18.05 Exam 2 Solutions

Cheat Sheet

Problem 0. (5 pts) **Be sure to attach your cheat sheet to your test.**

Part I: Concept Questions

Problem I.1. (10 pts: 2,2,2,2,2) *Determine which of the following concepts/statements are Frequentist and which are Bayesian. Note: a concept can be either or both. Give a short explanation for each answer.*

(a) P-value

(b) Prior distribution

(c) Average of data

(d) There is a 3% probability that the average weight is between 92.3 mg and 100 mg.

(e) The odds in favor of H_0 against H_A are 1 to 3.

(a) Solution: $\boxed{\text{Frequentist}}$: p-values are used in NHST

(b) Solution: $\boxed{\text{Bayesian}}$: Priors are used in Bayesian updating. Frequentists do not consider probabilities of hypotheses.

(c) Solution: $\boxed{\text{Both}}$: the average of data is a statistic. All statisticians compute statistics!

(d) Solution: $\boxed{\text{Bayesian}}$: To compute this probability requires knowing the probabilities of all the hypotheses.

(e) Solution: $\boxed{\text{Bayesian}}$: Odds are a way of presenting probabilities. Frequentists do not consider the probabilities of hypotheses.

Problem I.2. (6 pts: 3,3) *Suppose you run a significance test at significance level 0.025, and that the test has a power of 95%.*

For each part of this problem, give your answer and a short explanation.

(a) Assuming the null hypothesis, what is the probability of a type I error?

(i) *0.025*   **(ii)** *0.05*   **(iii)** *0.95*   **(iv)** *0.975*   **(v)** *cannot be determined from the information given.*

*(b) Assuming the alternative hypothesis, what is the probability of a type II error?*

(i) *0.025*   **(ii)** *0.05*   **(iii)** *0.95*   **(iv)** *0.975*   **(v)** *cannot be determined from the information given.*

(a) Solution: $\boxed{\text{(i) } 0.025}$: By definition $P(\text{type I error}) = P(\text{reject} \mid H_0) = \text{significance}.$

(b) Solution: $\boxed{\text{(ii) } 0.05 = 1 - 0.95}$: By definition $P(\text{type II error}) = P(\text{non-rejection} \mid H_A) = 1 - P(\text{reject} \mid H_A) = 1$ - power.

Problem I.3. (3 pts) *The following graphs show the rejection regions and pdfs of the null and alternative hypotheses for two different hypothesis tests. Which graph shows the test with the higher power?*

Your answer should be 'left graph' or 'right graph'. Give a short explanation.
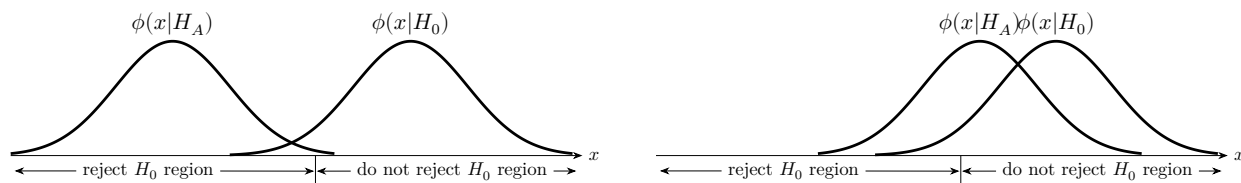
Solution: Left graph : power $= P(\text{reject}|H_A)$. The left hand graph has much more area under $\phi(x|H_A)$ and above the rejection region than the right hand graph.

Problem I.4. (3 pts) *You find a coin on the street, with some unknown probability θ of landing heads when tossed. Circle the only reasonable prior for θ. (No explanation needed.)*

(i) *Uniform*$([0, 0.5])$ **(ii)** *Beta*$(2, 2)$ **(iii)** $N(0.5, 0.25)$.

Solution: Beta$(2, 2)$: It has the correct range and no bias towards heads or tails.

Uniform$([0, 0.5])$ does not permit $\theta > 0.5$.

$N(0.5, 0.5^2)$ has significant amount of probability for $\theta < 0$ and $\theta > 1.0$. This is not allowed.

Problem I.5. (4 pts: 2,2) *For each of the following: Is the prior conjugate to the given likelihood? In each case, a and b are parameters for the priors.*

	hypothesis	data	prior	likelihood
(a)	$\theta \in [0, 1]$	x	$c_1\theta^a(1-\theta)^b$	$\binom{10}{x}\theta^x(1-\theta)^{10-x}$
(b)	$\lambda \in [0, \infty)$	x	$c_1\lambda^{a-1}e^{-b\lambda}$	$\lambda e^{-\lambda x}$

(a) Solution: Yes : the prior $\times$ likelihood has the form $c\theta^{a+x}(1-\theta)^{b+10-x}$. This is the same form as the prior, i.e. a constant times θ to a power times $(1-\theta)$ to a power. (In fact, a beta distribution.)

