Code Structure

/thesis-research-2.0

|-- data

|-- raw

|-- spectroscopy

|-- photometry

|-- processed

|-- spectro\_photo

|-- etg\_selected

|-- foundation

|-- fp\_sample

|-- pv\_sample

|-- smoothing

|-- bulk\_flow

|-- utils

|-- cosmoFunc.py

|-- calc\_kcor.py

|-- src

|-- preprocessing.py

|-- select\_etg.py

|-- fit\_fp.py

|--

|-- README.md

1. data
   1. raw
      1. spectroscopy

* LAMOST DR7: directly obtained from Khaled.
* 6dFGS: using table 2 and 8
* SDSS: pake query di bawah:

SELECT s.specobjid, p.objID, s.ra, s.dec, s.plate, s.instrument, s.mjd, s.fiberid, p.devMag\_u, p.devRad\_u, devAB\_u, p.devMag\_r, p.devRad\_r, devAB\_r, s.z,s.zErr,s.veldisp,s.veldispErr, em.sigmaStars, em.sigmaStarsErr

into mydb.SDSS\_spectro

From SpecObjAll as s

JOIN emissionLinesPort em ON (em.specObjID = s.specobjid)

JOIN PhotoObjAll p ON (p.specObjID = s.SpecObjID)

WHERE

(s.sdssPrimary = 1)

AND (s.z <= 0.1)

**Fotometri 2MASS**

Koordinat diinput ke 2MASS XSC IPAC. Dengan cone search radius 5”. Data yang diambil:

* designation: source designation formed from sexigesimal coordinates
* glon dan glat:  galactic longitude (decimal deg) based on peak pixel
* j\_ba, h\_ba, k\_ba: minor/major axis ratio fit to the 3-sigma isophote
* sup\_ba: minor/major axis ratio fit to 3-sig. super-coadd isophote
* r\_ext: extrapolation/total radius
* j\_m\_ext, h\_m\_ext, k\_m\_ext: mag from fit extrapolation
* j\_r\_eff, h\_r\_eff, k\_r\_eff: half-light (integrated half-flux point) radius
* cc\_flg: indicates artifact contamination and/or confusion
* j\_m\_5, h\_m\_5, k\_m\_5: 5 arcsec radius circular aperture magnitude
* j\_m\_10, h\_m\_10, k\_m\_10: 10 arcsec radius circular aperture magnitude
* j\_m\_20, h\_m\_20, k\_m\_20: 20 arcsec radius circular aperture magnitude

preprocessing.ipynb

PART 1: menghitung variabel-variabel relevan

* ngambil data dari raw\_data/spectrophotometry 🡪 udah kombinasi data spektroskopi dan fotometri. Step untuk mendapatkan data ini lupa ngapain aja, tapi kemungkinan:
* Ubah format file jadi csv (dari fits atau ascii)
* Join data spektroskopi dan fotometri (2MASS)
* Join dengan data grup dan cluster (buat dapetin redshift)
* Bubuhin ‘SDSS’ dan ‘LAMOST’ ke objID SDSS dan obsid LAMOST
* Untuk galaksi SDSS dan LAMOST, apply kriteria seleksi dari John:
  + fit\_ok\_j == ‘OK’
  + log\_r\_h\_model\_j > 0.
  + red\_chi\_j <= 2.
* Hitung PSF-corrected radii:
  + Data SDSS dan LAMOST:
    - Hitung besar koreksi: delta\_r\_j = 10^log\_r\_h\_smodel\_j – 10^log\_r\_h\_model\_j
    - Hitung radius terkoreksi: 10^log\_r\_h\_app\_j – delta\_r\_j
  + Data 6DFGS:
    - Radius terkoreksi langsung dihitung dari 10^Jlogr
* Hitung redshift dalam CMB frame untuk setiap galaksi menggunakan code dari Khaled, kemudian gunakan data redshift grup atau cluster jika tersedia (perlu join individual survey dengan data grup).
* Apply koreksi apertur terhadap dispersi kecepatan:
  + Pake rumus empiris buat konversi J band ke R band
  + Hitung koreksi apertur menggunakan nilai apertur masing-masing survei:
    - LAMOST: apertur 1.65”, pake kolom veldisp dan veldisp\_err
    - SDSS: apertur 1.5”, pake kolom sigmaStars dan sigmaStarsErr
    - 6dFGS: apertur 3.35”, pake kolom Vd dan e\_Vd
* Hitung ekstingsi Galaksi di pita JHK menggunakan dustmaps (masukin ke utils): input ra dec dalam degree (cek!)
* Hitung K-corrections (masukin ke utils)
  + Perlu input redshift heliosentris, warna J2-H2 dan J2-Ks2.

