Assumptions and Transformations

Bret Hanlon and Bret Larget

Department of Statistics
University of Wisconsin—Madison

November 10, 2011

The Big Picture

- The t-methods we have seen so far for one and two sample problems assume that underlying populations are normally distributed.
- Sometimes populations are not normal.
- There are three ways (at least) to handle this non-normality:
  - Just use the t-methods anyway: the methods are robust to nonnormality when the samples are large enough, because:
    - by the CLT, the sample mean is approximately normal;
    - the sample variance is approximately chi-square (scaled appropriately);
    - and the sample mean and sample variance are only very weakly dependent;
  - Use nonparametric methods (like randomization/permutation tests or the bootstrap);
  - *Transform* the variable so that it is more like a normal distribution, use the t-methods on the transformed data, and then transform back.

How to Decide if a Sample is Normal

- While there are formal methods to test for normality, we do not advocate their use for the following reasons:
  - No real biological distribution is exactly normal;
  - The real issue is to ascertain if the lack of normality in the populations will adversely affect methods based on that assumption—and formal tests do not test this;
  - For a small sample, there may be insufficient information to formally reject normality, but ignoring it could be perilous;
  - For a large sample, there may be enough data to demonstrate nonnormality, but the robustness of t-methods, especially for large samples, means that ignoring the nonnormality is not bad.

What to do

- Informal *graphical assessment and judgment* can help indicate when nonnormality is potentially problematic and when action (nonparametric methods or transformations) are warranted.
- Sample characteristics which indicate potential trouble are:
  - Strong skewness;
  - Extreme outliers.
- . . . especially for small samples.
- It never hurts to compare the inferences when using t-methods and nonparametric methods.

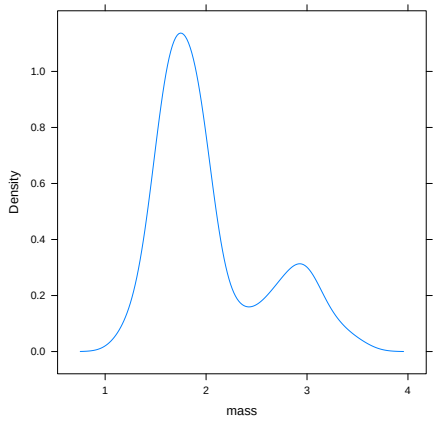
Quantile Plots

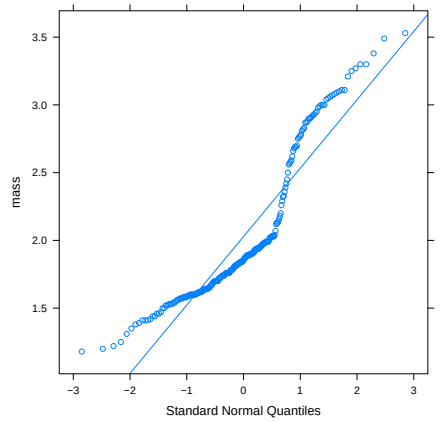
- Histograms and density plots show the shape of a distribution;
- One can see if a distribution is bell-shaped and symmetric, but subtle deviations from normality can be hard to see.
- A *quantile plot* plots *ordered sample values* against *quantiles of a standard normal distribution*.
- If the plotted points *form an approximate straight line*, then the sample is approximately normal.
- There are different ways to pick the quantiles; generally, they are spaced so that the area between them under a standard normal curve is equal.
- For example, with n points, the quantiles can be chosen so there is area $1/n$ in each of the $n-1$ gaps between quantiles and $1/(2n)$ in the two tails.
- In the case when there are 5 points, this corresponds to the 0.1, 0.3, 0.5, 0.7, and 0.9 quantiles.

Sockeye Salmon Revisited

Example

- Here is the female sockeye salmon mass example.
- It does not look normal.

