Introduction to data science & artificial intelligence (INF7100)

Arthur Charpentier

#241 Testing & *p*-values

été 2020

Testing

Drunkometer, How Police Nab Drunk Drivers.

Z-test (*how different is different*)

The z-statistic measures how many standard deviations away the observed value is from its expectation,

$$z = \frac{\text{observed} - \text{expected}}{\text{standard deviation}} \quad (\text{compare with } \mathcal{N}(0,1))$$

	randomized	
	size	rate*
treatment	200,745	28
control	201,229	71
no consent	338,778	46

	randomized	
	size	number
treatment	200,000	57
control	200,000	142
no consent	350,000	92

Hypothesis (H_0): groups have the same chance of getting polio (against ($H1$) chances are different)

$X_\cdot$: number of observed cases in a group, $X_\cdot \sim \mathcal{N}(np_\cdot, np_\cdot(1-p_\cdot))$

here $p_\cdot$ small, $\text{Var}(X_\cdot) \simeq np_\cdot$

X_t: number of observed cases in the treatment group

X_c: number of observed cases in the control group

Z-test

	randomized	
	size	rate*
treatment	200,745	28
control	201,229	71
no consent	338,778	46

	randomized	
	size	number
treatment	200,000	57
control	200,000	142
no consent	350,000	92

Two groups have similar size (n), $X_c - X_t \sim \mathcal{N}(\star, n(p_c + p_t))$

$$z = \frac{142 - 57}{\sqrt{142 + 57}} \simeq 6.1$$

(very unlikely under H_0, since Z should follow a $\mathcal{N}(0, 1)$)

χ^2-test

$$X^2 = \sum_{j=1}^{k} \frac{\big((\text{observed number of } i) - (\text{expected number of } i)\big)^2}{(\text{expected number of } i)}$$

compare with χ^2_{k-1}

	dice value					
	1	2	3	4	5	6
observed	4	6	17	16	8	9
expected	10	10	10	10	10	10

Hypothesis (H_0): dice is fair (against ($H1$) dice is unfair)

$$X^2 = \frac{6^2}{10} + \frac{4^2}{10} + \frac{7^2}{10} + \frac{6^2}{10} + \frac{2^2}{10} + \frac{1^2}{10} \simeq 14.2$$

The probability of getting a probability of 14.2 with a χ^2_5 is 1.4%

χ^2-test

$$X^2 = \sum_{j=1}^{k} \frac{\big(\text{observed number of } i) - (\text{expected number of } i)\big)^2}{(\text{expected number of } i)}$$

	observed			expected ($\perp$)	
	men	women	total	men	women
right-handed	934	1070	2004	956	1048
left-handed	113	92	205	98	107
ambidextrous	20	8	28	13	15
total	1067	1170	2237	1067	1170

$$n \cdot \mathbb{P}(N_{rm}^{\perp}) = n \cdot \mathbb{P}(N_r)\mathbb{P}(N_m) = n\frac{n_r}{n}\frac{n_m}{n} = 2237\frac{2004}{2237}\frac{1067}{2237} \simeq 956$$

Hypothesis: left-handedness equally common for men and women

$$X^2 = \frac{22^2}{956} + \frac{22^2}{1048} + \frac{15^2}{98} + \frac{15^2}{107} + \frac{7^2}{13} + \frac{7^2}{15} \simeq 12$$

The probability of getting a probability of 12 with a χ^2_2 is 0.2%

Acceptation / Rejection Regions

Consider n coin flipping. We observed 55% tails. Is the coin biased?

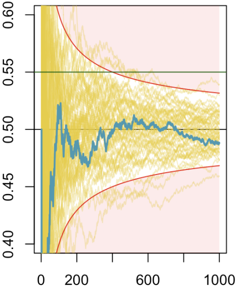
Not biased (H_0) means $p = p_0 = 50\%$

Under H_0, $\overline{x} \sim \mathcal{N}\left(\frac{1}{2}, \frac{1}{4n}\right)$, i.e. if H_0 is true

with 95% chance, $\overline{x} \in \left[\frac{1}{2} \pm \frac{1}{\sqrt{n}}\right]$

hence, 55% belongs to that interval if $\frac{1}{\sqrt{n}} \geq 5\%$

i.e. $n \leq 400$

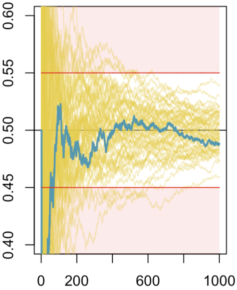


Equivalently, $z = 2\sqrt{n}\left(\overline{x} - \frac{1}{2}\right) \sim \mathcal{N}(0,1)$

We reject H_0 if $|z| > 2$ (or 1.96).

p-value

Consider n coin flipping. We observed 55% tails. Is the coin biased?



