18.05 Exam 2 in-class review problems
Spring 2022

1 Summary

- Data: $x_1, \ldots, x_n$
- Basic statistics: sample mean, sample variance, sample median
- Likelihood, maximum likelihood estimate (MLE)
- Bayesian updating: prior, likelihood, posterior, predictive probability, probability intervals; prior and likelihood can be discrete or continuous
- NHST: H_0, H_A, significance level, rejection region, power, type 1 and type 2 errors, p-values.

2 Short review

2.1 Basic statistics

Data: $x_1, \ldots, x_n$.

$$\text{sample mean} = \overline{x} = \frac{x_1 + \ldots + x_n}{n}$$
$$\text{sample variance} = s^2 = \frac{\sum_{i=1}^{n}(x_i - \overline{x})^2}{n-1}$$
$$\text{sample median} = \text{middle value}$$

Example. Data: 1, 2, 3, 6, 8.
$\overline{x} = 4$, $\quad s^2 = \frac{4+1+16+9+4}{4} = 8.5$, $\quad$ median $= 3$.

2.2 Likelihood

x = data

θ = parameter of interest or hypotheses of interest

Likelihood:
For a discrete distribution: Likelihood = probability of data given hypothesis i.e.

$$L(\theta) = p(x \mid \theta).$$

For a continuous distribution: Likelihood = density of data given hypothesis i.e.

$$L(\theta) = f(x \mid \theta)$$

Log likelihood : $\ln(p(x \mid \theta)) \qquad \ln(f(x \mid \theta))$.

2.3 Maximum likelihood estimates (MLE)

Methods for finding the maximum likelihood estimate (MLE).

- Discrete hypotheses: compute each likelihood
- Discrete hypotheses: maximum is obvious
- Continuous parameter: compute derivative (often use log likelihood)
- Continuous parameter: maximum is obvious

2.4 Bayesian updating: conjugate priors.

Beta prior, binomial likelihood

Data: $x \sim \text{binomial}(n, \theta)$. θ is unknown.
Prior: $f(\theta) \sim \text{Beta}(a, b)$
Posterior: $f(\theta \mid x) \sim \text{Beta}(a + x, b + n - x)$

Normal prior, normal likelihood

$$a = \frac{1}{\sigma^2_{\text{prior}}} \qquad\qquad b = \frac{n}{\sigma^2}$$

$$\mu_{\text{post}} = \frac{a\mu_{\text{prior}} + b\overline{x}}{a+b}, \qquad\qquad \sigma^2_{\text{post}} = \frac{1}{a+b}.$$

Notice: μ_{post} is between μ_{prior} and $\overline{x}$; σ^2_{post} smaller than σ^2_{prior}.

2.5 Null hypothesis significance testing (NHST)

2.5.1 NHST: Steps

1. Specify H_0 and H_A.
2. Choose a significance level α.
3. Choose a test statistic and determine the null distribution.
4. Determine how to compute a p-value and/or the rejection region.
5. Collect data.
6. Compute p-value or check if test statistic is in the rejection region.
7. Reject or fail to reject H_0.

Make sure you can use the probability tables.

2.5.2 NHST: One-sample t-test

- Data: we assume normal data with both μ and σ unknown:

$$x_1, x_2, \dots, x_n \sim N(\mu, \sigma^2).$$

- Null hypothesis: $\mu = \mu_0$ for some specific value μ_0.
- Test statistic:

$$t = \frac{\overline{x} - \mu_0}{s/\sqrt{n}}$$

  where

$$s^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \overline{x})^2.$$

- Null distribution: $t(n-1)$, Student t with $n-1$ degs of freedom.
- Student t is symmetric around 0, like standard normal.

2.5.3 Two-sample t-test: equal variances

Data: we assume normal data with μ_x, μ_y and (same) σ unknown:

$$x_1, \dots, x_n \sim \mathrm{N}(\mu_x, \sigma^2), \quad y_1, \dots, y_m \sim \mathrm{N}(\mu_y, \sigma^2)$$

Null hypothesis H_0: $\mu_x = \mu_y$.

Pooled variance: $s_p^2 = \dfrac{(n-1)s_x^2 + (m-1)s_y^2}{n+m-2}\left(\dfrac{1}{n} + \dfrac{1}{m}\right).$

Test statistic: $t = \dfrac{\bar{x} - \bar{y}}{s_p}$

Null distribution: $\phi(t \mid H_0)$ is drawn from $\sim t(n+m-2)$.

(More generally we can test H_0: $\mu_x - \mu_y = \Delta\mu$ using $t = \dfrac{\overline{x} - \bar{y} - \mu_0}{s_p}$)

2.5.4 F-test = one-way ANOVA

Like t-test but for n groups of data with m data points each.

$$y_{i,j} \sim N(\mu_i, \sigma^2), \qquad y_{i,j} = j^{\text{th}} \text{ point in } \mathrm{i}^{\text{th}} \text{ group}$$

Assumptions: data for each group is an independent normal sample with (possibly) different means but the same variance.

