# *Chapter 8: Tests of Hypotheses Based on a Single Sample*

*Curtis Miller*

*2019-01-02*

## *Introduction*

Statistics involes more than parameter estimation. We may not care about the actual value of the parameter but rather whether the parameter is a particular value or within some range. In this case we may prefer to perform a statistical hypothesis test rather than construct a confidence interval.

In this chapter we see for the first time statistical hypothesis testing, involving only a single sample. The hypotheses of interest will typically be making a statement about the value of a parameter, though other hypothesis tests make more general statements. The fundamental principles, though, are the same, along with the general format of a test.

Hypothesis testing is a popular procedure, which suggests it's also frequently abused. We should always remember that hypothesis testing is part of our toolset for reaching conclusions about a phenomenon using a dataset; it is not the *only* tool that should be used We should supplement hypothesis testing with other procedures, such as visualization and providing point estimates. Furthermore, we should be *honest* when collecting our data and be sure we are not "coercing" the dataset to get an answer we want.[1]

[1] As Nobel Prize winning economist Robert Coase said, "If you torture the data enough, nature will always confess."

## *Section 1: Hypotheses and Test Procedures*

A **statistical hypothesis** is a statement about the probabilistic properties of a data-generating process.[2] A **test of hypotheses** is a procedure where sample data is used to decide which of two competing hypotheses better describes the process that generated the data. The **null hypothesis** (usually denoted $H_0$) can be thought of as the current assumption about the data[3], while the **alternative hypothesis** (usually denoted $H_A$) is the assumption that will replace the null hypothesis if we reject the null hypothesis. If we don't reject $H_0$, we do not say that we accept $H_0$ but rather that we failed to reject $H_0$.

[2] This is usually a statement about a parameter, a collection of parameters, or even whether the data follows some distribution.

[3] Usually $H_0$ is the statement we seek to disprove, but this is not always the case; for example, tests for distribution, which intend to determine if the data follows a particular distribution, will often state that under the null hypothesis the data follows the distribution of interest.

Hypothesis testing is a form of *reductio ad absurdum* ("argument to absurdity"), similar to a proof by contradiction; the argument is that by assuming the null hypothesis is true we see a result in the data that is "absurd", so we should surrender our belief in $H_0$. If this

"absurdity" in the data does not appear, though, that does not mean $H_0$ is true; it just means we could not show it is false, or that $H_A$ is more correct.[4]

In this chapter we consider tests that make statements about a population parameter $\theta$. These tests almost always take the following form in practice:

We call $\theta_0$ the **null value** for $\theta$, and it is the assumed value of $\theta$ under $H_0$.[5]

Statistical tests (of any form) follow the procedure described below:

1. Identify $H_0$ and $H_A$.
2. Specify a number $\alpha \in (0,1)$, usually small (typical $\alpha$ are $\alpha \in \{0.1, 0.05, 0.01, 0.001\}$; there is an interpretation of $\alpha$ I will explain later that can guide this decision). This is called the **significance level** of the test.
3. Collect data and compute the test statistic; call the random version of the test statistic $T$ for now, and let the observed (computed) value of $T$ be $\hat{T}$. If $H_0$ is true, the distribution of $T$ is known.
4. Compute a quantity known at the **p-value**, denoted here $p_{\text{val}}$[6]. The definition of $p_{\text{val}}$ in general is[7]

$$p_{\text{val}} = \mathbb{P}\left(\text{Observe } T \text{ more contradictory to } H_0 \text{ than } \hat{T}\right)$$

5. If $p_{\text{val}} < \alpha$, reject $H_0$; otherwise ($p_{\text{val}} \geq \alpha$) do not reject $H_0$. (Because of this rule, $p_{\text{val}}$ is sometimes referred to as the **observed significance level** of the test, as it is the smallest $\alpha$ at which you would reject $H_0$.)
6. Conclude and interpret the results of the test.

Be clear that $p_{\text{val}}$ is the probability of observing a test statistic at least as contradictory to $H_0$ as the observed test statistic. If we were to say that large $T$ are evidence against $H_0$ (with larger $T$ meaning even more evidence against $H_0$), then $p_{\text{val}} = \mathbb{P}\left(T > \hat{T}\right)$; that is, it is the probability of seeing even more contradictory evidence than what was seen.

[4] The usual comparison is the ancient Roman legal principle (still in use in America) of "innocent until proven guilty"; we assume that the individual on trial is innocent (i.e. $H_0$ is "true"), and the burden of proof lies on the prosecution (the statistical test) to show that this assumption is "absurd" based on the evidence (the data) and we should then assume the individual is guilty ($H_0$ is "false", and $H_A$ better describes reality). Failure to prove guilt, though, does not imply innocence, and the "beyond reasonable doubt" criteria sets a high bar for proving guilt. It tilts justice in favor of letting guilty people go (a Type II error) as opposed to using the state's resources to punish the innocent (a Type I error).

[5] We amost never see $H_0$ of the form

$$H_0 : \theta \leq \theta_0$$

or

$$H_0 : \theta \geq \theta_0$$

This is because the statistical test does not change if we replace the inequality with equality. The border value $\theta_0$ is the most difficult case to check, and it can be shown that if we reject $H_0$ at the border we can safely reject for all other potential values of $\theta$, while if we could not reject $H_0$ when assuming $\theta = \theta_0$ we should reject $H_0$ at all. Consequently we can view $H_0$ as actually making a statement about all possible $\theta$ within a region when $H_A$ is one-sided.

[6] The usual notation for p-values is simply $p$, but we will run into situations in this class where the letter $p$ appears in many places, so I use this notation to keep all these different $p$'s straight.

[7] The classical approach to statistical testing does not involve p-values but instead a critical value, $T_0$, and if $\hat{T} > T_0$, $H_0$ would be rejected. This theory still underlies statistics; power/Type II error analysis and the formulas for computing p-values are derived with this theory in mind. However, there are advantages to referring to p-values. One is that software usually computes p-values. Another is that p-values have a universal interpretation; given any p-value you can decide whether to reject $H_0$ or not even if you don't know the context of the test. Additionally, readers can decide whether a reported p-value is convincing for themselves personally, regardless of what the authors of the study write. (Unfortunately, though, authors often don't write p-values but instead will write $p < 0.05$, which partially defeats the purpose of p-values.)

The following are *incorrect* interpretations of $p_{\text{val}}$:

- The probability $H_0$ is true or false.
- The probability the conclusion of the test is due to random chance alone.

