Week 8:
Statistical tests involving two variables (part II)

Phase 3

Recap from previous week

- Sometimes we want to make inferences that involve one categorical variable and one continuous variable (e.g differences in means between two or more populations).
- In this case, we could use a **two-sample t-test** (2 populations) or a **one-way ANOVA** (>= 2 populations).
- For ANOVA, we may need to run **post-hoc analyses** followed by a procedure to keep Type I error, α, under control.
- In this lecture, we will cover testing the relationship between **pairs of categorical variables** and **pairs of continuous variables**.

Your dataset

Inferences based on the relation between two variables

gender	education	age	ACT	SATV	SATQ
2	3	19	24	500	500
2	3	23	35	600	500
2	3	20	21	480	470
1	4	27	26	550	520
1	2	33	31	600	550
1	5	26	28	640	640
2	5	30	36	610	500
1	3	19	22	520	560
2	4	23	22	400	600
2	5	40	35	730	800

Your dataset

Inferences based on the relation between two variables

e.g. are there differences in ACT scores between men and women?

gender	education	age	ACT	SATV	SATQ
2	3	19	24	500	500
2	3	23	35	600	500
2	3	20	21	480	470
1	4	27	26	550	520
1	2	33	31	600	550
1	5	26	28	640	640
2	5	30	36	610	500
1	3	19	22	520	560
2	4	23	22	400	600
2	5	40	35	730	800

Your dataset

Inferences based on the relation between two variables

e.g. do men and women taking ACT exams exhibit differences in their education levels?

gender	education	age	ACT	SATV	SATQ
2	3	19	24	500	500
2	3	23	35	600	500
2	3	20	21	480	470
1	4	27	26	550	520
1	2	33	31	600	550
1	5	26	28	640	640
2	5	30	36	610	500
1	3	19	22	520	560
2	4	23	22	400	600
2	5	40	35	730	800

Your dataset

Inferences based on the relation between two variables

e.g. do people of greater ages tend to score ACT exams better?

gender	education	age	ACT	SATV	SATQ
2	3	19	24	500	500
2	3	23	35	600	500
2	3	20	21	480	470
1	4	27	26	550	520
1	2	33	31	600	550
1	5	26	28	640	640
2	5	30	36	610	500
1	3	19	22	520	560
2	4	23	22	400	600
2	5	40	35	730	800

Categorical vs categorical variable

Categorical vs categorical variable

- Here, instead of comparing means across levels of one of the categorical variables, we compare their **proportions**.
- Key concept: Each level in the categorical variable represents a **different population** (e.g. healthy and disease, education levels, etc)
- The most famous tests in this scenario are a χ^2**-test of independence** and the Fisher exact test.

Categorical vs categorical variable

<u>Research question</u>

do men and women taking ACT exams exhibit differences in their education levels?

gender	education	age	ACT	SATV	SATQ
2	3	19	24	500	500
2	3	23	35	600	500
2	3	20	21	480	470
1	4	27	26	550	520
1	2	33	31	600	550
1	5	26	28	640	640
2	5	30	36	610	500
1	3	19	22	520	560
2	4	23	22	400	600
2	5	40	35	730	800

Categorical vs categorical variable: Contingency Table

Research question
do men and women taking ACT exams exhibit differences in their education levels?

	1	2	Total
0	27	30	57
1	20	25	45
2	23	21	44
3	80	195	275
4	51	87	138
5	46	95	141
Total	247	453	700

Categorical vs categorical variable: Contingency Table

Notation:

- O_{ij} = Observed occurrences of the category *i* of the first variable in the category *j* of the second variable
- R_i = number of the category *i* of the first variable
- C_j = of the category *j* of the second variable

	1	2	Total
0	O_{11}	O_{12}	R_1
1	O_{21}	O_{22}	R_2
2	O_{31}	O_{32}	R_3
3	O_{41}	O_{42}	R_4
4	O_{51}	O_{52}	R_5
5	O_{61}	O_{62}	R_6
Total	C_1	C_2	N

χ^2-test of independence

- The hypothesis of a χ^2-test of independence is that there is no association between two categorical variables.
- Null hypothesis: the two categorical variables are independent of each other.
- Alternative hypothesis: there is a dependence between the two categorical variables.
- Evaluated by testing whether the number of occurrences in one variable change with the number of of occurrences in the other variable.
- Example: Is there a significant association between the size of the hippocampus (small, medium, or large) and memory performance (poor, fair, or excellent) in older adults?

