Claim Size Distributions in German Private Health Insurance

Jens Piontkowski

June 23, 2011

Abstract

We examine the yearly claim size distribution in dependence of age group and sex. Named distributions are fitted to the data. Special attention is given to the heaviness of the tails due to its importance for the assessment of risks inherent in health insurance. Finally, we caution against using the normal approximation without an error estimate because simulations show a bad fit of the Gaussian distribution to claim distributions of groups thousand or even ten thousand strong.

Key words. claim size distribution, distribution fitting, health insurance, normal approximation, tail index.

German private health insurance is health insurance with guaranteed renewal. A contract of an adult is intended to be lifelong and the premiums are calculated in such a way that they remain constant for life if no medical inflation occurs. The main basis of premium calculation are the expected claims in the entire future of the insured. These together with the expected lapse and mortality rates yield an expected cash flow. The net premium is the present value of this cash flow converted into a lifelong constant annual premium.

By German law [KalV, §6] the estimation of the expected claims must be based on previous experience taking into account the sex and the age of the insured. (In order to ensure gender equality costs for pregnancy and birth are distributed over both sexes, but we will ignore this for the purpose of this article.) In practice, the insured persons are split into cohorts depending on the sex and the age group, where the groups cover five consecutive ages, e.g., 20–24, 25–29, 30–34, etc. The mean value of the claims in a cohort is taken as a raw estimate for the expected claim. For premium calculation these raw estimates are smoothed over the ages for each of the sexes.

Already for a moderate number of observations the mean is in general a good estimator for the expected value. Due to this and a security margin of at least 5% in the premium calculation prescribed by law [KalV, §7], there was no pressing need to investigate the claim size distribution for the purpose of premium calculation. However, there are several areas where the knowledge of claim size distributions is useful:

- For pricing high–excess loss layers in reinsurance the tail of the distribution must be known.

- Recognition of an atypical number of high claims in a cohort which may have a large impact on the mean and hence impair the premium calculation. In practice, this is dealt with in an ad hoc manner.
- For the calculation of risk adjusted premiums more information about the claim size distribution besides the expected value is needed. In general, it is assumed that the mean of a cohort is normally distributed. This assumption must be justified — in particular, if there is only a moderate number of observations and the claim size distribution is highly skewed, as it is the case for medical insurance covering specifically hospital expenses.
- In recent years German health insurance companies are evaluated by analysts based on a stochastic simulation, called market consistent embedded value. So far only the income from capital investments is modeled stochastically, while the claims are modeled deterministically by their expected value. An understanding of claim size distributions enables an improved modeling of the liabilities.
- Academic interest in distributions which occur in practice.

In the first section we will discuss the data and describe the portion of the claim–free insured. In the second section we examine the conditional distribution of the non–zero claim sizes. We will show that after normalization these distributions are similar over the age groups. Then we describe these empirical distributions by named theoretical distributions. Finally, we study their tail. In the third and final section we warn against using the central limit theorem without an error estimation by giving unexpected examples.

The author is indebted to Deutsche Krankenversicherung AG (DKV) for providing the data. Further, the author thanks Rasmus Schlömer for several discussions.

1 The data set and general considerations

The examination is based on the claims in the year 2005 of a modern full cover insurance as well as a classical inpatient, outpatient, and dental insurance which are usually sold together. The data are grouped by the cohorts, i.e., by sex and age groups spanning five years each. In order to ensure that each group contains at least about one thousand members (more precisely at least 900), we consider in the case of the full cover insurance only the ages from 25 to 54 for both sexes, in the case of inpatient insurance the ages from 25 to 64 for men and from 30 to 59 for women, in the remaining cases the ages from 25 to 84 for both sexes. In full cover insurance the strongest age group is 35–39 with more than 4500 men and 2500 women. In the classical inpatient, outpatient, and dental insurance the strongest age groups are 40–44 with more than 9500, 27000 resp. 30000 men and 3500, 6500 resp. 8500 women. The fewer members in inpatient insurance are due to the fact that the insurance company offers several popular inpatient insurances, and we restrict our examination only to the most popular one.

In practice the members of a cohort are considered as a homogeneous risk group. As mentioned in the introduction, it is required by the German law that

the cohorts should not be split any further for the purpose of claim estimation in premium calculation. We follow this practice in the article. However, there are two well–known reasons why a cohort is not a homogeneous risk group:

- Before the signing of a health insurance contract the insurance company requires a health risk assessment. High risks are rejected, medium risks are charged an additional premium. The risk assessment leads to lower than average claims during the first three to seven years of a contract.

- The insurance company offers a premium refund to the insured if neither outpatient or dental treatments are reimbursed during a year. Thus, insureds with small expenses will appear claim–free because they will prefer the refund to the reimbursement. The premium refund increases if the insured is claim–free for several years in a row. Hence, there are several different thresholds at which the insured will ask for reimbursement.

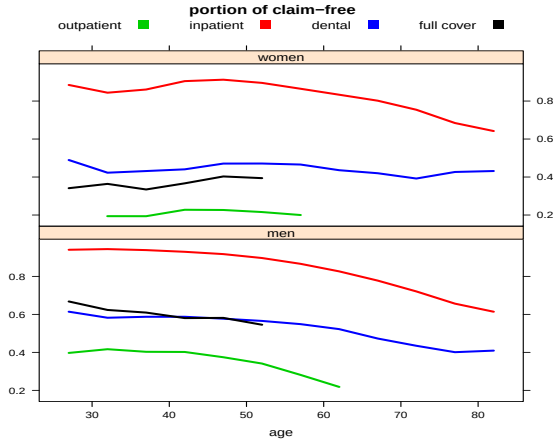
Resulting from this we cannot expect that the cohorts are really homogeneous. However, one might hope that a cohort has a stable mixture of risks after several years. Finally, we want to mention that the dental insurance has a 25 percentage deductible for dental prosthesis and the full cover insurance has a small fixed deductible. The latter increases the loss of information about the possible small claims which already occurs because of the offered premium refund for being claim–free.

How do we model the distribution of the claim sizes? Because of the claim–free insureds the distribution will have a large mass at zero. So it is natural to split the description of the claim size distribution into two parts: a distribution describing the occurrence of zero claims and the conditional distribution of non–zero claims. While examining the conditional distribution we will pay special attention to the high claims, i.e., to the tail of the distribution, because of its importance for the estimation of high quantiles, risk measures, pricing of reinsurance etc.

In non–life insurance the small and medium size claims are modeled for each cohort separately, while the extremely high claims are modeled by a Pareto distribution neglecting the grouping into cohorts. Here, this does not seem to be necessary as the distributions that we will fit to the claim size distributions in the cohorts possess sufficiently heavy tails to cover the extremely high claims appropriately. In addition, while in non–life insurance the extremely high claims are due to random accidents, in health insurance a large portion of the extremely high claims can be foreseen on the basis of the insureds' health states. Thus the claims are only partly random, and one should deal with the excepted high claims on a case by case basis.

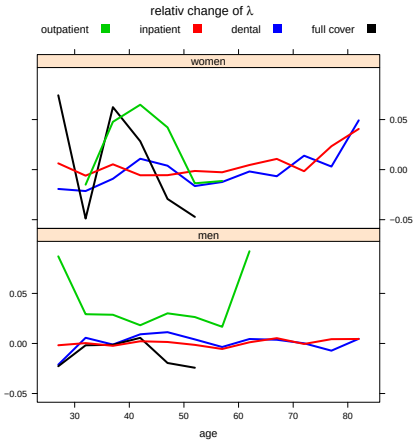
Let us start our examination with the description of the number of claim–free insureds. The graph below shows the dependence of the percentage of the claim–free insureds on insurance type, sex, and age based on data of the year

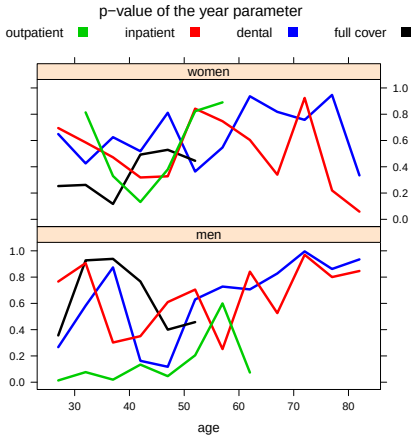
2005.



Naturally, the highest percentage of claim–free insured are in inpatient insurance. For men the number of claim–free decrease with age, for women this appears to be only true in inpatient insurance — the other types of insurances are roughly constant. The obvious model for the number of claim–free is a Binomial distribution $\mathrm{B}(n, \lambda)$ whose parameter λ can be read off the above graph.

To validate this model we compare the data of 2005 to the data of 2004 in the next two graphs. The first shows the relative change of λ in the year 2004 compared to the year 2005. For the second we fit a logistic regression model to each cohort with the observation year as the only independent variable and plot the p–value of the hypothesis test whether the observation year is relevant.





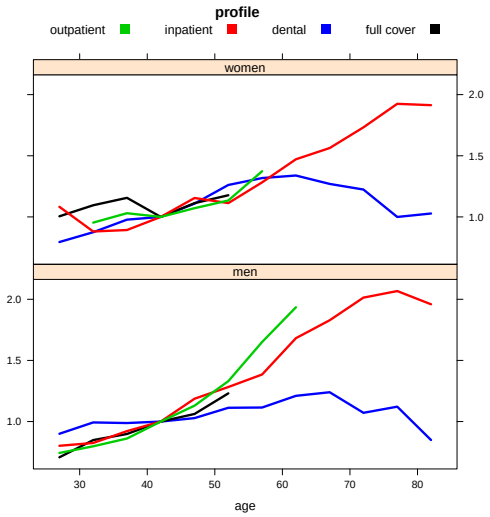
The small probabilities for the moderate relative changes are due to the high number of observations, which should have produced a very reliable estimate for the parameter λ. In general, the influence of the observation year does not seem to be significant — with the notable exception in outpatient insurance in the younger ages. This may be due to an epidemic of a minor illness. Here, the Binomial model should be used with caution.

2 The conditional distributions

Having dealt with the portion of claim–free insureds in the previous section we turn to the conditional distribution of non–zero claim sizes. Unfortunately, because of the large proportion of claim–free in inpatient insurance the data basis for our examination is greatly reduced. In the age group of 20–24 we have only about 100 observations for women and 200 for men, in the other age groups we have between 450 and 2100. For the other insurance types we have at least 950 observations with the exception of dental insurance men 80–84, women 25–29, 75–84 and full cover men 25–29, 50–54, women 25–29, 45–54, which still have at least 500 observations.

2.1 Comparison over the age groups

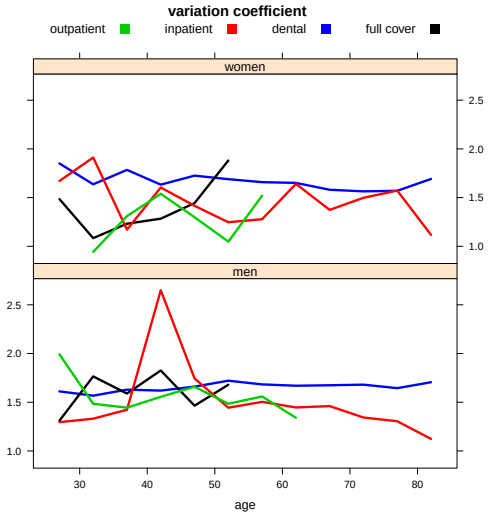
We start by plotting the means of the cohorts relative to the cohorts of age 40–44. This is called the profile:



We note that none of the graphs has a big jump and the development with increasing age has a clear trend. With the exception of dental insurance the profiles increase with age. For dental insurance the graphs are slightly concave with maximum at the age 62–67. We can observe this smooth behavior of the profiles despite the fact that we included the extremely high claims, which are usually ignored for premium calculation.

