4. Introduction to Statistics

Descriptive Statistics

Types of data

A *variate* or *random variable* is a quantity or attribute whose value may vary from one *unit* of investigation to another. For example, the units might be headache sufferers and the variate might be the time between taking an aspirin and the headache ceasing.

An *observation* or *response* is the value taken by a variate for some given unit.

There are various types of variate.

- *Qualitative* or *nominal;* described by a word or phrase (e.g. blood group, colour).

- *Quantitative*; described by a number (e.g. time till cure, number of calls arriving at a telephone exchange in 5 seconds).

- *Ordinal;* this is an "in-between" case. Observations are not numbers but they can be ordered (e.g. much improved, improved, same, worse, much worse).

Averages etc. can sensibly be evaluated for quantitative data, but not for the other two. Qualitative data can be analysed by considering the frequencies of different categories. Ordinal data can be analysed like qualitative data, but really requires special techniques called *nonparametric methods*.

Quantitative data can be:

- *Discrete:* the variate can only take one of a finite or countable number of values (e.g. a count)

- *Continuous:* the variate is a measurement which can take any value in an interval of the real line (e.g. a weight).

Displaying data

It is nearly always useful to use graphical methods to illustrate your data. We shall describe in this section just a few of the methods available.

Discrete data: frequency table and bar chart

Suppose that you have collected some discrete data. It will be difficult to get a "feel" for the distribution of the data just by looking at it in list form. It may be worthwhile constructing a frequency table or bar chart.

The *frequency* of a value is the number of observations taking that value.

A *frequency table* is a list of possible values and their frequencies.

A *bar chart* consists of bars corresponding to each of the possible values, whose heights are equal to the frequencies.

Example

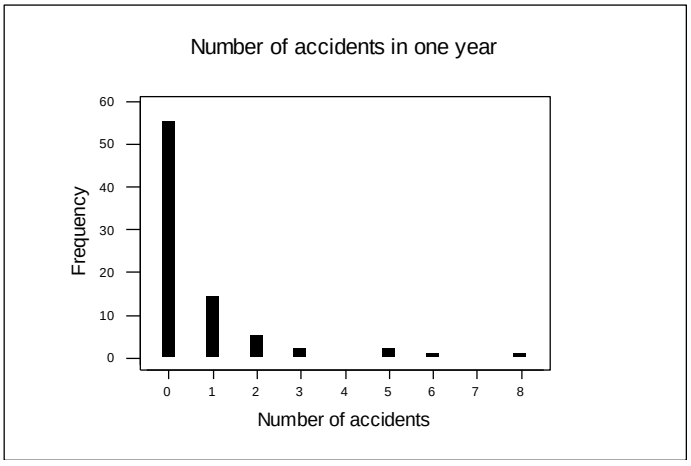
The numbers of accidents experienced by 80 machinists in a certain industry over a period of one year were found to be as shown below. Construct a frequency table and draw a bar chart.

```
2   0   0   1   0   3   0   6   0   0   8   0   2   0   1
5   1   0   1   1   2   1   0   0   0   2   0   0   0   0
0   0   0   0   0   0   0   0   0   0   0   0   1   0   1
0   0   0   5   1   0   0   0   0   0   0   0   0   1   1
0   3   0   0   1   1   0   0   0   2   0   1   0   0   0
0   0   0   0   0
```

Solution

Number of accidents	Tallies	Frequency
0	|||| |||| |||| |||| |||| |||| |||| |||| |||| |||| ||||	55
1	|||| |||| ||||	14
2	||||	5
3	||	2
4		0
5	||	2
6	|	1
7		0
8	|	1

Barchart



Continuous data: histograms

When the variate is continuous, we do not look at the frequency of each value, but group the values into intervals. The plot of frequency against interval is called a histogram. Be careful to define the interval boundaries unambiguously.

Example

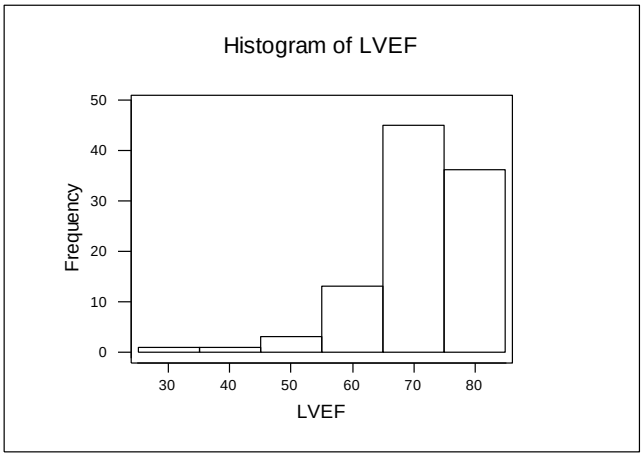
The following data are the left ventricular ejection fractions (LVEF) for a group of 99 heart transplant patients. Construct a frequency table and histogram.

