Hypothesis Testing

Big Data Lectures – Chapter 2

association between two random quantities

Lexin Li

Division of Biostatistics
University of California, Berkeley

Outline

- list of topics:
  - a comparative summary of statistical hypothesis tests
  - A/B testing
  - multiple testing
  - causal inference from observational data

Hypothesis Testing: A Comparative Summary

Hypothesis testing – a comparative summary

- summary:
  - test statistical significance of **relation** between random variables
  - this relation is about **association**, but not **causation**
- **confounding, confounding, confounding!**
- key elements of hypothesis testing:
  - a pair of hypothesis – formulate the right questions
  - test statistic – a measure of compatibility between the data and the null hypothesis – the smaller the (absolute) test statistic, the more compatible
  - *p*-value – a statistical significance measure of compatibility (reference: significance level α) – the bigger the p-value, the more compatible
  - sample size calculation – power / type I and type II error
- crucial to know **which test to use** given the data and the question

Hypothesis testing – a comparative summary

- one random variable:
  - one continuous – one sample t test
  - one discrete – χ^2 goodness-of-fit test
- two random variables:
  - one continuous *vs* one discrete (2 levels)
    - independent samples – two sample t test
    - independent but small samples – Wilcoxon-Mann-Whitney test

paired (i.e. dependence)

    - paired samples – paired t test
    - paired but small samples – Wilcoxon signed rank test
  - one continuous *vs* one discrete (> 2 levels) – one-way ANOVA
  - one discrete *vs* one discrete
    - 2-level *vs* 2-level – 2×2 contingency table
    - 2-level *vs* 2-level but small counts – Fisher's exact test
    - r-level *vs* c-level – $r \times c$ contingency table
    - paired samples – McNemar's test
  - one continuous *vs* one continuous – correlation/simple linear regression
- more than two random variables:
  - one discrete *vs* one discrete *vs* one continuous – two-way ANOVA
  - one continuous *vs* many – multiple linear regression
  - one discrete *vs* many – logistic regression

Example: χ^2 goodness-of-fit test

- example:
  - a cross between white and yellow summer squash gave progeny of the following colors
  - Q: consistent with the 12:3:1 ratio predicted by a genetic model?

COLOR	WHITE	YELLOW	GREEN
Number of progeny	155	40	10

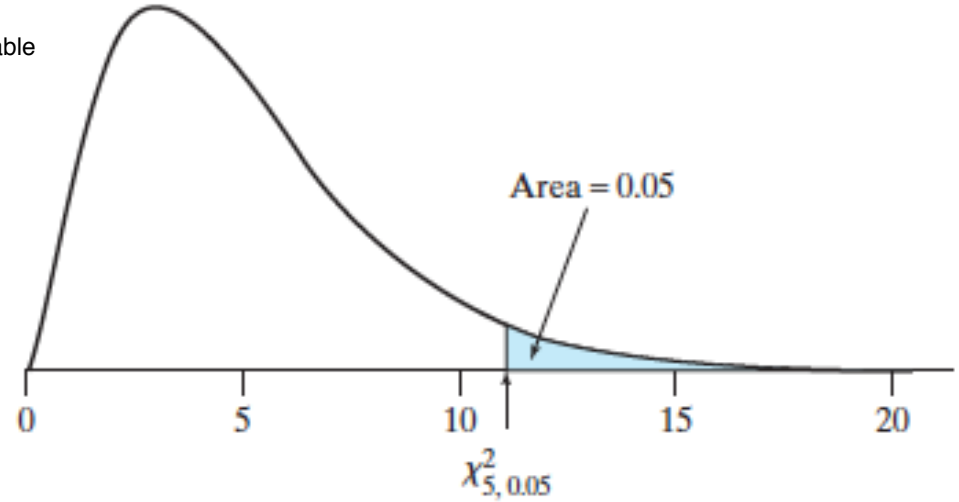
whats the random quantity we're looking at? colors of offspring
this random quantity can 3 different possible values
how many random quantities are we looking at? 1 random quantity (that can take 3 possible values)