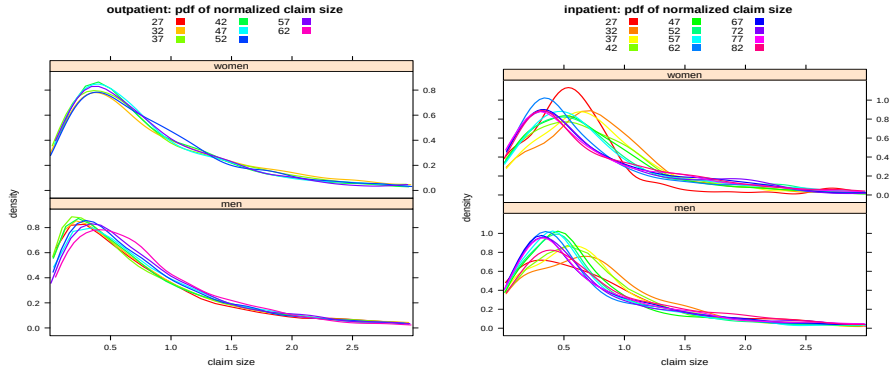
Next we plot the variance coefficient for the cohorts, i.e., the quotient of the

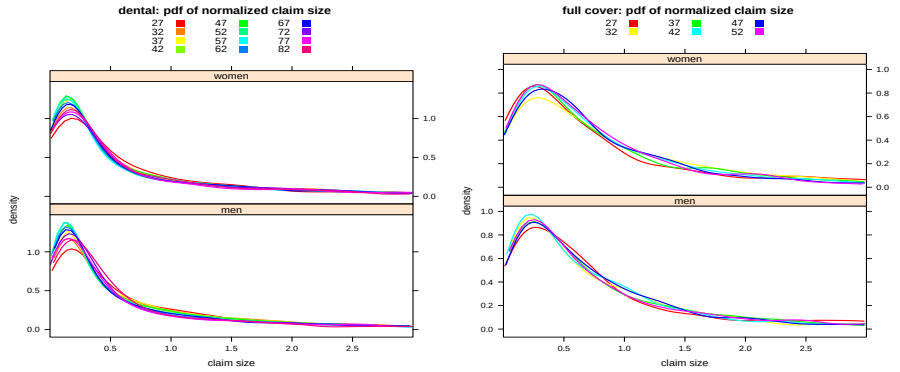
standard deviation by the mean:



The graphs are jumpy, but the overall tendency is that the variation coefficient is mainly independent of the age. The huge jump in the cohort of 40–44 year old men in inpatient insurance is due to two extremely high claims of 83 and 59 times the mean of this cohort, while for the other ages the maximum is nearly always below 25. The more prominent volatility of the graphs for women results from the fact that there are fewer observations for women and hence we have a larger error in the estimation of the standard deviation.

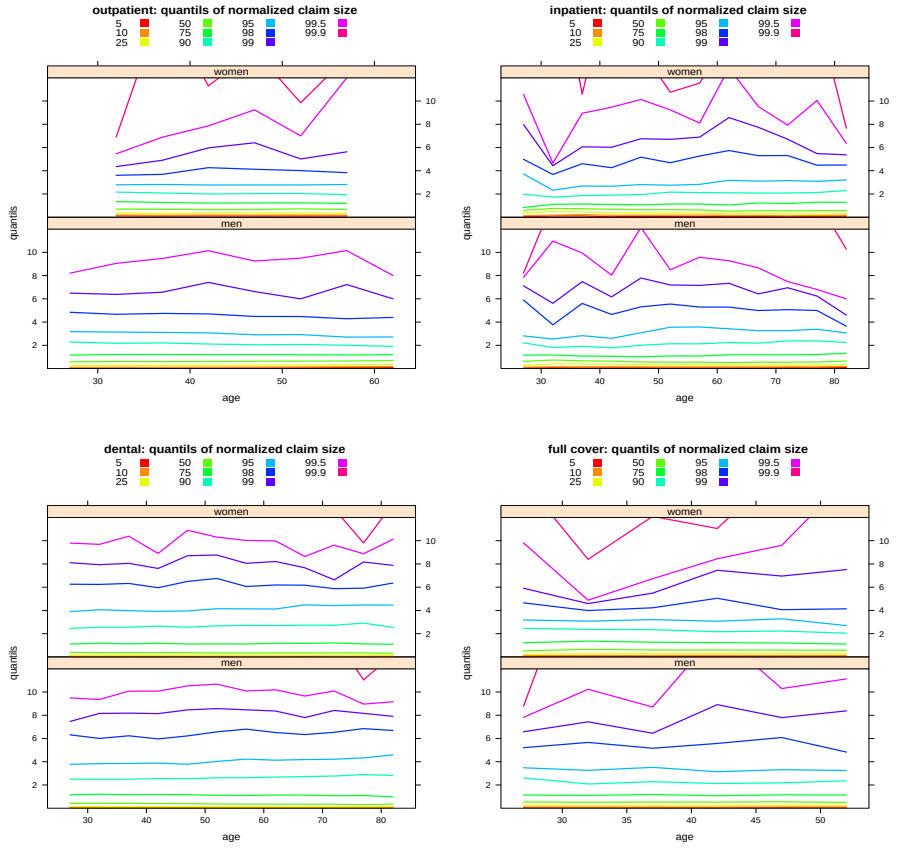
For all the following examinations we normalize the conditional distributions in the cohorts by their means. Thus we will have a mean of 1 in each cohort and our observation about the variation coefficients means that the standard deviation depends on the insurance type and the sex, but is roughly independent of the age group. Below we plot the pdfs of the cohorts using the **R** defaults. A reduction of the default bandwidth in the kernel estimator leads to small bumps in the graphs, thus the default value of **R** seems optimal.





To our surprise we have very similar pdfs over the age groups with the exception of inpatient insurance. Looking back at the huge change in the portion of not claim–free polices over the ages in inpatient insurance compared to the other insurance types a different behavior of inpatient insurance is unsurprising.

To give further evidence to the fact that the pdfs of the normalized claims are roughly independent of the age — even at the right tails — we plot some empirical quantiles of the distributions.

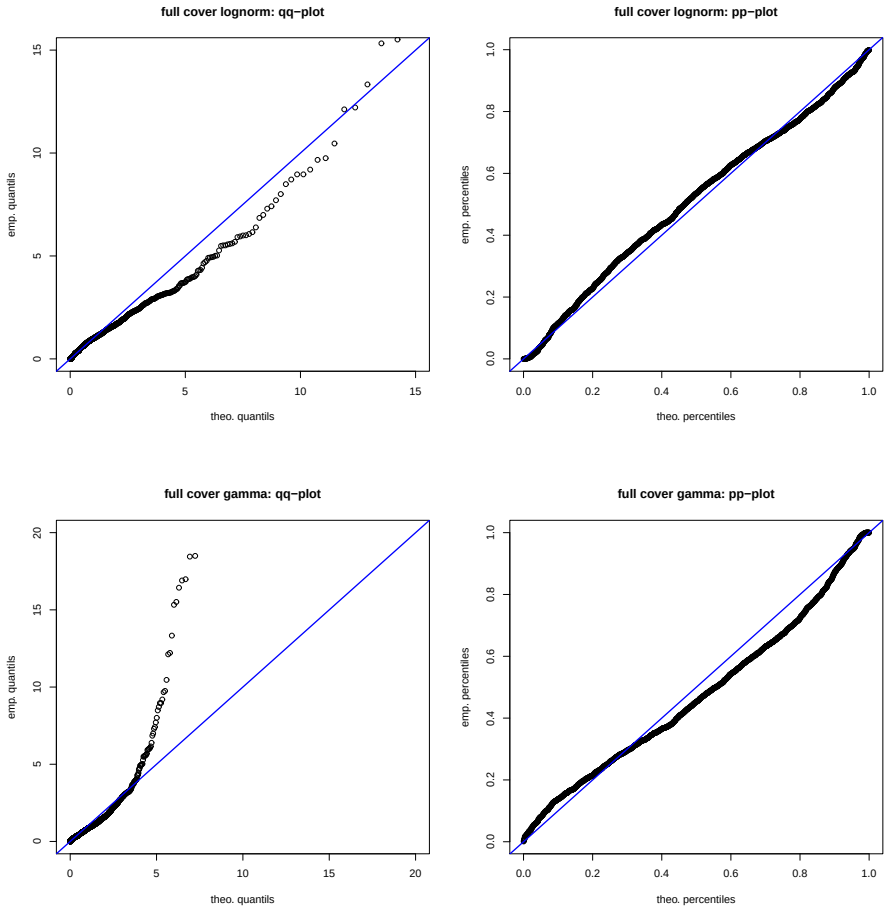


Again we note the quantiles are nearly constant — with the exception of the very high quantiles above 99%. The latter is caused by the fact that in

that range are at most 50, sometimes less than 10 observations. With these few observations volatility is to be excepted. To compute these quantiles one should use the advanced methods described in Subsection 2.3.

2.2 Description through a probability distribution function

Classically claim size distributions are modeled by a lognormal or gamma distribution in combination with a Pareto distribution for large claims. Let us fit a lognormal and gamma distribution with ML–estimation to the claim size distribution of 40–44 year old men in full cover insurance. The resulting qq– and pp–plots are:



Obviously, the gamma distribution fits very badly. This is also the case for other cohorts and insurance types, and we want to discourage its use in health insurance. The fit of the lognormal distribution is not good either, there are cohorts where the fit is worse. We see that the tail of the lognormal distribution is not thick enough in the range from 2.5 to 12.5. In practice, one models additional high claims with a Pareto distribution to compensate for this gap.

Instead of trying to fix the bad fit of the lognormal distribution by adding a Pareto distribution, we attempt to fit the following distributions to the claim size distribution in hope of getting a better fit: scaled Burr Type XII, generalized extreme value, gamma, generalized beta, generalized Birnbaum–Saunders with Laplace, logistic, normal, and Student–t kernel (see Appendix B), inverse Gauss type with Laplace, logistic, normal, and Student–t kernel (see Appendix A), lognormal, Pareto, transformed beta and transformed gamma.

As a measure for the quality of the fitting we looked at the histograms, qq–plots, pp–plots, and plot of the empirical cdfs together with the fitted cdfs. However, we can present here only the results of the Kolmogorov–Smirnov test. To reduce the amount of data we split the data by insurance type and sex and give only the mean and quantiles of the statistics and p–values over the age groups. The best fitting distributions as well as the lognormal and gamma distribution can be found in the following table:

