out what it should be by taking limits (using calculus). This tells us we should interpret $0^0 = 1$ because $\pi_i^0 = 1$ for all $\pi_i > 0$ by convention.

Multinomial distributions with some components of π equal to zero are only *partially degenerate*. Some components are not really random ($\pi_i = 0$ implies $Y_i = 0$ with probability one and $\pi_i = 1$ implies $Y_i = n$ with probability one). They are constant random variables. But other components are really random (the Y_i such that $0 < \pi_i < 1$). The only completely degenerate multinomial distributions are those having π with only one nonzero component (which must be equal to n in order that the components sum to n).

If Y has the Multi(n, π) distribution, then

$$\begin{aligned} E(Y_i) &= n\pi_i \\ \operatorname{var}(Y_i) &= n\pi_i(1 - \pi_i) \\ \operatorname{cov}(Y_i, Y_j) &= -n\pi_i\pi_j, \qquad i \neq j \end{aligned}$$

(Note that the formula for $\operatorname{cov}(Y_i, Y_j)$, $i \neq j$ does not work for the case $i = j$.)

If we define a diagonal matrix Π whose diagonal components are the corresponding components of π, then we can write these in matrix notation as

$$\begin{aligned} E(Y) &= n\pi \\ \operatorname{var}(Y) &= n(\Pi - \pi\pi^T) \end{aligned}$$

Addition rule: the sum of independent multinomial random vectors having the same dimension and same parameter vector is again multinomial: if $X_i \sim$ Multi(n_i, π) for $i = 1, ..., k$, then $\sum_i X_i \sim$ Multi$(n_1 + \cdots + n_k, \pi)$. (This is false if the word "independent" is omitted.)

Statistical Models for Categorical Data

We have already met all of our statistical models for categorical data. They are

- Poisson,
- multinomial, and
- product multinomial

The last we haven't officially met yet, but it is implied by the multinomial and the multiplication rule for independent random vectors.

We will call these our three different *sampling models* for categorical data. Much will be made of the relationship between them in this course.

- The *Poisson sampling model* says the category counts are independent Poisson random variables. Then the sample size is the sum of these random variables, which has a Poisson distribution.

  This model applies when the sample size is not determined in advance and can be considered Poisson. An example would be if you interviewed people passing on Northrop mall, interviewed for one hour, and the sample size is just however many people you managed to interview in that hour.

- The *multinomial sampling model* says the category counts are a multinomial random vector. The sample size is fixed in advance of collecting data. It is the multinomial sample size.

  This model applies when the sample size is determined in advance and can be considered constant. An example would be if you interviewed people passing on Northrop mall, interviewing until you got the predetermined sample size and then stopped.

- The *product multinomial sampling model* says the category counts are several independent multinomial random vectors.

  This model applies when several sample sizes for subgroups are determined in advance and can be considered constant. An example would be if you decided to have 100 males and 100 females in your data, interviewed people passing on Northrop mall, interviewed until you got the predetermined sample sizes, and stopped interviewing in each subgroup when the predetermined sample size was reached. (If 100 female interviews is reached first, then you stop interviewing females and continue interviewing males until 100 have been interviewed, and then stop.)

You have already met the latter two sampling models in intro stats, and we reviewed this subject

More on sampling schemes after we introduce some more theory.

One more issue is the warning issued by R function `chisq.test`. The P-value is based on asymptotic approximation, and it is warning that the approximation is not very good. What is asymptotic approximation? That is what we turn to now.

The Univariate Normal Distribution

As noted above, the section title is a misnomer, since "normal" denotes a parametric statistical model (a family of distributions) so there is not just one univariate normal distribution so the "the" in "the univariate normal distribution" is wrong.

Everyone should be familiar with the univariate normal distribution from intro stats courses and also familiar with the central limit theorem (CLT), which provides the rationale for the univariate normal distribution.

As we shall see throughout the course, many quantities, including all point estimators of interest, have approximately normal distributions for "large sample size" (the scare quotes are there to remind us that the theorem does not tell us how large sample size has to be to get good normal approximation).

The normal distribution has a PDF but the PDF will not be useful in this course. Knowing the PDF would not help us calculate any probabilities or expectations. We will have to use the computer for that.

The parameters of the univariate normal distribution are the mean and standard deviation (or variance can replace standard deviation). The mean can be any real number. The standard deviation must be positive. So the sample space is the set of all vectors (μ, σ) such that $\sigma > 0$.

We abbreviate the normal distribution with mean μ and variance σ^2 as $\text{Normal}(\mu, \sigma^2)$.

Taking the limit as $\sigma \to 0$ while μ is held fixed gives the discrete distribution concentrated at one point μ, thus the distribution of a constant random variable. But this distribution is often not considered a normal distribution; normal distributions are continuous but this limit is discrete.

Both the binomial and Poisson distributions have normal approximations, which are good for some values of the parameters. We already know the binomial distribution is not well approximated by the normal distribution for all values of the parameters because of its Poisson approximation.

These normal approximations come from the CLT. Normal approximation is good when sample size is large. But what is "sample size" for the Poisson distribution? It doesn't have one. But it does have an addition rule, which gives us a "sample size" of sorts.

The binomial distribution is approximately normal when its sample size is "large" (again scare quotes). But how large n has to be depends on what π is. A rule of thumb is that we need $n\pi$ and $n(1 - \pi)$ to be greater than 5. But like all rules of thumb, this is dumbed down to the point of being wrong. Asymptotics does not work like people want.

- It is not all or nothing. There is no sharp dividing line between good and bad normal approximation or Poisson approximation or chi-square approximation or any other kind of asymptotic approximation.
- The larger n is, the better the approximation. But the transition from bad to good is gradual.
- How good the approximation is also depends on what question you are asking (what probability or expectation you are trying to approximate). The approximation may work well in the middle of the distribution but very badly far out in the tails of the distribution.
- How good the approximation is also depends on details about the "population" (in scare quotes) distribution. Skewness or heavy tails or other features of the "population" distribution will require larger n than otherwise.
- Approximation error in asymptotic approximation is absolute not relative. Good asymptotic approximation means small probabilities are estimated to be small. If the true probability is 0.001 and the asymptotic approximation says 10^{-10}, that is what we mean by good approximation. There is enormous relative error, but small absolute error (the ratio is large but the difference small).

