1 Random sampling

1.1 Data display

1.1.1 Dotplots

1.1.2 Box plots

A box plot is a graphic summary of a set of data in terms of he median and the quartiles. Suppose we measure the pulse rate of 12 students following a mid term exam and we observe

$62, 64, 68, 70, 70, 74, 74, 76, 76, 78, 78, 80$

The Median is the mid value if n is odd or the average of the mid values if n is even. For the data above it is equal to $\frac{74+74}{2} = 74$

The first quartile denoted Q1 is the median of the lower 6 values, i.e. $Q1 = \frac{68+70}{2} = 69$

The third quartile denoted Q3 is the median of the upper 6 values, i.e. $Q3 = \frac{76+78}{2} = 77$

The inter quartile range denoted IQR is the difference $Q3 - Q1 = 77 - 69 = 8$. It is measure of the spread of the data.

1.1.3 Graphic displays

Example Car battery life data

2.2	4.1	3.5	4.5	3.2	3.7	3.0	2.6
3.4	1.6	3.1	3.3	3.8	3.1	4.7	3.7
2.5	4.3	3.4	3.6	2.9	3.3	3.9	3.1
3.3	3.1	3.7	4.4	3.2	4.1	1.9	3.4
4.7	3.8	3.2	2.6	3.9	3.0	4.2	3.5

Relative	Frequency	distribution	
Class Interval	Class midpoint	Frequency	Relative frequency
1.5-1.9	1.7	2	0.050
2.0-2.4	2.2	1	0.025
2.5-2.9	2.7	4	0.100
3.0-3.4	3.2	15	0.375
3.5-3.9	3.7	10	0.250
4.0-4.4	4.2	5	0.125
4.5-4.9	4.7	3	0.075

$Q1 = 3.1, Q2 = 3.4, Q3 = 3.875, IQR = Q3 - Q1 = 0.775$

We will now indicate the link between a sample and the population. We define a population to be the set of all observations with which we are concerned. A sample is a subset of a population chosen in some way. We shall formally represent a sample by means of random variables, $X_1, ..., X_n$. By a random sample of size n we shall mean that $X_1, ..., X_n$ are independent and identically distributed, i.i.d. for short. It is informative to consider the following example whereby three different individuals toss a coin 3 times each.

	Tom		Bill		Shelly
First toss X_1	H		H		H
Second toss X_2	T		H		H
Third toss X_3	T		T		H
Number of Heads $W = \sum_{i=1}^{3} X_i$	1		2		3

So Tom took a sample and observed HTT . Bill took a sample of the same size and observed HHT. Similarly Shelly observed HHH.

We denote specific values of a random variable W by w. So here W has taken observed values $w_1 = 1, w_2 = 2, w_3 = 3$.

The joint probability distribution of $X_1, ..., X_n$ is

$$f(x_1, ..., x_n) = f(x_1) ... f(x_n)$$

Definition A function of the random variables $X_1, ..., X_n$ is called a statistic.

For example the sample average is

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

The sample variance is

$$S^2 = \frac{1}{n\left(n-1\right)}\left[n\sum X_i^2 - \left(\sum_{i=1}^{n} X_i\right)^2\right]$$

Example Suppose we observe values $12, 15, 17, 20$. Then $\bar{x} = 16, \sum x^2 = 1058, s^2 =$

$\frac{34}{3}$

1.2 Asymptotic sampling distributions

In statistics, we are most often interested in the distribution of a function of the sample. For example, in coin tossing,we are interested in the distribution of the number of heads in n tosses. This is called the sampling distribution. In another example, we may be interested in the distribution of the average lifetime of a certain type of light bulb. A remarkable result in probability is the central limit theorem which states that for a large enough random sample, the distribution of the mean of the sample under certain conditions will be approximately Gaussian. This result does not require knowledge of the underlying distribution of the random variable under consideration.

1.2.1 Sampling distribution of means

Theorem Central Limit theorem

Let $X_1, ..., X_n$ be a random sample from a population with mean μ and variance σ^2. Then the asymptotic distribution of

$$Z = \sqrt{n}\left(\frac{\bar{X}_n - \mu}{\sigma}\right)$$

as $n \to \infty$ is the standard normal distribution.

The approximation is usually good whenever $n \geq 30$.

Example Suppose that we measure the diameter of a batch of 100 metal rods produced in a manufacturing process. We observe a batch mean value $\bar{x}_{100} = 5.027$.

