Stat 200: Introduction to Statistical Inference **Winter 2023**

Lecture 11: More tests, and some normal theory

Lecturer: Dominik Rothenhäusler *February 22*

Note: *These lecture notes were written by Art Owen. If you like the material, he gets the credit! These notes have not been subjected to the usual scrutiny reserved for formal publications. They are meant as a memory aid for students who took stat 200 at Stanford University. They may be distributed outside this class only with the permission of Art Owen. Also, Stanford University holds the copyright.*

Abstract

These notes are mnemonics about what was covered in class. They don't replace being present or reading the book. Reading ahead in the book is very effective.

This lecture was about adding a few more things about hypothesis testing. Then we started looking at some theory of distributions derived from the normal distribution in order to prepare for two sample tests and regression coming next week.

11.1 Poisson dispersion test

Suppose that we have independent random variables X_i that are counts of some phenomenon. The example I gave in class is that X_i is the number of flaws on a silicon wafer being used to make integrated circuits. There are a lot of locations on the wafer where a flaw might occur and (hopefully) most locations don't have flaws. That suggests that a Poisson distribution might be reasonable.

We can take our null hypothesis to be that X_i are independent $\mathrm{Poi}(\lambda)$ random variables. However it is possible that not all wafers have the same distribution for the number of flaws. Maybe some are handled on old equipment and some new. Maybe different people handle them. Maybe they came from different suppliers. Maybe Monday morning work is different from Thursday afternoon. An alternative hypothesis H_A is that $X_i \sim \mathrm{Poi}(\lambda_i)$ independently but the λ_i are not all equal. As usual with the GLRT we make H_1 the union of H_0 and H_A. So under H_1 $X_i \sim \mathrm{Poi}(\lambda_i)$ for any $\lambda_i \geqslant 0$.

We we already know the MLE under H_0. It is $\hat{\lambda} = \bar{X}$. The likelihood under H_1 is

$$\prod_{i=1}^{n} \frac{e^{-\lambda_i}\lambda_i^{x_i}}{x_i!}.$$

To maximize it, we can just maximize each factor. The i'th factor is $e^{-\lambda_i}\lambda_i^{x_i}/x_i!$. The log of that is $-\lambda_i + x_i \log(\lambda_i)$. Taking the derivative with respect to λ_i yields $-1 + x_i/\lambda_i$. We set that to zero getting $\hat{\lambda}_i = x_i$. The second derivative is $-x_i/\lambda_i^2$ which is never positive, so we have maximized the likelihood not minimized it. We then realize that we already knew this: we can think of $\hat{\lambda}_i$ as the parameter for a sample of just one observation which equals x_i. The MLE would be the average of that one observation which is itself.

So, going in slow motion,

$$\begin{aligned}\Lambda &= \frac{\text{Max likelihood under } H_0}{\text{Max likelihood under } H_1} \\ &= \frac{\prod_{i=1}^n e^{-\hat{\lambda}} \hat{\lambda}^{x_i}/x_i!}{\prod_{i=1}^n e^{-\hat{\lambda}_i} \hat{\lambda}_i^{x_i}/x_i!} \\ &= \frac{\prod_{i=1}^n e^{-\hat{\lambda}} \hat{\lambda}^{x_i}}{\prod_{i=1}^n e^{-\hat{\lambda}_i} \hat{\lambda}_i^{x_i}} \\ &= \frac{\prod_{i=1}^n e^{-\bar{x}} \bar{x}^{x_i}}{\prod_{i=1}^n e^{-x_i} x_i^{x_i}} \\ &= \frac{\prod_{i=1}^n \bar{x}^{x_i}}{\prod_{i=1}^n x_i^{x_i}},\end{aligned}$$

because the product of e^{-x_i} is $e^{-\sum_i x_i}$ which equals $e^{-n\bar{x}}$. Now

$$-2\log(\Lambda) = 2\sum_{i=1}^n x_i \log(x_i/\bar{x}) \doteq \frac{1}{\bar{x}} \sum_{i=1}^n (x_i - \bar{x})^2.$$

That last approximation comes from a Taylor expansion just like the one used to get $\sum_i (O_i - E_i)^2/E_i$ in the multinomial likelihood ratio test.