- How well asymptotic approximation works can be checked by simulation. But short of that, we never really know how good the approximation is.

We already know that if n is large but $n\pi$ is not large, we get good Poisson approximation, and if n is large and $n\pi$ is large (but not too large as we shall soon see), we get good normal approximation. But there is no sharp dividing line.

When X has the $\text{Bin}(n, \pi)$ distribution, $n - X$ has the $\text{Bin}(n, 1 - \pi)$ distribution. Hence when n is large and $n(1 - \pi)$ is moderate sized, $n - X$ has good Poisson approximation.

We know that the sum of IID $\text{Poi}(\mu)$ random variables is $\text{Poi}(n\mu)$. The CLT says this is approximately normal for sufficiently large n. Hence the Poisson distribution is approximately normal when the mean is large. The normal approximation is bad when the mean is not large. Again, there is no sharp dividing line where the normal approximation goes from bad to good.

Addition rule: see under the multivariate normal distribution.

The (univariate) central limit theorem if X_1, X_2, ... are IID random variables having mean μ and variance σ^2 and

$$\overline{X}_n = \frac{1}{n} \sum_{i=1}^{n} X_i$$

then

$$\overline{X}_n \approx \text{Normal}\left(\mu, \frac{\sigma^2}{n}\right)$$

and if

$$S_n = \sum_{i=1}^{n} X_i$$

then

$$S_n \approx \text{Normal}\left(n\mu, n\sigma^2\right)$$

There are central limit theorems that do not require IID but they are not needed in this class except for MCMC central limit theorems, which will be discussed when we talk about Bayesian inference.

The Multivariate Normal Distribution

As noted above, the section title is a misnomer, since "normal" denotes a parametric statistical model (a family of distributions) so there is not just one multivariate normal distribution so the "the" in "the multivariate normal distribution" is wrong.

The *standard* univariate normal distribution is the one having mean zero and standard deviation one (hence variance one).

The *standard* multivariate normal distribution is the one having mean vector zero (the zero vector) and variance matrix the identity matrix (ones on the diagonal, zeros off the diagonal). This says components of the normal distribution have variance one and covariance zero.

It is a very special property of the multivariate normal distribution that covariance zero implies independence (no other distribution has this property). Thus components of a standard multivariate normal random vector are IID univariate normal.

Linearity rule: any multivariate linear function of a multivariate normal random vector is another multivariate normal random vector. If X is a random vector, a is a constant (nonrandom) vector, and B is a constant (nonrandom) matrix, then

$$Y = a + BX$$

is another random vector, assuming the dimensions of a, B, and X are such that the vector addition and matrix multiplication in the formula make sense (the dimension of a is equal to the row dimension of B, and the dimension of X is equal to the column dimension of B). The preceding sentence is true for any

random vector X (normally distributed or not normally distributed). The linearity rule says that if X has a multivariate normal distribution then so does Y.

Depending on which prerequisites for the course you may or may not have had, you may or may not have been exposed to matrix multiplication. This is usually taught in calculus or linear algebra or theoretical probability and statistics. But even if you don't know, the computer does. The R expression `a + B %*% x` calculates $a + Bx$.

The parameters of the multivariate normal distribution are the mean vector and the variance matrix (there is no multivariate analog of standard deviation). Any vector can be a mean vector, but a variance matrix must be positive semidefinite, which we will not need to define. Any variance matrix, if computed correctly, will have this property. Thus the parameter space is the set of all parameter vectors (μ, Σ), where μ is the mean vector and Σ is the variance matrix, and the the notation (μ, Σ) means that we are thinking of the parameters as combined into one thing, which can abstractly be thought of as a vector (in abstract linear algebra, vectors can be anything things that satisfy the axioms for a vector space, any things that can be added and multiplied by scalars, and the addition and multiplication operations satisfy the axioms; so μ is a vector (of course) and Σ is a vector because matrices are a special case of vectors because they can be added and multiplied by scalars, and vector spaces can be combined as product spaces, so we can put μ and Σ together to make an abstract vector; we need this "abstract nonsense" to think of the set of possible (μ, Σ) values as a subset of a vector space, the *parameter space* of the multivariate normal family of distributions).

We abbreviate the multivariate normal distribution with mean vector μ and variance Σ as Normal(μ, Σ).

Repeating what what said in the preceding section, as we shall see throughout the course, many quantities, including all point estimators of interest, have approximately normal distributions for "large sample size" (the scare quotes are there to remind us that the theorem does not tell us how large sample size has to be to get good normal approximation). Random variables have approximate univariate normal distributions. Random vectors have approximate multivariate normal distributions.

The normal distribution has a PDF if its variance matrix is strictly positive definite, but we will not need to define that property. The variance matrices that arise in normal approximation of the sampling distributions of estimators, if computed correctly, will have this property. But the PDF will not be useful in this course. Knowing the PDF would not help us calculate any probabilities or expectations. We will have to use the computer for that.

Multivariate normal random vectors Y that do not have PDF because they do not have strictly positive variance matrices can always be written in the form $Y = a + BX$ as in the linearity rule, where X is multivariate normal and does have a PDF. But since we are not going to use normal PDF anyway, this is of little importance to us.

The rationale of the multivariate normal distribution is the multivariate CLT, which works just like the univariate CLT except for random vectors rather than random scalars.

Addition rule: this follows from the linearity rule because addition is a linear operation. If $X_i \sim \text{Normal}(\mu_i, \Sigma_i)$ for $i = 1, \ldots, k$, and these random variables are independent, then $\sum_i X_i \sim \text{Normal}(\mu_1 + \cdots + \mu_k, \Sigma_1 + \cdots + \Sigma_k)$. In particular if $X_1, \ldots, X_k$ are IID Normal(μ, Σ), then $\sum_i X_i \sim \text{Normal}(n\mu, n\Sigma)$.

