BITTIGER

DS 501 Data scientist express bootcamp

Week 2 [Ella]

版权声明

BITTIGER

Copyright Policy

All content included on the Site or third-party platforms as part of the class, such as text, graphics, logos, button icons, images, audio clips, video clips, live streams, digital downloads, data compilations, and software, is the property of BitTiger or its content suppliers and protected by copyright laws.

Any attempt to redistribute or resell BitTiger content will result in the appropriate legal action being taken.

We thank you in advance for respecting our copyrighted content.

For more info:
see https://www.bittiger.io/termsofuse
and https://www.bittiger.io/termsofservice

Open ended question

- FB launched new UI for news feed, but user interaction decreases, what would you do?
  - How should we proceed?
  - What questions are expected from us?
- General interview tips
  - Ask clarifying questions
  - Show positive attitude, brainstorm
  - Read interviewers' feedback, understand what interviewer is looking for, build chemistry

How to deal with open ended question?

- Understand the context
  - Which metric decreases? What about other metrics?
  - How much is decreased? Significant?
  - How many users? Same users or different users?
- Understand the goal
  - 1. Do we want to **verify the new UI is the cause**?
  - 2. Do we want to understand why new UI has such impact?
- How to achieve the goal
  - 1. Randomized experiments
  - 2. Check other metrics, slice dice users into groups, and etc

Causal inference

- Causality has always been studied
  - Does a decision to smoke cigarettes increase the likelihood of a person getting lung cancer?
  - Does number of homeless increase crime rate?
- Definition of causal inference
  - Process of drawing a conclusion about a causal connection
- Causal inference v.s. (statistical) inference
  - (Statistical) inference is drawn from observation
  - Causal inference cannot drawn from observation, but from experiment.

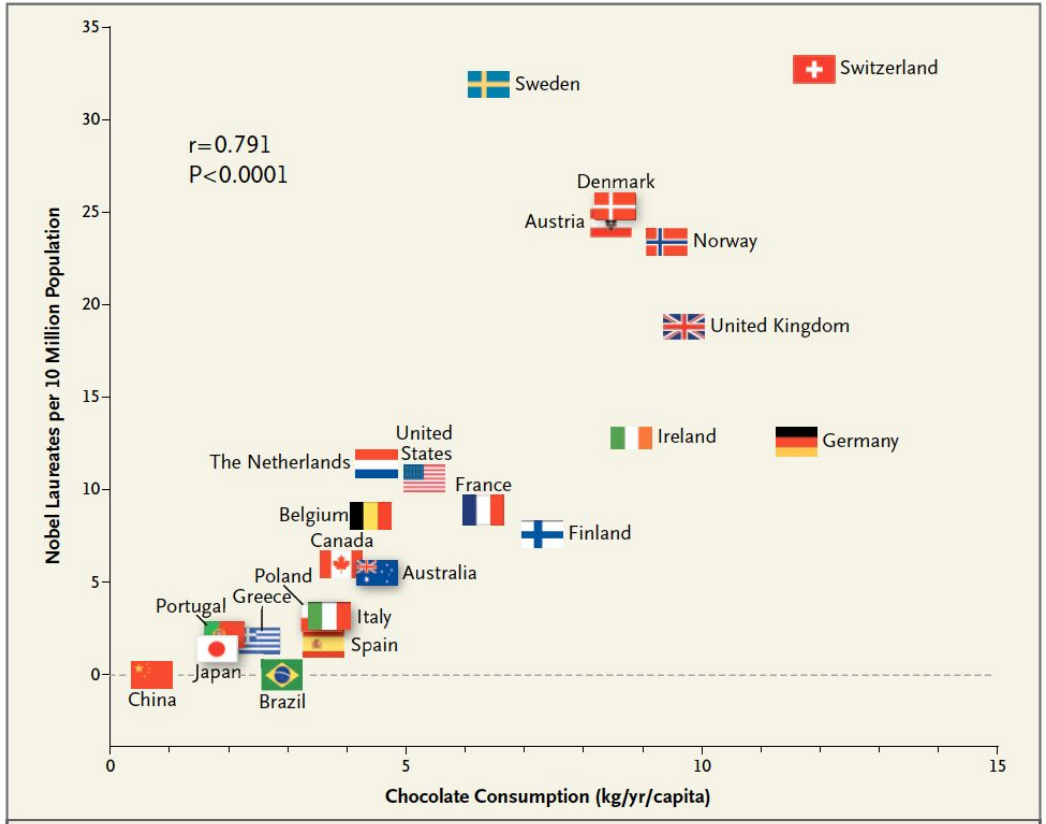
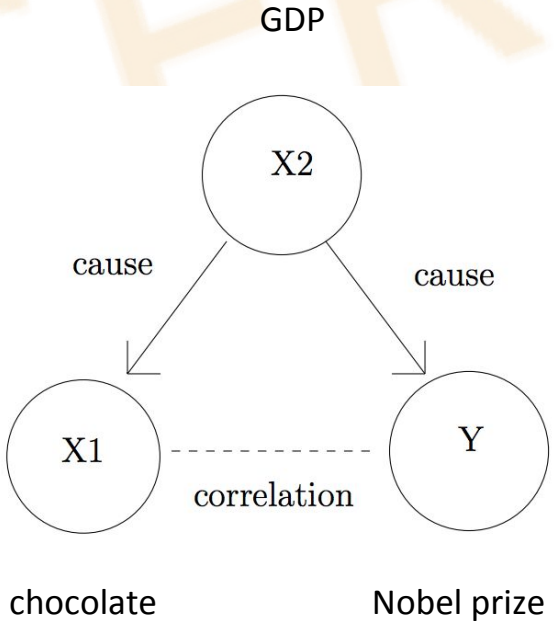