62	64	63	70	63	69	65	74	67	77	65	72	65
77	71	79	75	78	64	78	72	32	78	78	80	69
69	65	76	53	74	78	59	79	77	76	72	76	70
76	76	74	67	65	79	63	71	70	84	65	78	66
72	55	74	79	75	64	73	71	80	66	50	48	57
70	68	71	81	74	74	79	79	73	77	80	69	78
73	78	78	66	70	36	79	75	73	72	57	69	82
70	62	64	69	74	78	70	76					

Frequency table

LVEF	Tallies	Frequency
24.5 - 34.5	|	1
34.5 - 44.5	|	1
44.5 - 54.5	|||	3
54.5 - 64.5	|||| |||| |||	13
64.5 - 74.5	|||| |||| |||| |||| |||| |||| |||| |||| ||||	45
74.5 - 84.5	|||| |||| |||| |||| |||| |||| |||| |	36

Histogram



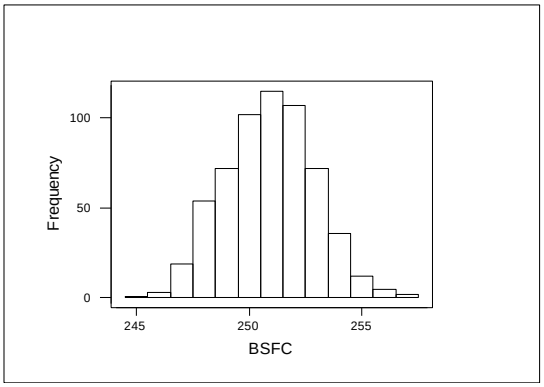
Note: if the interval lengths are unequal, the heights of the rectangles are chosen so that the area of each rectangle equals the frequency i.e. height of rectangle = frequency ÷ interval length.

Things to look out for

Bar charts and histograms provide an easily understood illustration of the distribution of the data. As well as showing where most observations lie and how variable the data are, they also indicate certain "danger signals" about the data.

Normally distributed data

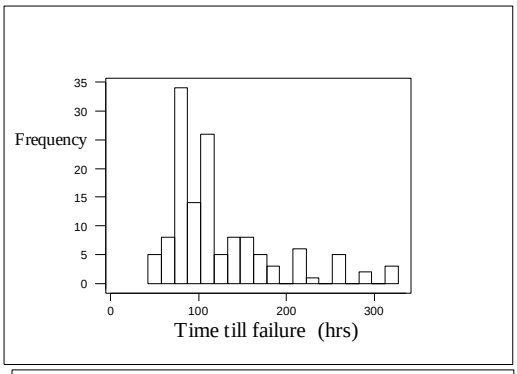
The histogram is bell-shaped, like the probability density function of a Normal distribution. It appears, therefore, that the data can be modelled by a Normal distribution. (Other methods for checking this assumption are available.)



Similarly, the histogram can be used to see whether data look as if they are from an Exponential or Uniform distribution.

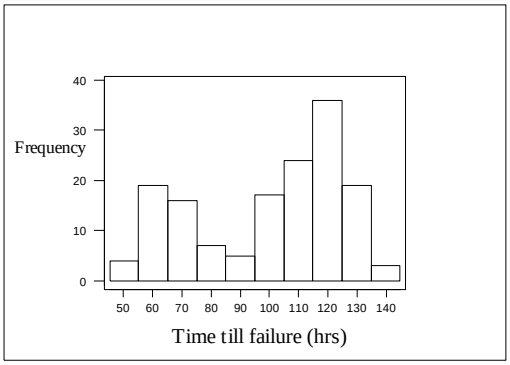
Very skew data

The relatively few large observations can have an undue influence when comparing two or more sets of data. It might be worthwhile using a transformation e.g. taking logarithms.



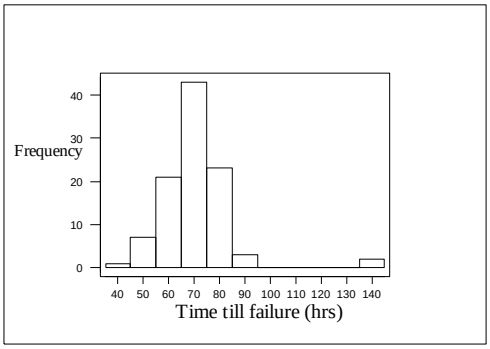
Bimodality

This may indicate the presence of two sub-populations with different characteristics. If the subpopulations can be identified it might be better to analyse them separately.



Outliers

The data appear to follow a pattern with the exception of one or two values. You need to decide whether the strange values are simply mistakes, are to be expected or whether they are correct but unexpected. The outliers may have the most interesting story to tell.



Summary Statistics

Measures of location

By a measure of location we mean a value which typifies the numerical level of a set of observations. (It is sometimes called a "central value", though this can be a misleading name.) We shall look at three measures of location and then discuss their relative merits.

Sample mean

The *sample mean* of the values $x_1, x_2, \ldots, x_n$ is

$$\bar{x} = \frac{x_1 + x_2 + .. x_n}{n} = \frac{1}{n}\sum_{i=1}^{n} x_i$$

This is just the *average* or *arithmetic mean* of the values. Sometimes the prefix "sample" is dropped, but then there is a possibility of confusion with the *population mean* which is defined later.

Frequency data: suppose that the frequency of the class with midpoint x_i is f_i, for i = 1, 2, ..., m). Then

$$\bar{x} = \frac{f_1 x_1 + f_2 x_2 + \cdots + f_m x_m}{n} = \frac{1}{n}\sum_{i=1}^{m} f_i x_i$$

Where $n = \sum_{i=1}^{m} f_i$ = total number of observations.

Example

Accidents data: find the sample mean.

Number of accidents, x_i	Frequency f_i	$f_i x_i$
0	55	0
1	14	14
2	5	10
3	2	6
4	0	0
5	2	10
6	1	6
7	0	0
8	1	8
TOTAL	80	54

$$\Rightarrow \bar{x} = \frac{54}{80} = 0.675$$

Sample median

The median is the central value in the sense that there as many values smaller than it as there are larger than it.

