STAT 513 fa 2019 Lec 03

Measuring strength of evidence against the null with p-values

Karl B. Gregory

Measuring strength of evidence against the null

- For some $\alpha \in (0,1)$, consider testing some null and alternate hypotheses H_0 versus H_1 based on a random sample $X_1, \dots, X_N$ with the test

$$\text{Reject } H_0 \text{ iff } T(X_1, \dots, X_n) \in \mathcal{R}_\alpha,$$

  where $T(X_1, \dots, X_n)$ is a test statistic and $\mathcal{R}_\alpha$ is a rejection region chosen such that the test has size less than or equal to α.

- We have seen that if we decrease the size of the test to protect against making a Type I error, the test will require stronger evidence against H_0 in order to reject it.

- If we choose a very small value of α and draw a random sample which leads us to H_0, this is a stonger result than if we reject H_0 using a large value of α.

- **Example:** Consider the case of Vinaya and her younger brother Anuj, who are interested in testing some hypotheses H_0 and H_1. Each gathers data, and

  - Anuj rejects H_0 based on a test which has size 0.10 and
  - Vinaya rejects H_0 based on a test which has size 0.01.

  Both Anuj and Vinaya have rejected their null hypotheses, but may we say that the result of one of them is stronger in some sense than that of the other? Since Vinaya set the size of her test at 0.01, she requires stronger evidence against H_0 in order to reject it than does Anuj, who set the size of his test at 0.10; Vinaya's test caps the probability of a Type I error at 0.01, while Anuj's test allows Type I errors to occur with probability as great as 0.10. We should have a greater suspicion that Anuj's claim is an error than that Vinaya's claim is an error. We shall say that Vinaya's result is more *significant*.

- If the size of a test is less than or equal to α, we will say that the test has *significance level* α.

- If we reject H_0 using a test with significance level α, we say that we "reject the null hypothesis at significance level α".

- A way to measure the strength of some observed evidence against H_0 is to find the smallest significance level α at which the observed data would lead to a rejection of H_0. This smallest significance level is called the p-value.

- **Exercise:** *Let $X_1, \dots, X_{10}$ be a random sample from the Normal(μ, σ^2) distribution, where μ and σ^2 are unknown, and suppose we wish to test H_0: $\mu \leq 5$ versus H_1: $\mu > 5$. Suppose $\sqrt{10}(\bar{X}_{10} - 5)/S_{10} = 2.63$.*

  i) *What is our decision about H_0 versus H_1 if we test with significance level $\alpha = 0.05$?*

  ii) *What is our decision about H_0 versus H_1 if we test with significance level $\alpha = 0.01$?*

  iii) *Compute the size of the test Reject H_0 iff $\sqrt{10}(\bar{X}_{10} - 5)/S_{10} > 2.63$.*

  iv) *What is the smallest significance level at which the observed random sample, for which $\sqrt{10}(\bar{X}_{10} - 5)/S_{10} = 2.63$, would lead to a rejection of H_0?*

  v) *Draw a plot of the density of the test statistic when $\mu = 5$ and shade the area corresponding to the p-value.*

  Answers:

  i) A size-0.05 test (thus having significance level 0.05) is

  $$\text{Reject } H_0 \text{ iff } \sqrt{10}(\bar{X}_{10} - 5)/S_{10} > t_{9,0.05} = \texttt{qt(.95,9)} = 1.833113,$$

  so we reject H_0 at the 0.05 significance level.

  ii) A size-0.01 test is

  $$\text{Reject } H_0 \text{ iff } \sqrt{10}(\bar{X}_{10} - 5)/S_{10} > t_{9,0.01} = \texttt{qt(.99,9)} = 2.821438,$$

  so we fail to reject H_0 at the 0.01 significance level.