Figure 1. **Correlation between Countries' Annual Per Capita Chocolate Consumption and the Number of Nobel Laureates per 10 Million Population.**

Correlation is not causation

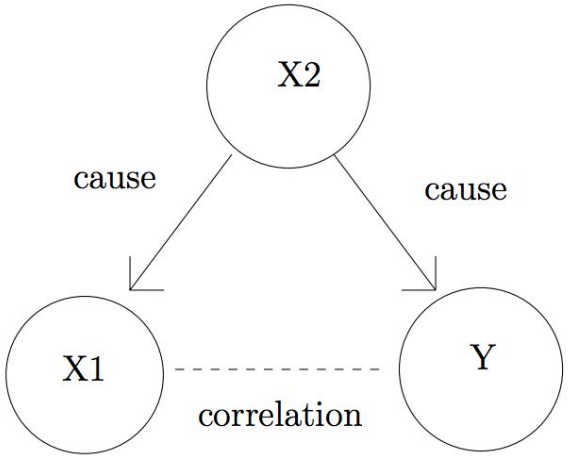
- Quote from chairman of Nobel chemistry committee
  - “Chocolate is a luxury. Wealthy individuals are more likely to be able to afford it.

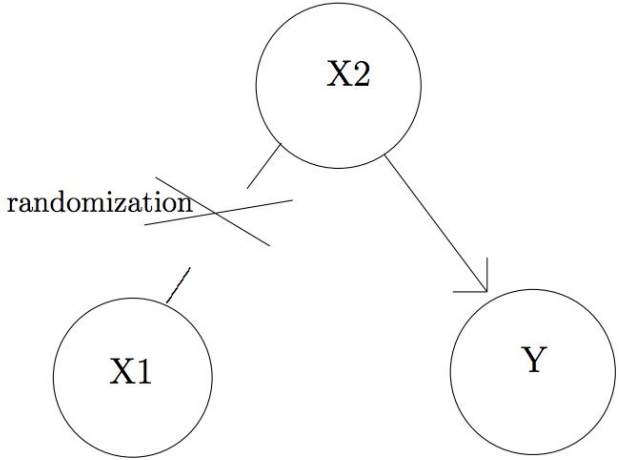
    Education is also a luxury. Poor people can’t afford to go to college for 10 years to get a PhD in chemistry. But you can’t win the Nobel prize in chemistry unless you’re a chemist. “
- Common factor
  - “GDP or wealth of a country will be correlated both with chocolate eating and with Nobel prizes. ”



Observational study v.s. Randomized experiment

- Observational studies can suggest good experiments to run, but can't definitively show causation.
- Randomization can eliminate correlation between X_1 and Y due to a different cause X_2(confounder).
  - FB example





