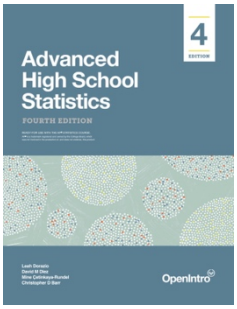
R FUNCTIONS FOR AP STATISTICS.

EXCERPTED FROM *ADVANCED HIGH SCHOOL STATISTICS, 4TH ED*, available for free at openintro.org/ahss.



- Summarizing a single Variable
- Binomial probabilities
- Normal probabilities and boundary values
- Inference for a population proportion
- Inference for a difference of population proportions
- Chi-square for one-way tables (special topic)
- Chi-square distribution probabilities
- Chi-square for two-way tables
- *t*-distribution probabilities and boundary values
- Inference for a population mean or a population mean of differences
- Inference for a difference of population means
- Scatterplots and regression

1.4.8 Technology: summarizing a single variable

Download the `loan50` CSV file from openintro.org/data. Open it and graphically and numerically summarize the `loan_amount` variable.

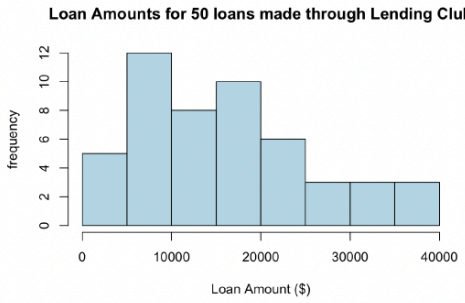
R: You will need to first download R and RStudio from: `posit.co/download/rstudio-desktop`.

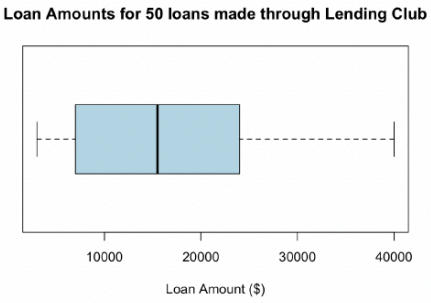
1. Open the CSV file. Copy the data that corresponds to the `loan_amount` variable. Do not include the header/variable name.
2. In RStudio, at the > prompt, type: `a = scan()` and hit return. Paste the copied data, then hit return and return again.

```
> a = scan()
1:  22000
2:  6000
...
49:  20000
50:  15000
51:
Read 50 items
```

3. Use `hist()` and `boxplot()` as shown below. Replace labels with appropriate values.

```
> hist(a, main = "Loan Amounts for 50 loans made through Lending Club",
xlab = "Loan Amount ($)", ylab = "frequency", col = "lightblue")

> boxplot(a, main = "Loan Amounts for 50 loans made through Lending Club", xlab
= "Loan Amount ($)", horizontal = TRUE, col = "lightblue")
```





4. Use `summary()`,[40] and `sd()` functions as shown. You can ignore the `[1]`.

```
> summary(a)
Min.    1st Qu.    Median   Mean    3rd Qu.   Max.
3000    7125       15500    17083    24000    40000
> sd(a)
[1] 10455.46
```

Other ways to read in data:

Manually enter data between the parentheses of `c( )` as shown below.

```
> a = c(22000, 6000, 25000, ..., 5825, 20000, 15000)
```

Access a CSV file on a computer or on the web. Adjust path and file name as necessary.

```
> loan50 = read.csv("Users/FirstnameLastname/Desktop/loan50.csv")
```

Install the `openintro` package to access *all* data sets found at openintro.org/data.

```
> install.packages("openintro")
> library(openintro)
```

[40] There are multiple definitions for computing quartiles. To ensure values match the definition used here, for small data sets of length n, you may need to add `quantile.type=6` when n is odd or `quantile.type=2` when n is even.

2.5.7 Technology: binomial probabilities

A spinner has four equally likely options: red, green, blue, yellow. If you spin the spinner 6 times, what is the probability you get:

a) Exactly 2 red?

b) Less than 2 red?

c) At least 3 red?

This scenario describes a **binomial distribution** with $n = 6$ and $p = 0.25$.