Now we see that this likelihood ratio test will reject for large values of (n times) the sample variance divided by the sample mean. The Poisson distribution has variance equal to the mean. If every X_i has a potentially different λ_i then that will increase the variance of the X_i so the criterion makes intuitive sense.

This is an odd example because the null hypothesis has one free parameter while the alternative has n. So from the usual GLRT we anticipate $n-1$ degrees of freedom. However the GLRT was given for a degrees of freedom quantity that does not depend on n.

Let's investigate non-rigorously with the usual approach of putting a CLT in the numerator and the LLN in the denominator. We know from the LLN that $\bar{x} \to \mathbb{E}(X) = \lambda$ under H_0. The numerator is n times the average of n squared deviations. Let's plug $\bar{x} = \lambda$ into the numerator as well. Then our statistic is roughly

$$\frac{1}{\lambda} \sum_{i=1}^n (x_i - \lambda)^2.$$

The expected value of $(X-\lambda)^2$ is the variance of X which is λ. So our statistic should have expectation roughly $n\lambda/\lambda = n$. If we had not plugged in $\bar{x} = \lambda$ for the numerator, we would have gotten $n-1$ (after lengthier algebra). So a first approximation is that

$$-2\log\Lambda \approx N(n, n * \mathrm{Var}(X^2)).$$

Then

$$\frac{-2\log\Lambda - n}{\sqrt{n}} \approx N(0, \mathrm{Var}(X^2)).$$

In order to get a critical threshhold we now need to work out the variance of X^2 when $X \sim \mathrm{Poi}(\lambda)$. You can be sure that answer will depend on λ. We would then plug in $\hat{\lambda}$.

11.2 One and two sided tests

Suppose $\theta \in \mathbb{R}$ and H_0 is that $\theta = \theta_0$. If H_A is $\theta > \theta_0$ we can use the likelihood ratio

$$\frac{\text{Max likelihood under } \theta = \theta_0}{\text{Max likelihood under } \theta \geqslant \theta_0}.$$

Notice that the denominator has $H_0 \cup H_A$. If instead, H_A is $\theta < \theta_0$ we can use the likelihood ratio

$$\frac{\text{Max likelihood under } \theta = \theta_0}{\text{Max likelihood under } \theta \leqslant \theta_0}.$$

Very commonly θ is the mean of X and the rejection region turns out to be made up of large $\bar{X}$ for the first test above and small $\bar{X}$ for the second.

For instance if $\hat{\theta} \sim N(\theta, \sigma^2/n)$ then we would reject H_0 in favor of $\theta > \theta_0$ at $\alpha = 0.05$ if

$$\hat{\theta} > \theta_0 + \frac{1.645\sigma}{\sqrt{n}}.$$

The constant 1.645 comes from the normal distribution. We would have a p-value of

$$\Pr\Big(N\Big(\theta_0, \frac{\sigma^2}{n}\Big) > \hat{\theta}\Big) = \Pr\Big(N(0,1) > \frac{\hat{\theta} - \theta_0}{\sigma/\sqrt{n}}\Big).$$

In the above probability $\hat{\theta}$ is the fixed known observed value of the estimate and the randomness comes from $N(\theta_0, \sigma^2/n)$.

To reject at $\alpha = 0.05$ versus $\theta < \theta_0$ we would need

$$\hat{\theta} < \theta_0 - \frac{1.645\sigma}{\sqrt{n}}.$$

The p-value for this test is

$$\Pr\Big(N\Big(\theta_0, \frac{\sigma^2}{n}\Big) < \hat{\theta}\Big) = \Pr\Big(N(0,1) < \frac{\hat{\theta} - \theta_0}{\sigma/\sqrt{n}}\Big) = \Phi\Big(\frac{\hat{\theta} - \theta_0}{\sigma/\sqrt{n}}\Big),$$

