ANALYSIS OF PROPORTIONS

Andrea Giansanti
Dipartimento di Fisica, Sapienza Università di Roma

Andrea.Giansanti@roma1.infn.it

DA_2022 Lecture n. 11, Rome 6th April 2022

DIPARTIMENTO DI FISICA

SAPIENZA
UNIVERSITÀ DI ROMA

OUTLINE

- proportions (and the binomial distribution)

- The Binomial distribution
- sampling the proportions: parameter estimates, uncertainty
- Testing proportions: the binomial test

Study materials: Whitlock and Schluter chap.7

further reading: interleaf 3: why statistical significance is not the same as biological importance?

Interleaf 4 Correlation does not require causation (see "Spuriuous Correlations" a nice website:

http://www.tylervigen.com/spurious-correlations

PROPORTIONS

SLA

What proportion of people with Lou Gehrig's disease will survive at least 10 years after diagnosis? What proportion of the North Carolina red wolf population is female? In what fraction of years does global temperature increase? Each of these questions is about a proportion, the fraction of the population that has a particular characteristic of interest. The proportion of individuals sharing some characteristic in a population is also the probability that an individual randomly sampled from that population will have that attribute. A proportion can range from zero to one.

. . . remember the toad handedness proble

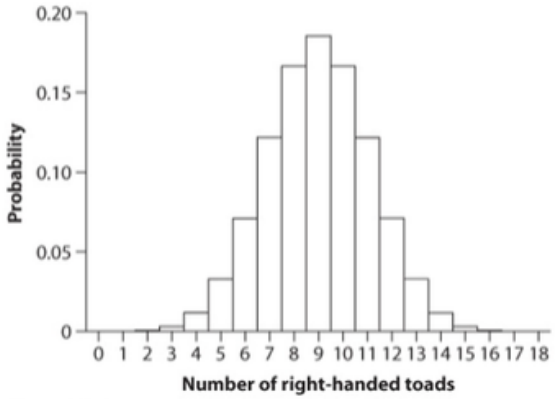

Figure 6.2-1
Whitlock et al., *The Analysis of Biological Data*, 2e, © 2015 W. H. Freeman and Company

The binomial distribution

Consider a measurement made on individuals that divides them into two mutually exclusive groups, such as success or failure, alive or dead, left-handed or right-handed, or diabetic or nondiabetic. In the population, a fixed proportion *p* of individuals fall into one of the two groups (call it "success") and the remaining individuals fall into the other group (call it "failure"). Calling one of the categories "success" and the other "failure" is a convenience, not a value judgment.[1]

If we take a random sample of *n* individuals from this population, the sampling distribution for the number of individuals falling into the success category is described by the **binomial distribution**. The term "binomial" reveals its meaning: there are only two (*bi-*) possible outcomes, and both are named (*-nomial*) categories.

The ***binomial distribution*** provides the probability distribution for the number of "successes" in a fixed number of independent trials, when the probability of success is the same in each trial.

the proportion, the probability p is the parameter of the distribution

Formula for the binomial distribution

The binomial formula gives the probability of X successes in n trials, where the outcome of any single trial is either success or failure. The binomial distribution assumes that

- the number of trials (n) is fixed,
- separate trials are independent, and
- the probability of success (p) is the same in every trial.

Under these conditions, the probability of getting X successes in n trials is

Pr[X successes]=(nX)pX(1-p)n-X. $\Pr[X \text{ successes}] = \binom{n}{X} p^X (1-p)^{n-X}$.

The left side of this equation, Pr[X *successes*], means the "probability of X successes," where X is an integer between 0 and n. On the right-hand side, the quantity (nX)$\binom{n}{X}$ is read "n choose X." This represents the number of unique ordered sequences of successes and failures that yield exactly X successes in n trials.[2] The term is shorthand for

(nX)=n!X!(n-X)!' $\binom{n}{X} = \frac{n!}{X!(n-X)!}$

where $n!$ is called "n factorial" and refers to the product

n! = n × (n − 1) × (n − 2) × (n − 3) × … × 2 × 1. $n! = n \times (n-1) \times (n-2) \times (n-3) \times \ldots \times 2 \times 1$.

Similarly, $X!$ is "X factorial" and $(n - X)!$ is "$(n - X)$ factorial." By definition, 0! = 1, so (n0)$\binom{n}{0}$ and (nn)$\binom{n}{n}$ are both equal to 1. Factorials get very large very fast. For example, 5! = 120, but 20! = 2,432,902,008,176,640,000. Even with a reasonably small number of trials, calculating the binomial coefficient can require a good calculator or computer.[3]

Sampling distribution of the proportion

If there are X successes out of n trials in a random sample, then the estimated proportion of successes is $\hat{p}$:

$\hat{p} = \frac{X}{n}$.

