Some stats topics in R

Jahred Liddie

2023-12-04

# An easy way to create a summary table

There are many packages that can help make tables containing summary statistics. Sometimes it can be easier to make them manually too. Below shows how to quickly get a summary of all the variables in a dataframe using skimr. Then, I show one example of how to make a more finalized summary table using qwraps2.

skim(lbw\_example)

Data summary

|  |  |
| --- | --- |
| Name | lbw\_example |
| Number of rows | 20 |
| Number of columns | 30 |
| \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_ |  |
| Column type frequency: |  |
| character | 2 |
| numeric | 28 |
| \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_ |  |
| Group variables | None |

**Variable type: character**

| skim\_variable | n\_missing | complete\_rate | min | max | empty | n\_unique | whitespace |
| --- | --- | --- | --- | --- | --- | --- | --- |
| adequate\_care | 0 | 1 | 8 | 12 | 0 | 3 | 0 |
| mother\_age\_cat | 0 | 1 | 3 | 5 | 0 | 3 | 0 |

**Variable type: numeric**

| skim\_variable | n\_missing | complete\_rate | mean | sd | p0 | p25 | p50 | p75 | p100 | hist |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| X.1 | 0 | 1 | 10.50 | 5.92 | 1 | 5.75 | 10.5 | 15.25 | 20 | ▇▇▇▇▇ |
| X | 0 | 1 | 10.50 | 5.92 | 1 | 5.75 | 10.5 | 15.25 | 20 | ▇▇▇▇▇ |
| child\_birthwt | 0 | 1 | 2732.00 | 834.19 | 533 | 2352.00 | 2870.0 | 3232.00 | 4470 | ▁▂▆▇▁ |
| use\_tobacco | 0 | 1 | 0.50 | 0.51 | 0 | 0.00 | 0.5 | 1.00 | 1 | ▇▁▁▁▇ |
| born\_in\_hospital | 0 | 1 | 1.00 | 0.00 | 1 | 1.00 | 1.0 | 1.00 | 1 | ▁▁▇▁▁ |
| MD\_at\_birth | 0 | 1 | 0.75 | 0.44 | 0 | 0.75 | 1.0 | 1.00 | 1 | ▂▁▁▁▇ |
| any\_drink\_on\_avg | 0 | 1 | 0.05 | 0.22 | 0 | 0.00 | 0.0 | 0.00 | 1 | ▇▁▁▁▁ |
| mother\_foreign\_born | 0 | 1 | 0.05 | 0.22 | 0 | 0.00 | 0.0 | 0.00 | 1 | ▇▁▁▁▁ |
| mother\_age | 0 | 1 | 26.55 | 4.81 | 19 | 23.50 | 26.5 | 30.25 | 36 | ▇▃▇▇▁ |
| father\_age | 0 | 1 | 28.95 | 4.94 | 20 | 26.00 | 27.0 | 32.00 | 37 | ▂▇▃▅▅ |
| num\_prenatal\_visits | 0 | 1 | 11.35 | 3.83 | 2 | 9.75 | 11.0 | 13.25 | 20 | ▁▃▇▆▁ |
| mother\_hispanic\_other | 0 | 1 | 0.00 | 0.00 | 0 | 0.00 | 0.0 | 0.00 | 0 | ▁▁▇▁▁ |
| mother\_black | 0 | 1 | 0.05 | 0.22 | 0 | 0.00 | 0.0 | 0.00 | 1 | ▇▁▁▁▁ |
| father\_hispanic\_other | 0 | 1 | 0.00 | 0.00 | 0 | 0.00 | 0.0 | 0.00 | 0 | ▁▁▇▁▁ |
| father\_black | 0 | 1 | 0.05 | 0.22 | 0 | 0.00 | 0.0 | 0.00 | 1 | ▇▁▁▁▁ |
| mother\_educ\_less\_than\_hs | 0 | 1 | 0.30 | 0.47 | 0 | 0.00 | 0.0 | 1.00 | 1 | ▇▁▁▁▃ |
| mother\_educ\_hs | 0 | 1 | 0.50 | 0.51 | 0 | 0.00 | 0.5 | 1.00 | 1 | ▇▁▁▁▇ |
| father\_educ\_less\_than\_hs | 0 | 1 | 0.20 | 0.41 | 0 | 0.00 | 0.0 | 0.00 | 1 | ▇▁▁▁▂ |
| father\_educ\_hs | 0 | 1 | 0.45 | 0.51 | 0 | 0.00 | 0.0 | 1.00 | 1 | ▇▁▁▁▆ |
| birth\_month | 0 | 1 | 6.05 | 2.91 | 1 | 4.00 | 6.0 | 7.25 | 12 | ▃▆▇▃▂ |
| birth\_weekday | 0 | 1 | 4.35 | 1.69 | 1 | 3.00 | 5.0 | 6.00 | 7 | ▅▂▃▇▇ |
| child\_female | 0 | 1 | 0.55 | 0.51 | 0 | 0.00 | 1.0 | 1.00 | 1 | ▆▁▁▁▇ |
| mother\_anemic | 0 | 1 | 0.00 | 0.00 | 0 | 0.00 | 0.0 | 0.00 | 0 | ▁▁▇▁▁ |
| mother\_cardiac\_disease | 0 | 1 | 0.00 | 0.00 | 0 | 0.00 | 0.0 | 0.00 | 0 | ▁▁▇▁▁ |
| mother\_lung | 0 | 1 | 0.05 | 0.22 | 0 | 0.00 | 0.0 | 0.00 | 1 | ▇▁▁▁▁ |
| mother\_herpes | 0 | 1 | 0.00 | 0.00 | 0 | 0.00 | 0.0 | 0.00 | 0 | ▁▁▇▁▁ |
| mother\_diabetes | 0 | 1 | 0.00 | 0.00 | 0 | 0.00 | 0.0 | 0.00 | 0 | ▁▁▇▁▁ |
| low\_weight | 0 | 1 | 0.45 | 0.51 | 0 | 0.00 | 0.0 | 1.00 | 1 | ▇▁▁▁▆ |