A/B Testing

- Define the causal relationship to be explored
  - New UI decreases user interaction
- Define metric
  - Number of posts per user
- Design randomized experiments (A/B test)
  - Two groups of users, comparable
  - control group: old UI, experiment group: new UI
- Collect data and conduct **hypothesis testing**
  - Compare the metrics using two sample t test
- Draw conclusion

Hypothesis testing

- Definition
  - Use sample of data to test an assumption regarding a population parameter, which could be
    - A population mean μ
    - The difference in two population means, $\mu 1 - \mu 2$
    - A population variance
    - The ratio of two population variances
    - A population proportion p
    - The difference in two population proportions, $p1 - p2$
    - Three (or more!) means, $\mu 1$, $\mu 2$, and $\mu 3$

Hypothesis testing (cont'd)

- Definition
  - Two opposing hypotheses about a population
    - Null hypothesis, H_0 , is usually the hypothesis that sample observations result purely from chance.
    - Alternative hypothesis, H_1 or H_a, is the hypothesis that sample observations are influenced by some non-random cause.

Bet over flip coin, win $1 if head, what do you feel after playing 10 times and losing $10?

Flip coin example

1

$$X_1, \ldots, X_n \sim \text{Bernoulli}(p)$$

Given sample data

p is population statistic of our interest

2

$$H_0 : p = \frac{1}{2} \qquad H_1 : p \neq \frac{1}{2}.$$

Null hypothesis

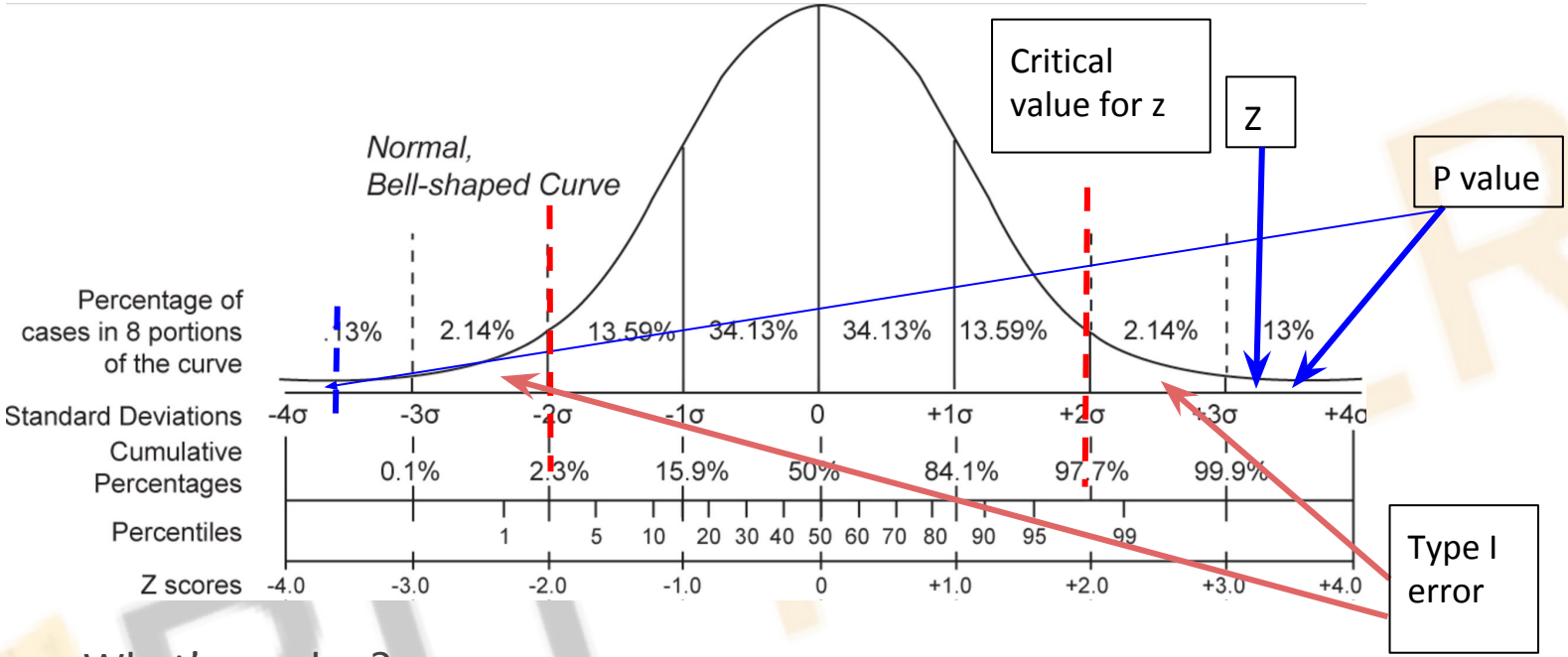
Alternative hypothesis

3 How to decide if we should reject or not reject null hypothesis?

