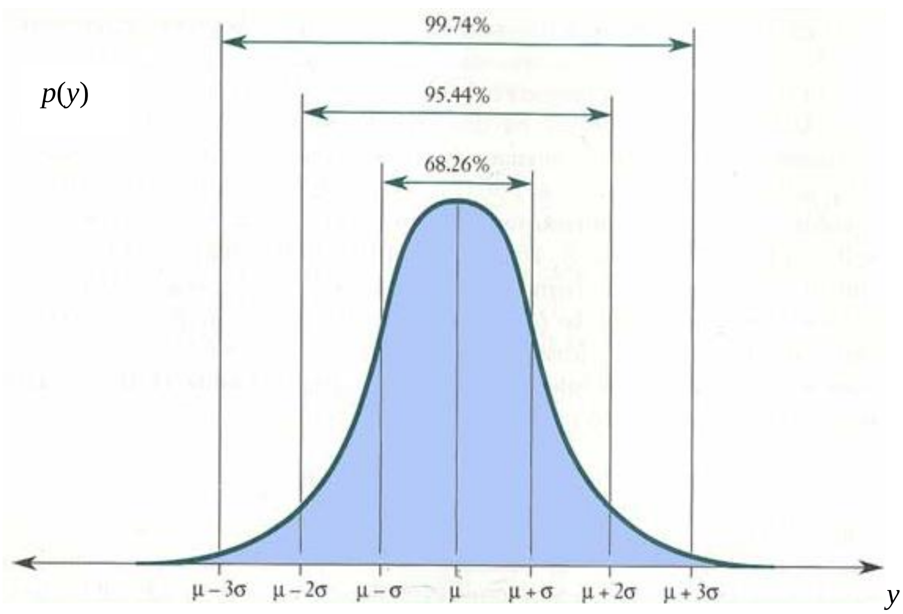
ISQS 5347 Midterm 2 Sp 2013

Closed book, notes, and no electronic devices.

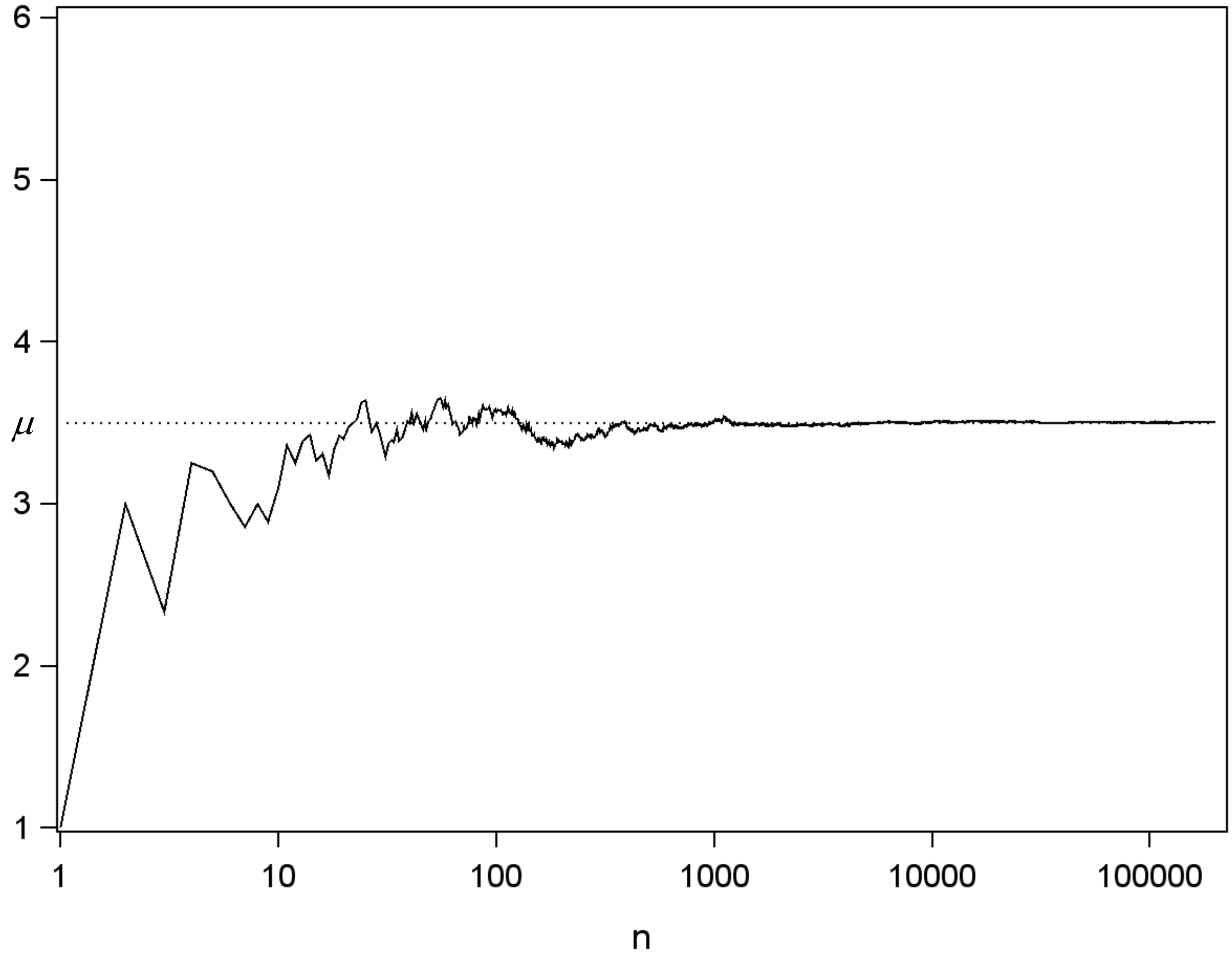
1. (10) Draw a graph of the normal probability distribution function with mean μ and standard deviation σ. Show on the graph where μ and σ appear.

