Permutation Tests

Context

- So far, all of our tests have been in a setting where:
  - we have identified a statistical model we are confident is correct
  - we had a test statistic whose sampling distribution under H_0 we could obtain either
    * analytically
    * via a large-sample approximation
- What if we are not confident we have a good statistical model, or we can't derive the sampling disribution of our statistic, and our sample size is small?
  - Use computational/sampling approaches to approximate the sampling distribution.
  - Many variations on this idea; here we will discuss permutation tests for paired data.

Example

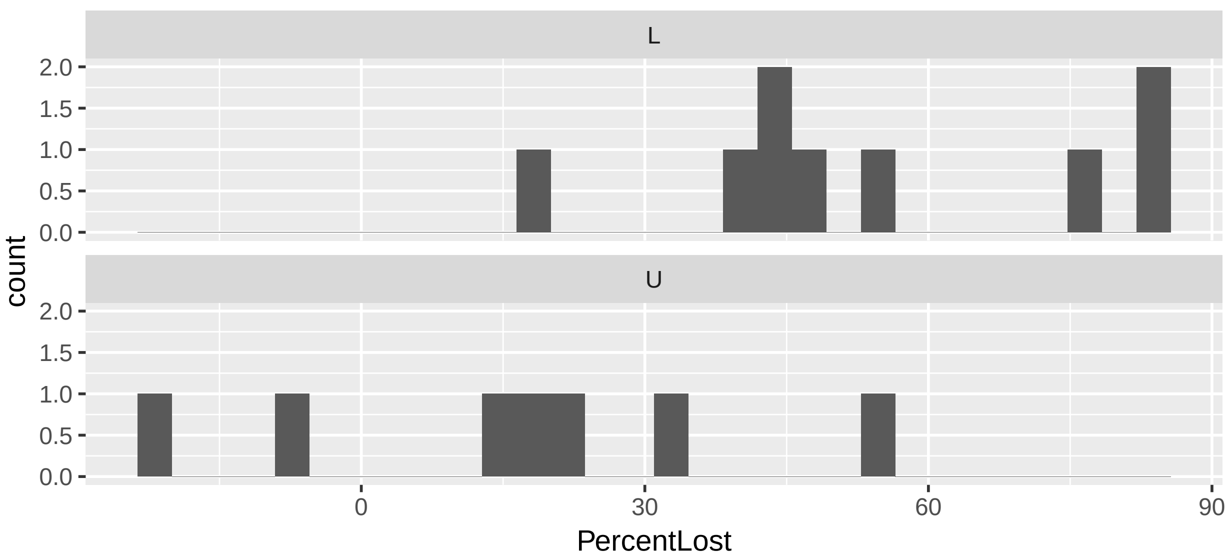
It is sometimes claimed that logging can help forests recover more quickly after forest fires. Is there evidence for this claim?

Here's a quote from the Statistical Sleuth (Ramsey and Schafer, 2013) describing our data:

> The 2002 Biscuit Fire in southwest Oregon provided a test case. Researchers selected 16 fire-affected plots in 2014 – before any logging was done – and counted tree seedlings along a randomly located transect pattern in each plot. They returned in 2005, after nine of the plots had been logged, and counted the tree seedlings along the same transects. (Data from D.C. Donato et al., 2006. "Post-Wildfire Logging Hinders Regeneration and Increases Fire Risk," *Science*, 311: 352.) The numbers of seedlings in the logged (L) and unlogged (U) plots are [loaded in the R code below].

```
logging <- read_csv("http://www.evanlray.com/data/sleuth3/ex0429_logging.csv")

ggplot(data = logging, mapping = aes(x = PercentLost)) +
  geom_histogram() +
  facet_wrap(vars(Action), ncol = 1)
```



Let μ_1 = mean percent lost in "population" of plots that are logged.

Let μ_2 = mean percent lost in "population" of plots that are unlogged.

Test

$H_0 : \mu_1 = \mu_2$ vs $H_A : \mu_1 \neq \mu_2$

Option 1: Likelihood Ratio Test based on parametric model

- If we assume that $X_{1i} \overset{\text{i.i.d.}}{\sim} \text{Normal}(\mu_1, \sigma^2)$ and $X_{2i} \overset{\text{i.i.d.}}{\sim} \text{Normal}(\mu_2, \sigma^2)$, we can derive a likelihood ratio test based on t distributions.
- If we don't trust that the data are normally distributed, this is risky (we have seen that t-based methods can fail with moderate sample sizes if conditions are not met).

Option 2: Large-sample χ^2 approximation to sampling distribution of likelihood ratio

- 15 is not large, this is a worse idea than #1

Option 3: Permutation test

- Easier to motivate with a slight modification to the hypotheses:

H_0: The distributions of percent lost are the same whether or not logging is done. (In particular, $\mu_1 = \mu_2$)

H_A: The distributions of percent lost are different depending on whether or not logging is done. (In particular, $\mu_1 \neq \mu_2$)

- Our test statistic will be the difference in means: $W = \bar{X}_1 - \bar{X}_2$.
- To calculate a p-value, we need an estimate of the sampling distribution of W under the condition that H_0 is true.
- Key ideas:
  - If H_0 is true, the observed percent lost for our plots would have been equally likely to be observed in either the logged or unlogged plots.
  - We will simulate many data sets that might have been observed if H_0 was true by permuting assignments of percent lost to different plots.

1. Allocate storage space for difference in group means from *nsims* different samples
2. For $i = 1, \dots, nsims$:
   a. Permute the assignments of observed values to groups
   b. Calculate the difference in group means, save in allocated space
3. Calculate *approximate* p-value as proportion of simulated samples with a difference in group means at least as large as our observed difference in group means