Null hypothesis is: the means are all equal: $\mu_1 = \cdots = \mu_n$

Test statistic is $\frac{\mathrm{MS}_B}{\mathrm{MS}_W}$ where:

MS_B = between group variance $= \dfrac{m}{n-1}\sum(\bar{y}_i - \bar{y})^2$

MS_W = within group variance = sample mean of $s_1^2, \ldots, s_n^2$

Idea: If μ_i are equal, this ratio should be near 1.

Null distribution is F-statistic with $n-1$ and $n(m-1)$ d.o.f.:

$$\frac{\text{MS}_B}{\text{MS}_W} \sim F_{n-1,\, n(m-1)}.$$

2.5.5 NHST: some key points

1. The significance level α is not the probability of being wrong. It's the probability of being wrong if the null hypothesis is true.

2. Likewise, power is not the probability of being right. It's the probability of being right if a particular alternate hypothesis is true.

Problem 1. The following data is from a random sample:

$$1, 1, 1, 2, 3, 5, 5, 8, 12, 13, 14, 14, 14, 14, 18, 100.$$

Find the first, second and third quartiles.

Problem 2. MLE examples. For each of the following, there is an unknown parameter and some data. Give the likelihood function and find the MLE.

(a) We have a coin with probability of heads θ. We toss it 10 times and get 3 heads.

(b) Wait time follows $\exp(\lambda)$. In 5 independent trials wait 3, 5, 4, 5, 2

(c) We have 4, 6, 8, 12 and 20-sided dice. One is chosen at random and rolled twice giving resulting in a 9 and a 5.

Problem 3. MLE examples
For each of the following, there is an unknown parameter and some data. Give the likelhood function and find the MLE.

(a) In this problem there are two unknown parameters μ and σ.

Independent samples $x_1, \ldots, x_n$ are drawn from a $\mathrm{N}(\mu, \sigma^2)$ distribution.

(b) One sample $x = 6$ drawn from a $\mathrm{uniform}(0, \theta)$ distribution.

(c) One sample x drawn from a $\mathrm{uniform}(0, \theta)$ distribution.

Problem 4. Discrete prior-discrete likelihood.
Jon has 1 four-sided, 2 six-sided, 2 eight-sided, 2 twelve sided, and 1 twenty-sided dice. He picks one at random and rolls a 7.

(a) For each type of die, find the posterior probability Jon chose that type.

(b) What are the posterior odds Jon chose the 20-sided die?

(c) Compute the prior predictive probability of rolling a 7 on the first roll.

(d) Compute the posterior predictive probability of rolling an 8 on the second roll.

Problem 5. Suppose $x \sim \mathrm{binomial}(30, \theta)$, $x = 12$. If we have a prior $f(\theta) \sim \mathrm{Beta}(1, 1)$ find the posterior for θ.

Problem 6. Suppose $x \sim \mathrm{geometric}(\theta)$, $x = 6$. If we have a prior $f(\theta) \sim \mathrm{Beta}(4, 2)$ find the posterior for θ.

Problem 7. In the population IQ is normally distributed: $\theta \sim \mathrm{N}(100, 15^2)$. An IQ test finds a person's 'true' IQ + random error $\sim N(0, 10^2)$. Someone takes the test and scores 120.

Find the posterior pdf for this person's IQ.

Problem 8. z and one-sample t-test. For both problems use significance level $\alpha = 0.05$.

Assume the data 2, 4, 4, 10 are independent draws from a $N(\mu, \sigma^2)$ distribution.

Take H_0: $\mu = 0$; H_A: $\mu \neq 0$.

(a) Assume $\sigma^2 = 16$ is known and test H_0 against H_A.

(b) Now assume σ^2 is unknown and test H_0 against H_A.

Problem 9.
Two-sample t-test Suppose that we have data from 1408 women admitted to a maternity hospital for (i) medical reasons or through (ii) unbooked emergency admission. The duration of pregnancy is measured in complete weeks from the beginning of the last menstrual period.

(i) Medical: 775 observations with $\bar{x} = 39.08$ and $s^2 = 7.77$.

(ii) Emergency: 633 observations with $\bar{x} = 39.60$ and $s^2 = 4.95$.

(a) Set up and run a two-sample t-test to investigate whether the duration differs for the two groups.

(b) What assumptions did you make?

Problem 10. Three treatments for a disease are compared in a clinical trial, yielding the following data:

	Treatment 1	Treatment 2	Treatment 3
Cured	50	30	12
Not cured	100	80	18

Use a chi-square test to compare the cure rates for the three treatments

Problem 11. ANOVA. The table shows recovery time in days for three medical treatments.

T_1	T_2	T_3
6	8	13
8	12	9
4	9	11
5	11	8
3	6	7
4	8	12

Note: For $\alpha = 0.05$, the critical value of $F_{2,15}$ is 3.68.

(a) Set up and run an F-test for H_0 vs. H_A.

(b) Based on the test, what might you conclude about the treatments?

MIT OpenCourseWare
https://ocw.mit.edu

18.05 Introduction to Probability and Statistics
Spring 2022

For information about citing these materials or our Terms of Use, visit: https://ocw.mit.edu/terms.