skim(lbw\_example$child\_birthwt)

Data summary

|  |  |
| --- | --- |
| Name | lbw\_example$child\_birthwt |
| Number of rows | 20 |
| Number of columns | 1 |
| \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_ |  |
| Column type frequency: |  |
| numeric | 1 |
| \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_ |  |
| Group variables | None |

**Variable type: numeric**

| skim\_variable | n\_missing | complete\_rate | mean | sd | p0 | p25 | p50 | p75 | p100 | hist |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| data | 0 | 1 | 2732 | 834.19 | 533 | 2352 | 2870 | 3232 | 4470 | ▁▂▆▇▁ |

# summary() is similar  
summary(lbw\_example)

## X.1 X child\_birthwt use\_tobacco born\_in\_hospital  
## Min. : 1.00 Min. : 1.00 Min. : 533 Min. :0.0 Min. :1   
## 1st Qu.: 5.75 1st Qu.: 5.75 1st Qu.:2352 1st Qu.:0.0 1st Qu.:1   
## Median :10.50 Median :10.50 Median :2870 Median :0.5 Median :1   
## Mean :10.50 Mean :10.50 Mean :2732 Mean :0.5 Mean :1   
## 3rd Qu.:15.25 3rd Qu.:15.25 3rd Qu.:3232 3rd Qu.:1.0 3rd Qu.:1   
## Max. :20.00 Max. :20.00 Max. :4470 Max. :1.0 Max. :1   
## MD\_at\_birth any\_drink\_on\_avg mother\_foreign\_born mother\_age   
## Min. :0.00 Min. :0.00 Min. :0.00 Min. :19.00   
## 1st Qu.:0.75 1st Qu.:0.00 1st Qu.:0.00 1st Qu.:23.50   
## Median :1.00 Median :0.00 Median :0.00 Median :26.50   
## Mean :0.75 Mean :0.05 Mean :0.05 Mean :26.55   
## 3rd Qu.:1.00 3rd Qu.:0.00 3rd Qu.:0.00 3rd Qu.:30.25   
## Max. :1.00 Max. :1.00 Max. :1.00 Max. :36.00   
## father\_age num\_prenatal\_visits mother\_hispanic\_other mother\_black   
## Min. :20.00 Min. : 2.00 Min. :0 Min. :0.00   
## 1st Qu.:26.00 1st Qu.: 9.75 1st Qu.:0 1st Qu.:0.00   
## Median :27.00 Median :11.00 Median :0 Median :0.00   
## Mean :28.95 Mean :11.35 Mean :0 Mean :0.05   
## 3rd Qu.:32.00 3rd Qu.:13.25 3rd Qu.:0 3rd Qu.:0.00   
## Max. :37.00 Max. :20.00 Max. :0 Max. :1.00   
## father\_hispanic\_other father\_black mother\_educ\_less\_than\_hs mother\_educ\_hs  
## Min. :0 Min. :0.00 Min. :0.0 Min. :0.0   
## 1st Qu.:0 1st Qu.:0.00 1st Qu.:0.0 1st Qu.:0.0   
## Median :0 Median :0.00 Median :0.0 Median :0.5   
## Mean :0 Mean :0.05 Mean :0.3 Mean :0.5   
## 3rd Qu.:0 3rd Qu.:0.00 3rd Qu.:1.0 3rd Qu.:1.0   
## Max. :0 Max. :1.00 Max. :1.0 Max. :1.0   