type	distribution	sex	d–statistic						p–value					
			mean	min	$q_{.25}$	$q_{.5}$	$q_{.75}$	max	mean	min	$q_{.25}$	$q_{.5}$	$q_{.75}$	max
outpatient	Burr	M	.014	.007	.010	.013	.016	.022	.391	.087	.206	.286	.607	.895
outpatient	Burr	W	.017	.012	.014	.015	.017	.027	.704	.236	.677	.768	.835	.947
outpatient	Burr adj.	M	.013	.007	.009	.013	.016	.020	.449	.067	.334	.401	.611	.859
outpatient	Burr adj.	W	.017	.013	.015	.016	.016	.026	.660	.262	.578	.654	.814	.967
outpatient	gamma	M	.057	.047	.052	.056	.060	.066	0	0	0	0	0	0
outpatient	gamma	W	.054	.038	.049	.054	.061	.068	.005	0	0	0	.001	.027
outpatient	IGT–St	M	.011	.007	.009	.010	.011	.018	.647	.19	.589	.662	.756	.967
outpatient	IGT–St	W	.016	.012	.014	.016	.017	.019	.708	.434	.586	.665	.883	.971
outpatient	IGT–St adj.	M	.017	.010	.012	.015	.023	.025	.326	0	.014	.203	.618	.911
outpatient	IGT–St adj.	W	.026	.016	.019	.022	.027	.047	.348	.003	.106	.400	.454	.805
outpatient	lognorm	M	.058	.049	.054	.055	.062	.070	0	0	0	0	0	0
outpatient	lognorm	W	.049	.037	.047	.050	.054	.057	.006	0	0	.001	.002	.031
outpatient	trbeta	M	.010	.005	.007	.010	.010	.019	.757	.583	.650	.743	.869	.968
outpatient	trbeta	W	.016	.013	.014	.014	.016	.024	.719	.359	.657	.706	.887	.955
inpatient	Burr	M	.049	.034	.040	.049	.054	.062	.094	0	0	.003	.034	.908
inpatient	Burr	W	.059	.035	.047	.052	.063	.118	.097	.003	.016	.069	.111	.332
inpatient	Burr adj.	M	.048	.036	.041	.049	.053	.059	.092	0	0	.003	.037	.891
inpatient	Burr adj.	W	.058	.035	.046	.052	.060	.122	.104	.005	.017	.069	.124	.353
inpatient	EVT	M	.037	.023	.029	.036	.043	.064	.168	.001	.016	.068	.288	.527
inpatient	EVT	W	.046	.028	.032	.041	.048	.110	.315	.063	.092	.229	.477	.799
inpatient	gamma	M	.087	.057	.070	.092	.098	.116	.024	0	0	0	.001	.285
inpatient	gamma	W	.096	.066	.073	.091	.104	.198	.001	0	0	0	.001	.003
inpatient	GBS–St	M	.034	.019	.024	.029	.042	.056	.248	.008	.043	.182	.432	.648
inpatient	GBS–St	W	.039	.029	.035	.038	.042	.050	.382	.161	.200	.294	.537	.953
inpatient	IGT–St	M	.037	.023	.035	.038	.044	.045	.191	0	.002	.063	.12	.935
inpatient	IGT–St	W	.042	.028	.035	.038	.044	.080	.361	.027	.168	.406	.511	.862
inpatient	IGT–St adj.	M	.047	.027	.039	.048	.052	.067	.133	0	0	.005	.058	.726
inpatient	IGT–St adj.	W	.051	.033	.037	.044	.046	.152	.211	.014	.095	.171	.294	.511
inpatient	lognorm	M	.096	.069	.082	.096	.106	.120	.011	0	0	0	0	.128
inpatient	lognorm	W	.111	.065	.086	.106	.129	.176	0	0	0	0	0	.003
inpatient	trbeta	M	.035	.030	.033	.035	.037	.046	.172	.003	.035	.039	.104	.948
inpatient	trbeta	W	.044	.034	.037	.039	.045	.087	.279	.044	.138	.293	.408	.558
inpatient	trgamma	M	.064	.047	.054	.065	.074	.082	.063	0	0	0	.003	.742
inpatient	trgamma	W	.076	.049	.058	.072	.079	.150	.018	0	.001	.007	.030	.068
dental	Burr	M	.043	.034	.041	.043	.045	.052	.048	0	0	0	.004	.53
dental	Burr	W	.047	.040	.045	.047	.050	.052	.028	0	0	0	.027	.169
dental	Burr adj.	M	.052	.040	.047	.050	.054	.070	.008	0	0	0	0	.068
dental	Burr adj.	W	.058	.050	.054	.057	.062	.068	.013	0	0	0	.001	.130
dental	EVT	M	.047	.030	.043	.047	.054	.060	.064	0	0	0	.006	.720
dental	EVT	W	.051	.044	.049	.050	.053	.057	.026	0	0	0	.030	.129
dental	gamma	M	.105	.085	.094	.103	.114	.133	0	0	0	0	0	0
dental	gamma	W	.114	.106	.107	.111	.117	.132	0	0	0	0	0	0
dental	GBS–logistic	M	.032	.016	.025	.030	.037	.047	.091	0	0	0	.009	.882
dental	GBS–logistic	W	.040	.031	.035	.037	.039	.068	.063	0	0	.003	.025	.451
dental	BS	M	.033	.018	.022	.030	.037	.062	.054	0	0	0	.023	.567
dental	BS	W	.040	.027	.034	.038	.041	.066	.053	0	0	.004	.016	.330
dental	GBS–St	M	.031	.014	.021	.030	.036	.050	.104	0	0	0	.031	.964
dental	GBS–St	W	.040	.032	.035	.036	.041	.070	.068	0	0	.003	.012	.508
dental	IGT–laplace	M	.050	.040	.046	.047	.053	.066	.005	0	0	0	0	.033
dental	IGT–laplace	W	.058	.051	.053	.055	.061	.083	.010	0	0	0	.001	.110
dental	IGT–logistic	M	.040	.031	.035	.039	.044	.052	.061	0	0	0	.019	.643
dental	IGT–logistic	W	.042	.032	.038	.040	.044	.052	.086	0	0	.001	.068	.394
dental	IGT–St	M	.038	.023	.036	.038	.042	.047	.091	0	0	0	.046	.938
dental	IGT–St	W	.043	.033	.040	.042	.046	.052	.077	0	0	0	.048	.407
dental	IGT–St adj.	M	.045	.026	.044	.046	.050	.057	.073	0	0	0	.007	.854
dental	IGT–St adj.	W	.050	.039	.045	.049	.053	.061	.046	0	0	0	.018	.245
dental	lognorm	M	.037	.029	.035	.038	.038	.045	.058	0	0	0	.018	.409
dental	lognorm	W	.044	.032	.042	.042	.044	.067	.039	0	0	.001	.022	.235
dental	lognorm adj.	M	.039	.032	.037	.038	.040	.050	.046	0	0	0	.010	.498
dental	lognorm adj.	W	.046	.037	.041	.043	.049	.065	.028	0	0	0	.008	.252
dental	Pareto	M	.044	.040	.042	.044	.045	.048	.024	0	0	0	.009	.214
dental	Pareto	W	.049	.038	.046	.047	.049	.068	.017	0	0	0	.012	.104
dental	trbeta	M	.037	.031	.034	.037	.038	.043	.057	0	0	0	.033	.430
dental	trbeta	W	.042	.036	.040	.041	.043	.054	.052	0	0	.001	.058	.339
dental	trgamma	M	.059	.050	.054	.058	.065	.072	.001	0	0	0	0	.014
dental	trgamma	W	.065	.054	.063	.064	.066	.091	.007	0	0	0	0	.082
full cover	Burr	M	.018	.011	.014	.016	.019	.031	.849	.658	.789	.844	.946	.999
full cover	Burr	W	.022	.014	.021	.022	.023	.029	.725	.285	.668	.736	.913	.975
full cover	Burr adj.	M	.020	.012	.014	.017	.022	.035	.797	.524	.687	.865	.898	.993
full cover	Burr adj.	W	.023	.015	.021	.022	.025	.031	.685	.200	.612	.723	.883	.951
full cover	gamma	M	.080	.074	.075	.078	.082	.093	0	0	0	0	0	0
full cover	gamma	W	.064	.039	.055	.065	.077	.081	.011	0	0	.001	.001	.061
full cover	IGT–St	M	.017	.012	.012	.016	.019	.029	.838	.511	.754	.901	.987	.997
full cover	IGT–St	W	.022	.013	.023	.024	.025	.025	.693	.261	.513	.765	.918	.972
full cover	IGT–St adj.	M	.022	.013	.019	.021	.022	.036	.634	.318	.418	.584	.874	.986
full cover	IGT–St adj.	W	.035	.019	.024	.033	.041	.059	.427	.001	.098	.451	.687	.910
full cover	lognorm	M	.042	.033	.037	.043	.047	.048	.104	.001	.011	.103	.168	.247
full cover	lognorm	W	.048	.040	.041	.045	.051	.064	.081	.002	.004	.009	.139	.281
full cover	trbeta	M	.019	.012	.014	.016	.019	.037	.784	.441	.593	.884	.973	.996
full cover	trbeta	W	.020	.014	.016	.021	.024	.025	.812	.505	.733	.883	.946	.954

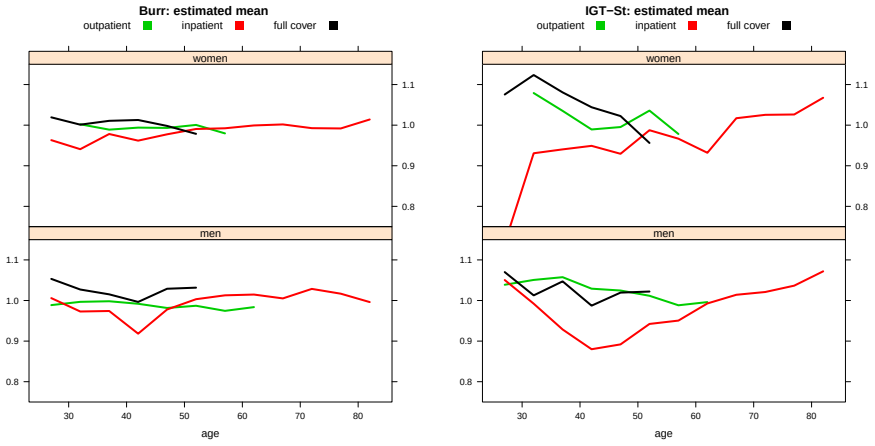
While considering this table we should keep in mind that we cannot hope that our claim size distributions are amongst the known named distributions. We want to find only a good approximation. Here, a comparison of the p–values only can be misleading because it is heavily influenced by the number of observations. In some cohorts we have a huge number of observations, hence a small deviation of the claim size distribution from the fitted one will lead to a small p–value. Therefore, it is important to look at the statistic of the KS–test as well, which is the maximum distance of the cdfs of the two distributions.

The first glimpse at the table confirms that the gamma distribution is unsuitable for approximating any of the claim size distributions. The lognormal distribution can only be useful in dental insurance. However, a closer look reveals that it overestimates the tail in this case.

The Burr, IGT–St, and transformed beta distributions are the most suitable distributions for the modeling of outpatient and full cover insurance whose claims are dominated by the outpatient claims. (Ignore the "adj." lines in the table for the time being.) The transformed beta distribution is the only four parameter distribution among them. It does not fit the claim size distribution notably better than the two other three parameter distributions. Under such circumstances the model with the least parameters should be chosen, following general principles. In addition, sometimes numerical difficulties arise during the ML–estimation of the four parameters. Thus we will consider only the Burr and IGT–St distribution further.

In inpatient insurance we are unable to get a fit of similar quality as in outpatient and full cover insurance. However, the same distributions still give a good fit and even the comparison of their fits is as above. The table shows that the EVT, GBS–St, and transformed gamma distributions provide a fit of the same quality. However, the qq– and pp–plots indicate an inferior fit: The GBS–St distribution fits the tail very badly and the the graph of the pp–plot of the transformed gamma distribution shows an S–shape. The EVT distribution provides a good fit in the younger ages, but has problems with the tail for the older ages. While from the quality of the fit it might be worthwhile to examine the EVT distribution further for the inpatient claim size distribution, we will neglect it because it fits the other insurance type badly.

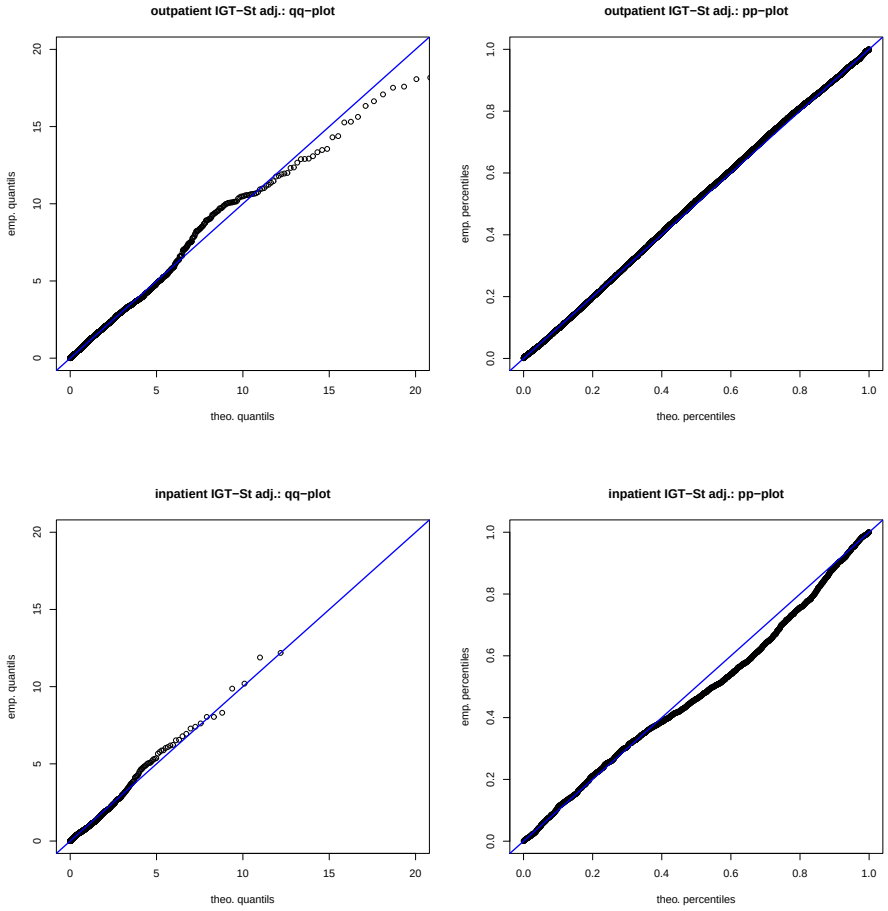
We estimate the parameters for the Burr and IGT–St distribution by the maximum likelihood method and plot the resulting estimated means. Due to our scaling they should be 1.

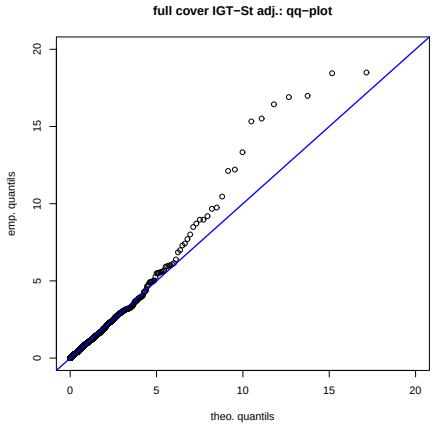


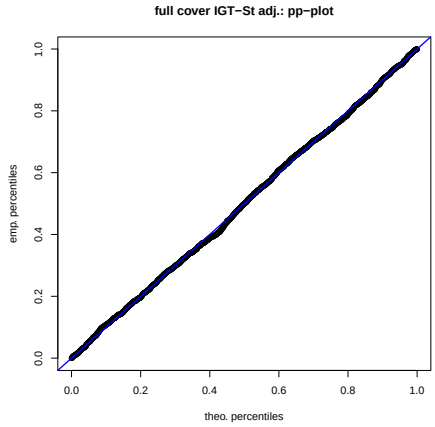
The IGT–St distribution obviously fits the cohorts of 25–29 year old women in inpatient insurance badly. This cohort is different from the other because it contains many claims due to pregnancy. In addition, it is the smallest one, thus we have the largest estimation error here. However, even apart from this the

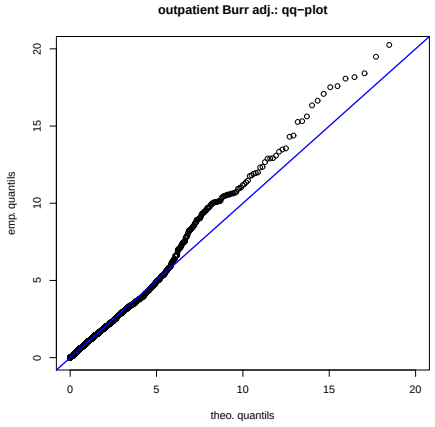
estimated mean deviates often by more than 5% for both distributions. As this effect varies slowly with age, we have to assume that it is not random, but that there is a systematic difference between the empirical claim size distributions and the theoretical ones. A systematic deviation of the expected mean from the true one is not acceptable for many applications, for example premium calculation. Therefore, we force the mean to be one and estimate only the remaining parameters by ML–estimation. For the IGT–St distribution we set $\mu = 1$; for the Burr distribution we estimate the parameters α, β by maximum likelihood while simultaneously choosing the scale parameter such that the mean is one. The resulting KS–test results are the "adjusted" lines in above table.

