4/16/19 Lecture 5

Take-Home Test 1
- Can do probs 1 & 3
- After class you can do prob 2
- 4 & 5 are extra hard

Death Penalty Case

	Death Penalty		
Defendent	Yes	No	Total
White	19	141	160
Black	17	149	166
Total	36	290	326

White Victim

	Death Penalty		
Defendent	Yes	No	Total
White	19	132	151
Black	11	52	63
Total	30	184	214

Black Victim

	Death Penalty		
Defendent	Yes	No	Total
White	0	9	9
Black	6	97	103
Total	6	106	112

2x2 Contingency tables

DP = Death Penalty
DW = Defendent White

$$\begin{array}{cc} DP? & DW? \\ \left[\begin{array}{c} no \\ | \end{array}\right. & \left.\begin{array}{c} no \\ | \end{array}\right] \end{array} \quad n = 326$$

In 1st table, 19 of the 326 cases the defendent was white and the death penalty was imposed.

Case chosen at random:

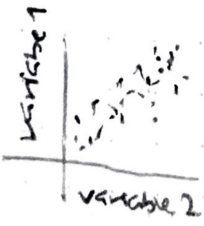
$$P(DP) = \frac{36}{326} = 11.0\%$$

$$P(DP \mid DW) = \frac{19}{160} = 11.9\%$$

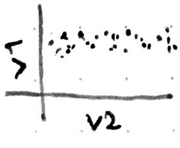
$$P(DP \mid DB) = \frac{17}{166} = 10.2\%$$

It appears that white defendents receive the death penalty more often than black defendents, which is odd because of racial bias.

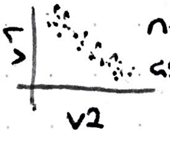
This is an observational study



positive association



no association



negative association

Experimental Design

Y = outcome variable = $\begin{cases} \text{DP} \\ \text{Not DP} \end{cases}$

X = defendent race = $\begin{cases} \text{DW} \\ \text{Not DW} \end{cases}$

enemy: Potential PCFs

Z = victim race = $\begin{cases} \text{VW} \\ \text{Not VW} \end{cases}$

Analysis of White Victim Table

$$P(DP \mid VW) = \frac{30}{214} = 14.0\%$$

$$P(DP \mid VW, DW) = \frac{19}{151} = 12.6\%$$

$$P(DP \mid VW, DB) = \frac{11}{63} = 17.5\%$$

Analysis of Black Victim Table

$$P(DP \mid VB) = \frac{6}{112} = 5.4\%$$

$$P(DP \mid VB, DW) = \frac{0}{9} = 0\%$$

$$P(DP \mid VB, DB) = \frac{6}{103} = 5.8\%$$

Holding the ethnicity of the victim constant of Black, the rate of imposition of the death penalty falls from 11% to 5.4% and again now Black defendents get the DP more often than White defendents.

Why did this happen?

1) Murder victims typically know their murderer
2) In the U.S., white people tend to hang out with white people and Black people with Black
3) Therefore white defendents are mostly murdering white victims
4) Judges and juries in the U.S. impose the death penalty more often when the victim is White than when the victim is Black.

$$P_{n,k} = \frac{n(n-1)(n-k+1)(n-k)!}{(n-k)!} = \frac{n!}{(n-k)!}$$

Convention $0! = 1$

<u>Combinations</u>

In the T-S case study
consider the special case in which the family ends up with exactly $k=1$ T-S baby. i.e. $(k=)1$ T and $(n-k)$ 4 Ns.

Let's initially imagine that all 5 of these T and N symbols are different (like different playing cards) by denoting them $\left\{ \begin{matrix} T_1 \\ N_1 \; N_2 \; N_3 \; N_4 \end{matrix} \right\}$

There would then be $n! = 5! = 120$ ways to arrange them in order left to right, eg. $N_3 \; T_1 \; N_4 \; N_1 \; N_2$

NOW take the subscripts away: There are $4!$ ways to rearrange the Ns among themselves and $1! = 1$ ways to "rearrange" the Ts among themselves

So $5!$ is way too big and needs to be divided by $4! \cdot 1!$

$$\frac{5}{1!\,4!} = \frac{n!}{k!(n-k)!} = \frac{5 \cdot \cancel{4!}}{\cancel{4!}} = 5 \quad \text{(the right answer)}$$

<u>Def</u>: Given a set with n distinct elements, each distinct subset of size k is called a <u>combination</u> of elements, and there are $C_{n,k} = \frac{n!}{k!(n-k)!}$ ways to do this

Notation: $\frac{n!}{k!(n-k)!} = \binom{n}{k}$ "n choose k"

binomial coefficient

So what we've shown is $P(Y=y) = \binom{n}{y} p^y (1-p)^{n-y}$

# of T-S babies

valid for all $y = 0, 1, \ldots, n$

$n \geq 1$ and $0 \leq p \leq 1$

(Refer to as: binomial distribution)

Birthday Problem

$P\left(\begin{array}{l}\text{at least 2 people registered for AMS 131 this term} \\ \text{have the same birthday}\end{array}\right) = ?$

Simplifying assumptions:

1) birth rate constant from Jan 1 – Jan 31
2) Feb 29 → randomize to another day
3) Only considering month and day, not birth year

Let k = # people registered

For AMS 131 = 93 as of July 29, 2016
132 " " August 2, 2017
242 " " April 16, 2019

birth dates

The sample size S is like filling in k slots, each of which has n possible values, so S contains n^k equally likely outcomes.

