Hypothesis testing

– Math 161a, Spring 2019, San Jose State University

Prof. Guangliang Chen

May 7, 2019

Introduction

Consider the brown eggs problem again.

Suppose the weights of the eggs produced at the farm (population) are normally distributed with unknown mean μ but known standard deviation $\sigma = 2$ g.

It is claimed by the manufacturer that $\mu = 65$ g.

You bought a carton of 12 eggs, with an average weight of 61.5 g.

Question. Is such a difference purely due to randomness or significant evidence against the claim?

The formal procedure of hypothesis testing

First, we set up the following hypothesis test:

$$H_0 : \mu = 65 \quad \text{vs} \quad H_1 \, (\text{or } H_a) : \mu \neq 65$$

in which

- H_0: **null hypothesis** (statement which we intend to reject)
- H_1: **alternative hypothesis** (statement we suspect to be true)

The goal is to make a decision, based on a random sample $X_1, \ldots, X_n$ from the population, whether or not to reject H_0 so as to correspondingly establish H_1.

There are two kinds of decisions:

- If the sample "strongly" contradicts H_0, then we reject H_0 and correspondingly accept H_1;
- If the sample "does not strongly" contradict H_0, then we fail to reject H_0, or equivalently we **retain** H_0.

Remark. This is essentially a proof by contradiction approach.

Remark. There is a perfect analogy to **courtroom trial**. In this scenario, the following two hypotheses are tested:

- H_0: *Defendant is innocent*;
- H_a: *Defendant is guilty*.

The prosecutor presents evidence to the court, examined by the jury:

- If the jury thinks the evidence is strong enough (significant), the defendant will be convicted (H_0 is rejected and H_a is then accepted);
- Otherwise, the defendant is not found guilty and will be acquitted (the prosecutor has thus failed to convict the defendant due to insufficient evidence).

Remark. It is also possible to use a one-sided alternative:

$$H_0 : \mu = 65 \quad \text{vs} \quad H_a : \mu < 65.$$

In this case, the null is understood as "μ is *at least* 65 ($\mu \geq 65$)".

For example, the FDA's main interest is to know whether the eggs are lighter than 65 g (on average). It is not an issue if they are actually heavier (good for customers).

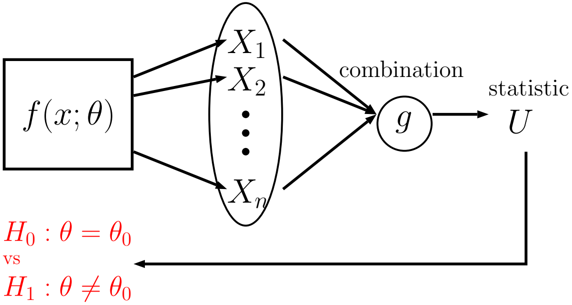
Similarly, for some other consideration, we might want to test

$$H_0 : \mu = 65 \quad \text{vs} \quad H_a : \mu > 65,$$

where the null is understood as "μ is *at most* 65 ($\mu \leq 65$)".

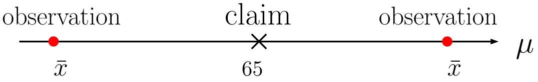
Test statistic

Typically, a test statistic needs to be specified to assist in making a decision. It is often a point estimator for the parameter being tested.

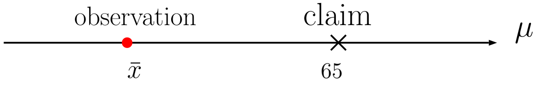


In the brown eggs example, we can use $\bar{X}$ as a test statistic to test $H_0 : \mu = 65$ against

- $H_1 : \mu \neq 65$: "very small or large" values of $\bar{X}$ are evidence against the null and correspondingly in favor of the alternative hypothesis.



- $H_1 : \mu < 65$: **only** "very small" values of $\bar{X}$ are evidence against the null and correspondingly in favor of the alternative hypothesis.

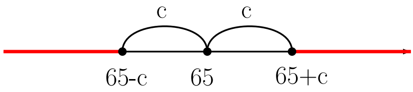


Decision rules

Clearly, a rule needs to be specified in order to decide **when to reject the null** $H_0 : \mu = 65$. This leads to a rejection region for the test.

