Continuous Random Variables

General Information

- A function $f\colon \mathbb{R} \to \mathbb{R}$ is a *probability mass function* (pdf) of a continuous random variable X iff f is nonnegative and $\int_{-\infty}^{\infty} f(x)\,dx = 1$.
- For any probability mass function f, we have $\mathrm{P}(a \leq X \leq b) = \int_a^b f(x)\,dx$. Whether the inequality is strict or nonstrict does not affect the above identity.
- A *mode* of X is any value m such that $f(m)$ is maximum.
- A *cumulative distribution function* (cdf) $F\colon \mathbb{R} \to [0,1]$ of a random variable X is defined by

$$F(x) \coloneqq P(X \leq x) = \int_{-\infty}^{x} f(x)\,dx.$$

- When writing out the cdf as a piecewise function, we explicitly write out the range of values for each case. We reserve the use of "otherwise" for pdf's.
- Any cdf is continuous and nondecreasing.
- Let X be a continuous random variable with cdf F. To find the pdf g of any $y(X)$, we first find its cdf, then differentiate. We achieve this by reverse engineering $y(X) \leq y$ to find an inequality that relates X with y. E.g. $e^X \leq y$ iff $X \leq \ln(y)$.
- A *median* of X is any value m such that $\mathrm{P}(X \leq m) = F(m) = 1/2$.
- Mean/Expectation:

$$\mu = \mathrm{E}(X) \coloneqq \int_{-\infty}^{\infty} x f(x)\,dx \qquad \text{and} \qquad \mathrm{E}(g(X)) = \int_{-\infty}^{\infty} g(x) f(x)\,dx.$$

- Important property:

$$\mathrm{E}(ag(X) \pm bh(x)) = a\,\mathrm{E}(g(X)) \pm \mathrm{E}(h(X)).$$

- Variance:

$$\mathrm{Var}(X) \coloneqq \mathrm{E}(X^2) - [\mathrm{E}(X)]^2.$$

- Important property:

$$\mathrm{Var}(aX \pm b) = a^2\,\mathrm{Var}(X).$$

Special Continuous Random Variables

Definition 2.1

A continuous random variable X has a *normal distribution* with mean μ and standard deviation σ, denoted by $X \sim \mathrm{N}(\mu, \sigma^2)$, iff its pdf f is such that

$$f(x) = \frac{1}{\sigma\sqrt{2\pi}} \exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right).$$

General Information

- A normal distribution is symmetrical about the line $x = \mu$. That is

$$\mathrm{P}(X \leq \mu - \delta) = \mathrm{P}(X \geq \mu + \delta)$$

  for each $\delta > 0$. Note that the mean, median, and mode coincide with μ.

- Properties of the normal distribution. Let X and Y be independent, such that $X \sim \mathrm{N}(\mu, \sigma^2)$ and $Y \sim \mathrm{N}(m, s^2)$. Then, for any $n \in \mathbb{N}$ and $x, y \in \mathbb{R}$,
  - $nX \sim \mathrm{N}(n\mu, n^2\sigma^2)$,
  - $X_1 + X_2 + \cdots + X_n \sim \mathrm{N}(n\mu, n\sigma^2)$,
  - $aX \pm bY \sim \mathrm{N}(a\mu \pm bm, a^2\sigma^2 + b^2s^2)$.

- At times, the question may be phrased in a misleading manner. Try using some inference to figure out the intended interpretation.

Example 2.1

"The mass of the padding is 30% of the mass of a randomly selected light bulb of mass L. Find the probability that a light bulb with padding has mass c."
Then for any light bulb of mass L_1, the mass of the padding is $0.3L_2$ (and *not* $0.3L_1$). i.e. we are to find $\mathrm{P}(L_1 + 0.3L_2)$.

- A variable $Z \sim \mathrm{N}(0, 1)$ is said to follow the *standard* normal distribution.

  Note: Z is reserved for this purpose.

- Let $X \in \mathrm{N}(\mu, \sigma^2)$. Then, $\frac{X-\mu}{\sigma}$ follows the standard normal distribution.

- What `Tail` do we select for `invNorm`?

