~~Tuesday~~ [redacted]

Chapter 5

A random variable (r.v.) is a numerical description of the outcome of a statistical experiment.

That is, it is a real-valued function defined on the sample space S. Let X denote a r.v.. Then

$$\boxed{S} \xrightarrow{X} \mathbb{R}\text{eal numbers}$$

$\boxed{\text{Example 1:}}$ Toss a fair coin twice. The sample space is

$$S = \{(H,H), (H,T), (T,H), (T,T)\}.$$

This is discrete and finite.

Let X = number of heads resulting on the two tosses.

Then X is a r.v. since it provides a numerical description of the experimental outcome. We use cap X to denote the r.v. and little x to denote its possible values or "realizations". Each x corresponds to a subset of the sample space. So X assumes its values with certain probability. That's why it is called a random variable.

S	$X=x$	$P(X=x)$
$\{(T,T)\}$ ⟷	0	1/4
$\{(T,H),(H,T)\}$ ⟷	1	1/2
$\{(H,H)\}$ ⟷	2	1/4

$\boxed{\text{Example 2:}}$ Toss a balanced die until a five occurs.

$$S = \{F, NF, NNF, NNNF, \ldots\}$$

Let X = number of tosses required to obtain a five
$= 1, 2, 3, 4, \ldots$

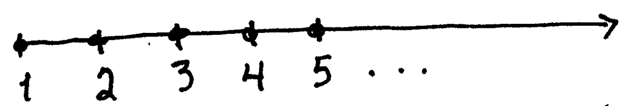
$P(X=1) = 1/6$, $P(X=2) = \frac{5}{6} \cdot \frac{1}{6} = \frac{5}{36}$ (tosses are independent)

$P(X=3) = \frac{5}{6} \cdot \frac{5}{6} \cdot \frac{1}{6} = \left(\frac{5}{6}\right)^2 \frac{1}{6}$; in general

$$P(X=x) = \left(\frac{5}{6}\right)^{x-1} \cdot \frac{1}{6}, \quad x = 1, 2, 3, \ldots$$

The r.v. X here is discrete and infinite.

A discrete r.v. may assume only a finite or countable # of values. Their values correspond to separated points on a number line.



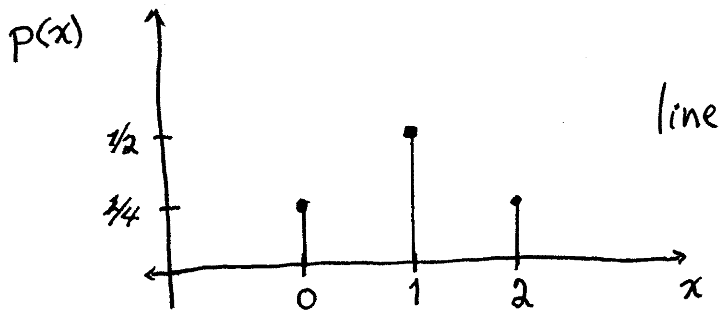
. For any discrete r.v. X, the probability distribution is a table, graph, or mathematical formula that specifies all possible values of the r.v. X and their associated probabilities. The prob. function, written $f(x)$, gives the prob. associated with value x.

$$f(x) = P(X=x) = p(x).$$

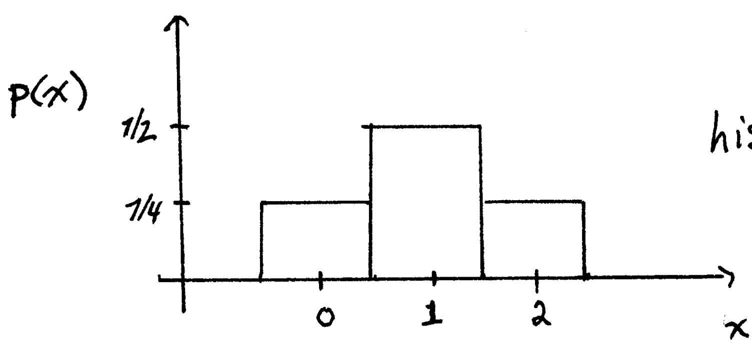
Graphical Displays of a probability distribution:
for a <u>discrete</u> <u>r.v.</u> X.