(We pronounce $\hat{p}$ as "*p*-hat." Recall that p refers to the proportion in the population, whereas $\hat{p}$ refers to the *sample* proportion.)

We can use the same hypothetical population of flowers, having a true proportion of p = 0.25 successes, to illustrate the sampling distribution for the sample proportion $\hat{p}$. The panel on the top in Figure 7.1-2 is the sampling distribution when n = 10 (a relatively small sample size), whereas the panel on the bottom is the distribution for a larger sample size, n = 100. Both are based on binomial distributions, but rather than showing the number of successes X, we have divided X by n to yield $\hat{p}$

This is a case of built-in self averaging

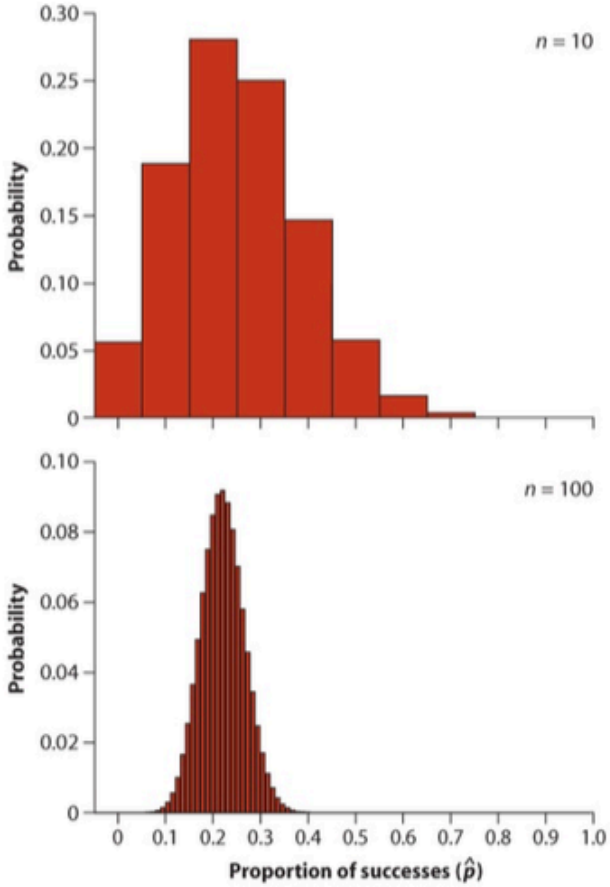

Figure 7.1-2
Whitlock et al., *The Analysis of Biological Data*, 2e, © 2015 W. H. Freeman and Company

FIGURE 7.1-2 The sampling distribution for the proportion of successes p^$\hat{p}$ for sample size n = 10 (*top*) and n = 100 (*bottom*). In both of these graphs, the population proportion is p = 0.25. The distribution is narrower (smaller standard deviation) when n is larger.

the effect of the sample size on the precision of the estimates

The mean of the sampling distribution of $\hat{p}$, the proportion of successes in a random sample of size *n*, is *p*. In other words, the proportion of successes in random samples is the same *on average* as the proportion of successes in the population. Therefore, $\hat{p}$ is an unbiased estimate of the population proportion—on average, it gives the right answer.

Notice in Figure 7.1-2 how the sample size affects the width of the sampling distribution for $\hat{p}$. When *n* is large (bottom panel), the sampling distribution is narrow. This effect is quantified in the formula for the standard error of $\hat{p}$ (Remember from Section 4.2 that the standard error of an estimate is the standard deviation of its sampling distribution.) The standard error of $\hat{p}$ (designated by σ_p) is

$$\sigma_p = \sqrt{\frac{p(1-p)}{n}}.$$

σp=p(1-p)n.

The sample size (*n*) is in the denominator of the standard error equation, so the standard error decreases as the sample size increases. That is why the estimates from samples of size 10 in Figure 7.1-2 (top panel) are more spread out than the estimates based on 100 individuals (bottom panel). Larger samples yield more precise estimates. The improvement in precision as sample size increases is called the **law of large numbers**.

Testing a proportion: the binomial test

The **binomial test** applies the binomial sampling distribution to hypothesis testing for a proportion. The types of questions it is suitable for have already been encountered in Chapter 6. The binomial test is used when a variable in a population has two possible states (i.e., "success" and "failure"), and we wish to test whether the relative frequency of successes in the population (p) matches a null expectation (p_0). The hypothesis statements look like this:

H_0: The relative frequency of successes in the population is p_0.

