

# *Introduction to the Theory of Inference*

Jon Cryer, University of Iowa  
Jeff Witmer, Oberlin College

Statistics is the systematic study of variation in data: how to display it, measure it, model it, and use it to gain new knowledge.

Data may come from a carefully designed experiment, a sample survey, or from a data bank of regularly kept records. But all data vary. Even in a carefully controlled experiment, results vary to some degree.

Measuring devices such as thermometers and scales are imperfect. Results also vary when an experiment is performed under the same conditions. Variability arises from seasonal factors such as the increase in retail sales preceding seasonal holidays and from environmental factors such as the natural warming that takes place during the day.

In surveys variability arises because different people have different opinions, different ages, different cultures, and so on. Survey results also vary because we only collect data on a part of the group we wish to study -- a sample.

Variability is inherent in various processes -- when a coin is tossed it sometimes comes up tails and sometimes heads. Variation, variation, variation! We need to learn in spite of variation.

## **Review of random variables**

We begin with a set  $S$  containing a finite or countable number of outcomes of a chance experiment, the atoms of individual things that can happen. This set  $S$  is known as the *discrete sample space*. For example, suppose we roll a standard die. The sample space associated with the die-tossing consists of six sample points, corresponding to the six simple events which naturally correspond to the number of pips showing on the die:

$$S = \{E_1, E_2, E_3, E_4, E_5, E_6\}$$

Any subset  $E$  of the sample space is known as an *event*. We may create new events by intersections, unions, and complements, of existing events. For example, the event  $E$  described by "getting an even number" is the union of three events from the sample space. That is,  $E = E_2 \cup E_4 \cup E_6$ . Once we have defined a simple event,  $E$ , we may associate a real number with it; a *random variable* is a real-valued function with the set of simple events as its domain. Associated with the random variable is a recipe for assigning probabilities to events, determining the *probability mass function* in the discrete case, or the *probability density function* for the continuous case. Note that these are sometimes referred to simply as *probability functions*.

There is a small set of circumstances where we can construct "random" processes and easily calculate the probability distributions of outcomes of those processes. We don't always have the basic tools for building random variables; we can study the "standard" random variables of statistics mostly because we have the mathematical tools necessary for the job. Some of these elementary situations will now be mentioned.

## Basic discrete distributions

Independent trials that result in a success with probability  $p$  and a failure with probability  $1-p$  are called *Bernoulli* trials. The outcomes, success and failure, form a dichotomy and are usually given numeric values of 0 for failure and 1 for success. Three commonly occurring discrete random variables are constructed from Bernoulli trials: the *binomial*, *geometric*, and *negative binomial*.

### Binomial random variable

Suppose that a fixed number of Bernoulli trials are performed, and the number of successes is recorded. If  $Y$  represents the number of successes that occur in  $n$  trials, then  $Y$  is said to be a *binomial random variable* with parameters  $(n, p)$ . This is often written as  $Y \sim B(n, p)$ . The probability mass function for a binomial random variable with parameters  $(n, p)$  is given by:

$$p(y) = \binom{n}{y} p^y (1-p)^{n-y} \quad y = 0, 1, \dots, n \text{ and } 0 \leq p \leq 1.$$

For example, if we are rolling a die and call a success having either 1 or 2 spots showing on a roll, then the probability of success is  $\frac{1}{3}$ . If we roll the die 5 times, we want to know the distribution of  $Y$ , the number of successes in 5 trials. Then  $Y \sim B(5, \frac{1}{3})$ . The sample space for  $Y$  is  $S = \{0, 1, 2, 3, 4, 5\}$ . Thus,

$$p(0) = \binom{5}{0} \left(\frac{1}{3}\right)^0 \left(\frac{2}{3}\right)^5 = 0.132$$

$$p(1) = \binom{5}{1} \left(\frac{1}{3}\right)^1 \left(\frac{2}{3}\right)^4 = 0.329$$

$$p(2) = \binom{5}{2} \left(\frac{1}{3}\right)^2 \left(\frac{2}{3}\right)^3 = 0.329$$

$$p(3) = \binom{5}{3} \left(\frac{1}{3}\right)^3 \left(\frac{2}{3}\right)^2 = 0.165$$

$$p(4) = \binom{5}{4} \left(\frac{1}{3}\right)^4 \left(\frac{2}{3}\right)^1 = 0.041$$

$$p(5) = \binom{5}{5} \left(\frac{1}{3}\right)^5 \left(\frac{2}{3}\right)^0 = 0.004$$

