Mean comparisons: *t*-test vs Bayesian-- or Let the Prior Beware!

A look at the perils of NHST and mis-steps one can take with Bayesian estimation

PA 412 TW-PA-14243 approved for public release;
distribution unlimited

James Brownlow, Tim Jorris, Edwards AFB, CA

A few observations

- Bayesian analysis is, of course, wrong… except for the fact that it works.
- (1968) Milton Friedman to Richard Nixon, “We’re all Keynesians now.” (2014) Dr. J to LtCol Jorris, “We’re all Bayesians now.”
- “Far better an approximate answer to the right question than an exact answer to the wrong question.” J.W. Tukey

Observations…

- “Everyone who is actually involved in the real world does things in a Bayesian way.” – Sharon McGrayne, “the theory that would not die,” page 209
- L. J. Savage: Personal probability a “natural” development from Neymen-Pearson ideas (*Foundations of Statistics,* p iv.)
- A good reference for the Bayesian aspects of today’s presentation is John Kruschke’s book, *Doing Bayesian Data Analysis*, Burlington, Elisver

Overview

- Compare a one-sample *t*-test to an equivalent test using Bayesian estimation
- Bayes Law:

$$f(\theta|\text{data}) \; \alpha \; L(\text{data}|\; \theta) * \text{prior}(\theta)$$

Where θ is the parameter (or parameters) of interest

Basic problem

- The Take-Off roll for a particular jet is, on average 2500 feet, with a standard deviation of 200 feet.
- A new inlet design for the jet may result in a longer TO roll. Given a “small” set of TO roll distances from the jet with the new inlet, can we assert that the average TO roll is >2500 feet?
- If the TO roll with the new inlet is, on average, greater than 2750 feet, flight manuals will have to be re-written

On to test…

- t – test provides information about the hypothesis that the average TO distance is greater than 2500 feet
- t – test is a typical “null hypothesis significance test” (NHST); the null is “no difference” (the average TO roll is no different with the new inlet) and we either “reject” or “fail to reject” this null based on observed data and an arbitrary p-value
- What if we want to affirm the null hypothesis? We want to assert that the average TO roll is 2500 feet. Can’t do that with NHST

t-test vs Bayesian posterior

- How certain are we about the magnitude of the difference in mean TO roll between jets with the old and new inlet?
  - *t-test*: confidence interval for difference between average TO roll of the jet with the new inlet and 2500 feet
  - Bayesian: predictive distribution of the difference between the TO roll of the jet with the new inlet and 2500 feet

Magnitude of the difference between average TO roll and 2500 feet

- Doesn’t the p-value tell us all we need to know?
  - NO! One of the perils of NHST is that the p-value is the probability of getting the observed data set, (or any future data set with the observed mean and standard deviation), given the null hypothesis is true
  - What?? I don’t care about this or future data sets, I want to know about the difference in TO roll between the jet with the new inlet and 2500 feet

NHST: *t*-test

- If I use a *t*-test to test the hypothesis H_0 : $\mu \leq 2500$ feet, and get a p-value of, say, 0.001, doesn't that mean that there is only one chance in 1000 that the null hypothesis is true?
  - That may be how the *t*-test results are interpreted, but it's wrong: p-values are not "strength of evidence"
  - If the p-value is > 0.05, doesn't that mean I can "accept" the null? – No!   p-value simply estimates the probability of getting the data you observed, given the null is true. This is not evidence for the null.

Bayesian approach is parameter based

- Bayesian posterior distribution gives an explicit description of credible parameter values without NHST or p-values- *i.e.* Bayes provides:
  - Ability to *accept* the null hypothesis, not just “fail to reject”
  - Ability to use hierarchical modeling and inclusion of prior information (empirical and “personal”)
  - Procedure to handle outliers (“heavy tail” distributions)
  - Power analysis and sample size determination based on highest probability density (HPD) intervals

Cut to the chase..