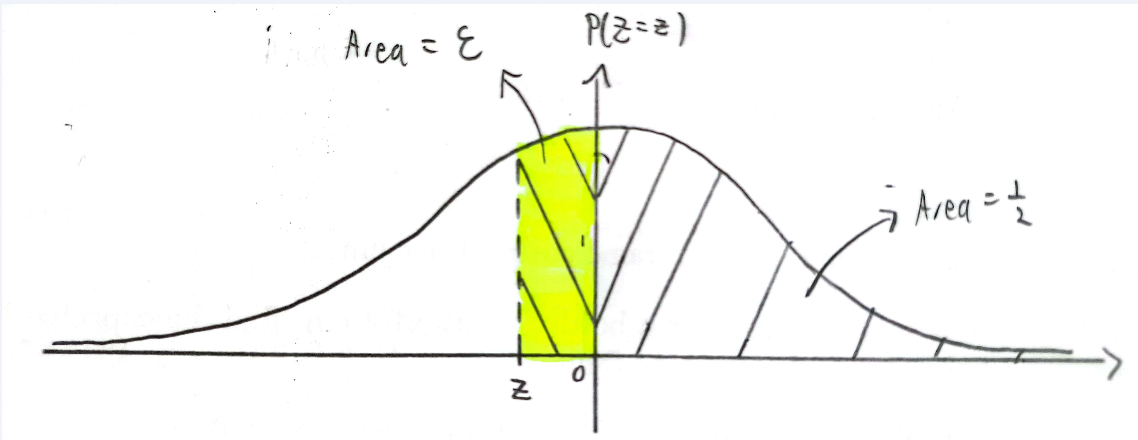
$\mathrm{P}(X < x) = p$	`LEFT`
$\mathrm{P}(-x < X < x) = p$	`CENTER`
$\mathrm{P}(X > x) = p$	`RIGHT`

- When using `invNorm` on an inequality, what should the sign be? For simplicity, we write $\mathscr{L}(p) = \texttt{invNorm}(p, 0, 1, \texttt{RIGHT})$, and $\mathscr{R}(p) = \texttt{invNorm}(p, 0, 1, \texttt{LEFT})$. Then,

$\mathrm{P}(Z > z) \geq p$	$z \leq \mathscr{L}(p)$
$\mathrm{P}(Z > z) \leq p$	$z \geq \mathscr{L}(p)$
$\mathrm{P}(Z < z) \geq p$	$z \geq \mathscr{R}(p)$
$\mathrm{P}(Z < z) \leq p$	$z \leq \mathscr{R}(p)$

Example 2.2

Suppose we want to find the least integer value of m for which $\mathrm{P}(Z > 1 - m) \geq 1/2$.
Then, using `invNorm (RIGHT)`, we infer that $z \leq 0$, *not* $z \geq 0$. An illustration:



Definition 2.2

A continuous random variable X has a *uniform distribution* over the interval (a, b), which is denoted by $X \sim \mathrm{U}(a, b)$, iff its pdf f is such that

$$f(x) = \begin{cases} \frac{1}{b-a} & \text{if } a < x < b, \\ 0 & \text{otherwise.} \end{cases}$$

Definition 2.3

A continuous random variable Y has an (negative) exponential distribution, which we denote with $Y \sim \mathrm{Exp}(\lambda)$, iff its pdf g is such that

$$g(Y) = \begin{cases} \lambda e^{-\lambda x} & \text{if } x \geq 0, \\ 0 & \text{otherwise.} \end{cases}$$

(An exponential distribution models time between occurrences.)

Note

Let $Y \sim \mathrm{Exp}(\lambda)$, then

$$\mathrm{P}(Y > z + y \mid Y > y) = \mathrm{P}(Y > z) \qquad \text{and} \qquad \mathrm{P}(Y < z + y \mid Y > y) = \mathrm{P}(Y < z).$$

- Expectation and variance:

Distribution	Expectation	Variance
$X \sim \mathrm{U}(a, b)$	$\dfrac{a+b}{2}$	$\dfrac{(b-a)^2}{12}$
$Y \sim \mathrm{Exp}(\lambda)$	$\dfrac{1}{\lambda}$	$\dfrac{1}{\lambda^2}$

  Note: We need to remember the expectation and variance for the uniform distribution, as it is not provided in the MF26 formula sheet (unlike all other distributions).

- *Warning*: The G.C. tends to incorrectly process an integral if its upper and lower bounds contain $\pm$E99.

- Let T be the time taken between two consecutive arrivals and $\# \sim \text{Po}(\lambda t)$ the number of arrivals in time t. Then,

  $$\text{P}(T > t) = \text{P}(\# = 0) = e^{-4t}.$$

  As such, the probability that there is at least one arrival in an interval of time t is

  $$\text{P}(T \leq t) = 1 - e^{-4t}.$$

Sampling and Estimation

Definition 3.1
A sample is a finite subset of the population.

Definition 3.2
A random sample is a sample selected such that each member of the population has an equal probability of being selected into the sample.

Note
State, in context, what it means for the sample to be random.

> It means that every [a member of the population] has an equal probability of being selected into the sample.

Note
Explain why the sample would actually not be random.

> [Contextual reason], so not all the [members of the population] have an equal probability of being selected into the sample.

Definition 3.3
Any statistic T derived from a random sample and used to estimated an unknown population parameter θ is known as an *estimator*. It is an *unbiased* estimator iff $\mathrm{E}(T) = \theta$. If T is unbiased we commonly write $\hat{\theta}$ for T.

