Chapter 6

Estimation

> A grade is an inadequate report of an inaccurate judgment by a biased and variable judge of the extent to which a student has attained an undefined level of master of an unknown proportion of an indefinite amount of material. (George W. Tauxe)

One of the objectives of statistical inference is estimation of population characteristics on the basis of limited information contained in a sample.

Figure: Population, Sample, Statistic, Parameter.

Usually, a sample is taken and a **statistic** (a function of observations) is calculated. The value of the statistic is a point estimator of the parameter.

For instance, responses in political pools observed as sample proportions are used to estimate population proportion of voters in favor of different candidates. To be added.

The following phenomenon exemplifies why we need a formal, mathematical framework for statistical estimation.

The Anchoring Phenomenon. The skills, background information, and biases of a data analyst may affect the conclusions drawn from a body of data. The anchoring phenomenon, studied by Kahneman et al. (1982),[1] exemplifies the effect of an individual's preliminary judgments. Subjects were asked to estimate various percentages (e.g., the percentages of African countries in the United Nations). For each quantity, a number between 0 and 100 was determined by spinning a wheel in view of the subject. The subjects were instructed first to indicate whether that number was higher or lower than the actual value of the quantity and then to estimate that actual value by moving up or down from the given number. The starting number had a marked effect on estimated. For example, the median estimates of the percentage of African countries in the United Nations were 25 and 45 for groups of subjects that received 10 and 65, respectively, as starting points.

Probabilistic reasoning is used to give a sensible procedure for estimating an unknown parameter p of Bernoulli(p) distribution in the following example.

[1] Kahneman, D., Slovic, P., and Tversky, A. (Eds.) (1982). *Judgment under Uncertainty: Heuristics and Biases.* Cambridge: Cambridge University Press.

Estimating the probability of heads: One hundred people are asked to toss four times each a coin giving an unknown probability p of heads. The exact count of how many times they obtained heads was not kept. However, six of them reported getting no tails in the four tosses. Give an estimate of the value p.

Solution: If p is the probability of heads, then event HHHH has probability of p^4. The relative frequency of this event is found to be $\frac{6}{100}$. Solving the equation $p^4 = 0.06$ we get a reasonable estimator of p, $\hat{p} = 0.4949$.

6.1 Point Estimators and Their Sample Distributions

We start with an estimation problem in the following military episode from Thucydides [Warner (1954), p. 172].

> [The problem for the Athenians was] ...to force their way over the enemies' surrounding wall .... Their methods were as follows: they constructed ladders to reach the top of the walls from the number of layers of bricks at a point which was facing in their direction and had not been plastered. The layers were counted by a lot of people at the same time, and though some were likely to get the figure wrong, the majority would get it right, especially as the counted the layers frequently and were not so far away from the wall that they could not see it well enough for their purpose. Thus, guessing what the thickness of a single brick was, they calculated how long their ladders would have to be ....[2]

6.1.1 Estimation of a mean

Suppose that the population is finite ($y_1, \ldots, y_N$, size=N). The population mean is $\mu = \frac{1}{N}\Sigma_{i=1}^{N} y_i$. Given a sample $X_1, \ldots, X_n$ of size n we apply the same arithmetic operations to obtain the sample mean:

$$\bar{X} = \frac{1}{n}\Sigma_{i=1}^{n} X_i.$$

We estimate μ by $\hat{\mu} = \bar{X}$. The estimator $\bar{X}$ is "optimal" in estimating a mean in many different models and for many different definitions of "optimality".

The estimator $\bar{X}$ varies from sample to sample. More precisely, $\bar{X}$ is a random variable with a fixed distribution depending on the distribution of addends, X_is.

> The following is true for **any** distribution in the population as long as $EX_i = \mu$ and $Var(X_i) = \sigma^2$ are finite.
>
> $$E\bar{X} = \mu, \qquad Var(\bar{X}) = \frac{\sigma^2}{n}.$$

The above two equations are very important in estimation theory. First says that $\bar{X}$ is an unbiased estimator of μ. The second assures that larger samples will make variability of estimator the smaller.

If $X_i \sim N(\mu, \sigma^2)$ then we know the distribution of $\bar{X}$ exactly. From the previous chapter we know $\bar{X} \sim N(\mu, \sigma^2/n)$.

