Lecture 10: Categorical Data and Non-parametric methods

STATS 101: Foundations of Statistics

Linh Tran
linh@thetahat.ai

February 5, 2019

Announcements

- Next '*assignment*' is posted (due 2/12 @ 9:00am)
- A *Colab* script is available for today's class.

Outline

Hypothesis testing

- Finish hypothesis test from last lecture
  - t-test
  - F-test
  - Likelihood ratio test
- Categorical data
- χ^2 test
- Kolmogorov-Smirnov test
- Wilcoxon-Mann-Whitney rank test
- Robust estimators
- Simpson's paradox

Recall

Given $X_n, \ldots, X_n \overset{iid}{\sim} P_0$, we can form an estimator

$$\hat{\theta}_n = \omega(X_1, \ldots, X_n) \tag{1}$$

of some underlying parameter on P_0.

- $\hat{\theta}$ has a sampling distribution
- We can try to find estimators that reach the CRLB
- We can opt for estimators that are ranges (i.e. CI's)
- We can conduct tests against a null hypothesis

The t-test

Let $X_1, ..., X_n \overset{iid}{\sim} N(\mu, \sigma^2)$.

Our hypothesis: $H_0 : \mu = \mu_0$ vs. $H_1 : \mu \neq \mu_0$

The t-test

Let $X_1, ..., X_n \overset{iid}{\sim} N(\mu, \sigma^2)$.

Our hypothesis: $H_0 : \mu = \mu_0$ vs. $H_1 : \mu \neq \mu_0$
We can form a test statistic

$$U = n^{1/2} \frac{\bar{X}_n - \mu_0}{\sigma_n'} \tag{2}$$

where $\sigma_n' = \left(\frac{1}{n-1} \sum_{i=1}^n (X_i - \bar{X}_n)^2 \right)^{1/2}$.

- When $\mu = \mu_0$, then $U \sim t(n-1)$.
- We therefore set up δ such that we reject H_0 if $|U| \geq T_{n-1}^{-1}(1 - \alpha_0/2)$

n.b. T is the quantile function of the $t - distribution$ with $n - 1$ degrees of freedom. *Colab link*

The t-test

Many variations exist for the t-test, e.g.

- **Two sample t-tests**: two different samples are drawn and we want to compare them to each other
- **Paired t-tests**: Matched units are compared (e.g. a before and after study on the same patients)

The t-test

Many variations exist for the t-test, e.g.

- **Two sample t-tests**: two different samples are drawn and we want to compare them to each other
- **Paired t-tests**: Matched units are compared (e.g. a before and after study on the same patients)

Many variations exist within each type of t-test as well, e.g. for the two sample t-test we have

- Equal sample sizes, equal variances
- Unequal sample sizes, equal variance
- Unequal sample sizes, unequal variances

The F-test

Commonly used for (i) testing variances of normal distributions, and (ii) testing means of more than two distributions.

The F-test

Commonly used for (i) testing variances of normal distributions, and (ii) testing means of more than two distributions.

Let $W \sim \chi^2(n)$ and $Y \sim \chi^2(m)$. Then

$$X = \frac{Y/m}{W/n} = \frac{nY}{mW} \sim F(m,n) \tag{3}$$

follows a *F-distribution* with m and n degrees of freedom.

Some notes:

- $F(m,n) \neq F(n,m)$
- If $X \sim F(m,n)$, then $1/X \sim F(n,m)$
- If $X \sim t(n)$, then $X^2 \sim F(1,n)$

The F-test

Let $X_1, ..., X_m \overset{iid}{\sim} N(\mu_1, \sigma_1^2)$ and $Y_1, ..., Y_n \overset{iid}{\sim} N(\mu_1, \sigma_2^2)$.

Suppose we want to test: $H_0 : \sigma_1^2 = \sigma_2^2$ vs $H_1 : \sigma_1^2 \neq \sigma_2^2$.

We can form a test statistic

$$V = \frac{S_X^2/(m-1)}{S_Y^2/(n-1)} \tag{4}$$

where $S_X^2/(m-1)$ and $S_Y^2/(n-1)$ are estimators of σ_1^2 and σ_2^2, respectively.