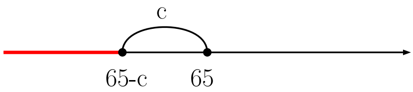
- For $H_1 : \mu \neq 65$:

  $$|\bar{x} - 65| > c$$



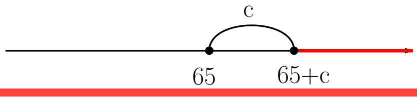
- For $H_1 : \mu < 65$:

  $$\bar{x} < 65 - c$$



- For $H_1 : \mu > 65$:

  $$\bar{x} > 65 + c$$



Test errors

There are two kinds of test errors depending on whether H_0 is true or not.

		Decision	
		Retain H_0	Reject H_0
H_0	**true**	Correct decision	Type I error
	false	Type II error	Correct decision

Remark. In the courtroom trial scenario, a type I error is convicting an innocent person, while a type II error is acquitting a guilty person.

Calculating the type-I error probability

Example 0.1. In the brown eggs problem, suppose the true population standard deviation is $\sigma = 2$ grams. A person decides to use the following decision rule (for a sample of size $n = 12$, i.e., a carton of eggs)

$$|\bar{x} - 65| > 1 \qquad \longleftarrow \text{rejection region of the test}$$

to conduct the two-sided test

$$H_0 : \mu = 65 \qquad \text{vs} \quad H_1 : \mu \neq 65.$$

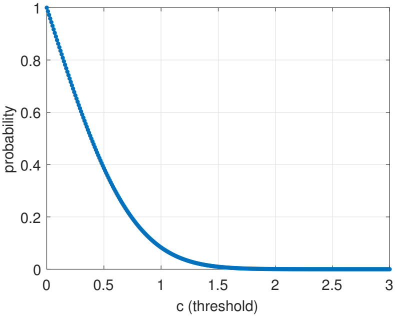
What is the probability of making a type-I error? (Answer: 0.0833)

Example 0.2. Redo the above example for a different decision rule:

$$|\bar{x} - 65| > 2$$

(Answer: 0.000532)

Type I error probabilities for using $|\bar{x} - 65| > c$ as rejection regions:



Observation. The larger the threshold (c), the smaller the rejection region (the harder to reject H_0), the smaller the type-I error.

Example 0.3. Redo the previous two examples for the one-sided test $H_1 : \mu < 65$ with corresponding decision rule

- $\bar{x} < 65 - 1 = 64$ $(\bar{x} - 65 < -1)$, or
- $\bar{x} < 65 - 2 = 63$ $(\bar{x} - 65 < -2)$

(Answers: $0.0416,\ 0.000266$)

Too easy, too good?

It seems that by increasing the threshold c (which would shrink the rejection region), we can make the type-I error probability arbitrarily small.

This seems a bit too easy and too good to be true.

This is indeed true, as far as only type-I error is concerned, but is this perhaps at the expense of something else?

How is the type-II error affected?

It turns out that reducing the rejection region will cause the probability of making a type-II error to increase:

- Making it hard to reject H_0 (by using a small rejection region) is good when H_0 is true (this corresponds to type-I errors).
- But it would be bad when H_0 is false (we actually want to reject H_0 in this case).

The thing is that we don't know which hypothesis is true, so we have to choose a rejection region carefully such that both errors are small.

Calculating the type-II error probabilities

Consider the two sided test

$$H_0 : \mu = 65 \qquad \text{vs} \quad H_1 : \mu \neq 65.$$

When $H_0 : \mu = 65$ is false (H_1 is correspondingly true), μ could be 64, or 68, or any other value.

Thus, there is a separate type-II error probability at each $\mu \neq 65$.

For any fixed decision rule $|\bar{x} - 65| > c$ (with c given), the probability of making a type-II error depends on the true value of μ:

$$\beta(\mu) = P(\text{Fail to reject } H_0 \mid H_0 \text{ false}) = P(|\bar{X} - 65| < c \mid H_1 \text{ true})$$

Remark.

- $1 - \beta(\mu)$ is the probability of making a correct decision by rejecting H_0 when it is false:

$$1 - \beta(\mu) = P(\text{Reject } H_0 \mid H_0 \text{ false}) = P(|\bar{X} - 65| > c \mid H_1 \text{ true})$$

- It is called the **power** (function) of the test.
- We would like
  - the type-II error probability $\beta(\mu)$ for a given μ to be small, and
  - the power of the test at the given μ to be large (80% or bigger).

