IDS 702 Midterm Practice Questions

Formula Sheet (this will be provided)

Any specific probability distribution functions that are needed will be given with the question (as seen in HW 2 question 1)

$Pr(A|B) = \frac{Pr(A \cap B)}{Pr(B)}$

$Pr(A \cup B) = Pr(A) + Pr(B) - Pr(A \cap B)$

Independence: $Pr(A|B) = Pr(A)$, $Pr(A \cap B) = Pr(A)Pr(B)$

Bayes Theorem: $Pr(B|A) = \frac{Pr(A|B)Pr(B)}{Pr(A)}$

General formula for confidence interval: $\bar{x} \pm z * \times \sigma / \sqrt{n}$

Practice Questions

Multiple choice

1. A researcher wants to assess the relationship between hospitalized patients' perceived quality of care and systolic blood pressure. The researcher would like to obtain a representative sample but cannot afford to travel to more than 10 hospitals. Which type of sampling would be most useful in this case?

   a. Simple random sampling

   b. Cluster sampling

   c. Stratified sampling

   d. A representative sample cannot be obtained in this case.

The table below presents table comparing two treatments to remove kidney stones. Use the table to answer questions 2 and 3.

	Successful	Not successful	Total
Open surgery	273	77	350
Small incision	289	61	350
Total	552	148	700

2. What is the conditional probability that a kidney stone is successfully removed for open surgery-treated patients?

   a. 273/350

   b. 273/552

   c. 273/700

   d. 552/700

3. What is the joint probability that a kidney stone is successfully removed for open surgery-treated patients?

a. 273/350

b. 273/552

c. 273/700

d. 552/700

4. What is a likelihood function?

a. A function of the observed data x_i based on the fixed parameter of interest

b. A function of the unobserved data x_i based on the joint data distribution

c. A function of the random population parameter based on the fixed joint data distribution

d. A function of the fixed population parameter based on the joint data distribution

5. Which of the following is true of bootstrapping?

a. We sample with replacement to simulate the sampling distribution of the sample statistic

b. We sample without replacement to simulate the population distribution of the population parameter

c. We sample with replacement to simulate the sampling distribution of the population parameter

d. We sample without replacement to simulate the population distribution of the sample statistic

6. Which of the following is NOT a requirement to conduct a two-sample parametric t-test?

a. Independent observations

b. Normal distribution of the population/large sample

c. The two samples are related

d. We want to assess the mean

7. Which of the following is an accurate distinction between simulation-based inference and parametric inference?

a. Simulation-based inference assumes that the null distribution is a normal distribution, while parametric inference assumes that the null distribution is a t-distribution.

b. Simulation-based inference uses resampling methods to approximate the null distribution, while parametric inference assumes that the test statistic takes a well-defined probability distribution.

c. Simulation-based inference can only be used for the mean, while parametric inference can be used for other sample statistics such as the median.

d. Simulation-based inference assumes that the parameter of interest is a fixed value, while parametric inference assumes that the parameter of interest is normally distributed.

8. The null and alternative hypotheses are always in terms of the _____.

a. sample statistic

b. population parameter

c. probability distribution

d. population mean

9. Increasing the confidence level _____ the width of the confidence interval and increasing the sample size _____ the width of the confidence interval.

   a. increases; increases

   b. increases; decreases

   c. decreases; increases

   d. decreases; decreases

10. Which type of question can NOT be answered with a linear model?

    a. On average, how much does Y change per increase in the variable X?

    b. On average, per increase in Y, how much does X change?

    c. What is the predicted value of Y for a given value of X?

    d. What is the average value of Y when $X = 0$?

