Poultry Science

Partition of variation for predicting experimental power with a broiler chicken example

G. M. Pesti ,[*,†,1] L. Billard,[†] S.-B. Wu,[‡] N. K. Morgan,[§] P. S. Taylor,[‡] and S. de las Heras-Saldana[#]

[*]*The Poultry Hub Australia, CJ Hawkins Homestead, University of New England, Armidale, New South Wales, 2351, Australia;* [†]*University of Georgia, Athens, Georgia, 30602, USA;* [‡]*School of Environmental and Rural Science, University of New England, Armidale, New South Wales, 2351, Australia;* [§]*Curtin University, School of Molecular and Life Sciences, Bentley, Western Australia, 6152, Australia; and* [#]*AGBU, NSW Department of Primary Industries and University of New England, Armidale, New South Wales, 2351, Australia*

ABSTRACT A 1932 editorial in *Poultry Science* stated that sampling theory, or experimental power, could be useful for "the investigator to know how many … birds to put into each experimental pen." Nevertheless, in the past 90 yr, appropriate experimental power estimates have rarely been applied to research with poultry. To estimate the overall variation and appropriate use of resources with animals in pens, a nested analysis should be conducted. Bird-to-bird and separate pen-to-pen variances were separated for 2 datasets, one from Australia and one from North America. The implications of using variances for birds per pen and pens per treatments are detailed. With 5 pens per treatment, increasing birds per pen from 2 to 4 decreased the SD from 183 to 154, but increasing birds/pen from 100 to 200 only decreased the SD from 70 to 60. With 15 birds per treatment, increasing pens/treatment from 2 to 3 decreased SD from 140 to 126, but increasing pens/treatment from 11 to 12 only decreased the SD from 91 to 89. Choosing the number of birds to include in any study should be based on expectations from historical data and the amount of risk investigators are prepared to accept. Too little replication will not allow relatively small differences to be detected. On the other hand, too much replication is wasteful in terms of birds and resources, and violates the fundamental principles of the ethical use of animals in research. Two general conclusions can be made from this analysis. First, it is very difficult to detect 1% to 3% differences in broiler chicken body weight with only one experiment consistently because of inherent genetic variability. Second, increasing either birds per pen or pens per treatment decreased the SD in a diminishing returns fashion. The example presented here is body weight, of primary importance to production agriculture, but it is applicable whenever a nested design is used (multiple samples from the same bird or tissue, etc.).

Key words: experimental design, experimental power, nested design, ethical animal use

2023 Poultry Science 102:102698
https://doi.org/10.1016/j.psj.2023.102698

INTRODUCTION

Institutional Animal Ethics Committees are entrusted by the public to approve the use of animals in teaching and research only when it is deemed ethical, humane and responsible (e.g., Australian Government, 2013; Rose and Grant, 2013). In most countries, legislation dictates that researchers are required to justify their use of animals in scientific research, including the number of experimental animals (Ibrahim, 2006). Typically, a power analysis is performed to calculate experimental power and justify the use of animals in proposed experiments. The chance of determining a given response difference in a future experiment is called experimental power. In the case of poultry, the choice of the number of birds in an experiment usually involves the number of pens to use for each treatment and the number of birds to put in each pen. This is particularly true for nutrition and behavior studies. Because of genetic diversity between birds and environmental differences between pens, experimental conclusions are always based on probabilities. To estimate the overall variation and appropriate use of resources with animals in pens, a nested analysis should be conducted (Krzywinski et al., 2014). Nested designs are "A class of experimental design in which every level of a given factor appears with only a single level of any other factor. Factors which are not nested are said to be crossed. If every level of one appears

Received February 18, 2023.
Accepted April 3, 2023.

[1]Corresponding author: gpesti@uga.edu

with every level of the others, the factors are said to be completely crossed: if not, they are partly crossed" (Marriott, 2002).

The objective of planning experiments should be to have adequate numbers of birds to ensure a high probability of finding real differences, without using excessive or unethical amounts of resources, be they birds or money. An editorial in *Poultry Science* (Hays, 1932) summarized statistical analytical techniques that could be useful in research with poultry: "Among the most useful applications of biometrics to poultry research may be mentioned: 1. The theory of sampling which enables the investigator to know how many and what kind of birds to put into each experimental pen." However, as Roush and Tozer (2004) observed: "With some exceptions, the power of tests is rarely formally considered or mentioned in poultry research." Upon searching literature in poultry research following 2004, it is evident that the situation has not changed significantly; there is scarce use of test power in poultry research, and a lack of detail presented when it is used (Sadurni et al., 2022). The important pieces of information needed to predict experimental power for a future experiment are the expected means and standard deviations from past experiments. The terms that need to be added to the Schroedek and Lawrence (1932) analysis of variance (**ANOVA**) are the variances due to the birds within a pen (the genetic variation) and the variances between the physical pens themselves. Pen-to-pen environmental variation can result from differences in ventilation within a house, lack of lighting uniformity, differences in noise, humidity, and arbitrary human disruptions.

It is always important to assess if the experiment is relevant to the intended application. Birds kept solely indoors usually have decreased exposure to many stressors, but more exposure to coccidiosis due to oocyst build up in the litter. Birds with outside access are more likely to be exposed to a variety of climatic conditions and other stressors: These stressors include predators and any number of diseases due to contact with wild birds and their excrements. Experiments conducted with more controlled conditions are more repeatable. Is an experiment under closely controlled (inside) conditions relevant to birds grown with outside access and subject to a variety of uncontrolled conditions and stressors? No, and this question raises another: Is there value in conducting an experiment with birds with access to the outside with uncontrolled conditions that is not likely to be repeated? If the experiment is not strictly repeatable, how can its value (validity) be assessed? If the outcomes of subsequent experiments are to be repeated from preliminary ones, great care must be taken to assure that the preliminary experiments' conditions are consistent with the application of the intended research.

Since the very beginning of trials with poultry to compare different feeds, there has been an interest in the statistical interpretation of experimental results (Parker, 1925), and in determining the optimum number of birds required to find significant differences (Schroedek and Lawrence, 1932). Schroedek and Lawrence (1932) demonstrated how to calculate ANOVA for results when males and females were kept in the same pens. The ANOVA was based on individual variation within a single pen per treatment. They used paired t tests for individual mean separation between 4 dietary treatments, the same procedure used currently with Proc LSMeans of the Statistical Analysis System (SAS, 2012). Schroedek and Lawrence (1932) emphasized the need to keep birds under identical conditions, presenting pictures of seemingly identical pens with identical sunporches. At that time, physical separation of birds on different treatments, whereas ensuring pen environments were as similar as possible, was considered adequate. The practice of keeping birds in replicate pens that are randomized, and including this information in the ANOVA, did not become common practice for several decades.

In the early 1930s, the concept of experimental power was brand new (Neyman and Pearson, 1928, 1933). Titus and Hammond (1935) published the first paper on power analyses for poultry experiments. Their discussion centers around the reasons experimental results are often not repeatable: 1) the variability of feed ingredients making replicating diets nearly impossible, and 2) an insufficient number of individuals used in a trial. They believed that it was necessary to have enough individuals in each treatment for the frequency plot of the data to appear normal. These conclusions were based on outputs from rudimentary simulations, many of which were insightful for the time and quite correct: "In a very general way, the accuracy of the results tends to increase as the square root of the number of individuals."

