Econ 123 Review Handout 3: Finite-Sample Inference

Edward Vytlacil, Yale University

January 9, 2023

This handout reviews exact, finite-sample inference, including both hypothesis testing, confidence intervals, and test-inversion which allows the construction of confidence intervals from hypothesis tests and vice versa.

Contents

Hypothesis Testing

Hypothesis testing is a formal way to reach a quantitative conclusion about two mutually exclusive hypotheses, namely the null hypothesis and the alternative hypothesis.

Definition 1: Null and Alternative Hypotheses

The *null hypothesis* $\mathbb{H}_0$ is the hypothesis we accept unless there is sufficiently strong evidence against it in favor of the *alternative hypothesis* $\mathbb{H}_1$. The null hypotheses we consider are statements about a parameter θ being equal to some hypothesized value θ_0 vs either a one-sided or two sided alternative:

$$\mathbb{H}_0 : \theta = \theta_0,$$

$$\text{vs.} \quad \begin{cases} \mathbb{H}_1 : \theta \neq \theta_0, & (\textit{two-sided alternative}), \text{ or} \\ \mathbb{H}_1 : \theta > \theta_0 & (\textit{one-sided alternative, greater}), \text{or} \\ \mathbb{H}_1 : \theta < \theta_0 & (\textit{one-sided alternative, less}). \end{cases}$$

Researchers in economics typically consider two-sided alternatives, though one-sided alternatives are appropriate in some contexts. We initially consider examples with one-sided alternatives for pedagogical reasons - it is easier to understand hypothesis testing in the simpler context of a one-sided alternative.

Example 1 (Probability of Outperforming Benchmark)**.** *Suppose we are interested in whether a mutual fund manager can systematically outperform a market benchmark. In particular, consider the null hypothesis that her fund is equally likely to out-perform as under-perform the benchmark, versus the alternative that her fund is more likely to outperform the benchmark. Let r_A denote the return on her fund in a given year, and let r_M denote the return on the market benchmark. Let X denote an indicator variable for her fund outperforming the market benchmark,*

$$X = \mathbb{1}[r_A - r_M > 0],$$

Indicator functions and Bernoulli and Binomial random variables are reviewed in Review Handout 2.

where $\mathbb{1}$ *is the logical indicator function, and thus*

$$X \sim Bernoulli(p),$$

with p *denoting the probability of the event that she outperforms the market,* $p = \Pr[r_A - r_M > 0]$*. Then our null and alternative hypotheses are:*

$$\begin{aligned} \mathbb{H}_0 : &\quad p = \tfrac{1}{2}, \\ \mathbb{H}_1 : &\quad p > \tfrac{1}{2}. \end{aligned}$$

Example 2 (Expected Excess Return)**.** *Continuing example 1, suppose we are interested in whether a mutual fund manager can systematically outperform a market benchmark in the sense that the expected excess return on her fund is positive, where excess return is defined as return on her fund minus that of the market benchmark. Let* Y *denote the excess return on her fund in a given year,* $Y \equiv r_A - r_M$*. Suppose that*

$$Y \sim N(\mu, \sigma^2).$$

Then our null and alternative hypotheses are:

$$\begin{aligned} \mathbb{H}_0 : &\quad \mu = 0, \\ \mathbb{H}_1 : &\quad \mu > 0. \end{aligned}$$

The normal distribution is reviewed in Review Handout 2.

Given a null and alternative hypothesis, we summarize evidence against the null by a test statistic:

Definition 2: Test Statistic

A *test statistic* T_n is a function of our random sample of size n (and thus a random variable) that summarizes evidence against the null, with larger values indicating stronger evidence against the null.

A test statistic is a function of the sample which we can calculate given the sample. Because the sample is random, the test statistic is random. In order to make statements as to how likely it would be to observe a given level of evidence against the null if the null were true, it will be important to choose a test statistic for which we know what it's distribution would be if the null hypothesis were true (called its *distribution under the null*).[1] Further, how we choose the test statistic depends on the alternative hypothesis. We want to choose a test statistic such that larger values are more evidence against the null in favor of the alternative – we wouldn't want to use a test statistic under which large values are evidence against the null and even stronger evidence against the alternative.

We can now define a hypothesis test:

[1] His handout covers *exact inference*, also called *finite-sample inference*, where the distribution of the test statistic is known exactly for any sample size n. We will later consider asymptotic inference where we may not know the exact distribution of the test statistic but can approximate it for large n using, e.g., the CLT.

Definition 3: Hypothesis Test

A *hypothesis test* is a decision rule that specifies a test statistic T_n and *critical value* c such that:

- If $T_n \leq c$, we accept the null $\mathbb{H}_0$ (fail to reject the null);
- If $T_n > c$, we reject the null $\mathbb{H}_0$ in favor of the alternative $\mathbb{H}_1$.

A hypothesis test can make two different types of errors: reject the null hypothesis when it is true, or fail to reject the null hypothesis when it is false.

Definition 4: Type I Error

A *Type I error* is incorrectly rejecting the null hypothesis when the null is true.

Definition 5: Type II Error

A *Type II error* is failing to reject the null hypothesis when the null is false.

		Decision	
		Fail to Reject $\mathbb{H}_0$	Reject $\mathbb{H}_0$
Truth	$\mathbb{H}_0$	Correct Decision	*Type I Error*
	$\mathbb{H}_1$	*Type II Error*	Correct Decision

Let Pr_θ denote the distribution of the test statistic index by the parameter θ, so that Pr_{θ_0} indicates the distribution of the test statistic under the null, i.e., $Pr_{\theta_0}[T_n > c]$ is the probability that the test statistic is larger than c if $\theta = \theta_0$. Then, if the null hypothesis is true, we falsely reject (make a Type I error) with probability $Pr_{\theta_0}[T_n > c]$, which is called the size of the test.

Definition 6: Size of a Test

The *size* of a test is the probability of a Type I error when the null hypothesis is true, and is given by $Pr_{\theta_0}[T_n > c]$.

Now consider the probability that we reject the null hypothesis when the null is false and some specified value under the alternative is true. If θ is some specific alternative (some specific value such that $\mathbb{H}_1$ is true), then $Pr_\theta[T > c]$ is the probability that we correctly reject the null and $1 - Pr_\theta[T_n > c]$ is the probability of a Type II error. We refer to $Pr_\theta[T > c]$ for a given θ such that $\mathbb{H}_1$ holds as the power of the test against θ.

Definition 7: Power of a Test

Let θ denote a specified parameter value such that $\mathbb{H}_1$ holds. Then the *power* of a test against alternative θ is one minus the probability of a Type II error when θ is the true parameter value, and is given by $Pr_\theta[T_n > c]$.

