EPSY 8264 Prerequisite Information

2018-05-02

This handout contains a list of basic results and concepts that should be familiar to each of you prior to taking EPsy 8264. Please review your previous texts/notes for information on these topics.

1. Summation rules. These rules form the basis for many of the earlier proofs that we will do for the class. In the following, X and Y are variables and a is a constant:

   a) $\sum X$
   b) $\sum(X + Y)$
   c) $\sum a$
   d) $\sum aX$
   e) $\sum(XY)$

2. Random Variables

   a) Expected value (definition/meaning)
   b) Variance (definition/meaning)
   c) Probability distribution—You should have a mathematical as well as a conceptual idea of each of the following distributions.
      i) Joint
      ii) Marginal
      iii) Conditional
   d) Covariance
   e) Correlation
   f) Statistical and conceptual idea of what it means for variables to be independent.
   g) Functions of random variables and the effect of transformations on:
      i) Expected values
      ii) Variances
      iii) Correlations (covariances)
   h) Distributions of data
      i) Number of variables (univariate, bivariate, multivariate)
      ii) Characteristics of distributions (location, spread, shape)
      iii) Shape (normal, chi-square, uniform)
   i) Linear interpolation

3. Central Limit Theorem ("general" definition and utility)

4. Normal probability distribution

   a) Link probabilities to test statistics (e.g., `pnorm()`, `qnorm()`)
   b) Define and understand implications of equivalent scores
   c) Understand implications of a standard score (especially one that is normally distributed)
   d) Perform hypothesis tests and accurately interpret p-values.

5. t-distributions

   a) Link probabilities to test statistics (e.g., `pt()`, `qt()`)
   b) Define degrees of freedom
   c) Understand how to find degrees of freedom (*df*) for a host of simple statistical methods
   d) Perform hypothesis tests using t-distributions.

6. F-distributions

    a) Link probabilities to test statistics (e.g., `pf()`, `qf()`)
    b) Numerator and denominator degrees of freedom
    c) Perform an overall test of a model using $F = \mathrm{MS}_{\mathrm{Model}}/\mathrm{MS}_{\mathrm{Error}}$
    d) Test of model comparison (e.g., Nested F test)

7. Chi-Square (χ^2) distributions

    a) Link probabilities to test statistics (e.g., `pchisq()`, `qchisq()`)
    b) Degrees of freedom
    c) Tests of independence
    d) Tests of a given hypothesis
    e) Test of model comparison (e.g., Deviance test)

8. Statistical estimation

    a) Understanding of what it means to be an estimator
    b) *Biased* versus *unbiased* estimators
    c) Consistent estimators
    d) Minimum variance estimators
    e) Confidence intervals

9. Least Squares Estimation

    a) Obtaining estimators from very simple situations via the use of algebra.
    b) Obtaining standard error estimates for least squares (and other) estimators

10. Inference

    a) Empirical versus theoretical distributions.
    b) Hypothesis testing (very general)
    c) Interval estimation (confidence intervals – very general)
    d) Type I, Type II errors and their relation to sample size, number of tests, and effect (relationship of the statistic under observation)
    e) Relationship between hypothesis testing and confidence intervals
    f) Test for the equality of variances
    g) Statistical (p-value) versus practical (effect size) difference

11. Mathematics to review

    a) Algebra 1
    b) Algebra 2
    c) Geometry
    d) Matrix Algebra

Summation Rules

Rule 1: When a quantity which is itself a sum or difference is to be summed, the summation sign may be distributed among the separate terms of the sum. That is:

$$\sum(X+Y) = \sum X + \sum Y$$

This rule can be verified using a numerical example.

Table 1: *Numeric Example of Rule 1. Gray Row is the Sum of Each Column*

X	Y	$(X+Y)$
2	4	6
3	5	8
3	4	7
1	2	3
2	3	5
4	6	10
5	6	11
4	2	6
24	**32**	**56**

Below, we carry out the computation in R.

```
X = c(2, 3, 3, 1, 2, 4, 5, 4)
Y = c(4, 5, 4, 2, 3, 6, 6, 2)

sum(X + Y)
```

```
## [1] 56
```

```
sum(X) + sum(Y)
```

```
## [1] 56
```

```
# Use a logical statement to check this is TRUE
sum(X + Y) == sum(X) + sum(Y)
```

```
## [1] TRUE
```

