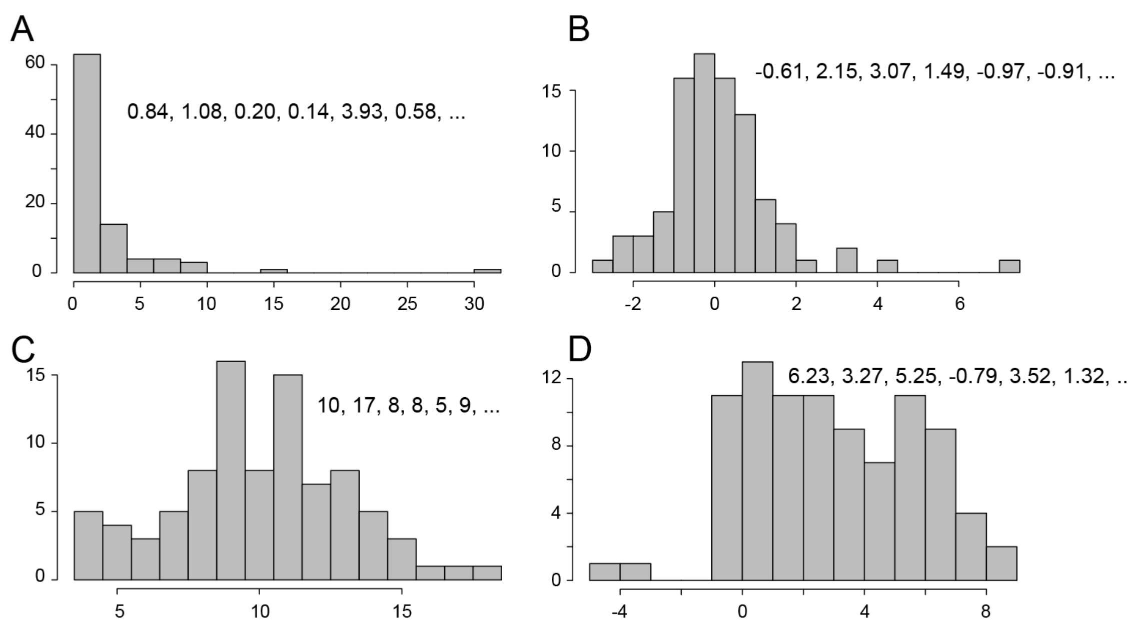
Biometry mid-term exam Spring 2016

(235 pts total)

Section 1 – Short answer

1. (20 pts) Below are four histograms of data, along with a sample of six values from that distribution. For each panel, list **all** the possible distributions from the table below from which these data may have originated. (This is different from asking which distributions might be used to describe the data in an analysis. Here I am asking which distributions, strictly speaking, could have actually generated the data in each plot.)

Normal	Standard normal	Log-normal
Poisson	Binomial	Gamma
Chi-squared	F	t
Beta	Multinomial	Negative binomial



A)

B)

C)

D)

2. (15 pts) X is a random variable drawn from the following discrete probability distribution

X	0	1	2	3
P(X)	0.4	0.1	0.2	0.3

Use the Central Limit Theorem to find the distribution of $\bar{X}$ as $n \to \infty$. (Your answer should include the sample size n since I have not specified it, other than to declare it large enough that the Central Limit Theorem can be applied.)

3. (15 pts) Assume two sets of independent random variables

$$X \sim N(\mu, \sigma^2)$$

and

$$Y \sim Pois(\lambda)$$

and a composite variable

$$Z = a + bX + cY$$

where a,b,c are constants, and X and Y are the random variables described above.

a) (7.5 pts) What is E[Z] in terms of μ, σ^2, λ? (Answer need not involve all three parameters.)

b) (7.5 pts) What is Var[Z] in terms of μ, σ^2, and λ? (Answer need not involve all three parameters.)

4. (10 pts) Describe Karl Popper's hypothetico-deductive method.

Section 2 – Long answer

5. (25 pts) Suppose X_1, X_2,…, X_n are i.i.d. random variables with probability density function

$$f(x|\sigma) = \frac{1}{2\sigma} e^{-\frac{|x|}{\sigma}}$$

A) (5 pts) Find the E[X]. (Hint: This does not require any integration.)

B) (20 pts) Find the maximum likelihood estimator for the parameter σ.