- *t*-test estimates the probability of getting the observed data set, given the null hypothesis is true
- Bayesian parameter estimation provides a more complete look at the parameter(s) of interest
- Bayesian inference re-allocates credibility across the parameter space for all parameters, not just the means

Example: TO distance

- Modification to the inlet of a trainer
  - Manual says average TO roll = 2500 feet
  - If new average TO roll ≥ 2500, we have to rewrite the manual
  - Standard deviation on a TO roll is typically 200 feet
- Collect 4 data points:
  - 2712, 2982, 2597, 2470 feet

Aside …

- The 4 TO roll distance data points were generated for this example from a normal distribution with mean=2750 feet and standard deviation=200 feet
- “Correct” thing to do is rewrite the manual

t-test

- Test whether the average TO roll is 2500 feet=
  - H_0: μ=2500 vs H_1: μ>2500
- Power of the test:
  - n=4, delta=275, sd=200, significance=0.05, one-sided alternative, power =0.67

t-test results

- One Sample t-test data:
  - t = 1.8176, df = 3, p-value = 0.1667
  - alternative hypothesis: true mean is not equal to 2500
  - 95 percent confidence interval: 2367 to 2986 feet
  - sample estimate of the mean: 2677
- Rewrite the manual? Insufficient evidence
- p-value is the probability of getting the observed sample given the null hypothesis is true (mean TO distance is 2500 feet)

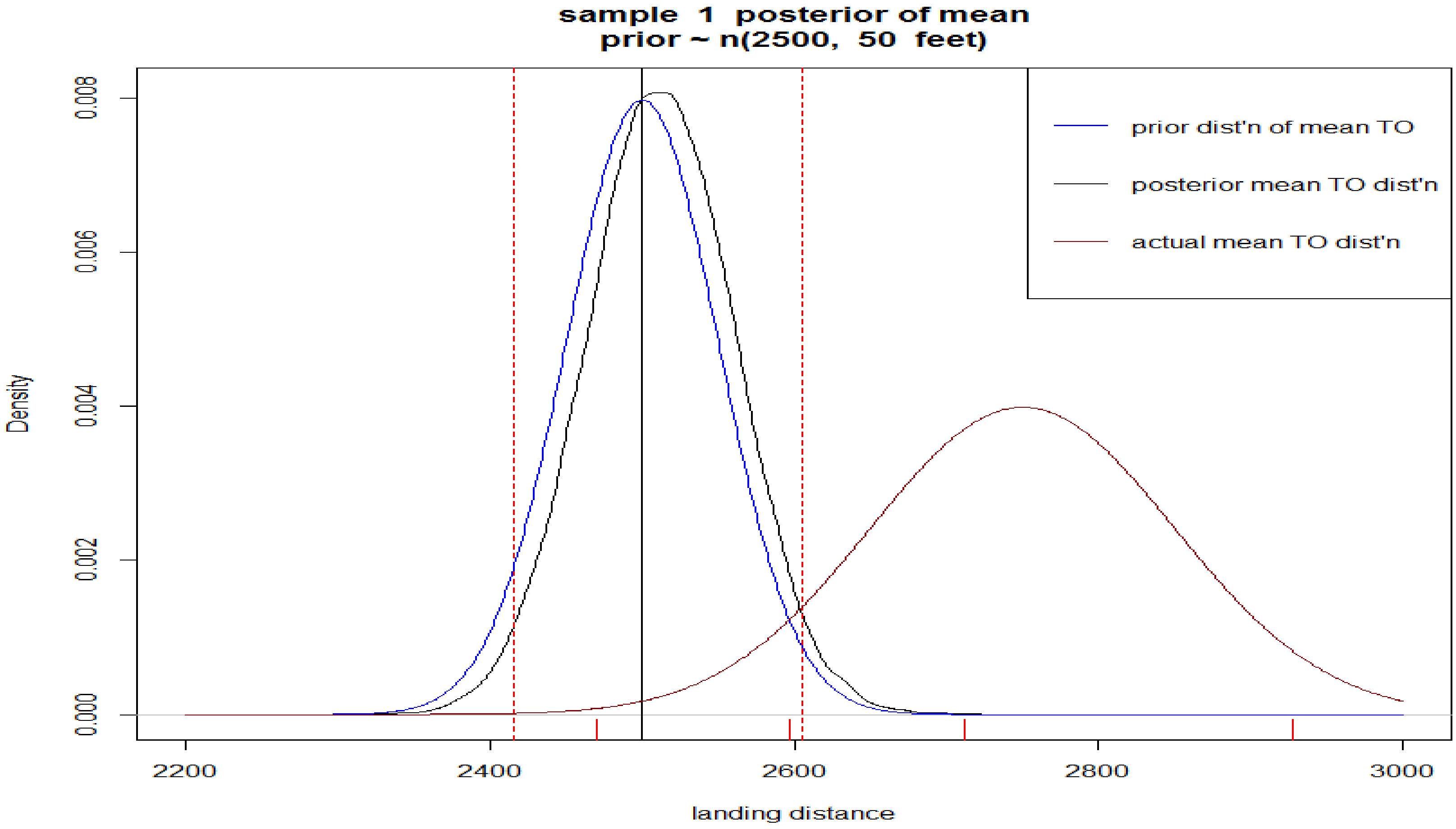
Well, small data sample…

- We have only 4 data points, and no evidence that the average TO roll > 2500 feet
- This test has very little power (remember the data were generated from a normal distribution with mean=2750, standard deviation=200): power = 0.67 (one sided)

And what does Bayes have to contribute?