The multivariate central limit theorem: if X_1, X_2, ... are IID random vectors having mean vector μ and variance matrix Σ and

$$\overline{X}_n = \frac{1}{n} \sum_{i=1}^{n} X_i$$

then

$$\overline{X}_n \approx \text{Normal}\left(\mu, \frac{\Sigma}{n}\right)$$

and if

$$S_n = \sum_{i=1}^{n} X_i$$

then

$$S_n \approx \text{Normal}\,(n\mu, n\Sigma)$$

The Chi-Squared Distribution

If X is a multivariate normal random vector with mean vector μ and variance matrix Σ that is strictly positive definite, then Σ is invertible, meaning it has a matrix inverse Σ^{-1} and

$$T = (X - \mu)^T \Sigma^{-1} (X - \mu)$$

is a random variable (scalar not vector) having the *chi-squared distribution* with *degrees of freedom* that is the dimension of X.

In the special case that X is a *standard* normal random vector, μ is the zero vector and Σ is the identity matrix Id, which is its own inverse, so

$$T = X^T \text{Id}\, X = X^T X = \sum_{i=1}^{k} X_i^2$$

where k is the dimension of X. Thus the chi-squared distribution with k degrees of freedom is often defined to be the distribution of the sum of squares of k IID standard normal random variables.

But our more general definition says why the chi-squared distribution is interesting. When X has close to a multivariate normal distribution, then T has close to a chi-squared distribution. It will turn out that test statistics for hypothesis tests involving multiple parameters will be approximately chi-squared distributed under the null hypothesis.

An abbreviation for the chi-squared distribution with k degrees of freedom is $\text{ChiSq}(k)$.

You will have already seen this in intro stats discussion of so-called chi-squared tests. But as we shall see throughout this course, those tests discussed in intro stats are just the tip of the iceberg. We will meet a great many more tests in this course having test statistics with asymptotic chi-squared distributions.

Unlike binomial, Poisson, multinomial, and normal the chi-squared distribution does not have any unknown parameters. However, the "the" in "the chi-squared distribution" is still wrong because there is a different chi-squared distribution for each degrees of freedom, and the degrees of freedom can be any positive integer. Actually, as you learn in theory courses, the chi-square distribution actually makes sense for non-integer degrees of freedom (it can be any positive real number). But we won't need that for our use of chi-square approximation of the distribution of test statistics.

When we use the chi-square distribution, we know the degrees of freedom so there are no unknown parameters.

How chi-squared distributions arise in hypothesis testing will be a topic throughout the course.

A Little Bit of Theory

The Linearity Rule and the Delta Method

As in the linearity rule for multivariate normal random vectors, if

$$Y = a + BX$$

with X and Y random vectors and a and B constant (nonrandom) objects, a a vector and B a matrix,

$$\begin{aligned} E(Y) &= a + BE(X) \\ \text{var}(Y) &= B\,\text{var}(X)B^T \end{aligned}$$

If X has mean vector μ and variance matrix Σ, then we can rewrite this

$$\begin{aligned} E(Y) &= a + B\mu \\ \operatorname{var}(Y) &= B\Sigma B^T \end{aligned}$$

In the special case where X and Y are both random variables (scalar-valued) the matrix multiplication becomes ordinary multiplication (of scalars) and the vector addition becomes ordinary addition (of scalars). Then we write

$$Y = a + bX$$

where now a and b are constant (nonrandom) scalars, and our formulas for mean and variance of Y become

$$\begin{aligned} E(Y) &= a + bE(X) \\ \operatorname{var}(Y) &= b^2 \operatorname{var}(X) \end{aligned}$$

or

$$\begin{aligned} E(Y) &= a + b\mu \\ \operatorname{var}(Y) &= b^2\sigma^2 \end{aligned}$$

where X has mean μ and variance σ^2.

The *delta method* says the linearity rule holds approximately for changes of variable that are approximately linear. If f is a vector-to-vector differentiable function having Jacobian matrix B and X is a random vector that is close in distribution to a constant vector θ, then

$$f(X) \approx f(\theta) + B(X - \theta)$$

If X is approximately multivariate normal with mean vector θ and variance matrix Σ, then $f(X)$ is approximately normal with mean vector $f(\theta)$ and variance matrix $B\Sigma B^T$. The form of the variance matrix arises from the linearity rule and the properties of multivariate differentiation.

When we use the delta method we either won't notice at all because it is entirely hidden inside some R function (typically R generic functions `predict` and `summary`) or we will at least not need to know multivariable calculus because we can use R function `jacobian` in R package `numDeriv` to calculate Jacobian matrices for us. So if `B` is the Jacobian matrix and `Sigma` is the variance matrix

```
B %*% Sigma %*% t(B)
```

is the R code to calculate $B\Sigma B^T$. Often R function `vcov` can be used to obtain Σ.

Conditional Probability

Conditional probability is just like unconditional probability except it depends on the observed value(s) of some random variable(s).

If there are two random variables X and Y, and you observe X before you observe Y, then what you find out about X may change what you expect to find out about Y when you observe it.

If the (joint) probability mass function (PMF) for X and Y is f, then the conditional PMF of Y given X is proportional to $f(x, y)$ thought of as a function of y for fixed x. So when we are conditioning on x, the variable x is no longer playing the role of a random variable. It is fixed at its observed value throughout the discussion. Thus the PMF of Y given X is

$$f(y \mid x) = c \cdot f(x, y)$$

where the constant c is chosen to make the left-hand side a probability distribution thought of as a function of y for fixed x, that is,

$$\begin{aligned} f(y \mid x) &\geq 0, \qquad \text{for all } y \\ \sum_y f(y \mid x) &= 1 \end{aligned}$$

The PMF $f(\cdot|x)$ depends on x, but x is not an argument of the function, so it is really more like a parameter. In fact, there is really no difference between a parametric family of probability distributions f_θ and a conditional distribution. In both case the distribution depends on something that is not considered an argument.

