Workbook: Statistical Thinking for Forensic Practitioners

Sam Tyner, Hal Stern, Alicia Carriquiry, Michael Daniels

Contents

Chapter 1

Introduction

This workbook is intended to accompany the Statistical Thinking for Forensic Practitioners workshop taught by members of the Center for Statistics and Applications in Forensic Evidence (CSAFE). The slides for this workshop were constructed by Hal Stern, Alicia Carriquiry, and Michael Daniels.

When taking the workshop, please follow along with the slides handout (if given) and this workbook. The workbook contains the same material as the slides, with room for you to take notes and to fill in the missing material.

Chapter 2

Statistical Preliminaries

Briefly, this section contains a broad review of probability concepts and of statistical inference concepts, with examples from the forensic science context. We will cover probability, data collection, statistical distributions, estimation, and hypothesis testing.

2.0.1 Definitions

- **population**: ____

- **sample**: ____

- **probability**: Using knowledge about the ____________________ to make statements describing the ______________________. Probability can loosely be thought of as a type of deductive reasoning, where we are applying general knowledge about the population of interest to make conclusions about a small part of that population.

- **statistics**: Using knowledge about the ____________________ to make statements describing the ______________________. Statistics can loosely be thought of as a type of inductive reasoning, where we are applying knowledge about a sample to state that something *may* be true about the population generally.

2.0.2 Forensic Science Examples

- Suppose 100 1-pound bags of heroin are seized on the US-Mexico border, and the FBI want to know the chemical composition of the confiscated drugs to store in their database.
  - Population: ______________________________
  - Sample: ______________________________

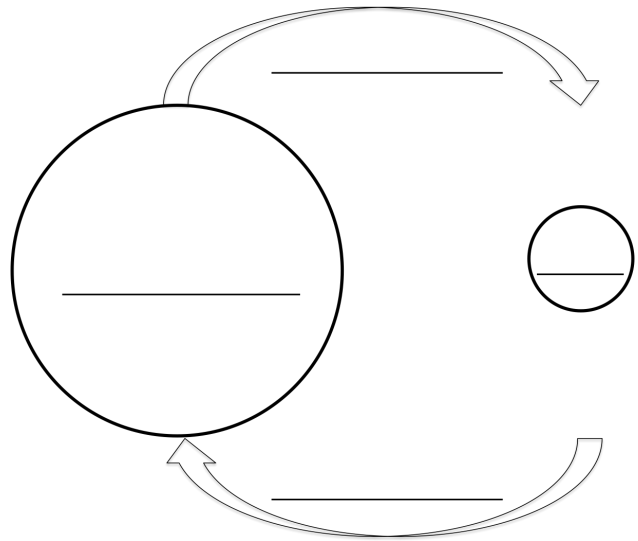

Figure 2.1: "The Big Picture"

- A window was broken in a robbery, and the suspect who was apprehended nearby had glass fragments lodged in the soles of their shoes. Do the fragments from the suspect's shoes have the same or similar chemical composition as the broken window?
  - Population 1: ______________________________
  - Sample 1: ______________________________

  - Population 2: ______________________________
  - Sample 2: ______________________________

- A city government employee is suspected of embezzling funds from the city's coffers. Forensic accountants examine a subset of the city's transactions to determine whether embezzling occurred and how much money was lost.
  - Population: ______________________________
  - Sample: ______________________________

How do you think this pertains to pattern evidence? List some possible relevant populations and samples below.

- Population 1: ______________________________
- Sample 1: ______________________________

- Population 2: ______________________________

- Sample 2: ____
- Population 3: ____
- Sample 3: ____

2.1 Probability

Probability concerns the *uncertainty* of outcomes. The set of all possible outcomes is called the _________________ space, and a particular outcome or set of outcomes of interest is referred to as an _______________.

2.1.1 Examples

1. Footwear
   - Sample Space = All shoe sizes e.g. $\{6, 6.5, 7, 7.5, 8, 8.5, \ldots\}$
   - Event = Shoe of size 9
2. Footwear
   - Sample Space = Brand of shoe e.g. { Nike, Vans, Converse, ...}
   - Event = Nike sneaker
3. Firearms
   - Sample Space = CMS (consecutive matching striae) for a pair of bullets e.g. $\{0, 1, 2, 3, 4, \ldots\}$
   - Event = CMS of 10 or more

2.1.2 Interpretation

The probability of observing an event in a sample space is a number less than or equal to 1 and greater than or equal to 0 that describes the __________________ that the event will occur.

There are two primary interpretations of probability:

1. The long run __________________ of occurrence of an event.
2. The ____________________ belief of likelihood of an event occurring.

2.1.3 Basic Notation and Laws of Probability

Let an event of interest be denoted by ___________. The probability of this event occurring is then denoted _______________. Recall that the probability of an event is always between 0 and 1. When $P(Y) = 0$, the event Y will never happen. When $P(Y) = 1$, the event Y will always happen. The sum of the probabilities of all possbile outcomes in the sample space always equal to ________.

The event of interest, Y, also has a complement event, $\overline{Y}$, which is read as "not Y". The complement, $\overline{Y}$, of an event, Y, is itself an event containing all outcomes in the sample space other than that initial event of interest, Y.

$$P(Y) + P(\overline{Y}) = \underline{\hspace{2em}}$$

The above equation also gives us the following rules:

$$\begin{aligned} P(Y) &= 1 - P(\overline{Y}) \\ P(\overline{Y}) &= 1 - P(Y) \end{aligned} \tag{2.1}$$

2.1.4 Probability and Odds

The probability of an event defines the odds of the event. The odds *in favor* of an event Y are defined as the probability of Y divided by the probability of everything except Y ("not Y"):

$$O(Y) = \frac{P(Y)}{P(\overline{Y})} = \frac{P(Y)}{1 - \underline{\hspace{1.5em}}}.$$

Conversely, the odds *against* a event Y are defined as the probability of everything except Y ("not Y") divided by the probability of Y:

$$O(\overline{Y}) = \frac{P(\overline{Y})}{P(Y)} = \frac{1 - \underline{\hspace{1.5em}}}{P(Y)}.$$

When we typically talk about odds, like in horse racing, the odds reported are the odds *against* the outcome of interest. Let's construct a horse race scenario using our probability notation to find the probability of a horse winning a race from the reported odds:

- Suppose you want to place a bet on a horse name Cleopatra winning the race. Odds for Cleopatra are reported as 4:1.
- Y = Cleopatra wins the race
- $\overline{Y}$ = Any horse in the race *other than* Cleopatra wins the race.
- $O(\overline{Y}) = \dfrac{P(\overline{Y})}{P(Y)} = \frac{4}{1} = 4$
- We know that $P(Y) + P(\overline{Y}) = 1$. With this information, we can determine $P(Y)$, which is the probability that Cleopatra wins the race:

$$
\begin{aligned}
O(\overline{Y}) &= \frac{P(\overline{Y})}{P(Y)} = 4 \\
\Rightarrow \frac{P(\overline{Y})}{P(Y)} &= 4 \\
\Rightarrow \frac{1-P(Y)}{P(Y)} &= 4 \qquad \textit{(See Equation 2.1)} \\
\Rightarrow \frac{1}{P(Y)} - 1 &= 4 \\
\Rightarrow \frac{1}{P(Y)} &= 5 \\
\Rightarrow P(Y) &= \frac{1}{5} = 0.2 \\
\Rightarrow P(\overline{Y}) &= 0.8
\end{aligned}
$$

- So, the odds for Cleopatra (4:1) mean that Cleopatra has a probability of 0.2 of winning the race. Because this outcome is not very likely (it will only happen in 1 race out of 5), you win money if Cleopatra wins simply because that is not a likely outcome.
- **Betting**: Suppose you bet \$1 on Cleopatra to win the race with 4:1 odds. You will win \$4 if Cleopatra wins, otherwise you've lost \$1.
- The amount you win (\$4) is determined so that you break even in the long run.
- Suppose 5 identical races are run. In 1 of those races, Cleopatra will win, and in the other 4, Cleopatra will lose. If you bet \$1 on Cleopatra in each race, you will lose that \$1 4 of 5 times. So, in order for you to break even, the designated amount you'll win when Cleopatra wins is \$4.
- This is a statistical concept known as *expected value*. Your expected value when placing the bet is \$0. We compute expected value by multiplying each possible outcome value by its probability and adding them all together:

$$
\begin{aligned}
\$4 \cdot P(Y) + (-\$1) \cdot P(\overline{Y}) &= 0 \\
\$4 \cdot 0.2 + (-\$1) \cdot 0.8 &= 0 \\
\$0.8 - \$0.8 &= 0
\end{aligned}
$$

2.1.5 Probability Math

Up until now, we have only considered one event, Y. Now, suppose we have another event that we are interested in, Z.

Let's consider the possibility of *either* of these two events, Y or Z, occurring. We'd write this as $Y \cup Z$, which is mathematical notation for "Y or Z occurs". There are two scenarios that arise:

1. Y and Z cannot occur together: they are ______________ ________________
2. Y and Z can occur together.

In scenario #1, computing the probability of either Y or Z happening is easy: we just add their respective probabilities together:

$$Y, Z \text{ mutually exclusive } \Rightarrow P(Y \cup Z) = P(Y) + P(Z)$$

In scenario #2, computing the probability of either Y or Z happening is more complicated because we know there is a chance that Y and Z can happen together. We'd write this as $Y \cap Z$, which is mathematical notation for "Y and Z occurs". In scenario #1, this event never occurred, so $P(Y \cap Z) = 0$ there. To compute the probability of Y or Z occurring in scenario #2, we have to consider the probability of Y, the probability of Z, and the probability of $Y \cap Z$. If we just add $P(Y) + P(Z)$ as in scenario #1, we include the event $Y \cap Z$ twice, so we have to subtract one instance of it:

$$Y, Z \text{ not mutually exclusive } \Rightarrow P(Y \cup Z) = P(Y) + P(Z) - P(Y \cap Z).$$

This probability is much easier to think about when illustrated. In Figure 2.2, we consider human blood types. There are four groups: A, B, O, and AB, and there are two RH types: + and −. We first consider the blood types A and B, represented by the two non-overlapping circles. Define:

- Event Y = a person has blood type A
- Event Z = a person has blood type B
- Event $Y \cup Z$ = a person has blood type A or blood type B

These two events are *mutually exclusive* because one person cannot have both blood type A and blood type B. (The circles don't overlap in the venn diagram) So, the probability that a randomly selected person has blood type A or B is:

$$P(Y \cup Z) = \_\_\_\_ \quad + \quad \_\_\_\_$$

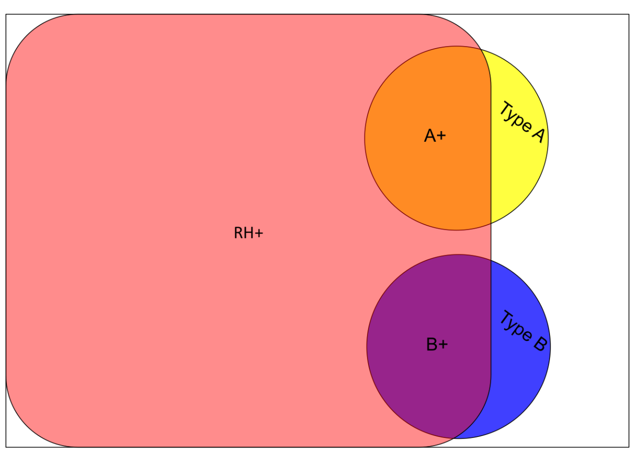

Figure 2.2: Probabilities of blood types in humans. Areas are approximate.

Return to Figure 2.2 and consider two other events: a person having blood type A or having the Rh factor (RH+). We see in Figure 2.2 that someone can have both type A blood and the Rh factor (blood type A+). Define:

- Event Y = a person has blood type A
- Event Z = a person has the Rh factor
- Event $Y \cup Z$ = a person has blood type A or the Rh factor
- Event $Y \cap Z$ = a person has blood type A and the Rh factor (they have A+ blood)

So, the probabilty that someone has either type A blood or has the Rh factor is the sum of probability of having type A blood (represented by the yellow circle) and the probability of having the Rh factor (represented by the red rectangle) minus the probability of having A+ blood (represented by the orange area of overlap that is counted twice) in Figure 2.2. So, the probability that a randomly selected person has blood type A or the Rh factor is:

$$P(Y \cup Z) = \_\_\_\_ \quad + \quad \_\_\_\_ \quad - \quad \_\_\_\_$$

2.1.6 Conditional Probability

Let's consider an event of interest Y which has probability $P(Y)$. Then, suppose we learn of another event of interest Z that has occurred. Knowing that Z has occurred already may change our opinion about the likelihood of _______ occurring. The key idea here is that the probability of an event often depends on other information, leading us to the definition of *conditional probability*:

$$P(Y|Z),$$

which is the conditional ____________________ that Y occurs given that we know Z has occurred. Return to Figure 2.2. Suppose we want to know the probability of a person having type A blood, represented by the yellow circle. But, if we already know that a person has the Rh factor, we are only interested in the part of the type A circle that overlaps with the Rh+ rectangle. Thus the probability of having type A blood is different with different knowledge. The formula for calculating conditional probability is:

$$P(Y|Z) = \frac{P(Y \cap Z)}{P(Z)} \tag{2.2}$$

Returning to the venn diagram, the value $P(Y \cap Z)$ is represented by the overlap of the type A circle and the Rh+ rectangle, and the value $P(Z)$ is represented by the Rh+ rectangle. Then, the value $P(Y|Z)$ is the ratio of the overlap (A+) to the Rh+ rectangle.

Equation 2.2 also gives us a multiplication rule for computing probabilities:

$$P(Y \cap Z) = P(Y|Z) \cdot P(Z) \tag{2.3}$$

RV	DP	NDP	Total
W	45	85	130
B	14	218	232
Total	59	303	362

Table 2.1: The results of the Baldus et al study for black defendants convicted of murder.

Philosophically speaking, it can be helpful to think of *all* probabilities as conditional. It is just a question of what information is assumed to be ________________.

2.1.6.1 Examples

Death Penalty Convictions

A study of sentencing of 362 black people convicted of murder in Georgia in the 1980s found that 59 were sentenced to death (Baldus, Pulaski, and Woodworth (1983)). They also examined the race of the murder victim, either black or white, and found some disparities. In Table 2.1, DP means the defendant received the death penalty, NDP means the defendant did not receive the death penalty. The race of the victim (RV) is either black (B) or white (W).

Returning to Section 2.0.1, let's define the problem:

- **Population**: All black people convicted of murder in Georgia in the 1980s
- **Sample**: N/A (the whole population was studied)

Using the numbers from Table 2.1, compute the following probabilities:

- $P(DP) = \text{—} = 0.$_____
- $P(DP|RV = W) = \text{—} = 0.$_____
- $P(DP|RV = B) = \text{—} = 0.$_____

Note: These numbers are selected from the study, and should not be considered a comprehensive summary of its results. There are a number of things not discussed here. The entire publication can be found online[1]

Consecutive Matching Striae

In firearms and toolmark analysis, the number of consecutive matching striae (CMS) between a crime scene sample and a lab sample is often used to help determine a match. Generally speaking, the higher the maximum number of CMS found in a pair, the more likely the two samples came from the same source. Several known match (KM) pairs and known non-match (KNM) pairs of bullets were examined, and the results are shown in Figure 2.3 (Hare, Hofmann, and Carriquiry (2017)). What is the probability of seeing two known matches (or two known non-matches) given the maximum number of CMS? Here, we condition on ____. Again, we briefly return to Section 2.0.1, let's define the problem:

[1] http://scholarlycommons.law.northwestern.edu/cgi/viewcontent.cgi?article=6378&context=jclc.

