A Quick Review of Classical Statistics

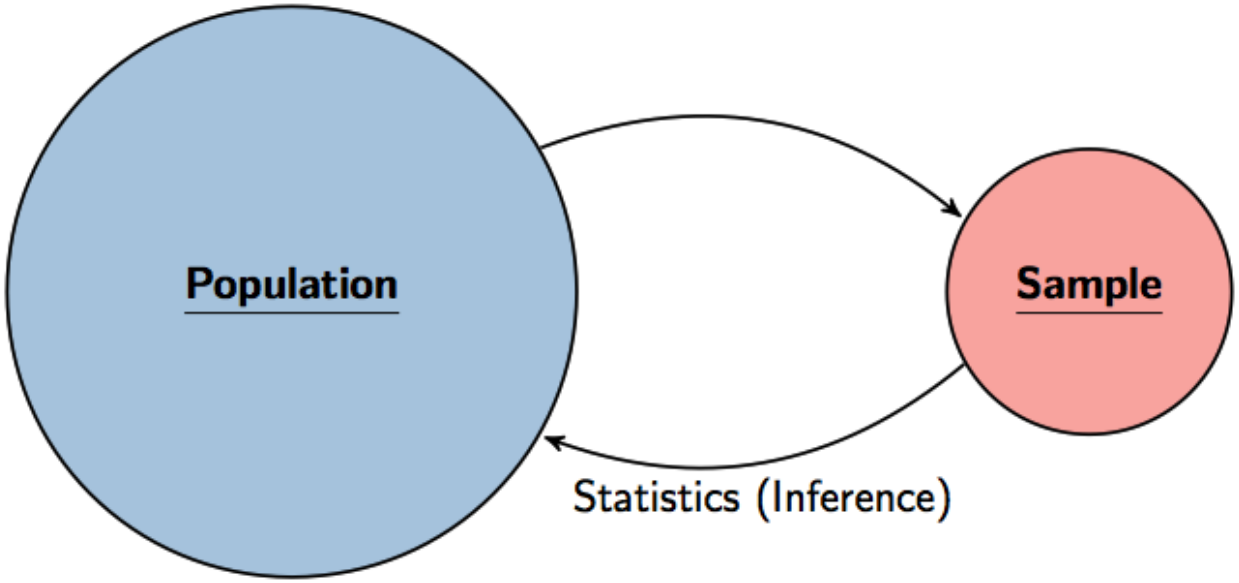
Yingbo Li

Southern Methodist University

STAT 4399

Outline

Populations and samples



- A *population* is a set of units (usually people, objects, transactions, or events) that we are interested in studying.
- A *sample* is a subset of the units of a population.
- A *variable* is a characteristic or property of an individual experimental unit (e.g., age, gender or income).
- A *statistical inference* is an estimate or prediction or some other generalization about a population based on information.

Sampling is natural

Example: you are cooking a pot of soup - you taste (examine) a small spoonful of it to get an idea about the saltiness of the entire pot.

Question: what's the population, sample and variable here?

- When you taste a spoonful of soup and decide it doesn't taste salty enough, that's *exploratory analysis*.
- If you generalize and conclude that your soup needs salt, that's an *inference*.
- For your inference to be valid the spoonful you tasted (the sample) needs to be *representative* of the entire pot (the population).
  - If your spoonful comes only from the surface and the salt is collected at the bottom of the pot, what you tasted is probably not representative of the whole pot.
  - If you first stir the soup thoroughly before you taste, your spoonful will more likely be representative of the whole pot.

Random vs. biased sampling

Over the last 5 years, what is the average time to degree for SMU undergraduate students?
Population: SMU students who have graduated in the last five years

- If we want to examine characteristics of all SMU undergraduates, we should take a *random sample* from the population.
- We could do this by getting a list of all SMU undergraduates from the registrar's office and randomly selecting a number of students from that list.
- If we employ a method where all students are equally likely to be selected, this would be a *simple random sampling* that is *representative* of the population.
  Can you suggest such a method?
- If instead we asked first-year students to sample a number of SMU students and they sample only from students in their classes and dorm, the sample would be *biased* towards first-year students.

