8.1 Random Sampling

Basic terminology

Population and Sample

- **Population:** A well-defined collection of all possible observations with which we are concerned.
- For instance, the population of U.S. registered voters as of November 1 in the most recent presidential election year.
- "Population $f(x)$" means a population whose observations are values of random variable having distribution $f(x)$.
- **Sample:** a subset of a population. It is often denoted as $X_1, X_2, \ldots, X_n$.
- For instance, a random sample of size 1000 from the above list of U.S. registered voters.

Parameter and Statistic

- **Parameter:** characteristic of a population.
- For instance, p =percentage of Democratic voters in the above list of U.S. registered voters.
- **Statistic:** characteristic of a sample, an estimate of the parameter.
- For instance, $\hat{p}$ = percentage of Democratic voters in the above sample.
- **A Parameter is to a Population as a Statistic is to a Sample.**
- **In statistics, we often rely on a sample to draw inferences about the population.**

Example 1. Examples of population, sample, parameter, and statistic.

CONCERN	POPULATION	SAMPLE	PARAMETER/ STATISTIC
Election	All voters	Gallup poll	% of votes for candidate A
Cancer	All cancer patients	20 cancer patients	Average cancer size
Water safety	Water in the well	Water in test cube	bacterial counts in unit volume
Pseudorandom numbers generator	The complete sequence	Subsequence	Distribution of pseudorandom numbers
Unemployment	Entire labor force	One million force	rate of unemployment

Random Sample

- **Definition:** If $X_1, X_2, \ldots, X_n$ are n random observations from population $f(x)$, and are independent, that is, they have joint distribution

$$f(x_1, \ldots, x_n) = f(x_1) \cdots f(x_n)$$

then $X_1, X_2, \ldots, X_n$ is said to be a **random sample** of a **size** n from the population $f(x)$.

- When we use capital letters, we treat $X_1, X_2, \ldots, X_n$ as n independent random variables having the same distribution $f(x)$;
- When we use lower case letters, we treat $x_1, x_2, \ldots, x_n$ as the numerical values of the n independent random variables;

Example 2. The number of traffics during weekday 11:30am to 12:30pm at an intersection of a city has Poisson distribution with mean λ.

- A random sample of size 20, $X_1, X_2, \ldots, X_{20}$ are 20 independent random variables having the same Poisson distribution $P(\lambda)$.
- The sample values $x_1, x_2, \ldots, x_{20}$ are actually observed data values in 20 weekdays.

8.2 Some Important Statistics

Definition Statistic Any function of random variables constituting a random sample is called a **statistic**, which is free of any unknown parameter.

Example 0. The number of traffics during weekday 11:30am to 12:30pm at an intersection of a city has Poisson distribution with mean λ.

- The mean value of a random sample of size 20, $X_1, X_2, \ldots, X_{20}$ is a statistic:

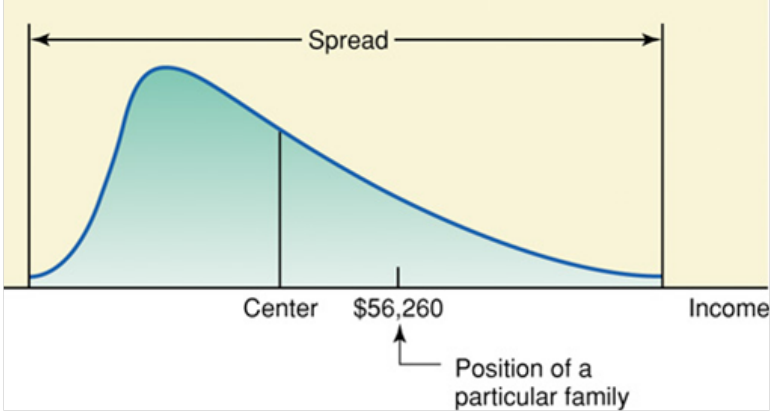
$$\bar{X} = \frac{1}{20}(X_1 + \cdots + X_{20})$$

  $P(\lambda)$.
- The mean value of the actual observed sample values $x_1, x_2, \ldots, x_{20}$ is denoted by

$$\bar{x} = \frac{1}{20}(x_1 + \cdots + x_{20})$$

Measure of Central Tendency for the Sample

Measures for a distribution



Mean is the average value

$$\text{Sample Mean} = \frac{\text{Sum of values}}{\text{Number of values}}$$

Let $X_1, X_2, \ldots, X_n$ be a sample of size n. Let $x_1, x_2, \ldots, x_n$ be the values of the sample data.

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

$$\bar{x} = \frac{\sum_{i=1}^{n} x_i}{n}$$

Properties of the Mean

- Uniqueness: For any data set, there is only one arithmetic mean.
- Simplicity: Easy to calculate.
- Non-robustness: A single extreme value in a sample could cause undesirable result. Isolated extreme values are called *outliers*.

Median is the middle value

(1) Sort the data values in increasing order.

(2) If the number of values is odd, then the middle term is the median.

(3) If the number of values is even, then the average of the two middle terms is the median.

Example 1: Find the median of each data set.

(a) Data set 1: 7, 2, -1, 5, 9, 2, 4.

(b) Data set 2: 1.2, 0.7, 3.5, 1.6, 0.3, 2.4.

Properties of the Median

- Uniqueness: For any data set, there is only one median.
- Simplicity: Easy to calculate.
- Robustness: It is not drastically affected by extreme values.

Mode is the most frequent value The mode of a dataset is the value that has the highest frequency.
Example 2: Find the mode of each data set.

(a) Find the mode of the 50 students status data.

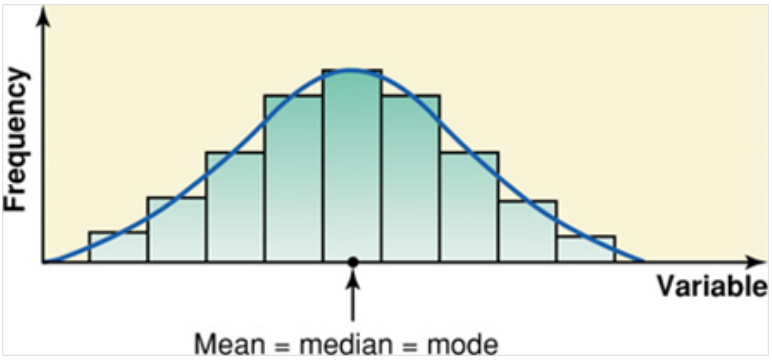
Table 2.2 **Status of 50 Students**

J	F	SO	SE	J	J	SE	J	J	J
F	F	J	F	F	F	SE	SO	SE	J
J	F	SE	SO	SO	F	J	F	SE	SE
SO	SE	J	SO	SO	J	J	SO	F	SO
SE	SE	F	SE	J	SO	F	J	SO	SO

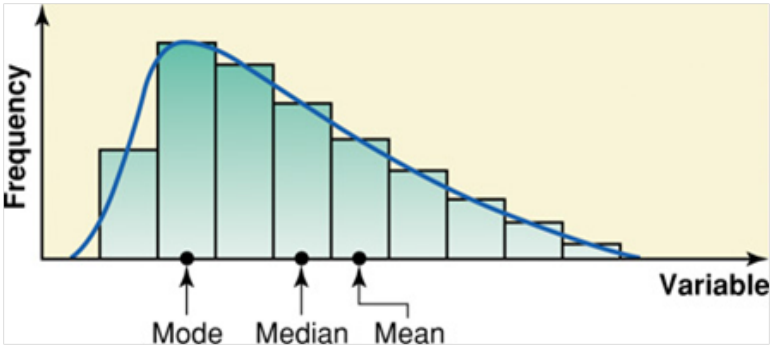
(b) The speeds(mph) of 8 cars stopped on I-95 for speeding: 77, 82, 74, 81, 79, 84, 74, 78.

(c) The ages of 10 randomly selected students are: 21, 19, 27, 22, 29, 19, 25, 21, 22, 30.

