The University of Melbourne

Semester 1 Examinations — June, 2018

School of Mathematics and Statistics

MAST90105 Methods of Mathematical Statistics Sample Final Exam

Exam Duration: 3 Hours

Reading Time: 15 Minutes

This paper has 10 pages

Authorised materials:
Hand-held electronic calculators (including graphics calculators) may be used.
Students may bring ONE double-sided A4 sheet of notes into the exam room.
A formula sheet is attached at the end of this paper.

Instructions to Invigilators:
A 16-page script book shall be supplied to each student.
Students should NOT include the examination paper with their script books.
Students may take this paper with them at the end of the exam.

Instructions to Students:
This paper has **9** questions.
Attempt as many questions, or parts of questions, as you can.
Questions carry marks as shown in the brackets after the question statement.
The total number of marks available for this examination is **100**.
Working and/or reasoning must be given to obtain full credit.
Answers may be given as fractions or decimals (with a specified accuracy).

This paper may be reproduced and lodged at the Baillieu Library.

1. The volume, V, in cubic ft. and height, H, in ft. of a random Black Cherry tree may be modelled as a bivariate normal distribution with correlation 0.6 and means and standard deviations given in the following table:

R.V.	Mean	Standard deviation
H	76	6
V	30	16

State the conditional distributions of V for a tree of

(a) average height,

(b) one standard deviation above the mean,

(c) one standard deviation below the mean.

[6]

2. A study measured the Girth (in inches) at 4 feet 6 inches above the ground, Height (in feet) and Volume of timber (in cubic feet) in 31 felled black cherry trees. The data is recorded in the R data set "trees". The R commands and output in the questions below aim to look at the relationship of Volume to Height, as well as the relationship of Volume to a new variable CylinderVol, defined as Girth^{2}Volume, and representing a constant times the volume of the tree were it a cylinder with radius proportional to the Girth.

(a) What do the following R commands and output, including Figure 1, reveal about the relationship between Volume and Height. In particular, what is the estimate of Volume for a tree of Height 10 feet? Is this estimate plausible? Comment. [5]

```
> attach(trees)
> FitHeight <- lm(Volume ~ Height)
> plot(Height, Volume)
> abline(FitHeight) # Figure 1
> summary(FitHeight)

Call:
lm(formula = Volume ~ Height)

Residuals:
    Min      1Q    Median    3Q     Max
-21.274  -9.894  -2.894  12.068  29.852

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) -87.1236    29.2731  -2.976 0.005835 **
Height        1.5433     0.3839   4.021 0.000378 ***
---
```

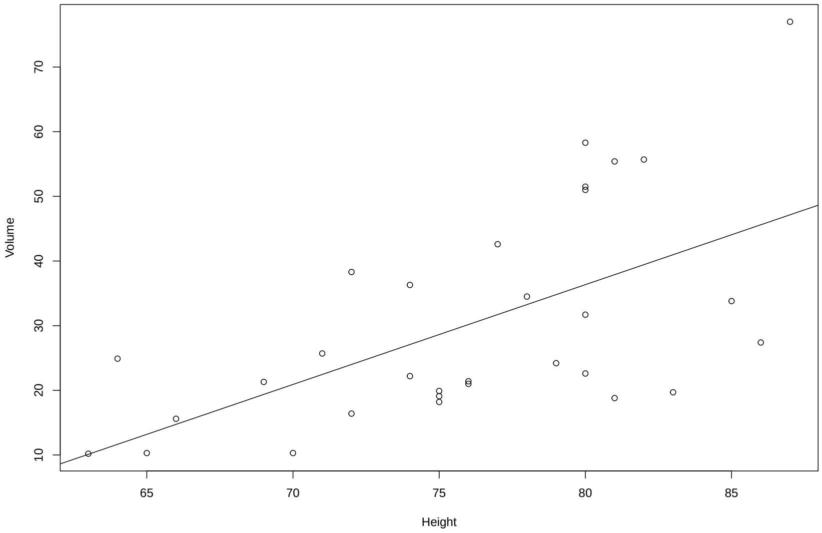

Figure 1: See Question 2a

(b) What do the following R commands and output, including Figure 2, reveal about the relationship between Volume and Height. In particular, what is the estimate of Volume for a tree of Cylinder Volume 50 cubic units? Is this estimate plausible? Which model is preferable? [5]

```
> CylinderVol <- Girth^2*Height
> FitCylinder <- lm(Volume ~ 0 + CylinderVol)
> plot(CylinderVol, Volume)
> abline(FitCylinder) # Figure 2
> summary(FitCylinder)

Call:
lm(formula = Volume ~ 0 + CylinderVol)

Residuals:
    Min      1Q  Median      3Q     Max
-4.6696 -1.0832 -0.3341  1.6045  4.2944

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
CylinderVol 2.108e-03  2.722e-05   77.44   <2e-16 ***
---
```