Additionally, practitioners should not fret over exactly what threshold a p-value passes (such as whether $p_{\text{val}} < 0.05$). While (5) in the above description of the statistical testing procedure suggests that certain p-value imply certain conclusions, p-values are more useful when considered as a measure of how strong the evidence against $H_0$ is.[8]

In hypothesis testing, there are two types of errors. A **Type I error** is rejecting $H_0$ when $H_0$ is true, while a **Type II error** is failing to reject $H_0$ when $H_0$ is false. The table below visualizes the relationship:

Immediately after a test, you do not know whether you committed an error or what the nature of the error is. Error analysis is part of study design, conducted before any data is collected. It determines what must be observed to reject $H_0$ and what sample size the study should use. There should be a discussion about what happens when a Type I or Type II error is made, what the consequences are, the relative severity of the consequences, and thus what the acceptable error rates should be.

$\alpha$ is the Type I error rate:[9]

In this context, the Type II error rate depends on what the true value of $\theta$ is; we call $\beta(\theta_A)$ the Type II error rate when $\theta = \theta_A$:

A related concept to the Type II error rate is the **power** of the statistical test; the power is the probability of rejecting $H_0$ when $\theta = \theta_A$. It is defined below:

[8] People have identified as one of the culprits of the so-called reproducibility crisis (many results in published scientific papers cannot be reproduced), and they are frequently abused. The problem has gotten severe enough that in 2006 the American Statistical Association (ASA) issued a statement about the appropriate use and interpretation of p-values [Wasserstein and Lazar, 2016]. However, the problems associated with p-values can be pinned (more fairly) on publishing practices and how publication decisions are made. Journals are biased to "positive results" (i.e. when $H_0$ is rejected) and have given $\alpha = 0.05$ unreasonable importance. This can lead to malicious practices such as p-hacking (rephrasing a statistical problem until "statistically significant" results are found), or ignoring the size of the effect found in the paper. See Aschwanden [2015] and Aschwanden [2016] for interesting discussions and even interactive demonstrations of these issues.

[9] Actually this is the case when the test statistic is a continuous variable. For discrete variables, we may choose a desired $\alpha$ but due to the discrete nature of the cdf the actual Type I error rate may be less than specified (when being conservative). We see this in Example 1.

Power relates to $\alpha$ and $\beta$ in the following way:

There is a general relationship between $\alpha$ and $\beta$:

After we have reflected on the consequences of Type I and Type II errors, we may decide on an acceptable Type I error rate, $\alpha$. We then focus on a particular $\theta_A$ we want our test to be able to detect and what the acceptable Type II error rate $\beta(\theta_A)$ should be. Once we have these pieces of information, we may find a sample size $n$ that achieves these two error rates for our test.

---

*Example 1*

I claim that I am an 80% freethrow shooter, but you don't believe me; you think I make less than 80% of freethrows. To settle the dispute, we agree that I will shoot 20 free-throws and you will count how many baskets I manage to make. Based on this you will decide whether you believe my claim. You decide to use $\alpha = 0.05$ as your significance level.

1. Identify $H_0$ and $H_A$.

2. What is the test statistic? What is its distribution under $H_0$?

3. Out of 20 baskets, I manage to make 11. Compute $p_{\text{val}}$.

4. What is the conclusion of the test?

5. Let $N_\alpha$ denote the fewest number of baskets I could make while still allowing you to believe my claim when you use significance level $\alpha$ (that is, if $X \sim \text{BIN}(20, 0.8)$, $N_\alpha$ is the largest number such that $\mathbb{P}(X < N_\alpha) \leq \alpha$). Find $N_{0.05}$.[10]

[10] (5) and on are questions we would ask before we observed any data and reached a conclusion.

6. While $\alpha = 0.05$ is the specified Type I error rate, due to the discrete nature of the test statistic, it is not the actual Type I error rate. What is the actual Type I error rate?

7. Suppose I were not an 80% freethrow shooter and instead only make 75% of my baskets. What is the Type II error rate in this case?

```r
pbinom(11, 20, .8)  # Part 3
```

```
## [1] 0.009981786
```

```r
(N <- qbinom(.05, 20, .8) - 1)  # Part 5
```

```
## [1] 12
```

```r
pbinom(N, 20, .8)  # Part 6
```

```
## [1] 0.03214266
```

```r
pbinom(N, 20, .75, lower.tail = FALSE)
```

```
## [1] 0.8981881
```

*Example 2*

Let $\mu$ denote the population mean. We wish to determine if the true population mean is greater than the specified value $\mu_0$.

1. State the null and alternative hypothesis.

2. We collect a dataset $X_1, \ldots, X_n$ from the population, with $\mathbb{E}[X] = \mu$, and $\text{SD}(X) = \sigma$. Consider the test statistic

$$Z = \frac{\bar{X} - \mu_0}{\sigma / \sqrt{n}}$$

According to the central limit theorem, what is the approximate distribution of $Z$ under $H_0$?

3. What is the (approximate) distribution when $H_A$ is true and $\mu = \mu_A > \mu_0$?

4. Let $\alpha$ be the level of significance of the test. We will reject $H_0$ when $Z$ is larger than some threshold value. Find this threshold value such that the Type I error rate is $\alpha$.

5. Compute $\beta(\mu_A)$. This is the probability of *not* rejecting $H_0$ when $\mu = \mu_A$; in other words, $Z$ is less than the threshold value even though $\mu = \mu_A > \mu_0$.

6. Given the answers to (4) and (5), find a sample size $n$ such that a test with Type I error rate $\alpha$ will have Type II error rate $\beta$ when $\mu = \mu_A$.

7. Let's now suppose that we are investigating whether men's average height is 5.9 ft., and under the alternative hypothesis men are taller than 5.9 ft. Phrase $H_0$ and $H_A$ below.

8. Let our significance level be $\alpha = 0.1$. The standard deviation of height is known to be $\sigma = 0.5$. Suppose the true mean height for men is 6 ft. What is the Type II error rate when $n = 100$? Repeat for a potential mean height of 6.5 ft.

9. Find the sample size that, for a test with $\alpha = 0.1$, would have a Type II error rate of $\beta = 0.1$ when the true average height is 6 ft.

10. A sample mean height of 5.97 ft. is observed, and the sample size is the one found in part (9) above. Compute $p_{\text{val}}$.

