

## Formulas and Terms in Math 102 / Core 143

In a histogram,

# of scores (or total probability of outcomes) in an interval  
= area under the histogram above that interval

For a list of  $n$  scores ( $x$ -values):

$$\text{Average } \bar{x} = \mu = \frac{\sum x}{n}, \quad \text{Standard deviation } \sigma = SD = \sqrt{\frac{\sum (x - \bar{x})^2}{n}}$$

$N$ -th percentile =  $x$ -value for which  $N\%$  of scores are  $\leq x$

Median = 50-th percentile,  $IQR = 75\text{-th} - 25\text{-th}$  percentile

Converting a score  $x$  to standard units ( $z$ -value or  $t$ -value):

$$z \text{ or } t = \frac{x - \bar{x}}{\sigma} \quad \text{or} \quad \frac{x - EV}{SE} \quad (\text{see below})$$

Correlation:  $r =$  average of products ( $z_x z_y$ ), where  $z_q$  means  $q$  in standard units

Regression line of  $y$  on  $x$  (for predicting  $y$  from  $x$ , or for estimating average  $y$  within a vertical strip at  $x$ ): Denote predicted  $y$ -value by  $\hat{y}$ . Then

$$\hat{y} - \bar{y} = r \left( \frac{\sigma_y}{\sigma_x} \right) (x - \bar{x})$$

Residual corresponding to data point  $(x, y)$ :  $y - \hat{y}$  or  $(x, y - \hat{y})$ .

RMS error for regression of  $y$  on  $x$  (= approximate standard deviation of data in any vertical strip, if scatter diagram is homoscedastic):

$$\sigma_y \sqrt{1 - r^2}$$

Multiplication rule:  $\text{Prob}(A \text{ and } B) = \text{Prob}(A) \cdot \text{Prob}(B \text{ given } A)$

Independent events:  $\text{Prob}(B \text{ given } A) = \text{Prob}(B)$

Addition rule:  $\text{Prob}(A \text{ or } B) = \text{Prob}(A) + \text{Prob}(B) - \text{Prob}(A \text{ and } B)$

Mutually exclusive events:  $\text{Prob}(A \text{ and } B) = 0$

Binomial probabilities: If an event has probability  $p$  on each trial, the probability of its occurring exactly  $k$  times in  $n$  independent trials:

$$C(n, k) \cdot p^k \cdot (1 - p)^{n-k}, \quad \text{where} \quad C(n, k) = \frac{n!}{k!(n - k)!}$$

For a sample of size  $n$  from a population with average  $\mu$  and standard deviation  $\sigma$ :

$EV$  of sum of scores in sample =  $n\mu$

$SE$  of sum =  $\sigma \cdot \sqrt{n}$

$EV$  of average of sample =  $\mu$

$SE$  of average =  $\sigma/\sqrt{n}$

For significance tests, approximate (bootstrap) population standard deviation  $\sigma$  with sample standard deviation  $s = SD^+ = \sqrt{\frac{\sum(x - \bar{x})^2}{n - 1}} = (SD \text{ of sample})\sqrt{\frac{n}{n-1}}$ . (The null hypothesis will give a value to use for  $\mu$ .) For large samples ( $n \geq 30$ ),  $s$  is close to  $\sigma$ .

For confidence intervals, also approximate population average  $\mu$  with sample average  $\bar{x}$ .

Special case: Population is 0's and 1's (or yeses and nos, or ins and outs, or ...), fraction of 1's is  $p$ , for a sample of size  $n$ :

$EV$  of count =  $np$

$SE$  of count =  $\sqrt{p(1-p)} \cdot \sqrt{n}$

$EV$  of % (or proportion) =  $p$

$SE$  of % (or proportion) =  $\sqrt{p(1-p)}/\sqrt{n}$

For CIs, approximate (bootstrap) population proportion  $p$  with sample proportion  $\hat{p}$ .

For use with confidence interval or  $t$ -test for significance on small ( $n < 30$ ) samples: degrees of freedom =  $n - 1$

$k\%$  confidence interval for the average of a population:

Let  $z_k$  denote the  $z$ -value for which  $k$  percent of the data is between  $-z_k$  and  $z_k$ . Then the CI is

$$\bar{x} \pm z_k \cdot (SE \text{ for average})$$

(Similar for "proportion" in place of "average".)

For  $n < 30$ : Let  $t_{(100-k)\%/2}$  denote the  $t$ -value for which  $(100-k)\%/2$  of the probability in the  $t$ -distribution is to its right. Then the CI is

$$\bar{x} \pm t_{(100-k)\%/2} \cdot (SE \text{ for average})$$

For CI or significance test for difference of  $\mu$ 's in two populations:

$SE$  for difference of averages of 2 samples =  $\sqrt{(SE \text{ of first})^2 + (SE \text{ of second})^2}$

For significance test,  $EV$  of difference = 0 by  $H_0$ . (For more than 2 samples, use one-way ANOVA.)

[ Technicality, mostly to be ignored: To get SE for difference of proportions in two populations:

For CI, use  $\sqrt{\frac{\hat{p}_1(1-\hat{p}_1)}{n_1} + \frac{\hat{p}_2(1-\hat{p}_2)}{n_2}}$ , where  $\hat{p}_i$ 's are proportions in samples, as above

For significance test: Pooled estimate for common population proportion is  $\hat{p} = \frac{n_1 p_1 + n_2 p_2}{n_1 + n_2}$ ,

and  $SE$  for difference =  $\sqrt{\hat{p}(1-\hat{p}) \left( \frac{1}{n_1} + \frac{1}{n_2} \right)}$  ]

For deciding significance of differences in frequency distributions among categories:

$\chi^2 = \sum[(\text{observed} - \text{expected})^2/\text{expected}]$

degrees of freedom: in "list" distributions, # in list - 1;

in "table" distributions, (# of rows - 1) · (# of columns - 1)