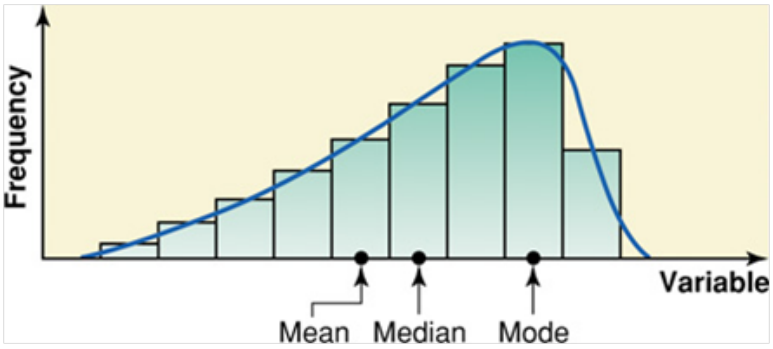
Relationships



Relationships



Relationships

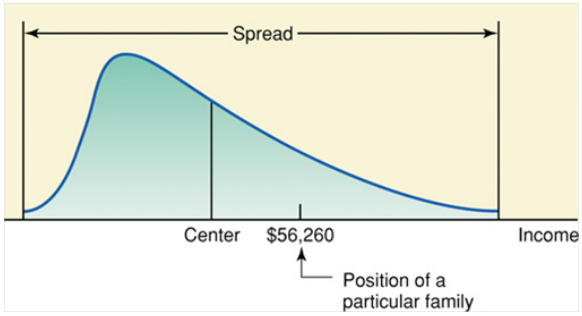


Trimmed mean Because mean value is not robust, easily impacted by outlier. If outliers present we can use either median or the following trimmed mean.
The $k\%$ trimmed mean is the mean of the data values after cutting off $k\%$ of the values from each end of the sorted data.
Example 3: The following are the money spent (in dollars) on books in 2015 by 10 randomly selected students from a small college.
890 1354 1861 1644 87 5403 1429 1993 938 2176
Find the 10% trimmed mean.
Solution The sorted data are
87 890 938 1354 1429 1644 1861 1993 2176 5403
Number of 10% value equals $n * k/100 = 10 * 10/100 = 1$. So we drop one value from each end and we have the trimmed data
890 938 1354 1429 1644 1861 1993 2176
Then the 10% trimmed mean is \$1535.625=(890 + 938+ 1354+ 1429+ 1644+ 1861 +1993+ 2176)/8.

Range and The Sample Variance

Range

$$\text{Range} = \text{Max} - \text{Min}$$



Variance Measures Dispersion from the mean Let $X_1, X_2, \ldots, X_n$ be a sample size n. Let $x_1, x_2, \ldots, x_n$ be the values of the sample data.

$$\text{Sample Variance } S^2 = \frac{\sum(X_i - \bar{X})^2}{n-1}$$

$$\text{The Value of Sample Variance } s^2 = \frac{\sum(x_i - \bar{x})^2}{n-1}$$

$$\text{The Value of Sample Standard Deviation } s = \sqrt{\frac{\sum(x_i - \bar{x})^2}{n-1}}$$

Short-Cut Formulas

$$\text{Sample Variance } s^2 = \frac{\sum x_i^2 - n\bar{x}^2}{n-1}$$

$$\text{Sample Standard Deviation } s = \sqrt{s^2}$$

Calculation using Short-Cut Formulas

Table 3.6

x	x^2
46.5	2162.25
18.0	324.00
16.0	256.00
7.8	60.84
7.2	51.84
$\Sigma x = 95.5$	$\Sigma x^2 = 2854.93$

n = 5

$\bar{x} = \frac{\sum x_i}{n} = 95.5/5 = 19.1$

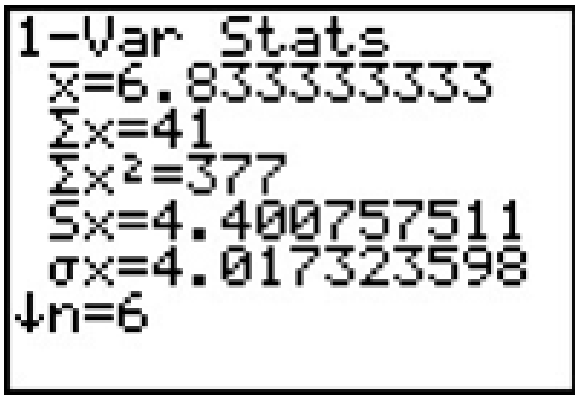
$s^2 = \frac{\sum x_i^2 - n\bar{x}^2}{n-1}$

= (2854.93 - 5*19.1^2)/4

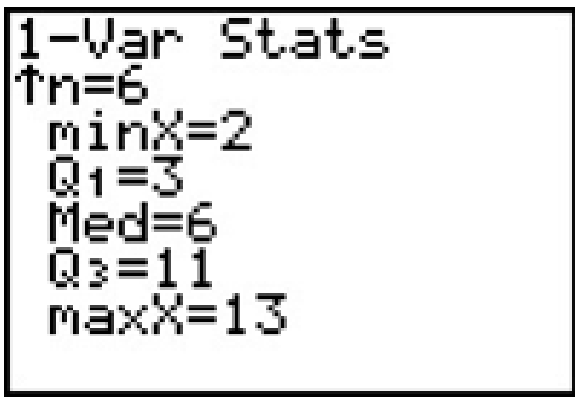
= 257.72

$s = \sqrt{257.72} = 16.0536$

Calculation using TI-83/84



Calculation using TI-83/84



Calculation using Excel

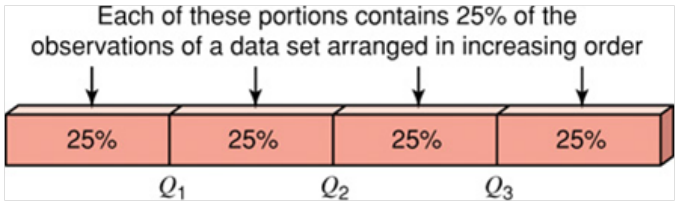
	A	B	C
1	Data	Average	
2			
3	2	=average(A3:A8)	
4	3		
5	5		
6	7		
7	11		
8	13		

Calculation using Excel

	A	B	C
1	Data	Average	
2			
3	2	6.833333	
4	3		
5	5		
6	7		
7	11		
8	13		

Quartiles & Interquartile Range

Quartiles



- First quartile: Q_1 the 25th percentile;
- Second quartile: Q_2 the 50th percentile, also the median;
- Third quartile: Q_3 the 75th percentile;
- 5 number summary: Min, Q_1, Q_2, Q_3, Max.

Interquartile-range:

$$IQR = Q_3 - Q_1;$$

Using TI-83

- First input the data as a list by Pressing button STAT and select "1: Edit..." then press ENTER.
- Second, Press button STAT, then choose "CALC" and "1-Var Stats" then ENTER.
- Press ENTER again you will get $\bar{x}$, $S_X = s$, $\sigma_X = \sigma$, and 5-number summary $minX$, Q_1, $Med = Q_2$, Q_3 and $maxX$.

Simplified Method for Finding Quartiles

Step 1. Sort the data in increasing order.

Step 2. Q_2 is the median of the complete data.

Step 3. Q_1 is the median of the subdataset that are smaller than or equal to Q_2, and

Step 4. Q_3 is the median of the subdataset that are greater than or equal to Q_2.