The numeric example, which is often used as a sanity check, verifies the result for a specific X and Y. A more powerful result can be a mathematical proof, which allows us to verify that the rule is true more generally (for all variables X and Y)

Proof.

$$\begin{aligned}\sum(X+Y) &= (X_1+Y_1)+(X_2+Y_2)+\ldots+(X_n+Y_n)\\ &= (X_1+X_2+\ldots X_n)+(Y_1+Y_2+\ldots Y_n)\\ &= \sum X+\sum Y\end{aligned}$$

By extension of the proof, it follows that:

$$\sum(X+Y+Z) = \sum X+\sum Y+\sum Z\text{ , etc.}$$

□

Rule 2: The sum of a random variable and a constant equals the sum of the variable plus n times the constant, where n is the length of the random variable.

$$\sum(X+a) = \sum X+na$$

First we verify this for a specific case, and then show the mathematical proof.

Table 2: *Numeric Example of Rule 2. Gray Row is the Sum of Each Column*

X	a	$(X+3)$
2	3	5
3	3	6
3	3	6
1	3	4
2	3	5
4	3	7
5	3	8
4	3	7
24	**24**	**48**

```
n = length(X)

# Use a logical statement to check this is TRUE
sum(X + 3) == sum(X) + (n * 3)
```

```
## [1] TRUE
```

To construct the mathematical proof, we first define $\sum a$:

$$\begin{aligned}\sum a &= n \times a \\ &= \underbrace{a + a + \ldots + a}_{n} \\ &= na\end{aligned}$$

We now use this definition along with Rule 1 in the proof.

Proof.

$$\begin{aligned}\sum(X + a) &= \sum X + \sum a \\ &= \sum X + na\end{aligned}$$

□

Rule 3: The sum of the product of a constant and a variable is equivalent to the product of the constant and the sum of the variable.

$$\sum(aX) = a\sum X$$

Again, we first verify this for a specific case, and then show the mathematical proof.

Table 3: *Numeric Example of Rule 3. Gray Row is the Sum of Each Column*

X	$3X$
2	6
3	9
3	9
1	3
2	6
4	12
5	15
4	12
24	**72**

```
# Use a logical statement to check this is TRUE
sum(3 * X) == 3 * sum(X)
```

```
## [1] TRUE
```

Proof.

$$\begin{aligned}\sum(aX) &= a(X_1) + a(X_2) + \ldots + a(X_n) \\ &= a(X_1 + X_2 + \ldots X_n) \\ &= a\sum X\end{aligned}$$

□

The rules can be combined, and extended. Some illustrations are given below and the rules involved are listed. In the illustrations, a and b represent constants, and X and Y are variables. Try to verify a specific case and prove a few of these.

1. $\sum(X + 2) = \sum X + 2n$ (using Rules 1 and 2)
2. $\sum(X^2 - 1) = \sum X^2 - n$ (using Rules 1 and 2); [Note: if X is a variable, so is X^2]
3. $\sum 2a = 2 \times a \times n$ (using Rule 2)
4. $\sum ab^2XY^2 = ab^2\sum XY^2$ (using Rule 3)
5. $\sum a(Y + 3)^2 = a\sum(Y + 3)^2$ (using Rule 3); [Note: if Y is a variable, so is $(Y + 3)$]
6. $\sum(2X - 3) = 2\sum X - 3n$ (using Rules 1, 2, and 3)
7. $\sum(Y - a)^2 = \sum Y^2 - 2a\sum Y + na^2$ (using Rules 1, 2, and 3)

Computational Formulas

In days of yesteryear, these summations were used to provide computational formulas that formed the basis for calculating sums of squares, variances, and other useful measures (see, http://www.ablongman.com/graziano6e/text_site/MATERIAL/Stats/manvar.htm). When pen-and-paper, not computers, were the way to do statistics, reducing quantities to easy to calculate summaries (e.g., $\sum X^2$) was critical. For example, in computing the sum of squared deviations from the mean, we were interested in

$$\sum(X - \bar{X})^2$$

Using the Example 7 (above), we could re-write this as

$$\sum X^2 - 2\bar{X}\sum X + n\bar{X}^2$$

Substituting $\sum X/n$ in for $\bar{X}$ and reducing we get:

$$\begin{aligned}&= \sum X^2 - 2\left(\frac{\sum X}{n}\right)\sum X + n\left(\frac{\sum X}{n}\right)^2 \\ &= \sum X^2 - \frac{2\left(\sum X\right)^2}{n} + \frac{\left(\sum X\right)^2}{n} \\ &= \sum X^2 - \frac{\left(\sum X\right)^2}{n}\end{aligned}$$

