Statistics for working dogs – how do you know if your test/assessments are better than a coin toss?

Arthur E. Dunham[1]
1-Department of Biology, University of Pennsylvania, Philadelphia, PA, USA,
adunham@sas.upenn.edu

Penn
University of Pennsylvania

Key statistical and experimental design concepts for the acquisition, training and certification of working dogs:

- **Terminology and important concepts:**
  - **Errors and error rates, false positives and negatives,**
  - **Sensitivity and specificity,**
  - **Odds and odds ratios.**

- **Guiding Principles:**
  - **One size doesn't fit all,**
  - **Laws of probability,**
  - **Statistical Independence,**
  - **Experimental design – implications for certification.**
  - **Cost of error minimization**

Key statistical and experimental design concepts for the acquisition, training and certification of working dogs:

- **Terminology and important concepts:**
  - **Errors and error rates, false positives and negatives,**
  - **Sensitivity and specificity,**
  - **Odds and odds ratios.**

- **Guiding Principles:**
  - ***One size doesn't fit all,***
  - **Laws of probability,**
  - **Statistical Independence,**
  - **Experimental design – implications for certification.**
  - **Cost of error minimization**

How good does a detection dog have to be?

Guiding Principle I: 'One size doesn't fit all':

- **Different types of training and certification for different types of working dogs:**
  - **Explosive detection**
  - **Drug detection**
  - **Firearms and related detection**
  - **Currency detection**
  - **Human trafficking**
  - **Other contraband**
  - **Assistance dogs**
  - **Other**

- **I will concentrate on detection dogs:**

Key statistical and experimental design concepts for the acquisition, training and certification of working dogs:

- **Terminology and important concepts:**
  - ***Errors and error rates, false positives and negatives,***
  - ***Sensitivity and specificity,***
  - **Odds and odds ratios.**

- **Guiding Principles:**
  - **One size doesn't fit all,**
  - **Laws of probability,**
  - **Statistical Independence,**
  - **Experimental design – implications for certification.**
  - **Cost of error minimization**

Terminology for Detection Dogs

	+/present in real world	-/absent in real world
+/present in dog's opinion	**A**	B – false positive
-/absent in dog's opinion	C – false negative	D

False positive rate = B/(B + D) – ***specific tests*** ***maximize the ability to ID 'blanks' or true negatives*** [D/(B + D)]

False negative rate = C/(A + C) – ***sensitive tests*** ***maximize the ability to ID a target substance*** [A/(A + C)]

Guiding Principle II: The laws of probability -

The probability (*P*) of choosing *K* correct bins (containing targets) from *N* total bins of which only *K* are *targets* and the rest are non-target substances or blanks is given by (*C* = # of combinations):

$$P(N,K) = \frac{1}{C(N,K)}$$

where,

$$C(N,K) = \frac{N!}{(N-K)!\,K!}$$

Allows rigorous interpretation of the results of training trials or certification tests and allows a quantitative assessment of how good a dog really is at his/her job.

- **Most training and certification programs for *detection dogs* permit some false positives and/or negatives. How do you choose how many of each?**
  - ***Guiding Principle V: Cost of Error Minimization:***
  - ***Bomb in checked luggage analogy –***
    - ***False negative error ⇒ plane blows up, many killed BIG COST!***
    - ***False positive error ⇒ isolate and examine suspect bag, possible flight delay SMALL COST!***
  - **So emphasize minimizing *false negative* error rate**
- **So…….How many target odors and how many blanks will you need for any given # of acceptable false positives and false negatives, given the calculated false positive or false negative rate?**
  - **Depends on the type of detection dog being trained / certified?**

Example:

- Constraint 1: The minimum criterion for passing is:
  - false positive error rate less than or equal to 10% ($\alpha \leq 0.10$)
  - false negative error rate of less than or equal to 5% ($\beta \leq 0.05$).

- Constraint 2: The dog can pass with:
  - 2 false positive errors ($\delta = 2$) and
  - 1 false negative error ($\Upsilon = 1$)

- What is the minimum test design that meets all of these criteria simultaneously?

