One sample location problems

S520

October 20, 2016

Questions to ask

Before rushing into doing a significance test or finding a confidence interval, here are some question you should ask. The answers will help you work out what method you should use.

1. What is the experimental unit? (The experimental units must be independent.)
2. From how many populations were the experimental units sampled? (Remember that the units within each population must be identically distributed.) What are the populations?
3. How many measurements were taken on each experimental unit? What are the measurements?
4. What are the parameters of interest for this problem?

For one-sample location problems, first define the random variable X_i in terms of the measurements taken on unit i. The parameter is either the population mean μ or the population median θ.

For two-sample location problems, define X_i in terms of the measurements taken on unit i in the first sample and Y_j in terms of the measurements taken on unit j in the second sample. The parameter of interest is usually the *difference* in population means:

$$\Delta = \mu_1 - \mu_2$$

where $\mu_1 = EX_i$ and $\mu_2 = EY_j$. Note that it doesn't which sample you call the X's and which you call the Y's, as long as you're consistent throughout the analysis.

5. Do you need to do a significance test? If so, what are appropriate null and alternative hypotheses? In a one-sample location problem, then the hypotheses should be statements about μ (or θ.) In a two-sample location problem, then the hypotheses should be statements about Δ.

Normal populations

In chapter 9, we did inference based on an assumption of an approximately normal distribution for the error $\bar{X} - \mu$, with a standard error of about $s/\sqrt{n}$. This requires a large sample for two reasons:

- The approximate normal distribution for the error is justified by the Central Limit Theorem.
- The real standard error is $\sigma/\sqrt{n}$, but with a large sample, s will be close to σ.

What if we don't have so large a sample? Then we need to make further assumptions. For example, if we have a distribution with extreme skewness or bad outliers, we might not be able to say anything accurate about the population mean from a small sample.

Example. Consider the population consisting of a STAT S-520 class plus Bill Gates. Suppose we wish to estimate the population mean wealth by taking the sample mean with $n = 10$. Well, if our sample of ten includes Bill Gates, we'll horrendously overestimate the average wealth, and if our sample of ten doesn't

include Bill Gates, we'll horrendously underestimate the average wealth. A simple random sample of size n just isn't big enough for this problem.

So for now let's deal with situations where we know the form of the population distribution. Suppose we take a sample of size n from a Normal distribution with mean μ and variance σ^2. Then the sample mean has distribution

$$\bar{X} \sim \text{Normal}(\mu, \sigma^2/n).$$

What if we don't know μ – what if μ is what we're trying to find out? Then

$$\text{Error} = \bar{X} - \mu \sim \text{Normal}(0, \sigma^2/n).$$

This will be easier to deal with mathematically if we rescale to get a standard normal distribution. So divide through by $\sigma/\sqrt{n}$:

$$\frac{\bar{X} - \mu}{\sigma/\sqrt{n}} \sim \text{Normal}(0, 1).$$

So if we know σ, we can do inference using `pnorm()` and `qnorm()`. However, most of the time we don't know σ. Instead we have to use s, the sample standard deviation instead. We've previously asserted that for large samples this is fine. But what happens for moderate or small n? Let's define the t-statistic:

