6. Hypothesis Testing in the Regime of Small Sample Sizes - Preparations

Exercises due Jul 29, 2020 08:59 JST Past Due

Bookmark this page

Concept check: Small Sample Sizes

1 point possible (graded)

Recall the clinical trials problem set-up.

Suppose that $n = 10$ and $m = 12$, so that the sample sizes are quite small. Consider the analysis of the one-sided, two-sample test performed in the videos in the previous vertical (and in Slide 7 of slides for Unit 4), where we defined a test statistic for $H_0 : \Delta_d = \Delta_c$, $H_1 : \Delta_d > \Delta_c$.

Should you expect this analysis to be accurate when the sample size is this small? (Choose all that apply. Correct responses must have a correct answer, 'Yes' or 'No', and also a correct explanation.)

- [ ] Yes, because Slutsky's theorem is a non-asymptotic result.
- [ ] Yes, because the asymptotic analysis given is independent of the sample size.
- [ ] No, because the calculation presented on the given slides was an asymptotic analysis (*i.e.*, we assumed $n \to \infty$). ✔
- [x] No, because, informally, Slutsky's theorem only gives a good approximation when the sample size is very large. ✔

Solution:

We first examine the correct choices, which are the third and fourth responses.

- The third choice "No, because the calculation presented on the given slides was an asymptotic analysis (*i.e.*, we assumed $n \to \infty$)." is correct. If $n = 10$ and $m = 12$, then both samples are quite small. Therefore, we cannot expect to apply limiting results, such as Slutsky's theorem, and derive accurate results.

- The fourth choice "No, because, informally, Slutsky's theorem only gives a good approximation when the sample size is very large." is also correct. We know that in the large n, m limit that the sample variance converges to the true variance, but we cannot assume that the same is true for relatively small m, n.

Now we examine the first and second choices, both of which are incorrect.

- "Yes, because Slutsky's theorem is a non-asymptotic result." is incorrect. We have shown in the previous part that the correct answer is "No." Moreover, the statement of Slutsky's theorem involves a limit as the sample size $n, m \to \infty$. Hence, this theorem is an *asymptotic* result.

- "Yes, because the asymptotic analysis given is independent of the sample size." We know that the correct answer is 'No' from above, and moreover, the asymptotic analysis depends on the sample size growing to infinity. Hence, the explanation is also incorrect.

7. The Chi-Squared Distribution and its Properties

Exercises due Jul 29, 2020 08:59 JST *Past Due*

Bookmark this page

The Chi-Squared Distribution and its Expectation

3 points possible (graded)

Note: This problem introduces the chi-squared distribution and is intended as an exercise in probability that you are encouraged to attempt before watching the following video.

The **χ_d^2 distribution with d degrees of freedom** is given by the distribution of

$$Z_1^2 + Z_2^2 + \cdots + Z_d^2,$$

where $Z_1, \ldots, Z_d \overset{iid}{\sim} \mathcal{N}(0,1)$.

What is the smallest possible sample space of χ_d^2?

- ○ $\mathbb{Z}_{\geq 0}$
- ○ $\mathbb{Z}$
- ● $\mathbb{R}_{\geq 0}$ ✔
- ○ $\mathbb{R}$

If $X \sim \chi_d^2$, what is $\mathbb{E}[X]$? Give your answer in terms of d.

0.5*d

Answer: d

$0.5 \cdot d$

STANDARD NOTATION

Let $\mathbf{Z} \sim \mathcal{N}(0, I_{d\times d})$ denote a random vector whose components are standard Gaussians: $Z^{(1)}, \ldots, Z^{(d)} \sim \mathcal{N}(0,1)$. Which one of the following random variables has a chi-squared distribution with d degrees of freedom?

- ( ) $\max(Z^{(1)}, \ldots, Z^{(d)})$
- (•) $|Z^{(1)}| + |Z^{(2)}| + \cdots |Z^{(d)}|$
- ( ) $\|\mathbf{Z}\|_2$
- ( ) $\|\mathbf{Z}\|_2^2$ ✔

Solution:

The smallest sample space of a Gaussian random variable Z is $\mathbb{R}$. Hence, the smallest possible sample space of Z^2 is $\mathbb{R}_{\geq 0}$. And the same holds for the sum

$$Z_1^2 + Z_2^2 + \cdots + Z_d^2,$$

so the smallest possible sample space for χ_d^2 is $\mathbb{R}_{\geq 0}$.

