Mann-Whitney U test

C. Durso

Application

- Given independent samples of continuous measurements from two populations, test the null hypothesis that the two populations have the same distribution.
- The test actually is more general than this.
- It is useful in the absence of normality.

Set-up

- Two sets of numerical values, $\{x_1, x_2, \ldots x_{n_X}\}$ and $\{y_1, y_2, \ldots y_{n_Y}\}$
- Alternatively, ranks of samples $\{x_1, x_2, \ldots x_{n_X}\}$ and $\{y_1, y_2, \ldots y_{n_Y}\}$ in pooled sample $\{x_1, x_2, \ldots x_{n_X}, y_1, y_2, \ldots y_{n_Y}\}$
- Null hypothesis: population distributions X and Y are such that $P(X > Y) = P(Y > X)$
- If, for some c, the distributions satisfy $X + c = Y$, null hypothesis becomes $c = 0$
- Primarily a test of null hypothesis

Test Statistic

$$w = \left|\left\{(i,j) \middle| x_i < y_j\right\}\right| + \frac{1}{2}\left|\left\{(i,j) \middle| x_i = y_j\right\}\right|$$

Intuition: Model each pair (i,j) equally likely. The test statistic divided by $n_X n_Y$ is $P\left(\left\{(i,j)|x_i < y_j\right\}\right) + \frac{1}{2} P\left(\left\{(i,j)|x_i = y_j\right\}\right)$.

Alternative Form

Define the rank function r:

$\{y_1, y_2, \ldots y_{n_Y}\} \rightarrow [1, n_X + n_Y]$ by

$r(y_j) =$ rank of y_j in $\{x_1, x_2, \ldots x_{n_X}, y_1, y_2, \ldots y_{n_Y}\}$.

Then $w = \sum_{j=1}^{n_Y} r(y_j) - n_Y(n_Y + 1)/2$

Justification

- The value $n_Y(n_Y + 1)/2$ is the sum of the ranks of the y_j's in just $\{y_1, y_2, \ldots y_{n_Y}\}$.
- Each x_i increases $r(y_j)$ by 1 for each $y_j > x_i$. Subtracting $n_Y(n_Y + 1)/2$ from $\sum_{j=1}^{n_Y} r(y_j)$ leaves just these increases.

Example

- Suppose $\begin{matrix}(x_1, & y_1, & x_2, & x_3, & y_2, & y_3, & x_4) \\ \mathbf{1} & \mathbf{2} & \mathbf{3} & \mathbf{4} & \mathbf{5} & \mathbf{6} & \mathbf{7}\end{matrix}$ is in ascending order.
- Subscripts of the y's are their ranks in $\{y_1, y_2, y_3\}$. Values of r appear below the y's.
- Check that each $r(y_j) = j +$ number of x_i's less than y_j.

Evaluation

Calculate (with software usually) the probability q of a value of $W \geq w$ under the assumption that all assignments of the ranks to the first or second populations are equally likely.

Set $p = 2\min(q, 1 - q)$.

χ^2 Test Motivation

C. Durso

Probability of Success
Large Sample Test

Given n Bernoulli trials, test if the probability of success is p:

- $Binomial(n, p)$ approximately $normal\big(np, np(1-p)\big)$
- Observed count k
- $\frac{k-np}{\sqrt{np(1-p)}}$ approximately $normal(0,1\ )$
- Use z-test
- Rule of thumb: $np(1-p) > 3$

χ^2 Distribution

Fact

The χ^2 distribution with n degrees of freedom is the distribution of $\Sigma_{i=1}^{n} X_i^2$ where $X_1, \ldots X_n$ $iid\ normal(0,1)$.

The χ^2 distributions are a 1-parameter family.

χ^2 Test One Proportion

Given n Bernoulli trials, test if the probability of success is p:

- $\frac{k-np}{\sqrt{np(1-p)}}$ approximately $normal(0,1)$
- $\left(\frac{k-np}{\sqrt{np(1-p)}}\right)^2$ approximately χ^2 distribution with 1 degree of freedom

$(O-E)^2/E$ Representation

Fact

$$\left(\frac{k-np}{\sqrt{np(1-p)}}\right)^2 = \frac{(k-np)^2}{np} + \frac{\big((n-k)-n(1-p)\big)^2}{n(1-p)}$$

The second form is a sum of $(observed - expected)^2/expected$ terms.