There is an inherent tension in how we pick the critical value c: the larger we choose c, the more evidence against the null we require for us to reject the null, and the less likely we are to make a Type I error if the null is true but the more likely we are to make a Type II error if the null is false. In most empirical work, we worry more about a Type I error than a Type II error, and thus choose the critical value c to ensure that the size of the test (probability of Type 1 error if null is true) is not larger than some given probability, called the significance level of the test.

Definition 8: Significance Level of a Test

A test has *significance level* α if its size is less than or equal to α.

In the social science, by convention, we typically choose $\alpha = 0.05$, or alternatively $\alpha = 0.01$ or 0.10. To construct a test with the desired significance level, we choose as the critical value the appropriate quantile of the test statistic under the null.

Remark 1 (Finding Critical Values:)**.** *Let $q_{1-\alpha}$ denote the $1-\alpha$ quantile of the distribution of our test statistic under the null. Choose the critical value $c_\alpha = q_{1-\alpha}$. Then the test that rejects the null when $T_n > c_\alpha$ is of level α.*

Example 1. *(continued) Suppose we observe whether the manager's fund outperforms the market benchmark each year for n years. In particular, suppose we observe a random sample $(X_1, ..., X_n)$, where X_t as the indicator variable for the manager outperforming the benchmark in year t, and suppose the X_t are independent Bernoulli(p) random variables. Let $S_n = \sum_{t=1}^{n} X_t$ denote the number of years out of n that the fund outperforms the market, so that $S_n \sim$Binomial(n, p). Choose $T_n = S_n$ as our test statistic. This choice of test statistic satisfies the three key properties we need in a test statistic:*

1. *we can calculate it from our sample;*
2. *larger values provides evidence against $\mathbb{H}_0$ in favor of $\mathbb{H}_1$;*

3. *we know its distribution under the null.*

Suppose that $n = 10$. *Suppose we choose significance level* 0.10, *in which case we choose* $c = 7$, *the* 0.90 *quantile of Binomial*$(10, 0.5)$. *We thus reject the null whenever the manager's fund outperforms the market in* 8 *or more years out of* 10. *We can now calculate size of the test as probability that* $S_{10} > 7$ *when* $p = 0.5$, *which must be less than or equal to the significance level* 0.10. *For any particular alternative, say* $p = 0.7$ *or* $p = 0.9$, *we can calculate the power of the the test against that alternative as probability that* $S_{10} > 7$ *when that value of* p *is true. We can make these calculations using* ***R***. *Notice that the power of the test depends on the alternative. If the manager is only slightly more likely to outperform than underperform the market, the null is false but we have low power (are likely to make a Type II error). If the manager is far more likely to outperform the market, the null is false and we have high power (are unlikely to make a Type II error).*

```
> # Critical Values if N=10, null p=0.5
> qbinom(.90, 10, .5) #for alpha=0.1
[1] 7
> # Size of test, note less than level
> # prob. of reject under null p=0.5
> 1 - pbinom(7,10,0.5)
[1] 0.0546875
> # Power of test against p=0.7
> 1 - pbinom(7,10,0.7)
[1] 0.3827828
> # Power of test against p=0.9
> 1 - pbinom(7,10,0.9)
[1] 0.9298092
```

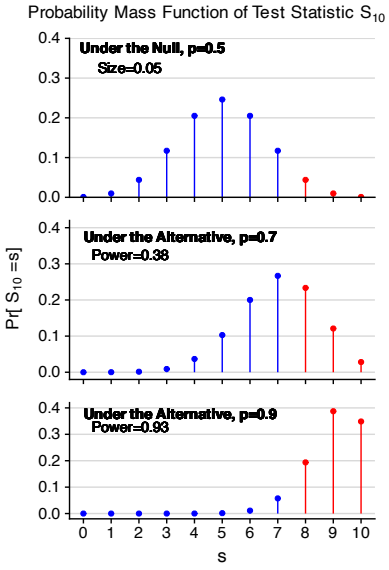


Figure 1: Example 1, Testing $\mathbb{H}_0 : p = 0.5$ vs $\mathbb{H}_1 : p > 0.5$, with test that fails to reject when $S_{10} \leq 7$ and rejects when $S_{10} \geq 8$. Notice that higher p shifts the distribution of S_{10} to the right and results in higher probability of $S_{10} \geq 8$.

Example 2. *(continued) Let* $Y_t = r_{A,t} - r_{M,t}$ *denote the fund managers excess return in year* T_n, *where we are assuming that* $Y_t \sim N(\mu, \sigma^2)$. *For simplicity, assume that* σ^2 *is known. Suppose that we observe returns on her fund over* n *years,* $X_1, X_2, ..., X_n$. *Since we are considering the alternative* $\mathbb{H}_1 : \mu > 0$, *it is natural to define the test statistic as*

$$T_n \equiv \frac{\bar{X}_n}{\sigma/\sqrt{n}},$$

where $\bar{X}_n = \frac{1}{n}\sum_{t=1}^{n} X_t$. *With this definition, we can calculate* T_n *from our sample (since* σ^2 *is assumed known), and the larger the value of* T_n *the more evidence against the null in favor of the alternative. Now consider the distribution of* T_n *under the null. From Remark* 3 *and Corollary* 5 *of Review Handout* 2 *we have that*

$$\frac{\bar{X}_n - \mu}{\sigma/\sqrt{n}} \sim N(0,1),$$

so that

$$T_n = \frac{\bar{X}_n}{\sigma/\sqrt{10}} = \frac{\bar{X}_n - \mu}{\sigma/\sqrt{n}} + \frac{\mu}{\sigma/\sqrt{n}} \sim N\left(\frac{\mu}{\sigma/\sqrt{n}}, 1\right).$$

Thus,

$$T_N \sim N(0,1) \;\; \textit{under } \mathbb{H}_0 : \; \mu = 0.$$

We consider later in this handout how the analysis of Example 2 changes with σ^2 unknown.

Recall that, by Theorem 3 of Review Handout 2, $\frac{\bar{X}_n - \mu}{\sigma/\sqrt{n}} \sim N(0,1)$ implies $\frac{\bar{X}_n - \mu}{\sigma/\sqrt{n}} + a \sim N(a,1)$ for any constant a.