All values known: if there are *n* observations then the median is:

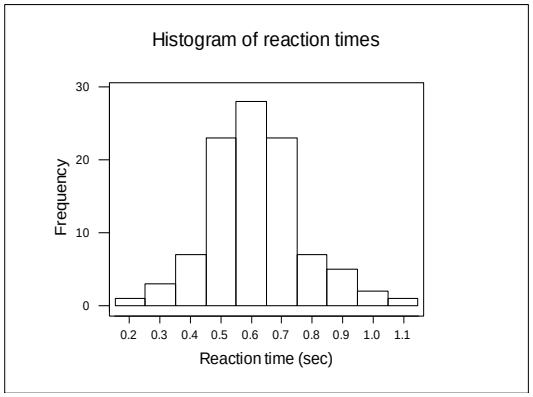
- the $\frac{n+1}{2}$ largest value, if *n* is odd;
- the sample mean of the $\frac{n}{2}$ largest and the $\frac{n}{2}+1$ largest values, if *n* is even.

Mode

The mode, or modal value, is the most frequently occurring value. For continuous data, the simplest definition of the mode is the midpoint of the interval with the highest rectangle in the histogram. (There is a more complicated definition involving the frequencies of neighbouring intervals.) It is only useful if there are a large number of observations.

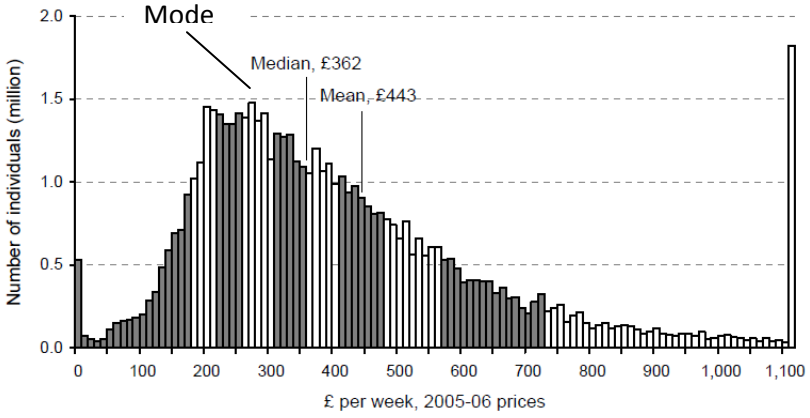
Comparing mean, median and mode

Symmetric data: the mean median and mode will be approximately equal.



Skew data: the median is less sensitive than the mean to extreme observations. The mode ignores them.

Figure 1. The income distribution in 2005–06 (UK) IFS Briefing Note No 73



Notes: Incomes have been measured before housing costs have been deducted. The right-most bar represents incomes of over £1,100.

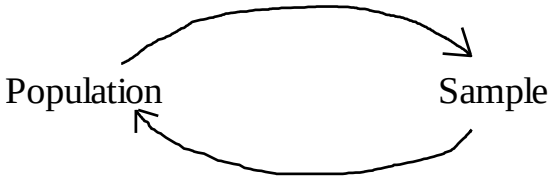
The mode is dependent on the choice of class intervals and is therefore not favoured for sophisticated work.

Sample mean and median: it is sometimes said that the mean is better for symmetric, well behaved data while the median is better for skewed data, or data containing outliers. The choice really mainly depends on the use to which you intend putting the "central" value. If the data are very skew, bimodal or contain many outliers, it may be questionable whether any single figure can be used, much better to plot the full distribution. For more advanced work, the median is more difficult to work with. If the data are skewed, it may be better to make a transformation (e.g. take logarithms) so that the transformed data are approximately symmetric and then use the sample mean.

Statistical Inference

Probability theory: the probability distribution of the population is known; we want to derive results about the probability of one or more values ("random sample") - *deduction*.

Statistics: the results of the random sample are known; we want to determine something about the probability distribution of the population - *inference*.



In order to carry out valid inference, the sample must be representative, and preferably a random sample.

Random sample: two elements: (i) no bias in the selection of the sample;

(ii) different members of the sample chosen independently.

Formal definition of a random sample: $X_1, X_2, \ldots X_n$ are a random sample if each X_i has the same distribution and the X_i's are all independent.

Parameter estimation

We assume that we know the type of distribution, but we do not know the value of the parameters θ, say. We want to estimate θ,on the basis of a random sample $X_1, X_2, \ldots, X_n$. Let's call the random sample $X_1, X_2, \ldots, X_n$ our data D. We wish to infer $P(\theta|D)$ which by Bayes' theorem is

$$\mathrm{P}(\theta|\mathrm{D}) = \frac{\mathrm{P}(\mathrm{D}|\theta)\mathrm{P}(\theta)}{\mathrm{P}(\mathrm{D})}$$

P(θ) is called the *prior*, which is the probability distribution from any prior information we had before looking at the data (often this is taken to be a constant). The denominator P(D) does not depend on the parameters, and so is just a normalization constant. $P(D|\theta)$ is called the *likelihood:* it is how likely the data is given a particular set of parameters.

The full distribution $P(\theta|D)$ gives all the information about the probability of different parameters values given the data. However it is often useful to summarise this information, for example giving a peak value and some error bars.

Maximum likelihood estimator: the value of θ that maximizes the likelihood $P(D|\theta)$ is called the *maximum likelihood* estimate: it is the value that makes the data most likely, and if P(θ) does not depend on parameters (e.g. is a constant) is also the most probable value of the parameter given the observed data.

