

7 7818 Interval estimation and hypothesis testing - Set

revised Nov 29, 2010

You might want to read some of the chapter in MGB on Parametric Interval Estimation.

There are subtle differences across questions. Make sure to look for and understand those differences.

1. Assume the random variable X is normally distributed with unknown mean μ and variance 16; that is $f_X(x : \mu) = \phi_{\mu,16}(x)$. Let X_1, X_2, \dots, X_n be a random sample from this population. Determine $\Pr[\bar{X} - 3 < \mu < \bar{X} + 3]$.

answer: we know that $Z = \frac{\bar{X} - \mu}{\sigma/\sqrt{n}}$ is normally distributed with mean zero and variance of one, so has no parameters. In this case, $Z = \frac{\bar{X} - \mu}{4/\sqrt{n}}$.¹ So, in this case

$$\begin{aligned}
 \Pr[\bar{X} - 3 < \mu < \bar{X} + 3] &= \Pr[(\bar{X} - \mu) - 3] < 0 < (\bar{X} - \mu) + 3] \\
 &= \Pr\left[\frac{(\bar{X} - \mu)}{4/\sqrt{n}} - \frac{3}{4/\sqrt{n}} < 0 < \frac{(\bar{X} - \mu)}{4/\sqrt{n}} + \frac{3}{4/\sqrt{n}}\right] \\
 &= \Pr\left[-\frac{3}{4/\sqrt{n}} < -\frac{(\bar{X} - \mu)}{4/\sqrt{n}} < +\frac{3}{4/\sqrt{n}}\right] \\
 &= \Pr\left[\frac{3}{4/\sqrt{n}} > \frac{(\bar{X} - \mu)}{4/\sqrt{n}} > -\frac{3}{4/\sqrt{n}}\right] \\
 &= \Pr\left[-\frac{3}{4/\sqrt{n}} < Z < +\frac{3}{4/\sqrt{n}}\right]
 \end{aligned}$$

This probability depends on the sample size, but not on μ . For any sample size, we could look up the answer in the standard-normal table. For example, if $n = 9$, $\Pr[-\frac{3}{4/\sqrt{9}} < Z < +\frac{3}{4/\sqrt{9}}] = \Pr[-\frac{3}{4/\sqrt{9}} < Z < +\frac{3}{4/\sqrt{9}}] = \Pr[-2.25 < Z < 2.25] = \text{NormalDist}(2.25) - \text{NormalDist}(-2.25) = 0.97555$. What does this mean? 97.555% of the random intervals $(\bar{X} - 3)$ to $(\bar{X} + 3)$ will contain μ , giving us confidence that μ lies in this interval. Note that this interval varies with \bar{X} , so is random.

If $n = 100$, $\Pr[-\frac{3}{4/\sqrt{100}} < Z < +\frac{3}{4/\sqrt{100}}] = \Pr[-\frac{3}{4/\sqrt{100}} < Z < +\frac{3}{4/\sqrt{100}}] = \Pr[-7.5 < Z < 7.5] = \text{NormalDist}(7.5) - \text{NormalDist}(-7.5) = 1.0000$ and effectively 100% of these random intervals will contain μ .

¹While μ appears in Z , Z is unit normal for all values of μ . Further note the significance of knowing σ .

2. Assume the random variable X is normally distributed with unknown mean μ and variance 16; that is $f_X(x : \mu) = \phi_{\mu,16}(x)$. Let X_1, X_2, \dots, X_n be a random sample from this population. Determine $\Pr[\bar{X} - 3 < \mu < \bar{X} - 1]$.

answer: we know that $Z = \frac{\bar{X} - \mu}{\sigma/\sqrt{n}}$ is normally distributed with mean zero and variance of one, and has no parameters. In this case, $Z = \frac{\bar{X} - \mu}{4/\sqrt{n}}$. So, in this case

$$\begin{aligned} \Pr[\bar{X} - 3 < \mu < \bar{X} - 1] &= \Pr[(\bar{X} - \mu) - 3 < 0 < (\bar{X} - \mu) - 1] \\ &= \Pr\left[\frac{(\bar{X} - \mu)}{4/\sqrt{n}} - \frac{3}{4/\sqrt{n}} < 0 < \frac{(\bar{X} - \mu)}{4/\sqrt{n}} - \frac{1}{4/\sqrt{n}}\right] \\ &= \Pr\left[-\frac{3}{4/\sqrt{n}} < -\frac{(\bar{X} - \mu)}{4/\sqrt{n}} < -\frac{1}{4/\sqrt{n}}\right] \\ &= \Pr\left[\frac{1}{4/\sqrt{n}} < Z < +\frac{3}{4/\sqrt{n}}\right] \end{aligned}$$

This probability depends on the sample size, but not on μ . For any sample size, we could look up the answer in the standard-normal table. For example, if $n = 9$, $\Pr[\frac{1}{4/\sqrt{9}} < Z < +\frac{3}{4/\sqrt{9}}] = \Pr[.75 < Z < +2.25] = \text{NormalDist}(2.25) - \text{NormalDist}(.75) = 0.2144$. What does this mean? 21.44% of the random intervals $(\bar{X} - 3)$ to $(\bar{X} - 1)$ will contain μ . In this problem, we have derived a confidence interval on a parameter, μ .

3. Assume the random variable X is normally distributed with unknown mean μ and variance 16; that is $f_X(x : \mu) = \phi_{\mu,16}(x)$. Let X_1, X_2, \dots, X_n be a random sample from this population. Specify three different .4 confidence intervals for μ in terms of the sample average, \bar{X} .

answer: we know that $Z = \frac{\bar{X} - \mu}{\sigma/\sqrt{n}}$ is normally distributed with mean zero and variance of one, and has no parameters. In this case, $Z = \frac{\bar{X} - \mu}{4/\sqrt{n}}$.

