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.