- **Population**: All pairs of fired bullets from unknown sources
- **Sample**: A sample of pairs of known matches and known non-matches

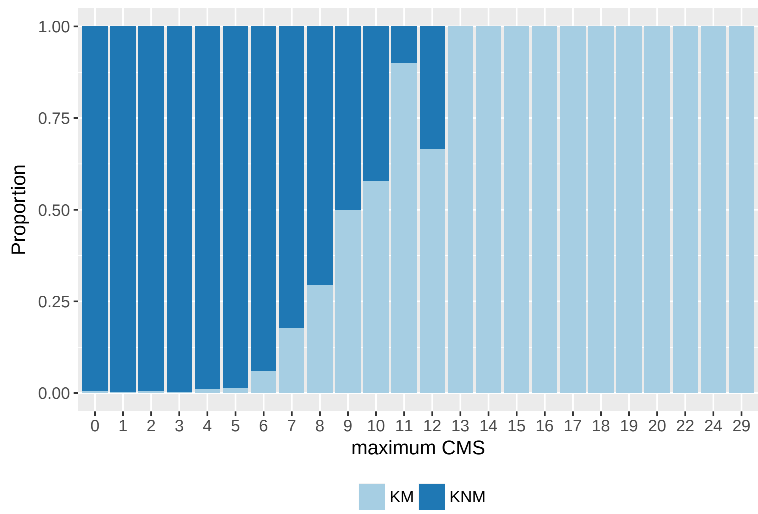

Figure 2.3: This bar chart represents the conditional probabilities of two bullets matching given the maximum number of CMS. The light blue represents known matches, while the dark blue represents known non-matches.

Generally, as seen in Figure 2.3, the probability of finding a match tends to increase with then number of maximum CMS. For ________ maximum CMS values is it much more likely that we have a ____________________________ pair.

2.1.7 Independence

If the likelihood of one event is *not* affected by knowing whether a second has occured, then the two events are said to be ____________________________. For example, the region of the country where you live and what color car you drive are (probably) not related.

The death penalty example from the previous section demonstrates that defendants receiving the death penalty is *not* independent of the race of the victim. In other words, a black defendant found guilty of murder in Georgia in the 1980s received a different penalty depending on the race of the victim.

Another example from DNA analysis relies on on independence across chromosomes. By using loci on different chromosomes, there is independence between the allele counts, allowing for simple calculation of random match probabilities.

2.1.8 Probability Math... Again

Recall Equation 2.3, which gives us the probability of two events, Y and Z occurring together:

$$P(Y \cap Z) = P(Z) \cdot P(Y|Z) = P(Y) \cdot P(Z|Y)$$

If Y and Z are *independent*, there is a simple formula:

$$P(Y \cap Z) = \underline{\hspace{2cm}} \cdot \underline{\hspace{2cm}}$$

This is because Z occurring does not effect the probability of Y occurring, and vice versa. Thus,

$$P(Y|Z) = P(Y) \quad \text{and} \quad P(Z|Y) = P(Z)$$

For example, the probability of being left-handed and from Florida is equal to the probability of being left-handed times the probability of being from Florida, assuming the events "being left-handed" and "being from Florida" are independent.

Multiplying probabilities of events directly like this is *only* applicable when the events are independent. When *dependent* events are treated as independent events, things can go terribly wrong. An infamous example of this in the courts is the case *People v. Collins*[2]. This was a robbery trial, where eyewitnesses described the robbers a "black male with a beard and a moustache, and a white female with a blonde ponytail, fleeing in a yellow car".

The prosecution provided estimated probabilities of each of these individual characteristic:

- P(black man with a beard) = ______________
- P(black man with a moustache) = ______________
- P(white woman with ponytail) = ______________
- P(white woman with blonde hair) = ______________
- P(yellow car) = ______________
- P(interratial couple in a car) = ______________

A mathematics "expert" talked about the so-called "multiplication rule for probability", and directly multiplied the above probabilities together without considering that the events could be *dependent*. i.e. a man with a beard probably has a much higher chance of having a moustache than a man with no beard. Due to this faulty math, the conviction was set aside and the statistical reasoning criticized for ignoring dependence among the characteristics.

In a courtroom situation, let S be the event that the suspect was present at the scene of the crime and $\overline{S}$ be the event that the suspect was not present at the scene. Assume that each juror has in mind an initial probability for the events S and $\overline{S}$. Then, a witness says they saw a tall Caucasian male running from the scene, and the defendant is a tall Caucasian male. After hearing the witness' testimony, the jurors ___________ their probabilities. Next, an expert witness testifies that fragments from a window broken during the crime and fragments found on the defendant's clothing match. Again, the jurors update their ______________. This process continues throughout the trial. There are some key questions to consider:

- How should jurors update their probabilities?
- Do jurors *actually* think this way?

[2] *People v. Collins*, 68 Cal.2d 319, 438 P.2d 33 (1968)

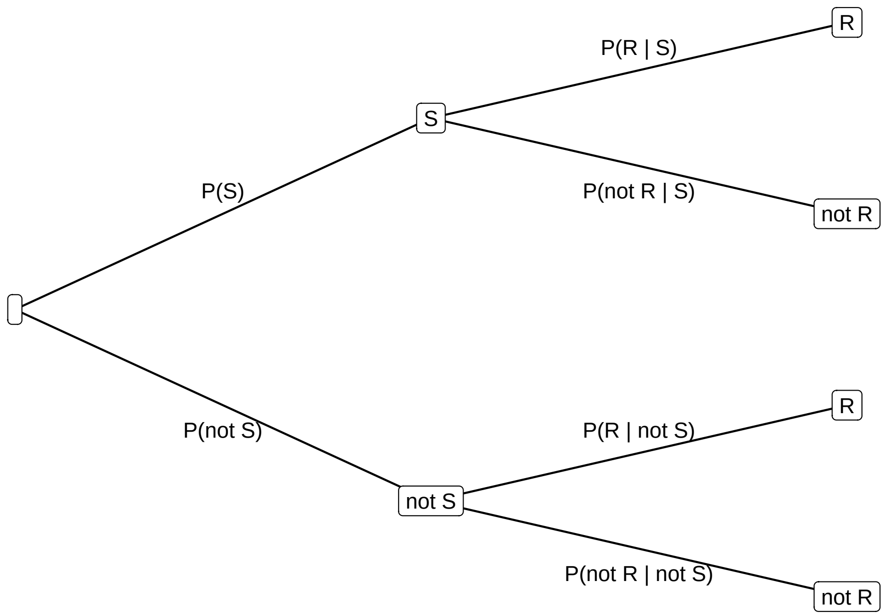

Figure 2.4: A probability tree showing the direction of flow when updating probabilities. Move from left to right on the tree through the events possible. Events are in boxes, probabilities are on the branches of the tree.

2.1.9 Bayes' Rule

Bayes' Rule provides an ______________ formula for probabilities. Like in the trial scenario above, suppose we have an initial estimate for the probability of event S, $P(S)$. Then, we learn that an event R has occurred and we want to update or probability of event S. To do this, we need to know about the __________________ of R and S. To update the probability of S, we can use Bayes' Rule, also called Bayes' _________________:

$$\begin{aligned} P(S|R) &= \frac{P(R \cap S)}{P(R)} = \frac{P(R|S)P(S)}{P(R)} \\ &= \frac{P(R|S)P(S)}{P(R|S)P(S) + P(R|\overline{S})P(\overline{S})} \end{aligned} \tag{2.4}$$

2.1.9.1 Examples

Consider performing diagnostic tests for gunshot residue.

- Let G denote the presence of gunshot residue
- Let $\overline{G}$ denote the __________________ of gunshot residue
- Let T denote a ____________________ diagnostic test
- Let $\overline{T}$ denote a negative diagnostic test

The values in the table can also be thought of as conditional probabilities:

Truth	T	$\overline{T}$
G	True Positive	False Negative
$\overline{G}$	False Positive	True Negative

Table 2.2: All potential outcomes of a diagnostic test for gunshot residue.

- The value $P(T|G)$ is the ____________________________ rate, also called *sensitivity* of the test
- The value $P(\overline{T}|\overline{G})$ is the _________________________ rate, also called the *specificity* of the test
- The value $P(T|\overline{G})$ is the ___________________________ rate, the Type I error rate
- The value $P(\overline{T}|G)$ is the ______________________________ rate, the Type II error rate

Studies of the diagnostics test usually tell us $P(T|G)$, ______________, and $P(\overline{T}|\overline{G})$, _____________ . Examiners may begin with some idea of $P(G)$, or the ______________ of gunshot residue in a similar situation. What is most relevent for the case is the *postitive predictive value*, or in probability notation, ______________. We can use _________________________________ to obtain this value:

$$P(G|T) = \frac{P(T|G)P(G)}{P(T|G)P(G) + P(T|\overline{G})P(\overline{G})}$$

Generally speaking, the most important thing to remember is that, in general, $P(T|G)$ _____ $P(G|T)$.

The careful application of Bayes' Rule can sometimes lead to surprising, non-intuitive results. Continuing with the gunshot residue test example, assume

- sensitivity is 98% ($P(\quad|\quad) = 0.98$)
- specificity is 96% ($P(\quad|\quad) = 0.96$)
- prevalence is 90% ($P(\quad) = 0.90$)
- Plug values into the Bayes' Rule formula to find $P(G|T)$:

$$\begin{aligned}
P(G|T) &= \frac{P(T|G)P(G)}{P(T|G)P(G) + P(T|\overline{G})P(\overline{G})} \\
&= \frac{0.98 \cdot 0.9}{0.98 \cdot 0.9 + (1 - 0.96) \cdot (1 - 0.9)} \\
&= \frac{0.882}{0.882 + 0.004} \\
&= 0.995
\end{aligned} \tag{2.5}$$

- Now assume prevalence is 10% ($P(\quad) = 0.10$) and plug in the values again

$$
\begin{aligned}
P(G|T) &= \frac{P(T|G)P(G)}{P(T|G)P(G) + P(T|\overline{G})P(\overline{G})} \\
&= \frac{0.98 \cdot 0.1}{0.98 \cdot 0.1 + (1 - 0.96) \cdot (1 - 0.1)} \\
&= \frac{0.098}{0.098 + 0.036} \\
&= \frac{0.098}{0.134} \\
&= 0.731
\end{aligned}
\tag{2.6}
$$

- So, even if there is a postive test, we are not really sure about whether gunshot residue is *actually* present.
- Why does this happen?? See Figure 2.5.

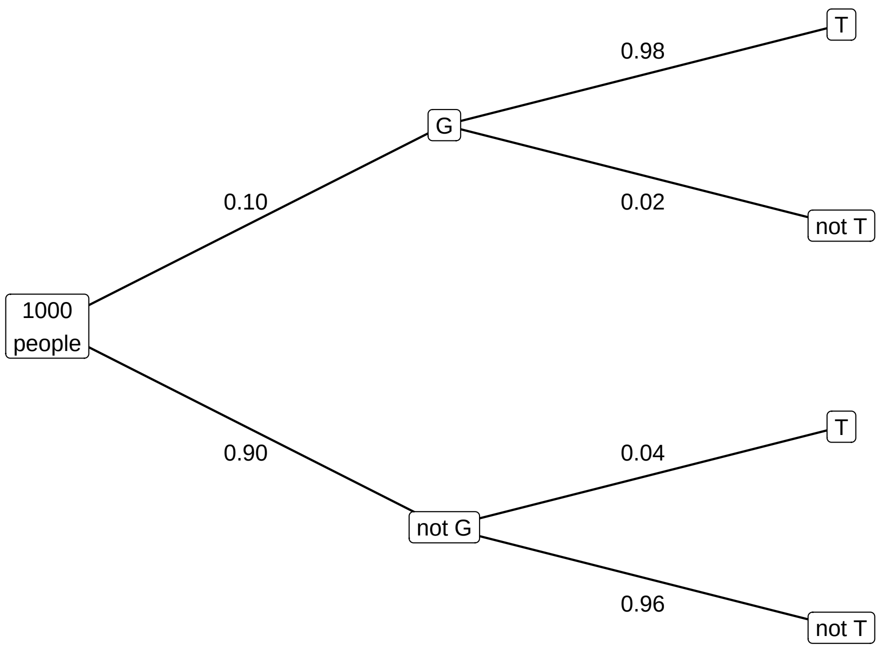

Figure 2.5: A probability tree showing the direction of flow when updating probabilities for the presence of gunshot residue. Suppose there are 1,000 people in the population you're considering. Write the number of people in the groups throughout the tree according to the probabilities indicated on the branches of the tree

2.1.10 Bayes' Rule to the Likelihood Ratio

In the general forensic setting, let S denote the event that the evidence from the scene and comparison sample are from the same source. Let E denote the evidence found at the scene. The formulation of Bayes' Rule for this situation is:

$$
P(S|E) = \frac{P(E|S)P(S)}{P(E|S)P(S) + P(E|\overline{S})P(\overline{S})}
$$

We can rewrite Bayes' Rule in terms of odds:

$$\frac{P(S|E)}{P(\overline{S}|E)} = \frac{P(E|S)}{P(E|\overline{S})}\frac{P(S)}{P(\overline{S})} \tag{2.7}$$

Derivation of Equation 2.7 is shown in Equation 2.8. For now, just consider Equation 2.7:

- On the left, $\frac{P(S|E)}{P(\overline{S}|E)}$ are the odds in favor of S given the evidence E.
- The last term on the right, $\frac{P(S)}{P(\overline{S})}$ are the odds in favor of S before seeing the evidence E (the "prior odds")
- The first term on the right $\frac{P(E|S)}{P(E|\overline{S})}$, is known as the ____________________ ratio
- The likelihood ratio (LR) is the factor by which we ________________ prior odds of two samples being from the same source to get ______________________ odds (after seeing evidence) of the same source.

$$\begin{aligned}
P(S|E) &= \frac{P(E|S)P(S)}{P(E|S)P(S) + P(E|\overline{S})P(\overline{S})} \\
\Rightarrow \frac{1}{P(S|E)} &= \frac{P(E|S)P(S) + P(E|\overline{S})P(\overline{S})}{P(E|S)P(S)} \\
&= 1 + \frac{P(E|\overline{S})P(\overline{S})}{P(E|S)P(S)} \\
\Rightarrow \frac{1}{P(S|E)} - 1 &= \frac{P(E|\overline{S})P(\overline{S})}{P(E|S)P(S)} \\
\frac{1}{P(S|E)} - \frac{P(S|E)}{P(S|E)} &= \\
\frac{1 - P(S|E)}{P(S|E)} &= \\
\frac{P(\overline{S}|E)}{P(S|E)} &= \frac{P(E|\overline{S})P(\overline{S})}{P(E|S)P(S)} \\
\Rightarrow \frac{P(S|E)}{P(\overline{S}|E)} &= \frac{P(E|S)P(S)}{P(E|\overline{S})P(\overline{S})}
\end{aligned} \tag{2.8}$$

2.1.10.1 Examples

Return to the gunshot residue (GSR) test example. Define:

- E = evidence = a positive test for (GSR)
- S = suspect has GSR on them

$$LR = \frac{P(E|S)}{P(E|\overline{S})} = \frac{0.98}{0.04} = 24.5$$

In a high prevalence case ($P(G) = 0.9$), the prior odds are $\frac{0.9}{0.1} = 9$. The posterior odds are $LR \times$prior odds $= 24.5 \times 9 = 220.5 : 1$.

In a low prevalence case ($P(G) = 0.1$), the prior odds are $\frac{0.1}{0.9} = \frac{1}{9}$. The posterior odds are $LR \times$prior odds $= 24.5 \times \frac{1}{9} = 24.5 : 9 = 2.72 : 1$.

We can also compute the likelihood ratio if the evidence were a negative test. This value turns out to be $\frac{1}{48}$, which is **not** the reciprocal of the LR for the positive test.

2.1.11 Recap

- Probability is the ____________________ language of ____________________
- Provides a common scale, from ________ to ________, for describing the chance that an event will occur
- **Conditional** probability is a key concept! The probabilitity of an event depends on what ___________________ is available
- Independent events can be powerful! They allow us to ______________ probabilities of events *directly*, as is common in ___________________________.
- _______________________________ is a mathematical result showing how we should _________________ our probabilities when available information changes.
  - This will later lead us to the likelihood ratio as a numerical ___________________ of the evidence.
  - Bayes' Rule does not necessarily describe how people operate in practice.

