INTRODUCTION TO STATISTICS

LECTURE 12

LAST TIME

- Hypothesis testing
- P-values
- Overview of some statistical tests
  - One-sample tests
    - t-test
    - One sample t-test
  - Two-sample tests
    - Two sample t-test
    - Welch's test
    - Pair samples t-test

TODAY

- Wrap-up hypothesis testing
  - Non-parametric tests: an overview
  - Practice

- Two random variables
  - Covariance
  - Correlation

QUICK QUIZ

You are checking a hypothesis H_0 against a two-sided alternative H_1 at the level of significance $\alpha = 0.05$.

After running a statistical test, you obtain a p-value of 0.001.

What is your conclusion?

You are checking a hypothesis H_0 against a two-sided alternative H_1 at the level of significance $\alpha = 0.01$.

After running a statistical test, you obtain a p-value of 0.1.

What is your conclusion?

You are checking a hypothesis H_0 against a two-sided alternative H_1 at the level of significance $\alpha = 0.05$.

After running a statistical test, you obtain a p-value of 0.001.

What is the probability to incorrectly reject the null hypothesis (Type I error)?

You are checking a hypothesis H_0 against a two-sided alternative H_1 at the level of significance $\alpha = 0.05$.

After running a statistical test, you obtain a p-value of 0.001.

What is the probability to obtain a value of the test statistic at least as extreme as the one you've got?

You are checking a hypothesis H_0 against a **one-sided** alternative H_1 at the level of significance $\alpha = 0.05$.

After running a statistical test, you obtain a **two-sided** p-value of 0.09.

What is your conclusion?

ONCE AGAIN ABOUT ONE- AND TWO-SIDED TESTS

- Example: one-sample test for population mean μ

$$H_0\colon \mu = \mu_0$$

ONCE AGAIN ABOUT ONE- AND TWO-SIDED TESTS

- Example: one-sample test for population mean μ

$$H_0\colon\ \mu = \mu_0$$

- Two-sided alternative:

$$H_1\colon \mu \neq \mu_0$$

ONCE AGAIN ABOUT ONE- AND TWO-SIDED TESTS

- Example: one-sample test for population mean μ

$$H_0\colon\ \mu = \mu_0$$

- Two-sided alternative:

$$H_1\colon \mu \neq \mu_0$$

- One-sided alternatives:

$$H_1\colon \mu < \mu_0 \qquad\qquad H_1\colon \mu > \mu_0$$

ONCE AGAIN ABOUT ONE- AND TWO-SIDED TESTS

- Example: one-sample test for population mean μ

$$H_0\colon\ \mu = \mu_0$$

- Two-sided alternative:

$$H_1\colon \mu \neq \mu_0$$

- One-sided alternatives:

$$H_1\colon \mu < \mu_0 \qquad\qquad H_1\colon \mu > \mu_0$$

<u>Pay attention: the null hypothesis is still just $\mu = \mu_0$.</u>

ONCE AGAIN ABOUT ONE- AND TWO-SIDED TESTS

- Example: one-sample test for population mean μ

$$H_0\colon\ \mu = \mu_0$$

- Two-sided alternative:

$$H_1\colon \mu \neq \mu_0$$

- One-sided alternatives:

$$H_1\colon \mu < \mu_0 \qquad\qquad H_1\colon \mu > \mu_0$$

<u>Pay attention: the null hypothesis is still just $\mu = \mu_0$.</u>

Never use a one-sided alternative unless absolutely sure that the opposite of it isn't possible.

NON-PARAMETRIC TESTS

NON-PARAMETRIC TESTS

- Up till now: **parametric** models
  - Example: mean μ and variance σ^2

NON-PARAMETRIC TESTS

- Up till now: **parametric** models
  - Example: mean μ and variance σ^2

- **Non-parametric statistics** doesn’t rely on data belonging to any parametric family of probability distributions:
  - distribution-free or
  - having a specified distribution but with the distribution's parameters unspecified.

NON-PARAMETRIC TESTS

PARAMETRIC HYPOTHESES	NON-PARAMETRIC HYPOTHESES
• Data comes from the normal distribution with specified mean and variance.	
• Data comes from the normal distribution with specified mean and unspecified variance.	

