Binomial Sampling and the Binomial Distribution

Characterized by two mutually exclusive “events."

Examples:

GENERAL: {success or failure} {on or off} {head or tail} {zero or one}

BIOLOGY: {dead or alive} {captured or not captured} {reported or not reported}

These events are “outcomes" from a single “trial." **Binomial or Bernoulli trials.** For n trials one has y “successes." This is standard, general symbolism. Then y is an integer,

$$0 \leq y \leq n.$$

The binomial parameter, denoted p, is the *probability of success*; thus, the *probability of failure* is $1–p$ or often denoted as q. Denoting success or failure to p is arbitrary and makes no difference. Obviously, $p+q = 1$ because the events are assumed mutually exclusive and exhaustive (a coin must be a head or a tail and cannot remain resting on its edge!).

This is a good place to point out that these are only (approximating) *models*, and not full truth. Truth is not a model. A model, by definition, is an approximation (and not truth). In fact, out of a few million coin flips, one or more coins will remain standing on edge – that is truth. The binomial model does not allow such reality. The idea is more easily understood for the outcomes “dead" or “alive." In many cases, it is clear that an animal is either dead or alive; however, such a simple classification might be difficult in cases where an animal is “close to death." Even here, the binomial model might be a very good approximation to truth or full reality. We are, of course, interested in models that are very good approximations!

Of course, p is continuous and able to take any value between between 0 and 1 and including 0 and 1. $0 \leq p \leq 1$. It is likewise somewhat obvious that an *estimator* of the probability of success is merely

$$\hat{p} = y/n \quad = \text{ number of successes/number of trials.}$$

The estimator $\hat{p}$ is unbiased; some other useful quantities are:

$$\mathrm{E}(y) = np$$

Here is an example where the expectation is symbolized – we will employ this in many ways starting with lecture 4. Let $n = 100$ flips of a fair coin (thus $p = 0.5$). Then $\mathrm{E}(y) = 100 \cdot 0.5 =$

50. This was a case where the expectation of a statistic y was used. This procedure is common in modeling data.

$$\text{var}(y) = npq = np(1-p)$$

$$\text{var}(\hat{p}) = (pq)/n$$

$$\hat{\text{var}}(\hat{p}) = (\hat{p}\hat{q})/n$$

$$\hat{\text{se}}(\hat{p}) = \sqrt{(\hat{p}\hat{q})/n} = \sqrt{\hat{\text{var}}(\hat{p})}\,.$$

You should recognize these from your earlier classes in statistics. These results show that **statistics** such as y = number of successes also have expected values. The use of "hats" to denote estimators or estimates is important. One must no confuse as estimate (e.g., 119) with the parameter (e.g., 188, but usually not known to the investigator). The expression var($\hat{p}$) denotes the sampling variance; the uncertainty associated with the use of an estimator and **sample data**. Finally, do not think that the standard error of the estimator, $\hat{p}$ in this case, is anything other than the square root of the variance of the estimator $\hat{p}$. Note too, that

$$\hat{\text{var}}(\hat{p}) = (\hat{\text{se}}(\hat{p}))^2\,.$$

Beginning classes in statistics often use simplified notation to ease learning; however, occasionally this becomes a hindrance in more advanced work. For example, often the population mean μ is estimated by the sample mean and denoted as $\bar{x}$; this seems awkward and we will merely use $\hat{\mu}$. The estimate of the population variance σ^2 is often denoted by s^2 or S^2 and this seems particularly poor, at least for FW663. Consider the usual simple example from ST001 class.

> A random sample of n = 30 pigs are drawn and weighed exactly. The sample data are the pig weights w_1, w_2, w_3, ..., w_{30}. An estimate of the variability in pig weights across the pigs sampled is the usual standard deviation,
>
> $$\text{Estimated standard deviation} = \hat{\sigma} = \sqrt{\sum(w_i - \bar{w})^2 \Big/ n-1}.$$
>
> This is an estimate of the population standard deviation, σ. Because a random sample of the population was taken, the sample standard deviation can be taken as a valid measure of the variation in pig weights in the population. Note, if only pigs were weighed close to the road, then the standard deviation would be a measure of the variation in the sample of 30 pigs; without a valid way to make an inductive inference to the population of pigs.