Example 1: It is believed that the IQ score for a particular age-group of children is a normal random variable with an unknown mean μ and the variance 100. Find an estimate of μ from the sample: 100, 89, 123, 109, 98, 94, 143, 96, 97, 108, 113.

[106.4]

Example 2: A gaging device produces a random error whose standard deviation is 1%. How many measurements should be averaged in order to reduce the standard deviation of the error to less than 0.05%. [Ans. ≥ 400]

6.1.2 Point Estimation of Variance

We will get an intuition starting again with a finite population, $y_1, \ldots, y_N$. The population variance is $\sigma^2 = \frac{1}{N}\Sigma_{i=1}^{N}(y_i - \mu)^2$, where $\mu = \frac{1}{N}\Sigma_{i=1}^{N} y_i$ is the population mean.

[2]Warner, Rex, trans. (1954). *Thucydides, History of the Peloponnesian War.* Penguin Books, Baltimore, Maryland.

An estimator of σ^2 is

$$\hat{\sigma}^2 = \frac{1}{n}\Sigma_{i=1}^n (X_i - \mu)^2,$$

if μ is known, or

$$\hat{\sigma}^2 = s^2 = \frac{1}{n-1}\Sigma_{i=1}^n (X_i - \bar{X})^2,$$

if μ is not known.

In the expression for s^2 we divide by $n-1$ instead of n since that assure unbiasedness of s^2. More precisely, $Es^2 = \sigma^2$.

When the population is normal $N(\mu, \sigma^2)$ then

$$(n-1)s^2/\sigma^2 \sim \chi^2_{n-1}.$$

i.e. the random quantity $(n-1)s^2/\sigma^2$ has χ^2 distribution with $n-1$ degrees of freedom.

The formula below is usually called the **calculational formula of** s^2. The savings in calculation time is substantial when n is large.

The calculational formula for s^2 :

$$s^2 = \frac{1}{n-1}(\Sigma_{i=1}^n (X_i^2) - n(\bar{X})^2).$$

6.1.3 Finite population correction factor

In finite populations, when the ratio $\frac{n}{N}$ is not negligible, the given estimator for $\sigma_{\bar{X}}$ is true when we sample **with replacement.** If the sampling is done **without replacement** then the sampling distribution of the variance of $\bar{X}$, is

$$\sigma^2_{\bar{X}} = \frac{N-n}{N-1} \cdot \frac{\sigma^2}{n}.$$

The factor $\frac{N-n}{N-1}$ is the **finite population correction factor.** The above formula is used if $20n > N$.

Reaction Time. A sample of 20 students is selected from the population of 400 students and given a test to determine their reaction time to respond to a given stimulus. If the mean reaction time is determined to be $\bar{X} = 0.9$ sec. and the standard deviation is $s = 0.12$ sec., find an estimator of $\sigma^2_{\bar{X}}$.

Solution: To be added.

6.1.4 Estimation of Median

As an estimator of the population median it is natural to take the median of a random sample.

For symmetric distributions the mean and the median coincide. One can use bothe, the mean and the median to estimate the center of the population. If the population is normal, the mean is more efficient[3] than the median. To achieve the same precision as that of the mean, the median needs 57% more observations.

6.2 Interval Estimation

A somewhat disturbing property of point estimates is the following. If the estimator has continuous distribution (like in the case of $\bar{X}$, when X_is are normal) then the estimator will never take the exact value of the parameter. Neyman and Pearson introduced the notion of **confidence interval** in 1950. The interval estimators are computed from the sample and cover "true" but unknown value of the parameter with a pre-assigned confidence probability. The confidence probability is usually selected to be 0.95 (or 95 %), 0.99, and 0.90.

[3]The estimator $\hat{\theta}_1$ is is more efficient than the estimator $\hat{\theta}_2$, if it achieves the same variance (precision) with smaller sample size.