$\text{NormalDist}(Z) - \text{NormalDist}(-Z) = .4$, Solution is: $\{[Z = 0.5244]\}$. So, one range on Z to get 40% of the distribution is -0.5244 to 0.5244 , symmetrical around zero.

$\text{NormalDist}(Z) - \text{NormalDist}(0) = .4$, Solution is: $\{[Z = 1.2816]\}$, So, another range on Z to get 40% of the distribution is 0 to 1.2816

$\text{NormalDist}(Z) - \text{NormalDist}(.1) = .4$, Solution is: $\{[Z = 1.5533]\}$, , So, another range on Z to get 40% of the distribution is 0.1 to 1.5533

Note that these three ranges on Z are of increasing length, and only the first one is symmetric around zero. Now we need to convert these ranges

in terms of Z into ranges in terms of \bar{X} . Start with the first.

$$\begin{aligned}
 .4 &= \Pr[-.5244 < Z < .5244] \\
 &= \Pr[-.5244 < \frac{\bar{X} - \mu}{4/\sqrt{n}} < .5244] \\
 &= \Pr[-.5244(4/\sqrt{n}) < \bar{X} - \mu < .5244(4/\sqrt{n})] \\
 &= \Pr[-\bar{X} - .5244(4/\sqrt{n}) < -\mu < -\bar{X} + .5244(4/\sqrt{n})] \\
 &= \Pr[\bar{X} - .5244(4/\sqrt{n}) < \mu < \bar{X} + .5244(4/\sqrt{n})] \\
 &= \Pr[\bar{X} - 2.0976/\sqrt{n} < \mu < \bar{X} + 2.0976/\sqrt{n}]
 \end{aligned}$$

So, 40% of the random intervals $\bar{X} - 2.0976/\sqrt{n}$ to $\bar{X} + 2.0976/\sqrt{n}$ will include μ . Now consider the second 40% range on Z

$$\begin{aligned}
 .4 &= \Pr[0 < Z < 1.2816] \\
 &= \Pr[0 < \frac{\bar{X} - \mu}{4/\sqrt{n}} < 1.2816] \\
 &= \Pr[0 < \bar{X} - \mu < 1.2816(4/\sqrt{n})] \\
 &= \Pr[-\bar{X} < -\mu < -\bar{X} + 1.2816(4/\sqrt{n})] \\
 &= \Pr[\bar{X} - 1.2816(4/\sqrt{n}) < \mu < \bar{X}] \\
 &= \Pr[\bar{X} - 5.1264/\sqrt{n} < \mu < \bar{X}]
 \end{aligned}$$

So, 40% of the random intervals $\bar{X} - 5.1264/\sqrt{n}$ to \bar{X} will include μ . Now consider the third 40% range on Z

$$\begin{aligned}
 .4 &= \Pr[-.1 < Z < 1.5533] \\
 &= \Pr[-.1 < \frac{\bar{X} - \mu}{4/\sqrt{n}} < 1.5533] \\
 &= \Pr[-.1(4/\sqrt{n}) < \bar{X} - \mu < 1.5533(4/\sqrt{n})] \\
 &= \Pr[-\bar{X} - .1(4/\sqrt{n}) < -\mu < -\bar{X} + 1.5533(4/\sqrt{n})] \\
 &= \Pr[\bar{X} - 1.5533(4/\sqrt{n}) < \mu < \bar{X} + .1(4/\sqrt{n})] \\
 &= \Pr[\bar{X} - 6.2132/\sqrt{n} < \mu < \bar{X} + .4/\sqrt{n}]
 \end{aligned}$$

So, 40% of the random intervals $\bar{X} - 6.2132/\sqrt{n}$ to $\bar{X} + .4/\sqrt{n}$ will include μ .

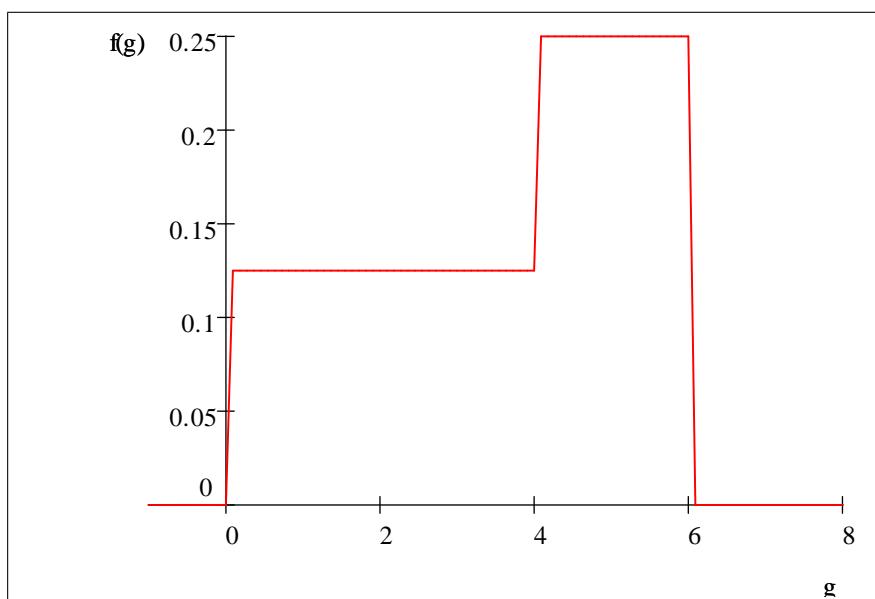
Note that in this problem we have identified three different .4 confidence interval on the parameter, u .