Example 0.4. Consider the two-sided test:

$$H_0 : \mu = 65 \qquad \text{vs} \quad H_1 : \mu \neq 65$$

along with the following decision rule:

$$|\bar{x} - 65| > c.$$

Find the probability of making a type-II error when $\mu = 64$ for each value of $c = .5, 1, 2$.

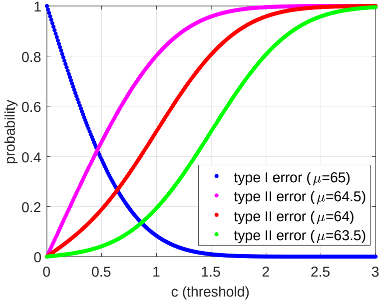
(Answer: $\beta(64) = P(|\bar{X} - 65| < c \mid \mu = 64) = 0.1886, 0.4997, 0.9584$)

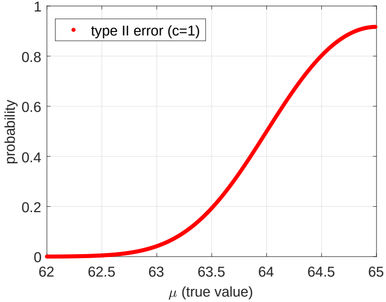
What about other values of $\mu \neq 65$?

Type-II errors	$c = 0.5$	1	2
$\mu = 63.5$	.0414	.1932	.8068
$\mu = 64$	.1886	.4997	.9584
$\mu = 64.5$	.4584	.8021	.9953

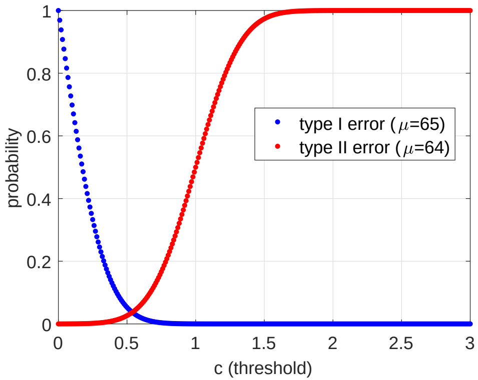
Observations.

- **For fixed value** μ: the larger c (the smaller the rejection region, and thus the harder to reject H_0), the larger the type-II error.
- **For fixed test** (c): the closer μ is to the value in H_0 (65), the larger the type II error.

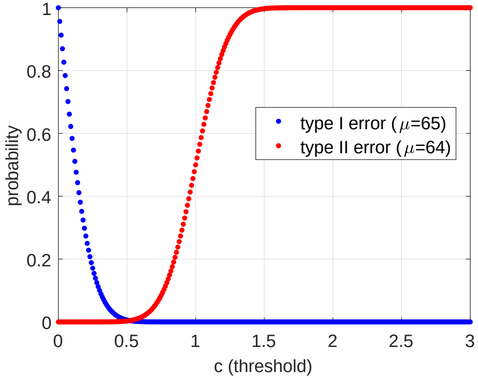

1
0.8
0.6
0.4
0.2
0
probability
0
0.5
1
1.5
2
2.5
3
c (threshold)
type I error (μ=65)
type II error (μ=64.5)
type II error (μ=64)
type II error (μ=63.5)


1
0.8
0.6
0.4
0.2
0
probability
62
62.5
63
63.5
64
64.5
65
μ (true value)
type II error (c=1)

How can you do better in both errors?



Increase the sample size!



How to control both errors together

Previously we assumed that both sample size n and test threshold c are fixed so as to evaluate the type-I and type-II errors of the test

$$H_0 : \mu = 65 \quad \text{vs} \quad H_a : \mu \neq 65$$

Here we consider the inverse design problem by assuming the two types of error probabilities are given first:

- type-I error probability α (called **level of the test**) $\longleftarrow$ typically 5%
- type-II error probability β (at specified location μ) $\longleftarrow$ typically 20%

and then trying to determine the required n and c as follows:

1. For the given level of the test i.e., α, solve

$$\alpha = P(\text{Reject } H_0 \mid H_0 \text{ true}) = P(|\bar{X} - 65| > c \mid \mu = 65)$$

to determine the required threshold c (dependent on n).

