The Standard errors of total employment in the QLFS

Andrew Kerr, DataFirst

DataFirst Data Quality conference

June 2017

Introduction

- This paper has its origins in a request for comment by Business Day on President Zuma's claim in the Feb 2014 State of the Nation address that in the previous year the economy had created 650 000 jobs.
- Conforming to the academic stereotype I can give them a partial answer 3 years later ☺
- The question made me realise no one in the media reports on or understands the statistical uncertainty in the estimates produced from surveys
    - this made me check the standard errors reported in the Stats SA release documentation.
- The main finding is that the standard errors I calculate are substantially higher than those reported by Stats SA, so that most quarter to quarter changes in employment are not statistically significant, whereas Stats SA reports imply the majority of changes are statistically significant.

QLFS widely used and reported

- National Treasury used the QLFS to estimate the unemployment rate for young people and to make the case for the Employment Tax Incentive.
- The Parliamentary Budget Office used the QLFS as part of its Quarterly Economic Briefs,
- Cosatu referred to the unemployment rate in the QLFS in its statement about the 2017 budget speech
- First National Bank provides a brief online summary of the QLFS results from various quarters
- None of these sources referred to statistical uncertainty of the estimates.

Overview of sampling methods

- The LFS and QLFS are two stage surveys in which households are selected by first selecting a sample of PSUs and then selecting a certain number of dwelling units from each PSU.
- A list of PSUs to be sampled in several surveys is created and is called the master sample.
- In each master sample around 3000 PSUs have been selected, with often 10 dwelling units selected per PSU, and thus a final sample of around 30000 dwelling units.
- In the new 2013 Master Sample, which was used from Q1 2015 onwards in the QLFS, about 3300 PSUs used so the sample size has increased by 10%.

Stratification

- The first Master sample in the early LFSs had 18 strata, corresponding to urban and rural strata for each of South Africa's nine provinces.
- In the second master sample, which was used from the September 2004 LFS, stratification was done for each of South Africa's 53 district councils.
- In the QLFSs a more complex stratification process was undertaken, with the aim of decreasing sampling error- 363 strata until end 2014, 248 since then.

Strata and PSU variable creation

- Stratum variable only released with QLFS in later rerelease (2013?)
- PSU almost never released- this is a problem for any analysis wanting to estimate correct standard errors for any estimate.
- I create a PSU using the first 7 digits of the hh id number- this matches the psu released in Q1 2009 and has the properties we would expect given what we know about the sampling methods.
- I use these plus the weights released by Stats SA to estimate total employment and the coefficient of variation=
- Standard Error$_{totalemployment}$/total employment