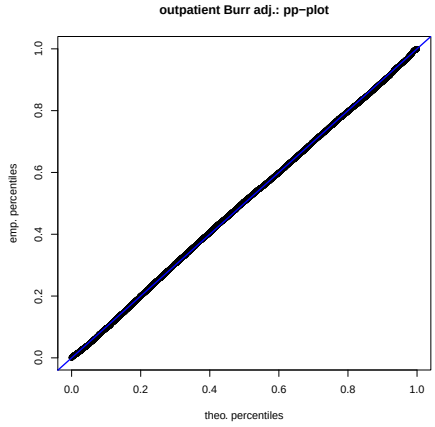
To get an impression of the fit we draw the qq– and pp–plot of the cohorts of the 40–44 year old men. We cut off the qq–plots at 20 because we have a detailed discussion about the tails in the next subsection.

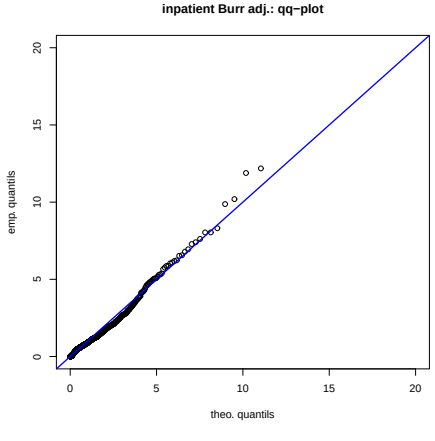


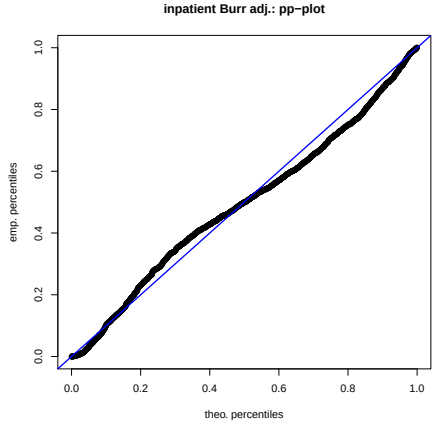

full cover IGT−St adj.: qq−plot
emp. quantils
theo. quantils
0
5
10
15
20

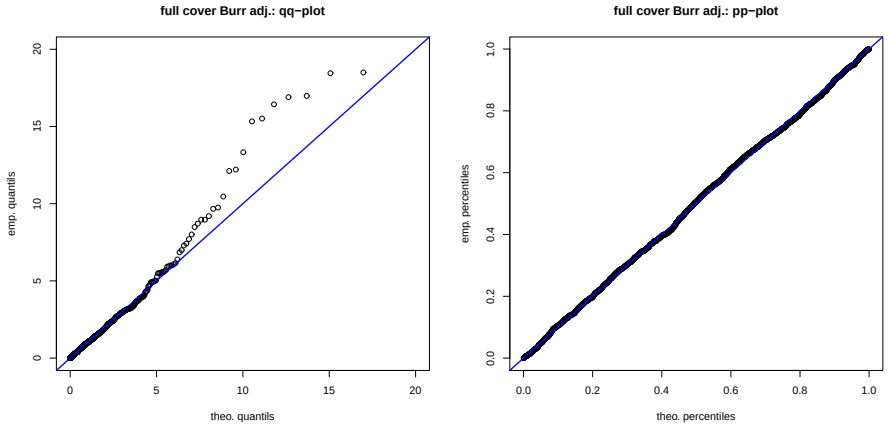

full cover IGT−St adj.: pp−plot
emp. percentiles
theo. percentiles
0.0
0.2
0.4
0.6
0.8
1.0


outpatient Burr adj.: qq−plot
emp. quantils
theo. quantils
0
5
10
15
20


outpatient Burr adj.: pp−plot
emp. percentiles
theo. percentiles
0.0
0.2
0.4
0.6
0.8
1.0

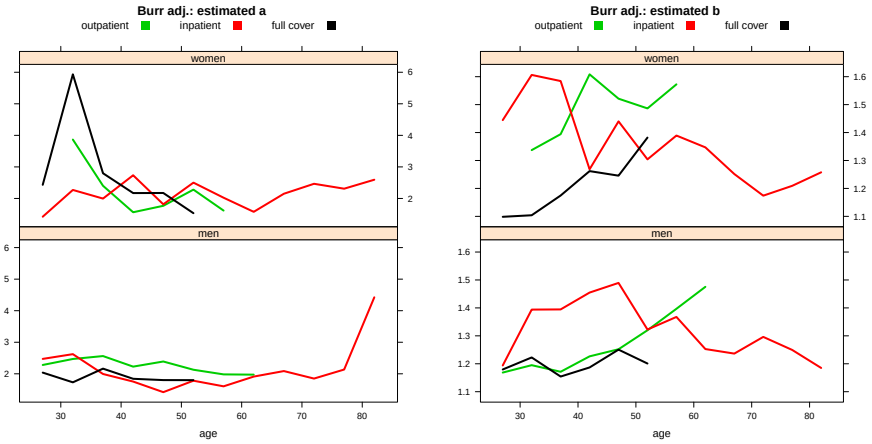

inpatient Burr adj.: qq−plot
emp. quantils
theo. quantils
0
5
10
15
20

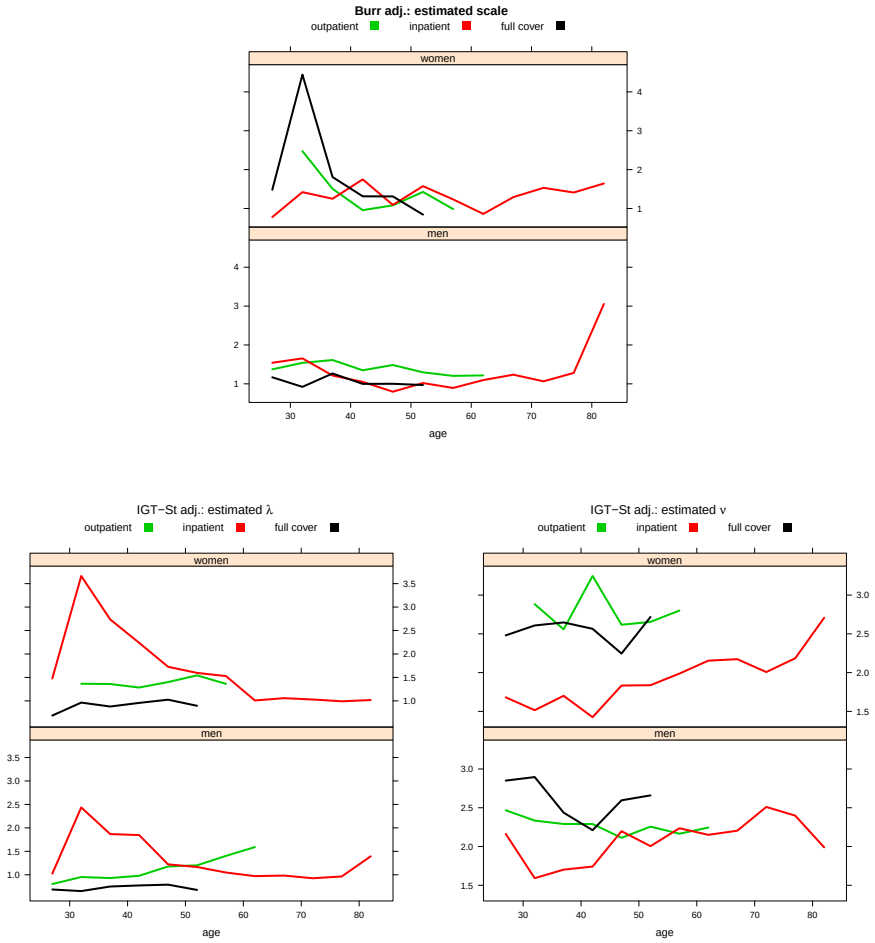

inpatient Burr adj.: pp−plot
emp. percentiles
theo. percentiles
0.0
0.2
0.4
0.6
0.8
1.0



The pp–plots confirm the impression we got from the table about the KS–test: both distributions fit the claim size distributions of outpatient and full cover insurance very well, but the inpatient insurance less well. The qq–plot shows that beyond six the empirical distributions deviate somewhat from the theoretical ones. However, this is already an area where there are only few observations and random effects will be noticeable. In the next subsection we will show that we have a good fit in the tail overall.

We plot the estimated parameters after the adjustment to mean one:

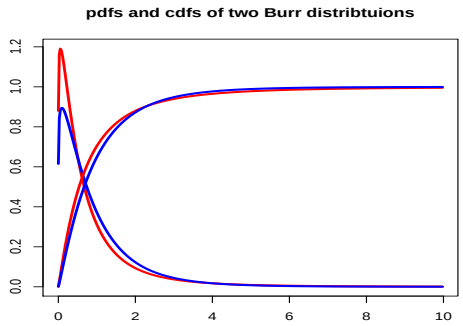




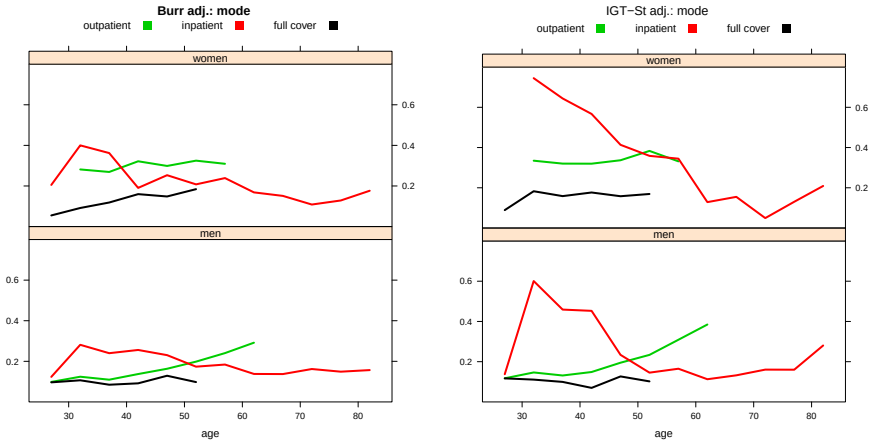
In general, the graphs are well–behaved, not too volatile, and mainly constant with the exception of impatient insurance, which shows a clear trend. Let us discuss the obvious exceptions first. We already noted above that the 25–29 year old women in inpatient insurance are exceptional, and we see that the IGT–St distribution has also exceptional parameters for the men of this age group. Since no obvious outliers were visible in the data, we can only speculate about the reasons: One possibility is that due to few observations in those cohorts an estimation of a distribution function is too unreliable, another is that this age group contains many newly insureds, which have thus recently passed a risk assessment and therefore distort the distribution.