General Information

- Either write $\hat{\mu} = \bar{x} = \ldots$ or write out "Unbiased estimate of the population mean μ, $\bar{x} = \ldots$" Same holds for other population parameters θ.
- Estimators you should know:

Parameter	Estimator	Unbiased?	Formula
Population Mean μ	Sample Mean $\overline{X}$	✓	$\dfrac{X_1 + X_2 + \cdots + X_n}{n}$
Population Variance σ^2	Sample Variance σ_n^2	×	$\dfrac{\sum (X_i - \overline{X})^2}{n}$ $\dfrac{\sum X_i^2}{n} - \overline{X}^2$
	S^2	✓	$\dfrac{n}{n-1}\sigma_n^2$ $\dfrac{\sum (X_i - \overline{X})^2}{n-1}$ $\dfrac{1}{n-1}\left[\sum X_i^2 - \dfrac{(\sum X_i)^2}{n}\right]$
Population Proportion p	Sample Proportion P_s	✓	$\dfrac{X}{n}$

- Let X be a random variable following *any distribution*, and suppose we have a random sample $X_1, X_2, \ldots, X_n$ of size $n \geq 50$. Then by CLT (Central Limit Theorem), since $n \geq 50$ is large,

$$\overline{X} \sim \mathrm{N}\left(\mu, \frac{\sigma^2}{n}\right) \qquad \text{and} \qquad X_1 + X_2 + \cdots + X_n \sim \mathrm{N}(n\mu, n\sigma^2)$$

approximately.

- Assumptions when using CLT:
  - The sample is random.
  - Each X_i is independent and identically distributed.
- Suppose $X \sim \mathrm{N}(\mu, \sigma^2)$ is known and we pick a *particular* sample. Then,

Distribution	Is An Approximation?
$\overline{X} \sim \mathrm{N}(\mu, \sigma^2)$	No
$\overline{X} \sim \mathrm{N}(\bar{x}, \sigma^2)$	Yes
$\overline{X} \sim \mathrm{N}(\mu, s^2)$	Yes
$\overline{X} \sim \mathrm{N}(\bar{x}, s^2)$	Yes

  So, if we obtain any of the latter three in solving a question, we must write "$X \sim \mathrm{N}(__, __)$approximately" (even though we knew X *exactly* follows a normal distribution!)

- Pooled estimators. First assume we have two populations, from which we select a random sample of size n_1 and n_2. We let $\overline{X}_1$ and S_1^2 denote the sample mean and unbiased estimator for variance, respectively, for the first sample. Similarly define $\overline{X}_2$ and S_2^2, for the second sample.

Parameter	Unbiased Pooled Estimator
Mean	$\hat{\mu} = \dfrac{n_1\overline{X}_1 + n_2\overline{X}_2}{n_1 + n_2}$
Variance	$S_p^2 = \dfrac{(n_1-1)S_1^2 + (n_2-1)S_2^2}{n_1+n_2-2}$

The following definition is found in Hogg-McKean-Craig. Similar definitions are also found in Wackerly-Mendenhall-Schaefer and Nitis Mukhopadhyay.

Definition 3.4

Let $X_1, X_2, \ldots, X_n$ be a sample on a random variable X, where X has pdf $f(x;\theta)$, $\theta \in \Omega$. Let $0 < \alpha < 1$ be specified. Let $L = L(X_1, X_2, \ldots, X_n)$ and $U = U((X_1, X_2, \ldots, X_n))$ be two statistics. We say that the interval (L, U) is a $(1-\alpha)100\%$ *confidence interval* for θ iff

$$1 - \alpha = P_\theta[\theta \in (L, U)].$$

That is, the probability that the interval contains θ is $1-\alpha$, which is called the *confidence coefficient* or *confidence level* of the interval.

- We cannot write "a $1-\alpha$ (e.g. 0.95) confidence interval". The $1-\alpha$ must always be expressed as a *percentage*.
- Let $\hat{\theta}$ be a statistic that is normally distributed with mean θ and standard error $\sigma_{\hat{\theta}}$. We see that

$$\frac{\hat{\theta} - \theta}{\sigma_{\hat{\theta}}} = Z \sim \mathrm{N}(0, 1).$$

  Rewriting $\mathrm{P}(-z_{1-\alpha/2} < Z < z_{1-\alpha/2}) = 1 - \alpha$ gives

$$\mathrm{P}(\hat{\theta} - z_{1-\alpha/2}\sigma_{\hat{\theta}} < \theta < \hat{\theta} + z_{1-\alpha/2}\sigma_{\hat{\theta}}) = 1 - \alpha.$$

Hence, a $(1-\alpha)100\%$ confidence interval for θ is

$$(\hat{\theta} - z_{1-\alpha/2}\sigma_{\hat{\theta}}\,,\ \hat{\theta} + z_{1-\alpha/2}\sigma_{\hat{\theta}}).$$

(Wackerly-Mendenhall-Schaefer)

- Let $0 < \alpha < 1$ and $X_1, X_2, \ldots, X_n$ be a sample on a random variable X with mean μ, where n is large. Then, an approximate $(1-\alpha)100\%$ confidence interval for μ is

$$\left(\bar{x} - z_{1-\alpha/2}\frac{s}{\sqrt{n}}\,,\ \bar{x} + z_{1-\alpha/2}\frac{s}{\sqrt{n}}\right).$$

When the variance σ^2 is known, we can replace s with σ. If the distribution of X is known to be normal, in addition to σ^2 being known exactly, then the confidence interval is exact; it is not just an approximation.