$n = 365$ = # possible birthdays

$365^{242} = 1.19 \times 10^{620}$

Let's try to work out $P(\text{not } A)$

If nobody has the same birthday, then a randomly chosen person 1 has $n = 365$ possibilities, a randomly chosen person 2 (distinct from person 1) has $(n-1) = 364$ possibilities, ... and finally the last person k (242) (no longer random) has $(n-k+1)$ $= 124$ possibilities

All together (not A) has $n(n-1)\ldots(n-k+1) = P_{n,k} = \frac{n!}{(n-k)!}$ equally likely outcomes favorable to it

and $P(A) = 1 - P(\text{not } A) = 1 - \frac{365!}{123! \cdot 365^{242}}$

$= 1 - \frac{n!}{(n-k)!\, n^k} = ?$

To compute:

1) Don't evaluate the numerator and denominator separately & then divide. Instead, cancel them against each other.

$$1 - \frac{365!}{272!\, 365^{93}} = 1 - \frac{(365)(364)\ldots(272)}{(365)(365)\cdots(365)} = 0.999997$$

2) Stirling's approximation: $\log n! = \frac{1}{2}\log 2\pi + (n+\frac{1}{2})\log n - n$

The log function goes to ∞ much slower than the x function.

So $P(A) = 1 - \exp\left\{\log\left[\frac{n!}{(n-k)!\, n^k}\right]\right\}$

Stirling simplification $\quad$ for any $x > 0$, $x = \exp[\log(x)]$

$= 1 - \exp\left\{(n-k+\frac{1}{2})[\log(n) - \log(n-k)] - k\right\} = 0.9999974$

3) The Gamma function is a generalization of $n!$, n integer, to all positive real #s : $n! = \underset{\text{gamma}}{\Gamma}(n+1)$

Many mathematical packages (R, Matlab, ...) have a log-gamma function

With 23 people, there's already a 50/50 chance of 2 ppl having the same birthday

$P(A) = 1 - \exp[\log n! - \log(n-k)! - k\log n]$

$= 1 - \exp[\log\Gamma(n+1) - \log\Gamma(n-k+1) - k\log n]$

Generalizing the binomial coefficients

What if there are more than 2 possible outcomes in a generalization of the T-S case study (T, N)?
T-S baby ↑ ↑ not T-S baby

We want n distinct elements to be divided into k different groups $(k \geq 2)$ so that n_j elements fall into group j, $\sum_{j=1}^{k} n_j = n$;

Q: In how many ways can this be done?
A: Follow the argument in the textbook pg. 42–43, which generalises the line of reasoning leading to the binomial coefficients $\binom{n}{y}$ when $k=2$

$$\binom{n}{y} = \frac{n!}{y!(n-y)!} \qquad \begin{matrix} n \geq 1 \\ k \geq 2 \end{matrix} \qquad 1 \leq n_j \leq k \qquad \sum_{j=1}^{k} n_j = n$$

<u>Def</u>: A multinomial coefficient is of the form

$$\binom{n}{n_1, n_2, \ldots, n_k} = \frac{n!}{n_1!\, n_2! \cdots n_k!}$$

This answers how many different ways to have people fall into different groups

<u>Ex</u>: 2016 presidential election (See pg. 333–334 from text)
Imagine randomly sampling with replacement n eligible perspective voters from all such people (the population) in the U.S.

<u>Possible outcomes $k=5$</u>

1 Clinton
2 Trump
3 Johnson
4 Stein
5 Undecided / no comment

Let X_i = # people in sample who say they will vote for candidate i, $i = 1, \ldots, k=5$

Suppose (unknown to us) that the proportion of voters who favor candidate i in the population (*) above is p_i

Because the people are chosen with independent identically distributed (IID) sampling (i.e. at random w/ replacement) each person's outcome will be independent of all other outcomes

Thus P(1st person favors candidate i and 2nd person favors i_2 and... and person n favors i_n) = product of all the p's that correspond to those candidates

$(p_i \cdot p_{i_2} \cdot \ldots \cdot p_{i_n})$

Therefore P(the sample has x people favoring candidate 1, x_2 people favoring candidate 2, ... x_k people favoring candidate k) (listed in a prespecified order) = ...

$= p_1^{x_1} \cdot p_2^{x_2} \ldots p_k^{x_k}$, with $0 \leq x \leq n$ and $\sum_{i=1}^{k} x_i = n$

Thus, P(exactly x people favor Clinton & ... & x_k people favor undecided) = $?\, p_1^{x_1} \ldots p_k^{x_k}$, where ? is the total # of different ways the order of the n people in the sample can be listed. But this ? is precisely

$\binom{n}{x_1, \ldots, x_k} = \dfrac{n!}{x_1!\, x_2! \ldots x_k!}$, the multinomial coefficient defined before

So $P(X_1 = x_1 \,\&\, \ldots \,\&\, X_n = x_n) = \dfrac{n!}{x_1! \ldots x_k!} p_1^{x_1} \ldots p_k^{x_k}$

Refer to this as the multinomial probability distribution