Next, by linearity of expectation,

$$\mathbb{E}\left[X\right] = \mathbb{E}\left[Z_1^2 + Z_2^2 + \cdots + Z_d^2\right] = d \cdot 1 = d,$$

because $Z_1, \ldots, Z_d \overset{iid}{\sim} \mathcal{N}\left(0, 1\right)$.

The ℓ_2 norm $\|\cdot\|_2$ measures the Euclidean distance from the origin. Hence, if $\mathbf{Z} \sim \mathcal{N}\left(0, I_{d \times d}\right)$, then

$$\|\mathbf{Z}\|_2^2 = \left(Z^{(1)}\right)^2 + \left(Z^{(2)}\right)^2 + \cdots + \left(Z^{(d)}\right)^2 \sim \chi_d^2.$$

Throwing Darts

1 point possible (graded)

You are playing darts on a dart-board that is represented by the entire plane, $\mathbb{R}^2$. You get a 'bullseye' if the dart lands inside of the unit disc $D^1 := \{(x,y) : x^2 + y^2 \leq 1\}$. You dart throws are modeled by a Gaussian random vector $\mathbf{Z}$, where $Z^{(1)}, Z^{(2)} \overset{iid}{\sim} \mathcal{N}(0,1)$.

Let f_d represent the density of the χ^2_d distribution.

Which of the following equals the probability of getting a bullseye?

- ◯ $\int_0^1 f_1(x)\ dx$
- ◯ $\int_0^1 f_2(x)\ dx$ ✔
- ◯ $\int_1^\infty f_2(x)\ dx$
- ◯ $\int\int_{D^1} f_2(x)\ dxdy$

Solution:

A bullseye is given by the event $\left(Z^{(1)}\right)^2 + \left(Z^{(2)}\right)^2 \leq 1$. Since $\left(Z^{(1)}\right)^2 + \left(Z^{(2)}\right)^2 \sim \chi^2_2$, it follows that

$$P(\text{bullseye}) = \int_0^1 f_2(x)\ dx.$$

Remark: The $d = 2$ case is special, because it turns out that $\chi^2_2 = \text{Exp}(1/2)$. This can be seen using the explicit formula for the density of a χ^2_2, but it is not necessary for this course to know the density of a chi-squared random variable with d degrees of freedom by heart.

The Chi-Squared Distribution and the Sample Second Moment

2 points possible (graded)

Let $X_1, \ldots, X_n \overset{iid}{\sim} \mathcal{N}(0, \sigma^2)$ and let

$$V_n = \frac{1}{n}\sum_{i=1}^{n} X_i^2$$

denote the sample second moment. For an appropriate expression A given in terms of n and σ^2, we have that $AV_n \sim \chi^2$.

What is A?

$A =$ n

n

Answer: n/sigma^2

How many degrees of freedom does the above χ-squared random variable have? (Give your answer in terms of n.)

n

Answer: n

Solution:

Observe that

$$\frac{n}{\sigma^2} V_n = \sum_{i=1}^{n} \frac{X_i^2}{\sigma^2} = \sum_{i=1}^{n} \left(\frac{X_i}{\sigma}\right)^2,$$

and $X_i/\sigma \sim \mathcal{N}(0,1)$ because $X_i \sim \mathcal{N}(0, \sigma^2)$. Hence, $\frac{n}{\sigma^2} V_n$ is a χ_n^2 random variable.

A Special Case of Cochran's Theorem I

3 points possible (graded)

Cochran's theorem states that if $X_1, \ldots, X_n \stackrel{iid}{\sim} \mathcal{N}(\mu, \sigma^2)$, then the sample variance

$$S_n := \frac{1}{n}\left(\sum_{i=1}^{n} X_i^2\right) - \left(\overline{X}_n\right)^2$$

satisfies:

- $\overline{X}_n$ is independent of S_n, and
- $\frac{nS_n}{\sigma^2} \sim \chi^2_{n-1}$.

In this problem, you will verify that Cochran's theorem holds when $n = 2$. Let $X_1, X_2 \stackrel{iid}{\sim} \mathcal{N}(\mu, \sigma^2)$.

The expression S_2 can be written in the form A^2 where A is a polynomial in X_1 and X_2.

What is A^2?

Type **X_1** for X_1 and **X_2** for X_2.

