Exam 2 Review

18.05 Spring 2018

- Cannot cover everything.
- You may bring a cheat sheet 5×7 inch index card (both sides) to the exam.
- You can also bring your cheat sheet from the first exam.
- Calculators are not allowed on the exam—they won't be needed.
- Get familiar with the probability tables for Z, t and χ^2. There are copies with the practice exam.

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, confidence intervals.

Basic statistics

Data: $x_1, \ldots, x_n$.

$$\text{sample mean} = \bar{x} = \frac{1}{n}(x_1 + \ldots + x_n)$$

$$\text{sample variance} = s^2 = \frac{1}{n-1}\left(\sum_{i=1}^{n}(x_i - \bar{x})^2\right)$$

$$\text{sample median} = \text{middle value (or average of two middle values)}$$

Example. Data: 6, 3, 8, 1, 2

$$\bar{x} = (6+3+8+1+2)/4 = 4$$

$$s^2 = ((6-4)^2 + (3-4)^2 + (8-4)^2 + (1-4)^2 + (2-4)^2)/4$$

$$= (4+1+16+9+4)/4 = 8.5$$

$$\text{median} = 3.$$

Likelihood

$x =$ data

$\theta =$ parameter of interest or hypotheses of interest

Likelihood = probability of data given hypothesis:

$$p(x \mid \theta) \quad \text{(discrete distribution)}$$
$$f(x \mid \theta) \quad \text{(continuous distribution)}$$

Log likelihood :

$$\ln(p(x \mid \theta)).$$
$$\ln(f(x \mid \theta)).$$

Likelihood examples

Examples. Find the likelihood function of each of the following.

1. Coin with probability of heads θ. Toss 10 times, get 3 heads.
2. Wait time $\sim$ exp(λ). In 5 independent trials wait $3, 5, 4, 5, 2$.
3. Usual 5 dice. Two independent rolls, $9, 5$. (Make a likelihood table.)
4. Independent $x_1, \ldots, x_n \sim \mathrm{N}(\mu, \sigma^2)$
5. $x = 6$ drawn from uniform$(0, \theta)$
6. x drawn from uniform$(0, \theta)$

In each case likelihood depends on data and unknown hypotheses.

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

Examples. Find the MLE for each example in the previous slide.

Bayesian updating: 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.

1. For each type, find the posterior probability Jon chose that type.
2. What are the posterior odds Jon chose the 20-sided die?
3. Compute the prior predictive probability of rolling 7 on roll 1.
4. Compute the posterior predictive probability of rolling 8 on roll 2.

Bayesian updating: conjugate priors

1. Beta prior, binomial likelihood

Data: $x \sim \text{binomial}(n, \theta)$. θ is unknown.

Prior: $f(\theta) \sim \text{beta}(a,\ b)$

Posterior: $f(\theta \,|\, x) \sim \text{beta}(a + x,\ b + n - x)$

Example. Suppose $x \sim \text{binomial}(30, \theta)$, $x = 12$.
If we have a prior $f(\theta) \sim \text{beta}(1, 1)$ find the posterior.

2. Beta prior, geometric likelihood

Data: x
Prior: $f(\theta) \sim \text{beta}(a,\ b)$
Posterior: $f(\theta \,|\, x) \sim \text{beta}(a + x,\ b + 1)$.

Example. Suppose $x \sim \text{geometric}(\theta)$, $x = 6$.
If we have a prior $f(\theta) \sim \text{beta}(4, 2)$ find the posterior.

Normal-normal

3. Normal prior, normal likelihood:

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

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

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

Example. In the population IQ is normally distributed:

$$\theta \sim \text{N}(100, 15^2).$$

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.

Bayesian updating: continuous prior-continuous likelihood

Examples. Update from prior to posterior for each of the following with the given data. Graph the prior and posterior in each case.

1. Romeo is late:

$$\text{likelihood: } x \sim U(0,\theta), \quad \text{prior: } U(0,1).$$
$$\text{data: } 0.3, 0.4. 0.4$$

2. Waiting times:

$$\text{likelihood: } x \sim \exp(\lambda), \quad \text{prior: } \lambda \sim \exp(2).$$
$$\text{data: } 1, 2$$

3. Waiting times:

$$\text{likelihood: } x \sim \exp(\lambda), \quad \text{prior: } \lambda \sim \exp(2).$$
$$\text{data: } x_1, x_2, \ldots, x_n$$

NHST: Steps

1. Specify H_0 and (perhaps) 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. (At least this deserves its own color.)
6. Compute p-value or see if test statistic is in rejection region.
7. Reject or fail to reject H_0.

It's very important that # 5 COMES AFTER #1–4!

Make sure you are familiar with the probability tables!

NHST: One-sample t-test

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

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

- 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.

Example: z and one-sample t-test

For both problems use significance level $\alpha = 0.05$.

Assume the data 2, 4, 4, 10 is drawn from a $N(\mu, \sigma^2)$.

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

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

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

Two-sample t-test: equal variances

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

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

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: $f(t \mid H_0)$ is the pdf of $t(n+m-2)$

More generally we can test H_0: $\mu_x - \mu_y = \mu_0$ using $t = \dfrac{\bar{x} - \bar{y} - \mu_0}{s_p}$.

Example: two-sample t-test

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 $s^2 = 4.95$

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

2. What assumptions did you make?

Chi-square test for goodness of fit

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

F-test = one-way ANOVA

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

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

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

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

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

$$\text{MS}_B = \text{between group variance} = \frac{m}{n-1} \sum (\bar{y}_i - \bar{y})^2$$

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.:

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

ANOVA example

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

1. Set up and run an F-test.

2. Based on the test, what might you conclude about the 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

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

NHST: right and wrong 1A.

1. Significance 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.