  iii) The size of the test is given by

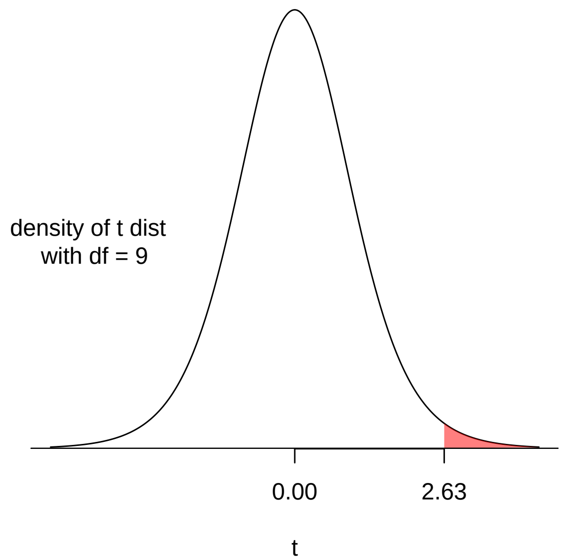
  $$\begin{aligned} P_{\mu=5}(\sqrt{10}(\bar{X}_{10} - 5)/S_{10} > 2.63) &= P(T > 2.63), \quad T \sim t_9 \\ &= \texttt{1 - pt(2.63,9)} \\ &= 0.01367933. \end{aligned}$$

  iv) A size-0.01367933 test is

  $$\text{Reject } H_0 \text{ iff } \sqrt{10}(\bar{X}_{10} - 5)/S_{10} > t_{9,0.01367933} = \texttt{qt(1-0.01367933,9)} = 2.63$$

  For any significance level less than 0.01367933, we get a larger critical value, and we do not reject H_0. So the p-value is 0.01367933.

  v) Shade under right tail beyond $\sqrt{10}(\bar{X}_{10} - 5)/S_{10} = 2.63$.



- We can interpret the p-value as the probability—if the null hypothesis is true—of observing a random sample that carries as much or more evidence against the null hypothesis as the observed sample.

- **Exercise:** *Let $X_1, \ldots, X_{10}$ be a random sample from the Normal(μ, σ^2) distribution, where μ and σ^2 are unknown, and suppose we wish to test H_0: $\mu = 8$ versus H_1: $\mu \neq 8$. Suppose $\sqrt{10}(\bar{X}_{10} - 8)/S_{10} = -3.12$.*

  i) What is our decision about H_0 versus H_1 if we test with significance level $\alpha = 0.05$?

  ii) What is our decision about H_0 versus H_1 if we test with significance level $\alpha = 0.01$?

  iii) If H_0 is true, what is the probability of getting a sample which carries as much or more evidence against H_0?

  iv) Draw a plot of the density of the test statistic when $\mu = 8$ and shade the area corresponding to the p-value.

  Answers:

  i) A size-0.05 test is

  $$\text{Reject } H_0 \text{ iff } |\sqrt{10}(\bar{X}_{10} - 8)/S_{10}| > t_{9,0.025} = \texttt{qt(.975,9)} = 2.262157,$$

  so we reject H_0 at the 0.05 significance level.

  ii) A size-0.01 test is

  $$\text{Reject } H_0 \text{ iff } |\sqrt{10}(\bar{X}_{10} - 8)/S_{10}| > t_{9,0.005} = \texttt{qt(.995,9)} = 3.249836,$$

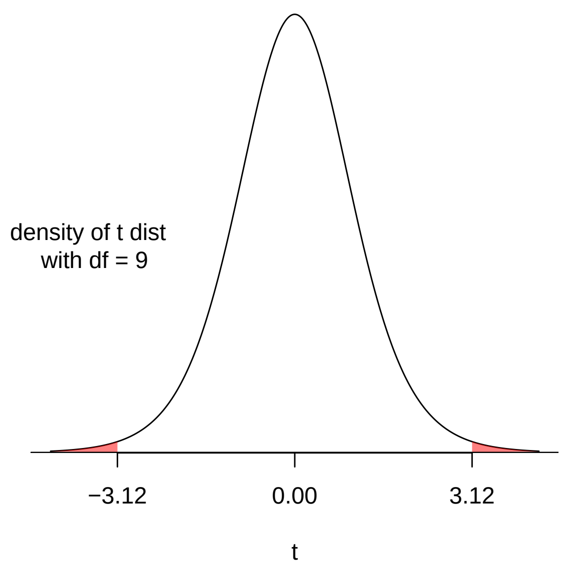
  so we fail to reject H_0 at the 0.01 significance level.