Now, instead of wanting a measure of variation in the weights of pigs in a population (based on a sample), suppose one wanted to estimate some *parameter* in the pig population? The obvious parameter here is the mean weight μ (or, perhaps μ_w to denote it is the mean *weight*). It turns out that the estimator of this mean is

$$\hat{\mu} = \sum w_i \Big/ n$$

and this estimator is unbiased, asymptotically minimum variance and asymptotically normal (it is an "MLE" see below). Of course, we need a measure of its precision or repeatability; this is the estimated *sampling variance*,

$$\hat{\text{var}}(\hat{\mu}) = \sum (w_i - \bar{w})^2 \Big/ n(n-1)$$

or its estimated *standard error*,

$$\hat{\text{se}}(\hat{\mu}) = \sqrt{\hat{\text{var}}(\hat{\mu})} \, .$$

Whenever we need a measure of the precision of an *estimator*, we turn to the sampling variance or standard error (and perhaps coefficients of variation or confidence intervals). Note, these measures of precision or uncertainty are, themselves, only estimates (note the "hat" to indicate this).

If a measure of the variation in the population members is desired, one is interested in estimating the population standard deviation, σ. The standard deviation does not change with sample size; it is an innate value of the population. It has nothing to do with sampling, except that large sample might often permit a better estimate of this population parameter.

You might turn to a random page in a scientific journal in your field of interest and ask if these concepts about population variation σ^2 versus sampling variation $\text{var}(\hat{\theta})$ are made clear in tables!

Quantities such as the sampling variance are parameters and they have estimators. For example, in the case of the binomial model, the sampling variance is

$$\text{var}(\hat{p}) = p(1-p)/n$$

and its estimator is

$$\hat{\text{var}}(\hat{p}) = \hat{p}(1-\hat{p})/n \, .$$

This ∧ notation might seem irritating at first, but it becomes essential in real world problems. Why? Perhaps the *estimate* of the sampling variance is severely negatively biased in a particular situation; then one must worry about confidence interval coverage being less than the nominal level (e.g., 0.90). It is easy to blur the distinction between a parameter and its estimate. For example, consider the person who has just weighted 50 male and 50 female pigs and computed their sample means: $\hat{\mu}_m$ = 34Kg and $\hat{\mu}_f$ = 37Kg. Are these sample means "significantly" different, he asks? No *t* or *F* test is needed to answer this question; of course they are different! $34 \neq 37$. Perhaps he *really* meant to ask if the estimated sample means provided any evidence that the **population parameters** (μ_m and μ_f) differed by sex? Of course, he did.

The var($\hat{p}$) is really shorthand for var($\hat{p} \mid$ *model*), a measure of the sampling variation (uncertainty) of $\hat{p}$ as an estimator of true *p*, given a model. If sample size $\uparrow \infty$, var($\hat{p}$) goes to 0. If true *p* varies from coin to coin, var($\hat{p}$) tells us nothing at all about that other source of variation. Here, there are two variance components.

Consider a "population" of loaded coins, each with its own probability, *p*, of landing heads up and *p* varies over coins (i.e., "individual heterogeneity"). Then there is a population variance of *p*, namely σ_p^2.

Draw a sample of coins of size *k*, flip each one *n* times, get

$$\hat{p}_1, \hat{p}_2, \ldots, \hat{p}_k.$$

$\hat{\text{var}}(\hat{p}_i)$ tells us nothing about σ_p^2. The variation among the $\hat{p}_1, \hat{p}_2, \ldots, \hat{p}_k$ reflects both variation among the set of true *p*'s (σ_p^2) and the *k* sampling variances (each one conditional on *p* of that coin),

$$\text{var}(\hat{p}_i \mid \textit{model}), \quad i = 1, 2, \ldots, k.$$

Here, there is a sampling variance and a variance across individuals. We will see other examples of "variance components."