- “LET THE PRIOR BEWARE” –LtCol Tim Jorris,
- Use the same 4 data points, and a normal likelihood for the data X_i ~normal(TO_mean, 200), i=1,2,..,4
- Use an informative prior, an engineering assessment is that the TO mean is normally distributed TO_mean ~ normal(2500, 50)
- This says prior belief is that TO_mean distance is 2500 feet, sigma = 50 feet

Bayesian estimate of mean TO distance, informative prior

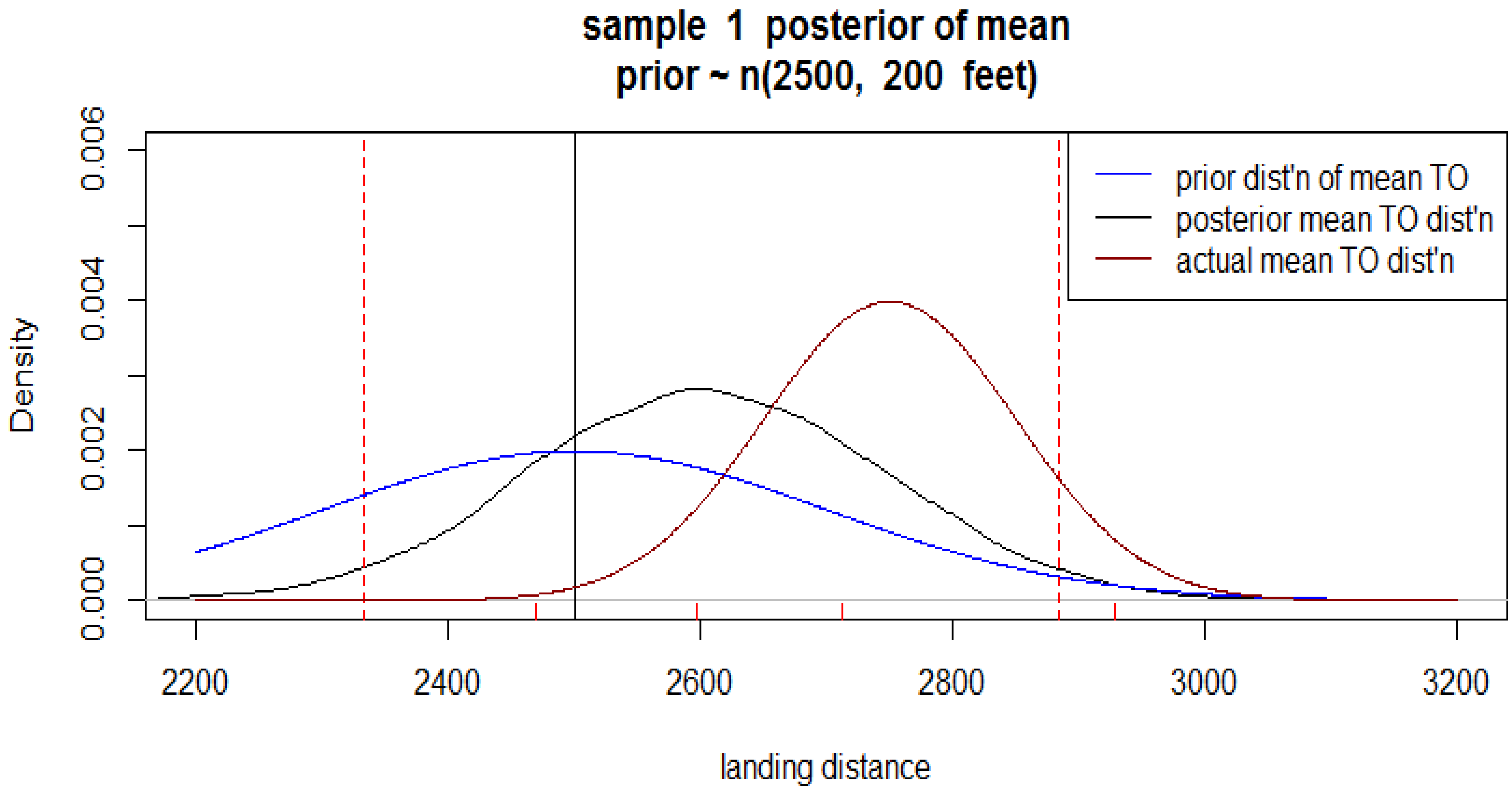


What does this say?

