**Inferential Statistics**

Kunyu HE

**Week 1 Central Limit Theorem (CLT) and Confidence Interval**

* Sampling Variability

Sampling variability: the variation in the distribution of sample statistics (or point estimates of the population parameters)

Parameters of sampling distribution:

* Mean:
* Standard deviation:
* Central Limit Theorem

Conditions:

* Independent: random sampling / random assignment; n < 10%N;
* Sample size / skew: nearly normal distribution or n > 30.

Theorem:

* Confidence Intervals

Definition: plausible range of values of a population parameter; for sample statistics, their values are certain (either 100% or 0%)

Conditions: same conditions with CLT

Confidence Interval:

* Accuracy against precision:

;

To increase both, increase the sample size n, but within 10%N

**Week 2 Inference and Significance**

* Hypothesis

Null hypothesis: =, set population parameter equal to the null value

Alternative hypothesis: >, <, , range of the population parameter

Only when the sample satisfy the CLT conditions can we use following hypothesis test methods

* P-value:

p−value = P(observed or more extreme sample statistic | H0 true)

=

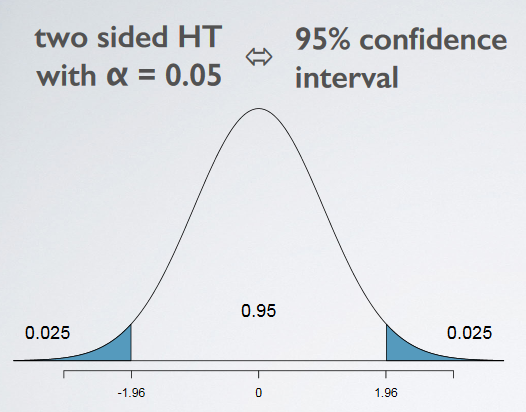
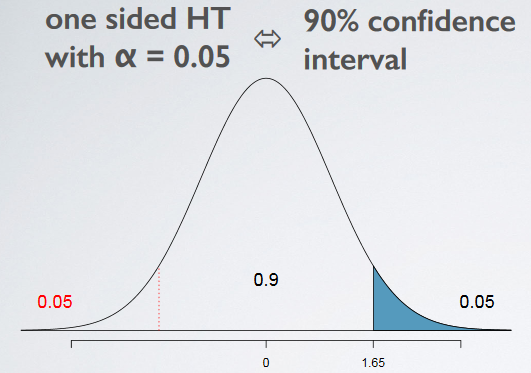
* Decision Errors

How to balance between the two kinds of decision errors: compare the cost of each;

depends on the effective level , which refers to the difference between the population parameter and the null value (the practical significance)



* Confidence Interval and Significance Level

* Statistical Significance and Practical Significance

Statistical significance: test statistic (for the case above, z = ) large enough;

Practical significance: the difference between the population parameter and the null value;

(Large sample would over-estimate the practical significance; in fact it matters only statistically)

##### Week 3 Inference for Comparing Means

* **T-distribution**

T-distribution: approximately normal with heavier tails; used when the population parameter unknown;

Degree of freedom (with sample mean as point estimate): n-1

Confidence interval:

Hypothesis test: critical value =

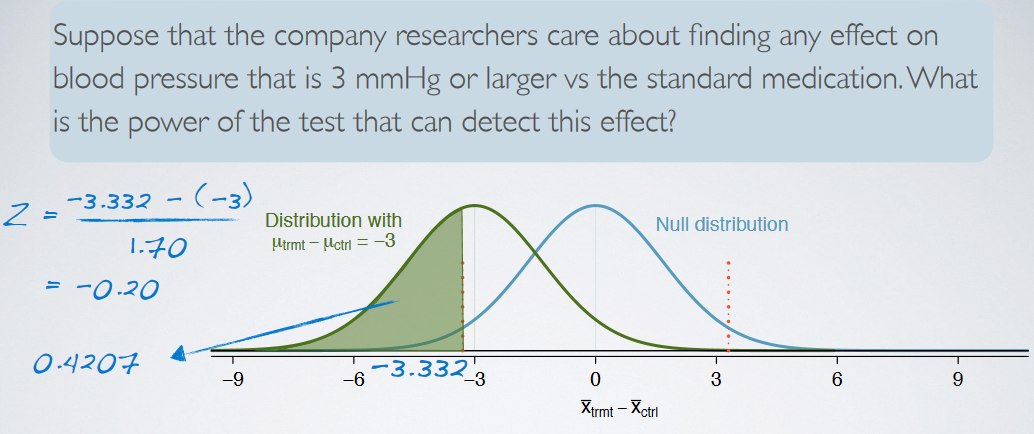
* Comparing Population Means

Conditions:

* Independence within groups: random sample / assignment; both <10% if without replacement
* Independence between groups: if dependent, then we call them paired, handle with diff;
* Skew: normal distribution or sample size larger than 30 if skewed.