- I. False positive constraint combination: The false positive error rate is less than or equal to 10% ($\alpha = 0.10$) and the dog can make 2 false positive errors ($\delta = 2$) and still pass the test.

- Let Θ = the number of opportunities to make a false positive error during the test.

- To meet the first constraint we require:
- $\delta / \Theta \leq \alpha$
- $[2 / \Theta] \leq 0.1$
- $\Theta = 2 / 0.1 =$ 20 empty boxes. If you tolerate only a 5% ($\alpha = 0.05$) false positive rate, you would need 40 empty boxes

- II. False negative constraint calculation: The false negative error rate is less than or equal to 5% ($\beta \leq 0.05$), and the dog can make 1 false negative error ($\Upsilon = 1$) and still pass.

- Let Ω = the number of opportunities to make a false negative error during the test.

- To meet the second constraint we require:
- $\Upsilon / \Omega \leq \beta$
- $[1 / \Omega] \leq 0.05$
- $\Omega = 1 / 0.05 =$ 20 boxes with target substance

- **The minimally acceptable test design is 40 boxes (20 empty and 20 with target) with target locations randomly assigned on each iteration of the test.**

- **Using these statistical concepts in *truly randomized, completely blinded tests* will enhance any program and provide much greater accuracy and reliability measures for dogs.**

- **Why use completely blinded certification tests? (Much resistance especially in the US.)**
  - **The bad guys don’t normally tell you how many targets you are looking for. So, in the field the task is conducted under blinded conditions. Consequently, certification of dog – handler teams should be completely blinded.**
  - **If the test is not blinded, all you are testing is the handler’s memory and how good he/she is at cueing the dog.**

General model – if you decide on or know 2 of the variables, you can solve for the third

False positives:

- false positive error rate ≤ α
- false positive errors allowed = δ
- number of opportunities to make a false positive error during the test = Θ
- δ / Θ ≤ α ← Put your tolerance for false positive error rate here (e.g., 5%/0.05)

Put # test containers here

Put # false positives errors allowed here

General model – if you decide on or know 2 of the variables, you can solve for the third

False negatives:

- false negative error rate ≤ β
- false negative errors allowed =ϒ
- number of opportunities to make a false negative error during the test = Ω
- ϒ / Ω ≤ β

Put your tolerance for false negative error rate here (eg, 5%/0.05)

