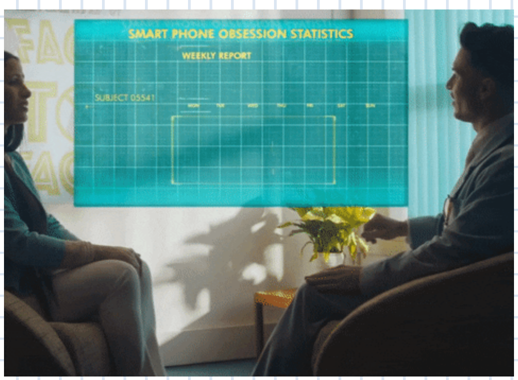
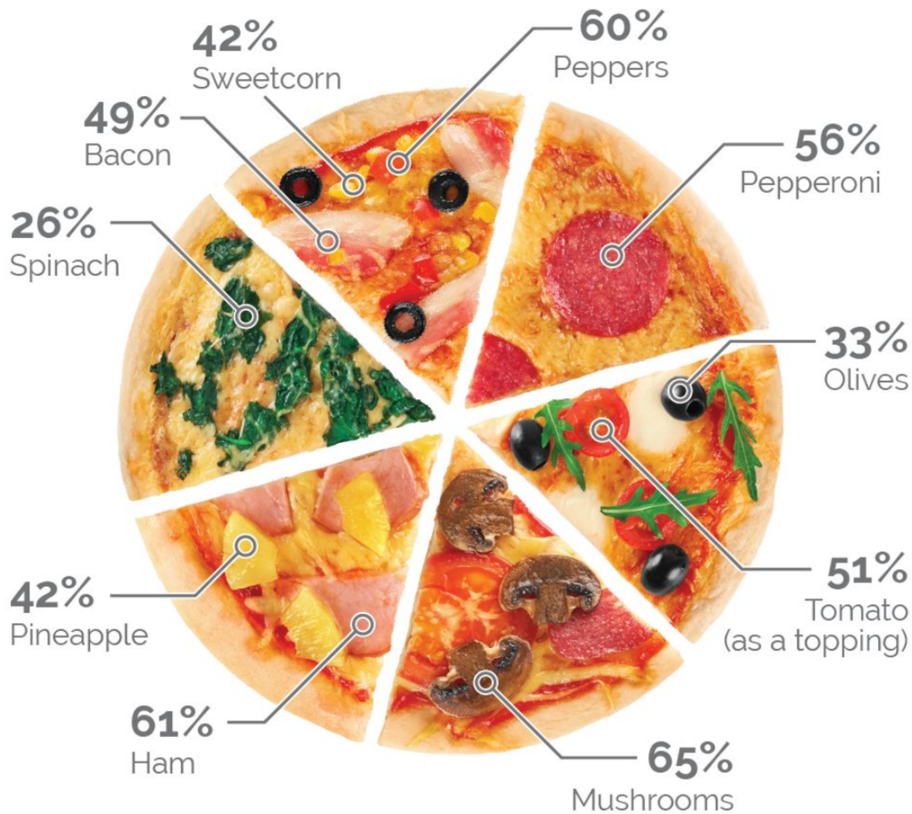
Merging Probability and Data Summarization



Mushroom is the UK's most liked pizza topping

Generally speaking, which of the following toppings do you like on a pizza? Select as many as you like



Other items not depicted include: onions (62%), chicken (56%), beef (36%), chillies (31%), jalapeños (30%), pork (25%), tuna (22%), anchovies (18%). 2% of people say they only like Margherita pizzas

YouGov | yougov.com

February 26-28, 2017

Random variables: Probability Distributions

- A random variable, X, is some <u>numerical</u> outcomes of a random process
- Toss a coin 10 times, X = # of heads
- Toss a coin until a head appears, X = # of tosses needed
- More random variables:

  1. Toss a die X=points showing
  2. Plant 100 seeds of pumpkins X=% germinating
  3. Test a light bulb X=lifetime of bulb
  4. Test 20 light bulbs X=average lifetime of bulbs

Types of random variables (look familiar?)