2.1.12 Probability and the Courts

Sally Clark was the only person in the house when her first child died unexpectedly at 3 months old. The cause of death was determined to be SIDS, sudden infant death syndrome. One year later, Sally and her husband had a second child, who died at 2 months old under similar circumstances. Sally was convicted of murder.

During her trial, a pediatrician testified that the probability of a single SIDS death for a family like the Clarks (similar income, etc.) was $\frac{1}{8500} \approx 0.0001$, and thus the probability of two SIDS death in the family was $\frac{1}{8500^2} = \frac{1}{73 \times 10^6} \approx 1.37 \times 10^{-8}$. There are several problems with this approach to evidence. What do you think? Jot down a few ideas below:

Issues with the evidence presented by the pediatrician:

1. Is the probability of a child dying of SIDS given, $\frac{1}{8500}$, correct for "families like the Clarks"?

2. The use of direct multiplication of probabilities assumes independence of the two deaths in the family. (Independence within the family is not a reasonable assumption.)
3. Alternative hypotheses (causes of death of the infants) were not considered. Did something else with perhaps a higher likelihood cause the children's deaths?

2.2 Probability to Statistical Inference

Probability is important, but it is only one tool in our toolbox. Another, more powerful tool is statistical inference.

2.2.1 Collecting Data

First, we consider data collection. Where do data come from? One data source is an *experiment*. An investigator designs a study and collects information on and maybe applies treatments to a *sample*, a subset of the population of interest. _____________ can tell us a great deal about how to design an ________________ or choose a _____________.

The area of statistics concerned with creating studies is called *experimental design*. The experimental design literature is extensive (see for example Morris (2011)). Here are a few crucial points:

- The goal of an experiment is to compare __________________
- Those __________________ must be ________________ assigned to units
- The _____________________ in the experiment must be large enough t obe able to make informed conclusions
- Blinding plays an important role in avoiding ___________. e.g. "double-blind" studies in medicine, where neither the patient nor the doctor administering the treatment know which treatment the patient is receiving

How is experimental design relevant to forensic science?

- Experiments are used to evaluate process improvements
- Blinding is used in "black box" studies, where examiners do not know ground truth

Experiments almost always involve *sampling* from the population of interest. Why?

- We sample because it is too _____________ or ___________________ to study the *entire* population
- A _____________ sample allows us to use the laws of ________________ to describe how certain we are that our __________________ answer reflects the __________________.
- There are many famous failures (cautionary tales) with __________________________ sampling. (See Figure 2.6.)

How is sampling relevant to forensic science?

- Sampling techniques used to determine which and how many bags of suspect powder collected from a crime scene to test.

Figure 2.6: This picture from the US presidential election of 1948 shows President Harry Truman, who won the election, holding a newspaper that went to print with the headline "Dewey Defeats Truman!" The headline was based on biased sampling that favored typically Republican demographics. Image Source: https://blogs.loc.gov/loc/2012/11/stop-the-presses/

All data collected can be divided into one of two groups: qualitative or quantitative.

- **Qualitative** data describe qualities about the observations. For example, the race of a suspect, or their level of education. There are two subcategories of qualitative data:
  - ____________________: the data belong to one of a discrete number of groups or categories. For example: blood type (A, B, AB, or O)
  - ____________________: the data belong to one group in a set of ordered values. For example, the evaluation of a teacher (poor, average, excellent). The categories have an inherent ordering, unlike in categorical data.
- **Quantitative** data describe quantities that can be measured on the observations. These are numerical data. There are also two subcategories of quantitative data:
  - ______________________: the values are distinct or separate. An easy-to-understand example is integer observations: $\{0, 1, 2, 3, 4, \ldots\}$. A forensic science example is consecutive matching striae on bullets or toolmarks. (See Figure 2.3)
  - ________________________: the values can take on any value in a finite or infinite interval. Continuous values fall anywhere on the number line. A forensic science example is the refractive index of a glass fragment.

2.2.2 Probability Distributions

Suppose we are to collect data on some characteristic for a sample of individuals or objects (e.g. weight, trace element concentration). A probability ________________ is used to describe these possible values and how ________________ each value is to occur. There are many, many possible probability distributions, but some of the most common are the Binomial, Poisson, Normal, and Lognormal distributions. The probabilities associated with each of these distributions and their possible outcomes are plotted in Figures 2.7-2.8.

- Discrete distributions
  - ________________________: counts the number of _______________ in a fixed number (n) of ________________. Possible values are $\{0, 1, 2, \ldots, n\}$.
  - ____________________: counts the number of _______________ occurring. Possible values are $\{0, 1, 2, 3, 4, 5, \ldots\}$

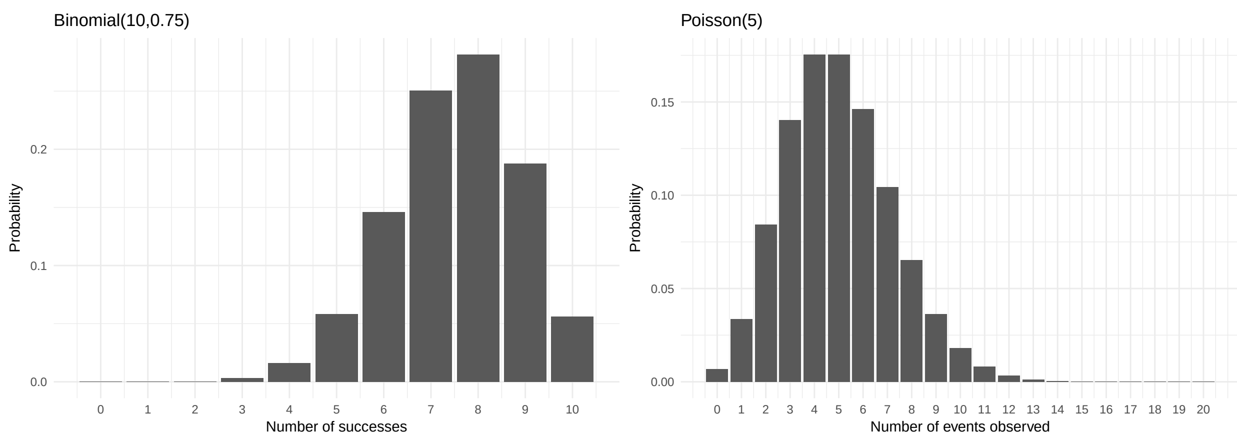

Figure 2.7: On the left, the probability of each possible outcome for variable with binomial distribution with 10 trials and probability of success 0.75. On the right, the probability of each possible outcome for variable with Poisson distribution with mean value 5.

- Continuous distributions
  - ______________________: the famous, symmetric "bell-shaped" ______________. Possible values are all real numbers, $(-\infty, \infty)$
  - ____________________: the (natural) logarithm of observations from this distribution follow a _______________ distribution. Possible values are all positive real numbers, $(0, \infty)$

2.2.2.1 Normal

You may already be familiar with this distribution with the bell-shaped curve. Measurement error is one example of something often assumed to follow a normal distribution. The normal distribution is described by two parameters: the ____________, denoted by μ, and the ______________________________, denoted by σ. If we have a variable (say, something observed in our data like weight), we give the variable a capital letter,

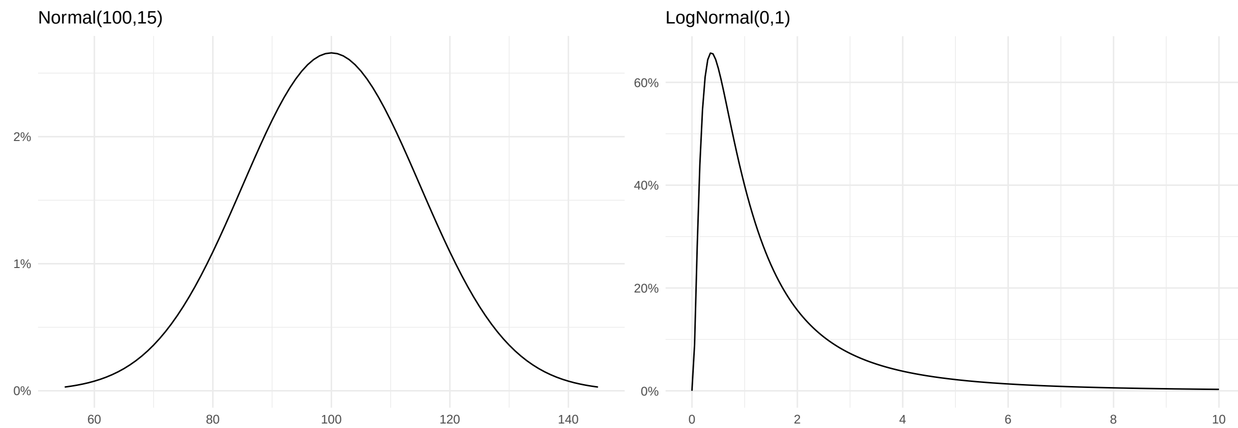

Figure 2.8: On the left, the probability distribution curve of possible outcomes for a variable with Normal distribution with mean value 100 and standard deviation 15. On the right, the probability distribution curve of possible outcomes for a variable with Lognormal distribution with mean value 0 and standard deviation 1.

typically X. If this variable X is normally distributed with mean μ and standard deviation σ, we write this as:

$$X \sim N(\underline{\hspace{1cm}}, \underline{\hspace{1cm}})$$

In measurement error, for example, we typically assume that the mean is 0. So, if X represents measurement error, we'd write $X \sim N(0, \sigma)$.

There are many nice properties of the normal distribution. For instance, we know that ________% of observable values lie within ________ standard deviations of the mean ($\mu \pm 2\sigma$), and also that ________% of observable values lie within ________ standard deviations of the mean ($\mu \pm 3\sigma$). When working with the normal distribution, we use software (such as Excel, Matlab, R, SAS, etc.), tables[3], or websites like onlinestatbook.com or stattrek.com to compute probabilities of events.

2.2.2.2 Lognormal

We often act as if everything is normally distributed, but of course this is not true. For instance, a quantity that is certain to be ________________ (greater than or equal to zero) cannot possible be normally distributed. Consider trace element concetration: either none is detected, or there is some amount greater than 0 detected.

In cases where nonnegative values are not possible, we may believe that the (natural) ______________ of the quantity is normal, which gives us a ________________ distribution for the quantity itself. The lognormal distribution, like the normal, has two parameters: mean (on the log scale), denoted ________, and standard deviation (on the log scale), denoted ________.

[3]See for example http://www.stat.ufl.edu/~athienit/Tables/Ztable.pdf

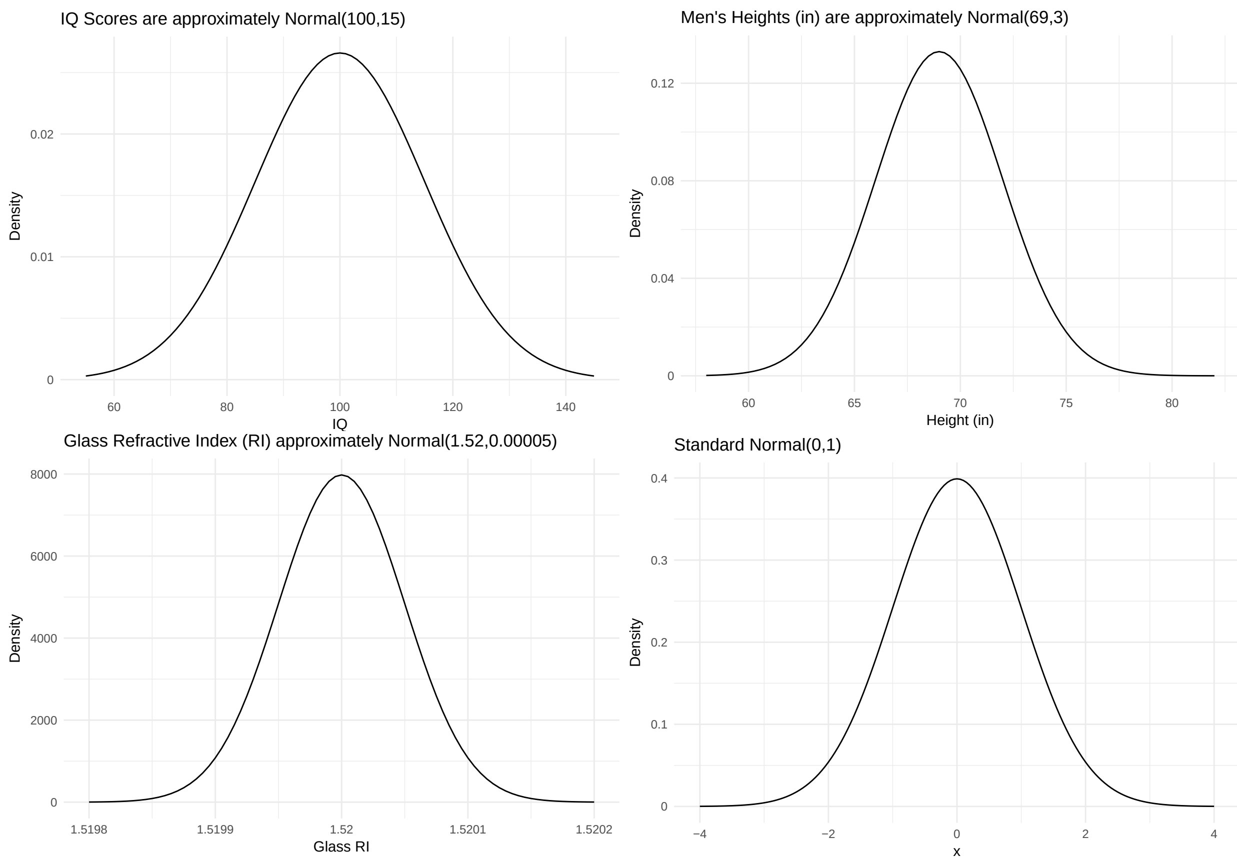

Figure 2.9: Four examples of the probability distribution functions for normally distributed variables

2.2.2.3 Discrete

Some quantities take on very few possible values. These are *discrete* data.

Recall the two common discrete distributions from section 2.2.2:

- Binomial:
  - Data are ____________ (two categories: "success" or "failure")
  - Data are a result of n independent ______________
  - $P(\text{success}) = p$ on each trial. (Same ______________ of success each time)
  - Expected number of successes you expect to see out of n trials: ________ $\times$ ________
  - Example: Suspect a student of cheating on an exam, response is the number of correct answers.
- Poisson:
  - Data are counts: number of events occurring in a ______________ time
  - The mean and the ___________________ of this distribution are the same, so the variablility in responses increases as the ______________ increases.

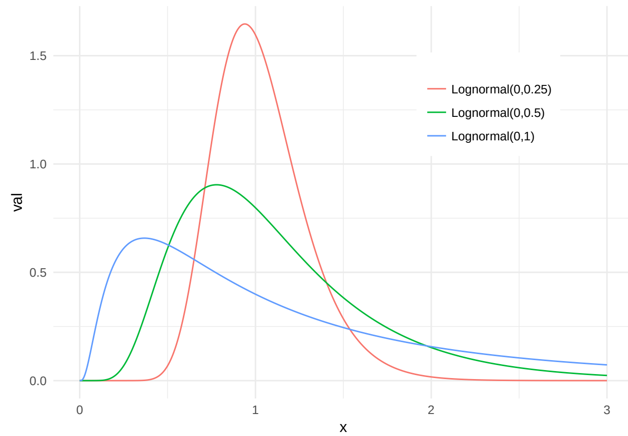

Figure 2.10: Three lognormal distributions with the same mean (on the log scale) and different standard deviations (on the log scale)

- Example: number of calls to 911 between 10:00 and midnight on Friday nights. See Figure 2.11 for a forensics example.

