SAMPLING DISTRIBUTIONS AND THE CENTRAL LIMIT THEOREM (Chapter 7)

Some theory regarding the normal distribution

Theorem 1 Suppose that $Y_1,\ldots,Y_n \sim$ iid $N(\mu,\sigma^2)$.

Let $\bar{Y} = \frac{1}{n}\sum_{i=1}^{n} Y_i$ (the *sample mean*).

Then $\bar{Y} \sim N(\mu, \sigma^2/n)$.

Proof In Chapter 6 we showed that linear combinations of normal rv's are also normal. So it only remains to find the mean and variance of $\bar{Y}$.

$$E\bar{Y} = \frac{1}{n}\sum_{i=1}^{n} EY_i = \frac{1}{n}\sum_{i=1}^{n} \mu = \frac{1}{n} n\mu = \mu .$$

$$Var\bar{Y} = \frac{1}{n^2}\sum_{i=1}^{n} VarY_i = \frac{1}{n^2}\sum_{i=1}^{n} \sigma^2 = \frac{1}{n^2} n\sigma^2 = \frac{\sigma^2}{n}. \quad \text{QED}$$

More direct proof: $m_{\bar{Y}}(t) = Ee^{\bar{Y}t} = Ee^{t\left(\frac{Y_1+\ldots+Y_n}{n}\right)} = E\left(e^{t\frac{Y_1}{n}} \times \ldots \times e^{t\frac{Y_n}{n}}\right) = \left(Ee^{t\frac{Y_1}{n}}\right)\ldots\left(Ee^{t\frac{Y_n}{n}}\right)$

$$= m_{Y_1}\left(\frac{t}{n}\right)\ldots m_{Y_n}\left(\frac{t}{n}\right) = \left\{m_{Y_1}\left(\frac{t}{n}\right)\right\}^n = \left\{e^{\mu\left(\frac{t}{n}\right)+\frac{1}{2}\sigma^2\left(\frac{t}{n}\right)^2}\right\}^n = e^{\mu t+\frac{1}{2}\frac{\sigma^2}{n}t^2} \Rightarrow \bar{Y} \sim N\left(\mu, \frac{\sigma^2}{n}\right).$$

Corollary $Z = \dfrac{\bar{Y}-\mu}{\sigma/\sqrt{n}} \sim N(0,1)$.

Z may be called the *standardised sample mean*.

The sample mean $\bar{Y}$ is an example of a statistic.

Definition: A *statistic* is any function of the observable random variables in a sample and known constants.

The probability distribution of a statistic is sometimes referred to as the *sampling distribution* of that statistic.

For example, $\bar{Y}-\mu$ above is not a statistic (unless μ is known).
Also, Z above is not a statistic (unless both μ and σ are known).

We call: μ the *population mean*

$\bar{Y}$ the *sample mean*

$E\bar{Y}$ the *mean of the sample mean*

$N(\mu, \sigma^2/n)$ the *sampling distribution of the sample mean.*

Example 1 A bottling machine discharges volumes of drink that are independent and normally distributed with standard deviation 1 ml.

Find the sampling distribution of the mean volume of 9 randomly selected bottles that are filled by the machine.

Hence find the probability that this sample mean will be within 0.3 ml of the mean volume of all bottles filled by the machine.

(NB: The latter mean is the population mean, μ.)

Let Y_i be the volume of the *i*th bottle in the sample, *i* = 1,...,*n*, where *n* = 9.
Then $Y_1,\ldots,Y_n \sim$ iid $N(\mu, \sigma^2)$, where $\sigma^2 = 1$ and μ is unknown.
So $\bar{Y} \sim N(\mu, 1/9)$. (This is the sampling distribution of the sample mean.)

Hence $P(|\bar{Y}-\mu|<0.3) = P\left(\left|\frac{\bar{Y}-\mu}{\sigma/\sqrt{n}}\right| < \frac{0.3}{1/3}\right) = P(|Z|<0.9)$ where $Z \sim N(0,1)$

$= 1 - 2P(Z > 0.9)$

$= 1 - 2(0.1841)$ by tables

$= 0.6318.$

(Note that this calculation did not require knowledge of μ.)