Suppose we choose significance level 0.05*, in which case we choose cutoff* c *to be the* .95 *quantile of* $N(0,1)$*. We can calculate the power of the test for any alternative* μ *by*

$$\Pr[T_n \geq c] = 1 - \Phi\left(c - \frac{\mu}{\sigma/\sqrt{n}}\right).$$

```
> # Critical Values
> c <- qnorm(.95) #for alpha=0.05
> c
[1] 1.644854
> # Size of test,
> 1 - pnorm(c)
[1] 0.05
> # Suppose n=10, sigma=0.15
> a<-0.15/sqrt(10)
> # Power of test against mu=0.05
> 1 - pnorm(c -0.05 /a)
[1] 0.2773403
> # Power of test against mu=0.10
> 1 - pnorm(c -0.10 /a)
[1] 0.6784366
```

We can make these size and power calculations using ***R****. Notice that the power of the test depends depends on how far the alternative is from the null, with higher power against alternatives further from the null. For example, suppose* $n = 10, \sigma = 0.15$*. Then the the power of the test against alternative* $\mu = 0.05$ *is only* 0.28*, while the power of the test against alternative* $\mu = 0.10$ *is* 0.68*.*

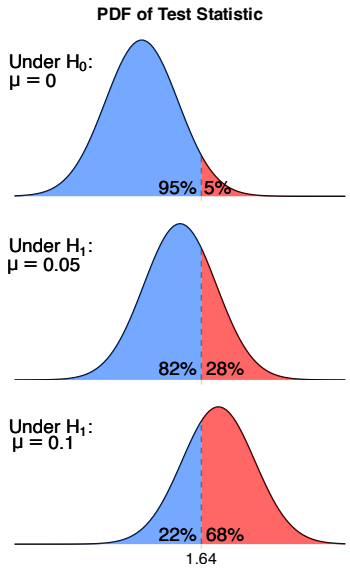


Figure 2: Example 2, Testing $\mathbb{H}_0 : \mu = 0$ vs $\mathbb{H}_1 : \mu > 0$, with test with size 0.05 that fails to reject when $T_n \leq 1.64$ and rejects when $T_n > 1.64$. Plotting example with $n = 10$ and $\sigma = 0.15$. Notice that higher μ shifts the distribution of T_n to the right and results in higher probability of $T_n > 1.64$.

Remark 2 (Size vs Level)**.** *Choosing a critical value following Remark 1 by the appropriate quantile of the test statistic under the null results in a test with the size as close as possible to the significance level while being weakly less than the significance level. If the test statistic has a continuous distribution under the null, as in Example 2, then the size of the resulting test will equal the significance level. In contrast, if the test statistic has a discrete distribution under the null, as in Example 1, then it will typically not be possible to chose a critical value such that the size equals the chosen significance level, and choosing a critical value by the appropriate quantile typically results in a test with size strictly less than the significance level.*

As illustrated by Figure 3 for Example 2, tests typically have higher power against alternatives further from the null for any given sample size, while having higher power the larger the sample size for any given alternative. One implication is that we should expect a test to accept the null even when the null is false if the alternative is sufficiently close to the null given the sample size. Another implication is that we should expect to be able to reject relatively small deviations from the null if the sample size is sufficiently large. If, for a given sample size, the test is unlikely to be able to correctly reject the null when the alternative of interest is true, we say that the test is *underpowered*, and failing to reject the null is not particularly strong evidence in favor of the null. For example, in Ex 2, suppose $n = 10$ and we are interested in the alternative $\mu = 0.05$, then the test with significance level 0.05 only has power 0.28 against that alternative and we would expect to be unable to correctly reject the null at that deviation from the null (see Figure 3, top). One response to a test being underpowered is to collect more data. For example, if we wish

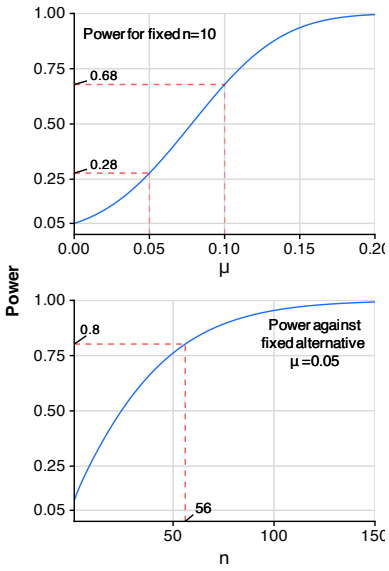


Figure 3: Example 2 with $\sigma = 0.15$, power for fixed $n = 10$ as a function of the alternative (top), and power against fixed alternative $\mu = 0.05$ as a function of sample size n (bottom).

to have power at least 0.80 against alternative $\mu = 0.05$, we would need at least $n = 56$ years of data instead of 10 years of data (see Figure 3, bottom). These issues are especially important in experiment design, where researchers choose sample size based on such power calculations, to make sure that they have sufficient power against alternatives of interest.

As a result of such considerations, researchers sometimes distinguish between *statistical significance* (rejecting the null hypothesis at the given significance level) versus *economic or substantive significance*. For example, in Example 2 with $\sigma = 0.15$, suppose the researcher observes $n = 10$ years of data, computes $\bar{X}_{10} = 0.05$ so that $T_{10} = 1.05$, in which case we estimate that the fund manager's expected excess return is 0.05 which is of economic significance, but we can't reject the null of zero expected excess return at the 0.05-level (not statistically significant). In contrast, suppose we somehow could observe $n = 10,000,000$ observations on the performance of her fund, and computed $\bar{X}_{10,000,000} = 0.0001$ so that $T_{10,000,000} = 2.11$, in which case we estimate that the fund manager's expected excess return is 0.0001 which is not meaningfully different from zero (not economically significant), but we can reject the null of zero expected excess return at the 0.05-level (is statistically significant). The larger the sample size, the more likely tests are able to detect even trivially small deviations from the null.

Stephen Ziliak and Deidre McCloskey have made these points particularly forcibly in a series of publications, arguing against what they call "asterisk econometrics". See, e.g., Ziliak and McCloskey (2004), "Size matters: the standard error of regressions in the American Economic Review".

Misconception: Accepting $\mathbb{H}_0$ as Evidence $\mathbb{H}_0$ is True

A common misconception is to interpret acceptance of the null hypothesis as evidence that the null is true. Such an interpretation is only warranted if the power of the test against alternatives of interest is high. If the test has low power against an alternative of interest, then it is unlikely to reject the false null when that alternative it true, and we should expect the test to incorrectly accept the null even when that alternative is true instead.