Solution:



2. (10) Suppose $Y_1, Y_2, Y_3, \ldots \sim_{\text{iid}} p(y)$, with $E(Y_1) = \mu$ (finite). Consider the successive averages $Ave_1 = Y_1/1$, $Ave_2 = (Y_1 + Y_2)/2$, $Ave_3 = (Y_1 + Y_2 + Y_3)/3, \ldots$ Draw a graph with the values of Ave_n on the vertical axis and n on the horizontal axis, where n ranges from 1 to 100,000. Show on the graph where μ appears.

Solution:



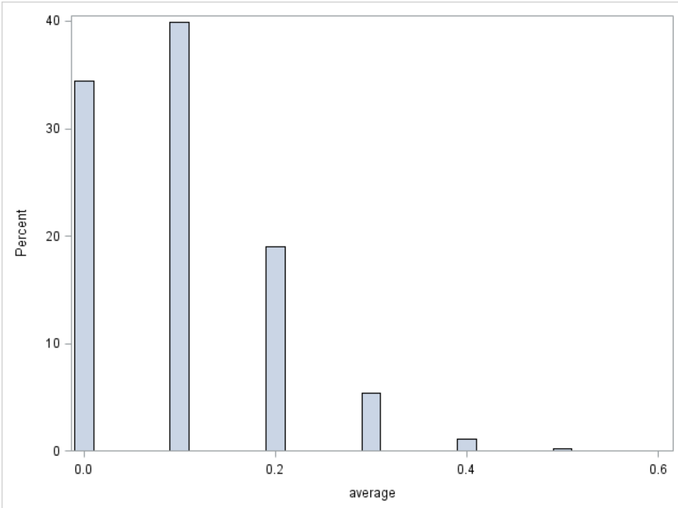
3. (20) Explain, step by step, how you would demonstrate the Central Limit Theorem to someone who does not know what it is, by using simulation. You can assume this person knows how to simulate data using random number generators, so you don't have to explain what simulation means in your answer.

