

Statistics Coop Workshop-May 8, 2002
CLAS 108; 1:00 - 5:30 p.m.

Organizer: Nalini Ravishanker
Associate Professor and Undergraduate Program Director,
Department of Statistics, UConn
e-mail: nalini@stat.uconn.edu
URL: <http://www.stat.uconn.edu/~nalini>

Session I: Instructional Material

A. Probability and Distributions

B. Statistical Inference

A. Probability and Distributions

1. Counting Rules for Computing Probabilities
2. Conditional Probability
3. Properties of Random Variables

1. Counting Rules for Computing Probabilities

Probability is the bridge that will enable us to make inference from a sample to the population and to measure how accurate and reliable the inference is. To study basic ideas in probability, we define an experiment, sample space, events, probabilities of events, and rules of probability. Probability helps us in the study of uncertainty in a systematic manner.

Probability computations require an introduction to several combinatorial formulas.

Fundamental Principle of Counting.

Suppose there are k separate steps in an experiment, and the i th step can be done in n_i ways, $i = 1, \dots, k$. The entire experiment can then be done in

$$n_1 \times n_2 \times \cdots \times n_k \text{ ways.}$$

Example 1.1. How many possible telephone numbers are there within one area code?

$$\#\#\#-\#\#\#\#.$$

Each position can be any one of 10 digits. So, there are 10^7 possible telephone numbers.

Example 1.2. In a lottery, suppose the first number can be chosen in 44 ways, and the second number in 43 ways. There are $44 \times 43 = 1892$ ways of choosing the first two numbers. However, if a person is allowed to choose the “same number” again, there will be $44^2 = 1936$ possibilities.

Also, suppose the winning numbers were selected in order as

$$12, 22, 28, 8, 13, 10$$

Does the sequence

$$8, 12, 22, 13, 28, 10$$

qualify as a winner?

Important questions in counting experiments:

- a) with replacement, or without replacement?
- b) objects ordered, or unordered?

Possible Methods of Counting: k objects out of n

		without replacement	with replacement
objects	Ordered	$\frac{n!}{(n-k)!}$	n^k
	Unordered	$\binom{n}{k}$	$\binom{n+k-1}{k}$

A discussion of the previous table with counting rules as applied to the lottery example:

Recall that

$$\begin{aligned} n! &= n(n-1) \times \cdots \times 1, \\ n! &= n(n-1)!, \text{ and} \\ 0! &= 1. \end{aligned}$$

a) Ordered, without replacement.

By the Fundamental Principle of Counting, the first object can be selected in n ways, the second object in $(n - 1)$ ways, etc. There are

$$n(n - 1)(n - 2) \times \cdots \times (n - k + 1) = \frac{n!}{(n - k)!} \text{ ways.}$$

In Example 1.2, there are $44 \times 43 \times 42 \times 41 \times 40 \times 39 = \frac{44!}{38!} = 5,082,517,440$ ways.

b) Ordered, with replacement.

Each of the k objects can be chosen in n ways, so there are n^k possible ways.

In Example 1.2, there are $44^6 = 7,256,313,856$ ways.

c) Unordered, without replacement.

By the Fundamental Principle of Counting, k objects can be arranged in $k(k - 1) \times \cdots \times 2 \times 1$ ways, and constitute redundant orderings, which must be divided out from (a).

The number of unordered objects is

$$\frac{n(n-1)(n-2)\times\cdots\times(n-k+1)}{k(k-1)\times\cdots\times 2\times 1} = \frac{n!}{(n-k)!k!} = \binom{n}{k}$$

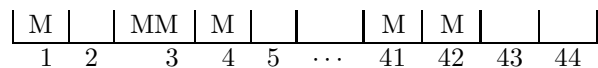
[Recall Binomial Coefficients]

In Example 1.2, there are $\frac{44!}{38!6!} = 7,059,052$ ways.

d) Unordered, with replacement.