11. Based on this data, what is the conclusion of the test?

```r
alpha <- 0.1
beta <- 0.1
sigma <- 0.5
mu0 <- 5.9
muA <- 6.0
xbar <- 5.97
(za <- qnorm(alpha, lower.tail = FALSE))
```

```
## [1] 1.281552
```

```r
(zb <- qnorm(beta, lower.tail = FALSE))
```

```
## [1] 1.281552
```

```r
pnorm(za + (mu0 - muA)/(sigma/sqrt(100)))  # Part 8
```

```
## [1] 0.2362404
```

```r
pnorm(za + (mu0 - 6.5)/(sigma/sqrt(100)))
```

```
## [1] 4.169523e-27
```

```r
(n <- ceiling((sigma * (za + zb) / (mu0 - muA))^2))  # Part 9
```

```
## [1] 165
```

```r
(z <- (xbar - mu0)/(sigma/sqrt(n)))  # Part 10
```

```
## [1] 1.798333
```

```r
(pval <- pnorm(z, lower.tail = FALSE))
```

```
## [1] 0.03606216
```

```r
(pval < alpha)  # Part 11
```

```
## [1] TRUE
```

---

## *Section 2: z Tests for Hypotheses about a Population Mean*

From here, in order to perform a hypothesis test, we only need the following bits of information:

- The null hypothesis $H_0$, and potential $H_A$
- Assumptions about the data made by the test
- The test statistic and how to compute it
- How to compute $p_{\text{val}}$ based on the test statistic

Our first case is a test for the mean $\mu$ when $\sigma$ is known. This test is exact when the data was drawn from a Normal distribution, and asymptotically correct when the data is not Normally distributed.

Suppose $\sigma$ is not known. If $n$ is large[11], we can replace $\sigma$ with the sample standard deviation $S$ and thus use the test statistic

[11] Let's say $n > 40$.

The test is otherwise the same.

Below are formulas for computing Type II errors. If $\sigma$ is not known, you will need to guess it.

The formulas below allow for sample size planning. Overestimating $\sigma$ will produce large $n$ and thus produce tests that may do better than specified.

---

*Example 3*

A factory that produces ball bearings is testing its assembly line to see whether the line produces ball bearings with the specified diameter of 10 mm or whether the line is not properly calibrated. The managers believe that the standard deviation of bearings produced by this line is $\sigma = 0.1$ mm. They want tests that are significant at the $\alpha = 0.01$ significance level.

1. State the null and alternative hypothesis.

2. What is the probability of a Type I error?

3. What is the probability of a Type II error when the ball bearings' mean diameter differ from the specified diameter by 0.1 mm and the sample size is $n = 20$?

4. The managers want to be able to detect a difference of 0.1 mm from the specified diameter with probability 0.9995. Find a sample size that guarantees this under our assumptions.

5. Using the sample size $n = 41$, experimenters run the line and produce some ball bearings. The following sample was observed:

```
bearings <- c(10.11, 9.858, 10.072, 10.007, 10.158, 9.878, 9.935, 9.787,
              9.993, 10.008, 9.927, 9.959, 10.086, 10.001, 9.881, 10.057, 9.913,
              9.744, 10.136, 10, 9.988, 10.022, 10.112, 10.013, 9.809, 10.014,
              10.036, 9.977, 9.952, 9.963, 9.955, 9.926, 10.095, 10.076, 9.994,
              9.93, 10.057, 9.923, 9.954, 9.969, 10.124)
mean(bearings)
```

```
## [1] 9.985341
```

```
sd(bearings)
```

```
## [1] 0.09381434
```

Using the sample standard deviation rather than $\sigma$, perform the appropriate statistical test to decide between $H_0$ and $H_A$, computing $p_{\text{val}}$.

6. Conclude.

```r
suppressPackageStartupMessages(library(BSDA))
alpha <- 0.01
beta <- 0.0005
sigma <- 0.1
mu0 <- 10
muA <- 10.1  # We could also use 9.9 and be fine
(za2 <- qnorm(alpha/2, lower.tail = FALSE))
```

```
## [1] 2.575829
```

```r
(zb <- qnorm(beta, lower.tail = FALSE))
```

```
## [1] 3.290527
```

```r
pnorm(za2 + (mu0 - muA)/(sigma/sqrt(10))) +
  pnorm(-za2 + (mu0 - muA)/(sigma/sqrt(10)))  # Part 3
```

```
## [1] 0.2787871
```

```r
(n <- ceiling((sigma * (za2 + zb)/(mu0 - muA))^2))  # Part 4
```

```
## [1] 35
```

```r
z.test(bearings, mu = 10, sigma.x = sd(bearings))  # Part 5
```

```
##
##  One-sample z-Test
##
## data:  bearings
## z = -1.0005, p-value = 0.3171
## alternative hypothesis: true mean is not equal to 10
## 95 percent confidence interval:
##   9.956625 10.014058
## sample estimates:
## mean of x
##  9.985341
```

---

## *Section 3: The One-Sample t Test*

If we assume our data follows a Normal distribution, then the distribution of

$$T = \frac{\bar{X} - \mu_0}{S/\sqrt{n}}$$

is $t(n-1)$ when $H_0$ is true. Based on this we can describe a test based on the $t$ distribution.

This test works better than the test described in the previous section when the data follows a Normal distribution, and the difference is noticeable for small $n$.[12]

Table A.5 isn't well suited for hypothesis testing; instead, use Table A.8.

---

*Example 4*

Repeat the test performed in Example 3 but using the $t$-test instead. Does your conclusion change?

[12] What's the difference between these two tests? What is the penalty for using the $z$-test rather than the $t$-test for Normally distributed data? Notice that $z_\alpha < t_{\alpha,n-1}$ for all $n$. Since the random variable $T$ follows the $t(n-1)$ distribution, we can conclude that when we use the $z$-test instead of the $t$-test, $p_{\text{val}}$ will be inappropriately small, and thus we are more likely to reject the null hypothesis. The true Type I error rate is *greater* than $\alpha$! This phenomenon is known as **size inflation**. When $n$ is large the inflation is negligible, but for small $n$ it could be a problem.

```
t.test(bearings, mu = 10)
```

```
##
##  One Sample t-test
##
## data:  bearings
## t = -1.0005, df = 40, p-value = 0.3231
## alternative hypothesis: true mean is not equal to 10
## 95 percent confidence interval:
##   9.95573 10.01495
## sample estimates:
## mean of x
##  9.985341
```

---

Type II error analysis (including sample size planning) is more complicated for $t$-testing, and we do not have clean formulas like we did when $\sigma$ was known. We either need to use software or graphs like those provided in Table A.17. When using software, the power $\pi(\mu_A)$ is usually referred to rather than $\beta(\mu_A)$, and the input is usually not $\mu_A$ but $d = (\mu_0 - \mu_A)/\sigma$. (Table A.17 also uses $d$.) Notice that a guess of $\sigma$ needs to be made.