- line graph
- histogram

Return to Example 1: X = # of heads observed in two tosses of a fair coin.



line graph $\qquad p(x) = P(X=x)$



histogram

The <u>mean value</u>, μ_x, also called the <u>expected</u> value, $E(X)$, is the point (value) where the histogram is balanced. We define it as follows:

$$\mu_x = \sum_{\text{all } x} x \cdot p(x)$$

The variation (or spread) in the histogram is classically measured by the <u>variance</u> of X. We denote it by σ_x^2 and is defined as follows:

$$\sigma_x^2 = \operatorname{var}(X) = \sum_{\substack{\text{all}\\ x}} (x - \mu_x)^2 \cdot p(x)$$

algebraic identity $$= \left(\sum_{\substack{\text{all}\\ x}} x^2 \cdot p(x)\right) - (\mu_x)^2$$

Back to Example 1:

<u>mean:</u> $\mu_x = \sum_{\text{all } x} x \cdot p(x) = 0p(0) + 1p(1) + 2p(2)$

$= 0 + 1(1/2) + 2(1/4)$

$= \frac{1}{2} + \frac{1}{2} = 1$

<u>variance:</u> $\sum_{\substack{\text{all}\\ x}} x^2 p(x) = 0^2 p(0) + 1^2 p(1) + 2^2 p(2)$

$= 0 + 1/2 + 4(1/4) = 1.5$

Then, $\sigma_x^2 = (1.5) - (1)^2 = .5$

<u>standard deviation:</u> $\sigma_x = \sqrt{\sigma_x^2} = \sqrt{.5} \approx .707$

When a statistical experiment is designed in such a way that on any one trial, one of two possible outcomes will result, we call this a Bernoulli trial.

For example:

1. Toss a coin once. $S = \{H, T\}$.

2. Draw a marble at random from an urn with 4 yellows, 5 reds, and 6 greens. Our interest is in observing a yellow marble.

   We can let yellow = a success

   and red or green = a failure.

   Then on any one draw (with replacement), we observe a success or a failure.

   If we draw 3 marbles in succession, each time we replace the marble, then on each trial

   $$p = P(\text{success}) = P(\text{yellow}) = \frac{4}{15}.$$

   $$q = P(\text{failure}) = 1 - p = 1 - \frac{4}{15} = \frac{11}{15}.$$

Binomial Experiment

We perform a sequence of n independent and identical Bernoulli trials. On each trial $p = P(\text{success})$ and $q = 1-p$.

Now, the random variable X = # of successes observed in the n trials. The X is called a binomial random variable.

Note: $X = x$, where $x = 0, 1, 2, \ldots, n$.

In example 2, let X count the # of yellow marbles drawn in trials.

$$X = 0, 1, 2, 3.$$

That's it. Here $n = 3$, and

$$p = P(\text{success}) = P(\text{yellow}) = \frac{4}{15}, \quad q = \frac{11}{15}.$$

Since the marble is replaced after each draw, p (and hence q) stays the same from draw to draw (a draw is a trial).

Thus, example 2 is an example of a binomial experiment, an X is a binomial r.v.

Let's compute the probabilities associated with each possible value of X.

$X=0$: $\underline{\bar{Y}}\ \underline{\bar{Y}}\ \underline{\bar{Y}}$ $\quad \therefore\ P(X=0) = \frac{11}{15} \times \frac{11}{15} \times \frac{11}{15} = \left(\frac{11}{15}\right)^3$

by <u>independence</u> of draws.

$X=1$: ⓐ $\underline{Y}\ \underline{\bar{Y}}\ \underline{\bar{Y}}$ $\quad P(ⓐ) = \frac{4}{15} \times \left(\frac{11}{15}\right)^2$

or ⓑ $\underline{\bar{Y}}\ \underline{Y}\ \underline{\bar{Y}}$ $\quad P(ⓑ) = \frac{11}{15} \times \frac{4}{15} \times \frac{11}{15} = \frac{4}{15} \times \left(\frac{11}{15}\right)^2$

or ⓒ $\underline{\bar{Y}}\ \underline{\bar{Y}}\ \underline{Y}$ $\quad P(ⓒ) = \frac{11}{15} \times \frac{11}{15} \times \frac{4}{15} = \frac{4}{15} \times \left(\frac{11}{15}\right)^2$

$$\therefore\ P(X=1) = 3 \cdot \left(\frac{4}{15}\right) \cdot \left(\frac{11}{15}\right)^2$$