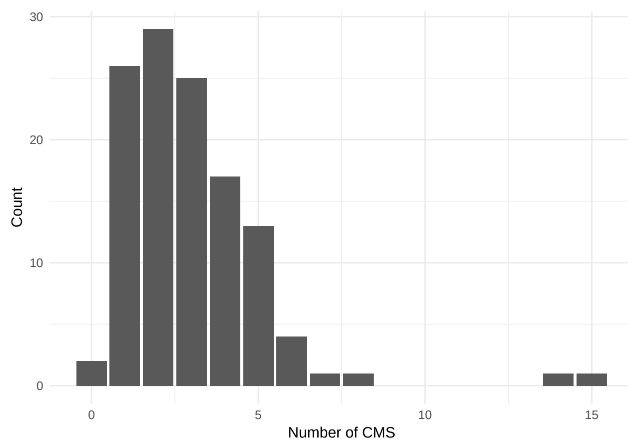

Figure 2.11: Distribution of the maximum number of CMS for a randomly selected bullet compared to 118 known lands approximately follows a Poisson distribution.

2.3 Statistical Inference - Estimation

Recall from Section 2.0.1:

- The ____________________ is the universe of objects of interest.
- The ____________________ is comprised of the objects available for study.
- ____________________ is deductive: use knowledge about the population to make statements describing the sample

- ___________________ is inductive: use knowledge about the sample to make statements describing the population
- Probability and statistics are used together!
  1. Build or assume a ____________ for a population
  2. Assess the ______________ using the model
  3. Refine the model, return to step 2.

2.3.1 Background

A __________________ is a numerical characteristic of the population, e.g. the population mean. Statistical methods are usually concerned with learning about population parameters from ________________________________.

Note: The mean of a *sample* and the mean of a *population* are differenct concepts. The mean of a sample can be calculated exactly, while the mean of a population is (usually) unknown, because there are too many objects in the population to record and calculate the mean.

The idea underlying statistical inference is that we can apply laws of probability to draw _______________ about a population from a sample. This process is briefly summarized below:

- Observe ______________ mean
- If we have a "good" sample this sample mean should be close to the _________________ mean.
- The laws of __________________ tell us how close we can expect them to be.

For example, suppose we are interested in the average height of the adult population in the U.S.

- Population: _______________________
- Sample: _______________________
- We can take the average height of everyone here and use this sample _________________ to make _________________ about the __________________ mean of all U.S. adults.
- Note: This approach will work if our sample is a ______________ sample from the population. This assumption may be questionable, so it should be verified.

The *goal* of statistical inference is __________________ about a ______________________. Different possible parameters are:

- Mean
- Variance
- Proportion

We can also make different types of inferential statements, depending on what question we are trying to answer and how we are going to report our results. We will talk about:

- __________________ estimate: an estimate of a parameter value

- __________________ estimate: a range of plausible values for a parameter
- Hypothesis __________________: examine a specific hypothesis about the true value of a parameter

When you want to do statistical inference, it is always inportant to look at your sample data before proceeding directly to inference. We do this because we want to

1. See general ____________________ in the data
2. Get an idea of the _________________ of the distribution of the data
3. Identify __________________ values and/or errors.

How we look at our data to check for these three things? If our data are ___________________, we look at a table of frequencies or a bar chart of the different outcomes. If our data are __________________, we look at histograms of the values, or numerical summaries such as mean, median, standard deviation, or percentiles.

A quick example shows why it can be important to examine your data before a formal statistical analysis:

- Suppose the data are (19,20,21,22,23). Then, the mean is $\frac{19+20+21+22+23}{5} = 21$, the median is 21, and the standard deviation is 1.58.
- But what if the data you receive are (19,20,21,22,93)? Then, the mean is $\frac{19+20+21+22+93}{5} = 35$, the median is 21, and the standard deviation is 32.4.
- There could have been a typo, or someone interpreted some handwriting wrong, etc. The moral of the story is ALWAYS look at your data first!

2.3.2 Point Estimation

An _______________ is a rule for estimating a population _______________ from a sample. We evaluate the quality of the estimator by considering two key properties:

- Bias: how close ________________________ an estimator is to the true population mean
- Variability: how _______________ is the estimate?

For the population mean, we might use sample mean as an estimator because it has ________ bias and ____________ variability is the sample is ______________. There are other possible estimators for the mean such as:

- The median (good for skewed data or data with outliers)
- The midrange ($\frac{\max + \min}{2}$)
- 47 (obviously this is just guessing and is not advised)

Let θ denote an unknown population parameter that we wish to estimate. The letter θ represents the true value of the parameter. In Figure 2.12, we see what would happen in many repeated attempts to estimate θ using estimators with different properties.

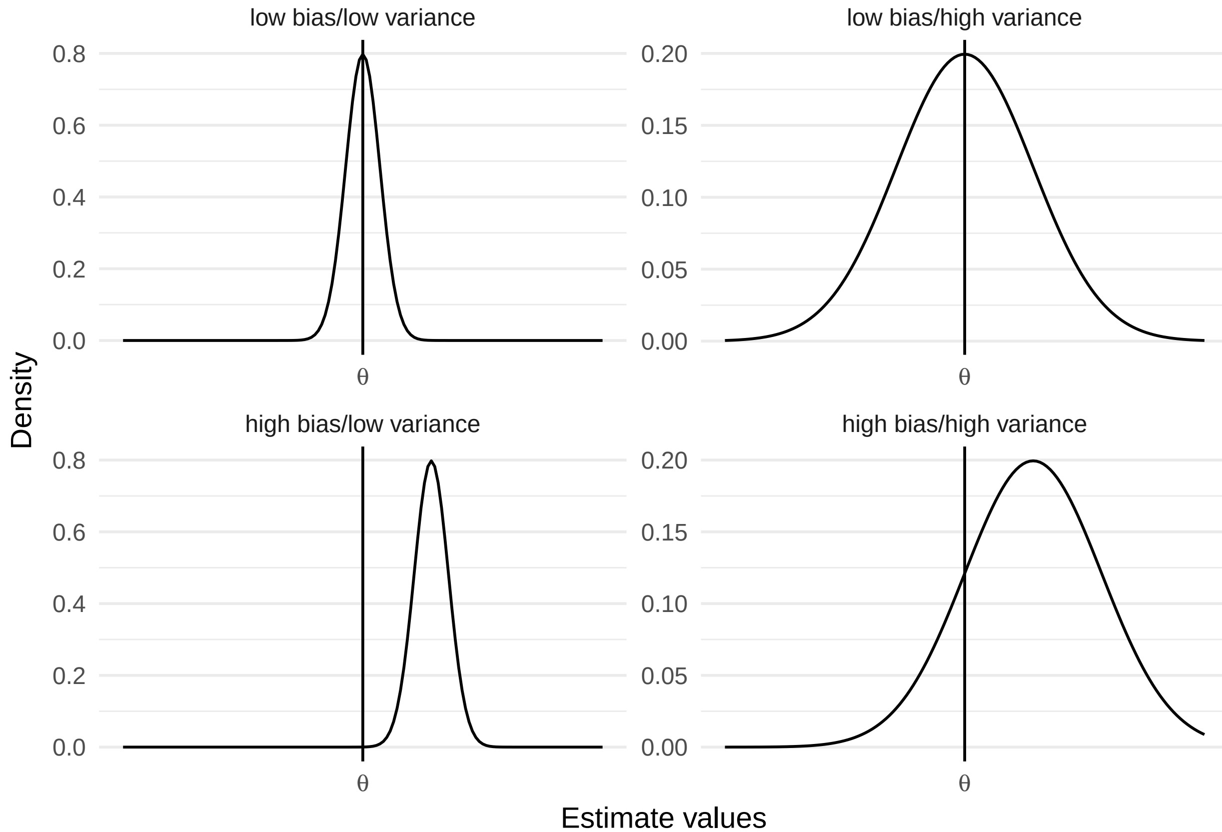

Figure 2.12: The curve in each plot shows the distribution of estimates we would see under each condition shown.

2.3.3 Standard Errors

One limitation of just providing a point estimate is that it doesn't give us any indication of ____________________. As we saw in Figure 2.12, a point estimate alone can be very different from the true mean. We can do better than this!

The ____ of an estimator measures the uncertainty in our estimate. When looking at a summary statistic, like mean, median, or percentiles, that statistic is also a __________________________ quantity. This means that if we had observed a different set of sample values, we would observe different values of the summary statistics. The idea of standard error is similar to the idea of standard deviation. Both are measures of spread, or variability. The difference is that standard deviation is a measure of variability of a sample or population, while standard error is a measure of variability of an _______________________.

Consider a population with that is normally distributed with mean 100 and standard deviation 15. Recall from Section 2.2.2.1 that 95% of observations from a normal distribution fall within two standard deviations of the mean. So, in this example we expect 95% of observations to be between 70 and 130. This distribution is shown on left in Figure 2.13. Now suppose we want to look at the distribution of the *estimates* of the population mean from several samples. This will demonstrate the idea of standard error, using sample size of

$n = 25$. The formula to compute standard error is:

$$se = \frac{\quad}{\sqrt{\quad}}$$

The standard error for this example is $\frac{15}{\sqrt{25}} = 3$. The mean of a sample of size 25 from this population should be about ____________. And about 95% of the time, the sample mean will be between ________ and ________. This distribution is shown on right in Figure 2.13.

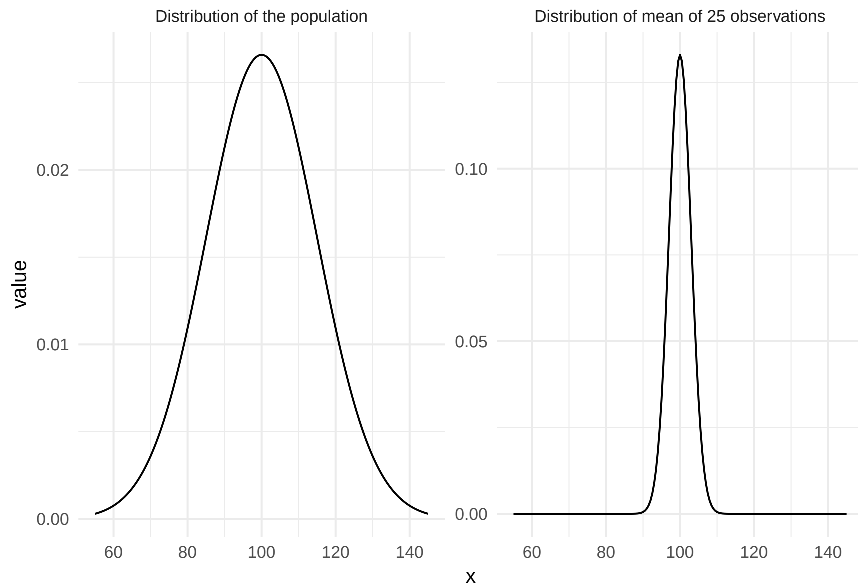

Figure 2.13: On the left, the distribution of the population, which is distributed N(100,15). On the right, the distribution of the sample mean for samples of size 25 from the populations, which is distributed N(100,3).

2.3.4 Sample Size

The size of a sample plays a *critical* role in determining how accurate we can be. Again, consider a population with distribution $N(100, 15)$. We can use _______________ to examine the effect of sample size. We simulate samples from a normal distribution with mean 100 and standard deviation 15. We use four different sample sizes: 10, 25, 50, and 100. We take 500 samples of each size and compute the mean for each sample, leaving us with 2,000 means that we have calculated. We show histograms of the means of these samples for each sample size in Figure 2.14.

2.3.5 Interval Estimation

A _________________ interval is an interval based on sample data that, with some specified confidence _____________, contains a population parameter. Essentially, a confidence interval takes a _________________ estimate and then adds some information about ______________________. Typically, we get an approximate

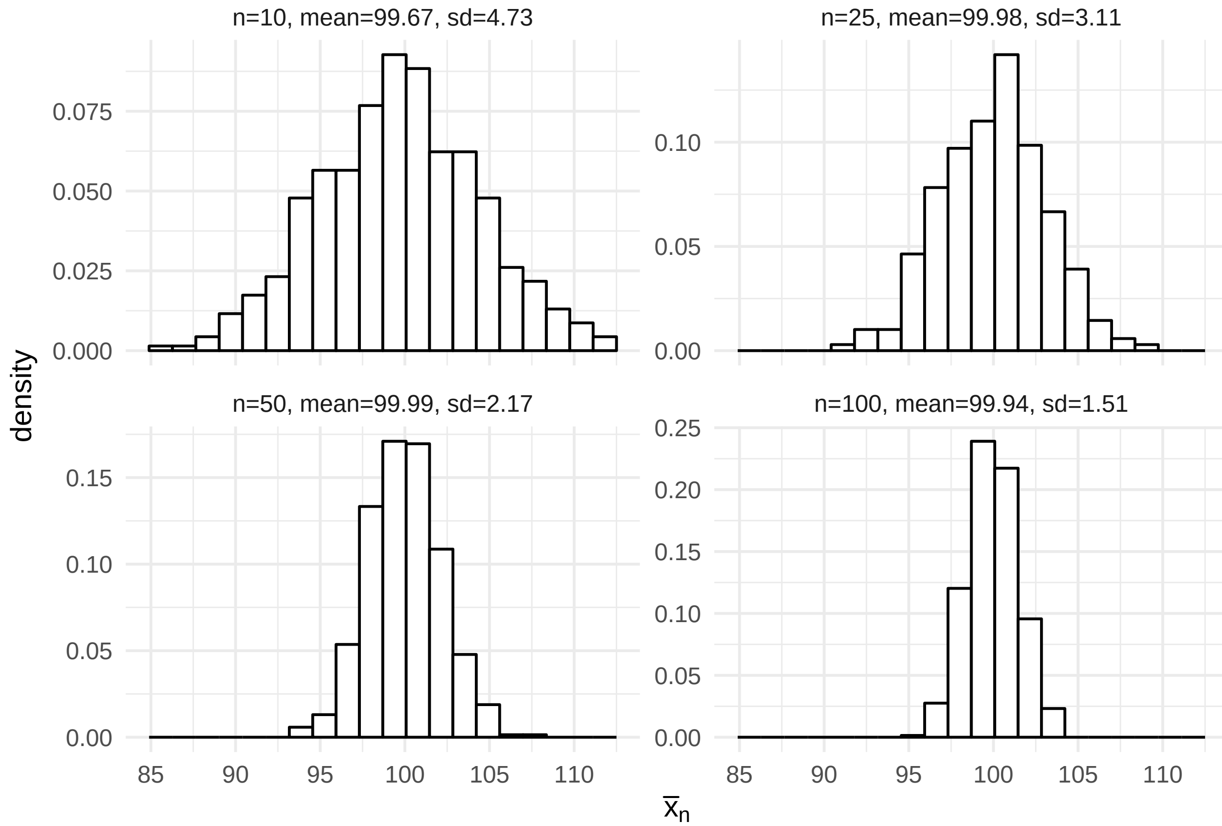

Figure 2.14: How does sample size affect the sampling distribution? As sample size increases, the standard error decreases, so the distribution of sample means becomes more narrow.

_______% confidence interval for a quantity by taking a point estimate and adding and subtracting 2 standard errors from it.

The most well-established procedures for finding confidence intervals are those related to drawing conclusion about the mean of a ________________ population. Suppose that we have acquired a random sample of n observations from a normal population.

- **Point** estimate: the natural point estimate of the population mean is the sample mean, as we've seen already. The sample mean is often denoted by _______. To calculate the sample mean, simple add up all the values and divide by the number of values you have, n. In mathematical notation this is written:

$$\overline{X} = \frac{1}{n} \cdot \sum_{i=1}^{n} X_i$$

- **Interval** estimate: denote the standard error of the sample mean by $SE(\overline{X})$, and the standard deviation of the population by SD(population). Then, compute the standard error by:

$$SE(\overline{X}) = \frac{SD(\text{population})}{\sqrt{n}}$$

Then, an approximate 95% confidence interval is computed:

$$\overline{X} \pm 2 \cdot SE(\overline{X}) \tag{2.9}$$

A key result is that these procedures for point and interval estimation work well even if the population does **not** follow a ________________ distribution **as long as the sample is ___________________**.