- $V \sim F(m-1, n-1)$ when $\sigma_1^2 = \sigma_2^2$

Colab link

The Likelihood ratio test

Let $X_1, ..., X_n \overset{iid}{\sim} P_\theta$

Suppose we want to test: $H_0 : \theta \in \Theta_0$ vs $H_1 : \theta \in \Theta_1$.

Recall that the likelihood of our data is

$$L(\theta|x_1, ..., x_n) = \prod_{i=1}^{n} f(x_i|\theta) \tag{5}$$

We can use this to form a "likelihood ratio" statistic, i.e.

$$\begin{aligned} \Gamma(x) &= -2\log\left[\frac{\sup_{\theta\in\Theta_0} L(\theta|x_1, ..., x_n)}{\sup_{\theta\in\Theta} L(\theta|x_1, ..., x_n)}\right] & (6) \\ &= 2\left[\ell(\hat{\theta}_{MLE}|x_1, ..., x_n) - \sup_{\theta\in\Theta_0} \ell(\theta|x_1, ..., x_n)\right] & (7) \end{aligned}$$

It can be shown that $\Gamma(x) \sim \chi^2(p)$. *Colab link*

Categorical data

Let $X_1, ..., X_n \overset{iid}{\sim} P_\theta : X_i \in \{1, ..., k\}$.

We can think of the probability function as

$$p_j = P(X_i = j) : j = 1, ..., k \tag{8}$$

where $\sum_{j=1}^{k} = 1$

Some examples:

- Male vs Female
- Level of education
- Genre of book

χ^2 Test

Let $X_1, ..., X_n \overset{iid}{\sim} P_\theta : X_i \in \{1, ..., k\}$.

Let $p_i^0 \in (0,1) : i = 1, ..., k$ denote some fixed value such that $\sum_{i=1}^k p_i^0 = 1$.

We can set up the following hypothesis test:

$$\begin{array}{lll} H_0 : & p_i = p_i^0 & \text{for } i = 1, ..., k \\ H_1 : & p_i \neq p_i^0 & \text{for at least one value of } i \end{array}$$

using the following statistic:

$$Q = \sum_{i=1}^{k} \frac{(N_i - np_i^0)^2}{np_i^0} \tag{9}$$

where N_i is the total number of category i observed and $Q \sim \chi^2(k-1)$

Example: Proportions

Let p denote the proportion of defective cells in a sample and suppose

$$\begin{aligned} H_0 : &\quad p = 0.1 \\ H_1 : &\quad p \neq 0.1 \end{aligned} \tag{10}$$

Example: Proportions

Let p denote the proportion of defective cells in a sample and suppose

$$\begin{aligned} H_0 &: \quad p = 0.1 \\ H_1 &: \quad p \neq 0.1 \end{aligned} \tag{10}$$

We can frame this in our setting, i.e.

$$\begin{aligned} H_0 &: \quad p_1 = 0.1 \text{ and } p_2 = 0.9 \\ H_1 &: \quad p_1 \neq 0.1 \text{ or } p_2 \neq 0.9 \end{aligned} \tag{11}$$

Colab link

Contingency tables

Used a lot when dealing with categorical variables, e.g.

	No	Yes
Male	a	b
Female	c	d

Contingency tables

Used a lot when dealing with categorical variables, e.g.

	No	Yes
Male	a	b
Female	c	d

n.b. Many times we'll want to test for independence in these tables, e.g.

$$\begin{aligned} H_0 &: \quad p_{ij} = p_{i\cdot} p_{\cdot j} \text{ for } i = 1, ..., R \text{ and } j = 1, ..., C \\ H_1 &: \quad \text{The hypothesis } H_0 \text{ is not true.} \end{aligned}$$

where $p_{i\cdot} = \sum_{j=1}^{C} p_{ij}$ and $p_{\cdot j} = \sum_{i=1}^{R} p_{ij}$.

χ^2 test of independence

Under H_0, we have only $p_{i.}p_{.j}$ for all the values if i, j.
Furthermore, because $\sum_i p_{i.} = \sum_j p_{.j} = 1$:

- we only need to estimate $(R-1)+(C-1) = R+C-2$ probabilities

Our statistic is:

$$Q = \sum_{i=1}^{R} \sum_{j=1}^{C} \frac{(N_{ij} - \hat{E}_{ij})^2}{\hat{E}_{ij}} \tag{12}$$

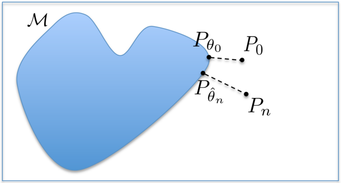
where $\hat{E}_{ij} = np_{i.}p_{.j} = \frac{N_{i.}N_{.j}}{n}$

n.b. $Q \sim \chi^2([RC-1]-[R+C-2])$

Parametric vs non-parametric statistics

Let $X_1, ..., X_n \overset{iid}{\sim} P_0$.

