# III. Sampling

## 1 Overview of Sampling, Error, Bias

### 1.1 Biased vs. random sampling

### 1.2 Biased vs. unbiased statistic (or estimator)

### 1.3 Precision vs. accuracy

## 2 Error Estimates With Assumed Sampling Distribution

### 2.1 Standard Error:

Standard deviation of distribution of sample statistics that would result from infinite number of trials of drawing sample from underlying probability distribution and calculating the sample statistic.

### 2.2 In practice we generally do not estimate error by repeated sampling from the underlying distribution (expensive and time-consuming), although there are exceptions.

### 2.3 Approximations based on sample distribution (from Sokal and Rohlf):

**BOX 7.1 Standard Errors for Common Statistics**

| | (1) Statistic | (2) Estimate of standard error | (3) *df* | (4) Comments on applicability |
|---|---|---|---|---|
| 1 | $\bar{Y}$ | $s_{\bar{Y}} = \frac{s}{\sqrt{n}} = \frac{s_Y}{\sqrt{n}} = \sqrt{\frac{s^2}{n}}$ | $n-1$ | True for any population with finite variance |
| 2 | Median | $s_{\text{med}} = (1.2533)s_{\bar{Y}}$ | $n-1$ | Large samples from normal populations |
| 3 | Average deviation (AD) | $s_{AD} \approx (0.602,810,3)\frac{s}{\sqrt{n}}$ | $n-1$ | For large $n$( > 100) |
| | | $= \frac{s}{n}\sqrt{\frac{2(n-1)}{\pi}\left[\frac{\pi}{2} + \sqrt{n(n-1)} - n - \arcsin\left\{\frac{1}{(n-1)}\right\}\right]}$ | $n-1$ | Samples from normal populations |
| 4 | $s$ | $s_s = (0.707,106,8)\frac{s}{\sqrt{n}}$ | $n-1$ | Samples from normal populations ($n$ > 15) |
| 5 | $V$ | $s_V \approx \frac{V}{\sqrt{2n}}\sqrt{1+2\left(\frac{V}{100}\right)^2}$ } Kendall & Stuart, v. I, ex. 19.5 (2nd ed.) | $n-1$ | Samples from normal populations |
| | | $s_V \approx \frac{V}{\sqrt{2n}}$ | $n-1$ | Used when $V$ < 15 |
| 5* | $V^*$ | $s_{V^*} = \left(1+\frac{1}{4n}\right)s_V$ | $n-1$ | Samples from normal populations |
| 6 | $g_1$ | $s_{g_1} = \sqrt{\frac{6n(n-1)}{(n-2)(n+1)(n+3)}} \approx \sqrt{\frac{6}{n}}$ | $\infty$ | Samples from normal populations. The approximate formula is for large $n$ ($n$ > 150) |
| 7 | $g_2$ | $s_{g_2} = \sqrt{\frac{24n(n-1)^2}{(n-3)(n-2)(n+3)(n+5)}} \approx \sqrt{\frac{24}{n}}$ | $\infty$ | Samples from normal populations. The approximate formula is for large $n$ ($n$ > 150) |

### 2.4 Limitations:

**2.4.1 Many approximation formulae make assumptions about shape of distribution and sample size.**

**2.4.2 We may be interested in novel statistic or one whose sampling distribution is not well characterized.**

## 3 Bootstrap Error Estimates

### 3.1 Estimate standard error by resampling from the single sample we have.

### 3.2 This approach uses sampling with replacement from observed sample to simulate sampling without replacement from the underlying distribution.

### 3.3 Procedure

**3.3.1 Start with observed sample of size $n$ and observed sample statistic, call it $Z$.**

**3.3.2 Randomly pick a sample of size $n$, with replacement, from the observed sample.**

**3.3.3 Calculate the sample statistic of interest on this random sample; call is $Z_{boot}$.**

**3.3.4 Repeat many times (generally hundreds to thousands, ideally until estimate of $SE$ stabilizes).**

**3.3.5 Calculate standard deviation of the $Z_{boot}$.**

This is an estimate of the standard error of the observed sample statistic $Z$:
$SD(Z_{boot}) \approx SE(Z)$.

### 3.4 Simple (but not necessarily most useful) example: trimmed mean

- Define $p$-% trimmed mean as mean of sample with $p$% lowest and $p$% highest observations discarded. (Idea is to try to reduce effect of outliers.)
- Suppose data consist of 10 (ordered) observations: 1,2,3,4,8,10,12,15,20,30. Let the trimmed mean be denoted $Z$. Then $Z = (3+4+8+10+12+15)/6 = 8.67$.