- We say not reject (or retain) null hypothesis, but not say accept null hypothesis. Why?

How to decide reject or not reject?

- According to CLT, sample mean p = y/n = 1 following $N(\mu, \sigma^2/n)$
  - Population mean $p_0 = 0.5$, $\sigma^2/n = p_0(1 - p_0)/n$
  - So $z = (p - p_0)/\sqrt{(p_0(1 - p_0)/n)}$ following N(0, 1)
  - Is p so far from p_0 (or z too large) that true population mean p_0 is not 0.5 ? (when null hypothesis should be rejected.)
  - Two ways to decide
    - "Critical value" approach, compare z with critical value
    - "P value" approach, compare p value with threshold



- What's p value?
  - Assuming H_0 is true, what's the probability of observing a more extreme test stats in the direction of H_a
- No matter how large threshold is, there is still error of rejecting H_0 when we shouldn't (i.e. H_0 is true), type I error, α.
  - Type I error should be small, like 0.1, 0.05, 0.01.
  - Type I error is threshold for p value to compare with.
  - Type I error determines critical value for z score to compare with.

How to construct hypothesis testing

- Define H_0, H_a
- Decide which test statistics to be calculated, e.g., z score
  - Option 1

$$z = \frac{(\bar{x} - \mu)}{\sigma / \sqrt{n}}$$

    - Calculate z score and compare with critical value ($z^* = z_{1-\alpha/2}$)
    - If $|z| > z^*$, reject H_0; else do not reject H_0
  - Option 2
    - Calculate p value and compare with type I error (α)
    - If p is smaller, reject H_0; else do not reject H_0

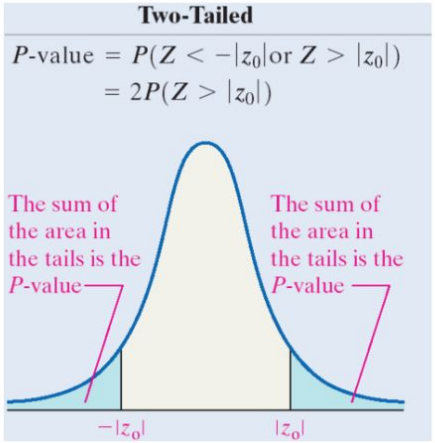
p value of z^* is α

- Other ways to define alternative hypothesis?