- **Parametric models**: We assume $P_0 = P_\theta : \theta \in \Theta \subset \mathbb{R}^p$.
- **Non-parametric models**: We allow θ to be infinite dimensional (but typically have 'restrictions').



χ^2 as a goodness-of-fit test

Let $X_1, ..., X_n \overset{iid}{\sim} P_0 : X_i \in (0,1)$. Where P_0 is some (unknown) pdf.

χ^2 as a goodness-of-fit test

Let $X_1, ..., X_n \overset{iid}{\sim} P_0 : X_i \in (0, 1)$. Where P_0 is some (unknown) pdf.

We can compare $X_1, ..., X_n$ against the uniform distribution via a goodness-of-fit test.

- Divide the interval (0,1) into 20 subintervals of equal length
- Calculate the expected count for each interval under the assumed distribution (i.e. uniform).
- Apply the χ^2 test

Colab link

The Kolmogorov-Smirnov test

Let $X_1, ..., X_n \overset{iid}{\sim} P_0 : X_i \in (-\infty, \infty)$. Where P_0 is some (unknown) pdf. For some fixed $F^*(x)$, we can test:

$$\begin{array}{ll} H_0: & F(x) = F^*(x) \text{ for } -\infty < x < \infty \\ H_1: & \text{The hypothesis } H_0 \text{ is not true.} \end{array}$$

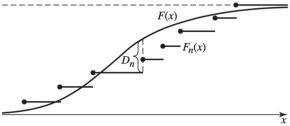
The Kolmogorov-Smirnov test

Let $X_1, ..., X_n \overset{iid}{\sim} P_0 : X_i \in (-\infty, \infty)$. Where P_0 is some (unknown) pdf. For some fixed $F^*(x)$, we can test:

$$\begin{aligned} H_0: &\quad F(x) = F^*(x) \text{ for } -\infty < x < \infty \\ H_1: &\quad \text{The hypothesis } H_0 \text{ is not true.} \end{aligned}$$

An approach of testing this:

$$D_n^* = \sup_{-\infty<x<\infty} |F_n(x) - F^*(x)| \tag{13}$$



- D_n^* has a unique cdf (i.e. $1 - 2\sum_{i=1}^{\infty}(-1)^{i-1}e^{-2i^2t^2}$)
- The KS-test is commonly used for two-sample tests.

The Wilcoxon-Mann-Whitney Rank test

Let $X_1, ..., X_m \overset{iid}{\sim} F$ and $Y_1, ..., Y_n \overset{iid}{\sim} G$. We can test:

$$\begin{aligned} H_0 :& \quad F = G \\ H_1 :& \quad F \neq G \end{aligned}$$

The Wilcoxon-Mann-Whitney Rank test

Let $X_1, ..., X_m \overset{iid}{\sim} F$ and $Y_1, ..., Y_n \overset{iid}{\sim} G$. We can test:

$$\begin{aligned} H_0 : &\quad F = G \\ H_1 : &\quad F \neq G \end{aligned}$$

The procedure:

1. Arrange the $m + n$ observations in order.
2. Record the rank of each ordered observation (along with whether it's from the X's or Y's).
3. Calculate $S = \sum_{i=1}^{m} = rank(X_i)$
4. Reject H_0 if $|S - (1/2)m(m + n + 1)| \geq c$

n.b. Intuition: If $F = G$, then $X_1, ..., X_m$ will be dispersed evenly throughout the $m + n$ observations, and

$$\mathbb{E}[S] = m\frac{m+n+1}{2} \quad (14)$$

$$\text{var}(S) = mn\frac{m+n+1}{12} \quad (15)$$

Robust estimators

Most of the time, our data does not follow standard distributions.

- Robust estimators perform well irrespective of the distribution it's applied to.