NON-PARAMETRIC TESTS

PARAMETRIC HYPOTHESES	NON-PARAMETRIC HYPOTHESES
• Data comes from the normal distribution with specified mean and variance.	• Data comes from a normal distribution form with both mean and variance unspecified.
• Data comes from the normal distribution with specified mean and unspecified variance.	• Two unspecified continuous distributions are identical.

MANN-WHITNEY U TEST

An alternative to the two-sample t-test
when the distribution of the data cannot be assumed to be normal

MANN-WHITNEY U TEST

- Two independent i.i.d. sets of data:

$$X_1, X_2, \dots, X_n$$

$$Y_1, Y_2, \dots, Y_m$$

MANN-WHITNEY U TEST

- Two independent i.i.d. sets of data:

$$X_1, X_2, \ldots, X_n$$

$$Y_1, Y_2, \ldots, Y_m$$

- Null hypothesis:

H_0: For randomly selected values *X* and *Y* from two populations,

$$P(X > Y) = P(X < Y)$$

MANN-WHITNEY U TEST

- Two independent i.i.d. sets of data:

$$X_1, X_2, \ldots, X_n$$

$$Y_1, Y_2, \ldots, Y_m$$

- Null hypothesis:

H_0: For randomly selected values *X* and *Y* from two populations,

$$P(X > Y) = P(X < Y)$$

- Alternatives
  - Two-sided: $P(X > Y) \neq P(X < Y)$
  - One-sided: $P(X > Y) > P(X < Y)$, $\quad P(X > Y) < P(X < Y)$

MANN-WHITNEY U TEST

- Test statistic:

$$U = \sum_{i=1}^{n} \sum_{j=1}^{m} S(X_i, Y_j)$$

where $S(X,Y) = \begin{cases} 1, & Y < X \\ 1/2\,, & Y = X \\ 0, & Y > X \end{cases}$

MANN-WHITNEY U TEST

- Test statistic:

$$U = \sum_{i=1}^{n} \sum_{j=1}^{m} S(X_i, Y_j)$$

$$\text{where } S(X,Y) = \begin{cases} 1, & Y < X \\ 1/2\,, & Y = X \\ 0, & Y > X \end{cases}$$

- Null distribution:

For small n, tabulated

For large n, $(n \geq 30)$ $U \sim \text{normal}$

WILCOXON SIGNED RANK TEST

An alternative to the paired t-test
when the distribution of the differences cannot be assumed to be normal

WILCOXON SIGNED RANK TEST

- Two paired i.i.d. sets of data:

$$X_1, X_2, \dots, X_n$$

$$Y_1, Y_2, \dots, Y_m$$

WILCOXON SIGNED RANK TEST

- Two paired i.i.d. sets of data:

$$X_1, X_2, \dots, X_n$$

$$Y_1, Y_2, \dots, Y_m$$

- Null hypothesis:

H_0: difference between the pairs follows a symmetric distribution around zero

WILCOXON SIGNED RANK TEST

- Two paired i.i.d. sets of data:

$$X_1, X_2, \dots, X_n$$

$$Y_1, Y_2, \dots, Y_m$$

- Null hypothesis:

H_0: difference between the pairs follows a symmetric distribution around zero

- Alternatives
  - Two-sided: difference between the pairs doesn't follow a symmetric distribution around zero
  - One-sided: distribution is skewed to one particular side

WILCOXON SIGNED RANK TEST

- Test statistic:

WILCOXON SIGNED RANK TEST

- Test statistic:

Compute $|X_i - Y_i|$ and $\text{sgn}(X_i - Y_i)$

WILCOXON SIGNED RANK TEST

- Test statistic:

Compute $|X_i - Y_i|$ and $\mathrm{sgn}(X_i - Y_i)$

Exclude pairs (X_i, Y_i) where $|X_i - Y_i| = 0$

WILCOXON SIGNED RANK TEST

- Test statistic:

Compute $|X_i - Y_i|$ and $\mathrm{sgn}(X_i - Y_i)$

Exclude pairs (X_i, Y_i) where$|X_i - Y_i| = 0$

Rank the remaining pairs from smallest to largest $|X_i - Y_i| \rightarrow R_i$

WILCOXON SIGNED RANK TEST

- Test statistic:

Compute $|X_i - Y_i|$ and $\text{sgn}(X_i - Y_i)$

Exclude pairs (X_i, Y_i) where$|X_i - Y_i| = 0$

Rank the remaining pairs from smallest to largest $|X_i - Y_i| \rightarrow R_i$

$$W = \sum_{i=1}^{n} R_i \,\text{sgn}(X_i - Y_i)$$

WILCOXON SIGNED RANK TEST

- Test statistic:

Compute $|X_i - Y_i|$ and $\mathrm{sgn}(X_i - Y_i)$

Exclude pairs (X_i, Y_i) where$|X_i - Y_i| = 0$

Rank the remaining pairs from smallest to largest $|X_i - Y_i| \to R_i$

$$W = \sum_{i=1}^{n} R_i \,\mathrm{sgn}(X_i - Y_i)$$

- Null distributions:
  - some specific tabulated distribution.

χ^2-TEST

Independence of two categorical variables based on contingency table

χ^2-TEST

- Motivating example: consider the following **contingency table**:

Education	Married once	Married multiple times	Total
College	550	61	611
No college	681	144	825
Total	1231	205	1436

χ^2-TEST

- Motivating example: consider the following **contingency table**:

Education	Married once	Married multiple times	Total
College	550	61	611
No college	681	144	825
Total	1231	205	1436

- Are education levels and number of marriages (one / many) independent?

χ^2-TEST

Education	Married once	Married multiple times	Total
College	550	61	611
No college	681	144	825
Total	1231	205	1436

- Are education levels and number of marriages (one / many) independent?

If so, cell probabilities are the product of the marginal ones:

Education	Married once	Married multiple times	Total
College			$611/1436$
No college			$825/1436$
Total	$1231/1436$	$205/1436$	1

χ^2-TEST

Education	Married once	Married multiple times	Total
College	550	61	611
No college	681	144	825
Total	1231	205	1436

- Are education levels and number of marriages (one / many) independent?

If so, cell probabilities are the product of the marginal ones:

Education	Married once	Married multiple times	Total
College	0.365		611/1436
No college			825/1436
Total	1231/1436	205/1436	1

χ^2-TEST

Education	Married once	Married multiple times	Total
College	550	61	611
No college	681	144	825
Total	1231	205	1436

- Are education levels and number of marriages (one / many) independent?

If so, cell probabilities are the product of the marginal ones:

Education	Married once	Married multiple times	Total
College	0.365	0.061	$611/1436$
No college			$825/1436$
Total	$1231/1436$	$205/1436$	1

χ^2-TEST

Education	Married once	Married multiple times	Total
College	550	61	611
No college	681	144	825
Total	1231	205	1436

- Are education levels and number of marriages (one / many) independent?

If so, cell probabilities are the product of the marginal ones:

Education	Married once	Married multiple times	Total
College	0.365	0.061	$611/1436$
No college	0.492		$825/1436$
Total	$1231/1436$	$205/1436$	1

χ^2-TEST

Education	Married once	Married multiple times	Total
College	550	61	611
No college	681	144	825
Total	1231	205	1436

- Are education levels and number of marriages (one / many) independent?

If so, cell probabilities are the product of the marginal ones:

Education	Married once	Married multiple times	Total
College	0.365	0.061	$611/1436$
No college	0.492	0.082	$825/1436$
Total	$1231/1436$	$205/1436$	1

χ^2-TEST

Education	Married once	Married multiple times	Total
College	0.365	0.061	611/1436
No college	0.492	0.082	825/1436
Total	1231/1436	205/1436	1

- From this, we can get ***expected counts*** *(multiplying cell probabilities by the total number of women surveyed):*

Education	Married once	Married multiple times
College	550, 523.8	61,
No college	681,	144,

χ^2-TEST

Education	Married once	Married multiple times	Total
College	0.365	0.061	$611/1436$
No college	0.492	0.082	$825/1436$
Total	$1231/1436$	$205/1436$	1

- From this, we can get ***expected counts*** *(multiplying cell probabilities by the total number of women surveyed):*

Education	Married once	Married multiple times
College	550, 523.8	61, 87.2
No college	681,	144,

χ^2-TEST

Education	Married once	Married multiple times	Total
College	0.365	0.061	$611/1436$
No college	0.492	0.082	$825/1436$
Total	$1231/1436$	$205/1436$	1