For this reason, Bayesian statisticians always write parametric families as conditional distributions. They write $f(x \mid \theta)$ instead of $f_\theta(x)$, but that is getting a bit ahead of ourselves.

The main point of this section is that conditioning changes the distribution, and the conditioning variable(s) is/are treated as *fixed* in the conditional distribution.

The Relationship of the Various Sampling Schemes

Here we are expanding on Agresti Section 1.2.5.

Our three sampling schemes — Poisson, multinomial, and product multinomial — are related by conditional probability. All of the theorems in this section are proved in the notes on sampling schemes.

> **Theorem** If we start with Poisson sampling (the data vector Y has independent Poisson distributed components) and condition on the event $\text{sum}(Y) = n$, then we get multinomial sampling with n the multinomial sample size and the relation $\mu = n\pi$ between the Poisson parameter vector μ and the multinomial parameter vector π.
>
> Conversely, if we start with multinomial sampling and then let the multinomial sample size be a Poisson random variable, we get Poisson sampling.

Proving this theorem is two homework problems when I teach Stat 5101. It is beyond the scope of this course (we give the proof in the notes on sampling schemes but won't go over it, similarly for all proofs in this section).

Let Y be the data vector, and let $\mathcal{A}$ be a partition of its index set. (Partition in math means that $\mathcal{A}$ is a set of sets — a set whose elements are themselves sets — and every element of the set which is partitioned is found in exactly one element of $\mathcal{A}$.)

For each $A \in \mathcal{A}$, let Y_A denote the subvector of Y whose components are Y_i for $i \in A$ (for much more on subvectors and sampling schemes, see the notes on sampling schemes). Then we say Y has a product multinomial distribution for the partition $\mathcal{A}$ if the random vectors Y_A for $A \in \mathcal{A}$ have independent multinomial distributions. By the product rule, the joint distribution is the product of the marginal distributions

$$f(y) = \prod_{A \in \mathcal{A}} f_A(y_A)$$

and each of these marginal distributions is multinomial. The multinomial sample sizes for these multinomial distribution are, of course, the sum of the counts of these multinomial random vectors

$$n_A = \sum_{i \in A} Y_i, \qquad A \in \mathcal{A}$$

This may seem very abstract, but the point is it does not matter how you partition the data vector. For any partition, there is a product multinomial distribution. In two-way tables, the usual way to partition is by rows or columns (one or the other but not both). But the usual way is not the only way. We need the abstractness to allow for any partition.

> **Theorem** If we start with Poisson sampling (the data vector Y has independent Poisson distributed components) and condition on the events $\text{sum}(Y_A) = n_A$, $A \in \mathcal{A}$, then we get product multinomial sampling with n_A, $A \in \mathcal{A}$, the product multinomial sample sizes and the relation $\mu_i = n_A \pi_i$ for $i \in A$ between the Poisson parameter vector μ which has components μ_i and the product multinomial parameter vectors which have components π_i.
>
> Conversely, if we start with product multinomial sampling and then let the product multinomial sample sizes be a Poisson random variables, we get Poisson sampling.

Theorem If we start with multinomial sampling (the data vector Y has a multinomial distribution with sample size n) and condition on the events $\operatorname{sum}(Y_A) = n_A$, $A \in \mathcal{A}$, then we get product multinomial sampling with n_A, $A \in \mathcal{A}$, the product multinomial sample sizes and the relation $n\pi_i = n_A\psi_i$ for $i \in A$ between the multinomial parameter vector π which has components π_i and the product multinomial parameter vectors which have components ψ_i.

There is a converse, but we won't bother to state it (too complicated).

Now things get worse. Readers probably want to skip to the summary paragraph at the end of this section. The details of the theorems above and below do not matter for data analysis, just the summary. The details are there in case anyone ever asks you what you are talking about.

If $\mathcal{A}$ and $\mathcal{B}$ are two different partitions of the index set of the data vector, we say that $\mathcal{B}$ is *finer* than $\mathcal{A}$ if every $B \in \mathcal{B}$ is contained in some $A \in \mathcal{A}$. (Hence by the nature of partitions every $B \in \mathcal{B}$ is contained in exactly one $A \in \mathcal{A}$.)

When $\mathcal{B}$ is *finer* than $\mathcal{A}$, we can also say this as $\mathcal{A}$ is *coarser* than $\mathcal{B}$.

Theorem If $\mathcal{A}$ and $\mathcal{B}$ are partitions with $\mathcal{A}$ coarser than $\mathcal{B}$ and we start with product multinomial sampling for the partition $\mathcal{A}$ and condition on the events $\operatorname{sum}(Y_B) = n_B$, $B \in \mathcal{B}$, then we get product multinomial sampling for the partition $\mathcal{B}$ with the relation $n_A\pi_i = n_B\psi_i$ for $i \in B \subset A$ between the two product multinomial parameter vectors.

There is a converse, but we won't bother to state it (too complicated).

Summary On the list

- Poisson sampling
- multinomial sampling
- product multinomial sampling
- product multinomial sampling with a finer partition

we go down in the list by conditioning and up in the list by "unconditioning" (in scare quotes because this is not a technical term). The theorems in which the converse is stated say what that means. Otherwise we have just what conditional distribution means. We always have

$$\text{conditional} = \frac{\text{joint}}{\text{marginal}}$$

so we always also have

$$\text{joint} = \text{conditional} \times \text{marginal}$$

and the latter is what "unconditioning" means.

In all cases, no matter where we start, when we condition on more stuff we get the same thing as if we had decided to fix that stuff in advance.

Whether you consider this deep or trivial is up to you.

We will find that most of the time it doesn't matter what sampling scheme we "assume" because the same asymptotic statistical inference results.

If we could do exact statistical inference, the results would be different because the different sampling schemes are actually different. But usually we have no idea how to obtain exact sampling distributions and must depend on asymptotic approximations.

And these asymptotic approximations give the same results, if we are careful (more on this later).