Recall (lecture 7):
A *parameter* is a mapping, $\theta : \mathbb{R}^p \to P_\theta \ni$

$$\mathcal{M} = \{P_\theta : \theta \in \Theta\} \tag{16}$$

Robust estimators

Most of the time, our data does not follow standard distributions.

- Robust estimators perform well irrespective of the distribution it's applied to.

Recall (lecture 7):
A *parameter* is a mapping, $\theta : \mathbb{R}^p \to P_\theta \ni$

$$\mathcal{M} = \{P_\theta : \theta \in \Theta\} \tag{16}$$

More generally, we can define a *parameter* as the mapping

$$\Psi : \mathcal{M} \to \mathbb{R} \tag{17}$$

This allows us to define a target parameter $\psi_0 = \Psi(P_0)$.

Example: The median, i.e. $\Psi(P_0) = \inf_{x \in \mathbb{R}} 0.5 \leq F(x)$

Robust estimators

M-estimators: (Robust) estimators that maximize a function

- Not necessarily the likelihood

Example of a function:

$$g_k(x|\theta,\sigma) = c_k e^{h_k([x-\theta]/\sigma)} : h_k(y) = \begin{cases} -0.5y^2 & \text{if } -k < y < k \\ 0.5k^2 - k|y| & \text{otherwise} \end{cases}$$

n.b. This function is maximized iteratively.

Simpson's paradox

Scenario: You want to study the effect of some exposure (A) on some outcome (Y) using observational data. Our current example is:

$$A = \text{matches} \tag{18}$$

$$Y = \text{lung cancer} \tag{19}$$

Our null hypothesis is therefore: "Matches are not associated with lung cancer."

$$\begin{tikzcd} U_A \arrow[r] & A \arrow[r] & Y & U_Y \arrow[l] \end{tikzcd}$$

Proposed structural causal model (A=matches, Y=lung cancer).

Simpson's paradox

To test the hypothesis, you decide to do a "*case control*" study (e.g. collect $1,000$ cases & $1,000$ controls from observed data).

- Among the $1,000$ cases of lung cancer, 820 carry matches.
- Among the $1,000$ reference patients, 340 carry matches.

Simpson's paradox

To test the hypothesis, you decide to do a "*case control*" study (e.g. collect $1,000$ cases & $1,000$ controls from observed data).

- Among the $1,000$ cases of lung cancer, 820 carry matches.
- Among the $1,000$ reference patients, 340 carry matches.

We can compute an "*odds ratio*" from this:

$$\text{Odds ratio} = \frac{820/180}{340/660} = 8.84 \ (95\% \ \text{CI} : 8.64, 9.05) \tag{20}$$

Question: Do we conclude that matches are associated with lung cancer?

Simpson's paradox

We decide to look at the relationship of matches and lung cancer separately in (i) smokers, and (ii) nonsmokers.

In doing so, we see that:

- Among the $1,000$ cases of lung cancer, 900 are smokers and 810 (of the 900) carry matches.
- Among the $1,000$ reference patients, 300 are smokers and 270 (of the 300) carry matches.

Simpson's paradox

We decide to look at the relationship of matches and lung cancer separately in (i) smokers, and (ii) nonsmokers.

In doing so, we see that:

- Among the $1,000$ cases of lung cancer, 900 are smokers and 810 (of the 900) carry matches.
- Among the $1,000$ reference patients, 300 are smokers and 270 (of the 300) carry matches.

We can compute stratified odds ratio's from this:

$$\begin{aligned} \text{Odds ratio}_{smokers} &= 1 \text{ (95\% CI} : 0.56, 1.44) \qquad (21) \\ \text{Odds ratio}_{nonsmokers} &= 1 \text{ (95\% CI} : 0.3, 1.7) \qquad (22) \end{aligned}$$

Question: What should we conclude now?

Simpson's paradox

To be complete, we decide to look at the relationship of smoking and lung cancer separately in patients with (i) matches, and (ii) no matches.

Using the same counts, we can compute stratified odds ratio's from this:

$$\begin{aligned} \text{Odds ratio}_{matches} &= 21 \text{ (95\% CI} : 20.32, 21.68) \qquad (23) \\ \text{Odds ratio}_{no\_matches} &= 21 \text{ (95\% CI} : 20.53, 21.47) \qquad (24) \end{aligned}$$

Question: What should we conclude now?

References

- DeGroot & Schervish Chapters 10.1, 10.3, 10.5-10.9