- From this, we can get ***expected counts*** *(multiplying cell probabilities by the total number of women surveyed):*

Education	Married once	Married multiple times
College	550, 523.8	61, 87.2
No college	681, 707.2	144,

χ^2-TEST

Education	Married once	Married multiple times	Total
College	0.365	0.061	611/1436
No college	0.492	0.082	825/1436
Total	1231/1436	205/1436	1

- From this, we can get ***expected counts*** *(multiplying cell probabilities by the total number of women surveyed):*

Education	Married once	Married multiple times
College	550, 523.8	61, 87.2
No college	681, 707.2	144, 117.8

χ^2-TEST

Education	Married once	Married multiple times
College	550, 523.8	61, 87.2
No college	681, 707.2	144, 117.8

χ^2-TEST

Education	Married once	Married multiple times
College	550, 523.8	61, 87.2
No college	681, 707.2	144, 117.8

- Null hypothesis:

H_0: cell probabilities are the product of the marginal ones, the difference, so the difference between them should be small

χ^2-TEST

Education	Married once	Married multiple times
College	550, 523.8	61, 87.2
No college	681, 707.2	144, 117.8

- Null hypothesis:

H_0: cell probabilities are the product of the marginal ones, the difference, so the difference between them should be small

H_1: the difference between observed and expected counts is large **(!!! one-sided)**

χ^2-TEST

Education	Married once	Married multiple times
College	550, 523.8	61, 87.2
No college	681, 707.2	144, 117.8

- Test statistic: Pearson's chi-square statistic

$$\chi^2 = \sum \frac{(O_i - E_i)^2}{E_i} \sim \chi^2(df), \qquad df = (n-1)(m-1)$$

where O_i - observed count, and E_i - expected count

χ^2-TEST

Education	Married once	Married multiple times
College	550, 523.8	61, 87.2
No college	681, 707.2	144, 117.8

- In this case

$$\chi^2 =$$

χ^2-TEST

Education	Married once	Married multiple times
College	550, 523.8	61, 87.2
No college	681, 707.2	144, 117.8

- In this case

$$\chi^2 = 16.01, \ df \ = 1$$

χ^2-TEST

Education	Married once	Married multiple times
College	550, 523.8	61, 87.2
No college	681, 707.2	144, 117.8

- In this case

$$\chi^2 = 16.01, \ df \ = 1$$

χ^2-TEST

Education	Married once	Married multiple times
College	550, 523.8	61, 87.2
No college	681, 707.2	144, 117.8

- In this case

$$\chi^2 = 16.01, \; df = 1$$

$$\chi^2_{1-\alpha}(1) = \chi^2_{0.95}(1) = 7.879$$

χ^2-TEST

Education	Married once	Married multiple times
College	550, 523.8	61, 87.2
No college	681, 707.2	144, 117.8

- In this case

$$\chi^2 = 16.01, \; df = 1$$

$$\chi^2_{1-\alpha}(1) = \chi^2_{0.95}(1) = 7.879$$

$$16.01 > 7.879 \Longrightarrow \text{reject } H_0$$

NON-PARAMETRIC TESTS

- Fewer assumptions, wider applicability.

NON-PARAMETRIC TESTS

- Fewer assumptions, wider applicability.

- This comes with the cost: when parametric tests are applicable, non-parametric ones have less *power*
  - *a larger sample size can be required to draw conclusions with the same degree of confidence.*

PRACTICE!

Google Classroom -> Lecture 11 -> Two-sample tests

TWO RANDOM VARIABLES

Covariance and correlation

REMINDER: INDEPENDENCE

Random variables X and Y are independent if and only if

REMINDER: INDEPENDENCE

Random variables X and Y are independent if and only if

- Discrete case:

$$P(X = x, Y = y) = P(X)P(Y)$$

REMINDER: INDEPENDENCE

Random variables X and Y are independent if and only if

- Discrete case:

$$P(X = x, Y = y) = P(X)P(Y)$$

- Continuous case:

$$p_{xy}(x, y) = p_x(x)p_y(y)$$

REMINDER: INDEPENDENCE

If X and Y are independent, then

$$var(aX + bY) = a^2 var(X) + b^2 var(Y)$$

REMINDER: INDEPENDENCE

If X and Y are independent, then

$$var(aX + bY) = a^2 var(X) + b^2 var(Y)$$

But what if X and Y are dependent?