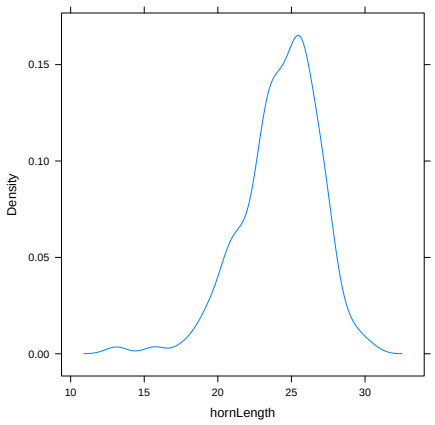


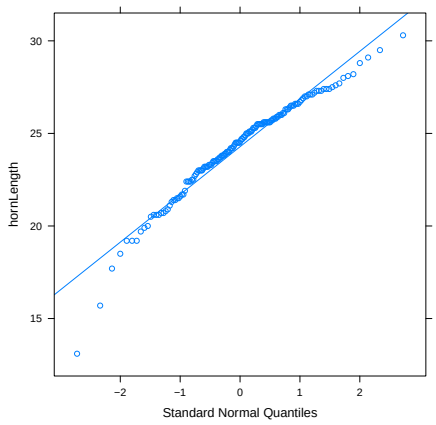


Lizard Horn Length Revisited

Example

- Here is the lizard horn length example for the living lizards.
- It is more normal than the salmon, but skewed a bit left.

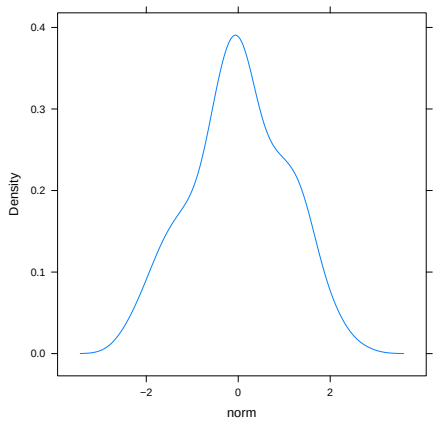


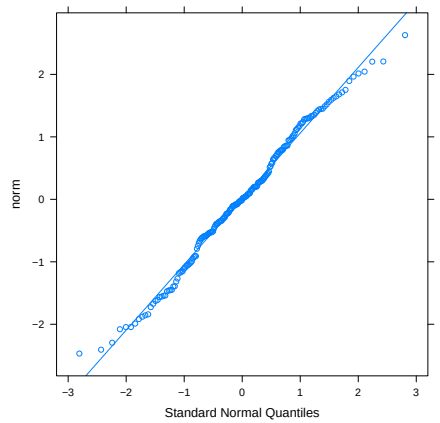


Example with Simulated Normal Data

Example

- This example is with 200 simulated normal data points.
- Even with truly normal sample data, there is some deviation from the line at the ends and some wiggle in the middle.





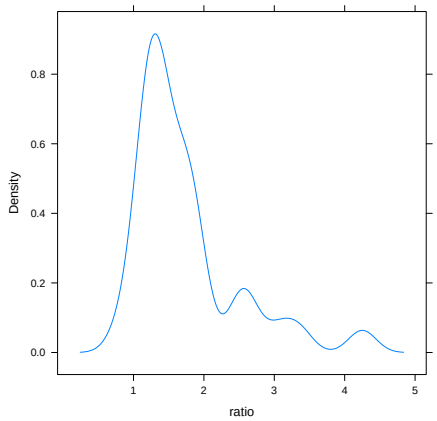
Biomass

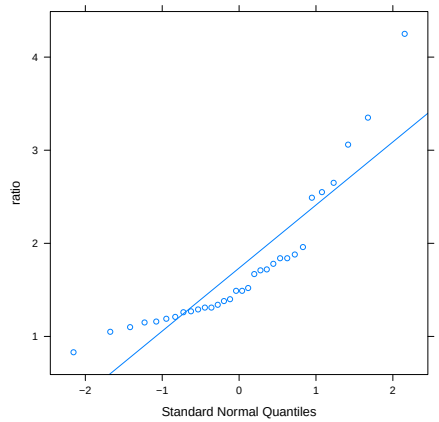
Case Study

Halpern (2003) compared each of 32 marine reserves to a control location (either a similar location or the same location prior to it being protected). For each pair of locations, a biomass ratio was calculated which was the ratio of the total mass of all marine plants and animals per unit area for the protected area over the same quantity for its control. Ratios larger than one are consistent with the protection leading to more abundant life (by mass).

Plots of Biomass Ratio

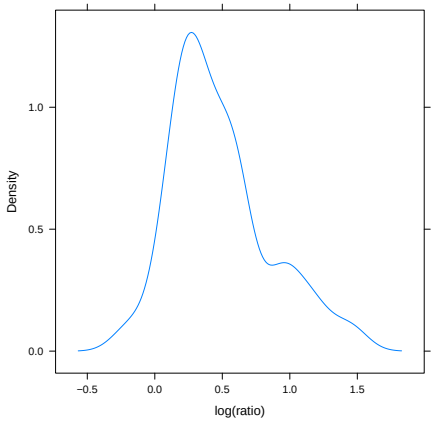
- The biomass ratio is skewed right.
- This often occurs with ratio data.

