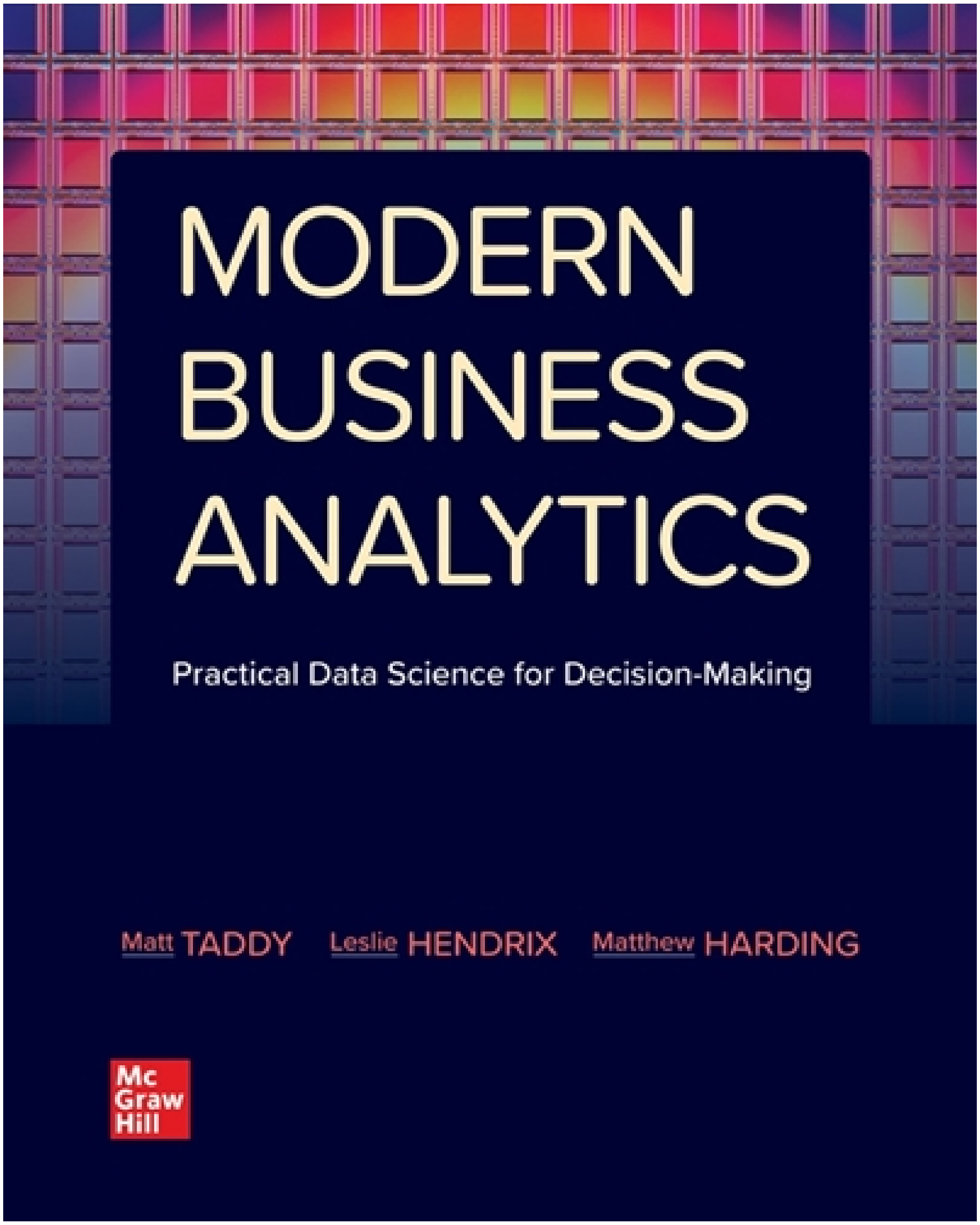
Test Bank for Modern Business Analytics 1st Edition by Taddy


MODERN BUSINESS ANALYTICS
Practical Data Science for Decision-Making
Matt TADDY
Leslie HENDRIX
Matthew HARDING
Mc Graw Hill

Test Bank

Modern Business Analytics Edition 1 by Taddy

CORRECT ANSWERS ARE LOCATED IN THE 2ND HALF OF THIS DOC.

CHECK ALL THE APPLY. Choose all options that best completes the statement or answers the question.

1) Which of the following are dangers of focusing exclusively on statistically significant *p*-values?