LFS Total employment and CVs

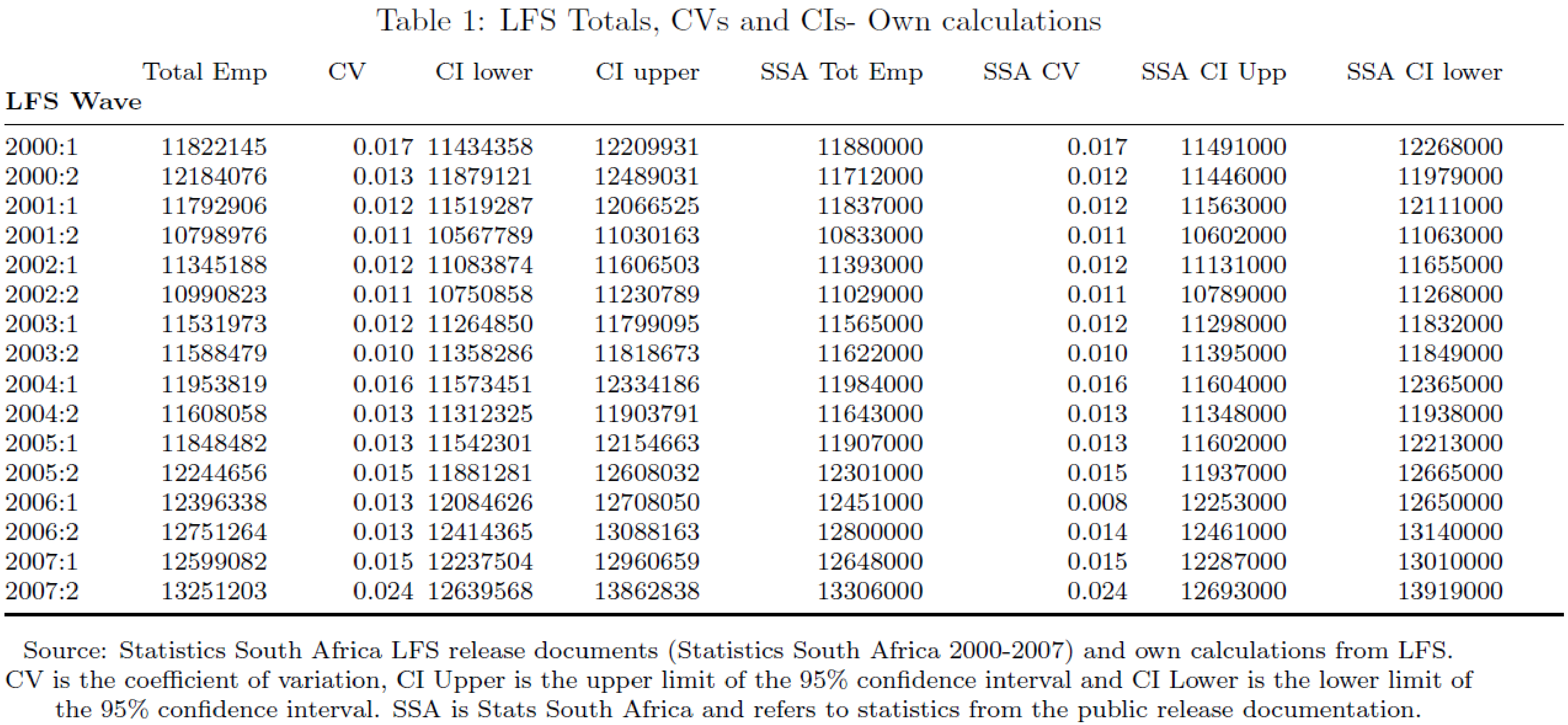

Table 1: LFS Totals, CVs and CIs- Own calculations

LFS Wave	Total Emp	CV	CI lower	CI upper	SSA Tot Emp	SSA CV	SSA CI Upp	SSA CI lower
2000:1	11822145	0.017	11434358	12209931	11880000	0.017	11491000	12268000
2000:2	12184076	0.013	11879121	12489031	11712000	0.012	11446000	11979000
2001:1	11792906	0.012	11519287	12066525	11837000	0.012	11563000	12111000
2001:2	10798976	0.011	10567789	11030163	10833000	0.011	10602000	11063000
2002:1	11345188	0.012	11083874	11606503	11393000	0.012	11131000	11655000
2002:2	10990823	0.011	10750858	11230789	11029000	0.011	10789000	11268000
2003:1	11531973	0.012	11264850	11799095	11565000	0.012	11298000	11832000
2003:2	11588479	0.010	11358286	11818673	11622000	0.010	11395000	11849000
2004:1	11953819	0.016	11573451	12334186	11984000	0.016	11604000	12365000
2004:2	11608058	0.013	11312325	11903791	11643000	0.013	11348000	11938000
2005:1	11848482	0.013	11542301	12154663	11907000	0.013	11602000	12213000
2005:2	12244656	0.015	11881281	12608032	12301000	0.015	11937000	12665000
2006:1	12396338	0.013	12084626	12708050	12451000	0.008	12253000	12650000
2006:2	12751264	0.013	12414365	13088163	12800000	0.014	12461000	13140000
2007:1	12599082	0.015	12237504	12960659	12648000	0.015	12287000	13010000
2007:2	13251203	0.024	12639568	13862838	13306000	0.024	12693000	13919000

Source: Statistics South Africa LFS release documents (Statistics South Africa 2000-2007) and own calculations from LFS. CV is the coefficient of variation, CI Upper is the upper limit of the 95% confidence interval and CI Lower is the lower limit of the 95% confidence interval. SSA is Stats South Africa and refers to statistics from the public release documentation.

