Statistical Regression Analysis 09/11/2018

Lecture Notes [Basic]: Basic Statistics with R

Teacher: Prof. Ernest Fokoué **Note**: Reminders of basic statistics

Note: *This part of the course presupposes that you fully remember your basic probability concepts. We therefore delve straight into statistical inference right after probabilistic reminders and a bit of descriptive statistics.*

1 Basic Statistics and Probability Reminders

1.1 Typical Statistical Analysis Procedure

We shall adopt a typical stepwise when using data mining techniques to solve problems. You will probably notice that the steps are really commonsense and not very different typical problem solving.

[Step 1:] Problem Description and Understanding and Statistical Formulation

- ⊞ What is the context of the problem, and what is the overarching question of interest?
- ⊞ How large is the scope of the study? Has anyone tackled it before?
- ⊞ How will the answer to the questions be measured? Are the measures clear and valid?
- ⊞ How can this problem be formulated in statistical terms? Formulate it!

[Step 2:] Data Acquisition and Data Exploration

- ⊞ Is the data easily available? How much data is there? How large is the data set?
- ⊞ Get the data and by all means have a look at it. Was it collected correctly?
- ⊞ Explore the data as thoroughly and carefully as possible

[Step 3:] Informal Analysis

- ⊞ Use Graphical and Numerical summaries to gain insight relative to the question of interest
- ⊞ Use Extensive (sophisticated if need be) Exploratory data analysis techniques
- ⊞ Does your informal analysis seem to provide a particular answer? Is the answer plausible?

[Step 4:] Formal Analysis

- ⊞ What is the appropriate statistical framework
- ⊞ What is the ideal (if any) statistical technique in this framework
- ⊞ What is the best (if any) implementation of the chosen technique

- ⊞ Describe (clearly) the answers that you get.
- ⊞ Are the answers plausible, correct, valid?

[Step 5:] Report

- ⊞ Interpret your Results in light of the key question and (if need be) compare techniques used
- ⊞ Contrast with previous studies, the indicate possible extension and future work
- ⊞ Make Recommendations in plain and intelligible non technical language

2 Simple Example of Statistical Analysis

2.1 Step 1: Problem Description and Understanding, and Statistical Formulation

Question: `Is professional golf lucrative?` Without any further clarification and explanation, this question is at best vague. This question might have come up in a discussion, or just from a remark made anecdoctally by someone upon noticing the wealth of a couple of golfers. Now, to attempt to answer the question, it will help to define a measure of lucrativeness across all fields of endeavors. Let *Is golf lucra*

$$X \equiv \text{random variable whose value is the annual earning of a golfer}$$

Let $\mu = \mu_X = \mathbb{E}[X]$ represent the average annual earning in golf. One million dollars is still - even in these days of trillion dollars deficit - a pretty sizeable amount of money, both psychological and practically. I suggest that we use $\$1M$ average annual earning as our cut-off. My initial inclination is that `professional golf is lucrative`, i.e. I claim that

$$\mu > \$1M$$

Note: *We will see later that the distribution of X is heavily skewed to the right. This clearly means that while golf itself may have a lot of money flowing in it, not all golfers do enjoy the benefits of that flow equitably. The issue here is not about golfers per se, but golf as a profession, and for us statistically, it matters not, as long as our formal techniques do not violate their underlying assumptions.*
To find out if my initial inclination is plausible, we will need to perform a thorough statistical analysis of golfers' earnings, and that requires collecting some data on golfers earnings. Before that, let's note what we have achieved so far:

(a) We have clearly defined a variable who measurements help answer the question of interest

(b) We have chosen a population parameter, namely the population mean (μ) on whose estimation and inference we can provide an statistical sound answer to the question of interest.

(c) We have agreed on (of I have suggested) a criterion for determining the answer to our question, namely ($\mu > \$1M$)

2.2 Data Acquisition and Data Exploration

Ideally, we would like to collect a random sample $X_1, X_2, \cdots, X_n$ of annual earnings from n randomly selected golfers. The dataset we have, namely golf2008.csv contain annual earnings of golfers for the year 2008 only. You may argue that the sample is really not all that random, or even that only one year is taken into consideration. You will be right on both counts. However, in the spirit of cluster sampling, it is not too farfetched for me to consider this dataset and its content a valid sample for this study. As we indicate at the end of this analysis, one could consider: (a) sampling from the present sample (b) sample randomly from many different years of golfers earnings. We first open the data set containing information on golfers along with their earnings.

```
golf <- read.csv('golf2008.csv', header=T)
```

We have `n = 187` golfers in this data set. We also have 23 columns, but one of them really just made up of the names of golfers. We re-arrange this data to conform to the way `R` handles things.

```
golfers <- golf[,1]
golf    <- golf[,-1]
rownames(golf) <- golfers
```

All in all, there does not seem to be any missing values, and the column `Earnings` contains exactly the observations we need for our analysis.

2.3 Informal Analysis of Golf Earnings

The very first step in informal analysis is the use of meaningful graphical summaries. In this case, we construct both the histogram and boxplot of annual earnings. The histogram

graphical summaries

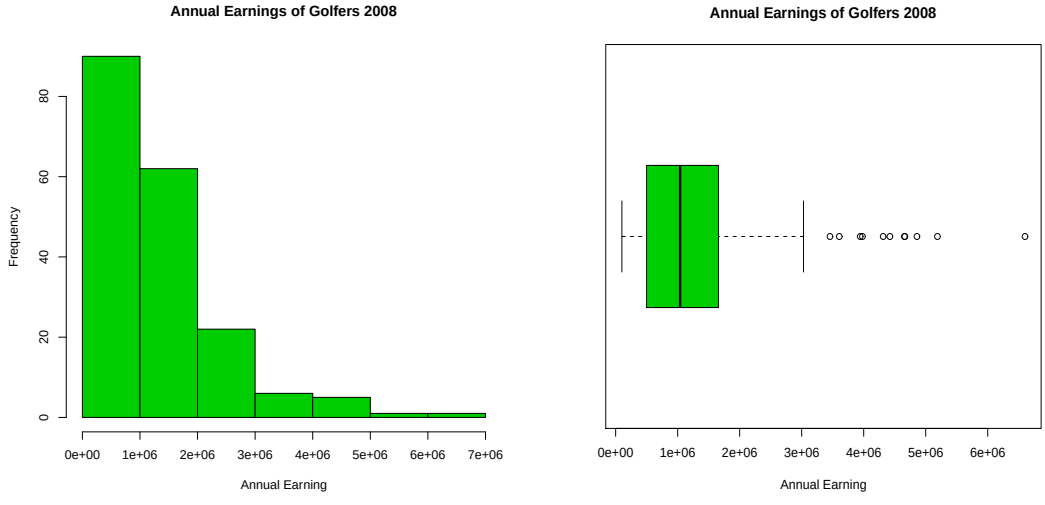


Figure 1: Histogram and box plot showing the distribution of annual earnings of golfers

reveals a distribution that is heavily skewed to the right, and the boxplot adds to that by revealing a fair number of outliers. In fact, one wonders what the shape of the distribution would be without these outliers. Our basic descriptive statistics are given below:

```
Descriptive statistics on golf earnings
~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~

 Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
100600  498500 1040000 1313000 1659000 6601000
```