(answer is not $44^6/6!$).

Think of the n objects defining bins in which we place k markers (more than one marker per bin is allowed). Count the number of ways of placing k markers on the n objects. We only need to keep track of the arrangement of markers and the walls of the bins. For Example 1.2, suppose



The two outermost walls do not matter. We count all arrangements of 43 walls and 6 markers, i.e., 49 objects, which may be arranged in $49!$ ways. We must divide this by $6! \times 43!$ in order to eliminate redundant orderings. The total number of arrangements is then

$$\frac{49!}{6! \times 43!} = 13,983,816.$$

The general formula is

$$\binom{n + k - 1}{r}.$$

Partitions Counting Rule

Partitioning a set of N different elements into k sets with n_1, n_2, \dots, n_k elements respectively, can be done in $\frac{N!}{n_1!n_2!\cdots n_k!}$ ways where $N = n_1 + n_2 + \cdots + n_k$.

(Recall multinomial coefficients)

Example 1.3. A software consulting company employs 10 consultants. They need to assign 3 consultants to job 1, 2 to job 2 and 5 to job 3. To find the number of different ways of making this assignment, see that $k = 3$, $N = 10$, $n_1 = 3$, $n_2 = 2$, and $n_3 = 5$. Hence, the total number of ways is

$$\frac{10!}{2!3!5!} = 2,520.$$

2. Conditional Probability

Example 2.1. Consider a population of students, of whom 40% are female. Suppose 15% of females and 10% of males smoke.

Find the fraction of smokers that are female, i.e., find $P(F|S)$.

$$P(F) = 0.4, P(M) = 0.6, P(S|F) = 0.15, \text{ and } P(S|M) = 0.10.$$

	S	NS	
M	0.06	0.54	0.6
F	0.06	0.34	0.4
Total	0.12	0.88	1.0

$$\text{Then, } P(F|S) = \frac{P(S|F)P(F)}{P(S)} = \frac{(0.15)(0.4)}{0.12} = 0.5.$$

Example 2.2. Morse code uses “dots” and “dashes”, which occur in the proportion 3 : 4. Sometimes, there may be errors in transmission. Suppose due to interference, a dot is mistakenly received as a dash with probability 1/8. Similarly, a dash is mistakenly received as a dot with probability 1/8. If we receive a dot, how sure are we that a dot was sent, and not a dash? That is, find $P(\text{dot sent}|\text{dot recd})$.

We are given the following probabilities:

$$P(\text{dot sent}) = 3/7; P(\text{dash sent}) = 4/7;$$

$$P(\text{dash recd} | \text{dot sent}) = 1/8; P(\text{dot recd} | \text{dot sent}) = 7/8;$$

$$P(\text{dot recd} | \text{dash sent}) = 1/8; P(\text{dash recd} | \text{dash sent}) = 7/8;$$

$$\begin{aligned} P(\text{dot recd}) &= P(\text{dot recd and dot sent}) + P(\text{dot recd and dash sent}) = \\ &= P(\text{dot recd} | \text{dot sent})P(\text{dot sent}) + P(\text{dot recd} | \text{dash sent})P(\text{dash sent}) = \\ &= (7/8)(3/7) + (1/8)(4/7) = 25/56. \end{aligned}$$

Hence

$$P(\text{dot sent} | \text{dot recd}) = P(\text{dot recd} | \text{dot sent})P(\text{dot sent})/P(\text{dot recd}) = \frac{(7/8)(3/7)}{25/56} = 21/25.$$

3. Properties of Random Variables.

Example 3.1. A contestant on the TV game show “Let’s Make a Deal” has just won \$9,000. He is offered the option of trading this in by choosing 1 of 3 doors, in which case he would receive the prize behind that door. He cannot know behind which of the doors are the prizes of \$20,000, \$5,000 and \$2,000. Should he STOP or GO?