(b) Solution: Yes : the prior $\times$ likelihood has the form $c\lambda^a e^{-(b+x)\lambda}$. This is the same form as the prior.

Part II: Problems

Problem II.1. (12 pts.)
The gamma distribution with shape parameter 3 and unknown rate parameter β has range $(0, \infty)$ and pdf

$$f(x) = \frac{\beta^3 x^2}{2} e^{-\beta x}.$$

Suppose the data

$$1, 1, 2, 3, 5$$

was drawn independently from such a distribution. Find the maximum likelihood estimate (MLE) of β.

(Note: we shouldn't expect to get integer values for the data. So either our measurements were quite crude or we didn't want you to have to do arithmetic with fractions.)

Solution: Call the 5 data values: x_1, x_2, x_3, x_4, x_5. We'll use their numerical values when needed. The likelihood of the given data is

$$f(x_1)f(x_2)f(x_3)f(x_4)f(x_5) = \beta^{15} \cdot \frac{x_1^2 \cdot x_2^2 \cdot x_3^2 \cdot x_4^2 \cdot x_5^2}{2^5} \cdot e^{-\beta(x_1+x_2+x_3+x_4+x_5)}$$

So the log likelihood is $(\beta) = 15\ln(\beta) - \beta(x_1 + x_2 + x_3 + x_4 + x_5) + c$, where c is a constant. Taking the derivative and setting it to 0, we get

$$l'(\beta) = \frac{15}{\beta} - (x_1 + x_2 + x_3 + x_4 + x_5) = 0.$$

Solving for β gives $\boxed{\hat{\beta} = \dfrac{15}{x_1 + x_2 + x_3 + x_4 + x_5} = \dfrac{15}{12} = \dfrac{5}{4}}$, where in the last step, we used the given numerical values of the data.

Problem II.2. (15 pts)
A random process produces outcomes labeled A, B and C with probabilities $\theta/2$, $\theta/2$, $1 - \theta$ respectively. Here θ is an unknown parameter with value between 0 and 1. You want to know the value of θ.

Before running any experiments you have a prior pdf for θ of $f(\theta) = 3\theta^2$. You then run the process five times producing data A, B, C, A, B.

Find the posterior probability density for θ.

Solution: As usual, we make a Bayesian update table. The data $ABCAB$ has probability

$$p(ABCBA \mid \theta) = \frac{\theta}{2} \cdot \frac{\theta}{2} \cdot (1 - \theta) \cdot \frac{\theta}{2} \cdot \frac{\theta}{2} = \left(\frac{\theta}{2}\right)^4 (1 - \theta)$$

hypothesis	prior	likelihood P(data | hypoth.)	BN	posterior
θ	$3\theta^2 \, d\theta$	$\left(\frac{\theta}{2}\right)^4 (1-\theta)$	$\frac{3}{16}\theta^6(1-\theta) \, d\theta$	$56 \cdot \theta^6(1-\theta) \, d\theta$
sum	1		$p(\text{data}) = \frac{3}{16} \cdot \frac{1}{56}$	1

Here, the total probability $p(\text{data})$ is computed by integrating the Bayes numerator column

$$p(\text{data}) = \int_0^1 \frac{3}{16}\theta^6(1 - \theta) \, d\theta = \frac{3}{16} \cdot \frac{1}{56}.$$

So, the posterior pdf for θ is $\boxed{f(\theta|\text{data}) = 56 \cdot \theta^6(1 - \theta).}$

(We could also have found the normalizing factor by recognizing the posterior as a Beta$(7, 2)$, so the factor is $\dfrac{8!}{6!1!} = 56$.)

Problem II.3. (12 pts.) *It is the year 2122 and a small percentage of children are born with an array of superpowers. The usual super strength and ability to make bad jokes in the direst of dire situations won't manifest themselves till puberty. The one superpower that*

manifests at age 7 is known as Bayesian intelligence. They can think clearly about statistics and can answer virtually any stats question.

So, a screening test was developed that asks 7 year-olds to compute the posterior odds that a 7 year-old who correctly answers the screening question has superpowers. The test is quite accurate, but some ordinary children do answer correctly

Suppose the odds that a random 7 year-old has superpowers is 1/100. Suppose the screening test has a 100% true positive rate and a 10% false positive rate. If a randomly chosen child correctly answers the question, what are the posterior odds that they have superpowers?

Solution: The easiest way to do this is to use the formula:

$$\text{posterior odds} = \text{prior odds} \times \text{likelihood ratio.}$$

The prior odds are given as 1/100.

The likelihood ratio $= \dfrac{P(\text{correct answer} \mid \text{superpower})}{P(\text{correct answer} \mid \text{no superpower})} = \dfrac{1}{1/10} = 10.$

So: $\boxed{\text{posterior odds} = \dfrac{1}{100} \cdot 10 = \dfrac{1}{10}.}$

Problem II.4. (15 pts: 3,3,3,3,3)
You collect data from an experiment and do a one-sided Z-test with the rejection region in the right tail and significance level 0.1. You find the Z-value is 2.