With computers, computational formulas are rarely used in practice; it is easier to calculate quantities directly from the data.

```r
X = c(2, 3, 3, 1, 2, 4, 5, 4)

# Compute from data
sum( (X - mean(X)) ^ 2 )
```

```
## [1] 12
```

```r
# Use computational formula
n = length(X)
sum(X^2) - (sum(X)^2) / n
```

```
## [1] 12
```

Basic Proofs

Here are several definitions and basic proofs which should be helpful as you are reading the text and trying to follow some of the mathematical arguments presented there.

Quantity	Notation	Formula
Population size	N	
Mean	μ	$\frac{\sum x_i}{N}$
Mean deviation		$x_i - \mu$
Sum of squared deviations		$\sum (x_i - \mu)^2$
Variance (Population)	σ^2	$\frac{\sum (x_i - \mu)^2}{N}$

Biased and Unbiased Estimators

An estimator whose average value over all possible samples is equal to its population value is *unbiased*. An estimator that is consistently too high or too low is biased (high or low).

Consider a population with $N = 4$ values: 3, 5, 7, 9. This population has the following parameters:

$$\mu = \frac{3 + 5 + 7 + 9}{4} = 6$$

$$\sigma^2 = \frac{(3 - 6)^2 + (5 - 6)^2 + (7 - 6)^2 + (9 - 6)^2}{4} = 5$$

Now, consider sampling two observations at random ($n = 2$). For example, suppose you randomly chose $\{3, 5\}$. Based on this simple random sample (SRS) we can estimate the population mean and variance,

$$\hat{\mu} = \bar{x} = \frac{3 + 5}{2} = 4$$

and

$$\hat{\sigma}^2 = s^2 = \frac{(3 - 4)^2 + (5 - 4)^2}{2} = 1$$

The sample mean and variance are referred to as *estimators*. (They are estimates of the population parameters.) When we refer to the biased-ness or unbiased-ness of an estimator, we need to consider the average value across ALL POSSIBLE samples of a particular size; in our case $n = 2$. With only four different values in the population, we can actually list all the possible samples of size 2 (we need to sample *with replacement*):

Table 5: *The Estimated Mean and Variance for All Possible SRSs of Size n=2 Drawn from 3, 5, 7, 9.*

SRS	Observations	x1	x2	Mean	Variance
1	{3, 3}	3	3	3	0
2	{3, 5}	3	5	4	1
3	{3, 7}	3	7	5	4
4	{3, 9}	3	9	6	9
5	{5, 3}	5	3	4	1
6	{5, 5}	5	5	5	0
7	{5, 7}	5	7	6	1
8	{5, 9}	5	9	7	4
9	{7, 3}	7	3	5	4
10	{7, 5}	7	5	6	1
11	{7, 7}	7	7	7	0
12	{7, 9}	7	9	8	1
13	{9, 3}	9	3	6	9
14	{9, 5}	9	5	7	4
15	{9, 7}	9	7	8	1
16	{9, 9}	9	9	9	0

Below we plot the 16 estimated mean values.

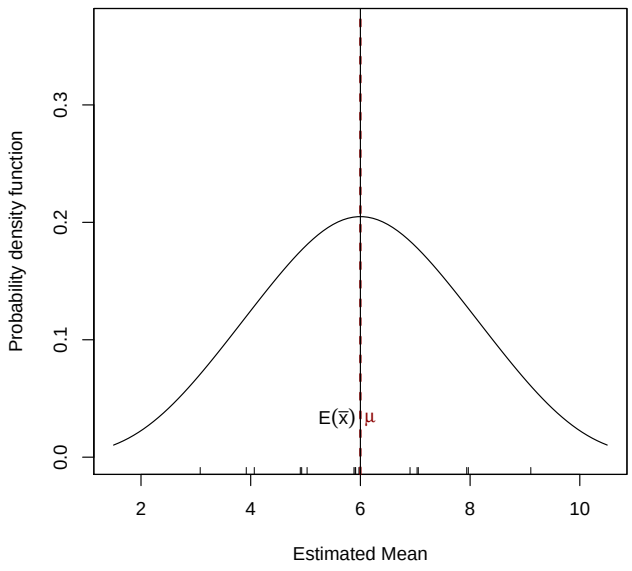

Figure 1: *Dotplot of the 16 estimated mean values. The average of these values (black line) and the population variance (darkred line) are also displayed.*