H_A: The relative frequency of successes in the population is not p_0.

The null expectation (p_0) can be any specific proportion between zero and one, inclusive.

The ***binomial test*** uses data to test whether a population proportion (p) matches a null expectation (p_0) for the proportion.

Example 7.2 shows how to apply the binomial test to real data.

EXAMPLE 7.2 Sex and the X

A study of 25 genes involved in spermatogenesis (sperm formation) found their locations in the mouse genome. The study was carried out to test a prediction of evolutionary theory that such genes should occur disproportionately often on the X chromosome.[4] As it turned out, 10 of the 25 spermatogenesis genes (40%) were on the X chromosome (Wang et al. 2001; see Figure 7.2-1). If genes for spermatogenesis occurred "randomly" throughout the genome, then we would expect only 6.1% of them to fall on the X chromosome, because the X chromosome contains 6.1% of the genes in the genome. Do the results support the hypothesis that spermatogenesis genes occur preferentially on the X chromosome?

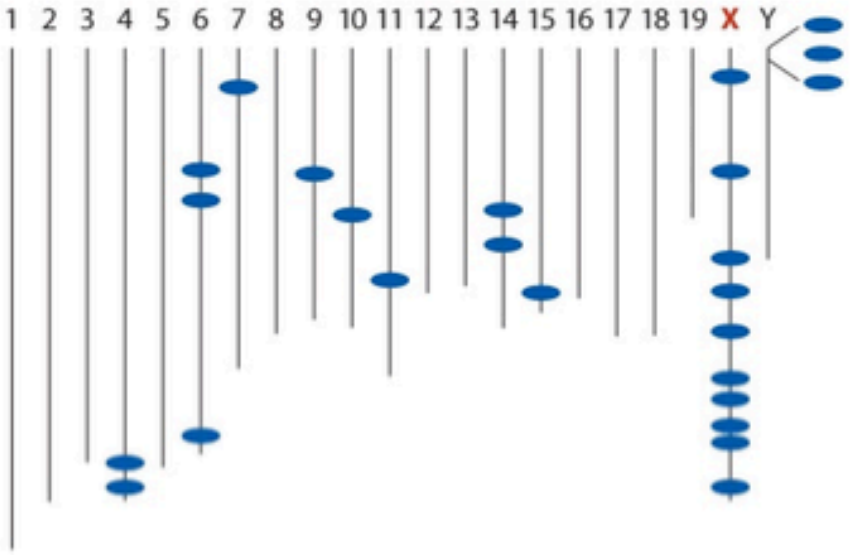

Figure 7.2-1
Whitlock et al., *The Analysis of Biological Data*, 2e, © 2015
W. H. Freeman and Company

FIGURE 7.2-1 Cartoon of the mouse genome. Each vertical line represents one of the mouse chromosomes and indicates its length relative to the others. Each mark on a line indicates a single gene involved in spermatogenesis. Note the abundance of

The null hypothesis is that the spermatogenesis genes would be on the X chromosome about 6.1% of the time, if they were randomly spread around the genome. To express this in terms of the binomial distribution, let's call the placement of each gene in the sample a "trial," and if the gene is on the X chromosome we'll call it a success. The null hypothesis is that the probability of success (p) is 0.061. The more interesting alternative hypothesis is that the probability of success (p) is *not* 0.061—that is, either spermatogenesis genes occur *more* frequently than 0.061 or they occur *less* frequently on the X chromosome than expected by chance.

We can write these hypotheses more formally as follows:

H_0: The probability that a spermatogenesis gene falls on the X chromosome is $p = 0.061$.

H_A: The probability that a spermatogenesis gene falls on the X chromosome is something other than 0.061 ($p \neq 0.061$).

Note once again the asymmetry of these two hypotheses. The null hypothesis is very specific, while the alternative hypothesis is not specific, referring to every other possibility. Also note that there are two ways to reject the null hypothesis: there can be an excess of spermatogenesis genes on the X chromosome (i.e., $p > 0.061$) or there can be too few (i.e., $p < 0.061$). Too few is not inconceivable, so it should also be included in the alternative hypothesis. Therefore, the test is two-sided.

the test statistic

The next step is to identify the test statistic that will be used to compare the observed result with the null expectation. In the case of the binomial test, the test statistic is the observed number of successes. For the data in Example 7.2, that would be 10 spermatogenesis genes on the X chromosome. The null expectation is $0.061 \times 25 = 1.525$. On average, we expect the fraction 0.061 of the 25 spermatogenesis genes sampled—namely, 1.525—to be located on the X chromosome if H_0 is true. Therefore, we know that in the *sample* more genes were found on the X chromosome than were expected by the null hypothesis.

