Statistical Thinking for Forensic Practitioners

Quiz on Part 5: Probability Models and Uncertainty

1 Identifying probability models

Probability models allow us to treat seemingly disparate situations using the same mathematical tools. A handful of probability models are used so often that they are given names. The probability models discussed in the Part 5 lecture slides are examples of such probability models. Dr. Rice explained, in general terms, situations in which these models are most often applied. We will consider specific examples here. Statisticians often consider a series of simple questions to determine which model adequately describes a situation. In the following exercises, you will practice identifying which probability model discussed in the lecture notes best describes each situation.

1. A footwear impression analyst from Irvine, California is interested in determining the local prevalence of Carhartt boots. They randomly survey 10,000 people living in Irvine and record whether they own Carhartt boots.

   (a) Identify the quantity of interest from the sample that we can model using a probability distribution. (Hint: it may be helpful to start your answer with "the number of...").
   The number of people who own Carhartt boots.

   (b) Is the quantity of interest you identified above discrete or continuous?
   Discrete

   (c) What are the possible outcomes of one survey response?
   A respondent either owns Carhartt boots or does not own Carhartt boots.

   (d) What probability distribution should be used to model the quantity of interest?
   Binomial (sum of independent, binary random trials)

   (e) Interpret the unknown parameter of the probability distribution identified above within the context of the problem.
   The unknown parameter is the probability of success p (the number of trials n is known to be 10,000 in this problem). The parameter p represents the probability that a randomly selected person living in Irvine owns Carhartt boots.

   (f) Assume the quantity of interest you identified in part (a) is a random variable that follows the probability distribution you identified in part (d). Interpret the expectation of this random variable in the context of the problem.
   The expectation of a binomial random variable, X, with sample size $n = 10000$ and probability of success p (unknown in our problem) is $E(X) = n * p = 10000p$. This represents the expected number of people from the sample of size 10,000 to own Carhartt boots.

2. A researcher is interested in the prevalence of deltas in the latent prints from a population of interest. The researcher considers a sample of 122 latent prints from this population.

   (a) Identify the quantity of interest from the sample that we can model using a probability distribution. (Hint: it may be helpful to start your answer with "the number of...").
   The number of deltas in a latent fingerprint.

(b) Is the quantity of interest you identified above discrete or continuous?

Discrete

(c) What types of values can the quantity of interest you identified above take on?

Non-negative whole numbers (i.e., 0 deltas, 1 delta, 2 deltas, etc.)

(d) What probability distribution should be used to model the quantity of interest?

Poisson (which can be used to model count data within a spatial region such as a latent print). Note that the Poisson distribution assigns non-zero probability to all non-negative whole numbers (although the probability assigned decays with larger numbers). While this isn't physically realistic for this problem, we can nonetheless use the Poisson distribution to represent the fact that there isn't a physical law that limits the number of deltas that can appear on latent prints from this population of interest.

(e) Assume the quantity of interest you identified in part (a) is a random variable that follows the probability distribution you identified in part (d). Interpret the expectation of this random variable in the context of the problem.

For a random variable X that follows a Poisson distribution with parameter λ, the expectation is $E(X) = \lambda$. This represents expected number of deltas per fingerprint (λ is sometimes called the "rate" parameter.

3. Forensic examiners consider the frequency of an allele in a sample of 381 genotypes.

(a) Identify the quantity of interest from the sample that we can model using a probability distribution. (Hint: it may be helpful to start your answer with "the number of...").

The number of genotypes with the allele.

(b) Is the quantity of interest you identified above discrete or continuous?

Discrete

(c) What are the possible outcomes for a single genotype?

A genotype either has the allele or does not have the allele.

(d) What probability distribution should be used to model the quantity of interest?

Binomial (sum of independent, binary random trials)

(e) Interpret the unknown parameter of the probability distribution identified above within the context of the problem.

The unknown parameter is the probability of success p (the number of trials n is known to be 381 in this problem). The parameter p represents the probability that a randomly selected genotype as the allele.

(f) Assume the quantity of interest you identified in part (a) is a random variable that follows the probability distribution you identified in part (d). Interpret the expectation of this random variable in the context of the problem.

The expectation of a binomial random variable, X, with sample size $n = 381$ and probability of success p (unknown in our problem) is $E(X) = n * p = 381p$. This represents the expected number of genotypes from the sample of size 381 to have the allele.

4. Investigators receive a tip of an unknown number of contraband shipping containers on a barge of 2,000 containers. Due to the large number of containers, investigators decide to randomly select 20 containers without replacement to search. They intend to use the resulting number of contraband containers as an estimate for the number of contraband containers on the whole ship.

(a) Identify the quantity of interest from the sample that we can model using a probability distribution. (Hint: it may be helpful to start your answer with "the number of...").

The number of contraband shipping containers drawn from the sample of 20 shipping containers.

(b) Is the quantity of interest you identified above discrete or continuous?

Discrete

(c) What types of values can the quantity of interest you identified above take on?

Whole numbers between 0 and 20.

(d) What probability distribution should be used to model the quantity of interest?

Hypergeometric. The investigators are interested in the number of "successes" (i.e., contraband containers) out of a sample of $n = 20$ containers drawn without replacement from a population of $N = 2000$ containers.