Table 2: QLFS Totals, CVs and CIs- Own calculations

	Total Emp	CV	CI lower	CI upper	SSA Tot Emp	SSA CV	SSA CI Upper	SSA CI lower
QLFS Wave								
2008:1	13727018	0.013	13383166	14070871	13623000	0.006	13462794	13783206
2008:2	13862039	0.013	13497286	14226793	13729000	0.006	13567547	13890453
2008:3	13798507	0.013	13440831	14156182	13655000	0.006	13494417	13815583
2008:4	14013371	0.013	13655802	14370941	13844000	0.006	13681195	14006805
2009:1	13827225	0.013	13471446	14183005	13636000	0.006	13475641	13796359
2009:2	13583269	0.014	13220805	13945732	13369000	0.006	13211781	13526219
2009:3	13123545	0.014	12761876	13485214	12885000	0.007	12708218	13061782
2009:4	13243208	0.014	12880011	13606405	12974000	0.006	12821426	13126574
2010:1	13060112	0.014	12711421	13408803	12803000	0.031	12025090	13580910
2010:2	13050562	0.014	12695393	13405731	12742000	0.007	12567180	12916820
2010:3	12958495	0.015	12579774	13337216	12975000	0.007	12796983	13153017
2010:4	13119402	0.015	12745609	13493194	13132000	0.007	12951829	13312171
2011:1	13102691	0.015	12714276	13491107	13118000	0.007	12938021	13297979
2011:2	13113575	0.015	12730596	13496554	13125000	0.007	12944925	13305075
2011:3	13305855	0.015	12925151	13686559	13318000	0.007	13135277	13500723
2011:4	13488584	0.014	13122594	13854575	13497000	0.007	13311821	13682179
2012:1	13393459	0.015	13005818	13781101	13422000	0.007	13237850	13606150
2012:2	13426581	0.014	13052173	13800989	13447000	0.007	13262507	13631493
2012:3	13621412	0.015	13232152	14010672	13645000	0.007	13457791	13832209
2012:4	13549756	0.015	13158001	13941510	13577000	0.007	13390724	13763276
2013:1	13597662	0.015	13202711	13992613	13621000	0.007	13434120	13807880
2013:2	13693133	0.014	13304924	14081341	13721000	0.007	13532748	13909252
2013:3	14003099	0.015	13596692	14409505	14029000	0.007	13836522	14221478

Source: Statistics South Africa QLFS release documents (Statistics South Africa 2008-2014) and own calculations from QLFS. CV is the coefficient of variation, CI Upper is the upper limit of the 95% confidence interval and CI Lower is the lower limit of the 95% confidence interval. SSA is Stats South Africa. There was 1 stratum in Q1 2011 with only 1 PSU- meaning standard errors could not be calculated. I excluded this stratum, which contained 6 employed individuals, in the calculations for this table.

Table 3: QLFS Totals, CVs and CIs- Own calculations using Revised data

	Total Emp	CV	CI lower	CI upper	SSA Tot Emp	SSA CV	SSA CI Upp	SSA CI lower
QLFS Wave								
2012:4	14523850	0.015	14102546	14945155	14524000	.	.	.
2013:1	14558375	0.015	14140380	14976370	14558000	.	.	.
2013:2	14691538	0.014	14285551	15097525	14692000	.	.	.
2013:3	15035843	0.015	14605838	15465848	15036000	0.007	14829706	15242294
2013:4	15176755	0.015	14735016	15618494	15177000	0.007	14968772	15385228
2014:1	15054791	0.015	14623231	15486351	15055000	0.007	14848445	15261555
2014:2	15094243	0.014	14671031	15517455	15094000	0.007	14886910	15301090
2014:3	15116569	0.014	14694186	15538952	15117000	0.007	14909595	15324405
2014:4	15287197	0.014	14857579	15716815	15320000	0.007	15109810	15530190
2015:1	15459420	0.011	15138547	15780292	15459000	0.007	15246903	15671097
2015:2	15657003	0.010	15337933	15976072	15657000	0.007	15442186	15871814
2015:3	15828439	0.010	15505092	16151786	15828000	0.006	15641863	16014137
2015:4	16018068	0.010	15693858	16342278	16018000	0.006	15829628	16206372
2016:1	15674513	0.011	15347915	16001112	15675000	0.006	15490662	15859338
2016:2	15545447	0.011	15211338	15879557	15575000	0.006	15391838	15758162
2016:3	15833195	0.011	15495444	16170946	15833000	0.006	15646804	16019196
2016:4	16068612	0.011	15720864	16416360	16018000	0.006	15829628	16206372

Source: Statistics South Africa QLFS release documents (Statistics South Africa 2008-2014) and own calculations from QLFS. CV is the coefficient of variation, CI Upper is the upper limit of the 95% confidence interval and CI Lower is the lower limit of the 95% confidence interval. SSA is Stats South Africa. In Q4 2014 there were two strata with a single PSU and standard errors could thus be computed. I excluded the 9 employed individuals in these two strata from the table.