The maximum likelihood estimator is usually the best estimator, though in some instances it may be numerically difficult to calculate. Other simpler estimators are sometimes possible. Estimates are typically denoted by: $\hat{\theta}, \theta^*$, etc. Note that since P(D|θ) is positive, maximizing P(D|θ) gives the same as maximizing log P(D|θ).

Example Random samples $X_1, X_2, \ldots X_n$ are drawn from a Normal distribution. What is the maximum likelihood estimate of the mean μ?

Solution

$$P(D|\mu,\sigma^2) = P(X_1|\mu,\sigma^2)P(X_2|\mu,\sigma^2)..P(X_n|\mu,\sigma^2) \propto \boldsymbol{e}^{-\frac{(\boldsymbol{X_1}-\boldsymbol{\mu})^2 + (\boldsymbol{X_2}-\boldsymbol{\mu})^2+\cdots(\boldsymbol{X_n}-\boldsymbol{\mu})^2}{\boldsymbol{2\sigma^2}}}$$

We find the maximum likelihood by maximizing the log likelihood, here log $P(D|\mu,\sigma^2)$. So for a maximum likelihood estimate of μ we want

$$\frac{\partial \log P(D|\mu,\sigma^2)}{\partial \mu} = -\frac{\partial}{\partial \mu}\sum_i \frac{(X_i-\mu)^2}{2\sigma^2} = \sum_i \frac{(X_i-\mu)}{\sigma^2} = 0$$

The solution is the maximum likelihood estimator $\hat{\mu}$ with

$$\frac{\sum_i X_i}{\sigma^2} = \frac{\sum_i \hat{\mu}}{\sigma^2} = \frac{n\hat{\mu}}{\sigma^2}$$

$$=> \hat{\mu} = \bar{X} = \frac{1}{n}\sum_{i=1}^{n} X_i$$

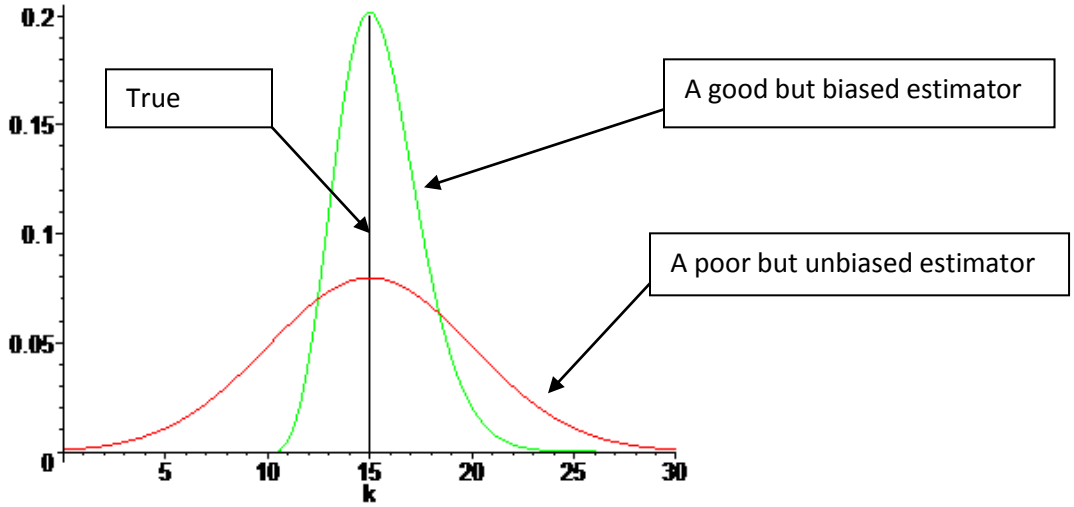
So the maximum likelihood estimator of the mean is just the sample mean we discussed before. We can similarly maximize with respect to σ^2 when the mean is the maximum likelihood value $\mu = \hat{\mu}$. This gives

$$\widehat{\sigma^2} = \frac{1}{n}\sum_{i=1}^{n} X_i^2 - \bar{X}^2.$$

Comparing estimators

A good estimator should have as narrow a distribution as possible (i.e. be close to the correct value as possible). Often it is also useful to have it being unbiased, that on average (over possible data samples) it gives the true value:

The estimator $\hat{\theta}$ is *unbiased* for θ if $E(\hat{\theta}) = \theta$ for all values of θ.



Result: $\hat{\mu} = \bar{X}$ is an *unbiased* estimator of μ.

$$\langle \bar{X} \rangle = \langle \frac{1}{n}(X_1 + \cdots + X_n) \rangle = \frac{1}{n}(\langle X_1 \rangle + \langle X_2 \rangle + \ \ldots + \langle X_n \rangle) = \frac{1}{n}(\mu + \mu + \cdots + \mu)$$
$$= \frac{1}{n}(n\mu) = \mu$$

Result: $\widehat{\sigma^2}$ is a *biased* estimator of σ^2.

