Name:

- The exam is open book and open notes.
- Do all your work in the spaces provided. If you need additional space for your work, indicate **clearly** where the additional work can be found.
- The parts within a problem are not necessarily sequential.
- To receive full credit, you must show your work.
- Do not dwell too long on any one question. Answer as many questions as you can.

For instructor's use:

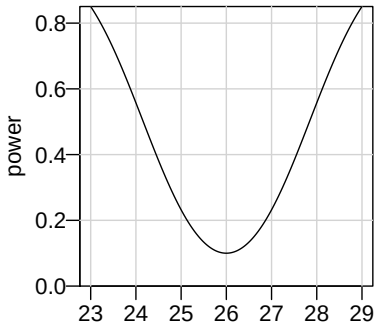
1	8	
2	22	
3	14	
4	6	
5	27	
6	7	
7	16	
Total	100	

1. An experiment is planned to evaluate the mean number of song birds in a certain tropical forest. The researchers will sample several 1km^2 plots. In each plot, they will count the number of bird species present through song recording. They will test the null hypothesis that the mean number of bird species per km^2 equals some fixed value. With their previous experience, they already have a good idea of the variability in species number. Therefore, they made the power curve below.

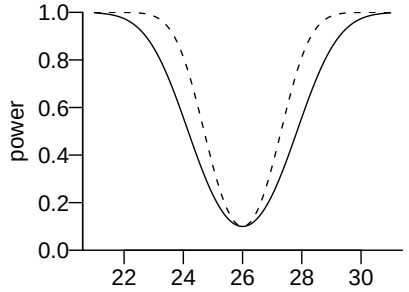
(a) What is the fixed value of the mean that the researchers are using in their null hypothesis?

(b) What level α are the researchers planning to use?

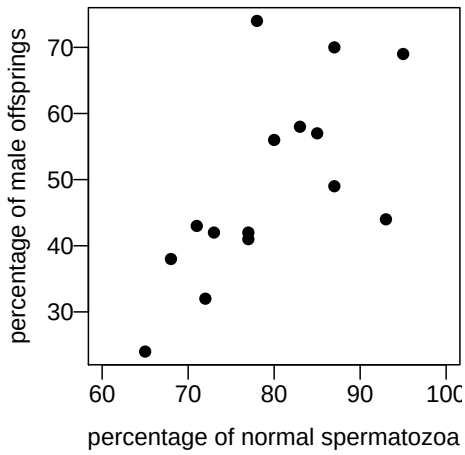
(c) What size of difference between the null hypothesis and the alternate hypothesis will have an 80% chance to be detected (if true)?



(d) The researchers are at the stage of planning their experiment and deciding how many plots should be surveyed. They made these two power curves below. Which one corresponds to the largest number of plots? ☐ the black curve ☐ the dotted curve.



2. It is now well known that mothers can manipulate sex ratios at birth in many species, and researchers wanted to know if fathers can also influence offspring sex ratio. A study was conducted on red deers (*Cervus elaphus*). Sperm from 15 wild males was collected. 360 hinds (females) were kept under similar environmental conditions and provided with unlimited food supply. Each hind was artificially inseminate once. Sperm quality was measured as the percentage of morphologically normal spermatozoa. The number of sons and daughters each male had was counted in order to obtain the percentage of male offspring for each male. The data is presented here. Calculations yield $\bar{x} = 79.4$, $\bar{y} = 49.27$, $\sum(x_i - \bar{x})^2 = 1105.6$, $\sum(y_i - \bar{y})^2 = 2916.93$ and $\sum(x_i - \bar{x})(y_i - \bar{y}) = 1192.4$. Also, SSErr = 1630.92.



Male	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
Sperm quality (x)	65	72	68	73	71	77	93	77	87	85	80	83	87	78	95
Proportion of male offsprings (y)	25	33	38	42	43	41	44	42	49	57	56	58	70	74	69

(a) Calculate the slope and intercept of the least squares regression line.

(b) Test the null hypothesis that the slope is significantly different from zero. Calculate an approximate p-value and state the conclusion.

(c) Calculate a 95% prediction interval for the percentage of male offsprings for a new red deer male with 65% normal spermatozoa.

3. Individuals from the fish species *Telmatherina sarasinorum* are of 3 different colors: females are grey, whereas males are either blue or yellow. It is hypothesized that fish with different colors may prefer to stay in different environments. Two types of environments were surveyed: shallow beach sites (yellow environment), and deeper sites with overhanging roots (blue environment). Many sites from each type were surveyed the same day, and at the same hour of the day, and fish from each color were counted. Below are the data.

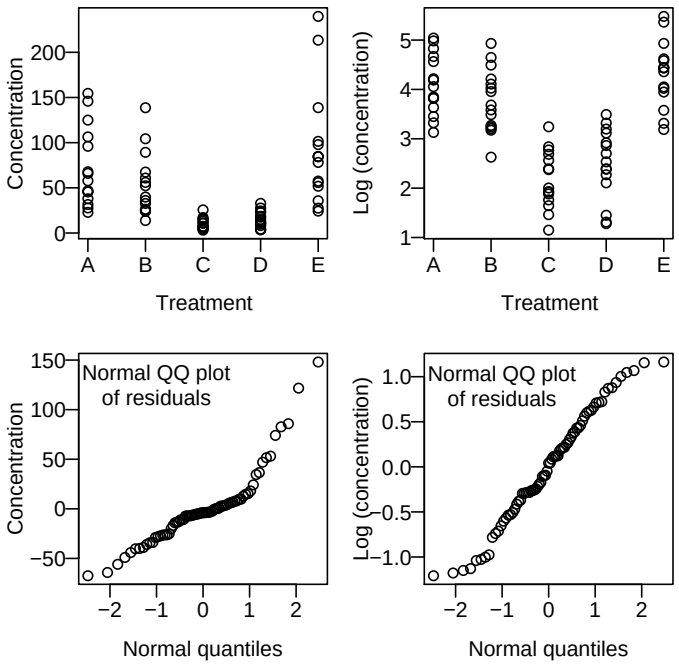
	Fish color			
Environment	blue	yellow	grey	Total
blue	14	22	36	72
yellow	25	12	31	68
Total	39	34	67	140

(a) A chi-square test of this data yields $X^2 = 6.3077$. Calculate the associated p-value, and state the conclusion of the test in the context of the study.