The skewness to the right is further confirmed by the fact that the sample mean $\bar{\mathbf{x}} = \$1313000$ is substantially larger than the sample median $\tilde{\mathbf{x}} = 1040000$. Now, since we decided ahead of time that the population mean μ was our parameter of interest, we look at this informal stage to the counterpart of μ in the provided sample. In this case, the sample mean $\bar{\mathbf{x}}$ is an estimate of μ (more about this later). Now, since $\bar{\mathbf{x}} = \$1313000$ is far greater than our cut off of \$1000000, we can informally assert than professional golf is lucrative by the standards we set. In the interest of a more thorough analysis however, let's pay attention to the outliers. Let's name them `Too-Rich`.

```
too.rich  <-  which(x>3000000)
```

We made now reveal who this people are, and probably any golf enthusiast will recognize them

```
 Here are our truly rich golfers  2008
~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~

[1] "Anthony Kim"         "Camilo Villegas"    "Jim Furyk"
[4] "Justin Leonard"      "Kenny Perry"        "Mike Weir"
[7] "Padraig Harrington" "Phil Mickelson"     "Robert Allenby"
[10] "Ryuji Imada"         "Sergio Garcia"      "Stewart Cink"
[13] "Vijay Singh"
```

It is indeed shocking to notice that the richest golfer in the world, Mr Tiger Woods is not on the list. My statistics 315 student - Mr Simon Stam - who downloaded this data preferred the 2008 data precisely because Mr Woods was not active that year due to injury and Mr Stam liked that because he feared that Mr Woods huge earnings would be too extreme an outlier. Unfortunately for Mr Stam, Mr Vijay Singh was threatening $\$7M$ that very same year. Should we be concerned that any findings we make here are invalid because the data was somewhat influenced by the experimenter? You be the judge. Let's continue! We said informally earlier that our statistics reveal that professional golf is lucrative. Could it be that these heavily rich outliers made the mean fat. Let's compute the sample mean for ordinary golfers. Interesting, $\bar{\mathbf{x}}_{\texttt{ordinary}} = \$1090000$, which still makes professional golf lucrative by our set standard. Unsurprisingly, the truly rich golfers have an average that $\bar{\mathbf{x}}_{\texttt{too.rich}} = \$4287631$. There seems to be a double world in golfers earnings, with both groups revealing skewness (see plots).
Technically, One could reasonably think of the distribution of annual golfers earnings as a mixture of two exponential distributions or a mixture of two lognormal distributions.

2.4 Formal Analysis of Golf Earnings

First of all, the way we formulated our problem earlier, makes it a typical problem of statistical inference on a single population parameter, namely the population mean μ. In our particular case, we can perform one or both of the following statistical inference activities:

- ⊡ Interval estimation for μ

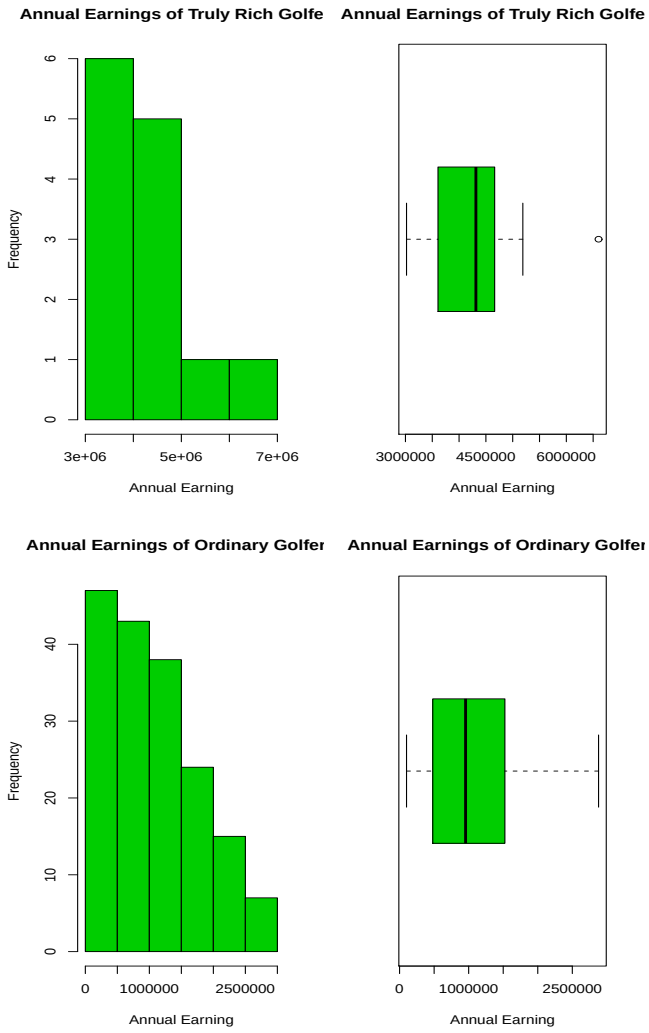


Figure 2: Histogram and box plot showing the distribution of annual earnings of golfers

⊡ Hypothesis testing about μ

Whichever one we choose, we need the sampling distribution of $\bar{X}$, the sample mean. In this particular case, since the sample size $n = 187$ is very large, the central limit theorem applies, namely

Key 1: *S pling dist tion of* $\bar{X}$

$$\bar{X} \sim \mathbf{N}\left(\mu, \frac{\sigma^2}{n}\right)$$

where the sample mean is defined by

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

Therefore, although we do not know the true population mean μ, we know, by virtue the distribution of $\bar{X}$ to be normal, that

$$\Pr\left[\mu - z_{\alpha/2}\left(\frac{\sigma}{\sqrt{n}}\right) \leq \bar{X} \leq \mu + z_{\alpha/2}\left(\frac{\sigma}{\sqrt{n}}\right)\right] = 1 - \alpha$$

which after algebraic cookery can be rewritten as

$$\Pr\left(\left[\bar{X} - z_{\alpha/2}\left(\frac{\mathtt{S}_\mathrm{x}}{\sqrt{n}}\right), \bar{X} + z_{\alpha/2}\left(\frac{\mathtt{S}_\mathrm{x}}{\sqrt{n}}\right)\right] \ni \mu\right) = 1 - \alpha \tag{1}$$

where the sample standard deviation

$$\mathtt{S}_\mathrm{x} = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(X_i - \bar{X})^2}$$

is now used as an estimator of the population standard deviation σ. Equation (1) is basically saying the following: The random interval

$$\left[\bar{X} - z_{\alpha/2}\left(\frac{\mathtt{S}_\mathrm{x}}{\sqrt{n}}\right), \bar{X} + z_{\alpha/2}\left(\frac{\mathtt{S}_\mathrm{x}}{\sqrt{n}}\right)\right] \tag{2}$$

will contain μ with a probability of $1 - \alpha$. When actually collect a sample and compute the number $\bar{\mathrm{x}}$ and $\mathtt{s}_\mathrm{x}$ for that particular sample, the particular interval

$$\left[\bar{\mathrm{x}} - z_{\alpha/2}\left(\frac{\mathtt{s}_\mathrm{x}}{\sqrt{n}}\right), \bar{\mathrm{x}} + z_{\alpha/2}\left(\frac{\mathtt{s}_\mathrm{x}}{\sqrt{n}}\right)\right] \tag{3}$$

is a $100(1-\alpha)\%$ confidence interval for μ. The following statements are equivalent:

⊡ We are $100(1-\alpha)\%$ confident that μ lies between $\bar{\mathrm{x}} - z_{\alpha/2}\left(\frac{\mathtt{s}_\mathrm{x}}{\sqrt{n}}\right)$ and $\bar{\mathrm{x}} + z_{\alpha/2}\left(\frac{\mathtt{s}_\mathrm{x}}{\sqrt{n}}\right)$

⊡ We are $100(1-\alpha)\%$ confident that μ lies in the interval $\left[\bar{\mathrm{x}} - z_{\alpha/2}\left(\frac{\mathtt{s}_\mathrm{x}}{\sqrt{n}}\right), \bar{\mathrm{x}} + z_{\alpha/2}\left(\frac{\mathtt{s}_\mathrm{x}}{\sqrt{n}}\right)\right]$

⊡ The interval $\left[\bar{\mathrm{x}} - z_{\alpha/2}\left(\frac{\mathtt{s}_\mathrm{x}}{\sqrt{n}}\right), \bar{\mathrm{x}} + z_{\alpha/2}\left(\frac{\mathtt{s}_\mathrm{x}}{\sqrt{n}}\right)\right]$ contains μ with $100(1-\alpha)\%$

Note that from Equation (3), we went from capital letters representing random variables, to small letters, representing the values of those random variables. In other words, once the data is collected, the interval we form is a `fixed` interval. As such it either contains μ or it does not.

`[Confidence is not probability]`: When we say that we are $100(1-\alpha)\%$ confident that μ lies in the interval

$$\left[\bar{\mathrm{x}} - z_{\alpha/2}\left(\frac{\mathtt{s}_{\mathrm{x}}}{\sqrt{n}}\right), \bar{\mathrm{x}} + z_{\alpha/2}\left(\frac{\mathtt{s}_{\mathrm{x}}}{\sqrt{n}}\right)\right]$$

we are NOT saying that the ~~probability is $1-\alpha$ that μ is in that interval~~[1]. We are just saying that $100(1-\alpha)\%$ of the intervals constructed in a similar way using samples from the same population will capture μ, and the remaining $100\alpha\%$ intervals will miss μ.

`[Margin of error]`: The most important inferential ingredient in the above CI is the margin of error

$$\mathtt{E}_{100(1-\alpha)\%} = z_{\alpha/2}\left(\frac{\mathtt{s}_{\mathrm{x}}}{\sqrt{n}}\right)$$

where

$$z_{\alpha/2} = \Phi^{-1}\left(1 - \frac{\alpha}{2}\right)$$

is the $100(1-\frac{\alpha}{2})\%$ quantile of the standard normal distribution, i.e. the point on the standard normal distribution range (domain) such

$$\Pr[Z \leq z_{\alpha/2}] = \Phi(z_{\alpha/2}) = 1 - \frac{\alpha}{2}$$

With `R`, this is obtained quite simply by the command

```
z.alpha2 <-qnorm(1-alpha/2)
```

In general, if we let X be a random variable with probability density function $f_X(\mathrm{x})$ and (continuous and monotonic) cumulative distribution function $F_X(\mathrm{x})$. Let $0 < p < 1$ and define x_p to be such that

$$\mathrm{x}_p = F_X^{-1}(p)$$

which means that

$$\Pr[X \leq \mathrm{x}_p] = F_X(\mathrm{x}_p) = p$$

Then x_p is called the $100p$th quantile of the distribution of X.

- ⊡ `[Importance of the CDF]`. Clearly, one needs the CDF $F_X\mathrm{x}$ to be able to compute the quantiles.

- ⊡ If the distribution is discrete

$$\mathrm{x}_p = F_X^{-1}(p) = \inf\{\mathrm{x} \in R_X : p \leq F_X(\mathrm{x})\}$$

[1] *In fact, if μ happens to be in the fixed interval, then the probability is 1. If it is not in the fixed interval, the probability is 0. We are making a statement about the technique for building such intervals, and not about the fixed interval in hand.*

- ⊡ For some distributions, this calculation is straightforward. However, in some cases, getting quantiles can be very difficult and require computationally intensive simulations.

- ⊡ `[Sample quantiles]:` The mighty R provides the very handy command `quantile` that computes the quantiles for a given sample. Let $\mathbf{x} = (\mathrm{x}_1, \mathrm{x}_2, \cdots, \mathrm{x}_n)$ be the vector that contains the values observed in my sample. With `R`, it is easy to obtain sample quantiles. For instance, the 95th quantile is

```
x.95<-quantile(x, 0.95)
```

`[Crucial role of quantiles]:` *It cannot be overemphasized that obtaining quantiles is one of the most important part of building confidence intervals.*

`Exercise:` Let X be a random variable with pdf

$$f(\mathrm{x}) = \begin{cases} \frac{1}{\beta} e^{-\frac{1}{\beta}\mathrm{x}} & \mathrm{x} > 0 \\ 0 & \text{otherwise} \end{cases}$$

- ⊡ Show that f is a valid probability density function.

- ⊡ Write down the cdf of X.

- ⊡ Derive the general formula for the $100p$th quantile for this distribution.

Precision-accuracy dilemma and Precision-accuracy tradeoff

We cannot finish this subsection about confidence intervals without mentioning one of the most interesting aspect of it, namely the `precision-accuracy dilemma`.

- ⊡ The confidence level $1-\alpha$ represents the `accuracy` of our estimation. Clearly, if $1-\alpha$ is large, i.e. close to 1, then almost all our confidence intervals will contain the true parameter. For illustration, 95% confidence intervals are more accurate than 90% CI since the latter miss 10% of the time while the former only misses 5% of the time. Therefore, high confidence is synonymous with high accuracy.

- ⊡ On the other hand, the margin of error $\mathtt{E}_{100(1-\alpha)\%}$ represents the `precision` of the estimation because when it is small, the confidence intervals are narrow and one somewhat zeroes in on the true parameter, since the width of the intervals are small. Small margin of error is synonymous with high precision and is therefore desirable.

