# 208dowels-bootstrap

B-MAT-400

# Frequency distribution

- A frequency distribution is a series that represents the frequency of various observations in an experiment.
- Example: number of times there is $x$ defective pieces among in 100 samples of 100 pieces
- Notation
  - $x$ is the observed class (number of defective pieces among 100 pieces)
  - $O_x$ is the observed size of the class (number of samples where there was $x$ defective pieces)
  - $N = \sum_x O_x$ is the number of total observations

| $x$ | 0 | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8+ | Total |
|---|---|---|---|---|---|---|---|---|---|---|
| $O_x$ | 2 | 7 | 14 | 21 | 19 | 17 | 11 | 5 | 4 | 100 |

# Distribution fitting

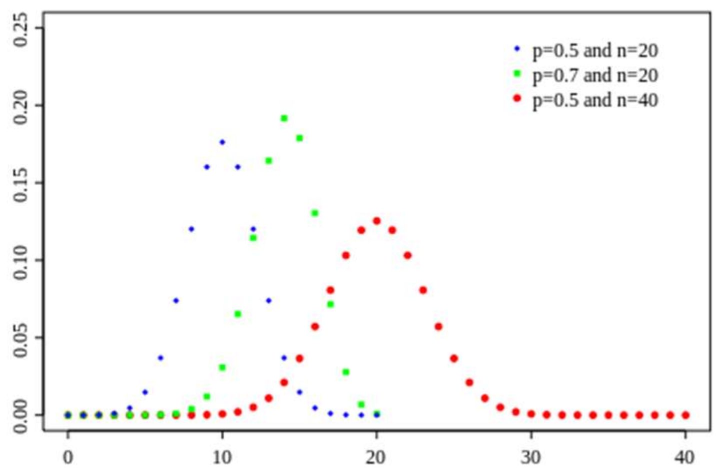
- Fitting observations to a probability distribution:
  - Normal distribution
  - Binomial distribution
  - Poisson distribution
  - …
- This allows to replace the values of the observations with only the parameters of the distribution (mean, variance, etc.)
- Notation:
  - $T_x$ is the theoretical size of the class

# Reminder: Binomial distribution

- $B(n, p)$ is the probability distribution of the number of successes in $n$ independent Bernouilli trials of probability $p$

$$P(X = k) = C_n^k p^k (1-p)^{n-k}$$

- $C_n^k$ is the binomial coefficient: $C_n^k = \frac{n!}{k!(n-k)!}$



# Binomial distribution fitting

- Finding $n$ and $p$ so that $B(n, p)$ is a good fit to our observations:

$$T_x = N \times C_n^x \, p^x (1-p)^{n-x}$$

- Example: Probability of having $x$ defective pieces in a sample
  - If each observed piece is a trial, then $n$ is the total number of pieces in a sample
  - The average number of defective pieces is $\bar{x} = \frac{1}{N}\sum_x xO_x$, so the probability that one piece is defective is:

$$p = \frac{\bar{x}}{n} = \frac{1}{nN}\sum_x xO_x$$

# Example

| $x$ | 0 | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8+ | Total |
|---|---|---|---|---|---|---|---|---|---|---|
| $O_x$ | 2 | 7 | 14 | 21 | 19 | 17 | 11 | 5 | 4 | 100 |

- $n = 100$
- $p = \frac{0\times2+1\times7+\cdots+7\times5+8\times4}{100\times100} = 0.0392$
- $T_x = 100 \times C_{100}^{x}\,(0.0392)^x(1-0.0392)^{100-x}$

| $x$ | 0 | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8+ | Total |
|---|---|---|---|---|---|---|---|---|---|---|
| $T_x$ | 1.8 | 7.5 | 15.1 | 20.1 | 19.9 | 15.6 | 10.1 | 5.5 | 4.3 | 100 |

- To ensure that $N = \sum_x T_x$, the last class size is computed by subtracting the other sizes to $N$.