   A) In large sample sizes, tiny effects may be statistically significant but practically insignificant.

   B) In small sample sizes, tiny effects may become statistically and practically insignificant.

   C) Ignoring *p*-values below a certain threshold (*p*-hacking) can lead to bias.

   D) *p*-values cannot help you determine whether to reject the null hypothesis.

   E) A false positive rate can be much higher than stated *p*-values.

2) Call the amesHousing.csv data into R. Which of the following regression coefficients are significant when you control for a 5% false discovery rate? (Note: You can use the code below or create your own.)

```
pvals <- summary(amesfit)$coef[-1,"Pr(>|t|)"]
par(mfrow=c(1,2))
hist(pvals, col=8, breaks=10, xlab="p-values", main="", freq=FALSE)
plot(sort(pvals), xlab="rank", ylab="p-values")
fdr_cut <- function(pvals, q){
pvals <- pvals[!is.na(pvals)]
N <- length(pvals)
k <- rank(pvals, ties.method="min")
max(pvals[ pvals<= (q*k/N) ])
cutoff5 <- fdr_cut(pvals,q=.05)}
```

   A) Overall.Cond

   B) Year.Built

   C) Central.Air

   D) Electrical

   E) Gr.Liv.Area

3) Call the amesHousing.csv data into R. Which of the following regression coefficients are significant when you control for a 5% false discovery rate? (Note: You can use the code below or create your own.)

```
pvals <- summary(amesfit)$coef[-1,"Pr(>|t|)"]
par(mfrow=c(1,2))
hist(pvals, col=8, breaks=10, xlab="p-values", main="", freq=FALSE)
plot(sort(pvals), xlab="rank", ylab="p-values")
fdr_cut <- function(pvals, q){
pvals <- pvals[!is.na(pvals)]
N <- length(pvals)
k <- rank(pvals, ties.method="min")
max(pvals[ pvals<= (q*k/N) ])}
cutoff5 <- fdr_cut(pvals,q=.05)
```

A) Full.Bath
B) Half.Bath
C) Bedroom.AbvGr
D) Kitchen.AbvGr
E) TotRms.AbvGrd

MULTIPLE CHOICE - Choose the one alternative that best completes the statement or answers the question.

4) How does frequentist uncertainty quantification differ from Bayesian?

A) Frequentist approaches are based on considering how estimates would change if a new sample of data were generated by the same processes and scenarios; the Bayesian framework is based on a probabilistic construct based on the variance estimate from the data obtained.

B) Frequentist approaches are based on the assumption that the data generating process will change with each sample; the Bayesian approach assumes the data generating process is consistent.

C) Frequentist approaches rely on a probabilistic construct for the distribution based on the data obtained; the Bayesian approach asks how estimates would change if a new sample of data were generated using the same processes and scenarios.

D) Frequentist approaches update the variance estimate based on the data obtained from samples using different data generating processes; the Bayesian framework introduces a probabilistic construct for the distribution of the parameter and does not change the variance estimate.

5) Which of the following is a drawback of the frequentist approach?
   - A) It requires the assumption that the data generating process is consistent, but this is not always the case.
   - B) It is more dependent on the initial distribution assumptions, which makes it less accurate at predicting how estimates would change.
   - C) It assumes that the data generating process changes with each sample, so variance estimates are inconsistent.
   - D) It requires the variance estimate to be updated based on new data obtained.

6) Which of the following is drawback of the Bayesian approach?
   - A) It is more dependent on the initial distribution assumptions than is the frequentist approach.
   - B) It requires the assumption that the data generating process does not change.
   - C) It cannot update the variance estimate based on new data that is obtained.
   - D) It is based on a thought experiment rather than on real data.

7) What assumptions do the frequentist and Bayesian approaches rely on?
   - A) The frequentist approach assumes that the data generating process is consistent; the Bayesian approach is dependent on the initial distribution assumptions.
   - B) The frequentist approach is dependent on the initial distribution assumptions; the Bayesian approach assumes that the data generating process is consistent.
   - C) The frequentist approach and Bayesian approaches both assume that the data generating process is consistent.
   - D) The frequentist and Bayesian approaches both assume that the initial distributions will remain unchanged.

8) What is a sampling distribution?
   - A) The hypothetical distribution of a sample statistic computed from many theoretical random samples from the same data generating process
   - B) The actual distribution of the changes to the variance estimate from actual samples
   - C) The distribution of any statistic from actual and theoretical samples using various data generating processes
   - D) The hypothetical distribution of any statistic computed from many nonrandom samples selected using a changing data generating process

9) Consider the following statement: *A sampling distribution is a sample of data displayed on a histogram.* What is wrong with this statement?