The other prominent exception are the parameters for the Burr distribution for the 30–34 year old women in full cover insurance. Clearly, claims due to pregnancy will play an important role here. However, here we see also a disadvantage of the Burr distribution, namely that the interpretation of its parameters is difficult. The parameters of the exceptional cohort are $(a, b, scale) = (5.93, 1.10, 4.44)$ and the ones of a neighboring cohort are $(2.43, 1.10, 1.48)$. While the values are very different, the following plot shows

that the resulting pdfs and cdfs are similar:



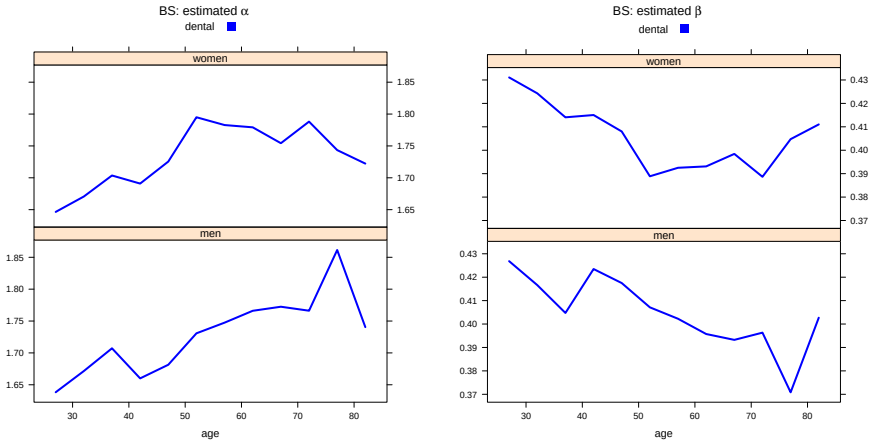
Apart from these exceptions the graphs are well–behaved. To get a better intuitive understanding and to be able to compare both distributions we plot the mode, i.e., the point of maximum of the pdf:

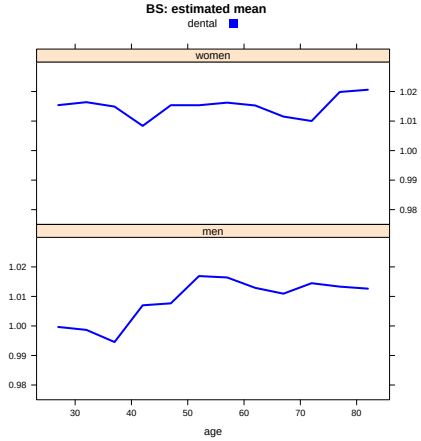


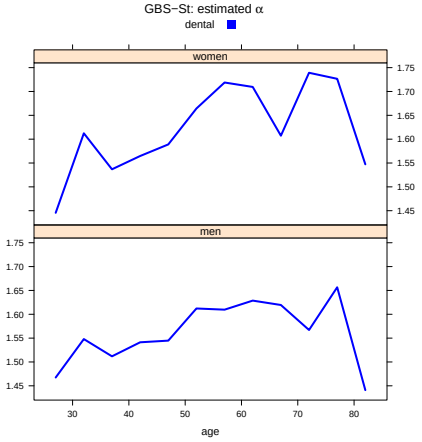
Most prominent are the decrease in inpatient insurance as well as the increase in outpatient insurance for men. Looking back at the plots of the empirical pdfs in Subsection 2.1 we see that these cases are in fact the ones where the profiles and the pdfs vary at most with age. We also note that the modes for the Burr and the IGT–St distribution fits differ somewhat, especially in inpatient insurance where the fit of both distributions is not as good as in inpatient and full cover insurance.

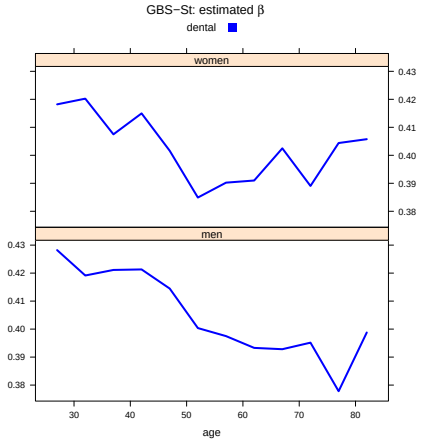
The claim size distribution in dental insurance behaves very differently from the remainder. This is due to fact that very high claims are nearly impossible, and thus the distribution has at most a slightly heavy tail, see also the discussion in the next subsection. The best fit is achieved by the generalized Birnbaum–Saunders distribution with normal, logistic, and Student–t kernel. As the fits with normal and logistic kernel are very similar, we will focus on the classical Birnbaum–Saunders distribution with normal kernel. The GBS distribution with Student–t kernel usually has a slightly better fit in the tail, but it comes at

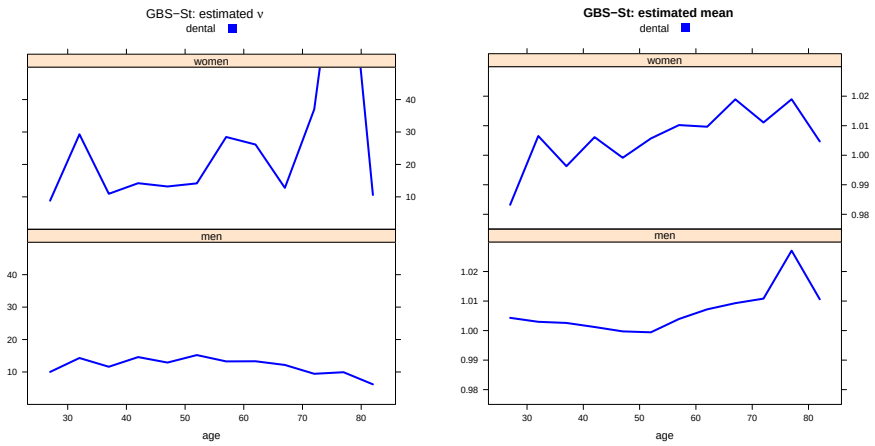
the disadvantage of an additional parameter. The parameters by ML–estimation and the resulting expected mean are as follows:



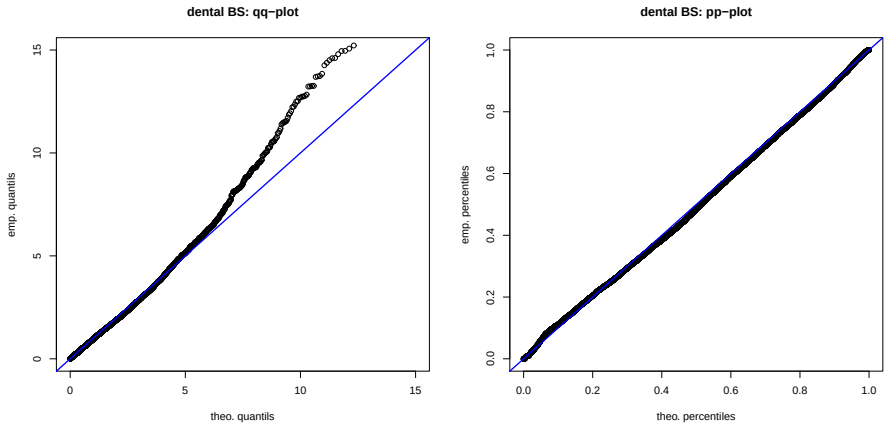


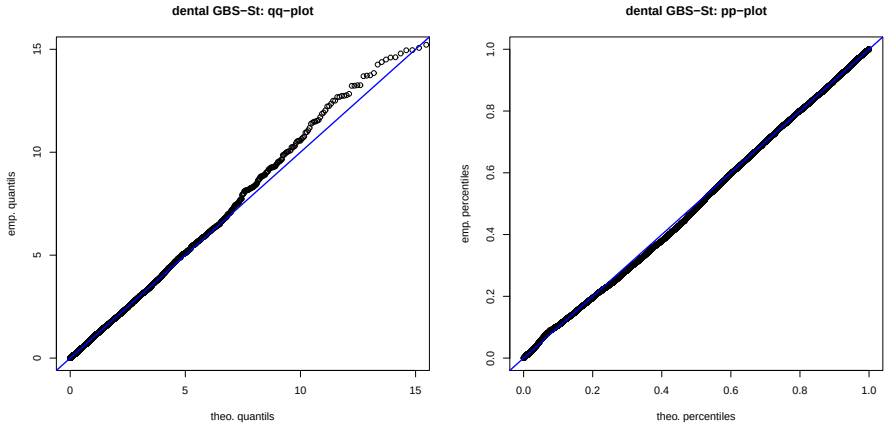




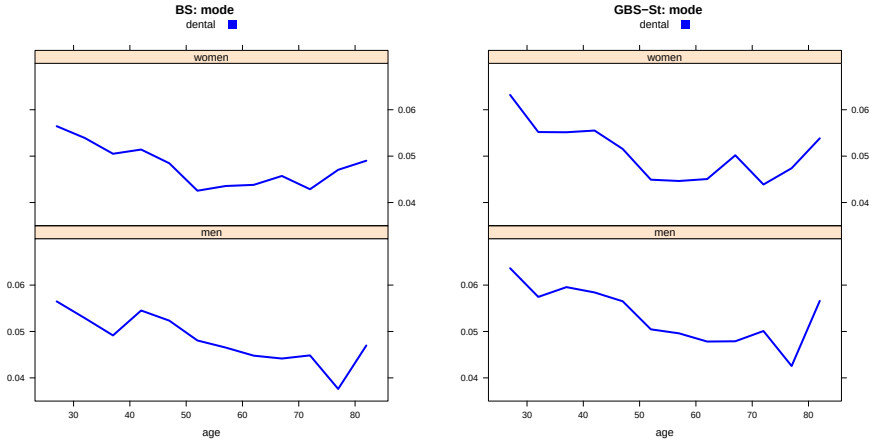


The graphs are in general well–behaved and in addition the resulting expected mean is very close to 1, thus we do not make any correction of the parameters this time. The only problem is that the ML–estimation the ν parameter for the GBS–St distribution is not robust for ν in the above range, see Appendix B. In the cohort of 75–79 year old women we have $\nu > 100$. However, this is acceptable because the changes of ν in this range have only a minor impact on the shape of the pdf. To get a better impression of the quality of the fit, we show again the qq– and pp–plots for the 40–44 year old men:



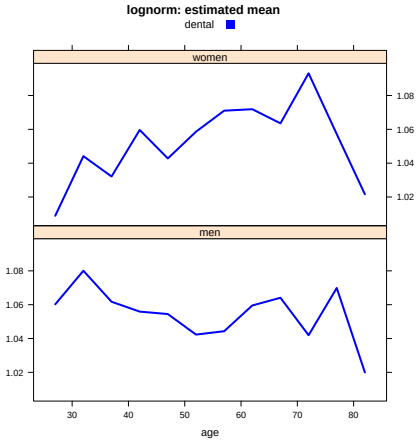


We compare the fits of the distributions by their modes:

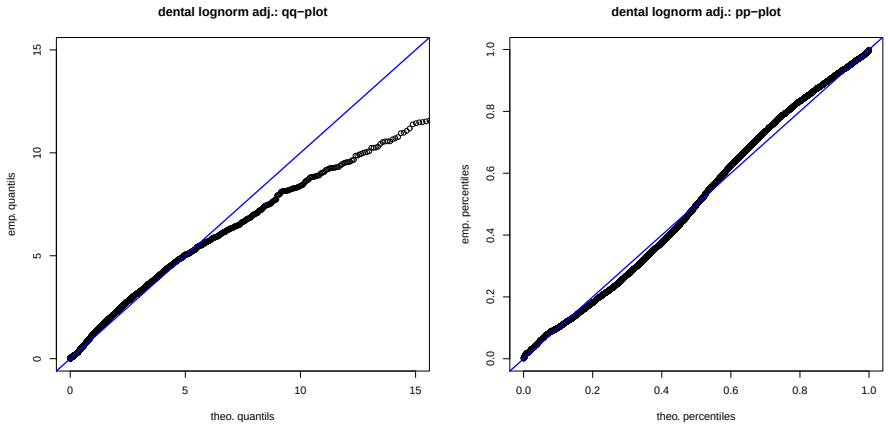


The graphs are similar and both reveal a slight shift of the mode to the left with increasing age. In fact, we can see this trend in the plot of empirical distribution functions in Subsection 2.1.

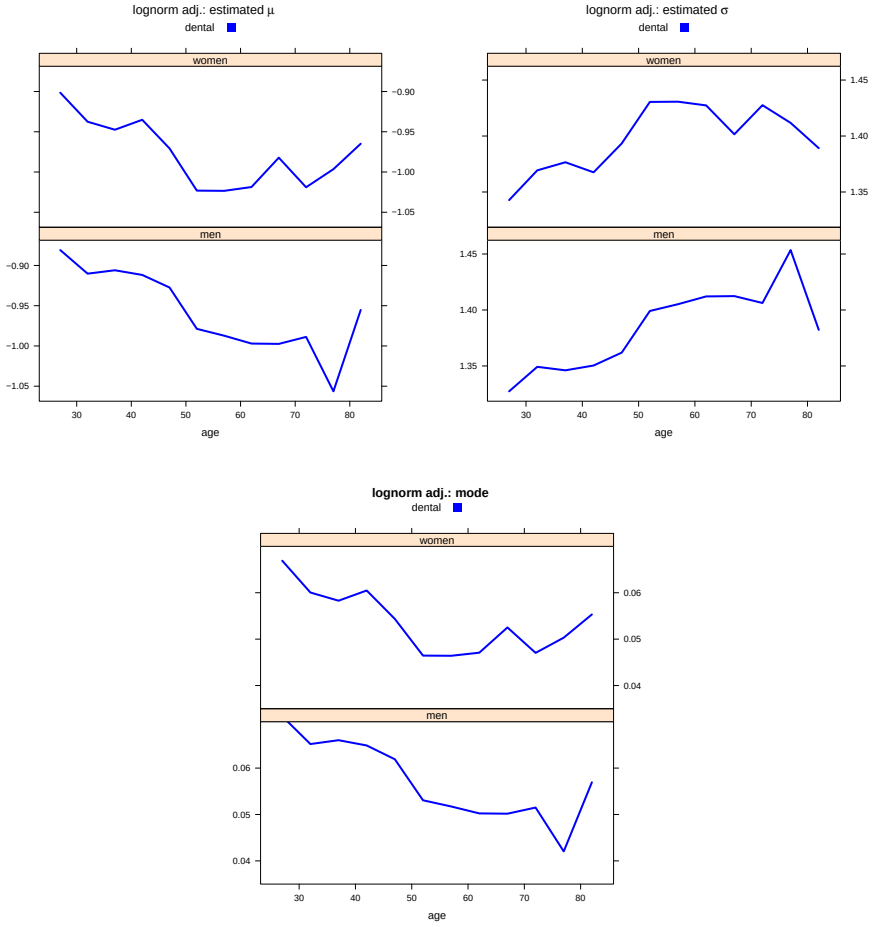
If only the traditional distributions are available for modeling the claim size distributions one should choose the lognormal distribution. All other distributions except the gamma distribution fit the tail very badly — as qq–plots show — and the above table shows that the lognormal distribution provides a much better fit than the gamma distribution. We estimate the parameters by applying the maximum likelihood method and plot the resulting means:



As the mean is often off by more than 5% we adjust the parameters such that the mean is 1. This is done by a ML–estimation of σ while simultaneously choosing μ such that the mean is one. For comparison we give again the qq– and pp–plots:



The resulting parameters and modes are:



The graphs are again well–behaved. Like the other two distributions the fit reveals a slight shift of the mode to the left with increasing age.