Notice that he must choose between the “fixed” (non-random) amount of \$9,000 and the random value of the door he would choose. Since each door is equally likely, the contestant would “expect to get”

$$1/3(20,000 + 5,000 + 2,000) = \$9,000.$$

Whether he should be indifferent between the choices, or go for the chance at \$20,000 will depend on his “utility function”.

Suppose the contestant MUST pay off a \$20,000 loan the next day, with no other source of funds! Suppose his utility function has the form

$$u(M) = 0 \text{ if } M < 20,000, \quad 1 \text{ if } M \geq 20,000.$$

Hence, $u(9000) = 0$, and the expected utility of selecting a door is

$$\begin{aligned} E[u(X)] &= [u(2000) + u(5000) + u(20000)] \times \frac{1}{3} \\ &= 1/3. \end{aligned}$$

In this case, it is in his best interests to opt to choose a door.

3.2. Some properties of the Binomial Distribution

a) The Binomial distribution applies to sampling with replacement from a finite Bernoulli population. Note that when sampling without replacement is considered, we obtain the hypergeometric distribution. If we extend the notion of a Bernoulli variable to a categorical variable which can assume k (> 2) possible values, and consider sampling with replacement, the result is a Multinomial distribution.

b) Computation done in the 18th century by Pascal at the request of the gambler de Mere: compare the probabilities of two events. De Mere thought the two events had the same probability.

Event 1: Roll a fair die 4 times. Find $P(\text{at least 1 six})$.

Event 2: Throw a pair of dice 24 times. Find $P(\text{a double six})$.

Under Event 1, let X : Number of sixes in 4 rolls. X is $Bin(4, 1/6)$, so that

$$P(\text{at least 1 six}) = P(X > 0) = 1 - P(X = 0) = 1 - \left(\frac{5}{6}\right)^4 = 0.518$$

Under Event 2, let Y : Number of double sixes in 24 rolls.

Y is $Bin(24, 1/36)$, so that

$$P(\text{at least 1 double six}) = P(Y > 0) = 1 - P(Y = 0) = 1 - \left(\frac{35}{36}\right)^{24} = 0.491$$

No wonder De Mere found himself losing on the second bet!

c) For any n , because $\binom{n}{k} = \binom{n}{n-k}$, the Binomial distribution with $p = 1/2$ is symmetric about $n/2$. This can be verified numerically. Also, see this pattern in tables of the Binomial distribution. For $p > 1/2$, the distribution is skewed to the left. For $p < 1/2$, the distribution is skewed to the right.

d) A useful way to “compute” the mean and variance of the Binomial r.v. Y . Let X_1, \dots, X_n be n independent *Bernoulli*(p) random variables, with $E(X_i) = p$, and $Var(X_i) = pq$ for each i . Writing $Y = X_1 + \dots + X_n$, we obtain $E(Y) = np$, and $Var(Y) = npq$ (using properties of expectation).

e) Reproductive property: if Y_1 is $Bin(n_1, p)$, Y_2 is $Bin(n_2, p)$, and Y_1 and Y_2 are independent, then $Y_1 + Y_2$ is $Bin(n_1 + n_2, p)$.

f) Provided n is large, and p is not too extreme (i.e., near 0 or 1), we may use the normal approximation to the Binomial distribution. This approximation works best, of course when p is close to $1/2$. A conservative rule to follow is $\min(np, n(1-p)) \geq 5$. This approximation is especially useful for large values of n for which Binomial probability tables are not readily available. We must remember to use the correction for continuity.