2. Choose sample size n to achieve type-II error probability β at given location μ ($\mu \neq 65$):

$$\beta = P(\text{Fail to reject } H_0 \mid H_0 \text{ false}) = P(|\bar{X} - 65| < c \mid \mu)$$

Example 0.5. Assume the setting of the brown eggs example (with known $\sigma = 2$, but sample size n TBD). Consider the following two-sided test

$$H_0 : \mu = 65 \quad \text{vs} \quad H_a : \mu \neq 65$$

with decision rule

$$|\bar{x} - 65| > c$$

Choose n, c so that the test has level 5% and power 80% (at $\mu = 64$).

Answer: $c = z_{\alpha/2}\frac{\sigma}{\sqrt{n}} = 0.693,\ n \approx \left(\frac{\sigma(z_{\alpha/2}+z_\beta)}{\mu_0-\mu'}\right)^2 = 32$

Connection to confidence intervals

In the last example, the rejection region (for $\alpha = 5\%$) is

$$\bar{x} > 65 + z_{.025}\frac{\sigma}{\sqrt{n}}, \qquad \text{or} \quad \bar{x} < 65 - z_{.025}\frac{\sigma}{\sqrt{n}}$$

which is equivalent to

$$65 \notin (\bar{x} - z_{.025}\frac{\sigma}{\sqrt{n}}, \bar{x} + z_{.025}\frac{\sigma}{\sqrt{n}}) = \bar{x} \pm z_{.025}\frac{\sigma}{\sqrt{n}} \quad (95\% \text{ CI})$$

That is, we reject the null at level α if and only if the $1-\alpha$ confidence interval fails to capture the claimed value 65.

One can thus use a $1-\alpha$ confidence interval to conduct the hypothesis test at level α:

- Confidence interval captured $\mu = 65$: Do not reject H_0
- Confidence interval failed to capture $\mu = 65$: Reject H_0

Note the relationship between and interpretation of:

$1-\alpha$ (confidence level) and α (level of the test).

Remark. For a one-sided test

$$H_0 : \mu = 65 \quad \text{vs} \quad H_a : \mu < 65$$

with corresponding decision rule

$$\bar{x} < 65 - c$$

the two equations (for determining n, c) become

$$\alpha = P(\text{Reject } H_0 \mid H_0 \text{ true}) = P(\bar{X} < 65 - c \mid \mu = 65)$$
$$\beta = P(\text{Fail to reject } H_0 \mid H_0 \text{ false}) = P(\bar{X} > 65 - c \mid \mu)$$

Example 0.6. Redo the preceding example but instead for a one-sided test

$$H_0 : \mu = 65 \quad \text{vs} \quad H_a : \mu < 65$$

with corresponding decision rule

$$\bar{x} < 65 - c$$

Answer: $c = z_\alpha \frac{\sigma}{\sqrt{n}} = 0.658,\ n = \left(\frac{\sigma(z_\alpha + z_\beta)}{\mu_0 - \mu'}\right)^2 = 25$

i-Clicker quiz 10 (extra credit)

Consider the decision rule $\bar{x} > 65 + c$ for a one-sided test

$$H_0 : \mu = 65 \quad \text{vs} \quad H_a : \mu > 65$$

As c is increased, which one of the following statements is WRONG?

A. The rejection region becomes smaller.

B. It gets harder to reject H_0.

C. Type-I error probability becomes smaller.

D. Type-II error probability becomes smaller.

E. None of the above

Summary

A hypothesis test has the following components:

- **Population**: e.g., all brown eggs produced by the farm, whose weights have a normal distribution with unknown mean μ but known variance σ^2
- **Null and alternative hypotheses**: $H_0 : \mu = \mu_0 \text{ vs } H_a : \mu \neq \mu_0$;
- **Random sample** from the population: $X_1, \ldots, X_n \overset{iid}{\sim} N(\mu, \sigma^2)$
- **Test statistic**: e.g., $\bar{X}$
- **Decision rule** (based on a specified **rejection region**): $|\bar{x} - \mu_0| > c$

Evaluation of the test:

- **Type-I error**:

$$\alpha = P(\text{Reject } H_0 \mid H_0 \text{ true}) = P(|\bar{X} - \mu_0| > c \mid \mu = \mu_0)$$

If α is specified first as the level of the test, then set $c = z_{\alpha/2}\frac{\sigma}{\sqrt{n}}$ (for a two sided test)

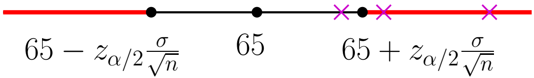
- **Type-II errors** (at any $\mu \neq \mu_0$)

$$\beta(\mu) = P(\text{Fail to reject } H_0 \mid H_0 \text{ false}) = P(|\bar{X} - \mu_0| < c \mid H_1 \text{ true})$$

To control both errors, we first choose c (dependent on n) to attain level α, then choose sample size n to achieve certain power $1 - \beta(\mu)$.