The average of the 16 estimated mean values is 6. This is the value of the population parameter. The sample mean is an unbiased estimator of the population mean. Mathematically, we write

$$E(\bar{x}) = \mu$$

Now let's look at the 16 estimated variances

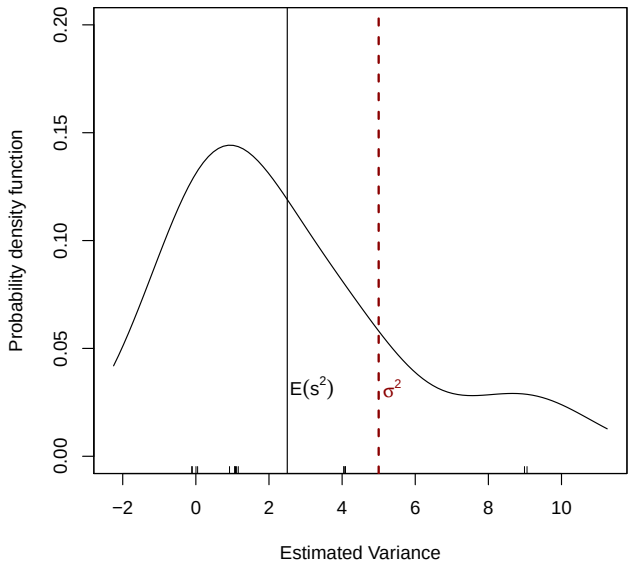

Figure 2: *Dotplot of the 16 estimated variance values. The average of these values (black line) and the population variance (darkred line) are also displayed.*

The average of the 16 estimated variance values is 2.5. This is NOT the value of the population parameter. The sample variance is a biased estimator of the population variance. Mathematically, we write

$$E(s^2) \neq \sigma^2$$

Since the average value is LESS than the value of the population variance ($2.5 < 5$), we say that the sample variance UNDERESTIMATES the population variance.

To account for this, when we are estimating the population variance from a sample, we typically divide the sum of squared deviations by $n - 1$ rather than n.

$$s^2 = \frac{\sum(x_i - \hat{\mu})^2}{n - 1}$$

Updating this in the 16 samples,

Table 6: *The Estimated Mean and Variance for All Possible SRSs of Size n=2 Drawn from 3, 5, 7, 9.*

SRS	Observations	x1	x2	Mean	Variance	Variance (n-1)
1	{3, 3}	3	3	3	0	0
2	{3, 5}	3	5	4	1	2
3	{3, 7}	3	7	5	4	8
4	{3, 9}	3	9	6	9	18
5	{5, 3}	5	3	4	1	2
6	{5, 5}	5	5	5	0	0
7	{5, 7}	5	7	6	1	2
8	{5, 9}	5	9	7	4	8
9	{7, 3}	7	3	5	4	8
10	{7, 5}	7	5	6	1	2
11	{7, 7}	7	7	7	0	0
12	{7, 9}	7	9	8	1	2
13	{9, 3}	9	3	6	9	18
14	{9, 5}	9	5	7	4	8
15	{9, 7}	9	7	8	1	2
16	{9, 9}	9	9	9	0	0

Now the average of the 16 estimated variance values is 5, the value of the population parameter. The sample variance is a biased estimator of the population variance when we divide by n, but is unbiased when we divide by $n - 1$.