A) A histogram of sample data shows one iteration of the hypothetical samples. A sampling distribution shows a sample statistic from many hypothetical examples.
B) Each bar on a histogram becomes one datapoint on a sampling distribution.
C) A histogram contains actual data, while a sampling distribution contains theoretical data.
D) A histogram contains theoretical data, while a sampling distribution contains actual data.

10) What is the relationship between a histogram and a sampling distribution?

A) A histogram is distilled into a single point in the sampling distribution.
B) A histogram contains data from many theoretical samples, while a sampling distribution contains data from many actual samples.
C) Each bar on a histogram becomes one data point on a sampling distribution.
D) A sampling distribution and a histogram show the exact same data in different forms.

11) Define *p*-value.

A) The probability of obtaining a sample statistic as extreme, or more extreme than what was observed in the sample, assuming the null hypothesis is true
B) The probability of obtaining an actual statistic that is more extreme than what was observed in the sample, assuming the null hypothesis is true
C) The probability that the null hypothesis is true
D) The probability that the elasticity is statistically different from 0

12) What is the term that describes the probability of obtaining a sample statistic as extreme compared to what was observed in the sample, assuming the null hypothesis is true?

A) *p*-value
B) Confidence interval
C) Elasticity
D) Variance

13) What does $P(H_0 \textit{ is true})$ describe?

A) The probability of observing the data (or more extreme data), given the null
B) The probability that the null hypothesis is true, given the data
C) The sampling distribution, assuming the null hypothesis is true
D) The *p*-value

Modern Business Analytics Edition 1 by Taddy

14) The 95% confidence interval for the coefficient representing the difference in elasticity of Minute Maid orange juice from Dominick's was (−0.056, 0.169). Based on this information alone, what can we conclude about elasticity at the $\alpha = 0.05$ level?

A) We can conclude that the elasticity is not statistically different from 0.
B) We can conclude that the elasticity is not statistically different from 1.
C) We can conclude that the elasticity is 0.113.
D) We can conclude that the elasticity is 0.05.

15) Assume the 95% confidence interval for the coefficient representing the difference in elasticity is (−0.056, 0.169). At the $\alpha = 0.05$ level, why can we conclude that the elasticity is not statistically different from 0?

A) Because the confidence interval includes 0.
B) Because the upper bound of the confidence interval is less than 1.
C) Because the lower bound of the confidence interval is less than 1.
D) Because the mean of the confidence interval is 0.05.

16) Suppose a regression is run with 100 independent variables. If the significance level is set at 0.05 for each test, what would the false discovery rate (FDR) be if none of the variables were truly significant?

A) FDR = 1 − $(1 - 0.05)^{100}$(1 − 0.05)power of 100 = 0.0994
B) FDR = 1 ÷ (100 − 0.05) = 0.010
C) FDR = (100 × 0.05) ÷ 100 = 0.05
D) FDR = 100(1 − (1 − 0.05)) = 5

17) For a given set of predictor variable values, which is wider: the 95% confidence interval for the mean or the 95% prediction interval?

A) The 95% prediction interval is wider.
B) The 95% confidence interval for the mean is wider.
C) They are equal.
D) It depends on the value of the standard error.

18) For a given set of predictor variable values, why is the 95% prediction interval wider than the 95% confidence interval (CI) for the mean?

A) Because the uncertainty (standard error) of the prediction includes both the standard error for the fit (from the 95% CI) and the error in the observations via the residuals
B) Because the CI includes the error in observations via the residuals
C) Because the standard error of predictions is always less certain than that of the confidence interval
D) Because a prediction interval is always more accurate than a confidence interval

Modern Business Analytics Edition 1 by Taddy

19) How is the bootstrap utilized to generate a sampling distribution?

- A) Resample with replacement from your current sample many times. Calculate the statistic of interest from each of the resamples and present them in a single distribution.
- B) Use the same sample to calculate various statistics of interest and determine which best supports your hypothesis.
- C) Divide your current sample into many smaller subsamples. Create a histogram for each subsample.
- D) Resample with replacement from your current sample many times. Create a histogram for each subsample and use only the dataset that best fits your hypothesis.

20) What does the central limit theorem state?

- A) The average of independent random variables becomes normally distributed if your sample size is large enough.
- B) The standard error increases as the sample size increases.
- C) The variance of the sum of independent variables equals the standard error of the normal distribution.
- D) The average of independent random variables increases if the bootstrap is utilized.

