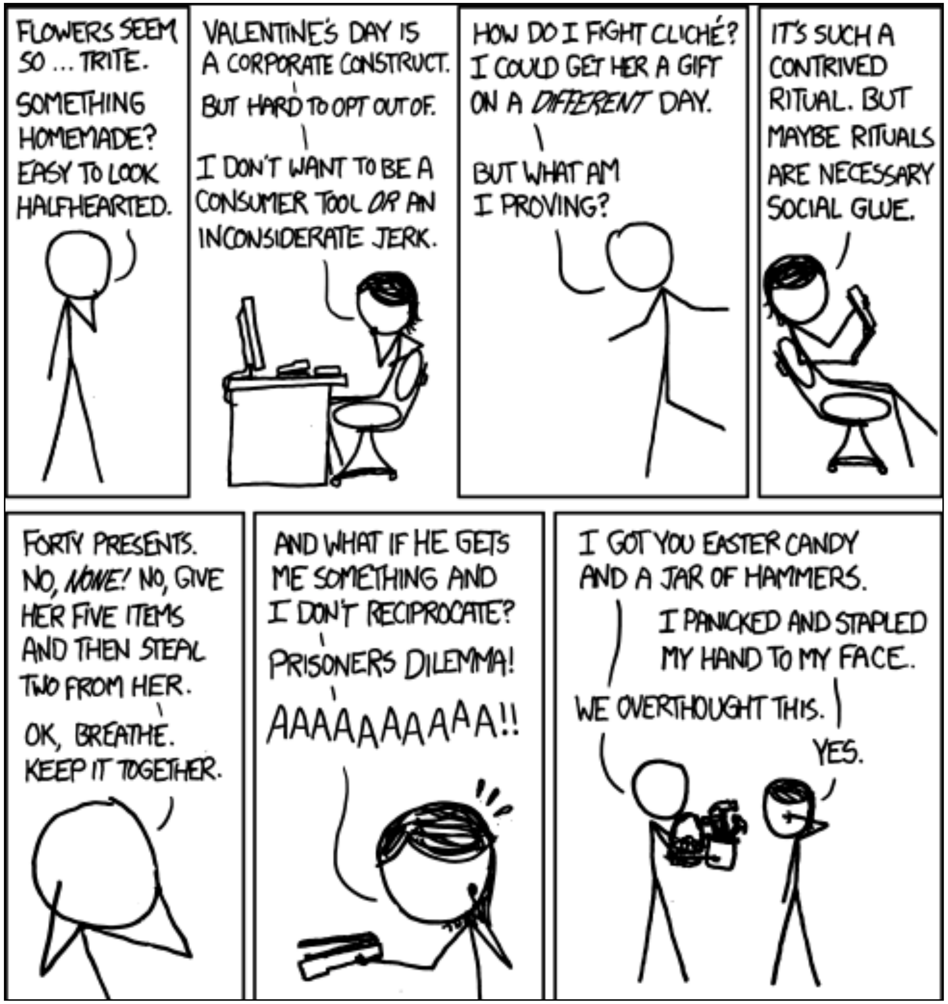
Stat 88: Probability & Math. Stat in Data Science



https://xkcd.com/1016/

Lecture 13: 2/14/2024

Examples, waiting times

4.2

Shobhana Stoyanov

Randomized Controlled Experiments

Two randomized controlled experiments are being run independently of each other. In each experiment, a simple random sample of **half** the participants will be assigned to the treatment group and the other half to control. Expt 1 has 100 participants of whom 20 are men. Expt 2 has 90 participants of whom 30 are men.

What is the chance that the treatment and control groups in Experiment 1 contain the same number of men?

Expt 1 (100): 20 M, 80 M^c

Treatment 50 — Ctrl 50

T_1 = # of men, C_1 = # of men

$T_1 \sim HG(N=100, G=20, n=50)$

$$P(T_1=10) = \frac{\binom{20}{10}\binom{80}{40}}{\binom{100}{50}}$$

Expt 2 (90): 30 Men (M), 60 M^c

Trt 45 T_2 — Ctrl 45 C_2

Problems, continued

What is the chance that the treatment groups in the two experiments have the **same** number of men?

- Notice this is a bit tricky. There are many disjoint cases (each of the treatment groups has 1 man, or 2 men or 3 men etc. What is the max?
- We will have to split the chance into the chance of each of the cases and add them.
-

(1) $P(T_1 = T_2 = 0) = P(T_1 = 0 \,\&\, T_2 = 0)$

$= P(T_1 = 0)\, P(T_2 = 0)$

$= \dfrac{\binom{20}{0}\binom{80}{50}}{\binom{100}{50}} \cdot \dfrac{\binom{30}{0}\binom{60}{45}}{\binom{90}{45}}$

$T_1 \sim HG(N=100, n=50, G=20)$

$T_2 \sim HG(N=90, n=45, G=30)$

$P(T_1 = T_2 = 1) = \dfrac{\binom{20}{1}\binom{80}{49}}{\binom{100}{50}} \cdot \dfrac{\binom{30}{1}\binom{60}{44}}{\binom{90}{45}}$

$P(T_1 = k = T_2)$, $0 \le k \le 20$

$= P(T_1 = k) P(T_2 = k)$

$$= \sum_{k=0}^{20} \frac{\binom{20}{k}\binom{80}{50-k}}{\binom{100}{50}} \cdot \frac{\binom{30}{k}\binom{60}{45-k}}{\binom{90}{45}}$$

( ) ( )

Did the treatment have an effect?