χ^2-test of independence

Assumptions:

1. <u>Independence</u>: Observations in your sample are not correlated with each other (e.g. in *cross-sectional* studies).

2. <u>Sufficiently large expected frequencies</u>.

    Rule of thumb: expected frequencies larger than about 5, or at least 80% of the the expected frequencies are above 5 and none of them are below 1 (larger tables)

χ^2-test of independence

Assumptions:

1. Independence: Observations in your sample are not correlated with each other (e.g. in *cross-sectional* studies). What if they are? → McNemar test!!!

2. Sufficiently large expected frequencies.

   Rule of thumb: expected frequencies larger than about 5, or at least 80% of the the expected frequencies are above 5 and none of them are below 1 (larger tables)

χ^2-test of independence

- Null Hypothesis $\mathbf{H_0}$: the two variables are independent of each other.

- Test statistic: $\chi^2 = \sum_{i=1}^{r} \sum_{j=1}^{c} \frac{(E_{ij} - O_{ij})^2}{E_{ij}}$

- Alternative Hypothesis $\mathbf{H_A}$:

  there is a significant association between the two variables

χ^2-test of independence

- Null Hypothesis **H_0**: the two variables are independent of each other.

$$E_{ij} = \frac{C_j \times R_i}{N}$$

No input, we have to estimate this from the data!

- Test statistic: $\chi^2 = \sum_{i=1}^{r} \sum_{j=1}^{c} \frac{(E_{ij} - O_{ij})^2}{E_{ij}}$

- Alternative Hypothesis **H_A**:

  there is a significant association between the two variables

χ^2-test of independence

- Null Hypothesis $\mathbf{H_0}$: the two variables are independent of each other.

$$E_{ij} = \frac{C_j \times R_i}{N}$$

No input, we have to estimate this from the data!

- Test statistic: $\chi^2 = \sum_{i=1}^{r} \sum_{j=1}^{c} \frac{(E_{ij} - O_{ij})^2}{E_{ij}}$ ~ $\square^2$ distribution (df)

- Alternative Hypothesis $\mathbf{H_A}$:

  there is a significant association between the two variables

χ^2-test of independence

- Null Hypothesis $\mathbf{H_0}$: the two variables are independent of each other.

$$E_{ij} = \frac{C_j \times R_i}{N}$$

- Test statistic: $\chi^2 = \sum_{i=1}^{r} \sum_{j=1}^{c} \frac{(E_{ij} - O_{ij})^2}{E_{ij}}$ ~ χ^2 distribution (df)

$\boxed{df=(r-1)\cdot(c-1)}$

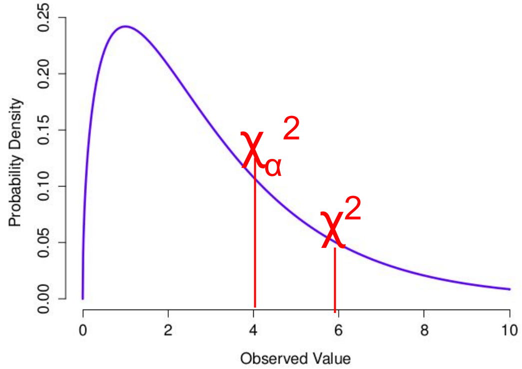
Read 12.1.5 from the book…

- Alternative Hypothesis $\mathbf{H_A}$:

there is a significant association between the two variables

χ^2-test of independence

- Null Hypothesis $\mathbf{H_0}$: the two variables are independent of each other.



