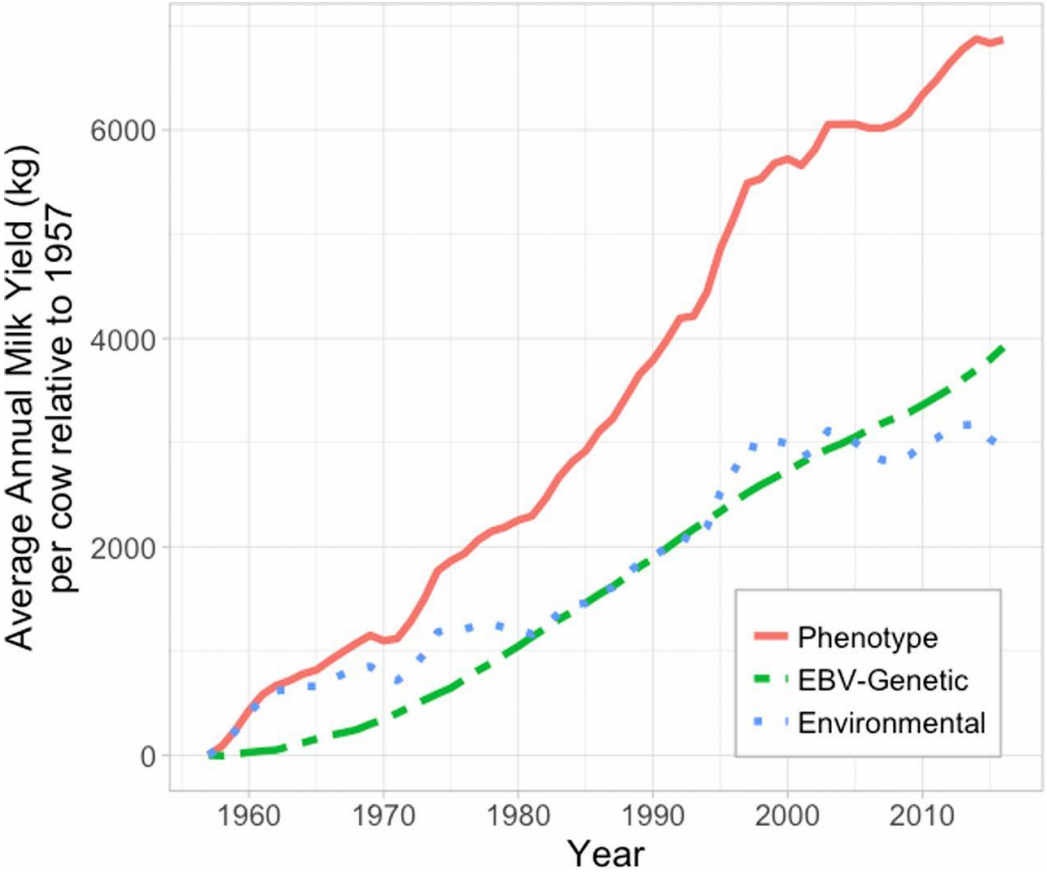
UCDAVIS
Department of Animal Science
Quantitative Genetics Lab

