

# BIOS 600: Principles of Statistical Inference

## Discrete Probability Distributions

Fall 2012

# Reading

- ▶ Pagano and Gauvreau, Chapter 7, Sections 7.1-7.2

# Probability Distributions

When we learned how to find probabilities by applying basic principles, we focused on one particular outcome or event, like the probability of having a preterm baby, or the probability of having breast cancer given a positive mammogram. We now look at the bigger picture by considering all the possible values of a discrete random variable, along with their associated probabilities. This list of possible values and probabilities is called the *probability distribution* of the random variable.

# Why do we care about probability distributions?

We need probability distributions to make statements about how likely an event may be. For example, suppose we are interested in comparing two groups of pancreatic cancer patients, with one group on the standard treatment and one group on an experimental new treatment. The statistical test we use to compare the two groups depends on the probability distribution of  $X$ . For example, one of you could define  $X$  to be a 0/1 variable indicating whether each patient survives 1 year, and someone else could instead define  $X$  as the number of months a patient survives. In these two different cases,  $X$  has two different probability distributions, and you would use a different type of test for each case.

## Example: Gender

Consider gender of newborns in the US. 51.2% of births are boys, and 48.8% of births are girls. Thus the probability of having a boy is 0.512.

- ▶ The probability of an event is  $\geq 0$  and  $\leq 1$ .
- ▶ The sum of the probabilities of all possible events is exactly 1 (e.g.,  $0.512+0.488=1$ ).

# Bernoulli Distribution

Consider a dichotomous (two-level) random variable  $Y$ . By definition,  $Y$  must assume one of two possible values:

- ▶ Failure or success
- ▶ Dead or alive
- ▶ Male or female
- ▶ Current smoker or not
- ▶ Heads or tails (coin flip)

A random variable of this type is known as a *Bernoulli* random variable, and we describe the probability of response using the parameter  $\pi$ .

# Bernoulli Random Variable

Often coded so that  $Y = 1$  is called an event or success and  $Y = 0$  is called a failure, and  $\pi$  is defined as the probability of a success. This is not required, but you do need to know which level has probability corresponding to  $\pi$  (the other level will have probability  $1 - \pi$  by definition). Examples:

- ▶ Coin flip: let  $Y = 1$  if heads and  $Y = 0$  if tails (can flip those), then  $\Pr(Y = 1) = \pi = 0.50 = \Pr(Y = 0)$
- ▶ Gender at birth in US: let  $Y = 1$  if male and  $Y = 0$  if female, then  $\Pr(Y = 1) = \pi = 0.512$  and  $\Pr(Y = 0) = 1 - \pi = 0.488$
- ▶ Gender at birth in China:  $\Pr(Y = 1) = \pi = 0.53$  and  $\Pr(Y = 0) = 1 - \pi = 0.47$
- ▶ Vegetarian in US:  $Y = 1$  if vegetarian and  $Y = 0$  if not,  $\Pr(Y = 1) = \pi = 0.03$  and  $\Pr(Y = 0) = 1 - \pi = 0.97$
- ▶ Vegetarian in India:  $\Pr(Y = 1) = \pi = 0.31$  and  $\Pr(Y = 0) = 1 - \pi = 0.69$

# Bernoulli Distribution

- ▶  $Y$  takes value 1 with probability  $\pi$  and 0 with probability  $1 - \pi$
- ▶  $\Pr(Y = y) = \pi^y(1 - \pi)^{1-y}$
- ▶ So  $\Pr(Y = 1) = \pi^1(1 - \pi)^0 = \pi$  as  $x^0 = 1$  for any  $x$
- ▶ Similarly,  $\Pr(Y = 0) = \pi^0(1 - \pi)^1 = 1 - \pi$
- ▶ We don't really need the formality of stating this distribution, as it is probably simpler just to keep track of  $\pi$  and  $1 - \pi$
- ▶ However, we want to extend this to a more complex setting: in a randomly selected group of 3 students, how surprising would it be to get 2 smokers?