$$\langle \widehat{\sigma^2} \rangle = \langle \frac{1}{n}\sum_{i=1}^{n} X_i^2 \ - \ \hat{\mu}^2 \rangle = \frac{1}{n}\sum_{i=1}^{n} \langle X_i^2 \rangle \ - \langle \hat{\mu}^2 \rangle$$
$$= \frac{1}{n}\sum_{i=1}^{n} \langle X_i^2 \rangle \ - \langle \left(\frac{1}{n}\sum_{i=1}^{n} X_i \right)^2 \rangle = \frac{1}{n}\sum_{i=1}^{n} (\mu^2 + \sigma^2) - \langle \frac{1}{n^2}\sum_{i=1}^{n} X_i \ \sum_{j=1}^{n} X_j \ \rangle$$
$$= \mu^2 + \sigma^2 - \frac{1}{n^2}\sum_{i=1}^{n} \langle X_i^2 \rangle - \frac{1}{n^2}\langle \sum_{i=1}^{n}\sum_{j \neq i} X_i X_j \ \rangle$$
$$= \ \mu^2 + \sigma^2 - \frac{1}{n}(\mu^2 + \ \sigma^2) - \frac{n(n-1)}{n^2}\mu^2 = \sigma^2 - \frac{\sigma^2}{n} = \frac{n-1}{n}\sigma^2$$

where we used $\langle X_i X_j \rangle = \langle X_i \rangle \langle X_j \rangle = \mu^2$ for independent variables ($i \neq j$).

Sample variance

Since $\widehat{\sigma^2}$ is a *biased* estimator of σ^2 it is common to use the *unbiased* estimator of the variance, often called the *sample variance:*

$$s^2 = \frac{n}{n-1}\widehat{\sigma^2} = \frac{1}{n-1}\sum(X_i - \bar{X})^2 = \frac{1}{n-1}\sum(X_i^2 - \bar{X}^2) = \frac{\sum_i X_i^2 - n\bar{X}^2}{n-1}$$

The last form is often more convenient to calculate, but also less numerically stable (you are taking the difference of two potentially large numbers).

Why the $n-1$?

We showed that $\langle\widehat{\sigma^2}\rangle = \frac{n-1}{n}\sigma^2$, and hence that $s^2 = \frac{n}{n-1}\langle\widehat{\sigma^2}\rangle$ is an unbiased estimate of the variance.

Intuition: the reason the estimator is biased is because the mean is also estimated from the same data. It is not biased if you know the true mean and can use μ instead of $\bar{X}$: One unit of information has to be used to estimate the mean, leaving *n-1* units to estimate the variance. This is very obvious with only one data point X_1: if you know the true mean this still tells you something about the variance, but if you have to estimate the mean as well – best guess X_1 – you have nothing left to learn about the variance. This is why the unbiased estimator s^2 is undefined for n=1.

Intuition 2: the sample mean is closer to the centre of the distribution of the samples than the true (population) mean is, so estimating the variance using the r.m.s. distance from the *sample* mean underestimates the variance (which is the scatter about the *population* mean).

For a normal distribution the estimator $\widehat{\sigma^2}$ is the maximum likelihood value when $\mu = \hat{\mu}$ - i.e. the mean fixed to its maximum value. If we averaged over possible values of the true mean (a process called *marginalization*), and then maximized this averaged distribution, we would have found s^2 is the maximum likelihood estimator. i.e. s^2 accounts for uncertainty in the true mean. For large n the mean is measured accurately, and $\frac{n}{n-1} \approx 1$.

Measures of dispersion

A measure of dispersion is a value which indicates the degree of variability of data. Knowledge of the variability may be of interest in itself but more often is required in order to decide how precisely the sample mean – and *estimator* of the mean - reflects the population (true) mean.

A measure of dispersion in the original units as the data is the *standard deviation*, which is just the (positive) square root of the sample variance: $s = \sqrt{\text{sample variance}}$.

For frequency data, where f_i is the frequency of the class with midpoint x_i (i = 1, 2, ..., m):

$$s^2 = \frac{f_1(x_1 - \bar{x})^2 + f_2(x_2 - \bar{x})^2 + \cdots + f_m(x_m - \bar{x})^2}{n-1} = \frac{1}{n-1}\sum_{i=1}^{m} f_i(x_i - \bar{x})^2$$

$$= \frac{1}{n-1}\left(\sum_{i=1}^{m} f_i x_i^2 - n\bar{x}^2\right) \quad (\text{where } n = \sum_{i=1}^{m} f_i)$$

Example Find the sample mean and standard deviation of the following: 6, 4, 9, 5, 2.

Example Evaluate the sample mean and standard deviation, using the frequency table.

LVEF	Midpoint, x_i	Frequency, f_i	$f_i x_i$	$f_i x_i^2$
24.5 - 34.5	29.5	1	29.5	870.25
34.5 - 44.5	39.5	1	39.5	1560.25
44.5 - 54.5	49.5	3	148.5	7350.75
54.5 - 64.5	59.5	13	773.5	46023.25
64.5 - 74.5	69.5	45	3127.5	217361.25
74.5 - 84.5	79.5	36	2862.0	227529.00
	TOTAL	**99**	**6980.5**	**500695.00**

Sample mean, $\bar{x} = \frac{69805}{99} = 70.510$.

