國立清華大學
NATIONAL TSING HUA UNIVERSITY

EE 306001
Probability

Lecture 25: Introduction to Statistics

李祈均

Maximum likelihood estimation

- One way: pick θ, that means pick a specific probability model that the data we observe, $X's$, most likely have occurred

Mathematically:

- Model with unknown parameter(s), $X \sim p_X(x;\theta)$
- ML: pick θ that *"makes the data most likely"*

$$\hat{\theta}_{ML} = \arg\max_{\theta} p_X(x;\theta)$$

Simple example

- Example: $X_1, \ldots, X_n$: i.i.d. exponential(θ)

Try to find a good estimate of θ using the ML approach

$$\hat{\theta}_{ML} = \arg\max_{\theta} p_X(x;\theta)$$

$$p_X(x;\theta) = \prod_{i=1}^{n} \theta e^{-\theta x_i}$$

Take log

$$\ln p_X(x;\theta) = n \ln \theta - \theta \sum_{i=1}^{n} x_i$$

$$\hat{\theta}_{ML} = \arg\max_{\theta} p_X(x;\theta)$$

$$\hat{\theta}_{ML} = \arg\max_{\theta} \ln p_X(x;\theta) = \arg\max_{\theta} \left(n \ln \theta - \theta \sum_{i=1}^{n} x_i \right)$$

How to solve?

Take derivative with respect to θ and set it equals to 0 and solve:

$$\hat{\theta}_{ML} = \frac{n}{x_1 + \cdots + x_n}$$

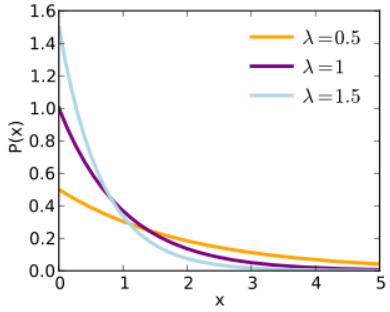
Let's abstractly this about this:

In fact, we have just designed an estimator (function on data) of the following form:

$$\widehat{\Theta}_n = \frac{n}{X_1 + \cdots + X_n}$$

Imagine this as an experiment, once you do the experiment, that each X_i will output a number, and you have the ML-estimate

Desired probability of an estimator



- This estimator, $\widehat{\Theta}_n$, is random
- **Unbiased**: $E\left[\widehat{\Theta}_n\right] = \theta$
  - $\widehat{\Theta}_n$ is a function of data X
  - X is affected by true parameter θ (each θ corresponds to a different model)
  - Is it always true? Not necessary

  Exponential example that we just demonstrated (take $n = 1$)

  $$E[1/X_1] = \infty \neq \theta$$

  This ML estimate is biased estimator (biased upward)

  In general ML estimate is just like this, under some condition, it will turn out to be unbiased

- **Consistent**: $\widehat{\Theta}_n \to \theta$ (convergence in probability)
  - This is good property especially if you have large amount of data
  - ML estimate tend to have this properties (given independent data)!

Exponential example:

$$\frac{(X_1 + \cdots + X_n)}{n} \to E[X] = 1/\theta$$

Knowing, we can look at our estimator:

$$\widehat{\Theta}_n = \frac{n}{X_1 + \cdots + X_n}$$

$$\widehat{\Theta}_n = \frac{n}{X_1 + \cdots + X_n} \to \frac{1}{E[X]} = \theta$$

Weak law of large number, this is true no matter what the true theta is

- $\widehat{\Theta}_n$ is a function of data X
- X is affected by true parameter θ (each θ corresponds to a different model)

One more desired property for an estimator: **small mean square error (MSE)**

$$E\left[\left(\widehat{\Theta}-\theta\right)^2\right] = var\left(\widehat{\Theta}-\theta\right) + \left(E\left[\widehat{\Theta}-\theta\right]\right)^2$$
$$= var\left(\widehat{\Theta}\right) + (\text{bias})^2$$

- Ideally, we want $\widehat{\Theta}$ to be very close θ, so we like the biased term to be zero, and at the same time, the fluctuation (variance of our estimator) is small too!

$$E\left[\left(\widehat{\Theta}-\theta\right)^2\right] = var\left(\widehat{\Theta}-\theta\right) + \left(E\left[\widehat{\Theta}-\theta\right]\right)^2$$
$$= var\left(\widehat{\Theta}\right) + (\text{bias})^2$$

- Let's do a silly example

Assume we have distribution that is normal with unknown mean θ and variance 1

Let's design a simple estimator: just keep saying that mean equals to 100 no matter what

- This estimator has 0 variance, but huge bias term!