$X=2$: a prototype outcome is

$\underline{Y}\ \underline{Y}\ \underline{\bar{Y}}$ $\quad P(\ \ ) = \left(\frac{4}{15}\right)^2 \times \left(\frac{11}{15}\right)$

How many sequences would correspond to $X=2$?

$${}_3C_2 = \frac{3!}{2!\,1!} = 3.$$

$$\therefore\ P(X=2) = 3 \cdot \left(\frac{4}{15}\right)^2 \cdot \left(\frac{11}{15}\right)$$

$X=3$: only one sequence $\underline{Y}\ \underline{Y}\ \underline{Y}$

$$P(X=3) = \left(\frac{4}{15}\right)^3.$$

Let x denote # of observed successes in n Bernoulli trials in a binomial experiment. Then there are

$${}_nC_x \text{ ways to observe } x \text{ success;}$$

i.e., there are ${}_nC_x$ sequences yielding x successes.

Let p = prob(success) and $q = 1 - p$. Then the <u>general</u> <u>math</u> formula for a <u>binomial</u> <u>probability</u> <u>function</u> is given by

$$f(x) = P(X = x) = {}_nC_x \cdot p^x \cdot q^{n-x}, \quad x = 0, 1, 2, \ldots, n.$$

This completely describes the distribution.

The <u>expected</u> <u>number</u> <u>number</u> of successes in n trials is

$$\mu_x = n \cdot p.$$

and the variance of X, σ_x^2, is given by

$$\sigma_x^2 = npq.$$

Hence, standard deviation is $\sigma_x = \sqrt{npq}$.

In example 2, X = # of yellow marbles in three draws.

Then

$$f(x) = P(X=x) = {}_3C_x \left(\frac{4}{15}\right)^x \left(\frac{11}{15}\right)^{3-x}, \quad x = 0, 1, 2, 3.$$

mean: $\mu_x = np = 3\left(\frac{4}{15}\right) = \frac{12}{15} = \frac{4}{5} = .8$

var: $\sigma_x^2 = npq = 3\left(\frac{4}{15}\right)\left(\frac{11}{15}\right) = \frac{4}{5} \times \frac{11}{15} = \frac{44}{75} = .587$

st.dev: $\sigma_x = \sqrt{\frac{44}{75}} = .766.$

MINITAB Example 2 : YELLOW MARBLE

① Enter into C1 x = 0, 1, 2, 3
Name C1 x

② Name C2 p(x)

③ Main menu bar
Calc → Probability Distributions → Binomial

⊙ Probability

Number of trials : $\boxed{3}$
Prob. of Success: $\boxed{.2667}$

⊙ Input Column : $\boxed{C1}$ x
Optional Storage: $\boxed{C2}$ p(x)