Example

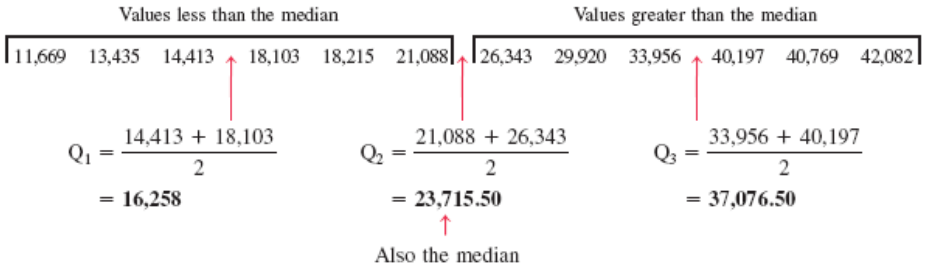
City	Number of Car Thefts
Phoenix-Mesa, Arizona	40,769
Washington, D.C.	33,956
Miami, Florida	21,088
Atlanta, Georgia	29,920
Chicago, Illinois	42,082
Kansas City, Kansas	11,669
Baltimore, Maryland	13,435
Detroit, Michigan	40,197
St. Louis, Missouri	18,215
Las Vegas, Nevada	18,103
Newark, New Jersey	14,413
Dallas, Texas	26,343

Source: National Insurance Crime Bureau.

(a) Find the values of the three quartiles. Where does the number of car thefts of 40,197 fall in relation to these quartiles?

(b) Find the interquartile range.

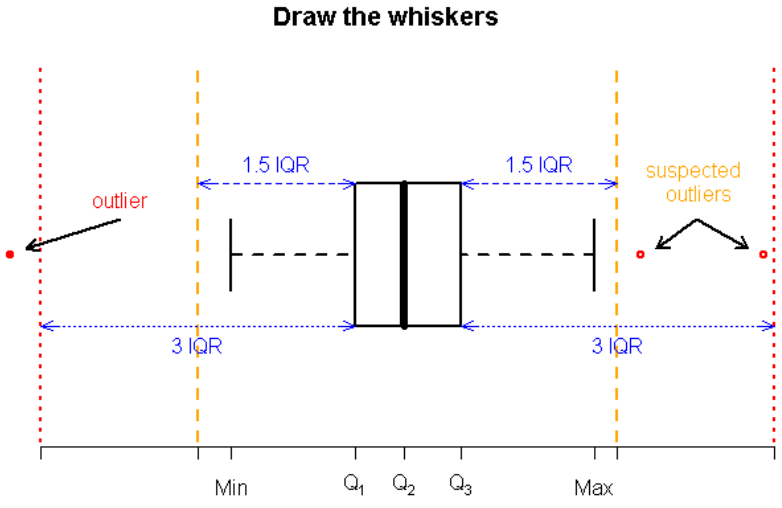
Solution of Example

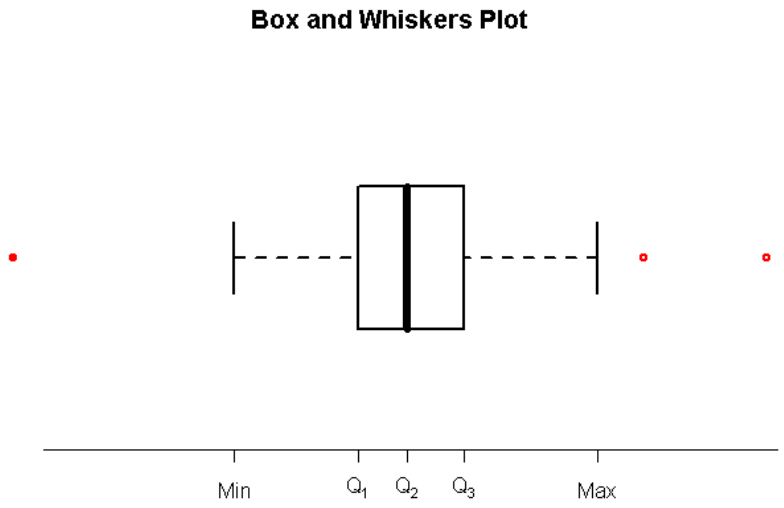


(a) The three quartiles are $Q_1 = 16,258$, $Q_2 = 23,715.5$, and $Q_3 = 37,076.5$. The number of car thefts of 40,197 falls in the top 25%.

(b) The interquartile range: $IQR = Q_3 - Q_1 = 37,076.50 - 16,258 = 20,818.50$ car thefts.

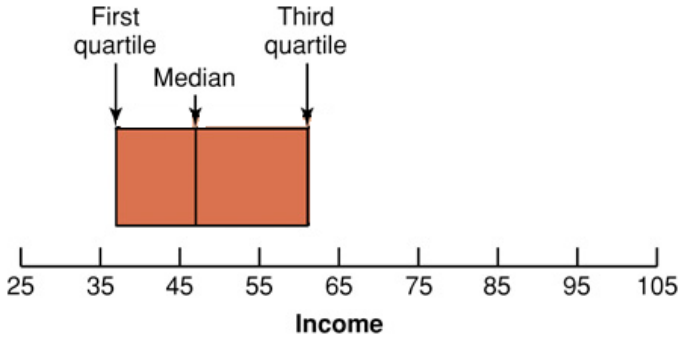
Box-and-Whisker Plot



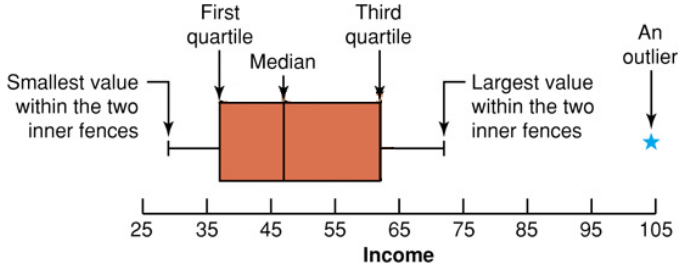


Example of Box-and-Whisker Plot Example: The following data are the incomes(in thousands of dollars) for a sample of 12 households.
35, 29, 44, 72, 34, 64, 41, 50, 54, 104, 39, 58
Construct a box-and-whisker plot.
Solution: (1) The sorted data:
29, 34, 35, 39, 41, 44, 50, 54, 58, 64, 72, 104
(2) $Q_2 = (44 + 50)/2 = 47$, $Q_1 = (35 + 39)/2 = 37$, $Q_3 = (58 + 64)/2 = 61$
(3) $IQR = Q_3 - Q_1 = 61 - 37 = 24$.
(4) $1.5 \times IQR = 1.5 \times 24 = 36$.
(4) $3 \times IQR = 3 \times 24 = 72$.

Example of Box-and-Whisker Plot



Example of Box-and-Whisker Plot



Quantiles

Definition:
Sample Quantile: A quantile of a sample is a value, $q(p)$, for which a specific proportion p of the data values is less than or equal to $q(p)$.
Specifically, Let $x_1, \ldots, x_n$ be the sample data. Then $q(p)$ is the smallest data value such that the proportion

of the data values less than or equal to $q(p)$ is at least p.

Finding Quantiles $q(p)$

Step 1. Sort the data in increasing order: $y_{(1)} \leqslant y_{(2)} \leqslant \cdots \leqslant y_{(n)}$;

Step 2. Calculate $m = pn$;

Step 3. If m is integer, then $q(p)$ is the mth term $y_{(m)}$;

Step 4. If m is NOT integer, say $m = i + r$ where i is an integer and r is a fraction, then

$$q(p) = y_i + r(y_{i+1} - y_i) = (1 - r)y_i + ry_{i+1}.$$

Note: $y_{(i)}$ is $q(i/n)$.
MINITAB uses the same formula. But some other software, e.g. TI-83, Excel and R, use different formula to calculate percentiles. For large n, they are close.