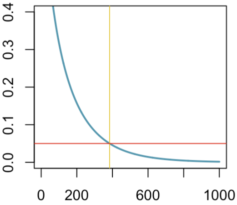
Conversely, we can compute

$$\mathbb{P}(|\overline{X}| > 55\%) \text{ when } \overline{X} \sim \mathcal{N}\left(\frac{1}{2}, \frac{1}{4n}\right)$$

called p-value.

We reject H_0 if $p < 5\%$.

If we could replicate experiments of this sample size, how often will we see a statistic this extreme, assuming that H_0 is true ?



Surgery versus Radiation Therapy

Source McNeil *et al.* (1982)
Form A: Of 100 people having surgery, 10 will die during treatment, 32 will have died by one year, and 66 will have died by five years. Of 100 people having radiation therapy, none will die during treatment, 23 will die by one year, and 78 will die by five years.

Form B: Of 100 people having surgery, 90 will survive the treatment, 68 will survive one year or longer, and 34 will survive five years or longer. Of 100 people having radiation therapy, all will survive the treatment, 77 will survive one year or longer, and 22 will survive five years or longer.

	Forms	
	A	B
surgery	40	73
radiation	40	14
total	80	87
surgery	50%	84%

Surgery versus Radiation Therapy

Let $\widehat{p}_A$ and $\widehat{p}_B$ be the empirical frequency favoring surgery.

- $\widehat{p}_A$ is (roughly) normally distributed, with mean p_A and standard deviation $\sqrt{p_A(1-p_A)/n}$, that can be approximated by $\sqrt{\widehat{p}_A(1-\widehat{p}_A)/n} \simeq \sqrt{0.5^2/80} = 0.056$,
- $\widehat{p}_B$ is (roughly) normally distributed, with mean p_B and standard deviation $\sqrt{p_B(1-p_B)/n}$, that can be approximated by $\sqrt{\widehat{p}_B(1-\widehat{p}_B)/n} \simeq \sqrt{0.84 \cdot 0.16/87} = 0.039$,

Assuming that $p_A = p_B$ (assumption H_0), $\widehat{p}_A - \widehat{p}_B$ is (roughly) normally distributed, with mean 0 and standard deviation approximated by $\sqrt{0.056^2 + 0.039^2} = 0.068$.

$$Z = \frac{\widehat{p}_A - \widehat{p}_B}{0.068} = \frac{0.50 - 0.84}{0.068} = -5$$

Regression Test and Significance

In a (simple) linear regression, $y = \alpha + \beta x + \varepsilon$

$$\widehat{\beta} = \frac{\sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^n (x_i - \bar{x})^2} = \text{cor}[x, y]\sqrt{\frac{s_y^2}{s_x^2}}$$

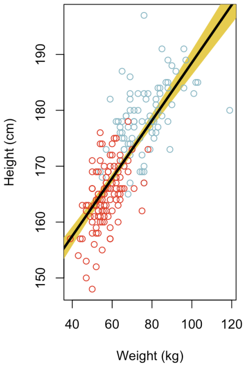
then

$$z = \frac{\widehat{\beta} - \beta}{s_{\widehat{\beta}}} \sim \mathcal{N}(0, 1),$$

where $s_{\widehat{\beta}} = \sqrt{\frac{\sum_{i=1}^n \widehat{\varepsilon}_i^2}{(n-2)\sum_{i=1}^n (x_i - \bar{x})^2}}$.

We want to test $H_0 : \beta = 0$ (x is not significant)

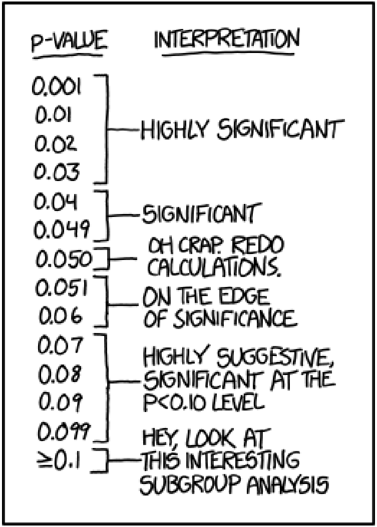
If $\left|\frac{\widehat{\beta}}{s_{\widehat{\beta}}}\right| > 2$, then $p < 5\%$ and x is significant



