

# Continuous Random Variables

## General Information

- A function  $f: \mathbb{R} \rightarrow \mathbb{R}$  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  $P(a \leq X \leq b) = \int_a^b f(x) dx$ . 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)  $F: \mathbb{R} \rightarrow [0, 1]$  of a random variable  $X$  is defined by

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

- 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  $P(X \leq m) = F(m) = 1/2$ .
- Mean/Expectation:

$$\mu = E(X) := \int_{-\infty}^{\infty} x f(x) dx \quad \text{and} \quad E(g(X)) = \int_{-\infty}^{\infty} g(x) f(x) dx.$$

- Important property:

$$E(ag(X) \pm bh(x)) = a E(g(X)) \pm E(h(X)).$$

- Variance:

$$\text{Var}(X) := E(X^2) - [E(X)]^2.$$

- Important property:

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

# Special Continuous Random Variables

## Definition 2.1

A continuous random variable  $X$  has a *normal distribution* with mean  $\mu$  and standard deviation  $\sigma$ , denoted by  $X \sim N(\mu, \sigma^2)$ , 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

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

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

- Properties of the normal distribution. Let  $X$  and  $Y$  be independent, such that  $X \sim N(\mu, \sigma^2)$  and  $Y \sim N(m, s^2)$ . Then, for any  $n \in \mathbb{N}$  and  $x, y \in \mathbb{R}$ ,
  - $nX \sim N(n\mu, n^2\sigma^2)$ ,
  - $X_1 + X_2 + \dots + X_n \sim N(n\mu, n\sigma^2)$ ,
  - $aX \pm bY \sim N(a\mu \pm bm, a^2\sigma^2 + b^2s^2)$ .
- 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  $P(L_1 + 0.3L_2)$ .

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

*Note:*  $Z$  is reserved for this purpose.

- Let  $X \in N(\mu, \sigma^2)$ . Then,  $\frac{X-\mu}{\sigma}$  follows the standard normal distribution.
- What **Tail** do we select for **invNorm**?

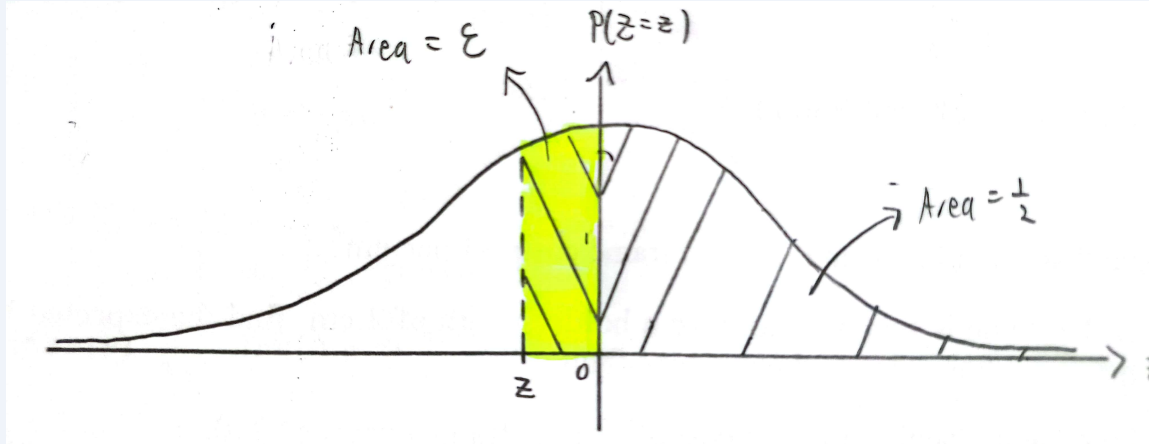
$P(X < x) = p$	LEFT
$P(-x < X < x) = p$	CENTER
$P(X > x) = p$	RIGHT

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

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

**Example 2.2**

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

**Definition 2.2**

continuous random variable  $X$  has a *uniform distribution* over the interval  $(a, b)$ , which is denoted by  $X \sim U(a, b)$ , 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**

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

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

**Note**

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

$$P(Y > z + y \mid Y > y) = P(Y > z).$$

- Expectation and variance:

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

*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  $\pm E99$ .

# 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.

## Definition 3.3

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

## General Information

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

Parameter	Estimator	Unbiased?	Formula
Population Mean $\mu$	Sample Mean $\bar{X}$	✓	$\frac{X_1 + X_2 + \dots + X_n}{n}$
Population Variance $\sigma^2$	Sample Variance $\sigma_n^2$	×	$\frac{\sum (X_i - \bar{X})^2}{n}$ $\frac{\sum X_i^2}{n} - \bar{X}^2$
	$S^2$	✓	$\frac{n}{n-1} \sigma_n^2$ $\frac{\sum (X_i - \bar{X})^2}{n-1}$ $\frac{1}{n-1} \left[ \sum X_i^2 - \frac{(\sum X_i)^2}{n} \right]$
Population Proportion $p$	Sample Proportion $P_s$	✓	$\frac{X}{n}$

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

$$\bar{X} \sim N\left(\mu, \frac{\sigma^2}{n}\right) \quad \text{and} \quad X_1 + X_2 + \dots + X_n \sim N(n\mu, n\sigma^2)$$

*approximately.*

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

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

So, if we obtain any of the latter three in solving a question, we must write “ $X \sim N(\_, \_) \text{approximately}$ ” (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  $\bar{X}_1$  and  $S_1^2$  denote the sample mean and unbiased estimator for variance, respectively, for the first sample. Similarly define  $\bar{X}_2$  and  $S_2^2$ , for the second sample.