Example 1.: The following data are the incomes(in thousands of dollars) for a sample of 12 households.
35, 29, 44, 72, 34, 64, 41, 50, 54, 104, 39, 58

(a) Find and interpret the 15% quantile $q(0.15)$;

(b) Find and interpret the 75% quantile $q(0.75)$;

Definition:
Population Quantile: A quantile of a population with pmf/pdf $f(x)$ is a value, $q_f(p)$, for which the probability that any population data value is less than or equal to $q_f(p)$ is about p, that is

$$P(X \leqslant q_f(p)) \approx p, \quad q_f(p) = \min\{x : P(X \leqslant x) \geqslant p\}.$$

For discrete distribution

$$\sum_{x \leqslant q_f(p)} f(x) \approx p$$

For continuous distribution

$$\int_{-\infty}^{q_f(p)} f(x)dx = p, \quad \text{i.e.} \quad q_f(p) = F^{-1}(p)$$

Finding Population Quantiles $q_f(p)$
Example 2. Let the population distribution of X have pmf

$$f(x) = \frac{x}{10}, \quad x = 1, 2, 3, 4$$

Find $q_f(0.1)$ and $q_f(0.6)$.
Example 3. Let the population distribution of X have pdf

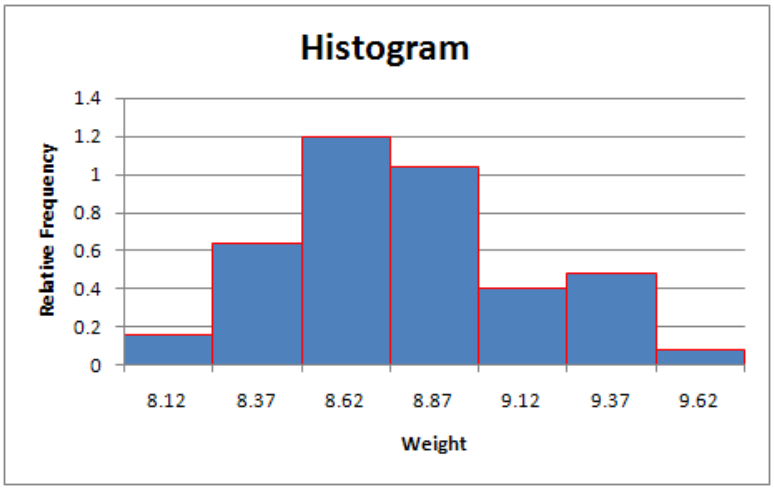
$$f(x) = \begin{cases} e^{-x}, & x > 0; \\ 0, & \text{elsewhere.} \end{cases}$$

Find $q_f(p)$ for $0 < p < 1$.

Frequency Table and Histogram

1. Find Min & Max values and **range** $R = Max - Min$;
2. Find k non-overlapping intervals (class intervals) of equal length h by the endpoints (**class boundaries**)
$$c_0 < c_1 < c_2 < \cdots < c_{k-1} < c_k$$
The endpoints should contain one more decimal place than the data values and $c_0 \lessapprox Min < Max \lessapprox c_k$.
3. Find the **class mark** for each class: the midpoint of the class interval: $m_i = \frac{c_{i-1}+c_i}{2}$

Frequency Table					
	Class	Interval	Frequency	Rel. Freq.	Class
i	c(i-1)	c(i)	fi	h(x)	Mark
1	7.995	8.245	2	0.16	8.12
2	8.245	8.495	8	0.64	8.37
3	8.495	8.745	15	1.2	8.62
4	8.745	8.995	13	1.04	8.87
5	8.995	9.245	5	0.4	9.12
6	9.245	9.495	6	0.48	9.37
7	9.495	9.745	1	0.08	9.62



4. Calculate relative frequency (density) for each class

$$h(x) = \frac{f_i}{n(c_i - c_{i-1})}, \text{ for } c_{i-1} < x \leqslant c_i, \quad i = 1, 2, \ldots, k.$$

5. Suggested $k \approx R/h$, $h = 2IQR/n^{1/3}$.

Relative Frequency

Example 1. (Exercise 3-1-4) The weights (in grams) of 50 nails:

```
8.05 8.31 8.51 8.56 8.66 8.76 8.85 8.90 9.20 9.34
8.24 8.36 8.51 8.57 8.69 8.79 8.85 8.93 9.21 9.40
8.27 8.38 8.51 8.58 8.69 8.79 8.85 8.98 9.21 9.41
8.27 8.41 8.55 8.58 8.71 8.82 8.88 9.08 9.25 9.42
8.29 8.43 8.56 8.59 8.73 8.82 8.88 9.15 9.26 9.63
```

1. $n = 50$, $Min = 8.05$, $Max = 9.63$, $R = Max - Min = 1.58$, $IQR = 0.4475$;
2. $h = \lceil 2\frac{IQR}{n^{1/3}} \rceil = 0.25$. $k = \lceil \frac{R}{h} \rceil = 7$.
3. Choose $c_0 = 7.995$. $c_i = c_0 + ih$, $i = 1, \ldots, k$. The class boundaries are 7.995 8.245 8.495 8.745 8.995 9.245 9.495 9.745
4. Calculate frequency f_i and density $h(x)$
5. The class marks are 8.12 8.37 8.62 8.87 9.12 9.37 9.62

Example 1. Heights of 5000 female students

Lower Bound	Upper Bound	Frequency	Class	Rel. Freq.
XL	XU	f	Mark	Density
59	60	0	59.5	0
60	61	90	60.5	0.018
61	62	170	61.5	0.034
62	63	460	62.5	0.092
63	64	750	63.5	0.15
64	65	970	64.5	0.194
65	66	760	65.5	0.152
66	67	640	66.5	0.128
67	68	440	67.5	0.088
68	69	320	68.5	0.064
69	70	220	69.5	0.044
70	71	180	70.5	0.036
71	72	0	71.5	0
Total		5000		1

Relative frequency density for class i with frequency f_i is

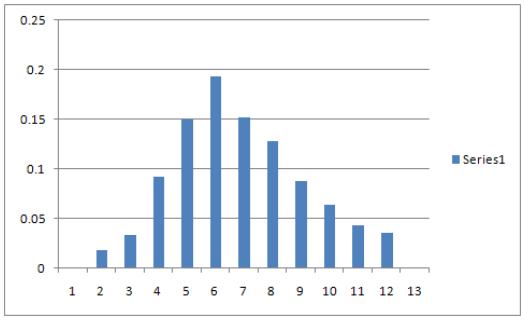
$$\frac{f_i}{Nw}$$

where w is the class width.

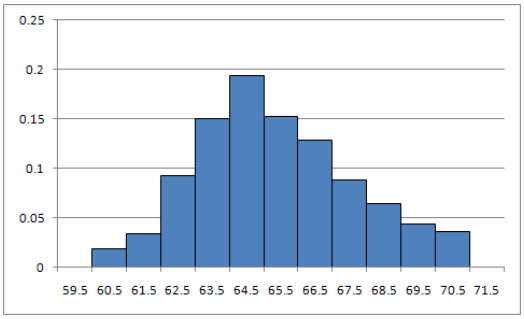
Histogram and Polygon Using Excel Step 1: Make a frequency table in Excel:

Lower Bound	Upper Bound	Frequency	Class	Rel. Freq.
XL	XU	f	Mark	Density
59	60	0	59.5	0
60	61	90	60.5	0.018
61	62	170	61.5	0.034
62	63	460	62.5	0.092
63	64	750	63.5	0.15
64	65	970	64.5	0.194
65	66	760	65.5	0.152
66	67	640	66.5	0.128
67	68	440	67.5	0.088
68	69	320	68.5	0.064
69	70	220	69.5	0.044
70	71	180	70.5	0.036
71	72	0	71.5	0
Total		5000		1

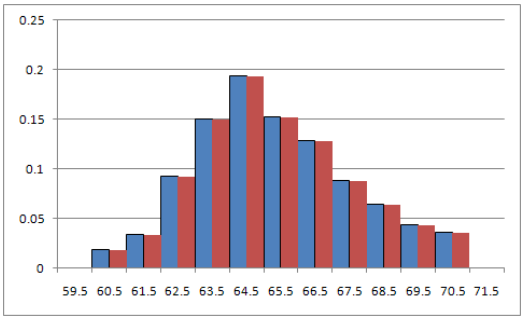
Histogram and Polygon Using Excel Step 2: Highlight column "Relative Frequency Density", then insert a 2-d column chart:



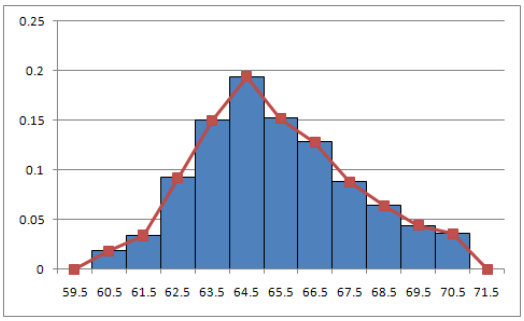
Histogram and Polygon Using Excel Step 3: Edit the chart: reduce the gap width to 0% and replace "the Horizontal (Category) Axis Lable" with the "class mark" column.