- ⊡ It is obvious from the expression of the margin of error that it depends of the confidence level. Unfortunately the dependence is such that high values of $1-\alpha$ lead to high margins of error. In other words, our ideal scenario of `high accuracy and high precision` runs into problems. This is what we term the `precision-accuracy dilemma`.

- ⊡ Since the formula for the margin of error also has quantities that do not depend on the confidence level, the `precision-accuracy dilemma` is addressed via a trade-off achieved by playing on the sample size.

- ⊞ Fix the desired confidence level $1-\alpha$. `[accuracy]`
- ⊞ Fix the desired margin of error $\mathtt{E}_{100(1-\alpha)\%}$. `[precision]`
- ⊞ Determine the sample size that allows the above two to be realized. `[trade-off]`

The only trouble with the above `precision-accuracy tradeoff` is that collecting samples may be practically too difficult or too expensive, rendering the trade-off unrealizable.

Confidence Interval for Average Annual Golf Earnings

For our golf earnings example, we have $n = 187$ and $\bar{\mathrm{x}} = 1313000$ along with $\mathtt{s}_{\mathrm{x}} = 1079247$. Therefore the estimated standard error on $\bar{\mathbf{x}}$ is

$$\mathtt{ese}(\bar{\mathbf{x}}) = \frac{\mathtt{s}_{\mathrm{x}}}{\sqrt{n}} = \frac{1079247}{\sqrt{187}} = 78922.38$$

Let's now build a 95% confidence interval. We have $\alpha = 0.05$ and therefore $\alpha/2 = 0.025$ so that we need $z_{0.025} = \Phi^{-1}(0.975) = 1.96$. This number can be found in `R` quite readily

```
z.0025 <- qnorm(0.975)
```

Therefore our 95% margin of error in this case is

$$\mathtt{E}_{95\%} = z_{0.025} \times \mathtt{ese}(\bar{\mathbf{x}}) = 1.96 \times 78922.38 = 154687.9$$

Now our 95% confidence interval for the average annual earnings of golfers is

$$[\bar{\mathbf{x}} - \mathtt{E}_{95\%}, \bar{\mathbf{x}} + \mathtt{E}_{95\%}] = [1157908, 1467283]$$

We are 95% confident that the average annual earnings of golfers is between \$1157908 and \$1467283.

Technically, I phrased my concept of lucrativeness in such a way that a double sided confidence interval like the one above is not entirely appropriate. To rigorously address the question, I need to build a one sided confidence interval, namely a $100(1-\alpha)\%$ lower confidence bound for μ. The main difference here is that the margin on error is now

$$\mathtt{E}_{95\%} = z_{0.05} \times \mathtt{ese}(\bar{\mathbf{x}}) = 1.65 \times 78922.38 = 1182374$$

The percentage of wrong samples is concentrated on one side. For the annual golf earnings, we have

$$[\bar{\mathbf{x}} - \mathtt{E}_{95\%}, +\infty] = [1182374, +\infty)$$

We are 95% confident that the average annual earnings of golfers is at least equal to \$1182374. Thanks to these results, we can now say infer that `professional golf is lucrative` by the standard we set.

Hypothesis Testing: It is interesting to provide a difference analysis based on hypothesis testing. Let's consider the following upper tail (right sided) test on a single unknown population mean μ:

$$\left[\begin{array}{lll} \mathtt{H}_0: & \mu \leq \$1M & \texttt{Professional golf is not lucrative} \\ \mathtt{H}_\mathtt{a}: & \mu > \$1M & \texttt{Professional golf is lucrative} \end{array}\right.$$

Here, instead of formulating our question as a statistical estimation problem, we have formulated it as hypothesis testing. Think of it as a debate between two contradicting sides, where one side's hypothesis (view) necessarily contradicts the other side. Remember that we do not know the true value of μ. In this case, I hypothesized something - we shall call it from now on the `null value`. In hypothesis testing, statements are made by the two parties.

- ⊡ The null hypothesis $\mathtt{H_0}$ is sometimes referred to as the `status quo` hypothesis
- ⊡ The alternative hypothesis $\mathtt{H_a}$ is known as the `research` hypothesis.

Grosso modo, the procedure of hypothesis testing has the following steps:

1. `Clearly specify the parameter of interest`
2. `Specify the null and the alternative hypotheses`
3. `Specify the desired significance level` α `(to be clarified later).`
4. `Obtain or collect the data on which to base the test`
5. `Find a point estimator for the parameter of interest`
6. `Devise the sampling distribution of that estimator`
7. `Specify a strategy for deciding which hypothesis to reject`
8. `Make the desired inference i.e. decide.`

For our particular annual golf earnings case study, we have covered all the steps except for 7 and 8. However since we know the sampling distribution of $\bar{X}$ from our previous task of confidence interval construction, we can go straight to the computation of the so-called P-value. Basically, we have here an upper tail test and we can use the command provided readily by `R`, namely `t.test()`[2]. Before, using th command `R` command however, let's explain a little bit what a P-value really is.

`P-value`: The calculation of the P-value proceeds as follows: *Assuming that the null hypothesis is true, what is the probability of finding a random sample whose observed statistic goes in the direction of the alternative hypothesis like the value observed in the present sample?* [3]

For instance, for a upper tail test on a population mean μ like the one we have now, the P-value is simply,

$$\mathtt{P\text{-}value} = \Pr[\bar{X} > \mathrm{x}_{\mathtt{obs}} | \mathtt{H_0\ \ is\ true}] = \Pr[\bar{X} > \mathrm{x}_{\mathtt{obs}} | \mu = \mu_0].$$

Now, since the CLT allows us in this case to have a normal distribution for $\bar{X}$, we have

$$\mathtt{P\text{-}value} = \Pr[\bar{X} > \mathrm{x}_{\mathtt{obs}} | \mu = \mu_0] = 1 - \Phi\left(Z \leq z_{\mathtt{obs}}\right) = 1 - \Phi(3.961) \approx 0.$$

[2] Rigorously speaking, we should be performing a Z-test in this case, because the sample size is very large and the CLT applies. However, the t.test we are conducting yields the same result as the Z.test when samples are large.

[3] *Always remember that a P-value is a tail probability, and represents the observed significance level. Hence the decision rule that compares the P-value to the specified significance level* α.

where

$$z_{\texttt{obs}} = \frac{\bar{\texttt{x}}_{\texttt{obs}} - \mu_0}{\frac{\texttt{s}_{\texttt{x}}}{\sqrt{n}}}$$

is the so-called test statistic[4].

Clearly, the knowledge of the sampling distribution of the statistic $\bar{X}$ is central to the calculation of the P-value. Note also that the knowledge of the null hypothesis is CRUCIAL, as it determines what direction the probabilistic statement must go in. Hence the following cardinal rule:

Cardinal Rule for P-values:

- Before even attempting to interpret a P-value, ALWAYS be sure to first find out clearly what are the hypotheses being tested.
- Reject the null hypothesis $\texttt{H}_0$ if the P-value is less than the significance level α, i.e

  $$\texttt{If Pvalue} < \alpha \quad \texttt{then Reject H}_0$$

  When this happens, the test is said to be statistically significant, because the research hypothesis wins over the status quo hypothesis.