The question now is whether we are likely to get such an excess by chance alone if the null hypothesis were true. To decide this we need the null distribution, the sampling distribution for the test statistic assuming that the null hypothesis is true. As mentioned previously, the sampling distribution for the number of successes X in a random sample of n individuals from a population having the proportion p of successes is described by the binomial distribution. Under the null hypothesis, the proportion is $p = 0.061$, so, for the above data (where $n = 25$ genes), the null distribution is given by

$$\Pr[X \text{ successes}] = \binom{25}{X} (0.061)^X (1 - 0.061)^{25-X}.$$

Pr[X successes]=(25X)(0.061)X(1 - 0.061)25-X.

p-value

This null distribution allows us to calculate the *P*-value, the probability of getting a result as extreme as, or more extreme than, 10 spermatogenesis genes on the X chromosome when the null expectation is 1.525. Because the test is two-tailed, *P* is the probability of getting 10 or more genes on the X chromosome plus the probability of similarly extreme results at the other tail of the null distribution, corresponding to too *few* genes on the X chromosome. We account for all the extreme outcomes by doubling the probability of getting 10 or more:

P=2Pr[number of successes ≥ 10].$P = 2\,\Pr\left[number\ of\ successes\ \geq\ 10\right].$

The probability of getting exactly 10 out of 25 on the X chromosome, when the probability of being on the X chromosome is 0.061, is

Pr [10 successes]=(2510)(0.061)10(1 -0.061)15 =9.07×10-7.

$$\Pr\left[10\ successes\right] = \binom{25}{10}(0.061)^{10}(1-0.061)^{15} = 9.07\times10^{-7}.$$

The probability of getting 10 *or more* spermatogenesis genes on the X chromosome, assuming the null hypothesis is true, is the sum over all of these mutually exclusive possibilities:

Pr [number of successes ≥10]=Pr [10] +Pr [11] +Pr [12]+ ⋯ +Pr [25]=9.9×10-7.

$$\begin{aligned}\Pr\left[number\ of\ successes\ \geq 10\right] &= \Pr\left[10\right] + \Pr\left[11\right] + \Pr\left[12\right] + \cdots + \Pr\left[25\right] \\ &= 9.9\times10^{-7}.\end{aligned}$$

The final *P*-value is

P=2Pr[number of successes≥10]=2(9.9×10−7)=1.98×10−6.

$$P = 2\,\Pr\left[number\ of\ successes\ \geq 10\right] = 2\left(9.9\times10^{-7}\right) = 1.98\times10^{-6}.$$

Conclusion

This *P*-value[5] is well below the conventional significance level of $\alpha = 0.05$. The probability of getting a result as extreme as, or more extreme, than the observed result is very low if the null hypothesis were true. Therefore, we reject the null hypothesis and conclude that there *is* a disproportionate number of spermatogenesis genes on the X chromosome. Our best estimate of the proportion of spermatogenesis genes that are located on the mouse X chromosome is p^=1025=0.40, $\hat{p} = \frac{10}{25} = 0.40,$
which is much greater than 0.061, the proportion stated in the null hypothesis. These results might be stated in a scientific report: "A disproportionately large proportion of spermatogenesis genes occur on the X chromosome (0.40, SE = 0.10; binomial test, $n = 25$, $P < 0.001$)." This statement includes the standard error of the proportion, which we show you how to calculate in the next section.

Problem 13 chap 7 from W&S

13. We all believe that we see most of what goes on around us, at least the most obvious things. Recently, however, psychologists have identified a phenomenon called "selective looking" which means that, if our attention is drawn to one aspect of what we see, we can miss even seemingly obvious features presented at the same time. In a striking demonstration of this phenomenon, a series of randomly chosen students was shown a video of six people throwing a basketball around, and they were asked to count how many times the people in white shirts threw the ball (Simons and Chabris 1999). In the middle of this video,[9] a woman dressed as a gorilla walked through the shot, pausing in the center to thump her chest, and then walked out of the shot. Look at the photo, and you will realize that nothing could be more obvious. Or was it? Of the 12 students watching the video, only five noticed the gorilla.

a. What is the best estimate from these data of the proportion of students in the population who notice the woman in the gorilla suit?

b. What is the 95% confidence interval for the proportion of students in the population who notice the woman in the gorilla suit?

c. What is the best estimate from these data of the proportion of students who *fail to notice* the woman in the gorilla suit?

© Photo courtesy of Daniel Simons [Simons and Chabris (1999)] "Figure provided by Daniel Simons, www .theinvisiblegorilla.com.