4. Let G represent gubers, where G is a random variable with

$$f_G(g) = \begin{cases} .125 & \text{if } 0 \leq g \leq 4 \\ .25 & \text{if } 4 < g \leq 6 \\ 0 & \text{otherwise} \end{cases}$$

Charlotte tells you that she has randomly sampled one G from this distribution, and its realized value is 7. How likely is it that she is either mistaken or lying? Jesse tells you that he has randomly sampled one G from this distribution, and its realized value is 5.6. How likely is it one will get a random draw from this distribution that is greater than or equal to 5.6 (what is the probability of drawing a $g \geq 5.6$ if the true distribution is $f_G(g)$)? Explain.

answer: The 7 could not have come from this population. The 5.6 could have. Picture $f_G(g)$



50% of the density is in the range $4 < g \leq 6$. Break this up into 5 equal ranges with 10% in each: $4 < g \leq 4.4$, $4.4 \leq g \leq 4.8$, $4.8 \leq g \leq 5.2$, $5.2 \leq g \leq 5.6$, and $5.6 \leq g \leq 6.0$. So there is a 10% chance of randomly drawing a $g \geq 5.6$ from $f_G(g)$. Said the other way, there is a 90% chance that the drawn g is between 0 and 5.6.

If one set her level of significance at .10 one would reject, just, the null hypothesis that 5.6 is a random draw from $f_G(g)$. If one set her significance level at .05, one would not reject the null hypothesis.

5. Consider again, the old issue of how many time one gets married. Assume again that the number of times one gets married, M , has a Poisson distribution. Further assume you have randomly sampled one member of the population, Ralph, and that he has been married one time. Find a $x\%$ confidence interval for μ . You choose x .

answer: So what do we know? The Poisson density is $f_M(m) = \frac{e^{-\mu}\mu^m}{m!}$ for integer values of $m = 0, 1, 2, \dots$. With a sample of one individual married once, the maximum likelihood estimate of μ is 1. And the estimated density function for M is $f_M(m) = \frac{e^{-1}1^m}{m!} = \frac{1}{m!}e^{-1} = \Pr[M = m]$. Consider, what this means, the probability of never marrying is

$\text{PoissonDen}(0; 1) = e^{-1} = 0.36788$. Once is $\text{PoissonDen}(1; 1) = 0.36788$ and twice is $\text{PoissonDen}(2; 1) = 0.18394$

The estimated CDF is $\text{PoissonDist}(x; \mu) = \sum_{k=0}^x \frac{\mu^k e^{-\mu}}{k!}$. So, for example, the estimated probability of being married one time or less is $\text{PoissonDist}(1; 1) = 0.73576$.

Now that we have a feel for things, let's see how the probability of being married one time or less varies with μ . If, for example, our estimate of μ was 4, the estimated probability of being married one time or less is $\text{PoissonDist}(1; 4) = 9.1578 \times 10^{-2} = .092$, a little over 9%.

Now let's ask another question. What is the probability that one will be married one or more times as a function of μ . It is $1 - \text{PoissonDist}(1; \mu)$. For example the probability of being married one or more time if μ is .5 is $1 - \text{PoissonDist}(1; .5) = 9.0204 \times 10^{-2} = .09$, 9%.

Flipping these questions, what would the estimated μ have to be for the estimated probability of someone being married one or less times be .05? I kept playing around with values of μ in $\text{PoissonDen}(1; \mu)$ until I stumbled on $\text{PoissonDen}(1; 4.5) = 0.04999$. That is, if μ is 4.5 there is only a 5% chance that one will be married one or less times, pretty low.

Now let's figure out what μ would have to be for .05 to be the probability of being married one or more times. We want to figure what value of μ will make $1 - \text{PoissonDen}(1; \mu) = 0.05$. Trying different numbers I got $1 - \text{PoissonDist}(1; .36) = 0.05116$

So, what has been determined? A 90% confidence interval for μ is .36 to 4.5. Note that this interval is a rv; it varies from sample to sample, holding constant the number of observations at one. Interpreting, 90% of these estimated intervals will contain λ .

What else can we say about intervals. A 95% confidence interval for λ is $\geq .36$, another is $\lambda \leq 4.5$. A 5% confidence interval is $\lambda \geq 4.5$ another is $\lambda \leq .36$.

To make sure you understand, redo this problem assuming Ralph was married twice.

6. Redo the previous problem assuming the sample has two rather than one observation.
7. As $P[X \geq y]$ decreases it becomes less and less likely that y is a random draw from the population described by $f_X(x; \mu_x, \sigma_x^2)$. Explain to me why this is a reasonable way to decide whether y is a random draw from population described by $f_X(x; \mu_x, \sigma_x^2)$.

8. For fun, calculate the 95% and 99% confidence intervals for X for a few different specific specifications of $f_X(x; \mu_x, \sigma_x^2)$. For example, do it assuming X has a Poisson distribution with a mean of 3 (and a mean of 6), or assuming X has a t distribution with some specific parameter. Consider both one-sided and two-sided intervals. (Note that the Poisson is a discrete distribution, so the confidence interval will be a finite set of integers that span some range.) (Further note that in this problem we are deriving confidence intervals on X , not confidence intervals on a parameter.)
9. Grade point average in the Department's undergraduate math-econ course is continuously distributed between 0 and 4, where 0 is an F and 4 is an A . The distribution of grades in Weird Shirley's section is always symmetric and bowl shaped and we know its functional form. Put simply, Shirley gives lots of high grades, lots of low grades and few grades in between. Imagine that one randomly draws the grade for one undergraduate in one undergraduate economics course. Discuss in general terms how you would decide/test whether this grade was a grade in Weird Shirley's math-econ class. Discuss in general terms why and how what you would do differs from a case where Shirley's grades are normally distributed.

answer: The null hypothesis would be the grade is from Shirley's class. The more the grade is in the middle of the range, the greater reason for thinking the grade is not from Shirley's class. One might reject the null if the probability that the grade was given by Shirley is less than 5%. I would find the the GPA range for Shirley's classes such that 5% of the grades were in that range and the midpoint of that range was the average grade in Shirley's classes, μ . If the kid's grade was in this range, I would reject the null that he was in Shirley's class. Here the critical region is in the middle, we are accustomed to seeing the critical range at one or both ends of the distribution. In more detail: the confidence interval is two intervals: an interval to the right of the mean and extending to 4 and an interval to the left of the mean extending to 0. We want to choose these two intervals such that they are symmetric (the distribution is symmetric) and cover 95% of the density. And subject to these two constraints, the length of these two intervals together is minimized.