Things to ponder:

Do var($\hat{p}$) and σ^2 differ? Why?

How does one know which estimator to use? Are there general methods for finding such estimators?

What is the sampling variance of some parameter if a census is conducted?

Why is it called a *sampling variance*?

Review some old quizzes as a way to bring up other issues and consider them.

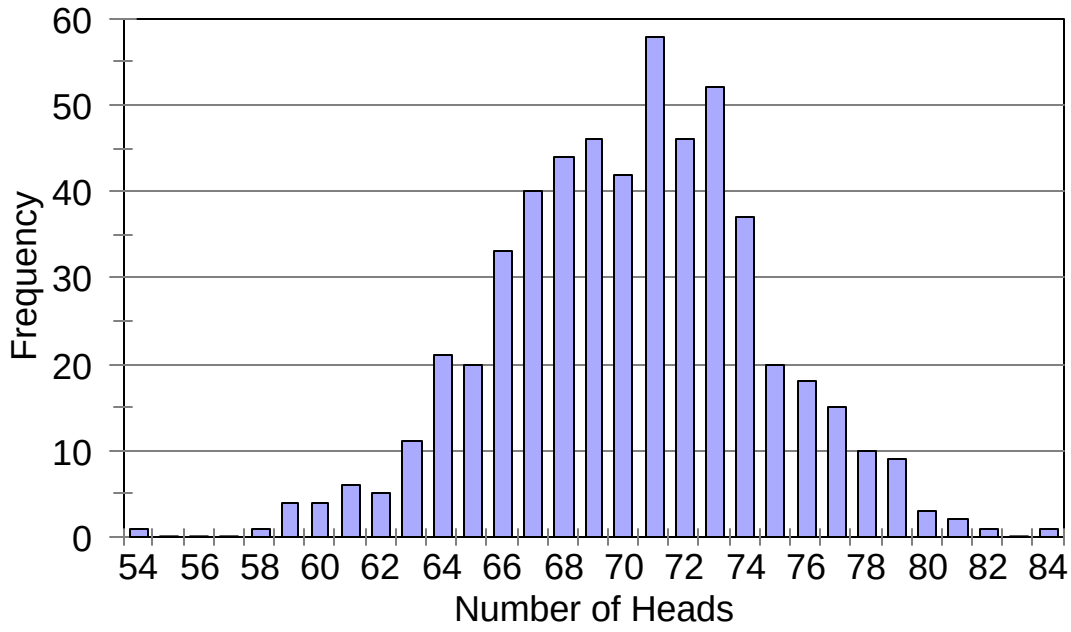
Summary – Binomial Random Variables:

1. n identical trials (e.g. a flip of 20 pennies)
2. each Bernoulli trial results in one of two mutually exclusive outcomes (e.g., head or tail)
3. the Prob{success} = p on each trial and this remains constant
4. Trials are independent events
5. y = number of successes; the *random* variable after n trials.
6. p is the probability of success
7. $\hat{p}$ is the *estimator* of the probability of success; $\hat{p} = y/n$.
8. $\hat{p} = 0.73$ is an *estimate.*
9. The precision of the estimator is measured by $\hat{\text{var}}(\hat{p}) = (\hat{p}\hat{q})/n$, the estimated sampling variance (or, the square root of $\hat{\text{var}}(\hat{p})$, the estimated standard error $\hat{\text{se}}(\hat{p})$).

Extended Example:

Consider n = 100 unfair pennies where the underlying binomial parameter is p = 0.70 (we know this value in this example). In a prior FW-663 class 11 students each made 50 surveys each involving flipping 100 unfair pennies. Thus, we have the results of 11x50=550 surveys, where each survey involved flipping 100 pennies. For each survey, we have n=100, y= the number of heads observed.

