MAT 2377
Probability and Statistics for Engineers

Chapter 4
Descriptive Statistics and Sampling Distributions

P. Boily (uOttawa)

Winter 2021

Based on course notes by Rafał Kulik

Contents

4.3 – Sampling Distributions (p.29)

- Sum of Independent Random Variables (p.31)
- Independent and Identically Distributed Random Variables (p.32)
- Sample Mean (Reprise) (p.35)
- Sum of Independent Normal Random Variables (p.37)

4.4 – Central Limit Theorem (p.40)

4.5 – Sampling Distributions (Reprise)

- Difference Between 2 Means (p.49)
- Sample Variance S^2 (p.51)
- Sample Mean With Unknown Population Variance (p.54)
- $F-$Distribution (p.60)

4.1 – Data Descriptions

In a sense, the underlying reason for statistical analysis is to reach an **understanding of the data**.

Studies and experiments give rise to **statistical units**.

These units are typically described with **variables** (and measurements).

Variables are either **qualitative** (categorical) or **quantitative** (numerical).

Categorical variables take values (**levels**) from a finite set of **categories** (or classes).

Numerical variables take values from a (potentially infinite) set of **quantities**.

Examples:

1. Age is a numerical variable, measured in years, although is is often reported to the nearest year integer, or in an age range of years, in which case it is an **ordinal** variable (mixture of qualitative or quantitative).

2. Typical numerical variables include distance in m, volume in cm^3, etc.

3. Disease diagnosis is a categorical variable with (at least) 2 categories (positive/negative).

4. Compliance with a standard is a categorical variable: there could be 2 levels (compliant/non-compliant) or more (compliance, minor non-compliance issues, major non-compliance issues).

5. Count variables are numerical variables.

Numerical Summaries

As a first pass, a variable can be described along 2 dimensions: **centrality**, **spread** (**skew** and **kurtosis** are also used sometimes).

- **Centrality** measures: (sample) **median**, (sample) **mean**, (mode, less frequent).
- Spread (or **dispersion**) measures: **standard deviation** (sd), **quartiles**, **inter-quartile range** (IQR), range (less frequent).

The median, range and the quartiles are easily calculated from an **ordered** list of the data.

(Sample) Median

The **median** $\text{med}(x_1, \ldots, x_n)$ of a sample of size n is a numerical value which splits the ordered data into 2 equal subsets: half the observations are below the median, **and** half above it.

- If n is **odd**, then the **position** of the median is $(n+1)/2$, that is to say, the median observation is the $\frac{n+1}{2}^{\text{th}}$ ordered observation.

- If n is **even**, then the median is the average of the $\frac{n}{2}^{\text{th}}$ and the $(\frac{n}{2}+1)^{\text{th}}$ ordered observations.

The procedure is simple: order the data, and follow the even/odd rules **to the letter**.

Examples:

1. $\text{med}(4, 6, 1, 3, 7) = \text{med}(1, 3, 4, 6, 7) = x_{(5+1)/2} = x_3 = 4$. There are 2 observations below 4 $(1, 3)$, and 2 observations above 4 $(6, 7)$.

2. $\text{med}(1, 3, 4, 6, 7, 23) = \frac{x_{6/2}+x_{6/2+1}}{2} = \frac{x_3+x_4}{2} = \frac{4+6}{2} = 5$. There are 3 observations below 5 $(1, 3, 4)$, and 3 observations above 4 $(6, 7, 23)$.

3. $\text{med}(1, 3, 3, 6, 7) = x_{(5+1)/2} = x_3 = 3$. There seems to be only 1 observation below 3 (1), but 2 observations above 3 $(6, 7)$.

   This is not quite the correct interpretation of the median: **above** and **below** in the definition should be interpreted as **after** and **before**, respectively. In this example, there are 2 observations $(x_1 = 1, x_2 = 3)$ before the median $(x_3 = 3)$, and 2 after $(x_4 = 6, x_5 = 7)$.

(Sample) Mean

The **mean** of a sample is simply the arithmetic average of its observations. For observations $x_1, x_2, \ldots, x_n$, the sample mean is

$$\mathsf{AM}(x_1, \ldots, x_n) = \overline{x} = \frac{x_1 + x_2 + \ldots + x_n}{n} = \frac{1}{n}\left(\sum_{i=1}^{n} x_i\right)$$

Other means exist, such as the **harmonic** mean and the **geometric** mean:

$$\mathsf{HM}(x_1, \ldots, x_n) = \frac{n}{\frac{1}{x_1} + \cdots + \frac{1}{x_n}} \quad \text{and} \quad \mathsf{GM}(x_1, \ldots, x_n) = \sqrt[n]{x_1 \cdots x_n}.$$

Examples:

1. $\mathsf{AM}(4,6,1,3,7) = \frac{4+6+1+3+7}{5} = \frac{21}{5} = 4.2 \approx 4 = \mathsf{med}(4,6,1,3,7)$.

2. $\mathsf{AM}(1,3,4,6,7,23) = \frac{1+3+4+6+7+23}{6} = \frac{44}{6} \approx 7.3$, which is not nearly as close to $\mathsf{med}(1,3,4,6,7,23) = 5$.

3. $\mathsf{HM}(4,6,1,3,7) = \frac{5}{\frac{1}{4}+\frac{1}{6}+\frac{1}{1}+\frac{1}{3}+\frac{1}{7}} = \frac{5}{53/28} = \frac{140}{53} \approx 2.64$.

4. $\mathsf{GM}(4,6,1,3,7) = \sqrt[5]{4 \cdot 6 \cdot 1 \cdot 3 \cdot 7} \approx \sqrt[5]{(504)} \approx 3.47$.

If $x = (x_1, \ldots, x_n)$ and $x_i > 0$ for all i,

$$\min(x) \leq \mathsf{HM}(x) \leq \mathsf{GM}(x) \leq \mathsf{AM}(x) \leq \max(x).$$

Mean or Median?

Which measure of centrality should be used to report on the data?

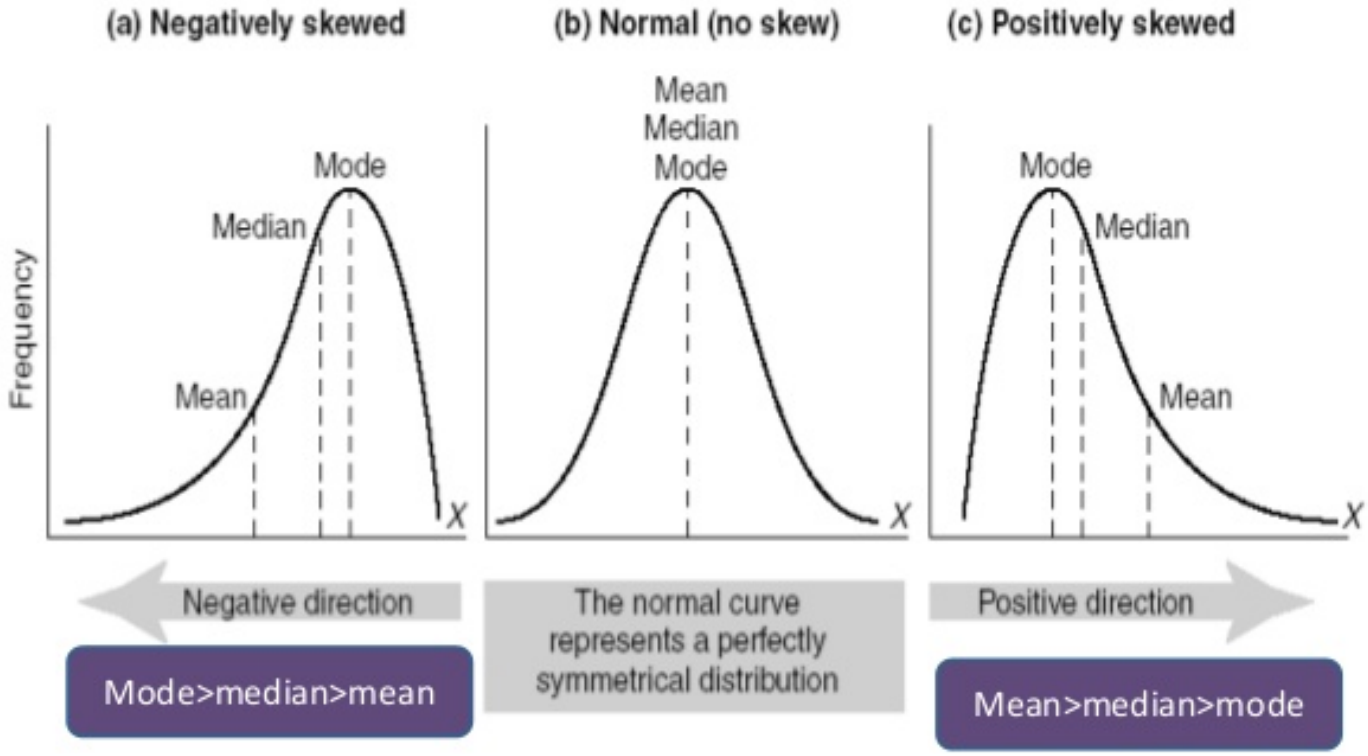
1. The mean is **theoretically supported** (see Central Limit Theorem).

2. If the data distribution is roughly symmetric then both values will be near one another.

3. If the data distribution is **skewed** then the mean is pulled toward the long tail and as a result gives a distorted view of the centre. Consequently, medians are generally used for house prices, incomes etc.