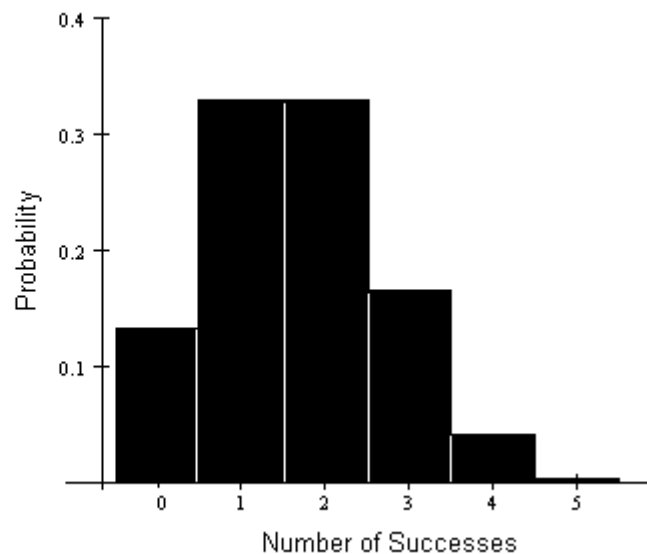


Figure 1: Probability Distribution for Binomial Random Variable  $X \sim B(5, \frac{1}{3})$

**Geometric random variable**

Suppose that Bernoulli trials are performed until a success occurs, and the number of trials is recorded. If  $Y$  represents the number of the trial on which the first success occurs, then  $Y$  is said to be a *geometric random variable* with parameter  $p$ . The probability mass function for a geometric random variable with parameter  $p$  is given by:

$$p(y) = (1 - p)^{y-1} p \quad y = 1, 2, 3, \dots \text{ and } 0 \leq p \leq 1$$

To have the first success on the  $y$ th trial requires a failure in the first  $y - 1$  trials, followed by a success. Notice that the number of trials until the first success is unbounded.

Suppose we roll a die until the first 6 shows. The initial portion of the distribution of the number of trials until the first success is shown below.

$$p(1) = \left(\frac{5}{6}\right)^0 \left(\frac{1}{6}\right) = 0.167$$

$$p(2) = \left(\frac{5}{6}\right)^1 \left(\frac{1}{6}\right) = 0.139$$

$$p(3) = \left(\frac{5}{6}\right)^2 \left(\frac{1}{6}\right) = 0.116$$

$$p(4) = \left(\frac{5}{6}\right)^3 \left(\frac{1}{6}\right) = 0.096$$

$$p(5) = \left(\frac{5}{6}\right)^4 \left(\frac{1}{6}\right) = 0.080$$

$$p(6) = \left(\frac{5}{6}\right)^5 \left(\frac{1}{6}\right) = 0.067$$

$$p(7) = \left(\frac{5}{6}\right)^6 \left(\frac{1}{6}\right) = 0.056$$

$$p(8) = \left(\frac{5}{6}\right)^7 \left(\frac{1}{6}\right) = 0.047$$

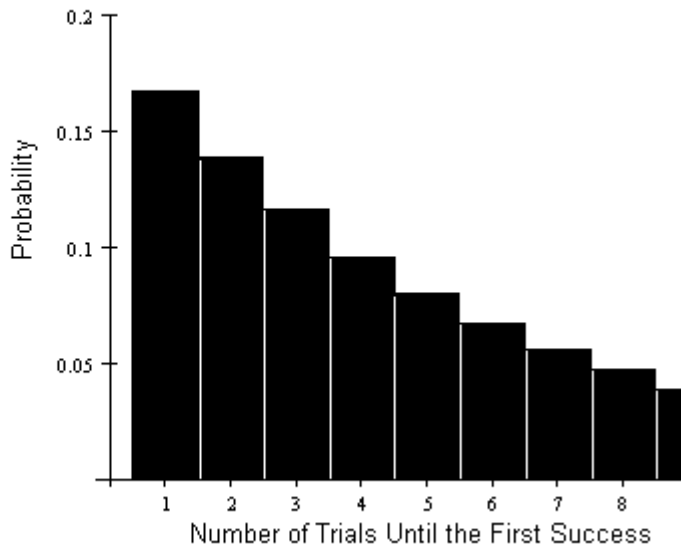


Figure 2: Probability Distribution for Geometric Random Variable with  $p = \frac{1}{6}$

Caution: Some authors define a Geometric random variable as the number of failures preceding the first success. Then the set of possible value is  $\{0, 1, 2, \dots\}$ .