Theorem 2 Suppose that $Y_1,\ldots,Y_n \sim$ iid $N(\mu, \sigma^2)$.

Let $S^2 = \frac{1}{n-1}\sum_{i=1}^{n}(Y_i-\bar{Y})^2$ (the *sample variance*).

Then: **(a)** $\frac{(n-1)S^2}{\sigma^2} \sim \chi^2(n-1)$

(b) $S^2 \perp \bar{Y}$.

(*Note*: The proof of this theorem is beyond the scope of this course and therefore non-assessable.)

Example 2 Refer to Example 1. Find an interval which we can be 90% sure will contain the sample variance of the 9 sampled volumes.

We will solve $P(a < S^2 < b) = 0.9$ for a and b.

$$0.9 = P\left(\frac{(9-1)a}{1} < \frac{(n-1)S^2}{\sigma^2} < \frac{(9-1)b}{1}\right) = P(8a < U < 8b)$$

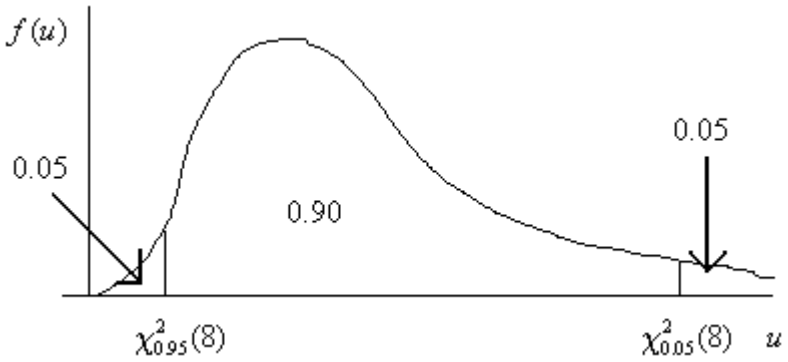
where $U = \frac{(n-1)S^2}{\sigma^2} \sim \chi^2(8)$.

Now 0.9 = $P(2.73264 < U < 15.5073)$ by the χ^2 tables on pages 850-851 in the text.

Therefore $8a = 2.73264$ and $8b = 15.5073$. Hence $a = 0.34158$ and $b = 1.93841$.

So the required interval is (0.342, 1.938).

Note 1: We say that 15.5073 is the *upper* 0.05-*quantile* of the chi square distribution with 8 degrees of freedom. This quantile may be denoted $\chi^2_{0.05}(8)$.



Note 2: The above solution is not unique. A different interval is obtained by seeing from the tables that $0.9 = P(0 < U < 13.3616)$. We now equate 13.3616 with $8b$ to get the interval (0,1.6702).

As a third example, we see from the tables that $0.9 = P(3.48954 < U < \infty)$. We now equate 3.48954 with $2a$ to get the interval $(0.4362, \infty)$.

Note 3: A related problem is: "Find the *shortest* interval which we can be 90% sure will contain the sample variance of the 9 sampled volumes."

The solution to this problem is unique and the required interval (a,b) satisfies two equations: $P(a < S^2 < b) = 0.9$ and $f_{S^2}(a) = f_{S^2}(b)$.

This solution will have to be found numerically, e.g. trial and error or the Newton Raphson algorithm. (Don't worry about this for now.)

The *t* distribution

Definition: Suppose that $Z \sim \text{N}(0,1)$, $U \sim \chi^2(k)$ and $Z \perp U$. We say that the rv

$$Y = \frac{Z}{\sqrt{U/k}}$$

has the *t-distribution* with *k* degrees of freedom. The pdf of *Y* is

$$f(y) = \frac{\Gamma\left(\frac{k+1}{2}\right)}{\sqrt{k\pi}\,\Gamma(k/2)}\left(1+\frac{y^2}{k}\right)^{-\frac{1}{2}(k+1)}, \quad -\infty < y < \infty.$$

(See Exercise 7.98 for a proof, but note that this proof is non-assessable.)

