Lesson 12. Chi-Square Goodness-of-Fit Test

1 Using Excel's functions

- Last time: Inverse transform method
  - X is a continuous random variable with cdf F
  - U is a Uniform$[0,1]$ random variable
  - $F^{-1}(U)$ is a random variate generator for X
  - Generate observations of $U \Rightarrow$ generate observations of X
- Normal distribution with mean μ and standard deviation σ
  - $F^{-1}(u) =$ NORMINV(u, μ, σ)
- Gamma distribution with parameters α, β
  - Mean $\mu = \alpha\beta$, variance $\sigma^2 = \alpha\beta^2$
  - $\alpha = 1 \Rightarrow$ exponential distribution with rate parameter $\lambda = 1/\beta$ (mean β)
  - $F^{-1}(u) =$ GAMMAINV(u, α, β)
- Let's use Excel to generate
  - 10 normal random variates with mean 2 and standard deviation 0.5
  - 10 gamma random variates with parameters $\alpha = 2$ and $\beta = 3$

2 Chi-square goodness-of-fit test

- How do we know if a set of observations are in fact generated from a particular random variable?
- The **Chi-square goodness-of-fit test**: examines the expected number of observations in a family of subintervals to determine how closely the observations fit a particular distribution
- Works essentially the same way as the Chi-square test for uniformity
- Let F be a cdf of a random variable X
- Let $Y_1, \ldots, Y_n$ be n independent random variables
- Do $Y_1, \ldots, Y_n$ share a common cdf F?
- Divide real line into m subintervals: $[a_1, b_1], [a_2, b_2], \ldots, [a_m, b_m]$
- The null hypothesis for this test: if Y represents any of the Y_j's,

- Let $y_1, \ldots, y_n$ be observations of $Y_1, \ldots, Y_n$
- Let e_i = expected number of observations in interval $[a_i, b_i]$

    - Rule of thumb: set up intervals so $e_i \geq 5$ for $i = 1, \ldots, m$
- Let o_i = observed number of observations in interval $[a_i, b_i]$
- The observed test statistic is
- The p-value is
    - Small p-values ($< \alpha$, where α is typically 0.05, or even 0.01) $\Rightarrow$ reject H_0

3 Conducting the Chi-square goodness-of-fit test in Excel

- In the chi-square sheet in the Excel workbook for today's lesson, there are 100 numbers
- Are they from a Normal distribution with mean 2 and standard deviation 0.5?
- Use MIN and MAX function to determine interval of observation values
- Determine subintervals (don't forget $-\infty$ and $+\infty$ if applicable)
- Use FREQUENCY function to determine the number of observations in subinterval i
- Compute the expected number of observations in subinterval i
    - cdf F of Normal random variable with mean μ and standard deviation σ

$$F(x) = \text{NORMDIST(x, mean, stdev, TRUE)}$$

- Merge subintervals so that the expected number of observations ≥ 5 for each resulting subinterval
- Compute the observed test statistic and p-value

4 On your own

- Generate 100 exponentially distributed random variates with mean 1/2 (Remember that the gamma distribution with $\alpha = 1$ and β is the exponential distribution with mean β.)
- Use a Chi-square goodness-of-fit test to test how well your random variates fit with an exponential distribution with mean 1/2.