## Case Study: Smoking

- ▶ The CDC reports that roughly 20% of US adults are smokers.
- ▶  $\Pr(Y = 1) = \Pr(\text{Smoker}) = \pi = 0.2$  and  $\Pr(Y = 0) = 1 - 0.2 = 0.8$
- ▶ Now suppose we randomly select two adults in the US and let a new random variable  $X$  represent the number of smokers:  $X$  can be 0, 1, or 2. Assume these persons are independent.

First Person's $Y$	Second Person's $Y$	Number of Smokers $X$	Probability of These Outcomes
0	0	0	
1	0	1	
0	1	1	
1	1	2	

## Case Study: Smoking

Let  $y_1$  take value 1 if the first person smokes and take value 0 otherwise, and define  $y_2$  similarly for the second person.

Let  $A_1$  be the event that  $y_1 = 1$  and  $A_2$  be the event that  $y_2 = 1$ .

Because the people are independent, then

$$\begin{aligned} Pr(\text{both smokers}) &= Pr(A_1 \cap A_2) \\ &= Pr(A_1)Pr(A_2 \mid A_1) \\ &= Pr(A_1)Pr(A_2) \\ &= \pi\pi = 0.2(0.2) = 0.04. \end{aligned}$$

## Case Study: Smoking

Row 1 asks for  $Pr(\bar{A}_1 \cap \bar{A}_2) = Pr(\bar{A}_1)Pr(\bar{A}_2) = (1 - \pi)(1 - \pi)$

(recalling  $Pr(\bar{A}_2 | \bar{A}_1) = Pr(\bar{A}_2)$  due to independence of observations)

First Person's $Y$	Second Person's $Y$	Number of Smokers $X$	Probability of These Outcomes
0	0	0	$(1 - \pi)(1 - \pi) = 0.8(0.8) = 0.64$
1	0	1	
0	1	1	
1	1	2	

## Case Study: Smoking

Row 2 asks for  $Pr(A_1 \cap \bar{A}_2) = Pr(A_1)Pr(\bar{A}_2) = (\pi)(1 - \pi)$

First Person's $Y$	Second Person's $Y$	Number of Smokers $X$	Probability of These Outcomes
0	0	0	$(1 - \pi)(1 - \pi) = 0.8(0.8) = 0.64$
1	0	1	$\pi(1 - \pi) = 0.2(0.8) = 0.16$
0	1	1	
1	1	2	

## Case Study: Smoking

Row 3 asks for

$$Pr(\bar{A}_1 \cap A_2) = Pr(\bar{A}_1)Pr(A_2) = (1 - \pi)(\pi) = 0.8(0.2) = 0.16$$

First Person's $Y$	Second Person's $Y$	Number of Smokers $X$	Probability of These Outcomes
0	0	0	$(1 - \pi)(1 - \pi) = 0.8(0.8) = 0.64$
1	0	1	$\pi(1 - \pi) = 0.2(0.8) = 0.16$
0	1	1	$(1 - \pi)\pi = 0.8(0.2) = 0.16$
1	1	2	

## Case Study: Smoking

Row 4 asks for

$$Pr(A_1 \cap A_2) = Pr(A_1)Pr(A_2) = (\pi)(\pi) = 0.2(0.2) = 0.04$$

First Person's $Y$	Second Person's $Y$	Number of Smokers $X$	Probability of These Outcomes
0	0	0	$(1 - \pi)(1 - \pi) = 0.8(0.8) = 0.64$
1	0	1	$\pi(1 - \pi) = 0.2(0.8) = 0.16$
0	1	1	$(1 - \pi)\pi = 0.8(0.2) = 0.16$
1	1	2	$\pi\pi = 0.2(0.2) = 0.04$

## Case Study: Smoking

Thus the probability distribution of number of smokers out of two people is given by

$X$	0	1	2
$Pr(X = x)$	0.64	0.32	0.04

So if we randomly sample two people from the US population, the probability that both are smokers is 0.04 or a 4% chance. The probability both are nonsmokers is 0.64 or a 64% chance. The probability that only one smokes is  $0.16+0.16=0.32$  or a 32% chance (there are two ways this can happen).

