EDA – Part II

MF9130E – Introductory Course in Statistics
08.05.2023

Chi Zhang
chi.zhang@medisin.uio.no
Oslo Center for Biostatistics and Epidemiology
Department of Biostatistics, UiO

Outline

8:30-9:15	Exploratory data analysis II
9:30-10:15	Transformations, non-parametric tests
Demontration & Practice	**Demonstration in R**

Lab notes for today:
(under *R Lab and Code* tab)

EDA II

Non-parametric tests

Exploratory vs Confirmatory

Confirmatory data analysis
- focus on inference: hypothesis testing
- parameter estimation, uncertainty
- model selection

(Table 2, 3…)

Initial data analysis (**exploratory**)
- describe data and collection procedures
- scrutinise data for **errors, outliers**, missing
- check **assumptions** needed for confirmatory analysis to hold

(Table 1, or none)

John Tukey (1977): "Too much emphasis in statistics was placed on **hypothesis testing**; more emphasis needed to suggest **what hypothesis** to test"

Hypothesis doesn't generate by itself; also, your data won't be perfect - EDA helps!

Exploratory data analysis

Check data type and quality
- **coding**: e.g. are numerical numbers coded as a character?
- does data fall within a **plausible range**? (e.g. weight, height)
- are there too many **missing** data? Missing at random or not?

Check whether **assumptions** hold
- for example, does a continuous variable look ‘normal’?
- do linear relationships hold?

If not, consider alternative methods.

Note: not all data *should* be normal (e.g. time data, only positive and often skewed)

Dataset: length of hospital stay (liggetid)

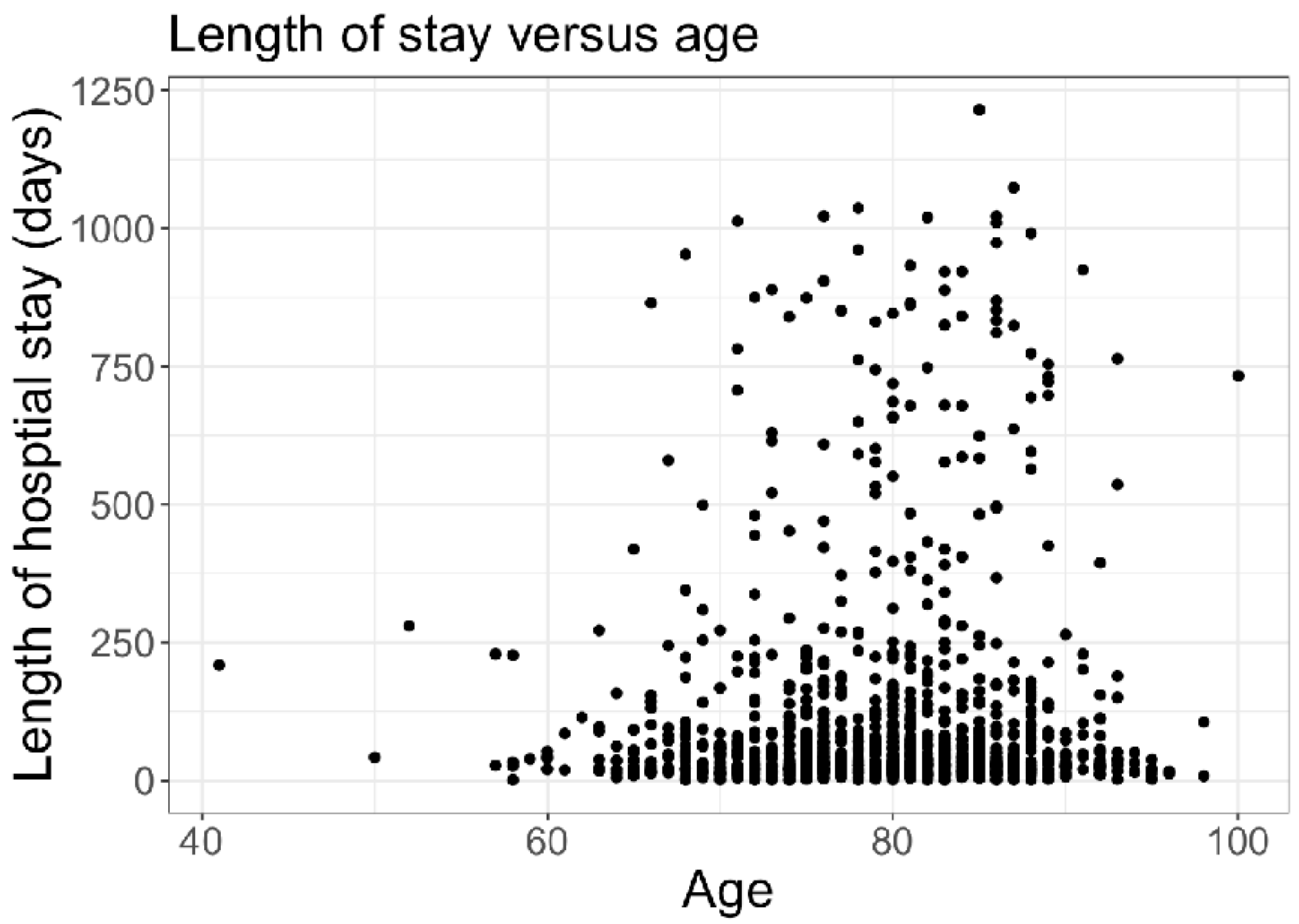
Data collected at Ullevål hospital; 1139 observations, 21 variables
Main outcome of interest: **length of hosptal stay** (liggetid)
Other variables to look at: **year of admission, age, gender, type of admission, stroke**