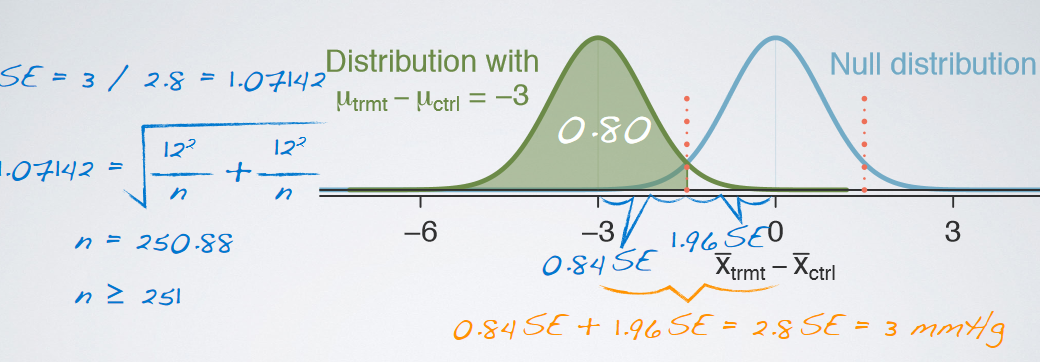
Comparing:

* With paired data, make a difference between the variables to be compared,
* H0 : ; H1 : ;
* ;
* SE;
* test statistic: t =
* df (degree of freedom) = min()
* Power
* : correctly rejecting H0 when H1 is true;
* Calculating:



Given the desired effective level to test out (-3 in the case above), adjust the null distribution and calculate the test statistic, using the critical value in the null distribution (-3.332 in the case above)

* Determine the minimum sample size for a desired power



Use a similar approach with an assumption of the desired sample size n.

* **ANOVA (Analysis of Variance)**
* H0: no difference among the multiple population means across groups

H1: at least one mean is different

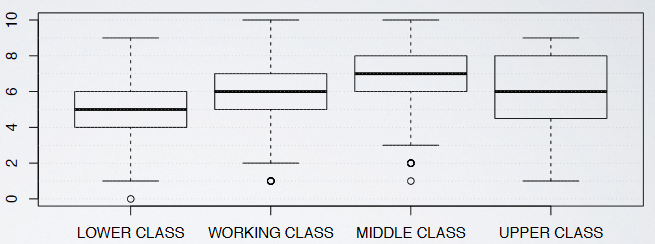
* Conditions:

1. Independence within groups: random sample/assignment; < 10%

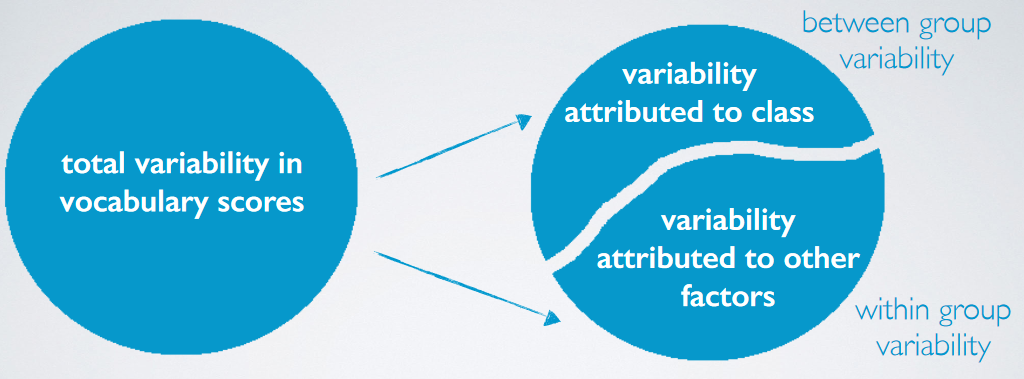
2. Independence across groups: multiple comparisons;

3. Approximate normality, or sample size large enough;

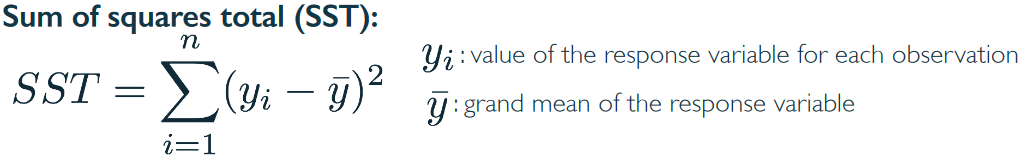
4. Constant variance among groups (homoscedastic): use tables or box plots to present.