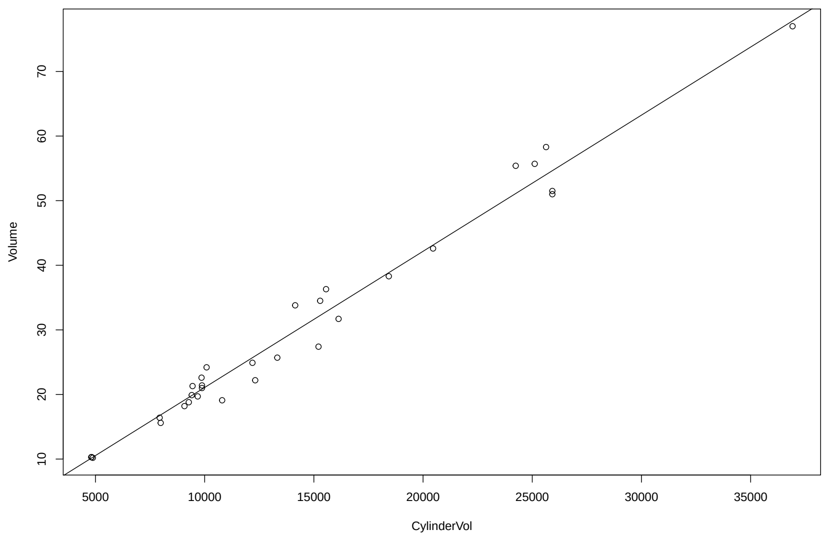

Figure 2: See Question 2b

(c) Explain the plots in the following R commands and what their output demonstrates. [3]

```
> qqnorm(FitHeight$residuals)
> qqline(FitHeight$residuals)  # Figure 3
> qqnorm(FitCylinder$residuals)
> qqline(FitCylinder$residuals)  # Figure 4
```

