Section 24

Lecture 8

Section 25

Estimation (learning)

Lecture 8: Plan for estimation (learning)

- Review foundations of estimation theory that are relevant to causal inference.
  - Statistical models (Parametric and non-parametric).
  - Correctly specified models.
- Motivate why we need to study certain estimation problems.
  - Convergence of conditional means.
- Introduce some commonly used estimators: Regression estimators and inverse probability weighted estimators.
  - Brief summary of linear models.
  - Logistic regression models.
  - M-estimators.
  - Link this back to counterfactuals.

My take on data science

1. Start with the question.
   (Design your target trial)
2. Formalize the question in mathematical language.
   (Define your estimand)
3. Display the assumptions that are needed to identify your estimand.
   (Present your identifiability conditions)
4. Compute estimates of your estimands from your data.
   (Do your estimation)

$\Longrightarrow$ we **never** start the process by considering a regression model (linear, logistic, Cox model, neural net, random forest, ..., whatever).

Finite sample inference: Where does randomness come from?

- We will mostly consider superpopulation inference, where the randomness comes from the fact that we have a random draw from the superpopulation.
- However, in a randomised trial, we do not necessarily need to consider a superpopulation at all.
- In these (simple) settings, we can do *finite* sample inference.
- Yet, we shall see that to generalize the results outside of the study – which is really what researcher would like to do in most settings – it is necessary to consider large sample extensions (which usually end up being superpopulations).

Superpopulation inference and finite sample inference

- We will most often suppose that our study population is sampled at random from an (essentially) infinite superpopulation, sometimes referred to as the target population.
- Broadly speaking, we aimed to generalize our results to this superpopulation.
- It is possible to take a different point of view in randomised trials, often called "design-based inference", which we will study now. This does not require the consideration of a superpopulation at all.[34]

Definition (Design-based inference)

Inference is drawn from a *finite* population, where the potential outcomes of the experimental units are fixed and the randomness comes solely from the treatment assignment.

[34] However, to generalize results from finite samples to settings outside of the experiment – even if we start in the design based setting – it is necessary to rely on superpopulation inference. Thus, if we are interested in using the results from the trials for decisions (or rigorous reasoning more broadly) outside of the experiment, it seems that we need to rely on superpopulation inference anyway.

Fisher Randomisation inference

- Key idea: do inference based solely on the assignment mechanism.
- The counterfactuals $Y_i^{a=1}$, $Y_i^{a=0}$ are considered to be *fixed* variables.
- All the *randomness* comes from the random assignment of A.
- Fisher's aim was to test the sharp null hypothesis, using Fisher exact test.
- The idea is basically a stochastic proof by contradiction...
- Fisher's null hypothesis is $H_0 : Y_i^{a=1} \equiv Y_i^{a=0}$ for all $i \in \{1, 2, \ldots, n\}$. In words, the treatment has no effect of the outcomes in no individual. Under the null hypothesis (but of course not under the alternative) $Y_i^{a=1} = Y_i^{a=0} = Y_i$.
- This null hypothesis is called a **sharp** null hypothesis because it is specified such that it allows the researcher to fill in a hypothetical value for each unit's missing counterfactual outcome

Fisher's exact test: A test of individual effects

- Define the *sharp null hypothesis* $H_0 : Y_i^{a=1} = Y_i^{a=0}$ for all $i \in \{1, 2, \ldots, n\}$.
- Define a test statistic[35], e.g. $S^{diff} = \frac{1}{n_1} \sum_{i:A_i=1} Y_i - \frac{1}{n_0} \sum_{i:A_i=0} Y_i$.
- Let s^* be an observed test statistic. Then $P(S \geq s^*)$ is a p-value, where the probability is under the law that describes the null hypothesis.
- Fisher suggested an exact test.
  - The idea is to ask the following question: How unusual or extreme is the observed statistic (say, absolute difference), assuming that the null hypothesis is true?
- Intuitively, we want to have *power* against alternative hypotheses, but this is somehow complicated because there are many alternative hypotheses. It seems reasonable to have good power against alternative hypotheses that are substantively most interesting.