- Moral of the story;
  - You can make variance extremely small but pay the price in the bias term
  - There is certain tradeoff between the two
  - We won't cover this further in this class

Revisit our estimation of mean

- $X_1, \ldots, X_n$: iid mean θ, variance σ^2

$$X_i = \theta + W_i$$

W_i: iid, mean 0, variance σ^2

Design an estimator to estimate mean θ using sample mean:

$$\widehat{\Theta}_n = \text{sample mean} = M_n = \frac{X_1 + \cdots + X_n}{n}$$

Revisit this sample mean estimator

- **Unbiased:**

$$\widehat{\Theta}_n = M_n = \frac{X_1 + \cdots + X_n}{n}$$

$$E[M_n] = \theta$$

- **Consistent:**

By weak law of large number:

Sample mean converge to true mean in probability

$$\widehat{\Theta}_n \to \theta$$

- **MSE**

$$var\left(\widehat{\Theta}\right) + (\text{bias})^2 = \frac{\sigma^2}{n} + 0$$

ML estimate of normal distribution parameters

Consider estimating the mean and variance of a normal distribution using n independent observations $X_1, \ldots, X_n$

The parameter that is unknown as $\theta = (\mu, v)$

To do ML estimation, we need get the data likelihood:

$$f_X(x; \mu, v) = \prod_{i=1}^{n} f_{X_i}(x_i; \mu, v) = \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi v}} e^{-(x_i-\mu)^2/2v}$$

$$f_X(x;\mu,v) = \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi v}} e^{-(x_i-\mu)^2/2v}$$

Note:

$$(x_i - \mu)^2 = (x_i - m_n + m_n - \mu)^2$$
$$= (x_i - m_n)^2 + (m_n - \mu)^2 + 2(x_i - m_n)(m_n - \mu)$$

For $i = 1, \dots, n$

$$\sum_{i=1}^{n} (x_i - m_n)(m_n - \mu) = (m_n - \mu) \sum_{i=1}^{n} (x_i - m_n)$$

Define:

$$m_n = \frac{1}{n} \sum_{i=1}^{n} x_i$$

$$\sum_{i=1}^{n}(x_i - m_n)(m_n - \mu) = (m_n - \mu)\sum_{i=1}^{n}(x_i - m_n)$$

Due to our definition,

This term = 0

Now define:

$$s_n^2 = \frac{1}{n}\sum_{i=1}^{n}(x_i - m_n)^2$$

Now we can re-write: $f_X(x;\mu,v) = \prod_{i=1}^{n}\frac{1}{\sqrt{2\pi v}}e^{-(x_i-\mu)^2/2v}$

$$f_X(x;\mu,v) = \frac{1}{(2\pi v)^{n/2}}\exp\left\{-\frac{ns_n^2}{2v}\right\}\exp\left\{-\frac{n(m_n-\mu)^2}{2v}\right\}$$

Now to find ML estimate, find μ, v that maximize that data likelihood function:

$$\log f_X(x; \mu, v) = -\frac{n}{2}\log(2\pi) - \frac{n}{2}\log v - \frac{ns_n^2}{2v} - \frac{n(m_n - \mu)^2}{2v}$$

Differentiate respect to each parameter and set to zero:

$$\mu = m_n, v = s_n^2$$

Sample mean is also the ML estimate for mean!

Sample variance ($s_n^2 = \frac{1}{n}\sum_{i=1}^{n}(x_i - m_n)^2$) is also the ML estimate for variance!