(Hogg-McKean-Craig)

- Let X be a Bernoulli random variable with probability of success p, where X is 1 or 0 if the outcome is success or failure, respectively. Suppose $X_1, X_2, \ldots, X_n$ is a random sample from the distribution of X, where n is large. Let $\hat{p} = \overline{X}$ be the sample proportion of successes. Then, an approximate $(1-\alpha)100\%$ confidence interval for p is given by

$$\left(\hat{p} - z_{1-\alpha/2}\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}\,,\ \hat{p} + z_{1-\alpha/2}\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}\right).$$

(Letting $Y = X_1 + X_2 + \cdots + X_n \sim \mathrm{B}(n,p)$ gives $\hat{p} = Y/n$, which is the presentation used in the school's notes.)

(Hogg-McKean-Craig)

Note

Standard phrasing for the interpretation of a $(1-\alpha)100\%$ confidence interval (a, b).

The probability that the interval (a, b) contains the true value of the [population mean/proportion in context] is $1-\alpha$.

Note

Standard phrasing for what is a $(1-\alpha)100\%$ confidence interval for θ?

It is an interval which has probability $1-\alpha$ of containing the true value of θ.

Note

Standard phrasing for whether mean/proportion in context has likely increased/decreased, when given suitable confidence intervals.

1. There is no conclusive result.

   Since the old and new $(1-\alpha)\%$ confidence intervals overlap, we are unable to conclude whether the [mean/proportion in context] has decreased or not. Hence, it is inconclusive from these figures as to whether the [context (e.g. an awareness campaign)] has been effective.

2. It has likely increased/decreased.

   The old $(1-\alpha)\%$ confidence interval is to the left/right of the new $(1-\alpha)\%$ confidence interval, such that they do not overlap. So, can conclude that the [mean/proportion in context] likely increased/decreased. Hence, these figures suggests that the [context (e.g. an awareness campaign)] has been effective.

Note

Advantage and disadvantage of a $(1-\beta)\%$ confidence interval compared to a $(1-\alpha)\%$ confidence interval, where $\beta < \alpha$.

Advantage:	A $(1-\beta)\%$ CI is more likely to contain the true mean.
Disadvantage:	A $(1-\beta)\%$ CI is less precise (or wider).

Note. Clearly state which is the advantage and disadvantage, as illustrated above.

G.C. Skills

Calculating statistics (i.e. $\bar{x}$, s, etc) by G.C. given data for a sample.

1. Keying in the data: `stat` $\Longrightarrow$ `1:Edit` $\Longrightarrow$ Key in the data into one of the lists $\mathtt{L}_i$.
2. Calculating the statistic: `stat` $\Longrightarrow$ `CALC` $\Longrightarrow$ `1-Var Stats` (`List:`$\mathtt{L}_i$) $\Longrightarrow$ `Calculate`.
3. Getting the statistic for further calculations: `vars` $\Longrightarrow$ `5:Statistics` $\Longrightarrow$ Select the desired statistic.

G.C. Skills

Calculating the symmetric confidence interval for a normally distributed random variable.

Mean:	`stat` $\Longrightarrow$ `TESTS` $\Longrightarrow$ `7:ZInterval...`
Proportion:	`stat` $\Longrightarrow$ `TESTS` $\Longrightarrow$ `A:1-PropZInt...`

Statistics: Hypothesis Testing

4.1 General Information

Definition 4.1

The *null hypothesis* H_0 and *alternative hypothesis* H_1 are the hypotheses that we hope to reject and accept, respectively.

General Information

- Without going into details, a *critical region* C is just a set that defines the decision rule / test

$$\text{Reject } H_0 \text{ (Accept } H_1) \quad \text{if } (X_1, X_2, \ldots, X_n) \in C,$$

for any random sample $X_1, X_2, \ldots, X_n$ from the distribution of a random variable X.

Definition 4.2

The *significance level* $100\alpha\%$ of a test is the probability of rejecting H_0 when it is in fact true. i.e. $\alpha = \mathrm{P}(H_0 \text{ is rejected} \,|\, H_0 \text{ is true})$.

Note

Explain, in context, the meaning of 'at the $\alpha\%$ level of significance'.

The probability that [H_1 in context], when actually [H_0 in context], is $\alpha\%$.