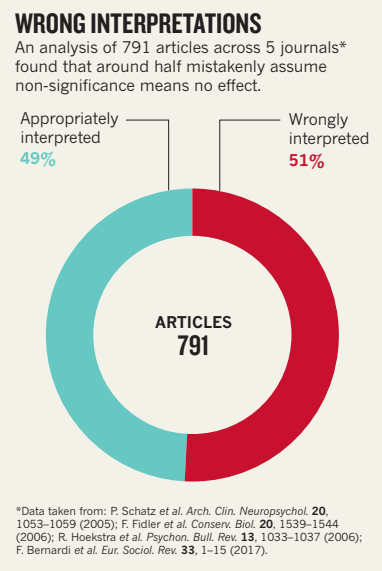


Figure 4: Figure from Amrhein et al. (2019), who provide a particularly cogent discussion of this misconception.

Other Examples

Thus far, in Examples 1 and 2, we considered one particular null hypothesis ($p = 0.5$ in Examples 1 and $\mu = 0$ in Example 2), and one-sided (greater) alternatives. We can easily adapt the analysis to other null hypotheses and other alternatives. For example, consider Example 2, but with $\mathbb{H}_0 : \mu = \mu_0$, $\mathbb{H}_1 : \mu > \mu_0$ for some μ_0, where our previous analysis was for the special case of $\mu_0 = 0$. Then,

it is natural to construct our test statistic as

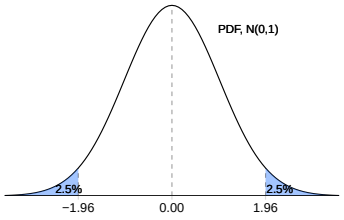
$$T_n \equiv \frac{\bar{X}_n - \mu_0}{\sigma/\sqrt{n}}.$$

Under the null $\mu = \mu_0$, this test statistic is distributed $N(0,1)$, and all of the previous analysis goes through unchanged. For example, if we wished to test at the 0.05 level the null that the fund's expected excess return is 0.01 against the alternative that it is greater than 0.01, we would construct our test statistic as $T_n = \frac{\bar{X}_n - 0.01}{\sigma/\sqrt{n}}$, and we would reject the null when T_n is greater than the .95 quantile of $N(0,1)$. Now suppose in Example 2 that we wished to test the null that $\mathbb{H}_0 : \mu = \mu_0$ vs $\mathbb{H}_1 : \mu < \mu_0$. We would redefine the test statistic as $-T_n \equiv -\frac{\bar{X}_n - \mu_0}{\sigma/\sqrt{n}}$ so that larger values are more evidence against the null in favor of the alternative $\mu < \mu_0$, but otherwise proceed as before.

Now suppose in Example 2 that, as if often the case in practice, we wish to test the null that $\mathbb{H}_0 : \mu = \mu_0$, vs the two-sided alternative $\mathbb{H}_1 : \mu \neq \mu_0$. We now redefine the test statistic as

$$|T_n| = \left| \frac{\bar{X}_n - \mu_0}{\sigma/\sqrt{n}} \right|,$$

so that values of T_N far above or far below 0 are evidence against the null in favor of the alternative. We would again proceed as before except that we need to adjust the critical values. Under the null hypothesis, $T_n \sim N(0,1)$, so that $|T_n| \sim |\, N(0,1) \,|$, i.e., distributed as the absolute value of a standard normal.[2] Thus, for a test with significance level α, we would reject the null hypothesis $\mu = \mu_0$ in favor of $\mu \neq \mu_0$ when $T_n > q_{1-\alpha/2}$ where $q_{1-\alpha/2}$ is the $1 - \alpha/2$ quantile of a standard normal. For example, for a test with level $\alpha = 0.05$, we would use the critical value $c = 1.96$ which is the 0.975 quantile of a standard normal and thus the 0.95 quantile of the absolute value of a standard normal.



[2] As discussed in Review Handout 2, Remark 4, this results follow from symmetry of the standard normal distribution. Note that, under the null, $T_N \sim N(0,1)$, so that

$$\Pr[T_N < -1.96] = \Pr[T_N > 1.96] = 0.025,$$

$$\begin{aligned} \Rightarrow \Pr[|T_N| > 1.96] &= \Pr[T_N < -1.96] + \Pr[T_N > 1.96] \\ &= 0.05. \end{aligned}$$

p-values

In addition to (or instead of) reporting the results of hypothesis tests, researchers often report p-values. p-values answer the question: if the null hypothesis is true, how likely are we to observe as much evidence against the null as what we actually observe? In particular, we use the test statistic to measure strength of evidence against the null, and so the p-value answers the question: if the null hypothesis is true, how likely are we to observe a test statistic as least as large as what we actually observe?

Definition 9: p-values

The *p-value* is the probability under the null of observing a test statistic as large or larger as the one observed in the sample.

Note that the test statistic is a random variable since it depends on the random sample, and thus the p-value is a random variable as it is a function of the random test statistic. The smaller the p-value, the less likely we are to observe that much evidence against the null if the null was true, and thus the more evidence against the null. Recall that to construct a test with significance level α we reject the null only when, if the null were true, we would observe that much evidence against the null with probability α or less. We can thus use p-values to implement tests:

Theorem 1:

Suppose p_n is a p-value for the null hypothesis $\mathbb{H}_0$. Then the test that rejects the null whenever $p_n \leq \alpha$ has significance level α.

Thus, for example, if we observe a p-value of 0.06, we would know that we would not reject the null hypothesis at significance level 0.05 but would reject the null at level 0.10. Reporting p-values thus allows the reader to know whether the null would be rejected at any given α, and thus provides more information than simply reporting whether the null hypothesis would be rejected at one given significance level.

Example 1. *(continued) Recall that, in this example,* $\mathbb{H}_0 : p = 0.5$ *vs* $\mathbb{H}_1 : p > 0.5$*, with test statistic* $T_n = S_n$*,* $S_n \sim$ *Binomial*$(n, 0.5)$ *under the null, so the p-value is the probability that a Binomial*$(n, 0.5)$ *takes a value greater than or equal to the value of* S_n *we observe in the data. For example, suppose we observe the manager's fund outperforms the market benchmark in* 9 *years out of* 10*, so that* $S_{10} = 9$*, and thus the p-value is the probability that a Binomial*$(10, 0.5)$ *variable takes a value greater than or equal to* 9*. From discreteness of the Binomial distribution,* $\Pr[S_{10} \geq 9] = \Pr[S_{10} > 8] = 1 - \Pr[S_{10} \leq 8]$*, which we can compute using* ***R*** *by* `1-pbinom(8,10,0.5)` *resulting in a p-value of* 0.011*. We would conclude that we would observe as much or more evidence against the null* $p = 0.5$ *as what we actually observe* 1.1% *of the time if the null were true, and thus we reject the null at the* 0.05 *level but not quite at the* 0.01 *level. Notice that the p-value depends on the null. For example, if our null was that* $p = 0.7$*, we would have that* $S_n \sim$ *Binomial*$(n, 0.7)$

under the null. In that case, with $S_{10} = 9$, we could compute the p-value by `1-pbinom(8,10,0.7)` *resulting in a p-value of* 0.15 *and we would conclude that we cannot reject the null $p = 0.7$ at the* 0.10 *level.*