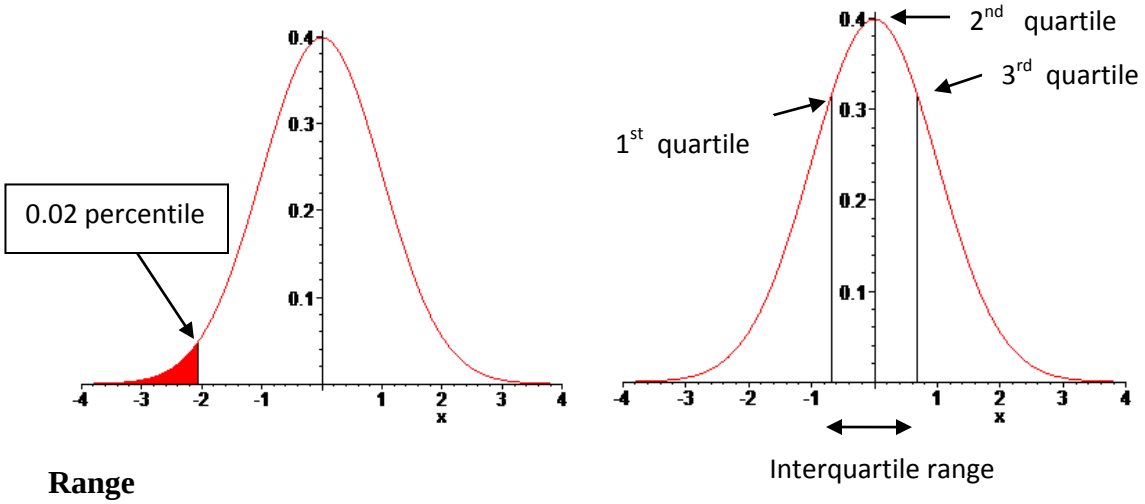
Sample variance, $s^2 = \frac{1}{98}(500695.00 - 99 \times 70.510^2) = 86.727$

Sample standard deviation, $s = \sqrt{86.727} = 9.313$.

Note: when using a calculator, work to full accuracy during calculations in order to minimise rounding errors. If your calculator has statistical functions, s is denoted by σ_{n-1}.

Percentiles and the interquartile range

The kth percentile is the value corresponding to cumulative relative frequency of k/100 on the cumulative relative frequency diagram e.g. the 2nd percentile is the value corresponding to cumulative relative frequency 0.02. The 25th percentile is also known as the first quartile and the 75th percentile is also known as the third quartile. The *interquartile range* of a set of data is the difference between the third quartile and the first quartile, or the interval between these values. It is the range within which the "middle half" of the data lie, and so is a measure of spread which is not too sensitive to one or two outliers.



Range

The *range* of a set of data is the difference between the maximum and minimum values, or the interval between these values. It is another measure of the spread of the data.

Comparing sample standard deviation, interquartile range and range

The range is simple to evaluate and understand, but is sensitive to the odd extreme value and does not make effective use of all the information of the data. The sample standard deviation is also rather sensitive to extreme values but is easier to work with mathematically than the interquartile range.

Confidence Intervals

Estimates are "best guesses" in some sense, and the sample variance gives some idea of the spread. Confidence intervals are another measure of spread, a range within which we are "pretty sure" that the parameter lies.

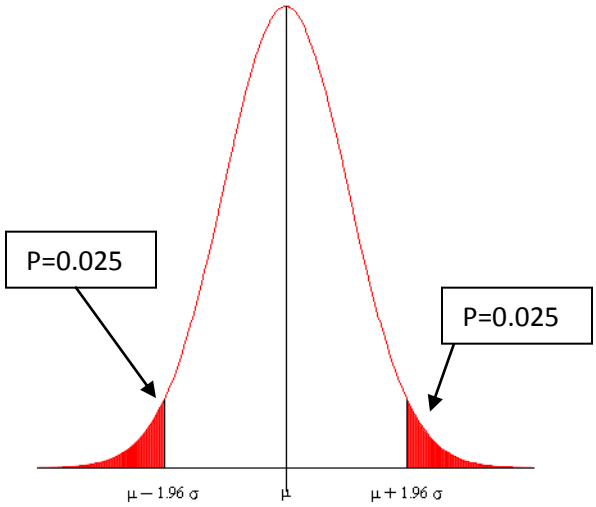
Normal data, variance known

Random sample $X_1, X_2, \ldots X_n$ from $N(\mu, \sigma^2)$, where σ^2 is known but μ is unknown. We want a confidence interval for μ.

Recall:

(i) $\bar{X} \sim N(\mu, {\sigma_{\bar{X}}}^2)$

(ii) With probability 0.95, a Normal random variables lies within 1.96 standard deviations of the mean.



$$P(\mu - 1.96\sigma_{\bar{X}} \leq \bar{X} \leq \mu + \ 1.96\sigma_{\bar{X}} \,|\mu) = 0.95$$

Since the variance of the sample mean is $\sigma_{\bar{X}}^2 = \sigma^2/n$ this gives

$$P\left(\mu - 1.96\sqrt{\frac{\sigma^2}{n}} \leq \bar{X} \leq \mu + \ 1.96\sqrt{\frac{\sigma^2}{n}}\,\middle|\,\mu\right) = 0.95$$

To infer the distribution of μ given $\bar{X}$ we need to use Bayes' theorem

$$P(\mu|\bar{X}) = \frac{P(\bar{X}\,|\mu)P(\mu)}{P(\bar{X})}$$

If the prior on μ is constant, then $P(\mu|\bar{X})$ is also Normal with mean $\bar{X}$ so

$$P(\bar{X} - 1.96\,\sigma_{\bar{X}} \leq \mu \leq \bar{X} + 1.96\sigma_{\bar{X}}) = 0.95$$