For example, suppose there are 10 glass fragments found at a crime scene, and the concentration of aluminum in each one is measured. The mean aluminum concentration of the sample was 0.73 and the standard deviation was 0.04. The standard error is thus:

$$SE(\overline{X}) = \frac{\quad}{\sqrt{\quad}} = 0.013$$

The approximate 95% confidence interval for the mean aluminum concentration in the crime scene window is:

$$0.73 \pm 1.96 \cdot 0.013 = (\underline{\quad\quad}, \underline{\quad\quad})$$

The *interpretation* of the confidence interval is important: 95% of the intervals _____________ in this way will contain the _________ population parameter.

2.4 Statistical Inference - Hypothesis Testing

Sometimes we wish to formally _____________ a hypothesis about a population parameter. The hypothesis to be evaluated is known as the __________ hypothesis. This hypothesis is usually the status quo, or what is assumed to be true. When hypothesis testing we look for evidence _____________ the null. There is also a an __________________ or research hypothesis that helps us design the test. If we _____________ the null hypothesis, then we say that we have a statistically __________________ result.

As with anything in life, errors are possible in hypothesis testing. There are two main typs of errors that we care about:

1. Type I: _______________ the null hypothesis when it is _____________. (false positive)
2. Type II: _______________________________ the null hypothesis when it is _____________. (false negative)

Of these two errors, Type I error is often considered more serious. We only want to ________________ the null hypothesis is there is *strong* evidence against it. These statistical testing ideas are closely related to concepts in the ___________________________________:

- The null hypothesis: the defendant is ________________
- The alternative hypothesis: the defendant is _______________
- In court, a Type I error is to find guilty when the defendant is __________________

- In court, a Type II error is to find innocent when the defendant is __________________

Ultimately, the basic idea of hypothesis testing is to compute a ___________________________ that measures "distance" between the _________ we have collected and what we expect under the _________ hypothesis. Typically, we use a test statistic of the form:

$$\frac{\text{point estimate} - \text{null hypothesis value}}{SE(\text{estimate})} \tag{2.10}$$

This test statistic can be interpreted as the number of _________________________ the sample estimate is from the ___________________ value under the null hypothesis.

A common way to to summarize hypothesis tests is by attaching a ________________ to the test statistic. This probability is called a __________________. The p-value of a hypothesis test gives the probability that we would get data like that in our sample (or something even more ______________), given our assumption that the null hypothesis is ___________. This idea is demonstrated in Figure 2.15.

Small p-values mean that we have observed ____________ data that lead us to question the ____________ hypothesis, which we have *assumed to be true*. Small p-values tell us that the sample data are unlikely to happen by chance under the null. A p-value, however, *only* addresses the ____________ hypothesis. It does *not* speak to the likelihood of the _______________ hypothesis being true.

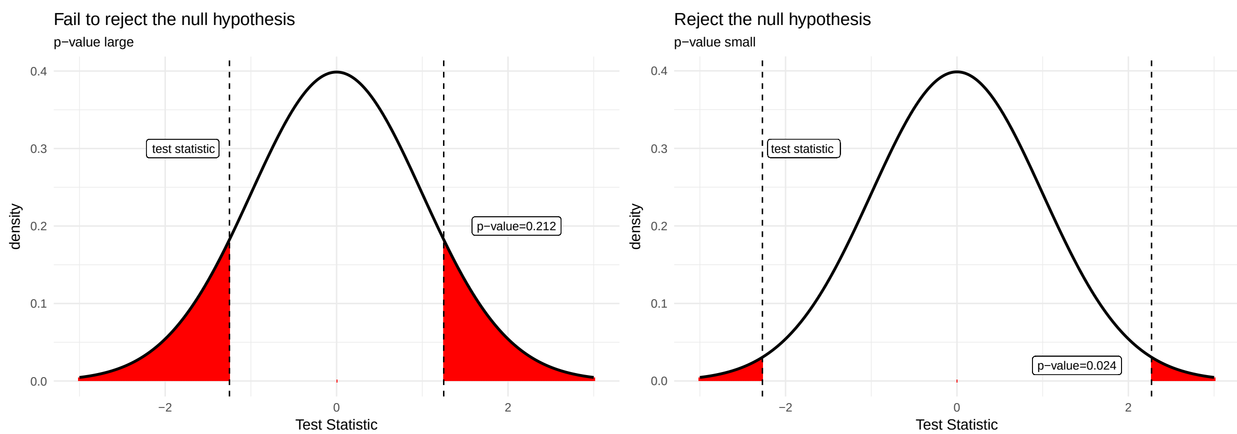

Figure 2.15: On the left, an example of a p-value that is large (test statistic small), leading us to fail to reject the null hypothesis. On the right, an example of a p-value that is small (test statistic large), leading us to reject the null hypothesis.

2.4.1 Normal Data

The most well-established hypothesis testing procedure is for testing a hypothesis about the _____________ of a __________________ population. This population parameter is denoted μ. We can test the hypothesis that the population mean, μ, is equal to some specified values μ_0. The hypotheses for this test are:

- Null hypothesis - $H_0 : \mu = \mu_0$

- Alternative hypothesis - $H_A : \mu \neq \mu_0$

The test statistic, call it T, to perform this hypothesis follows the form of Equation 2.10:

$$t = \frac{\_\_\_\_ - \_\_\_\_}{SE(\_\_\_\_)} \tag{2.11}$$

The p-value for this hypothesis test is obtained from a t distribution (see Figure 2.16) using software, a table of values, or an online calculator. These hypothesis testing procedures work well even if the population is *not* normally distributed **as long as the sample size is ____________________**.

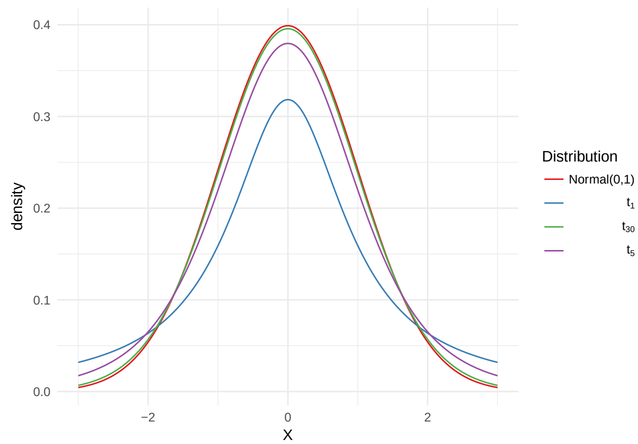

Figure 2.16: The t distribution is simlar to the Normal distribution (red line) is shape, but the tails are higher. As the sample size increases, the t distribution approaches the the Normal distribution.

2.4.1.1 Example

Suppose we want to estimate the mean amount of a trace element for the population of all bullets in Iowa. We get a random sample of 400 bullets from the state.

- The sample mean concentration is $\overline{X} = 55.5$
- The sample standard deviation is $s = 22.0$
- The standard error of the mean is $SE(\overline{X}) = \frac{22}{\sqrt{400}} = \frac{22}{20} = \_\_\_\_\_\_$

Suppose we have reason to believe that Remington (mean = 58) is the main producer in this area. We can check this idea with a hypothesis test.

- Null hypothesis - $H_0 : \mu = 58$
- Alternative hypothesis - $H_A : \mu \neq 58$
- Test statistic - $t = \frac{\overline{X} - \mu_0}{SE(\overline{X})} = \frac{55.5 - 58}{1.10} = -2.27$

The value of the test statistic is more than two standard errors away from the mean under the null hypothesis. The exact p-value is 0.023. This means that if the null hypothesis is true, then observing a value 2.27 standard errors or more away from the mean happens only 2.3% of the time. So, we reject our assumption that the population mean concentration is 58.

We can also calculate a 95% confidence interval for the mean using Equation 2.9:

$$55.5 \pm 1.96 \cdot 1.10 = (53.3, 57.7)$$

The hypothesized value (58) is is not in the 95% confidence interval, which also suggests that population mean concentration equal to 58 is not possible.

2.4.2 Confidence Intervals

There is a very close relationship between tests and interval estimates. Recall that a confidence interval (CI) gives a range of plausible values for the true population parameter, which here is the mean. A hypothesis test evaluates whether a specified (μ_0) is a __________________ value for the mean. A CI collects all values of μ_0 that we would find plausible in a test.

Statistical hypothesis test are *very* popular in practice. Sometimes, they address the scientific question of interest, but often they do not.[4]

2.4.3 Comparing Two Means

In section 2.4.1, we discussed hypothesis testing methods for one sample. In practice, we are often interested in comparing _________ samples from ________ different populations. For now, assume we have random samples from each of the two populations that we are interested in. The test we want to do is a test for _____________________ of parameters in the two populations.

2.4.3.1 Example

Suppose, for example, that we have collected broken glass at a crime scene, and glass fragments on a suspect. Define μ_{scene} to be the mean trace element level for population of glass at the scene. Define $\mu_{suspect}$ to be the mean element level for the population of glass on the suspect. We can compare the means to address the question of whether or not the glass fragments on the suspect could plausibly have come from the crime scene.

Hypotheses:

- $H_0 : \mu_{scene}$ _____ $\mu_{suspect}$
- $H_A : \mu_{scene}$ _____ $\mu_{suspect}$

Suppose 10 glass fragments are taken from the glass found at the scene (denote these by Y), and 9 fragments are found on the suspect (denote these by X). Concentrations of a trace element were measured in each fragment of glass. Summary values from the samples are:

[4] For more reading on this topic, consider this article from *Nature*: http://www.nature.com/news/psychology-journal-bans-p-values-1.17001

- $\overline{X} = 5.3$
- $s_X = 0.9$
- $SE(\overline{X}) = \frac{s_X}{\sqrt{n_X}} = \frac{0.9}{\sqrt{10}} = 0.28$
- $\overline{Y} = 5.9$
- $s_Y = 0.85$
- $SE(\overline{Y}) = \frac{s_Y}{\sqrt{n_Y}} = \frac{0.85}{\sqrt{9}} = 0.28$
- Obeserved difference $= \overline{X} - \overline{Y} =$ ____ $-$ ____ $= 0.6$
- The standard error for the *difference*, $\overline{X} - \overline{Y}$, is

$$SE(\{\overline{X} - \overline{Y}\}) = \sqrt{SE(\overline{X})^2 + SE(\overline{Y})^2} = \sqrt{\_\_\_\_ + \_\_\_\_} = 0.4$$

- The test statistic for this hypothesis test is:

$$t = \frac{(\overline{X} - \overline{Y}) - 0}{SE(\{\overline{X} - \overline{Y}\})} = \frac{0.6}{0.4} = 1.5$$

- The corresponding p-value for this statistic is 0.15.
- So, we fail to reject the null hypothesis that the two glass population means are equal.
- Interpretation is a *key* issue. When we say we fail to reject the null, we are saying there is a possibility of a common source.

2.4.4 Discussion

There are three key points for you to take away:

1. Hypothesis testing does *not* treat the two hypotheses ____________________. The null hypothesis is given priority. This is appropriate when there is reason to ________________ the null hypothesis until there is significant evidence _____________________ it. We don't necessarily always want this to be the case. (We will discuss this more later on in a forensic context.)

2. The p-values that result from hypothesis tests depend heavily on the sample size. If you have the same ___________ and standard deviation, but _________________ the sample size, the result will me more significant, due to the sample size alone.

3. Interpreting the results of the hypothesis test can be tricky. If we _______________ the null hypothesis, this does not necessarily mean that we have found an important difference in the context of our problem. In addition ____ the null hypothesis does not necessarily mean the null hypothesis is true.

2.5 Overview of Statistical Preliminaries

- We reviewed the basics of probability

  - Probability is the language of uncertainty
  - It is important to understand what is being assumed when talking about probability
  - For instance, the probability of having disease given a positive test is different than the probability of having a positive test given the disease
  - Probability distributions describe the variability in a population or in a series of measurements
- We reviewed basics of statistical inference
  - Statistical inference uses sample data to draw conclusions about a population
  - Point estimation, interval estimation, and hypothesis tests are main tools
  - It is critical that our procedures account for variation that could be observed due to chance

Chapter 3

Statistics for Forensic Science

In this section, we will first discuss a brief review of probability and statistics. Then, we will outline the forensic examination and discuss where probability and statistics can be added, and two different approaches to consider.

3.1 Brief Review of Probability and Statistics

- Probability is...
  - the language for describing uncertainty.
  - a number, always between 0 and 1, to describe the likelihood of an event.
  - dependent on the information available (information conditioned on)
  - useful for deducing likely values for individuals or samples from given or hypothesized information about the population
- Probability distributions...
  - suppose we have a *random* quantity, like a trace element concentration in a glass fragment.
  - give possible values and relative likelihood of each value observed or observable.
- Statistics...
  - draws inferences about a population (usually some characteristic of the population) based on sample data.
  - relies on careful definition of the "population" of interest.
  - relies on the method of data collection.
  - is made up of a variety of *inference* procedures, like...
    * point estiamtes,
    * confidence intervals, and
    * hypothesis tests.

3.2 The Forensic Examination

As you know there are a range of question that arise in forensic examinations, such as source conclusion, the timing of events, and cause & effect. We focus in this section on sources conclusions.

The evidence, which we will denote E, are items or objects found at the crime scene and on a suspect. E can also denote the measurements of these items. For evidence found at the crime scene, we will occasionally write __________, and for evidence found on the suspect or in the suspect's possession, we will occasionally write _________. There is also other information available to us, denoted I, such as the race of the perpetrator (according to a witness) or evidence substrate.

In the source conclusions piece of the forensic examination, we can divide the possible events into two groups:

- S : the items from the crime scene and from the suspect have _______________ source. In other words, the suspect is the ________________ of a crime scene item.
- $\overline{S}$: the items from the crime scene and from the suspect ________________ have common source.

The goal of the forensic examination is the assessment of evidence. There are two primary questions:

1. Do the items found at the crime scene and with the suspect appear to have a common source?
2. How unusual is it to observe source agreement *by chance*?

Obviously, there are many different types of forensic evidence:

- biological evidence (such as blood type or DNA)
- glass fragments
- fibers
- latent prints
- shoe prints or tire tracts
- and others

Different probability & statistics related issues will inevitably arise for different evidence types. For example:

- Discrete and continuous variables are treated differently.
- What information is available about the probability distribution of observable measurements?
- A reference database may or may not exist.
- What role does the manufacturing process play in ability to make a match?

The Daubert standard[1] identifies the the judge as a ________________________________ to determine the admissibility of expert scientific testimony. In order to determine admissibility, the judge can apply ____________________ factors:

- The theory or method should be ________________
- The theory or method should be subject to _________________ and ___________________.
- There are known or potential _______________ rates.

[1] *Daubert v. Merrell Dow Pharmaceuticals*, 509 U.S. 579

- The theory or method has ________________ and controls.
- The theory or method is generally ________________ by the _________________ scientific community.

The National Research Council (Committee on Identifying the Needs of the Forensic Sciences Community and Council (2009)) found:

- ____________________ provider community (federal, state, & local)
- ____________________ across disciplines
- Lack of ___________________________ in practices
- Insufficient _______________________
- Questions underlying ______________________ basis for some conclusions

This led to *single source* DNA's emergence as a __________________________________.

In 2016, the President's Council of Advisors on Science and Technology released a report on the state of forensic science. (Advisors on Science & Technology (2016)). This report…

- focused on the _______________ of ________________ matching disciplines:
  - examined foundational validity and the use of ____________________________ studies
  - examined validity as _______________, including information at the examiner level.