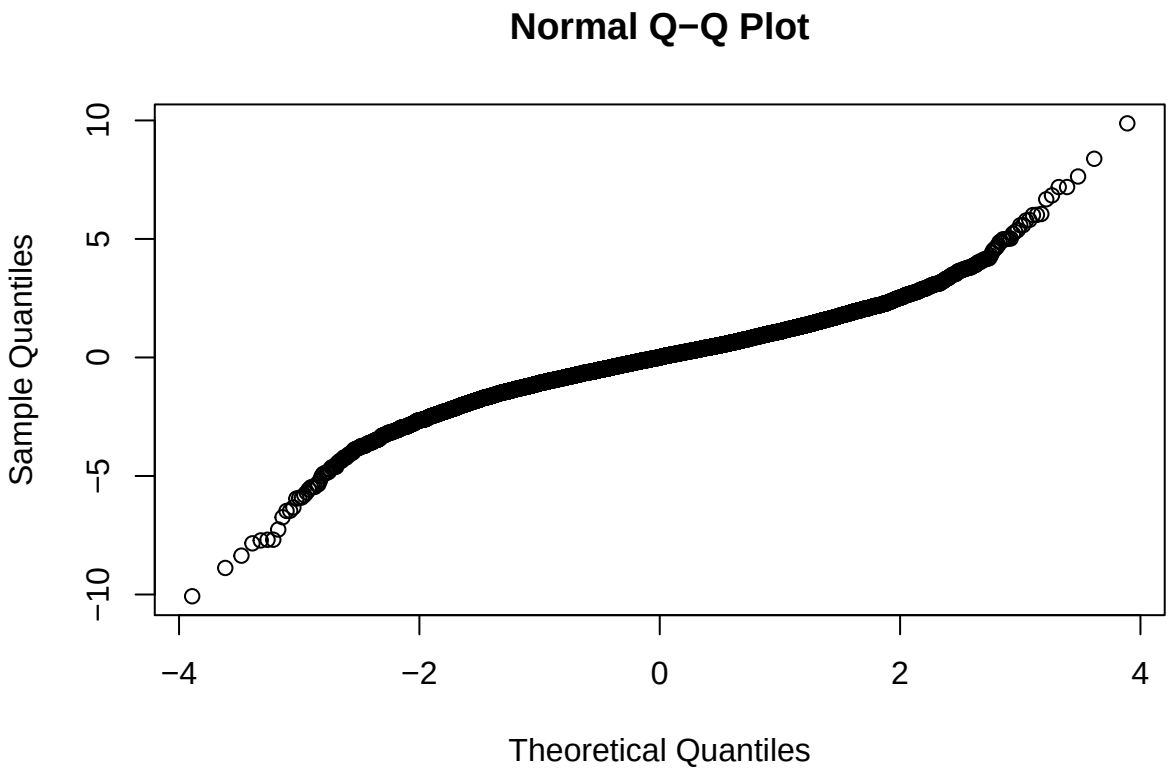
$$t_n = \frac{\bar{x} - \mu}{s/\sqrt{n}}$$

Trosset p. 240 goes through some of the theory, but I'll proceed by simulation. Consider the case where we take a sample from a normal population and calculate this statistic t_n.

```
t.statistic = function(n, mu=0, sigma=1){
  my.sample = rnorm(n, mu, sigma)
  x.bar = mean(my.sample)
  s = sd(my.sample)
  t = (mean(my.sample) - mu) / (s/sqrt(n))
  return(t)
}
```

Now replicate this lots of times with $n = 6$. Is the distribution normal, or something else?

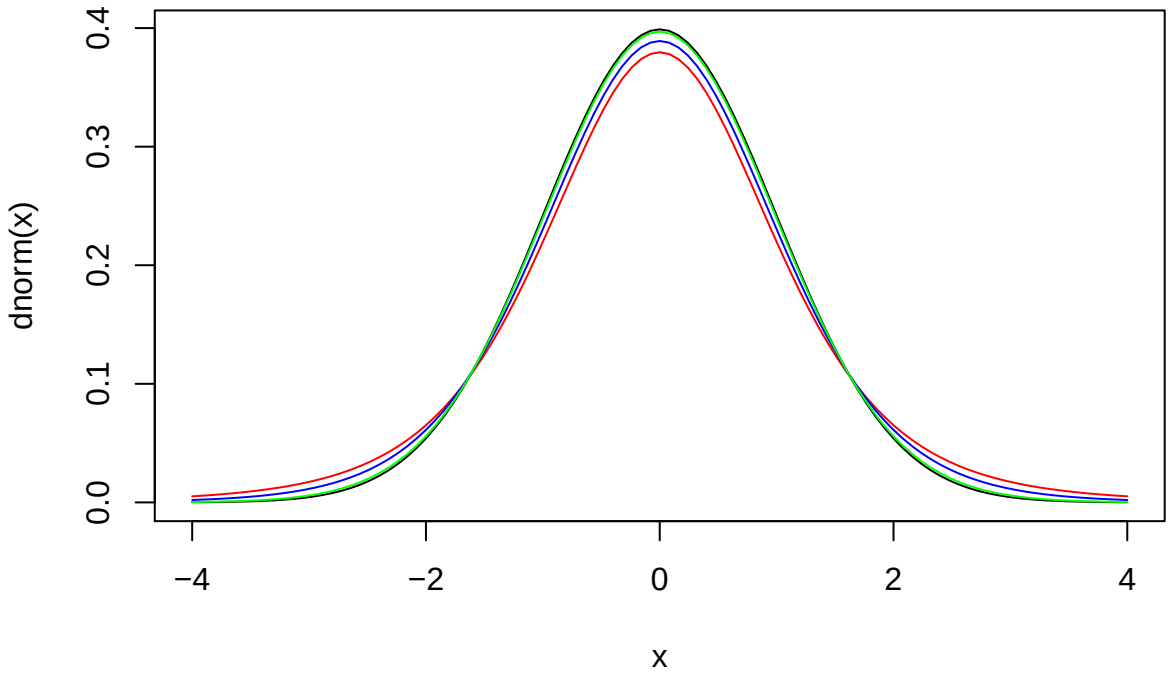
```
t.list = replicate(10000, t.statistic(n = 6))
qqnorm(t.list)
```