$\boxed{ok}$

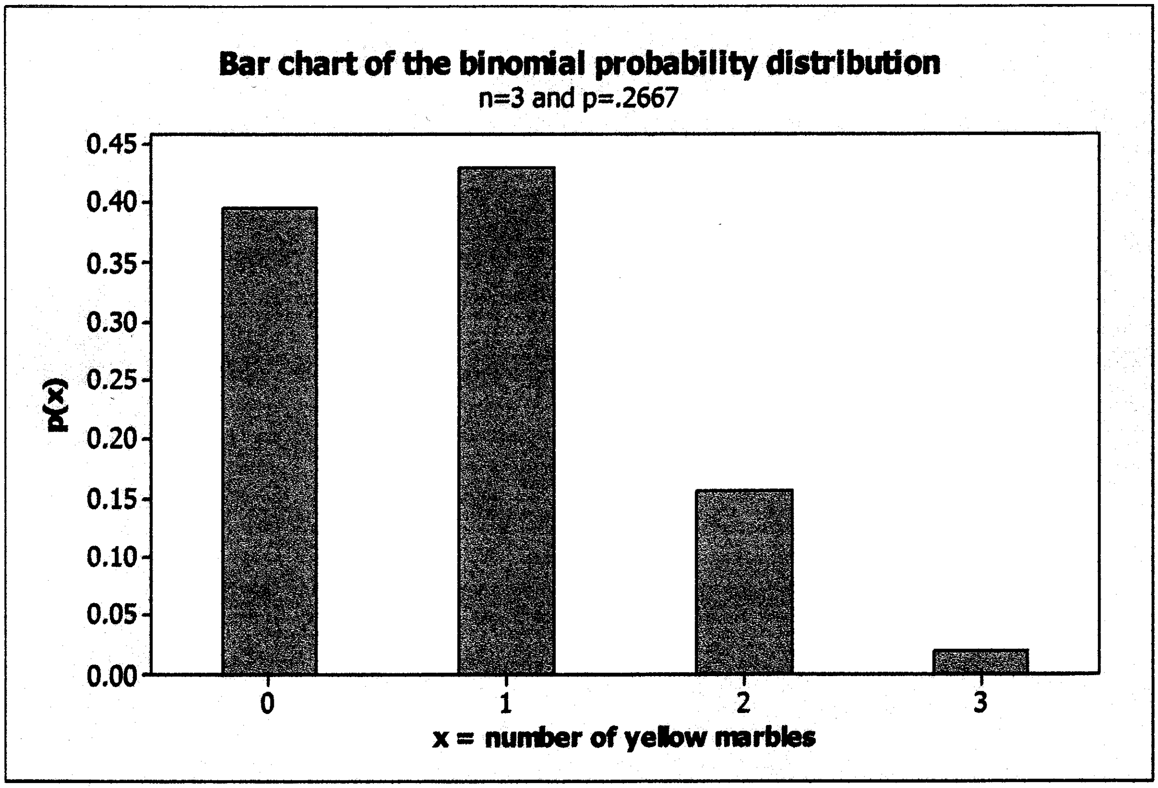
④ Graph → Bar chart
bar represents
$\boxed{\text{values from a table}}$
Graph variable
$\boxed{p(x)}$
categorical variable
x

