STAT 401A - Statistical Methods for Research Workers

Modeling assumptions

Jarad Niemi (Dr. J)

Iowa State University

last updated: September 9, 2014

Normality assumptions

In the paired t-test, we assume

$$D_i \overset{iid}{\sim} N(\mu, \sigma^2).$$

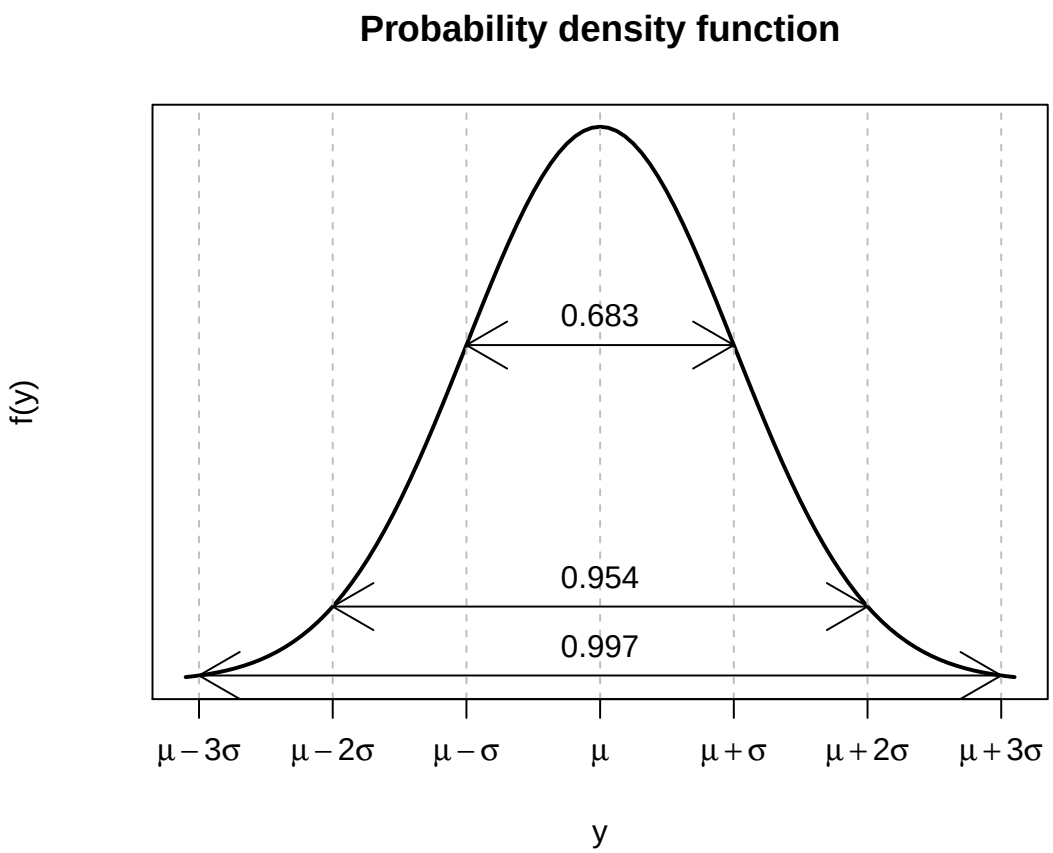
In the two-sample t-test, we assume