The R command for performing the desired hypothesis test is

```
t.test(golf$Earnings, conf.level = 0.95, mu = 1000000, alternative="greater")
```

The output yielded by R for the above upper tail test on μ is

```
        One Sample t-test
data:  golf$Earnings
t = 3.961, df = 186, p-value = 5.316e-05
alternative hypothesis: true mean is greater than 1e+06
95 percent confidence interval:
      1182130      Inf
```

Since the P-value is less than any reasonable significance level, we have a strong evidence to reject the null hypothesis, and to retain $\texttt{H}_\texttt{a}$ which states that professional golf is lucrative.[5]

[4] Statistical Significance vs Practical Significance: If you look carefully at the expression for the P-value above, you will notice that it depends on n. More specifically, the test statistic can be made arbitrarily large by increasing the sample size n. Indeed, the larger n, the easier it is to reject the null hypothesis, leading to the conclusion of statistical significance. Making n arbitrarily large may therefore lead to a statistical significance that has no practical significance.

[5] You might have noticed that the lower confidence bound for μ produced by R is slightly different from the one we computed by hand. This is not an inconsistency. It is simply due to the fact quantiles from the t-distribution were used instead of Gaussian quantiles. Specifically,

$$t_{n-1,\alpha} = t_{186,.95} = 1.653$$

It can be obtained in R by using the command for Student's t-distribution quantiles, namely

$$\texttt{t.95} < -\texttt{qt}(186, 0.95)$$

Be mindful of errors: When assessing the output of a test, it is crucial to always first find out what are the hypotheses being tested: what is null hypothesis stating? It is also primordial to be mindful of the fact that the decision of the test could be wrong.

Courtroom analogy: To clearly illustrate what a hypothesis test really entails, consider the following courtroom scenario.

$$\begin{cases} H_0: \text{ The defendant is innocent} \\ H_a: \text{ The defendant is NOT innocent (he/she is guilty)} \end{cases}$$

When we are conducting a hypothesis test, there is the true state of nature that we do not know in practice, and there is the decision that we ultimately make based on the evidence provide by the sample. It goes without saying that we will either make the correct decision and make an error, as summarized in the following table.

		Decision	
		H_0 is not rejected	H_0 is rejected
State of nature	H_0 True	Correct decision	Type I Error
	H_0 False	Type II Error	Correct decision

Table 1: Possible outcomes of a Hypothesis test.

- ⊡ **[Correct decision]:** If the null hypothesis is true and we do not reject it, or the null hypothesis is false and we do reject it, then we make the correct decision

  The defendant was innocent and the judge did set him free, or the defendant was guilty and the judge convicted him

- ⊡ **[Type I Error]:** If the null hypothesis is true, and we mistakenly reject it, then we are said to have committed a Type I Error

  The defendant was innocent but the judge mistakenly convicted him

- ⊡ **[Type II Error]:** If the null hypothesis is false, and we mistakenly fail to reject it, then we are said to have committed a Type II Error

  The defendant was guilty but the judge mistakenly let him/her go

Errors are inevitable

- ⊡ **[Errors are inevitable]:** The truth is that there will be errors. It is just a fact of life. No one in his/her right mind should think of a judicial system that has it right all the time. Therefore, a judge is deemed great, not because he/she makes no errors (that's impossible) - but because he/she makes the correct decision most of the time.

- ⊡ [Random variation]: Statistically, errors due to random variation CANNOT be avoided. Therefore, even the most powerful test will occasionally deliver the wrong (incorrect) decision, i.e rejecting a true null or failing to reject a false null.

Goal: *The focus should be on the proportion of correct decisions made in the long run*

$$\text{maximize the probability of correct decision} \tag{4}$$

Hypothesis testing dilemma: However, there is another problem, probably one of the most central theme to statistics. Let's call it Hypothesis testing dilemma, and let's explain why it is a dilemma. Clearly there is a dilemma. To see the dilemma mathematically, let

$$\alpha = \Pr[\text{H}_0 \text{ is rejected when } \text{H}_0 \text{ is true}] = \Pr[\text{Type I Error}]$$

and let

$$\beta = \Pr[\text{H}_0 \text{ is not rejected when } \text{H}_0 \text{ is false}] = \Pr[\text{Type II Error}]$$

Here is the Ideal for Hypothesis Testing:

Ideally, one wants to achieve both small α and small β simultaneously.

Ideal is unreacheable:

Unfortunately, since the decision to reject or not to reject the null hypothesis depends on α, it follows that β depends on α in a way that makes the above ideal difficult to achieve.

The quantity

$$1 - \beta = \Pr[\text{H}_0 \text{ is rejected when } \text{H}_0 \text{ is false}]$$

is known as the power of the test. The quantity

$$1 - \alpha = \Pr[\text{H}_0 \text{ is not rejected when } \text{H}_0 \text{ is true}]$$

is known as the confidence level. α is the significance level of the test.

`Example of Hypothesis testing dilemma situation`: Let's reconsider our courtroom analogy:

- ⊡ If you concentrate on guaranteeing that no truly guilty person ever gets mistakenly set free, then you will quite frequently send innocent people to jail.
- ⊡ On the other hand, if you are Mr nice judge and adamantly want make sure that no innocent person even gets mistakenly convicted, you are going to set a fair number of criminals free to room society.

$$\texttt{if } \alpha \texttt{ is set too small, then } \beta \texttt{ will grow, and vice versa.}$$

This dilemma can be termed the `significance-power dilemma`. This dilemma can be addressed in a manner similar to the way we addressed the `accuracy-precision dilemma` in interval estimation.

- ⊡ Fix the desired significance level α.
- ⊡ Fix the desired power $1 - \beta$.
- ⊡ Determine the sample size that allows the above two to be realized. `[trade-off]`

The only trouble with the above `significance-power tradeoff` is that collecting samples may be practically too difficult or too expensive, rendering the trade-off unrealizable.