PART 2: MENURUNKAN r, s, I (mungkin codenya dipisah aja)

* Generate redshift-distance lookup table (masukin ke utils/cosmoFunc)
* Hitung r:
  + Hitung estimasi jarak dari redshift cmb cluster
  + Hitung radius menggunakan circularized apparent radii: radius sudut dikali akar rasio b/a
  + Hitung menggunakan rumus r (konversi ke jarak diameter sudut)
* Hitung i:
  + Hitung menggunakan rumus i yang panjang itu (yang ada koreksi macem2), tapi koreksi evolusi diabaikan
* Hitung s: tinggal log10 aja, tapi untuk SDSS ada nilai negatif dan outlier ekstrim (nilainya sangat tinggi), jadi waktu itu diseleksi kuantil < 0.995
* Save data yang sudah ada rsi 🡪 masuk ke first\_step

PART 3: KALIBRASI DISPERSI KECEPATAN

PART 3.1: TENTUKAN GALAKSI YANG ADA PENGUKURAN BERULANG

* Untuk mendapatkan kalibrasi yang lebih baik, tidak dilakukan seleksi morfologi terlebih dahulu agar menggunakan data sebanyak mungkin
* Langkah pertama: ambil data dari folder 1st\_step: 2MASS\_id, masing2 id survei, dispersi kecepatan dan errornya
* Join ketiga survei untuk mendapatkan galaksi yang overlap: 6df-lamost, 6df-sdss, dan sdss-lamost. Kemudian join lagi untuk mendapatkan list galaksi yang ada di ketiga survei.
* Buang galaksi yang ada 3 repeat measurements dari list 2 itu, kemudian semuanya concat dan join dengan data dispersi kecepatan
* Hasilnya disave agar bisa generate grafik dan melakukan fine tuning

PART 3.2: CARI ERROR SCALING (HANYA UNTUK SDSS DAN LAMOST) 🡪 sepertinya perlu dirombak, pelajari lagi algoritmanya

* Buang galaksi yang eror dan offsetnya sangat tinggi
* Pertama, hitung scaling eror lamost: tebakan awal 1.0
* kemudian hitung statistik epsilon
* Hitung scaling yang baru
* Repeat until convergent (di code sekarang kriteria stopping hanya Nmax)
* Print nilainya

PART 3.3: HITUNG OFFSET DENGAN ALGORITMA WEGNER

* Sepertinya tinggal copas
* Print nilainya
* Inspeksi distribusi chi setelah apply scaling dan offset
* Apply kalibrasi dan simpan ke kolom baru, save hasilnya

PART 4: APPLY SELECTION CRITERIA

* Seleksi:
  + Upper redshift limit z\_cmb\_group < 16120
  + Lower redshift limit z\_cmb\_group > 3000
  + Magnitude limit < 13.65
  + Veldisp limit > log(112) + offset yang diperoleh
* Hitung galaksi yang diexclude di tiap seleksi
* Lakukan yang sama untuk galaksi low redshift -> diexclude dari fitting FP karena scatternya besar, tapi untuk perhitungan PV dimasukkan kembali karena harusnya sangat akurat.

PART 5: TURUNKAN EROR FOTOMETRI (kenapa ini gak sebelum selection criteria? find out why)

* Sama seperti kalibrasi veldisp, kali ini pilih galaksi yang unik saja (karena fotometrinya harusnya sama persis).
* Kolom yang diperlukan: 2MASS id, j\_m\_ext, extinction\_j, i\_j, i\_h, i\_k. Semua diconcat, terus drop duplicates 2MASS id
* Hitung warna, kalkulasi variansi tiap bin magnitudo, plot
* Fitting terhadap piecewise linear: konstan di m<m\_0, linear setelahnya. Perlu automasi buang outlier
* Hitung eror fotometri i dan r: eror r = 0.5 eror i

PART 6: SELECT ETG

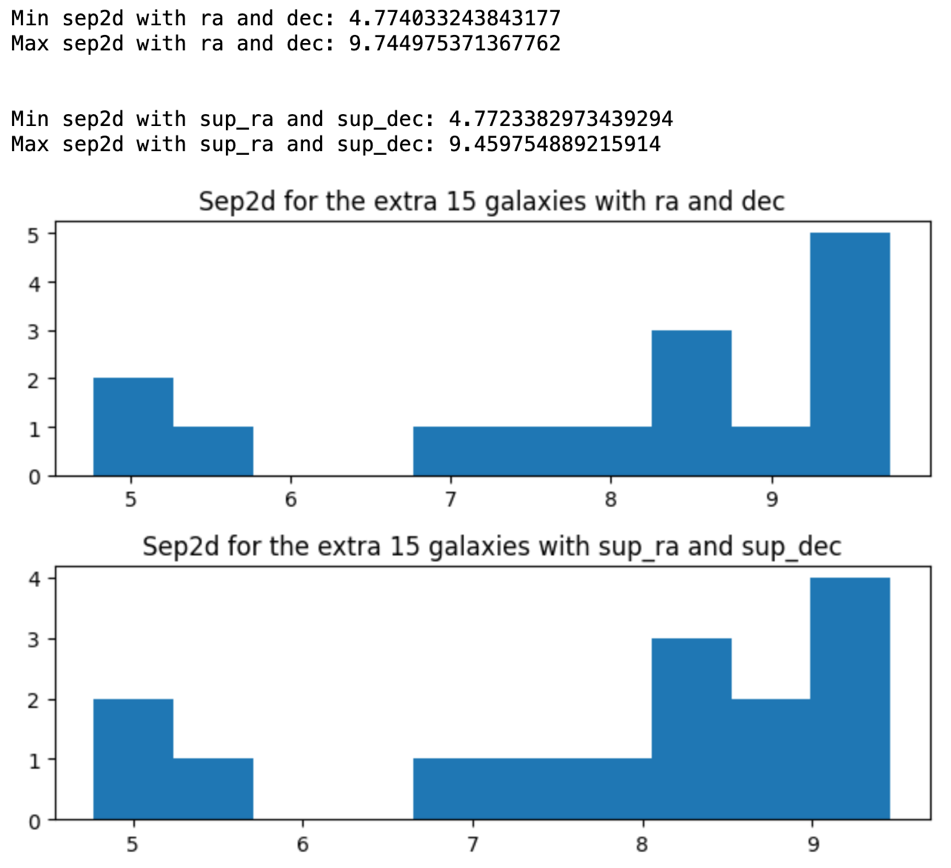
* 6dFGS: pake sampel terbaru Khaled
* SDSS: pake sampel Cullan
* LAMOST: pake seleksi visual dari John