10. Assume X_1, X_2, \dots, X_n is a random sample from a normal distribution with unknown μ but known σ_x^2 . Find a 95% confidence interval for μ in terms of \bar{X} and interpret this interval.

answer: (See MGB page 374 and page 381) Consider the statistic $S = \frac{\bar{X} - \mu}{\frac{\sigma_x}{\sqrt{n}}}$, where $\frac{\sigma_x}{\sqrt{n}}$ is the standard deviation of \bar{X} . If X is normally distributed then $\frac{\bar{X} - \mu}{\frac{\sigma_x}{\sqrt{n}}}$ has a standard normal distribution, and

$$\Pr \left[-1.96 < \frac{\bar{X} - \mu}{\frac{\sigma_x}{\sqrt{n}}} < 1.96 \right] = .95$$

Rearranging (pivoting around μ)

$$\begin{aligned}
 .95 &= \Pr \left[-1.96 < \frac{\bar{X} - \mu}{\frac{\sigma_x}{\sqrt{n}}} < 1.96 \right] \\
 &= \Pr \left[-1.96 \frac{\sigma_x}{\sqrt{n}} < \bar{X} - \mu < 1.96 \frac{\sigma_x}{\sqrt{n}} \right] \\
 &= \Pr \left[-\bar{X} - 1.96 \frac{\sigma_x}{\sqrt{n}} < -\mu < -\bar{X} + 1.96 \frac{\sigma_x}{\sqrt{n}} \right] \\
 &= \Pr \left[\bar{X} + 1.96 \frac{\sigma_x}{\sqrt{n}} > \mu > \bar{X} - 1.96 \frac{\sigma_x}{\sqrt{n}} \right] \\
 &= \Pr \left[\bar{X} - 1.96 \frac{\sigma_x}{\sqrt{n}} < \mu < \bar{X} + 1.96 \frac{\sigma_x}{\sqrt{n}} \right]
 \end{aligned}$$

So, what does this mean? the interval $\bar{X} - 1.96 \frac{\sigma_x}{\sqrt{n}}$ to $\bar{X} + 1.96 \frac{\sigma_x}{\sqrt{n}}$ varies from sample to sample and 95% of these intervals will contain μ . For a given sample, $\bar{x} - 1.96 \frac{\sigma_x}{\sqrt{n}}$ to $\bar{x} + 1.96 \frac{\sigma_x}{\sqrt{n}}$ is a numerical range (e.g. 2 to 8) and it highly likely that μ will fall in this range, but not for sure. (As an aside, note the distinction between $\frac{\bar{X} - \mu}{\frac{\sigma_x}{\sqrt{n}}}$ and $\frac{X - \mu}{\sigma_x}$ and note that both have a standard normal distribution if X is normally distributed.)

11. Assume that X is normally distributed with known μ and known σ_x^2 . Determine and interpret the 95% confidence interval for X .

answer: If X is normally distributed then $\frac{X - \mu}{\sigma_x}$ has a standard-normal distribution. So

$$\Pr \left[-1.96 < \frac{X - \mu}{\sigma_x} < 1.96 \right] = .95$$

Pivoting this around X

$$\begin{aligned}
 .95 &= \Pr \left[-1.96 < \frac{X - \mu}{\sigma_x} < 1.96 \right] \\
 &= \Pr [-1.96 \sigma_x < X - \mu < 1.96 \sigma_x] \\
 &= \Pr [\mu - 1.96 \sigma_x < X < \mu + 1.96 \sigma_x]
 \end{aligned}$$

The range $\mu - 1.96 \sigma_x$ to $\mu + 1.96 \sigma_x$ is a specific numerical range determined by μ and σ^2 ; it does not vary from sample to sample. If one randomly draws an X from the population, there is a 95% chance that the drawn X , x , will fall in this range.

12. What's the difference between

$$.95 = \Pr [\mu - 1.96 \sigma_x < X < \mu + 1.96 \sigma_x]$$

and

$$.95 = \Pr \left[\mu - 1.96 \frac{\sigma_x}{\sqrt{n}} < \bar{X} < \mu + 1.96 \frac{\sigma_x}{\sqrt{n}} \right]$$

Why? As part of your answer, interpret both intervals.

answer: The range $\mu - 1.96\sigma_x$ to $\mu + 1.96\sigma_x$ is a specific numerical range determined by μ and σ^2 ; it does not vary from sample to sample. If one randomly draws an X from the population, there is a 95% chance the drawn X , x , will fall in this range. The interval $\mu - 1.96\frac{\sigma_x}{\sqrt{n}}$ to $\mu + 1.96\frac{\sigma_x}{\sqrt{n}}$ is a numerical range; it does not vary from sample to sample. How to interpret? If one draws a random sample of size n and takes its mean, there is a 95% chance that the observed mean will be in this range. The range on the latter is obviously smaller than the range on the former.

13. Assume that X is normally distributed with unknown μ and known σ_x^2 . Interpret

$$.95 = \Pr [X - 1.96\sigma_x < \mu < X + 1.96\sigma_x]$$

answer: Everytime one draws a new X one gets a different interval, and 95% of these intervals will contain the population mean, μ .