4. The median is **robust** against outliers and incorrect readings whereas the mean is not.

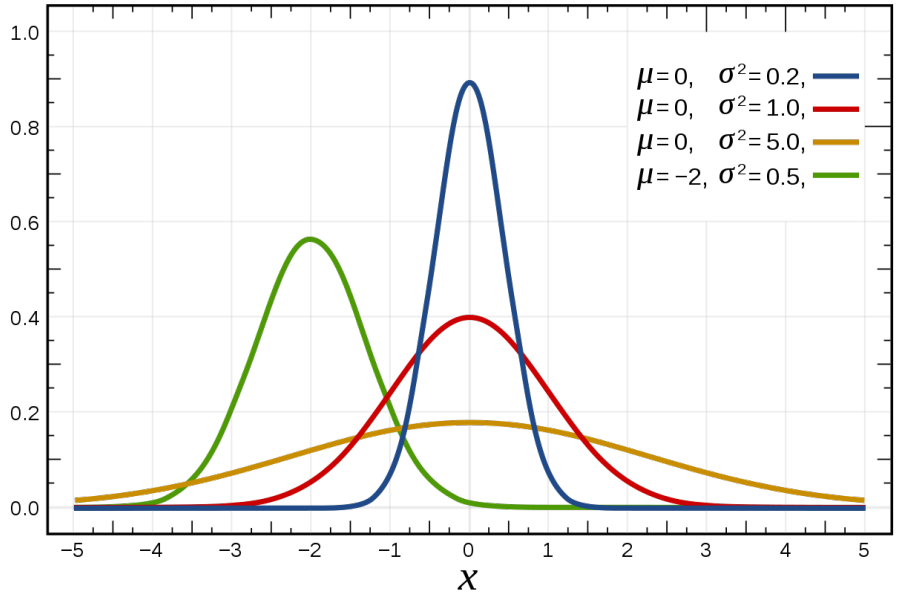
Mean or Median?



Standard Deviation (Reprise)

The mean, the median, and the mode provide an idea as to where some of the distribution's "mass" is located.

The standard deviation provides some notion of its spread.



Quartiles

Another way to provide information about the spread of the data is with the help of **centiles**, **deciles**, or **quartiles**.

The **lower quartile** $Q_1(x_1, \ldots, x_n)$ of a sample of size n, or Q_1, is a numerical value which splits the ordered data into 2 unequal subsets: 25% of the observations are below Q_1, **and** 75% of the observations are above Q_1.

Similarly, the **upper quartile** Q_3 splits the ordered data into 75% of the observations below Q_3, **and** 25% of the observations above Q_3.

The median can be interpreted as the **middle quartile**, Q_2, of the sample, the minimum as Q_0, and the maximum as Q_4.

Centiles $p_i,\ i = 0, \ldots, 100$ and deciles $d_j,\ j = 0, \ldots, 10$ run through different splitting percentages $\Longrightarrow p_{25} = Q_1, p_{75} = Q_3, d_5 = Q_2$, etc.

Sort the sample observations $\{x_1, x_2, \ldots, x_n\}$ in an **increasing order** as

$$y_1 \leq y_2 \leq \ldots \leq y_n.$$

The smallest y_1 has **rank** 1 and the largest y_n has **rank** n.

The lower quartile Q_1 is computed as the average of ordered observations with ranks $\lfloor \frac{n}{4} \rfloor$ and $\lfloor \frac{n}{4} \rfloor + 1$. Similarly, Q_3 is computed as the average of ordered observations with ranks $\lceil \frac{3n}{4} \rceil$ and $\lceil \frac{3n}{4} \rceil + 1$.

Examples:

$$Q_1(1, 3, 4, 6, 7, 10, 12, 23) = 3.5, \quad Q_3(1, 3, 4, 6, 7, 10, 12, 23) = 11.$$

Example: a dataset describes the daily number of accidents in Sydney:

```
> accident
 6, 3, 2, 24, 12, 3, 7, 14, 21, 9, 14, 22, 15, 2,
 17, 10, 7, 7, 31, 7, 18, 6, 8, 2, 3, 2, 17, 7, 7,
 21, 13, 23, 1, 11, 3, 9, 4, 9, 9, 25
> sort(accident)
 1  2  2  2  2  3  3  3  3  4  6  6  7  7  7  7  7  8  9
 9  9  9 10 11 12 13 14 14 15 17 17 18 21 21 22 23 24 25 31
> summary(accident)
   Min.  1st quartile  Median  Mean  3rd quartile   Max.
   1.00     5.50        9.00   10.78    15.50      31.00
> var(accident) 58.7
```

Now, replace the 31 with 130. The new mean is 13.28 and the new variance is 412.4, but the median is the same.

Dispersion Measures

The **sample range** is range$(x_1, \ldots, x_n) = \max\{x_i\} - \min\{x_i\} = y_n - y_1$, where $y_1 \leq \ldots \leq y_n$ is the ranked data.

The **inter-quartile range** is IQR $= Q_3 - Q_1$.

The **sample standard deviation** s and **sample variance** s^2 are estimates of the underlying distribution's σ and σ^2.

For observations $x_1, x_2, \ldots, x_n$, we have

$$s^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \overline{x})^2 = \frac{1}{n-1}\left(\sum_{i=1}^{n} x_i^2 - \frac{1}{n}\left(\sum_{i=1}^{n} x_i\right)^2\right).$$

Outliers

An outlier is an observation that lies outside the overall pattern in a distribution. Let x be an observation in the sample. It is a **suspected outlier** if

$$x < Q_1 - 1.5\,\text{IQR} \quad \text{or} \quad x > Q_3 + 1.5\,\text{IQR},$$

where $\text{IQR} = Q_3 - Q_1$ it the **inter-quartile range** $Q_3 - Q_1$.

This definition only applies with certainty to **normally distributed** data, although it is often used as a first outlier analysis method.

Exercise: Consider a sample of $n = 10$ observations displayed in ascending order.

$$15, 16, 18, 18, 20, 20, 21, 22, 23, 75.$$

1. Verify that the standard deviation for this sample is $s = 17.81884$.

2. Verify that $Q_1 = 17.5$ and $Q_3 = 22.25$.

3. Are there any likely outliers in the sample? If so, indicate their values.

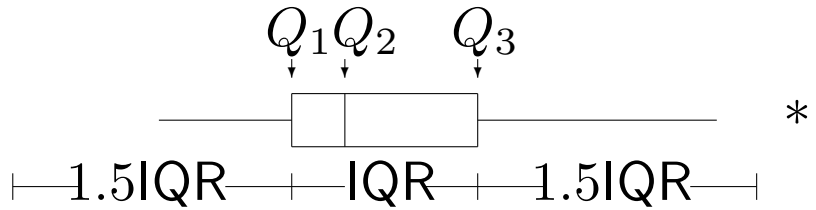
4.2 – Visual Summaries

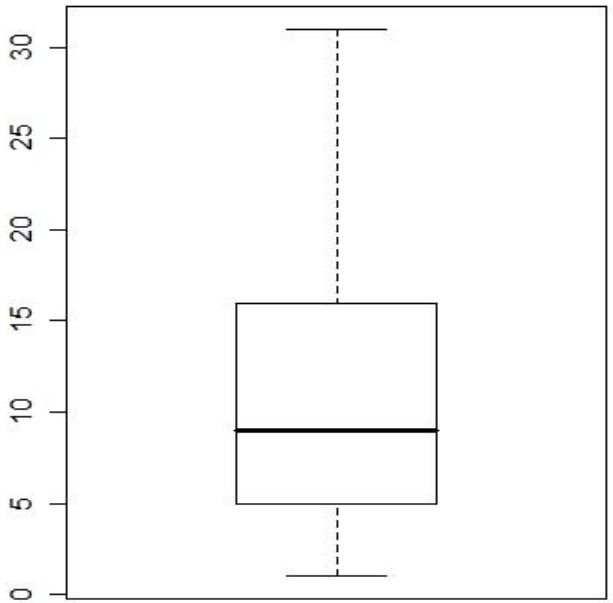
The **boxplot** is a quick and easy way to present a graphical summary of a univariate distribution.

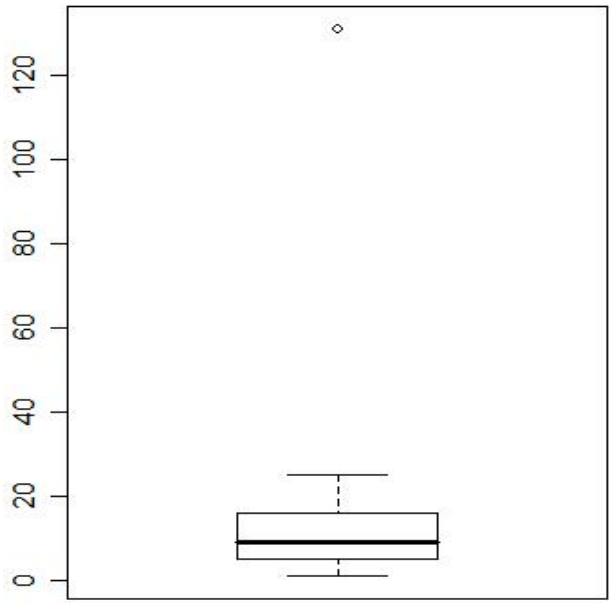
Draw a box along the observation axis, with endpoints at the lower and upper quartiles, and with a "belt" at the median.