where Φ is the $N(0,1)$ CDF. The test vs $\theta > \theta_0$ has

$$p = \Phi\Big(\frac{\theta_0 - \hat{\theta}}{\sigma/\sqrt{n}}\Big).$$

The above are one sided tests sometimes called one-tailed tests. The more usual two sided test rejects H_0 in favor of $H_A : \theta \neq \theta_0$ if

$$|\hat{\theta} - \theta_0| > \frac{1.96\sigma}{\sqrt{n}}$$

and has a p value of

$$\begin{aligned}
\Pr\Big(\big|\, N\Big(\theta_0, \frac{\sigma^2}{n}\Big) - \theta_0 \,\big| > |\hat{\theta} - \theta_0|\Big) &= \Pr\Big(|N(0,1)| > \frac{|\hat{\theta} - \theta_0|}{\sigma/\sqrt{n}}\Big) \\
&= \Pr\Big(N(0,1) > \frac{|\hat{\theta} - \theta_0|}{\sigma/\sqrt{n}}\Big) + \Pr\Big(N(0,1) < -\frac{|\hat{\theta} - \theta_0|}{\sigma/\sqrt{n}}\Big) \\
&= 2\Pr\Big(N(0,1) < -\frac{|\hat{\theta} - \theta_0|}{\sigma/\sqrt{n}}\Big) \\
&= 2\Phi\Big(-\frac{|\hat{\theta} - \theta_0|}{\sigma/\sqrt{n}}\Big).
\end{aligned}$$

This can turn out to be double the one sided p-value (depending on the side).

Here are three reasons why researchers might use a one sided test and corresponding p-value:

- they are sure that "if anything" $\theta > \theta_0$ and $\theta < \theta_0$ is simply not possible
- $\theta < \theta_0$ is inconsequential while $\theta > \theta_0$ would have important ramifications worth discovering, and
- they prefer a p-value half as large.

The third choice is not meant to be interpreted as a good practice. There is a danger that one could get for instance $p = 0.08$ and then be tempted to argue after the fact that one of the two other arguments is valid and switch to $p = 0.04$. The same dynamic can of course take place at α values other than 0.05. The two somewhat better reasons are still problematic. The p-value is meant to describe a result so unlikely that it is a surprise under H_0. Somebody might be wrong about which direction away from θ_0 is the only possible or only consequential one. Even a small probability of error there would be important.

11.3 Testing for normality

The skewness of a random variable X is

$$\gamma = \frac{\mathbb{E}((X-\mu)^3)}{\sigma^3}$$

when X has mean μ variance $\sigma^2 > 0$ and the numerator expectation exists. We interpret $\gamma > 0$ as X having a heavier right tail than left tail and $\gamma < 0$ has the opposite interpretation. A normal random variable has $\gamma = 0$.

For $n \geqslant 2$, we can estimate the skewness of X by

$$\hat{\gamma} = \frac{\frac{1}{n}\sum_{i=1}^n (X_i - \bar{X})^3}{s^3}$$

where $s^2 = (1/(n-1))\sum_{i=1}^n (X_i - \bar{X})^2$. Let $Z_i = (X_i - \mu)/\sigma$. The sample skewness of the Z_i is exactly the same as the one for the X_i (which we can compute). If $X_i \sim N(\mu, \sigma^2)$ then $Z_i \sim N(0,1)$. If we know the distribution of $\hat{\gamma}$ for $N(0,1)$ data then we know it for any normal data with $\sigma > 0$.

We might have to do a computer simulation to find the distribution of $\hat{\gamma}$.

Knowing that distribution we could reject normality if $|\hat{\gamma}|$ was unusually large. That test would not be very effective against an alternative distribution that was not normal but had $\gamma = 0$ anyway (like a uniform distribution has).

Another way to test for normality is to make up a statistic that describes how close to linear the QQ plot of $X_{(i)}$ versus $\Phi^{-1}(i/(n+1))$ is. The Shapiro-Wilk test does this. That test would easily detect that a uniformly distributed sample was not normal. It might not be as good as the test based on $\hat{\gamma}$ for some other alternatives, such as skewed distributions that $\hat{\gamma}$ is designed to detect.

Picking a test involves tradeoffs.