- There is no chance that the average TO roll distance is 2750, no need to re-write the manual- WRONG!
- The informative prior overpowered information from the data!
- Prior information: we said we’re 95% certain that the average TO roll with the new inlet is between 2400 and 2600 feet- insufficient data to counter this

OK, suppose we go with the status quo

- Assume the prior has mean 2500 and sd = 200



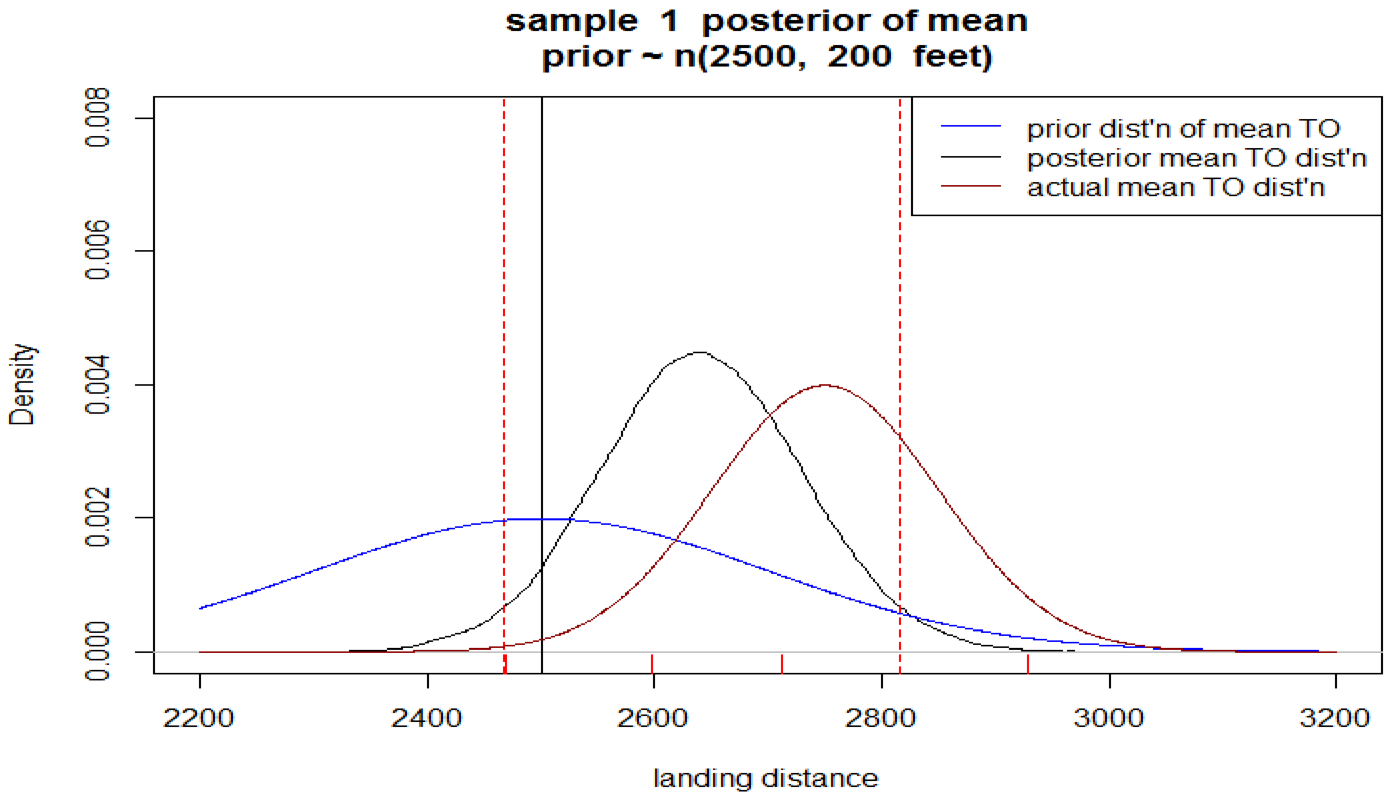
“Less informative” prior

- There is a 77% chance that the average TO roll distance is greater than 2500, probably need to re-write the manual
- The informative prior is not as overpowering
- Prior information: we’re 95% certain average TO roll for the new inlet is between 2100 and 2900 feet
- Posterior: 95% certain that the average TO roll with the new inlet is between 2330 and 2880 feet

What if we use a “non-informative” prior, or increase the uncertainty ?