The process has a specified mean length $\mu = 5$ cm and the standard deviation is $\sigma = 0.1$cm. The manufacturer wishes to ensure that the sample averages do not exceed the set mean of 5 cm. Does the batch meet the specifications?

To answer the question we compute the probability that the batch mean differs from the process mean by more than 0.027.

$$\begin{aligned} P\left(|\bar{X}_{100} - 5| > 0.027\right) &= P\left(|Z| > \frac{0.027\sqrt{100}}{0.1}\right) \\ &= 2P\left(Z > 2.7\right) \\ &= 2\left(0.0035\right) = 0.007 \end{aligned}$$

Since it is highly unlikely that the mean of the sample will differ from the process mean by more than 0.027, we conclude that the batch does not meet specifications. Note that we did not make any distributional assumptions.

Example A label on a standard bottle of beer in Canada states that it contains 341 ml by volume. A company claims it is complying with label and that the standard deviation is 5 ml. To verify the company claim, we take a random sample of 100 bottles and observe $\bar{x} = 339.5$ml. The probability of observing an average value less than or equal to 339.5 is

$$P\left(\bar{X}_{100} \leq 339.5\right) = P\left(Z \leq \frac{339.5 - 341}{5/\sqrt{100}}\right) = \Phi\left(-3\right) = 0.0013$$

Hence it is highly unlikely to observe an average as small as or smaller than the one observed. We conclude the company is not complying.

Example Rounding errors. Suppose that in making numerical computations either manually or by computer, only a certain finite number of significant digits can be retained at each step. If computations are carried to the closest $10^{-4,}$then all numbers involved will be restricted to 4 digits to the right of the decimal. That is, $3\frac{1}{3}$ is 3.3333; $\frac{2}{3}$ is 0.6667; π is 3.1416.

The difference between the number written in full as an infinite decimal and its rounded equivalent is rounding error. For $\frac{2}{3}$ it is $0.666...-0.6667 = -0.333...\times10^{-4}$

.

Suppose that in general the largest and smallest rounding errors are uniformly distributed in the interval -0.5×10^{-k} , 0.5×10^{-k}. If n numbers are computed and added together, the rounding error is a sum $\sum_{i=1}^{n} X_i$ which is approximately normally distributed with mean 0 and variance $\frac{\left(10^{-k}\right)^2}{12}n$.

If $n = 12, k = 4$ $P\left(-10^{-4} < \sum_{i=1}^{n} X_i < 10^{-4}\right) \simeq \Phi\left(\frac{10^{-4}-0}{10^{-4}}\right)$-$\Phi\left(\frac{-10^{-4}-0}{10^{-4}}\right) = \Phi(1) - \Phi(-1) = 0.6836$

Theorem Central Limit theorem for samples from different populations

Let $\bar{X}_{n_1}$ be the mean of a random sample of size n_1 from a population with mean μ_1 and variance σ_1^2. Let $\bar{X}_{n_2}$ be the mean of a random sample of size n_2 from another population with mean μ_2 and variance σ_2^2. Then the asymptotic distribution of

$$Z = \left(\frac{\left(\bar{X}_{n_1} - \bar{X}_{n_2}\right) - (\mu_1 - \mu_2)}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}} \right)$$

for large n_1, n_2 is the standard normal distribution.

Example The target thickness for both fruit flavored gum and for fruit flavored bubble gum is 6.7 hundredths of an inch. Hence, $\mu_1 - \mu_2 = 0$. We take random samples of $n_1 = 50, n_2 = 40$ respectively and we observe

$$\bar{x}_{50} = 6.701, \bar{y}_{40} = 6.841.$$

Suppose we know $\sigma_1^2 = 0.108^2$, $\sigma_2^2 = 0.155^2$. The probability of observing a difference as small as $\bar{x}_{50} - \bar{y}_{40} = -0.14$ is

$$P\left(\bar{X}_{50} - \bar{Y}_{40} \leq -0.14\right) = P\left(Z \leq -4.848\right) \approx 0$$

We conclude it is unlikely to observe this difference.

1.3 Exact distributions

In the following theorem, we state the exact sampling distribution of S^2 when we have a random sample from a normal distribution

Theorem The statistic

$$\chi^2 = \frac{(n-1)\,S^2}{\sigma^2} = \frac{\sum \left(X_i - \bar{X}_n\right)^2}{\sigma^2}$$

where S^2 is the sample variance from a random sample of size n drawn from a normal population has a chi-squared distribution with $\nu = n - 1$ degrees of freedom.