11.4 Normal distribution and friends

We will need some distributions related to the normal distribution. Those are what we will use in order to get tests and confidence intervals in problems described by normally distributed data. The highlight is that we will be able to get around the problem of infering on μ as if we knew σ in a $N(\mu, \sigma^2)$ model. Getting

around that is useful because if we don't know μ we ordinarily don't know σ either. Student's t test lets us plug in an estimate for σ and still get an exact inference.

If $X \sim N(\mu, \sigma^2)$ and $\sigma > 0$ then $Z = (X - \mu)/\sigma \sim N(0, 1)$. We proved this by writing $\Pr(Z \leqslant z)$ as an integral, plugging in $Z = (X - \mu)/\sigma$, simplifying and recognizing the $N(0, 1)$ CDF.

Next we saw that if $Z \sim N(0, 1)$ and $W = Z^2$ then $W \sim \chi^2_{(1)}$. This was given as the definition of the $\chi^2_{(1)}$ distribution. So we have to check that it matches the distribution as given in Appendix A of Rice. If $w < 0$ then of course $\Pr(W \leqslant w) = 0$ because Z^2 cannot be negative. For $w \geqslant 0$,

$$\begin{aligned}
\Pr(W \leqslant w) &= \Pr(Z^2 \leqslant w) \\
&= \int_{-\sqrt{w}}^{\sqrt{w}} \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}z^2}\,\mathrm{d}z \\
&= 2\int_{0}^{\sqrt{w}} \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}z^2}\,\mathrm{d}z, \qquad \text{let } v = z^2,\ \mathrm{d}v = 2z\mathrm{d}z,\ \mathrm{d}z = \mathrm{d}v/(2\sqrt{v}) \\
&= 2\int_{0}^{w} \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}v} \frac{1}{2} v^{-1/2}\,\mathrm{d}v, \\
&= \int_{0}^{w} \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}v} v^{-1/2}\,\mathrm{d}v.
\end{aligned}$$

Therefore the PDF of W is

$$\frac{\mathrm{d}}{\mathrm{d}w}\Pr(W \leqslant w) = \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}w} w^{-1/2}, \quad \text{for } w > 0.$$

This is proportional to $e^{-\frac{1}{2}w} w^{-1/2}$. So it has to be the $\mathrm{Gam}(1/2, 1/2)$ distribution. That is the same as the $\chi^2_{(1)}$ distribution from the book, so the W^2 definition matches what we had before.

If $U_1, \ldots, U_n$ are IID $\chi^2_{(1)}$, then $V = \sum_{i=1}^n U_i \sim \chi^2_{(n)}$ (where that distiribution means $\mathrm{Gam}(n/2, 1/2)$). We can prove this by using the MGF. A $\chi^2_{(1)}$ random variable has $M(t) = (1 - 2t)^{-1/2}$, and more generally, a $\chi^2_{(n)}$ random variable has $M(t) = (1 - 2t)^{-n/2}$ and if we add independent random variables then the MGF of the sum is the product of the individual variables' MGFs.

If $U_i \sim \chi^2_{(m_i)}$ (independently) then $\sum U_i \sim \chi^2_{(\sum m_i)}$. This follows from the MGFs. We can remember it by thinking of each U_i as a sum of m_i squared independent $N(0, 1)$ random variables.

The t ***distribution*** with n degrees of freedom is defined to be the distribution of

$$t = \frac{Z}{\sqrt{U/n}}$$

where $Z \sim N(0, 1)$ independently of $U \sim \chi^2_{(n)}$. The PDF of the $t_{(n)}$ distribution is

$$\frac{\Gamma((n+1)/2)}{\sqrt{n\pi}\Gamma(n/2)} \Big(1 + \frac{t^2}{n}\Big)^{-(n+1)/2}, \quad -\infty < t < \infty.$$

For the special case $n = 1$ we see that the PDF is proportional to $(1 + t^2)^{-1}$. The Cauchy PDF is $(1/\pi)(1 + t^2)^{-1}$ so $t_{(1)}$ has to be the Cauchy distribution. [Both on $-\infty < t < \infty$.]