Look closer at the variance estimate

$$v = s_n^2$$

Essentially, we have constructed a function of estimator of the following form:

$$S_n^2 = \frac{1}{n}\sum_{i=1}^{n}(X_i - M_n)^2$$

Is it biased?

$$E[S_n^2] = \frac{1}{n}E\left[\sum_{i=1}^{n}(X_i - M_n)^2\right]$$

$$E[S_n^2] = \frac{1}{n} E\left[\sum_{i=1}^{n} (X_i - M_n)^2\right] = \frac{1}{n} E\left[\sum_{i=1}^{n} X_i^2 - 2M_n \sum_{i=1}^{n} X_i + nM_n^2\right]$$

$$= E\left[\frac{1}{n}\sum_{i=1}^{n} X_i^2 - 2M_n^2 + M_n^2\right] = E\left[\frac{1}{n}\sum_{i=1}^{n} X_i^2 - M_n^2\right]$$

Now noting that:

$$E[M_n] = \mu, E[M_n^2] = \mu^2 + \frac{v}{n}, E\left[X_i^2\right] = \mu^2 + v$$

$$E\left[\frac{1}{n}\sum_{i=1}^{n} X_i^2 - M_n^2\right] = \mu^2 + v - \left(\mu^2 + \frac{v}{n}\right) = \frac{n-1}{n} v$$

This estimator is biased estimator, (though asymptotically unbiased)

Can we get an unbiased estimator?

Sure, just with proper scaling:

$$\hat{S}_n^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - M_n)^2$$

This is an unbiased variance estimator!

- Note sample mean does not always correspond to ML estimate

- Which one to report when you are forced to pick one?
  - Probably sample mean… much easier!

Desired probability of an estimator

- This estimator, $\widehat{\Theta}_n$, is random
- Unbiased: $E\left[\widehat{\Theta}_n\right] = \theta$
- Consistent: $\widehat{\Theta}_n \to \theta$ (convergence in probability)
- Small mean square error (MSE)
  - $E\left[\left(\widehat{\Theta} - \theta\right)^2\right] = var\left(\widehat{\Theta} - \theta\right) + \left(E\left[\widehat{\Theta} - \theta\right]\right)^2 = var\left(\widehat{\Theta}\right) + (\text{bias})^2$

Estimator you already know in classical sense

- Sample mean
  - This is an estimator with very good property
  - Very easy and good estimator for mean of a distribution
- ML estimate
  - Maximum data likelihood
  - Observable evidence is the KING
- They don’t necessary coincide

Confidence interval

- Idea: you want to know how much you can trust a given estimate (say, for example: you estimate the mean to be 2.37)

- Can we construct an interval to say the likely value of ‘true thetas’?

Confidence interval

- Design a $1 - a$ confidence interval
  - $\left[\widehat{\Theta}_n^-, \widehat{\Theta}_n^+\right]$
  - Such that: $P\left(\widehat{\Theta}_n^- \widehat{\Theta}_n^+\right) \geq 1 - a$ for all θ
  - Often this $a = 0.05, 0.025, \text{or } 0.01$

- Note: this interval is 'random' – uppercase rvs!
  - So what you are doing exactly is to construct two other random variables as estimators!

- Interpretation (subtle)
  - Say you have an interval, and given the data you observe, you realize the value of the interval (uppercase -> lowercase)
  - Say it's between 1.97 – 2.56 ($a = 0.05$)
  - Can you say:
    - With probability 0.95, the true theta falls in that interval (1.97, 2.56)?
    - Nope, probability statement is associated with randomness statement
    - θ is a number, the two realized intervals are numbers, so it's either you are 'IN' the interval or 'NOT'
  - Proper way to state this:
    - the interval, that's being constructed by our procedure, should have the property that, with probability 95%, it's going to fall on top of the true value of theta

- Imagine this procedure as experiment
  - You do it once on a day, seeing the data, construct the interval, and yes, the true theta is in
  - You do it on another day, seeing different data, construct the interval, and yes, the true theta is in
  - You do it on another way, seeing different data, but this time, nope!
  - 95% of the days when I use this procedure to construct the confidence interval, I got it right!