[35] A statistic is a known, real-valued function of the data (here, $Y_1, A_1, L_1, \ldots, Y_n, A_n, L_n$)

*Examples of statistics

- Averages (like above)
- Trimmed means
- Quantiles (medians)
- T-statistics
- Rank statistics (perhaps good when heavy-tailed distributions)

One example is the Kolmogorov-Smirnov Statistic. Define, the empirical distributions

$$\hat{F}_{a=1}(y) = \frac{1}{n_1} \sum_{i:A_i=1} I(Y_i \leq y) \quad \hat{F}_{a=0}(y) = \frac{1}{n_0} \sum_{i:A_i=1} I(Y_i \leq y).$$

The Kolmogorov-Smirnov Statistic is

$$S^{ks} = \sup_y |\hat{F}_{a=1}(y) - \hat{F}_{a=0}(y)| = \max_i |\hat{F}_{a=1}(Y_i) - \hat{F}_{a=0}(Y_i)|.$$

*We can combine statistics

- Fisher's exact p-value inference is valid when there is one test statistic and one null hypothesis.
- However, we can combine test statistics.
  - Consider two statistics S^1 and S^2.
  - The combine $S^{comb} = g(S^1, S^2)$. (e.g. $S^{comb} = \max(S^1, S^2)$)
  - Then we can calculate a p-value

$$P(S^{comb} \leq s^{\star,comb})$$

Illustration of Fisher's exact test

Under the sharp H_0, we can impute missing values of the counterfactuals

i	$Y_i^{a=1}$	$Y_i^{a=0}$	A_i	Y_i
1	-5	-5	1	-5
2	6	6	0	6
3	8	8	1	8
4	0	0	0	0

Table 2: Fisher's idea

The idea is resampling without replacement

Consider the estimator $\frac{1}{n_1}\sum_{i:A_i=1} Y_i - \frac{1}{n_0}\sum_{i:A_i=0} Y_i$. Because we have a completely randomised experiment, the following $\binom{4}{2} = 6$ scenarios are equally possible under H_0,

$$\boldsymbol{A} = (1,1,0,0),\ \hat{\tau} = \frac{-5+6-8-0}{2} = -3.5$$

$$\boldsymbol{A} = (1,0,1,0),\ \hat{\tau} = \frac{-5-6+8-0}{2} = -1.5$$

$$\boldsymbol{A} = (1,0,0,1),\ \hat{\tau} = \frac{-5-6-8+0}{2} = -9.5$$

$$\boldsymbol{A} = (0,1,1,0),\ \hat{\tau} = \frac{5+6+8-0}{2} = 9.5$$

$$\boldsymbol{A} = (0,1,0,1),\ \hat{\tau} = \frac{5+6-8+0}{2} = 1.5$$

$$\boldsymbol{A} = (0,0,1,1),\ \hat{\tau} = \frac{5-6+8+0}{2} = 3.5$$

One way of explaining Fisher's exact test

1. Do the randomization.
2. Calculate a statistic S, a function of the observed data.
3. Under the assumption of H_0, i.e. no individual level causal effect, fill in missing potential outcomes.
4. Under the assumption of H_0, generate many hypothetical replications of the randomization, and in each of which calculate a statistic S_{rep}.
5. Compare S with the values S_{rep}

This is an example of a permutation test.

More formally

- Define $H_0 : Y_i^{a=1} = Y_i^{a=0}$.
- Now, consider the randomisation distribution of two statistics S
- Define $\mathcal{F} = (\boldsymbol{Y^0}, \boldsymbol{Y^1})$. In this case, the randomization distributions of $S = S(\boldsymbol{A}, \boldsymbol{Y}, \boldsymbol{L})$ is

$$F(s) = P(S \leq s \mid \mathcal{F})$$

- Then the one-sided p-value of observing the same value or more extreme of the observed statistics S is $F(S)$.
- In our example, the one-sided p-value is $1 - F(-1.5) = 1 - 0.5$.

Fisher's randomization test formally

Theorem (Nominal coverage of the exact test)

Under consistency and H_0, $P(F(S) \leq \alpha \mid \mathcal{F}) \leq \alpha$ for all $\alpha \in (0, 1)$.

Proof.