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

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  $X_1, X_2, \dots, X_n$  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  $L = L(X_1, X_2, \dots, X_n)$  and  $U = U((X_1, X_2, \dots, X_n))$  be two statistics. We say that the interval  $(L, U)$  is a  $(1 - \alpha)100\%$  *confidence interval* for  $\theta$  iff

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

That is, the probability that the interval contains  $\theta$  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  $\theta$  and standard error  $\sigma_{\hat{\theta}}$ . We see that

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

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

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

Hence, a  $(1 - \alpha)100\%$  confidence interval for  $\theta$  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  $X_1, X_2, \dots, X_n$  be a sample on a random variable  $X$  with mean  $\mu$ , where  $n$  is large. Then, an approximate  $(1 - \alpha)100\%$  confidence interval for  $\mu$  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  $\sigma^2$  is known, we can replace  $s$  with  $\sigma$ . If the distribution of  $X$  is known to be normal, in addition to  $\sigma^2$  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  $X_1, X_2, \dots, X_n$  is a random sample from the distribution of  $X$ , where  $n$  is large. Let  $\hat{p} = \bar{X}$  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  $Y = X_1 + X_2 + \dots + X_n \sim B(n, p)$  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  $\theta$ ?

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

#### 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.

#### 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**  $\Rightarrow$  **1:Edit**  $\Rightarrow$  Key in the data into one of the lists  $L_i$ .
2. Calculating the statistic: **stat**  $\Rightarrow$  **CALC**  $\Rightarrow$  **1-Var Stats (List:L<sub>i</sub>)**  $\Rightarrow$  **Calculate**.
3. Getting the statistic for further calculations: **vars**  $\Rightarrow$  **5:Statistics**  $\Rightarrow$  Select the desired statistic.

**G.C. Skills**

Calculating the symmetric confidence interval by G.C.

Mean: `stat`  $\Rightarrow$  TESTS  $\Rightarrow$  7:ZInterval...

Proportion: `stat`  $\Rightarrow$  TESTS  $\Rightarrow$  A:1-PropZInt...

# Statistics: Hypothesis Testing

## 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

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

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

## Definition 4.2

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

## 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

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

we have

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

- A large sample hypothesis test for the mean.

Test  $H_0: \mu = \mu_0$

- against  $H_1: (a) \mu < \mu_0, \quad (b) \mu \neq \mu_0, \quad \text{or} \quad (c) \mu > \mu_0,$   
at the  $100\alpha\%$  significance level.

- Under  $H_0$ , we have  $\bar{X} \sim N(\mu_0, \hat{\sigma}^2/n)$  approximately. Or, if  $\sigma^2$  is known exactly, then by CLT  $\bar{X} \sim N(\mu_0, \sigma^2/n)$  approximately.

- Test statistic:

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



4. Find  $z_{1-\alpha}$  or  $z_{1-\alpha/2}$ , which satisfies

(a)  $P(Z < z_{1-\alpha}) = \alpha$ ,

(b)  $P(-z_{1-\alpha/2} < Z < z_{1-\alpha/2}) = \alpha$ ,  
or

(c)  $P(Z > z_{1-\alpha})$ .

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)  $z_{\text{calc}} > z_{1-\alpha}$ .

7. Write a **conclusion**.

- If we have a null hypothesis, such as

$$H_0: \mu \leq \mu_0 \quad \text{or} \quad H_0: \mu \geq \mu_0,$$

we can just use  $H_0: \mu = \mu_0$  instead.

4. Find the  $p$ -value using GC.

5. Reject  $H_0$  iff  $p$ -value is less than  $\alpha$ .

#### G.C. Skills

Calculating the  $p$ -value of a sample.

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

#### Note

Standard phrasing for rejecting  $H_0$ .

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].

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

# 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.

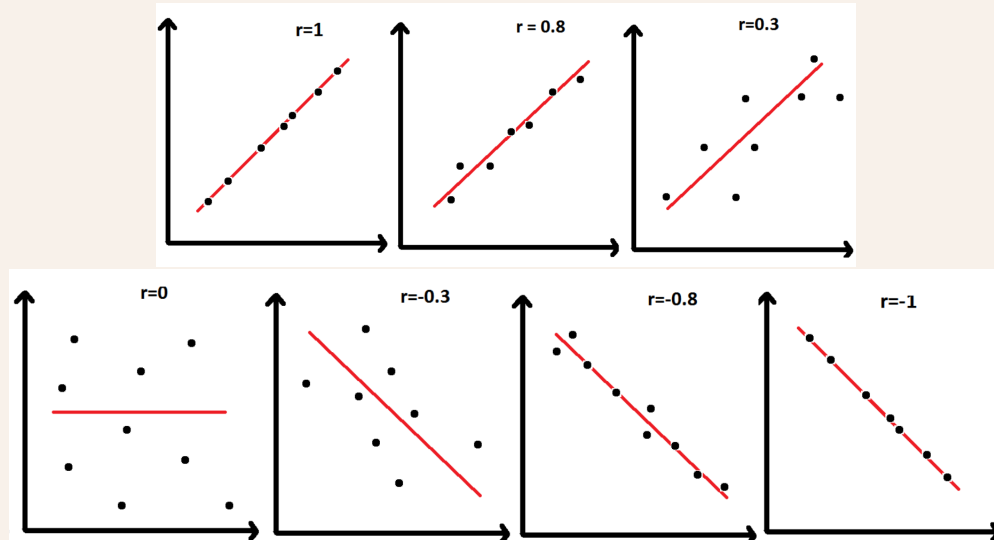
## General Information

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

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

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

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

- 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 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.