Different types of alternative hypothesis

- Two tailed v.s. one tailed

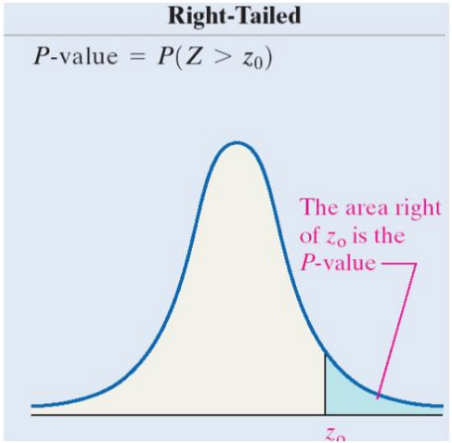
Two-tail



$H_0 : \mu = \mu_0$

$H_1 : \mu \neq \mu_0$

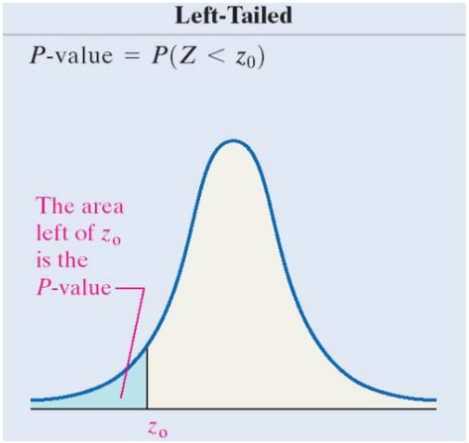
Right Tail



$H_0 : \mu = \mu_0$

$H_1 : \mu > \mu_0$

Left Tail



$H_0 : \mu = \mu_0$

$H_1 : \mu < \mu_0$

- When to use one sided test? Reading

Type I, II error, power

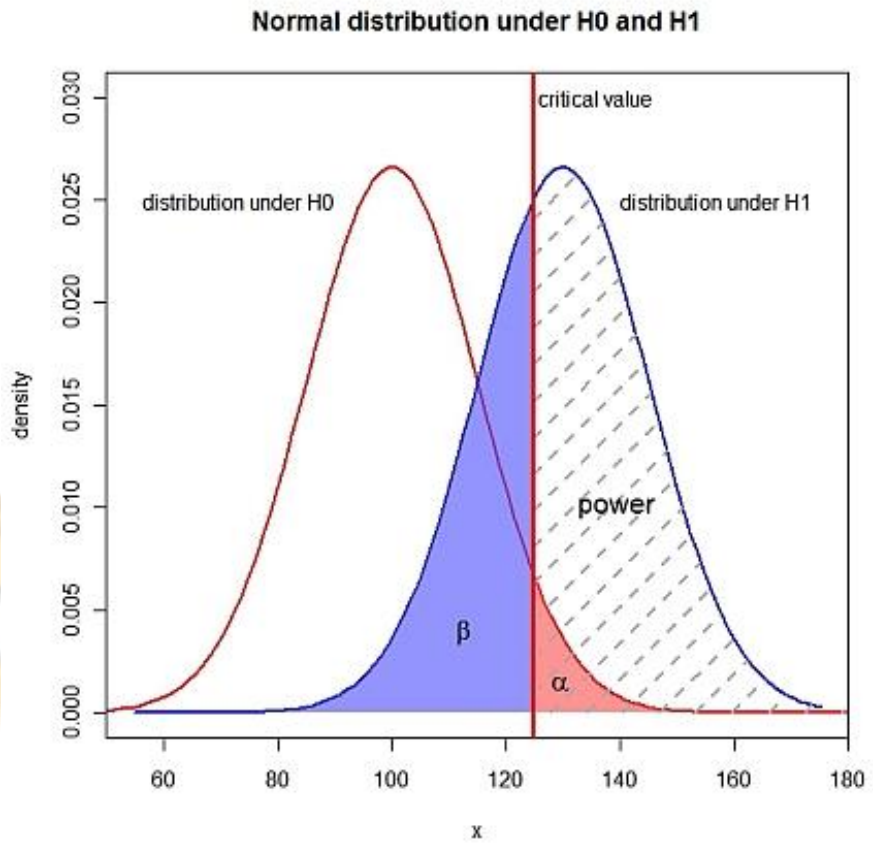
	Ground Truth	
Decision	H_0	H_a
Not reject H_0	Correctly (not reject null)	Type II error, β
Reject H_0	Type I error, α	Correctly (reject null), power

$$\alpha = P(reject\ H_0 \mid H_0)$$

$$\beta = P(not\ reject\ H_0 \mid H_a)$$

$$power = P(reject\ H_0 \mid H_a)$$

Type I, II error, power



- Power and power function
- Trade off between Type I error and Type II error
- How can we get small type I error and type II error?
- Sample size calculation

Difference between flip coin and FB example?

One sample v.s. Two sample tests

- One sample test
  - One population, compare test statistic, e.g, sample mean, with a known number
- Two sample test
  - Two populations, compare two population means
  - Paired test
    - Two dependent groups, for example, same group been measured at two different times.
    - Essentially one sample test
  - Unpaired test
    - Two independent groups, may have different sample sizes.

