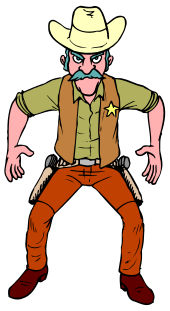
CoSMo Exercises Weds01

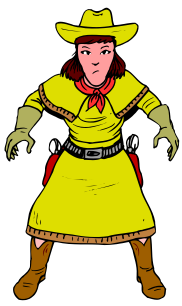
1. **Cue combination.** Paul is windsurfing parallel to the beach. Suppose that he has two depth cues to the distance to the beach. Both are distributed as independent Gaussian random variables.

a. Assume the cues are unbiased. The variance of the first is 4, that of the second is 1. What weights w and 1-w would an optimal observer give to the two cues to get an unbiased minimum variance estimate? What would the variance of his resulting estimate be?
b. Suppose that Paul just averages the unbiased two cues (giving them equal weight). What is the variance of his estimate? How much worse is it than the variance of the minimum variance unbiased estimator?
c. Compared to the "average cue" in 1c would Paul be better off in just using the lower variance cue and discarding the other?
d. For what values of $0 < w < 1$ is the combined cue lower variance than the lower variance single cue? Worse?
e. Suppose you knew nothing about the variances of the cues except that the variance of the second cue was lower than that of the first and the ratio of the variances was less than 4 to 1. You could just use the lower variance cue…. or would it make sense to just guess a value for w and combine the cues anyway?

2. **Dueling Estimators**. In this exercise, we consider two estimators of the parameter μ in the Normal$\left(\mu,\sigma^2\right)$. The first was the *mean* of a sample from the distribution and the second was the *median*. You have probably heard that the mean is good and the median is bad. In this exercise we investigate what happens to the mean or median as we change distribution.
For convenience, we set $\mu = 0$ and $\sigma^2 = 1$ and we see how well each of the estimators does at estimating the true value $\mu = 0$ from a sample.



2A. Generate 10000 samples, each of size 10, from the Normal(0,1) distribution (as a 10x10000 matrix). Compute the means of each of the 10000 samples. Call this vector **est1**. Histogram **est1** (use axis to set the horizontal axis limits to (-2,2) and the vertical axis to convenient values; use about 50 'bars').



2B. You expect the histogram in 1A to look Normal (Explain why). What is the standard deviation of the means (**std(est1)**)? You should be able to calculate theoretically what the standard deviation of the average of a sample with size 10 from Normal(0,1) should be. What is it? How does your calculation compare to **std(est1)**?

2C. Once again, generate 10000 samples, each of size 10, from the Normal(0,1) distribution (as a 10x1000 matrix). Compute the median of each of the 10000 samples. Call this vector **est2**. Histogram **est2** with the

same axis limits (and with 50 or more bars). You have no reason to expect the distribution to look Normal. Does it? Try a **normplot** to check your intuition and comment on what you find.

2D. These estimators are both unbiased**.** Our criterion for goodness of estimators will be their standard deviation. A good estimator hovers around its expected value, a bad estimator is all over the place.
What is the standard deviation of the medians (**std(est2)**)? Which has a higher standard deviation, the mean or the median? Can you see it in the histograms? Compute the ratio of the estimate of the standard deviation of the mean to the estimate of the standard deviation of the median (you could use the theoretical value of the standard deviation of the mean instead of the estimate if you like).

2E. (harder) You know that if your sample size were smaller or larger than 10, then the standard deviation of the mean would be smaller or larger. Adjust the sample size so that the standard deviation of the mean with the new sample size is close to the standard deviation of the median with sample size 10. Hint: square-root-of-N.
Moral: You can collect fewer data points with the better estimator and get the same performance.

3. Robustness. If you try distributions other than the Normal, though, the estimator you picked in 1E may not always be the one with the lower standard deviation. Redo Problem 2 with the t-distribution with 3 degrees of freedom which you can get through the function **trnd(3,M,N).** Report your results.

Moral: If you don't know the distribution of your data, then you might worry that your 'really good' estimator is really not good at all! This topic in statistics is called 'distributional robustness' You might wonder whether people use the mean or the median to estimate the center of a cluster of dots on a computer screen ….