R: Given a binomial distribution with $n = 6$ and $p = 0.25$, calculate the following probabilities.

a) P(X = 2). Use `dbinom(x, size, prob)` for the probability of exactly `x`.
Note that "`size =`" and "`prob =`" can be omitted, but the labels are often included for clarity.

```
> dbinom(2, size = 6, prob = 0.25)    or    > dbinom(2, 6, 0.25)
[1] 0.2966309
```

b) P(X ≤ 1). Use `pbinom(q, size, prob)` for the probability of ≤ `q` as shown.

```
> pbinom(1, size = 6, prob = 0.25)
[1] 0.5339355
```

c) P(X ≥ 3).

```
> 1 - pbinom(2, size = 6, prob = 0.25)
[1] 0.1694336
```

or

```
> pbinom(2, size = 6, prob = 0.25, lower.tail = FALSE)
[1] 0.1694336
```

2.6.4 Technology: normal probabilities and boundary values

Given a standard normal distribution ($\mu = 0$, $\sigma = 1$),

(i) find the following probabilities:

(a) Probability of a value between -2 and 2.

(b) Probability of a value less than -1.5.

(ii) find the following boundary values:

(a) The value that corresponds to the 40th percentile.

(b) The value $z^\star$ such that 95% of the distribution lies between $-z^\star$ and $z^\star$.

*With the standard normal distribution ($\mu = 0$, $\sigma = 1$), the values of the distribution correspond to Z-scores. If a problem involves a normal distribution with a different mean and standard deviation (e.g. battery life), we can use the mean and standard deviation of that distribution in place of 0 and 1.

R: Normal probabilities and boundary values for the standard normal distribution.

(i) Finding probabilities/areas. Use the `pnorm(x, mean, stdev)` function, which gives the area to the *left* of the `x` value entered, unless specified otherwise. Note that "`mean =`" and "`sd =`" can be omitted, but the labels are often included for clarity.

(a) $P(-2 \le X \le 2)$.

```
> pnorm(2, 0, 1) - pnorm(-2, 0, 1)
[1] 0.9544997
```

(b) $P(X \le 1.5)$.

```
> pnorm(-1.5, mean = 0, sd = 1)
[1] 0.0668072
```

(ii) Finding boundary values. Use the `qnorm(p, mean, stdev)` function. `qnorm()` returns the value that has the entered probability `p` to the *left* of it, unless specified otherwise.

(a) Find the value that corresponds to the 40th percentile.

```
> qnorm(0.40, 0, 1)
[1] -0.2533471
```

(b) Find the value $z^\star$ such that 95% of the distribution lies between $-z^\star$ and $z^\star$. This implies that 2.5% in the upper (and lower) tail.

```
> qnorm(0.025, mean = 0, sd = 1, lower.tail = FALSE)
[1] 1.959964
```

3.4.8 Technology: the one-sample Z-interval and Z-test for p

Evaluate the 95% confidence interval from Example 3.29 to estimate the proportion of US adults who think there is intelligent life on other planets. Also find the test statistic and p-value for a test to see if there is evidence that the true proportion differs from 0.70. The sample percent was 68% and the sample size was 1,033. Conditions were verified to be met.

R: 1-sample Z-interval/test for p

CONFIDENCE INTERVAL.

```
> prop.test(x = 702, n = 1033, correct = FALSE, conf.level = 0.95)[39]
1-sample proportions test without continuity correction
data:  702 out of 1033, null probability 0.5
X-squared = 133.24, df = 1, p-value < 0.00000000000000022
alternative hypothesis:  true p is not equal to 0.5
95 percent confidence interval:
 0.6504973 0.7073202
sample estimates:
p
0.6795741
```

HYPOTHESIS TEST.
Make sure to specify `p`, the hypothesized or null proportion.
`alternative` can be `"two.sided"`, `"greater"`, or `"less"`.

```
> prop.test(x = 702, n = 1033, p = 0.70, correct = FALSE, alternative = "two.sided")
1-sample proportions test without continuity correction
data:  702 out of 1033, null probability 0.7
X-squared = 2.0523, df = 1, p-value = 0.152
alternative hypothesis:  true p is not equal to 0.7
95 percent confidence interval:
0.6504973 0.7073202
sample estimates:
p
0.6795741
```

This test returns X-squared instead of Z.
$Z = +\sqrt{\texttt{X-squared}}$ or $-\sqrt{\texttt{X-squared}}$, depending on if `(sample p - null p)` is +. or −.
Here $(0.6504973 - 0.70)$ is negative, so $Z = -\sqrt{2.0523}$.

```
> Z = -sqrt(2.0523)
> Z
[1] -1.432585
```

A note on default values:

- If `p` is omitted a default null value of 0.5 is used.
- If `correct = FALSE` is omitted, what is called the "continuity correction" will be applied.
- If `alternative` is omitted, a default value of `"two.sided"` is used.
- If `conf.level` is omitted, a default value of 0.95 is used.