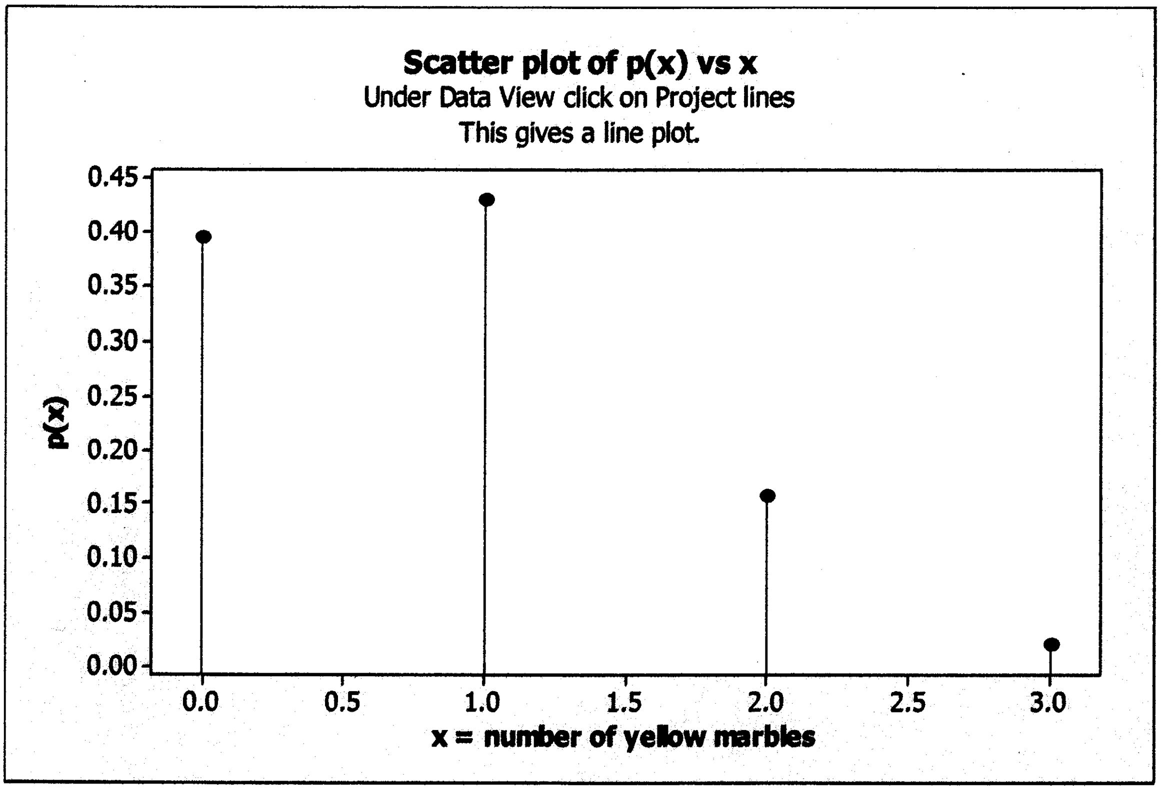
$\boxed{ok}$

Example 2: We draw three marbles at random in succession with replacement from an urn filled with 4 yellow, 5 red, and 6 green marbles. Let X count the number of yellow marbles observed.

The probability distribution of X is binomial with parameters n=3 and p=4/15=.2667. The results from MINITAB are:

Row	x	p(x)
1	0	0.394317
2	1	0.430237
3	2	0.156476
4	3	0.018970




Scatter plot of p(x) vs x
Under Data View click on Project lines
This gives a line plot.
p(x)
0.45
0.40
0.35
0.30
0.25
0.20
0.15
0.10
0.05
0.00
0.0
0.5
1.0
1.5
2.0
2.5
3.0
x = number of yellow marbles

Let's consider $\boxed{\text{Example 3}}$ again.

The population: A box contains 6 blue marbles and 4 red marbles.

BOX

6B
4R

10 marbles

We can think of B as a success
" " " " R then as a failure.

When a marble is drawn (sampled) there are only <u>two</u> possible outcomes S(B) or F(R).

Can use this as a model for a survey sample.
Let N = # of registered voters in Multnomah County.
Let S = voter is registered Democratic
F = " " NOT Registered DEMOCRATIC.
Select a registered voter at random. See if s/he is registered D or $\overline{D}$.

$\boxed{\text{Experiment 1.}}$

Marbles are chosen at random, color is observed, but <u>not</u> replaced, from a <u>finite</u> population of size N. Success here is the color is BLue.

Let n = # of trials

Let X = # of blue marbles in a sample of $n=5$.

<u>Hypergeometric Experiment</u>

X is hypergeometric random variable (r.v.)

$\boxed{\text{Experiment 2.}}$

Marbles are chosen at random, color is observed, but then the marble is replaced.

Success ≡ color is Blue

Let n = # of trials

Let X = # of blue marbles in a sample of $n=5$

<u>Binomial Experiment</u>

X is a binomial r.v.

$\boxed{\text{When } X \text{ is a hypergeometric r.v.:}}$

Let $a = \#$ of successes

$b = \#$ of failures. Let $n =$ sample size (or # of trials)

Then, $N =$ total # of items in population

$= a+b.$

$X = x$

The probability function is:

$$P(X=x) = \frac{{}_aC_x \cdot {}_bC_{n-x}}{{}_NC_n} = \frac{\binom{a}{x}\cdot\binom{b}{n-x}}{\binom{N}{n}}.$$

MEAN: $= E(X) = n \cdot \frac{a}{N}$, Var(X) $= \frac{N-n}{N-1} \cdot n \cdot \frac{a}{N} \cdot \frac{b}{N}$

$\boxed{\text{When } X \text{ is a BINOMIAL r.v.}}$

Let $n = \#$ of trials $=$ sample size

let $p =$ proportion (probability) of successes on each trial.

$q = 1-p.$

$X = x. = 0, 1, 2, \cdots, n$

The probability function is:

$$P(X=x) = {}_nC_x \cdot p^x \cdot q^{n-x} = \frac{n!}{x!(n-x)!} \cdot p^x \cdot q^{n-x}$$

mean(X) $= E(X) = np$; var(X) $= npq$.

<u>Question</u>: When can sampling <u>without</u> replacement from a finite population (the hypergeometric experiment) be ignored; i.e. be nicely approximated by the Binomial Experiment (i.e, as if we did replace each item) ??