A general way of obtaining many interval estimators is the following. One starts with a point estimator as a center of the interval. The width is determined by the variability of the estimator and some theoretical distribution quantile that incorporates the wanted confidence. For example:

		$(z_{\alpha/2})$		(Standard error of the point estimate)
(Point estimate)	$\pm$	or	$\cdot$	or
		$(t_{\nu,\alpha/2})$		(Sample standard error of the point estimate)

6.2.1 Confidence intervals for a normal mean

Let $X_1, \ldots, X_n$ be a sample from $\mathcal{N}(\mu, \sigma^2)$ distribution where the parameter μ is to be estimated and σ^2 is known.

Starting from the identity

$$P(-z_{\alpha/2} \le Z \le z_{\alpha/2}) = 1 - \alpha,$$

and the fact that $\bar{X}$ has $\mathcal{N}(\mu, \frac{\sigma^2}{n})$ distribution, we can write:

$$P(-z_{\alpha/2}\frac{\sigma}{\sqrt{n}} + \mu \le \bar{X} \le z_{\alpha/2}\frac{\sigma}{\sqrt{n}} + \mu) = 1 - \alpha,$$

Simple algebra gives:

$$\bar{X} - z_{\alpha/2}\frac{\sigma}{\sqrt{n}} \le \mu \le \bar{X} + z_{\alpha/2}\frac{\sigma}{\sqrt{n}}$$

If σ^2 is not known, then the confidence interval with the sample standard deviation s in place of σ can be used. The z-cut points are valid for large n, for small n ($n < 30$) we use t_{n-1} cut-points. Thus for σ^2 unknown:

$$\bar{X} - t_{n-1,1-\alpha/2}\frac{s}{\sqrt{n}} \le \mu \le \bar{X} + t_{n-1,1-\alpha/2}\frac{s}{\sqrt{n}}$$

We summarize above

> The $(1 - \alpha)$ 100% confidence interval for the unknown normal mean μ when variance σ^2 is known on basis of sample of size n is
>
> $$[\bar{X} - z_{\alpha/2}\frac{\sigma}{\sqrt{n}}, \bar{X} + z_{\alpha/2}\frac{\sigma}{\sqrt{n}}].$$
>
> If the variance is not known, then the $(1 - \alpha)$ 100% confidence interval is
>
> $$[\bar{X} - t_{n-1,1-\alpha/2}\frac{s}{\sqrt{n}}, \bar{X} + t_{n-1,1-\alpha/2}\frac{s}{\sqrt{n}}].$$

Useful Table

Confidence $(1 - \alpha)$ 100 %	$\Phi(z) = 1 - \alpha/2$	$z_{\alpha/2}$
0.80	0.90	1.28
0.90	0.95	1.64
0.975	0.9875	2.24
0.98	0.99	2.33
0.99	0.995	2.58
0.999	0.9995	3.30

Vehicles on highway. A Florida State Highway Department inspector is interested in knowing the mean weight of commercial vehicles traveling on the roads in his state. He takes a sample of 36 randomly selected trucks passing through state weigh stations and finds the mean gross weight $\bar{X} = 15.8$ tons. The standard deviation is known and equals $\sigma = 6.0$ tons. Assume that the weights are normally distributed.

Construct a 99% confidence interval for the mean gross weight of commercial vehicles traveling the highways of this state.

Next we give a simple Splus program that calculates confidence intervals.

```
> zint_function(x, sig, n = length(x), alpha = 0.05, r = 3)
 {
      center <- mean(x)
      abb <- (qnorm(1 - alpha/2) * sig)/sqrt(n)
      lower <- round(center - abb, r)
      upper <- round(center + abb, r)
      cat("[", lower, ",", upper, "]\n")
 }
```

The function `zint` (short for z-interval), gives $(1-\alpha)$ 100% confidence interval for the following input.

`x` - (1) a sequence of observations or (2) a single number $\bar{X}$. That means that the function `zint` can take two types of data to process: already summarized by $\bar{X}$ and row (as a vector, say `c(2,3,2,3,4,3,4,3,3,4)`.

`sig` - known σ. This argument has no default and it has to be given on input.

`n` - sample size (if omitted it will be taken as length of `x`, by default).

`alpha` - default 0.05 corresponding to confidence level 95%. The relation between `alpha` and confidence level is the following: confidence level = 1 - `alpha`.

`r` - number of decimal places of the result. Default is rounding to 3 decimal places. Rounding is done by the function `round`.