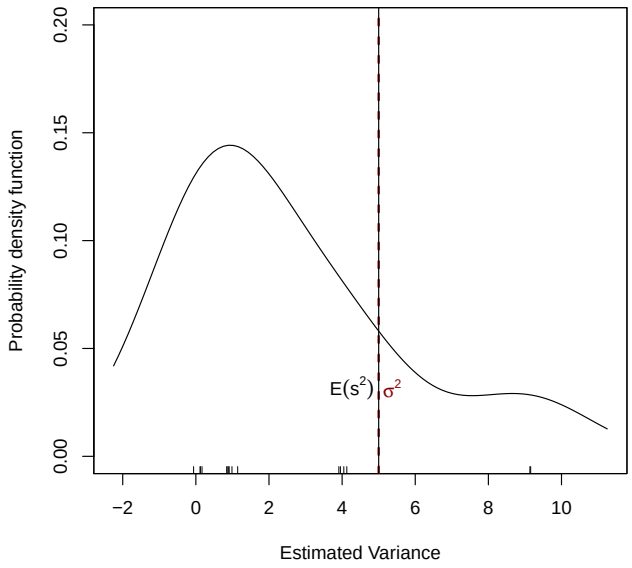

Figure 3: *Dotplot of the 16 estimated variance values using a denominator of n-1. The average of these values (black line) and the population variance (darkred line) are also displayed.*

Standard Deviation

What about the standard deviation? Is it an unbiased estimator? The population standard deviation is $\sqrt{5} = 2.24$; this is the square root of the population variance. If we compute the standard deviation for each of the 16 samples, we get:

Table 7: *The Estimated Mean and SD (using n and n-1 in the denominator) for All Possible SRSs of Size n=2 Drawn from 3, 5, 7, 9.*

SRS	Observations	x1	x2	Mean	SD (n)	SD (n-1)
1	{3, 3}	3	3	3	0	0.000000
2	{3, 5}	3	5	4	1	1.414214
3	{3, 7}	3	7	5	2	2.828427
4	{3, 9}	3	9	6	3	4.242641
5	{5, 3}	5	3	4	1	1.414214
6	{5, 5}	5	5	5	0	0.000000
7	{5, 7}	5	7	6	1	1.414214
8	{5, 9}	5	9	7	2	2.828427
9	{7, 3}	7	3	5	2	2.828427
10	{7, 5}	7	5	6	1	1.414214
11	{7, 7}	7	7	7	0	0.000000
12	{7, 9}	7	9	8	1	1.414214
13	{9, 3}	9	3	6	3	4.242641
14	{9, 5}	9	5	7	2	2.828427
15	{9, 7}	9	7	8	1	1.414214
16	{9, 9}	9	9	9	0	0.000000

The average of the 16 values for the SD based on dividing by n is 1.25. This is a biased estimate (underestimate) of the population standard deviation. The average of the 16 values for the SD based on dividing by $n - 1$ is 1.77. This is also a biased (under-) estimate of the population standard deviation.

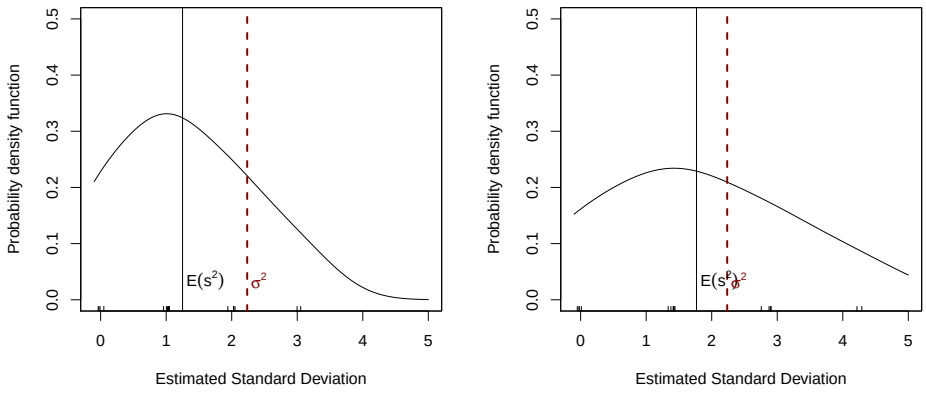

Figure 4: *Dotplot of the 16 estimated standard deviation values using a denominator of n (left-hand plot) and n-1 (right-hand plot). The average of these values (black line) and the population variance (darkred line) are also displayed in each plot, respectively.*

KEY POINT: It is the variances which are unbiased (not the standard deviations)! This is why statisticians like to estimate variances rather than standard deviations; despite their unwieldy (squared) metric.

References