Nope! Instead the distribution is something called a t-distribution, and here in particular it's a t-distribution with 5 degrees of freedom. Why 5? <handwave> The sample size is 6, so if the mean is given, then 5 observations can vary freely; the sixth is determined by the other 5.</handwave> In general, if the sample size (from a Normal population) is n, the t-statistic follows a t-distribution with $n-1$ degrees of freedom.

We can get a better idea of what this is by plotting some PDFs:

```
curve(dnorm, from=-4, to=4)
curve(dt(x, df=5), col="red", add=TRUE)
curve(dt(x, df=10), col="blue", add=TRUE)
curve(dt(x, df=50), col="green", add=TRUE)
```



The PDF of the t-distribution with 5 degrees of freedom is given by the red curve. Like the normal, it's symmetric. Unlike the normal, it doesn't die away so quickly and you get further away from the center. The greater the "degrees of freedom", the closer you get to the Normal curve. We can see the difference by finding some tail probabilities:

```
pnorm(-1.75)
```

```
## [1] 0.04005916
```

```
pt(-1.75, df=5)
```

```
## [1] 0.07026118
```

```
pt(-1.75, df=10)
```

```
## [1] 0.05534047
```

```
pt(-1.75, df=50)
```

```
## [1] 0.04312663
```

To summarize: When the population is Normal, the variable $T_n = \frac{\bar{X}-\mu}{S/\sqrt{n}}$ has a t-distribution with $n-1$ degrees of freedom. Thus the tail probabilities of the t-distribution can be used to construct a P-value for a test of the hypothesis that the population mean takes a certain value.