Practically speaking, when interested in the # of ~~observed~~ "successes" obtained in a survey (sampling <u>without</u> replacement), when can we treat this X has a binomial and hence use the binomial prob. function and the binomial expression for variance?

Recall:

$var(\text{binomial } X) = npq$ $\qquad$ $var(\text{hypergeo. } X) = \dfrac{N-n}{N-1} \cdot n p q$

$p \leftrightarrow \dfrac{a}{N}$, $q \leftrightarrow \dfrac{b}{N}$

finite population correction factor when sampling w.o. replacement from finite pop of size N. (f.p.c.f.)

Answer: When $\dfrac{n}{N} \leq .05 = 5\%$ (The 5% rule.)

Box

$$\boxed{\begin{matrix} 6B \\ 4R \end{matrix}}$$

Experiment 1. Hypergeometric.

let $n = 5$

let X = # of blue marbles after 5 trials.

$$P(X=x) = \frac{\binom{6}{x}\cdot\binom{4}{5-x}}{\binom{10}{5}}, \quad x = 0,1,2,3,4,5$$

$$P(X=3) = \frac{\binom{6}{3}\binom{4}{2}}{\binom{10}{5}} = \frac{(20)(6)}{252} = \frac{20}{42} = \frac{10}{21} = .47619 \approx .4762$$

$E(X) = 5 \cdot \frac{6}{10} = 3$ blue marbles are expected

$$Var(X) = \frac{10-5}{10-1} \cdot 5 \cdot \frac{6}{10} \cdot \frac{4}{10} = \frac{5}{9} \cdot 5 \cdot \frac{3}{5} \cdot \frac{2}{5} = \frac{2}{3}$$

$$SD(X) = \sigma_x = \sqrt{\frac{2}{3}} \approx .8165$$

Experiment 2. Binomial

Let $n = 5$, Let X = # of blue marbles after 5 trials.

$$P(X=x) = \frac{5!}{x!(5-x)!} \cdot \left(\frac{6}{10}\right)^x \left(\frac{4}{10}\right)^{5-x}, \quad x = 0,1,2,\dots,5,$$

$$P(X=3) = \frac{5!}{3!2!} \cdot \left(\frac{3}{5}\right)^3 \left(\frac{2}{5}\right)^2 = 10 \cdot \frac{27}{125} \cdot \frac{4}{25} = .3456$$

$$\mu = E(X) = 5 \cdot \frac{6}{10} = 3$$

$$\sigma^2 = Var(X) = 5 \cdot \frac{6}{10} \cdot \frac{4}{10} = 5 \cdot \frac{3}{5} \cdot \frac{2}{5} = \frac{6}{5}$$

$$SD(X) = \sigma = \sqrt{\frac{6}{5}} \approx 1.095$$

NOTE! The Binomial prob. does not approx. the hypergeo. prob. well. Why?

$\frac{n}{N} = \frac{5}{10} = .50$ > way larger than .05!!

Poisson Distribution

Let X count the number of "events" that occur over a period (interval) of time, or over a given area or volume. Let λ denote the rate of occurrences per unit (time, area, volume, etc.). Then the probability ~~function~~ distribution of such a r.v. can be nicely modeled

by the Poisson ~~probability~~ probability function:

$$P(X=x) = \frac{\lambda^x e^{-\lambda}}{x!}, \quad x = 0, 1, 2, 3, \ldots$$

(Note: $e \approx 2.7183$) Round the prob. to four decimal places.

Note: $\text{mean}(X) = \mu_x = \text{Var}(X) = \sigma_x^2 = \lambda$.

The mean and variance are the same and it is equal to λ!!

EXAMPLE

Phone calls arrive at the rate of 48 per hour at the reservation desk for Regional Airways.

a. Find the probability of receiving three calls in a five-minute interval of time.

b. Find the probability of receiving exactly 10 calls in 15 minutes.

c. Suppose no calls are currently on hold. If the agent takes five minutes to complete the current call, how many callers do you expect to be waiting by that time? What is the probability that none will be waiting?

d. If no calls are currently being processed, what is the probability that the agent can take three minutes for personal time without being interrupted?

The Poisson distribution can be used to approximate the binomial distribution when n is large and p = prob. of success is small.