Then, plot a line extending from Q_1 to the smallest value less than 1.5IQR to the left of Q_1 and from Q_3 to the largest value less than 1.5IQR to the right of Q_3.

Any suspected outlier is plotted separately.




0
5
10
15
20
25
30


0
20
40
60
80
100
120

Skewness

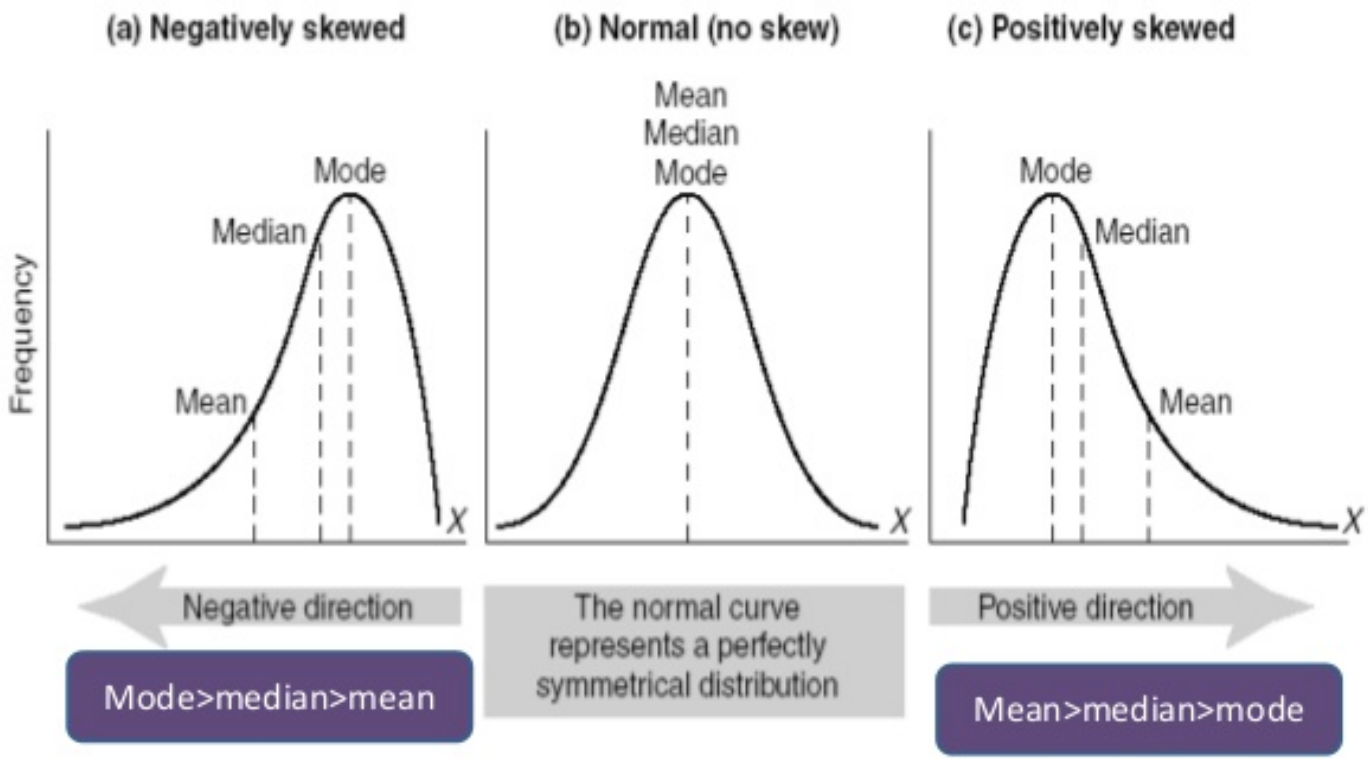
If the data distribution is symmetric then the (population) median and mean are equal and the first and third (population) quartiles are equidistant from the median.

If $Q_3 - Q_2 > Q_2 - Q_1$ then the data distribution is **skewed to the right**.

If $Q_3 - Q_2 < Q_2 - Q_1$ then the data distribution is **skewed to left**.

In both of the examples in the previous slide, the distributions are skewed to the right.

Skewness


(a) Negatively skewed
(b) Normal (no skew)
(c) Positively skewed
Frequency
Mode
Median
Mean
Mean
Median
Mode
Mode
Median
Mean
X
X
X
Negative direction
The normal curve represents a perfectly symmetrical distribution
Positive direction
Mode>median>mean
Mean>median>mode

Histograms

Histograms also provide an indication of the distribution of the sample.

Histograms should contain the following information:

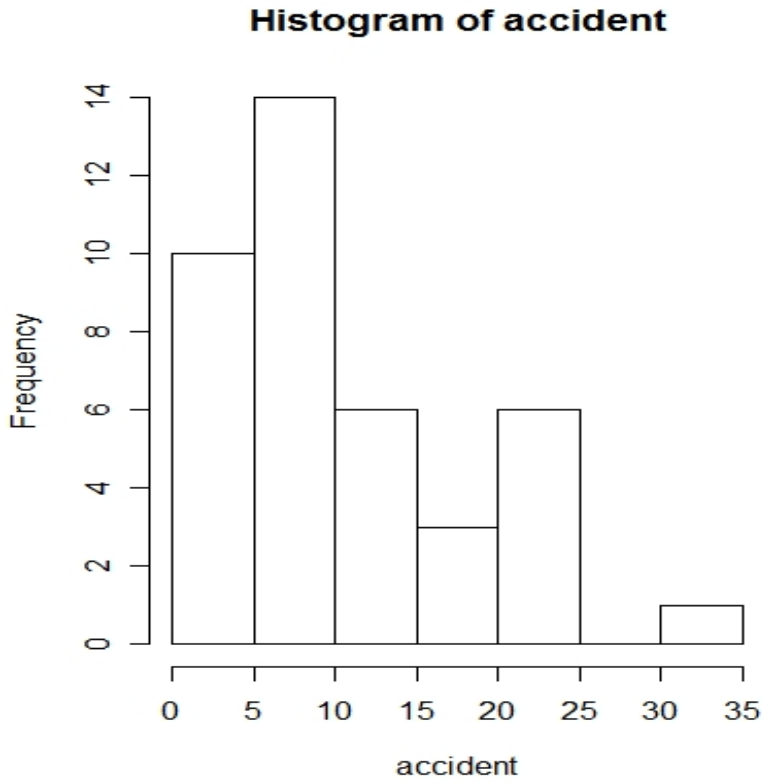
- the range of the histogram is $r = \max\{x_i\} - \min\{x_i\}$;
- the number of bins should approach $k = \sqrt{n}$, where n is the sample size;
- the bin width should approach r/k,
- and the frequency of observations in each bin should be added to the chart.

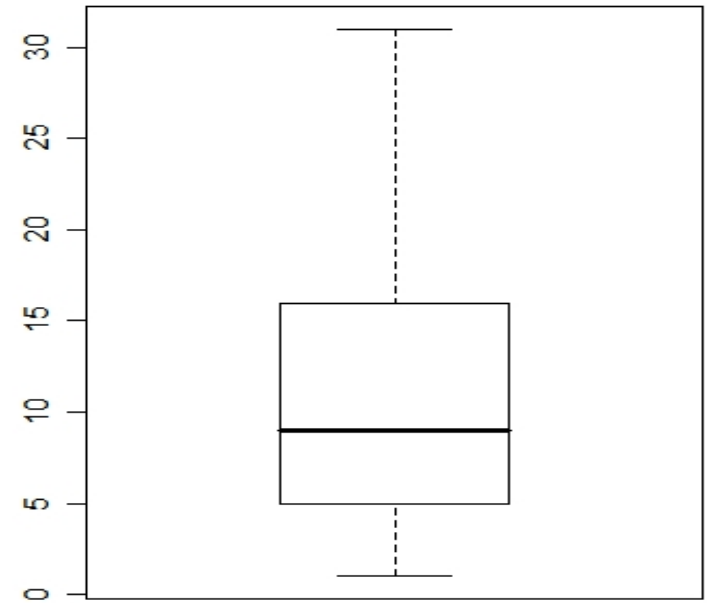
Shapes of Datasets

Boxplots give an easy graphical means of getting an impression of the shape of the data set. The shape is used to suggest a mathematical model for the situation of interest.

The data set is **right skewed** if the boxplot is stretched to the right.

Similar observations can be inferred from the histogram.