21) The central limit theorem provides a theoretical framework for the sampling distribution of the sample mean. How is the sampling distribution discovered for other sample statistics that may not have a defined theoretical sampling distribution?

- A) The bootstrap
- B) The glm function
- C) The Gaussian distribution
- D) Conjugate models

22) Which of the following is true about the bootstrap?

- A) It is most practical with low-dimensional statistics.
- B) It is most practical with high-dimensional statistics.
- C) It eliminates the need to make modeling assumptions.
- D) It cannot be used in combination with Monte Carlo.

23) Why is the bootstrap not practically useful for approximating high-dimensional statistics?

- A) It requires an enormous observed sample to get enough information to summarize the covariances between the variables.
- B) It cannot be used with parallel computing.
- C) It has a high level of precision that cannot be changed to fit what is needed for a specific application.
- D) There is no way to observe the variability that occurs across resamples.

Modern Business Analytics Edition 1 by Taddy

24) Define robustness in terms of the bootstrap.

A) As long as the observations are independent, the bootstrap will provide a sampling distribution of the statistic, even if some model assumptions are incorrect.

B) As long as the observations are dependent, the bootstrap can provide a sampling distribution of the statistic without additional model assumptions.

C) As long as the bootstrap makes parametric model assumptions, it will provide a sampling distribution of the statistic.

D) As long as the bootstrap provides a sampling distribution of the statistic, you can be certain that the model assumptions are correct.

25) What key assumption of the bootstrap is often untrue in practice?

A) The bootstrap assumes that the observations are independent.

B) The bootstrap assumes that the observations are dependent.

C) The bootstrap assumes all models are parametric.

D) The bootstrap assumes that random sampling guarantees replicability.

26) When using the bootstrap, how can you avoid potential problems related to the assumption that observations are independent?

A) Resample blocks of observations that are potentially correlated with each other.

B) Resample individual observations to avoid correlation.

C) Ensure that the allowed error variance is constant.

D) Resample without replacement to ensure datapoints are not duplicated across samples.

27) When constructing a 90% confidence interval using the nonparametric bootstrap, what are the appropriate percentiles to choose?

A) The 5th and 95th percentiles

B) The 10th and 90th percentiles

C) The 2.5th and 97.5th percentiles

D) The 15th and 85th percentiles

28) What is the term for bootstrap estimates that consistently over/underestimate the truth?

A) Biased

B) Variable

C) Low confidence

D) Convex

Modern Business Analytics Edition 1 by Taddy

29) When do biased bootstrap estimates occur?
   A) When the log() transformation is applied
   B) When the function is convex
   C) When exponentiation is applied
   D) When the bootstrap is parametric

30) What is a key characteristic of the nonparametric bootstrap?
   A) It makes no model assumptions about the data generating process.
   B) It cannot provide standard errors that are robust to nonconstant variance.
   C) It provides accurate estimates of sampling distribution for datasets with very heavy tails.
   D) It is not affected by model selection.

31) What is a benefit of the parametric bootstrap?
   A) It generates bootstrap samples from a fitted model.
   B) It restricts its resamples to the observed data.
   C) It does not introduce new model assumptions.
   D) It works best when models are not flexible.

32) What is a potential negative effect of the parametric bootstrap's sample generation?
   A) It adds new assumptions.
   B) It removes correct assumptions.
   C) It restricts resamples to only observed data values.
   D) It can make the data seem less noisy than it really is.

33) What is a benefit of the parametric bootstrap's sample generation?
   A) It removes the restriction of only having observed data values in the resample.
   B) It makes the data appear less noisy than it actually is.
   C) It makes your uncertainty quantification less sensitive to new model assumptions.
   D) It removes randomness from sample generation.

34) Suppose you believe a customer will buy your product with a probability of 0.5. You then collect data on 20 incoming (independent) customers and find that 7 of them will purchase your product. Using the beta-binomial model (with a Beta(1,1) prior) on the probability *q* that a customer will buy your product, what is the updated expected value of the posterior probability of purchase?
   A) 0.364
   B) 0.444
   C) 0.175
   D) 0.700

Modern Business Analytics Edition 1 by Taddy

35) Call the amesHousing.csv data into R. Determine the 95% confidence interval for the effect of having central air on the expected log sale price using glm. What is the standard error?