Definition 4.3

The *p-value* is the lowest level of significance for which the null hypothesis will be rejected. In other words, for the null hypotheses

$$\text{(a) } \mu < \mu_0, \quad \text{(b) } \mu \neq \mu_0, \quad \text{(c) } \mu > \mu_0,$$

we have

$$\text{(a) } p\text{-value} = \mathrm{P}(Z \leq z_{\text{calc}}), \quad \text{(b) } p\text{-value} = \mathrm{P}(|Z| \geq |z_{\text{calc}}|), \quad \text{(c) } p\text{-value} = \mathrm{P}(Z \geq z_{\text{calc}}).$$

Note

Explain what the p-value means in context.

The p-value is the least level of significance to conclude that [H_1 in context].

- One sample z-test. There are various combination of assumptions for which this test applies. For brevity, we shall avoid restating it, instead directing the reader to table 4.2

1. Let [X in context] and μ be the population mean.
2. $\boxed{\begin{array}{ll} \text{Test} & H_0\colon \mu = \mu_0 \\ \text{against} & H_1\colon \text{ (a) } \mu < \mu_0, \quad \text{(b) } \mu \neq \mu_0, \quad \text{or} \quad \text{(c) } \mu > \mu_0, \\ \multicolumn{2}{l}{\text{at the } 100\alpha\% \text{ significance level.}} \end{array}}$
3. Under H_0, we have $\overline{X} \sim \mathrm{N}(\mu_0, \hat{\sigma}^2/n)$ approximately. Or, if σ^2 is known exactly, then by CLT $\overline{X} \sim \mathrm{N}(\mu_0, \sigma^2/n)$ approximately.
4. Test statistic:

$$Z = \frac{\overline{X} - \mu_0}{\sigma/\sqrt{n}} \sim \mathrm{N}(0, 1).$$

4. Find $z_{1-\alpha}$ or $z_{1-\alpha/2}$, which satisfies
   (a) $\mathrm{P}(Z < z_{1-\alpha}) = \alpha$,
   (b) $\mathrm{P}(-z_{1-\alpha/2} < Z < z_{1-\alpha/2}) = 1 - \alpha$, or
   (c) $\mathrm{P}(Z > z_{1-\alpha})$.
5. Find the test statistic value
$$z_{\text{calc}} = \frac{\hat{\mu} - \mu_0}{\sigma/\sqrt{n}}.$$
6. Reject H_0 iff
   (a) $z_{\text{calc}} < z_{1-\alpha}$,
   (b) $|z_{\text{calc}}| > z_{1-\alpha/2}$, or
   (c) $z_{\text{calc}} > z_{1-\alpha}$.

4. Find the p-value using GC.
5. Reject H_0 iff p-value is less than α.

G.C. Skills

Calculating the p-value of a sample.

`stat` $\Longrightarrow$ `TESTS` $\Longrightarrow$ `1:Z-Test...`

7. Since (a) $z_{\text{calc}} < z_{1-\alpha}$, (b) $|z_{\text{calc}}| > z_{1-\alpha/2}$, (c) $z_{\text{calc}} > z_{1-\alpha}$, or p-value $< \alpha$, we reject H_0. There is sufficient evidence at the significance level $100\alpha\%$ that [H_1 in context].

   Note. For *not* rejecting H_0, simply change to the appropriate inequality (such that z_{calc} is outside the critical region) and write “insufficient” instead of “sufficient”.

- If we have a null hypothesis, such as
$$H_0\colon \mu \le \mu_0 \quad \text{or} \quad H_0\colon \mu \ge \mu_0,$$
  we can just use $H_0\colon \mu = \mu_0$ instead.

Note

Explain why there is no need to assume that the distribution of X is normal/know anything about the population distribution of X.

> As the sample size n is large, by the Central Limit Theorem, the sample mean of [random variable X in context] will approximately follow a normal distribution.

Note. Spell “Central Limit Theorem” and “the sample mean” out *in full.* Do not use CLT or $\overline{X}$ for this question.

Definition 4.4

random variable X follows Student’s t-distribution with ν degrees of freedom iff its pdf is
$$f(t) = \frac{\Gamma\left(\frac{\nu+1}{2}\right)}{\sqrt{\pi\nu}\,\Gamma\left(\frac{\nu}{2}\right)}\left(1 + \frac{t^2}{\nu}\right)^{-\frac{1}{2}(\nu-1)}.$$

This is denoted by $X \sim t(\nu)$.

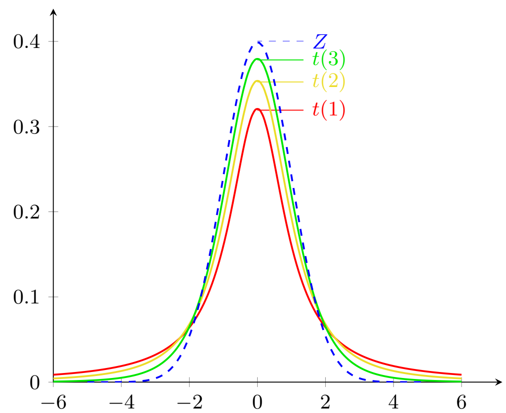


Figure 4.1: Student's t-distribution compared to the standard normal distribution.

