STA101 Formula Sheet

Luc Hens
Vrije Universiteit Brussel

Reset the *TI-84*: Mem → Reset → Reset RAM.

Scientific notation of numbers on *TI-84*: `E` stands for "10 to the power of":

`2.3E+04` $= 2.3 \times 10^4 = 2.3 \times 10\,000 = 23\,000$ in R: `2.3e+4`

`2.3E-04` $= 2.3 \times 10^{-4} = 2.3 \times \frac{1}{10^4} = 2.3 \times \frac{1}{10\,000} = 0.00023$ in R: `2.3e-4`

Descriptive statistics

$$\text{density} = \frac{\text{relative frequency}}{\text{width of interval}} \qquad \text{(in a density histogram)}$$

$$\text{IQR} = \text{Q3} - \text{Q1}$$

$$\text{mean} = \bar{x} = \frac{\sum x}{n}$$

$$\text{deviation (from the mean)} = x - \bar{x}$$

$$\text{standard deviation} = \begin{pmatrix} \text{quadratic mean} \\ \text{of the deviations} \end{pmatrix} \quad : \quad s = \sqrt{\frac{\sum (x - \bar{x})^2}{n-1}}$$

(quadratic mean = root-mean-square)

$$z\text{-score} = \frac{x - \bar{x}}{s}$$

TI-84: Enter the data (e.g., in list L_1): STAT → EDIT ; STAT → CALC → 1-Var Stats L_1

R: Enter the data: `x <- c(2.3, 3.4, ...)` ; `summary(x)` ; `sd(x)`

Correlation, regression:

$$\text{correlation coefficient} = \text{mean of } [(x \text{ in } z\text{-scores}) \times (y \text{ in } z\text{-scores})] \quad : \quad r = \frac{\sum z_x z_y}{n-1}$$

$$\text{regression line (line of best fit):} \quad \hat{y} = b_0 + b_1 x \qquad b_1 = r\frac{s_y}{s_x} \qquad b_0 = \bar{y} - b_1 \bar{x}$$

TI-84: enter the data (x in list L_1, y in list L_2): STAT → EDIT
CATALOG → DiagnosticOn → ENTER → ENTER
STAT → CALC → LinReg(ax+b) L_1, L_2 (x list first, y list second)
Output: `a` = slope = b_1 in $\hat{y} = b_0 + b_1 x$; `b` = intercept = b_0 in $\hat{y} = b_0 + b_1 x$

R: Enter the data: `x <- c(2.3, 3.4, ...)` ; `y <- c(5.1, 6.9, ...)`
`cor(x,y)` ; `lm(y ~ x)` ; `summary(lm(y ~ x)`

Probability

Complement rule:	$P(\textit{not } \mathbf{A}) = 1 - P(\mathbf{A})$
General multiplication rule:	$P(\mathbf{A} \textit{ and } \mathbf{B}) = P(\mathbf{A}) \times P(\mathbf{B}\mid\mathbf{A})$
Multiplication rule for independent events:	$P(\mathbf{A} \textit{ and } \mathbf{B}) = P(\mathbf{A}) \times P(\mathbf{B})$
General addition rule:	$P(\mathbf{A} \textit{ or } \mathbf{B}) = P(\mathbf{A}) + P(\mathbf{B}) - P(\mathbf{A} \textit{ and } \mathbf{B})$
Addition rule for disjoint events:	$P(\mathbf{A} \textit{ or } \mathbf{B}) = P(\mathbf{A}) + P(\mathbf{B})$
Conditional probability:	$P(\mathbf{B}\mid\mathbf{A}) = \dfrac{P(\mathbf{A} \textit{ and } \mathbf{B})}{P(\mathbf{A})}$
Events **A** and **B** are independent when	$P(\mathbf{B}\mid\mathbf{A}) = P(\mathbf{B})$

Discrete probability distributions

$E(X) = \mu = \sum xP(x)$
$Var(X) = \sigma^2 = \sum (x-\mu)^2 P(x)$ $\quad SD(X) = \sigma = \sqrt{\sum (x-\mu)^2 P(x)}$
$Cov(X,Y) = E[(X-\mu_X)(Y-\mu_Y)]$

$E(X \pm c) = E(X) \pm c$	$Var(X \pm c) = Var(X)$	$SD(X \pm c) = SD(X)$
$E(aX) = aE(X)$	$Var(aX) = a^2 Var(X)$	$SD(aX) = \lvert a \rvert SD(X)$
$E(X \pm Y) = E(X) \pm E(Y)$	$Var(X \pm Y) = Var(X) + Var(Y) \pm 2Cov(X,Y)$	

Binomial model: For Bernoulli trials, with p = probability of success:

X = number of successes in n trials

$P(X = x) = P(\text{exactly } x \text{ successes in } n \text{ trials}) = \dfrac{n!}{x!(n-x)!}\, p^x (1-p)^{n-x}$

$E(X) = np \qquad SD(X) = \sqrt{np(1-p)}$

TI-84: DISTR → binompdf(n, p, x) R: `dbinom(x,n,p)`
P(2 successes in 5 trials when $p = 0.10$): *TI-84*: binompdf(5, 0.10, 2) R: `dbinom(2,5,0.10)`

Continuous probability distributions

Normal distribution: Use *TI-84*, R to find **areas** under the normal pdf:

to the left of 2:	*TI-84*: DISTR → normalcdf($-10 \wedge 99$,2)	R: `pnorm(2)`
between -2 and 1:	*TI-84*: DISTR → normalcdf(-2,1)	R: `pnorm(1) - pnorm(-2)`
to the right of 1:	*TI-84*: DISTR → normalcdf(1, $10 \wedge 99$)	R: `1-pnorm(1)`

Use *TI-84*, R to find a **percentile** of the normal distribution:

95^{th} percentile	*TI-84*: DISTR → invNorm(.95)	R: `qnorm(0.95)`

Sampling distributions

Provided that the sampled values are independent and the sample size is large enough (*), in repeated samples the **sample proportion** $\hat{p}$ is approximately normally distributed with

$E(\hat{p}) = p \quad \text{and} \quad SD(\hat{p}) = \sqrt{\dfrac{p(1-p)}{n}} \qquad (*)\ n\hat{p} > 10 \text{ and } n(1-\hat{p}) > 10$

Provided that the sample size is large enough, in repeated samples the **sample mean** $\bar{y}$ is approximately normally distributed with

$E(\bar{y}) = \mu \quad \text{and} \quad SD(\bar{y}) = \dfrac{\sigma}{\sqrt{n}}$

Confidence intervals

(Approximate) 95% confidence interval for a **proportion** if the normal approximation works:

$\hat{p} \pm 2 \times SE(\hat{p}) \qquad \hat{p} \pm 2 \times \sqrt{\dfrac{\hat{p}(1-\hat{p})}{n}}$

TI-84: STAT → TESTS → 1-PropZInt (`x`: number of successes in the sample)
R: `binom.test(k, n=...)` (`k`: number of successes in the sample)

(Approximate) 95% confidence interval for a **mean** (only valid for a **large sample**):

$\bar{y} \pm 2 \times SE(\bar{y}) \qquad \bar{y} \pm 2 \times \dfrac{s}{\sqrt{n}}$

TI-84: STAT → TESTS → ZInterval
R: If you have the data in a list `x`: `t.test(x)`. Otherwise: enter the values of $\bar{y}$, s, n:

```
y.bar <- 170 ; s <- 10 ; n <- 90
margin.of.error <- qnorm(0.975)*s/sqrt(n)
y.bar - margin.of.error ; y.bar + margin.of.error
```