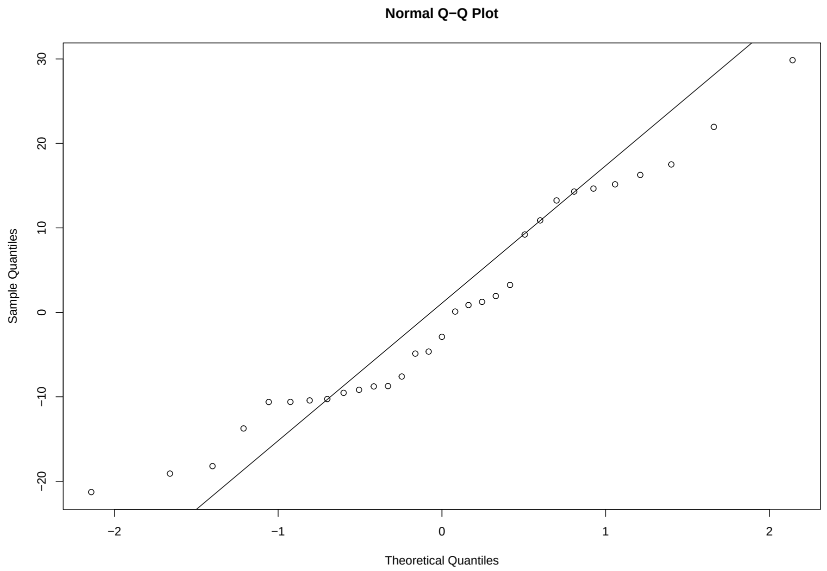

Figure 3: See Question 2c

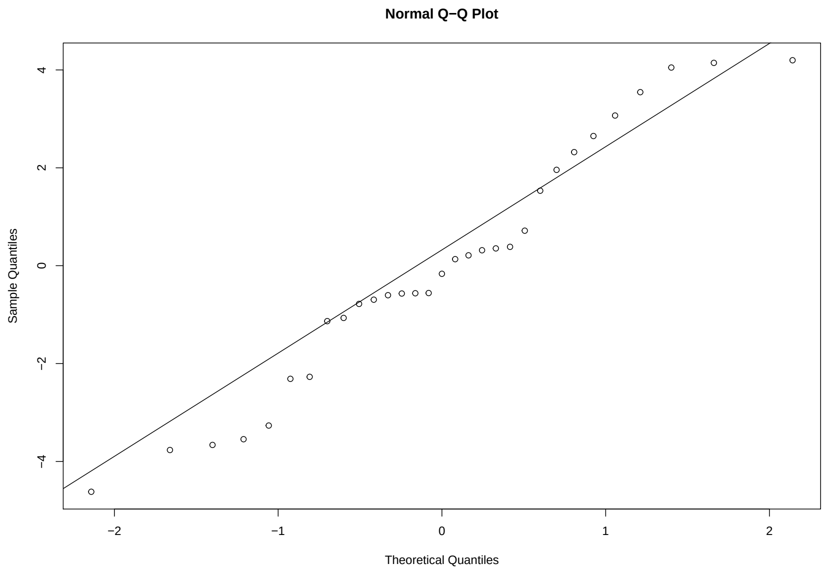

Figure 4: See Question 2c

3. Let $X_1, \ldots, X_n$ be a random sample from the density:

$$f(x;\theta) = \frac{x}{\theta^2} e^{-x/\theta}, \quad 0 < x < \infty, \quad 0 < \theta < \infty.$$

   (a) Write the log-likelihood function and the score function. [2]

   (b) Determine the maximum likelihood estimator of θ, checking that your estimator is a maximum for the likelihood function. [3]

   (c) Find the Fisher Information. [2]

   (d) Give the Cramér-Rao lower bound of the variance of unbiased estimators of θ. [2]

   (e) Is the maximum likelihood estimator unbiased? What is its variance? Does it achieve the the Cramér-Rao lower bound [4]

   (f) A random sample of size $n = 35$ gave $\overline{x} = 10.5$. Determine the maximum likelihood estimate of θ and an approximate 95% confidence interval for θ. Some R output that may help. [2]

```
>  z <- c(0.95,0.975,0.99,0.995)
> qnorm(z)
[1] 1.644854 1.959964 2.326348 2.575829
```

4. Let $X_1, \ldots, X_n$ be a random sample from the a Gamma(α, θ) distribution with pdf $(x^{\alpha-1}e^{-x/\theta})/(\theta^\alpha\Gamma(\alpha))$, with $x, \alpha, \theta > 0$.

   (a) Derive an estimator for α and θ using the method of moments. [3]

   (b) Find points estimates of α and β from for the following observations of X: [2]

```
3.51 3.27 4.90 5.27 4.25
```

5. Let $Y = \sum_{i=1}^n X_i$ be the sum of the observations of a random sample $X_1, \dots, X_n$, where X_i follows a Poisson distribution with mean θ:

$$f(x|\theta) = \frac{e^{-\theta}\theta^x}{x!}, \quad x \in \{0, 1, \dots\}, \quad \theta > 0.$$

Suppose the prior distribution of θ has a gamma density:

$$f(\theta|\alpha, \beta) = \frac{1}{\Gamma(\alpha)\beta^\alpha}\theta^{\alpha-1}e^{-\theta/\beta}, \quad 0 < \theta < \infty,$$

.

(a) Determine the posterior distribution of θ, given that $Y = y$. [4]

(b) Show that point estimate $\hat{\theta}$ under the quadratic loss function is a weighted average of the maximum likelihood estimate y/n and the prior mean $\alpha\beta$ with weights $n\beta/(n\beta + 1)$ and $1/(n\beta + 1)$, respectively. [2]

(c) From past experience, a lecturer believes that the weekly number of emails he receives from students in statistics is a random variable with a Poisson distribution with parameter following a gamma distribution with $\alpha = 10$ and $\beta = 2$.

i. What would his estimate of the mean number of weekly emails be if he only considered the prior information? [1]

ii. Suppose in four weeks he receives $y = 75$ emails. What would his estimate of the mean number of weekly emails be if he considered both the prior information and the observed data? [1]

6. Let $X_1, \dots, X_n$ and $Y_1, \dots, Y_m$ be independent sequences of independent normally distributed random variables. Suppose that $E(X_i) = \mu$, $i = 1, \dots, n$ and $E(Y_j) = \mu + \theta$, $j = 1, \dots, m$, and $\text{Var}(X_i) = \text{Var}(Y_j) = \sigma^2$.