The Splus program works the same way as if we used a calculator. The `center` is $\bar{X}$ and `abb` is $z_{\alpha/2}\frac{\sigma}{\sqrt{n}}$.

For the above problem, the Splus session is:

```
> zint(x=15.8, sig=6, n=36, alpha=0.01)
[ 13.224 , 18.376 ]
```

Of course, it was enough to say `zint(15.8,6,36,0.01)`, but because of the danger of mixing-up the arguments we recommend naming them.

When the variance is not known a confidence interval can be found by the following Splus function:

```
> tint_function(x, sig = sqrt(var(x)), n = length(x), alpha = 0.05, r = 3)
 {
      center <- mean(x)
      abb <- (qt(1 - alpha/2, n - 1) * sig)/sqrt(n)
      lower <- round(center - abb, r)
      upper <- round(center + abb, r)
      cat("[", lower, ",", upper, "]\n")
 }
```

Compare the functions `zint` and `tint`. The function `tint` has t cut-points and an option to calculate the sample standard deviation. When the entry for `x` is a vector `sig` will be calculated from `x` by default.

Naphtha drums. A company purchases large quantities of naphtha in 50-gallon drums. Because the purchases are ongoing, small shortages in the drums can represent a sizable loss to the company. The weights of the drums vary slightly from drum to drum, so the weight of the naphtha is determined by removing it from the drums and measuring it. Suppose the company samples the content of 15 drums, measures the naphtha in each, and calculates $\bar{x} = 49.70$ and $s = 0.32$ gallon.

1. Find a 95% confidence interval for the mean number of gallons of naphtha per drum.
2. What assumptions are necessary to assure validity of the confidence interval?

```
> tint(x=49.7, s=0.32, n=15)
[ 49.523 , 49.877 ]
```

Apparel and Shoe Trade. An apparel and shoe trade association has 8,900 member firms. As a part of a study of the impact of new minimum-wage legislation on association members, a random sample of 400 member firms was selected to estimate μ, the mean number of hourly-paid employees in member firms. A computer analysis of the sample results showed: $\bar{X} = 8.31$, $s = 4.16$, where X denotes the number of hourly paid employees in a firm.

(a) Construct 94% confidence interval for μ.

```
> tint(x=8.31,s=4.16, n=400, alpha=0.06)
[ 7.918 , 8.702 ]
```

(b) Explain why the confidence level of your interval is only approximately 94%.

(c) If you had constructed a 99% confidence interval, would it have been wider or narrower then one in (a)? Wider.

```
> tint(x=8.31,s=4.16, n=400, alpha=0.01)
[ 7.772 , 8.848 ]
```

Depression. In a study of 180 female psychiatric patients suffering from recurrent depression, the mean age at their first depressive episode was 26 years, with a standard deviation of 11.2 years (Frank, Carpenter, and Kupfer, 1988). Assuming that these 180 cases can be viewed as a random sample, find and interpret a 95% confidence interval for the mean age of onset. [24.353, 27.647]

The most recent depressive episodes of the 180 patients lasted an average of 22.1 weeks, with a standard deviation of 17.4 weeks. Find and interpret a 95% confidence interval for the mean length of the most recent episode. [19.541, 24.659]

6.2.2 Confidence Intervals for a normal variance

In the previous section we saw that $\frac{(n-1)s^2}{\sigma^2}$ has χ^2 distribution with $n-1$ degrees of freedom. Thus, from the definition of χ^2 quantiles

$$1 - \alpha = P(\chi^2_{n-1,\alpha/2} \le \chi^2_{n-1} \le \chi^2_{n-1,1-\alpha/2}).$$

From the distribution of $\frac{(n-1)s^2}{\sigma^2}$,

$$1 - \alpha = P(\chi^2_{n-1,\alpha/2} \le \frac{(n-1)s^2}{\sigma^2} \le \chi^2_{n-1,1-\alpha/2}).$$

Simple algebra with the above inequalities (take reciprocal of all three parts, being careful about inequalities, and multiply everything by $(n-1)s^2$)

$$\frac{(n-1)s^2}{\chi^2_{n-1,1-\alpha/2}} \le \sigma^2 \le \frac{(n-1)s^2}{\chi^2_{n-1,\alpha/2}}.$$

The $(1-\alpha)$ 100% confidence interval for an unknown normal variance is

$$[\frac{(n-1)s^2}{\chi^2_{n-1,1-\alpha/2}}, \frac{(n-1)s^2}{\chi^2_{n-1,\alpha/2}}].$$

We did not say explicitly if we knew the mean or not. As given, the formula does not assume the mean is known. If we knew the mean μ then the χ^2 cut-points, will have one degree of freedom more (n instead of $n-1$) making the confidence a bit tighter. **Example:** To be added.