- key observations:
  - test statistic: the chi-square test stat is a random variable

o - observed for some color
e - expected for some color

$$\chi_s^2 = \sum_{i=1}^{c} \frac{(o_i - e_i)^2}{e_i}$$

  - compare to χ^2 distribution with $c - 1$ df



Example: two-sample t test

- example:

two random quantities: blood flow and air-type
one is continuous (blood flow) and one is discrete (air-type - 2 levels)

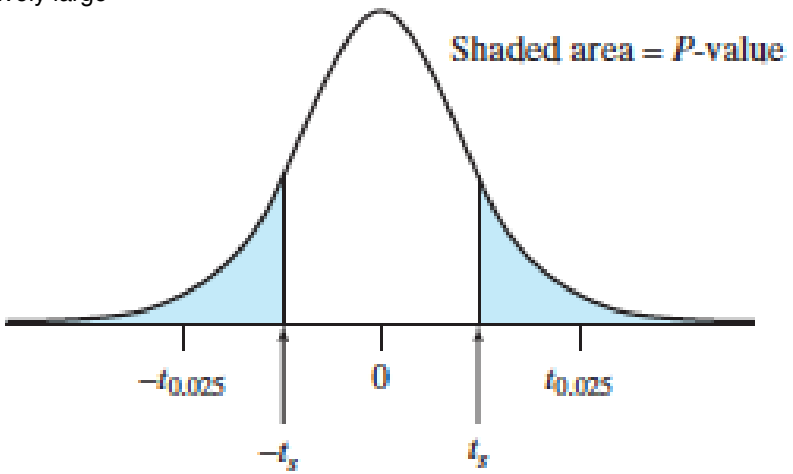
  - myocardial blood flow was measured for two groups of subjects after five minutes of bicycle exercise
  - "normoxia" group was provided normal air to breathe; "hypoxia" group air with reduced oxygen

	NORMOXIA	HYPOXIA
	3.45	6.37
	3.09	5.69
	3.09	5.58
	2.65	5.27
	2.49	5.11
	2.33	4.88
	2.28	4.68
	2.24	3.50
	2.17	
	1.34	
n	10	8
$\bar{y}$	2.51	5.14
s	0.60	0.84

- key observations:

2 key assumptions:
the data should follow an approximately normal distribution w/i each group
sample size from each group is relatively large

  - compare to t distribution with $n_1 + n_2 - 2$ df, approximately
  - assume a fully randomized design, so no other confounders
  - $n_1 \geq 20, n_2 \geq 20$
  - Wilcoxon-Mann-Whitney test is the nonparametric alternative



Example: paired t test

- example:

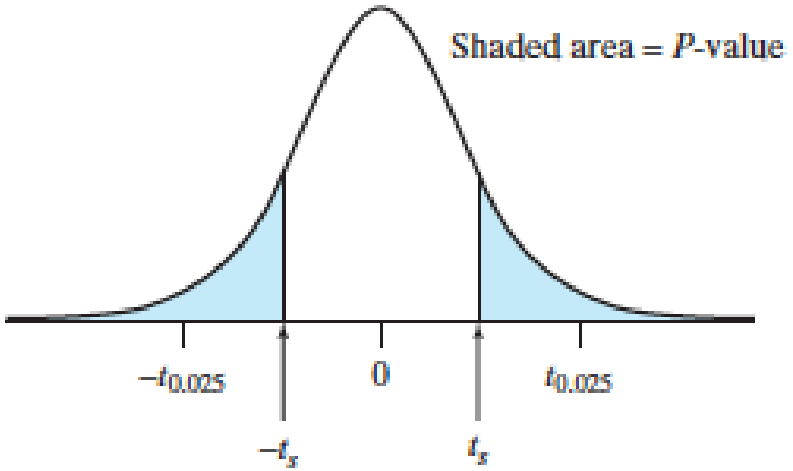
two random quantities: blood flow and caffeine intake
one is continuous (blood flow) and one is discrete (caffeine intake - 2 levels)