## father\_educ\_less\_than\_hs father\_educ\_hs adequate\_care birth\_month   
## Min. :0.0 Min. :0.00 Length:20 Min. : 1.00   
## 1st Qu.:0.0 1st Qu.:0.00 Class :character 1st Qu.: 4.00   
## Median :0.0 Median :0.00 Mode :character Median : 6.00   
## Mean :0.2 Mean :0.45 Mean : 6.05   
## 3rd Qu.:0.0 3rd Qu.:1.00 3rd Qu.: 7.25   
## Max. :1.0 Max. :1.00 Max. :12.00   
## birth\_weekday child\_female mother\_anemic mother\_cardiac\_disease  
## Min. :1.00 Min. :0.00 Min. :0 Min. :0   
## 1st Qu.:3.00 1st Qu.:0.00 1st Qu.:0 1st Qu.:0   
## Median :5.00 Median :1.00 Median :0 Median :0   
## Mean :4.35 Mean :0.55 Mean :0 Mean :0   
## 3rd Qu.:6.00 3rd Qu.:1.00 3rd Qu.:0 3rd Qu.:0   
## Max. :7.00 Max. :1.00 Max. :0 Max. :0   
## mother\_lung mother\_herpes mother\_diabetes low\_weight mother\_age\_cat   
## Min. :0.00 Min. :0 Min. :0 Min. :0.00 Length:20   
## 1st Qu.:0.00 1st Qu.:0 1st Qu.:0 1st Qu.:0.00 Class :character   
## Median :0.00 Median :0 Median :0 Median :0.00 Mode :character   
## Mean :0.05 Mean :0 Mean :0 Mean :0.45   
## 3rd Qu.:0.00 3rd Qu.:0 3rd Qu.:0 3rd Qu.:1.00   
## Max. :1.00 Max. :0 Max. :0 Max. :1.00

summary(lbw\_example$child\_birthwt)

## Min. 1st Qu. Median Mean 3rd Qu. Max.   
## 533 2352 2870 2732 3232 4470

options(qwraps2\_markup = 'markdown')  
  
# qwraps2 requires a list of lists to make the table, like this:  
summary.stats1 <-   
 list("Birthweight [g]" =   
 list("Min" = ~round(min(child\_birthwt)),  
 "Median" = ~round(median(child\_birthwt)),  
 # mean\_sd comes from qwraps2  
 "Mean" = ~qwraps2::mean\_sd(child\_birthwt, digits = 0),  
 "Max" = ~round(max(child\_birthwt))))  
  
qwraps2::summary\_table(lbw\_example, summary.stats1)

|  | lbw\_example (N = 20) |
| --- | --- |
| **Birthweight [g]** |  |
| Min | 533 |
| Median | 2870 |
| Mean | 2,732 ± 834 |
| Max | 4470 |

# can also report broken down by group:   
summary\_table(lbw\_example, summary.stats1, by = "use\_tobacco")

|  | 0 (N = 10) | 1 (N = 10) |
| --- | --- | --- |
| **Birthweight [g]** |  |  |
| Min | 1752 | 533 |
| Median | 3052 | 2410 |
| Mean | 3,056 ± 722 | 2,408 ± 845 |
| Max | 4470 | 3515 |

# Overview of non-parametric and parametric testing

You may already be familiar with a lot of this, so hopefully this is a good recap if so.