14. Assume X_1, X_2, \dots, X_5 is a random sample from a normal distribution with unknown μ and unknown σ_x^2 . Find a 95% confidence interval for μ in terms of \bar{X} and interpret this interval.

answer: Define the statistic $M = \frac{\bar{X} - \mu}{\frac{\hat{\sigma}_x}{\sqrt{n}}}$ where $\hat{\sigma}_x^2 = \frac{\sum_{i=1}^5 (X_i - \bar{X})^2}{(n-1)}$ is an estimate of σ_x^2 . The statistic M has a t distribution with 4 degrees of freedom ($n - 1$) – see MGB page ??? So,²

$$\begin{aligned} .95 &= \Pr \left[-2.776 < \frac{\bar{X} - \mu}{\frac{\hat{\sigma}_x}{\sqrt{n}}} < 2.776 \right] \\ &= \Pr \left[-2.776 \frac{\hat{\sigma}_x}{\sqrt{n}} < \bar{X} - \mu < 2.776 \frac{\hat{\sigma}_x}{\sqrt{n}} \right] \\ &= \Pr \left[-\bar{X} - 2.776 \frac{\hat{\sigma}_x}{\sqrt{n}} < -\mu < -\bar{X} + 2.776 \frac{\hat{\sigma}_x}{\sqrt{n}} \right] \\ &= \Pr \left[\bar{X} + 2.776 \frac{\hat{\sigma}_x}{\sqrt{n}} > \mu > \bar{X} - 2.776 \frac{\hat{\sigma}_x}{\sqrt{n}} \right] \\ &= \Pr \left[\bar{X} - 2.776 \frac{\hat{\sigma}_x}{\sqrt{n}} < \mu < \bar{X} + 2.776 \frac{\hat{\sigma}_x}{\sqrt{n}} \right] \end{aligned}$$

95% of the intervals, $\bar{X} - 2.776\frac{\hat{\sigma}_x}{\sqrt{n}}$ to $\bar{X} + 2.776\frac{\hat{\sigma}_x}{\sqrt{n}}$ will contain μ .

15. Assume X_1, X_2, \dots, X_5 is a random sample from a normal distribution with known μ and unknown σ^2 . Find a 95% confidence interval for σ^2 and interpret it.

²2.766 is the critical t value for .025 when parameter of the t is 4.

answer: Consider the statistic $Q = \frac{\sum_{i=1}^5 (X_i - \mu)^2}{\sigma_x^2} = \sum_{i=1}^5 \left(\frac{X_i - \mu}{\sigma}\right)^2$. This statistic has a Chi-squared distribution with parameter (degrees for freedom) n – see MGB page 243. Therefore

$$.95 = \Pr \left[.831 < \frac{\sum_{i=1}^5 (X_i - \mu)^2}{\sigma_x^2} < 12.8 \right]$$

In explanation, with 5 degrees of freedom, 2.5% of the values of Q are less than .831 and 2.5% are greater than 12.8. Note that all of the terms are positive. Rearranging

$$\begin{aligned} .95 &= \Pr \left[.831 < \frac{\sum_{i=1}^5 (X_i - \mu)^2}{\sigma_x^2} < 12.8 \right] \\ &= \Pr \left[\frac{1}{.831} > \frac{\sigma_x^2}{\sum_{i=1}^5 (X_i - \mu)^2} > \frac{1}{12.8} \right] \\ &= \Pr \left[1.2034 > \frac{\sigma_x^2}{\sum_{i=1}^5 (X_i - \mu)^2} > .078123 \right] \\ &= \Pr \left[1.2034 \sum_{i=1}^5 (X_i - \mu)^2 > \sigma_x^2 > .078123 \sum_{i=1}^5 (X_i - \mu)^2 \right] \\ &\quad \Pr \left[.078123 \sum_{i=1}^5 (X_i - \mu)^2 < \sigma_x^2 < 1.2034 \sum_{i=1}^5 (X_i - \mu)^2 \right] \end{aligned}$$

The interval $.078123 \sum_{i=1}^5 (X_i - \mu)^2$ to $1.2034 \sum_{i=1}^5 (X_i - \mu)^2$ varies across samples (each with 5 observations). 95% of these intervals will contain σ_x^2 . Note that the interval depends on μ , which was assumed known.

16. Assume X_1, X_2, \dots, X_5 is a random sample from a normal distribution with unknown μ and unknown σ^2 . Find a 95% confidence interval for σ^2 in terms of \bar{X} and interpret it.

answer: Consider the statistic $W = \frac{\sum_{i=1}^5 (X_i - \bar{X})^2}{\sigma_x^2} = \frac{(n-1)\hat{\sigma}_x^2}{\sigma_x^2}$ where $\hat{\sigma}_x^2 = \frac{\sum_{i=1}^5 (X_i - \bar{X})^2}{(n-1)}$ is an estimate of σ_x^2 . The statistic W has a Chi-square distribution with parameter (degrees of freedom) $n - 1$ – see MGB page ???. Therefore, for five observations

$$.95 = \Pr \left[.484 < \frac{(n-1)\hat{\sigma}_x^2}{\sigma_x^2} < 11.1 \right]$$

In explanation, with 5 observations, 4 degrees of freedom, 2.5% of the values of Q are less than .484 and 2.5% are greater than 11.1. Note that all of the terms are positive. Rearranging

$$\begin{aligned} .95 &= \Pr \left[.484 < \frac{(n-1)\hat{\sigma}_x^2}{\sigma_x^2} < 11.1 \right] \\ &= \Pr \left[\frac{1}{.484} > \frac{\sigma_x^2}{(n-1)\hat{\sigma}_x^2} > \frac{1}{11.1} \right] \\ &= \Pr \left[2.0661 > \frac{\sigma_x^2}{(n-1)\hat{\sigma}_x^2} > 0.09009 \right] \\ &= \Pr \left[2.0661(n-1)\hat{\sigma}_x^2 > \sigma_x^2 > 0.09009(n-1)\hat{\sigma}_x^2 \right] \\ &= \Pr \left[0.09009(n-1)\hat{\sigma}_x^2 < \sigma_x^2 < 2.0661(n-1)\hat{\sigma}_x^2 \right] \end{aligned}$$

The interval $0.09009(n-1)\hat{\sigma}_x^2$ to $2.0661(n-1)\hat{\sigma}_x^2$ varies across samples. 95% of these intervals will include σ_x^2 . For fun, compare this interval to $.078123 \sum_{i=1}^5 (X_i - \mu)^2$ to $1.2034 \sum_{i=1}^5 (X_i - \mu)^2$, which is the comparable interval when μ is known—derived in the previous question. The second one will be a tighter interval, because more is known.