B. Statistical Inference

1. The Central Limit Theorem
2. Large sample sampling distribution of \bar{X} and \hat{p} . (Video clip)
3. Hypothesis Testing. (Video clip)

1. Central Limit Theorem (CLT).

a) Motivation: In statistical inference, we derive sample statistics which are themselves random variables, and hence admit probability distributions called “sampling distributions”. Confidence limits for the interval estimation problem and decision rules for the hypothesis testing problem must be based on “critical values” derived from such distributions. However, in some situations, it is only possible to do this by “approximating” the distribution of the sample statistic, using a limiting argument. The CLT is concerned with a limiting property of a “sum” of random variables. In particular, if X_1, X_2, \dots are iid random variables from some distribution with mean μ and variance σ^2 (both finite), then

$$\sqrt{n}(\bar{X} - \mu)/\sigma$$

has a limiting $N(0, 1)$ distribution, as $n \rightarrow \infty$. Note that normality comes from sums of small, independent disturbances.

b) Although the CLT is a useful general approximation, there is no way to tell how good the approximation is. Since the “goodness” of the approximation depends on the original distribution, the adequacy of the approximation must be checked case by case.

Session 2: Integration of Interesting Examples into the Course

1. The Birthday Problem.
2. The Capture/Recapture Method

1. The Birthday Problem

Suppose there are n people in a room. What is the probability that at least two of them have a common birthday?

Assume each day of the year is equally to be a birthday.
Let A : there are at least 2 people with a common birthday.
Then A^c : no two people have the same birthday.
Note that A^c is a simpler event than A .
Total number of possible outcomes: 365^n . A^c can happen in $365 \times 364 \times \cdots \times (365 - n + 1)$ ways, so that

$$P(A^c) = \frac{365 \times 364 \times \cdots \times (365 - n + 1)}{365^n}; \text{ hence}$$
$$P(A) = 1 - \frac{365 \times 364 \times \cdots \times (365 - n + 1)}{365^n}$$

We can evaluate this for various n values:

n	$P(A)$
4	0.016
16	0.284
23	0.507
32	0.753
40	0.891
56	0.988

How many people in the room must be ask in order to have a 50 : 50 chance of finding someone who shares your birthday?

Suppose you ask n people.

Let A : someone shares your birthday.

Then, A^c : no one in the room shares your birthday.

Total number of outcomes is 365^n .

A^c can happen in 364^n ways; hence

$$P(A) = 1 - \frac{364^n}{365^n} = 1 - \left(1 - \frac{1}{365}\right)^n.$$

For $P(A)$ to be 0.5, n should be 253!

2. Capture/Recapture Method

Useful in estimating the size of a wildlife population.

Suppose 10 animals of a given species are captured from a forest, tagged and released. One year later, 20 animals are captured from the same area; 4 of them are tagged. How large is the species population in that area?

Assume there are n animals in the population. Of these 10 are tagged. Suppose we make this big assumption: the 20 animals captured later are such that all $\binom{n}{20}$ possible groups are equally likely. The probability that 4 of these are tagged is

$$\frac{\binom{10}{4} \binom{n-10}{16}}{\binom{n}{20}}.$$

In this problem, we cannot solve for n explicitly.

However, we can “estimate” n via a value that makes the observed outcome “most probable”. That is, we can use the Method of Maximum Likelihood, and find the maximum likelihood estimate, MLE , of n . Suppose that t animals are tagged. Let the second sample size be m , and suppose r tagged animals are recaptured. We must estimate n by maximizing the likelihood function (recall the hypergeometric probability distribution):

$$L(n) = \frac{\binom{t}{r} \binom{n-t}{m-r}}{\binom{n}{m}}.$$

Consider the ratio of successive terms:

$$\frac{L(n)}{L(n-1)} = \frac{(n-t)(n-m)}{n(n-t-m+r)}.$$

$L(n)$ is increasing if this ratio is > 1 . This occurs when

$$(n-t)(n-m) > n(n-t-m+r), \text{ i.e.,} \\ n^2 - nm - nt + mt > n^2 - nm - nt - nr, \text{ i.e., } mt > nr, \text{ i.e., } \frac{mt}{r} > n.$$

Since $L(n)$ is increasing for $n < mt/r$, and $L(n)$ is decreasing for $n > mt/r$, we estimate $n = \lceil mt/r \rceil$, i.e., the greatest integer not exceeding mt/r .