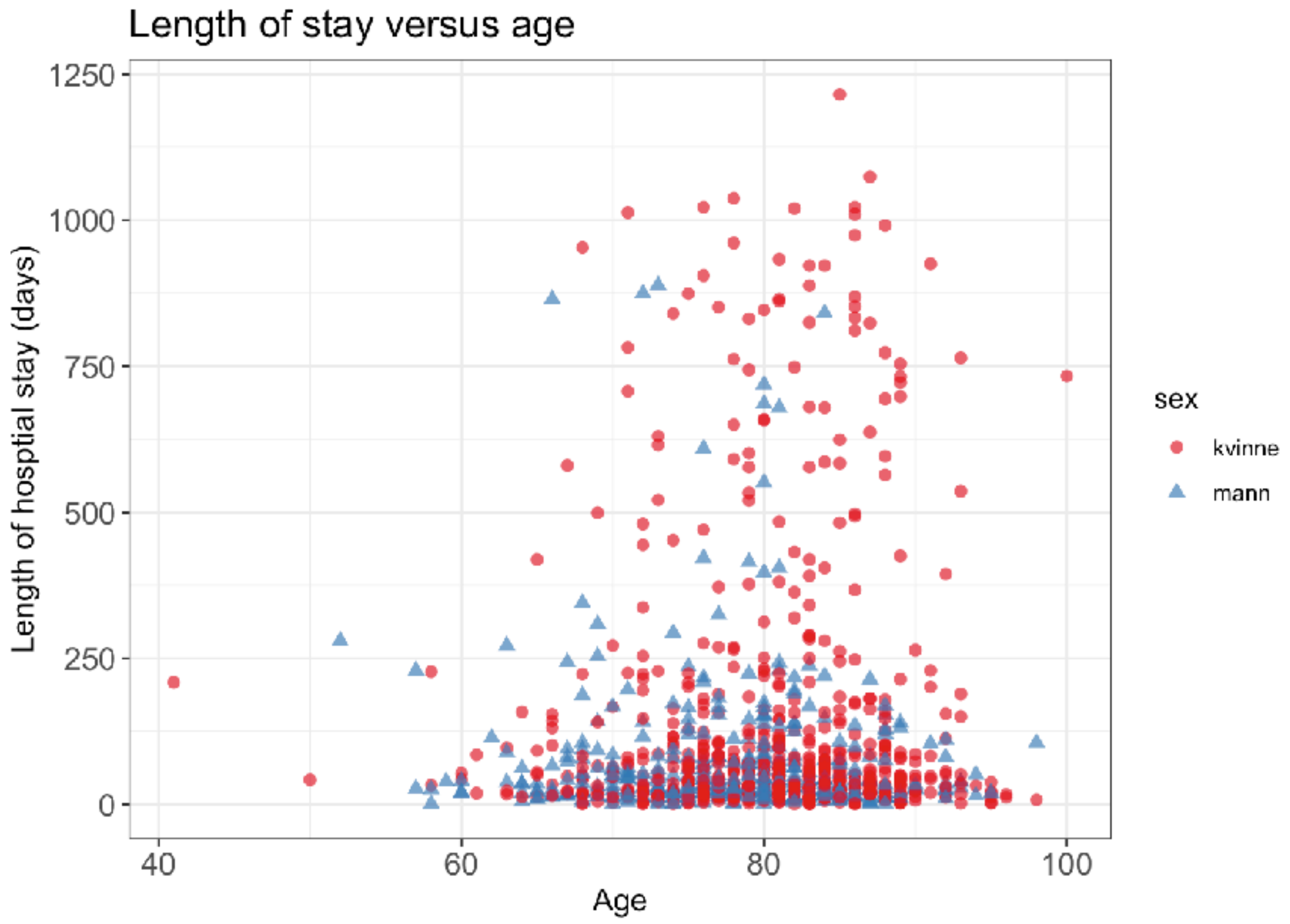
	faar	fmaan	fdag	innaar	innmaan	inndag	utaar	utmaan	utdag	kjoenn	kom_fra	slag	alder	liggetid	nliggti	kom_fra2
327	1909	1	2[illegible]	1985	11	28	85	12	20	mann	2	0	76	22	3.0910425	1
328	1908	1	1[illegible]	1985	12	3	85	12	17	kvinne	3	0	77	14	2.6390573	0
329	1902	9	[illegible]	1985	12	4	85	12	17	kvinne	1	0	83	13	2.5649494	0
330	1900	1	[illegible]	1985	12	5	85	12	16	mann	1	0	85	11	2.3978953	0
331	1911	12	2[illegible]	1985	12	12	87	3	9	kvinne	2	0	74	452	6.1136822	1
332	1901	9	2[illegible]	1985	12	13	87	1	22	kvinne	2	0	84	405	6.0038871	1
333	1908	7	2[illegible]	1985	12	13	85	12	19	kvinne	2	0	77	6	1.7917595	1
334	1906	4	[illegible]	1985	12	20	87	1	1	kvinne	2	0	79	377	5.9322452	1
335	1893	2	1[illegible]	1985	12	23	87	1	21	kvinne	1	0	92	394	5.9763509	0
336	1899	10	[illegible]	1986	1	3	86	1	29	mann	4	NA	87	26	3.2580965	0
337	1893	12	2[illegible]	1986	1	3	86	9	3	kvinne	2	NA	93	243	5.4930614	1
338	1911	4	[illegible]	1986	1	3	86	2	3	kvinne	1	NA	75	31	3.4339872	0
339	1908	8	2[illegible]	1986	1	3	86	3	19	kvinne	1	NA	78	75	4.3174881	0
340	1906	12	1[illegible]	1986	1	7	86	4	2	kvinne	6	NA	80	85	4.4426513	0
341	1913	10	2[illegible]	1986	1	9	86	4	7	kvinne	2	NA	73	88	4.4773368	1
342	1908	3	2[illegible]	1986	1	9	86	3	14	kvinne	2	NA	78	64	4.1588831	1
343	1905	9	1[illegible]	1986	1	10	86	2	12	kvinne	5	NA	81	33	3.4965076	0
344	1903	6	2[illegible]	1986	1	10	86	1	17	mann	4	NA	83	7	1.9459101	0
345	1895	2	1[illegible]	1986	1	14	86	1	17	kvinne	1	NA	91	3	1.0986123	0
346	1907	2	[illegible]	1986	1	14	86	4	1	kvinne	2	NA	79	77	4.3438054	1
347	1909	9	[illegible]	1986	1	15	86	3	5	kvinne	2	NA	77	49	3.8918203	1

Explore: age vs los (length of stay)

Examine the relationship between age and length of stay, via scatter plot.

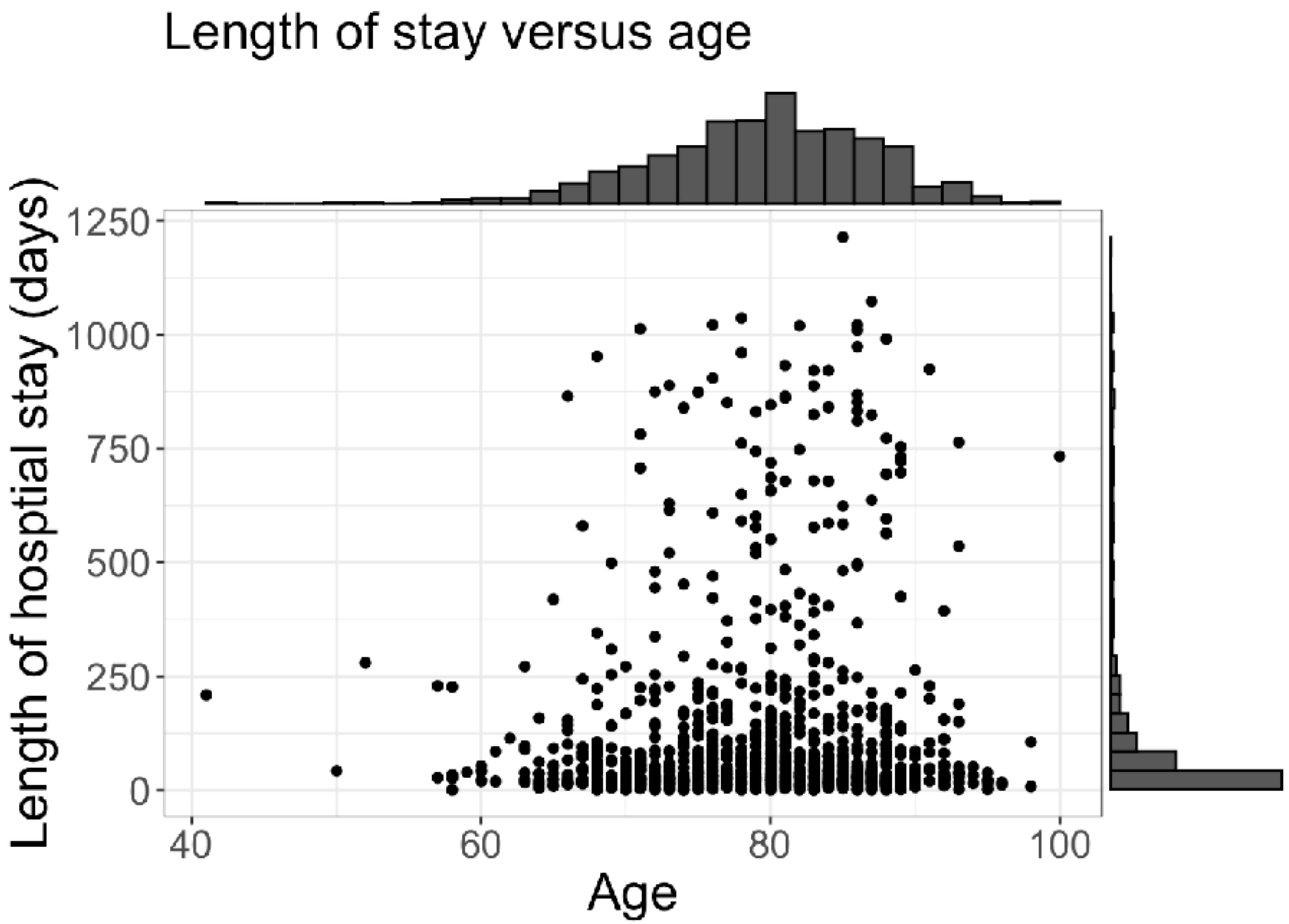
Can we add more information to the plot, such as gender?