iii) A sample which carries as much or more evidence against H_0 than the observed sample will have $|\sqrt{10}(\bar{X}_{10} - 8)/S_{10}| > 3.12$. The probability of this when H_0 is true is

$$\begin{aligned} P_{\mu=8}(|\sqrt{10}(\bar{X}_{10} - 8)/S_{10}| > 3.12) &= P(|T| > 3.12), \quad T \sim t_9 \\ &= 2(1 - P(T < 3.12)) \\ &= \texttt{2*(1-pt(3.12,9))} \\ &= 0.01231863. \end{aligned}$$

This is the p-value.

iv) Shade under both tails beyond $|\sqrt{10}(\bar{X}_{10} - 8)/S_{10}| = 3.12$.



- Using p-values, we can reformulate the test

$$\text{Reject } H_0 \text{ iff } T(X_1, \ldots, X_n) \in \mathcal{R}_\alpha$$

as

$$\text{Reject } H_0 \text{ iff } p\text{-value} \leq \alpha.$$

If the p-value is less or equal to the the significance level α, it indicates that $T(X_1, \ldots, X_n) \in \mathcal{R}_\alpha$ and vice versa. The smaller the p-value, the further inside the rejection region the test statistic lies.

- Researchers often report only the p-value without reporting the value of the test statistic. This is okay; they are not hiding anything, because the value of the test statistic could be computed going backwards from the p-value. The reason for reporting the p-value is that it is a succinct and standard measure of how strong the evidence is against H_0.

- Sometimes the p-value is referred to as the *observed significance level*.

- The p-value is sometimes misinterpreted as "the probability that H_0 is true". The p-value is *not* the probability that H_0 is true! Rather, we may think of it as measuring the plausibility of H_0 in light of the data. If the p-value is small, then H_0 is implausible in light of the data; if the p-value is large, H_0 is plausible in light of the data.

- **Formulas for p-values of tests about Normal mean:** Suppose $X_1, \dots, X_n$ is a random sample from the Normal(μ, σ^2) distribution, where μ and σ^2 are unknown. Then we have the following, where $F_{t_{n-1}}$ is the cdf of the t_{n-1} distribution and $T_n = \sqrt{n}(\bar{X}_n - \mu_0)/S_n$:

H_0	H_1	Reject H_0 iff	p-value
$\mu \leq \mu_0$	$\mu > \mu_0$	$T_n > t_{n-1,\alpha}$	$1 - F_{t_{n-1}}(T_n)$
$\mu \geq \mu_0$	$\mu < \mu_0$	$T_n < -t_{n-1,\alpha}$	$F_{t_{n-1}}(T_n)$
$\mu = \mu_0$	$\mu \neq \mu_0$	$\lvert T_n \rvert > t_{n-1,\alpha/2}$	$2(1 - F_{t_{n-1}}(\lvert T_n \rvert))$

- **Exercise:** *Let $X_1, \dots, X_{10}$ be a random sample from the Normal(μ, σ^2) distribution, where μ and σ^2 are unknown. Suppose $S_{10}^2 = 3$. Give the p-values for testing the following sets of hypotheses:*