If $U \sim \chi^2_{(m)}$ and $V \sim \chi^2_{(n)}$ independently then

$$F = \frac{\frac{1}{m}U}{\frac{1}{n}V}$$

has Fisher's $F_{m,n}$ distribution, with m numerator degrees of freedom and n denominator degrees of freedom.

We can tell from the definitions that $t_{(n)}^2$ is $F_{1,n}$.

The way we use these facts and definitions is through the following. If $X_1, X_2, \ldots, X_n$ are IID $N(\mu, \sigma^2)$ for $n \geqslant 2$ and $\sigma > 0$ and

$$\bar{X} = \frac{1}{n}\sum_{i=1}^{n} X_i \quad \text{and} \quad s^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \bar{X})^2,$$

then $\bar{X}$ is independent of s^2. This is: Rice Ch 6.3 Theorem A. He proves it via MGFs. It is worth reading how. What he shows is that the vector

$$(X_1 - \bar{X}, X_2 - \bar{X}, \ldots, X_n - \bar{X})$$

is independent of $\bar{X}$ so of course s^2 is independent of $\bar{X}$. That is pretty interesting. You compute these two things from the exact same set of data and they come out independent. That is a special property of the normal distribution.

Next he shows that $(n-1)s^2/\sigma^2 \sim \chi^2_{(n-1)}$ again via MGFs. We already know that $\bar{X} \sim N(\mu, \sigma^2/n)$. Now here is the punch line

$$\frac{\bar{X} - \mu}{s/\sqrt{n}} \sim t_{(n-1)}. \tag{11.1}$$

We can get this from the definition

$$\frac{\bar{X} - \mu}{s/\sqrt{n}} = \frac{\frac{\bar{X}-\mu}{\sigma/\sqrt{n}}}{s/\sigma} = \frac{\frac{\bar{X}-\mu}{\sigma/\sqrt{n}}}{\sqrt{s^2/\sigma^2}}.$$

The numerator has the $N(0,1)$ distribution. The denominator is $\sqrt{\chi^2_{(n-1)}/(n-1)}$ and they are independent. So the ratio has the $t_{(n-1)}$ distribution.

If we look at

$$\frac{\bar{X} - \mu}{s/\sqrt{n}}$$

we see a ratio with known values $\bar{X}$, s and n and one unknown value, μ. Ordinarily when you put together a formula mixing knowns and unknowns, you get an unknown. That happens here too. However the huge difference is that our unknown has a ***known distribution***. The ratio $(\bar{X} - \mu)/(s/\sqrt{n})$ is called a ***pivotal quantity*** because it contains the unknown but has a known distribution.

If we know the value Δ such that $\Pr(|t_{(n)}| > \Delta) = \alpha$, then

$$\Pr\Big(\Big|\frac{\bar{X} - \mu}{s/\sqrt{n}}\Big| \leqslant \Delta\Big) = 1 - \alpha.$$

We can rearrange this to get a confidence interval for μ,

$$\Pr\Big(\bar{X} - \frac{\Delta s}{\sqrt{n}} \leqslant \mu \leqslant \bar{X} + \frac{\Delta s}{\sqrt{n}}\Big) = 1 - \alpha.$$

We don't know σ. So we plug in s instead of σ. Now the ratio $(\bar{X} - \mu)/(s/\sqrt{n})$ ***does not*** have the $N(0,1)$ distribution. It has a wider distribution because we are using a random s instead of the true σ. The t distribution is exactly that wider distribution.

For a given α we know that

$$1 - \alpha = \int_{-\Delta}^{\Delta} f_{(n-1)}(t)\,\mathrm{d}t$$

where $f_{(n-1)}$ is the PDF of the $t_{(n-1)}$ distribution. That distribution is symmetric, so we know that

$$1 - \alpha/2 = \int_{-\infty}^{\Delta} f_{(n-1)}(t)\,\mathrm{d}t.$$

So we want $\Delta = F_{(n)}^{-1}(1-\alpha/2)$ where $F_{(n-1)}$ is the CDF of the $t_{(n-1)}$ distribution. In R this is `qt(1-alpha/2,n-1)`.