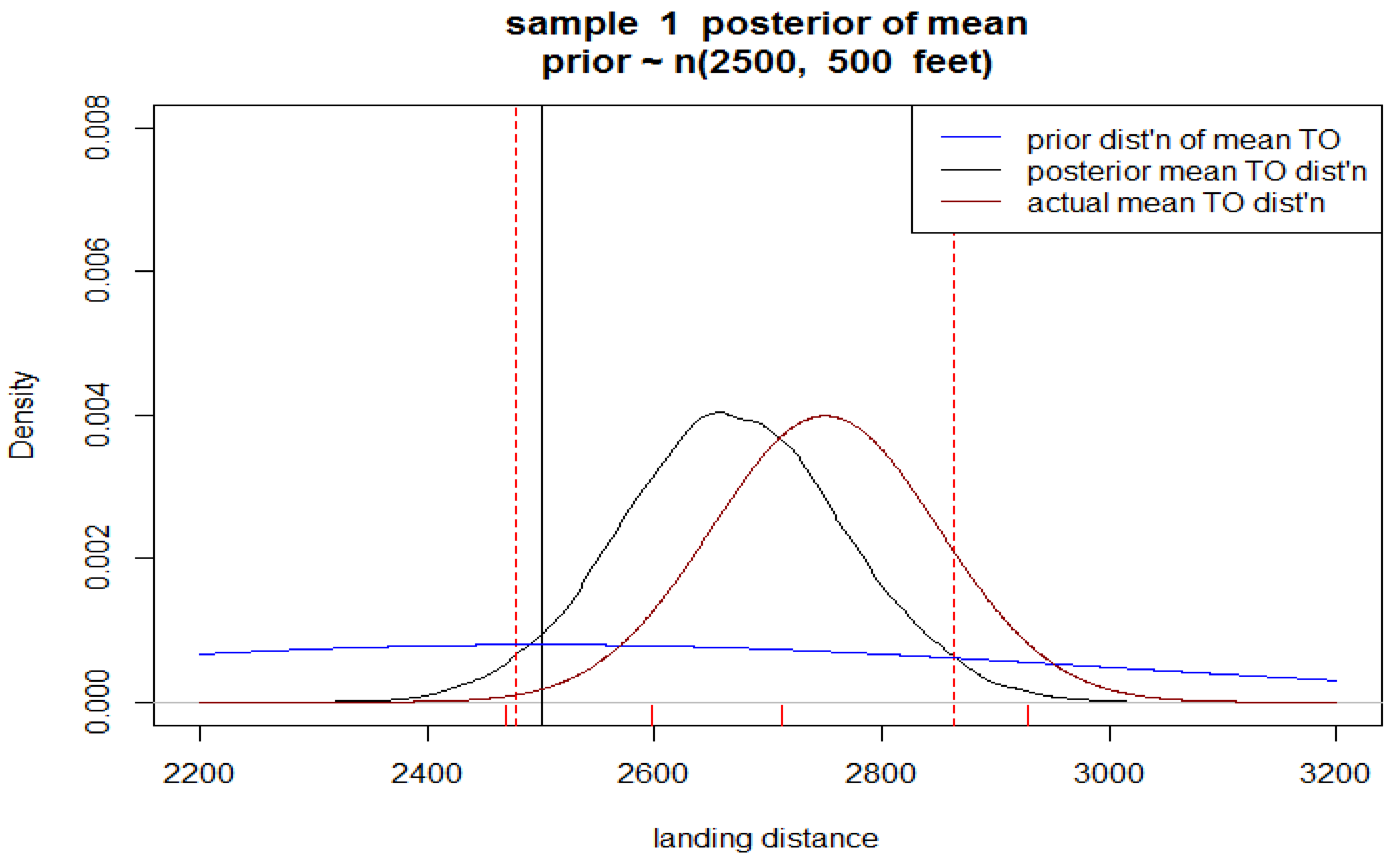
- Use the same distribution for the likelihood associated with the data
- Widen the uncertainty in the average TO roll distance: same normal prior, but with a greater standard deviation (mean=2500, sigma = 200, and then 500 feet)

- Average TO roll, prior is normal, mean=2500, sd = 200



Non-informative prior on average TO roll

- Average TO roll prior normal, mean=2500, sd=500 feet.