  i) H_0: $\sigma^2 \leq 5$ *versus* H_1: $\sigma^2 > 5$

  ii) H_0: $\sigma^2 \geq 5$ *versus* H_1: $\sigma^2 < 5$

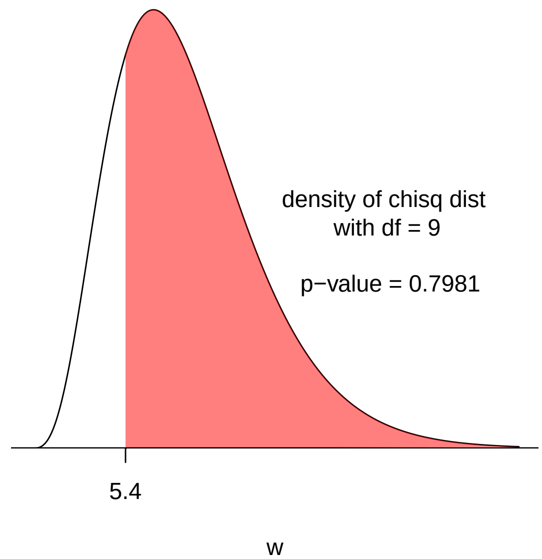
  iii) H_0: $\sigma^2 = 5$ *versus* H_1: $\sigma^2 \neq 5$

  Answers:

  i) For this set of hypotheses, larger values of $(10-1)S_{10}^2/5$ carry more evidence against H_0. For this sample $(10-1)S_{10}^2/5 = 5.4$. If we were to reject H_0 iff $(10-1)S_{10}^2/5 > 5.4$, the size of the test would be

$$\begin{aligned} P_{\sigma^2=4}((10-1)S_{10}^2/5 > 5.4) &= P(W > 5.4), \quad W \sim \chi_9^2 \\ &= 1 - P(W < 5.4) \\ &= 1 \texttt{ - pchisq(5.4,9)} \\ &= 0.7981391. \end{aligned}$$

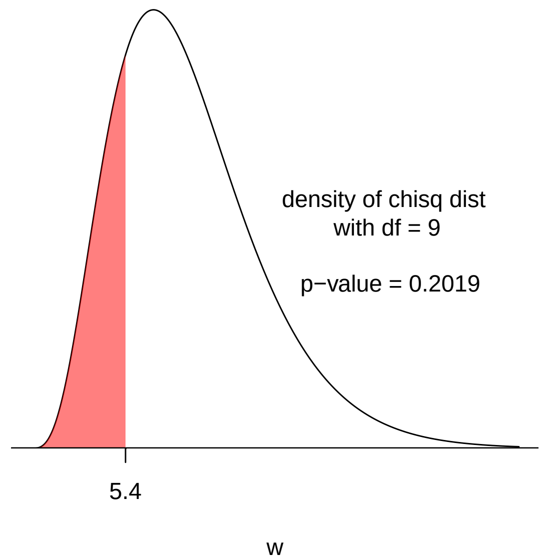
  We can depict the p-value as the area under the χ_9^2 density to the right of 5.4.



ii) For this set of hypotheses, smaller values of $(10-1)S_{10}^2/5$ carry more evidence against H_0. For this sample $(10-1)S_{10}^2/5 = 5.4$. If we were to reject H_0 iff $(10-1)S_{10}^2/5 < 5.4$, the size of the test would be

$$\begin{aligned} P_{\sigma^2=4}((10-1)S_{10}^2/5 < 5.4) &= P(W < 5.4), \quad W \sim \chi_9^2 \\ &= P(W < 5.4) \\ &= \texttt{pchisq(5.4,9)} \\ &= 0.2018609. \end{aligned}$$

We can depict the p-value as the area under the χ_9^2 density to the left of 5.4.



iii) For the two-sided hypothesis test, we reject H_0 at significance level α iff

$$\frac{(10-1)S_{10}^2}{5} < \chi^2_{9,1-\alpha/2} \text{ or } \frac{(10-1)S_{10}^2}{5} > \chi^2_{9,\alpha/2},$$

so that smaller or larger values of $(10-1)S_{10}^2/5$ carry more evidence against H_0.