PART 7: SELEKSI KOLOM FINAL UNTUK FP

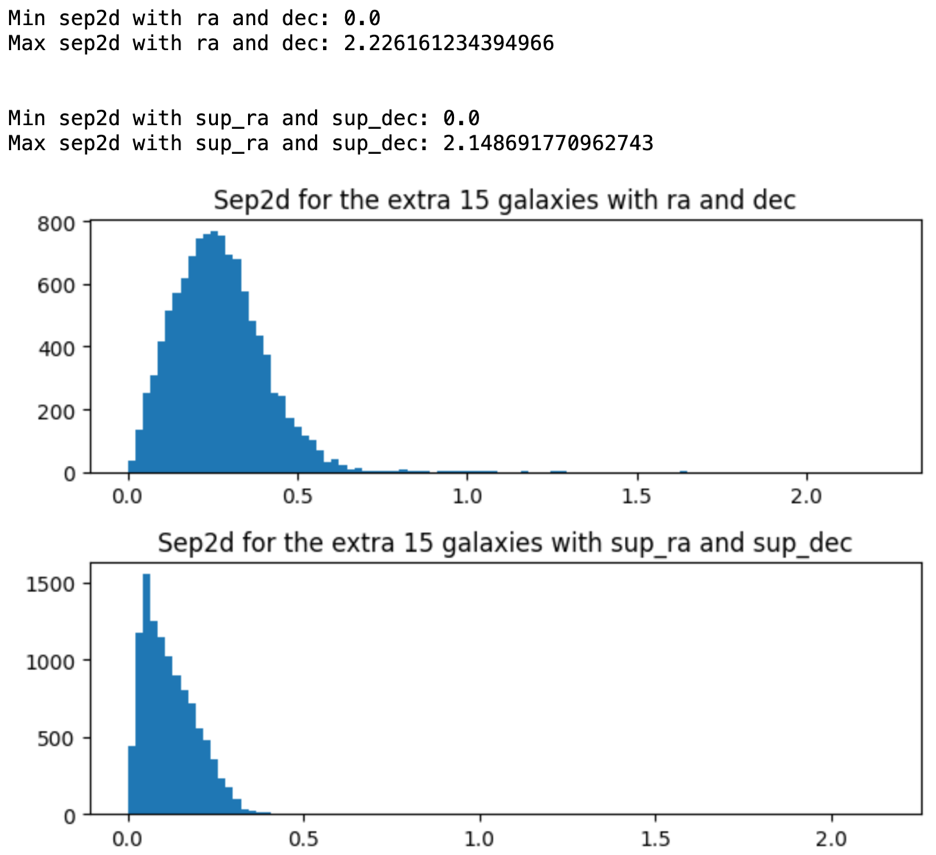
* Kolom yang diperlukan untuk fitting FP: ['2MASS', ‘survei id’, 'ra', 'dec', 'z\_hel', 'z\_cmb', 'z\_cmb\_group', 'j\_m\_ext', 'extinction\_j', 'kcor\_j', 'r\_j', 'er\_j', 's\_corr', 'es\_corr', 'i\_j', 'ei\_j']
* Combine data SDSS dan LAMOST