17. So you meet a guy at a bar. He tells you that the random variable X has some density $f_X(x : \theta)$ but that he does not tell you its form or the value of θ . After another drink he tell that he has a derived an estimator for θ , $\hat{\theta} = \hat{\theta}(X_1, X_2, \dots, X_n)$ where X_1, X_2, \dots, X_n is a random sample. He does not tell you the form of the estimator but does tell you the sampling distribution of $\hat{\theta}$, $f_{\hat{\theta}}(\cdot : E[\hat{\theta}], \sigma_{\hat{\theta}}^2)$ is Guber distributed (the famous Guber distribution, named after its founder Philoneus Guber) with parameters 4 and 9.

You look in the Guber table in the back of some statistics book and determine a 95% confidence interval for $\hat{\theta}$ is -1 to 9 . In words, what does this confidence interval tell you?

Given the information provided, what do you know about θ ? Explain

answer: The confidence interval tell you that everytime you draw a random sample and calculate $\hat{\theta}$ it has a 95% chance of being between -1 and 9 . (95% of all samples will generate $\hat{\theta}$ that are between -1 and 9). The answer to the second question is *nothing*. We don't know anything about the guy's estimator, so don't know if it has any desirable properties. For example, if we knew it was an unbiased estimator, we would know that $\theta = 4$, but we don't know this.

18. Imagine that men come from both Mars and Pluto, and no where else. We know exactly how IQ (the intelligence coefficient) varies in each population. That is, the researcher knows $f_{IQ_m}(IQ_m, \theta_m)$ and $f_{IQ_p}(IQ_p, \theta_p)$. To be explicit, the two functional forms are known as well as the values of the two parameters (θ_m and θ_p). Further assume that the IQ of men from Mars varies from 50 to 130 ($50 \leq IQ_m \leq 130$) and the IQ of men from Pluto varies from 97 to 160 ($97 \leq IQ_m \leq 160$). One man is randomly sampled from one of the two populations and presented to you as a gift. You don't know whether he is from Mars or Pluto; all you know is that his IQ is 125. You are indifferent to where he is from; you just want to know.

(Part 1) Describe a rule or rules you might use to estimate whether he is from Mars or Pluto, where he is most likely to come from. Drawing and presenting some example overlapping density functions will clarify things for both you and the reader

answer to part 1: A simple rule, with a sound theoretical foundation, would be to compare the numerical values of $f_{IQ_m}(125, \theta_m)$ and $f_{IQ_p}(125, \theta_p)$, and choose the largest. For example, if $f_{IQ_m}(125, \theta_m) = .07 > .05 = f_{IQ_p}(125, \theta_p)$ estimate that he is from Mars. In more detail, $f_{IQ_m}(IQ_m, \theta_m)$ is the likelihood function for a sample of one if the population randomly sampled is Mars and, $f_{IQ_p}(IQ_p, \theta_p)$ is the likelihood function for a sample of one if the population randomly sampled is Pluto. Which population is more likely to generate an observation of 125, the one with the larger $f_{IQ_i}(125, \theta_i)$, $i = 1$ or 2 .

The first graph is an example where where you would estimate Mars. The second graph is an example where you would estimate Pluto.

There are more complicated estimation techniques that one might imagine. For example, one might choose planet i over j if $f_{IQ_i}(125, \theta_i) > f_{IQ_j}(125, \theta_j)$ by at least $x\%$, otherwise choose by flipping a fair coin. Or, one might just flip a fair coin (probably not the most efficient estimation technique).

(Part 2) Now make the problem more interesting by assuming two guy are randomly sampled from one of the two planets, What is your rule for choosing which planet they are from if one has an IQ of 104 and the other an IQ of 119?

answer to part 2: I would compare $f_{IQ_m}(104, \theta_m) \times f_{IQ_m}(119, \theta_m)$ with $f_{IQ_p}(104, \theta_p) \times f_{IQ_p}(119, \theta_p)$. Each is the likelihood of the sample give the planet.

some thoughts after reading your answers: $f_{IQ_m}(125, \theta_m)$, for example, is a number, a likelihood, but for a continuous distribution it is not a probability. Some of you turned it into a probability by assuming some small ε and then integrating $\int_{-\varepsilon}^{+\varepsilon} f_{IQ_m}(125 + \alpha, \theta_m) d\alpha$ to turn it into a probability. You then compared, for the first part of the question, $\int_{-\varepsilon}^{+\varepsilon} f_{IQ_m}(125 + \alpha, \theta_m) d\alpha \leq \int_{-\varepsilon}^{+\varepsilon} f_{IQ_p}(125 + \alpha, \theta_p) d\alpha$, and choose the planet with the larger one. This is fine, but the integration is unnecessary. This integration approach becomes a bit more complicated when one has two observations.

Many people wanted to answer the question with a confidence interval. This tack is bit problematic: Imagine assuming the null hypothesis is that the guys are from Mars, and then generating the 95% confidence interval for IQ_m , and easy task.³ One then finds one of two things: the observed IQ is or is not in the interval, or it is. What does one conclude in each case, not much. If the observed IQ is outside the interval one might be tempted to conclude the guy is from Pluto, but that would be premature - maybe you could also reject the null that the guy is from Pluto. There is no natural null hypothesis for this problem, unless you want him to be from Mars, or from Pluto. Alternatively, if the guy is in the interval for Mars he might also be in the interval for Pluto.