[39] R uses a different formula for calculating the one proportion interval/test than what is presented in this textbook. Results may differ slightly from Desmos and handheld calculators.

3.7.4 Technology: the two-sample Z-interval and Z-test for $p_1 - p_2$

Figure 3.16, summarizes an experiment involving patients admitted to a hospital for a heart attack. 14 out of the 40 patients assigned to the treatment (blood thinner) group survived versus 11 out of 50 assigned to the control group. Calculate a 95% confidence interval for a difference in the proportion (treatment – control) that would survive. Also calculate the test statistic and p-value to test whether there is evidence that the true proportions *differ*. Conditions have been verified.

R: 2-proportion Z-interval/test for $p_1 - p_2$

Here we use `prop.test()` with `x = c(x1, x2)` and `n = c(n1, n2)`.

CONFIDENCE INTERVAL.

```
> prop.test(x=c(14, 11), n=c(40, 50), correct = FALSE, conf.level = 0.95)
2-sample test for equality of proportions without continuity correction
data:  c(14, 11) out of c(40, 50)
X-squared = 1.872, df = 1, p-value = 0.1712
alternative hypothesis:  two.sided
95 percent confidence interval:
-0.05716886 0.31716886
sample estimates:
prop 1 prop 2
0.35 0.22
```

HYPOTHESIS TEST.

```
> prop.test(x = c(14, 11), n=c(40, 50), correct = FALSE, alternative = "two.sided")
2-sample test for equality of proportions without continuity correction
data:  c(14, 11) out of c(40, 50)
X-squared = 1.872, df = 1, p-value = 0.1712
alternative hypothesis:  two.sided
95 percent confidence interval:
-0.05716886 0.31716886
sample estimates:
prop 1 prop 2
0.35 0.22
```

This test returns X-squared instead of Z.
Z = $+\sqrt{\texttt{X-squared}}$ or $-\sqrt{\texttt{X-squared}}$, depending on if sample `(prop 1 - prop 2)` is + or −.
Here (0.35 − 0.22) is positive, so Z = $\sqrt{1.872}$.

```
> Z = sqrt(1.872)
> Z
[1] 1.368211
```

3.8.5 Technology: the chi-square goodness of fit test

A spinner has four colors that are supposed to be equally likely: red, green, blue, yellow. Someone records the number of each color after many spins in a game. After 200 spins, there are 59 red, 44 green, 54 blue and 41 yellow. Use technology to find the test statistic, df, and p-value for a chi-square goodness of fit test, testing whether there is evidence that the colors do not have the same likelihood of coming up. Also find the expected values and chi-square contributions for the test. Assume conditions for the test are met.

R: Chi-square goodness of fit test
Use `chisq.test(x = c(observed counts), p = c(expected proportions))`.

```
> X = chisq.test(c(59, 44, 54, 41), c(0.25, 0.25, 0.25, 0.25)) or
> X = chisq.test(x = c(59, 44, 54, 41), p = c(0.25, 0.25, 0.25, 0.25))

> X
Chi-squared test for given probabilities
data:  c(59, 44, 54, 41)
X-squared = 4.303, df = 3, p-value = 0.2305

> X$expected
[1] 49.5 49.5 49.5 49.5

> (X$residuals)^2
[1] 1.8232323 0.6111111 0.4090909 1.4595960
```

3.9.7 Technology: chi-square distribution probabilities

Use technology to find the upper tail area for a chi-square distribution with 5 degrees of freedom and a cutoff of 5.1.

R:
`pchisq(q, df)` will give the area to the left of `q`, so we add `lower.tail = FALSE` to get the area to the right.

```
> pchisq(5.1, df = 5, lower.tail = FALSE)
[1] 0.4037985
```

3.9.8 Technology: the chi-square test for two-way tables

Consider the data from the search algorithm experiment introduced in Example 3.9.1. Use technology to find the test statistic, *df*, and p-value for a chi-square test for homogeneity, testing whether there is evidence that the algorithms do not perform equally well. Also find the expected values and chi-square contributions for this test. Conditions were verified to be met.