  - myocardial blood flow was measured during bicycle exercise before and after giving the subjects a dose of caffeine that was equivalent to drinking two cups of coffee
  - Q: drinking coffee affects blood flow during exercise?

behavior of subject with one treatment and behavior of same subject with a second treatment

Subject	MBF	
	Baseline y_1	Caffeine y_2
1	6.37	4.52
2	5.69	5.44
3	5.58	4.70
4	5.27	3.81
5	5.11	4.06
6	4.89	3.22
7	4.70	2.96
8	3.53	3.20
Mean	5.14	3.99
SD	0.83	0.86

- key observations:
  - check the difference; compare to t with $n_1 - 1$ df; $n_1 = n_2$
  - randomized pairs design
  - Wilcoxon signed rank test is the alternative nonparametric test



Example: one-way ANOVA

for > 2 levels of discrete variable

considering two levels with effect modification on top, could do ANOVA or two-sample t-test, would get the same result

- example:
  - compare the weights of ears of sweet corn under five treatment conditions

one continuous - weights and one discrete - treatment with 5 possible values

sample variation is not that different (there are tests for this)
independence assumption
large-ish sample size or normality

	Treatment				
	1	2	3	4	5
	16.5	11.0	8.5	16.0	13.0
	15.0	15.0	13.0	14.5	10.5
	11.5	9.0	12.0	15.0	11.0
	12.0	9.0	10.0	9.0	10.0
	12.5	11.5	12.5	10.5	14.0
	9.0	11.0	8.5	14.0	12.0
	16.0	9.0	9.5	12.5	11.0
	6.5	10.0	7.0	9.0	9.5
	8.0	9.0	10.5	9.0	18.5
	14.5	8.0	10.5	9.0	17.0
	7.0	8.0	13.0	6.5	10.0
	10.5	5.0	9.0	8.5	11.0
Mean	11.5	9.6	10.3	11.1	12.3
SD	3.5	2.4	2.0	3.1	2.9
n	12	12	12	12	12

- key observations:
  - compare to F distribution with $I - 1$, $n_\cdot - I$ dfs
  - box-plot is the corresponding graphical representation
  - don't forget the assumptions!

ANOVA Quantities with Formulas

Source	df	SS (Sum of Squares)	MS (Mean Square)
Between groups	$I - 1$	$\sum_{i=1}^{I} n_i(\bar{y}_i - \bar{\bar{y}})^2$	SS/df
Within groups	$n_\cdot - I$	$\sum_{i=1}^{I} (n_i - 1)s_i^2$	SS/df
Total	$n_\cdot - 1$	$\sum_{i=1}^{I}\sum_{j=1}^{n_i} (y_{ij} - \bar{\bar{y}})^2$	

Example: contingency table

- example:
  - a clinical trial to assess an experimental surgery for patients who suffered from moderate to severe migraine headache
- key observations:
  - $\chi^2_s = \sum \frac{(o_i - e_i)^2}{e_i}$
  - $e = \frac{row\ total \times column\ total}{grand\ total}$
  - compare to χ^2 distribution with $(r-1) \times (c-1)$ df
  - assumption: okay to have some $1 \leq e < 5$ and at least 80% of e's are ≥ 5
  - alternative: Fisher's exact test

2 random variables - both discrete & each has 2 levels

		Surgery	
		Real	Sham
Substantial reduction in migraine headaches?	Success	41	15
	No success	8	11
	Total	49	26

	Surgery		
	Real	Sham	Total
Success	41 (36.59)	15 (19.41)	56
No success	8 (12.41)	11 (6.59)	19
Total	49	26	75

Example: McNemar's test

- example:
  - 114 HIV-infected women who gave birth to two children
  - Q: H0: the probability of HIV infection is the same for older and younger siblings