Nonparametric tests put few assumptions, if any, on the underlying population that a sample is drawn from. In contrast, parametric tests use distributional assumptions. However, the general approach to using either set of tests is the same:

* Specify null () and alternative hypotheses ()
* Set the level of significance (commonly )
* Calculate a test statistic
* Based on the value of the test statistic, either reject or do not reject

A flowchart with suggestions for which tests to use is available at the end of this file.

# Examples of common parametric tests and their counterparts

You’re probably familiar with the t-test. The two-sample t-test is used to compare sample means. Two-sample t-tests can be paired or unpaired, depending on whether the observations in each sample have a corresponding observation in the other sample. For unpaired tests, you assume that the observations in each group were independently sampled.

For a two-sample t-test, our null hypothesis is (typically) that the means are equal and the alternative hypothesis is that the means are not equal:

$$
H\_0: \mu\_1 = \mu\_2 \\
H\_A: \mu\_1 \ne \mu\_2
$$

We’re able to reject if the difference in the sample means, , is far enough from 0. The t-statistic is used to make that decision:

is the pooled estimate of the variance, which combines information from each sample. This assumes that the variances are equal, although there are versions of the t-test available in your programming software assume unequal variances. Ultimately, the t-statistic calculated above is compared to a distribution to find a -value. The degrees of freedom (and width of the tails) of the distribution depend on and . Assuming the null hypothesis is true, this is the probability of observing a difference in means as large or larger than . If , the null hypothesis is rejected; otherwise, you fail to reject . Your software will likely do all of these steps on its own.

This t-test also assumes that the means in each of the populations are normally distributed. This assumption can be hard to meet if the sample sizes are small (and especially if small and non-normal). (In such situations, the mean of the sample may also not be very informative.) For a situation like this, a non-parametric test like the Wilcoxon rank sum test (AKA Mann-Whitney U test) may be helpful. For continuous data, the null hypothesis of this test is that the median of each population is equal and the alternative is that the medians are not equal. The test does the following:

* Combine the data from the two groups and rank all the observations from smallest to largest
  + Tied values are assigned an average rank
* Calculate , which the sum of the ranks corresponding to group 1

Under the null hypothesis (), we would expect the sum of the ranks in each group to be about equal. If is too small or too large than what we would expect under , we have evidence to reject . Again, your software does all of these calculations. (Note that when ties occur, you may have to specify how the p-values should be calculated.)

Unlike the t-test, the Wilcoxon rank sum test can be extended to ordinal data. Because the Wilcoxon rank sum test only uses ranks of the observations, it is less sensitive to outliers, especially when the sample size is small. One last note: it’s important to recognize that these tests have different null and alternative hypotheses.

Now here is an example of each of these tests with the test dataset. We’ll compare birthweights among children born to mothers who used/did not use tobacco.

# the stats library has both the t-test and wilcoxon tests  
t.test(formula = child\_birthwt ~ use\_tobacco,   
 data = lbw\_example)

##   
## Welch Two Sample t-test  
##   
## data: child\_birthwt by use\_tobacco  
## t = 1.8454, df = 17.574, p-value = 0.0819  
## alternative hypothesis: true difference in means between group 0 and group 1 is not equal to 0  
## 95 percent confidence interval:  
## -91.11146 1388.31146  
## sample estimates:  
## mean in group 0 mean in group 1   
## 3056.3 2407.7

# by default, this is a unpaired two-sided t-test with unequal variance  
  
example.t <- t.test(formula = child\_birthwt ~ use\_tobacco,   
 data = lbw\_example)  
  
# this function from broom cleans up the output of statistical objects  
example.t <- tidy(example.t)  
  
# this is equivalent to the above  
t.test(x = lbw\_example$child\_birthwt[lbw\_example$use\_tobacco==0],   
 y = lbw\_example$child\_birthwt[lbw\_example$use\_tobacco==1],   
 data = lbw\_example)