The forensic science community as a whole is a community in transition. The National Commission on Forensic Science, which was established in 2013 to advise the U.S. Attorney General, was not renewed for a third term after its second term expired on April 23, 2017.[2] To accompany the end of this commission, the Department of Justice (DOJ) released a call for comments on advancing forensic science, with comments closing on June 9, 2017.[3]

There are additional federal organizations that are also involved in advancing the field of forensic science. For instance, Organization of Scientific Area Committees (OSAC) for Forensic Sciences is also a part of NIST.[4] In addition, the NIJ and NIST have forensic Centers of Excellence, the Forensic Technology Center of Excellence[5] (FTCoE) and the Center for Statistics and Applications in Forensic Evidence.

3.3 Common Approaches to Assessing Forensic Evidence

In this section, we consider the two primary approaches to assessing forensic evidence: expert assessment based on experience and training, and statistical approaches, including statistical testing and likelihood ratio

[2] https://www.washingtonpost.com/local/public-safety/sessions-orders-justice-dept-to-end-forensic-science-commission-suspend-review-policy/2017/04/10/2dada0ca-1c96-11e7-9887-1a5314b56a08_story.html?utm_term=.858a461dec18

[3] https://www.justice.gov/opa/press-release/file/956146/download

[4] https://www.nist.gov/topics/forensic-science/organization-scientific-area-committees-osac

[5] NIJ: https://forensiccoe.org/

based approaches.

Ultimately, the goal of the collaboration between academics and practitioners is to come to a combination of the two as the gold standard for assessing forensic evidence.

3.3.1 Significance Testing / Coincidence Probability

One common statistical approach[6] solves the forensic problem in two stages:

1. First, we determine if the crime scene and suspect objects ____________ on characteristic of _______________. This is typically done using a hypothesis or a significance ___________.
2. Second, we assess the ________________________ of thi agreement by finding the likelihood of such agreement occurring by ____________.

Note: DNA analysis can be categorized in this way, but is usually thought of as a *likelihood ratio* approach. See Section 3.3.4 for more.

In the significance testing or coincidence probability approach, determining agreement is straightforward for ____________ data like blood type or gender. There are still ___________________ in these cases due to the possibility of laboratory or measurement error. Usually, it is easier or more straightforward to think about discrete data in terms of the likelihood ratio. Again, see Section 3.3.4 for more. Statistical significance tests can also be used for ___________________ data like trace element concentrations (e.g., in glass fragments).

For this approach, the testing procedure is outlined below:

1. _______________________ each object by a __________ value. For example, consider the mean trace element concentration in a *population* of glass fragments. This is the _______________ mean in statistics terminology: one mean for glass from the crime scene, one mean for glass from the suspect)
2. Obtain _____________ values from the ________________ object.
3. Obtain _____________ values from the ________________ object.
4. Use the sample values from steps 2-3 to test the ______________________ that the two objects have the same population _______________. The common tool for testing is the t-test demonstrated earlier in Section 2.4.3.
5. Summarize the test with the p-value, the _________________ of data like the observed data or more extreme, assuming population means are the same.
6. A small p-value, typically small means $p < .05$ or _______, indicates there is no agreement between the two objects. If the p-value is larger, we can't ________________ the hypothesis that the two means are equal.[7]

[6] This approach is also known as the *comparison/significance approach*.
[7] But is this evidence that they came from the same population. . . ?

3.3.1.1 Examples

First, consider two glass samples: one from a crime scene, and one from a sample recovered from the suspect. These data can be found in Curran et al. (1997). The null hypothesis (H_0) is that the two samples come from the same source, and the alternative hypothesis (H_A) is that the two samples are from different sources.

- Five measurements of ____________________ concentration in crime scene sample:

$$0.751, 0.659, 0.746, 0.772, 0.722$$

- Five measurements of ____________________ concentration in recovered sample:

$$0.752, 0.739, 0.695, 0.741, 0.715$$

- Sample means:
  - Crime Scene: 0.730
  - Recovered Sample: 0.728
- Standard errors
  - Crime Scene: $\frac{0.0435}{\sqrt{5}} = 0.019$
  - Recovered Sample: $\frac{0.023}{\sqrt{5}} = 0.010$
- Test statistic (see Section 2.4.3):

$$\frac{0.730 - 0.728}{\sqrt{0.019^2 + 0.010^2}} = \frac{0.002}{0.0215} = 0.0931 \approx 0.1$$

- p-value $= 0.70 \Rightarrow$ fail to reject the null hypothesis that the two samples come from the same source

In fact, ground truth is known here: these measurements did come from the same bottle.

Next, consider a different recovered sample. Again, the data are from Curran et al. (1997). The crime scene sample remains the same as the prior example.

- Five measurements of aluminum concentration in the second recovered sample:

$$0.929, 0.859, 0.845, 0.931, 0.915$$

- Sample means:
  - Crime Scene: 0.730
  - Recovered Sample: 0.896
- Standard errors
  - Crime Scene: $\frac{0.0435}{\sqrt{5}} = 0.019$
  - Recovered Sample: $\frac{0.0408}{\sqrt{5}} = 0.018$
- Test statistic (see Section 2.4.3):

$$\frac{0.730 - 0.896}{\sqrt{0.019^2 + 0.018^2}} = \frac{-0.166}{0.0262} = -6.38$$

- p-value $= 0.0015 \Rightarrow$ Reject the null hypothesis that the two samples come from the same source.

In fact, ground truth is again known, and these two samples are from two different bottles.

3.3.1.2 Other Significance Testing Approaches

Many other alternative, related methods exist for assessing forensic evidence. For instance, 4-σ (4-sigma) methods create an interval for each element in each sample, which are formed by taking the mean concentration of each element $\pm$ four standard errors of those means. Then, check each interval for overlap, using "control" sample to obtain an expected range and checking whether the "test" samples are in/out of the control range. Hotelling's T^2 (T-squared) test compares all elements simultaneously to account for the within-sample dependence.

There are some **technical** concerns about the aforementioned procedures. The formal tests, the t-test and Hotelling's T^2 test, require assumptions about the probability distribution of the data. In addition, univariate procedures such as the t-test are repeated on multiple elements, and the existence of multiple comparisons should be accounted for. Furthermore, univariate procedures ignore information in the correlation of elements multivariate procedures (like Hotelling's test) require large samples.

The bigger concerns with these procedures are **conceptual**. First, significance tests do not treat the two hypotheses (equal means vs. unequal means) symmetrically:

- The null hypothesis, that the means are equal, is *assumed* true unless the data *rejects* the null hypothesis
- Failing to reject the null in this hypothesis test setting is taken as evidence *against* the suspect. This is the opposite of the courtroom setup, where failing to reject the null is taken as *lack of* evidence against the suspect. Thus, the asymmetry of the null and alternative hypotheses is an issue here. In addition, the binary decision to reject the null fail to reject the null requires an arbitrary cutoff value. e.g. Why 4σ rather than 3σ? Why $p = 0.05$ rather than $p = 0.01$ or $p = 0.10$? Lastly, the match decision from the hypothesis test is separated from assessment of the practical significance of the match. How different two samples are, according to the hypothesis test, may not correspond to the ground truth. e.g. a large sample size will decrease the threshold for rejection, which could cause true matches to be misclassified as different sources if the within-population variation is large enough. And conversely, a small sample size will increase the rejection threshold and will be unable to distinguish true non-matches from matches because of a lack of information about the two truly different populations.

3.3.1.3 Alternatives to Significance Tests

Another route to investigate is **equivalence testing** instead of significance testing. Equivalence testing changes the null hypothesis and addresses the first concern regarding asymmetric hypostheses. The Bayesian approach and the likelihood ratio approah address the other concerns: these methods avoid the binary decision and the separation of match and significance. We discuss these methods further later on in Section 3.3.4 and focus on equivalence testing for now.

The usual hypotheis test assumes the null hypothesis is true until proven otherwise. **Equivalence testing** is an alternative approach that *assumes the population means are different. This becomes the null hypothesis, but it also requires us to specify a "practically" important difference, Δ. We then write the hypotheses as:

$$\begin{aligned} H_0 &: |\mu_{scene} - \mu_{suspect}| > \Delta \\ H_A &: |\mu_{scene} - \mu_{suspect}| < \Delta \end{aligned} \tag{3.1}$$

This requires that we test two different hypotheses:

1. the means differ by more than _________ vs the alternative that they don't
2. the means differ by less than _________ vs the alternative that they don't

Here, we reject the null hypothesis and conclude the samples are equivalent ONLY if we get a small p-value for both hypothesis tests.

Recall the example from Section 2.4.3 comparing two glass samples 10 glass fragments that were taken from glass at the scene (Y) and 9 fragments that were found on the suspect (X). The statistics recorded were:

- $\overline{X} = 5.3$; $s_X = 0.9$; $SE(\overline{X}) = 0.28$
- $\overline{Y} = 5.9$; $s_Y = 0.85$; $SE(\overline{Y}) = 0.28$
- $SE(\{\overline{X} - \overline{Y}\}) = 0.4$

In Section 2.4.3 we tested the hypothesis $H_0 : \mu_X = \mu_Y$ and obtained a p-value of 0.15. So, we failed to reject the null, meaning there was not evidence the two samples came from poplations with identical means. (i.e. They have the same source.)

In the *equivalence testing* approach, let's suppose a difference of 1.0 or more ($\Delta = 1.0$) is considered "distinguishable". Then our hypotheses are:

$$\begin{aligned} H_0 &: \mu_y - \mu_x \geq 1 \\ H_A &: \mu_y - \mu_x < 1 \end{aligned} \tag{3.2}$$

The observed difference is 0.6, with a standard error of 0.4. This observed difference ce is one standard error *below* the practically important difference of $\Delta = 0.1$, which results in a p-value of 0.32. This equivalence test does not reject **its** corresponding null hypothesis, and thus we cannot reject the possibility that the two samples come from populations with distinguishable means.

Now, let's return to the usual significance testing approach and assume we have found a statistical "match", meaning we could not reject the null hypothesis. The second stage of our analysis is assessing the "______________________" of the match. Note that we put significance in quotes above because the word "significance" has a formal statistical meaning, and so we try not to use that term here. Other terms we could use in place of "significance" are:

- Strength of _________________
- _______________ of evidence
- Usefulness of ________________

- ________________ value

Consider, for an example, the suspect in the movie *The Fugitive* (1993). (See Figure 3.1.) If we know that the suspect and the criminal are...

1. both male, that provides us with limited evidence. (About 50% of the population is male. This "clue" is not very informative.)
2. both one-armed males, that provides us with stronger evidence. (Of that 50% of the population, some unknown but assumed very small proportion of them has only one arm. This "clue" is much more informative.)

This step, where we quantify the strength or probative value of evidence is *crucial* for the courtroom setting.

Figure 3.1: Harrison Ford's character on the hunt for the one-armed man who killed his wife. Image Source: http://www.imdb.com/title/tt0106977/

3.3.2 Strength of Evidence: Discrete Data

For discrete data like blood type and DNA, when we want to find the probability of a match by chance, there are several important considerations:

1. This evidence is usually ____________________ centered: material from scene is considered ______________ and we want to compute the likelihood that an individual would have a similar object/characteristic.

2. This evidence depends on relevant "________________". For instance, the suspect is male, the suspect is Chinese, etc.

3. We have to consider *where* our data come from. This could be from population records or convenience samples.

These concerns are equally relevant to the likelihood ratio approach so we don't discuss them further here. See 3.3.4 for more.

3.3.3 Strength of Evidence: Continuous Data

For continuous data, finding the probability of a match by chance is typically a bit harder to do. We need the likelihood that objects (e.g. glass fragments) selected at random would match crime scene sample. The basic idea is outlined below, using terms from the *t*-test context. (See Section 2.4.1 for reference.)

1. Suppose for the moment we know the "population" mean of a randomly chosen glass source.
2. We can find the probability that a *t*-test based on a sample from this *random* object will result in agreement with the *crime scene* sample.
3. The total coincidental agreement probability is an average over *all possible choices* for the random source:

$$\text{coincidence probability} = \sum_{\text{pop. means}} Pr(\text{a mean})Pr(\text{match to scene sample}|\text{a mean})$$

This procedure is technically challenging but it can be done! The key question is: Where does the information about the set of possible random sources (i.e. the relevant population) come from?

3.3.3.1 Example

Let's illustrate this idea with an example. Consider again the data from Curran et al. (1997). Recall the crime scene example, call it X, has 5 observations with a mean aluminum concentration of $\overline{X} = 0.730$ and standard deviation of 0.04.

Assume we will apply a standard statistical test with 5 samples from an unknown randomly chosen glass source, with a cutoff corresponding to a p-value of 0.05. So, we will only reject the null hypothesis (that the random sample comes from the same source as the crime scene sample) if the test statistic is so large that it has a less than 5% chance of occurring at random. Any value that falls in the 95% non-rejection range will be deemed "indistinguishable" from the crime scene sample.

Next, suppose that the means of *all* randomly chosen glass sources from the population of interest can be descried using a normal distribution with mean $\mu = 0.83$ and standard deviation $\sigma = 0.10$. This distribution is shown in Figure 3.2. In the population of glass sources, some randomly chosen sources will have means near 0.73 and will be ____________ to distinguish from the control sample. Some randomly chosen sources will have means near 0.83 and will likely be ____________________ from the control sample. Some randomly chosen sources will have means near .93 and will be ____________ to distinguish from the control sample.

Under this setup, we can compute how likely it is to find a randomly chosen source which provides a sample that is indistinguishable from the crime scene sample. The answer is 0.24 or 24% of the time

We can repeat the same idea for finding coincidence probabilities for different population means and standard deviations. The table shown below gives the different coincidence probabilities (cp) for different population means and standard deviations.

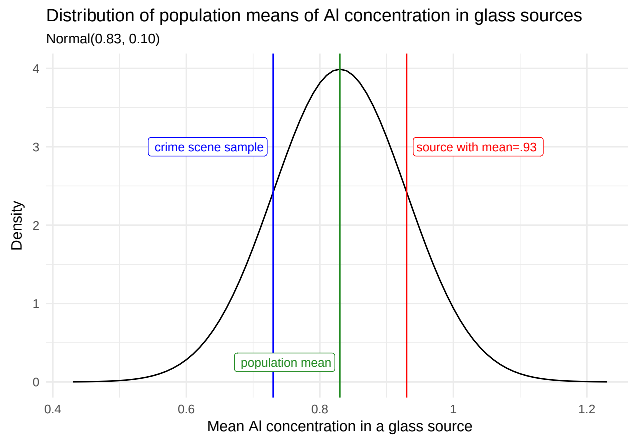

Figure 3.2: The population distribution of mean aluminum concentrations in all randomly chosen glass sources.

Table 3.1: Coincidence probabilities for different population distributions.

mean	sd	cp
0.73	0.2	0.2
0.83	0.2	0.37
0.93	0.2	0.65
0.73	0.1	0.37
0.83	0.1	0.24
0.93	0.1	0.17
0.73	0.05	0.12
0.83	0.05	0.06
0.93	0.05	0.002

The moral of the story, is that the probability of a coincidental match is high when there is…

- a small difference between control sample and the population of randomly chosen sources (i.e. the crime scene/control sample is “ordinary”).
- a large amount of heterogeneity among the potential sources in the population.
- a large amount of variability among the fragments in an individual source.

3.3.4 Likelihood Ratio

The goal for the __________ of ___________ in a courtroom setting is to make a decision about the relative likelihood of two hypotheses (e.g. same source or different source) *given* the data. In statistical terms, this is a **Bayesian formulation** because we ask for probabilities about the state of the world *given* observed data. Recall from Section 2.1.9 Bayes’ Rule (or Bayes’ Theorem): given two events A and B we have the probability of A *given* B:

$$P(A|B) = \frac{P(A)P(B|A)}{P(B)}$$

Bayes' rule is a way of _______________ the direction of conditional probabilities. We can go from statements about the likelihood of the **evidence** given the ________________ to statements about the likelihood of the **hypotheses** given the ________________.

