# Statistics formulae booklet

6 February 2015

## 1 Quantitative variables

Where $x_n$ is the indexed element in the data set, $n$ is the number of elements themselves, $\mu$ is the mean, and $\sigma$ is the standard deviation. Note that variance $\left(\sigma^2\right)$ is equal to standard deviation squared.

| | | |
|---|---|---|
| Mean | $\bar{x} = \frac{\sum_{i=0}^{n} x_i}{n}$ | the sum of all elements divided by the number of elements. Note that for populations, $\bar{x}$ should be $\mu$ |
| Sample std dev | $s = \sqrt{\frac{\sum_{i=1}^{n} (x_i - \bar{x})^2}{n-1}}$ | the square root of the sum of the squares of the difference of all elements and the mean divided by $n-1$ |
| Population std dev | $\sigma = \sqrt{\frac{\sum_{i=1}^{n} (x_i - \mu)^2}{n}}$ | the square root of the sum of the squares of the difference of all elements and the mean divided by $n$ |

## 2 Normal model

The normal model has two parameters, $\sigma$ and $\mu$. They are standard deviation and mean, respectively. Be aware that the normal model has no endpoints.

| | | |
|---|---|---|
| Standardised normal variable (z-score) | $z = \frac{x - \mu}{\sigma}$ | the difference between the element and the mean, divided by the standard deviation |

## 3 Linear regression

Remember that for all linear models, the model predicts that such a data value will occur. Values of $r$, if not given, should be found using the calculator or by reversing the slope formula. Remember that the negativity of $r$ matters in determining the slope of the line.

| | | |
|---|---|---|
| Equation of a line | $\hat{y} = b_0 + b_1 x$ | $b_0$ is the intercept, and $b_1$ is the slope of the line |
| Intercept ($b_0$) | $b_0 = \bar{y} - b_1 \bar{x}$ | Algebraic manipulation of the equation of a line |
| Slope ($b_1$) | $b_1 = r\frac{s_y}{s_x}$ | Where $s_y$ and $s_x$ are the standard deviations of $y$ and $x$, respectively |

# 4 Probability

| | | |
|---|---|---|
| General addition | $P(A \cup B) = P(A) + P(B) - P(A \cap B)$ | the probability of A or B occurring is the sum of the two probabilities without the probability that both occur |
| General multiplication | $P(A \cap B) = P(A\|B) \cdot P(B)$ | the probability that both occur is the probability of the second multiplied by the probability of the first given the second |

# 5 Random variables

## 5.1 Expected values and measures of spread

Where $p$ is the probability of success, $n$ is the number of trials, and if you find an instance of $q$, remember that $q = 1 - p$. The following is a table for expected values and variance.

| | | |
|---|---|---|
| General expected value | $E(X) = \sum_{i=1}^{n} x_i P_i$ | sum of all relevant frequencies multiplied by all relevant outputs starting from 1 to the number of relevant data values |
| Geometric $E(X)$ | $E(X) = p^{-1}$ | the reciprocal of the probability of occurrence |
| Binomial $E(X)$ | $E(X) = np$ | $E(X)$ is the probability multiplied by the number of trials |
| General variance | $\mathrm{Var}(X) = \sum_{i=1}^{n} (x_i - \mu_i)^2 p_i$ | the sum of all the elements' differences from the mean squared which is then multiplied by the probability of that element |
| Pythagorean theorem of statistics | $\mathrm{SD}(X)^2 + \mathrm{SD}(Y)^2 = \mathrm{SD}(X+Y)^2$ | used when adding or subtracting two different random variables, here, $X$ and $Y$. It can also be expressed as $\mathrm{Var}(X) + \mathrm{Var}(Y) = \mathrm{Var}(X+Y)$, using variance instead of standard deviation |
| Geometric $\sigma$ | $\sigma = \sqrt{(1-p)\,p^{-2}}$ | the square root of the quantity $q$ divided by $p^2$ |
| Binomial $\sigma$ | $\sigma = \sqrt{np(1-p)}$ | the square root of $p$ multiplied by $q$ and $n$ |

## 5.2 Probabilities

For determining the probabilities of these random variables, determine whether it is geometric or binomial. The formulae for those is provided here. It is good to remember that as the number of trials approaches infinity, it becomes closer to the expected value.

| | | |
|---|---|---|
| Binomial $P(X = x)$ | $P(X = x) = \binom{n}{x} p^x (1-p)^{n-x}$ | this states that the probability of getting a parameter number of $X$ of $n$ trials |
| Geometric $P(X = x)$ | $P(X = x) = (1-p)^{n-1} p$ | this states the probability of not getting a success for $n-1$ times until getting an $n$ |

# 6 Sample distributions

This deals with the Central limit theorem and the sample of the population (the sample proportion, listed as $\hat{p}$). The mean of multiple samples, the foundation of the Central limit theorem, is shown as $\bar{x}$ and its standard deviation as $\sigma_{\bar{x}}$

| | | |
|---|---|---|
| SD($\hat{p}$) | $\sigma_{\hat{p}} = \sqrt{\frac{p(p-1)}{n}}$ | This is for determining the standard deviation of sample distributions |
| SD($\bar{x}$) | $\sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}}$ | This is for determining the standard deviation of multiple sample means |
| Standard error | $SE(\hat{p}) = \sqrt{\frac{\hat{p}(1-\hat{p})}{n}}$ | Standard error is an estimation of standard deviation when we do not know $p$ but do know $\hat{p}$ |