Example A professor claims that IQ scores for college students have a variance of 100. To test the claim, he takes a random sample of 23 students and computes the sample variance to be 147.82.The chi squared value is then

$$\begin{aligned} \chi^2 &= \frac{(n-1)\,S^2}{\sigma^2} \\ &= 22\frac{147.82}{100} = 32.52 \end{aligned}$$

Is this value too large or too small? From the Chi square table A.5, we see that 32.52 is to the left of the upper 5% point given to be 33.924. We conclude that there is substance to the claim.

1.3.1 t-distribution

In the next theorem, we state the exact distribution of the sample mean when we have a random sample from a normal distribution with unknown variance.

Theorem Let Z be a standard normal random variable and V a chi-squared random variable with ν degrees of freedom. If Z, V are independent, then

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

has a Student's t-distribution with ν degrees of freedom. This density is given by

$$h(t) = \frac{\Gamma\left(\frac{\nu+1}{2}\right)}{\Gamma\left(\frac{\nu}{2}\right)\sqrt{\pi\nu}}\left(1+\frac{t^2}{\nu}\right)^{-\left(\frac{\nu+1}{2}\right)}, -\infty < t < \infty$$

Corollary Let $X_1, ..., X_n$ be a random sample from a normal distribution with mean μ and variance σ^2. Then

$$T = \sqrt{n}\left(\frac{\bar{X}_n - \mu}{S}\right)$$

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

Example A man throws a shot put n =6 times with distances in feet equal to $58, 69, 62, 55, 64, 65$. Let X be the random variable representing the distance. Assume X is normally distributed with mean μ and variance σ^2. The man claims that his average throw distance is $\mu = 65$. Is he telling the truth?

To test the claim, we compute $\bar{x}_6 = 62.2, s = 5.04$. Then

$$\begin{aligned} P\left(|\bar{X}_6 - 65| > 2.8\right) &= P\left(|T| > \frac{2.8}{5.04}\sqrt{6}\right) \\ &= P\left(|T| > 1.36\right) \\ &= 2P\left(T > 1.36\right) \end{aligned}$$

We see that from the Student t distribution with $n - 1$ =5 degrees of freedom in Table A.4 p.439

$$(0.10) < P\left(T > 1.36\right) < (0.15)$$

There does not appear to be any evidence that the man is not telling the truth.

1.3.2 Fisher distribution

In the next theorem, we state the exact distribution of the ratio of sample variances when we have two independent random samples from normal distributions.

Theorem Let U and V be two independent random variables having chi-squared distributions with ν_1, ν_2 degrees of freedom respectively. The distribution of

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

is the Fisher distribution with ν_1 degrees of freedom in the numerator and ν_2 degrees of freedom in the denominator.

Corollary If $S_1^2$and S_2^2 are the sample variances with sizes n_1, n_2 based on two independent normal populations having variances σ_1^2, σ_2^2 respectively, then

$$F = \frac{\frac{S_1^2}{\sigma_1^2}}{\frac{S_2^2}{\sigma_2^2}}$$

has an F-distribution with $\nu_1 = n_1 - 1$,$\nu_2 = n_2 - 1$ degrees of freedom.

Example In a midterm there were two versions of the same exam. The following data was recorded

	n	$\bar{x}$	s^2
Version 1	**82**	**9.171**	**9.946**
Version 2	**73**	**9.438**	**9.194**

Is it reasonable to assume that the variances are equal?

We calculate

$$F = \frac{9.946}{9.194} = 1.08$$

The probability of observing a value larger is about 0.01 from table A.6. We conclude that it is not reasonable to assume the variances are the same for both versions.

Some data

Tensile strength

Cotton %	Tensile Strength
15	7, 7, 9, 8, 10
20	19, 20, 21, 20, 22
25	21, 21, 17, 19, 20
30	8, 7, 8, 9, 10

Nicotine data

1.09	1.92	2.31	1.79	2.28	1.74	1.47	1.97
0.85	1.24	1.58	2.03	1.70	2.17	2.55	2.11
1.86	1.90	1.68	1.51	1.64	0.72	1.69	1.85
1.82	1.79	2.46	1.88	2.08	1.67	1.37	1.93
1.40	1.64	2.09	1.75	1.63	2.37	1.75	1.69