6.2.3 Confidence Intervals for a median

The confidence interval for the median replaces confidence intervals for the mean if normality is not satisfied. Let $X_{(1)}, X_{(2)}, \ldots, X_{(n)}$ be the order statistics of the sample $X_1, \ldots, X_n$. Then a $(1-\alpha)$ 100% confidence interval for the median Me is

$$X_{(h)} \le Me \le X_{(n-h+1)}.$$

The value of h is usually given by tables. For n large ($n > 50$) good approximation for h is

$$h = \frac{n - z_{\alpha/2}\sqrt{n} - 1}{2}.$$

For example, if $n = 300$, the 95% confidence interval for the median is $[X_{(133)}, X_{(168)}]$.

6.2.4 Confidence Intervals for Proportions

Let p be the population proportion and $\hat{p}$ the observed sample proportion. Assume that the smaller of numbers $\frac{np}{q}$, $\frac{nq}{p}$is larger than 9 (rule of thumb). Then, CI of confidence level $(1-\alpha)100\%$ for the unknown p is

$$[p - z_{\alpha/2}\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}, p + z_{\alpha/2}\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}]$$

Elections. Project yourself back in time to six recent U.S. presidential elections. In parentheses we give the results of the Gallup pre-election poll of 1500 voters. (As we mentioned already, each sample has about the same accuracy as a simple random sample. And we continue to ignore third parties.)

Year	Democrat	Republican
1960	Kennedy (51%)	Nixon (49%)
1964	Johnson (64%)	Goldwater (36%)
1968	Humphrey (50%)	Nixon (50%)
1972	McGovern (38%)	Nixon (62%)
1976	Carter (51%)	Ford (49%)
1980	Carter (48%)	Reagan (52%)

(a) In each case, construct a 95% confidence interval for the proportion of Democratic supporters in the whole population.

(b) Mark each case where the interval is wrong - that is, fails to include the true proportion p given in the following list of actual voting results:

1960	Kennedy	50.1%
1964	Johnson	61.3%
1968	Humphrey	49.7%
1972	McGovern	38.2%
1976	Carter	51.1%
1980	Carter	44.7%

Simple Splus function for confidence interval on proportions for large sample sizes is:

```
> pint_function(p, n, alpha = 0.05, r = 3)
{
        abb <- qnorm(1 - alpha/2) * sqrt((p * (1 - p))/n)
        lower <- round(p - abb, r)
        upper <- round(p + abb, r)
        cat("[", lower, ",", upper, "]\n")
}
```

For McGovern's proportion (38%):

```
> pint(0.38, 1500)
[ 0.355 , 0.405 ]
```

Alcoholism. In a study to determine whether alcoholism has (in part) a genetic basis, genetic markers were observed for a group of 50 Caucasian alcoholics. For 5 alcoholics the antigen (marker) B15 was present. Estimate with 99% confidence interval the proportion of Caucasian alcoholics having this antigen. [0.017, 0.183]

If p or q is close to 0, then more precise confidence interval for unknown proportion p is[a]

$$\left[\frac{(x-0.5)+\frac{z^2_{\alpha/2}}{2}-z_{\alpha/2}\sqrt{(x-0.5)-\frac{(x-0.5)^2}{n}+\frac{z^2}{4}}}{n+z^2_{\alpha/2}},\ \frac{(x+0.5)+\frac{z^2_{\alpha/2}}{2}+z_{\alpha/2}\sqrt{(x+0.5)-\frac{(x+0.5)^2}{n}+\frac{z^2}{4}}}{n+z^2_{\alpha/2}}\right]$$

[a] Blyth, C. and Still, H. (1983). Binomial confidence intervals, *JASA* **78**, *108-116.*

Aids. A random sample of 500 individuals was taken from a large city to estimate the incidence of the disease AIDS. The disease AIDS was detected among 6 of the sample members. Find 95% confidence interval for the true proportion. [0.002, 0.022]

Solution:

6.3 Designing the sample size

In all previous examples it is assumed that we have data in hand. Thus, we look at the data after sampling procedure has been completed. It is often the case that we have a control what sample size to adopt before the sampling. How big should the sample be? Too small a sample may influence the validity of our statistical conclusions. On the other hand an unnecessarily large sample wastes money, time and resources.