**Data 6dFGS**

* Ada 2 sumber data: dari Vizier ([di sini](https://vizier.cds.unistra.fr/viz-bin/VizieR-3?-source=+J%2FMNRAS%2F443%2F1231%2FFPsample&-from=nav&-nav=cat%3AJ%2FMNRAS%2F443%2F1231%26tab%3A%7BJ%2FMNRAS%2F443%2F1231%2FFPsample%7D%26key%3Asource%3DJ%2FMNRAS%2F443%2F1231%2Ftable2%26HTTPPRM%3A%26)) atau dari paper MNRAS supplementary data (table 2, 4, 8).
* Schema keduanya beda, yang vizier tampaknya lebih lengkap (ada raw Vd). Yang dipake dulu itu data dari Vizier, jadi semua parameter FP diturunkan dari awal (lupa kenapa, most likely karena ada step yang berbeda).
* Data 6dFGS tetep perlu ambil data fotometri 2MASS (mainly buat surface brightness, extrapolated magnitude, dan axial ratio). 2MASS harus query pake sky coordinates (opsi lain: download semua data 2MASS, merge by 2MASS id). Di vizier, table2 dan FPsample punya RA DEC masing2, FPsample nama kolomnya RAJ2000 dan DEJ2000, sementara table veldisp Namanya \_RA dan \_DE, dan ada perbedaan (max 1.6”). Sementara RAJ2000 dan DEJ2000 nilainya konsisten dengan kolom ra dan dec di tabel FP Campbell (tabel8).
* Query ra dec ke 2mass return semua galaksi (11102) kalo cone searchnya 2.5”. Keliatannya harus 2 step: cone search besar (maybe 10”?), terus crossmatch lagi berdasarkan sup\_ra dan sup\_dec. Also for some reason jumlah row hasil query bisa lebih banyak daripada jumlah row input. Kalo 10”, dapet 11117 galaksi (ekstra 15 galaksi)
* Eksperimen: pake ra dec langsung (1 step) atau sup\_ra dan sup\_dec (2 step):
  + Pertama cari 15 galaksi ekstranya (semua response dari 2mass unik, ga ada id ganda). Terus crossmatch ke tabel input awal (11102 galaksi) pake 2mass id, dapetlah 15 galaksi yang ekstra. Terus liat distribusi separasi sudut 15 galaksi ini, hasilnya:



Sementara untuk galaksi common (all 11102):



Jadi mungkin memang perlu 2 step: pertama cone search yang besar (5”) biar ga ada galaksi yang kelewat, lalu crossmatch lagi dengan sup\_ra dan sup\_dec dengan threshold 2.2”

* Tapi bukannya masalah output lebih banyak daripada input bisa diselesaikan via one-to-one match? Jadi jumlah row output dipastikan sama dengan jumlah row input. Kemudian untuk mergingnya juga lebih gampang karena tinggal merge by index (data gak keacak, udah dicek).
  + Sudah dicek di data 6dFGS. Dengan cone 10”, semua galaksi ada pasangannya 🡪 tapi untuk SDSS dan LAMOST sepertinya harus diturunkan ke sekitar 1”-2” agar mendapatkan galaksi yang akurat.
* Sepertinya final stepnya: dari data koordinat ra dec, feed ke 2MASS XSC, centang 1 to 1 match, cone search 2.5” (separasi max di 6dfgs 2.23”). Merge dengan data spektroskopi awal by index
* Jadi cara merging data 6dFGS:
  + Data FP, buang kolom2:

**18 Feb 2024**

* Dulu buat perhitungan (kayak koreksi apertur gitu2) pake circularized radius (radius \* sqrt(axial ratio)). Tapi ketika dibandingin ke data bersih Campbell hasilnya beda.
* Terus masih ada sedikit perbedaan antara hasil aperture correction dengan data campbell (kemungkinan bedanya di galaksi yang punya beberapa pengukuran, belum dicek)

A graph with a line

Description automatically generated

A graph with numbers and lines

Description automatically generated

**19 Feb 2024**

* Ganti query SDSS, seleksi sigma > 0 dan sigma\_err > 0. Lalu sigma\_err < 1000.

SELECT s.specobjid, p.objID, s.ra, s.dec, s.plate, s.instrument, s.mjd, s.fiberid, p.devMag\_u, p.devRad\_u, devAB\_u, p.devMag\_r, p.devRad\_r, devAB\_r, s.z,s.zErr,s.veldisp,s.veldispErr, em.sigmaStars, em.sigmaStarsErr

into mydb.SDSS\_spectro\_20240219

From SpecObjAll as s

JOIN emissionLinesPort em ON (em.specObjID = s.specobjid)

JOIN PhotoObjAll p ON (p.specObjID = s.SpecObjID)

WHERE

(s.sdssPrimary = 1)

AND (s.z <= 0.1)

AND (em.sigmaStars > 0)

AND (em.sigmaStarsErr > 0)

AND (em.sigmaStarsErr < 1000)

* Hasilnya 280k galaksi SDSS.
* rsi udah berhasil diturunkan (derive\_rsi.py). Setelah ngecek hasil rsi 6dFGS terhadap Campbell, ada sistematis. Kayaknya karena kita pake circularized radius (dikali akar axial ratio).

A graph of a function

Description automatically generated with medium confidence

**20 Feb 2024**

* Lagi ngebersihin code buat nyari error scaling. Hasilnya lumayan bisa direproduce (mirip2 dengan hasil sebelumnya).
* Concern: dulu LAMOST duluan yang scalingnya dicari, nilainya 2 koma, sementara SDSS 1 koma. Kalo urutannya dibalik (SDSS yang dicari duluan), jadinya SDSS yang 2 koma, LAMOST 1 koma. Apakah ini masalah?