- Test statistic: $$\chi^2 = \sum_{i=1}^{r}\sum_{j=1}^{c}\frac{(E_{ij}-O_{ij})^2}{E_{ij}}$$

qchisq(α, df=*(r-1)·(c-1)*, lower.tail =FALSE)

- Alternative Hypothesis $\mathbf{H_A}$:

Significant association between the two variables

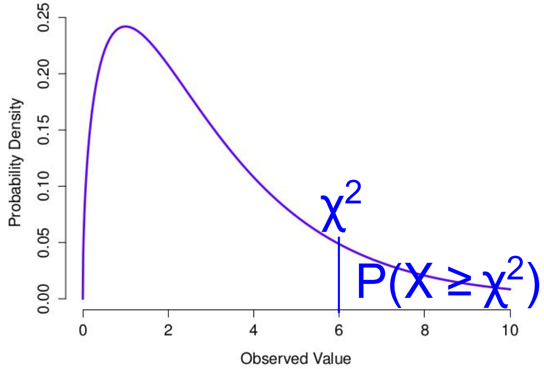
<u>Rejection region for α</u>

$\chi^2 \geq \chi^2_{\alpha,k-1}$

χ^2-test of independence

- Null Hypothesis $\mathbf{H_0}$: the two variables are independent of each other.

- Test statistic: $\chi^2 = \sum_{i=1}^{r} \sum_{j=1}^{c} \frac{(E_{ij} - O_{ij})^2}{E_{ij}}$



$\boxed{\textit{pchisq}(\chi^2,\ \text{df}=(r\text{-}1)\cdot(c\text{-}1),\ \text{lower.tail}=\text{FALSE})}$

- Alternative Hypothesis $\mathbf{H_A}$:

  Significant association between the two variables

<u>Rejection for α in P-values</u>

$P(X \geq \chi^2) \leq \alpha$

χ^2-test of independence

- Null Hypothesis $\mathbf{H_0}$: the two variables are independent of each other.

- Test statistic: $\chi^2 = \sum_{i=1}^{r}\sum_{j=1}^{c}\frac{(E_{ij}-O_{ij})^2}{E_{ij}}$

***R**: chisq.test(O_{ij})*

Input as table, see tutorial!!

- Alternative Hypothesis $\mathbf{H_A}$:

  Significant association between the two variables

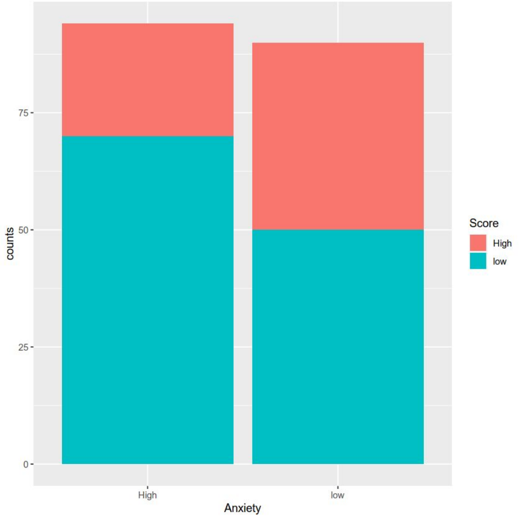
<u>Rejection for α in P-values</u>

$P(X \geq \chi^2) \leq \alpha$

Practice example

<u>Problem</u>: Relationship between anxiety and test performance in college students.

	High Anxiety	Low Anxiety	Total
High Score	24	40	64
Low Score	70	50	120
Total	90	90	184

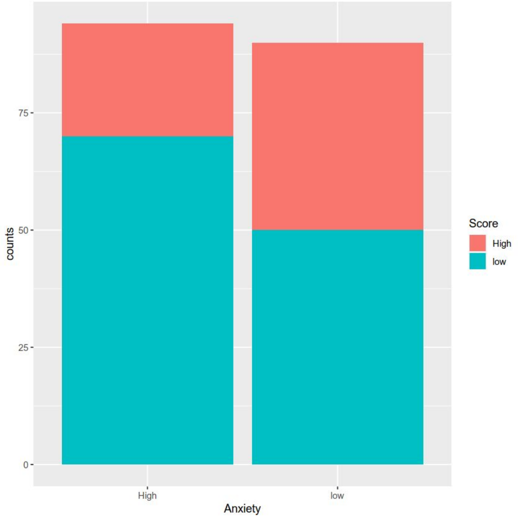


<u>Question</u>: At a significance level of 0.05, are students with high anxiety levels more likely to perform poorly on tests compared to students with low anxiety levels? (Hint: $\chi^2_{0.05,1} \sim 3.84$)

Practice example

<u>Problem</u>: Relationship between anxiety and test performance in college students.