The basic concepts needed to estimate treatment replication before conducting an experiment were detailed in *Poultry Science* (Demetrio et al., 2013). The expected variation in measured responses (e.g., growth, feed utilization efficiency) between experimental units (tissues, individuals, pens of individuals, etc.) is used to estimate experimental power. Figure 1 is from a Microsoft Excel application, where the user can input the mean and standard deviation expected for a future experiment and calculate what the expected power would be with different numbers of replications (Pesti et al., 2018). This example is from an experiment with 3 treatments and 7 pens of five 36-day-old broilers per treatment (Supplemental material). The standard deviation (**SD**) is based on pen means. The usual way to express experimental power is the number of replicates necessary to detect a real 5% difference 80% of the time (while declaring a false significant difference no more than 5% of the time). Such representation of experimental power may be misleading since it only represents one point of the sigmoidal line for each number of replicates. In this example, 20 replicate pens would be necessary to find a 10% difference in 9 of 10 identical experiments (the orange line in Figure 1 crosses the 0.9 horizontal line just below 10% on the horizontal axis). From the graph, it can be seen that: 1) it is practically impossible to detect a 5% difference with such a mean and SD from only 1 experiment more than 40% of the time; 2) the effect of increasing replication is

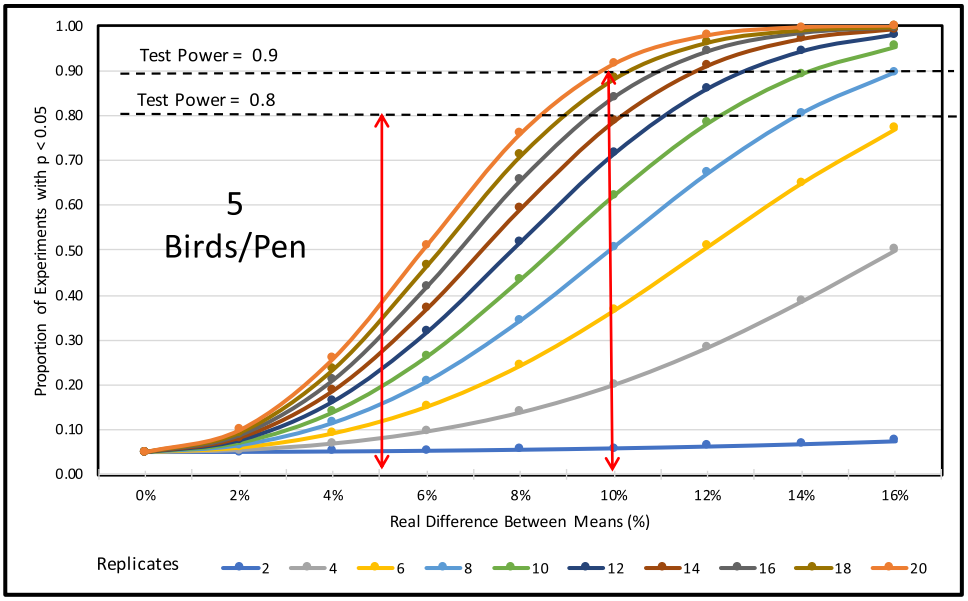

Figure 1. The proportions of experiments expected to have significant differences with different numbers of replicates and different real differences between 2 means. The historical mean = 2,353 and bird-to-bird standard deviation within a pen = 217.

a diminishing-returns phenomena; and 3) with 14 to 20 replicates and a real difference of 10%, a significant difference could be expected almost every time.

In explaining basic power concepts, Demetrio et al. (2013) wrote: "Determining the sample size is complicated because it involves 2 sources of uncontrolled variation: i. between-pen variation, σ_M^2, and ii. between-bird within-pen variation, σ_S^2, and requires a guess of the values of these 2 variances." The guess is only required if individual bird responses are not measured or no other estimate of bird-to-bird variation is available. The difficulty in estimating the 2 variances is due to the fact that birds are most often measured together as group within a pen, so the bird-to-bird variation is not commonly known. Similarly, when birds are sub-sampled the pen mean is typically used as the experimental unit, so variation between individual birds is rarely known. A limitation of the results in Figure 1 is that the SD was only based on pen means. To estimate the overall variation and apply appropriate use of resources with animals in pens, a nested analysis should be conducted (Krzywinski et al., 2014). The observed variation in the presented example is among pens containing fixed numbers of birds, and so contains both sources of variation, as explained by Demetrio et al. (2013). That is appropriate for comparing treatment means from past experiments, but not for estimating variance for future experiments, in which there is possibility of changing both pens/treatment and birds/pen.

The objective of this paper is to demonstrate how to partition variances into bird-to-bird within pen (genetic) and pen-to-pen (micro-environmental) sources (Figure 2). Data from 2 experiments with growing broilers is used to show the practical application of the results. Growth, or body weight, was used in the example as it is the most important attribute for production agriculture. The principles apply for any experiment where birds are kept in pens, or multiple samples are taken from the same bird or tissue, etc. (nested design).

MATERIALS AND METHODS

The first example dataset was derived from the Rob Cumming Poultry Innovation Centre at The University of New England (Armidale, Australia). Broiler chickens were raised from hatch to 35 d of age in 21 pens. There were 3 treatments (2 therapeutic agents and a control) with a 1-way nested design. There were 7 replicate pens of 10 birds each per treatment. Final body weight was the response variable investigated. By 35 d of age, the pens contained different numbers of males, due to sexing errors at placement and random mortality. To simplify this example, the first 5 male broilers from each pen were chosen on the assumption that they were randomly recorded and thus remained random (210 birds total).

The second example dataset is from J-House at the University of Georgia Poultry Research Center (Da Costa et al., 2017). Broiler chickens were raised from hatching to 48 d of age in 48 pens (28 broilers per pen, nested in pens at hatching). There were 4 treatments with a 2 × 2 factorial nested design (2 genetic strains and males vs. females). Responses of the 2 strains were very similar and strain differences were not considered in this analysis. Body weight at 12, 17, 25, 32, 41, and 48 d of age was the response variable investigated. For hypothesis testing this experiment should be analyzed with a repeated measures design. Our purpose was to estimate variances at each age, so the data were analyzed separately for each time.

The response model is:

$$\text{Response}\,(Y) = \mu + T_i + P_{j(i)} + \varepsilon_{ijk},$$

$$k = 1, \ldots, b,\ j = 1, \ldots, p,\ i = 1, \ldots t,$$

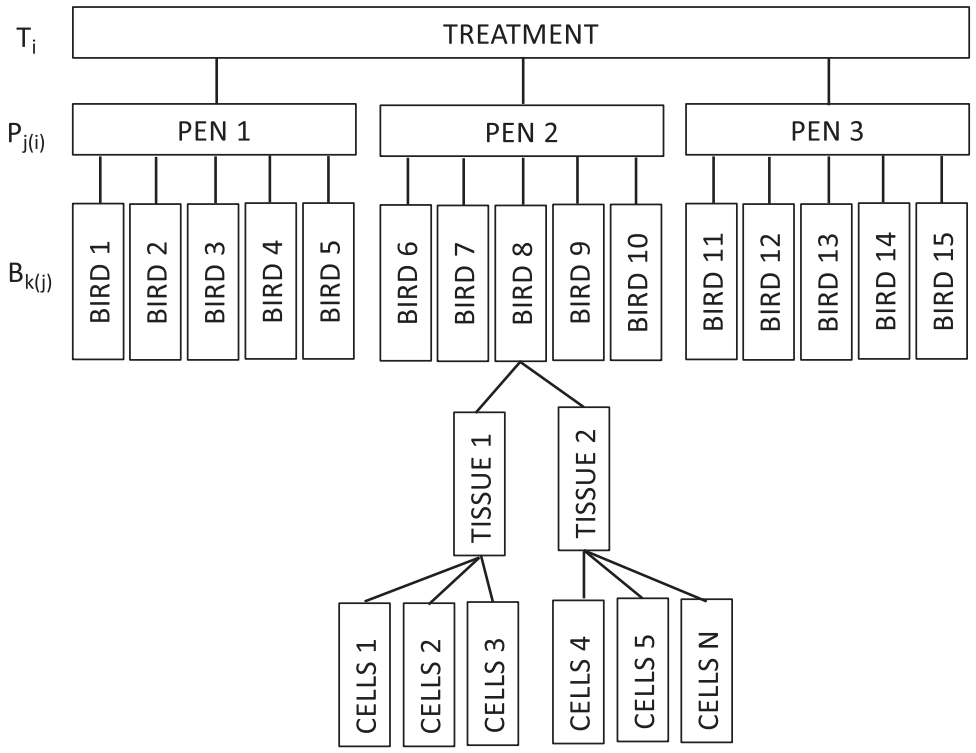

Figure 2. A diagram showing nesting of birds within pens and cells nested within tissues, nested within pens.

where $Y_{ijk} = k^{\text{th}}$ observation from the j^{th} bird in the i^{th} pen,

$T_i \sim N(0,{\sigma_2}^2)$

$P_{j(i)} \sim N(0,{\sigma_1}^2)$-Random Model

$\varepsilon_{ijk} \sim N(0,\sigma^2)$ i.e., ${\sigma_2}^2$ is the variation between treatments T_i,

${\sigma_1}^2$ is the variation between pens within treatments $P_{j(i)}$,

σ^2 is the experimental error; or, effectively, T_i is the effect of the i^{th} treatment and $P_{j(i)}$ is the effect of the j^{th} pen within the i^{th} treatment, and ε_{ijk} is the observed error within replicate pens (between birds).