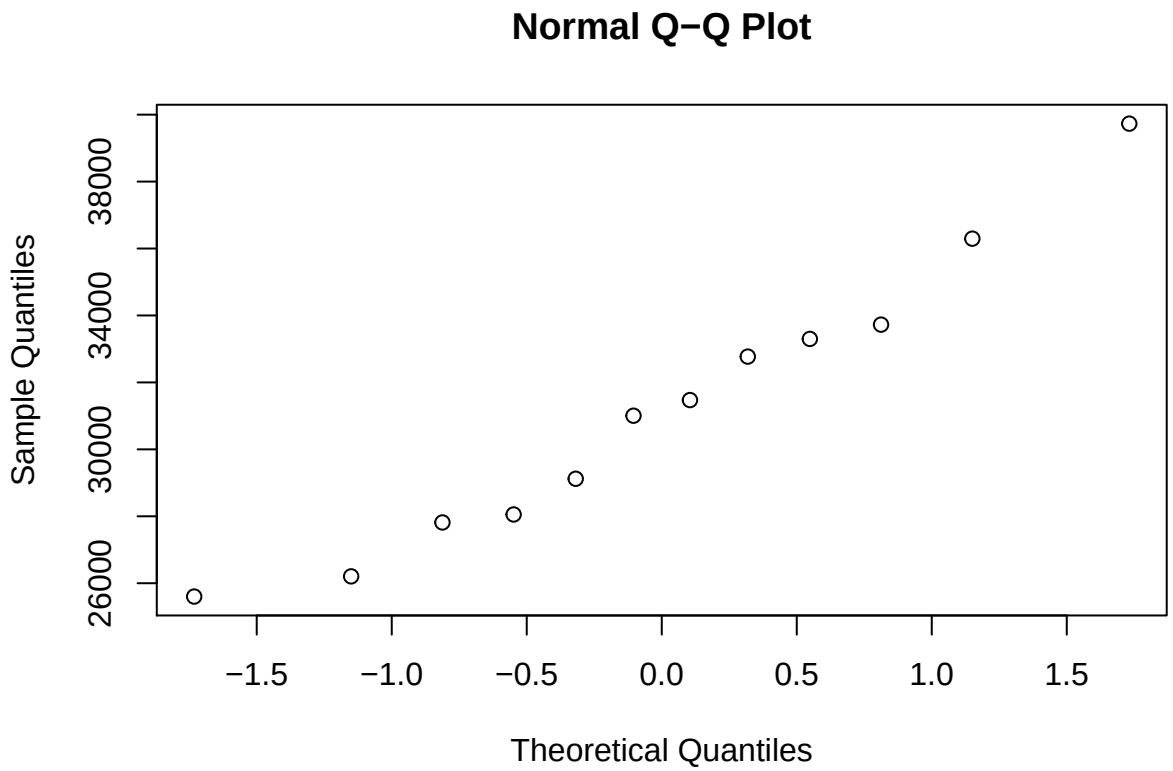
Example: Country club fees

Here's an example from a business statistics textbook I used to use. In 2008, the average country club fee was \$31,912. We have a sample of twelve country club fees from 2009. Has the average fee changed? Here's the 2009 data:

```
fees = c(29121, 31472, 28054, 31005, 36295, 32771,
         26205, 33299, 25602, 33726, 39731, 27816)
```

Are the fees normal?

```
qqnorm(fees)
```



The QQ plot is fairly close to a straight line. Now, there's no way of being sure that the population is approximately Normal just from a sample of size 12. There could be huge outliers outside of our sample – this seems quite possible when it comes to country club fees. Nevertheless, we'll make a leap of faith, and assume the population is approximately Normal.

```
mean(fees)
```

```
## [1] 31258.08
```

```
sd(fees)
```

```
## [1] 4199.802
```

```
t = (mean(fees) - 31912) /
  (sd(fees) / sqrt(12))
# Two-tailed P-value
2 * pt(t, df=11)
```

```
## [1] 0.6003808
```

```r
# or
2 * (1 - pt(abs(t), df=11))
```

```
## [1] 0.6003808
```

```r
# Find 95% CI
mean(fees) - qt(.975, df=11) * sd(fees)/sqrt(12)
```

```
## [1] 28589.66
```

```r
mean(fees) + qt(.975, df=11) * sd(fees)/sqrt(12)
```

```
## [1] 33926.51
```

What can we conclude?

- The data is compatible with the hypothesis that the average country club fee hasn't changed from 2008 to 2009.
- But the confidence interval is fairly wide. It could be that the average fee has dropped $3000, or it could be that it's risen $2000.
- The most glaring problem is the small sample size. With a sample of size 12, then you would be unlikely to detect a change of, say, a couple of thousand dollars.
- And this is with a suspiciously small standard deviation. Surely country club fees vary a lot more than this!

Conclusions: (1) Get more data. (2) Get better data. (3) Get a better business statistics textbook.

We tested the hypothesis that the population mean was $31,912. Since the sample was small, we couldn't rely on the Central Limit Theorem, so instead we needed to assume the data came from an approximately normal population. Although the QQ plot looked straightish, this was still a leap of faith, because it's hard to show something's normal from a small sample – what if there are large outliers that we didn't happen to get in our particular sample? This is particularly easy to believe for something like country club fees, where there might be club that charge **much** more than most others. A priori, we might expect the distribution of fees to be right-skewed.

So what if you don't want to assume normality? There are a couple of options.

1. Get more data. This is the most honest approach, but may not be practically feasible.
2. *Transform* the data to have a closer-to-normal distribution. This is particularly useful with positive, right-skewed data, where a log transformation often makes the data a lot closer to normal (and is able to be interpreted.) However, with small samples, we still face the problem that it's hard to know if it's normal whether or not you do a transformation.
3. Instead of studying the mean, study the **median**. Alternatively, maybe the median really is of more interest than the mean, which is often the case with skewed distributions.

The sign test

We wish to test the null hypothesis that the population median is θ_0, and as usual observe iid data $X_1, \ldots, X_n$. Let Y be the number of observations **greater** than the hypothesized median. If this null is true, then Y has a binomial distribution with sample size n and $p = 0.5$, because there's a 50-50 chance of being above the median (assuming the X's are continuous.)