Many people imposed a normal, or some some other density on the two distributions. Doing so correctly is not answering the question but a restrictive form of the question, so deserving of only partial credit. As an aside, neither density function can be normal. Also note that there was nothing in the question that implied that the two density functions had the same functional form.

Hakon had an interesting conjecture that got me going for a bit. He felt we had to take account of the fact the the male populations on the two planets might differ substantively in size, so a person from the planet with the larger population was more likely to be drawn. Whether this is correct depends on how one samples. If one puts both populations in a big urn and randomly draws one individual, Hakon would be correct, but that is not what the question implies. It says, "One man is randomly sampled from one of the two populations.." that is, the chooser picks a planet, randomly or otherwise, and then randomly selects an individual

19. Denote a random sample of size n from the normal distribution, $f_X(x; \mu, \sigma_x^2) = \phi_X(\mu, \sigma^2)$ as X_1, X_2, \dots, X_n .

³As an aside a number of you thought that to do this one needed to first determine the variance of IQ_m . While OK, this is not necessary: one has the distribution, just find the the shortest span that captures 95% of the density.

Describe, in words, the interval $\bar{X} - 2.776 \frac{\hat{\sigma}_x}{\sqrt{n}} < \mu < \bar{X} + 2.776 \frac{\hat{\sigma}_x}{\sqrt{n}}$ where $\Pr \left[\bar{X} - 2.776 \frac{\hat{\sigma}_x}{\sqrt{n}} < \mu < \bar{X} + 2.776 \frac{\hat{\sigma}_x}{\sqrt{n}} \right] = .95$.

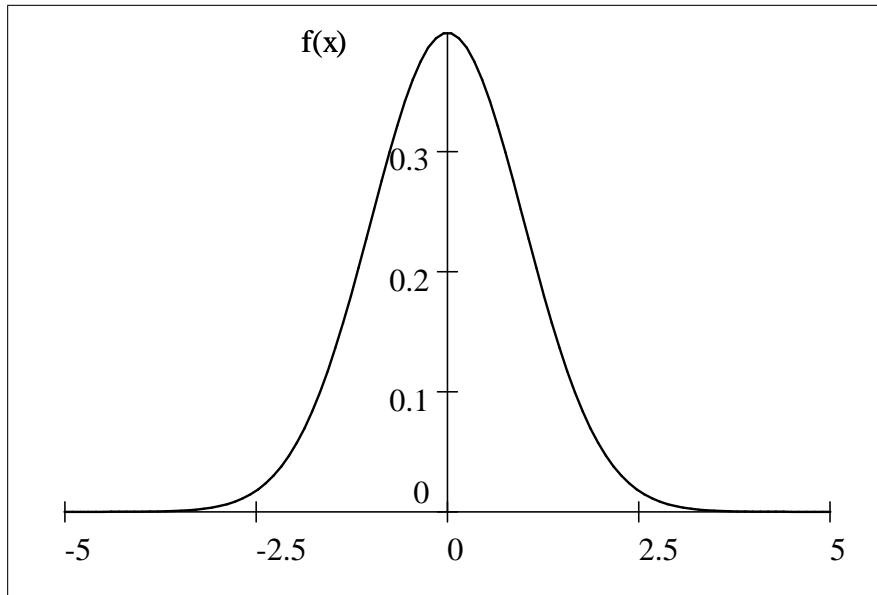
Now describe in words the interval $\mu - 1.96\sigma_x < X_i < \mu + 1.96\sigma_x$ where $\Pr [\mu - 1.96\sigma_x < X_i < \mu + 1.96\sigma_x] = .95$

answer: The interval $\bar{X} - 2.776 \frac{\hat{\sigma}_x}{\sqrt{n}} < \mu < \bar{X} + 2.776 \frac{\hat{\sigma}_x}{\sqrt{n}}$ varies from sample to sample and 95% of these intervals will contain μ . For example, for a specific sample, this interval is a specific numerical range (e.g. 2 to 8) and it highly likely that μ will fall in this range, but not for sure.

The interval $\mu - 1.96\sigma_x < X_i < \mu + 1.96\sigma_x$ is not a rv, but rather a fixed interval that depends on the values of μ and σ_x , it does not vary from sample to sample. X_i is the i^{th} draw from a sample. The i^{th} draw in 95% of the samples will be in this fixed interval. So, for this interval, but not the first interval, one can say that "there is a 95% chance that the i^{th} draw from a random sample will be in this interval.

20. Assume you are teaching econometrics. For a long time, I thought that for a specific random variable, X with density $f_X(x)$, there was only one 50% confidence interval. Explain, in a way your students would understand (try not to be too technical) why there can be many 50% confidence intervals. Graphs might help. Now that your students understand that there is likely more than one 50% confidence interval, explain how one might choose a specific 50% confidence interval. Is the 'best' confidence interval always continuous (no gaps)? Is it always symmetric around the mean? Explain.

answer: Assume some random variable X has some density function $f_X(x)$. Because it is a density function the area under it, by definition, is one. A 50% confidence interval is any range of values of X such that the area under that range is .50. There are often an infinite number of ranges that have this property. Consider for example a 50% confidence interval for the standard normal. `NormalDen(x; 0, 1)`



In my software the command $\text{NormalDist}(x)$ is the CDF of the standard normal. E.g. $\text{NormalDist}(0) = \frac{1}{2}$. So, one can express a 50% confidence interval as

$$\text{NormalDist}(x) - \text{NormalDist}(y) = .50 \text{ where } x > y$$

For example $(\text{NormalDist}(10) - \text{NormalDist}(0)) = 0.5$, so 0 to 10 is a 50% confidence interval for X (50% of randomly drawn X 's will be in this range)