Then the total sum of squares $= \sum\sum\sum (Y_{ijk} - \overline{Y})^2$ = Between treatments SS + between pen within treatment SS + residual SS (Table 1) where the between treatment SS $= bp \sum_{i=1}^{t} (\overline{Y}_{i\cdot\cdot} - \overline{Y})^2 = \sum_{i\cdot\cdot} \frac{Y_{i\cdot\cdot}^2}{bp} - \text{CF} = \sum_i \frac{T_i^2}{bp} - CF$ with $\text{CF} = (\sum\sum\sum Y_{ijk})^2/tpb$.

Between pens within treatment SS = b $\sum_i \sum_j (\overline{Y}_{ij\cdot} - \overline{Y}_{i\cdot\cdot})^2$

$$= \sum_i \sum_j \frac{P_{ij}^2}{b} - \sum_i \frac{T_i^2}{bp}$$

$$= \sum_i \sum_j P_{ij}^2 - \sum_i \frac{T_i^2}{bp}$$

$$= \sum_i \sum_j \frac{P_{ij}^2}{b} - CF - \text{between treatment SS}$$

where $P_{ij} = j^{\text{th}}$ pen total in i^{th} treatment $= Y_{ij\cdot}$,

T_i = total of i^{th} treatment = $Y_{i\cdot\cdot}$,and residual SS $= \sum_i \sum_j \sum_k (Y_{ijk} - \overline{Y}_{ij\cdot})^2$.

The residual SS can also be thought of as between birds.

It can be shown that:

$$\text{Between pens SS} = \text{b} \sum_i \sum_j (\overline{Y}_{ij} - \overline{Y})^2$$
$$= \sum_i \sum_j (\overline{Y}_{ij\cdot} - \overline{Y}_{i\cdot\cdot})^2 + \sum_i (\overline{Y}_{i\cdot\cdot} - \overline{Y})^2$$
$$= \text{Between pens within treatments SS} + \text{between treatments SS.}$$

Table 1. Analysis of variance table for partitioning variances between pens and birds.

Source	df	SS		MS
Between treatments	$t-1$	$\sum_i \frac{T_i^2}{pb} - CF$	j	j
Between pens within treatments	$t(p-1)$	$\sum_i \sum_j \frac{P_{ij}^2}{pb} - \sum_i \frac{T_i^2}{pb}$	k	=m-j
Between pens	$tp-1$	$\sum_i \sum_j \frac{T_{ij}^2}{pb} - CF$		m
Residual	$tp(b-1)$	Difference	l	
Total	tpb	$\sum\sum\sum {Y_{ijk}}^2$ - CF		

Between pens mean square (**MS**) was calculated from the between treatments MS (${\sigma^2} + b{\sigma_1}^2 + bp{\sigma_2}^2$) minus the between pens within treatments MS (${\sigma^2} + b{\sigma_1}^2$).

Further, SDs for future experiments were estimated as follows: $\sqrt{(((\text{MS}_{\text{P2P}} \times \sqrt{p_h})/\sqrt{p_f}) + \sqrt{((\text{MS}_{\text{B2B}} \times \sqrt{p_h})/(\sqrt{p_f} \times b_f))}}$, where p = number of pens/treatment, b = number of birds/pen, MS = mean square, P2P = pen-to-pen, B2B = bird-to-bird, subscripts h and f indicate from historical ($\boldsymbol{h}$) and future ($\boldsymbol{f}$) experiments. Sample size to detect a given difference between 2 means was estimated by Lehr's method (Lehr, 1992; van Belle, 2008).

Table 2. Analysis of variance table showing calculation of the variation from individuals nested within pens for an experiment with a 1-way ANOVA table showing calculation of the variation from individuals nested within pens for an experiment with a 1-way design, 3 treatments, 7 pens per treatment and 5 birds per pen.

Source	df	SS	MS	E(MS)
Between treatments	2	796,057.62	398,028.81	$\sigma^2 + b\sigma_1{}^2 + bp\sigma_2{}^2$
Between pens within treatments	18	864,397.14	48,022.06	$\sigma^2 + b\sigma_1{}^2$
Between pens	20	1,660,454.76	83,022.74	
Between birds	84	3,957,170.00	47,109.17	σ^2
Total SS	104	5,617,624.76	54,015.62	

The response variable is 35-day body weight.

RESULTS

The MS for between pen variation within treatments in the first dataset was found to be: (MS between pens within treatments − MS between birds)/(birds/pen) = 182.578 (Table 2). The between pen and between bird MSs were used to create Table 3. Table 3 demonstrates the relative importance of the number of pens and the number of birds per pen in this particular facility.

Increasing either birds per pen or pens per treatment decreased the SD in a diminishing returns fashion (Table 3, Figure 3). With 5 pens per treatment, increasing birds per pen from 2 to 4 decreased the SD from 183 to 154, but increasing birds/pen from 100 to 200 only decreased the SD from 70 to 60. With 15 birds per treatment, increasing pens/treatment from 2 to 3 decreased SD from 140 to 126, but increasing pens/treatment from 11 to 12 only decreased the SD from 91 to 89. More than approximately 5 pens/treatment gave relatively little difference between SDs, and the total number of birds was the most important factor in reducing SDs. There was little difference in SDs from 5 pens of 100 or 10 pens of 50 (70.19 vs. 69.74).

The same phenomenon in diminishing returns for increasing birds/pen and pens/treatment was observed for the second facility, where measurements from a larger experiment were taken over time (Table 4, Figures 5 and 6). Body weights and predicted SDs increased over time, with males exhibiting higher levels of both. The slopes of the lines depicting the effects of using the standard method of analyzing pen means vs. individuals nested within pen means are quite different

Table 3. The influence of observed bird-to-bird mean squares (MS_{B2B}) and pen-to-pen (MS_{P2P}) variations on the predicted variation (SD_{TOTAL}) and total numbers of birds per treatment for future experiments.

			Pens per Treatment											
			1	2	3	4	5	6	7	8	9	10	11	12
Birds /Pen	MS_{B2B}	MS_{P2P} ...	69.01	48.80	39.84	34.50	30.86	28.17	26.08	24.40	23.00	21.82	20.81	19.92
1	105,339	SD_{TOTAL}	8.31	6.99	6.31	5.87	5.56	5.31	5.11	4.94	4.80	4.67	4.56	4.46
		Birds	324.67	273.01	246.69	229.57	217.12	207.44	199.60	193.05	187.45	182.57	178.27	174.44
		Samples	*1*	*2*	*3*	*4*	*5*	*6*	*7*	*8*	*9*	*10*	*11*	*12*
2	74,486	SD_{TOTAL}	122	86	70	61	54	50	46	43	41	39	37	35
		Birds	273.05	229.60	207.47	193.07	182.60	174.46	167.87	162.36	157.64	153.55	149.93	146.70
		Samples	*2*	*4*	*6*	*8*	*10*	*12*	*14*	*16*	*18*	*20*	*22*	*24*
4	52,670	SD_{TOTAL}	86	61	50	43	39	35	33	30	29	27	26	25
		Birds	229.65	193.11	174.50	162.39	153.58	146.73	141.19	136.55	132.59	129.14	126.10	123.39
		Samples	*4*	*8*	*12*	*16*	*20*	*24*	*28*	*32*	*36*	*40*	*44*	*48*
8	37,243	SD_{TOTAL}	61	43	35	30	27	25	23	22	20	19	18	18
		Birds	193.16	162.43	146.77	136.59	129.18	123.42	118.75	114.86	111.52	108.62	106.07	103.78
		Samples	*8*	*16*	*24*	*32*	*40*	*48*	*56*	*64*	*72*	*80*	*88*	*96*
10	33,311	SD_{TOTAL}	43	30	25	22	19	18	16	15	14	14	13	12
		Birds	182.70	153.63	138.82	129.19	122.18	116.74	112.32	108.64	105.48	102.74	100.32	98.16
		Samples	10	20	30	40	50	60	70	80	90	100	110	120
15	27,198	SD_{TOTAL}	39	27	22	19	17	16	15	14	13	12	12	11
		Birds	165.13	138.86	125.47	116.76	110.43	105.51	101.52	98.19	95.34	92.86	90.67	88.72
		Samples	*15*	*30*	*45*	*60*	*75*	*90*	*105*	*120*	*135*	*150*	*165*	*180*
20	23,555	SD_{TOTAL}	32	22	18	16	14	13	12	11	11	10	10	9
		Birds	153.70	129.25	116.79	108.68	102.79	98.21	94.49	91.39	88.74	86.43	84.40	82.58
		Samples	*20*	*40*	*60*	*80*	*100*	*120*	*140*	*160*	*180*	*200*	*220*	*240*
25	21,068	SD_{TOTAL}	27	19	16	14	12	11	10	10	9	9	8	8
		Birds	145.39	122.25	110.47	102.80	97.22	92.89	89.38	86.45	83.94	81.76	79.83	78.11
		Samples	*25*	*50*	*75*	*100*	*125*	*150*	*175*	*200*	*225*	*250*	*275*	*300*
30	19,232	SD_{TOTAL}	24	17	14	12	11	10	9	9	8	8	7	7
		Birds	138.93	116.82	105.56	98.24	92.91	88.77	85.41	82.61	80.21	78.13	76.29	74.64
		Samples	*30*	*60*	*90*	*120*	*150*	*180*	*210*	*240*	*270*	*300*	*330*	*360*
50	14,897	SD_{TOTAL}	22	16	13	11	10	9	8	8	7	7	7	6
		Birds	122.34	102.87	92.96	86.51	81.81	78.17	75.21	72.74	70.63	68.79	67.18	65.73
		Samples	*50*	*100*	*150*	*200*	*250*	*300*	*350*	*400*	*450*	*500*	*550*	*600*
100	10,534	SD_{TOTAL}	17	12	10	9	8	7	7	6	6	5	5	5
		Birds	102.97	86.59	78.24	72.81	68.86	65.79	63.31	61.23	59.45	57.90	56.54	55.32
		Samples	*100*	*200*	*300*	*400*	*500*	*600*	*700*	*800*	*900*	*1,000*	*1,100*	*1,200*
200	7,449	SD_{TOTAL}	12	9	7	6	5	5	5	4	4	4	4	4
		Birds	86.70	72.91	65.88	61.31	57.98	55.40	53.30	51.55	50.06	48.76	47.61	46.58
		Samples	*200*	*400*	*600*	*800*	*1,000*	*1,200*	*1,400*	*1,600*	*1,800*	*2,000*	*2,200*	*2,400*