## Case Study: Smoking

If we randomly sample 3 people, what is the chance all 3 are smokers?

First Person's $Y$	Second Person's $Y$	Third Person's $Y$	Number of Smokers $X$	Probability of These Outcomes
0	0	0	0	
1	0	0	1	
0	1	0	1	
0	0	1	1	
1	1	0	2	
1	0	1	2	
0	1	1	2	
1	1	1	3	



## Case Study: Smoking

Row 1 asks for  $Pr(\bar{A}_1 \cap \bar{A}_2 \cap \bar{A}_3)$ , which due to independence of individuals can be written

$$Pr(\bar{A}_1)Pr(\bar{A}_2)Pr(\bar{A}_3) = (1-\pi)(1-\pi)(1-\pi) = 0.8(0.8)(0.8) = 0.512.$$

First Person's $Y$	Second Person's $Y$	Third Person's $Y$	Number of Smokers $X$	Probability of These Outcomes
0	0	0	0	$(1 - \pi)(1 - \pi)(1 - \pi) = 0.8(0.8)(0.8) = 0.512$
1	0	0	1	
0	1	0	1	
0	0	1	1	
1	1	0	2	
1	0	1	2	
0	1	1	2	
1	1	1	3	

## Case Study: Smoking

Row 2 asks for

$$\begin{aligned}Pr(A_1 \cap \bar{A}_2 \cap \bar{A}_3) &= Pr(A_1)Pr(\bar{A}_2)Pr(\bar{A}_3) \\&= (\pi)(1 - \pi)(1 - \pi) = 0.2(0.8)(0.8) = 0.128.\end{aligned}$$

First Person's $Y$	Second Person's $Y$	Third Person's $Y$	Number of Smokers $X$	Probability of These Outcomes
0	0	0	0	$(1 - \pi)(1 - \pi)(1 - \pi) = 0.8(0.8)(0.8) = 0.512$
1	0	0	1	$\pi(1 - \pi)(1 - \pi) = 0.128$
0	1	0	1	
0	0	1	1	
1	1	0	2	
1	0	1	2	
0	1	1	2	
1	1	1	3	

Similarly, we can fill in the rest of the table.

## Case Study: Smoking

If we randomly sample 3 people, what is the chance all 3 are smokers?

First Person's Y	Second Person's Y	Third Person's Y	Number of Smokers X	Probability of These Outcomes
0	0	0	0	$(1 - \pi)(1 - \pi)(1 - \pi) = 0.8(0.8)(0.8) = 0.512$
1	0	0	1	$\pi(1 - \pi)(1 - \pi) = 0.128$
0	1	0	1	$(1 - \pi)\pi(1 - \pi) = 0.128$
0	0	1	1	$(1 - \pi)(1 - \pi)\pi = 0.128$
1	1	0	2	$\pi\pi(1 - \pi) = 0.032$
1	0	1	2	$\pi(1 - \pi)\pi = 0.032$
0	1	1	2	$(1 - \pi)\pi\pi = 0.032$
1	1	1	3	$\pi\pi\pi = 0.008$

So this probability is fairly small, 0.008 or  $< 1\%$ . The chance that 2 of 3 are smokers is  $0.032 + 0.032 + 0.032 = 0.096$  or  $9.6\%$ .

## Case Study: Smoking

Thus the probability distribution of number of smokers out of three people is given by

$X$	0	1	2	3
$Pr(X = x)$	0.512	0.384	0.096	0.008

## Case Study: Smoking

If we randomly sample 4 people, what is the chance all 4 are smokers?

This is getting ridiculous, now we need a formula! We can use the *binomial distribution* to help determine this probability.

# Binomial distribution

The *binomial distribution* is used to give us the probability of  $X$  'successes' from a sequence of  $n$  independent Bernoulli trials. In our example, each person would be an independent Bernoulli trial (either a smoker, or not). This distribution involves three assumptions.