	High Anxiety	Low Anxiety	Total
High Score	24 (32.7)	40 (31.3)	64
Low Score	70 (61.3)	50 (52.7)	120
Total	94	90	184



<u>Question</u>: At a significance level of 0.05, are students with high anxiety levels more likely to perform poorly on tests compared to students with low anxiety levels? (Hint: $\chi^2_{0.05,1} \sim 3.84$) $\chi^2 \sim 7.25$ → Reject the null!!

Fisher's Test

- The χ^2-test works reasonably well when there are sufficiently large observations expected in each category.
- If our two variable has only two categories, that is, we have 2 x 2 contingency tables we could use the **Fisher's test**.

	1	2	Total
0	O_{11}	O_{12}	R_1
1	O_{21}	O_{22}	R_2
Total	C_1	C_2	N

Fisher's Test

- If our two variable has only two categories, that is, we have 2 x 2 contingency tables we could use the **Fisher's test**.
- Here probabilities are **exact**, so for small samples, it might be more powerful than the χ^2-test.
- Example: Are more proportions of smokers in men than in women? .
- It's a **two-tailed** test: we may test for greater, less, or unequal.

Fisher's Test

- If our two variable has only two categories, that is, we have 2 x 2 contingency tables we could use the **Fisher's test**.
- Here probabilities are **exact**, so for small samples, it might be more powerful than the χ^2-test.
- Example: Are more proportions of smokers in men than in women? .
- It's a **two-tailed** test: we may test for greater, less, or unequal.

R: *fisher.test(O_{ij} , alternative=”greater”) or*

R: *fisher.test(X, Y, alternative=”greater”)*

Fisher's Test

- If our two variable has only two categories, that is, we have 2 x 2 contingency tables we could use the **Fisher's test**.
- Here probabilities are **exact**, so for small samples, it might be more powerful than the χ^2-test.
- <u>Example</u>: Are more proportions of smokers in men than in women? .
- It's a **two-tailed** test: we may test for greater, less, or unequal.

R: *fisher.test(O_{ij} , alternative="less") or*

R: *fisher.test(X, Y, alternative="less")*

Fisher's Test

- If our two variable has only two categories, that is, we have 2 x 2 contingency tables we could use the **Fisher's test**.
- Here probabilities are **exact**, so for small samples, it might be more powerful than the χ^2-test.
- Example: Are more proportions of smokers in men than in women? .
- It's a **two-tailed** test: we may test for greater, less, or unequal.

R: *fisher.test(O_{ij}, alternative="two.sided") or*

R: *fisher.test(X, Y, alternative="two.sided")*

Paired data

- What if your categorical variables are **paired** (e.g. in longitudinal studies)?
- In this case we can't use either χ^2-test or the Fisher test, because they assume that observations are not correlated.
- In this case, we can use the **McNemar test**.

Paired data

<u>Example</u>: a study investigates the effectiveness of a mindfulness intervention for reducing symptoms of depression. Participants are recruited and complete a depression symptom questionnaire before and after the intervention. The study aims to determine whether the proportion of participants who report a decrease in depression symptoms after the intervention is significantly different from the proportion who report no change or an increase in symptoms.

	After Intervention: Reduced Anxiety	After Intervention: No Change/Increased Anxiety
Before Intervention: High Anxiety	20	10
Before Intervention: Low Anxiety	5	15

McNemar Test

- The test is only applicable to a 2 × 2 contingency table. For example:

	Before: Yes	Before: No	Total
After: Yes	a	b	$a+b$
After: No	c	d	$c+d$
Total	$a+c$	$b+d$	n

- <u>Null hypothesis</u>: row totals and column totals come from the same distribution, i.e. $p_a + p_b = p_c + p_d$ and $p_c + p_d = p_b + p_d$

McNemar Test

➢ The test is only applicable to a 2 × 2 contingency table. For example:

	Before: Yes	Before: No	Total
After: Yes	a	b	$a+b$
After: No	c	d	$c+d$
Total	$a+c$	$b+d$	n

➢ <u>Null hypothesis</u>: row totals and column totals come from the same distribution, i.e. $p_a + p_b = p_c + p_d$ and $p_c + p_d = p_b + p_d$

$$\boxed{p_b = p_c}$$

Only the off-diagonal entries matter!