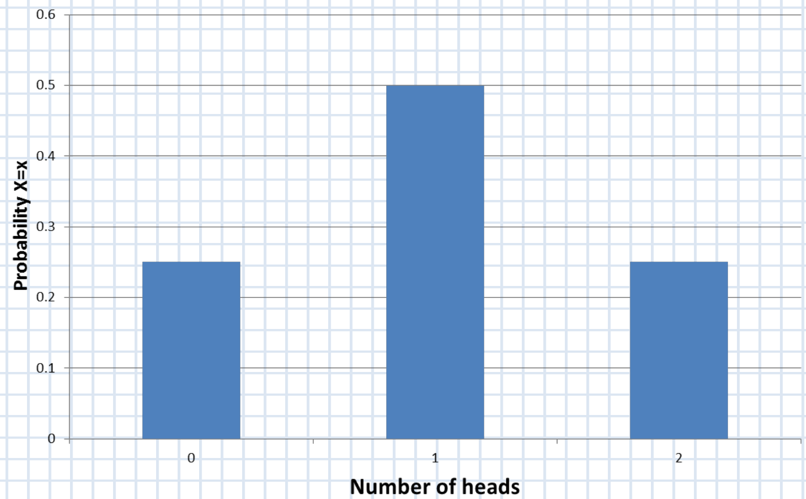
- Discrete or Countable and Counts = finite-possible values
- Continuous = time, lifetimes
- For a discrete random variable, the probability of for each outcome x to occur is denoted by f(x), with properties: $0 \leq f(x) \leq 1$, $\Sigma f(x)=1$
- Example: Roll a (fair) 6-sided die, X = # showing

x	1	2	3	4	5	6
f(x)	1/6	1/6	1/6	1/6	1/6	1/6

Probability distributions

- The table looks similar to a frequency table of the data, but it is actually the theoretical distribution
- If you perform an infinite number of coin tosses, your data may look like this table

Number of heads	Relative frequency
0	.25
1	.5
2	.25



- Note that the probabilities add to 1.
- We are assuming that the probability of a head on each toss is .5
- We are assuming that probability of heads on the first toss is independent of the second toss.

Probability distributions

- We can write down a formula for the probability of event X, which is observing a certain # of heads. This is written in general form $P(X=x)$
  - For 0 heads: $P(X=0)$
  - For 1 head: $P(X=1)$
  - For 2 heads: $P(X=2)$
  - For 1 or more heads: $P(X\geq 1)$
- If you performed the experiment (i.e. 2 coin tosses) once, you'd get 0,1, or 2 heads. If you performed the experiment 10 times and got: 2, 1, 1, 1, 1, 0, 0, 0, 1, 1 heads in each of the 2 tosses, your distribution would be:

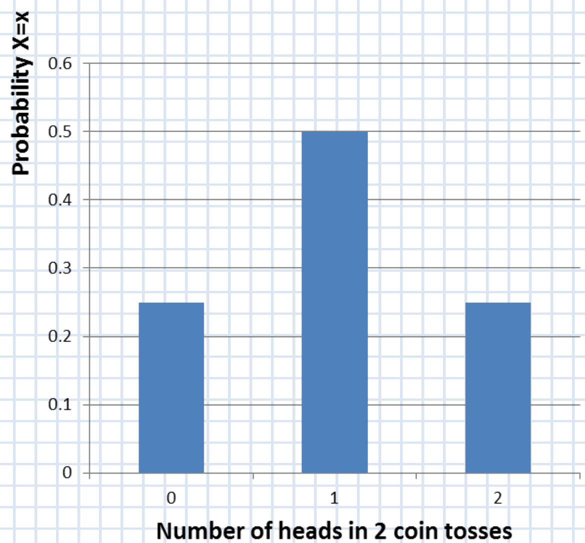
Number of heads	Frequency (%)
0	3 (30)
1	6 (60)
2	1 (10)

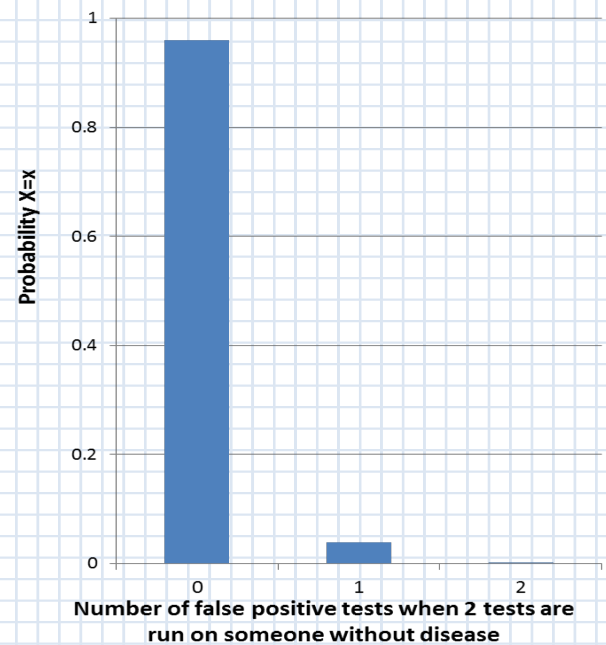
Probability distributions

- In the last video, we looked at two independent diagnostic tests with 2% false positive probability.
  - 2% of the time the test randomly gives a positive result when the true result is negative
- The possible outcomes were NegNeg, NegPos, PosNeg, PosPos
  - Calculating the probability 0, 1, or 2 positive tests:

Number of positive tests	Probability
0	$.98^2 = .9604$
1	$.98*.02*2 = .0392$
2	$.02^2 = .0004$

Probability distributions





These both represent distributions that show the probability of 0, 1, or 2 events occurring in 2 independent trials. Why do they look so different?

Bernoulli (who was he?) random variables

- If you have a variable that can take on one of two values with a constant probability p, then it is a Bernoulli random variable

- If the proportion of people in the population with a disease (the prevalence) is 15%, then when you randomly select one person, the probability that he/she has the disease is $P(Y=1)=p= 0.15$

- The probability that a randomly selected person does not have the disease is

$$P(Y=0)=1-p =0.85$$

Why is this kind of distribution important?

Important Distributions: Binomial distribution

- In many applied problems, we are interested in the probability that an event will occur x times out of n.
- Roll a die 3 times. X = # of sixes. S=a six, N=not a six

No six: (x=0) NNN → (5/6)(5/6)(5/6) = **$(5/6)^3$**

One six: (x=1) NNS → (5/6)(5/6)(1/6) NSN → same SNN → same = **$3(1/6)(5/6)^2$**

Two sixes: (x=2) NSS → (5/6)(1/6)(1/6) SNS → same SSN → same = **$3(1/6)^2(5/6)$**

Three sixes: (x=3) SSS → (1/6)(1/6)(1/6) = **$(1/6)^3$**

Binomial distribution

- Example: The proportion of people in the population with the disease (the prevalence) is 15%, then P(Y=1)=0.15 and P(Y=0)=0.85.
- One random draw will produce either 0 or 1 person with disease
- If we take a random sample of 5 people from this population, there will be 0,1,2,3,4, or 5 people with the disease.
- If the probability of disease in each person is independent, then we can write down the probability of each of these outcomes even before we draw the sample.
- The probability that ALL of them will have the disease is P(X=5):

  $=P(X_1=1)*\ P(X_2=1)*\ P(X_3=1)*\ P(X_4=1)*\ P(X_5=1)$

  $= 0.15 \times 0.15 \times 0.15 \times 0.15 \times 0.15 = 0.00008$

Binomial distribution

- The probability that NONE of them will have the disease is P(X=0):
  $=P(X_1=0)^* P(X_2=0)^* P(X_3=0)^* P(X_4=0)^* P(X_5=0) = (0.85)^5 = 0.444$
- The probability that <u>exactly one person</u> P(X=1) of 5 has the disease = .392
- The probability that <u>exactly two people</u> P(X=2) of 5 have the disease = .138
- The probability that <u>exactly three people</u> P(X=3) of 5 have the disease = .024
- The probability that <u>exactly four people</u> P(X=4) of 5 have the disease = .002
- The probability that <u>exactly five people</u> P(X=5) of 5 have the disease = .00008

Each of these would be = $(0.85)^x * (0.15)^{5-x}$

Binomial distribution

- Assumptions:
  - There are a fixed number of trials n
  - Within each trial, there are two mutually exclusive outcomes (heads vs. tails, disease vs. no disease)
  - The outcomes of the n trials are independent
  - The probability of success p is constant for each trial

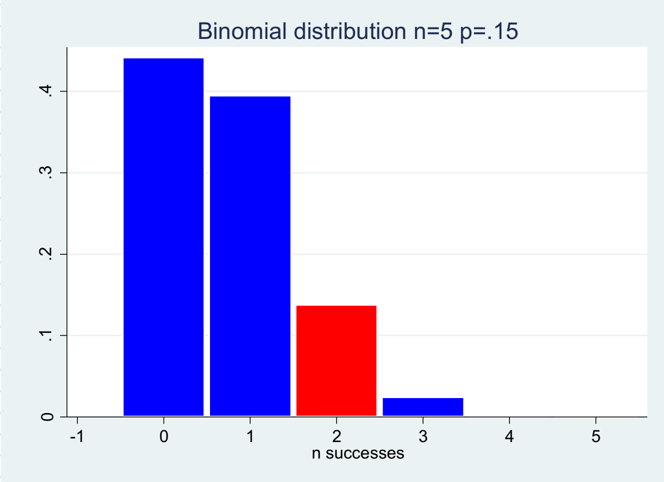
Using Stata for the Binomial distribution

Stata binomialp()

What is the probability of exactly 2 cases of disease in a sample of n=5 where p=0.15?