##   
## Welch Two Sample t-test  
##   
## data: lbw\_example$child\_birthwt[lbw\_example$use\_tobacco == 0] and lbw\_example$child\_birthwt[lbw\_example$use\_tobacco == 1]  
## t = 1.8454, df = 17.574, p-value = 0.0819  
## alternative hypothesis: true difference in means is not equal to 0  
## 95 percent confidence interval:  
## -91.11146 1388.31146  
## sample estimates:  
## mean of x mean of y   
## 3056.3 2407.7

wilcox.test(formula = child\_birthwt ~ use\_tobacco,   
 data = lbw\_example, exact = FALSE)

##   
## Wilcoxon rank sum test with continuity correction  
##   
## data: child\_birthwt by use\_tobacco  
## W = 70, p-value = 0.1402  
## alternative hypothesis: true location shift is not equal to 0

# by default, this is a unpaired two-sided test  
  
example.wilcox <- wilcox.test(formula = child\_birthwt ~ use\_tobacco,   
 data = lbw\_example, exact = FALSE)  
  
tidy(example.wilcox)

## # A tibble: 1 × 4  
## statistic p.value method alternative  
## <dbl> <dbl> <chr> <chr>   
## 1 70 0.140 Wilcoxon rank sum test with continuity correcti… two.sided

# Additional materials

## More than 2 groups

Below is also an example of an analysis of variance (ANOVA) test that compares birthweights between several categories (based on mothers’ ages). ANOVA (see diagram) is useful when comparing 3 or more groups. A non-parametric counterpart is the Kruskall-Wallis test, which also compares ranks between each group.

When comparing several groups, you may also be interested in pairwise comparisons. In this case, it’s advisable to adjust for having done multiple comparisons to keep the (type I) error rate across all tests at less than or equal to . A common method is to use a “Bonferroni correction”, which just divides by the number of tests that will be conducted. (Note: This can be overly conservative if you do a large number of tests or if the outcomes you’re investigating are correlated.)

Another option is [Tukey’s honest significant differences](https://personal.utdallas.edu/~herve/abdi-HSD2010-pretty.pdf) method, which is commonly done after conducting an ANOVA. An example using this (with tukeyHSD) is shown below. Based on my understanding, this method requires a larger difference in the mean (compared to a t-test) to reject the null hypothesis.

lbw\_example <- lbw\_example %>%  
mutate(mother\_age\_cat = case\_when(mother\_age >= min(mother\_age) & mother\_age < 25 ~ "<25",  
 mother\_age >= 25 & mother\_age < 30 ~ "25-29",  
 mother\_age >= 30 ~ "30+")  
 )  
  
bw\_aov <- aov(child\_birthwt ~ mother\_age\_cat, data = lbw\_example)  
  
# fail to reject H0 that the mean birthweights are equal  
summary(bw\_aov)

## Df Sum Sq Mean Sq F value Pr(>F)  
## mother\_age\_cat 2 1064357 532179 0.744 0.49  
## Residuals 17 12157277 715134

# same here  
kruskal.test(child\_birthwt ~ mother\_age\_cat, data = lbw\_example)

##   
## Kruskal-Wallis rank sum test  
##   
## data: child\_birthwt by mother\_age\_cat  
## Kruskal-Wallis chi-squared = 0.7249, df = 2, p-value = 0.696

# here is the same, but with a larger version of the birthweight dataset  
lbw\_data <- lbw\_data %>%  
mutate(mother\_age\_cat = case\_when(mother\_age >= min(mother\_age) & mother\_age < 25 ~ "<25",  
 mother\_age >= 25 & mother\_age < 30 ~ "25-29",  
 mother\_age >= 30 ~ "30+")  
 )  
  
bw\_aov <- aov(child\_birthwt ~ mother\_age\_cat, data = lbw\_data)  
  
# here, we reject H0 that the mean birthweights are equal  
summary(bw\_aov)

## Df Sum Sq Mean Sq F value Pr(>F)   
## mother\_age\_cat 2 3.771e+07 18855325 24.67 2.06e-11 \*\*\*  
## Residuals 9997 7.642e+09 764382   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1

# same here  
kruskal.test(child\_birthwt ~ mother\_age\_cat, data = lbw\_data)

##   
## Kruskal-Wallis rank sum test  
##   
## data: child\_birthwt by mother\_age\_cat  
## Kruskal-Wallis chi-squared = 48.044, df = 2, p-value = 3.694e-11