Example 2. *(continued) Recall that, in this example, $\mathbb{H}_0 : \mu = 0$ vs $\mathbb{H}_1 : \mu > 0$, with test statistic $T_n \equiv \frac{\bar{X}_n}{\sigma/\sqrt{n}}$, with $T_n \sim N(0,1)$ under the null, where we are taking σ to be known. Then the p-value is the probability that a standard normal variable takes a value greater than or equal to the value of T_n that we observe, $1 - \Phi(T_n)$. For example, suppose over a* 10 *year period we observe that the fund's average excess return is* 0.06*, and suppose we know that $\sigma = 0.15$, so that we observe*

$$T_n = \frac{0.06}{0.15/\sqrt{10}} = 1.26.$$

*In this case, the p-value is $1 - \Phi(1.26)$, which we can compute using **R** by* `1-pnorm(1.26)` *resulting in a p-value of* 0.104*. We would conclude that we would observe as much evidence against the null as what we actually observe* 10.4% *of the time, and thus we cannot quite reject the null at the* 0.10 *level. Notice that the p-value depends on the null. For example, if our null was that $\mu = 0.05$, we would have $T_n = 0.21$, and we could compute p-value by* `1-pnorm(0.21)` *resulting in a p-value of* 0.58*. In that case, we would conclude that more than half the time we would observe more evidence against the null $\mu = 0.05$ than we actually observed, and we would not reject the null at any standard significance level.*

In Example 2, if we consider the two sided alternative, $\mathbb{H}_1 : \mu \neq 0$, then we choose the test statistic $|T_n|$ which is distributed $|\, N(0,1) \,|$ under $\mathbb{H}_0$, resulting in p-values of the form

$$p_n = 2 \cdot (1 - \Phi(T_n)).$$

See Review Handout 2, Remark 4 for a review of the $|\, N(0,1) \,|$ distribution.

Remark 3 (Default p-values)**.** ***R*** *(as well as STATA, SAS, etc) typically reports p-values by default for the null hypothesis that the parameter of interest equals* 0 *(or sometimes a different default value, such as $p = 0.5$ for tests specifically for binary random variables) versus a two-sided alternative. Such nulls/alternatives are not always of interest. For example, the null hypothesis that average earnings is zero is typically not of interest, nor is the null that the probability of being a CEO is a half. Likewise, in regression models, say regressing consumption on income, the null that the coefficient on income is zero is not of interest. However, it is in some contexts standard to report such p-values in empirical work (particularly for regression coefficients), and your readers may expect you to do so, but you need not spend time focused on such p-values.*

Misconception: p-values as $\Pr\{\mathbb{H}_0$ is true.$\}$

A common misconception is to interpret the p-value as the probability that the null hypothesis is true. That interpretation is incorrect. The p-value is the probability that we would observe as much evidence against the null (a value of the test statistic at least as large) as what we actually observe if the null hypothesis were true.

Confidence Intervals

Instead of reporting p-values or the results of a specific hypothesis test (or in addition to reporting them), researchers often report confidence intervals:

Definition 10: Confidence Interval

Consider the interval $\mathbf{C}_n = [L_n, U_n]$ where L_n and U_n can be computed by the sample of size n. Then $\mathbf{C}_n$ is called a valid $1-\alpha$ *confidence interval* for θ if

$$\mathbb{P}_\theta[\theta \in \mathbf{C}_n] \geq 1-\alpha \quad \text{for all} \quad \theta.$$

A $1-\alpha$ confidence interval is an interval that can be computed from the random sample and which, for any possible value of the parameter, contains that value with probability at least $1-\alpha$ if that value is the true parameter value. We previously saw that choosing a smaller α in hypothesis testing results in a smaller probability of incorrectly rejecting the null if the null is true at the cost of a smaller probability of correctly rejecting the null when the null is false. Likewise, choosing a smaller α for a $1-\alpha$ confidence interval increases the probability that the CI contains the true parameter value at the cost of a wider interval. Just as economists typically choose $\alpha = 0.01, 0.05$, or 0.10 for hypothesis testing (with $\alpha = 0.05$ most comment), they typically choose $1-\alpha = 0.99, 0.95$, or 0.90 for confidence intervals (with $1-\alpha = .95$ most common).

There is a deep connection between hypothesis testing and confidence intervals. A test of the null $\mathbb{H}_0 : \theta = \theta_0$ at level α will falsely reject the null with probability at most α when the θ_0 is the true parameter value. A confidence interval $\mathbf{C}_n$ of level $1-\alpha$ will include the value θ_0 with probability at least $1-\alpha$ when θ_0 is the true value, or, in other words, will exclude the value θ_0 with probability at most α when θ_0 is the true value. These thoughts lead to test inversion:

Theorem 2: Test Inversion

Any $1-\alpha$ confidence interval can be used to construct an α level hypothesis test, and any α-level hypothesis test can be used to construct a $1-\alpha$ confidence interval. In particular:

- Let $\mathbf{C}_n$ denote a $1-\alpha$ confidence interval. Then, for any θ_0, the test that rejects the null $\mathbb{H}_0 : \theta = \theta_0$ when $\theta_0 \notin \mathbf{C}_n$ and fails to reject when $\theta_0 \in \mathbf{C}_n$ is a level α test.
- Consider a level α test of $\mathbb{H}_0 : \theta = \theta_0$ for each possible θ_0. Then the interval that contains all parameter values θ_0 not rejected by the test and excludes all parameter values rejected by the test is a $1-\alpha$ confidence interval.