- RCE with 100 participants, 60 in Treatment, 40 in Control
- T: 50 recover, out of 60 (83%), C: 30 recover out of 40 (75%)

- Suppose treatment had no effect, and these 80 just happened to recover. What is the chance they would have recovered no matter what and 50 were assigned to the treatment group by chance?

100 → Trt 60 → 50 Recovered
100 → Ctrl 40 → 30 recovered

T = # who recover in ~~Ct~~ Trt SP.

So total 80 participants recover no matter what. What is prob that 50 of them would have been selected in trt SP.

$$\frac{\binom{80}{50}\binom{20}{10}}{\binom{100}{60}}$$

Hypergeometric but don't know N

- A state has several million households, half of which have annual incomes over 50,000 dollars. In a simple random sample of 400 households taken from the state, what is the chance that more than 215 have incomes over 50,000 dollars?

? How should we do this? $n = 400$, $k = 215$, $G = N/2$, $N = ???$?

Treat it as Binomial $\left(400, \frac{1}{2}\right)$

4.2: Waiting times

$P(R) = \frac{18}{38}$

- Say Ali keeps playing roulette, and betting on red each time. The waiting time of a red win is the number of spins until they see a red (so the number of spins until and including the time the ball lands on a red pocket).

Let T = # of spins until & including the first spin that lands R

$T = 1, 2, \ldots$

What is the probability that Ali will wait for 4 spins before their first win? (That is, the first time the ball lands in red is the 4th spin or trial)

$P(T=1) = \frac{18}{38}$ $\quad P(T=2) = \frac{20}{38} \cdot \frac{18}{38} = \frac{NR}{\frac{20}{38}} \cdot \frac{R}{\frac{18}{38}}$

$P(T=4) = \left(\frac{20}{38}\right)^3 \left(\frac{18}{38}\right)$

- Say we have a sequence of ***independent*** trials (roulette spins, coin tosses, die rolls etc) each of which has outcomes of success or failure, and $\boldsymbol{P(S) = p}$ on each trial.

- Let T_1 be the number of trials up to and including the first success. Then T_1 is the ***waiting time until the first success***.

- What are the values T_1 takes? What is its pmf $f(x)$?

$f(k) =$

T_1 takes values $1, 2, \ldots$

$P(S) = p \quad P(F) = 1-p$

$P(T_1 = 27) = \left(\frac{20}{38}\right)^{26} \left(\frac{18}{38}\right)$

$= (1-p)^{26}(p)^1$

Geometric distribution

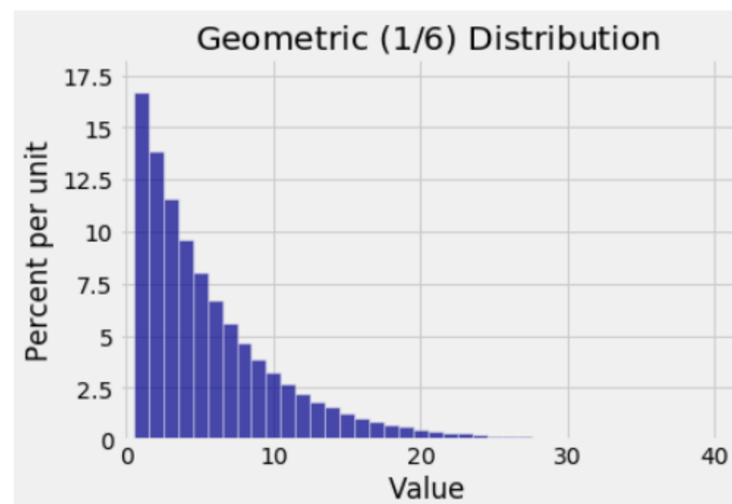
- Say T_1 has the **geometric distribution**, denoted $\boldsymbol{T_1 \sim Geom(p)}$ on $\{1, 2, 3, \ldots\}$

- $f(k) = P(T_1 = k) = (1-p)^{k-1}p$

- Read Check that it sums to 1. What is the cdf for this distribution? Can you think of an easy way to write down the cdf?

$f(x) = (1-p)^{x-1}p$

$F(x) = P(T_1 \leq x)$



Waiting time until r^{th} success

$\{1, 2, \ldots 8\}$

- Say we roll a 8 sided die. $P(\text{"}8\text{"}) = \frac{1}{8}$
- What is the chance that the first time we roll an eight is on the 11^{th} try?

Neg Binomial

$= P(\underbrace{FFFFFFFFFF}S) = \left(\frac{7}{8}\right)^{10}\left(\frac{1}{8}\right)$

- What is the chance that it takes us 15 times until the 4^{th} time we roll eight? (That is, the waiting time until the 4^{th} time we roll an eight is 15)

$= P(\underbrace{\overset{\left(\frac{1}{8}\right)^3\left(\frac{7}{8}\right)^{11}}{\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_}}_{14 \text{ rolls} \quad 3S \; 11F}\overset{\left(\frac{1}{8}\right)}{S})$ $\rightarrow$ $\binom{14}{3}\underbrace{\left(\frac{1}{8}\right)^3\left(\frac{7}{8}\right)^{11}\left(\frac{1}{8}\right)}_{\text{Total } 4S}$

$\uparrow$ 4^{th} success

- What is the chance that we need **more** than 15 rolls to roll an eight 4 times?
- Notice that the ***right-tail*** probability of T_4 is a left hand (cdf) of the Binomial distribution for (15,1/8), and where k=3.

- In general, $P(T_r = k) =$
- And $P(T_r > k) =$