		Search algorithm			
		current	test 1	test 2	Total
outcome	success	3511	1749	1818	7078
	failure	1489	751	682	2922
	Total	5000	2500	2500	10000

R: Chi-square test for two-way tables
Use `chisq.test()` with `cbind()` as illustrated below to bind the *columns* into a table.

```
> X = chisq.test(cbind(c(3511,1489), c(1749,751), c(1818,682)))
> X
Pearson's Chi-squared test
data:  cbind(c(3511, 1489), c(1749, 751), c(1818, 682))
X-squared = 6.1203, df = 2, p-value = 0.04688
```

```
> X$expected
[,1] [,2] [,3]
[,1] 3539 1769.5 1769.5
[,2] 1461 730.5 730.5
> (X$residuals)^2
[,1] [,2] [,3]
[1,] 0.2215315 0.2374965 1.329330
[2,] 0.5366188 0.5752909 3.220055
```

4.2.3 Technology: t-distribution probabilities and boundary values

Given a t-distribution with 3 degrees of freedom:

(i) Find the area to the right of $t = 1$.

(ii) Find the boundary value $t^\star$ such that 95% of the area is between $-t^\star$ and $t^\star$.

R: Probabilities and boundary values for a t-distribution with a given df.

(i) `pt(q, df)` gives the area to the left of `q`, so we add `lower.tail = FALSE` to get the area to the right.

```
> pt(1, df = 3, lower.tail = FALSE)
[1] 0.1955011
```

(ii) `qt(p, df)` gives the $t^\star$ value that has the probability `p` to the *left* of it, unless specified otherwise. To have 95% in the middle implies that there is 2.5% in the upper (and lower) tail.

```
> qt(0.025, df = 3, lower.tail = FALSE)
[1] 3.182446
```

4.3.4 Technology: the one-sample t-interval and t-test for μ

The data set `loan50`, introduced in Chapter 1, contains information on randomly sampled loans. Download the `loan50` CSV file from openintro.org/data. Open it and calculate a 95% confidence interval for the true mean of `loan_amount`. Also find the test statistic, df, and p-value for a test with the alternative hypothesis that the true mean is less than $20,000.

R: 1-sample t-interval/test for μ

First store the data into a variable as described on page 52.
`> loan_amount = scan()` Hit return, paste the numerical data, then hit return and return again.

You can also manually type in data as follows:
`> loan_amount = c(22000, 6000, 25000, 6000...)`

CONFIDENCE INTERVAL.
`t.test(data, conf.level = )`

```
> t.test(loan_amount, conf.level = 0.95)
One Sample t-test
data:  loan_amount
t = 11.553, df = 49, p-value = 0.000000000000001348
alternative hypothesis:  true mean is not equal to 0
95 percent confidence interval:
 14111.59 20054.41
sample estimates:
mean of x
17083
```

HYPOTHESIS TEST.
`t.test(data, mu = , alternative = "two.sided","greater","less")`

If a hypothesized value for μ is not entered, a default of 0 is used. If `alternative` is not specified, a default of `"two.sided"` is used. Here our alternative hypothesis is $\mu < 20,000$.

```
> t.test(loan_amount, mu = 20000, alternative = "less")
One Sample t-test
data:  loan50$loan_amount
t = -1.9728, df = 49, p-value = 0.02709
alternative hypothesis:  true mean is less than 20000
95 percent confidence interval:
-Inf 19561.99
sample estimates:
mean of x
17083
```

If you have installed the `openintro` package as described on page 52, you can skip the first step of storing the data into the variable `loan_amount`. Instead, simply reference `loan50$loan_amount`:
`> t.test(loan50$loan_amount, conf.level = 0.95)`

4.6.4 Technology: the two-sample t-interval and t-test for $\mu_1 - \mu_2$

Use technology to find a 95% confidence interval for a difference in true average scores between Version A and Version B of the exam, as described in Example 4.5.4. Also find the test statistic and p-value for a two-sided test to evaluate whether there is evidence that the true means differ. Conditions were verified to be met.