Statistics and Data Science

Hao Cheng
Department of Animal Science
University of California, Davis

UCDAVIS Department of Animal Science Quantitative Genetics Lab

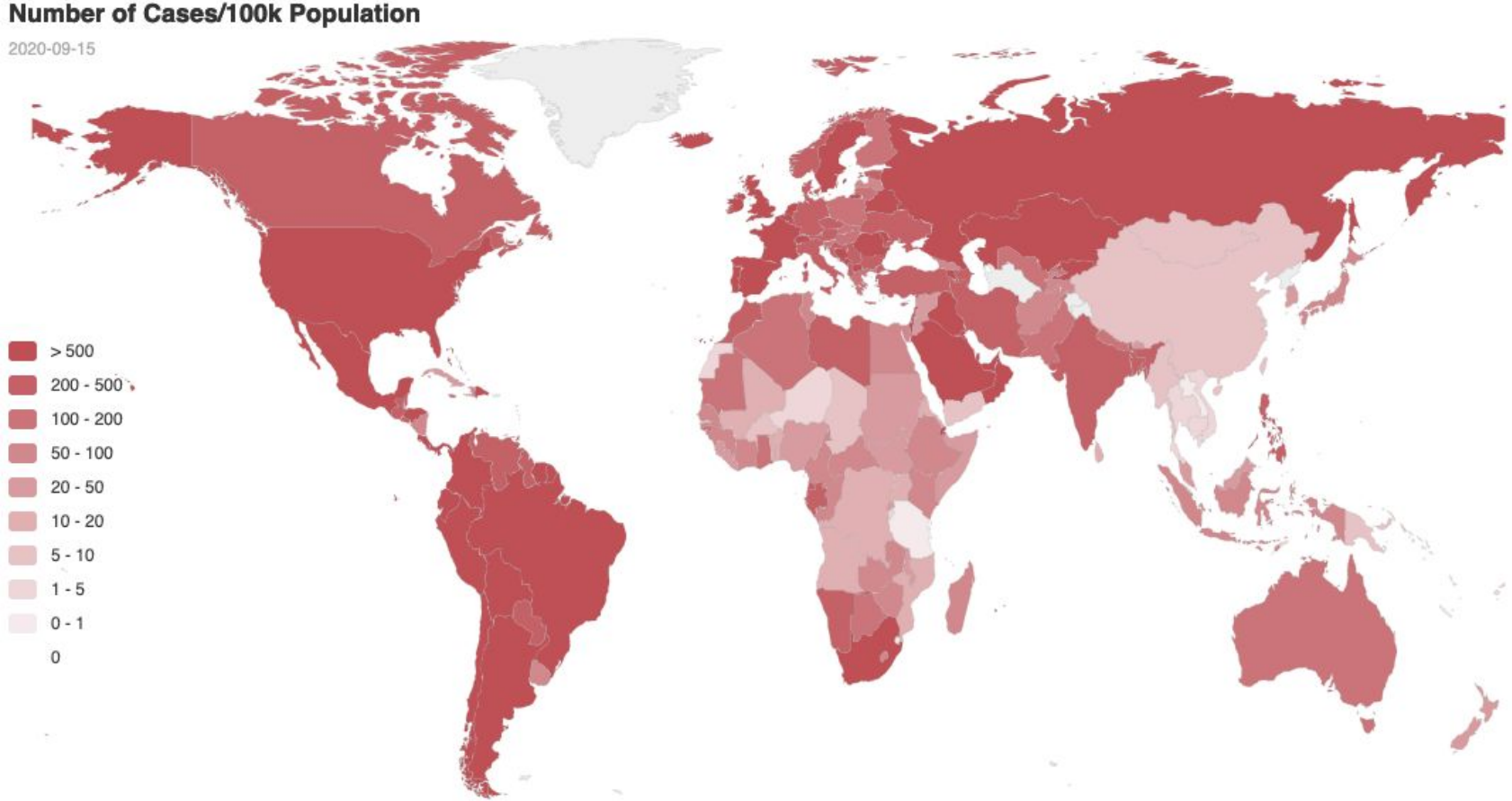
Why is Statistics important?


Average Annual Milk Yield (kg)
per cow relative to 1957
6000
4000
2000
0
1960
1970
1980
1990
2000
2010
Year
Phenotype
EBV-Genetic
Environmental