χ^2 test of independence

- Null Hypothesis **H_0**: $p_b = p_d$

- Test statistic: $\chi^2 = \frac{(b-c)^2}{b+c}$

- Alternative Hypothesis **H_A**:

  $p_b \neq p_d$

χ^2 test of independence

- Null Hypothesis $\mathbf{H_0}$: $p_b = p_d$

- Test statistic: $\chi^2 = \frac{(b-c)^2}{b+c}$ ~ □^{2}-distribution (df=1)

- Alternative Hypothesis $\mathbf{H_A}$:

  $p_b \neq p_d$

χ^2 test of independence

- Null Hypothesis **H_0**: $p_b = p_d$

- Test statistic: $\chi^2 = \frac{(b-c)^2}{b+c}$ ~ $\square^2$-distribution (df=1)

- Alternative Hypothesis **H_A**:

  $p_b \neq p_d$

R: *mcnemar.test(O_{ij}) or*

R: *mcnemar.test(X, Y)*

Paired data

<u>Example</u>: a study investigates the effectiveness of a mindfulness intervention for reducing symptoms of depression. Participants are recruited and complete a depression symptom questionnaire before and after the intervention. The study aims to determine whether the proportion of participants who report a decrease in depression symptoms after the intervention is significantly different from the proportion who report no change or an increase in symptoms.

	After Intervention: Reduced Anxiety	After Intervention: No Change/Increased Anxiety
Before Intervention: High Anxiety	20	10
Before Intervention: Low Anxiety	5	15

Paired data

<u>Example</u>: a study investigates the effectiveness of a mindfulness intervention for reducing symptoms of depression. Participants are recruited and complete a depression symptom questionnaire before and after the intervention. **For a type I error α=0.05, can we say mindfulness intervention was effective for reducing symptoms of depression? (Hint: $\square^{2}_{0.05,1} \sim 3.84$)**

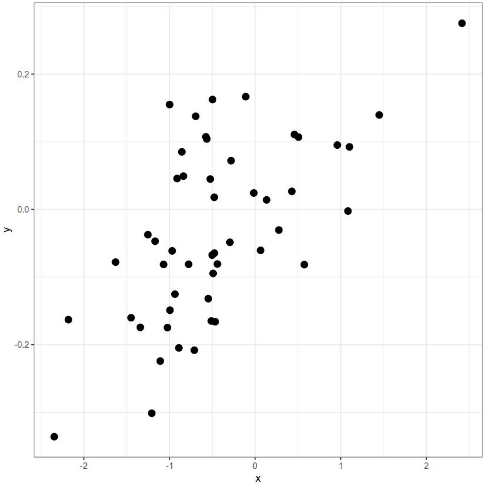
	After Intervention: Reduced Anxiety	After Intervention: No Change/Increased Anxiety
Before Intervention: High Anxiety	20	10
Before Intervention: Low Anxiety	5	15

Continuous vs Continuous variable

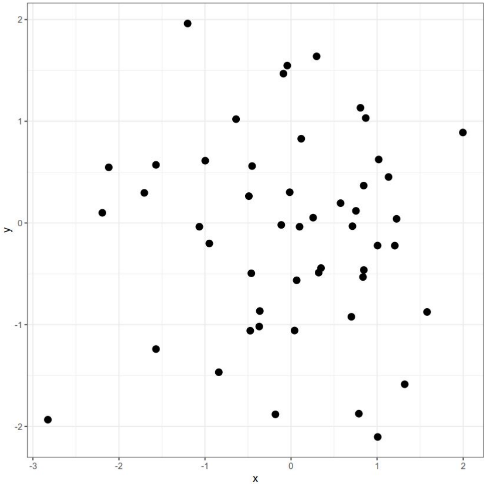
Continuous vs continuous variable

- ➢ Here, we want to test whether one continuous variable changes as the other continuous variable changes as well.
- ➢ We are mainly going to use a test for association based on **Pearson's correlation.**

Association



No Association



Continuous vs continuous variable

e.g. do people of greater ages tend to score ACT exams better?

gender	education	age	ACT	SATV	SATQ
2	3	19	24	500	500
2	3	23	35	600	500
2	3	20	21	480	470
1	4	27	26	550	520
1	2	33	31	600	550
1	5	26	28	640	640
2	5	30	36	610	500
1	3	19	22	520	560
2	4	23	22	400	600
2	5	40	35	730	800

Covariance and correlation

- The **covariance** is the generalization of the variance and it is just the average cross-product between two variables X and Y:

$$Cov(X,Y) = \frac{1}{N-1}\sum_{i=1}^{N}(X_i - <X>)(Y_i - <Y>)$$

- It is a measure **co**-variability.
- **Property:** if X and Y are entirely unrelated, Cov(X,Y) is exactly zero.