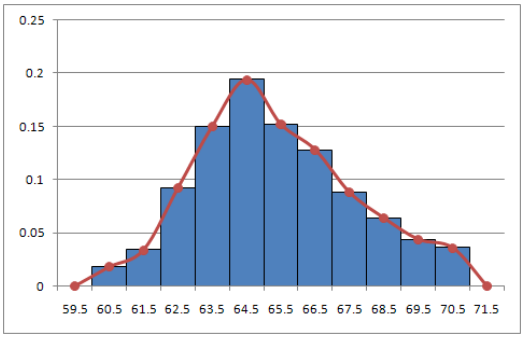
Histogram and Polygon Using Excel Step 4: Right click the chart and choose "Select Data...", add a new series using "Relative Frequency Density" column.



Histogram and Polygon Using Excel Step 5: Right click the chart on the new column chart(one of the dark red bars) and choose "Change Series Chart Type...", select a line chart.



Relative Frequency Density Histogram and Polygon

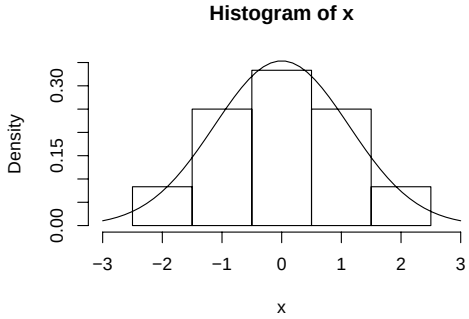


Total area of all shaded rectangles equals 1.
The smooth curve is called the *probability density function* of random variable X, Height of randomly selected female student.

8.3 Sampling Distributions

Definition:
Sampling distribution is the probability distribution of a **statistic**.
For example, the probability distribution of the sampling mean $\bar{X}$ is called the sampling distribution of the mean.

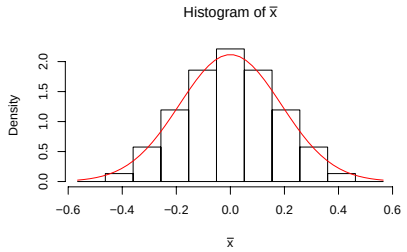
Example 1. Suppose we have a population containing the following values:
-2, -1, -1, -1, 0, 0, 0, 0, 1, 1, 1, 2
Distribution of the data is close to: $N(0, 7/6)$.



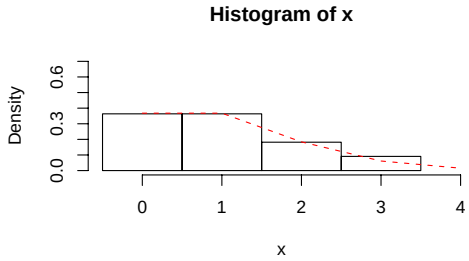
Distribution of All Sample Means
Let $X_1, X_2, \ldots, X_n$ be a sample of size $n = 9$.

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

There are $\binom{12}{9} = 220$ sample means. The distribution of all the sample means is also normal $N(0, 0.0354)$.



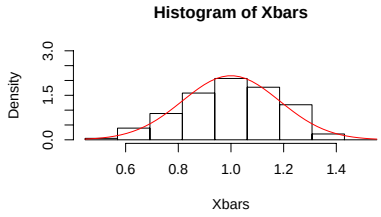
Example 2. Suppose we have a skewed population containing the following values:
0, 0, 0, 0, 1, 1, 1, 1, 2, 2, 3
Distribution of the data is close to Poisson: $P(1)$.



Distribution of All Sample Means
Let $X_1, X_2, \ldots, X_n$ be a sample of size $n = 8$.

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

There are $\binom{11}{8} = 165$ sample means. The distribution of all the sample means is also close to normal $N(1, 0.034)$.



8.4 Sampling Distribution of Means

Central Limit Theorem

Let $X_1, X_2, \ldots, X_n$ be a sample of size n. That is, $X_1, X_2, \ldots, X_n$ are independent random variables having the same distribution with mean μ and variance σ^2.

$$\overline{X} = \frac{1}{n}\sum_{i=1}^{n} X_i = \frac{1}{n}(X_1 + X_2 + \cdots + X_n)$$

$$E(\overline{X}) = \mu, \quad Var(\overline{X}) = \frac{\sigma^2}{n}$$

If the sample is from $N(\mu, \sigma^2)$, then exactly

$$\overline{X} \sim N\big(\mu, \tfrac{\sigma^2}{n}\big) \quad \text{and} \quad \tfrac{\overline{X}-\mu}{\sigma/\sqrt{n}} \sim N(0,1).$$

What if the sample is not from $N(\mu, \sigma^2)$?

Central Limit Theorem If $\overline{X}$ is the sample mean of a random sample $X_1, X_2, \ldots, X_n$ of size n from a distribution, discrete or continuous, with mean μ and variance σ^2, then the distribution of

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

approaches the standard normal distribution $N(0,1)$ as $n \to \infty$, i.e.,

$$\lim_{n\to\infty} P(Z \leqslant z) = \Phi(z).$$

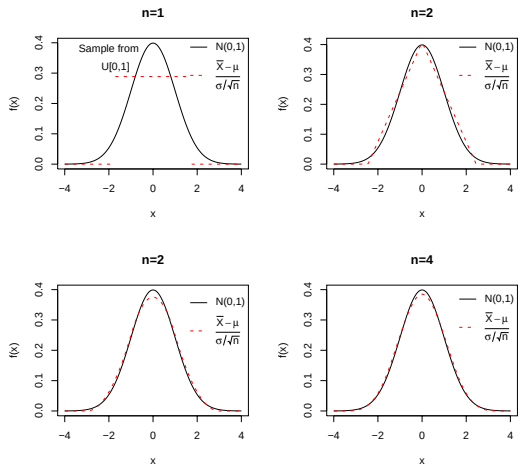
So if $n \geqslant 30$ (or smaller for symmetric distribution), then the distribution of $\overline{X}$ is approximately $N(\mu, \sigma^2/n)$ and

$$P(a < Z \leqslant b) \approx \Phi(b) - \Phi(a), \quad a < b.$$

Example 1. Let $\overline{X}$ be the sample mean of a random sample $X_1, X_2, \ldots, X_n$ of size n from $U[0,1]$, the uniform distribution on $[0,1]$.

$$\mu = \mathrm{E}(X_1) = \tfrac{1}{2}, \quad \sigma^2 = \mathrm{Var}(X_1) = \tfrac{1}{12}.$$

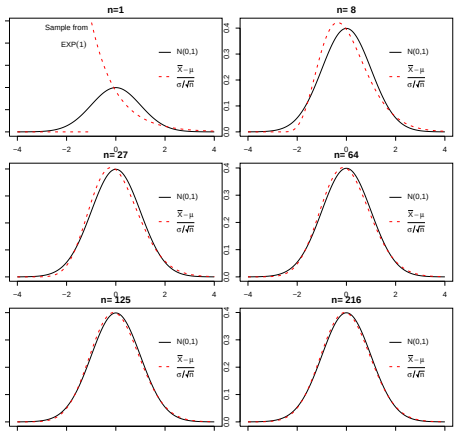
$$Z = \frac{\overline{X}-\mu}{\sigma/\sqrt{n}} = \sqrt{12n}(\overline{X} - \tfrac{1}{2})$$