UCDAVIS Department of Animal Science Quantitative Genetics Lab

Data Science


#30 Bojangles'
Famous Chicken 'n Biscuits
749 US Locations
0:00 / 2:28

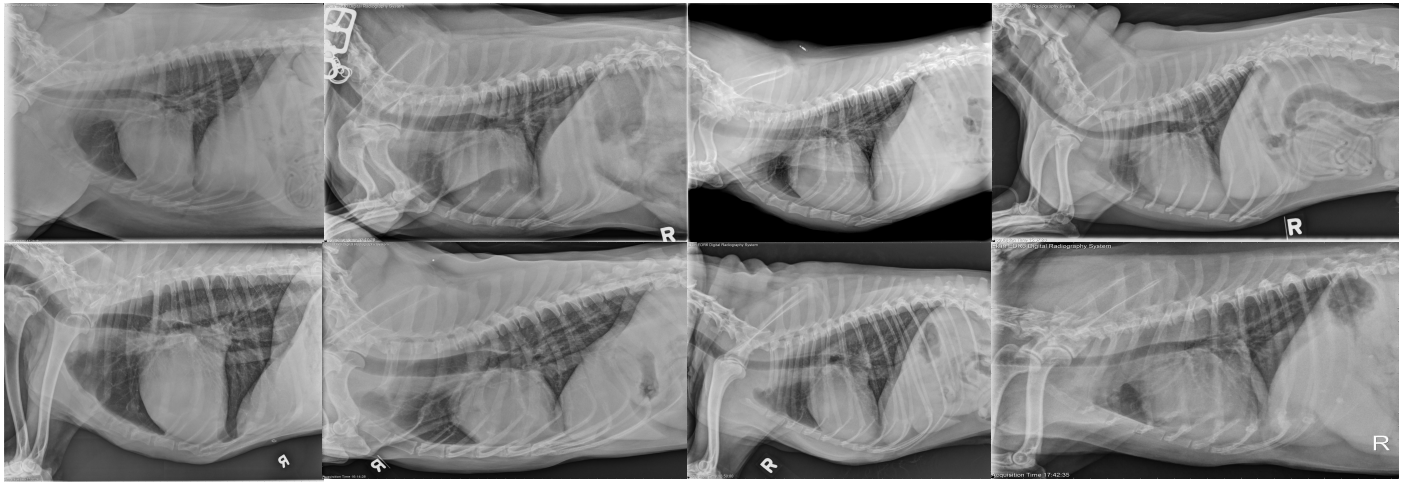

Number of Cases/100k Population
2020-09-15
> 500
200 - 500
100 - 200
50 - 100
20 - 50
10 - 20
5 - 10
1 - 5
0 - 1
0

UCDAVIS Department of Animal Science Quantitative Genetics Lab

Pilot study: Application of artificial intelligence for detecting left atrial enlargement on canine thoracic radiographs

Shen Li, Zigui Wang, Lance C. Visser, Erik R. Wisner, Hao Cheng

First published: 11 August 2020 | https://doi.org/10.1111/vru.12901



Quiz 1: What is p-value?

The p-value is the probability of obtaining a value of the test statistic as or more extreme than the observed value of the test statistic when the null hypothesis is true.

UCDAVIS Department of Animal Science Quantitative Genetics Lab

Creative Writing Study

- Experimental units: writers
- Treatment: type of questionnaire given at the beginning of the study (questions emphasize either intrinsic or extrinsic rewards)
- Random assignment: 24 intrinsic, 23 extrinsic
- Response: average of 12 evaluations of creativity on a 40 point scale

UCDAVIS Department of Animal Science Quantitative Genetics Lab

Observed data

intrinsic:	12.0	12.0	12.9	13.6	16.6	17.2
	17.5	18.2	19.1	19.3	19.8	20.3
	20.5	20.6	21.3	21.6	22.1	22.2
	22.6	23.1	24.0	24.3	26.7	29.7
extrinsic:	5.0	5.4	6.1	10.9	11.8	12.0
	12.3	14.8	15.0	16.8	17.2	17.2
	17.4	17.5	18.5	18.7	18.7	19.2
	19.5	20.7	21.2	22.1	24.0	

null hypothesis:

Treatments have the same effect.

model-based

(relies on the specification of a model)

model-based

- two independent random samples
- homogeneous population variances
- normality

Randomization

(design-based)

Quiz 2: What is type I (II) error?