(a) Show that the roots of the score function equations for μ and θ respectively are

$$\hat{\mu} = \bar{X} \text{ and } \hat{\theta} = \bar{Y} - \bar{X}$$

[3]

(b) Show that these estimators are unbiased. [1]

(c) Show that

$$\text{Var}(\hat{\mu}) = \frac{\sigma^2}{n} \text{ and } \text{Var}(\hat{\theta}) = \frac{\sigma^2}{m} + \frac{\sigma^2}{n}$$

[1]

(d) Show that

$$\sum_{i=1}^n (X_i - \mu)^2 = \sum_{i=1}^n (X_i - \hat{\mu})^2 + n(\hat{\mu} - \mu)^2$$

and

$$\sum_{j=1}^m (Y_j - \mu - \theta)^2 = \sum_{j=1}^m (Y_j - \hat{\mu} - \hat{\theta})^2 + m(\hat{\mu} + \hat{\theta} - \mu - \theta)^2$$

[3]

(e) Hence show that

i. $\hat{\mu}$ and $\hat{\theta}$ are the maximum likelihood estimators of μ and θ [3]

ii.
$$E\left[\sum_{i=1}^{n}(X_i - \bar{X})^2 + \sum_{j=1}^{m}(Y_j - \bar{Y})^2\right] = \sigma^2(m+n-2)$$
[4]

and deduce an unbiased estimator $\hat{\sigma}^2$ of σ^2 based on the sample variances of the two samples [3].

(f) Given that the unbiased estimator $\hat{\sigma}^2$ is independent of the pair of sample means, find the distribution of
$$T = \frac{\hat{\theta} - \theta}{\hat{\sigma}\sqrt{\frac{1}{n} + \frac{1}{m}}}$$
clearly indicating results that you use. [3]

(g) An eating attitude test (EAT) was administred to both a sample of models and a control group, resulting in the following summary statistics.

	Sample size	Sample mean	Sample standard deviation
Models	30	8.63	4.501
Controls	30	10.97	4.832

You may assume the EAT scores of both samples are normally distributed.

i. Is there sufficient evidence to justify claiming that a difference exists in the mean EAT score between models and controls? Assume that the two populations have equal variance and use a test with significance level $\alpha = 0.05$ and clearly state your null and alternative hypotheses [3]

ii. Give an approximate p-value for the test in (i). [2]

The following R output may be useful.

```
> t=c(0.005,0.01,0.025,0.05, 0.950, 0.975, 0.990, 0.995)
> qt(t, 58)
-2.66 -2.39 -2.00 -1.67  1.67  2.00  2.39  2.66
> qnorm(t)
-2.58 -2.33 -1.96 -1.64  1.64  1.96  2.33  2.58
> qt(t, 29)
-2.76 -2.46 -2.05 -1.70  1.70  2.05  2.46  2.76
```

7. Suppose that $X_1, \ldots, X_n$ are a random sample from an Exp(θ) distribution with pdf:
$$f(x;\lambda) = \lambda e^{-\lambda x}, x > 0.$$

(a) Show that the random variable $W = \min_{i=1,\ldots,n} X_i$ has an Exp($n\lambda$) distribution. [3]

(b) Determine the distribution of $n\lambda W$ [1]

(c) Using the statistic W, construct a $100(1-\alpha)\%$ confidence interval for λ. [4]

(d) Let X be the times between successive emissions of alpha particles from a source. Consider the following sample of $n = 5$ observations on X: 1.440.560.161.600.35. Find the endpoints for a 95% confidence interval for the reciprocal of the mean $\lambda = 1/E[X]$. [1]

Some R output that may help.

```
> t=c(0.01,0.025,0.05,0.1, 0.90, 0.95, 0.975, 0.99)
> -log(1-t)
[1] 0.01005034 0.02531781 0.05129329 0.10536052 2.30258509
[6] 2.99573227 3.68887945 4.60517019
```

8. Researchers were interested in whether chocolate milk might be a better recovery drink for athletes. So they took 16 cyclists and had them do a strenuous interval workout. Then they were allowed a four-hour recovery period where eight were randomly assigned to drink chocolate milk and the others were randomly assigned to drink a carb-replacement drink. After the recovery period, they were asked to bike until exhaustion, and their time to exhaustion was measured.