The average body weight was 2,353, SD = 220 g. Sample size (*Samples*) was estimated by Lehr's method (Lehr, 1992) to have an 80% chance of detecting a real 5% difference in body weight ($\text{ß} = 0.20$) and with a 5% chance of declaring a difference significant when none exists ($\alpha = 0.05$).

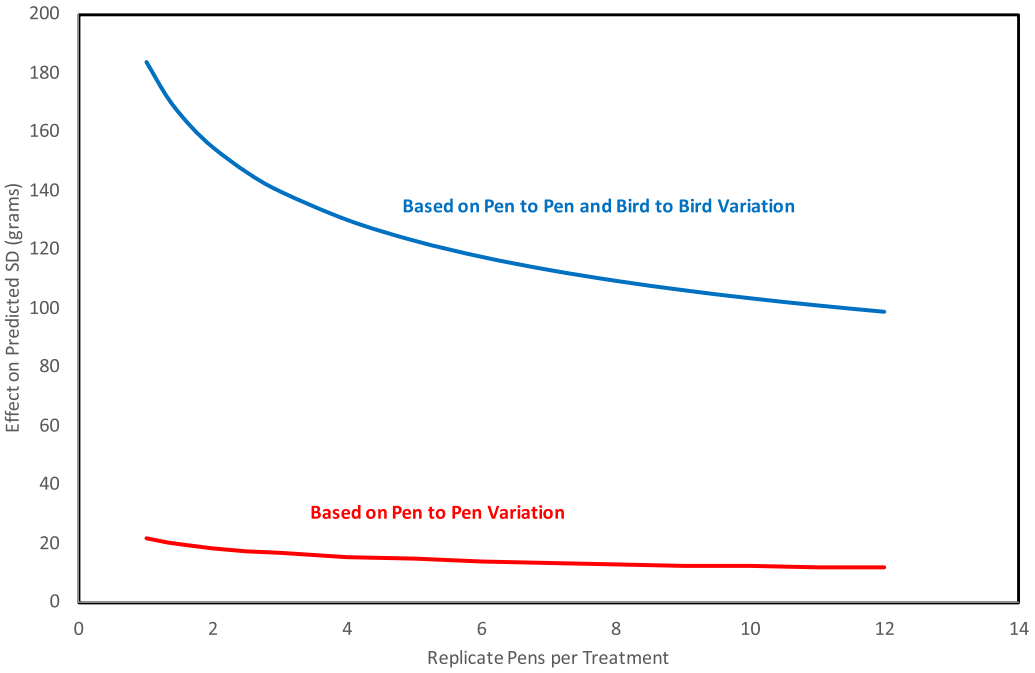

Figure 3. Comparison of the effects of changing the number of replicates when the variances are based only on pen means (pen-to-pen and bird-to-bird variation not partitioned) vs. partitioning variation so that only pen-to-pen variation is considered

(Figure 3). Twenty simulated experiments with an experimental power of 0.8 resulted in probabilities between $P = 0.000$ and $P = 0.125$ (Figure 6). Twenty simulated experiments with an experimental power of 0.8 resulted in probabilities between $P = 0.003$ and $P = 0.717$ (Figure 7).

DISCUSSION

The primary goal of experimental power analyses is to balance the number of experimental units, birds in this case, with the risk of not finding a real difference if one exists, or declaring a significant difference when none exists. The cost of the experiment, both monetary and animal lives, has to be weighed against the value of the expected outcome. The process of estimating experimental power is clearly very complex. Without considering nesting, it is a 4-dimensional mathematical problem with 4 variables: 1) The previously observed mean and SD; 2) The proportion of experiments with $P < 0.05$; 3) The potential number of replicates; and 4) The detectable difference (Figure 1). Partitioning the variance into bird-to-bird and pen-to-pen portions adds an additional dimension to interpret (Figures 4 and 5). Figures 4 and 5 are only 2 of n possible figures with n = number of birds per pen. It would take n such figures to illustrate all the possible choices of the number of birds per pen and pens per treatment and their effects on the probabilities of detecting differences of various sizes. The choice of 25 and 50 birds per pen was not entirely arbitrary. They were chosen to illustrate the relatively small effect of doubling the size of an experiment can have on experimental power. Had lower numbers of birds per pen been chosen, say 2 vs. 4 birds in each of 3 pens per treatment, the number of samples to detect a specified difference would be much greater, 50 vs. 35 (Table 3).

A tabular presentation, like Table 3, may be helpful for comparing the effects of birds per pen and pens per treatment vs. sample size. Such a presentation could be helpful for budgeting purposes, by including costs for birds, pen space, feed and labor, for example. The estimate of sample size ($\boldsymbol{sz}$) in Table 3 is only an estimate of one arbitrarily chosen point on the lines presented in Figures 1 to 3, presenting an 80% chance of detecting a real 5% difference in body weight ($ß = 0.20$), with a 5% chance of declaring a difference significant when none exists ($\alpha = 0.05$). For practical purposes, it may be prudent to consider some arbitrary difference ($\boldsymbol{d}$) that could be detected some proportion of the time with some alpha and beta errors for each cell. After such an initial screening, plots of probability lines (Figures 4 and 5) vs. costs in money or birds could then be considered.

For the predicted experimental power illustrated in Figures 3 and 4, the mean squares for the number of birds per pen were changed independently of the number of pens, conversely, theoretically resulting in an increase in accuracy vs. simply the pen mean approach in Figure 1. The question is: by how much? This is answered by the slope of the lines presented in Figure 3, which shows the magnitude of the differences in the 2 approaches for this example. This demonstrates that the 2 approaches lead to different numbers of replicates being proposed. In this example, the effects on predicted variation were much greater with fewer than 5 or 6 pens/treatment compared with more than 8 pens/treatment.

The same general patterns were found with the second dataset over time (Table 4 vs. Table 3). The pen-to-pen variation was greater at similar ages for the experiment

Table 4. Descriptive statistics and prediction of SD for use in estimating experimental power for future experiments.

