Lecture 27

We now construct a confidence interval for the unknown p **in the model** $X \sim \text{Binomial}(n, p)$**. We require** $np \geq 5$ **and** $n(1-p) \geq 5$ **(ie.** n **large and** p **moderate) so that we can use the approximation** $X \sim \text{Normal}(np, np(1-p))$**. We denote** $\hat{p} = X/n$ **and** $\hat{p}_{\text{obs}} = X_{\text{obs}}/n$**. A** $(1-\alpha)$**% confidence interval for** p **is obtained via:**

$$
\begin{aligned}
& \mathrm{P}\left(-z_{\frac{\alpha}{2}} < \frac{X - np}{\sqrt{np(1-p)}} < z_{\frac{\alpha}{2}}\right) = 1 - \alpha \\
\Leftrightarrow\ & \mathrm{P}\left(-z_{\frac{\alpha}{2}} < \frac{\hat{p} - p}{\sqrt{p(1-p)/n}} < z_{\frac{\alpha}{2}}\right) = 1 - \alpha \\
\Leftrightarrow\ & \mathrm{P}\left(-z_{\frac{\alpha}{2}}\sqrt{p(1-p)/n} - \hat{p} < -p < z_{\frac{\alpha}{2}}\sqrt{p(1-p)/n} - \hat{p}\right) = 1 - \alpha \\
\Leftrightarrow\ & \mathrm{P}\left(\hat{p} - z_{\frac{\alpha}{2}}\sqrt{p(1-p)/n} < p < \hat{p} + z_{\frac{\alpha}{2}}\sqrt{p(1-p)/n}\right) = 1 - \alpha
\end{aligned}
$$

Therefore,

$$\hat{p}_{\text{obs}} \pm z_{\frac{\alpha}{2}}\sqrt{\hat{p}_{\text{obs}}(1 - \hat{p}_{\text{obs}})/n} \tag{1}$$

is an approximate $(1-\alpha)100\%$ **CI for** p**. The CI (1) is based on two approximations:**

1. **approximating the Binomial with the Normal**
2. **replacing** p **with** $\hat{p}$

Example: From a sample of 1250 BC voters, 420 indicate that they support the NDP. Obtain an approximate 95% CI for the proportion of BC voters who support the NDP.

CI's based on the Student distribution: Suppose $X_1, \ldots, X_n$ are iid $\mathrm{Normal}(\mu, \sigma^2)$ where σ is unknown (the realistic case). It can be shown that

$$\frac{\bar{X} - \mu}{s/\sqrt{n}} \sim t_{n-1}$$

where t_{n-1} denotes the t distribution with $n-1$ degrees of freedom. The pdf of $Y \sim t_{n-1}$ is

$$f(y) = \frac{\Gamma(n/2)}{\Gamma((n-1)/2)\Gamma(1/2)}\left(1 + \frac{y^2}{n-1}\right)^{-n/2}$$

Here, the $(1-\alpha)$100% confidence interval for μ is

$$\bar{X} \pm t_{\frac{\alpha}{2},n-1}\frac{s}{\sqrt{n}}$$

Discussion points:

- **the t_{n-1} distribution is symmetric on $\mathcal{R}$**
- **the t_{n-1} has longer tails than the normal**
- **as $n \to \infty$, $t_{n-1} \to Z \sim \mathrm{Normal}(0,1)$**
- **for $n \geq 30$, you can replace t_{n-1} with Z**
- **the t distribution is intractable; no need to memorize pdf**
- **Table B.1 in the text gives points $t_{\frac{\alpha}{2},n-1}$**

The logic of hypothesis testing: We view the testing of hypotheses as consisting of three steps. We discuss the three steps in some detail.

1. **The experimenter forms a *null hypothesis* H_0 to test against an *alternative hypothesis* H_1.**
2. **The experimenter collects data.**
3. **In the inference step, the question is asked "Are the data compatible wrt H_0?" If yes, H_0 is not rejected. If no, H_0 is rejected.**

Example: In this informal example, we go over the three steps of hypothesis testing. Imagine a court of law where a defendent is accused of a crime.

Example: In this informal example, we go over the three steps of hypothesis testing. Imagine that you are playing cards and that your friend has obtained a royal flush three hands in a row.

In the inference step, if we answer "yes" to the key question (Are the data compatible wrt H_0?), we conclude using the curious language, "H_0 is not rejected". We discuss why this does not mean the same thing as "H_0 is accepted".