Chocolate milk (X):	49.1	51.0	50.9	53.6	50.6	49.8	50.0	19.8
Carb-replacement (Y):	59.0	20.3	19.8	22.2	20.8	19.6	20.0	22.9

(a) Use the sign test to determine if there is evidence that the time to exhaustion after the recovery period was longer for chocolate milk drinkers. Use $\alpha = 0.05$ and clearly state your hypotheses. [5]

Some R output that may be useful is

```
> pbinom(0:8, 8 ,.5)
0.00 0.04 0.14 0.36 0.64 0.86 0.96 1.00 1.00
> dbinom(0:8, 8 ,.5)
0.00 0.03 0.11 0.22 0.27 0.22 0.11 0.03 0.00
```

(b) Take $\alpha = 0.05$ and use the statistic W – defined by the sum of the ranks of the observations of Y (carb-replacement drinkers) in the combined sample – to conduct the test. Recall that when $Z \sim N(0, 1)$, we have $P(Z \leq 1.645) \approx 0.95$, $P(Z \leq 1.96) \approx 0.975$. You may use the fact that $E(W) = \frac{n_2(n_2+n_1+1)}{2}$ and $Var(W) = \frac{n_2 n_1(n_2+n_1+1)}{12}$ if W is the sum of randomly chosen ranks for a group of size n_2 chosen amongst the ranks of a combined group of size $n_1 + n_2$. What is the reason for any difference between the results of the two tests? [5]

9. A researcher has developed a theoretical model for the number of DNA mutations in humans, say X, occurring in the DNA region R. She predicted that X follows a Poisson distribution with mean $E(X) = 2$ (the corresponding pmf is $2^x e^{-2}/x!$, $x = 0, 1, \ldots$). To test her claim, she collected observations based on 100 healthy subjects and obtained the following .

Number of mutations:	0	1	2	3	4	≥ 5
Observed frequency:	10	30	28	15	9	8

(a) Clearly state appropriate hypotheses to test whether the researcher's prediction is correct. [2]

(b) Clearly state the test statistic and its approximate distribution under the null hypothesis [3]

(c) Find the observed value of the test statistic [1]

(d) Draw a conclusion for these data [2]

Some R output that may be useful is

```
> dpois(0:8, 2)
0.14 0.27 0.27 0.18 0.09 0.04 0.01 0.00 0.00
> ppois(0:8, 2)
0.14 0.41 0.68 0.86 0.95 0.98 1.00 1.00 1.00
> qchisq(c(0.90, 0.95, 0.975), 5)
[1]  9.236357 11.070498 12.832502
> qchisq(c(0.90, 0.95, 0.975), 4)
[1]  7.779440  9.487729 11.143287
```

Total marks = 100

End of the exam questions.
Formulas are on the next page.