Summary so far

- LFS CVs in the public release very similar to the ones I calculate.
- QLFS CVs I calculate are about double those in the public release.
- This is despite both sets of surveys having very similar structures and complex sampling procedures, which I incorporate in both cases.
- No change in Stats SA CVs with introduction of new sample in 2015- but mine do change (decreasing by about 20%).
- Now I move to looking at changes in employment between quarters and years.

Table 4: QLFS Quarter to Quarter Employment Changes

	Own Calculations					Stats SA release documentation				
	Emp Δ	CV	CI lower	CI upper	p value	Emp	CV	CI lower	CI upper	p value
QLFS Wave		Emp Δ	Emp Δ	Emp Δ		Δ	Emp Δ	Emp Δ	Emp Δ	
2008:2	129004	0.97	−117338	375346	0.30	106000	.	.	.	.
2008:3	−67905	-1.84	−312436	176625	0.59	−74000	.	.	.	0.35
2008:4	215578	0.53	−6761	437917	0.06	189000	.	.	.	0.01
2009:1	−183627	-0.85	−490656	123402	0.24	−208000	-0.36	−353000	−63000	0.01
2009:2	−247569	-0.74	−606444	111307	0.18	−267000	-0.25	−400000	−134000	0.00
2009:3	−462015	-0.41	−833991	−90038	0.01	−484000	-0.18	−652000	−318000	0.00
2009:4	116796	1.75	−284935	518526	0.57	89000	1.05	−95000	273000	0.34
2010:1	−173891	-1.00	−516039	168257	0.32	−171000	-2.33	−954000	611000	0.67
2010:2	−15647	-10.24	−329702	298408	0.92	−61000	-1.21	−206000	84000	0.41
2010:3	−86098	-2.11	−441739	269543	0.64	233000	-1.48	−337000	164000	0.50
2010:4	157632	0.97	−143297	458561	0.30	157000	0.77	−82000	398000	0.20
2011:1	−13820	-12.31	−347357	319718	0.94	−14000	-5.50	−165000	137000	0.86
2011:2	7053	26.43	−358309	372416	0.97	7000	20.68	−257000	270000	0.96
2011:3	192939	1.06	−209483	595360	0.35	193000	0.45	23000	363000	0.03
2011:4	178985	1.00	−173525	531494	0.32	179000	0.47	15000	343000	0.03
2012:1	−75568	-2.07	−382122	230986	0.63	−75000	-1.03	−228000	77000	0.33
2012:2	24867	6.32	−283383	333117	0.87	25000	3.28	−133000	182000	0.76
2012:3	198715	0.82	−119932	517362	0.22	198000	0.39	46000	352000	0.01
2012:4	−68032	-2.84	−447389	311324	0.73	−68000	-1.16	−222000	86000	0.39
2013:1	43719	4.06	−303967	391405	0.81	44000	2.01	−129000	217000	0.62
2013:2	99930	1.76	−244321	444180	0.57	100000	0.71	−40000	240000	0.16
2013:3	307654	0.60	−56480	671789	0.10	308000	0.28	136000	480000	0.00

Source: Statistics South Africa QLFS release documents (Statistics South Africa 2008-2014) and own calculations from QLFS.

otes: Changes in Employment not given in the first year of the QLFSs. CV is the coefficient of variation, CI upper is the upper limit of the 95% confidence interval and CI lower is the lower limit of the 95% confidence interval. SSA is Stats South Africa. p val is the p value for the quarter to quarter employment change.