Solution:

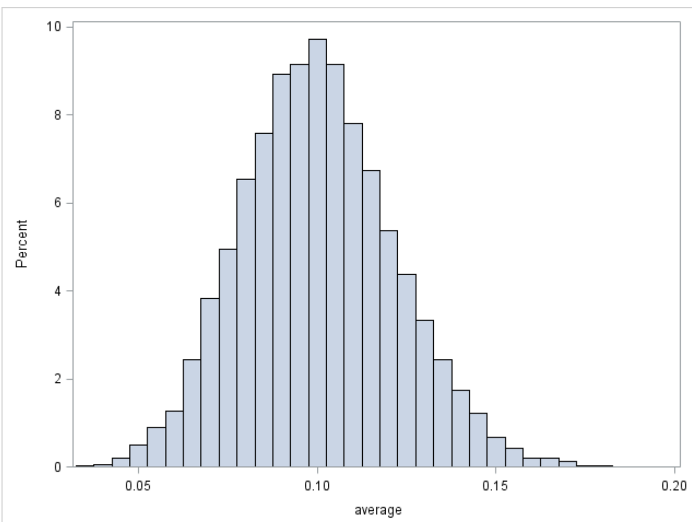
(1) Pick a non-normal distribution, say Bernoulli with p=0.1.

(2) Simulate n of these Bernoulli values, call them Y_1^*, Y_2^*, …, Y_n^*. To start, pick a small number for n, like n = 10. Calculate the average of these n values; call it Ave_1^*.

(3) Repeat (2) many times, say 10,000, getting Ave_1^*, Ave_2^*, …, $\text{Ave}_{10,000}^*$. Draw the histogram of these 10,000 numbers. Note that with n = 10, the distribution looks pretty skewed and discrete – not normal. In case you are curious, here it is for this particular distribution:



(4) Repeat (2) and (3), but using a larger *n*, say *n* = 200. Note that with *n* = 200, the distribution of the average is closer to normal. In case you are curious, here's what happens:



4. (16) Explain briefly the meanings of the lines below labeled a, b, c, and d.

```
data work.mt2;    /* a */
  input  George Barb;
  datalines;
  1 3
  4 3
  4 5
  4 4
  3 4
;
```

```
proc univariate data=mt2;   /* b */
   var George;               /* c */
   histogram George;         /* d */
run;
```

Solution:
/* a */ Opens a data set called "mt2" in the "work" library.
/* b */ Tells SAS to run the "univariate" data analysis procedure using the data set work.mt2. (Since the library is excluded, it picks the data set in the "work" library by default.)
/* c */ Tells PROC UNIVARIATE to analyze just the data in the first column of the data set; the first column values are called "George".
/ * d */ Tells PROC UNIVARIATE to draw a histogram of the data in the first column of the data set; the first column values are called "George".

5. (10) Suppose Y_1, Y_2, … $\sim_{\text{iid}}$ $p(y)$, with E(Y_1) = μ. Explain why Y_2 is an unbiased estimator of μ, using the definitions of both "iid" and "unbiased" in your answer.

   Solution: Since the random variables are all identically distributed, they all have the same expected value; namely, E(Y_1) = $\mu = \int y\, p(y)\, dy$ (assuming a continuous $p(y)$). Hence E(Y_2) = μ as well, which implies by definition that Y_2 is an unbiased estimator of μ. (Note that independence is not needed here).

6. (14) Suppose that $\hat{\theta}_1$ and $\hat{\theta}_2$ are both unbiased estimators of θ. Show, step by step, with justifications for each step, that $(\hat{\theta}_1 + \hat{\theta}_2)/2$ is also an unbiased estimator of θ.

Solution: Let's calculate E{ $(\hat{\theta}_1 + \hat{\theta}_2)/2$} and see what we get:

E{ $(\hat{\theta}_1 + \hat{\theta}_2)/2$}

= (1/2) E{ $(\hat{\theta}_1 + \hat{\theta}_2)$} (by the linearity property of expectation)

= (1/2){ E$(\hat{\theta}_1)$ + E{$(\hat{\theta}_2)$} (by the additivity property of expectation)

= (1/2){ θ + θ} (since $\hat{\theta}_1$ and $\hat{\theta}_2$ are both unbiased estimators of θ)

= θ (by algebra).

