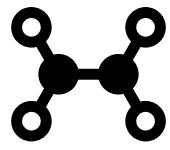
QUIZ 3

ONE PAGER



CHAPTER 1

1. INDIVIDUALS AND VARIABLES

1. Categorical variable and Quantitative variable

2. REPRESENTING CATEGORICAL VARIABLE

1. Bar chart, pie chart

Categorical data uses percentages, or proportions, to make inference.

3. REPRESENTING QUANTITATIVE VARIABLE

1. Histogram, bar chart, Sten plot， dot plot, box plot

quantitative data is **average-able.**

Quantitative data uses means, or averages, to make inference

4. DESCRIBING DISTRIBUTION OF A QUANTITATIVE VARIABLE

1. Shape, center, spread, outlier

It is important to include context in your answer. For example, "The mean number of bananas purchased was 5 bananas, There was one outlier when a customer purchased a bunch of 12 bananas. The shape of our data distribution was fairly symmetric. The range of bananas per bunch was 10, with the largest bunch being 12 and the smallest bunch being 2."

CHAPTER 2

CATEGORICAL VARIABLES

Two-way Table

QUANTITATIVE VARIABLES

Scatterplots: Form, Direction, Strength, outliers （strong/medium/weak + positive/negative + linear/nonlinear + outlier ）

Correlation Coefficient

Residual plot

Linear Regression (Least Squares Regression)

regression: interpretation, in context, of

1. r – positive or negative, weak or strong linear association between explanatory variable
and response variable

2. r^2 – x percent of the variation in the response variable can be explained by the approximate linear relationship with the explanatory variable.

3. **slope** – for every 1 unit increase in the explanatory variable, our model predicts an average increase of y units in the response variable.

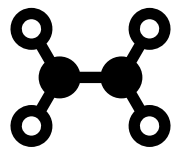
4. y**-intercept** – at an explanatory variable value of 0 units, our model predicts a response variable value of y units. (Does this make any sense?)

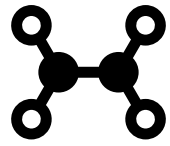
Extrapolation

Residuals

Influential Points

Transforming Data Sets





CHAPTER 3

DATA COLLECTION

simple random sample

stratified random sample

cluster sample

convenience sampling

systematic sample

multistage sampling

BIAS

voluntary response bias

under-coverage bias

nonresponse bias

response bias

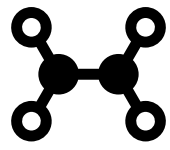
PRINCIPLES OF EXPERIMENTAL DESIGN

control, randomize, replicate, block

confounding variable

prospective study / retrospective study

matched pairs design



CHAPTER 4

INDEPENDENT

Two events are said to be independent if the outcome of one does not affect the outcome of the other.

MUTUALLY EXCLUSIVE

Another key concept in probability is when two events are **mutually exclusive.** When two events are mutually exclusive, it means that it is impossible for them to occur at the same time.

NORMAL DISTRIBUTION

The most popular type of distribution in all data situations is the normal distribution. Whether it be ACT scores, heights of people or blood pressure levels, these all follow normal distributions and make it much easier to calculate where one data point compares to the rest of our data.

BINOMIAL DISTRIBUTION

Binomial distributions are events that involve four conditions:

Two possible outcomes (binary)

Independent trials

Fixed number of trials

All trials are equally likely of occurring

GEOMETRIC DISTRIBUTION

A **geometric distribution** is very similar to a binomial distribution, with the only difference being that *we do not have a fixed number of trials*. A geometric distribution typically involves repeating an action until you get a success.

For example, if we flip a coin *until we get a heads* this would represent a geometric distribution.

CHAPTER 5

confidence interval – I'm __% confident that the population proportion/mean of _____ is between ___ and ___.
- or -
I am __% confident that the interval (__, __) captures the true proportion/mean of ____.
If slope CI, combine slope and CI interpretation.

confidence level – If this poll/experiment were repeated many times, then about __% of the resulting confidence intervals would contain the true proportion/mean of _____.