Example 1. *(continued) Consider Example 1 with null hypotheses:* $\mathbb{H}_0 : p = p_0$ *vs* $\mathbb{H}_1 : p > p_0$*, with test statistic* $T_n = S_n$*,* $S_n \sim$ *Binomial*(n, p_0) *under the null, so the p-value is the probability that a Binomial*(n, p_0) *takes a value greater than or equal to the value of* S_n *we observe in the data. Consider an* $\alpha = 0.05$ *level test, which we can invert to a* $1-\alpha = 0.95$ *CI. Thus, for each* p_0*, we can calculate the corresponding p-value for that null, and we reject the null when that p-value is less than* 0.05*, and the set of all* p_0 *for which we don't reject the null is a 95% CI on* p*. For example, suppose* $n = 10$ *and* $S_{10} = 9$*. To determine whether* 0.5 *should be in our confidence interval, we can compute the p-value by* `1-pbinom(8,10,0.5)` *which returns a value of* 0.011*, so that we reject the null* $p = 0.5$ *and thus exclude* 0.5 *from our 95% CI. In contrast,* `1-pbinom(8,10,0.7)` *returns* 0.149 *so that we fail to reject that null and we include* 0.7 *in our CI. Further, testing the null* $p = p_0$ *for any* $p_0 < 0.6058$ *results in a p-value less than* 0.05 *and is thus rejected at the* 0.05 *level, while testing the null* $p = p_0$ *for any* $p_0 > 0.6058$ *results in a p-value greater than* 0.05 *and is thus accepted at the* 0.05 *level, and we conclude that the interval* $[0.6058, 1]$ *is a valid* 95% *CI on* p*.*

Example 2. *(continued) Suppose we wish to invert a test the null* $\mathbb{H}_0 : \mu = \mu_0$*, vs the alternative* $\mathbb{H}_1 : \mu > \mu_0$*. We have the test statistic as*

$$T_n = \frac{\bar{X}_n - \mu_0}{\sigma/\sqrt{n}},$$

with $T_n \sim N(0,1)$ *under the null. Consider an* $\alpha = 0.05$ *level test, which we can invert to a* $1-\alpha = 0.95$ *CI. We fail to reject the null* $\mu = \mu_0$ *at the* 0.05 *level whenever the corresponding p-value* $1-\Phi(T_N)$ *is greater than* 0.05*, equivalently whenever* $|T_N|$ *is less than* 1.64*, the* 0.95 *quantile of a*

$N(0,1)$. *We thus fail to reject the null for any* μ_0 *such that*

$$\frac{\bar{X}_n - \mu_0}{\sigma/\sqrt{n}} \leq 1.64,$$

which we can rewrite to say that we fail to reject the null $\mathbb{H}_0 : \mu = \mu_0$ *at the* 0.05 *level for any* μ_0 *such that*

$$\bar{X}_n - 1.64\frac{\sigma}{\sqrt{n}} \leq \mu_0.$$

We thus conclude that

$$\mathbf{C}_n = \left[\bar{X}_n - 1.64\frac{\sigma}{\sqrt{n}},\ \infty\right]$$

is a valid 95% *CI on* μ. *For example, suppose over a* 10 *year period we observe that the fund's average excess return is* 0.06, *and suppose we know that* $\sigma = 0.15$, *then our* 95% *confidence interval is* $[-0.02,\ \infty]$.

Example 2. *(continued, two sided alternative) Again consider example 2, but now consider a test of the null* $\mathbb{H}_0 : \mu = \mu_0$, *vs the two-sided alternative* $\mathbb{H}_1 : \mu \neq \mu_0$. *Redefine the test statistic to be*

$$|T_n| = \left|\frac{\bar{X}_n - \mu_0}{\sigma/\sqrt{n}}\right|,$$

where $T_n \sim N(0,1)$ *under the null. Consider an* $\alpha = 0.05$ *level test, which we can invert to a* $1 - \alpha = 0.95$ *CI. We fail to reject the null* $\mu = \mu_0$ *at the* 0.05 *level whenever the corresponding p-value* $2 \cdot (1 - \Phi(T_N))$ *is greater than* 0.05, *equivalently whenever* $|T_N|$ *is greater than* 1.96, *the* 0.975 *quantile of a* $N(0,1)$. *We thus fail to reject for any* μ_0 *such that*

$$-1.96 \leq \frac{\bar{X}_n - \mu_0}{\sigma/\sqrt{n}} \leq 1.96,$$

i.e., for any μ_0 *such that*

$$\bar{X}_n - 1.96\frac{\sigma}{\sqrt{n}} \leq \mu_0 \leq \bar{X}_n + 1.96\frac{\sigma}{\sqrt{n}}.$$

We thus conclude that

$$\left[\bar{X}_n - 1.96\frac{\sigma}{\sqrt{n}},\ \bar{X}_n + 1.96\frac{\sigma}{\sqrt{n}}\right]$$

is a valid 95% *CI on* μ. *For example, suppose over a* 10 *year period we observe that the fund's average excess return is* 0.06, *and suppose we know that* $\sigma = 0.15$, *then our* 95% *confidence interval is* $[-0.03,\ 0.15]$.

Remark 4 (Upper/Lower/Two-Sided CIs)**.** *When we invert a test to form a confidence interval, the form of the CI depends on the test's alternative. Notice that, in Example 2, when we invert a* 0.05 *level test of*

$\mathbb{H}_0 : \mu = \mu_0$ vs the one-sided (greater) alternative $\mathbb{H}_1 : \mu > \mu_0$, we obtain the 95% CI: $[\bar{X}_n - 1.64\frac{\sigma}{\sqrt{n}}, \infty]$. A confidence interval of this form is called an upper one-sided confidence interval. However, when we invert a 0.05 level test of $\mathbb{H}_0 : \mu = \mu_0$ vs the two-sided alternative $\mathbb{H}_1 : \mu \neq \mu_0$, we obtain the 95% CI: $[\bar{X}_n - 1.96\frac{\sigma}{\sqrt{n}}, \bar{X}_n + 1.96\frac{\sigma}{\sqrt{n}}]$. A confidence interval of this form is called a two-sided confidence interval. If instead we invert a 0.05 level test of $\mathbb{H}_0 : \mu = \mu_0$ vs the one-sided (less) alternative $\mathbb{H}_1 : \mu < \mu_0$ we obtain the 95% CI: $[-\infty, \bar{X}_n + 1.64\frac{\sigma}{\sqrt{n}}]$. A confidence interval of this form is called a lower one-sided confidence interval. All three confidence intervals are valid, just based on inverting tests with different alternatives.