6. (50 pts) 4. Bliss and R.A. Fisher (1953) counted female European red mites (*Panonychus ulmi*) on McIntosh apple trees (*Malus domestica*). They counted 172 mites distributed across 150 leaves as described in the following table:

Mites per leaf	0	1	2	3	4	5	6	7	8+
Leaves observed	70	38	17	10	9	3	2	1	0

A) (10 pts) Assume that Bliss and Fisher set out to collect 150 leaves (no more, no less) and record the number of mites on each of these 150 leaves. What is the best statistical distribution for describing the data they collected, under the null hypothesis that all the mites are independent from one another?

B) (15 pts) Describe a parametric statistical test that you could use to test the null hypothesis of independence among the mites. You must state the test statistic, but are not required to state its distribution under the null hypothesis.

C) (15 pts) Describe a randomization based non-parametric statistical test that you could use to test the null hypothesis of independence among the mites. You must state the test statistic AND describe how you would generate the distribution of that test statistic under the null hypothesis. You may (but are not required) include R code to help describe your approach.

D) (10 pts) Assume that Bliss and Fisher had actually set out to collect 172 mites (no more, no less) and record the number of mites on each leaf collected. What is the best statistical distribution for describing the data they collected under this experimental design, under the null hypothesis that all the mites are independent of one another?

7. (40 pts) Brosi and Biber wrote a review in 2009 on the use of statistical inference in conservation. I show the abstract below to provide some context for their discussion.

REVIEWS REVIEWS REVIEWS

Statistical inference, Type II error, and decision making under the US Endangered Species Act

Berry J Brosi[1] and Eric G Biber[2]

Critical conservation decisions have been made based on the spurious belief that "no statistically significant difference between two groups means the groups are the same". We demonstrate this using the case of the Preble's meadow jumping mouse (*Zapus hudsonius preblei*), an endangered species in the US. Such faulty statistical logic has been recognized before, but ecologists have typically recommended assessing post hoc statistical power as a remedy. Statisticians, however, have shown that observed power will *necessarily* be low when no differences are found between two populations. Alternatives to assessments of statistical power include equivalence testing (a method rarely used by ecologists) and Bayesian or likelihood methods. Although scientists play a central role in ameliorating this problem, the courts could also assist by requiring litigated federal agency decisions to consider the risks of both Type I and Type II errors.

Front Ecol Environ 2009; 7(9): 487–494, doi:10.1890/080003 (published online 6 Nov 2008)

In their paper, the authors make the following comments in regard to the difficulty of interpreting genetic tests (to determine if two populations are genetically distinct) that fail to find statistically significant differences:

> The fundamental cause of this confusion is that standard statistical tests are set up in a way that gives researchers relative certainty about the result only when the test shows a significant difference between two groups. By definition, you can be at least 95% sure that you are correct when a hypothesis test finds a significant difference (assuming the standard $\alpha = 0.05$; Figure 2, top row). But when the test outcome is not significant (Figure 2, bottom row), there is no way to reliably estimate how likely you are to be wrong (from Type II error) if you conclude that the populations in question are homogeneous (eg Hoenig and Heisey 2001). Thus, if a significant difference is not found in a statistical test, the only appropriate conclusion is that the *null hypothesis cannot be rejected.*

Biometry mid-term exam Spring 2016

a) (15 points) The authors make some valid points in this paragraph: Name one of them.

b) (15 pts) The authors make at least one major error in this paragraph: What is it? Why are the authors incorrect?

8. (60 points) Assume that we have data on the probability of plant germination across species, and we are going to model that probability as a Beta distributed random variable.

$$X \sim Beta(\alpha, \beta)$$

and we use maximum likelihood to estimate the parameters of this Beta distribution $\hat{\alpha}$ and $\hat{\beta}$. We have two ways to express our uncertainty regarding our estimates of $\hat{\alpha}$ and $\hat{\beta}$: standard errors and confidence intervals.

a) (10 points) Describe in words the interpretation of the standard error of a parameter estimate ($\hat{\alpha}$ or $\hat{\beta}$) [Hint: The answer I'm looking for has something to do with *repeating* your experiment.]

b) (10 pts) Describe in words the interpretation of the confidence interval associated with your parameter estimates ($\hat{\alpha}$ or $\hat{\beta}$). [Hint: Once again, the answer I'm looking for has something to do with *repeating* your experiment.]

c) (20 points each = 40 pts) Describe two ways to construct confidence intervals for a parameter estimate? [I'm looking for one parametric and one non-parametric approach.]