Covariance and correlation

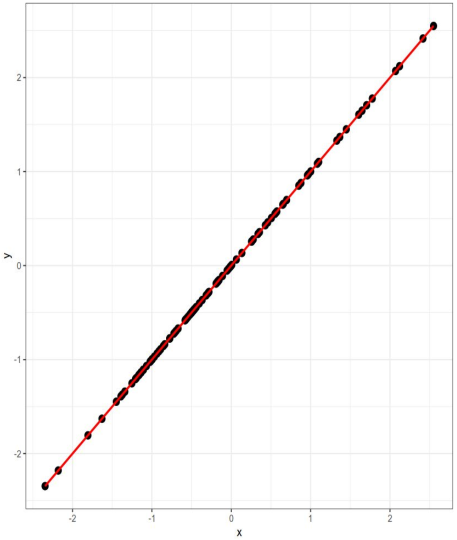
➢ The Pearson's **correlation,** r_{XY} , is just the standardization of the covariance:

$$r_{XY} = \frac{Cov(X,Y)}{\hat{\sigma}_X \hat{\sigma}_Y}$$

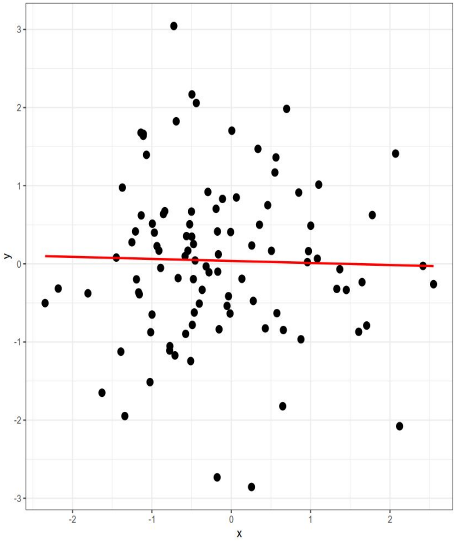
➢ **Properties:**

A. r_{XY} is symmetrical
B. The value r_{XY} is independent of the units of X and Y (WHY?)
C. $-1 \leq r_{XY} \leq 1$

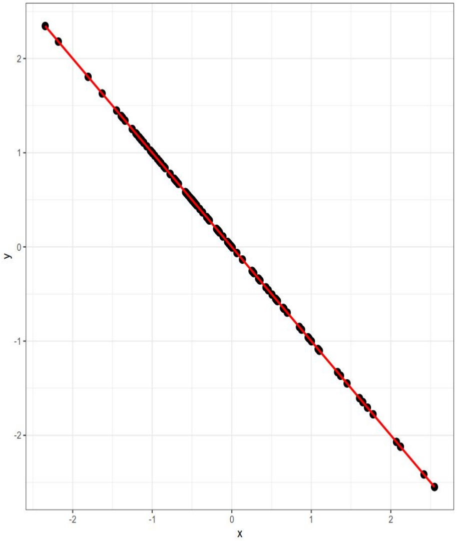
Pearson's correlation



$r_{XY} = 1$; perfect association

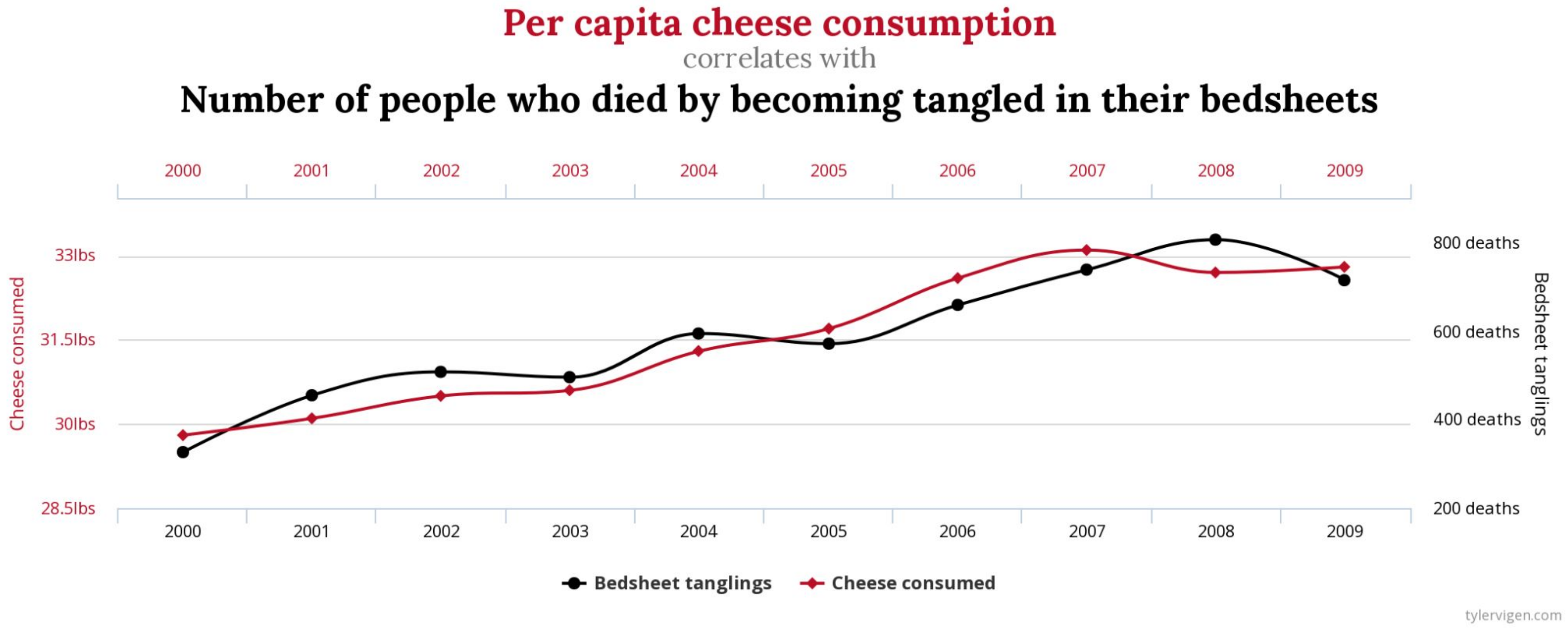


$r_{XY} = 0$; no association



$r_{XY} = -1$; perfect anti-association

WARNING: Correlation does not imply causation!



Test for Pearson's correlation

- It tests whether there is a significant correlation between two variables.

$$H_0: r_{XY}=0$$

$$H_A: r_{XY} > 0 \text{ or } r_{XY} < 0 \text{ (one-sided)}$$

$$r_{XY} \neq 0 \text{ (two-sided)}$$