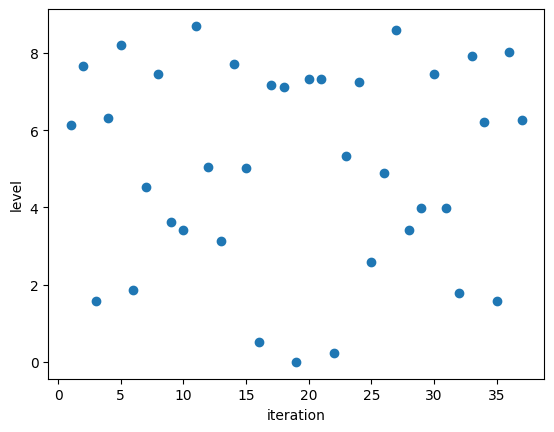
**21 Feb 2024**

* Ternyata hasil output 2MASS XSC gak unique. Setelah dicek, semua data raw itu galaksinya unik (berdasarkan masing2 ID, i.e. 6dFGS\_id, objID, dan obsid). Tapi di data raw 2mass, uniqueness field ‘designation’ nggak 100%. Mungkin kasusnya begini:

A group of planets in space

Description automatically generated

* Sepertinya perlu langkah yang lebih jelas untuk combine data spektroskopi dan fotometri.
* Untuk saat ini, untuk setiap nilai ‘designation’/’tmass’, pilih row dengan separasi terkecil di script combine\_spectrophoto.py.
* Kayaknya script sigcorr2.awk udah berhasil diconvert ke python (dengan vectorized). Beberapa concern:
  1. Line 41: ssig -= totoff 🡪 kenapa dilakukan? Bukannya akan ngaruh ke iterasi berikutnya (lama2 totoff akan semakin kecil)
  2. Dari sekian banyak simulasi, totoffnya diapain? Sekarang Cuma make totoff hasil simulasi dan iterasi terakhir. Kayaknya logicnya: iterasi while itu buat nyari total offset yang menghasilkan residual yang tidak signifikan (maxrat < target), tapi entah kenapa totoffnya nggak kumulasi dari iterasi2 sebelumnya.
  3. ‘level’ gak keliatan konvergen untuk tiap iterasi. By chance dapet yang ketiga levelnya kecil 🡪 estimasi terbaik.



* 1. Sepertinya karena pergeseran totoff untuk setiap iterasi besar, jadinya overshoot dari nilai optimalnya.
* FOUND THE ERROR! Karena line 41 harusnya sig = ssig – totoff. Essentially harus selalu pake nilai ssig awal, bukan terus2an dikurangin.

**25 Feb 2024**

* Bikin adjustment: totoff dari semua simulasi disave ke file totoffs.csv. Terus dibikin histogramnya untuk 6dFGS dan LAMOST. Nilai offset diambil dari mean distribusi masing2, erornya dari standard deviation (jadi eror dari simulasi Monte Carlo).
* Plot totoffs:

A graph of a graph

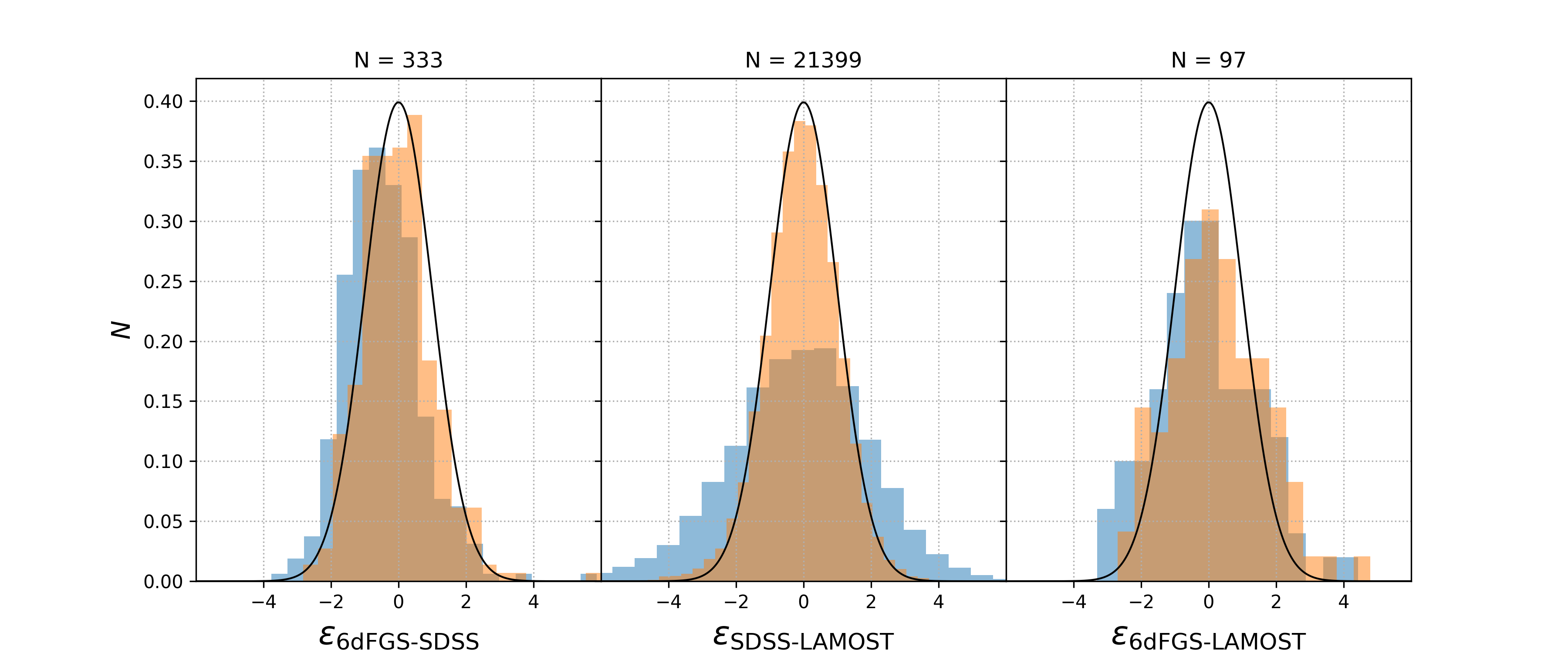
Description automatically generated with medium confidence