---

*Example 5*

Use Table A.17 to answer the following:

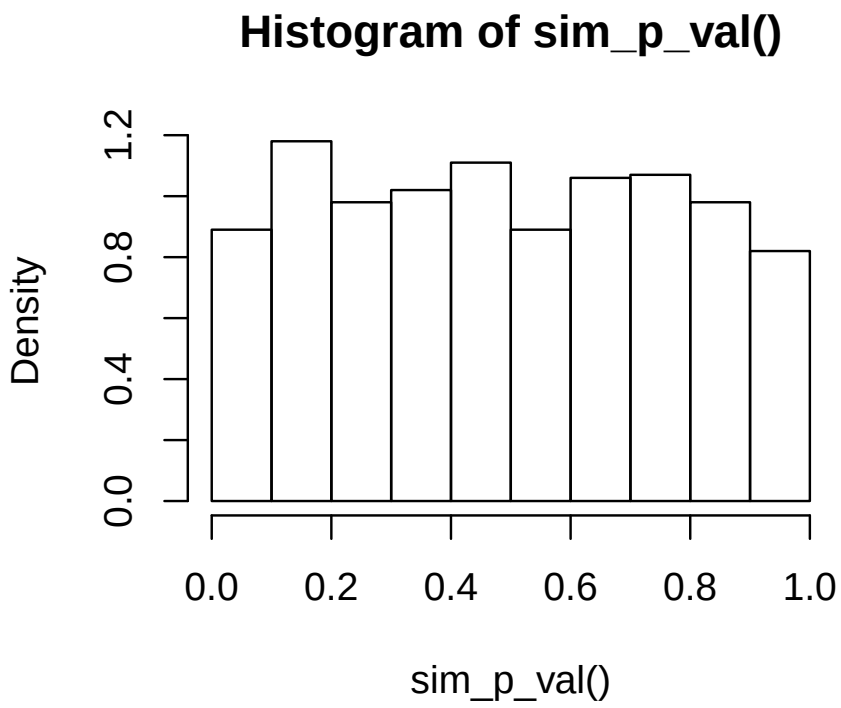
1. For a one-tailed $t$-test, what is the probability of a Type II error when the degrees of freedom is $\nu = 9$ and $|\Delta| = 0.6$? Repeat with $\nu = 29$.

2. For a two-tailed test, what sample size is needed so that a test will have a Type II error rate of 0.1 when $|\Delta| = 0.5$? Choose the smallest listed degrees of freedom.

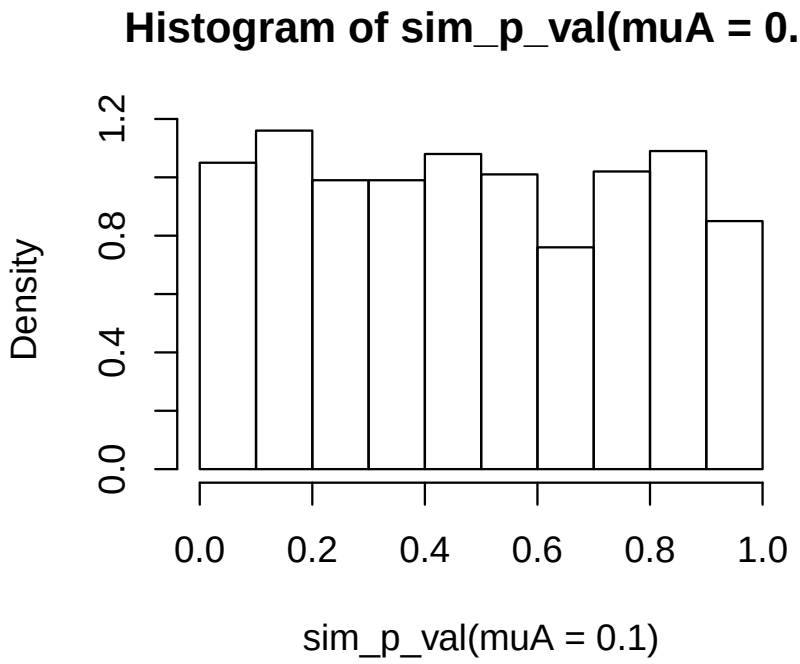
Something to consider when talking about $p_{\text{val}}$: this number is a statistic like any other quantity we compute from data, and thus it has a sampling distribution. Under the null hypothesis, if the assumptions of the $t$-test are met, then it can be shown that $p_{\text{val}} \sim \text{UNIF}(0,1)$. Under the alternative hypothesis, though, $p_{\text{val}}$ follows a distribution other than $\text{UNIF}(0,1)$, and the sampling distribution concentrates near 0 as $n$ grows or as $\Delta$ grows.

Below I simulate the distribution of $p_{\text{val}}$ in different scenarios.

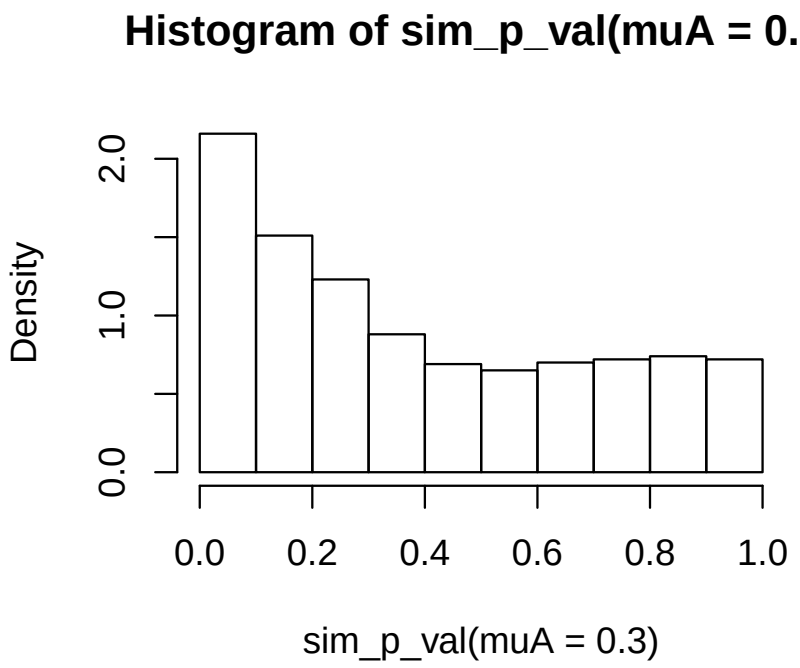
```
# I write a function to perform these simulations
sim_p_val <- function(M = 1000,    # Number of replications
                      mu0 = 0,     # Hypothesized mean
                      muA = NULL,  # True mean; if null, same as mu0
                      n = 10,
                      sd = 1,
                      alternative = c("two.sided", "less", "greater")) {
  if (is.null(muA)) {
    muA <- mu0
  }
  alternative <- alternative[1]

  replicate(M, {
    dat <- rnorm(n, mean = muA, sd = sd)
    return(t.test(dat, alternative = alternative, mu = mu0)$p.value)
  })
}

hist(sim_p_val(), freq = FALSE)
```