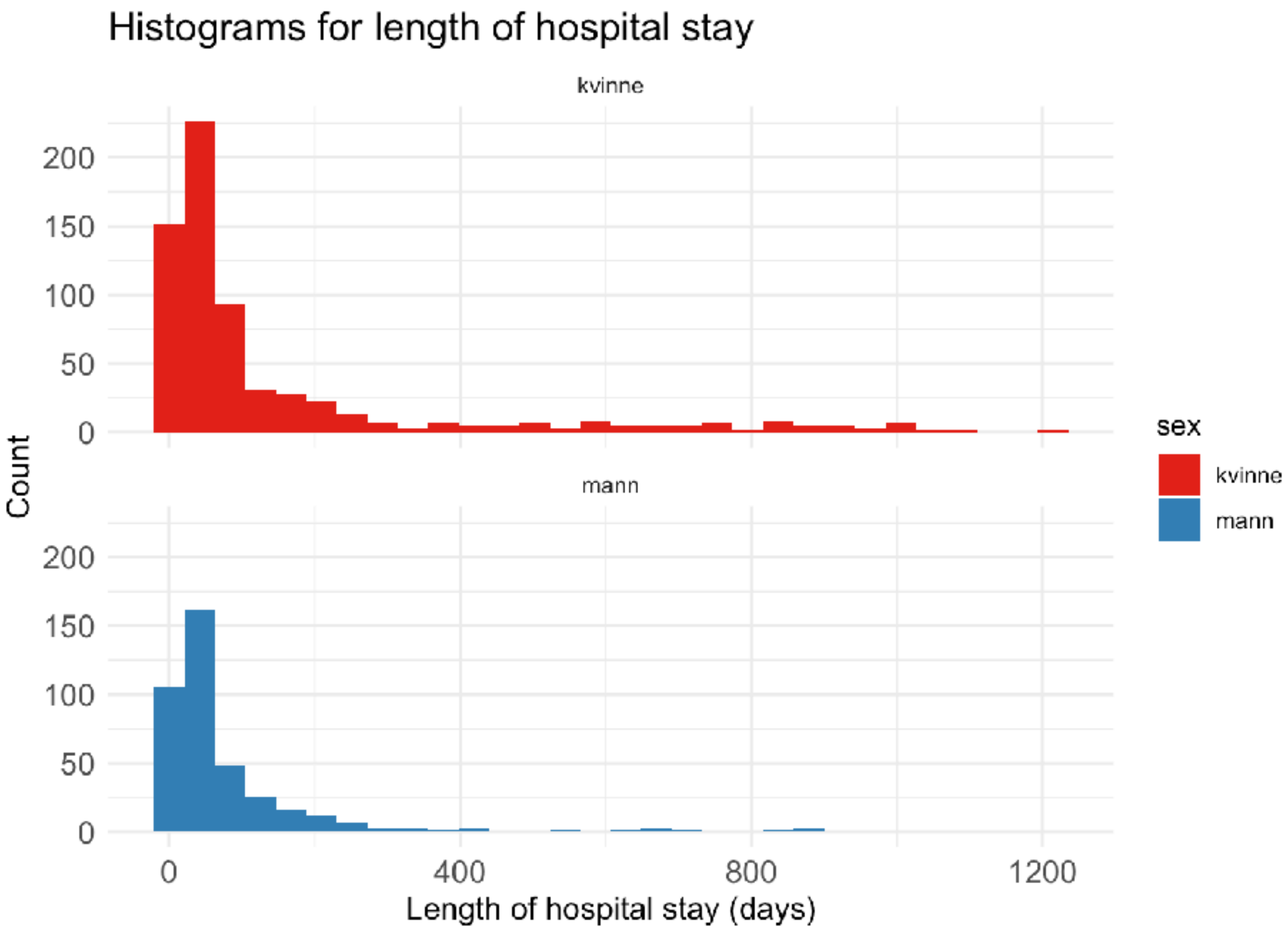


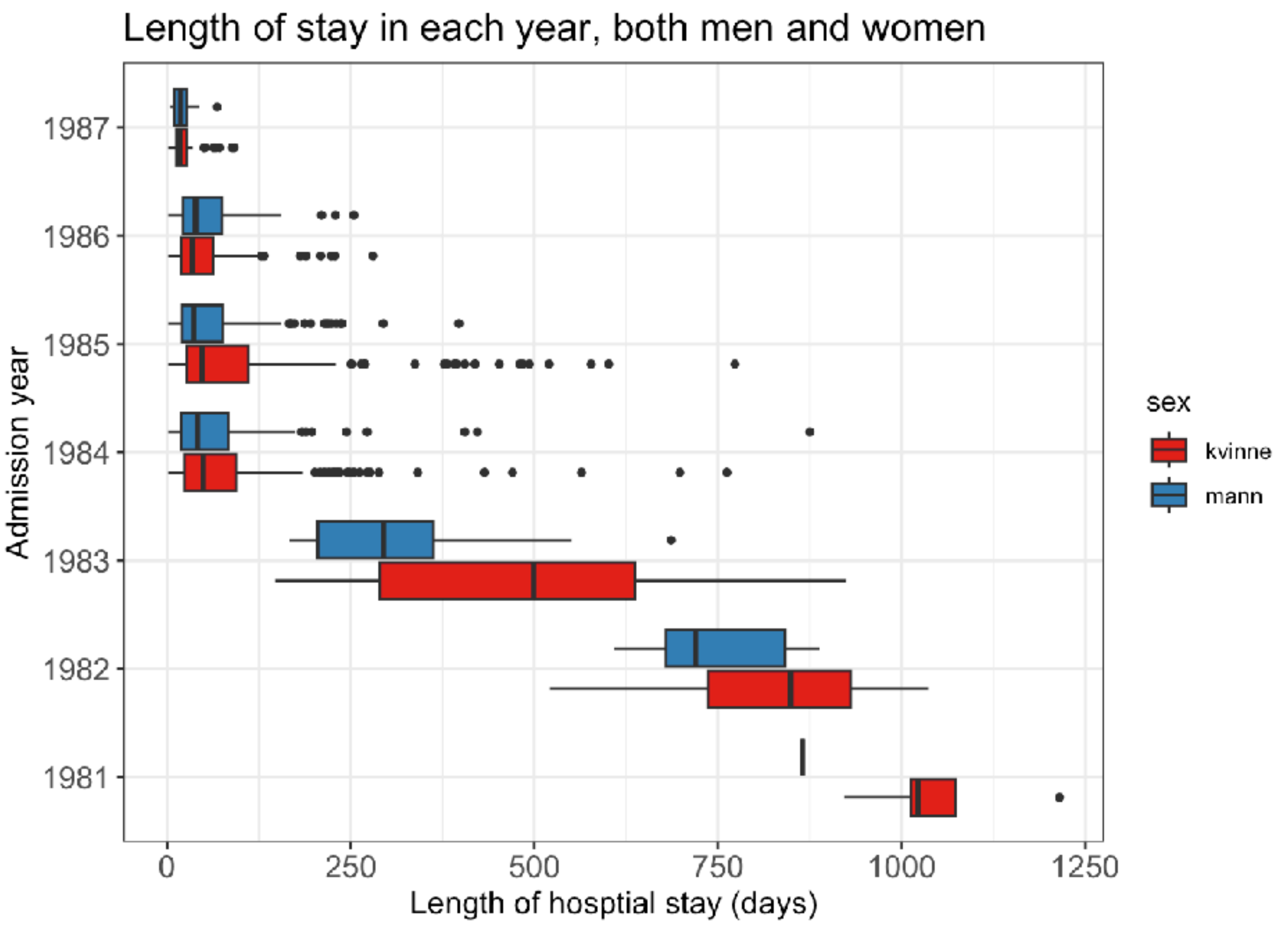
Explore: age vs los (length of stay)

Add histogram on top of the scatter plot: we can see the distribution for each variable.
Length of stay is not normally distributed! (As is often the case with 'time' data)