$$Y_{ij} \overset{ind}{\sim} N(\mu_j, \sigma^2).$$

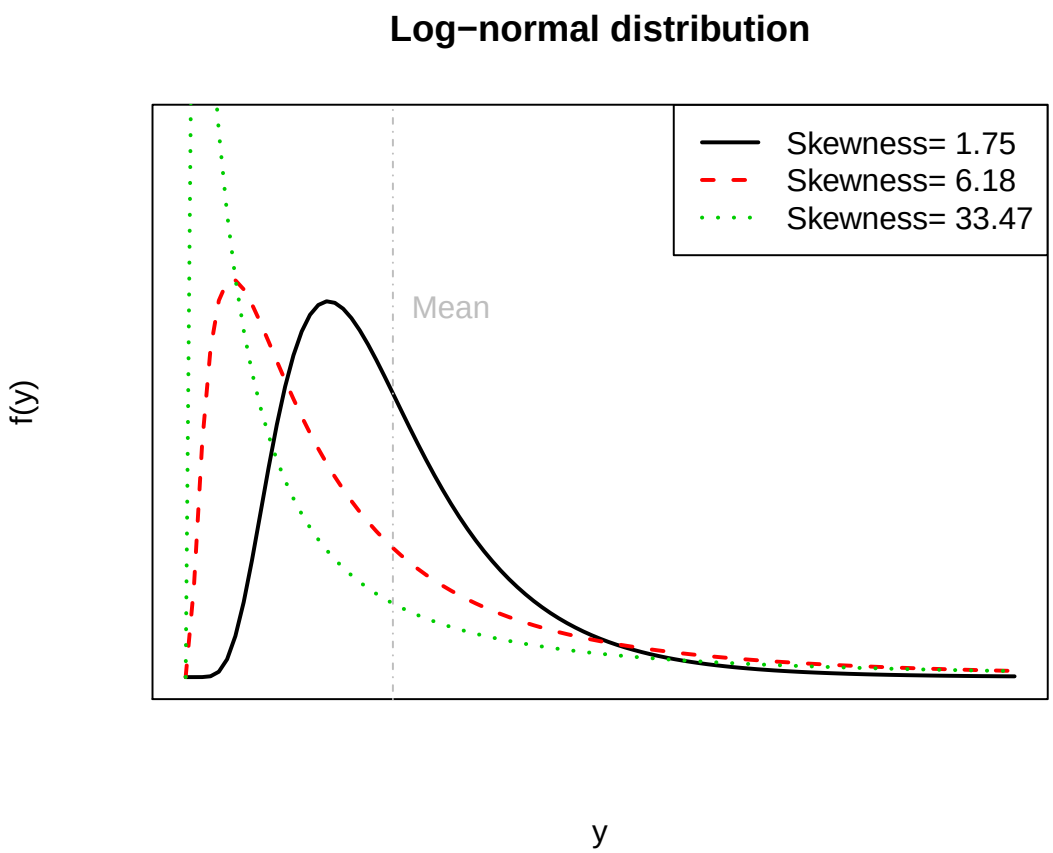
Key features of the normal distribution:

- Centered at the mean (expectation) μ
- Standard deviation describes the spread
- Symmetric around μ (no skewness)
- Non-heavy tails, i.e. outliers are rare (no kurtosis)