* Kalo cari scaling LAMOST dulu (5sigma clipping)

A graph of a function

Description automatically generated

* Kalo cari SDSS scaling dulu (5sigma clipping)



* Mungkin akan pake LAMOST dulu 🡪 keliatannya LAMOST yang sangat underestimated, dilihat dari komparasi 6dFGS-LAMOST (needs extra umph). Tapi kalo cuma liat 2 komparasi lainnya, keliatan lebih bagus.

**26 Feb 2024 (bimbingan)**

* Data Vizier sangat reliable, gapapa ambil data dari situ
* Pake circularized radii untuk perhitungan apapun yang berkaitan dengan radius, tapi harus konsisten dari awal sampe akhir
* Untuk perbandingan veldisp, harus pake pengukuran galaksi yang pressure-supported, jadi memang dari awal perlu seleksi ETG

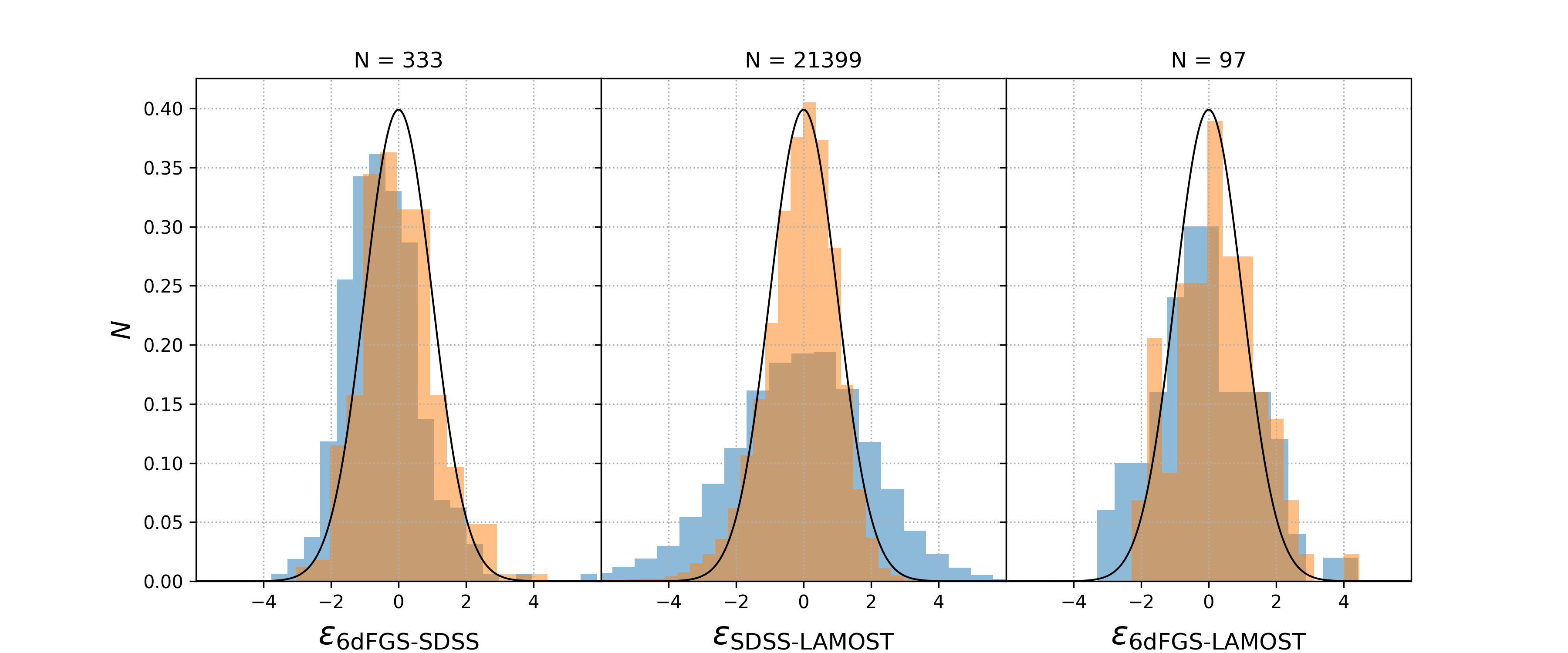
**2 Mar 2024**

* Perlu mikirin cara paling efisien buat dapetin offset 6dFGS (karena dipake untuk seleksi veldisp). Sekarang sih mikirnya ambil aja langsung dari list totoff 🡪 done
* Galaksi redshift rendah disimpen karena nanti akan digunakan untuk perhitungan logdist. Tapi diexclude di fitting FP karena scatternya besar (PV besar).