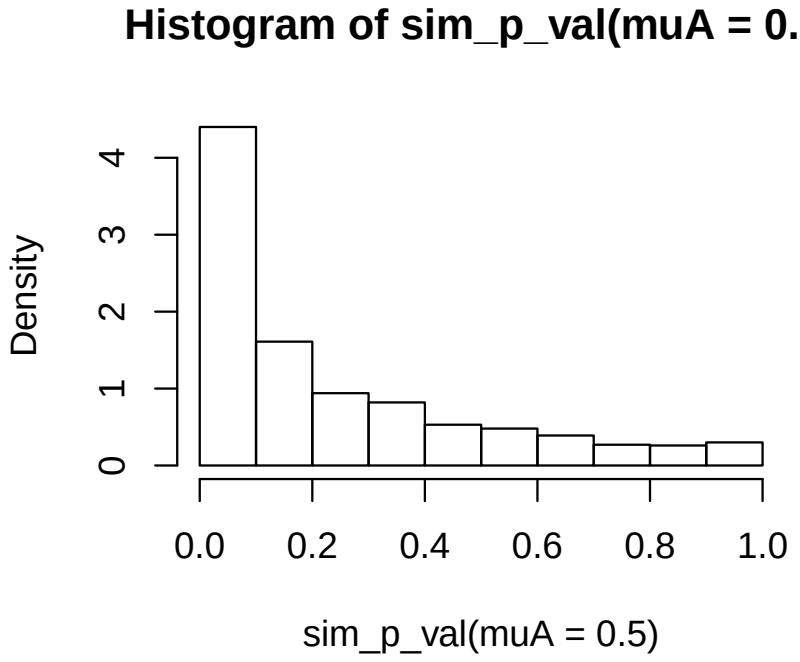

```
hist(sim_p_val(muA = 0.1), freq = FALSE)
```



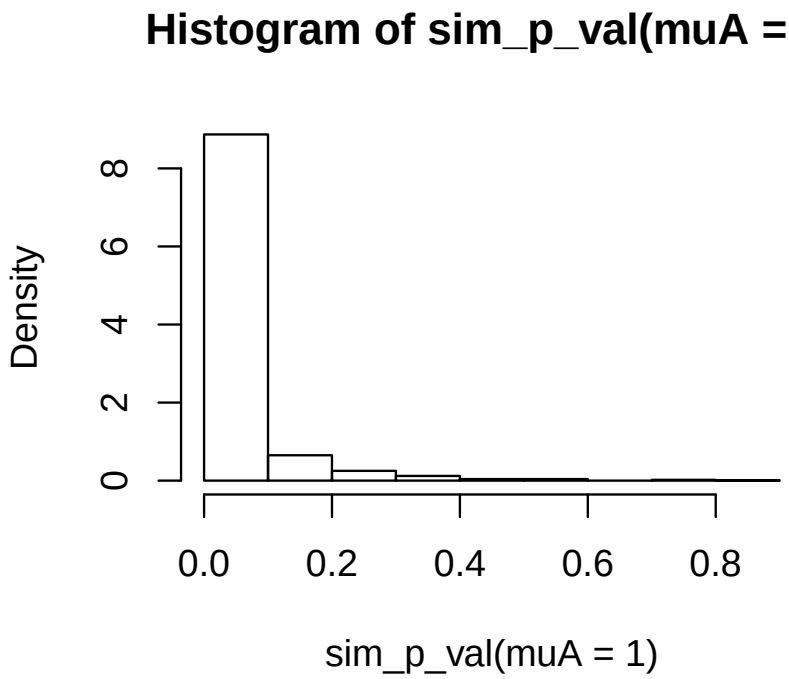

```
hist(sim_p_val(muA = 0.3), freq = FALSE)
```




```r
hist(sim_p_val(muA = 0.5), freq = FALSE)
```




```r
hist(sim_p_val(muA = 1), freq = FALSE)
```




When we compute a $p_{\text{val}}$ and get a statistically significant result we may be interested in whether others repeating our study will also get a statistically significant result; in other words, whether they will be able to replicate our result. This issue is discussed in Boos and Stefanski [2011]. They noted that for p-values that are near-misses (that is, $p_{\text{val}} < \alpha$ but only barely) there are good odds that replication studies will not also reject $H_0$, but when the p-value is much smaller than $\alpha$, the odds of replication should be good. They even recommend reporting estimates of the replication probability to signal how fragile the results of the study are.

### *Section 4: Tests Concerning a Population Proportion*

In Example 1 we saw what a small sample test for a population proportion looks like. When our data follows a Bernoulli distribution, we first state our null and alternative hypothesis:

Then we identify the distribution of the number of "successes" in the sample if $H_0$ is true:

Finally, we can provide a formula for computing $p_{\text{val}}$.

In this section I consider the large-sample version of the test. First, consider the sample proportion $\hat{p}$ computed from Bernoulli data $X_1, \ldots, X_n$, $X_i \sim \text{Ber}(p)$. Assume $H_0 : p = p_0$ is true. What then is $\mathbb{E}[\hat{p}]$ and $\text{Var}(\hat{p})$?

Based on this, what is the approximate distribution for $\hat{p}$ for large $n$?

Using this, we can create a large-sample test for sample proportions[13], described below.

[13] We can extend this reasoning to other statistics that asymptotically follow the Normal distribution. Suppose $\hat{\theta}$ is a consistent estimator of $\theta$, and let $\text{SD}(\hat{\theta}) = \sigma_{\hat{\theta}}$. If we have

$$Z = \frac{\hat{\theta} - \theta}{\sigma_{\hat{\theta}}}$$

and the approximate distribution of $Z$ is $N(0,1)$, then we can test $H_0 : \theta = \theta_0$ against some alternative using the statistic

$$Z = \frac{\hat{\theta} - \theta_0}{\sigma_{\hat{\theta}}}$$

In this case, under $H_0$, $\sigma_{\hat{p}} = \sqrt{np_0(1-p_0)}$, thus producing the large-sample test statistics described.

Below are large-sample Type II error analysis formulas:

---

*Example 6*

Jack Johnson and John Jackson are running for President of Earth. You work for the Johnson campaign and want to determine whether Johnson is currently the candidate with the most support. You plan on conducting a survey asking potential voters who they plan to vote for in the election.

1. Let $p$ represent the proportion of potential voters who support Johnson. State an appropriate null and alternative hypothesis.

2. You want to be able to detect when Johnson has a 1% lead among potential voters with a probability of 95%. Find a sample size capable of achieving this when your level of significaince is $\alpha = 0.05$.

