Prof. David Draper
Department of Applied Mathematics and Statistics
University of California, Santa Cruz

AMS 131: Quiz 9

Name: Jeremy Lafond

You're about to take an IID sample $(X_1, \ldots, X_n)$ from a distribution with variance $V(X_i) = \sigma^2$ that exists and is finite, which implies that the mean $E(X_i) = \mu$ also exists and is finite. The purpose of the sampling is to use the sample mean $\bar{X}_n = \frac{1}{n}\sum_{i=1}^n X_i$ as an estimator of μ, and you're wondering what value you should use for the sample size n. We've seen in class that *Chebyshev's Inequality* can help when no other details about the distribution of the X_i are available: if Y is any random variable whose variance $V(Y)$ exists, this inequality states that for any $t \geq 0$

$$P(|Y - E(Y)| \geq t) \leq \frac{V(Y)}{t^2} \, . \tag{1}$$

(a) Using basic facts about $E(\bar{X}_n)$ and $V(\bar{X}_n)$, show that inequality (1) implies, in the random sampling problem considered here, that for any $k > 0$

$$P(|\bar{X}_n - \mu| < k\,\sigma) \geq 1 - \frac{1}{n\,k^2} \, , \tag{2}$$

and show further that, if we want the probability in (2) to be at least $(1 - \alpha)$ for some $0 < \alpha < 1$, we should choose

$$n_{Chebyshev} \geq \frac{1}{\alpha\,k^2} \, . \tag{3}$$

$$\bar{X}_n = \frac{1}{n}\sum_{i=1}^n X_i$$

$$\sum(\bar{X}_n) = \frac{1}{n}\sum_{i=1}^n (X_i) = \mu$$

$$\ell V(\bar{X}_n) = \frac{1}{n^2}\sum_{\ell=1}^n V(X_1) = \frac{1}{n^2}(n\sigma^2) = \frac{\sigma^2}{n}$$

$$V(\bar{X}_n) = \frac{\sigma^2}{n}$$

By substitution(Let Y = X_n) :$P(|\bar{X}_n - E(\bar{X}_n)| \geq t) \leq \frac{V(\bar{X}_n)}{t^2} Lett = k\sigma,$

$$P(|\bar{X}_n - \mu| \geq k\sigma) \leq \frac{\sigma^2}{n(k\sigma)^2}$$

$$P(|\bar{X}_n - \mu| \geq k\sigma) \leq \frac{1}{nR^2}$$

$$1 - P(|\bar{X}_n - \mu| \geq k\sigma) \geq 1 - \frac{1}{nR^2}$$

$$P(|\bar{X}_n - \mu| < k\sigma) \geq 1 - \frac{1}{nR^2} - eq(2)$$

if probability in eq(2) is at least $(1 - \alpha)$ then:

$$P(|\bar{X}_n - \mu| \geq k\sigma) \leq 1 - \alpha$$

$$1 - \frac{1}{nR^2} \geq 1 - \alpha$$

$$\frac{1}{nR^2} \geq 1 - \alpha$$

$$\frac{1}{nR^2} \leq \alpha$$

$$n_{Chebyshev} \geq \frac{1}{\alpha k^2} - eq(3)$$

I also mentioned in class that Chebyshev's Inequality can be quite conservative. Let's quantify this: suppose for the rest of the problem that $(X_i \,|\, \mu\, \sigma^2) \overset{IID}{\sim} N(\mu, \sigma^2)$. Then we've seen in class that $\bar{X}_n$ also follows a Normal distribution, with mean μ and standard error $SE\left(\bar{X}_n\right) = \frac{\sigma}{\sqrt{n}}$. As usual let $\Phi(x)$ be the standard Normal CDF; in other words, $\Phi(x)$ is the area to the left of x under the standard Normal PDF.

