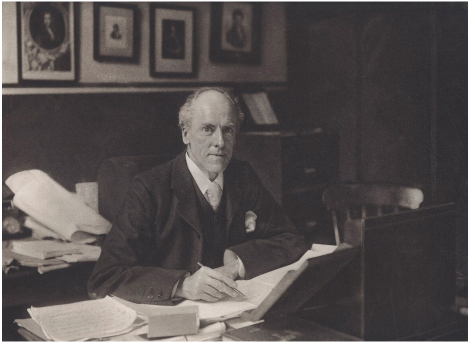
Cannlytics

Cannabis Data Science

Saturday Morning Statistics

November 13th, 2021



Karl Pearson (1857 - 1936)

Karl Pearson's contributions:

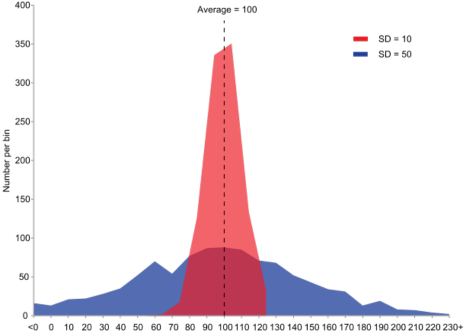
- Pearson correlation coefficient.
- p-value.
- Principal component analysis.
- Created the first histogram.

Sir Ronald Fisher (1890 - 1962)

Ronald Fisher's contributions:

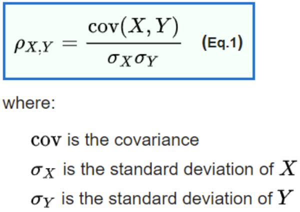
- Introduced the term variance.
- Analysis of variance (ANOVA).
- Popularized the maximum likelihood estimation method.
- The 'F' of the F distribution is named in his honor.

Variance



Example of samples from two populations with the same mean but different variances. The red population has mean 100 and variance 100 (SD=10) while the blue population has mean 100 and variance 2500 (SD=50).

Population correlation coefficient



$$\rho_{X,Y} = \frac{\operatorname{cov}(X,Y)}{\sigma_X \sigma_Y} \quad \textbf{(Eq.1)}$$

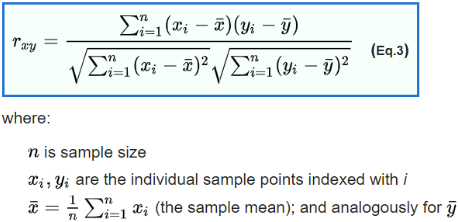
where:

cov is the covariance

σ_X is the standard deviation of X

σ_Y is the standard deviation of Y

Sample correlation coefficient



$$r_{xy} = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2}\sqrt{\sum_{i=1}^{n}(y_i - \bar{y})^2}} \quad \textbf{(Eq.3)}$$

where:

n is sample size

x_i, y_i are the individual sample points indexed with *i*

$\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$ (the sample mean); and analogously for $\bar{y}$

Analysis of variance (ANOVA)

- A statistical test of whether two or more population means are equal.
- Generalizes the *t*-test beyond two means,

Hypothesis Test Type Errors

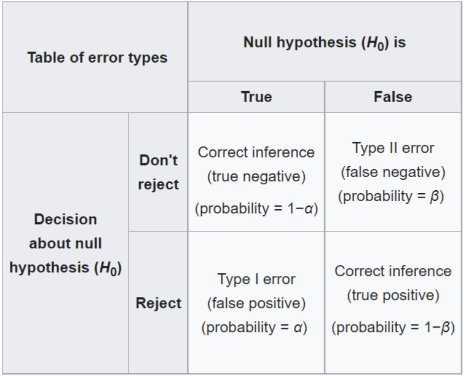


Table of error types		Null hypothesis (H_0) is	
		True	**False**
Decision about null hypothesis (H_0)	**Don't reject**	Correct inference (true negative) (probability = 1−α)	Type II error (false negative) (probability = β)
	Reject	Type I error (false positive) (probability = α)	Correct inference (true positive) (probability = 1−β)

Modern Statistical Models

- **Fixed-effects models** - Apply to situations in which the experimenter applies one or more treatments to the subjects of the experiment to see whether the response variable values change. This allows the experimenter to estimate the ranges of response variable values that the *treatment* would generate in the population as a whole.
- **Random-effects models** - Used when the treatments are not fixed. This occurs when the various factor levels are sampled from a larger population. Because the levels themselves are random variables, some assumptions and the method of contrasting the treatments differ from the fixed-effects model.

Thank you for coming.