Formally, for the evidence (E) and having same source (S), we write:

$$P(S|E) = \frac{P(E|S)P(S)}{P(E|S)P(S) + P(E|\overline{S})P(\overline{S})}.$$

Recall from Section 2.1.10 that Bayes' Rule can be rewritten in terms of the odds in favor of the same source hypothesis (left-hand side of the equation):

$$\frac{P(S|E)}{P(\overline{S}|E)} = \frac{P(E|S)}{P(E|\overline{S})} \frac{P(S)}{P(\overline{S})}$$

In words, we can describe the above equation as "the posterior odds of the same source hypothesis is equal to the likelihood ratio of the evidence times the prior odd of the same source hypothesis." The *likelihood ratio* (sometimes called the Bayes Factor) is a measure of the value of the evidence. It does *not* depend on the prior beliefs with regards to the same source hypothesis.

$$LR = \frac{P(E|S)}{P(E|\overline{S})} \tag{3.3}$$

The term "likelihood" is used because if E includes continuous measurements, then we cannot talk about probability of single events. The likelihood ratio could, in principle, be used with E representing "all" evidence of all types (more on this later). In addition, other available information (e.g. background) can be incorporated into the LR (more on this later as well).

The **interpretation** of the likelihood ratio is crucial. The derivation of the LR (see 2.8) shows that the LR is a factor that we should use to change our same source odds. Furthermore, there are some proposals for scales (e.g. ENFSI) that map LRs to words:

- LR from 2-10 implies ________ support of the same source hypothesis
- LR from 10-100 implies ____________ support of the same source hypothesis
- See page 17 of http://enfsi.eu/wp-content/uploads/2016/09/m1_guideline.pdf for the complete scale from ENFSI (European Network of Forensic Science Institutes)

There is some confusion about terminology. To clear this up, it is important to understand how the LR approach and the Bayesian approach relate. The LR (often called the Bayes Factor) plays a central role in a Bayesian approach to forensic evidence: it is the quantity used to update *a priori* odds in order to obtain posterior odds. The true distinction between the LR and the Bayes factor is technical and has to do with how statistical parameters are treated.

Let's return to Equation 3.3. The numerator, $P(E|S)$ assumes ______________ source and asks about the likelihood of the evidence in that case. This value is somewhat related to finding a p-value for testing the hypothesis of equal means but, there is no binary decision regarding a match in the LR approach. Instead, think of the LR as a quantitative measure of likelihood of evidence under the same source hypothesis, S. The denominator, $P(E|\overline{S})$, assumes no ____________ ___________ and asks about the likelihood of the evidence in that case. This is analogous to finding coincidence probability like we did in Section 3.3.3. Here too, as in the numerator, we do not require a binary decision regarding a match. The denominator is a quantitative measure of likelihood of evidence under _________.

The LR approach makes explicit the need to consider the evidence under ________ different hypotheses. It also separates "__________________" information about evidence from "_________________" assessments of the same source hypothesis S. There is some subtlety here. Do not fall prey to the...

- Prosecutor's ______________: interpreting $P(E|\overline{S})$ as $P(\overline{S}|E)$. i.e. if evidence is unlikely under $\overline{S}$ that fact is mistakenly interpreted as saying that $\overline{S}$ is unlikely.

- ___________________ attorney's fallacy: other misinterpretations of $P(E|\overline{S})$, such as if $P(E|\overline{S}) = \frac{1}{1{,}000{,}000}$, then there are 300 other people in the U.S. who could have been the source (and thus committed the crime).

Let's define $E = (x, y)$, where y represents the measurement of evidence from the crime scene, and x represents the measurement of evidence from the suspect. Then, we can rewrite the likelihood ratio using laws of probability from Section 2.1.6:

$$
\begin{aligned}
LR &= \frac{P(E|S)}{P(E|\overline{S})} \\
&= \frac{p(x, y|S)}{p(x, y|\overline{S}} \\
&= \frac{p(y|x, S)}{p(y|x, \overline{S}} \cdot \frac{p(x|S)}{p(x|\overline{S}}
\end{aligned}
\tag{3.4}
$$

Often, the likelihood of x is the same for _________ and _________, i.e. $p(x|S) = p(x|\overline{S})$. In other words, the distribution of the suspect's data does not depend on who committed the crime. This leads to yet another way to write the likelihood ratio:

$$
LR = \frac{p(y|x, S)}{p(y|x, \overline{S}} \tag{3.5}
$$

So, the likelihood ratio is the ratio of the probability of finding the evidence at the crime scene *given* the evidence from the suspect **and** the same source assumption, to the probability of finding the evidence at the crime scene *given* the evidence from the suspect **and** the different source assumption.

Let's assume we start with Equation 3.5:

- In the _________________ case, the numerator (probability of finding the evidence at the crime scene *given* the evidence from the suspect **and** the same source assumption) is typically 0 or 1, or values really close to those. (We should also consider the possibility of a lab error or contamination.) The denominator (probability of finding the evidence at the crime scene *given* the evidence from the suspect **and** the different source assumption) is then the probability of a ____________________ match.

- In the ____________________ case, the numerator is a measure of how likely it is to observe the numbers (x, y) if they represent multiple measures from the same source. The denominator is a measure of how likely it is to observe the numbers (x, y) if they are measures from two different sources.

3.3.4.1 Example

Suppose the evidence is the blood type for a crime scene sample and the blood type for a suspect sample. We have information about the distribution of blood types in the population:

Type	A	B	AB	O
U.S. frequency	0.42	0.10	0.04	0.44

Suppose both samples are observed to be of blood type O. Then, $Pr(y = O|x = O, S) \approx 1$. We'd expect to see the same blood type if the samples come from the same source. Conversely, $Pr(y = O|x = O, \overline{S}) = 0.44$. Blood type O is fairly common in the U.S. So the LR is $LR \approx \frac{1}{0.44} \approx 2.27$. Following the ENFSI language, the evidence provides ___________ support for the "same source" hypothesis. Blood type AB, however, is rare in the U.S. If the two samples were both AB, then LR would indicate stronger evidence (LR would be $\approx \frac{1}{.04} \approx 25$).

3.3.4.2 Where it works: DNA

A DNA profile identifies alleles at a number of different locations along the genome. For example, alleles at location TH01 are 7,9. As with blood type, we may see matching profiles from the crime scene and from the suspect. The numerator is also approximately one as in blood type example because we expect DNA samples from the same source to be indistinguishable. We can the determine the probability of a coincidental match for each marker or location:

TH01	4	5	6	7	8	9	9.3	10	11
Freq.	0.001	0.001	0.266	0.160	0.135	0.199	0.200	0.038	0.001

For TH01 agreeing on alleles 7, 9, the probability of a random agreement is $2.16.199 = .064$. Thus, the LR in this case is $\frac{1}{0.064} \approx 15$.

DNA evidence consists of data for a number of locations (CODIS used 13 locations pre-2017). The locations on different chromosomes are independent. Recall that if events are independent, then we can multiply probabilities (which basically means multiplying likelihood ratios). So, a match at *all* locations can lead to likelihood ratios in the **billions** (or even larger).

The likelihood ratio approach works well for DNA because the underlying biology is well understood. The probability model for the evidence follows directly from genetic theory, and population databases are available for computing random match probabilities. In addition, for single-source DNA sample, the methodology has been peer–reviewed and is well accepted by the scientific community. But, even with the above information, there are still problems arising in the DNA forensic field. For example, allele calling still has some subjective elements despite the reputation of pure objectivity, and DNA samples containing multiple sources (i.e. DNA mixtures) still are not as well-understood as single-source samples.

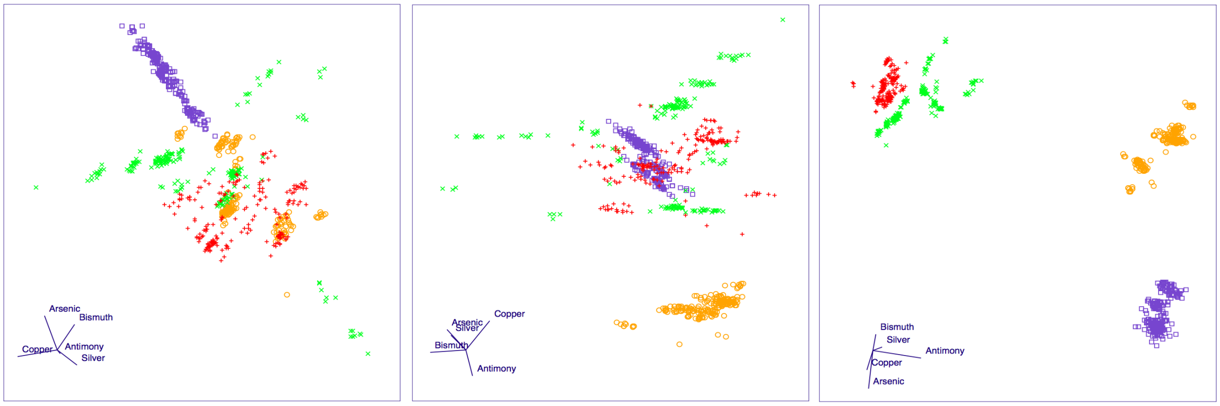

Figure 3.3: Three projections of 5-dimensional bullet lead trace element data.

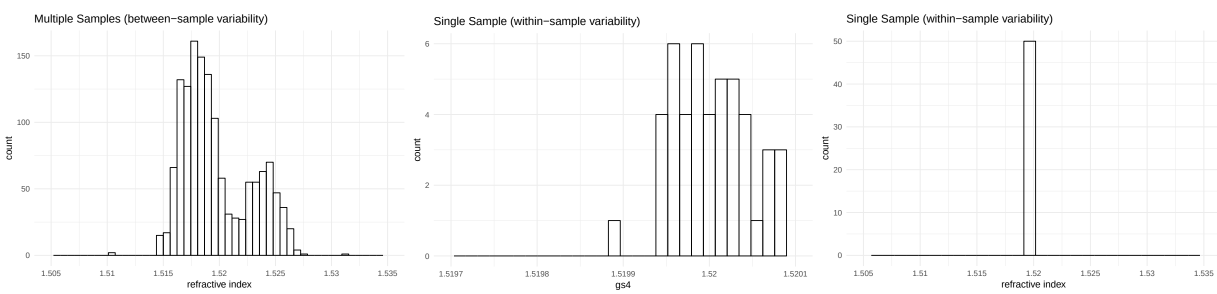

Figure 3.4: From left to right: multiple glass source samples where refractive index was reported, the same measurements for just one of those sources, and the data from the middle source on the same x axis as all three sources.

3.3.4.3 Where it can work: Trace evidence

Glass and bullet lead are examples of where the likelihood ratio approach can potentially work. We can measure chemical concentrations of elements in glass (or bullet lead). There may be broken glass at the crime scene and glass fragments on a suspect. But can we construct a likelihood ratio for evidence of this type? Perhaps… We can motivate this with some pictures from elemental analyses of bullet lead (see Figure 3.3) and refractive indices of glass (see Figure 3.4).

A conceptual discussion of the likelihood ratio approach is described by Aitken and Lucy (2004). They take y and x to be trace element concentrations for a single element (or multiple elements) from several glass fragments at the scene (y) and on the subject (x). Then, they assume a normal distribution for trace element concentrations, though this assumption may be more reasonable for the natural log values of the concentrations. In this setup, under the same source hypothesis S, x and y are two sets of measurements from a single source (i.e. from a single normal distribution). But, under the different source hypothesis, $\overline{S}$, x and y are sets of measurements from two different sources (i.e. from two different normal distributions drawn from the relevant population of possible sources).

It is then possible compute a likelihood ratio in this scenario if we have information about:

1. The ______________ of repeated measurements from a ______________ source.

2. The variability among _________ measurements from _______________ sources in the population of interest.

3. The average or typical measurement for random sources in the population of interest.

Aitken and Lucy's key findings were:

1. LR is small if y and x are very different (i.e. different source hypothesis)
2. LR is big if y and x are similar and y is unusual for the population of interest (i.e. they are indistinguishable on an unusual value)
3. Their examples find typical LRs in 100s or 1000s.
4. Aitken and Lucy also show how to take the likelihood ratio approach without the strong normal (or lognormal) distribution assumptions.

In conclusion, the LR approach can work for trace evidence when

1. there is a well-defined set of ________________ (e.g. chemical concentrations)

2. there is plausible probability models to describe __________________ within a sample (e.g. normal distribution or less restrictive models)

3. it is possible to sample from a population (e.g. other windows) to assess variation across different sources.

This can be and has been done:

- In Aitken and Lucy (2004) with glass
- In Carriquiry, Daniels, and Stern (2000) et al with bullet lead

But, likelihood ratios can be very sensitive to assumptions that are made, and assessing the relevant "population" is hard, and may vary from case to case.

3.3.4.4 Where it might work: Pattern evidence

Many forensic disciplines are focused on comparing a sample, or ______________, at the crime scene (the "unknown" or "__________________") and a potential source (the "known"). The goal is to assess whether two samples have the _______________ source or two different sources. There are many examples of trace evidence in forensic science (See Figure 3.5):

- Latent print examinations[8]
- Shoe prints[9]
- Tire tracks
- Questioned documents[10]
- Firearms
- Tool marks

[8]Image source: http://science.sciencemag.org/content/309/5736/892/F4
[9]Image source: Smith (2009)
[10]Image source: http://forensicunit.weebly.com/evidence.html

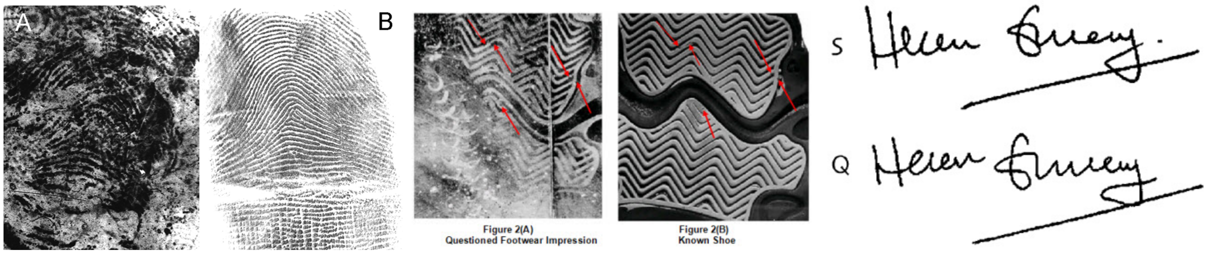

Figure 3.5: Examples of pattern evidence comparisons where there is potential application of the likelihood ratio approach for source determination. From left to right: latent print, shoe print, and questioned signature comparisons.

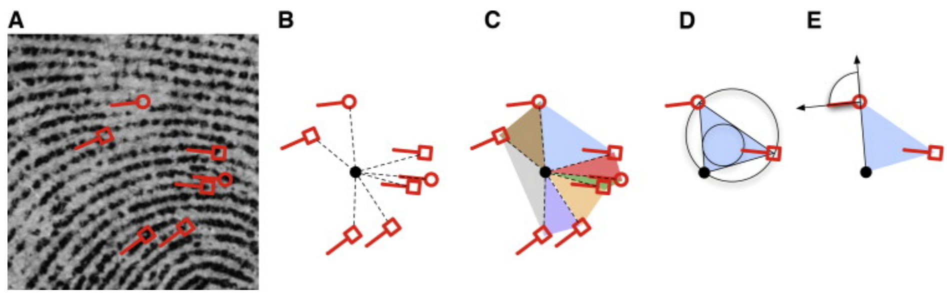

Figure 3.6: 'Extraction of the variables considered by the model from the raw information available on the image of a finger impression. From left to right: (a) annotation of the minutiae on the fingerprint image distinguishing ridge endings (round) and bifurcations (square); (b) definition of the centroid and organization of the minutiae with respect to the centroid; (c) creation of the triangles; (d) extraction of shape variables for one triangle and (e) extraction of the type and direction variables of the minutiae for one triangle (the variables for all triangles are similarly extracted).' (Caption from Neumann et al, p. 158)

There are a number of challenges in constructing likelihood ratios for pattern evidence:

1. The data are very _________ _____________ (often are images).
2. There is a great deal of flexibility (subjectivity?) in defining the numbers or types of _______________ to look at.
3. There is a lack of probability models for ___________________ features or patters.
4. There is still a need to study ________________ across a relevant population.