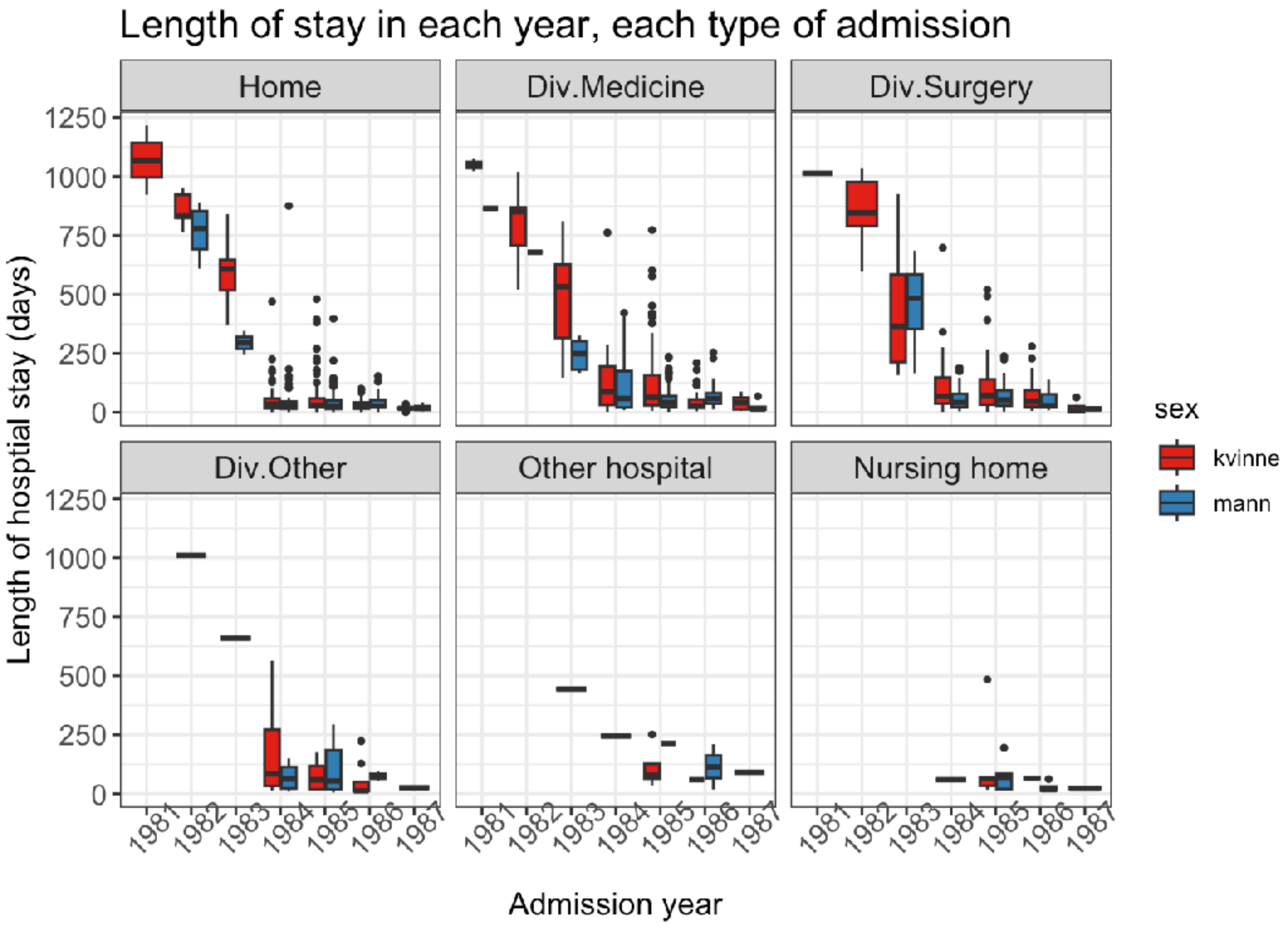


Explore: los, gender, year





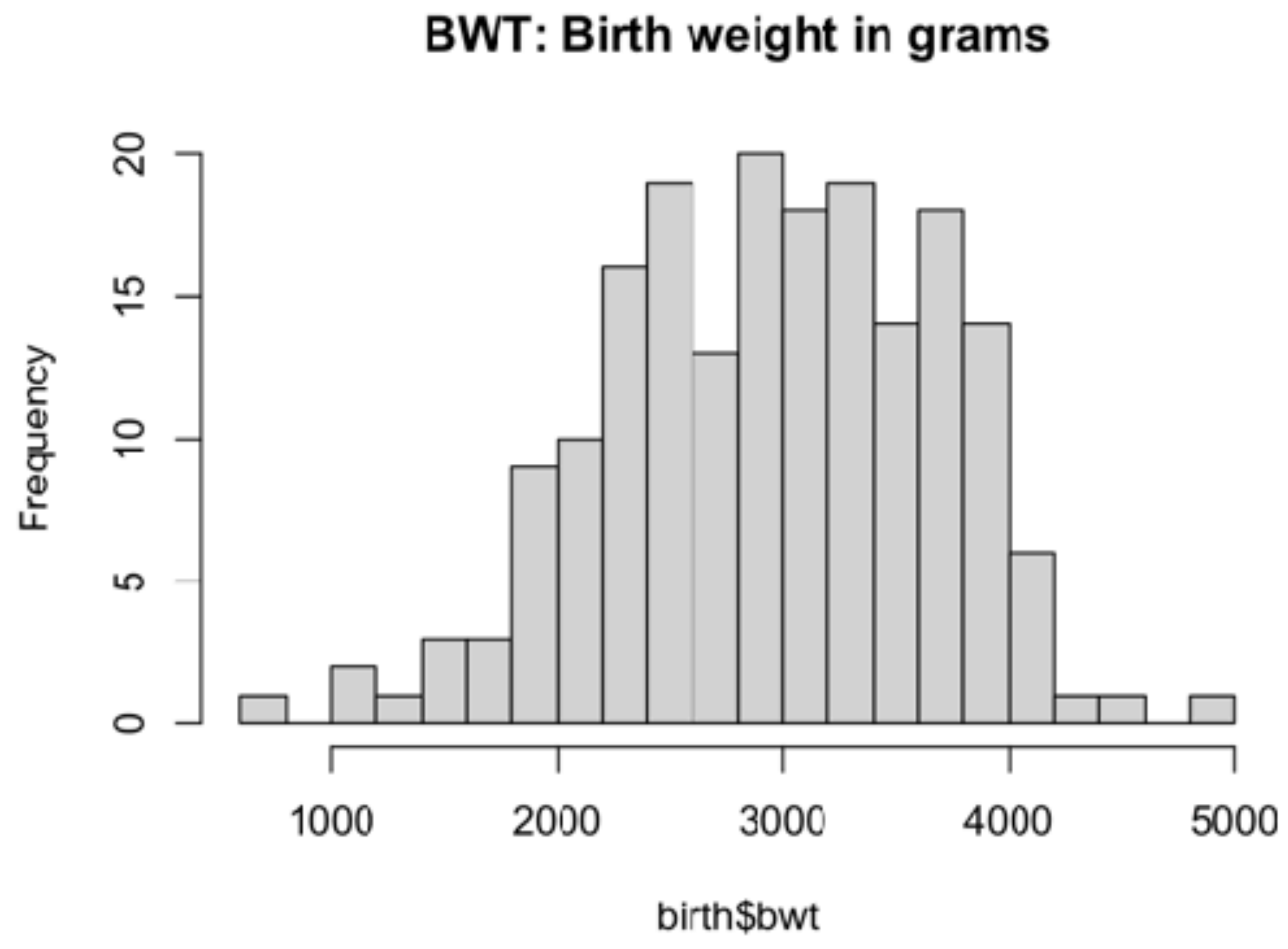
Explore: los, gender, year, type admission