**Negative binomial random variable**

Suppose that Bernoulli trials are performed until a total of  $r$  successes occurs, and the number of trials is calculated. If  $Y$  represents the number of trials that occur until there are  $r$  successes, then  $Y$  is said to be a *negative binomial random variable* with parameters

$(p, r)$ . The probability mass function for a negative binomial random variable with parameters  $(p, r)$  is given by:

$$p(y) = \binom{y-1}{r-1} p^r (1-p)^{y-r} \quad y = r, r+1, \dots \text{ and } 0 \leq p \leq 1$$

To have  $r$  success in  $y$  trials requires us to first have  $y-r$  failures. The last trial must be a success, so the  $r-1$  other successes could come in any one of the  $y-1$  previous trials. Notice that the number of trials is theoretically infinite for the negative binomial.

Return to the first example of rolling a 1 or 2 on a die. If we require 3 successes, the number of rolls must be at least three. The initial portion of the distribution of the number of rolls until three successes are found is shown below.

$$p(3) = \binom{2}{2} \left(\frac{1}{3}\right)^3 \left(\frac{2}{3}\right)^0 = 0.037$$

$$p(4) = \binom{3}{2} \left(\frac{1}{3}\right)^3 \left(\frac{2}{3}\right)^1 = 0.074$$

$$p(5) = \binom{4}{2} \left(\frac{1}{3}\right)^3 \left(\frac{2}{3}\right)^2 = 0.099$$

$$p(6) = \binom{5}{2} \left(\frac{1}{3}\right)^3 \left(\frac{2}{3}\right)^3 = 0.110$$

$$p(7) = \binom{6}{2} \left(\frac{1}{3}\right)^3 \left(\frac{2}{3}\right)^4 = 0.110$$

$$p(8) = \binom{7}{2} \left(\frac{1}{3}\right)^3 \left(\frac{2}{3}\right)^5 = 0.102$$

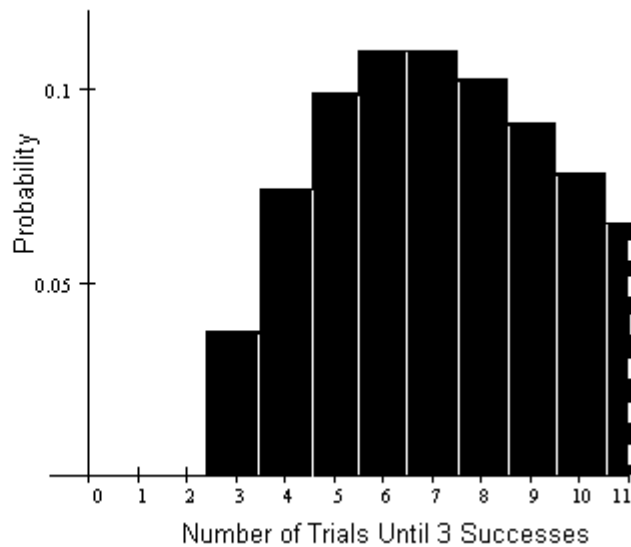


Figure 3: Negative Binomial Distribution with  $p = \frac{1}{3}, r = 3$

### Sampling Without Replacement and the Binomial Distribution

In an introductory course, we typically use the normal approximation to the binomial when considering confidence intervals and hypothesis tests on proportions. Students often question why we can use the binomial distribution when we are sampling without replacement. Doesn't the probability change after each element of the sample is taken? We often answer this question by assuming the population is sufficiently large so taking one or 50 from the population does not effectively change the situation. If the population is small, we will include the "correction factor" for finite populations. How does this all

work out theoretically? We will now discuss an interesting discrete random variable that is formed by a process similar to the process that leads to a binomial random variable. It will turn out that this random variable will have other interesting properties similar to the binomial!