**3 Mar 2024**

* Kayaknya implementasi finding error scaling salah… kalo di artikel Matthew, perhitungan scaling harusnya hanya dilakukan terhadap survei yang dijadiin fiducial (6dFGS). Tapi yang saya kerjain, 6dFGS malah ga dipake buat jadi komparasi (cuma diset k6dFGS = 1), tapi klamost dan ksdss dicarinya dari kedua survei itu sendiri, jadi mungkin menghasilkan solusi yang tidak unik.
* Waktu itu (mungkin) rationale nya:
  + Make 6dFGS sebagai fiducial terlalu sedikit, after all komparasi yang paling parah itu SDSS – LAMOST.
  + Dari ketiga survei, eror SDSS dan LAMOST jauh lebih kecil disbanding 6dFGS 🡪 jadi yang 6dFGS ga mau diutak atik.
* Kayaknya pemilihan sampel yang jelas harusnya dilakukan dari awal? Terkhusus SDSS dan 6dFGS, karena pada akhirnya dicrossmatch dengan data eksternal. Kalo LAMOST emang beneran seleksi ETG aja di akhir.

**4 Mar 2024**

* Tried the other two approaches:
  1. First approach: assume ksdss = k6df = 1, so we need to only find klamost. By comparing with SDSS, I obtained klamost = 2.578, whereas with 6dFGS, I obtained klamost = 3.912. The SDSS comparison yields better pairwise distribution

The one from 6dFGS seems to overestimate the error for SDSS-LAMOST comparison

**A graph of a function

Description automatically generated**

* 1. Second approach: only assume k6df = 1.0 or ksdss = 1.0, the other 2 scalings to be determined. Using 6dFGS as the fiducial, I obtained ksdss = 5.349 and klamost = 3.913 (definitely will overestimate the error). Whereas I got NaN for k6df if SDSS is used as the fiducial 🡪 not comfortable comparing with SDSS since k6df > 1 by construct, inflating the errors even more.
  2. Argument for using the first approach (and comparing with SDSS): the number of galaxies used in this comparison is the largest (20k galaxies).

**7-8 Mar 2024**

* I’m doing it… dari awal pake data yang selection criterianya lebih jelas aja, yaitu:
  1. 6dFGS pake 8803 galaksi (atau 8885? lupa bedanya apa.. need to make sure) (Magoulas et al., Said et al. 2020)
  2. SDSS pake 34059 data Cullan? 🡪 kenapa ga pake 24848 data Khaled ya?
  3. LAMOST pake visual inspection John langsung?
* HAJAR DULU LAH, GA KELAR2 NANTI.. SO:
  1. 6dFGS pake 8885 dulu (data dari email Khaled subject ‘New 6dF’)
  2. SDSS pake 34059 data Cullan
  3. LAMOST pake full DR7 + 2MASS + visual inspection John
* Berarti code dan data yang sekarang harus di-backup. Bikin folder ‘archive’ di root directory, isinya ‘data/’ dan ‘src/’, masing2 archive (dalam zip) si data lama sama code lama.
* Untuk LAMOST berdasarkan inspeksi John yang bagus cuma 2 ribu… itu dari sampel yang udah censored sih, jadi mungkin untuk perbandingan veldisp pake yang full aja (toh dicrossmatch dengan 2 tabel yang seharusnya udah ETG semua isinya).
* Jadi ceklis yang harus dilakukan sekarang:
  1. Pindahin data dan code sekarang ke archive folder ✅

