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

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

- 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:

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

- Important property:

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

- Variance:

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

- 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 $x, y \in \mathbb{R}$,
  - $nX \sim \mathrm{N}(n\mu, n^2\sigma^2)$,
  - $X_1 + X_2 + \cdots + X_n \sim \mathrm{N}(n\mu, n\sigma^2)$,
  - $aX \pm bY \sim \mathrm{N}(a\mu \pm bm, a^2\sigma^2 + b^2 s^2)$.

- Question phrasing may be misleading at times. Try to use some inference as to what exactly does the setter mean.

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.

- 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

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

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

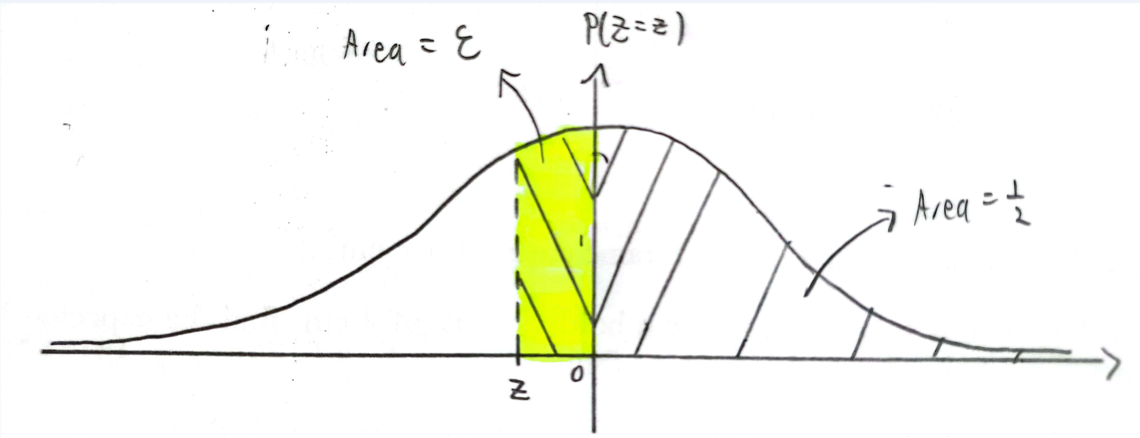
$\mathrm{P}(X < x) = p$	`LEFT`
$\mathrm{P}(-x < X < x) = p$	`CENTER`
$\mathrm{P}(X > x) = p$	`RIGHT`

- 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:



- 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

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

- 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.

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 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 $\hat{\mu}$ 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,

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

  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. Let the unbiased estimates for the population mean and variance of this sample be $\bar{x}$ and s^2, respectively. 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 you should know. 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. Similarly define $\overline{X}_2$ and S_2^2.

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

- Let $0 < \alpha < 1$

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 phrasing used by 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 [μ 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.

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 by G.C.

`stat` $\Longrightarrow$ `TESTS` $\Longrightarrow$ `7:ZInterval...`

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.

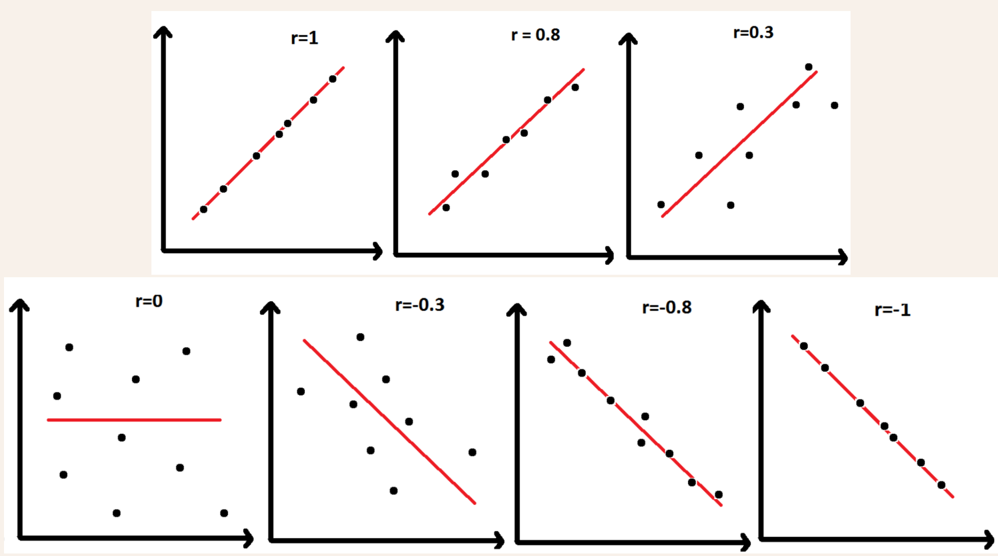
General Information

- 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.