With the non-informative prior

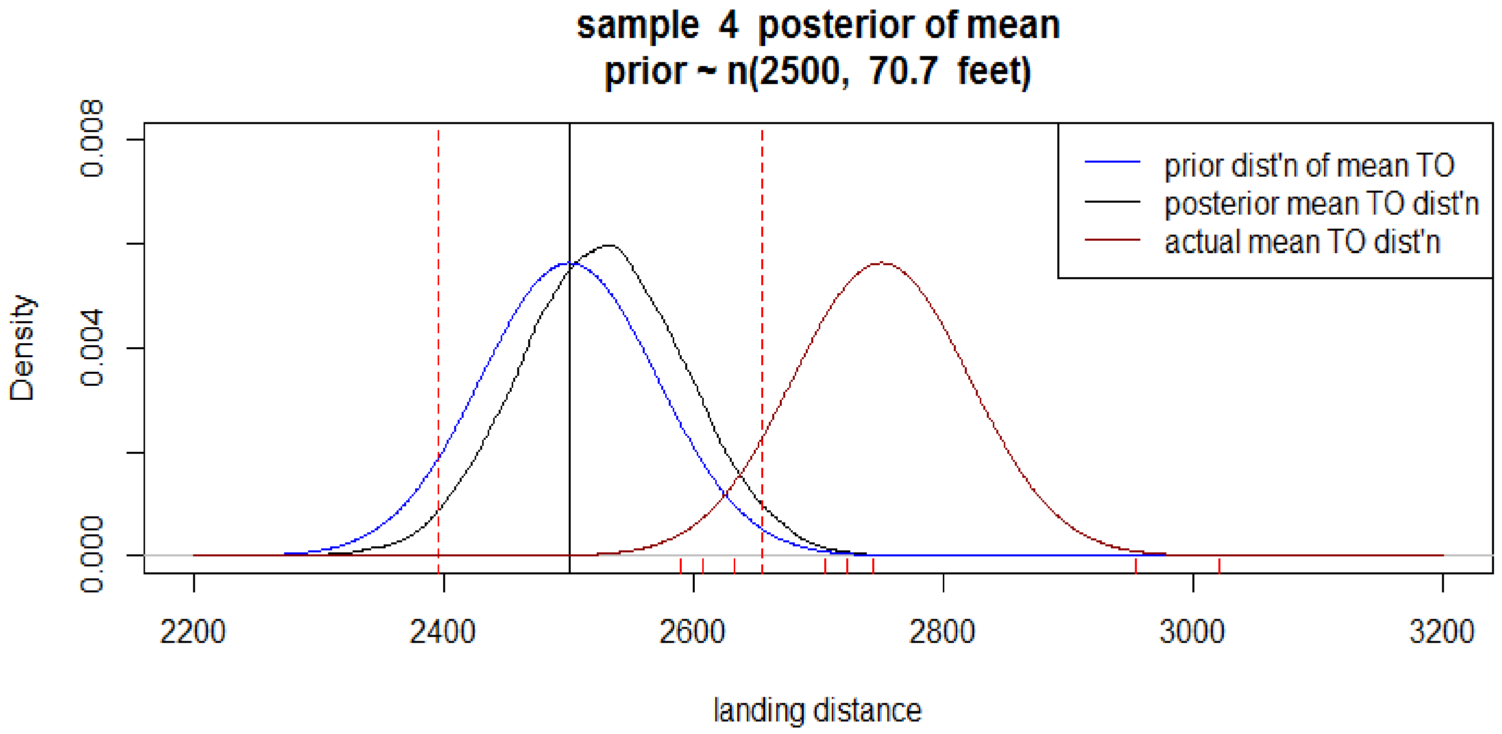
- There is a 95% chance that the average TO roll is between 2326 and 3054 feet
- The probability that the average TO roll is greater than 2500 feet is 0.83
- Bayesian result with only a few data points is heavily dependent on prior information

Well, get more data!