2.3 Examination of the Tail

The tail of a claim size distribution is essential for pricing high–excess loss layers in reinsurance or estimating the risk borne by the insurance company. To examine the tails of our distribution we apply Extreme Value Theory in the form of the excess over threshold approach. We will recall briefly the facts of interest to us, see [MFE, Sec. 7] for an extensive introduction and [CDL] for an application to large claims in American health insurance.

The excess distribution of a random variable X with cdf F over threshold u has cdf

$$F_u(x) = \mathrm{P}(X - u \leq x | X > u) = \frac{F(x+u) - F(u)}{1 - F(u)}$$

for $x \leq x_F - u$ where $x_F \leq \infty$ is the right endpoint of F.

Under reasonable assumptions which hold generally in practice the theorem of Pickands, Balkema, and de Haan states that for $u \to x_F$ the excess distribution converges to a generalized Pareto distribution (GPD), whose cdf is given by

$$G_{\xi,\beta}(x) = \begin{cases} 1 - (1 + \xi x/\beta)^{-1/\xi} & \text{for } \xi \neq 0, \\ 1 - \exp(-x/\beta) & \text{for } \xi = 0, \end{cases}$$

where ξ is called the shape and $\beta > 0$ the scale parameter. The x–value must be in the range $[0, \infty[$ for $\xi \geq 0$ and $[0, -\beta/\xi]$ for $\xi < 0$. Note that for $\xi = 0$ the GPD is the scaled exponential distribution and that this is the limit of $G_{\xi,\beta}$ for $\xi \to 0$. The GDP has the property that its excess distribution over threshold u is again a GPD of the same shape. More precisely, if $F = G_{\xi,\beta}$ then $F_u = G_{\xi,\beta(u)}$ with $\beta(u) = \beta + \xi u$.

An important consequence is that the GPDs approximating the excess distributions of some distribution F over a sufficient high threshold will have similar shapes, and we can define the limit of $\alpha = 1/\xi$ for $u \to x_F$ to be the tail index of the distribution F. For $\xi > 0$ the distribution is called heavy tailed. In this case the higher moments $\mathrm{E}(X^k)$ for $k \geq 1/\xi = \alpha$ are infinite.

If F is given analytically the easiest way to compute the tail index it to use a theorem of Gnedenko, which says for a heavy tail distribution

$$\overline{F}(x) := 1 - F(x) = x^{-\alpha} L(x)$$

with L a slowly varying function at ∞, i.e., $\lim_{x\to\infty} L(tx)/L(x) = 1$ for all $t > 0$. By Karamata's Theorem this is equivalent to

$$f(x) = x^{-\alpha-1} \tilde{L}(x)$$

for the pdf f with another slowly varying function $\tilde{L}$ if f is finally monotonically decreasing.

The main practical application of Extreme Value Theory is to improve the estimates related to the tail of a distribution. If the excess distribution of F over the threshold u is $G_{\xi,\beta}$ we obtain the tail probability

$$\overline{F}(x) = \mathrm{P}(X > u)\mathrm{P}(X > x | X > u) = \overline{F}(u)\overline{F}_u(x-u) = \overline{F}(u)\left(1 + \xi\frac{x-u}{\beta}\right)^{-1/\xi}.$$

The value at risk is for $z \geq F(u)$

$$\mathrm{Var}_z = u + \frac{\beta}{\xi}\left(\left(\frac{1-z}{\overline{F}(u)}\right)^{-\xi} - 1\right).$$

For $\xi < 1$ the expected shortfall exists as well and can be calculated as

$$\mathrm{ES}_z = \frac{\mathrm{Var}_z}{1-\xi} + \frac{\beta - \xi u}{1-\xi}.$$

In practice, one uses for $\overline{F}(u)$ the estimator n_u/n where n_u is the number of observations above u and n is the total number of observations. Such an estimator becomes unreliable if there are only few observations above the threshold.

This is the reason why one does not use such an estimator for $\overline{F}(x)$ with $x > u$, but instead the above formula where the estimator for $\overline{F}(u)$ is still reasonably reliable.

Now we apply the theory to our distributions. A major difficulty is to choose the threshold u because we have two conflicting goals: The higher u the better will be the GPD approximation to F_u; the lower u the more data are available to fit the GPD to F_u. There is no canonical way for this choice. The best known methods are graphical, like the Hill plot, the sample mean excess plot, or simply plotting the ML–estimator for ξ in dependence of u.

In the sample mean plot one plots an empirical estimate for $e(u) = \mathrm{E}(X - u|X > u)$. In case of a GPD one finds

$$e(u) = \frac{\beta(u)}{1-\xi} = \frac{\beta + \xi u}{1-\xi},$$

i.e., a straight line with ascending slope for a heavy tailed distribution. In the plots our distributions clearly exhibit this tail behavior with the exception of dental insurance. There one finds that in the middle age groups the graph of $e(u)$ becomes flat beyond 6 and for older ages the flat part starts already at 2–3. This indicates that in contrast to the other insurance types dental insurance is not heavy tailed — at least in the middle and older ages.

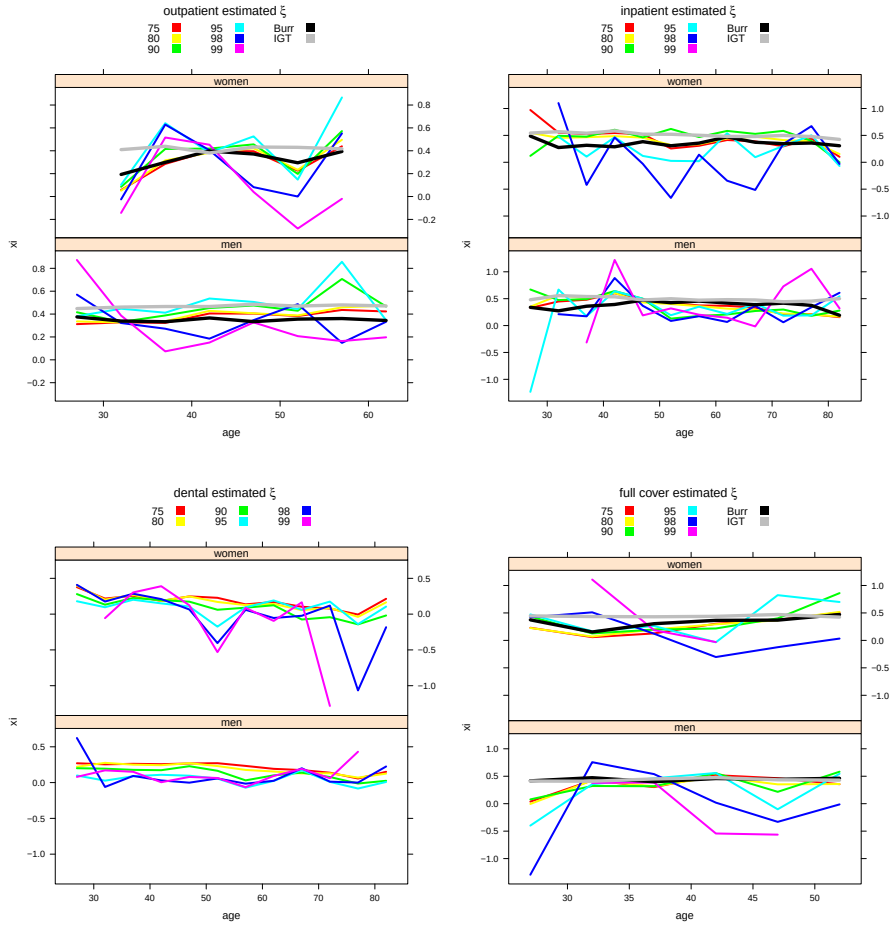
Of course, in practice it is well–known that extremely high claims in dental insurance are impossible and thus the distribution should not be heavy tailed.

Because we have so many claim size distributions we cannot choose our threshold based on plots. We will use thresholds based on the empirical 75%, 80%, 90%, 95%, 98%, 99% quantile of the distributions. We fit the GPD by a ML–estimation to the tail, which is well–behaved for $\xi > -0.5$. The table below gives the statistics and the p–values of the Kolmogorov–Smirnov tests of the GPD fits to the distribution tails. To reduce the amount of data we group the data by insurance type, sex, and threshold quantile and give the mean and quantiles of the statistics and the p–values over the age groups.

type	sex	thres.	d-statistic						p–value					
		quant.	mean	min	$q_{.25}$	$q_{.5}$	$q_{.75}$	max	mean	min	$q_{.25}$	$q_{.5}$	$q_{.75}$	max
outpatient	M	75	.02	.01	.02	.02	.03	.04	.51	.11	.39	.60	.67	.79
outpatient	M	80	.03	.01	.02	.03	.04	.05	.49	.05	.31	.58	.70	.82
outpatient	M	90	.03	.02	.02	.03	.04	.05	.68	.30	.57	.72	.86	.93
outpatient	M	95	.05	.02	.04	.04	.05	.07	.67	.16	.56	.77	.88	.99
outpatient	M	98	.06	.03	.05	.06	.07	.12	.71	.34	.55	.67	.99	1
outpatient	M	99	.09	.05	.07	.10	.12	.13	.72	.23	.50	.84	.96	.98
outpatient	F	75	.03	.02	.02	.03	.03	.04	.86	.53	.85	.91	.96	.99
outpatient	F	80	.03	.02	.02	.03	.03	.04	.89	.69	.78	.98	.99	1
outpatient	F	90	.05	.03	.04	.04	.05	.06	.82	.71	.76	.81	.89	.95
outpatient	F	95	.05	.04	.05	.06	.06	.07	.92	.76	.91	.96	.98	.98
outpatient	F	98	.09	.07	.08	.08	.09	.15	.85	.70	.75	.87	.92	.99
outpatient	F	99	.15	.09	.13	.14	.18	.22	.68	.56	.58	.65	.68	1
inpatient	M	75	.04	.02	.03	.03	.05	.07	.76	.43	.59	.77	.94	.95
inpatient	M	80	.04	.02	.03	.04	.05	.08	.73	.26	.58	.82	.93	1
inpatient	M	90	.06	.03	.04	.05	.06	.14	.82	.57	.73	.80	.96	1
inpatient	M	95	.09	.04	.06	.07	.09	.26	.78	.37	.70	.86	.88	1
inpatient	M	98	.11	.07	.08	.10	.14	.15	.82	.60	.67	.88	.96	.99
inpatient	M	99	.15	.10	.12	.13	.17	.24	.80	.26	.73	.90	.97	.99
inpatient	F	75	.06	.04	.05	.05	.07	.14	.70	.10	.61	.76	.90	.96
inpatient	F	80	.06	.04	.05	.06	.07	.13	.73	.18	.57	.78	.92	1
inpatient	F	90	.09	.05	.06	.08	.11	.14	.72	.27	.47	.80	.96	.98
inpatient	F	95	.11	.07	.09	.11	.13	.16	.74	.36	.58	.66	.95	.98
inpatient	F	98	.17	.09	.14	.17	.21	.26	.73	.38	.60	.72	.93	1
dental	M	75	.03	.01	.02	.02	.03	.07	.59	.14	.35	.61	.82	.98
dental	M	80	.03	.01	.02	.02	.03	.08	.60	.09	.39	.63	.80	1
dental	M	90	.04	.02	.03	.03	.05	.09	.57	.03	.35	.61	.80	1
dental	M	95	.05	.02	.03	.04	.06	.15	.74	.21	.54	.90	.95	.99
dental	M	98	.08	.03	.04	.06	.10	.17	.77	.10	.65	.95	.97	.99
dental	M	99	.09	.05	.06	.08	.10	.17	.79	.41	.61	.87	.96	.99
dental	F	75	.03	.01	.02	.03	.04	.06	.77	.08	.77	.85	.93	.98
dental	F	80	.04	.02	.02	.03	.06	.06	.74	.15	.67	.77	.90	.99
dental	F	90	.05	.03	.04	.05	.07	.09	.70	.24	.64	.71	.89	.99
dental	F	95	.07	.03	.04	.06	.10	.15	.78	.33	.71	.84	.91	.98
dental	F	98	.11	.04	.07	.09	.14	.25	.76	.15	.65	.85	.97	1
dental	F	99	.12	.07	.09	.10	.13	.26	.74	.28	.48	.88	.95	.98
full cover	M	75	.04	.02	.03	.03	.04	.06	.85	.52	.82	.93	.96	.99
full cover	M	80	.04	.03	.04	.04	.04	.06	.88	.75	.86	.89	.93	.94
full cover	M	90	.07	.05	.06	.06	.07	.09	.70	.51	.69	.71	.78	.81
full cover	M	95	.10	.04	.09	.10	.12	.15	.60	.39	.53	.55	.58	.99
full cover	M	98	.13	.08	.10	.12	.14	.25	.74	.40	.53	.82	.92	.99
full cover	M	99	.17	.14	.16	.17	.18	.18	.73	.54	.67	.76	.82	.85
full cover	F	75	.05	.02	.04	.05	.05	.08	.66	.24	.43	.76	.85	1
full cover	F	80	.05	.03	.03	.05	.07	.09	.72	.27	.61	.75	.92	.99
full cover	F	90	.07	.05	.06	.06	.07	.09	.76	.40	.71	.80	.90	.97
full cover	F	95	.09	.08	.08	.09	.10	.11	.75	.40	.67	.79	.88	.99
full cover	F	98	.15	.11	.13	.14	.17	.20	.70	.50	.58	.64	.83	.95
full cover	F	99	.12	.11	.11	.12	.12	.12	.98	.96	.97	.97	.98	.99

Clearly, the p–values are all very good. However, the statistics for the quantiles 99% and 98% are notably larger than for lower quantiles. This suggests that one should base the estimation of ξ on a quantile threshold below or equal to 95%.