Z test v.s. T test

- Hypothesis test for the population mean
  - If population variance σ^2 is known, z test
  - If the population variance σ^2 is unknown (most of the time), or n is small, t test
    - Consider CLT
    - T score calculation, same as z score
    - Only that critical value $t^* = t_{1 - \alpha;\, n - 1}$ from t distribution, not normal distribution.
- One sample t test
- Two sample t test
  - Paired
  - Unpaired

Two sample t-test

- Compares the means of the two groups of data
  - X_1 random sample from $N(\mu_1, \sigma_1^2)$

    X_2 random sample from $N(\mu_2, \sigma_2^2)$
  - $H_0 : \mu_2 = \mu_1 \quad H_a : \mu_2 != \mu_1$ (other H_a?)

$$t = \frac{\bar{x}_1 - \bar{x}_2}{\sqrt{Var(\bar{x}_1 - \bar{x}_2)}} = \frac{\bar{x}_1 - \bar{x}_2}{\sqrt{Var(\bar{x}_1) + Var(\bar{x}_2)}} = \frac{\bar{x}_1 - \bar{x}_2}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}}$$

  - When $\sigma_1^2 = \sigma_2^2$

$$S^2_{\text{pooled}} = \frac{\sum_{k=1}^{n_1}(X_{1,k} - \bar{X}_1)^2 + \sum_{k=1}^{n_2}(X_{2,k} - \bar{X}_2)^2}{n_1 + n_2 - 2} = \frac{(n_1 - 1)S_1^2 + (n_2 - 1)S_2^2}{n_1 + n_2 - 2}$$

Student t test v.s. Welch t test

- If population variance from two samples are equal, use pooled variance (student t test)

$$t^* = \frac{\bar{x}_1 - \bar{x}_2}{s_p \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}} \qquad s_p = \sqrt{\frac{(n_1 - 1)s_1^2 + (n_2 - 1)s_2^2}{n_1 + n_2 - 2}} \qquad df = n_1 + n_2 - 2$$

- If population variance from two samples are not equal, use unpooled variance (Welch t test)

$$t^* = \frac{\bar{x}_1 - \bar{x}_2}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}} \qquad df = \frac{(n_1-1)\cdot(n_2-1)}{(n_2-1)C^2+(1-C)^2(n_1-1)}$$

$$C = \frac{s_1^2/n_1}{\frac{s_1^2}{n_1}+\frac{s_2^2}{n_2}}$$

Assumptions of t test

- Student t test
  - Normality
  - Independence
  - Equal variance (two sample test)
- What if normality is violated
  - Just do it (CLT)
  - Transformation
  - Other nonparametric methods
- What if equal variance is violated
  - Welch t test, preferred over student t test reading

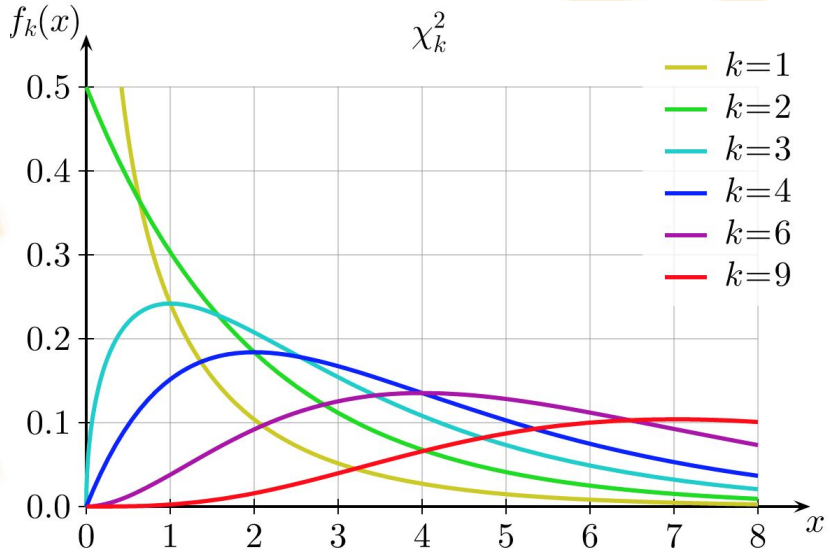
Test about variance

Chi square distribution

- Definition
  - Sum of square of k standard normal random variables, N (0, 1)

$$\chi^2(k) = \sum_{i=1}^{k} Z_i^2$$

  - k-1 degree of freedom



- Chi square test for one variance

$H_0 : \sigma^2 = \sigma_0^2$ $\quad H_A : \sigma^2 \neq \sigma_0^2, \quad H_A : \sigma^2 < \sigma_0^2,$ or $H_A : \sigma^2 > \sigma_0^2$

$$\chi^2 = \frac{(n-1)S^2}{\sigma_0^2}$$

What can Chi-square test be used for?