5.1PARAMETER AND STATISTICS

Parameter-Population，Statistic-Sample

Sampling Variability -sampling error，sample variability

sample size larger，sampling variability smaller，shape normal

Sampling Distribution：mean=population mean

5.2 SAMPLING PROPORTION

Conditions：

$$\mu_{\hat{P}} = p$$

$$\sigma_{\hat{P}} = \sqrt{\frac{pq}{n}}$$

1 Randomization

2 Sample size <10% population, independent

3 np>10, nq>10, 推出sample size large enough, sampling distribution of sample proportion nearly normal

all conditions met, safe to use one sample proportion model

normal cdf->p

5.3 SAMPLE MEANS

Conditions:

1. Randomization
2. Independent
3. n>30, according to CLT, the sampling distribution of sample mean is approximately normal; n<30, histogram ,histogram nearly normal

normal cdf -> p

CLT sample size, not population size

5.4 TWO INDEPENDENT SAMPLE PROPORTIONS

Condition：【independent group assumption】2x【random+10% condition （independent）+ np nq （normal）】. All conditions met, the sampling distribution can be seen as nearly normal model.

$$(\widehat{p}_1 - \widehat{p}_2) \pm z^* \sqrt{\frac{\widehat{p}_1(1-\widehat{p}_1)}{n_1} + \frac{\widehat{p}_2(1-\widehat{p}_2)}{n_2}}$$

5.5 TWO INDEPENDENT SAMPLE MEANS

Condition：【independent group assumption】2x【random+10% condition（independent）+ n1，n2>30， according to CLT…】

	Confidence Interval	Test Statistic
Two Independent Proportions At least 10 successes and 10 failures in both samples.	$(\hat{p}_1 - \hat{p}_2) \pm z^* \sqrt{\frac{\hat{p}_1(1-\hat{p}_1)}{n_1} + \frac{\hat{p}_2(1-\hat{p}_2)}{n_2}}$	$z = \frac{\hat{p}_1 - \hat{p}_2}{\sqrt{\hat{p}(1-\hat{p})\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}}$
Two Independent Means Both sample size are at least 30 OR populations are normally distributed.	$(\bar{x}_1 - \bar{x}_2) \pm t^* \sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}$ $Estimated\ df = smallest\ n - 1$	$t = \frac{\bar{x}_1 - \bar{x}_2}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}}$ $Estimated\ df = smallest\ n - 1$

CHAPTER 6

p-**value** – the probability of getting a result as extreme or more extreme than the one observed if the null hypothesis is correct.

fail to reject the null hypothesis – We do **not** have sufficient evidence at the ___ level (or PVal=.__) to support the alternative hypothesis. *Remember, you*

have not proved the null hypothesis is true—just failed to prove it false!

reject the null hypothesis – At the ___ level of significance, there is sufficient evidence to support the alternative hypothesis.

6.1 ONE PROPORTION Z INTERVAL

1-proportion z-interval

$$\hat{p} \pm z^* \sqrt{\frac{\hat{p}(1-\hat{p})}{n}}$$

Statistic ± Critical value × Standard deviation of the statistic

POINT ESTIMATE

STANDARD ERROR (an estimate for std. dev)

Conclusion: "There are n% confident that the interval from (a to b) captures the true proportion of …"

Confidence level: "If we take many samples of the same size from this population, about 95% of them will result in an interval that captures the actual percentage of …"

6.2 ONE PROPORTION Z TEST

P value: the probability of getting a result as extreme or more extreme than one observed if the null hypothesis is correct.

P>0.05, fail to reject hypothesis

P<0.05, reject hypothesis

6.3 ERRORS

P=0.05 means, “Given the null hypothesis, there is a n% chance of observing the statistic value that we have actually observed”

Large P value does not prove H0 true, but it offers no evidence that it’s not true—therefore we reject null hypothesis cuz sample variability

Alpha level （significant level ）,

Type 1 error, Ho true yet you reject H0, accept Ha

Type 2 error, Ho is false yet you accept H0, reject Ha

Alpha level+ , Beta level -