COVARIANCE

$$var(aX + bY) = a^2 var(X) + b^2 var(Y) - ab \cdot cov(X, Y)$$

COVARIANCE

- Covariance σ_{XY}^2 is a measure of the joint variability of two random variables X and Y:

$$\sigma_{XY}^2 = E[(X - \bar{X})(Y - \bar{Y})] = E(XY) - \bar{X}\bar{Y}$$

COVARIANCE

- Covariance σ_{XY}^2 is a measure of the joint variability of two random variables X and Y:

$$\sigma_{XY}^2 = E[(X - \bar{X})(Y - \bar{Y})] = E(XY) - \bar{X}\bar{Y}$$

- Example:

Height and weight of a giraffe have a positive covariance: when one is large, the other also tends to be large.

COVARIANCE

- Covariance σ_{XY}^2 is a measure of the joint variability of two random variables X and Y:

$$\sigma_{XY}^2 = E[(X - \bar{X})(Y - \bar{Y})] = E(XY) - \bar{X}\bar{Y}$$

$$\sigma_{XX}^2 =$$

COVARIANCE

- Covariance σ_{XY}^2 is a measure of the joint variability of two random variables X and Y:

$$\sigma_{XY}^2 = E[(X - \bar{X})(Y - \bar{Y})] = E(XY) - \bar{X}\bar{Y}$$

- Note that covariance of a variable with itself is its variance:

$$\sigma_{XX}^2 = E(X - \bar{X})^2 = E(X^2) - \bar{X}^2 = \sigma_X^2$$

COVARIANCE & INDEPENDENCE

$$\sigma_{XY}^2 = E[(X - \bar{X})(Y - \bar{Y})] = E(XY) - \bar{X}\bar{Y}$$

COVARIANCE & INDEPENDENCE

$$\sigma_{XY}^2 = E[(X - \bar{X})(Y - \bar{Y})] = E(XY) - \bar{X}\bar{Y}$$

If X and Y are independent, the covariance

COVARIANCE & INDEPENDENCE

$$\sigma_{XY}^2 = E[(X - \bar{X})(Y - \bar{Y})] = E(XY) - \bar{X}\bar{Y}$$

If X and Y are independent, the covariance $\sigma_{XY}^2 = 0$.

COVARIANCE & INDEPENDENCE

$$\sigma_{XY}^2 = E[(X - \bar{X})(Y - \bar{Y})] = E(XY) - \bar{X}\bar{Y}$$

If X and Y are independent, the covariance $\sigma_{XY}^2 = 0$.

Is the opposite true?

COVARIANCE: EXAMPLE

- Let X take values $-2, -1, 0, 1$ and 2 with equal probabilities. Compute covariance between X and $Y = X^2$

$Y \backslash X$	-2	-1	0	1	2	$p(y_j)$
0						
1						
4						
$p(x_i)$	$1/5$	$1/5$	$1/5$	$1/5$	$1/5$	1

COVARIANCE: EXAMPLE

- Let X take values $-2, -1, 0, 1$ and 2 with equal probabilities. Compute covariance between X and $Y = X^2$

$Y \backslash X$	-2	-1	0	1	2	$p(y_j)$
0	0	0	1/5	0	0	
1						
4						
$p(x_i)$	1/5	1/5	1/5	1/5	1/5	1

COVARIANCE: EXAMPLE

- Let X take values $-2, -1, 0, 1$ and 2 with equal probabilities. Compute covariance between X and $Y = X^2$

$Y \backslash X$	-2	-1	0	1	2	$p(y_j)$
0	0	0	$1/5$	0	0	
1	0	$1/5$	0	$1/5$	0	
4						
$p(x_i)$	$1/5$	$1/5$	$1/5$	$1/5$	$1/5$	1

COVARIANCE: EXAMPLE

- Let X take values $-2, -1, 0, 1$ and 2 with equal probabilities. Compute covariance between X and $Y = X^2$

$Y \backslash X$	-2	-1	0	1	2	$p(y_j)$
0	0	0	1/5	0	0	
1	0	1/5	0	1/5	0	
4	1/5	0	0	0	1/5	
$p(x_i)$	1/5	1/5	1/5	1/5	1/5	1

COVARIANCE: EXAMPLE

- Let X take values $-2, -1, 0, 1$ and 2 with equal probabilities. Compute covariance between X and $Y = X^2$.