- One sample test
  - A population variance
  - Compare categorical variable with known distribution (goodness of fit)
  - Null and alternative hypothesis?
- Two sample test
  - Compare two population variance
  - Compare two categorical variables to test if they have different distribution.
  - Chi-square independence test
  - Null and alternative hypothesis?

Chi square test for goodness of fit

- Test if X follows certain distribution F, X is categorical var
  - H_0: X follows F
    H_a: X does not follow F
  - Calculate (chi-square) statistics
    - Recall t/z score, which measures the (normalized) distance between observed stats and expected stats given H_0 is true
    - Similarly, $chi\ square = \sum_{all\ cells} \frac{(observed - expected)^2}{expected}$
    - Flip coin example

	head	tail
observed	10	0
expected	5	5

$$\chi^2 = \frac{(10-5)^2}{5} + \frac{(5-0)^2}{5} = 10$$

Critical value chart

Connection between binomial, normal, chi square

- N trial, m success, p is success rate, q = 1 - p

	success	fail
observed	m	N-m
expected	Np	Nq

$$\chi^2 = \frac{(m - Np)^2}{(Np)} + \frac{(N - m - Nq)^2}{(Nq)}$$

$$\chi^2 = \frac{(m - Np)^2}{(Npq)}$$

$$N = Np + N(1 - p)$$
$$N = m + (N - m)$$
$$q = 1 - p$$

$$\chi = \frac{m - Np}{\sqrt{(Npq)}}$$

CLT

$$\chi^2 = \sum_{i=1}^{n} \frac{(O_i - E_i)^2}{E_i}$$

O_i = the number of observations of type *i*.
$E_i = N * p_i$, the expected frequency of type *i*,
n = the number of cells in the table.

Chi square test for independence

- Test if X, Y are independent
  - H_0: X and Y are independent.
    H_a: X and Y are dependent.
  - Example
    - Two types of bidder, human and computer. Different categories of bidding product. Test if bidder type is independent with product category.

$$chi\ square = \sum_{all\ cells} \frac{(observed - expected)^2}{expected}$$

How to calculate the expected stats?

Chi square test (cont'd)

	Auto parts	Books music	cloth	computer	furniture	home goods	jewelry	mobile	Office equip	Sporting goods	**Row sum**
Human	9757	13733	476	9733	87807	389249	555634	492350	160671	939398	2658808
Robot	0	1509	0	11667	0	18708	37101	105138	7967	230326	412416
column sum	9757	15242	476	21400	87807	407957	592735	597488	168638	1169724	