Explain the sample size for point estimators. Example.

(i) Equation for sample size for estimating the mean: σ^2 known.

$$n=\frac{4z^2_{\alpha/2}\sigma^2}{w^2}$$

where w is the width of the interval (w=Upper bound–Lower bound)
(ii) Equation for sample size for estimating the proportion

$$n=\frac{4z^2_{\alpha/2}\hat{p}(1-\hat{p})}{w^2}$$

where $\hat{p}$ is the sample proportion. In absence of data $\hat{p}$ is our best guess. In absence of any information the most conservative choice is $\hat{p}=0.5$.

The Splus function calculating a sample size for the normal mean problem is:

```
> size.mean_function(w, sig, alpha = 0.05)
  {
         return(ceiling((4 * qnorm(1 - alpha/2)^2 * sig^2)/w^2))
  }
```

The function `ceiling` takes a larger integer of its argument. For example `ceiling(12.459)=13.` Using `size.mean` as a template write and test the function `size.prop` which designs a sample size for a proportion. Put `p` as a input argument with a default 1/2.

Credit manager. The credit manager of Durham's Circuit City Store would like to know what proportion of the charge customers take advantage in the "0 interest for 6 months" plan each year. The manager would like to estimate this figure within 10% at a 90% confidence level, but she has no idea right now about what this proportion might be.

Solution. $w = 2 \cdot 0.10 = 0.2$ and $z_{\alpha/2} = z_{0.05} = 1.645$. Since she has no preliminary knowledge about $\hat{p}$ the most conservative choice is $\hat{p} = 0.5$. Then

$$n \geq \frac{4z^2_{\alpha/2}\hat{p}(1-\hat{p})}{w^2} = \frac{4 \cdot 1.645^2 \cdot 0.5(1-0.5)}{0.2^2} = 67.7.$$

Rounded to the nearest larger integer the sample size is 68. The number seems a reasonable sample size to take but some work can be saved if the manager had some prior knowledge about $\hat{p}$. For instance if the manager believes that $\hat{p} = 0.3$ then the sample size is $n = 56.83$.

You, the company statistician. You are designing a 99% confidence interval for the population mean. From earlier experience you know the estimate of the population standard deviation, $s = 0.25$. Your boss prefers short intervals, at most 0.01 in length, but he is concerned about the cost of the sampling.

Find the minimal sample size that will yield the desired interval.

```
> size.mean(w=0.01, s=0.25, alpha=0.01) #or size.mean(0.01,0.25,0.01)
[1] 16588
```

EPA. The EPA standard on the amount of suspended solids discharged into rivers and streams is a maximum of 60 milligrams per liter daily, with a maximum monthly average of 30 milligrams per liter. Suppose you want to test a randomly selected sample of n water specimens and estimate the mean daily rate of pollution produced by the mining operation. If you want 95% confidence interval estimate of width 2 milligrams, how many specimens you need to sample? Assume prior knowledge indicates that pollution readings in water samples taken during a day are approximately normally distributed with a standard deviation equal to 5 milligrams.

```
> size.mean(w=2, sig=5, alpha=0.05) #or simply size.mean(2,5)
[1] 97
```

Homicides. In a random sample of 122 U.S. homicides in 1984, 44 were committed using a handgun.

(a) Find a 99% confidence interval for the proportion of all homicides during this year that were committed using a handgun.

(b) Interpret this interval in formal statistical terms, with reference to the population proportion π.

(c) Interpret the interval less formally, as you might if you were trying to explain it to an audience of non-statisticians.