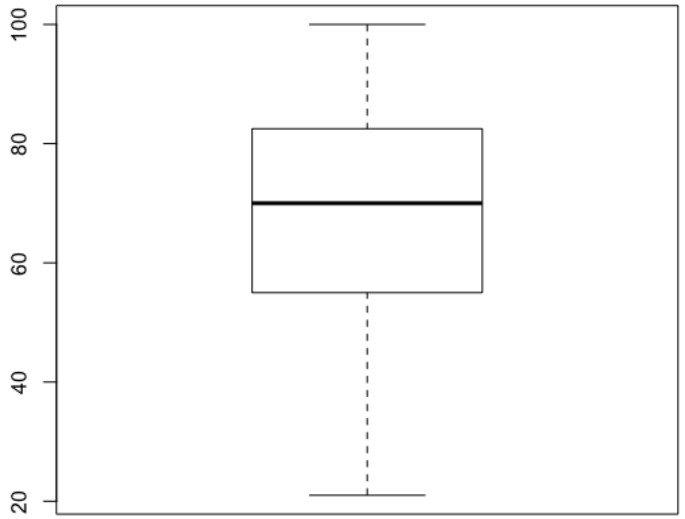

Histogram of accident
Frequency
0
2
4
6
8
10
12
14
0
5
10
15
20
25
30
35
accident

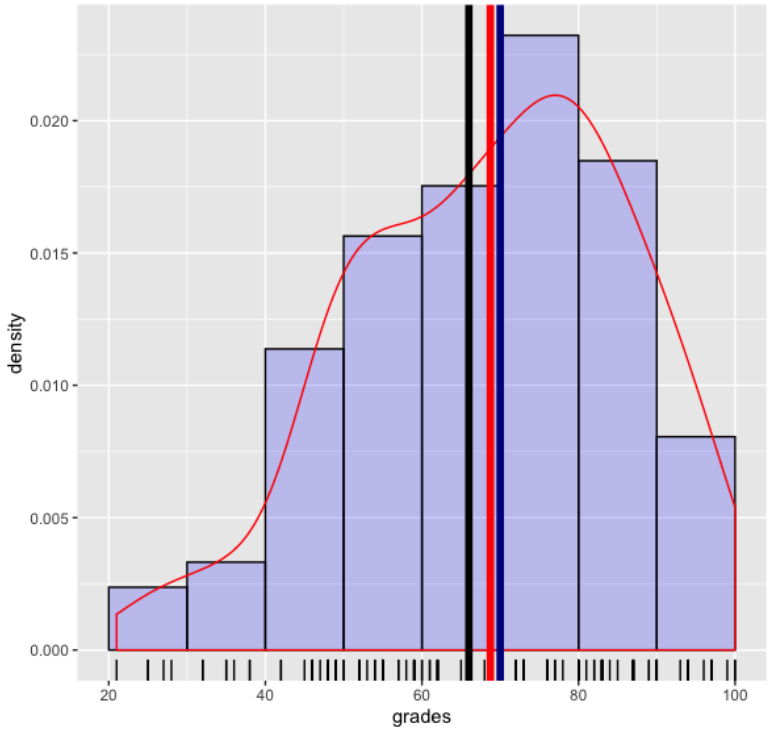

0
5
10
15
20
25
30

Example: the grades for the combined first assignment and midterm are shown below. Discuss the results.

```
> grades<-c(80,73,83,60,49,96,87,87,60,53,66,83,32,80,66,90,72,55,76,46,48,69,45,48,77,
52,59,97,76,89,73,73,48,59,55,76,87,55,80,90,83,66,80,97,80,55,94,73,49,32,76,57,42,94,
80,90,90,62,85,87,97,50,73,77,66,35,66,76,90,73,80,70,73,94,59,52,81,90,55,73,76,90,46,
66,76,69,76,80,42,66,83,80,46,55,80,76,94,69,57,55,66,46,87,83,49,82,93,47,59,68,65,66,
69,76,38,99,61,46,73,90,66,100,83,48,97,69,62,80,66,55,28,83,59,48,61,87,72,46,94,48,59,
69,97,83,80,66,76,25,55,69,76,38,21,87,52,90,62,73,73,89,25,94,27,66,66,76,90,83,52,52,
83,66,48,62,80,35,59,72,97,69,62,90,48,83,55,58,66,100,82,78,62,73,55,84,83,66,49,76,73,
54,55,87,50,73,54,52,62,36,87,80,80)

> hist(grades)

> # function to calculate mode
> fun.mode<-function(x){as.numeric(names(sort(-table(x)))[1])}
```

```
> library(ggplot2)
> ggplot(data=data.frame(grades), aes(grades)) + geom_histogram(aes(y =..density..),
                 breaks=seq(20, 100, by = 10),
                 col="black",
                 fill="blue",
                 alpha=.2) +
    geom_density(col=2) + geom_rug(aes(grades)) +
    geom_vline(aes(xintercept = mean(grades)),col='red',size=2) +
    geom_vline(aes(xintercept = median(grades)),col='darkblue',size=2) +
    geom_vline(aes(xintercept = fun.mode(grades)),col='black',size=2)

> boxplot(grades)

> summary(grades)
Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
21.00   55.00   70.00   68.74   82.50  100.00

> library(psych)
> describe(grades)
n   mean    sd median trimmed   mad min max range  skew kurtosis  se
211 68.74 17.37     70   69.43 19.27  21 100    79 -0.37    -0.46 1.2
```


100
80
60
40
20


density
0.020
0.015
0.010
0.005
0.000
20
40
60
80
100
grades

4.3 – Sampling Distributions

A **population** is a set of similar items which is of interest in relation to some questions or experiments. In some situations, it is impossible to observe the entire set of observations that make up a population. In these cases, we must consider a **sample** (subset) of the population in order to make inferences about the population.

Suppose that $X_1, \ldots, X_n$ are n **independent** random variables, each having the same distribution function F, i.e. they are **identically distributed**. Then, $\{X_1, \ldots, X_n\}$ is a **random sample** of size n from the population with distribution function F.

Any function of a random sample is called a **statistic** of the sample.

The probability distribution of a statistic is called a **sampling distribution**.

Linear Properties of Expectation and Variance

Recall: if X is a random variable, and $a, b \in \mathbb{R}$, then

$$\begin{aligned} \mathrm{E}\left[a + bX\right] &= a + b\mathrm{E}[X]\,, \\ \mathrm{Var}\left[a + bX\right] &= b^2\mathrm{Var}[X]\,, \\ \mathrm{SD}\left[a + bX\right] &= |b|\mathrm{SD}[X]\,. \end{aligned}$$

Sum of Independent Random Variables

For any random variables X and Y, we have

$$\mathrm{E}[X+Y] = \mathrm{E}[X] + \mathrm{E}[Y]$$

If **in addition** X and Y are **independent** random variables, then

$$\mathrm{Var}[X+Y] = \mathrm{Var}[X] + \mathrm{Var}[Y]\,.$$

More generally, if $X_1, X_2, \ldots, X_n$ are **independent** random variables, then

$$\mathrm{E}\left[\sum_{i=1}^{n} X_i\right] = \sum_{i=1}^{n} \mathrm{E}[X_i] \quad \text{and} \quad \mathrm{Var}\left[\sum_{i=1}^{n} X_i\right] = \sum_{i=1}^{n} \mathrm{Var}[X_i]\,.$$

The IID Case

A special case of the above occurs when all of $X_1, \ldots, X_n$ have **exactly the same distribution** (i.e. same c.d.f.). In that case we say they are **independent and identically distributed**, which is traditionally abbreviated to **iid**.

If $X_1, \ldots, X_n$ are iid, and

$$\mathrm{E}[X_i] = \mu \quad \text{and} \quad \mathrm{Var}[X_i] = \sigma^2 \quad \text{for } i = 1, \ldots, n,$$

then

$$\mathrm{E}\left[\sum_{i=1}^{n} X_i\right] = n\mu \quad \text{and} \quad \mathrm{Var}\left[\sum_{i=1}^{n} X_i\right] = n\sigma^2 .$$

Examples

1. A random sample of size 100 is to be taken from a population with mean-value 50 and variance 0.25. Find the expected value and variance of the **sample total**.

   Solution: this problem translates to "if $X_1, \ldots, X_{100}$ are iid with $\mathrm{E}[X_i] = \mu = 50$ and $\mathrm{Var}[X] = \sigma^2 = 0.25$ for $i = 1, \ldots, 100$, find $\mathrm{E}[\tau]$ and $\mathrm{Var}[\tau]$ for $\tau = \sum_{i=1}^{n} X_i$". According to the iid formulas,

$$\mathrm{E}\left[\sum_{i=1}^{n} X_i\right] = 100\mu = 5000 \text{ and } \mathrm{Var}\left[\sum_{i=1}^{n} X_i\right] = 100\sigma^2 = 25\,.$$

2. The mean value of potting mix bags weights is 5 kg, with standard deviation 0.2. If a shop assistant carries 4 bags (selected independently from stock) then what is the expected value and standard deviation of the total weight carried?

   Solution: there is an implicit "population" of bag weights. Let X_1, X_2, X_3, X_4 be iid with $\mathrm{E}[X_i] = \mu = 5$ and $\mathrm{SD}[X_i] = \sigma = 0.2$ (and thus $\mathrm{Var}[X_i] = \sigma^2 = 0.2^2 = 0.04$). Let $\tau = X_1 + X_2 + X_3 + X_4$. According to the iid formulas,

$$\mathrm{E}[\tau] = n\mu = 4 \cdot 5 = 20$$
$$\mathrm{Var}[\tau] = n\sigma^2 = 4 \cdot 0.04 = 0.16.$$

   So $\mathrm{SD}[\tau] = \sqrt{0.16} = 0.4$.

Sample Mean (Reprise)

The **sample mean** is a typical statistic of interest:

$$\overline{X} = \frac{1}{n}\sum_{i=1}^{n} X_i \,.$$

If $X_1, \ldots, X_n$ are iid with $\mathrm{E}[X_i] = \mu$ and $\mathrm{Var}[X_i] = \sigma^2$ for all $i = 1, \ldots, n$, then

$$\mathrm{E}\left[\overline{X}\right] = \mathrm{E}\left[\frac{1}{n}\sum_{i=1}^{n} X_i\right] = \frac{1}{n}\mathrm{E}\left[\sum_{i=1}^{n} X_i\right] = \frac{1}{n}\left(n\mu\right) = \mu$$

$$\mathrm{Var}\left[\overline{X}\right] = \mathrm{Var}\left[\frac{1}{n}\sum_{i=1}^{n} X_i\right] = \left[\frac{1}{n}\right]^2 \mathrm{Var}\left[\sum_{i=1}^{n} X_i\right] = \frac{1}{n^2}\left(n\sigma^2\right) = \frac{\sigma^2}{n}\,.$$

Example: a set of scales returns the true weight of the object being weighed plus a random error with mean 0 and standard deviation 0.1 g. Find the standard deviation of the average of 9 such measurements of an object.