- We use a sample size of 8:
  - 2743, 2590, 2954, 2723, 2607, 3020, 2705, 2632 feet
  - 8 samples, power = 0.97
- *t*-test results:
  - t = 4.38, df = 7, p-value = 0.003
  - alternative hypothesis: true mean is greater than 2500
  - 95 percent confidence interval: 2615 to 2880 feet mean of = 2745

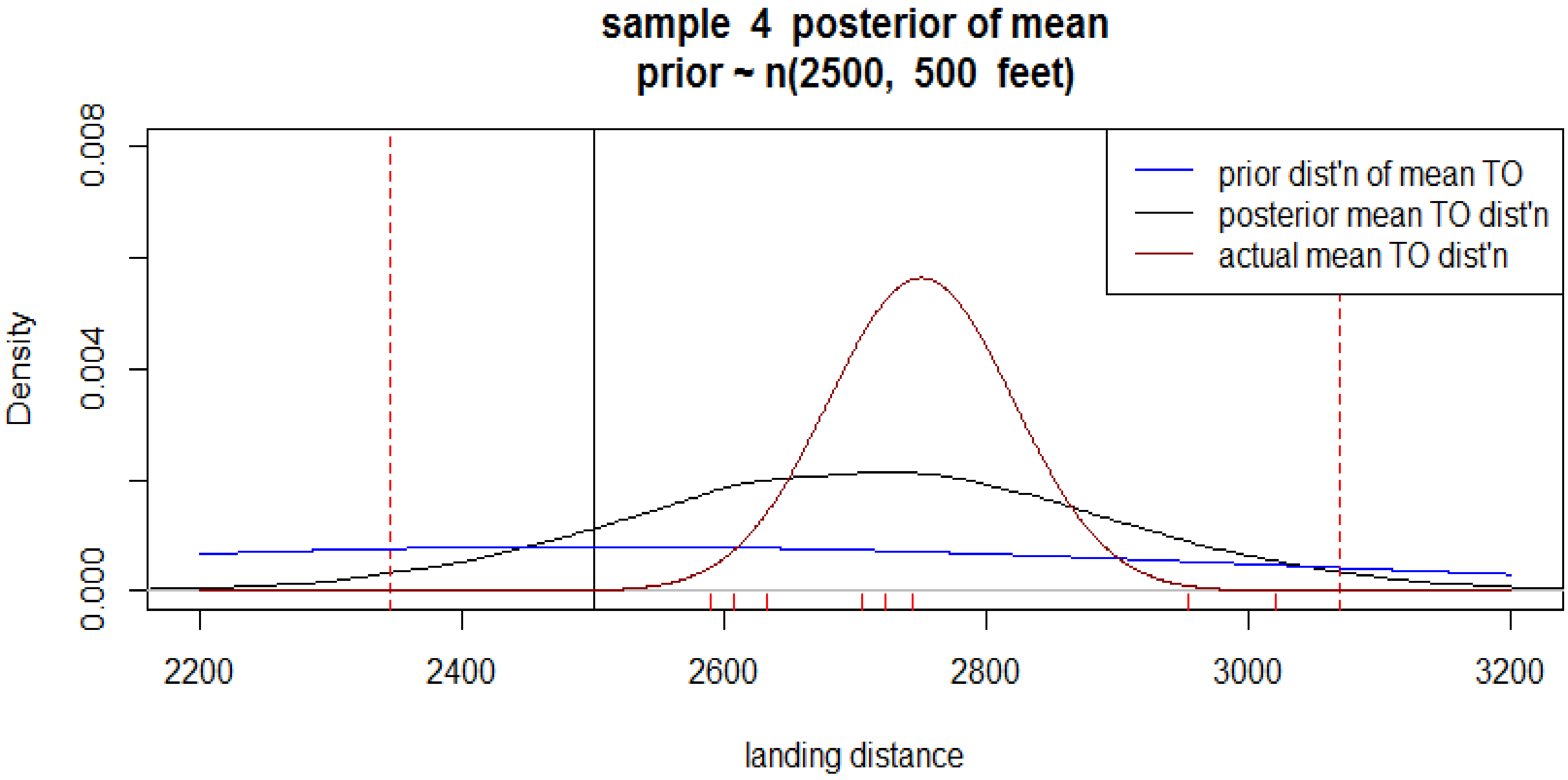
What do we do now?