Bayesian inference is exact and does differ for the different sampling schemes.

Likelihood Inference

Likelihood Function

For a parametric family of distributions with PMF f_θ, the *likelihood* for the model when the observed data are x is just PMF

$$L_x(\theta) = f_\theta(x)$$

with the roles of the parameter and data interchanged. In the PMF the data x is the variable and the parameter θ is fixed. In the likelihood (left-hand side) the data x is fixed at its observed value and the parameter θ is the variable.

You may think this is trivial, but everybody in statistics observes this pedantic distinction. You can really see the difference when it comes to calculus. The derivative of the likelihood function L requires us to differentiate with respect to θ (because that is the variable in that function). The derivative of the PMF f_θ requires us to differentiate with respect to x (because that is the variable in that function).

Log Likelihood Function

For mathematical convenience, the log likelihood is often preferred.

$$l_x(\theta) = \log L_x(\theta)$$

Modifications

For reasons that cannot be fully understood until we are done with both frequentist likelihood inference and Bayesian inference (all of which is likelihood-based) it makes no difference whatsoever to statistical inference (frequentist or Bayesian) if

- additive terms that do not contain the parameter(s) are dropped from the log likelihood
- multiplicative terms that do not contain the parameter(s) are dropped from the likelihood

For example, for the binomial distribution, the likelihood is

$$L_x(\pi) = \binom{n}{x} \pi^x (1 - \pi)^{n - x}$$

but we are allowed to drop the multiplicative term that does not contain the parameter π obtaining

$$L_x(\pi) = \pi^x (1 - \pi)^{n - x}$$

(either likelihood function is as good as the other). And the log likelihood is

$$l_x(\pi) = \log \binom{n}{x} + x \log(\pi) + (n - x) \log(1 - \pi)$$

but we are allowed to drop the additive term that does not contain the parameter π obtaining

$$l_x(\pi) = x \log(\pi) + (n - x) \log(1 - \pi)$$

"Principle" (in Scare Quotes) of Maximum Likelihood

You might see in places where they dumb down things to the point of being wrong that the maximizer of the likelihood is a good point estimate of the parameter.

This is not a "principle" you should find in any statistics book.

There are statistical models such that the more data you have the worse the maximum likelihood estimator (MLE) is.

What is true is that for statistical models satisfying "suitable regularity conditions" (and neither we nor Agresti go into what that is) we have the following results.

1. the MLE is a consistent and asymptotically normal (CAN) estimator of the parameter, and no other estimator can do better asymptotically than the MLE, except perhaps at a negligible set of true unknown parameter values.

2. the MLE is a root-n-consistent estimator, that is,

$$\sqrt{n}(\hat{\theta}_n - \theta)$$

converges in distribution to a mean-zero normal distribution (a multivariate normal distribution if θ is a vector), where $\hat{\theta}_n$ is the MLE for sample size n.

3. (under somewhat weaker regularity conditions than for item 1) if $\tilde{\theta}_n$ is any root-n-consistent estimator, that is,

$$\sqrt{n}(\tilde{\theta}_n - \theta)$$

converges to any distribution whatsoever, and we define $\hat{\theta}_n$ to be the nearest *local* maximum of the likelihood to $\tilde{\theta}_n$, then $\hat{\theta}_n$ is again CAN and best possible (except perhaps for a negligible set of parameter values).

4. The asymptotic variance in the asymptotic (normal) distribution of the MLE is inverse Fisher information, either observed or expected.

   - Observed Fisher information is minus the Hessian (second derivative) matrix of the log likelihood, evaluated at the MLE.
   - Expected Fisher information is the expectation of observed Fisher information (treating the data as random).

5. Log-linear models for categorical data analysis that are full exponential families (more on that later) *always* satisfy the regularity conditions for item 1.

6. Curved submodels of those in item 4 (for example Agresti Sec. 1.5.4) *always* satisfy the regularity conditions for item 3 (but not necessarily for item 1).

Item 4 means that, so long as we can write down the log likelihood and calculate two derivatives, we know the asymptotic distribution of the MLE (under "suitable regularity conditions"). Item 5 says that for most models used in this class (but not all), the MLE can be defined as the global maximizer of the log likelihood. Item 6 says that for all models used in this class, the MLE can be defined as the the nearest local maximum to a root-n-consistent estimator.

So there is no "principle" that says the global maximizer is best (or even exists). (http://www.stat.umn.edu/geyer/5102/examp/like.html#mix discusses a statistical model for which the global maximizer never exists because the supremum of the log likelihood is infinity. Unfortunately, that model is not categorical data analysis.) But at least for categorical data analysis we do have the estimator of item 6, which we can call the MLE and does have desirable properties.

Except all of this is as sample size goes to infinity. Nothing says the exact sampling distribution of the MLE (for the n we are at) is well approximated by the asymptotic distribution no matter what n is.

But Geyer (2013, DOI: 10.1214/12-IMSCOLL1001) following earlier authors shows that if the log likelihood is well approximated by a quadratic function, then the sampling distribution of the MLE is close to its asymptotic distribution (Agresti also mentions this). So that gives us some purchase on what we need to know about whether asymptotic approximation is good.

We can also use simulation (the so-called *parametric bootstrap*, more on this later) to check how good asymptotic approximation is, and also to make the approximation better if it isn't good already.

Agresti Section 1.3.2

One quibble. It so happens that the setting the first derivative of the log likelihood equal to zero and solving for the parameter seems to give the MLE $\hat{\pi} = x/n$. But this is sloppy. The derivative does not exist at the

boundary of the parameter space, and even if one uses one-sided derivatives, calculus does not say that the derivative is zero if the maximum occurs on the boundary. Thus to be careful, one needs better analysis (http://www.stat.umn.edu/geyer/5102/slides/s3.pdf slides 20, 21, 25, and 31).

