Test 2 Information

Stat 145

Fall 2020

Logistics

Date: Tuesday, October 27

Material Covered:

The test is cumulative, but it will emphasize the more recent topics (probability, confidence intervals, and hypothesis tests).

- The following sections of of *ISLBS*: Chapter 2 (probability)
- The following sections of *IMS*: 1.1-4, 2.1-4, 3.1-3, 5.1-2, 6.1-2, 7.1-7.3, but omitting the sub-sections labeled "Mathematical model".

Technology allowed: You will be allowed to use RStudio via your laptop. No other software may be running, and you may not browse to other sites. (Close everything else down prior to the test.)

Accommodations: If you require testing accommodations, please contact me so we can make the appropriate arrangements.

Topics

Here is a list of things you should be sure you know how to do. It is not intended to be an exhaustive list, but it is an important list. You should be able to:

- Understand, use and explain the statistical **vocabulary/terminology**.
  - Be sure to focus on important distinctions being made by terms like case vs. variable, statistic vs. parameter, sample vs. population, categorical vs. quantitative, sample vs. sampling distribution, sampling distribution vs. bootstrap distribution, etc.
  - Some other important terms: significance level, confidence level, margin of error, statistically significant, type I error, type II error, ~~critical value~~, paired design, blinding
- Work with **normal distributions** and **z-scores**
  - 68-95-99.7 Rule, `pnorm()`, `qnorm()`
- Understand the issues involved in collecting good data and the **design of studies**, including the distinction between observational studies and experiments, and when and how to use a paired design.
- Create and understand **confidence intervals**
  - what resampling is an why we use it to generate the bootstrap distribution
  - computing a CI from a bootstrap distribution (percentile method and standard error methods)
  - what the resulting interval tells you (the meaning of the confidence level, coverage rate, etc.)
  - recognizing incorrect ways to interpret a confidence interval and what is wrong with them
  - how to get RStudio to generate a bootstrap distribution
  - describing how to generate a bootstrap sample without a computer (cards, dice, coins, etc.)
  - resampling within groups
- Use the 4-step process for conducting a **hypothesis test**, including

  - expressing null and alternative hypotheses
  - determining a p-value from a randomization distribution or using SE and normal distributions
  - expressing the logic of a p-value in words (in the context of a particular example)
  - properties a randomization distribution must have
  - how to get RStudio to generate a randomization distribution
  - 1-tailed and 2-tailed tests (difference, when to use which)
- Connections between **p-values and confidence intervals**
  - Differences between randomization and bootstrap samples
  - Differences between randomization and bootstrap distributions
  - What confidence intervals tell you about p-values and vice versa (and why)
- **Use R** to compute numerical summaries, make plots, compute probabilities, create bootstrap and randomization distributions, compute p-values and confidence intervals.
  - Important functions to review include `gf_histogram()`, `gf_boxpot()`, `gf_point()`, `gf_lm()`, `gf_bar()`, `do()`, `df_stats()`, `diffmean()`, `diffprop()`, `cor()`, `pnorm()`, `qnorm()`, `cnorm()`, `xpnorm()`, `xqnorm()`, `xcnorm()`, `mutate()`, `lm()`.
  - Be sure to write the R command used and the result it produced on your test paper.
- **Probablity**
  - Make use of the rules of probability to determine probabilities of events.
    - Use probability rules to compute probabilities of combinations of events like
      - P(A and B)
      - P(A or B)
      - P(not A)
      - P(A if B)
  - Compute, interpret, and use conditional probabilities (tree diagrams, optional)
  - Sensitivity, specificity, prevalence, PPV (positive predictive value), NPV (negative predictive value)

Some sample problems

1. **What do I do?**

   In each of the following situations, pretend you want to know some information and you are designing a statistical study to find out about it. Give the following pieces of information for each: (i) what **variables** you would need to have in your data set, (ii) whether each variable is **categorical** or **quantitative**, (iii) the null and alternative hypotheses or the parameter for which you would create a confidence interval, (iv) whether the study is an observational study or an experiment, (v) particular design elements you would use (randomization (of what? how?), blinding, matched pairs design, etc.).

   a. You want to know if boys or girls score better on reading tests in Kent County grade schools.

2. For the same sorts of scenarios as above, you should be able to create a bootstrap or randomization distribution and use it to get a confidence interval or a p-value.

3. Be sure to **show some work** as you answer the following questions.

   A certain test is standardized in such a way that the mean score is 40 and the standard deviation is 5.

   a. What Z-score is associated with a test score of 48.5?
   b. Approximately what percentage of people score above 48.5 on the test?
   c. Approximately what percentage of people score between 37.0 and 48.5 on the test?
   d. Fred scored in the 65th percentile. What was his test score? What percent of the test takers did better than Fred?

4. A certain IQ test is standardized in such a way that the mean score is 100 and the standard deviation is 10. For parts b), c) and d) please sketch the distribution and shade the appropriate region.

   a. What Z-score is associated with a score of 117?

b. Aproximately what percentage of people have IQs between 90 and 110?
c. Approximately what percentage of people have IQs above 120?
d. Approximately what percentage of people have IQs between 90 and 120?

Solutions

1.

- Variables: Sex (Cat.); Reading Score (Quant.)
- $H_0 : \mu_m - \mu_f = 0$; $H_a : \mu_m - \mu_f \neq 0$
- Observational study (we can't assign genders to kids, and we can't assign their test scores either)
- We should make sure that the boys and girls are similar in terms of age, schools attended, etc. We might sample an equal number of boys and girls from the 3rd grade class at each school, for example.

