Hypothesis Testing (Part I)

A. Hypothesis Testing

Hypothesis testing is concerned with making decisions using data. To make decisions using data, we need to characterize the kinds of conclusions we can make. Classical hypothesis testing is concerned with deciding between two decisions (things get much harder if there's more than two). The first, a null hypothesis is specified that represents the status quo. This hypothesis is usually labeled, H_0. This is what we assume by default. The alternative or research hypothesis is what we require evidence to conclude. This hypothesis is usually labeled, H_a or sometimes H_1 (or some other number other than 0).

So to reiterate, the null hypothesis is assumed true and statistical evidence is required to reject it in favor of a research or alternative hypothesis

B. Example

A respiratory disturbance index (RDI) of more than 30 events / hour, say, is considered evidence of severe sleep disordered breathing (SDB). Suppose that in a sample of 100 overweight subjects with other risk factors for sleep disordered breathing at a sleep clinic, the mean RDI was 32 events / hour with a standard deviation of 10 events / hour.

We might want to test the hypothesis that

$$H_0: \mu = 30$$

versus the hypothesis

$$H_a: \mu > 30$$

where μ is the population mean RDI. Clearly, somehow we must figure out a way to decide between these hypotheses using the observed data, particularly the sample mean.

Before we go through the specifics, let's set up the central ideas.

C. Types of errors in hypothesis testing

The alternative hypotheses are typically of the form of the true mean being $<$, $>$ or $\neq$ to the hypothesized mean, such as $H_a: \mu > 30$ from our example. The null typically sharply specifies the mean, such as $H_0: \mu = 30$ in our example.

Note that there are four possible outcomes of our statistical decision process:

Truth	**Decide**	**Results**
H_0	H_0	Correctly accept null
H_0	H_a	Type I error
H_a	H_a	Correctly reject null
H_a	H_0	Type II error

We will perform hypothesis testing by forcing the probability of a Type I error to be small.

D. Discussion relative to court cases

Consider a court of law and a criminal case. The null hypothesis is that the defendant is innocent. The rules requires a standard on the available evidence to reject the null hypothesis (and the jury to convict). The standard is specified loosely in this case, such as convict if the defendant appears guilty "Beyond reasonable doubt". In statistics, we can be mathematically specific about our standard of evidence.

Note the consequences of setting a standard. If we set a low standard, say convicting only if there circumstantial or greater evidence, then we would increase the percentage of innocent people convicted (type I errors). However, we would also increase the percentage of guilty people convicted (correctly rejecting the null).

If we set a high standard, say the standard of convicting if the jury has "No doubts whatsoever", then we increase the the percentage of innocent people let free (correctly accepting the null) while we would also increase the percentage of guilty people let free (type II errors).

E. Building up a standard of evidence

Consider our sleep example again. A reasonable strategy would reject the null hypothesis if the sample mean, $\overline{X}$, was larger than some constant, say C. Typically, C is chosen so that the probability of a Type I error, labeled α, is 0.05 (or some other relevant constant) To reiterate, α = Type I error rate = Probability of rejecting the null hypothesis when, in fact, the null hypothesis is correct

Let's see if we can figure out what C has to be. The standard error of the mean is $10/\sqrt{100} = 1$. Furthermore, under H_0 we know that $\overline{X} \sim N(30,1)$ (at least approximately) via the CLT. We want to chose C so that:

$$P(\overline{X} > C; H_0) = 0.05$$

The 95th percentile of a normal distribution is 1.645 standard deviations from the mean. So, if C is set 1.645 standard deviations from the mean, we should be set since the probability of getting a sample mean that large is only 5%. The 95th percentile from a N(30, 1) is:

$$C = 30 + 1 \times 1.645 = 31.645$$

So the rule "Reject H_0 when $\overline{X} \geq 31.645$" has the property that the probability of rejection is 5% when H_0 is true.

In general, however, we don't convert C back to the original scale. Instead, we calculate how many standard errors the observed mean is from the hypothesized mean

$$Z = \frac{\overline{X} - \mu_0}{s/\sqrt{n}}$$

This is called a Z-score. We can compare this statistic to standard normal quantiles.

To reiterate, the Z-score is how many standard errors the sample mean is above the hypothesized mean. In our example:

$$\frac{32 - 30}{10/\sqrt{100}} = 2$$

Since 2 is greater than 1.645 we would reject. Setting the rule "We reject if our Z-score is larger than 1.645" controls the Type I error rate at 5%. We could write out a general rule for this alternative hypothesis as reject whenever $\sqrt{n}(\overline{X} - \mu_0)/s > Z_{1-\alpha}$ where α is the desired Type I error rate.

Because the Type I error rate was controlled to be small, if we reject we know that one of the following occurred:

the null hypothesis is false, we observed an unlikely event in support of the alternative even though the null is true, our modeling assumptions are wrong. The third option can be difficult to check and at some level all bets are off depending on how wrong we are about our basic assumptions. So for this course, we speak of our conclusions under the assumption that our modeling choices (such as the iid sampling model) are correct, but do so wide eyed acknowledging the limitations of our approach.

F. T test in R

Let's try the t test on the pairs of fathers and sons in Galton's data.

```
library(UsingR); data(father.son)
t.test(father.son$sheight - father.son$fheight)
##
##  One Sample t-test
##
## data:  father.son$sheight - father.son$fheight
## t = 11.789, df = 1077, p-value < 2.2e-16
## alternative hypothesis: true mean is not equal to 0
## 95 percent confidence interval:
##  0.8310296 1.1629160
## sample estimates:
## mean of x
## 0.9969728
```