Table XII: Discrete Distributions

Probability Distribution and Parameter Values	Probability Mass Function	Moment-Generating Function	Mean $E(X)$	Variance $\mathrm{Var}(X)$	Examples
Bernoulli $0 < p < 1$ $q = 1 - p$	$p^x q^{1-x}, \quad x = 0, 1$	$q + pe^t$	p	pq	Experiment with two possible outcomes, say success and failure, $p = P(\text{success})$
Binomial $n = 1, 2, 3, \ldots$ $0 < p < 1$	$\binom{n}{x} p^x q^{n-x}$, $x = 0, 1, \ldots, n$	$(q + pe^t)^n$	np	npq	Number of successes in a sequence of n Bernoulli trials, $p = P(\text{success})$
Geometric $0 < p < 1$ $q = 1 - p$	$q^{x-1}p$, $x = 1, 2, \ldots$	$\dfrac{pe^t}{1 - qe^t}$	$\dfrac{1}{p}$	$\dfrac{q}{p^2}$	The number of trials to obtain the first success in a sequence of Bernoulli trials
Hypergeometric $x \le n, x \le N_1$ $n - x \le N_2$ $N = N_1 + N_2$ $N_1 > 0, \quad N_2 > 0$	$\dfrac{\binom{N_1}{x}\binom{N_2}{n-x}}{\binom{N}{n}}$		$n\left(\dfrac{N_1}{N}\right)$	$n\left(\dfrac{N_1}{N}\right)\left(\dfrac{N_2}{N}\right)\left(\dfrac{N-n}{N-1}\right)$	Selecting r objects at random without replacement from a set composed of two types of objects
Negative Binomial $r = 1, 2, 3, \ldots$ $0 < p < 1$	$\binom{x-1}{r-1} p^r q^{x-r}$, $x = r, r + 1, \ldots$	$\dfrac{(pe^t)^r}{(1 - qe^t)^r}$	$\dfrac{r}{p}$	$\dfrac{rq}{p^2}$	The number of trials to obtain the rth success in a sequence of Bernoulli trials
Poisson $0 < \lambda$	$\dfrac{\lambda^x e^{-\lambda}}{x!}$, $x = 0, 1, \ldots$	$e^{\lambda(e^t - 1)}$	λ	λ	Number of events occurring in a unit interval, events are occurring randomly at a mean rate of λ per unit interval
Uniform $m > 0$	$\dfrac{1}{m}, \quad x = 1, 2, \ldots, m$		$\dfrac{m+1}{2}$	$\dfrac{m^2 - 1}{12}$	Select an integer randomly from $1, 2, \ldots, m$

Table XIII: Continuous Distributions

Probability Distribution and Parameter Values	Probability Density Function	Moment-Generating Function	Mean $E(X)$	Variance $\mathrm{Var}(X)$	Examples
Beta $0 < \alpha$ $0 < \beta$	$\dfrac{\Gamma(\alpha + \beta)}{\Gamma(\alpha)\Gamma(\beta)} x^{\alpha-1}(1 - x)^{\beta-1}$, $0 < x < 1$		$\dfrac{\alpha}{\alpha + \beta}$	$\dfrac{\alpha\beta}{(\alpha + \beta + 1)(\alpha + \beta)^2}$	$X = X_1/(X_1 + X_2)$, where X_1 and X_2 have independent gamma distributions with same θ
Chi-square $r = 1, 2, \ldots$	$\dfrac{x^{r/2-1}e^{-x/2}}{\Gamma(r/2)2^{r/2}}$, $0 < x < \infty$	$\dfrac{1}{(1 - 2t)^{r/2}}, \quad t < \dfrac{1}{2}$	r	$2r$	Gamma distribution, $\theta = 2$, $\alpha = r/2$; sum of squares of r independent $N(0, 1)$ random variables
Exponential $0 < \theta$	$\dfrac{1}{\theta} e^{-x/\theta}, \quad 0 \le x < \infty$	$\dfrac{1}{1 - \theta t}, \quad t < \dfrac{1}{\theta}$	θ	θ^2	Waiting time to first arrival when observing a Poisson process with a mean rate of arrivals equal to $\lambda = 1/\theta$
Gamma $0 < \alpha$ $0 < \theta$	$\dfrac{x^{\alpha-1}e^{-x/\theta}}{\Gamma(\alpha)\theta^\alpha}$, $0 < x < \infty$	$\dfrac{1}{(1 - \theta t)^\alpha}, \quad t < \dfrac{1}{\theta}$	$\alpha\theta$	$\alpha\theta^2$	Waiting time to αth arrival when observing a Poisson process with a mean rate of arrivals equal to $\lambda = 1/\theta$
Normal $-\infty < \mu < \infty$ $0 < \sigma$	$\dfrac{e^{-(x-\mu)^2/2\sigma^2}}{\sigma\sqrt{2\pi}}$, $-\infty < x < \infty$	$e^{\mu t + \sigma^2 t^2/2}$	μ	σ^2	Errors in measurements; heights of children; breaking strengths
Uniform $-\infty < a < b < \infty$	$\dfrac{1}{b - a}, \quad a \le x \le b$	$\dfrac{e^{tb} - e^{ta}}{t(b - a)}, \quad t \ne 0$ $1, \quad t = 0$	$\dfrac{a + b}{2}$	$\dfrac{(b - a)^2}{12}$	Select a point at random from the interval $[a, b]$