Example 2. Let $\overline{X}$ be the sample mean of a random sample $X_1, X_2, \ldots, X_n$ of size n from $EXP(\theta)$, the exponential distribution with

$$\mu = \mathrm{E}(X_1) = \theta, \quad \sigma^2 = \mathrm{Var}(X_1) = \theta^2.$$

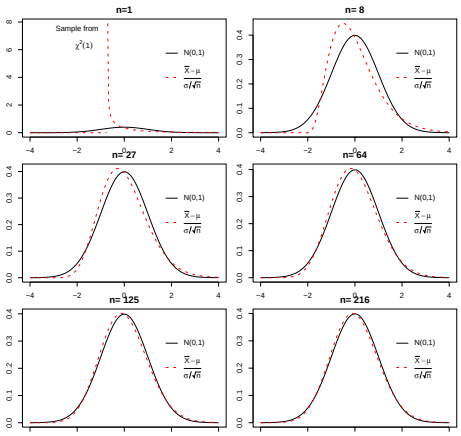
$$Z = \frac{\overline{X}-\mu}{\sigma/\sqrt{n}} = \frac{\overline{X}-\theta}{\theta/\sqrt{n}}$$



Example 3. Let $\overline{X}$ be the sample mean of a random sample $X_1, X_2, \ldots, X_n$ of size n from $\chi^2(k)$, the chi-square distribution with

$$\mu = \mathrm{E}(X_1) = k, \quad \sigma^2 = \mathrm{Var}(X_1) = 2k.$$

$$Z = \frac{\overline{X}-\mu}{\sigma/\sqrt{n}} = \frac{\overline{X}-k}{2k/\sqrt{n}}$$



The central limit theorem is also valid for discrete distribution.

Example 4. Let $\overline{X}$ be the sample mean of a random sample $X_1, X_2, \ldots, X_n$ of size n from $b(1, p)$ with

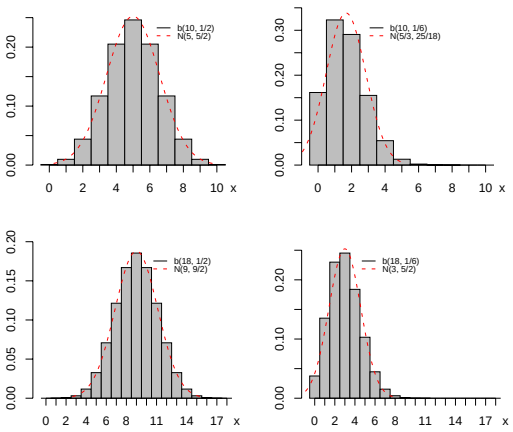
$$\mu = p, \quad \sigma^2 = p(1-p).$$

$$Y = \sum_{i=1}^{n} X_i \sim b(n, p).$$

By the CLT,

$$\overline{X} \sim N(\mu, \sigma^2/n), \text{ approximately.}$$

$$Y = n\overline{X} \sim N(n\mu, n^2\sigma^2/n) = N[np, np(1-p)], \text{ approximately.}$$



Example 5. If a certain machine makes resistors have a mean resistance of 40 ohms and standard deviation of 2 ohms, what is the probability that a random sample of 36 of these resistors will have a average resistance of more than 40.5 ohms.

Solution of Example 5. $n = 36$, $\mu = 40$, $\sigma = 2$. By the CLT, $\overline{X}$ has normal distribution with mean $\mu_{\bar{x}} = \mu = 40$ and standard deviation $\sigma_{\bar{x}} = \sigma/\sqrt{n} = 2/\sqrt{36} = 2/6 = 1/3$. The z value of $x = 40.5$ is

$$z = \frac{x-\mu_{\bar{x}}}{\sigma_{\bar{x}}} = \frac{40.5-40}{1/3} = 3(0.5) = 1.5$$

Using Table A.3

$$P(\overline{X} > 40.5) = P(Z > 1.5) = 1 - P(Z \leqslant 1.5) = 1 - 0.9332 = 0.0668$$

Inferences on the Population Mean

Example 6.[Example 5. (Cont.)] An observed sample of 36 of these resistors indicates a sample average of 39.1 ohms. Does this sample information appear to support or refute the conjecture that $\mu = 40$ ohms?

Solution of Example 6. If the conjecture $\mu = 40$ is true, then by the CLT $\overline{X}$ with $n = 36$ is approximately normal with mean $\mu = 40$ and standard deviation $\sigma/\sqrt{n} = 2/\sqrt{36} = 1/3$.
How likely can the value of $\overline{X}$ be as far away from the center $\mu = 40$ as the observed $\bar{x} = 39.1$?
That is, if $\mu = 40$,

$$P(|\overline{X} - 40| \geqslant |39.1 - 40|) = P(|\overline{X} - 40| \geqslant 0.9) = ?$$

By the CLT, if $\mu = 40$, $Z = \frac{\overline{X}-\mu}{\sigma/\sqrt{n}} = \frac{\overline{X}-40}{2/\sqrt{36}} = 3(\overline{X} - 40)$.

$$P(|\overline{X} - 40| \geqslant 0.9) = P(|Z| \geqslant 3(0.9)) = P(|Z| \geqslant 2.7)$$

$$= P(Z \geqslant 2.7) + P(Z \leqslant -2.7)$$

$$= P(Z \geqslant 2.7) + P(-Z \geqslant 2.7) = 2P(Z \geqslant 2.7)$$

$$= 2[1 - P(Z < 2.7)] = 2(1 - 0.996533) \approx 0.007$$

One would experience by chance that an $\bar{x}$ is 0.9 ohms from the mean in only 7 in 1000 samples of size 36. This sample is an evidence against the conjecture $\mu = 40$ ohms.

- The probability $P(|\overline{X} - 40| \geqslant 0.9|\mu = 40)$ is called the p-value of the sample mean $\bar{x} = 39.1$.
- Under the condition that the conjecture or hypothesis $\mu = 40$ is true, p-value is the probability that we can observed an $\bar{X}$ as extreme as $\bar{x} = 39.1$.
- p-value is NOT the probability that the conjecture or hypothesis $\mu = 40$ is true.

Sampling Distribution of Difference between Two Means

Exact Distribution of Difference between Two Means

If $\overline{X}_1$ and $\overline{X}_2$ are the sample means of independent random samples of size n_1 and n_2 from two normal distributions with means μ_1 and μ_2 and variances σ_1^2 and σ_2^2, respectively, then the distribution of $\overline{X}_1 - \overline{X}_2$ is exactly normal with mean $\mu_1 - \mu_2$ and variance $(\sigma_1^2/n_1) + (\sigma_2^2/n_2)$. So the distribution of

$$Z = \frac{(\overline{X}_1 - \overline{X}_2) - (\mu_1 - \mu_2)}{\sqrt{(\sigma_1^2/n_1) + (\sigma_2^2/n_2)}}$$

exactly standard normal $N(0, 1)$.

Central Limit Theorem for Difference between Two Means If $\overline{X}_1$ and $\overline{X}_2$ are the sample means of independent random samples of size n_1 and n_2 from two nonnormal distributions, discrete or continuous, with means μ_1 and μ_2 and variances σ_1^2 and σ_2^2, respectively, then the distribution of

$$Z = \frac{(\overline{X}_1 - \overline{X}_2) - (\mu_1 - \mu_2)}{\sqrt{(\sigma_1^2/n_1) + (\sigma_2^2/n_2)}}$$

is approximately standard normal distribution $N(0, 1)$ if $n_1, n_2 \geqslant 30$.