$(\text{NormalDist}(0) - \text{NormalDist}(-10)) = 0.5$, so -10 to 0 is a 50% confidence interval

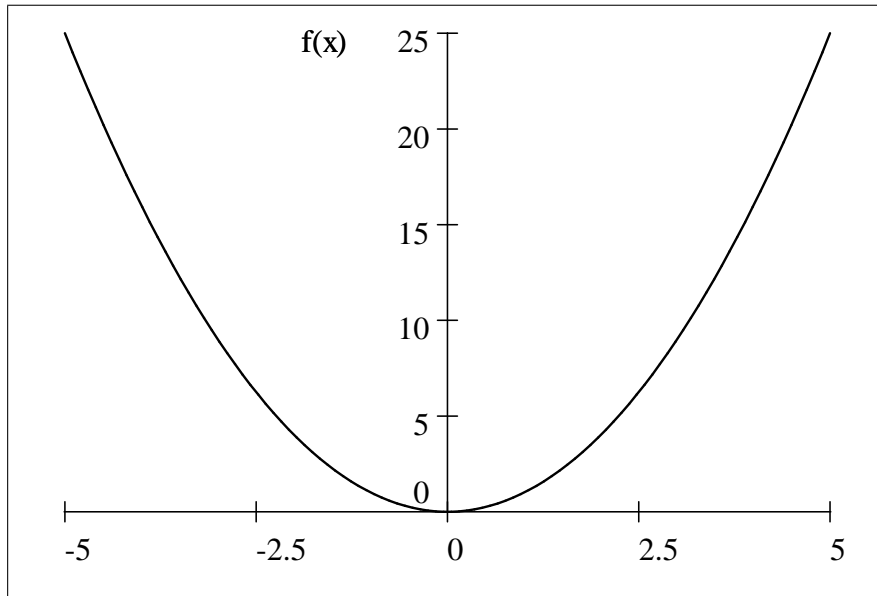
$(\text{NormalDist}(.69) - \text{NormalDist}(-.69)) = 0.50981$, so $-.69$ to $.69$ is a 50% confidence interval.

$(\text{NormalDist}(.5) - \text{NormalDist}(-.9)) = 0.5074$, so $-.9$ to $.5$ is a 50% confidence interval.

So, at this point, I am hopefully convinced that there can be more than one $m\%$ confidence intervals.

Which one to choose? Typically, the shortest one. Why? Because the shortest one is the one that provides the most accurate information about where a drawn X is most likely to fall. For example, in our above example, one could say that there is a 50% chance that a randomly drawn X will fall between zero and 10, but one can also say there is a 50% chance that a randomly drawn X will fall between $-.69$ and $.69$. The first interval covers a range of 10 the second interval a range of 1.38. The shorter one provides much more accurate information about where the X 's in this distribution "lie".

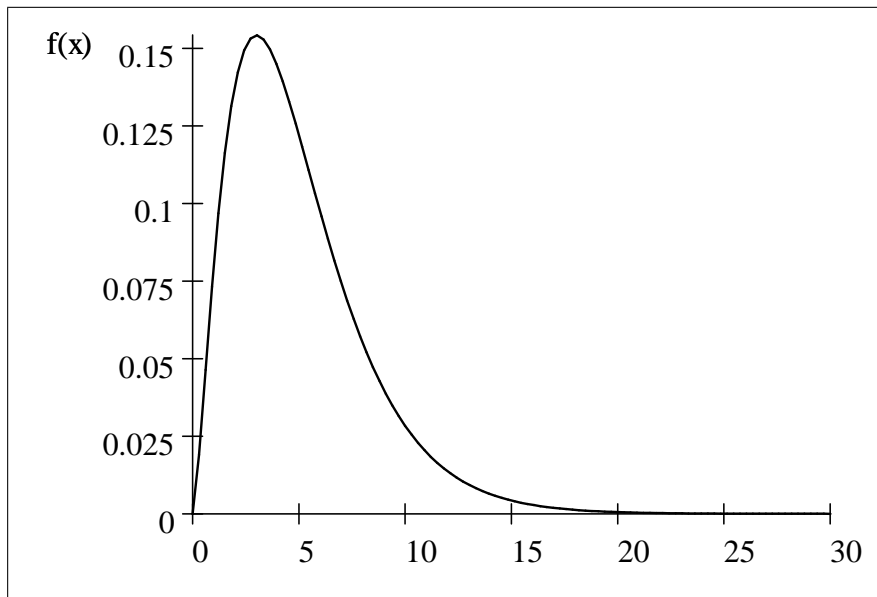
Note the best interval (the shortest one) will not always be a continuous interval. Consider for example the shortest 50% confidence interval for a density function that is *U*-shaped. For example. $f_X(x) =$

$$\begin{cases} \frac{x^2}{83.334} & \text{if } x > 0 \\ 0 & \text{if } 0 \\ \frac{x^2}{83.334} & \text{if } x < 0 \end{cases} \quad -5 \leq x \leq 5$$


By trial and error I determined that $\int_{-5}^{-3.965} (x^2/83.334)dx = 0.25066$, so

25% of the density is between -5 and -3.965 . So by symmetry, 25% of the density is between 3.965 and 5 . So, the shortest 50% confidence interval for this density is -5 to -3.965 and 3.965 to 5 . So, this is an example of an interval that is not continuous. Note, however that it is symmetric around the mean.

Consider the following example where the shortest 50% confidence interval is not symmetric around the mean. Assume $f_X(x) = \text{ChiSquareDen}(x; 5)$



In my software the command for the CDF of the Chi-Square is $\text{ChiSquareDist}(x; \mu)$
 $\text{ChiSquareDist}(4.4; 5) = 0.50663$. So, the shortest 50% confidence interval is 0 to 4.4. Since the expected value of this Chi-Square is 5 this interval is not centered on the mean and does not even include the mean.

21.