X_1 → Human

Y_1 → Robot

Col sum_1 → column sum

Row sum_1

Row sum_2

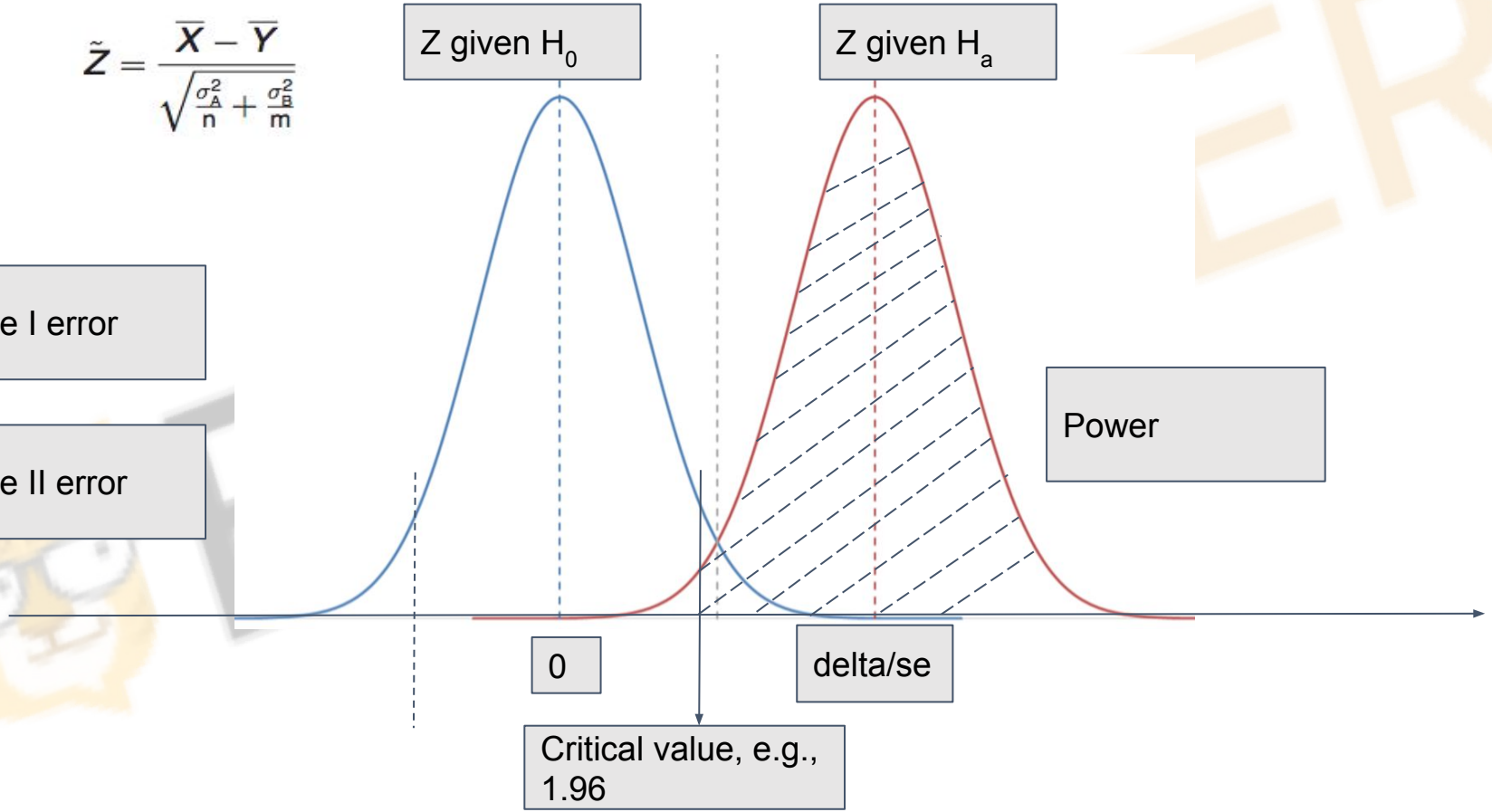
- If independent
  - $X_1 : Y_1$ = row sum_1 : row sum_2
  - $X_1 + Y_1$ = col sum_1
  - Expected value X_1 = col sum_1 * row sum_1 / (row sum_1 + row sum_2)

$$\chi^2 = \sum_i \frac{(O_i - E_i)^2}{E_i}$$

df: (row num - 1) (column num - 1)

One more word about power

$$\tilde{Z} = \frac{\overline{X} - \overline{Y}}{\sqrt{\frac{\sigma_A^2}{n} + \frac{\sigma_B^2}{m}}}$$



Power analysis (cont'd)

- Larger type I error (alpha), greater power
- Larger delta/se, greater power
  - Large delta
  - Smaller standard error
- Larger sample size (n, m), greater power
  - Mostly influence n, m to control power
- Check the relationships in R

- Suppose we want power = 0.8 to detect a particular value of $\mu_A - \mu_B = \Delta$ with known σ_A^2, σ_B^2, calculate n
  - $\Delta = 1$
  - $\sigma_A^2 = \sigma_B^2 = 1$
  - n = m

HW1

- Calculate n
- Check how n change along with Δ ?
- Check how n change along with σ_A^2 ?

HW3 Feature engineering

- How would you transform existing features if needed? Any new features could be generated?

- A new casino game involves rolling 3 dice. The winnings are directly proportional to the total number of sixes rolled. Suppose a gambler plays the game 101 times, with the following observed counts:

Number of Sixes	Number of Rolls
0	48
1	35
2	15
3	3

HW2: test if this is fair dice.

- What test to use?
- Calculate stats and p value