Test 1 will cover the material in Chapters 7, 8 and 9 from the book that was discussed in class. First a summary of each chapter, and then some sample test questions.

Chapter 7 Chapter 7 was about summaries of data sets. We discussed methods to summarize univariate (single variable) numeric data sets. These included histograms, density estimates, boxplots, stem and leaf diagrams, and dot plots. We discussed the numeric summaries of center: mean and median; spread: standard deviation, variance, IQR; and position: order statistics and quantiles.

For bivariate data, we discussed the graphical techniques of side-by-side boxplots and scatterplots. As well for paired data, we discussed the correlation coefficient.

Chapter 8 Chapter 8 introduced some key concepts about statistics.

First we learned about the likelihood function: this reverses the role of the parameters and the data values typical of the pdf.

We looked at the idea of a population or parent distribution describing a data set. We defined a random sample from a population and the definition of a statistic summarizing a random sample. We tied this together with the concept of a sampling distribution.

We mentioned the central limit theorem which characterized the sampling distribution of the sample mean, when n is large. We also computed a few other sampling distributions, often for the statistic $X_{(n)}$.

Some language about statistics was introduced. A statistic in general is a function of the values in a data set. However, the statistics we consider are usually *estimators* for a population parameter: $\bar{x}$ for μ, $\hat{p}$ for p, etc.

An estimator T is *sufficient* if the data part of the likelihood function can be written in terms of T.

An estimator is *consistent* if it concentrates on the parameter it is estimating. By concentrating, visualize the density of the sampling distribution having more and more of its total area mostly above the parameter. We saw that if the variance of the estimator goes to 0 then the estimator is consistent. In general the definition involved showing

$$P(|T - \theta| > \varepsilon) \to 0.$$

An estimator is unbiased if $E(T) = \theta$, otherwise biased.

Chapter 9 In chapter 9 we learn about statistical inference through confidence intervals. The key idea is to use probability to make precise statements about how a statistic estimates a parameter.

We discussed the m.s.e. to measure "closeness"

$$\text{m.s.e.} = E((T-\theta)^2) = \text{VAR}(T) + \text{bias}_\theta(T)^2.$$

For unbiased estimators this is just the variance, otherwise we add the bias squared. Related to this are the r.m.s.e. and then the more widely used s.e. or *standard error*. This latter, is the standard deviation with any population parameters replaced by estimates from the data.

The notion of confidence intervals results in a statement like:

A $(1-\alpha)100\%$ confidence interval for μ based on $\bar{x}$ is

$$\bar{x} \pm t^* SE.$$

However, there is a bit implied by this statement.

First it begins with understanding the sampling distribution of the following statistic

$$T = \frac{\bar{x}-\mu}{SE(\bar{x})}.$$

Under the assumption of a large sample, or a normal population this has a known sampling distribution allowing use to solve

$$P(-t^* \leq T \leq t^*) = 1-\alpha$$

for (t^*, α). Once this was done, we simply rewrite the above as an interval for μ. This interval is random, and if the assumptions to "know" the sampling distribution are met, then it has a $(1-\alpha)$ chance of containing μ.

We discussed a similar CI for p based on $\widehat{p}$ using the statistc

$$\frac{\hat{p}-p}{SE(\hat{p})}$$

We discussed a pivotal quantity, allowing use to find CIs in more general cases.

Finally we discussed methods to find estimators for parameters: the method of moments, the maximum likelihood method, and a means to compare the performance of estimators called the efficiency.

1. For the data set

   ```
   2,3,5,7,11
   ```

   Find the sample mean, the sample median, the sample variance, and the IQR.

2. For the paired data set

   ```
   x: 1 2 3 4
   ----------
   y: 1 3 3 5
   ```

   Find the correlation coefficient

3. From this stem and leaf output, find the sample median and range

   ```
   > stem(x)


     The decimal point is 1 digit(s) to the right of the |

      6 | 127
      8 | 0024568345566689
     10 | 06049
     12 |
     14 | 7
   ```

4. For a distribution with p.d.f.

   $$f(x|\theta) = cx^2, \quad 0 \leq x \leq \theta$$

   (a) Find c

   (b) Find μ

   (c) Find the likelihood function $L(\theta)$.

   (d) Find a sufficient statistic for θ

   (e) Find a maximum likelihood estimator for θ

(f) Is this estimator sufficient? consistent?

(g) Describe the asymptotic sampling distribution of the estimator

5. A survey of 1,000 college students finds 643 think Survivor has “jumped the shark.” If the sample is a random sample, find a 90% CI for the population proportion who think this. As well answer: what is the population? How does p summarize the population? Why did I slip in the phrase “random sample”? What does this phrase mean?

6. If Gossett had found these heights for 4 of his prisoners what would he have calculated for a 95% CI for μ?

```
68, 69, 69, 72
```

7. Suppose for your random sample you knew that

$$T = \frac{(n-1)s^2}{\sigma^2}$$

had a known sampling distribution such that you could solve (if you had to)

$$P(a \leq T \leq b) = 1 - \alpha.$$

($T > 0$, so instead of $-b$, I used $1/b$.)

Write down a $(1-\alpha)100\%$ CI for s.

8. For a uniform(0,1) population, the median, M, of a sample of size 3 has density given by

$$f(x) = cx(1-x)$$

(a) Find c

(b) Verify that $E(M) = 1/2$

(c) Show that the variance of M is less than $1/12$

(d) Find a so that $P(a < M < 1-a) = 0.90$.

9. Two competing surveys of presidential approval had these results

Who	Approve	n
AP	39%	1000
WSJ	37%	1100

Find a 95% CI for the difference of the population means.

10. Are students getting taller? Suppose two random samples of CSI students were conducted in different years. Among the questions asked was the height of the respondent. The following numbers were recorded

year	xbar	s	n
1995	69	3	16
2005	70	2.5	20

Assume both populations are assumed normal. Find a (conservative) 95% CI for the difference in population means. By conservative, I mean use the minimum of 15 or 19 degrees of freedom.

11. Some early statistical questions (in the development of the field) involved measuring properties of the earth. One question was whether the earth was perfectly spherical or not. Suppose the question boils down to seeing if 0 is in the 95% confidence for the population mean that produced this random sample

```
.5 .7 .3 -.1 .6
```

What do we conclude? (Actually the question was settled by teams sent off to Ecuador and above the artic circle to survey a fixed distance.)