`Computing the power of a simple test:` Recall that the power of a test is defined as the probability of rejecting $\texttt{H}_0$ when $\texttt{H}_0$ is false, that is

$$\texttt{Power} = \Pr[\texttt{H}_0 \texttt{ is rejected when } \texttt{H}_0 \texttt{ is false}] = 1 - \beta(\alpha)$$

For instance, for a large sample upper tail test, the expression of power is given by

$$\texttt{Power}(\alpha) = 1 - \Phi\left(z_\alpha + \sqrt{n}\left[\frac{\mu_0 - \mu'}{\sigma}\right]\right) = 1 - \beta(\alpha)$$

Note that the power of a test clear depends on:

- ⊡ `[The magnitude of the alternative value]`. If the alternative is really distinctly (substantially) different from the null, then the power grows, as it is easy to reject the null in such cases. If however the difference is not substantial, the test will struggle. *If the characteristics of the suspect are very similar to those of the true criminal, then even the best judge will have a hard time not making errors.* In other words, *If the alternative value μ' is very close to μ_0, then $\mu_0 - \mu'$ becomes small and even the best test will confuse the wheat and the chaff. If however μ' is really far greater than μ_0, then $\Phi\left(z_\alpha + \sqrt{n}\left[\frac{\mu_0 - \mu'}{\sigma}\right]\right)$ becomes very small, leading to an increase in power.*
- ⊡ `[The value of the significance level` α`]`. If α is set small, then $\Phi\left(z_\alpha + \sqrt{n}\left[\frac{\mu_0 - \mu'}{\sigma}\right]\right)$ becomes too large, leading to a decrease in power.

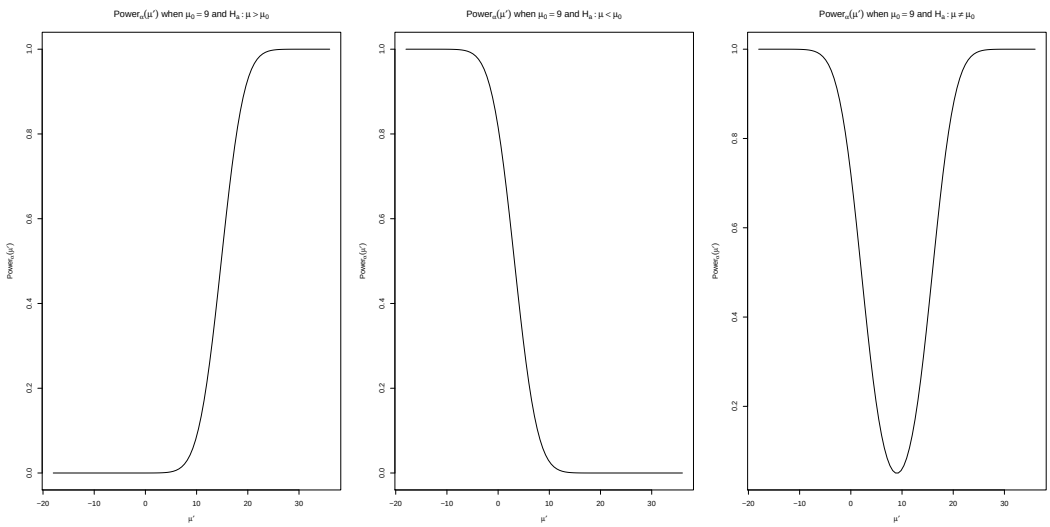


Figure 3: Curves of the power of a test (left) Upper sided test (Middle) Lower sided test (right) Double sided test

- ⊡ `[The variability in the population being studied (`σ`)]`. If σ turns out to be too large, then $\Phi\left(z_\alpha + \sqrt{n}\left[\frac{\mu_0 - \mu'}{\sigma}\right]\right)$ becomes large too large also, leading to a decrease in power.
- ⊡ `[The sample size` n`]`. If the sample size is large, then $\Phi\left(z_\alpha + \sqrt{n}\left[\frac{\mu_0 - \mu'}{\sigma}\right]\right)$ gets small, leading to an increase in power.

Similar reasoning can be made for tests other than the current upper tail test. Note that the computation of the power requires the knowledge of the sampling distribution of the test statistic. In practice, one often has to approximate the sampling distribution using such techniques as the bootstrap.

2.5 Conclusion

We have used statistical inference, namely interval estimation and hypothesis testing to answer the question: `Is professional golf lucrative`? Our statistical findings revealed `professional golf to be lucrative` indeed based on the standard we set.

- ⊡ Since we also noticed in our informal analysis that there were possible two groups in the golfers population as far as earnings were concerned, a reasonable follow up to our present analysis would be to consider studying the two hypothetical groups separately, maybe comparing for those groups features other than earnings.
- ⊡ It will also be good to consider taking many different years into consideration rather than just focusing on one single year.
- ⊡ Finally, since the mean is not the best measure of typicalness for a skewed distribution like the one we had, it is worth considering a nonparametric test that does not use the

mean but instead uses other measures like the median annual earnings of golfers.

$$\begin{cases} H_0: \text{ Median} \leq \$1M & \text{Professional golf is not lucrative} \\ H_a: \text{ Median} > \$1M & \text{Professional golf is lucrative} \end{cases}$$

```
wilcox.test(golf$Earnings, mu = 1000000, alt="greater")
Wilcoxon signed rank test with continuity correction
data:  golf$Earnings
V = 10484, p-value = 0.01112
alternative hypothesis: true location is greater than 1e+06
```

With `P-value` equal to 0.01112, we reject the null hypothesis at significance level 0.05 and conclude that `Professional Golf is lucrative indeed.`

When we call the command

```
test.golf<-t.test(golf$Earnings, conf.level = 0.95, mu = 1000000, alternative="greater")
```

The `R` object `test.golf` contains fields that can be accessed for used beyond the test itself. For instance, one may desire to store the p-values for a series of tests. It is therefore desirable to be able to extract the p-value and store it. Now, the object has the following fields.,

```
             Length Class  Mode
statistic    1      -none- numeric
parameter    1      -none- numeric
p.value      1      -none- numeric
conf.int     2      -none- numeric
estimate     1      -none- numeric
null.value   1      -none- numeric
alternative  1      -none- character
method       1      -none- character
data.name    1      -none- character
```

Therefore, I may perform desired operations like

```
p.value.golf <- test.golf$p.value   # Extract the p-value
print(test.golf$conf.int)           # display the confidence interval
```

2.6 The Two-sample t-test

The famous two sample test

$$\begin{cases} \mathtt{H_0}: & \mu_1 - \mu_2 = 0 \\ \mathtt{H_a}: & \mu_1 - \mu_2 \neq 0 \end{cases}$$

Are those golfers labelled as too rich by our standards better technically than those golfers we named ordinary? Before we attempt to answer this question, let's label golfers accordingly

```
rich.label<-rep(1,187)
rich.label[too.rich]<-2
boxplot(golf$Par.5.Birdies~rich.label, names=c('Ordinary','Too Rich'),
        main = 'Comparing Par 5 Birdies Performance', col = c(4,3))
```