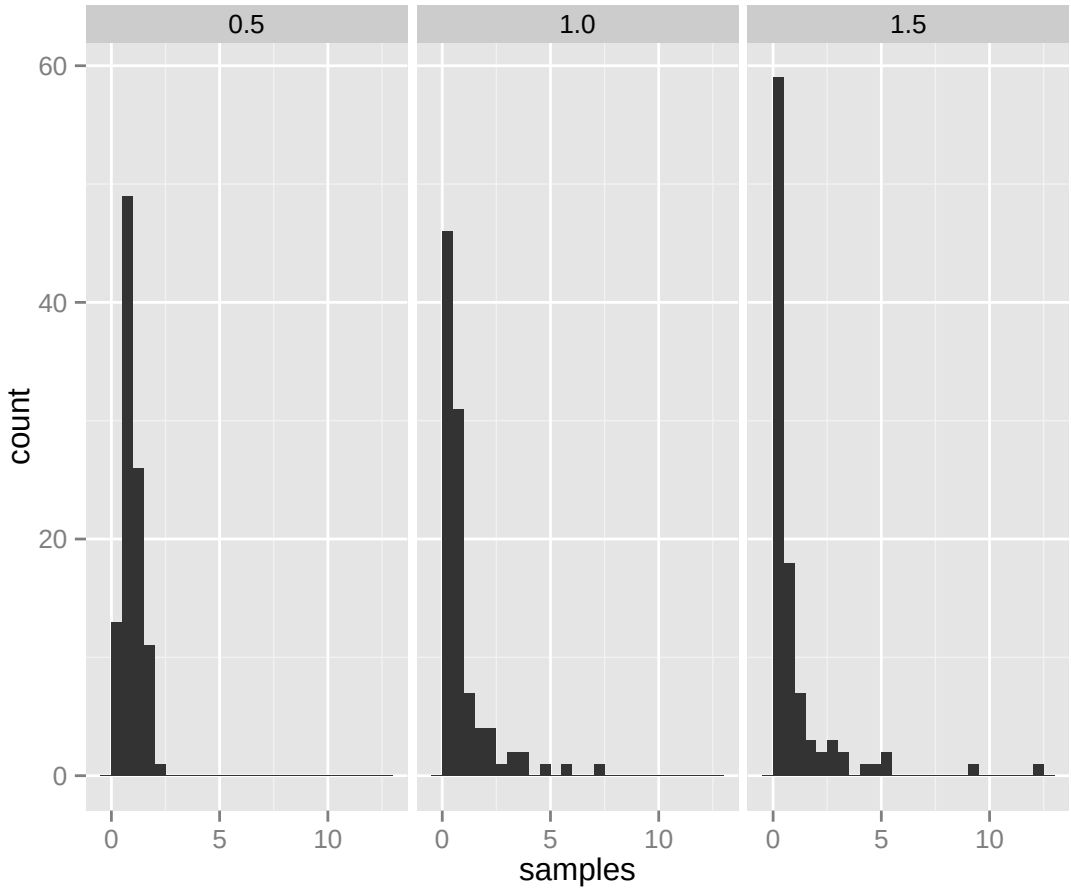
Normality assumptions



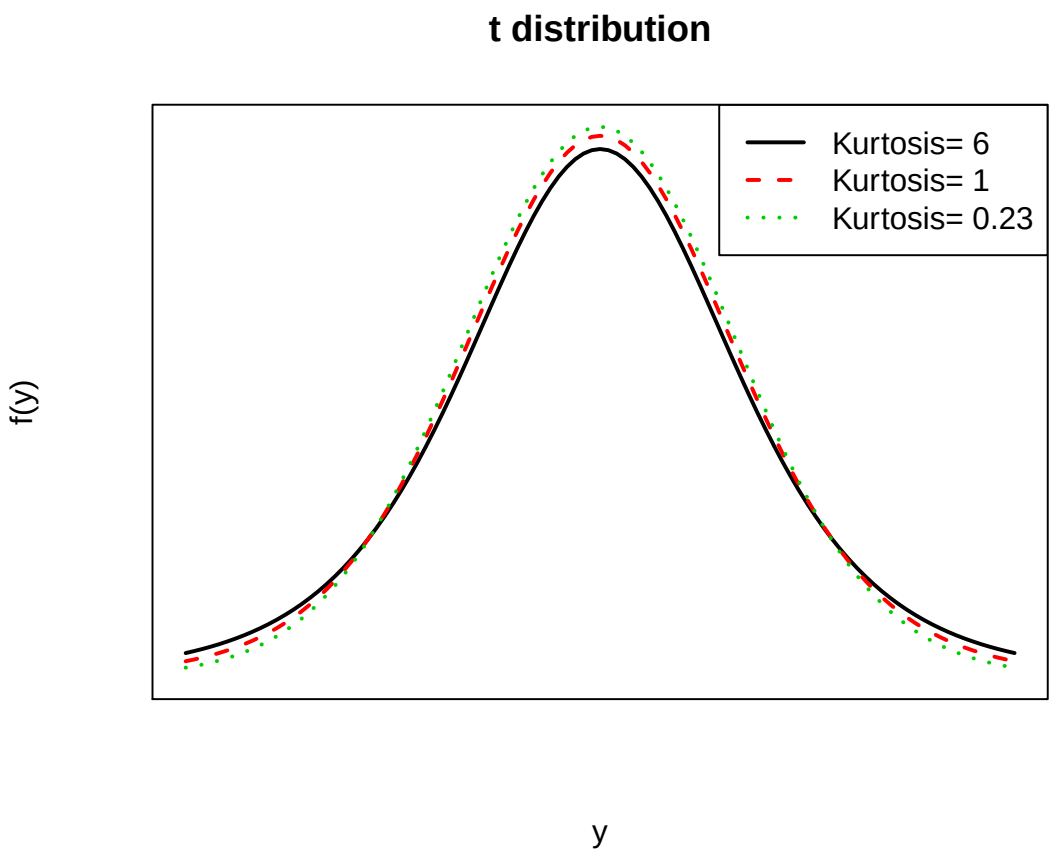
Skewness


Log−normal distribution
Skewness= 1.75
Skewness= 6.18
Skewness= 33.47
Mean
f(y)
y