- Example: Is there a significant correlation between brain volume and cognitive performance in older adults?

Test for Pearson's correlation

- Null Hypothesis $\mathbf{H_0}$: $r_{XY} = 0$

- Test statistic: $t = \dfrac{r_{XY}\sqrt{N-2}}{\sqrt{1-r_{XY}^2}}$

- Alternative Hypothesis $\mathbf{H_A}$

  $r_{XY} > 0$ (one-sided right tail)

  $r_{XY} < 0$ (one-sided left tail)

  $r_{XY} \neq 0$ (two-sided)

Test for Pearson's correlation

- Null Hypothesis $\mathbf{H_0}$: $r_{XY} = 0$

- Test statistic: $t = \frac{r_{XY}\sqrt{N-2}}{\sqrt{1-r_{XY}^2}}$ ~ Student's t (df=N-2)

- Alternative Hypothesis $\mathbf{H_A}$

  $r_{XY} > 0$ (one-sided right tail)

  $r_{XY} < 0$ (one-sided left tail)

  $r_{XY} \neq 0$ (two-sided)

Test for Pearson's correlation

- Null Hypothesis $\mathbf{H_0}$: $r_{XY} = 0$

- Test statistic: $t = \frac{r_{XY}\sqrt{N-2}}{\sqrt{1-r_{XY}^2}}$ ~ Student's t (df=N-2)