Note that something is fishy about $\pi = 0$ and $\pi = 1$ anyway. In these cases the asymptotic variance is zero ($\sqrt{\pi(1 - \pi)} = 0$ in either case). So all the asymptotics says is

$$\sqrt{n}(\hat{\pi}_n - \pi) \xrightarrow{\mathcal{D}} 0$$

(the right-hand side is the distribution concentrated at zero). It doesn't tell us much that is useful. (More on this later.)

Agresti Section 1.3.3

Likelihood-based hypothesis tests come in three kinds.

- Likelihood Ratio Tests, also called Wilks tests.
- Wald Tests.
- Score Tests, also called Rao tests, also called Lagrange Multiplier tests (the latter name mostly used by economists).

These tests are all *asymptotically equivalent* in the sense that (under suitable regularity conditions) for very large sample sizes they will have nearly equal values of the test statistic and P-value for the same data. (The test statistic and P-value are random quantities when the data are considered random, but for the same data all three tests will have nearly the same test statistics and P-values. The difference between test statistics for any two of these tests will be negligible (for large n) compared to either test statistic itself.)

Thus asymptotics gives us no reason to choose among these.

Recommendations about which to use are based on mathematical convenience, pedagogical convenience, philosophical ideas, or simulations. Simulations, of course, must be based on one particular model (or perhaps a few models) and so cannot prove any general conclusions about all models.

Under mathematical convenience we have

- assuming that one can calculate MLE for both the null and alternative hypotheses, the likelihood ratio test statistic is
$$2[l(\hat{\theta}_{\text{alternative}}) - l(\hat{\theta}_{\text{null}})]$$
and the asymptotic distribution is chi-square with degrees of freedom that is the difference of dimensions of the models. So this is actually easiest if one can fit both null and alternative models.
- the score test requires only the MLE for the null model, not the alternative. However the test statistic is complicated to calculate. Thus it is most useful when the MLE for the alternative model is difficult or impossible to fit or when the user does not want to bother with fitting the alternative model.
- the Wald test requires only the MLE for the alternative model, not the null. However the test statistic is complicated to calculate. Thus it is most useful when the MLE for the null model is difficult or impossible to fit or when the user does not want to bother with fitting the null model.

In particular, R function `summary` computes lots of P-values for lots of tests all based on having fit only one model. All of the P-values are for tests of null hypotheses that set one of the coefficients equal to zero (and have alternative hypothesis that is the model that was fit). Thus these must all be Wald tests.

The most famous example of score tests (and one which was invented long before general score tests) is the Pearson Chi-Square test for categorical data analysis. So whenever we use that, we are using a score test.

Exact formulas for the test statistics for general Wald and Rao tests are given in the course notes on likelihood. But they are hard to apply except in simple special cases (and sometimes not even then). So mostly when

we do one or the other of these tests, we will have a computer do all the hard work (as will be seen in the notes for Section 1.4 in Agresti).

TL;DR There is no one right way to do a hypothesis test.

Coverage of Confidence Intervals

Because the data (in this course) are discrete (counts), only a finite set of data values contains almost all of the probability. It follows that the *actual coverage probability* of a confidence interval (no matter what the recipe is) cannot be a flat function. It cannot be 0.95 for all values of the parameter (or any other confidence level other than 0.95). As the parameter (say θ) moves from inside to outside (or vice versa) the confidence interval for some data value (say x) the coverage probability must jump by $f_\theta(x)$.

http://www.stat.umn.edu/geyer/5102/examp/coverage.html illustrates this for a variety of confidence intervals for the binomial distribution, those covered in Section 1.4 of Agresti plus some more.

TL;DR There is no one right way to do a confidence interval. Which is better than the others is a matter of opinion.

No confidence interval (recipe) for discrete data can be exact (actually achieve its nominal coverage for all values of the true unknown parameter).

Fuzzy confidence intervals (Geyer and Meeden, 2005, *Statistical Science*, 20, 358–387) are exact (actually achieve their nominal coverage for all values of the true unknown parameter) but are more complicated to use and harder to interpret and explain. We will spend a little time on them (5421 notes on binomial and 5421 notes on fuzzy) but not now.

Agresti Section 1.3.4

Frequentist statistical inference procedures come in trios

- hypothesis test
- confidence interval
- point estimate

Every hypothesis test procedure determines a confidence interval produced by "inverting" the test. And vice versa.

- Given a hypothesis test and a significance level α, the set of points θ_0 such that a test at level α with null and alternative hypotheses

  $$H_0 : \theta = \theta_0$$
  $$H_1 : \theta \neq \theta_0$$

  is a confidence interval for θ with coverage probability $1 - \alpha$. The confidence interval will be exact if the hypothesis test is exact. The confidence interval will be approximate (large n, asymptotic) if the hypothesis test is approximate (large n, asymptotic).

- Given a confidence interval with coverage probability $1-\alpha$, the hypothesis test with null and alternative hypotheses in the preceding item that accepts H_0 if the confidence interval covers θ_0 and rejects H_0 (and accepts H_1) if the confidence interval does not cover θ_0 has level α. The hypothesis test will be exact if the confidence interval is exact. The hypothesis test will be approximate (large n, asymptotic) if the confidence interval is approximate (large n, asymptotic).

- The same relationships hold between one-tailed tests and one-sided confidence intervals.

If one has a confidence interval procedure (that constructs confidence intervals with any coverage probability), then the *Hodges-Lehmann estimator* associated with this procedure is the point to which these confidence intervals shrink as the coverage probability goes to zero.

In this course we will always use MLE for frequentist point estimates (Bayes is different, more on that later).

But we see that we have three kinds of hypothesis tests (Wald, Wilks, Rao) and so three different confidence interval procedures. But they all give back the MLE as the Hodges-Lehmann estimator.

Agresti Section 1.4

For this section we move to specific notes on the binomial distributions.

Agresti Section 1.4.3

I have to disagree with Agresti here. When the MLE is on the boundary of the parameter space, it is true that the score and likelihood confidence intervals are not completely ridiculous, and the Wald interval is completely ridiculous (zero width).