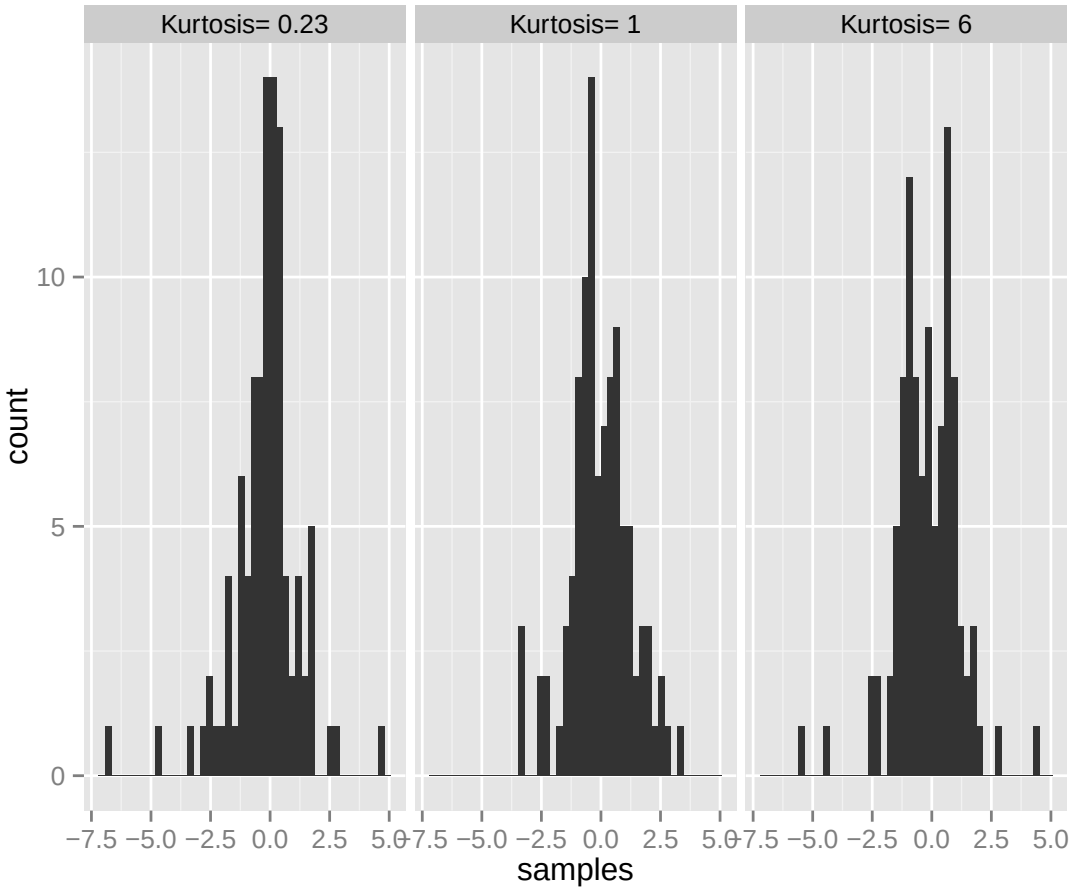
Samples from skewed distributions



Kurtosis (heavy-tailedness)



Kurtosis (heavy-tailedness)



Robustness

Definition

A statistical procedure is robust to departures from a particular assumption if it is valid even when the assumption is not met.

Remark If a 95% confidence interval is robust to departures from a particular assumption, the confidence interval should cover the true value about 95% of the time.

Robustness to skewness and kurtosis

Proportion of 95% confidence intervals that cover the true difference in means in an equal-sample two-sample t-test.

	sample.size	strongly.skewed	moderately.skewed	mildly.skewed	heavy.tailed	short.tailed
1	5	95.5	95.4	95.2	98.3	94.5
2	10	95.5	95.4	95.2	98.3	94.6
3	25	95.3	95.3	95.1	98.2	94.9
4	50	95.1	95.3	95.1	98.1	95.2
5	100	94.8	95.3	95.0	98.0	95.6

Robustness to differences in variances

Proportion of 95% confidence intervals that cover the true difference in means in an equal-sample two-sample t-test.

	n1	n2	r0.25x	r0.5x	r1x	r2x	r4x
1	10	10	95.2	94.2	94.7	95.2	94.5
2	10	20	83.0	89.3	94.4	98.7	99.1
3	10	40	71.0	82.6	95.2	99.5	99.9
4	100	100	94.8	96.2	95.4	95.3	95.1
5	100	200	86.5	88.3	94.8	98.8	99.4
6	100	400	71.6	81.5	95.0	99.5	99.9

Outliers

Definition

A statistical procedure is resistant if it does not change very much when a small part of the data changes, perhaps drastically.

Identify outliers:

1. If recording errors, fix.
2. If outlier comes from a different population, remove and report.
3. If results are the same with and without outliers, report with outliers.
4. If results are different, use resistant analysis or report both analyses.

Common ways for independence to be violated

- Cluster effect
  - e.g. pigs in a pen
- Correlation effect
  - e.g. measurements in time
- Spatial effect
  - e.g. corn fields (drainage)