Or

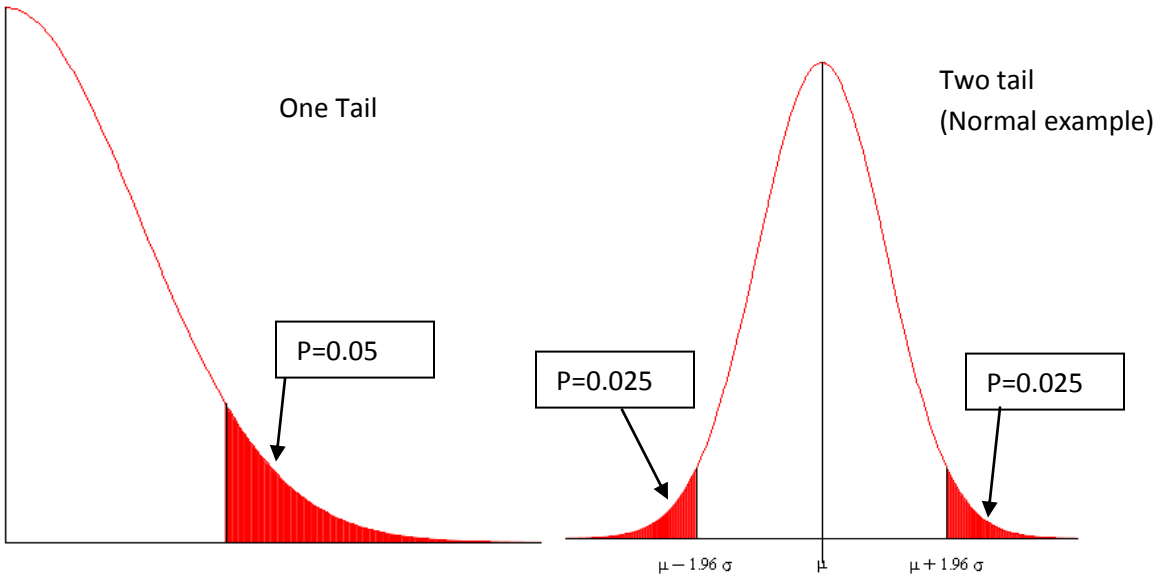
$$P\left(\bar{X} - 1.96\sqrt{\frac{\sigma^2}{n}} \leq \mu \leq \bar{X} + 1.96\sqrt{\frac{\sigma^2}{n}}\,\middle|\,\bar{X}\right) = 095$$

A 95% confidence interval for μ is: $\bar{X} - 1.96\sqrt{\frac{\sigma^2}{n}}$ to $\bar{X} + 1.96\sqrt{\frac{\sigma^2}{n}}$.

Two tail versus one tail

When the distribution has two ends (tails) where the likelihood goes to zero, the most natural choice of confidence interval is the regions excluding both tails, so a 95% confidence region means that 2.5% of the probability is in the high tail, 2.5% in the low tail. If the distribution is one sided, a *one tail* interval is more appropriate.

Example: 95% confidence regions



Example: Polling (Binomial data)
[unnecessarily complicated example… but a useful general result for poll error bars]

A sample of 1000 random voters were polled, with 350 saying they will vote for the Conservatives and 650 saying another party. What is the 95% confidence interval for the Conservative share of the vote?

Solution:
n Bernoulli trials, *X* = number of people saying they will vote Conservative; $X \sim B(n, p)$.

If *n* is large, *X* is approx. $N\big(np, np(1-p)\big)$. The mean is $X = np$ so we can estimate $\bar{p} = X/n$. The variance of *X* is $np(1-p) = X(1-\frac{X}{n})$ and hence the variance of $\bar{p}$ can be taken to be $\sigma_{\bar{p}}^2 = \frac{X\left(1-\frac{X}{n}\right)}{n^2}$ or a standard deviation of $\sigma_{\bar{p}} = \frac{\sqrt{X\left(1-\frac{X}{n}\right)}}{n} = \sqrt{\frac{\bar{p}(1-\bar{p})}{n}}$. Hence the 95% confidence (two-tail) interval is $\bar{p} - 1.96\sigma_{\bar{p}} < p < \bar{p} + 1.96\sigma_{\bar{p}}$ or

$$\bar{p} - 1.96\sqrt{\frac{\bar{p}(1-\bar{p})}{n}} < p < \bar{p} + 1.96\sqrt{\frac{\bar{p}(1-\bar{p})}{n}}$$

With $X = 350, n = 1000$ this corresponds to standard deviation of 0.015 and a 95% confidence plus/minus error of 3%:

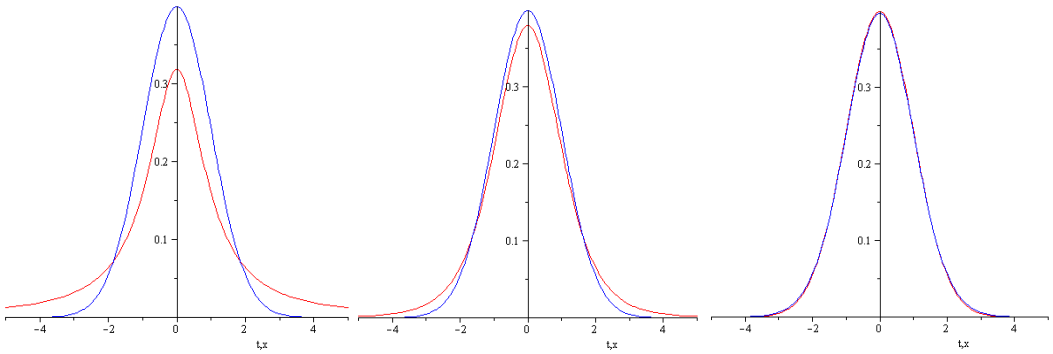
0.35–0.03 < p < 0.35+0.03 so 0.32 < p < 0.38.