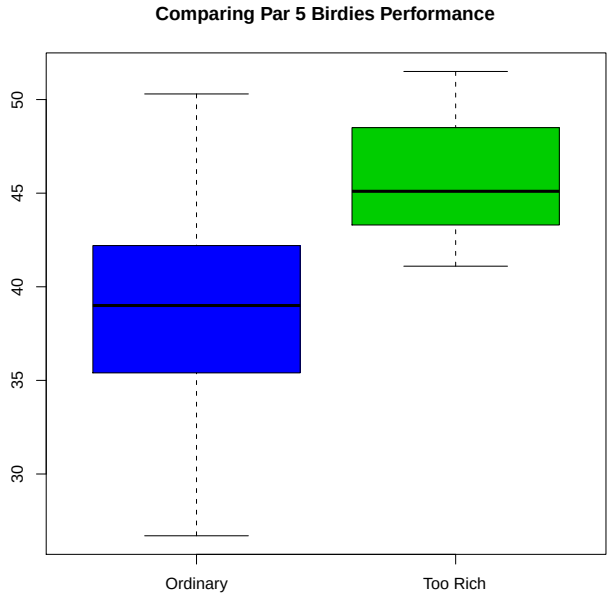


Figure 4: Comparative box plot for Par 5 Birdies. The truly rich golfers seem better than the rest.

```
> test.golf<-t.test(x[too.rich], x[-too.rich],conf.level = 0.95,
                    mu = 3000000, alternative="greater")

            Welch Two Sample t-test

data:  x[too.rich] and x[-too.rich]
t = 0.7184, df = 12.91, p-value = 0.2427
alternative hypothesis: true difference in means is greater than 3e+06
  95 percent confidence interval:
         2710645     Inf
        sample estimates:
      mean of x mean of y
       4287631   1090323
```

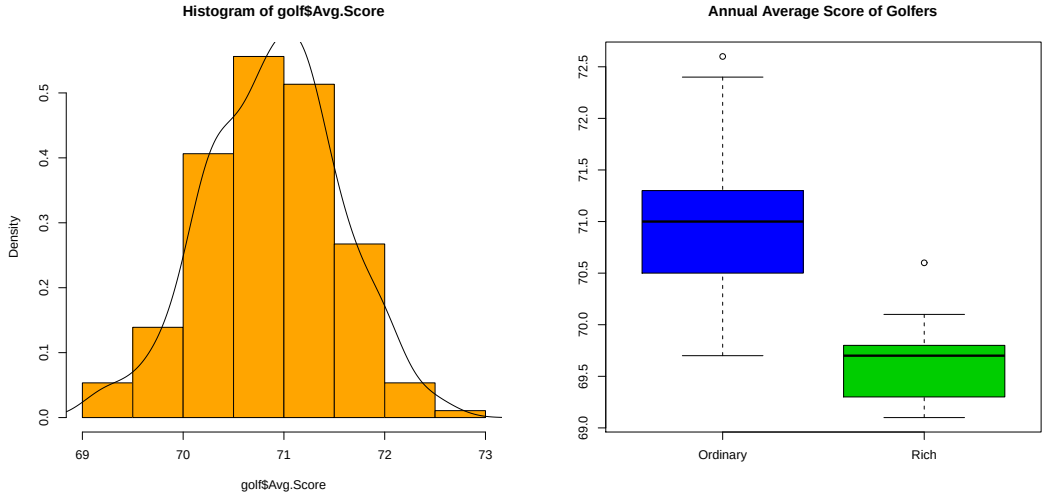
Do rich golfers really have a mean Annual Average Score less than that of ordinary golfers? To find, let μ_1 denote the average Annual average score of rich golfers, and μ_2 denote the average Annual average score of ordinary golfers. Then we need to perform the following test:

$$\left[\begin{array}{ll} \mathtt{H_0}: & \mu_1 - \mu_2 \geq 0 \\ \mathtt{H_a}: & \mu_1 - \mu_2 < 0 \end{array}\right.$$

```
        Welch Two Sample t-test

data:  golf$Avg.Score[too.rich] and golf$Avg.Score[-too.rich]
t = -10.69, df = 15.75, p-value = 6.292e-09
alternative hypothesis: true difference in means is less than 0
95 percent confidence interval:
   -Inf -1.097
sample estimates:
mean of x mean of y
    69.67     70.98
```

With $\mathtt{P\text{-}value} = 10^{-9} \approx 0 < \alpha(\mathtt{any})$, we reject the null hypothesis $\mathtt{H_0}$ and conclude that the average annual average score of rich golfers is indeed less than that of ordinary golfers. It



is clear from the histogram, that annual average scores are normally distributed. For good measure, we perform the shapiro test of normality and obtain the large $\mathtt{P\text{-}value} = 0.532$ that confirms the plausibility of normality.

```
> shapiro.test(golf$Avg.Score)

        Shapiro-Wilk normality test

data:  golf$Avg.Score
W = 0.9931, p-value = 0.532
```

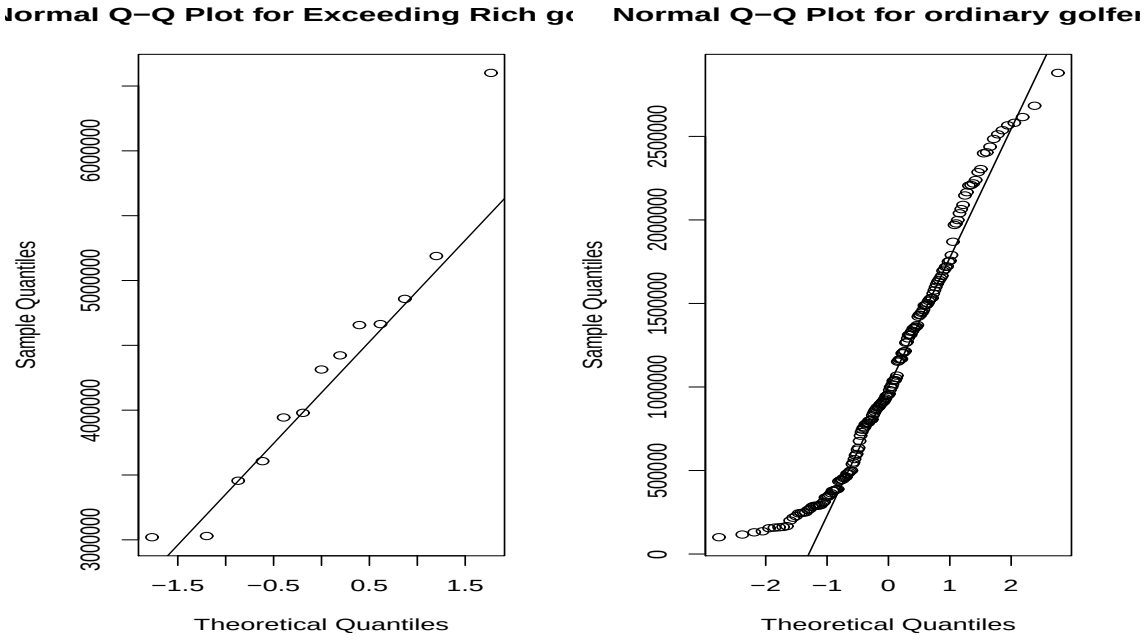
In fact, by the CLT, it is not surprising that annual Average Score is normally distributed because it is an average.