# when we can reject H0, we may want to conduct pairwise comparisons (t-test or KW test)  
# or using Tukey HSD  
pairwise\_lbw <- tidy( TukeyHSD(bw\_aov) ) # tidy cleans output into a dataframe  
  
# View(pairwise\_lbw)

## Demonstrating the Central Limit Theorem

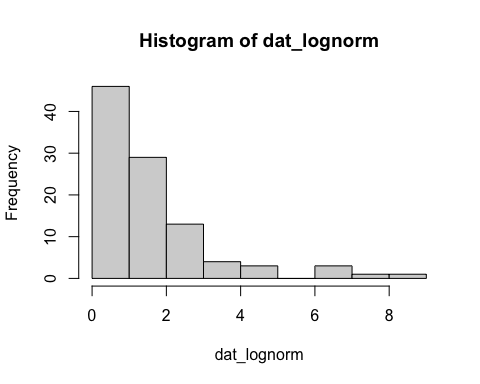
The example below is meant to show some of the power of the Central Limit Theorem (CLT). The CLT states that, as the sample sizes increases, the distribution of sample averages generated by random draws from a population becomes increasingly normal with a mean equal to the population mean and variance proportional to the sample size. This holds even if the underlying populations are not normal.

Mathematically, this looks like the following, as increases:

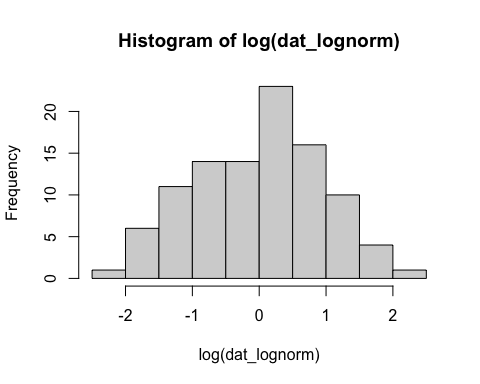
where is the sample mean, is the population mean, and is the variance of the sample mean.

The example below demonstrates this for sample draws from a lognormal distribution.

# draw 100 observations from lognormally distributed population  
dat\_lognorm <- exp(rnorm(n = 100, mean = 0, sd = 1))  
  
# this is the geometric mean, which should be close to exp(0) = 1  
# exp(mean(log(dat\_lognorm)))  
  
# verify that this looks lognormal  
hist(dat\_lognorm)



hist( log(dat\_lognorm) )



Now let’s repeatedly draw samples from a log-normally distributed population. In each sample, we’ll calculate the (sample) average and save it for later. Then we’ll see the distribution of these averages across all of our draws from the population.

For fun, you can try changing the distribution of the underlying population or the parameter that is calculated for each sample (e.g., geometric mean, median, etc). The CLT also supports that the difference in sample means is also approximately normally distributed, even if the underlying population distributions are not normal. This is why the t-test can still be okay to use if the sample size is large.

set.seed(12042023)  
  
one.sample\_lognorm.f <- function(n.obs = NULL) {  
   
 one.samp <- exp(rnorm(n = n.obs, mean = 0, sd = 1))  
   
 one.sample.mean <- mean(one.samp)  
   
 return(one.sample.mean)  
  
}  
  
# check that it works  
one.sample\_lognorm.f(n.obs = 1)  
  
# now we'll do 10k draws, increasing the sample size each time  
 # one observation in each sample  
 sample.means <- replicate(n = 10000, one.sample\_lognorm.f(n.obs = 1))  
   
 hist(sample.means, breaks = 50)  
   
 sample.means <- replicate(n = 10000, one.sample\_lognorm.f(n.obs = 10))  
   
 hist(sample.means, breaks = 50)  
   
 sample.means <- replicate(n = 10000, one.sample\_lognorm.f(n.obs = 30))  
   
 hist(sample.means, breaks = 50)  
   
 sample.means <- replicate(n = 10000, one.sample\_lognorm.f(n.obs = 100))  
   
 hist(sample.means, breaks = 50)  
   
 # all the way up to sample sizes of 1000  
 sample.means <- replicate(n = 10000, one.sample\_lognorm.f(n.obs = 1000))  
   
 hist(sample.means, breaks = 50)

Regression methods are not covered in this doc, but the CLT is especially powerful because it is also often used to construct confidence intervals for regression coefficients.

# Hypothesis testing flowchart