$Y \backslash X$	-2	-1	0	1	2	$p(y_j)$
0	0	0	$1/5$	0	0	$1/5$
1	0	$1/5$	0	$1/5$	0	$2/5$
4	$1/5$	0	0	0	$1/5$	$2/5$
$p(x_i)$	$1/5$	$1/5$	$1/5$	$1/5$	$1/5$	1

COVARIANCE: EXAMPLE

- Let X take values $-2, -1, 0, 1$ and 2 with equal probabilities. Compute covariance between X and $Y = X^2$.

$Y \backslash X$	-2	-1	0	1	2	$p(y_j)$
0	0	0	1/5	0	0	1/5
1	0	1/5	0	1/5	0	2/5
4	1/5	0	0	0	1/5	2/5
$p(x_i)$	1/5	1/5	1/5	1/5	1/5	1

$$\sigma_{XY}^2 = E(XY) - \bar{X}\,\bar{Y} =$$

COVARIANCE: EXAMPLE

- Let X take values $-2, -1, 0, 1$ and 2 with equal probabilities. Compute covariance between X and $Y = X^2$.

$Y \backslash X$	-2	-1	0	1	2	$p(y_j)$
0	0	0	1/5	0	0	1/5
1	0	1/5	0	1/5	0	2/5
4	1/5	0	0	0	1/5	2/5
$p(x_i)$	1/5	1/5	1/5	1/5	1/5	1

$$\sigma_{XY}^2 = E(XY) - \bar{X}\,\bar{Y} = \frac{1}{5}(0 - 1 + 1 - 2 + 2) - 0 \cdot 2 = 0$$

COVARIANCE & INDEPENDENCE

$$\sigma_{XY}^2 = E[(X - \bar{X})(Y - \bar{Y})] = E(XY) - \bar{X}\bar{Y}$$

If X and Y are independent, the covariance $\sigma_{XY}^2 = 0$.

The inverse is not true: $\sigma_{XY} = 0$ does not imply that If X and Y are independent.

CORRELATION

- Covariance: $\sigma_{XY}^2 = E(XY) - \bar{X}\,\bar{Y}$

CORRELATION

- Covariance: $\sigma_{XY}^2 = E(XY) - \bar{X}\,\bar{Y}$
- What are the measurement units of it?

CORRELATION

- Covariance: $\sigma_{XY}^2 = E(XY) - \bar{X}\,\bar{Y}$
- What are the measurement units of it?
  - *'units of X times units of Y'*

CORRELATION

- Covariance: $\sigma_{XY}^2 = E(XY) - \bar{X}\ \bar{Y}$
- What are the measurement units of it?
  - *'units of X times units of Y'*
- Hard to compare covariances.

CORRELATION

- Covariance: $\sigma_{XY}^2 = E(XY) - \bar{X}\,\bar{Y}$
- What are the measurement units of it?
  - *'units of X times units of Y'*
- Hard to compare covariances.
- Correlation removes scale from covariance:

$$\rho = \frac{\sigma_{XY}^2}{\sigma_X \sigma_Y} = \frac{E(X-\bar{X})(Y-\bar{Y})}{\sqrt{E(X-\bar{X})^2 E(Y-\bar{Y})^2}}$$

PROPERTIES OF CORRELATION

$$\rho = \frac{\sigma_{XY}^2}{\sigma_X \sigma_Y} = \frac{E(X - \bar{X})(Y - \bar{Y})}{\sqrt{E(X - \bar{X})^2 E(Y - \bar{Y})^2}}$$

PROPERTIES OF CORRELATION

$$\rho = \frac{\sigma_{XY}^2}{\sigma_X \sigma_Y} = \frac{E(X-\bar{X})(Y-\bar{Y})}{\sqrt{E(X-\bar{X})^2 E(Y-\bar{Y})^2}}$$

1. Correlation is dimensionless (it's a ratio!)

PROPERTIES OF CORRELATION

$$\rho = \frac{\sigma_{XY}^2}{\sigma_X \sigma_Y} = \frac{E(X - \bar{X})(Y - \bar{Y})}{\sqrt{E(X - \bar{X})^2 E(Y - \bar{Y})^2}}$$