3. It turns out that Johnson is very rich and can afford to conduct a survey of the size found above, and in the survey of 61,000 potential voters, 30,698 reported they were planning on voting for Johnson. Compute $p_{\text{val}}$ and conclude.

```r
alpha <- 0.05
beta <- 0.05
p0 <- 0.5
pA <- 0.51
n <- 61000
x <- 30698
(za <- qnorm(alpha, lower.tail = FALSE))
```

```
## [1] 1.644854
```

```r
(zb <- qnorm(beta, lower.tail = FALSE))
```

```
## [1] 1.644854
```

```r
# Part 2
ceiling(((za * sqrt(p0 * (1 - p0)) + zb * sqrt(pA * (1 - pA)))/(pA - p0))^2)
```

```
## [1] 27051
```

```r
prop.test(x, n, p = 0.5, alternative = "greater", correct = FALSE)
```

```
##
##  1-sample proportions test without
##  continuity correction
##
## data:  x out of n, null probability 0.5
## X-squared = 2.5708, df = 1, p-value =
## 0.05443
## alternative hypothesis: true p is greater than 0.5
## 95 percent confidence interval:
##  0.499916 1.000000
## sample estimates:
##         p
## 0.5032459
```

---

## *Section 5: Further Aspects of Hypothesis Testing*

Suppose we want to perform hypothesis tests for describing the value of the population variance: that is, we wish to test

Assume that the data $X_1, \ldots, X_n$ is an i.i.d. sample from the $N(\mu, \sigma)$ distribution[14]. Then se can describe the distribution of $S^2$, the sample variance:

[14] The $t$-test we saw in Section 3 was somewhat robust to the Normality assumption, working well for large sample sizes even when the assumptions of the test are not met. However, the $\chi^2$ test is *not* robust to this assumption. As discussed before, for non-Normal data, inference regarding $\sigma^2$ may not even be very useful when the data doesn't follow a Normal distribution.

Using this we can formulate a statistical test for inference for $\sigma^2$ (and thus $\sigma$ as well):

Below are formulas for Type II error analysis:

Sample size planning for this test is difficult and may require the use of numerical techniques.

---

*Example 7*

1. Suppose we plan to use ten days of returns from a stock price series to test whether the volatility (that is, $\sigma$) of the stock is greater than 10% or not. State the null and alternative hypotheses.

2. Suppose the true volatility of the stock is 15%. Estimate the probability of comitting a Type II error when $n = 10$ and $\alpha = 0.1$.

3. We have the following returns from the previous ten days of the stock with ticker symbol CGM:

```
cgm2 <- c(-0.2264,  0.0188, -0.0496,  0.1990,  0.1941,
          -0.0219, -0.0177,  0.0847,  0.0167, -0.0736)
```

The standard deviation and variance of the stock's daily returns are given below:

```
var(cgm2)
```

```
## [1] 0.01594101
```

```
(vol <- sd(cgm2))
```

```
## [1] 0.1262577
```

Test whether the volatility (that is, $\sigma$) of the stock is greater than 10% or not with significance level $\alpha = 0.1$.

```
sigma0 <- .1
# Part 2
pchisq(sigma0^2/.15^2 * qchisq(.1, df = 10 - 1, lower.tail = FALSE),
       df = 10 - 1)
```

```
## [1] 0.3136713
```

```
# Part 3
pchisq((10 - 1) * var(cgm2)/sigma0^2, df = 10 - 1, lower.tail = FALSE)
```

```
## [1] 0.1105089
```

---

In hypothesis testing, we can find **statistically significant** results (where $H_0$ was rejected) that are not **practically significant**. That is, we might conclude that $H_0$ is false, but the difference between $\theta_0$ and our best estimate of the true value of the parameter of interest are barely worth mentioning. Large sample sizes produce tests so powerful they can detect even tiny divergences from $H_0$, even if the actual effect is barely worth mentioning. Thus we should be cautious and not overstate the importance of our test's conclusions.

---

*Example 8*

Suppose we are testing to see if the proportion of individuals who have some rare disease is more than $p = 0.007$. We have a lot of funding and conduct a massive study and can conclude that, in fact, the true proportion of the population with the disease is more than 0.007. But our point estimate for this proportion is $\hat{p} = 0.00711$; this is barely larger than the hypothesized value, so the test's results are not noteworthy.

---

Statistical tests and confidence intervals have a connection. If we have a $100(1-\alpha)\%$ confidence interval $(l(x_1, \ldots, x_n), u(x_1, \ldots, x_n))$ and consider the set of hypotheses:

The CI can be interpreted as the set of $\theta_0$ for which $H_0$ would not be rejected at significance level $\alpha$. $100(1-\alpha)\%$ confidence bounds have a similar interpretation for the alternative hypotheses:

---

*Example 9*

Compute a 95% confidence interval for the mean diameter of ball bearings using the data mentioned in Example 3 (using $t$-procedures). Does the confidence interval agree with the conclusion of the test conducted in Example 4?

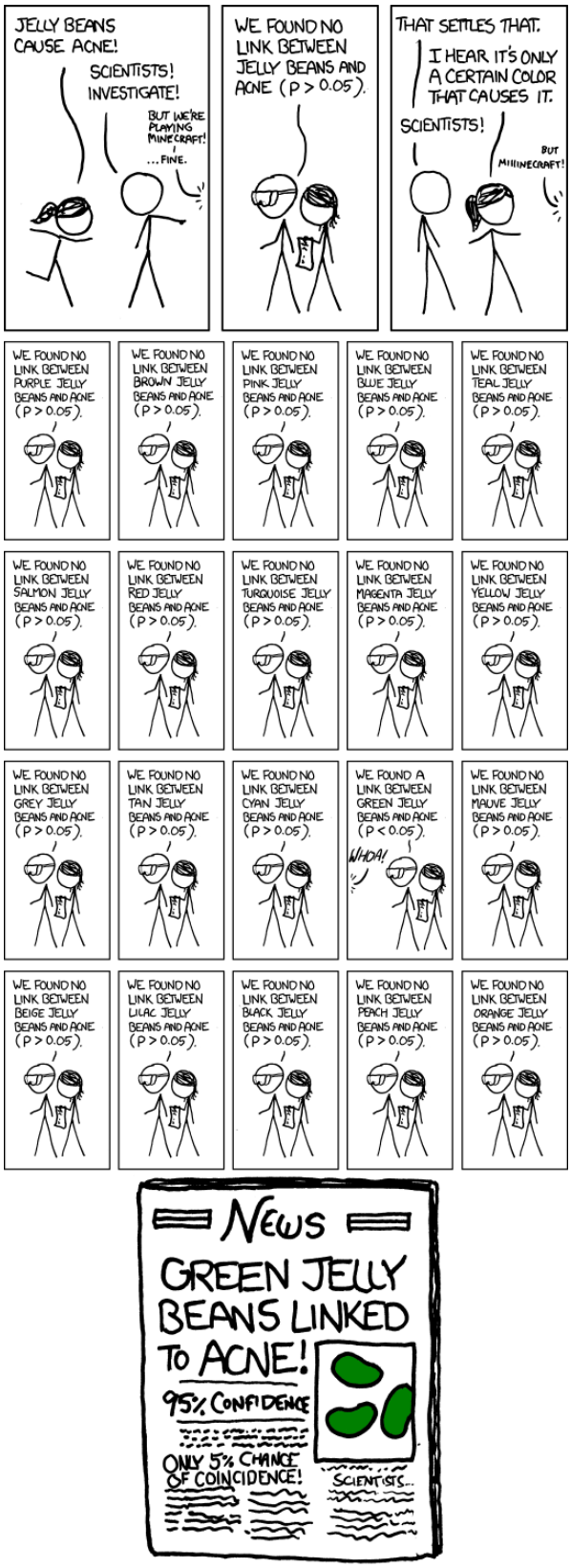
---

In many situations we don't want to conduct just one statistical test, but many. When we do so, the probability of making a Type I error in *any* test increases.