		Older sibling	Younger sibling
HIV?	Yes	19	20
	No	95	94
	Total	114	114

- key observations:
  - test statistic:

$$\chi_s^2 = \frac{(n_{12} - n_{21})^2}{n_{12} + n_{21}}$$

  - compare to χ^2 distribution with 1 df, approximately

		Younger sibling HIV?	
		Yes	No
Older sibling	Yes	2	17
HIV?	No	18	77

A/B Testing & Multiple Testing

A/B testing

= fully randomized experimental design

- what it is:
  - a methodology in advertising using randomized experiments with two variants, A (control) and B (treatment)
  - commonly used in web development, marketing, and other forms of advertising
  - the goal is to identify changes to web pages that increase or maximize an outcome of interest, e.g., click-through rate for an advertisement
  - two versions (A and B) are compared, which are identical except for one variation that might impact a user's behavior

- what to know:
  - how to design an A/B test
  - decide test size
  - interpret test results
  - understand error bound

Multiple testing

- why a problem:
  - setup: test a number of, say m, hypotheses **simultaneously**
  - probability of claiming at least one significant result while all results are truly insignificant:

$$\begin{aligned} & \Pr(\text{at least one result claimed significant}) \\ = \; & 1 - \Pr(\text{all results are claimed insignificant}) \\ = \; & 1 - (1-\alpha)^m \end{aligned}$$

  - $\alpha = 0.05$: if $m = 20$, this probability is 0.64; if $m = 100$, it is 0.99
- solutions:
  - Bonferroni correction; false discovery rate control; ...

Causal Inference from Observational Data

Causal inference from observational data

- example:
  - new drug / treatment / exposure
  - online ads display format / campaign

- counterfactual framework:
  - Z = a binary indicator; 1 treated/exposed; 0 control
  - $\boldsymbol{X}$ = a vector of covariates measured prior to receipt of treatment (baseline) or, if measured post treatment, not affected by either treatment
  - **counterfactual responses / potential outcome:**
    Y_1 = the response value that would be seen if, possibly contrary to the fact of what actually happened, the subject were to receive **treatment**
    Y_0 = the response value that would be seen if, possibly contrary to the fact of what actually happened, the subject were to receive **control**
  - **actually observed response:**

$$Y = ZY_1 + (1 - Z)Y_0$$

Causal inference from observational data

- quantity of interest:
  - **causal effect:**

$$\Delta = \mu_1 - \mu_0 = E(Y_1) - E(Y_0)$$

  - sample average response: does this give us what we want?

$$E(Y|Z = 1), \quad E(Y|Z = 0)$$

Causal inference from observational data

- quantity of interest:
  - **causal effect:**
$$\Delta = \mu_1 - \mu_0 = E(Y_1) - E(Y_0)$$
  - sample average response: does this give us what we want?
$$E(Y|Z=1), \quad E(Y|Z=0)$$
- **randomized study**:
  Z is determined for each subject at random, so it is unrelated to how s/he might potentially respond
$$(Y_0, Y_1) \perp\!\!\!\perp Z$$

Causal inference from observational data

- quantity of interest:
  - **causal effect:**
$$\Delta = \mu_1 - \mu_0 = E(Y_1) - E(Y_0)$$
  - sample average response: does this give us what we want?
$$E(Y|Z=1), \quad E(Y|Z=0)$$
- **randomized study**:
  Z is determined for each subject at random, so it is unrelated to how s/he might potentially respond
$$(Y_0, Y_1) \perp\!\!\!\perp Z$$
  therefore,
$$E(Y|Z=1) = E(Y_1|Z=1) = E(Y_1), \; E(Y|Z=0) = E(Y_0|Z=0) = E(Y_0)$$
- **observational data?**

Causal inference from observational data

- observational data:
  - assumption: **no unmeasured confounders**

$$(Y_0, Y_1) \perp\!\!\!\perp Z | \boldsymbol{X}$$

  - **propensity score**: the probability of treatment given the observed covariates

$$e(\boldsymbol{X}) = P(Z = 1 | \boldsymbol{X})$$

    then, under the no unmeasured confounders assumption,

$$(Y_0, Y_1) \perp\!\!\!\perp Z | e(\boldsymbol{X})$$

    so that treatment exposure is unrelated to the counterfactuals for individuals sharing the same propensity score