					Mean square			Predicted SD (pens × birds/pen)					
Age (d)	Pens (p_h)	Birds/pen (b_h)	Avg. BW (grams)	Birds	Pens + birds	Birds (MS_{B2B})	Pens (MS_{P2P})	2 × 10	4 × 10	4 × 25	4 × 50	10 × 50	5 × 100
Females													
0	8	28.0	41	448	6	10	N/A	N/A	N/A	N/A	N/A	N/A	N/A
12	8	27.8	301	444	3,952	1,060	2,892	83	74	52	49	43	40
17	8	27.9	522	446	9,443	2,255	7,188	129	113	80	76	65	61
25	8	27.9	1,156	445	24,123	8,821	15,302	205	185	134	123	110	101
32	8	27.8	1,793	445	62,614	19,968	42,646	331	295	213	198	174	161
41	8	23.9	2,613	382	69,377	43,061	26,316	337	316	239	211	198	175
48	8	19.6	3,260	314	141,024	59,323	81,701	479	432	314	289	258	236
Males													
0	8	28.0	42	448	16	11	N/A	N/A	N/A	N/A	N/A	N/A	N/A
12	8	28.0	306	448	5,698	1,042	4,656	100	87	61	58	49	47
17	8	27.8	541	445	8,447	2,734	5,713	122	109	78	73	64	59
25	8	27.7	1,255	443	33,058	10,959	22,098	240	215	155	144	127	117
32	8	27.3	1,985	437	129,672	26,966	102,705	478	418	294	279	238	225
41	8	22.9	2,613	366	138,271	69,086	69,185	478	438	324	293	267	240
48	8	18.4	3,830	294	309,657	94,863	214,795	716	633	450	423	367	343

N/A: not available, birds were weighed prior to being placed in pens.

Data from an experiment at the University of Georgia's Poultry Research Center (Da Costa et al., 2017).

SDs for future experiments were estimated as follows: $\sqrt{(((MS_{P2P} \times \sqrt{p_h})/\sqrt{p_f}) + \sqrt{((MS_{B2B} \times \sqrt{p_h})/(\sqrt{p_f} \times b_f))}}$, where p = pens/treatment, b = birds/pen, MS = mean square, P2P = pen-to-pen, B2B = bird-to-bird, subscripts h and f indicate from historical and future experiments.

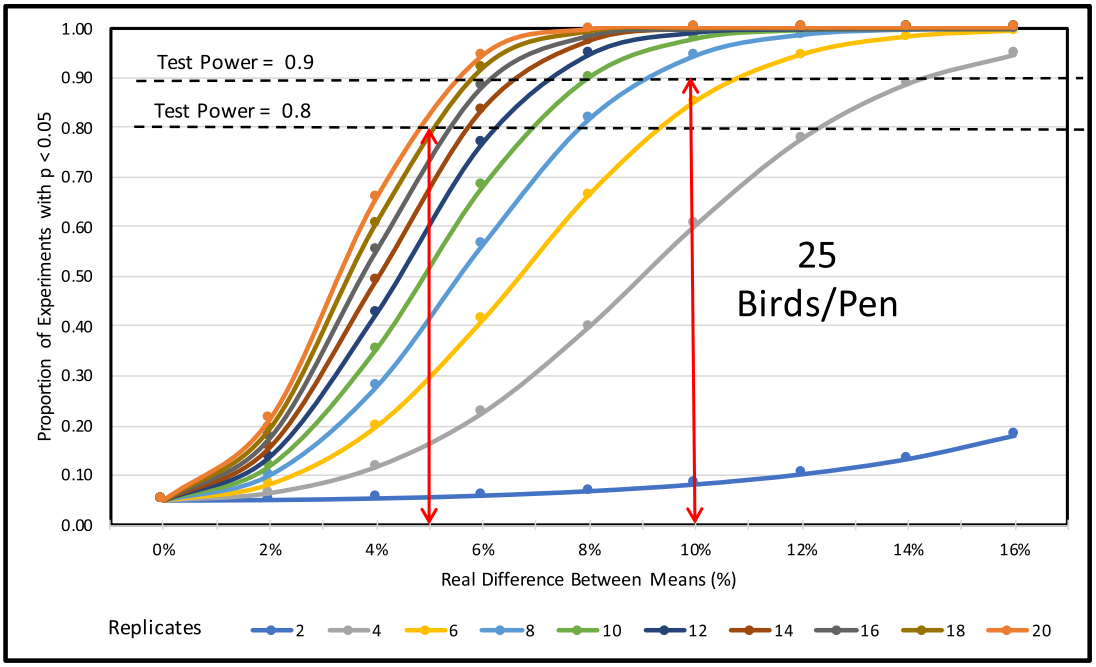

Figure 4. The proportions of experiments expected to have significant differences with different numbers of replicates and different real differences between 2 means. The historical mean = 2,353 and bird-to-bird standard deviation within a pen = 217.

in Table 4 than Table 3. The estimated SDs were very similar for 10 pens of 50 birds vs. 5 pens of 100 birds for both datasets (69.9 vs. 70.3 in Table 3, and 238 vs. 225 for male birds at d 32 in Table 4).

The ethical considerations for the use of animals in research dictate that a minimal number of animals should be used while ensuring the validity of the results (e.g., Australian Government, 2013). The reality of research involving sampling populations of animals with inherent variability is that there is no "minimal number of animals ... to ensure the validity of results." There are only different numbers of birds that lead to different probabilities of declaring results significant. Appropriate experimental designs can be chosen to increase the odds of making suitable statistical inferences, but the conclusions should only be stated in terms of the odds that conclusions are correct, not binary concepts like valid or invalid. Results should never be regarded as valid or invalid, only likely or unlikely to be repeatable to some specified degree or probability. Researchers are always faced with the dilemma of balancing type I vs. type II error; using more animals decreases the chances of

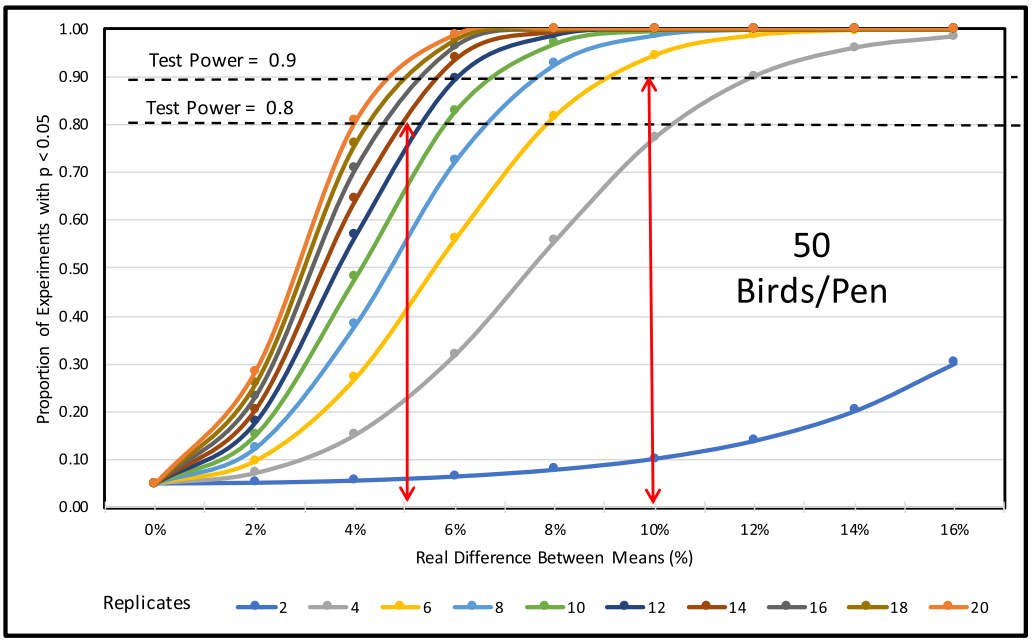

Figure 5. The proportions of experiments expected to have significant differences with different numbers of replicates and different real differences between 2 means. The historical mean = 2,353 and bird-to-bird standard deviation within a pen = 217.