Normal data, variance unknown: Student's *t*-distribution

Random sample $X_1, X_2, \ldots, X_n$ from $N(\mu, \sigma^2)$, where σ^2 and μ are unknown. We want a confidence interval for μ.

The distribution of $\frac{\bar{X}-\mu}{s/\sqrt{n}}$ is called a *t-distribution* with $n-1$ degrees of freedom.
So the situation is like when we know the variance, when $\frac{\bar{X}-\mu}{\sigma/\sqrt{n}}$ is normally distributed, but now replacing σ^2 by the sample estimate s^2. We have to use the t-distribution instead.

$\nu = n-1 = 1$ $\nu = n-1 = 5$ $\nu = n-1 = 50$



The fact that you have to estimate the variance from the data --- making true variances larger than the estimated sample variance possible --- broadens the tails significantly when there are not a large number of data points. As *n* becomes large, the *t*-distribution converges to a normal.

Derivation of the t-distribution is a bit tricky, so we'll just look at how to use it.

If σ^2 is known, confidence interval for μ is $\bar{X} - z\sqrt{\frac{\sigma^2}{n}}$ to $\bar{X} + z\sqrt{\frac{\sigma^2}{n}}$, where z is obtained from Normal tables.

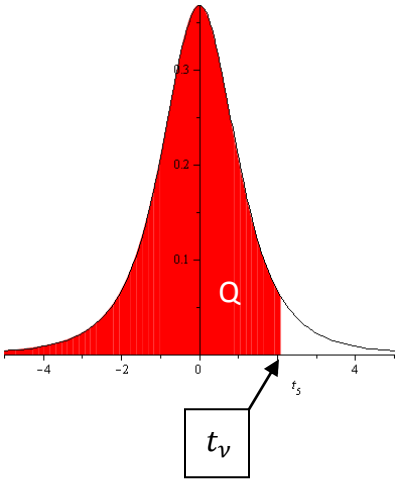
If σ^2 is unknown, we need to make two changes:

(i) Estimate σ^2 by s^2, the sample variance;

(ii) replace z by t_{n-1}, the value obtained from t-tables,

The confidence interval for μ is: $\bar{X} - t_{n-1}\sqrt{\frac{s^2}{n}}$ to $\bar{X} + t_{n-1}\sqrt{\frac{s^2}{n}}$.

t-tables: these give t_ν for different values Q of the cumulative Student's *t*-distributions, and for different values of ν. The parameter ν is called the number of degrees of freedom. When the mean and variance are unknown, there are *n-1* degrees of freedom to estimate the variance, and this is the relevant quantity here.



$$Q(t_\nu) = \int_{-\infty}^{t_\nu} f_\nu(t)dt$$

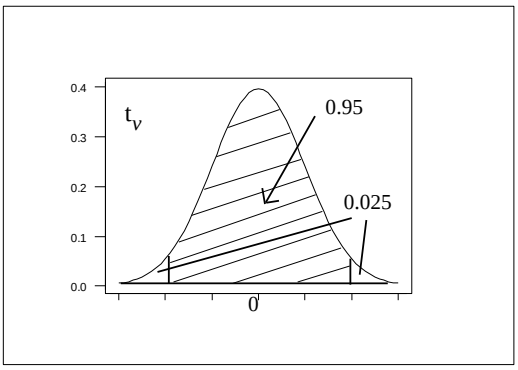
The t-tables are laid out differently from N(0,1).

Etymology (Wikipedia: Beer is good for statistics!)

The "Student's" distribution was actually published in 1908 by William Sealy Gosset. Gosset, however, was employed at a brewery that forbade members of its staff publishing scientific papers due to an earlier paper containing trade secrets. To circumvent this restriction, Gosset used the name "Student", and consequently the distribution was named "Student's t-distribution".[2]

For a 95% confidence interval, we want the middle 95% region, so Q = 0.975 (i.e. 0.05/2=0.025 in both tails).

Similarly, for a 99% confidence interval, we would want Q = 0.995.



Example: From n = 20 pieces of data drawn from a Normal distribution have sample mean $\bar{X} = 10$, and sample variance $s^2 = 2$. What is the 95% confidence interval for the population mean μ?

From t-tables, t_{19}, Q = 0.975, t = 2.093.

95% confidence interval for μ is: $10 - 2.093\sqrt{\frac{2}{20}} < \mu < 10 + 2.093\sqrt{\frac{2}{20}}$

i.e. 9.34 to 10.66

Sample size

When planning an experiment or series of tests, you need to decide how many repeats to carry out to obtain a certain level of precision in you estimate. The confidence interval formula can be helpful.

For example, for Normal data, confidence interval for μ is $\bar{X} \pm t_{n-1}\sqrt{\frac{s^2}{n}}$.

Suppose we want to estimate μ to within $\pm\delta$, where δ (and the degree of confidence) is given. We must choose the sample size, *n*, satisfying:

$$\delta = t_{n-1}\sqrt{\frac{s^2}{n}} \Rightarrow n = \frac{t_{n-1}^2 s^2}{\delta^2}$$

To use this need:

(i) an estimate of s^2 (e.g. results from previous experiments);

(ii) an estimate of t_{n-1}. This depends on *n*, but not very strongly. You will not go far wrong, in general, if you take $t_{n-1} = 2.1$ for 95% confidence.

Rule of thumb: for 95% confidence, choose $n = \frac{2.1^2 \times \text{Estimate of variance}}{\delta^2}$