# *Exercises 8 · Parametric inference*

**Due Monday, March 30, 2015**

*(1) Chi-squared tests*

Consider the data at right, from a hypothetical clinical trial. The cell counts are the number of patients that fell into each category. The question at issue is: is a patient's clinical outcome independent of the treatment regime?

| | Clinical outcome | | |
|---|---|---|---|
| Regime | Good | Average | Bad |
| Treatment | 203 | 198 | 145 |
| Control | 156 | 221 | 162 |

(A) Under the Neyman–Pearson testing framework, use a chi-squared test to assess whether the null hypothesis—that there is no difference between treatment and control—can be rejected at the $\alpha = 0.05$ level. Yes, you will have to look the details of this test up yourself, which will require a bit of nerd-work. Specifically, I am asking for Pearson's chi-squared test.[1] No, I haven't taught you this explicitly. Luckily, you are highly resourceful, and have access to the OpenIntro: Stats book (to say nothing of the web). And just as luckily, you also understand the unifying framework of all Neyman–Pearson hypothesis tests.[2] Make sure you show each step of the Neyman–Pearson test. Be very explicit about the test statistic, its sampling distribution under $H_0$, and so forth.

(B) Now try a permutation test instead. A key aspect in getting this test correct is to recognize that the row totals of this table—that is, the number of people assigned to the treatment and the control—are fixed by design. That is, we run a Monte Carlo procedure where, in each iteration, you fix the row totals (that is, the number of treatments and controls), and randomly assign each person a clinical outcome in proportion to the marginal column totals. For example, there are 546 treatment subjects and 539 control subjects in the data set; likewise, there are 359 Good outcomes, 419 Average outcomes, and 307 Bad outcomes. Therefore a single Monte Carlo iteration will involve the following steps:

```
treat=rmultinom(1,size=546,prob=c(359,419,307))
control = rmultinom(1,size=539,prob=c(359,419,307))
mytab = t(cbind(treat,control))
```

Those are random multinomial draws in the first two lines, where the probabilities of the categories are automatically normalized to sum to one.[3] Use an appropriate test statistic to measure dependence across

[1] This Pearson (Karl) is the father of the Pearson after whom Neyman–Pearson (Egon) tests are named.

[2] Why do I make you do this? Well, there will be times in your professional life where someone asks you a question such as, "Did you try the Fixitol-Wagamama test?" You should be unashamed to admit that you have no idea what a Fixitol-Wagamama test is. Frankly, this happens to me all the time. But you should also be able to look up the details and understand the context, assumptions, and so forth of a previously unknown (to you) test. Time to practice.

[3] `http://en.wikipedia.org/wiki/Multinomial_distribution`

rows and columns of each random table; you might use the same chi-squared statistic used above, or maybe something else seems sensible to you.[4] For whatever test statistic $t$ you choose, conduct a Neyman–Pearson test of the same null hypothesis, at $\alpha = 0.05$. Briefly compare this result to that from Part A above.

[4] You might find this interesting: `http://arxiv.org/pdf/1108.4126v2.pdf`.

*(2) The power function of a test*

Consider the following (highly stylized) hypothesis testing probem. Suppose that we observe a binomial random variable $X \sim \text{Binomial}(n, w)$ for sample size $n$ and success probability $w$. For example, $X$ might be the number of heads we see in in $n$ flips of a coin. The goal is to test whether the null hypothesis that $w = 0.5$ is plausible in light of the data.

(A) Suppose that $n = 100$. What should our rejection region $R$ be if we want to conduct a Neyman–Pearson test at the $\alpha = 0.05$ level, using $X$ as a test statistic? Hint: the R function `qbinom` will help here. Try reading up on the help file using `?qbinom`

(B) Suppose that the true success probability is actually $w = 0.55$, so that $X \sim \text{Binomial}(n = 100, w = 0.55)$. How likely is $X$ to fall in the rejection region you calculated in Part A? This quantity $P(X \in R \mid w = 0.55)$ is called the *power* of the test at the alternatve hypothesis $w = 0.55$, because it is the chance we will reject the null hypothesis (which we want to do when $w = 0.55$, because the null hypothesis is false.) Note: you can calculate this using probability theory, but you might find it easier to approximate it using Monte Carlo simulation; R's `rbinom` function will allow you to simulation binomial random variables if you decide to take this route.

(C) Now consider range of alternative $w$ values spanning the interval $(0, 1)$. For example, you can create a sequence $0, 0.01, 0.02, \ldots, 0.99, 1)$ using the R function `seq(0,1,by=0.01)`. For each $w$ in this range, calculate the power $P(X \in R \mid w)$ and plot the power as a function of $w$. This is called a power curve or a power function.

(D) Use the technique you developed in Part C to determine how big (approximately) the sample size must be in order for you to have an 80% chance of rejecting the null hypothesis of $w = 0.5$ if the truth is $w = 0.55$.

*(3) Uncertainty quantification using normality*

Suppose, as Gauss did 200 years ago, that the residuals in a simple linear regression model follow a normal distribution with mean 0 and variance $\sigma^2$:

$$\begin{aligned} y_i &= \beta_0 + \beta_1 x_i + \epsilon_i \\ \epsilon_i &\sim N(0, \sigma^2)\,. \end{aligned}$$

As an aside, to connect this to some terms you've met before: this means we have specified a normal likelihood for the data, given the parameters. That is, for a data set of size $n$,

$$L(\beta_0, \beta_1) = \prod_{i=1}^{n} \phi(y_i - \beta_0 - \beta_1 x_i \mid 0, \sigma^2)\,,$$

where $\phi(t \mid m, v)$ is the notation typically used to denote the probability density function of the normal distribution having mean $m$ and variance $v$, evaluated at the point $t$.

Your task: use these assumptions to derive explicit formulas[5] for the mean and variance of the sampling distribution for the maximum likelihood estimators of $\beta_0$ and $\beta_1$. Recall that these are the same as the least-squares estimators:

$$\begin{aligned} \hat{\beta}_1 &= \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{n}(x_i - \bar{x})^2} \\ \hat{\beta}_0 &= \bar{y} - \hat{\beta}_1 \bar{x}\,. \end{aligned}$$

That is, there are four important numbers here: the mean and variance for the sampling distribution of two different quantities. You should give these expressions in terms of the observed data, the sample size, and the unknown parameters $\beta_0$, $\beta_1$, and $\sigma^2$. Remember that the design points $x_i$ are given—that is, they are constant, not random. The only source of randomness is the residuals $\epsilon_i$.

If you feel stuck, try the following warm-up problem first. Suppose that you observe $n$ data points $y_i \sim N(\theta, \sigma^2)$, where $\sigma^2$ is assumed known, and where $\theta$ is some unknown mean common to all the observations. You choose to estimate $\theta$ using the estimator $\hat{\theta} = \bar{y}$, the sample mean of the observations. We gave expressions in class for both the mean and variance of the sampling distribution of $\hat{\theta}$. (These expressions involve both the sample size and the unknown parameters.) Prove and/or re-derive them. You will find the material starting on page 7 of the "Random Variables" notes helpful here.

If you can do this warm-up problem but still feel stuck on the main problem involving the least-squares estimator, try assuming that $\beta_0 = 0$.

[5] These formulas are given in Chapter 5, page 131 of the course packet. Your job is to derive them, i.e. to proove that these formulas are correct.