We compare the observed value of Y to this binomial distribution. For a two-tailed test, the P-value will be double the small tail. There are three cases:

- If $y < n/2$, the two-tailed P-value is $2 \times P(Y \leq y) =$ `2 * pbinom(y, n, 0.5)`.
- If $y > n/2$, the two-tailed P-value is $2 \times P(Y \geq y) =$ `2 * (1 - pbinom(y-1, n, 0.5))`.
- If $y = n/2$, the two-tailed P-value is 1.

If you get a P-value bigger than 1, you've messed up.

Example. Test the null hypothesis that the population **median** country club fee is $31,912:

```
plusses = sum(fees > 31912)
plusses
```

```
## [1] 5
```

```
2 * pbinom(plusses, length(fees), 0.5)
```

```
## [1] 0.7744141
```

Note that this method (1) does not require you to know the form of the population distribution, and (2) does not require a large sample. So it can be used very generally. However, it's not magic – if your sample is too small, your test is still unlikely to be able to reject the null even when it's false. That is, the sign test solves the assumptions problem, but not the more important "small samples don't tell us much" problem.

Sign test confidence intervals (will probably skip this)

We can, in general, construct confidence intervals by *inversion*: a two-sided $(1-\alpha)$ confidence interval consists of all the parameter values we would **not** reject in a two-tailed level α test.

For the sign test, our intervals for the median are of the form

$(k+1)$th smallest observation to $(n-k)$th largest observation

The level of confidence associated with such an interval is

`1 - 2 * pbinom(k, n, 0.5)`

Example. For the country club data, suppose we want an interval for median at a level of confidence of about 95%. First we need to find a value of k:

```
1 - 2*pbinom(0:5, 12, 0.5)
```

```
## [1] 0.9995117 0.9936523 0.9614258 0.8540039 0.6123047 0.2255859
```

We see that $k = 2$ gives an interval at 96% confidence, which is as close as we can get. (In general it's better to have too much confidence than too little.) $k = 2$ means our interval goes from the 3rd smallest to the 3rd largest of the twelve values:

```
sorted.fees = sort(fees)
sorted.fees[c(3, 10)]
```

```
## [1] 27816 33726
```

So a 96% confidence interval for the median country club fee is $27,800 to $33,700.

Dealing with ties

The above method works for continuous data. For discrete data, we need to adjust the method slightly to deal with the possibility that an observation could be **exactly** equal to the hypothesized mean. The rule of thumb is to do whatever gives you the **biggest** P-value.

Example: Ergonomic keyboards.

```
ergonomic = c(69, 80, 60, 71, 73, 64, 63, 70, 63, 74)
standard = c(70, 68, 54, 56, 58, 64, 62, 51, 64, 53)
differences = ergonomic - standard
differences
```

```
##  [1] -1 12  6 15 15  0  1 19 -1 21
```

We have seven observations greater than zero, one equal to zero, and two less than zero.

Suppose we wish to prove the alternative that the median difference (ergonomic minus standard) is greater than zero. In the continuous case, if y was the number of observations greater than zero, our one-tailed P-value would be:

```
1 - pbinom(y-1, n, 0.5)
```

(It's the right tail because our alternative is a "greater than"; if the alternative were true, we'd expect lots of observations above zero, hence a big y and a small P-value.)

Whether due to rounding or otherwise, we have a zero in the data set. Do we treat this observation as "just above zero" or "just below zero"? We can try both options:

```
1 - pbinom(7-1, 10, 0.5)
```

```
## [1] 0.171875
```

```
1 - pbinom(8-1, 10, 0.5)
```

```
## [1] 0.0546875
```

The bigger P-value is 0.17. Since this is not especially small, we have not shown beyond reasonable doubt that the median difference is positive.

If we didn't skip the section on the confidence interval for the median, we should probably find one of those as well. What confidence levels are achievable?

```
1 - 2 * pbinom(0, 10, 0.5)
```

```
## [1] 0.9980469
```

```
1 - 2 * pbinom(1, 10, 0.5)
```

```
## [1] 0.9785156
```

```r
1 - 2 * pbinom(2, 10, 0.5)
```

```
## [1] 0.890625
```

89% confidence seems a bit low, so let's do a 97.85% confidence interval ($k = 1$). This runs from the second-lowest to the second-highest value:

```r
sort(differences)
```

```
##  [1] -1 -1  0  1  6 12 15 15 19 21
```

```r
sort(differences)[c(2,9)]
```

```
## [1] -1 19
```