The resulting ξ are plotted below. The figures also contain the estimation of ξ based on the Burr and IGT–St fit to the entire distribution as thick lines.



We note that the estimates based on the 99% and 98% quantiles are very erratic over the age groups and therefore are not trustworthy. The remaining estimates are very close together and define the interval where the true ξ should be. We see that the two extremely high claims for the 40–44 year old males in inpatient insurance, which already disturbed the estimate of the variance coefficient, lead to very different estimates for ξ based on the 99%, 98%, and 95% quantile threshold.

The most remarkable fact is that our estimation for ξ based on the ML–estimation of the Burr and IGT–St distribution adjusted to mean one fits very well into our empirically estimated ξ. The estimated ξ is smoothed over age groups compared to the empirically estimated; this effect is even more prominent in case of the IGT–St distribution. Note in particular the very welcome effect that the two extremely high claims for the 40–44 year old males in inpatient insurance do not seem to disturb the estimation of ξ by this method. Further, the IGT–St distribution leads to a slightly larger ξ than the one based on the Burr distribution. Summing up we see that both distributions approximate even the tails of the claim size distributions very well.

For dental insurance we fitted the BS, GBS–St, and lognormal distribution to the claim size data. The BS distribution has an exponential tail, thus $\xi = 0$. The lognormal distribution has $\xi = 0$ as well. For the GBS–St distribution we found mostly $\nu > 10$, i.e., $0.2 \geq \xi \geq 0$. These distribution based estimates match the small values of ξ that we find empirically. This mostly confirms the a priori expectation that dental insurances should not have a heavy tail. Note however that the data indicate a slightly heavy tail — in particular for the younger ages — and thus suggest the use of the GBS–St distribution.

3 The quality of the normal approximation

This final section only wants to caution about the use of the central limit theorem without any error estimation. While we know from textbook examples that the mean of ten uniform random variables yields a nearly perfect Gaussian distribution, this will not be the case for the unconditional claim size distributions. The best known estimate about the speed of convergence in the central limit theorem is the theorem of Berry–Essen: Let X_n be i.i.d. random variables with $\mu = \mathrm{E}(X_n)$ and $\sigma^2 = \mathrm{Var}(X_n)$. Define $Z_n = \sum_{i=1}^{n}(X_i - \mu)/\sigma\sqrt{n}$ as the standardized mean of the first n variables. Then we have for the cdf F_n of Z_n

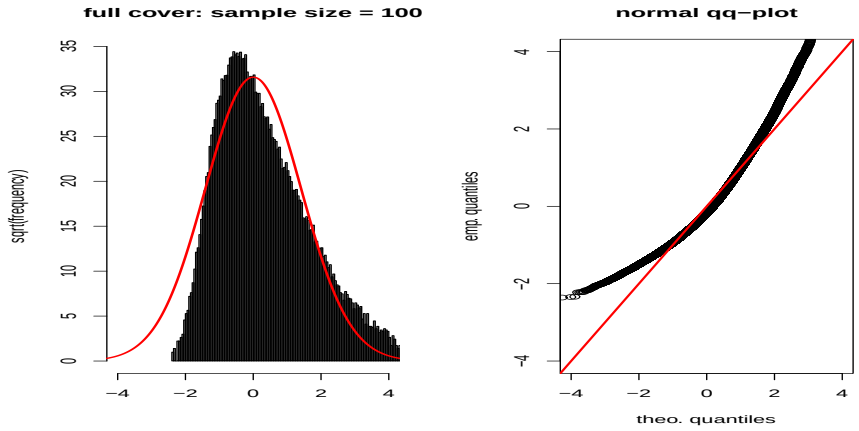
$$|F_n(x) - \Phi(x)| \leq \frac{C\varrho}{\sigma^3\sqrt{n}} \quad \text{for all } x \in \mathbb{R},$$

where Φ is the cdf of the standard Gaussian, C the universal Berry–Essen constant with a value between 0.39 and 0.77 and $\varrho = \mathrm{E}(|X_n - \mu|^3)$ the third absolute moment. The speed of convergence $1/\sqrt{n}$ cannot be improved as the example of the Bernoulli distribution with mean $1/2$ shows. Note that the unconditional claim size distribution has a high weight at 0 and therefore the characteristic of a Bernoulli distribution. Hence we expect to have a similar speed of convergence. In addition, with the exception of dental insurance the claim size distributions have a heavy tail and hence possess a large third absolute moment. Thus we have two reasons to expect a slow convergence in the central limit theorem.

We confirm this with a simulation study on the claims of the 40–44 year old men. We standardize the claims by their empirical mean and variance. Then we examine the distribution of 50000 means of k randomly chosen values of this standardized claim size distribution.

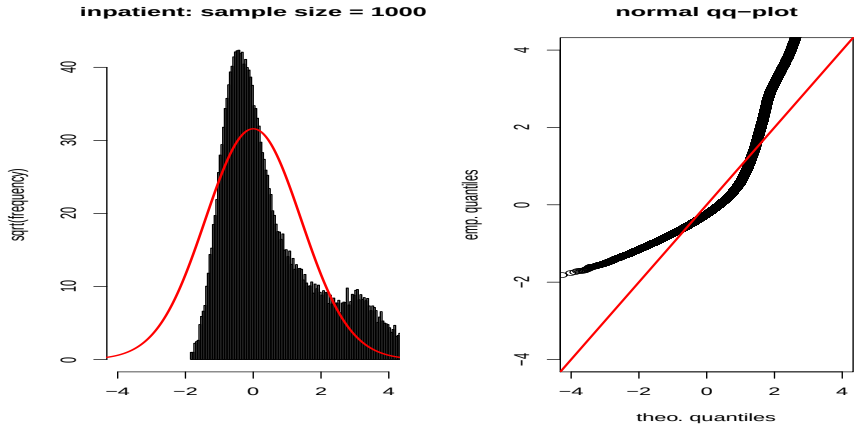
In case of full cover insurance a sample size of $k = 100$ still leads to a clearly

skewed distribution.

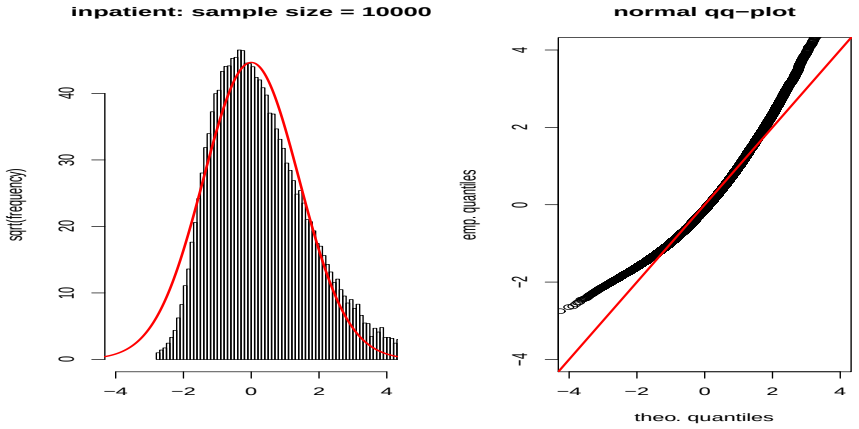


For a sample size of 1000 the resulting distribution looks Gaussian in the range of $[-2, 2]$, while the tail clearly diverges from Gaussian behavior.

The claims of inpatient insurance exhibit the slowest convergence to the Gaussian distribution due to the highest weight at 0 and the very heavy tail. For a sample size of $k = 1000$ the distribution bears nearly no resemblance to the Gaussian distribution:



For a sample size of 10000 the Gaussian distribution can be recognized in the interval $[-1, 2]$, but the tails are clearly not Gaussian:



In praxis the claims of an entire group of policy holders are approximated by a Gaussian distribution. Usually a risk margin for a premium or a risk capital requirement for solvency purposes are based on high quantiles. In case of Solvency II the 99.5% quantile is used. For the Gaussian distribution this is 2.576; however, our simulation study for inpatient insurance of 1000 and 10000 policy holders yields the higher values of 4.246 and 3.218, respectively. This shows again that the normal approximation should be used with caution.

The following table gives an overview of the simulation study. Beside the 95%, 99%, and 99.5% quantiles it contains the statistic and p–value of the Kolmogorov–Smirnov test against the Gaussian and the Anderson–Darling normality test. For comparison the textbook examples of the uniform and exponential distribution are included.

					Kolmogorov–Smirnov		Anderson–Darling	
type	k	$q_{.95}$	$q_{.99}$	$q_{.995}$	statistic	p–value	statistic	p–value
normal theo.	—	1.645	2.326	2.576	0	1	0	1
uniform	10	1.632	2.279	2.517	0.004	0.537	1.02	0.011
exp	10	1.800	2.784	3.188	0.041	0	208.4	0
exp	10^2	1.695	2.478	2.778	0.015	0	21.6	0
exp	10^3	1.652	2.362	2.635	0.005	0.106	2.36	0
exp	10^4	1.636	2.325	2.563	0.006	0.096	0.929	0.018
full cover	10	1.716	4.444	5.282	0.163	0	Inf	0
full cover	10^2	1.863	2.937	3.418	0.070	0	530.3	0
full cover	10^3	1.723	2.521	2.856	0.0176	0	42.3	0
full cover	10^4	1.681	2.419	2.688	0.009	0.001	9.03	0
full cover	10^5	1.645	2.335	2.612	0.0038	0.455	0.597	0.119
outpatient	10	1.808	3.744	4.778	0.1262	0	Inf	0
outpatient	10^2	1.808	3.081	3.593	0.057	0	392.5	0
outpatient	10^3	1.708	2.553	2.879	0.0208	0	43.1	0
outpatient	10^4	1.677	2.402	2.683	0.0075	0.007	7.40	0
outpatient	10^5	1.639	2.342	2.607	0.0038	0.462	0.597	0.119
inpatient	10	0.818	2.250	2.978	0.3829	0	Inf	0
inpatient	10^2	0.960	4.095	6.920	0.198	0	Inf	0
inpatient	10^3	2.183	3.706	4.246	0.1612	0	Inf	0
inpatient	10^4	1.820	2.848	3.218	0.0499	0	297.3	0
inpatient	10^5	1.698	2.439	2.746	0.0153	0	26.1	0
inpatient	10^6	1.662	2.372	2.655	0.0039	0.433	1.42	0.001
dental	10	1.953	3.608	4.332	0.1285	0	Inf	0
dental	10^2	1.795	2.788	3.200	0.0425	0	204.5	0
dental	10^3	1.705	2.482	2.787	0.0115	0	22.3	0
dental	10^4	1.670	2.375	2.621	0.0075	0.007	3.40	0
dental	10^5	1.653	2.343	2.600	0.0027	0.862	0.468	0.249

A The IGT–St Distribution

Sanhueza, Leiva, and Balakrishnan invented a new class of distributions by generalizing the Inverse Gaussian distribution in the following way [SLB]: Let Z be a random variable with standard symmetrical distribution, i.e., $\mathrm{E}(Z) = 0$, $\mathrm{Var}(Z) = 1$, and its pdf can be written as $h(t) = g(t^2)$, $t \in \mathbb{R}$. Then the random variable X with pdf

$$f(t) = h\left(\frac{\sqrt{\lambda}(t-\mu)}{\mu\sqrt{t}}\right)\frac{\sqrt{\lambda}}{\sqrt{t^3}} \quad \text{for } t > 0$$

is called an inverse Gauss type (IGT) distribution with parameters $\mu, \lambda > 0$ and kernel g. We abbreviate this as $X \sim \mathrm{IGT}(\mu, \lambda; g)$. If g is the kernel of the standard normal distribution then we obtain the well–known inverse Gauss distribution. The most important properties of the IGT distribution are

1. $\mathrm{E}(X) = \mu$
2. If $X \sim \mathrm{IGT}(\mu, \lambda; g)$ then $cX \sim \mathrm{IGT}(c\mu, c\lambda; g)$ for $c > 0$, i.e., a IGT distribution belongs to a scale family.