- it's a statement about the distribution of these random confidence intervals, how likely are they to fall on top of the true theta
  - It's a statement about probabilities associated with a confidence interval (intervals are random variables)
  - Not about the theta!

Quick example

- Let's construct a CI in the estimation of the mean

$$\widehat{\Theta}_n = (X_1 + X_2 + \cdots + X_n)/n$$

Note, we should probably already know this

- From standard normal table
- For $z = 1.96$: corresponds to .975 (right tail portion 0.025, and both right-left tail portion 0.05)
- The random variable here is sample mean: $\widehat{\Theta}_n$

From CLT:

$$P\left(\frac{|\widehat{\Theta}_n - \theta|}{\sigma/\sqrt{n}} \leq 1.96\right) \approx 0.95$$

Or:

$$P\left(\widehat{\Theta}_n - \frac{1.96\sigma}{\sqrt{n}} \leq \theta \leq \widehat{\Theta}_n + \frac{1.96\sigma}{\sqrt{n}}\right) \approx 0.95$$

Here, we have our confidence interval, so for this exercise of estimating the mean, we can report two things:

1. Sample mean , $\widehat{\Theta}_n$
2. 95% CI:

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

Confidence interval gets smaller as n gets larger, this interval is an interval derived from invoking CLT

- This can be done generally if we continue using CLT as the approach:

Let z be s.t. $\Phi(z) = 1 - a/2$

remember if we set $a = 0.95, 1 - \frac{a}{2} = 0.975, \Phi(z) = 0.975$ for $z = 1.96$

Now we can easily derive the CI:

$$P\left(\widehat{\Theta}_n - \frac{z\sigma}{\sqrt{n}} \leq \theta \leq \widehat{\Theta}_n + \frac{z\sigma}{\sqrt{n}}\right) \approx 1 - a$$

This is an 'approximate' $1 - a$ CI interval, why?

Because of CLT – that limiting property need to be used with care!

All is good, but do we know σ in general?

- Usually no, so we have a couple options to go to:

1. Use upper bound on σ

if X_i Bernoulli random variable, then the standard is at most ½ ($\sigma \leq \frac{1}{2}$)

2. Use ad-hoc estimate

if X_i Bernoulli random variable, we know standard deviation should be $\sqrt{p(1-p)}$, so if we have an estimate of p using $\widehat{\Theta}_n$, then estimated standard deviation can be $\sqrt{\widehat{\Theta}_n(1-\widehat{\Theta}_n)}$

3. Use generic estimate of variance
   - Start from $\sigma^2 = E[(X_i - \theta)^2]$
   - Variance (law of large number, averaging variance goes to true variance)

$$\hat{\sigma}_n^2 = \frac{1}{n}\sum_{i=1}^{n}(X_i - \theta)^2 \to \sigma^2$$

   - Problem? Don't know θ, use estimate, also make it unbiased (like we talk about in previous pages!)

$$\hat{S}_n^2 = \frac{1}{n-1}\sum_{i=1}^{n}\left(X_i - \widehat{\Theta}_n\right)^2$$

Now back to CI

$$P\left(\widehat{\Theta}_n - \frac{z\sigma}{\sqrt{n}} \leq \theta \leq \widehat{\Theta}_n + \frac{z\sigma}{\sqrt{n}}\right) \approx 1 - a$$

If we approximate σ (this is the true standard deviation of the sample mean) using $\hat{S}_n$

Then we obtain the following CI interval

$$\widehat{\Theta}_n - \frac{z\hat{S}_n}{\sqrt{n}} \leq \theta \leq \widehat{\Theta}_n + \frac{z\hat{S}_n}{\sqrt{n}}$$

$$\widehat{\Theta}_n - \frac{z\hat{S}_n}{\sqrt{n}} \leq \theta \leq \widehat{\Theta}_n + \frac{z\hat{S}_n}{\sqrt{n}}$$

Look at this closer, there are TWO approximation going on,

1. We assume $\widehat{\Theta}_n$ behave like normal using CLT
2. The true variance of $\widehat{\Theta}_n$ (which should be v/n, where v is the variance of individual X_i) is approximated using $\hat{S}_n^2/n$