### Hypergeometric random variable

Suppose that a population consists of  $N$  members. Each member is either a "success" (a red marble or a vote for candidate A, or ...) or a "failure". Suppose there are  $r$  successes or proportion of successes  $p = r / N$  in the population. To make inferences about  $r$  or  $p$  we will take a simple random sample without replacement of size  $n$ . Interest centers here on the distribution of the number of successes,  $H$ , in the sample. We will consider two different ways of carrying out the sampling. Note that sampling without replacement is appropriate since it is inefficient to ever collect redundant data from the same member of the population.

Suppose we sample by randomly "scooping out" a subset of size  $n$  ensuring that all subsets of size  $n$  have an equal chance of being selected. With this method, there is no first member drawn, no second member, etc. (With this sampling method it may also be argued that each member of the population has an equal chance of being selected. However, each member being equally likely is not sufficient to ensure that all subsets of size  $n$  are equally likely.) Since all subsets of size  $n$  are equally likely we have the *Hypergeometric Distribution for  $H$*

$$P(H = k) = \frac{\binom{r}{k} \times \binom{N-r}{n-k}}{\binom{N}{n}}, \text{ for } k = 0, 1, \dots, \min(r, n).$$

Alternatively, we could sample sequentially without replacement. Let  $Y_i$  be the number of successes of the  $i^{\text{th}}$  draw (trial). The  $Y_i$  are binary valued since on each draw we get either a success ( $Y_i = 1$ ) or we don't ( $Y_i = 0$ ). The  $Y_i$  are not independent so the trials are not Bernoulli trials. Clearly,  $P(Y_i = 1) = \frac{r}{N} = p$  = the proportion of successes in the population. Now consider  $Y_2$ . We have

$$P(Y_2 = 1 | Y_1 = 1) = \frac{r-1}{N-1}$$

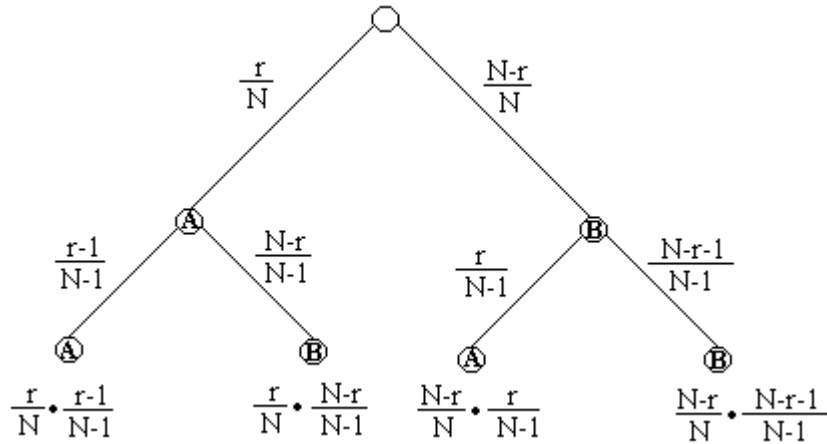
$$P(Y_2 = 1 | Y_1 = 0) = \frac{r}{N-1}$$

According to the Law of Total Probability  $P(A) = P(A|B) \cdot P(B) + P(A|B^C) \cdot P(B^C)$ .

So  $P(Y_2 = 1) = P(Y_2 = 1 | Y_1 = 1) \cdot P(Y_1 = 1) + P(Y_2 = 1 | Y_1 = 0) \cdot P(Y_1 = 0)$

$$\begin{aligned}
 &= \left(\frac{r-1}{N-1}\right) \cdot \frac{r}{N} + \left(\frac{r}{N-1}\right) \left(1 - \frac{r}{N}\right) \\
 &= \frac{r(N-1)}{(N-1)N} \\
 &= \frac{r}{N} = p
 \end{aligned}$$

A tree diagram often helps to illustrate the situation.