1. Correlation is dimensionless (it's a ratio!)

2. -$1 \leq \rho \leq 1$

WHAT IS CORRELATION

- Degree to which a pair of variables are *linearly* related.
  - *The higher $|\rho|$ is, the greater the degree of linear dependency is.*

- Sign:

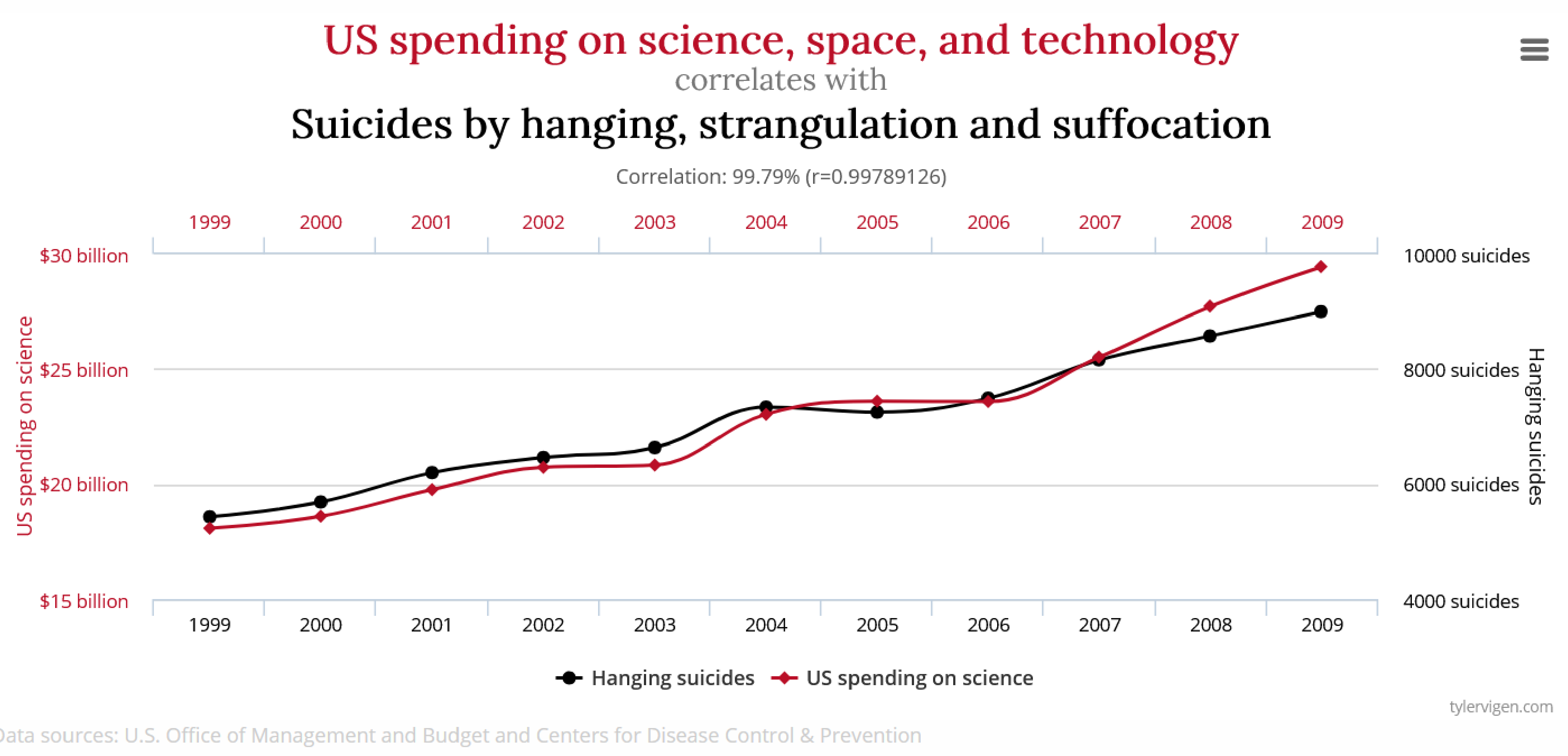
  - $\rho(X, Y) > 0$

    "The larger X is, the larger Y tends to be"