- For the CI that we talked about earlier, essentially, we have treated the following random variable as normal:

$$\frac{\widehat{\Theta}_n - \theta}{\sqrt{var\,(\widehat{\Theta}_n)}}$$

When $\sqrt{var\,(\widehat{\Theta}_n)}$ is unknown and using the approximation $\hat{S}_n^2$

$$\widehat{\Theta}_n = \frac{(X_1 + \cdots + X_n)}{n}$$

$$T_n = \frac{\widehat{\Theta}_n - \theta}{\sqrt{var\,(\widehat{\Theta}_n)}} = \frac{\widehat{\Theta}_n - \theta}{\sqrt{\hat{S}_n^2/n}} = \frac{\sqrt{n}(\widehat{\Theta}_n - \theta)}{\hat{S}_n}$$

In general, this is not normal, too many approximation going on

- We won't go into details, just to demonstrate the mechanics of workings out the problems:

In general, we are out of luck of working out this problem easily, but if we have the following condition:

$$X_i \sim \text{normal}$$

Then, we can show that T_n PDF does not depend on the mean θ, and true variance v

It is essentially called t-distribution with $n-1$ degree of freedom

t-distribution is just like normal distribution:

- Bell shape with heavier tail
- CDF is pre-computed just like normal distribution

- In practical life:

When X_i is normal-like and n relatively small, a confidence interval is the following:

$$\widehat{\Theta}_n - \frac{z\hat{S}_n}{\sqrt{n}} \leq \theta \leq \widehat{\Theta}_n + \frac{z\hat{S}_n}{\sqrt{n}}$$

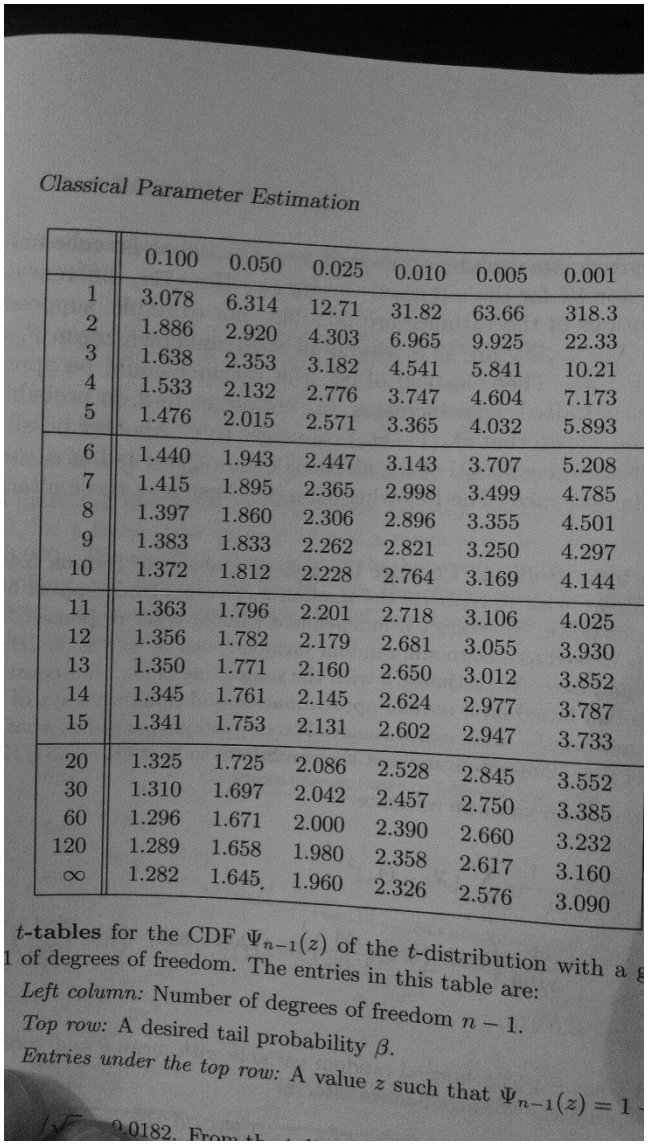
In order to find z, instead of looking at normal table, look at t table with $n - 1$ degree of freedom