Example 7. Two independent samples of size 18 are selected from two types of paints, A and B. The average drying time, in hours, is recorded for each sample. Assume that the the populations are normal with the same means and the population standard deviations are both known to be 1.0.
Find the probability that the difference between the two means is greater than 1.0.

Solution of Example 7. $\mu_A = \mu_B$, $\sigma_A = \sigma_B = 1$, and $n_A = n_B = 18$. Because the populations are normal

$$Z = \frac{(\overline{X}_A - \overline{X}_B) - (\mu_A - \mu_B)}{\sqrt{(\sigma_A^2/n_A) + (\sigma_B^2/n_B)}} = \frac{(\overline{X}_1 - \overline{X}_2) - 0}{\sqrt{(1/18) + (1/18)}} = 3(\overline{X}_1 - \overline{X}_2)$$

is exactly standard normal.

So

$$P(|\overline{X}_A - \overline{X}_B| \geqslant 1) = P(|Z| \geqslant 3) = 2(1 - P(Z < 3)) = 2(1 - 0.9987)$$

$$= 2(0.0013) = 0.0026$$

One would experience by chance that a difference between the two means is bigger than 1 in only 2.6 in 1000 pairs of samples of size 18.

Example 7.(Cont.) Suppose a difference of 1.0 in means was observed in real samples.
(a) Does this seem to be a reasonable results if the two population mean drying times truly are equal?
No.
(b) If someone selected 10,000 pairs of samples of size 18 under the condition that $\mu_A = \mu_B$, in how many of these 10,000 experiments would there be a difference $\bar{x}_A - \bar{x}_B$ is as large as 1.0?
26.

What if σ^2 is unknown?

- In the previous section, we assume σ is known. By the CLT, the distribution of

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

  is approximately standard normal.

- However, in application, we rarely know σ^2. We estimate σ^2 by $S^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \bar{X})^2$.

- What is the sampling distribution of S^2?

- What is the sampling distribution of $\frac{\overline{X} - \mu}{S/\sqrt{n}}$?

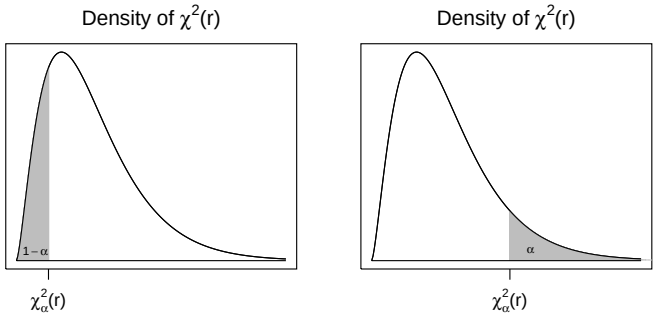
8.5 Sampling Distribution of S^2.

Theorem 1. If $X_1, X_2, \ldots, X_n$ is a random sample of size n from $N(\mu, \sigma^2)$, then

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

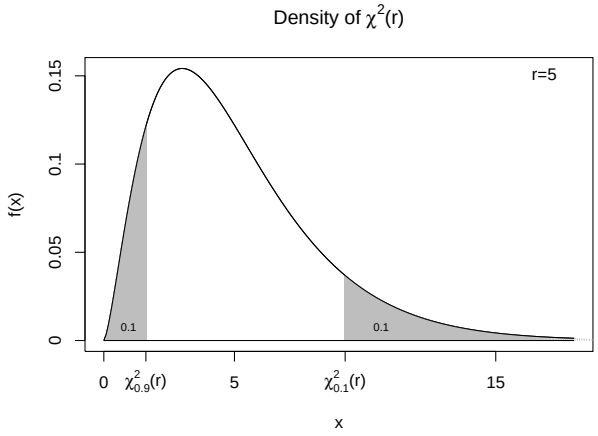
See Section 6.8 for information about χ^2 distribution.

Percentage Point of χ^2-Distribution The Percentage points of the χ^2-distribution are given in Table A-5.



Percentage Point of χ^2-Distribution

The Percentage points of the χ^2-distribution are given in Table A-5.



Example 1. A manufacturer of car batteries guarantees that his batteries will last, on the average, 3 years with a standard deviation of 1 year. If five of these batteries have lifetimes of 1.9, 2.4, 3.0, 3.5, and 4.2 years, is the manufacturer still convinced that his batteries have a standard deviation of 1 year? Assume that the battery lifetime follows a normal distribution.

Solution: $\bar{x} = \frac{1.9+2.4+3.0+3.5+4.2}{5} = 3,$

$$\sum x_i^2 = 1.9^2 + 2.4^2 + 3.0^2 + 3.5^2 + 4.2^2 = 48.26$$

$$s^2 = \frac{1}{n-1}\left(\sum x_i^2 - n\bar{x}^2\right) = \frac{1}{4}[48.26 - (5)(3^2)] = 0.815$$

Since $\sigma = 1$ $\chi^2 = \frac{(n-1)S^2}{\sigma^2} = (5-1)(0.815) = 3.26.$
Using Table A.5,

$$P(\chi^2 \leqslant 3.26) = 1 - P(\chi^2 > 3.26) > 1 - P(\chi^2 > 2.195) = 0.3$$

Using Excel

$$P(\chi^2 \leqslant 3.26) = 0.485$$

So it is likely to have an observed χ^2 as small as 3.26. There is no strong evidence against the hypothesis $\sigma = 1$.

Degrees of Freedom

Degrees of Freedom–Measure of Sample Information

- If μ is known, we estimate σ^2 by

$$\hat{\sigma}^2 = \frac{1}{n}\sum_{i=1}^{n}(X_i - \mu)^2$$

Let $Y_i = X_i - \mu$, $i = 1, 2, \ldots, n$. Y_i's are n independent normal r.v.'s. So

$$\hat{\sigma}^2 = \frac{1}{n}\sum_{i=1}^{n} Y_i^2$$

We have n independent *pieces of information* to estimate σ^2.

- If μ is unknown, we first use one independent piece of information to estimate μ by $\bar{X}$ and then estimate σ^2 by

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

So S^2 has $n-1$ degrees of freedom to estimate σ^2.

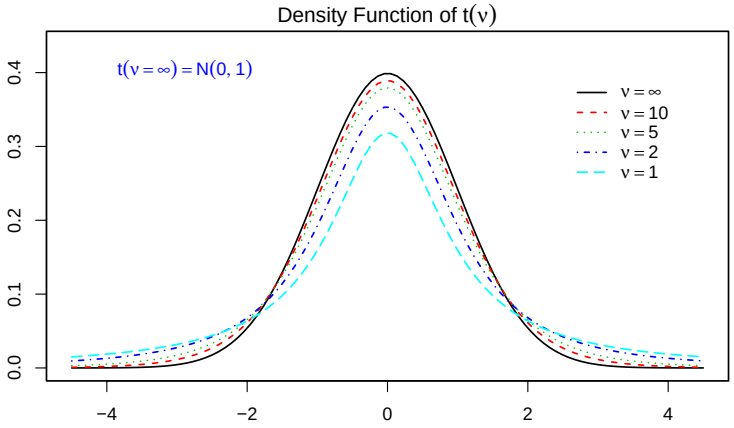
8.6 t-Distribution

Sampling Distribution of $\overline{X}$ Theorem 2. If $X_1, X_2, \ldots, X_n$ is a random sample of size n from $N(\mu, \sigma^2)$, then
(a) $\overline{X} = \frac{1}{n}\sum_{i=1}^{n} X_i$ and $S^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \overline{X})^2$ are independent.
(b) $T = \sqrt{n}(\overline{X} - \mu)/S$ has a **Student's t distribution** with d.f. $n-1$ and p.d.f.

$$h(t) = \frac{\Gamma\left(\frac{n}{2}\right)}{\Gamma\left(\frac{n-1}{2}\right)}\left(1 + \frac{t^2}{n-1}\right)^{-n/2}, \quad -\infty < t < \infty$$

Graph of Standard normal and Student's t-Distribution

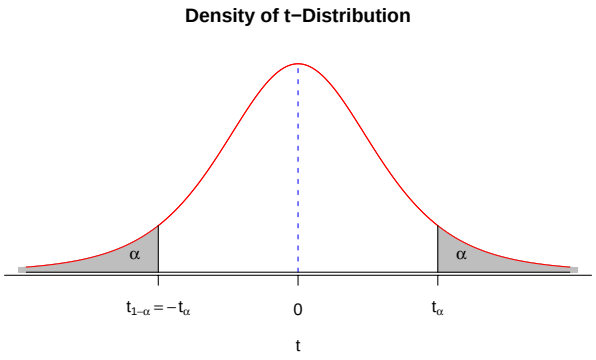
Graph of Student's t-Distribution Graph of Student's t-Distribution



Similar to $N(0, 1)$ and Symmetric about 0.