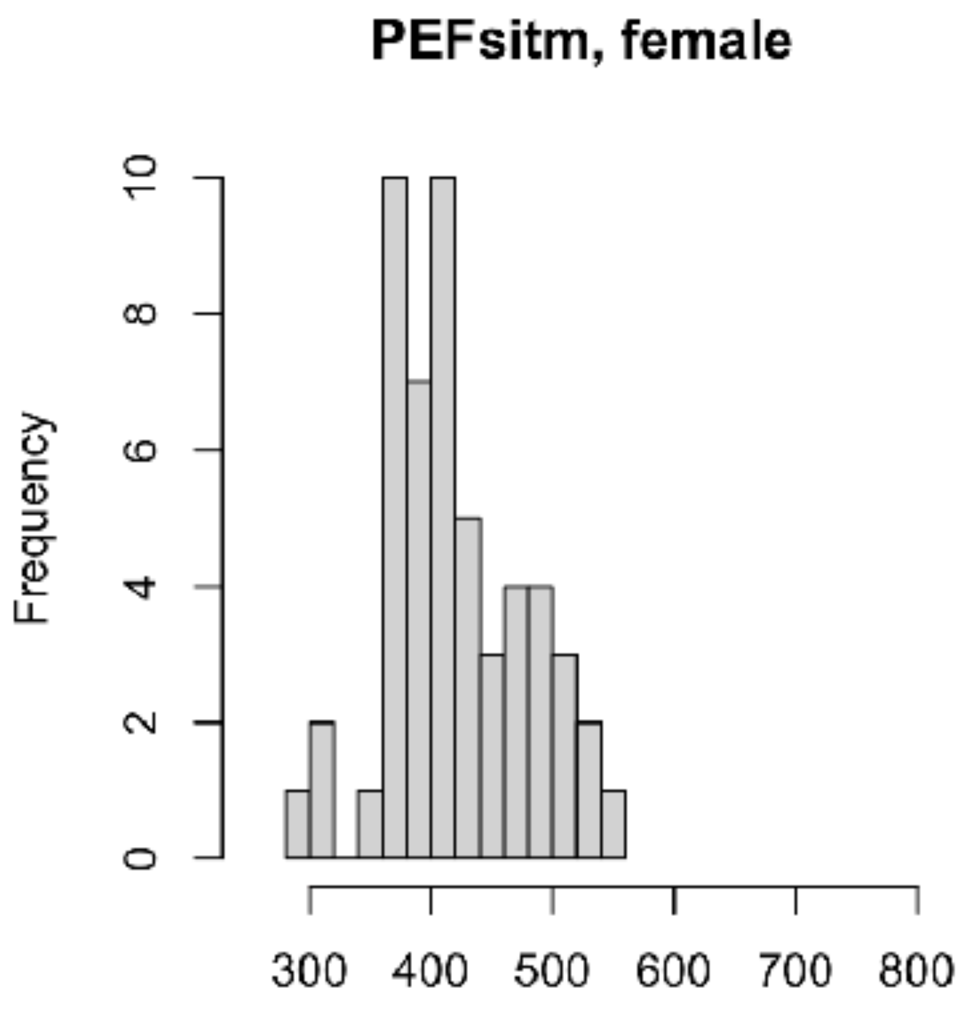


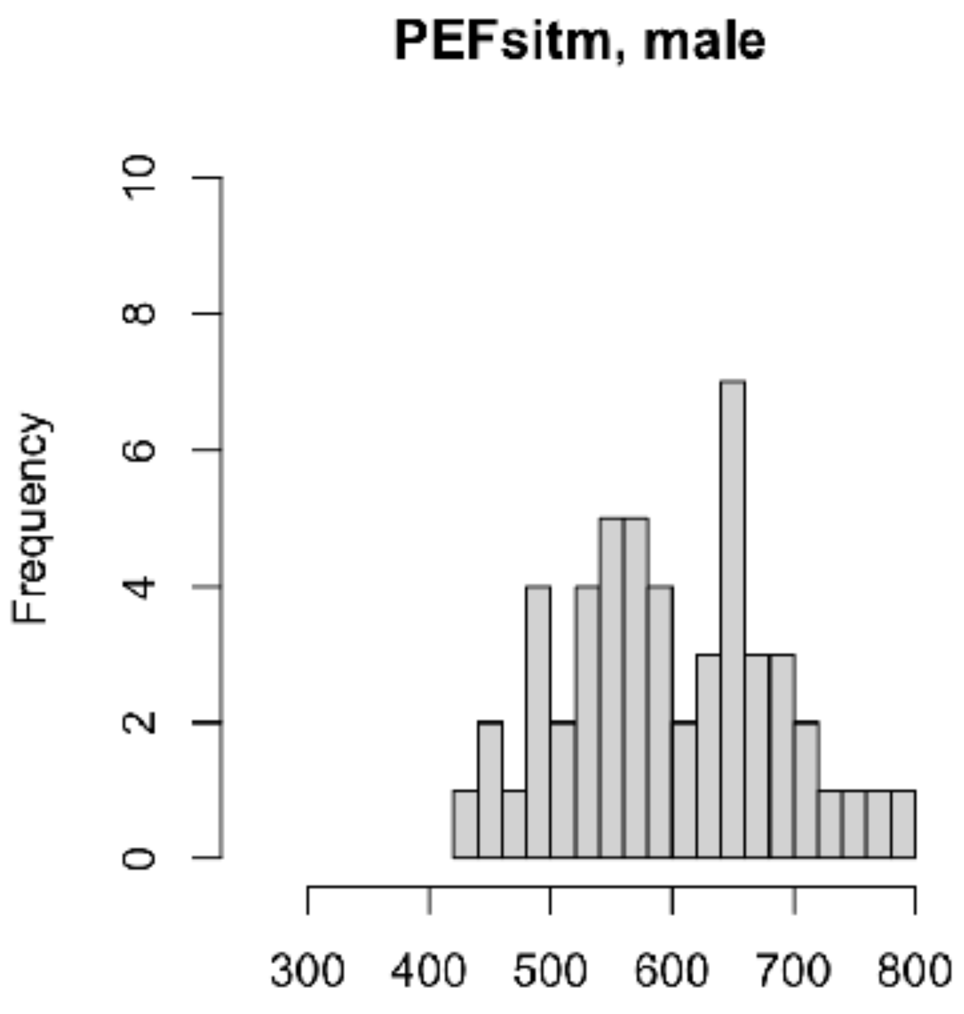
Non–normal data is common

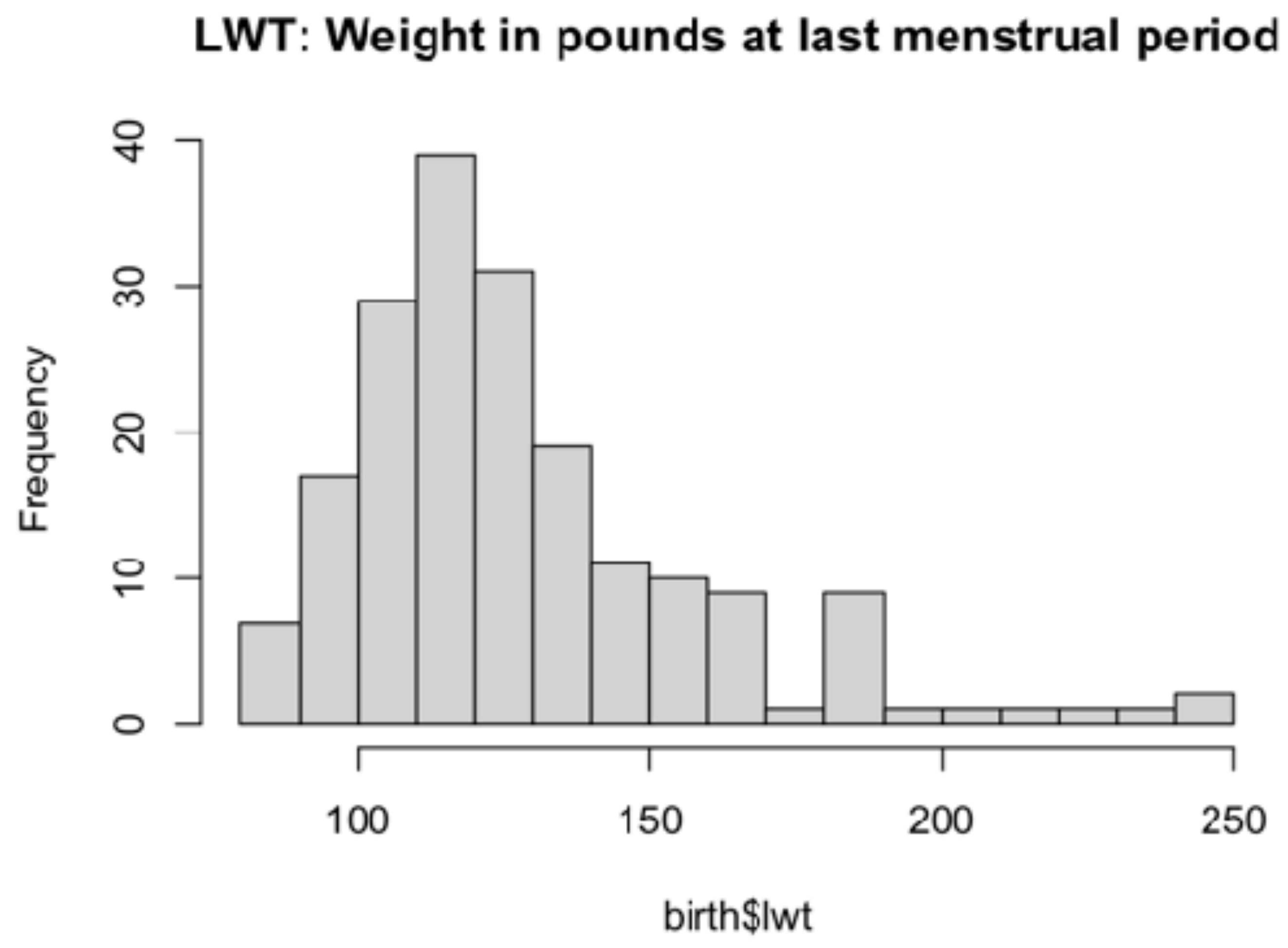
Example: 2 variables from birth data (right)
- BWT: approximately symmetrical
- LWT: not symmetrical - ‘right skewed’