Data collection

- *Observational study*: researchers collect data in a way that does not directly interfere with how the data arise, i.e. they merely observe, and can only establish an **association** between the explanatory and response variables.
  Eg. closing stock prices, profit trends, survey
- *Designed experiment*: researchers randomly assign subjects to various treatments in order to be able to establish **causal connections** between the explanatory and response variables.
  Eg. medical study on a particular drug, which is the best fertilizer
- **"Correlation does not imply causation"**.

Mean and variance

X is a continuous rv with pdf $f(x)$.

Definition

We define the mean or expected value of X as μ or $E(X)$, is

$$\mu = E(X) = \int_{-\infty}^{\infty} x f(x) dx$$

Think of mean as a "weighted average": use the probability model to weight the possible values of X.

Definition

We define the variance of X as σ^2 or $V(X)$, is

$$\sigma^2 = V(X) = E[(X-\mu)^2] = \int_{-\infty}^{\infty} (x-\mu)^2 f(x) dx = \int_{-\infty}^{\infty} x^2 f(x) dx - \mu^2$$

The standard deviation of X is σ.

Functions of random variables

For rv's X, X_1, X_2, and any constant $c, c_1, c_2 \in \mathbb{R}$,

1. Let $Y = X + c$, then

$$E(Y) = E(X) + c, \quad V(Y) = V(X)$$

2. Let $Y = cX$, then

$$E(Y) = cE(X), \quad V(Y) = c^2 V(X)$$

3. Let $Y = c + c_1 X_1 + c_2 X_2$, then

$$E(Y) = c + c_1 E(X_1) + c_2 E(X_2)$$

If X_1 and X_2 are **independent**, then

$$V(Y) = c_1^2 V(X_1) + c_2^2 V(X_2)$$

Point estimation

Definition

A point estimate of some population parameter θ is a single numerical value $\hat{\theta}$ of a sample statistic $\hat{\Theta} = h(X_1, X_2, \ldots, X_n)$.

- Estimator of the population mean μ

$$\bar{X} = \frac{\sum_{i=1}^{n} X_i}{n}$$

- Estimator of the population variance σ^2

$$\hat{\sigma}^2 = \frac{\sum_{i=1}^{n} (X_i - \bar{X})^2}{n-1}$$

Desirable properties of estimators

Definition

The bias of an estimator $\hat{\Theta}$ for parameter θ is:

$$bias(\hat{\Theta}) = E[\hat{\Theta}] - \theta$$

An estimator $\hat{\theta}$ is unbiased for population parameter θ if

$$E[\hat{\Theta}] = \theta, \quad \text{in other words, the bias of } \hat{\Theta} \text{ is 0.}$$

- ☞ Suppose $X_1, \ldots, X_n$ are independent random variables, and they all have the same mean μ and variance σ^2. Then
  1. $\bar{X}$ is an unbiased estimator of μ.
  2. $\hat{\sigma}^2$ is an unbiased estimator of σ^2.

Mean squared error (MSE) and standard error

Definition

The mean squared error of a point estimate $\hat{\theta}$ is:

$$MSE(\hat{\Theta}) = E[(\hat{\Theta} - \theta)^2].$$

- ☞ Variance-bias trade-off

$$\text{MSE}(\hat{\Theta}) = V(\hat{\Theta}) + (\text{bias})^2$$

Standard error of a statistic is the standard deviation of its sampling distribution. The smaller the SE is, the preciser the estimation is.

- The standard error of $\bar{X}$ is

$$\frac{\hat{\sigma}}{\sqrt{n}}$$

Hypothesis testing framework

- We start with a *null hypothesis* (H_0) that represents the status quo.
- We also have an *alternative hypothesis* (H_1) that represents our research question, i.e. what we're testing for.
- We conduct a hypothesis test under the assumption that the null hypothesis is true, either via simulation or theoretical methods.
- If the test results suggest that the data do not provide convincing evidence for the alternative hypothesis, we stick with the null hypothesis. If they do, then we reject the null hypothesis in favor of the alternative.

