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 scores:
  Average (AV) = (sum of scores in the list)/(# of scores in the list)
  $SD = \sqrt{\text{average of the new list (score} - AV)^2}$
  N-th percentile = score x for which $N\%$ of scores are $\leq x$
  Median = 50-th percentile, IQR = 75-th $-$ 25-th percentile

Converting a score to standard units (z-value):

$$z = (\text{score} - AV)/SD \qquad \text{or} \qquad (\text{score} - EV)/SE$$

Correlation coefficient:

$$r = \text{average of } ((x \text{ in standard units})\cdot(y \text{ in standard units}))$$
$$= \frac{(\text{average of } x \cdot y) - (\text{average of } x) \cdot (\text{average of } y)}{(SD \text{ of } x) \cdot (SD \text{ of } y)}$$

Regression line of y on x (for predicting y from x, or for estimating average y within a vertical strip at x):

$$y - (AV \text{ of } y) = r\left(\frac{SD \text{ of } y}{SD \text{ of } x}\right)(x - (AV \text{ of } x))$$

R.M.S. error for regression line of y on x (= SD of data in any vertical strip, if scatter diagram is homoscedastic):

$$\sqrt{1 - r^2} \cdot (SD \text{ of } y)$$

Multiplication rule: Prob(A and B) = Prob(A)·Prob(B given A)

Independent events: Prob(B given A) = Prob(B)

Mutually exclusive events: Prob(A and B) = 0

Addition rule for mutually exclusive events: Prob(A or B) = Prob(A) + Prob(B)
(If not mutually exclusive, subtract Prob(A and B) from right side.)

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

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

For a sample of n draws <u>with</u> replacement from a box of tickets:
  EV of sum of draws in sample = $n \cdot (AV \text{ of box})$
  SE of sum = $\sqrt{n} \cdot (SD \text{ of box})$
  EV of average draw in sample = AV of box
  SE of average = $(SD \text{ of box})/\sqrt{n}$

Special case: Box of 0's and 1's (for counting numbers of 1's and finding percentage of 1's in n draws:

EV of count $= n\cdot$(fraction of tickets with 1's)
SE of count $= \sqrt{n} \cdot (SD$ of box) (see shortcut formula below)
EV of % = fraction with 1's
SE of % $= (SD$ of box$)/\sqrt{n}$ (see shortcut formula below)

Shortcut formula for SD of a box with only 2 values of tickets (e.g., 0's and 1's):

$$(\text{larger value} - \text{smaller value})\cdot\sqrt{(\text{fraction with larger}) \cdot (\text{fraction with smaller})}$$

Bootstrap method: using an average or SD of a sample as an estimate for the average or SD of the population. (For a small sample, use SD^+ of the sample to approximate SD of the population — see below.)

95% confidence interval for the average of a population (if it isn't too "abnormal"):

$$(\text{average of sample}) \pm 2 \cdot (SE \text{ for average})$$

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

For a sample that is a large fraction of the population and sampling is without replacement, to approximate SE for percent or average (using the "correction factor"):

$$SE \text{ without replacement} = \sqrt{\frac{\#\text{ in box} - \#\text{ of draws}}{\#\text{ in box} - 1}} \cdot (SE \text{ with replacement})$$

For use with t-test for significance on small ($n < 25$) samples: First, if population SD is implied by null hypothesis, use it; otherwise, estimate with:

$$SD^+ \text{ of sample} = \sqrt{\frac{\#\text{ in sample}}{\#\text{ in sample} - 1}} \cdot (SD \text{ of sample})\ .$$ Then:

$$t = \frac{\text{average of sample} - EV}{SE}$$ (EV from null hypothesis;

$$SE = \frac{\text{population } SD}{\sqrt{\#\text{ in sample}}}$$, maybe bootstrapped)

degrees of freedom: (# in sample) -1

SE for difference of averages of 2 samples $= \sqrt{(SE \text{ of first})^2 + (SE \text{ of second})^2}$
(For more than 2 samples, use one-way ANOVA.)

For deciding significance of differences in frequency distributions:

$\chi^2 =$ sum of $[(\text{observed} - \text{expected})^2/\text{expected}]$
degrees of freedom: in "list" distributions, # of terms -1;
in "table" distributions, (# of rows -1) $\cdot$ (# of columns -1)