- Properties of Student's t-distribution.

  1. It is continuous and symmetric about the vertical axis, i.e. $t = 0$.
  2. From Figure 4.1, we see that the t-distribution has a flatter peak and fatter tails, than the standard normal distribution.
  3. As $\nu \to \infty$, we have $t(\nu) \to \mathrm{N}(0, 1)$.

- Let $T \sim t(n-1)$ and $t_{(n-1,1-\alpha/2)}$ be such that $\mathrm{P}\left(-t_{(n-1,1-\alpha/2)} < T < t_{(n-1,1-\alpha/2)}\right) = 1 - \alpha$. A $(1-\alpha)100\%$ confidence interval, for the population mean μ of T, is

$$\left(\bar{x} - t_{(n-1,1-\alpha/2)}\frac{s}{\sqrt{n}}\,,\ \bar{x} - t_{(n-1,1-\alpha/2)}\frac{s}{\sqrt{n}}\right).$$

- Suppose we are conducting the following test:

Test	$H_0\colon \mu = \mu_0$
against	$H_1\colon \mu \neq \mu_0$
at a $100\alpha\%$ significance level.	

  Then, we reject H_0 iff the appropriate symmetric interval (z or t-interval) does *not* contain μ_0.

G.C. Skills

Calculating the symmetric t-confidence interval, for the population mean, of a random variable following Student's t-distribution.

`stat` $\Longrightarrow$ `TESTS` $\Longrightarrow$ `8:TInterval...`

- A one sample t-test. Again, see table 4.2 for the necessary assumptions.

  1. Let [X in context], *which we assume to be normally distributed*, and μ be the population mean.
  2.

Test	$H_0\colon \mu = \mu_0$
against	$H_1\colon$ (a) $\mu < \mu_0$, (b) $\mu \neq \mu_0$, or (c) $\mu > \mu_0$,
at the $100\alpha\%$ significance level.	

3. Under H_0, the test statistic
$$T = \frac{\overline{X} - \mu}{s/\sqrt{n}} \sim t(n-1).$$
4. Continue as per usual, calculating the critical region or the p-value.

G.C. Skills

Calculating, for a one sample t-test, the

p-value:	`stat` $\Longrightarrow$ `TESTS` $\Longrightarrow$ `2:T-Test...`
critical region:	`2nd` $\Longrightarrow$ `vars` $\Longrightarrow$ `4:invT(`

Note

In the GC, `invT` is always 'to the `LEFT`'. That is, the output t of

```
invT
area:A
df:ν
Paste
```

is such that $\mathrm{P}(T < t) = A$.

- A two-sample z-test. Again, see table 4.3 for the necessary assumptions.
  (i) σ_1 and σ_2 are known, in addition to
    (1) X_1 and X_2 being normally distributed, or
    (2) both sample sizes, n_1 and n_2, being large.
  (ii) σ_1 and σ_2 are unknown, but X_1 and X_2 are normally distributed, and both samples are large (so we can use the fact that a t-distribution approximates to a normal distribution with large sample sizes).

1. Let [X_1, X_2 in context], (which we assume to be normally distributed)[a] and μ be the population mean.
2. | | | |
   |---|---|---|
   | Test | H_0: $\mu_1 - \mu_2 = c$ | |
   | against | H_1: | (a) $\mu_1 - \mu_2 < c$, (b) $\mu_1 - \mu_2 = c$, or (c) $\mu_1 - \mu_2 > c$, |
   | at the $100\alpha\%$ significance level. | | |
3. Under H_0, the test statistic
   (i)
   $$Z = \frac{(\overline{X}_1 - \overline{X}_2) - (\mu_1 - \mu_2)}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}} \sim \mathrm{N}(0,1).$$
   (ii)(1)
   $$Z = \frac{(\overline{X}_1 - \overline{X}_2) - (\mu_1 - \mu_2)}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}} \sim \mathrm{N}(0,1).$$
   (ii)(2)
   $$Z = \frac{(\overline{X}_1 - \overline{X}_2) - (\mu_1 - \mu_2)}{s_p\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}} \sim \mathrm{N}(0,1) \quad \text{where} \quad s_p^2 = \underline{\qquad}.$$
   Case (ii)(2) is used when the population variances coincide, i.e. $\sigma_1 = \sigma_2$.
4. Continue as per usual, calculating the critical region or the p-value.

[a] if applicable

Recall

$$s_p^2 = \frac{(n_1-1)s_1^2 + (n_2-1)s_2^2}{n_1+n_2-2}.$$

- A two-sample t-test. Again, see table 4.3 for the necessary assumptions.