By induction we can show that  $P(Y_i = 1) = \frac{r}{N} = p$  for all trials  $i$ . The (unconditional) success probability is the same for all trials just like in Bernoulli trials! These are not Bernoulli trials however since the trials are dependent. (Note that if we were sampling **with replacement** the trials would be Bernoulli trials and  $H$  would have a Binomial Distribution with  $n$  trials and success probability  $p = r / N$  .

The distribution of  $H$  is the same whether we sample randomly by scooping out a subset of members or sample one-at-a-time without replacement. Thinking one-at-a-time permits us to calculate the mean and variance of  $H$  fairly easily. Compare the following calculation with finding the moments directly from the Hypergeometric probability function.

First,  $E(Y_i) = 0(1-p) + 1p = p = \frac{r}{N}$  for all trials. Thus, since  $H = Y_1 + Y_2 + \dots + Y_n$ , we have  $E(H) = E(Y_1 + Y_2 + \dots + Y_n) = E(Y_1) + E(Y_2) + \dots + E(Y_n) = p + p + \dots + p = np = \frac{nr}{N}$  just like the Binomial distribution.

Finding  $Var(H)$  is more difficult since the  $Y$ 's are not independent. We have

$$Var(H) = Var\left(\sum_{i=1}^n Y_i\right) = \sum_{i=1}^n Var(Y_i) + 2 \sum_{i < j} Cov(Y_i, Y_j) .$$

Now, since the  $Y$ 's are binary and all have the same success probability  $p$ ,  $\text{Var}(Y_i) = p(1-p)$  for all  $i$ . In general,

$$\begin{aligned} \text{Cov}(Y_i, Y_j) &= E[(Y_i - \mathbf{m}_i)(Y_j - \mathbf{m}_j)] \\ &= E[Y_i Y_j - \mathbf{m}_i Y_j - \mathbf{m}_j Y_i + \mathbf{m}_i \mathbf{m}_j] \\ &= E(Y_i Y_j) - \mathbf{m}_i E(Y_j) - \mathbf{m}_j E(Y_i) + \mathbf{m}_i \mathbf{m}_j \\ &= E(Y_i Y_j) - \mathbf{m}_i \mathbf{m}_j - \mathbf{m}_j \mathbf{m}_i + \mathbf{m}_i \mathbf{m}_j \\ &= E(Y_i Y_j) - \mathbf{m}_i \mathbf{m}_j \end{aligned}$$

To evaluate  $E(Y_i Y_j)$ , we note that in this case  $Y_i Y_j$  must be either 0 or 1 since the  $Y$ 's are binary. Thus,

$$\begin{aligned} E(Y_i Y_j) &= 1 \cdot P(Y_i Y_j = 1) \\ &= P(Y_i = 1, Y_j = 1) \\ &= P(Y_j = 1 | Y_i = 1) \cdot P(Y_i = 1) \\ &= \frac{(r-1)}{(N-1)} \frac{r}{N} \end{aligned}$$

and

$$\begin{aligned} \text{Cov}(Y_i, Y_j) &= \frac{(r-1)}{(N-1)} \frac{r}{N} - \left(\frac{r}{N}\right)^2 \\ &= \frac{r[N(r-1) - r(N-1)]}{(N-1)N^2} \\ &= -\frac{r(N-r)}{(N-1)N^2} \\ &= -\frac{p(1-p)}{N-1} \end{aligned}$$

(A negative covariance and hence correlation makes sense since a success on one trials tends to reduce slightly the likelihood of a success on another trial.)

Finally, there are  $n$  identical variances and  $\frac{n(n-1)}{2}$  identical covariances to add to obtain  $\text{Var}(H)$ . Thus we have

$$\text{Var}(H) = np(1-p) + 2 \frac{n(n-1)}{2} \left( -\frac{p(1-p)}{N-1} \right)$$

$$= np(1-p) \left( 1 - \frac{n-1}{N-1} \right).$$

The factor  $\sqrt{1 - \frac{n-1}{N-1}}$  which multiplies the standard deviation of the Binomial to obtain the standard deviation for the corresponding Hypergeometric is called the **finite population correction factor**, or *fpc*. (Note: Some authors call  $1 - (n-1)/(N-1)$  the *fpc*.)