Suppose, for example, that we perform $K$ tests that are *independent* of each other (a strong and likely incorrect assumption). The following calculations show the probability of making a Type I error in the *study*:

The problem is explained well by Munroe [2011] in the comic xkcd.

One approach to this problem is to adjust the significance levels of the tests to achieve a study Type I error rate. For example, we could work with the above expression to find an appropriate $\alpha$ for each test.


JELLY BEANS CAUSE ACNE!
SCIENTISTS! INVESTIGATE!
BUT WE'RE PLAYING MINECRAFT!
...FINE.
WE FOUND NO LINK BETWEEN JELLY BEANS AND ACNE (P > 0.05).
THAT SETTLES THAT.
I HEAR IT'S ONLY A CERTAIN COLOR THAT CAUSES IT.
SCIENTISTS!
BUT MIIIINECRAFT!
WE FOUND NO LINK BETWEEN PURPLE JELLY BEANS AND ACNE (P > 0.05).
WE FOUND NO LINK BETWEEN BROWN JELLY BEANS AND ACNE (P > 0.05).
WE FOUND NO LINK BETWEEN PINK JELLY BEANS AND ACNE (P > 0.05).
WE FOUND NO LINK BETWEEN BLUE JELLY BEANS AND ACNE (P > 0.05).
WE FOUND NO LINK BETWEEN TEAL JELLY BEANS AND ACNE (P > 0.05).
WE FOUND NO LINK BETWEEN SALMON JELLY BEANS AND ACNE (P > 0.05).
WE FOUND NO LINK BETWEEN RED JELLY BEANS AND ACNE (P > 0.05).
WE FOUND NO LINK BETWEEN TURQUOISE JELLY BEANS AND ACNE (P > 0.05).
WE FOUND NO LINK BETWEEN MAGENTA JELLY BEANS AND ACNE (P > 0.05).
WE FOUND NO LINK BETWEEN YELLOW JELLY BEANS AND ACNE (P > 0.05).
WE FOUND NO LINK BETWEEN GREY JELLY BEANS AND ACNE (P > 0.05).
WE FOUND NO LINK BETWEEN TAN JELLY BEANS AND ACNE (P > 0.05).
WE FOUND NO LINK BETWEEN CYAN JELLY BEANS AND ACNE (P > 0.05).
WE FOUND A LINK BETWEEN GREEN JELLY BEANS AND ACNE (P < 0.05).
WHOA!
WE FOUND NO LINK BETWEEN MAUVE JELLY BEANS AND ACNE (P > 0.05).
WE FOUND NO LINK BETWEEN BEIGE JELLY BEANS AND ACNE (P > 0.05).
WE FOUND NO LINK BETWEEN LILAC JELLY BEANS AND ACNE (P > 0.05).
WE FOUND NO LINK BETWEEN BLACK JELLY BEANS AND ACNE (P > 0.05).
WE FOUND NO LINK BETWEEN PEACH JELLY BEANS AND ACNE (P > 0.05).
WE FOUND NO LINK BETWEEN ORANGE JELLY BEANS AND ACNE (P > 0.05).
NEWS
GREEN JELLY BEANS LINKED TO ACNE!
95% CONFIDENCE
ONLY 5% CHANCE OF COINCIDENCE!
SCIENTISTS...

The above assumption of independence, though, is strong and unrealistic. Another approach is to use the **Bonferroni inequality**:

This inequality suggests that our $\alpha$ for each test should be:

This may be too strong a correction, though; imagine if we were doing 1000 tests! Thus we don't see this approach used when $K$ is large.

---

*Example 10*

A medical researcher tests 1000 genes to see if there is a relationship between gene expression and rate of occurance of cancer. The researcher wants a study Type I error rate of $\alpha = 0.1$. How should we choose $\alpha$ for each test if we assume each test is independent? What if we use the Bonferroni inequality approach?

```
alpha <- 0.1
K <- 1000
1 - (1 - alpha)^(1/K)  # Independence approach
```

```
## [1] 0.000105355
```

```
alpha/K  # Bonferroni approach
```

```
## [1] 1e-04
```

---

There are other approaches to multiple hypothesis testing. Procedures such as ANOVA and the $\chi^2$ test have the following approach:

1. Execute an overall test to see if any effect is present.
2. If the null hypothesis of no effect is rejected, do a detailed analysis to see where the divergence from this null hypothesis occurs.

Some of the tests discussed in this chapter follow from the **likelihood ratio principle**. The **likelihood ratio statistic** is defined below:

Tests based on the likelihood ratio reject $H_0$ when the likelihood ratio statistic is "small". The statistic is useful for generating new statistical tests when data follows particular distributions. We can also find more expressive hypotheses using the likelihood ratio.

---

*Example 11*

Suppose $X_1, \ldots, X_n$ is an i.i.d. sample, with $X_i \sim \text{Exp}(\mu_i)$. $H_0$ and $H_A$ are described below:

1. It can be shown that the MLE for $\mu_i = \mu$ when $H_0$ is true is $\hat{\mu} = \bar{X}$, while the MLE for $\mu_i$ in general is $\mu_i = X_i$. Find the corresponding likelihood ratio.