"Experiment" Simulation Number	TREATMENT A						Mean = 2353 SD = 217 CV = 9.22		TREATMENT B						Mean = 2047 SD = 217 CV = 10.60		A vs B t-test	p > 0.05 ?
	Replicate Number								Replicate Number									
	1	2	3	4	5	6	Mean	SEM	1	2	3	4	5	6	Mean	SEM		
1	2619	2383	2490	2360	2606	2030	2415	97	2109	2124	1907	1747	1957	1991	1972	62	0.007	1
2	2231	2071	2454	2464	2312	2626	2360	88	2202	2246	1924	2213	1748	1872	2034	95	0.074	0
3	2501	2694	2328	2175	3014	2244	2493	142	2363	2284	2387	1715	2152	1979	2147	116	0.043	1
4	2545	2616	2152	2324	2287	2582	2417	85	2311	2390	2363	2104	2152	1813	2189	97	0.135	0
5	2384	2269	2145	2193	2545	2283	2303	65	2044	2210	2241	2185	2135	1780	2099	76	0.097	0
6	2645	2214	2503	2725	2598	2380	2511	84	1682	1998	2139	1856	2174	2001	1975	82	0.008	1
7	2560	2771	2171	2408	2550	2466	2488	89	2271	2499	2317	1692	2004	1888	2112	135	0.031	1
8	2447	2471	2476	2361	2366	2570	2448	35	2088	1672	2151	2067	1929	1830	1956	81	0.003	1
9	2290	2470	2267	2310	2118	2655	2352	83	2001	1872	1869	2081	2031	1916	1962	40	0.011	1
10	2136	2277	1990	2131	2524	2098	2193	83	2155	2058	2202	2036	2425	2141	2170	63	0.717	0
11	2444	2375	2592	2546	2384	2531	2479	41	1983	2032	1743	1649	2341	1993	1957	109	0.010	1
12	2278	2230	2529	2215	2490	2600	2390	75	1888	2130	1901	1861	2277	2139	2033	77	0.005	1
13	2420	2461	2385	2459	2227	2119	2345	63	1712	2322	1961	2244	2170	2017	2071	99	0.043	1
14	2209	2305	2544	2306	2357	2054	2296	73	2344	2063	1926	2311	2504	2151	2216	94	0.546	0
15	2335	2609	1996	2386	2404	2281	2335	89	1831	1791	2203	2189	2250	1789	2009	101	0.074	0
16	2268	2411	2342	2622	2302	2532	2413	62	2283	1846	1959	1749	1664	2116	1936	104	0.011	1
17	2339	2484	2510	2210	2190	1957	2281	93	2042	2102	2038	2386	1800	2383	2125	101	0.345	0
18	2527	2363	2461	2326	2115	2835	2438	108	1808	2214	1690	2275	1921	2302	2035	117	0.025	1
19	2308	2388	2180	2477	2361	2343	2343	44	2095	1700	2245	2143	1905	2097	2031	88	0.029	1
20	2514	2071	2770	2082	2818	2427	2447	144	2305	1965	1990	2240	1929	2289	2120	79	0.110	0
					Mean =		2387	82					Mean =		2057	91	0.116	
					SD =		83	28					SD =		86	22	0.190	
					Minimum =		2193	35					Minimum =		1936	40	0.000	
					Maximum =		2511	144					Maximum =		2216	135	0.144	
															Sum/Count = Power =			0.500

Figure 6. Twenty simulated experiments comparing Treatments A and B by Student's t test at $P < 0.05$. The number of simulated "Experiments" with H_0: Treatment A $\neq$ Treatment B were summed to estimate experimental power for experiments with mean body weights of 2,353 and 1,911 g and standard deviations of 217 g (18.8% difference). Responses were simulated with Microsoft Excel's random number generator (Pesti et al., 2018).

"Experiment" Simulation Number	TREATMENT A						Mean = 2353 SD = 217 CV = 9.22		TREATMENT B						Mean = 1911 SD = 217 CV = 11.36		A vs B t-test	p > 0.05 ?
	Replicate Number								Replicate Number									
	1	2	3	4	5	6	Mean	SEM	1	2	3	4	5	6	Mean	SEM		
1	2427	2314	2198	2489	2742	2455	2437	82	2129	1418	1857	2114	2288	1809	1936	139	0.003	1
2	2514	2544	2381	2388	2259	2529	2436	50	1741	1756	1787	1733	1469	1800	1714	55	0.000	1
3	2351	2528	2267	2418	2218	2736	2420	85	2108	2227	1820	1726	2035	1820	1956	88	0.011	1
4	2190	2881	2221	2130	2370	2019	2302	137	2089	1956	2113	1977	1659	2213	2001	86	0.144	0
5	2259	2239	2565	2324	2266	2103	2292	68	1926	1923	2011	2029	2358	1916	2027	76	0.028	1
6	2482	1978	2632	2567	2692	2693	2508	121	2185	1725	2291	1840	2085	2253	2063	103	0.002	1
7	2757	2245	2386	1894	2111	2541	2322	138	2336	1709	2118	1680	1797	2046	1948	117	0.001	1
8	2625	2383	2185	2133	2343	2303	2329	78	2146	2035	1650	1910	1503	2142	1898	120	0.008	1
9	1890	2346	2280	1927	2204	2221	2144	85	1919	1751	1883	2166	1607	2007	1889	87	0.125	0
10	2109	2084	2535	2109	2675	2378	2315	113	2051	1607	1742	1659	1528	1811	1733	83	0.011	1
11	2390	2459	2466	2238	2470	2646	2445	59	2397	1751	1681	2182	2082	1732	1971	131	0.031	1
12	2138	1878	2477	2177	2472	2232	2229	101	1914	1864	2271	1726	1839	1847	1910	84	0.016	1
13	2358	2148	2254	2132	2259	2461	2269	56	1869	1825	2102	1964	2318	2348	2071	100	0.050	1
14	2043	2401	1927	2717	1879	2653	2270	166	2082	1724	1564	2116	1525	1980	1832	117	0.011	1
15	2192	2412	2514	2302	2432	2031	2314	80	2241	1737	2121	2265	1625	1924	1986	120	0.073	0
16	2249	2465	2665	1927	2235	2366	2318	111	1888	2088	2440	2005	1802	1500	1954	140	0.034	1
17	2062	2057	2337	2499	2442	2210	2268	85	2265	1710	1927	2157	1993	1924	1996	87	0.039	1
18	2652	2184	2155	2166	2569	2103	2305	107	1777	2068	1720	1920	2188	1333	1834	135	0.012	1
19	2394	2434	2635	2070	2186	2069	2298	102	2120	2173	2089	1873	1740	2172	2028	80	0.032	1
20	2363	2104	2535	2836	2251	2219	2385	119	2099	1972	2289	1822	1943	2258	2064	83	0.082	0
					Mean =		2330	97					Mean =		1941	102	0.036	
					SD =		83	29					SD =		99	24	0.040	
					Minimum =		2144	29					Minimum =		1714	55	0.000	
					Maximum =		2508	132					Maximum =		2071	140	0.144	
															Sum/Count = Power =			0.800

Figure 7. Twenty simulated experiments comparing Treatments A and B by Student's t test at $P < 0.05$. The number of simulated "Experiments" with H_0: Treatment A $\neq$ Treatment B were summed to estimate experimental power for experiments with mean body weights of 2,353 and 2,047 g and standard deviations of 217 g (13% difference). Responses were simulated with Microsoft Excel's random number generator (Pesti et al., 2018).

declaring real differences not significant if they exist, and also decreases the chances of declaring significant differences if none exist.

From Figures 4 and 5, we show that predicting power considering the number of birds per pen is a critical exercise to meet the ethical requirements for the use of animals in research. However, the practicality of this approach can be problematic, as the number of birds that can be housed in a pen is dictated by the facilities available. For example, to detect a 10% difference 500 birds are required for 2 treatments; either 5 replicates of 50 birds per pen, or 10 replicates of 25 birds per pen. Housing these birds in a way that practices "refinement" must then be considered. Refinement takes into consideration the space available for appropriate pen size and stocking densities, social dynamics related to group size, and the ease of managing 25 birds compared to 50. Should the researcher require detection of a 5% difference, 18 replicates pens of 25 birds per pen is needed (Figure 4), or 14 replicate pens of 50 birds per pen. This results in a total of 450 birds vs. 700 birds, respectively. Calculating the error based on individual birds, that is, predicting power based on individual bird variation in addition to pen variation, ensures the appropriate amount of birds are used, considering the researchers adversity to risk. It is difficult to interpret risk assessment in terms of legislative standards required for ethical research using terms like "valid" results. As such, this analysis demonstrates considering individual bird data is best to predict the sample size required for ethical research in the future by clarifying just what the risks involved are.

The sampling of normally distributed measurements of experimental subjects, such as chickens, results in power curves that are sigmoidal in nature, as presented in Figures 2 to 4. This makes determining the optimum number of animals difficult. Choosing one arbitrary point on one of the lines in Figure 4 or 5 to be the standard for experimental power decisions is obviously a great over-simplification of the problem. It might be helpful if the mean difference between treatments was known, but in this scenario outcomes will not be known, because we are dealing with research. Knowing a minimum difference for economic importance of mean differences could be helpful to decision makers in some cases.