That is find z such that

$$\Psi_{n-1}(z) = 1 - \frac{a}{2}$$

n is the number of data points

For a given confidence a



Classical Parameter Estimation

	0.100	0.050	0.025	0.010	0.005	0.001
1	3.078	6.314	12.71	31.82	63.66	318.3
2	1.886	2.920	4.303	6.965	9.925	22.33
3	1.638	2.353	3.182	4.541	5.841	10.21
4	1.533	2.132	2.776	3.747	4.604	7.173
5	1.476	2.015	2.571	3.365	4.032	5.893
6	1.440	1.943	2.447	3.143	3.707	5.208
7	1.415	1.895	2.365	2.998	3.499	4.785
8	1.397	1.860	2.306	2.896	3.355	4.501
9	1.383	1.833	2.262	2.821	3.250	4.297
10	1.372	1.812	2.228	2.764	3.169	4.144
11	1.363	1.796	2.201	2.718	3.106	4.025
12	1.356	1.782	2.179	2.681	3.055	3.930
13	1.350	1.771	2.160	2.650	3.012	3.852
14	1.345	1.761	2.145	2.624	2.977	3.787
15	1.341	1.753	2.131	2.602	2.947	3.733
20	1.325	1.725	2.086	2.528	2.845	3.552
30	1.310	1.697	2.042	2.457	2.750	3.385
60	1.296	1.671	2.000	2.390	2.660	3.232
120	1.289	1.658	1.980	2.358	2.617	3.160
∞	1.282	1.645	1.960	2.326	2.576	3.090

t-tables for the CDF $\Psi_{n-1}(z)$ of the t-distribution with a g
1 of degrees of freedom. The entries in this table are:
Left column: Number of degrees of freedom $n-1$.
Top row: A desired tail probability β.
Entries under the top row: A value z such that $\Psi_{n-1}(z) = 1$ -

$\sqrt{\quad}$ 2.0182. From th

Example

The weight of an object is measured eight times using an electronic scale that reports the true weight plus a random error that is normally distributed with zero mean and unknown variance.

Assume that the errors in the in the observations are independent. We obtain the following results:

$$0.5547, 0.5404, 0.6364, 0.6438, 0.4917, 0.5674, 0.5564, 0.6066$$

Let's compute a 95% confidence interval ($a = 0.05$) using the t-distribution

First let's compute the sample mean of the results obtained:

$$\widehat{\Theta}_n = 0.5747$$

Now, lets' compute the sample variance associated with $\widehat{\Theta}_n$

$$\frac{\hat{S}_n^2}{n} = \frac{1}{n(n-1)} \sum_{i=1}^{n} \left(X_i - \widehat{\Theta}_n\right)^2 = 3.2952 * 10^{-4}$$

Then the sample standard deviation:

$$\frac{\hat{S}_n}{\sqrt{n}} = 0.0182$$

- Okay, now we are ready to use t-distribution with n-1 degree of freedom

$n = 8$ (degree of freedom $= 7$)