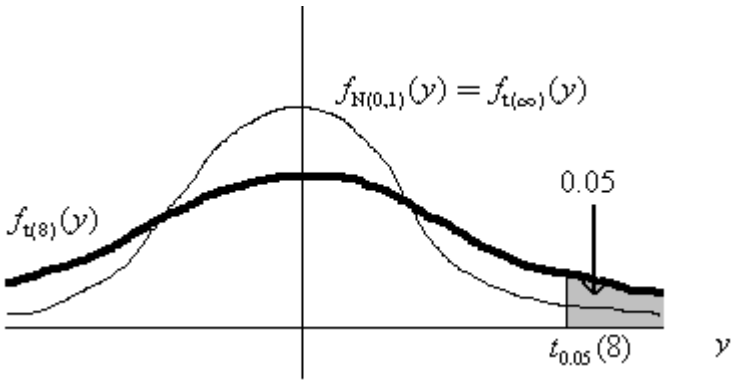
We write $Y \sim t(k)$ (or $Y \sim t_k$), and $f(y)$ as $f_{t(k)}(y)$.

The *t* pdf looks like a standard normal pdf but with 'fatter' tails.
The *t* dsn converges to the standard normal dsn as *k* tends to infinity.

Proof: The pdf of *Y* can be written $f(y) = c_k A_k(y) B_k(y)$, where c_k is a constant (which does not depend on *y*), $A_k(y) = \left\{\left(1+\frac{y^2}{k}\right)^k\right\}^{-1/2}$ and $B_k(y) = \left(1+\frac{y^2}{k}\right)^{-1/2}$.

Now, as $k \to \infty$, $B_k(y) = \left(1+\frac{y^2}{k}\right)^{-1/2} \to 1$ and $A_k(y) \to \left\{e^{(y^2)}\right\}^{-1/2} = e^{-\frac{1}{2}y^2}$.

So as $k \to \infty$, the pdf of $Y \sim t(k)$ converges proportionally to $e^{-\frac{1}{2}y^2}$, which is the kernel of the $N(0,1)$ distribution. QED. (It can also be shown that $c_k \to 1/\sqrt{2\pi}$.)



Upper quantiles of the *t* distribution are tabulated on the inside front cover of the text (and elsewhere). For example, $t_{0.05}(8) = 1.860$. As regards notation, we may also write $F_{t(8)}(1.86) = 0.95$, $F^{-1}_{t(8)}(0.95) = 1.86$, $t_p(k) = F^{-1}_{t(k)}(1-p)$, etc. $F^{-1}_{t(k)}(p)$ is the (lower) *p*-quantile function of the $t(k)$ dsn, and $t_p(k)$ is the upper *p*-quantile of the $t(k)$ dsn.

Theorem 3 Suppose that $Y_1,\ldots,Y_n \sim$ iid $N(\mu,\sigma^2)$. Let $T = \dfrac{\bar{Y}-\mu}{S/\sqrt{n}}$.

Then $T \sim t(n-1)$.

Proof Observe that: **(a)** $Z = \dfrac{\bar{Y}-\mu}{\sigma/\sqrt{n}} \sim N(0,1)$ by the Corollary to Theorem 1

(b) $U = \dfrac{(n-1)S^2}{\sigma^2} \sim \chi^2(n-1)$ by Theorem 2 Part (a)

(c) $Z \perp U$ by Theorem 2 Part (b).

It follows by definition of the t distribution that

$$Y = \frac{Z}{\sqrt{U/(n-1)}} \sim t(n-1).$$

But $Y = \dfrac{\dfrac{\bar{Y}-\mu}{\sigma/\sqrt{n}}}{\sqrt{\dfrac{(n-1)S^2}{\sigma^2}\Big/(n-1)}} = \dfrac{\bar{Y}-\mu}{S/\sqrt{n}} = T.$

Hence $T \sim t(n-1)$.

Example 3 Refer to Example 1.

Find the probability that the mean of the 9 sample volumes will be distant from the population mean by no more than half the sample standard deviation of those 9 volumes.

$$P(|\bar{Y}-\mu| < 0.5S) = P\left(\left|\frac{\bar{Y}-\mu}{S/\sqrt{n}}\right| < \frac{0.5\cancel{S}}{\cancel{S}/3}\right) = P(|T|<1.5) \quad \text{where } T \sim \text{t(8)}$$

$$= 1 - 2P(T > 1.5).$$

Now by tables, $P(T > 1.397) = 0.10$ and $P(T > 1.860) = 0.05$.
Therefore $P(T > 1.5)$ is between 0.05 and 0.10.
So $1 - 2P(T > 1.5)$ is between $1 - 2(0.10) = 0.8$ and $1 - 2(0.05) = 0.9$.