–Use binomialp(n,k,p) to get P(X=k) in n trials with probability of success in each trial=p

- `di binomialp(5,2,.15)`
  `.13817813`



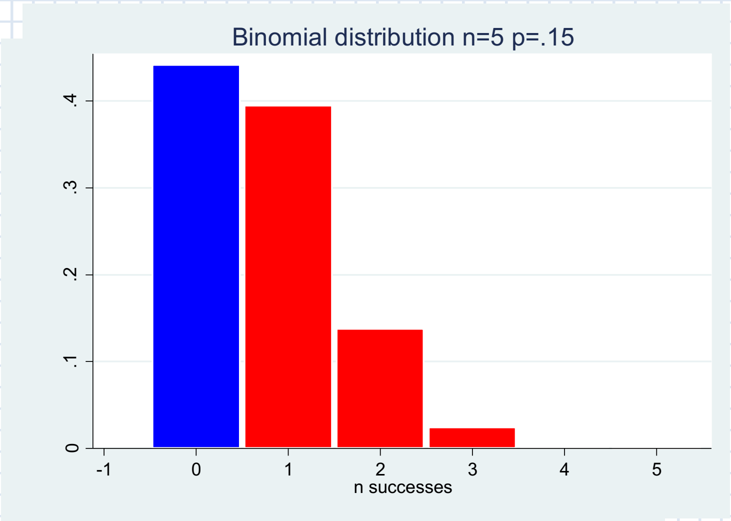
Using Stata for the Binomial distribution

Stata binomialp()

What is the probability of one or more cases of disease in a sample of n=5 where p=0.15?

–Use binomialtail(n,k,p) to get P(X>=k) in n trials with probability of success in each trial=p

- `display binomialtail(5,1,.15)`
  `.55629469`



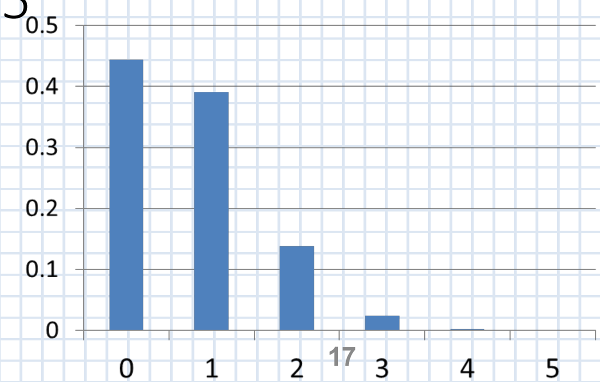
Summary Statistics: Means and Variances

- In Module 1 we learned how to calculate the sample mean, variance, and standard deviation
- Theoretical distributions have theoretical means, variances, and standard deviations
- For a discrete distribution, the mean of the distribution is

  $\sum_{i=0}^{n} x_i \, P(X = x_i)$ over all the possible values of x.

Binomial distribution

- The <u>mean</u> of a binomially distributed random variable X is *np*
- This means that over an large number of samples of size n with probability p of success, the mean number of successes ($\bar{x}$) over the samples will be approximately *np*
- In shorthand, *np* is the mean of the binomial distribution
- For our example with n=5 and p=P(disease)=.15 the mean of the distribution of the number of cases is 5*.15 = .75