(b) Calculate the contribution of the cell "grey fish in blue environment" to the X-squared value.

(c) In question 3(a), what were the assumptions of the chi-square test, and were they met? Justify your answers shortly.

4. The following data represent concentration of a chemical after 5 specific treatments, with 15 data points per treatment. The raw data is shown, along with the data transformed by taking logarithms. The analysis of variance on the raw data gives an F value of 13.1, and an Anova on the log-transformed data yields an F value of 34.7. Would you choose to perform the F test on the ☐ raw data or on the ☐ log-transformed data? Give two reasons why.



5. The following data represent the survival times (hours) of creatures created by a mad scientist using four different reanimation techniques. Seven (7) creatures were animated using each technique. Means and sample standard deviations from each method are given here.

Technique	Mean	Sample SD
A	1.1	0.84
B	2.6	1.30
C	6.8	1.15
D	10.3	1.26

(a) A partial ANOVA table is provided. Fill in the missing boxes (show your work!), and perform an F-test at $\alpha = 0.01$ to determine if any of the techniques result in significantly different survival times. Note: the normal scores plot and the residuals vs. fitted plot show no obvious deviations from normality or constant variance.

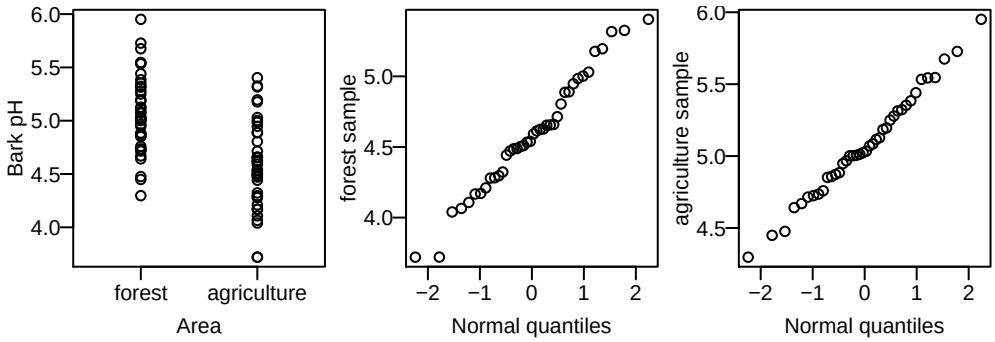
Source	df	SS	MS	F
Treatment			92.17	
Error				
Total				

(c) The mad scientist declared ahead of time that he only wanted to make the comparisons μ_A vs μ_B, μ_A vs μ_D, and μ_B vs μ_C. Test these three hypotheses using the three following methods: Fisher's LSD, Bonferroni's method, and (Tukey's) QD at $\alpha = 0.01$.

6. It was hypothesized that fertilizers may affect tree bark pH in agriculture areas. To assess this hypothesis, 10km by 10km regions were randomly selected in Wisconsin, and 20 plots were randomly selected in each region: 10 plots in forest areas and 10 plots in agriculture areas. In each plot, red oak trees were surveyed and their bark pH was measured. Measures from different red oak trees were averaged to yield a single measure for each plot. The data from all plots were used in a two independent sample t-test (assuming equal variances). Here is an output obtained with R as well as plots of the data.

```
        Two Sample t-test

data:  bark.forest and bark.agriculture
t = -5.6417, df = 78, p-value = 2.599e-07
alternative hypothesis: true difference in
                    means is not equal to 0
95 percent confidence interval:
 -0.6617658 -0.3165402
sample estimates:
mean of x mean of y
 4.586584  5.075737
```



(a) How many 10km by 10km regions were sampled?

(b) Assess the assumptions underlying the statistical analysis: what are they, and are they met in this case?

7. True/False questions.

(a) Probability uses a sample to make inferences about a population. ☐ True ☐ False

(b) For continuous random variables W, X and Y, the variance $\text{var}(W + X + Y)$ is always equal to the sum $\text{var}(W)$+ $\text{var}(X)$+ $\text{var}(Y)$. ☐ True ☐ False

(c) The sample standard deviation can sometimes be negative. ☐ True ☐ False

(d) For a random variable X with a binomial distribution $\mathcal{B}(n, p)$, as np and $n(1-p)$ get big, the distribution of X looks more and more like that of a normal random variable. ☐ True ☐ False

(e) The mean of the t distribution is always zero, regardless of its degrees of freedom. ☐ True ☐ False

(f) The p-value is defined to be the probability that the null hypothesis is true. ☐ True ☐ False

(g) Suppose $[12, 32]$ is a 95% confidence interval for μ calculated from a single sample. It means $P\{12 \leq \mu \leq 32\} = 0.95$. ☐ True ☐ False

(h) If the 3/4 sample quantile of a set of data is 20, then approximately one quarter of the data is greater than 20. ☐ True ☐ False