11. What does it mean to fit a linear model?

    a. Calculate the residuals using the fitted model

    b. Compute the coefficient estimates by minimizing the sum of squared residuals

    c. Calculate the outcome estimates by minimizing the coefficient estimates

    d. Compute the residuals by maximizing the objective function

12. Why is it often important to use multiple linear regression instead of simple linear regression?

    a. Multiple linear regression extends simple linear regression by assessing the relationship between a numeric predictor and a categorical outcome

    b. Multiple linear regression extends simple linear regression by assessing the relationship between a numeric predictor and a numeric outcome

    c. Multiple linear regression allows for the control of variables that are related to both the independent variable and the outcome

    d. Multiple linear regression allows for the control of correlated predictor variables

13. Which of the following is NOT true about interaction terms in a linear model?

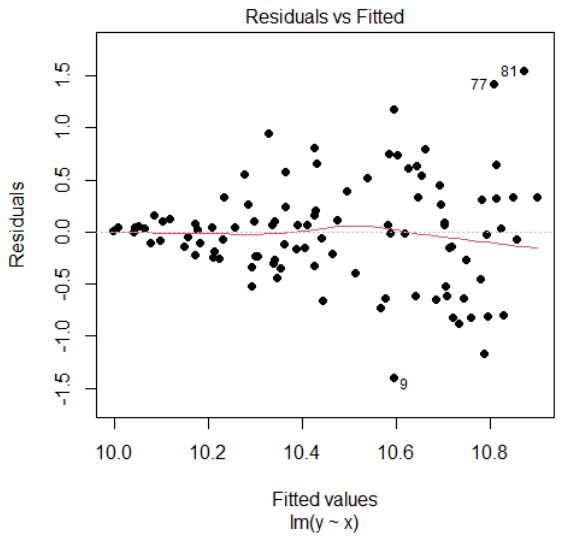
a. Interaction terms assess the effect of one predictor on the outcome based on the value of a second predictor

b. Interaction terms can be included in a model based on the research question or exploratory data analysis

c. Interaction terms can generate different slopes for a continuous predictor for different levels of a categorical predictor

d. Interaction terms should be used if two predictors both have an effect on the outcome

14. The diagnostic plot below indicates that which assumption of linear regression may be violated?



a. Linear relationship between the outcome and predictors

b. Independence of error terms

c. Errors are normally distributed

d. Errors have equal variance

15. A researcher fits a model regressing house listing price on number of bedrooms, number of bathrooms, city per capita crime rate, and distance to nearest grocery store (stored categorically as 0-1 mile, 1-5 miles, or more than 5 miles). The researcher has data on 1500 homes. What is the dimension of the design matrix for this model?

    a. 1500×4

    b. 1500×5

    c. 1500×6

    d. 1500×7

Short answer

16. This question uses a dataset containing information on 150 births in North Carolina in 2004.

Codebook:

m_age Mother's age.

weeks Weeks at which the mother gave birth.

premature Indicates whether the baby was premature or not.

weight Birth weight of the baby (lbs).

Smoke Whether or not the mother was a smoker.

```
Rows: 150
Columns: 5
$ m_age     <int> 30, 36, 35, 40, 37, 28, 35, 21, 20, 25, 19, 34, 19, 33, 27, ~
$ weeks     <int> 39, 39, 40, 40, 40, 40, 28, 35, 32, 40, 32, 40, 41, 38, 39, ~
$ premature <fct> full term, full term, full term, full term, full term, full ~
$ weight    <dbl> 6.88, 7.69, 8.88, 9.00, 7.94, 8.25, 1.63, 5.50, 2.69, 8.75, ~
$ smoke     <fct> smoker, nonsmoker, nonsmoker, nonsmoker, nonsmoker, smoker, ~
```

a. Which two variables in the births dataset can NOT be used as an outcome in a linear regression model? Justify your answer and assume that you have cleaned the dataset.

b. Write the theoretical model that regresses `weight` on `m_age`, `weeks`, and `premature`. Be sure to define each term (i.e., "Y= ——").

c. Using the output below, write the fitted model.