Note that if, as is usually the case, the sample size is much smaller than the population size,  $n \ll N$ , the *fpc* will be close to 1 so that the standard deviation from a Binomial distribution gives an excellent approximation to the correct standard deviation. In fact, the Binomial distribution will give an excellent approximation to the correct Hypergeometric distribution.

We very often go one step further. If  $n$  is large but still  $n \ll N$ , we use the Central Limit Theorem to approximate the distribution of  $H$  with a normal distribution with mean  $np$  and variance  $np(1-p)$ . Equivalently, we approximate the distribution of the sample proportion,  $\hat{p} = \frac{H}{n}$  by a normal distribution with mean  $p$  and variance  $\frac{p(1-p)}{n}$ .

## Basic Continuous Distributions

There are many families of continuous random variables, including the familiar normal distributions and uniform distributions. Another important family is the family of gamma distributions.

### Gamma Distributions

We want to define a continuous random variable  $Y$  that has the gamma distribution. We must begin by defining a *gamma function*.

$$\text{Gamma function: } \Gamma(\mathbf{a}) = \int_0^{\infty} y^{\mathbf{a}-1} e^{-y} dy$$

The Gamma function is defined for all  $\mathbf{a} > 0$ . It also has a recursive relationship:

$$\Gamma(\mathbf{a}) = (\mathbf{a}-1)\Gamma(\mathbf{a}-1) \text{ with } \Gamma(1) = 1.$$

To see this, use the technique of integration by parts on

$$\begin{aligned} \Gamma(\mathbf{a}) &= \int_0^{\infty} y^{\mathbf{a}-1} e^{-y} dy \\ u &= y^{\mathbf{a}-1} & dv &= e^{-y} dy \\ du &= (\mathbf{a}-1)y^{\mathbf{a}-2} dy & v &= -e^{-y} \end{aligned}$$

so

$$\int_0^{\infty} y^{a-1} e^{-y} dy = -e^{-y} y^{a-1} \Big|_{y=0}^{y \rightarrow \infty} + (a-1) \int_0^{\infty} y^{a-2} e^{-y} dy.$$

The first part of the sum goes away since  $-e^{-y} y^{a-1}$  is zero when  $y=0$  and goes to zero when  $y \rightarrow \infty$ . This means

$$\Gamma(a) = (a-1) \int_0^{\infty} y^{a-2} e^{-y} dy = (a-1) \Gamma(a-1).$$

We also have  $\Gamma(1) = \int_0^{\infty} y^0 e^{-y} dy = 1$ . One consequence of this is that  $\Gamma(a) = (a-1)!$  when  $a$  is an integer.

Also note that: 
$$\int_0^{\infty} y^{a-1} e^{-y/b} dy = b^a \Gamma(a)$$

Let  $w = y/b$ . Then  $y = bw$  and  $dy = b dw$ .

$$\text{So } \int_0^{\infty} y^{a-1} e^{-y/b} dy = \int_0^{\infty} w^{a-1} b^{a-1} e^{-w} b dw = b^a \int_0^{\infty} w^{a-1} e^{-w} dw = b^a \Gamma(a)$$

A continuous random variable  $Y$  is said to have a gamma distribution with parameters  $a$  and  $b$  if and only if the density function of  $Y$  is  $f(y) = \frac{y^{a-1} e^{-y/b}}{b^a \Gamma(a)}$  for  $y > 0$

and  $a, b > 0$ . It can be shown that  $m = E(Y) = ab$  and  $s^2 = V(Y) = ab^2$

(Reference: *Mathematical Statistics with Applications*, Wackerly, Mendenhall, and Scheaffer, Duxbury Press, 1996, page 159.)

There are several special cases of gamma distributions:

If  $a = 1$  then  $f(y) = \frac{1}{b} e^{-y/b}$  where  $b > 0, y \geq 0$ , and the random variable  $Y$  is said to have an exponential distribution with parameter  $b$ .

If  $b = 2$  and  $a = \frac{u}{2}$ , with  $u$  an integer, then  $f(y) = \frac{y^{\frac{u}{2}-1} e^{-y/2}}{2^{\frac{u}{2}} \Gamma\left(\frac{u}{2}\right)}$ , and the random

variable  $Y$  is said to have a chi-square distribution with  $u$  degrees of freedom.