(e) Only one of the parameters of the distribution you identified in part (d) is unknown. Interpret this unknown parameter within the context of the problem.

The unknown parameter is K, the number of contraband containers in the overall population of 2,000 containers.

5. Electron microscopy is used to digitally scan the surface of a cartridge case. The result is a "surface matrix" representation of the cartridge case containing surface height values. Practitioners visually compare variations in the height values (e.g., breech face impressions) to determine whether two cartridge cases were fired from the same firearm. Researchers are interested in modeling the height values of a set of 40 cartridge cases.

(a) Identify the quantity of interest from the sample that we can model using a probability distribution.

The height values of a cartridge case

(b) Is the quantity of interest you identified above discrete or continuous?

Continuous

(c) The researchers are unsure whether to use a normal model or a log-normal model. Consider the following histogram of height values for a cartridge case scan. Based on what you know about the normal and log-normal probability models, which would make the most sense to model the height values? Explain your reasoning.

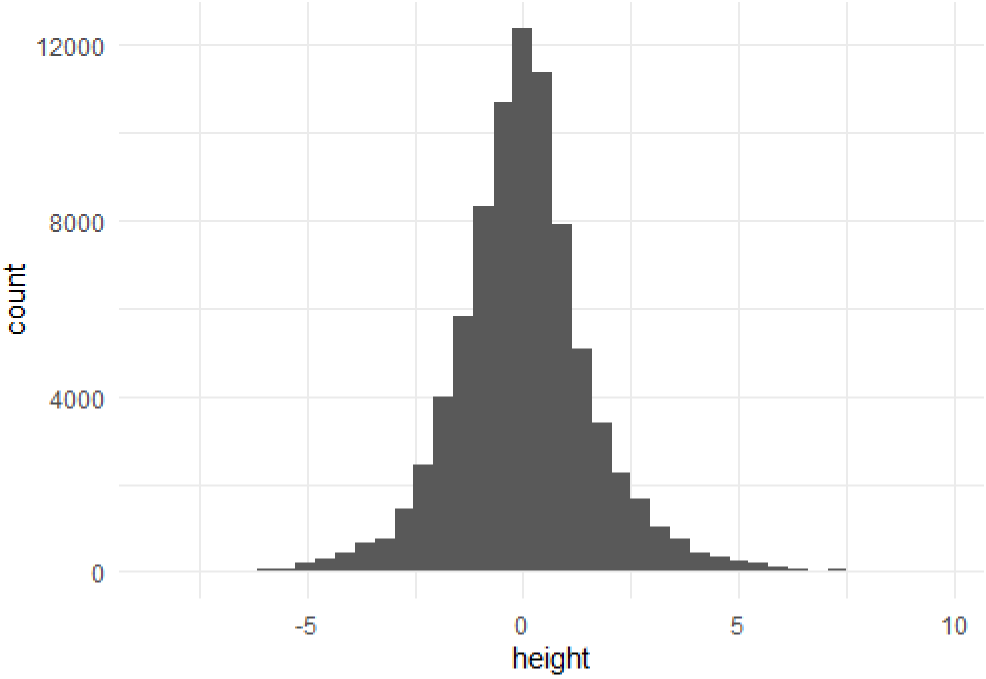

Figure 1: Histogram of height values of a cartridge case scan centered on the average height.

Normal. Note that the histogram of height values is approximately symmetric about 0, which is similar to the normal probability density function with mean 0. In contrast, the log-normal probability density function is right-skewed on the positive real line. Based on this, the normal model seems to be a more adequate representation of the height values.

(d) There are two unknown parameters for the distribution you chose in part (c). Interpret the parameters of the probability distribution identified above within the context of the problem.

The two unknown parameters are mu, the average height value, and σ^2, the variance of the height values. Alternatively, we could parameterize the normal using μ and σ, the standard deviation of the height values.

6. A longitudinal study (a study performed over time) was performed over 6 months to examine how shoes develop Randomly Acquired Characteristics (RACS). The researchers were interested in the number of RACs developed on 200 pairs of sneakers over the 6 month period.

(a) Identify the quantity of interest from the sample that we can model using a probability distribution. (Hint: it may be helpful to start your answer with "the number of...").

The number of Randomly Acquired Characteristics developed on sneakers over 6 months.

(b) Is the quantity of interest you identified above discrete or continuous?

Discrete

(c) What types of values can the quantity of interest you identified above take on?

Non-negative whole numbers (i.e., 0 RACS, 1 RAC, 2 RACS, etc.)

(d) What probability distribution should be used to model the quantity of interest?

Poisson (which can be used to model count data within a time frame such as over 6 months). Note that the Poisson distribution assigns non-zero probability to all non-negative whole numbers (although the probability assigned decays with larger numbers). While this isn't physically realistic for this problem, we can nonetheless use the Poisson distribution to represent the fact that there isn't a physical law that limits the number of RACs that can appear on a sneaker over 6 months.

(e) Assume the quantity of interest you identified in part (a) is a random variable that follows the probability distribution you identified in part (d). Interpret the expectation of this random variable in the context of the problem.

For a random variable X that follows a Poisson distribution with parameter λ, the expectation is $E(X) = \lambda$. This represents the expected number of RACs developed on a pair of sneakers per 6 month time period.