		True condition			
	Total population	Condition positive	Condition negative	Prevalence $= \frac{\Sigma \text{ Condition positive}}{\Sigma \text{ Total population}}$	Accuracy (ACC) $= \frac{\Sigma \text{ True positive} + \Sigma \text{ True negative}}{\Sigma \text{ Total population}}$
Predicted condition	Predicted condition positive	**True positive**	**False positive**, Type I error	Positive predictive value (PPV), Precision $= \frac{\Sigma \text{ True positive}}{\Sigma \text{ Predicted condition positive}}$	False discovery rate (FDR) $= \frac{\Sigma \text{ False positive}}{\Sigma \text{ Predicted condition positive}}$
	Predicted condition negative	**False negative**, Type II error	**True negative**	False omission rate (FOR) $= \frac{\Sigma \text{ False negative}}{\Sigma \text{ Predicted condition negative}}$	Negative predictive value (NPV) $= \frac{\Sigma \text{ True negative}}{\Sigma \text{ Predicted condition negative}}$
		True positive rate (TPR), Recall, Sensitivity, probability of detection, Power $= \frac{\Sigma \text{ True positive}}{\Sigma \text{ Condition positive}}$	False positive rate (FPR), Fall-out, probability of false alarm $= \frac{\Sigma \text{ False positive}}{\Sigma \text{ Condition negative}}$	Positive likelihood ratio (LR+) $= \frac{\text{TPR}}{\text{FPR}}$	Diagnostic odds ratio (DOR) $= \frac{\text{LR+}}{\text{LR−}}$
		False negative rate (FNR), Miss rate $= \frac{\Sigma \text{ False negative}}{\Sigma \text{ Condition positive}}$	Specificity (SPC), Selectivity, True negative rate (TNR) $= \frac{\Sigma \text{ True negative}}{\Sigma \text{ Condition negative}}$	Negative likelihood ratio (LR−) $= \frac{\text{FNR}}{\text{TNR}}$	F_1 score $= 2 \cdot \frac{\text{Precision} \cdot \text{Recall}}{\text{Precision} + \text{Recall}}$

https://en.wikipedia.org/wiki/Sensitivity_and_specificity

UCDAVIS Department of Animal Science Quantitative Genetics Lab

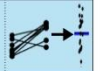
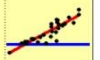
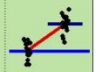
Common Statistical Tests

UC DAVIS Department of Animal Science Quantitative Genetics Lab

Common statistical tests are linear models

Last updated: 28 June, 2019.Also check out the Python version!

See worked examples and more details at the accompanying notebook: https://lindeloev.github.io/tests-as-linear

	Common name	Built-in function in R	Equivalent linear model in R	Exact?	The linear model in words	Icon
Simple regression: lm(y ~ 1 + x)	**y is independent of x** P: One-sample t-test N: Wilcoxon signed-rank	t.test(y) wilcox.test(y)	lm(y ~ 1) lm(signed_rank(y) ~ 1)	✓ for N >14	One number (intercept, i.e., the mean) predicts **y**. - (Same, but it predicts the *signed rank* of **y**.)	
	P: Paired-sample t-test N: Wilcoxon matched pairs	t.test(y_1, y_2, paired=TRUE) wilcox.test(y_1, y_2, paired=TRUE)	lm(y_2 - y_1 ~ 1) lm(signed_rank(y_2 - y_1) ~ 1)	✓ for N >14	One intercept predicts the pairwise $\mathbf{y_2}$-$\mathbf{y_1}$ differences. - (Same, but it predicts the *signed rank* of $\mathbf{y_2}$-$\mathbf{y_1}$.)	
	y ~ continuous x P: Pearson correlation N: Spearman correlation	cor.test(x, y, method='Pearson') cor.test(x, y, method='Spearman')	lm(y ~ 1 + x) lm(rank(y) ~ 1 + rank(x))	✓ for N >10	One intercept plus **x** multiplied by a number (slope) predicts **y**. - (Same, but with *ranked* **x** and **y**)	
	y ~ discrete x P: Two-sample t-test P: Welch's t-test N: Mann-Whitney U	t.test(y_1, y_2, var.equal=TRUE) t.test(y_1, y_2, var.equal=FALSE) wilcox.test(y_1, y_2)	lm(y ~ 1 + G_2)[A] gls(y ~ 1 + G_2, weights=…[B])[A] lm(signed_rank(y) ~ 1 + G_2)[A]	✓ ✓ for N >11	An intercept for **group 1** (plus a difference if **group 2**) predicts **y**. - (Same, but with one variance *per group* instead of one common.) - (Same, but it predicts the *signed rank* of **y**.)	
Multiple regression: lm(y ~ 1 + x_1 + x_2 +…)	P: One-way ANOVA N: Kruskal-Wallis	aov(y ~ group) kruskal.test(y ~ group)	lm(y ~ 1 + G_2 + G_3 +…+ G_N)[A] lm(rank(y) ~ 1 + G_2 + G_3 +…+ G_N)[A]	✓ for N >11	An intercept for **group 1** (plus a difference if group ≠ 1) predicts **y**. - (Same, but it predicts the *rank* of **y**.)	
	P: One-way ANCOVA	aov(y ~ group + x)	lm(y ~ 1 + G_2 + G_3 +…+ G_N + x)[A]	✓	- (Same, but plus a slope on **x**.) *Note: this is discrete AND continuous. ANCOVAs are ANOVAs with a continuous x.*	
	P: Two-way ANOVA	aov(y ~ group * sex)	lm(y ~ 1 + G_2 + G_3 +…+ G_N + S_2 + S_3 +…+ S_K + G_2*S_2 + G_3*S_3 + … + G_N*S_K)	✓	Interaction term: changing **sex** changes the **y ~ group** parameters. *Note: $\mathbf{G_{2\,to\,N}}$ is an indicator (0 or 1) for each non-intercept levels of the **group** variable. Similarly for $\mathbf{S_{2\,to\,K}}$ for sex. The first line (with G_i) is main effect of group, the second (with S_i) for sex and the third is the **group × sex** interaction. For two levels (e.g. male/female), line 2 would just be "S_2" and line 3 would be S_2 multiplied with each G_i.*	*[Coming]*
	Counts ~ discrete x N: Chi-square test	chisq.test(groupXsex_table)	**Equivalent log-linear model** glm(y ~ 1 + G_2 + G_3 + … + G_N + S_2 + S_3+ … + S_K + G_2*S_2 + G_3*S_3 +…+ G_N*S_K, family=…)[A]	✓	Interaction term: (Same as Two-way ANOVA.) *Note: Run glm using the following arguments: `glm(model, family=poisson())`* *As linear-model, the Chi-square test is log(y_i) = log(N) + log(α_i) + log(β_i) + log($\alpha_i\beta_i$) where α_i and β_i are proportions. See more info in the accompanying notebook.*	*Same as Two-way ANOVA*
	N: Goodness of fit	chisq.test(y)	glm(y ~ 1 + G_2 + G_3 +…+ G_N, family=…)[A]	✓	(Same as One-way ANOVA and see Chi-Square note.)	*1W-ANOVA*