Put # test containers here

Put # false negatives errors allowed here

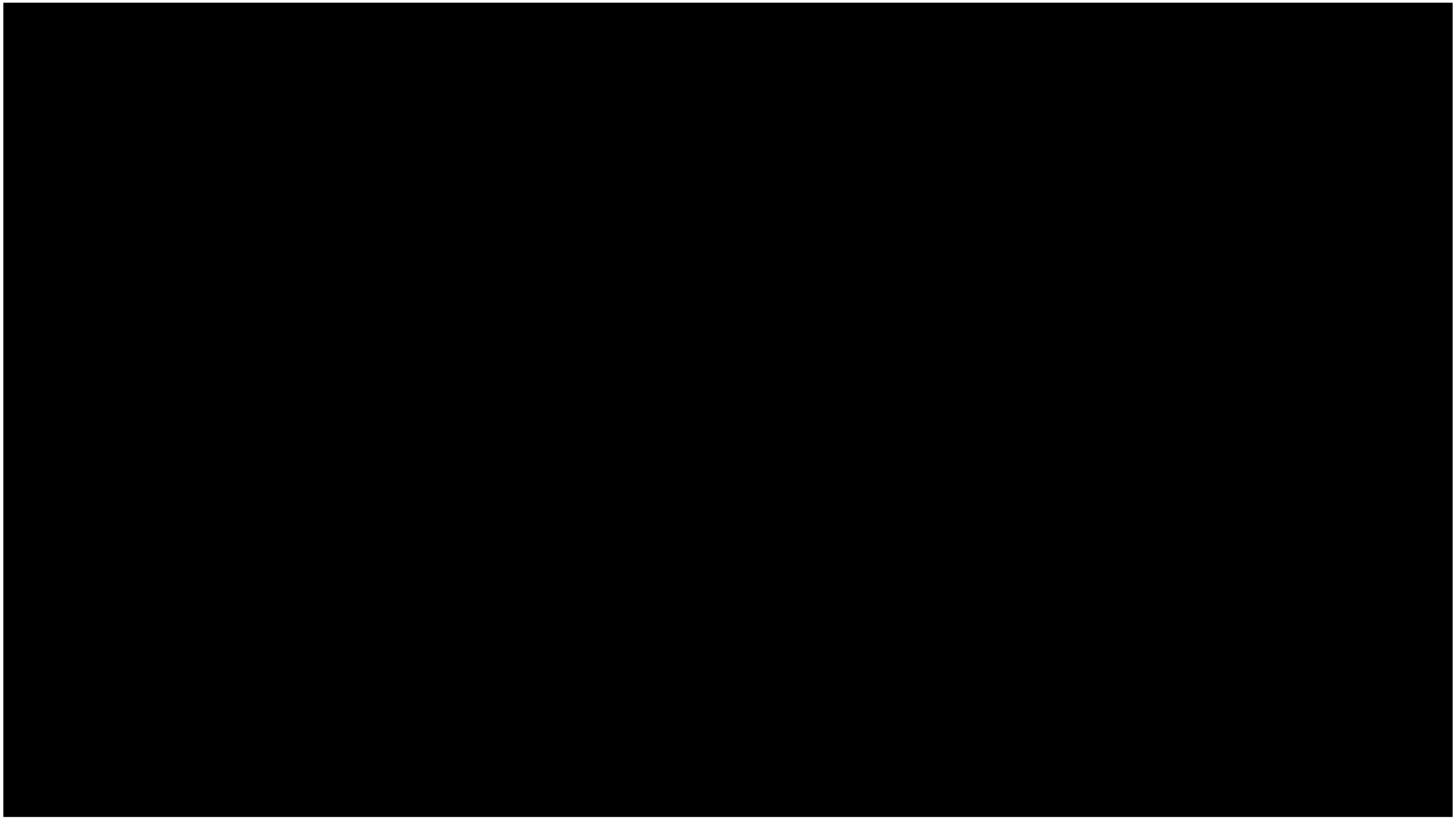
A Very Good Design

- 96 slot wheel
- Stainless steel design and containers food-grade stainless steel jars
- Industrial dishwasher – containers used once; apparatus cleaned each use
- 3 internal standards
- 13 non-target compounds plus empty containers (container odor, only)
- 11 target compounds (some newly being trained and added incrementally over 17 trials)
- Randomized container assignment using a random number table
- Randomized location of wheel by spinning
- Randomized start quadrant
- One way glass
- 2 way radio
- *Completely double blinded* – separate tasks/people and handler signal as only indication, with tester blind to any info except the start quadrant and handler signal

Penn
University of Pennsylvania

TOLL CUSTOMS

Tullverket



The probability (*P*) of choosing *K* correct bins (containing targets) from *N* total bins of which only *K* are *targets* and the rest are non-target substances or blanks is given by (*C* = # of combinations):