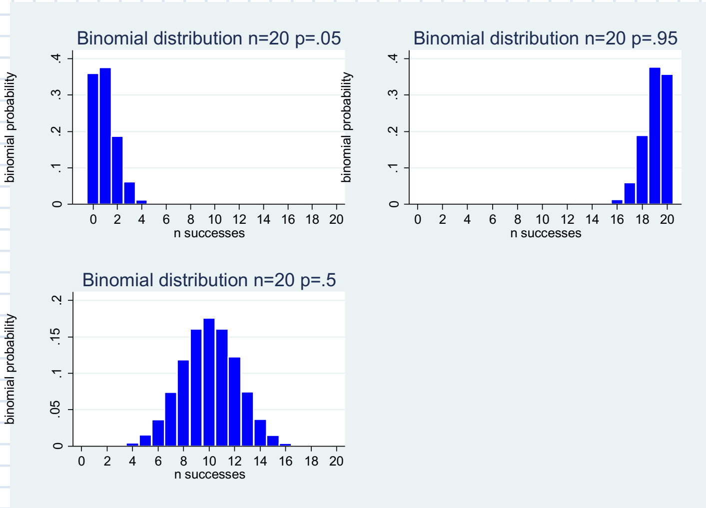


Binomial distribution

- The variance of a binomially distributed random variable X is *n*p*(1-p)*
- This means that over a large number of samples of size n, the *sample variance* of the X's, the number of successes, will be approximately *n*p*(1-p)*
- The standard deviation is $\sqrt{np(1-p)}$
- For our example with n=5 and p=P(disease)=.15, the variance is
  - n*p*(1-p) = 5*.15 *.85
  - If you took a sample of n=5, repeatedly, the sample variance in the number of cases over these samples would be this number

Binomial distribution: Summary

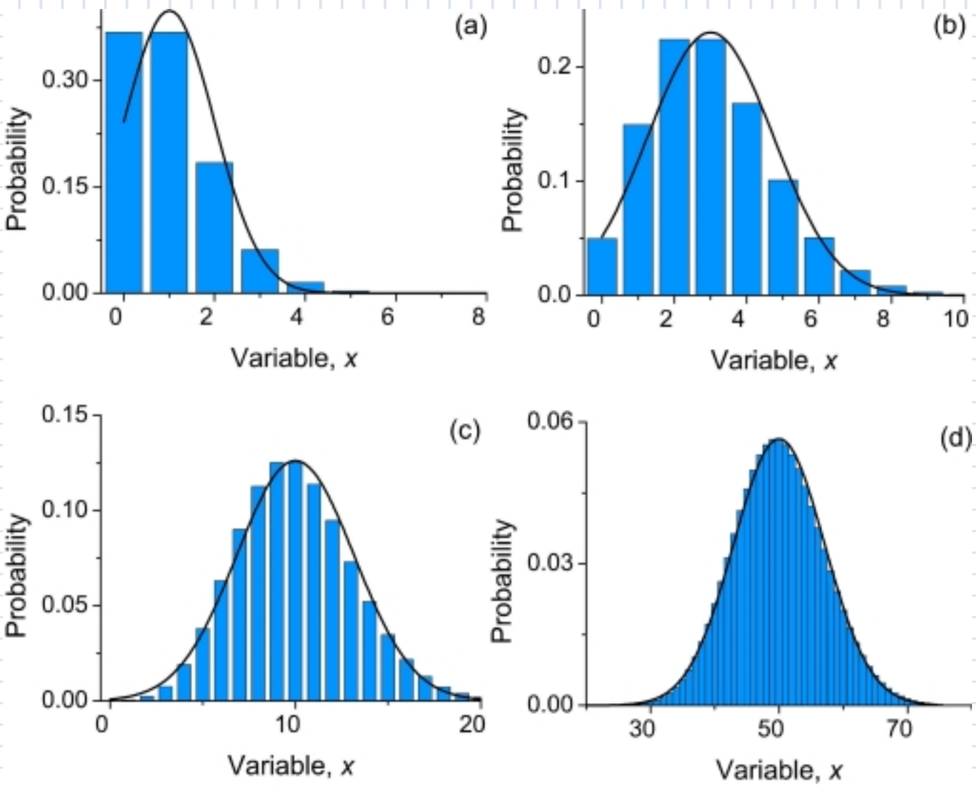
- Binomial mean = np
- Binomial variance= np(1-p)
  - Variance is largest when p=0.5, smaller when p closer to 0 or 1
  - The distribution is symmetric when p=0.5
  - The distribution is a mirror image for 1-p: the distribution for p=0.05 is the mirror image of the one for p=0.95



Important Distributions: Poisson distribution

- In general, this discrete distribution counts the number of events in a continuous interval of time
- Events happen independently in time or space with, on average, λ events per unit time or space.
  - Waiting for BART λ=1 train per 10 minutes
  - Radioactive decay λ=2 particles per minute
  - Lightening strikes λ=0.01 strikes per acre
- The Poisson has only one parameter, λ, that is the mean number of events and also the variance
- Described as the distribution of rare events

Poisson Distribution with varying means



$$P(X = x) = \frac{\lambda^{x} e^{-\lambda}}{x!}$$

Other Important Related Distributions

- The companion distribution to the Poisson is the Exponential distribution (continuous). It calculates the probability of the length of time until a certain number of events. You will see this often used as a transformation for your data (and with odds ratios)

- And extension of the Binomial distribution is the Hypergeometric distribution. It calculates probabilities when additional information about the outcomes is available and samples without replacement (what does that mean?)

We will use these in Stata in your Lab.