This follows from some basic properties of the distribution function: indeed, $F^{-1}(\alpha) = \sup\{s : F(s) \leq \alpha\}$. Also $F(s)$ is non-decreasing and right-continuous and therefore

$$P(F(S) \leq \alpha) = P(S < F^{-1}(\alpha)) = \lim_{s \to F^{-1}(\alpha)} P(S \leq s) \leq \alpha.$$

□

PS: you may have seen the probability integral transform before, i.e. if X is continuous, then $Z = F(X) \sim U(0, 1)$

$$P(F(X) \leq \alpha) = P(X \leq F^{-1}(\alpha)) = F(F^{-1}(\alpha)) = \alpha.$$

Conservative or good?

Conservative does not necessarily mean appropriate. Consider a confidence interval formed by stating that a random 95% of the time, the interval is any positive or negative number, and that 5% of the time, the interval is the number 0. Such an interval would cover the true value of any quantity of interest at least 95% of the time, and thus would also be a "conservative" interval. It would not, however, be of any use....
Guido W Imbens and Donald B Rubin. *Causal inference in statistics, social, and biomedical sciences*. Cambridge University Press, 2015

Checking for no causal effect (hypothesis testing)

- Suppose we want to check if there is no causal effect.
- A classical frequentist approach goes as follows
  - Assume no effect (the null hypothesis).
  - Calculate a statistic,[36] and see how surprising the statistics is, under the assumption of no effect.
  - If it is very surprising, we reject.
- This is contrapositive logic, applied to probabilities.

[36] A statistic is a known, real-valued function of the data

We should be careful with this (Example from Shpitser)

Suppose we do cancer screening.

- Consider a rare cancer, our outcome Y, such that $P(Y = 1) = 0.00001$
- Consider also a cancer lab test T. And suppose
  - Test false positive $P(T = 1 \mid Y = 0) = 0.01$.
  - Test false negative $P(T = 0 \mid Y = 1) = 0.001$.
- Suppose we had a positive test, $Y = 1$. Should we worry?
- Just use Bayes theorem,

$$P(Y = 1 \mid T = 1) = \frac{P(T = 1 \mid Y = 1)P(Y = 1)}{P(T = 1)} \approx 0.001.$$

- What would the Frequentist do? Assume $Y = 0$, and check how surprised we would be, that is, calculate $P(T = 1 \mid Y = 0) = 0.01$, which is surprising....
- Lesson learned, if hypothesis probabilities are uneven, hypothesis testing is not ideal..

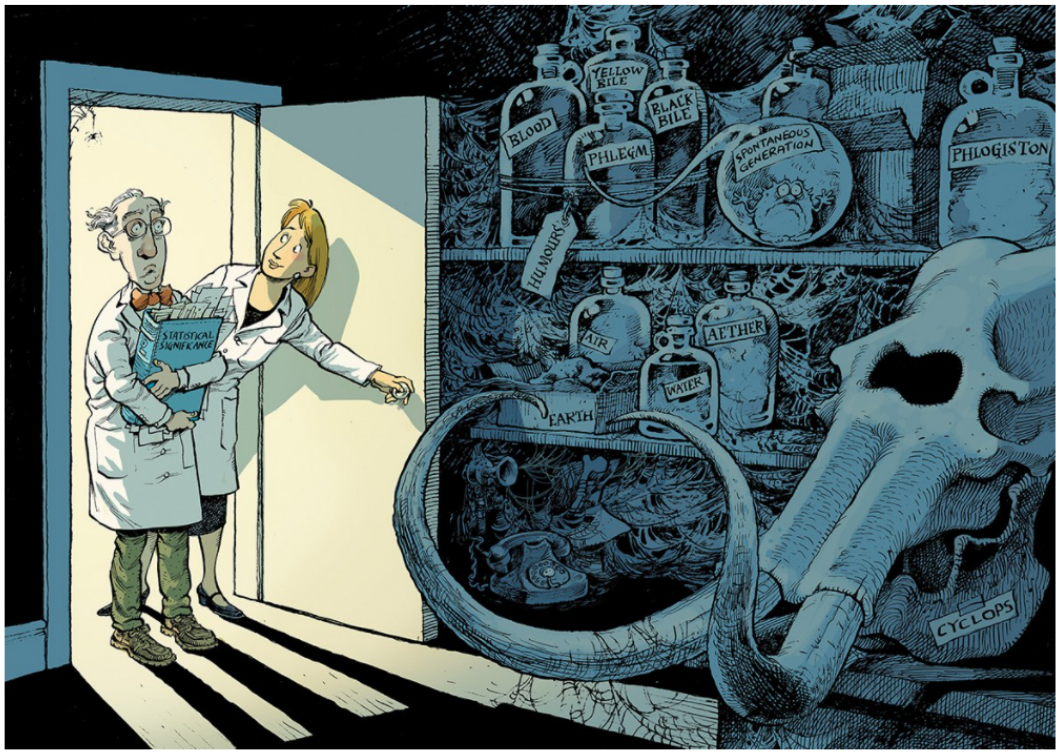
Abandon Statistical Significance?