Normal example with unknown σ

In Example 2, we assumed that $X_1, X_2, ..., X_n$ are i.i.d. with $X_i \sim N(\mu, \sigma^2)$, and further assumed that σ^2 is known. Because we assumed that σ^2 is known, we could calculate $T_n = \frac{\bar{X}_n - \mu_0}{\sigma/\sqrt{n}}$, which is distributed $N(0,1)$ under the null that $\mu = \mu_0$. Now suppose, as is the case in practice, that we do not know σ^2. If we do not know σ^2 we cannot calculate T_n and thus cannot use it as a test statistic. However, we can estimate σ^2. Thus, redefine $T_n = \frac{\bar{X}_n - \mu_0}{s_n/\sqrt{n}}$ where $s_n^2 = \frac{1}{n-1}\sum_{i=1}^n (X_i - \bar{X}_n)^2$. Then, by Review Handout 2, Theorem 6, with this redefinition, $T_n \sim t_{n-1}$ under the null. We can then redo the previous analysis, using as critical values the quantiles of the t_{n-1} distribution instead of the $N(0,1)$ distribution, and using p-values based on the t_{n-1} CDF instead of the $N(0,1)$ cdf. All analysis is otherwise unchanged. Recall the discussion from Review Handout 2 that a t_{n-1} distribution has heavier tails, is more likely to take extreme values, than a $N(0,1)$ distribution. In the present context, these heavier tails are generated by the random chance that our estimated s.d. is close to zero, which is especially likely when n is small. An implication of these heavier tails is that the critical values relevant for testing will be larger, especially when n is small, compared to the critical values based on normality that would be valid if we were dividing by true σ. However, the difference between those critical values and those of a normal distribution vanishes at n goes to infinity.

Quantiles of t_ν

ν	0.90	0.95	0.975
1	3.08	6.31	12.71
2	1.89	2.92	4.30
3	1.64	2.35	3.18
4	1.53	2.13	2.78
5	1.48	2.02	2.57
6	1.44	1.94	2.45
7	1.41	1.89	2.36
8	1.40	1.86	2.31
9	1.38	1.83	2.26
10	1.37	1.81	2.23
20	1.33	1.72	2.09
∞	1.28	1.64	1.96

*Implementation in **R***

We have considered two classes of examples for inference: inference on probability of success for a binomial distribution (Example 1), and inference on mean of i.i.d. normal observations with known or unknown variance (Example 2). We can implement hypothesis

testing, p-values, and confidence intervals in these examples using `pbinom` and `qbinom` for the binomial example, using `pnorm` and `qnorm` for the normal mean example with known variance, and using `pt` and `qt` for the normal mean example with unknown variance.

These **R** functions are covered in Review Handout 2.

There are **R** functions specifically for inference including for these examples. For inference on probability of success in the binomial example, `binom.test` is convenient, and especially makes inference versus a two-sided alternative far easier to implement than using `pbinom` and `qbinom` directly. Likewise, for inference on mean of i.i.d. normal observations with unknown variance, using `t.test` is convenient as opposed to implementing inference with `pt` and `qt` directly.

You can learn more about the `binom.test` and `t.test` functions in **R** using the `help()` function or equivalently the `?` help operator, for example, by entering `help(binom.test)` or `?t.test`.

Summary

Table 1 provides optional reading for this handout.[3]

Source	Chapters
Hogg et. al (2019)	7.1, 8.1

Table 1: Optional reading

[3] Hogg, R. V., E. A. Tanis, and D. L. Zimmerman (2020). *Probability and statistical inference* (10 ed.). Pearson

Important Definitions

Def 1: The **null hypothesis** $\mathbb{H}_0$ is the hypothesis we accept unless there is sufficiently strong evidence against it in favor of the **alternative hypothesis** $\mathbb{H}_1$. We consider null hypotheses of the form $\mathbb{H}_0 : \theta = \theta_0$, vs alternative hypotheses of the form:

- $\mathbb{H}_1$ $\theta \neq \theta_0$ **two-sided alternative**,
- $\mathbb{H}_1 : \theta > \theta_0$ (**one-sided alternative, greater**), or
- $\mathbb{H}_1 : \theta < \theta_0$ (**one-sided alternative, less)**.

Def 2: A **test statistic** T_n is a function of our random sample that summarizes evidence against the null, with larger values indicating stronger evidence against the null.

Def 3: A **hypothesis test** is a decision rule that specifies a test statistic T_n and **critical value** c such that, if $T_n > c$, we reject the null $\mathbb{H}_0$ in favor of the alternative $\mathbb{H}_1$, and otherwise we accept the $\mathbb{H}_0$.

Def 4: A **Type I error** is incorrectly rejecting the null hypothesis when the null is true.

Def 5: A **Type II error** is failing to reject the null hypothesis when the null is false.

Important Definitions (cont'd)

Def 6: The **size** of a test is the probability of a Type I error when the null hypothesis is true, and is given by $Pr_{\theta_0}[T > c]$.

Def 7: Then the **power** of a test against alternative θ is one minus the probability of a Type II error when θ is the true parameter value, and is given by $Pr_{\theta}[T > c]$.

Def 8: A test has **significance level** α if its size is less than or equal to α.

Def 9: The **p-value** is the probability under the null of observing a test statistic as large or larger as the one observed in the sample.

Def 10: The interval $\mathbf{C}_n = [L_n, U_n]$ is a valid $1-\alpha$ **confidence interval** for θ if $\mathbb{P}_\theta[\theta \in \mathbf{C}_n] \geq 1-\alpha$ for all θ.

Important Results

Remark 1: Let $q_{1-\alpha}$ denote the $1-\alpha$ quantile of the distribution of the test statistic under the null. Then the test that rejects the null when $T_n > q_{1-\alpha}$ is of level α.

Thm 1: Suppose p_n is a p-value for the null hypothesis $\mathbb{H}_0$. Then the test that rejects the null whenever $p_n \leq \alpha$ has significance level α.

Thm 2: Test Inversion:

- Let $\mathbf{C}_n$ denote a $1-\alpha$ confidence interval. Then, for any θ_0, the test that rejects the null $\mathbb{H}_0 : \theta = \theta_0$ when $\theta_0 \notin \mathbf{C}_n$ and fails to reject when $\theta_0 \in \mathbf{C}_n$ is a level α test.
- Consider a level α test of $\mathbb{H}_0 : \theta = \theta_0$ for each possible θ_0. Then the interval that contains all parameter values θ_0 not rejected by the test and excludes all parameter values rejected by the test is a $1-\alpha$ confidence interval.

Steps for Hypothesis Testing

1. State $\mathbb{H}_0$ and $\mathbb{H}_1$ hypotheses;
2. Decide significance level α;
3. Formulate test statistic, determine distribution of test statistic under the null;
4. Either:
   (a) Choose $1-\alpha$ quantile of distribution of test statistic under the null as critical value, reject null at level α if test statistic bigger than this critical value, or
   (b) Construct p-value as probability under the null of observing a test statistic at least as large as the one observed in the sample, reject null at level α if p-value less than α.