(a) Which R code computes the critical value for the rejection region?

(i) `pnorm(0.1, 0, 1)`   (ii) `pnorm(0.9, 0, 1)`   (iii) `pnorm(0.95, 0, 1)`
(iv) `pnorm(2.0, 0, 1)`   (v) `1 - pnorm(2.0, 0, 1)`   (vi) `qnorm(0.05, 0, 1)`
(vii) `qnorm(0.1, 0, 1)`   (viii) `qnorm(0.9, 0, 1)`
(ix) `qnorm(0.95, 0, 1)`

(b) Using the tables at the end of the exam, compute this critical value.

(c) Which R code computes the p-value for this experiment?

(i) `pnorm(0.1, 0, 1)`   (ii) `pnorm(0.9, 0, 1)`   (iii) `pnorm(0.95, 0, 1)`
(iv) `pnorm(2.0, 0, 1)`   (v) `1 - pnorm(2.0, 0, 1)`   (vi) `qnorm(0.05, 0, 1)`
(vii) `qnorm(0.1, 0, 1)`   (viii) `qnorm(0.9, 0, 1)`
(ix) `qnorm(0.95, 0, 1)`

(d) Using the tables at the end of the exam, compute the p-value.

(e) Should you reject the null hypothesis?

(a) Solution: (viii) `qnorm(0.9, 0, 1)`: the critical value for the right-tail should have 0.1 probability to the right, so 0.9 probability to the left.

(b) Solution: 1.28: the place where the left tail of the normal is 0.9. The table says that's between $z = 1.25$ and 1.30; interpolation says that it's about 1.28. (R says it's about 1.281552.)

(c) Solution: (v) `1 - pnorm(2.0, 0, 1)`: $p = P(Z > 2|H_0) = 1 - P(Z < 2|H_0)$. This is exactly what the code in (v) computes.

(d) Solution: 0.0228. This is the right tail of the standard normal at 2, or 1 minus the left tail; the table says the left tail is 0.9772.

(e) Solution: Yes, we should reject: according to (b), the value $z = 2$ is well inside the rejection region $z \geq 1.28$. Alternatively, the p-value computed in (d) is much smaller than the significance 0.1. Alternatively, the rule of thumb for 2 standard deviations above the mean for a normal distribution says $p = P(Z \geq 2|H_0) \approx 0.025 < 0.1$.

Problem II.5. (15 pts)

Adult onset diabetes is known to be highly genetically determined. A study was done comparing frequencies of a particular allele in a sample of diabetics and a sample of nondiabetics. The data are shown in the following table. (We adjusted the data to make hand calculation easier.)

	Diabetic	*Normal*
Bb or bb	*15*	*5*
BB	*35*	*45*

Do a significance test for whether the frequencies of the alleles is different in the two groups at a significance level of 0.05.

Solution: We will use a chi-square test for homogeneity. Remember we need to use all the data!. For hypotheses we have:

H_0: the frequency of alleles is the same for both groups: diabetic and nondiabetic.

H_A: the frequency of alleles is different between the two groups.

Here is the table of counts. The computation of the expected counts is explained below.

	Diabetic		Nondiabetic		
	observed	expected	observed	expected	
Bb, bb	15	10	5	10	20
BB	35	40	45	40	80
	50		50		100

The expected counts are computed as follows. Under H_0 the frequency of alleles is the same, call them θ. To find the expected counts we find the MLE of θ using the combined data:

$$\hat{\theta} = \frac{\text{total Bb and bb}}{\text{total subjects}} = \frac{20}{100} = 0.2.$$

Then, for example, the expected number of Bb and bb alleles in the diabetic group is $50 \cdot \hat{\theta} = 10$. The other expected counts are computed in the same way.

The chi-square test statistic is

$$X^2 = \sum \frac{(\text{observed - expected})^2}{\text{expected}} = \frac{5^2}{10} + \frac{5^2}{10} + \frac{5^2}{40} + \frac{5^2}{40} \approx 2.5 + 2.5 + 0.625 + 0.625 \approx 6.25.$$

Finally, we need the degrees of freedom: $df = 1$. This is because we have a two-by-two table and $(2-1) \cdot (2-1) = 1$. (Or because we can freely fill in the count in one cell and still be consistent with the marginal counts 20, 80, 50, 50, 100, which are all used to compute the expected counts.)

From the χ^2 table: $p = P(X^2 > 6.25|df = 1) < 0.05$, in particular interpolated from the table, $0.01 < p < 0.025$, but we can also see this directly by noting that the critical value is 3.84.

Conclusion: Since $p < \alpha$, we reject H_0 in favor of H_A, that the frequency of alleles is not the same between the diabetic and nondiabetic groups.

MIT OpenCourseWare

https://ocw.mit.edu

18.05 Introduction to Probability and Statistics

Spring 2022