6.3.1 Jackknife

Jackknife in general. The jackknife is an all-purpose technique for robust confidence intervals; its great virtue lies in its almost universal applicability. Once the estimate has been calculated from the sample, the jackknife provides the confidence interval around it, in 4 steps:

1. Denote the estimate (based on all the data) by X_{all}.

2. Calculate the estimated "sample value" X_{-1} based on all the data except the 1st observation. Similarly calculate X_{-2}, X_{-3}, and so on.

3. Calculate the "pseudo-value" $X_{(1)}$, which is just the X_{-1} spread further from X_{all}:

$$X_{(1)} \equiv X_{all} + (n-1)(X_{all} - X_{-1})$$

Similarly calculate $X_{(2)}, X_{(3)}$, and so on. They will be spread far enough apart to act as if they were independent. Thus they constitute a "pseudo-sample" that acts like a random sample.

3. For the pseudo-sample, calculate the mean $\bar{X}$ and standard deviation s, and substitute them into the 95% confidence interval:

$$\text{Population parameter} = \bar{X} \pm t_{.025} \frac{s}{\sqrt{n}}$$

Along with the confidence interval, jackknifing provides a bonus: If the original estimate is slightly biased (but asymptotically unbiased), jackknifing will often eliminate the bias (Wonnacott, 1984 IM). [4]

Two unbiased estimators $\bar{X}$ and s^2 are sitting in a bar and drinking.
s^2: Hey, Unbiasy. So you are getting married next week.
$\bar{X}$: Yes, I decided to lose one degree of freedom.

6.4 Exercises

Sheriff again. Suppose the volume of noise in dB produced by Sheriff of Nottingham when his arrow misses the target has a standard deviation of 70 dB. Robin personally makes 49 measurements and finds a sample mean of 500. Determine 90% confidence limits for the mean noise level. [483.551, 516.449]

Speed of sound. Thirty students in a physics laboratory make determinations of the speed of sound. The average of their determinations is 11,000 ft/sec, and the sample standard deviation is 200 ft/sec. Find a 50% confidence interval for the "true" speed of sound in the laboratory at the time. [10975.371, 11024.629]

3. **Little John's Report:** Of 50 arrows shot at a target, 35 were hits. Determine 80% confidence limits for the probability of a hit. [0.617, 0.783]

Steaks. How large a sample would you require to determine a 95% confidence interval of length not more than 0.05 for the proportion of people in a given large group who prefer steak well done to rare?

Cereal. The weights of full boxes of Kellogg's Nutri-Grain Cereal are normally distributed with a standard deviation $\sigma = 0.18$ ounce. If a sample of 12 randomly selected boxes produced a mean weight of 17.28 ounces, find

(a) The 90% and 95% confidence intervals for the true weight of a box of this cereal. [17.187, 17.373], [17.166, 17.394].

(b) Discuss the effect of different confidence levels on the width of the interval in (a).

LC50 The Environmental Protection Agency has collected data on the LC50 (concentration killing 50% of the test animals in a specified time interval) measurements for certain chemicals likely to be found in freshwater rivers and lakes. For a certain species of fish, the LC50 measurements for DDT in 4 experiments yielded the following: 13, 9, 21, 19. (Measurements are in parts per million). Assuming such LC50 measurements to be normally distributed, estimate the true mean of LC50 for DDT with confidence level 98%.

[Solution: $\bar{X} = 16,\ s^2 = 18$]

```
> tint(c(13, 12, 21, 18), alpha=0.02)
[ 6.368 , 25.632 ]
```

LC50 continued. The variance of LC50 measurements is important because it may reflect ability (or inability) to reproduce similar results in identical experiments. Find 95% confidence interval for σ^2, the true variance of LC50 measurements for DDT. [9.249, 22.751]

[4] Wonnacott, T. H., and R. J. Wonnacott (1984 IM), Instructor's Manual for Introductory Statistics for Business and Economics. New York: Wiley.

Cholesterol Level. Suppose you are designing a cholesterol study experiment and would like to estimate the mean cholesterol level of all Purdue students. You will take a random sample of n students and measure their cholesterol level. Previous studies have shown that the standard deviation is about 25 and you will use this value in planning your study.

If you want a 99% confidence interval with an approximate length of 12, how many students should you include in your sample?