Version	n	$\bar{x}$	s	min	max
A	30	79.4	14	45	100
B	30	74.1	20	32	100

R: 2-sample t-interval/test for $\mu_1 - \mu_2$
With all the data, use: `t.test(data 1, data 2, conf.level = , alternative = )`

Here we do not have all the data, so we install the BSDA (Basic Statistics and Data Analysis) package. You only need to enter the first line once on a computer and the second line once per R session.

```
> install.packages("BSDA")
> library(BSDA)
```

This allow us to use:

```
tsum.test(mean.x, sd.x, n.x, mean.y, sd.y, n.y, conf.level = , alternative = )
```

CONFIDENCE INTERVAL.

```
> tsum.test(79.4, 14, 30, 74.1, 20, 30, conf.level = 0.95)
Welch Modified Two-Sample t-Test
data:  Summarized x and y
t = 1.1891, df = 51.918, p-value = 0.2398
alternative hypothesis:  true difference in means is not equal to 0
95 percent confidence interval:
 -3.644372 14.244372
sample estimates:
mean of x mean of y
79.4 74.1
```

HYPOTHESIS TEST.

```
> tsum.test(79.4, 14, 30, 74.1, 20, 30, alternative = "two.sided")
Welch Modified Two-Sample t-Test
data:  Summarized x and y
t = 1.1891, df = 51.918, p-value = 0.2398
alternative hypothesis:  true difference in means is greater than 0
95 percent confidence interval:
-2.164646 NA
sample estimates:
mean of x mean of y
79.4 74.1
```

Note that `conf.level = 0.95` and `alternative = "two.sided"` are the default values if those arguments are omitted. You can also use `alternative = "less"` or `alternative = "greater"` .

5.3.5 Technology: Scatterplots and regression analysis

The data set `loan50`, introduced in Chapter 1, contains information on randomly sampled loans. Download the `loan50` CSV file from openintro.org/data. Open it and perform linear regression analysis for predicting `total_credit_utilized` from `total_credit_limit`.

R: Scatterplots and Linear Regression Analysis
First read in the x and y values as described on page 52. For simplicity, we use `scan()`. However, if you have loaded the `openintro` package as described near the bottom of page 53 you can use the `dataset$variable` structure, e.g. `loan50$total_credit_utilized`.

```
> x = scan()
1:  95131
2:  51929
...
50:  390156
51:
Read 50 items
```

```
> y = scan()
1:  32894
2:  78341
...
50:  52534
51:
Read 50 items
```

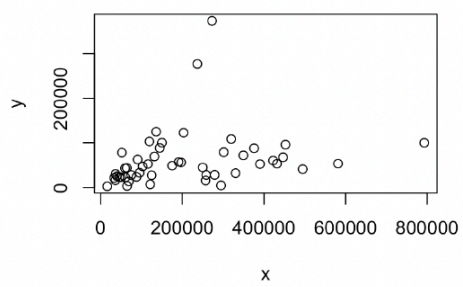
Use the `lm()` function to build a linear model and `summary()` to summarize it.

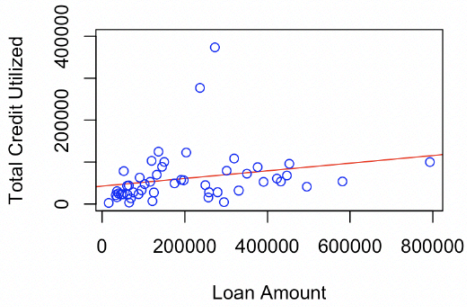
```
> model = lm(y ~ x)
> summary(model)
Residuals:
Min      1Q      Median   3Q      Max
-64713  -28247  -16934   10568  305912

Coefficients:
.              Estimate   Std.Error    t value   Pr(>|t|)
(Intercept) 42410.57946   14210.43201   2.984    0.00446 **
.         x     0.09176       0.05333   1.720    0.09179
---
Residual standard error:  62540 on 48 degrees of freedom
Multiple R-squared:  0.05808, Adjusted R-squared:  0.03846
```

Continued on next page…

Use `plot(x, y)` to make a scatterplot. For fun, type `help(plot)` at the `>` prompt and investigate the optional `type` argument.

```
> plot(x, y)
> plot(x, y, ylim=c(0,400000), col = "purple", xlab = "Loan Amount ($)",
ylab = "Total Credit Utilized", abline(model, col="red"))
```





Make a histogram of the residuals and a residual plot, plotting the residuals versus x. Add `xlab = " "` and `ylab = " "` as desired.

```
> hist(model$residuals)
> plot(x, model$residuals, abline(h=0))
```