Since the χ^2 distributions are not symmetric, we need to think more carefully about how to compute the p-value than when we are dealing with the t-distribution. For example, when testing H_0: $\mu = 0$ versus H_1: $\mu \neq 0$, a sample with $\bar{X}_n = 1$ and a sample with $\bar{X}_n = -1$ carry the same amount of evidence against the null (assuming the samples have the same value of S_n); however, when testing H_0: $\sigma^2 = 5$ versus H_1: $\sigma^2 \neq 5$, a sample with $S_n^2 = 3$ carries less evidence against H_0 than a sample with $S_n^2 = 7$, even though the two sample variances are the same distance from the null value $\sigma_0^2 = 5$. This is due to the skewness of the sampling distribution of S_n^2. Thus, to compute the p-value for a two-sided test about the variance, we ask: on which side of the null value does S_n^2 lie, and which value of S_n^2 on the opposite side carries the same amount of evidence against H_0? Then, finally, what is the sum of the areas in the two tails beyond these values?

For this sample, $S_{10}^2 = 3 < \sigma_0^2 = 5$, so that if we reject H_0, we will reject due to

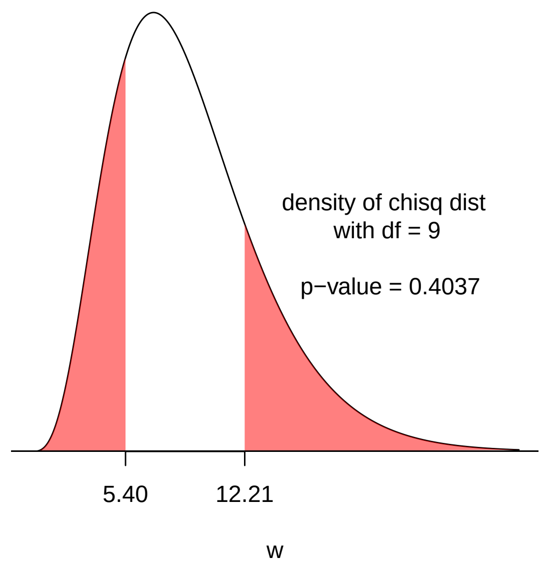
$$\frac{(10-1)S_{10}^2}{5} < \chi^2_{9,1-\alpha/2},$$

that is, our test statistic will lie in the left tail of the null distribution. So we compute the area under the χ_9^2 density to the left of $(10-1)S_{10}^2/5 = 5.4$, and this is equal to $1/2$ of the

p-value. So to get the p-value, we multiply this by 2. That is, we compute the p-value as

$$\begin{aligned} 2P_{\sigma^2=5}((10-1)S_{10}^2/5 < 5.4) &= 2P(W < 5.4), \quad W \sim \chi_9^2 \\ &= \texttt{2*pchisq(5.4,9)} \\ &= 0.4037219. \end{aligned}$$

We can depict the p-value as the sum of the area under the χ_9^2 density to the left of 5.4 and to the right of the value `qchisq(1 - pchisq(5.4,9),9)` $= 12.20753$, which is the value such that the area under the χ_9^2 density to the right of it is equal to the area under the χ_9^2 density to the left of 5.4. A sample with $(10-1)S_{10}^2/5 = 12.20753$, corresponding to $S_n^2 = 6.781962$, carries the same amount of evidence against H_0 as the observed sample with $S_n^2 = 3$.



In short, compute the tail probability for the tail in which the test statistic lies and then multiply it by 2.

- **Formulas for p-values of tests about Normal variance:** Suppose $X_1, \dots, X_n$ is a random sample from the Normal(μ, σ^2) distribution, where μ and σ^2 are unknown. Then we have the following, where $F_{\chi^2_{n-1}}$ is the cdf of the χ^2_{n-1} distribution and $W_n = (n-1)S_n^2/\sigma_0^2$:

H_0	H_1	Reject H_0 at α iff	p-value
$\sigma^2 \leq \sigma_0^2$	$\sigma^2 > \sigma_0^2$	$W_n > \chi^2_{n-1,\alpha}$	$1 - F_{\chi^2_{n-1}}(W_n)$
$\sigma^2 \geq \sigma_0^2$	$\sigma^2 < \sigma_0^2$	$W_n < \chi^2_{n-1,1-\alpha}$	$F_{\chi^2_{n-1}}(W_n)$
$\sigma^2 = \sigma_0^2$	$\sigma^2 \neq \sigma_0^2$	$W_n < \chi^2_{n-1,1-\alpha/2}$ or $W_n > \chi^2_{n-1,\alpha/2}$	$2 \cdot \min\{F_{\chi^2_{n-1}}(W_n), 1 - F_{\chi^2_{n-1}}(W_n)\}$

- **Exercise:** *Let $X_1, \ldots, X_{50}$ be a random sample from the Bernoulli(p) distribution, where p is unknown.*

  i) *Suppose you wish to test H_0: $p \le 1/3$ versus H_1: $p > 1/3$ and that you observe $X_1 + \cdots + X_{50} = 25$. Discuss the strength of evidence against H_0 and whether to reject it.*

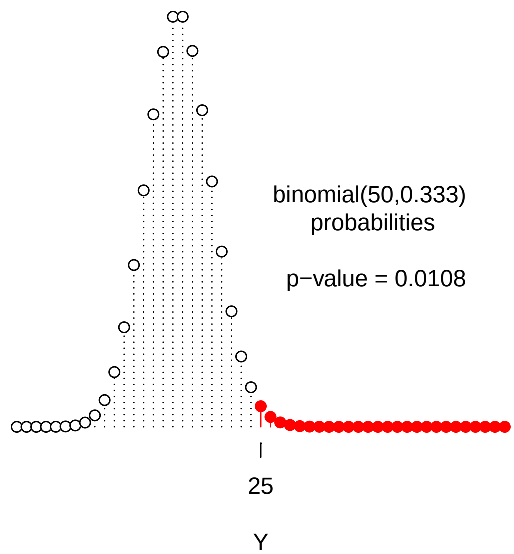
  ii) *Suppose we wish to test H_0: $p = 1/3$ versus H_1: $p \neq 1/3$ and that you observe $X_1 + \cdots + X_{50} = 20$. Discuss the strength of evidence against H_0 and whether to reject it.*

  Answers:

  i) Since $25/50 > 1/3$, the random sample gives some evidence against H_0 and therefore some evidence in favor of H_1. We can measure the strength of the evidence against H_0 by computing a p-value, which we do as follows: Suppose we were to reject H_0 any time we observed $X_1 + \cdots + X_{50} \ge 25$ (such samples would carry as much or more evidence against H_0 than the sample for which $X_1 + \cdots + X_{50} = 25$). This test would have size equal to

$$\begin{aligned} P_{p=1/3}(X_1 + \cdots + X_{50} \ge 25) &= P(Y \ge 25), \quad Y \sim \text{Binomial}(50, p) \\ &= 1 - P(Y \le 24) \\ &= 1 \text{ - } \texttt{pbinom(24,50,1/3)} \\ &= 0.01082668, \end{aligned}$$

  so this is the p-value. Since the p-value is rather small, H_0 seems rather implausible. The sum of the heights of the red points in the plot below represents the p-value:



ii) We can compute a p-value by considering all samples which would carry as much or more evidence against H_0. Since this is a two-sided test and the binomial distribution is discrete and asymmetric, we must proceed very carefully.

The sample outcome $X_1+\cdots+X_{50} = 20$ supports $p > 1/3$, so any sample with $X_1+\cdots+X_{50} \geq 20$ carries as much or more evidence against H_0. The probability of getting any of these samples when $p = 1/3$ is

$$\sum_{y=20}^{50} \binom{50}{y}(1/3)^y(1-1/3)^{50-y} = \texttt{1 - pbinom(19,50,1/3)} = 0.1964139.$$

Now we must consider samples which carry as much or more evidence against H_0 in the opposite direction—that is, which support $p < 1/3$. If we consider the probabilities $P(X_1 + \cdots + X_{50} \leq y)$ for $y = 0, 1, \ldots, n$, we find