COMMENT · 20 MARCH 2019

Scientists rise up against statistical significance

Valentin Amrhein, Sander Greenland, Blake McShane and more than 800 signatories call for an end to hyped claims and the dismissal of possibly crucial effects.

Valentin Amrhein, Sander Greenland & Blake McShane



Reasons (may seem obvious, but worth a reminder)

- There is nothing wrong with the p-value itself, as a mathematical construct.
- However, it is often misused.
- $p < 0.05$ is an arbitrary threshold.
- P-hacking is frequently done in practice.

Blakeley B McShane et al. "Abandon statistical significance". In: *The American Statistician* 73.sup1 (2019), pp. 235–245

Estimation (learning) in causal inference settings (informal motivation)

- An identification formula motivates estimators.
- Estimation in causal inference settings is, in principle, identical to the inverse problem you have studied in previous machine learning or statistics classes.
- However, the functionals we are estimating are sometimes unusual, and therefore we sometimes need new estimators. Indeed, a lot of identification results in causal inference have motivated new estimation theory.
- Broadly speaking, causal inference researchers are concerned about bias.
  - After doing the hard work of deriving an identification formula, we do not want to induce bias in the estimation step.
- I remind you about how we divide the causal inference into different tasks: (i) Define your question of interest (estimand), (ii) Evaluate whether the estimand is identified, (iii) if it is identified, we proceed with estimation.

Estimation vs. identification

- We have considered identification assumptions that are necessary even if we had an infinite amount of data.
- The statistical modeling assumption we consider now are invoked because we do not have infinite amount of data.

PS: In this course we will mainly consider frequentist inference: probability is defined as a limiting frequency. An alternative is Bayesian inference,[37] which defines probability as a degree of belief.

[37] Again, this is not the same as a Bayesian network

Where does randomness come from?

In causal inference

- Sampling variability
  Like classical statistics
  - Sample from superpopulation (classical inference)
  - Sample of counterfactuals (e.g. Fisher Randomization test)
- Non-deterministic counterfactuals
  But we have assumed that the counterfactuals are deterministic. And, in practice, that doesn't change anything when we do superpopulation inference (we will get to it).

Where do we do inference

Suppose we estimate the proportion of treated individuals who develop the outcome (say, death) as

$$\hat{p} = \hat{P}(Y = 1 \mid A = 1) = 7/13,$$

and I get 95% confidence intervals in the usual way (called Wald intervals) as

$$\hat{p} \pm 1.96\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}.$$

When is this confidence interval valid and what does it mean?

`Example from Hernan & Robins, Chapter 10.3`

There are two options

1. Individuals are sampled at random from an essentially infinite super-population, sometimes referred to as the source or target population. Then, if we repeatedly draw random samples of size 13 from the treated individuals in the super-population, the number of individuals who develop the outcome among the 13 is a binomial random variable with success probability equal to the true $P(Y = 1 \mid A = 1)$.
   This is the model we will consider most of the time.
2. We are not considering a super-population; we are doing inference in the sample we have. We assume that every individual i has a non-deterministic probability $p_i^{a=1}$ of experiencing $Y = Y_i^{a=1} = 1$ (because we consider those with $A = 1$). However, for the confidence interval to be correct, we must assume that $p_i^{a=1}$ is constant in i, say, $p_i^{a=1} = p$. Think about the idea that $p_i^{a=1}$ is constant in i. This seems very contrived, as we would believe that individuals have different risk of the outcome, due to genetics, life style factors etc.

Motivation for regression modelling and the curse of dimensionality

Definition (Statistical model)

A statistical model $\mathcal{P}$ is a collection of laws, $\mathcal{P} = \{P_\eta : \eta \in \Gamma\}$.

PS: Statistical models are sometimes called probabilistic hypothesis classes in the machine learning literature.