Derivation

$$\frac{(k-np)^2}{np}+\frac{\big((n-k)-n(1-p)\big)^2}{n(1-p)}$$

Derivation

$$\frac{(k-np)^2}{np}+\frac{\big((n-k)-n(1-p)\big)^2}{n(1-p)}$$

$$\frac{(k-np)^2(1-p)}{np(1-p)}+\frac{(n-k-n+np)^2p}{np(1-p)}$$

Derivation

$$\frac{(k-np)^2}{np}+\frac{\big((n-k)-n(1-p)\big)^2}{n(1-p)}$$

$$\frac{(k-np)^2(1-p)}{np(1-p)}+\frac{(n-k-n+np)^2p}{np(1-p)}$$

$$\frac{(k-np)^2(1-p)}{np(1-p)}+\frac{(-k+np)^2p}{np(1-p)}$$

Derivation

$$\frac{(k-np)^2}{np}+\frac{\big((n-k)-n(1-p)\big)^2}{n(1-p)}$$

$$\frac{(k-np)^2(1-p)}{np(1-p)}+\frac{(n-k-n+np)^2p}{np(1-p)}$$

$$\frac{(k-np)^2(1-p)}{np(1-p)}+\frac{(-k+np)^2p}{np(1-p)}$$

$$\frac{(k-np)^2(1-p)}{np(1-p)}+\frac{(k-np)^2p}{np(1-p)}$$

Derivation

$$\frac{(k-np)^2(1-p)+(k-np)^2p}{np(1-p)}$$

Derivation

$$\frac{(k-np)^2(1-p)+(k-np)^2p}{np(1-p)}$$

$$\frac{(k-np)^2}{np(1-p)}$$

Fisher’s Exact Test

C. Durso

Purpose

Fisher's exact test of independence:

- Applies to contingency tables, example:

	BA	no BA
Over 60	11	93
60 or under	2	83

- Assumes row and column totals are fixed
- Does not require large expected values, unlike a χ^2 test of independence
- Is computationally intensive, unlike a χ^2 test of independence

Concept

- Calculate all probabilities of all tables assuming:
  - Row and column totals are preserved.
  - Row and column variables are independent.
- Sum probabilities of configuration as likely as or less likely than observed.
- Use sum as p-value for null hypothesis of independence.

Calculation Example: Set-up

Consider:

	Factor 1	Factor 2	Row sum
Factor X	a	b	r_1
Factor Y	c	d	r_2
Column sum	c_1	c_2	n

$$n = c_1 + c_2 = r_1 + r_2 = a + b + c + d$$

All Arrangements

- Random permutations of all observations
- $\binom{n}{c_1}$ possible locations for factor 1 observations, equally likely
- $\binom{n}{r_1}$ possible locations for factor X observations, equally likely
- $\binom{n}{c_1}\binom{n}{r_1}$ possible row and column assignments, equally likely given independence

Observed Table: Number of Arrangements That Produce the Observed Count in the Cells

- Construct arrangements with observed cell totals

Observed Table: Number of Arrangements That Produce the Observed Count in the Cells

- Construct arrangements with observed cell totals
- $\binom{n}{r_1}$ possible locations for factor X observations, equally likely (others are factor Y)

Observed Table: Number of Arrangements That Produce the Observed Count in the Cells

- Construct arrangements with observed cell totals
- $\binom{n}{r_1}$ possible locations for factor X observations, equally likely (others are factor Y)
- $\binom{r_1}{a}$ possible locations for factor 1 intersect factor X observations, equally likely (others in row 1 are factor 2)

Observed Table: Number of Arrangements That Produce the Observed Count in the Cells

- Construct arrangements with observed cell totals
- $\binom{n}{r_1}$ possible locations for factor X observations, equally likely (others are factor Y)
- $\binom{r_1}{a}$ possible locations for factor 1 intersect factor X observations, equally likely (others in row 1 are factor 2)
- $\binom{r_2}{c}$ possible locations for factor 1 intersect factor Y observations, equally likely (others in row 2 are factor 2)

Observed Table: Number of Arrangements That Produce the Observed Count in the Cells

- Construct arrangements with observed cell totals
- $\binom{n}{r_1}$ possible locations for factor X observations, equally likely (others are factor Y)
- $\binom{r_1}{a}$ possible locations for factor 1 intersect factor X observations, equally likely (others in row 1 are factor 2)
- $\binom{r_2}{c}$ possible locations for factor 1 intersect factor Y observations, equally likely (others in row 2 are factor 2)
- $\binom{n}{r_1}\binom{r_1}{a}\binom{r_2}{c}$ equally likely assignments of factor X, factor Y, factor 1, and factor 2 produce the observed counts in each cell