However, when the true unknown parameter value is on the boundary of the parameter space, the "usual regularity conditions" for maximum likelihood are not satisfied. Thus, in this situation, none of the three tests, Wald, Wilks, or Rao, have any theoretical justification whatsoever.

Hence we recommend the intervals proposed by Geyer (2009, *Electronic Journal of Statistics*, **3**, 259–289) and illustrated on the web page about coverage of confidence intervals. For the binomial distribution these intervals for coverage $1 - \alpha$ are

- when $x = 0$, the confidence interval is $(0, 1 - \alpha^{1/n})$, and
- when $x = n$, the confidence interval is $(\alpha^{1/n}, 1)$.

These, at least, have a theoretical justification. For these data this interval is

```
alpha <- 0.05
n <- 25
c(0, 1 - alpha^(1 / n))
```

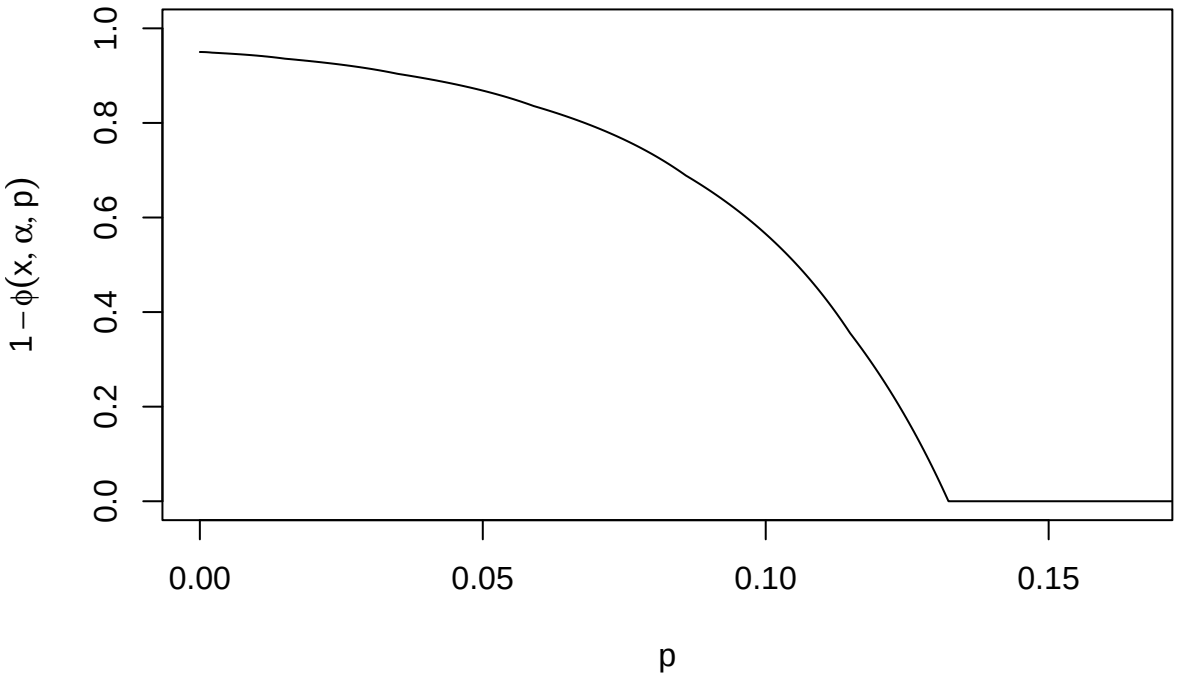
```
## [1] 0.0000000 0.1129281
```

Fuzzy intervals (Geyer and Meeden, 2005, *Statistical Science*, **20**, 358–387; notes on fuzzy for this course) also do the right thing (are exact-exact) for this case.

Here is the 95% fuzzy confidence interval for the data discussed by Agresti ($x = 0$, $n = 25$).

```
library(ump)
fci.binom(0, 25)
```

```
## 95 percent fuzzy confidence interval
## core is empty
## support is [0, 0.1322)
```



The interpretation is that the fuzzy interval is like part credit on a test question. Instead of points being either in or out of the interval, they are considered partially in and partially out. The fuzzy confidence interval is a function saying how much each point is partially in. This ranges from 0.95 (for 0.95 coverage) down to zero. This fuzzy confidence interval is zero for p above 0.1322 (we had to use cut-and-paste for this number, which one should never do, because of bad design of R function `fci.binom` which gives no programmatic access to this number; it just prints the number and does not return anything as a value). If one wanted a conventional confidence interval corresponding to this fuzzy interval it would be the reported support $[0, 0.1322)$, but this would be only conservative-exact rather than exact-exact. It would have coverage higher than the specified 0.95. The fuzzy interval says we don't need to consider all of these points to be fully in the interval to get the required coverage. (The fact that no point is considered fully in the interval is just a curious fact about the way these intervals work. It is related to the probability of $x = 0$ goes to 1 as $p \to 0$, so one would get more than the specified 0.95 coverage if points near zero had the graph of the fuzzy confidence interval greater than 0.95.)

Note that these intervals, which are theoretically justified, say that the score and likelihood intervals reported by Agresti, which are not theoretically justified (for true unknown parameter on the boundary of the parameter space) are too short, in fact, way too short. This can be seen from the fact that their coverage dips way below 0.95 for p near zero and one (web page about coverage of confidence intervals).

Agresti Section 1.4.4

Use fuzzy (preceding section) instead of mid-P-value.

Agresti Section 1.5

We have our own notes for this.

But there is a lot more to say on this subject. Basically, this whole course is about data that can be taken to arise from either Poisson, multinomial, or product multinomial sampling. Basically, the whole textbook (Agresti) and all of the lecture notes for this course are on this subject.

Agresti Section 1.5.6: Structural Zeros

Sometimes when a contingency table is laid out in an array, some cells have probability zero by design. They cannot occur. R function `chisq.test` cannot handle this case. Neither can R function `glm`.