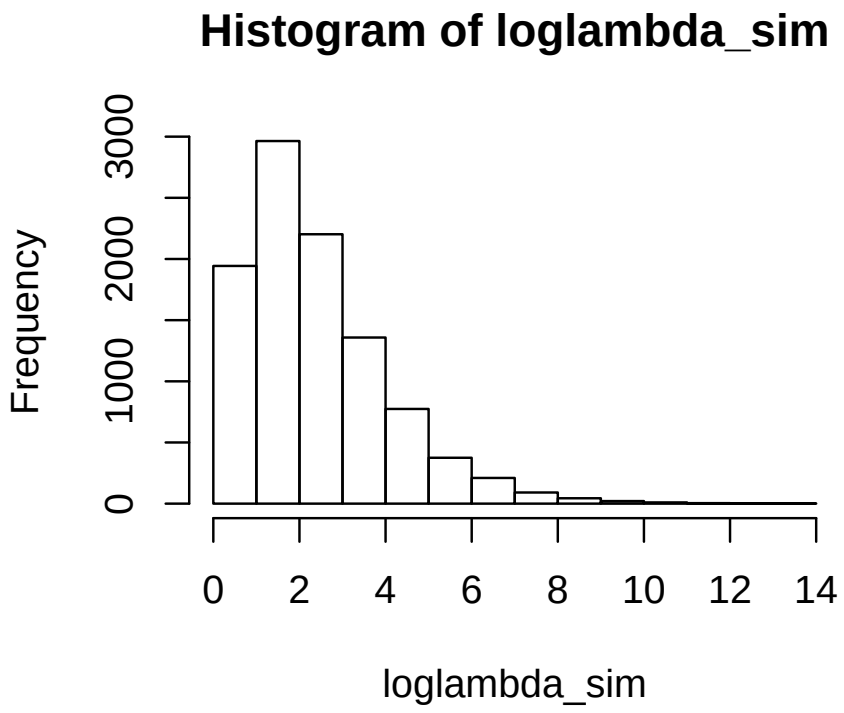
2. Show that if $H_0$ is true, the distribution of the likelihood ratio does not depend on $\mu$. (This involves showing that $\mu$ is a scale parameter.)

3. It is often numerically easier to work with the negative log-likelihood ratio than the likelihood ratio since for large $n$ the likelihood ratio can be very small, and the negative log-likelihood fits with our intuition that large test statistics are evidence against $H_0$. Write an expression for the negative log-likelihood ratio.

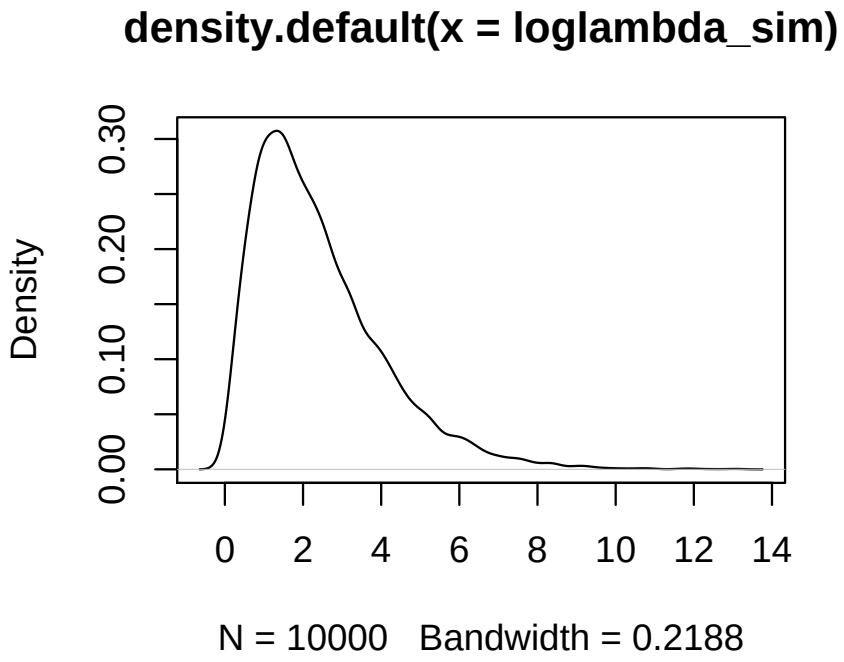
4. Below I estimate the sampling distribution of the negative log-likelihood ratio under $H_0$ in order to estimate critical values for the negative log-likelihood ratio statistic when $n = 5$:

```
n <- 5
loglambda_sim <- replicate(10000, {
  dat <- rexp(n)  # Can use rate = 1 since the rate does not matter under H0
  -(sum(log(dat)) - n * log(mean(dat)))  # Simulated test statistics
```

```
})
hist(loglambda_sim)
```




```
plot(density(loglambda_sim))
```



```
quantile(loglambda_sim, c(.9, .95, .99, .995, .999, .9995))
```

```
##       90%       95%       99%     99.5%
##  4.597638  5.595884  7.707713  8.505277
##     99.9%    99.95%
## 10.769724 11.840371
```

Suppose now that we are tracking the time between eruptions of a geyeser and we want to know whether the number of eruptions can be modelled with a Poisson process. If that is the case, the time between eruptions is i.i.d. and follow from an exponential distribution with some mean. All we wish to know is whether i.i.d. exponential time is an appropriate model for the time between eruptions (we don't necessarily care about the parameters of the model).

We watch our geyeser and observe the following times (in hours) between eruptions:

```
erupt_time <- c(1, 0.1, 27.6, 6.5, 16.3)
```

Test the appropriate hypotheses and estimate $p_{\text{val}}$. What's the conclusion of the test at a significance level of $\alpha = 0.01$?

```
(teststat <- -(sum(log(erupt_time)) -
                 length(erupt_time) * log(mean(erupt_time))))  # Test statistic
```

```
## [1] 5.982522
```

```
mean(loglambda_sim > teststat)  # Estimated p-value
```

```
## [1] 0.0385
```

```
mean(loglambda_sim > teststat) < 0.01
```

```
## [1] FALSE
```

---

Methods based on likelihood ratios make strong assumptions about the distribution of the data, specifying the distribution the data takes save for information about the values of some of the parameters. These methods are known as **parametric methods** since they are ultimately probing about the value of parameters of some assumed distribution. **Distribution-free** methods, also known as **non-parametric methods**, make fewer assumptions about the distribution of the data. These methods will not be considered in this course.

## *References*

Christie Aschwanden. Science isn't broken, 2015. URL https://fivethirtyeight.com/features/science-isnt-broken/.

Christie Aschwanden. Failure is moving science forward, 2016. URL https://fivethirtyeight.com/features/failure-is-moving-science-forward/.

Dennis D. Boos and Leonard A. Stefanski. P-value precision and reproducibility. *The American Statistician*, 65(4):213–221, 2011. DOI: 10.1198/tas.2011.10129. URL https://doi.org/10.1198/tas.2011.10129.

Randall Munroe. Significant, 2011. URL https://xkcd.com/882/.

Ronald L. Wasserstein and Nicole A. Lazar. The ASA's statement on p-values: Context, process, and purpose. *The American Statistician*, 70(2):129–133, 2016. DOI: 10.1080/00031305.2016.1154108. URL https://doi.org/10.1080/00031305.2016.1154108.