Solution: suppose the object has true weight μ. The "random error" indicates that each measurement $i = 1, \ldots, 9$ is written as $X_i = \mu + Z_i$ where $\mathrm{E}[Z_i] = 0$ and $\mathrm{SD}[Z_i] = 0.1$ and the Z_i's are iid.

Hence the X_i's are iid with $\mathrm{E}[X_i] = \mu$ and $\mathrm{SD}[X_i] = \sigma = 0.1$ (why?). If we average $X_1, \ldots, X_n$ (with $n = 9$) to get $\overline{X}$, then

$$\mathrm{E}\left[\overline{X}\right] = \mu \text{ and } \mathrm{SD}\left[\overline{X}\right] = \frac{\sigma}{\sqrt{n}} = \frac{0.1}{\sqrt{9}} = \frac{1}{30} \approx 0.033\,.$$

Note that we didn't know the distribution of the X_i, only the mean and variance.

Sum of Independent Normal RVs

Another interesting case occurs when we have **multiple independent normal** random variables on the same experiment.

Suppose $X_i \sim \mathcal{N}\left(\mu_i, \sigma_i^2\right)$ for $i = 1, \ldots, n$, and all the X_i are independent. We already know that for the sum $\tau = X_1 + \cdots + X_n$, we have

$$\mathrm{E}[\tau] = \mathrm{E}[X_1 + \cdots + X_n] = \mathrm{E}[X_1] + \cdots + \mathrm{E}[X_n] = \mu_1 + \cdots + \mu_n\,;$$

$$\mathrm{Var}[\tau] = \mathrm{Var}[X_1 + \cdots + X_n] = \mathrm{Var}[X_1] + \cdots + \mathrm{Var}[X_n] = \sigma_1^2 + \cdots + \sigma_n^2\,.$$

It turns out that τ is **also normally distributed**, i.e.

$$\tau = \sum_{i=1}^{n} X_i \sim \mathcal{N}\left(\mu_1 + \cdots + \mu_n, \sigma_1^2 + \cdots + \sigma_n^2\right).$$

If $\{X_1, \ldots, X_n\}$ is a random sample from a population **with mean μ and variance σ^2**, then

- $\mathrm{E}\left[\sum_{i=1}^n X_i\right] = n\mu$ and $\mathrm{Var}\left[\sum_{i=1}^n X_i\right] = n\sigma^2$;
- $\mathrm{E}\left[\overline{X}\right] = \mu$ and $\mathrm{Var}\left[\overline{X}\right] = \sigma^2/n$;
- furthermore, if the population distribution is **normal**, then $\sum_{i=1}^n X_i$ and $\overline{X}$ are also normal, i.e.

$$\sum_{i=1}^n X_i \sim \mathcal{N}\left(n\mu, n\sigma^2\right) \quad \text{and} \quad \overline{X} \sim \mathcal{N}\left(\mu, \frac{\sigma^2}{n}\right).$$

Example: suppose that the population of students' weights is normal with mean 75 kg and standard deviation 5 kg. If 16 students are picked at random, what is the distribution of the (random) total weight τ? What is the probability that the total weight exceeds 1250 kg?

Solution: If $X_1, \ldots, X_{16}$ are iid $\mathcal{N}(75, 25)$, then the sum $\tau = X_1 + \cdots + X_{16}$ is also normally distributed with

$$\tau = \sum_{i=1}^{16} X_i \sim \mathcal{N}(16{\cdot}75, 16{\cdot}25) = \mathcal{N}(1200, 400) \text{ and } Z = \frac{\tau - 1200}{\sqrt{400}} \sim \mathcal{N}(0, 1).$$

Thus,

$$\begin{aligned} P(\tau > 1250) &= P\left(\frac{\tau - 1200}{\sqrt{400}} > \frac{1250 - 1200}{20}\right) \\ &= P(Z > 2.5) = 1 - P(Z \leq 2.5) \approx 1 - 0.9938 = 0.0062\,. \end{aligned}$$

4.4 – Central Limit Theorem

Motivation: a professor has been teaching a course for the last 20 years. For every class during that period, the mid-term exam grades of all the students have been recorded.

Let $X_{i,j}$ be the grade of student i in year j. Looking back on the class lists, the professor finds that

$$\mathrm{E}[X_{i,j}] = 56 \quad \text{and} \quad \mathrm{SD}[X_{i,j}] = 11.$$

This year, there are 49 students in the class. What should the professor expect that the class average on the mid-term examination will be?

Of course, the professor is not sure what will happen, but they will try the following approach:

1. simulate the results of the class of 49 students by generating sample grades $X_{1,1}, \ldots, X_{1,49}$ from a **normal** distribution $\mathcal{N}(65, 15^2)$;

2. compute the sample mean for the sample and record it as $\overline{X}_1$;

3. repeat steps 1-2 m times and compute the standard deviation of the sample means $\overline{X}_1, \ldots, \overline{X}_m$;

4. plot the histogram of the sample means $\overline{X}_1, \ldots, \overline{X}_m$.

What do you think is going to happen?

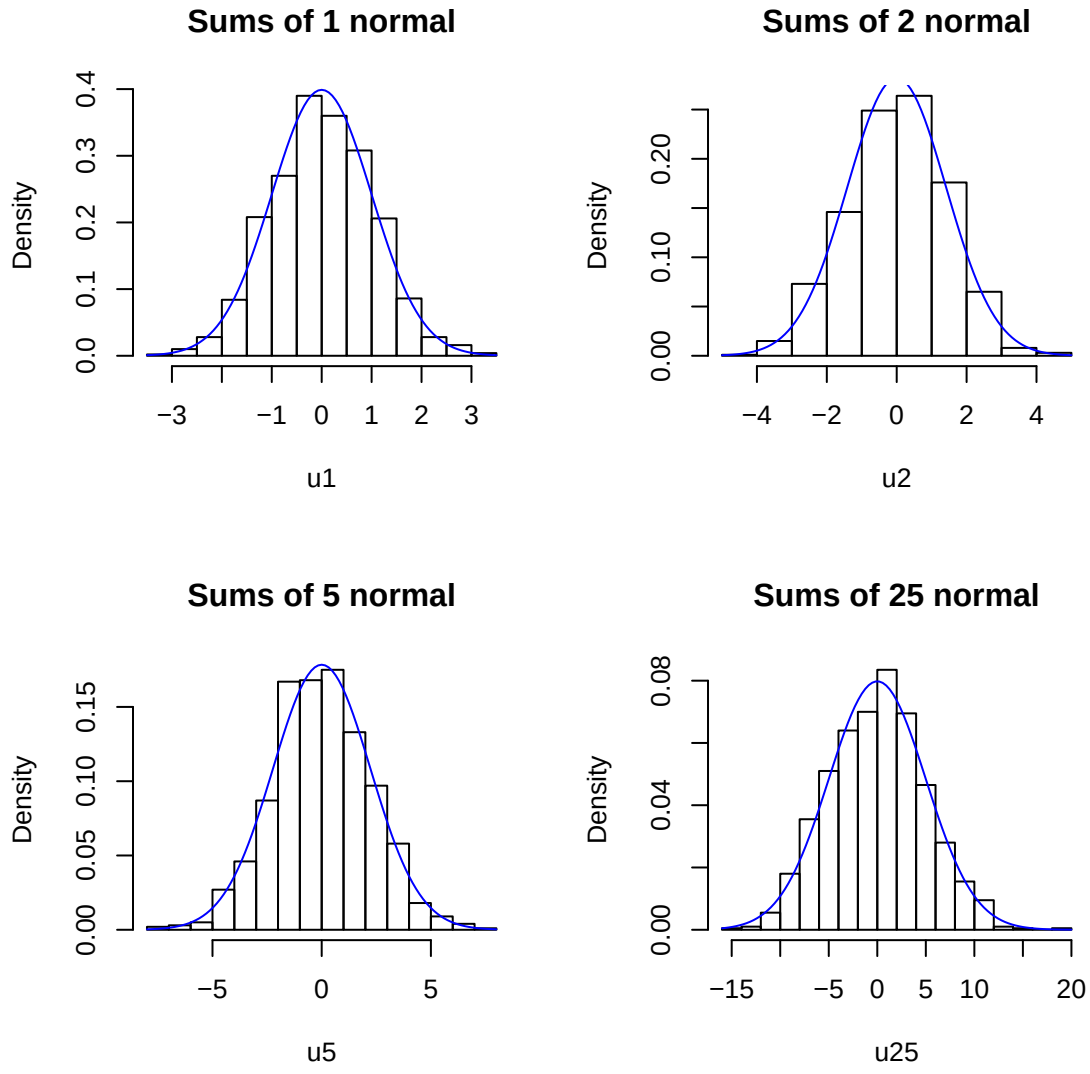
Theorem: If $\overline{X}$ is the mean of a random sample of size n taken from an **unknown** population with mean μ and finite variance σ^2, then $Z = \frac{\overline{X}-\mu}{\sigma/\sqrt{n}}$, has the standard normal distribution $\mathcal{N}(0,1)$ as $n \to \infty$.