Example: PEF data (below)
- PEF sitting mean for 2 genders look different

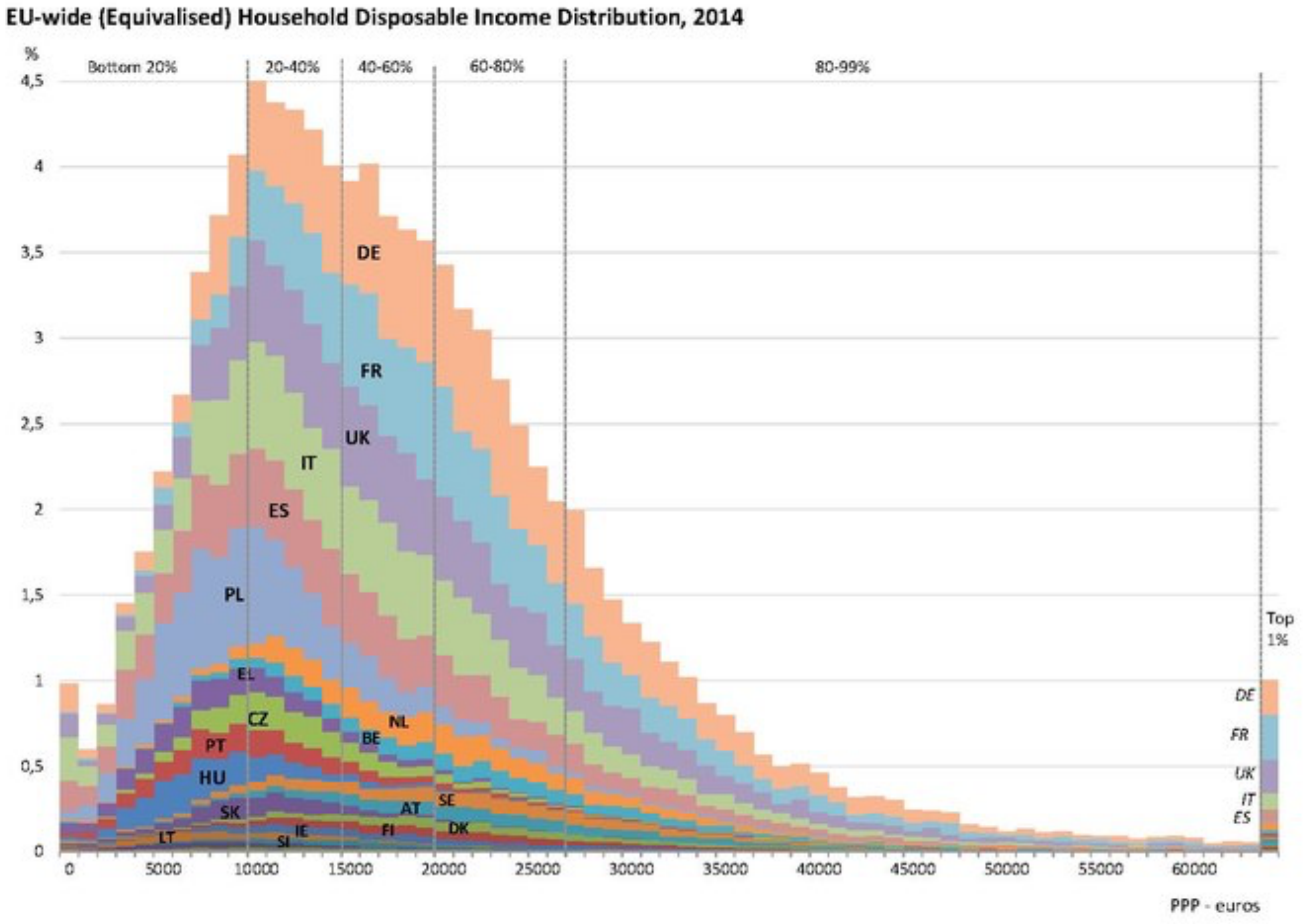




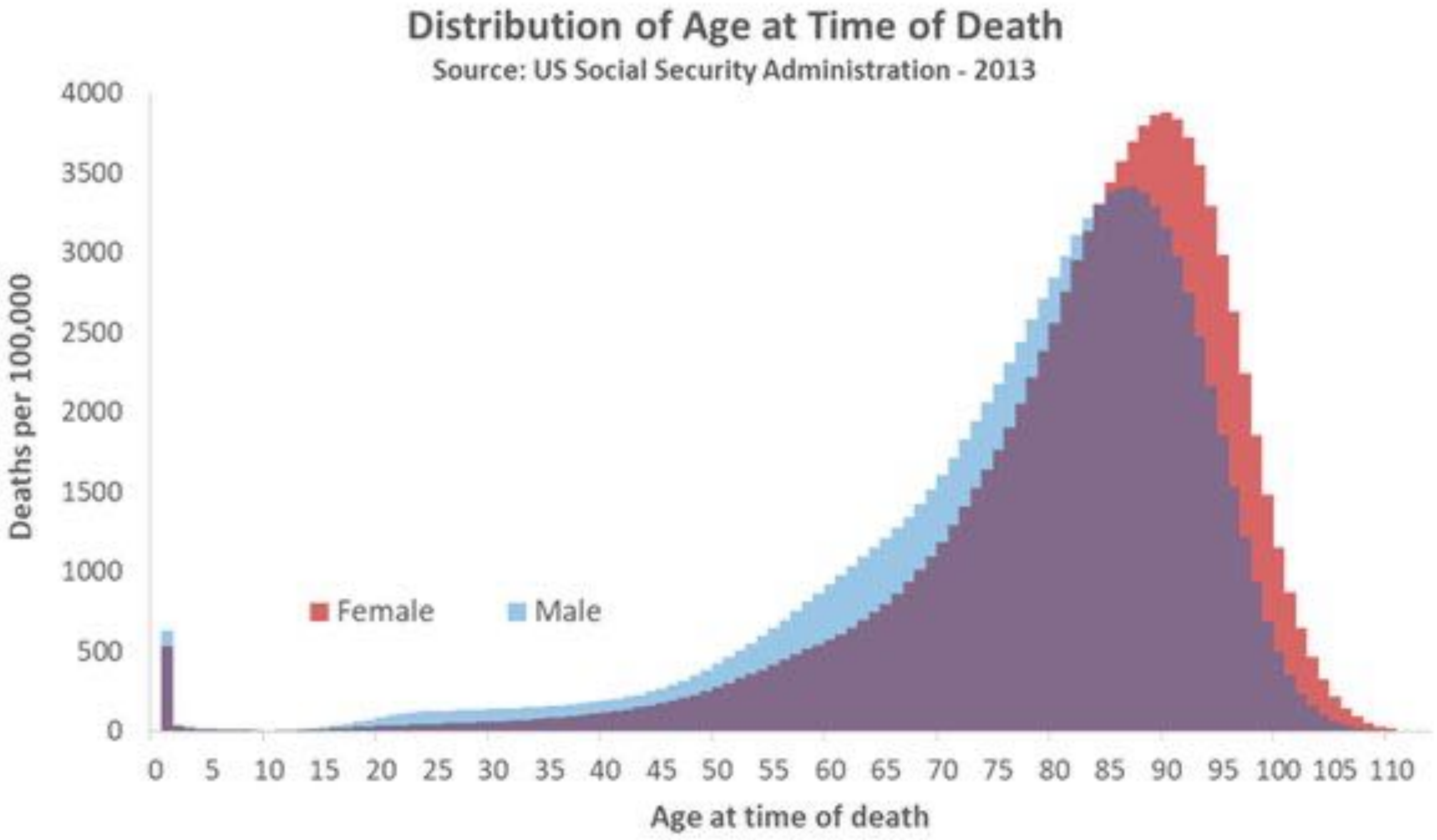




Non-normal data is common



Data for the EU aggregate excludes Bulgaria, Croatia, Malta and Romania.
Source: EU-SILC.



Q–Q plot

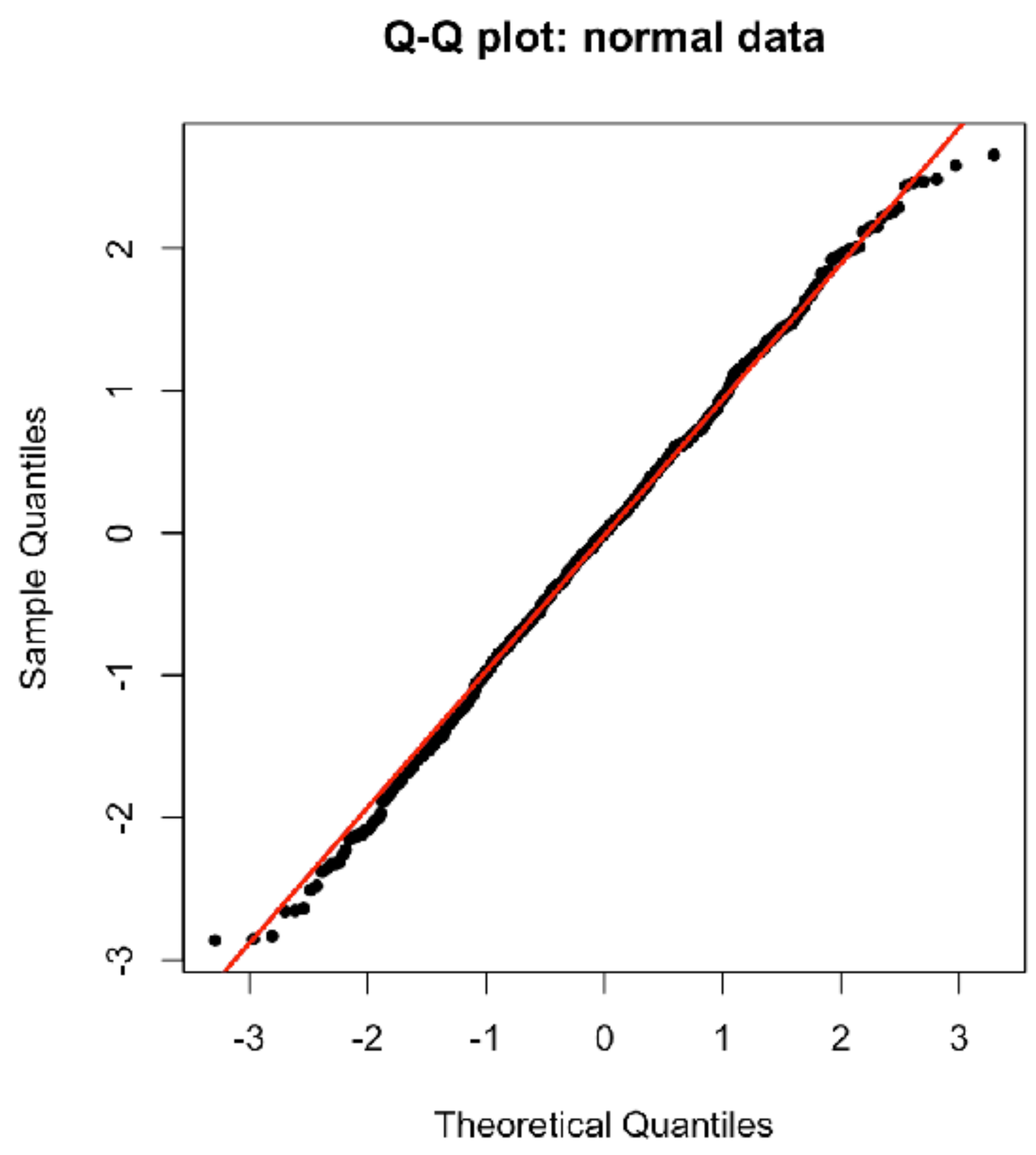
Q-Q (quantile-quantile) plots: graphical way of comparing two distributions