Thus the required probability is between 80% and 90%. (That's the best we can do using the tables available in the text book. The exact answer is 0.8279967, obtained using the *R* code `1-2*(1-pt(1.5,8))`. Note that this *R* code is non-assessable.)

The *F* distribution

Definition: Suppose that $U \sim \chi^2(a)$, $V \sim \chi^2(b)$ and $U \perp V$.

We say that the random variable $Y = \dfrac{U/a}{V/b}$

has the *F-distribution* with *a* numerator and *b* denominator degrees of freedom.

The pdf of *Y* is $f(y) = \dfrac{\Gamma\left(\frac{a+b}{2}\right)}{\Gamma(a/2)\Gamma(b/2)} a^{a/2} b^{b/2} y^{\frac{a}{2}-1} (b+ay)^{-\frac{1}{2}(a+b)}$, $y > 0$.

We write $Y \sim F(a,b)$ (or $Y \sim F_{a,b}$), and $f(y)$ as $f_{F(a,b)}(y)$.

The *F* pdf looks like a gamma pdf. Upper quantiles of the *F* distribution are tabulated on pages 852-861 in the text. For example, $F_{0.025}(4,17) = 3.66$.

As for other dsns, we can also write $F_{F(4,17)}(3.66) = 0.975$, $F^{-1}_{F(4,17)}(0.975) = 3.66$, $F_p(a,b) = F^{-1}_{F(a,b)}(1-p)$, etc. $F^{-1}_{F(a,b)}(p)$ is the (lower) *p*-quantile of the *F*(*a*,*b*) dsn.

(*Note*: "*F*" here is used to denote two different things, a cdf and a dsn, respectively.)

An important fact is that if $Y \sim F(a,b)$, then $X = 1/Y \sim F(b,a)$. (This is eaily proved.)

Theorem 4 Suppose that: $X_1,\ldots,X_n \sim$ iid $N(\mu_X, \sigma_X^2)$ (1st sample)

$Y_1,\ldots,Y_m \sim$ iid $N(\mu_Y, \sigma_Y^2)$ (2nd sample)

$(X_1,\ldots,X_n) \perp (Y_1,\ldots,Y_m)$

(the two samples are independent)

$\bar{X} = \dfrac{1}{n}\sum_{i=1}^{n} X_i$ (1st sample mean)

$\bar{Y} = \dfrac{1}{m}\sum_{i=1}^{m} Y_i$ (2nd sample mean)

$S_X^2 = \dfrac{1}{n-1}\sum_{i=1}^{n} (X_i - \bar{X})^2$ (1st sample variance)

$S_Y^2 = \dfrac{1}{m-1}\sum_{i=1}^{m} (Y_i - \bar{Y})^2$ (2nd sample variance).

Let $W = \dfrac{S_X^2/\sigma_X^2}{S_Y^2/\sigma_Y^2}$.

Then $W \sim F(n-1, m-1)$. (Prove as an exercise.)

Example 4 Refer to Example 1.

Suppose that another sample of 5 bottles is to be taken from the output of the same bottling machine.

Find the probability that the sample variance of the volumes in these 5 bottles will be at least 7 times as large as the sample variance of the volumes in the 9 bottles that were initially sampled.

$$P(S_X^2 > 7S_Y^2) = P\left(\frac{S_X^2}{S_Y^2} > 7\right)$$

$$= P\left(\frac{S_X^2/\sigma_X^2}{S_Y^2/\sigma_Y^2} > 7\right) \quad \text{since } \sigma_X^2 = \sigma_Y^2$$

$= P(U > 7)$ where $U \sim F(4,8)$

$\approx P(U > 7.01)$

$= 0.010$ by the F tables.

(Note that to obtain this answer we didn't need to know the common population variance, $\sigma^2 = \sigma_X^2 = \sigma_y^2$, nor the common population mean, $\mu = \mu_X = \mu_Y$.)