A) 0.02
B) 0.01
C) 0.03
D) 0.04

36) Call the amesHousing.csv data into R. Determine the 95% confidence interval for the effect of having central air on the expected log sale price using bootstrap. What is the standard error of the bootstrap confidence interval?

A) 0.55
B) 0.24
C) 0.06
D) 0.48

Modern Business Analytics Edition 1 by Taddy

Answer Key

Test name: Chapter 02

1) [A, C, E]

With large sample sizes, tiny effects may become statistically significant, but are practically insignificant. In addition, there is the additional danger of ‘ *p*-hacking’ as it is common (unfortunately) for individuals or journals to only present significant *p*-values and ignore non-significant findings leading to bias.

2) [A, B, E]

```
pvals <- summary(amesfit)$coef[-1,"Pr(>|t|)"]
par(mfrow=c(1,2))
hist(pvals, col=8, breaks=10, xlab="p-values", main="", freq=FALSE)
plot(sort(pvals), xlab="rank", ylab="p-values")
fdr_cut <- function(pvals, q){
pvals <- pvals[!is.na(pvals)]
N <- length(pvals)
k <- rank(pvals, ties.method="min")
max(pvals[ pvals<= (q*k/N) ])
cutoff5 <- fdr_cut(pvals,q=.05)}
```

3) [A, B, C, D, E]

```
pvals <- summary(amesfit)$coef[-1,"Pr(>|t|)"]
par(mfrow=c(1,2))
hist(pvals, col=8, breaks=10, xlab="p-values", main="", freq=FALSE)
plot(sort(pvals), xlab="rank", ylab="p-values")
fdr_cut <- function(pvals, q){
pvals <- pvals[!is.na(pvals)]
N <- length(pvals)
k <- rank(pvals, ties.method="min")
max(pvals[ pvals<= (q*k/N) ])}
cutoff5 <- fdr_cut(pvals,q=.05)
```

4) A

Frequentist approaches examine uncertainty via the thought experiment “If I were able to see a new sample of data generated by the same processes and scenarios as my current data, how would my estimates change?”

5) A

Frequentist approaches require the assumption that the data generating process is consistent and Bayesian approaches are somewhat dependent on the initial distribution assumptions.

6) A
The Bayesian framework introduces a probabilistic construct for the distribution of the parameter and updates the variance estimate based on the data obtained.

7) A
The Bayesian framework introduces a probabilistic construct for the distribution of the parameter and updates the variance estimate based on the data obtained. Frequentist approaches require the assumption that the data generating process is consistent and Bayesian approaches are somewhat dependent on the initial distribution assumptions.

8) A
The sampling distribution is the distribution of a sample statistic (in most cases, the sample average) computed from many theoretical random samples from the same data generating process.

9) A
The sampling distribution is the hypothetical distribution of the sample averages from many different samples from the DGP. The histogram of the sample data is simply one iteration of the hypothetical samples. It would be distilled into a single point in the sampling distribution (the sample mean).

10) A
The sampling distribution is the hypothetical distribution of the sample averages from many different samples from the DGP. The histogram of the sample data is simply one iteration of the hypothetical samples. It would be distilled into a single point in the sampling distribution (the sample mean).

11) A
The *p*-value is the probability of obtaining a sample statistic as extreme or more extreme than what you observed in the sample assuming the null hypothesis is true. This is not the probability of the null being true given the data but rather the probability of observing the data (or more extreme) given the null.

12) A

The *p*-value is the probability of obtaining a sample statistic as extreme or more extreme than what you observed in the sample assuming the null hypothesis is true. This is not the probability of the null being true given the data but rather the probability of observing the data (or more extreme) given the null.

13) A

The *p*-value is the probability of obtaining a sample statistic as extreme or more extreme than what you observed in the sample assuming the null hypothesis is true. This is not the probability of the null being true given the data but rather the probability of observing the data (or more extreme) given the null.

14) A

Because the confidence interval includes 0, we can conclude (at the 0.05 significance level) that the elasticity is not statistically different from 0.

15) A

Because the confidence interval includes 0, we can conclude (at the 0.05 significance level) that the elasticity is not statistically different from 0.

16) A

Suppose H_0: slope = 0 is true for all 100 independent variables in regression, i.e., none of the variables were truly significant. The probability that all 100 *p*-values would be greater than 0.05 is $(1 - 0.05)^{100}$(1 – 0.05)power of 100 and the probability that at least one *p*-value would be less than 0.05 is $1 - (1 - 0.05)^{100}$(1 – 0.05)power of 100 = 0.994, the false discovery rate (FDR).