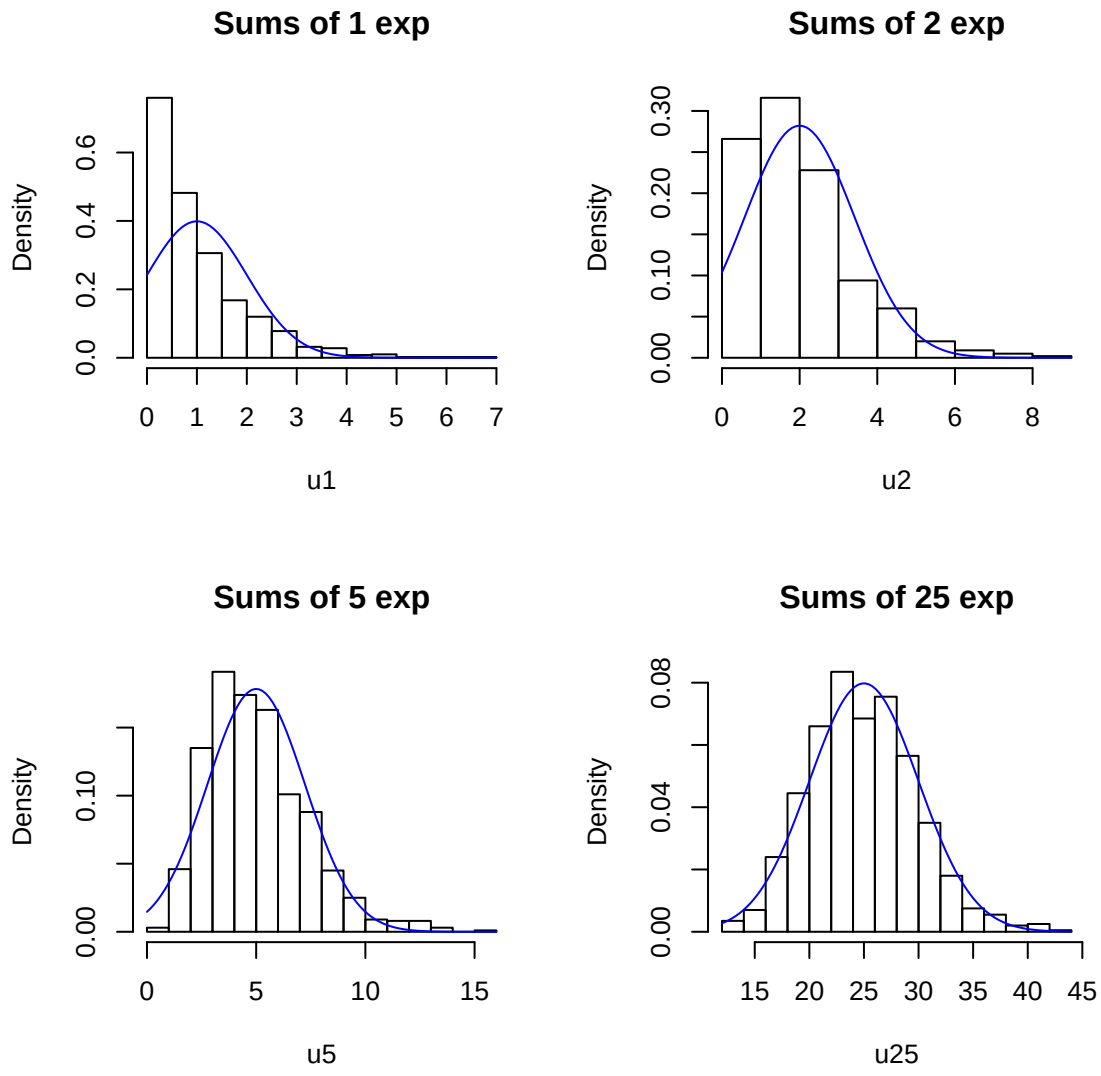
More precisely, the result is a **limiting** result. If we view the standardized

$$Z_n = \frac{\overline{X} - \mu}{\sigma/\sqrt{n}},$$

as functions of n, **regardless of whether the original X_i's are normal or not,** for each z we have

$$\lim_{n\to\infty} P\left(Z_n \leq z\right) = \Phi(z) \quad \text{and} \quad P\left(Z_n \leq z\right) \approx \Phi(z) \text{ if } n \text{ is large enough.}$$


Sums of 1 normal
Density
0.0
0.1
0.2
0.3
0.4
−3
−1
0
1
2
3
u1
Sums of 2 normal
Density
0.00
0.10
0.20
−4
−2
0
2
4
u2
Sums of 5 normal
Density
0.00
0.05
0.10
0.15
−5
0
5
u5
Sums of 25 normal
Density
0.00
0.04
0.08
−15
−5
0
5
10
20
u25


Sums of 1 exp
Density
0.0
0.2
0.4
0.6
0 1 2 3 4 5 6 7
u1
Sums of 2 exp
Density
0.00
0.10
0.20
0.30
0 2 4 6 8
u2
Sums of 5 exp
Density
0.00
0.10
0 5 10 15
u5
Sums of 25 exp
Density
0.00
0.04
0.08
15 20 25 30 35 40 45
u25

Examples:

1. The examination scores in an university course have mean 56 and standard deviation 11. In a class of 49 students, what is the probability that the average mark is below 50? What is the probability that the average mark lies between 50 and 60?

   Solution: Let the marks be $X_1, ..., X_{49}$ and assume the performances are independent. According to the central limit theorem (CLT),

   $$\overline{X} = (X_1 + X_2 + \cdots + X_{49})/49, \text{ with } \mathrm{E}[\overline{X}] = 56, \ \mathrm{Var}[\overline{X}] = 11^2/49.$$

   We thus have

   $$P(\overline{X} < 50) \approx P\left(Z < \frac{50 - 56}{11/7}\right) = P(Z < -3.82) = 0.0001$$

and

$$P(50 < \overline{X} < 60) \approx P\left(\frac{50-56}{11/7} < Z < \frac{60-56}{11/7}\right)$$
$$= \quad P(-3.82 < Z < 2.55) = \Phi(2.55) - \Phi(-3.82) = 0.9945.$$

Note: this says nothing about whether the scores are normally distributed or not. If they were, the $\approx$ would be replaced by $=$.

2. Systolic blood pressure readings for pre-menopausal, non-pregnant women aged $35-40$ have a mean of 122.6 mm Hg and a standard deviation of 11 mm Hg. An independent sample of 25 women is drawn from this target population and their blood pressure is recorded. What is the probability that the average blood pressure is greater than 125 mm Hg? How would the answer change if the sample size increases to 40?

Solution: according to the CLT, $\overline{X} \sim \mathcal{N}(122.6, 121/25)$, approximately. Thus

$$P(\overline{X} > 125) \approx P\left(Z > \frac{125 - 122.6}{11/\sqrt{25}}\right)$$
$$= P(Z > 1.09) = 1 - \Phi(1.09) = 0.1378.$$

If the sample size is 40, then

$$P(\overline{X} > 125) \approx P\left(Z > \frac{125 - 122.6}{11/\sqrt{40}}\right) = 0.0838.$$

Increasing the sample size reduces the probability that the average is far from the expectation of each original measurement.

3. Suppose that we select the random sample $X_1, \ldots, X_{100}$ from a population with mean 5 and variance 0.01. What is the probability that the difference between the sample mean of the random sample and the mean of the population exceeds 0.027?

   Solution: according to the CLT, we know that $Z = \frac{\overline{X}-\mu}{\sigma/\sqrt{n}}$ has standard normal distribution, approximatively. The desired probability is thus

$$P(|\overline{X} - \mu| \geq 0.027) = P(\overline{X} - \mu \geq 0.027 \text{ or } \mu - \overline{X} \geq 0.027)$$

$$= P\left(\frac{\overline{X} - 5}{0.1/\sqrt{100}} \geq \frac{0.027}{0.1/\sqrt{100}}\right) + P\left(\frac{\overline{X} - 5}{0.1/\sqrt{100}} \leq \frac{-0.027}{0.1/\sqrt{100}}\right)$$

$$\approx P\left(Z \geq 2.7\right) + P\left(Z \leq -2.7\right) = 2P\left(Z \geq 2.7\right) \approx 2(0.0035) = 0.007.$$

4.5 – Sampling Distributions (Reprise)

Difference Between 2 Means

Theorem: Let $X_1, \ldots, X_n$ be a random sample from a population with mean μ_1 and variance σ_1^2, and $Y_1, \ldots, Y_m$ be another random sample, independent of X, from a population with mean μ_2 and variance σ_2^2. If $\overline{X}$ and $\overline{Y}$ are the respective sample means, then

$$Z = \frac{\overline{X} - \overline{Y} - (\mu_1 - \mu_2)}{\sqrt{\frac{\sigma_1^2}{n} + \frac{\sigma_2^2}{m}}}$$

has standard normal distribution as $n, m \to \infty$. This is also a **limiting** result.

Example: two different box-filling machine are used to fill cereal boxes on an assembly line. The critical measurement influenced by these machines is the weight of the product in the boxes. For both machines, the variance of these weights is $\sigma^2 = 1$. Each machine produces a sample of 36 boxes, and the weights are recorded. What is the probability that the difference between the respective averages is less than 0.2, assuming that the true means are identical?