$A^2 =$ **Answer:** (X_1 - X_2)^2/4

The expression A from the previous question is a random variable, and moreover is distributed as $\mathcal{N}\left(\mu^*, (\sigma^*)^2\right)$ for some μ^* and σ^* that can be expressed in terms of the original parameters μ and σ. (Note: A can have two forms, but both would have the same distribution by symmetry).

What is μ^* expressed in terms of μ and σ?

$\mu^* =$

Answer: 0.0

What is $(\sigma^*)^2$ expressed in terms of μ and σ?

$(\sigma^*)^2 =$

Answer: sigma^2/2

Solution:

Observe that

$$S_n = \frac{X_1^2 + X_2^2}{2} - \left(\frac{X_1 + X_2}{2}\right)^2 = \frac{X_1^2}{4} + \frac{X_2^2}{4} - \frac{1}{2}X_1 X_2 = \left(\frac{X_1 - X_2}{2}\right)^2.$$

Hence, we can take $A = \pm\frac{X_1 - X_2}{2}$ (either choice has the same distribution, by symmetry). Next,

$$\mathbb{E}[A] = \frac{1}{2}\mathbb{E}[X_1 - X_2] = \frac{1}{2}(\mu - \mu) = 0,$$

and

$$\mathrm{Var}(A) = \mathrm{Var}\left(\frac{X_1 - X_2}{2}\right) = \frac{1}{4}(\mathrm{Var}(X_1) + \mathrm{Var}(X_2)) = \frac{\sigma^2}{2}.$$

A Special Case of Cochran's Theorem II

4 points possible (graded)

As above, let $X_1, X_2 \overset{iid}{\sim} \mathcal{N}(\mu, \sigma^2)$.

Recall the random variable A that you found in the previous problem in terms of X_1 and X_2.

Let $\overline{X}_2 = \dfrac{X_1 + X_2}{2}$, i.e. $\overline{X}_n$ when $n = 2$.

What is $\mathbb{E}[A\overline{X}_2]$?

0

Answer: 0.0

Using the answer above, which of the following are true? (Choose all that apply.)

- [ ] A and $\overline{X}_2$ are independent. ✔
- [ ] A and $\overline{X}_2$ are not independent.
- [ ] $A, \overline{X}_2 \sim \mathcal{N}(0, 2\sigma^2)$.
- [ ] $A \sim \mathcal{N}(0, \sigma^2/2)$ and $\overline{X}_2 \sim \mathcal{N}(\mu, \sigma^2/2)$. ✔

For some expression B in terms of σ^2, the random variable $BS_2 \sim \chi^2$. What is B?

$B =$

Answer: 2/sigma^2

How many degrees of freedom does the χ^2 random variable BS_2 have?

Answer: 1

Solution:

Recall that $A = \frac{X_1 - X_2}{2}$ and $\overline{X}_2 = \frac{X_1+X_2}{2}$. Hence,

$$\mathbb{E}[A\overline{X}_2] = \frac{1}{4}\mathbb{E}\left[(X_1 - X_2)(X_1 + X_2)\right] = \frac{1}{4}(\sigma^2 - \sigma^2) = 0.$$

As jointly Gaussian variables (why is it that A and $\overline{X}_2$ are jointly Gaussian?) that are uncorrelated are also independent, A and $\overline{X}_2$ are independent. By the previous problem, we know $A \sim \mathcal{N}(0, \sigma^2/2)$. A quick calculation shows that $\overline{X}_2 \sim \mathcal{N}(\mu, \sigma^2/2)$. Hence, the first and last choices are correct in the multiple choice question.

Observe that

$$\frac{2}{\ } S_2 = \frac{2}{\ }\left(\frac{X_1 - X_2}{\ }\right)^2 = \left(\frac{X_1 - X_2}{\ }\right)^2.$$

Concept Check: Cochran's Theorem and Unbiased Sample Variance

1 point possible (graded)

Let $X_1, \ldots, X_n$ be i.i.d. and distributed according to $\mathcal{N}(0, \sigma^2)$. Let $\overline{X}_n = \frac{X_1+X_2+\cdots+X_n}{n}$. What is the distribution of $\frac{(n-1)\widetilde{S}_n}{\sigma^2}$, where $\widetilde{S}_n$ is the the unbiased sample variance of $X_1, \ldots, X_n$:

$$\widetilde{S}_n = \frac{1}{n-1}\sum_{i=1}^{n}\left(X_i - \overline{X}_n\right)^2$$

Type **Cn** for chi-squared distribution with n degrees of freedom, **Cn1** for chi-squared distribution with $n-1$ degrees of freedom.