For details and additional properties we refer to the original article [SLB].

We are interested in the IGT distribution with Student–t kernel because it fits several of our claim size distributions particularly well. The pdf of the Student–t distribution is

$$h(t) = \frac{\Gamma(\frac{\nu+1}{2})}{\sqrt{\nu\pi}\Gamma(\frac{\nu}{2})}\left(1 + \frac{t^2}{\nu}\right)^{-\frac{\nu+1}{2}}.$$

Thus the pdf of the IGT–St distribution is

$$f(t) = \frac{\Gamma(\frac{\nu+1}{2})}{\sqrt{\nu\pi}\Gamma(\frac{\nu}{2})}\left(1 + \frac{\lambda}{\nu\mu}\left(\frac{t}{\mu} + \frac{\mu}{t} - 2\right)\right)^{-\frac{\nu+1}{2}}\frac{\sqrt{\lambda}}{\sqrt{t^3}}.$$

We will allow any positive real number for ν and not only integers. For $t \to 0$ the pdf behaves like the power function

$$\frac{\Gamma(\frac{\nu+1}{2})}{\sqrt{\pi}\Gamma(\frac{\nu}{2})}\left(\frac{\nu}{\lambda}\right)^{\frac{\nu}{2}} \cdot t^{\frac{\nu}{2}-1}.$$

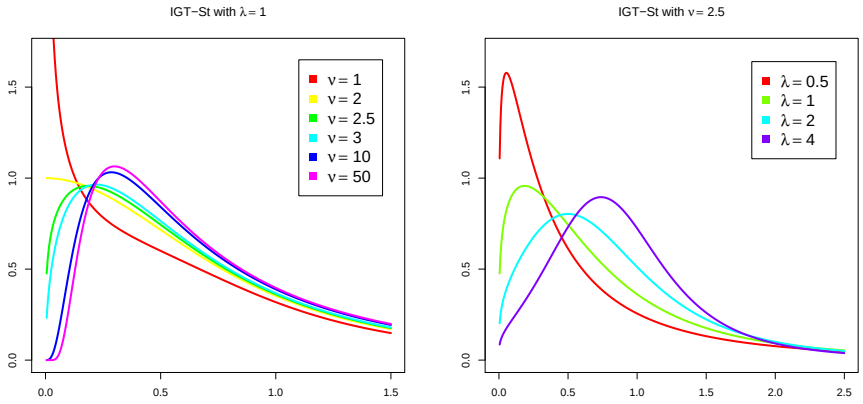
In particular, for $v < 2$ it tends to infinity, while for $v > 2$ it tends to 0. For $t \to \infty$ the pdf behaves like the power function

$$\frac{\Gamma(\frac{\nu+1}{2})}{\sqrt{\pi}\Gamma(\frac{\nu}{2})}\mu^{\nu+1}\left(\frac{\nu}{\lambda}\right)^{\frac{\nu}{2}} \cdot t^{-\frac{\nu}{2}-2}.$$

Thus the distribution is asymptotically a Pareto distribution and has a heavy tail with tail index $\alpha = \nu/2 + 1$, see Subsection 2.3.

The IGT distributions with normal, logistic, Laplace, and Student–t kernel have been implemented by Leiva, Hernández, and Sanhuezain in **R** as the **ig**

package [LHS]. Thus we can plot some graphs to get a better intuitive understanding of the IGT–St distribution. As a IGT–St distribution belongs to a scale family we may restrict ourselves to the case $\mu = 1$.



The graphs confirm that the parameter ν controls the head and tail behavior of the distribution. The shape parameter λ can be used to move and flatten the peak of the pdf.

Finally, we note that in the case of the IGT–St distribution the ML–estimation of the parameters μ, λ is robust, i.e., exceptional high or low observations influence the estimation only moderately [LHS, SLB]. The picture shows that the pdfs for large ν are very similar, thus the estimation of a high ν will not be robust. Luckily, in our applications we encounter only ν up to 3.5.

B The GBS Distributions

Díaz–García and Leiva–Sánchez generalized the Birnbaum–Saunders distribution in the following way [DL]: Let Z be a random variable with a standard symmetrical distribution, i.e., $\mathrm{E}(Z) = 0$, $\mathrm{Var}(Z) = 1$, and its pdf can be written as $h(t) = g(t^2)$, $t \in \mathbb{R}$. Then the generalized Birnbaum–Saunders distribution with parameters $\alpha, \beta > 0$ and kernel g is the random variable X with pdf

$$f(t) = h\left(\frac{1}{\alpha}\left(\sqrt{\frac{t}{\beta}} - \sqrt{\frac{\beta}{t}}\right)\right) \cdot \frac{1}{2\alpha\beta}\left(\sqrt{\frac{\beta}{t}} + \sqrt{\frac{\beta^3}{t^3}}\right) \qquad \text{for } t > 0.$$

We use the notation $X \sim \mathrm{GBS}(\alpha, \beta; g)$. If Z is the standard normal distribution we obtain the original Birnbaum–Saunders distribution. The most interesting properties of the GBS distributions are

1. β is the median.
2. If $X \sim \mathrm{GBS}(\alpha, \beta; g)$ then $cX \sim \mathrm{GBS}(\alpha, c\beta; g)$ for $c > 0$ and $1/X \sim \mathrm{GBS}(\alpha, 1/\beta; g)$, i.e., a GBS distribution belongs to a scale family, which is closed under reciprocation.
3. The cdf and quantiles of X can easily be computed in terms of the cdf and quantiles of Z.

For details and additional properties we refer to [DL, SLB2].

For our examination we are interested in the original BS distribution and the GBS with Student–t kernel. Their pdfs are

$$f(t) = \frac{1}{\sqrt{8\pi}\alpha\beta} \exp\left(-\frac{1}{2\alpha^2}\left(\frac{t}{\beta} + \frac{\beta}{t} - 2\right)\right)\left(\sqrt{\frac{\beta}{t}} + \sqrt{\frac{\beta^3}{t^3}}\right) \qquad \text{[BS]}$$

$$f(t) = \frac{\Gamma(\frac{\nu+1}{2})}{2\sqrt{\pi\nu}\Gamma(\frac{\nu}{2})\alpha\beta}\left(1 + \frac{\left(\frac{t}{\beta} + \frac{\beta}{t} - 2\right)}{\alpha^2\nu}\right)^{-\frac{\nu+1}{2}}\left(\sqrt{\frac{\beta}{t}} + \sqrt{\frac{\beta^3}{t^3}}\right) \qquad \text{[BS–St]}$$

and their means are

$$\frac{\beta}{2}\left(2 + \alpha^2\right) \quad \text{resp.} \quad \frac{\beta}{2}\left(2 + \frac{\nu}{\nu - 2}\alpha^2\right).$$

We will allow any positive real number for ν and not only integers.

The behavior of the BS pdf for $t \to 0$ as well as for $t \to \infty$ is dominated by the exponential part, which converges to 0. In particular, the BS distribution has no heavy tail. The pdf of the GBS–St distribution behaves for $t \to 0$ like the power function

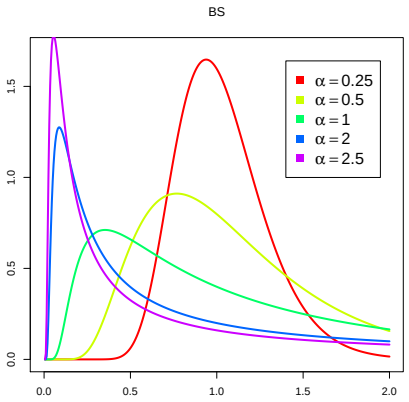
$$\frac{\Gamma(\frac{\nu+1}{2})}{2\sqrt{\pi}\Gamma(\frac{\nu}{2})}\frac{\nu^{\frac{\nu}{2}}\alpha^\nu}{\beta^{\frac{\nu}{2}}} \cdot t^{\frac{\nu}{2}-1}.$$

Hence, like the IGT–St distribution it tends to infinity for $v < 2$, while for $v > 2$ it tends to 0. For $t \to \infty$ the pdf behaves like the power function

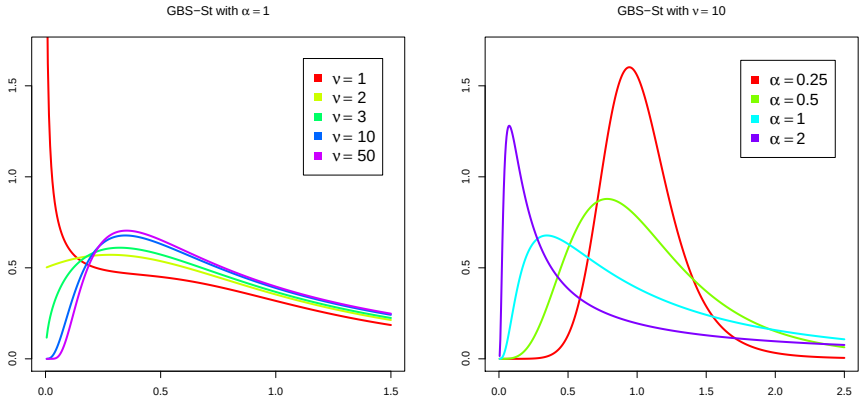
$$\frac{\Gamma(\frac{\nu+1}{2})}{2\sqrt{\pi}\Gamma(\frac{\nu}{2})}\alpha^\nu\beta^{\frac{\nu}{2}}\nu^{\frac{\nu}{2}} \cdot t^{-\frac{\nu}{2}-1}.$$

Thus the distribution is asymptotically a Pareto distribution and has a heavy tail with tail index $\alpha = \nu/2$.

Barros, Paula, and Leiva implemented the GBS distributions with normal, logistic, Laplace, and Student–t kernel in **R** as the **gbs** package [BPL]. We plot some graphs of the original BS and the GBS–St distribution. As a GBS distribution belongs to a scale family we may restrict ourselves to the case scale parameter $\beta = 1$.



In the case of the BS distribution the shape parameter can be used to move and flatten the peak of the pdf. We note that for $\alpha < 1$ the pdf has a very flat head. This is atypical for a claim size distribution, thus we expect $\alpha > 1$ in applications.



For the GBS–St distribution the graphs confirm that the parameter ν controls the head and tail behavior of the distribution. We note that the pdfs for $v = 10$ and $\nu = 50$ are very similar. This indicates that that ML–estimation of a large parameter ν is not robust. The shape parameter α can be used to move and flatten the peak of the pdf. We note the flat head of the pdf for $\alpha \ll 1$ and $\nu \gg 1$ due to the term α^ν in the pdf. As remarked in the case of the BS distribution, this is atypical for a claim size distribution, and we will not see this combination of parameters in applications.

Finally, we note that the ML–estimation of the parameters α, β is robuster for the GBS–St than for the GB distribution [BPL, SLB2].

References

[BPL] Barros, M., G. Paula, and V. Leiva: *An R implementation for generalized Birnbaum–Saunders distributions.* Comput. Stat. Data Anal. **53** (2009), 1511–1528.

[CDL] Cebrián, A., M. Denuit, and Ph. Lambert: *Generalized Pareto Fit to the Society of Actuaries' Large Claims Database.* North American Actuarial Journal **7** (2003), 18–36.

[DL] Díaz–García, J., and V. Leiva–Sánchez: *A new family of life distributions based of elliptically contoured distributions.* J. Statist. Plann. Inference **128** (2005), 445–457. Erratum **137** (2007), 1512–1513.

[KalV] *Verordnung über die versicherungsmathematischen Methoden zur Prämienkalkulation und zur Berechnung der Alterungsrückstellung in der privaten Krankenversicherung (Kalkulationsverordnung — KalV)*, http://www.gesetze-im-internet.de/kalv/BJNR178300996.html (2009).

[LHS] Leiva, V., H. Hernández, and A. Sanhueza: *An R Package for a General Class of Inverse Gaussian Distributions.* J. Statistical Software **26** (2008).

[MFE] McNeil, A., R. Frey, and P. Embrechts. *Quantitative Risk Management: Concepts, Techniques and Tools.* Princeton University Press, Princeton (2005).

[SLB] Sanhueza, A., V. Leiva, and N. Balakrishnan: *A new class of inverse Gaussian type distributions.* Metrika **68** (2008), 31–49.

[SLB2] Sanhueza, A., V. Leiva, and N. Balakrishnan: *The Generalized Birnbaum–Sauders Distribution and Its Theory, Methodology, and Application.* Commun. Sat. Theor. Meth. **37** (2008), 645–670.

ERGO Versicherungsgruppe AG
Aachener Strasse 300
50933 Köln
Germany