List of common parametric (P) non-parametric (N) tests and equivalent linear models. The notation y ~ 1 + x is R shorthand for y = 1·b + a·x which most of us learned in school. Models in similar colors are highly similar, but really, notice how similar they *all* are across colors! For non-parametric models, the linear models are reasonable approximations for non-small sample sizes (see "Exact" column and click links to see simulations). Other less accurate approximations exist, e.g., Wilcoxon for the sign test and Goodness-of-fit for the binomial test. The signed rank function is `signed_rank = function(x) sign(x) * rank(abs(x))`. The variables G_i and S_i are "dummy coded" indicator variables (either 0 or 1) exploiting the fact that when Δx = 1 between categories the difference equals the slope. Subscripts (e.g., G_2 or y_1) indicate different columns in data. lm requires long-format data for all non-continuous models. All of this is exposed in greater detail and worked examples at https://lindeloev.github.io/tests-as-linear.

[A] See the note to the two-way ANOVA for explanation of the notation.
[B] Same model, but with one variance per group: `gls(value ~ 1 + G2, weights = varIdent(form = ~1|group), method="ML")`.

Jonas Kristoffer Lindeløv
https://lindeloev.net

https://lindeloev.github.io/tests-as-linear/#51_independent_t-test_and_mann-whitney_u

Some Matrix Algebra

		movie		
		1	2	3
customer	1	4	1	?
	2	?	3	5
	3	?	?	3
	4	3	1	?

Can we guess ratings for customer/movie combinations not in the dataset?

y_{ij} = customer i's rating of movie j

Which movie is best?

$$y_{ij} = \mu + c_i + m_j + \epsilon_{ij}$$

		movie		
		1	2	3
customer	1	4	1	?
	2	?	3	5
	3	?	?	3
	4	3	1	?

Can we guess ratings for customer/movie combinations not in the dataset?

y_{ij} = customer i's rating of movie j

Which movie is best?

$$y_{ij} = \mu_{ij} + \epsilon_{ij}$$