Cn1

Answer: Cn1 + 0*Cn

$Cn1$

STANDARD NOTATION

Solution:

By Cochran's theorem,

$$\begin{aligned} \frac{nS_n}{\sigma^2} &\sim \chi^2_{n-1} \\ \iff \quad \frac{(n-1)\widetilde{S}_n}{\sigma^2} &\sim \chi^2_{n-1} \end{aligned}$$

Remark: We will use the random variable $\frac{\widetilde{S}_n}{\sigma^2}$ in the upcoming videos in what is called the Student's T Test. The point of this problem was to show that $\frac{(n-1)\widetilde{S}_n}{\sigma^2}$ is a χ^2_{n-1} random variable, thereby showing that the distribution of $\frac{\widetilde{S}_n}{\sigma^2}$ is the distribution of a χ^2_{n-1} random variable scaled by $\frac{1}{n-1}$.

Comparing Chi-Squared and Student's T Distribution

2 points possible (graded)

Consider the distribution χ^2_n (χ-squared with n degrees of freedom). Let $f_n : \mathbb{R} \to \mathbb{R}$ denote the pdf of χ^2_n, and let A_n denote the maximizer of f_n (*i.e.*, the peak of the pdf of the distribution χ^2_n is located at A_n).

What is $\lim_{n\to\infty} A_n$? (Answer heuristically, based on discussions in the lecture video about how the shape of the chi-squared distribution evolves with n.)

- [x] 0
- [ ] 1
- [ ] ∞ ✔
- [ ] None of the above

Consider the **Student's T Distribution** , which is defined to be the distribution of

$$T_n := \frac{Z}{\sqrt{V/n}}$$

where $Z \sim \mathcal{N}(0,1)$, $V \sim \chi^2_n$, and Z and V are independent. Let g_n denote the pdf of T_n, and let B_n denote the maximizer of g_n (*i.e.*, the peak of the pdf of the distribution T_n is located at B_n).

∨ **Solution:**

The graph of the pdf of χ^2_n in the slides shows that the peak of the distribution moves to the right as $n \to \infty$. Hence

$$\lim_{n\to\infty} A_n = \infty.$$

This is intuitive since we showed in a previous problem that $\mathbb{E}[X] = n$ if $X \sim \chi^2_n$.

As $n \to \infty$, the random variable V/n converges to 1 in probability. Hence, as $n \to \infty$,

$$T_n \xrightarrow[n\to\infty]{(d)} \mathcal{N}(0,1).$$

Since the distribution $\mathcal{N}(0,1)$ is peaked at the origin, this implies

$$\lim_{n\to\infty} B_n = 0.$$

Concept Check: Student's T Distribution

3 points possible (graded)

Consider the statistic

$$T_n := \sqrt{n}\left(\frac{\overline{X}_n - \mu}{\sqrt{\frac{1}{n-1}\sum_{i=1}^{n}\left(X_i - \overline{X}_n\right)^2}}\right),$$

where $\bar{X}_n$ is the sample mean of i.i.d. Gaussian observations with mean μ and variance σ^2.

For all $n \geq 2$, the distribution of T_n is a standard Gaussian $\mathcal{N}(0,1)$.

- (●) True
- ( ) False ✔

As $n \to \infty$, what does

$$\frac{1}{n-1}\sum_{i=1}^{n}\left(X_i - \overline{X}_n\right)^2$$

converge to...

- ( ) The number μ (weakly)
- ( ) The number σ^2 (weakly) ✔
- (●) The distribution $\mathcal{N}(0,1)$
- ( ) The distribution χ^2_{n-1}

As $n \to \infty$, the statistic T_n converges in distribution to

- ( ) $\mathcal{N}(0,1)$ ✔
- ( ) $\mathcal{N}(\mu, \sigma^2)$
- (●) χ^2_{n-1}
- ( ) χ^2_n

Solution:

The definition of the student's T distribution with $n-1$ degrees of freedom is that it is given by the distribution of $\frac{Z}{\sqrt{V/(n-1)}}$ where $Z \sim \mathcal{N}(0,1)$, $V \sim \chi^2_{n-1}$ and Z and V are independent. Since we are dividing by V, a χ^2 random variable, then T_n will not have the same distribution as $\mathcal{N}(0,1)$ for all $n \geq 2$.