filenya data\_20240308.zip dan src\_20240308.zip

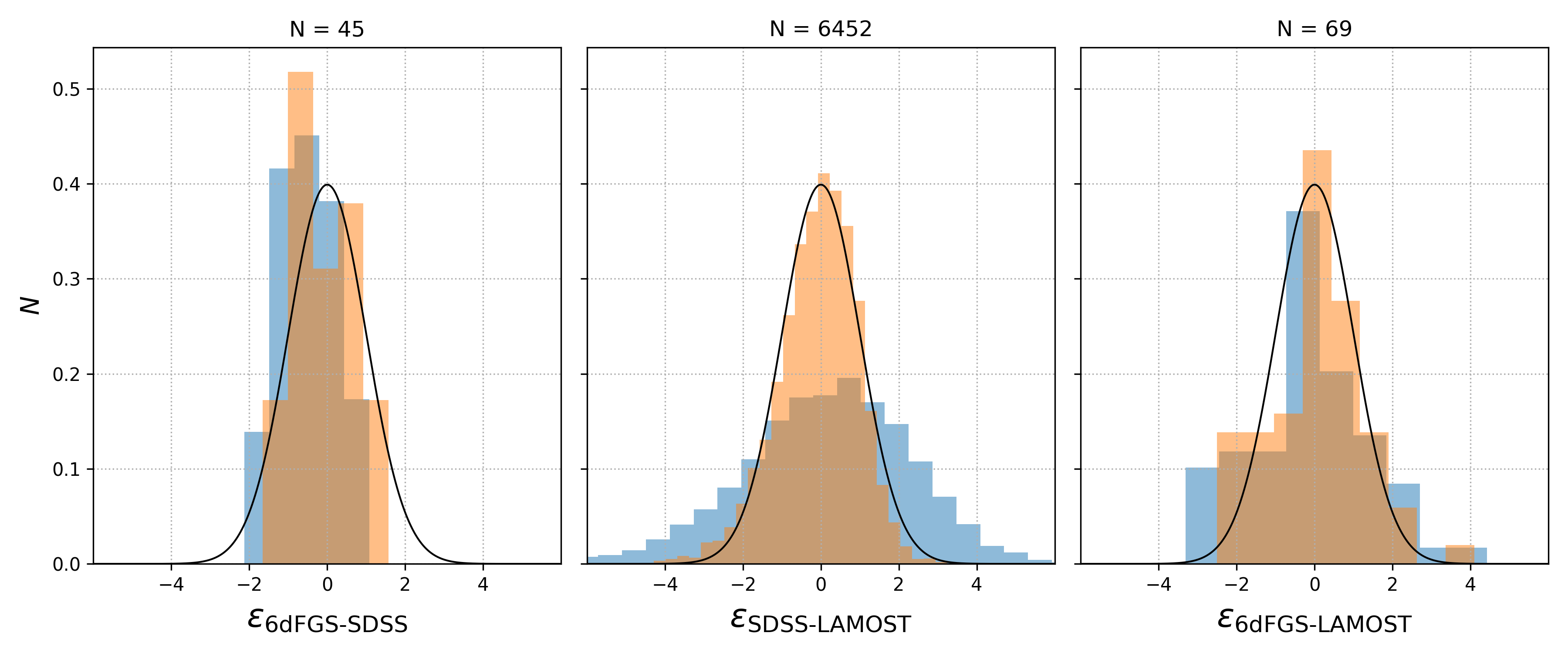
* 1. Untuk SDSS, fetch ra dan dec baru (34059) ✅
  2. Untuk SDSS, dapetin data fotometri 2MASS ✅
  3. Di combine spectro\_photometry, tambahin step buat masukin kolom is\_etg. Yang 6dFGS dicrossmatch sama 8885 tea, yang LAMOST sama seleksi visual John.
  4. Benerin code veldisp calibration 🡪 LAMOST aja yang diinflate? (eksperimen dulu)
  5. Do everything the same, baru seleksi ETG LAMOST.
* Penemuan: ra, dec, sigmastars, sigmastarserr yang ada di tabel Cullan beda dengan yang dari casjobs.
* Setelah diselidiki, yang saya pake itu ra dec dari tabel SpecObjAll, yang mana itu adalah koordinat fiber, bukan objeknya
* Kalo yang sigma stars ga tau kenapa.
* My solution: dari objID setiap galaksi di tabel Cullan, ambil data originalnya langsung dari CasJobs DR14.
* TERNYATA VELDISP GA ADA PERBEDAAN, FALSE ALARM! BISA LANGSUNG PAKE DATA SPEKTROSKOPI DARI TABEL CULLAN. Ra dec tetep beda sedikit sih, tapi kayaknya ga pengaruh banyak
* Setelah dicrossmatch dengan 2MASS, jumlah galaksi SDSS yang tersisa: 31631. Dari sini, setelah dicrossmatch dengan pengukuran GALFIT John, yang tersisa 29677 galaksi (kurang 2000 galaksi yang harusnya ada pengukuran, minta tolong lagi?). Wait.. yang data Cullan ini z nya masih sampe 0.1, tapi setelah histogram redshiftnya dicek, sepertinya tetep butuh..

A blue graph with black text

Description automatically generated

**9 Mar 2024**

* Kalibrasi veldisp selesaiii. Hasilnya:



* Akhirnya yang di-scaling cuma error LAMOST (pake metode pertama), sementara kalo offset 6dFGS dan LAMOST.
* Nilai scaling:
  + klamost (SDSS fiducial) = 2.64
  + klamost (6dFGS fiducial) = 4.21
* Nilai offset:
  + off6df = -0.019
  + offlamost = -0.006
* Cek: apakah total galaksi LAMOST yang diinspeksi John itu ada semua di sampel terbaru
* Catatan selection criteria: buat 6dFGS di-bypass aja. Sebenernya seleksi cz <= 16120 ada 5 galaksi yang terseleksi, max czgroup = 16133 ish. Kemungkinan besar karena beda nilai dipol yang dipake. Sementara seleksi yang lain gak ada yang mengubah jumlah sampel (in short, gak guna).