Clearly, $p = p(\text{H}) = 0.7$, $q = 0.3$, and $p+q=1$. The estimator is $\hat{p} = y/n$. The histogram below shows the frequency for the 550 independent surveys.



Note that y is a random variable and has a probability distribution (as above). It is actually a discrete random variable (y cannot be 37.54 heads). In fact, we will see that this binomial random variable is approximately normally distributed under certain conditions. Does the distribution of y above look somewhat "normal"?

$\bar{\hat{p}} = \hat{E}(\hat{p}) = 0.699$ and bias is *estimated* as $\hat{E}(\hat{p}) - p = 0.699 - 0.70 = -0.001$. This *estimate* of the bias suggests that bias is trivial; in fact, we know that the bias of this estimator is 0. Our *estimate* of bias is quite good in this case.

Note, percent relative bias (PRB) and relative bias (RB) are often useful:

$$\text{PRB} = 100 \cdot [\hat{E}(\hat{\theta}) - \theta] \Big/ \theta \qquad \text{and} \qquad \text{RB} = [\hat{E}(\hat{\theta}) - \theta] \Big/ \theta \, .$$

Binomial Probability Function

This function is of passing interest on our way to an understanding of likelihood and log-likehood functions. We will usually denote probability functions as *f* and, in this case, *f*(*y*) which is strictly positive and a function of the random variable *y*, the number of successes observed in *n* trials. We will return to a coin flipping survey where the outcomes are head (H) with probability *p* or tail (T) with probability 1–*p*. The binomial probability function is

$$f(y \mid n, p) = \binom{n}{y} p^y (1-p)^{n-y} \ .$$

The left hand side is read "the probability of observing *y*, given *n* flips with underlying parameter *p*. Thus, it is assumed that we know the exact values of *n* and *p*; only *y* is a random variable. How is this useful? Let us start with a coin presumed to be "fair" i.e., $p = 0.5$.

Let $n = 11$ trials or flips and *y* is the number of heads. The outcomes are below:

HHTHHHTTHHT; thus $y = 7$ = the number of heads (H).

The order is not important, thus the outcomes could have been written as

HHHHHHHTTTT

and, in terms of probabilities (which are multiplicative under the assumption of independence), we could write

$$p^y (1-p)^{n-y} \qquad \text{as} \qquad p^7 (1-p)^{11-7} \ .$$

Of course, *p* is assumed to be 0.5 if the coin is fair. Then, part of the probability function is written simply as

$$0.5^7 (1-0.5)^{11-7} \ .$$

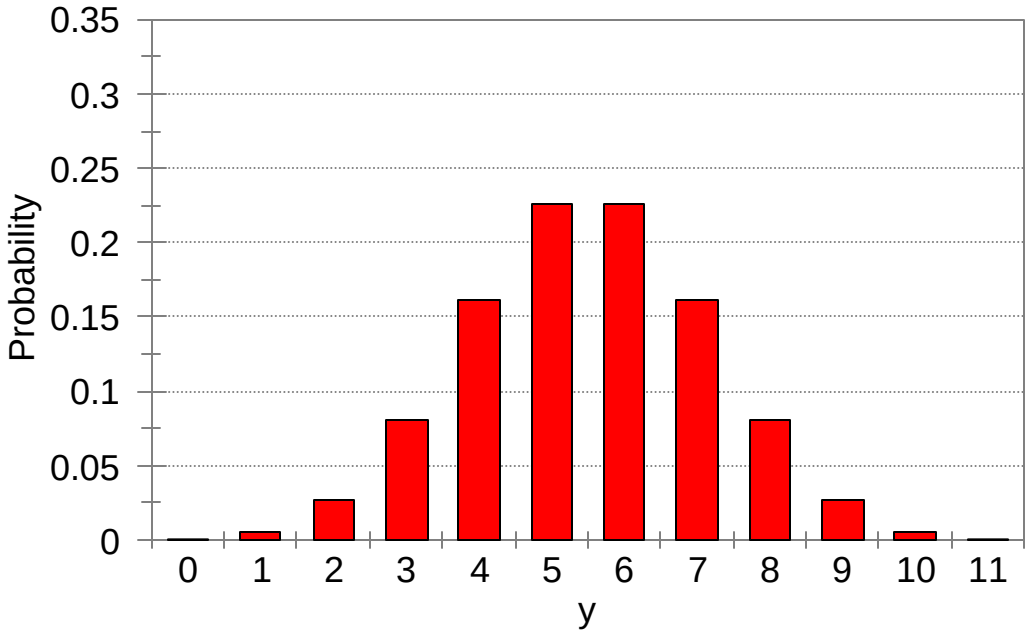
The full probability function, including the binomial coefficient, is then