Toxins. An investigation on toxins produced by molds that infect corn crops was performed. A biochemist prepared extrats of the mold culture with organic solvents and then measures the amount of toxic substance per gram of solution. From 6 preparations of the mold culture the following measurements of the toxic substance (in milligrams) are obtained: 3 2 5 3 2 6.

(a) Calculate the mean $\bar{X}$ and the standard deviation s from the data.

(b) Compute a 99% confidence interval for the mean weight of toxic substance per gram of mold culture. State the assumption you make about the population.

Limnologist. A limnologist wishes to estimate the mean phosphate content per unit volume of a lake water. It is known from studies in previous years that the standard deviation has fairly stable value $\sigma = 4$. How many water samples must the limnologist analyze to be 90% certain that the error of estimation does not exceed 0.8 (i.e. the corresponding confidence interval has the half-length of 0.8).

$n = (\frac{1.645*4}{.8})^2 = 67.65.$

Anxiety Measure. Morelli et al. (1982) report on a study in which extrovert and introvert groups were measured on anxiety scale[5].

Assume that the data were as follows:

Extroverts:	5.5	7.0	8.8	5.3	6.9	8.4	9.1	7.0	5.8	6.0	7.3	9.8
	5.2	6.7	8.2	7.9	6.1	7.4	6.6	4.7	7.4	7.6	6.4	6.6
Introverts:	7.1	7.6	9.0	8.0	5.4	8.7	9.3	8.2	8.5	6.7	8.9	9.6
	8.2	9.2	9.7	8.7	9.5							

(a) Find the 5-number-summary for both data sets. *You may find this info useful:*
Extroverts (ordered): 4.7 5.2 5.3 5.5 5.8 6.0 6.1 6.4 6.6 6.6 6.7 6.9 7.0 7.3 7.4 7.4 7.6 7.8 7.9 8.2 8.4 8.8 9.1 9.8
Introverts (ordered): 5.4 6.7 7.1 7.6 8.0 8.2 8.2 8.5 8.7 8.7 8.9 9.0 9.2 9.3 9.5 9.6 9.7

(b) Draw a stem-and-leaf display of each groups anxiety score.

(c) If you want to give an interval estimator of the population mean μ for Extroverts anxiety score, what sample size do you need to achieve 99% confidence with an interval of total length 0.1. Assume that the scores come from a normal population with known standard deviation $\sigma = 1$?

GREATP. The following is a sample of 10 scores obtained on the Graduate Record Examination Advanced Test in Psychology (GREATP) by graduates with BS in Psychology from a state university

520	470	410	480	440
500	460	510	390	500

(a) Find the mean and standard deviation for the sample.

(b) Find 99 % confidence interval for the population mean score.

(c) If the population standard deviation were known $\sigma = 40$, say, what sample size would you need to obtain a 99 % confidence interval of the length 2.

Probability of heads. A coin is tossed 40 times with the following result:

H	H	T	T	H	T	T	T	H	H
T	T	H	T	H	H	H	T	T	T
T	H	T	T	H	H	H	T	H	H
H	T	H	T	T	T	H	H	T	H

[5]Morelli, G., Andrews, L., and Morelli, R. (1982). The relation involving personality variables, problem relevance, rationality, and anxiousness among college men, *Cognitive Therapy and Research, 6, 57-62.*

(a) Estimate the probability of obtaining a head.
(b) Give 95 % confidence interval for the probability from (a).

Sleep deprivation again. [**15pt**] If you want to give an interval estimator of the population mean for reaction times on the motor task after 36-hour sleep deprivation, what sample size do you need to achieve 95% confidence with an interval of total length 0.1. Assume the reaction times are normally distributed with a standard deviation $\sigma = 0.3$.

Right to Die. A Gallup Poll estimated the support among Americans for "right to die" laws. For the survey, 1528 adults were asked whether they favor voluntary witholding of life-support systems from the terminally ill. The results: 1238 Said yes.

(a) Find a 99% confidence interval for the percentage of all adult Americans who are in favor of "right to die" laws.

(b) If the margin error is to be smaller than 0.01, what sample size is needed to achieve that requirement. Assume $\hat{p} = 0.8$.