1. Let [X_1, X_2 in context], *which we assume to be normally distributed*, and μ be the population mean.
2. Test $H_0\colon \mu_1 - \mu_2 = c$ against $H_1\colon$ (a) $\mu_1 - \mu_2 < c$, (b) $\mu_1 - \mu_2 = c$, or (c) $\mu_1 - \mu_2 > c$, at the $100\alpha\%$ significance level.
3. Under H_0, the test statistic

$$T = \frac{(\overline{X}_1 - \overline{X}_2) - (\mu_1 - \mu_2)}{s_p\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}} \sim t(n_1+n_2-2) \quad \text{where} \quad s_p^2 = \underline{\quad}.$$

4. Continue as per usual, calculating the critical region or the p-value.

G.C. Skills

Calculating the p-value for a

two-sample z-test: `stat` $\Longrightarrow$ `TESTS` $\Longrightarrow$ `3:2-SampZTest...`

two-sample t-test: `stat` $\Longrightarrow$ `TESTS` $\Longrightarrow$ `4:2-SampTTest...` $\Longrightarrow$ `Pooled:Yes`

- A paired sample t-test. Again, see table 4.3 for the necessary assumptions.

1. Let $D =$ [X in context] $-$ [Y in context], and μ_D be the population mean.
2. Test $H_0\colon \mu_D = \mu_0$ against $H_1\colon$ (a) $\mu_D < \mu_0$, (b) $\mu_D \neq \mu_0$, or (c) $\mu_D > \mu_0$, at the $100\alpha\%$ significance level.
3. Under H_0, the test statistic

$$T = \frac{\overline{D} - \mu_0}{s_D/\sqrt{n}} \sim t(n-1).$$

4. $d = x_1 - y_1, x_2 - y_2, \ldots, x_n - y_n$ (insert contextual values) so

$$\bar{d} = \underline{\quad} \qquad \text{and} \qquad s_d^2 = \frac{1}{n-1}\left(\sum d^2 - \frac{\left(\sum d\right)^2}{n}\right) = \underline{\quad}.$$

5. Continue as per usual, calculating the critical region or the p-value.

Note

How does the question signal the use of a paired sample t-test? It would be done in one of the following ways:

(a) Via a table

Index	1	2	$\cdots$	n
X	x_1	x_2	$\cdots$	x_n
Y	Y_1	Y_2	$\cdots$	Y_n

Table 4.1: Table containing data of two paired samples.

(b) Stated very explicitly. For instance, "The two sets of data are arranged according to respective students."

Note

Explain why a two-sample t-test would be better than a paired sample t-test.

A two-sample t-test would be better since the *samples are independent*, and we do not know if the data is organised such that each pair comes from the same column.

Note

If it were required to test whether the [population mean μ_1 of X in context] is k, give a reason, whether it would be correct to use the [pooled estimate of variance in context] or an estimate based on the [sample from the distribution of X].

It would be correct to use the estimate of variance based on [sample from the distribution of X], since the test statistic

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

involves only the [sample from the distribution of X].

Note

4.2 Summary

Throughout the two tables, we *always* assume that the (both) sample(s) independent and random. Square brackets indicate "and", while round brackets indicate "or".

Assumptions/Reasons	Test (Statistic)
[ii] The variance σ^2 is known. [ii](1) Sample size n is large (so CLT applies). [ii](2) Sample size n is small, but we assume X is normally distributed.	One-sample z-test $Z = \dfrac{\overline{X} - \mu_0}{\sigma/\sqrt{n}} \sim \mathrm{N}(0,1)$ (approximately if CLT was used)
[i] The variance σ^2 is unknown. [ii] Sample size n is large. [iii](1) X is known to be normally distributed. (FM) So $t(n-1)$ approximates to $\mathrm{N}(0,1)$. (H2 Math) No specific reason, just write "approximately.". [iii](2) X is not known to be normally distributed. (H2 Math Handwaving) CLT applies.	One-sample z-test $Z = \dfrac{\overline{X} - \mu_0}{s/\sqrt{n}} \sim \mathrm{N}(0,1)$ (approximately)
[i] The variance σ^2 is unknown. [ii] Sample size n is small. [iii] Assume X is normally distributed.	One-sample t-test $T = \dfrac{\overline{X} - \mu_0}{s/\sqrt{n}} \sim t(n-1)$

Table 4.2: Summary table for one-sample hypothesis testing.