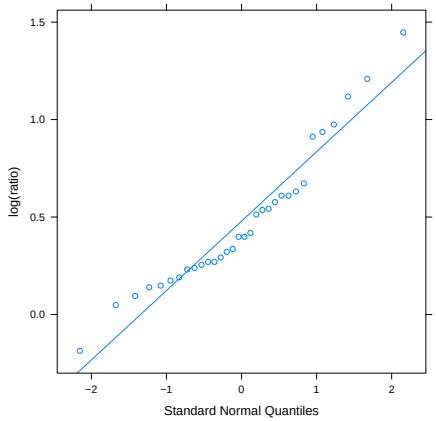




A natural log-transformation

- For positive right-skewed data, a natural log transformation often results in a more symmetrical distribution.





Confidence Interval

- When finding a confidence interval for transformed data, it is best to back transform the result to the original units.
- For the biomass ratio example, if we are 95% confident that

$$a < \ln \mu < b$$

then we are also 95% confident that

$$e^a < \mu < e^b$$

Numerical Example with R

```
> y = log(biomass$ratio)
> y.mean = mean(y)
> y.sd = sd(y)
> y.n = length(y)
> t.crit = qt(0.975, y.n - 1)
> a = y.mean - t.crit * y.sd/sqrt(y.n)
> b = y.mean + t.crit * y.sd/sqrt(y.n)
> ea = exp(a)
> eb = exp(b)
> out = c(y.mean, y.sd, y.n, t.crit, a, b, ea, eb)
> names(out) = c("mean", "sd", "n", "t*", "a", "b",
+     "exp(a)", "exp(b)")
> print(out)
      mean         sd          n         t*          a
 0.4791272  0.3664220 32.0000000  2.0395134  0.3470180
         b     exp(a)     exp(b)
 0.6112365  1.4148422  1.8427084
```


Interpretation in Context

We are 95% confident that the mean biomass ratio of protected over unprotected controls in marine reserves comparable to those included in the study is between 1.41 and 1.84. This suggests that protecting a marine environment may lead to an increase in the biomass between about 40 and 80 percent, on average.

Another Way with R

```
> results = t.test(log(biomass$ratio))
> exp(results$conf.int)
[1] 1.414842 1.842708
attr(,"conf.level")
[1] 0.95
```


What you should know

You should know:

- how to interpret quantile plots to assess normality;
- how to transform a variable before carrying out t-method inference;
- why transformations may lead to improved inference;
- how to back transform confidence intervals to improve interpretation.

Extensions

- Chapter 13 also describes nonparametric methods based on *ranks* of data.
- You are not responsible for this material: we prefer permutation/randomization methods or the bootstrap with the original data.
- Methods that use ranks allow p-values to be computed from tables, but simulation removes the need for this to make inference practical.

Other Transformations

- Different types of data are often better analyzed with different transformations: examples include:

  arcsine transformation $p' = \arcsin(\sqrt{p})$ (only for proportions);

  square root transformation $y' = \sqrt{y}$, often used for count data (the text suggests $\sqrt{y + 0.5}$);

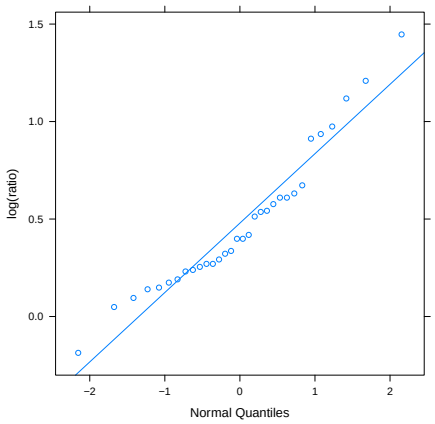
  reciprocal transformation $y' = 1/y$, sometimes useful for ratios or strongly right-skewed data—even more extreme than ln;

  square transformation $y' = y^2$, sometimes helps with left-skewed data;

  exponential transformation $y' = e^y$, sometimes helps with left-skewed data.

R for Quantile Plots

- The `lattice` library has the function `qqmath()` which can be used for normal quantile plots.
- Here is an example (plus signs are prompts for command over multiple lines).




```
> plot(qqmath(~log(ratio),
+ data = biomass,
+ type = c("p", "r"),
+ xlab = "Normal Quantiles"))
```