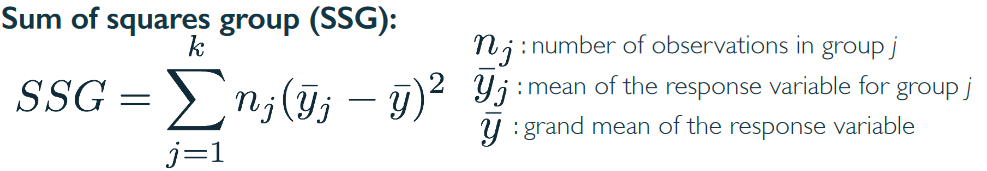


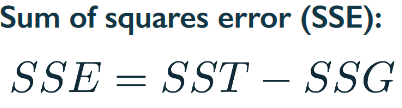
* Variability partitioning:

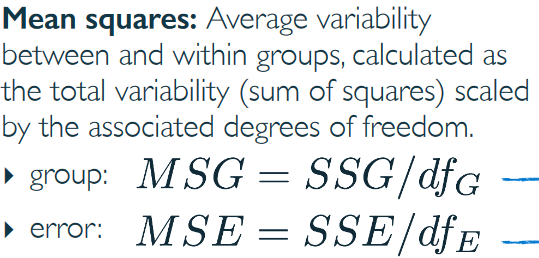


* F-test:









Pr(>F): p-value

(in R studio)

> pf(F-value, dfG, dfE, lower.tail=FALSE)

* Multiple Comparisons
* Multiple comparisons

Definition: the two-sample t-test to check whether their means differ;

Aim: find which means are different.

* The significance level for each t-test:

Bonferroni correction: , where K.

* T-test

SE;

T-statistic

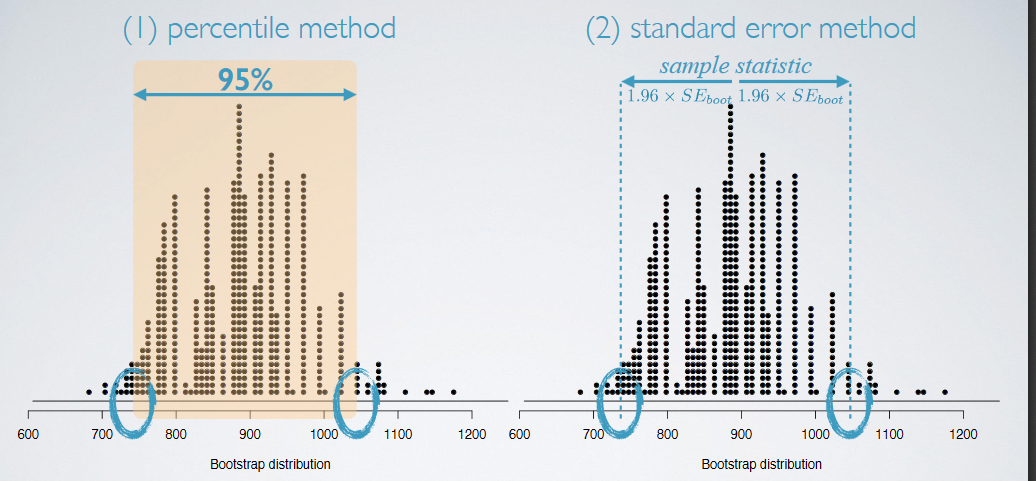
* **Bootstrapping**
* Serve: for small (<30) and skewed samples;
* Bootstrapping distribution: sample with replacement from the small sample for as many times (stimulate a distribution to make inference about the population parameters, even they’re other than population mean, e.g. median)
* Constructing bootstrapping intervals:

1. Percentile methods: find the desired percentiles as interval bounds

2. SD method:

sample statistic ;

, and df = n-1, where n denotes the original sample size



* Limitation: the bootstrapping sample should be nearly normal

##### Week 4 Inference for Proportions

* Central Limit Theorem for Proportions
* Conditions:

Independence: random sample/assignment; <10%

Sample size: ;

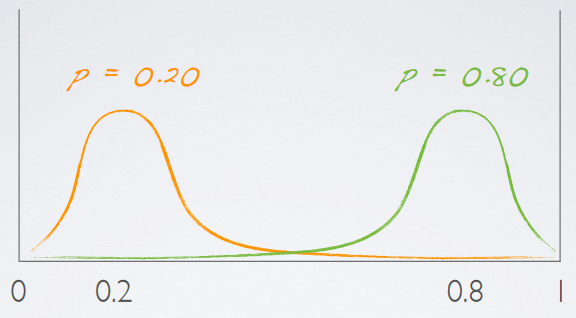
if population proportion p is unknown, use the sample proportion

* Theorem

.

* If the conditions fail to stand:

The sampling distribution would be skewed instead of nearly normal, where the tail depends on the population proportion



* HT and CI for a single proportion
* Hypothesis test

;

SE

* Confidence interval

SE

Interval: ;

If there is no previous study revealing use as a best guess.

* Comparison of proportions under CLT
* Conditions

Independence within group: random sample/assignment; <10%

Independence between groups: non-paired

Sample size: both meet , and

* CI

SE;

Confidence interval:

* HT

;

As the null hypothesis hasn’t given a null value to calculate with, we use:

pool proportion:

Adjust the sample size conditions: , and

Adjust the standard error: SE

* Inference via simulation (when the sample size is small)
* Intuition:

P-value is defined as P(observed or more extreme outcome | H0 true);

Assuming H0 true, devise a simulation and repeat for many times;

Calculate p-value as the proportion of simulations that yield an extreme result.

* R console codes
* Comparing two small sample proportions
* **Chi-square Test** (for categorical variables with more than two levels)
* For a single categorical variable: Goodness of Fit test

1. Goodness of fit

Intended to test how well the observed data fit with the expected results.

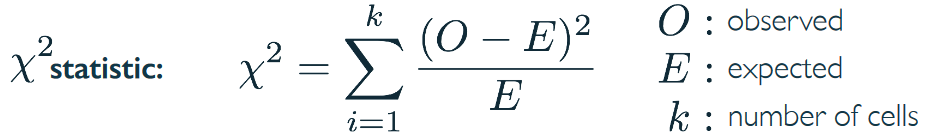
2. Conditions:

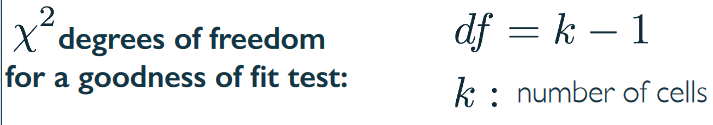
;

3. Hypotheses

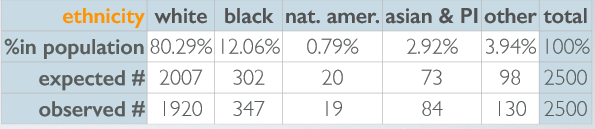
;

4. Test statistics





5. Report and R console



> pchisq(χ2, df, lower.tail = FALSE)

* For two categorical variables: Independence test

1. Independence:

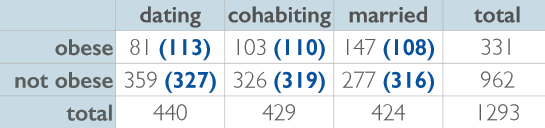
Check if two categorical variables are independent; dependent when large deviations from the expected are observed.

2. Conditions (same with chi-square GOF test)

3. Hypotheses:

;

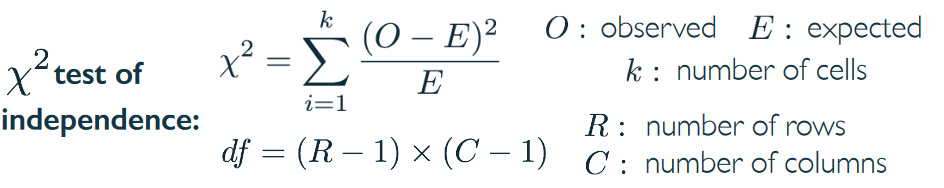
4. Report



Calculate the expected counts in () with:

Expected Count = expected proportion \* column total =

5. Test statistics



6. R console

> pchisq(χ2, df, lower.tail = FALSE)