# $\chi^2$ test

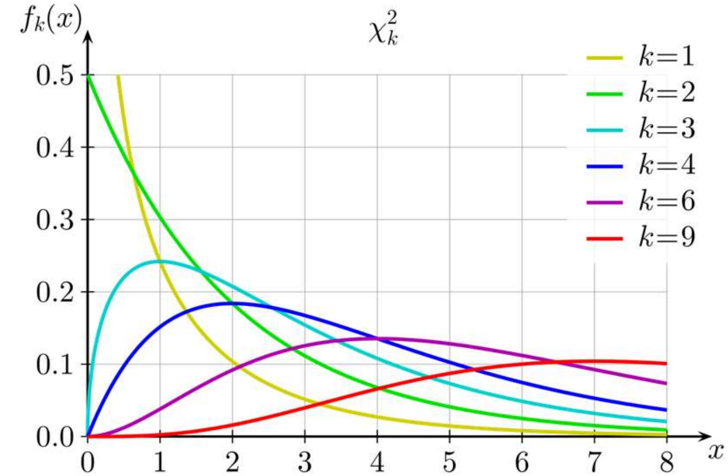
- How to evaluate the goodness of the fit?
- Measure of the deviation between observed theoretical sizes of the classes:

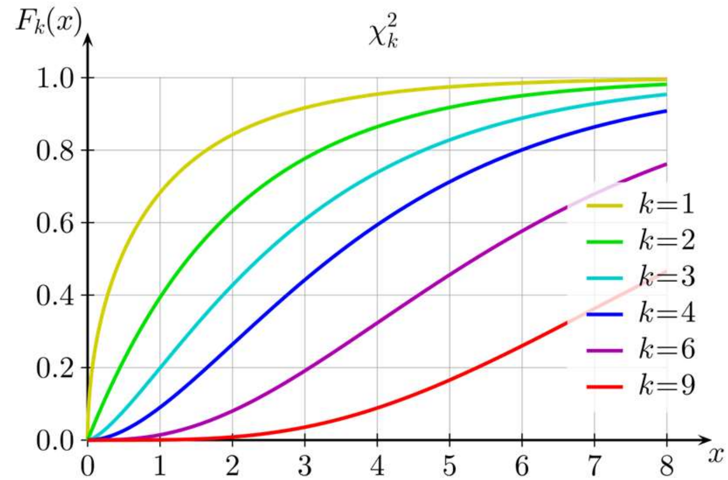
$$\chi^2 = \sum_x \frac{(O_x - T_x)^2}{T_x}$$

- If $\chi^2 = 0$, there is no deviation, and the fit is perfect
- The greater $\chi^2$ is, the worst the fit is.
- Classes must have a significant size, so they may need to be merged before computing $\chi^2$

# $\chi^2$ distribution

- The value of $\chi^2$ follows the $\chi^2$-distribution, which also depends on a parameter called the degrees of freedom
- Comparing the value of $\chi^2$ with the cumulative distribution function will indicate how likely the observed size will deviate from the theoretical sizes.



# Degrees of freedom

- $\nu$ is the number of independent classes $x$ when applying the probability distribution
- If $k$ is the number of classes and $p$ the number of constraints use to compute the fitting:
$$\nu = k - p$$
- In the binomial fitting, we fixed $N = \sum_x O_x = \sum_x T_x$ and we used $\bar{x}$ to compute $B(n, p)$, so we have:
$$\nu = k - 2$$