2. Randomization

```
Randomization <-
  do (1000) * diffmean(reading_score ~ shuffle(sex), data = ReadingStudy)
Bootstrap <-
  do (1000) * diffmean(reading_score ~ sex, data = resample(ReadingStudy))
```

3. Standardized test

```
# z-score
(48.5 - 40) / 5
```

```
## [1] 1.7
```

```
# score above 48.5
1 - pnorm(48.5, mean = 40, sd = 5)
```

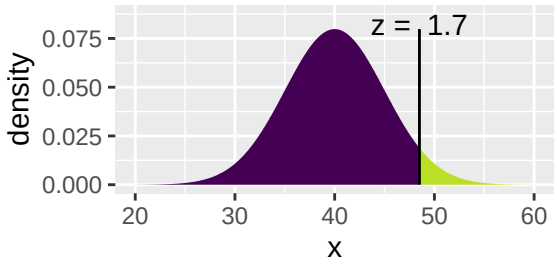
```
## [1] 0.0446
```

```
1 - pnorm(1.7)
```

```
## [1] 0.0446
```

```
xpnorm(48.5, mean = 40, sd = 5)
```

```
##
## If X ~ N(40, 5), then
##  P(X <= 48.5) = P(Z <= 1.7) = 0.9554
##  P(X >  48.5) = P(Z >  1.7) = 0.04457
##
```




```
## [1] 0.955
```

```
# between 37 and 48.5
pnorm(48.5, mean = 40, sd = 5) - pnorm(48.5, mean = 40, sd = 5)
```

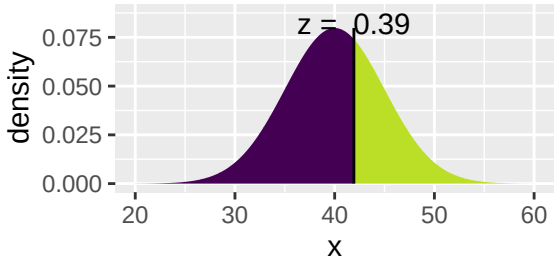
```
## [1] 0
```

```
# 65th percentile
qnorm(.65, mean = 40, sd = 5)
```

```
## [1] 41.9
```

```
xqnorm(.65, mean = 40, sd = 5)
```

```
##
## If X ~ N(40, 5), then
##  P(X <= 41.9) = 0.65
##  P(X >  41.9) = 0.35
##
```




```
## [1] 41.9
```

4. IQ test

These can all be approximated without using R.

a. $z = 17/10 = 1.7$
b. That's the middle 68%
c. That's 2 standard deviations out so 2.5% (95% in the middle, 2.5% in each tail)
d. 1/2 of 68% + 1/2 of 95% = 81.5%

Or we can use R.

```
# z-score
(117 - 100) / 10
```

```
## [1] 1.7
```

```
# between 90 and 100 -- should be about 68%
pnorm(110, mean = 100, sd = 10) - pnorm(90, mean = 100, sd = 10)
```

```
## [1] 0.683
```

```
pnorm(1) - pnorm(-1)
```

```
## [1] 0.683
```

```
# above 120 -- should be about 2.5%
1 - pnorm(120, mean = 100, sd = 10)
```

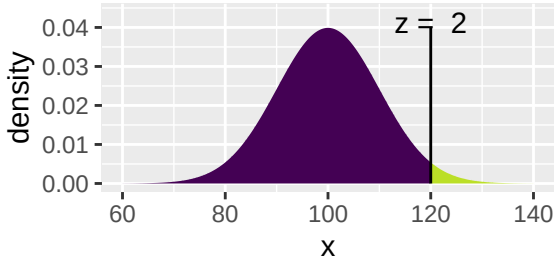
```
## [1] 0.0228
```

```
1 - pnorm(2)
```

```
## [1] 0.0228
```

```r
1 - xpnorm(120, mean = 100, sd = 10)
```

```
##
## If X ~ N(100, 10), then
##  P(X <= 120) = P(Z <= 2) = 0.9772
##  P(X >  120) = P(Z >  2) = 0.02275
##
```



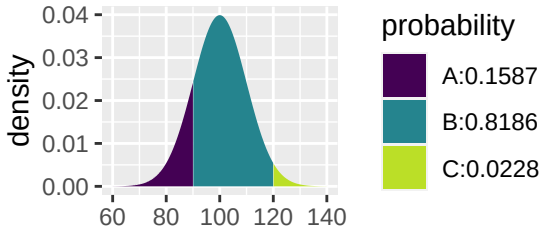

```
## [1] 0.0228
```

```r
# between 90 and 120
pnorm(120, mean = 100, sd = 10) - pnorm(90, mean = 100, sd = 10)
```

```
## [1] 0.819
```

```r
xpnorm(c(90, 120), mean = 100, sd = 10)
```

```
##
## If X ~ N(100, 10), then
##  P(X <=  90) = P(Z <= -1) = 0.1587   P(X <= 120) = P(Z <=  2) = 0.9772
##  P(X >   90) = P(Z >  -1) = 0.84134  P(X >  120) = P(Z >   2) = 0.02275
##
```




```
## [1] 0.159 0.977
```