Set the hypotheses

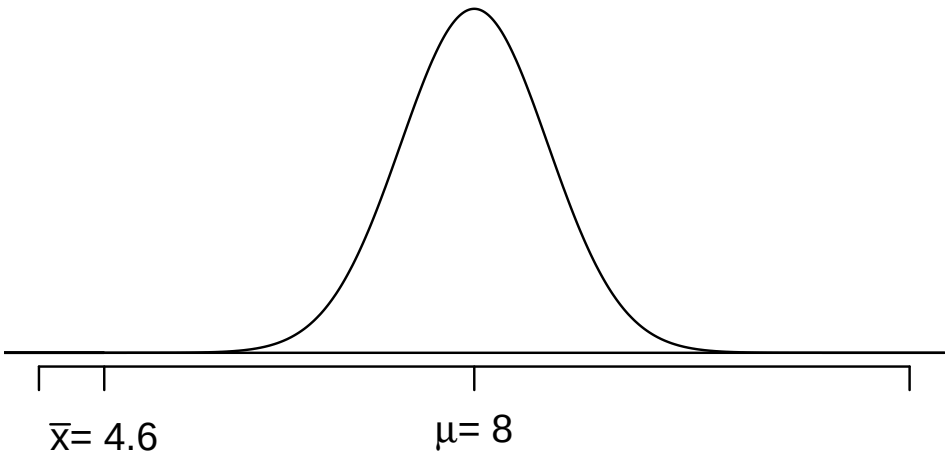
- Null hypothesis is a very specific statement, and usually includes $=$
- Alternative hypothesis claims other possibilities, and usually includes $>$, $<$, or $\neq$

A survey asked how many colleges students applied to, and 35 students responded to this question. This sample yielded an average of 4.6 college applications with a standard deviation of 4. College Board website states that counselors recommend students apply to roughly 8 colleges. What are the correct set of hypotheses to test if these data provide convincing evidence that the average number of colleges all SMU students apply to is **lower than** recommended.

$H_0 : \mu = 8$
$H_1 : \mu < 8$

http://www.collegeboard.com/student/apply/the-application/151680.html

Number of college applications - test statistic



When $n \geq 30$, according to CLT,

$$SE(\bar{X}) = \frac{\sigma}{\sqrt{n}} \text{ or } \frac{\hat{\sigma}}{\sqrt{n}},$$

whichever of σ and $\hat{\sigma}$ is available.

And

$$\bar{X} \sim N\left(\mu = 8,\ SE = \frac{4}{\sqrt{35}} = 0.7\right)$$

$$z_0 = \frac{\bar{x} - \mu_0}{SE} = \frac{4.6 - 8}{0.7} = -4.9$$

The sample mean is -4.9 standard errors away from the hypothesized value. Is this considered unusually (significantly) high?

Yes, but we can quantify how unusual it is using a p-value.

p-values

Definition

The p-value is the probability of observing data at least as favorable to the alternative hypothesis as our current data set, if the null hypothesis is true.

- If the p-value is *low* (lower than the significance level, α, which is usually 5%), then it would be very unlikely to observe the data if the null hypothesis were true, and hence *reject* H_0.
- If the p-value is *high* (higher than α), then it is likely to observe the data even if the null hypothesis were true, and hence *do not reject* H_0.

Probability of observing data at least as favorable to H_1 as our current data set (a sample mean smaller than 4.6), if in fact H_0 was true (the true population mean was 8):

$$P(\bar{X} < 4.6 \mid \mu = 8) = P(Z_0 = \frac{\bar{X} - \mu_0}{SE} < -4.9) = 0.0000005$$

Number of college applications - making a decision

- p-value $= 0.0000005$
  - If the true average of the number of colleges students applied to is 8, there is only 0.00005% chance of observing a random sample of 35 SMU students who on average apply to 4.6 or less schools.
  - This is a pretty low probability for us to think that a sample mean of 4.6 or less schools is likely to happen simply by chance.
- We never accept H_0 since we're not in the business of trying to prove it. We simply want to know if the data provide convincing evidence to support H_1.
- Since p-value is *low* (lower than $\alpha = 5\%$), we *reject* H_0.
- The data provide convincing evidence that SMU students average apply to less than 8 schools.
- The difference between the null value of 8 schools and observed sample mean of 4.6 schools is *not due to chance* or sampling variability.