~~[illegible]~~
A good rule-of-thumb for when n is "large enough" and p is "small enough" is when

$$\lambda = np < 5.$$

Then can use $P(X=x) \approx \dfrac{\lambda^x e^{-\lambda}}{x!}$

to approximate $\dfrac{n!}{x!\,(n-x)!} \cdot p^x \cdot q^{n-x}$.

Example.

Example 6–29

If approximately 2% of the people in a room of 200 people are left-handed, find the probability that exactly five people there are left-handed.

Solution

Since $\lambda = n \cdot p$, then $\lambda = (200)(0.02) = 4$. Hence,

$$P(X; \lambda) = \frac{(2.7183)^{-4}(4)^5}{5!} = 0.1563$$

which is verified by the formula ${}_{200}C_5(0.02)^5(0.98)^{195} \approx 0.1579$. The difference between the two answers is based on the fact that the Poisson distribution is an approximation and rounding has been used.

The Multinomial Distribution:

When we extend the binomial experiment to include three or more possible outcomes on each trial, then this is called a multinomial experiment.

Let $E_1, E_2, \ldots, E_K$ represent the k-possible outcomes. Let $p_1, p_2, \ldots, p_K$ represent the corresponding probabilities that the respective event E_i occurs. Then $p_1 + p_2 + \cdots + p_K = 1$. Perform n trials. Let X_1 = # of trials resulting in E_1, X_2 = # of trials resulting in E_2, ..., X_K = # of trials resulting in E_K. Then $X_1 + X_2 + \cdots + X_k = n$ = total # of trials.

READ This in your book

Formula for the Multinomial Distribution

If X consists of events $E_1, E_2, E_3, \ldots, E_k$, which have corresponding probabilities $p_1, p_2, p_3, \ldots, p_k$ of occurring, and X_1 is the number of times E_1 will occur, X_2 is the number of times E_2 will occur, X_3 is the number of times E_3 will occur, etc., then the probability that X will occur is

[illegible]

where [illegible] $+ p_k = 1$.

The expected number of trials resulting in cell i (outcome i) is np_i

$$np_1 + np_2 + \cdots + np_K$$

$$= n(\underbrace{p_1 + p_2 + \cdots + p_K}_{1}) = n$$

E_1 = movie X_1 = counts = 3
E_2 = dinner + play X_2 = counts = 1
E_3 = shopping X_3 = counts = 1

Example 6–24

In a large city, 50% of the people choose a movie, 30% choose dinner and a play, and 20% choose shopping as a leisure activity. If a sample of five people is randomly selected, find the probability that three are planning to go to a movie, one to a play, and one to a shopping mall.

$P_1 = .5$ $P_2 = .3$ $P_3 = .2$

Solution

$n = 5, X_1 = 3, X_2 = 1, X_3 = 1, p_1 = 0.50, p_2 = 0.30$, and $p_3 = 0.20$. Substituting in the formula gives

$$P(X) = \frac{5!}{3! \cdot 1! \cdot 1!} \cdot (0.50)^3(0.30)^1(0.20)^1 = 0.15$$

Again, note that the multinomial distribution can be used even though replacement is not done, provided that the sample is small in comparison with the population.

W	R	B
II	III	I
2	2	1
X_1	X_2	X_3

$P_1 = \frac{4}{10}$ $\frac{3}{10} = P_2$ $\frac{3}{10} = P_3$

Example 6–26

A box contains four white balls, three red balls, and three blue balls. A ball is selected at random, and its color is written down. It is replaced each time. Find the probability that if five balls are selected, two are white, two are red, and one is blue.

Solution

W R B

$n = 5, X_1 = 2, X_2 = 2, X_3 = 1; p_1 = \frac{4}{10}, p_2 = \frac{3}{10}$, and $p_3 = \frac{3}{10}$; hence,

$$P(X) = \frac{5!}{2!\,2!\,1!} \cdot \left(\frac{4}{10}\right)^2\left(\frac{3}{10}\right)^2\left(\frac{3}{10}\right)^1 = \frac{81}{625}$$

Thus, the multinomial distribution is similar to the binomial distribution but has the advantage of allowing one to compute probabilities when there are more than two outcomes for each trial in the experiment. That is, the multinomial distribution is a general distribution, and the binomial distribution is a special case of the multinomial distribution.