Definition (Parametric statistical model)

A statistical model $\mathcal{P}$ is parametric $\mathcal{P} = \{P_\theta : \theta \in \Theta\}$, where $\Theta \subseteq \mathbb{R}^k$ for a positive integer k.

So far we have been non-parametric: we have not restricted ourselves to parametric models. This is arguably desirable, because then we *do not* impose parametric restrictions on the data generating mechanism.

Consistency of an estimator

Here is an informal definition. Consistency of an estimator with respect to a parameter (the estimand) means that, when the sample size increases, the estimates get arbitrarily close to the parameter.
PS: This definition is with respect to an estimator. We have previously discussed consistency as an identification conditions, concerning interventions, which is a different thing.

More formal definition of consistent estimator (not strictly needed, but for your information)

Let $\{P_\eta : \eta \in \Gamma\}$ is a family of distributions (laws), and $X_\eta = \{X_1, X_2, \ldots : X_i \sim P_\eta\}$ is an infinitely large sample from the law P_η. Let $\{\hat{\mu}_n(\eta)\}$ be a sequence of estimators for $\mu(\eta)$, where e.g. $\hat{\mu}_n$ is an estimator based on the first n observations of a sample. Then the sequence $\{\hat{\mu}_n(\eta)\}$ is said to be (weakly) consistent if

$$\operatorname*{plim}_{n\to\infty} \hat{\mu}_n(\eta) = \mu(\eta), \quad \text{for all } \eta \in \Gamma.$$

where plim denotes convergence in probability, that is,

$$P_\eta(\mid \hat{\mu}_n(\eta) - \mu(\eta) \mid > \epsilon) \to 0 \text{ as } n \to \infty \text{ for all } \epsilon > 0, \eta \in \Gamma.$$

Motivation: Simple mean estimation

- Suppose we are interested in estimating a parameter, say, $h(L, A, Y)$ from an observed sample of n observations, (L_i, A_i, Y_i), $i = 1, \ldots, n$.
- Suppose we would like to ignore the assumptions encoded in our model $\mathcal{P}$ when we study $h(L, A, Y)$; more precisely, we will only use the fact that we have draws from i.i.d. individuals where $\mathbb{E}(Y) = \mu$ and that Y is continuous with finite variance $\sigma^2 < \infty$.
- Our statistical model is non-parametric; $\mathcal{P} = \{P(Y = y) : \int y^2 f(y) dy < \infty\}$. For $h(L, A, Y) \equiv \mathbb{E}(Y)$, we would simply do the empirical mean (sample mean) $\mathbb{E}_n(Y) = \frac{1}{n} \sum_{i=1}^{n} Y_i$. By the weak law of large numbers (WLLN),

$$\lim_{n \to \infty} P(|\mathbb{E}_n(Y) - \mu| > \epsilon) = 0.$$

  So the estimator is consistent. Indeed, the estimator is $\sqrt{n}$-consistent, and by the CLT $\sqrt{n}(\mathbb{E}_n(Y) - \mu) \sim \mathcal{N}(0, \sigma^2)$.
- Because $\mathbb{E}_n(Y)$ has variance σ^2/n, which is $O_P(1/n)$, then $\sqrt{n}(\mathbb{E}_n(Y) - \mu)$ has variance σ^2 which is $O_P(1)$, i.e. "bounded in probability" or "uniformly tight": A sequence $\{Q_n\}$ is uniformly tight if for all $\epsilon > 0$ there exists an M s.t. $\sup_n P(|Q_n| > M) < \epsilon$.

Motivation continues

- Now, suppose L is continuous and our parameter of interest is the conditional mean $h(L, A, Y) \equiv \mathbb{E}(Y \mid L)$.
- In particular, to estimate $\mathbb{E}(Y \mid L = l)$ there exists at most one individual l with $L_i = l$ and $\mathbb{E}_n(Y \mid L = l) = Y_i$, regardless of n, and clearly we do not have $\sqrt{n}$-consistency.
- Thus, we have to do something else...

Parameteric modelling

- Can we really say that the distribution that generated the data belongs to a parametric model?
- The answer is **no** in most settings. Therefore many argue that non-parametric methods are more desirable. And this is why machine learning methods are blooming.
- However, it is often argued that studying parametric models is useful (i) because they can be good approximations, (ii) sometimes we have knowledge about the data generating mechanism and (iii) they provide the background for understanding non-parametric methods.