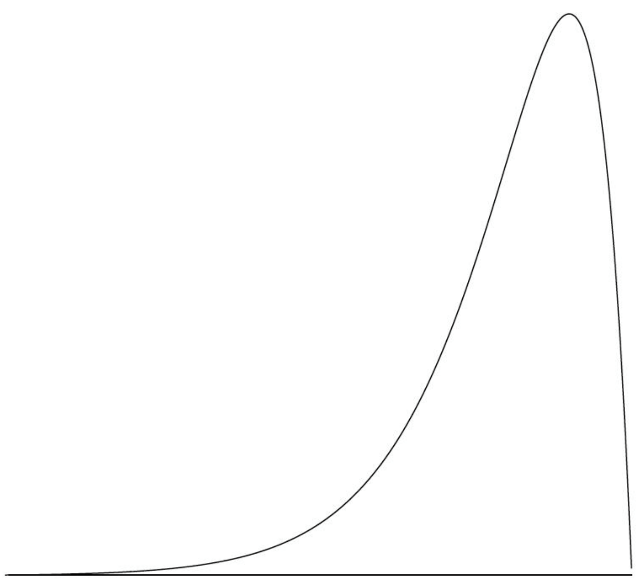
Hence, E{ $(\hat{\theta}_1 + \hat{\theta}_2)/2$} = θ, which implies that $(\hat{\theta}_1 + \hat{\theta}_2)/2$ is an unbiased estimator of θ.

Multiple choice questions. (2 points each)

7. The function $f(y) = y^2$ is
   A. convex B. concave C. linear

Solution: $f''(y) = 2 > 0$; hence it is convex.

8. Where are the most extreme outliers that are produced by a distribution that has negative skewness?
   A. Above the mean B. Below the mean C. Equally on either side of the mean



9. A non-normal distribution has mean 10 and standard deviation 1. What percent of the data that it produces will be between 7 and 13?
   A. 68% B. 95% C. 99.7% D. At least 88.9%

Solution: Since it is a non-normal distribution, you can't apply the 68-95-99.7 rule, which would give 99.7. Chebychev states > $1 - 1/3^2$, or > 88.9%.

10. The data set is 2, 2, 2, 2, 7. What is the bootstrap plug-in estimate of the mean?
    A. 3.0 B. 4.5 C. 1.8 D. It's an unknown parameter

Solution: It's just the ordinary sample average, 3.0.

11. Which is the law of the unconscious statistician?
    A. $E\{f(Y)\} = f\{E(Y)\}$ B. $E\{f(Y)\} = \int f(y)dy$ C. $E\{f(Y)\} = \int f(y)p(y)dy$

Solution: A is obviously wrong because of Jensen's inequality. B. is wrong because expected value has to consider the distribution.

12. Nature favors ________________________
    A. continuity over discontinuity B. discontinuity over continuity C. a vacuum

Solution. Hmmm… C sounds right. Oh no, that's "abhors" not "favors"! Gotta be A.

13. It is most logical to interpret the model $p(y)$ that you assume to produce your data as a
    A. population model B. bootstrap model C. process model D. regression model

Solution: In general, the data are not produced by a population, even when randomly sampled from a population, because there are design and measurement elements such as nonresponse and measurement error, which are not part of the population at all, which affect how the data will appear. The bootstrap is obviously wrong because it states that your data set define everything that could possibly be observed. Regression models are usually pretty logical, except then it would be called $p(y \mid x)$, not $p(y)$. But you knew it was C, even without those explanations, didn't you?

14. If the distribution of Y_t when $Y_{t-1} > 10$ is quite different from the distribution of Y_t when $Y_{t-1} \le 10$, then the data $Y_1, Y_2, \ldots$ are
    A. iid B. dependent C. not identically distributed

Solution: The given statement describes two conditional distributions that differ. This is the definition of dependence. The "identical distribution" part of "iid" refers to marginal distributions, not conditional distributions.

15. From the population perspective, the conditional distributions $p(y \mid x)$
    A. always exist B. sometimes exist C. never exist

Solution: As long as there is that particular value of x in the population (e.g., $x = 2.3$), then the population version exists. As shown in the book, it might not make any sense whatsoever, but it does exist. But if that particular value of x (e.g. $x = 2.3$) is not in the population, then the distribution does not exist at all.

16. Suppose $Y \sim N(10, 2)$. Then $E(Y) =$ ___________.
    A. 2 B. 4 C. 6 D. 8 E. 10

Solution: E