When checking normality, we plot
- the quantiles of the **observed** data
- against the quantiles of the corresponding normal distribution (**theoretical**)

If two distributions are identical, their quantiles are the same;

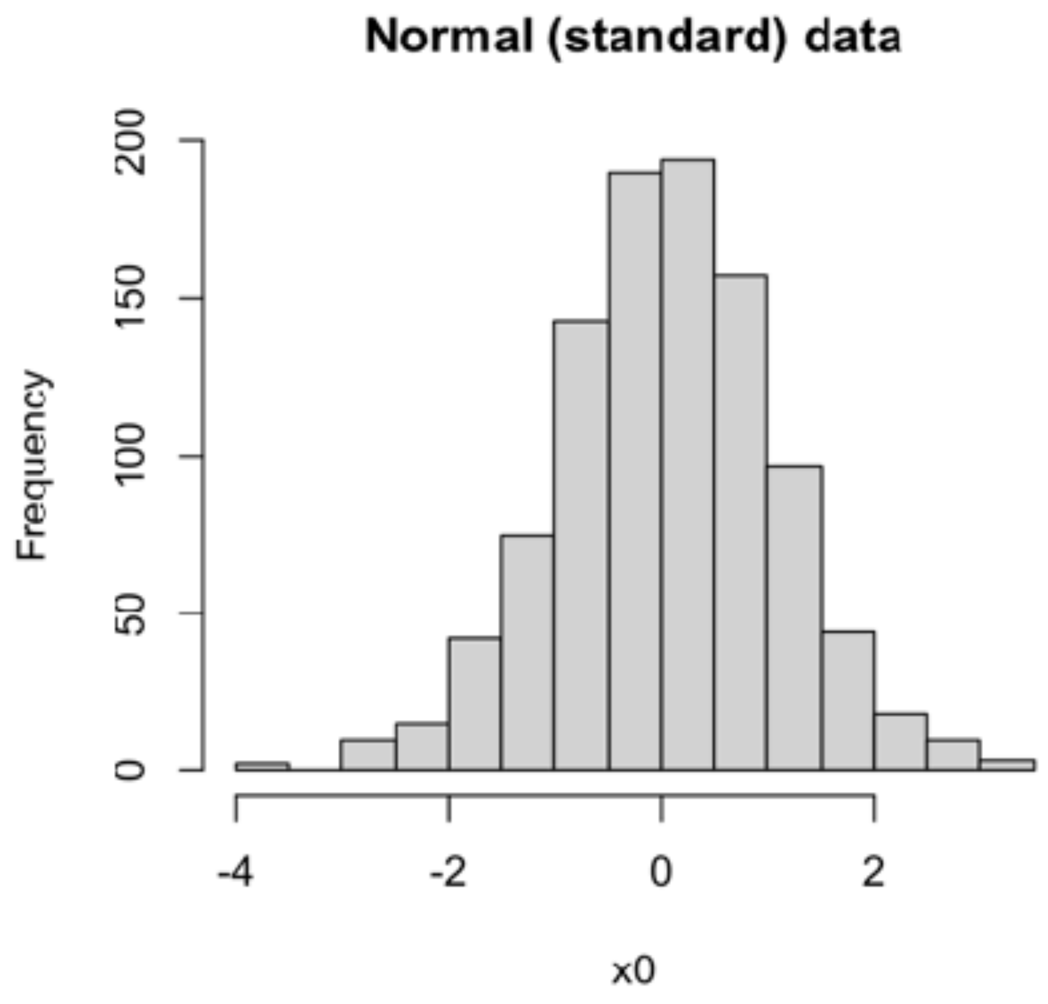
Q-Q plot should follow a straight 45 degree line

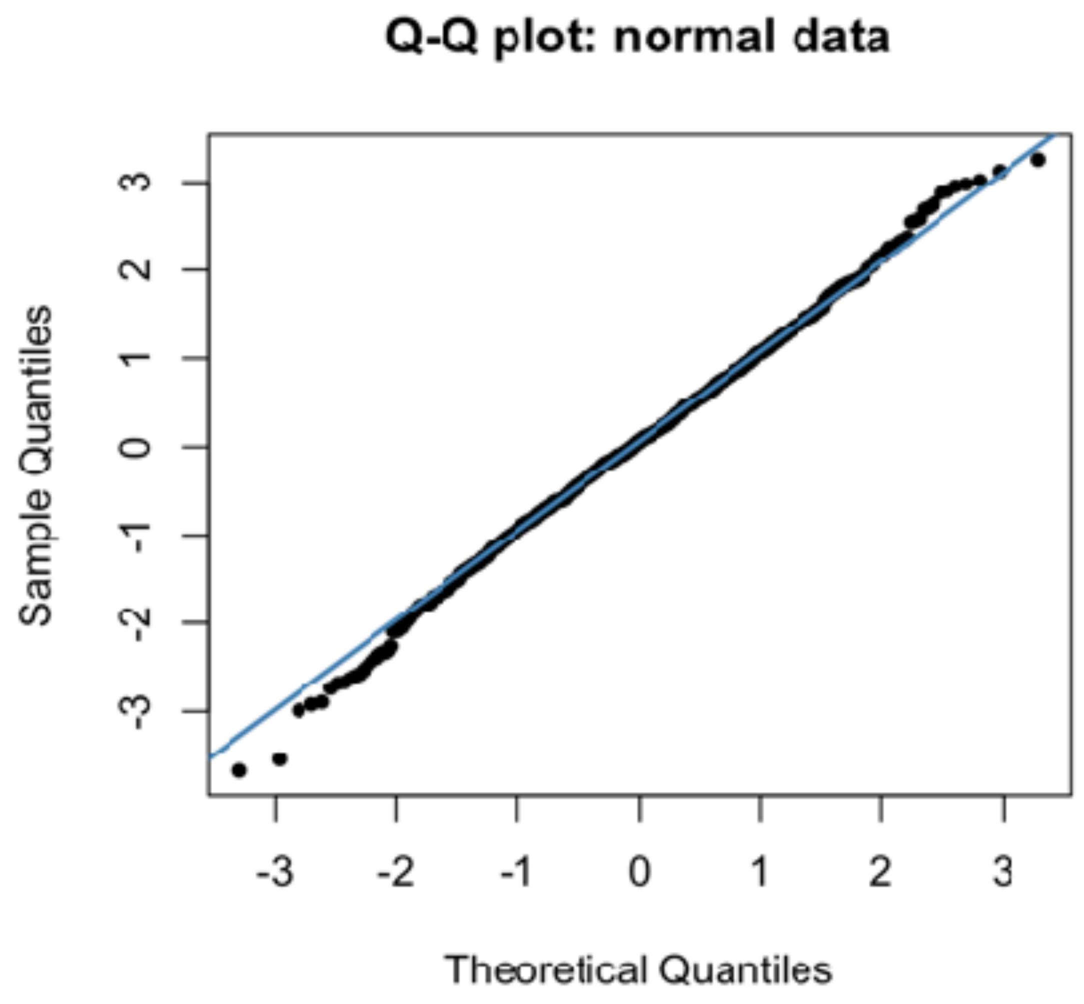
For ‘simulated’ normal data, there are some small deviation from the line; not a problem.