The "$p < 5\%$" Dogma

"if p is between 10% and 90% there is certainly no reason to suspect the hypothesis tested. If it is below 2% it is strongly indicated that the hypothesis fails to account for the whole of facts [...] We shall not often be astray if we draw a conventional line at 5%"
Ronald Fisher

see La guerre des étoiles, *p*-value and statistical practice or It's time to talk about ditching statistical significance



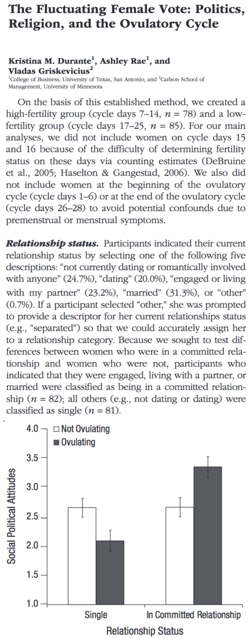
P-VALUE	INTERPRETATION
0.001	HIGHLY SIGNIFICANT
0.01	
0.02	
0.03	
0.04	SIGNIFICANT
0.049	
0.050	OH CRAP. REDO CALCULATIONS.
0.051	ON THE EDGE OF SIGNIFICANCE
0.06	
0.07	HIGHLY SUGGESTIVE, SIGNIFICANT AT THE P<0.10 LEVEL
0.08	
0.09	
0.099	
≥0.1	HEY, LOOK AT THIS INTERESTING SUBGROUP ANALYSIS

p-hacking (?)

See The Fluctuating Female Vote: Politics, Religion, and the Ovulatory Cycle, Durante, Rae & Griskevicius (2013)
Dataset = survey, and several criteria where considered to get a *valid* subset

- Exclusion criteria based on cycle length (3)
- Exclusion criteria based on "How sure are you?" response (2)
- Cycle day assessment (3)
- Fertility assessment (4)
- Relationship status assessment (3)

(from Andrew Gelman's slides)



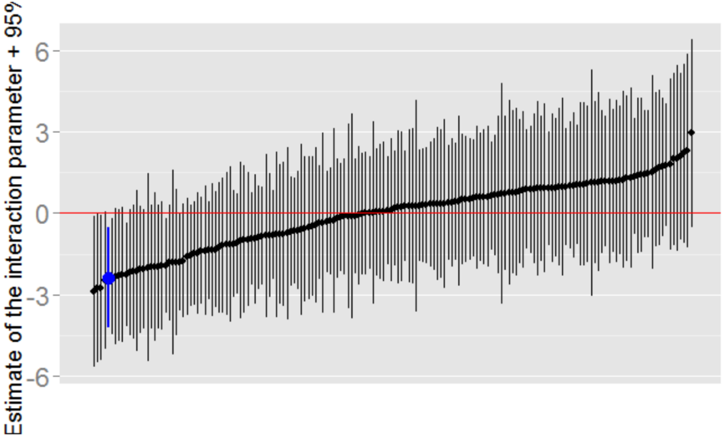
The Fluctuating Female Vote: Politics, Religion, and the Ovulatory Cycle

Kristina M. Durante[1], Ashley Rae[1], and Vladas Griskevicius[2]
[1]College of Business, University of Texas, San Antonio, and [2]Carlson School of Management, University of Minnesota

On the basis of this established method, we created a high-fertility group (cycle days 7–14, n = 78) and a low-fertility group (cycle days 17–25, n = 85). For our main analyses, we did not include women on cycle days 15 and 16 because of the difficulty of determining fertility status on these days via counting estimates (DeBruine et al., 2005; Haselton & Gangestad, 2006). We also did not include women at the beginning of the ovulatory cycle (cycle days 1–6) or at the end of the ovulatory cycle (cycle days 26–28) to avoid potential confounds due to premenstrual or menstrual symptoms.

Relationship status. Participants indicated their current relationship status by selecting one of the following five descriptions: "not currently dating or romantically involved with anyone" (24.7%), "dating" (20.0%), "engaged or living with my partner" (23.2%), "married" (31.3%), or "other" (0.7%). If a participant selected "other," she was prompted to provide a descriptor for her current relationships status (e.g., "separated") so that we could accurately assign her to a relationship category. Because we sought to test differences between women who were in a committed relationship and women who were not, participants who indicated that they were engaged, living with a partner, or married were classified as being in a committed relationship (n = 82); all others (e.g., not dating or dating) were classified as single (n = 81).

p-hacking (?)



See The garden of forking paths: Why multiple comparisons can be a problem, even when there is no “fishing expedition” or “p-hacking” and the research hypothesis was posited ahead of time