- ▶ There is a fixed number of Bernoulli trials,  $n$ , each of which results in one of two mutually exclusive outcomes.
- ▶ The outcomes of the  $n$  trials are independent.
- ▶ The probability of success  $\pi$  is the same for each trial.

The distribution is

$$\Pr(Y = y) = \binom{n}{y} \pi^y (1 - \pi)^{n-y},$$

and it has mean  $n\pi$  and variance  $n\pi(1 - \pi)$ .

## Binomial distribution

$$\Pr(Y = y) = \binom{n}{y} \pi^y (1 - \pi)^{n-y}$$

Wait, it's really not so bad!

First, let's look at the second part,  $\pi^y (1 - \pi)^{n-y}$ . This is just multiplying the right combination of  $\pi$  and  $1 - \pi$  as in the previous tables.

- ▶ If we are interested in the probability of three smokers ( $Y = 3$ ), the second part is just  $\pi^y (1 - \pi)^{n-y} = 0.2^3 (0.8)^0 = 0.008(1) = 0.008$  (same as in table).

## Binomial distribution

$$\Pr(Y = y) = \binom{n}{y} \pi^y (1 - \pi)^{n-y}$$

- ▶ If we are interested in the probability of two smokers ( $Y = 2$ ), the second part is just  $0.2^2(0.8)^{3-2} = 0.2^2(0.8)^1 = 0.032$ , which is what we see in any single row in which we have two smokers and one nonsmoker.
- ▶ This is the probability of any one specific combination of two smokers and one nonsmoker. Then we need to figure out how many combinations of two smokers and one nonsmoker we could get.
- ▶ The first part,  $\binom{n}{y}$ , accounts for all the possible ways in which we can have two smokers.



## Combinations

The first part is  $\binom{n}{y}$  (pronounced “n choose y”). This is called a *combination* (and in this setting is called the *binomial coefficient*) and is a formula for the number of ways to pick  $y$  subjects from a larger group of  $n$ . It is defined as

$$\binom{n}{y} = \frac{n!}{y!(n-y)!}.$$

Uhh, what's  $n!$ ?  $n!$  is just  $n(n-1)(n-2)\dots(1)$ . So  $3! = 3(2)(1) = 6$ ,  $4! = 4(3)(2)(1) = 24$ , and so forth. We define  $0! = 1$ .

How many ways can we pick 3 subjects from a group of three? Well, that's easy, just one way. We verify using

$$\binom{3}{3} = \frac{3!}{3!0!} = \frac{3(2)(1)}{3(2)(1)(1)} = 1.$$

# Combinations

$$\binom{n}{y} = \frac{n!}{y!(n-y)!}$$

How many ways can we pick 2 subjects from a group of 3? Well, from the table, we can see there are three ways.

First Person's $Y$	Second Person's $Y$	Third Person's $Y$	Number of Smokers $X$	Probability of These Outcomes
0	0	0	0	$(1 - \pi)(1 - \pi)(1 - \pi) = 0.8(0.8)(0.8) = 0.512$
1	0	0	1	$\pi(1 - \pi)(1 - \pi) = 0.128$
0	1	0	1	$(1 - \pi)\pi(1 - \pi) = 0.128$
0	0	1	1	$(1 - \pi)(1 - \pi)\pi = 0.128$
1	1	0	2	$\pi\pi(1 - \pi) = 0.032$
1	0	1	2	$\pi(1 - \pi)\pi = 0.032$
0	1	1	2	$(1 - \pi)\pi\pi = 0.032$
1	1	1	3	$\pi\pi\pi = 0.008$

# Combinations

$$\binom{n}{y} = \frac{n!}{y!(n-y)!}$$

How many ways can we pick 2 subjects from a group of 3?

We verify using  $\binom{3}{2} = \frac{3!}{2!1!} = \frac{3(2)(1)}{2(1)(1)} = 3$ .