Solution: we have $\mu_1 = \mu_2$, $\sigma_1^2 = \sigma_2^2 = 1$, and $n = m = 36$. The desired probability is

$$P\left(|\overline{X} - \overline{Y}| < 0.2\right) = P\left(-0.2 < \overline{X} - \overline{Y} < 0.2\right)$$

$$= P\left(\frac{-0.2 - 0}{\sqrt{1/36 + 1/36}} < \frac{\overline{X} - \overline{Y} - (\mu_1 - \mu_2)}{\sqrt{1/36 + 1/36}} < \frac{0.2 - 0}{\sqrt{1/36 + 1/36}}\right)$$

$$= P(-0.8485 < Z < 0.8485) = \Phi(0.8485) - \Phi(-0.8485) \approx 0.6.$$

Sample Variance S^2

Theorem: if

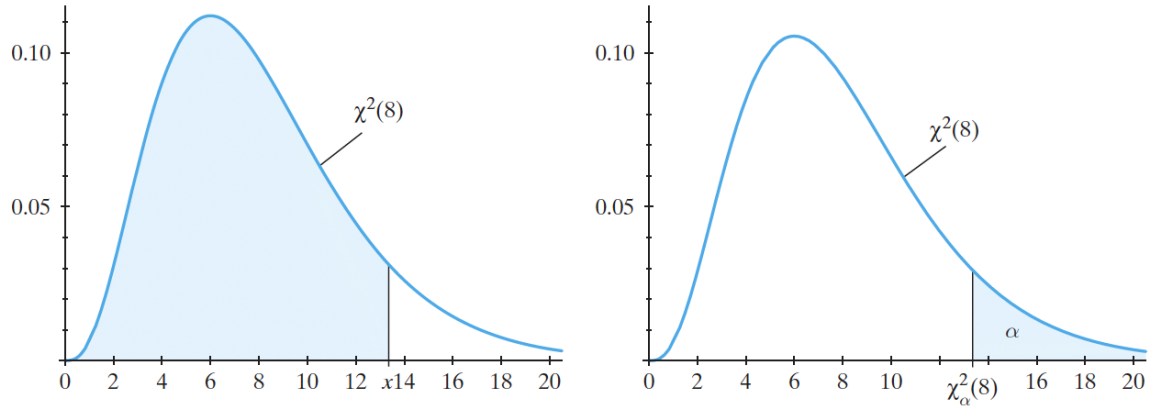
$$S^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \overline{X})^2$$

is the variance of a random sample of size n taken from a normal population with variance σ^2, then the statistic

$$\chi^2 = \frac{(n-1)S^2}{\sigma^2} = \sum_{i=1}^{n}\frac{(X_i - \overline{X})^2}{\sigma^2}$$

has a **chi-squared distribution with $\nu = n - 1$ degrees of freedom**, where $\chi^2(\nu) = \Gamma(1/2, \nu)$.

Table IV The Chi-Square Distribution



$$P(X \le x) = \int_0^x \frac{1}{\Gamma(r/2)2^{r/2}} w^{r/2-1} e^{-w/2} dw$$

	$P(X \le x)$							
	0.010	0.025	0.050	0.100	0.900	0.950	0.975	0.990
r	$\chi^2_{0.99}(r)$	$\chi^2_{0.975}(r)$	$\chi^2_{0.95}(r)$	$\chi^2_{0.90}(r)$	$\chi^2_{0.10}(r)$	$\chi^2_{0.05}(r)$	$\chi^2_{0.025}(r)$	$\chi^2_{0.01}(r)$
1	0.000	0.001	0.004	0.016	2.706	3.841	5.024	6.635
2	0.020	0.051	0.103	0.211	4.605	5.991	7.378	9.210
3	0.115	0.216	0.352	0.584	6.251	7.815	9.348	11.34
4	0.297	0.484	0.711	1.064	7.779	9.488	11.14	13.28
5	0.554	0.831	1.145	1.610	9.236	11.07	12.83	15.09
6	0.872	1.237	1.635	2.204	10.64	12.59	14.45	16.81
7	1.239	1.690	2.167	2.833	12.02	14.07	16.01	18.48
8	1.646	2.180	2.733	3.490	13.36	15.51	17.54	20.09
9	2.088	2.700	3.325	4.168	14.68	16.92	19.02	21.67
10	2.558	3.247	3.940	4.865	15.99	18.31	20.48	23.21

Notation: for $0 < \alpha < 1$ and $\nu \in \mathbb{N}^*$, $\chi^2_\alpha(\nu)$ is the **critical value** for which

$$P(\chi^2 > \chi^2_\alpha(\nu)) = \alpha\,,$$

where $\chi^2 \sim \chi^2(\nu)$ follows a chi-squared distribution with ν degrees of freedom. We can find the value of $\chi^2_\alpha(\nu)$ in Table A.5 of the textbook (this table will be made available to you on the final exam, if needed).

For instance, when $\nu = 7$ and $\alpha = 0.95$, we have $\chi^2_{0.95}(7) = 2.167\,,$ therefore $P(\chi^2 > 2.167) = 0.95\,,$ where $\chi^2 \sim \chi^2(7)$, i.e. χ^2 has a chi-squared distribution with $\nu = 7$ degrees of freedom.

In other words, 95% of the area under the curve of the probability density function of $\chi^2(7)$ is found to the right of 2.167.

Sample Mean with Unknown Population Variance

Suppose that $Z \sim \mathcal{N}(0,1)$ and $V \sim \chi^2(\nu)$. If Z and V are independent, then the distribution of the random variable

$$T = \frac{Z}{\sqrt{V/\nu}}$$

is a **Student $t-$distribution with ν degrees of freedom**, which we denote by $T \sim t(\nu)$.

The probability density function of $t(\nu)$ is

$$f(x) = \frac{\Gamma(\nu/2 + 1/2)}{\sqrt{\pi\nu}\Gamma(\nu/2)(1 + x^2/\nu)^{\nu/2+1/2}}.$$

Theorem: let $X_1, \ldots, X_n$ be independent normal random variables with mean μ and standard deviation σ. Let $\overline{X}$ and S^2 be the sample mean and sample variance, respectively. Then the random variable

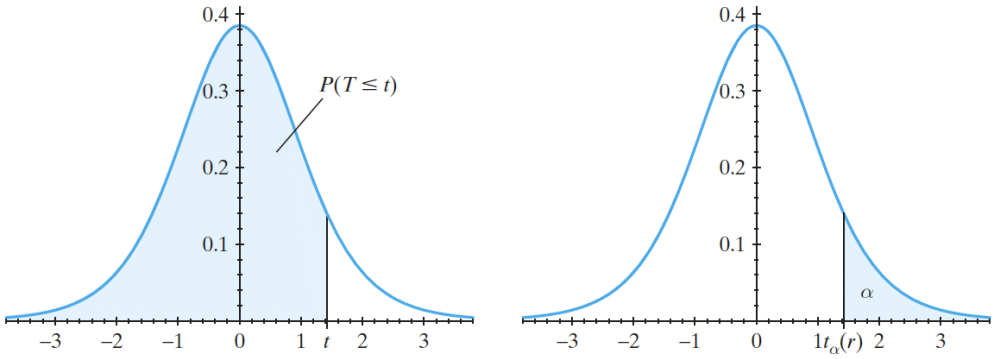
$$T = \frac{\overline{X} - \mu}{S/\sqrt{n}} \sim t(n-1),$$

follows a **Student t-distribution with $\nu = n - 1$ degrees of freedom**.

t-**Table:** let $t_\alpha(\nu)$ represent the critical t-value above which we find an area equal to α, i.e. $P(T > t_\alpha(\nu)) = \alpha$, where $T \sim t(\nu)$.

For all ν, the Student t-distribution is a symmetric distribution around zero, so we have $t_{1-\alpha(\nu)} = -t_\alpha$.

Table VI The t Distribution



$$P(T \le t) = \int_{-\infty}^{t} \frac{\Gamma[(r+1)/2]}{\sqrt{\pi r}\,\Gamma(r/2)(1+w^2/r)^{(r+1)/2}}\,dw$$

$$P(T \le -t) = 1 - P(T \le t)$$

	$P(T \le t)$						
	0.60	0.75	0.90	0.95	0.975	0.99	0.995
r	$t_{0.40}(r)$	$t_{0.25}(r)$	$t_{0.10}(r)$	$t_{0.05}(r)$	$t_{0.025}(r)$	$t_{0.01}(r)$	$t_{0.005}(r)$
1	0.325	1.000	3.078	6.314	12.706	31.821	63.657
2	0.289	0.816	1.886	2.920	4.303	6.965	9.925
3	0.277	0.765	1.638	2.353	3.182	4.541	5.841
4	0.271	0.741	1.533	2.132	2.776	3.747	4.604
5	0.267	0.727	1.476	2.015	2.571	3.365	4.032
6	0.265	0.718	1.440	1.943	2.447	3.143	3.707
7	0.263	0.711	1.415	1.895	2.365	2.998	3.499
8	0.262	0.706	1.397	1.860	2.306	2.896	3.355
9	0.261	0.703	1.383	1.833	2.262	2.821	3.250
10	0.260	0.700	1.372	1.812	2.228	2.764	3.169