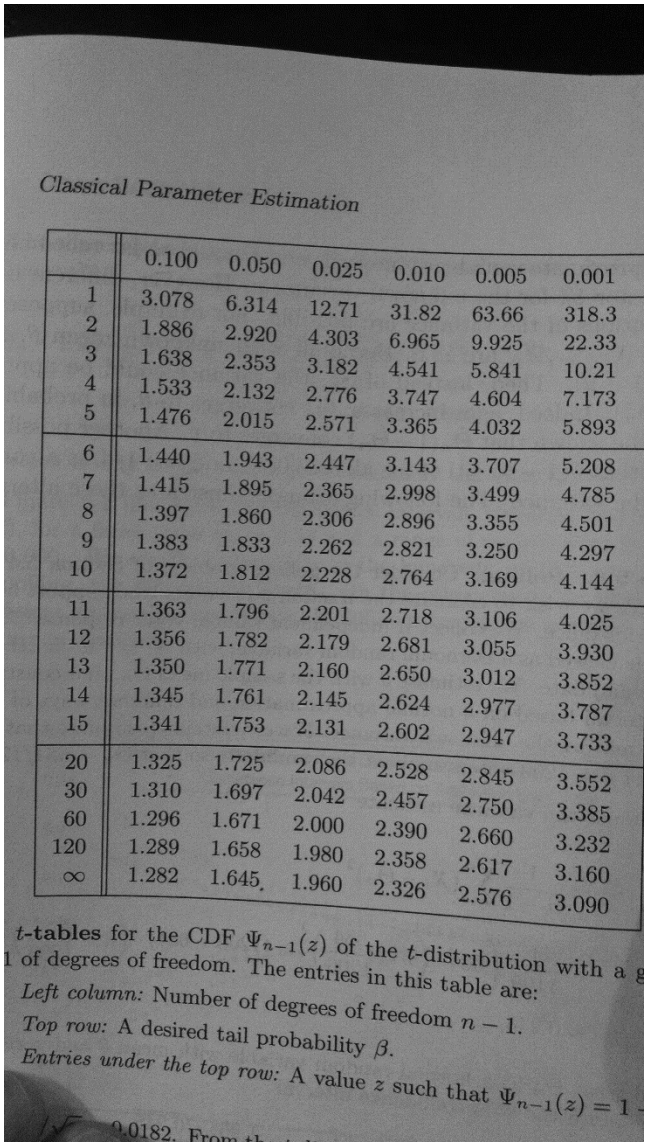
So if we set $a = 0.05$

Find z in the following:

$$P\left(\frac{\left|\widehat{\Theta}_n - \theta\right|}{\hat{S}_n/\sqrt{n}} \leq z\right) = 0.95$$

Find z such that $1 - \Psi_7(z) = \frac{a}{2} = 0.025$

$z = 2.365$



Classical Parameter Estimation

	0.100	0.050	0.025	0.010	0.005	0.001
1	3.078	6.314	12.71	31.82	63.66	318.3
2	1.886	2.920	4.303	6.965	9.925	22.33
3	1.638	2.353	3.182	4.541	5.841	10.21
4	1.533	2.132	2.776	3.747	4.604	7.173
5	1.476	2.015	2.571	3.365	4.032	5.893
6	1.440	1.943	2.447	3.143	3.707	5.208
7	1.415	1.895	2.365	2.998	3.499	4.785
8	1.397	1.860	2.306	2.896	3.355	4.501
9	1.383	1.833	2.262	2.821	3.250	4.297
10	1.372	1.812	2.228	2.764	3.169	4.144
11	1.363	1.796	2.201	2.718	3.106	4.025
12	1.356	1.782	2.179	2.681	3.055	3.930
13	1.350	1.771	2.160	2.650	3.012	3.852
14	1.345	1.761	2.145	2.624	2.977	3.787
15	1.341	1.753	2.131	2.602	2.947	3.733
20	1.325	1.725	2.086	2.528	2.845	3.552
30	1.310	1.697	2.042	2.457	2.750	3.385
60	1.296	1.671	2.000	2.390	2.660	3.232
120	1.289	1.658	1.980	2.358	2.617	3.160
∞	1.282	1.645	1.960	2.326	2.576	3.090

t-tables for the CDF $\Psi_{n-1}(z)$ of the t-distribution with a g
of degrees of freedom. The entries in this table are:
Left column: Number of degrees of freedom $n-1$.
Top row: A desired tail probability β.
Entries under the top row: A value z such that $\Psi_{n-1}(z) = 1$

0.0182. From th

- So the interval is the following:

$$[0.531, 0.618]$$

Compare it with using normal table:

$$\widehat{\Theta}_n - \frac{z\hat{S}_n}{\sqrt{n}} \leq \theta \leq \widehat{\Theta}_n + \frac{z\hat{S}_n}{\sqrt{n}}$$

$$z = 1.96$$

We have the following interval:

$$[0.539, 0.610]$$