## Probability of Getting 3 Smokers in a Group of 3

So the probability of getting three smokers is

$$\begin{aligned}\binom{n}{y} \pi^y (1 - \pi)^{n-y} &= \frac{n!}{y!(n-y)!} \pi^y (1 - \pi)^{n-y} \\ \binom{3}{3} 0.2^3 (1 - 0.2)^{3-3} &= \frac{3!}{3!(3-3)!} 0.2^3 (1 - 0.2)^{3-3} \\ &= \frac{(3)(2)(1)}{(3)(2)(1)(1)} 0.2^3 (0.8)^0 \\ &= 0.008.\end{aligned}$$

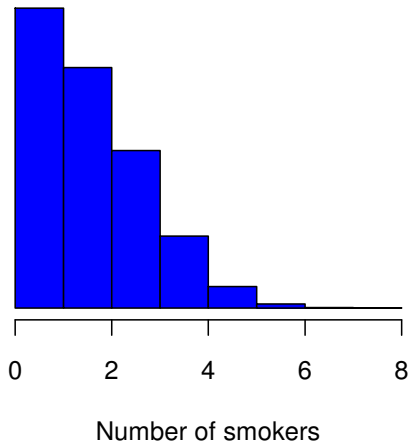
## Probability of Getting 2 Smokers in a Group of 3

The probability of getting 2 smokers is

$$\begin{aligned}\binom{n}{y} \pi^y (1 - \pi)^{n-y} &= \frac{n!}{y!(n-y)!} \pi^y (1 - \pi)^{n-y} \\ \binom{3}{2} 0.2^2 (1 - 0.2)^{3-2} &= \frac{3!}{2!(3-2)!} 0.2^2 (1 - 0.2)^{3-2} \\ &= \frac{(3)(2)(1)}{(2)(1)(1)} 0.2^2 (0.8)^1 \\ &= 3(0.032) = 0.096.\end{aligned}$$

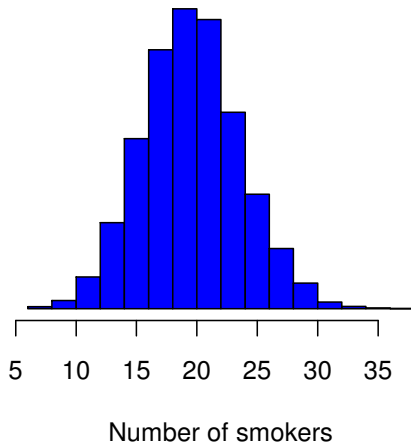
## Distribution of smokers in a group size $n = 10$

**Binomial  $n=10$   $p=.20$**



# Distribution of smokers in a group size $n = 100$

**Binomial  $n=100$   $p=.20$**



## In Vitro Fertilization (IVF)

Some estimates indicate that up to 1-2% of conceptions in the Western world are through IVF. Suppose the average success (live birth) rate of this procedure per embryo is 35%. IVF is expensive and not covered by health insurance in most US states. Because it is expensive, we wish to minimize the needed number of IVF treatments (i.e., maximize the success rate of a given IVF attempt).

To maximize the success rate, specialists usually implant more than one embryo. We want to find the optimum number of embryos to implant so that the likelihood of having at least one baby is high, but the likelihood of having triplets, quadruplets, etc. is low. Ideally we want a singleton birth, twins are acceptable (though not without risk), and the risk of higher-order births should be minimized. Assume that the probability of each embryo surviving is independent of the others.



# Learning by Doing: Probability Distributions

How do we address this mathematically? Here's what we need to know.

- ▶ Probability of having at least one baby (how to write mathematically?)
  - ▶  $Pr(Y \geq 1)$
- ▶ Probability of having more than twins (how to write mathematically?)
  - ▶  $Pr(Y > 2)$
- ▶ Let  $n$  represent the number of embryos implanted in one woman

# Learning by Doing: Probability Distributions

Binomial distribution:  $\Pr(Y = y) = \binom{n}{y} \pi^y (1 - \pi)^{n-y}$

- ▶  $y$  is the number of live births from  $n$  implanted embryos
- ▶ The probability of a live birth is  $\pi = 0.35$ .
- ▶ If we implant only 1 embryo,  $n = 1$ , if we implant two embryos,  $n = 2$ , if we implant 10 embryos,  $n = 10$ . ASRM (American Society for Reproductive Medicine) has maternal age-dependent guidelines on the number of embryos to implant, ranging from 1-5.
- ▶ We can use the binomial distribution to find  $\Pr(Y \geq 1)$  (probability of live birth) and  $\Pr(Y > 2)$  (probability of too many multiples) for  $n = 1, 2, 3, 4, 5$ .