Table 5: QLFS Quarter to Quarter Employment Changes using Revised data

				Own Calculations			**Stats SA release documentation**			
	Emp Δ	CV	CI lower	CI upper	p value	Emp	CV	CI lower	CI upper	p value
QLFS Wave		Emp Δ	Emp Δ	Emp Δ		Δ	Emp Δ	Emp Δ	Emp Δ	
2013:1	34525	8.77	−558955	628004	0.91	.	.	.	.	.
2013:2	133163	2.23	−449542	715868	0.65	.	.	.	.	.
2013:3	344305	0.88	−247075	935684	0.25	.	.	.	.	.
2013:4	140912	2.23	−475560	757383	0.65	−13500	−0.52	−27000	309	0.10
2014:1	−121963	-2.58	−739521	495594	0.70	−146000	−0.71	−292000	48000	0.16
2014:2	39452	7.82	−564992	643896	0.90	−66000	2.19	−132000	212000	0.65
2014:3	22326	13.66	−575601	620252	0.94	−73500	3.84	−147000	192000	0.80
2014:4	170628	1.80	−431849	773106	0.58	−37500	0.70	−75000	481000	0.15
2015:1	172223	1.59	−363996	708442	0.53	.	.	.	.	.
2015:2	197583	1.17	−254926	650092	0.39	10000	0.46	20000	375000	0.03
2015:3	171436	1.35	−282830	625703	0.46	−1500	0.52	−3000	346000	0.05
2015:4	189629	1.23	−268263	647521	0.42	14500	0.43	29000	350000	0.02
2016:1	−343555	-0.68	−803749	116639	0.14	−259000	−0.24	−518000	−191000	0.00
2016:2	−129066	-1.85	−596287	338155	0.59	−151500	−0.69	−303000	45000	0.15
2016:3	287748	0.84	−187336	762832	0.24	55500	0.49	111000	4650000	0.00
2016:4	235417	1.05	−249355	720190	0.34	35000	0.36	70000	401000	0.01

Source: Statistics South Africa QLFS release documents (Statistics South Africa 2008-2014) and own calculations from QLFS. CV is the coefficient of variation, CI upper is the upper limit of the 95% confidence interval and CI lower is the lower limit of the 95% confidence interval. SSA is Stats South Africa. p value is the p value for the quarter to quarter employment change. In Q4 2014 there were two strata with a single PSU and standard errors could thus be computed. I excluded the 9 employed individuals in these two strata from the table.

Notes: Changes in Employment were not provided in the Stats SA release document in the first year of the revised QLFSs and also in Q1 2015.

Summary of quarter to quarter results

- In the older data to 2013 I find only 1 of 21 quarter to quarter changes statistically significant at 5%, whereas Stats SA finds 8 statistically significant.
- In the newer revised data from 2013 I find none of the 16 quarter to quarter changes statistically significant, whilst Stats SA finds 6 of 12 quarter to quarter changes that they estimated statistically significant.
- The quarterly changes are regularly reported by the media and the lack of statistical significance is never mentioned by media (or Stats SA?) and is not (publicly?) corrected by Stats SA.

Table 6: QLFS Year to Year Employment Changes

	Emp Δ	CV(Emp Δ)	CI lower Emp Δ	CI upper Emp Δ	p value
QLFS Wave					
2009:1	93050	2.19	−306502	492601	0.65
2009:2	−283523	−0.89	−775993	208946	0.26
2009:3	−677633	−0.43	−1248797	−106469	0.02
2009:4	−776415	−0.43	−1427445	−125385	0.02
2010:1	−766679	−0.42	−1394344	−139014	0.02
2010:2	−534757	−0.56	−1124027	54513	0.08
2010:3	−158840	−1.86	−737798	420118	0.59
2010:4	−118004	−2.31	−652173	416166	0.67
2011:1	42068	6.60	−502107	586242	0.88
2011:2	64768	4.41	−495450	624986	0.82
2011:3	343804	0.85	−226627	914236	0.24
2011:4	365157	0.82	−223811	954125	0.22
2012:1	303409	1.01	−294952	901770	0.32
2012:2	321222	0.90	−242957	885402	0.26
2012:3	326999	0.85	−219533	873530	0.24
2012:4	79982	3.54	−475223	635186	0.78
2013:1	199268	1.46	−370092	768628	0.49
2013:2	274331	1.08	−305055	853718	0.35
2013:3	383271	0.81	−222576	989118	0.22

Source: own calculations from QLFS. CV is the coefficient of variation, CI upper is the upper limit of the 95% confidence interval and CI lower is the lower limit of the 95% confidence interval. p val is the p value for the year to year employment change.