In the example with $t = 10$, $m = 20$, and $r = 4$, we estimate $n = 50$.

Session III: Relevance of Statistics to the Real World

1. Conditional Probability in Understanding Medical Procedures

1. Conditional Probability in Understanding Medical Procedures

Recall that if A and B are two events such that $P(B) > 0$,

$$P(A|B) = \frac{P(A \cap B)}{P(B)}$$

defines the conditional probability of the event A given that B has occurred.

Medical testing (drug testing or disease testing) involves a tradeoff – benefit of detection (of drug abuse or disease) versus the cost and potential invasion of privacy to the subject. Two important features related to any diagnostic test are based on conditional probabilities

- sensitivity of the test, and
- specificity of the test.

Sensitivity = $P(\text{Positive test result} \mid \text{Disease present})$,
i.e., Sensitivity denotes the probability of a “true positive”.

Specificity = $P(\text{Negative test result} \mid \text{Disease absent})$,
i.e., Specificity denotes the probability of a “true negative”.

Let us discuss this in terms of a specific disease, say Down’s Syndrome (DS). DS is a genetic disorder which may lead to physical and mental retardation. It is caused by the inheritance of an extra copy of chromosome 21. Medical science now believes that the risk of giving birth to a baby with DS increases with the age of the mother. For instance, over the entire population, the risk is 1 in 700; however, for women at age 35, the risk is 1 in 270. That is, for women over 35, $P(DS) = 1/270 = 0.003704$

Two procedures useful for detecting DS are –

(a) Amniocentesis

- routinely recommended for women over 35
- almost 100% detection of DS
- small risk of miscarriage
- expensive ($> \$1000$).

(b) Blood tests

- investigated in a study “How Effective are Blood Tests for the Detection of Down’s Syndrome?” The New England Journal of Medicine, April 1994.
- No known risk of miscarriage
- Inexpensive ($< \$100$)
- How effective?

In the study, 5385 pregnant women over age 35 were given the blood tests and the amniocentesis.

Sensitivity of blood tests = $P(\text{Positive test result} \mid \text{DS}) = 0.89$.

Specificity of blood tests = $P(\text{Negative test result} \mid \text{No DS}) = 0.75$.

Recall that $P(\text{DS}) = 0.003704$.

Find the proportion of women over age 35 that would test positive on the blood tests.

$$P(\text{test positive}) = P(\text{test positive and DS}) + P(\text{test positive and No DS}) = P(\text{test positive} \mid \text{DS})P(\text{DS}) + P(\text{test positive} \mid \text{No DS})P(\text{No DS}) = (0.89)(0.003704) + (1 - 0.75)(1 - 0.003704) = (0.89)(0.003704) + (0.25)(0.996296) = 0.003297 + 0.249074 = 0.2524.$$

What proportion of these positive blood test results are “false positives”? That is, find $P(\text{No DS} \mid \text{test positive})$.

$$\text{Note that } P(\text{DS} \mid \text{test positive}) = P(\text{DS and test positive})/P(\text{test positive}) = 0.003297/0.2524 = 0.0131.$$

$$\text{Hence, } P(\text{No DS} \mid \text{test positive}) = 1 - 0.0131 = 0.9869.$$

The “false positive rate” is about 98.7%. This means that almost 98.7% of positive test results are given when there is no DS.

False Negative Rate:

$$P(\text{DS} \mid \text{test negative}) = P(\text{DS and test negative})/P(\text{test negative}) = P(\text{test negative} \mid \text{DS})P(\text{DS})/P(\text{test negative}) = \frac{(1-0.89)(0.003704)}{(1-0.2524)} = \frac{(0.11)(0.003704)}{0.7476} = 0.000545.$$

The false negative rate is 0.0545%, a 1 in 1800 chance.