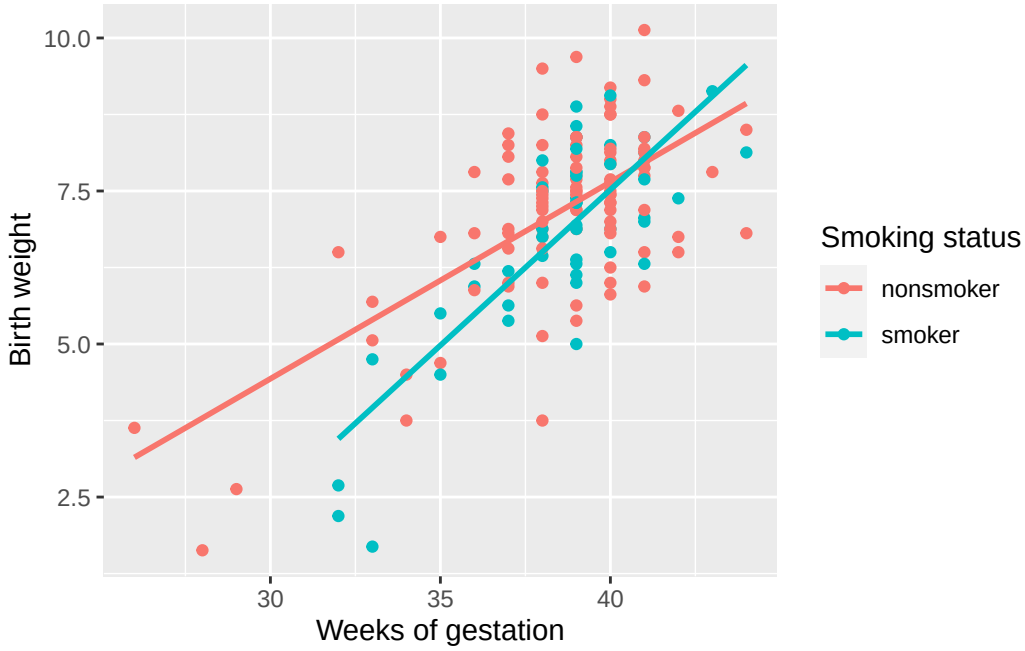
```
# A tibble: 4 x 5
  term             estimate std.error statistic     p.value
  <chr>               <dbl>     <dbl>     <dbl>       <dbl>
1 (Intercept)      -4.35      2.10       -2.07 0.0404
2 m_age             0.0270    0.0142      1.90 0.0594
3 weeks             0.281     0.0509      5.52 0.000000153
4 prematurepremie  -1.01      0.398      -2.54 0.0121
```

d. Write interpretations for the coefficient estimates of the `weeks` and `premature` variables.

`weeks`:

`premature`:

e. Write the theoretical model in mathematical notation that best matches the plot shown below. Be sure to define each term.



f. Write the implied linear models in mathematical notation for each level of smoking status. You do not need to define the terms again.

- Smoker:

- Nonsmoker:

g. In 2-3 sentences, describe what is being compared with the code below and the conclusion you can draw using the output.

```
Mod1 <- lm(weight~weeks+smoke, data=births)

Mod2 <- lm(weight~weeks*smoke, data=births)

summary(Mod1)$adj.r.squared
```

```
[1] 0.4760701
```

```
summary(Mod2)$adj.r.squared
```

```
[1] 0.496444
```

```
anova(Mod1, Mod2, test="F")
```

```
Analysis of Variance Table

Model 1: weight ~ weeks + smoke
Model 2: weight ~ weeks * smoke
  Res.Df    RSS Df Sum of Sq      F Pr(>F)
1    147 172.65
2    146 164.80  1    7.8424 6.9476 0.0093 **
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

17. **Answer each question in no more than 1-2 sentences.**

a. What is a sampling distribution?

b. How are the concepts of bootstrapping and sampling distributions related?

c. How is the central limit theorem related to the concept of a sampling distribution?