- Alternative Hypothesis $\mathbf{H_A}$ | Rejection region for α

Alternative Hypothesis $\mathbf{H_A}$	<u>Rejection region for α</u>
$r_{XY} > 0$ (one-sided right tail)	$t \geq t_{\alpha,N-2}$
$r_{XY} < 0$ (one-sided left tail)	$t \leq t_{\alpha,N-2}$
$r_{XY} \neq 0$ (two-sided)	$\lvert t \rvert \geq \lvert t_{\alpha/2,N-2} \rvert$

Test for Pearson's correlation

- Null Hypothesis $\mathbf{H_0}$: $r_{XY} = 0$

- Test statistic: $t = \dfrac{r_{XY}\sqrt{N-2}}{\sqrt{1-r_{XY}^2}}$

$\boxed{\textit{cor.test}\text{(X, Y, alternative="greater")}}$

- Alternative Hypothesis $\mathbf{H_A}$

Alternative Hypothesis $\mathbf{H_A}$	Rejection region for α
$r_{XY} > 0$ (one-sided right tail)	$P(T \geq t \mid H_0) \leq \alpha$
$r_{XY} < 0$ (one-sided left tail)	$t \leq t_{\alpha,N-2}$
$r_{XY} \neq 0$ (two-sided)	$\lvert t\rvert \geq \lvert t_{\alpha/2,N-2}\rvert$

Test for Pearson's correlation

- Null Hypothesis $\mathbf{H_0}$: $r_{XY} = 0$

- Test statistic: $t = \dfrac{r_{XY}\sqrt{N-2}}{\sqrt{1-r_{XY}^2}}$

$\boxed{\textit{cor.test}\text{(X, Y, alternative="less")}}$

- Alternative Hypothesis $\mathbf{H_A}$ — <u>Rejection region for α</u>

Alternative Hypothesis $\mathbf{H_A}$	Rejection region for α
$r_{XY} > 0$ (one-sided right tail)	$t \geq t_{\alpha,N-2}$
$r_{XY} < 0$ (one-sided left tail)	$P(T \leq t \mid H_0) \leq \alpha$
$r_{XY} \neq 0$ (two-sided)	$\lvert t\rvert \geq \lvert t_{\alpha/2,N-2}\rvert$

Test for Pearson's correlation

- Null Hypothesis $\mathbf{H_0}$: $r_{XY} = 0$

- Test statistic: $t = \dfrac{r_{XY}\sqrt{N-2}}{\sqrt{1-r_{XY}^2}}$

cor.test(X, Y, alternative=”two.sided”)

- Alternative Hypothesis $\mathbf{H_A}$ — Rejection region for α

Alternative Hypothesis $\mathbf{H_A}$	Rejection region for α
$r_{XY} > 0$ (one-sided right tail)	$t \geq t_{\alpha,N-2}$
$r_{XY} < 0$ (one-sided left tail)	$t \leq t_{\alpha,N-2}$
$r_{XY} \neq 0$ (two-sided)	$P(T \geq \lvert t \rvert \mid H_0) \leq \alpha$

Practice question

Problem: Is there an association between brain activity in the amygdala and self-reported anxiety levels?

Data analysis: In order to address this question, a lab collected functional magnetic resonance imaging (fMRI) data from **10** subjects while completing a task designed to activate the amygdala. The lab also measured the subjects' self-report anxiety levels through a questionnaire. After data collection, the lab calculated the Pearson correlation between the mean activity in the amygdala during the task and the self-reported anxiety scores. The result was **r ~ 0.69**.

Question: At a significance level 0.05, can they conclude that higher levels of amygdala activity associates with higher levels of reported anxiety? (Hint: $t_{0.05,8} \sim 1.85$)

Recap

- ➢ **Two categorical** variables and sufficiently **large expected frequencies** → χ^2 test
- ➢ **Small sample** sizes and only **2 x 2 contingency** tables → Fisher's test
- ➢ **Paired** data and only **2 x 2 contingency** tables → Mcnemar test
- ➢ For **two continuous** variable → Test of association for Pearson's correlation
- ➢ Next week we will study a generalization of all of these and previous tests: **Regression**!