Table 7: QLFS Year to Year Employment Changes using revised data

	Emp Δ	CV(Emp Δ)	CI lower Emp Δ	CI upper Emp Δ	p value
QLFS Wave					
2013:4	652904	0.477	42469	1263339	0.036
2014:1	496416	0.617	−104386	1097219	0.105
2014:2	402705	0.743	−183754	989163	0.178
2014:3	80725	3.810	−522028	683479	0.793
2014:4	110442	2.847	−505760	726644	0.725
2015:1	404628	0.678	−133148	942405	0.140
2015:2	562760	0.481	32747	1092772	0.037
2015:3	711871	0.381	179930	1243811	0.009
2015:4	730871	0.376	192649	1269094	0.008
2016:1	215094	1.086	−242756	672943	0.357
2016:2	−111555	−2.113	−573545	350434	0.636
2016:3	4756	50.162	−462822	472334	0.984
2016:4	50544	4.799	−424894	525981	0.835

Source: own calculations from QLFS. CV is the coefficient of variation, CI upper is the upper limit of the 95% confidence interval and CI lower is the lower limit of the 95% confidence interval. p value is the p value for the year to year employment change. In Q4 2014 there were two strata with a single PSU and standard errors could thus be computed. I excluded the 9 employed individuals in these two strata from the table.

Summary of year to year changes

- The changes the large decreases around the financial crisis were statistically significant.
- As was the year to year change that President Zuma reported in his state of the nation address- but not at the 1% level!
- And the large increases in employment in the change over from the master sample (which were not sustained into 2016 once the year to year changes were estimated on samples both from the new master sample).

QLFS QES comparisons

- QES has in the SAMPLE firms that account for 45-55% of total formal non agricultural employment (see Kerr et al 2014).
- We might thus expect that the CVs should be lower than the QLFS- which only has 30000 households!
- But actually the QES CV is higher than the CV in the QLFS release in every quarter, although lower than the one I calculate using the QLFS and that I reported above.

Explanations for differences between Stats SA release and my estimates

- PSU incorrect since not actually released with the data?
  - Unlikely since the PSU I construct matches that released in Q1 2009 when a PSU variable was released publicly.
- Differences in variance estimation methods matter?
- But Heeringa et al. (2010) note that the three main methods of variance estimation are Taylor series linearisation, Jack Knife repeated replication and balanced repeated replication and that these three methods "are unbiased and produce identical results in the special case where the estimator of interest is a linear statistic such as a weighted sample total."

Explanations continued

- Stats SA is collapsing PSUs together, so is incorrectly specifying the complex sample design by making PSUs larger (and more heterogeneous?!) than they are in reality-
- "The Fays BRR method [of variance estimation] on the other hand is applicable when two primary sampling units (PSUs) are sampled from each stratum. Therefore we decided to use Fays BRR method by collapsing PSUs into two groups of PSUs within each stratum"
- To check this we would need to know how Stats SA collapsed the PSUs in their variance calculations.
- The main issue though is that most methods are simple and easy to do and should give the same answer.
- What happens when there is only 1 PSU in a stratum, which happens when no hh in some PSU respond? Does Stats SA further collapse strata?

More explanations

- The rotating panel nature of the QLFS lowers the standard errors of changes over waves and that I am calculating these assuming that the data are a set of independent cross sections.
- The calibration undertaken by Stats SA in adjusting some of the marginal totals to match their best estimates of the population reduces the uncertainty in their estimates- not possible to explore with the publicly available data.

Conclusions

- I can replicate LFS variance estimates but not those in the QLFS.
- In the QLFSs I find that the Stats SA estimates of the coefficients of variation of total employment is around half those I estimate until Q1 2015 and around 70% of my estimates after that.
- A number of possible reasons are possible for these differences, most likely being an incorrect collapsing of PSUs by Stats SA or the use of calibration that I cannot do with the publicly available data.
- I estimate quarter to quarter changes in employment are statistically significant only in 1 out of 34 quarters, whereas the Stats SA documentation suggests that the quarter to quarter changes are statistically significant in around 41% of quarters (14 out of 34).
- This matters because these changes are widely reported and discussed, but actually are just noise if my estimates are correct.
- Having a less frequent survey with a larger sample might be more appropriate, something suggested by Simkins (2004)- new master sample moves a small way in this direction.
- I suggest making data and documentation that allow researchers to replicate the variance estimation results publicly available.