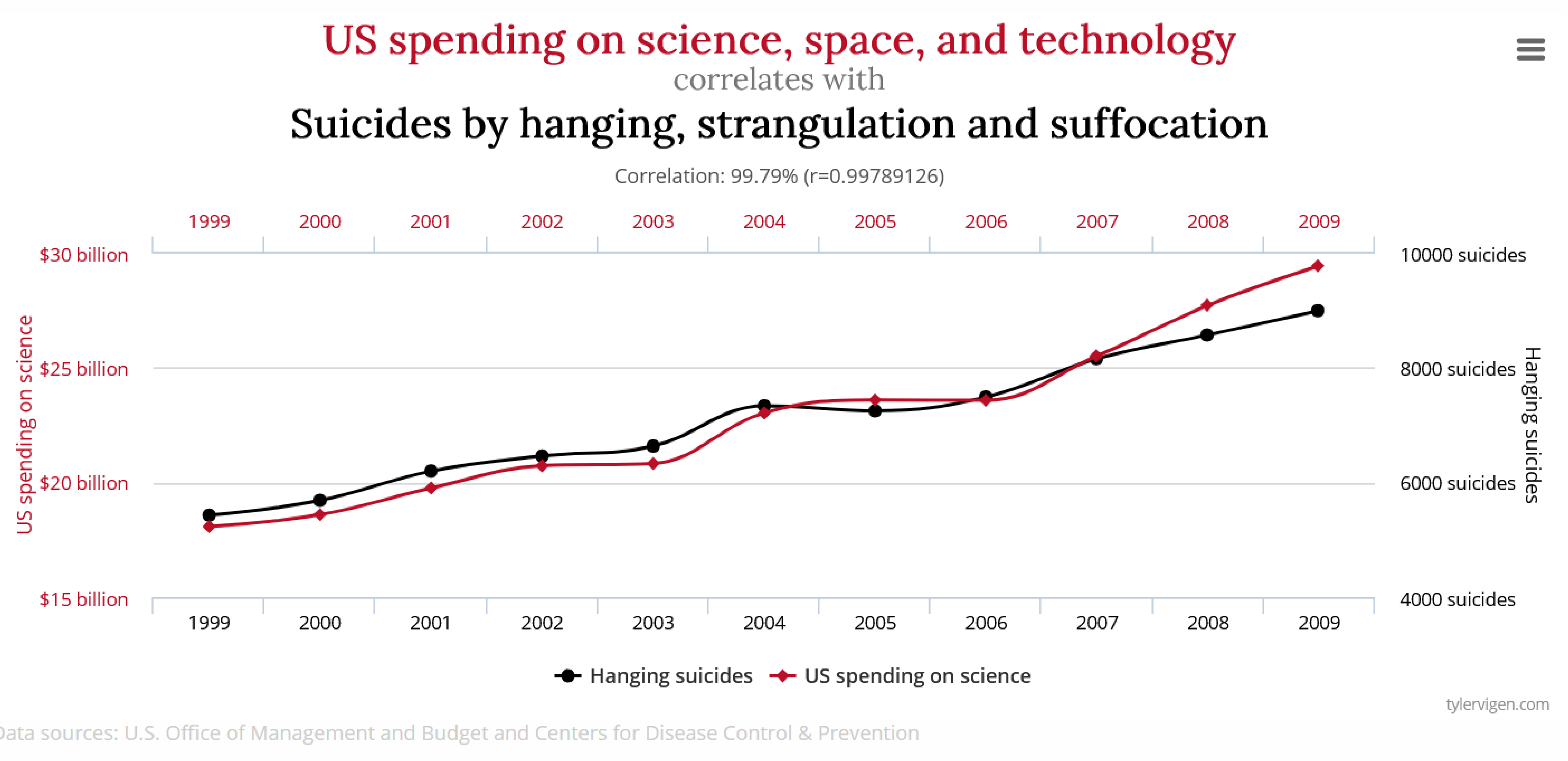
  - $\rho(X, Y) < 0$

    "The larger X is, the smaller Y tends to be"

WHAT IS CORRELATION



WHAT IS CORRELATION NOT



WATCH THE VIDEO

https://youtu.be/6RzDMEW5omc

SPURIOUS CORRELATIONS

More:

http://www.tylervigen.com/spurious-correlations

WHAT IS CORRELATION NOT