- P=0.003, re-write the manual
- Bayes result: prior sigma = 200/sqrt(8)



Bayes with a non-informative prior

- Average TO roll normal(2500, 500)

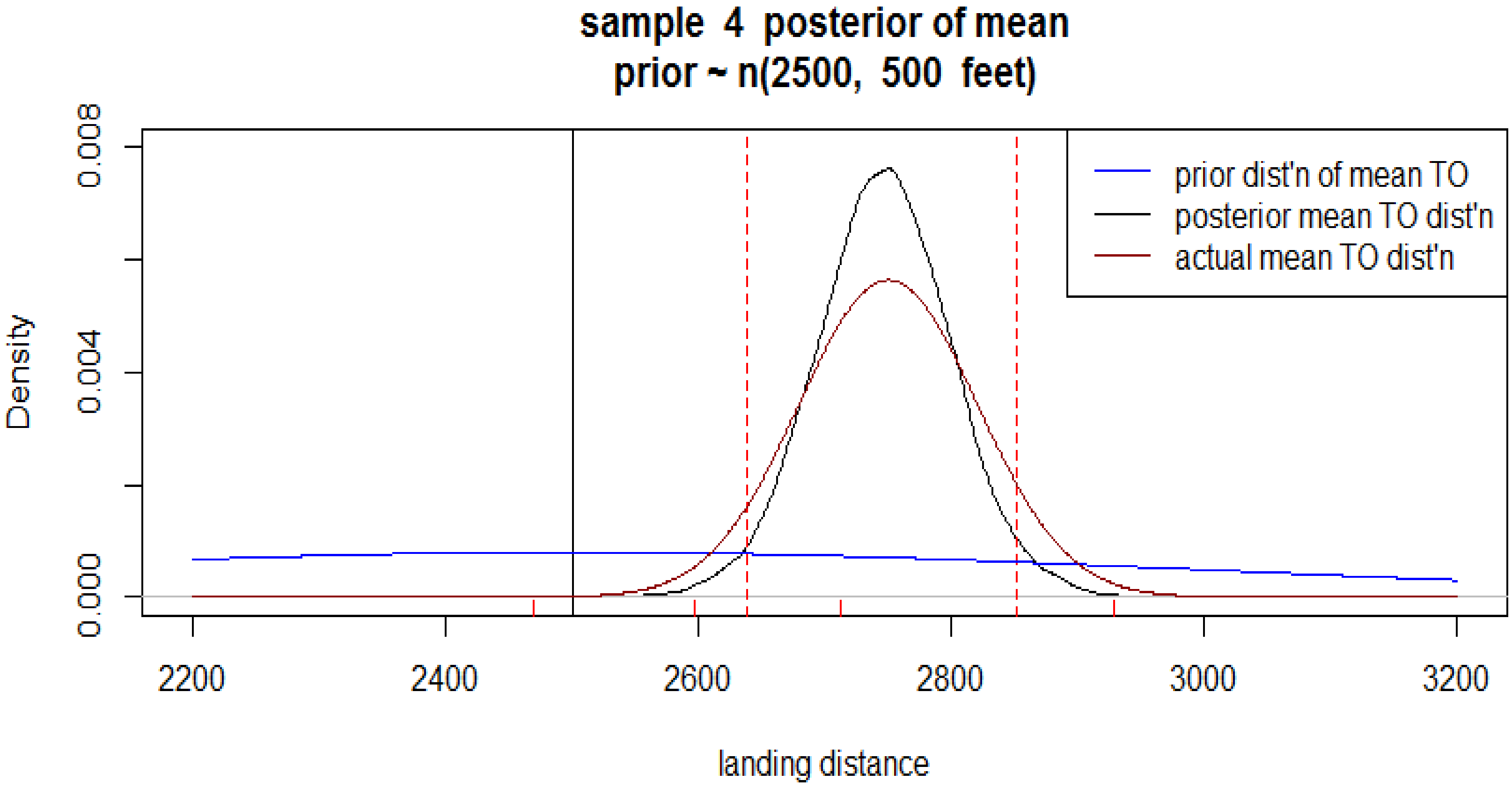


Bayes with non-informative prior

- Probability average TO roll ≥ 2500, is 0.87
- Bayesian result based on the assumption that the observed data were from normal(2500, 200) – what if we put a prior on the distribution of the observations (as well as on the distribution of the mean TO distance)

Change likelihood distribution of the data

- Data likelihood normal(h.mean, h.sigma)



Conclusions

- Results are only as good as the data and associated assumptions