There are ways to decrease variation amongst sampled birds. One method is to choose only one sex to study greatly decreases within pen (genetic) variation. The drawback of this approach is that it is then applicable to industries that house mixed-sex flocks (i.e., meat chickens) as it is not known if the results are also applicable to the other sex. Another way is to truncate distributions and choose only birds close to the mean. However, then it is not known the results are applicable to large and small birds in mixed flock scenarios.

There are 2 situations in practical experiments with poultry that would especially benefit from attempts to estimate, experimental power. The first is when the treatments being compared are feed additives and the objective is to show that one additive, or diet, is just as good as the other. In this case, experimental power should be linked to the level of difference that is economically important. Conclusions of no statistical difference should not be the result of inadequate bird numbers and replication (Greenland, 2011). Consider, for example, if Treatment A were declared to be just as good as Treatment B (no significant difference), but the actual mean difference was 50 g body weight per bird. If 50 g per bird meant a very significant increase or decrease in profits to a company, the declaration of no significant statistical difference would be entirely misleading. The second case is in determining responses to an input such as an environmental constraint or nutrient level in the feed. It is important to have small confidence limits on any response to make further economic modeling meaningful. For instance, in many nutritional requirement experiments the requirement is listed without a confidence interval. However, an estimate of the confidence interval for the requirement is absolutely necessary to understand the value of the requirement, and apply it to feed formulation in a meaningful way. Although experimental power may be considered for funding and animal care committees, it is not often discussed in published papers where it would be helpful to readers. When experimental power has been considered, in our experience it has always been with commonly accepted, and arbitrary, levels of significance.

Commonly accepted levels of significance for α and β error are 0.05 and 0.80 (Hartnell, 2007, FEEDAP, 2011). That is, researchers would expect to wrongly declare significant differences when none exist about 5% of the time, but only detect real differences (of a specified size) 80% of the time. These values were arbitrarily chosen at a time when calculating actual probabilities was very time consuming. The actual calculations of experimental power are now based on the significance levels and the differences that the researcher would like to be declared significant in the experiment. With modern computing capacity, it is possible to perform many thousands of such calculations each second, allowing researchers to visualize experimental power as a 3-dimensional surface instead of a static point (Figure 1). It is tempting to conclude that experimental conclusions of no significant differences are justified based on power considerations from previous research. Any such conclusions are not valid. The conclusions from each experiment should be based solely on the probabilities calculated from the actual variation observed in that experiment (Greenland, 2011).

A limitation of the prediction presented in Figure 1 is that the SD was based on pen means. Since poultry scientists changed the experimental unit from individual birds on a treatment in one pen to the mean of several birds in multiple pens there is little data available on the individual variation of birds within experimental pens. From the perspective of geneticists, all variation in responses is either due to genetics or the environment. The genetic component of our experiments is straight forward, it is the bird (or animal) that we choose to use. The choice of genetic strain determines the amount of

inherent genetic variability. The remainder of the variation, the environmental factors, include the imposed treatments and the microenvironment (the pens) in which our birds are kept.

From the traditional statistical perspective, experiments have been conducted with the pen average as the experimental unit in poultry studies, with the average responses of birds in each pen providing the experimental observations (body weight in the present example). In this scenario, variation not attributed to the treatment is considered to be random error. When planning future experiments, the error mean square is regarded as the standard deviation squared for the purpose of describing and estimating experimental variation. The SD of pens of n replicate birds, being normally distributed, is then proportional to $n^{0.5}$. For example, if a historical experiment had 5 birds per pen, and an SD = 7, then the SD of experiments with 10 birds per pen would be predicted to be $SD_x = (7 \times 5^{0.5})/10^{0.5} = 4.95$. This assumption is based on the birds not becoming crowded, limited by feeder space, altering their own micro-environment by producing heat, presence of ammonia build up in the house, or becoming subject to social stressors.

With this traditional approach, the error that should be attributed to pen-to-pen variation is not totally ignored, it is simply included within the random error. This is entirely appropriate for hypothesis testing of an experiment that has already been conducted. However, it may not be appropriate for estimating variation in future experiments if there is a possibility of having different numbers of observations per pen and/or different numbers of pens per treatment. For future planning purposes, the error attributable to pens and the random error should be separated. Just as increasing the number of birds per pen decreases the error mean square relative to any observed mean differences, increasing the number of pens decreases the error mean square relative to any observed mean differences. By determining the contribution of pen-to-pen variation independent of bird-to-bird variation, the accuracy of predicting overall SDs for future experiments should be improved. Demetrio et al. (2013) wrote that bird-to-bird and pen-to-pen variations had to be guessed. For many response variables, there are historical data on bird-to-bird variation that could and should be used when estimating future responses.

Ethical decisions are often said to be underpinned by the 3 Rs framework, which state: 1) Where possible the use of animals should be *replaced* (i.e., in vitro experiments); 2) Methods should be *refined* to safeguard animal welfare; and 3) The minimum number of animals should be used to produce a valid result (Fenwick et al., 2019). First, regarding point 1. Above, the quote often attributed to Albert Einstein should be considered "If we knew what it was we were doing, it would not be called research, would it?" Computer simulations can be very helpful in refining (planning) research, as evidenced by the power graphs in Figures 1, 4, and 5; however, they cannot replace it. Computer modeling must be based on what we know, and projections made from our current knowledge base. Even when we have an excellent understanding of current interrelationships, we may not know if the projections should be linear, log transformed, or sine wave. It is only research if it is trying to understand the unknowns yet to be solved and modeled. Computer-based techniques like holo- and meta-analyses can be very helpful in refining experiments. They can indicate where researchers should look for cause and effect relationships, make experiments more meaningful, and reduce the number of animals used, based on accurate test power predictions.

Second, only by using nested designs when the data are nested, can refinements be made for future experiments and the appropriate balance be struck. Refining experiments is the real key to safeguarding or improving animal welfare and use of resources. Very thoughtful consideration should be given prior to running each experiment regarding how the resulting data will be analyzed (Shim and Pesti, 2013), the potential outcomes, and their interpretations. Third, the latter R (*reduction*) suggests that too many animals used in an experiment is unethical, but too few animals will result in incorrect conclusions, which is also unethical (i.e., a type I error of no effect found when an effect is present). Therefore, R's for Replace and Reduction should be replaced by B for Balance. Research efforts have to use a balanced approach, accounting for numbers of birds and pens, acceptable type I and type II errors, economic costs, potential outcomes, and chances of improving bird welfare while improving food production efficiency.

The implications of various experimental designs are often hard to predict because of the complexity of biological systems. It may be tempting to over-simplify experimental power considerations for experiments where careful consideration of nested designs and proper calculation of the sources of variation involved can be most helpful. For instance, cost and benefit analysis decision making is particularly difficult for animal experiments. Many problems are multidimensional, like maximizing the bone health of growing broiler chickens. The interacting factors and interrelationships that must be understood include dietary calcium and calcium solubility, dietary phosphorus and its chemical form, dietary vitamin D, ultraviolet light exposure, bird activity, genetics (which are constantly changing), and exogenous dietary enzymes. There is great potential to exaggerate the importance of any single experiment on the bone health and welfare of billions of birds grown world-wide each month. Computer simulations and projections of experimental power can only help answer these questions.

Another aspect of research where computers can be particularly helpful is in aiding researchers to visualize and understand the different aspects of experimental power. Experiments can be simulated to illustrate the magnitude of different outcomes obtained from identical experiments (Figures 6 and 7). The proportion of experiments finding $P < 0.05$ with a real difference of 18.8% is 16 of 20, indicating a power of 0.80 (Figure 6). If the real difference is 13.0%, only half the experiments result in $P < 0.05$, indicating an experimental power of 0.50 (Figure 7).

These figures show examples of the ranges of typical simulations, presenting a range of $P < 0.001$ to 0.144 with a real difference of 18.8%, and $P < 0.003$ to 0.717 with a real difference of 13.0%. When 10,000 simulations were conducted, the range for both real differences was $P < 0.000$ to 0.998. When conducting such simulations with Microsoft Excel running on Windows the operator need only press the <F9> to almost immediately create another set of 10,000 randomly generated samples based on the mean and standard deviations inputted.

