Is Free Beer Good For Tips?

Goal

The goal of this project is to test a hypothesis using a variety of techniques: "eyeball" test, t-test, and bootstrapping. Hypothesis testing with p-values is the inverse problem of confidence intervals. Hypothesis testing examines the area outside the, say, 95% interval; i.e., the α=5% significance level.

Discussion

Here is a typical statistics question (derived from one by Jeff "The Hammer" Hamrick) that we will solve in multiple ways.

Q. *Psychologists studied the size of the tips in a restaurant on a given day when the waitron gave the patron a free beer. Here are tips from 20 patrons, measured in percent of the total bill: 20.8, 18.7, 19.1, 20.6, 21.9, 20.4, 22.8, 21.9, 21.2, 20.3, 21.9, 18.3, 21.0, 20.3, 19.2, 20.2, 21.1, 22.1, 21.0, and 21.7. Does a beer-inspired tip exceed 20 percent or perhaps dip below 20 percent (maybe patrons get drunk and can't do math)? Use a significance level equal to α = 0.06.*

[Always pick the α significance level before you run your experiment. It is really bad mojo to pick your significance after you know what the p-value is.]

Before starting on this exercise, let's interpret the question. It asks whether the mean of the given experimental sample differs significantly from the usual 20% tip. By "significantly" we mean "statistically distinguishable" not "a lot." By "usual" we mean our *control* situation in which customers tip according to a normal distribution distribution centered at $\mu = 20$ with variance σ^2. Formally,

$H_0 : m = 20.0$ (non-free beer situation)
$H_1 : m \neq 20.0$ (free beer situation; two-sided alternative hypothesis)

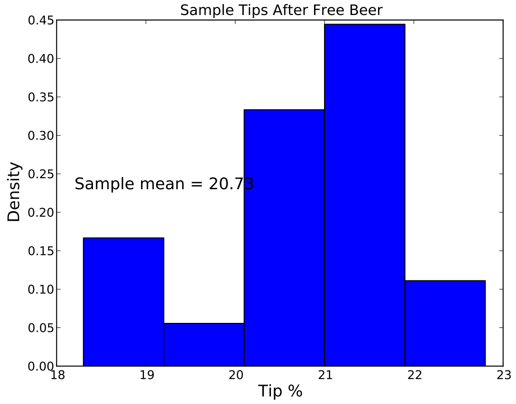
We could also say that $H_0 : m - \mu = 0$ and $H_1 : |m - \mu| > 0$.

On a typical day before running this experiment, we would expect tips to bounce to the left and right of the population mean of 20.0 with some variance σ^2. The average on a given day would therefore follow distribution $N(20.0, \sigma^2/n)$ for, let's say, a fixed n customers per day. The question is, does this particular experiment's sample mean, $m = 20.725$, fall outside of the typical variability of the sample means? **Note that we are comparing sample means not tips**. The **null hypothesis** is that the mean for the specified sample does not differ significantly from $\mu = 20.0$. The **alternate hypothesis** is that the sample mean differs significantly above or below the population mean.

Steps

Eyeballing it

1. First, just draw a histogram of the tips to see what it looks like. I used `plt.hist(tips, bins=5, normed=1)`. For this exercise, create a file called *userid*`-hyp/hist.py`.

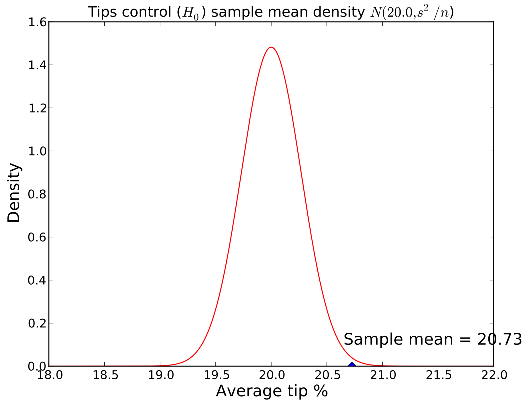


For your convenience, here are the tips in python format:

```
tips = [20.8, 18.7, 19.1, 20.6, 21.9, 20.4, 22.8,
        21.9, 21.2, 20.3, 21.9, 18.3, 21.0, 20.3,
        19.2, 20.2, 21.1, 22.1, 21.0, 21.7]
```

To me, there is a lot of "mass" to the right of the usual 20% tip but my eyeball is not a rigorous significance mechanism.

2. To get a better idea, let's plot the distribution of the *sample means* given our H_0 assumption: $N(20.0, s^2/n)$. (We can use the sample variance s^2 from our experimental sample for σ^2 because we don't know the variance of the original distribution. It safe to assume that the variance is similar.) This is our "control" or the usual tipping distribution: the distribution of the average tips per day if H_0, the control, is true. Please use file *userid*`-hyp/sample-mean-dist.py`.