By the law of large numbers and Slutsky's lemma,

$$\frac{1}{n-1}\sum_{i=1}^{n}\left(X_i - \overline{X}_n\right)^2 = \frac{n}{n-1}\left[\left(\frac{1}{n}\sum_{i=1}^{n}X_i^2\right) - \left(\overline{X}_n\right)^2\right] \to \sigma^2$$

in probability.

By the central limit theorem,

Quantiles of the T Distribution:

The $(1-\alpha)$-quantile of the t_{n-1} (corresponding to a one-sided test with statistic T_n) can be computed using standard computational tools such as R. One can also find online tables for the quantiles via a simple Google search, which yields results such as this, this, and this.

As a reminder, in this class the $(1-\alpha)$ quantile of the distribution of a random variable T is the number q_α such that

$$P\left(T \leq q_\alpha\right) = 1-\alpha.$$

Concept Check: T Test

1 point possible (graded)

Let $X_1, \ldots, X_n \stackrel{iid}{\sim} \mathcal{N}(\mu^*, \sigma^2)$ for some unknown $\mu^* \in \mathbb{R}$ and $\sigma^2 > 0$. You want to decide between the following null and alternative hypotheses on the mean of $X_1, \ldots, X_n$:

$$\begin{aligned} H_0 &: \mu^* = 0 \\ H_1 &: \mu^* \neq 0. \end{aligned}$$

To do so, you define the student's T statistic

$$T_n = \sqrt{n} \frac{\overline{X}_n}{\sqrt{\widetilde{S}_n}}$$

where

$$\widetilde{S}_n = \frac{1}{n-1} \sum_{i=1}^{n} \left(X_i - \overline{X}_n\right)^2$$

is the unbiased sample variance.

The student's T test of level α is specified by

$$\psi_\alpha = \mathbf{1}\left(|T_n| > q_{\alpha/2}\right)$$

where $q_{\alpha/2}$ is the unique number such that $P\left(T_n < q_{\alpha/2}\right) = 1 - \frac{\alpha}{2}$.

Which of the following are true about the student's T test? (Choose all that apply.)

- [ ] The statistic T_n is distributed as a standard Gaussian.
- [ ] The test requires the data $X_1, \ldots, X_n$ to be Gaussian. ✔
- [ ] The distribution of T_n is pivotal, *i.e.*, its quantiles may be found in tables. ✔
- [ ] The test is non-asymptotic. That is, for any fixed n, we can compute the level of our test rather than the *asymptotic* level. ✔

Solution:

We examine the choices in order.

- The first choice is incorrect. Due to the fact that T_n has the sample variance $\widehat{S}_n$ in the denominator and not the *true* variance σ^2, the statistic T_n will **not** be standard Gaussian.

- The second choice is correct. It is a key assumption that the data is Gaussian. Otherwise, the test statistic T_n will not necessarily follow the student's T distribution and, hence, may not even be pivotal.

- The third choice is correct. For any fixed n, we may find the quantiles of the student's T distribution in tables. Since the distribution does not depend on the value of the true parameter, the test statistic T_n is indeed pivotal.

- The last choice is also correct. As stated in the previous bullet, for any fixed n, the quantiles of the student's T distribution may be found in tables. Hence, we can find the non-asymptotic level of this test.

Remark: Assuming the data is Gaussian, the student's T test is useful in situations where the sample size is not very large, since the level may be precisely quantified even for small n.

11. Back to Clinical Trials: Two Sample T-Test and the Welch-Satterthwaite Formula

Exercises due Jul 29, 2020 08:59 JST

Bookmark this page

Video Note: In the following video, the calculated values for Slide 27/62 starting at 3:00 mark contain errors. The following are the correct values:

- The correct value of the test statistic is

$$\frac{156.4 - 132.7}{\sqrt{\frac{5198.4}{70} + \frac{3867}{50}}} \approx 1.9248.$$

- The p-value using the shorthand formula is 0.029974 and the p-value using the W-S formula is 0.02832.

Further, at the 3:46 mark, there is a reference to "49 degrees of freedom". The correct number of degrees of freedom is 50.