Limitation of the rejection region approach

The rejection region approach to conducting a hypothesis test at a given level makes sense, but the decision is discrete (reject or retain the null).



It does not reflect the strength of the evidence against H_0 (when rejecting it) or the closeness to the rejection region (when failing to reject it).

Another way of performing the hypothesis test is to assign a **score of extremeness** (relative to the null), called p-**value**, to any observed value of the test statistic in a continuous way.

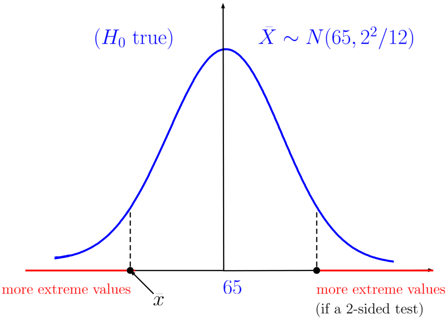
Logic behind the p-value approach to hypothesis testing

Consider the two-sided test again (in same setting but with a fresh mind):

$$H_0 : \mu = 65 \quad \text{vs} \quad H_a : \mu \neq 65 \text{ (or } H_a : \mu < 65)$$

We adopt a proof-by-contradiction procedure:

- **Assume H_0 is true**. Then $\mu = 65$ and $\bar{X} \sim N(65, 2^2/12)$.
- Intuitively, most observed values of $\bar{X}$ should be "around 65", while "extreme" values should be rare.
- For every observation $\bar{x}$ of $\bar{X}$, we assign an **extremeness score**, called p-value (e.g., most extreme 5%):



$$\text{pval}(\bar{x}) = \begin{cases} \text{left tail area only}, & \text{for } H_a : \mu < 65 \\ \text{total area of both tails}, & \text{for } H_a : \mu \neq 65 \end{cases}$$

- If for a specific sample, $\bar{x}$ is extreme (with small p-value), we have two possible explanations: **bad luck** or **wrong assumption** (H_0 does not hold true).
- If "very bad luck" is needed to explain the extreme observation, we choose to believe instead that the assumption must be wrong, and consequently H_0 should be rejected.
- Thus, small p-values lead to rejections of the null.
- Apparently, such a decision possesses a risk of making a type-I error (when H_1 is actually true).

The formal definition of p-value

Definition 0.1. The p-**value** of an observed value $\bar{x}$ of the test statistic $\bar{X}$ is the probability of observing $\bar{x}$, or values that are "more contradictory" to H_0, when assuming H_0 is true:

$$\text{pval}(\bar{x}) = P(\bar{X} \text{ is at least as contradictory as } \bar{x} \mid H_0 \text{ true})$$

We will reject H_0 if and only if the observed value of $\vec{X}$ corresponding to a sample is "very extreme".

Remark. The more extreme the observation, the smaller the p-value, the stronger the evidence against H_0.

Example 0.7. In the brown eggs example, suppose we observed $\bar{x} = 63.8$.

- $H_1 : \mu \neq 65$: The more contradictory values are $\bar{x} < 63.8$ and $\bar{x} > 66.2$ (mirror point). Thus, for a 2-sided test,

$$\begin{aligned} \text{pval}(63.8) &= 2 \cdot P(\bar{X} \leq 63.8 \mid H_0 \text{ true}) \\ &= 2 \cdot P\left(\frac{\bar{X} - \mu}{\sigma/\sqrt{n}} \leq \frac{63.8 - 65}{2/\sqrt{12}} \mid \mu = 65 \right) \\ &= 2 \cdot P(Z \leq -2.08) = 2 \cdot .019 = .038 \end{aligned}$$

- $H_1 : \mu \neq 65$: The more contradictory values are only $\bar{x} < 63.8$. In this case, the p-value is

$$\text{pval}(63.8) = P(\bar{X} \leq 63.8 \mid H_0 \text{ true}) = .019$$

Significance level

Definition 0.2. The cutoff p-value at which we choose to reject the null is called the **significance level** of the test. We denote it by α.

p-values that are smaller than the significance level (α) are said to be **significant** and will lead to the rejection of the null:

$$\text{Reject } H_0 \text{ if and only if } p\text{-value} \le \alpha.$$

Example 0.8. In the previous example, what is your conclusion if $\alpha = 5\%$? 1%?