  - in practice, one estimates the propensity score, by imposing a model, e.g., a logistic regression model, on Z given $\boldsymbol{X}$, and hopes the model is correctly specified

Causal inference from observational data

- some common solutions:
  - **stratification**:
    - form K strata according to the sample quantiles of the estimated $\hat{e}(\boldsymbol{X})$
    - within each stratum, calculate the difference of sample means of the observed response Y
    - estimate Δ by a weighted sum of the differences of sample means across strata, where weighting is by the proportion of observations falling in each stratum

Causal inference from observational data

- some common solutions:
  - **stratification**:
    - form K strata according to the sample quantiles of the estimated $\hat{e}(\boldsymbol{X})$
    - within each stratum, calculate the difference of sample means of the observed response Y
    - estimate Δ by a weighted sum of the differences of sample means across strata, where weighting is by the proportion of observations falling in each stratum
  - **inverse probability weighting**:

still the same goal of estimating the ATE :
if we don't observed the response then first fraction is 0

$$\frac{1}{n}\sum_{i=1}^{n}\frac{Z_i Y_i}{\hat{e}_i} - \frac{1}{n}\sum_{i=1}^{n}\frac{(1-Z_i)Y_i}{1-\hat{e}_i}$$

    - why does this work?

$$E\left\{\frac{ZY}{e(\boldsymbol{X})}\right\} = E\left[E\left\{\frac{I(Z=1)Y_{\mathbf{1}}}{e(\boldsymbol{X})}|Y_{\mathbf{1}},\boldsymbol{X}\right\}\right] = E\left[\frac{Y_{\mathbf{1}}}{e(\boldsymbol{X})}E\left\{I(Z=1)|Y_{\mathbf{1}},\boldsymbol{X}\right\}\right] = E(Y_{\mathbf{1}})$$

Causal inference from observational data

- some common solutions:
  - **stratification**:
    - form K strata according to the sample quantiles of the estimated $\hat{e}(\boldsymbol{X})$
    - within each stratum, calculate the difference of sample means of the observed response Y
    - estimate Δ by a weighted sum of the differences of sample means across strata, where weighting is by the proportion of observations falling in each stratum
  - **inverse probability weighting**:

$$\frac{1}{n}\sum_{i=1}^{n}\frac{Z_i Y_i}{\hat{e}_i} - \frac{1}{n}\sum_{i=1}^{n}\frac{(1-Z_i)Y_i}{1-\hat{e}_i}$$

    - why does this work?

$$E\left\{\frac{ZY}{e(\boldsymbol{X})}\right\} = E\left[E\left\{\frac{I(Z=1)Y_{\mathbf{1}}}{e(\boldsymbol{X})}|Y_{\mathbf{1}},\boldsymbol{X}\right\}\right] = E\left[\frac{Y_{\mathbf{1}}}{e(\boldsymbol{X})}E\left\{I(Z=1)|Y_{\mathbf{1}},\boldsymbol{X}\right\}\right] = E(Y_{\mathbf{1}})$$

    - **issues**: what if $\hat{e}_i$ is close to 0 or 1? what if the model for $e(\boldsymbol{X})$ is incorrect? how to make sure the assumption is satisfied?

Additional readings

- Samuels, M.L., Witmer, J.A., and Schaffner, A.A. (2012). *Statistics for the Life Sciences*, 4th edition, Prentice Hall
- Wikipedia entry on "A/B testing"
- Free EBook: An Introduction to Using A/B Testing for Marketing Optimization, HubSpot
- Chan, D., Ge, R., Gershony, O., Hesterberg, T., and Lambert, D. (2010). Evaluating online Ad campaigns in a pipeline: causal models at scale. *Proceedings of KDD*, 7-16.
- Lunceford, J.K., and Davidian, M. (2004). Stratification and weighting via the propensity score in estimation of causal treatment effects: a comparative study. *Statistics in Medicine*, **23**, 2937-2960.