Assumptions/Reasons	Test (Statistic)
[i] Both variances σ_1 and σ_2 are known. [ii](1) Both sample sizes n_1 and n_2 are large (so CLT applies). [ii](2) Either sample size n_1 or n_2 is small, but we assume X_1 and X_2 are normally distributed.	Two-sample z-test $Z = \dfrac{\overline{X}_1 - \overline{X}_2 - (\mu_1 - \mu_2)}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}} \sim \mathrm{N}(0,1)$ (approximately if CLT was used)
[i] One of the variances σ_1 and σ_2 are unknown. [ii] Both sample sizes n_1 and n_2 are large. [iii] Assume X_1 and X_2 are normally distributed. So $t(n_1 + n_2 - 2)$ approximates to $\mathrm{N}(0,1)$.	Two-sample z-test $Z = \dfrac{\overline{X}_1 - \overline{X}_2 - (\mu_1 - \mu_2)}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}} \sim \mathrm{N}(0,1)$ approximately
[i] Both variances σ_1^2 and σ_2^2 coincide. [ii] Assume X_1 and X_2 are normally distributed. (Alt: Both samples come from normal populations.)	Two-sample t-test $T = \dfrac{\overline{X}_1 - \overline{X}_2 - (\mu_1 - \mu_2)}{s_p\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}} \sim t(n_1{+}n_2{-}2)$
[i] Assume that $D_1, D_2, \ldots, D_n$ are normally distributed. [ii] Assume that the data within each pair $\{X_i, Y_i\}$ are dependent on each other, but pairs $\{X_i, Y_i\}$ and $\{X_j, Y_j\}$ are independent of each other, for $i \neq j$.	Paired-sample t-test $T = \dfrac{\overline{D} - \mu_D}{s_D/\sqrt{n}} \sim t(n-1).$

Table 4.3: Summary table for two-sample hypothesis testing.

Correlation and Linear Regression

Note

A good scatter diagram should follow the guidelines below.

- The relative position of each point on the scatter diagram should be clearly shown.
- The range of values for the set of data should be clearly shown by marking out the extreme x and y values on the corresponding axis.
- The axes should be labeled clearly with the variables.

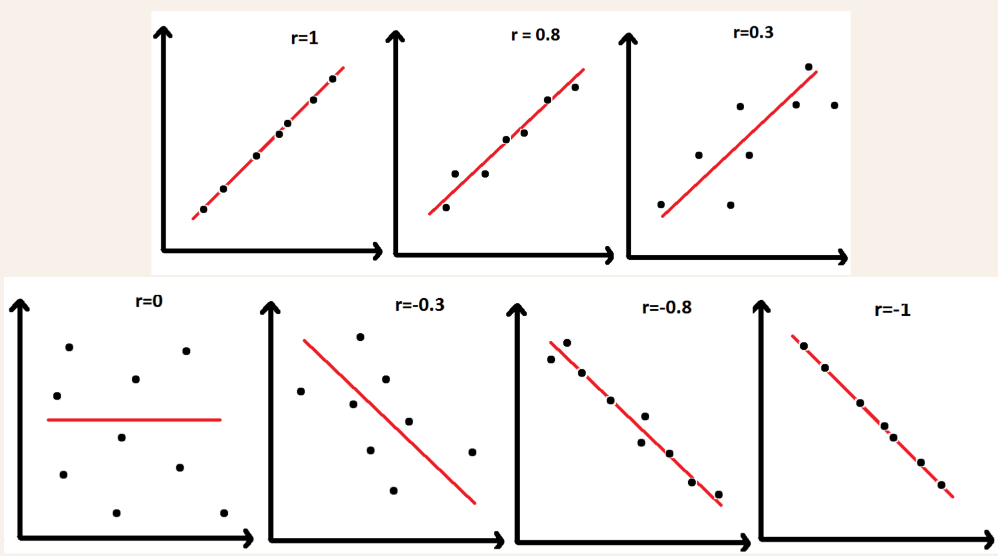
General Information

- The Product Moment Correlation Coefficient is a measure of the linear correlation between two variables. It is defined by

$$r = \frac{\sum(x-\bar{x})(y-\bar{y})}{\sqrt{\sum(x-\bar{x})^2\sum(y-\bar{y})^2}} = \frac{\sum xy - \dfrac{\sum x \sum y}{n}}{\sqrt{\left[\sum x^2 - \dfrac{\left(\sum x\right)^2}{n}\right]\left[\sum y^2 - \dfrac{\left(\sum y\right)^2}{n}\right]}},$$

  which takes on a value from 0 to 1.

- When $r = 0$, there is no linear relationship. But, a nonlinear relationship may be present. Additionally, the regression lines are perpendicular.
- The closer the value of r is to 1 (or -1), the stronger the positive (or negative) linear correlation. Furthermore, the regression lines coincide.



- The regression line of y on x minimises the sum of squares deviation (error) in the y-direction. (i.e. we are assuming x is the independent variable whose values are known exactly.) It is given by

$$y = \bar{y} + b(x-\bar{x}), \qquad \text{where} \qquad b = \frac{\sum(x-\bar{x})(y-\bar{y})}{\sum(x-\bar{x})^2} = \frac{\sum xy - \dfrac{\sum x \sum y}{n}}{\sum x^2 - \dfrac{\left(\sum x\right)^2}{n}}.$$

- The point $(\bar{x}, \bar{y})$ always lies on both the regression lines of y on x, and x on y.
- Say we are given the value of one variable, and asked to approximate the the value of the other variable. Then, we should always use the line of the *dependent* variable on the *independent*.
- Estimations should not be taken for data outside the range of the sample provided, even if the value of r is close to 1.