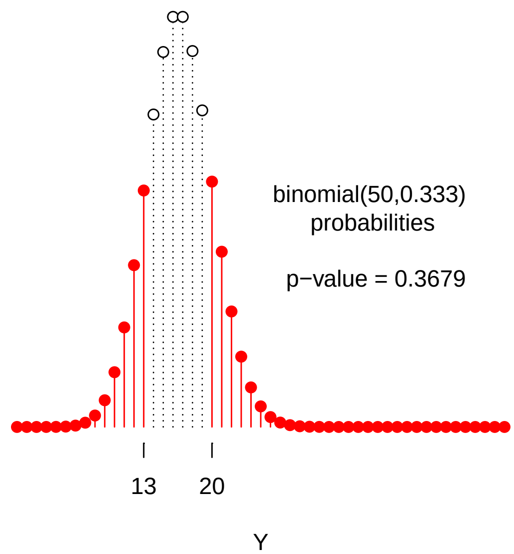
$$
\begin{aligned}
P(X_1 + \cdots + X_{50} \leq 0) &= \texttt{pbinom(0,50,1/3)} = 1.568329 \times 10^{-9} < 0.1964139 \\
&\vdots \\
P(X_1 + \cdots + X_{50} \leq 13) &= \texttt{pbinom(13,50,1/3)} = 0.1714651 < 0.1964139 \\
P(X_1 + \cdots + X_{50} \leq 14) &= \texttt{pbinom(14,50,1/3)} = 0.2612386 > 0.1964139 \\
&\vdots \\
P(X_1 + \cdots + X_{50} \leq 50) &= \texttt{pbinom(50,50,1/3)} = 1 > 0.1964139,
\end{aligned}
$$

so that any sample with $X_1 + \cdots + X_{50} \leq 13$ carries as much or more evidence against H_0 (is as rare or rarer under H_0) than the sample with $X_1 + \cdots + X_{50} = 20$.

So in the end, the two-sided p-value is $P(X_1 + \cdots + X_{50} \leq 13) + P(X_1 + \cdots + X_{50} \geq 20)$, which we can compute in R with

$$\texttt{pbinom(13,50,1/3) + 1 - pbinom(19,50,1/3)} = 0.3678789.$$

In the plot below, the p-value is represented as the sum of the heights of the red lines.



- **Formulas for p-values of tests about Bernoulli success probability:** Suppose $X_1, \dots, X_n$ is a random sample from the Bernoulli(p) distribution, where p is unknown. Then we have the following, where $B_{n,p}$ is the cdf of the Binomial(n, p) distribution, of which we let $B_{n,p,\alpha}$ denote the upper α quantile, and where $Y_n = X_1 + \dots + X_n$:

H_0	H_1	Reject H_0 at α iff	p-value
$p \leq p_0$	$p > p_0$	$Y_n \geq B_{n,p,\alpha}$	$1 - B_{n,p}(Y_n - 1)$
$p \geq p_0$	$p < p_0$	$Y_n < B_{n,p,1-\alpha}$	$B_{n,p}(Y_n)$
$p = p_0$	$p \neq p_0$	$Y_n < B_{n,p,1-\alpha/2}$ or $Y_n \geq B_{n,p,\alpha/2}$	$\begin{cases} 1, & Y_n = np_0 \\ B_{n,p}(Y_n) + 1 - B_{n,p}(Y_n^r - 1), & 0 \leq Y_n < np_0 \\ 1 - B_{n,p}(Y_n - 1) + B_{n,p}(Y_n^l), & np_0 < Y_n \leq n \end{cases}$,

where

$$Y_n^r = \min\{y : 1 - B_{n,p}(y - 1) \leq B_{n,p}(Y_n)\}$$
$$Y_n^l = \max\{y : B_{n,p}(y) \leq 1 - B_{n,p}(Y_n - 1).$$

These p-values can be computed using the R function `binom.test()`. See R documentation for details. Very soon we will cover a much simpler (but approximate) test for a proportion based on the central limit theorem.

- As we encounter more testing situations, we will for each of them consider how the p-value is to be computed.