Probability of Observed Counts

$$\frac{\binom{n}{r_1}\binom{r_1}{a}\binom{r_2}{c}}{\binom{n}{r_1}\binom{n}{c_1}}$$

Probability of Observed Counts

$$\frac{\binom{n}{r_1}\binom{r_1}{a}\binom{r_2}{c}}{\binom{n}{r_1}\binom{n}{c_1}}$$

$$\frac{\binom{r_1}{a}\binom{r_2}{c}}{\binom{n}{c_1}}$$

Probability of Observed Counts

$$\frac{\binom{n}{r_1}\binom{r_1}{a}\binom{r_2}{c}}{\binom{n}{r_1}\binom{n}{c_1}}$$

$$\frac{\binom{r_1}{a}\binom{r_2}{c}}{\binom{n}{c_1}}$$

$$\frac{(a+b)!\,(c+d)!\,(a+c)!\,(b+d)!}{a!\,b!\,c!\,d!\,n!}$$

p-value

The p-value for the null hypothesis that factor X and factor 1 are independent: sum of all probabilities $\frac{(a'+b')!(c'+d')!(a'+c')!(b'+d')!}{a'!b'!c'!d'!n'!}$ with:

- Original row and column sums: $a' + b' = r_1, c' + d' = r_2$, $a' + c' = c_1, b' + d' = c_2$
- Lower or equal than probability of observed counts: $\frac{(a'+b')!(c'+d')!(a'+c')!(b'+d')!}{a'!b'!c'!d'!n'!} \leq \frac{(a+b)!(c+d)!(a+c)!(b+d)!}{a!b!c!d!n!}$

Covariance

C. Durso

Sample Covariance

Definition

If $\langle x_1, x_2, \dots x_n \rangle$ and $\langle y_1, y_2, \dots y_n \rangle$ are two vectors of numerical data values, the sample covariance of $\langle x_1, x_2, \dots x_n \rangle$ and $\langle y_1, y_2, \dots y_n \rangle$ is $cov(\vec{x}, \vec{y}) = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{n-1}$ where $\bar{x}$ and $\bar{y}$ are the sample means of their respective vectors.

Alternate Form of Sample Covariance

Theorem

The sample covariance of $\langle x_1, x_2, \dots x_n \rangle$ and $\langle y_1, y_2, \dots y_n \rangle$ is equal to

$$\frac{\sum_{i=1}^{n} x_i y_i - n\bar{x} * \bar{y}}{n-1}$$

$$\frac{\sum_{i=1}^{n} (x_i - \bar{x})(y_i - \bar{y})}{n-1} = \frac{\sum_{i=1}^{n} (x_i y_i - x_i \bar{y} - \bar{x} y_i + \bar{x} * \bar{y})}{n-1}$$

Alternate Form of Sample Covariance

Theorem

The sample covariance of $\langle x_1, x_2, \dots x_n \rangle$ and $\langle y_1, y_2, \dots y_n \rangle$ is equal to

$$\frac{\sum_{i=1}^{n} x_i y_i - n\bar{x} * \bar{y}}{n-1}$$

$$\frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{n-1} = \frac{\sum_{i=1}^{n}(x_i y_i - x_i\bar{y} - \bar{x}y_i + \bar{x} * \bar{y})}{n-1}$$

$$= \frac{\sum_{i=1}^{n} x_i y_i}{n-1} - \frac{2n\bar{x} * \bar{y}}{n-1} + \frac{n\bar{x} * \bar{y}}{n-1}$$

Alternate Form of Sample Covariance

Theorem

The sample covariance of $\langle x_1, x_2, \ldots x_n \rangle$ and $\langle y_1, y_2, \ldots y_n \rangle$ is equal to

$$\frac{\sum_{i=1}^{n} x_i y_i - n\bar{x} * \bar{y}}{n-1}$$

$$\frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{n-1} = \frac{\sum_{i=1}^{n}(x_i y_i - x_i \bar{y} - \bar{x} y_i + \bar{x} * \bar{y})}{n-1}$$

$$= \frac{\sum_{i=1}^{n} x_i y_i}{n-1} - \frac{2n\bar{x} * \bar{y}}{n-1} + \frac{n\bar{x} * \bar{y}}{n-1}$$

$$= \frac{\sum_{i=1}^{n} x_i y_i - n\bar{x} * \bar{y}}{n-1}$$

Population or Distribution Covariance

Definition

Given a probability space (S, M, P) and functions $X: S \longrightarrow \mathbb{R}$ and $Y: S \longrightarrow \mathbb{R}$ giving rise to jointly distributed random variables, the covariance of X and Y is the expected value of
$(X - E[X])(Y - E[Y]), E[(X - E[X]), (Y - E[Y])] = Cov[X, Y]$

Note that the covariance may not be well-defined. The sum or integral may not converge.