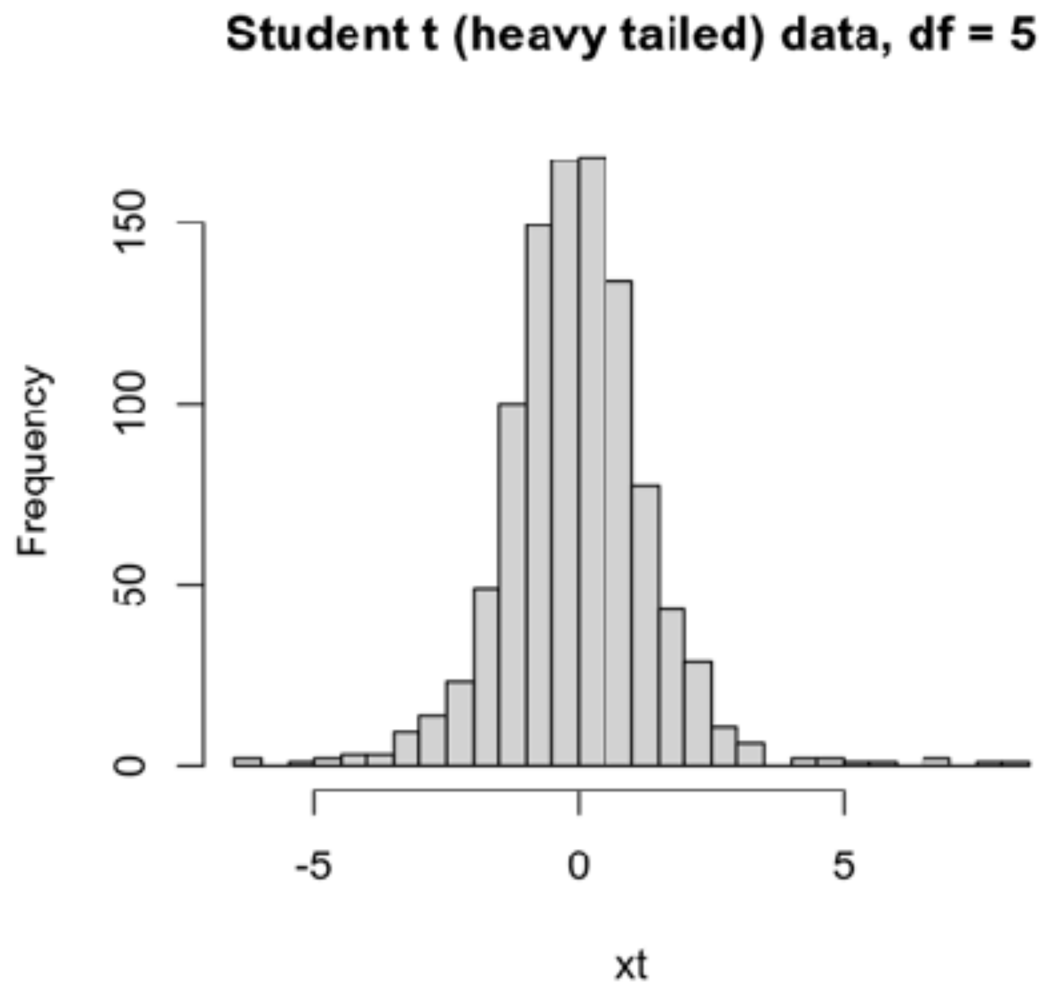
Can also carry out statistical tests for normality; but QQ plot is usually sufficient.
Kolmogorov-Smirnov test, Shapiro-Wilk test

Q–Q plot





Standard normal data N(0,1);
most points on 45 degree line in QQplot.



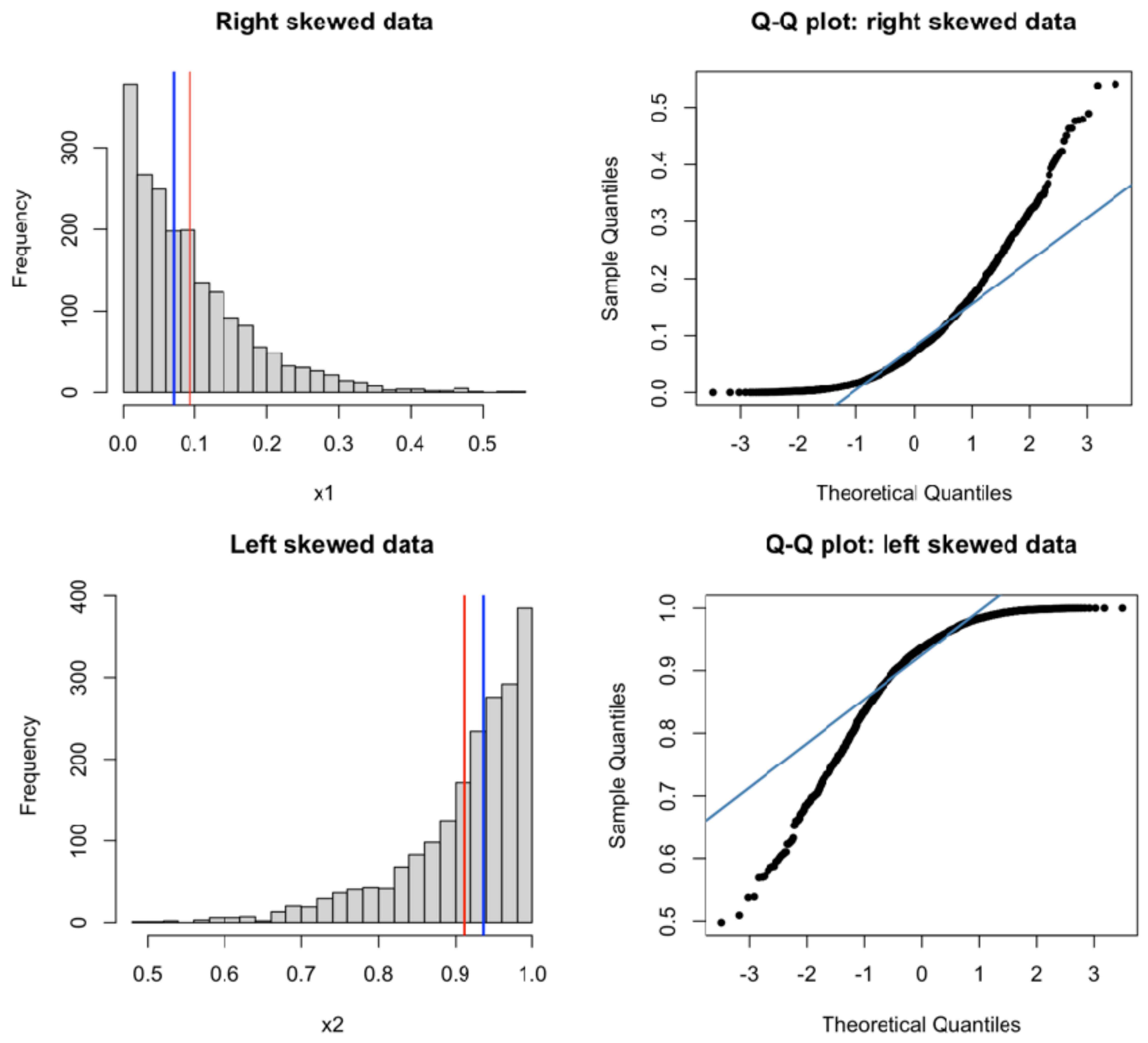
T-distributed data (df=5);

heavier tail (i.e. more points far from the average 0 compared to normal)

Points deviate from the line more strongly on QQplot.

Compare the quantiles for N(0,1) and t(5):
- 0.975: 1.96 vs 2.57
- 0.999: 3.09 vs 5.89

Q–Q plot



Right-skewed data (positively skewed)

longer tail to the right

Examples:
- **time**: can not be negative, but no upper bound
- **income**: (most people earn much less than a few rich people) - hence ‘median income’ is often used

Median < mean

Left-skewed data (negatively skewed)

longer tail to the left

Less common than right-skewed

Examples:
- age of death (there is an upper-bound)

Median > mean

Real data is messy

Real data is imperfect - doesn’t *necessarily* mean there is something wrong with the data; but need to choose methods carefully.

For one variable (univariate data), we have seen non-normal data:
- left or right skewed
- heavy tails (e.g. outliers)

If you use methods based on **normality and symmetry** (e.g +-1.96 s.e. for 95% CI), the results are wrong.

When you have more variables,
- the relationship between 2 variables might be **non-linear**

You might need to
- transform your data (apply a non-linear function), e.g. log transformation
- rank-based non-parametric methods
- bootstrap (resampling) to get a confidence interval
- generalized linear models (logistic regression rather than linear, etc)