$$P(N,K) = \frac{1}{C(N,K)}$$

where,

$$C(N,K) = \frac{N!}{(N-K)!\,K!}$$

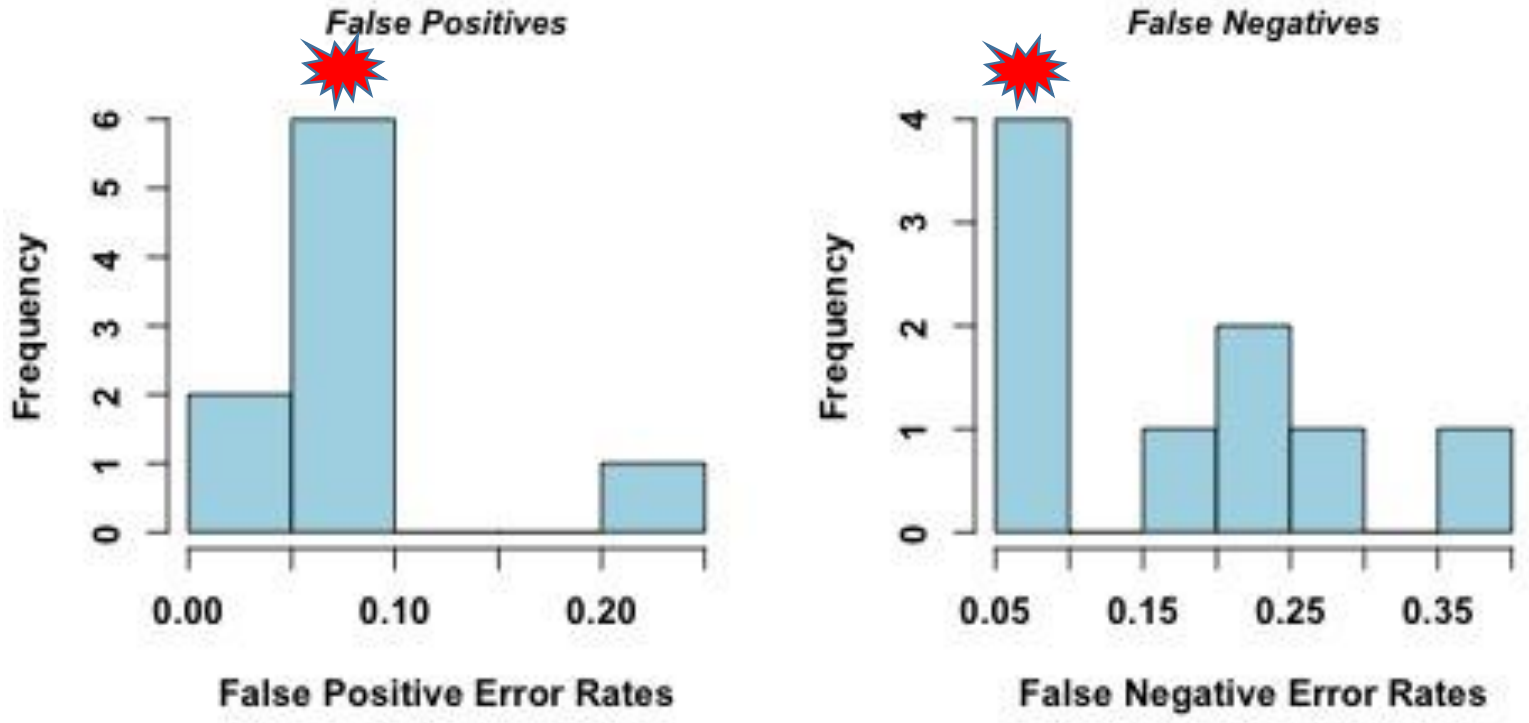
For 3 targets in a total of 96 bins:

$$C(96,3) = \frac{96!}{(93!)3!} = 142{,}880$$

$$P(96,3) = \frac{1}{142880} = 6.99888 \times 10^{-5}$$

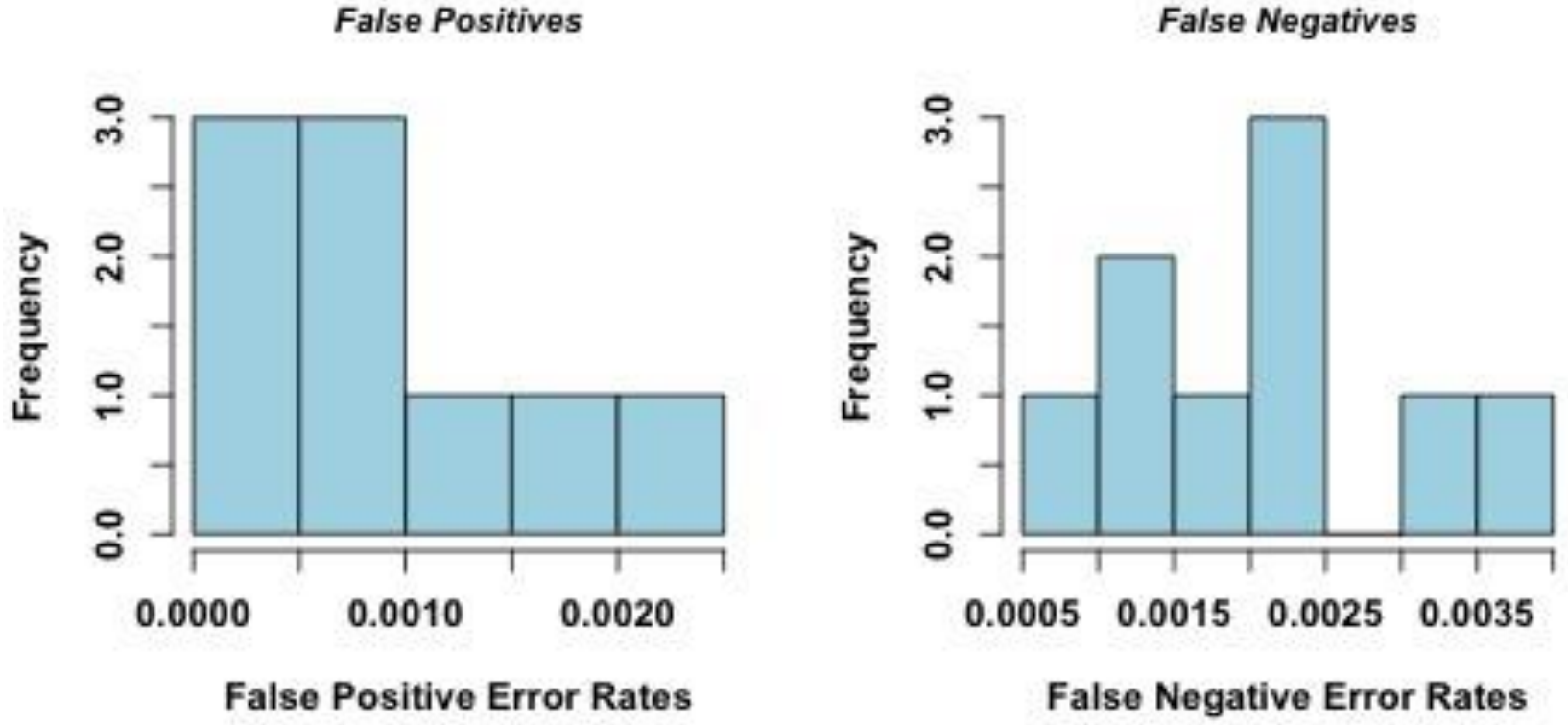
For example, in a test like the one you just saw one dog, Zara, correctly identified all 3 targets and had no false positives or false negatives. This is without the handler or anyone who could communicate with the handler knowing how many targets there were. The probability of this amazing performance being random is less than 1 in 6 million!

Error rates for each of 9 dogs over 17 trials, each of which consisted of 96 bins to be searched:



These rates can be interpreted as the probability that a dog will make a given type of error in one 96 bin trial.

Per bin **error rates for each of 9 dogs over 17 trials, each of which consisted of 96 bins to be searched**



These rates can be interpreted as the probability that a dog will make a given type of error on a given bin.

Conclusions – what this type of study can offer operational people:

- The use of such statistical designs, and the infrastructure that makes them possible, tells us what dogs know, what they don't know and where we need improvement. These studies also tell us about which dogs are learning and how fast they do so (and so…perhaps who to breed).

- Such designs teach handlers to watch and understand their dogs.

- Such designs give organizations confidence that their teams perform as demanded and promised, while identifying weaknesses to be addressed.

- The implementation of this type of strategy can keep us all safer, and allow handlers to do there jobs better, with maximal credibility.

Individual Dog Means over 17 Trials		Mean per bin Error Rates over 17 Trials	
r	P	r	P
0.0607	0.88 NS	-0.012	0.9754 NS

Within dog, there is no association between the probability of any dog making a false positive error and a false negative error for either measure of error. The error types are independent.

Questions?