Alternate Form of Covariance of Jointly Distributed Distributions

Theorem

$$Cov[X,Y] = E[XY] - E[X]E[Y]$$

Proof

$$Cov[X,Y] = E[(X - E[X])(Y - E[Y])]$$
$$= E[XY] - E\big[XE[Y]\big] - E[E[X]Y] + E\big[E[X]E[Y]\big]$$
$$= E[XY] - E[X]E[Y] - E[X]E[Y] + E[X]E[Y]$$
$$= E[XY] - E[X]E[Y]$$

Sample Variance of $\vec{x} + \vec{y}$

Theorem

Given two vectors $\langle x_1, x_2, \dots x_n \rangle$ and $\langle y_1, y_2, \dots y_n \rangle$ of numerical data values, the sample variance of $\langle x_1 + y_1, x_2 + y_2, \dots x_n + y_n \rangle$ equals $var(\vec{x}) + 2cov(\vec{x}, \vec{y}) + var(\vec{y})$ where $var(\vec{w})$ denotes the sample variance of the vector $\vec{w}$.

$$var(\vec{x} + \vec{y}) = \frac{\Sigma_{i=1}^{n} (x_i + y_i)^2 - n(\overline{x + y})^2}{n - 1} = \frac{\Sigma_{i=1}^{n} \left(x_i^2 + 2x_i y_i + y_i^2\right) - n(\bar{x} + \bar{y})^2}{n - 1}$$

Sample Variance of $\vec{x} + \vec{y}$

Theorem

Given two vectors $\langle x_1, x_2, \dots x_n \rangle$ and $\langle y_1, y_2, \dots y_n \rangle$ of numerical data values, the sample variance of $\langle x_1 + y_1, x_2 + y_2, \dots x_n + y_n \rangle$ equals $var(\vec{x}) + 2cov(\vec{x}, \vec{y}) + var(\vec{y})$ where $var(\vec{w})$ denotes the sample variance of the vector $\vec{w}$.

$$var(\vec{x} + \vec{y}) = \frac{\Sigma_{i=1}^{n} (x_i + y_i)^2 - n(\overline{x + y})^2}{n - 1} = \frac{\Sigma_{i=1}^{n} \left(x_i^2 + 2x_i y_i + y_i^2\right) - n(\bar{x} + \bar{y})^2}{n - 1}$$

$$= \frac{\Sigma_{i=1}^{n} \left(x_i^2 + 2x_i y_i + y_i^2\right) - n(\bar{x}^2 + 2\bar{x} * \bar{y} + \bar{y}^2)}{n - 1}$$

Sample Variance of $\vec{x} + \vec{y}$

Theorem

Given two vectors $\langle x_1, x_2, \ldots x_n \rangle$ and $\langle y_1, y_2, \ldots y_n \rangle$ of numerical data values, the sample variance of $\langle x_1 + y_1, x_2 + y_2, \ldots x_n + y_n \rangle$ equals $var(\vec{x}) + 2cov(\vec{x}, \vec{y}) + var(\vec{y})$ where $var(\vec{w})$ denotes the sample variance of the vector $\vec{w}$.

$$var(\vec{x} + \vec{y}) = \frac{\Sigma_{i=1}^{n}(x_i + y_i)^2 - n(\overline{x+y})^2}{n-1} = \frac{\Sigma_{i=1}^{n}\left(x_i^2 + 2x_iy_i + y_i^2\right) - n(\bar{x} + \bar{y})^2}{n-1}$$

$$= \frac{\Sigma_{i=1}^{n}\left(x_i^2 + 2x_iy_i + y_i^2\right) - n(\bar{x}^2 + 2\bar{x} * \bar{y} + \bar{y}^2)}{n-1}$$