17) A

The 95% prediction interval will be wider because the uncertainty (standard error) of the prediction includes both the standard error for the fit (from the 95% CI) and the error in the observations via the residuals.

18) A

The 95% prediction interval will be wider because the uncertainty (standard error) of the prediction includes both the standard error for the fit (from the 95% CI) and the error in the observations via the residuals.

19) A

To generate a bootstrap distribution, resample *n* observations from your sample of *n* observations with replacement many times and calculate the statistic of interest from each of the resamples.

20) A

Modern Business Analytics Edition 1 by Taddy

Because not all statistics have a defined theoretical sampling distribution, the bootstrap can be utilized to discover the sampling distribution or other statistics.

21) A
Because not all statistics have a defined theoretical sampling distribution, the bootstrap can be utilized to discover the sampling distribution or other statistics.

22) A
Because not all statistics have a defined theoretical sampling distribution, the bootstrap can be utilized to discover the sampling distribution or other statistics. The bootstrap is fairly efficient for large dimension parameters. A rule of thumb states that having at least 100 observations per dimension of the parameter is sufficient for comfortable estimation via the bootstrap.

23) A
The bootstrap is fairly efficient for large dimension parameters. A rule of thumb states that having at least 100 observations per dimension of the parameter is sufficient for comfortable estimation via the bootstrap.

24) A
As long as the observations are independent, the bootstrap will provide a sampling distribution of the statistic without additional model assumptions about the data generating process.

25) A
The key assumption is that the observations are independent. If the observations aren’t independent, one can resample blocks of observations that are (potentially) correlated with each other.

26) A
The key assumption is that the observations are independent. If the observations aren’t independent, one can resample blocks of observations that are (potentially) correlated with each other.

27) A
A simple way to get a confidence interval after bootstrapping is to calculate the bounds directly on your sample of bootstrap estimates. This is called a *percentile* bootstrap confidence interval, as it is based on the percentiles of the sample of bootstrap estimates rather than some theoretical distribution, like the Gaussian. For example, you can use the quantile function to get the percentiles of the sampling distribution. The convention is to leave the same amount of probability in both upper and lower tails.

28) A

This occurs when the log() transformation is applied.

29) A
Biased estimates occur when the log() transformation is applied.

30) A
The nonparametric bootstrap makes no model assumptions about the DGP, whereas the parametric bootstrap generates bootstrap samples from a fitted model. This adds new assumptions but removes the restriction of only having observed data values in the resample.

31) A
The nonparametric bootstrap makes no model assumptions about the DGP, whereas the parametric bootstrap generates bootstrap samples from a fitted model. This adds new assumptions but removes the restriction of only having observed data values in the resample.

32) A
The nonparametric bootstrap makes no model assumptions about the DGP, whereas the parametric bootstrap generates bootstrap samples from a fitted model. This adds new assumptions but removes the restriction of only having observed data values in the resample.

33) A
The nonparametric bootstrap makes no model assumptions about the DGP, whereas the parametric bootstrap generates bootstrap samples from a fitted model. This adds new assumptions but removes the restriction of only having observed data values in the resample.

34) A
The updated expected value of the posterior probability of purchase is: (1 + 7) ÷ (1 + 1 + 20) = 8 ÷ 22 = 0.363636.

35) A

```
> amesfit <- lm(log(SalePrice) ~ Central.Air, data=ames)
> summary(amesfit)
Call:
lm(formula = log(SalePrice) ~ Central.Air, data = ames)
Residuals:
        Min          1Q      Median         3Q        Max
   -2.00124    -0.24833    -0.03071    0.23775    1.47311
    Coefficients:
                        Estimate   Standard    t value    Pr(>|t|)
                                    Error
      (Intercept)       11.45758    0.02705     423.57    <2e-16 ***
      Central.AirY       0.60378    0.02800      21.56    <2e-16 ***
    ---
    Significance codes: 0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
    Residual standard error: 0.3787 on 2928 degrees of freedom
    Multiple R-squared: 0.137, Adjusted R-squared: 0.1367
    F-statistic: 464.9 on 1 and 2928 DF, p-value: < 2.2e-16
    > confint(amesfit)
                                   2.5%        97.5%

      (Intercept)            11.4045384    11.510616
      Central.AirY            0.5488737     0.658688
```

36) A

```
( bstats <- summary(amesfit)$coef["Central.AirY",] ) #coefficient for Central Air
bstats["Estimate"] + c(-1,1)*1.96*bstats["Std. Error"] #95% CI
```