Remark. For a p-value test at significance level α, the following three are the same

- significance level
- probability of making a type-I error
- level of the test.

They all equal α.

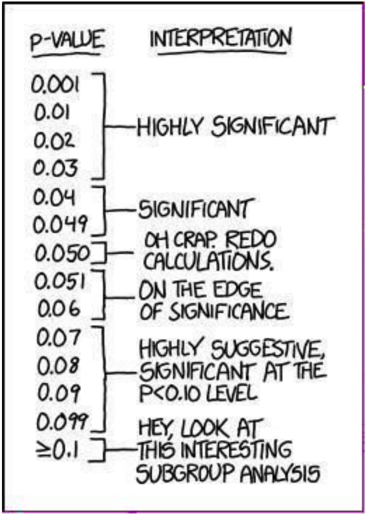
In theory, the p value is a continuous measure of evidence, but in practice it is typically trichotomized approximately into

- highly significant ($p \leq 0.01$)
- moderately significant ($0.01 < p \leq 0.03$)
- marginally significant ($p \approx 0.05$), and
- not statistically significant ($p > 0.06$)

Joke. What does a statistician call it when the heads of 10 rats are cut off and 1 survives?

Nonsignificant.

Hypothesis testing



When population variance is unknown

How do we conduct a hypothesis test for each of them?

- Population mean μ

- Population variance σ^2

Testing for μ with unknown variance

Recall that in the case of a normal population $N(\mu, \sigma^2)$ (with unknown μ and known σ^2), to conduct the two-sided test

$$H_0 : \mu = 65 \qquad vs \qquad H_1 : \mu \neq 65$$

at level α, one can use the following decision rule

$$|\bar{x} - 65| > z_{\alpha/2} \frac{\sigma}{\sqrt{n}}, \quad \text{or equivalently } \left| \frac{\bar{x} - 65}{\sigma/\sqrt{n}} \right| > z_{\alpha/2}$$

The test statistic $\frac{\bar{X}-65}{\sigma/\sqrt{n}}$ is correctly standardized (when H_0 is true), which has a standard normal distribution.

For the above reasons, the above test is called a (two-sided) z-test.

When σ is unknown, we can use the sample standard deviation S in place of σ (like the construction of confidence interval), yielding a t-test:

$$\left|\frac{\bar{x}-65}{s/\sqrt{n}}\right| > t_{\alpha/2,n-1}$$

Similarly, for a one-sided test, we can use a one-sided z-test (when σ known) or a one-sided t-test (when σ unknown).

Additionally, when σ is unknown, we can use the t distribution to calculate the p-value of a specific sample in order to conduct the hypothesis test at certain level α.

Example 0.9. Consider the egg-weight example again. Conduct the following test at level 95%

$$H_0 : \mu = 65 \qquad vs \qquad H_1 : \mu \neq 65$$

for a specific sample of 12 eggs with $\bar{x} = 64$ and $s^2 = 4.69$. Conduct the test at level $\alpha = .05$. What is the p-value of the sample?

(Answer: $|\frac{\bar{x}-65}{s/\sqrt{n}}| = 1.6 < t_{\alpha/2,n-1} = 2.201$, thus failing to reject the null. p-value=.138)

Testing for population variance

For population variance we are often only interested in a one-sided test of the form

$$H_0 : \sigma^2 = \sigma_0^2 \qquad vs \qquad H_1 : \sigma^2 > \sigma_0^2$$

Following previous reasoning, we write down a decision rule as follows

$$\frac{(n-1)s^2}{\sigma_0^2} > c$$

For a given level α, the cutoff c is determined from the following equation

$$\alpha = P\left(\frac{(n-1)s^2}{\sigma_0^2} > c \mid \sigma^2 = \sigma_0^2\right) \longrightarrow c = \chi^2_{\alpha,n-1}$$

Example 0.10 (Continuation of previous example). Conduct the following test at level 5%:

$$H_0 : \sigma^2 = 2^2 \qquad vs \qquad H_1 : \sigma^2 > 2^2$$

What is the p-value?

(Answer: $\frac{(n-1)s^2}{\sigma_0^2} = 12.9 < \chi^2_{\alpha,n-1} = 19.7$, thus failing to reject the null. $p$val$=.3$)