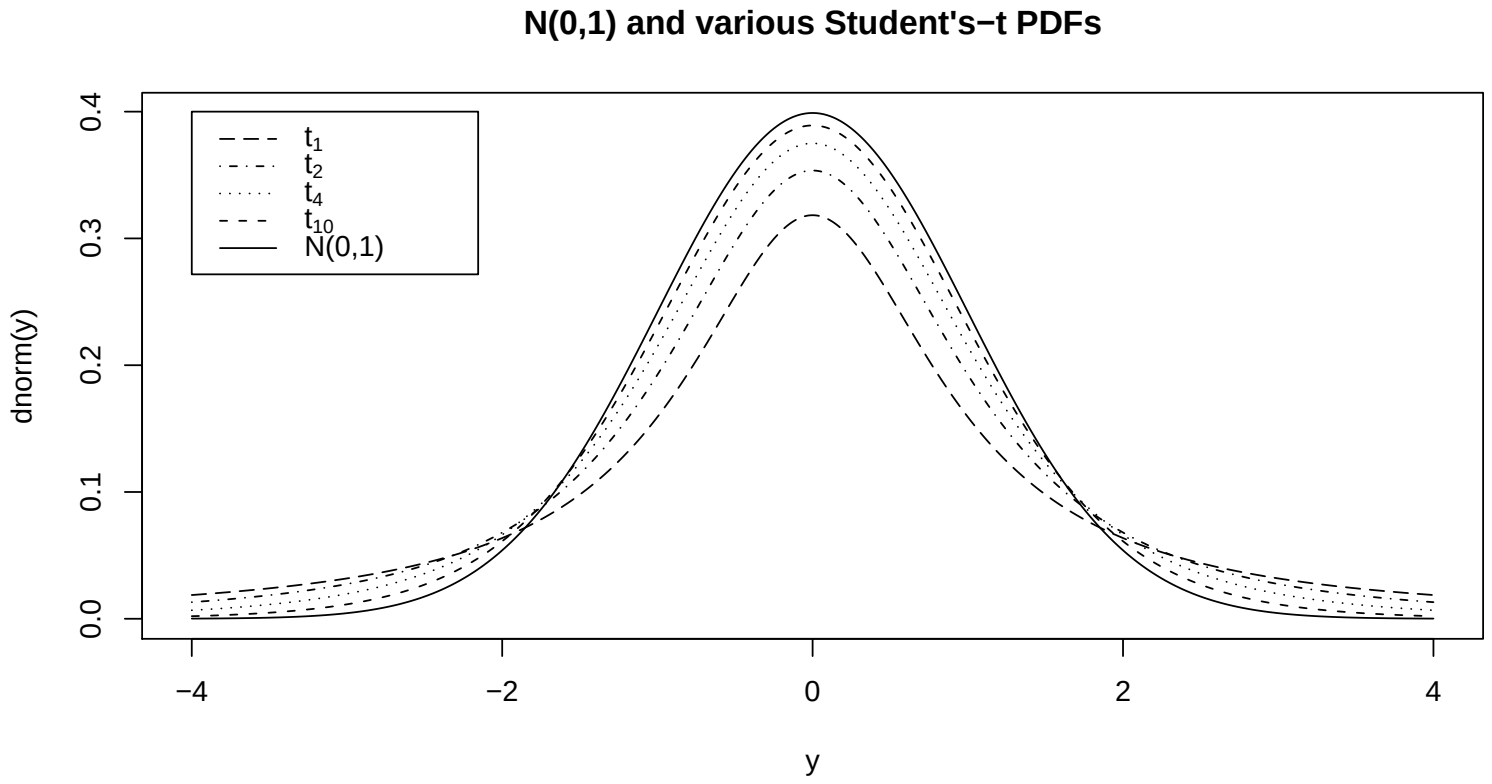
If $T \sim t(\nu)$, then for any $0 < \alpha < 1$, we have

$$\begin{aligned} P\left(-t_{\alpha/2}(\nu) < T < t_{\alpha/2}(\nu)\right) &= P\left(T < t_{\alpha/2}(\nu)\right) - P\left(T < -t_{\alpha/2}(\nu)\right) \\ &= 1 - P\left(T > t_{\alpha/2}(\nu)\right) - \left(1 - P\left(T > -t_{\alpha/2}(\nu)\right)\right) \\ &= 1 - P\left(T > t_{\alpha/2}(\nu)\right) - \left(1 - P\left(T > t_{1-\alpha/2}(\nu)\right)\right) \\ &= 1 - \alpha/2 - (1 - (1 - \alpha/2)) = 1 - \alpha\,, \end{aligned}$$

where the third equality follows by $t_{1-\alpha}(\nu) = -t_{\alpha}(\nu)$.

Consequently,

$$P\left(-t_{\alpha/2}(n-1) < \frac{\bar{X} - \mu}{S/\sqrt{n}} < t_{\alpha/2}(n-1)\right) = 1 - \alpha\,.$$



As $\nu \to \infty$, $t(\nu) \to \mathcal{N}(0, 1)$. This makes sense because the estimate S gets better at estimating σ when n increases.

Example: from the table, we can see that

$$P\left(T > 2.306\right) = 0.025 \Rightarrow P\left(T < -2.306\right) = 0.025\,,$$

where $T \sim t(8)$, so that $t_{0.025}(8) = 2.306$ and

$$\begin{aligned} P\left(|T| \leq 2.306\right) &= P\left(-2.306 \leq T \leq 2.306\right) \\ &= 1 - P\left(T < -2.306\right) - P\left(T > 2.306\right) = 0.95\,. \end{aligned}$$

The Student $t-$distribution will be helpful when the time comes to compute confidence intervals and to do hypothesis testing.

$F-$Distributions

Let $U \sim \chi^2(\nu_1)$ and $V \sim \chi^2(\nu_2)$. If U and V are independent, then the distribution of the random variable

$$F = \frac{U/\nu_1}{V/\nu_2}$$

is an F-**distribution with** ν_1 **and** ν_2 **degrees of freedom**, which we denote by $F \sim F(\nu_1, \nu_2)$.

The probability density function of $F(\nu_1, \nu_2)$ is

$$f(x) = \frac{\Gamma(\nu_1/2 + \nu_2/2)(\nu_1/\nu_2)^{\nu_1/2} x^{\nu_1/2-1}}{\Gamma(\nu_1/2)\Gamma(\nu_2/2)(1 + x\nu_1/\nu_2)^{\nu_1/2+\nu_2/2}}, \quad x \geq 0.$$

Table VII *continued*

$$P(F \le f) = \int_0^f \frac{\Gamma[(r_1 + r_2)/2](r_1/r_2)^{r_1/2} w^{r_1/2-1}}{\Gamma(r_1/2)\Gamma(r_2/2)(1 + r_1 w/r_2)^{(r_1+r_2)/2}}\, dw$$

α	$P(F \le f)$	Den. d.f. r_2	Numerator Degrees of Freedom, r_1									
			1	2	3	4	5	6	7	8	9	10
0.05	0.95	1	161.4	199.5	215.7	224.6	230.2	234.0	236.8	238.9	240.5	241.9
0.025	0.975		647.79	799.50	864.16	899.58	921.85	937.11	948.22	956.66	963.28	968.63
0.01	0.99		4052	4999.5	5403	5625	5764	5859	5928	5981	6022	6056
0.05	0.95	2	18.51	19.00	19.16	19.25	19.30	19.33	19.35	19.37	19.38	19.40
0.025	0.975		38.51	39.00	39.17	39.25	39.30	39.33	39.36	39.37	39.39	39.40
0.01	0.99		98.50	99.00	99.17	99.25	99.30	99.33	99.36	99.37	99.39	99.40
0.05	0.95	3	10.13	9.55	9.28	9.12	9.01	8.94	8.89	8.85	8.81	8.79
0.025	0.975		17.44	16.04	15.44	15.10	14.88	14.73	14.62	14.54	14.47	14.42
0.01	0.99		34.12	30.82	29.46	28.71	28.24	27.91	27.67	27.49	27.35	27.23
0.05	0.95	4	7.71	6.94	6.59	6.39	6.26	6.16	6.09	6.04	6.00	5.96
0.025	0.975		12.22	10.65	9.98	9.60	9.36	9.20	9.07	8.98	8.90	8.84
0.01	0.99		21.20	18.00	16.69	15.98	15.52	15.21	14.98	14.80	14.66	14.55
0.05	0.95	5	6.61	5.79	5.41	5.19	5.05	4.95	4.88	4.82	4.77	4.74
0.025	0.975		10.01	8.43	7.76	7.39	7.15	6.98	6.85	6.76	6.68	6.62
0.01	0.99		16.26	13.27	12.06	11.39	10.97	10.67	10.46	10.29	10.16	10.05

Theorem: If S_1^2 and S_2^2 are the sample variances of independent random samples of size n and m, respectively, taken from normal populations with variances σ_1^2 and σ_2^2, respectively, then

$$F = \frac{S_1^2/\sigma_1^2}{S_2^2/\sigma_2^2} \sim F(n-1, m-1)$$

follows an **F-distribution with $\nu_1 = n-1$ and $\nu_2 = m-1$ d.f.**

Notation: for $0 < \alpha < 1$ and $\nu_1, \nu_2 \in \mathbb{N}^*$, $f_\alpha(\nu_1, \nu_2)$ is the **critical value** for which $P(F > f_\alpha(\nu_1, \nu_2)) = \alpha$ where $F \sim F(\nu_1, \nu_2)$.

It can be shown that $f_{1-\alpha}(\nu_1, \nu_2) = \frac{1}{f_\alpha(\nu_2, \nu_1)}$.

We can find the value of $f_\alpha(\nu_1, \nu_2)$ in Table A.6 of the textbook. For instance, we have $f_{0.95}(6, 10) = \frac{1}{f_{0.05}(10,6)} = \frac{1}{4.06} = 0.246$.