For the example in Section 1.5.6 in Agresti the data are

```
n <- matrix(c(30, 0, 63, 63), nrow = 2)
n
```

```
##      [,1] [,2]
## [1,]   30   63
## [2,]    0   63
```

And the log likelihood for the alternative hypothesis is

$$\begin{aligned} l(\pi) &= n_{11} \log(\pi^2) + n_{12} \log[\pi(1 - \pi)] + n_{22} \log(1 - \pi) \\ &= 2n_{11} \log(\pi) + n_{12} \log(\pi) + n_{12} \log(1 - \pi) + n_{22} \log(1 - \pi) \end{aligned}$$

(middle unnumbered displayed equation on p. 21 in Agresti). Applying calculus we get the following displayed equation in Agresti

$$l'(\pi) = \frac{2n_{11}}{\pi} + \frac{n_{12}}{\pi} - \frac{n_{12}}{1 - \pi} - \frac{n_{22}}{1 - \pi}$$

If we are shaky on the calculus, even R can do this

```
D(quote(n11 * log(pi^2) + n12 * log(pi * (1 - pi)) + n22 * log(1 - pi)),
    name = "pi")
```

```
## n11 * (2 * pi/pi^2) + n12 * (((1 - pi) - pi)/(pi * (1 - pi))) -
##     n22 * (1/(1 - pi))
```

But it gets a somewhat messier formula. Setting our derivative equal to zero and multiplying both sides by $\pi(1 - \pi)$ gives

$$2n_{11}(1 - \pi) + n_{12}(1 - \pi) - n_{12}\pi - n_{22}\pi = 0$$

or

$$2n_{11} + n_{12} - (2n_{11} + n_{12} + n_{12} + n_{22})\pi = 0$$

the unique solution of which is

$$\hat{\pi} = \frac{2n_{11} + n_{12}}{2n_{11} + n_{12} + n_{12} + n_{22}}$$

and this agrees with the bottom unnumbered displayed equation on p. 21 in Agresti.

```
pihat <- (2 * n[1,1] + n[1,2]) / (2 * n[1,1] + 2 * n[1,2] + n[2,2])
pihat
```

```
## [1] 0.4939759
```

That is the MLE for π for the null hypothesis. Then the MLE for the expected cell counts is the vector

```
e <- sum(n) * c(pihat^2, pihat * (1 - pihat), 1 - pihat)
e
```

```
## [1] 38.06590 38.99434 78.93976
```

And this agrees with Agresti. Hence the Pearson chi-square test statistic is

```
o <- as.vector(n)
o <- o[o > 0]
tstat <- sum((o - e)^2 / e)
tstat
```

```
## [1] 19.70606
```

```
pchisq(tstat, lower.tail = FALSE, df = 1)
```

```
## [1] 9.031458e-06
```

The null hypothesis has one parameter. The alternative has two parameters (because the probabilities for the three cells that are not the structural zero must sum to one). Hence the degrees of freedom is the difference $2 - 1 = 1$.

We could also do a likelihood ratio test. The value of the log likelihood for the null hypothesis is

```
pihat.vec.0 <- e / sum(n)
pihat.vec.1 <- o / sum(n)
logl0 <- sum(o * log(pihat.vec.0))
logl1 <- sum(o * log(pihat.vec.1))
tstat <- 2 * (logl1 - logl0)
tstat
```

```
## [1] 17.73787
```

```
pchisq(tstat, lower.tail = FALSE, df = 1)
```

```
## [1] 2.535289e-05
```

We will mercifully not try a Wald test.

None of this may make much sense. To do an analysis of these data one should really have already taken Stat 5101-5102 but that is not a prerequisite for the course.

TL;DR If there are structural zeros, then you either really need to know your theoretical statistics or you need help from someone who does. Standard software isn't set up for this.

Agresti Section 1.6

Skip for now, we will take up Bayes later. For just a taste, see Chapter 0.

Concepts

- probability model
- statistical model
- discrete data
- continuous data
- discrete probability model
- continuous probability model
- frequentist inference
- Bayesian inference
- parametric statistical model

- nonparametric statistical model
- parameter
- statistic (a function of data but not a function of parameters)
- parameter space
- sample space
- probability mass function (PMF)
- probability density function (PDF)
- outcome
- event
- probability
- expectation
- $\Pr(A)$
- $E(X)$
- mean
  - of random variable
  - of random vector
- variance
  - of random variable
  - of random vector
- covariance
- standard deviation
- addition rule
  - for mean of sums of random variables
  - for variance of sums of independent random variables
  - for sum of Bernoulli random variables
  - for sum of binomial random variables
  - for sum of Poisson random variables
  - for sum of multinomial random vectors
- multiplication rule
  - for PMF
  - for PDF
  - for means of products of independent random variables
- independent also called *stochastically independent* or *statistically independent*
- dependent
- independent and identically distributed (IID)
- Bernoulli distribution

- binomial distribution (see also handout on inference for the binomial distribution)
- binomial coefficient
- Poisson distribution (see also handout on inference for the Poisson distribution)
- Poisson process
- multinomial distribution
- multinomial coefficient
- sampling models (see also the section about theorems relating sampling models)
  - Poisson
  - multinomial
  - product multinomial
- normal distribution
  - univariate
  - multivariate
- chi-squared distribution
- linearity rule
  - for mean and variance for a linear function of a random variable or random vector
  - for distribution of a linear function of a multivariate normal random vector or univariate normal random variable
- central limit theorem (CLT)
  - univariate
  - multivariate
- delta method
- conditional probability
- joint distribution
- conditional distribution
- marginal distribution
- likelihood (see also handout on likelihood theory and handout on likelihood computation)
- Fisher information
- hypothesis tests
  - Wald
  - likelihood ratio (also called *Wilks*)
  - score (also called *Rao*, also called *Lagrange multiplier*)
  - fuzzy (see also handout on fuzzy tests and confidence intervals)
- confidence intervals
- abbreviations
  - Ber(π)

- Bin(n, π)
- Poi(μ)
- Multi(n, π)
- Normal(μ, σ^2)
- Normal(μ, Σ)
- ChiSq(k)