Part 2: Probability, Probability Distributions, and Sampling Distributions


**Probability**

Suppose that $A$ and $B$ are events in a sample space $S$. Then,

\[ P(A \cup B) = P(A) + P(B) - P(A \cap B) \]

$P(A \cap B) = P(A)P(B)$ if $A$ and $B$ are independent events

$P(A^c) = 1 - P(A)$

$P(A \mid B) = \frac{P(A \cap B)}{P(B)}$ if $P(B) \neq 0$

**Normal Probability Distribution**

Suppose that $X \sim N(\mu, \sigma)$. Be able to calculate $P(a \leq X), P(X \leq b)$ and $P(a \leq X \leq b)$ for specific $a$ and $b$. Be able to calculate $q$ such that $P(X \leq q) = p$ for a given probability $p$. Be able to illustrate each of these calculations with an appropriate drawing of a normal probability distribution. Given $x$, calculate its $z$-score $z = (x - \mu) / \sigma$. Given $z$, be able to calculate $x = \mu + z \times \sigma$. Explain what the $z$-score of $x$ represents. Be able to use the R commands `pnorm`, `qnorm`.

**Binomial Probability Distribution**

If $X \sim \text{Binomial}(n, p)$, then be able to calculate the probability of $k$ successes in $n$ trials, $P(X = k)$, by using the R command `dbinom(k, n, p)`. Understand that the mean of such a binomial distribution is $np$ and the standard deviation is $\sqrt{np(1 - p)}$.

**Distribution of a statistic (mean or proportion)**

Two scenarios: (1) If we are studying a quantitative variable and we know the population parameters $\mu$ and $\sigma$, what can be said of the sample statistics? (2) If we know the sample statistics $\bar{x}$ and $s$, what can be said of the population parameters? The first question is easiest to answer, but is rarely the case. The second question leads to much of contemporary statistics.

We might be studying a quantitative variable in a population (normal or not) with population parameters $\mu$ and $\sigma$, and we wish to know the sampling distribution of the sample mean $\bar{x}$. Or we might be studying a categorical variable in a population with proportion $p$, and we wish to know the sampling distribution of the sample proportion $\hat{p}$. The following table summarizes the sample distributions in both cases (Probability and Statistics, Open Learning Initiative, CMU).

<table>
<thead>
<tr>
<th>Variable</th>
<th>Statistic</th>
<th>Shape</th>
<th>Center</th>
<th>Standard Error</th>
<th>Conditions</th>
</tr>
</thead>
<tbody>
<tr>
<td>quantitative ($\sigma$ known)</td>
<td>$\bar{x}$</td>
<td>Normal</td>
<td>$\mu$</td>
<td>$\sigma/\sqrt{n}$</td>
<td>$n \geq 30$ or approx. normal</td>
</tr>
<tr>
<td>quantitative ($\sigma$ unknown)</td>
<td>$\bar{x}$</td>
<td>$t$</td>
<td>$\mu$</td>
<td>$s/\sqrt{n}$</td>
<td>$n \geq 30$ or approx. normal</td>
</tr>
<tr>
<td>categorical</td>
<td>$\hat{p}$</td>
<td>Normal</td>
<td>$p$</td>
<td>$\sqrt{p(1-p)/n}$</td>
<td>$\min(np, n(1-p)) \geq 15$</td>
</tr>
</tbody>
</table>
Review exercises for Part 2

Consolidate your understanding of the concepts in these chapters by working through a number of these exercises.

Exercises for Chapter 5:
5.75 (health insurance), 5.78 (detergent), 5.88 (screening), 5.90 (color blind), 5.100 (heroin), 5.101 (school)

Exercises for Chapter 6:
6.35 (symmetric binomial), 6.36 (girls), 6.43 (jury duty), 6.58 (cholesterol), 6.61 (gestation), 6.62 (water), 6.67 (tennis balls)

Exercises for Chapter 7:
7.30 (exam), 7.32 (Alzheimer’s), 7.37 (home runs), 7.38 (physicians assistants), 7.39 (withdrawals)

Exercises for the Review of Part 2:
R2.3 (stem cell), R2.4 (happy), R2.8 (SAT), R2.13 (war), R2.15 (ice cream)