Looking at that graph, it seems that a sample mean of 20.73 is pretty far in the right tail of a normal curve centered at the control average 20% tip. It looks to be a few standard deviations away from the mean. My gut says that it's pretty likely that giving people a free beer increases tips significantly.

t-test

1. Let's use a *t-test* now to test for significance, just like we would do in statistics class. The t value measures the number of standard deviations a sample mean, m, is away from our presumed population mean μ:

$$t = \frac{m - \mu}{s/\sqrt{N}} \qquad \text{(t-value)}$$

It's just the difference between the means scaled to be in units of standard deviations. Write some code to compute the t-value. When computing s, the sample standard deviation, note that the numpy `std()` function returns a biased estimate of the standard deviation. Use `np.std(tips, ddof=1)` instead of just `std(tips)`. Please create file *userid*`-hyp/t-test.py`.

2. Print out the value of t. I get $t = 2.69417199392$. That means that m is about 2.7 standard deviations away from μ, which is a very significant departure.

3. To get a p-value, the likelihood that we would see such a t value in the nonfree beer situation, look up t in a t-distribution c.d.f. using `1-scipy.stats.t.cdf(t,N-1)`. (You might need to install scipy.) You should get 0.0071844. Since we need to check both tails, the probability is actually 2× that, or, p-value=0.014369 (1.4%). The definition of significance is $\alpha = 0.06$, which means that our sample mean is definitely a significant departure from the control mean 20.0 since $1.4\% < 6\%$. There is only a 1.4% chance that the control could generate a value that extreme or beyond. Here is my program output:

```
t is 2.6941720
one-sided p-value is 0.0071844
two-sided p-value is 0.0143687
p-value (from t-test) = 0.014369, Reject H0
```

We must conclude that m differs significantly from $\mu = 20.0$ based upon the significance of $\alpha = 0.06$ and, therefore, we reject H_0 in favor of H_1. Giving out free beers is likely to have increased the average tip in that experiment. Again, the p-value doesn't say anything about the magnitude of the difference, only that they are statistically distinguishable. An average tip of 20.7 may not be a huge increase but giving free beer does increase tip size.

Boostrapping for empirical hypothesis testing

Ok, now, let's use bootstrapping to estimate a *p-value*. Just to hammer it home, a p-value for some point statistic is the probability that the control (null hypothesis H_0) could generate that statistic. In our case, a p-value can tell us the likelihood that tips drawing from a normal distribution centered around $\mu = 20.0$ with $s^2 = var(tips)$ could result in a daily sample mean of 20.725. We are sampling from $N(\mu = 20.0, s^2)$ to conjure up samples from the control situation. We are not resampling from the tips list as we are trying to see how the observed sample mean, 20.725, fits within the control distribution not the test distribution. *We are also not generating samples from the distribution of the sample mean random variable,* $N(\mu = 20.0, s^2/n)$.

We know that the sample mean must follow distribution $N(20.0, \sigma^2/n)$, but there are point statistics where the central limit theorem does not apply. That motivates our examination of bootstrapping for empirical p-values. We know that the central limit theorem applies to the sample mean and so we could go directly to the $N(20.0, \sigma^2/n)$ distribution in our simulation. For other point statistics, we might need bootstrapping.

Please use file *userid*`-hyp/bootstrap.py`.

1. Bootstrap TRIALS=5000 samples of size $n = len(tips)$ from $N(\mu, s^2)$ using `numpy.random.normal()`. It's very important that we use the same sample size as $len(tips)$ so we are comparing the same thing. Compute the mean of each sample, X, and add to $\overline{X}$ as you generate samples from the normal distribution.

2. Compute how many means in $\overline{X}$ are greater than or equal to `mean(tips)`:

```
greater = np.sum(X_ >= np.mean(tips))
```

or

```
greater = sum([x>=np.mean(tips) for x in X_]) # the number of true values
```

3. The (one-sided) p-value is just the ratio of values above the observed mean, mean(tips), to the number of trials. Double that because we're doing a two-sided test. With 5000 trials, I see around 13 values greater than $m = 20.725$. That gives us a p-value of $2 * \frac{13}{5000} = 0.0052$ or .52%. That means that, empirically, we find that there is an extremely small probability that the control could generate an extreme value like $m = 20.725$. Certainly the likelihood is less than the required 6% significant value. Your output should look like:

```
observed mean = 20.725
num greater than mean(tips) = 14
p-value from bootstrapping (ratio of X_ >= mean(tips)) is 0.0056
```

Note: we would expect the empirical p-value (.52%) and the p-value derived from the t-test (1.4%) to be very close to each other when the number of trials is large with bootstrapping. Statistician Jeff Hamrick explains that the difference is not a problem with our bootstrapping solution and is ok.

"A student t distribution with dof=19, is pretty close to a normal. But the differences are most greatly felt in the tails, and we're in the tails (rejection H_0), thus casting a little bit of sketchiness or your choice to draw the simulated raw data from a normal random variable. If we were performing this exact same operation on a data set with reasonably large size (say, 40 or 50 or 75) the differences would still exist but would be even more minute."

Again, we easily reject the control and conclude that giving out free beers increases tips.

Deliverables. *userid*`-hyp/hist.py`, *userid*`-hyp/sample-mean-dist.py`, *userid*`-hyp/t-test.py`, and *userid*`-hyp/bootstrap.py`.