Percentage Point of t-Distribution

The Percentage points of the t-distribution are given in Table A-4.



Example 2. A certain machine makes resistors have a conjectured mean resistance of 40 ohms but unknown standard deviation. Assume resistance has normal distribution. An observed sample of 36 of these resistors indicates a sample average of 39.1 ohms and sample standard deviation of 1.7 ohms. Does this sample information appear to support or refute the conjecture that $\mu = 40$ ohms?

Solution of Example 2. If the conjecture $\mu = 40$ is true, then by the t statistic

$$T = \sqrt{n}(\overline{X} - \mu)/S$$

has a **Student's t distribution** with d.f. 35.
The observed value of $|T|$ is

$$|t| = \frac{\sqrt{n}|\bar{x} - \mu|}{s} = \frac{\sqrt{36}|39.1 - 40|}{1.7} = 3.18$$

How likely can the value of the t statistic T be as far away from its center 0 as the observed $t = 3.18$? That is, if $\mu = 40$,

$$P(|T| \geqslant 3.18) = ?$$
$$P(|T| \geqslant 3.18) = P(T \geqslant 3.18) + P(T \leqslant -3.18)$$
$$= P(T \geqslant 3.18) + P(-T \geqslant 3.18) = 2P(T \geqslant 3.18)$$
$$\approx 2[1 - P(Z < 3.18)] = 2(1 - 0.9984605) \approx 0.003079$$

One would experience by chance that a t is 3.18 from 0 in only 3 in 1000 samples of size 36. This sample is an evidence against the conjecture $\mu = 40$ ohms.

- The probability $P(|T| \geqslant 3.18|\mu = 40)$ is called the p-value of the t statistic $t = 3.18$.
- Under the condition that the conjecture or hypothesis $\mu = 40$ is true, p-value is the probability that we can observed a T as extreme as $t = 3.18$.
- p-value is NOT the probability that the conjecture or hypothesis $\mu = 40$ is true.

8.7 F-Distribution

Sampling distribution of s_1^2/s_2^2

- Let $x_{11}, \ldots, x_{1n_1}$ be a sample from normal population with variance σ_1^2, and $x_{21}, \ldots, x_{2n_2}$ be a sample from normal population with variance σ_2^2.
- Assume the two samples are independent and have sample variances s_1^2 and s_2^2, respectively.
- The distribution of $\frac{s_1^2/\sigma_1^2}{s_2^2/\sigma_2^2}$ is F-distribution with numerator degrees of freedom $\nu_1 = n_1 - 1$ and denominator degrees of freedom $\nu_2 = n_2 - 1$.
- Generally, if U and V are independent chi-squared random variables with degrees of freedom ν_1 and ν_2, respectively, then $F = \frac{U/\nu_1}{V/\nu_2}$ has an F-distribution with numerator and denominator degrees of freedom ν_1 and ν_2.

The F Critical Value (Table A.6)

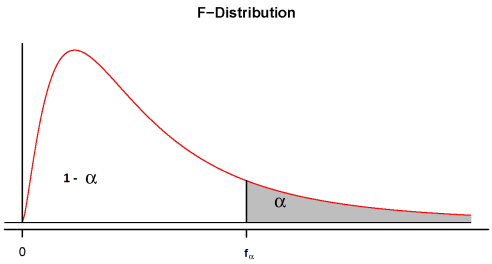
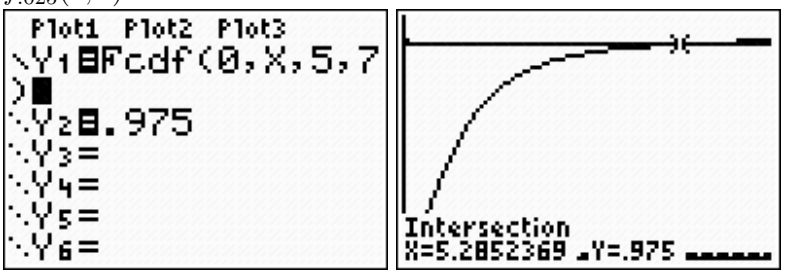


Table A.6 gives $f_\alpha(\nu_1, \nu_2)$ for $\alpha = .05, .01$.
For example, $\nu_1 = 5$, $\nu_2 = 7$, $f_{.05}(5, 7) = 3.97$ by Table A.6.

$$f_{1-\alpha}(\nu_1, \nu_2) = \frac{1}{f_\alpha(\nu_2, \nu_1)}$$

The F Critical Value using Technology

- In Excel: $f_{.025}(5, 7)$=F.INV.RT(0.025,5,7)=5.285236852, or =F.INV(1-0.025,5,7)=5.285236852.
- In TI-8x: Graph Y1=Fcdf(0, X, 5,7) and Y2=1-.025=0.975, the x-coordinate of the intersection is $f_{.025}(5, 7)$=5.2852369.



- In R: $f_{.025}$= "qf(.975,5,7)=5.285237".

8.8 Quantile-Quantile Plot: Q-Q Plot

Definition:
Q-Q Plot is a plot of $y_{(i)}$ against $q_f(p_i)$, where $p_i = \frac{i-3/8}{n+1/4}$, $i = 1, 2, 3, \ldots, n$.
Usage If the points of the q-q plot are close to a straight line, then the data is likely from the distribution f.

Exponential Q-Q Plot

Definition:
Exponential Q-Q Plot is a plot of $y_{(i)}$ against $q_f(p_i)$, where $p_i = \frac{i-3/8}{n+1/4}$, $i = 1, 2, 3, \ldots, n$, and f is the exponential distribution with mean =1, i.e., $q_f(p) = -\ln(1-p)$, $0 < p < 1$.
Usage If the points of the q-q plot are close to a straight line with y-intercept 0, then the data is likely from an exponential distribution with mean being estimated by the slope.

Example 4. Construct an exponential q-q lot for the data:
1.094, 2.630, 0.882, 1.885, 0.721, 1.290, 0.019.

Normal Q-Q Plot

Definition:
Normal q-q plot is a plot of $y_{(i)}$ against $q_f(p_i)$, where f is the standard normal distribution $N(0, 1)$ and $p_i = \frac{i-3/8}{n+1/4}$, $i = 1, 2, 3, \ldots, n$. $q_f(p) = \Phi^{-1}(p)$=invNorm(p).

Usage If the points of the q-q plot are close to a straight line, then the data is likely from a normal distribution $N(\mu, \sigma^2)$. The slope is an estimate of σ and y-intercept is an estimate of μ.

Example 5. Construct a normal q-q lot for the data:

6.72	6.77	6.82	6.70	6.78	6.70	6.62	6.75	6.66
6.66	6.64	6.76	6.73	6.80	6.72	6.76	6.76	6.68
6.66	6.62	6.72	6.76	6.70	6.78	6.76	6.67	6.70
6.72	6.74	6.81	6.79	6.78	6.66	6.76	6.76	6.72

Relative Frequency and Histogram To describe continuous-type data, we group the data values into classes (intervals) and count the (relative) frequency of the data values in each class.