Many variables need to be considered when planning for future experiments. Decisions have to be made based on the probability of finding differences, meaning there is no single correct answer to how many birds per pen and pens per treatment should be used. Decisions must be based on the risk of finding differences that the experimenter finds acceptable. An 80% chance of finding a real difference of 5% seems to be acceptable for many trials, but it is totally arbitrary, albeit accepted practice. Therefore, the level of difference to be detected must be determined by the purpose of the experiment, with its impact and application in mind. Experimental power considerations should also include costs. Since researchers do not know the expected outcome of their experiments, it would be prudent to look at the slopes of the lines for different numbers of replicates depicting possible real differences vs. the chances of declaring a difference significant. In different intervals of these curves, it may or may not be deemed valuable to increase sample sizes to decrease the chances of a wrong conclusion.

Different and perfectly valid outcomes can result from seemingly identical experiments. This is due to random sampling from the same population (Pesti et al., 2018). When researchers accept the null hypothesis that there are no differences ($P < 0.05$) when they expect to find some, they are faced with choices. They may re-examine their hypothesis and determine that it was incorrect in the first place. Alternatively, they may still think that their hypothesis was correct and the $P > 0.05$ was the result of random error. They can therefore repeat the experiment, often with increased replication. The choice between these 2 approaches will be influenced by the actual probability of real differences, and not by reliance on an arbitrary standard like 0.05 (1 chance in 20). If the probability that there were real differences was 0.99, the researcher's conclusion may be the opposite had the probability been 0.051. The response would be different if the probability of real differences was 0.10 or 0.20, or even 0.50 (half a chance of a real difference). Because different outcomes can come from seemingly identical experiments, many researchers will not accept the results of one "valid" experiment, be they positive or negative, and insist on replicating experiments twice, or several times, before making conclusions that they deem "valid" for their application. Adding covariates, such as sex or feed intake, is another possibility when attempting to design experiments to maximize experimental power (Bloom et al., 2007).

The need to repeat experiments comes from several sources. Research is always a matter of chance when randomly choosing experimental subject samples from a large population. There are also other possible sources of variation that are not perfectly controlled, especially dietary composition. For instance, chicks may come from different breeder flocks of different ages and be fed different feeds. Similarly, feed ingredients may be sourced from widely different localities and their compositions cannot be perfectly controlled for different experiments. Simply repeating (doubling the amount of birds) an experiment greatly decreases the chances of type I error, declaring significance when none exists. If $P < 0.05$ is the standard for one experiment, then repeating the experiment decreases the odds of declaring significance by mistake twice to $0.05 \times 0.05 = 0.0025$. Similarly, simply repeating an experiment could decrease the chance of not finding a difference twice if one really exists (type II error) from (1-0.β) to $(1-\beta)^2 = 0.04$ for $\beta = 0.2$ or 0.01 for $\beta = 0.1$.

From the example datasets examined here, there are broad conclusions that can be drawn. First, with a single experiment it is practically impossible to consistently declare small differences significant. Of course this depends on the means and SDs. For the body weight example, up to approximately 3% differences in broiler chicken growth were impossible to consistently declare significant, because of inherent bird-to-bird individual variation. Second, as illustrated in Figures 1, 4, and 5 (with example means and SD's) there are great differences in the slopes of the response lines and thus great differences in how changing the number of birds in a nest or nests in a treatment will affect SDs and experimental outcomes.

It is important to note that these examples have been for comparing multiple means with a 1-way design, not mean separation. When further tests to separate multiple means are being applied, or there are repeated measures, the same general assumptions apply. However, consideration must be given to the nature of the particular design being used. Particularly, when multiple comparisons need to be made, whether the researchers are willing to accept liberal or conservative tests must be considered.

ACKNOWLEDGMENTS

This research did not receive any specific grant from funding agencies in the public, commercial, or not-for-profit sectors.

DISCLOSURES

The authors declare no conflicts of interest.

SUPPLEMENTARY MATERIALS

Supplementary material associated with this article can be found in the online version at doi:10.1016/j.psj.2023.102698.

REFERENCES

Australian Government. 2013. Australian code of practice for the care and use of animals for scientific purposes. 8th Edition. Accessed Dec. 2022. https://www.nhmrc.gov.au/about-us/publications/australian-code-care-and-use-animals-scientific-purposes/australian-code-care-and-use-animals-scientific-purposes-code.

Bloom, H. S., L. Richburg-Hayes, and A. R. Black. 2007. Using covariates to improve precision for studies that randomize schools to evaluate educational interventions. Ed. Eval. Policy Anal. 29:30–59.

Da Costa, M. J., S. Zaragoza-Santacruz, T. J. Frost, J. Halley, and G. M. Pesti. 2017. Straight-run vs. sex separate rearing for 2 broiler genetic lines part 1: live production parameters, carcass yield, and feeding behavior. Poult. Sci 96:2641–2661.

Demetrio, C. G., J. F. Menten, R. A. Leandro, and C. Brien. 2013. Experimental power considerations: justifying replication for animal care and use committees. Poult. Sci. 92:2490–2497.

Fenwick, N., G. Griffin, and C. Gauthier. 2019. The welfare of animals used in science: How the "Three Rs" ethic guides improvements. Can. Vet. J. 50:523–530.

Greenland, S. 2011. Null misinterpretation in statistical testing and its impact on health risk assessment. Prev. Med. 53:225–228.

Hartnell, G. F., G. L. Cromwell, G. R. Dana, A. J. Lewis, D. H. Baker, M. R. Bedford, K. C. Klasing, F. N. Owens, and J. Wiseman. 2007. Best Practices for the Conduct of Animal Studies to Evaluate Crops Genetically Modified for Output Traits. International Life Sciences Institute, Washington DC.

Hays, F. A. 1932. Using biometry. Poult. Sci 11:128–129.

Ibrahim, D. M. 2006. Reduce, Refine, Replace: The Failure of the Three R's in the Future of Animal Experimentation. University of Chicago Legal Forum. Arizona Legal Studies Discussion Paper 06−17. Accessed Feb. 2022. http://ssrn.com/abstract=888206.

Krzywinski, M., N. Altman, and P. Blainey. 2014. Points of significance: nested designs. For studies with hierarchical noise sources, use a nested analysis of variance approach. Nat. Meth. 1:977–978.

Lehr, R. 1992. Sixteen s-squared over d-squared: S relation or crude sample size estimates. Stat. Med. 11:1099–1102.

Marriott, F. H. C. 2002. A Dictionary of Statistical Terms. 5th ed Longman Scientific & Technical, Harlow, Essex.

Neyman, J., and E. S. Pearson. 1928. On the use and interpretation of certain test criteria for purposes of statistical inference. Biometrica 20:175–240 263-294.

Neyman, J., and E. S. Pearson. 1933. On the problem of the most efficient tests of statistical hypotheses. Trans. R. Soc. Lond. Series A 231:289–337.

Parker, S. L. 1925. An Interpretation of the correlation coefficient. Poult. Sci. 4:232–241.

Pesti, G., D. Vedenov, R. V. Nunes, and R. Alhotan 2018. Three Workbooks to Help Estimate Experimental Power and Normalize Experimental Data. University of Georgia Extension. UGA Cooperative Extension Bulletin 1491. Accessed Dec. 2022. https://extension.uga.edu/publications/detail.html?number=B1491.

Rose, M., and E. Grant. 2013. Australia's ethical framework for when animals are used for scientific purposes. Anim. Welfare 22:315–322.

Roush, W. B., and P. R. Tozer. 2004. The power of tests for bioequivalence in feed experiments with poultry. J. Anim. Sci. 82(E-Suppl):E110–E118.

Sadurní, M., A. C Barroeta, R. Sala, C. Sol, M. Puyalto, and L. Castillejos. 2022. Impact of dietary supplementation with sodium butyrate protected by medium-chain fatty acid salts on gut health of broiler chickens. Animals 12:2496–2514.

Schroedek, C. H., and H. B. Lawrence. 1932. The number of chicks required to demonstrate the significance of growth differences. Poult. Sci. 11:208–218.

Shim, M.-Y., and G. M. Pesti. 2013. The use of a pen-size optimization workbook for experiment research design using the Visual Basic for Applications in Excel for poultry. J. Appl. Poult. Res 23:315–322.

Titus, H. W., and J. C. Hammond. 1935. A method of analyzing the data of chick nutrition experiments. Poult. Sci. 14:164–173.

van Belle, G. 2008. Statistical Rules of Thumb, Second Edition. Wiley Series in Probability and Statistics. John Wiley & Sons, Inc, Hoboken, NJ.