$$f(y \mid 11, 0.5) = \binom{11}{y} 0.5^7 (1-0.5)^{11-7}$$

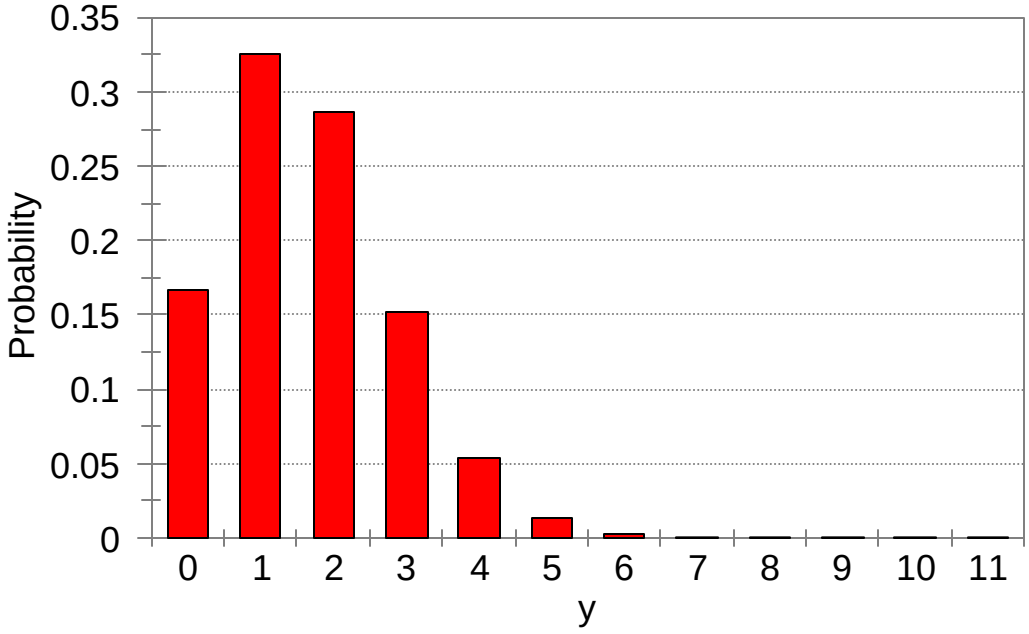
This function involves only one variable, $y = 7$. This can be evaluated numerically; $f(y \mid 11, 0.5) = (330) \times 0.5^{11} = 0.16$. So, in summary, the probability of getting 7 heads out of 11 flips of a fair coin is 0.16. Common sense leads one to believe that this is not an unusual event, the mathematics allows one to quantify the issue. Th ability to quantify probabilities is especially important when the model and data are much more complex (as we shall soon see). In such case, intuition is often of little help.

The theory of probability was well advanced 200 years ago so, by now, you can only imagine that many probability functions exist and are useful in a huge array of practical problems. However, probability functions are of little direct interest in biology because we rarely know the parameters. In fact, biologists have a reverse problem – they have data (the n and the y) but do not know the parameters (the p). This leads to the likelihood function!

Binomial distribution for $n = 11,\ p = 0.5$.



Binomial distribution for $n = 11,\ p = 0.15$.



Comparison of binomial distributions for $n = 11$, $p = 0.5$ (red) and $n = 11$, $p = 0.15$ (green).