Normal example with known σ^2:

Suppose $X_1, X_2, ..., X_n$ are i.i.d. with $X_i \sim N(\mu, \sigma^2)$ and σ^2 known. Define $T_n = \frac{\bar{X}_n - \mu_0}{\sigma/\sqrt{n}}$. Let $q_{1-\alpha}$ denote the $1-\alpha$ quantile of a $N(0,1)$, and let c_α denote the critical value for a level α test. Let F_ν denote the cdf of t_ν distribution. Then:

$\mathbb{H}_0$	$\mathbb{H}_1$	Test Stat.	c_α	p-value	$1-\alpha$ CI
$\mu = \mu_0$	$\mu > \mu_0$	T_n	$q_{1-\alpha}$	$1-\Phi(T_n)$	$\left[\bar{X}_n - q_{1-\alpha}\frac{\sigma}{\sqrt{n}}, \infty\right]$
$\mu = \mu_0$	$\mu < \mu_0$	$-T_n$	$q_{1-\alpha}$	$\Phi(T_n)$	$\left[-\infty, \bar{X}_n + q_{1-\alpha}\frac{\sigma}{\sqrt{n}}\right]$
$\mu = \mu_0$	$\mu \neq \mu_0$	$\lvert T_n \rvert$	$q_{1-\alpha/2}$	$2(1-\Phi(T_n))$	$\left[\bar{X}_n - q_{1-\alpha/2}\frac{\sigma}{\sqrt{n}}, \bar{X}_n + q_{1-\alpha/2}\frac{\sigma}{\sqrt{n}}\right]$

Normal example with unknown σ^2:

Suppose $X_1, X_2, ..., X_n$ are i.i.d. with $X_i \sim N(\mu, \sigma^2)$ and σ^2 unknown. Define $T_n = \frac{\bar{X}_n - \mu_0}{s_n/\sqrt{n}}$ where $s_n^2 = \frac{1}{n-1}\sum_{i=1}^n (X_i - \bar{X}_n)^2$. Let $q_{\nu,1-\alpha}$ denote the $1-\alpha$ quantile of a t_ν, and let c_α denote the critical value for a level α test. Then:

$\mathbb{H}_0$	$\mathbb{H}_1$	Test St.	c_α	p-value	$1-\alpha$ CI
$\mu = \mu_0$	$\mu > \mu_0$	T_n	$q_{\nu,1-\alpha}$	$1-F_\nu(T_n)$	$\left[\bar{X}_n - q_{\nu,1-\alpha}\frac{\sigma}{\sqrt{n}}, \infty\right]$
$\mu = \mu_0$	$\mu < \mu_0$	$-T_n$	$q_{\nu,1-\alpha}$	$F_\nu(T_n)$	$\left[-\infty, \bar{X}_n + q_{\nu,1-\alpha}\frac{\sigma}{\sqrt{n}}\right]$
$\mu = \mu_0$	$\mu \neq \mu_0$	$\lvert T_n \rvert$	$q_{\nu,1-\alpha/2}$	$2(1-F_\nu(T_n))$	$\left[\bar{X}_n - q_{\nu,1-\alpha/2}\frac{\sigma}{\sqrt{n}}, \bar{X}_n + q_{\nu,1-\alpha/2}\frac{\sigma}{\sqrt{n}}\right]$

Remark 5 (Normal Example, Testing with One-Sided Lower Alternatives)**.** *For the normal example with known σ^2, note that, by symmetry of the normal distribution, taking the test statistic as $-T_n$ and rejecting when $-T_n > q_{1-\alpha}$ is equivalent to rejecting when $T_N < -q_{1-\alpha} = q_\alpha$. Likewise, by symmetry of the normal distribution, the p-value is given by $1-\Phi(-T_n) = \Phi(T_n)$. Parallel statements hold for testing with unknown σ^2 by symmetry of the t-distribution.*

Self-Study Questions

1. Recall Review Handout 2, Self-Study Q4. Let X_t denote a indicator variable for a flood occurring in Houston in year t of such severity that, based on historical data, such a flood is considered a five hundred year flood. Consider Houston's record of floods from 2015 to 2022, $X_{2015}, ..., X_{2022}$, and take X_t to be i.i.d. over time with $p = \Pr[X_t = 1]$. Over that 8-year period, there were five floods of severe enough to be considered five hundred year floods based on historical data, $\sum_{t=2015}^{2022} X_t = 5$. Consider the null hypothesis that a flood of that severity is still a five hundred year flood, $\mathbb{H}_0 : p = 0.002$, versus the alternative that due to climate change a flood of that severity now occurs with higher probability, $\mathbb{H}_1 : p > 0.002$.
   (a) Using **R**, test this null versus this alternative:
      i. What is your p-value?
      ii. Would you reject the null hypothesis at the 0.01 level?
   (b) Using **R**, construct an upper one-sided 0.99-level confidence interval on p.
      i. What is your CI?
      ii. Are the results from your CI consistent with your answers to Q 1a above?
2. Recall Review Handout 2, Self-Study Q6. Suppose you are interested in investing in a mutual fund B with return $r_{B,t}$. Suppose $r_{B,t} \sim N(\mu, \sigma^2)$ and that returns are i.i.d. over time. The fund's manager tells you that $\mu = 0.12$ and $\sigma^2 = 0.01$, and you believe what the manager tells you about the variance of returns though you worry that the actual expected return might be lower. Suppose you are interested in the asset's excess return above the risk-free rate r_f, where $r_f = 0.02$. In the following, take r_f to be a constant. Let $X_t = r_{B,t} - r_f$ denote the fund's excess return in year t. Suppose you observe excess returns over four years, $X_1, ..., X_4$, and take X_t to be i.i.d. over t. Consider testing the null that what the manager told you about the mean return is true versus the one sided alternative that the true mean return is smaller.
   (a) Formally state the null and alternative hypotheses.
   (b) What is the distribution of X_t under the null?
   (c) Consider the test maintaining that what the manager told you about the riskiness of the fund is true, i.e., taking as known that $\sigma^2 = 0.01$.
      i. What would be your test statistic?
      ii. What is the distribution of your test statistic under the null?
      iii. What would be your critical value?
      iv. Suppose that $\bar{X}_4 = 0$. Using **R**,
         A. what is the p-value for your test?
         B. would you reject the null at the 0.10 level?
         C. Construct the 0.90-level lower confidence interval on μ. Based on the CI, what range of value of μ would you not reject at the 0.10 level?
   (d) Suppose you worry that your test is under-powered, in that your test might be unlikely to reject the null even if what the manager told you is wrong because you only have four years of returns. Consider the particular alternative $\mu = -0.08$. For your test taking $\sigma^2 = 0.01$ known,
      i. what is the power of your test against the alternative $\mu = -0.08$?

ii. Given your power calculate above, how strongly does failure to reject the null provide evidence that the null is true.

iii. How would your answers to (i) and (ii) above change if you had $n = 36$ years of returns?