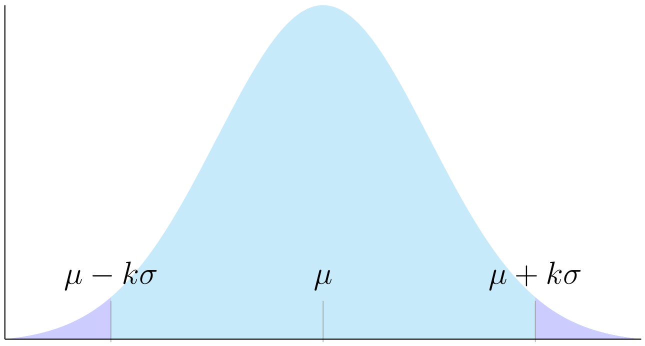
(b) Sketch the PDF of $\bar{X}_n$, shading in the area corresponding to $(*) = (|\bar{X}_n - \mu| < k\,\sigma)$ and identifying the places on both the raw-units and standard-units axes corresponding to the endpoints of $(*)$.

$$X_i \sim N(\mu, \sigma^2)$$

$$\bar{X}_n \sim N(\sigma, \frac{\sigma^2}{n})$$

$$Z' = \bar{X}_n - \mu \sim N(0, \frac{\sigma^2}{n})$$

$$Z = \frac{\bar{X}_n - \mu}{\sigma\sqrt{n}} \sim N(0, 1)$$



Sort of a rough picture. the idea is there though.

$$
\begin{gathered}
|\bar{X}_n - \mu| < k\sigma - k\sigma < \bar{X}_n - \mu < k\sigma \\
\mu - k\sigma < \bar{X}_n < \mu + k\sigma
\end{gathered}
$$

(c) Using a basic fact about $\Phi(x)$ and your sketch in part (b), show that under the Normality assumption for the X_i,

$$P(|\bar{X}_n - \mu| < k\,\sigma) \geq 2\,\Phi(k\,\sqrt{n}) - 1\,. \tag{4}$$

$$
\begin{gathered}
P(|\bar{X}_n - \mu| < k\sigma) = P(-k\sigma < |\bar{X}_n - \mu| < k\sigma) = P(-k\sigma < Z' < k\sigma) \\
\geq P(Z' \leq k\sigma) - P(Z' \leq -k\sigma) \\
\geq P(Z' \leq k\sigma) - (1 - P(Z' \leq -k\sigma)) \\
\geq 2P(Z' \leq k\sigma) - 1 \\
\geq 2P(\frac{Z'}{\sigma/\sqrt{n}}) \leq (\frac{k\sigma}{\sigma/\sqrt{n}}) - 1 \\
\geq 2P(\frac{Z'}{\sigma/\sqrt{n}}) \leq k\sqrt{n} - 1 \\
P(|\bar{X}_n - \mu| < k\sigma) \geq 2P(Z \leq k\sqrt{n}) - 1 \\
P(|\bar{X}_n - \mu| < k\sigma) \geq 2\Phi(k\sqrt{n}) - 1
\end{gathered}
$$

(d) Set the probability on the right-hand side of the inequality in (4) equal to $(1 - \alpha)$ and solve for n, thereby showing that under the Normality assumption the required sample size corresponding to $n_{Chebyshev}$ in equation (3) above is

$$n_{Normality} \geq \frac{\left[\Phi^{-1}\left(1 - \frac{\alpha}{2}\right)\right]^2}{k^2}\,, \tag{5}$$

in which (as usual) $\Phi^{-1}(p)$ (for $0 < p < 1$) is the inverse CDF (quantile function) for the standard Normal distribution.

Let Φ^{-1} be the inverse CDF of N(0,1). Setting the probability to $(1 - \alpha)$ gives us:

$$2\Phi(k\sqrt{n}) - 1 \geq 1 - \alpha$$

$$\Phi(k\sqrt{n}) \geq 1 - \frac{\alpha}{2}$$

$$\sqrt{n} \geq \frac{\Phi^{-1}(1 - \frac{\alpha}{2})}{k}$$

$$n_{normal} \geq (\frac{\Phi^{-1}(1 - \frac{\sigma}{2})}{k})^2$$

(e) Using the table on page 861 of Degroot and Schervish or (better) an online inverse Normal CDF calculator (e.g., there's one provided by `Wolfram Alpha`), complete the rest of the table below.

α	$\frac{1}{\alpha}$	$\left[\Phi^{-1}\left(1 - \frac{\alpha}{2}\right)\right]^2$
0.1	10	2.7
0.05	20	**3.88**
0.01	**100**	6.6
0.005	**200**	**7.88**
0.001	1000	**10.83**

(f) If the data values X_1 really did come from a Normal distribution, would you describe the Chebyshev Inequality sample size calculation as highly conservative, not too conservative, or in between? Explain briefly.

If the distribution of the samples is normalized, then it's best to use conditions that come from the properties of normal distribution to determine sample size. Using Chebyshev's Inequality will give you **highly conservative** results because it does not use the distribution information of the parent population.