$$= \frac{\Sigma_{i=1}^{n}x_i^2 - n\bar{x}^2 + 2\Sigma_{i=1}^{n}2x_iy_i - 2n\bar{x} * \bar{y} + \Sigma_{i=1}^{n}y_i^2 - n\bar{y}^2}{n-1}$$

$$= var(\vec{x}) + 2cov(\vec{x}, \vec{y}) + var(\vec{y})$$

Population or Distribution Variance of $X + Y$

Theorem

If X and Y are jointly distributed random variables and $Var[X+Y]$ is defined, then $Var[X+Y] = Var[X] + 2cov[X,Y] + Var[Y]$

$$Var[X+Y] = E[(X+Y)^2] - E[X+Y]^2$$

$$= E[X^2] + E[2XY] + E[Y^2] - (E[X] + E[Y])^2$$

$$= E[X^2] - E[X]^2 + 2E[XY] - 2E[X]E[Y] + E[Y^2] - E[Y]^2$$

$$Var[X] + 2Cov[X,Y] + Var[Y]$$

Correlation

C. Durso

Correlation

Definition

If $\langle x_1, x_2, \dots x_n \rangle$ and $\langle y_1, y_2, \dots y_n \rangle$ are two vectors of numerical data values, the sample correlation of $\langle x_1, x_2, \dots x_n \rangle$ and $\langle y_1, y_2, \dots y_n \rangle$ is

$$Cor[\vec{x}, \vec{y}] = \frac{Cov[\vec{x}, \vec{y}]}{\sqrt{var[\vec{x}]}\sqrt{var[\vec{y}]}}$$

Definition

If X and Y are jointly distributed random variables,

$$Cor[x, y] = \frac{Cov[x, y]}{\sqrt{var[x]}\sqrt{var[y]}}$$

Correlation Range

Both sample correlation and distribution correlation take values in $[-1,1]$. The fact underlying these restrictions is the Cauchy-Schwarz Inequality:

Theorem

(from Wikipedia) If u and v are vectors in a vector space $\mathbb{F}$ with an inner product $\langle u, v\rangle$ and corresponding norm $\|u\|^2 = \langle u, u\rangle$, then $\langle u, v\rangle^2 \leq \|u\|^2\|v\|^2$ with equality only if u and v are linearly dependent.

Assume v does not equal 0. Set $\lambda = \frac{\langle u,v\rangle}{\|v\|}$

$0 \leq \|u - \lambda v\|^2$, with equality only if $u - \lambda v = 0$

$$= \langle u - \lambda v, u - \lambda v\rangle = \langle u, u\rangle - 2\lambda\langle u, v\rangle + \lambda^2\langle v, v\rangle$$

$$= \|u\|^2 - 2\frac{\langle u, v\rangle}{\|v\|^2}\langle u, v\rangle + \left(\frac{\langle u, v\rangle}{\|v\|^2}\right)^2\|v\|^2 = \|u\|^2 - 2\frac{\langle u, v\rangle}{\|v\|^2} + \frac{\langle u, v\rangle^2}{\|v\|^2}$$

$= \|u\|^2 - \frac{\langle u,v\rangle^2}{\|v\|^2}$, so$\frac{\langle u,v\rangle^2}{\|v\|^2} \leq \|u\|^2$. Multiplying through by $\|v\|^2$gives the desired conclusion.

Link

For the sample correlation, take the vector $u = \langle x_1 - \bar{x}, x_2 - \bar{x}, \dots x_n - \bar{x} \rangle$ and $v = \langle y_1 - \bar{y}, y_2 - \bar{y}, \dots y_n - \bar{y} \rangle$. Then the Cauchy-Schwarz Inequality become $\left((n-1)cov(\vec{x}, \vec{y})\right)^2 \leq (n-1)var(\vec{x})(n-1)var(\vec{y})$.

Interpretation

- Since equality occurs only if $\langle x_1 - \bar{x}, x_2 - \bar{x}, \ldots x_n - \bar{x} \rangle$ and $\langle y_1 - \bar{y}, y_2 - \bar{y}, \ldots y_n - \bar{y} \rangle$ are colinear, the correlation is 1 or -1 if the scatter plot of $\vec{x}$ and $\vec{y}$ is a line.
- Correlation indicates the extent to which $\vec{x}$ and $\vec{y}$ lie along a line.
- Values near 1 or -1 indicate a high degree of colinearity.
- Values near 0 indicating a low degree of colinearity.

Interpretation

- Note that very diverse relationships can produce equal correlations.