# $\chi^2$ table

- Knowing the value of $\chi^2$ and the degrees of freedom, we can use a $\chi^2$ table to determine the goodness of the fit

| $\nu$ | **99%** | **90%** | **80%** | **70%** | **60%** | **50%** | **40%** | **30%** | **20%** | **10%** | **5%** | **2%** | **1%** |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| **1** | 0.00 | 0.02 | 0.06 | 0.15 | 0.27 | 0.45 | 0.71 | 1.07 | 1.64 | 2.71 | 3.84 | 5.41 | 6.63 |
| **2** | 0.02 | 0.21 | 0.45 | 0.71 | 1.02 | 1.39 | 1.83 | 2.41 | 3.22 | 4.61 | 5.99 | 7.82 | 9.21 |
| **3** | 0.11 | 0.58 | 1.01 | 1.42 | 1.87 | 2.37 | 2.95 | 3.66 | 4.64 | 6.25 | 7.81 | 9.84 | 11.34 |
| **4** | 0.30 | 1.06 | 1.65 | 2.19 | 2.75 | 3.36 | 4.04 | 4.88 | 5.99 | 7.78 | 9.49 | 11.67 | 13.28 |
| **5** | 0.55 | 1.61 | 2.34 | 3.00 | 3.66 | 4.35 | 5.13 | 6.06 | 7.29 | 9.24 | 11.07 | 13.39 | 15.09 |
| **6** | 0.87 | 2.20 | 3.07 | 3.83 | 4.57 | 5.35 | 6.21 | 7.23 | 8.56 | 10.64 | 12.59 | 15.03 | 16.81 |
| **7** | 1.24 | 2.83 | 3.82 | 4.67 | 5.49 | 6.35 | 7.28 | 8.38 | 9.80 | 12.02 | 14.07 | 16.62 | 18.48 |
| **8** | 1.65 | 3.49 | 4.59 | 5.53 | 6.42 | 7.34 | 8.35 | 9.52 | 11.03 | 13.36 | 15.51 | 18.17 | 20.09 |
| **9** | 2.09 | 4.17 | 5.38 | 6.39 | 7.36 | 8.34 | 9.41 | 10.66 | 12.24 | 14.68 | 16.92 | 19.68 | 21.67 |
| **10** | 2.56 | 4.87 | 6.18 | 7.27 | 8.30 | 9.34 | 10.47 | 11.78 | 13.44 | 15.99 | 18.31 | 21.16 | 23.21 |

# Example

- First, we merge our classes to get sizes large enough

| $x$ | 0-1 | 2 | 3 | 4 | 5 | 6 | 7+ | Total |
|---|---|---|---|---|---|---|---|---|
| $O_x$ | 9 | 14 | 21 | 19 | 17 | 11 | 9 | 100 |
| $T_x$ | 9.3 | 15.1 | 20.1 | 19.9 | 15.6 | 10.1 | 9.8 | 100 |

- Then we compute $\chi^2 = \sum_x \frac{(O_x - T_x)^2}{T_x} = 0.45$
- We have $\nu = 7 - 2 = 5$ degrees of freedom
- Using the table, we can see that our fit is valid with a probability larger than 99%.

# 208dowels

- Goal: Compute a binomial fit for defective pieces and validate it with a $\chi^2$ test
- Inputs: sizes of the 9 observed classes
- Outputs:
  - Frequency distribution table with observed and theoretical sizes
  - Binomial distribution used for the fit
  - Value of $\chi^2$
  - Value of $\nu$
  - Probability range of the fit validity

# Exercise: Binomial distribution fit

- Given the observed sizes of classes and the number of total pieces per observed sample, compute $n$ and $p$ for the binomial distribution fit $B(n, p)$

# Exercise: Theoretical sizes

- Given a binomial distribution fit, compute the theoretical sizes of each class

$$T_x = N \times C_n^x \, p^x (1-p)^{n-x}$$

# Exercise: Classes merge

- Given the observed sizes of classes and a minimum size, return the list of merged classes so that each size is at least the minimum value.
- Smallest classes must be merged first

# Exercise: $\chi^2$

- Given the observed and theoretical sizes of merged classes, compute the value of $\chi^2$

$$\chi^2 = \sum_x \frac{(O_x - T_x)^2}{T_x}$$