PS: a saturated model, because it does not impose restrictions on the data; we just call it a model because it looks like a model, but the model does not put any restrictions on the data generating mechanism.

What is bias

- Systematic bias: We say there is *systematic bias* if the causal estimand of interest is not identified.
  Informally, any structural association between the treatment and the outcome that does not arise from the causal effect of treatment on the outcome.
- Bias due to model misspecification: When we use a statistical model that is misspecified (I give a formal definition of model mis-specification in a later slide).

Motivating example

Suppose the counterfactual data are:

Group:	A			B			C		
Response Y^1:	1	1	1	2	2	2	3	3	3
Response Y^0:	0	0	0	1	1	1	2	2	2

and the average treatment effect $\mathbb{E}(Y^{a=1}) - \mathbb{E}(Y^{a=0}) = 1$.
but we observe:

Group:	A			B			C		
Response Y^1:	1	1	?	?	2	?	3	?	?
Response Y^0:	?	?	0	1	?	1	?	2	2

The naive contrast $\mathbb{E}(Y \mid A = 1) - \mathbb{E}(Y \mid A = 0) = \frac{7}{4} - \frac{6}{5} = 0.55$.

Example from Oliver Dukes.

Example continues

- However, from the table we see that,

$$\hat{\pi}(1, \text{group A}) = \frac{2}{3}$$
$$\hat{\pi}(1, \text{group B}) = \frac{1}{3}$$
$$\hat{\pi}(1, \text{group C}) = \frac{1}{3}$$

- Let us estimate $\mathbb{E}(Y^{a=1})$ by a weighted average, where each observation is weighted by $\frac{1}{\hat{\pi}(1,\text{group X})}$,Group X $\in \{\text{Group A}, \text{Group B}, \text{Group C}\}$,

$$\frac{(1+1)\frac{3}{2} + 2\frac{3}{1} + 3\frac{3}{1}}{\frac{3}{2} + \frac{3}{2} + \frac{3}{1} + \frac{3}{1}} = 2$$

and estimate $\mathbb{E}(Y^{a=0})$ by weighting each observation by $\frac{1}{\hat{\pi}(0,\text{Group X})}$, Group X $\in \{\text{Group A}, \text{Group B}, \text{Group C}\}$,

$$\frac{0\frac{3}{1} + (1+1)\frac{3}{2} + (2+2)\frac{3}{2}}{\frac{3}{1} + \frac{3}{2} + \frac{3}{2} + \frac{3}{2} + \frac{3}{2}} = 1.$$

Estimation when the propensity score is known

When $\pi(a \mid l)$ is a known function, the estimator of $\mathbb{E}(Y^a)$ is

$$\hat{\mu}_{IPW}(a) = \frac{1}{n} \sum_{i=1}^{n} \frac{I(A_i = a) Y_i}{\pi(A_i \mid L_i)}.$$

The propensity score $\pi(a \mid l)$, unlike the function $Q(l, a)$, is known in randomised experiments (it is determined by the investigator). However, in most observational data settings, it is unknown.
PS: This estimator has been known for a long time and is often called the Horvitz Thompson estimator in survey sampling[38].

[38] Daniel G Horvitz and Donovan J Thompson. "A generalization of sampling without replacement from a finite universe". In: *Journal of the American statistical Association* 47.260 (1952), pp. 663–685.

Estimation when the propensity score is unknown

More generally, we can propose a regression model $\pi(A \mid L; \gamma)$ for $\pi(A \mid L)$, and we can consider the estimator

$$\hat{\mu}_{IPW}(a) = \frac{1}{n} \sum_{i=1}^{n} \frac{I(A_i = a) Y_i}{\pi(A_i \mid L_i; \gamma)}.$$

For example, suppose that we fit a logistic regression model and find the MLE $\hat{\gamma}$ of γ, which is the solution to the estimating equation (See slide 248)

$$\sum_{i=1}^{n} \begin{pmatrix} 1 \\ L_i \end{pmatrix} \left(A_i - \frac{\exp(\gamma_1 + \gamma_2^T L_i)}{1 + \exp(\gamma_1 + \gamma_2^T L_i)} \right) = 0.$$