2.7 Normality Test

All the above t test procedures assume that the data comes from a normally distributed population. Of course, we here have large samples that allow us to use the CLT. However, let's see what the normality test does.

$$\left[\begin{array}{l} \texttt{H}_0: \texttt{ The sample } (x_1, x_2, \cdots, x_n) \texttt{ came from a normally distributed population} \\ \texttt{H}_a: \texttt{ The sample } (x_1, x_2, \cdots, x_n) \texttt{ did NOT come from a normally distributed population} \end{array}\right.$$

Before performing the formal test of normality, a plot known as the Q-Q plot for the normal distribution is generated. If the points on the plot are reasonably lined up with the straight line, then it is reasonable to conclude that normality of the generating population is plausible.



Clearly, plot for ordinary golfers reveals deviations that warrant rejecting the hypothesis of normality. Let's conduct a formal test to determine this clearly.

```
library(nortest)
shapiro.test(x[-too.rich])
     Shapiro-Wilk normality test

data:  x[-too.rich]
W = 0.948, p-value = 5.258e-06
```

However, since there are a large number of ordinary golfers, this failure of the normality test is not crucial. For the super rich, the following test delivers a `P-value` $= 0.4 > 0.05 = \alpha$. We therefore conclude that the distribution of annual earnings for all super rich golfers can be assumed to be normal.

```
shapiro.test(x[too.rich])
     Shapiro-Wilk normality test
```

```
data:  x[too.rich]
W = 0.9356, p-value = 0.4020
```

2.8 Test about a population proportion p

Let

$$p \equiv \texttt{Proportion of all annual golfers' earnings that exceed } \$3,000,000.$$

We are interested in finding out if really more than 10% of golfers exceed $\$3,000,000$ in their annual earnings. Statistically, that translates into the following lower tail test

$$\left[\begin{array}{ll} \texttt{H}_0: & p \geq p_0 \\ \texttt{H}_\texttt{a}: & p < p_0 \end{array}\right.$$

where $p_0 = 0.1$. With `R`, the above is a simple test about a population proportion p.

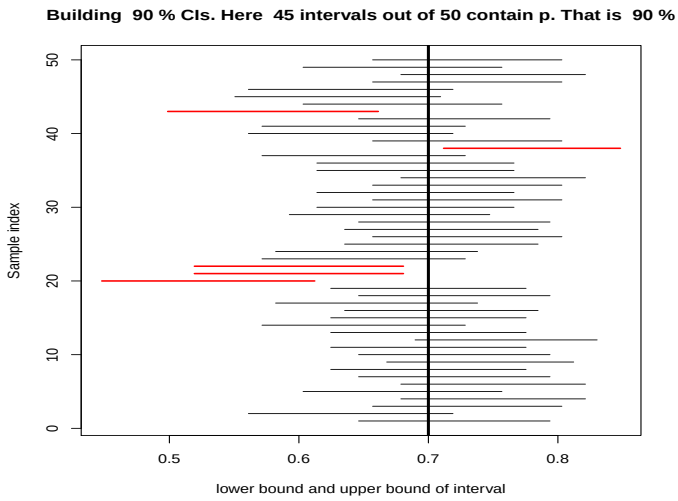
```
binom.test(length(rich.label[too.rich]), length(rich.label),
              p=0.1, alternative="less")

               Exact binomial test

data:  length(rich.label[too.rich]) and length(rich.label)
number of successes = 13, number of trials = 187, p-value = 0.09814
alternative hypothesis: true probability of success is less than 0.1
              95 percent confidence interval:
                   0.0000 0.1083
```

- ⊡ The above result reveals that `P-value` $= 0.09814 > 0.05 = \alpha$. Since `P-value` $> \alpha$, the null hypothesis cannot be reject. Therefore, we conclude that at significance level $\alpha = 0.05$, at least 10% of annual golfers' earnings exceed $\$3,000,000$.
- ⊡ Now, if you use the significance level $\alpha = 0.1$, then `P-value` $= 0.09814 < 0.1 = \alpha$. With that, the null hypothesis is rejected. Therefore, we conclude that at significance level $\alpha = 0.1$, there is not sufficient evidence from the data to support the assertion that at least 10% of annual golfers' earnings exceed $\$3,000,000$.

Below is a graphical demonstration of the confidence interval concept
`Question:` Clearly and succintly explain what this plot says.



3 Common Statistical Tests and their R Commands

Exercises on Basic Statistical Concepts

3-1. Find the expression of the power of the large sample lower tail test about the population mean μ.

3-2. `Demonstration of the central limit theorem:` Explore the `R` script `clt-702.R` for various distributions.

3-3. `Demonstration of confidence intervals:` Modify the `R` script `conf-int-prop-1.R` to demonstrated the concept of confidence intervals for a population mean μ.

(a) Try $\alpha = 0.1$ and $\alpha = 0.05$

(b) Describe what you notice.

3-4. `Computational demonstration of the power of a test`. Generate $m = 100$ samples, each of size $n = 2$, from a normal distribution with mean $\mu = 9$ and variance $\sigma^2 = 2^2$.

(a) For each sample, perform the lower tail test

$$\left[\begin{array}{ll} \mathtt{H_0}: & \mu \geq \mu_0 \\ \mathtt{H_a}: & \mu < \mu_0 \end{array}\right.$$

with $\mu_0 = 10$, and store the corresponding `P-value`.

(b) Compute the proportion `p` of times the null hypothesis is rejected.

(c) Repeat (a) and (b) for the following: $\mu = 8$, $\mu = 7$, $\mu = 6$, $\mu = 5$, $\mu = 4$.

- ⊡ Store the proportions in a vector, and plot them as a function of the μ's.
- ⊡ What do you notice? Does it confirm your intuition?

(d) For a significance level α, a lower tail test based on a `large` random sample of n observations from a population with spread σ, has power given by

$$\mathtt{Power}(\alpha) = \Phi\left(-z_\alpha + \sqrt{n}\left[\frac{\mu_0 - \mu'}{\sigma}\right]\right)$$

Compare the values obtained earlier with this one.

3-5. `Computational demonstration of the power of a test for test of a population proportion.` Consider the upper tail test (right sided test)

$$\left[\begin{array}{ll} \mathtt{H_0}: & p \leq p_0 \\ \mathtt{H_a}: & p > p_0 \end{array}\right.$$

Provide both the theoretical and the computational determination of the power of this test for very large samples.

3-6. `Computational demonstration of the power of a double sided test.` Consider the following double sided test on a single unknown population mean μ:

$$\left[\begin{array}{ll} \mathtt{H_0}: & \mu = \mu_0 \\ \mathtt{H_a}: & \mu \neq \mu_0 \end{array}\right.$$

(a) Find the theoretical expression of the power of the above test for large samples

(b) Perform computations similar to the ones done earlier

(c) Compare computational values to theoretical values