# Learning by Doing: Probability Distributions

Binomial distribution:  $\Pr(Y = y) = \binom{n}{y} 0.35^y (1 - 0.35)^{n-y}$

- ▶ We can use the binomial distribution to find  $\Pr(Y \geq 1)$  (probability of live birth) and  $\Pr(Y > 2)$  (probability of too many multiples) for  $n = 1, 2, 3, 4, 5$ .
- ▶  $\Pr(Y \geq 1) = 1 - \Pr(Y = 0)$  (either you have no live births, or you have 1+ live births)
- ▶  $\Pr(Y > 2) = 1 - \Pr(Y = 0) - \Pr(Y = 1) - \Pr(Y = 2)$
- ▶ In the simple case that  $n = 1$ , then  $\Pr(Y \geq 1) = 0.35$  and  $\Pr(Y > 2) = 0$

# Learning by Doing: Probability Distributions

Binomial distribution:  $\Pr(Y = y) = \binom{n}{y} 0.35^y (1 - 0.35)^{n-y}$

- ▶  $\Pr(Y \geq 1) = 1 - \Pr(Y = 0) = 1 - \binom{n}{0} 0.35^0 (0.65)^{n-0} = 1 - 0.65^n$
- ▶  $\Pr(Y > 2) = 1 - \Pr(Y = 0) - \Pr(Y = 1) - \Pr(Y = 2)$   
 $= 1 - 0.65^n - \binom{n}{1} 0.35^1 (0.65)^{n-1} - \binom{n}{2} 0.35^2 (0.65)^{n-2}$   
 $= 1 - 0.65^n - n(0.35)(0.65)^{n-1} - \frac{n!}{2!(n-2)!} 0.35^2 (0.65)^{n-2}$   
 $= 1 - 0.65^n - n(0.35)(0.65)^{n-1} - \frac{n(n-1)}{2} 0.35^2 (0.65)^{n-2}$
- ▶ Suppose  $n = 5$  then we have
  - ▶  $\Pr(Y \geq 1) = 1 - \Pr(Y = 0) = 1 - 0.65^5 = 0.88$  (good chances of a baby!)
  - ▶  $\Pr(Y \geq 2) = 1 - \Pr(Y = 0) - \Pr(Y = 1) - \Pr(Y = 2) = 1 - 0.65^5 - 5(0.35)(0.65)^{5-1} - 10(.35)^2 (.65)^3 = 0.23$  (and of lots o' babies!)

## Learning by Doing: Probability Distributions

Binomial distribution:  $\Pr(Y = y) = \binom{n}{y} 0.35^y (1 - 0.35)^{n-y}$

Embryos implanted ( $n$ )	$\Pr(\text{at least one baby})$	$\Pr(\text{more than 2 babies})$
1	0.35	0
2	0.58	0
3	0.73	0.04
4	0.82	0.13
5	0.88	0.23

Using this table of probabilities, and keeping in mind we want at least one baby but not triplets or higher-order births, how many embryos would you implant, and why?

## Other Discrete Distributions

The binomial is not the only probability distribution for discrete data, though you will most likely see it MUCH more often than any other. You may encounter the Poisson distribution, which is sometimes appropriate for count data. This distribution is named for the French mathematician and statistician Poisson.

Interestingly, Poisson wanted to be a surgeon, but due to his extreme lack of coordination killed his first patient. He redirected his efforts and became an extremely successful mathematics professor! There are a large number of other discrete probability distributions, including the geometric and negative binomial.

## For next time

In the next class we'll work some problems, just review the lecture notes before you come and be sure to bring pencil and paper.