These challenges are very hard overcome, but there is work underway, including at CSAFE institutions, to create the necessary statistical foundations for pattern evidence.

Consider the latent fingerprint example in Figure 3.6 from Neumann et al. (2015). In the authors' approach, each minutiae is characterized by direction/angle, type of minutiae, and shape/configuration. Neumann et al compute separate likelihood ratios for each of these characteristics, where the ________________ is based on variation within the same finger (obtained from a distortion model) and the ____________________ is based variation across different fingers (using the nearest non-match from a database search).

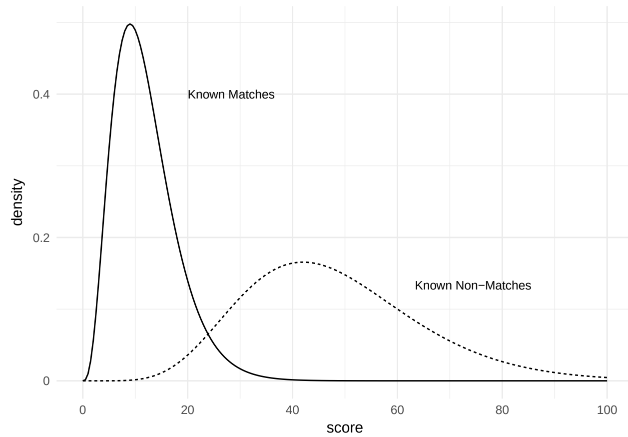

Figure 3.7: For a sample of known matches and known non-matches, the distribution of scores for the different populations.

3.3.4.5 Score-based likelihood ratios

Given the challenge in developing LRs for pattern evidence, there is some interest in a *score-based* approach. With this approach, we define a *score* measuring the "_________________" between the questioned and the known samples. We then obtain the ________________________ of scores for a sample of *known matches*, and we also obtain the distribution of scores for a sample of known ____________________. See Figure 3.7.

The basic idea behind the score-based likelihood approach is outlined as follows:

1. Fit a probability ________________________ to the scores of known matches ($Pr(S|H_p)$)[11]
2. Fit a probability distribution to the scores of known nonmatches ($Pr(S|H_d)$)[12]
3. Calculate the score-based likelihood ratio when we observe score S as $SLR = \frac{Pr(S|H_p)}{Pr(S|H_d)}$. (See Figure 3.8).

An example where the score-based likelihood approach can work for patter evidence is with bullet land signatures, as shown by Hare, Hofmann, and Carriquiry (2017). Hare et all calculated values for many characteristics, such as the number of consecutive matching striae and the value of the cross-correlation function, between two bullet land comparisons. Their empirical score distributions are shown in Figure 3.9 and an example of a comparison they performed is shown in Figure 3.10.

Across a number of existing examples the score distribution for known matches seems relatively straightforward to characterize. There are, however, challenges in defining the relevant *non-match* population:

- Is there a single non-match score distribution?
- Should the non-match score distribution depend on characteristics of the crime scene sample?

Another example of score-based likelihoods comes from the Defense Forensic Science Center's (DFSC)

[11] H_p is the Prosecution's hypothesis, that the defendant is guilty.
[12] H_d is the Defense's hypothesis, that the defendant is not guilty.

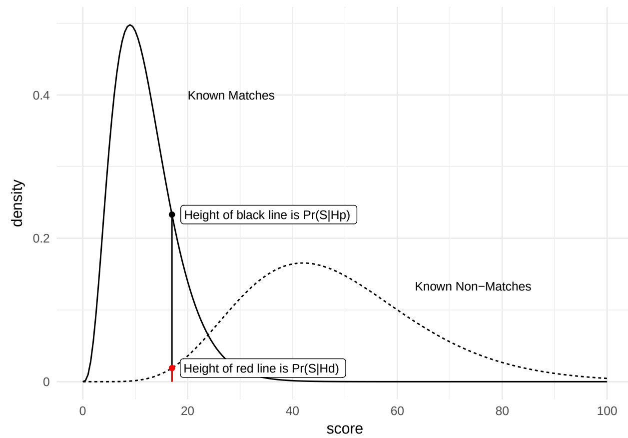

Figure 3.8: The score distributions for known matches and known non-matches with the values in the score-based likelihood ratio calculation shown on the distributions.

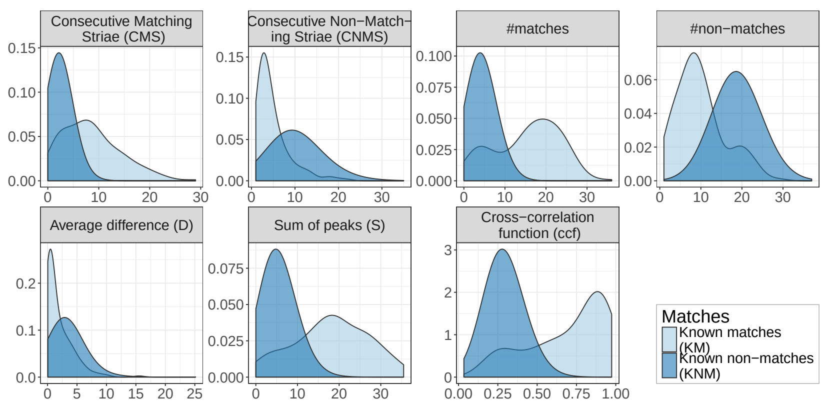

Figure 3.9: Distributions of various "scores" when comparing bullet lands. Known matches in light blue, known non-matches in dark blue. Figure from Hare et al.

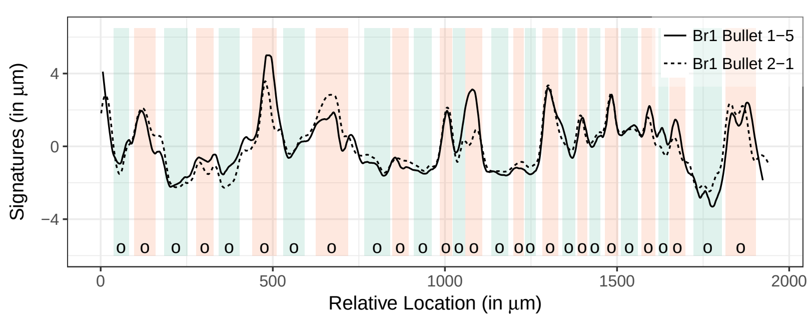

Figure 3.10: An example of a bullet land comparison used to calculate the scores. Figure from Hare et al.

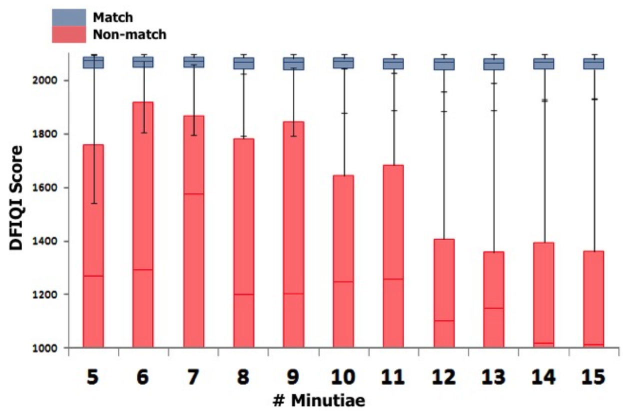

Figure 3.11: The score function values for known matches (blue) and known non-matches (red) by the number of matching minutiae.

evaluation of a latent print score function.[13] Boxplots of score function values for known matches and non-matches by number of minutiae matched are shown in Figure 3.11.

3.3.4.6 Closing thoughts and summary

There is also some discussion in the community about the role of contextual bias, task-relevant/task-irrelevant information, etc. An advantage of the likelihood ratio framework is that it can accommodate this discussion. Let E represent the evidence, S represent the same source, and I represent other information that is being considered. We can condition on this information in the LR:

$$LR = \frac{Pr(E|S, I)}{Pr(E|\overline{S}, I)}$$

The information I should include task-relevant information (e.g. substrate used), but I should *not* include results of other forensic examinations or other case information.

In addition, the likelihood ratio can accomodate multiple types of evidence, say E_1 and E_2:

$$LR = \frac{Pr(E_1, E_2|S)}{Pr(E_1, E_2|\overline{S})}$$

If these evidence types are independent, then the above expression simplifies to:

$$LR = \frac{Pr(E_1|S)}{Pr(E_1|\overline{S})} \cdot \frac{Pr(E_2|S)}{Pr(E_2|\overline{S})}$$

If the evidence types are dependent, however, determining their joint probability (the probability of E_1, E_2 occuring together) can be incredibly difficult.

The ENFSI has released guidelines for evaluative reporting:

- All reporting requires:

[13] Source: https://www.samsi.info/wp-content/uploads/2016/03/SAMSI-2016-Swofford-DFIQI-A__and__C-Combined__HJS__REVISED.ppt

- ________________: need to consider two propositions
- ________________: need to focus on the likelihood of *evidence* given hypothesis, not the other way around
- ________________: should withstand scrutiny
- ________________: there should be a clear case file and report

- Concerns about the Propositions in the report:
  - Level in the hierarchy
  - Absence of an alternative proposition
  - Absence of specified propositions
  - Changing propositions
- Assignment of the likelihood ratio:
  - data and/or expert knowledge used to assign the ____________________ required for likelihood ratio
  - subjective elements can be used
  - avoid undefined ____________________ (e.g., rare)
  - account for _______________________
- The LR then forms the basis for evaluation through verbal equivalents.

Many issues can complicate the calculation of LRs in practice. For example:

- accounting for the transfer process with glass or fibers
- accounting for heterogeneity due to packaging of bullets into boxes
- accounting for usage/lifetime of products (e.g. sneakers)

Though good work is being done, it seems likely that it will be some time before LRs are available for pattern evidence. It is important to remember that there is not one single LR for a given item of evidence. The LR calculation depends on assumptions made and models for the measured data and for the relevant population. Lund and Iyer (Journal of Research of NIST, forthcoming)[14] show that the range of plausible LRs can be extremely wide!

In summary, there are some clear advantages and disadvantages to the likelihood ratio approach. Some advantages are:

- Explicit comparison of the two relevant hypotheses/propositions
- Provide a quantitative summary of the evidence
- There is no need for arbitrary match/non-match decisions when faced with continuous data
- It can accommodate a wide range of factors
- There is enough flexibility to accommodate multiple pieces and multiple types of evidence

Some disadvantages are:

- Requirement of assumptions about distributions

[14] https://www.nist.gov/nist-research-library/journal-research-nist#vol

- The need for reference distributions to define the denominator (although this needs to be done implicitly in any examination).
- It can be difficult to account for all relevant factors
- Unclear how this information should be conveyed to the trier of fact

3.3.5 Forensic Conclusions as Expert Opinion

The previous sections have focused on statistical approaches such as statistical tests and likelihood ratios to make forensic conclusions. The status quo in pattern evidence disciplines, however, does not use such methods. Instead, forensic evidence enters the courtroom through expert testimony. Expert analysis is based on experience, training, and use of accepted methods. There are some issues to consider with this approach:

- What is the range of conclusions reported?
  - identification, inconclusive, exclusion ?
  - multi-point scales: some support, strong support, very strong support, etc ?
- Testifying in this way requires assessing the reliability and validity of the expert opinion (from the PCAST report).

Statistical methods are relevant to carrying out reliabaility and validation studies. Reproducibility and reliability are extremely important to validate the expert methods:

- Reproducibility - how often would the ______________ examiner reach the _______________ conclusion for given evidence

- Reliability - how often would __________________ examiners reach the _______________ conclusion for given evidence

- "White Box" studies - studies of repeatability and reproducibility of different aspects of the forensic examination
- Validation studies
  - "Black Box" studies of performance - examiners given cases with known "ground truth" to assess frequency of different types of errors e.g. Ulery et al. (2011) for latent prints.
  - One way to think about this is that now E = examiner's conclusion, and we need to assess $P(E|S)$ and $P(E|\overline{S})$
  - Study design is extremely important (see Section 2.2).

Reproducibility, reliability, and validity are likely to depend on characteristics of the evidence. For example,

- The quality of latent prints
- The complexity of signature

Ideally such characteristics can be integrated into reliability/validity studies, which would enable reports of the kind "for evidence of this type. . . ."

There will always be unique situations (e.g., did this typewriter produce this note?) for which there are no relevant validation/reliability studies. This is not a problem, but the conclusions expressed by the expert in such settings must acknowledge __________________ about the likelihood of a coincidental agreement.

3.4 Workshop Summary / Conclusions

1. Quantitative analysis of forensic evidence requires some familiarity with concepts from probability and statistics
2. The workshop reviewed basics of probability and statistics
3. Reviewed testing-based approaches and likelihood ratios to forensic examinations
4. We took away some key points:
   a. Any approach must account for the two (or more) competing hypotheses about how the data was generated
   b. Need to be explicit about reasoning and data on which reasoning is based
   c. Need to describe the level of certainty associated with a conclusion
5. There is ongoing discussion about a framework for forensic source conclusions in the OSAC

References

Advisors on Science & Technology, President's Council of. 2016. *Forensic Science in Criminal Courts: Ensuring Scientific Validity of Feature-Comparison Methods*. Executive Office of the President of the United States. https://obamawhitehouse.archives.gov/sites/default/files/microsites/ostp/PCAST/pcast_forensic_science_report_final.pdf.

Aitken, C.G.G, and D. Lucy. 2004. "Evaluation of Trace Evidence in the Form of Multivariate Data." *Applied Statistics* 53 (4): 109–22.

Baldus, David C., Charles Pulaski, and George Woodworth. 1983. "Comparative Review of Death Sentences: An Empirical Study of the Georgia Experience." *Journal of Criminal Law and Criminology*.

Carriquiry, Alicia, Michael Daniels, and Hal Stern. 2000. "Statistical Treatment of Class Evidence: Trace Element Concentrations in Bullet Lead." May. http://www.public.iastate.edu/~alicia/Papers/Forensic%20Statistics/finalreport.pdf.

Committee on Identifying the Needs of the Forensic Sciences Community, and National Research Council. 2009. *Strengthening Forensic Science in the United States : A Path Forward*. Washington: National Academy Press.

Curran, J.M., C.M. Triggs, J.R. Almirall, J.S. Buckleton, and K.A.J. Walsh. 1997. "The Interpretation of Elemental Composition Measurements from Forensic Glass Evidence: I." *Science & Justice* 37 (4): 241–44. doi:http://dx.doi.org/10.1016/S1355-0306(97)72197-X.

Hare, Eric, Heike Hofmann, and Alicia Carriquiry. 2017. "Automatic Matching of Bullet Land Impressions." *The Annals of Applied Statistics* Upcoming.

Morris, Max D. 2011. *Design of Experiments: An Introduction Based on Linear Models*. Chapman; Hall.

Neumann, Cedric, Christophe Champod, Mina Yoo, Thibault Genessay, and Glenn Langenburg. 2015. "Quantifying the Weight of Fingerprint Evidence Through the Spatial Relationship, Directions and Types of Minutiae Observed on Fingermarks." *Forensic Science International* 248: 154–71. doi:http://dx.doi.org/10.1016/j.forsciint.2015.01.007.

Smith, Michael B. 2009. "The Forensic Analysis of Footwear Impression Evidence." *Forensic Science Communications* 11 (3). https://archives.fbi.gov/archives/about-us/lab/forensic-science-communications/fsc/july2009/review/2009_07_review02.htm.

Ulery, Bradford T., R. Austin Hicklin, Joann Buscaglia, Maria Antonia Roberts, and Stephen E. Fienberg.

2011. "Accuracy and Reliability of Forensic Latent Fingerprint Decisions." *Proceedings of the National Academy of Sciences of the United States of America* 108 (19). National Academy of Sciences: 7733–8.